Stanford Encyclopedia of Philosophy Abridged Table of Contents

Assigned Entries Only

Welcome to the Stanford Encyclopedia of Philosophy, which was designed from its inception (September 1995) as a *dynamic* reference work. In a dynamic reference work, each entry is maintained and kept up to date by an expert or group of experts in the field. These authors are given remote electronic access to copies of their entries on our server and they can update those copies any time the need arises. Moreover, all entries and updates are refereed by the members of a distinguished Editorial Board *before* they are made public. Whenever an author uploads a new entry or modifies an existing entry, the new material is stored off-line until it is approved by the Editorial Board member in charge of that entry. Consequently, our dynamic reference work is *responsive* to new research, for it can change at any time with the addition of new entries or the modification of existing entries. You can, however, cite fixed editions which are made on a quarterly basis and stored in our Archives. Thank you for your patience as our Encyclopedia develops. (Many of the assigned entries below have not yet been written.) See the Unabridged Table of Contents for the complete list of projected and assigned entries.

Search Encyclopedia	**Editorial Information**
What's New	**How to Cite This Encyclopedia**
Encyclopedia Archives	**Unabridged Table of Contents**

Navigation Panel:

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

A

B

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

C

D

E

F

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

G

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

H

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

I

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

J

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

K

L

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

M

N

O

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

P

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Q

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

R

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

S

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

T

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

U

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

V

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

W

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

X

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Y

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Z

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Editorial Information

Search Stanford Encyclopedia of Philosophy

WebGlimpse Search

String to search for: [ ] Submit

☐ Case Sensitive ☐ Partial Match ☐ Jump to Line

Misspellings Allowed: 0

Maximum number of files returned: 50

Maximum number of matches per file returned: 30

Maximum number of characters output per file: 10000

Stanford Encyclopedia of Philosophy Editorial Information

Basic Information

The *Stanford Encyclopedia of Philosophy* is a dynamic reference work and is a publishing project of the Metaphysics Research Lab at the Center for the Study of Language and Information (CSLI) at Stanford University. The concept of a dynamic reference work was implemented in the design of the Encyclopedia by Edward N. Zalta (Director of the Metaphysics Research Lab). The project began when John Perry was the Director of the Center for the Study of Language and Information. All correspondence should be directed to:

editors@plato.stanford.edu

Information about our dynamic reference work can be found in the following papers and abstracts:

- "The Stanford Encyclopedia of Philosophy: A Developed Dynamic Reference Work", by Colin Allen, Uri Nodelman, and Edward N. Zalta, forthcoming *Metaphilosophy* (285K PDF file); to be reprinted in *CyberPhilosophy: The Intersection of Philosophy and Computing*, James H. Moor and Terrell Ward Bynum, (eds.), Oxford: Blackwell, 2002
- "A Solution to the Problem of Updating Encyclopedias", by Eric Hammer and Edward N. Zalta *Computers and the Humanities*, **31**/1 (1997): 47-60 [Note: The ftp-based file upload system described in this paper has been superseded by a browser-based file upload system which uses special password-protected web interfaces for the authors and editors.]
- "Why Philosophy Needs a 'Dynamic' Encyclopedia", by John Perry and Edward N. Zalta, URL = <http://plato.stanford.edu/why.html>, November 1997.
- Abstract, "Stanford Encyclopedia of Philosophy: A Dynamic Reference Work", in *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries* (June 24-28, 2001), New York: Association for Computing Machinery Publications, p. 457
- Abstract, "Digital Workflow Concepts for Dynamic Reference Works", abstract of talk delivered at the *Ancient Studies -- New Technology Conference*, Salve Regina University, December 2000

Principal Editor*:*

Edward N. Zalta(CSLI, Stanford University)

Associate Editor/Principal Perl Programmer:

Colin Allen(Philosophy/Texas A&M University)

Assistant Editor/Associate Perl Programmer:

Uri Nodelman(Computer Science/Stanford University)

Advisory Board:

Department of Philosophy, Stanford University

Editorial Board/Subject Editors

List of Authors

Occasional Referees

***Information for Authors*:**

- *Entry Content: Guidelines and Policies*
- *HTML Editors and Web Authoring Tools*
  - (create HTML easily without editing HTML sourcefiles directly)
- *HTML Guides and OnLine Help for Writing HTML*
  - (for editing HTML sourcefiles directly)
- *Recent Access Statistics for the Stanford University Server*
  - Weekly Statistical Summary (September 16, 2001 -- present)
    (Here is a list of the WWW worms excluded from these statistics.)
  - General Statistical Summary (September 16, 2001 -- present)
  - Weekly Statistical Summary (September 17, 2000 -- September 15, 2001)
    (Here is a list of the WWW worms excluded from these statistics.)
  - General Statistical Summary (September 17, 2000 -- September 15, 2001)
- *Recent Access Statistics for the University of Sydney Library (SETIS) Server (mirror site)*
  - Weekly Statistical Summaries
  - Weekly Statistical Summaries Archive
  - General Statistical Summary
  - General Statistical Summary Archive
- *Recent Access Statistics for the University of Amsterdam (ILLC) Server (mirror site)*

- Weekly Statistical Summary
- General Statistical Summary

- *Recent Access Statistics for the University of Leeds (LTSN) Server (mirror site)*
  - Weekly Statistical Summary
  - General Statistical Summary

Editorial Policies:

1. Editorial decisions concerning the *Encyclopedia*, including decisions concerning its content, format and distribution, are made by the Principal Editor in consultation with the Associate Editor, Assistant Editor, and the Board of Editors.
2. The members of the Board of Editors are selected in consultation with the Stanford Department of Philosophy, which serves as the *Encyclopedia*'s Advisory Board. The Advisory Board also advises the Principal Editor on the basic policies governing the operation of the *Encyclopedia*.
3. Contributions to the *Encyclopedia* are normally solicited by invitation from a member of the Board of Editors. However, *qualified* potential contributors may send a proposal to write on an *Encyclopedia topic*, and a *curriculum vitae* or other description of their qualifications, to an appropriate member of the Editorial Board.
   - By *qualified*, we mean those persons with accredited Ph.D.s in Philosophy (or a related discipline) who have published *refereed works* on the topic of the proposed entry. By *refereed works* we mean either articles in respected, refereed journals or books which have been published by respected publishing houses and which have undergone the usual peer review process prior to publication. (Notes in newsletters, proceedings, unpublished dissertations, etc., do not count as much.) However, if a member of our Board of Editors is familiar with the work of the potential contributor, the latter may be certified as qualified.
   - By *Encyclopedia topic*, we mean a topic that is suitable for a reference work in philosophy and is (a) either listed in our Unabridged Table of Contents or (b) falls within the area of expertise of one of the members of our Editorial Board. Since the *Encyclopedia* currently does not yet have subject editors for every specialized area of philosophy, some topics suitable for a reference work in philosophy might fail condition (b) -- we reserve the right to determine whether such entry proposals (in specialized areas for which the Encyclopedia lacks subject editors) should be pursued at this time.

   The Board of Editors reserves the right to compare the qualifications of any person submitting an unsolicited request with those of other potential authors who might naturally come to mind for the entry in question.
4. Readers of the *Encyclopedia* are encouraged to contact authors directly with comments, corrections, and other suggestions for improvements.
5. It remains the responsibility of authors to maintain their entries and to keep them current. Authors are expected: (1) to update their entries regularly, especially in response to important new research on the topic of the entry, and (2) to revise their entries *in a timely way* in light of any valid criticism they receive, whether it comes from the subject editors on our Editorial Board, other members of the profession, or interested readers. In connection with (1), authors should

update the Bibliography and Other Internet Resources sections of their entries regularly, to keep pace with significant new publications, both in print and on the web. In connection with (2), the validity of criticism shall be determined by the Principal Editor, typically in consultation with the relevant members of the Editorial Board. The length of time required for a "timely" revision will be negotiable and will both respect the author's current commitments and reflect the seriousness of the criticism. However, entries which require revision but which are not revised within the negotiated timetable may be retired from the active portion of the *Encyclopedia* and left in the *Encyclopedia* Archives until such time as the entry is revised so as to engage the valid criticisms in question.
6. The views expressed by the authors in their entries are their own and do not necessarily reflect those of Stanford University, the Stanford University Philosophy Department, the *Encyclopedia*'s Editor or of anyone else associated with the *Encyclopedia*.

Copyright Information

Copyright Notice. Authors contributing an entry or entries to the *Stanford Encyclopedia of Philosophy* retain copyright to their entry or entries but grant to the Metaphysics Research Lab at Stanford University an exclusive license to publish their entry or entries on the Internet and the World Wide Web, including any future technologies or media that develop to supplement or replace the Internet or World Wide Web. All rights not expressly granted to the Metaphysics Research Lab at Stanford University, including the right to publish an entry or entries in other print media, are retained by the authors. Copyright of the *Stanford Encyclopedia of Philosophy* itself is held by the Metaphysics Research Lab at Stanford University. All rights are reserved. No part of the *Encyclopedia* (excluding individual contributions and works derived solely from those contributions, for which rights are reserved by the individual authors) may be reprinted, reproduced, stored, or utilized in any form, by any electronic, mechanical, or other means, now known or hereafter invented, including printing, photocopying, saving (on disk), broadcasting or recording, or in any information storage or retrieval system, other than for purposes of fair use, without written permission from the copyright holder. (All communications should be directed to the Principal Editor.)

Licensing Agreement. By contributing to the *Stanford Encyclopedia of Philosophy* authors grant to the Metaphysics Research Lab at Stanford University a perpetual, exclusive, worldwide right to copy, distribute, transmit and publish their contribution on the Internet and World Wide Web. The authors also grant to the Metaphysics Research Lab at Stanford University a perpetual, non-exclusive, worldwide right to copy, distribute, transmit and publish any and all derivative works prepared or modified by the Editors from the original contribution, in whole or in part, by any variety of methods on all types of publication and broadcast media other than the Internet, now known or hereafter invented. Authors also grant to the Metaphysics Research Lab at Stanford University a perpetual, non-exclusive, worldwide right to translate their contribution, as well as any modified or derivative works, into any and all languages for the same purposes of copying, distributing, transmitting and

publishing their work.

Statement of Liability. By contributing to the *Stanford Encyclopedia of Philosophy* authors grant, to the Principal Editor, the Associate Editor, the Assistant Editor, members of the Advisory and Editorial Boards, the Metaphysics Research Lab, CSLI, and Stanford University, immunity from all liability arising from their work. All authors are responsible for securing permission to use any copyrighted material, including graphics, quotations, and photographs, within their contributions. The Principal Editor, Associate Editor, Assistant Editor, members of the Advisory and Editorial Boards, CSLI, and Stanford University therefore disclaim any and all responsibility for copyright violations and any other form of liability arising from the content of the *Encyclopedia* or from any material linked to the *Encyclopedia*. Alleged copyright violations should be brought to the attention of the author and the Principal Editor, so that such issues may be dealt with promptly.

Acknowledgements*:*

The Stanford Encyclopedia of Philosophy project is indebted to various people, both at Stanford and elsewhere, all of whom deserve acknowledgement. Although the Associate Editor and the Assistant Editor have been the main Perl programmers on this project since 1998, Eric Hammer, programmed on the project from 1995 to 1997. During the 2000-2001 and 2001-2002 academic years, David James Anderson wrote important Perl programs and has made other contributions to the project. We'd also like to thank David Barker-Plummer, Mark Greaves, Emma Pease, and Susanne Riehemann for their many helpful suggestions concerning the *Encyclopedia* project and the construction of this Web site.

The Principal Editor would also like to acknowledge the contributions of: Javier Ergueta, for his efforts and work in developing a business plan for the Stanford Encyclopedia of Philosophy during the first six months of 2002, and Nathan Tawil, who helped design the Encyclopedia entry format when the project started in 1995. Finally, the South Korean company C.O.Tech, Inc., deserves acknowledgement for their expert advice concerning XML and for working on a Java-based, graphical XML/XHTML-editing program for our consideration and possible use by the authors of the Encyclopedia.

Table of Contents

The Stanford Encyclopedia of Philosophy: A Developed Dynamic Reference Work

Colin Allen*
Philosophy Department
Texas A&M University

Uri Nodelman†
Computer Science Department
Stanford University

Edward N. Zalta‡
Center for the Study of Language and Information
Stanford University

1 Introduction

A fundamental problem faced by the general public and the members of an academic discipline in the information age is how to find the most authoritative, comprehensive, and up-to-date information about an important topic. The present information explosion is the source of this problem—more ideas than ever before are being published in print, on CD-ROM, and in a variety of forms on the Internet. One can nowadays use library search engines and web-indexing engines to generate lists of publications and websites about a topic and then access them immediately if they are

*Professor of Philosophy at Texas A&M University, Principal Programmer and Associate Editor of the Stanford Encyclopedia of Philosophy

†Graduate Student in Computer Science at Stanford University, Associate Programmer and Assistant Editor of the Stanford Encyclopedia of Philosophy

‡Senior Research Scholar at Stanford University, Project Director and Principal Editor of the Stanford Encyclopedia of Philosophy

online. But even limited area search engines can produce thousands of matches to keywords and even with new interface tools to narrow the search, one is typically confronted with a list that is not informed by human judgment. If one wants an introduction to a topic that is organized by an expert, if one wants a summary of the current state of research, or if one wants a bibliography of print and online works that has been filtered on the basis of informed human judgment, there are few places to turn. One might try a standard reference work, but the main problem with reference works is that they quickly go out of date (even before they are published) and don't reflect the latest advances in research. So the following questions arise: How can an academic discipline maintain a reference work which introduces the significant topics in the field (for those who wish to learn the basics), but which tracks, evaluates, and changes in response to new publications and new research being presented in a variety of media (for those with advanced knowledge on a given topic)? How can this be done so that access to the reference work is low-cost, if not free?

Members of our project started thinking about these questions in 1995, and in order to answer them, we developed and implemented the concept of a 'dynamic reference work' (DRW). A DRW is much more than a web-based encyclopedia. The most important features of a DRW are that: (1) it provides the authors (who may be scattered in universities all over the world) with electronic access to their entries, so that they can update those entries at any time to reflect advances in research, (2) it provides the subject editors (wherever they are located) with administrative access to those entries and updates, by which they can referee them prior to publication (and by which they can add new topics, commission new authors, etc.), and (3) it provides automated tools by which a principal editor can oversee administrative control of (1) and (2) with only a small staff. Thus, on our conception, a DRW includes a highly customized work-flow system by which the members of an entire discipline are empowered to collaboratively write *and maintain* a refereed resource. Such a resource would not only introduce traditional topics in the discipline, but would also track the (new) ideas that are constantly being published on those topics in a variety of media. With this concept of a DRW, all sorts of new and interesting questions arise concerning how to best design, program, and administer such a resource and work-flow system.

No electronic journal or preprint exchange in the sciences or human-

ities approaches this concept in scope. Electronic journals: (1) typically do not update the articles they publish, (2) do not aim to publish articles on a comprehensive set of topics, but rather, for the most part, publish articles that are arbitrarily submitted by the members of the profession, (3) typically serve a narrow audience of specialists, and (4) do not have to deal with the *asynchronous* activity of updating, refereeing, and tracking separate deadlines for entries, since they are published on a *synchronized* schedule. Preprint exchanges not only exhibit features (1), (2), and (3), but also do not referee their publications and so need not incorporate a work-flow system that handles the asynchronous referee process that occurs between upload and publication in a DRW. None of this is to say that electronic journals and preprint exchanges have a faulty design, but rather that a DRW is a distinctive new kind of publication that represents a new digital library concept.

Although commercial publishers have built web-based reference works and claim that they are dynamic, they lack some of the principal design features of a DRW, namely, (1) that authors should have electronic access to copies of their entries and be able to modify them, and (2) that subject editors and the principal editor should have electronic access to the encyclopedia databases and unrefereed entries, so that they can directly carry on the task of adding and commissioning new entries, refereeing entries and updates, etc. These commercial publishers typically don't give academics accounts on their computers, or access to their databases. Instead, the authors and editors must provide/referee content by first interacting with the staff of the publishing house (managing editors, copy editors, computer web specialists, computer markup specialists and others) before changes to the encyclopedia can be made public. On our model, however, the *publishing house* becomes inessential to the process of maintaining a DRW. Academics have direct electronic access to the entries, and can engage and manage the process of writing, refereeing, and updating entries without intermediaries.

Our implementation of a DRW is embodied by the Stanford Encyclopedia of Philosophy (SEP) <http://plato.stanford.edu/>. In the remainder of this paper, we document this particular DRW and then discuss some of the outstanding questions and problems it faces.

2 The Implementation of a Dynamic Reference Work

The SEP first came online in September 1995 with 2 entries! Since then, we have designed a workflow system which attempts to maximize efficiency among those involved in its production. The most important parts of this system are the password-protected web interfaces to the central server, which can be accessed by any author, subject editor, or the principal editors from any where in the world there is a computer with an internet connection.[1]

The web interface for authors allows them to download our HTML templates, to upload their new entries into a private area of our web server, and to remotely edit copies of their entries stored in this private area. So if an author is lecturing outside her university and encounters a reader of her entry who points out an error or omission, she can sit down at the next net-connected computer (possibly at an Internet cafe), contact the Stanford server using the machine's web browser and, after supplying her ID and password, remotely edit the content of her piece and submit it for editorial review. The web interface for subject editors allows them to enter new topics, commission authors for those topics, referee and comment on entries and updates submitted for review, and communicate their decisions to the editor. So, for example, if a subject editor is visiting another university and learns by email that an entry has

[1]These web interfaces, and the file download and file upload capacities which they enable, are the principal enhancements we've made to the SEP since the publication of the paper 'A Solution to the Problem of Updating Encyclopedias', by E. Hammer and E. Zalta, in *Computers and the Humanities*, **31**/1 (1997): 47–60. When that paper was published, the SEP still used an ftp-based file-upload system. We gave authors system accounts on our Unix server, linked their home directories into webspace, and allowed authors to transfer their files by ftp to our server. However, subsequent to the 1997 paper, when browser-based file upload had become a widely adopted and supported standard, we switched to the new technology. Authors and subject editors no longer needed system accounts on our Unix server, and indeed we determined that maintaining Unix accounts for all participants would introduce problems of scale when dealing with hundreds of accounts. Furthermore, we improved security on our machine by deleting those accounts. Instead, authors were given passwords for the browser-based file-uploads. Moreover, subsequent to the 1997 paper, we distinguished a private 'upload-space' (which includes 'revision space') from our public 'web-space'. The former contains private copies of the entries accessible only to authenticated users so that newly uploaded entries, and newly revised entries, do not become publicly viewable until after they have passed through the referee process.

been revised and submitted for review (see the discussion of our tracking and reminder system below), she can use a web browser to log onto the subject editors web interface, display the original and revised versions of the entry side-by-side *with the differences highlighted*, easily determine where the changes are located, referee them, and then accept or reject revised version.

The principal editor also has a special, secure web interface, by which this collaborative process is administered. The principal editor can easily add people to the project, add entries to the database, assign editorial control for entries to the subject editors, issue invitations, track deadlines (for new entries and for updates), and publish entries and updates when they are ready. Many of these things can be done with just the press of a few electronic buttons. For example, when a subject editor submits (through her web interface) a suggestion to commission an author on a particular topic, the suggestion gets entered into a database, and the principal editor is notified and prompted to log onto his web interface. He simply hits the New Invitation button, selects the entry in question, and is then prompted to invite the person listed in the database for that entry by hitting the Invite button.

Finally, we should mention that we have designed and implemented a web interface for *prospective* authors. When a prospective author receives an invitation, they are directed to log on to a special web interface to obtain information about the project, to set up an account with us if they plan to accept, and to set a deadline of up to a year for completing the entry (or else write to us with a counterproposal).

These 'front-end' web interfaces supply data to the 'back-end' processing programs and databases in our system. In particular, actions taken, and information entered, by authors, editors, and prospects are communicated to our tracking and logging system. This system can identify the state of any given entry, recognize who now owes work on an entry and which deadlines have or haven't been met, and pass this information to our automated email reminder system, which has recently been developed, initialized, and put into continuous operation. When an entry changes state and another person must now act to continue the publication process, the reminder system will prompt this person about what needs to be done and by when. It will continue to send reminders (on a fixed, inoffensive schedule) until the work is done (or notify the principal editor that that all reminders have been ignored and that human

intervention needs to take place). Finally, when any entry or substantive revision is published, the entry is scheduled for revision within 3-5 years (depending on how swiftly the field moves). Actually, some authors update once a year, but all authors are notified by our reminder system well in advance of any scheduled revision.

The use of these web and computer technologies offers considerable savings over more traditional publishing methods, and has enabled us to develop the Encyclopedia with a small staff and budget. The importance of this project for our profession cannot be overstated. As new ideas in logic, ethics, political philosophy, philosophy of science, philosophy of cognitive science, etc., are published in books and journals of philosophy, both in print and on the web, the SEP provides a rational and efficient system by which the new information is assimilated, digested, and disseminated in entries which are *responsive* to new research.

Here is a basic quantitative analysis of the effectiveness of our design which is justified by the above. Consider first the fact that there was a 30-year gap between philosophy encyclopedias (the Macmillan Encyclopedia was published in 1967, the Routledge Encyclopedia in 1998). So there was no up-to-date encyclopedia for at least 25 years (9125 days). By contrast, a typical Encyclopedia author is regularly visiting the library to read journals or receiving new journal issues at her office. As soon as she realizes that a recently published article advances the topic of her Encyclopedia entry (and, in principle, this could be the day that an article is published, and in some cases, she might even have advance knowledge of the publication if the author has sent her a preprint), she can use her computer to call up the Encyclopedia server and modify her entry accordingly (maybe by adding a paragraph and a Bibliographic item). The next day, assuming that the change is a substantive one, the relevant subject editor(s) will be notified that the revision must be refereed. Suppose it takes a week to referee the minor changes to her entry. Then we have reduced the length of time required for the update process from about 9,000 days to about 9 days, or by 3 orders of magnitude. Even if it were to take up to a year for a new idea (in a book, say) to become reflected in the Encyclopedia after the new idea is published, that would still constitute a 25-fold decrease in the length of time it takes for a philosophy reference work to reflect the advance.

We should mention two other features of the SEP which should be part of any DRW. The first is the fact that authors are encouraged to

write nested, as opposed to linear, documents. That is, we encourage our authors to put highly technical, scholarly, or highly detailed information into supplementary documents and to link these into the main part of the entry. These supplementary documents can have supplements as well, and so forth, and so the reader can then choose the level of detail they wish to explore. Such nested entries become useful to a wider range of readers — intelligent undergraduates should be able to get through the main entry by skipping the links to the supplements, while graduate students and colleagues may skip the basics and follow the links to the supplementary documents, to find the cutting edge material.

The second noteworthy feature concerns archiving. For purposes of citation, a DRW is a moving target, since the entries are always being corrected, updated, improved, etc. It is difficult to cite such a moving target. For example, a reader might quote a passage from a DRW entry in a research article, and after publishing the research article, discover that the author of the DRW entry has altered the passage in question. To solve this problem, we make quarterly archives of the SEP. On the equinoxes and solstices, we make an electronic copy or 'snapshot' of the entire encyclopedia as it exists on that day and link that complete copy into our special Archives page. We explain to users that the proper way to cite an SEP entry is to cite the most recent archived version. These archived versions will not be updated or changed in any way, and so scholars can rest assured that the passages they quote will be available for scholarly purposes. Note that every entry in the SEP contains a section called Other Internet Resources which contains links to offsite web-based material and these links may eventually break in the archived entries (especially if the links do not point to similarly archived material). That is a danger of the web. But we do attempt to minimize the problem, however. We have designed and programmed a 'link-rot' detection system which automatically notifies the authors anytime links break in the dynamic versions of their SEP entries. The authors are asked to revise or delete the link.

3 Statistics about the SEP as a Dynamic Reference Work

As of September 21, 2001, the SEP had 213 entries online. We had 69 subject editors overseeing 513 authors currently working on a total of 600

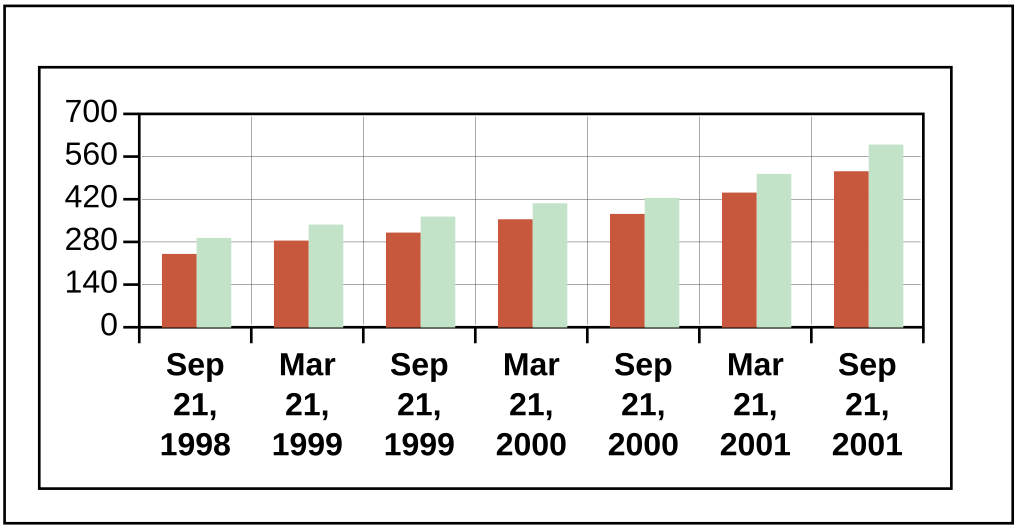


Figure 1: Number of Authors/Commissioned Entries

commissioned entries.

Over 10% of our entries make use of some hierarchical document structure (i.e., they involve more documents than simply a main text and a footnotes page) and just under half of our entries have been updated since they were first published.

The rate at which we commission entries increased by a factor of 3, from about 5 per month in 1999 and 2000 to about 15 per month in 2001. See Figure 1. During the same period, our publishing rate increased by a factor of 6, from about 1.5 entries per month in 1999 to 9 entries per month in 2001. See Figure 2. The average length of our entries also increased from approximately 6800 words per entry in September 1998 to 8900 words per entry currently.[2] We estimate that, in print, the current version of the SEP would fill over 3000 pages.[3]

Between September 1997 and September 2001, the content of the SEP grew from about 3 megabytes to about 26 megabytes. See Figure 3. During that same period our average accesses per month increased by an order of magnitude, from about 5000 to 57000. See Figure 4.

[2]These word counts are slightly inflated due to the presence of HTML tags in the text. We estimate that the tags add about 300 to 500 words per entry.

[3]This estimate is based on an assumption of 600 words per page.

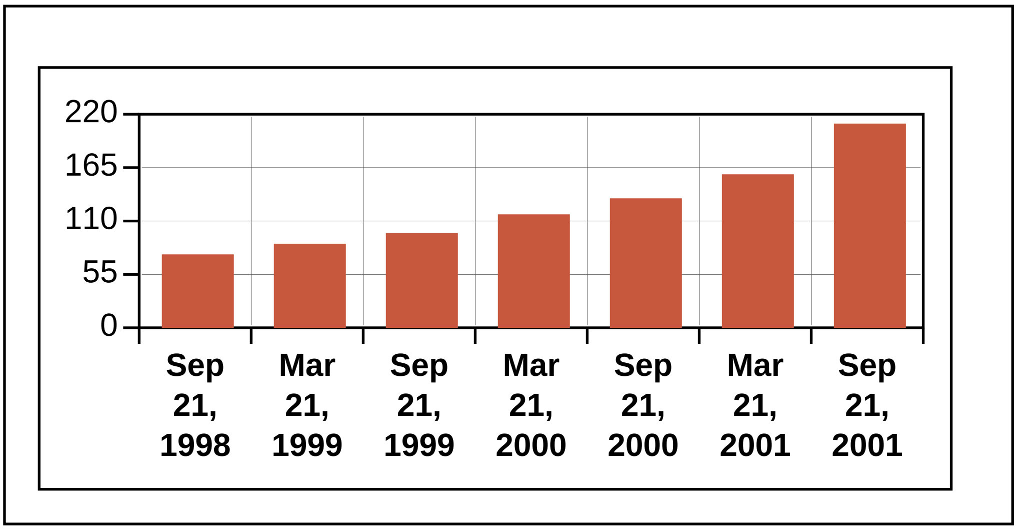


Figure 2: Number of Entries Online

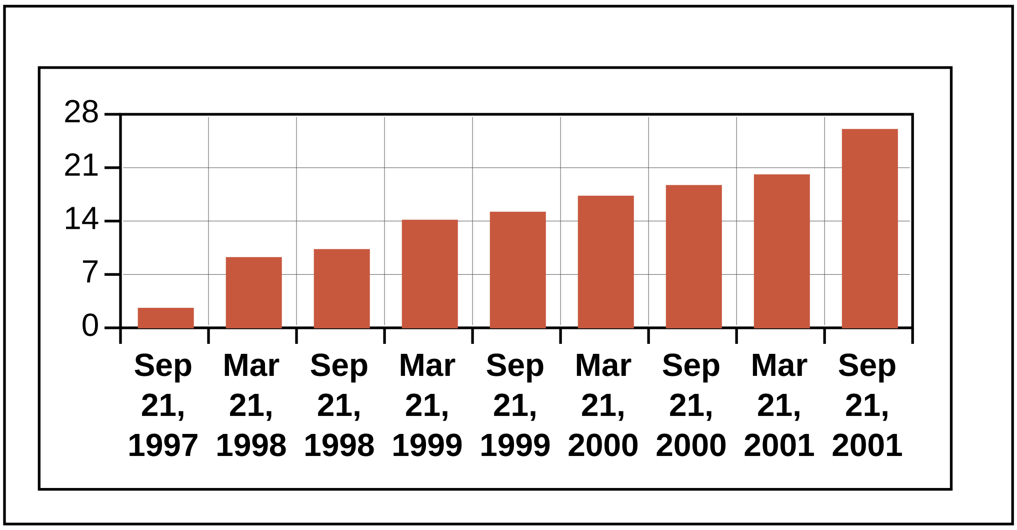


Figure 3: Content in Megabytes

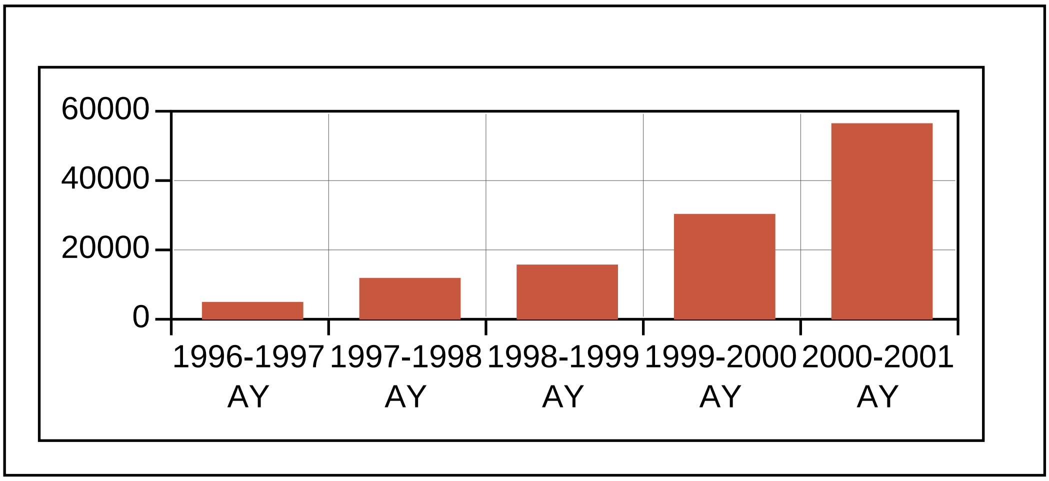


Figure 4: Weekly Average Number of Users (Excluding Mirror Sites)

4 Why the SEP (and other DRWs) Should Be Free

After 6 years of operation and numerous discussions with authors and subject editors involved with our project, with faculty members and colleagues around the world, with publishers, and with university librarians, we have come to the conclusion that if the means are possible, the SEP should try to remain a free resource. One may think of this as an idealistic goal, but there are several problems associated with a subscription-based, access-restricted models for funding the SEP. We discuss some of these problems in what follows.

The first problem arises from the fact that academics who author entries in scholarly reference works traditionally receive a fee for their efforts. They are, after all, providing a service to the publisher. However, authors of SEP entries are volunteering their time. There are various reasons why they do this. One might be the fact that they will reach a large number of readers. (As long as the SEP remains free or low-cost, it will have a large readership.) Another might be the intellectual obligation academics might feel they have to contribute to the profession and world at large by playing a role in maintaining an up-to-date resource

in philosophy. Another might be the prestige they might acquire should they become known widely as an expert on a certain topic, or by becoming associated with a Stanford University project. And, of course, authors may be motivated by the fact that they can list these entries as invited publications on their curricula vitae, which play an important role in promotion, hiring, and tenure.

This volunteer arrangement, however, might become compromised if the SEP were required to charge subscriptions in the attempt to make a profit. Even if the SEP charged subscriptions at a rate that simply recovers costs, authors might suggest both that we should recognize their efforts as part of that cost and that we should therefore increase prices a little so as to collect enough revenue to distribute royalties to the authors. The situation is complicated, moreover, by the fact that authors might *additionally* argue that maintaining an entry is also a service. In any year they update their entry, they might be owed remuneration. As we scale up to 1000+ authors, this problem becomes even more acute. And a final wrinkle is the fact that our authors currently provide us with entries formatted in HTML and nearly 'web-ready'. If we charged a subscription, if only to recover costs, authors might claim that we are *offloading our typesetting costs* onto them. It would be best to sidestep all of these questions by finding a model on which the SEP could remain a free resource and our authors never feel that they are being exploited.

As second, somewhat similar, problem concerns our subject editors. These subject editors are constantly suggesting new entries, commissioning entries, and refereeing entries and updates. If we required subscriptions for access to the SEP, if only to recover costs, the subject editors might argue that their services should be recognized as part of those costs, and that we should increase our subscription rates to pay them accordingly. This problem is magnified by the fact that the work involved in maintaining a dynamic reference work continues from year to year, with no fixed endpoint, and it is complicated by the fact that some subject editors have higher workloads than others.

A third problem concerns the difficulty of finding a subscription model that everyone can live with. University libraries and other institutions are reluctant to pay for something the rest of the world can get for free. Consequently, it would be difficult to implement a model in which these institutions would pay a subscription fee while everyone else would be allowed free access. So a subscription model might eventually force us

to require *everyone* to subscribe (albeit, at appropriately proportional rates).

A fourth problem is the fact that many deserving groups of people would be disenfranchised by a subscription model. Small colleges, public libraries, and K-12 school libraries usually can not afford even a modest subscription fee. Moreover, people who are accessing the web from home over an ISP seem reluctant to pay for content, much less for a subscription to a *philosophy* encyclopedia. And last, but not least, our users in the developing parts of the world would become disenfranchised; this includes university students and colleagues in developing countries, as well as lay persons in those countries. Since the SEP has been accessed by users in over 150 countries around the world, academics associated with the project as authors would lose a significant base of readers.

A fifth problem concerns search engines. When people search Google, Yahoo, etc., for philosophical topics, our pages figure prominently in the results. That is because our pages are free to everyone, and web-indexing spiders can access and index our pages. But on a subscription model, our server would restrict access to those who pay. This would make it nearly impossible for search engines to index our site, and consequently people wouldn't find the SEP pages when they conduct web searches. Even if some arrangement could be made to allow indexing spiders to index our site, access restrictions would make the links returned by a search engine useless to the majority of web users, and make it difficult and pointless for non-paying users and institutions to create links to our site on their own pages. But without those thousands of links to the entries on our site, the prominence of our pages in the results of certain web search engines (e.g., Google) would greatly diminish, since those search engines often prioritize the websites which are returned as matches to keyword searches by the number of cross-referencing links to those websites that exist on the web.

A sixth problem concerns our mirror sites. Currently, the SEP has mirror sites (U. of Sydney, U. of Amsterdam, and, soon, U. of Leeds) which perform the following functions. (1) They guarantee our users access to the Encyclopedia pages when the Stanford server temporarily goes down. (If anything happens to our server, our readers use a mirror site until we fix the problem and reboot.) (2) These mirror sites give our users in other parts of the world faster access to our pages. (3) The mirror sites provide very important layers of digital preservation for the SEP pages by keeping complete copies of our data. These institutions

completely underwrite the costs of the mirror sites and provide this service for free. But this arrangement would come to an end if we started charging a subscription for our site. On a subscription model, either we would have to pay the mirror sites for their costs in manpower, equipment, and overhead (assuming that some monetary arrangement could be made), or we would have to face the much more expensive proposition of running 'mission critical' servers that provide fast, world-wide access 24 hours of every day. (If institutions have to pay for the SEP, they would expect a level of service at least comparable that which they now enjoy.)

A seventh problem is the fact that if the SEP were forced to rely on cost-recovery to survive, the decisions which are now made *in-part* on the basis of the interests of the profession may have to be made instead on the basis of cost recovery *alone*. This might have a negative impact on the quality of the Encyclopedia, such as placing strict word limitations on the length of entries and bibliographies (to save disk space or stay within bandwidth limitations), banner advertising, links to online booksellers, etc. The latter, for example, would compromise the integrity of the Encyclopedia, since a user might wonder whether the link to an online bookseller is present because of the merits of the book or because the online bookseller was kicking-back some of the profits to the publisher.

In addition to the above seven problems that would arise if the SEP adopted a access-restricted, subscription-based cost-recovery model, we believe that there is a positive reason for remaining free, namely, that it would be an outstanding legacy for the SEP and profession as a whole if it could provide both academics and non-academics around the world with a free resource by which they could satisfy their intellectual curiosity from an authoritative source on philosophical questions of all kinds and, in particular, those concerning the human condition.

5 Costs of the SEP

Of course, it does cost money to produce the SEP and if it is to remain free, those costs will somehow have to be underwritten. Currently, the SEP costs are underwritten by a grant from the NSF, which funds the SEP from October 2000 to September 2003.[4]

[4]The NSF grant (#IIS-9981549) was made possible by a significant financial contribution from the NEH. The NEH previously funded the Encyclopedia project from September 1998 to August 2000 (#PA-23167-98). From September 1995 to August

The NSF grant pays for a Principal Editor (working 50% time), a Consultant Perl Programmer (working 33% time), an Associate Programmer (working 20% time), an Assistant Programmer (working 10% time) and an Administrative Assistant (working 10% time).[5] Clearly, this adds up to a tiny staff, since the percentages total to one person at 100% employment and one at 23% employment. We believe our accomplishments are significant when viewed in light of our tiny staff, though it should be mentioned that these accomplishments were made possible by the fact that our personnel regularly put in longer hours than official records indicate. We are currently seeking grant money to hire a business consultant to determine exactly what costs are required to run/staff the SEP and what funding models are available to ensure that the SEP's long-term survival.

We believe that the SEP could evolve into an even greater publication (i.e., with even higher quality entries, with fewer typographical and other errors, etc.) if it had more adequate staffing. It does not strike us as unreasonable to think that the SEP should, for the long term, require a Principal Editor at 50% time, an administrative assistant at 50% time, a programmer at 50% time, and an HTML (XML) copy-editor at 50% time. With an administrative assistant and HTML copy-editor working at these levels, the Principal Editor would be relieved of certain routine tasks and could concentrate more on content issues, on supporting subject editors for those subjects of the SEP which still aren't very far along, etc. Similarly, with a programmer working half-time, we could re-engineer many of the compromises we made in designing and programming our workflow system, we could adapt the SEP to new technologies that improve web-based publication, etc. Although our estimate of these long-term staffing requirements for the SEP will need to be subjected to a strict business analysis, it should give the reader some idea of the money for salary income that the SEP will need. Once money for overhead costs (including hardware and software) are included in the equation, it should be clear that the SEP will have to find sources of income if it hopes to remain a free publication after the NSF grant expires.

1998, the SEP was funded by seed money from the Center for the Study of Language and Information (Stanford University), where the project was conceived and developed. CSLI still contributes some cost-sharing funds under the terms of the NSF grant.

[5] After we complete the task of programming the essential workflow-systems underlying the SEP, there will be less programming time required. However, if we are to keep up with changes in technology, programming time will always be necessary.

6 Future Challenges

Clearly then, the principal future challenge for the SEP is to find a source of income by which it can continue to be universally accessible. In this concluding section, we discuss this challenge, along with two others.

6.1 Funding Models

We will, of course, attempt to raise an endowment. Indeed, we plan to prepare grant proposals to foundations, asking them to give us money to hire a fund-raiser for a year, at full or half-time. But we anticipate that it could take up to 10 years to raise an endowment large enough to cover our operating costs.[6]

Our present focus, however, is to explore funding models intermediate between universal free access and full-fledged subscription models. For example, we plan to study the following two models, both of which might be adopted without incurring all of the problems outlines in Section 4.

1. *Voluntary Archive-Acquisition Program.* On this income-producing plan, the SEP would remain free to everyone, but institutions which subscribe to this program would be entitled to download and store our 4 quarterly archives each year they subscribe. (Those who choose not to pay would still be able to access all of our pages.) There would be 2 advantages to subscribing: (a) the institutions which subscribe would be entitled to serve their copies of the archives whenever our server and its mirror sites are down, and (b) should the SEP project ever cease to exist, these institutions would own, and be able to locally circulate, copies of our archives even though our servers are no longer active. For an extra fee, we could enhance this service by burning and distributing yearly CDs of the archives for that year.

2. *Archive-Access Program.* On this income-producing plan, *only* those institutions/users who pay a yearly subscription fee would be able to access our archives for that year. The SEP's dynamic entries (which are always changing) would still be available free to everyone. However, since citation can take place only to fixed, archived versions of

[6] Remember that if endowments are managed properly and make 8 to 10% a year, then approximately one-half of the 8 to 10% would be returned to the SEP and the remaining half of the 8-10% would be reinvested.

the entries, the motivation for joining this program would be clear to scholars and librarians. On this plan, we could again create the following levels of service: (a) for a basic yearly fee, users at subscribing institutions would receive web-based access to the archives on our server, (b) for a higher fee, the institution would obtain the right to download and store the archives, and (c) for an even higher, but still moderate, fee, we would burn and distribute CDs to the subscribers.

We plan to investigate these and other models by which income could be raised. They strike us as offering the best hope of raising money in a way that addresses the problems discussed in Section 4.

6.2 Institutional Home

The question of funding models is connected to the question, where is the best institutional home for the Encyclopedia? We believe that the SEP is better off in an academic environment than in the hands of a commercial or even non-profit publisher/publishing house. Normally there would be three good reasons for joining forces with a non-profit, university-based publisher, namely, that such a publisher would: (1) offer expertise in the business of producing publishable material, (2) offer a more stable institutional home, and (3) have the mechanism for marketing and collecting revenues for the materials it produces. We consider these in turn.

Expertise?

We are not aware of any non-profit publishers which have the kind of expertise which already resides with the SEP project. Since 1995, we have been perfecting our model for dynamic reference works which can be run on a low-cost basis. Grants from the NEH and NSF for 1998-2003 have given us the financial resources to program a workflow system which would compare, on the open market, to \$200,000 – \$300,000 off-the-shelf 'web content management' software systems. The Stanford Encyclopedia of Philosophy has been successful to date as a project designed, programmed, and run by academics (not specialists) who have acquired the tools necessary to make use of the power that computers, the Internet, and web-based technologies provide. In particular, the most highly technical positions required by our project, Unix system administration and

Perl/CGI programming, are filled by *academics* working part-time.

A More Stable Institutional Home?

Currently, the SEP is published at Stanford University's Center for the Study of Language and Information (CSLI), and the Stanford Philosophy Department serves as its Advisory Board. Our analysis suggests that the best institutional home for the SEP is an academic setting of this kind. Numerous reasons for this are readily apparent. An institution such as CSLI or an academic philosophy department would have a more intimate and direct concern for the academic excellence of the project and for safeguarding a resource for the profession. Such a concern might not be shared by a non-profit, university-based publisher, since traditionally they let titles go out of print. Moreover, academic setting of a philosophy department or research institute has a concern for educating its graduate students and graduate students could play various and vital roles in the SEP project.

Many graduate students in philosophy have a background in mathematics and computer science. They arrive at graduate school with enough knowledge about Unix, web servers, etc., to work on or consult for our project. Here are two ways in which these students can provide a steady stream of innovative ideas for the future of the project. First, as paid part-time members of the SEP staff. Graduate students would make excellent part-time staff. As such, they could (a) help subject editors plan and commission entries (thereby becoming known to distinguished members of the profession outside their home institution), (b) acquire and use HTML or XML skills and help the Encyclopedia with content mark-up, (c) acquire and use their knowledge of Unix and web-servers to help adminster the project (thereby preparing them to use those skills in their later academic life), and (d) work as office staff, handling correspondence with the authors and relieving the Principal Editor of routine tasks in the administration of the Encyclopedia.

Second, work on the SEP could be made part of the graduate curriculum. A philosophy department, without exploiting the graduate students in any way, could create a one-hour/week proseminar for all first year graduate students. Each week, the students would be required to read and report on one article from the SEP in their field or in a related field. In this way, graduate students would enhance their breadth and analytic

skills as philosophers, improve their writing skills by focusing on whether entries are well-written from a pedagogical point of view, and suggest ways to improve and/or update entries. They could bring their talents with web-based searching to identify whether any related material on the web should be linked into the entries they consider.

Finally, it is important to note that a research institute such as CSLI can be an important collaborator in this project. For instance, CSLI has researchers with expertise in linguistics, computation, data-mining, etc. It also has a highly-skilled technical staff, which can deal with any technical issues (such as those connected with server operation, backup, etc.) that might go beyond the expertise of a typical academic department.

Marketing and Income Collection?

Of course, if we can find a way to keep the SEP a free resource, then the fact that a publisher offers marketing and income collection becomes moot. However, as we have seen, we may be forced to adopt an operating model intermediate between that of keeping the Encyclopedia free and requiring universal subscription. Such models would require some marketing and income collection.

Basic marketing would not be problematic for the Philosophy Department and CSLI. When our NSF grant runs out in 2003, the SEP will have been on the Web as a *free* resource for 8 years. During our first 6 years, we have become well-known throughout the philosophical world (both academic and non-academic) and many thousands of individuals know about, read, and rely upon our pages. If our service were to be shut-off or restricted in some way, libraries and other institutions would certainly hear about it from their constituents. Moreover, most of our 500+ authors and 70 subject editors would notify the libraries at their own institutions. In addition, the fact that the SEP is a free resource means that there are thousands (if not tens of thousands) of links to our web pages. If we were to start charging subscriptions for access, these links would all end up in an *advertisement to subscribe*, as soon as anyone followed the link.

Income collection could also be easily accomplished, and could be done by using a university-based eCommerce center, such as the one at Stanford University.[7] These eCommerce centers can provide subscription services via the Web and can handle subscription payments for departments that

[7] See <http://www.stanford.edu/group/itss-ccs/project/ecommerce/>.

create journals or other publications for sale. These eCommerce centers are relatively inexpensive to use.

Finally, we should mention that if a more sophisticated system for marketing and income collection is required, an alliance with the *non-profit* Philosophy Documentation Center might be possible. Discussions to this effect have already taken place.

6.3 A Move to Newer Technologies?

The SEP was designed to run on proven, free technologies. Of course, since we began the project in 1995, some of these technologies are now "legacies". But, in many cases, we chose to use certain technologies over others because they made the most sense given our budgetary constraints. We did have to make compromises in some cases.

We use HTML rather than XML for entry markup. We do not use any heavy-duty 'application server technology', since performance is excellent with our Apache server and Perl/CGI scripts. We do not rely on any heavy-duty data-base technology (such as Oracle), because our main data base has (or will have) only 1,000 to 2,000 records, as opposed to hundreds of thousands of records. We don't rely on Java or Javascript, though our authors are free to use it in their entries if needed. We avoid frames on the main pages of the Encyclopedia, though we use them in the web interfaces.

Those who follow trends on the web, however, will know that this implementation is relatively 'low-tech'. But this low-tech approach does have several advantages. One is that the system can run on any PC running a free Unix-based operating system such as Linux or FreeBSD. Our low-tech approach also does not require that we purchase any licenses or require that we become dependent on any commercial software. Just as important is the advantage that both philosophers with limited computer savvy and expertise, and philosophers in other parts of the world where access to hi-tech or up-to-date computer systems may be limited, can participate in the collaborative production of our DRW.

In addition, the newer technologies always seem to increase the costs of production (they often require specialized personnel, for example), and until we are satisfied that the benefits outweigh the costs, we will exercise caution when considering the latest technologies. But eventually some of these newer technologies may supercede the older ones and the SEP

will have to make the needed adjustments, assuming it is in a financial position to do so.

The question we are asked most frequently is, when do we plan to adopt XML as a markup standard? XML is now highly touted as the markup language to use in web publications. As a markup language, XML offers some serious advantages over HTML. For one thing, the tags are constructed on the basis of the kinds of content that appear in a document. For example, in HTML, one would format book titles using the italicizing tags, such as <i>...</i> or .... But in XML, one could format book titles with the tag <booktitle>...</booktitle>. This would allow one to search of the SEP in more sophisticated ways. One could tell the search engine to search only keywords in the <booktitle> environment, whereas in HTML, there is no way for a search engine to distinguish a keyword found in a book title from one found in other italicized environments, such as emphasized text or foreign words and phrases.

Another advantage of XML is the promise of more sophisticated mathematical and logical formatting. HTML has only weak resources for formatting sophisticated mathematical and logical notation. There is some promise that MathML (a formatting language which is a special instance of XML), when supported by MathML-aware browsers, will give web publishers the ability to publish professional-looking mathematical and logical notation.

However, these virtues of XML come at a cost. The first and foremost of these is the fact that since XML is a more sophisticated markup language, the costs of production rise significantly when taking proper advantage of XML's extended capabilities. Currently, our authors provide us with nearly 'web ready' HTML documents, which they produce with freely available HTML-editing software.[8] Indeed, they are free to use any HTML-editor to compose their entries—we do not want to force all authors to have to learn and/or use the same composition software. Until XML-editing software tools become widely available and easily configurable, our authors will not be willing to spend extra time using all the new tags provided by XML to format their entries. (For example, authors will understandably be reluctant to familiarize themselves with

[8] It must be mentioned, however, that we always have to spend time to bring their documents into compliance with international standards (e.g., removing proprietary HTML formatting codes that the HTML-creation software introduces) and to make them consistent with our other entries.

all the special tags such as <booktitle> and use them in their entries.)

There may be some ways, however, to ameliorate these costs. Suppose, for example, that there was a free software application which the authors could use to help them to graphically and automatically format entries in XML without learning the new tags. For example, any time an author wanted to insert a new item into the Bibliography, such a piece of software would, in response to a click on an 'Add Bibliography Citation' menu item, pop-up a window containing all the relevant fields (book title, article title, author name, date, city, publisher, etc.) of a typical citation. When an author inserts the information into these fields, the software would then mark the information with the appropriate XML tags in the sourcefile. There is very little extra cost to the author, since they have to type in the information in the Bibliography citations anyway. Of course, such a piece of software would be expensive to design, produce, and support on the major computer platforms. But until such software is widely and freely available or the SEP has the financial resources to hire XML-markup specialists, the move to XML will be problematic.

There is also a second problem with XML, which has to do with the fact that our authors now have electronic access to their entries and can keep them up to date. As we mentioned much earlier, authors can use their browser to contact the SEP's server through a web interface. When they activate our 'Make Changes' function, their browser will divide up a private copy of the entry into sections, and for each section, display both the rendered HTML and an editing box to the HTML sourcefile. The author can then edit the HTML sourcefile and redraw the screen to see that the HTML is rendered correctly. It is not difficult for authors to read past the HTML formatting tags to edit the text they wish to change, or even add basic HTML formatting to their updated text.

But this procedure becomes more difficult in XML, and especially, MathML. XML sourcefiles are much more highly formatted than HTML sourcefiles. A simple equation such as $2^2 + x = 8$ requires numerous MathML formatting tags and it is much harder to read past these tags to find and edit the text. (MathML was designed with the idea that authors would never actually edit the sourcefile, but always edit the file through a graphical interface.) So a move to XML would make it more difficult for our authors to update their entries since the sourcefiles would become much more difficult to edit. Again, there may be a way to get around this through a Java-based applet/servlet system which presented authors, no

matter where they are located in the world, with a graphical editing interface to their XML sourcefiles. But such a Java applet/servlet combination is extremely difficult to program so that it works with all combinations of computer architectures, operating systems (and their different versions), web browsers (and their different versions), etc. The costs would be exorbitant for a project that hopes to keep costs to a minimum. It is doubtful that even a single full-time programmer could design, produce, maintain, and support such an application. Consequently, a premature move to XML would interfere with the ease of scholarly communication which the SEP now enjoys.

As one can see, then, there are many challenges facing the SEP. We have a system that works reasonably well now, and we are working now to put ourselves secure the SEP's future for the long-term.

A Solution to the Problem of Updating Encyclopedias

Eric M. Hammer and Edward N. Zalta*

Center for the Study of Language and Information
Stanford University
(ehammer,zalta)@csli.stanford.edu

Abstract: This paper describes a way of creating and maintaining a `dynamic encyclopedia', i.e., an encyclopedia whose entries can be improved and updated on a continual basis without requiring the production of an entire new edition. Such an encyclopedia is therefore responsive to new developments and new research. We discuss our implementation of a dynamic encyclopedia and the problems that we had to solve along the way. We also discuss ways of automating the administration of the encyclopedia.

Note: This paper appeared in *Computers and the Humanities* (Volume 31/1, 1997, pp. 47-60). It is reprinted here with permission from the publisher. You may also access the paper in both (compressed) Postscript and Adobe Acrobat (PDF) formats.

The greatest problem with encyclopedias is that they tend to go out of date. Various solutions to this problem have been tried. One is to produce new editions in rapid succession.* Another is to publish supplements or yearbooks on a regular basis.* Another is to publish the encyclopedia in loose-leaf format.* In this paper, we propose a solution to this problem, namely, a `dynamic' encyclopedia that is published on the Internet.* Unlike static encyclopedias (i.e., encyclopedias that will become fixed in print or on CD-ROM), the dynamic encyclopedia allows entries to be improved and refined, thereby becoming *responsive* to new research and advances in the field. Though there are Internet encyclopedias which are being updated on a regular basis, typically none of these projects gives the authors direct access to the material being published. However, we have developed a dynamic encyclopedia which gives the authors direct access to their entries and the means to update them whenever it is needed, and which does so without sacrificing the quality of the entries. In the effort to produce a dynamic encyclopedia of high quality, we discovered that numerous problems had to be solved and that routine editorial and administrational functions could be automated. By reporting on our project, we hope to facilitate the creation of such reference works in other fields.

Next Up Previous
Next: Basic Description of Dynamic

Eric Hammer and Edward N. Zalta
Wed May 14 17:44:00 PDT 1997

Basic Description of Dynamic Encyclopedias

We have recently developed the Stanford Encyclopedia of Philosophy (URL = http://plato.stanford.edu/). The principal innovative feature of this dynamic encyclopedia is that authors have an ftp (`file transfer protocol') account on the multi-user computer that runs the encyclopedia's World Wide Web server. This feature not only enables the encyclopedia to become functional quickly, but also gives the authors of the entries the ability to revise, expand, and update their entries whenever needed.

Traditionally, encyclopedias have not been very responsive to new research and developments in the field--it is just too expensive to publish regularly new editions in a fixed medium such as print and CD-ROM. However, a dynamic encyclopedia simply *evolves* and quickly adapts to reflect advances in research. We believe that the process of updating individual entries never ceases, and that any encyclopedia which takes account of this fact will necessarily be more useful in the long run than those which don't.

Authors who have a strong interest in and commitment to the topics on which they write will be motivated to keep their entries abreast of the latest advances in research. Indeed, dynamic encyclopedias may speed up the dissemination of new ideas. Of course, there may come a time when an author wants to transfer responsibility for maintaining the entry to someone else. In such cases, there is the possibility of having multiple entries on a single topic, and this is one of the new possibilities that can be explored in a dynamic encyclopedia.

Here is how we implemented our dynamic encyclopedia. We connected a multi-user (UNIX) workstation to the Internet and installed a World Wide Web server. We then created a cover page, a table of contents, an editorial page, and a directory entitled *entries*. We recruited Editorial Board members for the job of identifying topics, soliciting authors, and reviewing the the entries and updates when they are received. Once an Editorial Board member decides on a topic and has found an author to write it, he or she passes on the information to the Editor of the encyclopedia, who creates an ftp account and home directory for the author on the workstation and then sends the author the information on how to ftp the entries and updates when they are ready. So when authors ftp an entry or an update to their home directory, it becomes part of the encyclopedia* and the Board member responsible for that entry is automatically notified. It is then his or her responsibility to evaluate the (modified) entry and notify the author of any changes that should be made.

The innovative features of a dynamic encyclopedia that has been organized on the above plan are:

1. It can be expanded indefinitely; there is no limit to its inclusiveness or size. New or previously

unrecognized topics within a given discipline can be included as they are discovered or judged to be important.
2. It eliminates the lag time between the writing and publication of the entries.
3. It eliminates many of the expenses of producing a printed document or CD-ROM: typesetting, copy-editing, printing, and distribution expenses are no longer necessary.
4. It can change in response to new technology as the latter develops, such as new tools, languages, and techniques.

In addition, statistics software for the encyclopedia can maintain logs of access to the encyclopedia, such as which sites users access it from, which entries they access most, which topics they search for, etc. Such information can help inform decisions about which additional entries to solicit, which authors to recruit to write them, etc.

An important motivating feature of using the Internet as a medium is that the encyclopedia can reach a wider audience than is possible with traditional academic journals and books. Because of this, we are recruiting authors capable of writing articles that are of interest not only to specialists.

Next Up Previous
Next: Computer Supported Collaborative Work **Up:** Introduction **Previous:** Introduction

Eric Hammer and Edward N. Zalta
Wed May 14 17:44:00 PDT 1997

Computer Supported Collaborative Work

Encyclopedias are, in some sense, a collaborative effort. It seems natural, therefore, to analyze the task of building a dynamic encyclopedia in terms of `computer supported collaborative work' (CSCW).* For example, since both the Editor and the author will have write access to an entry, the place on the disk where the entry is stored constitutes a `group workspace'.* Thus version control may seem necessary to prevent simultaneous editing by different `group members'.

Version control could prove useful on those rare occasions when the Editor, as opposed to the author, changes an entry to repair a typographical error or fix some problematic HTML code. Although the Editor will typically leave such tasks to the authors, there may be times when quick action by the Editor is necessary. On such occasions, authors and Editor could find themselves in the situation of attempting to modify the entry simultaneously. However, to avoid such conflicts, we instruct our authors to follow a protocol for revising their work, namely, to begin both by notifying the Editor of their intentions and by downloading the current version of their entry from the Encyclopedia. Such a procedure will prevent author and editor from overwriting each others modifications.*

Coauthored entries will obviously be highly collaborative, but these constitute only a very small percentage of the entries. If we ignore coauthored entries, it is striking that some of the distinguishing features of CSCW are absent. For example, no member of the group of authors requires information on the current status of the work being done by other group members.* Moreover, no member of the group of authors requires information about the *history* of other authors' collaborative activities. Nor do members of the group of authors require information about the process of collaboration (e.g., the roles and responsibilities of other members, and which group members fit into which roles).

These features of CSCW, however, do apply to the Editor, who requires information on the current status of the work by the authors, on aspects of the history of the authors' activities, and on the process of collaboration. In addition, members of the Board of Editors will need information about the history of the activities of those authors writing on topics under their editorial control; for example, a board member needs to know as soon as such an author has updated an entry. And, finally, if the encyclopedia project has the financial resources to maintain a large central staff, then such CSCW concepts as conferencing, bulletin boards, structured messaging, meeting schedulers, and organizational memory could play a role in the design of administrative procedures.

Since we are operating on a much smaller scale, these last CSCW concepts will play almost no role in what follows. The CSCW features that do apply will become features of the central administrative

control of the encyclopedia and can be managed by properly defined databases and updating procedures. Thus, the CSCW concept most relevant to our enterprise is `work flow management'. By analyzing the way in which the Encyclopedia would typically function (i.e., the sequence of tasks of the parties involved and the sequence of transactions among the parties), one can predict and address many of the problems that would affect the smooth operation of the Encyclopedia. These will be discussed in the next two sections. Even the choice of technologies was to some extent dictated by this analysis of work-flow. For example, we investigated SGML as a possible markup language for the Encyclopedia entries and we created a Document Type Definition for a typical encyclopedia entry (thereby defining tags that the authors would use to mark up their entries). Although SGML is superior in many respects, several factors prompted us to choose standard HTML, including (i) the availability of HTML editors and guides (which makes it easy for authors to produce entries in the proper format without extensive training), and (ii) the availability of good, free HTML search engines. Many other choices about the construction of the encyclopedia were made on the basis of such work-flow considerations.

It should be clear from our brief description that a dynamic encyclopedia poses very interesting questions concerning work-flow management. With adequate financial resources, a project of this type might consider buying, adapting, and/or modifying some off-the-shelf commercial workflow management system.* But few of the systems available seem to be designed to solve the specific problems of the dynamic encyclopedia concept that we wanted to implement. We therefore decided to develop our own solution to the problems of work-flow, one tailored to our specific needs. Having Unix and perl as resources, we have been able to address the special problems that arise in working out the idea of a dynamic encyclopedia.

Next Up Previous

Next: Problems Facing Dynamic Encyclopedias **Up:** Introduction **Previous:** Basic Description of Dynamic Encyclopedias

Eric Hammer and Edward N. Zalta
Wed May 14 17:44:00 PDT 1997

Next Up Previous

Next: Solutions to the Problems **Up:** Introduction **Previous:** Computer Supported Collaborative Work

Problems Facing Dynamic Encyclopedias

There are a number of problems that face the production of a dynamic encyclopedia. First and foremost is the problem of quality control. Whereas all encyclopedias face the problem of choosing high quality board members and authors and the problem of editing entries, the dynamic encyclopedia has the further problem of evaluating changes to entries because authors have the right to access and change their entries when the occasion arises. In a static encyclopedia, once board members and authors are chosen, there is a single further step of quality control which involves the careful editing of submitted entries, so that errors are not published in the fixed medium. In contrast, a dynamic encyclopedia needs a systematic method of evaluating both the new entries posted to the encyclopedia and the subsequent changes made to those entries.

Second, there are the problems involved in producing an electronic work, such as maintaining a uniform entry style and familiarizing authors with markup languages and electronic file transfer.

Third, there are the problems of automating routine editorial and administrative tasks so that the encyclopedia can be set-up and maintained without a large staff. For example, the following processes can be automated: creating accounts for the authors, sending them email about their accounts and the ftp commands they might need, monitoring changes in the content to entries, updating the table of contents, cross-referencing entries, modifying the email aliases (such as the list of the authors' email addresses), notifying the board members that entries for which they are responsible have been changed, etc.

Fourth, there are the issues of copyright. Who should own the copyright to individual entries? Who has the responsibility for obtaining permission to display photographs? What rights do the authors have over their entries? What rights does the encyclopedia have to republish entries in altered form?

Fifth, there are the problems of maintaining the encyclopedia. How often should authors be expected to update their entries? What happens when an author no longer wants to be responsible for updating his or her entry? How do we turn over an entry to a new author? Under what conditions should the encyclopedia allow multiple entries for a single topic?

Sixth, there are the problems of site security. How does one prevent authors or anyone else from gaining access to other parts of the encyclopedia. What if an article is accidentally deleted or damaged?

Finally, there are the issues of citation and digital preservation. How should people using the Encyclopedia cite the articles? What happens if the cited material is subsequently deleted when an author updates or modifies the entry? How will the Encyclopedia be preserved so that the material will always

be available for scholarly research in the same way that the citations to current and past encyclopedias are available?

Next Up Previous

Next: Solutions to the Problems **Up:** Introduction **Previous:** Computer Supported Collaborative Work

Eric Hammer and Edward N. Zalta
Wed May 14 17:44:00 PDT 1997

Solutions to the Problems

Quality Control

Like other high-quality reference works, the authors of entries will be nominated and/or approved by a carefully selected board of editors and the entries themselves will be subject to critical evaluation. But given that the authors have the right to access and change their entries at will, the dynamic encyclopedia has the special problem of how to evaluate updates to entries. Our solution is to monitor changes to each entry and to notify both the Editor and the editorial board member responsible for that particular entry. When notified of a change, the Editor immediately verifies that the entry has not been accidentally or maliciously damaged. More importantly, however, we have written a script that will send out email notices to the relevant board member automatically, not only when the entry is first transferred to the encyclopedia, but also when any changes are made thereafter.* A problem with this procedure is that Board members will be notified even if there have been trivial modifications to entries. Though we have configured our script so that changes that the Editor makes to an entry (to fix typographical errors, HTML formatting errors, etc.) are not reported, we are planning to make our script `smarter', so that it reports to the Board member only significant changes to content made by the author.*

Given that entries in the dynamic encyclopedia can be modified, the authors can improve their entries not only in response to comments from the relevant Board member, but also in response to comments received from colleagues in the field. The latter may also be aware of relevant research not mentioned in the article. However, this introduces a controversial element, since commentators might not be satisfied by the modifications, if any, that authors make in response to their comments and may therefore write to the Editors to make their case. So the Editors and Board members of a dynamic encyclopedia must be prepared to moderate between authors and such commentators.

As a final resort, the Editors can always remove entries should the authors fail to respond to valid criticism, from whatever source.

Production

To solve the problems of production, we have created an annotated HTML sourcefile of a sample entry. The authors may use this sourcefile as a model, replacing its content with their own content.* We created a list of HTML manuals available on the World Wide Web and linked this list into the Editorial Information page of the Encyclopedia. For those authors with HTML experience, we created a empty *template* sourcefile defining the basic entry format, which they can download and simply fill in with their content. Recently, however, a wide variety of HTML-editors have become available and we have created

a special page containing links directly to the download archives containing these editors. So the simplest way for an author with no HTML experience to create an entry would be for him or her to first download Netscape Navigator Gold from the archive, download our HTML template from the Encyclopedia, load the template into Navigator Gold, and then complete their entry simply by selecting text that they have entered and using menu items provided by Navigator Gold to format the text automatically.

Instructions which explain these options are automatically sent to the authors when we set up their accounts. These instructions also explain to the authors how to ftp their entry to our machine and get them into webspace once they have created the HTML sourcefile for their entry and tested it locally on their own computer. We have organized the author accounts in such a way that files transferred into the author's home directory immediately become a part of the encyclopedia.*

Automation

We have automated many of the routine editorial tasks so that the encyclopedia can be administered without a large staff. We have written UNIX and perl scripts to do the following: create accounts for the authors (from keyboard input by the Editors), send the authors email about their account and the ftp commands they might need, take notice of newly submitted entries, monitor changes in the content to entries, manage the cross-referencing between encyclopedia entries by linking keywords of new entries to other entries, modify the email aliases such as `authors' (which contains a list of the email addresses of all the authors), and notify the board members that entries for which they are responsible have been changed. Here is a more detailed description of some of the scripts that have been written:

new-author script: This script will perform the system tasks necessary to add a new author to the encyclopedia. The script automatically sets up an account and home directory for the author with the proper access privileges (i.e., `write' privileges for the author and the editors only), updates the encyclopedia archives (databases with information about authors and their entries), and mails customized information to the author about how to prepare his or her entry, access his or her account, and transfer the new entry to the encyclopedia's machine.

asterisks script: When an entry is assigned but not yet written, the name of the entry in the table of contents is marked with an asterisk. The `asterisks' script notices when an author has ftp'd a new entry to the encyclopedia and then removes the asterisk from the table of contents.

modifications script: This script sends email on a regular schedule to the Editorial Board members indicating which entries have been modified on which date. It determines which Board member is in charge of the entry and updates that Board member's log file with the filename, author, and date the file was modified.

encyclopedia script: This script is a database manager. It extracts and modifies information in the encyclopedia's databases. Among the tasks it performs are: (a) provide information about an author, (b) provide information about a board member, (c) provide information about an entry, (d) list authors by

last name, (e) list keywords to be used for cross-referencing completed entries, (f) add a keyword to the database, (g) remove a keyword from the database, (h) list the entry associated with a keyword, and (i) list all keywords for a given entry.

keyword script: This script verifies and, if necessary, updates the keyword cross-referencing links between entries. When a new entry is submitted, the script notifies the Editor if the author has included keywords for which there are no entries in the table of contents. The Editor can then decide either to add the entry to the encyclopedia (or associate the keyword with an existing entry) or to remove the keyword. The script also verifies that keywords for which authors have included links are linked to the correct entries. Finally, any keyword references to the new entry in previously written entries are automatically linked to the new entry by the script.

It should be mentioned that the selection of keywords is, in the first instance, carried out by the members of the Board of Editors at the stage when they identify topics for inclusion in the Encyclopedia. Since each board member will be chosen for his or her expertise in a philosophy subspecialty, the selection of topics and their corresponding keywords will be driven initially by the perspective that the board members have on their fields. However, the authors will also determine and list the concepts that are essential to understanding the entry they have contributed. When there are discrepancies between the concepts listed by the author and the topics identified by the board member, it will be the job of the Editor to work with these individuals and find the best way to organize the Encyclopedia. These judgements cannot always be made *a priori* and the keyword script identifies when such judgements have to be made.

Copyright Protection

Authors are instructed to read the encyclopedia's copyright notice before transferring their entry to the encyclopedia. The transfer of their entry constitutes an implicit acceptance of the copyright terms stated. The notice has three parts:*

> *Copyright Notice*. Authors contributing an entry or entries to the *Stanford Encyclopedia of Philosophy* retain copyright to their entry or entries but grant to Stanford University and the Editor an exclusive license to publish their entry or entries on the Internet. All rights not expressly granted to the University and Editor are retained by the authors. Copyright of the *Stanford Encyclopedia of Philosophy* itself is held by Stanford University and the Editor. All rights are reserved. No part of the *Encyclopedia* (excluding individual contributions and works derived solely from those contributions, for which rights are reserved by the individual authors) may be reprinted, reproduced, stored, or utilized in any form, by any electronic, mechanical, or other means, now known or hereafter invented, including printing, photocopying, saving (on disk), broadcasting or recording, or in any information storage or retrieval system, other than for purposes of fair use, without written permission from the Editor.

While this part gives authors copyright over their entries, the authors in turn give Stanford University and the Editor an exclusive license to publish the entry on the Internet. Note that to view an entry, the web browser accessing it makes a complete copy of the entry somewhere in the user's machine. We are assuming that such copying of entries qualifies as fair use, and is not ruled out by this portion of the copyright notice.

> *Licensing Agreement*. By contributing to the *Stanford Encyclopedia of Philosophy* authors grant to Stanford University and the Editor a perpetual, exclusive, worldwide right to copy, distribute, transmit and publish their contribution on the Internet. The authors also grant to the University and Editor a perpetual, non-exclusive, worldwide right to copy, distribute, transmit and publish any and all derivative works prepared or modified by the Editor from the original contribution, in whole or in part, by any variety of methods on all types of publication and broadcast media, now known or hereafter invented. Authors also grant to Stanford University and the Editor a perpetual, non-exclusive, worldwide right to translate their contribution, as well as any modified or derivative works, into any and all languages for the same purposes of copying, distributing, transmitting and publishing their work.

This part gives the Editor a license to use and modify submitted entries. The license gives the Editor the exclusive right to publish the entry on the Internet, using whatever technology is currently available, and a non-exclusive right to publish the entry in other media. It also gives the Editor the right to publish portions of an entry. For example, if someone searches the encyclopedia, a search engine will return only those portions of an entry relevant to the search keyword(s). The Editor may also wish to include a portion of an entry in an advertisement for the encyclopedia or in a description of the encyclopedia. Finally, it gives the Editor the right to modify entries, for example, to add links in the sourcefile to other entries or change the way entries are formatted.

> *Statement of Liability*. By contributing to the Encyclopedia authors grant to Stanford University and the Editor immunity from all liability arising from their work. All authors are responsible for securing permission to use any copyrighted material, including graphics, quotations, and photographs, within their articles. The University and the Editor of the Encyclopedia therefore disclaim any and all responsibility for copyright violations and any other form of liability arising from the content of the Encyclopedia or from any material linked to the Encyclopedia.

Because authors have access to their entries, they could include copyrighted material in an entry without the Editor's knowledge. Moreover, there is an interval between the time when an entry is modified and the time when it is checked. This clause protects the encyclopedia and its Editors from any problems with entries arising from these situations.

Maintenance

Dynamic encyclopedias require infrequent but regular maintenance by the authors and Board members,

and require only moderate maintenance by the Editor. Once the Board and authors have been selected and the entries have been written, maintenance of the encyclopedia will primarily involve revisions by authors and examinations of the revisions by the board members. The Editor will only need to handle activities that are not automated, such as communicating with authors and the board concerning any problems that arise, troubleshooting the operation of the encyclopedia, and commissioning new entries as new concepts become important.

We suggest that authors update their entries at least once every year. When an author no longer wishes to maintain his or her entry, the Editors and author have several options. One is to leave it in the encyclopedia, indicating that no further revisions will be made. It may come to be of historical interest. The Editor will then have to commission another author to write a second entry on the same topic. A second option is to transfer maintenance of the original entry to someone else, with the details to be worked out between the original author and the new author.

Security

For the most part, the security problems of a dynamic encyclopedia are the usual security problems of system administration. We have given our authors an `ftp account' on our machine rather than setting up an anonymous ftp server.* So only authors and the Editor can submit or modify entries. Moreover, an author can only modify entries in his or her own home directory.

The only way to protect against malicious and unauthorized access to the machine is to back it up on a regular basis. This also protects the encyclopedia against machine failures. We back up our encyclopedia onto tape and also onto an external hard drive.* This external hard drive has been configured as a boot disk and contains all the system software necessary to run the Encyclopedia. In case the machine that runs the Encyclopedia experiences catastrophic failure, we can install the external hard drive into one of our backup UNIX workstations and reboot, a process that takes fifteen minutes.

Citation and Digital Preservation

We propose that citations to our Encyclopedia conform to the Modern Languages Association style for the citation of electronic sources. The `MLA-style' format for citation is:*

> Author's Lastname, Author's Firstname. "Title of Document." Title of Complete Work (if applicable). Version or File Number, if applicable. Document date or date of last revision (if different from access date). Protocol and address, access path or directories (date of access).

So, for example, a citation to our entry on Bertrand Russell, would look like this:

> Irvine, Andrew. "Bertrand Russell." *Stanford Encyclopedia of Philosophy*. January 28,

1997. http://plato.stanford.edu/entries/russell/ (October 12, 1997)

So that cited material does not disappear when entries are revised, we have decided to fix a quarterly edition of the Encyclopedia and store those editions online on a special `Archive Page' of the Encyclopedia. By checking and citing the most recent quarterly edition, one can be sure that the material being cited won't disappear. Thus, the citation to the entry on Bertrand Russell becomes:

> Irvine, Andrew. Bertrand Russell." *Stanford Encyclopedia of Philosophy*. Fall 1997 Edition. http://plato.stanford.edu/archives/fall1997/entries/russell/ (October 12, 1997)

We are currently exploring whether there are any other alternatives to fixing a quarterly edition.*

Long term preservation of digital information is a somewhat more global problem than secure backup. From the previous section, it should be clear that on any given day, there exist three copies of the Encyclopedia (one on the principal computer, one on external hard drive and one recoverable from the backup tapes).* We maintain an archive of the backup tapes of the Encyclopedia in a separate building. We also have several similar UNIX workstations in the lab housing the main Encyclopedia workstation and each of these computers could serve as a backup machine. As long as we maintain the present edition and past quarterly editions on 3 separate hardware devices (transferring the data to new technology as it becomes available) and follow the security measures outlined above (employing whatever new backup systems become available), we will have adequately safeguarded the material that appears in our Encyclopedia for scholarly research far into the future.

Next Up Previous

Next: Conclusion **Up:** Introduction **Previous:** Problems Facing Dynamic Encyclopedias

Eric Hammer and Edward N. Zalta
Wed May 14 17:44:00 PDT 1997

Conclusion

A dynamic encyclopedia following the above plan, therefore, needs the following administrative staff: an Editor, a computer consultant, and an Editorial Board. The Editor will coordinate the activities of the encyclopedia and maintain the encyclopedia's host machine. Maintenance of the host machine does require some general system administration skills, such as updating the httpd installation and search engines, preparing a sample entry that demonstrates entry style, and maintaining the authors' accounts. A computer consultant will write the scripts described above, oversee the technical development of the project, and apprise the Editor of new developments taking place on the Internet.* Though an advisory board is not necessary, we have one to help us choose the members of our Editorial Board. The Editorial Board will be responsible for soliciting qualified authors to write entries on appropriate topics, and also for evaluating the entries contributed by the authors they solicit.

With a larger budget and support staff, a complete `work-flow' analysis could be developed, which noted and recorded the various (kinds of) transactions between editor and authors and between editor and board member. The Encyclopedia database should keep track of more information about the state of an entry than ours does.* At some point, we plan to develop a program which automatically sends out notices when it is time for the author of a particular entry to update their entry or bibliography. No doubt there are other ways to automate administrative tasks, and when time and money permit, we plan to implement them.

Although we have designed our dynamic encyclopedia principally with an eye toward solving the update problem, such an encyclopedia has other advantages. One is that there are no constraints on the length or number of entries other than that imposed by disk space. This feature easily accommodates multiple entries on a single topic (each reflecting a separate perspective). Another advantage is ease of distribution. By distributing the encyclopedia over the World Wide Web, it becomes accessible to anyone with Internet access. A third advantage is that the pace at which the encyclopedia can be published is limited by the fastest rather the slowest authors. There is no longer a lag between the time the entry is sent to the Editors and the time the entry can be published. Finally, since entries can be improved over time, any biases they may reflect can be found and eliminated. Thus, our solution to the problem of updating encyclopedias also provides a solution to the problem of avoiding bias in encyclopedias.

Eric Hammer and Edward N. Zalta
Wed May 14 17:44:00 PDT 1997

About this document ...

A Solution to the Problem of Updating Encyclopedias

The translation was initiated by Edward N. Zalta on Thu Sep 12 11:44:00 PDT 1996

Eric Hammer and Edward N. Zalta
Wed May 14 17:44:00 PDT 1997

...Zalta.

The authors would like to thank David Barker-Plummer, Mark Greaves, Andrew Irvine, Emma Pease, Susanne Riehemann, and Nathan Tawil for critical suggestions which often led to improvements in the Encyclopedia's design. We would also like to thank the anonymous referees at the journal *Computing and the Humanities* for their suggestions on how to improve the paper.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...succession.

For example, Louis Moréri tried this solution with his *Grand Dictionnaire Historique* of 1674, as did Arnold Brockhaus, in his *Konversations-Lexikon*, 1796-1811.

.
.

...basis.

So, for example, there were 11 supplementary volumes to the ninth Edition of the *Encyclopaedia Britannica* (1875-1889). These constituted the `tenth edition'.

...format.

For example, the second edition of Nelson's *Perpetual Loose Leaf Encyclopaedia* of 1920. The *Encyclopédie française* is still available in loose-leaf format.

.
.
.
.
.
.
.
.
.
.
.
.

...Internet.

We conceived of this solution in our effort to implement John Perry's suggestion that the Center for the Study of Language and Information develop an Internet encyclopedia of philosophy.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

.
.
.
.

...encyclopedia
: The way we have set things up, each entry is given its own subdirectory in the *entries* directory, and that subdirectory is then linked into the author's home directory. So any files that the author transfers into that subdirectory can be accessed over the World Wide Web.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...(CSCW)
: See R. Baecker, *Readings in Groupware and Computer Supported Cooperative Work: Assisting*

Human-Human Collaboration, Morgan Kaufman Press, 1993; R. Baecker, J. Grudin, W. Buxton, and S. Greenberg, (eds.), *Human-Computer Interaction: Toward the Year 2000*, Morgan Kaufman Press, 1995; S. Greenberg, (ed.), *Computer-Supported Cooperative Work and Groupware*, Academic Press, 1991; and I. Greif, (ed.), *Computer-Supported Cooperative Work: A Book of Readings*, Morgan Kaufman Press, 1988.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...workspace'

Only the principal author of coauthored entries will have ftp access to an entry.

.
.
.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...modifications.

To be absolutely safe, the Editor can always invoke superuser priveleges and prevent the author from further altering the file until the editing process is complete and a local backup is made.

.
.
.
.
.
.
.
.
.
.
.

...members
: If an author needs information about what topics the encyclopedia will include, this can be obtained directly by examining the Encyclopedia website or by asking the Editor.

.
.
.
.
.
.
.
.
.
.
.

...system
: See, for example, R. Medina-Mora, T. Winograd, R. Flores, and F. Flores, `The Action Workflow Approach to Workflow Management Technology', in *Proceedings of the (1992) Conference on Computer Supported Cooperative Work*, Association of Computing Machinery Press, 1992. It is unclear to us whether such software as the freely-distributed Egret (http://www.ics.hawaii.edu/~csdl/egret/) or the commercial Lotus `Notes' (http://www2.lotus.com/notes.nsf) would be helpful in this regard.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

.
.
.
.
.
.
.

...thereafter.

We have taken advantage of the UNIX `find' program; it is invoked in a script (`modifications') that runs each night and makes note of which entries have been changed in the past 24 hours. The `find' command is invoked with the following flags:

```
find entries -ctime -1 -name '*.html' -print
```

This causes `find' to print a list of all the HTML files in the `entries' directory that were altered in the last day. For each HTML file in the list, the `modifications' script then determines which Board member is responsible for the entry and places a time-stamped line in that Board member's log file (the log file is simply a list of entries along with the date they were modified and the author of the entry). On a fixed schedule, another script (`send-notifications') then sends the log file to the Board member in an email message. This notifies the Board member that he or she should evaluate the modified entries.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

.
.
.
.
.
.
.
.
.

...author.

For example, we are considering ways to use the UNIX `diff' command to tell us which lines in the file are different from the most recent backup copy. The problem with `diff' is the output, which is difficult to read. But there may be a way to convert the output into a more readable format.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

.
.
.

...content.
: The annotations in the sourcefile consist of both instructions and comments. The instructions tell the authors how to eliminate the dummy content and replace it (by cutting and pasting) with the genuine content of their entries. The comments serve to indicate what the special HTML formatting commands are doing.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...encyclopedia.
: We have things arranged so that the author of the entry `entryname.html' will ftp that entry not

just to his or her home directory, but to the special subdirectory of his or her home directory entitled `entryname'. This latter directory is created by our *new-author* script (see below) as a subdirectory of the *entries* directory and then linked into the author's home directory. Thus, any files the author ftp's into this special subdirectory are available to the httpd server.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...parts:

We would like to thank Andrew Irvine, a Stanford Encyclopedia Board member, for his assistance in the formulation of the three parts to this Statement of Copyright.

.
.
.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...server.

To be precise, we gave each author a login account with a home directory but made it impossible for the author to actually telnet, log on, and run processes on our machine. We did this by assigning a nonexistent UNIX shell `/bin/nosh' as their login shell. When an author ftp's to the machine, the ftp daemon checks to make sure that he or she has been assigned a login shell, but it doesn't require that the shell be a serviceable one. Thus, authors have ftp privileges to and from their home directories, but no login privileges, thereby reducing the load on our server and increasing security. Furthermore, each author's name not only serves to identify his or her home directory but also serves to identify a UNIX `group' (of users), of which only the author and the Editor are members. The author's home directory is assigned to this group, thus allowing only the author and the Editor write privileges to the author's home directory. Even if a password is stolen, at most one entry could be damaged.

.
.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...drive.

The tape backup is on an incremental dump schedule, with a full dump occurring every two weeks. The daily backup onto the external disk makes a new copy of the users' home directories, the HTML sourcefiles of the encyclopedia entries, and the various programs and support data needed to run a web server.

.
.
.
.
.
.
.
.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...is:

See Janice Walker. "MLA-Style Citations of Electronic Sources". Version 1.1. January, 1995 (Rev. 8/96). http://www.cas.usf.edu/english/walker/mla.html (May 12, 1997).}

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...edition:

The idea of fixing a quarterly edition has the added virtue of providing quarterly deadlines for the authors. This might help the Editors set specific goals for the authors and timetables for completing certain sections of the Encyclopedia.

.
.
.
.
.
.
.

...edition:
Actually, there are four copies, for a second copy of each entry is kept in the Editor's home directory on the principal computer. Whenever the Editor makes any modifications to an entry, a copy is immediately placed in this directory. By contrast, the backups on the external drive and tape drive are made once a day, in the early morning hours.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

.

...Internet.
If the Editor has no interest or skills in UNIX system administration, the computer consultant could be assigned these tasks as well.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

...does.
For example, we don't currently record when an entry is first put online, whether the last update was a substantive update to the content or an editorial update to fix poorly written HTML code, the amount of time elapsed since the entry was commissioned, how frequently the entry has been updated, when the Board member responsible for the entry last commented on it, etc. Given our limited budget, we have relied on our email record and and calendar to keep track of many of

these transactions.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

Ed Zalta
Thu Sep 12 11:44:00 PDT 1996

Why Philosophy Needs A `Dynamic' Encyclopedia of Philosophy

John Perry and Edward N. Zalta
CSLI/Philosophy Department
Stanford University

Contents

Introduction

At the Center for the Study of Language and Information (CSLI), we have developed a solution to the problem of updating encyclopedias, namely, a `dynamic' encyclopedia that is published on the Internet. Unlike static encyclopedias (i.e., encyclopedias that will become fixed in print or on CD-ROM), a dynamic encyclopedia allows entries to be updated, thereby becoming *responsive* to new research and advances in the field. A dynamic encyclopedia gives the authors direct electronic access to their entries and the means to update them whenever it is needed, and it does this without compromising the quality of the entries. Whereas static encyclopedias must publish supplements or an entire new edition to become current, a dynamic encyclopedia simply *evolves* and quickly adapts to reflect advances in research. The process of updating an encyclopedia never ceases, and the very concept of a dynamic encyclopedia takes account of this fact.

The *Stanford Encyclopedia of Philosophy* is our working prototype of a dynamic encyclopedia. This prototype implements the technical specifications developed in the paper `A Solution to the Problem of Updating Encyclopedias', which was coauthored by Eric Hammer and Edward N. Zalta (*Computers and the Humanities*, Volume 31/1, 1997, pp. 47-60).

We are now working towards the goal of turning this working prototype into a *mature* dynamic encyclopedia that will be useful both to professional scholars and the general public. It will not be a `finished' product because dynamic encyclopedias are never finished. But within a few years, we will have refined the operation of this new and innovative kind of dynamic reference work so that it will continue to prove its worth far into the future. We believe that the principles upon which our

encyclopedia is based will contribute to a reconceptualization the nature of reference works for the age of the Internet.

Basic Description

To implement our dynamic encyclopedia, we connected a multi-user Unix workstation (plato.stanford.edu) to the Internet and installed a World Wide Web server. We then created a cover page, a table of contents, an editorial page, and a directory entitled *entries* accessible to the web server. We recruited (and are still recruiting) Editorial Board members for the job of identifying topics, soliciting authors, and reviewing the the entries and updates when they are received. These Board members have begun recruiting authors who have a strong interest in and commitment to the topics on which they write and who will be motivated to keep their entries abreast of the latest advances in research.

However, the innovative technical feature of our encyclopedia is that authors are given a "file-upload" account on plato.stanford.edu, i.e., the computer that runs the encyclopedia's World Wide Web server. Once an Editorial Board member decides on a topic and finds or approves an author to write it, he or she then forwards the information (via email) to the Editor of the encyclopedia, who then creates a file-upload account for the author on plato.stanford.edu and sends the author detailed information on how to prepare the entry and upload the entry (or update) when it is ready. When an author uploads an entry or an update (using a web browser) to his or her private directory on plato.stanford.edu, only the editorial staff can view it (they have password protected access). The Editorial Board member responsible for that entry is automatically notified that the entry has been put online or changed. The Board member then inspects the new material, and if he/she approves it, the new material is then published in public webspace; otherwise, the author is sent suggestions for improvements (and in some cases, the entries must be rejected and a new author must be commissioned).

Innovative Features

The most obvious innovative feature of an encyclopedia fitting this description is that it will not go out of date. When new ideas are published in books and journals of philosophy, the authors of our encyclopedia can summarize the ideas and update their entries. Since the encyclopedia will be accessible to web-browsers such as Netscape or Internet Explorer, a wide audience (academics, students, and the general public) may become informed of the latest advances in thought more quickly. We believe that this will speed up the dissemination of new philosophical ideas to a wide audience.

The other innovative features of a dynamic encyclopedia that has been organized on the above plan are:

1. The entries in a dynamic encyclopedia can be *improved* and *refined* over the course of time. As Board members and colleagues read the entry and give the author feedback, the author can enhance the article to make it more readable, make it more logical in form, repair typographical errors, etc.

2. The encyclopedia eliminates the often large span of time that appears between productions of large-scale `static' encyclopedias. The Macmillan *Encyclopedia of Philosophy* (edited by Paul Edwards) was published in 1967. The Routledge *Encyclopedia of Philosophy* (edited by Edward Craig) will come out in 1998. During the 30 year gap, no up-to-date full-scale encyclopedia has been available. In 2003, when the Routledge Encyclopedia starts to show a few signs of going out of date, ours will still be current. In 2008, the differences will be even more striking.
3. The encyclopedia eliminates the lag time between the writing and publication of the entries. There won't be the usual two year delay between the time the article is written and the time the encyclopedia is published. Some of the articles written for the forthcoming Routledge Encyclopedia will already be several years old when they appear in print in 1998.
4. The encyclopedia can be expanded indefinitely; there is no limit to its inclusiveness or size. New or previously unrecognized topics within a given discipline can be included as soon as they are discovered and judged to be important.
5. For certain controversial topics, the encyclopedia can include more than one entry on a single topic, each one representing a different point of view. This feature reduces the bias that affects other encyclopedias. The forthcoming Routledge Encyclopedia, by contrast, will contain just a single entry on each topic. Biases of the author may sometimes go unremarked and unsophisticated readers will be unaware of the alternative point of view. We are in a position to publish two or more entries on controversial topics in the philosophy of mind, medical ethics (euthanasia, abortion), political philosophy, etc.
6. Not only will the Bibliography sections be up to date when entries first come online (since there is no lag time between the writing and publication of the entry), but also, for the first time, encyclopedia entries can be accompanied by more comprehensive bibliographies. Print- and CD-ROM-based encyclopedias have a fixed amount of space to publish bibliographies and so the authors can list only a small number of first-rank articles in the field. Our authors can give a complete listing of first-rank articles and books. If the list proves to be overwhelming, they can further divide the list into first-rank articles and books that must be read and those that should be read. If the list of first-rank articles and books is not overwhelming, the author may develop a more inclusive bibliography. The bibliographies of our entries will therefore become *the* place to look for references to a given philosophical topic. We expect our authors to update their bibliography once a year.
7. The encyclopedia will have an *automatic* cross-referencing system. When new entries come online, the keywords in the section `Related Entries' will be automatically linked to the other entries in the Encyclopedia and the keyword in the other entries referring to the new entry will be linked to the new entry.
8. The encyclopedia will include not only internal links that cross-reference the entries, but each entry will contain *updatable* external links in a special section called `Other Internet Resources'. The authors may judge which other Internet sites are worthy of being linked into this section of their entry. For example, Andrew Irvine's entry on Bertrand Russell has both internal and external links. In the section `Other Internet Resources' in this entry, there are links to the online Russell Archives at McMaster University, the Bertrand Russell Society web page, the web page of the Bertrand Russell Editorial Project, etc. *These updatable links cannot be included in print- or CD-ROM-based encyclopedias,* much less be maintained and kept up to date. The entry on Russell in

the forthcoming Routledge Encyclopedia will lack such connections to these invaluable Internet resources!

9. The internet format avoids most of the expenses of production and distribution connected with a major reference work. It eliminates many of the expenses of producing a printed document or CD-ROM: typesetting, copy-editing, printing, and distribution expenses are no longer necessary. We provide authors with an HTML template so that they can easily typeset their entries in HTML using an HTML-editor such as Netscape Navigator Gold (there are now over 10 such HTML-editors on the market, many of which are free for academic use).(1) The editing will be done collaboratively by the editorial assistants, authors, board members, and Editor as a routine function of the quality control systems.
10. A dynamic encyclopedia can change in response to new technology as new tools and languages develop. For example, if Java becomes more widespread, Encyclopedia entries on neural nets, turing machines, prisoner's dilemmas, evolutionary game theory, etc., can include demonstrations and `applets' that the user of the Encyclopedia can download to learn interactively the concepts being discussed.
11. Statistics software can process the information in the access log of the encyclopedia web server and identify which sites users access it from, which entries they access most, *which topics they search for*, etc. Such information can help inform decisions about which entries to add, where to advertise, etc. This information is absolutely invaluable to the editors of an encyclopedia, and yet, no such access statistics can be maintained for static encyclopedias. Since numerous copies of static encyclopedias exist, there is no way to track the frequency individual entries are read. Our access log, which excludes the accesses by our own computers and recognizable WWW indexing worms, indicates that just in the period from September 15, 1996 to September 20, 1997, there were over 19,000 accesses from the U.S. educational (edu) domain, over 24,000 from the commercial (com) domain, over 1000 from the U.S. government (gov) domain, over 300 from the U.S. military (mil) domain, close to 20,000 from the network (net) domain, and over 1000 from the non-profit organization (org) domain (another 27,000 accesses were unresolved). These logs also show that during that same period, the Encyclopedia has been accessed from over 90 different countries, that Andrew Irvine's entry on Bertrand Russell has been accessed nearly 10,000 times, and that our encyclopedia has recently started to average over 8000 accesses per week. (See the Recent Access Statistics page (General Statistical Summary for the period of September 15, 1996 -- September 20, 1997), which is linked into the Editorial Information page of the Encyclopedia.)

No print or CD-ROM based edition of an encyclopedia has the above features and this includes the Routledge Encyclopedia forthcoming in 1998.

Future Significance

What is the future significance of this project, in the context of present availability of encyclopedias in philosophy? In 1998, we will see the publication of a major new reference work, the vast Routledge Encyclopedia of Philosophy. A new supplementary volume for 1967 Macmillan Encyclopedia was

released last year. At least ten other reference works in philosophy have been released recently.[(2)] It might seem, then, that the reference needs of philosophers are well-served, and that there is little need for the project of an online Encyclopedia of Philosophy.

This is not so, however. The last major, comprehensive encyclopedia before Routledge's was the 1967 Macmillan Encyclopedia. Even in a relatively slow moving field such as philosophy, 30 years is a long time to wait between encyclopedias. As fine as the Macmillan Encyclopedia was, by the time it was five years old, its value as a reference work for the philosophy profession had greatly diminished. By the mid to late 1970s, a student might go to an encyclopedia looking for information about the Kripke/Donnellan theory of reference, the Lewis and Armstrong identity theories of mind, Putnam's functionalist theory of mind, or Rawls's theory of justice and find nothing of value in the Macmillan Encyclopedia. So the philosophy community went for 20 -- 25 years before the recent handbooks and dictionaries attempted to fill the gap.

In this context, the Routledge Encyclopedia will be a welcome addition and a wonderful new tool. But within the next five to ten years, it too will be out of date, along with the other reference works in philosophy cited. There are more professional philosophers, more journals, more articles, more books than every before in history. There is much valuable and interesting work being done that will be cited in undergraduate classes long before another encyclopedia is due on the present thirty year cycle.

Moreover, the philosophy profession cannot be confident that there will ever be another major print encyclopedia. The Routledge encyclopedia represents an enormous investment. The cost of purchase is also enormous, about $2500 for the CD-ROM edition. Routledge will not be quick to replace it. And yet even the CD-ROM technology is already dated for reference tools. By contrast, the Internet is ideal for reference tools, but not ideal (as far as one can see at present) for traditional private print publishers to make money.

Those academic disciplines deemed more scientific and essential to national goals or business interests than philosophy will have many more ways to subsidize their reference needs. We believe a partnership between CSLI and the philosophy profession to develop an online philosophy encyclopedia represents the best chance for the profession not to be left behind and without adequate reference tools in the future.

Our online encyclopedia, by its nature, will never be `finished'. But we project that it will be mature, in the sense of having virtually all presently planned entries complete, in about four years. Within two years thereafter, its inventory of articles should be very close to that of the Routledge. That means our Encyclopedia will be available as a full-fledged alternative for philosophy students *at about the time* that the Routledge encyclopedia begins to show signs of age, and at a time when students will increasingly be familiar with the web and expect to use it for reference work. Because of the fact that our Encyclopedia is `alive', it will stay current.

Thus, we propose not to redo the work that has already been done for the Routledge Encyclopedia, but to to use new technology to begin working on the next generation of reference tools for philosophers and

their students.

Funding the Encyclopedia

Most of the effort that goes into the creation and maintenance of the Encyclopedia will be donated by those who welcome the opportunity to advance the serious study of their favorite topics and figures. Our plan is that the authors, the Board of Editors and the Advisory Board will be unpaid (unless the Encyclopedia resolves its funding problems). For the foreseeable future, CSLI will provide an office for the Encyclopedia staff and computer, backup systems, and cover other indirect costs. If in the future CSLI goes out of existence or is unable to provide this support, it seems likely some other institution could be found that would do so, either here at Stanford or on some other campus. The amount of space and computer power are not that great, by the standards of late twentieth century American academia.

However, we do not believe the Encyclopedia can function on a completely volunteer basis. At a minimum, we must pay for a part-time Editor, part-time editorial assistants and at least one part-time computer programmer. Some equipment and supplies have to be purchased and telephone and other services (e.g., networking) need to be paid for. We estimate costs as follows:

1. Staff Salaries: $100,000
2. Staff Benefits: $25,000
3. Supplies, equipment, services, etc.: $25,000

----

4. Total: $150,000

Print encyclopedias are funded by sales to individual and institutional users. Our hope is our dynamic encyclopedia will remain free to all individual users. There should be as few obstacles to the serious study of philosophy as possible. We hope that the online encyclopedia will be available to anyone anywhere with access to the net.

The institutions most closely associated with users include Philosophy departments, universities (and their libraries), and professional organizations for philosophers (such as the APA, CPA, AAP, etc.). Universities should be able, in the long run, to support online encyclopedias and other online resource materials from the same budget lines used for print materials that now serve these purposes. One strategy for funding the Encyclopedia would be to charge university libraries a nominal yearly access (subscription) fee. This subscription fee would give computers based at those universities the right to access the Encyclopedia. If 1000 universities each pay $150/year for unrestricted access to the Encyclopedia, the Encyclopedia's projected annual budget would be covered. The $150/year access fee could simply represent an average---the fee could be set so that universities with larger (smaller) budgets or student populations pay a proportionally larger (smaller) access fee. This would keep the Encyclopedia free to the public.

However, in the short run, we do not expect to be able to tap these resources. Our Encyclopedia won't be

mature for another 3 - 5 years. Right now, we want University libraries to use their funds to purchase the much anticipated Routledge Encyclopedia. Until our encyclopedia has matured, we will need to look elsewhere for support.

We feel that Philosophy departments and professional organizations are natural sources of funding. Philosophy departments differ enormously in the amount of discretionary funds available to them, with the mean amount extremely low. Some departments do have such funds, however, and a few hundred dollars from a number of such departments could help significantly. However, we prefer not to approach individual departments until there is mature product that will be useful to them. (We might make an exception for a number of the most well-endowed departments, some of which are represented on our Editorial Board.) At the startup phase, it seems more appropriate to approach those institutions part of whose mission, or enlightened self-interest, is to help projects like this get off the ground. Within philosophy, this means professional organizations like the APA. At this point, the Pacific APA and the CPA have made generous contributions, and we have approached the Central and Eastern APA. Of course, a commitment at the national level would be preferable.

Among the funding sources that are not directly connected to Encyclopedia users, we distinguish between government sources, foundations, and businesses. We have received a 2-year grant from the NEH. They will give us $120,000 to develop the Encyclopedia, to be delivered over two years (AY 1998-99 and 1999-2000). We will apply for an NSF grant through the Digital Libraries II initiative. If this proposal proves successful, this would take care of our funding needs until 2003, at which point the Encyclopedia will be close to maturity.

We are just beginning to solicit funds from appropriate foundations. The development and fund-raising staffs at CSLI and Stanford are being most helpful in this respect.

It seems to us that Netscape, Microsoft, Yahoo, and other companies with a stake in the future of the Internet might have multiple reasons for investing in our Encyclopedia. In this regard, it would be very helpful to discover philosophers within such companies who would be willing to help us identify and approach the relevant people with funding authority.

(Some of these companies have been indirectly funding the Encyclopedia for the past two years through CSLI's Industrial Affiliates program. CSLI was founded in 1983 with a grant from the Systems Development Foundation. These funds were exhausted long ago. Since 1993, the Industrial Affiliates program has been the principal source of CSLI's operating budget.)

A final possibility for money from the private sector is advertising. One can imagine tasteful and wholly appropriate links from the Encyclopedia home page to the websites of publishers of philosophy books. Whether these publishers will find this to be a economical use of their advertising budgets remains to be seen.

We would welcome comments from the APA membership on any aspect of fund-raising. Anything from

leads concerning fund-raising in industry to comments about the appropriateness of advertising would be welcome.

Footnotes

1. When using such HTML-editors, the author simply selects (highlights) text and then uses `menu' functions to format the text in various ways (such as to italicize, to put the text in a list environment, to create a link and use the text as the label on that link, etc). The HTML-editor then creates the appropriate HTML code in the underlying HTML file. (return to text)

2. The list includes: *A Companion to the Philosophy of Language* (Blackwell, 1997), *Companion Encyclopedia of Asian Philosophy* (Routledge, 1997), *Routledge History of Philosophy* (Routledge, 1997), *Oxford Dictionary of Philosophy* (Oxford University Press, 1996), *Blackwell Companion to Philosophy* (Blackwell, 1996), *Biographical Dictionary of Twentieth Century Philosophers* (Routledge, 1995), *Cambridge Dictionary of Philosophy* (Cambridge University Press, 1995), *Oxford Companion to Philosophy* (Oxford University Press, 1995), *A Companion to Metaphysics* (Blackwell, 1995), *A Companion to Epistemology* (Blackwell, 1992), *A Companion to the Philosophy of Mind* (Blackwell, 1994), and *Handbook on Metaphysics and Ontology* (Philosophia Verlag, 1991). (return to text)

Stanford Encyclopedia of Philosophy: A Dynamic Reference Work

Edward N. Zalta[*], Colin Allen[†], Uri Nodelman[*]
[*]Stanford University and [†]Texas A&M University
zalta@stanford.edu, colin›allen@tamu.edu, nodelman@stanford.edu

The primary goal of the Stanford Encyclopedia of Philosophy project <http://plato.stanford.edu/> is to produce an authoritative and comprehensive reference work devoted to the academic discipline of philosophy that will be kept up to date *dynamically* so as to remain useful to those in academia *and* the general public. To accomplish this goal we have designed and implemented web-based software by which academic philosophers can collaboratively write and maintain such a 'dynamic reference work'. Our implementation has features that are not found in any other online reference work in any discipline, and that enable the profession of philosophy to maintain such a reference work without the cost or level of staff support required for traditional reference work publishing.

A *dynamic reference work* is a new concept in digital libraries technology. We define this concept as follows: (1) it is published in a continuously revisable electronic medium, (2) it offers a comprehensive set of entries on topics in a target discipline, (3) it *provides the authors of the entries with electronic access* to the reference work's central web server, so that they can remotely edit and update private copies of their entries and submit them for publication according to a regular update schedule and at any other time it becomes necessary to revise, (4) it maintains quality by way of a distinguished Board of Editors, the members of which commission the entries and referee both the initial versions of the entries and subsequent substantive modifications, *prior to publication* on the web, and (5) it creates, and makes publicly available, archives of the entries on at least a quarterly basis (i.e., these contain fixed versions of the entries, which can be cited in scholarly publications). A dynamic reference work based on this model constantly evolves and becomes *responsive* to new research.

Clauses (3) and (4) of this definition indicate that a *dynamic* reference work is not merely a revisable reference work or one that is published online. Successful implementation of the dynamic aspects of this definition depend upon the ease with which the authors, subject editors, and the principal editor have access to the tools and information that allow entries at all stages of the work flow to be managed asynchronously. In such an environment, each entry has its own deadlines and it is necessary to track electronically the location of every entry in the work flow and provide automated reminders to individuals with work pending. Our success in organizing 75 distinguished editorial board members to oversee 450 expert authors, working to a set of deadlines specified by our work flow, distinguishes our project from other, less formal, on-line encyclopedia projects.

We will demonstrate how our password-protected web interfaces and our back-end processing system and workflow system work together to facilitate this collaborative effort. Our system has several unique features designed to simplify the collaborative production of a dynamic reference work including tools for web-based editing of html source and side-by-side comparison of file versions. The web interface for authors allows them to: (1) download our templates and style sheets, (2) to upload their new entries into a private area of our web server, (3) to remotely edit copies of their entries which are stored in this private area, and (4) to submit their entries for editorial review. The web interface for subject editors allows them to: (1) add important or new topics in philosophy to our database, (2) commission authors to write entries, (3) examine and comment on the new or revised entries submitted by the authors prior to publication, (4) display, in their web browser, revised and original entries side-by-side, *with the differences highlighted*, and (5) accept or reject entries and revisions. The principal editor has a web interface, by which this collaborative workflow system is administered. Among other things, it tracks deadlines for each stage of the publication and revision process for each entry.

We have designed our software in such a way that authors and subject editors require only basic computer skills to participate effectively in the project. Authors may also use any html-authoring software to create their entries, which then are automatically formatted by our software and may be easily maintained by authors through the web interface. (This author-friendliness is an advantage of our html-based system over currently available xml-based alternatives.) Our use of cgi scripts also means that our browser-based interfaces are accessible from any internet-connected workstation without the installation of any special software. Thus the system enables an international collection of authors and subject editors to solicit, submit, review, revise, publish, and update high quality articles under the guidance of a principal editor, with minimal support from a programmer and other staff.

The technical specifications of our system are described in the project description of our NSF/DLI-2 proposal (accepted in August 2000), which may be found at the URL: <http://plato.stanford.edu/NSF/project-description.pdf>

This research is supported by NSF grant #IIS-9981549.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
ACM 1-58113-345-6/01/0006.

Stanford Encyclopedia of Philosophy
Editorial Board/Subject Editors

17th Century Philosophy	Alan Nelson (University of California/Irvine) Nicholas Jolley (University of California/Irvine)
18th Century Philosophy	Margaret Atherton (University of Wisconsin/Milwaukee) David Owen (University of Arizona)
19th Century Philosophy	Allen Wood (Stanford University)
Continental Philosophy	Robert Pippin (University of Chicago)
20th Century Philosophy	Andrew Irvine (University of British Columbia/Vancouver) Fred Kroon (University of Auckland) Huw Price (University of Sydney)
Continental Philosophy	Dagfinn Føllesdal (University of Oslo/Stanford University) Thomas Flynn (Emory University)

CURRICULUM VITAE

ALASDAIR IAN FENTON URQUHART
Professor of Philosophy
University of Toronto
Date of latest revision: April 1998

A. Biographical Information

Personal
Date of Birth: 20 December, 1945

Citizenship: British (Landed Immigrant 1970)

Home address: 54 Boustead Avenue, Toronto, Ontario, M6R 1Y9

Home phone: 767-5240 (unlisted)

University address: Department of Philosophy, University of Toronto,

215 Huron Street, Toronto, Ontario M5S 1A1

Office phone: 978-6789

Degrees

M.A. Hons., University of Edinburgh, 1967

M.A. University of Pittsburgh, 1969

Ph.D. University of Pittsburgh, 1973

Title of Ph.D.Thesis: "The Semantics of Entailment"

Supervisors: Profs. Nuel D. Belnap Jr. and Alan Ross Anderson

Employment

Present Appointment

Professor, University of Toronto, 1986 -

Associate Professor, University of Toronto, Erindale College, 1975-86

Date of appointment to Graduate School, 1973

Date of tenure award, Spring, 1975

Assistant Professor, University of Toronto, Erindale College, 1973-75

Lecturer, University of Toronto, Erindale College, 1970-73

Teaching Fellow, University of Pittsburgh, 1967-70

Professional Affiliations and Activities

Consulting editor Journal of Symbolic Logic 1983-89

Former editor of Canadian Philosophical Monographs

Associate of the Institute for the History and Philosophy of Science and Technology

Referee for Canada Council, NSERC, NSF etc.

Referee for JSL, JPL, Studia Logica etc. etc.

Member, Executive Committee, Association for Symbolic Logic 1987-90 Member, Advisory Editorial board, Studia Logica 1991- Member, ASL committee on meetings in N. America 1992- Member, ASL nominating committee, 1993 -

Programme comittee, joint meeting of ASL and SEP, York University May 1993.

Board of consultant editors, Russell. Editor, Lecture Notes in Logic (Springer Verlag), 1994-

Editor, Trends in Logic, 1994-

Consulting editor, Studia Logica.

Member of editorial board, International Studies in the Philosophy of Science.

Editor for non-classical logics, Stanford On-Line Encyclopedia.

B. Academic History

a. Research Endeavours

Non-classical logics, lattice theory, philosophy of logic, foundations of mathematics, history of logic, theory of computation, computational complexity theory.

b. Research Awards

Canada Council Leave Fellowship 1976-77 (one-year) - $9,200.
SSHRC Leave Fellowship 1983-84 $11,500.
SSHRC Research Grant 1984-85. Project: ``Investigations in logic" $6,920.
SSHRC Research Grant 1986-87. Project: "Studies in Complexity Theory", $10,185.
SSHRC Research Grant 1986-87. Project: "Editing Volume 4 of Bertrand Russell's Papers", $45,615.
SSHRC Research Grant 1989-91. Project: "Editing Volume 4 of The Collected Papers of Bertrand Russell", $124,700.
SSHRC Research Grant 1991-2. Project: "Editing Volume 4 of the Collected Papers of Bertrand Russell." $46,490.
SSHRC Research Grant 1992-3. Project: "Editing Volume 4 of the Collected Papers of Bertrand Russell." $10,400.
NSERC Operating Grant 1991-4. $24,000 p.a.
NSERC Operating Grant 1994-8. $27,000 p.a.

C. Scholarly and Professional Work

Refereed Publications

a. Articles in refereed journals

``Semantics for relevant logics'', Journal of Symbolic Logic, Vol. 37, No. 1, March 1972.
``Completeness of weak implication'', Theoria Vol. 37 (1971).
``A semantical theory of analytic implication'', Journal of Philosophical Logic, Vol. 2, April, 1973.
``An interpretation of many-valued logic'', Zeitschrift fur mathematische logik und grundlagen der Mathematik, Vol. 19, pp. 111-114.
``Free distributive pseudocomplemented lattices'', Algebra Universalis, Vol. 3, pp. 13-15.
``Free Heyting algebras'', Algebra Universalis, Vol. 3, pp. 94-97.
``Implicational formulas in intuitionistic logic'', Journal of Symbolic Logic, Vol. 39, No. 4, December, 1974.
``Popper's logical conceptions'', Communication and Cognition, Vol. 8 (1975), 237-242.
``Proofs, snakes and ladders'', Dialogue, Vol. 13, pp. 723-731 (1974).
Critical notice of Richard Montague's Formal Philosophy, Canadian Journal of Philosophy, Vol. 4, pp. 573-578 (1975).
Critical notice of Entailment, Vol. 1, by Anderson and Belnap. Canadian Journal of Philosophy, Vol. 7, 405-411.
``A Finite Matrix Whose Consequence Relation is not Finitely Axiomatizable'', Reports on Mathematical Logic, Vol. 9 (1977), 71-73.
Review of Meaning and Modality by Casimir Lewy, Journal of Philosophy, Vol. 75 (1978).
``A topological representation theory for lattices'', Algebra Universalis, Vol. 8 (1978), 45-58.
``Distributive lattices with a dual homomorphic operation', Studia Logica, Vol. 38 (1979), 201-209.
``Equational classes of distributive double p-algebras'', Algebra Universalis, Vol. l4 (l982), 235-243.
``Distributive lattices with a dual homomorphic operation II'', Studia Logica, Vol. XL, 1981-4, 391-404.
``The undecidability of entailment and relevant implication'', Journal of Symbolic Logic, Vol. 49 (1984), 1059-1073.
``Relevant implication and projective geometry'', Logique et Analyse, special issue on Canadian logic, Vol. 103-104, September 1983, pp. 345-357.
Critical notice of Handbook of Mathematical Logic (ed. Barwise) Canadian Journal of Philosophy, Vol. XIV, December 1984, 675-682.
Critical notice of Beyond Analytic Philosophy by Hao Wang, Canadian Journal of Philosophy, Vol. 17, 477-482 (June 1987).
``A Contractionless Semilattice Semantics'', (with S. Giambrone & Meyer), Journal of Symbolic Logic, Vol. 52 526-529 (June 1987).
``Proof Theories for Semilattice Logics'' (with S. Giambrone), Zeitschrift fur Math. Logik und Grundlagen der Mathematik, Vol. 33, 433-9 (1987).
``Hard examples for resolution'', Journal of the Association for Computing Machinery, Vol. 34, 209-219, (1987).
``Further Results on Proof Theories for Semilattice Logics'' (co-authors Meyer, Giambrone, Martin), Zeitschrift fur Math. Logik u. Grundlagenforschung, Vol. 34, 1988, 301-4.
``The Complexity of Gentzen Systems for Propositional Logic'', Theoretical Computer Science, Vol. 66,

1989, 87-97.
``What is relevant implication?'' in: Directions in Relevant Logic ed. by Norman and Sylvan (Kluwer 1989), 167-74.
``Functional Interpretations of Feasibly Constructive Arithmetic'' (extended abstract co-authored with S.A. Cook), 21st Annual ACM Symposium on theory of computing May 1989.
``The Logic of Physical Theory'', in: Physicalism in Mathematics, ed. A.D. Irvine, Kluwer 1990, 145-154.
``The complexity of decision problems in relevance logic'', in Truth and Consequences ed. by Dunn and Gupta, Kluwer 1990, 61-76.
``Complexity of proofs in classical propositional logic'', in Logic from Computer Science ed. Y.N. Moschovakis, Springer-Verlag 1992, 597-608.
Review of papers by Arnon Avron, J. Symbolic Logic, Vol. 57 (1992), 1481-2.
``Approximations and small-depth Frege proofs'' (co-authors Stephen Bellantoni and Toni Pitassi), extended abstract in conference proceedings Structures in Complexity Theory 1991, Springer Lecture Notes in Computer Science 1991.
``The relative complexity of resolution and cut-free Gentzen systems'', Annals of Mathematics and Artificial Intelligence, 6 (1992), 157-68.
``Approximations and small-depth Frege proofs'' (co-authors Stephen Bellantoni and Toni Pitassi), SIAM J. of Computing, Vol. 21 (1992), 1161-79.
``The Complexity of the Hajos calculus'' (co-author Toniann Pitassi), Proceedings of Symposium on the Theory of Computing, Pittsburgh 1992. See below for published complete version.
``Failure of interpolation in relevant logics'', J. of Philosophical Logic, Vol. 22 (1993), 449-479.
``Functional interpretations of feasibly constructive arithmetic'' (co-author S.A. Cook), Annals of Pure and Applied Logic, Vol. 63 (1993), 103-200.
``Russellian Propositions'' (co-author J. Pelham), Proceedings of the conference on Logic, Methodology and Philosophy of Science, Uppsala 1991, 307-326.
``Upper and lower bounds for tree-like cutting-plane proofs,'' (co-authors Russell Impagliazzo and Toni Ann Pitassi), Ninth I.E.E.E. Symposium on Logic in Computer Science (1994), pp. 220-228.
``Decision problems for distributive lattice-ordered semigroups'', Algebra Universalis, Vol. 33 (1995), 399-418.
``The Complexity of the Hajos calculus'' (co-author Toniann Pitassi). SIAM J. of Discrete Mathematics, Vol. 8 (1995), 464-483.
``Duality for algebras of relevant logics'', Studia Logica, Vol. 56 (1996), 263-276.
``G.F. Stout and the theory of descriptions,'' *Russell* Vol. 14, 163-171.
``The complexity of propositional proofs,'' *Bulletin of Symbolic Logic*, Vol. 1 (1995), 425-467.
``Simplified lower bounds for propositional proofs,'' (co-author Xudong Fu), *Notre Dame J. of Formal Logic*, Vol. 37 (1996), 523-544.
``The number of lines in Frege proofs with substitution,'' *Archive for Mathematical Logic*, (1997), Vol. 37, 15-19.
``Beth's definability theorem in relevant logics,'' forthcoming in Festschrift for Helena Rasiowa.

b. Books and Chapters in books

Temporal Logic (joint author with Nicholas Rescher), Springer Verlag New York and Vienna 1971.
Section on Many Valued Logic in Gabbay and Guenthner (eds.), Handbook of Philosophical Logic, Vol. III. Reidel l984.
Sections 47 and 65 of Entailment, Vol. 2, by Anderson, Belnap and Dunn. Princeton University Press 1992.
The Collected Papers of Bertrand Russell, Volume 4 : Foundations of Logic 1903-05 (Editor).
``Zeit und Zeitlogik'' (co-author Nicholas Rescher), in Zustand und Ereignis, ed. by Bertram Kienzle, Suhrkamp 1994 (partial translation of Temporal Logic).

Non-refereed Publications

a. Articles

About 100 short reviews on papers in logic and mathematics Mathematical Reviews, 1973-85.
Review of The Paradox of the Liar, ed. by Robert Martin, Dialogue, Vol. 10, (1971), pp. 823-825.
Review of The Development of Mathematical logic, by R.L. Goodstein, Historia Mathematica, Vol. 11, pp. 212-214.
Review of Distributive Lattices, by Balbes and Dwinger, Journal of Symbolic logic, 40 (1975).
Review of Classical Propositional Operators, by Krister Segerberg, Canadian Philosophical Review, Vol. III, No. 6.
``Intensional Languages via Nominalisation'', Pacific Journal of Philosophy, 1981.
Review of Mathematics in Philosophy, by Charles Parsons, History and Philosophy of Logic, (1985).
Review of two papers on symbolic logic, JSL 1985.
Review of Routley et al. "Relevant Logics and their Rivals I" Studia Logica (1988).
Review of Rachel Garden "Modern Logic and quantum mechanics", Journal of Symbolic Logic, Vol. 53, 648.
Review of J.M. Dunn, J. of Symbolic Logic, Vol. 54 (1989), pp. 615-16.
``Russell's zigzag path to the ramified theory of types'', Russell, N.S. Vol. 8 (1988), pp. 82-91.
Review of Troelstra and van Dalen "Constructivism in Mathematics Volume 1", Studia Logica 1989.
Review of Troelstra and van Dalen "Constructivism in Mathematics" Volume 2, Studia Logica, June 1991.
Review of Per Martin-Lof "Intuitionistic Type Theory", Studia Logica 1990.
Review of Stephen Read "Relevant Logic", History and Philosophy of Logic, Vol. 11 98-99.
Review of Jon Barwise "The situation in logic", Can. Phil. Rev., Vol. X, 96-8.
Review of Stuart Shapiro "Intensional Mathematics", Studia Logica, April 1990.
Review of paper by D. Deutsch, J. Symbolic Logic, Vol. 55, 1309-10.
``Functional interpretations of feasibly constructive arithmetic'' (extended abstract) in Feasible Mathematics,ed. Buss and Scott, Birkhauser 1990, 97-8.

Review of Hermes No. 7 , in Russell, Vol. 11 (1991), 103-5.
Review of M. Detlefsen, "Proof and Knowledge in Mathematics", Canadian Phil. Reviews, Vol. 12 (1992), 237-8.
Review of Rodriguez-Consuegra, "The Mathematical Philosophy of Bertrand Russell", Philosophia Mathematica Vol. 1 (1993), 90-93.
Review of Hartry Field "Realism, mathematics and modality", History and Philosophy of Logic, Vol. 14 (1993), 117-119.
Review of "Lectures on Linear Logic" by Anne Troelstra, Can. Phil. Reviews, Vol. 13, 126-128.
Review of "Russell's Idealist Apprenticeship" by Nicholas Griffin, Russell, Vol. 13, 104-108.
Review of "Substructural Logics", History and Philosophy of Logic, Vol. 16 (1995), 138-9.
Review of G. Malinowski, "Many-valued logics", Notre Dame Journal of Formal Logic, Vol. 35, 469-70.
Review of "Russell and Analytic Philosophy," J. Symbolic Logic, Vol. 61, 1391-2.
Review of "Feasible Mathematics II," J. of the I.G.P.L. Vol. 5, 301-2.
Review of Jagdish Mehra, "The Beat of a Different Drummer," International Studies in the Philosophy of Science, Vol. 11 (1997), 311-313.

Manuscripts in preparation and submitted

``Weakly additive operators in distributive lattices", in preparation.
``The symmetry rule in propositional logic," submitted.
``The graph constructions of Hajós and Ore," submitted.

Papers presented at Meetings and Symposia

``A general theory of implication", Association for Symbolic Logic Annual Meeting, December 1971, New York.
``Free Heyting Algebras", Association for Symbolic Logic Annual Meeting, Dallas, January 1973.
``Implicational formulas in intuitionistic logic", Association of Symbolic Logic Annual Meeting, Atlanta, December 1973.
``Elementary classes in infinitary logic", Association for Symbolic Logic, Spring Meeting, Chicago, April 1975.
``A representation theory for lattices", Universal Algebra Conference, Oberwolfach, West Germany, August l976.
``Congruence latices and Heyting algebras", Universal Algebra Conference, Esztergom, Hungary, July 1977.
``Ockham lattices", Canadian Mathematical Society Annual Meeting, Calgary, December 1977.
``Equational classes of distributive double p-algebras", read to American Mathematical Society special

session on Varieties, Claremont, California, October 1978.
``Projective distributive p-lattices'', American Mathematical Society session on Universal Algebra, Boulder, Colorado, March 1979.
``Word problems for distributive lattice-ordered semigroups'' Charleston Conference on Universal Algebra, July 1984.
Talks to graduate forum (Philosophy) Fall 1982 and Fall 1983 on logic.
``How do we know mathematical proofs are correct?'', talk to IHPST October 31, 1985.
``The complexity of decision procedures in relevant logic'', presented at meeting of the Association for Symbolic Logic, Los Angeles, January 1989.
``Are there absolutely undecidable mathematical propositions?'' Talk to graduate forum, October 19 1989.

Invited Lectures

``What is Relevant Implication?'', International Conference on Relevance Logics, St. Louis, September 1974.
``How to put Routley and Meyer into a Comer'', University of Waterloo Conference on the Foundations of Logic, April 1982.
``Undecidability of Relevant Implication'', Wollongong, Australia, July 1982.
``Does Many-valued Logic make Sense?'', University of Melbourne, July 1982.
``The Undecidability of Entailment and Relevant implication'', University of Western Ontario, November 1982, also University of Buffalo, February, 1983, University of Alberta, February 1984.
``Russell's Zig-Zag Path to the Ramified Theory of Types'', Conference on Russell's Early Philosophy, University of Toronto, Summer 1984.
``Intensional Languages via Nominalization'', Society for Exact Philosophy, Halifax 1980.
``The Algebra of Entailment'', Asilomar conference on algebra and logic, July 1987.
``The Geometry of Entailment'', Indiana University, 12 November 1987.
``Complexity of Proofs in Propositional Logic'', Indiana University, 13 November 1987.
``Functional Interpretations of Feasible Arithmetic'', U. Pennsylvania, Mathematics Department, 7 December, 1987
``Functional Interpretations of Fesibly Constructive Arithmetic'', Conference on Feasible Mathematics, Cornell, June, 1989.
``Translation Problems in Modal Logic'', Society for Exact Philosophy, Edmonton, Alberta, August 1989.
``The complexity of propositional proof systems'', Berkeley conference on Logic from Computer Science, MSRI, November 1989.
``Failure of interpolation in relevant logics'', Kleene '90, Chaika, Bulgaria, July 1990.
``Complexity of proofs in classical propositional calculus'', Department of Mathematics, University of Waterloo, October 30 1990.
``Recent results in propositional complexity'', Dept. of Mathematics, McMaster University March 25 1991.
``Why are some tautologies harder to prove than others?'' Society for Exact Philosophy, Victoria B.C. May 1991.

``Russellian Propositions'' (co-author Judy Pelham), Conference on Logic, Methodology and Philosophy of Science, Uppsala, Sweden, August 1991.
``Russellian Propositions'' (co-author Judy Pelham), U. of Buffalo Logic Colloquium October 1991.
"Propositional Complexity", Talk to Mathematics Seminar, U. of Buffalo, October 1991.
"Complexity of decision procedures in relevant logic", McGill University Mathematics Seminar November 1991.
"The complexity of the Hajos calculus", "Mathematische Logik", Mathematisches Forschungsinstitut Oberwolfach April 14 1992.
"Complexity of the Hajos construction", logic workshop, U. of Victoria, B.C. July 3 1992.
"Recent results and open problems in classical propositional logic", Association for Symbolic Logic Annual Meeting March 13 1993, Notre Dame, Indiana.
"Recent work in complexity theory", talk to Logic and Philosophy of Science group, Toronto February 1994.
"Recent results in classical propositional logic", talk to Toronto Set Theory seminar, April 6 1994.
"Complexity of propositional proofs", invited lecture, Conference on Computational Logic, Indianapolis, October 14 1994.
``Russell and the Foundations of Logic 1903-05,'' invited lecture, Annual Meeting, Association for Symbolic Logic, Irvine, California April 2 1995.
Lectures on propositional proof complexity, DIMACS Workshop on Feasible Mathematics and Proof complexity, Rutgers University, Center for Discrete Mathematics and Theoretical Computer Science, August 21-25 1995.
``The symmetry rule in propositional logic,'' conference on the satisfiability problem, Siena, Italy, April 29 - May 3 1996.
``The graph constructions of Hajós and Ore,'' talk to the University of Toronto combinatorics seminar October 12 1996.
``Complexity of Propositional Proofs,'' talk to Philosophy Department Carnegie Mellon University, April 12 1997.
``Solution of a 2000-year-old logic problem,'' Philosophy Department, University of Waterloo, November 14 1997.
``Is Gödel's Theorem a red herring?'' Talk to the University of Toronto Cognitive Science and Artificial Intelligence Students' Association, January 15 1998.

D. List of courses (in preceding 5 years)

a. Undergraduate courses taught

PHL 245: Modern Symbolic Logic

PHL 246: Probability and inductive logic PHL 345: Intermediate Logic PHL 346: Philosophy of Logic

and Mathemataics

PHL 349: Set Theory PHL 342: Minds and Machines PHL 350: Philosophy of Language PHL 231: Existence and Reality PHL 344: Metalogic

b. Graduate courses taught

PHL 2121: Modern Logic PHL 2124: Seminar in Logic (theory of entailment) PHL 2124: Seminar in Logic (algebraic logic) PHL 2087: Russell PHL 2126: Foundations of Mathematics

PHL 2122: Advanced Logic (Seminar on Russellian propositions) PHL 2192: Frege

c. Theses Supervised

Masters Student:
Giovanni Panti (Mathematics) 1990.
David McClurkin (Computer Science) 1991.
Oliver Schulte (Computer Science).
Francois Pitt (Computer Science).
Doctoral Students:
Gerald Charlwood. Thesis topic: Relevance Logic. Period of supervision: 1974-78. Thesis completed 1978. Primary supervisor
Pamela Ely. Thesis topic: Medieval Logic. Period of supervision: 1975-80. Thesis completed 1980. Secondary supervisor.
Merrie Bergmann. Thesis topic: Metaphysics. Period of supervision: 1975-77. Secondary supervisor.
Ben Russell. Thesis topic: Wittgenstein's Philosophy. Period of supervision: 1978-80. Completion date: 1980. Secondary supervisor.
Mark Vorobej. Thesis topic: Deontic Logic. Period of supervision: 1979-83. Primary supervisor.
Alejandro Garciadiego. Thesis topic: History of Foundations of Mathematics. Period of supervision: 1978-1980. Secondary supervisor.
Peter Turney. Thesis topic: Inductive Inference and Stability. Thesis completed 1988. Primary supervisor.
Arnold Silverberg. Thesis topic: Anti-Realism in Semantics and Logic. Thesis completed 1988. Primary supervisor.
Andre Vellino. Thesis topic: The complexity of automated reasoning. Thesis completed 1989. Primary supervisor.

Judith Pelham. Thesis topic: Russell on propositions and objects. Thesis completed September 1993. Primary supervisor.
Kent Peacock. Thesis topic: Space-time view of non-locality. Secondarysupervision.
Andrew Malton (Computer Science Dept). Thesis topic: Functional Interpretation of Programming Methods. Secondary supervision.
Arvind Gupta (Computer Science Dept). Thesis topic: Constructivity in Tree Minors. Secondary supervision.
Frederic Portoraro. Thesis Topic: Automated theorem proving. Primary supervisor.
Xudong Fu (Computer Science). Thesis topic: Complexity of propositional proofs. Secondary supervision. Thesis completed February 1996.
Jeffrey Denson (Philosophy). Thesis topic: Objectivity in Frege. Primary supervisor. Withdrew from doctoral program 1995.
Arnold Rosenbloom (Computer Science). Secondary supervisor.
Achille Varzi (Philosophy). Secondary supervision. Thesis topic: General semantics. Thesis completed: October 1994.
Francois Pitt (Computer Science). Secondary supervision. Thesis topic: "Bounded arithmetic."
Anthony Jenkins. "The Nature of Logical Inquiry." Primary supervision.
C.K. Poon (Computer Science). Thesis topic: "The Complexityof the st-connectivity problem." PhD completed August 1995.
David Mitchell (Computer Science). Thesis topic: "Propositional Satisfiability Testing." Secondary supervision.
Tomoyuki Yamakami (Computer Science). Thesis topic: "Structure of average case hierarchies." Secondary supervision. PhD completed February 1997.
David Hyder (Philosophy). Thesis topic: "Helmholtz and Wittgenstein." PhD completed December 1996. Secondary supervision.

Administrative Positions

1970-72 Scholarships and Awards Committee, Erindale College

1973-74, 1974-75 Search Committee

1974-75, 1977-80 Personnel Committee 1978-79 Discipline Representative, Erindale College

1980-83 UTFA Representative

1979 Tenure Committee 1980 Promotion Committee

Served on thesis committee, James Hoover, Department of Computer Science.

Wrote internal appraisal for Tom Archibald (IHPST, May 1987) Ph.D. Thesis.

Martha Lile Love Award Committee 1992.

Promotion Committee, J.R. Brown 1991.

Search Committee, 1991-2.

Co-ordinator, Logic and Philosophy of Science group 1994 - Member, Mathematical Sciences Review Panel, 1994.

Member, committee for Connaught awards, 1994-7; chair 1995-6

Member, reading committee, P. Apostoli tenure decision, 1996-7.

Jack Zupko
Assistant Professor

Department of Philosophy
Emory University
214 Bowden Hall
Atlanta, GA 30322
(tel. 404 727-0104; 727-6577)

Education

1982 B.A. in Latin and Philosophy (Joint Honours), University of Waterloo
1984-85 Visiting research student in Philosophy, King's College, University of London
1986 M.A. in Philosophy, Cornell University
1989 Ph.D. in Philosophy, Cornell University

Doctoral Dissertation

"John Buridan's Philosophy of Mind: An Edition and Translation of Book III of his `Questions on Aristotle's De anima' (Third Redaction), with Commentary and Critical and Interpretative Essays", under the direction of Norman Kretzmann (UMI #9001313)

Academic Positions Held

1989-93 Assistant Professor of Philosophy at San Diego State University
1993-95 Associate Professor of Philosophy (with tenure) at SDSU
1994-95 Visiting Professor of Franciscan Studies, The Franciscan Institute, St. Bonaventure University
1995- Assistant Professor of Philosophy, Emory University

Areas of Specialization: medieval philosophy; metaphysics and epistemology; philosophy of religion

Areas of Competence: history of philosophy; ethics; logic; philosophy of mind

Selected Publications

Articles

"The Parisian School of Science in the Fourteenth Century" in *Contemporary Philosophy: A New Survey*, ed. G. Fllistad, Vol. 6/1: *Philosophy and Science in the Middle Ages* (Boston: Kluwer, 1990): 495-509

"How Are Souls Related to Bodies?: A Study of John Buridan," *The Review of Metaphysics* 46.3 (March 1993): 575-601

"Buridan and Skepticism," *Journal of the History of Philosophy* 31.2 (April 1993): 191-221

"Nominalism Meets Indivisibilism," *Medieval Philosophy and Theology* III (1993 Annual): 158-185

"How It Played in the *rue de Fouarre*: The Reception of Adam Wodeham's Theory of the *Complexe Significabile* in the Arts Faculty at Paris in the Mid-Fourteenth Century," *Franciscan Studies* 54 (1994-97): 211-225

"Bonaventure"; "Buridan, John"; "Nicholas of Autrecourt"; and "William of Auxerre," in *The Cambridge Dictionary of Philosophy*, ed. Robert Audi (New York: Cambridge University Press, 1995): 81-82; 93-94; 531; 854 (respectively)

"Freedom of Choice in Buridan's Moral Psychology," *Mediaeval Studies* 57 (1995 Annual): 75-99

"What Is the Science of the Soul?: A Case Study in the Evolution of Late Medieval Natural Philosophy," *Synthese* 110.2 (February 1997): 297-334

Book Reviews

E. J. Ashworth, *Studies in Post-Medieval Semantics* (London: Variorum, 1985), in *Eidos* 5.1 (June 1986): 97-105

Alexander Broadie, *Notion and Object: Aspects of Late Medieval Epistemology* (Oxford: Clarendon Press, 1989), in *The Philosophical Review* 101.3 (July 1992): 641-644

Adam de Wodeham, Lectura secunda in librum primum sententiarum (3 vols.), 1: Prologus et distinctio prima; 2: Distinctiones II-VII; 3: Distinctiones VIII-XXVI, ed. Rega Wood and Gedeon GÆl, O.F.M., Franciscan Institute Publications (St. Bonaventure, NY: St. Bonaventure University, 1990), in *Speculum* 68.1 (January 1993): 95-97

M. J. F. M. Hoenen, *Marsilius of Inghen: Divine Knowledge in Late Medieval Thought* (Leiden: E. J. Brill, 1993), in *Journal of the History of Philosophy* 32.2 (April 1994): 301-303

John Buridan's Tractatus de infinito: Quaestiones super libros Physicorum secundum ultimam lecturam, liber III, quaestiones 14-19, ed. J. M. M. H. Thijssen (Nijmegen: Ingenium, 1991), in *Speculum* 69.2 (April 1994): 438-439

Risto Saarinen, *Weakness of the Will in Medieval Thought: From Augustine to Buridan* (Leiden: E. J. Brill, 1994), in *The Review of Metaphysics* 49.2 (December 1995): 434-435

Rolf Schünberger, *Relation als Vergleich: Die Relationstheorie des Johannes Buridan im Kontext seines Denkens und der Scholastik* (Leiden: E. J. Brill, 1994), in *The Thomist* 60.3 (July 1996): 497-502

Forthcoming

"John Buridan," in *The Routledge Encyclopedia of Philosophy*, ed. Edward Craig (London: Routledge, to appear in 1998)

"Substance and Soul: The Late Medieval Origins of Early Modern Psychology," in *Meeting of the Minds: The Relations Between Medieval and Classical Modern European Philosophy*, ed. Stephen F. Brown (to appear in 1998)

"Duns Scotus," in the Blackwell *Companion to the Philosophers*, ed. Robert Arrington (Oxford: Basil Blackwell, to appear in 1999)

"Sacred Doctrine; Secular Practice: The Changing Role of Theology in Parisian Natural Philosophy, 1325-1400," in *Miscellanea Mediaevalia 26: What is Philosophy in the Middle Ages? Proceedings of the Tenth International Congress of Medieval Philosophy (SIEPM)* (Berlin: Walter de Gruyter, to appear in 1998)

Review of *Nicolai Oresme: Exposito et Quaestiones in Aristotelis De Anima*, ed. Benoôt Patar (Louvain-Paris: ²ditions Peeters, 1995), in *Early Science and Medicine* (Fall 1998)

In Progress

"Whole Souls and Body Parts: John Buridan's Logic of Inherence" (research article)

"The Metaphysics of Virtue in John Buridan" (research article)

Critical Latin edition of John Buridan's "*Quaestiones in libros Aristotelis `De anima' secundum tertiam (ultimam) lecturam*" (in collaboration with Dr. Peter Sobol, Department of the History of Science, University of Wisconsin)

"John Buridan" (book-length manuscript on Buridan's philosophy)

English translation of Duns Scotus's *Quaestiones in librum Aristotelis Perihermenias* (*opus primum et secundum*); *Quaestiones in libros Aristotelis De sophisticis elenchis*

Selected Presentations

"John Buridan's Theory of Universal Cognition," at the 8th International Congress of Medieval Philosophy (SIEPM), University of Helsinki, 8/28/87

"Buridan's Epistemology," at the Annual Meeting of the Canadian Philosophical Association, Queen's University, Kingston, Ontario, 5/26/91

"William of Ockham, Adam Wodeham, and John Buridan on the Question of Contact Between Continuous Magnitudes," at the Annual Meeting of the Medieval Association of the Pacific, University of California, Irvine CA, 2/21/92

"Freedom of Choice in Buridan's Moral Psychology," at the 9th International Congress of Medieval Philosophy (SIEPM), University of Ottawa, Ottawa, Ontario, 8/19/92

"Intellectual Cognition Explained: The Case of Buridan and Oresme," at the Annual Meeting of the Medieval Academy of America, University of Arizona, Tucson AZ, 4/3/93

"The Metaphysics of Virtue in John Buridan," at a session sponsored by the Society for Medieval and Renaissance Philosophy at the American Philosophical Association Annual Meeting (Eastern Division), Atlanta GA, 12/30/93

"Whole Souls and Body Parts: Buridan's Logic of Inherence," at a colloquium on medieval philosophy at the American Philosophical Association Annual Meeting (Pacific Division), Los Angeles CA, 3/31/94

"What Is the Science of the Soul?: A Case Study in the Evolution of Late Medieval Natural Philosophy," to the Philosophy Department at the University of Toronto, Canada, 1/23/95

"Choosing With Reasons: The Voluntarist/Intellectualist Controversy in Buridan's Philosophy of Action," to the Philosophy Department at Emory University, Atlanta GA, 2/02/95

"The Methodology of Representation: Abelard's Parisian Legacy," at an interdisciplinary conference, `Representation and Interpretation in the 12th Century,' Canisius College, Buffalo NY, 4/21/95

"Augustinianism in Buridan's Theory of the Will," at the 30th International Congress of Medieval Studies, Western Michigan University, 5/06/95

"Substance and Soul in Late Medieval Natural Philosophy," at a conference, "Meeting of the Minds: The Relations between Medieval and Classical Modern European Philosophy," co-sponsored by The Boston College Institute for Medieval Philosophy and Theology and the *Societe Internationale pour l'etude de la Philosophie Medievale* (SIEPM), Boston College, Boston, MA, 6/15/96

"On What Is Truly Trivial," to the Emory University Medieval Studies Roundtable Discussion Group, 3/18/97

"Sacred Doctrine; Secular Practice: The Changing Role of Theology in Parisian Natural Philosophy, 1325-1400," at the 10th International Congress of Medieval Philosophy (SIEPM), Erfurt, Germany, 8/25/97

Grants, Awards, and Academic Distinctions

1989	Research Grant, College of Arts and Letters,
San	Diego State University
1990	Research, Scholarship, and Creative Activity
Award,	SDSU

1991 Research Grant, College of Arts and Letters, SDSU
1994 Named Honorary Member, National Residence Halls Honor Society
1994 Outstanding Teacher Award, San Diego State University

1994-95 NEH Fellowship for College Teachers and Independent Scholars: "The Philosophy of John Buridan"
1995-96 Massee-Martin/NEH Teaching Observation Grant, Emory University

1996 Emory College Faculty Development Award
1997 Arthur M. Blank/NEH Teaching Observation Award, Emory University
1997 Emory College Faculty Development Award

Professional

Executive Board, Society for Medieval and Renaissance Philosophy (1997-present)

Memberships
American Philosophical Association (Newark, DE)
Aristotelian Society (London, England)
Georgia Philosophical Society (Atlanta, GA)
Medieval Association of the Pacific (San Bernardino, CA)
Societe Internationale pour l'etude de la Philosophie Medievale (Louvain, Belgium)
Society for Medieval and Renaissance Philosophy (Fairfield, CT)

Curriculum Vitae
Jill Adrian Kraye

Personal

b. 27 August 1947 Chicago, Illinois, U.S.A.
American citizen with permanent residency in U.K.
Married

Education

B.A. in History (Departmental Honors with Great Distinction and University Honors with Great Distinction): University of California at Berkeley, 1969

M.A. in History: Columbia University, 1970

M.Phil in History: Columbia University, 1973

Employment

Assistant Librarian (Academic) at the Warburg Institute: 1974-1986

Lecturer in the History of Philosophy at the Warburg Institute: 1987-1996

Senior Lecturer in the History of Philosophy at the Warburg Institute: 1996-1998

Reader in the History of Renaissance Philosophy: 1999-

Publications

'Francesco Filelfo's Lost Letter *De ideis*', *Journal of the Warburg and Courtauld Institutes*, 42 (1979), pp. 236-49

'Francesco Filelfo on Emotions, Virtues and Vices: A Re-examination of his Sources', *Bibliothèque d'humanisme et Renaissance*, 43 (1981), pp. 129-40

'Cicero, Stoicism and Textual Criticism: Poliziano on κατόρθωμα', *Rinascimento*, 23 (1983), pp. 79-110

‘The Pseudo-Aristotelian *Theology* in Sixteenth- and Seventeenth-Century Europe’ in *Pseudo-Aristotle in the Middle Ages: The ‘Theology’ and Other Texts*, ed. J. Kraye, W.F. Ryan, and C.B. Schmitt, London, 1986, pp. 265-86

‘Moral Philosophy’ in *The Cambridge History of Renaissance Philosophy*, gen. ed. C.B. Schmitt; ed. Q. Skinner and E. Kessler; assoc. ed. J. Kraye, Cambridge, 1988, pp. 303-86

‘Daniel Heinsius and the Author of *De mundo*’, in *The Uses of Greek and Latin: Historical Essays*, ed. A. C. Dionisotti, A. T. Grafton and J. Kraye, London, 1988, pp. 171-97

‘Aristotle’s God and the Authenticity of *De mundo*: An Early Modern Controversy’, *Journal of the History of Philosophy*, 28 (1990), pp. 339-58

‘Erasmus and the Canonization of Aristotle: The Letter to John More’, in *England and the Continental Renaissance: Essays in Honour of J. B. Trapp*, ed. E. Chaney and P. Mack, Woodbridge, 1990, pp. 37-52

‘Alexander of Aphrodisias, Gianfrancesco Beati and the Problem of *Metaphysics* α’, in *Renaissance Society and Culture: Essays in Honor of Eugene F. Rice, jr.*, ed. J. Monfasani and R. Musto, New York, 1991, pp. 137-60

‘History of Western Ethics: The Renaissance’, in *Encyclopedia of Ethics*, ed. L. C. Becker, New York and London, 1992; reprinted (with additional bibliography) in *A History of Western Ethics*, ed. L. C. Becker, New York and London, 1992, pp. 63-79

‘The Philosophy of the Italian Renaissance’ in *The Routledge History of Philosophy*, vol. IV, ed. G. H. R. Parkinson, London and New York, 1993, pp. 16-69

‘The Transformation of Platonic Love in the Italian Renaissance’ in *Platonism and the English Imagination*, ed. A. Baldwin and S. Hutton, Cambridge, 1994, pp. 76-85; reprinted in *The Renaissance in Europe: A Reader*, ed. K. Whitlock (New Haven and London, 2000), pp. 81-7

‘Like Father, Like Son: Aristotle, Nicomachus and the *Nicomachean Ethics*’, in *Aristotelica et Lulliana magistro doctissimo Charles H. Lohr septuagesimum annum feliciter agenti dedicata*, ed. R. Imbach, F. Dominguez, T. Pindl-Büchel, P. Walter, Turnhout, 1995, pp. 155-80

‘Renaissance Commentaries on the *Nicomachean Ethics*’, in *The Vocabulary of Teaching and Research between Middle Ages and Renaissance*, Proceedings of the Colloquium: London, Warburg Institute, 11-12 March 1994, ed. O. Weijers, CIVICIMA: Études sur le vocabulaire du moyen âge, VIII, Turnhout, 1995, pp. 96-117

‘The Printing History of Aristotle in the Fifteenth Century: A Bibliographical Approach to Renaissance

Philosophy', *Renaissance Studies*, 9 (1995), pp. 189-211

Entries on G. Bruno, M. Ficino, G. Pico della Mirandola, P. Pomponazzi and L. Valla in *The Oxford Companion to Philosophy*, ed. T. Honderich (Oxford, 1995)

'Philologists and Philosophers', in *The Cambridge Companion to Renaissance Humanism*, ed. J. Kraye, Cambridge, 1996, pp. 142-60

'Lorenzo and the Philosophers', in *Lorenzo the Magnificent: Culture and Politics in Medicean Florence*, ed. M. Mallett and N. Mann (Warburg Institute Colloquia, 3), London, 1996, pp. 151-66

Seven entries on humanists and philosophers (Angelo Decembrio, Bartolomeo Fazio, Marsilio Ficino, Cristoforo Landino, Justus Lipsius, Agostino Nifo and Angelo Poliziano) in *The Dictionary of Art*, ed. J. S. Turner, 34 vols (London and New York, 1996); reprinted in *Encyclopedia of Italian Renaissance & Mannerist Art*, ed. J. S. Turner (London and New York, 2000).

'Melanchthons ethische Kommentare und Lehrbücher', in *Melanchthon und das Lehrbuch des 16. Jahrhunderts*, ed. J. Leonhardt (Rostock, 1997), pp. 195-214

Five translations in *Cambridge Translations of Renaissance Philosophical Texts*, ed. J. Kraye, 2 vols (Cambridge, 1997), vol. I: *Moral Philosophy*, pp. 47-87, 192-99

'Conceptions of Moral Philosophy' in *The Cambridge History of Seventeenth-Century Philosophy*, ed. D. Garber and M. Ayers, 2 vols (Cambridge, 1998), II, pp. 1279-1316

Philosophy: Ancient, Medieval and Modern, Units 26-8 of *Incunabula, The Printing Revolution in Europe 1455-1500* (Research Publications Inc., 1998)

'The Contribution of Renaissance Humanists to the Neostoic Revival', *Wątki neostoickie w literaturze polskiego renesansu i baroku*. (Papers of the conference *Neostoicyzm w literaturze i kulturze staropolskiej*, Szczecin, 20-22 October 1997), ed. Piotr Urbanski (Szczecin, 1999), pp. 15-41. (Published in Polish translation as 'Wkład renesansowych humanistów w odrodzenie neostoicyzmu', *Odrodzenie i Reformacja w Polsce*, 42, 1998, pp. 5-23)

'Stoicism and Epicureanism: Philosophical Revival and Literary Repercussions', *The Cambridge History of Literary Criticism*, vol. III, ed. G. P. Norton (Cambridge, 1999), pp. 458-65.

'Pietro Pomponazzi (1462-1525): Weltlicher Aristotelismus in der Renaissance', in *Philosophen der Renaissance: Eine Einführung*, ed. P. R. Blum (Darmstadt, 1999), pp. 87-103.

'"Ethnicorum omnium sanctissimus": Marcus Aurelius and His *Meditations* from Xylander to Diderot',

in *Humanism and Early Modern Philosophy*, ed. J. Kraye and M. Stone, London Studies in the History of Philosophy, 1 (London, 2000), pp. 107-34

‘Classical Scholarship’, ‘Moral Philosophy’ and ‘Stoicism’, in *The Encyclopedia of the Renaissance*, ed. P. F. Grendler, 6 vols (New York, 1999)

‘Sir Richard Barckley’, in *Dictionary of Seventeenth-Century British Philosophers*, ed. Andrew Pyle 2 vols (Bristol, 2000)

‘The Immortality of the Soul in the Renaissance: Between Natural Philosophy and Theology’, *Signatures*, 30 November 2000, pp. 51-68

‘In Praise of Reason: From Humanism to Descartes’, in *The Discovery of Happiness*, ed. S. McCready (London: MQ Publications, 2001), pp. 136-51

‘L’interprétation platonicienne de l’*Enchiridion* d’Épictète proposée par Politien: Philologie et philosophie dans la Florence du XVème siècle, à la fin des années 70’, in *Penser entre les lignes: Philologie et philosophie au Quattrocento*, ed. F. Mariani Zini (Villeneuve d’Ascq: Presses Universitaires du Septentrion, 2001), pp. 161-77

‘Lorenzo Valla and Changing Perceptions of Renaissance Humanism’, *Comparative Criticism*, 23 (2001), pp. 37-55.

‘Ficino in the Firing Line: A Renaissance Neoplatonist and his Critics’, in *Marsilio Ficino 1433-1499: His Sources, His Circle, His Legacy*, ed. M. J. B. Allen and V. Rees (Leiden: Brill, 2001), pp. 377-97

‘Humanism’, in *The Oxford Companion to the Body*, ed. C. Blakemore and S. Jennett (Oxford: Oxford University Press, 2001), p. 368

‘Neo-Stoicism’, in *Encyclopedia of Ethics*, second edition, ed. L. C. Becker and C. B. Becker, 3 vols (Routledge: London and New York, 2001), II, pp. 1228-32

‘Renaissance Philosophy: Between the Late Middle Ages and Early Modern Era’, *Internationale Zeitschrift für Philosophie* (2001), pp. 187-98

‘Thomas Gataker (1574-1654) and his Edition of the *Meditations* of Marcus Aurelius’, in *Aspects du néo-stoicisme en Europe aux XVI*[e] *et XVII*[e] *siècles*, ed. Pierre Maréchaux and Michel Simonin (Centre d’Études Supérieures de la Renaissance, Tours, forthcoming)

‘The Role of Medieval Philosophy in Renaissance Thought: The Evidence of Early Printed Books’, in *Actes du F.I.D.E.M. Congrès de Barcelone 1999*, ed. J. Hamesse (Brepols, forthcoming)

'Eclectic Aristotelianism in the Moral Philosophy of Francesco Piccolomini', in *The Presence of Paduan Aristotelianism in Early Modern Philosophy*, ed. E. Kessler *et al*. (Antenore, forthcoming)

'British Philosophy before Locke', in *A Companion to Early Modern Philosophy*, ed. S. Nadler (Blackwell, forthcoming)

21 entries for *The Oxford Companion to Italian Literature*, ed. David Robey and Peter Hainsworth (Oxford University Press, forthcoming)

Articles on Haly Heron, James Martin and Thomas Palfreyman for the *New DNB* (Oxford University Press, forthcoming)

Reviews for *Journal of Roman Studies*, *Modern Language Review*, *Journal of the History of Philosophy*, *Bibliothèque d'humanisme et Renaissance*, *The Library*, *Heythrop Journal*, *American Historical Review*, *Times Literary Supplement*, *Review of Metaphysics*, *Notes and Queries*, *Isis*, *Renaissance Quarterly*, *Speculum*, *History*, *The Slavonic and East European Review*, *Studi umanistici*, *Sixteenth-Century Journal*, *Journal of Ecclesiastical History*, *Religious Studies*.

Publications in Progress

The Classical Tradition in Renaissance Philosophy (Variorum Collected Studies Series)

Editor and translator, *Renaissance Philosophy* (Penguin Books)

Editor and translator: Cristoforo Landino, *Disputationes Camaldulenses* (I Tatti Renaissance Library)

Editorial Work

Editor, jointly with W.F. Ryan and C.B. Schmitt, *Pseudo-Aristotle in the Middle Ages: The 'Theology' and Other Texts* (London, 1986)

Editor, jointly with A. C. Dionisotti and A. Grafton , *The Uses of Greek and Latin: Historical Essays*, (London, 1988)

Associate editor, *The Cambridge History of Renaissance Philosophy*, gen. ed. C.B. Schmitt; ed. Q. Skinner and E. Kessler (Cambridge, 1988)

Editor, *The Cambridge Companion to Renaissance Humanism* (Cambridge, 1996); Spanish translation, *Introducción al humanismo renacentista* (Madrid, 1998)

Editor, *Cambridge Translations of Renaissance Philosophical Texts,* vol. I: *Moral Philosophy*; vol. II: *Political Philosophy* (Cambridge, 1997)

Joint editor, with Martin Stone, *Humanism and Early Modern Philosophy*, London Studies in the History of Philosophy, 1 (London, 2000)

Joint editor, with W. F. Ryan and C. Burnett, *Warburg Institute Surveys and Texts*, 1986-

Advisory Board: *Journal of the Warburg and Courtauld Institutes*, 1985-95

Associate Editor: *Journal of the Warburg and Courtauld Institutes*, 1995-6

Joint Editor: *Journal of the Warburg and Courtauld Institutes*, 1997-

Editorial consultant for philosophy: Bodleian Incunable Project, ed. K. Jensen: 1991-

Editorial consultant: *British Journal for the History of Philosophy*, 1991-

Editorial consultant: *Bruniana & Campanelliana*, 1995-

Editorial consultant: *Letteratura italiana antica*, 1999-

Advisory committee: I Tatti Renaissance Library, 1997-

Colloquia and Seminars Organized

Seminar on Medieval and Renaissance Intellectual History at the Warburg Institute (1982-4), jointly with C. B. Schmitt

Colloquia on Greek and Latin Scholarship at the Warburg Institute (1986), jointly with A. C. Dionisotti and Anthony Grafton

History of Seventeenth-Century Philosophy at the Warburg Institute (1988), jointly with A. Clericuzio

History of Philosophy Seminar at the Warburg Institute (1991-5), jointly with R. Sorabji and C. Burnett

History of the Problems of Philosophy Seminar at the University of London (1995-7), jointly with M. Stone and J. Wolff.

Humanism and Early Modern Philosophy at the Warburg Institute (1997), jointly with M. Stone

The New Historiography of Early Modern Philosophy (2000-2002), jointly with Tom Sorell and G. J. A. Rogers.

Thomas Robert Flynn
Curriculum Vitae
(Fall 2000)

Mailing Address

1278 Oakdale Road, NE
Atlanta, Georgia 30307
Telephone: 404/378-7321

Position

Samuel Candler Dobbs Professor of Philosophy
Department of Philosophy
Emory University
Atlanta, Georgia 30322
Telephone: 404/727-4316
Email: tflynn@emory.edu

Areas of Research and Teaching Specialization

20th Century Continental Philosophy
Political and Social Philosophy
Ethics (Theory of Responsibility)

Areas of Competence

History of Philosophy (Modern period and 19th Century)
Theory of Knowledge
Aesthetics

Higher Education and Degress

B.A., summa cum laude (Philosophy and History), Carroll College, Helena, Montana, 1958.

S.T.L., (*Licentiate* in Theology) summa cum laude, Gregorian University, Rome, Italy, 1962.

Ph.D., awarded "with distinction" (Philosophy), Columbia University, New York, New York, 1970.

Ph.D. Dissertation: Jean-Paul Sartre and the Problem of Collective Responsibility, written under the direction of Professors Robert D. Cumming (major professor), Charles Frankel, and David Sidorsky.

Teaching Experience

1962-66, Carroll College, Helena, Montana, Instructor (Philosophy and French).

1968-70, Columbia University, New York, New York, Preceptor (Philosophy and Contemporary Civilization).

1970-71, Carroll College, Montana, Assistant Professor (Philosophy and Humanities).

1971-75, Catholic University of America, Washington, D.C., Assistant Professor (Philosophy).

1975-76, St. Mary's University, Baltimore, Maryland, Scholar in Residence.

1976-78, Carroll College, Helena, Montana, Assistant Professor (Philosophy).

1978-82, Emory University, Atlanta, Georgia, Assistant Professor (Philosophy).

1982-86, Emory University, Atlanta, Georgia, Associate Professor (Philosophy).

1987-88, Emory University, Atlanta, Georgia, Professor (Philosophy).

1988- , Emory University, Atlanta, Georgia, Samuel Candler Dobbs Professor of Philosophy.

1995 (Spr), Villanova University, Visiting Professor

Awards and Distinctions

Listed in the Millennial Edition of Who's Who in the World, 2000.

Selected by Nominating Committee of the APA as one of three candidates for Vice President, President-elect of the American Philosophical Association in 1999-2000.

Outstanding Faculty Member in Service Award for 1999, Omicron Delta Kappa (ODK).

Selected to have class videotaped and interviewed as part of the Paradigm in Teaching Project, Emory College, 1998-99.

Seminar Leader of three week Summer Seminar for Liberal Arts College faculty on the topic "The Foucault Effect" at the National Humanities Center, Research Triangle Park, N.C., June, 1999.

D.V.S. Award of Recognition for Service to the Emory Community, Fall, 1998.

Emory University research Committee, Maximum Regular Grant for release time, 1998-99.

Fellowship from the National Endowment for the Humanities awarded via the Institute for Advanced Study for academic year 1998-99.

Member, Institute for Advanced Study, Princeton, 1998-99.

First annual George P. Cuttino Award for Student Mentoring, 1997.

Andrew W. Mellon Fellow at the National Humanities Center, North Carolina, 1991-92.

Emory University Research Committee, Maximum Regular Grant, 1991 (awarded but declined).

Emory Williams Teaching Award for Excellence in Undergraduate Teaching (awarded an unprecedented second time), 1990.

Alumni Academic Achievement Award, Carroll College, Helena, MT, October 1987.

Emory University Teacher/Scholar of the Year, September 1986.

Senior Council of Emory College Teaching Award (awarded an unprecedented second time), 1985.

Lilly Foundation Fellow, 1984-85.

A.C.L.S. Senior Research Fellow, Paris, France, 1983-84.

Senior Council of Emory College Teaching Award, 1983.

Emory Williams Teaching Award for Excellence in Undergraduate Teaching, 1981.

Emory University, Summer Research Grant, 1981.

N. E. H., Summer Research Grant, 1980.

Preceptorship in Philosophy (Columbia), 1968-70.

Danforth Teacher Grant for excellence in undergraduate teaching, 1966-67; renewed 1967-68.

Publications

Books

Sartre and Marxist Existentialism. The Test Case of Collective Responsibility. University of Chicago Press, 1984; paperback edition, 1986.

Dialectic and Narrative. Ed. and intros. with Dalia Judovitz. State University of New York Press, 1993.

Sartre, Foucault, and Historical Reason. Two vols. Vol 1 Toward an Existentialist Theory of History. University of Chicago Press, 1997. Vol. Two, A Poststructuralist Mapping of History (in progress).

The Ethics of History, Ed. and intros. with David Carr and Rudolf Makkreel, submitted to major university press.

Articles

77. "Pyramids and Prisms: Reading Foucault in 3-D," (University of South Carolina, Program in Comparative Literature Web Site) Web-publication, 20 pp.

76. Foreword to Kevin Boileau, Genuine Reciprocity and Group Authenticity: Foucault's Developments or Sartre's Social Ontology (Lanham, MD: University Press of America, 2000)5-8.

75. "Postmodernism and the Catholic Tradition: A Response to Kenneth Schmitz," American Catholic Philosophical Quarterly, 73:2 (Spring 1999), 261-266.

74. "Toward an Existentialist Philosophy of History: Responses to Matuštík and McBride," (Symposium on the first volume of my Sartre, Foucault and Historical Reason), in Human Studies: A Journal for Philosophy and the Social Sciences, forthcoming.

73. New Articles on "Bad Faith" and "The Absurd" for The Encyclopedia of Ethics 2nd ed. (New York: Garland), projected publication date, 1999. Three entries from the 1st ed., "Authenticity," "de Beauvoir, Simone" and "Sartre, Jean-Paul," are revised for the second edition.

72. The Philosopher-Historian as Cartographer," Research in Phenomenology, vol. 29 (1999), 31-50 special issue on "The Claims of History," ed. James Risser (invited).

71. "Existentialism and Beyond: Camus, de Beauvoir, Merleau-Ponty and Sartre," in Richard Popkin, ed. The Columbia History of Western Philosophy (New York: Columbia University Press, 1999), 698-705.

70. A Return to Foucault's Partially Desacralized Spaces With Jean-Luc Nancy," in Hugh J. Silverman, ed., Reading Jean-Luc Nancy, forthcoming.

69. Objectivity: The View from the Left Bank," Proceedings of the 20th World Congress of Philosophy (Bowling Green: Philosophical Documentation Center), forthcoming.

68. "Athens and Jerusalem; Paris and Rome," in Faith and the Intellectual Life, ed. Curtis Hancock and Brendan Sweetman, volume submitted for review

67. "Committed History," in Carr, Flynn and Makkreel eds. The Ethics of History submitted for review.

66. "Introduction" (with Carr and Makkreel) to The Ethics of History, ed. David Carr, Thomas Flynn and Rudolf Makkreel, submitted for review (see "Books").

65. "Who Inherits the Historical Dialectic? Derrida and the Spirits of Marx" in Reading Specters of Marx, ed. Leonard Lawlor and Hugh Silverman, Humanities Press, forthcoming.

64. "Sartre on Violence, Foucault on Power -- A Diagnostic,"Bulletin de la Socie'te' des Philosophes de la Langue Francaise, vol.X, no. 2 (Fall 1998), 129-151.

63. "Foucault and Justice: The Problem of the Floating Transcendental," Phenomenology and Politics, ed. Lester Embree and Keith Thompson (Dordrecht: Kluwer), forthcoming.

62. "Sartre, Jean-Paul," (1,500 words) Encyclopedia of Aesthetics, 4 vols. ed. Michael Kelly (Oxford: Basil Blackwell), 1998.

61. "Event, History, Totality: Lyotard and History without Witnesses," In Reading Lyotard ed. Hugh J. Silverman, forthcoming

60. "Reconstituting Praxis: An Existentialist Theory of History, Special Sartre issue of the American Catholic Philosophical Quarterly, 60/4 (Autumn 1996), 597-618.

59. "Times Squared: The Historical Times of Sartre and Foucault." In Recent Phenomenologies of Time, ed. John Brough. Kluwer Academic Publishers, forthcoming.

58. "Alterity and Appropriation: The Successful Failure of Historical Dialectic." In Reading Kristeva's Strangers to Ourselves, ed. Gary E. Aylesworth and Hugh J. Silverman. Humanities Press, forthcoming.

57. "Continental Philosophy on Another Continent." In Kontinentalphilosophie aus Amerika: 22 Photogrammische Portraits, ed. James R. Watson, Vienna: Turia and Kant Verlag, 1998. English version

with Northwestern University Press, 1999, 54-63.

56. "Authenticity." In Dictionary of Business Ethics, ed. R. Edwards Freeman and Patricia H. Werhane. Oxford: Blackwell Publishers, 1997, 30.

55. "Phenomenology of Ethics, Sartrean" (+3000 words). In The Encyclopedia of Phenomenology, ed. Lester Embree. Dordrecht: Kluwer Publishing Co., 1997, 184-189.

54. "Sartre (1905-1980)" (+5300 words). In A Companion to Continental Philosophy, ed. Simon Critchley and William Schroeder. (Oxford: Blackwell Publishers, 1998): 256-68.

53. "Benign Nihilism: Violence and Power." For an as yet untitled volume, ed. Hugh J. Silverman and Wilhelm Wurtzer. SUNY Press, forthcoming.

52. Six brief contributions to Dictionary of Existentialism listed above: "Dialectical Reason," "Facticity," "The Look," "Bad Faith," "Nothingness," and "Negation," (Greenwood Press, 1999).

51. "Authenticity." Short essay for Dictionary of Existentialism, ed. Haim Gordon. (Westport, CT: Greenwood Press, 1999), 24-26.

50. "Sartre, Jean-Paul." Invited essay for Dictionary of Existentialism, ed. Haim Gordon. (Westport, CT: Greenwood Press, 1999), 418-25.

49. "Truth(s) in Painting: Sartre, Foucault, and Derrida. In Reading Truth in Painting, ed. Hugh J. Silverman and Wilhelm S. Wurzer. Albany, N.Y.: SUNY Press, forthcoming.

48. "The Essence of Man," invited essay in The World & I, monthly book review and essay magazine of The Washington Times (February, 1996): 253-59.

47. "History and Histories: Weiss and the Problematization of the Historical." Invited essay in The Philosophy of Paul Weiss, Schilpp Library of Living Philosophers, ed. Louis Hahn. Chicago: Open Court, 1995, 183-200.

46. "Inauthentic and Authentic Love in Sartrean Existentialism." In The Nature and Pursuit of Love, ed. David Goicoechea. Amherst, NY: Prometheus Books, 1995, 208-220.

45. "The Future Perfect and the Perfect Future: History Has Its Reasons....," (Presidential Address to the American Catholic Philosophical Association, 1994), American Catholic Philosophical Quarterly, Supplement (Proceedings, 1994), 1-15.

44. Three entries: "Existentialism," "Foucault," and "Sartre," for The Encyclopedia of Time, ed. Sam

Macey (Garland Publishing), announced for February, 1994.

43. "Foucault and the Mapping of History." In The Cambridge Companion to Foucault, ed. Gary Gutting. London: Cambridge University Press, 1994, 28-46.

42. "Existential Philosophy II: Jean-Paul Sartre." Continental Philosophy in the Twentieth Century, ed. Richard Kearney. Routledge History of Philosophy 8. London: Routledge, 1994, 74-104.

41. "Sartre and the Paradox of Committed History." In Modern Concepts of Existentialism: Essays on Sartrean Problems in Philosophy, Political Theory and Aesthetics, ed. Peter L. Eisenhardt, et al. Jyväskylä Studies in Education, Psychology and Social Research 102. Jyväskylä, Finland: University of Jyväskylä Press, 1993, 97-107.

40. "Partially Desacralized Spaces: The Religious Availability of Foucault's Thought," Faith and Philosophy 10: 4 (Oct. 1993), 471-85.

39. "Truth is a Thing of This World." Review essay of James Bernauer's Michel Foucault's Force of Flight: Toward an Ethics For Thought. Research in Phenomenology 23 (1993), 193-201.

38. "Foucault and the Eclipse of Vision." In Modernity and the Hegemony of Vision, ed. David Michael Levin. Berkeley: Univ. of California Press, 1993, 273-286.

37. "Sartre, The Last Ten Years," review essay in Research in Phenomenology 22 (November, 1992), 210-216.

36. "The Possibility/Impossibility of a Foucauldian Ethic." In Joyful Wisdom: Sorrow and An Ethic of Joy, ed. David Goicoechea and Marko Zlomislic. St. Catharines, Ont.: Thought House, 1992, 127-43.

35. "Phenomenology and Faith: From Description to Explanation and Back." American Catholic Philosophical Quarterly 64, supplement (1990), 40-50.

34. "Sartre and the Poetics of History." In The Cambridge Companion to Sartre, ed. Christina Howells. Cambridge: Cambridge University Press, 1992, 213-260.

33. "Jean-Paul Sartre." In Encyclopedia of Ethics. New York: Garland Press, 1992, 2:1121-24.

32. "Simone de Beauvoir." In Encyclopedia of Ethics. New York: Garland Press, 1992, 1:241-42.

31. "Authenticity." In Encyclopedia of Ethics. New York: Garland Press, 1992, 1:67-69.

30. "Foucault and the Spaces of History" (invited essay for special issue on theories of history), The

Monist 74: 2 (April 1991): 165-86.

29. "Sartre and the Poetics of History." In Proceedings of the Sartre Society of Canada, ed. A. van den Hoven and W. Skakoon, Vol. 1, No. 1. Ontario: University of Windsor, 1990, 132-41.

28. "Sartre and Technological Being-in-the-World." In Lifeworld and Technology, ed. Lester Embree and Tim Casey. Washington, DC: University Press of America, 1990, 271-87.

27. "Foucault and the Politics of Postmodernity," Nous 23 (1989): 187- 98.

26. "History as Fact and as Value: The Posthumous Sartre." In Inquiries and Values, edited by Sander H. Lee. Lewiston, NY: Edwin Mellen Press, 1988, 375-90.

25. "Skizze für eine Theorie der Geschichte in der Philosophie Sartres: Die Carnets und die Cahiers." In Sartre-ein Kongress, ed. Traugott König. Reinbeck, W. Germany: Rowohlt Verlag, 1988, 201-25.

24. "Time Redeemed: Maritain's Christian Philosophy of History." In Understanding Maritain, edited by Diel Hudson and Matthew Mancini, 306-24. Macon, GA: Mercer University Press, 1988.

23. "Foucault and Historical Nominalism." In Phenomenology and Beyond: The Self and It's Language, edited by Harold A. Durfee and David F. T. Rodier. Brussels/Boston: Kluwer, 1989. pp. 134-47.

22. "Foucault as Parrhesiast: His Last Course at the College de France" in a special Foucault issue of Philosophy and Social Criticism (Vol 12:2-3). Reprinted in The Final Foucault, edited by James Bernauer, 102-18. Cambridge, MA: M.I.T. Press, 1988. Japanese translation, 1991. Reprinted in Michel Foucault: Critical Assessments Barry Smart, ed. 7 vols. (London, Routledge, 1998) 3: 302-325.

21. "Dying as Doing: Philosophical Thoughts on Death and Authenticity." In Death: Completion and Discovery, edited by Charles A. Corr and Richard A. Pacholski, 261-69. Lakewood, OH: Association for Death Education and Counseling, 1987.

20. "Foucault and the Career of the Historical Event." In At the Nexus of Philosophy and History, edited by Bernard P. Dauenhauer, 178- 200. Athens, GA: University of Georgia Press, 1987.

19. "Merleau-Ponty and the Critique of Dialectical Reason." In Hypatia, edited by William Calder, Ulrich Goldsmith and Phyllis Kenevan, 241-50. Boulder, CO: Colorado Associated University Press, 1985. Reprinted in The Debate Between Sartre and Merleau-Ponty, ed. Jon Stewert (Evanston, IL: Northwestern University Press, 1998).

18. "Truth and Subjectivation in the Later Foucault," invited symposium paper for annual meeting of the A.P.A., Eastern Division, December, 1985, published in The Journal of Philosophy 82 (October 1985): 531-40.

17. "Collective Responsibility and Obedience to the Law," invited essay on obedience to the law for special issue of the Georgia Law Review 18 (Summer 1984): 845-61.

16. "Jean-Paul Sartre." Major article in Thinkers of the 20th Century. A Biographical, Bibliographical and Critical Dictionary, edited by Elizabeth Devine, Michael Held, James Vinson and George Walsh, 497-500. Detroit: Gale Research, Macmillan, 1983.

15. "Beginnings Without End" (10-page review essay of R. D. Cummings's Starting Point: An Introduction to the Dialectic of Existence), Man and World 15 (1982): 197-205.

14. "From 'Socialisme et Liberte' to 'Pouvoir et Liberte': Sartre and Political Existentialism." In Phenomenology in a Pluralistic Context, edited by William McBride and Calvin Schrag, 26-38. Albany, New York: State University of New York Press, 1983.

13. "Existential Hermeneutics: The Progressive-Regressive Method," Eros 8 (Spring 1981): 3-24 (lead article in special Sartre issue).

12. "Another Sartrean Torso: Critique of Dialectical Reason," Social Theory and Practice 6 (Spring 1980): 91-107.

11. "Angst and Care in the Early Heidegger: The Ontic/Ontologic Aporia," International Studies in Philosophy 12 (Spring 1980): 61-76.

10. "Sartre-Flaubert and the Real/Unreal," in Existence and Dialectic: Contemporary Approaches to Jean-Paul Sartre, edited by Hugh Silverman and Frederick Elliston, 105-23. Pittsburgh: Duquesne University Press, 1980.

9. "Mediated Reciprocity and the Genius of the Third." In Sartre volume of Paul Arthur Schilpp's Library of Living Philosophers, 345-70. In the same volume Sartre responds to questions I have posed regarding my essay. La Salle, Illinois: Open Court Publishing Co., 1981.

8. "Vision, Responsibility, and Factual Belief," Journal of Chinese Philosophy 7 (1980): 27-36.

7. "L'Imagination au Pouvoir: the Evolution of Sartre's Political and Social Thought," lead essay in Sartre issue of Political Theory 7 (May 1979): 157-180.

6. "Catholic Charities and the American Experience: Suggestions for the Inner Dialogue," Social Thought 3 (Summer 1977): 31-40.

5. "An End to Authority: Epistemology and Politics in the Later Sartre," Man and World 10, no. 4

(1977): 448-65.

4. "Praxis and Vision: Elements of a Sartrean Epistemology," The Philosophical Forum 8 (Fall 1976): 21-43.

3. "The Use and Abuse of Utopias," The Modern Schoolman 43 (March 1976): 235-64.

2. "The Role of the Image in Sartre's Aesthetic," The Journal of Aesthetics and Art Criticism 33 (Summer 1975): 431-42.

1. "The Alienating and the Mediating Third in the Social Philosophy of Jean-Paul Sartre," in John K. Ryan, Ed., Studies in Philosophy and in the History of Philosophy 6 (1973): 3-38.

III. Books Reviewed

Numerous reviews in The Review of Metaphysics, Ethics, The American Catholic Philosophical Quarterly, Research in Phenomenology and elsewhere.

Review of Annette Aronowicz, Jews and Christians on Time and Eternity: Charles Péguy's Portrait of Bernard-Lazare (Stanford), The Jewish Quarterly Review XC, nos. 1-2 (July-October 1999) 173-175.

Papers Read and Conference Participation

Keynote address at Brock University, St. Catherine's,Ont., Feb.10, 2001

Address to Dept. of Philosophy, Texas A&M University, Feb. 15, 2001

Chair Session of Kierkegaard Conference, Augusta State Univ. GA, March 2, 2000

Invited address sponsored by Several Departments of Cornell University (Philosophy, Political Science, Womens' Studies) on "Government of Self and Others: Reading Foucault in 3-D," April 26, 2001

Organize and Chair Invited Session of International Association for Philosophy and Literature (IAPL) on "Origin/Birth/Provenance" in Nietzsche and Foucault"at Spelman College, Atlanta, May 3, 2000.

Chair of plenary session of annual meeting of the *Groupe d'études sartriennes*, Univ. of Paris (Sorbonne), Paris, France, June 25, 2000. Speakers from France, Belgium and Italy.

"If Everything Were a Dream,...: Recalling Baudrillard's Forgetting Foucault," International Philosophical Seminar, Alto Adige, Italy, July 7, 2000.

Comment on presentation on Sartre and Deleuze by professor from Concordia College, Montreal, at annual meeting of the Sartre Society of North America, Wilfred Laurier University, Waterloo, Ontario, September 17, 2000.

"Post-structuralist Spirituality: The Case of Michel Foucault," Aquinas Center, Emory, October 13, 2000.

"The Soul of the University," Discussion of my essay in Graduate Teaching Fellows' Seminar, Graduate Division of Religion (GDR), Oct. 18, 2000.

"Existential Psychoanalysis: Its Use and Abuse: The Case of Jean-Paul Sartre," Psychoanalytic Studies Colloquium, Emory, Oct. 18, 2000.

Commentary on presentation by Thomas Anderson, Marquette Univ., on Faith and Reason in Gabriel Marcel," Annual Meeting of the American Catholic Philosophical Association, Dallas, TX Nov. 4, 1000.

"Sartre on Violence, Foucault on Power: A Diagnostic," University of North Carolina-Asheville, Nov. 16, 2000.

"The Use and Abuse of Utopias," S. & H. Foundation Lecture, Washington, D.C., March 22, 1973.

"Sartre-Flaubert and the Real/Unreal," annual meeting of the American Philosophical Association (A.P.A.), Western Division, 1975 (blind review).

"Praxis and Vision: Elements of a Sartean Epistemology," A.P.A., Eastern Division meeting, 1977 (blind review).

"An End to Authority. Epistemology and Politics in the Later Sartre," annual meeting of the Northwest Conference on Philosophy, 1977 (invited).

"The Tension Theory of Metaphor," commentator at Georgia Philosophical Society meeting, 1978.

"Basic Themes in Existential Philosophy," invited paper at Morehouse College, 1979.

"Problems in Collective Responsibility," the Radical Caucus meeting, A.P.A., Eastern Division, 1979 (invited).

"Collective Responsibility," invited paper at University of Georgia, 1980.

"From 'Socialisme et Liberte' to 'Pouvoir et Liberte', The Case of Jean-Paul Sartre," annual meeting of the Society for Phenomenology and Existential Philosophy, Ottawa, Ontario, Canada, November 4, 1980 (invited).

"Collective Responsibility as a Problem in Social Ontology," annual meeting of the International Society for Metaphysics, King's College, University of London, London, England, July 19, 1980 (invited).

"Existentialist Hermeneutics: The Progressive-Regressive Method," The Collegium Phenomenologicum, Perugia, Italy, June 1980 (invited).

"Collective Responsibility as a Problem In Social Ontology," University of South Carolina, Columbia, South Carolina, 1981 (invited).

Commentator on the two principal speakers at Sartre Memorial, general session of the annual meeting of Society for Phenomenology and Existential Philosophy, Northwestern University, October 28-November 1, 1981.

"The Primacy of Praxis in the Later Philosophy of Sartre," University of Colorado at Boulder, March 4, 1982 (invited).

"The Primacy of Praxis in the Later Philosophy of Sartre," University of Kansas Public Colloquium Series, March 2, 1982 (invited).

"Foucault and Nominalist Historiography," read at Philosophy Colloquium, Emory University, January 24, 1985.

"Is There (Intellectual) Life After Tenure?" talk given to general session of Lilly Foundation Fellows and Directors, University of Indiana, March 31, 1985.

"Foucault and the Career of the Historical Event," read at symposium on the philosophy and history, University of Georgia, May 2, 1985 (invited).

"Learning the Language of Humanity." Convocation address, Emory College and Graduate School, September 5, 1985.

"Time Redeemed: Maritain's Christian Philosophy of History," American Maritain Association, Annual Meeting, Atlanta, GA, October 11, 1985.

"Foucault as Critical Theorist," Association for Humanist Sociology, November 9, 1985.

"Truth and Subjectivation in the Later Foucault." Invited Symposium paper, American Philosophical Association, Washington, D.C., December 28, 1985.

"Foucault and the Paradox of Creativity," address to Faculty Seminar of the Department of Modern Languages and Classics, February 12, 1986.

"Life, Death and Authenticity: Some Existentialist Reflections," the annual Philosophy Lecture, Siena College, Loudonville, NY, March 17, 1986.

Respondent to Richard Rorty, "The Contingency of the Self," Emory University, March 28, 1986.

Commentator on "Sartre, Maritain and the Contradiction of Conscious Being," Maritain Association in conjunction with American Catholic Philosophical Association, Baltimore, MD, April 5, 1986.

"Dying as Doing: Philosophical Reflections on Death and Authenticity," Keynote address given to plenary session of annual meeting of the Forum for Death Education and Counselling, Atlanta, GA, April 19, 1986.

Commentator on "Foucault and Feminism," annual meeting of American Philosophical Association (Western Division), St. Louis, MO, May 2, 1986.

"Learning the Language of Humanity," address to Emory Alumni Meeting, Harvard Faculty Club, Cambridge, MA, September 18, 1986.

"(Postmodern) Faith and (Enlightenment) Reason," faculty seminar on faith and reason, Candler School of Theology, Emory University, February 19, 1987.

"Posthumous Sartre and the Philosophy of History," keynote address to Southeast Conference for Graduate Students in Philosophy, The University of Georgia, Athens, GA, April 24, 1987.

"Can the University be Committed to Values?" delivered to faculty symposium on values in higher education, The University of Georgia, Athens, GA, May 12, 1987.

"Life, Death and Authenticity: Some Existentialist Reflections," annual Toulouse Lecture, Seattle University, Seattle, WA, May 19, 1987.

"The Posthumous Sartre and the Philosophy of History," address to Northwest Society for Phenomenology, Existentialism and Hermeneutics, Seattle, WA, May 20, 1987.

"(Postmodern) Faith and (Enlightenment) Reason," to the faculty seminar in philosophy and theology, Seattle University, Seattle, WA, May 21, 1987.

"Learning the Language of Humanity: Humanizing Legal Education," address to Thomas More Society of Lawyers, Seattle, WA, May 22, 1987.

"Sketch for a Theory of History in the Philosophy of Sartre," address to International Sartre Congress, University of Frankfurt, Frankfurt, Germany, July 11, 1987.

Respondent to Algis Mickunas, "The Emergence and Structure of Political Technocracy," Duquesne University, Pittsburgh, PA, October 25, 1987.

"Sartre and the Philosophy of History: Elements of an Existentialist Theory of History in Sartre's Posthumous Works," invited address, Duquesne University, Pittsburgh, PA, October 26, 1987.

Panel discussant of the Galileo case for series on Science and Society: Conflict and Dialogue, Georgia College, Milledgeville, GA, November 5, 1987.

"Posthumous Sartre and the Philosophy of History," address to the Georgia Philosophical Society, Atlanta, GA, November 14, 1987.

"Sartre and the Philosophy of History," invited address, Creighton University, Omaha, NE, March 17, 1988.

"Foucault and Historical Nominalism," invited address to conference on "Text and Context," University of Georgia, May 12, 1988.

"History as Fact and as Value: The Posthumous Sartre," invited address to the inaugural session of the International Society for Value Inquiry, Arundel and Brighton, England, August 24, 1988.

"Foucault and the Politics of Postmodernity," symposium address at the World Congress of Philosophy, Brighton, England, August 25, 1988.

"Sartre and the Poetics of History," Emory University, September 8, 1988.

"The Violence of Love." Response to John Caputo at Conference on Natural Theology, University of South Carolina, Columbia, October 22, 1988.

Commentator on Sharon Welsh's presentation at conference, "For the Trumpet Shall Sound," Aquinas Center, Emory University, Atlanta, GA, October 27, 1988.

"Foucault and the Politics of Postmodernity," invited paper for symposium at annual meeting of the American Philosophical Association (Central Division), Chicago, IL, April 29, 1989.

"Foucault and the Politics of Postmodernity," address to faculty and graduate students of the Heinrich Heine University, Dusseldorf, W. Germany, June 21, 1989.

Draft Lecture in Philosophy delivered at Ripon College, Ripon, WI, on the topic "Existentialism Alive: Five Philosophical Themes," September 28, 1989.

"Authenticity and the Tradition in Continental Ethics," invited address to annual meeting of the Society for Phenomenology and Existential Philosophy, Duquesne University, Pittsburgh, October 13, 1989.

Commentator on David Detmer's paper at meeting of Sartre Circle in conjunction with the APA, Atlanta, December 29, 1989.

"Phenomenology and Faith: From Description to Explanation and Back," invited address to plenary session of American Catholic Philosophical Association, Toronto, March 31, 1990.

Organized and led panel discussion "Technologies of Truth: Foucault and His Critics" at annual meeting of the International Association for Philosophy and Literature, University of California/Irvine, April 27, 1990.

"Foucault and the Spaces of History," Department of Philosophy Colloquium, Emory University, Atlanta, September 27, 1990.

"Foucault and the Philosophy of History," address to faculty seminar, University of Arkansas, Little Rock, November 3, 1990.

"Phenomenology and Religion: From Description to Explanation and Back," Rockefeller lecture delivered at the University of Arkansas, Little Rock, November 4, 1990.

"Phenomenology and Religious Belief," Rockefeller Lecture delivered at the University of Arkansas/Little Rock, November 14, 1990.

"Current Research on Foucault," Faculty Seminar presentation, University of Arkansas/Little Rock, November 14, 1990.

"Foucault and the Spaces of History," Canisius College, Buffalo, NY, February 13, 1991.

"Inauthentic and Authentic Love in the Thought of Jean-Paul Sartre," Brock University, St. Catharines, Ontario, Canada, February 16, 1991.

Organized and led panel discussion, "Origins: The Senses of a Beginning," annual meeting of the International Association for Philosophy and Literature, University of Montreal, Canada, May 17, 1991.

"Truth(s) in Painting: Derrida, Sartre and Foucault," International Philosophical Seminar, Castelrotto, Italy, June 29, 1991.

"Sartre, The Last Ten Years," Sartre Society, University of Dayton, Ohio, September 21, 1991.

"Reading The Political Theory of Jean-Paul Sartre by William McBride," University of Dayton, OH, September 29, 1991.

Chaired a session, "Foucault and Ethics," at Meeting of the Society for Phenomenology and Existential Philosophy, Memphis, TN, October 18, 1991.

"Foucault and the Possibility/Impossibility of an Ethics," Brock University, St. Catherine's, Ontario, Canada, November 9, 1991.

"Making Sens of History with Jean-Paul Sartre," annual Phi Kappa Phi Lecture, University of North Carolina/ Wilmington, November 14, 1991.

"Foucault and the Eclipse of Vision," Duke University, November 21, 1991.

"Sartre, Foucault and Reason in History," Duke University (Institute for Learning in Retirement), January 30, 1992.

"The Timetable and the Map: Existentialist and Postmodern Theories of History," U.N.C.- Chapel Hill, February 13, 1992.

"Foucault and the Eclipse of Vision" (faculty seminar, Dept. of Philosophy/Religion), George Mason Univ., Fairfax, VA, April 7, 1992.

"Benign Nihilism: Power and Violence," public lecture, George Mason University, April 7, 1992.

"Foucault and the Eclipse of Vision," University of Richmond, VA, April 7, 1992.

"Benign Nihilism," International Philosophical Seminar, Castelrotto, Italy, June 22, 1992.

"Benign Nihilism," Emory University, October 1, 1992.

"Foucault, Power and Critique," S.P.E.P., Boston, October 10, 1992.

"Recent French Philosophy: The Foucault Effect," Pennsylvania State University, October 24, 1992.

"Benign Nihilism: Violence (Sartre) and Power (Foucault)," Denison University, November 13, 1992.

"Sartre and the Problem of Committed History," Sartre Circle, in conjunction with the APA Eastern Division, Washington, DC, December 29, 1992.

Chaired session on Martha Nussbaum's Love's Body, annual meeting of American Philosophical

Association, Washington, DC, December 29, 1992.

Organized colloquium, "The Thinker from the Thought: Can We Separate Them?" in conjunction with CLLC, Emory University, March 3, 1993.

"The Future Perfect and the Perfect Future: History Has Its Reasons...." Presidential Address, Annual Meeting of the American Catholic Philosophical Association, Atlanta, March 27, 1993.

"Sartre and the Paradox of Committed History," Annual Thomasfest lecture, Xavier University, Cincinnati, OH, March 29, 1993.

"Foucault and the Eclipse of Vision," Faculty seminar, Xavier University, Cincinnati, OH, March 29, 1993.

Panel discussant of English translation of Sartre's Cahiers pour une Morale, meeting of Sartre Society of North America, Trent University, Peterborough, Ont., May 8, 1993.

"Postmodern Space(s)," panel organized for annual meeting of International Association for Philosophy and Literature, Pittsburgh, PA, May 13, 1993.

"Event, History, Totality: Lyotard and the Meaning(s) of It All," International Philosophical Seminar, Alto-Adige, Italy, July 8, 1993.

Introduction of Plenary Speaker, Jean-Luc Marion, annual meeting of the Society for Phenomenology and Existential Philosophy (SPEP), October 23, 1993.

Member of panel on "(un)Housing (the) Discipline," ILA, Emory University, October 27, 1993.

Commentator on two papers at meetings in conjunction with the American Philosophical Association Meeting, Atlanta: Steve Hendley on Sartre and Levinas (Sartre Circle) December 30, 1993, and Joseph Godfrey on Marcel's Interpersonalism (Marcel Society) December 28, 1993.

"(Postmodern) Faith and (Enlightenment) Reason," Aquinas Center, Emory University, January 18, 1994.

"The Future Perfect and the Perfect Future: History Has Its Reasons....," Presidential Address, 68th annual meeting of the American Catholic Philosophical Association, Atlanta, March 27, 1994.

Member of panel on "Choices and Responsibilities in a Changing University," Emory Symposium, April 14-15, 1994.

Organizer of Colloquium, "Thinking History: Poiesis and Praxis," International Association for

Philosophy and Literature, Edmonton, Alberta, May 4-7, 1994.

Chaired session at annual meeting of the Sartre Society of North America, DePaul University, Chicago, October 22, 1994.

Responded to "Sartre and Hegel" by R. Williams, Hiram College, Sartre Circle at the APA, Boston, December 29, 1994.

"Atheism, Modern and Postmodern," St. Charles Seminary of the Archdiocese of Philadelphia, February 27, 1995.

"Foucault in the Eyes of Sartre in the Eyes of Foucault," Villanova University, Villanova, PA, March 23, 1995.

"Foucault in the Eyes of Sartre in the Eyes of Foucault" (invited), regional meeting of the ACPA, Fordham University, New York, April 1, 1995.

"The Aporia of an Existentialist Philosophy of History" (invited), Georgetown University, Washington, DC, April 7, 1995.

"Atheism, Modern and Postmodern," St. Joseph's College, Philadelphia, PA, April 14, 1995.

Faculty seminar on "Patriotism and Cosmopolitanism" (Martha Nussbaum, The Boston Review), Montclair State University, Montclair, NJ, April 20, 1995.

"Foucault in the Eyes of Sartre in the Eyes of Foucault" (invited), regional meeting of the ACPA, LaSalle University, Philadelphia, April 22, 1995.

Comment on Foucault paper at the APA Central Division meeting, Chicago, April 28, 1995.

"Alterity and Appropriation: The Successful Failure of Historical Dialectic," International Philosophy Seminar, Kastelroto, Italy, July 8, 1995.

"Travel as Integrating a Liberal Education," "Philosophizing in the Mountains," and "What is Postmodeernism?" given to members of Association of Emory Alumni (Swiss Alumni College), Meiringen, Switzerland, August 3, 5, 7, 1995.

"Teaching as Vocation," Chaplain's Tea Talks, Emory University, October 3, 1995.

Comment on paper by Merald Westphal (Fordham University), "Faith as the Overcoming of Ontological Xenophobia," annual meeting of the Society for Phenomenology and Existential Philosophy, Chicago,

IL, October 13, 1995.

"The Diary and the Map: Sartre and Foucault on the Sense(s) of History," Emory University Philosophy Colloquium, October 19, 1995.

"Sartre and Foucault: History and Structure," colloquium on Recent Phenomenologies of Time, Center for Advanced Research in Phenomenology, Florida Atlantic University, Boca Raton, FL, November 19, 1995.

"Foucault and the Politics of Postmodernity" (invited), annual Brantl Lecture, Montclair State University, Montclair, NJ, November 30, 1995.

"Marcel and Postmodern Politics," Marcel Society in conjunction with the annual meeting of The American Philosophical Association, December 29, 1995.

"The Diary and the Map: Sartre and Foucault on Reason in History," invited lecture to faculty and students of University of Tennessee, Chattanooga, February 7, 1996.

"The Diary and the Map" (invited), Hendle Lecture, Creighton University, Omaha, NE, April 25, 1996.

"Reconstituting Praxis: Repetition as Re-enactment," International Association for Philosophy and Literature, George Mason University, Fairfax, VA, May 10, 1996.

"Eco's (of) Foucault: The Limits of Interpretation," International Philosophy Seminar, Bolzano, Italy, July 1, 1996.

"Foucault as Critic of Phenomenolgy: The Problem of Cross-Cultural Communication," Tohoku University, Sendai, Japan, Sept. 18, 1997

"A Matter of Justice: Foucault and the Problem of the Floating Transcendental," Center for Advanced Research in Phenomenology, Florida Atlantic University, Boca Raton, FL, October 26, 1997.

"Marcel's (Necessary) Misreading of Sartre," Marcel Society in conjunction with the APA, Atlanta, December 29, 1997.

"Who Inherits the Historical Dialectic? The Relevance of the New History to the Ideal of a Liberal Education," Society of Teachers of Philosophy in Jesuit Higher Education, in conjunction with the APA, Atlanta, December 29, 1997.

"Sartre, Foucault and Historical Reason: The Diary and the Map," Boston College, February 20, 1997.

"Does Emory Have a Soul?" (Plenary address to 16th Alumni Assembly), Emory University, April, 1997.

"History and Structure: Sartre and Foucault," University of Helsinki, Finland, June 20, 1997.

"Derrida and the Spirits of Marx," International Philosophy Seminar, Bolzano, Italy, July 2, 1997.

Response to Thomas Busch, Author meets Critic Session of Biennial Meeting ofSartre Society of North America (session on my recent book with the University of Chicago Press), October 5, 1997.

"Has the Academy Lost Its Soul?" (Plenary address to Minary Conference), Stanford University Center, Lake Tahoe, CA, October 20, 1997.

"At the Crossroads of Philosophy and Religion: The Test Case of Religious Existentialism," Georgia Southern University (invited address), November 6, 1997.

"Sartre on Violence and Foucault on Power: A Diagnostic," annual meeting of the Sartre Circle in conjunction with the APA, Philadelphia, PA, December 28, 1997. (invited)

"Religion and Community in Jean-Luc Nancy," International Philosophical Seminar, Castelrotto, Italy, June 25, 1998.

"Objectivity: The View from the Left Bank," invited session at 20th World Congress of Philosophy, Boston, MA, August 14, 1998.

"Elements of an Existentialist Theory of History," Sartre Society of North America in conjunction with the 20th World Congress of Philosophy, Boston, August 14, 1998 (invited).

"Postmodernism and the Catholic Tradition: A Response to Kenneth Schmitz," American Catholic Philosophical Association, session in conjunction with the 20th World Congress of Philosophy, Boston, August 11, 1998.

"The Philosopher as Cartographer: Foucault and the Mapping of History," Institute for Advanced Study, November 5, 1998.

Response to commentators at Session on my recently published book at annual meeting of the Society for Phenomenology and Existential Philosophy (SPEP), Denver, Colorado, October 8, 1998.

"Having Stars in Your Eyes," Commentary on Invited Paper by Charles Scott of Pennsylvania State University. At annual meeting of the American Philosophical Association (APA) in Washington, DC, December 30, 1998.

Commentary on papers by Matt Eschleman (Duquesne), Brian Seitz (Babson College) and Robert Boilou (University of New Mexico) on Sartre and Foucault at the biennial Meeting of the Sartre Society of North America, Los Angeles, CA, Feb. 13, 1999.

"The Philosopher-Historian as Cartographer," University of Western Ontario, London, Ontario, March 17, 1999 (invited).

"The Philosopher-Historian as Cartographer," University of Waterloo, Waterloo, Ontarion, March 18, 1999 (invited).

"Sartre and the Paradox of Committed History," University of Western Ontario, March 19, 1999 (invited).

"Ethics and Sexual Difference: Foucault and Irigaray," International Philosophical Seminar, Castelrotto, Italy, July 5, 1999.

"Time Squared: Historical Time in Sartre and Foucault," Emory University, September 16, 1999.

"Sartre and Foucault: The Diary and the Map" comparative Lieterature Roundtable, Emory, October 21, 1999.

"Sartre and the Paradox of Committed History," Boston College, Chestnut Hill, MA, Oct. 29, 1999.

Professional Service

Outside reviewer for Department of Philosophy, University of Notre Dame, November 5-8, 2000.

Outside reviewer for Department of Philosophy, Villanova University, Philadelphia, PA, December 3-5, 200.

Outside reviewer for Department of Philosophy, University of Colorado, Colorado Springs, 1991 and again in February, 1998.

Nominated by Nominating Committee of the American Philosophical Association as candidate for Vice President, President-elect, 1999-2000.

External examiner for Ph.D. dissertations at McGill University (Montreal) on Foucault and at the University of Helsinki (Finland) on Sartre in 1997.

Member, Advisory Committee to the Program Committee, American Philosophical Association, 1994-97.

President, American Catholic Philosophical Association, 1993-94.

Chair of Review Committee for "Work in Progress" sessions of the annual meeting, Society for Phenomenology and Existential Philosophy, 1992-95.

Referee for Journal of the History of Philosophy, American Catholic Philosophical Quarterly, International Philosophical Quarterly, The Journal of Aesthetics and Art Criticism, Faith and Philosophy.

Outside reviewer of Department of Philosophy, Vanderbilt University, February 2000.

Outside reviewer of Department and Graduate Program of Philosophy, Brock University, St. Catherines, Ontario, November 24-25, 1994.

Outside reviewer of University of Dallas for the Southern Association of Colleges and Schools (Commission on Colleges), April 10-13, 1994.

Outside Review Committee, Graduate Program in Philosophy, Fordham University, March 7-8, 1994.

Member of evaluation team for the Department of Philosophy and Relgion, Montclair State College, Upper Montclair, NJ, April 13, 1987.

Review fellowship applications to the National Humanities Center for 1994 and again for 1998.

Member Editorial Board of series, Perspectives in Continental Philosophy, ed.John D. Caputo (Fordham Univ. Press), 1993-.

Member of Board of Editorial Consultants of Faith and Philosophy, journals of the Society of Christian Philosophers, 1995-.

Editorial Consultant to The American Catholic Philosophical Quarterly, 1993-

Outside reviewer for candidates for promotion/tenure at the following universities:

- Purdue University (promotion and tenure), 1995
- DePaul University (promotion and tenure), 1995
- Carleton University, Canada, (promotion and tenure), 1995
- University of Michigan-Dearborn (tenure), 1991
- University of Colorado/Colorado Springs (tenure), 1991
- University of Colorado/Colorado Springs (promotion to professor), 1993.
- University of Notre Dame (promotion to full professor), 1991
- University of Michigan-Dearborn (promotion to professor), 1999

- Purdue University (promotion to professor), 1999
- Purdue University (promotion to Distinguished professor), 1999

Chair of the Committee on Committees of the Board of Officers of the American Philosophical Association for their general meeting in Philadelphia, October 3-5, 1991.

Chair of the Alfred Schutz Lectureship Committee of the American Philosophical Association (summer, 1991).

Co-Chair, Program Committee for the 1989 meeting of the International Association for Philosophy and Literature, Atlanta, GA, May 4-6, 1989.

Member of the Board of Officers of the American Philosophical Association (Representative of the Eastern Division), 1989-92.

Member Executive Committee of the American Catholic Philosophical Association, 1988-91.

Member Executive Committee of the International Association for Philosophy and Literature, 1988-93.

Chair, Executive Board of the Sartre Society of North America, 1988-91.

Chair, Program Committee for the 1987 meeting of the American Philosophical Association, Eastern Division. Responsible for the program of the largest annual meeting of philosophers in North America.

Program Committee for the 1986 meeting of the American Philosophical Association, Eastern Division (assess numerous submissions).

Program Committee for the 1986 meeting of the Society for Phenomenology and Existential Philosophy (assess scores of submissions).

Board of Directors of the Society for Phenomenology & Existential Philosophy, 1983-86.

Review Editor for Man and World, an international journal of philosophy, 1978-1997.

Referee for National Endowment for the Humanities (Special Programs) July 1986 and May 1992, and on a review panel, March 5 & 6, 1987; again (Research Programs) June, 1987.

Referee for University Presses: Chicago, Indiana, SUNY, Mercer, Penn State.

Founding member of Committee to establish an international Sartre Society. Inaugural conference, October 4-6, 1985, The New School, New York City.

Referee of submitted papers for the conference. Introduce keynote speaker. Referee of submitted papers for the 1987 and 1988 conferences.

Program Chairman of 1981 national meeting of the Metaphysical Society of America.

Member of:

- The American Philosophical Association.
- The Society for Phenomenology and Existential Philosophy.
- The Metaphysical Society.
- The American Catholic Philosophical Association.

National Treasurer of The American Catholic Philosophical Association and Business Manager of its review, The New Scholasticism, 1971-75.

International Association for Philosophy and Literature.

Assistant to the Dean, School of Philosophy, Catholic University, coordinating the programs of all undergraduate majors in philosophy, 1971-74.

Exchange graduate professor of philosophy, American University, Washington, DC, Fall term, 1973.

Member of curriculum committees both of the College of Arts and Sciences and of the School of Philosophy, Catholic University, 1971-75.

Chairman of the Curriculum Committee, School of Philosophy, Catholic University of America, 1973-74.

Service to the University

Chairman, Committee to Plan a Core Course for Freshmen, 1981-82.

Faculty Senate, Emory University, 1982-84.

University Research Committee, 1982-83.

Member of Committee on General Education, reviewing the general distribution requirements of Emory College, 1984-86.

Academic advisor to Merit Scholars, Emory College, 1980-83.

Member of the President's Committee on Undergraduate Education, Emory University.

Member, Admissions & Scholarship Committee, Emory College, 1979-83.

Chairman, Subcommittee of the President's Committee on Undergraduate Education, dealing with the Freshman Experience, 1980-81.

Chaplain's Advisory Committee on Religious Policy, 1979- .

Committee on the Emory College Advisory Program, 1980-81.

Lilly Endowment Summer Research Seminar, June, 1981.

Dobbs Hall Experiment in advising, 1981-83.

Chair of Emory Williams Teaching Award Committee 1984-85.

Academic advisor to Honors Students, Emory College, 1985- .

President's Advisory Committee on South Africa, 1985-86.

Consultant on Ford Foundation Proposal, 1985-86.

Consultant on Proposal for a Henry Luce Professorship at Emory, 1985-86.

Member of Search Committee for a Henry Luce Professor, 1986-87.

Member of Educational Policy Committee, Acting Chair, Fall 1985.

Member of Scholar/Teacher of the Year nominating committee, 1987, 1988, and 1990.

Member of the Business School Ethics Task Force, 1988-89.

Member of Search Committee for University Chaplain, 1989-90.

Member of College Committee to evaluate the fraternity/sorority system on campus (1990-91).

Member of Dean's Committee to review leave and release time policy in the College (1992).

Chair, Department of Philosophy (1988-91).

Chair, Center for Language, Literature, and Culture, Emory Univ., 1992-94.

Fulbright Interview Committee, 1992- .

Advisory Committee, Culture, History, Theory Program (ILA), 1993- .

Steering Committee, Friends of Theater Emory, 1993-95.

Member, Cannon Chapel Committee, 1993-94.

Member, University Committee on Law and Religion, 1995- .

Member, Advisory Committee on Center for Curricular Renewal and Teaching Enhancement, Emory College, 1995- .

Member, ad hoc Advisory Committee to Provost on disputed tenure decision, Dec.-Jan. 1996-97.

Associated Faculty of the Aquinas Center, Emory, 1997- .

Addendum

Radio interview with Wayne Pond for "Soundings" on PBS. Topic, "The Public Philosophy." Aired over two hundred PBS stations in March, 1992.

Television interview with Canadian television company for a 2-hour program on"Life After Death," aired on Canadian television last Spring and in the United States, on the Learning Channel, December, 1997.

Stanford Encyclopedia of Philosophy
List of Authors

Navigation Panel:

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

A

- **Alexander, J. McKenzie**
  - Evolutionary Game Theory
- **Allen, Anita**
  - Feminist Philosophy of Law
- **Allen, Colin**
  - Animal Consciousness
  - Teleological Notions in Biology
- **Allen, James**
  - Antiochus of Ascalon
  - Carneades
  - Zeno of Citium
- **Allen, Richard**
  - David Hartley
- **Altman, Andrew**
  - Civil Rights
- **Anderson, Elizabeth**
  - Feminist Epistemology and Philosophy of Science
- **Anderson, Lanier**
  - Heinrich Rickert
  - Hermann Cohen
  - Wilhelm Windelband
- **Andrews, Robert**
  - Medieval Theories of the Categories
- **Antonelli, Aldo**
  - Non-monotonic Logic
- **Archard, David William**
  - Children's Rights
- **Arneson, Richard**
  - Egalitarianism

  - Equality of Opportunity
- **Arntzenius, Frank**
  - Reichenbach's Common Cause Principle
  - Time Travel and Modern Physics (with Tim Maudlin)
- **Arpaly, Nomy**
  - Weakness of Will
- **Ashworth, E. Jennifer**
  - Medieval Theories of Analogy
  - Medieval Theories of Singular Terms
- **Avigad, Jeremy**
  - The Epsilon Calculus (with Richard Zach)
- **Aydede, Murat**
  - The Language of Thought Hypothesis
  - Pain

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

B

- **Bacciagaluppi, Guido**
  - The Problem of the Classical Limit in Quantum Mechanics
  - The Role of Decoherence in Quantum Mechanics
- **Bacon, John**
  - Tropes
- **Baird, Davis**
  - Scientific Instruments
- **Baltzly, Dirk**
  - Stoicism
- **Barber, Alex**
  - Idiolects
- **Barber, Michael**
  - Alfred Schütz
- **Barney, Rachel**
  - Callicles and Thrasymachus
- **Barrett, Jeffrey**
  - Everett's Relative-State Formulation of Quantum Mechanics
- **Barry, Christian**
  - Redistribution
- **Batterman, Robert**

  - Intertheory Relations in Physics
- **Bayne, Tim**
  - Parenthood (with Avery Kolers)
- **Beall, JC**
  - Curry's Paradox
  - Logical Consequence (with Greg Restall)
- **Beaney, Michael**
  - Analysis
- **Bedau, Hugo Adam**
  - Punishment
- **Beiser, Fred**
  - August Wilhelm Rehberg
- **Bell, Daniel**
  - Communitarianism
- **Bell, John L.**
  - Infinitary Logic
- **Berkovitz, Joseph**
  - Action at a Distance
- **Bermúdez, José**
  - Nonconceptual Mental Content
- **Bett, Richard**
  - Pyrrho
  - Timon of Phlius
- **Beyer, Christian**
  - Edmund Husserl
- **Biard, Joël**
  - Albert of Saxony
- **Bickle, John**
  - Multiple Realizability
  - The Philosophy of Neuroscience (with Peter Mandik)
- **Bigelow, John**
  - Simpson's Paradox (with Gary Malinas)
- **Biletzki, Anat**
  - Ludwig Wittgenstein (with Anat Matar)
- **Billington, David** ERROR Billington, David
- **Bishop, Michael** ERROR Bishop, Michael
- **Bix, Brian**
  - John Austin
- **Blackburn, Simon** ERROR Blackburn, Simon
- **Blake, Michael**
  - International Justice

- **Bobonich, Chris**
  - Plato on Utopia
- **Bodnar, Istvan**
  - Aristotle's Physics
- **Bohman, James**
  - Critical Theory
  - Jürgen Habermas
- **Bok, Hilary**
  - Baron de Montesquieu
- **BonJour, Laurence**
  - Epistemological Problems of Perception
- **Bouchard, Frederic**
  - Fitness (with Alexander Rosenberg)
- **Bowie, Andrew**
  - Friedrich Schlegel
  - Friedrich Wilhelm Joseph von Schelling
  - Novalis [Friedrich Leopold, Baron von Hardenberg]
- **Boyd, Richard**
  - Scientific Realism
- **Bradie, Michael**
  - Evolutionary Epistemology (with William Harms)
- **Brading, Katherine**
  - Symmetry and Symmetry Breaking (with Elena Castellani)
- **Brandl, Johannes**
  - Brentano's Theory of Judgement
- **Brandon, Robert**
  - Adaptation and Adaptationism
  - Natural Selection
- **Braun, David**
  - Indexicals
- **Breazeale, Dan**
  - Johann Gottlieb Fichte
  - Karl Leonhard Reinhold
- **Brennan, Andrew**
  - Environmental Ethics (with Yeuk-Sze Lo)
  - Necessary and Sufficient Conditions
- **Bridges, Douglas**
  - Constructive Mathematics
- **Bringsjord, Selmer**
  - Artificial Intelligence
- **Brittain, Charles**

    - Arcesilaus
    - Philo of Larissa
    - Saint Augustine and Greek Philosophy
- **Broadie, Alexander**
    - Scottish Philosophy in the 18th Century
- **Brogaard, Berit**
    - Fitch's Paradox of Knowability (with Joe Salerno)
- **Brook, Andrew**
    - The Unity of Consciousness
- **Brower, Jeffrey**
    - Medieval Theories of Relations
- **Brown, James R.**
    - Thought Experiments
- **Brown, Curtis**
    - Narrow Mental Content
- **Brown, Eric**
    - Cosmopolitanism (with Pauline Kleingeld)
    - Plato's Ethics and Politics in *The Republic*
- **Bub, Jeffrey**
    - Quantum Entanglement and Information
- **Buchanan, Allen**
    - Secession
- **Bueno, Otávio**
    - Nominalism in the Philosophy of Mathematics
- **Burch, Robert**
    - Charles Sanders Peirce
- **Busch, Thomas** ERROR Busch, Thomas
- **Buss, Sarah**
    - Personal Autonomy
- **Bynum, Terrell**
    - Computer Ethics: Basic Concepts and Historical Overview
- **Byrne, Alex**
    - Inverted Qualia

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

C

- **Callender, Craig**

- Thermodynamic Asymmetry in Time
- **Campbell, Richmond**
    - Moral Epistemology
- **Campbell, Kenneth**
    - Legal Rights
- **Candlish, Stewart**
    - Francis Herbert Bradley
    - The Identity Theory of Truth
    - Private Language
- **Card, Claudia**
    - Feminist Moral Psychology
- **Carlin, Laurence**
    - Leibniz's Philosophy of Mind (with Mark Kulstad)
- **Carroll, John W.**
    - Laws of Nature
- **Carruthers, Peter**
    - Higher-Order Theories of Consciousness
- **Carter, Ian**
    - Positive and Negative Liberty
- **Casati, Roberto**
    - Events (with Achille Varzi)
    - Holes (with Achille Varzi)
- **Castellani, Elena**
    - Symmetry and Symmetry Breaking (with Katherine Brading)
- **Caston, Victor**
    - Ancient Theories of Intentionality
- **Celano, Anthony**
    - Medieval Theories of Practical Reason
- **Chakrabarti, Arindam**
    - Epistemological Problems of Testimony
- **Chan, Alan**
    - Laozi
- **Christiano, Tom**
    - Authority
    - Democracy
- **Christman, John**
    - Autonomy in Moral and Political Philosophy
- **Clarke, Randolph**
    - Incompatibilist (Nondeterministic) Theories of Free Will
- **Cleary, Denis**
    - Antonio Rosmini

- **Clifton, Rob**
  - The Einstein-Podolsky-Rosen Argument in Quantum Theory
- **Code, Alan**
  - Aristotle
- **Cohen, S. Marc**
  - Aristotle's Metaphysics
- **Cohon, Rachel**
  - Hume's Moral Philosophy
- **Cole, David**
  - The Chinese Room Argument
- **Coleman, Jules**
  - Theories of Tort Law
- **Colyvan, Mark**
  - Indispensability Arguments in the Philosophy of Mathematics
- **Conee, Earl** ERROR Conee, Earl
- **Conti, Alessandro**
  - Johannes Sharpe
  - John Wyclif
  - Paul of Venice
  - Robert Alyngton
  - Walter Burley
  - William Penbygull
- **Copeland, B. Jack**
  - Arthur Prior
  - The Church-Turing Thesis
  - The Modern History of Computing
- **Corcoran, John**
  - Argument
  - Function
  - History of Logic
  - *Omega*
  - Schema
- **Corrigan, Kevin**
  - Pseudo-Dionysius the Areopagite
- **Cortens, Andrew**
  - Monism
- **Costelloe, Timothy**
  - Giambattista Vico
- **Cover, Jan**
  - Leibniz's Modal Metaphysics
- **Cowie, Fiona**

- Linguistic Innatism
- **Cox, Damian**
  - Integrity (with Marguerite La Caze and Michael Levine)
- **Crane, Tim**
  - Perception
- **Crisp, Roger**
  - Well Being
- **Cross, Richard**
  - Medieval Theories of Haecceity
- **Crowell, Steven**
  - Existentialism
- **Cudd, Ann**
  - Contractarianism
- **Cunning, David**
  - Descartes' Modal Metaphysics
- **Curd, Patricia**
  - Anaxagoras
  - Presocratic Philosophy
- **Curiel, Erik**
  - Singularities and Black Holes

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

D

- **D'Agostino, Fred**
  - Contemporary Approaches to the Social Contract
  - Original Position
  - Public Justification
- **D'Arms, Justin**
  - Envy
- **Dahlstrom, Daniel**
  - Moses Mendelssohn
- **Dancy, Russell**
  - Speusippus
  - Xenocrates
- **Dancy, Jonathan**
  - Moral Particularism
- **Daniels, Norman**
  - Reflective Equilibrium

- **Darden, Lindley**
  - Molecular Biology
- **Dauenhauer, Bernard**
  - Paul Ricoeur
- **David, Marian**
  - The Correspondence Theory of Truth
- **de Sousa, Ronald**
  - Emotion
- **DeCew, Judith**
  - Privacy
- **Des Chene, Dennis**
  - Aristotelianism in the Renaissance
  - Francesco Patrizi
  - Francisco Suárez
- **Deutsch, Harry**
  - Free Logic
  - Relative Identity
- **Deutscher, Penelope**
  - Feminist Approaches to Continental Philosophy
- **di Giovanni, George**
  - Friedrich Heinrich Jacobi
- **Dickson, Michael**
  - Modal Interpretations of Quantum Mechanics
- **Dickson, Julie**
  - Interpretation and Coherence in Legal Reasoning
- **Dillon, Robin S.**
  - Respect
- **DiSalle, Robert**
  - Space and Time: Inertial Frames
- **Divers, John**
  - Possible Worlds
- **Dombrowski, Dan**
  - Charles Hartshorne
- **Dowe, David**
  - The Turing Test (with Graham Oppy)
- **Dowe, Phil**
  - Causal Processes
- **Downes, Steve**
  - Heritability
- **Downing, Lisa**
  - George Berkeley

- **Duff, Antony**
  - Legal Punishment
  - Theories of Criminal Law
- **Dworkin, Gerald**
  - Paternalism

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

E

- **Easton, Patricia**
  - Antoine Le Grand
  - Robert Desgabets
- **Edgington, Dorothy**
  - Conditionals
- **Ehrlich, Philip**
  - Non-Archimedean Geometry
- **Endicott, Timothy**
  - Law and Language
- **Ereshefsky, Marc**
  - Species
- **Eshleman, Andrew**
  - Moral Responsibility

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

F

- **Føllesdal, Andreas**
  - Federalism
- **Falkenstein, Lorne**
  - Étienne Bonnot de Condillac
- **Faye, Jan**
  - Backward Causation
  - Copenhagen Interpretation of Quantum Mechanics
- **Feldman, Richard**
  - Naturalized Epistemology

- **Fisher, Saul**
  - Pierre Gassendi
- **Fitch, Greg**
  - Singular Propositions
- **Fitelson, Branden**
  - Confirmation
- **Flanagan, Owen**
  - Moral Psychology
- **Flores, Francisco**
  - The Equivalence of Mass and Energy
- **Floridi, Luciano**
  - Semantic Conceptions of Information
- **Flynn, Thomas**
  - Jean-Paul Sartre
- **Forrest, Peter**
  - The Epistemology of Religion
  - The Identity of Indiscernibles
- **Forster, Michael**
  - Friedrich Daniel Schleiermacher
  - Johann Gottfried von Herder
- **Franklin, Allan**
  - Experiment in Physics
- **Fraser, Chris**
  - Mohism
  - Mohist Canons
  - School of Names
- **Frede, Dorothea**
  - Plato's Ethics and Cosmology
- **French, Steven**
  - Identity and Individuality in Quantum Theory
- **Friedman, Michael**
  - Ernst Cassirer
- **Friedman, Russell L.**
  - Peter Auriol
- **Fullinwider, Robert**
  - Affirmative Action
- **Fumerton, Richard**
  - Foundationalist Theories of Epistemic Justification

G

- **Gale, George**
  - Cosmology: Methodological Debates in the 1930s and 1940s
- **Gallagher, Shaun**
  - Self-Consciousness
- **Galton, Antony**
  - Temporal Logic
- **Gannett, Lisa**
  - The Human Genome Project
- **Garry, Ann**
  - Feminist Approaches to Analytic Philosophy
- **Garson, James**
  - Connectionism
  - Modal Logic
- **Gaus, Gerald**
  - Liberalism
- **Gelber, Hester**
  - Robert Holkot
- **George, Robert**
  - Natural Law Theories
  - The Rule of Law and Procedural Fairness
- **Gerson, Lloyd**
  - Plotinus
- **Gert, Bernard**
  - The Definition of Morality
- **Gertler, Brie**
  - Self-knowledge
- **Ghirardi, Giancarlo**
  - Collapse Theories
- **Gill, Michael**
  - Lord Shaftesbury [Anthony Ashley Cooper, 3rd Earl of Shaftesbury]
- **Gjesdal, Kristin**
  - Hermeneutics (with Bjørn Ramberg)
- **Godfrey-Smith, Peter**
  - Biological Information (with Kim Sterelny)
- **Goldman, Alvin**
  - Social Epistemology
- **Goldstein, Sheldon**

  - Bohmian Mechanics
- **Goodman, Russell**
  - Chauncey Wright
  - Ralph Waldo Emerson
  - Transcendentalism
  - William James
- **Gordon, Robert M.**
  - Folk Psychology as Mental Simulation
- **Gosepath, Stefan**
  - Equality
- **Gotthelf, Allan**
  - Aristotle's Biology
- **Gottwald, Siegfried**
  - Many-Valued Logic
- **Graham, George**
  - Behaviorism
- **Graham, Gordon**
  - Scottish Philosophy in the 19th Century
- **Grattan-Guinness, Ivor**
  - Benjamin Peirce (with Alison Walsh)
- **Greco, John**
  - Virtue Epistemology
- **Green, Leslie**
  - Legal Obligation and Authority
  - Legal Positivism
- **Griffith-Dickson, Gwen**
  - Johann Georg Hamann
- **Griffiths, Paul**
  - Developmental Biology (with Lenny Moss)
  - The Distinction Between Innate and Acquired Characteristics
  - Philosophy of Biology (with Sahotra Sarkar)
- **Griswold, Charles**
  - Plato on Rhetoric and Poetry
- **Groarke, Leo**
  - Ancient Skepticism
  - Informal Logic
- **Gruen, Lori**
  - Animal Rights
- **Gruenbaum, Adolf**
  - Cosmology and Theology
- **Gutting, Gary**

- Michel Foucault

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

H

- **Hájek, Alan**
  - Interpretations of the Probability Calculus
  - Pascal's Wager
- **Hajek, Petr**
  - Fuzzy Logic
- **Hammer, Eric**
  - Peirce's Logic
  - The Revision Theory of Truth
- **Hansen, Chad**
  - Taoism
- **Hardin, Russell**
  - Free Rider Problem
- **Harms, William**
  - Evolutionary Epistemology (with Michael Bradie)
- **Harris, Ian**
  - Edmund Burke
- **Harrison, Ross**
  - James Mill
  - Jeremy Bentham
- **Harvey, Irene**
  - Jacques Derrida
- **Haslanger, Sally**
  - Approaches to Feminism (with Nancy Tuana)
  - Feminist Interventions (with Nancy Tuana)
  - Topics in Feminism (with Nancy Tuana)
- **Hawley, Katherine**
  - Temporal Parts
- **Hazen, Allen**
  - Plural Quantification
- **Healey, Richard**
  - Holism and Nonseparability in Physics
- **Heil, John**
  - Mental Causation

- **Held, Carsten**
  - The Kochen-Specker Theorem
- **Heyd, David**
  - Supererogation
- **Heyes, Cressida**
  - Identity Politics
- **Hilgevoord, Jan**
  - The Uncertainty Principle (with Jos Uffink)
- **Hilpinen, Risto**
  - Artifact
- **Hitchcock, Christopher**
  - Probabilistic Causation
- **Hodges, Wilfrid**
  - First-order Model Theory
  - Logic and Games
  - Model Theory
  - Tarski's Truth Definitions
- **Hodges, Andrew**
  - Alan Turing
- **Hoefer, Carl**
  - Causal Determinism
- **Hoenen, Maarten**
  - Marsilius of Inghen
- **Hoffman, Joshua**
  - Omnipotence (with Gary Rosenkrantz)
- **Hofweber, Thomas**
  - Logic and Ontology
- **Hogarth, Mark**
  - Malament-Hogarth Spacetimes and the New Computability
- **Holcomb, Harmon**
  - Sociobiology and Evolutionary Psychology
- **Holton, Richard**
  - Trust
- **Homiak, Marcia**
  - Moral Character
- **Honoré, Antony**
  - Causation in the Law
- **Hooker, Brad**
  - Henry Sidgwick
  - Rule Consequentialism
- **Horst, Steven**

- Computational Models of Mind
- **Howard, Don**
  - The Einstein-Bohr Debates
  - Einstein's Philosophy of Science
- **Howard-Snyder, Frances**
  - Doing vs. Allowing Harm
- **Huemer, Wolfgang**
  - Franz Brentano
- **Huggett, Nick**
  - Zeno's Paradoxes
- **Hull, David**
  - Replication
- **Hursthouse, Rosalind**
  - Virtue Ethics
- **Hussain, Nadeem J. Z.**
  - Friedrich Albert Lange
- **Hutton, Sarah**
  - The Cambridge Platonists
  - Lady Anne Conway
  - Lady Damaris Masham
- **Hyde, Dominic**
  - Sorites Paradox
- **Hyslop, Alec**
  - Other Minds

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

I

- **Irvine, A. D.**
  - Alfred North Whitehead
  - Bertrand Russell
  - *Principia Mathematica*
  - Russell's Paradox
- **Ismael, Jenann**
  - Quantum Mechanics

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

J

- **Jacob, Pierre**
  - Intentionality
- **Janis, Allen**
  - Conventionality of Simultaneity
- **Jech, Thomas**
  - Set Theory
- **Jennings, Ray**
  - Connectives
  - Disjunction
- **Jeshion, Robin**
  - *A Priori* Justification and Knowledge
- **Jeske, Diane**
  - Special Obligations
- **Johnson, Robert**
  - Kant's Moral Philosophy
- **Jollimore, Troy**
  - Impartiality
- **Jones, Martin**
  - Bell's Theorem
- **Joyce, James**
  - Bayes' Theorem
  - Causal Decision Theory
- **Jung-Palczewska, Elzbieta**
  - Richard Kilvington

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

K

- **Kahane, David**
  - Diversity
- **Kalderon, Mark Eli**
  - Fictionalism
- **Katz, Leonard D.**
  - Pleasure
- **Keller, Pierre**

    - Dialectic
- **Kellner, Douglas**
    - Jean Baudrillard
- **Khlentzos, Drew**
    - Semantic Challenges to Realism
- **Kim, Alan**
    - Paul Natorp
- **King, Peter**
    - Peter Abelard
- **King, Jeffrey C.**
    - Anaphora
    - Structured Propositions
- **Kirk, Robert**
    - Zombies
- **Klein, Juergen**
    - Francis Bacon
- **Klein, Peter**
    - Skepticism
- **Kleingeld, Pauline**
    - Cosmopolitanism (with Eric Brown)
- **Klima, Gyula**
    - The Medieval Problem of Universals
- **Knight, Jack**
    - Social Institutions
- **Knuuttila, Simo**
    - Medieval Theories of Modality
- **Kolers, Avery**
    - Parenthood (with Tim Bayne)
- **Korcz, Keith Allen**
    - The Epistemic Basing Relation
- **Kornhauser, Lewis**
    - Legal Philosophy: The Economic Analysis of Law
- **Kraus, Jody**
    - Theories of Contracts
- **Kraut, Richard**
    - Aristotle's Ethics
    - Plato
- **Krips, Henry**
    - Measurement in Quantum Theory
- **Kronz, Fred**
    - Quantum Theory: von Neumann vs. Dirac

- **Kuhn, Steven**
  - Prisoner's Dilemma
- **Kulstad, Mark**
  - Leibniz's Philosophy of Mind (with Laurence Carlin)

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

L

- **La Caze, Marguerite**
  - Integrity (with Damian Cox and Michael Levine)
- **Lambertini, Roberto**
  - Giles of Rome
- **Lamond, Grant**
  - Precedent and Analogy in Legal Reasoning
- **Lamont, Julian**
  - Distributive Justice
- **Langston, Douglas**
  - Medieval Theories of Conscience
- **Laraudogoitia, Jon Pérez**
  - Supertasks
- **Lau, Joe**
  - Externalist Theories of Mental Content
- **Laubichler, Manfred**
  - Character/Trait
- **Laudisa, Federico**
  - Relational Quantum Mechanics (with Carlo Rovelli)
- **Laurence, Stephen**
  - Concepts (with Eric Margolis)
- **Lawlor, Leonard**
  - Henri Bergson
- **Laycock, Henry**
  - Object
- **Le Poidevin, Robin**
  - The Experience and Perception of Time
- **LeBuffe, Michael**
  - Paul-Henri Dietrich (Baron) d'Holbach
  - Spinoza's Psychological Theory
- **Leftow, Brian**

- Divine Simplicity
- Eternity
- Immutability
- Ontological Dependence

- **Leiter, Brian**
  - Legal Realisms
  - Naturalism in Legal Philosophy
- **Lemon, Oliver**
  - Diagrams (with Sun-Joo Shin)
- **Lennon, Thomas M.**
  - Pierre Bayle
- **Lennox, James**
  - Darwinism
- **Leopold, David**
  - Max Stirner
- **Lesher, James**
  - Xenophanes
- **Leslie, John**
  - Cosmology and Theology
- **Levin, Janet**
  - Functionalism
- **Levine, Michael**
  - Integrity (with Damian Cox and Marguerite La Caze)
  - Miracles
  - Pantheism
- **Lewontin, Richard**
  - The Genotype/Phenotype Distinction
- **Linsky, Bernard**
  - Logical Constructions
- **Lloyd, Sharon A.**
  - Hobbes's Moral and Political Philosophy
- **Lloyd, Lisa**
  - Units (and Levels) of Selection
- **Lo, Yeuk-Sze**
  - Environmental Ethics (with Andrew Brennan)
- **Loar, Brian**
  - Mental Content
- **Loewer, Barry**
  - Quantum Theory and Free Will
- **Lokhorst, Gert-Jan**
  - Mally's Deontic Logic

- **Long, Anthony**
  - Epictetus
- **Longeway, John**
  - Medieval Theories of Demonstration
  - Simon of Faversham
  - William Heytesbury
- **Longino, Helen**
  - The Social Dimensions of Scientific Knowledge
- **Lorenz, Hendrik**
  - Ancient Theories of Soul
- **Ludlow, Peter**
  - Descriptions
- **Luper, Steven**
  - Death
  - The Epistemic Closure Principle
- **Lycan, William**
  - Representational Theories of Consciousness

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

M

- **Macbride, Fraser**
  - Neologicism
- **MacIntosh, J. J.**
  - Robert Boyle
- **Malinas, Gary**
  - Simpson's Paradox (with John Bigelow)
- **Malpas, Jeff**
  - Donald Davidson
  - Hans-Georg Gadamer
- **Mandik, Peter**
  - The Philosophy of Neuroscience (with John Bickle)
- **Mann, Wolfgang**
  - Plato's Academy
- **Mares, Edwin**
  - Relevance Logic
- **Margolis, Eric**
  - Concepts (with Stephen Laurence)

- **Markosian, Ned**
    - Time
- **Marmo, Costantino** ERROR Marmo, Costantino
- **Marmor, Andrei**
    - The Nature of Law
    - The Pure Theory of Law
- **Marquis, Jean-Pierre**
    - Category Theory
- **Martin, Christopher**
    - Anicius Manlius Severinus Boethius
    - Logic in the Twelfth Century
- **Martin, Robert**
    - The St. Petersburg Paradox
- **Matar, Anat**
    - Ludwig Wittgenstein (with Anat Biletzki)
- **Matheson, Carl**
    - Historicist Theories of Rationality
- **Matthews, Gareth**
    - The Philosophy of Childhood
- **Maudlin, Tim**
    - Time Travel and Modern Physics (with Frank Arntzenius)
- **Maund, Barry**
    - Color
- **McCann, Hugh J.**
    - Divine Providence
- **McCarthy, John**
    - Logic and Artificial Intelligence
- **McCluskey, Colleen**
    - Philip the Chancellor
- **McConnell, Terrance**
    - Moral Dilemmas
- **McDermott, John** ERROR McDermott, John
- **McDonald, William**
    - Søren Kierkegaard
- **McInerny, Ralph**
    - Saint Thomas Aquinas
- **McIntyre, Alison**
    - Doctrine of Double Effect
- **McKay, Thomas**
    - Propositional Attitude Reports
- **McKenna, Michael**

- Compatibilism
- **McLaughlin, Brian**
  - Supervenience
- **McLeod, Owen**
  - Desert
- **McNamara, Paul**
  - Deontic Logic
- **McNaughton, David**
  - Deontological Ethics (with Piers Rawling)
  - Morality and Practical Reason (with Piers Rawling)
- **Melamed, Yitzhak**
  - Salomon Maimon (with Peter Thielke)
- **Mendell, Henry**
  - Aristotle and Mathematics
- **Mendelson, Michael**
  - Saint Augustine
- **Menzel, Christopher**
  - Actualism
- **Menzies, Peter**
  - Counterfactual Theories of Causation
  - Counterfactuals
- **Meyer, Lukas**
  - Intergenerational Justice
- **Meyers, Diana**
  - Feminist Perspectives on the Self
- **Miller, Alexander**
  - Realism
- **Miller, Fred**
  - Aristotle's Political Theory
- **Miller, Barry**
  - Existence
- **Miller, Dale E.**
  - Harriet Taylor Mill
- **Miscevic, Nenad**
  - Nationalism
- **Mitchell, Sandra**
  - Pluralism in Biology
  - The Unity of Science
- **Moggach, Douglas**
  - Bruno Bauer

- **Monk, J. Donald**
  - The Mathematics of Boolean Algebra
- **Moore, Andrew**
  - Hedonism
- **Moran, Dermot**
  - John Scottus Eriugena
- **Morris, William Edward**
  - David Hume
- **Morris, Christopher**
  - Game Theory and Ethics (with Bruno Verbeek)
- **Morscher, Edgar**
  - Bernard Bolzano
- **Mortensen, Chris**
  - Change
  - Inconsistent Mathematics
- **Moschovakis, Joan**
  - Intuitionistic Logic
- **Moss, Lenny**
  - Developmental Biology (with Paul Griffiths)
- **Most, Glenn**
  - Aristotle's Poetics
- **Moulakis, Athanasios**
  - Civic Humanism
- **Mueller-Vollmer, Kurt**
  - Wilhelm von Humboldt
- **Murphy, Mark**
  - The Natural Law Tradition in Ethics
  - Theological Voluntarism
- **Murray, Michael**
  - Leibniz on the Problem of Evil
  - Philosophy and Christian Theology

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

N

- **Nadler, Steven**
  - Baruch Spinoza
- **Nambiar, Sriram**

- George Boole
- **Neander, Karen**
  - Teleological Theories of Mental Content
- **Nelkin, Dana K.**
  - Moral Luck
- **Nelson, Alan**
  - Gottfried Wilhelm Leibniz
  - René Descartes
- **Newman, Lex**
  - Descartes' Epistemology
- **Nickel, James**
  - Human Rights
- **Nida-Rümelin, Martine**
  - Qualia: The Knowledge Argument
- **Niiniluoto, IIkka**
  - Scientific Progress
- **Nolan, Lawrence**
  - Descartes' Ontological Argument
  - Malebranche's Theory of Ideas and Vision in God
- **Nolan, Daniel**
  - Modal Fictionalism
- **Noone, Tim**
  - Saint Bonaventure
- **Normore, Calvin**
  - Medieval Theories of Future Contingents
  - Medieval Theories of Intentionality
- **Norton, John**
  - The Hole Argument

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

O

- **O'Connor, Timothy**
  - Emergent Properties (with Hong Yu Wong)
  - Free Will
- **Oddie, Graham**
  - Truthlikeness
- **Okasha, Samir**
  - Biological Altruism

- **Olson, Eric T.**
    - Personal Identity
- **Oppy, Graham**
    - Ontological Arguments
    - The Turing Test (with David Dowe)
- **Orend, Brian**
    - War

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

P

- **Pagin, Peter**
    - Assertion
- **Pappas, George**
    - Internalist vs. Externalist Conceptions of Epistemic Justification
- **Parker, Kelly A.**
    - Josiah Royce
- **Parry, Richard**
    - Ancient Ethical Theory
    - Empedocles
    - *Episteme* and *Techne*
- **Parsons, Terence**
    - The Traditional Square of Opposition
- **Pasnau, Robert**
    - Divine Illumination
    - Peter John Olivi
- **Perring, Christian**
    - Mental Illness
- **Pessin, Sarah**
    - Maimonides [Moses ben Maimon]
    - Saadiah Gaon
- **Pettit, Philip**
    - Republicanism
- **Philp, Mark**
    - William Godwin
- **Philpott, Dan**
    - Sovereignty
- **Pickett, Brent**

    - Homosexuality
- **Pietroski, Paul**
    - Logical Form
- **Pigden, Charles**
    - Russell's Moral Philosophy
- **Pironet, Fabienne**
    - *Sophismata*
- **Pitt, David**
    - Mental Representation
- **Pohlers, Wolfram**
    - Proof Theory
- **Portoraro, Frederic**
    - Automated Reasoning
- **Preston, John**
    - Paul Feyerabend
- **Priest, Graham**
    - Dialetheism
    - Paraconsistent Logic (with Koji Tanaka)
- **Pritchard, Michael**
    - Philosophy for Children

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Q

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

R

- **Ramberg, Bjørn**
    - Hermeneutics (with Kristin Gjesdal)
    - Richard Rorty
- **Ramsey, William**
    - Eliminative Materialism
- **Rapp, Christof**
    - Aristotle's Rhetoric
- **Ravenscroft, Ian**

    - Folk Psychology as a Theory
- **Rawling, Piers**
    - Deontological Ethics (with David McNaughton)
    - Morality and Practical Reason (with David McNaughton)
- **Read, Stephen**
    - Medieval Theories: Properties of Terms
- **Reath, Andrews**
    - Constructivism
- **Redding, Paul**
    - Georg Wilhelm Friedrich Hegel
- **Reeve, C. D. C.**
    - Plato on Friendship and Eros
- **Reimer, Marga**
    - Reference
- **Rescher, Nicholas**
    - Process Philosophy
- **Restall, Greg**
    - Logical Consequence (with JC Beall)
    - Substructural Logics
- **Rey, Georges**
    - The Analytic/Synthetic Distinction
- **Rheinberger, Hans-Joerg**
    - Gene
- **Rice, Hugh**
    - Fatalism
- **Richardson, Henry**
    - Moral Reasoning
- **Ricketts, Thomas**
    - Rudolf Carnap
- **Ridge, Michael**
    - Moral Non-Naturalism
- **Riegel, Jeffrey**
    - Confucius
- **Robert, Jason Scott**
    - Evolution and Development
- **Robins, Dan**
    - Xunzi
- **Robinson, William**
    - Epiphenomenalism
- **Robinson, Howard**
    - Dualism

- **Rosen, Gideon**
  - Abstract Objects
- **Rosenberg, Jay**
  - Wilfrid Sellars
- **Rosenberg, Alexander**
  - Fitness (with Frederic Bouchard)
- **Rosenkrantz, Gary**
  - Omnipotence (with Joshua Hoffman)
- **Ross, Don**
  - Game Theory
- **Roth, Harold**
  - Zhuangzi
- **Rovelli, Carlo**
  - Relational Quantum Mechanics (with Federico Laudisa)
- **Rowe, William**
  - Divine Freedom
- **Rowe, Christopher**
  - Plato on the Sophist and the Statesman
- **Rudavsky, Tamar**
  - Gersonides
- **Ruse, Michael**
  - Creationism
- **Rutherford, Donald**
  - Descartes' Ethics
  - Leibniz's Ethics
- **Ryckman, Thomas A.**
  - Early Philosophical Interpretations of General Relativity
- **Rynasiewicz, Robert**
  - Newton's Views on Space, Time, and Motion

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

S

- **Saatkamp, Herman**
  - George Santayana
- **Sahlin, Nils-Eric** ERROR Sahlin, Nils-Eric
- **Salerno, Joe**
  - Fitch's Paradox of Knowability (with Berit Brogaard)

- **Sanford, David H.**
  - Determinates vs. Determinables
- **Sarkar, Sahotra**
  - Conservation Biology
  - Ecology
  - Philosophy of Biology (with Paul Griffiths)
- **Sassen, Brigitte**
  - 18th Century German Philosophy Prior to Kant
- **Satz, Debra**
  - Feminist Perspectives on Reproduction and the Family
- **Saul, Jennifer**
  - Feminist Philosophy of Language
- **Savitt, Steven**
  - Being and Becoming in Modern Physics
- **Sayre-McCord, Geoff**
  - Metaethics
  - Moral Realism
- **Schabel, Christopher**
  - Francis of Marchia
  - Gregory of Rimini
- **Schaffer, Jonathan**
  - The Metaphysics of Causation
- **Schauer, Fred**
  - Rights
- **Scheuerman, William**
  - Globalization
- **Schmaltz, Tad**
  - Nicolas Malebranche
- **Schmidt, Heinz-Juergen**
  - Structuralism in Physics
- **Schulte, Oliver**
  - Formal Learning Theory
- **Schwitzgebel, Eric**
  - Belief
- **Seager, William**
  - Panpsychism
- **Sedley, David**
  - Lucretius
  - Plato on Naming and Knowledge
- **Senor, Tom**
  - Epistemological Problems of Memory

- **Shapiro, Stewart**
    - Classical Logic
- **Shaver, Robert**
    - Egoism
- **Sheehan, Thomas**
    - Martin Heidegger
- **Sheridan, Patricia**
    - Catharine Trotter Cockburn
- **Shields, Christopher**
    - Aristotle's Psychology
- **Shin, Sun-Joo**
    - Diagrams (with Oliver Lemon)
- **Shun, Kwong Loi**
    - Mencius
- **Siewert, Charles**
    - Consciousness and Intentionality
- **Silverman, Allan**
    - Plato's Metaphysics and Epistemology
- **Sinnott-Armstrong, Walter**
    - Consequentialism
    - Moral Skepticism
- **Sklar, Lawrence**
    - Philosophy of Statistical Mechanics
- **Sloan, Phillip**
    - Evolution
- **Slote, Michael**
    - Justice as a Virtue
- **Smart, J. J. C.**
    - The Identity Theory of Mind
- **Smith, Kurt**
    - Descartes' Life and Works
- **Smith, Barry**
    - Ontology and Information Science
- **Smith, Michael J.**
    - Collective Responsibility
- **Smith, Robin**
    - Ancient Logic
    - Aristotle's Logic
- **Smith, David Woodruff**
    - Phenomenology
- **Smith, Kelly**

- Epigenesis and Preformationism
- **Snyder, Laura J.**
  - William Whewell
- **Sorensen, Roy**
  - Vagueness
- **Sowaal, Alice**
  - Mary Astell
- **Spade, Paul Vincent**
  - *Binarium Famosissimum*
  - Insolubles
  - Medieval Theories of *Obligationes*
  - Medieval Philosophy
  - William of Ockham
- **Spruyt, Joke**
  - Peter of Spain
- **Stanton-Ife, John**
  - The Limits of Law
- **Stavropoulos, Nicos**
  - Interpretivist Theories of Law
- **Sterelny, Kim**
  - Biological Information (with Peter Godfrey-Smith)
- **Steup, Matthias**
  - The Analysis of Knowledge
- **Stoljar, Daniel**
  - The Deflationary Theory of Truth
  - Physicalism
- **Stone, Martin**
  - Legal Philosophy
- **Streveler, Paul**
  - Richard the Sophister
- **Stubenberg, Leopold**
  - Neutral Monism
- **Sullivan, David**
  - Rudolf Hermann Lotze
- **Sullivan, Shannon**
  - Feminist Approaches to the Intersection of Pragmatism and Continental Philosophy
- **Sutton, John**
  - Memory
- **Sweeney, Eileen**
  - Literary Forms of Medieval Philosophy
- **Sweet, William**

- Bernard Bosanquet
- British Idealism
- Jacques Maritain

- **Swoyer, Chris**
  - Properties
  - Relativism
- **Sypnowich, Christine**
  - Law and Ideology

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

T

- **Tabarroni, Andrea**
  - Medieval Theories of Fallacies
- **Talbott, William**
  - Bayesian Epistemology
- **Tanaka, Koji**
  - Paraconsistent Logic (with Graham Priest)
- **Tauber, Alfred**
  - The Biological Notion of Self and Non-self
- **Thagard, Paul**
  - Cognitive Science
- **Thielke, Peter**
  - Salomon Maimon (with Yitzhak Melamed)
- **Thijssen, Hans**
  - Condemnation of 1277
  - Nicholas of Autrecourt
- **Thomas, Nigel**
  - Mental Imagery
- **Thomasson, Amie**
  - Roman Ingarden
- **Thornton, Stephen**
  - Karl Popper
- **Tong, Rosemarie**
  - Feminist Ethics
- **Tooley, Michael**
  - The Problem of Evil

- **Torretti, Roberto**
  - Nineteenth Century Geometry
- **Trout, J.D.** ERROR Trout, J.D.
- **Tuana, Nancy**
  - Approaches to Feminism (with Sally Haslanger)
  - Feminist Interventions (with Sally Haslanger)
  - Feminist Perspectives on Sexuality
  - Topics in Feminism (with Sally Haslanger)
- **Tye, Michael**
  - Qualia
- **Tyler, Colin**
  - Thomas Hill Green

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

U

- **Uffink, Jos**
  - Boltzmann's Work in Statistical Physics
  - The Uncertainty Principle (with Jan Hilgevoord)
- **Uzgalis, William**
  - Anthony Collins
  - John Locke

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

V

- **Vaidman, Lev**
  - Many-Worlds Interpretation of Quantum Mechanics
- **Vailati, Ezio**
  - Samuel Clarke
- **Vallentyne, Peter**
  - Libertarianism
- **van Atten, Mark**
  - Luitzen Egbertus Jan Brouwer
- **Van Bendegem, Jean-Paul**

  - Finitism in Geometry
- **Van Gulick, Robert**
  - Consciousness
- **van Mill, David**
  - Freedom of Speech
- **van Roojen, Mark**
  - Moral Cognitivism vs. Non-Cognitivism
- **Vanderschraaf, Peter**
  - Common Knowledge
- **Varzi, Achille**
  - Boundary
  - Events (with Roberto Casati)
  - Holes (with Roberto Casati)
- **Verbeek, Bruno**
  - Game Theory and Ethics (with Christopher Morris)
- **Verbrugge, Rineke**
  - Provability Logic
- **Vihvelin, Kadri**
  - Arguments for Incompatibilism
- **Villa, Dana**
  - Hannah Arendt

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

W

- **Wade, Michael**
  - Evolutionary Genetics
- **Wainwright, William**
  - Jonathan Edwards
- **Waldron, Jeremy**
  - Property
- **Wallace, Jay**
  - Practical Reason
- **Wallis, Charles**
  - Causal Theories of Mental Content
- **Walsh, Alison**
  - Benjamin Peirce (with Ivor Grattan-Guinness)
- **Waluchow, Wil**
  - Constitutionalism

- **Warnke, Georgia**
  - Feminist Approaches to the Intersection of Analytic and Continental Philosophy
- **Warren, Karen**
  - Feminist Environmental Philosophy
- **Wartenberg, Thomas**
  - Philosophy of Film
- **Waters, Ken**
  - Genetics
- **Waters, Ken**
  - Molecular Genetics
- **Weatherson, Brian**
  - Intrinsic vs. Extrinsic Properties
  - The Problem of the Many
- **Weber, Bruce**
  - Life
- **Wedin, Michael**
  - Aristotle on Non-contradiction
- **Weinstein, David**
  - Herbert Spencer
- **Weinstein, Steven**
  - Quantum Gravity
- **Weinstock, Daniel**
  - Citizenship
- **Wertheimer, Alan**
  - Exploitation
- **West, Caroline**
  - Pornography and Censorship
- **Westerståhl, Dag**
  - Generalized Quantifiers
- **Wetherbee, Winthrop**
  - Dante Alighieri
- **Wetzel, Thomas**
  - State of Affairs
- **White, Stuart**
  - Social Minimum
- **White, Graham**
  - Medieval Theories of Causation
- **Whiting, Jennifer**
  - Personal Identity and Ethics
- **Wicks, Robert**
  - Arthur Schopenhauer
  - Friedrich Nietzsche

- **Wilce, Alexander**
  - Quantum Logic and Probability Theory
- **Wildberg, Christian**
  - David
  - Elias
  - Olympiodorus
  - Philoponus
  - Simplicius
  - Syrianus
- **Williams, Thomas**
  - John Duns Scotus
  - Saint Anselm
- **Williams, Michael**
  - Contextualist Theories of Epistemic Justification
- **Williams, Melissa**
  - Political Representation
- **Wilson, Jack**
  - The Biological Notion of Individual
- **Wilson, George**
  - Action
- **Wilson, Fred**
  - John Stuart Mill
- **Wimsatt, William**
  - Cultural Evolution
- **Wippel, John**
  - Godfrey of Fontaines
- **Witt, Charlotte**
  - Feminist History of Philosophy
- **Wolenski, Jan**
  - Lvov-Warsaw School
- **Wolff, Jonathan**
  - Justice and Bad Luck
  - Marxism
- **Wolff, Robert Paul**
  - Anarchism
- **Wong, David**
  - Comparative Philosophy: Chinese and Western
- **Wong, Hong Yu**
  - Emergent Properties (with Timothy O'Connor)
- **Woodruff, Paul**
  - Plato's Shorter Ethical Works
- **Woodward, James**

  - Causation and Manipulability
  - Scientific Explanation

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

X

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Y

- **Yaffe, Gideon**
  - Thomas Reid
- **Yalowitz, Steven**
  - Anomalous Monism
- **Young, James O.**
  - The Coherence Theory of Truth
- **Young, Robert**
  - Voluntary Euthanasia

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Z

- **Zach, Richard**
  - The Epsilon Calculus (with Jeremy Avigad)
  - Hilbert's Program
- **Zalta, Edward N.**
  - Frege's Logic, Theorem, and Foundations for Arithmetic
  - Gottlob Frege
- **Zangwill, Nick**
  - Aesthetics and Objectivity
- **Zank, Michael**
  - Martin Buber
- **Zimmerman, Michael J.**

    - Intrinsic vs. Extrinsic Value
- **Zuidervaart, Lambert**
    - Theodor Adorno
- **Zupko, Jack**
    - John Buridan
    - Thomas of Erfurt

Evolutionary Game Theory

Evolutionary game theory originated as an application of the mathematical theory of games to biological contexts, arising from the realization that frequency dependent fitness introduces a strategic aspect to evolution. Recently, however, evolutionary game theory has become of increased interest to economists, sociologists, and anthropologists--and social scientists in general--as well as philosophers. The interest among social scientists in a theory with explicit biological roots derives from three facts. First, the ‘evolution’ treated by evolutionary game theory need not be biological evolution. ‘Evolution’ may, in this context, often be understood as *cultural* evolution, where this refers to changes in beliefs and norms over time. Second, the rationality assumptions underlying evolutionary game theory are, in many cases, more appropriate for the modelling of social systems than those assumptions underlying the traditional theory of games. Third, evolutionary game theory, as an explicitly dynamic theory, provides an important element missing from the traditional theory. In the preface to *Evolution and the Theory of Games*, Maynard Smith notes that "[p]aradoxically, it has turned out that game theory is more readily applied to biology than to the field of economic behaviour for which it was originally designed." It is perhaps doubly paradoxical, then, that the subsequent development of *evolutionary* game theory has produced a theory which holds great promise for social scientists, and is as readily applied to the field of economic behaviour as that for which it was originally designed.

1. Historical Development

Evolutionary game theory was first developed by R. A. Fisher [see *The Genetic Theory of Natural Selection* (1930)] in his attempt to explain the approximate equality of the sex ratio in mammals. The puzzle Fisher faced

was this: why is it that the sex ratio is approximately equal in many species where the majority of males never mate? In these species, the non-mating males would seem to be excess baggage carried around by the rest of the population, having no real use. Fisher realized that if we measure individual fitness in terms of the expected number of *grandchildren*, then individual fitness depends on the distribution of males and females in the population. When there is a greater number of females in the population, males have a higher individual fitness; when there are more males in the population, females have a higher individual fitness. Fisher pointed out that, in such a situation, the evolutionary dynamics lead to the sex ratio becoming fixed at equal numbers of males and females. The fact that individual fitness depends upon the relative frequency of males and females in the population introduces a strategic element into evolutions.

Fisher's argument can be understood game theoretically, but he did not state it in those terms. In 1961, R. C. Lewontin made the first explicit application of game theory to evolutionary biology in "Evolution and the Theory of Games" (not to be confused with the Maynard Smith work of the same name). In 1972, Maynard Smith defined the concept of an *evolutionarily stable strategy* (hereafter ESS) in the article "Game Theory and the Evolution of Fighting." However, it was the publication of "The Logic of Animal Conflict," by Maynard Smith and Price in 1973 that introduced the concept of an ESS into widespread circulation. In 1982, Maynard Smith's seminal text *Evolution and the Theory of Games* appeared, followed shortly thereafter by Robert Axelrod's famous work *The Evolution of Cooperation* in 1984. Since then, there has been a veritable explosion of interest by economists and social scientists in evolutionary game theory (see the bibliography below).

2. Two Approaches to Evolutionary Game Theory

There are two approaches to evolutionary game theory. The first approach derives from the work of Maynard Smith and Price and employs the concept of an evolutionarily stable strategy as the principal tool of analysis. The second approach constructs an explicit model of the process by which the frequency of strategies change in the population and studies properties of the evolutionary dynamics within that model.

As an example of the first approach, consider the problem of the Hawk-Dove game, analyzed by Maynard Smith and Price in "The Logic of Animal Conflict." In this game, two individuals compete for a resource of a fixed value *V*. (In biological contexts, the value *V* of the resource corresponds to an increase in the Darwinian fitness of the individual who obtains the resource; in a cultural context, the value *V* of the resource would need to be given an alternate interpretation more appropriate to the specific model at hand.) Each individual follows exactly one of two strategies described below:

Hawk	Initiate aggressive behaviour, not stopping until injured or until one's opponent backs down.
Dove	Retreat immediately if one's opponent initiates aggressive behaviour.

If we assume that (1) whenever two individuals both initiate aggressive behaviour, conflict eventually results and the two individuals are equally likely to be injured, (2) the cost of the conflict reduces individual fitness by some constant value *C*, (3) when a Hawk meets a Dove, the Dove immediately retreats and the Hawk obtains the resource, and (4) when two Doves meet the resource is shared equally between them, the fitness payoffs for the Hawk-Dove game can be summarized according to the following matrix:

	Hawk	**Dove**

Hawk	½(V - C)	V
Dove	0	V/2

Figure 1: The Hawk-Dove Game

(The payoffs listed in the matrix are for that of a player *using* the strategy in the appropriate row, playing against someone using the strategy in the appropriate column. For example, if you play the strategy Hawk against an opponent who plays the strategy Dove, your payoff is V; if you play the strategy Dove against an opponent who plays the strategy Hawk, your payoff is 0.)

In order for a strategy to be evolutionarily stable, it must have the property that if almost every member of the population follows it, no mutant (that is, an individual who adopts a novel strategy) can successfully invade. This idea can be given a precise characterization as follows: Let $\Delta F(s_1,s_2)$ denote the change in fitness for an individual following strategy s_1 against an opponent following strategy s_2, and let $F(s)$ denote the total fitness of an individual following strategy s; furthermore, suppose that each individual in the population has an initial fitness of F_0. If σ is an evolutionarily stable strategy and μ a mutant attempting to invade the population, then

$$F(\sigma) = F_0 + (1-p)\Delta F(\sigma,\sigma) + p\Delta F(\sigma,\mu)$$

$$F(\mu) = F_0 + (1-p)\Delta F(\mu,\sigma) + p\Delta F(\mu,\mu)$$

where p is the proportion of the population following the mutant strategy μ.

Since σ is evolutionarily stable, the fitness of an individual following σ must be greater than the fitness of an individual following μ (otherwise the mutant following μ would be able to invade), and so $F(\sigma) > F(\mu)$. Now, as p is very close to 0, this requires that *either* that

$$\Delta F(\sigma,\sigma) > \Delta F(\sigma,\mu)$$

or that

$$\Delta F(\sigma,\sigma) = \Delta F(\sigma,\mu) \text{ and } \Delta F(\sigma,\mu) > \Delta F(\mu,\mu)$$

(This is the definition of an ESS that Maynard Smith and Price give.) In other words, what this means is that a strategy σ is an ESS if one of two conditions holds: (1) σ does better playing against σ than any mutant does playing against σ, or (2) some mutant does just as well playing against σ as σ, but σ does better playing against the mutant than the mutant does.

Given this characterization of an evolutionarily stable strategy, one can readily confirm that, for the Hawk-Dove game, the strategy Dove is not evolutionarily stable because a pure population of Doves can be invaded by a Hawk mutant. If the value V of the resource is greater than the cost C of injury (so that it is worth risking injury in order to obtain the resource), then the strategy Hawk is evolutionarily stable. In the case where the value of the resource is *less* than the cost of injury, there is no evolutionarily stable strategy if individuals are restricted to following pure strategies, although there is an evolutionarily stable strategy if players may use mixed strategies.[1]

As an example of the second approach, consider the well-known Prisoner's Dilemma. In this game, individuals choose one of two strategies, typically called "Cooperate" and "Defect." Here is the general form of the payoff matrix for the prisoner's dilemma:

	Cooperate	**Defect**
Cooperate	(R,R')	(S,T')
Defect	(T,S')	(P,P')

Figure 2: Payoff Matrix for the Prisoner's Dilemma.
Payoffs listed as (row, column).

where $T > R > P > S$ and $T' > R' > P' > S'$. (This form does not require that the payoffs for each player be symmetric, only that the proper ordering of the payoffs obtains.) In what follows, it will be assumed that the payoffs for the Prisoner's Dilemma are the same for everyone in the population.

How will a population of individuals that repeatedly plays the Prisoner's Dilemma evolve? We cannot answer that question without introducing a few assumptions concerning the nature of the population. First, let us assume that the population is quite large. In this case, we can represent the state of the population by simply keeping track of what proportion follow the strategies Cooperate and Defect. Let p_c and p_d denote these proportions. Furthermore, let us denote the average fitness of cooperators and defectors by W_C and W_D, respectively, and let $\overline{W}$ denote the average fitness of the entire population. The values of W_C, W_D, and $\overline{W}$ can be expressed in terms of the population proportions and payoff values as follows:

$$W_D = F_0 + p_c \triangle F(C,C) + p_d \triangle F(C,D)$$
$$W_D = F_0 + p_c \triangle F(D,C) + p_d \triangle F(D,D)$$
$$\overline{W} = p_c W_C + p_d W_D$$

Second, let us assume that the proportion of the population following the strategies Cooperate and Defect in the next generation is related to the proportion of the population following the strategies Cooperate and Defect in the current generation according to the rule:

$$p_c' = \frac{p_c W_C}{\overline{W}} \qquad p_d' = \frac{p_d W_D}{\overline{W}}$$

We can rewrite these expressions in the following form:

$$p_c' - p_c = \frac{p_c(W_C - \overline{W})}{\overline{W}} \qquad p_d' - p_d = \frac{p_d(W_D - \overline{W})}{\overline{W}}$$

If we assume that the change in the strategy frequency from one generation to the next are small, these difference

equations may be approximated by the differential equations:

$$\frac{dp_c}{dt} = \frac{p_c(W_C - \overline{W})}{\overline{W}} \qquad \frac{dp_d}{dt} = \frac{p_d(W_D - \overline{W})}{\overline{W}}$$

These equations were offered by Taylor and Jonker (1978) and Zeeman (1979) to provide continuous dynamics for evolutionary game theory and are known as the *replicator dynamics*.

The replicator dynamics may be used to model a population of individuals playing the Prisoner's Dilemma. For the Prisoner's Dilemma, the expected fitness of Cooperating and Defecting are:

$$\begin{aligned} W_C &= F_0 + p_c \triangle F(C,C) + p_d \triangle F(C,D) \\ &= F_0 + p_c R + p_d S \end{aligned}$$

and

$$\begin{aligned} W_D &= F_0 + p_c \triangle F(D,C) + p_d \triangle F(D,D) \\ &= F_0 + p_c T + p_d P. \end{aligned}$$

Since $T > R$ and $P > S$, it follows that $W_D > W_C$ and hence $W_D > \overline{W} > W_C$. This means that

$$\frac{W_D - \overline{W}}{\overline{W}} > 0$$

and

$$\frac{W_C - \overline{W}}{\overline{W}} < 0$$

Since the strategy frequencies for Defect and Cooperate in the next generation are given by

$$p'_d = p_d \cdot \frac{W_D - \overline{W}}{\overline{W}}$$

and

$$p'_c = p_c \cdot \frac{W_C - \overline{W}}{\overline{W}}$$

respectively, we see that over time the proportion of the population choosing the strategy Cooperate eventually becomes extinct. Figure 3 illustrates one way of representing the replicator dynamical model of the prisoner's

dilemma, known as a state-space diagram.

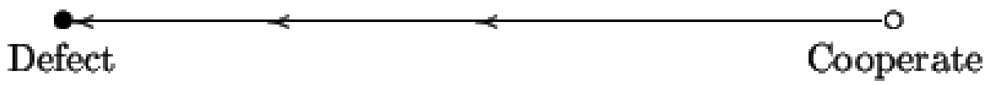


Figure 3: The Replicator Dynamical Model of the Prisoner's Dilemma

We interpret this diagram as follows: the leftmost point represents the state of the population where everyone defects, the rightmost point represents the state where everyone cooperates, and intermediate points represent states where some proportion of the population defects and the remainder cooperates. (One maps states of the population onto points in the diagram by mapping the state when *N*% of the population defects onto the point of the line *N*% of the way to the leftmost point.) Arrows on the line represent the evolutionary trajectory followed by the population over time. The open circle at the rightmost point indicates that the state where everybody cooperates is an unstable equilibrium, in the sense that if a small portion of the population deviates from the strategy Cooperate, then the evolutionary dynamics will drive the population away from that equilibrium. The solid circle at the leftmost point indicates that the state where everybody Defects is a stable equilibrium, in the sense that if a small portion of the population deviates from the strategy Defect, then the evolutionary dynamics will drive the population back to the original equilibrium state.

At this point, one may see little difference between the two approaches to evolutionary game theory. One can confirm that, for the Prisoner's Dilemma, the state where everybody defects is the only ESS. Since this state is the only stable equilibrium under the replicator dynamics, the two notions fit together quite neatly: the only stable equilibrium under the replicator dynamics occurs when everyone in the population follows the only ESS. In general, though, the relationship between ESSs and stable states of the replicator dynamics is more complex than this example suggests. Taylor and Jonker (1978), as well as Zeeman (1979), establish conditions under which one may infer the existence of a stable state under the replicator dynamics given an evolutionarily stable strategy. Roughly, if only two pure strategies exist, then given a (possibly mixed) evolutionarily stable strategy, the corresponding state of the population is a stable state under the replicator dynamics. (If the evolutionarily stable strategy is a mixed strategy *S*, the corresponding state of the population is the state in which the proportion of the population following the first strategy equals the probability assigned to the first strategy by *S*, and the remainder follow the second strategy.) However, this can fail to be true if more than two pure strategies exist.

The connection between ESSs and stable states under an evolutionary dynamical model is weakened further if we do not model the dynamics by the replicator dynamics. For example, suppose we use a local interaction model in which each individual plays the prisoner's dilemma with his or her neighbors. Nowak and May (1992, 1993), using a spatial model in which local interactions occur between individuals occupying neighboring nodes on a square lattice, show that stable population states for the prisoner's dilemma depend upon the specific form of the payoff matrix.[2]

When the payoff matrix for the population has the values $T = 2.8$, $R = 1.1$, $P = 0.1$, and $S = 0$, the evolutionary dynamics of the local interaction model agree with those of the replicator dynamics, and lead to a state where each individual follows the strategy Defect--which is, as noted before, the only evolutionarily stable strategy in the prisoner's dilemma. The figure below illustrates how rapidly one such population converges to a state where everyone defects.

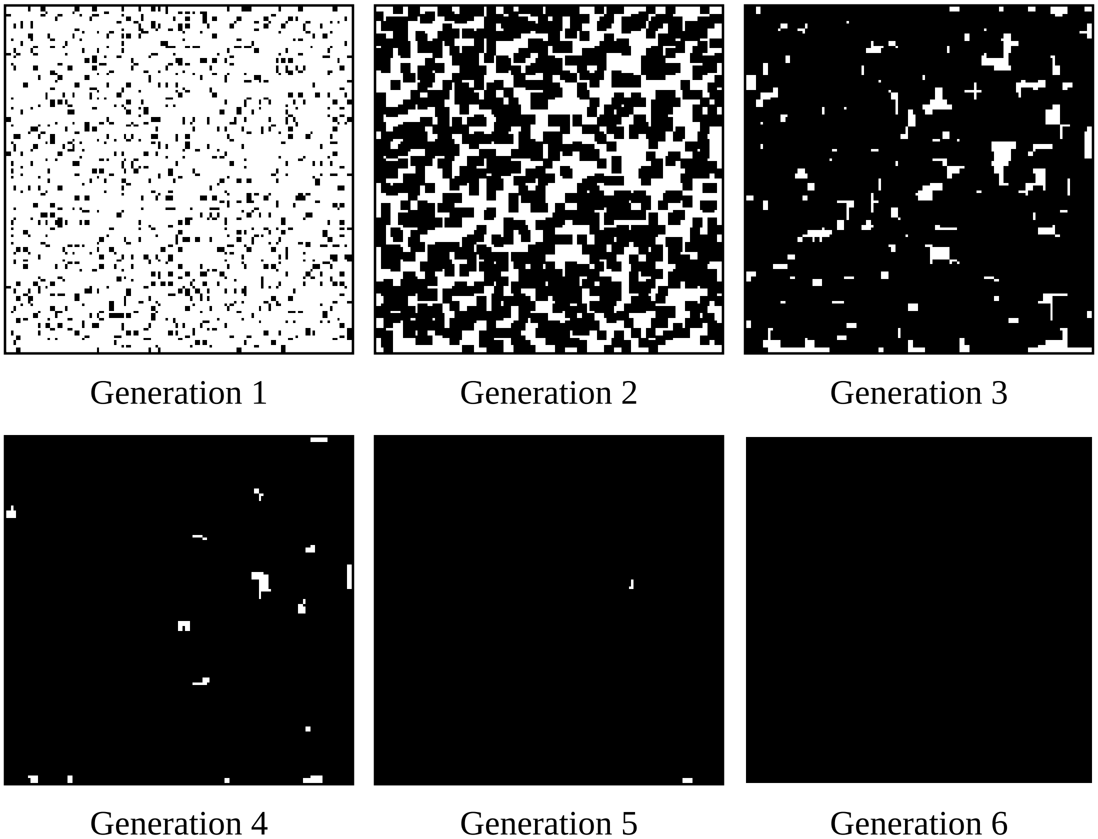


Figure 4: Prisoner's Dilemma: All Defect
[view a movie of this model]

However, when the payoff matrix has values of T = 1.2, R = 1.1, P = 0.1, and S = 0, the evolutionary dynamics carry the population to a stable cycle oscillating between two states. In this cycle cooperators and defectors coexist, with some regions containing "blinkers" oscillating between defectors and cooperators (as seen in generation 19 and 20).

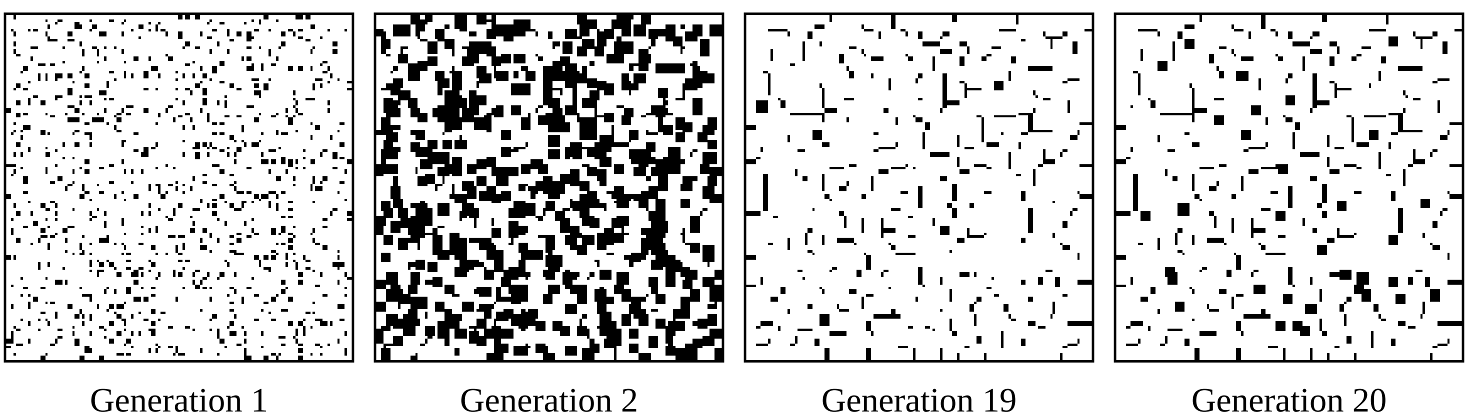


Figure 5: Prisoner's Dilemma: Cooperate
[view a movie of this model]

Notice that with these particular settings of payoff values, the evolutionary dynamics of the local interaction model differ significantly from those of the replicator dynamics. Under these payoffs, the stable states have no

corresponding analogue in either the replicator dynamics nor in the analysis of evolutionarily stable strategies.

A phenomenon of greater interest occurs when we choose payoff values of $T = 1.61$, $R = 1.01$, $P = 0.01$, and $S = 0$. Here, the dynamics of local interaction lead to a world constantly in flux: under these values regions occupied predominantly by Cooperators may be successfully invaded by Defectors, and regions occupied predominantly by Defectors may be successfully invaded by Cooperators. In this model, there is no "stable strategy" in the traditional dynamical sense.[3]

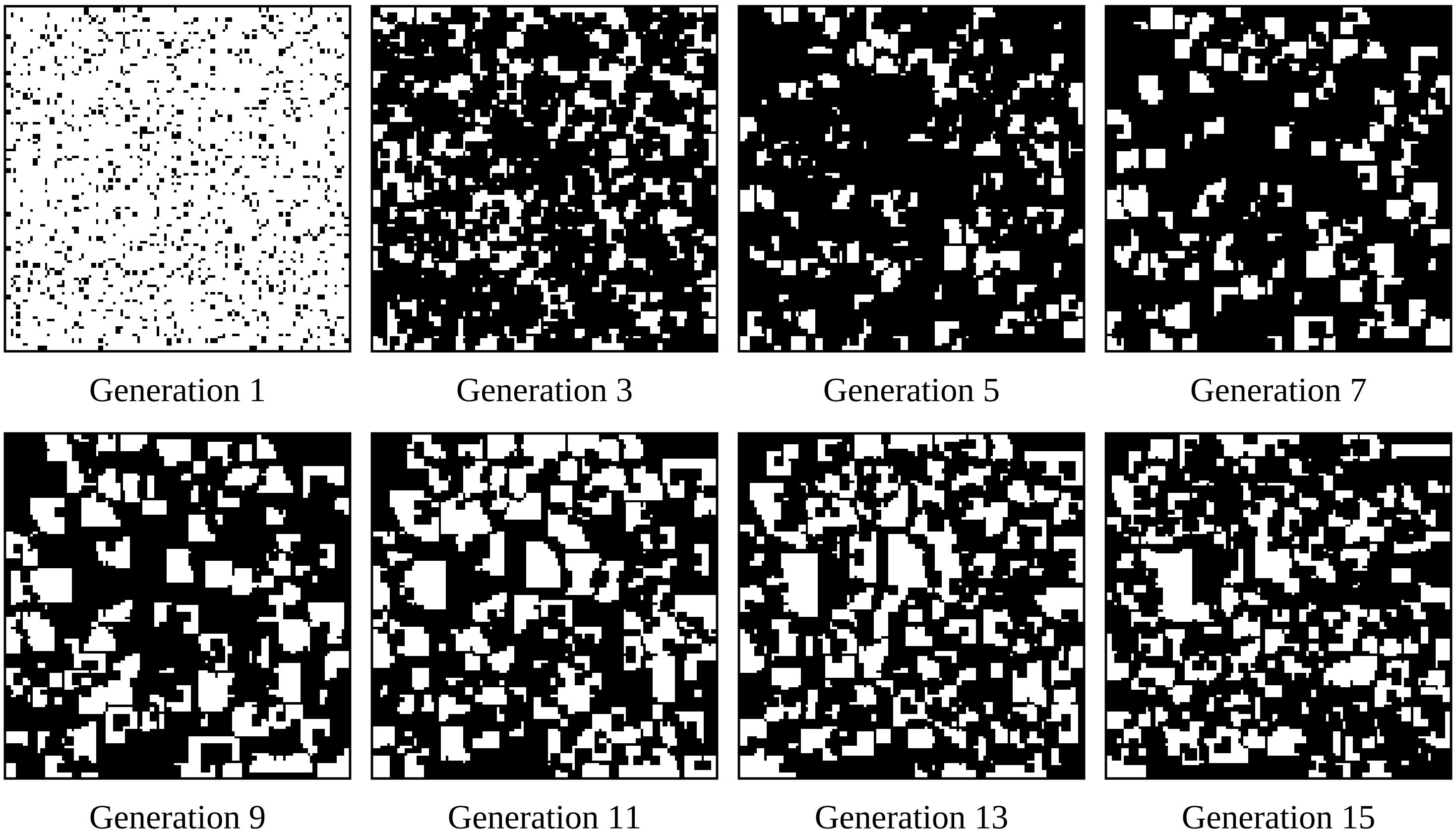


Figure 6: Prisoner's Dilemma: Chaotic
[view a movie of this model]

These models demonstrate that, although numerous cases exist in which both approaches to evolutionary game theory arrive at the same conclusion regarding which strategies one would expect to find present in a population, there are enough differences in the outcomes of the two modes of analysis to justify the development of each program.

3. Why Evolutionary Game Theory?

Although evolutionary game theory has provided numerous insights to particular evolutionary questions, a growing number of social scientists have become interested in evolutionary game theory in hopes that it will provide tools for addressing a number of deficiencies in the traditional theory of games, three of which are discussed below.

3.1 The equilibrium selection problem

The concept of a Nash equilibrium (see the entry on game theory) has been the most used solution concept in game theory since its introduction by John Nash in 1950. A selection of strategies by a group of agents is said to be in a Nash equilibrium if each agent's strategy is a best-response to the strategies chosen by the other players. By best-response, we mean that no individual can improve her payoff by switching strategies unless at least one other individual switches strategies as well. This need not mean that the payoffs to each individual are optimal in a Nash equilibrium: indeed, one of the disturbing facts of the prisoner's dilemma is that the only Nash equilbrium of the game--when both agents defect--is suboptimal.[4]

Yet a difficulty arises with the use of Nash equilibrium as a solution concept for games: if we restrict players to using pure strategies, not every game has a Nash equilbrium. The game "Matching Pennies" illustrates this problem.

	Heads	**Tails**
Heads	(0,1)	(1,0)
Tails	(1,0)	(0,1)

Figure 7: Payoff matrix for the game of Matching Pennies
(Row wins if the two coins do not match, whereas Column wins if the two coins match).

While it is true that every noncooperative game in which players may use mixed strategies has a Nash equilibrium, some have questioned the significance of this for real agents. If it seems appropriate to require rational agents to adopt only pure strategies (perhaps because the cost of implementing a mixed strategy runs too high), then the game theorist must admit that certain games lack solutions.

A more significant problem with invoking the Nash equilibrium as the appropriate solution concept arises because games exist which have multiple Nash equilibria (see Section 2.5, Solution Concepts and Equilibria, in the entry on game theory). When there are several different Nash equilibria, how is a rational agent to decide which of the several equilibria is the "right one" to settle upon?[5] Attempts to resolve this problem have produced a number of possible refinements to the concept of a Nash equilibrium, each refinement having some intuitive purchase. Unfortunately, so many refinements of the notion of a Nash equilibrium have been developed that, in many games which have multiple Nash equilibria, each equilibrium could be justified by some refinement present in the literature. The problem has thus shifted from choosing among multiple Nash equilibria to choosing among the various refinements. Some (see Samuelson (1997), *Evolutionary Games and Equilibrium Selection*) hope that further development of evolutionary game theory can be of service in addressing this issue.

3.2 The problem of hyperrational agents

The traditional theory of games imposes a very high rationality requirement upon agents. This requirement originates in the development of the theory of utility which provides game theory's underpinnings (see Luce (1950) for an introduction). For example, in order to be able to assign a cardinal utility function to individual agents, one typically assumes that each agent has a well-defined, consistent set of preferences over the set of "lotteries" over the outcomes which may result from individual choice. Since the number of different lotteries over outcomes is uncountably infinite, this requires each agent to have a well-defined, consistent set of uncountably infinitely many

preferences.

Numerous results from experimental economics have shown that these strong rationality assumptions do not describe the behavior of real human subjects. Humans are rarely (if ever) the hyperrational agents described by traditional game theory. For example, it is not uncommon for people, in experimental situations, to indicate that they prefer *A* to *B*, *B* to *C*, and *C* to *A*. These "failures of the transitivity of preference" would not occur if people had a well-defined consistent set of preferences. Furthermore, experiments with a class of games known as a "beauty pageant" show, quite dramatically, the failure of common knowledge assumptions typically invoked to solve games.[6] Since evolutionary game theory successfully explains the predominance of certain behaviors of insects and animals, where strong rationality assumptions clearly fail, this suggests that rationality is not as central to game theoretic analyses as previously thought. The hope, then, is that evolutionary game theory may meet with greater success in describing and predicting the choices of human subjects, since it is better equipped to handle the appropriate weaker rationality assumptions.

3.3 The lack of a dynamical theory in the traditional theory of games

At the end of the first chapter of *Theory of Games and Economic Behavior,* von Neumann and Morgenstern write:

> We repeat most emphatically that our theory is thoroughly static. A dynamic theory would unquestionably be more complete and therefore preferable. But there is ample evidence from other branches of science that it is futile to try to build one as long as the static side is not thoroughly understood. (Von Neumann and Morgenstern, 1953, p. 44)

The theory of evolution is a dynamical theory, and the second approach to evolutionary game theory sketched above explicitly models the dynamics present in interactions among individuals in the population. Since the traditional theory of games lacks an explicit treatment of the dynamics of rational deliberation, evolutionary game theory can be seen, in part, as filling an important lacuna of traditional game theory.

One may seek to capture some of the dynamics of the decision-making process in traditional game theory by modeling the game in its extensive form, rather than its normal form. However, for most games of reasonable complexity (and hence interest), the extensive form of the game quickly becomes unmanageable. Moreover, even in the extensive form of a game, traditional game theory represents an individual's strategy as a specification of what choice that individual would make at each information set in the game. A selection of strategy, then, corresponds to a selection, prior to game play, of what that individual will do at any possible stage of the game. This representation of strategy selection clearly presupposes hyperrational players and fails to represent the process by which one player observes his opponent's behavior, learns from these observations, and makes the best move in response to what he has learned (as one might expect, for there is no need to model learning in hyperrational individuals). The inability to model the dynamical element of game play in traditional game theory, and the extent to which evolutionary game theory naturally incorporates dynamical considerations, reveals an important virtue of evolutionary game theory.

4. Philosophical Problems of Evolutionary Game Theory

The growing interest among social scientists and philosophers in evolutionary game theory has raised several philosophical questions, primarily stemming from its application to human subjects.

4.1 The meaning of fitness in cultural evolutionary interpretations

As noted previously, evolutionary game theoretic models may often be given both a biological and a cultural evolutionary interpretation. In the biological interpretation, the numeric quantities which play a role analogous to "utility" in traditional game theory correspond to the fitness (typically Darwinian fitness) of individuals.[7] How does one interpret "fitness" in the cultural evolutionary interpretation?

In many cases, fitness in cultural evolutionary interpretations of evolutionary game theoretic models directly measures some objective quantity of which it can be safely assumed that (1) individuals always want more rather than less and (2) interpersonal comparisons are meaningful. Depending on the particular problem modeled, money, slices of cake, or amount of land would be appropriate cultural evolutionary interpretations of fitness. Requiring that fitness in cultural evolutionary game theoretic models conform to this interpretative constraint severely limits the kinds of problems that one can address. A more useful cultural evolutionary framework would provide a more general theory which did not require that individual fitness be a linear (or strictly increasing) function of the amount of some real quantity, like amount of food.

In traditional game theory, a strategy's fitness was measured by the expected utility it had for the individual in question. Yet evolutionary game theory seeks to describe individuals of limited rationality (commonly known as "boundedly rational" individuals), and the utility theory employed in traditional game theory assumes highly rational individuals. Consequently, the utility theory used in traditional game theory cannot simply be carried over to evolutionary game theory. One must develop an alternate theory of utility/fitness, one compatible with the bounded rationality of individuals, that is sufficient to define a utility measure adequate for the application of evolutionary game theory to cultural evolution.

4.2 The explanatory irrelevance of evolutionary game theory

Another question facing evolutionary game theoretic explanations of social phenomena concerns the kind of explanation it seeks to give. Depending on the type of explanation it seeks to provide, are evolutionary game theoretic explanations of social phenomena irrelevant or mere vehicles for the promulgation of pre-existing values and biases? To understand this question, recognize that one must ask whether evolutionary game theoretic explanations target the etiology of the phenomenon in question, the persistence of the phenomenon, or various aspects of the normativity attached to the phenomenon. The latter two questions seem deeply connected, for population members typically enforce social behaviors and rules having normative force by sanctions placed on those failing to comply with the relevant norm; and the presence of sanctions, if suitably strong, explains the persistence of the norm. The question regarding a phenomenon's etiology, on the other hand, can be considered independent of the latter questions.

If one wishes to explain how some currently existing social phenomenon came to be, it is unclear why approaching it from the point of view of evolutionary game theory would be particularily illuminating. The etiology of any phenomenon is a unique historical event and, as such, can only be discovered empirically, relying on the work of sociologists, anthropologists, archaeologists, and the like. Although an evolutionary game theoretic model may exclude certain historical sequences as possible histories (since one may be able to show that the cultural evolutionary dynamics preclude one sequence from generating the phenomenon in question), it seems unlikely that an evolutionary game theoretic model would indicate a unique historical sequence suffices to bring about the phenomenon. An empirical inquiry would then still need to be conducted to rule out the extraneous historical

sequences admitted by the model, which raises the question of what, if anything, was gained by the construction of an evolutionary game theoretic model in the intermediate stage. Moreover, even if an evolutionary game theoretic model indicated that a single historical sequence was capable of producing a given social phenomenon, there remains the important question of why we ought to take this result seriously. One may point out that since nearly any result can be produced by a model by suitable adjusting of the dynamics and initial conditions, all that the evolutionary game theorist has done is provide one such model. Additional work needs to be done to show that the underlying assumptions of the model (both the cultural evolutionary dynamics and the initial conditions) are empirically supported. Again, one may wonder what has been gained by the evolutionary model--would it not have been just as easy to determine the cultural dynamics and initial conditions beforehand, constructing the model afterwards? If so, it would seem that the contributions made by evolutionary game theory in this context simply are a proper part of the parent social science--sociology, anthropology, economics, and so on. If so, then there is nothing *particular* about evolutionary game theory employed in the explanation, and this means that, contrary to appearances, evolutionary game theory is really irrelevant to the given explanation.

If evolutionary game theoretic models do not explain the etiology of a social phenomenon, presumably they explain the persistence of the phenomenon or the normativity attached to it. Yet we rarely need an evolutionary game theoretic model to identify a particular social phenomenon as stable or persistent as that can be done by observation of present conditions and examination of the historical records; hence the charge of irrelevancy is raised again. Moreover, most of the evolutionary game theoretic models developed to date have provided the crudest approximations of the real cultural dynamics driving the social phenomenon in question. One may well wonder why, in these cases, we should take seriously the stability analysis given by the model; answering this question would require one engage in an empirical study as previously discussed, ultimately leading to the charge of irrelevance again.

4.3 The value-ladenness of evolutionary game theoretic explanations

If one seeks to use an evolutionary game theoretic model to explain the normativity attached to a social rule, one must explain how such an approach avoids committing the so-called "naturalistic fallacy" of inferring an ought-statement from a conjunction of is-statements.[8] Assuming that the explanation does not commit such a fallacy, one argument charges that it must then be the case that the evolutionary game theoretic explanation merely repackages certain key value claims tacitly assumed in the construction of the model. After all, since any argument whose conclusion is a normative statement must have at least one normative statement in the premises, any evolutionary game theoretic argument purporting to show how certain norms acquire normative force must contain--at least implicitly--a normative statement in the premises. Consequently, this application of evolutionary game theory does not provide a neutral analysis of the norm in question, but merely acts as a vehicle for advancing particular values, namely those smuggled in the premises.

This criticism seems less serious than the charge of irrelevancy. Cultural evolutionary game theoretic explanations of norms need not "smuggle in" normative claims in order to draw normative conclusions. The theory already contains, in its core, a proper subtheory having normative content--namely a theory of rational choice in which boundedly rational agents act in order to maximize, as best as they can, their own self-interest. One may challenge the suitability of this as a foundation for the normative content of certain claims, but this is a different criticism from the above charge. Although cultural evolutionary game theoretic models do act as vehicles for promulgating certain values, they wear those minimal value commitments on their sleeve. Evolutionary explanations of social norms have the virtue of making their value commitments explicit and also of showing how other normative commitments (such as fair division in certain bargaining situations, or cooperation in the prisoner's dilemma) may

be derived from the principled action of boundedly rational, self-interested agents.

Bibliography

The following bibliography, although it tries to be comprehensive, is by no means complete. If you are aware of articles, books, monographs, etc. which you believe should be included, but are not, please notify the author.

- Ackley, David and Michael Littman (1994) "Interactions Between Learning and Evolution," in Christopher G. Langton, ed., *Artificial Life III*. Addison-Wesley, pp. 487-509.
- Adachi, N. and Matsuo, K. (1991) "Ecological Dynamics Under Different Selection Rules in Distributed and Iterated Prisoner's Dilemma Games," *Parallel Problem Solving From Nature*, Lecture Notes in Computer Science Volume 496 (Berlin: Springer-Verlag), pp. 388-394.
- Alexander, J. McKenzie (2000) "Evolutionary Explanations of Distributive Justice," *Philosophy of Science* 67:490-516.
- Alexander, Jason and Brian Skyrms (1999) "Bargaining with Neighbors: Is Justice Contagious?" *Journal of Philosophy* 96, 11: 588-598.
- Axelrod, R. (1984) *The Evolution of Cooperation*. New York: Basic Books.
- Axelrod, Robert (1986) "An evolutionary approach to norms," *American Political Science Review* 80, 4: 1095-1111.
- Axelrod, Robert M. and Dion, Douglas (1988) 'The Further Evolution of Cooperation', *Science*, **242**(4884), 9 December, pp. 1385-1390.
- Axelrod, Robert M. and Hamilton, William D. (1981) 'The Evolution of Cooperation', *Science*, **211**(4489), pp. 1390-1396.
- Banerjee, Abhijit V. and Weibull, Jo:rgen W. (1993) "Evolutionary Selection with Discriminating Players," Research Paper in Economics, University of Stockholm.
- Bergin, J. and Lipman, B. (1996) "Evolution with State-Dependent Mutations," *Econometrica*, **64**, pp. 943-956.
- Binmore, Kenneth G. and Larry Samuelson (1994) "An Economist's Perspective on the Evolution of Norms," *Journal of Institutional and Theoretical Economics* 150, 1: 45-63.
- Binmore, Ken and Samuelson, Larry (1991) "Evolutionary Stability in Repeated Games Played By Finite Automata," *Journal of Economic Theory*, **57**, pp. 278-305.
- Binmore, Ken and Samuelson, Larry (1994) "An Economic Perspective on the Evolution of Norms," *Journal of Institutional and Theoretical Economics*, **150**(1), pp. 45-63.
- Björnerstedt, J. and Weibull, J. (1993) "Nash Equilibrium and Evolution by Imitation," in Arrow, K. and Colombatto, E. (eds.) *Rationality in Economics* (New York, NY: Macmillan).
- Blume, L. (1993) "The Statistical Mechanics of Strategic Interaction," *Games and Economic Behaviour*, **5**, pp. 387-424.
- Blume, Lawrence E. (1997) "Population Games," in W. Brian Arthur, Steven N. Durlauf, and David A. Lane, eds., *The Economy as an Evolving Complex System II*, Addison-Wesley, volume 27 of *SFI Studies in the Sciences of Complexity*, pp. 425-460.
- Bögers, Tilman and Sarin, R. (1993) "Learning Through Reinforcement and Replicator Dynamics," Technical Report, University College London.
- Bögers, Tilman and Sarin, R. (1996a) "Naive Reinforcement and Replicator Dynamics," ELSE Working Paper.
- Bögers, Tilman and Sarin, R. (1996b) "Learning Through Reinforcement and Replicator Dynamics," ELSE Working Paper.

- Boyd, Robert and Lorberbaum, Jeffrey P. (1987) "No Pure Strategy is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game," *Nature*, **327**, 7 May, pp. 58-59.
- Boylan, Richard T. (1991) "Laws of Large Numbers for Dynamical Systems with Randomly Matched Individuals," *Journal of Economic Theory*, **57**, pp. 473-504.
- Busch, Marc L. and Reinhardt, Eric R. (1993) "Nice Strategies in a World of Relative Gains: The Problem of Co-operation under Anarchy," *Journal-of-Conflict-Resolution*, **37**(3), September, pp. 427-445.
- Cabrales, A. and Ponti, G. (1996) "Implementation, Elimination of Weakly Dominated Strategies and Evolutionary Dynamics," ELSE Working Paper.
- Canning, David (1988) "Rationality and Game Theory When Players are Turing Machines," ST/ICERD Discussion Paper 88/183, London School of Economics, London.
- Canning, David (1990c) "Rationality, Computability and the Limits of Game Theory," Economic Theory Discussion Paper Number 152, Department of Applied Economics, University of Cambridge, July.
- Canning, David (1992) "Rationality, Computability and Nash Equilibrium," *Econometrica*, **60**(4), July, pp. 877-888.
- Cho, I.-K. and Kreps, David M. (1987) "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, **102**(1), February, pp. 179-221.
- Cowan, Robin A. and Miller, John H. (1990) "Economic Life on a Lattice: Some Game Theoretic Results," Working Paper 90-010, Economics Research Program, Santa Fe Institute, New Mexico.
- D'Arms, Justin, Robert Batterman, and Krzyzstof Górny (1998) "Game Theoretic Explanations and the Evolution of Justice," *Philosophy of Science* 65: 76-102.
- D'Arms, Justin (1996) "Sex, Fairness, and the Theory of Games," *Journal of Philosophy* 93, 12: 615-627.
- ----- (2000) "When Evolutionary Game Theory Explains Morality, What Does It Explain?" *Journal of Consciousness Studies* 7, 1-2: 296-299.
- Danielson, P. (1992) *Artificial Morality: Virtuous Robots for Virtual Games* (Routledge).
- Danielson, Peter (1998) "Critical Notice: *Evolution of the Social Contract," Canadian Journal of Philosophy* 28, 4: 627-652.
- Dekel, Eddie and Scotchmer, Suzanne (1992) "On the Evolution of Optimizing Behavior," *Journal of Economic Theory*, **57**, pp. 392-406.
- Eaton, B. C. and Slade, M. E. (1989) "Evolutionary Equilibrium in Market Supergames," Discussion Paper, University of British Columbia, November.
- Ellingsen, Tore (1997) "The Evolution of Bargaining Behavior," *The Quarterly Journal of Economics* pp. 581-602.
- Ellison, G. (1993) "Learning, Local Interaction and Coordination," *Econometrica* 61: 1047-1071.
- Epstein, Joshua A. (1998) "Zones of Cooperation in Demographic Prisoner's Dilemma," *Complexity* 4, 2: 36-48.
- Eshel, Ilan, Larry Samuelson, and Avner Shaked (1998) "Altruists, Egoists, and Hooligans in a Local Interaction Model," *The American Economic Review* 88, 1: 157-179.
- Fisher, R. A. (1930) *The Genetic Theory of Natural Selection*, Oxford, Clarendon Press.
- Fogel, David B. (1993) "Evolving Behaviours in the Iterated Prisoner's Dilemma," *Evolutionary Computation*, **1**(1), April, pp. 77-97.
- Forrest, Stephanie and Mayer-Kress, G. (1991) "Genetic Algorithms, Nonlinear Dynamical Systems, and Global Stability Models," in Davis, L. (ed.) *The Handbook of Genetic Algorithms* (New York, NY: Van Nostrand Reinhold).
- Foster, Dean and Young, H. Peyton (1990) "Stochastic Evolutionary Game Dynamics," *Journal of Theoretical Biology*, **38**, pp. 219-232.
- Friedman, Daniel (1991) "Evolutionary Games in Economics," *Econometrica*, **59**(3), May, pp. 637-666.
- Fudenberg, Drew and Maskin, Eric (1990) "Evolution and Cooperation in Noisy Repeated Games,"

American Economic Review (Papers and Proceedings), **80**(2), May, pp. 274-279.

- Gintis, Herbert (2000) "Classical Versus Evolutionary Game Theory," *Journal of Consciousness Studies* 7, 1-2: 300-304.
- Guth, Werner and Kliemt, Hartmut (1994) "Competition or Co-operation - On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes," *Metroeconomica*, **45**, pp. 155-187.
- Guth, Werner and Kliemt, Hartmut (1998) "The Indirect Evolutionary Approach: Bridging the Gap Between Rationality and Adaptation," *Rationality and Society*, **10**(3), pp. 377 - 399.
- Hamilton, W. D. (1963) "The Evolution of Altruistic Behavior," *The American Naturalist* 97: 354-356.] - (1964) " The Genetical Evolution of Social Behavior. I," *J. Theoret. Biol.* 7: 1-16.
- ----- (1964) "The Genetical Evolution of Social Behavior. II," *J. Theoret. Biol.* 7: 17-52.
- Hammerstein, P. and Selten, R. (1994) "Game Theory and Evolutionary Biology," in Auman, R. and Hart, S. (eds.) *Handbook of Game Theory with Economic Applications* (Elsevier Science), volume 2, pp. 931-962.
- Hansen, R. G. and Samuelson, W. F. (1988) "Evolution in Economic Games," *Journal of Economic Behavior and Organization*, **10**(3), October, pp. 315-338.
- Harms, William (1997) "Evolution and Ultimatum Bargaining," *Theory and Decision* 42: 147-175.
- ----- (2000) "The Evolution of Cooperation in Hostile Environments," *Journal of Consciousness Studies* 7, 1-2: 308-313.
- Harrald, Paul G. (in press) "Evolving Behaviour in Repeated Games via Genetic Algorithms," in Stampoultzsis, P. (ed.) *The Applications Handbook of Genetic Algorithms* (Boca Raton, FA: CRC Publishers). Hassell, Michael P., Hugh N. Comins, and Robert M. May (1991) "Spatial structure and chaos in insect population dynamics," *Nature* 353: 255-258.
- Hegselmann, Rainer (1996) "Social Dilemmas in Lineland and Flatland," in Liebrand and Messick, eds., *Frontiers in Social Dilemmas Research*, Springer, pp. 337-361.
- Hiebeler, David (1997) "Stochastic Spatial Models: From Simulations to Mean Field and Local Structure Approximations," *Journal of Theoretical Biology* 187: 307-319.
- Hines, W. G. (1987) "Evolutionary Stable Strategies: A Review of Basic Theory," *Theoretical Population Biology*, **31**, pp. 195-272.
- Hirshleifer, Jack and Martinez-Coll, Juan Carlos (1988) "What Strategies can Support the Evolutionary Emergence of Cooperation?," *Journal of Conflict Resolution*, **32**(2), June, pp. 367-398.
- Hirshleifer, Jack and Marti/nez-Coll, Juan Carlos (1992) "Selection, Mutation and the Preservation of Diversity in Evolutionary Games," Papers on Economics and Evolution, Number 9202, edited by the European Study Group for Evolutionary Economics.
- Howard, J. V. (1988) "Cooperation in the Prisoner's Dilemma," *Theory and Decision*, **24**, pp. 203-213.
- Huberman, Bernardo A. and Glance, Natalie S. (1993) "Evolutionary Games and Computer Simulations," *Proceedings of the National Academy of Sciences of the USA*, **90**(16), August, pp. 7716-7718.
- Ikegami, Takashi (1993) "Ecology of Evolutionary Game Strategies," in [ECAL 93], pp. 527-536.
- Kandori, Michihiro, Mailath, George J. and Rob, Rafael (1993) "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica*, **61**, pp. 29-56.
- Kreps, David M. (1990) *Game Theory and Economic Modelling* (Oxford: Clarendon Press).
- Kreps, David M. and Fudenberg, Drew (1988) *Learning, Experimentation, and Equilibrium in Games* (Cambridge, MA: MIT Press).
- Iwasa, Yoh, Mayuko Nakamaru, and Simon A. Levin (1998) "Allelopathy of bacteria in a lattice population: Competition between colicin-sensitive and colicin-producing strains," *Evolutionary Ecology* 12: 785-802.
- Kandori, Michihiro, George J. Mailath, and Rafael Rob (1993) "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica* 61, 1: 29-56.

- Kaneko, Kunihiko and Junji Suzuki (1994) "Evolution to the Edge of Chaos in an Imitation Game," in Christopher G. Langton, ed., *Artificial Life III*. Addison-Wesley, pp. 43-53.
- Kephart, Jeffrey O. (1994) "How Topology Affects Population Dynamics," in Christopher G. Langton, ed., *Artificial Life III*. Addison-Wesley, SFI Studies in the Sciences of Complexity, pp. 447-463.
- Kitcher, Philip (1999) "Games Social Animals Play: Commentary on Brian Skyrms' *Evolution of the Social Contract*," *Philosophy and Phenomenological Research* 59, 1: 221-228.
- Krebs, Dennis (2000) "Evolutionary Games and Morality," *Journal of Consciousness Studies* 7, 1-2: 313-321.
- Levin, B. R. (1988) "Frequency-dependent selection in bacterial populations," *Philosophical Transactions of the Royal Society of London* B, 319: 469-472.
- Lewontin, R. C. (1961) "Evolution and the Theory of Games" *J. Theor. Biol.* 1:382-403.
- Liebrand, Wim B. G. and Messick, David M. (eds.) (1996) *Frontiers in Social Dilemmas Research* (Berlin: Springer-Verlag).
- Lindgren, Kristian (1990) "Evolution in a Population of Mutating Strategies," NORDITA Preprint 90/22 S, Copenhagen.
- Lindgren, Kristian and Nordahl, Mats G. (1993) "Evolutionary Dynamics of Spatial Games," in *Self Organization and Life: From Simple Rules to Global Complexity, Proceedings of the Second European Conference on Artificial Life, Brussels, Belgium 24-26 May 1993* (Cambridge, MA: MIT Press), pp. 604-616.
- Lindgren, Kristian and Mats G. Nordahl (1994) "Evolutionary dynamics of spatial games," *Physica D* 75: 292-309.
- Lindgren, K. (1991) "Evolutionary phenomena in simple dynamics," in C.G. Langton, J.D. Farmer, S. Rasmussen, and C. Taylor, eds., *Artificial Life II*, Redwood City, CA: Addison-Wesley, pp. 295-312.
- Lomborg, Bjorn (1992) "Cooperation in the Iterated Prisoner's Dilemma," Papers on Economics and Evolution, Number 9302, edited by the European Study Group for Evolutionary Economics.
- Lomborg, Bjorn (1996) "Nucleus and Shield: The Evolution of Social Structure in the Interated Prisoner's Dilemma," *American Sociological Review*, **61**(xx), April, pp. 278-307.
- Macy, Michael (1989) "Walking Out of Social Traps: A Stochastic Learning Model for the Prisoner's Dilemma," *Rationality and Society*, **1**(2), pp. 197-219.
- Mailath, George J. (1992) "Introduction: Symposium on Evolutionary Game Theory," *Journal of Economic Theory*, **57**, pp. 259-277.
- Mailath, George J., Samuelson, Larry and Shaked, Avner (1992) "Evolution and Endogenous Interaction," Draft Paper, Department of Economics, University of Pennsylvania, latest version 24 August 1995.
- Matsui, Akihiko (1993) "Evolution and Rationalizability," Working Paper: 93-19, Center for Analytic Research in Economics and the Social Sciences (CARESS), University of Pennsylvania, May.
- Mar, Gary (2000) "Evolutionary Game Theory, Morality, and Darwinism" *Journal of Consciousness Studies* 7, 1-2: 322-326.
- May, R. M., Bohoeffer, S. and Nowak, Martin A. (1995) "Spatial Games and the Evolution of Cooperation," in Mora/n, F., Moreno, A., Morelo, J. J. and Chaco/n, P. (eds.) *Advances in Artificial Life: Proceedings of the Third European Conference on Artificial Life (ECAL95)* (Berlin: Sprnger-Verlag), pp. 749-759.
- Maynard-Smith, John (1976) "Evolution and the Theory of Games," *American Scientist*, **64**(1), January, pp. 41-45.
- Maynard-Smith, John (1982) *Evolution and the Theory of Games* (Cambridge: Cambridge University Press).
- Maynard Smith, John and George Price (1973) "The Logic of Animal Conflict" *Nature*:146, pp. 15-18.
- Miller, John H. (1988) "The Evolution of Automata in the Repeated Prisoner's Dilemma," in *Two Essays on*

the Economics of Imperfect Information, Doctoral Dissertation, Department of Economics, University of Michigan (Ann Arbor).

- Miller, John H. (1989) "The Coevolution of Automata in the Repeated Prisoner's Dilemma," Working Paper 89-003, Santa Fe Institute, New Mexico.
- Miller, John H. (1996) "The Coevolution of Automata in the Repeated Prisoner's Dilemma," *Journal of Economic Behavior and Organization*, **29**(1), January, pp. 87-112.
- Miller, John H. and Shubik, Martin (1992) "Some Dynamics of a Strategic Market Game with a Large Number of Agents," Working Paper 92-11-057, Santa Fe Institute, New Mexico.
- Miller, John H. and Shubik, Martin (1994) "Some Dynamics of a Strategic Market Game," *Journal of Economics*, **60**.
- Miller, J. H. and J. Andreoni (1991) "Can Evolutionary Dynamics Explain Free Riding in Experiments?" *Econ. Lett.* 36: 9-15.
- Nachbar, John H. (1990) "'Evolutionary' Selection Dynamics in Games: Convergence and Limit Properties," *International Journal of Game Theory*, **19**, pp. 59-89.
- Nachbar, John H. (1992) "Evolution in the Finitely Repeated Prisoner's Dilemma: A Methodological Comment and Some Simulations," *Journal of Economic Behaviour and Organization*, **19**(3), December, pp. 307-326.
- Neyman, A. (1985) "Bounded Complexity Justifies Cooperation in the Finitely Repeated Prisoner's Dilemma," *Economics Letters*, **19**, pp. 227-229.
- Nowak, Martin A. and May, Robert M. (1992) "Evolutionary Games and Spatial Chaos," *Nature*, **359**(6398), 29 October, pp. 826-829.
- Nowak, Martin A. and Sigmund, K. (1992) "Tit For Tat in Heterogenous Populations," *Nature*, **359**, pp. 250-253.
- Nowak, Martin A. and May, Robert M. (1993) "The Spatial Dilemmas of Evolution," *International Journal of Bifurcation and Chaos*, **3**, pp. 35-78.
- Nowak, Martin A., Sebastian Bonhoeffer, and Robert M. May (1994) "More Spatial Games," *International Journal of Bifurcation and Chaos* 4, 1: 33-56.
- Ockenfels, Peter (1993) "Cooperation in Prisoner's Dilemma - An Evolutionary Approach," *European Journal of Political Economy*, **9**, pp. 567-579.
- Reijnders, L. (1978) "On the Applicability of Game Theory to Evolution," *Journal of Theoretical Biology*, **75**(1), pp. 245-247.
- Robles, J. (1998) "Evolution with Changing Mutation Rates," *Journal of Economic Theory*, **79**, pp. 207-223.
- Robson, Arthur J. (1990) "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology*, **144**, pp. 379-396.
- Samuelson, Larry and J. Zhang (1992) "Evolutionary Stability in Asymmetric Games," *J. Econ. Theory* 57: 363-391. Samuelson, Larry (1993) "Does Evolution Eliminate Dominated Strategies?" in Kenneth G. Binmore, A. Kirman, and P. Tani, eds., *Frontiers of Game Theory*, Cambridge, MA: MIT Press, pp. 213-235.
- **-----** (1997). *Evolutionary Games and Equilibrium Selection*. MIT Press series on economic learning and social evolution. Cambridge, Massachusetts: MIT Press.
- Schlag, Karl H. (1998) "Why Imitate, and If So, How? A Boundedly Rational Approach to Multi-armed Bandits," *Journal of Economic Theory* 78: 130-156.
- Schuster, P. and Sigmund, K. (1983) "Replicator Dynamics," *Journal of Theoretical Biology*, pp. 533-538.
- Selten, Reinhard (ed.) (1991) *Game Equilibrium Models I: Evolution and Game Dynamics* (New York, NY: Springer-Verlag).
- Selten, Reinhard (1993) "Evolution, Learning, and Economic Behaviour," *Games and Economic*

Behaviour, 3(1), February, pp. 3-24.
- Sinclair, P. J. N. (1990) "The Economics of Imitation," *Scottish Journal of Political Economy*, 37(2), May, pp. 113-144.
- Skyrms, Brian (1992) "Chaos in Game Dynamics," *Journal of Logic, Language, and Information* 1: 111-130.
- ----- (1993) "Chaos and the Explanatory Significance of Equilibrium: Strange Attractors in Evolutionary Game Dynamics," in *Proceedings of the 1992 PSA*. volume 2, pp. 374-394.
- ----- (1994a) "Darwin Meets *The Logic of Decision*: Correlation in Evolutionary Game Theory," *Philosophy of Science* 61: 503-528.
- ----- (1994b) "Sex and Justice," *Journal of Philosophy* 91: 305-320.
- ----- (1996) *Evolution of the Social Contract*. Cambridge University Press.
- ----- (1997) "Game Theory, Rationality and Evolution," in M. L. Dalla Chiara et al., eds., *Structures and Norms in Science*, Kluwer Academic Publishers, pp. 73-85.
- ----- (1998) "Salience and symmetry-breaking in the evolution of convention," *Law and Philosophy* 17: 411-418.
- ----- (1999) "Précis of *Evolution of the Social Contract*," *Philosophy and Phenomenological Research* 59, 1: 217-220.
- ----- (2000) "Game Theory, Rationality and Evolution of the Social Contract," *Journal of Consciousness Studies* 7, 1-2: 269-284.
- ----- (2000) "Adaptive Dynamic Models and the Social Contract," *Journal of Consciousness Studies* 7, 1-2: 335-339.
- Smale, Steve (1980) "The Prisoner's Dilemma and Dynamical Systems Associated to Non-cooperative Games," *Econometrica*, 48, pp. 1617-1634.
- Maynard Smith, John and George Price (1973) "The Logic of Animal Conflict," *Nature* 246: 15-18.
- Maynard Smith, John (1982) *Evolution and the Theory of Games*. Cambridge University Press.
- Stanley, E. Ann, Dan Ashlock, and Leigh Tesfatsion (1994) "Iterated Prisoner's Dilemma with Choice and Refusal of Partners," in Christopher G. Langton, ed., *Artificial Life III*. Addison-Wesley, pp. 131-175.
- Suleiman, Ramzi and Ilan Fischer (1996) "The Evolution of Cooperation in a Simulated Inter-Group Conflict," in Liebrand and Messick, eds., *Frontiers in Social Dilemmas Research*, Springer.
- Taylor, Peter D. and Leo B. Jonker (1978) "Evolutionary Stable Strategies and Game Dynamics," *Mathematical Biosciences* 40: 145-156.
- Tomochi, Masaki and Mitsuo Kono (1998) "Social Evolution Based on Prisoner's Dilemma with Generation Dependent Payoff Matrices," *Research on Policy Studies* 3: 79-91.
- Trivers, Robert L. (1971) "The evolution of reciprocal altruism," *The Quarterly Review of Biology* 46: 35-57.
- Vanderschraaf, Peter (2000) "Game Theory, Evolution, and Justice," *Philosophy and Public Affairs* 28, 4: 325-358.
- Vega-Redondo, Fernando (1996) *Evolution, Games, and Economic Behaviour* (Oxford: Oxford University Press).
- Vega-Redondo, Fernando (1997) "The Evolution of Walrasian Behavior," *Econometrica*, 65(2), pp. 375-384.
- Weibull, Juergen W. (1995) *Evolutionary Game Theory* (Cambridge, MA: The M.I.T. Press).
- Witt, Ulrich (1989a) "The Evolution of Economic Institutions as a Propagation Process," *Public Choice*, 62(2), August, pp. 155-172.
- Young, H. Peyton. (1993) "An Evolutionary Model of Bargaining," *Journal of Economic Theory* 59: 145-168.
- Young, H. Peyton (1993) "The Evolution of Conventions," *Econometrica* 61, 1: 57-84. Young, H. Peyton

(2001) *Individual Strategy and Social Strategy: An Evolutionary Theory of Institutions*, Princeton, NJ: Princeton University Press.

Other Internet Resources

- Center for Learning, Evolutionary Game Theory, and Economics University of Vienna
- Brookings Center on Social and Economic Dynamics
- Complexity of Cooperation, website on Robert Axelrod's book (Center for the Study of Complex Systems, U. Michigan)

[Please contact the author with other suggestions.]

Related Entries

evolution: cultural | game theory | prisoner's dilemma

J. McKenzie Alexander
jalex@lse.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 14, 2002
Content last modified: January 14, 2002

Stanford Encyclopedia of Philosophy
Occasional Referees

The following philosophers have served as referees on those occasions when their expertise on a topic made them particularly appropriate.

- David Barker-Plummer (Stanford University)
- Jeffrey Blustein (Albert Einstein College of Medicine)
- Stuart Brock (Western Washington University)
- Johan van Benthem (University of Amsterdam and Stanford University)
- Earl Conee (University of Rochester)
- Max Cresswell (Victoria University of Wellington)
- John Divers (University of Leeds)
- Greg Fitch (Arizona State University)
- Branden Fitelson (Stanford University)
- Lou Goble (Willamette University)
- Ruth Grant (Duke University)
- Lori Gruen (Wesleyan University)
- Bob Hale (University of Glasgow)
- Thomas Hofweber (University of Michigan)
- David Israel (SRI/Menlo Park)
- Ernie Lepore (Rutgers University)
- Bernard Linsky (University of Alberta)
- Kirk Ludwig (University of Florida)
- Paul McNamara (University of New Hampshire)
- Christopher Menzel (Texas A&M University)
- Karl-Georg Niebergall (Universität München)
- Francis Jeffry Pelletier (University of Alberta)
- William Throop (Green Mountain College)
- Peter Vanderschraaf (Carnegie-Mellon University)
- Gary Varner (Texas A&M University)
- Albert Visser (University of Utrecht)

Entry Content: Guidelines and Policies

In this document, we develop guidelines and policies concerning the content of entries written for the *Stanford Encyclopedia of Philosophy*.

Entry Substance, Style, and Length

The *Stanford Encyclopedia of Philosophy* is intended to serve as an authoritative reference work suitable for use by professionals and students in the field of philosophy, as well as by all others interested in authoritative discussions on philosophical topics. Entries should therefore be written with the highest of professional standards, and be of interest to as wide an audience as possible. Entries should focus on the philosophical issues and arguments rather than on sociology and individuals, particularly in discussions of topics in contemporary philosophy. In other words, entries should be "idea-driven" and not "people-driven". Bibliographic entries should be critical and not hagiographical. Authors should strive to *minimize the use of jargon*.

All entries should provide objective, neutral analyses or surveys of particular topics, rather than promoting idiosyncratic or controversial points of view. Authors should see their task as one of offering a broad or insightful perspective which introduces both the topic and the literature and which puts the reader into position to read the primary and secondary sources cited in the entry. (To this end, the sources of all quotations should be clearly identified.) Clarity of substance and style should also be one of the most important goals.

Encyclopedia entries should therefore not be polemical. Controversial claims should be identified as such. Authors should not use the first person pronoun "I", and should avoid such constructions as "as I have argued elsewhere/previously...". In addition, authors should not cite or refer to unpublished or inaccessible materials, and in particular, to unpublished dissertations or talks. Authors should also be circumspect in the number of references to their own work, though obviously, since they are experts on the topic and typically will have written widely on it, occasional references will be in order and appropriate. The editors of the *Encyclopedia* will ensure that entries do not overstep the bounds of propriety in this regard.

The length of entries should depend on the topic. Entries will typically range from 7,000-8,000 words, but may be anywhere from half this long to twice this long depending on the how broadly the topic is focused and how much literature there is to introduce and explain. We encourage authors to organize longer entries by writing a set of nested, cross-linked documents rather than by writing a single, linear document. By this we mean that overly detailed, highly technical, or highly scholarly material should be put into separate HTML ("supplementary") documents and linked into the main entry. (See below.) This way, the main entry should become readable by an intelligent undergraduate in a sitting of about an hour or two. More advanced readers can follow the links to the highly technical, detailed or scholarly material. Such a cross-linked set of documents will therefore be accessible to a wide audience. However, authors should create such "nested" entries only if it seems unlikely that a separate entry in the *Encyclopedia* will be created to discuss the supplementary material.

Writing Your Entry in HTML

Because the *Encyclopedia* is being served over the World Wide Web, all entries must be written in HTML (HyperText Markup Language). This is the formatting language that controls the way text, graphics, and links are displayed in Web browsers. There are now available numerous software programs known as "HTML-editors" or "web authoring tools" and these allow the authors to format text and graphics in HTML easily, without learning any arcane commands. These programs are now as easy to use as Microsoft Word.

To begin writing an entry, authors should follow our instructions for downloading the sourcefile of the "Entry Template". Once the template is downloaded, the author may simply "Open" that file using an HTML-editor. This Entry Template will ensure that there is a uniform entry style, which is described below, in the section on "Entry Format".

For those authors who prefer to create an HTML sourcefile directly, without the assistance of an HTML-editor, we've indicated, in our instructions, how to obtain the "Annotated Sourcefile". After downloading the sourcefile, authors can replace the sample text in this file with their own content. This will minimize the number of HTML commands authors will need to learn.

Note that it is easy to create a link from your main entry to supplementary documents containing overly technical or scholarly material that would interfere with the presentation of the main ideas. To create a link from the main document to a supplementary document, suppose that the main document is entitled "index.html" and the supplementary document is entitled "supplement.html". Now suppose that at the end of a paragraph, you wish place a labeled link to the supplementary document. Here is what the link might be displayed as:

> ... . We discuss this last point in further detail in the following supplementary document:

> Supplement on [Title of Supplement]

Authors using an HTML-editor (such as Netscape Composer, Adobe Page Mill, Front Page Express, etc.) will have to use the functions provided by their software to create such a link. However, authors who can edit their HTML sourcefiles directly would insert the following HTML code in order to produce this link to the supplementary document:

```
...in the following supplementary document.
<blockquote>
<a href="supplement.html" name="return-1">Supplement on [Title of Supplement]</a>
</blockquote>
```

Note that this can be done in other places in your main document, if you have more than one supplementary document. (You will have to name your supplements as "supplement1.html", "supplement2.html", etc.) You can find the sourcefile for the supplement template here (use the View Source function of your browser after following this link).

Entry Format

The Entry Template and Annotated Sourcefile are formatted in HTML so that the following divisions are preserved in every entry:

- Introduction (Definition)
- Internal Links
- Main Sections of the Entry
- Bibliography
- Other Internet Resources
- Related Entries

These are discussed in turn.

Introduction. The Introduction should contain a brief definition of the subject. This may take one or two paragraphs, and if possible, these paragraphs should contain some statement of the subject's interest and significance. The main topics to be covered in the body of the entry may be mentioned here, so that the reader will get some idea of what is to follow.

Internal Links. The internal links should be a list of the main sections of the entry, and each item in the list should be a link to that section. The HTML commands needed to do this are included in the template and in the annotated sourcefile.

Main Sections. The sectioning of the entry is at the discretion of the author. However, we encourage

authors to include a Chronology or "Life" section in Biographical entries. Moreover, a "History" section is called for in the discussion of many topics.

Bibliography. Please use the following bibliographic format:

- Dodgson, H., 1885, ‘The Evidence for the Existence of Snarks’, *Journal of Ornithology,* 25: 22-44
- Hanes, A., 1999a, *Deliverance from Evil Bandersnatches,* London: Houghton & Miflin
- -----, (ed.), 1999b, *Papers on Alice*, Penrith: Bilgewater Press
- Madsen, B., 1924, ‘Slithy Toves’, in *History of Poetry*, S. Johnson (ed.), Cambridge: Cambridge University Press
- Terrell, N., 1888, ‘How to Gimble’, Paris: Longine; page reference is to the reprint in Hanes 1999b.

With this style of Bibliography, you can then cite these sources in your main text with the indication "(Dodson 1885, 32-33)".

Please note: (1) The Bibliography section may be divided into subsections such as Primary Literature and Secondary Literature, or References Cited and Other Important Works, etc. and (2) he Bibliography is reserved primarily for *refereed* material, whether print-based or on the web. Books, journal articles, e-journal articles, etc., which have undergone the normal referee process of legitimate presses should be included. Whenever a cited article or book is available online, a link may be included to the URL.

Other Internet Resources. The author should cite material on the web that is of excellent value but which may not have undergone a referee process. The author serves as referee for these materials (and our subject editors will referee these choices). To complete this section, authors are encouraged to conduct an on-line search of the Web for websites with *high-quality, academic content* on the topic in question. Such websites should be written and maintained by qualified individuals having a clear expertise on the topic. The task of finding such external websites is made considerably simpler by using the Limited Area Search Engines for philosophy, such as Hippias and Noesis. They can be found at the URLs:

> http://hippias.evansville.edu/
> http://noesis.evansville.edu

If this doesn't yield any results, you should try one of the wider area search engines, such as Google or Alta Vista, which can be found at:

> http://www.google.com/
> http://www.altavista.com/

Other search engines include www.lycos.com and www.yahoo.com. Please do not create links to websites that are not maintained by qualified individuals.

Related Entries. Please list the names of the most important concepts and philosophers that occur in your entry. You may list keywords that do not appear as topics in our Table of Contents if you feel that they are important. We are running software which will notice the discrepancy and alert the Editor. A decision will be made whether or not to include a new entry on that topic. If we decide that the topic is too specialized or otherwise inappropriate for the *Encyclopedia*, we will eliminate this keyword from your list in the Related Entries section.

Entry Revision

Because the *Encyclopedia* is designed to be a dynamic reference work, authors are responsible for maintaining and periodically updating their entries. Specifically, authors are expected: (1) to update their entries regularly, especially in response to important new research on the topic of the entry, and (2) to revise their entries *in a timely way* in light of any valid criticism they receive, whether it comes from the subject editors on our Editorial Board, other members of the profession, or interested readers. In connection with (1), authors should update the Bibliography and the Other Internet Resources sections of their entries regularly, to keep pace with significant new publications, both in print and on the web. In connection with (2), the validity of criticism shall be determined by the Editor, typically in consultation with the relevant members of the Editorial Board. The length of time required for a "timely" revision will be negotiable and will both respect the author's current commitments and reflect the seriousness of the criticism. However, entries which require revision but which are not revised within the negotiated timetable may be retired from the active portion of the *Encyclopedia* and left in the *Encyclopedia* Archives until such time as the entry is revised so as to engage the valid criticisms in question.

Making Modifications: There is a preferred, recommended, and easy to use protocol for making changes to your entry. If you would like to add/revise a paragraph, add an item to the Bibliography, fix a typo, etc., the the proper procedure to follow is to log in to our Author Area:

> http://plato.stanford.edu/cgi-bin/encyclopedia/authors.cgi

and initiate the action "Revise Entry on Server". This will allow you to directly edit a copy of your entry on our machine (from whichever browser you are using). There is an Instructions/Help file for using this software. However, as you will see, when you use the Revise Entry function, you will be prompted to select the file you wish to edit. In most cases, you will select the main document, which is called "index.html" (some entries have multiple files, e.g., a main document "index.html" and supplementary documents). Once you have made your selection, a new window will open on your browser and you will be presented with a page on which the file you wanted to edit is divided up into segments, each containing a "View" box and an "Edit" box. You will find the material you wish to edit in the View box, since the text in this box is rendered, or formatted, HTML. Then you edit in the corresponding "Edit" box, which contains the HTML sourcefile (which is plain text with markup tags). It should be clear how add content in the Edit box. Every so often, you should SAVE your work, using the Save buttons at the left or at the top of the page. You may SAVE your work without submitting it for review, but when you

believe you have completed the revisions you need to make, use the Save/Submit option, or return to the authors main menu page and use the Submit/Resubmit Privates Files to Editor function.

Making Major Modifications: There are only two conditions under which it is acceptable to follow a somewhat different procedure. (1) If you know the difference between a simple text editor and an HTML-editor, and you know that your simple text editor can be used to edit an HTML sourcefile without damaging or rewriting the HTML, then you may use the Download Existing Entry File function, edit the file using your simple text editor in plain text mode, save as plain text, and then reupload the file. (2) If you need to make major modifications to your entry, such as a structural reorganization, then it may be inappropriate to use the Revise Entry on Server function. In this case, you would use the Download Existing File function and edit it locally on your computer. However, before you begin to make major modifications to your entry (e.g., by downloading and editing locally), please note that many HTML-editors (Word, Netscape Composer, Dreamweaver, Adobe Page Mill, etc.) do not follow international standards for producing correct HTML. Some add special control characters to the file; others make changes to the HTML, following their own conception of how HTML should be written. So, if you can, please make major modifications to your entry by using a *simple text editor* and by saving the file as plain text (an HTML file is a plain text file in which special pieces of plain text, e.g., "tags" such as <em>, are used to "markup the text" in a way that gives formatting instructions to the web browser). Though a simple text editor will show you all the HTML markup tags, it should be easy to find your way around the file and edit the portions you are interested in. By using a simple text editor, you save us the trouble of having to reprocess (sometimes by hand) the HTML produced by these unfriendly HTML-editors.

After after editing your file, it should then be uploaded back to the *Encyclopedia* by using our Author Area and the "Upload Single File" action.

The Use of Footnotes

Footnotes may be included. Then can often help to shorten the main page of the entry, to make it more readable. The footnotes themselves should be put into a separate html file called "notes.html" and these should be placed into the same directory on plato.stanford.edu as the entry. If you are using an HTML-editor to create your entry, then you will need to use the functions provided by your software in order to create foonotes as links to a footnote file.

However, for those authors who wish to edit their HTML sourcefile directly in order to create links from the text to the footnotes, here are some guidelines to follow. Suppose you want to add footnote number *x* at a point in the text:

> ...some text.[x]

To produce this in your HTML sourcefile, use the following HTML code at the point in the text where

the footnote should occur:

```
...some text.<sup>[<a href="notes.html#x" name="note-x">x</a>]</sup>
```

This will place "[x]" as a superscript in the text, with "x" a link to the place in the notes.html file identified as name="x" (see below). The name="note-x" marks the spot in the current file to which a "Return" link from the notes.html file will return (again, see below). Then, create another HTML file named "notes.html" and in that file you will try to produce something that looks like this:

> x. Begin the body of the footnote here.

To produce this line in the notes.html file, you would add the following HTML code:

```
<p>
<a href="index.html#note-x" name="x">x.</a> Begin the body of the footnote.
```

This will start a new paragraph, start the footnote with the symbols "x." ("x." will be a link back to name="note-x" in the main index.html file; the name="x" identifies the place in the notes.html file to which the footnote in the main text will be linked). Note: Users of the *Encyclopedia* can always use the "Back" or "Return" button on their browsers to get back to the text.

The Use of Special Symbols

Although the specifications for the HTML4.0 language does include support for a variety of special symbols, including Greek, logic and math symbols, there is little support for arranging these symbols on the page in the variety of ways needed by mathematicians and logicians. A new standard for typesetting mathematical and logical formulas is developing, namely, MathML. See W3C Math: MathML puts Math on the Web. However, it will be sometime before web browsers are reprogrammed to meet this standard. So, for now, we would ask you not to employ the MathML standard or the symbols that are supported in HTML4.0.

In the meantime, if the special symbol you need is not on this list of special HTML characters which is widely supported, we have created a wide variety of small graphics of the Greek, logic, and math symbols that logicians and others typically use. These symbols are located at the URL:

> http://plato.stanford.edu/symbols/

For example, we have produced the following graphic of the symbol for the "set membership" relation:

> $\in$

This symbol can be found at the URL displayed above and you can download it onto your machine from there. Follow the link to "element.gif" in the table for Math symbols. You can then save that graphic onto the drive of your local computer.

If you are using an HTML-editor to create your entry, then you simply use your software's "add image" ("add graphic") function to place this graphic in your entry. However, those authors who are editing their HTML sourcefiles directly should use the following guidelines. To produce the formatted line:

$x \in y$

place the file "element.gif" into the directory containing your HTML entry and use the following HTML code in your entry:

```
<em>x</em> <img src="element.gif"> <em>y</em>
```

Note that you will not need to transfer the graphics to the same directory on plato when you transfer your entry to us -- the first time you view your file on our server after uploading it, our software will recognize the links to our graphics and it will automatically install copies of the needed graphic files into your upload directory.

You may use any graphic found in our symbols directory in this way. If you need a symbol not found in that directory, write to the editor---they are easily constructed.

[Title of Supplement]

[Supplementary Material]

[Title of Subsection]

[Supplementary Material]

[Title of Subsection]

[Supplemental Material]

[Link to Author Home Page]
[Author email address]

Return to [Title of Entry]

First published: [Month Day, Year]
Content last modified: [Month Day, Year]

Snark

Snarks are a species of bandersnatch. They are thought to have evolved from the slithy toves along with the other bandersnatches. Snarks were common in New England forests until the late 19th century, when it was thought that they were hunted to extinction. Recent ethological expeditions, however, have uncovered evidence that there may be a few extant individuals. Snarks are favorites with animal lovers everywhere because in the spring, they gyre and gimble on small ponds in a brilliant mating display. In reference works on bandersnatches, snarks are referred to as a group by the Latin name *Snarkidae*.

Evolutionary History and Adaptive Features

Snarks are a species of bandersnatch and are thought to have evolved from the slithy toves along with the other bandersnatches. Snarks evolved in the Devonian period and successfully adapted themselves in several ecological niches until the late 19th century. The fact that they have a dorsal sigmoid bone establishes their descent from the slithy toves.

Snarks bear young just at the point in early summer when their food source is most abundant. Their biological clocks are very well tuned and biologists study them as the principal model of such clocks in the animal kingdom.

Mating Display

There is nothing like the wonder and pageantry of snarks in mating season, as they gyre and gimble on the ponds of New England forests.

Bibliography

- Doe, J., "The Ecological Range of the Snark," *Ecology Today* **3**/1 (January 1992): 15-30
- Dodgson, C., "The Origins of Bandersnatches," *Annals of the Society for the Investigation of the Descendants of the Slithy Toves* **XL** (1993): 1-20

Other Internet Resources

- A Way Station for Snark Hunters
- Cosmological Snark Hunts

Related Entries

bandersnatch | gyration | toves

Charles Dodgson
dodgson@alice.uwonderland.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 14, 1995
Content last modified: January 9, 2001

Stanford Encyclopedia of Philosophy

Author Interface

Instructions/Help

Welcome. To use this system you need to know your login id and password.
You can find this information in our correspondence.
For additional assistance please send email to editors@plato.stanford.edu.

Please enter your login id:

and password:

Login

To maximize security Close Window when you have finished

(Information current as of Wed Aug 7 18:03:26 2002 GMT -- don't forget to reload page if you want latest information.)

Stanford Encyclopedia of Philosophy

Table of Symbols

NOTE: So that the symbols will load faster, we have placed the tables of Greek, Logical and Math symbols on separate HTML pages:

- Table of Greek Symbols:
- Table of Logical Symbols
- Table of Math Symbols

Stanford Encyclopedia of Philosophy

Table of Greek Symbols

Symbol	Name	Symbol	Name	Symbol	Name
α	alpha.gif	β	beta.gif	γ	gamma.gif
δ	delta.gif	Δ	Delta.gif	Γ	Gamma.gif
ε	epsilon.gif	ζ	zeta12.gif	ζ	zeta10.gif
θ	theta.gif	Θ	Theta.gif	ι	iota.gif
κ	kappa.gif	$\varkappa$	varkappa.gif	λ	lambda.gif
Λ	Lambda.gif	μ	mu.gif	ξ	xi.gif
π	pi.gif	Π	Pi.gif	ρ	rho.gif
σ	sigma.gif	Σ	Sigma.gif	τ	tau.gif
Φ	Phi.gif	φ	phi.gif	Ψ	Psi.gif
ψ	psi.gif	ω	omega.gif	Ω	Omega.gif
ν	nu.gif	χ	chi.gif		

Note: Authors who need other Greek symbols for their entry should send email to the Editors.

- Table of Logical Symbols
- Table of Math Symbols

α

β

γ

δ

Δ

Γ

ε

ζ

ζ

θ

Θ

ι

κ

$\varkappa$

λ

Λ

μ.

ξ

π

Π

ρ

σ

Σ

τ

Φ

φ

Ψ

ψ

ω

Ω

ν

χ

Stanford Encyclopedia of Philosophy

Table of Logical Symbols

Symbol	Name	Symbol	Name	Symbol	Name
$\exists$	exists.gif	$\forall$	forall.gif	$\rightarrow$	ra.gif
$\leftrightarrow$	lra.gif	$\neg$	neg.gif	$\neq$	not-equal.gif
$\supset$	righthook.gif	$\vee$	vel.gif	$\models$	models.gif
$\vdash$	proves.gif	$\Rightarrow$	Rightarrow.gif	$\Diamond$	Diamond.gif
$\Box$	Box.gif	$\equiv$	equiv.gif	$\sim$	sim.gif
$'$	prime.gif	$''$	doubleprime.gif	$\therefore$	therefore.gif
$\bot$	perp.gif	⥽	fishhook.gif	$\wedge$	wedge.gif
$\Leftrightarrow$	Leftrightarrow.gif	$\top$	top.gif		langl.gif
$\rangle$	rangl11.gif	$\rangle$	rangl10.gif	$\leftarrow$	la.gif
$\leftarrow$	leftarrow11pt.gif	$\not\models$	not-models.gif	$\bot$	small-perp.gif
$\Leftarrow$	Leftarrow.gif	$\mathcal{M}$	calM.gif	$\mathcal{S}$	calS.gif
$\models\!\sim$	sim-models.gif	$\vert$	vertical-bar11.gif	$\vert$	vertical-bar10.gif
$\langle$	langl11.gif	$\langle$	langl10.gif		_

Note: Authors who need other logical symbols for their entry should send email to the Editors.

- Table of Greek Symbols
- Table of Math Symbols

∃

$\forall$

$\rightharpoonup$

↔

¬

$\neq$

⊃

∨

$\models$

⊢

⇒

◇

□

≡

∼

′

″

∴

⊥

∧

⇔

⊤

〉

〉

$\leftharpoonup$

$\nvDash$

⊥

⇐

$\mathcal{M}$

$\mathcal{S}$

$\mid\!\sim$

|

|

⟨

⟨

Stanford Encyclopedia of Philosophy

Table of Mathematical Symbols

Symbol	Name	Symbol	Name	Symbol	Name
$\geq$	geq.gif	$\leq$	leq.gif	∞	infinity.gif
$\in$	element.gif	$\notin$	not-element.gif	$\subseteq$	subset.gif
$\subset$	prop-subset.gif	$\emptyset$	nullset.gif	$\mathbb{R}$	real.gif
$\int$	int.gif	∂	partial.gif	$\aleph$	aleph.gif
$\cap$	intersect.gif	$\bigcup$	bigcup.gif	$\cup$	cup.gif
$\circ$	circ.gif	$\hbar$	hbar.gif	$\cdot$	cdot.gif
$\otimes$	Cprod.gif	$\surd$	Sqrt.gif		blank5.gif
$-$	minus.gif	$\mapsto$	mapsto.gif	$\sum$	sum.gif
$\pm$	pm.gif	$\approx$	approx.gif	$\Im$	frak-I.gif
$\oplus$	oplus.gif	$\bigcap$	bigcap.gif	$\mathcal{H}$	calH.gif
$\mathbb{N}$	Num.gif	$\boldsymbol{\mathcal{R}}$	bold-calR.gif	$\mathcal{R}$	calR.gif
$\mathcal{A}$	calA.gif	$\mathcal{B}$	calB.gif	$\mathcal{P}$	calP.gif
$\bigwedge$	bigwedge.gif	$\bigvee$	bigvee.gif	$\prod$	prod.gif
$\times$	times.gif				

Note: Authors who need other math symbols for their entry should send email to the Editors.

- Table of Greek Symbols
- Table of Logical Symbols

≥

≤

∞

∈

∉

$\subseteq$

⊂

∅

$\mathbb{R}$

∫

∂

ℵ

∩

$\bigcup$

∪

∘

ħ

▪

$\otimes$

√

—

$\mapsto$

$\sum$

±

≈

⊕

$\bigcap$

$\mathcal{H}$

ℕ

$\mathcal{R}$

$\mathcal{R}$

$\mathcal{A}$

$\mathcal{B}$

$\mathcal{P}$

$\bigwedge$

$\bigvee$

$\prod$

×

Stanford Encyclopedia of Philosophy

HTML Editors and Web Authoring Tools

- Yahoo's Web Page on HTML Editors
- Netscape Product Archive
- Microsoft's Front Page 97
- Adobe's GoLive
- My Internet Business Page
- SoftQuad's HoTMetaL Pro

Stanford Encyclopedia of Philosophy

HTML Guides and Online Help for Writing HTML

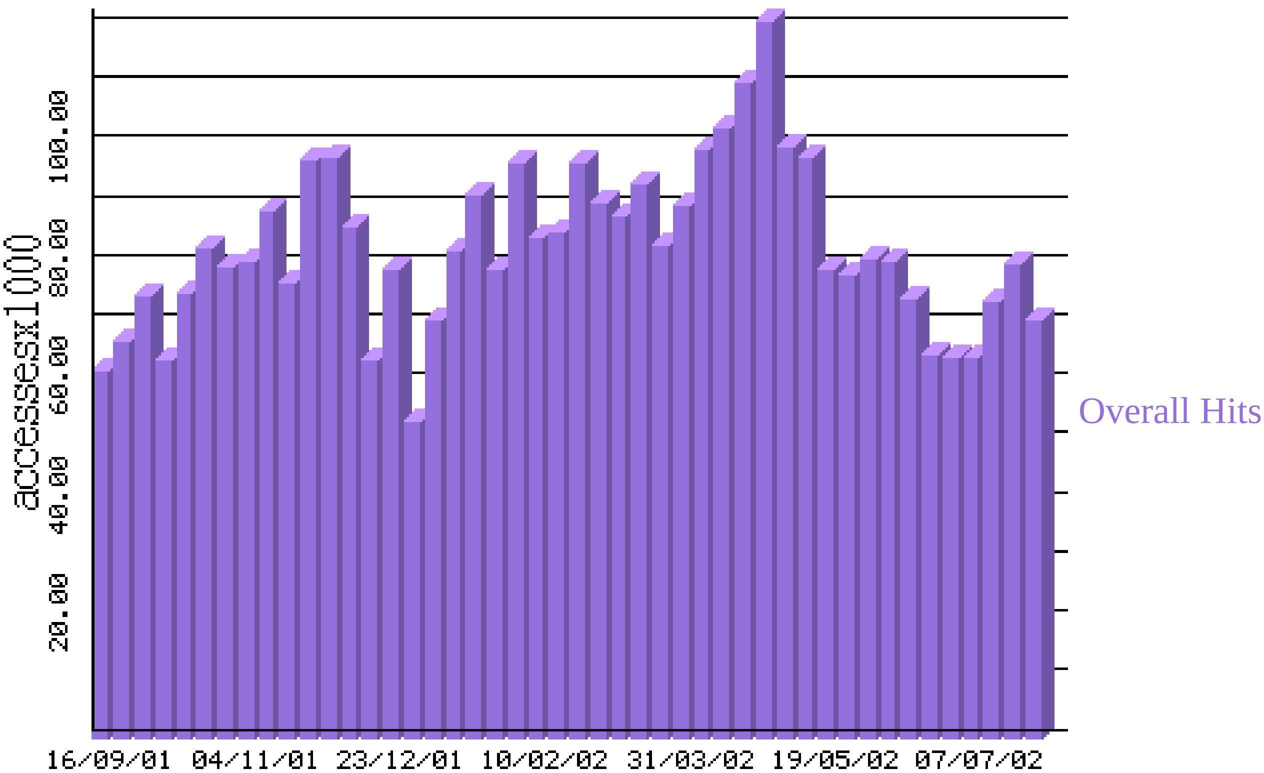
- The W3C HTML 3.2 Reference Specification
- The HTML 3.2 Standard
- Yahoo Page of links concerning HTML
- Yahoo List of HTML Guides and Tutorials
- WebMonkey's HTML Basics (An Excellent HTML Guide)
- List of Special HTML Characters in HTML 3.2
- A Beginner's Guide to HTML (NCSA)
- Beginners' Introduction to HTML (by Peter Flynn)
- HTML Quick Reference (by Michael Grobe)
- W3 Consortium Information Page on HTML
- Composing Good HTML (by James "Eric" Tilton)
- Spinning the Web: An Introduction to HTML (by James Powell)

Web Server Statistics for The Stanford Encyclopedia of Philosophy

History from 16/09/01 -- 03/08/02

Skip to weekly reports

Totals



Item	Total Accesses	Total Bytes	Average Accesses	Average Bytes	Latest Accesses	Latest Bytes
Overall Hits	3,822,120	186,289,035,054	83,090	4,049,761,632	70,970	3,577,136,244

Reports

Web Server Statistics for The Stanford Encyclopedia of Philosophy

Reports for the year 2002

Week of 28/07/02 to 03/08/02
Week of 21/07/02 to 27/07/02
Week of 14/07/02 to 20/07/02
Week of 07/07/02 to 13/07/02
Week of 30/06/02 to 06/07/02
Week of 23/06/02 to 29/06/02
Week of 16/06/02 to 22/06/02
Week of 09/06/02 to 15/06/02
Week of 02/06/02 to 08/06/02
Week of 26/05/02 to 01/06/02
Week of 19/05/02 to 25/05/02
Week of 12/05/02 to 18/05/02
Week of 05/05/02 to 11/05/02
Week of 28/04/02 to 04/05/02
Week of 21/04/02 to 27/04/02
Week of 14/04/02 to 20/04/02
Week of 07/04/02 to 13/04/02
Week of 31/03/02 to 06/04/02
Week of 24/03/02 to 30/03/02
Week of 17/03/02 to 23/03/02
Week of 10/03/02 to 16/03/02
Week of 03/03/02 to 09/03/02

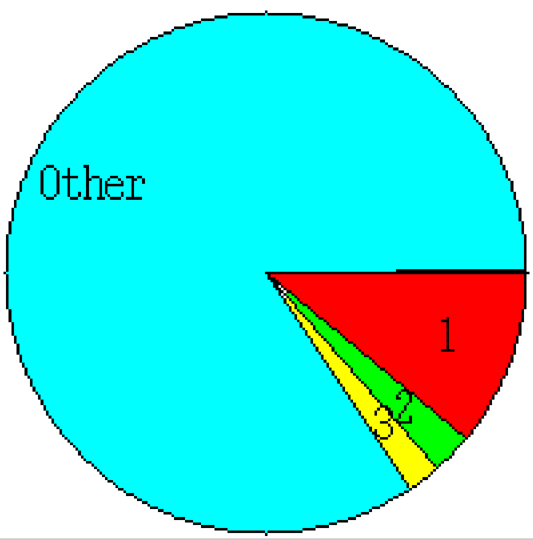
Web Server Statistics for The Stanford Encyclopedia of Philosophy

Week of 28/07/02 to 03/08/02

Totals

Item	Accesses	Bytes
Overall Hits	70,970	3,577,136,244

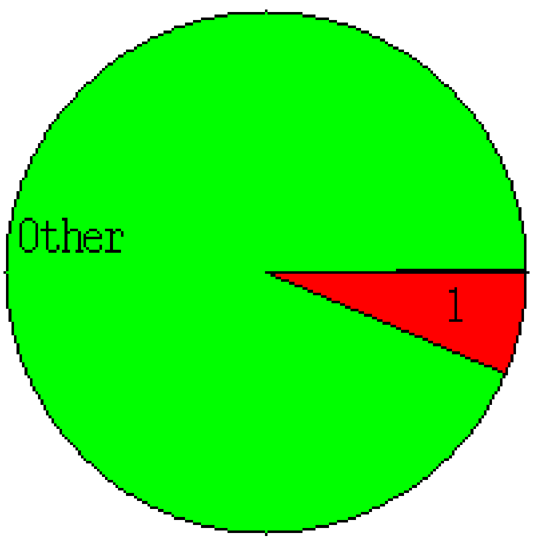
Top 20 of 1271 Documents Sorted by Access Count



Rank	URL	Accesses	Bytes
1	/contents.html	8,044	541,270,953
2	/entries/pascal-wager	1,661	70,522,174
3	/entries/nietzsche	1,428	63,873,656
4	/entries/russell	1,160	47,772,921
5	/entries/russell-paradox	915	11,512,369
6	/entries/popper	819	53,881,917
7	/entries/game-theory	740	84,822,433

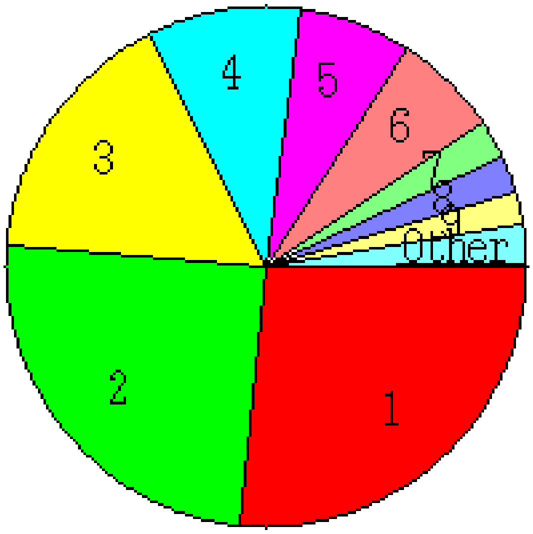
8	/entries/kierkegaard	644	30,733,406
9	/entries/descartes-epistemology	623	64,816,958
10	/entries/euthanasia-voluntary	581	20,914,368
11	/entries/hegel	564	35,634,570
12	/entries/turing-machine	563	3,499,272
13	/entries/aristotle-politics	523	15,069,471
14	/entries/time-travel-phys	508	34,961,729
15	/entries/liberalism	488	18,585,506
16	/entries/aquinas	469	29,273,638
17	/contents-unabridged.html	455	47,134,936
18	/entries/aristotle-logic	453	34,932,094
19	/entries/prisoner-dilemma	445	40,709,539
20	/entries/miracles	442	23,528,953

Top 10 of 19869 Sites by Access Count



Rank	Site	Accesses	Bytes
1	oeri3.ed.gov	4,676	193,953,730
2	crawler12.googlebot.com	1,300	68,712,880
3	199.4.154.11	1,012	40,893,656
4	213.189.174.196	723	28,279,569
5	64.41.39.158	722	27,403,750
6	crawler11.googlebot.com	681	36,732,738
7	207-121-0-68.sapient.com	679	37,939,693
8	cache.mylinuxisp.com	676	28,632,419
9	200.201.164.10	674	28,480,706
10	202.142.72.109	674	29,194,850

Top 10 of 29 Domains by Access Count



Rank	Domain	Accesses	Bytes
1	Unknown	19,066	936,999,567
2	com	17,406	862,801,922
3	net	11,573	616,207,820
4	europe	6,951	370,886,250
5	gov	4,854	203,173,652
6	edu	4,513	236,238,009
7	australia	1,692	92,366,420
8	namerica	1,570	81,859,021
9	asia	1,498	79,213,196
10	samerica	681	34,795,049

Pascal's Wager

"Pascal's Wager" is the name given to an argument due to Blaise Pascal for believing, or for at least taking steps to believe, in God. The name is somewhat misleading, for in a single paragraph of his *Pensées*, Pascal apparently presents at least *three* such arguments, each of which might be called a ‘wager’---it is only the final of these that is traditionally referred to as "Pascal's Wager". We find in it the extraordinary confluence of several strands in intellectual thought: the justification of theism; probability theory and decision theory, used here for almost the first time in history; pragmatism; voluntarism (the thesis that belief is a matter of the will); and the use of the concept of infinity.

We will begin with some brief stage-setting: some historical background, some of the basics of decision theory, and some of the exegetical problems that the *Pensées* pose. Then we will follow the text to extract three main arguments. The bulk of the literature addresses the third of these arguments, as will the bulk of our discussion here. Some of the more technical and scholarly aspects of our discussion will be relegated to lengthy footnotes, to which there are links for the interested reader. All quotations are from §233 of *Pensées* (1910, Trotter translation), the ‘thought’ whose heading is "*Infinite---nothing*".

---

1. Background

It is important to contrast Pascal's argument with various putative ‘proofs’ of the existence of God that had come before it. Anselm's ontological argument, Aquinas' ‘five ways’, Descartes' ontological and cosmological arguments, and so on, purport to give *a priori* demonstrations that God exists. Pascal is apparently unimpressed by such attempted justifications of theism: "Endeavour ... to convince yourself,

not by increase of proofs of God..." Indeed, he concedes that "we do not know if He is ...". Pascal's project, then, is radically different: he seeks to provide *prudential* reasons for believing in God. To put it crudely, we should wager that God exists because it is the *best bet*. Ryan 1994 finds precursors to this line of reasoning in the writings of Plato, Arnobius, Lactantius, and others; we might add Ghazali to his list---see Palacios 1920. But what is distinctive is Pascal's explicitly decision theoretic formulation of the reasoning. In fact, Hacking 1975 describes the Wager as "the first well-understood contribution to decision theory" (viii). Thus, we should pause briefly to review some of the basics of that theory.

In any decision problem, the way the world is, and what an agent does, together determine an outcome for the agent. We may assign *utilities* to such outcomes, numbers that represent the degree to which the agent values them. It is typical to present these numbers in a decision matrix, with the columns corresponding to the various relevant states of the world, and the rows corresponding to the various possible actions that the agent can perform.

In *decisions under uncertainty*, nothing more is given---in particular, the agent does not assign subjective probabilities to the states of the world. Still, sometimes rationality dictates a unique decision nonetheless. Consider, for example, a case that will be particularly relevant here. Suppose that you have two possible actions, A1 and A2, and the worst outcome associated with A1 is at least as good as the best outcome associated with A2; suppose also that in at least one state of the world, A1's outcome is strictly better than A2's. Let us say in that case that A1 *superdominates* A2. Then rationality surely requires you to perform A1.

In *decisions under risk*, the agent assigns subjective probabilities to the various states of the world. Assume that the states of the world are independent of what the agent does. A figure of merit called the *expected utility*, or the *expectation* of a given action can be calculated by a simple formula: for each state, multiply the utility that the action produces in that state by the state's probability; then, add these numbers. According to decision theory, rationality requires you to perform the action of maximum expected utility (if there is one).

Example. Suppose that the utility of money is linear in number of dollars: you value money at exactly its face value. Suppose that you have the option of paying a dollar to play a game in which there is an equal chance of returning nothing, and returning three dollars. The expectation of the game itself is

$$0*(1/2) + 3*(1/2) = 1.5,$$

so the expectation of paying a dollar for certain, then playing, is

$$-1 + 1.5 = 0.5.$$

This exceeds the expectation of not playing (namely 0), so you should play. On the other hand, if the game gave an equal chance of returning nothing, and returning two dollars, then its expectation would be:

> 0*(1/2) + 2*(1/2) = 1.

Then consistent with decision theory, you could either pay the dollar to play, or refuse to play, for either way your overall expectation would be 0.

Considerations such as these will play a crucial role in Pascal's arguments. It should be admitted that there are certain exegetical problems in presenting these arguments. Pascal never finished the *Pensées*, but rather left them in the form of notes of various sizes pinned together. Hacking 1972 describes the "Infinite---nothing" as consisting of "two pieces of paper covered on both sides by handwriting going in all directions, full of erasures, corrections, insertions, and afterthoughts" (24).[1] This may explain why certain passages are notoriously difficult to interpret, as we will see. Furthermore, our formulation of the arguments in the parlance of modern Bayesian decision theory might appear somewhat anachronistic. For example, Pascal did not distinguish between what we would now call *objective* and *subjective* probability, although it is clear that it is the latter that is relevant to his arguments. To some extent, "Pascal's Wager" now has a life of its own, and our presentation of it here is perfectly standard. Still, we will closely follow Pascal's text, supporting our reading of his arguments as much as possible.

There is the further problem of dividing the *Infinite-nothing* into separate arguments. We will locate three arguments that each conclude that rationality requires you to wager for God, although they interleave in the text.[2] Finally, there is some disagreement over just what "wagering for God" involves---is it *believing* in God, or merely *trying* to? We will conclude with a discussion of what Pascal meant by this.

2. The Argument from Superdominance

Pascal maintains that we are incapable of knowing whether God exists or not, yet we must "wager" one way or the other. Reason cannot settle which way we should incline, but a consideration of the relevant outcomes supposedly can. Here is the first key passage:

> "God is, or He is not." But to which side shall we incline? Reason can decide nothing here. There is an infinite chaos which separated us. A game is being played at the extremity of this infinite distance where heads or tails will turn up... Which will you choose then? Let us see. Since you must choose, let us see which interests you least. You have two things to lose, the true and the good; and two things to stake, your reason and your will, you knowledge and your happiness; and your nature has two things to shun, error and misery. Your reason is no more shocked in choosing one rather than the other, since you must of necessity choose... But your happiness? Let us weigh the gain and the loss in wagering that God is... If you gain, you gain all; if you lose, you lose nothing. Wager, then, without hesitation that He is.

There are exegetical problems already here, partly because Pascal appears to contradict himself. He speaks of "the true" as something that you can "lose", and "error" as something "to shun". Yet he goes on

to claim that if you lose the wager that God is, then "you lose nothing". Surely in that case you "lose the true", which is just to say that you have made an error. Pascal believes, of course, that the existence of God is "the true"---but *that* is not something that he can appeal to in this argument. Moreover, it is not because "you must of necessity choose" that "your reason is no more shocked in choosing one rather than the other". Rather, by Pascal's own account, it is because "[r]eason can decide nothing here". (If it could, then it might well be shocked - namely, if you chose in a way contrary to it.)

Following McClennen 1994, Pascal's argument seems to be best captured as presenting the following decision matrix:

	God exists	***God does not exist***
Wager for God	Gain all	Status quo
Wager against God	Misery	Status quo

Wagering for God superdominates wagering against God: the worst outcome associated with wagering for God (status quo) is at least as good as the best outcome associated with wagering against God (status quo); and if God exists, the result of wagering for God is strictly better that the result of wagering against God. (The fact that the result is *much* better does not matter yet.) Pascal draws the conclusion at this point that rationality requires you to wager for God.

Without any assumption about your probability assignment to God's existence, the argument is invalid. Rationality does *not* require you to wager for God if you assign probability 0 to God existing. And Pascal does not explicitly rule this possibility out until a later passage, when he assumes that you assign positive probability to God's existence; yet this argument is presented as if it is self-contained. His claim that "[r]eason can decide nothing here" may suggest that Pascal regards this as a decision under uncertainty, which is to assume that you do *not* assign probability at all to God's existence. If that is a further premise, then the argument is valid; but that premise contradicts his subsequent assumption that you assign positive probability. See McClennen for a reading of this argument as a decision under uncertainty.

Pascal appears to be aware of a further objection to this argument, for he immediately imagines an opponent replying:

> "That is very fine. Yes, I must wager; but I may perhaps wager too much."

The thought seems to be that if I wager for God, and God does not exist, then I really do lose something. In fact, Pascal himself speaks of *staking something* when one wagers for God, which presumably one loses if God does not exist. (We have already mentioned ‘the true’ as one such thing; Pascal also seems to regard one's worldly life as another.) In other words, the matrix is mistaken in presenting the two outcomes under ‘God does not exist’ as if they were the same, and we do not have a case of superdominance after all.

Pascal addresses this at once in his second argument, which we will discuss only briefly, as it can be thought of as just a prelude to the main argument.

3. The Argument From Expectation

He continues:

> Let us see. Since there is an equal risk of gain and of loss, if you had only to gain two lives, instead of one, you might still wager. But if there were three lives to gain, you would have to play (since you are under the necessity of playing), and you would be imprudent, when you are forced to play, not to chance your life to gain three at a game where there is an equal risk of loss and gain. But there is an eternity of life and happiness.

His hypothetically speaking of "two lives" and "three lives" may strike one as odd. It is helpful to bear in mind Pascal's interest in gambling (which after all provided the initial motivation for his study of probability) and to take the gambling model quite seriously here. Recall our calculation of the expectations of the two dollar and three dollar gambles. Pascal apparently assumes now that utility is linear in number of *lives*, that wagering for God costs "one life", and then reasons analogously to the way we did! This is, as it were, a warm-up. Since wagering for God is rationally required even in the hypothetical case in which one of the prizes is three lives, then all the more it is rationally required in the actual case, in which one of the prizes is *eternal* life (salvation).

So Pascal has now made two striking assumptions:

> (1) The probability of God's existence is 1/2.
> (2) Wagering for God brings *infinite* reward if God exists.

Morris 1994 is sympathetic to (1), while Hacking 1972 finds it "a monstrous premiss". It apparently derives from the classical interpretation of probability, according to which all possibilities are given equal weight. Of course, unless more is said, the interpretation yields implausible, and even contradictory results. (You have a one-in-a-million chance of winning the lottery; but either you win the lottery or you don't, so each of these possibilities has probability 1/2?!) Pascal's best argument for (1) is presumably that "[r]eason can decide nothing here". (In the lottery ticket case, reason can decide *something*.) But it is not clear that complete ignorance should be modeled as sharp indifference. In any case, it *is* clear that there are people in Pascal's audience who do not assign probability 1/2 to God's existence. This argument, then, does not speak to them.

However, Pascal realizes that the value of 1/2 actually plays no real role in the argument, thanks to (2). This brings us to the third, and by far the most important, of his arguments.

4. The Argument From Generalized Expectations:

"Pascal's Wager"

We continue the quotation.

> But there is an eternity of life and happiness. And this being so, if there were an infinity of chances, of which one only would be for you, you would still be right in wagering one to win two, and you would act stupidly, being obliged to play, by refusing to stake one life against three at a game in which out of an infinity of chances there is one for you, if there were an infinity of an infinitely happy life to gain. But there is here an infinity of an infinitely happy life to gain, a chance of gain against a finite number of chances of loss, and what you stake is finite. It is all divided; wherever the infinite is and there is not an infinity of chances of loss against that of gain, there is no time to hesitate, you must give all...

Again this passage is difficult to understand completely. Pascal's talk of winning two, or three, lives is at best misleading. By his own decision theoretic lights, you would *not* act stupidly "by refusing to stake one life against three at a game in which out of an infinity of chances there is one for you"---in fact, you should not stake more than an infinitesimal amount in that case (an amount that is bigger than 0, but smaller than every positive real number). The point, rather, is that the prospective prize is "an infinity of an infinitely happy life". In short, if God exists, then wagering for God results in infinite utility.

What about the utilities for the other possible outcomes? There is some dispute over the utility of "misery". Hacking interprets this as "damnation", and Pascal does later speak of "hell" as the outcome in this case. Martin 1983 among others assigns this a value of *negative infinity*. Sobel 1996, on the other hand, is one author who takes this value to be finite. There is some textual support for this reading: "The justice of God must be vast like His compassion. Now justice to the outcast is less vast ... than mercy towards the elect". As for the utilities of the outcomes associated with God's non-existence, Pascal tells us that "what you stake is finite". This suggests that whatever these values are, they are finite.

Pascal's guiding insight is that the argument from expectation goes through equally well *whatever* your probability for God's existence is, provided that it is non-zero and finite (non-infinitesimal)---"a chance of gain against a finite number of chances of loss".[3]

With Pascal's assumptions about utilities and probabilities in place, he is now in a position to calculate the relevant expectations. He explains how the calculations should proceed:

> ... the uncertainty of the gain is proportioned to the certainty of the stake according to the proportion of the chances of gain and loss... [4]

Let us now gather together all of these points into a single argument. We can think of Pascal's Wager as having three premises: the first concerns the decision matrix of rewards, the second concerns the

probability that you should give to God's existence, and the third is a maxim about rational decision-making. Specifically:

1. Either God exists or God does not exist, and you can either wager for God or wager against God. The utilities of the relevant possible outcomes are as follows, where f_1, f_2, and f_3 are numbers whose values are not specified beyond the requirement that they be finite:

	God exists	***God does not exist***
Wager for God	∞	f_1
Wager against God	f_2	f_3

2. Rationality requires the probability that you assign to God existing to be positive, and not infinitesimal.
3. Rationality requires you to perform the act of maximum expected utility (when there is one).
4. Conclusion 1. Rationality requires you to wager for God.
5. Conclusion 2. You should wager for God.

We have a decision under risk, with probabilities assigned to the relevant ways the world could be, and utilities assigned to the relevant outcomes. The conclusion seems straightforwardly to follow from the usual calculations of expected utility (where p is your positive, non-infinitesimal probability for God's existence):

$$E(\text{wager for God}) = \infty * p + f_1 * (1 - p) = \infty$$

That is, your expected utility of belief in God is infinite---as Pascal puts it, "our proposition is of infinite force". On the other hand, your expected utility of wagering against God is

$$E(\text{wager against God}) = f_2 * p + f_3 * (1 - p)$$

This is finite.[5] By premise 3, rationality requires you to perform the act of maximum expected utility. Therefore, rationality requires you to wager for God.

We now survey some of the main objections to the argument.

5. Objections to Pascal's Wager

Premise 1: The Decision Matrix

Here the objections are manifold. Most of them can be stated quickly, but we will give special attention to what has generally been regarded as the most important of them, ‘the many Gods objection’ (see also the link to footnote 7).

1. Different matrices for different people. The argument assumes that the same decision matrix applies to everybody. However, perhaps the relevant rewards are different for different people. Perhaps, for example, there is a predestined infinite reward for the Chosen, whatever they do, and finite utility for the rest, as Mackie 1982 suggests. Or maybe the prospect of salvation appeals more to some people than to others, as Swinburne 1969 has noted.

Even granting that a single 2 x 2 matrix applies to everybody, one might dispute the values that enter into it. This brings us to the next two objections.

2. The utility of salvation could not be infinite. One might argue that the very notion of infinite utility is suspect---see for example Jeffrey 1983 and McClennen 1994.[6] Hence, the objection continues, whatever the utility of salvation might be, it must be finite. Strict finitists, who are chary of the notion of infinity in general, will agree---see. Dummett 1978 and Wright 1987. Or perhaps the notion of infinite utility makes sense, but an infinite reward could only be finitely appreciated by a human being.

3. There should be more than one infinity in the matrix. There are also critics of the Wager who, far from objecting to infinite utilities, want to see *more* of them in the matrix. For example, it might be thought that a forgiving God would bestow infinite utility upon wagerers-for and wagerers-against alike---Rescher 1985 is one author who entertains this possibility. Or it might be thought that, on the contrary, wagering against an existent God results in *negative* infinite utility. (As we have noted, some authors read Pascal himself as saying as much.) Either way, f_2 is not really finite at all, but ∞ or $-\infty$ as the case may be. And perhaps f_1 and f_3 could be ∞ or $-\infty$. Suppose, for instance, that God does not exist, but that we are reincarnated *ad infinitum*, and that the total utility we receive is an infinite sum that does not converge.

4. The matrix should have more rows. Perhaps there is more than one way to wager for God, and the rewards that God bestows vary accordingly. For instance, God might not reward infinitely those who strive to believe in Him only for the very mercenary reasons that Pascal gives, as James 1956 has observed. One could also imagine distinguishing belief based on faith from belief based on evidential reasons, and posit different rewards in each case.

5. The matrix should have more columns: the many Gods objection. If Pascal is really right that reason can decide nothing here, then it would seem that various other theistic hypotheses are also live options. Pascal presumably had in mind the Catholic conception of God---let us suppose that this is the God who either ‘exists’ or ‘does not exist’. By excluded middle, this is a partition. The objection, then, is that the partition is not sufficiently fine-grained, and the ‘(Catholic) God does not exist’ column really subdivides into various *other* theistic hypotheses. The objection could equally run that Pascal's argument ‘proves too much’: by parallel reasoning we can ‘show’ that rationality requires believing in various incompatible

theistic hypotheses. As Diderot 1875-77 puts the point: "An Imam could reason just as well this way".[7]

Since then, the point has been represented and refined in various ways. Mackie 1982 writes, "the church within which alone salvation is to be found is not necessarily the Church of Rome, but perhaps that of the Anabaptists or the Mormons or the Muslim Sunnis or the worshippers of Kali or of Odin" (203). Cargile 1966 shows just how easy it is to multiply theistic hypotheses: for each real number *x*, consider the God who prefers contemplating *x* more than any other activity. It seems, then, that such ‘alternative gods’ are a dime a dozen---or aleph one, for that matter.

Premise 2: The Probability Assigned to God's Existence

There are four sorts of problem for this premise. The first two are straightforward; the second two are more technical, and can be found by following the link to footnote 8.

1. Undefined probability for God's existence. Premise 1 presupposes that you should *have* a probability for God's existence in the first place. However, perhaps you could rationally *fail* to assign it a probability---your probability that God exists could remain *undefined*. We cannot enter here into the thorny issues concerning the attribution of probabilities to agents. But there is some support for this response even in Pascal's own text, again at the pivotal claim that "[r]eason can decide nothing here. There is an infinite chaos which separated us. A game is being played at the extremity of this infinite distance where heads or tails will turn up..." The thought could be that any probability assignment is inconsistent with a state of "epistemic nullity" (in Morris' 1986 phrase): to assign a probability at all---even 1/2---to God's existence is to feign having evidence that one in fact totally lacks. For unlike a coin that we know to be fair, this metaphorical ‘coin’ is ‘infinitely far’ from us, hence apparently completely unknown to us. Perhaps, then, rationality actually requires us to *refrain* from assigning a probability to God's existence (in which case at least the Argument from Superdominance would be valid). Or perhaps rationality does not require it, but at least *permits* it. Either way, the Wager would not even get off the ground.

2. Zero probability for God's existence. Strict atheists may insist on the rationality of a probability assignment of 0, as Oppy 1990 among others points out. For example, they may contend that reason alone *can* settle that God does not exist, perhaps by arguing that the very notion of an omniscient, omnipotent, omnibenevolent being is contradictory. Or a Bayesian might hold that rationality places no constraint on probabilistic judgments beyond coherence (or conformity to the probability calculus). Then as long as the strict atheist assigns probability 1 to God's non-existence alongside his or her assignment of 0 to God's existence, no norm of rationality has been violated.

Furthermore, an assignment of p = 0 would clearly block the route to Pascal's conclusion. For then the expectation calculations become:

$$E(\text{wager for God}) = \infty * 0 + f_1 * (1 - 0) = f_1$$

$$E(\text{wager against God}) = f_2{*}0 + f_3{*}(1 - 0) = f_3$$

And nothing in the argument implies that $f_1 > f_3$. (Indeed, this inequality is questionable, as even Pascal seems to allow.) In short, Pascal's wager has no pull on strict atheists.[8]

Premise 3: Rationality Requires Maximizing Expected Etility

Finally, one could question Pascal's decision theoretic assumption that rationality requires one to perform the act of maximum expected utility (when there is one). Now *perhaps* this is an analytic truth, in which case we could grant it to Pascal without further discussion---perhaps it is *constitutive* of rationality to maximize expectation, as some might say. But this premise has met serious objections. The Allais 1953 and Ellsberg 1961 paradoxes, for example, are said to show that maximizing expectation can lead one to perform intuitively sub-optimal actions. So too the St. Petersburg paradox, in which it is supposedly absurd that one should be prepared to pay any finite amount to play a game with infinite expectation. (That paradox is particularly apposite here.)[9]

Finally, one might distinguish between *practical* rationality and *theoretical* rationality. One could then concede that practical rationality requires you to maximize expected utility, while insisting that theoretical rationality might require something else of you---say, proportioning belief to the amount of evidence available. This objection is especially relevant, since Pascal admits that perhaps you "must renounce reason" in order to follow his advice. But when these two sides of rationality pull in opposite directions, as they apparently can here, it is not obvious that practical rationality should take precedence. (For a discussion of pragmatic, as opposed to theoretical, reasons for belief, see Foley 1994.)

Is the Argument Valid?

A number of authors who have been otherwise critical of the Wager have explicitly conceded that the Wager is valid---e.g. Mackie 1982, Rescher 1985, Mougin and Sober 1994, and most emphatically, Hacking 1972. That is, these authors agree with Pascal that wagering for God really is rationally mandated by Pascal's decision matrix in tandem with positive probability for God's existence, and the decision theoretic account of rational action.

However, Duff 1986 and Hájek 2001 argue that the argument is in fact invalid. Their point is that there are strategies besides wagering for God that also have infinite expectation---namely, *mixed* strategies, whereby you do not wager for or against God outright, but rather choose which of these actions to perform on the basis of the outcome of some chance device. Consider the mixed strategy: "Toss a fair coin: heads, you wager for God; tails, you wager against God". By Pascal's lights, with probability 1/2 your expectation will be infinite, and with probability 1/2 it will be finite. The expectation of the entire strategy is:

$$1/2{*}\infty + 1/2[f_2{*}p + f_3{*}(1 - p)] = \infty$$

That is, the ‘coin toss’ strategy has the same expectation as outright wagering for God. But the probability 1/2 was incidental to the result. Any mixed strategy that gives positive and finite probability to wagering for God will likewise have infinite expectation: "wager for God iff a fair die lands 6", "wager for God iff your lottery ticket wins", "wager for God iff a meteor quantum tunnels its way through the side of your house", and so on.

The problem is still worse than this, though, for there is a sense in which *anything* that you do might be regarded as a mixed strategy between wagering for God, and wagering against God, with suitable probability weights given to each. Suppose that you choose to ignore the Wager, and to go and have a hamburger instead. Still, you may well assign positive and finite probability to your winding up wagering for God nonetheless; and this probability multiplied by infinity again gives infinity. So ignoring the Wager and having a hamburger has the same expectation as outright wagering for God. Even worse, suppose that you focus all your energy into *avoiding* belief in God. Still, you may well assign positive and finite probability to your efforts failing, with the result that you wager for God nonetheless. In that case again, your expectation is infinite again. So even if rationality requires you to perform the act of maximum expected utility when there is one, here there isn't one. Rather, there is a many-way tie for first place, as it were.[10]

Moral Objections to Wagering for God

Let us grant Pascal's conclusion for the sake of the argument: rationality requires you to wager for God. It still does not obviously follow that you *should* wager for God. All that we have granted is that one norm---the norm of rationality---prescribes wagering for God. For all that has been said, some *other* norm might prescribe wagering against God. And unless we can show that the rationality norm trumps the others, we have not settled what we should actually do.

There are several arguments to the effect that *morality* requires you to wager against God. Pascal himself appears to be aware of one such argument. He admits that if you do not believe in God, his recommended course of action will "deaden your acuteness." One way of putting the argument is that wagering for God may require you to corrupt yourself, thus violating a Kantian duty to yourself. Clifford 1986 argues that an individual's believing something on insufficient evidence harms society by promoting credulity. Penelhum 1971 contends that the putative divine plan is itself immoral, condemning as it does honest non-believers to loss of eternal happiness, when such unbelief is in no way culpable; and that to adopt the relevant belief is to be complicit to this immoral plan. See Quinn 1994 for replies to these arguments. For example, against Penelhum he argues that as long as God treats non-believers justly, there is nothing immoral about him bestowing special favor on believers, more perhaps than they deserve. (Note, however, that Pascal leaves open in the Wager whether the payoff for non-believers *is* just, even though as far as his argument goes, it may be extremely poor.)

Finally, Voltaire protests that there is something unseemly about the whole Wager. He suggests that Pascal's calculations, and his appeal to self-interest, are unworthy of the gravity of the subject of theistic

belief. This does not so much support wagering against God, as dismissing all talk of ‘wagerings’ altogether.

What Does It Mean to "Wager for God"?

Let us now grant Pascal that, all things considered (rationality and morality included), you should wager for God. What exactly does this involve?

A number of authors read Pascal as arguing that you should *believe* in God---see e.g. Quinn 1994, and Jordan 1994a. But perhaps one cannot simply believe in God at will; and rationality cannot require the impossible. Pascal is well aware of this objection: "[I] am so made that I cannot believe. What, then, would you have me do?", says his imaginary interlocutor. However, he contends that one can take steps to cultivate such belief:

> You would like to attain faith, and do not know the way; you would like to cure yourself of unbelief, and ask the remedy for it. Learn of those who have been bound like you, and who now stake all their possessions. These are people who know the way which you would follow, and who are cured of an ill of which you would be cured. Follow the way by which they began; by acting as if they believed, taking the holy water, having masses said, etc...
>
> But to show you that this leads you there, it is this which will lessen the passions, which are your stumbling-blocks.

We find two main pieces of advice to the non-believer here: act like a believer, and suppress those passions that are obstacles to becoming a believer. And these are actions that one *can* perform at will.

Believing in God is presumably one way to wager for God. This passage suggests that even the non-believer can wager for God, by striving to become a believer. Critics may question the psychology of belief formation that Pascal presupposes, pointing out that one could strive to believe (perhaps by following exactly Pascal's prescription), yet fail. To this, a follower of Pascal might reply that the act of genuine striving already displays a pureness of heart that God would fully reward; or even that genuine striving in this case is itself a form of believing.

Pascal's Wager vies with Anselm's Ontological Argument for being the most famous argument in the philosophy of religion. As we have seen, it is also a great deal more besides.

Bibliography

- Allais, Maurice. 1953. "Le Comportment de l'Homme Rationnel Devant la Risque: Critique des Postulats et Axiomes de l'École Américaine", *Econometrica* 21: 503-546.
- Broome, John. 1995. "The Two-Envelope Paradox", *Analysis* 55: 1, 6-11.

- Brown, Geoffrey. 1984. "A Defence of Pascal's Wager", *Religious Studies* 20: 465-79.
- Cain, James. 1995. "Infinite Utility", *Australasian Journal of Philosophy*, Vol. 73, No. 3, 401-404.
- Cargile, James. 1966. "Pascal's Wager", *Philosophy*, 35: 250-7.
- Castell, Paul and Diderik Batens. 1994. "The Two-Envelope Paradox: the Infinite Case", *Analysis* 54: 46-49.
- Chalmers, David. 1997. "The Two-Envelope Paradox: A Complete Analysis?", manuscript, http://ling.ucsc.edu/~chalmers/papers/envelope.html (and envelope.ps)
- Clifford, William K. 1986. "The Ethics of Belief", *The Ethics of Belief Debate*, ed. Gerald D. McCarthy, Scholars Press.
- Conway, John. 1976. *On Numbers and Games*, Academic Press.
- Cutland, Nigel, ed. 1988. *Nonstandard Analysis and its Applications*, London Mathematical Society, Student Texts 10.
- Diderot, Denis. 1875-1877. *Pensées Philosophiques*, LIX, *Oeuvres*, ed. J. Assézat, Vol. I.
- Duff, Antony. 1986. "Pascal's Wager and Infinite Utilities", *Analysis* 46: 107-9. n
- Dummett, Michael. 1978. "Wang's Paradox", in *Truth and Other Enigmas*, Harvard University Press.
- Ellsberg, D.. 1961. "Risk, Ambiguity and the Savage Axioms", *Quarterly Journal of Economics* 25: 643-669.
- Feller, William. 1971. *An Introduction to Probability Theory and its Applications*, Vol. II, 2nd edition, Wiley.
- Flew, Anthony. 1960. "Is Pascal's Wager the Only Safe Bet?", *The Rationalist Annual*, 76: 21-25.
- Foley, Richard. 1994. "Pragmatic Reasons for Belief", in Jordan 1994b.
- Hacking, Ian. 1972. "The Logic of Pascal's Wager", *American Philosophical Quarterly* 9/2, 186-92. Reprinted in Jordan 1994b.
- Hacking, Ian. 1975. *The Emergence of Probability*, Cambridge University Press.
- Hájek, Alan. 1997a. "Review of *Gambling on God*" (Jordan 1994b), *Australasian Journal of Philosophy*, Vol. 75, No. 1, March 1997, 119-122.
- Hájek, Alan. 1997b. "The Illogic of Pascal's Wager", *Proceedings of the 10th Logica International Symposium*, Liblice, ed. T. Childers et al, 239-249.
- Hájek, Alan. 2000. "Objecting Vaguely to Pascal's Wager", *Philosophical Studies*, vol. 82.
- Hájek, Alan. 2001. "Waging War on Pascal's Wager: Infinite Decision Theory and Belief in God", manuscript.
- Jackson, Frank, Peter Menzies and Graham Oppy. 1994. "The Two Envelope 'Paradox'", *Analysis* 54: 46-49.
- James, William. 1956. "The Will to Believe", in *The Will to Believe and Other Essays in Popular Philosophy*, Dover Publications.
- Jeffrey, Richard C.. 1983. *The Logic of Decision*, 2nd edition, University of Chicago Press.
- Jordan, Jeff. 1994a. "The Many Gods Objection", in Jordan 1994b.
- Jordan, Jeff, ed.. 1994b. *Gambling on God: Essays on Pascal's Wager*, Rowman & Littlefield.
- Lewis, David. 1981. "Causal Decision Theory", *Australasian Journal of Philosophy* 59, 5-30; reprinted in *Philosophical Papers*, Volume II, Oxford University Press, 1986.
- Lindstrom, Tom. 1988. "Invitation to Non-Standard Analysis", in Cutland 1988.

- Mackie, J. L.. 1982. *The Miracle of Theism*, Oxford.
- Martin, Michael. 1983. "Pascal's Wager as an Argument for Not Believing in God", *Religious Studies* 19: 57-64.
- Martin, Michael. 1990. *Atheism: a Philosophical Justification*, Temple University Press.
- McClennen, Edward. 1994. "Finite Decision Theory", in Jordan 1994b.
- Morris, T. V. 1986. "Pascalian Wagering", *Canadian Journal of Philosophy* 16, 437-54.
- Morris, Thomas V. 1994. "Wagering and the Evidence", in Jordan 1994b.
- Mougin, Gregory, and Elliot Sober. 1994. "Betting Against Pascal's Wager", *Nous* XXVIII: 382-395.
- Nalebuff, B. 1989. "Puzzles: The Other Person's Envelope is Always Greener", *Journal of Economic Perspectives* 3: 171-91.
- Nelson, Edward. 1987. *Radically Elementary Probability Theory*, Annals of Mathematics Studies, Princeton University Press.
- Nelson, Mark T.. 1991. "Utilitarian Eschatology", *American Philosophical Quarterly*, 339-347.
- Ng, Yew-Kwang. 1995. "Infinite Utility and Van Liedekerke's Impossibility: A Solution", *Australasian Journal of Philosophy*, 73: 408-411.
- Oppy, Graham. 1990. "On Rescher on Pascal's Wager", *International Journal for Philosophy of Religion*, 30: 159-68.
- Palacios, M. Asin. 1920. "Los Precedentes Musulmanes del ‘Pari’ de Pascal", Santander.
- Pascal, Blaise. 1910. *Pascal's Pensées*, translated by W. F. Trotter.
- Penelhum, Terence. 1971. *Religion and Rationality*, Random House.
- Rescher, Nicholas. 1985. *Pascal's Wager*, Notre Dame.
- Robinson, Abraham. 1966. *Non-Standard Analysis*, North Holland.
- Ryan, John. 1945. "The Wager in Pascal and Others", *New Scholasticism* 19/3, 233-50. Reprinted in Jordan 1994 b.
- Quinn, Philip L. 1994. "Moral Objections to Pascalian Wagering", in Jordan 1994b.
- Schlesinger, George. 1994. "A Central Theistic Argument", in Jordan 1994b.
- Skalia, H. J.. 1975. *Non-Archimedean Utility Theory*, D. Reidel.
- Sobel, Howard. 1994. "Two Envelopes", *Theory and Decision*, 69-96.
- Sobel, Howard. 1996. "Pascalian Wagers", *Synthese* 108: 11-61.
- Sorensen, Roy. 1994. "Infinite Decision Theory", in Jordan 1994b.
- Swinburne, R. G.. 1969. "The Christian Wager", *Religious Studies* 4: 217-28.
- Vallentyne, Peter. 1993. "Utilitarianism and Infinite Utility", *Australasian Journal of Philosophy* 71: 212-217.
- Vallentyne, Peter. 1995. "Infinite Utility: Anonymity and Person-Centredness", *Australasian Journal of Philosophy* 73: 413-420.
- Vallentyne, Peter and Shelly Kagan. 1997. "Infinite Value and Finitely Additive Value Theory", *The Journal of Philosophy*, Vol. XCIV, 1: 5-27
- Van Liedekerke, Luc. 1995. "Should Utilitarians Be Cautious About an Infinite Future?", *Australasian Journal of Philosophy*, Vol. 73, No. 3, 405-407.
- Weirich, Paul. 1984. "The St. Petersburg Gamble and Risk", *Theory and Decision* 17: 193-202.
- Wright, Crispin. 1987. "Strict Finitism", in *Realism, Meaning and Truth*, Blackwell.

Other Internet Resources

- Stephen R. Welch's page on Pascal's Wager

Related Entries

decision theory: causal | paradox: St. Petersburg paradox | probability calculus: interpretations of

Alan Hájek
ahajek@hss.caltech.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 2, 1998
Content last modified: October 26, 2001

Stanford Encyclopedia of Philosophy
Notes to Pascal's Wager

Notes

1. Those interested in the reconstruction over the years of the text itself should consult Lafuma 1954.

2. Our demarcation of the arguments follows that of Hacking 1972, although we will differ on certain points of detail.

3. Unfortunately, he squanders this insight when he lapses back to the assumption that the probability is 1/2 shortly thereafter: "And so our proposition is of infinite force, when there is the finite to stake in a game where there are equal risks of gain and of loss, and the infinite to gain."

4. We know from Pascal's other writings that he understood the decision theoretic formula of expectation. Note, though, that this is a rather curious way of presenting it. Consider a gamble in which there is a probability p of some good outcome g, and probability 1 − p of some bad outcome b. Then the expectation e of the gamble is given by

$$e = g*p + b*(1 - p)$$

Rearranging this, we have:

$$p*(g - e) = (e - b)*(1 - p),$$

or

```
e - b     p
----- = -----
g - e   1 - p
```

That is, compared to the expectation, the shortfall of the bad outcome is proportioned to surplus of the good outcome according to the proportion of the chances of gain and loss. It is not obvious that this amounts to the same thing as what Pascal says.

5. In the basic version of decision theory that we have presented, states are assumed to be independent of actions. Evidential decision theory generalizes this. It replaces in its expectation calculation for a given

action the unconditional probabilities of states by the conditional probabilities of the states, given the action---see Jeffrey 1983. Now perhaps what you do is not independent of whether God exists. For instance, maybe God helps people wager for Him, so that P(God exists|you wager for God) > P(God exists|you wager against God). Still, the expected utility calculations are as before, provided the first conditional probability is positive and finite: infinite for wagering for God, finite for wagering against God.

Causal decision theory replaces evidential decision theory's conditional probabilities with probabilities that capture the degree of causal relevance of an action to each state. There are various versions of causal decision theory---see Lewis 1981. Using some such version would presumably not significantly affect matters here. We would just replace the assumption that your probability is positive and finite with the same assumption about whatever probability is used instead.

6. After all, infinite utilities would run afoul of the Archimedean, or continuity axiom that is commonly assumed in decision theory:

> If you prefer A to B, and prefer B to C, then there is a gamble between A and C (A with probability p, C with probability 1 − p, for some real-valued p) that you regard as equally desirable as B for sure.

For suppose that salvation, say, has infinite utility for you. You prefer salvation to $1, and prefer $1 to nothing; but there is no such gamble that rewards you with salvation if you win, and nothing if you lose, that you value at $1. Indeed, assuming that the probability of winning remains positive, you prefer the gamble to any finite reward; but if the probability of winning drops to 0, your preference discontinuously switches to the finite reward.

The issue then becomes whether continuity is a requirement on rational preference. Hájek 1997a and 2000a argues that it is not, and gives further positive arguments for allowing infinite utilities into decision theory. Sorensen 1994 likewise argues for "infinite decision theory". For a highly technical presentation of ‘non-Archimedean’ decision theory, see Skalia 1975. For related work on infinite utilities that is more philosophical, see Cain 1995, Nelson 1991, Ng 1995, Vallentyne 1993, Vallentyne 1995, Vallentyne and Kagan 1997, and van Liedekerke 1995. Some of the literature on the so-called two-envelope problem is also relevant---see, for example, Broome 1995, Castell and Batens 1994, Chalmers 1997, Jackson et al. 1994, Nalebuff 1989.

7. It should be pointed out that the rival Gods must award infinite utility for salvation in order to create a problem---otherwise they will be trumped by the ones that do. (It seems that Kali and Odin thus drop out of consideration, for example.) And to be damaging to the Wager, the alternative hypotheses about how salvation is achieved should be mutually exclusive. If there is some common core to the theistic hypotheses, and it suffices to (strive to) believe that in order to be saved, then there is no problem. For instance, it will not matter that you do not know what God's favorite real number is, if it turns out that you are saved as long as your belief is adequate in other respects. So it is crucial that salvation hinges on

getting the details of the belief right. What, then, should we believe? To settle this question, it seems we get nowhere with Pascal-style practical reasoning.

One response is that we are therefore in a position somewhat like that of Buridan's ass, unable to settle which course of action is best; and that like the ass we are better off doing something rather than nothing, and in this case that means choosing one of the theistic hypotheses, and hoping we choose the right one. So it might still be rationally required to be a theist. See Jordan 1994a for a version of this "ecumenical" response. There are at least two counter-responses. Firstly, the assumption that there are alternative Gods who offer infinite rewards really plays no role in the many-Gods objection argument. All that matters is that there are sources of infinite reward besides Pascal's God. These sources could even be inanimate---as it might be, supreme pleasure machines, which offer infinite utility irrespective of one's beliefs. Secondly, one of the alternative Gods might punish those who wager for him, and reward those who don't---see Martin's 1983 "perverse master".

At this point it can be replied that these various other hypotheses lack the backing of tradition that genuine religions have, and thus should be disregarded---see especially Jordan 1994a and Schlesinger 1994. More precisely, these other hypotheses should be assigned zero (or perhaps at most infinitesimal) probability, so that they do not upset Pascal's expectation calculations. The debate then turns once again on what exactly rationality requires of one's probability assignments.

8. Here are the third and fourth problems for Premise 2.

3. Infinitesimal probability for God's existence.

One might reply that you can rationally assign infinitesimal probability to God's existence---see e.g. Oppy 1990. The argument might run, for example, that there are infinitely many possible Gods to consider (see our discussion of the many Gods objection), and for some infinite subset of them that includes Pascal's God, rationality does not favor any one over the rest. Treating them even-handedly then requires assigning infinitesimal probability to each. Or again, a Bayesian might say that you could coherently assign to God's existence an infinitesimal probability, provided that you also assign a probability to God's non-existence that falls short of 1 by the same infinitesimal.

It is remarkable that Pascal anticipated the notion of infinitesimal probability, when he says: "if there were an infinity of chances, of which one only would be for you, you would still be right in wagering one [life] ... if there were an infinity of an infinitely happy life to gain." But what he says here is far from obvious. If ∞ is a legitimate utility value, then offhand it would seem that $1/\infty$ is a legitimate probability value, and indeed it seems to be the very one that he is considering. However, then we have:

$$E(\text{wager for God}) = \infty * (1/\infty) + f_1 * [1 - (1/\infty)] \approx 1 + f_1$$

And it is not clear that this should exceed f_3.

All of this treats ∞ as if it is a number, subject to ordinary arithmetic operations, such as taking reciprocals, multiplying and adding. Perhaps, for example, $\infty*(1/\infty)$ is not defined, much as $\infty - \infty$ is not. But that is just another way in which a probability of $(1/\infty)$ might thwart Pascal's reasoning. We will say more below about infinite numbers for which such arithmetic operations are unproblematic.

4. Vague probability for God's existence

So far we have presupposed that probability assignments are sharp. However, Pascal's argument is addressed to us---mere humans. And it is apparently a fact about us that our belief states are irremediably vague: we cannot assign probability, precise to indefinitely many decimal places, to all propositions. Perhaps, then, rationality permits us to assign vague probability to God's existence. If it moreover permits us to assign it probability that is vague over an interval that includes 0, then the Wager fails---see Hájek 2000. Indeed, Pascal's claim that "[reason] can decide nothing here" might be thought to support a probability assignment to God's existence that is vague over the entire [0, 1] interval. [Return to text]

9. One could also insist that rational choices must be ratifiable (à la Jeffrey 1983 or Sobel 1996), and that the act of maximum expectation might not be.

The usual rationale for maximizing expectation comes from the various laws of large numbers. Their content is roughly that under suitable circumstances, in the limit, one's average reward tends to the expectation; and of course one wants to maximize one's average reward. But the strong law of large numbers assumes that the expectation is finite, and since the expectation of wagering for God is putatively infinite, it clearly cannot be appealed to here. (See e.g. Feller 1971, 236.) Perhaps an appeal to the weak law of large numbers, which allows infinite expectation, would suffice. But being a limit theorem, it concerns infinitely long runs of trials. Far from having such a long run here, we have just a single-shot decision problem. This is a decision that you do not get to repeat. This is not so troubling, perhaps, when the variance (a measure of the spread of the distribution of outcomes) is small, so that getting an outcome close to the expectation is probable; but what about when the variance is large?

This brings us to yet another problem for Pascal's third premise. To be sure, the expectation of wagering for God is infinite, if we accept Pascal's earlier assumptions; but so too is the variance. Expectation does not seem to be such a good guide to choiceworthiness when the variance is large---for what one might end up getting can then be much worse than the expectation---let alone when the variance is infinite. (See Weirich 1984 and Sorensen 1994 for versions of this last point.) Indeed, the lower one makes f_2 (or more generally, some highly dispreferred outcome), the less compelling premise 3 seems; and the lower one makes the probability of salvation (or more generally, of some highly desired outcome), the less compelling premise 3 seems. Yet consistent with premise 1, f_2 could be (almost) as low as one likes, and consistent with premise 2, the probability of salvation could be (almost) as low as one likes.

10. Schlesinger 1994 offers a tie-breaking criterion: "try and increase the probability of obtaining the prospective prize" (97). Of course, "the prospective prize" here is salvation. Schlesinger is suggesting

that decision theory should be supplemented with a new principle. In our present case, it amounts to this: rationality requires you to perform the action that maximizes your probability of salvation. This clearly rules out the coin-tossing strategy, the die-tossing strategy, and all the other mixed strategies, since these have lower probabilities of your achieving salvation than outright wagering for God does. Sorensen 1994 objects to Schlesinger's new principle as being ad hoc. In any case, Pascal does not appeal to the principle in his argument. As it stands, the argument is apparently invalid.

The problem is that multiplying ∞ by any positive, finite probability again yields ∞. Let us call this property of ∞ *reflexivity under multiplication* (by such a probability). Such reflexivity is at once the strength of the Wager (for then Pascal does not need to say anything more about your probability of God's existence), and its weakness (for then all the various mixed strategies get maximal expectation also). One could try to fix the weakness, while saving as much one can of the strength. This would involve finding a utility for salvation that is not reflexive under multiplication, yet which is still sufficiently large to swamp your probability, whatever it is, in the expectation calculation.

For instance, if the utility of salvation were enormous, but finite, then the mixed strategies would yield lower expectation than outright wagering for God (multiplying that utility by 1/2, 1/6, etc. makes a difference). And the utility could be made enormous enough to offset any actual person's probability assignment, however small (provided it is positive and finite), so that the expectation of outright belief is maximal for everybody. Or suppose that the utility of salvation were an infinite number that is not reflexive under multiplication. Consider, for example, the infinite numbers of non-standard analysis (see Robinson 1966, Nelson 1987), or the surreal infinite numbers of Conway 1976. Multiplication of such a utility by a positive, finite probability (less than 1) yields another, smaller infinite number. So the expectation of wagering for God again exceeds that of wagering against God, whatever your probability is (provided it is positive and finite), and also that of each mixed strategy. See Hájek 2000a for further devices along these lines.

These proposals appear to yield valid arguments for wagering for God, where Pascal's argument was invalid. The trouble is that they do not seem adequately to capture Pascal's reasoning. He writes: "Unity added to infinity adds nothing to it". Let us call this property of infinity *reflexivity under addition*. We can see why Pascal would want the utility of salvation to be reflexive under addition: salvation is supposed to be the best possible thing. But if that utility is finite, or non-standard infinite, or surreal infinite, then adding one to it does make a difference. What is wanted, then, is the seemingly impossible: a representation of the reward of salvation that is reflexive under addition (so that it cannot be bettered), but not reflexive under multiplication by positive, finite probabilities (so that the mixed strategies can be distinguished in expectation from outright belief).

Alan Hájek
ahajek@hss.caltech.edu

First published: October 26, 2001
Content last modified: October 26, 2001

The St. Petersburg Paradox

The St. Petersburg game is played by flipping a fair coin until it comes up tails, and the total number of flips, n, determines the prize, which equals $\$2^n$. Thus if the coin comes up tails the first time, the prize is $\$2^1 = \$2$, and the game ends. If the coin comes up heads the first time, it is flipped again. If it comes up tails the second time, the prize is $\$2^2$, = \$4, and the game ends. If it comes up heads the second time, it is flipped again. And so on. There are an infinite number of possible ‘consequences’ (runs of heads followed by one tail) possible. The probability of a consequence of n flips (‘P(n)’) is 1 divided by 2^n, and the ‘expected payoff’ of each consequence is the prize times its probability. The following table lists these figures for the consequences where n = 1 ... 10:

n	P(n)	Prize	Expected payoff
1	1/2	\$2	\$1
2	1/4	\$4	\$1
3	1/8	\$8	\$1
4	1/16	\$16	\$1
5	1/32	\$32	\$1
6	1/64	\$64	\$1
7	1/128	\$128	\$1
8	1/256	\$256	\$1
9	1/512	\$512	\$1
10	1/1024	\$1024	\$1

The ‘expected value’ of the game is the sum of the expected payoffs of all the consequences. Since the expected payoff of each possible consequence is \$1, and there are an infinite number of them, this sum is an infinite number of dollars. A rational gambler would enter a game iff the price of entry was less than the expected value. In the St. Petersburg game, any finite price of entry is smaller than the expected value of the game. Thus, the rational gambler would play no matter how large the finite entry price was. But it seems obvious that some prices are too high for a rational agent to pay to play. Many commentators agree with Hacking's (1980) estimation that "few of us would pay even \$25 to enter such a game." If this is correct, then something has gone wrong with the standard decision-theory calculations of expected value above. This problem, discovered by the Swiss eighteenth-century mathematician Daniel Bernoulli (1738;

English trans. 1954) is the St. Petersburg paradox.

Decreasing Marginal Utility

Bernoulli responded to this problem with the observation that the calculations err by adding expected payoffs in dollars, whereas what should be added are the expected utilities of each consequence. He proposed the widely-accepted principle that (roughly speaking) money has a decreasing marginal utility, and suggested that a realistic measure of the utility of money might be given by the logarithm of the money amount. Here are the first few lines in the table for this gamble if utiles = log($):

n	P(n)	Prize	Utiles	Expected Utility
1	1/2	$2	0.301	0.1505
2	1/4	$4	0.602	0.1505
3	1/8	$8	0.903	0.1129
4	1/16	$16	1.204	0.0753
5	1/32	$32	1.505	0.0470
6	1/64	$64	1.806	0.0282
7	1/128	$128	2.107	0.0165
8	1/256	$256	2.408	0.0094
9	1/512	$512	2.709	0.0053
10	1/1024	$1024	3.010	0.0029

The sum of expected utilities is not infinite: it reaches a limit of about 0.60206 utiles (worth $4.00). The rational gambler, then, would pay any sum less than $4.00 to play.

This response to the paradox is, however, unsatisfactory. Let us agree that money has a decreasing marginal utility, and accept (for the purposes of argument) that a reasonable calculation of the utility of any dollar amount takes the logarithm of the amount in dollars. The St. Petersburg game as proposed, then, presents no paradox, but it is easy to construct another St. Petersburg game which is paradoxical, merely by altering the dollar prizes. Suppose, for example, that instead of paying $\$2^n$ for a run of n, the prize were \$10 to the power 2^n. Here is the table for this game:

n	P(n)	Prize	Utiles of Prize	Expected utility
1	1/2	$\$10^2$	2	1
2	1/4	$\$10^4$	4	1
3	1/8	$\$10^8$	8	1
4	1/16	$\$10^{16}$	16	1
5	1/32	$\$10^{32}$	32	1
6	1/64	$\$10^{64}$	64	1
7	1/128	$\$10^{128}$	128	1
8	1/256	$\$10^{256}$	256	1
9	1/512	$\$10^{512}$	512	1
10	1/1024	$\$10^{1024}$	1024	1

This version contains much larger prizes than the original version, and one would presumably be willing to pay more to play this version than the original. But the expected value of this game - the sum of the infinite series of numbers in the last column - is infinite, and the paradox returns.

Of course, it is not clear how in fact dollar values relate to utility, but we can imagine a generalized paradoxical St. Petersburg game (suggested by Paul Weirich, 1984, following Menger, 1967) which offers prizes in utiles at the rate of 2^n *utiles* for a run of n, however that number of utiles is to be translated into dollars or other goods. This game would have infinite expected value, and the rational gambler should pay any amount, however large, to play. For simplicity, we shall ignore the generalized version of the game, and continue to discuss it in terms of the original dollar prizes, recognizing, however, that the diminishing marginal utility of dollars may make some revision of the prizes necessary to produce the paradoxical result.

Risk-Aversion

Consider the following argument. The St. Petersburg game offers the possibility of huge prizes. A run of forty would, for example, pay a whopping \$1.1 trillion. Of course, this prize happens rarely: only once in about 1.1 trillion times. Half the time, the game pays only \$2, and you're 75% likely to wind up with a

payment of $4 or less. Your chances of getting more than the entry price of $25 which Hacking suggests are less than one in 25. Very low payments are very probable, and very high ones very rare. It's a foolish risk to invest more than $25 to play.

This sort of reasoning is appealing, and may very well account for intuitions that agree with Hacking's. Many of us are risk-averse, and unwilling to gamble for a very small chance of a very large prize, because the chance is so small. Weirich claims that this sort of consideration in fact solves the St. Petersburg paradox. He offers a complicated way (which we need not go into here) of including a risk-aversion factor in our rational calculations, with the result that there is a finite upper limit to the rational entrance fee for the game.

But there are objections to this approach. For one thing, a factor for risk-aversion is not a generally applicable consideration in making rational decisions, because some people are not risk averse. In fact, some people may enjoy risk. What should we make, for example, of those people who routinely play state lotteries, or who gamble at pure games of chance in casinos? (In these games, the entry fee is greater than the expected utility.) It's possible to dismiss such behaviour as merely irrational, but sometimes these players offer the explanation that they enjoy the excitement of risk. In any case, it's not at all clear that risk-aversion can explain why the St. Petersburg game would be widely intuited to have a fairly small maximum rational entry fee, while so many people at the same time are not averse to the huge risk entailed by the very small expected probability of large prizes in lotteries.

But for the purposes of argument, let's assume that risk-aversion is what's responsible for the rational intuition that the appropriate entrance-fee for the St. Petersburg game is finite and small. But this will not make the paradox go away, for we can again adjust the prizes to take account of this risk-aversion.

Suppose you don't like to gamble, and wouldn't risk an entry fee in a game that offers a small possibility of a large prize, even when the odds were in your favour. For example, imagine that you were offered a lottery ticket costing $1, which gave you a one-in-ten chance at a prize of $20. Playing costs you utility, because you hate risk. But presumably, we could compensate you for this utility-loss by making the prize even bigger. Maybe you would invest $1 for a one-in-ten chance at making $100. If not, how about $1000? It appears that there is some prize large enough to compensate you for your risk aversion.

Now let's imagine that you consider the St. Petersburg game, and suppose that you're willing to pay an entrance fee of only $20 to play. The reason you're not willing to go higher is your risk-aversion. We can imagine that the increasing risk of large payments subtracts from their utility, and the result is that the last column contains numbers that decrease as probabilities shrink, and the sum of the last column reaches a limit - perhaps $25. But now, the game can be reformulated to repay you for the risk inherent in each outcome, by correspondingly increasing the prizes. For example, suppose we square each dollar-prize in compensation for the increasing risk - the lower probability - of the larger prizes. If this doesn't provide sufficient compensation for your risk-aversion, then we can make the prizes even higher. In any case, there seems to be some prize scheme huge enough to compensate you for your risk-aversion - one which makes the dollar utility of each prize minus its risk-factor equal 1 utile. A game with these larger prizes is again paradoxical.

But Weirich argues that offering increased prizes cannot sufficiently compensate for risk-aversion in such a way as to make the sum of the series unlimited. He appears to suggest that increasing the prize for an outcome may increase one's cost in terms of dread of risk. In the lottery example, then, increasing the prize to $1000 would correspondingly increase the risk for you, so you still wouldn't bet. No matter how high a prize you are offered, you still are unwilling to buy the ticket for $1, because the higher prizes raise the risk for you. Putting it ‘picturesquely,’ he says, ‘there is some number of birds in hand worth more than any number of birds in the bush.’

But one might doubt that risk-aversion works this way - or, anyway, that this sort of risk-aversion can be justified as rational. It's highly implausible to claim that an increase in prize-size increases the risk of a game. In the lottery example, the only sort of risk-aversion that would make one refuse to play no matter how high the prize is pathological, not rational. There must be some prize which is so valuable to any rational but risk-averse person that the person would see it as compensating him for the risk of $1, (where that dollar has the usual small utility). If someone prefers $1 worth of birds in hand to any value of birds in the bush, then that person needs psychiatric help; this is not a rational decision strategy.

One might argue, however, for a risk-aversion factor such that increase in prizes makes certain games more attractive, but never attractive enough to override the risk factor. The most obvious way to take risk-aversion into consideration in calculating the utility of a game would be to add the negative utility of its risk to the positive utility of each prize, as if its risk were a negative aspect of the prize. On this way of calculating, it is always possible to compensate for risk by increasing the utility of the payoff. But Weirich's proposal appears to make risk a function of the whole gamble and of one's current utility level, in such a way that no addition to the prizes can make a game desirable to a risk-averse agent. This might seem implausible: if one's risk-aversion to a game is finite, and does not increase merely because of increase in payoffs, and if the utility of the prizes can be increased without limit, it would seem that some prize-increase can always compensate for the risk-aversion, however it is reasonably calculated. But if prize-increase, while increasing the expected utility (ignoring the risk-factor) of the whole game, nevertheless cannot make it sufficient to overcome the finite risk-aversion, then something else (such as the diminishing marginal utility of prizes, or an upper limit on their utility) is operating. These possibilities are discussed elsewhere in this article.

But the St. Petersburg game is supposed to justify even an enormously high entry price, so the lottery example is not precisely germane. Let's consider examples with a high entry price, for example, your entire life-savings. Would it be rational always to refuse to risk this, no matter what the gamble is? It doesn't seem so. When you deposit your life-savings in a solid bank for a year, you are in fact accepting a gamble. There's a very high probability of the consequence that at the end of the year, you can get your savings back with interest, but there's also an extremely low probability that the bank and the deposit insurance will both collapse, and you'll be wiped out. Someone who refused to run this very tiny risk no matter how high the interest and how low the probability of disaster is clearly irrational. Everyone who crosses a street is, in effect, gambling his life, because crossing a street increases, to a small extent, the risk of being run down and killed. But to refuse to cross any street on these grounds is irrational. This sort of risk-aversion, when generally applied, would paralyze anyone. It is central to rationality that one take

account of the actual risks, and run suitably small ones.

The counter argument we have been considering is that risk-aversion is irrational when it refuses to gamble a small entry-price for no matter how high a prize, or when it refuses to gamble a large entry-price for no matter how high a probability of prize. But this does not answer all St. Petersburg objections, for here we imagine a gamble with a large entry-price and a small probability of large prizes. The most compelling examples of the rational unacceptability of risk no matter how high the prize, are the ones in which the entry price is high and the prize improbable. Imagine, for example, that you are risk averse, and are offered a gamble in which the entry price is your life-savings of (say) $100,000, and the chances of the prize are one-in-a-million. It seems rational to refuse, no matter how huge the prize. The reason for this is worth considering.

Classical decision theory says that, for this gamble to be rational, the prize must be enormous: in the imagined case, at least one million times the value of your life-savings - more than a hundred billion dollars. Compensating you for your risk-aversion by increasing the size of this prize makes it even larger - two hundred billion? a trillion? - so huge that our intuitions are inadequate to appreciate such a value. What is worth much more than a million times your life-savings? You don't know. Your intuitions boggle when considering this gamble. The diminishing marginal utility of money and of ordinary goods operates here as well. You might suppose that nothing would give you that much utility. But the facts that the world happens not to contain such huge utilities, or that one's intuitions get unreliable when considering them, are not difficulties with classical decision theory per se. More will be said about this below, when the argument will be proposed that these these sorts of practical considerations don't show that there's something wrong with classical decision theory.

Let us, then, increase the prizes in the St. Petersburg game to compensate a rational potential player for his risk-aversion, and the game once again has infinite expected value for that person.

An Upper Bound on Utility

The two reformulations of the game proposed so far share the feature that the dollar values of the prizes are increased as compensation (in the first case, for the diminishing marginal value of money, and, in the second case, for their improbability and risk-aversion). In both cases, it is assumed that the utility of each outcome can be increased without limit; but perhaps this assumption is incorrect, and there is an upper limit on the utility of the prizes. Then the sum of the series will reach a limit. In his classical treatment of the problem, Menger argues that the assumption that there is an upper limit to utility is the only way that the paradox can be resolved. Assume, for example, that utility = dollar value, except with an upper limit of 100 utiles. The chart for the game then looks like this:

n	P(n)	Prize	Utiles of Prize	Expected utility
1	1/2	$2	2	1
2	1/4	$4	4	1

3	1/8	$8	8	1
4	1/16	$16	16	1
5	1/32	$32	32	1
6	1/64	$64	64	1
7	1/128	$128	100	0.78
8	1/256	$256	100	0.391
9	1/512	$512	100	0.195
10	1/1064	$1064	100	0.098

The sum of the infinite series in the right-hand column reaches a limit of about 7.56, and the rational entry price is anything under $7.56.

The assumption that maximum utility is reached by any dollar-prize over $100 is implausible because it means that the value of $100, $1000, and $1,000,000 are all the same - the maximum. A more plausible point for maximum utility of dollars is much higher. Setting it at 16,000,000 makes the maximum rational entry price of the game close to $25, which is Hacking's guess at what our intuitions would accept.

Some people think that it is reasonable to set an upper limit on utility. Russell Hardin (1982), for example, calls this assumption "compelling in its own right." William Gustason (1994) suggests that one restrict the expected value concept by stipulating that values of any consequence have an upper bound. Richard Jeffrey (1983) agrees.

But the idea of an upper limit on utility might not be seen to be compelling in its own right. Note that this idea must be distinguished from the diminishing marginal value of money. Perhaps you find it reasonable to think that, once one had (say) $16,000,000 in the bank, you'd be able to buy anything you could possibly want; but this is not to say that that sum of money provides the maximum permissible utility. We can readily imagine someone with that amount of money - or any amount of money - still short of utility, due to lack of certain goods that money can't buy. What the idea of an upper limit on utility means is that there is some amount of utility which is so high that no additional utility is possible - that nothing additional adds any value at all. Imagine someone with all the wealth he could use: still he might have unfulfilled desires, for example, that his friends and relations be as fortunate as he. If this desire were fulfilled, then he might still desire that strangers be as fortunate; and that there be more people on earth than there currently were, to share his happiness, and more populated planets full of happy people. How many more? Why, the more the better - indefinitely more. If there is an upper limit on utility, then there is some finite amount of utility which is maximally good, an amount for which one would rationally trade anything else. It doesn't appear plausible to think that there is any such amount.

One might imagine that some people have an upper limit on the utility they can enjoy - people who have a finite number of desires, and whose desires can each be completely satisfied by some finite state. For these people, the utility of prizes does not increase without limit, and the St. Petersburg game has some

finite expected utility. Do such people exist? This is an empirical question. In any case, there surely are some people with some ‘the-more-the-better’ desires, and the theory of rational choice ought not to be restricted by the empirical and doubtful propositions that there aren't any, and that value cannot increase without limit. And these propositions are surely insufficiently well-founded to invoke to solve the St. Petersburg paradox.

Gustason says that "the upshot of the paradox is that if there is such a thing as an infinite value, then acts and consequences that involve it are beyond the scope of the expected value concept." Jeffrey states that the evaluation theory we are applying here has "from its inception...been closely tied to the notion that desirabilities are ... bounded." But the fact that the theory wasn't designed with such a result in mind is not a very good reason to try to resist its application in this case. The main reason both authors give for excluding unbounded-desirability games is that otherwise the St. Petersburg game has infinite expected utility. But this ad-hoc rationale is not compelling unless one can't bear this result. The acceptability of this result will be considered later.

Hardin offers the opinion that whether utility is bounded "is more a factual than a logical issue," and that its invocation to resolve the St. Petersburg paradox "is to grant that the paradox is not an antinomy." He may mean that the difficulty posed by the game is a result of a factual assumption that utility is unbounded (and not merely by its logical features), and can be removed by rejecting that assumption. But if one finds no no difficulty posed by the game, one is not tempted to reject the assumption.

Finitely Many Consequences

Gustason suggests that, in order to avoid the St. Petersburg problem, one has the choice between two restrictions on the expected value concept:

> (a) Each act has only finitely many consequences, or
>
> (b) Values must be ‘bounded,’ i.e., there are numbers n and m such that no value to be assigned a consequence exceeds n or is less than m.

He points out that imposing either restriction will suffice to rule out the St. Petersburg result. If one resists imposing restriction (b), in this case, by setting an upper bound to the value of consequences, perhaps restriction (a) might be found plausible.

One way to impose restriction (a) is merely to insist that the St. Petersburg game fails to meet it, so its value, and fair entrance price, cannot be calculated using standard expected value theory. How then, if at all, can it be calculated? Where does the intuition that $25 is too much to pay come from (if anywhere)? How (if at all) can it be justified?

Another way is to assume that the way the game will take place is not exactly as described, and that there are some possible very long strings that would never be carried out - i.e., that there is only a finite number

of prizes to be considered when calculating the expected value of the game. Presumably, this would be applied by setting some upper limit L to the number of flips which would be considered; after a run of L heads in a row, the game would be terminated and payment made for the run so far, despite the fact that tails hadn't yet come up. If L were set at 25, then the game would have an expected value of $25, and that would be the maximum entry price which a rational agent would pay to play (as in Hacking's intuition). Do we, perhaps unconsciously, assume that any run of 25 heads would be truncated, and paid off, at that point?

Many authors have pointed out that, practically speaking, there must be some point at which a run of heads would be truncated without a final tail. For one thing, the patience of the participants of the game would have to end somewhere. If you think that this sets too narrow a limit L, consider the considerably higher limit set by the life-spans of the participants, or the survival of the human race; or the limit imposed by the future time when the sun explodes, vaporizing the earth. Any of these limits produces a finite expected value for the game, but sets an L which is higher than 25; what, then, explains Hacking's $25 intuition?

Another fact that would set a limit on L is the finitude of the bankroll necessary to fund the game. Any casino that offers the game must be prepared to truncate any run that, were it to continue, would cost them more than the total funds they have available for prizes. A run of 25 would require a prize of a mere $33,554,432, possibly within the reach of a larger casino. A run of 40 would require a prize of about 1.1 trillion dollars.

Other facts make an upper limit L plausible, such as the limit on the amount of money available in the world. Perhaps all these financial limits can be overridden if we conceive of the game's being offered by a state capable of printing all the money it wanted to. This state could pay any prize whatever; but printing up and handing out a huge amount of cash would create havoc with any economy, so no rational state would.

Hardin claims that "the slightest bit of realism is sufficient to do in the St. Petersburg Paradox." But is the slightest bit of realism a justifiable consideration? The fact of which we are sure, that some upper limit on L, and thus a finite number of possible consequences of the game, would certainly be imposed, does not really solve the St. Petersburg problem because it does not show that the expected value of the game as described is not infinite. After all, any game with a limit L is not the game we have been talking about. Our question was about the St. Petersburg game, not about its relative.

One might argue: we are considering the St. Petersburg game, but under realistic conditions. Realistically, the game would be truncated, whether this is mentioned in its rules or not. Thus there is a finite and realistic price for entry. But if this is the case, why isn't the game offered, with an entry price somewhat above this (to produce a profit in the long run) by casinos (who after all, are quite realistic)?

Do these realistic considerations show that the genuine St. Petersburg game - exactly as originally described - can never be encountered in real life? Jeffrey says: "Put briefly and crudely, our rebuttal of the

St. Petersburg paradox consists in the remark that anyone who offers to let the agent play the St. Petersburg game is a liar, for he is pretending to have an indefinitely large bank."

It can be quibbled that Jeffrey is not exactly right: that someone can offer a game even though he is aware that there's a possibility that this offer involves the possibility of requiring consequences he cannot fulfill. Compare my offer to drive you to the airport tomorrow. I realise that there's a small possibility my car will break down between now and then, and thus that I'm making an offer I might not be able to fulfill. But the conclusion is not that I'm not really offering what I appear to be. If someone invites you to play St. Petersburg, we can't conclude that he's in fact not offering the St. Petersburg game, that he's really offering some other game.

Real casinos right now play games that offer the extremely remote possibility of continuing too long for anyone to complete, or of prizes too large to be managed. Casinos can go ahead and play these games anyway, confident that the risk of running into an impossible situation is very very small. They need not lose any sleep worrying about incurring a debt they can't manage. They live, and prosper, on probabilities, not certainties.

If these considerations are persuasive, then what Jeffry gives then is not a rebuttal of the paradox. In effect, he accepts the fact that the game offers the possibility of indefinitely large payoffs. The reason the game is not offered by casinos is that they realise that sooner or later (probably much later) the game will bankrupt them. This is correct reasoning - but it is done using the ordinary, general theory of choice. When casinos reason about the game, they do not decide that, since ordinary theory shows that the game has infinite value, ordinary theory should be restricted to exclude its consideration.

There are other reasons why we should not restrict theory to exclude consideration of the game. This ruling, in order to be theoretically acceptable, ought not merely rule out the St. Petersburg game in particular, ad hoc; it ought to be general in scope. And if it is, it will also rule out perfectly acceptable calculations. Michael Resnik (1987) notes that utility theory "is easily extended to cover infinite lotteries, and it must be in order to handle more advanced problems in statistical decision theory" but he gives no examples.

Imagine a life insurance policy bought for a child at its birth, which pays to the child's estate, when it eventually dies, $1000 for each birthday the child has passed, without limit. What price should an insurance company charge for this policy? (For simplicity, we shall ignore possible effects of inflation, and profits from investing the entry price.) Standard empirically-based mortality charts give the chances of living another year at various ages. Of course, they don't give the chances of surviving another year at age 140, because there's no empirical evidence available for this; but a reasonable function to extend the mortality curve indefinitely beyond what's provided by available empirical evidence can be produced; this curve asymptotically approaches zero. On this basis, ordinary mathematical techniques can give the expected value of the policy. But note that it promises to pay off without limit. If we think that, for each age, there is a (large or small) probability of living another year, then there are an indefinitely large number of consequences to be considered when doing this calculation, but mathematics can calculate the limit of this infinite series; and (ignoring other factors) an insurance company will make a profit, in the

long run, buy charging anything above this amount. There's no problem in calculating its expected value.

This insurance policy (call it Policy 1) offers an indefinite number of outcomes; but consider a different one (call it Policy 2) which would truncate the series at age 140, and offer only 140 outcomes. The probability of reaching age 140 is so tiny that the difference in expected value between the two policies is negligible, a tiny fraction of 1 cent. If you don't like infinite lotteries, you might claim that Policy 1 is ill-formed, and suggest substitution of Policy 2, pointing out that the expected value of this one is, for all practical purposes, equal to that of Policy 1. But note that your judgment that the two are virtually identical in expected value depends on your having calculated the expected value of Policy 1. So your statement presupposes that the expected value of Policy 1 is calculable, after all.

Infinite Value?

The St. Petersburg game is sometimes dismissed because it is has infinite expected value, which is thought not merely practically impossible, but theoretically objectionable - beyond the reach even of thought-experiment. But is it?

Imagine you were offered the following deal. For a price to be negotiated, you will be given permanent possession of a cash machine with the following unusual property: every time you punch in a dollar amount, that amount is extruded. This is not a withdrawal from your account; neither will you later be billed for it. You can do this as often as you care to. Now, how much would you offer to pay for this machine? Do you find it impossible to perform this thought-experiment, or to come up with an answer? Perhaps you don't, and your answer is: any price at all. Provided that you can defer payment for a suitable time after receiving the machine, you can collect whatever you need to pay for it from the machine itself.

Of course, there are practical considerations: how long would it take you to collect, say, a trillion dollars from the machine, if this were its price? Would you be worn out or dead by then? Any bank would be crazy to offer to sell you an infinite cash machine, and unfortunately I seem to have lost the address of the crazy bank which has made this offer. Anyway, there appears to be nothing wrong with this thought experiment: it imagines an action (buying the machine) with no upper limit on expected value. We easily ignore practical considerations when calculating the expected value (in this case, merely potential withdrawals minus purchase price), which is infinite.

Do your intuitions tell you to offer (say) $25 at most for this machine? I doubt that they do. But the only difference between this machine and a single-play St. Petersburg game is that this machine guarantees an indefinitely large number of payouts, while the game offers a one-time lottery from among an indefinitely large number of possible payouts, each with a certain probability. The only difference between them is the probability factor: the same difference that exists between a game which gives you a guaranteed prize of $5, and one which gives you half a chance of $10, and half a chance of $0. The expected value of both the St. Petersburg game and the infinite cash machine are both indefinitely large. You should offer any price at all for either. It appears, then, that the notion of infinite expected value is perfectly reasonable.

In a sense, the counter-intuitiveness of the St. Petersburg result is a special case of a general and familiar objection to classical decision theory. Someone might object that it would be perfectly rational for you to prefer a guaranteed prize of $5 to a gamble which offers half a chance of $10, and half a chance of $0, despite the theory's claim that they have equal value. If you intuit that an infinite cash machine has infinite expected value, but the St. Petersburg game does not, you're probably relying on this more general objection to classical theory. Arguments can be made that your preference of the guaranteed prize to the gamble is irrational; or attempts can be made to "fix" the theory to account for the rationality of this preference (for example, by allowing adjustments for risk-aversion.) However this more general "problem" with classical theory is dealt with, it is not a problem with St. Petersburg alone, and making ad-hoc fixes to theory to rule out St. Petersburg will not help in dealing with this more general problem.

Theory and Practicality

The St. Petersburg game is one of a large number of examples which have been brought against standard (unrestricted) Bayesian decision theory. Each example is supposed to be a counter-example to the theory because of one or both of these features: (1) the theory, in the application proposed by the example, yields a choice people really do not, or would not make; thus it is descriptively inadequate. (2) the theory, in the application proposed by the example, yields a choice people really ought not to make, or which a fully, ideally rational person, would not make; thus it is normatively inadequate.

If you see standard theory as normative, you can ignore objections of the first type. People are not always rational, and some people are rarely rational, and an adequate descriptive theory must take into account the various irrational ways people really do make decisions. It's no surprise that the classical rather a-prioristic theory fails to be descriptively adequate, and to criticize it on these grounds rather misses its normative point.

T he objections from standpoint (2) need to be taken more seriously; and we have been treating the responses to St. Petersburg as cases of this sort. Various sorts of "realistic" considerations have been adduced to show that the result the theory draws in the St. Petersburg game about what a rational agent should do is incorrect. It's concluded that the unrestricted theory must be wrong, and (sometimes) that it must be restricted to exclude consideration of the game as invented. We'll now consider the general plausibility of restricting the theory in these ways.

When considering the plausibility of restricting expected value calculations in various ways that would rule out the St. Petersburg calculation, Amos Nathan (1984) remarks, "it ought, however, to be remembered that important and less frivolous application of such games have nothing to do with gambling and lie in the physical world where practical limitations may assume quite a different dimension." Nathan doesn't mention any physical applications of analogous infinite value calculations. But it's nevertheless plausible to think that imposing restrictions on theory to rule out St. Petersburg bath water would throw out some valuable babies as well.

Any theoretical model is an idealization, leaving aside certain practicalities. "From the mathematical and

logical point of view," observes Resnick, "the St. Petersburg paradox is impeccable." But this is the point of view to be taken when evaluating a theory per se (though not the only point of view ever to be taken). By analogy, the aesthetic evaluation of a movie does not take into account the facts that the only local showing of the movie is far away, and that finding a baby sitter will be impossible at this late hour. If aesthetic theory tells you that the movie is wonderful, but other considerations show you that you shouldn't go, this isn't a defect in aesthetic theory. Similarly, the mathematical/logical theory for explaining ordinary casino games is not defective because it ignores practicalities such as a particular limit on a casino's bankroll, or on participants' patience.

There are all sorts of practical considerations which must be considered in making a real gambling decision. For example, in deciding whether to raise, see, fold, or cash in and go home, in a particular poker game, you must consider not only probability and expected value, but also the facts that it's 5 A.M. and you are cross-eyed from fatigue and drink; but it's not expected that classical decision theory has to deal with these.

The St. Petersburg game commits participants to doing what we know they will not. The casino may have to pay out more than it has. The player may have to flip a coin longer than physically possible. But this may not show a defect with choice theory. Classical unrestricted theory is still serving its purpose, which is modeling the abstract ideal rational agent. It tells us that no amount is too great to pay as a ideally rationally acceptable entrance fee, and this may be right. What it's reasonable for real agents, limited in time, patience, bankroll, and imaginative capacity to do, given the constraints of the real casino, the real economy, and the real earth, is another matter, one that the theoretical core of decision theory can be forgiven for not specifying. From this point of view, the St. Petersburg ‘paradox’ does not, after all, point out any defect with classical decision theory, and is not a paradox after all.

Bibliography

Works Cited

- Bernoulli, Daniel: 1738, "Exposition of a New Theory on the Measurement of Risk", *Econometrica* 22 (1954), 23-36.
- Gustason, William: 1994, *Reasoning from Evidence*. Macmillan College Publishing Company.
- Hacking, Ian: 1980, ‘Strange Expectations’, *Philosophy of Science* 47, 562-567.
- Hardin, Russell: 1982, *Collective Action* The Johns Hopkins University Press.
- Jeffrey, Richard C.: 1983, *The Logic of Decision*, Second Edition. University of Chicago Press.
- Menger, Karl: 1967 [1934], ‘The Role of Uncertainty in Economics’, in *Essays in Mathematical Economics in Honor of Oskar Morgenstern* (ed. Martin Shubik), Princeton University Press.
- Nathan, Amos: 1984, ‘False Expectations’ *Philosophy of Science* 51, 128-136.
- Resnik, Michael D.: 1987, *Choices: An Introduction to Decision Theory*. University of Minnesota Press.
- Weirich, Paul: 1984, ‘The St. Petersburg Gamble and Risk’, *Theory and Decision* 17, 193-202.

Other Discussions

- Ball, W. W. R. and Coxeter, H. S. M.: 1987, *Mathematical Recreations and Essays*, 13th ed., Dover.
- Gardner, M. 1959, *The Scientific American Book of Mathematical Puzzles & Diversions*. Simon and Schuster.
- Kamke, E.: 1932, *Einführung in die Wahrscheinlichkeitstheorie*. S. Hirzel..
- Keynes, J. M. K.: 1988, "The Application of Probability to Conduct", in *The World of Mathematics*, Vol. 2 (Ed. K. Newman). Microsoft Press.
- Kraitchik, M.: 1942, "The Saint Petersburg Paradox", §6.18 in *Mathematical Recreations*. W. W. Norton.
- Todhunter, I.: 1949, §391 in *History of the Mathematical Theory of Probability*, Chelsea.
- Peter Bernstein: 1996, *Against The Gods: the Remarkable Story of Risk*, John Wiley & Sons.

Other Internet Resources

- A Proposed ‘Solution’ to the Paradox, by Eric Arlet (Delft University of Technology)
- A Discussion in a Puzzle Archive, maintained by David Moews (University of Connecticut)
- ‘Two Lessons from Fractals and Chaos’, a preprint of a paper in *Complexity*, Vol. 5, No. 4, 2000, pp. 34-43, by Larry S. Liebovitch and Daniela Scheurle (Florida Atlantic University).

Related Entries

decision theory: causal | Pascal’s wager

Robert M. Martin
martin@is.dal.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 4, 1998
Content last modified: June 14, 2001

Friedrich Nietzsche

Friedrich Nietzsche was a German philosopher of the late 19th century who challenged the foundations of traditional morality and Christianity. He believed in life, creativity, health, and the realities of the world we live in, rather than those situated in a world beyond. Central to Nietzsche's philosophy is the idea of "life-affirmation," which involves an honest questioning of all doctrines which drain life's energies, however socially prevalent those views might be. Often referred to as one of the first "existentialist" philosophers, Nietzsche has inspired leading figures in all walks of cultural life, including dancers, poets, novelists, painters, psychologists, philosophers, sociologists and social revolutionaries.

Life: 1844-1900

In the small German town of Röcken bei Lützen, located in a rural farmland area southwest of Leipzig, Friedrich Wilhelm Nietzsche was born at approximately 10:00am on October 15, 1844. The date coincided with the 49th birthday of the Prussian King, Friedrich Wilhelm IV, after whom Nietzsche was named, and who had been responsible for Nietzsche's father's appointment as Röcken's town minister. Nietzsche's grandfathers were also Lutheran ministers, and his paternal grandfather, Friedrich August Ludwig Nietzsche, was further distinguished as a Protestant scholar, one of whose books (1796) affirmed the "everlasting survival of Christianity." When Nietzsche was 4 years old, his father, Karl Ludwig Nietzsche (1813-1849) died from a brain ailment, and the death of Nietzsche's two-year-old brother, Joseph, followed six months later. Having been living only yards away from Röcken's church in the

house reserved for the pastor and his family, the remaining Nietzsche family left their home soon after Karl Ludwig’s death. They moved to nearby Naumburg an der Saale, where Nietzsche (called "Fritz" by his family) lived for the next eight years with his mother, Franziska (1826-1897), his paternal grandmother, Erdmuthe, his father’s two sisters, Auguste and Rosalie, and his younger sister, Therese Elisabeth Alexandra (1846-1935).

From the ages of 14 to 19, Nietzsche attended a first-rate boarding school, Schulpforta, located not far from Naumburg, where he prepared for university studies. Here he met his lifelong acquaintance, Paul Deussen, who was confirmed at Nietzsche’s side in 1861, and who was to become an Orientalist, historian of philosophy, and in 1911, the founder of the Schopenhauer Society. During his summers in Naumburg, Nietzsche led a small music and literature club named "Germania," and became acquainted with Richard Wagner’s music through the club’s subscription to the *Zeitschrift für Musik.* The teenage Nietzsche also read the German romantic writings of Friedrich Hölderlin and Jean-Paul Richter, along with David Strauss’s controversial and demythologizing *Life of Jesus Critically Examined* (*Das Leben Jesu kritisch bearbeitet*, 1848).

After graduating from Schulpforta, Nietzsche entered the University of Bonn in 1864 as a theology and philology student, but his interests gravitated more exclusively towards philology -- a discipline which then centered upon the interpretation of classical and biblical texts. As a philology student, Nietzsche attended lectures by Otto Jahn (1813-1869) and Friedrich Wilhelm Ritschl (1806-1876). Jahn was a biographer of Mozart who had studied at the University of Berlin under Karl Lachmann (1793-1851) -- a philologist known both for his studies of the Roman philosopher Lucretius and for having developed the genealogical method in textual recension; Ritschl was a classics scholar whose work centered on the Roman comic poet Plautus (254-184 BC). Inspired by Ritschl, and following him to the University of Leipzig in 1865 -- an institution located closer to Nietzsche’s hometown of Naumburg -- Nietzsche quickly established his own academic reputation through his published essays on Aristotle, Theognis and Simonides. In Leipzig, he developed a close friendship with Erwin Rohde, a fellow philology student, with whom he would correspond extensively in later years. Momentous for Nietzsche in 1865 was his accidental discovery of Arthur Schopenhauer’s *The World as Will and Representation* (1818) in a local bookstore. He was then 21. Schopenhauer’s atheistic and turbulent vision of the world, in conjunction with his highest praise of music as an art form, captured Nietzsche’s imagination, and the extent to which the "cadaverous perfume" of Schopenhauer’s world-view continued to permeate Nietzsche’s mature thought is still a matter of scholarly debate. After discovering Schopenhauer, Nietzsche read F.A. Lange’s newly-published *History of Materialism and Critique of its Present Significance* (1866) -- a work which criticized materialist metaphysical theories from the standpoint of Kant’s critique of metaphysics in general, and attracted Nietzsche’s interest in its view that metaphysical speculation is an expression of poetic illusion.

In 1867, as he approached the age of 23, Nietzsche entered his required military service and was assigned to an equestrian field artillery regiment close to Naumburg, during which time he lived at home with his mother. While attempting to leap-mount into the saddle upon a particularly unruly horse, he suffered a serious chest injury and was put on sick leave after his chest wound refused to heal. He returned shortly thereafter to the University of Leipzig, and in November of 1868, met the composer Richard Wagner

(1813-1883) at the home of Hermann Brockhouse, an Orientalist who was married to Wagner's sister, Ottilie. Wagner and Nietzsche shared an enthusiasm for Schopenhauer, and Nietzsche -- who had been composing piano, choral and orchestral music since he was a teenager -- admired Wagner for his musical genius and magnetic personality. Wagner was exactly the age Nietzsche's father would have been, and Wagner had also attended the University of Leipzig many years before. The Nietzsche-Wagner relationship was quasi-familial, sometimes-stormy, and it affected Nietzsche deeply: twenty years later, he would still be assessing Wagner's cultural significance. During the months surrounding Nietzsche's initial meeting with Wagner, Ritschl strongly recommended Nietzsche for a position on the classical philology faculty at the University of Basel. The Swiss university offered Nietzsche the position, and he began teaching there in May, 1869, at the extraordinary age of 24.

At Basel, Nietzsche's satisfaction with his life among his philology colleagues was limited, and he established closer intellectual ties to the historians Franz Overbeck and Jacob Burkhardt, whose lectures he attended. Nietzsche also cultivated his friendship with Wagner and visited him often at his Swiss home in Tribschen, a small town near Lucerne. Never in outstanding health, further complications arose from Nietzsche's August-October 1870 service as a hospital attendant during the Franco-Prussian War (1870-71). He witnessed the traumatic effects of battle, took close care of wounded soldiers, contracted diphtheria and dysentery, and subsequently experienced a painful variety of health difficulties for the rest of his life.

Nietzsche's enthusiasm for Schopenhauer, his studies in classical philology, his inspiration from Wagner, his reading of Lange, and his frustration with the contemporary German culture, coalesced in his first book -- *The Birth of Tragedy* (1872) -- which was published when he was 28. Wagner showered the book with unqualified praise, but a biting critical reaction by the young and promising philologist, Ulrich von Wilamowitz-Möllendorff (1848-1931), dampened the book's reception among scholars.

As he continued his residence in Switzerland between 1872 and 1879, Nietzsche often visited Wagner at his new (1872) home in Bayreuth, Germany. In 1873, Nietzsche met Paul Rée, who, while living in close company with Nietzsche, would write *On the Origin of Moral Feelings* (1877). In 1876, at age 32, Nietzsche made an unsuccessful marriage proposal to a Dutch piano student in Geneva named Mathilde Trampedach. During this time, Nietzsche completed a series of four studies on contemporary German culture -- the *Unfashionable Observations* (1873-76) -- which focussed, respectively, upon the historian of religion and culture critic, David Strauss, issues concerning the social value of historiography, and Arthur Schopenhauer and Richard Wagner as inspirations for new cultural standards. Near the end of his university career, Nietzsche completed *Human, All-Too-Human* (1878) -- a book which marked a turning point in his philosophical style, and which signalled the end of his friendship with Wagner, who came under attack in Nietzsche's thinly-disguised characterization of "the artist." Despite the unflattering review of *The Birth of Tragedy,* Nietzsche remained respected in his professorial position in Basel, but his ailing health, which led to migraine headaches, eyesight problems and vomiting, necessitated his resignation from the university in June, 1879.

From 1880 until his collapse in January 1889, Nietzsche led a wandering, gypsy-like existence as a "stateless" person (having given up his German citizenship, and not having acquired Swiss citizenship),

circling almost annually between his mother's house in Naumburg and various French, Swiss, German and Italian cities. His travels took him through the Mediterranean seaside city of Nice (during the winters), the Swiss alpine village of Sils-Maria (during the summers), Leipzig (where he had attended university), Turin, Genoa, Recoaro, Messina, Rapallo, Florence, Venice, and Rome, never residing in any place longer than several months at a time. On a visit to Rome in 1882, Nietzsche, now at age thirty-seven, met Lou Salomé, a twenty-one-year-old Russian woman who was studying philosophy and theology in Zurich. He soon fell in love with her, and offered his hand in marriage. She declined, and the future of Nietzsche's friendship with her and Paul Rée appears to have suffered as a consequence. In the years to follow, Salomé would become an associate of Sigmund Freud, and would write with psychological insight of her association with Nietzsche. These nomadic years were the occasion of Nietzsche's main works, among which are *Daybreak* (1881), *The Gay Science* (1882), *Thus Spoke Zarathustra* (1883-85), *Beyond Good and Evil* (1886), and *On the Genealogy of Morals* (1887). Nietzsche's final active year, 1888, saw the completion of *The Case of Wagner* (May-August 1888), *Twilight of the Idols* (August-September 1888), *The Antichrist* (September 1888), *Ecce Homo* (October-November 1888) and *Nietzsche Contra Wagner* (December 1888).

On the morning of January 3, 1889, while in Turin, Nietzsche experienced a mental breakdown which left him an invalid for the rest of his life. Upon witnessing a horse being whipped by a coachman at the Piazza Carlo Alberto, Nietzsche threw his arms around the horse's neck and collapsed, never to return to full sanity. Some argue that Nietzsche was afflicted with a syphilitic infection (this was the original diagnosis of the doctors in Basel and Jena) contracted either while he was a student or while he was serving as a hospital attendant during the Franco-Prussian War; some claim that Nietzsche's use of chloral hydrate, a drug which he had been using as a sedative, deteriorated his already-weakened nervous system; some speculate that Nietzsche's collapse was due to a brain disease he inherited from his father; some maintain that a mental illness gradually drove him insane. The exact cause of Nietzsche's incapacitation still remains unclear. That Nietzsche had an extraordinarily sensitive nervous constitution and took an assortment of medications is well-documented as a more general fact.

During his creative years, Nietzsche struggled to bring his writings into print and never doubted that his books would have a lasting cultural effect. He did not live long enough to experience his world-historical influence, but he had a brief glimpse of his growing intellectual importance in discovering that he was the subject of 1888 lectures given by Georg Brandes (Georg Morris Cohen) at the University of Copenhagen, with whom he corresponded. Nietzsche's collapse, however, followed soon thereafter. After a brief hospitalization in Basel, he spent 1889 in a sanatorium in Jena at the Binswanger Clinic, and in March 1890 his mother took him back home to Naumburg, where he lived under her care for the next seven years. After his mother's death in 1897, his sister Elisabeth -- having previously returned home from Paraguay, where she had been working with her husband Bernhard Förster to establish an Aryan, anti-Semitic German colony called "New Germany" ("Nueva Germania") -- assumed responsibility for Nietzsche's welfare. In an effort to promote her brother's philosophy, she rented a large house on a hill in Weimar, called the "Villa Silberblick," and moved both Nietzsche and his collected manuscripts to the residence. This became the new home of the Nietzsche Archives (which was previously located at the family home in Naumburg), where Elisabeth received visitors who wanted to observe the now-incapacitated philosopher. On August 25, 1900, Nietzsche died in the villa as he approached his 56th

year, apparently of pneumonia in combination with a stroke. His body was then transported to the family gravesite directly beside the church in Röcken bei Lützen, where his mother and sister now also rest.

Early Writings: 1872-1876

Nietzsche's first book was published in 1872: *The Birth of Tragedy, Out of the Spirit of Music* (*Die Geburt der Tragödie aus dem Geiste der Musik*). It was reissued in 1886 with the title *The Birth of Tragedy, Or: Hellenism and Pessimism* (*Die Geburt der Tragödie, Oder: Griechentum und Pessimismus*), and contained a prefatory essay -- "An Attempt at Self-Criticism" -- which expressed Nietzsche's own critical reflections on his earlier work. *The Birth of Tragedy* set forth an alternative conception to the late 18th/early 19th century understanding of Greek culture -- a conception largely inspired by Johann Winckelmann's *History of Ancient Art* (1764) -- which hailed ancient Greece as the epitome of noble simplicity, calm grandeur, clear blue skies, and rational serenity. Nietzsche, having by this time absorbed the German romanticist, and specifically Schopenhauerian, view that non-rational forces reside at the foundation of all creativity and of reality itself, identified a strongly instinctual, wild, amoral, "Dionysian" energy within pre-Socratic Greek culture as an essentially creative and healthy force. Surveying the history of Western culture since the time of the Greeks, Nietzsche lamented over how this "Dionysian," creative energy had been submerged and weakened as it became overshadowed by the "Apollonian" forces of logical order and stiff sobriety. He concluded that European culture since the time of Socrates had remained one-sidedly Apollonian and relatively unhealthy. As a means towards cultural rebirth, Nietzsche advocated the resurrection and fuller release of Dionysian artistic energies -- those which he associated with primordial creativity, joy in existence and ultimate truth. The seeds of this rebirth Nietzsche perceived in the contemporary German music of his time, and the concluding part of *The Birth of Tragedy,* in effect, adulates the German artistic spirit as the potential savior of European culture.

Some scholars regard Nietzsche's 1873 unpublished essay, "On Truth and Lies in an Nonmoral Sense" ("*Über Wahrheit und Lüge im außermoralischen Sinn*") as a keystone in his thought. In this essay, Nietzsche rejects the idea of universal constants, and claims that what we call "truth" is only "a mobile army of metaphors, metonyms, and anthropomorphisms." His view at this time is that arbitrariness completely prevails within human experience: concepts originate via the very artistic transference of nerve stimuli into images; "truth" is nothing more than the invention of fixed conventions for merely practical purposes, especially those of repose, security and consistency. Viewing human existence from a great distance, Nietzsche further notes that there was an eternity before human beings came into existence, and believes that after humanity eventually dies out, nothing significant will have changed in the great scheme of things.

Between 1873 and 1876, Nietzsche wrote the *Unfashionable Observations* (*Unzeitgemässe Betrachtungen*). These are four (of a projected, but never completed, thirteen) studies concerned with the quality of European, and especially German, culture during Nietzsche's time. They are unfashionable and nonconformist (or "untimely," or "unmodern") insofar as Nietzsche regarded his standpoint as culture-critic to be in tension with the self-congratulatory spirit of the times. The four studies were: *David*

Strauss, the Confessor and the Writer (*David Strauss, der Bekenner und der Schriftsteller,* 1873); *On the Uses and Disadvantages of History for Life* (*Vom Nutzen und Nachteil der Historie für das Leben*, 1874); *Schopenhauer as Educator* (*Schopenhauer als Erzieher,* 1874); *Richard Wagner in Bayreuth* (1876). The first of these attacked David Strauss, whose popular six-edition book, *The Old and the New Faith: A Confession* (1871) encapsulated for Nietzsche the general cultural atmosphere in Germany. Responding to Strauss's advocacy of a "new faith" grounded upon a scientifically-determined universal mechanism -- one, however, lubricated by the optimistic, "soothing oil" of historical progress -- Nietzsche unmercifully attacked Strauss's view as a vulgar and dismal sign of cultural degeneracy. The second "untimely meditation" surveyed alternative ways to write history, and discussed how these ways could contribute to a society's health. Here Nietzsche claimed that the principle of "life" is a more pressing and higher concern than that of "knowledge," and that the quest for knowledge should serve the interests of life. The third and fourth studies -- on Schopenhauer and Wagner, respectively -- addressed how these two thinkers, as paradigms of philosophic and artistic genius, held the potential to inspire a stronger, healthier and livelier German culture.

Middle-Period Writings: 1878-1882

In 1878, Nietzsche completed *Human, All-Too-Human,* supplementing this with a second part in 1879, *Mixed Opinions and Maxims* (*Vermischte Meinungen und Sprüche*), and a third part in 1880, *The Wanderer and his Shadow* (*Der Wanderer und sein Schatten*). The three parts were published together in 1886 as *Human All-Too-Human, A Book for Free Spirits* (*Menschliches, Allzumenschliches, Ein Buch für freie Geister*). Reluctant to construct a philosophical "system," and sensitive to the importance of style in philosophic writing, Nietzsche composed these works as a series of several hundred aphorisms whose typical length ranges from a line or two to a page or two. Here, he often reflects upon cultural and psychological phenomena in reference to individuals's organic and physiological constitutions. The idea of power (for which he would later become known) sporadically appears as an explanatory principle, but Nietzsche tends at this time to invoke hedonistic considerations of pleasure and pain in his explanations of cultural and psychological phenomena.

In *Daybreak, Reflections on Moral Prejudices* (*Morgenröte. Gedanken über die moralischen Vorurteile*, 1881), Nietzsche continued writing in his aphoristic style, but began accentuating the importance of the "feeling of power," as opposed to pleasure, in his understanding of human, and especially of so-called "moral" behavior. In this respect, *Daybreak* contains the seeds of Nietzsche's doctrine of the "will to power" -- a doctrine which would appear explicitly for the first time two years later in *Thus Spoke Zarathustra* (1883-85). *Daybreak* is also one of Nietzsche's clearest, intellectually calmest, and most intimate, volumes, providing many social-psychological insights, in conjunction with some of his first sustained critical reflections on the cultural relativity at the basis of Christian moral evaluations.

In a more well-known aphoristic work, *The Gay Science* (*Die fröhliche Wissenschaft*, 1882) -- whose title was inspired by the troubadour songs of southern-French Provence (1100-1300) -- Nietzsche set forth some of the existential ideas for which he became famous, namely, the proclamation that "God is dead" and the doctrine of "eternal recurrence"-- the idea that one is, or might be, fated to relive forever

every moment one one's life, with no omission whatsoever of any pleasurable or painful detail. Nietzsche's atheism -- his account of "God's murder" (section 125) -- was voiced in reaction to the conception of a single, ultimate, judgmental authority who is privy to everyone's hidden, and personally embarrassing, secrets; his atheism also aimed to redirect people's attention to their inherent freedom, the presently-existing world, and away from all escapist, pain-relieving, heavenly otherworlds. To a similar end, Nietzsche's doctrine of eternal recurrence (sections 285 and 341) was formulated to draw attention away from all worlds other than the one in which we presently live, since eternal recurrence precludes the possibility of any final escape from the present world. The doctrine also functions as a measure for judging someone's overall psychological strength and mental health, since Nietzsche believed that the doctrine of eternal recurrence was the hardest world-view to accept and affirm. In 1887, *The Gay Science* was reissued with an important preface, an additional fifth Book, and an appendix of songs, reminiscent of the troubadours.

Later-Period Writings: 1883-1887

Thus Spoke Zarathustra, A Book for All and None (*Also Sprach Zarathustra, Ein Buch für Alle und Keinen*, 1883-85), is one of Nietzsche's most famous works, and Nietzsche himself regarded it as among his most significant. It is, in effect, a manifesto of personal self-overcoming. Thirty years after its initial publication, 150,000 copies of the work were printed by the German government and issued as inspirational reading, along with the Bible, to the young soldiers during WWI. Though *Thus Spoke Zarathustra* is antagonistic to the Judeo-Christian world-view, its poetic and prophetic style relies upon many, often inverted, Old and New Testament allusions. Nietzsche also filled the work with nature metaphors, almost in the spirit of pre-Socratic naturalist philosophy, which invoked animals, earth, air, fire, water, celestial bodies, plants, all in the service of describing the spiritual development of Zarathustra, a solitary, reflective, exceedingly strong-willed, sage-like, laughing and dancing voice of self-mastery who, accompanied by a proud, sharp-eyed eagle and a wise snake, envisioned a mode of psychologically healthier being beyond the common human condition. Nietzsche refers to this higher mode of being as "superhuman" (*übermenschlich*), and associates the doctrine of eternal recurrence -- a doctrine for only the healthiest who can love life in its entirety -- with this spiritual standpoint, in relation to which all-too-often downhearted, all-too-commonly-human attitudes stand as a mere bridge to be crossed and overcome.

In *Beyond Good and Evil, Prelude to a Philosophy of the Future* (*Jenseits von Gut und Böse. Vorspiel einer Philosophie der Zukunft*, 1886), Nietzsche identified imagination, self-assertion, danger, originality and the "creation of values" as qualities of genuine philosophers, as opposed to incidental characters who engage in dusty scholarship. Nietzsche also took aim at some of the world's great philosophers's key presuppositions, who grounded their outlooks wholeheartedly upon concepts such as "self-consciousness," "free will," and "either/or" bipolar thinking. Alternatively, Nietzsche philosophizes from "the perspective of life" which he regards as "beyond good and evil," and challenges the deeply-entrenched moral idea that exploitation, domination, injury to the weak, destruction and appropriation are universally objectionable behaviors. Above all, Nietzsche believes that living things aim to discharge their strength and express their "will to power" -- a pouring-out of expansive energy which, quite

naturally, can entail danger, pain, lies, deception and masks. As he views things from the perspective of life, he further denies that there is a universal morality applicable indiscriminately to all human beings, and instead designates a series of moralities in an order of rank ranging from the noble to the plebeian: some moralities are more appropriate for dominating and leading social roles; some are more suitable for subordinate roles. So what counts as a preferable and legitimate action depends upon the kind of person one is. The deciding factor is whether one is strong, healthy, powerful and overflowing with ascending life, or whether one is weak, sick and on the decline.

On the Genealogy of Morals, A Polemic (*Zur Genealogie der Moral, Eine Streitschrift*, 1887), is composed of three sustained essays which advance the critique of Christianity expressed in *Beyond Good and Evil*. The first essay continues the discussion of master morality versus servant morality, and maintains that the traditional ideals set forth as holy and morally good within Christian morality are products of self-deception, since they were forged in the bad air of revenge, resentment, hatred, impotence, and cowardice. In this essay, as well as the next, Nietzsche's controversial references to the "blond beast" akin to master morality often appear. In the second essay, Nietzsche continues with an account of how feelings of guilt, or the "bad conscience," arise merely as a consequence of an unhealthy Christian morality which turns an "evil eye" towards our natural inclinations. He also discusses how punishment, conceived as the infliction of pain upon someone in proportion to their offense, is likely to have been grounded in the contractual economic relationship between creditor and debtor. In the third essay, Nietzsche focusses upon the ascetic ideals typical of the social representatives of art, religion and philosophy, and he offers a particularly scathing critique of the priesthood: the priests are allegedly a group of weak people who shepherd even weaker people as a way to experience power for themselves. The third essay also contains one of Nietzsche's clearest expressions of "perspectivism" (section 12) -- the idea that there is no absolute, "God's eye" standpoint from which one can survey everything that is.

Final Writings of 1888

The Case of Wagner, A Musician's Problem (*Der Fall Wagner, Ein Musikanten-Problem,* May-August 1888), compares well with Nietzsche's 1873 meditation on David Strauss in its devastating and unbridled attack on a popular cultural figure. In *The Case of Wagner,* Nietzsche "declares war" upon Richard Wagner, whose music is characterized as both the epitome of modern cultural achievement and as thoroughly sick and decadent. The work is a brilliant display of Nietzsche's talents as a music critic, and includes memorable mockings of Wagner's theatrical style, reflections on redemption via art, a "physiology of art," and the virtues associated, respectively, with ascending and descending life energies.

The title, *Twilight of the Idols, or How One Philosophizes with a Hammer* (*Götzen-Dämmerung, oder Wie man mit dem Hammer philosophiert*, August-September 1888), word-plays upon Wagner's opera, *The Twilight of the Gods* (*Die Götterdämmerung*). Nietzsche reiterates and elaborates some of the criticisms of Socrates, Plato, Kant and Christianity found in earlier works, criticizes the then-contemporary German culture as being unsophisticated and too-full of beer, and shoots some disapproving arrows at key French, British, and Italian cultural figures such as Rousseau, Hugo, Sand, Michelet, Zola, Renan, Carlyle, Mill, Eliot, Darwin, and Dante. In contrast to all these alleged

representatives of cultural decadence, Nietzsche applauds Caesar, Napoleon, Goethe, Dostoevski, Thucydides and the Sophists as healthier and stronger types.

In *The Antichrist, Curse on Christianity* (*Der Antichrist. Fluch auf das Christentum,* September 1888), Nietzsche expresses his disgust over the way noble values in Roman Society were "corrupted" by the rise of Christianity, and he discusses specific aspects and personages in Christian culture -- the Gospels, Paul, the martyrs, priests, the crusades -- with a view towards showing that Christianity is a religion for weak and unhealthy people, whose general historical effect has been to undermine the healthy qualities of the more noble cultures.

Nietzsche describes himself as "a follower of the philosopher Dionysus" in *Ecce Homo, How One Becomes What One Is* (*Ecce Homo, Wie man wird, was man ist*, October-November 1888) -- a book in which he examines retrospectively his entire corpus, work by work, offering critical remarks, details of how the works were inspired, and explanatory observations regarding their philosophical contents. He begins this fateful intellectual autobiography -- he was to lose his mind little more than a month later -- with three eyebrow-raising sections entitled, "Why I Am So Wise," "Why I Am So Clever," and "Why I Write Such Good Books." Nietzsche claims to be wise as a consequence of his acute aesthetic sensitivity to nuances of health and sickness in people’s attitudes and characters; he claims to be clever because he knows how to choose the right nutrition, climate, residence and recreation for himself; he claims to write such good books because they allegedly adventurously open up, at least for a very select group of readers, a new series of noble and delicate experiences. After examining each of his published works, Nietzsche concludes *Ecce Homo* with the section, "Why I Am a Destiny." He claims that he is a destiny because he regards his anti-moral truths as having the annihilating power of intellectual dynamite; he expects them to topple the morality born of sickness which he perceives to have been reigning within Western culture for the last two thousand years. In this way, Nietzsche expresses his hope that Dionysus, the god of life’s exuberance, would replace Jesus, the god of the heavenly otherworld, as the premier cultural standard for future millennia.

Nietzsche Contra Wagner, Out of the Files of a Psychologist (*Nietzsche contra Wagner, Aktenstücke eines Psychologen,* December 1888) is a short, but classic, selection of passages Nietzsche extracted from his 1878-1887 published works. Many concern Wagner, but the excerpts serve mostly as a foil for Nietzsche to express his own views against Wagner’s. In this self-portrait, completed only a month before his collapse, Nietzsche characterizes his own anti-Christian sentiments, and contemplates how even the greatest people usually undergo significant corruption. In Wagner’s case, Nietzsche claims that the corrupting force was Christianity. At the same time, Nietzsche describes how he truly admired some of Wagner’s music for its deep expressions of loneliness and suffering -- expressions which Nietzsche admitted were psychologically impossible for he himself to articulate.

Nietzsche’s Unpublished Notebooks

Nietzsche’s unpublished writings often reveal his more tentative and speculative ideas. This material is surrounded by controversy, however, since some of it conflicts dramatically with views Nietzsche

expresses in his published works. Disagreement regarding Nietzsche's notebooks, also known as his *Nachlass*, centers around the degree of interpretive priority which ought to be given to the unpublished versus the published manuscripts. One popular approach in the tradition of classical scholarly interpretation is to maintain that Nietzsche's published works express his more considered and polished views, and that these should take precedence over the unpublished manuscripts when conflicts arise; a second attitude, given voice by Martin Heidegger, and broadly consistent with a psychoanalytic approach as well, is to regard what Nietzsche published as representative of what he decided was publicly presentable, and what he kept privately to himself in unpublished form as containing his more authentic views; a third, more comprehensive, interpretive style tries to grasp all of Nietzsche's texts together in an effort to form the most coherent interpretation of Nietzsche's thought, judging the priority of published versus unpublished works on a thematic, or case-by-case basis; a fourth position influenced by the French deconstructionist perspective maintains that any rigid prioritizing between published and private works is impossible, since all of the texts embody a comparable multidimensionality of meaning.

In his unpublished manuscripts, Nietzsche sometimes elaborates the topics found in the published works, such as his early 1870's notebooks, where there is important material concerning his theory of knowledge. In the 1880's notebooks -- those his sister collected together after his death under the title, *The Will to Power: Attempt at a Revaluation of all Values* -- Nietzsche adopts a more metaphysical orientation towards the doctrines of Eternal Recurrence and the Will to Power, speculating upon their intellectual strength as interpretations of reality itself. Side-by-side with these speculations, and complicating efforts towards developing an interpretation which is both comprehensive and coherent, Nietzsche's 1880's notebooks also repeatedly state that "there are no facts, only interpretations."

Nietzsche's Influence Upon 20th Century Thought

Nietzsche's thought extended a deep influence during the 20th century, especially in Continental Europe. In English-speaking countries, his positive reception has been less resonant. During the last decade of Nietzsche's life and the first decade of the 20th century, his thought was particularly attractive to avant-garde artists who saw themselves on the periphery of established social fashion and practice. Here, Nietzsche's advocacy of new, healthy beginnings, and of creative artistry in general stood forth. His tendency to seek explanations for commonly-accepted values and outlooks in the less-elevated realms of sheer animal instinct was also crucial to Sigmund Freud's development of psychoanalysis. Later, during the 1930's, aspects of Nietzsche's thought were espoused by the Nazis and Italian Fascists, partly due to the encouragement of Elisabeth Förster-Nietzsche through her solicitations with Adolf Hitler and Benito Mussolini. It was possible for the Nazi interpreters to assemble, quite selectively, various passages from Nietzsche's writings whose juxtaposition appeared to justify war, aggression and domination for the sake of nationalistic and racial self-glorification. Until the 1960's in France, Nietzsche appealed mainly to writers and artists, since the academic philosophical climate was dominated by G.W.F. Hegel's, Edmund Husserl's and Martin Heidegger's thought, along with the structuralist movement of the 1950's. Nietzsche became especially influential in French philosophical circles during the 1960's-1980's, when his "God is dead" declaration, his perspectivism, and his emphasis upon power as the real motivator and explanation for people's actions revealed new ways to challenge established authority and launch

effective social critique.

Specific 20th century figures who were influenced, either quite substantially, or in a significant part, by Nietzsche include painters, dancers, musicians, playwrights, poets, novelists, psychologists, sociologists, literary theorists, historians, and philosophers: Alfred Adler, Georges Bataille, Martin Buber, Albert Camus, E.M. Cioran, Jacques Derrida, Gilles Deleuze, Isadora Duncan, Michel Foucault, Sigmund Freud, Stefan George, André Gide, Hermann Hesse, Carl Jung, Martin Heidegger, Gustav Mahler, André Malraux, Thomas Mann, Rainer Maria Rilke, Jean-Paul Sartre, Max Scheler, Giovanni Segantini, George Bernard Shaw, Lev Shestov, Georg Simmel, Oswald Spengler, Paul Tillich, Ferdinand Tönnies, Mary Wigman, William Butler Yeats and Stefan Zweig.

That Nietzsche was able to write so prolifically and profoundly for years, while remaining in a condition of ill-health and often intense physical pain, is a testament to his spectacular mental capacities and willpower. Lesser people under the same physical pressures might not have had the inclination to pick up a pen, let alone think and record thoughts which -- created in the midst of striving for healthy self-overcoming -- would have the power to influence an entire century.

Bibliography

A. Nietzsche’s Writings

- *Kritische Gesamtausgabe Briefwechsel.* ed. G. Colli and M. Montinari, 24 vols. in 4 parts. Berlin: Walter de Gruyter, 1975.
- *The Antichrist.* trans. Walter Kaufmann, in *The Portable Nietzsche*, ed. Walter Kaufmann. New York: Viking Press, 1968.
- *Beyond Good and Evil*. trans. Walter Kaufmann. New York: Random House, 1966.
- *The Birth of Tragedy*. trans. Walter Kaufmann, in *The Birth of Tragedy and The Case of Wagner*. New York: Random House, 1967.
- *The Case of Wagner*. trans. Walter Kaufmann, in *The Birth of Tragedy and The Case of Wagner*. New York: Random House, 1967.
- Daybreak: Thoughts on the Prejudices of Morality. trans. R. J. Hollingdale. Cambridge: Cambridge University Press, 1982.
- *Ecce Homo: How One Becomes What One Is*. trans. Walter Kaufmann, in *On the Genealogy of Morals and Ecce Homo*. New York: Random House, 1967.
- *The Gay Science, with a Prelude of Rhymes and an Appendix of Songs*. tr. Walter Kaufmann. New York: Random House, 1974.
- *Human, All Too Human: A Book for Free Spirits*. trans. R. J. Hollingdale. Cambridge: Cambridge University Press, 1986.
- *Nietzsche Contra Wagner*. trans. Walter Kaufmann, in *The Portable Nietzsche*. New York: Viking Press, 1968.
- *On the Genealogy of Morals*. trans. Walter Kaufmann and R.J. Hollingdale, in *On the Genealogy*

of Morals and Ecce Homo. New York: Random House, 1967.

- *Philosophy and Truth: Selections from Nietzsche's Notebooks of the Early 1870's*. trans. and ed. Daniel Breazeale. Atlantic Highlands, N.J.: Humanities Press, 1979.
- *Philosophy in the Tragic Age of the Greeks*. trans. Marianne Cowan. Chicago: Henry Regnery Company, 1962.
- *Thus Spoke Zarathustra*. trans. Walter Kaufmann, in *The Portable Nietzsche*. New York: Viking Press, 1968.
- *Twilight of the Idols*. trans. Walter Kaufmann, in *The Portable Nietzsche*. New York: Viking Press, 1968.
- *Untimely Meditations*. trans. R. J. Hollingdale. Cambridge: Cambridge University Press, 1983.
- *The Will to Power*. trans. Walter Kaufmann. New York: Random House, 1967.

B. Books About Nietzsche

- Allison, David. *Reading the New Nietzsche*. Lanham, Maryland: Rowman & Littlefield Publishing, 2000.
- Aschheim, Steven E. *The Nietzsche Legacy in Germany, 1890-1990*. Berkeley and Los Angeles: University of California Press, 1992.
- Babich, Babette E. *Nietzsche's Philosophy of Science*. Albany: State University of New York Press, 1994.
- Bataille, Georges. *On Nietzsche*. trans. Bruce Boone. London: Athlone Press, 1992.
- Clark, Maudemarie. *Nietzsche on Truth and Philosophy*. Cambridge: Cambridge University Press, 1990.
- Danto, Arthur C. *Nietzsche as Philosopher: An Original Study*. New York: Columbia University Press, 1965.
- Deleuze, Gilles. *Nietzsche and Philosophy*. trans. Hugh Tomlinson. New York: Columbia University Press, 1983.
- Derrida, Jacques. *Spurs: Nietzsche's Styles*. trans. Barbara Harlow. Chicago: University of Chicago Press, 1979.
- Gilman, Sander L., ed., and David J. Parent, trans. *Conversations with Nietzsche: A Life in the Words of his Contemporaries*. New York: Oxford University Press, Inc., 1987.
- Hayman, Ronald. *Nietzsche, a Critical Life*. New York: Oxford University Press, 1980.
- Heidegger, Martin. *Nietzsche, Vol. I: The Will to Power as Art*. trans. David F. Krell. New York: Harper & Row, 1979. *Nietzsche, Vol. II: The Eternal Recurrence of the Same*. trans. David F. Krell. San Francisco: Harper & Row, 1984. *Nietzsche, Vol. III.: Will to Power as Knowledge and as Metaphysics*. trans. Joan Stambaugh and Frank Capuzzi. San Francisco: Harper & Row, 1986. *Nietzsche, Vol. IV: Nihilism*. trans. David F Krell. New York: Harper & Row, 1982.
- Higgins, Kathleen Marie. *Nietzsche's "Zarathustra."* Philadelphia: Temple University Press, 1987.
- Hollingdale, R. J. *Nietzsche*. London and New York: Routledge and Kegan Paul, 1973.
- Hunt, Lester H. *Nietzsche and the Origin of Virtue*. London: Routledge, 1991.
- Irigaray, Luce. *Marine Lover of Friedrich Nietzsche*. trans. Gillian C. Gill. New York: Columbia University Press, 1991.

- Jaspers, Karl. *Nietzsche: An Introduction to the Understanding of His Philosophical Activity*. trans. Charles F. Wallraff and Frederick J. Schmitz. South Bend, Indiana: Regentry/Gateway, Inc., 1979.
- Jung, Carl G., *Nietzsche's "Zarathustra."* ed. James L. Jarrett. Princeton: Princeton University Press, 1988.
- Kaufmann, Walter. *Nietzsche: Philosopher, Psychologist, Antichrist.* Princeton: Princeton University Press, 1950.
- Klossowski, Pierre. *Nietzsche and the Vicious Circle*. London: Athlone, 1993.
- Kofman, Sarah. *Nietzsche and Metaphor*. ed. and trans., Duncan Large. London: Athlone Press; Stanford, CA: Stanford University Press, 1993
- Krell, David Farrell. *Postponements: Women, Sensuality, and Death in Nietzsche.* Bloomington: Indiana University Press, 1986.
- Lambert, Laurence. *Nietzsche's Teaching: An Interpretation of "Thus Spoke Zarathustra."* New Haven: Yale University Press, 1987.
- Löwith, Karl. *Nietzsche's Philosophy of the Eternal Recurrence of the Same*." [1956], translated by J. Harvey Lomax, foreword by Bernd Magnus. Berkeley: University of California Press, 1997.
- Macintyre, Ben. *Forgotten Fatherland: The Search for Elisabeth Nietzsche.* London: Macmillan, 1992.
- Magnus, Bernd, Stanley Stewart, and Jean-Pierre Mileur. *Nietzsche's Case: Philosophy as/and Literature*. New York and London: Routledge, 1993.
- Magnus, Bernd. *Nietzsche's Existential Imperative*. Bloomington: Indiana University Press, 1978.
- Mandel, Siegfried. *Nietzsche & the Jews.* New York: Prometheus Books, 1998.
- Nehamas, Alexander. *Nietzsche: Life as Literature*. Cambridge, Mass.: Harvard University Press, 1985.
- Oliver, Kelly. *Womanizing Nietzsche: Philosophy's Relation to the "Feminine."* New York and London: Routledge, 1995.
- Parkes, Graham. *Composing the Soul: Reaches of Nietzsche's Psychology*. Chicago and London: University of Chicago Press, 1994.
- Pletch, Carl. *Young Nietzsche: Becoming a Genius*. New York: Free Press, 1991.
- Rosen, Stanley. *The Mask of Enlightenment: Nietzsche's Zarathustra*. Cambridge: Cambridge University Press, 1995.
- Salomé, Lou. *Nietzsche*. ed. and trans. Siegfried Mandel. Redding Ridge, Connecticut: Black Swan Books, Ltd., 1988.
- Schacht, Richard. *Nietzsche*. London: Routledge and Kegan Paul, 1983.
- Shapiro, Gary. *Nietzschean Narratives.* Bloomington: Indiana University Press, 1989.
- Simmel, Georg. *Schopenhauer and Nietzsche*. trans. Helmut Loiskandle, Deena Weinstein, and Michael Weinstein. Urbana and Chicago: University of Illinois Press, 1991.
- Schrift, Alan D. *Nietzsche and the Question of Interpretation: Between Hermeneutics and Deconstruction*. New York: Routledge, 1990.
- Stambaugh, Joan. *The Problem of Time in Nietzsche.* trans. John F. Humphrey. Philadelphia: Bucknell University Press, 1987.
- White, Alan. *Within Nietzsche's Labyrinth*. New York and London: Routledge, 1990.
- Wilcox, John T. *Truth and Value in Nietzsche.* Ann Arbor: University of Michigan Press, 1974.

- Young, Julian. *Nietzsche's Philosophy of Art.* Cambridge: Cambridge University Press, 1992.

C. Collected Essays on Nietzsche

- Allison, David B., ed., *The New Nietzsche: Contemporary Styles of Interpretation.* Cambridge, Massachusetts: The MIT Press, 1985.
- Bloom, Harold., ed., *Modern Critical Views: Friedrich Nietzsche.* New York, New Haven, Philadelphia: Chelsea House Publishers, 1987.
- Koelb, Clayton., ed., *Nietzsche as Postmodernist: Essays Pro and Contra.* Albany: State University of New York Press, 1990.
- Magnus, Bernd, and Higgins, Kathleen M., eds., *The Cambridge Companion to Nietzsche.* Cambridge: Cambridge University Press, 1996.
- Parkes, Graham., ed., *Nietzsche and Asian Thought.* Chicago: The University of Chicago Press, 1991.
- Sedgwick, Peter R., ed., *Nietzsche: A Critical Reader*. Oxford, UK and Cambridge, USA: Blackwell, 1995.
- Solomon, Robert C., and Higgins, Kathleen M., eds., *Reading Nietzsche.* New York and Oxford: Oxford University Press, 1988.
- Solomon, Robert. ed., *Nietzsche: A Collection of Critical Essays.* Garden City, New York: Anchor Books, 1973.
- Yovel, Yirmiyahu., ed., *Nietzsche as Affirmative Thinker*. Dordrecht: Martinus Nihoff Publishers, 1986.

Other Internet Resources

- The Friedrich Nietzsche Society (UK)
- The Nietzsche Society
- North American Nietzsche Society
- Journal of Nietzsche Studies
- New Nietzsche Studies
- Associazione Internazionale di Studie Ricerche Federico Nietzsche (Italy)
- Grupo de Estudos Nietzsche (Brazil)
- Society for Phenomenology and Existential Philosophy

Related Entries

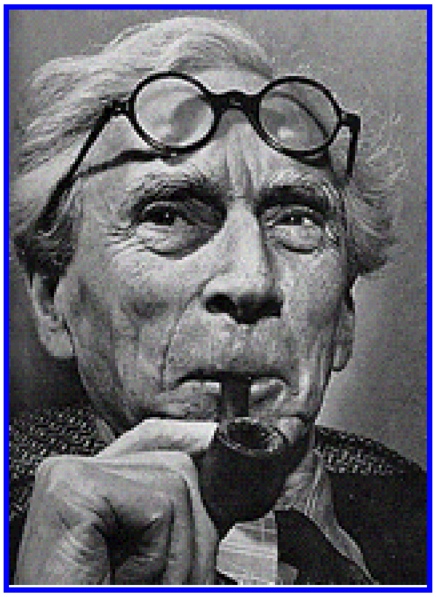
existentialism | relativism | Schopenhauer, Arthur

r.wicks@auckland.ac.nz

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 30, 1997
Content last modified: August 6, 2001



Bertrand Russell

Bertrand Arthur William Russell (b.1872 - d.1970), British philosopher, logician, essayist, and social critic, best known for his work in mathematical logic and analytic philosophy. His most influential contributions include his defense of logicism (the view that mathematics is in some important sense reducible to logic), and his theories of definite descriptions and logical atomism. Along with G.E. Moore, Russell is generally recognized as one of the founders of analytic philosophy. Along with Kurt Gödel, he is also often credited with being one of the two most important logicians of the twentieth century.

Over the course of his long career, Russell made significant contributions, not just to logic and philosophy, but to a broad range of other subjects (including education, politics, history, religion and science), and many of his writings on a wide variety of topics have influenced generations of general readers. After a life marked by controversy (including dismissals from both Trinity College, Cambridge, and City College, New York), Russell was awarded the Order of Merit in 1949 and the Nobel Prize for Literature in 1950. Also noted for his many spirited anti-war and anti-nuclear protests, Russell remained a prominent public figure until his death at the age of 97.

For an excellent short introduction to Russell's life, work and influence the reader is encouraged to consult John Slater's accessible and informative *Bertrand Russell* (Bristol: Thoemmes, 1994).

For a complete list of Russell's publications see *A Bibliography of Bertrand Russell* (3 vols, London: Routledge, 1994), by Kenneth Blackwell and Harry Ruja. A less detailed, but still comprehensive, list appears in Paul Arthur Schilpp, *The Philosophy of Bertrand Russell*, 3rd ed., (New York: Harper and

Row, 1963), pp. 746-803.

For a bibliography of the secondary literature surrounding Russell, see A.D. Irvine (ed.), *Bertrand Russell: Critical Assessments*, Vol. 1, (London: Routledge, 1998), pp. 247-312.

A Chronology of Russell's Life

A short chronology of the major events in Russell's life is as follows:

- (1872) Born May 18 at Ravenscroft, Wales.
- (1874) Death of mother and sister.
- (1876) Death of father; Russell's grandfather, Lord John Russell (the former Prime Minister), and grandmother succeed in overturning his father's will to win custody of Russell and his brother.
- (1878) Death of grandfather; Russell's grandmother, Lady Russell, supervises his upbringing.
- (1890) Enters Trinity College, Cambridge.
- (1893) Awarded first class B.A. in Mathematics.
- (1894) Completed the Moral Sciences Tripos (Part II)
- (1894) Marries Alys Pearsall Smith.
- (1900) Meets Peano at International Congress in Paris.
- (1901) Discovers Russell's paradox.
- (1902) Corresponds with Frege.
- (1908) Elected Fellow of the Royal Society.
- (1916) Fined 110 pounds and dismissed from Trinity College as a result of anti-war protests.
- (1918) Imprisoned for six months as a result of anti-war protests.
- (1921) Divorce from Alys and marriage to Dora Black.
- (1927) Opens experimental school with Dora.
- (1931) Becomes the third Earl Russell upon the death of his brother.
- (1935) Divorce from Dora.
- (1936) Marriage to Patricia (Peter) Helen Spence.

- (1940) Appointment at City College New York revoked following public protests.
- (1943) Dismissed from Barnes Foundation in Pennsylvania.
- (1949) Awarded the Order of Merit.
- (1950) Awarded Nobel Prize for Literature.
- (1952) Divorce from Peter and marriage to Edith Finch.
- (1955) Releases Russell-Einstein Manifesto.
- (1957) Organizes the first Pugwash Conference.
- (1958) Becomes founding President of the Campaign for Nuclear Disarmament.
- (1961) Imprisoned for one week in connection with anti-nuclear protests.
- (1970) Dies February 02 at Penrhyndeudraeth, Wales.

For a chronology of Russell's major publications, consult the section below entitled Russell's Writings.

For more detailed information about Russell's life, the reader is encouraged to consult, in the first instance, Russell's four autobiographical volumes, *My Philosophical Development* (London: George Allen & Unwin, 1959) and *The Autobiography of Bertrand Russell* (3 vols, London: George Allen & Unwin, 1967, 1968, 1969).

Other sources of biographical information include Ronald Clark, *The Life of Bertrand Russell* (London: J. Cape, 1975), A.D. Irvine (ed.), *Bertrand Russell: Critical Assessments*, Vol. 1, (London: Routledge, 1998), and Ray Monk, *Bertrand Russell: The Spirit of Solitude* (London: J. Cape, 1996).

Russell's Work in Logic

Russell's contributions to logic and the foundations of mathematics include his discovery of Russell's paradox, his defense of logicism (the view that mathematics is, in some significant sense, reducible to formal logic), his development of the theory of types, and his refining of the first-order predicate calculus.

Russell discovered the paradox that bears his name in 1901, while working on his *Principles of Mathematics* (1903). The paradox arises in connection with the set of all sets that are not members of themselves. Such a set, if it exists, will be a member of itself if and only if it is not a member of itself. The paradox is significant since, using classical logic, all sentences are entailed by a contradiction. Russell's discovery thus prompted a large amount of work in logic, set theory, and the philosophy and foundations of mathematics.

Russell's own response to the paradox came with the development of his theory of types in 1903. It was clear to Russell that some restrictions needed to be placed upon the original comprehension (or abstraction) axiom of naive set theory, the axiom which formalized the intuition that any coherent condition may be used to determine a set (or class). Russell's basic idea was that reference to sets such as the set of all sets that are not members of themselves could be avoided by arranging all sentences into a

hierarchy (beginning with sentences about individuals at the lowest level, sentences about sets of individuals at the next lowest level, sentences about sets of sets of individuals at the next lowest level, etc.). Using a vicious circle principle similar to that adopted by the mathematician Henri Poincaré, and his own so-called "no class" theory of classes, Russell was able to explain why the unrestricted comprehension axiom fails: propositional functions, such as the function "*x* is a set", may not be applied to themselves since self-application would involve a vicious circle. Thus, on Russell's view, all objects for which a given condition (or predicate) holds must be at the same level or of the same "type".

Although first introduced in 1903, the theory of types finds its mature expression in Russell's 1908 article "Mathematical Logic as Based on the Theory of Types" and in the monumental work he co-authored with Alfred North Whitehead, *Principia Mathematica* (1910, 1912, 1913). Thus the theory admits of two versions, the "simple theory" and the "ramified theory". Both versions of the theory later came under attack for being both too weak and too strong. For some, the theory was too weak since it failed to resolve all of the known paradoxes. For others, it was too strong since it disallowed many mathematical definitions which, although consistent, violated the vicious circle principle. Russell's response was to introduce the axiom of reducibility, an axiom that lessened the vicious circle principle's scope of application, but which many people claimed was too ad hoc to be justified philosophically.

Of equal significance during this same period was Russell's defense of logicism, the theory that mathematics was in some important sense reducible to logic. First defended in his *Principles of Mathematics*, and later in greater detail in *Principia Mathematica*, Russell's logicism consisted of two main theses. The first is that all mathematical truths can be translated into logical truths or, in other words, that the vocabulary of mathematics constitutes a proper subset of that of logic. The second is that all mathematical proofs can be recast as logical proofs or, in other words, that the theorems of mathematics constitute a proper subset of those of logic.

Like Gottlob Frege, Russell's basic idea for defending logicism was that numbers may be identified with classes of classes and that number-theoretic statements may be explained in terms of quantifiers and identity. Thus the number 1 would be identified with the class of all unit classes, the number 2 with the class of all two-membered classes, and so on. Statements such as "There are two books" would be recast as statements such as "There is a book, *x*, and there is a book, *y*, and *x* is not identical to *y*". It followed that number theoretic operations could be explained in terms of set theoretic operations such as intersection, union, and difference. In *Principia Mathematica*, Whitehead and Russell were able to provide many detailed derivations of major theorems in set theory, finite and transfinite arithmetic, and elementary measure theory. A fourth volume on geometry was planned but never completed.

Russell's most important writings relating to these topics include not only *Principles of Mathematics* (1903), "Mathematical Logic as Based on the Theory of Types" (1908), and *Principia Mathematica* (1910, 1912, 1913), but also his *An Essay on the Foundations of Geometry* (1897), and *Introduction to Mathematical Philosophy* (1919).

Russell's Work in Analytic Philosophy

In much the same way that Russell used logic in an attempt to clarify issues in the foundations of mathematics, he also used logic in an attempt to clarify issues in philosophy. As one of the founders of analytic philosophy, Russell made significant contributions to a wide variety of areas, including metaphysics, epistemology, ethics and political theory, as well as to the history of philosophy. Underlying these various projects was not only Russell's use of logical analysis, but also his long-standing aim of discovering whether, and to what extent, knowledge is possible. "There is one great question," he writes in 1911. "Can human beings *know* anything, and if so, what and how? This question is really the most essentially philosophical of all questions."[1]

More than this, Russell's various contributions were also unified by his views concerning both the centrality of scientific knowledge and the importance of an underlying scientific methodology that is common to both philosophy and science. In the case of philosophy, this methodology expressed itself through Russell's use of logical analysis. In fact, Russell often claimed that he had more confidence in his methodology than in any particular philosophical conclusion.

Russell's conception of philosophy arose in part from his idealist origins.[2] This is so even though he believed that his one, true revolution in philosophy came with his break from idealism. Russell saw that the idealist doctrine of internal relations led to a series of contradictions regarding asymmetrical (and other) relations necessary for mathematics. Thus, in 1898, he abandoned idealism and his Kantian methodology in favour of a pluralistic realism. Emerging from the idealism that he had encountered as a student at Cambridge, Russell became famous for his defense of the "new realism" and for his "new philosophy of logic", emphasizing as it did the importance of modern logic for philosophical analysis. The underlying themes of this "revolution", such as his belief in pluralism, his emphasis upon anti-psychologism, and the importance of science, remained central to Russell's philosophy for the remainder of his life.[3]

Russell's methodology consisted of the making and testing of hypotheses through the weighing of evidence (hence Russell's comment that he wished to emphasize the "scientific method" in philosophy[4]), together with a rigorous analysis of problematic propositions using the machinery of first-order logic. It was Russell's belief that by using the new logic of his day, philosophers would be able to exhibit the underlying "logical form" of natural language statements. A statement's logical form, in turn, would help philosophers resolve problems of reference associated with the ambiguity and vagueness of natural language. Thus, just as we distinguish three separate sense of "is" -- the is of predication, the is of identity, and the is of existence -- and exhibit these three senses by using three separate logical notations -- Px, $x=y$, and $\exists x$ respectively -- we will also discover other ontologically significant distinctions by being aware of a sentence's correct logical form. On Russell's view, the subject matter of philosophy is then distinguished from that of the sciences only by the generality and the *a prioricity* of philosophical statements, not by the underlying methodology of the discipline. In philosophy, as in mathematics, Russell believed that it was by applying logical machinery and insights that advances would be made.

Russell's most famous example of his "analytic" methodology concerns denoting phrases such as

descriptions and proper names. In his *Principles of Mathematics*, Russell had adopted the view that every denoting phrase (for example, "Scott", "blue", "the number two", "the golden mountain") was assumed to refer to an existing entity. By the time his landmark article, "On Denoting", appeared two years later, in 1905, Russell had modified this extreme realism and had instead become convinced that denoting phrases need not possess a theoretical unity. While logically proper names (words such as "this" or "that" which refer to sensations of which an agent is immediately aware) do have referents associated with them, descriptive phrases (such as "the smallest number less than pi") should be viewed a collection of quantifiers (such as "all" and "some") and propositional functions (such as "x is a number"). As such, they are not to be viewed as referring terms but, rather, as "incomplete symbols". In other words, they should be viewed as symbols that take on meaning within appropriate contexts, but that are meaningless in isolation.

Thus, in the sentence

> (1) The present King of France is bald,

the definite description "The present King of France" plays a role quite different from that of a proper name such as "Scott" in the sentence

> (2) Scott is bald.

Letting K abbreviate the predicate "is a present King of France" and B abbreviate the predicate "is bald", Russell assigns sentence (1) the logical form

> (1′) There is an x such that (i) Kx, (ii) for any y, if Ky then $y=x$, and (iii) Bx.

Alternatively, in the notation of the predicate calculus, we have

$$\exists x[Kx \ \& \ \forall y(Ky \rightarrow y=x) \ \& \ Bx].$$

In contrast, by allowing s to abbreviate the name "Scott", Russell assigns sentence (2) the very different logical form

> (2′) Bs.

This distinction between various logical forms allows Russell to explain three important puzzles. The first concerns the operation of the Law of Excluded Middle and how this law relates to denoting terms. According to one reading of the Law of Excluded Middle, it must be the case that either "The present King of France is bald" is true or "The present King of France is not bald" is true. But if so, both sentences appear to entail the existence of a present King of France, clearly an undesirable result. Russell's analysis shows how this conclusion can be avoided. By appealing to analysis (1′), it follows that there is a way to deny (1) without being committed to the existence of a present King of France,

namely by accepting that "It is not the case that there exists a present King of France who is bald" is true.

The second puzzle concerns the Law of Identity as it operates in (so-called) opaque contexts. Even though "Scott is the author of *Waverley*" is true, it does not follow that the two referring terms "Scott" and "the author of *Waverley*" are interchangeable in all contexts. Thus although "George wanted to know whether Scott was the the author of *Waverley*" is true, "George wanted to know whether Scott was Scott" is, presumably, false. Russell's distinction between the logical forms associated with the use of proper names and definite descriptions shows why this is so.

To see this we once again let s abbreviate the name "Scott". We also let w abbreviate "*Waverley*" and A abbreviate the two-place predicate "is the author of". It then follows that the sentence

(3) $s=s$

is not equivalent to the sentence

(4) $\exists x[Axw \,\&\, \forall y(Ayw \rightarrow y=x) \,\&\, x=s]$.

The third puzzle relates to true negative existential claims, such as the claim "The golden mountain does not exist". Here, once again, by treating definite descriptions as having a logical form distinct from that of proper names, Russell is able to give an account of how a speaker may be committed to the truth of a negative existential without also being committed to the belief that the subject term has reference. That is, the claim that Scott does not exist is false since

(5) $\sim\exists x(x=s)$

is self-contradictory. (After all, there must exist something that is identical to s since it is a logical truth that s is identical to itself!) In contrast, the claim that a golden mountain does not exist may be true since, assuming that G abbreviates the predicate "is golden" and M abbreviates the predicate "is a mountain", there is nothing contradictory about

(6) $\sim\exists x(Gx \,\&\, Mx)$.

Russell's emphasis upon logical analysis also had consequences for his metaphysics. In response to the traditional problem of the external world which, it is claimed, arises since the external world can be known only by inference, Russell developed his famous 1910 distinction between "knowledge by acquaintance and knowledge by description". He then went on, in his 1918 lectures on logical atomism, to argue that the world itself consists of a complex of logical atoms (such as "little patches of colour") and their properties. Together they form the atomic facts which, in turn, are combined to form logically complex objects. What we normally take to be inferred entities (for example, enduring physical objects) are then understood to be "logical constructions" formed from the immediately given entities of sensation, viz., "sensibilia". It is only these latter entities that are known non-inferentially and with

certainty. According to Russell, the philosopher's job is to discover a logically ideal language that will exhibit the true nature of the world in such a way that the speaker will not be misled by the casual surface structure of natural language. Just as atomic facts (the association of universals with an appropriate number of individuals) may be combined into molecular facts in the world itself, such a language would allow for the description of such combinations using logical connectives such as "and" and "or". In addition to atomic and molecular facts, Russell also held that general facts (facts about "all" of something) were needed to complete the picture of the world. Famously, he vacillated on whether negative facts were also required.

Russell's most important writings relating to these topics include not only "On Denoting" (1905), but also his "Knowledge by Acquaintance and Knowledge by Description" (1910), "The Philosophy of Logical Atomism" (1918, 1919), "Logical Atomism" (1924), *The Analysis of Mind* (1921), and *The Analysis of Matter* (1927).

Russell's Social and Political Philosophy

Russell's social influence stems from three main sources: his long-standing social activism, his many writings on the social and political issues of his day, and his popularizations of technical writings in philosophy and the natural sciences.

Among Russell's many popularizations are his two best selling works, *The Problems of Philosophy* (1912) and *A History of Western Philosophy* (1945). Both of these books, as well as his numerous but less famous books popularizing science, have done much to educate and inform generations of general readers. Naturally enough, Russell saw a link between education, in this broad sense, and social progress. At the same time, Russell is also famous for suggesting that a widespread reliance upon evidence, rather than upon superstition, would have enormous social consequences: "I wish to propose for the reader's favourable consideration," says Russell, "a doctrine which may, I fear, appear wildly paradoxical and subversive. The doctrine in question is this: that it is undesirable to believe a proposition when there is no ground whatever for supposing it true."[5]

Still, Russell is best known in many circles as a result of his campaigns against the proliferation of nuclear weapons and against western involvement in the Vietnam War during the 1950s and 1960s. However, Russell's social activism stretches back at least as far as 1910, when he published his *Anti-Suffragist Anxieties,* and to 1916, when he was convicted and fined in connection with anti-war protests during World War I. Following his conviction, he was also dismissed from his post at Trinity College, Cambridge. Two years later, he was convicted a second time. The result was six months in prison. Russell also ran unsuccessfully for Parliament (in 1907, 1922, and 1923) and, together with his second wife, founded and operated an experimental school during the late 1920s and early 1930s.

Although he became the third Earl Russell upon the death of his brother in 1931, Russell's radicalism continued to make him a controversial figure well through middle-age. While teaching in the United States in the late 1930s, he was offered a teaching appointment at City College, New York. The

appointment was revoked following a large number of public protests and a 1940 judicial decision which found him morally unfit to teach at the College.

In 1954 he delivered his famous "Man's Peril" broadcast on the BBC, condemning the Bikini H-bomb tests. A year later, together with Albert Einstein, he released the Russell-Einstein Manifesto calling for the curtailment of nuclear weapons. In 1957 he was a prime organizer of the first Pugwash Conference, which brought together a large number of scientists concerned about the nuclear issue. He became the founding president of the Campaign for Nuclear Disarmament in 1958 and was once again imprisoned, this time in connection with anti-nuclear protests in 1961. The media coverage surrounding his conviction only served to enhance Russell's reputation and to further inspire the many idealistic youths who were sympathetic to his anti-war and anti-nuclear protests.

During these controversial years Russell also wrote many of the books that brought him to the attention of popular audiences. These included his *Principles of Social Reconstruction* (1916), *A Free Man's Worship* (1923), *On Education* (1926), *Why I Am Not a Christian* (1927), *Marriage and Morals* (1929), *The Conquest of Happiness* (1930), *The Scientific Outlook* (1931), and *Power: A New Social Analysis* (1938).

Upon being awarded the Nobel Prize for Literature in 1950, Russell used his acceptance speech to emphasize, once again, themes related to his social activism and to warn of the dangers associated with nuclear war.

Russell's Writings

A Selection of Russell's Articles

- (1905) "On Denoting", *Mind,* 14, 479-493. Repr. in Russell, Bertrand, *Essays in Analysis,* London: Allen & Unwin, 1973, 103-119.
- (1908) "Mathematical Logic as Based on the Theory of Types", *American Journal of Mathematics,* 30, 222-262. Repr. in Russell, Bertrand, *Logic and Knowledge,* London: Allen & Unwin, 1956, 59-102, and in van Heijenoort, Jean, *From Frege to Gödel,* Cambridge, Mass.: Harvard University Press, 1967, 152-182.
- (1910) "Knowledge by Acquaintance and Knowledge by Description", *Proceedings of the Aristotelian Society,* 11, 108-128. Repr. in Russell, Bertrand, *Mysticism and Logic,* London: Allen & Unwin, 1963, 152-167.
- (1912) "On the Relations of Universals and Particulars", *Proceedings of the Aristotelian Society,*

12, 1-24. Repr. in Russell, Bertrand, *Logic and Knowledge,* London: Allen & Unwin, 1956, 105-124.

- (1918, 1919) "The Philosophy of Logical Atomism", *Monist*, 28, 495-527; 29, 32-63, 190-222, 345-380. Repr. in Russell, Bertrand, *Logic and Knowledge,* London: Allen & Unwin, 1956, 177-281.
- (1924) "Logical Atomism", in Muirhead, J.H., *Contemporary British Philosophers,* London: Allen & Unwin, 1924, 356-383. Repr. in Russell, Bertrand, *Logic and Knowledge,* London: Allen & Unwin, 1956, 323-343.

A Selection of Russell's Books

- (1896) *German Social Democracy*, London: Longmans, Green.
- (1897) *An Essay on the Foundations of Geometry,* Cambridge: At the University Press.
- (1900) *A Critical Exposition of the Philosophy of Leibniz*, Cambridge: At the University Press.
- (1903) *The Principles of Mathematics,* Cambridge: At the University Press.
- (1910, 1912, 1913) (with Alfred North Whitehead) *Principia Mathematica,* 3 vols, Cambridge: Cambridge University Press. Second edition, 1925 (Vol. 1), 1927 (Vols 2, 3). Abridged as *Principia Mathematica to *56,* Cambridge: Cambridge University Press, 1962.
- (1912) *The Problems of Philosophy*, London: Williams and Norgate; New York: Henry Holt and Company.
- (1914) *Our Knowledge of the External World,* Chicago and London: The Open Court Publishing Company.
- (1916) *Principles of Social Reconstruction,* London: George Allen & Unwin. Repr. as *Why Men Fight*, New York: The Century Company, 1917.
- (1917) *Political Ideals,* New York: The Century Company.
- (1919) *Introduction to Mathematical Philosophy,* London: George Allen & Unwin; New York: The Macmillan Company.
- (1921) *The Analysis of Mind,* London: George Allen & Unwin; New York: The Macmillan Company.
- (1923) *A Free Man's Worship*, Portland, Maine: Thomas Bird Mosher. Repr. as *What Can A Free Man Worship?,* Girard, Kansas: Haldeman-Julius Publications, 1927.
- (1926) *On Education, Especially in Early Childhood,* London: George Allen & Unwin. Repr. as *Education and the Good Life*, New York: Boni & Liveright, 1926. Abridged as *Education of Character*, New York: Philosophical Library, 1961.
- (1927) *The Analysis of Matter*, London: Kegan Paul, Trench, Trubner; New York: Harcourt, Brace.
- (1927) *An Outline of Philosophy*, London: George Allen & Unwin. Repr. as *Philosophy,* New York: W.W. Norton, 1927.
- (1927) *Why I Am Not a Christian,* London: Watts, New York: The Truth Seeker Company.
- (1929) *Marriage and Morals,* London: George Allen & Unwin; New York: Horace Liveright.
- (1930) *The Conquest of Happiness,* London: George Allen & Unwin; New York: Horace Liveright.
- (1931) *The Scientific Outlook*, London: George Allen & Unwin; New York: W.W. Norton.

- (1938) *Power: A New Social Analysis,* London: George Allen & Unwin; New York: W.W. Norton.
- (1940) *An Inquiry into Meaning and Truth*, London: George Allen & Unwin; New York: W.W. Norton.
- (1945) *A History of Western Philosophy*, New York: Simon and Schuster; London: George Allen & Unwin, 1946.
- (1948) *Human Knowledge: Its Scope and Limits,* London: George Allen & Unwin; New York: Simon and Schuster.
- (1949) *Authority and the Individual,* London: George Allen & Unwin; New York: Simon and Schuster.
- (1949) *The Philosophy of Logical Atomism,* Minneapolis, Minnesota: Department of Philosophy, University of Minnesota. Repr. as *Russell's Logical Atomism,* Oxford: Fontana/Collins, 1972.
- (1954) *Human Society in Ethics and Politics,* London: George Allen & Unwin; New York: Simon and Schuster.
- (1959) *My Philosophical Development,* London: George Allen & Unwin; New York: Simon and Schuster.
- (1967, 1968, 1969) *The Autobiography of Bertrand Russell,* 3 vols, London: George Allen & Unwin; Boston and Toronto: Little Brown and Company (Vols 1 and 2), New York: Simon and Schuster (Vol. 3).

Major Anthologies of Russell's Writings

- (1910) *Philosophical Essays,* London: Longmans, Green.
- (1918) *Mysticism and Logic and Other Essays,* London and New York: Longmans, Green. Repr. as *A Free Man's Worship and Other Essays,* London: Unwin Paperbacks, 1976.
- (1928) *Sceptical Essays,* London: George Allen & Unwin; New York: W.W. Norton.
- (1935) *In Praise of Idleness,* London: George Allen & Unwin; New York: W.W. Norton.
- (1950) *Unpopular Essays,* London: George Allen & Unwin; New York: Simon and Schuster.
- (1956) *Logic and Knowledge: Essays, 1901-1950,* London: George Allen & Unwin; New York: The Macmillan Company.
- (1956) *Portraits From Memory and Other Essays,* London: George Allen & Unwin; New York: Simon and Schuster.
- (1957) *Why I am Not a Christian and Other Essays on Religion and Related Subjects,* London: George Allen & Unwin; New York: Simon and Schuster.
- (1961) *The Basic Writings of Bertrand Russell, 1903-1959,* London: George Allen & Unwin; New York: Simon and Schuster.
- (1969) *Dear Bertrand Russell,* London: George Allen & Unwin; Boston: Houghton Mifflin.
- (1973) *Essays in Analysis,* London: George Allen & Unwin.
- (1992) *The Selected Letters of Bertrand Russell,* London: Penguin Press.

The Collected Papers of Bertrand Russell

The Bertrand Russell Editorial Project is currently in the process of publishing Russell's *Collected Papers.* When complete, these volumes will bring together all of Russell's writings, excluding his correspondence and previously published monographs.

In Print

- Vol. 1: *Cambridge Essays, 1888-99,* London, Boston, Sydney: George Allen & Unwin, 1983.
- Vol. 2: *Philosophical Papers, 1896-99,* London and New York: Routledge, 1990.
- Vol. 3: *Toward the Principles of Mathematics,* London and New York: Routledge, 1994.
- Vol. 4: *Foundations of Logic, 1903-05,* London and New York: Routledge, 1994.
- Vol. 6: *Logical and Philosophical Papers, 1909-13,* London and New York: Routledge, 1992.
- Vol. 7: *Theory of Knowledge: The 1913 Manuscript*, London, Boston, Sydney: George Allen & Unwin, 1984.
- Vol. 8: *The Philosophy of Logical Atomism and Other Essays, 1914-19,* London: George Allen & Unwin, 1986.
- Vol. 9: *Essays on Language, Mind and Matter, 1919-26,* London: Unwin Hyman, 1988.
- Vol. 10: *A Fresh Look at Empricism, 1927-42,* London and New York: Routledge, 1996.
- Vol. 11: *Last Philosophical Testament, 1943-68,* London and New York: Routledge, 1997.
- Vol. 12: *Contemplation and Action, 1902-14,* London, Boston, Sydney: George Allen & Unwin, 1985.
- Vol. 13: *Prophecy and Dissent, 1914-16,* London: Unwin Hyman, 1988.
- Vol. 14: *Pacifism and Revolution, 1916-18,* London and New York: Routledge, 1995.
- Vol. 15: *Uncertain Paths to Freedom: Russia and China, 1919-1922,* London and New York: Routledge, 2000.

Planned and Forthcoming

- Vol. 5: *Toward Principia Mathematica, 1906-08.*
- Vol. 16: *Labour and Internationalism, 1922-24.*
- Vol. 17: *Behaviourism and Education, 1925-28.*
- Vol. 18: *Science, Sex and Society, 1929-31.*
- Vol. 19: *Fascism and Other Depression Legacies, 1931-33.*
- Vol. 20: *Fascism and Other Depression Legacies, 1933-34.*
- Vol. 21: *How to Keep the Peace: The Pacifist Dilemma, 1934-36.*
- Vol. 22: *The Superior Virtue of the Oppressed and Other Essays, 1936-39.*
- Vol. 23: *The Problems of Democracy, 1940-44.*
- Vol. 24: *Civilization and the Bomb, 1944-47.*
- Vol. 25: *Civilization and the Bomb, 1948-50.*
- Vol. 26: *Respectability At Last, 1950-51.*
- Vol. 27: *Respectability At Last, 1952-53.*
- Vol. 28: *Man's Peril, 1954-56.*
- Vol. 29: *The Campaign for Nuclear Disarmament, 1957-60.*
- Vol. 30: *A New Plan for Peace and Other Essays, 1960-64.*

- Vol. 31: *The Vietnam Campaign, 1965-70.*
- Vol. 32: *Newly Discovered Papers.*
- Vol. 33: *Indexes.*

Bibliography

- Selected Articles
- Selected Books

Selected Articles

- Broad, C.D. (1973) "Bertrand Russell, as Philosopher", *Bulletin of the London Mathematical Society*, 5, 328-341.
- Carnap, Rudolf (1931) "The Logicist Foundations of Mathematics", *Erkenntnis,* 2, 91-105. Repr. in Benacerraf, Paul, and Hilary Putnam (eds), *Philosophy of Mathematics,* 2nd ed., Cambridge: Cambridge University Press, 1983, 41-52; in Klemke, E.D. (ed.), *Essays on Bertrand Russell,* Urbana: University of Illinois Press, 1970, 341-354; and in Pears, David F. (ed.), *Bertrand Russell: A Collection of Critical Essays,* Garden City, New York: Anchor Books, 1972, 175-191.
- Church, Alonzo (1976) "Comparison of Russell's Resolution of the Semantical Antinomies With That of Tarski", *Journal of Symbolic Logic,* 41, 747-760.
- Gandy, R.O. (1973) "Bertrand Russell, as Mathematician", *Bulletin of the London Mathematical Society*, 5, 342-348.
- Gödel, Kurt (1944) "Russell's Mathematical Logic", in Schilpp, Paul Arthur (ed.), *The Philosophy of Bertrand Russell,* 3rd ed., New York: Tudor, 1951, 123-153. Repr. in Benacerraf, Paul, and Hilary Putnam (eds), *Philosophy of Mathematics,* 2nd ed., Cambridge: Cambridge University Press, 1983, 447-469; and in Pears, David F. (ed.) (1972) *Bertrand Russell: A Collection of Critical Essays,* Garden City, New York: Anchor Books, 192-226.
- Hylton, Peter W. (1990) "Logic in Russell's Logicism", in Bell, David, and Neil Cooper (eds), *The Analytic Tradition: Philosophical Quarterly Monographs,* Vol. 1, Cambridge: Blackwell, 137-172.
- Irvine, A.D. (1989) "Epistemic Logicism and Russell's Regressive Method", *Philosophical Studies*, 55, 303-327.
- Irvine, A.D. (1996) "Bertrand Russell and Academic Freedom", *Russell,* n.s.16, 5-36.
- Kaplan, David (1970) "What is Russell's Theory of Descriptions?", in Yourgrau, Wolfgang, and Allen D. Breck, (eds), *Physics, Logic, and History,* New York: Plenum, 277-288. Repr. in Pears, David F. (ed.), *Bertrand Russell: A Collection of Critical Essays,* Garden City, New York: Anchor Books, 1972, 227-244.
- Lycan, William (1981) "Logical Atomism and Ontological Atoms", *Synthese*, 46, 207-229.
- Monro, D.H. (1960) "Russell's Moral Theories", *Philosophy,* 35, 30-50. Repr. in Pears, David F. (ed.), *Bertrand Russell: A Collection of Critical Essays,* Garden City, New York: Anchor Books, 1972, 325-355.

- Putnam, Hilary (1967) "The Thesis that Mathematics is Logic", in Schoenman, Ralph (ed.), *Bertrand Russell: Philosopher of the Century*, London: Allen & Unwin, 273-303. Repr. in Putnam, Hilary, *Mathematics, Matter and Method*, Cambridge: Cambridge University Press, 1975, 12-42.
- Quine, W.V. (1938) "On the Theory of Types", *Journal of Symbolic Logic*, 3, 125-139.
- Ramsey, F.P. (1926) "Mathematical Logic", *Mathematical Gazette*, 13, 185-194. Repr. in Ramsey, Frank Plumpton, *The Foundations of Mathematics*, London: Kegan Paul, Trench, Trubner, 1931, 62-81; in Ramsey, Frank Plumpton, *Foundations*, London: Routledge and Kegan Paul, 1978, 213-232; and in Ramsey, Frank Plumpton, *Philosophical Papers*, Cambridge: Cambridge University Press, 1990, 225-244.]
- Schultz, Bart (1992) "Bertrand Russell in Ethics and Politics", *Ethics*, 102, 594-634.
- Strawson, Peter F. (1950) "On Referring", *Mind*, 59, 320-344. Repr. in Flew, Anthony (ed.), *Essays in Conceptual Analysis*, London: Macmillan, 1960, 21-52, and in Klemke, E.D. (ed.), *Essays on Bertrand Russell*, Urbana: University of Illinois Press, 1970, 147-172.
- Weitz, Morris (1944) "Analysis and the Unity of Russell's Philosophy", in Schilpp, Paul Arthur (ed.), *The Philosophy of Bertrand Russell*, 3rd ed., New York: Tudor, 1951, 55-121.

Selected Books

- Blackwell, Kenneth (1985) *The Spinozistic Ethics of Bertrand Russell*, London: George Allen & Unwin.
- Blackwell, Kenneth, and Harry Ruja (1994) *A Bibliography of Bertrand Russell*, 3 vols, London: Routledge.
- Chomsky, Noam (1971) *Problems of Knowledge and Freedom: The Russell Lectures*, New York: Vintage.
- Clark, Ronald William (1975) *The Life of Bertrand Russell*, London: J. Cape.
- Clark, Ronald William (1981) *Bertrand Russell and His World*, London: Thames and Hudson.
- Dewey, John, and Horace M. Kallen (eds) (1941) *The Bertrand Russell Case*, New York: Viking.
- Eames, Elizabeth R. (1969) *Bertrand Russell's Theory of Knowledge*, London: George Allen & Unwin.
- Eames, Elizabeth R. (1989) *Bertrand Russell's Dialogue with his Contemporaries*, Carbondale: Southern Illinois University Press.
- Feinberg, Barry, and Ronald Kasrils (eds) (1969) *Dear Bertrand Russell*, London: George Allen & Unwin.
- Feinberg, Barry, and Ronald Kasrils (1973, 1983) *Bertrand Russell's America*, 2 vols, London: George Allen & Unwin.
- Grattan-Guinness, I. (1977) *Dear Russell, Dear Jourdain: A Commentary on Russell's Logic, Based on His Correspondence with Philip Jourdain*, New York: Columbia University Press.
- Griffin, Nicholas (1991) *Russell's Idealist Apprenticeship*, Oxford: Clarendon.
- Hager, Paul J. (1994) *Continuity and Change in the Development of Russell's Philosophy*, Dordrecht: Nijhoff.
- Hardy, Godfrey H. (1942) *Bertrand Russell and Trinity*, Cambridge: Cambridge University Press, 1970.

- Hylton, Peter W. (1990) *Russell, Idealism, and the Emergence of Analytic Philosophy,* Oxford: Clarendon.
- Irvine, A.D. (ed.) (1998) *Bertrand Russell: Critical Assessments*, 4 vols, London: Routledge.
- Irvine, A.D., and G.A. Wedeking (eds) (1993) *Russell and Analytic Philosophy,* Toronto: University of Toronto Press.
- Jager, Ronald (1972) *The Development of Bertrand Russell's Philosophy,* London: George Allen & Unwin.
- Klemke, E.D. (ed.) (1970) *Essays on Bertrand Russell,* Urbana: University of Illinois Press.
- Monk, Ray (1996) *Bertrand Russell: The Spirit of Solitude*, London: J. Cape.
- Monk, Ray, and Anthony Palmer (eds) (1996) *Bertrand Russell and the Origins of Analytic Philosophy*, Bristol: Thoemmes Press.
- Moorehead, Caroline (1992) *Bertrand Russell,* New York: Viking.
- Nakhnikian, George (ed.) (1974) *Bertrand Russell's Philosophy,* London: Duckworth.
- Park, Joe (1963) *Bertrand Russell on Education,* Columbus: Ohio State University Press.
- Patterson, Wayne (1993) *Bertrand Russell's Philosophy of Logical Atomism*, New York: Lang.
- Pears, David F. (1967) *Bertrand Russell and the British Tradition in Philosophy,* London: Collins.
- Pears, David F. (ed.) (1972) *Bertrand Russell: A Collection of Critical Essays,* New York: Doubleday.
- Roberts, George W. (ed.) (1979) *Bertrand Russell Memorial Volume,* London: Allen & Unwin.
- Rodriguez-Consuegra, Francisco A. (1991) *The Mathematical Philosophy of Bertrand Russell: Origins and Development,* Basel: Birkhauser Verlag.
- Ryan, Alan (1988) *Bertrand Russell: A Political Life,* New York: Hill and Wang.
- Savage, C. Wade, and C. Anthony Anderson (eds) (1989) *Rereading Russell: Essays on Bertrand Russell's Metaphysics and Epistemology,* Minneapolis: University of Minnesota Press.
- Schilpp, Paul Arthur (ed.) (1944) *The Philosophy of Bertrand Russell,* Chicago: Northwestern University; 3rd ed., New York: Harper and Row, 1963.
- Schoenman, Ralph (ed.) (1967) *Bertrand Russell: Philosopher of the Century*, London: Allen & Unwin.
- Slater, John G. (1994) *Bertrand Russell*, Bristol: Thoemmes.
- Tait, Katharine (1975) *My Father Bertrand Russell,* New York: Harcourt Brace Jovanovich.
- Vellacott, Jo (1980) *Bertrand Russell and the Pacifists in the First World War*, Brighton, Sussex: Harvester Press.
- Wittgenstein, Ludwig (1921) *Logisch-philosophische Abhandlung.* Trans. as *Tractatus Logico-Philosophicus,* London: Kegan Paul, Trench, Trubner, 1922.
- Wittgenstein, Ludwig (1956) *Remarks on the Foundations of Mathematics,* Oxford: Blackwell.
- Wood, Alan (1957) *Bertrand Russell: The Passionate Sceptic*, London: Allen & Unwin.

Other Internet Resources

- Bertrand Russell Archives
- Bertrand Russell Research Centre
- Bertrand Russell Society

- Ludwig Wittgenstein Page
- Russell: The Journal of Bertrand Russell Studies
- University of St Andrew's MacTutor History of Mathematics Archive -- Bertrand Russell
- Writings by Bertrand Russell

Related Entries

analytic philosophy | atomism: logical | descriptions | Frege, Gottlob | Gödel, Kurt | knowledge: by acquaintance vs. description | logic: classical | logical constructions | logicism | mathematics, philosophy of | Moore, George Edward | *Principia Mathematica* | propositional function | Russell's paradox | type theory | Whitehead, Alfred North | Wittgenstein, Ludwig

A. D. Irvine
andrew.irvine@ubc.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 7, 1995
Content last modified: January 2, 2001

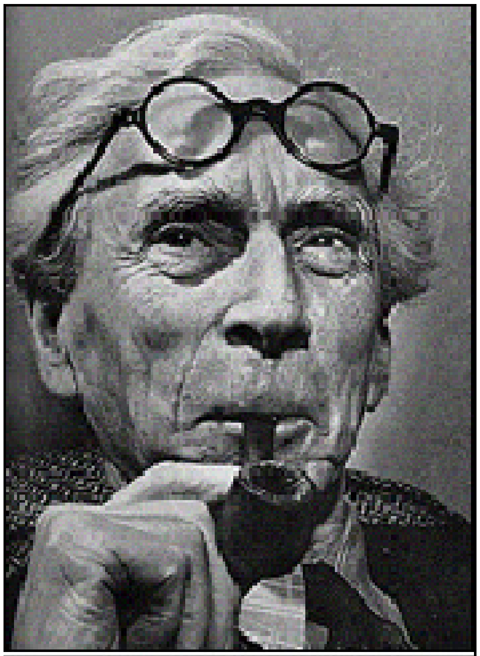


Photo credit: Larry Burrows

Stanford Encyclopedia of Philosophy
Audio Supplement to Bertrand Russell

Sound Clips of Bertrand Russell Speaking

The following two sound clips are from Bertrand Russell's Nobel Prize acceptance speech. They appear here courtesy of the United Nations (Unesco) Archives (tape #823, 18.12.50) and the Bertrand Russell Society Library.

Bertrand Russell on Desire

Soundclip in real time

- Real Audio format

Alternative formats

- QuickTime format (Apple)
- .wav format (Microsoft)
- .au format (Sun)
- .snd format (NeXT)

Bertrand Russell on Political Theory

Soundclip in real time

- Real Audio format

Alternative formats

- QuickTime format (Apple)
- .wav format (Microsoft)
- .au format (Sun)
- .snd format (NeXT)

Return to Bertrand Russell

First published: October 23, 1997
Content last modified: August 5, 2000

Russell's Paradox

Russell's paradox is the most famous of the logical or set-theoretical paradoxes. The paradox arises within naive set theory by considering the set of all sets that are not members of themselves. Such a set appears to be a member of itself if and only if it is not a member of itself, hence the paradox.

Some sets, such as the set of all teacups, are not members of themselves. Other sets, such as the set of all non-teacups, are members of themselves. Call the set of all sets that are not members of themselves *S*. If *S* is a member of itself, then by definition it must not be a member of itself. Similarly, if *S* is not a member of itself, then by definition it must be a member of itself. Discovered by Bertrand Russell in 1901, the paradox has prompted much work in logic, set theory and the philosophy and foundations of mathematics.

History of the paradox

Russell appears to have discovered his paradox in May of 1901[1] while working on his *Principles of Mathematics* (1903). Cesare Burali-Forti, an assistant to Giuseppe Peano, had discovered a similar antinomy in 1897 when he noticed that since the set of ordinals is well-ordered, it, too, must have an ordinal. However, this ordinal must be both an element of the set of all ordinals and yet greater than every such element.

Russell wrote to Gottlob Frege with news of his paradox on June 16, 1902. The paradox was of significance to Frege's logical work since, in effect, it showed that the axioms Frege was using to formalize his logic were inconsistent. Specifically, Frege's Rule V, which states that two sets are equal if and only if their corresponding functions coincide in values for all possible arguments, requires that an expression such as $f(x)$ be considered both a function of the argument f and a function of the argument x.

In effect, it was this ambiguity that allowed Russell to construct S in such a way that it could both be and not be a member of itself.

Russell's letter arrived just as the second volume of Frege's *Grundgesetze der Arithmetik* (*The Basic Laws of Arithmetic*, 1893, 1903) was in press. Immediately appreciating the difficulty that the paradox posed, Frege hastily added an appendix to the *Grundgesetze* to discuss Russell's discovery. In this appendix Frege observes that the consequences of Russell's paradox are not immediately clear. For example, "Is it always permissible to speak of the extension of a concept, of a class? And if not, how do we recognize the exceptional cases? Can we always infer from the extension of one concept's coinciding with that of a second, that every object which falls under the first concept also falls under the second? These are questions," Frege notes, that have been "raised by Mr Russell's communication."[2]

Because of these kinds of worries, Frege eventually felt forced to abandon many of his views. Russell himself was also concerned about the paradox and so, like Frege, he hastily composed an appendix for his soon to be released *Principles of Mathematics*. Entitled "Appendix B: The Doctrine of Types", the appendix represents Russell's first attempt at developing a workable theory of types.

Significance of the paradox

The significance of Russell's paradox can be seen once it is realized that, using classical logic, all sentences follow from a contradiction. (For example, assuming both P and $\sim P$, we can prove any arbitrary Q as follows: from P we can obtain $P \vee Q$ by the rule of Addition, and then from $P \vee Q$ and $\sim P$ we can obtain Q by the rule of Disjunctive Syllogism.) In the eyes of many, it therefore appeared that no mathematical proof could be trusted once it was discovered that the set theory underlying all of mathematics was contradictory.

Russell's paradox stems from the idea that any coherent condition may be used to determine a set. Attempts at resolving the paradox therefore typically have concentrated on various means of restricting the principles governing the existence of sets. Naive set theory contained the so-called unrestricted comprehension (or abstraction) axiom. This is an axiom to the effect that any predicate expression, $P(x)$, containing x as a free variable will determine a set. The set's members will be exactly those objects that satisfy $P(x)$, namely every x that is P.[3] It is now generally agreed that such an axiom must be either abandoned or modified.[4]

Russell's response to the paradox is contained in his so-called *theory of types*. His basic idea is that we can avoid reference to S (the set of all sets that are not members of themselves) by arranging all sentences into a hierarchy. This hierarchy will consist of sentences (at the lowest level) about individuals, sentences (at the next lowest level) about sets of individuals, sentences (at the next lowest level) about sets of sets of individuals, etc. It is then possible to refer to all objects for which a given condition (or predicate) holds only if they are all at the same level or of the same "type".

This solution depends upon the assumption, often called the *vicious circle principle*, that the meaning of a propositional function cannot be specified until one specifies the exact range of objects which are candidates for satisfying it. From this it follows that these objects cannot meaningfully include anything that is defined in terms of the function itself. The result is that propositional functions, and their corresponding propositions, will need to be arranged in a hierarchy of the kind Russell proposes.

Although Russell first introduced the idea of a theory of types in his *Principles of Mathematics*, type-theory found its mature expression five years later in his 1908 article "Mathematical Logic as Based on the Theory of Types" and in the monumental work he co-authored with Alfred North Whitehead, *Principia Mathematica* (1910, 1912, 1913). In its details, Russell's type theory thus came to admit of two versions, the "simple theory" and the "ramified theory". Both versions have been criticized for being too ad hoc to eliminate the paradox successfully.

Other responses to the paradox include those of David Hilbert and the formalists (whose basic idea was to allow the use of only finite, well-defined and constructible objects, together with rules of inference that were deemed to be absolutely certain), and of Luitzen Brouwer and the intuitionists (whose basic idea was that one cannot assert the existence of a mathematical object unless one can also indicate how to go about constructing it).

Yet a fourth response to the paradox was Ernst Zermelo's 1908 axiomatization of set theory. Zermelo's axioms were designed to resolve Russell's paradox by restricting the naive comprehension principle. ZF, the axiomatization generally used today, is a modification of Zermelo's theory developed primarily by Abraham Fraenkel.

These four responses to the paradox have helped logicians develop an explicit awareness of the nature of formal systems and of the kinds of metalogical results that are today commonly associated with them.

Bibliography

- Frege, Gottlob, 1902, "Letter to Russell", in van Heijenoort, Jean, *From Frege to Gödel*, Cambridge: Harvard University Press, 1967, 126-128.
- Frege, Gottlob, 1903, "The Russell Paradox", in Frege, Gottlob, *The Basic Laws of Arithmetic*, Berkeley: University of California Press, 1964, 127-143. Abridged and reprinted in Irvine, A.D., *Bertrand Russell: Critical Assessments*, vol. 2, London: Routledge, 1999, 1-3.
- Hallett, Michael, 1984, *Cantorian Set Theory and Limitation of Size*, Oxford: Clarendon.
- Menzel, Christopher, 1984, "Cantor and the Burali-Forti Paradox", *Monist*, 67, 92-107.
- Moore, Gregory, 1982, *Zermelo's Axiom of Choice*, New York: Springer.
- Russell, Bertrand, 1902, "Letter to Frege", in van Heijenoort, Jean, *From Frege to Gödel*, Cambridge, Mass.: Harvard University Press, 1967, 124-125.
- Russell, Bertrand, 1903, "Appendix B: The Doctrine of Types", in Russell, Bertrand, *Principles of Mathematics*, Cambridge: Cambridge University Press, 1903, 523-528.

- Russell, Bertrand, 1908, "Mathematical Logic as Based on the Theory of Types", *American Journal of Mathematics*, 30, 222-262. Repr. in Russell, Bertrand, *Logic and Knowledge,* London: Allen & Unwin, 1956, 59-102, and in van Heijenoort, Jean, *From Frege to Gödel*, Cambridge, Mass.: Harvard University Press, 1967, 152-182.
- Russell, Bertrand, 1944, "My Mental Development", in Schilpp, Paul Arthur, *The Philosophy of Bertrand Russell*, 3rd edn, New York: Tudor, 3-20.
- Russell, Bertrand, 1959, *My Philosophical Development,* London and New York: Routledge, 1995.
- Russell, Bertrand, 1967/1968/1969, *The Autobiography of Bertrand Russell,* 3 vols, Boston and Toronto: Little, Brown and Company.
- Whitehead, Alfred North, and Bertrand Russell, 1910/1912/1913, *Principia Mathematica,* 3 vols, Cambridge: Cambridge University Press. Second edition, 1925 (Vol. 1), 1927 (Vols 2, 3). Abridged as *Principia Mathematica to *56,* Cambridge: Cambridge University Press, 1962.

Other Internet Resources

- Bertrand Russell Archives
- Russell: The Journal of Bertrand Russell Studies
- Russell's Antinomy

Related Entries

Cantor, Georg | Frege, Gottlob | Frege, Gottlob: logic, theorem, and foundations for arithmetic | logic: paraconsistent | mathematics: inconsistent | Peano, Giuseppe | *Principia Mathematica* | Russell, Bertrand | type theory | Whitehead, Alfred North

Acknowledgements

My thanks goes to Chris Menzel for his helpful feedback on an earlier version of this entry.

A. D. Irvine
andrew.irvine@ubc.ca

Table of Contents

First published: December 7, 1995
Content last modified: June 29, 2002

Stanford Encyclopedia of Philosophy
Notes to Russell's Paradox

Notes

1. Exactly when the discovery of the paradox took place is not completely clear. Russell initially states that he came across the paradox "in June 1901" (see Russell 1944, p. 13). Later he reports that the discovery took place "in the spring of 1901" (see Russell 1959, p. 58). Later still he reports that he came across the paradox, not in June, but in May of that year (see Russell 1967/1968/1969, volume 3 (1969), p. 221).

2. See Frege 1903, p. 127.

3. It is worth noting that even prior to Russell's discovery this principle had not been universally accepted. Georg Cantor, for example, rejected it in favour of what was, in effect, a distinction between sets and classes, recognizing that some properties (such as the property of being an ordinal) produced collections that were too big to be sets, and that an assumption to the contrary would result in an inconsistent theory. (For further details see Menzel 1984, Moore 1982, and Hallett 1984.)

4. One exception is paraconsistent set theory. Paraconsistent set theory retains an unrestricted comprehension axiom but abandons classical logic, substituting a paraconsistent logic in its place. See the entries on paraconsistent logic and inconsistent mathematics in this Encyclopedia.

A. D. Irvine
andrew.irvine@ubc.ca

First published: June 29, 2002
Content last modified: June 29, 2002

Paraconsistent Logic

The development of *paraconsistent logic* was initiated in order to challenge the logical principle that anything follows from contradictory premises, *ex contradictione quodlibet (ECQ)*. Let $\models$ be a relation of logical consequence, defined either semantically or proof-theoretically. Let us say that $\models$ is *explosive* iff for every formula *A* and *B*, {*A* , ~*A*} $\models$ *B*. Classical logic, intuitionistic logic, and most other standard logics are explosive. A logic is said to be *paraconsistent* iff its relation of logical consequence is not explosive.

The modern history of paraconsistent logic is relatively short. Yet the subject has already been shown to be an important development in logic for many reasons. These involve the motivations for the subject, its philosophical implications and its applications. In the first half of this article, we will review some of these. In the second, we will give some idea of the basic technical constructions involved in paraconsistent logics. Further discussion can be found in the references given at the end of the article.

Motivation and Applications

Inconsistent but Non-Trivial Theories

A most telling reason for paraconsistent logic is the fact that there are theories which are inconsistent but non-trivial. Clearly, once we admit the existence of such theories, their underlying logics must be paraconsistent. Examples of inconsistent but non-trivial theories are easy to produce. An example can be derived from the history of science. (In fact, many examples can be given from this area.) Consider Bohr's theory of the atom. According to this, an electron orbits the nucleus of the atom without radiating energy. However, according to Maxwell's equations, which formed an integral part of the theory, an electron which is accelerating in orbit must radiate energy. Hence Bohr's account of the behaviour of the atom was inconsistent. Yet, patently, not everything concerning the behavior of electrons was inferred from it. Hence, whatever inference mechanism it was that underlay it, this must have been paraconsistent.

Dialetheias (True Contradictions)

The importance of paraconsistent logic also follows if, more contentiously, but as some people have argued, there are true contradictions (dialetheias), i.e., there are sentences, *A*, such that both *A* and ~*A* are true. If there are dialetheias then some inferences of the form $\{A, \sim A\} \models B$ must fail. For only true conclusions follow validly from the true premises. Hence logic has to be paraconsistent. A plausible example of dialetheia is the *liar paradox*. Consider the sentence: This sentence is not true. There are two options: either the sentence is true or it is not. Suppose it is true. Then what it says is the case. Hence the sentence is not true. Suppose, on the other hand, it is not true. This is what it says. Hence the sentence is true. In either case it is both true and not true.

Automated Reasoning

Paraconsistent logic is motivated not only by philosophical considerations, but also by its applications and implications. One of the applications is *automated reasoning* (*information processing*). Consider a computer which stores a large amount of information. While the computer stores the information, it is also used to operate on it, and, crucially, to infer from it. Now it is quite common for the computer to contain inconsistent information, because of mistakes by the data entry operators or because of multiple sourcing. This is certainly a problem for database operations with theorem-provers, and so has drawn much attention from computer scientists. Techniques for removing inconsistent information have been investigated. Yet all have limited applicability, and, in any case, are not guaranteed to produce consistency. (There is no algorithm for logical falsehood.) Hence, even if steps are taken to get rid of contradictions when they are found, an underlying paraconsistent logic is desirable if hidden contradictions are not to generate spurious answers to queries.

Belief Revision

As a part of artificial intelligence research, *belief revision* is one of the areas that have been studied widely. Belief revision is the study of rationally revising bodies of belief in the light of new evidence.

Notoriously, people have inconsistent beliefs. They may even be rational in doing so. For example, there may be apparently overwhelming evidence for both something and its negation. There may even be cases where it is in principle impossible to eliminate such inconsistency. For example, consider the "paradox of the preface". A rational person, after thorough research, writes a book in which they claim $A_1, \ldots, A_n$. But they are also aware that no book of any complexity contains only truths. So they rationally believe $\sim(A_1 \;\&\; \ldots \;\&\; A_n)$ too. Hence, principles of rational belief revision must work on inconsistent sets of beliefs. Standard accounts of belief revision, e.g., that of Gärdenfors *et al.*, all fail to do this since they are based on classical logic. A more adequate account is based on a paraconsistent logic.

Mathematical Significance

Other applications of paraconsistent logic concern theories of mathematical significance. Examples of such theories are formal *semantics* and *set theory*.

Semantics is the study that aims to spell out a theoretical understanding of meaning. Most accounts of semantics insist that to spell out the meaning of a sentence is, in some sense, to spell out its truth-conditions. Now, *prima facie* at least, truth is a predicate characterised by the Tarski T-scheme:

$$T(\boldsymbol{A}) \leftrightarrow A,$$

where A is a sentence and $\boldsymbol{A}$ is its name. But given any standard means of self-reference, e.g., arithmetisation, one can construct a sentence, B, which means that $\sim T(\boldsymbol{B})$. The T-scheme gives that $T(\boldsymbol{B}) \leftrightarrow \sim T(\boldsymbol{B})$. It then follows that $T(\boldsymbol{B}) \;\&\; \sim T(\boldsymbol{B})$. (This is, of course, just the liar paradox.)

The situation is similar in set theory. The naive, and intuitively correct, axioms of set theory are the *Comprehension Schema* and *Extensionality Principle*:

$$(\exists y)(x)(x \in y \leftrightarrow A)$$

$$(x)(x \in y \leftrightarrow x \in z) \rightarrow y = z$$

where x does not occur free in A. As was discovered by Russell, any theory that contains the Comprehension Schema is inconsistent. For putting ‘$y \notin y$’ for A in the Comprehension Schema and instantiating the existential quantifier to an arbitrary such object ‘r’ gives:

$$(y)(y \in r \leftrightarrow y \notin y)$$

So, instantiating the universal quantifier to ‘r’ gives:

$$r \in r \leftrightarrow r \notin r$$

It then follows that $r \in r$ & $r \notin r$.

The standard approaches to these problems of inconsistency are, by and large, ones of expedience. However, a paraconsistent approach makes it possible to have theories of truth and sethood in which the fundamental intuitions about these notions are respected. The contradictions may be allowed to arise, but these need not infect the rest of the theory.

The Philosophical Significance of Gödel's Theorem

Paraconsistent logic also has important philosophical ramifications. One example of this concerns Gödel's theorem. One version of Gödel's first incompleteness theorem states that for any consistent axiomatic theory of arithmetic, which can be recognised to be sound, there will be an arithmetic truth - viz., its Gödel sentence - not provable in it, but which can be established as true by intuitively correct reasoning. The heart of Gödel's theorem is, in fact, a paradox that concerns the sentence, *G*, 'This sentence is not provable'. If *G* is provable, then it is true and so not provable. Thus *G* is proved. Hence *G* is true and so unprovable. If an underlying paraconsistent logic is used to formalise the arithmetic, and the theory therefore allowed to be inconsistent, the Gödel sentence may well be provable in the theory (essentially by the above reasoning). So a paraconsistent approach to arithmetic overcomes the limitations of arithmetic that are supposed (by many) to follow from Gödel's theorem.

Systems of Paraconsistent Logic

The foregoing discussion indicates some of the motivations for paraconsistent logic, its applications and implications. We will now indicate some of the main approaches to paraconsistency. There are many different paraconsistent logics. Most of them can be defined in terms of a semantics which allows both *A* and ~*A* to hold in an interpretation. Validity is then defined in terms of the preservation of holding in an interpretation, and so ECQ fails. We will illustrate this with four kinds of propositional paraconsistent logics: *non-adjunctive*, *non-truth-functional*, *many-valued*, and *relevant*. (Paraconsistent quantified logics are straightforward extensions of these.) In all the following systems, not only ECQ fails, but so does the Disjunctive Syllogism (DS), defined as the following inference rule: $\{A, {\sim}A \vee B\} \vdash B$. In particular, then, if one defines the material conditional, $A \supset B$, as ${\sim}A \vee B$ (as usual) then *modus ponens* for this fails.

- Non-Adjunctive Systems
- Non-Truth-Functional Logics
- Many-Valued Systems
- Relevant Logics

Non-Adjunctive Systems

Let us start with non-adjunctive systems, so called because the inference from *A* and *B* to *A & B* fails. The first of these to be produced was also the first formal paraconsistent logic. This was Jaskowski's *discussive* (or *discursive*) *logic*. In a discourse, each participant puts forward some information, beliefs, or opinions. What is true in a discourse is the sum of opinions given by participants. Each participant's opinions are taken to be self-consistent, but may be inconsistent with those of others. To formalise this idea, take an interpretation, *I*, to be one for a standard modal logic, say *S5*. Each participant's belief set is the set of sentences true in a possible world in *I*. Thus, *A* holds in *I* iff *A* holds at *some* world in *I*. Clearly, one may have both *A* and ~*A* (but not *A & ~A*) holding in an interpretation. Since *modus ponens* for ⊃ fails, Jaskowski introduced a connective he called discussive implication, $\supset_d$, defined as ($\Diamond A \supset B$). It is easy to check that in *S5* discussive implication satisfies *modus ponens*.

Non-Truth-Functional Logics

The study of non-truth-functional systems was initiated by da Costa (who has also produced several other kinds of system). The main idea here was to maintain the apparatus of some positive logic, say classical or intuitionistic, but to allow negation in an interpretation to behave non-truth-functionally. Thus, take an interpretation to be a function which maps formulas to 1 or 0; *&* , ∨, and → behave in the usual (classical) way, but the value of ~*A* is independent of that of *A*. In particular, both may take the value 1. Negation has no significant properties under these semantics. Various properties of negation may be obtained by adding further constraints on interpretations. If we add the requirements that, for any *A*, either *A* or ~*A* must take the value 1 (giving the Law of Excluded Middle) and that whenever ~~*A* takes the value 1, so does *A*, we obtain the core of da Costa's systems C_i , for finite *i*. If we start with an appropriate semantics for positive intuitionist logic, and proceed in the same way, we obtain da Costa's logic $C\omega$. If we write A^{o} for *~(A & ~A)* then it is natural to take it as expressing the consistency of *A*. Further postulates constraining how A^{o} behaves differentiate between the C_i systems for finite *i*.

Many-Valued Systems

Perhaps the simplest way of generating a paraconsistent logic, first proposed by Asenjo, is to use a many-valued logic, that is, a logic with more than two truth values. The formulas which hold in a many-valued interpretations are those which have a value said to be *designated*. A many-valued logic will therefore be paraconsistent if it allows both a formula and its negation to be designated. The simplest strategy is to use three truth values: *true (only)* and *false (only)*, which function as in classical logic, and *both truth and false* (which, naturally, is a fixed point for negation). Both varieties of truth are designated. This is the approach of the paraconsistent logic *LP*. If one adds a fourth value, *neither true nor false*, which behaves in an appropriate way, one obtains Dunn's semantics for First Degree Entailment. If one takes the truth values to be the real numbers between 0 and 1, with a suitable set of designated values, the logic will be a natural paraconsistent fuzzy logic.

Relevant Logics

Relevant logics were pioneered by Anderson and Belnap. World-semantics for them were developed by R. and V.Routley and Meyer. In an interpretation for such logics, conjunction and disjunction behave in the usual way. But each world, *w*, has an associate world, *w**; and ~*A* is true at *w* iff *A* is false, not at *w*, but *w**. Thus, if *A* is true at *w*, but false at *w**, *A* & ~*A* is true at *w*. To obtain the standard relevant logics, one needs to add the constraint that *w*** = *w*. As is clear, negation in these semantics is an intensional operator. (There are also versions of world-semantics for relevant logics based on Dunn's four-valued semantics. In these, negation is extensional.)

The concern with relevant logics is not so much with negation as with a conditional connective, → (satisfying *modus ponens*). Semantics for this are obtained by furnishing each interpretation with a *ternary* relation, *R*. In the simplified semantics of Priest, Sylvan and Restall, worlds are divided into normal and non-normal. If *w* is a normal world, *A* → *B* is true at *w* iff at all worlds where *A* is true, *B* is true. If *w* is non-normal, *A* → *B* is true at *w* iff for all *x*, *y*, such that *Rwxy*, if *A* is true at *x*, *B* is true at *y*. (Validity is defined as truth preservation over *normal* worlds.) This gives the basic relevant logic, *B*. Stronger logics, such as the logic *R*, are obtained by adding constraints on the ternary relation. Further details concerning relevant logics can be found in the article on that topic in this encyclopedia.

Bibliography

For Paraconsistent Logic and Paraconsistency in general, see:

- Priest, G., Routley, R., and Norman, J. (eds.) *Paraconsistent Logic: Essays on the Inconsistent*, Philosophia Verlag, München, 1989.
- Priest, G. "Paraconsistent Logic", *Handbook of Philosophical Logic* (second edition), forthcoming.

On Dialetheism, see:

- Priest, G. "Logic of Paradox", *Journal of Philosophical Logic*, Vol. VIII, pp. 219-241, 1979.
- Priest, G. *In Contradiction: A Study of the Transconsistent*, Martinus Nijhoff, Dordrecht, 1987.

For Automated Reasoning, see:

- Belnap, N.D., Jr. "A Useful Four-valued Logic: How a computer should think", *Entailment: The Logic of Relevance and Necessity*, Vol II, A.R. Anderson, N.D. Belnap, Jr, and J.M. Dunn, Princeton University Press, 1992, first appeared as "A Usuful Four-valued Logic", *Modern Use of Multiple-valued Logic*, J.M. Dunn and G. Epstein (eds.), D.Reidel Publishing Company, Dordrecht, 1977, and "How a Computer Should Think", *Comtemporary Aspects of Philosophy*, G. Ryle (ed.), Oriel Press, 1977.

For Belief Revision, see:

- Restall, G. and Slaney, J. "Realistic Belief Revision", Technical Report: TR-ARP-2-95, Automated Reasoning Project, Australian National University, 1995.
- Tanaka, K. "Paraconsistent Belief Revision", to appear.

For Non-Adjunctive Systems, see:

- Jaskowski, S. "Propositional Calculus for Contradictory Deductive Systems", *Studia Logica*, Vol. XXIV, pp. 143-157, 1969, first published as "Rachunek zdah dla systemow dedukcyjnych sprzecznych", *Studia Societatis Scientiarun Torunesis*, Sectio A, Vol. I, No. 5, pp. 55-77, 1948.
- da Costa, N.C.A. and Dubikajtis, L. "On Jaskowski's Discussive Logic", *Non-Classical Logics, Modal Theory and Computability*, A.I. Arruda, N.C.A. da Costa and R. Chuaqui (eds.), North-Holland Publishing Company, Amsterdam, pp.37-56, 1977.
- Schotch, P.K. and Jennings, R.E. "Inference and Necessity", *Journal of Philosophical Logic*, Vol. IX, pp. 327-340, 1980.

For Non-Truth-Functional Systems, see:

- da Costa, N.C.A. "On the Theory of Inconsistent Formal Systems", *Notre Dame Journal of Formal Logic*, Vol. XV, No. 4, pp. 497-510, 1974.
- da Costa, N.C.A. and Alves, E.H. "Semantical Analysis of the Calculi Cn", *Notre Dame Journal of Formal Logic*, Vol. XVIII, No. 4, pp. 621-630, 1977.
- Loparic, A. "Une etude semantique de quelques calculs propositionnels", *Comptes Rendus Hebdomadaires des Seances de l'Academic des Sciences*, Paris 284, pp. 835-838, 1977.

For Many-Valued Systems, see:

- Asenjo, F.G. "A Calculus of Antinomies", *Notre Dame Journal of Formal Logic*, Vol. XVI, pp. 103-5, 1966.
- Dunn, J.M. "Intuitive Semantics for First Degree Entailment and Coupled Trees", *Philosophicl Studies*, Vol. XXIX, pp. 149-68, 1976.
- Kotas, J. and da Costa, N. "On the Problem of Jaskowski and the Logics of Lukasiewicz", *Non-Classical Logic, Model Theory and Computability*, A.I. Arruda, N.C.A da Costa, and R. Chuaqui (eds.), North Holland Publishing Company, Amsterdam, pp. 127-39, 1977.

For Relevant Systems, see:

- Dunn, J.M. "Relevant Logic and Entailment", *Handbook of Philosophical Logic, Vol. III: Alternatives to Classical Logic*, D. Gabbay and F. Guenthner (eds.), D.Reidel Publishing Company, Dordrecht, pp. 117-224, 1986.
- Routley, R., Plumwood, V., Meyer, R.K., and Brady, R.T. *Relevant Logics and Their Rivals*, Atascadero, Ridgeview, CA, 1982.

- Restall, G. "Simplified Semantics for Relevant Logics (and some of their rivals)", *Journal of Philosophical Logic*, Vol. XXII, pp. 481-511, 1993.

Other Internet Resources

[Please contact the authors with suggestions.]

Related Entries

dialetheism [dialethism] | logic: relevance | mathematics: inconsistent

Graham Priest
University of Melbourne
g.priest@unimelb.edu.au
and
Koji Tanaka
Macquarie University
Koji.Tanaka@mq.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 24, 1996
Content last modified: December 5, 2000

Relevance Logic

Relevance logics are non-classical logics. Called ‘relevance logics’ in North America and ‘relevant logics’ in Britain and Australasia, these systems developed as attempts to avoid the paradoxes of material and strict implication. Among the paradoxes of material implication are

- $p \rightarrow (q \rightarrow p)$.
- $\sim p \rightarrow (p \rightarrow q)$.
- $(p \rightarrow q) \vee (q \rightarrow r)$.

Among the paradoxes of strict implication are the following:

- $(p \,\&\, \sim p) \rightarrow q$.
- $p \rightarrow (q \rightarrow q)$.
- $p \rightarrow (q \vee \sim q)$.

Relevance logicians claim that what is unsettling about these so-called paradoxes is that in each of them the antecedent seems irrelevant to the consequent.

In addition, relevance logicians have had qualms about certain inferences that classical logic makes valid. For example, the inference

> The moon is made of green cheese. Therefore, either it is raining in Ecuador now or it is not.

Again here there seems to be a failure of relevance. The conclusion seems to have nothing to do with the premise. Relevance logicians have attempted to construct logics that reject theses and arguments that commit "fallacies of relevance".

At this point some confusion is natural about what relevant logicians have attempted to do. They have not given formal criteria of relevance that any true implication must meet, altough some relevant logicans have interpreted the semantics for relevance logic using informal notions of relevance (see the section "Semantics" below). Instead, relevant logic is relevant in two ways: (1) Relevance logics do not force us to accept any irrelevances. That is, they do not make valid any of the paradoxes. (2) Some relevance logics, through their proof theory, yield a relevant notion of proof (see the section "Proof Theory"

below).

In this article we will give a brief and relatively non-technical overview of the field of relevant logic.

Semantics

Our exposition of relevant logic is backwards to most found in the literature We will begin, rather than end, with the semantics, since most philosophers at present are semantically inclined.

The semantics that I present here is the ternary relation semantics due to Richard Routley and Robert K. Meyer. There are algebraic semantics due to J. Michael Dunn and Alasdair Urquhart and operational semantics produced by Kit Fine. These systems are interesting in their own right, but we do not have room to discuss them here.

The idea behind the ternary relation semantics is rather simple. Consider C.I. Lewis' attempt to avoid the paradoxes of material implication. He added a new connective to classical logic, that of strict implication. In post-Kripkean semantic terms, $A ⥽ B$ is true at a world w if and only if for all w' such that w' is accessible to w, either A fails in w' or B obtains there. Now, in Kripke's semantics for modal logic, the accessibility relation is a binary relation. It holds between pairs of worlds. Unfortunately, from a relevant point of view, the theory of strict implication is still irrelevant. That is, we still make valid formulae like $p ⥽ (q ⥽ q)$. We can see quite easily that the Kripke truth condition forces this formula on us.

Like the semantics of modal logic, the semantics of relevance logic relativises truth of formulae to worlds. But Routley and Meyer go modal logic one better and use a three-place relation on worlds. This allows there to be worlds at which $q \rightarrow q$ fails and that in turn allows worlds at which $p \rightarrow (q \rightarrow q)$ fails. Their truth condition for $\rightarrow$ on this semantics is the following:

> $A \rightarrow B$ is true at a world a if and only if for all worlds b and c such that $Rabc$ (R is the accessibility relation) either A is false at b or B is true at c.

For new customers, it takes some time to get used to this truth condition. But with a little work it can be seen to be just a generalisation of the Kanger-Kripke truth condition for strict implication (just set $b = c$).

But how should this accessibility relation be interpreted philosophically? Oddly, there has not been very much work on this, but there has been some. Mares (1997) uses a theory of information due to David Israel and John Perry (see Israel and Perry (1990)). On this interpreation, in addition to other information a world contains informational links, such as laws of nature, conventions, and so on. Thus, for example, a Newtonian world will contain the information that all matter attracts all other matter. In information-theoretic terms, this world contains the information that two things' being material carries the information that they attract each other. On this view, *Rabc* if and only if, according to the links in *a*, all the information carried by what obtains in *b* is contained in *c*. Thus, for example, if *a* is a Newtonian world and the information that *x* and *y* are material is contained in *b*, then the information that *x* and *y* attract each other is contained in *c*.

Another similar interpretation is given in Barwise (1993) and developed in Restall (1996). On this view, worlds are taken to be information-theoretic "sites". *Rabc* means that *a* is an information-theoretic channel between *b* and *c*. Both this channel-theoretic interpretation of the accessibiliity relation and the other information-theoretic interpretation attempt to provide the formal semantics with an informal notion of relevance. On these interpretations, what is needed for an implication to be true is that the antecedent carry the information that the consequent obtains. The antecedent must be informationally relevant to the consequent.

By itself, the ternary relation is not sufficient to avoid all the paradoxes of implication. Given what we have said so far, it is not clear how the semantics can avoid paradoxes such as $(p \,\&\, {\sim}p) \rightarrow q$ and $p \rightarrow (q \vee {\sim}q)$. These paradoxes are avoided by the inclusion of inconsistent and non-bivalent worlds in the semantics. For, if there were no worlds at which $p \,\&\, {\sim}p$ holds, then, according to our truth condition for the arrow, $(p \,\&\, {\sim}p) \rightarrow q$ would also hold everywhere. Likewise, if $q \vee {\sim}q$ held at every world, then $p \rightarrow (q \vee {\sim}q)$ would be universally true.

This brings us to the semantics for negation. The use of non-bivalent and inconsistent worlds requires a non-classical truth condition for negation. In the early 1970s, Richard and Val Routley invented their "star operator" to treat negation. The operator is an operator on worlds. For each world *a*, there is a world *a**. And

> ~*A* is true at *a* if and only if *A* is false at *a**.

Once again, we have the difficulty of interpreting a part of the formal semantics. What may be the nicest interpretation of the Routley star is that of Dunn (1993). Dunn uses a binary relation, *C*, on worlds. '*Cab*' means that *b* is compatable with *a*. *a**, then, is the maximal world (the world containing the most information) that is compatable with *a*.

There are other semantics for negation. One, due to Dunn and developed by Routley, is a four-valued

semantics. This semantics is treated in the entry on paraconsitent logics.

Proof Theory

There is now a lage variety of approaches to proof theory for relevant logics. There is a Gentzen system for the negation-free fragment of the logic **R** by J.M. Dunn and an elegant and very general Gentzen-style approach called "Display Logic" recently developed by Nuel Belnap. But here I will only deal with an treatment that most philosophers will find somewhat familiar, that is, the natural deduction system due to Anderson and Belnap.

Anderson and Belnap's natural deduction system is based on Fitch's natural deduction systems for classical and intuitionistic logic. The easiest way to understand this technique is by looking at an example.

1. $A_{\{1\}}$	hyp
2. $(A \rightarrow B)_{\{2\}}$	hyp
3. $B_{\{1,2\}}$	1,2, $\rightarrow$E
4. $((A \rightarrow B) \rightarrow B)_{\{1\}}$	2,3, $\rightarrow$I
5. $A \rightarrow ((A \rightarrow B) \rightarrow B)$	1,4, $\rightarrow$I

The numbers in set brackets indicate the assuptions used to prove the formula. We will call them ‘indices’. The idea here is that for an assumption to be counted as helping to generate the conclusion, an index denoting the assumption must be appear in the deduction and at some later point be discharged. This ensures that each premise is really used in the deduction. This natural deduction system gives an intuitive understanding of relevance in proofs. The indices keep track of which assumptions are used. For an argument to be valid in this system, all assumptions stated must really be used.

Now, it might seem that the system of indices allows irrelevant premises to creep in. One way in which it might appear that irrelevances can intrude is through the use of a rule of conjunction introduction. That is, it might seem that we can always add in an irrelevant premise by doing, say, the following:

1. $A_{\{1\}}$	hyp
2. $B_{\{2\}}$	hyp
3. $(A \& B)_{\{1,2\}}$	1,2, &I
4. $B_{\{1,2\}}$	3, &E
5. $(B \rightarrow B)_{\{1\}}$	2,4, $\rightarrow$I

6. $A \rightarrow (B \rightarrow B)$	1,5, $\rightarrow$I

To a relevance logician, the first premise is completely out of place here. To block moves like this, Anderson and Belnap give the following conjunction introduction rule:

From A_i and B_i to infer $(A \& B)_i$.

This rule says that two formulae must have the same index before the rule of conjunction introduction can be used.

There is, of course, a lot more to the natural deduction system, but this will suffice as for our current purposes. The theory of relevance that is captured by at least some relevant logics can be understood by how the corresponding natural deduction system understands a real use of a premise and how the rules are allowed to access premises.

Systems of Relevance Logic

Historically, the central systems of relevance logic have been the logic **E** of entailment and the system **R** or relevant implication. **E** was supposed to capture strict relevant implication. But, when a necessity operator and the appropriate modal axioms were added to **R** (to produce the logic **NR**), it was discovered that the resulting modal system was different from **E**. This has left some relevant logicians with a quandary. They have to decide whether to take **NR** to be the system of strict relevant implication, or to claim that **NR** was somehow defficient and that **E** stands as the system of strict relevant implication.

On the other hand, there are those relevance logicians who reject both **R** and **E**. On one hand there is Arnon Avron who has used semantic arguments to motivate logics stronger than **R**. On the other hand there are logicians like Ross Brady, John Slaney, Steve Giabrione, Richard Sylvan, Graham Priest and Greg Restall who have argued for the acceptance of systems weaker than **R** or **E**. Among the points in favour of weaker these systems is that, unlike **R** or **E**, many of them are decidable.

On an extreme end of the spectrum is the logic **S** of R.K. Meyer, Errol Martin and Robin Dwyer. This logic contains no theorems of the form $A \rightarrow A$. In other words, according to **S**, no proposition implies itself and no argument of the form '*A*, therefore *A*' is valid. Thus, this logic does not make valid any circular arguments.

Applications of Relevance Logic

Apart from the motivating applications of providing better formalisms of our pre-formal notions of implication and entailment, relevance logic has been put to various uses in philosophy and computer science. Here I will list just a few.

Dunn has developed a theory of intrinsic and essential properties based on relevant logic. This is his theory of *relevant predication*. Briefly put, a thing i has a property F relevantly if $\forall x(x=i \rightarrow F(x))$. Informally, an object has a property relevantly if being that thing relevantly implies having that property. Since the truth of the consequent of a relevant implication is by itself insufficient for the truth of that implication, things can have properties irrelevantly as well as relevantly. Dunn's formulation would seem to capture at least one sense in which we use the notion of an intrisic property. Adding modality to the language allows for a formilisation of the notion of an essential property.

Meyer has produced a variant of Peano arithmetic based on the relevance logic, **R**. Meyer gives a finitary proof that his relevant arithmetic does not have 0 = 1 as a theorem. Thus Meyer solves one of Hilbert's central problems in the context of relevant arithmetic: He shows using finitary means that relevant arithmetic is absolutely consistent. Unfortunately, as Meyer and Friedman have shown, relevant arithmetic does not contain all of the theorems of classical Peano arithmetic. Hence we cannot infer from this that classical Peano arithmetic is absolutely consistent (see Meyer and Friedman (1992)).

In a similar vein, Ross Brady and others have used weak relevant logics as bases for set theories. Brady shows that a weak relevant logic together with some set-theoretic axioms that include a naive comprehension principle is not trivial. That is, not every proposition can be proved in this system.

Anderson (1967) formulates a system of deontic logic based on **R**. This system avoids some of the standard problems with more traditional deontic logics. For example, the rule of necessitation from *A*'s being a theorem to *OA*'s being a theorem is rejected. Thus, it does not say that all theorems ought to be the case.

Mares and Fuhrmann (1995) present a theory of counterfactual conditionals based on relevant logic. This theory avoids the analogs of the paradoxes of implication that appear in standard logics of counterfactuals.

Relevant logics have been used in computer science as well as in philosophy. Linear logics, a branch of logic discoverd by the French logician Girard, is a logic of computational resources. Linear logic is, in fact, a weak relevant logic with the addition of two operators.

Bibliography

An extremely good, although slightly out of date, bibliography on relevance logic was put together by Robert Wolff and is in Anderson, Belnap and Dunn (1992). What follows is a brief list of some of the more influential works in the field and works that are referred to above.

Books on Relevance Logic and Introductions to the Field:

- Anderson, A.R. and N.D. Belnap, Jr. (1975) *Entailment: The Logic of Relevance and Necessity*, Princeton, Princeton University Press, Volume I. Anderson, A.R. N.D. Belnap, Jr. and J.M. Dunn (1992) *Entailment*, Volume II. [These are both collections of slightly modified articles on relevance logic together with a lot of material unique to these volumes. Excellent work and still the standard books on the subject. But they are very technical and quite difficult.]
- Read, S. (1988), *Relevant Logic*, Oxford: Blackwell. [A very interesting and fun book. Idiosyncratic, but philosophically adept and excellent on the pre-history and early history of relevance logic.]
- Routley, R., R.K. Meyer, V. Plumwood and R. Brady (1983), *Relevant Logics and its Rivals*, Volume I, Atascardero, CA: Ridgeview. [A very useful book for formal results especially about the semantics of relevance logics. The introduction and philosophical remarks are full of "Richard Routleyisms". They tend to be Routley's views rather than the views of the other authors and are fairly radical even for relevant logicians. Volume II is currently in the works. It is a large editorial project headed by Ross Brady.]
- Dunn, J.M. (1986) "Relevance Logic and Entailment" in F. Guenthner and D. Gabbay (eds.), *Handbook of Philosophical Logic*, Volume 3, Dordrecht: Reidel pp 117--224. [The best long introduction to relevance logic currently published. Dunn and Restall are currently rewriting this chapter for the next edition of the *Handbook*.]

Other Works Cited:

- Anderson, A.R. (1967) "Some Nasty Problems in the Formal Logic of Ethics" *Nous* 1 pp 354-360.
- Barwise, J. (1993) "Constraints, Channels and the Flow of Information" in P.Aczel, et al. (eds), *Situation Theory and Its Applications*, Volume 3, Stanford: CSLI pp 3-27.
- Brady, R.T. (1989) "The Non-Triviality of Dialectical Set Theory" in G. Priest, R. Routley and J. Norman (eds.), *Paraconsistent Logic*, Munich: Philosophia Verlag pp 437-470
- Dunn, J.M. (1993) "Star and Perp" *Philosophical Perspectives* 7 pp 331-357
- Israel, D. and J. Perry (1990) "What is Information?" in P.P. Hanson (ed.), *Information, Language, and Cognition*, Vancouver: University of British Columbia Press pp 1-19
- Mares, E.D. (1997) "Relevant Logic and the Theory of Information" *Synthese* 109 pp 345-360
- Mares, E.D. and A. Fuhrmann (1995) "A Relevant Theory of Conditionals" *Journal of Philosophical Logic* 24 pp 645-665
- Meyer, R.K. and H. Friedman (1992) "Whither Relevant Arithmetic?" *The Journal of Symbolic Logic* 57 pp 824-831
- Restall, G. (1996) "Information Flow and Relevant Logics" in J. Seligman and D. Westerstahl (eds), *Logic, Language and Computation*, Volume 1, Stanford: CSLI pp 463-478.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

logic: modal | logic: paraconsistent | mathematics: inconsistent

Edwin D. Mares
Edwin.Mares@vuw.ac.nz

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 17, 1998
Content last modified: June 17, 1998

Stanford Encyclopedia of Philosophy

Modal Logic

A modal is an expression (like 'necessarily' or 'possibly') that is used to qualify the truth of a judgement. Modal logic is, strictly speaking, the study of the deductive behavior of the expressions 'it is necessary that' and 'it is possible that'. However, the term 'modal logic' may be used more broadly for a family of related systems. These include logics for belief, for tense and other temporal expressions, for the deontic (moral) expressions such as 'it is obligatory that' and 'it is permitted that', and many others. An understanding of modal logic is particularly valuable in the formal analysis of philosophical argument, where expressions from the modal family are both common and confusing. Modal logic also has important applications in computer science.

1. What is Modal Logic?

Narrowly construed, modal logic studies reasoning that involves the use of the expressions 'necessarily' and 'possibly'. However, the term 'modal logic' is used more broadly to cover a family of logics with similar rules and a variety of different symbols.

A list describing the best known of these logics follows.

Logic	Symbols	Expressions Symbolized
Modal Logic	$\Box$	It is necessary that ..
	$\Diamond$	It is possible that ..
Deontic Logic	O	It is obligatory that ..
	P	It is permitted that ..
	F	It is forbidden that ..
Temporal Logic	G	It will always be the case that ..
	F	It will be the case that ..
	H	It has always been the case that ..
	P	It was the case that..
Doxastic Logic	Bx	x believes that ..

2. Modal Logics

The most familiar logics in the modal family are constructed from a weak logic called K (after Saul Kripke). Under the narrow reading, modal logic concerns necessity and possibility. A variety of different systems may be developed for such logics using K as a foundation. The symbols of K include '~' for 'not', '$\rightarrow$' for 'if...then', and '$\Box$' for the modal operator 'it is necessary that'. (The connectives '&', '$\vee$', and '$\leftrightarrow$' may be defined from '~' and '$\rightarrow$' as is done in propositional logic.) K results from adding the following to the principles of propositional logic.

Necessitation Rule: If A is a theorem of K, then so is $\Box$A.

Distribution Axiom: $\Box(A\rightarrow B) \rightarrow (\Box A\rightarrow\Box B)$.

(In these principles we use 'A' and 'B' as metavariables ranging over formulas of the language.) According to the Necessitation Rule, any theorem of logic is necessary. The Distribution Axiom says that if it is necessary that if A then B, then if necessarily A then necessarily B.

The operator $\Diamond$ (for 'possibly') can be defined from $\Box$ by letting $\Diamond A = {\sim}\Box{\sim}A$. In K, the operators $\Box$ and $\Diamond$ behave very much like the quantifiers $\forall$ (all) and $\exists$ (some). For example, the definition of $\Diamond$ from $\Box$ mirrors the equivalence of $\forall xA$ with ${\sim}\exists x{\sim}A$ in predicate logic. Furthermore, $\Box(A\&B)$ entails $\Box A\&\Box B$ and vice versa; while $\Box A\vee\Box B$ entails $\Box(A\vee B)$, but *not* vice versa. This reflects the patterns exhibited by the universal quantifier: $\forall x(A\&B)$ entails $\forall xA\&\forall xB$ and vice versa, while $\forall xA\vee\forall xB$ entails $\forall x(A\vee B)$ but not vice versa. Similar parallels between $\Diamond$ and $\exists$ can be drawn. The basis for this correspondence

between the modal operators and the quantifiers will emerge more clearly in the section on Possible Worlds Semantics.

The system K is too weak to provide an adequate account of necessity. The following axiom is not provable in K, but it is clearly desirable.

(M) $\Box A \rightarrow A$

(M) claims that whatever is necessary is the case. Notice that (M) would be incorrect were $\Box$ to be read 'it ought to be that', or 'it was the case that'. So the presence of axiom (M) distinguishes modal from other logics in the modal family. A basic modal logic M results from adding (M) to K. (Some authors call this system T.)

Many logicians believe that M is still too weak to correctly formalize the logic of necessity and possibility. They recommend further axioms to govern the iteration, or repetition of modal operators. Here are two of the most famous iteration axioms:

(4) $\Box A \rightarrow \Box\Box A$

(5) $\Diamond A \rightarrow \Box\Diamond A$

S4 is the system that results from adding (4) to M. Similarly S5 is M plus (5). In S4, the sentence $\Box\Box A$ is equivalent to $\Box A$. As a result, any string of boxes may be replaced by a single box, and the same goes for strings of diamonds. This amounts to the idea that iteration of the modal operators is superfluous. Saying that A is necessarily necessary is considered a uselessly long-winded way of saying that A is necessary. The system S5 has even stronger principles for simplifying strings of modal operators. In S4, a string of operators of *the same kind* can be replaced for that operator; in S5, strings containing both boxes and diamonds are equivalent to the last operator in the string. So, for example, saying that it is possible that A is necessary is the same as saying that A is necessary. A summary of these features of S4 and S5 follows.

S4: $\Box\Box...\Box = \Box$ and $\Diamond\Diamond...\Diamond = \Diamond$

S5: $00...\Box = \Box$ and $00...\Diamond = \Diamond$, where each 0 is either $\Box$ or $\Diamond$

One could engage in endless argument over the correctness or incorrectness of these and other iteration principles for $\Box$ and $\Diamond$. The controversy can be partly resolved by recognizing that the words 'necessarily' and 'possibly', have many different uses. So the acceptability of axioms for modal logic depends on which of these uses we have in mind. For this reason, there is no one modal logic, but rather a whole family of systems built around M. The relationship between these systems is diagrammed in Section 8, and their application to different uses of 'necessarily' and 'possibly' can be more deeply understood by studying their possible world semantics in Section 6.

The system B (for the logician Brouwer) is formed by adding axiom (B) to M.

(B) $A \rightarrow \Box\Diamond A$

It is interesting to note that S5 can be formulated equivalently by adding (B) to S4. The axiom (B) raises an important point about the interpretation of modal formulas. (B) says that if A is the case, then A is necessarily possible. One might argue that (B) should always be adopted in any modal logic, for surely if A is the case, then it is necessary that A is possible. However, there is a problem with this claim that can be exposed by noting that $\Diamond\Box A \rightarrow A$ is provable from (B). So $\Diamond\Box A \rightarrow A$ should be acceptable if (B) is. However, $\Diamond\Box A \rightarrow A$ says that if A is possibly necessary, then A is the case, and this is far from obvious. Why does (B) seem obvious, while one of the things it entails seems not obvious at all? The answer is that there is a dangerous ambiguity in the English interpretation of $A \rightarrow \Box\Diamond A$. We often use the expression ‘If A then necessarily B’ to express that the conditional ‘if A then B’ is necessary. This interpretation corresponds to $\Box(A \rightarrow B)$. On other occasions, we mean that if A, then B is necessary: $A \rightarrow \Box B$. In English, ‘necessarily’ is an adverb, and since adverbs are usually placed near verbs, we have no natural way to indicate whether the modal operator applies to the whole conditional, or to its consequent. For these reasons, there is a tendency to confuse (B): $A \rightarrow \Box\Diamond A$ with $\Box(A \rightarrow \Diamond A)$. But $\Box(A \rightarrow \Diamond A)$ is not the same as (B), for $\Box(A \rightarrow \Diamond A)$ is already a theorem of M, and (B) is not. One must take special care that our positive reaction to $\Box(A \rightarrow \Diamond A)$ does not infect our evaluation of (B). One simple way to protect ourselves is to formulate B in an equivalent way using the axiom: $\Diamond\Box A \rightarrow A$, where these ambiguities of scope do not arise.

3. Deontic Logics

Deontic logics introduce the primitive symbol O for ‘it is obligatory that’, from which symbols P for ‘it is permitted that’ and F for ‘it is forbidden that’ are defined: PA = ~O~A and FA = O~A. The deontic analog of the modal axiom (M): $OA \rightarrow A$ is clearly not appropriate for deontic logic. (Unfortunately, what ought to be is not always the case.) However, a basic system D of deontic logic can be constructed by adding the weaker axiom (D) to K.

(D) $OA \rightarrow PA$

Axiom (D) guarantees the consistency of the system of obligations by insisting that when A is obligatory, A is permissible. A system which obligates us to bring about A, but doesn't permit us to do so, puts us in an inescapable bind. Although some will argue that such conflicts of obligation are at least possible, most deontic logicians accept (D).

$O(OA \rightarrow A)$ is another deontic axiom that seems desirable. Although it is wrong to say that if A is obligatory then A is the case ($OA \rightarrow A$), still, this conditional *ought* to be the case. So some deontic logicians believe that D needs to be supplemented with $O(OA \rightarrow A)$ as well.

Controversy about iteration (repetition) of operators arises again in deontic logic. In some conceptions of obligation, OOA just amounts to OA. ‘It ought to be that it ought to be’ is treated as a sort of stuttering; the extra ‘ought’s do not add anything new. So axioms are added to guarantee the equivalence of OOA and OA. The more general iteration policy embodied in S5 may also be adopted. However, there are conceptions of obligation where distinction between OA and OOA is preserved. The idea is that there are genuine differences between the obligations we *actually* have and the obligations we *should* adopt. So, for example, ‘it ought to be that it ought to be that A’ commands adoption of some obligation which may not actually be in place, with the result that OOA can be true even when OA is false.

4. Temporal Logics

In temporal logic (also known as tense logic), there are two basic operators, G for the future, and H for the past. G is read ‘it always will be that’ and the defined operator F (read ‘it will be the case that’), can be introduced by FA = ~G~A. Similarly H is read: ‘it always was that’ and P (for ‘it was the case that’) is defined by PA=~H~A. A basic system of temporal logic called Kt results from adopting the principles of K for both G and H, along with two axioms to govern the interaction between the past and future operators:

> "Necessitation" Rules: If A is a theorem then so are GA and HA.
>
> Distribution Axioms: G(A→B) → (GA→GB) and H(A→B) → (HA→HB)
>
> Interaction Axioms: A→GPA and A→HFA

The interaction axioms raise questions concerning asymmetries between the past and the future. A standard intuition is that the past is fixed, while the future is still open. The first interaction axiom (A →GPA) conforms to this intuition in reporting that what is the case (A), will at all future times, be in the past (GPA). However A→HFA may appear to have unacceptably deterministic overtones, for it claims, apparently, that what is true now (A) has always been such that it will occur in the future (HFA). However, possible world semantics for temporal logic reveals that this worry results from a simple confusion, and that the two interaction axioms are equally acceptable.

Note that the characteristic axiom of modal logic, (M): □A→A, is not acceptable for either H or G, since A does not follow from ‘it always was the case that A’, nor from ‘it always will be the case that A’. However, it is acceptable in a closely related temporal logic where G is read ‘it is and always will be’, and H is read ‘it is and always was’.

Depending on which assumptions one makes about the structure of time, further axioms must be added to temporal logics. A list of axioms commonly adopted in temporal logics follows. An account of how they depend on the structure of time will be found in the section Possible Worlds Semantics.

GA→GGA and HA→HHA

GGA→GA and HHA→HA

GA→FA and HA→PA

It is interesting to note that certain combinations of past tense and futuere tense operators may be used to express complex tenses in English. For example, FPA, corresponds to sentence A in the future perfect tense, (as in ‘20 seconds from now the light will have changed’). Similarly, PPA expresses the past perfect tense.

For a more detailed discussion of temporal logic, see the entry on temporal logic.

5. Conditional Logics

The founder of modal logic, C. I. Lewis, defined a series of modal logics which did not have □ as a primitive symbol. Lewis was concerned to develop of logic of conditionals that was free of the so called Paradoxes of Material Implication, namely the classical theorems A→(~A→B) and B→(A→B). He introduced the symbol ⥽ for "strict implication" and developed logics where neither A⥽(~A⥽B) nor B ⥽(A⥽B) is provable. The modern practice has been to define A⥽B by □(A→B), and use modal logics governing □ to obtain similar results. However, the provability of such formulas as (A&~A)⥽B in such logics seems at odds with concern for the paradoxes. Anderson and Belnap (1975) have developed systems R (for Relevance Logic) and E (for Entailment) which are designed to overcome such difficulties. These systems require revision of the standard systems of propositional logic. (For a more detailed discussion of relevance logic, see the entry on relevance logic.)

David Lewis (1973) has developed special conditional logics to handle counterfactual expressions, that is, expressions of the form ‘if A *were* to happen then B *would* happen’. (Kvart (1980) is another good source on the topic.) Counterfactual logics differ from those based on strict implication because the former reject while the latter accept contraposition.

6. Possible Worlds Semantics

The purpose of logic is to characterize the difference between valid and invalid arguments. A logical system for a language is a set of axioms and rules designed to prove *exactly* the valid arguments statable in the language. Creating such a logic may be a difficult task. The logican must make sure that the system is *sound*, i.e. that every argument proven using the rules and axioms is in fact valid. Furthermore, the system should be *complete*, meaning that every valid argument has a proof in the system. Demonstrating soundness and completeness of formal systems is a logician's central concern.

Such a demonstration cannot get underway until the concept of validity is defined rigorously. Formal semantics for a logic provides a definition of validity by characterizing the truth behavior of the sentences of the system. In propositional logic, validity can be defined using truth tables. A valid argument is simply one where every truth table row that makes its premises true also makes its conclusion true. However truth tables cannot be used to provide an account of validity in modal logics because there are no truth tables for expressions such as ‘it is necessary that’, ‘it is obligatory that’, and the like. (The problem is that the truth value of A does not determine the truth value for □A. For example, when A is ‘Dogs are dogs’, □A is true, but when A is ‘Dogs are pets’, □A is false. Nevertheless, semantics for modal logics can be defined by introducing possible worlds. We will illustrate possible worlds semantics for a logic of necessity containing the symbols ~, →, and □. Then we will explain how the same strategy may be adapted to other logics in the modal family.

In propositional logic, a valuation of the atomic sentences (or row of a truth table) assigns a truth value (T or F) to each propositional variable p. Then the truth values of the complex sentences is calculated with truth tables. In modal semantics, a set W of possible worlds is introduced. A valuation then gives a truth value to each propositional variable *for each of the possible worlds* in W. This means that value assigned to p for world w may differ from the value assigned to p for another world w′.

The truth value of the atomic sentence p at world w given by the valuation v may be written v(p, w). Given this notation, the truth values (T for true, F for false) of complex sentences of modal logic for a given valuation v (and member w of the set of worlds W) may be defined by the following truth clauses. (‘iff’ abbreviates ‘if and only if’.)

(~) v(~A, w)=T iff v(A, w)=F.

(→) v(A→B, w)=T iff v(A, w)=F or v(B, w)=T.

(5) v(□A, w)=T iff for every world w′ in W, v(A, w′)=T.

Clauses (~) and (→) simply describe the standard truth table behavior for negation and material implication respectively. According to (5), □A is true (at a world w) exactly when A is true in *all* possible worlds. Given the definition of ◇, (namely, ◇A = ~□~A) the truth condition (5) insures that ◇A is true just in case A is true in *some* possible world. Since the truth clauses for □ and ◇ involve the quantifiers ‘all’ and ‘some’ (respectively), the parallels in logical behavior between □ and ∀x, and between ◇ and ∃x noted in section 2 will be expected.

Clauses (~), (→), and (5) allow us to calculate the truth value of any sentence at any world on a given valuation. A definition of validity is now just around the corner. An argument is *5-valid for a given set W* (of possible worlds) if and only if every valuation of the atomic sentences that assigns the premises T at a world in W also assigns the conclusion T at the same world. An argument is said to be *5-valid* iff it is valid for every non empty set of W of possible worlds.

It has been shown that S5 is sound and complete for 5-validity (hence our use of the symbol ‘5’). The 5-valid arguments are exactly the arguments provable in S5. This result suggests that S5 is the correct way to formulate a logic of necessity.

However, S5 is not a reasonable logic for all members of the modal family. In deontic logic, temporal logic, and others, the analog of the truth condition (5) is clearly not appropriate; furthermore there are even conceptions of necessity where (5) should be rejected as well. The point is easiest to see in the case of temporal logic. Here, the members of W are moments of time, or worlds "frozen", as it were, at an instant. For simplicity let us consider a *future* temporal logic, a logic where □A reads: ‘it *will* always be the case that’. (We formulate the system using □ rather than the traditional G so that the connections with other modal logics will be easier to appreciate.) The correct clause for □ should say that □A is true at time w iff A is true at all times *in the future of w*. To restrict attention to the future, the relation R (for ‘eaRlier than’) needs to be introduced. Then the correct clause can be formulated as follows.

(K) v(□A, w)=T iff for every w′, if wRw′, then v(A, w′)=T.

This says that □A is true at w just in case A is true at all times *after* w.

Validity for this brand of temporal logic can now be defined. A *frame* <W, R> is a pair consisting of a non-empty set W (of worlds) and a binary relation R on W. A *model* <F, v> consists of a frame F, and a valuation v that assigns truth values to each atomic sentence at each world in W. Given a model, the values of all complex sentences can be determined using (~), (→), and (K). An argument is K-valid just in case any model whose valuation assigns the premises T at a world also assigns the conclusion T at the same world. As the reader may have guessed from our use of ‘K’, it has been shown that the simplest modal logic K is both sound and complete for K-validity.

7. Modal Axioms and Conditions on Frames

One might assume from this discussion that K is the correct logic when □ is read ‘it will always be the case that’. However, there are reasons for thinking that K is too weak. One obvious logical feature of the relation R (earlier than) is transitivity. If wRv (w is earlier than v) and vRu (v is earlier than u), then it follows that wRu (w is earlier than u). So let us define a new kind of validity that corresponds to this condition on R. Let a 4-model be any model whose frame <W, R> is such that R is a transitive relation on W. Then an argument is 4-valid iff any 4-model whose valuation assigns T to the premises at a world also assigns T to the conclusion at the same world. We use ‘4’ to describe such a transitive model because the logic which is adequate (both sound and complete) for 4-validity is K4, the logic which results from adding the axiom (4): □A→□□A to K.

Transitivity is not the only property which we might want to require of the frame <W, R> if R is to be read ‘earlier than’ and W is a set of moments. One condition (which is only mildly controversial) is that there is no last moment of time, i.e. that for every world w there is some world v such that wRv. This condition on frames is called *seriality*. Seriality corresponds to the axiom (D): □A→◇A, in the same

way that transitivity corresponds to (4). A D-model is a K-model with a serial frame. From the concept of a D-model the corresponding notion of D-validity can be defined just as we did in the case of 4-validity. As you probably guessed, the system that is adequate with respect to D-validity is KD, or K plus (D). Not only that, but the system KD4 (that is K plus (4) and (D)) is adequate with respect to D4-validity, where a D4-model is one where <W, R> is *both* serial and transitive.

Another property which we might want for the relation ‘earlier than’ is density, the condition which says that between any two times we can always find another. Density would be false if time were atomic, i.e. if there were intervals of time which could not be broken down into any smaller parts. Density corresponds to the axiom (C4): $\Box\Box A \rightarrow \Box A$, the converse of (4), so for example, the system KC4, which is K plus (C4) is adequate with respect to models where the frame <W, R> is dense, and KDC4, adequate with respect to models whose frames are serial and dense, and so on.

Each of the modal logic axioms we have discussed corresponds to a condition on frames in the same way. The relationship between conditions on frames and corresponding axioms is one of the central topics in the study of modal logics. Once an interpretation of the intensional operator □ has been decided on, the appropriate conditions on R can be determined to fix the corresponding notion of validity. This, in turn, allows us to select the right set of axioms for that logic.

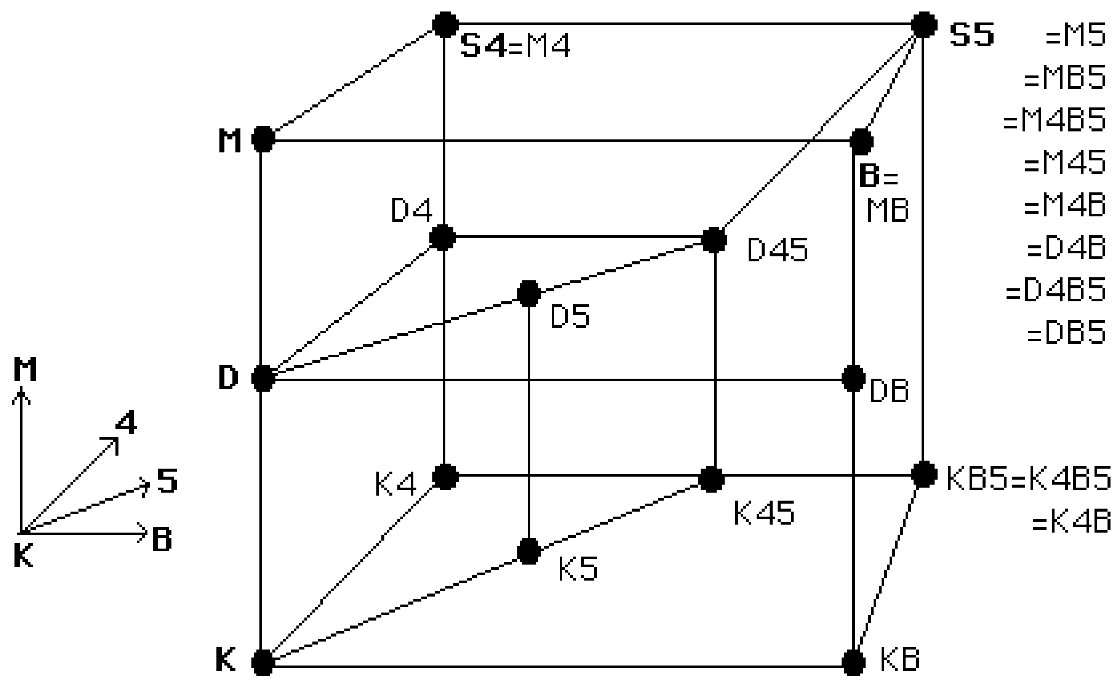
For example, consider a deontic logic, where □ is read ‘it is obligatory that’. Here the truth of □A does not demand the truth of A in *every* possible world, but only in a subset of those worlds where people do what they ought. So we will want to introduce a relation R for for this kind of logic as well, and use the truth clause (K) to evaluate □A at a world. However, in this case, R is not earlier than. Instead wRw' holds just in case world w' is a morally acceptable variant of w, i.e. a world that our actions can bring about which satisfies what is morally correct, or right, or just. Under such a reading, it should be clear that the relevant frames should obey seriality, the condition that requires that each possible world have a morally acceptable variant. The analysis of the properties desired for R makes it clear that a basic deontic logic can be formulated by adding the axiom (D) and to K.

Even in modal logic, one may wish to restrict the range of possible worlds which are relevant in determining whether □A is true at a given world. For example, I might say that it is necessary for me to pay my bills, even though I know full well that there is a possible world where I fail to pay them. In ordinary speech, the claim that A is necessary does not require the truth of A in *all* possible worlds, but rather only in a certain class of worlds which I have in mind (for example, worlds where I avoid penalties for failure to pay). In order to provide a generic treatment of necessity, we must say that □A is true in w iff A is true in all worlds *that are related* to w in the right way. So for an operator □ interpreted as necessity, we introduce a corresponding relation R on the set of possible worlds W, traditionally called the accessibility relation. The accessibility relation R holds between worlds w and w' iff w' is possible given the facts of w. Under this reading for R, it should be clear that frames for modal logic should be reflexive. It follows that modal logics should be founded on M, the system that results from adding (M) to K. Depending on exactly how the accessibility relation is understood, symmetry and transitivity may also be desired.

A list of some of the more commonly discussed conditions on frames and their corresponding axioms along with a map showing the relationship between the various modal logics can be found in the next section.

8. Map of the Relationships Between Modal Logics

The following diagram shows the relationships between the best known modal logics, namely logics that can be formed by adding a selection of the axioms (D), (M), (4), (B) and (5) to K. A list of these (and other) axioms along with their corresponding frame conditions can be found below the diagram.



In this chart, systems are given by the list of their axioms. So, for example M4B is the result of adding (M) (4) and (B) to K. In boldface, we have indicated traditional names of some systems. When system S appears below and/or to the left of S′ connected by a line, then S′ is an extension of S. This means that every argument provable in S is provable in S′, but S is weaker than S′, i.e. not all arguments provable in S′ are provable in S.

The following list indicates axioms, their names, and the corresponding conditions on the accessibility relation R, for axioms so far discussed in this encyclopedia entry.

Axiom Name	Axiom	Condition on Frames	R is...
(D)	$\Box A \rightarrow \Diamond A$	$\exists u\ wRu$	Serial
(M)	$\Box A \rightarrow A$	wRw	Reflexive

(4)	$\Box A \rightarrow \Box\Box A$	(wRv&vRu) $\Rightarrow$ wRu	Transitive
(B)	$A \rightarrow \Box\Diamond A$	wRv $\Rightarrow$ vRw	Symmetric
(5)	$\Diamond A \rightarrow \Box\Diamond A$	(wRv&wRu) $\Rightarrow$ vRu	Euclidean
(CD)	$\Diamond A \rightarrow \Box A$	(wRv&wRu) $\Rightarrow$ v=u	Unique
($\Box$M)	$\Box(\Box A \rightarrow A)$	wRv $\Rightarrow$ vRv	Shift Reflexive
(C4)	$\Box\Box A \rightarrow \Box A$	wRv $\Rightarrow$ ∃u(wRu&uRv)	Dense
(C)	$\Diamond\Box A \rightarrow \Box\Diamond A$	wRv&wRx $\Rightarrow$ ∃u(vRu&xRu)	Convergent

In the list of conditions on frames, the variables ‘w’, ‘v’, ‘u’, ‘x’ and the quantifier ‘∃u’ are understood to range over W. ‘&’ abbreviates ‘and’ and ‘ $\Rightarrow$ ’ abbreviates ‘if...then’.

9. The General Axiom

The correspondence between axioms and conditions on frames may seem something of a mystery. A beautiful result of Lemmon and Scott (1977) goes a long way towards explaining those relationships. Their theorem concerned axioms which have the following form:

(G) $\Diamond^h\Box^i A \rightarrow \Box^j\Diamond^k A$

We use the notation ‘$\Diamond^n$’ to represent n diamonds in a row, so, for example, ‘$\Diamond^3$’ abbreviates a string of three diamonds: ‘$\Diamond\Diamond\Diamond$’. Similarly ‘$\Box^n$’ represents a string of n boxes. When the values of h, i, j, and k are all 1, we have axiom (C):

(C) $\Diamond\Box A \rightarrow \Box\Diamond A \;=\; \Diamond^1\Box^1 A \rightarrow \Box^1\Diamond^1 A$

The axiom (B) results from setting h and k to 0, and letting j and k be 1:

(B) $A \rightarrow \Box\Diamond A \;=\; \Diamond^0\Box^0 A \rightarrow \Box^1\Diamond^1 A$

To obtain (4), we may set h and k to 0, set i to 1 and j to 2:

(4) $\Box A \rightarrow \Box\Box A \;=\; \Diamond^0\Box^1 A \rightarrow \Box^2\Diamond^0 A$

Many (but not all) axioms of modal logic can be obtained by setting the right values for the parameters in (G)

Our next task will be to give the condition on frames which corresponds to (G) for a given selection of values for h, i, j, and k. In order to do so, we will need a definition. The composition of two relations R and R′ is a new relation R∘R′ which is defined as follows:

$wR \circ R'v$ iff for some u, wRu and uR′v.

For example, if R is the relation of being a brother, and R′ is the relation of being a parent then R∘R′ is the relation of being an uncle, (because w is the uncle of v iff for some person u, both w is the brother of u and u is the parent of v). A relation may be composed with itself. For example, when R is the relation of being a parent, then R∘R is the relation of being a grandparent, and R∘R∘R is the relation of being a great-grandparent. It will be useful to write 'R^n', for the result of composing R with itself n times. So R^2 is R∘R, and R^4 is R∘R∘R∘R. We will let R^1 be R, and R^0 will be the identity relation, i.e. wR^0v iff w=v.

We may now state the Scott-Lemmon result. It is that the condition on frames which corresponds exactly to any axiom of the shape (G) is the following.

(hijk-Convergence) $wR^hv \ \& \ wR^ju \Rightarrow \exists x \ (vR^ix \ \& \ uR^kx)$

It is interesting to see how the familiar conditions on R result from setting the values for h, i, j, and k according to the values in the corresponding axiom. For example, consider (5). In this case i=0, and h=j=k=1. So the corresponding condition is

$wRv \ \& \ wRu \Rightarrow \exists x \ (vR^0x \ \& \ uRx).$

We have explained that R^0 is the identity relation. So if vR^0x then v=x. But ∃x (v=x & uRx), is equivalent to uRv, and so the Euclidean condition is obtained:

$(wRv \ \& \ wRu) \Rightarrow uRv.$

In the case of axiom (4), h=0, i=1, j=2 and k=0. So the corresponding condition on frames is

$(w=v \ \& \ wR^2u) \Rightarrow \exists x \ (vRx \ \& \ u=x).$

Resolving the identities this amounts to:

$vR^2u \Rightarrow vRu.$

By the definition of R^2, vR^2u iff ∃x(vRx & xRu), so this comes to:

$\exists x(vRx \ \& \ xRu) \Rightarrow vRu,$

which by predicate logic, is equivalent to transitivity.

$vRx \ \& \ xRu \Rightarrow vRu.$

The reader may find it a pleasant exercise to see how the corresponding conditions fall out of hijk-Convergence when the values of the parameters h, i, j, and k are set by other axioms.

The Scott-Lemmon results provides a quick method for establishing results about the relationship between axioms and their corresponding frame conditions. Since they showed the adequacy of any logic that extends K with a selection of axioms of the form (G) with respect to models that satisfy the corresponding set of frame conditions, they provided "wholesale" adequacy proofs for the majority of systems in the modal family. Sahlqvist (1975) has discovered important generalizations of the Scott-Lemmon result covering a much wider range of axiom types.

10. Provability Logics

Modal logic has been useful in clarifying our understanding of central results concerning provability in the foundations of mathematics (Boolos, 1993). Provability logics are systems where the propositional variables p, q, r, etc. range over formulas of some mathematical system, for example Peano's system PA for arithmetic. (The system chosen for mathematics might vary, but assume it is PA for this discussion.) Gödel showed that arithmetic has strong expressive powers. Using code numbers for arithmetic sentences, he was able to demonstrate a correspondence between sentences of mathematics and facts about which sentences are and are not provable in PA. For example, he showed there there is a sentence C that is true just in case no contradiction is provable in PA and there is a sentence G (the famous Gödel sentence) that is true just in case it is not provable in PA.

In provability logics, $\Box p$ is interpreted as a formula (of arithmetic) that expresses that what p denotes is provable in PA. Using this notation, sentences of provability logic express facts about provability. Suppose that $\bot$ is a constant of provability logic denoting a contradiction. Then $\sim\Box\bot$ says that PA is consistent and $\Box A \rightarrow A$ says that PA is sound in the sense that when it proves A, A is indeed true. Furthermore, the box may be iterated. So, for example, $\Box\sim\Box\bot$ makes the dubious claim that PA is able to prove its own consistency, and $\sim\Box\bot \rightarrow \sim\Box\sim\Box\bot$ asserts (correctly as Gödel proved) that if PA is consistent then PA is unable to prove its own consistency.

Although provability logics form a family of related systems, the system GL is by far the best known. It results from adding the following axiom to K:

(GL) $\Box(\Box A \rightarrow A) \rightarrow \Box A$

The axiom (4): $\Box A \rightarrow \Box\Box A$ is provable in GL, so GL is actually a strengthening of K4. However, axioms such as (M): $\Box A \rightarrow A$, and even the weaker (D): $\Box A \rightarrow \Diamond A$ are not available (nor desirable) in GL. In

provability logic, provability is not to be treated as a brand of necessity. PA may be defective (for all we can prove) so that if p is provable in PA (□p) it need not follow that ~p lacks a proof (~□~p = ◇p). PA might be inconsistent and so prove both p and ~p.

Axiom (GL) captures the content of Loeb's Theorem, an important result in the foundations of arithmetic. □A→A says that PA is sound for A, i.e. that if A were proven, A would be true. Such a claim might not be secure, since if PA goes awry, A might be provable and false. (GL) claims that if PA manages to prove the sentence that claims soundness for a given sentence A, then A is already provable in PA. Loeb's Theorem reports a kind of modesty on PA's part (Boolos, 1993, p. 55). PA never insists (proves) that a proof of A entails A's truth, unless it already has a proof of A to back up that claim.

It has been shown that GL is adequate for provability in the following sense. Let a sentence of GL be *always provable* exactly when the sentence of arithmetic it denotes is provable no matter how its variables are assigned values to sentences of PA. Then the provable sentences of GL are exactly the sentences that are always provable. This adequacy result has been extremely useful, since general questions concerning provability in PA can be transformed into easier questions about what can be demonstrated in GL.

GL can also be outfitted with a possible world semantics for which it is sound and complete. A corresponding condition on frames for GL-validity is that the frame be transitive, finite and irreflexive.

11. Quantifiers in Modal Logic

It would seem to be a simple matter to outfit a modal logic with the quantifiers ∀ (all) and ∃ (some). One would simply add the standard (or classical) rules for quantifiers to the principles of whichever propositional modal logic one chooses. However, systems of this kind create problems which have motivated some logicians to abandon classical quantifier rules in favor of the weaker rules of free logic (Garson, 1984). The controversy over whether classical principles should be adopted continues today.

The main points of disagreement can be traced back to decisions about how to handle the domain of quantification. The simplest alternative, the fixed-domain (sometimes called the possibilist) approach, assumes a single domain of quantification that contains all the possible objects. On the other hand, the world-relative (or actualist) interpretation, assumes that the domain of quantification changes from world to world, and contains only the objects that actually exist in a given world.

The fixed-domain approach requires no major adjustments to the classical machinery for the quantifiers. Modal logics that are adequate for fixed domain semantics can usually be axiomatized by adding principles of a propositional modal logic to classical quantifier rules together with the Barcan Formula (BF) (Barcan 1946). (For an account of some interesting exceptions see Cresswell (1995)).

(BF) ∀x□A→□∀xA.

The fixed-domain interpretation has advantages of simplicity and familiarity, but it does not provide a direct account of the semantics of certain quantifier expressions of natural language. We do not think that ‘Some man exists who signed the Declaration of Independence’ is true, at least not if we read ‘exists’ in the present tense. Nevertheless, this sentence was true in 1777, which shows that the domain for the natural language expression ‘some man exists who’ changes to reflect which men exist at different times. A related problem is that on the fixed-domain interpretation, the sentence $\forall y\Box\exists x(x=y)$ is valid. However, assuming that $\exists x(x=y)$ is read: y exists, $\forall y\Box\exists x(x=y)$ says that everything exists necessarily. However, it seems a fundamental feature of common ideas about modality that the existence of many things is contingent, and that different objects exist in different possible worlds.

The defender of the fixed-domain interpretation may respond to these objections by insisting that on his (her) reading of the quantifiers, the domain of quantification contains *all* possible objects, not just the objects that happen to exist at a given world. So the theorem $\forall y\Box\exists x(x=y)$ makes the innocuous claim that every *possible* object is necessarily found in the domain of all possible objects. Furthermore, those quantifier expressions of natural language whose domain is world (or time) dependent can be expressed using the fixed-domain quantifier $\exists x$ and a predicate letter E with the reading ‘actually exists’. For example, instead of translating ‘Some **M**an exists who **S**igned the Declaration of Independence’ by

$$\exists x(Mx\&Sx),$$

the defender of fixed domains may write:

$$\exists x(\mathbf{Ex\&}Mx\&Sx),$$

thus ensuring the translation is counted false at the present time. Cresswell (1991) makes the interesting observation that world-relative quantification has limited expressive power relative to fixed-domain quantification. World-relative quantification can be defined with fixed domain quantifiers and E, but there is no way to fully express fixed-domain quantifiers with world-relative ones. Although this argues in favor of the classical approach to quantified modal logic, the translation tactic also amounts to something of a concession in favor of free logic, for the world-relative quantifiers so defined obey exactly the free logic rules.

A problem with the translation strategy used by defenders of fixed domain quantification is that rendering the English into logic is less direct, since E must be added to all translations of all sentences whose quantifier expressions have domains that are context dependent. A more serious objection to fixed-domain quantification is that it strips the quantifier of a role which Quine recommended for it, namely to record robust ontological commitment. On this view, the domain of $\exists x$ must contain only entities that are ontologically respectable, and possible objects are too abstract to qualify. Actualists of this stripe will want to develop the logic of a quantifier $\exists x$ which reflects commitment to what is actual in a given world rather than to what is merely possible.

However, recent work on actualism tends to undermine this objection. For example, Linksy and Zalta

(1994) argue that the fixed-domain quantifier can be given an interpretation that is perfectly acceptable to actualists. Actualists who employ possible worlds semantics routinely quantify over possible worlds in their semantical theory of language. So it would seem that possible worlds are actual by these actualist's lights. By cleverly outfitting the domain with abstract entities no more objectionable than the ones actualists accept, Linsky and Zalta show that the Barcan Formula and classical principles can be vindicated. Note however, that actualists may respond that they need not be commited to the actuality of possible worlds so long as it is understood that quantifiers used in their theory of language lack strong ontological import. In any case, it is open to actualists (and non actualists as well) to investigate the logic of quantifiers with more robust domains, for example domains excluding possible worlds and other such abstract entities, and containing only the spatio-temporal particulars found in a given world. For quantifiers of this kind, a world-relative domains are appropriate.

Such considerations motivate interest in systems that acknowledge the context dependence of quantification by introducing world-relative domains. Here each possible world has its own domain of quantification (the set of objects that actually exist in that world), and the domains vary from one world to the next. When this decision is made, a difficulty arises for classical quantification theory. Notice that the sentence $\exists x(x=t)$ is a theorem of classical logic, and so $\Box\exists x(x=t)$ is a theorem of K by the Necessitation Rule. Let the term t stand for Saul Kripke. Then this theorem says that it is necessary that Saul Kripke exists, so that he is in the domain of every possible world. The whole motivation for the world-relative approach was to reflect the idea that objects in one world may fail to exist in another. If standard quantifier rulers are used, however, every term t must refer to something that exists in all the possible worlds. This seems incompatible with our ordinary practice of using terms to refer to things that only exist contingently.

One response to this difficulty is simply to eliminate terms. Kripke (1963) gives an example of a system that uses the world-relative interpretation and preserves the classical rules. However, the costs are severe. First, his language is artificially impoverished, and second, the rules for the propositional modal logic must be weakened.

Presuming that we would like a language that includes terms, and that classical rules are to be added to standard systems of propositional modal logic, a new problem arises. In such a system, it is possible to prove (CBF), the converse of the Barcan Formula.

(CBF) $\Box\forall xA\rightarrow\forall x\Box A.$

This fact has serious consequences for the system's semantics. It is not difficult to show that every world-relative model of (CBF) must meet condition (ND) (for ‘nested domains’).

(ND) If wRv then the domain of w is a subset of the domain of v.

However (ND) conflicts with the point of introducing world-relative domains. The whole idea was that existence of objects is contingent so that there are accessible possible worlds where one of the things in

our world fails to exist.

A straightforward solution to these problems is to abandon classical rules for the quantifiers and to adopt rules for free logic (FL) instead. The rules of FL are the same as the classical rules, except that inferences from $\forall$xRx (everything is real) to Rp (Pegasus is real) are blocked. This is done by introducing a predicate ‘E’ (for ‘actually exists’) and modifying the rule of universal instantiation. From $\forall$xRx one is allowed to obtain Rp only if one also has obtained Ep. Assuming that the universal quantifier $\forall$x is primitive, and the existential quantifier $\exists$x is defined by $\exists$xA =df ~$\forall$x~A, then FL may be constructed by adding the following two principles to the rules of propositional logic

> Universal Generalization. If $B \rightarrow A(y)$ is a theorem, so is $B \rightarrow \forall xA(x)$.

> Universal Instantiation. $(\forall xA(x)\ \&\ En) \rightarrow A(n)$

(Here it is assumed that A(x) is any well-formed formula of predicate logic, and that A(y) and A(n) result from replacing y and n properly for each occurrence of x in A(x).) Note that the principle of universal generalization is standard, but that the instantiation axiom is restricted by mention of En in the antecedent. In FL, proofs of formulas like $\exists x\Box(x=t)$, $\forall y\Box\ \exists x(x=y)$, (CBF), and (BF) which seem incompatible with the world-relative interpretation, are blocked.

One philosophical objection to FL is that E appears to be an existence predicate, and many would argue that existence is not a legitimate property like being green or weighing more than four pounds. So philosophers who reject the idea that existence is a predicate may object to FL. However in most (but not all) quantified modal logics that include identity (=) these worries may be skirted by defining E as follows.

> $Et\ =_{df}\ \exists x(x=t)$.

The most general way to formulate quantified modal logic is to create FS by adding the rules of FL to a given propositional modal logic S. In situations where classical quantification is desired, one may simply add Et as an axiom to FS, so that the classical principles become derivable rules. Adequacy results for such systems can be obtained for most choices of the modal logic S, but there are exceptions.

A final complication in the semantics for quantified modal logic is worth mentioning. It arises when non-rigid expressions such as ‘the inventor of bifocals’, are introduced to the language. A term is non-rigid when it picks out different objects in different possible worlds. The semantical value of such a term can be given by what Carnap (1947) called an individual concept, a function that picks out the denotation of the term for each possible world. One approach to dealing with non-rigid terms is to employ Russell's theory of descriptions. However, in a language that treats non rigid expressions as genuine terms, it turns out that neither the classical nor the free logic rules for the quantifiers are acceptable. (The problem can not be resolved by weakening the rule of substitution for identity.) A solution to this problem is to employ a more general treatment of the quantifiers, where the domain of quantification contains

individual concepts rather than objects. This more general interpretation provides a better match between the treatment of terms and the treatment of quantifiers and results in systems that are adequate for classical or free logic rules (depending on whether the fixed domains or world-relative domains are chosen).

Bibliography

An excellent bibliography of historical sources can be found in Hughes and Cresswell (1968).

- Anderson, A. and Belnap, N., *Entailment: The Logic of Relevance and Necessity*, Princeton: Princeton University Press vol. 1 (1975), vol. 2 (1992)
- Barcan, R., "A Functional Calculus of First Order Based on Strict Implication," *Journal of Symbolic Logic*, **11** (1946): 1-16
- Bencivenga, E., "Free Logics," in Gabbay, D., and Guenthner, F. (eds.) *Handbook of Philosophical Logic*, Dordrecht: D. Reidel (1986): **3.6**
- Bonevac, D., *Deduction*, Palo Alto, California: Mayfield Publishing Company (1987): Part II
- Boolos, G., *The Logic of Provability*, Cambridge, England: Cambridge University Press (1993)
- Bull, R. and Segerberg, Krister, "Basic Modal Logic," in Gabbay, D., and Guenthner, F. (eds.) *Handbook of Philosophical Logic*, Dordrecht: D. Reidel (1984): **2.1**
- Carnap, R., *Meaning and Necessity*, Chicago: U. Chicago Press, 1947
- Chellas, Brian, *Modal Logic: An Introduction*, Cambridge, England: Cambridge University Press (1980)
- Cresswell, M. J., "Incompleteness and the Barcan formula", *Journal of Philosophical Logic*, **24** (1995): 379-403.
- Cresswell, M. J., "In Defence of the Barcan Formula," *Logique et Analyse*, **135-136** (1991): 271-282
- Fitting, M. and Mendelsohn, R., *First Order Modal Logic*, Dordrecht: Kluwer, (1998)
- Gabbay, D., *Investigations in Modal and Tense Logics*, Dordrecht: D. Reidel (1976)
- Gabbay, D., *Temporal Logic: Mathematical Foundations and Computational Aspects*, New York: Oxford University Press (1994)
- Garson, James, "Quantification in Modal Logic," in Gabbay, D., and Guenthner, F. (eds.) *Handbook of Philosophical Logic*, Dordrecht: D. Reidel (1984): **2.5**
- Hintikka, J., *Knowledge and Belief: An Introduction to the Logic of the Two Notions*, Ithaca, N. Y.: Cornell University Press (1962)
- Hilpinen, R., *Deontic Logic: Introductory and Systematic Readings*, Dordrecht: D. Reidel (1971)
- Hughes, G. and Cresswell, M., *An Introduction to Modal Logic*, London: Methuen (1968)
- Hughes, G. and Cresswell, M., *A Companion to Modal Logic*, London: Methuen (1984)
- Hughes, G. and Cresswell, M., *A New Introduction to Modal Logic*, London: Routledge (1996)
- Kripke, Saul, "Semantical Considerations on Modal Logic," *Acta Philosophica Fennica*, **16**, (1963): 83-94
- Konyndik, K., *Introductory Modal Logic*, Notre Dame: University of Notre Dame Press (1986)
- Kvart, I., *A Theory of Counterfactuals*, Indianapolis: Hackett Publishing Company (1986)

- Lemmon, E. and Scott, D., *An Introduction to Modal Logic*, Oxford: Blackwell (1977)
- Lewis, C.I. and Langford, C. H., *Symbolic Logic*, New York: Dover Publications, 1959 (1932)
- Lewis, D., *Counterfactuals*, Cambridge, Massachusetts: Harvard University Press (1973)
- Linsky, B. and Zalta, E., "In Defense of the Simplest Quantified Modal Logic," *Philosophical Perspectives*, **8**, (Logic and Language), (1994): 431-458
- Prior, A. N., *Time and Modality*, Oxford: Clarendon Press (1957)
- Prior, A. N., *Past, Present and Future*, Oxford: Clarendon Press (1967)
- Quine, W. V. O., "Reference and Modality", in *From a Logical Point of View*, Cambridge, MA: Harvard University Press, 1953: 139-159
- Rescher, N, and Urquhart, A., *Temporal Logic*, New York: Springer Verlag (1971)
- Sahlqvist, H., "Completeness and Correspondence in First and Second Order Semantics for Modal Logic," in Kanger, S. (ed.) *Proceedings of the Third Scandanavian Logic Symposium*, Amsterdam: North Holland (1975): 110-143
- Van Benthem, J. F., *The Logic of Time*, Dordrecht: D. Reidel (1982)
- Zeman, J., *Modal Logic*, *The Lewis-Modal Systems*, Oxford: Oxford University Press (1973)

Other Internet Resources

- Advances in Modal Logic
- John Halleck's Logic System Interrelationships Home Page
- John McCarthy's Modal Logic Page

Related Entries

actualism | logic: classical | logic: deontic | logic: free | logic: intensional | logic: relevance | logic: temporal | possible worlds

James W. Garson
JGarson@uh.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 29, 2000
Content last modified: December 13, 2001

Temporal Logic

The term Temporal Logic has been broadly used to cover all approaches to the representation of temporal information within a logical framework, and also more narrowly to refer specifically to the modal-logic type of approach introduced around 1960 by Arthur Prior under the name of Tense Logic and subsequently developed further by logicians and computer scientists.

Applications of Temporal Logic include its use as a formalism for clarifying philosophical issues about time, as a framework within which to define the semantics of temporal expressions in natural language, as a language for encoding temporal knowledge in artificial intelligence, and as a tool for handling the temporal aspects of the execution of computer programs.

Modal-logic approaches to temporal logic

Tense Logic

Tense Logic was introduced by Arthur Prior (1957, 1967, 1969) as a result of an interest in the relationship between tense and modality attributed to the Megarian philosopher Diodorus Cronus (ca. 340-280 B.C.). For the the historical context leading up to the introduction of Tense Logic, as well as its subsequent developments, see Øhrstrøm and Hasle, 1995.

The logical language of Tense Logic contains, in addition to the usual truth-functional operators, four modal operators with intended meanings as follows:

P	"It has at some time been the case that ..."
F	"It will at some time be the case that ..."
H	"It has always been the case that ..."
G	"It will always be the case that ..."

P and F are known as the *weak tense operators*, while H and G are known as the *strong tense operators*. The two pairs are generally regarded as interdefinable by way of the equivalences

$$Pp \equiv \neg H \neg p$$
$$Fp \equiv \neg G \neg p$$

On the basis of these intended meanings, Prior used the operators to build formulae expressing various philosophical theses about time, which might be taken as axioms of a formal system if so desired. Some examples of such formulae, with Prior's own glosses (from Prior 1967), are:

$Gp \rightarrow Fp$	"What will always be, will be"
$G(p \rightarrow q) \rightarrow (Gp \rightarrow Gq)$	"If p will always imply q, then if p will always be the case, so will q"
$Fp \rightarrow FFp$	"If it will be the case that p, it will be --- in between --- that it will be"
$\neg Fp \rightarrow F \neg Fp$	"If it will never be that p then it will be that it will never be that p"

Prior (1967) reports on the extensive early work on various systems of Tense Logic obtained by postulating different combination of axioms, and in particular he considered in some detail what light a logical treatment of time can throw on classic problems concerning time, necessity and existence; for example, "deterministic" arguments that have been advanced over the ages to the effect that "what will be, will necessarily be", corresponding to the modal tense-logical formula $Fp \rightarrow \Box Fp$.

Of particular significance is the system of *Minimal Tense Logic* K_t, which is generated by the four axioms

$p \rightarrow HFp$	"What is, has always been going to be"
$p \rightarrow GPp$	"What is, will always have been"
$H(p \rightarrow q) \rightarrow (Hp \rightarrow Hq)$	"Whatever always follows from what always has been, always has been"

$G(p \rightarrow q) \rightarrow (Gp \rightarrow Gq)$	"Whatever always follows from what always will be, always will be"

together with the two rules of temporal inference:

RH: From a proof of p, derive a proof of Hp
RG: From a proof of p, derive a proof of Gp

and, of course, all the rules of ordinary Propositional Logic. The theorems of K_t express, essentially, those properties of the tense operators which do not depend on any specific assumptions about the temporal order. This characterisation is made more precise below.

Tense Logic is obtained by adding the tense operators to an existing logic; above this was tacitly assumed to be the classical Propositional Calculus. Other tense-logical systems are obtained by taking different logical bases. Of obvious interest is tensed predicate logic, where the tense operators are added to classical First-order Predicate Calculus. This enables us to express important distinctions concerning the logic of time and existence. For example, the statement *A philosopher will be a king* can be interpreted in several different ways, such as

∃x(Philosopher(x) & F King(x))	Someone who is now a philosopher will be a king at some future time
∃xF(Philosopher(x) & King(x))	There now exists someone who will at some future time be both a philosopher and a king
F∃x(Philosopher(x) & F King(x))	There will exist someone who is a philosopher and later will be a king
F∃x(Philosopher(x) & King(x))	There will exist someone who is at the same time both a philosopher and a king

The interpretation of such formulae is not unproblematic, however. The problem concerns the domain of quantification. For the second two formulae above to bear the interpretations given to them, it is necessary that the domain of quantification is always relative to a time: thus in the semantics it will be necessary to introduce a domain of quantification D(t) for each time t. But this can lead to problems if we want to establish relations between objects existing at different times, as for example in the statement "One of my friends is descended from a follower of William the Conqueror".

These problems are related to the so-called *Barcan formulae* of modal logic, a temporal analogue of which is

$F\exists x p(x) \rightarrow \exists x F p(x)$ ("If there will be something that is p, then there is now something that will be p")

For this formula to be true, we require the "domain cumulation" principle, which says that the whole domain of quantification D(t) at time t is included in all the domains D(t′) for times t′ later than t. For more on this and related matters, see van Benthem, 1995, Section 7.

Extensions to Tense Logic

Soon after its introduction, the basic "PFGH" syntax of Tense Logic was extended in various ways, and such extensions have continued to this day. Some important examples are the following:

The binary temporal operators S and U ("since" and "until"). These were introduced by Kamp (1968). The intended meanings are

Spq	"q has been true since a time when p was true"
Upq	"q will be true until a time when p is true"

It is possible to define the one-place tense operators in terms of S and U as follows:

$$Pp \equiv Sp(p \vee \neg p)$$
$$Fp \equiv Up(p \vee \neg p)$$

The importance of the S and U operators is that they are expressively complete with respect to first-order temporal properties on continuous, strictly linear temporal orders (which is not true for the one-place operators on their own).

Metric tense logic. Prior introduced the notation Fnp to mean "It will be the case the interval n hence that p". We do not need a separate notation Pnp since we can write F(-n)p for "It was the case the interval n ago that P". The case n=0 gives us the present tense. We can define the general, non-metric operators by

$$Pp \equiv \exists n(n<0 \; \& \; Fnp)$$
$$Fp \equiv \exists n(n>0 \; \& \; Fnp)$$
$$Hp \equiv \forall n(n<0 \rightarrow Fnp)$$
$$Gp \equiv \forall n(n>0 \rightarrow Fnp)$$

The "next time" operator O. This operator assumes that the time series consists of a discrete sequence of atomic times. The formula Op is then intended to mean that p is true at the immediately succeeding time step. Given that time is discrete, it can be defined in terms of the "until" operator U by

$$Op \equiv Up(p \& \neg p)$$

which says that p will be true at some future time, between which and the present time nothing is true. This can only mean the time immediately following the present in a discrete temporal order.

In discrete time, the future-tense operator F is related to the next-time operator by the equivalence

$$Fp \equiv Op \vee OFp.$$

Indeed, F can here be *defined* as the least fixed point of the transformation which maps an arbitrary propositional operator X onto the operator $\lambda p.Op \vee OXp$.

One could similarly define a past-time version of O; but since the main usefulness of this particular operator has been in relation to the logic of computer programming, where one is mainly interested in execution sequences of programs extending into the future, this has not so often been done.

Semantics of Tense Logic

The standard model-theoretic semantics of Tense Logic is closely modelled on that of Modal Logic. A *temporal frame* consists of a set T of entities called times together with an ordering relation < on T. This defines the "flow of time" over which the meanings of the tense operators are to be defined. An interpretation of the tense-logical language assigns a truth value to each atomic formula at each time in the temporal frame. Given such an interpretation, the meanings of the weak tense operators can be defined using the rules

> Pp is true at t if and only if p is true at some time t' such that $t'<t$
> Fp is true at t if and only if p is true at some time t' such that $t<t'$

from which it follows that the meanings of the strong operators are given by

> Hp is true at t if and only if p is true at all times t' such that $t'<t$
> Gp is true at t if and only if p is true at all times t' such that $t<t'$

We can now provide a precise characterisation of system K_t of Minimal Tense Logic. The theorems of K_t are precisely those formulae which are true at all times under all interpretations over all temporal frames.

Many tense-logical axioms have been suggested as expressing this or that property of the flow of time, and the semantics gives us a precise way of defining this correspondence between tense-logical formulae and properties of temporal frames. A formula p is said to *characterise* a set of frames F if

- p is true at all times under all interpretations over any frame in F.
- For any frame not in F, there is an interpretation which makes p false at some time.

Thus any theorem of K_t characterises the class of all frames.

A first-order formula in < determines a class of frames, namely those in which the formula is true. A tense-logical formula p corresponds to a first-order formula q just so long as p characterises the class of frames for which q is true. Some well-known examples of such pairs of formulae are:

$Hp \rightarrow Pp$	$\forall t \exists t'(t' < t)$	(unbounded in the past)
$Gp \rightarrow Fp$	$\forall t \exists t'(t < t')$	(unbounded in the future)
$Fp \rightarrow FFp$	$\forall t,t'(t < t' \rightarrow \exists t''(t < t'' < t'))$	(dense ordering)
$FFp \rightarrow Fp$	$\forall t,t'(\exists t''(t < t'' < t') \rightarrow t < t')$	(transitive ordering)
$FPp \rightarrow Pp \vee p \vee Fp$	$\forall t,t',t''((t < t'' \ \& \ t' < t'') \rightarrow (t < t' \vee t = t' \vee t' < t))$	(linear in the past)
$PFp \rightarrow Pp \vee p \vee Fp$	$\forall t,t',t''((t'' < t \ \& \ t'' < t') \rightarrow (t < t' \vee t = t' \vee t' < t))$	(linear in the future)

However, there are tense-logical formulae (such as $GFp \rightarrow FGp$) which do not correspond to any first-order temporal frame properties, and there are first-order temporal frame properties (such as *irreflexivity*, expressed by $\forall t \neg(t < t)$) which do not correspond to any tense-logical formula. For details, see van Benthem (1983).

Predicate-logic approaches to temporal logic

The method of temporal arguments

In this method, the temporal dimension is captured by augmenting each time-variable proposition or predicate with an extra argument-place, to be filled by an expression designating a time, as for example

Kill(Brutus,Caesar,44BC).

If we introduce into the first-order language a binary infix predicate < denoting the temporal ordering relation "earlier than", and a constant "now" denoting the present moment, then the tense operators can be readily simulated by means of the following correspondences, which not surprisingly bear more than a passing resemblance to the formal semantics for Tense Logic given above. Where p(t) represents the result of introducing an extra temporal argument place to the time-variable predicates occurring in p, we have:

Pp	$\exists t(t<now \,\&\, p(t))$
Fp	$\exists t(now<t \,\&\, p(t))$
Gp	$\forall t(t<now \rightarrow p(t))$
Hp	$\forall t(now<t \rightarrow p(t))$

Before the advent of Tense Logic, the method of temporal arguments was the natural choice of formalism for the logical expression of temporal information.

State and event-type reification

The method of temporal arguments encounters difficulties if it is desired to model *aspectual* distinctions between, for example, states, events and processes. Propositions reporting states (such as "Mary is asleep") have *homogeneous* temporal incidence, in that they must hold over any subintervals of an interval over which they hold (e.g., if Mary is asleep from 1 o'clock to 6 o'clock then she is asleep from 1 o'clock to 2 o'clock, from 2 o'clock to 3 o'clock, and so on). By contrast, propositions reporting events (such as "John walks to the station") have inhomogeneous temporal incidence; more precisely, such a proposition is not true of *any* proper subinterval of an interval of which it is true (e.g., if John walks to the station over the interval from 1 o'clock to a quarter past one, then it is not the case that he walks to the station over the interval from 1 o'clock to five past one --- rather, over that interval he walks part of the way to the station).

The method of state and event-type reification was introduced to cater for distinctions of this kind. It is an approach that has been especially popular in Artificial Intelligence, where it is particularly associated with the name of James Allen, whose influential paper (Allen 1984) is often cited in this connection. In this approach, state and event types are denoted by terms in a first-order theory; their temporal incidence is expressed using relational predicates "Holds" and "Occurs", as for example,

Holds(Asleep(Mary),(1pm,6pm))
Occurs(Walk-to(John,Station),(1pm,1.15pm))

where terms of the form (t,t') denote time intervals in the obvious way.

The homogeneity of states and inhomogeneity of events is secured by axioms such as

$$\forall s,i,i'(\mathrm{Holds}(s,i) \,\&\, \mathrm{In}(i',i) \rightarrow \mathrm{Holds}(s,i'))$$
$$\forall e,i,i'(\mathrm{Occurs}(e,i) \,\&\, \mathrm{In}(i',i) \rightarrow \neg\mathrm{Occurs}(e,i'))$$

where "In" expresses the proper subinterval relation.

Event-token reification

The method of event-token reification was proposed by Donald Davidson (1967) as a solution to the so-called "variable polyadicity" problem. The problem is to give a formal account of the validity of such inferences as

> John saw Mary in London on Tuesday.
> Therefore, John saw Mary on Tuesday.

The key idea is that each event-forming predicate is endowed with an extra argument-place to be filled with a variable ranging over event-tokens, that is, particular dated occurrences. The inference above is then cast in logical form as

> ∃e(See(John,Mary,e) & Place(e,London) & Time(e,Tuesday)),
> Therefore, ∃e(See(John,Mary,e) & Time(e,Tuesday)).

In this form, the inference does not require any additional logical apparatus over and above standard first-order predicate logic; on that basis, the validity of the inference is considered to be explained. This approach has also been used in a computational context in the Event Calculus of Kowalski and Sergot (1986).

Philosophical issues

Prior's motivation for inventing Tense Logic was largely philosophical, his idea being that the precision and clarity afforded by a formal logical notation was indispensible for the careful formulation and resolution of philosophical issues concerning time. See the article on Arthur Prior for a discussion of some of these.

The rivalry between the modal and first-order approaches to formalising the logic of time reflects an important set of underlying philosophical issues related to the work of McTaggart. This work is especially well-known, in the context of temporal logic, for introducing the distinction between the "A-series" and the "B-series". By the "A-series" is meant, essentially, the characterisation of events as Past, Present, or Future. By contrast, the "B-series" involves their characterisation as relatively "Earlier" or "Later". A-series representations of time inescapably single out some particular moment as present; of course, at different times, different moments are present --- a circumstance which, followed to what appeared to be its logical conclusion, led McTaggart to assert that time itself was unreal (see Mellor, 1981). B-series representations have no place for a concept of the present, instead taking the form of a synoptic view of all time and the (timeless) interrelations between its parts.

There is a clear affinity between the A-series and the modal approach and between the B-series and the first-order approach. In the terminology of Massey (1969), adherents of the former approach are called "tensers" while adherents of the latter are called "detensers". This issue is related in turn to the question of how seriously to take the representation of space-time as a single four-dimensional entity in which the four dimensions are at least in some respects on a similar footing. In view of the Theory of Relativity, it can be argued that this issue is not so much a matter for Philosophy as for Physics.

Applications of Temporal Logic

Applications to natural language

Prior (1967) lists amongst the precursors of Tense Logic Hans Reichenbach's (1947) analysis of the tenses of English, according to which the function of each tense is to specify the temporal relationships amongst a set of three times related to the utterance, namely S, the speech time, R, the reference time, and E, the event time. In this way Reichenbach was neatly able to distinguish between the simple past "I saw John", for which R=E<S, and the present perfect "I have seen John", for which E<R=S, the former statement referring to a past time coincident with the event of my seeing John, the latter referring to the present time, relative to which my seeing John is past.

Prior notes that Reichenbach's analysis is inadequate to account for the full range of tense usage in natural language. Subsequently much work has been done to refine the analysis, not only of tenses but also other temporal expressions in language such as the temporal prepositions and connectives ("before", "after", "since", "during", "until"), using the many varieties of temporal logic. For some examples, see Dowty (1979), Galton (1984), Taylor (1985), Richards *et al.* (1989).

Applications in artificial intelligence

We have already mentioned the work of Allen (1984), which is concerned with finding a general framework adequate for all the temporal representations required by AI programs. The Event Calculus of Kowalski and Sergot (1986) is pursued more specifically within the framework of logic programming, but is otherwise similarly general in character. A useful survey of issues involving time and temporal reasoning in AI is Galton (1995).

Much of the work on temporal reasoning in AI has been closely tied up with the notorious *frame problem*, which arises from the necessity for any automated reasoner to know, or be able to deduce, not only those properties of the world which *do* change as the result of any event or action, but also those properties which do *not* change. In everyday life, we normally handle such facts fluently without consciously adverting to them: we take for granted without thinking about it, for example, that the colour of a car does not normally change when one changes gear. The frame problem is concerned with how to formalise the logic of actions and events in such a way that indefinitely many inferences of this kind are

made available without our having to encode them all explicitly. A seminal work in this area is McCarthy and Hayes (1969). A useful recent reference for the frame problem is Shanahan, 1997.

Applications in computer science

Following Pnueli (1977), the modal style of Temporal Logic has found extensive application in the area of Computer Science concerned with the specification and verification of programs, especially concurrent programs in which the computation is performed by two or more processors working in parallel. In order to ensure correct behaviour of such a program it is necessary to specify the way in which the actions of the various processors are interrelated. The relative timing of the actions must be carefully co-ordinated so as to ensure that integrity of the information shared amongst the processors is maintained. Amongst the key notions here is the distinction between "liveness" properties of the tense-logical form Fp, which ensure that desirable states will obtain in the course of the computation, and "safety" properties of the form Gp, which ensure that undesirable states will never obtain.

Further information may be found in Galton (1987), Goldblatt (1987), Bolc and Szalas (1995).

Bibliography

- Allen, J. F., 1984,"Towards a general theory of action and time", *Artificial Intelligence*, volume 23, pages 123-154.
- van Benthem, J., 1983, *The Logic of Time*, Dordrecht, Boston and London: Kluwer Academic Publishers, first edition (second edition, 1991).
- van Benthem, J., 1995, "Temporal Logic", in D. M. Gabbay, C. J. Hogger, and J. A. Robinson, *Handbook of Logic in Artificial Intelligence and Logic Programming*, Volume 4, Oxford: Clarendon Press, pages 241-350.
- L. Bolc and A. Szalas (eds.), 1995, *Time and Logic: A Computational Approach*, London: UCL Press.
- Davidson, D., 1967, "The Logical Form of Action Sentences", in N. Rescher (ed.), *The Logic of Decision and Action*, University of Pittsburgh Press, 1967, pages 81-95. Reprinted in D. Davidson, *Essays on Actions and Events*, Oxford: Clarendon Press, 1990, pages 105-122.
- Dowty, D., 1979, *Word Meaning and Montague Grammar*, Dordrecht: D. Reidel.
- Gabbay, D. M., Hodkinson, I., and Reynolds, M., 1994, *Temporal Logic: Mathematical Foundations and Computational Aspects*, Volume 1,. Oxford: Clarendon Press.
- Galton, A. P., 1984, *The Logic of Aspect*, Oxford: Clarendon Press.
- Galton, A. P., 1987, *Temporal Logics and their Applications*, London: Academic Press.
- Galton, A. P., 1995, "Time and Change for AI", in D. M. Gabbay, C. J. Hogger, and J. A. Robinson, *Handbook of Logic in Artificial Intelligence and Logic Programming*, Volume 4, Oxford: Clarendon Press, pages 175-240.
- Goldblatt, R., 1987, *Logics of Time and Computation*, Center for the Study of Language and Information, CSLI Lecture Notes 7.

- Kamp, J. A. W., 1968. *Tense Logic and the Theory of Linear Order*, Ph.D. thesis, University of California, Los Angeles.
- Kowalski, R. A. and Sergot, M. J., 1986, "A Logic-Based Calculus of Events", *New Generation Computing*, volume 4, pages 67-95.
- Massey, G., 1969, "Tense Logic! Why Bother?", *Noûs*, volume 3, pages 17-32.
- McCarthy, J. and Hayes, P. J., 1969, "Some Philosophical Problems from the Standpoint of Artificial Intelligence", in D. Michie and B. Meltzer (eds.), *Machine Intelligence 4*, Edinburgh University Press, pages 463-502.
- Mellor, D. H., 1981, *Real Time*, Cambridge: Cambridge University Press. (Chapter 6 reprinted with revisions as "The Unreality of Tense" in R. Le Poidevin and M. MacBeath (eds.), *The Philosophy of Time*, Oxford University Press, 1993.)
- Øhrstrøm, P. and Hasle, P., 1995, *Temporal Logic: From Ancient Ideas to Artificial Intelligence*, Dordrecht, Boston and London: Kluwer Academic Publishers.
- Pnueli, A., 1977, "The temporal logic of programs", *Proceedings of the 18th IEEE Symposium on Foundations of Computer Science*, pages 46-67.
- Prior, A. N., 1957, *Time and Modality*, Oxford: Clarendon Press.
- Prior, A. N., 1967, *Past, Present and Future*, Oxford: Clarendon Press.
- Prior, A. N., 1969, *Papers on Time and Tense*, Oxford: Clarendon Press.
- Reichenbach, H., 1947, *Elements of Symbolic Logic*, New York: Macmillan
- Rescher, N. and Urquhart, A., 1971, *Temporal Logic*, Springer-Verlag.
- Richards, B., Bethke, I., van der Does, J., and Oberlander, J., 1989, *Temporal Representation and Inference*, London: Academic Press.
- Shanahan, M., 1997, *Solving the Frame Problem*, Cambridge MA and London: The MIT Press.
- Taylor, B., 1985, *Modes of Occurrence*, Aristotelian Society Series, Volume 2, Oxford: Basil Blackwell.

Other Internet Resources

- Foundations of Temporal Logic --- The WWW-site for Prior-studies, by Per Hasle and Peter Øhrstrøm.

Related Entries

logic: modal | Prior, Arthur

Antony Galton
A.P.Galton@ex.ac.uk

Table of Contents

First published: November 29, 1999
Content last modified: November 29, 1999

Arthur Prior

Arthur Prior (1914-1969) undertook pioneering work in intensional logic at a time when modality and intensional concepts in general were under attack. He invented tense logic and was principal theoretician of the movement to apply modal syntax to the formalisation of a wide variety of phenomena. Prior and Carew Meredith devised a version of the possible worlds semantics several years before Kripke published his first paper on the topic. An iconoclast and a resourceful innovator, Prior inspired many to undertake work in intensional logic.

Work on Tense Logic

Prior's most significant achievement was undoubtedly the invention and development of tense logic. His earliest mention of a logic of time-distinctions is to be found in the penultimate chapter of his unpublished manuscript *The Craft of Formal Logic* (completed in 1951). Following von Wright in ‘Deontic Logic’ he remarks that there are other groups of modal predicates to be set alongside the ordinary or ‘alethic’ modes of necessity and possibility. Prior refers to these non-alethic modalities as ‘quasi-modals’. After noting that Peter of Spain classified adverbial distinctions of time as modes he says (p. 750):

> That there should be a modal logic of time-distinctions has been suggested in our own day by Professor Findlay.

Findlay's paper ‘Time: A Treatment of Some Puzzles’ had appeared in the *Australasian Journal of Psychology and Philosophy* in 1941. Prior became aware of it as a result of its appearance in Flew's 1951 collection *Essays on Logic and Language*, which arrived in New Zealand just as Prior was writing the

final chapters of *The Craft.* The suggestion Prior refers to is barely more than a passing comment: "[O]ur conventions with regard to tenses are so well worked out that we have practically the materials in them for a formal calculus", wrote Findlay. He continued in a footnote:

> The calculus of tenses should have been included in the modern development of modal logics. It includes such obvious propositions as that
>
> x present ↔ (x present) present;
>
> x future ↔ (x future) present ↔ (x present) future;
>
> also such comparatively recondite propositions as that
>
> (x).(x past) future; i.e., all events, past, present and future, *will* be past.

Prior's first paper on the logic of tenses, ‘Three-Valued Logic and Future Contingents’, appeared in the *Philosophical Quarterly* for 1953. In 1949 he had learned from Geach's review of Julius Weinberg's *Nicolaus of Autricourt: A Study in 14th Century Thought* that for the scholastics an expression like ‘Socrates is sitting down’ is complete, in the sense of being assertable as it is, and is true at certain times, false at others. Prior had been brought up on the view -- prevalent even today -- that such an expression is incomplete until a time-reference is supplied, and hence that one cannot speak of the truth-value of the expression as altering with the passage of time. This was a crucial discovery for Prior: the idea that propositions which are subject to tense-inflections are liable to be true at one time and false at another was to become central to his philosophy. In a summary of his mature views on temporal realism composed nearly two decades later he wrote:

> Certainly there are unchanging truths, but there are changing truths also, and it is a pity if logic ignores these, and leaves it ... to comparatively informal ‘dialecticians’ to study the more ‘dynamic’ aspects of reality. (Prior 1996a: 46.)

Geach's review sent Prior to the sources, and he found that the ‘Socrates is sitting down’ example is not only in the scholastics but in Aristotle. Moreover, he discovered that Aristotle speaks of some propositions about the future - namely those about such events as are not already predetermined - as being neither true nor false when they are uttered, on the ground that there is as yet no definite fact with which they can accord or conflict. Prior quotes Aristotle's argument for believing that there are such events (De Interpretatione, ch.9): if there were not "there would be no need to deliberate or take trouble, on the supposition that if we should adopt a certain course, a certain result would follow, while, if we did not, the result would not follow" (1953: 232-3). This appealed to Prior, once a Barthian Calvinist but now on the side of indeterminism and freewill. There can be no doubt that Prior's interest in tense logic was bound up with his belief in the existence of real freedom (see, for example, Prior 1996b).

Over thirty years earlier, inspired by these same passages of Aristotle, Lukasiewicz had devised a three-

valued calculus whose third value, $\frac{1}{2}$, attached to propositions referring to future contingencies (Lukasiewicz 1920). Prior's 1953 article ‘Three-Valued Logic and Future Contingents’ is an exposition and defence of Lukasiewicz's system (which he had read about in Lewis and Langford's *Symbolic Logic* and in Jordan's monograph *The Development of Mathematical Logic and of Logical Positivism in Poland Between the Two Wars*). Prior thought, at this stage, that the logic of tensed propositions could be three-valued and that of tenseless propositions two-valued (1967: 16).

> In sum, three-valued logic does seem to bring new precision to our handling of statements with tenses (as opposed to the fundamentally tenseless propositions of the common systems); and we may say that Lukasiewicz has, by means of it, done for Aristotle's chapter on ‘future contingents’ what he has done elsewhere for the Aristotelian theory of the syllogism. (Prior 1953: 325.)

However, the Lukasiewicz system is far from being the ‘calculus of tenses’ at which Findlay had hinted. The system contains explicit syntactic representations of the alethic modalities but there is no syntactic representation of tense. Moreover, the simple addition of another truth value is hardly the means by which to do justice to Prior's key insight that propositions subject to tense-inflection can *change* in truth value with the passage of time: ‘Socrates is sitting down’ can be true at one time, false at another, and adding a third value does not help represent this change. It was some months before Prior realised that the use of off-the-shelf modal syntax was all that was required for the representation of this dynamical feature of tensed propositions. It was simply a matter of taking seriously the idea that he had discussed in *The Craft*: tense is a species of modality.

In 1953 Prior read of the work of the Megarian logician Diodorus Chronos in Benson Mates' book *Stoic Logic*. Prior later wrote of Diodorus that he "seems to have been an ancient Greek W.V. Quine, who regarded the Aristotelian logic of possibility and necessity with some scepticism, but offered nevertheless some ‘harmless’ senses that might be attached to modal words" (1967: 16). Diodorus defined the possible as what is or will be true: according to Diodorus, what actually happens is all that *can* happen. Prior found this deterministic definition uncongenial and set himself the task of locating a fallacy in the argument that Diodorus used to support it, the so-called Master Argument:

> The aim of the Master Argument, as I conceive it, was to refute the Aristotelian view that while it is now beyond the power of men or gods to affect the past, there are alternative futures between which choice is possible. Against this, Diodorus held that the possible is simply what either is or will be true. (1962a: 138; see also 1967: 33.)

Consideration of the Master Argument brought together three of Prior's great interests: indeterminism, modal logic, and the logic of time. In the course of his reflections on the argument, Findlay's footnote pushed its way to the front of his mind. His wife Mary remembers "his waking me one night, coming and sitting on my bed, and reading a footnote from John Findlay's article on time, and saying he thought one could make a formalised tense logic." His first explorations of this calculus of tenses appeared in his article ‘Diodoran Modalities’, completed by early 1954 and published in 1955. (‘Diodoran’ is Prior's

term, ‘Diodorean’ is Mates's. In 1958 Prior switched to the latter.) In it he wrote:

> I here propose to do something a little different, namely to employ the ordinary propositional variables ‘p’, ‘q’, ‘r’ etc., for ‘propositions’ in the Diodoran sense [i.e. propositions which ‘may be true at one time and false at another’] and to use certain operators which take such propositions as arguments, and which form functions taking such propositions as values. I shall use ‘Fp’ for ‘It will be the case that p’. (1955b: 205.)

The axioms and rules of the calculus that Prior sets out in this article owe much to von Wright's 1951 axiomatisation of Lewis' modal system S4 (*An Essay in Modal Logic*, Appendix II). Von Wright obtains S4 by adding the following axioms and rules to the classical propositional calculus (Lp is defined as -M-P):

> W1. p → Mp
>
> W2. M(pvq) ↔ (Mp)v(Mq)
>
> W3. MMp → Mp
>
> RE. If X ↔ Y is a thesis, so is MX ↔ MY
>
> RL. If X is a thesis so is LX.

Substituting his new operator F for M throughout the axioms produced, in the case of W1, a clear falsehood and in the case of W2 and W3, formulae which he found ‘obvious enough’. The result of substituting F into RE also made good sense to him. To deal with RL he introduced a tense-prefix G by means of a definition paralleling von Wright's definition of L, viz. Gp=-F-p; he read G as ‘it will be always be the case that’. Thus his 1953/54 calculus of tenses was the system produced by adding the F-analogues of W2, W3 and RE, and the G-analogue of RL, to the classical propositional calculus. (Prior remarks that an operator P may be introduced for ‘It has been the case that p’ but states no axioms or rules concerning it.) He shows that if, following Diodorus, Mp is defined as pvFp, von Wright's three modal axioms and two rules are all derivable in his calculus.

In a matchless piece of philosophical reconstruction Prior expresses the conclusion of the Master Argument, that what neither is nor will be true is not possible, as (-p&-Fp) → -Mp and derives it in his calculus from Diodorus' premisses, Pp → -M-Pp and -Mq → (L(p → q) → -Mp), together with two ‘broad assumptions about time, likely to have been taken for granted both by Diodorus and by his main opponents’: p → HFp and (-p&-Fp) → P-Fp. So the Master Argument is indeed valid. The fallacy, Prior tells us, lies with the second ‘broad assumption’, (-p&-Fp) → P-Fp (which says: when anything neither is nor will be the case, it has been the case that it will not be the case). This, Prior tells us, is not true if p refers to a future contingency (and thus has the truth value $\frac{1}{2}$ or ‘indeterminate’). Where p is indeterminate both Fp and -Fp are indeterminate, so the consequent of the disputed formula, P-Fp, is

false. -p must also be indeterminate (for if the negation of p were determinate p could not be indeterminate). Thus the antecedent of the disputed formula, -p&-Fp, is indeterminate, since both its conjuncts are indeterminate. According to the Lukasiewicz truth-table, an indicative conditional with a false consequent and an indeterminate antecedent is not true but indeterminate.

$\rightarrow$	1	$\frac{1}{2}$	0
1	1	$\frac{1}{2}$	0
$\frac{1}{2}$	1	1	$\frac{1}{2}$
0	1	1	1

Thus Prior is able to "deny that propositions of the form [(-p & -Fp) $\rightarrow$ P-Fp] are in all cases true".

This reasoning of Prior's is persuasive only if one accepts the Lukasiewicz truth-table. Yet are there not indicative conditionals with false consequents and indeterminate antecedents that are *true*? For example, let Fp be ‘I will kill myself’, and let us suppose that Fp is indeterminate (since I have not yet decided). So PFp is false: there has as yet been no time at which Fp is true. Yet it seems excessively harsh always to withhold the value True from the indicative conditional ‘If Fp then PFp’ (harsh given, at any rate, that temporal succession is transitive and that atypical moments such as the first moment of time are not under consideration). The difficulty is that in the conditional form a statement that is actually indeterminate in value, Fp, is supposed *true*. If Fp is now true then the present moment lies in an epoch throughout which it is the case that p's future truth is determined. Unless the moment of utterance happens to be the very first moment of this epoch there are past moments at which Fp is true, so PFp is now true. So, given some suitable assumption concerning such epochs, which for the present purposes may be chosen ad libitum (for example, it will suffice to assume that such epochs are always open intervals in continuous linear time), it is the case that if Fp is true then PFp is true. The claim that the conditional ‘if Fp then PFp’ is indeterminate simply because its antecedent happens to be so (and its consequent false) cannot be sustained in the present situation, for what the above reasoning shows is that (under the assumption) the combination antecedent True, consequent False cannot arise.

At bottom what this indicates, of course, is that the two-valued equivalence between -pvq and p $\rightarrow$ q does not extend to the three-valued case. -pÃq may be indeterminate and yet, as in the present example, there be no principled reason for withholding the value True from p $\rightarrow$ q. Lukasiewicz's idea that propositions about the future may be assigned a third truth-value is not as straightforward as it may seem. Prior himself later came to see that "[t]he truth-functional technique seems simply out of place" where indeterminate propositions are concerned (1967: 135). It was the treatment of conjunction that particularly troubled Prior. According to the Lukasiewicz truth-table, p&q is indeterminate where p and q are both indeterminate, yet as Prior remarks, if q is the negation of an indeterminate p then p&q is not indeterminate even though both its constituents are. The proposition ‘There both will and won't be a sea battle tomorrow’ is, Prior says, "plain false".

The Master Argument for determinism continued to exercise Prior for the rest of his life, and some of the

most useful and mathematically most interesting parts of his work were inspired by his thoughts on it. To choose just one example, from computer science, the calculi that Prior developed in response to the idea that the Master Argument is defeated if time is conceived as branching into the future have become useful for describing and verifying the behaviour of concurrent and distributed processing systems.

The tense calculus that Prior worked out in ‘Diodoran Modalities’ was hardly complete. For one thing there were no axioms or rules for the past tense operator P, and for another the calculus' two axioms, FFp → Fp and F(pvq) ↔ (Fp)v(Fq), had been chosen in an almost haphazard manner, by transforming certain axioms of a calculus designed for a rather different purpose. Prior duly set about expanding his calculus. He worked fast, and in August 1954 he presented a system of far greater sophistication in his Presidential Address to the second New Zealand Congress of Philosophy, held in Wellington. (Prior himself had organised the first Congress in Christchurch the previous year.)

His first step was to add two additional axioms concerning futurity, Gp → Fp and Fp → FFp. The former is the tense-analogue of Lp → Mp, a thesis whose derivation involves von Wright's axiom W1 (which is false under a tense-logical interpretation). The second of the new axioms is the converse of one of the original axioms. Prior remarks that it is suggested by ‘common notions on the subject of time’. Next he made a simplification which drew on a recent proof by Sobocinski that von Wright's weakest system M is equivalent to Feys' system T: by taking G as undefined and defining F as -G-, he was able to replace the somewhat unwieldy axiom F(pvq) ↔ (Fp)v(Fq) by G(p → q) → (Gp → Gq) (Sobocinski 1953).

This calculus of ‘pure futurity’ can be transformed into a calculus of ‘pure pastness’ by replacing F by P and G by H (‘It has always been the case that’) throughout the axioms, rules and definition. (Charles Hamblin was later to describe this process and its reverse as producing the ‘mirror-image’ of what one starts with.) Like the alethic systems, both these ‘pure’ calculi are *monomodal*. That is, each contains only one undefined modal operator. Prior wanted a ‘full tense calculus in F and P’: a *bimodal* logic (i.e. a logic containing two undefined modal operators). To obtain the bimodal system it was not enough to simply bundle together the futurity axioms and rules with their mirror images, for the two tense operators would then remain independent of one another. Some *interactive* axioms, "laws which relate to the interaction of pastness and futurity", were also required. Prior chose p → GPp and p → HFp. He found the first of these in William of Ockham's *Tractatus de Praedestinatione* and the second was one of the formulae he himself had used in his reconstruction of the Master-Argument. Thus Prior constructed the first multimodal system.

Prior honoured Findlay by proving one of the propositions stated in his footnote as the first theorem of the full tense calculus: all events, past, present and future, will be past. In tense-logical notation this is (pvPpvFp) → FPp, a formula that Prior was later to refer to simply as Findlay's law. (Prior describes Findlay's own suggested symbolisation of the proposition, (x).(x past) future, as ‘unfortunate’, since "the formula suggests that *everything* will have been the case (even permanent falsehoods)" (1967: 9).)

In ‘Diodoran Modalities’ Prior was content to describe FFp → Fp as ‘obvious enough’. By the time of the Wellington Congress his thinking had moved forward considerably. In a strikingly original section of

a strikingly original paper he set out what he called the l-calculus (he was later to prefer the term ‘U-calculus’). In the l-calculus the propositions of the tense calculus are treated as predicates expressing properties of dates, the latter being represented by name-variables x, y, z. The concatenation ‘px’ is read as ‘p at x’. ‘l’ is a binary relation taking dates as arguments and is read ‘is later than’. Using an arbitrary date z to represent the time of utterance, Fp is equated with (∃x)(lxz & px) (‘p at some time later than z’), and Pp with (∃x)(lzx & px) (‘p at some time earlier than z’). Gp and Hp are equated with the universal quantifications (x)(lxz → px) and (x)(lzx → px), respectively. Prior's key idea was that by imposing various conditions on the relation l, analogues of the axioms of the tense calculus can be derived in the l-calculus. (It was this idea that he and Carew Meredith were to exploit two years later in devising the possible worlds semantics for modal logic.)

Prior discovered that FFp → Fp and its mirror-image require for their derivation the condition lxy → (lyz → lxz), asserting that the later-than relation is transitive. Fp → FFp and its mirror-image require the condition lxz → (Ey)(lxy&lyz), which asserts that ‘between any two dates there is another date’. Gp → Fp requires the condition that ‘there is a date later than any given date’, (Ex)lxz (mutatis mutandis for the mirror-image). Prior notes that no conditions are required for the derivation in the l-calculus of the two interactive axioms p → GPp and p → HFp nor for the axioms G(p → q) → (Gp → Gq) and H(p → q) → (Hp → Hq): in the case of these formulae the standard machinery of truth-functional and quantificational logic suffices for their derivation (this applies also to the two rules of the calculus). It was these latter formulae that Lemmon later took as axioms for his minimal tense logic K_t, a system making no assumptions concerning contingent properties of the later-than relation.

Prior closes this section with a warning against regarding this interpretation of the tense calculus within the l-calculus as "a metaphysical explanation of what we mean by *is*, *has been* and *will be*": the l-calculus is not "metaphysically fundamental". His reason is that F(Socrates is sitting down) means ‘It is *now* the case that it will be the case that Socrates is sitting down’, and there is no genuine way of representing the indexical ‘now’ in the l-calculus (the free variable z is ‘a complete sham’). He continues: "If there is to be any ‘interpretation’ of our calculi in the metaphysical sense, it will probably need to be the other way round; that is, the l-calculus should be exhibited as a logical construction out of the PF-calculus rather than *vice versa*." This idea of the primacy of the tense calculus over the l-calculus - or, as he was later to put it, of McTaggart's A-series over the B-series - was to become a central and distinctive tenet of his philosophy. Prior's thesis is perhaps even more radical in its application to modal logic: the language of possible worlds is to be interpreted in terms of a language with modal operators and not - as is popularly held - *vice versa*. These issues form the theme of his final book *Worlds, Times and Selves* (edited by Kit Fine).

The text of the Wellington address was not published until 1958 (in the journal *Franciscan Studies*, under the title ‘The Syntax of Time-Distinctions’). It was the 1956 John Locke lectures and the ensuing book *Time and Modality* (published in 1957) that brought Prior's discoveries in tense and modal logic before a wider audience. A number of logicians - notably Thomas, Geach, Lemmon, Meredith and Kripke - took an immediate interest in Priorean modal logic, in particular his Diodorean system and his system Q, a multivalued logic admitting the existence of contingent beings. Less immediate attention was paid to his tense logic. The bibliography of the subject in Prior's 1968 volume *Papers on Time and Tense*

reveals that up until 1965 the only publications in the field were either by Prior himself or were reviews of his work (chiefly of *Time and Modality*). Yet a momentum was slowly gathering.

At a colloquium on modal and many-valued logics held in Helsinki in 1962 Hintikka proposed a tense-logical construal of his possible worlds semantics, maintaining that ‘if we do not want to tie our logic to old-fashioned physics, we are undoubtedly wiser if we ... no longer require that the alternativeness relation (in this case it could perhaps be more appropriately termed "futurity relation") effect a linear ordering’ (1963: 76). (Prior had happily tied his 1954 l-relation to ‘old-fashioned physics’. He made it clear that he did not think much of the view of time embodied in twentieth century physics (1996b: 49-51).) A pupil of von Wright, Hintikka had been stimulated by the latter's proposals for the wide application of modal logic (see the next section) and had come to appreciate the possibility of applying modal notions to the study of the logic of time before he read of Prior's sophisticated work in *Time and Modality* (which he reviewed in 1958). Hintikka was perhaps the first to stress the importance of a semantical approach to the tenses. During the early 1960s Hintikka travelled regularly between Helsinki and California. His ideas on tense influenced a number of logicians working in California, in particular Dana Scott.

Also in 1962, Scott gave a lecture on tense logic in Amsterdam. Among his audience was Hans Kamp, then an undergraduate. Scott's work on tense logic was one aspect of his study of the semantics of natural language, which he pursued in close collaboration with Richard Montague. Scott was aware of Prior's work, and was also influenced in his understanding of tense by Reichenbach, who had been a powerful figure at UCLA until his death in 1953. (Prior himself was critical of Reichenbach's analysis of the tenses, and described it as having been ‘in some ways a hindrance rather than a help to the construction of a logic of tenses’ (Prior 1967: 13, Reichenbach 1948).) Scott's tense logic was rather different in style from Prior's. Scott established the completeness and decidability of various axiomatic tense logics. He also showed that the temporal predicate logic of the reals is non-axiomatisable. His work in tense logic is cited widely but remains unpublished. Prior learned of Scott's work in a letter from Lemmon dated January 1964. (Lemmon had left Oxford in 1963 for Claremont, near Los Angeles. Scott was then at Stanford.)

In 1965 Prior visited California for several months, as Flint Professor of Philosophy at UCLA. For the first time Prior found himself among a group of enthusiasts for tense logic. Shortly after the visit ended he was to write: "I suppose that California is the most logically mature place in the world, and now that the logic of tenses is pursued so widely and so vigorously there, its raw pioneering days can be considered over" (1967: vi). When Prior arrived at UCLA Nino Cocchiarella was just completing a Ph.D. thesis on quantified modal and tense logic under Montague's supervision (‘Tense and Modal Logic: A Study in the Topology of Temporal Reference’). Cocchiarella's interest in the philosophy of time had initially been aroused by Reichenbach's work on space and time but it was his acquaintance with Prior's *Time and Modality* that swept him into tense-logical research. (Only later did he learn of Scott's work.) Prior's visit coincided with Hans Kamp's arrival at UCLA as a graduate student. Kamp attended Prior's lectures on tense logic in his first semester and became deeply interested in the subject. These lectures led more or less directly to the topic of Kamp's Ph.D. thesis, written under Montague's supervision and entitled ‘On Tense Logic and the Theory of Order’ (1968). In Kamp's work the development of tense

logic achieved a new level of formal sophistication. Segerberg, too, had just arrived in California, to study under Scott at Stanford. (Segerberg had become interested in tense logic in Finland in 1964 at a series of summer seminars given by von Wright, who was independently pursuing a tense logic that had arisen from his study of the logic of action and which was later shown to be equivalent to a system Prior had discussed in *Time and Modality* (Prior 1957: 23-4; see von Wright 1965 and Segerberg 1967, 1989).) In December of 1965 Scott delivered his famous talk to the Hume Society at Stanford entitled ‘The Logic of Tenses’. A multilith of Scott's handwritten notes for this talk has been circulating ever since among tense logicians. Four days later Prior himself addressed the Society, again on tense logic. It was in this fecund atmosphere that Prior completed the manuscript of his book *Past, Present and Future*, which remains to this day one of the most important references in the field.

The years 1965-7 saw the publication of work in tense logic by &Aqvist, Bull, Clifford, Cocchiarella, Garson, Geach, Hamblin, Luce, Makinson, Rescher, Segerberg, von Wright - and, of course, Prior. In a little over a decade Prior's invention had become an internationally pursued branch of logic.

Prior always had a firm belief that his tense logic would one day find useful application in other disciplines (possibly in mathematical physics, he thought). When the outside demand for tense logic did come it was from computer science. An early and influential application was by Pnuelli, who employed tense logic in formal reasoning about the behaviour of concurrent programs (Pnuelli 1977). (A concurrent program is one that governs the behaviour of a number of interacting processors running in parallel.) Pnuelli is sometimes mistakenly credited with having originated tense logic but in fact he first learned of it from the classic volume ‘Temporal Logic’ by Rescher and Urquhart (1971) (Ohrstrom and Hasle 1995: 344). This volume is dedicated to Prior and is an elegant introduction to his work.

Prior would not have been completely surprised to learn how useful tense logic is proving to be in computer science. He himself took little interest in computing beyond including material on elementary boolean circuit theory in his undergraduate lectures, but a number of the logicians with whom he was in touch were more deeply involved (Dov Gabbay and Dana Scott, for instance). Through others Prior knew something of the potential. He wrote "There are practical gains to be had from this study too, for example in the representation of time-delay in computer circuits" (1996a: 46). In *Past, Present and Future* he remarked concerning logics of discrete time that their usefulness "does not depend on any serious metaphysical assumption that time *is* discrete; they are applicable in limited fields of discourse in which we are concerned only with what happens next in a sequence of discrete states, e.g. in the workings of a digital computer" (1967: 67). Other logics from the group that he and von Wright pioneered are also finding computational applications, for example epistemic logic in Artificial Intelligence and knowledge-base engineering, and the logic of action in programming theory.

It is noteworthy that two of the major forces in the genesis of these software technologies were a love of ancient and medieval logic and a concern to make conceptual room for freedom of the human will. Only people who know little of the history of ideas will find any incongruity here. Would that those who now control and administer the funding of university research were less unaware of the oblique ways in which idea gives rise to idea.

Work on Modal Logic

Prior's own interest in modal logic arose chiefly from his study of the ancients. His earliest written piece on modal logic, the penultimate chapter of his manuscript *The Craft of Formal Logic* (completed in 1951) is largely historical in nature, with discussions of Aristotle, Peter of Spain, John Wallis, the *Port Royal Logic*, Isaac Watts' *Logick*, Hume and Mill on natural necessity, de Morgan, Whately, Aldrich. One of his conclusions, significant for his later work, is that "[t]here is everything to be said ... for the ... view that we may not only use devices developed in the study of quantity to throw light on modality, but also *vice versa*" (p.747). One of the most distinctive features of his mature philosophy was the view that quantification over possible worlds and instants is to be interpreted in terms of modality and tense, which constitute primitive notions - a view which he held in tandem with the belief that the study of such quantifications could usefully illuminate the study of modality and tense (as in his own U-calculi, described below). The chapter contains brief comparisons of the symbolic notations of Lukasiewicz, Feys and Lewis, but Prior makes very little use of the symbols he describes. At one point he explains that symbolic notation can be used to good advantage to bring out the difference between ‘Everything has something which is F to it’ and ‘There is something which is F to everything’. Probably Prior was just beginning to appreciate the power of the new symbolism to express the subtle distinctions demanded by his subject matter.

Early in 1951 Prior read von Wright's article ‘Deontic Logic’, and the penultimate chapter of *The Craft* contains a cameo discussion of this topic. Von Wright's influence is also clear in the paper Prior read in August of that same year to the AAP Conference in Sydney, entitled ‘The Ethical Copula’. Here Prior discusses and defends the parallel von Wright drew in ‘Deontic Logic’ between moral words and modal words. No doubt Prior found in deontic logic a significant connection between his existing interest in ethics and his fast-developing interest in modal logic.

Prior's reading of von Wright reinforced in his mind an idea that he had come across in Peter of Spain, Isaac Watts and the *Port Royal Logic*, an idea that was to be of considerable importance for his own future work. What von Wright calls the ‘alethic’ modes - necessity, possibility, impossibility and contingency - are members of an extended group of concepts that includes the epistemic modes (‘it is known that’, ‘it is not known to be false that’, etc.), the doxastic modes (for example ‘it is believed that’), and the deontic modes (such as ‘it is permitted that’ and ‘it is obligatory that’). In *The Craft* Prior also lists Watts' ‘it is written that’ and ‘it is said that’, noting that "one could think of innumerable others" (p.749). Later von Wright was to draw attention to what may be called the agentive modes: ‘the agent brings it about that’, ‘the agent makes it true that’, and the like (von Wright 1963). Prior introduces the collective term ‘quasi-modals’ for the non-alethic modes (p.749) and remarks, accurately, that "there is a hint of a large field here" (p.752). He was later to refer to his own tense operators as quasi-modal operators (1968: 138). By the time he wrote *Formal Logic* he was advocating the study of "the general modal form ‘It is -- that p’; ... as a distinct propositional form", observing that "this field has not been much cultivated" (1955a: 218). Between them Prior and von Wright pioneered the now much investigated field of general intensional logic, as it may be called, in which the syntax, and latterly the semantics, developed for the study of the alethic modalities is used in the analysis of a wide range of

quasi-modal concepts. Von Wright's deontic logic and Prior's tense logic were the first major successes in this field. Another has been the logic of action or the logic of the agentive modes.

Prior was convinced that no satisfactory metalinguistic analysis can be given of sentences having the general modal form ‘It is -- that p’. In *Formal Logic* he writes "It is quite plain, for example, that I am not talking about the sentence ‘Socrates is dead’; when I say ‘I wish that Socrates were dead’; " (1955a: 219). In *Time and Modality* he reiterates the point, now in connection with the tenses: "‘Professor Carnap will be flying to the moon’; ... is quite obviously a statement about Professor Carnap, and quite obviously not a statement about the statement ‘Professor Carnap is flying to the moon’" (1957: 8). What, then, is the semantic value of an expression replacing p in a sentence of the general modal form ‘It is -- that p’? Certainly not a truth value, as is the case with the standard extensional propositional calculus, for substituting a different expression with the same truth value into the sentence of the form ‘It is -- that p’ may alter the truth value of the latter sentence. Prior's answer - and in a sense it amounts to a rejection of the question - is that modal functions take propositions as arguments, but propositions are logical constructions. All sentences containing the word ‘proposition’ - including such sentences as ‘A modal operator expresses a function from propositions to truth values’ - mean no more and no less than sentences which contain neither that word nor an equivalent. In essence Prior's view is that there are intensional contexts but no intensions. For the last six years of his life Prior worked on a book that was to give systematic expression to his views on propositions. The incomplete manuscript, which Prior had entitled *Objects of Thought*, was published posthumously in 1971.

Of the four technical papers that marked the explosive beginning of Prior's career as a formal logician in 1952, two concern modal logic. ‘Modality De Dicto and Modality De Re’ is a discussion of this distinction as it appears in Aristotle, Ockham and Peter of Spain together with a comparison of these earlier views with those of von Wright in *An Essay on Modal Logic*. ‘In What Sense is Modal Logic Many-Valued?’ proposes an interpretation of Lukasiewicz's four-valued matrices for modal logic. This paper marked the beginning of Prior's study of Lukasiewicz's work on modality. Thereafter he read Lukasiewicz widely - even material in Polish, of which he said "the symbols are so illuminating that the fact that the text is incomprehensible doesn't much matter". In the Preface to *Time and Modality* he wrote "[W]hile I differed radically from the late Professor Lukasiewicz on the subject of modal logic, my debt to him will be obvious on almost every page".

Prior's detailed contributions to the development of modal logic are legion. At least one aspect of his work has not received the recognition it deserves. Prior and his collaborator Carew Meredith invented crucial elements of the possible worlds semantics for propositional modal logic several years ahead of Kripke, including the all-important binary relation which opens the way to modelling systems of different strengths. (Meredith was a lecturer in mathematics at Trinity College, Dublin, whose interest in logic was stimulated by the arrival of Lukasiewicz in Dublin shortly after the war.)

The invention is foreshadowed in the penultimate chapter of *The Craft*.

> For the similarity in behaviour between signs of modality and signs of quantity, various

> explanations may be offered. It may be, for example, that signs of modality are just ordinary quantifiers operating upon a peculiar subject-matter, namely possible states of affairs ... It would not be quite accurate to describe theories of this sort as ‘reducing modality to quantity’. They do reduce modal *distinctions* to distinctions of quantity, but the variables to which the quantifiers are attached retain something modal in their signification - they signify ‘possibilities’, ‘chances’, ‘possible states of affairs’, ‘possible combinations of truth-values’, or the like. (pp. 736-7.)

As sources for this idea Prior cites John Wallis (a seventeenth century logician) and the account of logically necessary and logically impossible propositions given by Wittgenstein in the *Tractatus* (p.737). Interestingly, he mentions Carnap only in a footnote: "Professor Carnap has a similar definition of logical necessity in terms of what he calls ‘state-descriptions’" (ibid). Prior does not refer to, and presumably had not at that time read, Carnap's 1946 paper ‘Modalities and Quantification’, which attempted a semantics for quantified S5 in terms of state-descriptions. (A state-description is a class of sentences satisfying certain conditions. Each state-description represents a possible state of affairs.) Carnap too cites the *Tractatus* account of modal propositions as his inspiration (1946: 47). Prior goes on to defend his account of modality as quantification over possible states of affairs against various alternatives, for example the Andersonian account, according to which ‘Every table here is necessarily brown’ means ‘There is a property which every table here in fact possesses, and of which it is true that everything that possesses it is in fact brown’. (John Anderson, Professor at the University of Sydney, was a leading figure in the development of philosophy in Australasia.)

In 1956 Prior wrote up his and Meredith's formal work on what he later described (1962a: 140) as the ‘logic of world-accessibility’, in a paper entitled ‘Interpretations of Different Modal Logics in the "Property Calculus"’ (Meredith and Prior 1996). It carries the attribution ‘C.A.M., August 1956; recorded and expanded A.N.P.’. Prior circulated the paper in mimeograph form. He mentions it in *Past, Present and Future* (1967: 42-5) and in his 1962 articles ‘Possible Worlds’ and ‘Tense-Logic and the Continuity of Time’. This paper is one of the earliest to employ a binary relation between possible states of affairs in order to discriminate between S5 and weaker systems. (Carnap's 1946 paper concerned only S5 and contained no such relation.)

The property calculus is essentially a variation of Prior's 1954 l-calculus described above. In the l-calculus tense-modal propositions are treated as predicates expressing properties of dates, and quantification theory is supplemented with various special axioms for a binary relation ‘l’ taking dates as arguments. In the modal version of the calculus sentences of modal logic are treated as if they express properties of certain objects a, b, c, etc. Objects are related by a binary relation U. (Prior and Meredith supply no account of what a formula of form ‘Uab’ might express.) The following definitions of necessity L and possibility M are given (‘pa’ indicates that object a has the property expressed by the sentence p):

$$(Lp)a = (b)(Uab \rightarrow pb)$$

$(Mp)a = (\exists b)(Uab \,\&\, pb)$.

The calculus consists of ordinary quantification theory supplemented by these definitions, together with certain axioms governing the relation U, and the following clauses:

$(-p)a = -(pa)$

$(p \rightarrow q)a = (pa) \rightarrow (qa)$.

It is implied that a modal proposition A is to be called a theorem of the calculus if and only if Aa is provable for an arbitrarily chosen object a.

Axioms for U are selected from a list containing (amongst others):

1. Uaa (U is reflexive)

2. Uab → (Ubc → Uac) (U is transitive)

3. Uab → Uba (U is symmetrical).

(Axiom 2 is also present in the l-calculus.) Prior and Meredith establish that the distribution principle L(p → q) → (Lp → Lq) is a theorem in the absence of any special axioms for U; that Lp → p is a theorem if axiom 1 is imposed; that axiom 2 gives the S4 principle Lp → LLp; and that 2 together with 3 give the S5 principle MLp → Lp. (Their approach is proof-theoretic in its basic orientation and they offer no completeness results.) In 1962a and 1962b Prior extends the approach to systems between S4 and S5 and systems independent of S4 between T and S5.

As previously remarked, the idea that the variables of quantification of the calculus should range over possible states of affairs or possible worlds is present in *The Craft*. In 1960, following a suggestion by Geach, Prior began thinking of U as a relation of *accessibility* between worlds. Prior tells us that Geach cashed out the notion of 'reaching' one world from another in terms of 'some dimension-jumping vehicle dreamed up by science fiction' (1962b: 36; see also 1962a: 140). (Geach referred to the whole business as 'Trans World Airlines'.) With this interpretation of U to hand, the property calculus can be viewed as treating (Lp)a - or 'Necessarily-p in world a' - as short for 'p is true in all worlds accessible from a'. Lemmon, in a draft of material intended for his and Dana Scott's projected book 'Intensional Logic', mistakenly credits Geach with the idea that the binary relation 'may be intuitively thought of as a relation between possible worlds'. In a letter to Scott, written after Lemmon's death in 1966, Prior remarked: "What Geach contributed was not the interpretation of [U] as a relation between worlds (God knows when *that* started), but the interpretation of [U] as a relation of accessibility". When Prior says "God knows when *that* started" he is presumably referring to the idea that the 'objects' of the calculus be regarded as possible worlds. Prior was right to think that the history of this idea is a tangled one. Priority is often assigned to Leibniz, but scholars have now traced the idea back to Duns Scotus and William of

Ockham (Knuuttila 1993).

It seems that the binary relation first made its appearance in a 1951 article by Jonsson and Tarski, ‘Boolean Algebras with Operators’. In their Theorem 3.14 they establish that every closure algebra is isomorphic to an algebraic system formed by a set and a reflexive and transitive relation between its elements; their Theorem 3.5 considers also a symmetry condition. In hindsight these theorems (which explicitly concern boolean algebras, of course) can be viewed as in effect a treatment of all the basic modal axioms and corresponding properties of the accessibility relation. Concerning this article Saul Kripke has remarked:

> Had they known they were doing modal logic, they would have had the completeness problem for many of the modal propositional systems wrapped up, and some powerful theorems. Mathematically they did this, but it was presented as algebra with no mention of semantics, modal logic, or possible worlds, let alone quantifiers. When I presented my paper at the conference in Finland in 1962, I emphasized the importance of this paper. Tarski was present, and said that he was unable to see any connection with what I was doing!

During the next eight years the binary relation was reinvented by a number of logicians. Prior, in his address to a conference in Wellington in 1954, seems to have been the first to use the binary relation in an explicitly tense-modal context. Other landmarks were an address by Montague to a conference held at UCLA in 1955, Prior and Meredith's property calculus of 1956, and Kanger (1957), Hintikka (1957, 1961) and Kripke (1959a, 1959b, 1963). Kripke was familiar with Kanger's work involving the binary relation at the time he obtained his own results. Kanger himself had read the 1951 paper by Jonsson and Tarski and he describes his results as similar to theirs (Kanger 1957: 39). My forthcoming paper ‘Possible Worlds: the Pre-Kripkean Era’ gives a fuller account of the history of the possible worlds semantics.

It was through reading Prior's paper ‘Modality and Quantification in S5’ in 1956 that Kripke first became interested in modal logic. He was at this time still a schoolboy, working on logic in almost complete isolation in Omaha, Nebraska. In 1958 he read *Time and Modality* and was impressed by the parallel Prior drew between tense and the alethic modalities. (At almost exactly the same time Prior was reading Kripke's first paper, ‘A Completeness Theorem in Modal Logic’, in his capacity as referee for *The Journal of Symbolic Logic*.) Kripke suspects that it was his reading of *Time and Modality* that first interested him in the problem of treating variable domains (a constant domain is assumed in ‘A Completeness Theorem in Modal Logic’). Kripke worked on Prior's idea that variable domains might lead to truth-value gaps even at the level of propositional logic, although he did not pursue this approach in his published material. (This idea was the motivation for Prior's system Q.) Kripke thinks it probable that it was Prior's work on many-valued matrices in *Time and Modality* which gave him the idea that a possible worlds model can be converted to a many-valued matrix (an idea he developed in his 1963 paper ‘Semantical Analysis of Modal Logic I: Normal Modal Propositional Calculi’).

Kripke wrote to Prior (September 3, 1958) pointing out an error in *Time and Modality*: contrary to Prior's claim, the Diodoran matrix (1957: 23) is not characteristic of S4. In his letter Kripke gave a characteristic matrix for S4 involving the idea that in a non-deterministic universe time branches into the future. He wrote:

> [I]n an indetermined system, we perhaps should not regard time as a linear series, as you have done. Given the present moment, there are several possibilities for what the next moment may be like - - and for each possible next moment, there are several possibilities for the moment after that. Thus the situation takes the form, not of a linear sequence, but of a ‘tree’.

This is essentially no more than a tensed interpretation of Kripke's relational semantics for S4, as he himself points out in the letter. Prior was excited by the letter and passed it on to Ivo Thomas and John Lemmon. Kripke wrote again on October 13, 1958. (The letters may still be seen in the Bodleian Library, Oxford. They are discussed in Ohrstrom and Hasle (1993).)

Prior's Life

In 1932, at the age of 17, Arthur Prior left his home town, sleepy Masterton in the North Island of New Zealand, and enrolled at the University of Otago. The son of a doctor, Prior's initial intention was to study medicine. He was soon beckoned away by philosophy, in which he gained a B.A. in 1935. It was John Findlay, then Professor of Philosophy at Otago, who introduced Prior to logic. A contemporary of Gilbert Ryle and William Kneale, Findlay himself had studied at Graz and at Oxford; his influential book *Meinong's Theory of Objects* was published during Prior's second year at Otago. Under Findlay's direction Prior cut his teeth on W.E. Johnson's classic text *Logic* and studied the 18th century British moralists. It was Findlay who first interested Prior in the history of logic. In 1949 Prior wrote of him ‘I owe to his teaching, directly or indirectly, almost all that I know of either Logic or Ethics’ (1949: xi) and he was later generously to describe Findlay as ‘the founding father of modern tense-logic’ (1967: 1).

Prior's M.A. thesis, in which he criticised subjectivist and formalist approaches to logic, was awarded only a second by the external examiner. Fortunately Findlay knew a budding logician when he saw one and secured Prior an assistant lectureship at Otago. During 1937 Prior gave courses on logic, ethics, and probability theory. In December of that year his first published paper - arguing that a nation is a logical construction out of individuals - appeared in the *Australasian Journal of Psychology and Philosophy*.

At this point Prior temporarily abandoned his academic career and spent three bohemian years wandering in Britain and Europe. He returned to New Zealand at the end of 1940, and on emerging from the air force in 1945 he applied for a vacant lectureship at Canterbury University College in Christchurch. By now he had a further three articles in the *Australasian Journal of Psychology and Philosophy* (‘Can Religion be Discussed?’, ‘The Meaning of Good’ and ‘The Subject of Ethics’) and with a strong recommendation from Findlay Prior got the job. He started work in February 1946. (The vacancy Prior filled was created by the departure from New Zealand of Karl Popper. Prior and Popper were never

colleagues. Apart from Prior's attendance at some of Popper's Workers' Educational Association lectures in 1943 there was no contact between the two men.)

At Canterbury Prior was thrown entirely on his own resources, being as he put it "the only philosopher about the place". He bore the responsibility for providing a broad and balanced philosophy curriculum, yet his own formal education in philosophy had stopped short nine years previously. Prior's one recourse in the face of isolation was to read, and read he did. In logic he began by returning to W.E. Johnson. Next came J.N. Keynes's *Studies and Exercises in Formal Logic* and then (in his own phrase) he got stuck into *Principia Mathematica*. He learned a lot about the history of the subject from Peirce, whom he found "unexpectedly magnificent". An important discovery, in 1950, was Bochenski's *Precis de Logique Mathematique*. Prior was fascinated by the ‘very neat symbolic notation’ due to Lukasiewicz, and before long he turned his back completely on the more usual Peano-Russell notation. Bochenski was later to describe Prior as even more of a ‘CCCC-logician’ than he was himself. (In Lukasiewicz's parenthesis-free notation Cpq is written for ‘If p then q’, Kpq for ‘Both p and q’, Apq for ‘Either p or q’, Epq for ‘If and only if p then q’, Np for ‘Not p’.) Lukasiewicz's own *Aristotle's Syllogistic* and Tarski's *Introduction to Logic* soon followed. By now the logic bug had well and truly bitten. Prior saw from the work of the Poles that formal precision is possible in philosophy and this delighted him. The upshot of Prior's reading for the curriculum was that his students learned Aristotelian and medieval logic, using Polish notation and with Bochenski's *Precis de Logique Mathematique* as a text. "Despite the language difficulty, I have found this a first-class textbook to accompany lectures to New Zealand students", he declared (1952c: 35).

An exuberant, playful man of seemingly inexhaustible vitality, Prior made an excellent teacher. He had no trace of pomposity or pretension. His students appreciated the friendly welcome they would receive at his home, not to mention his relaxed attitude toward the administrative paraphernalia of roll-taking and the like. In those days Canterbury University College was a formal, stuffy place and Prior was a breath of fresh air for his students. In a milieu where jacket and tie were the norm even in a sweltering New Zealand summer, Prior would lecture in baggy khaki shorts and roman sandles. Jim Wilson recalls the friendly informality of Prior's first-year classes:

> The strained precision of clock time was alien to him, so he was usually late for his own lectures (or anyone else's for that matter - he was very egalitarian about it). But he almost always turned up eventually, thinning hair blown vertical by his dash on his bike when he remembered the time. He would pull cycle clips off his trousers and plonk an ancient shopping bag on the desk in front of him. Out of this bag would come ... a cabbage, a bunch of carrots, a loaf of bread, a bottle of milk ... until, always at the bottom, he would find the book he was looking for. Back into the bag went the rest of the goodies, then he would look up at us, apologise for being late if he was more than usually so, and ask: "Now where were we last time?" Someone in the front row would consult her or his notes - Arthur couldn't as he never had any - and would say "You were just dealing with such and such." "Ah yes, thank you" Arthur would respond, and forthwith launch into an extempore exposition which followed on perfectly from the previous session and was beautifully structured and clear even though he was just thinking along with us. And of course we

> could stop him and ask for clarification or elaboration at any time, without in the slightest affecting the overall structure and direction of his thoughts.

Soon after his discovery of *Precis de Logique Mathematique* Prior wrote to Bochenski in Fribourg and then, a little later, to Lukasiewicz in Dublin. He was excited to receive replies. "We are, all of us, very isolated, being few and scattered", wrote Bochenski. "It is a real pleasure to hear that a Colleague so far away is interested in the same problems you are working at and that he finds one's little writings may be of some use." Thus began Prior's voluminous correspondence with logicians the world over. There were other ways, too, in which his isolation lessened. In 1951 he met and became friends with John Mackie and Jack Smart, at a conference in Sydney. This was Prior's first experience of being among a large gathering of philosophers and Mary Prior describes the conference as his ‘entry into a wider world’. In the same year George Hughes was appointed to the Victoria University of Wellington. Prior and Hughes had to make the most of their all-too-infrequent meetings, sometimes talking until the birds woke. Prior was fortunate in having a number of excellent students during these early years, among them Jonathan Bennett, Ronald Butler and (a little later) Robert Bull. For Prior they were oases in the desert. In 1952 he gained an assistant lecturer, Sandy Anderson. The following year philosophy became a department in its own right and Prior was made Professor.

1949 saw the publication of Prior's first book, a slim but potent volume entitled *Logic and the Basis of Ethics*. It was published by the Clarendon Press and soon became prominent in Oxford. Austin liked it and Ryle approved of "Prior's complete lack of mugwumpery". In the Introduction Prior explains that by the ‘logic of ethics’ he means "not a special kind of logic, nor a special branch of logic, but an application of it", and the book is a vigorous examination of the arguments of each side in the naturalism/anti-naturalism debate.

Logic and the Basis of Ethics contains no symbolism, and Prior's phrase ‘the logic of ethics’ is little more than a battle cry. The few technical concepts that are introduced all pertain to syllogistic logic. It was not until 1952 that Prior began publishing papers in symbolic logic - four of them, suddenly, in the same year. At the unusually late age of 38 Prior had become a formal logician. He wrote these papers while completing the manuscript of what was intended to be his second book, *The Craft of Formal Logic*. (The manuscript of *The Craft of Formal Logic* is deposited in the Bodleian Library, Oxford.) This began life in 1949 as a Dictionary of Formal Logic, but at the advice of the Clarendon Press Prior soon switched to a more orthodox format. His logical interests veered sharply while he was writing *The Craft*. To sixteen chapters on the logic of categoricals, hypotheticals, terms and relations are added, almost as an afterthought, one on modal logic and one on the axiomatic method. Prior finished the manuscript in December 1951 and sent it to the Clarendon Press; fourteen months later they wrote agreeing to publish the book if Prior would both shorten it and give greater emphasis to modern logic. He undertook to make the changes but ended up writing a completely different book. This was finally published in 1955 with the title *Formal Logic*; it ran into a second edition in 1962. Parts of *The Craft* not absorbed into the later work were published posthumously under the title *The Doctrine of Propositions and Terms*.

Steeped in Polish notation and the axiomatic method, *Formal Logic* typifies Prior's mature work. It teaches, enthusiastically yet without fuss, that there was life - fascinating life - before the here and now

of logic. What Prior once wrote admiringly of Lukasiewicz is no less true of Prior himself: ‘having done very distinguished work as a mathematical logician in the modern style, [he] is at the same time interested in the history of his subject ... and contrives both to use modern techniques to bring out more clearly what the ancients were driving at, and to learn from the ancients useful logical devices which the moderns have in general forgotten’ (1952c: 37).

After Findlay, Lukasiewicz was the greatest single influence on Prior's development as a logician. Prior's 1952 review article ‘Lukasiewicz's Symbolic Logic’ is one of the first papers in which he makes extensive use of symbolism. (He discusses Lukasiewicz's book *Aristotle's Syllogistic From the Standpoint of Modern Formal Logic* (published in 1951) and two articles, ‘The Shortest Axiom of the Implicational Calculus of Propositions’ and ‘On Variable Functions of Propositional Arguments’.) Prior seems to have first learned of Lukasiewicz's work through Bochenski's writings (Bochenski was a pupil of Lukasiewicz). Lukasiewicz had devised an axiomatic treatment of Aristotle's reduction of the imperfect syllogistic moods to those of the first figure, which Prior encountered in Bochenski's *Precis de Logique Mathematique* (published in 1949). This enchanted Prior. He was taking his students through the derivations as early as 1951, and he summarises Lukasiewicz's system in the final chapter of *The Craft.* Throughout this chapter he makes extensive use of Lukasiewicz's symbolic notation. It was Lukasiewicz's axiomatic treatment of traditional logic that fully brought home to Prior the power of modern symbolic methods. Moreover it was probably his reading of Lukasiewicz that made clear to him the fundamental importance of propositional logic. "It seems that Aristotle did not suspect the existence of another system of logic besides his theory of the syllogism", Lukasiewicz had written, "[y]et he uses intuitively the laws of propositional logic ..." (1951: 49). (Lukasiewicz's axiomatisation of the syllogistic incorporates his own three-axiom formalisation of propositional logic (1951: 80).) In his review Prior quotes approvingly Lukasiewicz's assertion that "the logic of the Stoics, the inventors of the ancient form of the propositional calculus, was much more important than all the syllogisms of Aristotle" (1951: 131). In *The Craft* propositional logic is barely mentioned until the final chapter, whereas *Formal Logic* begins with a thorough introduction to the subject. On page 3 of *Formal Logic* Prior states that the logic of propositions is "basic, and the rest of logic built upon it". Prior's interest in economical bases for propositional and pure implicational logics, initially aroused by his study of Peirce, was stimulated by Lukasiewicz's article ‘The Shortest Axiom of the Implicational Calculus of Propositions’ and the opening chapters of *Formal Logic* draw heavily on Lukasiewicz's work in this area.

In 1954 Gilbert Ryle visited New Zealand. He brought Prior an invitation to visit Oxford and deliver the John Locke lectures. Prior arranged a twelve month leave of absence from Canterbury and arrived in Oxford at the beginning of 1956. Rather quickly a small group began to form around him: Ivo Thomas, John Lemmon, Peter Geach. (These meetings with Prior were Lemmon's first introduction to modal logic.) Hughes summarises the news of him that was arriving back home: "this wild colonial boy just hit Oxford and started to gather around him the main people [interested in] logic, and he started to organise a lot of parties, almost, for the serious doing of logic." Prior kitted out his tiny rented flat with a toyshop blackboard and held open house. On Mondays during Hilary and Trinity terms Prior lectured on modal logic, his great passion, and on tense logic, his great invention. The lectures were published the following year, under the title *Time and Modality*.

In the summer break following the John Locke lectures Prior organised a Logic Colloquium in Oxford. In Britain in the 1950s logic was deeply out of fashion and its practitioners were isolated and somewhat demoralised. As Prior wrote shortly after the Colloquium "There *are* logicians in England and Ireland; but it must be admitted that they are somewhat scattered, and so far as I could gather they had never had any general get-together" (1956b: 186). Prior's Colloquium brought together Lemmon, Thomas, Geach, M. Kneale, W.C. Kneale, Lewy, Smiley, Bennett, Lejewski, M.W. Dick, Faris, Nidditch, Carew Meredith, David Meredith, and others. It was all a huge success, and the Colloquium became a regular fixture. Through his John Locke lectures, the Colloquium, and his numerous visits around the country, Prior helped to revitalise British logic. The group he left behind saw similarities between themselves and the close-knit group of researchers that existed in Warsaw before 1939.

Prior's heart may have been heavy as he journeyed back to New Zealand. After twelve months of logical companionship on a grand scale, life at Canterbury must have seemed a bleak prospect. He was seething with enthusiasm for logic and threw himself once more into a massive correspondence, but it could no longer satisfy him. Prior pined. When the offer arrived of a newly established second chair at the University of Manchester he snatched it up. Prior left New Zealand in December 1958.

He was at Manchester for seven years. In 1966 Anthony Kenny recommended him for a fellowship at Balliol. The move would mean a drop in both status and salary, not to mention an increase in teaching, but Prior did not hesitate. His sabbatical in Oxford had been one of his happiest years. ‘This is the good life’, he told George Hughes once he was settled in at Balliol. He felt he simply belonged. Prior soon built up a reputation for being one of the best teachers in Oxford - though his students were sometimes surprised to be given eighteenth century moralists to read instead of books by the currently fashionable.

Just before his departure from Manchester Prior told Tom Richards, a visiting New Zealander, that he was going to Oxford with a mission. Prior's own work was an exemplary fusion of philosophy and logic, and he went to Oxford with the intention of interesting the mathematical logicians in philosophy and the philosophers in mathematical logic. The time was right; and Prior spared no energy in preaching his message:

> [F]ormal logic and general philosophy have more to bring to one another than is sometimes supposed. I do not mean by saying this to underrate the work of those who have explored the properties of symbolic calculi without any concern as to what they might be used to mean ... Nor do I mean to underrate what recent philosophers have done in the way of exploring the obstinate and intricate ‘logic’ embedded in common discourse, even when they have not derived or sought to derive anything like a calculus from it ... But these activities are, or can be, related to one another very much as theory and observation are in the physical sciences; and I must confess to a hankering after well-constructed theories which much contemporary philosophy fails to satisfy. (1957: vii.)

Prior did not live to enjoy the *entente cordiale* between philosophy and logic that he helped usher in. His health began to let him down during his second year at Balliol. He was found to have both angina

pectoris and polymyalgic rheumatism. During the autumn of 1969 the rheumatism grew steadily worse. He was at this time on sabbatical at the University of Oslo. The pain left him with no zest for work. He dutifully gave his weekly seminars and spent the remainder of his time brooding savagely on how painful it was to do such elementary things as put on a coat. His hosts made him an appointment with a rheumatologist, who prescribed cortisone. In a letter written a few days later, and a few days before his heart failed, Prior described himself as one of the miracles of modern medicine. "I've been sleeping well ... running up and down stairs ... I can stand on one leg and put a sock on the other (first time for months) ... they've got me cured now and I'm fine."

Bibliography

- Bochenski, I.M. 1948. *Precis de Logique Mathematique*. Bussum, Pays-Bas: Kroonder.
- Carnap, R. 1946. 'Modalities and Quantification'. *The Journal of Symbolic Logic*, vol. 11, pp.33-64.
- Copeland, B.J. (ed.) 1996. *Logic and Reality: Essays on the Legacy of Arthur Prior*. Oxford: Clarendon Press.
- Cresswell, M., Crossley, J.N. (eds), 1989. 'Prior Postscript'. Unpublished. (An edited transcript of a panel discussion concerning Prior held at the 1981 Annual Conference of the Australasian Association for Logic.)
- Findlay, J.N. 1933. *Meinong's Theory of Objects*. Oxford: Clarendon Press.
- Findlay, J.N. 1941.'Time: A Treatment of Some Puzzles'. *Australasian Journal of Psychology and Philosophy*, vol.19, pp.216-35.
- Flew, A. (ed.) 1951. *Essays on Logic and Language*. Oxford: Blackwell.
- Geach, P. 1970. 'Arthur Prior: A Personal Impression'. *Theoria*, vol. 36, pp.186-8.
- Hintikka K.J.J. 1958. 'Review of *Time and Modality*'. *Philosophical Review*, vol. 67, pp.401-4.
- Hintikka, K.J.J. 1957. *Quantifiers in Deontic Logic*. *Societas Scientiarum Fennica, Commentationes Humanarum Litterarum*, vol. XXIII:4, Helsinki.
- Hintikka, K.J.J. 1961. 'Modality and Quantification'. *Theoria*, vol. 27, pp.119-28.
- Hughes, G.E. 1971. 'Arthur Prior (1914-1969)'. *Australasian Journal of Philosophy*, vol. 49, pp.241-3.
- Johnson, W.E. 1921-24. *Logic*. Vols 1-3. Cambridge: Cambridge University Press.
- Jonsson, B., Tarski, A. 1951. 'Boolean Algebras With Operators'. *American Journal of Mathematics*, 73, pp.891-939.
- Jordan, A.A. 1945. *The Development of Mathematical Logic and of Logical Positivism in Poland between the Two Wars*. Oxford: Oxford University Press.
- Kanger, S. 1957. *Provability in Logic*. Stockholm: Almqvist and Wiksell.
- Kenny, A. 1970. 'Arthur Norman Prior'. *Proceedings of the British Academy*, vol. LVI, pp.321-349.
- Keynes, J.N. 1906. *Studies and Exercises in Formal Logic*. London: Macmillan.
- Knuuttila, S. 1993. *Modalities in Medieval Philosophy*. London: Routledge.
- Kripke, S.A. 1959a. 'A Completeness Theorem in Modal Logic'. *The Journal of Symbolic Logic*, vol. 24, pp.1-14.

- Kripke, S.A. 1959b. ‘Semantical Analysis of Modal Logic’. (Abstract) *The Journal of Symbolic Logic*, vol. 24, pp.323-4.
- Kripke, S.A. 1963. ‘Semantical Analysis of Modal Logic I: Normal Modal Propositional Calculi’. *Zeitschr. f. math. Logik und Grundlagen d. Math.*, vol. 9, pp.67-96.
- Lewis, C.I., Langford, C.H. 1932. *Symbolic Logic*. London: Century.
- Lukasiewicz, J. 1920. ‘On Three-Valued Logic’. *Ruch Filozoficzny*, 5, pp.170-1. English translation in Borkowski, L. (ed.) 1970. *Jan Lukasiewicz: Selected Works*. Amsterdam: North Holland.
- Lukasiewicz, J. 1948. ‘The Shortest Axiom of the Implicational Calculus of Propositions’. *Proceedings of the Royal Irish Academy*, 52, pp.25-33.
- Lukasiewicz, J. 1951. *Aristotle's Syllogistic From the Standpoint of Modern Formal Logic*. Oxford: Clarendon Press.
- Mates, B. 1953. *Stoic Logic*. Berkeley: University of California Press.
- Meredith, C.A., Prior, A.N. 1996. ‘Interpretations of Different Modal Logics in the "Property Calculus"’. In Copeland, B.J. (ed.) 1996. *Logic and Reality: Essays on the Legacy of Arthur Prior*. Oxford: Clarendon Press.
- Meredith, D. 1977. ‘In Memoriam: Carew Arthur Meredith (1904-1976)’. *Notre Dame Journal of Formal Logic*, vol. 18, pp.513-16.
- Ohrstrom, P., Hasle, P. 1993. ‘A.N. Prior's Rediscovery of Tense Logic’. *Erkenntnis*, vol. 39, pp.23-50.
- Ohrstrom, P., Hasle, P. 1995. *Temporal Logic: From Ancient Ideas to Artificial Intelligence*. Dordrecht: Kluwer.
- Pnuelli, A. 1977. ‘The Temporal Logic of Programs’. *Proceedings of the Eighteenth Annual Symposium on Foundations of Computer Science*, New York: Institute of Electrical and Electronics Engineers.
- Prior, A.N. 1937. ‘The Nation and the Individual’. *Australasian Journal of Psychology and Philosophy*, vol.15, pp.294-8.
- Prior, A.N. 1942. ‘Can Religion be Discussed?’. *Australasian Journal of Psychology and Philosophy*, vol.20, pp.141-51.
- Prior, A.N. 1944. ‘The Meaning of Good’. *Australasian Journal of Psychology and Philosophy*, vol.22, pp.170-4.
- Prior, A.N. 1945. ‘The Subject of Ethics’. *Australasian Journal of Psychology and Philosophy*, vol.23, pp.78-84.
- Prior, A.N. 1949. *Logic and the Basis of Ethics*. Oxford: Clarendon Press.
- Prior, A.N. 1951. ‘The Ethical Copula’. *Australasian Journal of Philosophy*, vol.29, pp.137-54.
- Prior, A.N. 1952a. ‘Modality De Dicto and Modality De Re’. *Theoria*, vol.18, pp.174-80.
- Prior, A.N. 1952b. ‘In What Sense is Modal Logic Many-Valued?’. *Analysis*, vol.12, pp.138-43.
- Prior, A.N. 1952c. ‘Lukasiewicz's Symbolic Logic’. *Australasian Journal of Philosophy*, vol.30, pp.121-30.
- Prior, A.N. 1953. ‘Three-Valued Logic and Future Contingents’. *Philosophical Quarterly*, vol.3, pp.317-26.
- Prior, A.N. 1955a. *Formal Logic*. Oxford: Clarendon Press.
- Prior, A.N. 1955b. ‘Diodoran Modalities’. *Philosophical Quarterly*, vol.5, pp.205-13.

- Prior, A.N. 1956a. ‘Modality and Quantification in S5’. *The Journal of Symbolic Logic*, vol.21, pp.60-62.
- Prior, A.N. 1956b. ‘Logicians at play; or Syll, Simp and Hilbert’. *Australasian Journal of Philosophy*, vol.34, pp.182-92.
- Prior, A.N. 1957. *Time and Modality*. Oxford: Oxford University Press.
- Prior, A.N. 1958a. ‘The Syntax of Time-Distinctions’. *Franciscan Studies*, vol.18, pp.105-120.
- Prior, A.N. 1962a. ‘Tense Logic and the Continuity of Time’. *Studia Logica*, vol.13, pp.133-48.
- Prior, A.N. 1962b. ‘Possible Worlds’. *Philosophical Quarterly*, vol.12, pp.36-43.
- Prior, A.N. 1967. *Past, Present and Future*. Oxford: Clarendon Press.
- Prior, A.N. 1968. *Papers on Time and Tense*. Oxford: Clarendon Press.
- Prior, A.N. 1971. *Objects of Thought*. Oxford: Clarendon Press. (Edited by Geach, P.T., Kenny, A.J.P.)
- Prior, A.N. 1976a. *The Doctrine of Propositions and Terms*. London: Duckworth. (Edited by Geach, P.T., Kenny, A.J.P.)
- Prior, A.N. 1976b. *Papers in Logic and Ethics*. London: Duckworth. (Edited by Geach, P.T., Kenny, A.J.P.)
- Prior, A.N. 1977.*Worlds, Times and Selves*. London: Duckworth. (Edited by Fine, K.)
- Prior, A.N. 1996a. ‘A Statement of Temporal Realism’. In Copeland, B.J. (ed.) 1996. *Logic and Reality: Essays on the Legacy of Arthur Prior*. Oxford: Clarendon Press.
- Prior, A.N. 1996b. ‘Some Free Thinking about Time’. In Copeland, B.J. (ed.) 1996. *Logic and Reality: Essays on the Legacy of Arthur Prior*. Oxford: Clarendon Press.
- Reichenbach, H. 1948. *Elements of Symbolic Logic*. New York: Macmillan.
- Rescher, N., Urquhart, A. 1971. *Temporal Logic*. New York: Springer-Verlag.
- Segerberg, K. 1967, ‘On the Logic of "To-morrow"’, *Theoria*, vol.33, pp.45-52.
- Segerberg, K. 1989. ‘Von Wright's Tense Logic’. In Schilpp, P.A., Hahn, L.E. 1989, *The Philosophy of Georg Henrik von Wright*, Illinois: Open Court, pp.602-35.
- Sobocinski, B., 1953. ‘Note on a Modal System of Feys-von Wright’. *Journal of Computing Systems*, vol. 1, pp.171-8.
- Tarski, A. 1941. *Introduction to Logic and to the Methodology of Deductive Sciences*. New York: Oxford University Press.
- Thomas, I. 1968. ‘In Memoriam: Edward John Lemmon (1930-1966)’. *Notre Dame Journal of Formal Logic*, vol. 9, pp.1-3.
- Thomas, I. 1971. ‘In Memoriam: A.N. Prior’. *Notre Dame Journal of Formal Logic*, vol. 12, pp.129-30.
- Von Wright, G.H. 1951a *An Essay on Modal Logic*. Amsterdam: North-Holland.
- Von Wright, G.H. 1951b. ‘Deontic Logic’. *Mind*, vol. LX, pp.1-15.
- Von Wright, G.H. 1963. *Norm and Action*. London: Routledge and Kegan Paul.
- Von Wright, G.H., 1965. ‘"And Next"’, *Acta Philosophica Fennica*, fasc. XVIII, pp.293-304.
- Watts, I. 1726. *Logick, or the Right Use of Reason in the Enquiry After Truth*. 2nd edition. London: John Clark and Richard Hett.
- Weinberg, J. 1948. *Nicolaus of Autricourt: A study in 14th Century Thought* . New York: Greenwood Press.

Other Internet Resources

Related Entries

logic: modal | possible worlds

Acknowledgements

My sources for this essay -- other than Prior's papers and correspondence, which are held in the Bodleian Library, Oxford, and his published work -- are: conversations and/or correspondence with: Jonathan Bennett, Colin Brown, Robert Bull, Nino Cocchiarella, Vincent Denard, John Faris, Dov Gabbay, Peter Geach, Jaakko Hintikka, George Hughes, Hans Kamp, Saul Kripke, Peter Ohrstrom, Mary Prior, Stephen Read, Dana Scott, Krister Segerberg, Jack Smart, Richard Sylvan, Jim Thornton, Jim Wilson, Georg Henrik von Wright; Kenny (1970); Cresswell and Crossley (1989), which is an unpublished edited transcript of a panel discussion concerning Prior held at the 1981 Annual Conference of the Australasian Association for Logic in Wellington (the participants were Robert Bull, Martin Bunder, Max Cresswell, John Crossley, Charles Hamblin, George Hughes, John Kalman, David Lewis, Michael McRobbie, Wilf Malcolm, Ken Pledger, Tom Richards, Krister Segerberg and Pavel Tichy); Hughes (1971); Thomas (1968, 1971); Geach (1970); Meredith (1977); and the Annual Calendars of Canterbury University College. A number of people commented helpfully on earlier versions of this material: Colin Brown, George Hughes, Saul Kripke, David Lewis, Peter Ohrstrom, Diane Proudfoot, Stephen Read, Krister Segerberg, Miriam Solomon, and Bob Stoothoff.

B. Jack Copeland
J.Copeland@phil.canterbury.ac.nz

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 7, 1996
Content last modified: November 29, 1999

Actualism

To understand the thesis of actualism, consider the following example. Imagine a race of beings -- call them ‘Aliens’ -- that is very different from any life-form that actually exists anywhere in the universe; different enough, in fact, that no actually existing thing could have been an Alien, any more than a given gorilla could have been a fruitfly. Now, even though there are no Aliens, it seems intuitively true that there could have been such things. After all, life could have evolved very differently than the way it did in fact, differently enough, at least, that other kinds of things might have existed. So why is it true that there could have been Aliens when in fact there are none, and when, moreover, nothing that actually exists could have been an Alien?

To answer this question, a philosopher should try to identify the special features of the world that are responsible for the truth of claims about what could have been been the case. One group of philosophers, the *possibilists*, offer the following answer: ‘It is possible that there are Aliens’ is true because there are in fact individuals that could have been Aliens. By hypothesis, however, such individuals are simply possible and not actual. No actually existing thing could possibly have been an Alien. Hence, the truth of ‘It is possible that there are Aliens’ is, according to possibilism, grounded in the fact that there are possible-but-nonactual Aliens, i.e., things that are not actual but which could have been, and such that, moreover, if they had been actual, they would have been Aliens.

Actualists reject this answer; they deny that there are any nonactual individuals. Actualism is the philosophical position that everything there is -- everything that can be said to exist in any sense -- is *actual*. Put another way, actualism denies that there is any kind of being beyond actuality; to be is to be actual. Actualism therefore stands in stark contrast to possibilism, which, as we've seen, takes the things there are to include possible but non-actual objects.

Of course, actualists will agree that there could have been Aliens. Actualism, therefore, can be thought of as the metaphysical theory that attempts to account for the truth of claims like ‘It is possible that there are Aliens’ without appealing to any nonactual objects whatsoever. What makes actualism so philosophically interesting, is that there is no obviously correct way to account for the truth of claims like ‘It is possible that there are Aliens’ without appealing to possible but nonactual objects. In the rest of this article, we will lay out the various attempts to do so in some detail and assess their effectiveness.

---

§1: The Possibilist Challenge to Actualism

The fundamental thesis of actualism is:[1]

> (A) Everything that exists (i.e., everything there is) is actual.

Possibilism is the denial of this thesis and there are various forms of possibilism that correspond to the various ways in which one can deny Thesis (A). (This is explained in more detail in the supplementary document: Three Types of Possibilism.)

The possibilist challenge to actualism is to give an analysis of our ordinary modal beliefs which is consistent with Thesis (A), i.e., which doesn't appeal to possible but nonactual objects. There are two central aspects to the possibilist challenge: the challenge of possible worlds, and the challenge of possible objects. The latter will be the central focus of this article, but, for the sake of completeness, we begin with a brief discussion of the former.

Worlds

Claims such as ‘it is possible that there are Aliens’, ‘it is possible that there is a planet disturbing the orbit of planet X’, and ‘it is necessary that everything is self-identical’ are known as *modal* claims, because the sentential prefixes ‘it is possible that’ and ‘it is necessary that’ indicate a ‘mode’ in which the statements they precede are true. Modal claims are ubiquitous in our thought and discourse. Many of our reflective and creative thoughts seem to be about possibilities (consider, for example, the possibility that there are clean, fuel-efficient automobiles that cause no harm to the environment) and much of our logical

reasoning involves drawing conclusions which, in some sense, necessarily follow from premises that we already believe. Modal logic is the logic of possibility and necessity and the study of modal logic, as a discipline, has flourished in the latter half of the twentieth century. This was due in no small part to the introduction of the concept of *possible worlds* to investigate the truth conditions of modal claims. A large part of the logic of possibility and necessity seems to be captured by treating the modal operators ‘it is necessary that’ and ‘it is possible that’ as quantifiers over possible worlds. That is, the following semantic analyses seem to capture a large part of the logic of modality:

> (a) The statement ‘It is necessary that *p*’ (‘$\Box p$’) is true just in case *p* is true in all possible worlds.
>
> (b) The statement ‘It is possible that *p*’ (‘$\Diamond p$’) is true just in case *p* is true in some possible world.

Notice that it is a consequence of analysis (b) that true claims asserting a possibility imply the existence of possible worlds.

On the face of it, then, the possible worlds analysis of basic modal statements just sketched appears to entail the existence of nonactual possible worlds, and hence appears directly to contradict Thesis (A). Consequently, actualists either have to try to develop a semantics for modal statements in terms that do not entail the existence of nonactual possible worlds, or at least to provide an account of possible worlds on which this consequence is rendered metaphysically innocuous.

The power of the possible worlds semantics -- and the distinct lack of any persuasive alternatives -- is very attractive to many actualists, and they are loathe to give it up (so long, of course, as they do not have to abandon actualism). Consequently, actualists typically grasp the second horn of the above dilemma and adopt some sort of actualistically acceptable, "sanitized" version of this theory on which possible worlds are conceived as theoretical *abstract* objects which actually exist. Many such theories of abstractly-conceived worlds have been developed, some with better success than others (see, for example, Plantinga [1974] and [1976], Chisholm [1976], Fine [1977], Adams, [1974], van Inwagen [1986], or Zalta [1983] and [1993]). Some take worlds to be maximal possible states of affairs, others take them to be maximal possible properties or propositions, still others treat them as maximal consistent sets of some sort, and yet others treat them as part of a more general theory of abstract objects. For purposes here, it will serve well enough just to assume some generic version of this view on which such abstractly conceived worlds can perform their theoretical tasks in virtue of certain actualistically unobjectionable modal properties. A detailed version of such an account, and some of its philosophical ramifications, can be found in the supplementary document on An Account of Abstract Possible Worlds.

Mere *Possibilia*

The second step in the actualist analysis of modality is to find a way to do without possible but nonactual individuals or, at least, a way to replace them with less objectionable entities like properties of some ilk.

Possible but nonactual individuals -- also known as *mere possibilia* or *contingently nonactual* individuals -- seem to be required for the analysis of modal claims involving quantifiers such as ‘there is’ or ‘there exists’. Consider, first, a non-modal quantifier claim, such as ‘There are Aliens’. Such a claim might be regimented in first-order logic as "There exists an *x* such that *x* is an Alien", or in formal terms (in which ‘*Ax*’ abbreviates the predicate ‘*x* is an Alien’): $\exists xAx$.

Now consider the modal claim ‘There could have been Aliens’. It is natural to regiment this claim as "It is possible that there exists an *x* such that *x* is an Alien", which is typically formalized as follows:

(1) $\Diamond\exists xAx$.

Now, if we deploy some acceptable theory of possible worlds, we know that sentence (1) is true if and only if:

(2) There exists a possible world **w** and there exists an individual *x* such that *x* is an Alien at **w**.

But, it is a fact about the logic of the quantifier ‘there exists’ that such quantifiers ‘commute’ with one another. In other words, (2) implies (3):

(3) There exists an individual *x* and there exists a possible world **w** such that *x* is an Alien at **w**.

So the truth conditions of (1) imply (3). But if (3) is true, then so is the ordinary modal claim ‘Something is possibly an Alien’, i.e.,

(4) $\exists x\Diamond Ax$

for which (3) provides the truth conditions. Thus, given the simplest logic concerning modal and quantifier claims, (1) implies (4). In other words, the simplest quantified modal logic tells us that (5) implies (6):

(5) It is possible that there exists an *x* such that *x* is an Alien.

(6) There exists an *x* such that it is possible that *x* is an Alien.

The problem for the second step of the actualist treatment of modality may now be stated more precisely, namely, Thesis (A) is inconsistent with (6). Thesis (A) asserts that everything actually exists. But (6) seems to assert the existence of a possible Alien. There seem to be no candidates among the actually existing individuals which we might plausibly identify as a possible Alien.[2] Thus, the consequences of our ordinary modal beliefs that are valid according to the simplest quantified modal logic seems to be

inconsistent with actualism.[3]

Since it seems reasonable to want to hang on to such ordinary modal beliefs as (5), there is an apparent incompatibility between the simplest quantified modal logic and actualism. This is only the tip of the iceberg, however, for the problem described in the previous paragraph resurfaces each time we ‘nest’ or ‘iterate’ modalities. Consider, for example, the following sentences:

(7) The Pope (i.e., Karol Wojtyla) could have had a son who could have become a priest.

(8) There could be a planet disturbing the orbit of Pluto and it could have a period of *n* years.

Such sentences seems to be representable as follows:

(9) $\Diamond\exists x(Sxp \;\&\; \Diamond Px)$

(10) $\Diamond\exists x(Lx \;\&\; Dxp \;\&\; \Diamond Pxn)$

These cases pose a serious problem for any actualist metaphysics. Even if we assume that actualists can successfully develop a metaphysics and logic that explain the truth of the first occurrence of ‘could’ in (7) and (8), respectively, a serious question arises concerning the second occurrence. The simplest logic of the second occurrence of the "nested" modal operator in (9) and (10) would suggest that it describes a modal fact about a possible individual -- a possible son of the pope in (9), and a possible planet disturbing the orbit of Pluto in (10). (9) seems to assert that a possible son of the pope has the modal property of *possibly becoming a priest*. (10) seems to assert that the possible planet disturbing Pluto has the modal property *possibly having a period of n years* (for some *n*). These cases of ‘nested’ modalities and the problems they pose for actualism were first discussed in a forceful way in McMichael [1983]. We will return to this issue at several points below.

Where We Go From Here

As the reader who works through the remainder of this essay will discover, the simplest quantified modal logic has numerous consequences that seem incompatible in some way or another with actualism. In the next section, we will discover still other such consequences. Though we have succeeded in describing the issues surrounding actualism in more precise terms, we have only scratched the surface of the debate. Much of the debate turns on the precise characteristics of the modal logic being proposed as a logic for actualism. This debate can only be understood if one can contrast the characteristics of these proposed alternative logics with the characteristics of the simplest modal logic. Thus, we will spend the next section of this essay describing the characteristics of the simplest modal logic. Only then will we be in a position to evaluate the more complicated alternatives developed by actualists in the attempt to avoid commitment to nonactual possibles. For example, it is important to see just how Kripke's modal logic (Kripke [1963]) employs a variety of special techniques that yield a logic consistent with Thesis (A) (these will be documented below).

The remaining sections of this essay, therefore, contain the following material. In Section §2, we describe, in a precise way, both the characteristics of the simplest quantified modal logic and its controversial theorems. (As we acquire more sophisticated logical tools, we will revisit some of the examples already discussed; the redescription of these examples in more sophisticated logical terms may prove to be instructive.) We also show why each of the controversial theorems is objectionable from the standpoint of actualism. In Section §3, we outline a modal system developed by Saul Kripke that appears to be consistent with Thesis (A). However, in Section §4, we'll discover that Kripke's system introduces special problems of its own. Finally, in Section §5, we discuss the various attempts actualists have made to work within a Kripke-style framework to solve these problems and to find a metaphysical theory of necessity and possibility which is consistent with Thesis (A). However, we will also examine the attempts of some actualists who have recently discovered a new interpretation of the simplest quantified modal logic which is consistent with Thesis (A).

§2: The Simplest Quantified Modal Logic

A first-order quantified modal logic is a group of logical axioms and rules of inference that systematizes the logically true sentences of a standard first-order modal language with identity (**L**) relative to some class of interpretations of this language. The language **L** is defined just like the language of the predicate calculus with identity, but with the following additional clause in the definition of a 'formula': whenever φ is a formula, so is $\Box\varphi$. Thus, the language will have constants and variables for individuals, *n*-place predicates, atomic formulas such as '$P^n a_1 ... a_n$' and '$x=y$', and the usual molecular, quantified and modal formulas involving the logical notions expressed by '$\neg$' (the negation symbol), '$\rightarrow$' (the symbol for forming conditionals), '$\forall$' (the universal quantifier), and '$\Box$' (the modal operator). The other logical connectives, such as '&' (and), '$\vee$' (or), and '$\equiv$' (iff), and the 'existential' quantifier '$\exists$', are all defined in the usual way. The formula $\Diamond\varphi$ is defined as $\neg\Box\neg\varphi$. (To assert that φ is possibly true is to say that φ is not necessarily false.) From this definition, the equivalence of $\Box\varphi$ and $\neg\Diamond\neg\varphi$ also follows. For convenience, a complete specification is provided in the supplementary document A First-order, Quantified Modal Language. It would serve well to spend a moment or two examining these definitions to making sure that you understand the kinds of statements that are expressible in this language.

The simplest semantics for the language **L** defines a class of interpretations having two distinguishing features: (1) each interpretation **I** in the class has just two, mutually exclusive domains--a nonempty domain of possible worlds (which includes a distinguished "actual world" $\mathbf{w_0}$) and a nonempty domain of individuals, and (2) given any individuals $\mathbf{i}_1$,..., $\mathbf{i}_n$ in the domain and given any possible world **w**, each interpretation **I** specifies, for each *n*-place predicate '*R*', whether '*R*' applies to $\mathbf{i}_1$,..., $\mathbf{i}_n$ at **w** or not. Given that specification, the semantics then defines truth conditions for every formula of the language. The definition of truth even accomodates 'open formulas' (i.e., formulas with free variables) by appealing to assignment functions **f** which assign to each variable *x* some member **f**(*x*) of the domain of individuals. Moreover, given any interpretation **I** and assignment function **f** to the variables, a denotation function **d** relative **I** and **f** is defined for the terms (constants and variables) of the language. When τ is a constant,

$\mathbf{d}_{\mathbf{I},\mathbf{f}}(\tau)$ is the individual in the domain that **I** assigns to τ. When τ is a variable, $\mathbf{d}_{\mathbf{I},\mathbf{f}}(\tau)$ is $\mathbf{f}(\tau)$.

The semantic notion 'φ is true (under interpretation **I** and assignment **f**) at world **w**' ('true$_{\mathbf{I},\mathbf{f}}$ at **w**') is then defined recursively for all of the formulas of the language. The three mosts important parts of this definition for quantified modal logic are the clauses for atomic, quantified, and modal formulas. Here are examples of each:

1. The open, atomic formula '*Px*' is true$_{\mathbf{I},\mathbf{f}}$ at **w** just in case **I** specifies that '*P*' applies to $\mathbf{d}_{\mathbf{I},\mathbf{f}}(x)$ at **w**.
2. The quantified formula '$\forall xPx$' is true$_{\mathbf{I},\mathbf{f}}$ at **w** just in case, for all individuals **a**, '*Px*' is true$_{\mathbf{I},\mathbf{f}[x,\mathbf{a}]}$ at **w**, where $\mathbf{f}[x,\mathbf{a}]$ is **f** if $\mathbf{f}(x) = \mathbf{a}$, and otherwise is just like **f** except that it assigns **a** to *x* instead of $\mathbf{f}(x)$. (A little less formally, '$\forall xPx$' is true$_{\mathbf{I},\mathbf{f}}$ at **w** just in case, for all individuals **a**, the predicate '*P*' applies to **a** at **w**.)
3. The open, modal formula '$\Box Px$' is true$_{\mathbf{I},\mathbf{f}}$ at **w** just in case for every possible world $\mathbf{w}'$, '*Px*' is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}'$.

A formula φ (which may have the variable *x* free) is then defined to be true$_{\mathbf{I}}$ just in case for every assignment **f**, φ is true$_{\mathbf{I},\mathbf{f}}$ at the actual world $\mathbf{w}_0$. (Note that when φ is a closed formula (i.e., a sentence), then if φ is true relative to some assignment to the variables, it is true relative to all assignments to the variables.) A formula is logically true just in case it is true$_{\mathbf{I}}$ in all interpretations **I** (in this class of interpretations). For convenience, we reproduce here a precise definition in the supplementary document The Simplest Semantics for a Quantified Modal Language. It would serve well to study these definitions if they are unfamiliar. (Readers with some familiarity with modal logic will recognize that we have formulated the semantics without an accessibility relation. However, no such relation is required for a correct semantics of S5, which, in order to keep things simple, we will be presupposing henceforth.)

The simplest quantified modal logic (SQML) systematizes the logically true sentences of **L** relative to the simplest semantics. SQML combines the logical axioms and rules of inference from classical propositional logic, classical first-order quantification theory, the logic of identity, and S5 modal logic. We presuppose here that the laws of classical first-order logic with identity are known. The system S5 (aka KT5) adds three axioms to classical first-order logic with identity--the K axiom, the T axiom, and the 5 axiom (see below)--and adds the Rule of Necessitation (RN) (which states that whenever φ is a theorem, so is $\Box\varphi$). Each of the logical axioms of the resulting SQML is true in every interpretation in the class described in the previous paragraph. Moreover, the rules of inference 'preserve truth' (and preserve logical truth). That is, the rules of inference permit one to infer only (logical) truths from any set of premises consisting solely of (logical) truths. Notice that open formulas are assertible as axioms and provable as theorems in SQML. For convenience, we reproduce these in the supplementary document The Simplest Quantified Modal Logic. Familiarity with this logic will be presupposed in what follows.

The problem that SQML poses for actualist philosophers is that whereas all of the logical axioms appear to be true, some of the logical consequences of these axioms appear to be false. Consider first the fact that the new modal axioms added by SQML to classical first-order logic, i.e., the K, T, and 5 axioms, all seem

true. The K axiom asserts that if a conditional is necessary, then if the antecedent is necessary, so is the consequent:

K axiom: $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$

It is easy to see that this is true in every interpretation of the class of simplest interpretations: if a conditional is true in every possible world and the antecedent of the conditional is true in every world, then the consequent of the conditional is true in every world.

The ‘T’ axiom asserts that a formula true in every possible world is simply true:

T axiom: $\Box\varphi \rightarrow \varphi$

Clearly, this is true in all interpretations--if a sentence is true in every possible world, it is true in the distinguished actual world.

The ‘5’ axiom asserts that if a formula φ is possible, then it is necessarily the case that it is possible:

5 axiom: $\Diamond\varphi \rightarrow \Box\Diamond\varphi$

It is not hard to see that this is logically true. If a formula is true in some possible world, then from the point of view of every possible world, the formula is true in some possible world. That is, if a formula is true at some possible world, then at every possible world, there is some possible world where the formula is true. (The formal validity of the 5 Axiom is proved in the supplementary document The 5 Axiom is Logically True.)

Controversial Consequences of SQML

However, the controversies surrounding actualism and modal logic center on the following theorems, (the instances of) which are all logically true and provable from the axioms and rules of SQML:

BF: $\forall x\Box\varphi \rightarrow \Box\forall x\varphi$

NE: $\forall x\Box\exists y\, y{=}x$

CBF: $\Box\forall x\varphi \rightarrow \forall x\Box\varphi$

BF, NE, and CBF are, respectively, the Barcan Formula, Necessary Existence, and the Converse Barcan Formula. It is reasonably straightforward to establish that NE and the instances of BF and CBF are all logically true and that they are all derivable from the axioms and rules of SQML. The proofs of these claims are provided in following supplementary documents:

[The Barcan Formula is Logically True](#)
[The 'Necessary Existence' Principle is Logically True](#)
[The Converse Barcan Formula is Logically True](#)

We also show how to use the axioms and rules of SQML in the following documents:

[Proof of the Barcan Formula in S5](#)
[Proof of 'Necessary Existence' in S5](#)
[Proof of Converse Barcan Formula in S5](#)

Before we turn to the discussion of why actualists find BF, NE, and CBF objectionable, there are two minor details to attend to. The first is that the Barcan Formula is often discussed in the following equivalent form (indeed, this was the formulation that played a role in our discussion in the first section of the present essay):

$$\Diamond\exists x\varphi \rightarrow \exists x\Diamond\varphi$$

(In the supplementary document, [Proof of Barcan Formula Equivalent](#), we show that this statement is equivalent to BF.) Referring back to our Alien example, then, BF in this form, where φ is '*Ax*', asserts that proposition (1) of the previous section implies (4), or, in terms of their ordinary language counterparts, that (5) implies (6).

Why Actualists Find These Consequences Unacceptable

With the machinery of SQML laid out before us, it is instructive to consider why its consequences BF, NE, and CBF offend the actualist. Recall that, according to actualism, everything there is, in any sense, is actual. Consider first the Barcan Formula. We have seen how BF leads to problems in the case of the Alien example, but, to drive the point home, it is perhaps worth considering yet another case that has appeared in the literature. Linsky and Zalta (1994) clearly express the problem for actualists by having the reader consider a sisterless person *b*; presumably, though sisterless in fact, most everyone would agree that it is at least possible that *b* have a sister. But given that possibility,

> ... BF requires that *there exists* something that is possibly *b*'s sister. Since *b* has no sisters, which existing object is it that is possibly *b*'s sister? Some actualists, notably Ruth Marcus [1986], might defend BF by pointing to an existing woman (possibly one closely related to *b*) and suggesting that she is the thing which both exists and which is possibly *b*'s sister. But the great majority of actualists don't accept this idea, for they subscribe to certain essentialist views about the nature of objects. For example, they believe that women who aren't *b*'s sister could not have been (in a metaphysical sense) *b*'s sister. This is a fact about their very nature, one concerning their origins. ... Since there seems to be no *actually existing* thing which is possibly *b*'s sister, they conclude BF is false. We think the

> essentialist intuitions leading to this conclusion are not unreasonable, and so understand why these actualists take BF to be false. Indeed, it seems that BF, in general, is incompatible with the intuition that there might have been something distinct from every actual thing. It is hard to see how that intuition could be compatible with a principle which seems to require that every possibility be grounded in something that exists. This is further evidence actualists have against the acceptability of BF. But since they still want to make sense of modal discourse in terms of possible world semantics, they reject the Barcan formula as having unacceptable consequences, and search for a modal semantics on which it is not valid.

Consider next NE. For actualists, this explicitly says that for any object *x*, necessarily something exists that is identical with *x*. Now most actualists accept the following definition of what it is for an object to exist:

$$x \textit{ exists} =_{df} \exists y(y = x)$$

Given this definition, NE says that everything necessarily exists. (hence our abbreviation ‘NE’). Prior in [1957] was especially concerned by this, pointing out that classical quantified modal logic was "haunted by the myth that whatever exists exists necessarily." Note that NE applies even to those objects not named by a constant of the language. These consequences run counter to our ordinary (modal) intuitions. They are not inconsistent with Thesis (A), but are instead inconsistent with the reasonable belief that some objects might not have existed. Actualists see this as an additional and independent reason to abandon SQML.

Another problem with NE is that it leads to an even stronger result. By applying the Rule of Generalization and then the Rule of Necessitation to NE, one obtains:

NNE: $\Box\forall x\Box\exists y\, y=x$

This asserts that it is necessary that everything necessarily exists. It follows that it is not even possible for there to be contingent individuals. To see why, note that an actualist would define ‘contingent individual’ as follows:

$$x \textit{ is contingent} =_{df} \Diamond\neg\exists y(y = x)$$

Consequently, the claim that there are contingent individuals would be formulated as:

$$\exists x\Diamond\neg\exists y(y = x)$$

But it now follows from NNE that it is not possible that there are contingent individuals:

$\neg\Diamond\exists x\Diamond\neg\exists y(y = x)$

(The proof is left as an exercise.) If this is a consequence of SQML, it is no wonder actualists are dissatisfied.

Finally, there is CBF. For the actualist, the main problem with CBF is that in SQML it implies NE in conjunction with the thesis known as *Serious Actualism*. Serious Actualism is the thesis that it is not possible for an object to have a property without existing, i.e., that if an object exemplifies a property at a world, it exists at that world. [See Plantinga [1983], [1985], Menzel [1991], Pollock [1985], and Deutsch [1993] for various discussions of Serious Actualism.] In semantic terms, this amounts to the constraint that an object in the extension of a property at a world must fall under the range of the quantifier at that world. Serious Actualism is often expressed by the following schema of the object language:

SA: $\Box[\varphi \rightarrow \exists y\, y{=}x]$, where φ is atomic and contains x.

Note that SA is a theorem of SQML (the proof is a simple exercise) and seems consistent with the actualist point of view. But from SA and CBF, one can rederive NE from any (logically) necessary property in the system. (See A Derivation of NE from SA and CBF.) Thus, even if there were a way to block the direct derivation of NE, the alternative derivation of NE from CBF and SA shows that serious actualists could not accept SQML unless CBF is somehow invalidated.

As noted, Arthur Prior was the first to realize the controversial consequences of SQML. Prior himself was a staunch actualist, and dealt with the problem by developing an alternative quantified modal logic that differs significantly from SQML, and which does not have the controversial principles above as theorems. A more detailed account of Prior's approach can be found in the supplementary document Prior's Modal Logic.

§3: Kripke's System

Given the above consequences of SQML, one can understand why actualists would seek a reformulation of quantified modal logic that both: (1) defines interpretations so that BF, NE, and CBF are not logically true, and (2) weakens the proof theory of SQML so that BF, NE, and CBF are not derivable as theorems of the logic. Kripke's logic appealed to actualists and serious actualists for these very reasons. The system of Kripke [1963] invalidates BF, NE, and CBF both model-theoretically and proof-theoretically.

It is illuminating both to see exactly how Kripke was able to construct interpretations on which BF, NE, and CBF are not logically true and to see exactly how Kripke modified the logic of SQML so that these these schemata and sentences are no longer theorems. These techniques will be the subject of the next two subsections.

Kripke Models

The key insight in Kripke's quantified modal logic is the replacement of the single domain **D** of individuals in the interpretation of a first-order modal language with a function **dom** that assigns to each world **w** its own distinct domain of individuals **dom**(**w**). No restrictions are placed on the domain of a world; any set of individuals, including the empty set, will do. Thus, instead of a single domain common to all worlds, domains are permitted to vary from world to world. Intuitively, of course, **dom**(**w**) represents the objects that *exist* in **w**. In particular, the domain of the actual world represents--of course--the things that are actual, the things that exist simpliciter. Interpretations like this for first-order modal languages in which each world has its own domain are known as *Kripke models*.

The central semantic difference between Kripke models and interpretations for SQML is that, in a Kripke model, when a quantified formula $\forall x\varphi$ is evaluated at a world **w**, the quantifier ranges only over the objects that exist in the domain of **w**. Thus, in particular, the sample clause in the definition of truth for quantification above must be revised for Kripke models **M** as follows, where, again, **f** is an assignment function:

> The quantified formula '$\forall xPx$' is true$_{\mathbf{M},\mathbf{f}}$ at **w** just in case, for all individuals *a* in **dom**(**w**), 'Px' is true$_{\mathbf{M},\mathbf{f}[x,\mathbf{a}]}$ at **w**, where **f**[x,**a**] is **f** if **f**(x) = **a**, and otherwise is just like **f** except that it assigns **a** to x instead of **f**(x).

Again, a little less formally, '$\forall xPx$' is true$_{\mathbf{I},\mathbf{f}}$ at **w** just in case, for all individuals **a** *that exist in **w***, the predicate 'P' applies to **a** at **w**.

Kripke's changes to the model theory of first-order modal languages are relatively simple. Nonetheless, unlike the model theory for SQML, Kripke's model theory yields a set of logical truths that is fully compatible with actualism. In particular, all three of the principles with which the actualist takes issue -- BF, NE, and CBF -- turn out to be invalid in Kripke's semantics. Consider first BF in the form

> $\Diamond\exists x\varphi \rightarrow \exists x\Diamond\varphi$

For definiteness, let φ be the formula 'Ax' expressing that x is an Alien (in the sense of 'Alien' introduced above). As we noted, even though there are no Aliens in the actual world $\mathbf{w_0}$, there could have been; that is, there is a possible world **w** in which there are Aliens. Thus, on Kripke's way of evaluating quantified formulas, the antecedent to BF -- '$\Diamond\exists xAx$' -- comes out true at the actual world: '$\Diamond\exists xAx$' is true at $\mathbf{w_0}$ if and only if there is some world **u** at which '$\exists x$Ax' is true, and that, in turn, is true at such a **u** just in case some entity *that exists in **u*** is an Alien there. Assuming, as we are, that there is such a world, then, '$\Diamond\exists x$Ax' is indeed true, i.e., true at the actual world. However, as we also noted, no actually existing thing is an Alien. Thus, there is nothing in the domain of the actual world such that '$\Diamond Ax$' is true of it, that is, nothing **a** in the actual world and no world **u** are such that 'A' is true of **a** at **u**. So, in the case in question, the antecedent of BF is true, but the consequent is false. So BF is not valid, that is, there are models in which some of its instances are false.

It should be obvious why NE is also invalid in Kripke's semantics: domains of worlds can be empty. Thus, let $\mathbf{M_1}$ be a Kripke model containing at least one actual individual (i.e., at least one object **a** $\in$ **dom**($\mathbf{w_0}$)) and a world **w** that has an empty domain. Then, obviously, **a** does not exist in **w**, and hence '$\exists y\, y=x$' is false at **w** when '**a**' is assigned to *x*. Thus, '$\Box\exists y\, y=x$' is false at $\mathbf{w_0}$ when **a** is assigned to '*x*', and so, because **a** is in the domain of the actual world $\mathbf{w_0}$, NE (hence also NNE, of course) is false in $\mathbf{M_1}$.

Finally, we note that CBF is invalid in Kripke's semantics. First, let $\mathbf{M_2}$ be a Kripke model in which NE is false. (We just proved the existence of such a model in the previous paragraph, of course.) Next, let the predicate '*P*' express a property in $\mathbf{M_2}$ that is *universal* and *existence-entailing*, that is, a property which is exemplified at each world **w** by everything that exists at **w**. (Such a property is said to be existence-entailing because, necessarily, anything that has it exists. The property *existence*, of course, is the simplest example of such a property.) We can represent the universal, existence-entailing character of this property in the predicate '*P*' formally in $\mathbf{M_2}$ simply by stipulating that the extension of '*P*' at any world **w** of $\mathbf{M_2}$ is **dom**(**w**), so that all and only the things that exist at each world are in the extension of *P* at that world. Now, note that, under these conditions, it is true in $\mathbf{M_2}$ (i.e., true at the actual world $\mathbf{w_0}$ of $\mathbf{M_2}$) that $\Box\forall xPx$: intuitively, everything that exists in every possible world has the property *P* in that world. Let **a** be any object in the actual world $\mathbf{w_0}$ that fails to exist in some world **w**. (Since, by hypothesis, NE fails in $\mathbf{M_2}$, there must be such an object in **dom**($\mathbf{w_0}$).) Because *P* is existence-entailing, **a** is not in the extension of '*P*' at **w**. So there is something in the actual world $\mathbf{w_0}$ that does not have the property *P* in every possible world, i.e., '$\forall x\Box Px$' is false in $\mathbf{M_2}$. Hence, the instance '$\Box\forall xPx \rightarrow \forall x\Box Px$' of CBF is false in $\mathbf{M_2}$.

Note that the invalidity of CBF opens the door back up to serious actualism (SA) in Kripke's semantics, as it was the combination of SA with CBF that led to trouble (i.e., trouble for the actualist) in SQML. And it is easy to see formally that this is the case by constructing a Kripke model in which SA is true. Note first that the thesis of serious actualism can be expressed as the thesis that all properties are existence-entailing; there is no possible world in which something has a property but fails to exist in that world. In first-order languages, properties are represented by predicates, and having a property is represented semantically by being in the extension of a predicate. Thus, to represent a property as existence-entailing in a Kripke model, one simply ensures that, at every possible world **w**, the extension of the predicate representing that property consist only of things that exist in **w**. Hence, to represent *all* properties as existence entailing, and hence, to make SA true, one ensures that this is so for all the predicates of one's language. Formally, then, let $\mathbf{M_3}$ be any Kripke model satisfying the condition that, for every *n*-place predicate *F* and world **w**, *F* is interpreted so that, at **w**, *F* applies only to things that exist in **w**; more formally, for individuals $\mathbf{i}_1, \ldots, \mathbf{i}_n$ of $\mathbf{M_3}$, *F* applies to $\mathbf{i}_1, \ldots, \mathbf{i}_n$ at **w** only if $\mathbf{i}_1, \ldots, \mathbf{i}_n \in$ **dom**(**w**). This condition ensures that SA is true in $\mathbf{M_3}$.

The compatibility of SA with Kripke's semantics is yet further evidence of its suitability as a formal semantics for the actualist. A question that remains is: What sort of *logic* does this semantics yield?

Kripke's Quantified Modal Logic

A logic formulated in a given language **L*** is said to be *sound* and *complete* with respect to a semantics for **L*** if and only if all and only those formulas of **L*** that are valid relative to that semantics (i.e., true in every interpretation or model of the semantics) are theorems of the logic. A sound and complete logic for a semantics is a good thing, of course, as it provides a purely syntactic, proof-theoretic mechanism for demonstrating the semantic validity of formulas and arguments in the language.[4].

SQML is sound and complete relative to the semantics given for its language **L** above. Since BF, NE, and CBF are invalid in Kripke's semantics, SQML is obviously not sound and complete for it; more specifically, it is not sound: some of SQML's theorems--notably, BF, NE, and CBF--are not valid, as we saw in the previous section. Hence, Kripke must modify SQML to block their derivation without blocking the derivation of any valid formulas.

The key element to Kripke's solution to this problem is the generality interpretation of free variables. The proof of NE in SQML relies crucially on the ability to derive theorems involving free variables, and more specifically on the application of the rule of Necessitation to such theorems. As described in the supplementary document Proof of ‘Necessary Existence’ in S5, the proof makes use of the following instances of logical axioms:

- $x=x$
- $\forall y\, y \neq x \rightarrow x \neq x$.

By contraposition and quantifier exchange rules, the latter axiom is equivalent to $x=x \rightarrow \exists y\, y=x$. Thus, given $x=x$, we have $\exists y\, y=x$. The crucial step now is the application of Necessitation to this formula -- containing, we note, the free variable x -- to yield $\Box\exists y\, y=x$, which in turn yields NE, by Generalization.

To repair this "flaw" in SQML, Kripke proposes no changes to SQML other than this: a formula φ containing free variables $x_1,..., x_n$, when asserted as a theorem (and hence, in particular, as an axiom), is taken to be an abbreviation for its universal closure $\forall x_1...\forall x_n\varphi$. Thus, under this proposal, the axioms used in the proof of NE noted above are no longer assertible as theorems (they cannot appear in the lines of a proof) --- they must be taken to be abbreviations of:

- $\forall x\, x=x$
- $\forall x[\forall y\, y \neq x \rightarrow x \neq x]$

respectively. From the second axiom displayed above we can derive $\forall x(x=x \rightarrow \exists y\, y=x)$ and, from this $\forall x\, x=x \rightarrow \forall x\exists y\, y=x$ (by the quantifier distribution axiom). So using the first axiom displayed above, we can now derive $\forall x\exists y\, y=x$ by Modus Ponens, and, finally, by Necessitation we can derive only $\Box\forall x\exists y\, y=x$. This latter, for Kripke, is uproblematically and uncontroversially true -- in every possible world, every individual existing in that world is identical to something (viz., itself). To derive NE from

this, however, we need CBF -- in particular, the instance $\Box\forall x\exists y\, y=x \rightarrow \forall x\Box\exists y\, y=x$. But, as Kripke points out, the proof of CBF in SQML also depends essentially on an application of Necessitation to a theorem containing a free variable -- the same "flaw" that infects the proof of NE. (See the inference from line 1 to line 2 in the supplementary document Proof of the Converse Barcan Formula in S5.) Hence, it, too, fails under the generality interpretation of free variables. The proof of BF fails for the same reason. Hence, the SQML proofs of all three actualistically unacceptable principles fail in Kripke's system.

The failure of those particular proofs, of course, does not mean that the principles in question cannot be proved some other way. However, Kripke guarantees their unprovability in his system by showing that the system is sound and complete relative to his semantics. Soundness, in particular, tells us that no invalid formula is provable in the system. Hence, since NE, CBF, and BF are all invalid in Kripke's semantics, soundness guarantees that they are all unprovable in his system.

§4: Is Kripke's System Actualist?

On the face of it, Kripke's system provides the actualist with a powerful alternative to SQML. However, although BF, NE, and CBF are neither valid nor provable in Kripke's system, the system is open to several serious objections.

First, Kripke regards the loss of free variables from assertible sentences as a mere inconvenience. However, much of mathematical reasoning is carried out in terms of sentences with free variables, and one should at least wonder why modal logic, as opposed to classical logic, can't be formulated with free variables in assertible sentences. Far more serious, however, is the fact is that, as the system stands, one cannot add constants for contingent beings. For suppose we add a constant '*c*' to Kripke's system. As noted in the last section, $\forall x\, x=x \rightarrow \forall x\exists y\, y=x$ is a theorem of Kripke's logic. Hence, by the identity axiom $\forall x\, x=x$ we have $\forall x\exists y\, y=x$, and so by universal instantiation it follows that $\exists y\, y=c$. However, by necessitation it follows that $\Box\exists y\, y=c$, i.e., that *c*, whatever it may be, is a necessary being. Thus, in Kripke's system as it stands, one cannot consistently assert, e.g., that Socrates is a contingent being, $\neg\Box\exists y\, y=s$, surely a seriously undesirable feature in a logic that is intended properly to capture our modal intuitions.

Alarming as this problem might be, it is more a formal rather than a philosophical objection to Kripke's system. Though Kripke himself might not be particularly pleased at the prospect, it seems that the proper response to these problems is simply to alter those features of classical quantification theory and/or classical propositional modal logic that give rise to invalid inferences such as the above. (Arguably, Kripke has already made a similar move in adopting the generality interpretation of free variables.) Obvious suspects here are universal instantiation and necessitation. After all, there is nothing sacrosanct about either classical quantification theory or classical modal logic. If they are inconsistent with strong modal intuitions, then their revision is required and fully warranted.

So its current inability to name contingent beings does not of itself constitute much of an objection to

Kripke's system. It is likely that it could be patched up so as to allow it this expressive capacity. Far more serious is the fact that, despite the invalidity and unprovability of the actualistically objectionable principles BF, NE, and CBF, Kripke's system does not appear to have escaped ontological commitment to possibilia. A model theory provides a *semantics* for a language -- an account of how the truth value of a given sentence of the language is determined in a model by the meanings of its semantically significant component parts, notably, the meanings of its names, predicates, and quantifiers. Now, truth-in-a-model is not the same as truth simpliciter. However, truth simpliciter is usually understood simply to be truth in an *intended* model, a model consisting of the very things that the language is intuitively understood to be "about". So if we are to take Kripke models seriously as an account of truth for modal languages, then we must identify the intended models of those languages. And for this there seems little option but to take Kripke's talk of possible worlds literally: the set W in an intended Kripke model is the set of all possible worlds. If so, however, it appears that Kripke is committed to possibilia. For suppose the modal operators are literally quantifiers over possible worlds. And suppose it is possible that there be objects -- Aliens, for example -- that are distinct from all actually existing objects. Letting '*A*' express the property of being an Alien, we can represent this proposition by means of the sentence '$\Diamond\exists yAy$ & $\neg\exists x\Diamond Ax$', i.e., while there could be Aliens ($\Diamond\exists yAy$), no actually existing thing could be an Alien ($\neg\exists x\Diamond Ax$). On Kripke's semantics, the first conjunct of this sentence can be true only if there is a possible world *w* and an object *a* such that *a* is an Alien at *w*. But given the second conjunct, any such object *a* is distinct from all actually existing things. Hence, using Kripke's semantics to provide us with an account of truth, we find ourselves quantifying directly over possible worlds and mere *possibilia*. That BF, NE, and CBF are unprovable in Kripke's system, it seems, is metaphysically irrelevant. For it appears that, nonetheless, the semantics itself is wholly committed to possibilism.

An option for the actualist here, perhaps, is simply to deny that Kripke models have any genuine metaphysical bite. The real prize is the logic, which describes the modal facts of the matter *directly*. The model theory is simply a formal *instrument* that enables us to prove that the logic possesses certain desirable metatheoretic features, notably consistency. But this position is unsatisfying at best. Consider ordinary "Tarskian" model theory for nonmodal first-order logic. Intuitively, this model theory is more than just a formal artifact. Rather, when one constructs an intended model for a given language, it shows clearly how the semantic values of the relevant parts of a sentence of first-order logic -- the objects, properties, relations, etc. in the world those parts signify -- contribute to the actual truth value of the sentence. The semantics provides insight into the "word-world" connection that explains how it is that sentences can express truth and falsity, how they can carry good and bad information. The embarrassing question for the actualist who would adopt the proposed instrumentalist view of Kripke semantics is: what distinguishes Kripkean model theory from Tarskian? Why does the latter yield insight into the word-world connection and not the former? Distaste for the metaphysical consequences of Kripke semantics at best provides a motivation for finding an answer to these questions, but it is not itself an answer. The actualist owes us either an explanation of how Kripke's model theory provides a semantics for modal languages that does not commit us to possibilism, or else he owes us a semantical alternative.

§5: Actualist Responses to the Possibilist Challenge

We now turn to the work of actualists who have tried to address the possibilist challenge.

Individual Essences

One of the best known responses to the possibilist challenge was developed by Alvin Plantinga [1974]. The heart of Plantinga's approach is the notion of an *individual essence*. Plantinga's precise definition of this notion is a bit complex, but the idea itself is quite simple. Consider first the venerable distinction between *essential* and *accidental* properties. Intuitively, the essential properties of an object are those properties that make the object "what it is." More exactly, they are the properties that the object couldn't possibly have lacked. Its accidental properties, by contrast, are those that it just happens to have but might well have lacked. Thus, the property **being a horse** is intuitively not a property that the champion racehorse Secretariat could have lacked; he couldn't have been a rabbit, say, or a stone. The property **being a horse** is thus essential to Secretariat. By contrast, Secretariat could easily have lacked the property **being a racehorse**. Under different circumstaces -- if, say, he'd injured a leg as a colt -- he might have spent his days frolicking in the fields. That property is therefore accidental to Secretariat. So the first part of the definition of an individual essence -- the "essence" part -- is that an individual essence is an essential property of anything that has it. And the "individual" part of the definition is simply that if something has a given individual essence, then nothing else could possibly have that same individual essence. (We provide the definition Plantinga actually uses in the document Plantinga's Definition of an Individual Essence.)

Examples of individual essences are a little harder to come by than examples of essential properties. There are fairly strong intuitive grounds for the thesis that having arisen from the exact sperm and egg that one has is an individual essence of every human person, or at least of every human body. A different sperm and the same egg, say, would have resulted in a perhaps similar but numerically distinct person. Less controversial from a purely logical standpoint are what Plantinga calls *haecceities*, i.e., properties like **being Plantinga**, or perhaps, **being identical with Plantinga**, that are "directly about" some particular object. Pretty clearly, Plantinga has the property **being Plantinga** essentially -- he could not exist and lack it; any world in which he exists is, *ex hypothesi*, a world in which he is *Plantinga*, and hence a world in which he exhibits the property in question. Moreover, nothing but the individual Plantinga could have had that property; necessarily, anything that has it is identical to Plantinga. Hence, **being Plantinga** is an individual essence. Importantly, Plantinga takes individual essences, like all properties, to exist necessarily, even if they are not exemplified. (Interested readers may wish to read the supplementary document Background Assumptions for Plantinga's Account.)

Briefly put, Plantinga's solution to the possibilist challenge is to replace the possibilia of Kripke's semantics with individual essences. We follow the development of this solution found in Jager [1982]. Specifically, an interpretation **I** of the first-order modal language **L**, consists again of two mutually disjoint nonempty sets: the set of possible worlds and the set of individual essences. And, as with Kripke, there is a function **dom** that assigns to each possible world **w** its own distinct domain **dom**(**w**). However, instead of the possible individuals that exist in **w**, this domain consists of those individual essences that

are *exemplified* in **w**, or, more exactly, that *would have been exemplified* if **w** had been actual.

But how, exactly, does **I** assign values to predicates? After all, it is not individual essences to which predicates apply at worlds, it is the things that exemplify them; **being an Alien**, if it were exemplified, would not be a property of essences, but of individuals. Plantinga's trick is to talk, not about exemplification, but *coexemplification*. Properties **P** and **Q** are *coexemplified* just in case some individual has both **P** and **Q**. And for any world **w**, **P** and **Q** are *coexemplified in* ***w*** just in case, if **w** were actual, **P** and **Q** would be coexemplified. So, for example, given that there are men who are philosophers, the properties **being a man** and **being a philosopher** are coexemplified in the actual world. Again, any world in which the pope (i.e., Wojtyla) has a child is a world in which the property **being a child of Wojtyla** is coexemplified with an individual essence *E*; any such essence *E*, of course, is unexemplified in the actual world (assuming the lifelong chastity of the pope).

A relation **R** is coexemplified with properties $\mathbf{P}_1,\ldots, \mathbf{P}_n$ (in that order) just in case (i) there are individuals $\mathbf{i}_1,\ldots, \mathbf{i}_n$ that exemplify $\mathbf{P}_1,\ldots, \mathbf{P}_n$, respectively, and (ii) $\mathbf{i}_1,\ldots, \mathbf{i}_n$ stand in the relation **R**. And for any world **w**, **R** is coexemplified with properties $\mathbf{P}_1,\ldots, \mathbf{P}_n$ in **w** just in case, if **w** were actual, **R** would be coexemplified with $\mathbf{P}_1,\ldots, \mathbf{P}_n$. In Plantinga's system, then, a 1-place predicate *P applies to* a given individual essence **e** at a world **w** just in case the property expressed by *P* is coexemplified with the **e** at **w**. And an *n*-place predicate *R* applies to essences $\mathbf{e}_1,\ldots, \mathbf{e}_n$ at **w** just in case the relation expressed by *R* is coexemplified with $\mathbf{e}_1,\ldots, \mathbf{e}_n$ at **w**. For any individual essences $\mathbf{e}_1,\ldots, \mathbf{e}_\mathrm{n}$ and possible world **w** in our interpretation **I**, then, **I** specifies, for each *n*-place predicate *R*, whether or not *R* applies to $\mathbf{e}_1,\ldots, \mathbf{e}_n$ at **w**.

The denotation function **f** for **I** works just as in SQML and Kripke semantics, only now, of course, it assigns essences to variables instead of possibilia. Given this, we can now illustrate the definition of truth for this model theory by means of several instances of its most important clauses:

1. The open, atomic formula '*Px*' is $\text{true}_{\mathbf{I},\mathbf{f}}$ at **w** just in case **I** specifies that '*P*' applies to $\mathbf{d}_{\mathbf{I},\mathbf{f}}(x)$ at **w**.
2. The quantified formula '$\forall xPx$' is $\text{true}_{\mathbf{I},\mathbf{f}}$ at **w** just in case, for all individual essences **e** in **dom**(**w**), '*Px*' is $\text{true}_{\mathbf{I},\mathbf{f}[x,\mathbf{e}]}$ at **w**, where **f**[*x*,**e**] is **f** if **f**(*x*) = **e**, and otherwise is just like **f** except that it assigns **e** to *x* instead of **f**(*x*). (A little less formally, '$\forall xPx$' is $\text{true}_{\mathbf{I},\mathbf{f}}$ at **w** just in case, for all individual essences **e**, the predicate '*P*' applies to **e** at **w**.)
3. The open, modal formula '$\Box\varphi$' ('$\Diamond\varphi$') is $\text{true}_{\mathbf{I},\mathbf{f}}$ at **w** just in case for every (some) possible world **w**, 'φ' is $\text{true}_{\mathbf{I},\mathbf{f}}$ at **w**.

Referring back to our Alien example, then, the proposition that it is possible that there are Aliens, $\Diamond\exists xAx$, is true on this account if and only if there is a possible world **w** and a haecceity **e** such that '*A*' applies to **e** at **w**, i.e., if and only if the property **being an Alien** and **e** are coexemplified in **w**.

Notice that Plantinga's account also has no problem dealing with iterated modalities. The problem, recall, was that sentences like

(9) $\Diamond\exists x(Sxp \& \Diamond Px)$

appear to require mere possibilia to serve as the values of the quantifier, since no actually existing thing could be a son of the current pope. For Plantinga, the solution is simply that quantifiers range over haecceities, and that, in particular, (9) is true in virtue of their being an unexemplified haecceity which, in some possible world is coexemplified with the property **being a son of Wojtyla**, and in another world, that very same haecceity is coexemplified with **being a priest**.

In sum, then, in Plantinga's account there is an individual essence for every *possibile* in Kripke's. And for every property that every *possibile* enjoys at any given world **w** in Kripke's account, there is an individual essence that is coexemplified with that property in (the Plantingian counterpart of) **w**. Plantinga's semantics would thus appear to generate precisely the same truth values for the sentences of a modal language as Kripke's. Hence, it would appear that Plantinga has indeed successfully developed a semantics for modal languages that comports with actualist scruples.

Problems with this Account

Objections to Plantinga's account of actualism are addressed in the document Problems with the Actualist Accounts.

World Stories

Many propositions are *singular* in form. That is, as opposed to general propositions like **All men are animals** and **There are Aliens**, some propositions are, in the words of Arthur Prior, "directly about" specific individuals -- for example, the proposition **Winston Churchill was a German citizen**. Such propositions are typically expressed by means of sentences involving names, pronouns, indexicals, or other devices of direct reference. As we've seen, possibilists believe that there are singular possibilities (i.e., singular propositions that are possibly true) about things that don't actually exist, possibilities involving mere *possibilia*. Similarly, haecceitists also believe that there are singular possibilities that are, in a certain clear sense, directly about things that don't exist, viz., possibilities that, were they actual, would involve the exemplification of haecceities that are in fact actually unexemplified. Say that a *strong* actualist is someone who rejects both nonactual *possiblia* and unexemplified haecceities. For the strong actualist, then, there are no singular propositions directly about things that do not actually exist. Since this is a necessary truth for the strong actualist, it also follows that, had some actually existing individual failed to exist, there have been no singular propositions about that individual. Singular propositions about contingent beings are thus themselves likewise contingent for the strong actualist. A strong actualist, then, as we might put it, believes that all possibilities are either wholly general, or at most are directly about actually existing individuals only.

Several philosophers -- notably, Robert Adams and, building on work of Prior [1977], Kit Fine -- have developed possible world semantics that are strongly actualist. The approach in Adams [1974] centers

around Adams' notion of a possible world, or "world story". For Adams, a world story is a *maximally possible* set of propositions, that is, a set **s** of propositions such that (i) for any proposition **p**, s contains either **p** or its negation ¬**p**, and (ii) it is possible that all the memebers of **s** be true together.[5]. A proposition **p** is *true* in a world story **w**, then, just in case **p** is a member of **w**. Thus, Adams takes a proposition to be possible just in case it is true in some world story. In particular, then, the semantics of our paradigmatic proposition **Possibly, there are Aliens**, i.e., formally,

(1) $\Diamond\exists xAx$,

is straightforward and, on the face of it, innocuous from a strongly actualist perspective: (1) is true if and only if (the proposition expressed by)

(14) '$\exists xAx$' (i.e., the proposition **There are Aliens**) is true at some world.

Problems with this Account

Objections to world stories are addressed in the document Problems with the Actualist Accounts.

World Propositions

Inspired by the work of Arthur Prior, Kit Fine [1977] has developed an approach similar to Adams' which takes account of the contingency of singular propositions. For Adams, a possible world is a certain set of propositions. Fine, by contrast, drawing on an idea of Prior's, identifies a world with a certain type of proposition --- what he calls a *world proposition*. Roughly speaking, a world proposition might be thought of as the infinite conjunction of all of the proposition in one of Adams' world stories. More specifically, a world proposition is a proposition **q** such that it is possible both that **q** be true and that it entail all true propositions. A proposition **p** is then said to be *true in* a possible world, i.e., world proposition, **q** just in case **q** *entails* **p**, i.e., formally, just in case $\Box(\mathbf{q} \rightarrow \mathbf{p})$.

Now, if Fine's account were to parallel Adams', then Fine would now say that, for a proposition to be possible is for it to be true in some possible world, i.e., to be entailed by some world proposition. But Fine is more mindful of the problems that contingent propositions raise for actualism. Consequently, he suggests alternative truth conditions for propositions of the form **Possibly p**, namely, that it is *possible* that **p** be true in some possible world. Thus, for Fine, the full analysis of the iterated modal proposition (9) -- $\Diamond\exists x(Sxp \,\&\, \Diamond Px)$ -- is as follows: (9) is true if and only if

(18) It is possible that '$\exists x(Sxp \,\&\, \Diamond Px)$' (i.e., the proposition **Wojtyla has a son who could have become a priest**) is true at some world **w**.

(18), in turn, is true if and only if

(19) It is possible that, for some some individual *x*, '*Sxp* & ◇*Px*' (i.e., the proposition ***x* is a son of Wojtyla and *x* could have become a priest**) is true at some world **w**,

which, in turn, is true if and only if

(20) It is possible that, for some some individual *x*, '*Sxp*' (i.e., the proposition ***x* is a son of Wojtyla**) is true at some world **w** and, it is possible that, for some world **u**, '*Px*' (i.e., the proposition ***x* is a priest**) is true at **u**.

Thus, all that Fine's account requires in its analyses of (9) is the *possibility* that certain propositions exist -- notably, singular propositions (and, in particular, world propositions) that don't exist in fact but would exist if certain individuals did, as would be the case, e.g., if the pope were to have a son. Unlike Adams account, then, Fine's wears its fully intensional character on its sleeve. It abandons the idea that ordinary modal operators such as "possibly" and "necessarily" can, in general, be analyzed as extensional quantifiers over possible worlds. For some occurrences of those operators -- those in (9), for instance -- are ineliminable.

Problems with this Account

Objections to world propositions are addressed in the document Problems with the Actualist Accounts.

Roles

McMichael [1983a] has proposed an actualist semantics that avoids the objections to haecceities raised against Plantinga and both the loss of compositionality objection and the iterated modalities objection raised against Adams. Like Adams, McMichael rejects haecceities. However, like Plantinga, McMichael introduces a class of actualist surrogates for *possibilia*, which he calls *roles*. McMichael's account builds on a very rich and elaborate theory of relations, and it is necessary to lay out at least some of its basic concepts in order to understand the account.

For McMichael, a primitive logical relation of *inclusion* can hold between properties and relations. Because it is a primitive, it cannot be defined, but, intuitively, in the case of properties, the idea is that one property **P** includes another **Q** just in case, necessarily, anything that has **P** has **Q**. Thus, the property **being red** includes the property **being colored**. Again, intutively once again, one binary relation **R** includes another **R′** just in case, necessarily, for any objects **x** and **y**, if **x** bears **R** to **y**, then **x** bears **R′** to **y**. So, for example,the conjunctive relation **being both a child and an heir of** includes the relation **being a child of**.

Inclusion can also hold between an *n*+1-place relation and an *n*-place relation, relative to one of the argument places of the former.[6] Thus, in particular, a 2-place relation **Q** can include a property **P**, relative to one of its two argument places: intuitively, **Q** includes **P**, relative to its first argument place, if

and only if, necessarily, whenever two things **a** and **b** stand in the relation **Q**, **a** exemplifies **P**. And if the inclusion were with respect to the second argument place, of course, it would be **b** that exemplifies **P**. So, for example, the **child-of** relation, relative to its first argument place, includes the property **being a child of something**; whenever any object **a** bears the **child-of** relation to some object **b**, **a** has the existentially quantified property **being a child of something**. Similarly, **child-of** includes the property **being a parent of something** relative to its second argument place.

A (unary) role is just a "purely qualitative" property of a certain sort, where (as described in more detail in the supplementary document on Qualitative Essences) a purely qualitative property is a property that "involves" no particular individuals. Thus, such properties as **being a philosopher** and **being someone's mother or maternal aunt** are purely qualitative, while **being a student of Quine** and **being Johnson's mother or a friend of Boswell** are not. Given this, McMichael defines a property **P** to be a *unary role* if (i) it is exemplifiable, (ii) it is purely qualitative, and (iii) for any purely qualitative property **Q**, either **P** includes **Q** or **P** includes the complement **-Q** of **Q**. A role is thus a complete (nonmodal) "characterization" of the way something could be, qualitatively. Intuitively, then, the role of a given object is a "conjunction" of all of the purely qualitative, nonconjunctive properties the object exemplifies. Thus, for example, Socrates' role includes the properties **being a philosopher**, **being snub-nosed**, **being the most famous teacher of a famous philosopher**, **being condemned to death** and so on.[7] The notion of role generalizes in a natural way to all *n*-place relations, including, notably, propositions (i.e., 0-place relations) and binary (i.e., 2-place) relations. Thus, the binary role that Boswell bears to Johnson is, intuitively, a conjunction of all of the purely qualitative, nonconjunctive binary relations that Boswell bears to Johnson.[8] As one might suspect, it can be shown on McMichael's theory that a binary role includes a unique unary role with respect to each of its argument places. In particular, the binary role that Boswell bears to Johnson includes Boswell's unary role relative to its first argument place and Johnson's relative to its second.

Now (as also explained in the supplementary document on Qualitative Essences), Adams [1979] has argued persuasively that no purely qualitative property, no matter how complex, can serve as an individual essence for a contingent being. Hence, in general, roles -- being purely qualitative -- are not individual essences. Rather, they are general properties that are (in general) exemplifiable by different things (though not necessarily things in the same possible world). Because of this, none of the objections to Plantinga's haecceities is relevant to roles, as the fact that haecceities are individual essences lies at the heart of those objections. At the same time, McMichael is able to provide a semantics for (9) that does not run afoul of the iterated modalities objection. The basic trick is to

> ...alter the criterion for deciding what an individual might have done. Instead of saying that what an individual might have done is what *he* does in some possible world, let us say that what an individual might have done is what *any such* individual does in some possible world....To determine what Socrates might have done, we don't look for worlds in which he appears, but instead we look for roles in possible worlds which are accessible to Socrates' actual role. If one of these roles includes a certain property, then that property is one Socrates could have had; otherwise, it is not [ibid, 73].

Thus, a little more formally, where F is the property **being foolish**, and s is Socrates, a simple modal sentence such as

(21) Possibly, Socrates is foolish ($\Diamond Fs$)

is true just in case some unary role accessible to the actual role of Socrates includes the property of being foolish.

Similar to Plantinga's semantics, then, quantifiers do not range over individuals, but over roles. This enables McMichael to avoid the iterated modalities objection and provide a compositional semantics for our iterative paradigm (9). Specifically, (9) is true if and only if

(22) Some role **R** accessible to Wojtyla's actual role $\mathbf{R_k}$ includes the property **being a parent of someone** (i.e., the property [$\lambda y\ \exists xCxy$] expressed by the open formula ‘$\exists xCxy$’).

(22) captures the idea that an individual *such as* Wojtyla could have been a parent. Adams, of course, got this far in his account of (9). But, unlike Adams, with roles at his disposal, McMichael can continue his analysis of (9) and unpack the existentially quantified formula ‘$\exists xCxy$’. Specifically, (22) holds if and only if

(23) Some binary role **S** that includes the **child-of** relation (i.e., the relation expressed by the atomic formula ‘Cxy’) also includes, relative to its second argument place, the role **R** (a role accessible to Wojtyla's actual role $\mathbf{R_k}$).

That is, in accordance with McMichael's recursive definition of truth, (23) unpacks the quantified formula ‘$\exists xCxy$’ in terms of the **child-of** relation that is expressed by the embedded atomic formula ‘Cxy’. Specifically, the truth of (9) consists in the existence of a binary role **S** that includes the **child-of** relation and, relative to its second argument place, a role accessible to Wojtyla's role $\mathbf{R_k}$. Note that, being a binary role, **S** also includes a unique role relative to its first argument place. And because it includes the **child-of** relation and, relative to its second argument place, a role **R** accessible to Wojtyla's role $\mathbf{R_k}$, **S** will include, relative to its first argument place, a role $\mathbf{R'}$ that can only be exemplified by a child of whatever exemplifies **R**, i.e., a child of *such* an object as Wojtyla.

To capture the intuition that no such child could be identical to any actually existing thing, then, McMichael can simply deny that the role $\mathbf{R'}$ that would be exemplified by such a child is accessible to the role of any actually existing thing.

Problems with this Account

Objections to McMichael's role theory are addressed in the document Problems with the Actualist Accounts.

Dispensing with Worlds

A rather different approach to the possibilist challenge is broached by Menzel [1990]. This approach is clarified and refined by Ray [1995], and a very similar approach is elaborated in great formal and philosophical detail by Chihara [1998]. For ease of reference, call this the "no-worlds" approach.

All non-skeptical approaches to modality agree that Kripke models provide key insights into the meaning of our modal discourse and the nature of modal reality. However, as we have seen, the naive "intended" model of Kripke semantics leads to possibilism. The standard actualist response -- following David Lewis, "ersatzism" -- has been to define actualistically acceptable notions of possible worlds and possible individuals to serve as replacements for the elements of **W** and **D** in the naive intended model, thereby (or so it is argued) preserving the semantical and metaphysical benefits of Kripke models while avoiding ontological commitment to *possibilia*. As just seen above, however, ersatzism is still problematic. By contrast, the no-worlds account does not attempt to identify worlds as acceptable abstract entities of some sort. Rather, the notion of a possible world is abandoned altogether.

To get at the idea, note first that the notion of an intended Tarski model makes perfectly good sense for a formalization of nonmodal discourse about the actual world. To illustrate, suppose we have a given a piece of nonmodal discourse about a certain event, a baseball game, say. Suppose now we formalize that discourse in a nonmodal language L'; that is, for each referring expression in the disourse (e.g., ‘Mark McGwire’, ‘second inning’, etc.), there is a unique constant of L', and for every simple verb phrase in the discourse (‘is a home run’, ‘is out’, ‘relieves’, etc.) there is a unique predicate of L'. Then we can form a Tarski model $\mathbf{M}_{L'}$ for L' whose domain consists of the actual objects that the speakers are talking about in the discourse (fans, players, equipment, etc.) and which interprets the predicates of L' so that they are true of exactly those (*n*-tuples of objects in the domain that are in the extension of the corresponding verb phrases of the discourse. In this way we form the intended model of L', the piece of the world that it is intended to be about.

According to proponents of the no-worlds account, the fallacy of both ersatzism and possibilism is the inference that things must work in largely the same way with regard to Kripke models. A Kripke model is basically an indexed collection of Tarski models. Just as there is an intended Tarski model for a nonmodal language L' constructed from the actual world, the accounts above infer that, for a given formalization L of modal discourse, there must be an intended Kripke model constructed from all possible worlds. And, depending on one's tolerance for possibilia, this leads either to possibilism or one of its ersatz variations.

For no-worlders, the modal upshot of a Kripke model lies in its structure rather than its content. The specific elements of a Kripke model are irrelevant. Rather, under appropriate conditions, it is the form of a Kripke model alone that tells us something about modal reality. Specifically, the model theory of Kripke semantics is retained in the no-world account. The elements of a model are irrelevant; it is easiest

just to take them to be pure sets, or ordinal numbers, or some other type of familiar mathematical object. Consequently, there can be no notion of a single intended model, because, for every model, there are infinitely many others that are structurally isomorphic to it, and structure is all that matters on the no-worlds account. In place of intended models, the no-worlds account offers the notion of an *intended** model. To get at the idea, suppose one has an intended Tarski model **M** of the actual world, a model that actually contains entities in the world and assigns extensions to predicates that reflect the actual meanings of the adjectives and verb phrases those predicates formalize. Now replace the objects in that model with abstract objects of some ilk to obtain a new model $\mathbf{M'}$ that is structurally isomorphic to **M**. Then $\mathbf{M'}$ also models the world under a mapping, or embedding, that takes each element $\mathbf{e'}$ of $\mathbf{M'}$ to the element **e** that it replaced in **M**. We can thus justifiably think of $\mathbf{M'}$ as a sort of intended model because, even though it doesn't necessary *contain* anything but pure sets (or some other type of mathematical object), under an appropriate embedding it models the actual world no less than **M**. To distinguish models like $\mathbf{M'}$ that require a nontrivial embedding into the world from models like **M**, we call the former *intended** models.

For no-worlders, the notion of an intended* Tarski model is all that is needed for modeling the modal facts. From these a notion of an intended* Kripke model can be defined. Assume that **L** is a model language that formalizes some range of modal discourse about the world. Roughly, then, an intended* Kripke model **M** is simply a Kripke model such that (i) the Tarski model indexed by the distinguished index $\mathbf{w_0}$ is an intended* Tarski model of the actual world, (ii) every Tarski model $\mathbf{M'}$ in **M** *could have been* an intended* model of the world, that is, the world could have been as $\mathbf{M'}$ represents it, and (iii) necessarily, some Tarski model in **M** is an intended* model of the world, i.e., no matter how the world had been, there would have been an intended* Tarski model of it in **M**.[9]

Truth in a model on the the no-worlds account is defined as usual as truth at the distinguish index $\mathbf{w_0}$ of the model, hence, a sentence φ is true simpliciter if and only if it is true at the distinguished index of some (hence, any) intended* Kripke model. Given the definition of an intended* model, it follows that a modal formula '$\Diamond\varphi$' is true if and only if, for some intended* Kripke-model **M**, φ is true in some Tarski-model $\mathbf{M'}$ in **M**, that is, in some Tarski-model $\mathbf{M'}$ in **M** that could have been an intended* model of the world. Thus, in particular, (1) is true if and only if

> (20) For some intended* Kripke model **M**, there is a Tarski model $\mathbf{M'}$ in **M** in which '$\exists xAx$' is true,

that is, given the definition of an intended* Kripke model, if and only if there is a Tarski model $\mathbf{M'}$ that could have been an intended* Tarski model and such that, if it had been, there would have been Aliens.

For the no-worlder, then, intended* Kripke models adequately represent the modal structure of the world simply by virtue of their own modal properties. Since Kripke models are constructed entirely out of existing objects, the semantics for modal logic requires no distinction between what is actual and what is possible. It therefore conforms with the thesis of actualism, but does so without the elaborate metaphysical apparatus of the ersatzers.

Problems with this Account

Objections to the no-world account are addressed in the document Problems with the Actualist Accounts.

An Actualist Interpretation of the Simplest QML

Finally, we consider to a new form of actualism that has been proposed recently. This form of actualism refocuses our attention on the Simplest Quantified Modal Logic (SQML) and offers a way to reinterpret this formalism to eliminate its apparent commitment to *possibilia*. On this form of actualism, the truth conditions for the modal claim "There might have been Aliens" are just what they appear to be, namely, that in some possible world, there is an object that is an Alien at that world. However, these truth conditions do not commit us to possibilia. Instead, the new form of actualism is based on the idea that these truth conditions are committed only to the existence of *nonconcrete* objects which might have been Aliens. This theory has been recently put forward by Linsky and Zalta ([1994] and [1996]) and Williamson ([1998] and [1999]), though Williamson simply eschews the word ‘actual’ in his formulation of the theory. These philosophers claim that the nonconcrete objects in question are not Aliens, but instead have the modal property of possibly being an Alien. In other words, modal claims such as "There might have been Aliens" (formalized, once again, as (1)) can be interpreted to be true in virtue of the *actual* existence of objects that are nonconcrete (and hence which are not Aliens) at our world, but which are Aliens (and hence concrete) at some other possible world. Thus, the nonconcrete objects involved in the truth conditions of such modal claims do exist and are actual. To say this is to use a sense of "existence" and "actual" similar to that used by Platonists when they claim that mathematical objects exist and are actual. However, unlike mathematical objects, which are nonconcrete at *every* possible world, the actual objects required by the truth of modal claims are only *contingently* nonconcrete -- they are nonconcrete at our world but concrete at other possible worlds. Similarly, ordinary concrete objects (like the rocks, tables, planets, etc., of our world) are assumed to be contingently concrete--they are concrete at some worlds (including ours) and not at other worlds.

With this basic idea in hand, the ‘new actualists’ point out that our ordinary modal claims can be given a straightforward analysis by: (1) regimenting ordinary modal discourse in the simplest possible way using the language and logic of the simplest quantified modal logic (SQML) and (2) semantically interpreting SQML by appealing to contingently concrete objects and contingently nonconcrete objects (both of which are assumed to actually exist). This interpretation, the new actualist argues, reveals that the problematic theorems of SQML -- most notably, the Barcan Formula (BF), the Converse Barcan Formula (CBF), and the Necessary Existence (NE) theorems -- do not contradict our modal intuitions, once those intuitions are understood in terms of a more subtle conception of the abstract/concrete distinction.

To see why, reconsider the definition and discussion of SQML and reexamine BF. From the fact that there might have been aliens ($\Diamond\exists xAx$), BF requires only that there exist something that *could have been* an Alien ($\exists x\Diamond Ax$). But, it was asked, doesn't this contradict the intuition (described at the very outset) that *nothing could have been an Alien*? Here, the new actualist argues that this intuition is true only when it is properly understood as the intuition that *no concrete object could have been an Alien*. Recall our

thought experiment at the outset, which asked us to "imagine a race of beings that is very different from any life-form that actually exists anywhere in the universe; different enough, in fact, that no actually existing thing could have been an Alien, any more than a given gorilla could have been a fruitfly." The relevant intuition, that nothing could have been an Alien, is grounded in the fact that when we look around us and examine all the concrete objects that there are, we note that none of those objects could have been an Alien (just as no gorilla could have been a fruitfly). However, this leaves room to claim that there exist (contingently) *non*concrete objects which could have been Aliens. According to the new actualist, these contingently nonconcrete objects have been overlooked because (1) no one has correctly drawn the proper distinction between contingently nonconcrete and necessarily nonconcrete objects, and (2) everyone has assumed that concreteness was an *essential* property of concrete objects (see below).

Thus, according to the new actualist, whenever there is a true claim of the form "There might have been something which is *F*", BF doesn't imply anything that is incompatible with our modal intuitions. For the conclusion that it forces, namely, that "There exists something that might have been an *F*", does not require us to suppose that there is some concrete object which might have been *F*. BF need only require the existence of contingently nonconcrete objects which might have been *F*. Similar reasoning is developed in the supplementary document Why CBF and NE Are Harmless Consequences of SQML.

Once it is seen that BF requires only contingently nonconcrete objects and not possibilia, it is natural to reconceive the nature of concrete objects. According to new actualism, ordinary concrete objects are concrete at some worlds and not at others. This is the sense in which they are contingent objects. Traditional actualists have described worlds where these objects are not concrete as worlds where these objects don't exist or have any kind of being. By contrast, the new actualist just rests with their nonconcreteness at the world in question, and argues that that should suffice to account for our intuition that such objects "are not to be found" in such a world. Moreover, new actualists reconceive the idea of an "essential" property of a concrete object. Instead of saying that Socrates is essentially a person because he is a person in every possible world where he exists, new actualists say that he is essentially a person because he is a person in every world where he is concrete.

So by recognizing the existence of contingently nonconcrete objects and by reconceiving both the contingency of concrete objects and the notion of an essential property in what seem to be harmless ways, there appears to be a way to interpret SQML so that it is consistent with actualism. Specifically, in the "intended" new actualist model, everything in the one domain **D** both exists and is actual. **D** includes: (1) (contingently) concrete objects, (2) necessarily concrete objects (if there are such), (3) contingently nonconcrete objects, and (4) necessarily nonconcrete objects (if there are such). All of these objects in **D** are said to actually exist.

New actualism appears to lack the awkward features that plague other forms of actualism. In contrast to Kripke's System, the metalanguage does not quantify over possibilia and object language quantifiers can range over everything the metalanguage quantifiers range over. In contrast to Prior's approach, no distinction between two kinds of necessity is needed. In contrast to Plantinga's haecceitism, there is an objectual interpretation of quantified modal logic which is expressible in terms of the basic notion of an individual exemplifying properties (rather than in the less basic terminology of coexemplification). In

contrast to Adams' world story approach, there is no puzzle arising from propositions which do not exist at worlds where their constituents do not exist--propositions and their constituents exist necessarily, though the contingent constituents of propositions fail to be concrete at some worlds. In particular, all objects exist necessarily for the new actualist, and hence can be quantified over relative to any possible world, it should be clear that the new actualist has no trouble with the semantics of iterated modalities. (Details are provided in the supplementary document New Actualism and Iterated Modalities for the interested reader.) In contrast to Fine's world propositions approach, object language modal operators are fully interpreted in terms of worlds alone, and hence can be thought of as providing a genuine semantical analysis of the modal operators -- although, admittedly, this will depend on a conception of worlds that does not itself involve a primitive notion of propositional possibility. At the least, the account provides as much explanatory power as Fine's in a manner that is both ontologically simpler and more direct. In contrast to McMichael's role theory, modal truths such as "Socrates might have been a carpenter" are genuinely about Socrates. In contrast to the no-worlders, the idea that necessary truth is truth in all possible worlds is preserved. The intended interpretation has in its domain all of the objects that actually exist and extensions can be distributed to properties at the various worlds in just the way that is required by the modal facts. Modal discourse, then, is directly about an independent reality free of possibilia, and the relationship between the formal language and the intended model exactly mirrors the relationship between ordinary modal language and the reality that grounds modal truth.

Problems with this Account

Objections to new actualism are addressed in the document Problems with the Actualist Accounts.

Bibliography

- Adams, R. M., 1974, ‘Theories of Actuality’, *Nous* **8**: 211-31; reprinted in Loux [1979], pp. 190-209
- Adams, R. M., 1981, ‘Actualism and Thisness’, *Synthese* **49**: 3-41
- Adams, R. M., 1979, "Primitive Thisness and Primitive Identity," *Journal of Philosophy* **76**, 5-26
- Chihara, C., 1998, *The Worlds of Possibility*, Oxford: Clarendon Press
- Chisholm, R., 1976, *Persons and Objects*, La Salle: Open Court
- Deutsch, H., 1990, ‘Contingency and Modal Logic’, *Philosophical Studies***60**: 89-102
- Deutsch, H., 1993, ‘Logic for Contingent Beings’, *Journal of Philosophical Research*, forthcoming
- Field, H., 1989, *Realism, Mathematics and Modality*, Oxford: Blackwell
- Fine, K., 1977, ‘Postscript’, in Prior [1977]
- Fine, K., 1978, ‘Model Theory for Modal Logics: Part I-The *De Re*/*De Dicto* Distinction’, *Journal of Philosophical Logic* **7**: 125-56
- Fine, K., 1981, ‘Model Theory for Modal Logic-Part III: Existence and Predication’, *Journal of Philosophical Logic* **10**: 293-307
- Forbes, G., 1985, *The Metaphysics of Modality*, Oxford: Clarendon Press
- Frege, G., 1980, *Philosophical and Mathematical Correspondence*, Chicago: The University of

Chicago Press

- Garson, J., 1984, ‘Quantification in Modal Logic’, in *Handbook of Philosophical Logic: Volume II*, D. Gabbay and F. Guenthner (eds.), Dordrecht: D. Reidel
- Gödel, K., 1931, ‘On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I, in J. van Heijenoort (ed.), *From Frege to Gödel*, Cambridge, Massachusetts: Harvard University Press, 1981.
- Hughes, G. and Cresswell, M., 1968, *An Introduction to Modal Logic*, London: Methuen
- van Inwagen, P., 1986, ‘Two Concepts of Possible Worlds’, in *Midwest Studies in Philosophy, XI*, P. French, T. Uehling, and H. Wettstein (eds.), Minneapolis: University of Minnesota Press, 185-213
- Jager, T., 1982, ‘An Actualist Semantics for Quantified Modal Logic’, *Notre Dame Journal of Formal Logic* **23**/3 (July): 335-49
- Kaplan, D., 1975, ‘How to Russell a Frege-Church’, *Journal of Philosophy***72** 716-29; reprinted in Loux [1979], pp. 210-24
- Kripke, S., 1963, ‘Semantical Considerations on Modal Logic’, *Acta Philosophica Fennica* **16**: 83-94.
- Kripke, S., 1972, ‘Naming and Necessity’, Cambridge, Mass.: Harvard, 1980
- Lewis, D., 1986, *On The Plurality of Worlds*, Oxford: Blackwell
- Linsky, B., and Zalta, E., 1994, ‘In Defense of the Simplest Quantified Modal Logic’, *Philosophical Perspectives 8: Logic and Language*, J. Tomberlin (ed.), Atascadero: Ridgeview, pp. 431-458
- Linsky, B., and Zalta, E., 1996, ‘In Defense of the Contingently Concrete’, *Philosophical Studies* **84**: 283-294
- Loux, M., (ed.), 1979, *The Possible and the Actual*, Ithaca: Cornell
- Marcus, R. Barcan, 1986, ‘Possibilia and Possible Worlds’, *Grazer Philosophische Studien*, R. Haller (ed.), **25**/**26** (1985/1986): 107-33
- Marcus, R. Barcan, 1946, ‘A Functional Calculus of First Order Based on Strict Implication’, *Journal of Symbolic Logic***11**: 1-16
- McMichael, A., 1983a, ‘A New Actualist Modal Semantics’, *Journal of Philosophical Logic* **12**, 73-99
- McMichael, A., 1983b, ‘A Problem for Actualism about Possible Worlds’, *Philosophical Review* **92**, 49-66
- Meinong, A., 1904, ‘On Object Theory’, in *Realism and the Background of Phenomenology*, R. Chisholm (ed.), Glencoe: The Free Press, 1960; translation of ‘Über Gegenstandstheorie’, in *Untersuchungen zur Gegenstandstheorie und Psychologie*, A. Meinong (ed.), Leipzig: Barth, 1904
- Mendelson, E., 1964, *Introduction to Mathematical Logic*, New York: D. Van Nostrand
- Menzel, C., 1990, ‘Actualism, Ontological Commitment, and Possible Worlds Semantics’, *Synthese* **85**: 355-89
- Menzel, C., 1991, ‘The True Modal Logic’, *Journal of Philosophical Logic* **20**: 331-374
- Parsons, T., 1967, ‘Grades of Essentialism in Quantified Modal Logic’, *Nous* **1**/No. 2 (May): 181-91
- Parsons, T., 1969, ‘Essentialism and Quantified Modal Logic’, *The Philosophical Review* **78**/1 (January): 35-52

- Parsons, T., forthcoming, ‘An Experiment in Making Ontological Commitments’
- Plantinga, A., 1976, ‘Actualism and Possible Worlds’, *Theoria* **42**: 139-60; reprinted in Loux [1979], pp. 253-73
- Plantinga, A., 1974, *The Nature of Necessity*, Oxford: Oxford University Press
- Plantinga, A., 1983, ‘On Existentialism’, *Philosophical Studies* **44**: 1-20
- Plantinga, A., 1985, ‘Replies’, in *Alvin Plantinga*, J. Tomberlin and P. van Inwagen (eds.), Dordrecht: D. Reidel, pp. 313-96
- Pollock, J., 1985, ‘Plantinga on Possible Worlds’, in *Alvin Plantinga*, J. Tomberlin and P. van Inwagen (eds.), Dordrecht: D. Reidel, pp. 121-44
- Prior, A., 1977, *Worlds, Times, and Selves*, Amherst: University of Masssachusetts Press
- Prior, A., 1968, *Papers on Time and Tense*, Oxford: Clarendon
- Prior, A., 1967, *Past, Present, and Future*, Oxford: Clarendon
- Prior, A., 1957, *Time and Modality*, Oxford: Clarendon
- Prior, A., 1956, ‘Modality and Quantification in S5’, *Journal of Symbolic Logic* **21**: 60-2
- Quine, W. V. O., 1940, *Mathematical Logic*, Cambridge, MA: Harvard University Press; 2nd edition, revised, 1951
- Quine, W. V. O., 1948, ‘On What There Is’, in *From a Logical Point of View*, New York: Harper, 1953, 1-19
- Ray, G., 1996, ‘An Ontology-free Modal Semantics’, *Journal of Philosophical Logic* **25**: 333-361
- Russell, B., 1905, ‘On Denoting’, *Mind* **14** (October): 479-493
- Salmon, N., 1987, ‘Existence’, in *Philosophical Perspectives 1*, J. Tomberlin (ed.), Atascadero: Ridgeview Press
- Tomberlin, J., 1994, ‘Troubles with Actualism’, *Philosophical Perspectives 8: Logic and Language*, J. Tomberlin (ed.), Atascadero: Ridgeview
- Tomberlin, J., 1996, ‘Actualism or Possibilism?’, *Philosophical Studies* **84**: 263-281
- Tomberlin, J., and van Inwagen, P. (eds.), 1985, *Profiles: Alvin Plantinga*, Dordrecht: D. Reidel, pp. 121-44
- Williamson, T., 1998, ‘Bare Possibilia’, *Erkenntnis* **48**: 257-273
- Williamson, T., 1999, ‘Existence and Contingency’, *The Aristotelian Society: Supplementary Volume* **73**: 181-203; to be reprinted in the *Proceedings of the Aristotelian Society* **100**, forthcoming
- Zalta, E., 1993, ‘Twenty-Five Basic Theorems in Situation and World Theory’, *Journal of Philosophical Logic* **22**: 385-428
- Zalta, E., 1983, *Abstract Objects: An Introduction to Axiomatic Metaphysics*, Dordrecht: D. Reidel

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

logic: classical | logic: free | logic: modal | possible objects | possible worlds | Prior, Arthur | state of affairs

Christopher Menzel
cmenzel@tamu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Notes to Actualism

Notes

1. Some good sources for actualism are the following: Adams [1974], Plantinga [1976], Kaplan [1975], Loux [1979], and Tomberlin and van Inwagen [1985].

2. We assume here that, as with all natural kinds, being an Alien is an essential property of anything that has it.

3. A potential actualist move is worth addressing here. One might argue that, in fact, there is a straightforward actualist account of the possibility of Aliens. A widely-accepted contemporary metaphysical belief is that, for every set of objects, there exists the mereological sum consisting of exactly those objects. Given this, assuming that Aliens are composed of the same basic atomic stuff that we are, there are surely actual merological sums of atoms that are possible Aliens, i.e., that *could* have been Aliens if only they'd been properly arranged. Hence, there is no need to postulate possibilia to provide a semantics for claims like ‘It is possible that there are Aliens’. However, this objection misses the point. The general intuition that we are attempting to isolate with the Alien example is that

> (*) *There could have been things other than the things that actually exist*.

All that the actualist move just noted succeeds in showing is that perhaps the Alien example doesn't entail (*). But it does not succeed in accounting for the intuition that (*) is true. For suppose we accept the proposed mereological gambit, i.e., that certain mereological sums of actual atoms could have been Aliens, or instances of any other uninstantiated natural kind. Is it not still the case that there could have been *different atoms* (or quarks or whatever basic building blocks you choose) than there are in fact? Indeed, is it not logically possible that the universe could have been composed of entirely different stuff altogether? If so, then the actualist still needs to account for (*).

It should also be noted that the mereological gambit itself is dubious. Its basic premise -- that any collection of atoms constitutes a further physical object -- is far from uncontroversial. More seriously, it seems quite clear that no instance of a physical natural kind *is* identical with any given mereological sum of atoms, as physical bodies are constituted by many different sums of atoms across time as those bodies change. Perhaps, however, the actualist could come up with some more sophisticated mereological construct C to avoid this objection. Still, it seems, there are problems. For intuitively, it seems that the same C, structured one way, could have been an instance of one kind, and structured another, could have been an instance of a different kind. But then it seems to follow from the "modal transitivity" of identity (i.e., the principle:

$$\forall x \Box \forall y(y{=}x \rightarrow \Box \forall z(x{=}z \rightarrow y{=}z)))$$

that if a member of a natural kind is literally identical with a C, then it is possible that an instance of a given kind could have been an instance of a very different kind. But this conflicts with strong intuitions about the essentiality of kind membership. So even if the actualist's hypothetical construct C were plausible, rather than taking Cs to be actualist surrogates for possible Aliens (or whatever), it would be at least equally reasonable to claim that certain Cs are only possibly *co-located* with, or possibly *constitutive of*, but not possibly identical with, an Alien. For, in that case, all that follows is that the same C might have been co-located with (or constitutive of) instances of very different natural kinds, and intuitions about the essentiality of kind membership are preserved.

4. Unfortunately, for reasons rooted ultimately in the monumental work of Gödel [1931], a first-order logic cannot provide a completely *decidable* mechanism for determining validity. More exactly, while it is true that, if a formula is valid, one can eventually find a proof of it in the logic, there is in general no proof theoretic way to determine that a formula is *invalid*.

5. Adams [1974], p. 204. This is equivalent to the following, simpler definition: a world story is a set **s** of propositions such that it is possible that, for all propositions **p**, **s** contains **p** if and only if **p** is true. This account of world stories is significantly more accessible than the later account in Adams [1981]. The added subtleties of the later account are introduced to enable it to serve as a semantics for a broader range of modal statements, particular those involving contingent propositions, notably propositions about possible nonexistence. However, for the purposes of the present article, these subtleties add unnecessary complexity, as I believe that the 1981 account ultimately falls prey to essentially the same objections that are raised here against the earlier account.

6. McMichael does not actually use the idea of inclusion relative to an argument place. Rather, I have introduced it to simplify the presentation of the theory. It is an equivalent mechanism and so has no impact on the theory's content. McMichael's own account relies on an elegant, but conceptually more challenging permutation mechanism that shuffles argument places in relations.

7. For the sake of simplicity, we ignore temporal qualifications in these examples that would be needed in a fully accurate account, e.g., **being condemned to death as an adult**.

8. As indicated, order matters in our representation of relations: the binary role that Boswell bears to Johnson is distinct from the binary role that Johnson bears to Boswell, the latter, of course, being the converse of the former.

9. Things are a bit more subtle than this, for to have an intended* model, one also needs to do a little more to reflect the modal facts expressed by means of iterated modalities. See Menzel (1990) for details.

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

Three Types of Possibilism

A possibilist is someone who believes that there are things that are not actual. There are two ways to understand this claim. On the first, the possibilist distinguishes *what there is* from *what exists, or is actual*, and argues that the latter comprises a relatively small portion of the former. On this view, then, actual things exhibit a certain intrinsic, ontological property --- *existence*, or *actuality* --- which things that merely *are* happen to lack, things like Aliens, people that were never born, and so on. Such things *could* have exhibited this property, of course, and indeed would have had things been different, but, as it happens, they simply do not. This, then, on the first understanding of possibilism, is what the possibilist means when she says that there are things that are not actual.

The second way of understanding possibilism can be traced back to Quine [1948], who insisted that there is no coherent distinction between what there is and what exists. Thus, for Quine, possibilism, as understood in the previous paragraph, is straightforwardly inconsistent. For to say ‘there are things that don't exist’ would simply be to say ‘there exist things that don't exist’. He left a loophole, however, that allows a possibilist to skirt Quine's charge. Rather than distinguishing being from existence, the possibilist can agree with Quine that they are identical: everything there is exists. However, the possibilist will insist, not everything that exists is *actual*. To say that there are things that are not actual, then, is to say that there *exist* things that fail to be actual.

One might charge that this second variety of possibilist is playing mere word games: she has retained the metaphysics of classical possibilism, but has simply renamed being as ‘existence’ and existence as ‘actuality’. And indeed, that is perhaps the most natural way of construing the move -- the possibilist agrees to cede Quine's point about the comprehensiveness of the meaning of ‘existence’. The possibilist's rejoinder is that Quine's point is no threat to the coherence of a possibilist metaphysics.

Quine, of course, would likely reply that the wedge driven by this new variety of possibilism between existence and actuality is no more legitimate than the one formerly driven between being and existence. Thus, for Quine, the possibilist has not really addressed the original objection, which challenges the possibilist's introduction of two modes of being. The possibilist has simply replaced two modes of being with two modes of existence --- actual existence and possible existence. Again, though, the possibilist is not without a reply: she can simply deny that actuality is any sort of ontological mode. It's just a property that some things have, and other things lack. The debate at this point seems a stalemate.

There is one final notable form of possibilism that is truer in spirit to Quine's original objection, namely, David Lewis's. As with the second form of possibilism just discussed, on Lewis's view, being and existence coincide; there is no special ontological property, no distinct mode of being, that separates

merely possible objects from actual ones. Moreover, as with the second form, actuality does not coincide with existence. However, unlike the second form, actuality is not any sort of intrinsic ontological property. Indeed, actuality is not really a property at all, but a *relation*: *x* is actual relative to *y* just in case *x* and *y* occupy the same possible world; or, equivalently, just in case they are spatially or temporally related to one another. (*x* is spatially related to *y* just in case *x* and *y* occupy the same space or some distance separates them. Similarly, *x* is temporally related to *y* just in case *x* and *y* exist at exactly the same times or one existed at some time before or after the other.) On this view, then, to say that there are things that are not actual is simply to say that there are things that occupy other worlds than ours, things that exist, in a fully-fledged sense, but which are just spatially and temporally unrelated to *us*.

Lewis proposes a well-known, and natural, semantic corollary to this view about actuality, namely, that the word ‘actual’ is an *indexical*: its reference on any given occasion of utterance, like that of ‘I’, ‘now’, etc., is essentially determined by the context of the utterance, and in particular, the world in which the utterance occurs. Thus, what makes "Clinton is actual" (or, somewhat more naturally, "Clinton actually exists") true when I utter it is not some intrinsic property of Clinton --- actuality --- but rather simply the fact that he occupies the same world as the speaker, i.e., me.

As with the second version of possibilism, then, Lewis acknowledges the comprehensive character of existence, but without introducing any special primitive property, or existential mode, of actuality. Either way, modern day possibilists are able to follow Quine in rejecting the distinction between being and existence and still make sense of both possibilism and modality generally.

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

An Account of Abstract Possible Worlds

As an example of actualistically acceptable abstract worlds, we draw chiefly upon Plantinga's account. In this account, we replace the possibilist idea of merely possible worlds with that of a certain type of *state of affairs* (alternatively, a certain type of proposition) which exists but fails to obtain (alternatively, fails to be true). Where propositions are said to be true or false, states of affairs are said to *obtain* or not. Say that one state of affairs *s includes* another s' if and only if it is not possible that *s* obtain without s' obtaining. Importantly, note that a state of affairs can exist without obtaining, just as a proposition can exist without being true. States of affairs, like propositions, are taken to be necessary beings on this account.

We can now define several critical notions:

- A state of affairs *w* is a *world* just in case it is possible that *w* includes all and only states of affairs that obtain.
- For any state of affairs *s* and world *w*, *s obtains at w* just in case *w* includes *s*.
- A world *w* is *actual* just in case *w* obtains.
- An individual *x exists in* world *w* just in case the state of affairs *x's existing* obtains at *w*.

This theory is then to be applied as follows. The everyday claim ‘it is possible that there are Aliens’ can then be analyzed as: the state of affairs *There are Aliens* obtains at some world (i.e., there is some world, in the above sense of ‘world’, which implies that there are Aliens). If there are no Aliens, then no such world obtains. Similarly, an ordinary claim of the form ‘it is necessary that *p*’ can be analyzed as: *p* obtains at every possible world. Thus, in this first stage of the actualist treatment of modality, ordinary possibility claims are analyzed in terms of actually existing states of affairs. This step is, therefore, consistent with Thesis (A). So far, no possible-but-nonactual objects have been introduced for the analysis of modal claims.

In putting forward this theory, the actualist takes herself to be replacing an obscure distinction between two modes of being -- possible existence and actual existence -- with an intelligible distinction. This distinction is replaced by an allegedly clear distinction between two kinds of existing states of affairs -- those that obtain and those that don't). That the latter distinction is more intelligible than the former ones is often just assumed by the actualist without argument. This invites the question whether there are cogent arguments for this assumption. However, again, we will not pursue this question here.

Furthermore, in putting forward this theory, the actualist has not invoked any objects which have such

modal properties as being a possible million carat diamond, being a possible talking donkey, being a possible Alien, etc. The ‘worlds’ of the actualist do in fact have modal properties and the fact that they do is essential for them to do the work they have to do in the theory. A possible world is a state of affairs that *could* be such that it includes all and only states of affairs that obtain. Postulating objects with modal properties such as this seems less objectionable to the actualist than postulating objects with the modal properties described at the beginning of this paragraph. This of course invites a certain question, namely, just why is it less objectionable to have objects with the latter modal properties than the former one. But, again, we will not pursue this question here.

This latter point about the actualist theory of worlds brings us to the second step of their treatment of modality, namely, how to analyze ordinary modal claims that seem to require such possible individuals as possible million carat diamonds, possible talking donkeys, possible Aliens, etc. For the remainder of this essay, then, we assume that some actualist theory of worlds is viable and therefore concentrate our energies solely on the problems that arise in connection possible individuals rather than possible worlds.

Return to Actualism

Return to Background Assumptions for Plantinga's Account

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Background Assumptions for Plantinga's Account

There are several background assumptions that are required for Plantinga's solution to be adequate

> (P1) Worlds are (maximal possible) states of affairs.

That is, Plantinga adopts the approach sketched in the supplementary document An Account of Abstract Possible Worlds on which possible worlds are abstract states of affairs that actually exist but (in general) don't obtain. Plantinga thus avoids the first aspect of the possibilist challenge.

> (P2) Properties are "first-class citizens"; that is, they are legitimate objects of reference and (first-order) quantification;

If individual essences are to play the role that possibilia play in Kripke's account, then, clearly, it must be possible to quantify over them directly with first-order quantifiers. Plantinga is a platonist --- properties, propositions, and states of affairs are as real as any concrete particular, it's just that they are abstract. Hence, they can be quantified over no less than their concrete counterparts. The advantages of this move are clear. As noted above, Kripke's account fails to be actualist because it quantifies over possibilia in the metalanguage. By replacing possibilia with individual essences, quantifiers range over essences, and hence only over actually existing things.

> (P3) Properties, propositions, and states of affairs all exist necessarily.

As noted in the main document, unlike (most, at least) concrete particulars, properties are necessary beings for Plantinga; it is not possible that there be a property that might fail to exist. Consequently, necessarily, any property that exists in any possible world exists in all possible worlds. (That is, in terms of Plantinga's actualist reconstruction of worlds, necessarily, if a property's existence is entailed by any possible world, it is entailed by all possible worlds.) It follows, in particular, that individual essences are necessary beings.

Why is (P3) needed? After all, in general, not all possibilia exist in every possible world in Kripke's semantics and hence, it would seem, they are not necessary beings; more formally, NE is not a logical truth in Kripke's system. However, the fact that NE fails in Kripke's theory is simply an artifact of his semantics for the quantifiers --- quantifiers at a given world **w** range only over the things that exist in **w**, and hence, that NE is not a logical truth in Kripke's system only reflects the fact that not all possibilia exist *in*, i.e., are *actual* in, every possible world. However, all possibilia are in a clear sense *there* at each

world all the same; Kripke could just as easily have defined the semantics of the quantifiers so that they ranged over all of them. (Indeed, in his original model theory, a 1-place predicate at a world **w** can have possibilia that do not exist in **w** in its extension.) He chose not to define the semantics of the quantifiers in this way simply to avoid validating BF, CBF, and NE and thus to have at least an actualistically acceptable proof theory. Thus, although there is no object language theorem expressing that possibilia are necessary beings, this is clearly an important metaphysical consequence of his semantics, and hence it needs to be reflected in Plantinga's solution.

The necessity of individual essences is still not enough, though. For if individual essences are to replace possibilia, then we must be assured that there are enough of them, that, crudely put, for "every possible individual" there is a corresponding individual essence. This is guaranteed by the following principle:

> (P4) Necessarily, every object has an individual essence.

Christopher Menzel
cmenzel@tamu.edu

Return to Actualism

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

A First-Order, Quantified Modal Language

Primitive Vocabulary:

Terms:

Individual Constants: $a, b, c, \ldots$

individual Variables: $x, y, z, \ldots$

Predicates: $P^n, Q^n, R^n, \ldots \quad (n \geq 0)$

Atomic Formulas:

1. If $\tau_1, \ldots, \tau_n$ are any terms and ρ^n is any predicate, then $\rho^n \tau_1 \ldots \tau_n$ is an atomic formula. ('$\tau_1, \ldots, \tau_n$ exemplify ρ^n')
2. If τ_1 and τ_2 are any terms, then $\tau_1 = \tau_2$ is an atomic (identity) formula. ('τ_1 is identical to τ_2')

Formulas:

1. All atomic formulas are formulas.
2. If φ is a formula, then so is $\neg\varphi$. ('it is not the case that φ')
3. If φ and ψ are formulas, then so is $\varphi \rightarrow \psi$. ('if φ, then ψ')
4. If φ is a formula, and α is any variable, then so is $\forall\alpha\varphi$. ('every object α is such that φ')
5. If φ is a formula, then so is $\Box\varphi$. ('it is necessary that φ')

Definitions:

$$\varphi \,\&\, \psi \;=_{df}\; \neg(\varphi \rightarrow \neg\psi)$$

$$\varphi \,\&\, \psi \;=_{df}\; \neg(\varphi \rightarrow \neg\psi)$$

$$\varphi \vee \psi \;=_{df}\; \neg\varphi \rightarrow \psi$$

$$\varphi \equiv \psi \;=_{df}\; (\varphi \rightarrow \psi) \,\&\, (\psi \rightarrow \varphi)$$

$$\exists\alpha\varphi \;=_{df}\; \neg\forall\alpha\neg\varphi$$

$$\Diamond\varphi \;=_{df}\; \neg\Box\neg\varphi$$

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 200

Stanford Encyclopedia of Philosophy
Supplement to Actualism

The Simplest Semantics for a Quantified Modal Language

Interpretations: An interpretation $\mathbf{I}$ is any quadruple $\langle \mathbf{W}, \mathbf{w}_\alpha, \mathbf{D}, \mathbf{V} \rangle$ defined as follows: (1) $\mathbf{W}$ is a non-empty domain of possible worlds. (2) $\mathbf{w}_0$ is a distinguished member of $\mathbf{W}$ (the actual world). (3) $\mathbf{D}$ is a non-empty domain of individuals. (4) $\mathbf{V}$ is a function assigning semantic values to the constants and predicates as follows: (a) $\mathbf{V}(\tau) \in \mathbf{D}$, and (b) $\mathbf{V}(\rho^n) \in [\mathcal{P}(\mathbf{D}^n)]^{\mathbf{W}}$; i.e., $\mathbf{V}$ both (a) assigns each constant τ to an element of the domain $\mathbf{D}$ and (b) assigns to each n-place predicate ρ^n a function which maps each possible world $\mathbf{w}$ (supplied as an argument) to the set of n-tuples that ρ^n applies to at $\mathbf{w}$.

Assignments to Variables and the Denotation Function: (1) An *assignment* to the variables is any function $\mathbf{f}$ which maps each variable to a member of the domain $\mathbf{D}$. (2) If given an interpretation $\mathbf{I}$ and assignment function $\mathbf{f}$, we define *the denotation of* term τ *with respect to* interpretation $\mathbf{I}$ *and* assignment $\mathbf{f}$ ('$\mathbf{d}_{\mathbf{I},\mathbf{f}}$') as follows: (a) If τ is a constant, then $\mathbf{d}_{\mathbf{I},\mathbf{f}}(\tau)$ is $\mathbf{V}(\tau)$, and (b) If τ is a variable, then $\mathbf{d}_{\mathbf{I},\mathbf{f}}(\tau)$ is $\mathbf{f}(\tau)$.

Truth$_{\mathbf{I},\mathbf{f}}$ at a World (φ *is true*$_{\mathbf{I},\mathbf{f}}$ *at* $\mathbf{w}$):

(1) $\rho^n\tau_1 \ldots \tau_n$ is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$ iff $\langle \mathbf{d}_{\mathbf{I},\mathbf{f}}(\tau_1), \ldots, \mathbf{d}_{\mathbf{I},\mathbf{f}}(\tau_n) \rangle \in [\mathbf{V}(\rho^n)](\mathbf{w})$.
(2) $\tau_1 = \tau_2$ is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$ iff $\mathbf{d}_{\mathbf{I},\mathbf{f}}(\tau_1) = \mathbf{d}_{\mathbf{I},\mathbf{f}}(\tau_2)$.
(3) $\neg\psi$ is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$ iff ψ fails to be true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$.
(4) $\psi \rightarrow \chi$ is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$ iff either ψ fails to be true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$ or
χ is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$.
(5) $\forall x\psi$ is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$ iff for every $\mathbf{f}'$, if $\mathbf{f}' \stackrel{x}{=} \mathbf{f}$, then ψ is true$_{\mathbf{I},\mathbf{f}'}$ at $\mathbf{w}$.[1]
(6) $\Box\psi$ is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$ iff for every $\mathbf{w}'$, ψ is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}'$.

Truth$_\mathbf{I}$ at a World:

φ is *true*$_\mathbf{I}$ *at* $\mathbf{w}$ iff for every assignment $\mathbf{f}$, φ is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}$.

Truth$_\mathbf{I}$:

φ is *true*$_\mathbf{I}$ iff φ is true$_\mathbf{I}$ at $\mathbf{w}_0$.

Logical Truth:

φ is *logically true* iff for every interpretation $\mathbf{I}$, φ is true$_\mathbf{I}$.

---

[1]The clause '$f' \overset{x}{=} f$' abbreviates the claim that f' is an assignment function just like f except perhaps for what it assigns to the variable x.

Christopher Menzel
cmenzel@tamu.edu

Return to Actualism

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

The Simplest Quantified Modal Logic

Propositional Logic:

Axioms: Tautologies of propositional logic

Rule of Modus Ponens (MP): if $\vdash \varphi \rightarrow \psi$ and $\vdash \varphi$, then $\vdash \psi$

Classical First-Order Quantification Theory:

Axiom: $\forall\alpha\,\varphi \rightarrow \varphi^{\tau}_{\alpha}$, where τ is any term substitutable for α

Axiom: $\forall\alpha(\varphi \rightarrow \psi) \rightarrow (\varphi \rightarrow \forall\alpha\,\psi)$, where α is not free in φ

Rule of Generalization (GEN): if $\vdash \varphi$, then $\vdash \forall\alpha\,\varphi$

Logic of Identity:

Axiom: $x = x$

Axiom: $x = y \rightarrow (\varphi(x,x) \rightarrow \varphi(x,y))$, where $\varphi(x,y)$ is the result of substituting y for some, but not necessarily all, occurrences of x in $\varphi(x,x)$, provided that y is substitutable for x at those occurrences.

S5 Modal Logic:

K Axiom: $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$

T Axiom: $\Box\varphi \rightarrow \varphi$

5 Axiom: $\Diamond\varphi \rightarrow \Box\Diamond\varphi$

Rule of Necessitation (RN): if $\vdash \varphi$, then $\vdash \Box\varphi$

NOTE: Both the Barcan Formula and the Converse Barcan Formula are derivable from the above:

BF: $\forall x\Box\varphi \rightarrow \Box\forall x\varphi$

$$\text{BF: } \forall x \Box \varphi \rightarrow \Box \forall x \varphi$$

$$\Diamond \exists x \varphi \rightarrow \exists x \Diamond \varphi \quad \text{(equivalent formulation)}$$

$$\text{CBF: } \Box \forall x \varphi \rightarrow \forall x \Box \varphi$$

$$\exists x \Diamond \varphi \rightarrow \Diamond \exists x \varphi \quad \text{(equivalent formulation)}$$

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

The 5 Axiom is Logically True

Proof: To see that the 5 axiom is true in every interpretation, pick an arbitrary interpretation **I**. To show that a conditional sentence is true$_{\mathbf{I}}$, the definition tells us that we must show that it is true$_{\mathbf{I}}$ at the actual world $\mathbf{w}_0$. To do this, we assume that the antecedent is true$_{\mathbf{I}}$ at $\mathbf{w}_0$ and then show that the consequent is true$_{\mathbf{I}}$ at $\mathbf{w}_0$. So assume that the antecedent of the 5 axiom, namely, $\Diamond\varphi$, is true$_{\mathbf{I}}$ at $\mathbf{w}_0$. It follows, by the definition of truth, that φ is true$_{\mathbf{I}}$ at some possible world, say $\mathbf{w}_1$. Now to show that $\Box\Diamond\varphi$ is true$_{\mathbf{I}}$ at $\mathbf{w}_0$, we need to show that $\Diamond\varphi$ is true$_{\mathbf{I}}$ at all possible worlds. So pick an arbitrary possible world, say $\mathbf{w}_2$. Note that $\Diamond\varphi$ is true$_{\mathbf{I}}$ at $\mathbf{w}_2$, since φ is true$_{\mathbf{I}}$ at $\mathbf{w}_1$. But since $\mathbf{w}_2$ was chosen arbitrarily, it follows that $\Diamond\varphi$ is true$_{\mathbf{I}}$ in all possible worlds. So $\Box\Diamond\varphi$ is true$_{\mathbf{I}}$ at the actual world.

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

The Barcan Formula is Logically True

Proof: To see that the Barcan Formula is true in every interpretation, pick an arbitrary interpretation **I**. To show that a conditional sentence is true$_{\mathbf{I}}$, the definition tells us that we must show that it is true$_{\mathbf{I}}$ at the actual world $\mathbf{w}_0$. To do this, we assume that the antecedent is true$_{\mathbf{I}}$ at $\mathbf{w}_0$ and then show that the consequent is true$_{\mathbf{I}}$ at $\mathbf{w}_0$. So assume that the antecedent of BF axiom, namely, $\Diamond\exists x\varphi$, is true$_{\mathbf{I}}$ at $\mathbf{w}_0$. It follows, by the definition of truth, that $\exists x\varphi$ is true$_{\mathbf{I}}$ at some possible world, say $\mathbf{w}_1$. It follows from this that some individual in the domain, say **i**, satisfies$_{\mathbf{I}}$ φ at $\mathbf{w}_1$. [This conclusion takes a few liberties with the definition of satisfaction, but no harm comes from employing this simpler manner of speaking as if individuals satisfied formulas at worlds rather than strictly speaking in terms of assignment functions satisfying formulas at worlds.] Therefore, **i** satisfies$_{\mathbf{I}}$ $\Diamond\varphi$ at $\mathbf{w}_0$. So, $\exists x\Diamond\varphi$ is true$_{\mathbf{I}}$ at $\mathbf{w}_0$.

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

The 'Necessary Existence' Principle is Logically True

Proof Sketch: To see that NE is true in every interpretation, pick an arbitrary interpretation **I**. Note that it suffices to show that the open modal formula '$\Box\exists y(y = x)$' is true$_{\mathbf{I}}$ at $\mathbf{w}_0$ (for by the definition of truth, this open formula will have be true with respect to every assignment function **f**; but if so, then '$\forall x\Box\exists y(y = x)$' will be true). To do that, we have to show that '$\exists y(y = x)$' is true$_{\mathbf{I}}$ at every possible world. So pick an arbitrary possible world, say $\mathbf{w}_1$. We have to show that '$\exists y(y = x)$' is true$_{\mathbf{I}}$ at $\mathbf{w}_1$, i.e., that for every assignment to the variables **f**, '$\exists y(y = x)$' is true$_{\mathbf{I},\mathbf{f}}$ at $\mathbf{w}_1$. So, we show, for an arbitrarily chosen assignment, say $\mathbf{f}_1$, that '$\exists y(y = x)$' is true$_{\mathbf{I},\mathbf{f}_1}$ at $\mathbf{w}_1$. Now call the individual which $\mathbf{f}_1$ assigns to the variable x '$\mathbf{i}_1$' (we know there must be such an individual since the domain is nonempty). Now to show '$\exists y(y = x)$' is true$_{\mathbf{I},\mathbf{f}_1}$ at $\mathbf{w}_1$, we have to show that there is some assignment function **f′** which differs from $\mathbf{f}_1$ at most in what it assigns to y and such that '$y = x$' is true$_{\mathbf{I},\mathbf{f}'}$ at $\mathbf{w}_1$. But consider the assignment function $\mathbf{f}^*_1$, which is just like $\mathbf{f}_1$ but which assigns $\mathbf{i}_1$ to the variable y (we know there must be such an assignment function, by the definition of an assignment function). Then, clearly, '$y = x$' is true$_{\mathbf{I},\mathbf{f}^*_1}$ at $\mathbf{w}_1$, since $\mathbf{f}^*_1(x) = \mathbf{f}^*_1(y)$ (and hence $\mathbf{d}_{\mathbf{I},\mathbf{f}^*_1}(x) = \mathbf{d}_{\mathbf{I},\mathbf{f}^*_1}(y)$).

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

The Converse Barcan Formula is Logically True

Proof: To see that the Converse Barcan Formula is true in every interpretation, pick an arbitrary interpretation **I**. To show that a conditional sentence is true$_{\mathbf{I}}$, the definition tells us that we must show that it is true$_{\mathbf{I}}$ at the actual world $\mathbf{w}_0$. To do this, we assume that the antecedent is true$_{\mathbf{I}}$ at $\mathbf{w}_0$ and then show that the consequent is true$_{\mathbf{I}}$ at $\mathbf{w}_0$. So assume that the antecedent of CBF, namely, $\Box\forall x\varphi$, is true$_{\mathbf{I}}$ at $\mathbf{w}_0$. It follows, by the definition of truth, that at every possible world $\mathbf{w}$, $\forall x\varphi$ is true$_{\mathbf{I}}$ at $\mathbf{w}$. It follows from this that at every possible world $\mathbf{w}$, every individual $\mathbf{i}$ in the domain satisfies$_{\mathbf{I}}$ φ at $\mathbf{w}$. [This conclusion takes a few liberties with the definition of satisfaction, but no harm comes from employing this simpler manner of speaking as if individuals satisfied formulas at worlds rather than strictly speaking in terms of assignment functions satisfying formulas at worlds.] Therefore, every individual $\mathbf{i}$ in the domain is such that for every world $\mathbf{w}$, $\mathbf{i}$ satisfies$_{\mathbf{I}}$ φ at $\mathbf{w}$. That is, every individual $\mathbf{i}$ satisfies$_{\mathbf{I}}$ $\Box\varphi$ at $\mathbf{w}_0$. So, $\forall x\Box\varphi$ is true$_{\mathbf{I}}$ at $\mathbf{w}_0$.

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

Proof of Barcan Formula in S5

Proof of $\varphi \to \Box\Diamond\varphi$:

(1)	$\Box\neg\varphi \to \neg\varphi$	instance of T axiom
(2)	$\varphi \to \neg\Box\neg\varphi$	from (1), by contraposition
(3)	$\varphi \to \Diamond\varphi$	from (2), by definition of $\Diamond$
(4)	$\Diamond\varphi \to \Box\Diamond\varphi$	instance of 5 schema
(5)	$\varphi \to \Box\Diamond\varphi$	from (3),(4) by propositional logic

Proof of $\Diamond\Box\psi \to \psi$

(1)	$\neg\psi \to \Box\Diamond\neg\psi$	instance of $\varphi \to \Box\Diamond\varphi$ (above)
(2)	$\Box\Diamond\neg\psi \equiv \neg\Diamond\Box\psi$	instance of modal negation principles
(3)	$\neg\psi \to \neg\Diamond\Box\psi$	from (1),(2) by propositional logic
(4)	$\Diamond\Box\psi \to \psi$	from (3) by contraposition

Proof of Rule1: if $\vdash \chi \to \theta$, then $\vdash \Box\chi \to \Box\theta$

(1)	$\chi \to \theta$	Assume as theorem.
(2)	$\Box(\chi \to \theta)$	from (1) by RN
(3)	$\Box(\chi \to \theta) \to (\Box\chi \to \Box\theta)$	instance of K axiom
(4)	$\Box\chi \to \Box\theta$	from (2),(3) by MP

Proof of Rule2: if $\vdash \Diamond\chi \to \theta$, then $\vdash \chi \to \Box\theta$

(1)	$\Diamond\chi \to \theta$	Assume as theorem.
(2)	$\Box\Diamond\chi \to \Box\theta$	by Rule1 (above)
(3)	$\chi \to \Box\Diamond\chi$	instance of $\varphi \to \Box\Diamond\varphi$ (above)
(4)	$\chi \to \Box\theta$	from (2),(3) and propositional logic

Proof of Barcan Formula: $\forall x\Box\varphi \to \Box\forall x\varphi$

(1)	$\forall x\Box\varphi \to \Box\varphi$	by quantifier axiom
(2)	$\Box[\forall x\Box\varphi \to \Box\varphi]$	from (1) by RN
(3)	$\Box(\chi \to \theta) \to (\Diamond\chi \to \Diamond\theta)$	theorem of K
(4)	$\Diamond\forall x\Box\varphi \to \Diamond\Box\varphi$	from (2) given (3)
(5)	$\Diamond\Box\varphi \to \varphi$	instance of $\Diamond\Box\psi \to \psi$ (above)

(5)	$\Diamond\Box\varphi \rightarrow \varphi$	instance of $\Diamond\Box\psi \rightarrow \psi$ (above)
(6)	$\Diamond\forall x\Box\varphi \rightarrow \varphi$	from (4),(5) by propositional logic
(7)	$\forall x[\Diamond\forall x\Box\varphi \rightarrow \varphi]$	from (6) by GEN
(8)	$\Diamond\forall x\Box\varphi \rightarrow \forall x\varphi$	from (7) and quantifier axiom by MP
(9)	$\forall x\Box\varphi \rightarrow \Box\forall x\varphi$	from (8) by Rule2 (above)

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

Proof of 'Necessary Existence' in S5

Proof of $\forall x \Box \exists y(y = x)$**:**

(1)	$x = x$	instance of identity axiom
(2)	$\forall y(y \neq x) \rightarrow x \neq x$	instance of quantifier axiom
(3)	$x = x \rightarrow \neg \forall y(y \neq x)$	from (2), by contraposition
(4)	$x = x \rightarrow \exists y(y = x)$	from (3), by definition of $\exists$
(5)	$\exists y(y = x)$	from (1) and (4) by MP
(6)	$\Box \exists y(y = x)$	from (5) by RN
(7)	$\forall x \Box \exists y(y = x)$	from (6) by GEN

Return to Section 2 of Actualism (Controversial Consequences of SQML)
Return to Section 3 of Actualism (Kripke's Quantified Modal Logic)

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

Proof of Converse Barcan Formula in S5

Proof of $\Box\forall x\varphi \rightarrow \forall x\Box\varphi$:

(1)	$\forall x\varphi \rightarrow \varphi$	by quantifier axiom
(2)	$\Box(\forall x\varphi \rightarrow \varphi)$	from (1) by RN
(3)	$\Box\forall x\varphi \rightarrow \Box\varphi$	from (2) and K axiom by MP
(4)	$\forall x[\Box\forall x\varphi \rightarrow \Box\varphi]$	from (3) by GEN
(5)	$\Box\forall x\varphi \rightarrow \forall x\Box\varphi$	from (4) and quantifier axiom by MP

Return to Section 2 of Actualism (Controversial Consequences of SQML)
Return to Section 3 of Actualism (Kripke's Quantified Modal Logic)

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

Proof of Barcan Formula Equivalent

Proof of $[\forall x \Box \varphi \rightarrow \Box \forall x \varphi] \equiv [\Diamond \exists x \varphi \rightarrow \exists x \Diamond \varphi]$**:**

$(\rightarrow)$

(1)	$\forall x \Box \psi \rightarrow \Box \forall x \psi$	BF
(2)	$\forall x \Box \neg \varphi \rightarrow \Box \forall x \neg \varphi$	instance of (1)
(3)	$\neg \Box \forall x \neg \varphi \rightarrow \neg \forall x \Box \neg \varphi$	from (2) by contraposition
(4)	$\Diamond \neg \forall x \neg \varphi \rightarrow \exists x \neg \Box \neg \varphi$	from (3) by modal and quantifier negation rules
(5)	$\Diamond \exists x \varphi \rightarrow \exists x \Diamond \varphi$	from (4) by df of '∃' and '◇'

$(\leftarrow)$

(1)	$\Diamond \exists x \psi \rightarrow \exists x \Diamond \psi$	Assume alleged equivalent to BF.
(2)	$\Diamond \exists x \neg \varphi \rightarrow \exists x \Diamond \neg \varphi$	instance of (1)
(3)	$\neg \exists x \Diamond \neg \varphi \rightarrow \neg \Diamond \exists x \neg \varphi$	from (3) by contraposition
(4)	$\forall x \neg \Diamond \neg \varphi \rightarrow \Box \neg \exists x \neg \varphi$	from (3) by quantifier and modal negation rules
(5)	$\forall x \Box \varphi \rightarrow \Box \forall x \varphi$	from (4) by interdefinability of □ and ∀

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

A Derivation of NE from SA and CBF

Consider any property P such that $\Box\forall z Pz$. P might be the property of *being self-identical*, for example. From this, CBF yields $\forall z\Box Pz$. But instantiate this to an arbitrary object, say x, to get $\Box Px$. Then we have the following instance of SA: $\Box(Px \rightarrow \exists y\, y = x)$. Thus, by the K axiom, it follows that $\Box\exists y\, y = x$, i.e., NE. Note that even in a language without identity, in which *existence* is expressed by a predicate '$E!$' and serious actualism is expressed by the formula $\Box[\varphi(x) \rightarrow E!x]$, we would still get the result that $\forall x\Box E!x$.

Christopher Menzel
cmenzel@tamu.edu

Return to Actualism

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

Prior's Modal Logic

One of the first philosophical logicians to think about these problems was Arthur Prior. He pointed out that SQML is "haunted by the myth that whatever exists exists necessarily" ([1968], p. 48). His goal was to develop a modal logic for contingent beings (see Prior [1957] and [1967]). Prior developed a solution to these problems by rejecting the interdefinability of the ‘necessity’ and ‘possibility’ operators and restricting the Rule of Necessitation. The central notion that underlies Prior's system is the notion of *necessary statability*. Prior's idea is that when *a* is a contingent being, then propositions about *a* are not even statable in a world where *a* doesn't exist; i.e., such propositions are not necessarily statable. Intuitively, only those formulas which are about necessary beings or which are wholly general in character are necessarily statable. Thus, Prior introduces the modal claim $S\varphi$ to indicate that the proposition denoted by φ is necessarily statable.

Prior then undermines the interdefinability of possibility and necessity by distinguishing ‘not possibly false’ from ‘necessarily true’ as follows. Using the operator ‘$\Diamond$’ as primitive, Prior labels any statement φ for which there is no world where it is false as ‘not possibly false’ (i.e., $\neg\Diamond\neg\varphi$). For example, when *a* is a contingent being, the statement ‘$Pa \rightarrow Pa$’ is not possibly false, since there is no world in which the statement is false. However, since this statement will be neither true nor false (but rather unstatable) at worlds where *a* doesn't exist, it is not a ‘necessarily true’ statement in the full or strong sense of being true in every possible world. Prior defines this strong sense of necessity in his object language in terms of possibility and statability as follows: φ is necessarily true iff φ is necessarily statable and φ is not possibly false. In formal terms:

$$\Box\varphi \equiv S\varphi \,\&\, \neg\Diamond\neg\varphi$$

So, when *a* is a contingent being, ‘$Pa \rightarrow Pa$’ is not a necessary truth. But when ‘*a*’ denotes an object that necessarily exists, ‘$Pa \rightarrow Pa$’ will be necessarily true. Also, completely general sentences like ‘$\forall xPx \rightarrow \forall xPx$’ will be necessarily true.

Given this distinction between weak necessity (‘not possibly false’) and strong necessity, Prior must make sure that the Rule of Necessitation never allows us to infer the strong necessity of a theorem of logic that is only weakly necessary. Thus, for Prior, the Rule of Necessitation must be reformulated as follows:

Revised RN: If $\vdash\varphi$, then $\vdash(S\varphi \rightarrow \Box\varphi)$

Thus, after establishing that $Pa \rightarrow Pa$ is a theorem of logic, we cannot infer that $\Box(Pa \rightarrow Pa)$. However, we can infer that $\neg\Diamond\neg(Pa \rightarrow Pa)$, since Prior does accept the following rule of inference:

> If $\vdash \varphi$, then $\vdash \neg\Diamond\neg\varphi$

It is now easy to see that BF, NE, and CBF are no longer theorems of Prior's logic. The proof of BF relied both on the interdefinability of possibility and necessity as well as on the unrevised Rule of Necessitation. (Specifically, the interdefinability of possibility and necessity plays a role behind the scenes in the second line of the second subproof in the derivation of BF, and the Rule of Necessitation was used in line 2 in the third subproof and in line 2 of the final, assembled proof of BF.) The proof of NE is undermined since it relied on unrevised RN as well. (Specifically, Revised RN will not permit the inference on line 3 of the derivation of NE.) Finally, the proof of CBF cannot proceed in the usual way using Revised RN. (Specifically, we cannot appeal to Revised RN on line 2 of the derivation of CBF.)

Since BF, NE, and CBF are no longer theorems, one might think that Prior had succeeded in finding the correct modal logic for actualism. However, few actualists have adopted this logic. The interdefinability of possibility and necessity has seemed just too elegant to abandon---it seems to capture something deep about the logical relationships among our modal beliefs. Actualists are divided on the queston of Prior's restriction on RN---some see it as unnecessary while others accept some such weakening of RN. Probably the most awkward consequence of Prior's logic, however, is the fact that logical contradictions which aren't necessarily statable turn out to be weakly possible! Before we explain why this is so, note that when *a* is a contingent being, the claim $\exists y(y = a)$ is not possibly false --- it is true at any world where *a* exists, and is neither true nor false at worlds where *a* fails to exist. As Deutsch observes, "Yet surely there is a sense in which ‘Prior [*a*] exists’ might have been *false*; and there is no way to express this in Prior's system" ([1990], 92-93). Prior here would emphasize that this weak necessity of *a*'s existence doesn't contradict the fact that *a* is a contingent being, for the fact that *a* is contingent is properly expressed by the fact that neither $\exists y(y = a)$ nor its negation are necessarily true (since neither is necessarily statable). However, if that is what it is to say that a fact is contingent, then any literal contradiction mentioning *a* will be contingent! For example, the contradiction Pa & $\neg Pa$ is not necessarily true, nor is its negation, since neither is statable in a world where *a* doesn't exist. So this contradiction becomes a contingent claim. Indeed, if we suppose that $\neg\Box\neg\varphi$ (using Prior's defined sense of $\Box$) defines a weak sense in which φ is possibly true, it turns out that both the formulas $\neg\exists y(y = a)$ and the contradictory formula Pa & $\neg Pa$ are possibly true in exactly this sense. As I have put it elsewhere ([1991], 348), Prior's modal logic "cannot distinguish the expression of [*a*'s] contingency from the possibility of manifest repugnancies".

Christopher Menzel
cmenzel@tamu.edu

Return to Actualism

First published: June 16, 2000
Content last modified: June 16, 2000

Proof of Barcan Formula in S5

Proof of $\varphi \rightarrow \Box\Diamond\varphi$:

(1)	$\Box\neg\varphi \rightarrow \neg\varphi$	instance of T axiom
(2)	$\varphi \rightarrow \neg\Box\neg\varphi$	from (1), by contraposition
(3)	$\varphi \rightarrow \Diamond\varphi$	from (2), by definition of $\Diamond$
(4)	$\Diamond\varphi \rightarrow \Box\Diamond\varphi$	instance of 5 schema
(5)	$\varphi \rightarrow \Box\Diamond\varphi$	from (3),(4) by propositional logic

Proof of $\Diamond\Box\psi \rightarrow \psi$

(1)	$\neg\psi \rightarrow \Box\Diamond\neg\psi$	instance of $\varphi \rightarrow \Box\Diamond\varphi$ (above)
(2)	$\Box\Diamond\neg\psi \equiv \neg\Diamond\Box\psi$	instance of modal negation principles
(3)	$\neg\psi \rightarrow \neg\Diamond\Box\psi$	from (1),(2) by propositional logic
(4)	$\Diamond\Box\psi \rightarrow \psi$	from (3) by contraposition

Proof of Rule1: if $\vdash \chi \rightarrow \theta$, then $\vdash \Box\chi \rightarrow \Box\theta$

(1)	$\chi \rightarrow \theta$	Assume as theorem.
(2)	$\Box(\chi \rightarrow \theta)$	from (1) by RN
(3)	$\Box(\chi \rightarrow \theta) \rightarrow (\Box\chi \rightarrow \Box\theta)$	instance of K axiom
(4)	$\Box\chi \rightarrow \Box\theta$	from (2),(3) by MP

Proof of Rule2: if $\vdash \Diamond\chi \rightarrow \theta$, then $\vdash \chi \rightarrow \Box\theta$

(1)	$\Diamond\chi \rightarrow \theta$	Assume as theorem.
(2)	$\Box\Diamond\chi \rightarrow \Box\theta$	by Rule1 (above)
(3)	$\chi \rightarrow \Box\Diamond\chi$	instance of $\varphi \rightarrow \Box\Diamond\varphi$ (above)
(4)	$\chi \rightarrow \Box\theta$	from (2),(3) and propositional logic

Proof of Barcan Formula: $\forall x\Box\varphi \rightarrow \Box\forall x\varphi$

(1)	$\forall x\Box\varphi \rightarrow \Box\varphi$	by quantifier axiom
(2)	$\Box[\forall x\Box\varphi \rightarrow \Box\varphi]$	from (1) by RN
(3)	$\Box(\chi \rightarrow \theta) \rightarrow (\Diamond\chi \rightarrow \Diamond\theta)$	theorem of K
(4)	$\Diamond\forall x\Box\varphi \rightarrow \Diamond\Box\varphi$	from (2) given (3)
(5)	$\Diamond\Box\varphi \rightarrow \varphi$	instance of $\Diamond\Box\psi \rightarrow \psi$ (above)
(6)	$\Diamond\forall x\Box\varphi \rightarrow \varphi$	from (4),(5) by propositional logic
(7)	$\forall x[\Diamond\forall x\Box\varphi \rightarrow \varphi]$	from (6) by GEN
(8)	$\Diamond\forall x\Box\varphi \rightarrow \forall x\varphi$	from (7) and quantifier axiom by MP
(9)	$\forall x\Box\varphi \rightarrow \Box\forall x\varphi$	from (8) by Rule2 (above)

(6) $\forall x \Box \varphi \rightarrow \Box \forall x \varphi$ from (5) by Rule2 (above)

Proof of 'Necessary Existence' in S5

Proof of $\forall x \Box \exists y(y = x)$**:**

(1)	$x = x$	instance of identity axiom
(2)	$\forall y(y \neq x) \rightarrow x \neq x$	instance of quantifier axiom
(3)	$x = x \rightarrow \neg \forall y(y \neq x)$	from (2), by contraposition
(4)	$x = x \rightarrow \exists y(y = x)$	from (3), by definition of $\exists$
(5)	$\exists y(y = x)$	from (1) and (4) by MP
(6)	$\Box \exists y(y = x)$	from (5) by RN
(7)	$\forall x \Box \exists y(y = x)$	from (6) by GEN

Proof of Converse Barcan Formula in S5

Proof of $\Box\forall x\varphi \rightarrow \forall x\Box\varphi$:

(1)	$\forall x\varphi \rightarrow \varphi$	by quantifier axiom
(2)	$\Box(\forall x\varphi \rightarrow \varphi)$	from (1) by RN
(3)	$\Box\forall x\varphi \rightarrow \Box\varphi$	from (2) and K axiom by MP
(4)	$\forall x[\Box\forall x\varphi \rightarrow \Box\varphi]$	from (3) by GEN
(5)	$\Box\forall x\varphi \rightarrow \forall x\Box\varphi$	from (4) and quantifier axiom by MP

Stanford Encyclopedia of Philosophy
Supplement to Actualism

Plantinga's Definition of an Individual Essence

Essential Properties

A property *P* is *essential to* an individual *x* if and only if it is not possible that *x* exist and fail to exemplify *P*.

Individual Essences

A property *E* is an *individual essence of* an individual *x* if and only if (i) *E* is essential to *x* and (ii) necessarily, for all individuals *y*, *y* exemplifies *E* if and only if *y* = *x*. Hence, *E* is an individual essence, simpliciter, if and only if it is possible that there be an individual *x* such that *P* is an individual essence of *x*.

Note that, by the above definition, an individual could have several individual essences.

Christopher Menzel
cmenzel@tamu.edu

Return to Actualism

First published: June 16, 2000
Content last modified: June 16, 2000

Problems with the Actualist Accounts

Problems for Individual Essences

Plantinga's account is not without its problems. We will raise three that have appeared in the literature. Plantinga has reasonable responses to the first two. The third seems to be the most problematic.

An Extravagant Ontology

The first, a perhaps least pressing, problem of Plantinga's account is its commitment to a rich universe of fine-grained properties, relations, states of affairs, etc. However, the epistemological problems of platonism in general notwithstanding, abstract entities are largely recognized as, if perhaps not indispensible, extremely useful theoretical constructs, and are ubiquitous in contemporary philosophy of language, logic, mathematics, and linguistics, and in artificial intelligence. Hence, Plantinga's project is in no deeper water in this regard than many other well-regarded projects. For all but the most stringent nominalists, then, the only objection can be that the ontology is too rich, too fine-grained. However, once abstract entities are admitted at all, such a charge will stick only if the entities in question are not doing any reasonable philosophical work, and that is not the case in Plantinga's account; for all its richness, the elements of its ontology and their nature seem quite carefully chosen to play their respective roles. Hence, the objection here seems to boil down simply to a difference of philosophical taste. And that is not a serious objection.

Uninterpretability of the Semantics

Plantinga and Jager purport to be providing a genuine semantics for quantified modal languages, an account of the meaning of modal assertions. There are two related, and more serious, objections to their project, however. (These objections are based in part upon Linsky and Zalta [1994].)

First, Plantinga's modal semantics does not square with our basic semantic intuitions. In particular, on Plantinga's semantics, names denote, and quantifiers range over, individual essences. Intuitively, however (and also according to the current dominant theories of reference), names do no such thing; names denote individuals: ‘Quine’ denotes Quine, not his individual essence. Indeed, in order to denote haecceities, we resort to grammatically more complex constructions like gerunds that contain names referring to individuals, e.g., ‘being Quine’. The term ‘Quine’ here must be taken to refer to Quine, not an individual essence, lest the term denote, not an individual essence of Quine, but an individual essence of an individual essence of Quine.

This last point illustrates the second and more difficult objection, namely, that *Plantinga's semantics provides no way of interpreting the basic definitions and principles of the semantics itself.* According to the semantics, quantifiers range over individual essences and atomic formulas express coexemplification. Yet the metalinguistic quantifiers in the definitions of ‘individual essence’ and ‘coexemplification’ must, on pain of circularity, be taken to range over individuals. Again, the crucial principle P4 of Plantinga's semantics says that, necessarily, every object has an individual essence. But clearly, the universal quantifier here cannot itself be interpreted as a quantifier over individual essences, lest P4 express only the trivial proposition that, in every world **w**, every individual essence that is exemplified in **w** is coexemplified with the property **having an individual essence**. So, while P4 is intended to guarantee that there are enough individual essences, intepreted according to Plantinga's own semantics P4 turns out to be true even if there are no individual essences at all.

There are moves that one can imagine Plantinga and Jager making in response to these difficulties. For instance, perhaps they could add a domain of actual individuals to the semantics to serve as the referents for names. Again, perhaps Plantinga intends the semantics not to capture the denotation relation but rather only the relation of *expressing* that he has argued holds between names and individual essences (cf. Plantinga [1974]). The point, however, is that, if the account of Plantinga and Jager is to be understood as a genuine semantic theory as they seem to purport, numerous difficulties must be addressed before the account can be considered viable. If their semantics is to be understood in some other way, its nature and purpose need significant clarification and, where appropriate, modification.

Conceptual Difficulties with Individual Essences

More pernicious difficulties surround the notion of an individual essence itself, and in particular the notion of a haecceity. To get at the chief problem, first, define a property or relation to be *logically simple* (*simple*, for short) if it is not itself a negation, conjunction, disjunction, quantification, modalization, etc. of any other properties or relations. (The idea here, of course, is that logically simple properties correspond to basic predicates in a language, and logically complex properties are analogous to complex sentences. It is a bit difficult in fact to find any uncontroversial examples of logically simple properties. Perhaps certain fundamental mental states like happiness or physical states like *being a quark* or *having mass* qualify. For purposes here, however, we needn't delve deeper.) Next, say that a property **P** is *general* if it is possible both that (i) something **x** exemplify **P** and that (ii) possibly, something **y** distinct from **x** exemplify **P**. Intuitively, then, a property is general if it can be exemplified by more than one thing, albeit perhaps only at different times or in different possible worlds. The notion of generality can be extended to relations in an obvious way.

Now, haecceities are either simple or they are not. Both options are problematic. Plantinga refers to haecceities by means of two types of gerunds: grammatically simple gerunds like ‘being Quine’, and grammatically more complex gerunds like ‘being identical with Quine’. Those of the former sort suggest that haecceities are logically simple, the latter that they are logically complex. We consider them in turn.

If haecceities are logically complex, the central question is: In what does this logical complexity consist?

An appealing and quite popular answer dating back to Russell is that logical complexity, at least in part, involves a certain type of metaphysical complexity: a logically complex property, proposition, or relation is literally *constituted* by less complex metaphysical parts. (See Frege [1980], p. 169.) So, for example, the property **being human and over 6 feet tall** is constituted by the properties **being human** and **being over 6 feet tall**. And, most relevantly, *singular* properties and relations like **being a student of Quine** that involve expressions for a relation and an individual are constituted by those very entities, in this case, in this case, the relation **being a student of** and *Quine himself*.

If this account is correct, then Quine is a literal metaphysical component of the haecceity **being identical with Quine**. If so, however, then it seems that haecceities are ontologically dependent on their instances; no haecceity exists uninstantiated. For Quine is the very component that distinguishes **being identical with Quine** from every other haecceity, and hence he appears to be essential to its identity. But if that is correct, then haecceities cannot play the role of *possibilia*, for possibilia are, in a certain sense, necessary beings. Though perhaps not *actual* in every world, nontheless, for the possibilist, necessarily, for every *possibile* **x** and every world **w**, *there is* such a thing as **x** at **w**. But if haecceities are ontologically dependent upon their instances, then there are no uninstantiated haecceities. In particular, then, there are no haecceities that could be instantiated by Aliens, since, by hypothesis, no actual individual is possibly an Alien. Hence, Plantinga's semantics for (1) do not work, as they depend upon the existence of an uninstantiated haecceity. Plantinga, of course, could (and, in fact, does) just resist the idea that Quine is a constituent of **being identical with Quine**. However, he identifies no other problems with this conception of logical complexity, and provides no alternative account. Hence, this response seems ad hoc.

A somewhat stronger move for Plantinga is to deny that haecceities are logically complex and take them instead to be logically simple, as suggested by grammatically simple gerunds like ‘being Quine’ that do not involve reference to identity or any other property or relation. Now, the fact that one still has to refer to Quine in order to refer to his haecceity might suggest that **being Quine** is no less ontologically dependent upon Quine than is **being identical with Quine**. The important difference in this case, however, is that there is no apparent logical complexity that needs explaining: **being Quine** -- or perhaps better, **quineity** -- as it happens, simply holds essentially and uniquely of Quine. True enough, we can only refer to Quine's haecceity by referring, at least obliquely, to Quine. However, all that follows from that is that if Quine hadn't existed, it would not have been possible to refer to his haecceity, at least, not by means of a gerund involving a proper name of Quine. The haecceity itself is, arguably, no more ontologically dependent upon Quine than is the number 2. Hence, there is no reason to deny that logically simple haecceities, like other logically simple properties, are necessary beings, and hence no reason to think that they cannot play all of the metaphysical roles demanded of them in Plantinga's account.

The chief objection to this move now, however, is whether Plantinga has distinguished his own view sufficiently from possibilism. On his account, haecceities are logically simple but non-general properties. But this seems a very odd combination. Intuitively, at first blush anyway, properties and relations are common, general, repeatable characteristics of, or connections between, things -- redness, wisdom, humanity, marriage, adjacency, etc. Recognition of shareability among many particulars, awareness of a

one over many, is what gave rise to the concept of a property in the first place. In fact, of course, not all properties are general. But, intuitively again, non-generality comes about by virtue of logical complexity, by virtue of the manner in which the components of a complex property are "woven together" logically, e.g., **being smaller than every other prime number**, **possibly being the father of Xantippe**, or **being identical with Quine**. Hence, it follows from these intuitions that, necessarily, all logically simple properties are general. However, Plantinga flouts these intuitions in order to introduce an entirely new class of simple property whose sole function is to serve as an actualist counterpart to *possibilia*. But given their oddity, it is far from clear that there is any greater philosophical virtue in postulating logically simple but essentially non-general properties than in postulating that there are objects that are not actual. So whatever victory Plantinga can claim for actualism here seems Pyrrhic at best. (Some readers may be interested in the supplementary document Qualitative Essences and a Final Defense for Plantinga.)

Return to Actualism

Problems for World Stories

Though promising and intuitive, Adams' account, like Plantinga's, suffers from some serious objections.

Loss of Compositionality

One of the traditional strengths of possible worlds semantics is that it provides a compositional semantics for modal notions. A virtue of the Plantinga/Jager approach is that it preserves compositionality. Haecceities, however, are the key to their approach. Adams' account by contrast, with its rejection of haecceities, sacrifices compositionality. Notably, the semantics of (1) must stop at (14): Because there are in fact no Aliens, the proposition **There are Aliens** has no witnesses, no objects that, by virtue of their actual or possible properties, make it true. Hence, unlike compositional accounts, one cannot further analyze the right side of the biconditional in (14).

The Iterated Modalities Objection

However, perhaps the sacrifice of compositionality is not too high a price for the strong actualist to pay. After all, we *understand* quantification well enough from standard, nonmodal Tarskian semantics. And given strong actualism, it is no surprise that (14) should be unanalyzable, for there *are* no possible Aliens to serve as witnesses to (14). The real goal of a semantics of modality is to provide an analysis of our ordinary use of *modal* operators, not extensional ones like quantification. From that perspective, the unanalyzability of (14) is unproblematic, as the important work has been done in analyzing the modal operator in (1).

However, at this point Adams falls victim to the iterated modalities objection. Recall the following proposition:

(7) The pope could have had a son who could have become a priest

that is, formally,

(9) $\Diamond\exists x(Sxp \;\&\; \Diamond Px)$.

On Adams' semantics, (9) is true if and only if

(15) '$\exists x(Sxp \;\&\; \Diamond Px)$' (i.e., the proposition **Wojtyla has a son who could have become a priest**) is true at some world **w**.

The problem now is that, if we stop the analysis of (9) with (15), the semantic prize noted above -- the analysis of our ordinary modal locutions in terms of possible worlds -- is lost, as (15) contains an embedded modal operator that cannot be analyzed in terms of world stories. For to do so, unlike the case of (14), one must produce a witness for '$\exists x(Sxp \;\&\; \Diamond Px)$' about whom it is true at some world that *he* is a priest. That is, in order to analyze the embedded modal operator, (15) seems to require analysis along the following lines:

(16) For some some individual *x*, '$Sxp \;\&\; \Diamond Px$' (i.e., the proposition ***x* is a son of Wojtyla and *x* could have become a priest**) is true at some world **w**,

which then enables us to analyze the embedded modal operator:

(17) For some some individual *x*, '*Sxp*' (i.e., the proposition ***x* is a son of Wojtyla**) is true at some possible world **w** and, for some possible world **u**, '*Px*' (i.e., the proposition ***x* is a priest**) is true at **u**.

But by strong actualism, there is no such instance *x*, as the pope has no children and, assuming that one's actual parents are necessarily one's parents (provided one exists), nothing actual could have been the pope's child. Hence, given strong actualism, there is no information about any such *x*, no singular proposition that is directly about any such *x*. Hence, there is no proposition of the requisite form ***x* is a son of Wojtyla** specified in (17) to be true at some possible world, that is, to be a member of some world story. Granted, there *could* be such a proposition, but, given strong actualism, there isn't in fact. Hence, Adams' semantics fails to provide the analysis of our ordinary modal discourse that it purports to provide.

Return to Actualism

Problems for World Propositions

Fine's account leaves us with a general question: What role, then, are possible worlds playing in accounts like Plantinga's, Adams', and Fine's in which possible worlds are defined in terms of a primitive modality? If these accounts are not providing a semantical *analyses* of modality, what is their purpose? Why clutter our ontology with them if they do not lead to any genuine semantical insight?

A natural answer is that, even if worlds don't provide us with a semantical *analysis* of modality, possible worlds are still tremendously useful in the formal analysis of modality. Notably, the apparatus of possible world semantics enables us to define such critical properties as consistency and completeness for formal systems of modality. It also enables a definition of the critical notion of logical consequence for modal languages. However, the formal, mathematical *apparatus* of possible world semantics -- the name notwithstanding -- is completely independent of any philosophical conception of possible worlds. There need be no "intended" possible worlds model -- one actually containing possible worlds of some ilk -- in order for us to employ the apparatus to define and use the notions above. Hence, the usefulness of the formal semantics provides us with no reason to allow possible worlds into our ontology.

However, perhaps that is too hasty. For while it is true that possible world semantics is independent of the question of whether there are any possible worlds, by the same token, if there are no possible worlds in any sense, it is difficult to justify the use of possible worlds models as a reasonable formal mechanism for studying the semantical properties of modal languages. For if modal truth is not related in any way to truth in honest-to-goodness possible worlds, then it is hard to see any connection between the informal, intuitive notions modal truth and entailment and the formal notions of truth-in-a-(possible-worlds)-model and logical consequence. Arguably, then, even if they do not provide semantical analyses, possible worlds play an essential role in linking our formal possible world semantics to our ordinary modal concepts of truth and logical consequence. And this could be seen as a motivation for accounts like Adams', Plantinga's, and Fine's. The possible worlds of these accounts give substance to the important intuition that *there are* other ways that things could have been, that there is indeed an intuitive connection between modal truth and quantification over in possible worlds. Thus, the "worlds" in a formal possible worlds model can be viewed as something more than mere formal artifacts. They can be seen as actual representations of other ways the world could have been, and quantification over them as a reasonable representation of an important way in which we think intuitively about modal truth. This connection between the "worlds" of a formal model and genuine possible worlds thus establishes a connection between truth-in-a-model and honest-to-goodness modal truth. (Although without the actual existence of an *intended* model, as is the case for Adams and Fine, rather more would need to be said to flesh out the nature of this connection.) Hence, even though they do not provide semantical analyses, the accounts in question can, at least, be thought of as providing a philosophical justification for considering the apparatus of formal possible worlds models to be a reasonable formal mechanism for studying the semantical properties of modal languages. An actualist account that attempts to do without worlds but which retains an understanding of possible world semantics that ties it to our ordinary notion of modal truth is discussed in the section Dispensing With Worlds.

Return to Actualism

Problems for Roles

McMichael's semantics seems reasonably successful in coping with the problems facing the semantics of Plantinga and Adams. The account is actualist, but neither requires haecceities nor falls prey to the compositionality and iterated modalities objections. However, the success of the account comes at a fairly steep intuitive price, as it abandons strong intuitions about *de re* modality. McMichael would have us understand (21), the statement that Socrates could have been foolish, to be expressing a fact about a complex, abstract property, a role, that bears some sort of accessibility relation to another role that is exemplified by Socrates. actual role of Socrates. Similarly, (9) expresses a fact about a role accessible to the Pope's actual role. But one still wonders: what do such facts about roles and their accessibility to one another have to do with the properties *Socrates* could have had? What, in particular, is the connection between the accessibility of one role to another and the modal properties of individuals? What we'd like to say is that a role **S** is accessible to, say, Socrates' actual role $\mathbf{R_s}$ just in case it is a role Socrates could have exemplified. But, in McMichael's semantics, to say that Socrates could have exemplified any given property **P** (a role in particular) is only to say that some role that includes **P** is accessible to $\mathbf{R_s}$, and we are back where we started with no insight into the connection between accessibility and the modal properties of Socrates. On this score, both possibilism and haecceitism account far more satisfactorily for our modal intuitions.

Return to Actualism

Problems for the No-world Account

Perhaps the central problem for the no-world account is that is abandons traditional ideas about truth and modality. The no-worlder counts his eschewal of worlds as a virtue. However, as Linsky and Zalta [1996] note, this is problematic for several reasons. For if we abandon possible worlds, we must also abandon the "seminal insight" that necessary truth is truth in all possible worlds. And this is problematic for several reasons, for it not only undermines the elegant, extensional characterization of the truth conditions of modal claims, but also dismisses the intuition so strongly evident in ordinary thinking about modality that there exist alternative *ways the world might have been*. For the no-worlder, however, there are no worlds, and hence no intended model whose indices are genuine possible worlds. There are instead only purely formal intended* models that possibly model the structure of modal reality. But, Linsky and Zalta ask,

> ... surely there is something more to modal truth than this; surely *necessity* and *possibility* are about something besides the structure of intended* models, something which *grounds* modal truth and which is modeled by an intended model. ([1994], 444)

But intended* models are not really *models* of any*thing*. At best the have the property of being actual objects that *possibly* model the *structure* of modal reality. But a model of the pure structure of modal reality, the objection might continue, is not the same as a genuine model of modal reality. On the no-

worlds approach,

> ...we cannot say that modal discourse is in part *about* the objects over which the quantifiers range, at least not in the same way that we can say that nonmodal language is about these objects. (Linsky and Zalta, [1994], 444)

And if not, it is hard to see in what sense the no-worlds approach accounts for modal truth at all.

Return to Actualism

Problems for New Actualism

New actualism is a powerful and elegant solution to the problem of possibilism. It appears to have all the theoretical power of possibilism without possibilism's commitment to mere possibilia; everything there is is actual. However, one might argue that it is no surprise that new actualism retains the theoretical power of possibilism, because new actualism is nothing more than thinly veiled possibilism: the new actualist's "actuality" is just the possibilist's being, and contingent nonconcreteness is nothing but the possibilist's mere possibility; nothing but terminology distinguishes a mere *possibile* from a possibly (but not actually) concrete individual.

Even if this is all there were to new actualism, it would not be insignificant. For, in that case, new actualism shows that the standard definition of actualism has not gotten to the heart of the matter -- that actualism is not best characterized as the thesis that everything there *is*, in any sense, is actual. For new actualism demonstrates, at least, that there is a way of systematically reclothing possibilist statements in actualist guise.

This might prompt the "classical" actualist to try to get at the essence of her view in a slightly different way. The real target of the claim that everything is actual is the possibilist's division of being into two modes: actuality and mere possibility (i.e., contingent nonactuality). The new actualist certainly denies *that* division; there are indeed no mere *possibilia* for the new actualist, no objects that are, but which fail to be actual. However, the classical actualist would argue that the new actualist still violates the spirit of actualism. The new actualist does indeed maintain a single sense of being; but in place of the possibilist's division of being into two modes -- actuality and contingent nonactuality -- the new actualist substitutes a division of actuality into two modes: concreteness and contingent nonconcreteness. It is difficult not to see this as a mere relabeling of the possibilist's distinction. Most classical actualists will, therefore, regard new actualism as having the *form* of actualism without having the content necessary to serve as a genuine solution to the possibilist challenge.

For further debate on new actualism see Tomberlin [1996] and the reply by Zalta and Linsky [1996].

Return to Actualism

Christopher Menzel
cmenzel@tamu.edu

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

Qualitative Essences and a Final Defense for Plantinga

There is one final possibility that Plantinga could pursue. On the assumption that all logically simple properties are general, say that a property (or proposition) is *purely qualitative* if it is either logically simple or logically complex but "constructed" from purely qualitative properties, relations, and propositions, that is, if it is the conjunction, negation, modalization, quantification, etc of purely qualitative properties, relations, and propositions. For example, **being smaller than every other prime number** is a conjunction of the properties **being a prime number** and the quantified property (expressed somewhat awkwardly) **being something smaller than every prime number distinct from it** -- in the notation of the lambda calculus: $[\lambda x\ \ (Px\ \&\ \forall y((Py\ \&\ x \neq y) \rightarrow Sxy))]$. Again, **being the father of something** is the existential generalization of one of the "argument places" of the **father of** relation. By contrast, properties that "involve" a particular individual like **being Quine** and **being married to Xantippe** are not purely qualitative, as they involve the individuals Quine and Xantippe.

Now, there surely are some purely qualitative individual essences, e.g., **being smaller than every other prime number**. However, the only clear examples of such belong to necessary beings like the number 2. The real question is whether *contingent* beings have purely qualitative essences. Adams [1979] has argued convincingly that they do not. His central argument is that, given any possible world **w**, there is a world Sym(**w**) that is "symmetrical" with respect to **w**. The idea of symmetry here is difficult to define precisely, but the intuitive idea is that Sym(**w**) contains two parts, each of which is a sort of "copy" of **w**. One of these copies --- call it **C1** -- is (but for properties arising from the existence of the other copy) identical to **w** in both qualititative and nonqualitative respects, and in particular *contains exactly the same objects*as **w**. The other copy --- call it **C2** --- is an exact *qualitative replica* of **w**, i.e., a copy that is indistinguishable from **w** in all qualitative respects (other than those arising from the existence of **C1**). Every object in **C1**, and hence in **w** thus has a qualitative "doppelgänger" in **C2**, an exact replica that has all of its purely qualitative properties. Given this, it seems that there is a possible world **w′** such that Sym(**w′**) = Sym(**w**), but where now **w′** is identical with **C2**, and where **C1** is the replica. It follows that, for every possible world **w**, there is another world **w′** that is qualitatively identical with **w**, but which contains only doppelgängers of the objects in **w**.

Now, one might argue that all that follows from the example is that, for any object *x* and any world **w** containing *x*, there is a distinct object *y* in some other world **w′** that has all of *x*'s qualitative *nonmodal* properties in **w**. That is, Adams' argument from symmetrical worlds only makes plausible the idea that, necessarily, every object *x* has, as one might say, a *de facto* doppelgänger, something qualitatively identical to it with regard to the properties it just happens to exemplify. It does not follow that,

necessarily, there is a *complete*, or *modal* doppelgänger *y* of *x*, i.e., that, for any world **w** in which *x* exists, there is a *de facto* doppelgänger of *x* with respect to **w** such that, in every other world $\mathbf{w}'$ in which *x* exists, there is a world $\mathbf{w}''$ such that *y* is a *de facto* doppelgänger in $\mathbf{w}''$ of *x* with respect to $\mathbf{w}'$, and furthermore, that *x* is a modal doppelgänger of *y*. Hence it does not follow that there are no purely qualitative essences -- perhaps the purely qualitative *modal* properties of *x* are sufficient to distinguish it from any of its *de facto* doppelgängers, i.e., that none of its *de facto* doppelgängers are modal doppelgängers.

This is certainly a line worth pursuing. However, the central problem with it is that there appears to be no intuitive justification for this claim. Given that a doppelgänger *y* in $\mathbf{w}'$ is qualititatively identical to *x* with respect to the nonmodal properties *x* exemplifies in **w**, what possible ground could there be for asserting that the same couldn't to true of *y* with respect to any world in which *x* exists? Intuitively, it seems, there are no such grounds. A defender of this line would have to provide some.

Christopher Menzel
cmenzel@tamu.edu

Return to Problems with Haecceitism
Return to Section 5 of Actualism (Roles)

First published: June 16, 2000
Content last modified: June 16, 2000

Why NE and CBF Are Harmless Consequences of SQML

To see in more detail why the Necessary Existence theorem (NE) of SQML is a harmless consequence of SQML, note that, according to new actualism, Socrates does exist necessarily, but since he is not necessarily concrete, NE does not imply that Socrates is a "necessary being". As we saw above, Socrates' contingency lies in the fact that he is concrete at some possible worlds and nonconcrete at others. Necessary beings, by contrast, are objects that are either concrete at every possible world (like Spinoza's God) or nonconcrete at every possible world (like numbers, sets, etc.). This means that the "contingently nonconcrete" are aptly named, since they are not necessary beings in either of these senses. So the claim that "Everything necessarily exists" (NE), does not conflict with our intuition that some beings are contingent, once the notion of what it is to be contingent is properly understood. Nor does its necessitation NNE. So if NE and NNE are acceptable consequences of SQML, the worry about the Converse Barcan Formula (CBF) disappears as well, for the fact that it, together with Serious Actualism (SA), implies NE, is of little consequence.

Christopher Menzel
cmenzel@tamu.edu

Return to Actualism

First published: June 16, 2000
Content last modified: June 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Actualism

New Actualism and Iterated Modalities

To see that iterated modalities pose no problem for new actualism, recall first that the sentences

(7) The pope (i.e., Karol Wojtyla) could have had a son who could have become a priest.
(8) There could be a planet disturbing the orbit of planet Y and it could have a period of Z years.

can be represented as follows:

(9) $\Diamond \exists x(Sxp \;\&\; \Diamond Px)$
(10) $\Diamond \exists x(Dxy \;\&\; \Diamond Pxz)$

These have straightforward actualist truth conditions requiring no possibilia. For consider (9). The thing *x* which might have been both the pope's son and possibly a priest is a contingently nonconcrete object that at some other world is the pope's son and at yet another world is a priest.

Christopher Menzel
cmenzel@tamu.edu

Return to Actualism

First published: June 16, 2000
Content last modified: June 16, 2000

Classical Logic

Typically, a *logic* consists of a formal or informal language together with a deductive system and/or a model-theoretic semantics. The language is, or corresponds to, a part of a natural language like English or Greek. The deductive system is to capture, codify, or simply record which *inferences* are correct for the given language, and the semantics is to capture, codify, or record the meanings, or truth-conditions, or possible truth conditions, for at least part of the language.

The following sections provide the basics of a typical logic, sometimes called "classical elementary logic" or "classical first-order logic". Section 2 develops a formal language, with a rigorous syntax and grammar. The formal language is a recursively defined collection of strings on a fixed alphabet. As such, it has no meaning, or perhaps better, the meaning of the formulas is given by the deductive system and the semantics. Some of the symbols have counterparts in ordinary language. We define an *argument* to be a non-empty collection of formulas in the formal language, one of which is designated to be the conclusion. The other formulas (if any) in an argument are its premises. Section 3 sets up a deductive system for the language, in the spirit of natural deduction. An argument is *derivable* is there is a deduction from some of its premises to its conclusion. Section 4 provides a model-theoretic semantics. An argument is *valid* if there is no interpretation (in the semantics) in which its premises are all true and its conclusion false. This reflects the longstanding view that a valid argument is truth-preserving.

In Section 5, we turn to relationships between the deductive system and the semantics, and in particular, the relationship between derivability and validity. We show that an argument is derivable only if it is valid. This pleasant feature, called *soundness*, entails that no deduction takes one from true premises to a false conclusion. Thus, deductions preserve truth, and there aren't too many deductions. Then we establish a converse, called *completeness*, that an argument is valid only if it is derivable. This establishes that the deductive system is rich enough to provide a deduction for every valid argument. There are enough deductions. All and only valid arguments are derivable. We briefly indicate other features of the logic, some of which are corollaries to soundness and completeness.

---

1. Introduction

Today, logic is both a branch of mathematics and a branch of philosophy. In most large universities, both departments offer sequences of courses in logic, and there is usually a lot of overlap between them. Formal languages, deductive systems, and model-theoretic semantics are mathematical objects and, as such, the logician is interested in their mathematical properties and relations. Soundness, completeness, and most of the other results reported below are typical examples. Philosophically, logic is the study of *correct reasoning*. Reasoning is an epistemic, mental activity. This raises questions concerning the philosophical relevance of the mathematical aspects of logic. How do deducibility and validity, as properties of formal languages--sets of strings on a fixed alphabet--relate to correct reasoning? What do the mathematical results reported below have to do with the original philosophical issue? This is an instance of the philosophical problem of explaining how mathematics applies to non-mathematical reality.

Typically, ordinary reasoning takes place in a natural language, or perhaps a natural language augmented with some mathematical symbols. So our question begins with the relationship between a natural language and a formal language. Without attempting to be comprehensive, it may help to sketch several options on this matter.

One view is that the formal languages accurately exhibit actual features of certain fragments of a natural language. Some philosophers claim that declarative sentences of natural language have underlying *logical forms* and that these forms are displayed by formulas of a formal language. Other writers hold that (successful) declarative sentences express *propositions*; and formulas of formal languages somehow display the forms of these propositions. On views like this, the components of a logic provide the underlying deep structure of correct reasoning. A chunk of reasoning in natural language is correct if the forms underlying the sentences constitute a valid or deducible argument. See for example, Montague [1974], Davidson [1984], Lycan [1984].

Another view, held at least in part by Gottlob Frege and Wilhelm Leibniz, is that because natural languages are vague and ambiguous, they should be *replaced* by formal languages. A similar view, held by W. V. O. Quine (e.g., [1960], [1986]), is that a natural language should be *regimented*, cleaned up for serious scientific and metaphysical work. One desideratum of the enterprise is that the logical structures in the regimented language should be transparent. It should be easy to "read off" the logical properties of each sentence. A regimented language is similar to a formal language regarding, for example, the explicitly presented rigor of its syntax and its truth conditions.

On a view like this, deducibility and validity represent *idealizations* of correct reasoning in natural language. A chunk of reasoning is correct to the extent that it corresponds to, or can be regimented by, a valid or deducible argument in a formal language.

When mathematicians and many philosophers reason, they occasionally invoke formulas in a formal language to help disambiguate, or otherwise clarify what they mean. In other words, sometimes formulas in a formal language are *used* in ordinary reasoning. This suggests that one might think of a formal language as an *addendum* to a natural language. Then our present question concerns the relationship between this addendum and the original language. What do deducibility and validity, as sharply defined on the addendum, tell us about correct reasoning in general?

Another view is that a formal language is a *mathematical model* of a natural language in roughly the same sense as, say, a collection of point masses is a model of a system of physical objects, and the Bohr construction is a model of an atom. In other words, a formal language displays certain features of natural languages, or idealizations thereof, while ignoring or simplifying other features. The purpose of mathematical models is to shed light on what they are models of, without claiming that the model is accurate in all respects or that the model should replace what it is a model of. On a view like this, deducibility and validity represent mathematical models of (perhaps different aspects of) correct reasoning in natural languages. Correct chunks of reasoning correspond, more or less, to valid or deducible arguments; incorrect chunks of reasoning roughly correspond to invalid or non-deducible arguments. See, for example, Corcoran [1973] or Shapiro [1998].

There is no need to adjudicate this matter here. Perhaps the truth lies in a combination of the above options, or maybe some other option is the correct, or most illuminating one. I raise the matter only to lend some philosophical perspective to the formal treatment that follows.

2. Language

Here we develop the basics of a formal language, or to be precise, a class of formal languages. Again, a formal language is a recursively defined set of strings on a fixed alphabet. Some aspects of the formal languages correspond to, or have counterparts in, natural languages like English. Technically, this "counterpart relation" is not part of the formal development, but I will mention it from time to time, to motivate some of the features and results.

Building blocks

We begin with analogues of *singular terms,* linguistic items whose function is to denote a person or object. We call these *terms*. We assume a stock of *individual constants*. These are lower-case letters, near the beginning of the Roman alphabet, with or without numerical subscripts:

$a, a_1, b_{23}, c, d_{22}$, etc.

We envisage a potential infinity of individual constants. In the present system each constant is a single character, and so individual constants do not have an internal syntax. Thus we have an infinite alphabet. This last could be avoided by taking a constant like d_{22}, for example, to consist of three characters, a lowercase "*d*" followed by a pair of subscript "2"s.

We also assume a stock of *individual variables*. These are lower-case letters, near the end of the alphabet, with out without numerical subscripts:

w, x, y_{12}, z, z_4, etc.

Variables serve a dual function. Sometimes a variable is used as a singular term to denote a specific, but unspecified (or arbitrary) object. For example, a mathematician might start a derivation: "Let x be a natural number". Variables are also used to express generality, as in the mathematical assertion that for any natural number x, there is a natural number y, such that $y>x$ and y is prime. Some logicians employ different symbols for unspecified objects (sometimes called "individual parameters") and variables used to express generality.

Constants and variables are the only terms in our formal language, so all of our terms are simple, corresponding to proper names and pronouns. Some authors also introduce *function letters*, which allow complex terms corresponding to: "7+4" and "the wife of Bill Clinton", or complex terms containing variables, like "the father of x" and "x/y". Logic books aimed at mathematicians are likely to contain function letters, probably due to the centrality of functions to mathematical discourse. Books aimed at a more general audience (or at philosophy students), may leave out function letters, since it simplifies the syntax and theory. We follow the latter route here. This is an instance of a general tradeoff between presenting a system with greater expressive resources, at the cost of making its formal treatment more complex.

For each natural number n, we introduce a stock of n-place *predicate letters*. These are upper-case letters at the beginning or middle of the alphabet. A superscript indicates the number of places, and there may or may not be a subscript. For example,

A^3, $B^3{}_2$, P^3, etc.

are three-place predicate letters. We often omit the superscript, when no confusion will result. We also add a special two-place symbol "=" for identity.

Zero-place predicate letters are sometimes called "sentence letters". They correspond to free-standing sentences whose internal structure does not matter. One-place predicate letters, called "monadic predicate letters", correspond to linguistic items denoting properties, like "being a man", "being red", or "being a prime number". Two-place predicate letters, "binary predicate letters", correspond to linguistic items denoting binary relations, like "is a parent of" or "is greater than". Three-place predicate letters

correspond to three-place relations, like "lies on a straight line between". And so on.

The *non-logical terminology* of the language consists of its individual constants and predicate letters. The symbol "=", for identity, is not a non-logical symbol. In taking identity to be logical, we provide explicit treatment for it in the deductive system and the model-theoretic semantics. Most authors do the same, but there is some controversy over the issue (Quine [1986, Chapter 5]). If K is a set of constants and predicate letters, then we give the fundamentals of a language $\mathcal{L}1K=$ built on this set of non-logical terminology. It may be called the *first-order language with identity* on K. A similar language that lacks the symbol for identity (or which takes identity to be non-logical) may be called $\mathcal{L}1K$, first-order logic without identity.

Atomic formulas

If V is an n-place predicate letter in K, and $t_1, ..., t_n$ are terms of K (i.e., constants in K or variables), then $Vt_1... t_n$ is an *atomic formula* of $\mathcal{L}1K=$. Notice that the terms $t_1, ..., t_n$ need not be distinct. Examples of atomic formulas include:

P^4xaab, C^1x, C^1a, D^0, A^3abc.

The last one is an analogue of a statement that a certain relation (A) holds between three objects (a, b, c). If t_1 and t_2 are terms, then $t_1=t_2$ is an atomic formula of $\mathcal{L}1K=$. It corresponds to an assertion that t_1 is identical to t_2.

If an atomic formula has no variables, then it is called an *atomic sentence*. If it does have variables, it is called an *open formula*. In the above list of examples, the first and second are open; the rest are sentences.

Compound formulas

We now introduce the final items of the lexicon:

$\neg$, &, $\vee$, $\rightarrow$, $\forall$, $\exists$, (,)

We give a recursive definition of a *formula* of $\mathcal{L}1K=$:

1. All atomic formulas of $\mathcal{L}1K=$ are formulas of $\mathcal{L}1K=$.

2. If θ is a formula of $\mathcal{L}1K=$, then so is $\neg\theta$.

Asserting a sentence corresponding to $\neg\theta$ is tantamount to denying the sentence corresponding to θ. The symbol "$\neg$" is called "negation", and is a unary connective.

3. If θ and ψ are formulas of $\mathcal{L}1K$=, then so is $(\theta \,\&\, \psi)$.

The ampersand "&" corresponds to the English "and" (when "and" is used to connect sentences). So $(\theta \,\&\, \psi)$ can be read "θ and ψ". The formula $(\theta \,\&\, \psi)$ is called the "conjunction"of θ and ψ.

4. If θ and ψ are formulas of $\mathcal{L}1K$=, then so is $(\theta \vee \psi)$.

The wedge "$\vee$" corresponds to "either . . . or . . . or both", so $(\theta \vee \psi)$ can be read "θ or ψ". The formula $(\theta \,\&\, \psi)$ is called the "disjunction"of $\theta$and ψ.

5. If θ and ψ are formulas of $\mathcal{L}1K$=, then so is $(\theta \rightarrow \psi)$.

The arrow "$\rightarrow$" corresponds to "if . . . then . . . ", so $(\theta \rightarrow \psi)$ can be read "if θ then ψ" or "θ only if ψ".

The symbols "&", "$\vee$", and "$\rightarrow$" are called "binary connectives", since they serve to "connect" two sentences into one. Some authors introduce $(\theta \leftrightarrow \psi)$ as an abbreviation of $((\theta \rightarrow \psi) \,\&\, (\psi \rightarrow \theta))$. The symbol "$\leftrightarrow$" is an analogue of the locution "if and only if".

6. If θ is a formula of $\mathcal{L}1K$= and *v* is a variable, then $\forall v\, \theta$ is a formula of $\mathcal{L}1K$=.

The symbol "$\forall$" is called a *universal quantifier*, and is an analogue of "for all"; so $\forall v\theta$ can be read "for all *v*, θ".

7. If θ is a formula of $\mathcal{L}1K$= and *v* is a variable, then $\exists v\, \theta$ is a formula of $\mathcal{L}1K$=.

The symbol "$\exists$" is called an *existential quantifier*, and is an analogue of "there exists" or "there is"; so $\exists v\, \theta$ can be read "there is a *v* such that θ".

8. That's all folks. That is, all formulas are constructed in accordance with rules (1)-(7).

Clause (8) allows us to do inductions on the complexity of formulas. If a certain property holds of the atomic formulas and is closed under the operations presented in clauses (2)-(7), then the property holds of all formulas.

We next define the notion of an occurrence of a variable being *free* or *bound* in a formula. All variables that occur in an atomic formula are free. If a variable occurs free (or bound) in θ or in ψ, then that same occurrence is free (or bound) in $\neg\theta$, $(\theta \,\&\, \psi)$, $(\theta \vee \psi)$, and $(\theta \rightarrow \psi)$. That is, the (unary and binary) connectives do not change the status of variables that occur in them. All occurrences of the variable *v* in

θ are bound in $\forall v\,\theta$ and $\exists v\,\theta$. Any *free* occurrences of *v* in θ are bound by the initial quantifier. All other variables that occur in θ are free or bound in $\forall v\,\theta$ and $\exists v\,\theta$, as they are in θ. A variable that immediately follows a quantifier (as in "$\forall x$" and "$\exists y$") is neither free nor bound. We do not think of those as occurrences of the variable.

For example, in the formula ($\forall x(Axy \vee Bx)$ & Bx), the occurrences of "*x*" in *Axy* and in the first *Bx* are bound by the quantifier. The occurrence of "*y*" and last occurrence of "*x*" are free. In $\forall x(Ax \rightarrow \exists xBx)$, the "*x*" in *Ax* is bound by the initial universal quantifier, while the other occurrence of *x* is bound by the existential quantifier. The above syntax allows this "overlap" of bound variables, and it does not create an ambiguity, but we will avoid such formulas, as a matter of taste and clarity

Free variables correspond to place-holders, while bound variables are used to express generality. If a formula has no free variables, then it is called a *sentence*. If a formula has free variables, it is called *open*.

Features of the syntax

Before turning to the deductive system and semantics, I mention a few features of the language, as developed so far. This helps draw the contrast between formal languages and natural languages like English.

We assume at the outset that all of the categories are disjoint. For example, no connective is also a quantifier or a variable, and the non-logical terms are not also parentheses or connectives. Also, the items within each category are distinct. For example, the sign for disjunction does not do double-duty as the negation symbol, and perhaps more significantly, no two-place predicate is also a one-place predicate.

> **Theorem 1**. Every formula of $\mathcal{L}1K$= has the same number of left and right parentheses. Moreover, each left parenthesis corresponds to a unique right parenthesis, which occurs to the right of the left parenthesis. Similarly, each right parenthesis corresponds to a unique left parenthesis, which occurs to the left of the given right parenthesis. If a parenthesis occurs between a matched pair of parentheses, then its mate also occurs within that matched pair. In other words, parentheses that occur within a matched pair are themselves matched.
>
> **Proof**: By clause (8), every formula is built up from the atomic formulas using clauses (2)-(7). The atomic formulas have no parentheses (by the policy that the categories are disjoint). Parentheses are introduced only in clauses (3)-(5), and each time they are introduced as a matched set. So at any stage in the construction of a formula, the parentheses are paired off.

One difference between natural languages like English and formal languages like $\mathcal{L}1K$= is that the latter are not supposed to have any ambiguities. Our policy that the different categories of symbols do not overlap, and that no symbol does double-duty helps avoid the kind of ambiguity, sometimes called

"equivocation", that occurs when a single word has two meanings: "I'll meet you at the bank." But there are other kinds of ambiguity. Consider the English sentence:

> John is married, and Mary is single, or Joe is crazy.

It can mean that John is married and either Mary is single or Joe is crazy, or else it can mean that either both John is married and Mary is single, or else Joe is crazy. An ambiguity like this, due to different ways to parse the same sentence, is sometimes called an "amphiboly". If our formal language did not have the parentheses in it, it would have amphibolies. For example, there would be a "formula" $A \& B \vee C$. Is this supposed to be $((A \& B) \vee C)$, or is it $(A \& (B \vee C))$? The parentheses resolve what would be an amphiboly.

Can we be sure that there are no other amphibolies in our language? That is, can we be sure that each formula of $\mathcal{L}1K$= can be put together in only one way? Showing this is our next task.

Let us temporarily use the term "unary marker" for the negation symbol (¬) or a quantifier followed by a variable (e.g., $\forall x$, $\exists z$).

> **Lemma 2**. Each formula consists of a string of zero or more unary markers followed by either an atomic formula or a formula produced using a binary connective, via one of clauses (3)-(5).
>
> **Proof**: We proceed by induction on the complexity of the formula or, in other words, on the number of formation rules that are applied. The Lemma clearly holds for atomic formulas. Let n be a natural number, and suppose that the Lemma holds for any formula constructed from n or fewer instances of clauses (2)-(7). Let θ be a formula constructed from n+1 instances. The Lemma holds if the last clause used to construct θ was either (3), (4), or (5). If the last clause used to construct θ was (2), then θ is $\neg\psi$. Since ψ was constructed with n instances of the rule, the Lemma holds for ψ (by the induction hypothesis), and so it holds for θ. Similar reasoning shows the Lemma to hold for θ if the last clause was (6) or (7). By clause (8), this exhausts the cases, and so the Lemma holds for θ, by induction.
>
> **Lemma 3**. If a formula θ contains a left parenthesis, then it ends with a right parenthesis, which matches the leftmost left parenthesis in θ.
>
> **Proof**: Here we also proceed by induction on the number of instances of (2)-(7) used to construct the formula. Clearly, the Lemma holds for atomic formulas, since they have no parentheses. Suppose, then, that the Lemma holds for formulas constructed with n or fewer instances of (2)-(7), and let θ be constructed with n+1 instances. If the last clause applied was (3)-(5), then the Lemma holds since θ itself begins with a left parenthesis and ends with the matching right parenthesis. If the last clause applied was (2), then θ is $\neg\psi$, and

the induction hypothesis applies to ψ. Similarly, if the last clause applied was (6) or (7), then θ consists of a quantifier, a variable, and a formula to which we can apply the induction hypothesis. It follows that the Lemma holds for θ.

Lemma 4. Each formula contains at least one atomic formula.

The proof proceeds by induction on the number of instances of (2)-(7) used to construct the formula, and we leave it as an exercise.

Theorem 5. Let α, β be nonempty sequences of characters on our alphabet, such that $\alpha\beta$ (i.e α followed by β) is a formula. Then α is *not* a formula.

Proof: By Theorem 1 and Lemma 3, if α contains a left parenthesis, then the right parenthesis that matches the leftmost left parenthesis in $\alpha\beta$ comes at the end of $\alpha\beta$, and so the matching right parenthesis is in β. So, α has more left parentheses than right parentheses. By Theorem 1, α is not a formula. So now suppose that α does not contain any left parentheses. By Lemma 2, $\alpha\beta$ consists of a string of zero or more unary markers followed by either an atomic formula or a formula produced using a binary connective, via one of clauses (3)-(5). If the latter formula was produced via one of clauses (3)-(5), then it begins with a left parenthesis. Since α does not contain any parentheses, it must be a string of unary markers. But then α does not contain any atomic formulas, and so by Lemma 4, α is not a formula. The only case left is where $\alpha\beta$ consists of a string of unary markers followed by an atomic formula, either in the form t_1=t_2 or $Pt_1 \ldots t_n$. Again, if α just consisted of unary markers, it would not be a formula, and so α must consist of the unary markers that start $\alpha\beta$, followed by either t_1 by itself, t_1= by itself, or the predicate letter P, and perhaps some (but not all) of the terms $t_1, \ldots, t_n$. In the first two cases, α does not contain an atomic formula, by the policy that the categories do not overlap. Since P is an n-place predicate letter, by the policy that the predicate letters are distinct, P is not an m-place predicate letter for any $m \neq n$. So the part of α that consists of P followed by the terms is not an atomic formula. In all of these cases, then, α does not contain an atomic formula. By Lemma 4, α is not a formula.

We are finally in position to show that there is no amphiboly in our language.

Theorem 6. Let θ be any formula of $\mathcal{L}1K$=. If θ is not atomic, then there is one and only one among (2)-(7) that was the last clause applied to construct θ. That is, θ could not be produced by two different clauses. Moreover, no formula produced by clauses (3)-(7) is atomic.

Proof: By Clause (8), either is θ atomic or it was produced by one of clauses (2)-(7). Thus, the first symbol in θ must be either a predicate letter, a term, a unary marker, or a left

parenthesis. If the first symbol in θ is a predicate letter or term, then θ is atomic. In this case, θ was not produced by any of (2)-(7), since all such formulas begin with something other than a predicate letter or term. If the first symbol in θ is a negation sign "¬", then was θ produced by clause (2), and not by any other clause (since the other clauses produce formulas that begin with either a quantifier or a left parenthesis). Similarly, if θ begins with a universal quantifier, then it was produced by clause (6), and not by any other clause, and if θ begins with an existential quantifier, then it was produced by clause (7), and not by any other clause. The only case left is where θ begins with a left parenthesis. In this case, it must have been produced by one of (3)-(5), and not by any other clause. We only need to rule out the possibility that θ was produced by more than one of (3)-(5). To take an example, suppose that θ was produced by (3) and (4). Then θ is $(\psi_1 \,\&\, \psi_2)$ and θ is also $(\psi_3 \vee \psi_4)$, where ψ_1, ψ_2, ψ_3, and ψ_4 are themselves formulas. That is, $(\psi_1 \,\&\, \psi_2)$ is the very same formula as $(\psi_3 \vee \psi_4)$. By Theorem 5, ψ_1 cannot be a proper part of ψ_3, nor can ψ_3 be a proper part of ψ_1. So ψ_1 must be the same formula as ψ_3. But then "&" must be the same symbol as "$\vee$", and this contradicts the policy that all of the symbols are different. So θ was not produced by both Clause (3) and Clause (4). Similar reasoning takes care of the other combinations.

This result is sometimes called "unique readability". It shows that each formula is produced from the atomic formulas via the various clauses in exactly one way. If θ was produced by clause (2), then its *main connective* is the initial "¬". If θ was produced by clauses (3), (4), or (5), then its *main connective* is the introduced "&", "$\vee$", or "$\rightarrow$", respectively. If θ was produced by clauses (6) or (7), then its *main connective* is the initial quantifier. I apologize for the tedious details. I included them to indicate the level of precision and rigor for the syntax.

3. Deduction

We now introduce a *deductive system*, *D*, for our languages. As above, we define an *argument* to be a non-empty collection of formulas in the formal language, one of which is designated to be the *conclusion*. If there are any other formulas in the argument, they are its *premises*. By convention, we use "Γ", "Γ'", "Γ_1", etc, to range over sets of formulas, and we use the letters "φ", "ψ", "θ", uppercase or lowercase, with or without subscripts, to range over single formulas. We write "Γ, Γ'" for the union of Γ and Γ', and "Γ, φ" for the union of Γ with $\{\varphi\}$.

We write an argument in the form $<\Gamma, \varphi>$, where Γ is the set of premises and φ is the conclusion. Remember that Γ may be empty. We write $\Gamma \vdash \varphi$ to indicate that φ is deducible from Γ, or, in other words, that the argument $<\Gamma, \varphi>$ is deducible in *D*. We may write $\Gamma \vdash_D \varphi$ to emphasize the deductive system *D*. We write $\vdash \varphi$ or $\vdash_D \varphi$ to indicate that φ can be deduced (in *D*) from the empty set of premises.

The rules in *D* are chosen to match logical relations concerning the English analogues of the logical terminology in the language. Again, we define the deducibility relation by recursion. We start with a rule of assumptions:

> **(As)** If φ is a member of Γ, then $\Gamma \vdash \varphi$.

We thus have that $\{\varphi\} \vdash \varphi$; each premise follows from itself. We next present two clauses for each connective and quantifier. The clauses indicate how to "introduce" and "eliminate" formulas in which each symbol is the main connective.

First, recall that "&" is an analogue of the English connective "and". Intuitively, one can deduce a formula in the form (θ & ψ) if one has deduced θ and one has deduced ψ. Conversely, one can deduce θ from (θ & ψ) and one can deduce ψ from (θ & ψ):

> **(&I)** If $\Gamma_1 \vdash \theta$ and $\Gamma_2 \vdash \psi$, then $\Gamma_1, \Gamma_2 \vdash$ (θ & ψ).

> **(&E)** If $\Gamma \vdash$ (θ & ψ) then $\Gamma \vdash \theta$; and if $\Gamma \vdash$ (θ & ψ) then $\Gamma \vdash \psi$.

The name "&I" stands for "&-introduction"; "&E" stands for "&-elimination".

Since, the symbol "$\vee$" corresponds to the English "or", $(\theta \vee \psi)$ should be deducible from θ, and $(\theta \vee \psi)$ should also be deducible from ψ:

> ($\vee$**I**) If $\Gamma \vdash \theta$ then $\Gamma \vdash (\varphi \vee \psi)$; if $\Gamma \vdash \psi$ then $\Gamma \vdash (\theta \vee \psi)$.

The elimination rule is a bit more complicated. Suppose that "θ or ψ" is true. Suppose also that φ follows from θ and that φ follows from ψ. One can reason that if θ is true, then φ is true. If instead ψ is true, we still have that φ is true. So either way, φ must be true.

> ($\vee$**E**) If $\Gamma_1 \vdash (\theta \vee \psi)$, $\Gamma_2, \theta \vdash \varphi$ and $\Gamma_3, \psi \vdash \varphi$, then $\Gamma_1, \Gamma_2, \Gamma_3 \vdash \varphi$.

For the next clauses, recall that the symbol "$\rightarrow$" is an analogue of the English "if . . . then . . . " construction. If one knows, or assumes $(\theta \rightarrow \psi)$ and also knows, or assumes θ, then one can conclude ψ. Conversely, if one deduces ψ from an assumption θ, then one can conclude that $(\theta \rightarrow \psi)$.

> ($\rightarrow$**I**) If $\Gamma, \theta \vdash \psi$, then $\Gamma \vdash (\theta \rightarrow \psi)$.

> ($\rightarrow$**E**) If $\Gamma_1 \vdash (\theta \rightarrow \psi)$ and $\Gamma_2 \vdash \theta$, then $\Gamma_1, \Gamma_2 \vdash \psi$.

Our next clauses are for the negation sign, "¬". The underlying idea is that a formula ψ is inconsistent with its negation $\neg\psi$. They cannot both be true. We call a pair of formulas ψ, $\neg\psi$ *contradictory opposites*. If one can deduce such a pair from an assumption θ, then one can conclude that θ is false, or, in other words, one can conclude $\neg\theta$.

> **(¬I)** If $\Gamma_1, \theta \vdash \psi$ and $\Gamma_2, \theta \vdash \neg\psi$, then $\Gamma_1, \Gamma_2 \vdash \neg\theta$.

There is some controversy over the other rule for the negation sign.

By (As), we have that $\{A,\neg A\} \vdash A$ and $\{A,\neg A\} \vdash \neg A$. So by ¬I we have that $\{A\} \vdash \neg\neg A$. However, we do not have the converse yet. Intuitively, $\neg\neg\theta$ corresponds to "it is not the case that it is not the case that" . One might think that this last is equivalent to θ, and we have a rule to that effect:

> **(DNE)** If $\Gamma \vdash \neg\neg\theta$, then $\Gamma \vdash \theta$.

The name DNE stands for "double-negation elimination". This inference is rejected by philosophers and mathematicians who do not hold that each meaningful sentence is either true or not true. *Intuitionistic logic* does not sanction the inference in question (see, for example Dummett [1977]), but, again, classical logic does.

To illustrate the parts of the deductive system *D* presented thus far, I show that $\vdash(A \vee \neg A)$:

> (i) $\{\neg(A \vee \neg A), A\} \vdash \neg(A \vee \neg A)$, by (As)
>
> (ii) $\{\neg(A \vee \neg A), A\} \vdash A$, by clause (As).
>
> (iii) $\{\neg(A \vee \neg A), A\} \vdash (A \vee \neg A)$, by (∨I), from (ii).
>
> (iv) $\{\neg(A \vee \neg A)\} \vdash \neg A$, by (¬I), from (i) and (iii).
>
> (v) $\{\neg(A \vee \neg A), \neg A\} \vdash \neg(A \vee \neg A)$, by (As)
>
> (vi) $\{\neg(A \vee \neg A), \neg A\} \vdash \neg A$, by (As)
>
> (vii) $\{\neg(A \vee \neg A), \neg A\} \vdash (A \vee \neg A)$, by (∨I), from (vi).
>
> (viii) $\{\neg(A \vee \neg A)\} \vdash \neg\neg A$, by (¬I), from (v) and (vii).
>
> (ix) $\vdash \neg\neg(A \vee \neg A)$, by (¬I), from (iv) and (viii).

(x) $\vdash (A \vee \neg A)$, by (DNE), from (ix).

The principle $(\theta \vee \neg\theta)$ is sometimes called the *law of excluded middle*. It is not valid in intuitionistic logic.

Let θ, $\neg\theta$ be a pair of contradictory opposites, and let ψ be any formula at all. By (As) we have $\{\theta, \neg\theta, \neg\psi\} \vdash \theta$ and $\{\theta, \neg\theta, \neg\psi\} \vdash \neg\theta$. So by (¬I), $\{\theta, \neg\theta\} \vdash \neg\neg\psi$. So, by (DNE) we have $\{\theta, \neg\theta\} \vdash \psi$. That is, anything at all follows from a pair of contradictory opposites. Some logicians introduce a rule to codify a similar inference:

If $\Gamma_1 \vdash \theta$ and $\Gamma_2 \vdash \neg\theta$, then for any formula ψ, $\Gamma_1, \Gamma_2 \vdash \psi$

The inference is sometimes called *ex falso quodlibet*. Some call it "¬-elimination", but perhaps this stretches the notion of "elimination" a bit. We do not officially include *ex falso quodlibet* as a separate rule in *D*, but as will be shown below (Theorem 10), each instance of it is derivable.

Some logicians object to *ex falso quodlibet*, on the ground that the formula ψ may be *irrelevant* to any of the premises in Γ. Suppose, for example, that one starts with some premises Γ about human nature and facts about certain people, and then deduces both the sentence "Clinton had extra-marital sexual relations" and "Clinton did not have extra-marital sexual relations". One can surely conclude that there is something wrong with premises Γ. But should we be allowed to then deduce *anything at all* from Γ? Should we be allowed to deduce "The economy is sound"?

Deductive systems that demur from *ex falso quodlibet* are part of *relevance logic*. See Anderson and Belnap [1975], Anderson, Belnap, and Dunn [1992], and Tennant [1987]. Deep philosophical issues concerning the nature of logical consequence are involved. Far be it for an article in a philosophy encyclopedia to avoid philosophical issues, but space considerations preclude a fuller treatment of this issue here. Suffice it to note that the inference is sanctioned in systems of *classical logic*, the subject of this article. It is essential to establishing the balance between the deductive system and the semantics (see §5 below).

The next pieces of *D* are the clauses for the quantifiers. Let θ be a formula, *v* a variable, and *t* a term (i.e., a variable or a constant). The define $\theta(v|t)$ to be the result of substituting *t* for each *free* occurrence of *v* in θ. So, if θ is $(Qx \,\&\, \exists xPxy)$, then $\theta(x|c)$ is $(Qc \,\&\, \exists xPxy)$. The last occurrence of *x* is not free (but recall that we avoid using formulas like this).

We have one other nicety to attend to. Suppose that v_1 and v_2 are variables. It may happen that some of the substituted instances of v_2 are bound in $\theta(v_1|v_2)$. When this happens, we say that there is a *clash* of the variables. Suppose, for example, that θ is $\exists y(\neg x = y)$, and so $\theta(x|y)$ is $\exists y(\neg y = y)$. We say that a term *t* is *free for* a variable *v* in θ if either *t* is a constant or there is no clash of variables in $\theta(v|t)$. The idea is

that no substituted instance of t should become a bound variable in $\theta(v|t)$.

A formula in the form $\forall v\ \theta$ is an analogue of the English "for every v, θ holds". So one should be able to infer $\theta(v|t)$ from $\forall v\ \theta$:

> (∀**E**) If $\Gamma \vdash \forall v\ \theta$, then $\Gamma \vdash \theta(v|t)$, provided that t is free for v in θ.

The idea here is that if $\forall v\ \theta$ is true, then θ should hold of t, no matter what t is. We can illustrate the restriction on (∀E) as follows: The sentence $\forall x\ \exists\ y(\neg x{=}y)$ corresponds to an assertion that for every object x, there is an object different from x. This is a coherent, plausible assertion. It is true if and only if the universe has at least two objects. It should follow that no matter what object t may be, something is different from t. However, if we were allowed to substitute the variable y for x, we would conclude $\exists y(\neg y{=}y)$, which says that there is something which is different from itself, a blatant falsehood.

The introduction clause for the universal quantifier is a bit more complicated. Suppose that a formula θ has a variable v free, and that θ has been deduced from a set of premises Γ. If the variable v does not occur free in any member of Γ, then θ will hold no matter which object v may denote. That is, $\forall v\ \theta$ follows.

> (∀**I**) If $\Gamma \vdash \theta$ and the variable v does not occur free in any member of Γ, then $\Gamma \vdash \forall v\theta$.

This introduction rule corresponds to a common inference in mathematics. Suppose that a mathematician says "let n be a natural number" and goes on to show that n has a certain property P, without assuming anything about n (except that it is a natural number). She then reminds the reader that n is "arbitrary", and concludes that P holds for *all* natural numbers. The condition that the variable v not occur in any premise is what guarantees that it is indeed "arbitrary". It could be any object, and so anything we conclude about it holds for all objects.

The existential quantifier is an analogue of the English expression "there exists", or perhaps just "there is". If we have established (or assumed) that a given object t has a given property, then it follows that there is something that has that property. Again, we have to be careful with the syntax, and avoid clashes of variables.

> (∃**I**) If $\Gamma \vdash \theta$, then $\Gamma \vdash \exists v\ \theta'$, where θ' is obtained from θ by substituting the variable v for zero or more occurrences of a term t, provided that (1) if t is a variable, then all of the replaced occurrences of t are free in θ, and (2) all of the substituted occurrences of v are free in θ'.

The provision (1) keeps us from replacing bound variables. The provision (2) comes up only if v is bound by another quantifier in θ. As noted above, we avoid such formulas (since they appear to bind the same occurrence twice).

The elimination rule for ∃ is not quite as simple:

> (**∃E**) If $\Gamma_1 \vdash \exists v\, \theta$ and $\Gamma_2, \theta \vdash \varphi$, then $\Gamma_1, \Gamma_2 \vdash \varphi$, provided that *v* does not occur free in φ, nor in any member of Γ_2.

This elimination rule also corresponds to a common inference. Suppose that a mathematician asserts that there is a natural number with a given property *P*. She then says "let *n* be such a natural number, so that *Pn*", and goes on to establish a sentence φ, which does not mention the number *n*. If the derivation of φ does not invoke anything about *n* (other than the assumption that it has the given property *P*), then *n* could have been any number that has the property *P*. That is, *n* is an *arbitrary* number with property *P*. It does not matter which number *n* is. Since φ does not mention *n*, it follows from the assertion that something has property *P*. The provisions added to (∃E) are to guarantee that *x* is "arbitrary".

As noted in the previous section, some authors introduce different letters for bound variables and (what amounts to) free variables. This makes the syntax slightly more complex, but simplifies the provisions on some of the rules of inference. Writers of logic books often face tradeoffs like this.

The final items are the rules for the identity sign "=". The introduction rule is about a simple as can be:

> **(=I)** $\Gamma \vdash t{=}t$, where *t* is any term.

This "inference" corresponds to the truism that everything is identical to itself. The elimination rule corresponds to a principle that if *a* is identical to *b*, then anything true of *a* is also true of *b*, again paying attention to clashes of variables.

> **(=E)** If $\Gamma_1 \vdash t_1{=}t_2$ and $\Gamma_2 \vdash \theta$, then $\Gamma_1, \Gamma_2 \vdash \theta'$, where θ' is obtained from θ by replacing zero or more occurrences of t_1 with t_2, provided that no bound variables are replaced, and if t_2 is a variable, then all of its substituted occurrences are free.

The rule (=E) indicates a certain restriction in the expressive resources of our language. Suppose, for example, that Harry is identical to Donald (since his mischievous parents gave him two names). It would not follow from this and "Dick knows that Harry is wicked" that "Dick knows that Donald is wicked", for the reason that Dick might not know that Harry is identical to Donald. Contexts like this, in which identicals cannot safely be substituted for each other, are called "opaque". We assume that our language $\mathcal{L}1K{=}$ has no opaque contexts.

One final clause completes the description of the deductive system *D*:

> **(*)** That's all folks. $\Gamma \vdash \theta$ only if θ follows from members of Γ by the above rules.

Again, this clause allows proofs by induction on the rules used to establish an inference. If a property of arguments holds of all instances of (As) and (=I), and if the other rules preserve the property, then every argument that is deducible in *D* enjoys the property in question.

Before moving on to the model theory for $\mathcal{L}1K$=, we pause to note a few features of the deductive system.

Lemma 7. Suppose that $\Gamma \vdash_D \varphi$, and let v' be a variable that does not occur free in φ or in any member of Γ. Assume that v' is free for v in φ and in every member of Γ. Let Γ' be $\{\theta(v|v') \mid \theta \in \Gamma\}$. That is, Γ' is the result of replacing every free occurrence of a variable v with v' in every member of Γ. Then $\Gamma' \vdash_D \varphi(v|v')$.

Proof: The proof of this lemma is tedious, but we give its essentials. We proceed by induction on the number of rules that were used to arrive at $\Gamma \vdash \varphi$. Suppose that n>0 is a natural number, and that the lemma holds for any argument that was derived using fewer than n rules, and suppose that $\Gamma \vdash \varphi$ using exactly n rules. If n=1, then the rule applied is either (As) or (=I). In this case, $\Gamma' \vdash \varphi(v|v')$ by the same rule. If the last rule applied is (&I), then φ has the form $(\theta \,\&\, \psi)$, and we have $\Gamma_1 \vdash \theta$ and $\Gamma_2 \vdash \psi$, with $\Gamma = \Gamma_1, \Gamma_2$. We apply the induction hypothesis to the deductions of θ and ψ, and then apply (&I) to the result. If the last rule applied was (&E), we have two sub-cases, but they are symmetric. We have $\Gamma \vdash (\varphi \,\&\, \psi)$. There are two slight complications here: the new variable v' may occur free in ψ and it may not be free for v in ψ. In either case, first pick a new variable u that does not occur (free or bound) in $(\varphi \,\&\, \psi)$ or in any member of Γ. Now apply the induction hypothesis, substituting u for v' in the deduction $\Gamma \vdash (\varphi \,\&\, \psi)$. Since v' does not occur free in φ or in any member of Γ, those formulas are left unchanged. The maneuver removes any free occurrences of v' from the subformula ψ. Now apply the induction hypothesis to the result, substituting v' for v, and then apply (&E). The remaining cases are similar.

Theorem 8. The rule of Weakening. If $\Gamma_1 \vdash \varphi$ and $\Gamma_1 \subseteq \Gamma_2$, then $\Gamma_2 \vdash \varphi$.

Proof: Again, we proceed by induction on the number of rules that were used to arrive at $\Gamma_1 \vdash \varphi$. Suppose that n>0 is a natural number, and that the theorem holds for any argument that was derived using fewer than n rules. Suppose that $\Gamma_1 \vdash \varphi$ using exactly n rules. If n=1, then the rule is either (As) or (=I). In these cases, $\Gamma_2 \vdash \varphi$ by the same rule. If the last rule applied was (&I), then φ has the form $(\theta \& \psi)$, and we have $\Gamma_3 \vdash \theta$ and $\Gamma_4 \vdash \psi$, with $\Gamma_1 = \Gamma_3, \Gamma_4$. We apply the induction hypothesis to the deductions of θ and ψ, to get $\Gamma_2 \vdash \theta$ and $\Gamma_2 \vdash \psi$. and then apply (&I) to the result to get $\Gamma_2 \vdash \varphi$. Most of the other cases are

exactly like this. Slight complications arise only in the rules (∀I) and (∃E), because there we have to pay attention to the conditions for the rules. Starting with (∃E), we have $\Gamma_3 \vdash \exists v \theta$ and $\Gamma_4, \theta \vdash \varphi$, with Γ_1 being Γ_3, Γ_4, and v not free in φ, nor in any member of Γ_4. We apply the induction hypothesis to get $\Gamma_2 \vdash \exists v \theta$, and then (∃E) to end up with $\Gamma_2 \vdash \varphi$. Suppose that the last rule applied to get $\Gamma_1 \vdash \varphi$ is (∀I). So φ is a formula in the form $\forall v \theta$, and we have $\Gamma_1 \vdash \theta$ and the variable v does not occur free in any member of Γ_1. The problem is that v may occur free in a member of Γ_2, and so we cannot just invoke the induction hypothesis and apply (∀I) to the result. Let v' be a variable that does not occur (free or bound) in θ or in any member of Γ_2, and let Γ' be the result of substituting v for every free occurrence of v in Γ_2. Since v does not occur free in any member of Γ_1, we still have $\Gamma_1 \subseteq \Gamma$. The induction hypothesis gives us $\Gamma' \vdash \theta$, and now we apply (∀I) to get $\Gamma' \vdash \varphi$. We now apply Lemma 7, substituting v for the new variable v. The result is $\Gamma_2 \vdash \varphi$.

Theorem 8 allows us to add on premises at will. It follows that $\Gamma \vdash \varphi$ if and only if there is a subset $\Gamma' \subseteq \Gamma$ such that $\Gamma' \vdash \varphi$. By clause (*), all derivations are established in a finite number of steps. So we have

Theorem 9. $\Gamma \vdash \varphi$ if and only if there is a finite $\Gamma' \subseteq \Gamma$ such that $\Gamma' \vdash \varphi$.

Theorem 10. The rule of ex falso quodlibet is a "derived rule" of *D*. That is, if $\Gamma_1 \vdash \theta$ and $\Gamma_2 \vdash \neg\theta$, then $\Gamma_1, \Gamma_2 \vdash \psi$, for any formula ψ.

Proof: Suppose that $\Gamma_1 \vdash \theta$ and $\Gamma_2 \vdash \neg\theta$. Then by Theorem 8, $\Gamma_1, \neg\psi \vdash \theta$, and $\Gamma_2, \neg\psi \vdash \theta$. So by (¬I), $\Gamma_1, \Gamma_2 \vdash \neg\neg\psi$. By (DNE), $\Gamma_1, \Gamma_2 \vdash \psi$.

Theorem 11. The rule of Cut. If $\Gamma_1 \vdash \psi$ and $\Gamma_2, \psi \vdash \theta$, then $\Gamma_1, \Gamma_2 \vdash \theta$.

Proof: Suppose $\Gamma_1 \vdash \psi$ and $\Gamma_2, \psi \vdash \theta$. We proceed by induction on the number of rules used to establish $\Gamma_2, \psi \vdash \theta$. Suppose that *n* is a natural number, and that the theorem holds for any argument that was derived using fewer than *n* rules. Suppose that $\Gamma_2, \psi \vdash \theta$ was derived using exactly *n* rules. If the last rule used was (=I), then $\Gamma_1, \Gamma_2 \vdash \theta$ is also an instance of (=I). If $\Gamma_2, \psi \vdash \theta$ is an instance of (As), then either θ is ψ, or θ is a member of Γ_2. In the former case, we have $\Gamma_1 \vdash \theta$ by supposition, and get $\Gamma_1, \Gamma_2 \vdash \theta$ by Weakening (Theorem 8). In the latter case, $\Gamma_1, \Gamma_2 \vdash \theta$ is itself an instance of (As). Suppose that $\Gamma_2, \psi \vdash \theta$ was obtained using (&E). Then we have $\Gamma_2, \psi \vdash (\theta \& \varphi)$. The induction hypothesis gives us $\Gamma_1, \Gamma_2 \vdash (\theta \& \varphi)$, and (&E) produces $\Gamma_1, \Gamma_2 \vdash \theta$. The remaining cases

are similar.

Theorem 11 allows us to chain together inferences. This fits the practice of establishing theorems and lemmas and then using those theorems and lemmas later, at will. The cut principle is, I think, essential to reasoning. In some logical systems, the cut principle is a deep theorem. The system here was designed to make the proof of Theorem 11 straightforward.

If $\Gamma \vdash_D \theta$, then we say that the formula θ is a *deductive consequence* of the set of formulas Γ, and that the argument $<\Gamma,\theta>$ is *deductively valid*. A formula θ is a *logical theorem*, or a *deductive logical truth*, if $\vdash_D \theta$. That is, θ is a logical theorem if it is a deductive consequence of the empty set. A set Γ of formulas is *consistent* if there is no formula θ such that $\Gamma \vdash_D \theta$ and $\Gamma \vdash_D \neg\theta$. That is, a set is consistent if it does not entail a pair of contradictory opposite formulas.

Theorem 12. A set Γ is consistent if and only if there is a formula θ such that it is not the case that $\Gamma \vdash \theta$.

Proof: Suppose that Γ is consistent and let θ be any formula. Then either it is not the case that $\Gamma \vdash \theta$ or it is not the case that $\Gamma \vdash \neg\theta$. For the converse, suppose that Γ is inconsistent and let ψ be any formula. We have that there is a formula such that both $\Gamma \vdash \theta$ and $\Gamma \vdash \neg \theta$. By ex falso quodlibet (Theorem 10), $\Gamma \vdash \psi$.

Define a set Γ of formulas of the language $\mathcal{L}1K$= to be *maximally consistent* if Γ is consistent and for every formula θ of $\mathcal{L}1K$=, if θ is not in Γ, then Γ, θ is inconsistent. In other words, Γ is maximally consistent if Γ is consistent, and adding any formula in the language not already in Γ renders it inconsistent. Notice that if Γ is maximally consistent then $\Gamma \vdash \theta$ if and only if θ is in Γ.

Theorem 13. The Lindenbaum Lemma. Let Γ be any consistent set of formulas of $\mathcal{L}1K$=. Then there is a set Γ' of formulas of $\mathcal{L}1K$= such that $\Gamma \subseteq \Gamma'$ and Γ' is maximally consistent.

Proof: Although this theorem holds in general, we assume here that the set *K* of non-logical terminology is either finite or denumerably infinite (i.e., the size of the natural numbers, usually called $\aleph_0$). It follows that there is an enumeration $\theta_0, \theta_1, \ldots$ of the formulas of $\mathcal{L}1K$=, such that every formula of $\mathcal{L}1K$= eventually occurs in the list. Define a sequence of sets of formulas, by recursion, as follows: Γ_0 is Γ; for each natural number *n*, if Γ_n, θ_n is consistent, then let $\Gamma_{n+1} = \Gamma_n, \theta_n$. Otherwise, let $\Gamma_{n+1} = \Gamma_n$. Let Γ' be the union of all of the sets Γ_n. Intuitively, the idea is to go through the formulas of $\mathcal{L}1K$=, throwing each one into Γ' if doing so produces a consistent set. Notice that each Γ_n is consistent. Suppose that Γ' is inconsistent. Then there is a formula θ such that $\Gamma' \vdash \theta$ and

$\Gamma' \vdash \neg\theta$. By Theorem 9 and Weakening (Theorem 8), there is finite subset Γ'' of Γ' such that $\Gamma'' \vdash \theta$ and $\Gamma'' \vdash \neg\theta$. Because Γ'' is finite, there is a natural number *n* such that every member of Γ'' is in Γ_n. So, by Weakening again, $\Gamma_n \vdash \theta$ and $\Gamma_n \vdash \neg\theta$. So Γ_n is inconsistent, which contradicts the construction. So Γ' is consistent. Now suppose that a formula θ is not in Γ'. We have to show that Γ', θ is inconsistent. The formula θ must occur in the aforementioned list of formulas; say that θ is θ_m. Since θ_m is not in Γ', then it is not in Γ_{m+1}. This happens only if Γ_m, θ_m is inconsistent. So a pair of contradictory opposites can be deduced from Γ_m, θ_m. By Weakening, a pair of contradictory opposites can be deduced from Γ', θ_m. So Γ', θ_m is inconsistent. Thus, Γ' is maximally consistent.

Notice that this proof uses a principle corresponding to the law of excluded middle. In the construction of Γ', we assumed that, at each stage, either Γ_n is consistent or it is not. Intuitionists, who demur from excluded middle, do not accept the Lindenbaum lemma (see Shapiro [1988]).

4. Semantics

Let *K* be a set of non-logical terminology. An *interpretation* for the language $\mathcal{L}1K$= is a structure *M* = <*d*,*I*>, where *d* is a non-empty set, called the *domain-of-discourse*, or simply the *domain*, of the interpretation, and *I* is an *interpretation function*. Informally, the domain is what we interpret the language $\mathcal{L}1K$= to be about. It is what the variables range over. The interpretation function assigns appropriate extensions to the non-logical terms. In particular,

> If *c* is a constant in *K*, then *I*(*c*) is a member of the domain *d*.

> If P^0 is a zero-place predicate letter in *K*, then $I(P^0)$ is a truth value, either truth or falsehood.

> If Q^1 is a one-place predicate letter in *K*, then *I*(*Q*) is a subset of *d*. Intuitively, *I*(*Q*) is the set of members of the domain that the predicate *Q* holds of. If *Q* represents "red", then *I*(*Q*) might be the red members of the domain.

> If R^2 is a two-place predicate letter in *K*, then *I*(*R*) is a set of ordered pairs of members of *d*. Intuitively, *I*(*R*) is the set of pairs of members of the domain that the relation *R* holds between. If *R* represents "love", then *I*(*R*) might consist of the pairs <*a*,*b*>, such that *a* loves *b*.

> In general, if S^n is an *n*-place predicate letter in *K*, then *I*(*S*) is a set of ordered *n*-tuples of members of *d*.

Define *s* to be a *variable-assignment*, or simply an *assignment*, on an interpretation *M*, if *s* is a function

from the variables to the domain *d* of *M*. The role of variable-assignments is to assign denotations to the *free* variables of open formulas. (In a sense, the quantifiers determine the "meaning" of the bound variables.) Logical systems that dispense with free variables do not need variable-assignments, but some other device is employed.

We now define a relation of *satisfaction* between interpretations, variable-assignments, and formulas of $\mathcal{L}1K$=. If φ is a formula of $\mathcal{L}1K$=, *M* is an interpretation for $\mathcal{L}1K$=, and *s* is a variable-assignment on *M*, then we write $M,s \vDash \varphi$ for *M satisfies* φ *under the assignment s*. The idea is that $M,s \vDash \varphi$ is an analogue of "φ comes out true when interpreted as in *M* via *s*".

Let *t* be a term of $\mathcal{L}1K$=. We define the *denotation* of *t* in *M* under *s*, in terms of the interpretation function and variable-assignment:

> If *c* is a constant, then $D_{M,s}(c)$ is $I(c)$, and if *v* is a variable, then $D_{M,s}(v)$ is $s(v)$.

That is, the interpretation *M* assigns denotations to the constants, while the variable-assignment assigns denotations to the (free) variables. If the language contained function symbols, the denotation function would be defined by recursion.

We now proceed by recursion on the complexity of the formulas of $\mathcal{L}1K$=.

> If t_1 and t_2 are terms, then $M,s \vDash t_1=t_2$ if and only if $D_{M,s}(t_1)$ is the same as $D_{M,s}(t_2)$.

This is about as straightforward as it gets. An identity $t_1=t_2$ comes out true if and only if the terms t_1 and t_2 denote the same thing.

> If P^0 is a zero-place predicate letter in *K*, then $M,s \vDash P$ if and only if $I(P)$ is truth.

> If S^n is an *n*-place predicate letter in *K* and $t_1, \ldots, t_n$ are terms, then $M,s \vDash St_1 \ldots t_n$ if and only if the *n*-tuple $\langle D_{M,s}(t_1), \ldots, D_{M,s}(t_n)\rangle$ is in $I(S)$.

This takes care of the atomic formulas. We now proceed to the compound formulas of the language, following the meanings of the English counterparts of the logical terminology.

> $M,s \vDash \neg\theta$ if and only if it is not the case that $M,s \vDash \theta$.

> $M,s \vDash (\theta \& \psi)$ if and only if both $M,s \vDash \theta$ and $M,s \vDash \psi$.

> $M,s \vDash (\theta \vee \psi)$ if and only if either $M,s \vDash \theta$ or $M,s \vDash \psi$.

$M,s \vDash (\theta \rightarrow \psi)$ if and only if either it is not the case that $M,s \vDash \theta$, or $M,s \vDash \psi$.

$M,s \vDash \forall v\theta$ if and only if $M,s' \vDash \theta$, for every assignment s' that agrees with s except possibly at the variable v.

The idea here is that $\forall v\theta$ comes out true if and only if θ comes out true no matter what is assigned to the variable v. The final clause is similar.

$M,s \vDash \exists v\theta$ if and only if $M,s' \vDash \theta$, for some assignment s' that agrees with s except possibly at the variable v.

So $\exists v\theta$ comes out true if there is an assignment to v that makes θ true.

Theorem 6, unique readability, assures us that this definition is coherent. At each stage in breaking down a formula, there is exactly one clause to be applied, and so we never get contradictory verdicts concerning satisfaction.

As indicated, the role of variable-assignments is to give denotations to the free variables. We now show that variable-assignments play no other role.

Theorem 14. For any formula θ, if s_1 and s_2 agree on the free variables in θ, then $M,s_1 \vDash \theta$ if and only if $M,s_2 \vDash \theta$.

Proof: We proceed by induction on the complexity of the formula θ. The theorem clearly holds if θ is atomic, since in those cases only the values of the variable-assignments at the variables in θ figure in the definition. Assume, then, that the theorem holds for all formulas less complex than θ. And suppose that s_1 and s_2 agree on the free variables of θ. Assume, first, that θ is a negation, $\neg\psi$. Then, by the induction hypothesis, $M,s_1 \vDash \psi$ if and only if $M,s_2 \vDash \psi$. So, by the clause for negation, $M,s_1 \vDash \neg\psi$ if and only if $M,s_2 \vDash \neg\psi$. The cases where the main connective in θ is a binary connectives are also straightforward. Suppose that θ is $\exists v\psi$, and that $M,s_1 \vDash \exists v\psi$. Then there is an assignment $s_1{}'$ that agrees with s_1 except possibly at v such that $M,s_1{}' \vDash \psi$. Let $s_2{}'$ be the assignment that agrees with s_2 on the free variables not in and agrees with $s_1{}'$ on the others. Then, by the induction hypothesis, $M,s_2{}' \vDash \psi$. Notice that $s_2{}'$ agrees with s_2 on every variable except possibly v. So $M,s_2 \vDash \exists v\psi$. The converse is the same, and the case where θ begins with a universal quantifier is similar.

Recall that a sentence is a formula with no free variables. So by Theorem 14, if θ is a sentence, and s_1, s_2, are any two variable-assignments, then $M,s_1 \vDash \theta$ if and only if $M,s_2 \vDash \theta$. So we can just write $M \vDash \theta$ if

$M,s \vDash \theta$ for some, or all, variable-assignments s.

Suppose that $K' \subseteq K$ are two sets of non-logical terms. If $M = \langle d,I \rangle$ is an interpretation of $\mathcal{L}1K$=, then we define the *restriction* of M to $\mathcal{L}1K'$ be the interpretation $M'=\langle d,I' \rangle$ such that I' is the restriction of I to K'. That is, M and M' have the same domain and agree on the non-logical terminology in K'. A straightforward induction establishes the following:

> **Theorem 15**. If M' is the restriction of M to $\mathcal{L}1K'$, then for every formula θ of $\mathcal{L}1K'$, if s is any variable-assignment, $M,s \vDash \theta$ if and only if $M',s \vDash \theta$.

> **Theorem 16.** If two interpretations M_1, M_2 have the same domain and agree on the non-logical terminology of a formula θ, then if s is any variable-assignment, $M_1,s \vDash \theta$ if and only if $M_2,s \vDash \theta$.

In short, the satisfaction of a formula θ only depends on the domain of discourse, the interpretation of the non-logical terminology in θ, and the assignments to the free variables in θ.

We say that an argument $\langle \Gamma,\theta \rangle$ is *semantically valid*, or just *valid*, written $\Gamma \vDash \theta$, if for every interpretation M of the language and any variable-assignment s on M, if $M,s \vDash \psi$, for every member ψ of Γ, then $M,s \vDash \theta$. If $\Gamma \vDash \theta$, we also say that θ is a *logical consequence*, or *semantic consequence*, or *model-theoretic consequence* of Γ. The definition corresponds to the informal idea that an argument is valid if it is not possible for its premises to all be true and its conclusion false. Our definition of logical consequence also sanctions the common thesis that a valid argument is truth-preserving--to the extent that satisfaction represents truth. Officially, an argument in $\mathcal{L}1K$= is valid if its conclusion comes out true under every interpretation of the language in which the premises are true. Validity is the model-theoretic counterpart to deducibility.

A formula θ is *logically true*, or *valid*, if $M,s \vDash \theta$, for every interpretation M and assignment s. A formula is logically true if and only if it is a consequence of the empty set. If θ is logically true, then for any set Γ of formulas, $\Gamma \vDash \theta$. Logical truth is the model-theoretic counterpart of theoremhood.

A formula θ is *satisfiable* if there is an interpretation M and a variable-assignment s on M such that $M,s \vDash \theta$. That is, θ is satisfiable if there is an interpretation and assignment that satisfies it. A set Γ of formulas is satisfiable if there is an interpretation M and a variable-assignment s on M such that $M,s \vDash \theta$, for every formula θ in Γ. If Γ is a set of sentences and if $M \vDash \theta$ for each sentence θ in Γ, then we say that M is a *model of* Γ. So a set of sentences is satisfiable if it has a model. Satisfiability is the model-theoretic counterpart to consistency.

Notice that $\Gamma \vDash \theta$ if and only if the set $\Gamma,\neg\theta$ is not satisfiable. It follows that if a set Γ is not satisfiable, then if θ is any formula, $\Gamma \vDash \theta$. This is a model-theoretic counterpart to ex falso quodlibet (see Theorem 10). We have the following, as an analogue to Theorem 12:

Theorem 17. Let Γ be a set of formulas. The following are equivalent: (a) Γ is satisfiable; (b) there is no formula θ such that both $\Gamma \vDash \theta$ and $\Gamma \vDash \neg\theta$; (c) there is some formula ψ such that it is not the case that $\Gamma \vDash \psi$.

Proof: (a) $\Rightarrow$ (b): Suppose that Γ is satisfiable and let θ be any formula. There is an interpretation *M* and assignment *s* such that *M*,*s* $\vDash \psi$ for every member ψ of Γ. By the clause for negations, we cannot have both *M*,*s* $\vDash \theta$ and *M*,*s* $\vDash \neg\theta$. So either $\langle\Gamma,\theta\rangle$ is not valid or else $\langle\Gamma,\neg\theta\rangle$ is not valid. (b) $\Rightarrow$ (c): This is immediate. (c) $\Rightarrow$ (a): Suppose that it is not the case that $\Gamma \vDash \psi$. Then there is an interpretation *M* and an assignment *s* such that *M*,*s* $\vDash \theta$, for every formula θ in Γ and it is not the case that *M*,*s* $\vDash \psi$. A fortiori, *M*,*s* satisfies every member of Γ, and so Γ is satisfiable.

5. Meta-theory

We now present some results that relate the deductive notions to their model-theoretic counterparts. The first one is probably the most straightforward. We motivated both the various rules of the deductive system *D* and the various clauses in the definition of satisfaction in terms of the meaning of the English counterparts to the logical terminology. So one would expect that an argument is deducible, or deductively valid, only if it is semantically valid.

Theorem 18. Soundness. For any formula θ and set Γ of formulas, if $\Gamma \vdash_D \theta$, then $\Gamma \vDash \theta$.

Proof: We proceed by induction on the number of clauses used to establish $\Gamma \vdash \theta$. So let *n* be a natural number, and assume that the theorem holds for any argument established as deductively valid with fewer than *n* steps. And suppose that $\Gamma \vdash \theta$ was established using exactly *n* steps. If the last rule applied was (=I) then θ is a formula in the form *t*=*t*, and so θ is logically true. A fortiori, $\Gamma \vDash \theta$. If the last rule applied was (As), then θ is a member of Γ, and so of course any interpretation and assignment that satisfies every member of Γ also satisfies θ. Suppose the last rule applied is (&I). So θ has the form $(\varphi \& \psi)$, and we have $\Gamma_1 \vdash \varphi$ and $\Gamma_2 \vdash \psi$, with $\Gamma = \Gamma_1, \Gamma_2$. The induction hypothesis gives us $\Gamma_1 \vDash \varphi$ and $\Gamma_2 \vDash \psi$. Suppose that *M*,*s* satisfies every member of Γ. Then *M*,*s* satisfies every member of Γ_1, and so *M*,*s* satisfies φ. Similarly, *M*,*s* satisfies every member of Γ_2, and so *M*,*s* satisfies ψ. Thus, by the clause for "&" in the definition of satisfaction, *M*,*s* satisfies θ. So $\Gamma \vDash \theta$. Suppose the last clause applied was ($\exists$E). So we have $\Gamma_1 \vDash \exists v\psi$ and $\Gamma_2, \psi \vDash \theta$, where $\Gamma = \Gamma_1, \Gamma_2$, and *v* does not occur free in ψ, nor in any member of Γ_2. By the induction hypothesis, we have $\Gamma_1 \vdash \exists v\psi$ and $\Gamma_2, \psi \vDash \theta$. Let *M* be an interpretation and *s* an assignment such that *M*,*s* satisfies every member of Γ. Then *M*,*s* satisfies every member of

Γ_1, and so $M,s \vDash \exists v\psi$. So there is an assignment s' that agrees with s on every variable except possibly v such that $M,s' \vDash \psi$. We have that M,s satisfies every member of Γ_2. Since v does not occur free in any member of Γ_2, and s agrees with s' on everything else, we have that M,s' satisfies every member of Γ_2, by Theorem 14. So $M,s' \vDash \theta$. Since v does not occur free in θ, and s agrees with s on everything else, we have that $M,s \vDash \theta$, also by Theorem 14. So, in this case, $\Gamma \vDash \theta$. Notice the role of the restrictions on (∃E) here. The other cases are about as straightforward.

Corollary 19. Let Γ be a set of formulas. If Γ is satisfiable, then Γ is consistent.

Proof: Suppose that Γ is satisfiable. So let *M* be an interpretation and *s* an assignment such that *M*,*s* satisfies every member of Γ. Assume that Γ is inconsistent. Then there is a formula θ such that $\Gamma \vdash \theta$ and $\Gamma \vdash \neg\theta$. By soundness (Theorem 18), $\Gamma \vDash \theta$ and $\Gamma \vDash \neg\theta$. So we have that $M,s \vDash \theta$ and $M,s \vDash \neg\theta$. But this is impossible, given the clause for negation in the definition of satisfaction.

Even though the deductive system *D* and the model-theoretic semantics were developed with the meanings of the logical terminology in mind, one should not automatically expect the converse to soundness (or Corollary 19) to hold. For all we know so far, we may not have included enough rules of inference to deduce every valid argument. The converses to soundness and Corollary 19 are among the most interesting results in contemporary mathematical logic. We begin with the latter.

Theorem 20. Completeness. Gödel [1930]. Let Γ be a set of formulas. If Γ is consistent, then Γ is satisfiable.

Proof: The proof of completeness is rather complex. We only sketch it here. Let Γ be a consistent set of formulas of $\mathcal{L}1K$=. Again, we assume for simplicity that the set *K* of non-logical terminology is either finite or countably infinite (although the theorem holds even if *K* is uncountable). The task at hand is to find an interpretation *M* and a variable-assignment *s* on *M*, such that *M*,*s* satisfies every member of Γ. Consider the language obtained from $\mathcal{L}1K$= by adding a denumerably infinite stock of new individual constants c_0, c_1, . . . We stipulate that the constants, c_0, c_1, . . . , are all different from each other and none of them occur in *K*. We build an interpretation of the language from the language itself, using some of the constants as members of the domain of discourse. Let θ_0, θ_1, . . . be an enumeration of the formulas of the expanded language, so that each formula occurs in the list eventually. Let *x* be any variable, and define a sequence Γ_0, Γ_1, . . . of sets of formulas (of the expanded language) by recursion as follows: $\Gamma_0 = \Gamma$; and $\Gamma_{n+1} = \Gamma_n,(\exists x\theta_n \rightarrow \theta_n(x|c_i))$, where c_i is the first constant in the above list that does not occur in θ_n or in any member of Γ_n. The underlying idea here is that if $\exists x\theta_n$is true, then c_i is to be one such *x*. Let Γ be the union of the sets Γ_n.

I sketch a proof that Γ' is consistent. Suppose that Γ' is inconsistent. By Theorem 9, there is a finite subset of Γ that is inconsistent, and so one of the sets Γ_m is inconsistent. By hypothesis, $\Gamma_0 = \Gamma$ is consistent. Let *n* be the smallest number such that Γ_n is consistent, but $\Gamma_{n+1} = \Gamma_n, (\exists x\theta_n \rightarrow \theta_n(x|c_i))$ is inconsistent. By (¬I), we have that

(1) $\Gamma_n \vdash \neg(\exists x\theta_n \rightarrow \theta_n(x|c_i))$.

By *ex falso quodlibet* (Theorem 10), $\Gamma_n, \neg\exists x\theta_n, \exists x\theta_n \vdash \theta_n(x|c_i)$. So by (→I), $\Gamma_n, \neg\exists x\theta_n \vdash (\exists x\theta_n \rightarrow \theta_n(x|c_i))$. From this and (1), we have $\Gamma_n \vdash \neg\neg\exists x\theta_n$, by (¬I), and by (DNE) we have

(2) $\Gamma_n \vdash \exists x\theta_{\mathrm{n}}$.

By (As), $\Gamma_n, \theta_n(x|c_i), \exists x\theta_n \vdash \theta_n(x|c_i)$. So by (→I), $\Gamma_n, \theta_n(x|c_i) \vdash (\exists x\theta_n \rightarrow \theta_n(x|c_i))$. From this and (1), we have $\Gamma_n \vdash \neg\theta_n(x|c_i)$, by (¬I). Let *v* be a variable that does not occur (free or bound) in θ_n or in any member of Γ_n. By uniform substitution of *v* for c_i, we can turn the derivation of $\Gamma_n \vdash \neg\theta_n(x|c_i)$ into $\Gamma_n \vdash \neg\theta_n(x|v)$. By (∀I), we have

(3) $\Gamma_n \vdash \forall v\neg\theta_n(x|v)$.

By (As) we have $\{\forall v\neg\theta_n(x|v), \theta_n\} \vdash \theta_n$ and by (∀E) we have $\{\forall v\neg\theta_n(x|v), \theta_n\} \vdash \neg\theta_n$. So $\{\forall v\neg\theta_n(x|v), \theta_n\}$ is inconsistent. Let φ be any sentence of the language (so that φ has no free variables). By ex falso quodlibet (Theorem 10), we have that $\{\forall v\neg\theta_n(x|v), \theta_n\} \vdash \varphi$ and $\{\forall v\neg\theta_n(x|v), \theta_n\} \vdash \neg\varphi$. So with (2), we have that $\Gamma_n, \forall v\neg\theta_n(x|v) \vdash \varphi$ and $\Gamma_n, \forall v\neg\theta_n(x|v) \vdash \neg\varphi$, by (∃E). By Cut (Theorem 11), $\Gamma_n \vdash \varphi$ and $\Gamma_n \vdash \neg\varphi$. So Γ_n is inconsistent, contradicting the assumption. So Γ' is consistent.

Applying the Lindenbaum Lemma (Theorem 13), let Γ'' be a maximally consistent set of sentences (of the expanded language) that contains Γ'. So, of course, Γ'' contains Γ. We define an interpretation *M*, and a variable-assignment *s* on *M*, such that *M,s* satisfies every member of Γ''.

If we did not have a sign for identity in the language, we would let the domain of *M* be the collection of new constants $\{c_0, c_1, \ldots\}$. But as it is, there may be a sentence in the form $c_i{=}c_j$, with $i \neq j$, in Γ''. If so, we cannot have both c_i and c_j in the domain of the interpretation. So we define the domain *d* of *M* to be the set $\{c_i \mid$ there is no $j<i$ such that

$c_i=c_j$ is in Γ″}. In other words, a constant c_i is in the domain of *M* if Γ″ does not declare it to be identical to an earlier constant in the list. Notice that for each new constant c_i, there is exactly one $j \leq i$ such that c_j is in *d* and the sentence $c_i=c_j$ is in Γ″.

We now define the interpretation function *I*. Let *a* be any constant in the expanded language. By (=I) and (∃I), Γ″ ⊢ ∃*x x*=*a*, and so ∃*x x*=*a* ∈ Γ″. By the construction of Γ′, there is a sentence in the form (∃*x x*=*a* → c_i=*a*) in Γ″. We have that c_i=*a* is in Γ″. As above, there is exactly one c_j in *d* such that $c_i=c_j$ is in Γ″. Let $I(a)=c_j$. Notice that if c_i is a constant in the domain *d*, then $I(c_i)=c_i$. That is each c_i in *d* denotes itself.

Let *P* be a one-place predicate letter in *K*. Let *I*(*P*) be the set of constants {c_i | c_i is in *d* and the formula *Pc* is in Γ″}. Let *R* be a binary predicate letter in *K*. Let *I*(*R*) be the set of pairs of constants {<c_i,c_j> | c_i is in *d*, c_j is in *d*, and the formula Rc_ic_j is in Γ″}. Three-place predicates, etc. are interpreted similarly. In effect, *I* interprets the non-logical terminology as they are in Γ″.

The variable-assignment is similar. If *v* is a variable, then $s(v)=c_i$, where c_i is the first constant in *d* such that $c_i=v$ is in Γ″.

The final item in this proof is a tedious lemma that for every formula θ in the expanded language, *M*,*s*⊨θ if and only if θ is in Γ″. This proceeds by induction on the complexity of θ. The case where θ is atomic follows from the definitions of *M* (i.e., the domain *d* and the interpretation function *I*) and the variable-assignment *s*. The other cases follow from the various clauses in the definition of satisfaction.

Since Γ ⊆ Γ″, we have that *M,s* satisfies every member of Γ. By Theorem 15, the restriction of *M* to the original language $\mathcal{L}1K$= and *s* also satisfies every member of Γ. Thus Γ is satisfiable.

A converse to Soundness (Theorem 18) is a straightforward corollary:

Theorem 21. For any formula θ and set Γ of formulas, if Γ⊨θ, then $\Gamma \vdash_D \theta$.

Proof: Suppose that Γ⊨θ. Then there is no interpretation *M* and assignment *s* such that *M,s* satisfies every member of Γ but does not satisfy θ. So the set Γ,¬θ is not satisfiable. By Completeness (Theorem 20), Γ,¬θ is inconsistent. So there is a formula ψ such that Γ,¬θ ⊢ ψ and Γ,¬θ ⊢ ¬ψ. By (¬I), Γ ⊢ ¬¬θ, and by (DNE) Γ ⊢ θ.

Our next item is a corollary of Theorem 9, Soundness (Theorem 18), and Completeness:

> **Corollary 22. Compactness.** A set Γ of formulas is satisfiable if and only if every finite subset of Γ is satisfiable.
>
> **Proof:** If M,s satisfies every member of Γ, then M,s satisfies every member of each finite subset of Γ. For the converse, suppose that Γ is not satisfiable. Then we show that some finite subset of Γ is not satisfiable. By Completeness (Theorem 20), Γ is inconsistent. By Theorem 9 (and Weakening), there is a finite subset $\Gamma' \subseteq \Gamma$ such that Γ' is inconsistent. By Corollary 19, Γ' is not satisfiable.

Soundness and completeness together entail that an argument is deducible if and only if it is valid, and a set of formulas is consistent if and only if it is satisfiable. So we can go back and forth between model-theoretic and proof-theoretic notions, transferring properties of one to the other. Compactness holds in the model theory because all derivations use only a finite number of premises.

Recall that in the proof of Completeness (Theorem 20), we made the simplifying assumption that the set K of non-logical constants is either finite or denumerably infinite. The interpretation we produced was itself either finite or denumerably infinite. Thus, we have the following:

> **Corollary 23. Löwenheim-Skolem Theorem.** Let Γ be a satisfiable set of sentences of the language $\mathcal{L}1K=$. If Γ is either finite or denumerably infinite, then Γ has a model whose domain is either finite or denumerably infinite.

In general, let Γ be a satisfiable set of sentences of $\mathcal{L}1K=$, and let κ be the larger of the size of Γ and denumerably infinite. Then Γ has a model whose domain is at most size κ.

There is a stronger version of Corollary 23. Let $M_1=<d_1,I_1>$ and $M_2=<d_2,I_2>$ be interpretations of the language $\mathcal{L}1K=$. Define M_1 to be a *submodel* of M_2 if $d_1 \subseteq d_2$, $I_1(c) = I_2(c)$ for each constant c, and I_1 is the restriction of I_2 to d_1. For example, if R is a binary relation letter in K, then for all a,b in d_1, the pair $<a,b>$ is in $I_1(R)$ if and only if $<a,b>$ is in $I_2(R)$. If we had included function letters among the non-logical terminology, we would also require that d_1 be closed under their interpretations in M_2. Notice that if M_1 is a submodel of M_2, then any variable-assignment on M_1 is also a variable-assignment on M_2.

Say that two interpretations $M_1=<d_1,I_1>$, $M_2=<d_2,I_2>$ are *elementarily equivalent* if one of them is a submodel of the other, and for any formula of the language and any variable-assignment s on the submodel, $M_1,s\models\theta$ if and only if $M_2,s\models\theta$. Notice that if two interpretations are elementarily equivalent, then they satisfy the same sentences.

> **Theorem 25. Downward Löwenheim-Skolem Theorem.** Let $M = <d,I>$ be an interpretation of the language $\mathcal{L}1K=$. Let d_1 be any subset of d, and let κ be the maximum of the size of K, the size of d_1, and denumerably infinite. Then there is a submodel $M' = <d$

$'$,I'> of M such that (1) d' is not larger than κ, and (2) M and M' are elementarily equivalent. In particular, if the set K of non-logical terminology is either finite or denumerably infinite, then any interpretation has an elementarily equivalent submodel whose domain is either finite or denumerably infinite.

Proof: Like completeness, this proof is complex, and we rest content with a sketch. The downward Löwenheim-Skolem theorem invokes the axiom of choice, and indeed, is equivalent to the axiom of choice. So let C be a choice function on the powerset of d, so that for each non-empty subset $e \subseteq d$, $C(e)$ is a member of e. We stipulate that if e is the empty set, then $C(e)$ is $C(d)$.

Let s be a variable-assignment on M, let θ be a formula of $\mathcal{L}1K$=, and let v be a variable. Define the *v-witness of θ over s*, written $w_v(\theta,s)$, as follows: Let q be the set of all elements $c \in d$ such that there is a variable-assignment s' on M that agrees with s on every variable except possibly v, such that $M,s' \vDash \theta$, and $s'(v)=c$. Then $w_v(\theta,s) = C(q)$. Notice that if $M,s \vDash \exists v\theta$, then q is the set of elements of the domain that can go for v in θ. Indeed, $M,s \vDash \exists v\theta$ if and only if q is non-empty. So if $M,s \vDash \exists v\theta$, then $w_v(\theta,s)$ (i.e., $C(q)$) is a chosen element of the domain that can go for v in θ. In a sense, it is a "witness" that verifies $M,s \vDash \exists v\theta$.

If e is a non-empty subset of the domain d, then define a variable-assignment s to be an *e-assignment* if for all variables u, $s(u)$ is in e. That is, s is an e-assignment if s assigns an element of e to each variable. Define $sk(e)$, the *Skolem-hull* of e, to be the set:

> $e \cup \{w_v(\theta,s) \mid \theta$ is a formula in $\mathcal{L}1K$=, v is a variable, and s is an e-assignment$\}$.

That is, the Skolem-Hull of e is the set e together with every v-witness of every formula over every e-assignment. Roughly, the idea is to start with e and then throw in enough elements to make each existentially quantified formula true. But we cannot rest content with the Skolem-hull, however. Once we throw the "witnesses" into the domain, we need to deal with $sk(e)$ assignments. In effect, we need a set which is its own Skolem-hull, and also contains the given subset d_1.

We define a sequence of non-empty sets $e_0, e_1, \ldots$ as follows: if the given subset d_1 of d is empty and there are no constants in K, then let e_0 be $C(d)$, the choice function applied to the entire domain; otherwise let e_0 be the union of d_1 and the denotations under I of the constants in K. For each natural number n, e_{n+1} is $sk(e_n)$. Finally, let d' be the union of the sets e_n, and let I' be the restriction of I to d'. Our interpretation is $M' = \langle d',I' \rangle$.

Clearly, d_1 is a subset of d', and so M' is a submodel of M. Let κ be the maximum of the size of K, the size of d_1, and denumerably infinite. A calculation reveals that the size of d' is at most κ, based on the fact that there are at most κ-many formulas, and thus, at most κ-many witnesses at each stage. Notice, incidentally, that this calculation relies on the fact that a denumerable union of sets of size at most κ is itself at most κ. This also relies on the axiom of choice.

The final item is to show that M' is elementarily equivalent to M: For every formula θ and every variable-assignment s on M',

$$M,s \vDash \theta \text{ if and only if } M',s \vDash \theta.$$

We proceed by induction on the complexity of θ. If θ is atomic, then the definition of satisfaction entails the equivalence. So let θ be non-atomic, and assume that $M,s \vDash \psi$ if and only if $M',s \vDash \psi$, for all assignments s on M' and all formulas ψ less complex than θ. Let s be any such assignment. If the main connective of θ is the negation sign or a binary connective, then the induction hypothesis entails that $M,s \vDash \theta$ if and only if $M',s \vDash \theta$. The remaining cases are those in which θ begins with a quantifier, i.e., θ is either $\exists v\psi$ or $\forall v\psi$. Suppose that $M',s \vDash \exists v\psi$. Then there is a variable-assignment s' that agrees with s except possibly at v such that $M',s' \vDash \psi$. By the induction hypothesis, $M,s' \vDash \psi$ and so $M,s \vDash \exists v\psi$. The converse is a bit tricky, and amounts to showing that the Skolem-hull of d' is d'. Assume that $M,s \vDash \exists v\psi$. We are given that s is a variable-assignment on d'. Since there are only finitely many free-variables in ψ, let n be any natural number such that for all variables u that occur free in ψ, $s(u)$ is in e_n. Let s_1 be an e_n-assignment that agrees with s on all of the free variables in ψ. Then, by Theorem 14, $M,s_1 \vDash \exists v\psi$. Let c be $w_v(\theta,s_1)$, the *v-witness of θ over s_1*. Notice that c is in e_{n+1} and so c is in d'. Let s_1' agree with s_1, except possibly at v, and let $s_1'(v)=c$. So s_1' is a variable-assignment on M'. By the definition of the witness function, $M,s_1' \vDash \psi$. By the induction hypothesis, $M',s_1' \vDash \psi$, and so $M',s_1 \vDash \exists v\psi$. By Theorem 14, $M',s \vDash \exists v\psi$. The final case, where θ has the form $\forall v\psi$, is similar.

Another corollary to Compactness (Corollary 22) is the opposite of the Löwenheim-Skolem theorem:

Theorem 26. Upward Löwenheim-Skolem Theorem. Let Γ be any set of formulas of $\mathcal{L}1K{=}$, such that for each natural number n, there is an interpretation $M_n = \langle d_n, I_n \rangle$, and an assignment s_n on M_n, such that d_n has at least n elements, and M_n,s_n satisfies every member of Γ. In other words, Γ is satisfiable and there is no finite upper bound to the size of the interpretations that satisfy every member of Γ. Then for any infinite cardinal κ, there is an interpretation $M=\langle d,I \rangle$ and assignment s on M, such that the size of d is *at least* κ and M,s satisfies every member of Γ. In particular, if Γ is a set of sentences, then it has arbitrarily

large models.

Proof: Add a collection of new constants $\{c\alpha \mid \alpha<\kappa\}$, of size κ, to the language, so that if c is a constant in K, then $c\alpha$ is different from c, and if $\alpha<\beta<\kappa$, then $c\alpha$ is a different constant than $c\beta$. Consider the set of formulas Γ' consisting of Γ together with the set $\{\neg c\alpha=c\beta \mid \alpha\neq\beta\}$. That is, Γ' consists of Γ together with statements to the effect that any two different new constants denote different objects. Let Γ'' be any finite subset of Γ', and let m be the number of new constants that occur in Γ''. Then expand the interpretation M_m to an interpretation M_m' of the new language, by interpreting each of the new constants in Γ'' as a different member of the domain d_m. By hypothesis, there are enough members of d_m to do this. One can interpret the other new constants at will. So M_m is a restriction of M_m'. By hypothesis (and Theorem 15), M'_m,s_m satisfies every member of Γ. Also M'_m,s_m satisfies the members of $\{\neg c\alpha=c\beta \mid \alpha\neq\beta\}$ that are in Γ''. So M'_m,s_m satisfies every member of Γ''. By compactness, there is an interpretation $M = \langle d,I\rangle$ and an assignment s on M such that M,s satisfies every member of Γ'. Since Γ' contains every member of $\{\neg c\alpha=c\beta \mid \alpha\neq\beta\}$, the domain d of M must be of size at least κ, since each of the new constants must have a different denotation. By Theorem 15, the restriction of M to the original language $\mathcal{L}1K=$ satisfies every member of Γ, with the variable-assignment s.

The proofs of the downward and upward Löwenheim-Skolem theorems can be combined to show that for any satisfiable set Γ of sentences, if there is no finite bound on the models of Γ, then for any infinite cardinal κ, there is a model of Γ whose domain has size *exactly* κ. Moreover, if M is any interpretation whose domain is infinite, then for any infinite cardinal κ, there is an interpretation M' whose domain has size exactly κ such that M and M' are elementarily equivalent.

These results indicate a weakness in the expressive resources of first-order languages like $\mathcal{L}1K=$. No satisfiable set of sentences can guarantee that its models are all denumerably infinite, nor can any satisfiable set of sentences guarantee that its models are uncountable. So in a sense, first-order languages cannot express the notion of "denumerably infinite", at least not in the model theory.

Let A be any set of sentences in a first-order language $\mathcal{L}1K=$, where K includes terminology for arithmetic, and assume that every member of A is true of the natural numbers. We can even let A be the set of all sentences in $\mathcal{L}1K=$ that are true of the natural numbers. Then A has uncountable models, indeed models of any infinite cardinality. Such interpretations are sometimes called *unintended*, or *non-standard* models of arithmetic. Let B be any set of first-order sentences that are true of the real numbers, and let C be any first-order axiomatization of set theory. Then if B and C are satisfiable (in infinite interpretations), then each of them has denumerably infinite models. That is, any first-order, satisfiable set theory or theory of the real numbers, has (unintended) models the size of the natural numbers. This is despite the fact that a sentence (seemingly) stating that the universe is uncountable is provable in most set-theories. This situation, known as the *Skolem paradox*, has generated much discussion, but we must refer the reader elsewhere for a sample of it.

Bibliography

Cited Works

- Anderson, A. and N. Belnap [1975], *Entailment: The logic of relevance and necessity I*, Princeton: Princeton University Press.
- Anderson, A. and N. Belnap, and M. Dunn [1992], *Entailment: The logic of relevance and necessity II*, Princeton: Princeton University Press.
- Corcoran, J. [1973], "Gaps between logical theory and mathematical practice", *The methodological unity of science*, ed. by M. Bunge, Dordrecht: Holland, D. Reidel, 23-50.
- Church, A. [1956], *Introduction to mathematical logic*, Princeton: Princeton University Press. Classic textbook.
- Davidson, D. [1984], *Inquiries into truth and interpretation*, Oxford: Clarendon Press.
- Dummett, M. [1977], *Elements of intuitionism*, Oxford: Oxford University Press.
- Gödel, K. [1930], "Die Vollständigkeit der Axiome des logischen Funktionenkalkuls", *Montatshefte für Mathematik und Physik 37*, 349-360; translated as "The completeness of the axioms of the functional calculus of logic", in van Heijenoort [1967], 582-591.
- Lycan, W. [1984], *Logical form in natural language*, Cambridge, Massachusetts: The MIT Press.
- Montague, R. [1974], *Formal philosophy*, ed. by R. Thomason, New Haven: Yale University Press.
- Quine, W. V. O. [1960], *Word and object*, Cambridge, Massachusetts: The MIT Press.
- Quine, W. V. O. [1986], *Philosophy of logic*, second edition, Englewood Cliffs, New Jersey: Prentice-Hall.
- Shapiro, S. [1998], "Logical consequence: models and modality", in *The philosophy of mathematics today*, edited by M. Schirn, Oxford: Oxford University Press, 131-156.
- Tennant, N. [1987], "Conventional necessity and the contingency of deduction", *Dialectica 41*, 79-95.

Further Reading

- Boolos, G., and R. Jeffrey [1989], *Computability and logic*, third edition, Cambridge, England: Cambridge University Press. Textbook on mathematical logic.
- Enderton, H. [1972], *A mathematical introduction to logic*, New York: Academic Press. Textbook in mathematical logic.
- Forbes, G. [1994], *Modern Logic*, Oxford: Oxford University Press. Elementary textbook.
- Mendelson, E. [1987], *Introduction to mathematical logic*, third edition, Princeton: van Nostrand. Textbook in mathematical logic.
- Shapiro, S. [1996], *The limits of logic: Second-order logic and the Skolem paradox*, *The international research library of philosophy*, Dartmouth Publishing Company, 1996. An anthology containing many of the significant later papers on the Skolem paradox.
- Van Heijenoort, J [1967], *From Frege to Gödel*, Cambridge, Massachusetts: Harvard University

Press. An anthology containing many of the major historical papers on mathematical logic in the early decades of the twentieth century.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

logic: infinitary | logic: intuitionistic | logic: modal | logic: temporal

Stewart Shapiro
shapiro+@osu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 15, 2000
Content last modified: September 15, 2000

Infinitary Logic

Traditionally, expressions in formal systems have been regarded as signifying finite inscriptions which are--at least in principle--capable of actually being written out in primitive notation. However, the fact that (first-order) formulas may be identified with natural numbers (via "Gödel numbering") and hence with finite *sets* makes it no longer necessary to regard formulas as inscriptions, and suggests the possibility of fashioning "languages" some of whose formulas--such as that in the above quotation--would be naturally identified as *infinite sets*. A "language" of this kind is called an *infinitary language*: in this article we discuss those infinitary languages which can be obtained in a straightforward manner from first-order languages by allowing conjunctions, disjunctions and, possibly, quantifier sequences, to be of infinite length. In the course of the discussion we shall see that, while the expressive power of such languages far exceeds that of their finitary (first-order) counterparts, very few of them possess the "attractive" features (e.g., compactness and completeness) of the latter. Accordingly, the infinitary languages that do in fact possess these features merit special attention.

In §1 we lay down the basic syntax and semantics of infinitary languages and demonstrate their expressive power by means of examples. §2 is devoted to those infinitary languages which permit only finite quantifier sequences: these languages turn out to be relatively well-behaved. In §3 we discuss the *compactness problem* for infinitary languages and its connection with purely set-theoretical questions concerning "large" cardinal numbers. In §4 an argument is sketched which shows that most "infinite quantifier" languages have a *second-order* nature and are, *ipso facto*, highly incomplete. In §5 we give a brief account of a certain special class of sublanguages of infinitary languages for which a satisfactory generalization of the compactness theorem can be proved. (This section links to a Supplement on the definition of admissible sets.) We conclude with historical and bibliographical remarks in §6.

1. Definition and Basic Properties of Infinitary Languages

Given a pair κ, λ of infinite cardinals such that $\lambda \leq \kappa$, we define a class of infinitary languages in each of which we may form conjunctions and disjunctions of sets of formulas of cardinality < κ, and quantifications over sequences of variables of length < λ.

Let $\mathcal{L}$ -- the (finitary) *base language* -- be an arbitrary but fixed first-order language with any number of extralogical symbols. The infinitary language $\mathcal{L}(\kappa, \lambda)$ has the following *basic symbols*:

- All symbols of $\mathcal{L}$
- A set **Var** of individual variables, where the cardinality of **Var** (written: |**Var**|) is λ
- A logical operator $\bigwedge$ (*infinitary conjunction*)

The class of *preformulas* of $\mathcal{L}(\kappa, \lambda)$ is defined recursively as follows:

- Each formula of $\mathcal{L}$ is a preformula;
- if φ and ψ are preformulas, so are $\varphi \wedge \psi$ and $\neg\varphi$;
- if Φ is a set of preformulas such that $|\Phi| < \kappa$, then $\bigwedge\Phi$ is a preformula;
- if φ is a preformula and $X \subseteq$ **Var** is such that $|X| < \lambda$, then $\exists X\varphi$ is a preformula;
- all preformulas are defined by the above clauses.

If Φ is a set of preformulas indexed by a set *I*, say $\Phi = \{\varphi_i : i \in I\}$, then we agree to write $\bigwedge\Phi$ for:

$$\bigwedge_{i \in I} \varphi$$

or, if *I* is the set of natural numbers, we write $\bigwedge\Phi$ for:

$$\varphi_0 \wedge \varphi_1 \wedge \ldots$$

If *X* is a set of individual variables indexed by an ordinal α, say $X = \{x_\xi : \xi < \alpha\}$, we agree to write $(\exists x_\xi)_{\xi<\alpha}\varphi$ for $\exists X\varphi$.

The logical operators $\vee$, $\rightarrow$, $\leftrightarrow$ are defined in the customary manner. We also introduce the operators $\bigvee$

(*infinitary disjunction*) and $\forall$ (*universal quantification*) by

$$\bigvee\Phi =_{df} \neg\bigwedge\{\neg\varphi : \varphi \in \Phi\}$$

$$\forall X\varphi =_{df} \neg\exists X\neg\varphi,$$

and employ similar conventions as for $\bigwedge$, $\exists$.

Thus $\mathcal{L}(\kappa,\lambda)$ is the infinitary language obtained from $\mathcal{L}$ by permitting conjunctions and disjunctions of length < κ and quantifications[1] of length < λ. Languages $\mathcal{L}(\kappa,\omega)$ are called *finite-quantifier* languages, the rest *infinite-quantifier* languages. Observe that $\mathcal{L}(\omega,\omega)$ is just $\mathcal{L}$ itself.

Notice the following *anomaly* which can arise in an infinitary language but not in a finitary one. In the language $\mathcal{L}(\omega_1,\omega)$, which allows countably infinite conjunctions but only finite quantifications, there are preformulas with so many free variables that they cannot be "closed" into sentences of $\mathcal{L}(\omega_1,\omega)$ by prefixing quantifiers. Such is the case, for example, for the $\mathcal{L}(\omega_1,\omega)$-preformula

$$x_0 < x_1 \wedge x_1 < x_2 \wedge \ldots \wedge x_n < x_{n+1} \ldots,$$

where $\mathcal{L}$ contains the binary relation symbol <. For this reason we make the following

> **Definition**. A *formula* of $\mathcal{L}(\kappa,\lambda)$ is a preformula which contains < λ free variables. The set of all formulas of $\mathcal{L}(\kappa,\lambda)$ will be denoted by **Form**($\mathcal{L}(\kappa,\lambda)$) or simply **Form**($\kappa,\lambda$) and the set of all sentences by **Sent**($\mathcal{L}(\kappa,\lambda)$) or simply **Sent**($\kappa,\lambda$).

In this connection, observe that, in general, nothing would be gained by considering "languages" $\mathcal{L}(\kappa,\lambda)$ with $\lambda > \kappa$. For example, in the "language" $\mathcal{L}(\omega,\omega_1)$, formulas will have only finitely many free variables, while there will be a host of "useless" quantifiers able to bind infinitely many free variables.[2]

Having defined the syntax of $\mathcal{L}(\kappa,\lambda)$, we next sketch its *semantics*. Since the extralogical symbols of $\mathcal{L}(\kappa,\lambda)$ are just those of $\mathcal{L}$, and it is these symbols which determine the form of the structures in which a given first-order language is to be interpreted, it is natural to define an $\mathcal{L}(\kappa,\lambda)$-structure to be simply an $\mathcal{L}$-structure. The notion of a formula of $\mathcal{L}(\kappa,\lambda)$ being *satisfied* in an $\mathcal{L}$-structure ***A*** (by a sequence of elements from the domain of ***A***) is defined in the same inductive manner as for formulas of $\mathcal{L}$ except that we must add two extra clauses corresponding to the clauses for $\bigwedge\Phi$ and $\exists X\varphi$ in the definition of preformula. In these two cases we naturally define:

> $\bigwedge\Phi$ is satisfied in ***A*** (by a given sequence) $\Leftrightarrow$ for all $\varphi \in \Phi$, φ is satisfied in ***A*** (by the sequence);

$\exists X\varphi$ is satisfied in $\boldsymbol{A} \Leftrightarrow$ there is a sequence of elements from the domain of $\boldsymbol{A}$ in bijective correspondence with X which satisfies φ in $\boldsymbol{A}$.

These informal definitions need to be tightened up in a rigorous development, but their meaning should be clear to the reader. Now the usual notions of *truth*, *validity*, *satisfiability*, and *model* for formulas and sentences of $\mathcal{L}(\kappa,\lambda)$ become available. In particular, if $\boldsymbol{A}$ is an $\mathcal{L}$-structure and $\sigma \in$ **Sent**(κ,λ), we shall write $\boldsymbol{A}\vDash\sigma$ for **A** is a model of σ, and $\vDash\sigma$ for σ *is valid*, that is, for all $\boldsymbol{A}$, $\boldsymbol{A}\vDash\sigma$. If $\Delta \subseteq$ **Sent**(κ,λ), we shall write $\Delta\vDash\sigma$ for σ *is a logical consequence of* Δ, that is, each model of Δ is a model of σ.

We now give some examples intended to display the expressive power of the infinitary languages $\mathcal{L}(\kappa, \lambda)$ with $\kappa \geq \omega_1$. In each case it is well-known that the notion in question cannot be expressed in any first-order language.

Characterization of the standard model of arithmetic in $\mathcal{L}(\omega_1,\omega)$. Here the *standard model of arithmetic* is the structure **N** = (N, +, ·, s, 0), where N is the set of natural numbers, +, ·, and 0 have their usual meanings, and s is the successor operation. Let $\mathcal{L}$ be the first-order language appropriate for **N**. Then the class of $\mathcal{L}$-structures isomorphic to **N** coincides with the class of models of the conjunction of the following $\mathcal{L}(\omega_1,\omega)$ sentences (where **0** is a name of 0):

$$\bigwedge_{m\in\omega}\ \bigwedge_{n\in\omega} s^m\mathbf{0} + s^n\mathbf{0} = s^{m+n}\mathbf{0}$$

$$\bigwedge_{m\in\omega}\ \bigwedge_{n\in\omega} s^m\mathbf{0} \cdot s^n\mathbf{0} = s^{m\cdot n}\mathbf{0}$$

$$\bigwedge_{m\in\omega}\ \bigwedge_{n\in\omega-\{m\}} s^m\mathbf{0} \neq s^n\mathbf{0}$$

$$\forall x \bigvee_{m\in\omega} x = s^m\mathbf{0}$$

The terms $s^n x$ are defined recursively by $s^0x = x$; $s^{n+1}x = s(s^nx)$.

Characterization of the class of all finite sets in $\mathcal{L}(\omega_1,\omega)$. Here the base language has no extralogical symbols. The class of all finite sets then coincides with the class of models of the $\mathcal{L}(\omega_1,\omega)$-sentence

$$\bigvee_{n\in\omega}\exists v_0 \ldots \exists v_n\forall x(x = v_0 \vee \ldots \vee x = v_n).$$

Truth definition in $\mathcal{L}(\omega_1,\omega)$ for a countable base language $\mathcal{L}$. Let $\mathcal{L}$ be a countable first-order language (for example, the language of arithmetic or set theory) which contains a name ***n*** for each natural number *n*, and let σ_0, σ_1, ... be an enumeration of its sentences. Then the $\mathcal{L}(\omega_1,\omega)$-formula

$$\mathbf{Tr}(x) =_{df} \bigvee_{n\in\omega} (x = \boldsymbol{n} \wedge \sigma_n)$$

is a *truth predicate* for $\mathcal{L}$ inasmuch as the sentence

$$\mathbf{Tr}(\boldsymbol{n}) \leftrightarrow \sigma_n$$

is valid for each *n*.

Characterization of well-orderings in $\mathcal{L}(\omega_1,\omega_1)$. The base language $\mathcal{L}$ here includes a binary predicate symbol $\preceq$. Let σ_1 be the usual $\mathcal{L}$-sentence characterizing linear orderings. Then the class of $\mathcal{L}$-structures in which the interpretation of $\preceq$ is a well-ordering coincides with the class of models of the $\mathcal{L}(\omega_1,\omega_1)$ sentence $\sigma = \sigma_1 \wedge \sigma_2$, where

$$\sigma_2 =_{df} \forall(v_n)_{n\in\omega}\, \exists x\, [\bigvee_{n\in\omega} (x = v_n) \wedge \bigwedge_{n\in\omega} (x \preceq v_n)].$$

Notice that the sentence σ_2 contains an *infinite quantifier*: it expresses the essentially *second-order* assertion that every countable subset has a least member. It can in fact be shown that the presence of this infinite quantifier is essential: the class of well-ordered structures cannot be characterized in any finite-quantifier language. This example indicates that infinite-quantifier languages such as $\mathcal{L}(\omega_1,\omega_1)$ behave rather like second-order languages; we shall see that they share the latters' defects (incompleteness) as well as some of their advantages (strong expressive power).

Many extensions of first-order languages can be *translated* into infinitary languages. For example, consider the generalized quantifier language $\mathcal{L}(Q_0)$ obtained from $\mathcal{L}$ by adding a new quantifier symbol Q_0 and interpreting $Q_0x\varphi(x)$ as *there exist infinitely many x such that* $\varphi(x)$. It is easily seen that the sentence $Q_0x\varphi(x)$ has the same models as the $\mathcal{L}(\omega_1,\omega)$-sentence

$$\neg\bigvee_{n\in\omega}\exists v_0\ldots\exists v_n\forall x[\varphi(x) \to (x = v_0 \vee \ldots \vee x = v_n)].$$

Thus $\mathcal{L}(Q_0)$ is, in a natural sense, translatable into $\mathcal{L}(\omega_1,\omega)$. Another language translatable into $\mathcal{L}(\omega_1,\omega)$ in this sense is the *weak second-order language* obtained by adding a countable set of monadic

predicate variables to $\mathcal{L}$ which are then interpreted as ranging over all *finite* sets of individuals.

2. Finite-Quantifier Languages

We have remarked that infinite-quantifier languages such as $\mathcal{L}(\omega_1,\omega_1)$ resemble second-order languages inasmuch as they allow quantification over infinite sets of individuals. The fact that this is not permitted in finite-quantifier languages suggests that these may be in certain respects closer to their first-order counterparts than might be evident at first sight. We shall see that this is indeed the case, notably in the case of $\mathcal{L}(\omega_1,\omega)$.

The language $\mathcal{L}(\omega_1,\omega)$ occupies a special place among infinitary languages because--like first-order languages--it admits an effective *deductive apparatus*. In fact, let us add to the usual first-order axioms and rules of inference the new axiom scheme

$$\bigwedge \Phi \rightarrow \varphi$$

for any countable set $\Phi \subseteq$ **Form**(ω_1,ω) and any $\varphi \in \Phi$, together with the new rule of inference

$$\frac{\varphi_0, \varphi_1, \ldots, \varphi_n, \ldots}{\bigwedge_{n\in\omega}\varphi_n}$$

and allow deductions to be of countable length. Writing $\vdash^*$ for deducibility in this sense, we then have the

$\mathcal{L}(\omega_1,\omega)$-**Completeness Theorem**. For any $\sigma \in$ **Sent**(ω_1,ω), $\models\sigma \Leftrightarrow \vdash^*\sigma$

As an immediate corollary we infer that this deductive apparatus is *adequate for deductions from countable sets of premises in* $\mathcal{L}(\omega_1,\omega)$. That is, with the obvious extension of notation, we have, for any *countable* set $\Delta \subseteq$ **Sent**(ω_1,ω)

(2.1) $\Delta\models\sigma \Leftrightarrow \Delta\vdash^*\sigma$

This completeness theorem can be proved by modifying the usual Henkin completeness proof for first-order logic, or by employing Boolean-algebraic methods. Similar arguments, applied to suitable further augmentations of the axioms and rules of inference, yield analogous completeness theorems for many other finite-quantifier languages.

If just deductions of countable length are admitted, then no deductive apparatus for $\mathcal{L}(\omega_1,\omega)$ can be set up which is adequate for deductions from *arbitrary* sets of premises, that is, for which (2.1) would hold for every set $\Delta \subseteq \mathbf{Sent}(\omega_1,\omega)$, *regardless of cardinality*. This follows from the simple observation that there is a first-order language $\mathcal{L}$ and an uncountable set Γ of $\mathcal{L}(\omega_1,\omega)$-sentences such that Γ *has no model but every countable subset of* Γ *does*. To see this, let $\mathcal{L}$ be the language of arithmetic augmented by ω_1 new constant symbols $\{\boldsymbol{c}_\xi : \xi < \omega_1\}$ and let Γ be the set of $\mathcal{L}(\omega_1,\omega)$-sentences $\{\sigma\} \cup \{\boldsymbol{c}_\xi \neq \boldsymbol{c}_\eta : \xi \neq \eta\}$, where σ is the $\mathcal{L}(\omega_1,\omega)$-sentence characterizing the standard model of arithmetic. This example also shows that the *compactness theorem* fails for $\mathcal{L}(\omega_1,\omega)$ and so also for any $\mathcal{L}(\kappa,\lambda)$ with $\kappa \geq \omega_1$.

Another result which holds in the first-order case but fails for $\mathcal{L}(\kappa,\omega)$ with $\kappa \geq \omega_1$ (and also for $\mathcal{L}(\omega_1, \omega_1)$, although this is more difficult to prove) is the *prenex normal form theorem*. A sentence is *prenex* if all its quantifiers appear at the front; we give an example of an $\mathcal{L}(\omega_1,\omega)$-sentence which is not equivalent to a conjunction of prenex sentences. Let $\mathcal{L}$ be the first-order language without extralogical symbols and let σ be the $\mathcal{L}(\omega_1,\omega)$-sentence which characterizes the class of finite sets. Suppose that σ were equivalent to a conjunction

$$\bigwedge_{i\in I} \sigma_i$$

of prenex $\mathcal{L}(\omega_1,\omega)$-sentences σ_i. Then each σ_i is of the form

$$Q_1x_1 \ldots Q_nx_n\, \varphi_i(x_1,\ldots, x_n),$$

where each Q_k is $\forall$ or $\exists$ and φ_i is a (possibly infinitary) conjunction or disjunction of formulas of the form $x_k = x_l$ or $x_k \neq x_l$. Since each σ_i is a sentence, there are only finitely many variables in each φ_i, and it is easy to see that each φ_i is then equivalent to a first-order formula. Accordingly each σ_i may be taken to be a first-order sentence. Since σ is assumed to be equivalent to the conjunction of the σ_i, it follows that σ and the set $\Delta = \{\sigma_i : i \in I\}$ have the same models. But obviously σ, and hence also Δ, have models of all finite cardinalities; the compactness theorem for sets of first-order sentences now implies that Δ, and hence also σ, has an infinite model, contradicting the definition of σ.

Turning to the *Löwenheim-Skolem theorem*, we find that the *downward* version has adequate generalizations to $\mathcal{L}(\omega_1,\omega)$ (and, indeed, to all infinitary languages). In fact, one can show in much the same way as for sets of first-order sentences that if $\Delta \subseteq \mathbf{Sent}(\omega_1,\omega)$ has an infinite model of cardinality $\geq |\Delta|$, it has a model of cardinality the larger of $\aleph_0$, $|\Delta|$. In particular, any $\mathcal{L}(\omega_1,\omega)$-sentence with an infinite model has a countable model.

On the other hand, the *upward* Löwenheim-Skolem theorem in its usual form *fails* for all infinitary languages. For example, the $\mathcal{L}(\omega_1,\omega)$-sentence characterizing the standard model of arithmetic has a

model of cardinality $\aleph_0$ but no models of any other cardinality. However, all is not lost here, as we shall see.

We define the *Hanf number* **h(L)** of a language L to be the least cardinal κ such that, if an **L**-sentence has a model of cardinality κ, it has models of arbitrarily large cardinality. The existence of **h(L)** is readily established. For each **L**-sentence σ not possessing models of arbitrarily large cardinality let $\kappa(\sigma)$ be the least cardinal κ such that σ does not have a model of cardinality κ. If λ is the supremum of all the $\kappa(\sigma)$, then, if a sentence of **L** has a model of cardinality λ, it has models of arbitrarily large cardinality.

Define the cardinals $\mu(\alpha)$ recursively by

$$\mu(0) = \aleph_0$$

$$\mu(\alpha+1) = 2^{\mu(\alpha)}$$

$$\mu(\lambda) = \sum_{\alpha<\lambda}\mu(\alpha) \text{ for limit } \lambda.$$

Then it can be shown that

$$\mathbf{h}(\mathcal{L}(\omega_1,\omega)) = \mu(\omega_1),$$

similar results holding for other finite-quantifier languages. The values of the Hanf numbers of infinite-quantifier languages such as $\mathcal{L}(\omega_1,\omega_1)$ are sensitive to the presence or otherwise of large cardinals, but must in any case greatly exceed that of $\mathcal{L}(\omega_1,\omega)$.

A result for $\mathcal{L}$ which generalizes to $\mathcal{L}(\omega_1,\omega)$ but to no other infinitary language is the

> **Craig Interpolation Theorem**: If $\sigma,\tau \in$ **Sent**(ω_1,ω) are such that $\models \sigma \rightarrow \tau$, then there is $\theta \in$ **Sent**(ω_1,ω) such that $\models \sigma \rightarrow \theta$ and $\models \theta \rightarrow \tau$, and each extralogical symbol occurring in θ occurs in both σ and τ.

The proof is a reasonably straightforward extension of the first-order case.

Finally, we mention one further result which generalizes nicely to $\mathcal{L}(\omega_1,\omega)$ but to no other infinitary language. It is well known that, if ***A*** is any finite $\mathcal{L}$-structure with only finitely many relations, there is an $\mathcal{L}$-sentence σ characterizing ***A*** up to isomorphism. For $\mathcal{L}(\omega_1,\omega)$ we have the following generalization known as

> **Scott's Isomorphism Theorem**. If ***A*** is a countable $\mathcal{L}$-structure with only countably many relations, then there is an $\mathcal{L}(\omega_1,\omega)$-sentence whose class of countable models coincides with the class of $\mathcal{L}$-structures isomorphic with ***A***.

The restriction to *countable* structures is essential because countability cannot in general be expressed by an $\mathcal{L}(\omega_1,\omega)$-sentence.

3. The Compactness Property

As we have seen, the compactness theorem in its usual form fails for all infinitary languages. Nevertheless, it is of some interest to determine whether infinitary languages satisfy some suitably modified version of the theorem. This so-called *compactness problem* turns out to have a natural connection with purely set-theoretic questions involving "large" cardinal numbers.

We construct the following definition. Let κ be an infinite cardinal. A language **L** is said to be *κ-compact* (resp. *weakly κ-compact*) if whenever Δ is a set of **L**-sentences (resp. a set of **L**-sentences of cardinality $\leq \kappa$) and each subset of Δ of cardinality $< \kappa$ has a model, so does Δ. Notice that the usual compactness theorem for $\mathcal{L}$ is precisely the assertion that $\mathcal{L}$ is ω-compact. One reason for according significance to the κ-compactness property is the following. Call L *κ-complete* (resp. *weakly κ-complete*) if there is a deductive system ***P*** for **L** with deductions of length $< \kappa$ such that, if Δ is a ***P***-consistent[3] set of **L**-sentences (resp. such that $|\Delta| \leq \kappa$), then Δ has a model. Observe that such a ***P*** will be adequate for deductions from arbitrary sets of premises (of cardinality $\leq \kappa$) in the sense of §2. It is easily seen that if **L** is κ-complete or weakly κ-complete, then **L** is κ-compact or weakly κ-compact. Thus, if we can show that a given language is *not* (weakly) κ-compact, then there can be no deductive system for it with deductions of length $< \kappa$ adequate for deductions from arbitrary sets of premises (of cardinality $\leq \kappa$).

It turns out, in fact, that most languages $\mathcal{L}(\kappa, \lambda)$ fail to be even weakly κ-compact, and, for those that are, κ must be an exceedingly *large* cardinal. We shall need some definitions.

An infinite cardinal κ is said to be *weakly inaccessible* if

> (a) $\lambda < \kappa \rightarrow \lambda^+ < \kappa$, (where λ^+ denotes the cardinal successor of λ), and
>
> (b) $|I| < \kappa$ and $\lambda_i < \kappa$ (for all $i \in I$) $\Rightarrow \sum_{i \in I} \lambda_i < \kappa$.

If in addition

> (c) $\lambda < \kappa \Rightarrow 2\lambda < \kappa$,

then κ is said to be (*strongly*) *inaccessible*. Since $\aleph_0$ is inaccessible, it is normal practice to confine attention to those inaccessible, or weakly inaccessible, cardinals that exceed $\aleph_0$. Accordingly, inaccessible or weakly inaccessible cardinals will always be taken to be *uncountable*. It is clear that such cardinals--if they exist--must be extremely large; and indeed the Gödel incompleteness theorem implies that the existence of even weakly inaccessible cardinals cannot be proved from the usual axioms of set theory.

Let us call a cardinal κ *compact* (resp. *weakly compact*) if the language $\mathcal{L}(\kappa,\lambda)$ is κ-compact (resp. weakly κ-compact). Then we have the following results:

> (3.1) $\aleph_0$ *is compact*. This is, of course, just a succinct way of expressing the compactness theorem for first-order languages.
>
> (3.2) κ is *weakly compact* $\Rightarrow$ $\mathcal{L}(\kappa,\omega)$ is *weakly* κ*-compact* $\Rightarrow$ κ is *weakly inaccessible*. Accordingly, it is consistent (with the usual axioms of set theory) to assume that no language $\mathcal{L}(\kappa,\omega)$ with $\kappa \geq \omega_1$ is weakly κ-compact, or, *a fortiori*, weakly κ-complete.
>
> (3.3) Suppose κ is inaccessible. Then κ is *weakly compact* $\Leftrightarrow$ $\mathcal{L}(\kappa,\omega)$ is *weakly* κ*-compact*. Also, Also κ is weakly compact $\Rightarrow$ *there is a set of* κ *inaccessibles before* κ. Thus a weakly compact inaccessible cardinal is exceedingly large; in particular it cannot be the first, second, ..., n^{th}, ... inaccessible.
>
> (3.4) κ *is compact* $\Rightarrow$ κ *is inaccessible*. (But, by the result immediately above, the converse fails.)

Let **Constr** stand for Gödel's axiom of constructibility; recall that **Constr** is consistent with the usual axioms of set theory.

> (3.5) *If* **Constr** *holds, then there are no compact cardinals*.
>
> (3.6) *Assume* **Constr** *and let* κ *be inaccessible. Then* κ *is weakly compact* $\Leftrightarrow$ $\mathcal{L}(\omega_1,\omega)$ *is weakly* κ*-compact for all* $\mathcal{L}$.
>
> (3.7) *If* **Constr** *holds, then there are no cardinals* κ *for which* $\mathcal{L}(\omega_1,\omega)$ *is compact*. Accordingly, it is consistent with the usual axioms of set theory to suppose that there is no cardinal κ such that all languages $\mathcal{L}(\omega_1,\omega)$ are κ-complete. This result is to be contrasted with the fact that *all* first-order languages are ω-complete.

The import of these results is that the compactness theorem fails very badly for most languages $\mathcal{L}(\kappa,\lambda)$

with $\kappa \geq \omega_1$.

Some historical remarks are in order here. In the 1930s mathematicians investigated various versions of the so-called *measure problem* for sets, a problem which arose in connection with the theory of Lebesgue measure on the continuum. In particular, the following very simple notion of measure was formulated. If X is a set, a (countably additive two-valued nontrivial) *measure* on X is a map μ on the power set $\mathbf{P}X$ to the set {0, 1} satisfying:

(a) $\mu(X) = 1$,

(b) $\mu(\{x\}) = \mu(\varnothing) = 0$ for all $x \in X$, and

(c) if $\boldsymbol{A}$ is any countable family of subsets of X, then $\mu(\bigcup \boldsymbol{A}) = \sum\{\mu(Y): Y \in \boldsymbol{A}\}$.

Obviously, whether a given set supports such a measure depends only on its cardinality, so it is natural to define a cardinal κ to be *measurable* if all sets of cardinality κ support a measure of this sort. It was quickly realized that a measurable cardinal must be inaccessible, but the falsity of the converse was not established until the 1960s when Tarski showed that measurable cardinals are weakly compact and his student Hanf showed that the first, second, etc. inaccessibles are not weakly compact (cf. (3.3)). Although the conclusion that measurable cardinals must be monstrously large is now normally proved without making the detour through weak compactness and infinitary languages, the fact remains that these ideas were used to establish the result in the first instance.

4. Incompleteness of Infinite-Quantifier Languages

Probably the most important result about first-order languages is the *Gödel completeness theorem* which of course says that the set of all valid formulas of any first-order language $\mathcal{L}$ can be generated from a simple set of axioms by means of a few straightforward rules of inference. A major consequence of this theorem is that, if the formulas of $\mathcal{L}$ are coded as natural numbers in some constructive way, then the set of (codes of) valid sentences is *recursively enumerable*. Thus, the completeness of a first-order language implies that the set of its valid sentences is *definable* in a particularly simple way. It would accordingly seem reasonable, given an *arbitrary* language **L**, to turn this implication around and suggest that, if the set of valid **L**-sentences is *not* definable in some simple fashion, then *no* meaningful completeness result can be established for **L**, or, as we shall say, that **L** is *incomplete*. In this section we are going to employ this suggestion in sketching a proof that "most" *infinite quantifier* languages are incomplete in this sense.

Let us first introduce the formal notion of *definability* as follows. If **L** is a language, $\boldsymbol{A}$ an **L**-structure, and X a subset of the domain A of $\boldsymbol{A}$, we say that X is *definable in* $\boldsymbol{A}$ by a formula $\varphi(x, y_1, \ldots, y_n)$ of **L** if there is a sequence $a_1, \ldots, a_n$ of elements of A such that X is the subset of all elements x A for which $\varphi(x, a_1, \ldots, a_n)$ holds in $\boldsymbol{A}$.

Now write *Val*(**L**) for the set of all the *valid* **L**-sentences, i.e., those that hold in every **L**-structure. In order to assign a meaning to the statement "*Val*(**L**) is definable", we have to specify

(i) a structure ***C*(L)**--the *coding structure* for **L**;

(ii) a particular one-one map--the *coding map*--of the set of formulas of **L** into the domain of ***C*(L)**.

Then, if we identify *Val*(**L**) with its image in ***C*(L)** under the coding map, we shall interpret the statement "*Val*(**L**) is definable" as the statement "*Val*(**L**), regarded as a subset of the domain of ***C*(L),** is definable in ***C*(L)** by a formula of **L**."

For example, when **L** is the first-order language $\mathcal{L}$ of arithmetic, Gödel originally used as coding structure the standard model of arithmetic $\mathbb{N}$ and as coding map the well-known function obtained from the prime factorization theorem for natural numbers. The recursive enumerability of *Val*($\mathcal{L}$) then means simply that the set of codes ("Gödel numbers") of members of *Val*($\mathcal{L}$) is definable in $\mathbb{N}$ by an $\mathcal{L}$-formula of the form $\exists y \varphi(x, y)$, where $\varphi(x, y)$ is a recursive formula.

Another, equivalent, coding structure for the first-order language of arithmetic is the structure[4] $<H(\omega), \in | H(\omega)>$ of *hereditarily finite sets*, where a set *x* is *hereditarily finite* if *x*, its members, its members of members, etc., are all finite. This coding structure takes account of the fact that first-order formulas are naturally regarded as finite sets.

Turning now to the case in which **L** is an infinitary language $\mathcal{L}(\kappa,\lambda)$, what would be a suitable coding structure in this case? We remarked at the beginning that infinitary languages were suggested by the possibility of thinking of formulas as set-theoretical objects, so let us try to obtain our coding structure by thinking about what kind of set-theoretical objects we should take infinitary formulas to be. Given the fact that, for each $\varphi \in$ **Form**(κ,λ), φ and its subformulas, subsubformulas, etc., are all of length $< \kappa$[5], a moment's reflection reveals that formulas of $\mathcal{L}(\kappa,\lambda)$ "correspond" to sets *x hereditarily of cardinality* $< \kappa$ in the sense that *x*, its members, its members of members, etc., are all of cardinality $< \kappa$. The collection of all such sets is written $H(\kappa)$. $H(\omega)$ is the collection of *hereditarily finite* sets introduced above, and $H(\omega_1)$ that of all *hereditarily countable* sets.

For simplicity let us suppose that the only extralogical symbol of the base language $\mathcal{L}$ is the binary predicate symbol $\underline{\in}$ (the discussion is easily extended to the case in which $\mathcal{L}$ contains additional extralogical symbols). Guided by the remarks above, as coding structure for $\mathcal{L}(\kappa,\lambda)$ we take the structure,

$$\mathcal{H}(\kappa) =_{df} <H(\kappa), \in | H(\kappa)>.$$

Now we can define the coding map of **Form**(κ,λ) into $\mathcal{H}(\kappa)$. First, to each basic symbol *s* of $\mathcal{L}(\kappa,\lambda)$

we assign a code object $\ulcorner s \urcorner \in H(\kappa)$ as follows. Let $\{v_\xi : \xi < \kappa\}$ be an enumeration of the individual variables of $\mathcal{L}(\kappa,\lambda)$.

Symbol	Code Object	Notation
$\neg$	1	$\ulcorner \neg \urcorner$
$\wedge$	2	$\ulcorner \wedge \urcorner$
$\bigwedge$	3	$\ulcorner \bigwedge \urcorner$
$\exists$	4	$\ulcorner \exists \urcorner$
$\in$	5	$\ulcorner \in \urcorner$
$=$	6	$\ulcorner = \urcorner$
v_ξ	$<0,\xi>$	$\ulcorner v_\xi \urcorner$

Then, to each $\varphi \in$ **Form**(κ,λ) we assign the code object $\ulcorner \varphi \urcorner$ recursively as follows:

$$\ulcorner v_\xi = v_\eta \urcorner =_{df} <\ulcorner v_\xi \urcorner, \ulcorner = \urcorner, \ulcorner v_\eta \urcorner>,$$

$$\ulcorner v_\xi \in v_\eta \urcorner =_{df} <\ulcorner v_\xi \urcorner, \ulcorner \in \urcorner, \ulcorner v_\eta \urcorner>;$$

for $\varphi, \psi \in$ **Form**(κ,λ),

$$\ulcorner \varphi \wedge \psi \urcorner =_{df} <\ulcorner \varphi \urcorner, \ulcorner \wedge \urcorner, \ulcorner \psi \urcorner>$$

$$\ulcorner \neg \varphi \urcorner =_{df} <\ulcorner \neg \urcorner, \ulcorner \varphi \urcorner>$$

$$\ulcorner \exists X \varphi \urcorner =_{df} <\ulcorner \exists \urcorner, \{\ulcorner x \urcorner : x \in X\}, \ulcorner \varphi \urcorner>;$$

and finally if $\Phi \subseteq$ **Form**(κ,λ) with $|\Phi| < \kappa$,

$$\ulcorner \bigwedge \varphi \urcorner =_{df} <\ulcorner \bigwedge \urcorner, \{\ulcorner \varphi \urcorner : \varphi \in \Phi\}>.$$

The map $\varphi \mapsto \ulcorner \varphi \urcorner$ from **Form**(κ,λ) into $H(\kappa)$ is easily seen to be one-one and is the required coding map. Accordingly, we agree to identify $Val(\mathcal{L}(\kappa,\lambda))$ with its image in $H(\kappa)$ under this coding map.

When is $Val(\mathcal{L}(\kappa,\lambda))$ a *definable* subset of $\mathcal{H}(\kappa)$? In order to answer this question we require the following definitions.

An $\mathcal{L}$-formula is called a Δ_0-*formula* if it is equivalent to a formula in which all quantifiers are of the

form $\forall x{\in}y$ or $\exists x{\in}y$ (i.e., $\forall x(x{\in}y \rightarrow ...)$ or $\exists x(x{\in}y \wedge ...)$). An $\mathcal{L}$-formula is a Σ_1-*formula* if it is equivalent to one which can be built up from atomic formulas and their negations using only the logical operators $\wedge$, $\vee$, $\forall x{\in}y$, $\exists x$. A subset X of a set A is said to be Δ_0 (resp. Σ_1) *on A* if it is definable in the structure $\langle A, \in | A\rangle$ by a Δ_0- (resp. Σ_1-) formula of $\mathcal{L}$.

For example, if we identify the set of natural numbers with the set $H(\omega)$ of hereditarily finite sets in the usual way, then for each $X \subseteq H(\omega)$ we have:

X is Δ_0 on $H(\omega) \Leftrightarrow X$ is recursive

X is Σ_1 on $H(\omega) \Leftrightarrow X$ is recursively enumerable.

Thus the notions of Δ_0- and Σ_1-set may be regarded as generalizations of the notions of *recursive* and *recursively enumerable* set, respectively.

The completeness theorem for $\mathcal{L}$ implies that $Val(\mathcal{L})$ -- regarded as a subset of $H(\omega)$ -- is recursively enumerable, and hence Σ_1 on $H(\omega)$. Similarly, the completeness theorem for $\mathcal{L}(\omega_1,\omega)$ (see §2) implies that $Val(\mathcal{L}(\omega_1,\omega))$ -- regarded as a subset of $H(\omega_1)$ -- is Σ_1 on $H(\omega_1)$. However, this pleasant state of affairs collapses completely as soon as $\mathcal{L}(\omega_1,\omega_1)$ is reached. For one can prove

Scott's Undefinability Theorem for $\mathcal{L}(\omega_1,\omega_1)$. *$Val(\mathcal{L}(\omega_1,\omega_1))$ is not definable in $\mathcal{H}(\omega_1)$ even by an $\mathcal{L}(\omega_1,\omega_1)$-formula*; *hence* a fortiori *$Val(\mathcal{L}(\omega_1,\omega_1))$ is not Σ_1 on $H(\omega_1)$.*

This theorem is proved in much the same way as the well-known result that the set of (codes of) valid sentences of the second-order language of arithemetic $\mathcal{L}^2$ is not second-order definable in its coding structure $\mathbb{N}$. To get this latter result, one first observes that $\mathbb{N}$ is characterized by a single $\mathcal{L}^2$-sentence, and then shows that, if the result were false, then "truth in $\mathbb{N}$" for $\mathcal{L}^2$-sentences would be definable by an $\mathcal{L}^2$-formula, thereby violating Tarski's theorem on the undefinability of truth.

Accordingly, to prove Scott's undefinability theorem along the above lines, one needs to establish:

(4.1) *Characterizability of the coding structure $\mathcal{H}(\omega_1)$ in $\mathcal{L}(\omega_1,\omega_1)$*: there is an $\mathcal{L}(\omega_1,\omega_1)$-sentence τ_0 such that, for all $\mathcal{L}$-structures $\boldsymbol{A}$,

$$\boldsymbol{A} \models \tau_0 \Leftrightarrow \boldsymbol{A} \cong \mathcal{H}(\omega_1).$$

(4.2) *Undefinability of truth for $\mathcal{L}(\omega_1,\omega_1)$-sentences in the coding structure*: there is no $\mathcal{L}(\omega_1,\omega_1)$-formula $\varphi(v_0)$ such that, for all $\mathcal{L}(\omega_1, \omega_1)$-sentences σ,

$$\mathcal{H}(\omega_1) \models \sigma \leftrightarrow \varphi(\ulcorner\sigma\urcorner).$$

(4.3) *There is a term* $t(v_0, v_1)$ *of* $\mathcal{L}(\omega_1,\omega_1)$ *such that, for each pair of sentences* σ, τ *of* $\mathcal{L}(\omega_1,\omega_1)$,

$$\mathcal{H}(\omega_1) \models t(\ulcorner\sigma\urcorner,\ulcorner\tau\urcorner) = \ulcorner\sigma \rightarrow \tau\urcorner.$$

(4.1) is proved by analyzing the set-theoretic definition of $\mathcal{H}(\omega_1)$ and showing that it can be "internally" formulated in $\mathcal{L}(\omega_1,\omega_1)$. (4.2) is established in much the same way as Tarski's theorem on the undefinability of truth for first- or second-order languages. (4.3) is obtained by formalizing the definition of the coding map $\sigma \mapsto \ulcorner\sigma\urcorner$ in $\mathcal{L}(\omega_1,\omega_1)$.

Armed with these facts, we can obtain Scott's undefinability theorem in the following way. Suppose it were false; then there would be an $\mathcal{L}(\omega_1,\omega_1)$-formula $\theta(v_0)$ such that, for all $\mathcal{L}(\omega_1,\omega_1)$-sentences σ,

(4.4) $\mathcal{H}(\omega_1) \models \theta(\ulcorner\sigma\urcorner) \Leftrightarrow \sigma \in Val(\mathcal{L}(\omega_1,\omega_1))$.

Let τ_0 be the sentence given in (4.1). Then we have, for all $\mathcal{L}(\omega_1,\omega_1)$-sentences σ,

$$\mathcal{H}(\omega_1) \models \sigma \Leftrightarrow \tau_0 \rightarrow \sigma \in Val(\mathcal{L}(\omega_1,\omega_1)),$$

so that, by (4.4),

$$\mathcal{H}(\omega_1) \models \sigma \Leftrightarrow \mathcal{H}(\omega_1) \models \theta(\ulcorner\tau_0 \rightarrow \sigma\urcorner).$$

If t is the term given in (4.3), it would follow that

$$\mathcal{H}(\omega_1) \models \sigma \leftrightarrow \theta(t(\ulcorner\tau_0\urcorner, \ulcorner\sigma\urcorner)).$$

Now write $\varphi(v_0)$ for the $\mathcal{L}(\omega_1,\omega_1)$-formula $\theta(t(\ulcorner\tau_0\urcorner, \ulcorner\sigma\urcorner))$. Then

$$\mathcal{H}(\omega_1) \models \sigma \leftrightarrow \varphi(\ulcorner\sigma\urcorner),$$

contradicting (4.2), and completing the proof.

Thus $Val(\mathcal{L}(\omega_1,\omega_1))$ is not definable *even by an* $\mathcal{L}(\omega_1,\omega_1)$*-formula*, so *a fortiori* $\mathcal{L}(\omega_1,\omega_1)$ is incomplete. Similar arguments show that Scott's undefinability theorem continues to hold when ω_1 is replaced by any successor cardinal κ^+; accordingly the languages $\mathcal{L}(\kappa^+,\kappa^+)$ are all incomplete.[6]

5. Sublanguages of $\mathcal{L}(\omega_1,\omega)$ and the Barwise Compactness Theorem

Given what we now know about infinitary languages, it would seem that $\mathcal{L}(\omega_1,\omega)$ is the only one to be reasonably well behaved. On the other hand, the failure of the compactness theorem to generalize to $\mathcal{L}(\omega_1,\omega)$ in any useful fashion is a severe drawback as far as applications are concerned. Let us attempt to analyze this failure in more detail.

Recall from §4 that we may code the formulas of a first-order language $\mathcal{L}$ as hereditarily finite sets, i.e., as members of $H(\omega)$. In that case each finite set of (codes of) $\mathcal{L}$-sentences is also a member of $H(\omega)$, and it follows that the compactness theorem for $\mathcal{L}$ can be stated in the form:

> (5.1) If $\Delta \subseteq$ **Sent**($\mathcal{L}$) is such that each subset $\Delta_0 \subseteq \Delta$, $\Delta_0 \in H(\omega)$ has a model, so does Δ.

Now it is well-known that (5.1) is an immediate consequence of the *generalized completeness theorem* for $\mathcal{L}$, which, stated in a form similar to that of (5.1), becomes the assertion:

> (5.2) If $\Delta \subseteq$ **Sent**($\mathcal{L}$) and $\sigma \in$ **Sent**($\mathcal{L}$) satisfy $\Delta \vDash \sigma$, then there is a deduction ***D*** of σ from Δ such that ***D*** $\in H(\omega)$.[7]

In §2 we remarked that the compactness theorem for $\mathcal{L}(\omega_1,\omega)$ fails very strongly; in fact, we constructed a set $\Gamma \subseteq$ **Sent**(ω_1,ω) such that

> (5.3) Each countable subset of Γ has a model but Γ does not.

Recall also that we introduced the notion of *deduction* in $\mathcal{L}(\omega_1,\omega)$; since such deductions are of countable length it quickly follows from (5.3) that

> (5.4) There is a sentence[8] $\sigma \in$ **Sent**(ω_1,ω) such that $\Gamma \vDash \sigma$, but there is no deduction of σ in $\mathcal{L}(\omega_1,\omega)$ from Γ.

Now the formulas of $\mathcal{L}(\omega_1, \omega)$ can be coded as members of $\mathcal{H}(\omega_1)$, and it is clear that $\mathcal{H}(\omega_1)$ is closed under the formation of countable subsets and sequences. Accordingly (5.3) and (5.4) may be written:

> (5.3 *bis*) Each $\Gamma_0 \subseteq \Gamma$ such that $\Gamma_0 \in \mathcal{H}(\omega_1)$ has a model, but Γ does not;

(5.4 *bis*) There is a sentence $\sigma \in \mathbf{Sent}(\omega_1,\omega)$ such that $\Gamma \models \sigma$, but there is no deduction $\boldsymbol{D} \in H(\omega_1)$ of σ from Γ.

It follows that (5.1) and (5.2) fail when "$\mathcal{L}$" is replaced by "$\mathcal{L}(\omega_1,\omega)$" and "$\mathcal{H}(\omega)$" by "$\mathcal{H}(\omega_1)$". Moreover, it can be shown that the set $\Gamma \subseteq \mathbf{Sent}(\omega_1,\omega)$ in (5.3 *bis*) and (5.4 *bis*) may be taken to be Σ_1 on $H(\omega_1)$. Thus the compactness and generalized completeness theorems fail even for Σ_1-sets of $\mathcal{L}(\omega_1, \omega)$-sentences.

We see from (5.4 *bis*) that the reason why the generalized completeness theorem fails for Σ_1-sets in $\mathcal{L}(\omega_1,\omega)$ is that, roughly speaking, $H(\omega_1)$ is not "closed" under the formation of deductions from Σ_1-sets of sentences in $H(\omega_1)$. So in order to remedy this it would seem natural to replace $H(\omega_1)$ by sets A which are, in some sense, closed under the formation of such deductions, and then to consider just those formulas whose codes are in A.

We now give a sketch of how this can be done.

First, we identify the symbols and formulas of $\mathcal{L}(\omega_1,\omega)$ with their codes in $H(\omega_1)$, as in §4. For each countable transitive[9] set A, let

$$\mathcal{L}_A = \mathbf{Form}(\mathcal{L}(\omega_1,\omega)) \cap A.$$

We say that $\mathcal{L}_A$ is a *sublanguage* of $\mathcal{L}(\omega_1,\omega)$ if the following conditions are satisfied:

(i) $\mathcal{L} \subseteq \mathcal{L}_A$

(ii) if $\varphi, \psi \in \mathcal{L}_A$, then $\varphi \wedge \psi \in \mathcal{L}_A$ and $\neg\varphi \in \mathcal{L}_A$

(iii) if $\varphi \in \mathcal{L}_A$ and $x \in A$, then $\exists x\varphi \in \mathcal{L}_A$

(iv) if $\varphi(x) \in \mathcal{L}_A$ and $y \in A$, then $\varphi(y) \in \mathcal{L}_A$

(v) if $\varphi \in \mathcal{L}_A$, every subformula of φ is in $\mathcal{L}_A$

(vi) if $\Phi \subseteq \mathcal{L}_A$ and $\Phi \in A$, then $\bigwedge\Phi \in \mathcal{L}_A$.

The notion of deduction in $\mathcal{L}_A$ is defined in the customary way; if Δ is a set of sentences of $\mathcal{L}_A$ and $\varphi \in \mathcal{L}_A$, then a *deduction* of φ from Δ in $\mathcal{L}_A$ is a deduction of φ from Δ in $\mathcal{L}(\omega_1, \omega)$ every formula of

which is in $\mathcal{L}_A$. We say that φ is *deducible* from Δ in $\mathcal{L}_A$ if there is a deduction $\boldsymbol{D}$ of φ from Δ in $\mathcal{L}_A$; under these conditions we write $\Delta \vdash_A \varphi$. In general, $\boldsymbol{D}$ will not be a member of A; in order to ensure that such a deduction can be found in A it will be necessary to impose further conditions on A.

Let A be a countable transitive set such that $\mathcal{L}_A$ is a sublanguage of $\mathcal{L}(\omega_1, \omega)$ and let Δ be a set of sentences of $\mathcal{L}_A$. We say that A (or, by abuse of terminology, $\mathcal{L}_A$) is Δ-*closed* if, for any formula φ of $\mathcal{L}_A$ such that $\Delta \vdash_A \varphi$, there is a deduction $\boldsymbol{D}$ of φ from Δ such that $\boldsymbol{D} \in A$. It can be shown that the only countable language which is Δ-closed for *arbitrary* Δ is the first-order language $\mathcal{L}$, i.e., when $A = H(\omega)$. However J. Barwise discovered that there are countable sets $A \subseteq H(\omega_1)$ whose corresponding languages $\mathcal{L}_A$ differ from $\mathcal{L}$and yet are Δ-closed *for all* Σ_1-*sets of sentences* Δ. Such sets A are called *admissible sets*; roughly speaking, they are extensions of the hereditarily finite sets in which recursion theory--and hence proof theory--are still possible.[10]

From Barwise's result one obtains immediately the

> **Barwise Compactness Theorem.** *Let* A *be a countable admissible set and let* Δ *be a set of sentences of* $\mathcal{L}_A$ *which is* Σ_1 *on* A. *If each* $\Delta' \subseteq \Delta$ *such that* $\Delta' \in$ A *has a model, then so does* Δ.

The presence of "Σ_1" here indicates that this theorem is a generalization of the compactness theorem for *recursively enumerable* sets of sentences.

Another version of the Barwise compactness theorem, useful for constructing models of set theory, is the following. Let **ZFC** be the usual set of axioms for Zermelo-Fraenkel set theory, including the axiom of choice. Then we have:

> **5.5. Theorem**. *Let A be a countable transitive set such that* $\boldsymbol{A} = \langle A, \in | A\rangle$ *is a model of* ***ZFC***. *If* Δ *is a set of sentences of* $\mathcal{L}_A$ *which is definable in* $\boldsymbol{A}$ *by a formula of the language of set theory and if each* $\Delta' \subseteq \Delta$ *such that* $\Delta' \in A$ *has a model, so does* Δ.

To conclude, we give a simple application of this theorem. Let $\boldsymbol{A} = \langle A, \in | A\rangle$ be a model of **ZFC**. A model $\boldsymbol{B} = \langle B, E\rangle$ of **ZFC** is said to be a *proper end-extension* of $\boldsymbol{A}$ if (i) $\boldsymbol{A} \subseteq \boldsymbol{B}$, (ii) $\boldsymbol{A} \neq \boldsymbol{B}$, (iii) $a \in A$, $b \in B$, $bEa \Rightarrow b \in A$. Thus a proper end-extension of a model of **ZFC** is a proper extension in which no "new" element comes "before" any "old" element. As our application of **5.5** we prove

> **5.6. Theorem**. *Each countable transitive model of* ***ZFC*** *has a proper end-extension.*

> **Proof**. Let $\boldsymbol{A} = \langle A, \in | A\rangle$ be a transitive model of **ZFC** and let $\mathcal{L}$ be the first-order language of set theory augmented by a name **a** for each $a \in A$, and an additional constant **c**. Let Δ be the set of $\mathcal{L}_A$-sentences comprising:

- all axioms of **ZFC**;
- $\mathbf{c} \neq \mathbf{a}$, for each $a \in A$;
- $\forall x(x \in \mathbf{a} \rightarrow \bigvee_{b \in a} x = \mathbf{b})$, for each $a \in A$;
- $\mathbf{a} \in \mathbf{b}$, for each $a \in b \in A$.

> It is easily shown that Δ is a subset of A which is definable in $\boldsymbol{A}$ by a formula of the language of set theory. Also, each subset $\Delta' \subseteq \Delta$ such that $\Delta' \in A$ has a model. For the set C of all $a \in A$ for which **a** occurs in Δ' belongs to A -- since Δ' does -- and so, if we interpret **c** as any member of the (necessarily nonempty) set A - C, then $\boldsymbol{A}$ is a model of Δ'. Accordingly, (5.5) implies that Δ has a model <*B*, *E*>. If we interpret each constant **a** as the element $a \in A$, then <*B*, *E*> is a proper end-extension of $\boldsymbol{A}$. The proof is complete.

The reader will quickly see that the first-order compactness theorem will not yield this result.

[Supplement: Definition of the Concept of Admissible Set

6. Historical and Bibliographical Remarks

§§**1** and **2**. Infinitary propositional and predicate languages seem to have made their first explicit appearance in print with the papers of Scott and Tarski [1958] and Tarski [1958]. The completeness theorem for $\mathcal{L}(\omega_1,\omega)$, as well as for other infinitary languages, was proved by Karp [1964]. The Hanf number calculations for $\mathcal{L}(\omega_1,\omega)$ were first performed by Morley [1965]. The nondefinability of well-orderings in finite-quantifier languages was proved by Karp [1965] and Lopez-Escobar [1966]. The interpolation theorem for $\mathcal{L}(\omega_1,\omega)$ was proved by Lopez-Escobar [1965] and Scott's isomorphism theorem for $\mathcal{L}(\omega_1,\omega)$ by Scott [1965].

§**3**. Results (3.2) and (3.3) are due to Hanf [1964], with some refinements by Lopez-Escobar [1966] and Dickmann [1975], while (3.4) was proved by Tarski. Result (3.5) is due to Scott [1961], (3.6) to Bell [1970] and [1972]; and (3.7) to Bell [1974]. Measurable cardinals were first considered by Ulam [1930] and Tarski [1939]. The fact that measurable cardinals are weakly compact was noted in Tarski [1962].

§**4**. The undecidability theorem for $\mathcal{L}(\omega_1,\omega_1)$ was proved by Scott in 1960; a fully detailed proof first appeared in Karp [1964]. The approach to the theorem adopted here is based on the account given in Dickmann [1975].

§**5**. The original motivation for the results presented in this section came from Kreisel; in his [1965] he pointed out that there were no compelling grounds for choosing infinitary formulas solely on the grounds of "length", and proposed instead that definability or "closure" criteria be employed. Kreisel's suggestion

was taken up with great success by Barwise [1967], where his compactness theorem was proved. The notion of admissible set is due to Platek [1966]. Theorem (5.6) is taken from Keisler [1974]. For further reading on the subject of infinitary languages, see Aczel [1973], Dickmann [1975], Karp [1964], Keisler [1974], and Makkai [1977].

Bibliography

- Aczel, P., 1973, "Infinitary Logic and the Barwise Compactness Theorem", *Proceedings of the 1971 Bertrand Russell Memorial Logic Conference* (Uldum, Denmark), J. Bell, J. Cole, G. Priest, and A. Slomson (eds.), Leeds: published by the Bertrand Russell Memorial Logic Conference, 234-277.
- Barwise, J., 1967, *Infinitary Logic and Admissible Sets*. Ph.D. Thesis, Stanford University.
- Bell, J. L., 1970, "Weak Compactness in Restricted Second-Order Languages", *Bull. Pol. Acad. Sci.* 18: 111-114.
- ---------, 1972, "On the Relationship between Weak Compactness in $\mathcal{L}(\omega_1, \omega)$, $\mathcal{L}(\omega_1, \omega_1)$, *and Restricted Second-Order Languages*", *Arch. Math. Logik* 15: 74-78.
- ---------, 1974, "On Compact Cardinals", *Z. f. Math. Logik u. Grund. D. Math* 20: 389-393.
- Dickmann, M. A., 1975, *Large Infinitary Languages*, Amsterdam: North-Holland.
- Hanf, W. P., 1964, *Incompactness in Languages with Infinitely Long Expressions*, Amsterdam: North-Holland.
- Karp, C., 1964, *Languages with Expressions of Infinite Length*, Amsterdam: North-Holland.
- ---------, 1965, "Finite-Quantifier Equivalence" in *The Theory of Models*, J. Addison, L. Henkin, and A. Tarski (eds.), Amsterdam: North-Holland, 407-412.
- Keisler, H. J., 1974, *Model Theory for Infinitary Logic*, Amsterdam: North-Holland.
- Kreisel, G., 1965, "Model-Theoretic Invariants, Applications to Recursive and Hyperarithmetic Operations", in *The Theory of Models*, J. Addison, L. Henkin, and A. Tarski (eds.), Amsterdam: North-Holland, 190-205.
- Lopez-Escobar, E. G. K., 1965, "An Interpolation Theorem for Infinitely Long Sentences", *Fund. Math.* 57: 253-272.
- ---------, 1966, "On Defining Well-Orderings", *Fund. Math.* 59: 13-21.
- Makkai, M., 1977, "Admissible Sets and Infinitary Logic", *Handbook of Mathematical Logic*, J. Barwise (ed.), Amsterdam: North-Holland, 233-282.
- Morley, M., 1965, "Omitting Classes of Elements", *The Theory of Models*, J. Addison, L. Henkin, and A. Tarski (eds.), Amsterdam: North-Holland, 265-273.
- Platek, R., 1966, *Foundations of Recursion Theory*, Ph.D. Thesis, Stanford University.
- Scott, D., 1961, "Measurable Cardinals and Constructible Sets", *Bull. Acad. Pol. Sci.* 9: 521-524.
- ---------, 1965, "Logic with Denumerably Long Formulas and Finite Strings of Quantifiers", *The Theory of Models*, J. Addison, L. Henkin, and A. Tarski (eds.), Amsterdam: North-Holland, 329-341.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

logic: classical

John L. Bell
jbell@julian.uwo.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 23, 2000
Content last modified: September 19, 2000

Stanford Encyclopedia of Philosophy
Notes to Infinitary Logic

Notes

1. Observe, however, that while the formation rules for $\mathcal{L}(\kappa,\lambda)$ allow the deployment of infinitely many quantifiers, each preformula can contain only finitely many *alternations* of quantifiers. Languages permitting infinite quantifier alternations have been developed in the literature, but we shall not discuss them here.

2. This remark loses its force when the base language contains predicate symbols with infinitely many argument places. However, this possibility is excluded here since our base language is a conventional first-order language.

3. I.e., such that no contradictions can be derived from Δ using the deductive machinery in ***P***.

4. If *A* is a set, $\in | A$ denotes the membership relation on *A*, i.e., $\{<x, y> \in A \times A: x \in y\}$.

5. Strictly speaking, this is only the case when κ is *regular*, that is, not the limit of $< \kappa$ cardinals each of which is $< \kappa$. In view of the fact that "most" cardinals are regular, we shall take this as read.

6. It should be pointed out, however, that there are languages $\mathcal{L}(\kappa,\lambda)$ apart from $\mathcal{L}(\omega,\omega)$ and $\mathcal{L}(\omega_1,\omega)$ which are complete; for example, all languages $\mathcal{L}(\kappa^+,\omega)$ and $\mathcal{L}(\lambda,\lambda)$ with inaccessible λ.

7. This is just a consequence of the fact that a first-order deduction is a finite sequence, hence a member of $H(\omega)$.

8. Take σ to be any logically false sentence!

9. A set *A* is *transitive* if $x \in y \in A \Rightarrow x \in A$.

10. For the definition of admissible set, see the Supplement at the end of §5.

John L. Bell
jbell@julian.uwo.ca

First published: January 23, 2000
Content last modified: January 23, 2000

Definition of the Concept of Admissible Set

A nonempty transitive set A is said to be *admissible* when the following conditions are satisfied:

> (i) if $a, b \in A$, then $\{a, b\} \in A$ and $\bigcup A \in A$;
>
> (ii) if $a \in A$ and $X \subseteq A$ is Δ_0 on A, then $X \cap a \in A$;
>
> (iii) if $a \in A$, $X \subseteq A$ is Δ_0 on A, and $\forall x \in a \exists y (\langle x,y \rangle \in X)$, then, for some $b \in A$, $\forall x \in a \exists y \in b (\langle x,y \rangle \in X)$.

Condition (ii) -- the Δ_0-*separation scheme* -- is a restricted version of Zermelo's axiom of separation. Condition (iii) -- a similarly weakened version of the axiom of replacement -- may be called the Δ_0-*replacement scheme*.

It is quite easy to see that if A is a transitive set such that $\langle A, \in | A \rangle$ is a model of **ZFC**, then A is admissible. More generally, the result continues to hold when the power set axiom is omitted from **ZFC**, so that both $H(\omega)$ and $H(\omega_1)$ are admissible. However, since the latter is uncountable, the Barwise compactness theorem fails to apply to it.

John L. Bell
jbell@julian.uwo.ca

Return to Infinitary Logic

First published: January 23, 2000
Content last modified: January 23, 2000

Intuitionistic Logic

Intuitionistic logic encompasses the principles of logical reasoning which were used by L. E. J. Brouwer in developing his intuitionistic mathematics, beginning in [1907]. Because these principles also underly Russian recursive analysis and the constructive analysis of E. Bishop and his followers, intuitionistic logic may be considered the logical basis of constructive mathematics.

Philosophically, intuitionism differs from logicism by treating logic as a part of mathematics rather than as the foundation of mathematics; from finitism by allowing (constructive) reasoning about infinite collections; and from platonism by viewing mathematical objects as mental constructs with no independent ideal existence. Hilbert's formalist program, to justify classical mathematics by reducing it to a formal system whose consistency should be established by finitistic (hence constructive) means, was the most powerful contemporary rival to Brouwer's developing intuitionism. In his 1912 essay *Intuitionism and Formalism* Brouwer correctly predicted that any attempt to prove the consistency of complete induction on the natural numbers would lead to a vicious circle.

Brouwer rejected formalism *per se* but admitted the potential usefulness of formulating general logical principles expressing intuitionistically correct constructions, such as *modus ponens*. Formal systems for intuitionistic propositional and predicate logic were developed by Heyting [1930], Gentzen [1935] and Kleene [1952]. The Gödel-Gentzen negative translation interpreted classical predicate logic in its intuitionistic subsystem. In [1965] Kripke provided a semantics with respect to which intuitionistic predicate logic is complete.

Rejection of *Tertium Non Datur*

Intuitionistic logic can be succinctly described as classical logic without the Aristotelian law of excluded middle LEM (A ∨ ¬A), but with the law of contradiction (¬A → (A → B)). Brouwer [1908] observed that LEM was abstracted from finite situations, then extended without justification to statements about infinite collections. For example, if x, y range over the natural numbers 0, 1, 2, ... and B(x) abbreviates the property (there is a y > x such that both y and y+2 are prime numbers), then we have no general method for deciding whether B(x) is true or false for arbitrary x, so ∀x(B(x) ∨ ¬B(x)) cannot be asserted in the present state of our knowledge. And if A abbreviates the statement ∀xB(x), then (A ∨ ¬A) cannot be asserted because neither A nor (¬A) has yet been proved.

One may object that these examples depend on the fact that the Twin Primes Conjecture has not yet been settled. A number of Brouwer's original "counterexamples" depended on problems (such as Fermat's Last Theorem) which have since been solved. But to Brouwer the general LEM was equivalent to the *a priori* assumption that **every** mathematical problem has a solution -- an assumption he rejected, anticipating Gödel's incompleteness theorem by a quarter of a century.

Intuitionistic First-Order Predicate Logic

The Brouwer-Heyting-Kolmogorov explication of intuitionistic truth, outlined in Section 1 ("The Basis of Constructive Mathematics") of the article on Constructive Mathematics in this Encyclopedia, results in the constructive independence of the logical operations &, ∨, →, ∀, ∃. This contrasts with the classical situation, where e.g., (A ∨ B) is equivalent to ¬(¬A & ¬B), and ∃xA(x) is equivalent to ¬∀x¬A(x). Intuitionistically, a sentence of the form (A ∨ B) asserts that either a proof of A, or a proof of B, has been constructed; while ¬(¬A & ¬B) asserts that an algorithm has been constructed which would effectively convert any pair of constructions proving ¬A and ¬B respectively, into a proof of a known contradiction.

Following is a Hilbert-style formalism, from Kleene [1952], for intuitionistic first-order predicate logic. The language *L* has predicate letters P, Q(.),... of all arities, individual variables, and symbols for the logical operations. We use A, B, C as metavariables for well-formed formulas. There are three rules of inference:

- From A and (A → B), conclude B.
- From C and (C → A(x)), where x is a variable which does not occur free in C, conclude (C → ∀xA(x)).
- From C and (A(x) → C), where x is a variable which does not occur free in C, conclude (∃xA(x) → C).

The axioms are all formulas of the following forms, where in the last two schemas the subformula A(t) is

the result of substituting an occurrence of the term t for every free occurrence of x in A(x), and no variable free in t becomes bound in A(t) as a result of the substitution.

- A → (B → A).
- (A → B) → ((A → (B → C)) → (A → C)).
- A → (B → A & B).
- A & B → A.
- A & B → B.
- A → A ∨ B.
- B → A ∨ B.
- (A → C) → ((B → C) → (A ∨ B → C)).
- (A → B) → ((A → ¬B) → ¬A).
- ¬A → (A → B).
- ∀xA(x) → A(t).
- A(t) → ∃xA(x).

A *proof* is any sequence of formulas, each of which is an axiom or an immediate consequence, by a rule of inference, of preceding formulas of the sequence. Any proof is said to *prove* its last formula, which is called a *theorem* or *provable formula* of first-order intuitionistic predicate logic.

If, in the given list of axiom schemas, the law of contradiction:

- ¬A → (A → B).

is replaced by the law of double negation:

- ¬¬A → A.

(or, equivalently, by LEM), a formal system for classical first-order predicate logic results. Since the law of contradiction is a classical theorem, intuitionistic logic is contained in classical logic.

A Little Proof Theory

The Gödel-Gentzen negative translation associates with each formula A of the language *L* a formula *g*(A) with no ∨ or ∃, which is equivalent to A in classical predicate logic. We define *g*(A) by induction on the logical form of A, as follows:

- *g*(P) is ¬¬P, if P is prime.
- *g*(A & B) is (*g*(A) & *g*(B)).
- *g*(A ∨ B) is ¬(¬*g*(A) & ¬*g*(B)).
- *g*(A → B) is (*g*(A) → *g*(B)).

- $g(\neg A)$ is $\neg g(A)$.
- $g(\forall x A(x))$ is $\forall x\, g(A(x))$.
- $g(\exists x A(x))$ is $\neg\forall x \neg g(A(x))$.

For each formula A, *g*(A) is provable intuitionistically if and only if A is provable classically. In particular, if (B & ¬B) were classically provable for some formula B, then (*g*(B) & ¬*g*(B)) (which is *g*(B & ¬B)) would in turn be provable intuitionistically. Gödel [1933] interpreted these results as showing that intuitionistic logic is *richer* than classical logic, because

- intuitionistic logic distinguishes formulas which are classically equivalent, and
- intuitionistic logic is equiconsistent with classical logic.

Gödel extended the negative translation to number theory, but attempts to apply it to intuitionistic analysis failed because the negative translation of the countable axiom of choice has no intuitionistic justification.

Gödel [1932] observed that intuitionistic propositional logic has the *disjunction property*: If (A ∨ B) is a theorem, then A is a theorem or B is a theorem. Gentzen [1935] established the disjunction property for closed formulas of intuitionistic predicate logic. From this follows e.g., that (P ∨ ¬P) is not a theorem if P is prime. Kleene [1945, 1952] proved that intuitionistic first-order number theory also has the related (cf., Friedman [1975]) *existence property*: If ∃xA(x) is a closed theorem, then for some closed term t, A(t) is a theorem. Kleene and Friedman also proved existence properties for second-order intuitionistic systems.

Some Semantics

Intuitionistic systems have inspired a variety of interpretations, including Beth's tableaus, Rasiowa and Sikorski's topological models, and Kleene's recursive realizabilities. Kripke's [1965] possible-world semantics, with respect to which intuitionistic predicate logic is complete and consistent, most resembles classical model theory.

A *Kripke structure* **K** consists of a partially ordered set *K* of *nodes* and a *domain function D* assigning to each node *k* in *K* an inhabited set *D*(*k*) of constructive objects, such that if $k \le k'$, then $D(k) \subseteq D(k')$. In addition **K** has a *forcing* relation determined as follows.

For each node *k* let *L*(*k*) be the language extending *L* by new constants for all the elements of *D*(*k*). To each node *k* and each prime predicate P(x) assign a (possibly empty) subset *T*(P,*k*) of *D*(*k*) in such a way that if $k \le k'$ then $T(P,k) \subseteq T(P,k')$, and similarly for predicates of more variables. Say that *k forces* P(*d*) if and only if $d \in T(P,k)$. Now define forcing for compound sentences of *L*(*k*) by

- *k forces* (A & B) if *k forces* A and *k forces* B.

- k *forces* (A ∨ B) if k *forces* A or k *forces* B.
- k *forces* (A → B) if, for every $k' \geq k$, if k' *forces* A then k' *forces* B.
- k *forces* ¬A if for no $k' \geq k$ does k' *force* A.
- k *forces* ∀xA(x) if for every $k' \geq k$ and every $d \in D(k')$, k' *forces* A(d).
- k *forces* ∃xA(x) if for some $d \in D(k)$, k *forces* A(d).

Any such forcing relation is *consistent* and *monotone*:

- for no sentence A and no k does k force both A and ¬A.
- if $k \leq k'$ and k forces A then k' forces A.

Kripke's Soundness and Completeness Theorems establish that a sentence of L is provable in intuitionistic predicate logic if and only if it is forced by every node of every Kripke structure. Thus to show that (¬∀x¬P(x) → ∃xP(x)) is intuitionistically unprovable, it is enough to consider a Kripke structure with $K = \{k, k'\}$, $k < k'$, $D(k) = D(k') = \{0\}$, $T(\text{P},k)$ empty but $T(\text{P},k') = \{0\}$.

Additional Topics

While intuitionistic arithmetic is a proper part of classical arithmetic, the intuitionistic attitude toward mathematical objects results in a theory of real numbers diverging from the classical. For readers wishing to pursue this subject farther, the third edition [1971] of Heyting's classic [1956] is an authentic and accessible introduction to intuitionistic philosophy, logic and mathematical practice. Kleene and Vesley's [1965] provides a careful axiomatic treatment of, and a constructive consistency proof for, intuitionistic analysis. Troelstra and van Dalen's comprehensive [1988] brings the story nearly up to date.

Bibliography

- Brouwer, L. E. J., "On the Foundations of Mathematics," Thesis, Amsterdam (1907); English translation in A. Heyting, Ed. *L. E. J. Brouwer: Collected Works* **1**: *Philosophy and Foundations of Mathematics*, Amsterdam: North Holland / New York: American Elsevier (1975): 11-101.
- Brouwer, L. E. J., "The Unreliability of the Logical Principles," (1908); English translation in Heyting, Ed., *Op.Cit.*: 107-111.
- Brouwer, L. E. J., "Intuitionism and Formalism," Amsterdam: (1912); English translation by A. Dresden in *Bull. Amer. Math. Soc.* **20** (1913): 81-96, reprinted in Heyting, Ed., *Op.Cit.*: 123-138.
- Friedman, H., "The disjunction property implies the numerical existence property," *Proc. Nat. Acad. Sci.* **72** (1975): 2877-2878.
- Gentzen, G., "Untersuchungen über das logische Schliessen," *Math. Zeitschrift* **39** (1934-5): 176-210, 405-431.
- Gödel, K., "Zum intuitionistischen Aussagenkalkül," *Anzeiger der Akademie der Wissenschaftischen in Wien* **69** (1932): 65-66.
- Gödel, K., "Zur intuitionistischen Arithmetik und Zahlentheorie," *Ergebnisse eines*

mathematischen Kolloquiums **4** (1933): 34-38.

- Heyting, A., "Die formalen Regeln der intuitionistischen Logik," in three parts, *Sitzungsber. preuss. Akad. Wiss.* (1930): 42-71, 158-169.
- Heyting, A., *Intuitionism: An Introduction*, Amsterdam: North-Holland (1956). Third Revised Edition (1971).
- Kleene, S. C., "On the interpretation of intuitionistic number theory," *Jour. Symb. Logic* **10** (1945): 109-124.
- Kleene, S. C., *Introduction to Metamathematics*, Princeton: Van Nostrand (1952).
- Kleene, S. C. and Vesley, R. E., *The Foundations of Intuitionistic Mathematics, Especially in Relation to Recursive Functions*, Amsterdam: North-Holland (1965).
- Kripke, S. A., "Semantical analysis of intuitionistic logic," in J. Crossley and M. A. E. Dummett, eds., *Formal Systems and Recursive Functions* Amsterdam: North-Holland (1965): 92-130.
- Troelstra, A. S. and van Dalen, D., *Constructivism in Mathematics: An Introduction*, in two volumes, Amsterdam: North-Holland (1988).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Brouwer, Luitzen Egbertus Jan | finitism | formalism | logicism | mathematics: constructive | Platonism: in metaphysics

Joan R. Moschovakis
joan@math.ucla.edu

First published: September 1, 1999
Content last modified: September 1, 1999

Constructive Mathematics

Constructive mathematics is distinguished from its traditional counterpart, classical mathematics, by the strict interpretation of the phrase ‘there exists’ as ‘we can construct’. In order to work constructively, we need to re-interpret not only the existential quantifier but all the logical connectives and quantifiers as instructions on how to construct a proof of the statement involving these logical expressions.

Although certain individuals -- most notably Kronecker -- had expressed disapproval of the ‘idealistic’, nonconstructive methods used by some of their nineteenth century contemporaries, it is in the polemical writings of L.E.J. Brouwer (1881-1966), beginning with his Amsterdam doctoral thesis (Brouwer 1907) and continuing over the next forty-seven years, that the foundations of a precise, systematic approach to constructive mathematics were laid. In Brouwer's philosophy, known as intuitionism, mathematics is a free creation of the human mind, and an object exists if and only if it can be (mentally) constructed.

---

1. Introduction

Before mathematicians assert something they are supposed to have proved it true. What, then, do mathematicians mean when they assert a disjunction $P \vee Q$, where P and Q are syntactically correct statements in some (formal or informal) language that a mathematician can use? A natural -- although, as we shall see, not the unique -- interpretation of this disjunction is that not only does (at least) one of the statements P, Q hold, but also we can decide which one holds. Thus just as mathematicians will assert that P only when they have decided that P by proving it, they may assert $P \vee Q$ only when they either can decide -- that is, prove -- that P or decide (prove) that Q.

With this interpretation, however, mathematicians run into a serious problem in the special case where Q is the negation, $\neg P$, of P. To decide that $\neg P$ is to show that P implies a contradiction (such as 0=1). But it will often be that mathematicians have neither decided that P nor decided that $\neg P$. To see this, we need only reflect on the following:

> **Goldbach Conjecture**:
> Every even integer > 2 can be written as a sum of two primes,

which remains neither proved nor disproved despite the best efforts of many of the leading mathematicians since it was first raised in a letter from Goldbach to Euler in 1742. We are forced to conclude that, under the very natural interpretation of $P \vee Q$ just canvassed, only an optimist can retain a belief in the law of excluded middle, $P \vee \neg P$.

Traditional, or *classical*, mathematics gets round this by widening the interpretation of disjunction: it interprets $P \vee Q$ as $\neg(\neg P \wedge \neg Q)$, or in other words, "it is contradictory that both P and Q be false". In turn, this leads to the *idealistic* interpretation of existence, in which $\exists x P(x)$ means $\neg \forall x \neg P(x)$ ("it is contradictory that $P(x)$ be false for every x"). It is on these interpretations of disjunction and existence that mathematicians have built the grand, and apparently impregnable, edifice of classical mathematics which serves a foundation for the physical, the social, and (increasingly) the biological sciences. However, the wider interpretations come at a cost: for example, when we pass from our initial, natural interpretation of $P \vee Q$ to the unrestricted use of the idealistic one, $\neg(\neg P \wedge \neg Q)$, the resulting mathematics cannot generally be interpreted within computational models such as recursive function theory.

This point is illustrated by a well-worn example, the proposition:

> There exists irrational numbers a, b such that a^b is rational.

A slick classical proof goes as follows. Either $\sqrt{2}^{\sqrt{2}}$ is rational, in which case we take $a = b = \sqrt{2}$; or else $\sqrt{2}^{\sqrt{2}}$ is irrational, in which case we take $a = \sqrt{2}^{\sqrt{2}}$ and $b = \sqrt{2}$ (See Dummett 1977, 10). But as it stands, this proof does not enable us to pinpoint which of the two choices of the pair (a,b) has the required property. In order to determine the correct choice of (a,b), we would need to decide whether $\sqrt{2}^{\sqrt{2}}$ is rational or irrational, which is precisely to employ our initial interpretation of disjunction with P the statement "$\sqrt{2}^{\sqrt{2}}$ is rational".

Here is another illustration of the difference between interpretations. Consider the following simple statement about the set $\mathbb{R}$ of real numbers:

> (*) $\forall x \in \mathbb{R}\ (x = 0 \vee x \neq 0)$,

where, for reasons that we divulge shortly, $x \neq 0$ means that we can find a rational number r with $0 < r <$

|x|. A natural computational interpretation of (*) is that we have a procedure which, applied to any real number x, either tells us that x=0 or else tells us that $x \neq 0$. (For example, such a procedure might output 0 if x=0, and output 1 if $x \neq 0$.) However, because the computer can handle real numbers only by means finite rational approximations, we have the problem of *underflow*, in which a sufficiently small positive number can be misread as 0 by the computer; so there cannot be a decision procedure that justifies the statement (*). In other words, we cannot expect (*) to hold under our natural computational interpretation of the quantifier $\forall$ and the connective $\vee$.

Let's examine this from another angle. Let $G(n)$ act as shorthand for the statement “$2n + 2$ is a sum of two primes”, where n ranges over the positive integers, and define an infinite binary sequence **a** = $(a_1, a_2, \ldots)$ as follows:

$$a_n = \begin{cases} 0 \text{ if } G(n) \text{ holds for all } k \leq n \\ 1 \text{ if } \neg G(n) \text{ holds for some } k \leq n \end{cases}$$

There is no question that **a** is a computationally well-defined sequence, in the sense that we have an algorithm for computing a_n for each n: check the even numbers 4,6,8,…,$2n$+2 to determine whether each of them is a sum of two primes; in that case, set a_n=0; in the contrary case, set a_n=1. Now consider the real number whose n^{th} binary digit is a_n:

$$\begin{aligned} x &= (0 \cdot a_1 a_2 a_2 \cdots)_2 \\ &= 2^{-1}a_1 + 2^{-2}a_2 + \cdots \\ &= \sum_{n=1}^{\infty} 2^{-n}a_n. \end{aligned}$$

If (*) holds under our computational interpretation, then we can decide between the following two alternatives:

- $2^{-1}a_1 + 2^{-2}a_2 + \cdots = 0$, which implies that $a_n = 0$ for every n;
- we can find a positive integer N such that $2^{-1}a_1 + 2^{-2}a_2 + \cdots > 2^{-N}$.

In the latter case, by testing $a_1,\ldots,a_N$, we can find $n \leq N$ such that $a_n = 1$. Thus the computational interpretation of (*) enables us to decide whether there exists n such that $a_n = 1$; in other words, it enables us to decide the status of the Goldbach Conjecture.

The use of the Goldbach Conjecture here is purely dramatic. To avoid it, we define a function f on the set of binary sequences as follows:

$$f(a) = \begin{cases} 0 \text{ if } a_n = 0 \text{ for all } n \\ 1 \text{ if } a_n = 1 \text{ for some } n. \end{cases}$$

The argument of the preceding paragraph can then be modified to show that, under our computational interpretation, (*) provides us with a procedure for calculating *f*(*a*) for any computationally well-defined binary sequence *f*(*a*). Now, the computability of the function *f* can be expressed informally by the following:

> **Limited Principle of Omniscience (LPO)**:
> For each binary sequence $(a_1,a_2,\ldots)$ either $a_n = 0$ for all n or else there exists n such that $a_n = 1$,

which is generally regarded as an essentially nonconstructive principle, for several reasons. First, its recursive interpretation,

> There is a recursive algorithm which, applied to any recursively defined binary sequence $(a_1,a_2,\ldots)$, outputs 0 if $a_n = 0$ for all n, and outputs 1 if $a_n = 1$ for some n,

is provably false within recursive function theory, even with classical logic (see Bridges & Richman 1987, Chapter 3); so if we want to allow a recursive interpretation of all our mathematics, then we cannot use LPO. Secondly, there is a model theory (involving the use of Kripke models) in which it can be shown that LPO is not derivable in Peano arithmetic using the computational interpretation of the connectives and quantifiers that we state in more detail in the next section (Bridges & Richman 1987, Chapter 7).

2. Constructive Interpretation of Logic

It should, by now, be clear that a full-blooded computational development of mathematics disallows the idealistic interpretations of disjunction and existence upon which most classical mathematics depends. In fact, in order to work constructively, we need to return from the classical interpretations back to the natural, constructive ones, as follows.

$\vee$ (or):	to prove $P \vee Q$ we must have either a proof of P or a proof of Q.
$\wedge$ (and):	to prove $P \wedge Q$ we must have a proof of P and a proof of Q.
$\Rightarrow$ (implies):	a proof of $P \Rightarrow Q$ is an algorithm that converts a proof of P into a proof of Q.
$\neg$ (not):	to prove $\neg P$ we must show that P implies $0 = 1$.
$\exists$ (there exists):	to prove $\exists x P(x)$ we must construct an object x and prove that $P(x)$ holds.

$\forall$ (for each/all): a proof of $\forall x P(x)$ is an algorithm that, applied to any object x, proves that $P(x)$ holds.

These computational interpretations can be made more precise using Kleene's notion of *realizability*; see (Dummett 1977, 318-335; Beeson 1985, Chapter VII).

Why would we want to do this? First, there is the desire to retain, as far as possible, computational interpretations of our mathematics. Ideally, we are trying to develop mathematics in such a way that if a theorem asserts the existence of an object x with a property P, then the proof of the theorem embodies algorithms for constructing x and for demonstrating, by whatever calculations are necessary, that x has the property P. Here are some examples of theorems, each followed by an informal description of the requirements for its constructive proof.

(A) For each real number x, either $x = 0$ or $x \neq 0$.

Proof requirement: An algorithm which, applied to a given real number x, would decide whether $x = 0$ or $x \neq 0$. Note that, in order to make this decision, the algorithm might use not only the data describing x but also the data showing that x is actually a real number.

(B) Each nonempty subset S of *R that is bounded above has a least upper bound.*

Proof requirement: An algorithm which, applied to a set S of real numbers, a member s of S, and an upper bound for S,

i. computes an object b and shows that b is a real number;
ii. shows that $x \leq b$ for each $x \in$ S; and
iii. given a real number $b' < b$, computes an element x of S such that $x > b'$.

(C) If f is a continuous real-valued mapping on the closed interval [0,1] such that $f(0)f(1) < 0$, then there exists x such that $0 < x < 1$ and $f(x) = 0$.

Proof requirement: An algorithm which, applied to the function f, a modulus of continuity for f, and the values $f(0)$ and $f(1)$,

i. computes an object x and shows that x is a real number between 0 and 1, and
ii. shows that $f(x) = 0$.

(D) If f is a continuous real-valued mapping on the closed interval [0,1] such that $f(0)f(1) < 0$, then for each $\varepsilon > 0$ there exists x such that $0 < x < 1$ and $|f(x)| < \varepsilon$.

Proof requirement: An algorithm which, applied to the function f, a modulus of continuity for f, the values $f(0)$ and $f(1)$, and a positive number ε,

i. computes an object x and shows that x is a real number between 0 and 1, and
ii. shows that $|f(x)| < \varepsilon$.

We already have reasons for doubting that (A) has a constructive proof. If the proof requirements for (B) can be fulfilled, then, given a binary sequence $(a_1,a_2,\ldots)$, we can apply our proof of (B) to the set $\{a_1,a_2,\ldots\}$ in order to determine its supremum σ. Computing σ with an error < 1/2, we then determine whether $\sigma = 0$ or $\sigma = 1$; in the first case, $a_n = 0$ for all n, whereas in the second, we can easily find N such that $a_N = 1$. Thus (B) implies LPO and is therefore essentially nonconstructive. However, in Bishop's constructive theory of the real numbers, based on Cauchy sequences with a preassigned convergence rate, we can prove the following:

> **Constructive Least-Upper-Bound Principle**:
> Let S be a nonempty subset of R that is bounded above. Then S has a least upper bound if and only if it is *located*, in the sense that for all real numbers α, β with $\alpha < \beta$, either β is an upper bound for S or else there exists x in> S with $x > \alpha$ (Bishop & Bridges 1985, p. 37, Proposition (4.3))

Each of statements (C) and (D), which are classically equivalent, is a version of the Intermediate Value Theorem. In these statements, a *modulus of continuity* for f is a set Ω of ordered pairs (ε, δ) of positive real numbers with the following two properties:

- for each $\varepsilon > 0$ there exists $\delta > 0$ such that $(\varepsilon, \delta) \in \Omega$;
- for each $(\varepsilon, \delta) \in \Omega$, and for all $x,y \in [0,1]$ such that $|x - y| < \delta$, we have $|f(x) - f(y)| < \varepsilon$.

Statement (C) is essentially nonconstructive, since it entails the following nonconstructive principle:

> **Lesser Limited Principle of Omniscience (LLPO)**:
> For each binary sequence $(a_1,a_2,\ldots)$ with at most one term equal to 1, either $a_n = 0$ for all even n or else $a_n = 0$ for all odd n.

Statement (D), a weak form of (C), can be proved constructively, using an interval-halving argument of a standard type. The following stronger constructive intermediate value theorem, which suffices for most practical purposes, is proved using an approximate interval-halving argument:

> Let f be a continuous real-valued mapping on the closed interval [0,1] such that $f(0)$ and $f(1)$ have opposite signs. Suppose also that f is *locally nonzero*, in the sense that for each x

$\in$ [0,1] and each $r > 0$, there exists y such that $|x - y| < r$ and $f(y) \neq 0$. Then there exists x such that $0 < x < 1$ and $f(x)=0$.

The situation of the intermediate value theorem is typical of many in constructive analysis, where we find one classical theorem with several constructive versions, some or all of which may be equivalent under classical logic. (See also, for example, Bridges *et al*. 1982.)

There is one omniscience principle whose constructive status is less clear than that of LPO and LLPO, namely, the following:

Markov's Principle (MP):
For each binary sequence (a_n), if it is contradictory that all the terms a_n equal 0, then there exists a term equal to 1.

This principle is equivalent to a number of simple classical propositions, including the following:

- For each real number x, if it is contradictory that x equal 0, then $x \neq 0$ (in the sense we mentioned earlier).
- For each real number x, if it is contradictory that x equal 0, then there exists $y \in \mathbb{R}$ such that $xy = 1$.
- For each one-one continuous mapping $f : [0,1] \to \mathbb{R}$, if $x \neq y$, then $f(x) \neq f(y)$.

Markov's Principle represents an unbounded search: if you have a proof that all terms a_n being 0 leads to a contradiction, then, by testing the terms $a_1,a_2,a_3,\ldots$ in turn, you are 'guaranteed' to come across a term equal to 1; but this guarantee does not extend to an assurance that you will find the desired term before the end of the universe. Most practitioners of constructive mathematics view Markov's Principle with at least suspicion, if not downright disbelief. Such views are reinforced by the observation that there is a Kripke Model showing that MP is not derivable in Peano arithmetic under our computational interpretation of logic. (See Bridges & Richman 1987, 137-138.)

3. Varieties of Constructive Mathematics

The desire to retain the possibility of a computational interpretation is one motivation for using the constructive reinterpretations of the logical connective and quantifiers that we gave above; but it is not exactly the motivation of the pioneers of constructivism in mathematics.

In the late nineteenth century, certain individuals -- most notably Kronecker and Poincaré -- had expressed doubts, or even disapproval, of the idealistic, nonconstructive methods used by some of their contemporaries; but it is in the polemical writings of L.E.J. Brouwer (1881-1966), beginning with his Amsterdam doctoral thesis in 1907 and continuing over the next forty-seven years, that the foundations of a precise, systematic approach to constructive mathematics were laid. In Brouwer's philosophy, known

as *intuitionism*, mathematics is a free creation of the human mind, and an object exists if and only if it can be (mentally) constructed. If one takes that philosophical stance, then one is inexorably drawn to the foregoing constructive interpretation of the logical connectives and quantifiers: for how could a proof of the impossibility of the non-existence of a certain object x -- an idealistic proof of the existence of x -- describe a mental construction of x? Brouwer was not the clearest expositor of his ideas, as is shown by the following quotation:

> Mathematics arises when the subject of two-ness, which results from the passage of time, is abstracted from all special occurrences. The remaining empty form [the relation of n to n+1] of the common content of all these two-nesses becomes the original intuition of mathematics and repeated unlimitedly creates new mathematical subjects. (quoted in Kline 1972, pp. 1199-2000)

A modern precis of Brouwer's view was given by Errett Bishop (Bishop 1967, p. 2):

> The primary concern of mathematics is number, and this means the positive integers. We feel about number the way Kant felt about space. The positive integers and their arithmetic are presupposed by the very nature of our intelligence and, we are tempted to believe, by the very nature of intelligence in general. The development of the positive integers from the primitive concept of the unit, the concept of adjoining a unit, and the process of mathematical induction carries complete conviction. In the words of Kronecker, the positive integers were created by God.

However obscure Brouwer's writings could be, one thing was always clear: for him, mathematics took precedence over logic. One might say, as Hermann Weyl does in the following passage, that Brouwer saw classical mathematics as flawed precisely in its use of classical logic without reference to the underlying mathematics:

> According to [Brouwer's] view and reading of history, classical logic was abstracted from the mathematics of finite sets and their subsets. ... Forgetful of this limited origin, one afterwards mistook that logic for something above and prior to all mathematics, and finally applied it, without justification, to the mathematics of infinite sets. This is the Fall and original sin of set theory, for which it is justly punished by the antinomies. It is not that such contradictions showed up that is surprising, but that they showed up at such a late stage of the game. (quoted in Kline 1972, p. 2001)

In particular, this misuse of logic led to nonconstructive existence proofs which, in Hermann Weyl's words, “inform the world that a treasure exists without disclosing its location”.

In order to describe the logic used by the intuitionist mathematician, it was necessary first to analyse the mathematical processes of the mind, from which analysis the logic could be extracted. In 1930, Brouwer's most famous pupil, Arend Heyting, published a set of formal axioms which so clearly

characterise the logic used by the intuitionist that they have become universally known as the axioms for intuitionistic logic (Heyting 1930). These axioms captured the informal computational interpretations of the connectives and quantifiers that we gave earlier. Over the years, Brouwer added principles, suggested by his introspective analysis of the nature of the continuum, which made intuitionistic mathematics diverge dramatically from its classical counterpart. The reader is referred to the entry on intuitionistic logic for more information about such matters.

Unfortunately -- and perhaps inevitably, in the face of opposition from mathematicians of such stature as Hilbert -- Brouwer's intuitionist school of mathematics and philosophy became more and more involved in what, at least to classical mathematicians, appeared to be quasi-mystical speculation about the nature of constructive thought, to the detriment of the practice of constructive mathematics itself. On the other hand, in the late 1940s the Russian mathematician A.A. Markov (Markov 1954) began the development of a form of recursive constructive mathematics (RUSS), which was, essentially, recursive function theory with intuitionistic logic (see Kushner 1985). In this variety the objects are defined by means of Gödel-numberings, and the procedures are all recursive; the main distinction between RUSS and the classical recursive analysis developed after, in 1936, the work of Turing, Church, and others clarified the nature of computable processes, is that the logic used in RUSS is intuitionistic. Thus RUSS may be described as ‘recursive mathematics with intuitionistic logic’.

One obstacle faced by the mathematician attempting to come to grips with RUSS is that, expressed in the language of recursion theory, it is not easily readable; indeed, on opening a page of Kushner's excellent lectures, one might be forgiven for wondering whether this is analysis or logic. Fortunately, one can get to the heart of RUSS by an axiomatic approach, as shown in Richman 1983.

Progress in all varieties of constructive mathematics was relatively slow throughout the next decade and a half. What was needed to raise the profile of constructivism in mathematics was a top-ranking classical mathematician to show that a thoroughgoing constructive development of mathematics was possible without a commitment to Brouwer's non-classical principles or to the machinery of recursive function theory. This need was fulfilled in 1967, with the appearance of Errett Bishop's monograph *Foundations of Constructive Analysis* (Bishop 1967), the product of an astonishing couple of years in which, working in the informal but rigorous style used by normal analysts, Bishop provided a constructive development of a large part of twentieth-century analysis, including the Stone-Weierstrass Theorem, the Hahn-Banach and separation theorems, the spectral theorem for self-adjoint operators on a Hilbert space, the Lebesgue convergence theorems for abstract integrals, Haar measure and the abstract Fourier transform, ergodic theorems, and the elements of Banach algebra theory. (See also Bishop & Bridges 1985.) Thus, at a stroke, he gave the lie to the commonly-held view expressed so forcefully by Hilbert:

> Taking the principle of excluded middle from the mathematician would be the same, say, as proscribing the telescope to the astronomer or to the boxer the use of his fists. (Hilbert 1928)

Not only did Bishop's mathematics (BISH) have the advantage of readability -- if you open Bishop's

book at any page, what you see is clearly recognisable as analysis, even if, from time to time, his moves in the course of a proof may appear strange to one schooled in the use of the law of excluded middle -- but, unlike intuitionistic or recursive mathematics, it admits many different interpretations. Intuitionistic mathematics, Markov's recursive constructive mathematics, and even classical mathematics all provide models of BISH. In fact, the results and proofs in BISH can be interpreted, with at most minor amendments, in any reasonable model of computable mathematics, such as, for example, Weihrauch's Type Two Effectivity Theory (Weihrauch 1996, 2000).

How is this multiple interpretability achieved? At least in part by Bishop's refusal to pin down his primitive notion of ‘algorithm’ or, in his words, ‘finite routine’. This refusal has led to the criticism that his approach lacks the precision that a logician would normally expect of a foundational system. However, this criticism can be overcome by looking more closely at what practitioners of BISH actually do, as distinct from what Bishop may have thought he was doing, when they prove theorems: in practice, they are doing *mathematics with intuitionistic logic*. Experience shows that the restriction to intuitionistic logic always forces mathematicians to work in a manner that, at least informally, can be described as algorithmic; so *algorithmic mathematics appears to be equivalent to mathematics that uses only intuitionistic logic*. If that is the case, then we can practice constructive mathematics using intuitionistic logic on any reasonably defined mathematical objects, not just some class of ‘constructive objects’.

This view, more or less, appears to have first been put forward by Richman (1990, 1996). Taking the logic as the primary characteristic of constructive mathematics, it does not reflect the primacy of mathematics over logic that was part of the belief of Brouwer, Heyting, Markov, Bishop, and other pioneers of constructivism. On the other hand, it does capture the essence of constructive mathematics in practice.

Thus one might distinguish between the *ontological constructivism* of Brouwer and others who are led to constructive mathematics through a belief that mathematical objects are mental creations, and the *epistemological constructivism* of Richman and those who see constructive mathematics as characterised by its methodology, based on the use of intuitionistic logic. Of course, the former approach to constructivism inevitably leads to the latter; and the latter is certainly not inconsistent with a Brouwerian ontology.

4. Conclusion

From a technical point of view, there seem to be (at least) two ways of handling the computational content and interpretation of mathematics. One way uses classical logic. In order to avoid decisions, such as whether or not a real number equals 0, that cannot be made by a real computer, one then has to work within a carefully specified algorithmic framework such as that of recursive function theory. The second approach uses intuitionistic logic, which automatically takes care of computationally inadmissible decisions. It is then unnecessary to insist upon a closely circumscribed algorithmic framework, so one can work in the style of an analyst, algebraist, geometer, ... , the only constraints being those imposed by intuitionistic logic.

Bibliography

References

- Beeson, Michael, 1985, *Foundations of Constructive Mathematics*, Heidelberg: Springer-Verlag.
- Bishop, Errett, 1967, *Foundations of Constructive Analysis*, New York: McGraw-Hill.
- Bishop, Errett, 1973, *Schizophrenia in Contemporary Mathematics, Amer. Math. Soc. Colloquium Lectures*, Missoula: University of Montana; reprinted in *Errett Bisop: Reflections on Him and His Research*, Amer. Math. Soc. Memoirs 39.
- Bishop, E. and Bridges, D., 1985, , Grundlehren der math. Wissenschaften **279**, Heidelberg: Springer-Verlag.
- Bridges, D., Calder, A., Julian, W., Mines. R. and Richman, F., 1982, "Picard's Theorem", *Trans. Amer. Math. Soc.*, 269(2), 513-520.
- Bridges, D. and Richman, F., 1987, *Varieties of Constructive Mathematics*, London Math. Soc. Lecture Notes **97**, Cambridge: Cambridge University Press.
- Brouwer, L.E.J., 1907, *Over de Grondslagen der Wiskunde*, Doctoral Thesis, University of Amsterdam, 1907; reprinted with additional material, D. van Dalen (ed.), by Matematisch Centrum, Amsterdam, 1981.
- Dummett, Michael, 1977, *Elements of Intuitionism*, Oxford: Clarendon Press.
- Heyting, A., 1930, "Die formalen Regeln der intuitionistischen Logik", Sitzungsber. preuss. Akad. Wiss. Berlin, 42-56.
- Hilbert, David, 1928, "Die Grundlagen der Mathematik", Hamburger Mathematische Einzelschriften 5, Teubner, Leipzig. Reprinted in English translation in van Heijenoort 1967, in which the exact quotation appears on page 476.
- Kline, Morris, 1972. *Mathematical Thought from Ancient to Modern Times* , vol 3, Oxford: Clarendon Press.
- Kushner, B., 1985, *Lectures on Constructive Mathematical Analysis*, Providence RI: Amer. Math. Soc.
- Markov, A.A., 1954, *Theory of Algorithms*, Trudy Mat. Istituta imeni V.A. Steklova, 42, Moskva: Izdatel'stvo Akademii Nauk SSSR.
- Richman, Fred, 1983, "Church's Thesis Without Tears", *J. Symbolic Logic*, 48, 797-803.
- -----, 1990, "Intuitionism as Generalization", *Philosophia Math*, **5**, 124-128.
- -----, 1996, "Interview with a Constructive Mathematician", *Modern Logic*, **6**, 247-271.
- Weihrauch, Klaus, 1996, "A Foundation for Computable Analysis", in *Combinatorics, Complexity, & Logic*, D. Bridges, C. Calude, J. Gibbons, S. Reeves, and I. Witten (eds.), Singapore: Springer-Verlag.
- -----, 2000, *Computable Analysis*, EATCS Texts in Theoretical Computer Science, Heidelberg: Springer-Verlag.

Related Literature

- Bridges, Douglas, 1998, "Constructive Truth in Practice", in *Truth in Mathematics*, H. Dales and G. Oliveri (eds.), Oxford: Clarendon Press.
- Brouwer, L.E.J., 1908, "De onbetrouwbaarheid der logische principes", *Tijdschrift voor Wijsbegeerte*, 2, 152-158.
- Goodman, N.D., and Myhill, J., "Choice Implies Excluded Middle", *Zeit. Logik und Grundlagen der Math*, 24, 461.
- Hayashi, S., and Nakano, H., 1988, *PX: A Computational Logic*, MIT Press, Cambridge MA.
- Heijenoort, Jean van, 1967, *From Frege to Gödel: A Source Book in Mathematical Logic 1879-1931*, Harvard University Press, Cambridge, Mass.
- Heyting, A., 1971, *Intuitionism -- An Introduction*, 3rd edition, Amsterdam: North Holland.
- Hilbert, D., 1925, "Über das Unendliche", *Mathematische Annalen*, 95, 161-190; translation, "On the Infinite", by E. Putnam and G. Massey, in *Philosophy of Mathematics: Selected Readings*, P. Benacerraf and H. Putnam (eds.), 134-151, Englewood Cliffs, NJ: Prentice Hall, 1964.
- Mines, R., Richman, F., and Ruitenburg, W., 1988, *A Course in Constructive Algebra*, Universitext, Heidelberg: Springer-Verlag.
- Myhill, John, 1973, "Some Properties of Intuitionistic Zermelo-Fraenkel Set Theory", in *Cambridge Summer School in Mathematical Logic*, A. Mathias and H. Rogers (eds.), Lecture Notes in Mathematics, 337, Heidelberg: Springer-Verlag, 206-231.
- -----, 1975, "Constructive Set Theory", *J. Symbolic Logic*, 40/3, 347-382.
- Martin-Löf, P., 1975, "An intuitionistic theory of types: predicative part", in *Logic Colloquium 1973* (H.E. Rose and J.C. Shepherdson, eds), 73-118, North-Holland, Amsterdam.
- Richman, Fred, 2000, "The Fundamental Theorem of Algebra: A Constructive Treatment Without Choice", *Pacific J. Math.*, 196, 213-230.
- Troelstra, A.S., 1978, "Aspects of Constructive Mathematics", in *Handbook of Mathematical Logic*, J. Barwise (ed.), Amsterdam: North-Holland.
- Troelstra, A.S., and van Dalen, D., 1988, *Constructivity in Mathematics: An Introduction* (two volumes), Amsterdam: North Holland.
- van Dalen, D., 1981, *Brouwer's Cambridge Lectures on Intuitionism*, Cambridge: Cambridge University Press.
- -----, 1999, *Mystic, Geometer and Intuitionist: The Life of L.E.J. Brouwer*, vol. I, Oxford: Clarendon Press.
- van Stigt, W.P., 1990, *Brouwer's Intuitionism*, Amsterdam: North-Holland.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Brouwer, Luitzen Egbertus Jan | logic: intuitionistic

Douglas Bridges
d.bridges@math.canterbury.ac.nz

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 18,1997
Content last modified: March 26, 2002

Inconsistent Mathematics

Inconsistent mathematics is the study of the mathematical theories that result when classical mathematical axioms are asserted within the framework of a (non-classical) logic which can tolerate the presence of a contradiction without turning every sentence into a theorem.

Inconsistent Mathematics

Inconsistent Mathematics began historically with foundational considerations. Set-theoretic paradoxes such as Russell's led to attempts to produce a consistent set theory as a foundation for mathematics. But, as is well known, set theories such as ZF, NBG and the like were in various ways ad hoc. Hence, a number of people including da Costa (1974), Brady (1971), Priest, Routley, and Norman (1989), considered it preferable to retain the full power of the natural abstraction principle (every predicate determines a set), and tolerate a degree of inconsistency in set theory. This requires, of course, that one dispense with the logical principle *ex contradictione quodlibet* (ECQ) (from a contradiction every proposition may be deduced), as well as any principle which leads to it, such as disjunctive syllogism (DS) (from *A*-or-*B* and not-*A* deduce *B*). But considerable debate, in Burgess (1981) and Mortensen (1983), made it clear that dispensing with ECQ and DS was not so counter-intuitive, especially when a plausible story emerged about the special conditions under which they continue to hold.

In addition, mathematics has a metalanguage; that is, names for mathematical statements and other parts of syntax, self-reference, proof and truth. Gödel's contribution to the philosophy of mathematics was to show that the first three of these can be rigorously expressed in arithmetical theories, albeit in theories which are either inconsistent or incomplete. The possibility of a well-structured example of the former alternative was not taken seriously, again because of belief in ECQ. However, in addition natural languages seem to have their own truth predicate. Combined with self-reference this produces the Liar paradox, "This sentence is false", an inconsistency. Priest (1987) and Priest, Routley, and Norman (1989) argued that the Liar had to be regarded as a statement both true and false, a true contradiction. This

represents another argument for studying inconsistent theories, namely the claim that some contradictions are true. Kripke (1975) proposed instead to model a truth predicate differently, in a consistent incomplete theory. We see below that incompleteness and inconsistency are closely related.

But mathematics is not its foundations. Hence there is a further independent motive, to see what mathematical structure remains when the constraint of consistency is relaxed. But it would be wrong to regard this as in any way a loss of structure. If it is different at all, then it represents an addition to known structure.

Robert K. Meyer (1976) seems to have been the first to think of an inconsistent arithmetical theory. At this point, he was more interested in the fate of a consistent theory, his relevant arithmetic R#. There proved to be a whole class of inconsistent arithmetical theories; see Meyer and Mortensen (1984), for example. Meyer argued that these theories provide the basis for a revived Hilbert Program. Hilbert's program was widely held to have been seriously damaged by Gödel's Second Incompleteness Theorem, according to which the consistency of arithmetic was unprovable within arithmetic itself. But a consequence of Meyer's construction was that within his arithmetic R# it was demonstrable by simple finitary means that whatever contradictions there might happen to be, they could not adversely affect any numerical calculations. Hence Hilbert's goal of conclusively demonstrating that mathematics is trouble-free proves largely achievable. The arithmetical models used later proved to allow inconsistent representation of the truth predicate. They also permit representation of structures beyond natural number arithmetic, such as rings and fields, including their order properties.

One could hardly ignore the examples of analysis and its special case, the calculus. There prove to be many places where there are distinctive inconsistent insights; see Mortensen (1995) for example. (1) Robinson's non-standard analysis was based on infinitesimals, quantities smaller than any real number, as well as their reciprocals, the infinite numbers. This has an inconsistent version, which has some advantages for calculation in being able to discard higher-order infinitesimals. Interestingly, the theory of differentiation turned out to have these advantages, while the theory of integration did not. (2) Another place is topology, where one readily observes the practice of cutting and pasting spaces being described as "identification" of one boundary with another. One can show that this can be described in an inconsistent theory in which the two boundaries are both identical and not identical, and it can be further argued that this is the most natural description of the practice. (3) Yet another application is the class of inconsistent continuous functions. Not all functions which are classically discontinuous are amenable of inconsistent treatment; but some are, for example $f(x)=0$ for all $x<0$ and $f(x)=1$ for all $x \geq 0$. The inconsistent extension replaces the first $<$ by $\leq$, and has distinctive structural properties. These inconsistent functions may well have some application in dynamic systems in which there are discontinuous jumps, such as quantum measurement systems. Differentiating such functions leads to the delta functions, applied by Dirac to the study of quantum measurement also. (4) Next, there is the well-known case of inconsistent systems of linear equations, such as the system (i) $x+y=1$, plus (ii) $x+y=2$. Such systems can potentially arise within the context of automated control. Little work has been done classically to solve such systems, but it can be shown that there are well-behaved solutions within inconsistent vector spaces. (5) Finally, one can note a further application in topology and dynamics. Given a supposition which seems to be conceivable, namely that whatever happens or is true, happens or

is true on an open set of (spacetime) points, one has that the logic of dynamically possible paths is open set logic, that is to say intuitionist logic, which supports incomplete theories par excellence. This is because the natural account of the negation of a proposition in such a space says that it holds on the largest open set contained in the Boolean complement of the set of points on which the original proposition held, which is in general smaller than the Boolean complement. However, specifying a topological space by its closed sets is every bit as reasonable as specifying it by its open sets. Yet the logic of closed sets is known to be paraconsistent, ie. supports inconsistent theories; see Goodman (1981) for example. Thus given the (alternative) supposition which also seems to be conceivable, namely that whatever is true is true on a closed set of points, one has that inconsistent theories may well hold. This is because the natural account of the negation of a proposition, namely that it holds on the smallest closed set containing the Boolean negation of the proposition, means that on the overlapping boundary both the proposition and its negation hold. Thus dynamical theories determine their own logic of possible propositions, and corresponding theories which may be inconsistent, and are certainly as natural as their incomplete counterparts.

Category theory throws light on many mathematical structures. It has certainly been proposed as an alternative foundation for mathematics. Such generality inevitably runs into problems similar to those of comprehension in set theory, see eg. Hatcher (1982, p.255-260). Hence there is the same possible application of inconsistent solutions. There is also an important collection of categorial structures, the toposes, which support open set logic in exact parallel to the way sets support Boolean logic. This has been taken by many to be a vindication of the foundational point of view of mathematical intuitionism. However, it can be proved that that toposes support closed set logic as readily as they support open set logic. That should not be viewed as an objection to intuitionism, however, so much as an argument that inconsistent theories are equally reasonable as items of mathematical study.

Duality between incompleteness/intuitionism and inconsistency/paraconsistency has at least two aspects. First there is the above topological (open/closed) duality. Second there is Routley * duality. Discovered by the Routleys (1972) as a semantical tool for relevant logics, the * operation dualises between inconsistent and incomplete theories of the large natural class of de Morgan logics. Both kinds of duality interact as well, where the * gives distinctive duality and invariance theorems for open set and closed set arithmetical theories. On the basis of these results, it is fair to argue that both kinds of mathematics, intuitionist and paraconsistent, are equally reasonable.

A very recent development is the application to explaining the phenomenon of inconsistent pictures. The best known of these are perhaps M.C.Escher's masterpieces *Belvedere, Waterfall* and *Ascending and Descending*. In fact the tradition goes back millennia to Pompeii. Escher seems to have derived many of his intuitions from the Swedish artist Oscar Reutersvaard, who began in 1934. Escher also actively collaborated with the English mathematician Roger Penrose. There have been several attempts to describe the mathematical structure of inconsistent pictures using classical consistent mathematics, by theorists such as Cowan, Francis and Penrose. As argued in Mortensen (1997), however, no consistent mathematical theory can capture the sense that one is seeing an impossible thing. Only an inconsistent theory can capture the content of that perception. This amounts to an appeal to cognition, that is the epistemological justification of paraconsistency as above. One can then proceed to describe inconsistent

theories which are candidates for such inconsistent contents. There is an analogy with classical mathematics on this point. Projective geometry is a mathematical theory which is interesting because we are creatures with an eye, since it explains why it is that things look the way they do in perspective.

These constructions do not in any way challenge or repudiate existing mathematics, but extend our conception of what is mathematically possible.

Bibliography

- Brady, Ross, "The Consistency of the Axioms of Abstraction and Extensionality in a Three-Valued Logic", *Notre Dame Journal of Formal Logic (NDJFL)*, 12 (1971), 447-453.
- Burgess, John, "Relevance, a Fallacy?", *NDJFL*, 22 (1981), 97-104.
- Da Costa, Newton C.A. "On the Theory of Inconsistent Formal Systems", *NDJFL*, 15 (1974), 497-510.
- Goodman, Nicholas, "The Logic of Contradictions", *Zeitschrift fur Mathematische Logic und Grundlagen der Arithmetik*, 27 (1981), 119-126.
- Hatcher, W.S. *The Logical Foundations of Mathematics*, Oxford: Pergamon, 1982.
- Kripke, Saul, "Outline of a Theory of Truth", *The Journal of Philosophy*, 72 (1975), 690-716.
- Meyer, Robert K., "Relevant Arithmetic", *Bulletin of the Section of Logic of the Polish Academy of Sciences*, 5 (1976), 133-137.
- Meyer, Robert K. and Chris Mortensen, "Inconsistent Models for Relevant Arithmetics",*The Journal of Symbolic Logic*, 49 (1984), 917-929.
- Mortensen, Chris, "Reply to Burgess and to Read", *NDJFL*, 24 (1983), 35-40.
- Mortensen, Chris, *Inconsistent Mathematics*, Kluwer Mathematics and Its Applications Series, Kluwer, 1995. Errata
- Mortensen, Chris, "Peeking at the Impossible", *NDJFL*, 38 (1997), 527-534.
- Priest, Graham, *In Contradiction,* Dordrecht: Nijhoff, 1987.
- Priest, Graham, Richard Routley and Jean Norman (eds), *Paraconsistent Logic*, Munchen: Philosophia Verlag, 1989.
- Routley, Richard and Val, "The Semantics of First Degree Entailment", *Nous*, 6 (1972), 335-359.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

contradiction | logic: paraconsistent | mathematics, philosophy of

Chris Mortensen
cmortens@arts.adelaide.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 2, 1996
Content last modified: August 6, 2000

Dialetheism

A *dialetheia* is a true contradiction, a statement, *A*, such that both it and its negation, ¬*A*, are true. Hence, *dialeth(e)ism* is the view that there are true contradictions. Dialetheism opposes the so-called *Law of Non-Contradiction* (LNC) (sometimes also called the *Law of Contradiction*): for any *A*, it is impossible for both *A* and ¬*A* to be true. Since Aristotle's defence of the LNC, the Law has been orthodoxy in Western philosophy. Nonetheless, there are some dialetheists in the history of Western Philosophy. Moreover, since the development of paraconsistent logic in the second half of this century, dialetheism has now become a live issue once more. In the rest of this article, I will start by explaining the relationship between dialetheism and some other important concepts. Next, I will describe the history of dialetheism and the motivations for the modern dialetheic movement. I will then indicate and briefly discuss some of the objections to dialetheism, and its connections with the notion of rationality.

Some Basic Concepts

Though dialetheism is not a new view, the word itself is. It was coined by Graham Priest and Richard Routley (later Sylvan) in 1981 (see Priest, Routley and Norman, 1989, p. *xx*). The inspiration for the name was a passage in Wittgenstein's *Remarks on the Foundations of Mathematics*, where he describes the liar sentence (‘this sentence is not true’) as a Janus-headed figure facing both truth and falsity (1978, IV.59). Hence a di-aletheia is a two(-way) truth. Unfortunately, Priest and Routley forgot to agree how to spell the ‘ism’, and versions with and without the ‘e’ appear in print.

Dialetheism should be clearly distinguished from *trivialism*, the view that *all* contradictions are true (and hence, assuming that a conjunction entails its conjuncts, that everything is true). Though a trivialist must be a dialetheist, the converse is not the case. Dialetheism should also be clearly distinguished from *paraconsistency*. An inference relation, $\vdash$, is *explosive* if, according to it, a contradiction entails everything (for all A and B: $A, \neg A \vdash B$). It is paraconsistent iff it is not explosive. Dialetheists, unless they are also trivialists, must subscribe to the view that entailment (deductively valid inference) is paraconsistent. But one may subscribe to this view for other reasons; for example, that, though the actual truth is consistent, entailment must preserve what holds in non-actual situations, some of which may be inconsistent; or that entailment must preserve more than just truth, e.g., information content.

Dialetheism in the History of Philosophy

In Western Philosophy, a number of the Presocratics appear to have endorsed dialetheism. For example, in Fragment 49a, Heraclitus says: ‘We step and do not step into the same rivers; we are and we are not’ (Robinson, 1987, p. 35). At any rate, Aristotle appears to have taken Heraclitus, Protagoras and other Presocratics to be dialetheists, even trivialists. Their views triggered his attack in *Metaphysics*, Book Gamma. In Chapter 4 of this occurs Aristotle's defence of the LNC. Historically, this attack was almost completely successful: the LNC has been high orthodoxy in Western Philosophy ever since, as is witnessed by the fact that no one since Aristotle seems to have felt the need to provide a sustained defence of it. It is perhaps worth noting that in *Metaphysics* Gamma (Chapter 7) Aristotle also defends the dual of the LNC, the Law of Excluded Middle, LEM: for any A, it is necessary for (at least) one of A and $\neg A$ to be true. But the LEM has always had a less secure place in Western Philosophy than the LNC. Aristotle himself, in fact, attacks the Law in *De Interpretatione*, Ch.9.

Despite the orthodoxy about the LNC, there have been a few dialetheists since Aristotle. It is arguably the case that some of the Neoplatonists were dialetheists. Nicholas of Cusa, for example, held that God has all properties, including contradictory ones (Heron, 1954, I.4). And, according to some interpretations, Meinong was a dialetheist, holding that some non-existent objects, such as the round square, have inconsistent properties (see Routley, 1980, ch.5). But the most obvious dialetheists since the Presocratics and before the 20th century are Hegel and his successors in dialectics, such as Marx and Engels (see Priest 1990, 1991). According to these, reality (in the form of *Geist* for Hegel, or social structures for Marx) may be literally inconsistent. For example, in the *Logic*, Hegel says: ‘Something moves, not because at one moment it is here and another there, but because at one and the same moment it is here and not here, because in this "here", it at once is and is not’ (Miller, 1969, p. 440). Indeed, it is the resolution of these contradictory states that drives the development of the history of thought (or society) forwards.

Dialetheism appears to be a much more common and recurrent view in Eastern Philosophy than in the West. In ancient Indian logic/metaphysics, there were standardly four possibilities to be considered on any statement at issue: that it is true (only), false (only), neither true nor false, or *both*. Early Buddhist logic added a fifth possibility: none of these. (This was called the *catushkoti*.) The Jains went even

further and advocated the possibility of contradictory values of the kind: true (only) *and* both true and false. (Smart, 1964, has a discussion of the above issues.) Contradictory utterances are a commonplace in Taoism. For example, the *Chuang Tsu* says: ‘That which makes things has no boundaries with things, but for things to have boundaries is what we mean by saying "the boundaries between things". The boundaryless boundary is the boundary without a boundary’ (Mair, 1994, p. 218). When Buddhism and Taoism fused to form Chan (or Zen, to give it its Japanese name), a philosophy arose in which contradiction plays a central role. The very process for reaching enlightenment (*Prajna*) is a process, according to Suzuki (1969, p. 55), "which is at once above and in the process of reasoning. This is a contradiction, formally considered, but in truth, this contradiction is itself made possible because of Prajna."

Of course, interpreting the philosophers I have mentioned is a sensitive issue; and many commentators, especially Western ones who have wanted to make sense of their chosen philosopher whilst subscribing to the LNC, have suggested that the contradictory utterances of the philosopher in question are not really contradictory. There are a number of standard devices that may be employed here. One is to claim that the contradictory utterance is to be taken as having some non-literal form of meaning, e.g., that it is a metaphor. Another is to claim that the contradictory assertion is ambiguous in some way, and that it is true on one disambiguation (or in one respect) and false in another. It is certainly the case that contradictory utterances that one sometimes hears are best construed in some such way. Whether this is so in the case of the philosophers I have mentioned, is a matter for detailed case-by-case consideration. In most of these cases, it may be argued, such interpretations produce a manifestly inaccurate and distorted version of the views of the philosopher in question.

Modern Motivations for Dialetheism

Turning now to contemporary philosophy, the second half of the 20th century has seen a resurgence of dialetheism, driven by largely new considerations. This has been made possible by the modern construction and analysis of paraconsistent logics; in turn, it has been one of the motive forces behind this movement in logic.

Probably the major argument used by modern dialetheists invokes the paradoxes of self-reference, such as the liar paradox (‘this sentence is not true’), and Russell's paradox (concerning the set of all sets that are not members of themselves). Though paradoxes of this kind have been known since antiquity, they were thrown into prominence by developments in the foundations of mathematics at the turn of this century. In the case of each paradox, there appears to be a perfectly sound argument ending in a contradiction; and if the arguments are sound, then dialetheism is true. Of course, many have argued that the soundness of such arguments is merely an appearance, and that subtle fallacies may be diagnosed in them. Such suggestions were made in ancient and medieval logic; but many more have been made in modern logic--indeed, attacking the paradoxes has been something of a *leitmotiv* of modern logic. And one thing that appears to have come out of this is how resilient the paradoxes are: attempts to solve them often simply succeed in relocating the paradoxes elsewhere, as so called "strengthened" forms of the arguments show. There is, at any rate, no generally agreed upon solution to many of the paradoxes,

particularly those of a semantic (as opposed to set-theoretic) nature. It is these facts that give dialetheism about the paradoxes of self-reference one of its major appeals. It is not the only one, though: the simplicity of a dialetheic account of truth, to the effect that truth is simply characterised by the *T*-schema, is another.

The paradoxes of self reference are not the only examples of dialetheias that have been mooted. Others include the following. (1) Transition states: when I leave the room, for an instant I am both in it and not in it. (2) Some of Zeno's paradoxes: the moving arrow is both where it is, and where it is not. (3) Certain legal situations: there are laws to the effect that persons in category *A* must do something, and persons in category *B* may not do it. Someone in both categories then turns up. (4) Borderline cases of vague predicates: an adolescent is both an adult and not an adult. (5) Certain quantum mechanical states: a particle may go through two slits simultaneously, even though this is not possible. (6) Multi-criterial terms: where a term has more than one necessary and sufficient empirical criterion for application, and these fall apart in novel circumstances. The viability of all the preceding examples depends on detailed philosophical consideration, differing from case to case.

Objections to Dialetheism

I now turn to arguments against dialetheism. The only sustained defense of the LNC in the history of philosophy is, as I mentioned, that given by Aristotle in Chapter 4 of *Metaphysics*, Gamma. Given the influence this chapter has had, the arguments are surprisingly poor. Aristotle's main argument, which takes up the first half of chapter, is tangled and contorted. It is not clear what it is, let alone that it works. About the best one can say for it is that it depends on substantial and moot principles of Aristotelian metaphysics, and, in any case, as a suasive argument, begs the question. The six or seven arguments that Aristotle deploys in the second half of the chapter are varied, swift, and fare little better. Many of them seem also to beg the question. Worse: many of them simply confuse dialetheism and trivialism. (For an analysis of Aristotle's arguments, see Priest ,1998a.)

A standard modern argument against dialetheism is to invoke the logical principle of explosion, in virtue of which dialetheism would entail trivialism. Given that trivialism is absurd, which we may agree upon here (though why this is so is not as easy a question as it might appear), dialetheism must be rejected. It is clear that this argument will fail against someone who subscribes to a paraconsistent logic, as most dialetheists will.

Another argument that is sometimes deployed is as follows. A sentence is meaningful only if it rules something out. But if the LNC fails, *A* does not rule out $\neg A$ or, *a fortiori*, anything else. Hence meaningful language presupposes the LNC. There are many problems with this argument, but the central one is that the first premise is simply false. Consider the sentence ‘Everything is true’. This entails everything, and so rules out nothing. Yet it is quite meaningful. It is what everyone except a trivialist rejects.

Of the other arguments one might consider in this context, I will consider only one more, which goes as

follows. The truth conditions for negation are: $\neg A$ is true iff A is not true. Hence, if A and $\neg A$ were true, A would be both true and not true, which is impossible. The truth conditions for negation employed by this argument are contentious. (An alternative is that $\neg A$ is true iff A is *false*-- and in the semantics of many paraconsistent logics, truth and falsity may overlap.) But in any case, the argument fails, since it clearly begs the question at last step. Many other arguments for the LNC, whatever other they failings have, seem ultimately to beg the question in similar ways.

Dialetheism and Rationality

Some have felt that what is wrong with dialetheism is not so much violation of the LNC itself, as that an *acceptance* of the LNC is a precondition for rationality. For example, it is often suggested that it could not possibly be rational to accept a contradiction. Whilst the question of the conditions under which it is rational to accept something is a moot one, it is commonly agreed that, as Hume put it, the wise person 'proportions his beliefs to the evidence' (1955, p. 118). Hence, if a sufficient case can be made out for a contradiction, it will be rational to believe it. And cases there are. For example, the case for the truth of the liar sentence, 'this sentence is not true', was gestured at in the previous section. And whether or not one takes the case in question to be completely persuasive, it illustrates the fact that there is nothing in principle impossible about the existence of such a case. Of course, if there were *conclusive* evidence for the LNC, then no case for a contradiction could be strong enough. But conclusive evidence for any philosophical position is virtually impossible.

A more persuasive worry about dialetheism, relating to rationality, is the claim that if a person could legitimately accept a contradiction, then no one could be forced, rationally, to abandon a view held. For if a person accepts A then, when an argument for $\neg A$ is put up, they could simply accept both A and $\neg A$. But this is too fast. The fact that some contradictions are rationally acceptable does not entail that all are. There is certainly a case that the liar sentence is both true and false, but this in no way provides a case that Brisbane is and is not in Australia. (Of course, if one subscribes to the claim that entailment is explosive, a case for one contradiction is a case for all; but if entailment is paraconsistent, this argument is of no use.) As orthodox philosophy of science indicates, there are, in fact, many considerations that speak against the rational acceptability of a view: that it is unduly complex, that it is contrived, that it has observable consequences that are not observed. (*Why* these are negative criteria is a different--and often difficult--question.) And these criteria may speak against the acceptability of a view, whether it is consistent or inconsistent. In the end, the rational evaluation of a view must balance it against all criteria of this kind (of which, inconsistency is, arguably, one), each, on its own, being defeasible.

Conclusion

I think it fair to say that since Aristotle's defence of the LNC, consistency has been something of a shibboleth in Western philosophy. The thought that consistency is a *sine qua non* for central notions such as validity, truth, meaningfulness, rationality, is deeply ingrained into its psyche. One thing that has come out of the modern investigations into dialetheism appears to be how superficial such a thought is. If

consistency is, indeed, a necessary condition for any of these notions it would seem to be for reasons much deeper than anyone has yet succeeded in articulating. And if it is not, then the way is open for the exploration of all kinds of avenues and questions in philosophy and the sciences that have traditionally been closed off.

Bibliography

I break up the references into sections corresponding to those of the text. Where a reference is not explicitly referred to in the text, I add a sentence concerning its relevance.

Some Basic Concepts

- G. Priest, R. Routley, and J. Norman (eds.), *Paraconsistent Logic: Essays on the Inconsistent*, Philosophia Verlag, 1989.
- L. Wittgenstein, *Remarks on the Foundations of Mathematics*, Basil Blackwell, 3rd edition, 1978.

Dialetheism in the History of Philosophy

- V. H. Mair (trans.), *Wandering on the Way: Early Taoist Tales and Parables of Chuang Tzu*, Bantam Books, 1994.
- A. V. Miller (trans.), *Hegel's Science of Logic*, Allen and Unwin Ltd., 1969.
- G. Heron (trans.), *Of Learned Ignorance*, Routledge and Kegan Paul, 1954.
- G. Priest, ‘Dialectic and Dialetheic’, *Science and Society* 53 (1990), 388-415.
- G. Priest, ‘Was Marx a Dialetheist?’, *Science and Society* 54 (1991), 468-75.
- G. Priest, and R. Routley, ‘The History of Paraconsistent Logic’, ch.1 of Priest, Routley and Norman, 1989 (above). (An account of dialetheism and paraconsistency in the history of philosophy.)
- T. M. Robinson, *Heraclitus: Fragments*, University of Toronto Press, 1987.
- R. Routley, *Exploring Meingong's Jungle*, Australian National University, 1980.
- N. Smart, *Doctrine and Argument in Indian Philosophy*, Allen and Unwin, 1964.
- D. T. Suzuki, *The Zen Doctrine of No Mind*, Rider and Co., 1969.

Modern Motivations for Dialetheism

- C. Mortensen, *Inconsistent Mathematics*, Kluwer Academic Publishers, 1995. (The introduction contains a discussion of dialetheism.)
- G. Priest, *In Contradiction*, Martinus Nijhoff, 1987. (A sustained defence of modern dialetheism.)
- G. Priest and R. Routley, ‘Applications of Paraconsistent Logic’ and ‘The Philosophical Significance and Inevitability of Paraconsistency’, chs. 13 and 18 of Priest, Routley and Norman, 1989 (above). (These contain some discussion of most of the motivations for dialetheism.)
- R. Routley, ‘Dialectical Logic, Semantics and Metamathematics’, *Erkenntnis* 14 (1979), 301-31.

(A defence of a dialetheic account of the paradoxes of self-reference.)

Objections to Dialetheism
Dialetheism and Rationality

- N. Denyer, ‘Dialetheism and Trivialisation’, *Mind* 98 (1989), 259-63. (A critique of a dialetheic account of the paradoxes of self-reference.)
- D. Hume, *An Inquiry Concerning Human Understanding*, ed. C.W.Hendel, Bobbs-Merril Company Inc., 1955.
- A. D. Irvine, ‘Gaps, Gluts and Paradox’, *Canadian Journal of Philosophy*, Supplementary Volume 18 (1992), 273-99. (A critique of a dialetheic account of the paradoxes of self-reference.)
- T. Parsons, ‘True Contradictions’, *Canadian Journal of Philosophy* 20 (1990), 335-53. (A critique of a dialetheic account of the paradoxes of self-reference.)
- G. Priest, ‘Denyer's $ Not Backed by Sterling Arguments’, *Mind* 98 (1989), 265-8. (A reply to Denyer, 1989.)
- G. Priest, ‘Gaps and Gluts: Reply to Parsons’, *Canadian Journal of Philosophy* 25 (1995), 57-66. (A reply to Parsons, 1990.)
- G. Priest, ‘To Be and Not to Be?That is the Answer. On Aristotle on the Law of Non-Contradiction’, *Philosophiegeschichte und Logische Analyse* 1 (1998a), 91-130.
- G. Priest, ‘What's So Bad About Contradictions?’, *Journal of Philosophy* 95 (1998b), 410-26. (A detailed discussion of some modern objections to dialetheism.)
- G. Priest, and T. Smiley, ‘Can Contradictions be True?’, *Proceedings of the Aristotelian Society, Supplementary Volume* 68 (1993), 17-54. (A debate on the issue of dialetheism.)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle: metaphysics | Hegel, Georg Wilhelm Friedrich | liar paradox | logic: paraconsistent | Russell’s paradox

Graham Priest
University of Melbourne
g.priest@unimelb.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 4, 1998
Content last modified: December 4, 1998

Aristotle's Metaphysics

The first major work in the history of philosophy to bear the title “Metaphysics” was the treatise by Aristotle that we have come to know by that name. But Aristotle himself did not use that title or even describe his field of study as ‘metaphysics’; the name was evidently coined by the first century C.E. editor who assembled the treatise we know as Aristotle's *Metaphysics* out of various smaller selections of Aristotle's works. The title ‘metaphysics’ — literally, ‘after the *Physics*’ — very likely indicated the place the topics discussed therein were intended to occupy in the philosophical curriculum. They were to be studied after the treatises dealing with nature (*ta phusika*). In this entry, we discuss the ideas that are developed in Aristotle's treatise.

§1. The Subject Matter of Aristotle's Metaphysics

Aristotle himself described his subject matter in a variety of ways: as ‘first philosophy’, or ‘the study of being qua being’, or ‘wisdom’, or ‘theology’. A comment on these descriptions will help to clarify Aristotle's topic.

In *Metaphysics* A.1, Aristotle says that “all men suppose what is called wisdom (*sophia*) to deal with the first causes (*aitia*) and the principles (*archai*) of things” (981b28), and it is these causes and principles that he proposes to study in this work. It is his customary practice to begin an inquiry by reviewing the opinions previously held by others, and that is what he does here, as Book A continues with a history of the thought of his predecessors about causes and principles.

These causes and principles are clearly the subject matter of what he calls ‘first philosophy’. But this does not mean the branch of philosophy that should be studied first. Rather, it concerns issues that are in some sense the most fundamental or at the highest level of generality. Aristotle distinguished between things that are “better known to us” and things that are “better known in themselves,”[1] and maintained that we should begin our study of a given topic with things better known to us and arrive ultimately at an understanding of things better known in themselves. The principles studied by ‘first philosophy’ may seem very general and abstract, but they are, according to Aristotle, better known in themselves, however remote they may seem from the world of ordinary experience. Still, since they are to be studied only by one who has already studied nature (which is the subject matter of the *Physics*), they are quite appropriately described as coming “after the *Physics*.”

Aristotle's description ‘the study of being qua being’ is frequently and easily misunderstood, for it seems to suggest that there is a single (albeit special) subject matter — being qua being — that is under investigation. But Aristotle's description does not involve two things — (1) a study and (2) a subject matter (being qua being) — for he did not think that there is any such subject matter as ‘being qua being’. Rather, his description involves three things: (1) a study, (2) a subject matter (being), and (3) a manner in which the subject matter is studied (qua being).

Aristotle's Greek word that has been Latinized as ‘qua’ means roughly ‘in so far as’ or ‘under the aspect’. A study of *x* qua *y*, then, is a study of *x* that concerns itself solely with the *y* aspect of *x*. So Aristotle's study does not concern some recondite subject matter known as ‘being qua being’. Rather it is a study of being, or better, of beings — of things that can be said to be — that studies them in a particular way: as beings, in so far as they are beings.

Of course, first philosophy is not the only field of inquiry to study beings. Natural science and mathematics also study beings, but in different ways, under different aspects. The natural scientist studies them as things that are subject to the laws of nature, as things that move and undergo change. That is, the natural scientist studies things qua movable (i.e., in so far as they are subject to change). The mathematician studies things qua countable and measurable. The metaphysician, on the other hand,

studies them in a more general and abstract way — qua beings. So first philosophy studies the causes and principles of beings qua beings. In Γ.2, Aristotle adds that for this this reason it studies the causes and principles of substances (*ousiai*). We will explain this connection in §3 below.

In Book E, Aristotle adds another description to the study of the causes and principles of beings qua beings. Whereas natural science studies objects that are material and subject to change, and mathematics studies objects that although not subject to change are nevertheless not separate from (i.e., independent of) matter, there is still room for a science that studies things (if indeed there are any) that are eternal, not subject to change, and independent of matter. Such a science, he says, is theology, and this is the "first" and "highest" science. Aristotle's identification of theology, so conceived, with the study of being qua being has proved challenging to his interpreters. We will deal with this issue in §14 below.

Finally, we may note that in Book B, Aristotle delineates his subject matter in a different way, by listing the problems or perplexities (*aporiai*) he hopes to deal with. Characteristic of these perplexities, he says, is that they tie our thinking up in knots. They include the following, among others: Are sensible substances the only ones that exist, or are there others besides them? Is it kinds or individuals that are the elements and principles of things? And if it is kinds, which ones: the most generic or the most specific? Is there a cause apart from matter? Is there anything apart from material compounds? Are the principles limited, either in number or in kind? Are the principles of perishable things themselves perishable? Are the principles universal or particular, and do they exist potentially or actually? Are mathematical objects (numbers, lines, figures, points) substances? If they are, are they separate from or do they always belong to sensible things? And ("the hardest and most perplexing of all," Aristotle says) are unity and being the substance of things, or are they attributes of some other subject? In the remainder of Book B, Aristotle presents arguments on both sides of each of these issues, and in subsequent books he takes up many of them again. But it is not always clear precisely how he resolves them, and it is possible that Aristotle did not think that the *Metaphysics* contains definitive solutions to all of these perplexities.

§2. The Categories

To understand the problems and project of Aristotle's *Metaphysics*, it is best to begin with one of his earlier works, the *Categories*. Although placed by long tradition among his logical works (see the discussion in the entry on Aristotle's logic), due to its analysis of the terms that make up the propositions out of which deductive inferences are constructed, the *Categories* begins with a strikingly general and exhaustive account of the things there are (*ta onta*) — beings. According to this account, beings can be divided into ten distinct categories. (Although Aristotle never says so, it is tempting to suppose that these categories are mutually exclusive and jointly exhaustive of the things there are.) They include substance, quality, quantity, and relation, among others. Of these categories of beings, it is the first, substance (*ousia*), to which Aristotle gives a privileged position.

Substances are unique in being independent things; the items in the other categories all depend somehow on substances. That is, qualities are the qualities of substances; quantities are the amounts and sizes that substances come in; relations are the way substances stand to one another. These various non-substances

all owe their existence to substances — each of them, as Aristotle puts it, exists only 'in' a subject. That is, each non-substance "is in something, not as a part, and cannot exist separately from what it is in" (*Cat*. 1a25). Indeed, it becomes clear that substances are the subjects that these ontologically dependent non-substances are 'in'.

Each member of a non-substance category thus stands in this inherence relation (as it is frequently called) to some substance or other — color is always found in bodies, knowledge in the soul. Neither whiteness nor a piece of grammatical knowledge, for example, is capable of existing on its own. Each requires for its existence that there be some substance in which it inheres.

In addition to this fundamental inherence relation across categories, Aristotle also points out another fundamental relation that obtains between items within a single category. He describes this as the relation of "being said of a subject," and his examples make clear that it is the relation of a more general to a less general thing within a single category. Thus, man is 'said of' a particular man, and animal is 'said of' man, and therefore, as Aristotle points out, animal is 'said of' the particular man also. The 'said of' relation, that is to say, is transitive (cf. 1b10). So the genus (e.g., animal) is 'said of' the species (e.g., man) and both genus and species are 'said of' the particular. The same holds in non-substance categories. In the category of quality, for example, the genus (color) is 'said of' the species (white) and both genus and species are 'said of' the particular white. (There has been considerable scholarly dispute about these particulars in nonsubstance categories. For more detail, see the supplementary document Nonsubstantial Particulars.)

The language of this contrast ('in' a subject vs. 'said of' a subject) is peculiar to the *Categories*, but the idea seems to recur in other works as the distinction between accidental vs. essential predication. Similarly, in works other than the *Categories,* Aristotle uses the label 'universals' (*ta katholou*) for the things that are "said of many;" things that are not universal he calls 'particulars' (*ta kath' hekasta*). Although he does not use these labels in the *Categories*, it is not misleading to say that the doctrine of the *Categories* is that each category contains a hierarchy of universals and particulars, with each universal being 'said of' the lower-level universals and particulars that fall beneath it. Each category thus has the structure of an upside-down tree.[2] At the top (or trunk) of the tree are the most generic items in that category[3] (e.g., in the case of the category of substance, the genus plant and the genus animal); branching below them are universals at the next highest level, and branching below these are found lower levels of universals, and so on, down to the lowest level universals (e.g., such *infimae species* as man and horse); at the lowest level — the leaves of the tree — are found the individual substances, e.g., this man, that horse, etc.

The individuals in the category of substance play a special role in this scheme. Aristotle calls them "primary substances" (*prôtai ousiai*) for without them, as he says, nothing else would exist. Indeed, Aristotle offers an argument (2a35-2b7) to establish the primary substances as the fundamental entities in this ontology. Everything that is not a primary substance, he points out, stands in one of the two relations (inhering 'in', or being 'said of') to primary substances. A genus, such as animal, is 'said of' the species below it and, since they are 'said of' primary substances, so is the genus (recall the transitivity of the 'said

of' relation). Thus, everything in the category of substance that is not itself a primary substance is, ultimately, 'said of' primary substances. And if there were no primary substances, there would be no "secondary" substances (species and genera), either. For these secondary substances are just the ways in which the primary substances are fundamentally classified within the category of substance. As for the members of non-substance categories, they too depend for their existence on primary substances. A universal in a non-substance category, e.g., color, in the category of quality, is 'in' body, Aristotle tells us, and therefore in individual bodies. For color could not be 'in' body, in general, unless it were 'in' at least some particular bodies. Similarly, particulars in non-substance categories (although there is not general agreement among scholars about what such particulars might be) cannot exist on their own. E.g., a determinate shade of color, or a particular and non-shareable bit of that shade, is not capable of existing on its own — if it were not 'in' at least some primary substance, it would not exist. So primary substances are the basic entities — the basic "things that there are" — in the world of the *Categories*.

§3. The Role of Substance in the Study of Being Qua Being

The *Categories* leads us to expect that the study of being in general (being qua being) will crucially involve the study of substance, and when we turn to the *Metaphysics* we are not disappointed. First, in *Metaphysics* Γ Aristotle argues in a new way for the ontological priority of substance; and then, in Books Z, H, and Θ, he wrestles with the problem of what it is to be a substance. We will begin with Γ's account of the central place of substance in the study of being qua being.

As we noted above, metaphysics (or, first philosophy) is the science which studies being qua being. In this respect it is unlike the specialized or departmental sciences, which study only part of being (only some of the things that exist) or study beings only in a specialized way (e.g., only in so far as they are changeable, rather than in so far as they are beings).

But 'being', as Aristotle tells us in Γ.2, is "said in many ways". That is, the verb 'to be' (*einai*) has different senses, as do its cognates 'being' (*on*) and 'entities' (*onta*). So the universal science of being qua being appears to founder on an equivocation: how can there be a single science of being when the very term 'being' is ambiguous?

Consider an analogy. There are dining tables, and there are tide tables. A dining table is a table in the sense of a smooth flat slab fixed on legs; a tide table is a table in the sense of a systematic arrangement of data in rows and columns. But there is not a single sense of 'table' which applies to both the piece of furniture at which I am writing these words and to the small booklet that lies upon it. Hence it would be foolish to expect that there is a single science of tables, in general, that would include among its objects both dining tables and tide tables. Tables, that is to say, do not constitute a single kind with a single definition, so no single science, or field of knowledge, can encompass precisely those things that are correctly called 'tables'.

If the term ‘being’ were ambiguous in the way that ‘table’ is, Aristotle's science of being qua being would be as impossible as a science of tables qua tables. But, Aristotle argues in Γ.2, ‘being’ is not ambiguous in this way. ‘Being’, he tells us, is ‘said in many ways’ but it is not merely (what he calls) ‘homonymous’, i.e., sheerly ambiguous. Rather, the various senses of ‘being’ have what he calls a ‘*pros hen*’ ambiguity — they are all related to a single central sense. (The Greek phrase ‘*pros hen*’ means “in relation to one.”)

Aristotle explains his point by means of some examples that he takes to be analogous to ‘being’. Consider the terms ‘healthy’ and ‘medical’. Neither of these has a single definition that applies uniformly to all cases: not every healthy (or medical) thing is healthy (medical) in the same sense of ‘healthy’ (‘medical’). There is a range of things that can be called ‘healthy’: people, diets, exercise, complexions, etc. Not all of these are healthy in the same sense. Exercise is healthy in the sense of being productive of health; a clear complexion is healthy in the sense of being symptomatic of health; a person is healthy in the sense of having good health.

But notice that these various senses have something in common: a reference to one central thing, health, which is actually possessed by only some of the things that are spoken of as ‘healthy’, namely, healthy organisms, and these are said to be healthy in the primary sense of the term. Other things are considered healthy only in so far as they are appropriately related to things that are healthy in this primary sense.

The situation is the same, Aristotle claims, with the term ‘being’. It, too, has a primary sense as well as related senses in which it applies to other things because they are appropriately related to things that are called ‘beings’ in the primary sense. The beings in the primary sense are substances; the beings in other senses are the qualities, quantities, etc., that belong to substances. An animal, e.g., a horse, is a being, and so is a color, e.g, white, a being. But a horse is a being in the primary sense — it is a substance — whereas the color white (a quality) is a being only because it qualifies some substance. An account of the being of anything that is, therefore, will ultimately have to make some reference to substance. Hence, the science of being qua being will involve an account of the central case of beings — substances.

§4. The Fundamental Principles: Axioms

Before embarking on this study of substance, however, Aristotle goes on in Book Γ to argue that first philosophy, the most general of the sciences, must also address the most fundamental principles — the common axioms — that are used in all reasoning. Thus, first philosophy must also concern itself with the principle of non-contradiction (PNC): the principle that “the same attribute cannot at the same time belong and not belong to the same subject and in the same respect” (1005b19). This, Aristotle says, is the most certain of all principles, and it is not just a hypothesis. It cannot, however, be proved, since it is employed, implicitly, in all proofs, no matter what the subject matter. It is a first principle, and hence is not derived from anything more basic.

What, then, can the science of first philosophy say about the PNC? It cannot offer a proof of the PNC, since the PNC is presupposed by any proof one might offer — any purported proof of the PNC would therefore be circular. Aristotle thus does not attempt to prove the PNC; in the subsequent chapters of Γ he

argues, instead, that it is impossible to disbelieve the PNC. Those who would claim to deny the PNC cannot, if they have any beliefs at all, believe that it is false. For one who has a belief must, if he is to express this belief to himself or to others, say something — he must make an assertion. He must, as Aristotle says, signify something. But the very act of signifying something is possible only if the PNC is accepted. Without accepting the PNC, one would have no reason to think that his words have any signification at all — they could not mean one thing rather than another. So anyone who makes any assertion has already committed himself to the PNC. Aristotle thus does not argue that the PNC is a necessary truth (that is, he does not try to prove the PNC); rather, he argues that the PNC is indubitable. (For more on the PNC, see the discussion in the entry on Aristotle's logic)

§5. What is Substance?

In the seventeen chapters that make up Book Z of the *Metaphysics*, Aristotle takes up the promised study of substance. He begins by reiterating and refining some of what he said in Γ: that 'being' is said in many ways, and that the primary sense of 'being' is the sense in which substance are beings. Here, however, he explicitly links the secondary senses of 'being' to the non-substance categories. The primacy of substance leads Aristotle to say that the age-old question 'What is being?' "is just the question 'What is substance?'" (1028b4).

One might have thought that this question had already been answered in the *Categories*. There we were given, as examples of primary substances, an individual man or horse, and we learned that a primary substance is "what is neither in a subject nor said of a subject" (2a10). This would seem to provide us with both examples of, and criteria for being, primary substances. But in *Metaphysics* Z, Aristotle does not seem to take either the examples or the criteria for granted.

In Z.2 he recounts the various answers that have been given to the question of which things are substances — bodies (including plants, animals, the parts of plants and animals, the elements, the heavenly bodies), things more basic than bodies (surfaces, lines, and points), imperceptible things (such as Platonic Forms and mathematical objects) — and seems to regard them all as viable candidates at this point. He does not seem to doubt that the clearest examples of substances are perceptible ones, but leaves open the question whether there are others as well.

Before answering this question about examples, however, he says that we must first answer the question about criteria: what is it to be a substance (*tên ousian prôton ti estin*)? The negative criterion ("neither in a subject nor said of a subject") of the *Categories* tells us only which things are substances. But even if we know *that* something is a substance, we must still say what *makes* it a substance — what the cause is of its being a substance. This is the question to which Aristotle next turns. To answer it is to identify, as Aristotle puts it, the substance *of* that thing.

§6. Substance, Matter, and Subject

Z.3 begins with a list of four possible candidates for being the substance of something: essence, universal, genus, and subject. Presumably, this means that if *x* is a substance, then the substance of *x* might be either (i) the essence of *x*, or (ii) some universal predicated of *x*, or (iii) a genus that *x* belongs to, or (iv) a subject of which *x* is predicated. The first three candidates are taken up in later chapters, and Z.3 is devoted to an examination of the fourth candidate: the idea that the substance of something is a subject of which it is predicated.

A subject, Aristotle tells us, is “that of which everything else is predicated, while it is itself not predicated of anything else” (1028b36). This characterization of a subject is reminiscent of the language of the *Categories*, which tells us that a primary substance is not predicated of anything else, whereas other things are predicated of it. Candidate (iv) thus seems to reiterate the *Categories* criterion for being a substance. But there are two reasons to be wary of drawing this conclusion. First, whereas the subject criterion of the *Categories* told us that substances were the ultimate subjects of predication, the subject criterion envisaged here is supposed to tell us what the substance *of* something is. So what it would tell us is that if *x* is a substance, then the substance of *x* — that which makes *x* a substance — is a subject that *x* is predicated of. Second, as his next comment makes clear, Aristotle has in mind something other than this *Categories* idea. For the subject that he here envisages, he says, is either matter or form or the compound of matter and form. These are concepts from Aristotle's *Physics*, and none of them figured in the ontology of the *Categories*. To appreciate the issues Aristotle is raising here, we must briefly compare his treatment of the notion of a subject in the *Physics* with that in the *Categories*.

In the *Categories*, Aristotle was concerned with subjects of predication: what are the things we talk about, and ascribe properties to? In the *Physics*, his concern is with subjects of change: what is it that bears (at different times) contrary predicates and persists through a process of change? But there is an obvious connection between these conceptions of a subject, since a subject of change must have one predicate belonging to it at one time that does not belong to it at another time. Subjects of change, that is, are also subjects of predication. (The converse is not true: numbers are subjects of predication — six is even, seven is prime — but not of change.)

In the *Categories*, individual substances (a man, a horse) were treated as fundamental subjects of predication. They were also understood, indirectly, as subjects of change. (“A substance, one and the same in number, can receive contraries. An individual man, for example, being one and the same, becomes now pale and now dark, now hot and now cold, now bad and now good” 4a17-20.) These are changes in which substances move, or alter, or grow. What the *Categories* did not explore, however, are changes in which substances are generated or destroyed. But the theory of change Aristotle develops in the *Physics* requires some other subject for changes such as these — a subject of which substance is predicated — and it identifies matter as the fundamental subject of change (192a31-32). Change is seen in the *Physics* as a process in which matter either takes on or loses form.

The concepts of matter and form, as we noted, are absent from the *Categories*. Individual substances — this man or that horse — apart from their accidental characteristics — the qualities, etc., that inhere in them — are viewed in that work as essentially simple, unanalyzable atoms. Although there is metaphysical structure to the fact that, e.g., *this horse is white* (a certain quality inheres in a certain

substance), the fact that *this is a horse* is a kind of brute fact, devoid of metaphysical structure. This horse is a primary substance, and *horse,* the species to which it belongs, is a secondary substance. But there is no predicative complex corresponding to the fact that this is a horse in the way that there is such a complex corresponding to the fact that this horse is white.

But from the point of view of the *Physics,* substantial individuals are seen as predicative complexes (cf. Matthen 1987); they are hylomorphic compounds — compounds of matter and form — and the subject criterion looks rather different from the hylomorphic perspective. *Metaphysics* Z.3 examines the subject criterion from this perspective.

Matter, form, and the compound of matter and form may all be considered subjects, Aristotle tells us, (1029a2-4), but which of them is substance? The subject criterion by itself leads to the answer that the substance of *x* is an entirely indeterminate matter of which *x* is composed (1029a10). For form is predicated of matter as subject, and one can always analyze a hylomorphic compound into its predicates and the subject of which they are predicated. And when all predicates have been removed (in thought), the subject that remains is nothing at all in its own right — an entity all of whose properties are accidental to it (1029a12-27). The resulting subject is matter from which all form has been expunged. (Traditional scholarship calls this "prime matter," but Aristotle does not here indicate whether he thinks there actually is such a thing.) So the subject criterion leads to the answer that the substance of *x* is the formless matter of which it is ultimately composed.

But Aristotle rejects this answer as impossible (1029a28), claiming that substance must be "separate" (*chôriston*) and "some this" (*tode ti*, sometimes translated "this something"), and implying that matter fails to meet this requirement. Precisely what the requirement amounts to is a matter of considerable scholarly debate, however. A plausible interpretation runs as follows. Being separate has to do with being able to exist independently (*x* is separate from *y* if *x* is capable of existing independently of *y*), and being some this means being a determinate individual. So a substance must be a determinate individual that is capable of existing on its own. (One might even hold, although this is controversial, that on Aristotle's account not every "this" is also "separate." A particular color or shape might be considered a determinate individual that is not capable of existing on its own — it is always the color of shape of some substance or other.) But matter fails to be simultaneously both *chôriston* and *tode ti*. The matter of which a substance is composed may exist independently of that substance (think of the wood of which a desk is composed, which existed before the desk was made and may survive the disassembly of the desk), but it is not as such any definite individual — it is just a quantity of a certain kind of matter. Of course, the matter may be construed as constituting a definite individual substance (the wood just *is*, one might say, the particular desk it composes), but it is in that sense not separate from the form or shape that makes it that substance (the wood cannot be that particular desk unless it is *a* desk). So although matter is in a sense separate and in a sense some this, it cannot be both separate and some this. It thus does not qualify as the substance of the thing whose matter it is.

§7. Substance and Essence

Aristotle turns in Z.4 to a consideration of the next candidate for substance: essence. (‘Essence’ is the standard English translation of Aristotle's curious phrase *to ti ên einai*, literally “the what it was to be” for a thing. This phrase so boggled his Roman translators that they coined the word *essentia* to render the entire phrase, and it is from this Latin word that ours derives. Aristotle also sometimes uses the shorter phrase *to ti esti*, literally “the what it is,” for approximately the same idea.) In his logical works, Aristotle links the notion of essence to that of definition (*horismos*) — “a definition is an account (*logos*) that signifies an essence” (*Topics* 102a3) — and he links both of these notions to a certain kind of *per se* predication (*kath' hauto*, literally, “in respect of itself”) — “what belongs to a thing in respect of itself belongs to it in its essence (*en tôi ti esti*)” for we refer to it “in the account that states the essence” (*Posterior Analytics*, 73a34-5). He reiterates these ideas in Z.4: “there is an essence of just those things whose *logos* is a definition” (1030a6), “the essence of a thing is what it is said to be in respect of itself” (1029b14). It is important to remember that for Aristotle, one defines things, not words. The definition of tiger does not tell us the meaning of the word ‘tiger’; it tells us what it is to be a tiger, what a tiger is said to be in respect of itself. Thus, the definition of tiger states the essence — the “what it is to be” of a tiger, what is predicated of the tiger *per se*.

Aristotle's preliminary answer (Z.4) to the question “What is substance?” is that substance is essence, but there are important qualifications. For, as he points out, “definition (*horismos*), like ‘what it is’ (*ti esti*), is said in many ways” (1030a19). That is, items in all the categories are definable, so items in all the categories have essences — just as there is an essence of man, there is also an essence of white and an essence of musical. But, because of the *pros hen* equivocity of ‘is’, such essences are secondary — “definition and essence are primarily (*protôs*) and without qualification (*haplôs*) of substances” (1030b4-6). Thus, Z.4 tells us, it is only these primary essences that are substances. Aristotle does not here work out the details of this “hierarchy of essences” (Loux, 1991), but it is possible to reconstruct a theory of such a hierarchy on the basis of subsequent developments in Book Z.

In Z.6, Aristotle goes on to argue that if something is “primary” and “spoken of in respect of itself (*kath' hauto legomenon*)” it is one and the same as its essence. The precise meaning of this claim, as well as the nature and validity of the arguments offered in support of it, are matters of scholarly controversy. But it does seem safe to say that Aristotle thinks that an “accidental unity” such as a pale man is not a *kath' hauto legomenon* (since pallor is an accidental characteristic of a man) and so is not the same as its essence. *Pale man*, that is to say, does not specify the “what it is” of any primary being, and so cannot be an essence of the primary kind. As Z.4 has already told us, “only species of a genus have an essence” (1030a11-12) in the primary sense. *Man* is a species, and so there is an essence of man; but *pale man* is not a species and so, even if there is such a thing as the essence of pale man, it is not, at any rate, a primary essence.

At this point there appears to be a close connection between the essence of a substance and its species (*eidos*), and this might tempt one to suppose that Aristotle is identifying the substance of a thing (since the substance of a thing is its essence) with its species. (A consequence of this idea would be that Aristotle is radically altering his conception of the importance of the species, which in the *Categories* he called a secondary substance, that is, a substance only in a secondary sense.) But such an identification would be a mistake, for two reasons. First, Aristotle's point at 1030a11 is not that a species is an essence,

but that an essence of the primary kind corresponds to a species (e.g., *man*) and not to some more narrowly delineated kind (e.g., *pale man*). Second, the word ‘*eidos*’, which meant ‘species’ in the logical works, has acquired a new meaning in a hylomorphic context, where it means ‘form’ (contrasted with ‘matter’) rather than ‘species’ (contrasted with ‘genus’). In the conceptual framework of *Metaphysics* Z, a universal such as *man* or *horse* — which was called a species and a secondary substance in the *Categories* — is construed as “not a substance, but a compound of a certain formula and a certain matter, taken universally” (Z.10, 1035b29-30). The *eidos* that is primary substance in Book Z is not the species that an individual substance belongs to but the form that is predicated of the matter of which it is composed (Cf. Driscoll 1981).

§8. Substances as Hylomorphic Compounds

The role of form in this hylomorphic context is the topic of Z.7-9. (Although these chapters were almost certainly not originally included in Book Z — there is no reference to them, for example, in the summary of Z given in H.1, which skips directly from Z.6 to Z.10 — they provide a link between substance and form and thus fill what would otherwise be a gap in the argument.) Since individual substances are seen as hylomorphic compounds, the role of matter and form in their generation must be accounted for. Whether we are thinking of natural objects, such as plants and animals, or artifacts, such as houses, the requirements for generation are the same. We do not produce the matter (to suppose that we do leads to an infinite regress) nor do we produce the form (what could we make it out of?); rather, we put the form into the matter, and produce the compound (Z.8, 1033a30-b9). Both the matter and the form must pre-exist (Z.9, 1034b12). But the source of motion in both cases — what Aristotle calls the “moving cause” of the coming to be — is the form.

In artistic production, the form is found in the soul of the artisan, for “the art of building is the form of the house” (1034a24) and “the form is in the soul” (1032b23) of the artisan. For example, the builder has in mind the plan or design for a house and he knows how to build; he then “enmatters” that plan or design by putting it into the materials out of which he builds the house. In natural production, the form is found in the parent, where “the begetter is the same in kind as the begotten, not one in number but one in form — for man begets man” (1033b30-2). But in either case, the form pre-exists and is not produced (1033b18).

As for what is produced in such hylomorphic productions, it is correctly described by the name of its form, not by that of its matter. What is produced is a house or a man, not bricks or flesh. Of course, what is made of gold may still be described in terms of its material components, but we should call it not “gold” but “golden” (1033a7). For if gold is the matter out of which a statue is made, there was gold present at the start, and so it was not gold that came into being. It was a statue that came into being, and although the statue is golden — i.e., made of gold — it cannot be identified with the gold of which it was made.

The essence of such a hylomorphic compound is evidently its form, not its matter. As Aristotle says “by form I mean the essence of each thing, and its primary substance” (1032b1), and “when I speak of

substance without matter I mean the essence” (1032b14). It is the form of a substance that makes it the kind of thing that it is, and hence it is form that satisfies the condition initially required for being the *substance of* something. The substance of a thing is its form.

§9. Substance and Definition

In Z.10 and 11, Aristotle returns to the consideration of essence and definition left off in Z.6, but now within the hylomorphic context developed in Z.7-9. The main question these chapters consider is whether the definition of *x* ever includes a reference to the matter of *x*. If some definitions include a reference to matter, then the link between essence and form would seem to be weakened.

Aristotle begins Z.10 by endorsing the following principle about definitions and their parts: “a definition is an account, and every account has parts, and part of the account stands to part of the thing in just the same way that the whole account stands to the whole thing” (1034b20-22). That is, if *y* is a part of a definable thing *x*, then the definition of *x* will include as a part something *z* that corresponds to *y*. Indeed, *z* must stand to *y* in the same relation that the definition of *x* stands in to *x*; that is, *z* is the definition of *y*. So, according to this principle, the definition of a thing will include the definitions of its parts.

In a way, this consequence of the principle seems very plausible, given Aristotle's idea that it is universals that are definable (Z.11, 1036a29). Consider as a definiendum a universal, such as *man*, and its definiens, *rational animal*. The parts of this definiens are the universals *rational* and *animal*. If these parts are, in turn, definable, then each should be replaced, in the definition of *man*, with its own definition, and so on. In this way the complete and adequate definition of a universal such as *man* will contain no parts that are further definable. All proper, or completely analyzed, definitions are ultimately composed of simple terms that are not further definable.

But the implication of this idea for the definitions of hylomorphic compounds is obvious: since matter appears to be a part of such a compound, the definition of the compound will include, as a part, the definitions of its material components. And this consequence seems implausible to Aristotle. A circle, for example, seems to be composed of two semicircle (for it obviously may be divided into two semicircle), but the definition of *circle* cannot be composed of the definitions of its two semicircular parts. For, as Aristotle points out (1035b9), *semicircle* is defined in terms of *circle*, and not the other way around. His point is well taken, for if circles were defined in terms of semicircles, then presumably semicircles would be defined in terms of the quarter-circles of which they are composed, and so on, *ad infinitum*. The resulting infinite regress would make it impossible to define *circle* at all, for one would never reach the ultimate “simple” parts of which such a definition would be composed.

Aristotle flirts with the idea of distinguishing between different senses in which one thing can be a part of another (1034b33), but instead proposes a different solution: to specify carefully the whole of which the matter is allegedly a part. “The bronze is part of the compound statue, but not of the statue spoken of as form” (1035a6). Similarly, “the line when divided passes away into its halves, and the man into bones and muscle and flesh, but it does not follow that they are composed of these as parts of their essence”

(1035a17-20). Rather, “it is not the substance but the compound that is divided into the body and its parts as into matter” (1035b21-2).

In restating his point “yet more clearly” (1035b4), Aristotle notes parenthetically another important aspect of his theory of substance. He reiterates the priority of form, and its parts, to the matter into which a compound is divided, and notes that “the soul of animals (for this is the substance of living things) is their substance” (1035b15). The idea recurs in Z.11, where he announces that “it is clear that the soul is the primary substance and the body is matter” (1037a5). It is further developed, in the *Metaphysics*, in Z.17, as we will see below, and especially in *De Anima*. For more detail on this topic, see the entry on Aristotle's psychology.

Returning now to the problem raised by the apparent need to refer to matter in the definition of a substance, we may note that the solution Aristotle offered in Z.10 is only partially successful. His point seems to be that whereas bronze may be a part of a particular statue, neither that particular batch of bronze nor even bronze in general enters into the essence of statue, since being made of bronze is no part of what it is to be a statue. But that is only because statues, although they must be made of some kind of matter, do not require any particular kind of matter. But what about kinds of substances that do require particular kinds of matter? Aristotle's distinction between form and compound cannot be used in such cases to isolate essence from matter. Thus there may after all be reasons for thinking that reference to matter will have to intrude into at least some definitions.

In Z.11, Aristotle addresses just such a case (although the passage is difficult and there is disagreement over its interpretation). “The form of man is always found in flesh and bones and parts of this kind,” Aristotle writes (1036b4). The point is not just that each particular man must be made of matter, but that each one must be made of matter of a particular kind — flesh and bones, etc. “Some things,” he continues, “surely are a particular form in a particular matter” (1036b23), so that it is not possible to define them without reference to their material parts (1036b28). Nevertheless, Aristotle ends Z.11 as if he has defended the claim that definition is of the form alone. Perhaps his point is that whenever it is essential to a substance that it be made of a certain kind of matter (e.g., that man be made of flesh and bones, or that “a saw cannot be made of wool or wood,” H.4, 1044a28) this is in some sense a formal or structural requirement. A kind of matter, after all, can itself be analyzed hylomorphically — bronze, for example, is a mixture of copper and tin according to a certain ratio or formula (*logos*), which is in turn predicated of some more generic underlying subject. The reference to matter in a definition will thus always be to a certain kind of matter, and hence to a predicate, rather than a subject. At any rate, if by ‘matter’ one has in mind the ultimate subject alluded to in Z.3 (so-called ‘prime matter’), there will be no reference to it in any definition, “for this is indefinite” (1037a27).

Z.12 introduces a new problem about definitions — the so-called “unity of definition.” The problem is this: definitions are complex (a definiens is always some combination of terms), so what accounts for the definiendum being *one* thing, rather than many (1037b10)? Man, for example, is defined as *rational animal*; “why is this one and not many — *rational* and *animal*?” (1037b13-14). Presumably, Aristotle has in mind his discussion in Z.4 of such “accidental unities” as a pale man. The difference cannot be that our language contains a single word (‘man’) for a rational animal, but no single word for a pale man, for

Aristotle has already conceded (1029b28) that we might very well have had a single term (he suggests *himation*, literally ‘cloak’) for a pale man, but that would still not make the formula ‘pale man’ a definition nor *pale man* an essence (1030a2).

Aristotle proposes a solution that applies to definitions reached by the “method of division.” According to this method (see Aristotle's logic), one begins with the broadest genus containing the species to be defined, and divides the genus into two sub-genera by means of some differentia. One then locates the definiendum in one of the sub-genera, and proceeds to divide this by another differentia, and so on, until one arrives at the definiendum species. This is a classic definition by genus and differentia. Aristotle's proposal is that “the division should be by the differentia of the differentia” (1038a9). For example, if one uses the differentia *footed* to divide the genus *animal*, one then uses a differentia such as *cloven-footed* for the next division. If one divides in this way, Aristotle claims, “clearly the last (or completing, *teleutaia*) differentia will be the substance of the thing and its definition” (1038a19). For each “differentia of a differentia” entails its predecessor (being cloven-footed entails being footed), and so the long chain of differentiae can be replaced simply by the last differentia. As Aristotle points out, “saying *footed two-footed animal* . . . is saying the same thing more than once” (1038a22-24).

This proposal shows how a long string of differentiae in a definition can be reduced to one, but it does not solve the problem of the unity of definition. For we are still faced with the apparent fact that genus + differentia constitutes a plurality even if the differentia is the last, or “completing,” one. It is not surprising, then, that Aristotle returns to the problem of unity later (H.4) and offers a different solution.

§10. Substances and Universals

At this point, we seem to have a clear idea about the nature of substantial form as Aristotle conceives of it. A substantial form is the essence of a substance, and it corresponds to a species. Since it is an essence, a substantial form is what is denoted by the definiens of a definition. Since only universals are definable, substantial forms are universals. That substantial forms are universals is confirmed by Aristotle's comment, at the end of Z.8, that “Socrates and Callias are different because of their matter . . . but they are the same in form” (1034a6-8). For them to be the same in form is for them to have the same form, i.e., for one and the same substantial form to be predicated of two different clumps of matter. And being “predicated of many” is what makes something a universal (*De Interpretatione* 17a37).

But Z.13 throws our entire understanding into disarray. Aristotle begins by returning to the candidates for the title of *ousia* introduced in Z.3, and points out that having now discussed the claims of the subject and the essence, it is time to consider the third candidate, the universal. But the remainder of the chapter consists of a barrage of arguments to the conclusion that universals are not substances.

Z.13 therefore produces a fundamental tension in Aristotle's metaphysics that has fragmented his interpreters. Some maintain that Aristotle's theory is ultimately inconsistent, on the grounds that it is committed to all three of the following propositions:

(i) Substance is form.

(ii) Form is universal.

(iii) No universal is a substance.

Others have provided interpretations according to which Aristotle does not maintain all of (i) - (iii), and there is a considerable variety of such interpretations, too many to be canvassed here. But there are two main, and opposed, lines of interpretation. According to one, Aristotle's substantial forms are not universals after all, but each belongs exclusively to the particular whose form it is, and there are therefore as many substantial forms of a given kind as there are particulars of that kind. According to the other, Aristotle's arguments in Z.13 are not intended to show that no universal is a substance, *tout court*, but some weaker thesis that is compatible with there being only one substantial form for all of the particulars belonging to the same species. Proponents of particular forms (or essences) include Sellars 1957, Harter 1975, Hartman 1977, Irwin 1988, Witt 1989b. Opponents include Woods 1967, Owen 1978, Code 1986, Loux 1991, Lewis 1991.

It would be foolish to attempt to resolve this issue within the confines of the present entry, as it is perhaps the largest, and most disputed, single interpretative issue concerning Aristotle's *Metaphysics*. I will, instead, mention some of the main considerations brought up on each side of this dispute, and give my reasons for thinking that substantial forms are universals.

The idea that substantial forms are particulars is supported by Aristotle's claims that a substance is “separate and some this” (*chôriston kai tode ti*, Z.3), that there are no universals apart from their particulars (Z.13), and that universals are not substances (Z.13). On the other side, the idea that substantial forms are universals is supported by Aristotle's claims that substances are, *par excellence*, the definable entities (Z.4), that definition is of the universal (Z.11), and that it is impossible to define particulars (Z.15).

In my opinion, the indefinability of particulars makes it impossible for substantial forms to be particulars. If there were a substantial form that is unique to some sensible particular, say Callias, then the definition corresponding to that form, or essence, would apply uniquely to Callias — it would define him, which is precisely what Aristotle says cannot be done. The question, then, is whether the evidence against substantial forms being universals can be countered. This is less clear, but the following considerations are relevant. (1) Aristotle's claim that a substantial form is an individual (*tode ti*) does not exclude its being a universal (*katholou*). Universals are contrasted with particulars (*kath' hekasta*), not individuals (although Aristotle does sometimes ignore the distinction between *tode ti* and *kath' hekaston*). What makes something a *tode ti* is its being a fully determinate thing, not further differentiable; what makes something a *kath' hekaston* is its being a particular thing, unrepeatable, and not predicated of anything else. There is thus the possibility of a universal *tode ti* — a fully determinate universal not further divisible into lower-level universals, but predicated of numerous particulars. (2) The claim that there are no universals apart from particulars needs to be understood in context. When Aristotle asserts (1038b33) that “there is no animal apart from the particulars (*ta tina*)” he is just as likely to be referring to the particular *kinds* of animals as he is to particular specimens. If so, his point may be that a generic kind,

such as animal, is ontologically dependent on its species, and hence on the substantial forms that are the essences of those species. (3) The arguments of Z.13 against the substantiality of universals are presented as part of a give-and-take investigation of the perplexities involved in the notion of substantial form. It is not clear, therefore, whether the blanket claim “No universal is a substance” is intended to be accepted without qualification. Indeed, a closer examination of the arguments may show that qualifications are required if the arguments are to be cogent. For example, the argument at 1038b11-15 is based on the premise that the substance of *x* is peculiar (*idion*) to *x*. It then draws the conclusion that a universal cannot be the substance of all of its instances (for it could not be *idion* to all of them), and concludes that it must be the substance “of none.” But note that this conclusion does not say that no universal can be a substance, but only that no universal can be the substance of any of its instances (cf. Code 1978). Aristotle's point may be that since form is predicated of matter, a substantial form is predicated of various clumps of matter. But it is not the substance of those clumps of matter, for it is predicated accidentally of them. The thing with which it is uniquely correlated, and of which it is the substance, is not one of its instances, but is the substantial form *itself*. This conclusion should not be surprising in light of Aristotle's claim in Z.6 that “each substance is one and the same as its essence.” A universal substantial form just is that essence.

§11. Substance as Cause of Being

In Z.17 Aristotle proposes a new point of departure in his effort to say what sort of a thing substance is. The new idea is that a substance is a “principle and a cause” (*archê kai aitia*, 1041a9) of being. Before looking at the details of his account, we will need to make a brief detour into Aristotle's theory of causes. The relevant texts are *Physics* II.3, *Posterior Analytics* II.11, and *Metaphysics* A.3 and Δ.2. See also the entries on [*Aristotle's physics*] and [Aristotle's psychology].

The word *aitia* (“cause” or, perhaps better, “explanation”), Aristotle tells us, is “said in many ways.” In one sense, a cause is “that out of which a thing comes to be, and which persists; e.g., bronze, silver, and the genus of these are causes of a statue or a bowl” (*Physics* 194b24). A cause in this sense has been traditionally called a *material* cause, although Aristotle himself did not use this label. In a second sense, a cause is “the form . . . the account of the essence” (194b27), traditionally called the *formal* cause. A third sense, traditionally called the *efficient* cause, is “the primary source of change or rest” (194b30). In this sense, Aristotle says, an adviser is the cause of an action, a father is the cause of his child, and in general the producer is the cause of the product. Fourth is what is traditionally called the *final* cause, which Aristotle characterizes as “the end (*telos*), that for which a thing is done” (194b33). In this sense, he says, health is the cause of walking, since we might explain a person's walking by saying that he walks in order to be healthy — health is what the walking is *for*. Note that, as in this case, “things may be causes of one another — hard work of fitness, and fitness of hard work — although not in the same sense: fitness is what hard work is for, whereas hard work is principle of motion” (195a10). So hard work is the efficient cause of fitness, since one becomes fit by means of hard work, while fitness is the final cause of hard work, since one works hard in order to become fit.

Although Aristotle is careful to distinguish four different kinds of cause (or four different senses of

‘cause’), it is important to note that he claims that one and the same thing can be a cause in more than one sense. As he puts it, “form, mover, and *telos* often coincide” (198a25). And in *De Anima* he is perfectly explicit that the soul, which is the form or essence of a living thing, “is a cause in three of the ways we have distinguished” (415b10) — efficient, formal, and final.

Let us return to Aristotle's discussion in Z.17. The job of a cause or principle of being, he notes, is to explain why one thing belongs to another (1041a11); that is, it is to explain some predicational fact. What needs to be explained, for example, is why *this is a man*, or *that is a house*. But what kind of a question is this? The only thing that can be a man is a man; the only thing that can be a house is a house. So we would appear to be asking why a man is a man, or why a house is a house, and these seem to be foolish questions that all have the same answer: because each thing is itself (1041a17-20). The questions must therefore be rephrased by taking advantage of the possibility of a hylomorphic analysis. We must ask, e.g., “Why are these things, i.e., bricks and stones, a house?” (1041a26). The answer Aristotle proposes is that the cause of being of a substance (e.g., of a house) is the form or essence that is predicated of the matter (e.g., of the bricks and stones) that constitute that substance. The essence is not always just a formal cause; in some cases, Aristotle says, it is also a final cause (he gives the examples of a house and a bed), and in some cases an efficient cause (1041a29-30). But in any case “what we seek is the cause, i.e., the form, by reason of which the matter is some definite thing; and this is the substance of the thing” (1041b6-9) and “the primary cause of its being” (1041b27).

Notice that the explanandum in these cases (“why is this a man?” or “why is that a house?”) involves a species predication (“Callias is a man,” “Fallingwater is a house”). But the answer Aristotle proposes invokes a hylomorphic analysis of these questions, in which form is predicated of matter. So Callias is a man because the form or essence of man is present in the flesh and bones that constitute the body of Callias; Fallingwater is a house because the form of house is present in the materials of which Fallingwater is made. In general, a species predication is explained in terms of an underlying form predication, whose subject is not the particular compound but its matter. Form predications are thus more basic than their corresponding species predications. A substantial form, as a primary definable, is its own substance, for it is essentially predicated of itself alone. But the substantial form of a material compound, because it is predicated (accidentally) of the matter of the compound, is the cause of the compound's being the kind of thing that it is. The form is therefore, in a derivative way, the substance of the compound, as well.

§15. Glossary of Aristotelian Terminology

- Accident: *sumbebêkos*
- Accidental: *kata sumbebêkos*
- Account: *logos*
- Actuality: *energeia*, *entelecheia*
- Alteration: *alloiôsis*
- Affirmative: *kataphatikos*
- Assertion: *apophansis* (sentence with a truth value, declarative sentence)

- Assumption: *hupothesis*
- Attribute: *pathos*
- Axiom: *axioma*
- Be: *einai*
- Being(s): *on*, *onta*
- Belong: *huparchein*
- Category: *katêgoria*
- Cause: *aition*, *aitia*
- Come to be: *gignesthai*
- Coming to be: *genesis*
- Contradict: *antiphanai*
- Contradiction: *antiphasis* (in the sense “contradictory pair of propositions” and also in the sense “denial of a proposition”)
- Contrary: *enantion*
- Definition: *horos*, *horismos*
- Demonstration: *apodeixis*
- Denial (of a proposition): *apophasis*
- Dialectic: *dialektikê*
- Differentia: *diaphora*; specific difference, *eidopoios diaphora*
- Distinctive: *idios*, *idion*
- End: *telos*
- Essence: *to ti ên einai*, *to ti esti*
- Essential: *en tôi ti esti*, *en tôi ti ên einai* (of predications); *kath' hauto* (of attributes)
- Exist: *einai*
- Explanation: *aition*, *aitia*
- Final cause: *hou heneka* (literally, “what something is for”)
- Form: *eidos*, *morphê*
- Formula: *logos*
- Function: *ergon*
- Genus: *genos*
- Homonymous: *homônumon*
- Immediate: *amesos*
- Impossible: *adunaton*
- In respect of itself: *kath' hauto*
- Individual: *atomon*, *tode ti*
- Induction: *epagôgê*
- Infinite: *apeiron*
- Kind: *genos*, *eidos*
- Knowledge: *epistêmê*
- Matter: *hulê*
- Nature: *phusis*
- Negation (of a term): *apophasis*
- Particular: *en merei*, *epi meros* (of a proposition); *kath'hekaston* (of individuals)

- Peculiar: *idios, idion*
- Per se: *kath' hauto*
- Perception: *aisthêsis*
- Perplexity: *aporia*
- Possible: *dunaton, endechomenon*; *endechesthai* (verb: "be possible")
- Potentially: *dunamei*
- Potentiality: *dunamis*
- Predicate: *katêgorein* (verb); *katêegoroumenon*("what is predicated")
- Predication: *katêgoria* (act or instance of predicating, type of predication)
- Principle: *archê* (starting point of a demonstration)
- Qua: *hêi*
- Quality: *poion*
- Quantity: *poson*
- Refute: *elenchein*; refutation*, elenchos*
- Separate: *chôriston*
- Said in many ways: *pollachôs legetai*
- Science: *epistêmê*
- Soul: *psuchê*
- Species: *eidos*
- Specific: *eidopoios* (of a differentia that "makes a species", *eidopoios diaphora*)
- Subject: *hupokeimenon*
- Substance: *ousia*
- Term: *horos*
- This: *tode ti*
- Universal: *katholou* (both of propositions and of individuals)
- Wisdom: *sophia*

Bibliography

- Ackrill, J. L. 1963. *Aristotle: Categories and De Interpretatione*. Oxford: Clarendon Press.
- Addis, L. 1972. "Aristotle and the Independence of Substances." *Philosophy and Phenomenological Research* 54: 699-708.
- Allen, R. E. 1969. "Individual Properties in Aristotle's *Categories*." *Phronesis* 14: 31-39.
- Annas, J. 1974. "Individuals in Aristotle's *Categories*: Two Queries," *Phronesis* 19: 146-152.
- Annas, J. 1976. *Aristotle: Metaphysics Books M and N*. Oxford: Clarendon Press.
- Anscombe, G. E. M. 1953. "The Principle of Individuation." *Proceedings of the Aristotelian Society, Supplementary Volume* 27: 83-96. Reprinted in J. Barnes, M. Schofield, and R. R. K. Sorabji (eds.), *Articles on Aristotle, Vol 3. Metaphysics*. London: Duckworth (1979). 88-95.
- Block, I. 1978. "Substance in Aristotle." In G. C. Simmons (ed.), *Paideia: Special Aristotle Issue*. Brockport, NY. 59-64.
- Bolton, R. 1995. "Science and the Science of Substance in Aristotle's *Metaphysics* Z." *Pacific Philosophical Quarterly* 76: 419-469.

- Bostock, D. 1994. *Aristotle: Metaphysics Books Z and H*. Oxford: Clarendon Press.
- Brody, B. A. 1973. “Why Settle for Anything Less Than Good Old-fashioned Aristotelian Essentialism?” *Noûs* 7: 351-365.
- Burnyeat, M. F. et al. 1979. *Notes on Book Zeta of Aristotle's Metaphysics*. Oxford: Sub-faculty of Philosophy.
- Chappell, V. 1973. “Aristotle's Conception of Matter.” *Journal of Philosophy* 70: 679-696.
- Charlton, W. 1972. “Aristotle and the Principle of Individuation.” *Phronesis* 17: 239-249.
- Charlton, W. 1983. “Prime Matter: a Rejoinder.” *Phronesis* 28: 197-211.
- Chen, Chung-Hwan. 1957. “Aristotle's Concept of Primary Substance in Books Z and H of the *Metaphysics*.” *Phronesis* 2: 46-59.
- Code, Alan. 1978. “No Universal is a Substance: an Interpretation of *Metaphysics* Z 13, 1038b 8-15.” In G. C. Simmons (ed.), *Paideia: Special Aristotle Issue*. Brockport, NY. 65-74.
- Code, Alan. 1984. “The Aporematic Approach to Primary Being in *Metaphysics* Z.” *Canadian Journal of Philosophy* Suppl. Vol. 10: 1-20.
- Code, Alan. 1985. “On the Origins of Some Aristotelian Theses About Predication.” In J. Bogen and J. E. McGuire (eds.), *How Things Are: Studies in Predication and the History of Philosophy*. Dordrecht: Reidel. 101-131.
- Code, Alan. 1986. “Aristotle: Essence and Accident.” In R. Grandy and R. Warner (eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. Oxford: Clarendon Press. 411-439.
- Code, Alan. 1987. “Metaphysics and Logic.” In M. Matthen (ed.), *Aristotle Today: Essays on Aristotle's Ideal of Science*. Edmonton: Academic Printing and Publishing. 127-149.
- Code, Alan. 1995. “Potentiality in Aristotle's Science and Metaphysics.” *Pacific Philosophical Quarterly* 76:.
- Code, Alan. 1997. “Aristotle's Metaphysics as a Science of Principles.” *Revue Internationale de Philosophie* 51: 357-378.
- Code, Alan. 1999. “Monty Furth's Aristotle: 10 Years Later.” *Philosophical Studies*. 94: 69-80.
- Cohen, S. Marc. 1978a. “Essentialism in Aristotle.” *Review of Metaphysics* 31: 387-405.
- Cohen, S. Marc. 1978b. “Individual and Essence in Aristotle's *Metaphysics*.” In G. C. Simmons (ed.), *Paideia: Special Aristotle Issue*. Brockport, NY. 75-85.
- Cohen, S. Marc. 1984. “Aristotle and Individuation.” *Canadian Journal of Philosophy* Suppl. Vol. 10: 41-65.
- Cohen, Sheldon M. 1981. “Proper Differentiae, the Unity of Definition, and Aristotle's Essentialism.” *The New Scholasticism* 55: 229-240.
- Cohen, Sheldon M. 1984. “Aristotle's Doctrine of the Material Substrate.” *Philosophical Review* 93: 171-194.
- Cooper, John. 1988. “Metaphysics in Aristotle's Embryology.” *Proceedings of the Cambridge Philological Society* no. 214: 14-41. Reprinted in D. Devereux and P. Pellegrin (eds.), *Biologie, Logique et Metaphysique chez Aristote*. Paris: CNRS (1990). 55-84.
- Cresswell, M. J. 1971. “Essence and Existence in Plato and Aristotle.” *Theoria* 37: 91-113.
- Cresswell, M.J. 1975. “What Is Aristotle's Theory of Universals?” *American Philosophical Quarterly* 53: 238-247.
- Dahl, Norman. 1997. “Two Kinds of Essence in Aristotle: A Pale Man Is Not the Same as His

Essence." *Philosophical Review* 106: 233-265.

- Dahl, Norman. 1999. "On Substance Being the Same As Its Essence in *Metaphysics* Z 6: The Pale Man Argument." *Journal of the History of Philosophy* 37: 1-27.
- Dancy, R. 1975. "On Some of Aristotle's First Thoughts about Substances." *Philosophical Review* 84: 338-373.
- Dancy, R. 1978. "On some of Aristotle's Second Thoughts about Substances: Matter." *Philosophical Review* 87: 372-413.
- Dancy, R. 1983. "Aristotle on Existence." *Synthèse* 54: 409-442.
- Devereux, Daniel T. 1992. "Inherence and Primary Substance in Aristotle's *Categories*." *Ancient Philosophy* 12: 113-131.
- Driscoll, J. 1981. "*Eidê* in Aristotle's Earlier and Later Theories of Substance." In D. J. O'Meara (ed.), *Studies in Aristotle*. Washington: Catholic University Press. 129-159.
- Duerlinger, J. 1970. "Predication and Inherence in Aristotle's *Categories*." Phronesis 15: 179-203.
- Durrant, M. 1975. "Essence and Accident." *Mind* 84: 595-600.
- Ebert, T. 1998. "Aristotelian Accidents." *Oxford Studies in Ancient Philosophy* 16:.
- Engmann, J. 1973. "Aristotle's Distinction Between Substance and Universal." Phronesis 18: 139-155.
- Ferejohn, M.T. 1980. "Aristotle on Focal Meaning and the Unity of Science." Phronesis 25: 117-128.
- Ferejohn, M.T. 1994. "Matter, Definition and Generation in Aristotle's *Metaphysics*." *Proceedings of the Boston Area Colloquium in Ancient Philosophy* 10: 35-58.
- Fine, Gail. 1985. "Separation: a Reply to Morrison." *Oxford Studies in Ancient Philosophy* 3: 159-165.
- Fine, Kit. 1992. "Aristotle on Matter." *Mind* 101: 35-57.
- Frede, Michael. 1987. *Essays in Ancient Philosophy*. Minneapolis: University of Minnesota Press.
- Frede, Michael. 1990. "The Definition of Sensible Substances in *Met*. Z." In D. Devereux and P. Pellegrin (eds.), *Biologie, Logique et Metaphysique chez Aristote*. Paris: CNRS.
- Frede, Michael and Günther Patzig. 1988. *Aristoteles Metaphysik Z*. Munich: C. H. Beck.
- Furth, Montgomery. 1978. "Transtemporal Stability in Aristotelean Substances." *Journal of Philosophy* 75: 624-646.
- Furth, Montgomery. 1985. *Aristotle's Metaphysics 7-10*. Indianapolis: Hackett.
- Furth, Montgomery. 1987. "Aristotle on the Unity of Form." In M. Matthen (ed.), *Aristotle Today: Essays on Aristotle's Ideal of Science*. Edmonton: Academic Printing and Publishing. 77-102.
- Furth, Montgomery. 1988. *Substance, Form and Psyche: an Aristotelian Metaphysics*. Cambridge: Cambridge University Press.
- Furth, Montgomery. 1990. "Specific and Individual Forms in Aristotle." In D. Devereux and P. Pellegrin (eds.), *Biologie, Logique et Metaphysique chez Aristote*. Paris: CNRS.
- Gill, Mary Louise. 1989. *Aristotle on Substance: The Paradox of Unity*. Princeton: Princeton University Press.
- Gill, Mary Louise. 1993. "Matter against Substance." *Synthèse* 96: 379-397.
- Gill, Mary Louise. 1995. "Symposium on Aristotle on Substance and Predication." *Ancient Philosophy* 15: 511-520.
- Gotthelf, Allan. 1999. "A Biological Provenance." *Philosophical Studies*. 94: 35-56.

- Graham, D. W. 1987a. “The Paradox of Prime Matter.” *Journal of the History of Philosophy* 25: 475-490.
- Graham, D. W. 1987b. *Aristotle's Two Systems*. Oxford: Oxford University Press.
- Granger, H. 1980. “A Defense of the Traditional Position concerning Aristotle's non-substantial Particulars.” *Canadian Journal of Philosophy* 10: 593-606.
- Granger, H. 1984. “Aristotle on Genus and Differentia.” *Journal of the History of Philosophy* 22: 1-24.
- Granger, H. 1989. “Aristotle's Natural Kinds.” *Philosophy* 64: 245-247.
- Grene, M. 1974. “Is Genus to Species as Matter to Form? Aristotle and Taxonomy.” *Synthèse* 28: 51-69.
- Grice, H. P. 1988. “Aristotle on the Multiplicity of Being.” *Pacific Philosophical Quarterly* 69: 175-200.
- Halper, E.. 1987. “A Solution to the Problem of Sensible Substance.” *Journal of Philosophy* 84: 666-672.
- Halper, E. 1989. *One and Many in Aristotle's Metaphysics, the Central Books*. Columbus: Ohio State University Press.
- Harter, E. 1975. “Aristotle on Primary *ousia*.” *Archiv für Geschichte der Philosophie* 57: 1-20.
- Hartman, Edwin. 1976. “Aristotle on the Identity of Substance and Essence.” *Philosophical Review* 85: 545-561.
- Hartman, Edwin. 1977. *Substance, Body, and Soul: Aristotelian Investigations*. Princeton: Princeton University Press.
- Heinaman, R. 1981a. “Non-substantial Individuals in the *Categories*.” *Phronesis* 26: 295-307.
- Heinaman, R. 1981b. “Knowledge of Substance in Aristotle.” *Journal of Hellenic Studies* 101: 63-77.
- Heinaman, R. 1997. “Frede and Patzig on Definition in *Metaphysics* Z.10 and 11.” *Phronesis* 42: 283-298.
- Hetherington, S.C. 1984. “A Note on Inherence.” *Ancient Philosophy* 4: 218-223.
- Irwin, T. H. 1981. “Homonymy in Aristotle.” *Review of Metaphysics* 34: 523-544.
- Irwin, T. H. 1988. *Aristotle's First Principles*. Oxford: Clarendon Press.
- Jones, B. 1972. “Individuals in Aristotle's *Categories*,” *Phronesis* 17: 107-123.
- Jones, B. 1975. “An Introduction to the first five chapters of Aristotle's *Categories*.” *Phronesis* 20: 146-172.
- Kirwan, C. A. 1970. “How Strong are the Objections to Essence?” *Proceedings of the Aristotelian Society* 71: 43-59.
- Kirwan, C. A. 1971. *Aristotle: Metaphysics Books Gamma, Delta, and Epsilon*. Oxford: Clarendon Press.
- Kosman, L. A. 1984. “Substance, Being, and Energeia.” *Oxford Studies in Ancient Philosophy* 2: 121-149
- Kosman, L. A. 1999. “Aristotelian Metaphysics and Biology: Furth's *Substance, Form and Psyche*. *Philosophical Studies*. 94: 57-68.
- Kung, Joan. 1977. “Aristotelian Essence and Explanation.” *Philosophical Studies* 31: 361-383.
- Kung, Joan. 1978. “Can Substance be Predicated of Matter?” *Archiv für Geschichte der Philosophie* 60: 140-159.

- Kung, Joan. 1986. "Aristotle on 'Being is Said in Many Ways'." *History of Philosophy Quarterly* 3: 3-18.
- Lacey, A. R. 1965. "Ousia and Form in Aristotle." *Phronesis* 10: 54-69.
- Lear, Jonathan. 1988. *Aristotle: the Desire to Understand*. Cambridge: Cambridge University Press.
- LeBlond, J. M. 1979. "Aristotle on Definition." In J. Barnes, M. Schofield, and R. R. K. Sorabji (eds.), *Articles on Aristotle, Vol 3. Metaphysics*. London: Duckworth. 63-79.
- Lesher, James H. 1971. "Aristotle on Form, Substance, and Universals: a Dilemma." *Phronesis* 16: 169-178.
- Lewis, Frank A. 1982. "Accidental Sameness in Aristotle." *Philosophical Studies* 39: 1-36.
- Lewis, Frank A. 1983. "Form and Predication in Aristotle's *Metaphysics*." In J. Bogen and J. E. McGuire (eds.), *How Things Are: Studies in Predication and the History of Philosophy*. Dordrecht: Reidel. 59-83.
- Lewis, Frank A. 1984. "What is Aristotle's Theory of Essence?" *Canadian Journal of Philosophy* Suppl. Vol. 10: 89-132.
- Lewis, Frank A. 1991. *Substance and Predication in Aristotle*. Cambridge: Cambridge University Press.
- Lewis, Frank A. 1995a. "Aristotle on the Unity of Substance." *Pacific Philosophical Quarterly* 76: 222-265.
- Lewis, Frank A. 1995b. "Symposium on Substance, Predication, and Unity in Aristotle." *Ancient Philosophy* 15: 521-549.
- Lloyd, A. C. 1970. "Aristotle's Principle of Individuation." *Mind* 79: 519-529.
- Lloyd, A. C. 1981. *Form and Universal in Aristotle*. Liverpool: F. Cairns.
- Loux, Michael J. 1979. "Form, Species, and Predication in *Metaphysics* Z, H, and Θ." *Mind* 88: 1-23.
- Loux, Michael J. 1991. *Primary Ousia: An Essay on Aristotle's Metaphysics Z and H*. Ithaca, NY: Cornell University Press.
- Loux, Michael J. 1995a. "Symposium on Aristotle's *Metaphysics*." *Ancient Philosophy* 15: 495-510.
- Loux, Michael J. 1995b. "Composition and Unity: An Examination of *Metaphysics* H.6." In May Sim, *The Crossroads of Norm and Nature*. Lanham: Rowman and Littlefield.
- MacKinnon, D. M. 1965., "Aristotle's Conception of Substance." In R. Bambrough (ed.), New Essays on Plato and Aristotle. London: Routledge and Kegan Paul. 97-119.
- Malcolm, John. 1993. "On the Endangered Species of the *Metaphysics*." *Ancient Philosophy* 13: 79-93.
- Malcolm, John. 1996. "On the Duality of *Eidos* in Aristotle's *Metaphysics*." *Archiv für Geschichte der Philosophie* 78: 1-10.
- Mansion, S. 1979. "The Ontological Composition of Sensible Substances in Aristotle (*Metaphysics* VII, 7-9)." In J. Barnes, M. Schofield, and R. R. K. Sorabji (eds.), Articles on Aristotle, Vol 3. Metaphysics. London: Duckworth. 80-87.
- Matthen, Mohan. 1987. "Individual Substances as Hylomorphic Complexes." In M. Matthen (ed.), *Aristotle Today: Essays on Aristotle's Ideal of Science*. Edmonton: Academic Printing and Publishing. 151-176.

- Matthews, Gareth B. 1982. "Accidental Unities." In M. Schofield and M. C. Nussbaum, *Language and Logos: Studies in Ancient Greek Philosophy*. Cambridge: Cambridge University Press. 223-240.
- Matthews, Gareth B. 1986. "Gender and Essence in Aristotle." *American Philosophical Quarterly* Supp. 64: 16-25.
- Matthews, Gareth B. 1989. "The Enigma of *Categories* la20ff and Why it Matters." *Apeiron* 22: 91-104.
- Matthews, Gareth B. 1990. "Aristotelian Essentialism." *Philosophy and Phenomenological Research* Suppl. 50: 251-262.
- Matthews, Gareth B. 1991. "Container Metaphysics According to Aristotle's Greek Commentators." *Canadian Journal of Philosophy* Supp. vol. 17: 7-23.
- Matthews, Gareth B. and S. Marc Cohen. 1968. "The One and the Many." *Review of Metaphysics* 21: 630-655.
- Miller, F. D. 1978. "Aristotle's Use of Matter." In G. C. Simmons (ed.), *Paideia: Special Aristotle Issue*. Brockport, NY. 105-119.
- Modrak, Deborah K. 1979. "Forms, Types, and Tokens in Aristotle's *Metaphysics*." *Journal of the History of Philosophy* 17: 371-381.
- Modrak, Deborah K. 1983. "Forms and Compounds." In J. Bogen and J. E. McGuire (eds.), *How Things Are: Studies in Predication and the History of Philosophy*. Dordrecht: Reidel. 85-99.
- Modrak, Deborah K. 1989. "Aristotle on the difference between Mathematics and Physics and First Philosophy." *Apeiron* 22: 121-139.
- Moravcsik, J. M. E. 1967. "Aristotle on Predication." *Philosophical Review* 76: 80-96.
- Morrison, D. 1985b. "Separation in Aristotle's *Metaphysics*." *Oxford Studies in Ancient Philosophy* 3: 125-157.
- Morrison, D. 1985c. "Separation: a Reply to Fine." *Oxford Studies in Ancient Philosophy* 3: 167-173.
- Morrison, D. 1993. "The Place of Unity in Aristotle's Metaphysical Project." *Proceedings of the Boston Area Colloquium in Ancient Philosophy* 9: 131-156.
- Moser, P. 1983. "Two Notions of Substance in *Metaphysics* Z." *Apeiron* 17: 103-112.
- Owen, G. E. L. 1965a. "Inherence." *Phronesis* 10: 97-105.
- Owen, G. E. L. 1965b. "Aristotle on the Snares of Ontology." In R. Bambrough (ed.), *New Essays on Plato and Aristotle*. London: Routledge and Kegan Paul. 69-95.
- Owen, G. E. L. 1965c. "The Platonism of Aristotle." *Proceedings of the British Academy* 50 125-150. Reprinted in In J. Barnes, M. Schofield, and R. R. K. Sorabji (eds.), *Articles on Aristotle, Vol 1. Science*. London: Duckworth (1975). 14-34.
- Owen, G. E. L. 1978. "Particular and General." *Proceedings of the Aristotelian Society* 79: 1-21.
- Owens, Joseph. 1978. *The Doctrine of Being in the Aristotelian Metaphysics*. 3d ed., rev. Toronto: Pontifical Institute of Mediaeval Studies.
- Page, C. 1985. "Predicating Forms of Matter in Aristotle's *Metaphysics*." *Review of Metaphysics* 39: 57-82.
- Panayides, Christos. 1999. "Aristotle on the Priority of Actuality in Substance." *Ancient Philosophy* 19: 327-344.
- Pelletier, F. J. 1979. "Sameness and Referential Opacity in Aristotle." *Noûs* 13: 283-311.

- Pena, Lorenzo. 1999. “The Coexistence of Contradictory Properties in the Same Subject According to Aristotle.” *Apeiron* 32: 203-229.
- Reeve, C. D. C. 2000. *Substantial Knowledge: Aristotle's Metaphysics*. Indianapolis: Hackett.
- Regis, E. 1976. “Aristotle's ‘Principle of Individuation’.” *Phronesis* 21: 157-166.
- Robinson, H. M. 1974. “Prime Matter in Aristotle.” *Phronesis* 19: 168-188.
- Rorty, Richard. 1973. “Genus as Matter: a Reading of *Metaphysics* Z-H.” In Lee, Mourelatos, and Rorty (eds.) *Exegesis and Argument*, *Phronesis* Suppl. 1. Assen: Van Gorcum. 393-420.
- Rorty, Richard. 1974. “Matter as Goo: Comments on Grene's Paper.” *Synthèse* 28: 71-77.
- Ross, W. D. 1924. *Aristotle's Metaphysics*. Oxford: Clarendon Press.
- Scaltsas, T. 1985. “Substratum, Subject, and Substance.” *Ancient Philosophy* 5: 215-40.
- Scaltsas, T. 1994. *Substances and Universals in Aristotle's Metaphysics*. Ithaca: Cornell University Press.
- Schofield, M. 1972. “*Metaphysics* Z3: Some Suggestions.” *Phronesis* 17: 97-101.
- Sellars, Wilfrid. 1957. “Substance and Form in Aristotle.” *Journal of Philosophy* 54: 688-699.
- Shields, C. 1990. “The Generation of Form in Aristotle.” *History of Philosophy Quarterly* 7: 367-390.
- Smith, J. A. 1921. “*Tode ti* in Aristotle.” *Classical Review* 35: 19.
- Spellman, L. 1989. “Specimens of Natural Kinds and the Apparent Inconsistency of *Metaphysics* Z.” *Ancient Philosophy* 9: 49-65.
- Stahl, D. 1981. “Stripped Away: Some Contemporary Obscurities Surrounding *Metaphysics* Z3 (1029a10-26).” *Phronesis* 26: 177-180.
- Stough, C. L. 1972. “Language and Ontology in Aristotle's *Categories*.” *Journal of the History of Philosophy* 10: 261-272.
- Sykes, R. D. 1975. “Form in Aristotle: Universal or Particular?” *Philosophy* 50: 311-331.
- Teloh, H. 1979. “Aristotle's *Metaphysics* Z.13.” *Canadian Journal of Philosophy* 5: 77-79.
- Thorp, J. W. 1974. “Aristotle's Use of *Categories*.” *Phronesis* 19: 238-256.
- Tweedale, M. 1987. “Aristotle's Universals.” *American Philosophical Quarterly* 65: 412-426.
- Tweedale, M. 1988. “Aristotle's Realism.” *Canadian Journal of Philosophy* 18: 501-526.
- Verbeke, G. 1981. “Aristotle's *Metaphysics* viewed by the Ancient Greek Commentators.” In D. J. O'Meara (ed.), *Studies in Aristotle*. Washington: Catholic University Press. 107-127.
- Wedin, Michael V. 1991. “PARTisanship in *Metaphysics* Z.” *Ancient Philosophy* 11: 361-385.
- Wedin, Michael V. 1993. “Nonsubstantial Individuals.” *Phronesis* 38: 137-165.
- Wedin, Michael V. 1999. “The Scope of Non-Contradiction: A Note on Aristotle's ‘Elenctic’ Proof in *Metaphysics* gamma 4.” *Apeiron* 32: 231-242.
- Wheeler, Mark. 1999. “The Possibility of Recurrent Individuals in Aristotle's *Organon*.” *Gregorianum* 80: 539-551.
- Wheeler, S. 1977. “The Theory of Matter from *Metaphysics* ZHΘ.” *International Studies in Philosophy* 13-22.
- White, N. 1972. “The Origins of Aristotle's Essentialism.” *Review of Metaphysics* 26: 57-85.
- White, N. 1986. “Identity, Modal Individuation, and Matter in Aristotle.” *Midwest Studies in Philosophy* 11: 475-494.
- Whiting, J. E. 1986. “Form and Individuation in Aristotle.” *History of Philosophy Quarterly* 3: 359-377.

- Whiting, J. E. 1991. “Metasubstance: Critical Notice of Frede-Patzig and Furth.” *Philosophical Review* 100: 607-639.
- Williams, D. C. 1958. “Form and Matter.” *Philosophical Review* 67: 291-312, 499-521.
- Witt, Charlotte. 1987. “Hylomorphism in Aristotle.” *Journal of Philosophy* 84: 673-679.
- Witt, Charlotte. 1989a. “Aristotelian Essentialism Revisited.” *Journal of the History of Philosophy* 27: 285-298.
- Witt, Charlotte. 1989b. *Substance and Essence in Aristotle: an Interpretation of Metaphysics VII-IX*. Ithaca, NY: Cornell University Press.
- Woods, M. J. 1967. “Problems in *Metaphysics* Z, Chapter 13.” In J. Moravcsik (ed.), *Aristotle: A Collection of Critical Essay*s. New York: Anchor. 215-238.
- Woods, M. J. 1974. “Substance and Essence in Aristotle.” *Proceedings of the Aristotelian Society* 75: 167-180.
- Woods, M. J. 1991a. “Particular Forms Revisited.” *Phronesis* 26: 75-87.
- Woods, M. J. 1991b. “Universals and Particular Forms in Aristotle's *Metaphysics*.” *Oxford Studies in Ancient Philosophy* supplement. 41-56.

Other Internet Resources

- An Outline of Metaphysics Z. (by S. Marc Cohen)

Related Entries

Aristotle: logic | Aristotle: physics | Aristotle: psychology

S. Marc Cohen
smcohen@u.washington.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 8, 2000
Content last modified: September 29, 2001

Stanford Encyclopedia of Philosophy
Supplement to Aristotle's Metaphysics

Nonsubstantial Particulars

The precise nature of particulars in nonsubstance categories has been, and remains, a matter of considerable controversy. According to the traditional account (cf. Ackrill 1963), by “the particular white” (or “a certain white”, *to ti leukon*) Aristotle means a *trope*, something that is unique to the individual substance in which it inheres and is not repeatable elsewhere. If a particular white is in Socrates, then it is not in anything else. Indeed, even if Callias is of exactly the same shade of color as Socrates, the particular white in Callias is an entity distinct from the particular white in Socrates — they are two numerically distinct but qualitatively identical tropes. An alternative account (initially championed by Owen 1965a) takes “the particular white” to denote a repeatable entity: a fully determinate universal that is capable of being shared by distinct substances. It is only a particular in the sense that it is not essentially predicated (‘said of’) anything else in its own category. That is, a particular white is a fully determinate shade of color, rather than a determinable such as *white*, which is a generic classification of various determinate shades of white.

At the center of the controversy is the interpretation of Aristotle's definition of ‘in’. At *Cat*. 1a25 he says “by ‘in a subject’ I mean what is in something, not as a part, and cannot exist separately from what it is in.” The definition is clearly ambiguous. On the one hand, it might mean that what is ‘in’ a particular subject is incapable of existing separately from that subject. This is Ackrill's understanding:

> x is in $y =_{\mathrm{df}}$
>
> > (a) x belongs to y, and
> > (b) x is not a part of y, and
> > (c) x cannot exist separately from y.

On this understanding, the only thing that can, strictly speaking, be ‘in’ a particular subject (e.g., Socrates) is something that cannot exist separately from that subject. The color ‘in’ Socrates, in this sense, could not exist ‘in’ anything else. Indeed, the only thing that can be ‘in’ a particular substance, in this sense, is something that cannot exist separately from that substance.

The problem for this reading of Aristotle's definition is that specific or generic universals (such as *white* and *color*) could not be ‘in’ a particular substance; *color* could not be ‘in’ Socrates because *color* can surely exist separately from Socrates. Yet Aristotle says (2b2) that “color is in body, and therefore in an individual body (for if it were not in any individual, it would not be in body at all).” Unless Aristotle is speaking carelessly here (as Ackrill supposes), his claim cannot be consistent with the definition of ‘in’,

as Ackrill interprets it.

A second reading of Aristotle's definition is Owen's:

> *x* is in *y* $=_{\text{df}}$
>
> > (a) *x* belongs to *y*, and
> > (b) *x* is not a part of *y*, and
> > (c) *x* cannot exist on its own (i.e., *x* cannot exist unless there is something *z* such that *x* belongs to *z*)

On this understanding, it is possible for a generic quality, such as *white* or *color*, to be ‘in’ a particular substance. The reason that *white* can be ‘in’ Socrates (as well as in other individuals) is that *white* belongs to (i.e., is a property of) Socrates, not a part of him, and is incapable of existing unless it belongs to some substance or other.

A third reading of Aristotle's definition is that of Frede 1987:

> *x* is in a subject $=_{\text{df}}$
>
> > There is something, *y*, such that
> > (a) *x* is not a part of *y*, and
> > (b) *x* cannot exist independently of *y*.

Frede's reading is different from the other two in one important way, yet shares features of each. The difference is that on Frede's reading Aristotle is defining the one-place predicate ‘*x* is in a subject’, not the two-place predicate ‘*x* is in *y*’. That is, he is defining what it is to be an *accident*, the sort of thing that is ‘in’ a subject, rather than what it is for *x* to be ‘in’ *y*. This reading, like Owen's, allows a generic quality, such as *white* or *color*, to be ‘in’ a particular substance. *White* can be ‘in’ Socrates because it is an accident (i.e., it is an ‘in a subject’ sort of thing) and belongs to (i.e., is a property of) Socrates. Yet, like Ackrill's, it has a specific “inseparability” requirement. That is, in order for *x* to be an accident, there must be some sort of thing that *x* is incapable of existing separately from. In the case of *color*, that thing is *body*.

There is a vast literature on this dispute. See, in addition to Ackrill and Owen: Matthews and Cohen 1968, Allen 1969, Duerlinger 1970, Jones 1972, Annas 1974, Hartman 1977, Granger 1980, Heinaman 1981a, Frede 1987, Matthews 1989, Devereux 1992, and Wedin 1993.

S. Marc Cohen
smcohen@u.washington.edu

Return to Aristotle's Metaphysics

First published: November 1, 2000
Content last modified: November 1, 2000

Tropes

A trope is an instance or bit (not an exemplification) of a property or a relation; e.g. Clinton's eloquence, Sydney's beauty, or Pierre's love of Heloïse. Clinton's eloquence is understood here not as Clinton's participating in the universal eloquence, nor as the peculiar quality of Clinton's eloquence, but simply as Clinton's bit of eloquence, the eloquence that he and he alone has. Similarly, Pierre's love is not his participation in love as such, nor the special way he loves, but the loving peculiar to Pierre as directed toward Heloïse. The appeal of tropes for philosophers is as an ontological basis free of the postulation of obscure abstract entities such as propositions and universals.

Name and Incidence in Philosophy

The ontological theory of tropes holds that properties and relations subsist as so many instances or tropes, one for each exemplification. These tropes are particulars, not universals, distinct from the concrete particulars they characterize. By other names, trope ontologies have been espoused throughout the history of Western philosophy. According to D. W. Mertz (1996, ch. IV), variants can be found in the writings of Plato, Aristotle, Boëthius, Avicenna, Averroës, Thomas, Scotus, Buridan, Suárez, Leibniz, Husserl, the early Russell (1911), Stout, Cook Wilson, and Strawson. Tropes have been variously called "property (and relation) instances", "abstract particulars", "concrete properties", "unit properties (and relations)", "quality (and relation) bits", "individual accidents", and (in German) "*Momente*".

(Parenthesized years refer to the Bibliography below.)

The most compelling advocate of such objects in our time has been D. C. Williams (1953), who is responsible for the regrettable term *trope*. It has nothing to do with figures of speech in rhetoric, *Leitmotive* in music, or tropisms in plants. Williams coined it as a sort of philosophical joke: Santayana, he says, had employed ‘trope’ pointlessly for ‘essence of an occurrence’. Williams would go him one better and press it into service for ‘occurrence of an essence’ (1953: 78). [Far from poking fun at Santayana, Williams published an appreciation of his views on essence and occurrence in a memorial issue of the *Journal of Philosophy* (1954).] Ironically, the word ‘trope’ is to be heard correctly these days mainly from the lips of poststructuralists. Meanwhile, many trope theorists have adopted Williams’ usage, but some avoid it (e.g. Mertz). Williams acknowledged the close affinity between his trope theory and G. F. Stout’s theory of abstract particulars (1921, 1923).

Approaches to Univerals

Obviously one could see tropes as complexes of some sort, perhaps composed of particulars and universals. (I use ‘universal’ here to cover both properties and relations.) Such a construction is, indeed, very strongly suggested by the subject-predicate form of our language. Philosophical ontologists have, however, long since considered departing from this linguistic pattern in various ways. Nominalists recognize the particulars as subjects, but hold that there really are no universals beyond the linguistic predicates themselves. Plato held, by contrast, that certain universals, the Forms, are the only realities, the particulars being mere figments of belief (-380). A less radical variant of nominalism recognizes properties and relations, but as mere set-theoretic constructs out of individuals. This approach is usual in model-theoretic semantics. A less otherworldly version of Platonism takes particulars to be bundles of universals; cf. Russell (1940, ch. 6, 8, 24) and Blanshard (1939, ch. 16, 17). For those students of ontology who are not obsessed with parsimony, however, the most natural course would probably seem to be to take a leaf from our language and to recognize both exemplifying individuals and repeatedly exemplified universals. Such a view is so common that it has no particular name; Armstrong calls it the "substance-attribute view" (1989: 59 et seq.). This view need not deny that there are tropes, but it denies that they are basic or simple or primitive. Rather they must be composite structures involving a property or relation, some individuals, and an exemplification nexus. An ontology based on tropes takes the opposite approach. It recognizes tropes as basic, not as constructed. Individuals and properties then require further analysis. Ontological theories thus based on primitive tropes may be called versions of tropism or trope theory. A major attraction of tropism has been its promise of parsimony; some adherents go so far as to proclaim a one-category ontology (Campbell, Mertz).

Varieties of Trope Theory

Trope theories divide according to their treatment of universals and individuals. What I should like to regard as the classic trope theory (Stout, Williams) treats universals and individuals as constructs or bundles of tropes. This is the trope-bundle theory, called by some (Simons, Mertz) trope nominalism or

moderate nominalism (Hochberg). (‘Nominalism’ because it repudiates primitive universals; ‘moderate’ because it still recognizes unit properties). Then there are trope theories that retain either primitive individuals or primitive universals. The former position, substratum tropism, was taken in a way by Leibniz, who recognized individual substances (monads), but correlated with complete individual concepts comprising nonrelational tropelike representations of the whole world (1686: §§9, 14; 1714: §§8, 14, 17-19). A similar view is hailed by C. B. Martin (1980) in Locke (1690: 159) and noted approvingly by Armstrong (1989: 114, 136). The latter view, tropes plus primitive universals, was held by Cook Wilson (1926, vol. 2, 713 *et passim*) and may be represented also by Mertz (1996), with the important qualification that his universals are given conceptual, not Platonic status. Such a position might be called trope universalism; Mertz calls his version "moderate realism". (‘Realism’ because universals are recognized; ‘moderate’ because they are immanent: only their instances really exist.) Finally, there is the possibility of combining tropism with a full substance-attribute view. Husserl (1913-21: 430f, 436f) may perhaps be read in this way, and certain truthmaker theories may come close. (Truthmakers, like tropes, may be posited in addition to states of affairs, complexes made up of particulars and universals.)

Another significant division among trope theories separates the actualists from the meinongians. (The term alludes to no specific teaching of Meinong, just the preparedness to recognize nonexistents.) For the actualist, there is a trope, say, of Old Faithful’s heat, only if Old Faithful is actually hot. The only property instances are actual ones. For the meinongian, on the other hand, there are also tropes of Old Faithful’s coldness, Bill Clinton’s shyness, etc. (The contrast mirrors the traditional dispute over false facts or nonobtaining states of affairs.) These days actualism is popular. Meinongian tropism has, however, one great advantage: it affords a straightforward account of possible worlds (deemed by many hopelessly obscure). A possible world, on this approach, is simply a set of tropes. (There are problems with nonlogically incompatible tropes, such as *a*’s redness and *a*’s greenness, but similar problems beset other theories. Not every trope set need be a possible world.)

Trope-Bundle Theory

Classic tropism, the trope-bundle theory, would seem to hold the greatest promise of economy. For this theory dispenses with both primitive individuals and primitive universals, leaving at first glance only tropes. However, second-level bundling relations of tropes prove necessary. Tropes belong to the same individual if they are all *compresent* (concurrent) with one another. Tropes belong to the same universal (property or relation) if they *exactly resemble* one another. The two second-level trope relations of compresence and exact resemblance are essential to the bundle theory. They are similarity relations (reflexive and symmetric); compresence is also transitive, an equivalence relation on tropes. Thus universals become similarity classes and individuals equivalence classes of tropes: both are products of abstraction. (This is a first approximation: individuals may ultimately have be taken as more complicated; see Individuals Refined.) Exemplification (as expressed by predication) is then simply overlapping. On the actualist approach, Clinton is eloquent iff he (his compresence class) overlaps eloquence (the set of eloquences). The meinongian approach brings in possible worlds: Clinton is eloquent in *w* iff he, eloquence, and *w* all overlap.

Trope-bundle theory can be further developed to include a treatment of compound universals (also requiring further complications in the structure of individuals and universals) and a construction of what Bacon calls states of affairs (essentially, world sets) (1995, ch. III). The whole question of the relation of tropes to states of affairs is a vexed one, partly because intuitive conceptions of states of affairs diverge. For some, it is analytic that states of affairs are complexes, making it unthinkable for them to be tropes. Others see an extensive parallel between the two notions. The latter view is ruled out if the tropes are assumed as basic. But there is some interest in seeing what results if we plug states of affairs (complexes) into trope theory in place of tropes. Connections both to situation semantics and to Armstrong's later theory of universals (1989: 94) are revealed.

Relations

The seemingly parsimonious trope-bundle theory, as we saw, is pushed to acknowledge at least a second category besides tropes, the second-level relations. There are probably more such relations, e.g. temporal precedence and betterness. Williams advocated the obvious therapy here without working out the details. The second-level relations, he suggested, crumble into second-level tropes (1953: 84). But it should be clear that in order to bundle second-level tropes into the requisite relations, *third*-level relations will be needed, and so forth. It turns out that a significant simplification is actually achieved at the third or fourth level, so the regress is not vicious. At least one unpulverized relation is still needed though, and the third- or fourth-level tropes ultimately assumed are scarcely plausible candidates for basic constituents of reality.

Mertz points out how hostile the Western tradition has been to recognizing genuine relations (1996, ch. 6). Only Russell's early insistence on their importance turned the tide in our century. Few trope theories have a well worked out treatment even of first-level (ordinary) relations. Campbell holds that while relational *discourse* is ineliminable, relations themselves come down to their *foundations*, the properties of their relata in which they are grounded (1990: 98ff). As Mertz has pointed out (1996: 63-67), this general approach goes back at least as far as Ockham. Although Campbell does not give details, the project is not to be regarded as hopeless.

Bacon, on the other hand, retains first-level relations in the same status as properties, bundled into universals by exact resemblance (1995, ch. II). But whereas modern predicate logic treats the semantic values of relational predicates as complicated (as sets of *n*-tuples), Bacon complicates individuals. He multiplies compresence into indexed 1-compresence, 2-compresence, . . . An individual (in the new extended sense) is then a chain (sequence) of a 1-compresence-equivalence class, a 2-compresence-equivalence class, and so on. This inobvious extension makes a unified treatment of predication possible. On the actualist approach, Yeltsin is healthy iff his first compresence class overlaps health. Pierre loves Heloïse iff his first compresence class, her second compresence class, and love all overlap. The meinongian approach brings in possible worlds: Yeltsin is healthy in *w* iff his 1-compresence class, health, and w all overlap. The dyadic case is similar. Williams considered the explication of exemplification to be one of the important achievements of tropism, "do[ing] much to dispel the ancient mystery of predication" (1953: 82). Bacon extends that explication to relational predication.

Individuals Refined

For some trope theorists, a mere set of tropes, or even a chain of such, has too little inner coherence and unity to qualify as an individual. Thus Williams takes an individual to be the *mereological sum* of a compresence class (1953: 81). Martin writes, "An object is not a collectable out of its properties or qualities as a crowd is collectable out of its members. For each and every property of an object has to be had by that object to exist at all" (1980: 8). Mertz constructs individuals with the help of what he calls integrated networks (1996: 76). The integrated network of a particular *t* comprises all the atomic facts about *t.* Since the integrated network is itself a nonrepeatable individual, it can have its own integrated network, and so on. A hierarchy of such integrated networks is then an ordinary individual. Mertz appears to leave it open whether the hierarchy ever terminates. He is also vague about facts (states of affairs): they are complexes consisting of a trope and its exemplification or relata, the latter apparently also tropes. Facts serve as truthmakers. Mertz's account is developed partly to avoid positing individuals as bare particulars. The price would seem to be to obscure the truth condition for simple predication sentences.

A further refinement of bundles is offered in Simons' nuclear theory (1994). In place of compresence, Simons takes over Husserl's foundation relation (1913-22.478f). A trope s is *founded* on *t* if *t*'s existence is necessary for the existence of *s. s* and *t* are *directly foundationally related* if either is founded on the other. *Foundational relatedness*, the ancestral of direct foundational relatedness, is an equivalence relation on tropes. Its equivalence classes are *foundational systems.* An *integral whole* [Husserl: whole in the pregnant sense (1913-22.475)] is the mereological fusion of a foundational system. An integral whole forms the *nucleus* or individual nature of a substance. Its *accidents* are a nimbus of tropes dependent (founded) on the nucleus, generically though not individually required by it. Thus Simons envisions a tight bundle within a loose bundle, the whole constituting a thick particular. The tight bundle (the nucleus) is like a substratum, but is not assumed as basic.

Objections to the Bundle Theory

The assault on the trope-bundle theory has been led by Mertz. His objections appear to stem from two deeply held intuitions, which I will call the *predication intuition* and the *glue intuition.* According to the former, it is unacceptable to conceive of tropes as free-floating (Mertz 1996.26). They are not genuine property instances unless they are saturable, properties of something. Compresence classes do not possess enough unity to be genuine subjects of predication. At the same time, as we have seen, Mertz hesitates to posit primitive individuals lest they turn out to be bare particulars, which would be incoherent by his lights. Hence his hierarchies of integrated networks of tropes (see previous section).

According to the glue intuition, complexes need to be held together, and relations are the glue. They are "ontoglial", Mertz says, i.e., from the Greek, the glue of being (1996: 25). Sets and bundles as such lack unity. Thus Mertz is obliged to reject the bundle theory of relations as well as that of individuals. Only

genuine relations can be ontoglial. Together with the predication intuition, this yields Mertz's distinctive dualism about relations, his trope universalism or moderate realism. The basic universals do the gluing, but the basic tropes get predicated. What is the connection between the two? They are both aspects of the trope, the relation instance. The instance aspect is the fundamental ontic unit; the repeatable aspect is conceptual. It might seem that this makes the glue unreal, but Mertz speaks also of extra-conceptual intensions (universals) as goals of total science (1996: 32).

D. H. Mellor, citing Ramsey (1931), and Thomas Hofweber have objected that the above tropist account of predication in terms of overlapping makes exemplification symmetric: it fails to explain which is the subject and which is the the predicate, or which is the individual and which is the universal. So long as compresence classes can be distinguished from exact-resemblance classes (particulars from universals), there is no problem. But what if the same class could be both a particular (or a link in its bundle chain) and a universal? Bacon rules out this possibility, but seemingly *ad hoc*. Might it not be, for example, on a radically monotheistic scheme, that the trope God's divinity was the sole trope in the individual God as well as the sole trope in the property of divinity?

Applications of Trope Theory

Various applications have been proposed for trope theory. Campbell suggests that tropes are the natural relata of causation (1981: 480f). Although events are often cast in that role, Williams affirms that they are a kind of trope (1953: 90). It remains to see whether this insight will shed any real light on the nature of causation. [Bacon sketches a treatment of causation in trope theory, but it is not clear that he makes any essential use of tropes, other than to form possible worlds (1995, ch. VIII).] Campbell further suggests that tropes are the natural subjects of evaluation (1981: 481). Again, while this seems feasible, it is not clear where it takes us. [Bacon tries to develop this idea too (1995, ch. IX), but his treatment would seem to work equally well with states of affairs rather than tropes.] Campbell suggests a trope-theoretic interpretation of the fields recognized by modern physics, but a lot is expected of his field-tropes. Why not have just *one* trope, the-world's-being-the-way-it-is?

Mertz puts forward a distinctive system of logic, particularized predicate logic (PPL), exploiting the opportunity of quantifying over tropes in many places where we should expect second-order quantification over properties (1996, ch. IX). Impressive claims are made for PPL. It is said to be a provably consistent type-free extension of second-order logic, admitting impredicative definitions. Diagonal arguments and Gödel's incompleteness proofs are allegedly defeated, and solutions are proffered to Russell's paradox, the various liar paradoxes, and the generalized Fitch-Curry paradox.

While tropism, like any other theory, must stand or fall on its merits, it may be asking too much to expect metaphysical arguments to establish its pre-eminence. The substance-attribute view, the property-bundle theory, the trope-bundle theory, and even perhaps model-theoretic particularism are apparently all capable of modeling each other (Bacon 1988). If tropes deserve first place in first philosophy, it may be for epistemological or even pragmatic reasons. As we knock about the world, it is tropes we encounter in the first instance. An intelligible theory can start there.

Bibliography

- ARMSTRONG, D. M. 1989. *Universals: an Opinionated Introduction*, Boulder: Westview Press
- BACON, John 1988. "Four Modal Modelings", *Journal of Philosophical Logic* **17**: 91-114
- —— 1995. *Universals and Property Instances: the Alphabet of Being*, Oxford: Blackwell
- BLANSHARD, Brand 1939. *The Nature of Thought*, vol. I (London: Allen & Unwin)
- CAMPBELL, Keith 1981. "The Metaphysic of Abstract Particulars", *Midwest Studies in Philosophy* **6**: 477-488
- —— 1990. *Abstract Particulars*, Oxford: Blackwell
- COOK WILSON, John 1926. *Statement and Inference, with Other Philosophical Papers*, Oxford: Clarendon
- HOCHBERG, Herbert 1988. "A Refutation Of Moderate Nominalism", *Australasian Journal of Philosophy* **66**: 188-207
- HUSSERL, Edmund 1913-21. *Logische Untersuchungen*, 2nd ed., Halle: Niemeyer; 3rd ed. Ursula Panzer, The Hague: Nijhoff, 1984; 2nd ed. tr. J. N. Findlay, *Logical Investigations*, London: Routledge & Kegan Paul, 1970. [Engl. ed. cited.]
- LEIBNIZ, Gottfried Wilhelm von 1686. *Discours de la métaphysique*, tr. *Discourse on Metaphysics [abridged]*, in *Philosophical Writings*, ed. G. H. R. Parkinson, London/Toronto: Dent,1934 (Everyman's Library), 18-47
- —— 1714. *Monadologie*, tr. *Monadology*, *ibid.*, 179-194
- LOCKE, John 1690. *An Essay Concerning Humane Understanding*, ed. Peter H. Nidditch, Oxford: Clarendon 1975. [Modern ed. cited.]
- MARTIN, C. B. 1980. "Substance Substantiated", *Australasian Journal of Philosophy* **58**: 3-10
- MERTZ, D. W. 1996. *Moderate Realism and its Logic*, New Haven: Yale
- PLATO -380. *Republic* in *Works*, tr. Benjamin Jowett, London/Oxford: Macmillan, 1892/1920
- RAMSEY, F. P. 1931. "Universals", *The Foundations of Mathematics and Other Logical Essays*, ed. R. B. Braithwaite, New York & London: Routledge & Kegan Paul, 270-286; *Philosophical Papers*, ed. D. H. Mellor, Cambridge: University Press, 1990: 8-30
- RUSSELL, Bertrand 1911. "On the Relations of Universals and Particulars", *Logic and Knowledge: Essays 1901-1950*, ed. Robert C. Marsh, London: Allen & Unwin, 1956: 105-124
- —— 1940. *An Inquiry into Meaning and Truth*, London: Allen & Unwin
- SIMONS, Peter 1994. "Particulars in Particular Clothing: Three Trope Theories of Substance", *Philosophy and Phenomenological Research* **54**: 553-575
- STOUT, G. F. 1921. "The Nature of Universals and Propositions", *The Problem of Universals*, ed. Charles Landesman, New York: Basic Books, 1971: 154-166.
- —— 1923 "Are the Characteristics of Particular Things Universal or Particular?", *ibid.* 178-183.
- WILLIAMS, D. C. 1953. "The Elements of Being", *Review of Metaphysics* **7**: 3-18, 171-192; *Principles of Empirical Realism*, Springfield: Charles C Thomas, 1966: 74-109. [Reprint cited.]
- —— 1954. "Of Essence and Existence and Santayana", *Journal of Philosophy* **51**: 31-42

Other Internet Resources

- The Metaphysics Research Lab Web Pages

Related Entries

individual | Meinong, Alexius | nominalism: in metaphysics | ontology | Plato | Platonism: in metaphysics | predication and instantiation | properties | realism | Russell's paradox | situation | state of affairs | substance

John Bacon
john.bacon@philosophy.usyd.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 19, 1997
Content last modified: February 20, 1997

Properties

Questions about the nature and existence of properties are nearly as old as philosophy itself. Interest in properties has ebbed and flowed over the centuries, but they are now undergoing a resurgence. The last twenty five years have seen a great deal of interesting work on properties, and this entry will focus primarily on that work (thus taking up where Loux's (1972) earlier review of the literature leaves off).

When we turn to the recent literature on properties we find a confusing array of terminology, incompatible standards for evaluating theories of properties, and philosophers talking past one another. It will be easier to follow this literature if we begin by focusing on the *point* of introducing properties in the first place. Philosophers who argue that properties exist almost always do so because they think properties are needed to solve certain philosophical problems, and their views about the *nature* of properties are strongly influenced by the problems they think properties are needed to solve. So the discussion here will be organized around the tasks properties have been introduced to perform and the ways in which these tasks influence accounts of the nature of properties.

In §1 I introduce some distinctions and terminology that will be useful in subsequent discussion. The tasks properties are called on to perform are typically *explanatory*, and so §2 contains a brief discussion of explanation in ontology. §3 contains a discussion of traditional attempts to use properties to explain phenomena in metaphysics, epistemology, philosophy of language, and ethics. §4 focuses on the three areas where contemporary philosophers have offered the most detailed accounts based on properties: philosophy of mathematics, the semantics of natural languages, and topics in a more nebulous area that might be called *naturalistic ontology*. We then turn to issues about the nature of properties, including their existence conditions (§5), their identity conditions (§6), and the various sorts of properties there might be (§7). §8 provides an introductory, informal discussion of formal theories of properties. After §2 the sections, and in many cases the subsections, are relatively modular, and readers can use the detailed tables of contents and hyperlinks to locate those topics that interest them most.

---

1 Distinctions and Terminology

1.1 Properties: Basic Ideas

Properties include the attributes or qualities or features or characteristics of things. Issues in ontology are so vexed that even those philosophers who agree that properties exist often disagree about *which* properties there are. This means that there are no wholly uncontroversial examples of properties, but likely candidates include the color and rest mass of the apple on my desk, as well (more controversially) as the properties of *being an apple* and *being a desk*. For generality we will also take properties to include relations like *being taller than* and *lying between*.

Universals and Particulars. A fundamental question about properties--second only in importance to the question whether there are any--is whether they are universals or particulars. To say that properties are universals is to say that the selfsame property can be instantiated by numerically distinct things. On this view it is possible for two different apples to exemplify exactly the same color, a single universal. The competing view is that properties are just as much individuals or particulars as the things that have them. No matter how similar the colors of the two apples, their colors are numerically distinct properties, the redness of the first apple and the redness of the second. Such individualized properties are variously known as ‘perfect particulars’, ‘abstract particulars’, ‘quality instances’, ‘moments’, and ‘tropes’. Tropes have various attractions and liabilities, but since they are the topic of another entry, we will construe properties (save for any, perhaps those like *being identical with Socrates*, that could only be exemplified by one thing) as universals.

Properties and Relations. Properties are sometimes distinguished from relations. For example, a specific shade of red or a rest mass of 3 kilograms is a property, while *being smaller than* or *having a weight of 29.4 newtons* are typically regarded as relations (both of which relate my laptop computer to the Earth). Relations generate a few special problems of their own, but for the most part properties and relations raise the same philosophical issues and, except where otherwise noted, I will use ‘property’ as a generic

term to cover both monadic (one-place, nonrelational) properties and (polyadic, multi-place) relations.

Properties can be Instantiated. Properties are most naturally contrasted with particulars, i.e., with individual things. The fundamental difference between properties and individuals is that properties can be *instantiated* or *exemplified*, whereas individuals cannot. Furthermore, at least many properties are *general*; they can be instantiated by more than one thing.

The things that exemplify a property are called *instances* of it (the instances of a relation are the things, taken in the relevant order, that stand in that relation). It is a matter of controversy whether properties can exist without actually being exemplified and whether some properties can be exemplified by other properties (in the way, perhaps, that *redness* exemplifies the property of *being a color*). Some philosophers even hold that there are unexemplifiable properties, e.g., *being red and not red*, but even they typically believe that such properties are intimately related to other properties (here *being red* and *not being red*) that can be exemplified.

Realism, Nominalism, and Conceptualism. The deepest question about properties is whether there are any. Textbooks feature a triumvirate of answers: *realism*, *nominalism*, and *conceptualism*. There are many species of each view, but the rough distinctions come to this. Realists hold that there are universal properties. Nominalists deny this (though some hold that there are tropes). And conceptualists urge that words (like ‘honesty’) which might seem to refer to properties really refer to concepts. A few contemporary philosophers have defended conceptualism (cf. Cocchiarella, 1986, ch. 3), and recent empirical work on concepts bears on it. It is not a common view nowadays, however, and I will focus on realism here.

The Revival of Properties. Just a few decades ago many philosophers concurred with Quine's dismissal of properties as "creatures of darkness," but philosophers now widely invoke them without guilt or shame. For example, most current discussions of mental causation are couched in terms of the causal efficacy of mental properties, while discussions of supervenience often proceed by way of a claim that one family of properties (e.g., mental properties) is supervenient on some other family of properties (e.g., physical properties). But the resurgence of interest in properties has left us with widely varying accounts of their nature, and questions about their existence have by no means disappeared.

Properties are as Properties Do. It is possible to classify theories of properties in terms of their characterizations of the *nature* of properties *or* in terms of the *jobs* they introduce properties to do. The former kind of characterization is more fundamental, but since views about the nature of properties are typically motivated by accounts of the work properties are invoked to do, it will be more useful to begin with the latter. We will ask what *explanatory roles* properties have been introduced to fill, and we will then try to determine what something would *have* to be *like* in order to occupy those roles. This will also allow us to consider the sorts of arguments that are typically advanced for the claim that properties exist.

1.2 Talking about Properties

Philosophers do not have a settled idiom for talking about properties. Often they make do with a simple distinction between singular terms and predicates. Singular terms are words and phrases (like proper names and definite descriptions) that can occupy subject positions in sentences and that purport to denote or refer to a single thing. Examples include ‘Bill Clinton,’ ‘Chicago’, and ‘The first female Supreme Court Justice’. Predicates, by contrast, can be true of things. When we represent a sentence like ‘Quine is a philosopher’ in a standard formal language (like first-order logic) as ‘*Pq*’, we absorb the entire expression ‘is a philosopher’ into the predicate ‘*P*’ (though for some theoretical purposes it is more useful to count ‘philosopher’ or even ‘a philosopher’ as the predicate). The notion of a predicate is supplanted by the notion of a verb phrase in modern grammars, so we don't need to pursue this issue here, but we can raise our first question about property talk at this relatively atheoretical level.

Failed Substitutions

It is perfectly grammatical to say ‘Monica is honest’ or ‘Honesty is a virtue’, but your old English teacher will cringe if you say ‘Honest is a virtue’ or ‘Monica is honesty’. We must use ‘honest’ after the ‘Monica is’, and we have to use the nominalization, ‘honesty’, before ‘is a virtue’. The fact that ‘honest’ and ‘honesty’ cannot be interchanged without destroying the grammaticality of our original sentences has been thought to have various philosophical morals. Some philosophers take it to show that the two expressions cannot stand for the same thing; for example, ‘honest’ might stand for a property and ‘honesty’ might stand for a property-correlate of some sort (Frege draws roughly this moral from his discussion of ‘the concept horse’). Others take it to show that although both expressions are related to the same thing, the property *honesty*, they are related to by different semantic relations; for example, the nominalization denotes this property, whereas the predicate expresses it.

Frege's argument for the first sort of view is not compelling (see Parsons, 1986, for a good discussion); moreover, it would be desirable to avoid multiplying entities (e.g., property correlates) and semantic relations (e.g., expression) beyond necessity. And mere failures of substitutivity are not enough to show that they are necessary, since various syntactic features of sentences prohibit the exchange of terms that are clearly co-referential. Consider case forms of personal pronouns: ‘I’ and ‘me’ cannot be exchanged (without destroying grammaticality) in sentences like ‘*I* dropped the hammer, and he returned it to *me*’. But no one concludes that distinct objects (me and a me-correlate) or distinct semantic relations (nominative-case reference and accusative-case reference) are needed to account for this.

Predicative Expressions

The multiplicity of ways of talking about properties can be obscured when we use familiar formal languages to represent them. The constructions verb (‘lives’), verb + adverb (‘sings badly’), copula + adjective (‘is red’), copula + determiner + common noun (‘is a dog’), copula + noun phrase (‘is a Republican President), and (if Davidson's account of events is correct) even adverbs (‘slowly’) and prepositional phrases (‘in the bathroom’) all go over into the familiar ‘*F*’s and ‘*G*’s of standard logical notation. The fact that these expressions can often be handled in the same way without too much violence tells us that they have certain similarities, but there are also many differences, and some of them may turn out to be relevant to ontology.

Singular Terms

The complexities involving property words are even greater when we turn to singular terms. We can form singular terms from predicative expressions in many ways (different ways are appropriate for different predicates). To begin with, English contains a plethora of *suffixes* that we can append to predicative expressions (sometimes after minor surgery on the original) to form singular terms. These include '-hood' ('motherhood, 'falsehood), '-ness' ('drunkeness', 'betweeness'), '-ity' ('triangularity', 'solubility', 'stupidity'), '-kind' ('mankind'), '-ship' ('friendship', 'brinksmanship'), 'ing' ('walking', 'loving'), 'ment' ('commitment', 'judgment'), 'cy' ('decency', 'leniency'), and more.

Various philosophical terms of art serve a similar purpose. The word 'itself' plays this role in some translations of Plato ('The equal itself', 'Justice itself'), and contemporary authors use phrases like 'the property red', 'the property of being red', and 'the causal relation' to much the same end. Various gerundive phrases (e.g., 'being red' and 'being a red thing') and infinitive phrases ('to be happy', 'to be someone who is happy') work in a similar way. Finally, there are many less systematic ways of talking about properties; for example, we can use a definite description that a property just happens to satisfy ('the color of my true love's hair', 'John's favorite four-place relation').

The expressions formed in these ways occupy subject positions in sentences where they seem to denote to properties. It is worth noting, however, that it is often impossible to substitute some of these expressions for related ones without destroying the grammaticality or, in some cases, without altering the truth value of the original sentence. Consider 'wisdom', 'being wise', 'the property of being wise', and 'to be wise'. 'Wisdom is a virtue' is unexceptionable, but 'Being wise is a virtue' is shaky at best. On the other hand, 'To be wise is to be virtuous' and 'Being wise is a good thing' are fine, but 'Wisdom is to be virtuous' clearly won't do. And 'The property of being wise is a good thing' is grammatical, but has a different meaning from 'Being wise is a good thing'.

The phenomenon of case shows that lack of substitutivity alone doesn't have deep ontological consequences, but it is quite possible that the sorts of phenomena noted in the previous paragraph signal important differences in ontology. Some of these differences might begin to emerge from informal probing, but we cannot expect to settle such matters without detailed, philosophically-sensitive syntactic and semantic theories that are better supported than their rivals. Such theories do not yet exist, and so here I will be fairly cavalier about "property terms," using various phrases, e.g., 'redness' and 'the property of being red' indifferently to refer to the same property. But this expedient is not meant to suggest that subtle grammatical differences won't eventually turn out to have important ontological implications.

2 Philosophical Explanations: Why Think that Properties Exist?

2.1 Explanation in Ontology

Properties are typically introduced to help *explain* or *account for* phenomena of philosophical interest. The existence of properties, we are told, would explain qualitative recurrence or help account for our ability to agree about the instances of general terms like 'red'. In the terminologies of bygone eras, properties save the phenomena; they afford a *fundamentum in re* for things like the applicability of general terms. Nowadays philosophers make a similar point when they argue that some phenomenon holds *because of* or *in virtue of* this or that property, that a property is its *foundation* or *ground* for it, or that a property is the *truth maker* for a sentence about it. These expressions signify explanations.

When properties are introduced to help explain certain philosophically puzzling phenomena, we have a principled way to learn what properties are *like*: since they are invoked to play certain *explanatory roles*, we can ask what they would *have* to be like in order to play the roles they are introduced to fill. What, for example, would their existence or identity conditions need to be for them to explain the (putative) modal features of natural laws or the *a priori* status of mathematical truths?

The Limits of Explanation

Perhaps the deepest question in ontology is when (if ever) it is legitimate to postulate the existence of entities (like possible worlds, facts, or properties) that are not evident in experience. Some philosophers insist that it never is. Others urge that at least some entities of this sort, in particular properties, have no explanatory power and that appeals to them are vacuous or otherwise illegitimate (e.g., Quine, 1961, p. 10; Quinton 1973, p. 295).

The more heavy-handed dismissals of properties and other metaphysical creatures have often been based on faulty accounts of concept formation (which led Hume to counsel consignment of metaphysical works to the flames) or defective theories of meaning (which led many positivists to view metaphysics as a series of pseudo explanations offered to solve pseudo problems). Wittgenstein takes a more subtle approach, trying to show us that 'our disease is one of wanting explanations' (1991, Pt VI, 31) and striving to cure us of it. Swoyer (1999) has attempted some defense of explanation by postulation in ontology, but the issues are difficult ones that are not amenable to proof or disproof. Fortunately the present task is not to defend explanation in ontology, but it will be useful to briefly note two general views about such explanations.

Two Views of Explanation in Ontology

Metaphysics has traditionally been viewed as first philosophy, and some philosophers hold that its arguments should be demonstrative. Recently Linsky & Zalta (1995) have argued that it is possible to give a transcendental argument for the existence of properties; if this argument is successful, it is demonstrative, and they claim that its conclusion (that a wide range of properties exist) is synthetic *a priori*. Others (e.g., Swoyer, 1983; 1999) urge that most of the arguments advanced on behalf of properties appear anemic when judged by the demonstrative ideal, but that they look much better when

viewed as inferences to the best explanations. We will not pursue this issue, however, since it is impossible to form a satisfactory view about the nature of philosophical explanations in a vacuum. An account of metaphysical explanation should instead emerge from a consideration of the more plausible metaphysical explanations, and we will focus on such explanations here.

2.2 Constraints on Explanations Employing Properties

Parochial Constraints

Philosophical explanations are usually thought to be constrained in various ways, but beyond philosophical family values like consistency, parsimony and comprehensiveness these constraints will often seem parochial to those philosophers who are not committed to them. In Medieval disputations about universals, for example, religion and theology were fundamental, and it was widely held that any account of properties should be able to explain the Trinity, the Eucharist, and the absolutely unchanging nature of God (this last requirement often led to quite tortured accounts of the relations holding between protean finite beings and God). But few philosophers in our naturalistic era would give such considerations a second thought.

More General Constraints

Some proposed constraints on metaphysical explanation depend on more general philosophical orientations. For example, Russell's Principle of Acquaintance, the injunction that we only admit items into our ontology if we are directly acquainted with them, expresses an strong empiricist sentiment. Other constraints are more directly metaphysical. For example, Aristotle upbraids Plato for separating the Forms from their instances, suggesting that this renders them incapable of explaining anything (e.g., *Metaphysics*,1079b11--1080a10). His point seems to be that properties could explain things about individuals only if they were located *in* those individuals. The sentiment is that an individual, spatio-temporal object (like my cat) which stands in some obscure relation to some entity entirely outside of space and time (say the Form of the cat) cannot explain anything about the cat itself.

Mandatory Constraints

All accounts of properties must avoid various perennial objections to them. Three criticisms of this sort were anticipated by Plato (worrying about his own doctrines) in the *Parmenides*.

First, it appears that a universal property can be in two completely different places (i.e., in two different instances) at the same time, but ordinary things can never be separated from themselves in this way. There are scattered individuals (like the former British Empire), but they have different spatial parts in different places. Properties, by contrast, do not seem to have spatial parts; indeed, they are sometimes said to be wholly-present in each of their instances. But how could a single thing be wholly present in widely separated locations?

This conundrum has worried some philosophers so much that they have opted for an ontology of tropes in order to avoid it, but realists have two lines of reply (both of which commit us to fairly definite views about the nature of properties). One response is that properties are not located in their instances (or anywhere else), so they are never located in two places at once. The other response is that this objection wrongly judges properties by standards that are only appropriate for individuals. Properties are a very different sort of entity, and they *can* exist in more than one place at the same time without needing spatial parts to do so.

Second, some properties seem to exemplify themselves. For example, if properties are abstract objects, then the property of being abstract should itself exemplify the property of being abstract. In various passages throughout his dialogues Plato appears to hold that Forms (which are often taken to be his version of properties) participate in themselves. Indeed, this claim serves as a premise in what is known as his *Third-Man Argument* which, he seems to think, may show that the very notion of a Form is incoherent (*Parmenides*, 132ff). Russell's paradox raises more serious worries about self-exemplification. It shows that any account which allows properties to exemplify themselves must be carefully formulated if it is to avoid paradox (a polite word for inconsistency).

Third, many critics have charged that properties generate vicious regresses, e.g., the one exhibited in Plato's third man argument or Bradley's regress, and any viable account of properties must have the resources to avoid them.

The disputes about plausible constraints on property-invoking explanations, together with the obvious difficulty of settling such disputes, leave the situation murkier than we would wish. We will see that the use of properties to explain phenomena in the philosophy of mathematics or naturalistic ontology or the semantics of natural languages imposes additional, tighter, constraints that make it easier to evaluate competing accounts. But constraints of the sort noted here have played a central role in many philosophical discussions of properties, and we will often fail to understand those discussions if we forget this.

2.3 The Fundamental Ontological Tradeoff

Metaphysics, like life, is full of tradeoffs, cost-benefit analyses, the attempt to simultaneously satisfy competing constraints. In ontology we must frequently weigh tradeoffs between various desiderata, e.g., between simplicity and comprehensiveness, and even between different kinds of simplicity. But one tradeoff is so pervasive that it deserves a name, and I will call it *the fundamental ontological tradeoff*. The fundamental ontological tradeoff reflects the perennial tension between explanatory power and epistemic risk, between a rich, lavish ontology that promises to explain a great deal and a more modest ontology that promises epistemological security. The more machinery we postulate, the more we might hope to explain--but the harder it is to believe in the existence of all the machinery.

The dialectic between a realism with chutzpah and a diffident empiricism runs all through philosophy, from ethics to philosophy of science to philosophy of mathematics to metaphysics. Excessive versions of

each view are usually unappealing. Extreme realists ask us to believe in things many philosophers find it difficult to believe in; extreme empiricists wind up unable to explain much of anything. But the dialectic between power and risk remains even when we move in from the extremes. It often manifests itself in a yearning for parsimony, a desire for as few entities as we can scrimp by with. Such longings may seem prudish or stuffy or a bit too metaphysically correct. Often the desire is not to achieve parsimony for its own sake, however, but to find an ontology that is modest enough to provide a measure of epistemological security. Choices needn't be all or none, and a principled middle ground is always worth striving for. But no matter where a philosopher tries to stake her claim, the fundamental ontological tradeoff can rarely be avoided and we will encounter it frequently in what follows.

3 Traditional Explanations: An Unscientific Survey

Properties have been invoked to explain a very wide range of phenomena. The things to be explained (*explananda*; singular *explanandum*) are a mixed bag, and the explanations vary greatly in plausibility. To fix ideas, we will note several of the most common explanations philosophers have asked properties to provide (for a longer list see Swoyer, 1999, §3).

Resemblance and Recurrence

There are objective similarities or groupings in the world. Some things are alike in certain ways. They have the same color or shape or size; they are protons or lemons or central processing units. A puzzle, sometimes called the problem of the *One over the Many*, asks for an account of this. Possession of a common property (e.g., a given shade of yellow) or a common constellation of properties (e.g., those essential to lemons) has often been cited to explain such resemblance. Similarly, different groups of things, e.g., Bill and Hillary, George and Barbara, can be related in similar ways, and the postulation of a relation (here *being married to*) that each pair jointly instantiates is often cited to explain this similarity. Finally, having different properties, e.g., different colors, is often said to explain qualitative differences. A desire to explain qualitative similarity and qualitative difference has been a traditional motivation for realism with respect to universals, and it continues to motivate many realists today (e.g., Armstrong, 1984, p 250; Butchvarov, 1966; Aaron, 1967, ch. 9).

Recognition of New and Novel Instances

Many organisms easily recognize and classify newly encountered objects as yellow or round or lemons or rocks, they can recognize that one new thing is larger than a second, and so on. Some philosophers have urged that this ability is based partly on the fact that the novel instances have a property that the organism has encountered before--the old and new cases share a common property--and that the creature is somehow attuned to recognize *it*.

Meaning

Our ability to use general terms (like ‘yellow, ‘lemon’, ‘heavier than’, ‘between’) provides a linguistic counterpart to the epistemological phenomenon of recognition and to the metaphysical problem of the One over the Many. Most general terms apply to some things but not to others, and in many cases competent speakers have little trouble knowing when they apply and when they do not. Philosophers have often argued that possession of a common property (like *redness*), together with certain linguistic conventions, explains why general terms apply to the things that they do. For example Plato noted that ‘we are in the habit of postulating one unique Form for each plurality of objects to which we apply a common name’ (*Republic*, 596A; see also *Phaedo* 78e, *Timaeus*, 52a, *Parmenides*, 13; Russell, *Problems of Philosophy*, p. 93). Questions about the meanings (now often known as the ‘semantic values’) of singular terms like ‘honesty’ and ‘hunger’ and ‘being in love’ may be even more pressing, since the chief task of such terms seems to be to *refer* to things. But what could a word like ‘honesty’ refer to? If there are properties, it could refer to the property *honesty*.

Unification and Triangulation

In a brilliant paper on Plato's theory of Forms (which, as noted above, are often taken to be his version of properties), the classicist H. F. Cherniss (1936) argues that Plato intended his theory to solve three fundamental philosophical problems. By the end of the fifth century B.C. the arguments and conundrums of philosophers had cast doubt on several things that Plato thought were obviously true. In ethics Protagorean relativism threatened the view that ethical principles could be objective; in the clamor of individual disagreements, clashes between cultures, and the failure of philosophical inquiry to locate any firm ground, the challenge was to *explain* how ethical objectivity was possible. When Plato turned to epistemology, various considerations convinced him that there was an important difference between knowledge (*episteme*) and belief (*doxa*), even between knowledge and *true* belief (right opinion). But how could we explain that? Finally, in metaphysics it seemed clear that things change in various ways, but the arguments of Parmenides made even this seem mysterious.

Plato drew on his Forms to explain how these three phenomena were possible. On his view, the Forms exist pure and unadulterated by human thought, and some Forms, most prominently the Good, offer objective standards for values like goodness and justice. In epistemology Plato attempted to explain the difference between knowledge and belief by arguing that Forms are the objects of the former but not the latter (e.g., *Timaeus*, 51d3ff). In metaphysics Plato argued that change is only possible against a background of things that do not change, and he urged that the Forms provided this (*Cratylus*, 439d3ff). Finally, although Cherniss doesn't mention it, Plato's theory of Forms helped explain the semantics of general terms (as suggested in *Republic*, 596A).

This isn't to say that all, or indeed any, of Plato's explanations were successful. But it is worth noting that many philosophers still invoke properties to account for the sorts of things Plato struggled to explain. Early in this century G. E. Moore offered an alternative to ethical naturalism by claiming that goodness is a simple, non-natural property. Few contemporary philosophers would accept Moore's anti-naturalism or his account of non-natural properties, but many would defend ethical naturalism by arguing that moral properties supervene on naturalistically respectable properties.

Virtually no philosophers accept Plato's account of the difference between knowledge and belief, but many still hold that properties have an important role to play in explaining epistemological phenomena. For example, Russell (1912, ch. 10) argued that the only way to explain the possibility of *a priori* knowledge is to regard it as knowledge of relations among universals . Most philosophers today would question this, but many of them would agree that properties have an important role to play in explaining such epistemological phenomena as our ability to recognize and categorize things in the world around us.

Few contemporary philosophers would endorse Plato's claims about the need for some permanent backdrop for flux, but properties can still be cited to explain change. If my pet chameleon was brown all over yesterday and is green all over today, then the brute existence of the creature isn't enough to explain the change; after all, *he* persisted throughout. But, some philosophers urge, we can explain the alteration by noting that the chameleon exemplified the property *brownness* yesterday but he exemplifies the property *greenness* today.

Finally, many philosophers would concur that Plato's account of the meanings of general terms was on the right track, though as we shall see in §4.2, current accounts of meaning have moved far beyond Plato's in their detail and formal sophistication.

Explanation by Unification

This brief survey of putative explanations that rely on properties isn't meant to be detailed or exhaustive; the point is simply to illustrate how a range of accounts employ properties in an effort to explain philosophically puzzling phenomena. Just as importantly, Plato's account suggests an attractive model for philosophical explanation. A general pattern of explanation by *unification*, *integration* or *systematization* is at work in his attempt to solve three, superficially disparate, problems using the same resources. He attempts to show that at a fundamental level the three phenomena are related, linked by the Forms and the principles than govern them. This unification has explanatory value, since it allows us to see a single pattern or entity at work in a range of superficially diverse cases. At all events, this is one explanatory virtue in the natural sciences, clearly at work in the work of Newton and Maxwell and Darwin, and it is also a pattern we find in Plato's account.

An account that employs properties to do multiple tasks has two further virtues. First, insofar as each of the explanations is plausible, it serves as part of a *cumulative case* for the existence of properties. Second, if properties can perform multiple tasks, they must simultaneously satisfy multiple constraints, and so different sorts of data can be used to test a theory of properties. The hope is that by considering several tasks of this sort we could begin to triangulate in on the nature of properties; we could begin to see what features properties would need to have in order to play each of the different explanatory roles. It may turn out, of course, that entities well-suited to one explanatory role will be ill-suited to another. For example, we will see below that the existence and identity conditions of entities used to account for causation may be rather different from those needed by entities that could serve as the meanings of intentional idioms (like ‘is thinking of Vienna’). This might lead us to postulate the existence of several kinds of properties; alternatively, it might lead us to conclude that properties cannot do all of the things philosophers has hoped that they could. Either way, as fragmentation increases, cumulative support and triangulation on

the nature of properties will slip away.

4 What have you done for us lately? Recent Explanations

Properties alone cannot explain much of anything. A *theory* of properties--an account that tells us what properties are *like* and *how* they do what they are invoked to do--is required for that. A number of theories of properties have been developed over the last quarter century, and many of them possess much more depth, sophistication, and formal detail than the no-frills accounts alluded to in the previous section. I will focus on explanations in three areas where properties are often invoked today: philosophy of mathematics, semantics (the theory of meaning), and naturalistic ontology. These areas are also useful to consider, because if properties can explain things of interest to philosophers who don't specialize in metaphysics, things like mathematical truth or the nature of natural laws, then properties will seem more interesting. Unlike the substantial forms derided by early modern philosophers as dormitive virtues, properties will pay their way by doing interesting and important work.

My aim is to indicate the general lay of the land and point the way to more detailed discussions that interested readers can follow up. In each of the three cases I will indicate:

1. **What is to be explained**. As with most things in philosophy, there is often some controversy over which things in a given area stand in need of philosophical explanation. In some cases a few philosophers question the very existence of the things that other philosophers think require explanation; for example, able philosophers have denied that there are such things as mathematical truth (e.g., Field, 1980) or laws of nature (e.g., van Fraassen, 1989). And even those philosophers who think that we need to explain certain things, e.g., various features of mathematical truth, may disagree about precisely what those features are. In the three areas examined in this section, however, there is a reasonable degree of consensus about which things stand in need of explanation, and I will focus on these.
2. **How properties explain**. In some cases different philosophers use properties in different ways to explain the same phenomenon. I will focus on the simpler, more common approaches. We will also see that in most cases a theory of properties only explains things when it is conjoined with various background assumptions or auxiliary hypotheses.
3. **Beating the competition**. Arguments that properties exist because they explain some particular phenomenon (like qualitative recurrence or mathematical truth) are weak if other sorts of entities can account for it just as well. Arguments that alternative accounts don't work, especially when they involve alternative putative entities (like sets or tropes), are typically based on the claim that these entities lack the requisite features to account for the explanandum. I will also note a few cases where proponents of one account of properties argue against proponents of a rival account, since these arguments typically involve disputes over the nature of properties.
4. **Difficulties**. Almost all explanations that employ properties face difficulties, and I will briefly indicate the most serious of these.

5. **Lessons the explanations teach us about properties**. Properties often must have certain features in order to provide certain explanations. So once we have examined a given explanation, we will ask what properties would have to be *like* in order to provide it. In particular, we will ask what lessons are to be learned about the existence and identity conditions of properties, their structure (if any), and their modal and epistemic status.

4.1 Mathematics

Philosophers of mathematics have focused much (arguably too much) of their attention on number theory (arithmetic). Number theory is just the theory of the natural numbers, 0, 1, 2, ..., and the familiar operations (like addition and multiplication) on them. Many sentences of arithmetic, e.g., ‘7 + 5 = 12’ certainly seem to be true, but such truths present various philosophical puzzles and philosophers have tried to explain how they *could* have the features they seem to have.

Explananda in Philosophy of Mathematics

Most wish lists include hopes for explanations of at least five (putative) facts; philosophers want to know:

1. How the sentences of arithmetic can have truth values (how they can be true or false)
2. How the sentences of arithmetic can be objectively true (or false), independently of human language and thought
3. What the logical forms of the sentences of arithmetic are
4. How the sentences of arithmetic can be necessarily true (or necessarily false)
5. How the truth values of sentences of arithmetic can be known independently of experience (*a priori*), save for a modicum of experience needed to acquire mathematical concepts

Sample Explanations

Identificationism

Most attempts to use properties to explain the items on this list are versions of *identificationism*, the reductionist strategy that *identifies* numbers with things that initially seem to be different. This approach is familiar from the original versions of identificationism where numbers were identified with sets, but it is straightforward to adapt this earlier work to identify numbers with properties rather than with sets.

Properties vs. Sets

Sets are often contrasted with properties, and before proceeding it is important to note a fundamental difference between the two. If *x* and *y* are sets and have exactly the same members, then *x* and *y* are one and the same set. When *x* and *y* have precisely the same members they are said to have the same *extension*, and sets are often called *extensional* entities. Just as sets can have members, properties can have instances, things that exemplify or instantiate them, and this relation of exemplification is to

properties what the membership relation is to sets.

The identity conditions of properties are a matter of dispute. Everyone who believes there are properties at all, however, agrees that numerically distinct properties *can* have exactly the same instances without being identical. Even if it turns out that exactly the same things exemplify a given shade of green and circularity, these two properties are still distinct. For this reason properties are often said to be *intensional* entities, although people often concur with this because they agree about what properties' identity conditions are *not* (they aren't extensional), rather than because they agree about what their identity conditions *are*.

The ABCs of identificationism

If we have a rich enough theory of properties, it is possible to retrace the steps of earlier versions of identificationism using properties in place of sets. The property theorist can formulate axioms for property theory that parallel the axioms of standard set theories (save for replacing the axiom of extensionality with some other identity condition, perhaps omitting the axiom of foundations, and making other minor emendations to adapt the ideas better to properties; e.g., Jubien, 1989; cf. Bealer, 1982, Ch. 6; Pollard and Martin, 1986).

There are infinitely many natural numbers (the collection of natural numbers in fact has the smallest size an infinite collection can have), so the first step in identificationist programs is to find (or postulate, or imagine) an infinite realm of properties. The next step is to identify one denizen of this realm with the number zero and to identify some operation on this realm of entities with the successor function. The key here is that successive iterations of the function yield a new and different entity every time it's applied.

There are two major species of identificationism. The first views the reducing theory (of sets, or of properties) as a branch of logic; the second views it as a substantive theory (of sets, or of properties) that makes commitments over and above those made by logic. There are important differences between the two approaches, but given the very strong nature of the logic required for logicist identificationism, the differences do not matter greatly here so I will treat both approaches together. (For a discussion of the differences, see Section 1 ("Logicist Identificationism") of the supplementary document Uses of Properties in the Philosophy of Mathematics.)

Identificationist accounts treat ‘1’ and ‘2’ as singular terms that refer to properties (those properties that are identified as the numbers 1 and 2), and they treat predicates and function symbols as denoting relations and functions. Thus, since the semantics values of ‘1’ and ‘2’ are in the extension of the relation expressed by the predicate ‘<’, the sentence ‘1 < 2’ is true and, indeed, it has the simple logical form of a predication of a two-place predicate, ‘<’, with two singular terms, ‘1’ and ‘2’, i.e., it has the simple logical form *Rxy*. We apply this idea to all atomic sentences in the language of arithmetic and then extend the account to all sentences in this language by the usual recursive treatments of the logical constants.

This explains how sentences of arithmetic can be objectively true (wishes 1 and 2): they are true because

they describe an objective realm of mind-independent properties. And since the language we use has a straightforward referential semantics, it also supplies a very natural and straightforward account of the logical forms of the sentences of number theory (wish 3). Finally, if the properties identified with numbers are ones that exist necessarily, and if they necessarily stand in the arithmetical relations that they do, the truths of arithmetic will be necessarily true (wish 4). But taken alone property-based identificationism does not explain mathematical knowledge (wish 5; we will return to this matter below).

Some recent accounts identify numbers with properties that seem less other-worldly than those invoked by mainstream identificationists. For example, Bigelow and Pargetter (1990) argue that rational numbers are higher-order relations--*ratios*--among certain kinds of first-order relations. The leading idea is that if Bill is twice as tall as Sam, then Bill stands in the relation *twice as tall* to Sam. This relation in turn stands in the (second-order) ratio relation of 2:1 to the identity relation among objects. Such higher-order ratio relations are isomorphic to the rational numbers, and Bigelow and Pargetter go on to *identify* them with the rational numbers. Thus, the second-order relation 5:1 turns out to be the number five. It isn't clear how to extend the ideas to large infinite cardinals or to ordinal numbers, but they propose extending the idea to second-order relations of proportion, and identifying the reals with such proportions.

Other Property-based accounts in the Philosophy of Mathematics

There are also several non-identificationist accounts of mathematical truth that make use of properties.

Structuralism

The most important features, perhaps the only features, of the natural numbers are structural ones. These are the features that axiomatizations capture (zero is the first member of a countably infinite sequence, each member of the sequence has exactly one member that follows it, etc.). Such sequences are said to be omega-sequences. Structuralists (often inspired by Benacerraf, 1965) take this idea to heart and argue that any omega-sequence can play the role of the natural numbers (cf. Resnik, 1995). They claim that it's the *structure* that such sequences have in common, rather than the particular entities that happen to populate them, that are important for mathematics. And one way to develop this idea is to think of an omega-sequence as a very complex, relational property that could be instantiated by actual sequences of objects of the appropriate sort.

Structuralist accounts avoid one of the problems noted below (that of isomorphic models) which besets all versions of identificationism. They may also make the epistemology of mathematics slightly less puzzling, since many structural or pattern-like properties can be instantiated in the things we perceive (we perceive such a property when we recognize a melody played in different keys, for example). But they cannot deliver explanations of the truth conditions and logical forms of arithmetical sentences that are as straightforward as those provided by identificationist accounts since they don't offer us any objects to serve as the referents of the numerals.

Abstract individuals and situations

Linsky and Zalta (1995) develop a novel account of mathematical truth using Zalta's (1983) theory of abstract objects. (The account is developed in much more detail in Zalta (2000) and (1999).) It is relevant here because it is developed along side a formal account of properties that rivals Bealer's in scope and detail. Abstract objects are correlated with collections of properties (which needn't be either maximal or consistent), situations are defined as a special sort of abstract object, and mathematical theories are identified with situations that encode only propositional properties. The account is too detailed to present here, but we will discuss Zalta's basic ideas below when we turn to the identity conditions of properties.

Beating the Competition

The most obvious competitors to property-based accounts of mathematical truth identify numbers with sets, and as long as we focus *solely* on mathematics, sets may seem more appealing. After all, sets do have clearer identity conditions than properties. Moreover, the iterative conception of sets, a picture according to which they form a natural hierarchy, fits nicely with our picture of the structure of natural numbers, whereas an iterative conception of properties is less natural. Finally, set theory provides a powerful unifying framework in which all sorts of mathematical entities, like functions and spaces, can be reconstructed (or at least represented) in a common idiom and dealt with by a common stock of techniques (like proofs by mathematical induction).

The most compelling defense of the use of properties in the philosophy of mathematics urges that when we step back and consider the big picture we see that a rich enough stock of properties can do all the work of sets (and numbers--or that we can use them to define sets or numbers) *and* that properties can do further things that sets simply cannot. For example, it has been argued that properties can be used to give accounts of the semantics of English or explain the nature of natural laws. The appeal of sets, in short, results from a metaphysical myopia, but once we adopt a larger view of things we find that properties provide the best global, overall explanation.

Difficulties

The gravest threats to identificationism are posed by what might be called the *Benacerraf problems*. Authors who defend such accounts are aware of these difficulties and have proposed various responses to them, but the problems are serious and no solutions are generally accepted.

The Problem of Isomorphic Identifications

As Benacerraf (1965) noted, if there is one way to identify the natural numbers with sets, there are countless ways, e.g., Frege's, Zermelo's, von Neumann's, etc. (For a brief discussion of this, see Section 2 ("Set-theoretic Identificationism") of the supplementary document Uses of Properties in the Philosophy of Mathematics.) Some accounts are better for certain purposes than others. But no account is best for all purposes, and if one was, no one has ever explained how it would follow that *it* was the true story about numbers.

There is a similar arbitrariness in any particular identification of numbers with properties (as the fact that different property theorists identify numbers with different properties shows). The point is most obvious with those theories that treat properties as intensional analogues of sets, since it is well-known that numbers can be identified with sets in myriad ways. But it will be a problem for any identificatory program, since there will be many isomorphic models of number theory in the realm of properties (if it is commodious enough to provide any models at all). And there is no reason for thinking that any particular model gives The One True Story about what the numbers actually are.

This difficulty also threatens less formal property-based accounts. For example, there is some arbitrariness in Bigelow and Pargetter's identifications, since we can find many different models of the theory of rational numbers among the realm of ratio relations (e.g., we could identify *n/m* with the relation *n:m* or with the relation *m:n*), and there is no clear reason to suppose that one identification is the right one.

The Problem of Epistemic Access

The second problem, suggested by Benacerraf (1973) a few years later, is that most versions of identificationism propose to identify numbers with putative objects that lie outside the spatio-temporal, causal order. The problem is that we are physical organisms living in a spatio-temporal world who cannot interact causally (or in any other discernible way) with abstract, causally inert things. Few people are aware of having any special cognitive faculty that puts them in touch with a timeless realm of abstract objects, neuroscientists have never found any system in the brain that subserves such a capacity, such a story is not suggested by what is known about the ways in which children acquire numerical concepts, and nothing in physics remotely suggests any way in which a physical system (the brain) can make any sort of contact with causally inert, non-physical objects. None of this proves that we don't have some sort of access to an abstract realm of objects, but the claim that we do leaves the epistemology of mathematics a mystery and, more importantly, there seems to be little positive reason to suppose that it's true.

A few philosophers, e.g., Linsky & Zalta (1995) have taken the problem of epistemic access seriously, and proposed solutions that do not involve mysterious cognitive faculties. Philosophers remain divided on this issue, but it is safe to say that if the problem of epistemic access cannot be overcome, it in turn undermines identificationist attempts to use properties to explain arithmetic truth. If we cannot gain epistemic access to the realm of numbers, then there is no clear way for us to establish connections between the items of our language (e.g., ‘one’) and the numbers they denote. We can't, for example, say that zero is the first number until we manage to attach the word ‘number’ to the realm of numbers. It might seem that we could avoid this difficulty by using purely structural descriptions, ones employing only logical vocabulary, for the task. If such descriptions were couched in a sufficiently powerful language they could be used to characterize the natural numbers up to isomorphism. Such a characterization is all we could ask of any formalization, but it isn't enough to pick out the natural numbers themselves, since if there is one model of a purely structural sentence incorporating such a description, there will be many. For example, such a sentence will have models in the domains of the positive real numbers, the negative real numbers, many fragments of the iterative hierarchy of sets, and so on.

Once again we face the fundamental ontological tradeoff: A richer ontology offers to explain many things that might otherwise be mysterious. But in the view of many philosophers, it engenders epistemological mysteries of its own.

Excursus: Other Reductions

Identificationists sometimes speak of *reducing* numbers to properties. Similarly, one might hope to reduce other things, e.g., possible worlds, to properties (e.g., Zalta, 1983, §4.2; Forrest, 1986). The aim is to show that they such things are nothing over and above very complicated properties.

Bundle Theories

One of the most interesting reductionist programs attempts to reduce individuals or particulars to collections of properties. Such programs are often called *bundle theories*, since they identify ordinary individuals with bundles of properties. Russell (1948, Pt. IV, ch. 8) developed one account of this sort in which individuals were treated as properties linked together by a relation he called *compresence*. The evaluation of such accounts would require an excursus into the ontology of individuals where issues like the problem of individuation, the identity of indiscernibles, and identity through time loom large. Such matters lie outside the scope of our present discussion, though it is worth noting that they involve a purer version of ontology than theories of properties; they have relatively few implications outside of ontology itself.

Lessons About Properties

What do property-based versions of identificationism tell us about the *nature* of properties? We can read off minimum requirements from the fact that in this domain sets can do most of the work that properties are invoked to do.

> **Existence Conditions:** We require an infinite realm containing at least aleph-null (the smallest infinite cardinal number) many properties. Depending on our aspirations, we may need many more. For example, if we want to work with huge transfinite cardinal numbers, we will need a very large infinity of properties.
>
> **Identity Conditions:** Formalized mathematics is one of the few domains where extensionality reigns, and the fact that sets can be used as surrogates of the natural numbers tells us that entities with very coarse-grained identity conditions can do at least most of the work of numbers.
>
> **Structure:** The realm of properties has to include enough relations among properties to give it the structure of an omega-sequence. And if we want to identify others sorts of numbers, e.g., the real, or complex, or transfinite ordinal numbers, with properties, we will

require many additional properties as well as further relations to structure them in the right sorts of ways.

Modal Status: If the truths or arithmetic are necessarily true, then we need a realm of necessarily existing properties that necessarily stand in the (mathematically relevant) relations that they do.

Epistemic Status: If the truths of arithmetic can be known *a priori*, then the arithmetic features of those properties that play the role of numbers must be knowable *a priori*.

4.2 Semantics and Logical Form

Language and logic have long been an important source of data for ontologists. Many philosophers have contented themselves with fairly informal appeals to various features of language to support their claim that properties exist, but in the last two decades some philosophers (along with a few linguists and even computer scientists) have employed properties as parts of detailed accounts of the semantics (meaning) of large fragments of natural languages like English or Choctaw, and some of these accounts contain the most detailed formal theories of properties ever devised. Some property theorists are motivated almost exclusively by a desire to give a semantic account of natural language (e.g., Chierchia and Turner, 1988), others hold that this is but one of several motivations for developing an account of properties (e.g., Bealer, 1982; Zalta, 1993), but it should be noted that still others (e.g., Jubien, 1989; Armstrong, 1997; cf. Mellor, 1986, pp. 180ff) doubt that properties have any serious role to play in semantics at all.

Explananda in Semantics

Logical form

Semantic accounts often go hand in hand with theories of logical form. *Logical form* is a technical notion motivated by the observation that sentences with a similar surface structure may exhibit quite different logical behavior. For example, ‘John is tall and Tom is tall’ entails ‘Tom is tall’, but ‘You show me someone who dislikes John and I'll show you a real misanthrope’ does not entail ‘I'll show you a real misanthrope’. Furthermore, sentences that appear different on the surface may exhibit similar logical behavior. For example, ‘You show me someone who dislikes John and I'll show you a real misanthrope’ and ‘If you show me someone who dislikes John, then I'll show you a real misanthrope’ evince similar logical behavior.

Such facts led various philosophers to introduce a theoretical notion of logical form and to use it to provide theoretical redescriptions of sentences in terms of their logical form in a way that allows us to explain their logical features (e.g., why they are consistent with some sentences but not with others or why they entail the sentences they do). Although philosophers differ in how systematic they are in developing such accounts, most arguments to the effect that properties are needed to explain linguistic phenomena are linked to some conception of logical form.

Sample Explanations

Informal appeals to language and logic

We will begin with four linguistic phenomena that might be explained by a relatively informal and somewhat piecemeal account of properties.

1. General terms like ‘blue’ and ‘honest’ can apply to a variety of things, they apply to the things that they do partly because of their meanings, and in some cases where two predicates in fact apply to exactly the same things, they could have applied to different things.
2. Abstract singular terms like ‘courage’ can occupy subject position in true sentences (‘Courage is a virtue’), they seem to be referring singular terms, and many of sentences of this sort (e.g., ‘Courage is Tom's favorite virtue’) cannot be paraphrased in a way that eliminates the abstract singular term.
3. We can use pronouns (which certainly seem to be referring expressions) that are anaphorically linked back to predicates (‘Clinton is undisciplined, and *that* is a bad quality in a president’) or to terms in subject position like gerunds (‘Being undisciplined is deplorable, and *it* also endangers others’).
4. Many sentences of English appear to quantify over the semantics values of predicates (‘Clinton is tenacious, so there is at least one virtue that he has’) or abstract singular terms (‘Lethargy is a symptom of mononucleosis, so there is at least one symptom of that malady’). And although some of these sentences can perhaps be paraphrased or reconstrued in ways that dispel the appearance of quantification, many have resisted years of such attempts. For example, ‘There are some properties that will never be named’ cannot be interpreted as an ontologically harmless substitutional quantification. We can also count the things predicates or abstract singular terms stand for (e.g., ‘There are exactly two symptoms that mononucleosis and Barr-Epstein syndrome have in common’) and abstract singular terms can flank the identity predicate (e.g., ‘I believe in the unity of virtue: courage and temperance are the same thing’).

As long as we lack a precise mathematical characterization of English, it isn't possible to *prove* that certain idioms cannot be paraphrased away. But the use of abstract singular terms is so common and the failures of attempts to paraphrase them away are so clearcut that there is no reason to think that they could be eliminated from English without eviscerating it.

A relatively unsophisticated account of properties can be mobilized to explain the four phenomena listed above in a way that allows us to use a relatively straightforward referential semantics with objectual quantifiers. Such accounts explain the meanings of general terms (item 1) like ‘honest’ by claiming that they denote (or express) properties (like *honesty*), that a sentence like ‘Tom is honest’ has the logical form of a simple, subject-predicate sentence, and that it is true just in case the individual denoted by ‘Tom’ is in the extension of the property denoted (or expressed) by the predicate ‘honest’, which requires that there be a property expressed by this predicate (a slightly more formal account is given below; see Hochberg, 1968, for a good discussion of related issues).

In a similar spirit, some philosophers argue that abstract singular terms like ‘honesty’ (item 2) *denote* the property that the associated predicate (‘honest’) denotes or expresses, that sentences like ‘Honesty is a virtue’ have the simple logical form of a subject-predicate sentence, and that the sentence is true exactly when the word ‘honesty’ denotes a property that is in the extension of the property denoted by the verb phrase ‘is a virtue’.

Once we take these steps, it is also straightforward to explain the remaining items on our list. For example the validity of the argument: ‘Clinton is self-indulgent; therefore, there is at least one vice that Clinton has’ can be explained as follows: The logical form of the premise is that a simple subject-predicate sentence and the logical from of the conclusion is that of an existential quantification with a standard objectual quantifier. If the first sentence is true, then ‘self-indulgent’ expresses a property, and this property satisfies the open sentence ‘Clinton is X’. Hence, just as in standard first-order logic, the existential quantification is true. Similar maneuvers allows us to explain the remaining items on this list: if properties are genuine things, then we can count them and we can use different expressions to stand for the same property.

These explanations rely on little more than the following three claims. First, there is a rich enough stock of properties to provide a semantic value (meaning) for every predicate and abstract singular term of English (or better, for all of those that could have such semantic values without leading to paradox). Second, sentences like ‘Courage is a virtue’ and ‘John is courageous’ are simple subject-predicate sentences. Third, such sentences are true just in case the thing denoted by the subject is in the extension of the property denoted (or expressed) by the predicate.

These simple assumptions account for the phenomena on our list in a much better way than their more prominent rivals can. Some philosophers, for example, hold that predicates have a multiple denotation (multiply denoting all of the things to which they apply). Others hold that the semantic values of predicates are sets (the sets of things to which they apply). But these accounts cannot explain the fact that many pairs of predicates that in fact have the same extension (and hence the same multiple denotation‘) *could* have applied to different groups of things and that their meanings are precisely what allow them to do so. Even more seriously, these two rivals have no plausible account at all of the last three items on the list.

More formal accounts of language and logic

If the goal is simply to argue that there are properties because there is no other way to explain several obvious linguistic and logical phenomena (which is all many philosophers have aspired to show), then the simple accounts sketched above make a plausible (though certainly not unassailable) start. Some philosophers have set their sights higher, however, wanting to provide a rigorous and systematic account of the semantics of a large fragment of English. They try to work the above ideas out in a more detailed way and to extend them to deal with more complex phenomena, including the following:

1. Various English constructions are quite naturally interpreted as complex predicates: ‘Tom is a

boring but honest brother of Sam' is straightforwardly construed as a containing a compound predicate, 'is a boring but honest brother of Sam' that is predicated of the noun 'Tom' (and that could be predicated of other nouns too, e.g., 'Wilbur'). Other constructions are very naturally interpreted as complex singular terms (as in 'Being a boring but honest brother of Sam is no bed of roses'). Furthermore, these complex expressions are related to simpler expressions in systematic ways. For example, 'Tom is a boring but not dishonest brother of Sam' should entail 'Tom is not dishonest'.

2. English is full of intensional idioms like 'necessarily', 'believes' and 'imagines' that cannot be handled by any extensional semantics.

The simple, informal claim that there are properties cannot explain such phenomena in a systematic way, especially when they are combined (as in 'Tom believes that it is necessarily the case that being a seventh son is more like being a sixth son than like being a fifth son').

In recent years a number of philosophers (e.g., Bealer, 1982, 1994; Zalta, 1983, 1988; Chierchia & Turner, 1988; Menzel, 1993) have developed intricate accounts that include formal logics whose semantics provide systematic ways of forming "compound" properties (e.g., *loving Darla*) to serve as semantic values of complex predicates ('loves Darla') or complex singular terms ('loving Darla). The details of such accounts are too complex to pursue here (although a generic account of some of the central ideas will be sketched in §8). It should be noted, however, that most philosophers who aspire to a semantic account of large intensional fragments of English introduce propositions, which they treat as zero-place properties.

The proper treatment of intentional idioms like 'believes that' also require properties that are very finely individuated, probably as finely individuated as the linguistic expressions that denote or express them. For example Tom's grasp of logic may be so tenuous that he believes of Ortcutt that he is a spy and an auditor for the IRS but doubts that he is an auditor for the IRS and a spy. This is sometimes taken to suggest that *being a spy and an auditor for the IRS* is distinct from the (necessarily coextensive) property *being an auditor for the IRS and a spy*. To be sure, few people are guilty of such blatant lapses, but we can certainly make mistakes when necessarily coextensive properties are described in more complicated ways (such errors are routine in mathematics and logic).

On the plausible (though not inevitable) assumption that the structure of many of our thoughts is similar to the structure of the sentences we use to describe the contents of those thoughts ('Sam thinks Tom is boring but not dishonest'), we might also hope to use properties in an account of mental content that would in many ways parallel an account of the semantics of the more intensional fragments of English.

Beating the Competition

Accounts that treat the semantic values of predicates as sets can handle a certain amount of English if we are willing to twist ("regiment") it into a rather complex, even tortured logical form. But little is gained by this, since such approaches cannot accommodate such simple intensional phenomena as the fact that

two predicates might just happen to apply to exactly the same things even though they could have applied to different things. And extensional accounts do even worse with complex nominalizations or more complicated intensional idioms like ‘believes that’. Sets (of ordinary things) are simply too coarse-grained to make the fine distinctions semantic theories require.

Intensions

The only serious alternative to the use of properties in formal semantics treats the semantic values of noun phrases and verb phrases as *intensions*. Intensions are functions that assign a set to the expression at each possible world (or related set-theoretic devices that encode the same information). On such accounts, for example, the semantic value of ‘red’ is the function that maps each possible world to the set of things in that world that are red. Montague (1974) and linguists and philosophers inspired by his work have devised systems inspired by this idea that have great elegance and power. Nevertheless, properties are more natural and better suited to handle many linguistic constructions than intensions are.

Properties are more natural, because we learn the meanings of many predicates by ostension, and we group objects together when they share properties that seem salient or important. I recognize the sound of an oboe or the taste of rhubarb; these are very direct and simple experiences that seem completely unrelated to functions from huge infinite sets of possible worlds to objects therein. If we learn to recognize certain properties and categorize objects in terms of such properties, this is relatively easy to understand. But if the semantic values of predicates are intensions, meanings are now incredibly complicated set-theoretic objects that require a huge ontology of possible worlds and, often, merely possible individuals.

Properties are more useful in semantics than intensions because intensions are still too coarse-grained to explain many semantic phenomena involving intensional idioms. For example, semantic accounts that employ intensions would most naturally treat ‘lasted a fortnight’ and ‘lasted two weeks’ as having the same meaning (since they have the same intension), which makes it difficult for such accounts to explain how ‘Tom believes the battle lasted two weeks, but does not believe that it lasted a fortnight’ could be true. Various stratagems are available to deal with problematic cases like this, but they are much less natural and involve a much more dubious ontology (all those sets and possible worlds) than accounts that employ properties. Furthermore, intensions are unlikely to be able to perform tasks in areas outside semantics (like naturalistic ontology) that properties may be able to do. It is natural, for example, to suppose that things have the capacities that they do (e.g., the capacity to exert a force on a distant object) because of the properties they possess (e.g., gravitational mass). But it seems most unlikely that huge, set-theoretic intensions would be able to explain things like this.

Reductions of Properties

Some philosophers have construed intensions as providing a *reduction* of properties to intensions (properties are nothing over and above functions from the class of possible worlds to classes of objects). We have seen that this account has little to recommend it, and it is much better to view properties (including relations, and perhaps propositions) as primitive entities. Other philosophers, less concerned

with formal matters, have sometimes envisioned a reduction of properties to sets of tropes; a discussion of some of the issues this involves will be found in the entry on tropes.

Difficulties

Every large-scale theory of the semantics of English generates anomalies of one sort of another. Furthermore, some accounts require very large ontologies and very finely-drawn distinctions. For example, on really fine-grained accounts of the identity conditions of properties, the relations *loving* and the converse of its converse are distinct relations. Similarly, the properties *being red and square* and *being square and red* are distinct. We might wonder whether such distinctions exist and (if they do) what enables us to match the right linguistic expressions with the right relation? How *do* we match ‘red and square’ and ‘square and red’ with the correct members of the relevant pair of properties (we will return to this matter below)?

If the properties needed for semantics are completely isolated from the natural world, the epistemological problems noted in the previous subsection (on the philosophy of mathematics) resurface. We might hope to avoid this by holding that all properties are either instantiated or that they can be constructed by a series of applications of logical operations (like conjunction and negation) from properties that are instantiated. But it is far from clear that we can "construct" properties to serve as the semantic values for all English predicative expressions in this way. But could we define properties to serve as semantic values for all the predicates that lack instances? Expressions like ‘witch’ have a good bit of open texture, and it is at best an open question whether we can define them in terms of properties that are actually instantiated.

Current property-based semantic theories do not accommodate vagueness. This is a serious shortcoming, because vague predicates (like ‘bald’) and vague nominalizations (like ‘baldness’) are the rule, rather than the exception. When property-based semantic theories are modified to accommodate them, their proponents will have to decide whether vagueness is an objective feature in the world itself (so that some properties themselves are vague, in the sense of having vague or fuzzy extensions), or whether all vagueness resides in language (with properties having precise extensions and vagueness arising because it is sometimes somewhat indeterminate which sharp-edged property a given predicate or nominalization denotes).

Recent empirical work on concepts reinforces the point that many concepts (and, with them, predicates) have a graded membership and goes on to stress the importance of phenomena like typicality. Some creatures are more typical examples of birds than others, and there is some evidence that we determine whether something is a bird by assessing how similar (according to some psychological standard of similarity) it is to typical birds. This and various other phenomena have inspired a range of accounts of the structures of concepts, beginning with Rosch's (1978) account of prototypes and now including other accounts like exemplar theory (where we store exemplars of a concept in memory and determine what other things fall under that concept by assessing how similar they are to those exemplars).

Different accounts may well apply to different sorts of concepts (and perhaps, derivatively, to the predicates associated with them). For example, most mathematical concepts do have sharp boundaries, whereas many everyday concepts do not. On many recent psychological accounts, concepts involve features and similarity relations. Since features (e.g., having feathers, having a beak) are properties, there is no reason why current property theories could not be emended and extended to make contact with such accounts, and it seems likely that this will be a fruitful line of inquiry in the future (see Margolis & Laurence, 1999, for a useful selection of papers on concepts).

Lessons about Properties

What do semantic theories based on properties tell us about the *nature* of properties? The lessons here are less straightforward than in the philosophy of mathematics, partly because a detailed semantic theory must include a number of elements in addition to a theory of properties. For example, it must include a theory about the underlying logic in which the theory of properties is formulated, a theory about the logical forms of various English constructions (e.g., belief-sentences, gerundive phrases, parenthetical clauses), and perhaps claims that certain apparent entailment relations among English sentences don't really hold (e.g., because they are implicatures rather than logical entailments).

In short, we test a *total package* of such assumptions when we see how well a semantic theory accommodates our intuitions about what entails what or which groups of sentences are consistent. Moreover, somewhat different theories of properties may provide equally good accounts if we make compensatory adjustments in their underlying logics, in their accounts of the logical form of various constructions, or in our views about implicatures. Still, we have seen enough to draw some tentative lessons about properties from their use in semantics.

> **Existence Conditions:** If we want to account for the meanings of all predicates or all abstract singular terms (save for those which would lead to paradox), we need a very large stock of properties to serve as their semantic values (and since languages are extensible, we need properties to serve as the semantic values for any words that might ever be added).
>
> **Identity Conditions:** Even if we only aim to use properties as semantics values for run of the mill predicates, properties must be more finely individuated than sets. And if we hope to use properties as part of a systematic semantic account of belief attributions and other intensional idioms, they will have to be even more finely individuated than intensions. They will have to be (at least) nearly as finely individuated as the linguistic expressions that denote (or express) them.
>
> **Structure:** If we want to account for the behavior of complex predicates or complex singular terms in a systematic way, properties need to have something akin to a logical structure (we will explore the relevant notion of structure in §8.2).
>
> **Modal Status:** The use of properties in many parts of semantics does not obviously require

that properties exist necessarily. But when we turn to portions of English that explicitly involve the alethic modalities and related notions, i.e., when we turn to sentences (like ‘Red is necessarily a color’, ‘7 is necessarily prime’), the most natural accounts will involve properties that exist necessarily.

Epistemic Status: If properties are used to furnish semantic values for a multitude of expressions of a natural language like English or Choctaw, then we will need a lavish realm of properties that includes properties that are not instantiated. If such properties raise epistemological problems, then there will be difficulties explaining how our linguistic behavior, here in the natural world, involves properties we couldn't know much about. Furthermore, the more facts about language we can know *a priori*, the more likely it is that we will need some sort of *a priori* access to properties.

4.3 Naturalistic Ontology

In recent years properties have played a central role in philosophical accounts of scientific realism, measurement, causation, dispositions, and natural laws. This is a less unified set of concerns than those encountered in the previous two subsections, but it is still a clearly recognizable area, and I will call it *naturalistic ontology*. The use of properties in naturalistic ontology is often less formal and more varied than the work in the areas we have examined. I will indicate the flavor of this work by describing several noteworthy treatments of topics in the area.

Scientific Realism

Even quite modest and selective versions of scientific realism are most easily developed with the aid of properties. This is so because they offer a way to account for the following phenomena.

Quantification over Properties

Claims that appear to quantify over properties are common in science.

1. If one organism is fitter than a conspecific, then there is at least one property the first organism has that gives it a greater propensity to reproduce than the second.
2. There are many inherited characteristics, but there are no acquired characteristics that are inherited.
3. Properties and relations measured on an interval scale are invariant under positive linear transformations, but this isn't true of all properties and relations measured on ordinal scales.
4. In a Newtonian world all fundamental ("meaningful") properties are invariant under Galilean transformations, whereas the fundamental properties in a special-relativistic world are those that are invariant under Lorentz transformation.

No one has any idea how to paraphrase most of these claims in a non-quantificational idiom, and they

certainly seem to assert (or deny) the existence of various sorts of properties. The claim that this is in fact precisely what they do explains how they can be meaningful and, in many cases, true.

Functional Properties

Many important properties like *being a simple harmonic oscillator*, *being a gene*, *being an edge detector*, or *being a belief* are often thought to be functional properties. To be a gene, for example, is to play a certain causal role in the transmission of hereditary information, and it is in principle possible for quite disparate physical mechanisms to play this role. To say that something exemplifies a functional property is, roughly, to say that *there are* certain properties that it exemplifies and that together they allow it to play a certain causal role. For example, DNA molecules have certain properties that allow them to transmit genetic information in pretty much in the way described by Mendel's laws. Here again, we have quantifications over properties that seem unavoidable.

Causal Powers

Much explanation in science is causal explanation, and casual explanations often proceed by citing properties of the things involved in causal interactions. For example, electrons repel one another in the way that they do because they have *the same charge* (we will return to this below).

Reduction and Supervenience

A few decades ago claims that one sort of thing was *reducible* to a second were common; e.g., one often heard that the temperature of a gas is reducible to its mean molecular kinetic energy. Nowadays we are more likely to hear that one sort of thing *supervenes* on another: e.g., all biological (or all psychological) features of an organism supervene on its physical properties. Such claims make the best sense if we take them to involve properties. For example the claim that the psychological realm supervenes on the physical realm is plausibly construed as the claim that, necessarily, everything that has any psychological properties also has physical properties and any two things that have exactly the same physical properties will have exactly the same psychological properties. Disputes remain about the best way to spell out the fine print, but almost all of the candidates advert to properties.

Theory Change

Some philosophers of science, most notably Feyerabend and Kuhn, argue that theoretical terms draw their meaning from the theories within which they occur. Hence, they conclude, a change in theory causes a shift in the meanings of all of its constituent terms, and so different theories simply talk about different things. And since Newton's talk of ‘mass’ and Einstein's talk of ‘mass’ are about different things, their theories cannot be rationally compared; the theories are "incommensurable". The common realist rejoinder is that the reference of terms can remain the same even when the surrounding theory shifts (at least as long as it doesn't shift too much). Now it is certainly true that some realists have placed a greater explanatory burden on reference than it can bear. But for this response to work, even in cases of small shifts in theory, terms like ‘mass’ or ‘rest mass’ or ‘mass of 3.4kg’ must refer to something, and the most

plausible candidate for this is a property.

Measurement

Various features of measurement are most easily explained by invoking properties.

Different ways to Measure the Same Thing

Simpler anti-realist theories of measurement (like operationalism) cannot explain how we can use different methods to measure the same thing, e.g., how we can use such different methods to measure lengths and distances in cosmology, geology, histology, and atomic physics. By contrast, the view that measurements aim to discover objective properties can explain this.

Measurement Error is a Fact of Life

In many sciences it is expected that estimates of the magnitude of measurement error will be reported along with measurement results. Indeed, in fields like econometrics and psychometrics, extremely detailed theories of error are always near center stage. But such talk makes little sense unless there is a fact about what a correct measurement would be. Since an object can have one magnitude (e.g., a rest mass of 3kg) at one time and a different magnitude (e.g., a rest mass of 4kg) at another time, the object alone cannot explain this. But it is quite naturally explained by assuming that the object instantiates two different mass properties (namely a rest mass of 3 kg, and a rest mass of 4 kg) at the two different times. It also explains why later techniques for measuring things can be more accurate than earlier methods (e.g., why Atwood's machine allowed him to measure the value of the gravitational constant much more accurately than his predecessors could).

Measurement Units are Often Specified Using Properties

Nowadays measurement units are often specified directly in terms of properties. At one time the meter was specified as the length of the standard meter bar in Paris, But we now specify the meter in terms of something that can in principle be instantiated anywhere in the world, e.g., as *the length* equal to a certain number of wavelengths (in a vacuum) of a particular color of light emitted by krypton 86 atoms.

These facts have led to several adaptations of the representational theory of measurement developed by Suppes and his coworkers to a framework involving properties (Mundy, 1987; Swoyer, 1987). Among other things, these accounts offer characterizations of the algebraic structure of many of the properties involved in measurement.

Causal Powers

Some philosophers have employed properties in reductive accounts of causation (cf. Tooley; 1987; Fales, 1990). It would take us too far afield to explore this work here, but it is worth noting that it is never a

single, undifferentiated amorphous blob of an object (or blob of an event) that makes things happen. It is an object (or event) *with properties*. Furthermore, *how* it affects things depends on what these properties are. The liquid in the glass causes the litmus paper to turn blue because the liquid is an alkaline (and not because the liquid also happens to be blue). The Earth exerts a gravitational force of the moon because of their respective gravitational masses. And because explanations often cite causes, it is not surprising that explanations frequently cite properties: the liquid's being an alkaline explains why it turned the litmus paper blue (this doesn't preclude deeper explanations involving the molecular mechanisms that underlie this process, but they too will typically involve properties (like valence and charge)).

Some causal powers are deterministic: any object with a gravitational mass will exert a certain amount of force on an object with a certain gravitational mass at a certain distance from it. Others are indeterministic: photons can be prepared in a *state* that will give them a 50/50 chance of making it through a polarizer set at a certain angle. In some cases the *only* informative things we can say about a property are what tendencies or powers or capacities it confers on its instances. For example, the things we know about determinate charges have to do with the active and passive powers they confer on particles that instantiate them, their effects on the electromagnetic fields surrounding them, and the like. Thus, two negatively charged particles at a given distance will exert a force with a specific magnitude and direction on each other that depends on their respective charges (monadic properties) and the distance between them (a two-place relation) in accordance with Coulomb's law. Similar points hold for many other properties in science, including mass, momentum, force, electrical resistance, tensile strength, torque, and spin.

Such facts have led some philosophers to claim that properties are essentially dispositional, or even that properties *just are* dispositions. This led to a debate over whether all properties are dispositional (like charge and spin are) or whether some were non-dispositional (perhaps like squareness). The discussion here was considerably clarified by Shoemaker's (1984, p. 210ff) claim that it is linguistic items, rather than *properties*, that are dispositional or not. Some predicates, e.g., ‘fragile’, ‘flexible’, and ‘irascible’ are dispositional, whereas predicates, e.g., ‘square’ and ‘table’ arguably are not. But all properties confer causal powers on their instances; a square peg does not have the capacity to fit into a round hole (below a certain size).

Properties and Powers

Philosophers who focus on the causal or nomological capacities that properties confer on their instances often urge that properties are *identical* just in case they confer *the same* capacities on their instances (e.g., Achinstein, 1974; Armstrong, 1978, Ch. 16; Shoemaker 1984, Ch. 10-11). This general idea leaves us with questions about the relationship between properties and the capacities they bestow, but using fairly intuitive (though not incontrovertible) counting principles for properties and capacities, we can say the following:

> **Different Properties, Same Power:** Different properties can bestow the same powers on their instances. For example, charge and gravitational mass both bestow a power to exert a force on nearby objects (that have the right sorts of properties).

> **Same Property, Different Powers:** A single property can bestow different powers on its instances. For example, a determinate charge like the unit negative charge that characterizes electrons confers an ability to exert an attractive force on positively-charged particles and it confers an ability to exert a repulsive force on negatively-charged particles.

Although the connection between properties and powers is important, it isn't fully understood. Is a capacity an additional sort of property over and above the property that confers it? This sounds unduly complicated, but if this is not the case we need an account of the relationship.

Laws of Nature

Properties have played a central role in several recent accounts of natural laws. I will focus on two accounts that put properties at center stage; hybrids are possible, but the examples discussed here typify much recent work.

N-relation Theories

Laws of nature (e.g., the ideal gas laws, Newton's laws, Shrödinger's equation, Einstein's field equations for general relativity, conservation laws) have several important features, and the task of a philosophical account of laws is to explain how this is so. Different philosophers view different (and sometimes incompatible) features as central to laws, but those who favor what I will call *N-relation theories* agree that laws have (at least most of) the following five features. I will focus primarily on deterministic laws, not because they are more important than probabilistic laws, but because if an account cannot get deterministic laws right, it will have little chance with probabilistic laws.

1. Laws are objective. We don't invent laws, we discover them.
2. Laws have modal force. This shows up when we describe laws (or their implications) using words like ‘must’, ‘require’, ‘preclude’, and ‘impossible’.
3. Laws, unlike accidental generalizations, are confirmed by their instances and underwrite predictions.
4. The line between laws and non-laws is sharp; nomologicality does not come in degrees (this is implicit in the work of many *N*-relation theorists; Armstrong, 1983, p. 71 notes that his account depends on it).
5. Laws have genuine explanatory power. They play a central role in scientific explanation that accidental generalizations cannot.

N-relation theories have been defended by Armstrong (e.g., 1978, 1983), Dretske (1977), Tooley (1977) and others. Their accounts differ in detail, but they share the core idea that laws of nature are relations among properties. A law is a second-order relation of *nomic necessitation (N*, for short) holding among two or more first-order properties. Hence the *logical form* of a statement of a simple law is not ‘All Fs are Gs’; in the case of a law involving two first-order properties, it is a second-order atomic sentence of the

form ‘*N*(F,G)’.

In the more exact sciences these first-order properties (our *F*s and *G*s) will typically be determinate magnitudes like a kinetic energy of 1.6×10^{-2} joule or a force of 1 newton or an electrical resistance of 12.3 ohms (rather than mass or force or resistance simpliciter). Hence the laws specified by an equation (like Newton's second law) are really infinite families of specific laws where each specific, determinate mass *m* (a scalar, and so a monadic property) and total impressed force *f* (a vector, and so a relational property) stand in the *N*-relation to the appropriate relation (vector) of acceleration *a* (= *f/m*).

The Background: *N*-Relation Theories vs. Regularity Theories

The dominant accounts of laws during much of this century were *regularity theories*, and *N*-relation theories were originally devised to avoid perceived shortcomings of these earlier accounts. There are many versions of the regularity theory, but they share the core idea that laws are simply contingent regularities (or the sentences expressing them). On such views there is no *metaphysical* difference between genuine laws and true accidental generalizations (at least accidental generalizations involving purely qualitative predicates or properties) like ‘all cubes of pure gold weigh less than ten tons’ (which I'll assume is true). According to regularity theorists, the only difference between laws and accidental regularities is that laws have some special epistemic or pragmatic or logical trappings (e.g., they contain projectible predicates like ‘rest mass’ rather than ‘grue’ or they form part of a powerful deductive theory). The most prominent version of the regularity theory nowadays is the Ramsey-Lewis account, according to which laws are those universal generalizations that would be part of the overall systematization of our theories about the world that best combines simplicity and strength.

One of the chief attractions of regularity theories is that they have a relatively low epistemological cost. We observe instances of many regularities here in the actual world, and the additional features used to upgrade regularities to laws are not epistemically problematic in any deep way. Indeed, although there are various detailed problems with regularity theories, the major issues between *N*-relation theorists and regularity theorists involve the fundamental ontological tradeoff. According to *N*-relation theorists, regularity theories only achieve their epistemic security by being so weak that they cannot explain the fundamental features of laws. Regularity theorists counter that the *N*-relation is a mysterious bit of metaphysics, and that there is no way we could ever gain epistemic access to it. *N*-relation theorists respond that we should believe in it because it provides the best explanation of the five items on the above list. Is this response plausible? To evaluate it we need to look briefly at how those explanations are supposed to work.

N-relation Theories: Sample Explanations

Objectivity

According to *N*-relation theories, laws are objective because the *N*-relation relates those properties it does quite independently of our language and thought (in the case of properties that don't specifically involve

our language or thought). By contrast, the epistemic and pragmatic features used by regularity theorists to demarcate laws from accidental generalizations are too anthropocentric to account for the objectivity of laws.

Modal Force

Many laws seem to necessitate some things and to preclude others. Pauli's exclusion principle *requires* that two fermions occupy different quantum states. The special theory of relativity *doesn't allow* a signal to be propagated at a velocity exceeding that of light. The laws of thermodynamics show the *impossibility* of perpetual motion machines. Conservation laws assure us that such quantities as angular momentum, mass-energy, and charge *cannot* be created or destroyed. The modal force of laws is also said to manifest itself in the way laws support counterfactuals; had there been a tenth planet, it too would have obeyed Kepler's Laws. But, *N*-relation theorists insist, since regularity theorists forswear everything modal, they can never account for the modal aspects of laws.

Confirmation and Prediction

N-relation theorists often argue that their accounts can, and that regularity theories cannot, explain how laws are confirmed by their instances. *If* laws were mere regularities, then the fact that observed *F*s have been *G*s would give us no reason to conclude that those *F*s we haven't encountered will also be *G*. If the *F*s we have observed are to be *relevant* to our belief that unobserved *F*s are *G*s, then there needs to be *something about* an object's being *F* that requires (or, in the case of probabilistic laws, makes it probable) that it will also be *G*. And if the properties *F* and *G* stand in a nomic relation, then the properties themselves (and not merely their instances) are related in a law-like way. Hence, if *N*-relation accounts are right, there *will* be something about an object's being an *F* that will make it be a *G*, and the examined cases will be related to the unexamined cases in the relevant way.

A Nice Sharp Line

Properties either stand in the *N*-relation or they do not. When they do, we have a law; when they do not, we don't.

Explanation

The accidental regularity that all cubes of gold weigh less than ten tons doesn't explain why any particular cube of gold weighs less than ten tons. But, *N*-relation theorists often argue, if one property nomically necessitates a second, that does explain why anything having the first also has the second.

The Upshot

If *N*-relation accounts are on the right track, there is a reasonably rich realm of properties that is structured by one or more nomic relations. But before drawing this conclusion we should note that *N*-relation theories face difficulties of their own. Indeed, it is unclear whether *N*-relation theories can

successfully explain all of the things they were introduced to explain, but we will focus on two more general difficulties here. (A fuller discussion of the problems for *N*-relation theories can be found in the supplementary document [Difficulties for *N*-relation Accounts of Natural Laws].) First, it is not clear how to extend *N*-relation accounts to deal with several important kinds of laws, most prominently conservation laws and symmetry principles. Second, even in the case of laws that can be coaxed (or crammed) into the *N*-relation scheme, the account involves a highly idealized notion whose connection to the things that go by the name 'law' in labs and research centers is rather remote.

At this point some philosophers propose a distinction between the *current laws of science* and the *true laws of nature*. The former are approximate, idealized and provisional, whereas the later are precise, definite and unchanging. Furthermore, they continue, while it is perfectly respectable for philosophers to discuss the current laws of science, philosophy should also provide an account of the true laws of nature. But although some philosophers propose lists (like the one above) of features that are supposed to characterize the true laws of nature, it is not clear that there are any laws of this sort. At all events, current science doesn't force this conclusion on us, and the claim that there are such laws involves a bit of metaphysical speculation.

Properties, Powers and Laws

If we begin with actual scientific laws, we are likely to come up with quite different features from those on the [list above].

1. Laws almost always involve approximation and idealization Sometimes the idealization is so great that a law is quite inaccurate over parts of the range of phenomena it is supposed to cover (as is the law for the simple pendulum or the general gas laws). Most laws only hold *ceteris paribus*, "other things being equal," but other things rarely are.
2. When we apply a law to a situation, we often use a highly simplified version of the law that everyone acknowledges is false.
3. Laws are not in any straightforward way confirmed by their instances. Actual data and phenomena that provide evidence for a law rarely fit it exactly (even when we discount for measurement error).
4. We often explain things by citing the causal mechanisms and processes they involve, rather than by subsuming them under general laws. For example, we do not explain why all crows are black by saying (in some more idiomatic way) that the *N*-relation holds between the properties *being a crow* and *being black*. We explain it by finding causal (in this case genetic) mechanisms that link the two properties. In other cases we appeal to a deeper theory, e.g., we explain why Kepler's laws hold (to the extent that they do) by deriving (approximations of) them from Newton's laws.
5. The distinction between laws and accidental generalizations is a matter of degree. We often talk as though some laws (e.g., various conservation laws) are very fundamental and robust, while other laws (e.g., Hooke's Law, Boyle's Law, Gresham's Law) are less so.

A philosopher who sees 1-5 as central features of laws will be drawn to an account that is very different

from that proposed by *N*-relation theorists. Far from involving universal (or even precise probabilistic) nomological relations, actual laws are idealized, approximate, and limited in scope (often applying only to highly artificial systems created in laboratories or even just to simplified models of real systems).

When *N*-relation theories first appeared on the scene much of their appeal was that they promised a better account of the objectivity and (perhaps) the modal status of laws than regularity theories could provide. But it is now possible to discern the beginnings of an account that explains these things (to the limited degree that they hold) and that also explains how actual laws work. I will quickly sketch a generic version of such an account here (several versions are in the air, but most of them owe a large debt to Cartwright, 1983; 1989).

We have already noted that (at least many) properties confer causal capacities and tendencies on their instances. For example, electrically charged particles have a capacity to exert forces on other particles and to affect an electromagnetic field around them *in virtue of* the property of having a specific, determinate charge. This is a perfectly objective fact, and it has a certain modal force (if the particles had moved away from each other, the forces would have fallen off with the square of the distance between them). This suggests the view that laws result from the operations of capacities (including probabilistic capacities). Laws tell us what happens when we shield off (or hold constant) the influence of other capacities and allow a single capacity (or just a few capacities) to work without interference. In a few cases we can shield the operation of a single capacity from outside influences in a way that allows us to make fairly precise and accurate predictions, and cases like this may approximate the *N*-relation theorist's conception of a law. But most laws, and most applications of laws, aren't like this. The detailed behavior of most things, including many relatively simple physical systems, results from the joint operation of many capacities or tendencies, and often it cannot be predicted, or even explained, on the basis of such laws. Accounts like Cartwright's take science at face value and they leave room for laws in fields other than basic physics. But for our purposes the most important thing about them is that they, like *N*-relation theories, place properties at center stage.

Lessons About Properties

The work discussed in this subsection suggests that properties include determinate physical magnitudes like *being a mass of 3.7 kg* and *being an electrical resistance of 7 ohms*. Furthermore, such properties typically form families of ordered determinates (e.g., the family of determinate masses) that have a definite algebraic structure (Mundy, 1987; Swoyer, 1987). It also suggests that a fundamental feature of at least many properties is that they confer causal capacities on their instances. Work on naturalistic ontology doesn't entail detailed answers to every question about the nature of properties, but it does suggest answers to some of them.

> **Existence Conditions:** A natural, though not inevitable, conclusion to draw from the work discussed in this subsection is that properties exist only if they confer causal or nomological capacities on their instances. For properties: to be is to (be able to) confer causal capacities.

Identity Conditions: The most natural conclusion to draw here is that properties are identical just in case they confer exactly the same causal powers on their instances.

Structure: If *N*-relation theories are right, the realm of properties is structured in at least the minimal sense that many pairs of properties stand in one or more higher-order nomological relations. Properties are also related to the causal powers they confer on their instances in some intimate, though not clearly understood, way.

Modal Status: If laws of nature are metaphysically necessary, then properties that actually stand in the *N*-relation stand in that relation in all possible worlds in which they exist. The fact that properties confer causal powers on their instances is also naturally understood as the claim that the instances of a property have those powers in all possible worlds in which that property exists.

Epistemic Status: Philosophers who employ properties to provide explanations in naturalistic ontology typically hold that we learn about properties empirically. On some accounts all properties are instantiated, and we learn about them because their instances affect our sensory apparatus or our measuring instruments. On other accounts some properties are uninstantiated, but they are related to properties that are instantiated in systematic ways. For example, a specific determinate mass (e.g., 4 kg) might be uninstantiated, but we can describe it quite precisely (as twice as great a mass as 2 kg, which is, let us suppose, exemplified). Furthermore, we can say what causal powers it would have conferred on its instances, had it had any (e.g., we can say what gravitational force its instances would have exerted on a 2 kg object 5 meters away).

At this point it is useful to step back to note the fundamental way in which the *general* conception of properties discussed in this subsection differs from many of the conceptions discussed earlier. On those earlier conceptions at least many properties are causally inert, other-worldly, abstract entities that exist outside space and time; they are timeless, necessary beings, and since we cannot come into causal contact with them, our knowledge of them is problematic.

By contrast, the view that emerges from much of the work in naturalistic ontology treats properties as contingent beings that are intimately related to the causal, spatio-temporal order, and we learn what properties there are and what they are like through empirical investigation. Such properties are not much like meanings or concepts, and so it is possible to discover that a property conceived of in one way (e.g., the property of being water) is identical with a property conceived in some quite different way (e.g., the property of being H_2O). It might be misleading to call such properties ‘concrete’ (the standard antonym of the slippery word ‘abstract’), but it isn't quite right to call them ‘abstract’ either. Indeed, the stark dichotomy between the abstract and concrete is probably too simple to be useful here.

5 Existence Conditions

What properties are there? Under what conditions does a property exist? In formal accounts--those modeled on axiomatic set theory or axiomatic treatments of other mathematical entities--the goal is typically to find formal principles (like comprehension schemas) that state sufficient (and, with luck, necessary) conditions for the existence of properties. But the basic issues about the existence conditions of properties are not really formal ones. Indeed, views about their existence conditions typically derive from an interplay of views about the explanatory tasks of properties and legitimate constraints on philosophical explanation.

We can view the array of views about the existence conditions of properties as a continuum, with claims that the realm of properties is sparse over on the right (conservative) end and claims that it is bountiful over on the left (liberal) end.

5.1 Minimalism

According to minimalist conceptions of properties, the realm of properties is sparsely populated. This is a comparative claim (it is more thinly populated than many realists suppose), rather than a claim about cardinality. Indeed, a minimalist could hold that there is a large infinite number of properties, say that there are at least as many properties as real numbers. This would be a natural view, for example, for a philosopher who thought that each value of a physical magnitude is a separate property and that field theories of such properties as gravitational potentials are correct in their claim that the field intensity drops off continuously as we move away from the source of the field.

The best-known contemporary exponent of minimalism is David Armstrong (e.g., 1978, 1984, 1997), though it has also been defended by others (e.g., Swoyer, 1996). Specific reductionist motivations (e.g., a commitment to physicalism) can lead to minimalism, but here we will focus on more general motivations. These motivations typically involve some combination of the view that everything that exists at all exists in space and time (or space-time), a desire for epistemic security, and a distrust of modal notions like necessity. Hence, a minimalist is likely to subscribe to at least most of the following four principles.

1. The Principle of Instantiation

The *principle of instantiation* says that there are no uninstantiated properties. For properties: to be is to be exemplified. Taken alone the principle of instantiation doesn't enforce a strong version of minimalism, since it might be that a wide array of properties are exemplified. For example, someone who thinks that numbers or individual essences or other abstract objects exist would doubtless think that a vast number of properties are exemplified. So it is useful to distinguish two versions of the principle of instantiation.

> **Weak Instantiation:** All properties are instantiated; there are no uninstantiated properties.
>
> **Strong Instantiation:** All properties are instantiated by things that exist in space and time

(or, if properties can themselves instantiate properties, each property is part of a descending chain of instantiations that bottoms out in individuals in space and time).

Armstrong (1978) holds that properties enjoy a timeless sort of existence; *if* a property is ever instantiated, then it *always* exists. A more rigorous minimalism holds that properties are mortal; a property only exists *when* it is exemplified. This account has an admirable purity about it, but it is hard pressed to explain very much; for example, if laws are relations among properties, then a law would seem to come and go as the properties involved did.

2. Properties are Contingent Beings

Philosophers who subscribe to the strong principle of instantiation are almost certain to hold that *properties are contingent beings*. It is a contingent matter just which individuals exist and what properties they happen to exemplify, so it is a contingent matter what properties there are.

3. The Empirical Conception of Properties

A natural consequence of the view that properties are contingent beings is that questions about which properties exist are empirical. There are no logical or conceptual or any other *a priori* methods to determine which properties exist.

4. Properties are Coarse Grained

Those who hold that properties are very finely individuated will be inclined to hold that the realm of properties is fairly bountiful. For example, if the relation of *loving* and the converse of its converse (and the converse of the converse of that, and so on) are distinct, then properties will be plentiful. Minimalists, by contrast, are more likely to hold that properties are identical just in case they necessarily have the same instances or just in case they bestow the same causal powers on their instances. On these views the converse of the converse of a binary relation will just be that relation itself (we will return to this matter in the section on identity conditions of properties).

Locations?

The strong principle of instantiation opens the door to the claim that properties are literally located in their instances. This is a version of Medieval philosophers' doctrine of *universalia in rebus*, which was contrasted with the picture of *universalia ante rem*, the view that properties are transcendent beings that exist apart from their instances. With properties firmly rooted here in the spatio-temporal world, it may seem less mysterious how we could learn about them, talk about them, and use them to provide illuminating explanations. For it isn't some weird, other-worldly entity that explains why this apple is red; it is something *in* the apple, some aspect of it, that accounts for this. It is easier, however, to think of monadic properties as located in their instances than it is to view relations in this way (this may be why Aristotle and the moderate realists of the middle ages viewed a relation as an accident that inheres in a

single thing). Nevertheless, the general feeling that transcendent properties couldn't explain anything about their instances has figured prominently in many debates over properties.

Minimalists must pay a price for their epistemic security (there's no escaping the fundamental ontological tradeoff). They will have little hope of finding enough properties for a semantic account of even a modest fragment of any natural language and they will be hard pressed (though Armstrong, 1997, does try) to use properties to account for phenomena in the philosophy of mathematics. Minimalists may not be greatly bothered by this, however, for many of them are primarily concerned with issues in naturalistic ontology.

5.2 Maximalism

At the other, left, end of the spectrum we find maximalist conceptions of properties. Borrowing a term from Arthur Lovejoy, maximalists argue that properties obey a *principle of plenitude*. Every property that could possibly exist *does* exist. For properties: To be is to be possible (Linsky & Zalta, 1995; cf. Jubien, 1989). If one accepts the view that properties are necessary beings, then it is a simple modal fact that if a property is possible it is necessary and, hence, actual.

Just as the principle of instantiation alone does not guarantee minimalism, the principle of plenitude alone does not guarantee maximalism. One can endorse the former while holding that all sorts of properties are instantiated, and one can endorse the latter by holding that very few properties are possible (an actualist who subscribed to the strong principle of instantiation might hold this). So to get to the maximalist end of the spectrum we need to add the claim that a vast array of properties is possible. Various formal principles, e.g., a strong comprehension principle (e.g., Zalta, 1988) and axioms ensuring very finely-individuated properties (e.g., Bealer, 1982, p. 65; Menzel, 1986, p. 38) are two formal ways to achieve this.

Maximalist accounts are often propounded by philosophers who want to explain meaning and mental content, but since such accounts postulate so many properties that maximalists have the resources to offer accounts of other things (e.g., phenomena in the foundations of mathematics), and many do. Indeed, the great strength of maximalism is that its enormously rich ontology offers the resources to explain all sorts of things.

Epistemology is the Achilles' heel of maximalism. At least some philosophers find it difficult to see how our minds could make epistemic contact (and how our words could make semantic contact) with entities lying outside the spatio-temporal, causal order. But maximalism has its advantages. Those maximalists who are untroubled by epistemic angst typically remain maximalists. By contrast, philosophers who begin as minimalists sometimes feel pressure to move to a richer conception of properties, either to extend their explanations to cover more phenomena or, sometimes, even just to adequately explain the things they started out trying to explain (e.g., Armstrong's more recent work is somewhat less minimalist than his earliest work).

5.3 Centrism

There is a large middle ground between extreme minimalism and extreme maximalism. For example, several philosophers primarily concerned with physical ontology have urged that a limited number of uninstantiated properties are needed to account for features of measurement (Mundy, 1987), vectors (Bigelow and Pargetter, 1990, p. 77), or natural laws (Tooley, 1987). Such accounts can, like minimalism, treat properties as contingent, fairly coarsely individuated, and too sparse to satisfy any general comprehension principles (e.g., they may deny that there are negative or disjunctive properties). One can also arrive at a centrist position by endorsing a comprehension principle but adding that it only guarantees the existence of properties built up from an initial, sparse, stock of simple properties (cf. Bealer, 1994, p. 167).

Being moderate isn't always easy, and it can be difficult to stake out a position in the center that doesn't appear arbitrary. Once *any* uninstantiated properties are admitted, we are in much the same epistemological boat as the maximalist. No doubt the minimalist will see this as a reason to reject any uninstantiated properties, while the maximalist (who believes that epistemological problems can be overcome) will see it as a reason to admit as many as possible.

5.4 Dual-Entity Accounts

There is some reason to think that accounts in different fields (e.g., semantics and natural ontology) may call for entities with different identity conditions; for example, semantics requires very finely individuated properties whereas naturalist ontology may need more coarsely individuated ones. If this is so, then no single kind of entity could do both kinds of jobs. The minimalist is likely to conclude that it is a mistake to employ properties in semantics. But less squeamish philosophers may instead conclude that there are (at least) two different sorts of property-like entities. Bealer's (1982) qualities and concepts and Lewis's (1986) sparse and abundant properties are examples of this approach. But this happy hybrid won't satisfy everyone, since minimalists (and some centrists) will reject the view that there are any concepts or abundant properties.

Summary

Disputes over the existence conditions of properties splinter into several related, but distinct, issues.

1. Instantiation issues: Must a property be exemplified to exist?
2. Localization Issues: Do exemplified properties exist in space and time (namely where they are instantiated)?
3. Modal issues: Are properties contingent beings or are at least some necessary beings?
4. Epistemological issues: Is the only way to discover the existence of properties though empirical means?
5. Individuation issues: How finely individuated are properties?

Minimalists hold that *all* properties are contingent beings, that they must be exemplified in space and

time to exist, that we can only discover their existence empirically, and that they are fairly coarsely individuated. Some minimalists also hold that they exist in those locations where they are instantiated. Maximalists reject all of these views. These two sets of views form fairly natural packages, but other combinations are possible, and they lead to views falling between the two extremes.

6 Identity Conditions

What are the identity conditions for properties? An answer would give us necessary and sufficient conditions for the properties *x* and *y* to be one and the same property. Another way to pose the question is to ask how finely individuated properties are, and here we find a spectrum of views.

- **Infra-coarse:** Properties with the same extension are the same properties (this holds for Frege's concepts, but all contemporary property theories reject this claim; it is included simply as a point of reference).
- **Medium coarse:** Properties are identical just in case they necessarily have the same extension.
- **Medium Fine:**
  - Properties are identical just in case they confer the same causal or, more generally, the same nomological powers on their instances (i.e., they are identical if they play the same functional role), or
  - Properties are identical just in case they are encoded by exactly the same abstract objects (Zalta)
- **Ultra-fine:** Properties are individuated almost as finely as the linguistic expressions that express them. A natural way to work this out is to develop an account of the analysis of a property and to hold that properties are identical just in case they have the same analysis (cf. Bealer's account of concepts; Menzel, 1993).

On most medium-fine views, formal considerations alone do not determine identity conditions for properties.

6.1 Necessary Coextension

Probably the best-know candidate for an identity condition for properties is necessary coextensiveness. This seems to transpose the identity conditions for sets into an appropriately intensional key, and this is precisely how identity conditions for properties work in accounts that treat them as intensions (as functions from possible worlds to objects therein). Bealer also views this as the identity condition for his sparse properties, qualities and connections (though he is undogmatic about this).

Although necessary coextension may be the most-discussed candidate identity condition for properties, many realists reject it because it doesn't comport well with the explanations they want to develop. Identity conditions don't matter greatly when the aim is simply to explain mathematical phenomena; even extensional entities like sets could do that job, so we don't need necessary coextension as an identity

condition here. This proposal is also in tension with the picture that properties are individuated by their functional roles, at least on the assumption that necessarily coextensive properties can confer different causal powers on their instances (Sober, 1982, contains a strong argument that this can happen, though the jury is probably still out on this issue). And in semantics we need hyper-intensional properties that are individuated much more finely than the necessary-coextension condition allows.

6.2 Functional Role

The view that properties are identical just in case they confer the same causal, or more generally, the same nomological or functional roles on their instances has been endorsed by various philosophers who work primarily in scientific ontology. On this conception, there are few, if any, purely formal identity conditions for properties (unless, as seems unlikely, someone devises a purely formal account of causal roles). One cost of this view is that the notion of a causal role and the relationship between such roles and properties is not completely clear.

6.3 Property Identity in Terms of Encoding

We have encountered all of the major current views about the identity conditions of properties except for one in previous sections, so we will go into it in a bit more detail here. The *encoding account of property identity*, proposed in Zalta (1983; 1988), is developed in the context of a rich theory of properties that goes along with a rich theory of abstract objects. In this theory, ordinary objects (like Bill Clinton) exist, exemplify properties, but cannot encode properties. By contrast (on the most common interpretation of his system), abstract objects (like Pegasus and the Euclidean triangle) exist necessarily, but they necessarily fail to have a spatio-temporal location. Abstract objects encode, as well as exemplify, properties; indeed, an abstract object is *constituted* by the properties it encodes. For example, the abstract object the Euclidean triangle encodes all and only those properties implied by being a triangle of Euclidean geometry (e.g., being a closed three-side plane figure whose interior angles sum to 180 degrees). This abstract object also exemplifies properties, e.g., being mentioned in all textbooks on plane geometry.

The metaphysical theory of abstract objects is linked to our actual thought and talk by the bridge principle that the English copula is ambiguous; sometimes ‘is’ means ‘exemplifies’, sometimes ‘encodes’. The existence condition for abstract objects is given by a comprehension schema according to which there is (necessarily) a unique abstract object that encodes just those properties satisfying each condition on properties specifiable in the language of the theory, and abstract objects are identical just in case they encode exactly the same properties. More importantly, for current purposes, properties are identical just in case they are, necessarily and always, encoded by exactly the same individuals.

On this account, properties that necessarily have the same *encoding extensions* are *identical*, but properties that necessarily have the same *exemplification extensions* may be *distinct*. To see the difference, note that the property of being a round square and the property of being a round triangle necessarily have the same exemplification extension. Hence, accounts (like those noted in §6.3) that treat

necessary (exemplification) coextension as sufficient for identity tell us that they are one and the same property. But since these properties have different encoding extensions, the present account treats them as distinct. One can even make this work without any actual abstract objects, though the nonidentity of two properties would still require that there *could* have been an abstract object that encodes one but not the other. This is one of the few novel accounts of property identity to be proposed in recent decades. It has the virtue of being part of a detailed theory that has been employed to explain a wide range of phenomena, and it expresses the identity conditions for properties in terms of one of their most fundamental features, namely that they are predicable entities. The price is that it requires us to hold that there are two modes of predication and that there are, or at least that there could have been, abstract objects.

6.4 Ultra-fine Properties

The view that properties have ultra-fine identity conditions is typically developed in the context of a rich formal theory of properties. One could mandate fine-grained identity conditions by brute force, e.g., by laying down axioms that each set of objects is in the extension of more than one property, or in the extension of many different properties. But the intuitive idea behind property theories tailored to semantics is that there are "compound properties" which are built up from simpler properties by logical operations akin to conjunction, negation, and so on. On such accounts the property *being red and square* is distinct (because built up in a different way) from the property *being square and red*. Similarly, the converse of the converse of a two-place relation is distinct from that relation itself.

Such accounts may be well-suited for explaining strongly intensional phenomena (like belief sentences or mental content), but they also raise certain questions. For example, what is the difference between the property *being red and square* and the distinct property *being square and red*, and what allows us to link the right complex predicate (say 'is red and square') to the right property (*being red and square*) rather than to the wrong one (*being square and red*)? If properties literally had parts, the answer might be that the arrangements of things in the two properties is different (e.g., *being red* comes "first" in *being red and square*). Explaining what such structural differences amount to would not be easy, but at least such differences might point to the beginnings of a solution. By contrast, if properties lack genuine internal structure, it is less clear how an account of the difference between *being red and square* and *being square and red* would even begin. We will return briefly to such matters in the final section.

7 Kinds of Properties

Most realists agree that there are various sorts of properties, and in this section we will review the main kinds of properties they have proposed. But many realists are also selective; they believe that some, but not all, of these kinds of properties exist. Indeed, *almost none* of the putative kinds of properties discussed here is accepted by all realists, but to avoid constant qualifications (like 'putative kind of property') I will present each sort of property as though it were unproblematic.

First-order vs. Higher-order Properties

The first set of issues we will examine involve the most fundamental logical or structural features of properties. We will begin with a picture of a hierarchy of properties arranged according to order. First-order properties and relations are those that can only be instantiated by individuals. For example *redness* can be instantiated by the apple on my desk and *being married to* can be instantiated by Bill and Hillary, but no properties can be red or married. It is natural to suppose, however, that at least many first-order properties and relations can themselves have properties and relations. For example, *redness* might be thought to exemplify the property of *being a color* and *being married to* might be thought to exemplify the property of *being a symmetrical relation*. Once we think of second-order properties, it is natural to wonder whether there are third-order properties (properties of second- or, perhaps in cumulative fashion, of second- *and* first-order properties), and so on up through ever-higher orders.

This metaphysical picture finds a formal parallel in higher-order logics. On one common system of classification, we move from familiar first-order logic to second-order logic by adding first-order variables, from second- to third-order logic by adding second-order constants, from third- to fourth-order logic by adding second-order variables, and so on up, alternating constants and variables at successive steps.

Realists differ over which niches in this proposed hierarchy of orders are occupied. Proponents of the empirical conception of properties will hold that it is an empirical question whether there are second- or forth- or fifty-seventh-order properties. The issue for them is likely to be whether putative higher-order properties confer any causal powers on their instances over and above those already conferred by lower-order properties. But it is also possible to have less empirically motivated views about which parts of the hierarchy are occupied.

Elementarism (Bergmann, 1968) is the view that there are first-order properties but that there are no properties of any higher-order. There are first-order properties like various shades of red, but there is no higher-order property (like *being a color*) that such properties share nor are they related by any higher-order relations (like *being darker than*).

Elementarism has sometimes been defended by appealing to something like Russell's *principle of acquaintance*, the tenet that only things with which we are acquainted should be thought to exist, together with the claim that we are acquainted with first-order properties but not with those of any higher orders. To the extent that first-order properties are able to perform all of the tasks that properties are called on to do, elementarism could also be defended on grounds of parsimony. But it is now widely acknowledged, even by minimalists, that at the very least some higher-order relations are needed to confer structure on first-order properties.

Self-predication and Theories of Types

In May of 1901 Russell discovered his famous paradox. If every predicative expression determines or corresponds to a property, then the expressions 'is a property that does not instantiate itself' should do so. This raises the question: does this property instantiate itself? Suppose that it *does*. Then it is a property that does not instantiate itself; so if it does instantiate itself, it doesn't instantiate itself. Now suppose that it does *not* instantiate itself. Then it is one of those properties that does not instantiate itself; so it does instantiate itself. Such a property, which instantiates itself if and only if it does not instantiate itself, cannot exist (on pain of contradiction). This led Russell to introduce a theory of types which institutes a total ban on self-exemplification by a strict segregation of properties into orders (his account is actually even more restrictive than this; see the entry on Russell's paradox for details).

Untyped Conceptions of Properties

Russell's reaction seems extreme, because many cases of self-exemplification are innocuous. Furthermore, realists who are not minimalists or conservative centrists are likely to think that self-exemplification is common. For example, the property of being a property is itself a property, so it exemplifies itself. There also seem to be transcendental properties and relations. A transcendental relation like *thinks about* is one that can relate quite different types of things: Hans can think about Vienna and he can think about *triangularity*. But typed theories cannot accommodate transcendental properties without several epicycles.

Several recent accounts (e.g., Bealer, 1982; Menzel, 1993) treat properties as entities that can exemplify themselves. From this perspective, the picture of a hierarchy of orders is fundamentally misguided; there are simply properties (which can be exemplified--in many cases by other properties, even by themselves) and individuals (which cannot be exemplified). One challenge here is to develop formal accounts that allow as much self-exemplification as possible without teetering over the brink into paradox. In formal systems where abstract singular terms or predicates may (but need not) denote properties (cf. Swoyer, 1998), formal counterparts of (complex) predicates like 'being a property that does not exemplify itself' could exist in the object language without denoting properties; from this perspective Russell's paradox would merely show that this predicate lacked a denotation, rather than that the logic was inconsistent.

Relations

We have treated relations as properties (with more than one argument place); for example the properties of *loving* and *being shorter than* are two-place relations, that of being in *between* (two other things) is a three-place relation, and so on. On abundant conceptions of properties, there are relations of every finite number of argument places, but on sparse conceptions it is an empirical question whether there are relations of any particular degree (i.e., with any particular number of argument places). What we think of as genuine relations were not recognized by philosophers until about a century and a half ago (with the work of DeMorgan, Schroeder, and Peirce and, somewhat later, Russell). Until then relations were treated as a special sort of monadic property (when they were acknowledged at all).

We have seen how several selective realisms focus on the hierarchy of orders, but selective realisms can also focus on the *degree* (the number of argument places) of relations. Leibniz argued that relations could be reduced to monadic properties (though he never really explained just how this was to work) and so were dispensable. Some philosophers still hold that relations supervene on the monadic properties of their relata in a very strong sense that shows that relations are not actually real (some trope theorists hold this view; it is defended at length in Fisk, 1972). But no one has been able to show that all relations do supervene on monadic properties, and there are strong reasons for thinking that at least some sorts of relations, e.g., spatio-temporal ones, do not. Hence, most contemporary realists hold that there are genuine relations.

Other selective realisms are possible. For example, in contrast with Leibniz's view one might hold that there are relations but no monadic properties (this view is sometimes attributed, with very little textual support, to Peirce). And Armstrong (1978, ch. 24) proposes the tentative hypothesis that there are first- and second-level monadic properties, but no monadic properties of any higher-level.

Fixed-degree vs. Multigrade Properties

Many predicates are *multigrade* or *variably polyadic*; they can be true of various numbers of things. For example, the predicate ‘robbed a bank together’ is true of Bonnie and Clyde, Ma Barker and her two boys, Patti Hearst and three members of the Symbionese Liberation Army, and so on. Multigrade predicates are very common (e.g., ‘work well together’, ‘conspired to commit murder’, ‘are lovers’). Some of them can be analyzed as conjunctions of fixed-degree predicates, but many of them cannot. Standard logic does not accommodate multigrade predicates, but they are very common, and if the goal is to use properties as semantic values of English predicates, then multigrade properties are needed. They have also been used in naturalistic ontology in an ingenious treatment of measurement (Mundy, 1989). A truly flexible account of properties would abandon both the restrictive hierarchy of orders and the equally restrictive view that all properties come with a fixed number of argument places, but as yet little work of this sort has been done.

Propositions

In ancient and Medieval times propositions were not seen as a special kind of property (in those days philosophers didn't even recognize genuine autonomous relations), and many contemporary philosophers

who focus on physical ontology or philosophy of mathematics do not regard propositions as a kind of property (many of them doubt that there are any such things). But those who work on the semantics of natural language often postulate the existence of propositions, noting that we can think of them as a limiting case of a property. Consider a two-place property like *loves* and think of plugging one of its open places up with Darla to obtain the one-place property *loves Darla*. If we can do this, it is sometimes argued, then we can plug the remaining (last) open place up with Sam to get the zero-place property, or proposition, *that Sam loves Darla*.

Structured vs. Unstructured Properties

Some philosophers (e.g., Grossman, 1983, §§58-61) argue that all properties are simple. Others argue that there is a distinction between *simple properties* and *compound properties*, that some compound properties exist, and that they have a structure that involves or incorporates simpler properties. Properties might have different sorts of structure, including various sorts of algebraic or logical structure. Because such issues often arise in connection with formal accounts of properties, this issue is discussed in the section on formal theories of properties.

Instantiation

If instantiation or exemplification is just another run-of-the-mill relation, it appears to lead to a vicious regress. This is often known as *Bradley's regress*, although it is doubtful that Bradley himself had this particular regress in mind. The construal of the Bradley's regress that has passed into the literature goes like this. Suppose that the individual a has the property F. For a to instantiate F it must be *linked to F* by a relation of instantiation, I. But (here's where the trouble begins), this requires a further pair of relations, R_1 and R_2, one to connect a to I and a second to connect I to F. This in turn requires four additional relations to bind R_1 and R_2 to the things that they are supposed to relate, then eight further relations to fasten these four relations to their relata, and so on without end. It is sometimes suggested that the regress is innocuous, but the problem isn't simply that there is a regress. The problem is that at each "stage" further relations are required, but they are *never* able to link their would-be relata. The difficulty is that nothing *ever* gets connected to anything else.

The only way to avoid this difficulty is to insist that instantiation is not a relation, at least not a normal one. Some philosophers hold that it is a *sui generis* linkage that hooks things up without intermediaries. Strawson (1959), following W. E. Johnson, calls it a *non-relational tie*; others stress that it is a mode of predication. It may even be that there is no such thing as instantiation at all and that talk of it is just a misleading figure of speech. At this point it is natural to resort to metaphors like Frege's claim that properties have gaps that can be filled by objects or the early Wittgenstein's suggestion (if we read him as a realist about properties) that objects and properties can be hooked together like links in a chain. Broad likened instantiation to Metaphysical Glue, noting that when we glue two sheets of paper together we don't need additional glue, or mortar, or some other adhesive to bind the glue to the paper (Broad 1933, p.85). Glue just sticks. And instantiation just relates. It is metaphysically self-adhesive.

Particularizing Properties

Some properties are metaphysical analogues of count nouns. They have been called *sortal properties* (by Strawson) and *particularizing properties* (by Armstrong), but the ideas involved here have a long history. Strawson borrows the word 'sortal' from Locke, and at least some particularizing properties correspond closely to Aristotle's secondary substances. Particularizing properties provide counting principles, or principles on identity, in the sense that they allow us to count objects. For example, the properties of *being a table* and *being a cat* are particularizing properties; there are definite facts of the matter as to how many tables are in the kitchen and how many cats are on those tables. There are also properties, e.g., *intervention*, *bombing*, that particularize events.

Characterizing Properties

Particularizing properties are naturally contrasted with *characterizing properties*. Characterizing properties like *redness* and *triangularity* do not divide the world up into a definite number of things. To the extent that a property like *redness* seems to allow us to count red things, it is because we are relying on the umbrella count noun 'thing' to help with the count.

Mass Properties

Particularizing properties may also be contrasted with *mass properties*. These are properties, like *water*, *gold*, and *furniture*, that apply to stuff. Like characterizing properties, mass properties do not divide the world up into definite numbers of things, but many characterizing properties apply to individuals, whereas mass properties apply to stuff. It was noted above that different sorts of linguistic categories that might correspond to important ontological distinctions are run together we represent all of them as predicates of a formal language. It now appears, for example, that common nouns express particularizing properties, while adjectives typically express characterizing properties.

Genus and Species

Although the notions of *genus* and *species* play a relatively small role in contemporary metaphysics, they figured prominently in Aristotle's philosophy and in the many centuries of work inspired by it. When we construe these notions as properties (rather than as linguistic expressions), a genus is a general property and a species is a more specific subtype of it. The distinction is typically thought to be a relative one: *being a mammal* is a species relative to the genus *being an animal*, but it is a genus relative to the species *being a donkey*. It has usually been assumed that in such chains there is a top-most, absolute genus, and a bottom-most, absolute species.

It was traditionally supposed that a species could be uniquely specified or defined in terms of a genus and a differentia. For example, the property *being a human* is completely determined by the properties *being an animal* (genus) and *being rational* (differentia). It is difficult, by today's lights, to draw a principled distinction between genera and differentiae, but the idea that species properties are compound,

conjunctive properties remains a natural one. For example, the property of *being a human* might be identified with the conjunctive property *being a human and being rational*. But it is now rarely assumed, as it was for many centuries, that all compound properties are conjunctive.

Determinables and Determinates

The concepts of *determinables* and *determinates* were popularized by the Cambridge philosopher W. E. Johnson. Properties like *color* and *shape* are determinables, while more specific versions of these properties (like *redness* and *octagonality*) are determinates. Like the distinction between genus and species, the distinction between determinables and determinates is a relative one; redness is a determinate with respect to color but a determinable with respect to specific shades of red. The hierarchy is generally thought to bottom out, however, in completely specific, absolute determinates.

Species are often thought to be definable in terms of a genus and a differentia. But determinates are not definable in terms of a determinable and a differentia; indeed, they are not conjunctive properties of any obvious sort. Determinates under the same determinable are incompatible; nothing can instantiate both of them at the same time, and anything that exemplifies a determinate must exemplify its determinables as well. The distinction between determinables and determinates has played a larger role in recent metaphysics than the more venerable distinction between genus and species. For example, much recent work in naturalistic ontology treats physical magnitudes as absolute determinate properties (like *being a mass of 3 kg* or *being a force of 17 newtons*) falling under determinables like *mass* and *force*.

Any object that instantiates a determinate (e.g. red) must have the corresponding determinable (e.g., color), and Armstrong has raised the question whether determinables are genuine properties or whether we simply apply determinable predicates to things on the basis of the determinate properties that they have. The determinable properties of a thing (if there are any), are necessarily supervenient on the determinates of the thing, in the sense that two things that have exactly the same determinates must also have exactly the same determinables. For example, if two things exemplify the same determinate mass, then both must have a mass. Armstrong (1997, p. 45) urges that this issue is part of a more general issue as to whether necessarily supervenient properties are anything over and above the properties on which they supervene. His answer is that they are not; they are a "metaphysical free lunch." He offers little argument for this claim, however, and many philosophers would demur.

Natural Kinds

Natural Kind Properties are important properties that carve nature at its natural joints. Paradigms include the property of being a specific sort of elementary particle (e.g., the property of being a neutron), chemical element (e.g., the property of being gold), and biological species (e.g., the property of being a jackal). Natural kinds are often contrasted with artificial kinds (e.g., being a central processing unit).

In recent years a good deal of work has been done on the semantics of natural kind terms (involving such issues as whether they are rigid designators). Less work has been done on the ontology of natural kinds,

but it is clear that it is most plausible to speak of natural kinds in those cases where something has what Locke called a *real essence* (in the way that elementary particles or chemical elements probably do). In these cases it seems plausible to suppose that natural kind properties are compound properties that involve simpler properties (e.g., the quantum numbers, for elementary particles; being made of simpler parts standing in specific relations in the way that chemical elements are made up of atoms related by chemical bondings).

The distinctions between natural and artificial kinds and that between particularizing and mass properties are orthogonal to each other. Some natural kind properties, e.g., *being a dog*, are particularizing while others, e.g., *gold*, are not. Likewise, some artificial properties, e.g., *being a table*, are particularizing while others, e.g., *furniture*, are not. The chief issue here is whether there are any natural kinds or whether our classifications are primarily a matter of cultural and linguistic conventions that represent just one of many ways of classifying things (so that joints are a result of the way that we happen to carve things up).

Purely Qualitative Properties

Some properties involve or incorporate particulars. The properties of *being identical with Harry* and *being in love with Harry* involve Harry. Even those who think that lots of properties exist necessarily often believe that non-qualitative properties like this are contingent; they depend upon Harry, and they only exist in circumstances in which he exists. By contrast, purely qualitative properties (like *being a unit negative charge* or *being in love*) do not involve individuals in this way. The distinction between properties that are purely qualitative and those that are not is usually easy to draw in practice, but a precise characterization of it is elusive.

Essential Properties and Internal Relations

A (monadic) property is an essential property of an individual just in case that individual has the property in every possible circumstance in which the individual exists. Essential properties are contrasted with accidental properties, properties that things just happen, quite contingently, to have. My car is red but it could have been blue (had I painted it), so its color is an accidental property. It is doubtless an essential property of my car that it is extended, but interesting examples of essential properties are more controversial--so controversial that some philosophers have doubted whether there are any. It is sometimes suggested, though, that if something is a member of a natural kind, then being a member of that kind is essential to it; for example, *being human* is an essential property of Saul Kripke.

The phrase ‘internal relation’ has been used in different ways, but it is often used as the relational analogue of an essential (monadic) property. For example, if *a* bears the relation *R* to *b*, then *R* internally relates *a* to *b* just in case *a* bears this relation to *b* in every possible circumstance in which they both exist. Relations that are not internal, that contingently link their relata (the things they relate), are external. Bill and Hillary are married, but they might not have been, so this relation between them is external. By contrast, some philosophers have suggested that the relation *being a biological parent of* is an internal

relation. In every world in which Bill and Chelsea both exist, Bill is her father. If this is correct, then the relational property, *being a child of Bill* is essential to Chelsea, but *being the father of Chelsea* is not essential to Bill (he and Hillary might never have met, in which case they would not have had Chelsea).

Intrinsic vs. Extrinsic Properties

Some properties are instantiated by individuals because of the relations they bear to other things. For example the property *being married* is instantiated by Bill Clinton because he is *married to* Hillary Clinton. Such properties are sometimes called *extrinsic* or *relational properties*. Objects have them because of their relations to other things. By contrast, *intrinsic* or *non-relational properties* are properties that a thing has quite independently of its relationships to other things. Many properties that seem to be intrinsic turn out to be extrinsic when we examine them carefully. The main questions here are whether there are any interesting intrinsic properties and how the notions of intrinsicness and extrinsicness are to be explicated.

Primary vs. Secondary Properties

The distinction between primary and secondary properties goes back to the Greek atomists. It lay dormant for centuries, but was revived by Galileo, Descartes, Boyle, Locke, and others during the seventeenth century. Locke's influence is so pervasive that such properties still often go under the names he gave them, *primary* and *secondary qualities*.

The intuitive idea is that primary properties are objective features of the world; on many accounts they are also fundamental properties that explain why things have the other properties that they do. Early lists of primary properties included *shape*, *size*, and (once Newton's influence was absorbed) *mass*. Today we might add properties like *charge*, *spin* or the *four-vectors* of special relativity to the list of primary properties. By contrast, secondary properties somehow depend on the mind; standard lists of secondary properties include *colors*, *tastes*, *sounds*, and *smells*. On Locke's account secondary properties are powers of objects that are rooted in the primary properties. The most pressing question about the two kinds of properties is how (if at all) a precise and informative distinction can be drawn between them. Issues involving primary and secondary properties have been revived in the recent flurry of discussions of color.

Supervenient Properties

Supervenience is sometimes taken to be a relationship between two fragments of language (e.g., between psychological vocabulary and physical vocabulary), but it is increasingly taken to be a relationship between pairs of *families of properties*. To say that psychological properties supervene on physical properties, for example, is to say that, necessarily, everything that has any psychological properties also has physical properties and any two things that have exactly the same physical properties will have exactly the same psychological properties. There are no differences in psychological properties without some difference in physical properties.

There is no consensus as to how globally to construe supervenience claims. In the case of psychological and physical properties it seems too narrow to say that the non-relational psychological properties of a person supervene on her non-relational physical properties (too many important psychological properties involve relationships to things outside the organism for this to be right). And it seems unhelpfully narrow just to say that any two worlds that are just alike in their distributions of physical properties will be just alike in their distributions of psychological properties. But however we phrase the doctrine (and no doubt different versions will be useful for different purposes), supervenience is very naturally construed as a relation between pairs of families of properties.

In some cases it also seems plausible to think of the supervenient realm as linguistic and the supporting, subvenient realm in terms of properties: there can be no difference in truths in the upper realm (e.g., those employing psychological vocabulary) without a difference in properties (e.g., physical properties) at the lower level. But this hybrid approach has not received much attention.

Fictional Properties

We might arrive at a notion of *fictional properties* in the following way. In *Naming and Necessity* Kripke discusses what he calls ‘mythical species’ (1980, pp. 156-158). On Kripke's view, ‘Sherlock Holmes’ does not denote anyone, neither an actual individual or a merely possible one. The name fails to denote because all our descriptions of Holmes are essentially incomplete. They do not fully specify a unique person (not even a unique merely possible person), and so there are many different possible detectives who have all of the properties ascribed to Holmes but who differ in other ways. No one of them has any more claim on being Holmes than any of the rest, and so there are no counterfactual situations that could correctly be described as ones in which Holmes exists.

The properties ascribed to unicorns are similarly incomplete. There are many different possible species that have the properties we commonly ascribe to unicorns but which differ in other ways. No one of these species has any more claim on being unicorns than any of the rest, and so there are no counterfactual situations that could correctly be described as ones in which unicorns exists. It may also be the case, though Kripke doesn't discuss the matter, that notions like phlogiston and caloric fluid are similarly incomplete, and that there are no counterfactual situations that could correctly be described as ones in which anything has the property of *being phlogiston* or that of *being caloric fluid*. But even if this further claim isn't true, Kripke's example of unicorns suggests that there may be a distinction between actual properties and fictional properties. There are interesting philosophical issues about fictional characters, individuals like Holmes and Pegasus and the bride of Frankenstein, and there may be similarly interesting questions about fictional properties. Aside from Zalta (1983), however, little work has been done on this topic.

8 Formal Theories of Properties

In this section I will explain some of the most rudimentary ideas behind several recent formal theories of properties. The aim is to convey the intuitive flavor of this work, so I will proceed primarily by example

rather than with definitions and proofs (interested readers can find plenty of both in the works cited below).

There are several important choices that must be made in devising a formal theory of properties; they include the following:

1. Should the account be developed as a formal theory in a familiar logic (in the way set theories are now standardly axiomatized in standard first-order logic)? This was the approach in the early formal theories of Barcan-Marcus (1963) and Lemmon (1963), who used first-order modal logic (see Jubien, 1989, for a recent, more sophisticated (and non-modal) implementation of this approach). Alternatively, should a formal account of properties be developed as a richer "logic of properties" (e.g., Bealer, 1982; Zalta, 1983; Menzel, 1993; Swoyer, 1998)?
2. Should we employ a typed or an untyped conception of properties (the latter is much more flexible, but it must be handled with care to avoid paradox)?
3. Should we make provisions for complex predicates (or complex singular terms, or both) with something akin to a logical structure?
4. Should we require all of the expressions in the syntactic categories that can denote (or express) properties to do so or should we allow some or all of them not to?

Different choices are recommended by different conceptions of properties. For the sake of exposition I will begin with those choices that minimize the departures from familiar logical systems like non-modal first-order logic. This means a logic of properties that is typed, that does not include complex properties, and in which every predicate expresses a property. We will then see how to extend this approach in various ways.

8.1 A Bare-Bones Logic of Properties

Think of your favorite formulation of first-order logic with identity. Then just add *n*-place predicate variables (for all positive integers *n*) to its syntax and count any *n*-place predicate variable followed by *n* individual terms (i.e., by individual variables and constants) as a formula. Finally, allow *n*-place predicate variables to be bound by quantifiers containing an *n*-place predicate variable (in just the same way that individual variables can be), and count formulas with no free variables as sentences. For example, the expression ‘X^3abc’ is a formula consisting of the three-place predicate variable ‘X^3’ followed by the three individual terms ‘*a*’, ‘*b*’, and ‘*c*’; since the predicate variable is unbound, this formula is not a sentence. By contrast, ‘$(\exists X^2)X^2ab$’ is a sentence (here ‘$\exists$’ is the existential quantifier; this sentence says that there is at least one two-place relation that relates *a* to *b* or, more idiomatically, that *a* stands in some binary relation to *b*).

We interpret languages in this logical system over *Intensional Relational Structures* (*IRS*s). An *IRS* is an ordered triple:

$$\mathbf{I} = \langle \mathbf{D_I}, \mathbf{D_P}, \mathbf{ext} \rangle,$$

where $\mathbf{D_I}$ and $\mathbf{D_P}$ are non-empty, non-overlapping sets. On intended interpretations (which we will take for granted here) $\mathbf{D_I}$ is the domain of individuals and $\mathbf{D_P}$ is the domain of properties (including relations). $\mathbf{D_P}$ is in turn the union of an infinite number of non-overlapping, non-empty sets: ${}^1\mathbf{D_P}$ (the set of one-place properties), ${}^2\mathbf{D_P}$ (the set of two-place relations), ${}^3\mathbf{D_P}$ (the set of three-place relations), and so on for each positive integer *n*. Finally, **ext** is an extension assignment function; it assigns an extension to every property in the property domain $\mathbf{D_P}$ in accordance with the following rule:

> Where $\mathbf{P}^n$ is an *n*-place property, **ext** assigns $\mathbf{P}^n$ a (possibly empty) set of ordered *n*-tuples whose members are drawn from the individual domain $\mathbf{D_I}$

(we take an ordered one-tuple whose member is *a* to be *a* itself). Hence, the extension assignment assigns each one-place property a subset of the individual domain, each two-place property (i.e., each binary relation) a set of ordered pairs on individuals, and so on. In other words, it assigns exactly the same sorts of extensions to *n*-place properties that interpretations in standard first-order logic assign to *n*-place predicates.

Finally, a *model* or *interpretation* on an *IRS* interprets our formal language over it. It assigns a denotation to each individual constant (exactly as in standard first-order logic) and an *n*-place property to each *n*-place predicate (here we go beyond standard first-order logic). The fundamental idea is just that:

- An atomic sentence like 'F^1a', is *true in the model* just in case the object denoted by '*a*' is in the extension of the property denoted by 'F^1',
- An atomic sentence like 'F^2ab', is *true in the model* just in case the ordered pair containing the objects denoted by '*a*' and '*b*' (in that order) is in the extension of the (two-place) property denoted by 'F^2',
- ...and so on.[1]

As in first-order logic, we must add variable assignments (or their equivalents) to explain the workings of quantifiers. A variable assignment assigns an object of the appropriate sort to each variable of the language (it assigns individuals to individual variables, one-place properties to one-place predicate variables, two-place properties (binary relations) to two-place predicate variables, and so on). We then define satisfaction conditions for formulas (in a model and relative to a variable assignment) just as we do in standard first-order logic. Finally, a formula is true in a model just in case it contains no unbound variables and is satisfied by every variable assignment with respect to the model; a sentence is valid if is true in all models; and a set of sentences entails a sentence if that sentence is true in every model in which all of the sentences in the set are true. It is routine to extend such logics to higher orders, so that first-level properties can have second-level properties and stand in second-order relations, and so on up.

Untyped Variants

Philosophers, linguists, and computer scientists have increasingly chafed under the inflexible aspects of typed theories, and several recent accounts treat properties as individuals that are included in the range of the quantifiers for ordinary individuals. One way to accommodate this approach is to modify *IRS*s so that they include just a single domain that contains both properties and individuals. We then introduce *n*-place property-designating singular terms and require that an interpretation assign them denotations of the appropriate sort. For example, a one-place term of this sort might be used to represent the word ‘honesty’, and a three-place term of this sort might be used to represent ‘betweeness’. We can still allow predicates to denote properties (or we can introduce a new semantic relation, *expression*, which assigns properties to predicates). This allows something akin to self predication; if ‘*F*’ denotes (or expresses) the same property that the one-place singular term ‘*a*’ denotes, then ‘*Fa*’ will be true just in case the property denoted by ‘*a*’ is in the extension of itself. Quite intricate variations on this basic theme are possible; the most detailed is Bealer's (1982) first-order intensional logic that includes intensional abstracts among its singular terms.

Complex Terms and "Compound" Properties

It is much easier to deal with some features of natural language if we include complex predicates in our language and introduce a systematic way of interpreting them over *IRS*s. In the 1970s it occurred to several people that a rigorous formal system embodying this conception of properties could be constructed by investing the operations on linguistic expression in systems like Quine's predicate functor logic (1960) with an extra-linguistic, ontological status (e.g., Bealer, 1973, 1982; McMichael & Zalta, 1981; Leeds, 1978). Adding the machinery necessary to accommodate all of the complex predicates we might want is quite intricate (see Zalta, 1983 for a very readable account), and here I will just mention two examples to convey the general idea.

The first step is to introduce a variable-binding operator, λ, that allows us to construct complex predicates from open formulas. For example, we can apply ‘λ’ to the open formula, ‘*Rx* & *Sx*’ to form the one-place complex predicate ‘[λx (*Rx* & *Sx*)]’; if ‘*R*’ denotes *being red* and ‘*S*’ denotes *being square*, then this complex predicate denotes the compound, conjunctive property *being red and square* (a stilted, but sometimes useful rendering of this is ‘the property of being a thing that is both red and square’). Similarly, we can apply the operator to the open formula ‘∃*y*(*Lxy*)’ to form the one-place predicate ‘[λx ∃*y*(*Lxy*)]’; if ‘*L*’ stands for *loves*, this complex predicate denotes the compound property *loving someone* (whereas ‘[λy ∃*x*(*Lxy*)]’ would denote *being loved by someone*).[2]

Although the guiding ideas here are relatively straightforward, considerable delicacy is required to ensure that everything works out in precisely the right way. For example, an object should exemplify the conjunctive property denoted by ‘[λx (*Rx* & *Sx*)]’ if and only if it exemplifies both the properties *R* and *S*. And an object should exemplify the property denoted by ‘[λy ∃*x*(*Lxy*]’ just in case it loves some object (and just in case it does *not* exemplify the property denoted by ‘[λy ~∃*x*(*Lxy*]’). There are many systematic connections of this sort among complex predicates, compound properties, and the things that exemplify them, and some fairly heavy machinery is required to ensure that things work smoothly for properties of arbitrarily complexity. (Consider, for example, [λxyz ∃*w*(*Fxyw* & ~(*Gy* or ∃*v*(*Hzvy*)))].)

One way to make all of this work as it should is to add operations to *IRS*s that allow us to "build" compound properties up out of simpler ones. For example, to accommodate conjunctive properties we introduce an operation, **&**, that maps each pair of properties, **P** and **Q** (having the same number of argument places, though this restriction can be dropped at a slight cost of simplicity) to the conjunctive property **P & Q**. To ensure that things work properly we must *also* add a clause specifying how the **&**-operation interacts with the extension assignment function. In particular, we require that **ext**(**P & Q**) be the intersection of **ext**(**P**) and **ext**(**Q**).

By adding a handful of additional "property-building" operations (corresponding to the various connectives, quantifiers, operations on relations like conversion, and the syntactic operation of substitution), clauses specifying how each operation interlocks with the extension assignment, and a recursive definition of the denotations of predicates, we can ensure that complex predicates denote properties that behave as they should (Zalta, 1983 contains an elegant account of one way to do this; Swoyer, 1998 contains a slightly different approach in which assignments of denotations to complex predicates and assignments of extensions to the properties they denote are both treated as homomorphisms). With such a rich stock of properties we can add a comprehension schema to our logic which tells us that each condition (open formula) determines a property, i.e., there is at least one property that an ordered group of things has just in case the open formula is true of them.[3] It is also possible to add complex singular terms to the language; these are formal counterparts of nominalizations like *being poor but happy*. We can then set up the semantics so that these abstract singular terms denote compound properties.

Logic and the Empirical Conception of Properties

Some realists hold that it is an empirical question just which properties there are. On this view, there can be no logical or a priori existence conditions for properties. It is possible to formulate a very minimal logic (Swoyer, 1993; 1998) that fits nicely with this conception. Because it is so minimal, it has a philosophical neutrality that provides a framework in which various richer theories of properties (including ones with complex predicates) can be formulated, classified, and compared.

The Status of Formal Theories

There are two ways to view the kinds of formal systems described in this section (whether we call them ‘logics’ or not). We can view them as attempts to tell the One True Story about an abstract realm of properties (or the One True Story about the logical structure underlying a natural language like English or Hindi). On this construal the various systems are competitors. But it is also possible to view such formal systems in a more prosaic way, as abstract models that allow us to represent and reason about various phenomena involving properties (including various fragments of English). On this picture, such systems are similar in important respects to formal models in the sciences. They always involve simplifications and idealizations, and different models are useful for different purposes. Moreover, if a simpler model is enough to handle the phenomenon we are interested, it is overkill to employ more complex models even if they are available.

Future Directions

Various combinations of the features discussed in this section are possible. At this point several extensions also seem desirable, including allowing multigrade predicates like ‘had a knock-down, drag-out fight with each other’ and multigrade properties. It is also important to extend current accounts to deal with vagueness, and it would be gratifying to see them make contact with recent empirical work on concepts and categorization.

8.2 The Mereology of the Forms?

There has been some discussion of complex or structural properties in the recent literature, and certainly the metaphor of relation-building operations (like **&**) may suggest that some properties literally have parts. This idea can be traced back to Plato's later dialogues, where he speaks of one Form being part of another (*Sophist*, 257d). It has been defended recently by Armstrong (1978, pp. 36-39, pp. 67f) and Bigelow and Pargetter (1989). This view may have some plausibility for certain sorts of properties, e.g., conjunctive properties. But the general view (which these philosophers do not endorse) that properties literally have logical structures that mirror the syntactic structures of the complex predicates that denote them is less appealing. In the case of a negative property, for example, it would require us to think that the property *F* is somehow part of the negative property, *being a non-F*.

Structured Specifications vs. Specifications of Structure

On an alternative view, the appearance that some properties are literally structured is an artifact of our use of structured terms (complex, λ-predicates like ‘[λ*x* (*Rx* & *Sx*)]’) to denote them. But our use of structured terms and structural metaphors doesn't mean that the properties themselves are genuinely structured or that they literally have parts. There is a difference between *structured specifications*, which we do employ, and *specifications of structure*, which is another matter entirely. There *are* logical relations among properties; *being F* and *being not F* are inconsistent (in the sense that nothing could exemplify both at once); *being F and G* entails *being F* (in the sense that anything exemplifying the former property must also exemplify the latter). These logical relations do structure the realm of properties. This makes a structured specification a natural device for singling out a member of this structured realm of entities, since it identifies it by its place--its logical location--in that domain. But the role of the syntactic structure of a complex predicate is not to exhibit the internal structure of a property; it is to disclose that property's niche in the logical network of properties. We should add the cautionary note that a picture of compound properties needn't be a package deal. It is possible to argue that there are no compound properties, that there are some but not others (e.g., there might be conjunctive, but no disjunctive, properties), or that there is a multitude of them. Which view is correct? That is a philosophical question, and formal work alone cannot answer it.

Bibliography

- Achinstein, Peter (1974) "The Identity Conditions of Properties," *American Philosophical Quarterly*, 11; 257-275
- Armstrong, David (1978) *Universals and Scientific Realism, Vol. II. A Theory of Universals*. Cambridge: Cambridge University Press.
- ------ (1983) *What is a Law of Nature?* Cambridge: Cambridge University Press.
- ------ (1984) "Replies," in Radu Bogdan, ed. *D.M. Armstrong: Profiles*. Dordrecht: D. Reidel.
- ------ (1997) *A World of States of Affairs*. Cambridge: Cambridge University Press.
- Aaron, Richard (1967) *A Theory of Universals* 2nd/ed., Oxford: Clarendon Press.
- Barcan-Marcus, Ruth (1963) "Classes and Attributes in Extended Modal Systems," *Acta Philosophica Fennica*; 95-122.
- Bealer, George (1973) *A Theory of Qualities: A First-order Extensional Theory which includes a definition of analyticity, a one-level semantic method, and a derivation of intensional logic, set theory and modal logic*. Ph.D. Dissertations, University of California, Berkeley.
- ------ (1982) *Quality and Concept*. Oxford: Clarendon Press.
- ------ (1994) "Property Theory: The Type-free Approach vs. the Church Approach," *Journal of Philosophical Logic*, 23; 139-171.
- Benacerraf, Paul (1965) "What Numbers Could Not Be," *Philosophical Review*, 74; 47-73.
- ------ (1973) "Mathematical Truth," *Journal of Philosophy*, 70; 661-679.
- Bergmann, Gustav (1968) "Elementarism," Ch. 6 of *Meaning and Existence*. Madison: University of Wisconsin Press.
- Bigelow, John and Pargetter, Robert (1989) "A Theory of Structural Universals," *Australasian Journal of Philosophy*, 67; 1-11.
- ------ (1990) *Science and Necessity*. Cambridge: Cambridge University Press.
- Broad, C. D. (1933) *Examination of McTaggart's Philosophy: Volume I*. Cambridge: Cambridge University Press.
- Butchvarov, Panayot (1966) *Resemblance and Identity*. Bloomington: University of Indiana Press.
- Cartwright, Nancy (1983) *How the Laws of Physics Lie*. Oxford: Clarendon Press.
- ------ (1989) *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.
- Cherniss, H. F. (1936) "The Philosophical Economy of Plato's Theory of Ideas," *American Journal of Philology*, 57; 445-456.
- Chierchia, Gennaro & Turner, Raymond (1988) "Semantics and Property Theory," *Linguistics and Philosophy*, 11; 261-302
- Cocchiarella, Nino (1986) *Logical Investigations of Predication Theory and the Problem of Universals*. Napoli: Bibliopolis.
- Dretske, Fred (1977) "Laws of Nature," *Philosophy of Science* 44: 248-268.
- Fales, Evan (1990) *Causation and Universals*. New York: Routledge.
- Field, Hartry (1980) *Science without Numbers: A Defense of Nominalism*. Princeton: Princeton University Press.
- Fisk, Milton (1972) "Relatedness without Relations," *Noûs*, 6; 139-151.
- Forrest, Peter (1986) "Ways Worlds Could Be," *Australasian Journal of Philosophy*, 64; 15-24.
- Grossmann, R (1983)*The Categorial Structure of the World*. Bloomington: University of Indiana Press.
- Hochberg, Herbert (1968) "Nominalism, Platonism, and Being True of," *Noûs*, 2; 413-419.

- Jubien, Michael (1989) "On Properties and Property Theory," in Gennaro Chierchia, Barbara Partee and Raymond Turner, eds., *Properties, Types and Meanings, Vol I: Foundational Issues*. Dordrecht: Kluwer; 159-175.
- Kripke, Saul (1980) *Naming and Necessity*. Cambridge, Ma.: Harvard University Press.
- Leeds, Stephen (1978) "Quine on Properties and Meanings," *Southwestern Journal of Philosophy*, 9; 97-108.
- Lemmon, E. J. (1963) "A Theory of Attributes Based on Modal Logic," *Acta Philosophica Fennica*; 95-122.
- Lewis, David (1986) *On the Plurality of Worlds*. Oxford: Blackwell.
- Linsky, Bernard & Zalta, Edward (1995) "Naturalized Platonism vs. Platonized Naturalism," *Journal of Philosophy* 92; 525-555.
- Loux, Michael J. (1972) "Recent Work in Ontology," *American Philosophical Quarterly* 9; 119-138.
- Margolis, Eric & Laurence, Stephen eds., (1999) *Concepts: Core Readings*. Cambridge: MIT Press.
- McMichael, Alan & Zalta, Edward (1981) "An Alternative Theory of Nonexistent Objects," *Journal of Philosophical Logic*, 9; 297-313.
- Mellor, D. H. (1991) *Matters of Metaphysics*. Cambridge: Cambridge University Press.
- Menzel, Chris (1986) "A Complete Type-free Second Order Logic of Properties, Relations, and Propositions," Center for the Study and Language and Information, Technical Report #CSLI-86-40, Stanford University.
- ------ (1993) "The Proper Treatment of Predication in Fine-grained Intensional Logic," *Philosophical Perspectives*, 7; 61-87.
- Montague, Richard (1974) *Formal Philosophy: Selected Papers of Richard Montague*. New Haven: Yale University Press.
- Mundy, Brent (1987) "The Metaphysics of Quantity," *Philosophical Studies*, 51; 29-54.
- ------ (1990) "Elementary Categorial Logic, Predicates of Variable Degree, and Theory of Quantity," *Journal of Philosophical Logic*, 18; 115-140.
- Parsons, Terence (1986) "Why Frege Should not have Said 'The Concept *Horse* is not a Concept'," *History of Philosophy Quarterly*, 3; 449-465.
- Pollard, Stephen & Martin, Norman (1986) "Mathematics for Property Theorists," *Philosophical Studies*, 49; 177-186.
- Quine, W. V. O. (1961) "On What There is," in *From a Logical Point of View*. 2nd/ed. N.Y: Harper and Row.
- ------ (1960) "Variables Explained Away," *Proceedings of the American Philosophical Society*, 104; 343-347.
- Quinton, Anthony (1973) *The Nature of Things*. London: Routledge & Kegan Paul.
- Resnik, Michael (1997) *Mathematics as a Science of Patterns*. Oxford: Clarendon Press.
- Rosch, Eleanor (1978) "Principles of Categorization," in E. Rosch & B. Lloyd, eds. *Cognition and Categorization*. Hillsdale, N.J.: Lawrence Erlbaum; 27-48.
- Russell, Bertrand (1912) *The Problems of Philosophy*. London: Home University Library.
- ------ (1948) *Human Knowledge: Its Scope and Limits*. London: Allen and Unwin.

- Shoemaker, Sydney (1984) *Identity, Cause, and Mind: Philosophical Essays*. Cambridge: Cambridge University Press.
- Sober, Elliot (1982) "Why Logically Equivalent Predicates may Pick out Different Properties, *American Philosophical Quarterly*, 19; 183-189.
- Swoyer, Chris (1983) "Realism and Explanation," *Philosophical Inquiry*, 5; 14-28.
- ------ (1987) "The Metaphysics of Measurement," in *Measurement, Realism and Objectivity*, ed. John Forge. Dordrecht: D. Reidel; 235-290.
- ------ (1993) "Logic and the Empirical Conception of Properties," *Philosophical Topics*, 21; 199-231.
- ------ (1996) "Theories of Properties: From Plenitude to Paucity, *Philosophical Perspectives*, 10; 243-264.
- ------ (1998) "Complex Predicates and Theories of Properties and Relations," *Journal of Philosophical Logic*, 27: 295-325.
- ------ (1999) "How Ontology Might be Possible: Explanation and Inference in Metaphysics," *Midwest Studies in Philosophy*, P. A. French & H. K. Wettstein, eds. 23: 100-131.
- Strawson, P. F. (1959) *Individuals: An Essay in Descriptive Metaphysics*. Garden City, N.Y.: Doubleday.
- Tooley, Michael (1977) "The Nature of Laws," *Canadian Journal of Philosophy* 7: 667-698.
- ------ (1987) *Causation*. Oxford: Oxford University Press.
- Whitehead, Alfred North, & Bertrand Russell (1910, 1912, 1913) Principia Mathematica, 3 vols, Cambridge:
  Cambridge University Press. Second edition, 1925 (Vol. 1), 1927 (Vols 2, 3). Abridged as Principia Mathematica to *56,
  Cambridge: Cambridge University Press, 1962.
- Wittgenstein, Ludwig (1991) *Remarks on the Foundations of Mathematics*. Revised Edition. Cambridge: MIT Press.
- van Fraassen, Bas (1989) *Laws and Symmetry*. Oxford: Clarendon Press.
- Zalta, Edward (1983) *Abstract Objects: An Introduction to Axiomatic Metaphysics*. Dordrecht: D. Reidel.
- ------ (1988) *Intensional Logic and the Metaphysics of Intentionality*. Cambridge: MIT Press.
- ------ (1999) "Natural Numbers and Natural Cardinals as Abstract Objects: A Partial Reconstruction of Frege's Grundgesetze in Object Theory," *Journal of Philosophical Logic* 28/6: 619-660
- ------ (2000) "Neologicism? An Ontological Reduction of Mathematics to Metaphysics," *Erkenntnis* 53/1-2: 219-265

Other Internet Resources

- The Metaphysics Research Lab
- Internet Encyclopedia of Philosophy: Universals

Related Entries

Frege, Gottlob: logic, theorem, and foundations for arithmetic | propositional function | Russell's paradox | tropes | universals: the medieval problem of | vagueness

Chris Swoyer
cswoyer@ou.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 23, 1999
Content last modified: December 17, 2000

Stanford Encyclopedia of Philosophy
Supplement to Properties

Detailed Table of Contents (to the subsection level)

To Beginning of Entry

More Detailed Table of Contents (to subsubsection level)

To Beginning of Entry

Stanford Encyclopedia of Philosophy
Supplement to Properties

Uses of Properties in the Philosophy of Mathematics

1. Logicist Identificationism

The first detailed version of identificationism was developed by Gottlob Frege, in which numbers were identified with sets. (Frege actually identified numbers with the extensions of second-order concepts, but since these are extensional creatures, it isn't wildly anachronistic to think of them as sets. See the entry on Frege's logic, theorem and foundations for arithmentic.) But Frege made a further claim: these extensions, and hence the numbers, are *logical* entities, ones whose very existence is guaranteed by logic alone. Hence, truths of number theory are really just logical truths. This thesis, which came to be known as *logicism*, offered the hope of an extremely secure foundation for arithmetic that fulfilled wishes 4 (telling us how arithmetical truths can be necessarily true) and 5 (telling us how we have *a priori* knowledge in mathematics). After all, logical truths are surely necessarily true if anything is, and if anything can be known *a priori*, they surely can be. Indeed, they are usually thought to be analytic, and there is supposed to be little puzzle about how we can know such truths *a priori*; we just analyze the meanings of terms.

In their monumental, three-volume work *Principia Mathematica*, Whitehead and Russell develop logicism in a slightly different way. They identify numbers with propositional functions which, owing to unclarity of exposition, hover between properties, on the one hand, and linguistic expressions (open sentences, predicates roughly), on the other. But they quantify over propositional functions in a way that treats them as properties, so this early version of logicism actually identifies numbers with properties rather than with sets. Whitehead and Russell don't make use of the intensionality of properties, however, and they could as well have used sets. More recently several philosophers, particularly Bealer (1982, ch 5-6; cf. Menzel, 1986) have argued that logic alone guarantees the existence of enough properties with the right structure to play the role of the natural numbers. This requires a conception of logic much richer than many traditional ones; among other things, logic alone must ensure the existence of infinitely many things (such systems, which are quite intricate, are beyond the scope of the present article, but a few of the basic ideas are explained in §8).

Champions of a logicism based on properties can simply adapt earlier identificationist's explanations of truth, objectivity, and logical form, substituting properties for the sets employed in the earlier accounts. Moreover, they can hold that the existence of properties, at least those identified with the numbers, is guaranteed by logic. And since logical truths are necessarily true, these properties necessarily exist and they necessarily stand in the (mathematically relevant) relations that they do, which explains the modal

status of arithmetic claims. None of this tells us how mathematical knowledge is possible; any account of this will require substantive auxiliary hypotheses about human cognitive faculties.

This brief account of property-theoretic versions of logicism can't do justice to their details, but it tells us enough to spot several likely objections to them. They face problems peculiar to logicism as well as more general problems that threaten most versions of identificationism.

Return to Main Entry

2. Set-theoretic Identificationism

These steps are familiar from set-theoretic versions of identificationism in which numbers are identified with sets. For example, we might identify zero with the empty set, Ø, and for each x in the domain, the successor of x, $s(x)$, could be identified with $\{x\}$ (as in Zermelo's account), or with the union of x and $\{x\}$ (as in von Neumann's account), or with the result of applying some other operation of the appropriate sort to x (when there is one way to do this, there are lots of ways). Classical versions of identificationism begin with a formal theory of the objects we take to be numbers, and the task is to show that arithmetical truths match up with truths of this theory and that arithmetical falsehoods match up with its falsehoods. When the proposed identification is with sets, this involves deriving the set-theoretic translations of our favorite set of axioms for number theory in some version of set theory.

Return to Main Entry

3. Logicism and Logic

The exact character of logic is elusive; indeed, there is some reason to think that, like many other notions, it is vague around the edges. But one traditional hallmark of logic is that it is topic neutral; it applies to everything indifferently; it doesn't deal explicitly with specific sorts of things like planets or dogs or the number 16. The notion of a property is very general, however, since it is plausible to suppose that absolutely anything can exemplify a property, so there is some room to argue that logic can guarantee the existence of properties without violating its topic-neutrality.

A second traditional hallmark of logic is that it does not make existence commitments, at least not any specific ones. It is true that theorems of our most familiar logical system, first-order logic, are only true in circumstances in which at least one thing exists. But this one thing can be *any* sort of thing at all, and this system is sometimes criticized for assuming even this much (this is one reason why some philosophers prefer free logic, which doesn't make even this minimal commitment). In the end, though, it is less important to decide whether rich and powerful systems count as logic than to see whether they have the sorts of features that make logicism attractive in the first place.

It may seem plausible to view theorems of classical sentential logic as analytic, true in virtue of the

meanings of the logical constants they contain. But it is generally agreed that existence claims are not analytic (this is the root of the problem with the ontological argument). So the nice, intuitive view that mathematical knowledge is no great mystery (it's simply knowledge of analytic logical truths) seems less plausible once the notion of logic is stretched to become so powerful. Moreover, this stretching doesn't end with the requirement that our logic of properties guarantee the existence of infinitely many things. First-order logic is not strong enough to characterize the natural numbers even up to isomorphism, and logics that are do not have recursively enumerable sets of logical truths (by Gödel's theorem). Hence, such systems are epistemically untractable in the sense that their truths far outrun our ability to track (prove) them.

In short, it may well be a merely verbal dispute just which formal systems merit the honorific title of ‘logic’. The substantive point here is that if logicism requires a logic that is so lavish in its existence commitments and so computationally (and hence epistemologically) unruly, then the appeal that the slogan "truths of arithmetic are just truths of logic" had when we had a kinder and gentler vision of logic in mind begins to fade. Indeed, as we will now see, logicist identificationism does not turn out to be so very terribly different from its non-logicist counterparts.

4. Non-logicist identificationism

In view of the heavy existence commitments of number theory, most philosophers came to doubt that "reductions" of arithmetic to set theory really counted as reductions to logic (and parallel reasoning would lead them to doubt that using a logic of properties rather than set theory changes this). Instead, the reducing theory, that of sets, came to be seen as an independent formal theory with its own proprietary existence commitments that was best formulated in, but was not identical with, some logical system (as the decades rolled by this system was increasingly first-order logic).

The claim that numbers are identical with sets is really all we need to mobilize the logicist's accounts of mathematical truth, objectivity, and logical form. Furthermore, since the sets identified with numbers are all pure sets (they don't have any members that aren't themselves sets), it can be argued that they exist necessarily, which explains the modal status of mathematical claims. The important point here is that a property theorist can write a set of axioms for property theory that parallel the axioms of standard set theories (save for replacing the axiom of extensionality with some other identity condition, perhaps omitting the axiom of foundations, and making other minor emendations to adapt the ideas better to properties; e.g., Jubien, 1989; cf. Pollard and Martin, 1986). The resulting axioms are not viewed as part of logic, but as a characterization of an independently existing realm of properties. And once this is done, the property theorist can adapt the explanations given by those non-logicists who identify numbers with sets to explain all of the items on our list except for the one involving mathematical knowledge.

Chris Swoyer
cswoyer@ou.edu

First published: December 15, 2000
Content last modified: December 15, 2000

Gottlob Frege

Gottlob Frege (b. 1848, d. 1925) was a German mathematician, logician, and philosopher who worked at the University of Jena. Frege essentially recreated the discipline of logic by constructing the first ‘predicate calculus’. In this calculus, Frege developed a new analysis of atomic and quantified statements and formalized the notion of a ‘proof’ in terms that are still accepted today. Frege then demonstrated that one could use this calculus to resolve theoretical mathematical statements in terms of simpler logical and mathematical notions. One of the axioms that Frege later adopted for his system, in the attempt to derive significant parts of mathematics from logic, proved to be inconsistent. Nevertheless, his definitions (of the *predecessor* relation and of the concept of *natural number*) and methods (for deriving the axioms of number theory) were insightful and constituted a significant advance. To ground his views about the relationship of logic and mathematics, Frege conceived a comprehensive philosophy of language that many philosophers still find insightful. However, his lifelong project, of showing that mathematics was reducible to logic, was not successful.

Frege's Life

- 1848, born November 8 in Wismar (Mecklenburg-Schwerin)

- 1869, entered the University of Jena
- 1871, entered the University of Göttingen
- 1873, awarded Ph. D. in Mathematics (Geometry), University of Göttingen
- 1874, earned a Habilitation in Mathematics, University of Jena
- 1874, became Privatdozent, University of Jena
- 1879, became Professor Extraordinarius, University of Jena
- 1896, became ordentlicher Honorarprofessor, University of Jena
- 1917, retired from the University of Jena
- 1925, died July 26 in Bad Kleinen (now in Mecklenburg-Vorpommern)

Frege's Advances in Logic

Frege virtually founded the modern discipline of mathematical logic. He forever changed the way philosophers and mathematical logicians think about the predicate calculus, the analysis of simple sentences and quantifier phrases, proofs, the foundations of mathematics, definitions, and the ‘natural numbers’.

The Predicate Calculus

In an attempt to realize Leibniz's ideas for a language of thought and a rational calculus, Frege developed a formal notation for regimenting thought and reasoning (see his *Begriffsschrift*). Though we no longer use his notation, Frege's in effect developed the first predicate calculus. A predicate calculus is a formal system with two components: a formal language and a logic. The formal language Frege designed was capable of: (a) expressing predicational statements of the form ‘*x* falls under the concept *F*’ and ‘*x* bears relation *R* to *y*’, etc., (b) expressing complex statements such as ‘it is not the case that ...’ and ‘if ... then ...’, and (c) expressing ‘quantified’ statements of the form ‘Some *x* is such that ...*x*...’ and ‘Every *x* is such that ...*x*...’. The logic of Frege's calculus was a set of rules that govern when some statements of the language may be correctly inferred from others.

Frege's system was powerful enough to resolve the essential logic of mathematical reasoning. That was partly due to the fact that his predicate calculus was a ‘second-order’ predicate calculus, allowing statements of the form ‘Some concept *F* is such that ...*F*...’ and ‘Every concept *F* is such that ...*F*...’. However, the most important insight underlying Frege's calculus was his ‘function-argument’ analysis of sentences. This freed him from the limitations of the ‘subject-predicate’ analysis of sentences that formed the basis of Aristotelian logic and it made it possible for him him to develop a general treatment of quantification.

The Analysis of Atomic Sentences and Quantifier Phrases

In traditional Aristotelian logic, the subject of a sentence and the direct object of a verb are not on a logical par. The rules governing the inferences between statements with different but related subject

terms are different from the rules governing the inferences between statements with different but related verb complements. For example, in Aristotelian logic, the rule which permits the valid inference from ‘John loves Mary’ to ‘Something loves Mary’ is different from the rule which permits the valid inference from ‘John loves Mary’ to ‘John loves something’. The rule governing the first inference is a rule which applies only to the subject terms ‘John’ and ‘Something’. The rule governing the second inference applies only to the transitive verb complements ‘Mary’ and ‘something’. In Aristotelian logic, these inferences have nothing in common.

In Frege's logic, a single rule governs both the inference from ‘John loves Mary’ to ‘Something loves Mary’ and the inference from ‘John loves Mary’ to ‘John loves something’. This was made possible by Frege's analysis of atomic and quantified sentences. Frege took intransitive verb phrases such as ‘is happy’ to be functions of one variable (‘*x* is happy’), and resolved the sentence "John is happy" in terms of the application of the function denoted by ‘is happy’ to the argument denoted by ‘John’. In addition, Frege took the verb phrase ‘loves’ to be a function of two variables (‘*x* loves *y*’) and resolved the sentence ‘John loves Mary’ as the application of the function denoted by ‘*x* loves *y*’ to the objects denoted by ‘John’ and ‘Mary’ respectively. In effect, Frege saw no distinction between the subject ‘John’ and the direct object ‘Mary’. What is logically important is that ‘loves’ denotes a function of 2 arguments, that ‘gives’ denotes a function of 3 arguments (*x* gives *y* to *z*), that ‘bought’ denotes a function of 4 arguments (*x* bought *y* from *z* for amount *u*), etc.

This analysis allowed Frege to develop a more systematic treatment of quantification than that offered by Aristotelian logic. No matter whether the quantified expression ‘something’ appears within a subject ("Something loves Mary") or within a predicate ("John loves something"), it is to be resolved in the same way. In effect, Frege treated quantified expressions as variable-binding operators. The variable-binding operator ‘some *x* is such that’ can bind the variable ‘*x*’ in the expression ‘*x* loves Mary’ as well as the variable ‘*x*’ in the expression ‘John loves *x*’. Thus, Frege analyzed the above inferences in the following general way:

- John loves Mary. Therefore, some *x* is such that *x* loves Mary.
- John loves Mary. Therefore, some *x* is such that John loves *x*.

Both inferences are instances of a single valid inference rule.

Proof

As part of his predicate calculus, Frege developed a strict definition of a ‘proof’. In essence, he defined a proof to be any finite sequence of well-formed statements such that each statement in the sequence either is an axiom or follows from previous members by a valid rule of inference. A proof of the statement B from the premises $A_1,...,A_n$ is any finite sequence of statements (with B the final statement in the sequence) such that each member of the sequence: (a) is one of the premises $A_1,...,A_n$, or (b) is an axiom, or (c) follows from previous members of the sequence by a rule of inference. This is essentially the definition of a proof that logicians still use today.

Definition

Frege was extremely careful about the proper description and definition of logical and mathematical concepts. He developed powerful and insightful criticisms of mathematical work which did not meet his standards for clarity. For example, he criticized mathematicians who defined a variable to be a number that varies rather than an expression of language which can vary as to which determinate number it refers to. And he criticized those mathematicians who developed ‘piecemeal’ definitions or ‘creative’ definitions. In the *Grundgesetze* (Band II, Sections 56-67) Frege criticized the practice of defining a concept on a given range of objects and later redefining the concept on a wider, more inclusive range of concepts. Frequently, this ‘piecemeal’ style of definition led to conflict, since the redefined concept did not always reduce to the original concept when one restricts the range to the original class of objects. In that same work (Band II, Sections 139-147), Frege criticized the mathematical practise of introducing notation to name (unique) entities without first proving that there exist (unique) such entities. He pointed out that such ‘creative definitions’ were simply unjustified.

A Foundation for Mathematics

Frege attempted to construct a foundation for mathematics. His most comprehensive logical system was developed in his landmark work *Grundgesetze der Arithmetik*, in which he attempted to validate the philosophical doctrine known as *logicism*, i.e., the idea that mathematical concepts can be defined in terms of purely logical concepts and that mathematical axioms can be derived from the laws of logic alone. Unfortunately, Frege employed a principle in the *Grundgesetze* (Basic Law V) which turned out to be subject to Russell's Paradox. This paradox caused him to question the truth of logicism, and few philosophers today believe that mathematics can be reduced to logic. Mathematics seems to require some non-logical notions (such set membership) and some non-logical axioms (such as the non-logical axioms of set theory). Despite the fact that a contradiction invalidated a part of his system, the intricate theoretical web of definitions and proofs developed in the *Grundgesetze* produced a conceptual framework for mathematical logic that was nothing short of revolutionary. There is no doubt that the logical system and maze of definitions developed by Bertrand Russell and Alfred North Whitehead in *Principia Mathematica* owe a huge debt to the work found in Frege's *Grundgesetze*.

The Natural Numbers

In his seminal work *Die Grundlagen der Arithmetik*, Frege successfully defined the notion of a ‘cardinal number’ in terms of the primitive notion of an *extension* or *set*. The insight behind the definition is that a statement of cardinal number such as ‘There are *n* *F*-things’ predicates a higher-order concept of the concept *F*, namely, that it is a concept under which *n* things fall. Frege simply defines the (cardinal) number of the concept *F* (i.e., the number of *F*s) as the extension of the concept *being a concept equinumerous to F*. On this definition, the number of planets is identified as the extension of the concept *being a concept equinumerous to the concept of being a planet*. In other words, the number of planets is

an extension (or set) which contains all those concepts which, like the concept *being a planet*, are exemplified by nine objects.

Frege defined the concept of *natural number* by defining, for every relation *xRy*, the general concept '*x* is an ancestor of *y* in the R-series' (this new relation is called 'the ancestral of the relation R'). The ancestral of a relation R was first defined in Frege's *Begriffsschrift*. The intuitive idea is easily grasped if we consider the relation *x* is the father of *y*. Suppose that *a* is the father of *b*, that *b* is the father of *c*, and that *c* is the father of *d*. Then Frege's definition of '*x* is an ancestor of *y* in the fatherhood-series' ensured that *a* is an ancestor of *b*, *c*, and *d*, that *b* is an ancestor of *c* and *d*, and that *c* is an ancestor of *d*.

More generally, if given a series of facts of the form *aRb*, *bRc*, *cRd*, and so on, Frege showed how to define the relation *x is an ancestor of y in the R-series* (this is the ancestral of the relation R). To exploit this definition in the case of natural numbers, Frege had to define both the relation *x precedes y* and the ancestral of this relation, namely, *x is an ancestor of y in the predecessor-series*. He first defined the relational concept *x precedes y* as follows:

> *x precedes y* iff there is a concept *F* and an object *z* such that:
>
> - *z* falls under *F*,
> - *y* is the (cardinal) number of the concept *F*, and
> - *x* is the (cardinal) number of the concept *object other than z falling under F*

In the notation of the second-order predicate calculus, Frege's definition becomes:

$$Precedes(x,y) \quad =_{df} \quad \exists F \exists z (Fz \;\&\; y = \#F \;\&\; x = \#[\lambda u \; Fu \;\&\; u \neq z])$$

To see the intuitive idea behind this definition, consider how the definition is satisfied in the case of the number 1 preceding the number 2: there is a concept *F* (e.g., let *F*=the concept *being an author of Principia Mathematica*) and an object *z* (e.g. let *z*=Alfred North Whitehead) such that:

- Whitehead falls under the concept *being an author of Principia Mathematica*,
- *2* is the (cardinal) number of the concept *being an author of Principia Mathematica*, and
- *1* is the (cardinal) number of the concept *object other than Whitehead which falls under the concept being an author of Principia Mathematica*

Note that the last conjunct is true because there is exactly 1 object (namely, Bertrand Russell) which falls under the concept *object other than Whitehead which falls under the concept being an author of Principia Mathematica*.

Given this definition of *precedes,* Frege then defined the ancestral of this relation, namely, *x is an ancestor of y in the predecessor-series*. So, for example, if 10 precedes 11 and 11 precedes 12, it follows that 10 is an ancestor of 12 in the predecessor-series. Note, however, that although 10 is an ancestor of

12, 10 does not precede 12, for the notion of *precedes* is that of *strictly* precedes. Note also that by defining the ancestral of the precedence relation, Frege had in effect defined $x < y$.

Frege then defined the number 0 as the (cardinal) number of the concept *being an object not identical with itself*. The idea here is that nothing fails to be self-identical, so nothing falls under this concept. The number 0 is therefore identified with the extension of all concepts which fail to be exemplified.

Finally, Frege defined:

> *x is a natural number* iff either x=0 or 0 is an ancestor of x in the predecessor series

In other words, a natural number is any member of the predecessor series beginning with 0.

Using this definition, Frege derived many important theorems of number theory. It was recently shown by R. Heck [1993] that, despite the logical inconsistency in the system of his *Grundgesetze*, Frege validly derived the Dedekind/Peano Axioms for number theory from a powerful and consistent principle now known as Hume's Principle ("The number of Fs is equal to the number of Gs if and only if there is a one-to-one correspondence between the Fs and the Gs"). Although Frege used his inconsistent axiom Basic Law V to establish Hume's Principle, once Hume's Principle was established, Frege validly derived the Dedekind/Peano axioms without making any further essential appeals to Basic Law V. Following the lead of George Boolos, philosophers today call derivation of the Dedekind/Peano Axioms from Hume's Principle ‘Frege's Theorem’. The proof of Frege's Theorem was a *tour de force* which involved some of the most beautiful, subtle, and complex logical reasoning that had ever been devised. For a comprehensive introduction to the logic of Frege's Theorem, see the entry Frege's logic, theorem, and foundations for arithmetic.

Frege's Ontology and Philosophy of Language

While pursuing his investigations into mathematics and logic (and quite possibly, in order to ground those investigations), Frege was led to develop not only an ontology of *functions* and *objects* but also a philosophy of language. In his ontology, Frege asserted the existence of two special objects, namely, the truth values The True and The False and defined *concepts* as a special kind of function, namely, any function that mapped objects to truth values. He also suggested that *existence* is not a property of objects but rather of concepts. Frege is most well-known among philosophers, however, for the way these ideas were applied and extended in his philosophy of language. His seminal paper in this field ‘Über Sinn und Bedeutung’ (‘On Sense and Denotation’) is now a classic. In this paper, Frege considered two puzzles about language and noticed, in each case, that one cannot account for the meaningfulness or logical behavior of certain sentences simply on the basis of the denotations of the terms (names and descriptions) in the sentence. One puzzle concerned identity statements and the other concerned sentences with relative clauses such as propositional attitude reports. To solve these puzzles, Frege suggested that the terms of a language have both a sense and a denotation (i.e., that at least two semantic relations are

required to explain the significance or meaning of the terms of a language). This idea has inspired research in the field for over a century.

Frege's Ontology

In Frege's ontology, functions and objects were rigorously distinguished as two fundamentally different kinds of entity. Functions are the kind of thing that take objects as arguments and map those arguments to a value. Frege did not limit examples of functions to mathematical functions such as $x + 3$. He allowed the variable x to range over any object, and so *father of x* is a genuine example of a function---it maps certain biological offspring to their fathers and maps everything else to The False. Frege associates with every function, a *course-of-values*. The course-of-values of a function explicitly indicates the value of the function for each object that is supplied as an argument. In addition, Frege believed that there are two distinguished objects, namely, the truth value The True and the truth-value The False. Those functions which map objects to truth values are called *concepts*. For example, not only is the mathematical function $x + 3=5$ a concept (this concept maps the number 2 to The True and everything else to The False), but so is the function *x is happy* (which maps anything that is happy to The True and everything else to The False). Frege defines the *extension* of a concept to contain just those objects which the concept maps to The True (as indicated by the course-of-values associated with the concept).

Frege suggested that *existence* is not a property of objects but rather of concepts: it is the property a concept has just in case it has an non-empty extension (i.e., just in case it maps some object to The True). So the fact that the extension of the concept *martian* is empty underlies the ordinary claim "Martians don't exist." Frege therefore took *existence* to be a ‘second-level’ concept.

Frege's Puzzles

Frege's Puzzle About Identity Statements. Here are some examples of identity statements:

> 117 + 136=253.
> The morning star is identical to the evening star.
> Mark Twain is Samuel Clemens.
> Bill is Debbie's father.

Frege believed that these statements all have the form "$a=b$", where ‘*a* ’ and ‘*b* ’ are either names or descriptions that *denote* individuals. He naturally assumed that a sentence of the form "$a=b$" is true if and only if the object denoted by ‘*a* ’ is the same as the object denoted by ‘*b* ’. For example, "117 + 136=253" is true if and only if ‘117 + 236’ and ‘253’ denote the same number. And "Mark Twain is Samuel Clemens" is true if and only if ‘Mark Twain’ and ‘Samuel Clemens’ denote the same person. So the truth of "$a=b$" requires that the expressions flanking the identity sign denote the same object.

But Frege noticed that on this account of truth, the truth conditions for "$a=b$" are no different from the truth conditions for "$a=a$". For example, the truth conditions for "Mark Twain=Mark Twain" are the

same as those for "Mark Twain=Samuel Clemens"; not only do the names flanking the identity sign denote the same object in each case, but the object is the same between the two cases. The problem is that the cognitive significance (or meaning) of the two sentences differ. We can learn that "Mark Twain=Mark Twain" is true simply by inspecting it; but we can't learn the truth of "Mark Twain=Samuel Clemens" simply by inspecting it. Similarly, whereas you can learn that "117 + 136=117 + 136" and "the morning star is identical to the morning star" are true simply by inspection, you can't learn the truth of "117 + 136=253" and "the morning star is identical to the evening star" simply by inspection. In the latter cases, you have to do some arithmetical work or astronomical investigation to learn the truth of these identity claims.

So the puzzle Frege discovered is: if we cannot appeal to a difference in denotation of the terms flanking the identity sign, how do we explain the difference in cognitive significance between "*a*=*b*" and "*a*=*a*"?

Frege's Puzzle About Propositional Attitude Reports. Frege is generally credited with identifying the following puzzle about propositional attitude reports, even though he didn't quite describe the puzzle in the terms used below. A propositional attitude is a psychological relation between a person and a proposition. Belief, desire, intention, discovery, knowledge, etc., are all psychological relationships between persons, on the one hand, and propositions, on the other. When we report the propositional attitudes of others, these reports all have a similar logical form:

> *x* believes that *p*
> *x* desires that *p*
> *x* intends that *p*
> *x* discovered that *p*
> *x* knows that *p*

If we replace the variable '*x* ' by the name of a person and replace the variable '*p* ' with a sentence that describes the propositional object of their attitude, we get specific attitude reports. So by replacing '*x* ' by 'John' and '*p* ' by "Mark Twain wrote *Huckleberry Finn*" in the first example, the result would be the following specific belief report:

> John believes that Mark Twain wrote *Huckleberry Finn*.

To see the problem posed by the analysis of propositional attitude reports, consider what appears to be a simple principle of reasoning, namely, the Principle of Substitution. If a name, say *n*, appears in a true sentence S, and the identity sentence *n*=*m* is true, then the Principle of Substitution tells us that the substitution of the name *m* for the name *n* in S does not affect the truth of S. For example, let S be the true sentence "Mark Twain was an author", let *n* be the name 'Mark Twain', and let *m* be the name 'Samuel Clemens'. Then since the identity sentence "Mark Twain=Samuel Clemens" is true, we can substitute 'Samuel Clemens' for 'Mark Twain' without affecting the truth of the sentence. And indeed, the resulting sentence "Samuel Clemens was an author" is true. In other words, the following argument is valid:

Mark Twain was an author.
Mark Twain=Samuel Clemens.
Therefore, Samuel Clemens was an author.

Similarly, the following argument is valid.

4 > 3
4=8/2
Therefore, 8/2 > 3

In general, then, the Principle of Substitution seems to take the following form, where S is a sentence, *n* and *m* are names, and S(*n*) differs from S(*m*) only by the fact that at least one occurrence of *m* replaces *n*:

From S(*n*) and *n*=*m*, infer S(*m*)

This principle seems to capture the idea that if we say something true about an object, then even if we change the name by which we refer to that object, we should still be saying something true about that object.

But Frege, in effect, noticed the following counterexample to the Principle of Substitution. Consider the following argument:

John believes that Mark Twain wrote *Huckleberry Finn* .
Mark Twain=Samuel Clemens.
Therefore, John believes that Samuel Clemens wrote *Huckleberry Finn*.

This argument is not valid. There are circumstances in which the premises are true and the conclusion false. We have already described such circumstances, namely, one in which John learns the name ‘Mark Twain’ by reading *Huckleberry Finn* but learns the name ‘Samuel Clemens’ in the context of learning about 19th century American authors (without learning that the name ‘Mark Twain’ was a pseudonym for Samuel Clemens). John may *not* believe that Samuel Clemens wrote *Huckleberry Finn*. The premises of the above argument, therefore, do not logically entail the conclusion. So the Principle of Substitution appears to break down in the context of propositional attitude reports. The puzzle, then, is to say what causes the principle to fail in these contexts. Why aren't we still saying something true about the man in question if all we have done is changed the name by which we refer to him?

Frege's Theory of Sense and Denotation

To explain these puzzles, Frege suggested that in addition to having a denotation, names and descriptions also express a *sense*. The sense of a expression accounts for its cognitive significance---it is the way by

which one conceives of the denotation of the term. The expressions ‘4’ and ‘8/2’ have the same denotation but express different senses, different ways of conceiving the same number. The descriptions ‘the morning star’ and ‘the evening star’ denote the same planet, namely Venus, but express different ways of conceiving of Venus and so have different senses. The name ‘Pegasus’ and the description ‘the most powerful Greek god’ both have a sense (and their senses are distinct), but neither has a denotation. However, even though the names ‘Mark Twain’ and ‘Samuel Clemens’ denote the same individual, they express different senses. Using the distinction between sense and denotation, Frege can account for the difference in cognitive significance between identity statements of the form "a=a" and "a=b". The sense of the whole statement, on Frege's view, is a function of the senses of its component parts. Since the sense of ‘a’ differs from the sense of ‘b’, the components of "a=a" and "a=b" are different and so the sense of the whole expression will be different in the two cases. Since the sense of an expression accounts for its cognitive significance, Frege has an explanation of the difference in cognitive significance between "a=a" and "a=b", and thus a solution to the first puzzle.

Moreover, Frege proposed that when a term (name or description) follows a propositional attitude verb, it no longer denotes what it ordinarily denotes. Instead, Frege claims that in such contexts, a term denotes its ordinary sense. This explains why the Principle of Substitution fails for terms following the propositional attitude verbs in propositional attitude reports. The Principle asserts that truth is preserved when we substitute one name for another having the same denotation. But, according to Frege's theory, the names ‘Mark Twain’ and ‘Samuel Clemens’ denote different senses when they occur in the following sentences:

> John believes that Mark Twain wrote *Huckleberry Finn*.
> John believes that Samuel Clemens wrote *Huckleberry Finn*.

If they don't denote the same object, then there is no reason to think that substitution of one name for another would preserve truth.

Frege developed the theory of sense and denotation into a thoroughgoing philosophy of language. This philosophy can be explained, at least in outline, by considering a simple sentence such as "John loves Mary". In Frege's view, each word in this sentence is a name and, moreover, the sentence as a whole is a complex name. Each of these names has both a sense and a denotation. Then sense and denotation of the words are basic; but sense and denotation of the sentence as a whole can be described in terms of the sense and denotation of the words and the way in which those words are arranged in the sentence. Let us refer to the denotation and sense of the words as follows:

> **d**[j] refers to the denotation of the name ‘John’.
> **d**[m] refers to the denotation of the name ‘Mary’.
> **d**[L] refers to the denotation of the name ‘loves’.
> **s**[j] refers to the sense of the name ‘John’.
> **s**[m] refers to the sense of the name ‘Mary’.
> **s**[L] refers to the sense of the name ‘loves’.

We now work toward a theoretical description of the denotation of the sentence as a whole. On Frege's view, **d**[j] and **d**[m] are the real individuals John and Mary, respectively. **d**[L] is a function that maps **d**[m] (i.e., Mary) to a function which serves as the denotation of the predicate ‘loves Mary’. Let us refer to that function as **d**[Lm]. Now the function **d**[Lm] maps **d**[j] (i.e., John) to the denotation of the sentence "John loves Mary". Let us refer to the denotation of the sentence as **d**[jLm]. Frege identifies the denotation of a sentence as one of the two truth values. Because **d**[Lm] maps objects to truth values, it is a concept. Thus, **d**[jLm] is the truth value The True if the extension of the concept **d**[Lm] contains John; otherwise it is the truth value The False. So, on Frege's view, the sentence "John loves Mary" names a truth value.

The sentence "John loves Mary" also expresses a sense. Its sense may be described as follows. First, **s**[L] (the sense of the name "loves") is identified as a function. This function maps **s**[m] (the sense of the name "Mary") to the sense of the predicate ‘loves Mary’. Let us refer to the sense of ‘loves Mary’ as **s**[Lm]. Now the function **s**[Lm] maps **s**[j] (the sense of the name ‘John’) to the sense of the whole sentence. Let us call the sense of the entire sentence **s**[jLm]. Frege calls the sense of a sentence a *thought*, and whereas there are only two truth values, he supposes that there are an infinite number of thoughts.

On Frege's view, therefore, the sentences "4=8/2" and "4=4" both name the same truth value, but they express different thoughts. That is because **s**[4] is different from **s**[8/2]. Similarly, "Mark Twain=Mark Twain" and "Mark Twain=Samuel Clemens" denote the same truth value, but express different thoughts (since the sense of the names differ). Thus, Frege has a general account of the difference in the cognitive significance between identity statements of the form "*a*=*a*" and "*a*=*b*". Furthermore, recall that Frege proposed that terms following propositional attitude verbs denote not their ordinary denotations but rather the senses they ordinarily express. In fact, in the following propositional attitude report, not only do the words ‘Mark Twain’, ‘wrote’ and ‘*Huckleberry Finn* ’ denote their ordinary senses, but the entire sentence "Mark Twain wrote *Huckleberry Finn*" also denotes its ordinary sense (namely, a thought):

> John believes that Mark Twain wrote *Huckleberry Finn*.

Frege, therefore, would analyze this attitude report as follows: "believes that" denotes a function that maps the denotation of the sentence "Mark Twain wrote *Huckleberry Finn*" to a concept. In this case, however, the denotation of the sentence "Mark Twain wrote *Huckleberry Finn*" is not a truth value but rather a thought. The thought it denotes is different from the thought denoted by "Samuel Clemens wrote *Huckleberry Finn*" in the following propositional attitude report:

> John believes that Samuel Clemens wrote *Huckleberry Finn*.

Since the thought denoted by "Samuel Clemens wrote *Huckleberry Finn*" in this context differs from the thought denoted by "Mark Twain wrote *Huckleberry Finn*" in the same context, the concept denoted by ‘believes that Mark Twain wrote *Huckleberry Finn*’ is a different concept from the one denoted by ‘believes that Samuel Clemens wrote *Huckleberry Finn*’. One may consistently suppose that the concept denoted by the former predicate maps John to The True whereas the the concept denoted by the latter

predicate does not. Frege's analysis therefore preserves our intuition that John can believe that Mark Twain wrote *Huckleberry Finn* without believing that Samuel Clemens did. It also preserves the Principle of Substitution---the fact that one cannot substitute "Samuel Clemens" for "Mark Twain" when these names occur after propositional attitude verbs does not constitute evidence against the Principle. For if Frege is right, names do not have their usual denotation when they occur in these contexts.

Chronological Catalog of Frege's Work

Chronological Catalog of Frege's Work (PDF file=Adobe Acrobat file)

Bibliography

- Beaney, M., 1996, *Frege: Making Sense*, London: Duckworth
- Bell, D., 1979, *Frege's Theory of Judgment*, Oxford: Clarendon
- Boolos, G., 1986, "Saving Frege From Contradiction", *Proceedings of the Aristotelian Society*, **87** (1986/87): 137-151
- Boolos, G., 1987, "The Consistency of Frege's *Foundations of Arithmetic*", in J. Thomson (ed.), *On Being and Saying*, Cambridge, MA: The MIT Press, pp. 3-20
- Boolos, G., 1998, *Logic, Logic, and Logic*, Cambridge, MA: Harvard University Press
- Currie, G., 1982, *Frege: An Introduction to His Philosophy*, Brighton, Sussex: Harvester Press
- Demopoulos, W., (ed.), 1995, *Frege's Philosophy of Mathematics*, Cambridge, MA: Harvard
- Dummett, M., 1973, *Frege: Philosophy of Language*, London: Duckworth
- Dummett, M., 1981, *The Interpretation of Frege's Philosophy*, Cambridge, MA: Harvard University Press
- Dummett, M., 1991, *Frege: Philosophy of Mathematics*, Cambridge, MA: Harvard University Press
- Haaparanta, L., and Hintikka, J., (eds.), 1986, *Frege Synthesized*, Dordrecht: D. Reidel
- Heck, R., 1993, "The Development of Arithmetic in Frege's *Grundgesetze der Arithmetik*", *Journal of Symbolic Logic*, **58**/2 (June): 579-601
- Klemke, E. D. (ed.), 1968, *Essays on Frege*, Urbana, IL: University of Illinois Press
- Parsons, T., 1981, "Frege's Hierarchies of Indirect Senses and the Paradox of Analysis", *Midwest Studies in Philosophy: VI*, Minneapolis: University of Minnesota Press, pp. 37-57
- Parsons, T., 1987, "On the Consistency of the First-Order Portion of Frege's Logical System", *Notre Dame Journal of Formal Logic*, **28**/1 (January): 161-168
- Parsons, T., 1982, "Fregean Theories of Fictional Objects", *Topoi*, **1**: 81-87
- Pelletier, F.J., 2001, "Did Frege Believe Frege's Principle", *Journal of Logic, Language, and Information*, **10**/1: 87-114
- Perry, J., 1977, "Frege on Demonstratives", *Philosophical Review*, **86** (1977): 474-497
- Resnik, M., 1980, *Frege and the Philosophy of Mathematics*, Ithaca, NY: Cornell University Press
- Ricketts, T., 1997, "Truth-Values and Courses-of-Value in Frege's *Grundgesetze*", in *Early*

Analytic Philosophy, W. Tait (ed.), Chicago: Open Court, pp. 187-211
- Ricketts, T., 1986, "Logic and Truth in Frege", *Proceedings of the Aristotelian Society*, Supplementary Volume 70, pp. 121-140
- Salmon, N., 1986, *Frege's Puzzle*, Cambridge, MA: MIT Press
- Schirn, M., (ed.), 1996, *Frege: Importance and Legacy*, Berlin: de Gruyter
- Sluga, H., 1980, *Gottlob Frege*, London: Routledge and Kegan Paul
- Sluga, H., 1993, *The Philosophy of Frege*, New York: Garland, four volumes
- Wright, C., 1983, *Frege's Conception of Numbers as Objects*, Aberdeen: Aberdeen University Press

Other Internet Resources

- Die Grundlagen der Arithmetik, (528 KB PDF file), original German text (maintained by Alain Blachair, Académie de Nancy-Metz)
- MacTutor History of Mathematics Archive
- Metaphysics Research Lab Web Page on Frege
- Brian Carver's Web Page on Frege

Related Entries

denotation | Frege, Gottlob: logic, theorem, and foundations for arithmetic | logic: classical | logic: intensional | logicism | mathematics, philosophy of | *Principia Mathematica* | quantification | Russell, Bertrand | Russell’s paradox | sense/reference distinction

Edward N. Zalta
zalta@mally.stanford.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 14, 1995
Content last modified: January 30, 2002

Principia Mathematica

Principia Mathematica is the landmark work on mathematical logic and the foundations of mathematics written by Alfred North Whitehead and Bertrand Russell. It was first published in three volumes, in 1910, 1912 and 1913. Written as a defense of logicism (the view that mathematics is in some significant sense reducible to logic), the book was instrumental in popularizing modern mathematical logic. Next to Aristotle's *Organon*, it is the most influential book on logic ever written. Interested readers may wish to view the Title page of the 1st edition of *Principia Mathematica*, Volume 1 or the Cover of the 1st edition of *Principia Mathematica to *56*.

History of *Principia Mathematica*

Logicism is the view that (some or all of) mathematics can be reduced to (formal) logic, and it is often explained as a two-part thesis. First, it consists of the thesis that all mathematical truths can be translated into logical truths or, in other words, that the vocabulary of mathematics constitutes a proper subset of that of logic. Second, it consists of the thesis that all mathematical proofs can be recast as logical proofs or, in other words, that the theorems of mathematics constitute a proper subset of those of logic. In Bertrand Russell's words, it is the logicist's goal "to show that all pure mathematics follows from purely logical premises and uses only concepts definable in logical terms" (Russell [1955], p. 57).

In its essentials, this thesis was first advocated in the late 17th century by Gottfried Leibniz. Later, the idea was defended in much greater detail by Gottlob Frege. During the critical movement initiated in the 1820s, mathematicians such as Bernard Bolzano, Niels Abel, Louis Cauchy and Karl Weierstrass had succeeded in eliminating much of the vagueness and many of the contradictions present in the mathematical theories of their day. By the late 1800s, William Hamilton had also introduced ordered

couples of reals as the first step in supplying a logical basis for the complex numbers, and Weierstrass, Richard Dedekind and Georg Cantor had all developed methods for founding the irrationals in terms of the rationals. Using work by H.G. Grassmann and Dedekind, Guiseppe Peano had also gone on to develop a theory of the rationals based on his now famous axioms for the natural numbers. Thus, by Frege's day it was generally recognized that a large portion of mathematics could be derived from a relatively small set of primitive notions.

Even so, it was not until 1879, when Frege developed the necessary logical apparatus, that the project of logicism could be said to have become technically viable. Following another five years' work, Frege arrived at the definitions necessary for logicising arithmetic. During the 1890s he worked on many of the essential derivations. However, with the discovery of paradoxes such as Russell's paradox at the turn of the century, it appeared that additional resources would be required if logicism were to succeed.

By 1903, both Whitehead and Russell had come to the same conclusion. By this time, both men were also in the initial stages of preparing second volumes to earlier books on related topics: Whitehead's 1898 *A Treatise on Universal Algebra* and Russell's 1903 *The Principles of Mathematics.* Since their research overlapped considerably, they began collaboration on what was eventually to become *Principia Mathematica.*

Unfortunately, after almost a decade of difficult work on the part of both men, Cambridge University Press concluded that publishing *Principia* would result in an estimated loss of approximately 600 pounds. Although the press agreed to assume half this amount and the Royal Society agreed to donate another 200 pounds, that still left a 100-pound deficit. Only by each contributing 50 pounds were the authors able to see their work through to publication.

Today there is not a major academic library anywhere in the world that does not possess a copy of this landmark publication.

Significance of *Principia Mathematica*

Principia's main goal of showing the detailed deduction of mathematics from logic proved to be controversial. Primarily at issue were the kinds of assumptions that Whitehead and Russell used to complete their project. Although *Principia* succeeded in providing detailed derivations of major theorems in set theory, finite and transfinite arithmetic, and elementary measure theory, two axioms in particular were arguably non-logical in character: the axiom of infinity and the axiom of reducibility. The axiom of infinity assumed that there exists an infinity of objects. Thus, it made the kind of assumption that is generally thought to be empirical rather than logical in nature. The axiom of reducibility was introduced as a means of overcoming the not completely satisfactory effects of the theory of types, the theory that Russell and Whitehead used to restrict the notion of a well-formed expression, and so to avoid the paradoxes. Although technically feasible, many critics claimed that the axiom of reducibility was simply too ad hoc to be justified philosophically. As a result, the question of whether mathematics could be reduced to logic, or whether it could be reduced only to set-theory, remained open.

Despite these criticisms, *Principia Mathematica* proved to be remarkably influential in at least three other ways. First, it popularized modern mathematical logic to an extent undreamt of by its authors. By using a notation superior in many ways to that of Frege, Whitehead and Russell managed to convey the remarkable expressive power of modern logic in a way that previous writers had been unable to achieve. Second, by exhibiting so clearly the deductive power of the new logic, Whitehead and Russell were also able to show how powerful the modern idea of a formal system could be. Third, *Principia Mathematica* reaffirmed clear and interesting connections between logicism and two main branches of traditional philosophy, namely metaphysics and epistemology, thus initiating new and interesting work in both these and other areas.

Thus, not only did *Principia* introduce a wide range of philosophically rich notions (such as propositional function, logical construction, and type theory), it also set the stage for the discovery of classical metatheoretic results (such as those of Kurt Gödel and others) and initiated a tradition of common technical work in fields as diverse as philosophy, mathematics, linguistics, economics and computer science.

Contents of *Principia Mathematica*

Principia Mathematica appeared in three volumes which together are divided into six parts. Volume 1 begins with a lengthy Introduction containing sections entitled "Preliminary Explanations of Ideas and Notations", "The Theory of Logical Types" and "Incomplete Symbols". It also contains Part I, entitled "Mathematical Logic", which contains sections on "The Theory of Deduction", "Theory of Apparent Variables", "Classes and Relations", "Logic of Relations", and "Products and Sums of Classes"; and Part II, entitled "Prolegomena to Cardinal Arithmetic", which contains sections on "Unit Classes and Couples", "Sub-Classes, Sub-Relations, and Relative Types", "One-Many, Many-One and One-One Relations", "Selections", and "Inductive Relations".

Volume 2 begins with a "Prefatory Statement of Symbolic Conventions". It then continues with Part III, entitled "Cardinal Arithmetic", which itself contains sections on "Definition and Logical Properties of Cardinal Numbers", "Addition, Multiplication and Exponentiation", and "Finite and Infinite"; Part IV, entitled Relation-Arithmetic", which contains sections on "Ordinal Similarity and Relation-Numbers", "Addition of Relations, and the Product of Two Relations", "The Principle of First Differences, and the Multiplication and Exponentiation of Relations", and "Arithmetic of Relation-Numbers"; and the first half of Part V, entitled "Series", which contains sections on "General Theory of Series", "On Sections, Segments, Stretches, and Derivatives", and "On Convergence, and the Limits of Functions".

Volume 3 continues Part V with sections on "Well-Ordered Series", "Finite and Infinite Series and Ordinals", and "Compact Series, Rational Series, and Continuous Series". It also contains Part VI, entitled "Quantity", which itself contains sections on "Generalization of Number", "Vector-Families", "Measurement", and "Cyclic Families".

A fourth volume, on geometry, was planned but never completed. Even so, the book remains one of the great scientific documents of the twentieth century.

Bibliography

- Frege, Gottlob (1893, 1903) *Grundgesetze der Arithmetik*, Band I (1893), Band II (1903), Jena: Verlag Hermann Pohle. Ed. and trans. in part by M. Furth as *The Basic Laws of Arithmetic*, Berkeley: University of California Press, 1964.
- Russell, Bertrand (1903) *Principles of Mathematics*, Cambridge: Cambridge University Press.
- Russell, Bertrand (1919) *Introduction to Mathematical Philosophy*, London: George Allen & Unwin.
- Russell, Bertrand (1948) "Whitehead and *Principia Mathematica*", *Mind*, 57, 137-138.
- Russell, Bertrand (1955) *My Philosophical Development*, London and New York: Routledge
- Whitehead, Alfred North (1898) *A Treatise on Universal Algebra*, Cambridge: Cambridge University Press.
- Whitehead, Alfred North (1906) *On Mathematical Concepts of the Material World*, London: Dulau.
- Whitehead, Alfred North, and Bertrand Russell (1910, 1912, 1913) *Principia Mathematica*, 3 vols, Cambridge: Cambridge University Press. Second edition, 1925 (Vol. 1), 1927 (Vols 2, 3). Abridged as *Principia Mathematica to *56*, Cambridge: Cambridge University Press, 1962.

Other Internet Resources

- *Principia Mathematica*: Whitehead and Russell

Related Entries

Frege, Gottlob | Frege, Gottlob: logic, theorem, and foundations for arithmetic | Leibniz, Gottfried Wilhelm | logic: classical | logicism | propositional function | Russell, Bertrand | Russell's paradox | type theory | Whitehead, Alfred North

A.D. Irvine
andrew.irvine@ubc.ca

Table of Contents

First published: May 21, 1996
Content last modified: July 20, 2000

Alfred North Whitehead

Alfred North Whitehead (b.1861 - d.1947), British mathematician, logician and philosopher best known for his work in mathematical logic and who, in collaboration with Bertrand Russell, authored the landmark three-volume *Principia Mathematica* (1910, 1912, 1913).

Although there are significant continuities throughout his thought, Whitehead's intellectual life is often divided into three periods. The first corresponds roughly with his time at Cambridge, from 1884 to 1910, during which he worked primarily on logic and mathematics. The second corresponds roughly with his time at London, from 1910 to 1924, during which he concentrated mainly on issues in the philosophy of science. The third corresponds roughly with his time at Harvard, from 1924 onward, during which he worked on more general issues in philosophy, including the development of a comprehensive metaphysical system which has come to be known as process philosophy.

Whitehead's Chronology

A short chronology of the major events in Whitehead's life is as follows:

- (1861) Born February 15 in Ramsgate, Isle of Thanet, Kent, England.
- (1880) Enters Trinity College, Cambridge with a scholarship in mathematics.
- (1884) Elected a Fellow in Mathematics at Trinity.
- (1891) Marries Evelyn Wade.
- (1903) Elected a Fellow of the Royal Society as a result his work on universal algebra.
- (1910) Moves to University College London.
- (1914) Appointed Professor of Applied Mathematics at the Imperial College of Science and Technology.
- (1924) Appointed Professor of Philosophy at Harvard University.
- (1931) Elected a Fellow of the British Academy.
- (1937) Retires from Harvard.
- (1945) Awarded Order of Merit.
- (1947) Dies December 30 in Cambridge, Massachusetts, USA.

For a chronology of Whitehead's major publications, consult the section below entitled Whitehead's Writings.

Whitehead's Philosophical Influence

Whitehead's philosophical influence can be felt in all three of the main areas in which he worked -- logic and the foundations of mathematics, the philosophy of science, and metaphysics -- as well as in other areas such as ethics, education and religion.

Whitehead began his academic career at Trinity College, Cambridge, where, starting in 1885, he taught for twenty-five years. In 1890 Bertrand Russell arrived as a student at Trinity and during the 1890s the two men came into regular contact with one another. According to Russell, "Whitehead was extraordinarily perfect as a teacher"[1] and Whitehead soon became something of a mentor to the younger man.

By the early 1900s, both men had completed books on the foundations of mathematics. Whitehead's 1898 *A Treatise on Universal Algebra* had resulted in his election to the Royal Society. Russell's 1903 *The Principles of Mathematics* had marked a decisive break from his earlier neo-Kantian work, such as his 1897 *An Essay on the Foundations of Geometry*. Since the research for a proposed second volume of Russell's *Principles* overlapped considerably with Whitehead's own research for the proposed second volume of his *Universal Algebra*, the two men began collaboration on what was eventually to become *Principia Mathematica* (1910, 1912, 1913). According to Whitehead, they initially expected the research

to take about a year to complete. In the end, they worked together on the project for a full decade.

Logicism, the theory that mathematics is in some important sense reducible to logic, consists of two main theses. The first is that all mathematical truths can be translated into logical truths or, in other words, that the vocabulary of mathematics constitutes a proper subset of that of logic. The second is that all mathematical proofs can be recast as logical proofs or, in other words, that the theorems of mathematics constitute a proper subset of those of logic.

Like Gottlob Frege, Whitehead and Russell took the view that numbers could be identified with sets of sets, and that number-theoretic operations could be explained in terms of set theoretic operations such as intersection, union, and difference. Although Whitehead and Russell were able to provide many detailed derivations of major theorems in set theory, finite and transfinite arithmetic, and elementary measure theory, the issue of whether set theory itself can be said to have been successfully reduced to logic remained controversial.

Following the completion of *Principia*, Whitehead and Russell began to go their separate ways. Perhaps inevitably, Russell's anti-war activities during World War I (in which Whitehead lost his youngest son) also led to something of a split between the two men. Nevertheless, they remained on relatively good terms for the rest of their lives.

At the University of London, Whitehead turned his attention to issues in the philosophy of science. Of particular note was his rejection of the idea that each object has a simple spatial or temporal location. Instead, Whitehead advocated the view that all objects should be understood as fields having both temporal and spatial extensions. For example, just as we cannot perceive a Euclidean point that has position but no magnitude, or a line that has length but no breadth, it is impossible, says Whitehead, to conceive of a simple spatial or temporal location. To think that we can do so involves what he called "The Fallacy of Misplaced Concreteness," the error of mistaking the abstract for the concrete.[2]

As Whitehead explains, "I shall argue that among the primary elements of nature as apprehended in our immediate experience, there is no element whatever which possesses this character of simple location. ... [Instead,] I hold that by a process of constructive abstraction we can arrive at abstractions which are the simply-located bits of material, and at other abstractions which are the minds included in the scientific scheme."[3]

Whitehead's basic idea was that we obtain the abstract idea of a spatial point by considering a real-life series of volumes extending over each other, for example, a nested series of Russian dolls or a nested series of pots and pans. However, it would be a mistake to think of a spatial point as being anything more than an abstraction; instead, real positions involve the entire series of extended volumes. As Whitehead himself puts it, "In a certain sense, everything is everywhere at all times. For every location involves an aspect of itself in every other location. Thus every spatio-temporal standpoint mirrors the world."[4]

Further, according to Whitehead, every real-life object may be understood as a similarly constructed

series of events and processes. It is this latter idea that Whitehead later systematically elaborates in his imposing *Process and Reality* (1929), going so far as to suggest that process, rather than substance, should be taken as the fundamental metaphysical constituent of the world. Underlying this work was also the basic idea that, if philosophy is to be successful, it must explain the connection between objective, scientific and logical descriptions of the world and the more everyday world of subjective experience.

While at London, Whitehead also became involved in many practical aspects of tertiary education, serving as Dean of the Faculty of Science and holding several other senior administrative posts. Many of the essays in his *The Aims of Education and Other Essays* (1929) date from this time. It was also during his time in London that Whitehead also published several less well known books, including *An Inquiry Concerning the Principles of Natural Knowledge* (1919), *The Concept of Nature* (1920), and *The Principle of Relativity* (1922).

Upon being offered an appointment at Harvard, Whitehead moved to the United States in 1924. Given his prior training in mathematics and in the physical sciences, it was sometimes joked that the first philosophy lectures he ever attended were those that he delivered at Harvard in his new role as Professor of Philosophy. A year later he also delivered Harvard's prestigious Lowell Lectures which formed the basis for his first primarily metaphysical book, *Science and the Modern World* (1925). In it he again introduced several themes which later found fuller expression in *Process and Reality*. The same was true of the 1927/28 Gifford Lectures at the University of Edinburgh on which *Process and Reality* came to be based.

In *Process and Reality*, rather than assuming substance as the basic metaphysical category, Whitehead introduces the notion of an *actual occasion*. On Whitehead's view, an actual occasion is not an enduring substance, but a process of becoming. As Donald Sherburne points out, "It is customary to compare an actual occasion with a Leibnizian monad, with the caveat that whereas a monad is windowless, an actual occasion is ‘all window’. It is as though one were to take Aristotle's system of categories and ask what would result if the category of substance were displaced from its preeminence by the category of relation ... ."[5] As Whitehead himself explains, his "philosophy of organism is the inversion of Kant's philosophy ... For Kant, the world emerges from the subject; for the philosophy of organism, the subject emerges from the world."[6]

Significantly, this view runs counter to the more traditional view of material substance: "There persists," says Whitehead, "[a] fixed scientific cosmology which presupposes the ultimate fact of an irreducible brute matter, or material, spread through space in a flux of configurations. In itself such a material is senseless, valueless, purposeless. It just does what it does do, following a fixed routine imposed by external relations which do not spring from the nature of its being. It is this assumption that I call ‘scientific materialism.’ Also it is an assumption which I shall challenge as being entirely unsuited to the scientific situation at which we have now arrived."[7]

The assumption of scientific materialism is effective in many contexts, says Whitehead, only because it directs our attention to a certain class of problems that lend themselves to analysis within this framework.

However, scientific materialism is less successful when addressing issues of teleology and when trying to develop a comprehensive, intergrated picture of the universe as a whole. According to Whitehead, recognition that the world is organic rather than materialistic is therefore essential, and this change in viewpoint can result as easily from attempts to understand modern physics as from attempts to understand human psychology and teleology. Says Whitehead, "Mathematical physics presumes in the first place an electromagnetic field of activity pervading space and time. The laws which condition this field are nothing else than the conditions observed by the general activity of the flux of the world, as it individualises itself in the events."[8]

The end result is that Whitehead concludes that "nature is a structure of evolving processes. The reality is the process."[9]

Whitehead's ultimate attempt to develop a metaphysical unification of space, time, matter, events and teleology has proved to be controversial. In part this may be because of the connections that Whitehead saw between his metaphysics and traditional theism. According to Whitehead, religion is concerned with permanence amid change, and can be found in the ordering we find within nature, something he sometimes called the "primordial nature of God". Thus although not especially influential among contemporary Anglo-American secular philosophers, his metaphysical ideas have had greater influence among many theologians and philosophers of religion.

Whitehead's Writings

Whitehead's principal publications include the following:

- (1898) *A Treatise on Universal Algebra*, Cambridge: Cambridge University Press.
- (1906) *On Mathematical Concepts of the Material World*, London: Dulau.
- (1906) *The Axioms of Projective Geometry*, Cambridge: Cambridge University Press.
- (1907) *The Axioms of Descriptive Geometry*, Cambridge: Cambridge University Press.
- (1910, 1912, 1913) (with Bertrand Russell) *Principia Mathematica*, 3 vols, Cambridge: Cambridge University Press. Second edition, 1925 (Vol. 1), 1927 (Vols 2, 3). Abridged as *Principia Mathematica to *56*, Cambridge: Cambridge University Press, 1962.
- (1911) *An Introduction to Mathematics*, London: Williams & Norgate.
- (1919) *An Enquiry Concerning the Principles of Natural Knowledge*, Cambridge: Cambridge University Press.
- (1920) *The Concept of Nature*, Cambridge: Cambridge University Press.
- (1922) *The Principle of Relativity With Applications to Physical Science*, Cambridge: Cambridge University Press.
- (1925) *Science and the Modern World*, Cambridge: Cambridge University Press.
- (1926) *Religion in the Making*, New York: Macmillan.
- (1927) *Symbolism, Its Meaning and Effect*, New York: Macmillan.
- (1929) *The Aims of Education and Other Essays*, New York: Macmillan.
- (1929) *The Function of Reason*, Princeton: Princeton University Press.

- (1929) *Process and Reality,* New York: Macmillan.
- (1933) *Adventures of Ideas,* New York: New American.
- (1934) *Nature and Life,* Chicago: University of Chicago Press.
- (1938) *Modes of Thought*, New York: Macmillan.
- (1947) *Essays in Science and Philosophy,* New York: Philosophical Library.
- (1947) *The Wit and Wisdom of Whitehead,* Boston: Beacon Press.

Bibliography

- Bright, Laurence (1958) *Whitehead's Philosophy of Physics*, London: Sheed and Ward.
- Cobb, John B. (1965) *A Christian Natural Theology, Based on the Thought of Alfred North Whitehead,* Philadelphia: Westminster Press.
- Connelly, Robert Joseph (1981) *Whitehead vs Hartshorne*, Washington, D.C.: University Press of America.
- Dunkel, Harold Baker (1965) *Whitehead on Education,* Columbus: Ohio State University Press.
- Emmet, Dorothy Mary (1932) *Whitehead's Philosophy of Organism*, 2nd edition, London: Macmillan, 1966.
- Hartshorne, Charles (1972) *Whitehead's Philosophy: Selected Essays, 1935-1970,* Lincoln: University of Nebraska Press.
- Johnson, A.H. (1952) *Whitehead's Theory of Reality*, Boston: Beacon Press.
- Johnson, A.H. (1973) *Experiential Realism*, London: George Allen and Unwin.
- Kline, George Louis (1963) *Alfred North Whitehead*, Englewood Cliffs, N.J.: Prentice-Hall.
- Lango, John W. (1972) *Whitehead's Ontology*, Albany: State University of New York Press.
- Lawrence, Nathaniel Morris (1956) *Whitehead's Philosophical Development*, Berkeley: University of California Press.
- Lowe, Victor (1962) *Understanding Whitehead*, Baltimore: Johns Hopkins Press.
- Lowe, Victor (1985) *Alfred North Whitehead*, Baltimore: Johns Hopkins University Press.
- Lucas, George R. (1989) *The Rehabilitation of Whitehead*, Albany: State University of New York Press.
- Nobo, Jorge Luis (1986) *Whitehead's Metaphysics of Extension and Solidarity,* Albany: State University of New York Press.
- Pittenger, W. Norman (1969) *Alfred North Whitehead,* Richmond: John Knox Press.
- Pols, Edward (1967) *Whitehead's Metaphysics,* Carbondale: Southern Illinois University Press.
- Quine, Willard Van Orman (1941) "Whitehead and the Rise of Modern Logic", in Schilpp, Paul Arthur (ed.) *The Philosophy of Alfred North Whitehead*, La Salle: Open Court, 125-164.
- Ross, Stephen David (1983) *Perspectives in Whitehead's Metaphysics,* Albany: State University of New York Press.
- Russell, Bertrand (1903) *The Principles of Mathematics,* Cambridge: Cambridge University Press.
- Russell, Bertrand (1948) "Whitehead and *Principia Mathematica", Mind*, 57, 137-138.
- Russell, Bertrand (1952) "Alfred North Whitehead", *The Listener,* 48 (10 July), 51-52. Revised and reprinted in Russell, *Portraits From Memory,* New York: Simon and Schuster, 1956, 99-104; and in Russell, *The Autobiography of Bertrand Russell,* Vol. 1, London: George Allen & Unwin,

1967, 127-130.

- Schilpp, Paul Arthur (ed.) (1941) *The Philosophy of Alfred North Whitehead,* La Salle: Open Court.
- Sherburne, Donald W. (1966) *A Key to Whitehead's Process and Reality*, New York: Macmillian.

Other Internet Resources

- University of St Andrew's MacTutor History of Mathematics Archive -- Alfred North Whitehead

Related Entries

Frege, Gottlob | logic: classical | logicism | *Principia Mathematica* | process philosophy | Russell, Bertrand | substance

A. D. Irvine
andrew.irvine@ubc.ca

First published: May 21, 1996
Content last modified: July 20, 2000

Stanford Encyclopedia of Philosophy
Notes to Alfred North Whitehead

Notes

1. Bertrand Russell, *Portraits From Memory,* New York: Simon and Schuster, 1956, p. 104.

2. Alfred North Whitehead, *Science and the Modern World,* New York: Free Press, 1967, p. 58.

3. Alfred North Whitehead, *Science and the Modern World,* New York: Free Press, 1967, p. 58. See also, Alfred North Whitehead, *An Enquiry Concerning the Principles of Natural Knowledge*, Cambridge: Cambridge University Press, 1919, Part III.

4. Alfred North Whitehead, *Science and the Modern World,* New York: Free Press, 1967, p. 91.

5. Donald W. Sherburne, "Whitehead, Alfred North", in *The Cambridge Dictionary of Philosophy,* Robert Audi (ed.), Cambridge: Cambridge University Press, 1995, p. 852.

6. Donald W. Sherburne, "Whitehead, Alfred North", in *The Cambridge Dictionary of Philosophy,* Robert Audi (ed.), Cambridge: Cambridge University Press, 1995, p. 852.

7. Alfred North Whitehead, *Science and the Modern World,* New York: Free Press, 1967, p. 17.

8. Alfred North Whitehead, *Science and the Modern World,* New York: Free Press, 1967, pp. 152-3.

9. Alfred North Whitehead, *Science and the Modern World,* New York: Free Press, 1967, p. 72.

A. D. Irvine
andrew.irvine@ubc.ca

First published: May 21, 1996
Content last modified: July 20, 2000

Process Philosophy

The philosophy of process is a venture in metaphysics, the general theory of reality. Its concern is with what exists in the world and with the terms of reference in which this reality is to be understood and explained. The task of metaphysics is, after all, to provide a cogent and plausible account of the nature of reality at the broadest, most synoptic and comprehensive level. And it is to this mission of enabling us to characterize, describe, clarify and explain the most general features of the real that process philosophy addresses itself in its own characteristic way. The guiding idea of its approach is that natural existence consists in and is best understood in terms of *processes* rather than *things* -- of modes of change rather than fixed stabilities. For processists, change of every sort -- physical, organic, psychological -- is the pervasive and predominant feature of the real.

Process philosophy diametrically opposes the view -- as old as Parmenides and Zeno and the Atomists of Pre-Socratic Greece -- that denies processes or downgrades them in the order of being or of understanding by subordinating them to substantial things. By contrast, process philosophy pivots on the thesis that the processual nature of existence is a fundamental fact with which any adequate metaphysic must come to terms.

Process philosophy puts processes at the forefront of philosophical and specifically of ontological concern. Process should here be construed in pretty much the usual way -- as *a sequentially structured sequence of successive stages or phases*. Three factors accordingly come to the fore:

1. That a process is a complex -- a unity of distinct stages or phases. A process is always a matter of now this, now that.
2. That this complex has a certain temporal coherence and unity, and that processes accordingly have an ineliminably temporal dimension.
3. That a process has a structure, a formal generic format in virtue of which every concrete process is equipped with a shape or format.

From the time of Aristotle, Western metaphysics has had a marked bias in favor of *things* or *substances*. However, another variant line of thought was also current from the earliest times onward. After all, the concentration on perduring physical *things* as existents in nature slights the equally good claims of another ontological category, namely processes, events, occurrences -- items better indicated by verbs than nouns. And, clearly, storms and heat-waves are every bit as real as dogs and oranges.

What is characteristically definitive of *process* philosophizing as a distinctive sector of philosophical

tradition is not simply the commonplace recognition of natural process as the active initiator of what exists in nature, but an insistence on seeing process as constituting an essential aspect of everything that exists -- a commitment to the fundamentally processual nature of the real. For the process philosopher is, effectively by definition, one who holds that what exists in nature is not just originated and sustained by processes but is in fact ongoingly and inexorably *characterized* by them. On such a view, process is both pervasive in nature and fundamental for its understanding.

---

1. Historical Aspects

Like so much else in the field, process philosophy began with the ancient Greeks. The Greek theoretician Heraclitus of Ephesus (b. ca. 540 B.C.) -- known even in antiquity as "the obscure" -- is universally recognized as the founder of the process approach. His book "On Nature" depicted the world as a manifold of opposed forces joined in mutual rivalry, interlocked in constant strife and conflict. Fire, the most changeable and ephemeral of these elemental forces, is the basis of all: "This world-order . . . is . . . an ever living fire, kindling in measures and going out in measures" (Fr. 217, Kirk-Raven-Schofield). The fundamental "stuff" of the world is not a material substance of some sort but a natural process, namely "fire," and all things are products of its workings (*puros tropai*). The variation of different states and conditions of fire -- that most process-manifesting of the four traditional Greek elements -- engenders all natural change. For fire is the destroyer and transformer of things and "All things happen by strife and necessity" (Fr. 211, *ibid*). And this changeability so pervades the world that "one cannot step twice into the same river" (Fr. 215, *ibid*). As Heraclitus saw it, reality is at bottom not a constellation of *things* at all, but one of *processes*: we must at all costs avoid the fallacy of substantializing nature into perduring things (substances) because it is not stable things but fundamental forces and the varied and fluctuating activities

which they produce that make up this world of ours. Process is fundamental: the river is not an *object*, but an ever-changing flow; the sun is not a *thing*, but a flaming fire. Everything in nature is a matter of process, of activity, of change. Heraclitus taught that *panta rhei* ("everything flows") and this principle exerted a profound influence on classical antiquity. Even Plato, who did not much like the principle ("like leaky pots" he added at *Cratylus* 440 C), came to locate his exception to it -- the enduring and changeless "ideas" -- in a realm wholly removed from the domain of material reality.

Heraclitus may accordingly be seen as the founding father of process philosophy, at any rate in the intellectual tradition of the West. And the static system of Parmenides affords its sharpest contrast and most radical opposition. However, the paradigm substance philosophy of classical antiquity was the atomism of Leucippus and Democritus and Epicurus, which pictured all of nature as composed of unchanging and inert material atoms whose only commerce with process was an alteration of their positioning in space and time. Here the *properties* of substances are never touched by change, which effects only their *relations*. It was this sort of view that Heraclitus preeminently sought to oppose.

In recent years, "process philosophy" has virtually become a code-word for the doctrines of Alfred North Whitehead and his followers. But of course, this cannot really be what process philosophy actually is. If there indeed is a "philosophy" of process, it must pivot not a *thinker* but on a *theory*. What is at issue must, in the end, be a philosophical position that has a larger life of its own, apart from any particular exposition or expositor. And in fact process philosophy is a well-defined and influential tendency of thought that can be traced back through the history of philosophy to the days of the Pre-Socratics. Its leading exponents were Heraclitus, Leibniz, Bergson, Peirce, and William James -- and it ultimately moved on to include Whitehead and his school (Charles Hartshorne, Paul Weiss), but also other 20th Century philosophers such as Samuel Alexander, C. Lloyd Morgan, and Andrew Paul Ushenko.

2. How Process Philosophy Proceeds

Against this historical background, "process philosophy" may be understood as a doctrine invoking certain basic propositions: (1) That time and change are among the principal categories of metaphysical understanding, (2) That process is a principal category of ontological description, (3) That process is more fundamental, or at any rate not less fundamental than things for the purposes of ontological theory, (4) That several if not all of the major elements of the ontological repertoire (God, nature-as-a whole, persons, material substances) are best understood in process linked terms, and (5) That contingency, emergence, novelty, and creativity are among the fundamental categories of metaphysical understanding. A process philosopher, accordingly, is someone for whom temporality, activity, and change -- of alteration, striving, passage, and novelty-emergence -- are the cardinal factors for our understanding of the real.

The demise of classical atomism brought on by the dematerialization of physical matter through the rise of the quantum theory brings much aid and comfort to a process-oriented metaphysics. Matter in the small, as contemporary physics conceives it, is not a Rutherfordian planetary system of particle-like objects, but a collection of fluctuating processes organized into stable structures (insofar as there is indeed

stability at all) by statistical regularities -- i.e., by regularities of comportment at the level of aggregate phenomena. Twentieth century physics has thus turned the tables on classical atomism. Instead of very small *things* (atoms) combining to produce standard processes (windstorms and such) modern physics envisions very small processes (quantum phenomena) combining to produce standard things (ordinary macro-objects) as a result of their *modus operandi*.

For the process philosopher, the classical principle *operari sequitur esse* (functioning follows upon being) is reversed: his motto is the reverse *esse sequitur operari*. As he sees it, all is in the final analysis the product of processes. Process thus has priority over product -- both ontologically and epistemically. As process philosophers see it, processes are basic and things derivative, because it takes a mental process (of separation) to extract "things" from the blooming buzzing confusion of the world's physical processes. For process philosophy, what a thing *is* consists in what it *does*.

And insofar as reality itself is a vast macroprocess embracing a diversified manifold of microprocesses novelty, innovation, and the emergence of new focus is an inherent feature of the cosmic scene.

3. An Evolutionary Perspective

Evolution is an emblematic and paradigmatic process for process philosophy. For not only is evolution a process that makes philosophers and philosophy possible, but it provides a clear model for how processual novelty and innovation comes into operation in nature's self-engendering and self-perpetuating scheme of things. Evolution, be it of organism or of mind, of subatomic matter or of the cosmos as a whole, reflects the pervasive role of process which philosophers of this school see as central both to the nature of our world and to the terms in which it must be understood. Change pervades nature. The passage of time leaves neither individuals nor types (species) of things statically invariant. Process at once destabilizes the world and is the cutting-edge of advance to novelty. And evolution of every level, physical, biological, and cosmic carries the burden of the work here. But does it work blindly?

On the issue of purposiveness in nature, process philosophers divide into two principal camps. On the one side is the naturalistic (and generally secularist) wing that sees nature's processuality as a matter of an inner push or nisus to something new and different. On the other side is the teleological (and often theological) wing that sees nature's processuality as a matter of teleological directedness towards a positive destination. Both agree in according a central role to novelty and innovation in nature. But the one (naturalistic) wing sees this in terms of chance-driven randomness that leads away from the settled formulations of an established past, while the other (teleological) wing sees this in terms of a goal-directed purposiveness preestablished by some value-geared directive force.

Process philosophy correspondingly has a complex, two sided relationship with the theory of evolution. For secular, atheological processists evolution typifies the creative workings of a self-sustaining nature that dispenses with the services of God. For theological processists like Teilhard de Chardin, evolution exhibits God's handwriting in the book of nature. But processists of all descriptions see evolution not only as a crucial instrument for understanding the role of intelligence in the world's scheme of things but

also as a key aspect of the world's natural development. And, more generally, the evolutionary process has provided process philosophy with one of its main models for how large scale collective processes (on the order of organic development at large) can inhere in and result from the operation of numerous small-scale individual processes (on the order of individual lives), thus accounting for innovation and creativity also on a macro-level scale.

But there is one further complexity here. Where human intelligence is concerned, biological evolution is undoubtedly Darwinian, with teleologically blind natural selection operating with respect to teleologically blind random mutations. Cultural evolution, on the other hand, is generally Teilhardian, governed by a rationally-guided selection among purposefully devised mutational variations. Taken in all, cognitive evolution involves both components, superimposing rational selection on biological selection. Our cognitive capacities and faculties are part of the natural endowment we owe to biological evolution. But our cognitive methods, procedures, standards, and techniques are socio-culturally developed resources that evolve through *rational* selection in the process of cultural transmission through successive generations. Our cognitive hardware (mechanisms and capacities) develops through Darwinian natural selection, but our cognitive software (the methods and procedures by which we transact our cognitive business) develops in a Teilhardian process of rational selection that involves purposeful intelligence-guided variation and selection. Biology produces the instrument, so to speak, and culture writes the music -- where obviously the former powerfully constrains the latter. (You cannot play the drums on a piano.)

The ancient Greeks grappled with the question: Is anything changeless, eternal, and exempt from the seemingly all-destructive ravages of time. Rejecting the idea of eternal material atoms, Plato opted for eternal changeless universals ("form," "ideas") and the Stoics for eternal, changeless laws. But the world-picture of modern science has seemingly blocked these solutions. For, as it sees the matter, species (natural kinds) are also children of time, not changelessly present but ever-changingly emergent under the aegis of evolutionary principles. The course of cosmic evolution brings nature's laws also within the orbit of process, endowing these laws with a developmental dimension, (where, after all, was genetics in the microsecond after the big bang?). For process philosophy, nothing is eternal and secure from the changes wrought by time and its iron law that everything that comes into being must perish, so that mortality is omnipresent and death's cold hand is upon all of nature -- laws as well as things.

However, process philosophy does not see this gloomy truth as the end of the story. For process philosophy has always looked to evolutionary theory to pull the plum of collective progress from the pie of distributive mortality. In the small -- item by item -- nature's processes are self-canceling: what arises in the course of time perishes in the course of time. But nevertheless the overall course of processual change tends to the development of an ever richer, more complex and sophisticated condition of things on the world's ample stage. For there are processes and processes: processes of growth and decay, of expanding and contracting, of living and dying. Recognizing that this is so, process philosophy has always accentuated the positive and worn a decidedly optimistic mien. For it regards nature's microprocesses as components of an overall macroprocess whose course is upwards rather than downwards, so to speak. Hitching its wagon to the star of a creative evolutionism, process philosophy sees nature as encompassing creative innovation, productive dynamism and an emergent development of richer, more complex and sophisticated forms of natural existence.

To be sure, there are, in theory, both productive and destructive processes, degeneration and decay being no less prominent in nature than growth and development. Historically, however, most process philosophers have taken a decidedly optimistic line and have envisioned a close relationship between *process* and *progress*. For them, this relationship is indicated by the macro-process we characterize as evolution. At every level of world history -- the cosmic, the biological, the social, the intellectual -- process philosophers have envisioned a developmental dynamic in which later is *better* -- somehow superior in being more differentiated and sophisticated. Under the influence of Darwinian evolutionism, most process philosophers have envisioned a course of temporal development within which value is somehow survival-facilitative so that the arrangements which do succeed in establishing and perpetuating themselves will as a general tendency manage to have done so because they represent actual improvements in one way of another. (A decidedly optimistic tenor has prevailed throughout process philosophy.)

After all, differentiation is sophistication; detail is enrichment. The person who merely sees a bird does not see as much as the person who sees a finch, and she in turn does not see as much as the person who sees a Darwin finch. The realization and enhancement of detail bestows not just complexification as such but also sophistication. As process philosophy sees it, the world's processuality involves not only change but improvement -- the evolutionary realization -- at large and on the whole -- of what is not only different but also in some way better. Accordingly, novelty and fruitfulness compensate for transiency and mortality in process philosophy's scheme of things.

4. An Instructive Application

Recourse to process is a helpful device for dealing with the classical problem of universals. We are surrounded on all sides by items more easily conceived of as processes than as substantial things -- not only physical items like a magnetic field or an *aurora borealis*, but also conceptual artifacts like letters of the alphabet, words, and statements. That purported universal -- the opening line of a play, say, or a shade of phenomenal red -- now ceases to be a mysterious *object* of some sort and becomes a specifiable feature of familiar processes (readings, perceivings, imaginings). How distinct minds can perceive the same universal is now no more mysterious than how distinct walkers can share the same limp -- it is a matter of actions proceeding in a certain particular way. Since processes are structural in nature, universals are now pulled down from the Platonic realm to become generic features of the ways in which we concretely conduct our cognitive affairs.

The philosophy of mind is another strongpoint of process philosophizing. It feels distinctly uncomfortable to conceptualize *people* (persons) as *things* (substances) -- oneself above all -- because we resist flat-out identification with our bodies. However, there is no problem with experiential access to the processes and patterns of process that characterize us personally -- our doings and undergoings, either individually or patterned into talents, skills, capabilities, traits, dispositions, habits, inclinations, and tendencies to action and inaction are, after all, what characteristically define a person as the individual he or she is. Once we conceptualize the core "self" of a person as a unified manifold of actual and potential process -- of action

and capacities, tendencies, and dispositions to action (both physical and psychical) -- then we thereby secure a concept of personhood that renders the self or ego experientially accessible, seeing that experiencing itself simply *consists* of such processes. What makes my experience mine is not some peculiar qualitative character that it exhibits but simply its forming part of the overall ongoing process that defines and constitutes my life. The unity of person is a unity of experience -- the coalescence of all of one's diverse micro-experience as part of one unified macro-process. (It is the same sort of unity of process that links each minute's level into a single overall journey.) On this basis, the Humean complaint -- "One experiences feeling this and doing that, but one never experiences *oneself*" -- is much like the complaint of the person who says "I see him picking up that brick, and mixing that batch of mortar, and troweling that brick into place, but I never see him building a wall." Even as "building the wall" just exactly *is* the complex process that is *composed* of those various activities, so -- from the process point of view -- one's self just *is* the complex process *composed* of those various physical and psychic experiences and actions in their systemic interrelationship.

5. Diversity/Complexity

Like any philosophical tendency -- realism, idealism, materialism, etc. -- process philosophy is a fundamentally prismatic complex and has internal variations. The difference at issue is rooted in the issue of what type of process is taken as paramount and paradigmatic. Some contributors (especially A. N. Whitehead and Henri Bergson) see organic processes as central and other sorts of processes as modeled on or superengrafted upon them -- the conception of an all-integrating physical field being pivotal even for Whitehead's organic/biological reflections. Others (especially William James) based their ideas of process on a psychological model and saw human thought as idealistically paradigmatic. Methodologically, on the other hand, some (e.g. Whitehead) articulated their process philosophy in essentially scientific terms, while others (esp. Bergson) relied more on intuition and indeed an almost mystical sort of sympathetic apprehension. And then too, of course, there are cultural processists like John Dewey. But such differences notwithstanding there are family-resemblance commonalties of theme and emphasis that nevertheless leave the teachings of the several processists in the position of variations on a common approach. So in the end it is -- or should be -- clear that the unity of process philosophy is not doctrinal but thematic; it is not a consensus or a thesis but rather a mere diffuse matter of type and approach.

Accordingly, process philosophy as such is something rather schematic. There are distinct approaches to implementing its pivotal idea of the pervasiveness and fundamentality of process, ranging from a materialism of physical processes (as with Boscovitch) to a speculative idealism of psychic processes (as in some versions of Indian philosophy). There are rather different ways of being a process philosopher, varying drastically according to the nature of one's ideas regarding what process is all about. In historical perspective, process philosophy has accordingly run a somewhat meandering course that traces back more to the origins of philosophy in the days of Pre-Socratic philosophy.

As such considerations indicate, the process approach has many assets. But it has significant liabilities as well. It is not unfair to the historical situation to say that process philosophy at present remains no more

than a glint in the mind's eye of various philosophers. A full-fledged development of the process doctrine simply does not yet exist as an accomplished fact, its development to the point where it can be compared with other major philosophical projects like materialism or absolute idealism still remains to be realized.

The process approach has been a particularly important development in and for American philosophy -- especially owing to its increasingly close linkage to pragmatism in such thinkers as Peirce, James, and Dewey. In recent decades the great majority of its principal exponents have done their philosophical work in the United States, and it is here that interest in this approach to philosophy has been the most intense and extensive, constituting a considerable sub-sector within American philosophy at large. Like American philosophy in general, process philosophy is too complex and diversified an enterprise to be captured or even dominated by any one school of thought; it is a highly diversified manifold that encompasses tendencies of thought representing a wide variety of sources.

Regrettably, authors of histories and surveys not infrequently fail to give process philosophy the recognition that is its due. For example, the otherwise excellent survey of American philosophy by the able French scholar Gerard Deledalle omits all mention of process philosophy as such and takes only perfunctory notice of Whitehead in an Appendix. To take this line is not, perhaps, to give us Hamlet without the ghost, but is at least tantamount to omitting Horatio.

6. Strawson's Critique

P. F. Strawson has argued in his influential book on metaphysics that processism in all its versions is doomed to failure because physical objects -- and, in particular, material bodies -- are requisites for the idea of identifiable particulars in a way that is virtually indispensable to any viable metaphysical position. Strawson maintains that the identification of particulars in communication between speakers and hearers ("referential identification" as he terms it) necessarily requires reference to things possessed of material bodies, so that "we find that material bodies play a unique and fundamental role in particular identification" (*Ibid*., p. 56). As he sees it, processes will not do as basis for particular identification because: "If one had to give the spatial dimensions of such a process, say, [as] a death or a battle, one could only have the outline of the dying man or indicate the extent of the ground the battle was fought over" *Ibid*., p. 57. Strawson accordingly holds that material bodies are a necessary precondition for any setting in which objective knowledge of particulars is to be possible.

In brief outline, Strawson's argument runs essentially as follows:

1. For objective and identifiable particulars to be knowable, some items must be (1) distinguishable from other co-existents, and (2) reidentifiable over time.
2. These conditions (viz. distinguishability and reidentifiability) can only be met by material objects (i.e. particulars with material bodies).

If this line of reasoning is indeed correct, processism is untenable in metaphysics. For it is perfectly clear that any viable metaphysic must have room for identifiable particulars, and if these are to be had only on

the basis of a material-object substantialism then process metaphysics is a lost cause.

This argumentation, however, has its problems. To begin with, Strawson would have been well advised to add yet a third item, viz. (3) that individuals must not only be distinguishable and reidentifiable by a *particular* knower, but interpersonally and intersubjectively distinguishable and reidentifiable throughout a *community* of knowers. Yet even with premiss 1 strengthened in this way, premiss 2 does not hold water.

Strawson maintained that:

> The only objects which can constitute [the space-time framework essential to interpersonal communication] are those which confer upon it their own fundamental characteristics. That is to say they must be three dimensional objects with some endurance through time . . . They must collectively have enough diversity, richness, stability, and endurance to make possible just that conception of a single unitary [space-time] framework which we possess. [Page 39]

The process philosopher will have no quarrel with any of this. However, Strawson then proceeded straightaway to draw a deeply problematic conclusion:

> Of the categories of object which we recognize, only those satisfy these requirements which are, or possess, material bodies -- in the broad sense of the expression. Hence given a certain general feature of the [space-time committed] conceptual scheme which we possess, and given the character of available major categories, things which are, or possess, material bodies must be [epistemologically] basic particulars.

To its decisive detriment, Strawson's argument simply begs the question here. For all of the features that his analysis require (spatiotemporal stability and endurance, diversity, richness, interpersonal accountability, and the like) are possessed every bit as much by *physical processes* as by the things that "are or possess *material bodies*." It is not material substances (*things*) that can be distinguished and reidentified within nature's spatiotemporal framework, but occurrence-contexts (processes) as well. Processes are physically realized without being literally embodied. And the one is no less confrontable and capable of ostensive indication than the other ("that lion"; "that yawning"). Only by an act of deeply problematic fiat is Strawson able -- even within the restricted confines of his own analysis -- to advantage and prioritize material bodies over physical processes. Even Strawson's insistence that epistemically basic particulars must be identifiable by ostension holds every bit as much for instances of physical process as for particular *constrained* material bodies. (Indeed, as we shall see, it is theoretically possible to reconceptualize material bodies as complexes of physical processes, while the reverse -- the general reconceptualization of physical processes as complexes of material objects -- is just not all that plausible (the "Reism" or "Concretism" of Kotarbinski and of the later Brentano notwithstanding).)

Strawson's reasoning sets out from the quite appropriate Kantian observation that objective

distinguishably and reidentifiability requires the machinery of a spatiotemporal matrix for the emplacement of our experiential encounters with objects in a unified all-encompassing framework of coordination viz. space-time. But at this point his reasoning goes astray. For as he sees it, a spatiotemporal framework demands -- and can only be determined in terms of -- ordering relations among material objects. But there are in fact other physically "embodied" items distinct from material bodies that can serve this function equally well -- to wit, *processes*. For as long as processes have both position and duration -- as long as like a flame (rather than a sound) or a wedding ceremony (rather than something more ethical like a divorcement) -- there are items that have a sufficiently definite place and a sufficiently long lifespan to serve as coordinate markers. Processes too, in sum, can serve to define and constitute the required spatio-temporal framework.

Strawson's position is plausible only because he accepts the question-begging Process Reducibility Thesis that insists on seeing all processes in terms of the activities of things (substances). From this standpoint, all processes are owned and we are to look at them from a specifically genitive point of view: the death *of* Caesar, or the great clash *of* the armies of Napoleon and Tsar Alexander I at Borodino. But this of-indicated object-correlativity (*of* that person, *of* these two armies) takes too narrow a view of the matter. It reflects only the particular (i.e. owned) sort of processes at issue, and not their processuality as such. Where processes are more basically concerned, their object-correlativity can disappear from view.

The point is that while we can indication-identify various concrete processes *genitively* -- as per "this birth" = "the birth of Julius Ceasar" -- proceeding in terms of process-type plus substance-correlative possession, we can no less easily in dualism identify them *positionally* in terms of process-type plus location: "this birth" = "the birth at such-and-such a space-time location." And of course the referential markers that orient us in space-time need not be substantial (the town center of Greenwich) but can be processual (the pole = the place where the compass needle spins around evenly).

Accordingly, Strawson's argumentation misses its target. It is simply not the case that material objects are the indispensable basis for a framework of knowable particulars. Physical processes of a suitable sort can accomplish this essential task equally well.

7. Processes and Stability

As process ontologists see it, enduring things are never more than patterns of stability in a sea of process. Like a wave pattern in water they are simply pending configurations in a realm of change.

The very idea of a process involves trans-temporal constancies. Water evaporates. That is to say, the evaporation of water is a generic process. It has many instances, occurring alike after rainstorms in 16th century Lima and in 20th century Atlanta. Any and every particular process is always an instantiation of a general pattern. One just simply cannot identify a process that fails to be of a (processual) *type* and which, in consequence, is not -- at that level of abstraction -- capable of repetition. And so the concreta of history, viewed in an epistemic perspective, can in fact manage to transcend their space-time settings to instantiate general patterns. Although their manifestations are inevitably temporal and concrete, those

processes themselves can be atemporal and generic.

And of course different concrete instances of a process can produce products of exactly the same generic type. Different factories can and often do produce the same model of car, different cooks can and do produce the same variety of soup. And this is strikingly so when the product happens to be information: different presses can print the same text, different respondents can give the same answer to the same question, different mouths can utter the same sentence, different minds can entertain the same idea.

The point is that in the realm of informational abstractness products can escape the limitations of their (invariably relativized) productive origins. The historical relativization of the production *process* to a particular historico-cultural context -- the fact that the thinking or the assertion of a truth is so relativized -- of itself does nothing to limit the *product* (the truth that is so thought or asserted) to a historico-cultural context. Once produced, it is generally available -- and (insofar as abstract) will be cross-temporally accessible via its exemplifications and manifestations at different times and places.

Some sorts of things exist out of space but not time -- one's ownership of a piece of jewelry, for example, or one's right to exercise an option to purchase a tract of land. Other sorts of things exist neither in space nor in time -- numbers, facts, and generalized relationships for example. (The Eiffel Tower was erected in Paris in the 19th century, but the fact that Julius Caesar did not realize this is something that has no spatiotemporal emplacement.) And information is like that. The things that information may be about may be spatiotemporal, as will be the speech or writing by which the information is conveyed from one person to another. But the information itself is altogether nonspatiotemporal. It simply lies in the nature of certain sorts of things, information included, not to be located in space and time -- to be "abstract."

Admittedly, when we are viewing something, the only views we can possibly obtain are views from somewhere (and from view-points belonging to us and not to God). But when the viewing is done with the eyes of the mind, and its object is the realm of information rather than the realm of physical reality then what the view is a view of is something ahistorical. For information as such exists outside of history even though our acquiring it is invariably an historical transaction. We must avoid the category mistake of confusing process with product here: of conflating the information that we access with the historical actions and events of our accessing it.

Of course we have no way to get to the abstract (the belief) save via the historical (the believing). But what we achieve (the product) is something of a nature different and status distinct from the mode of its realization (the process). When we engage ourselves in intellectual processes that carry us into the informational domain we impel ourselves from history into an ahistorical sphere. The same idea (the same thought-process, the same belief) is accessible to people at different times and places. Were it not so, communication would be altogether impossible.

The overall situation in matters of abstraction is triadic (to use the term favored by C. S. Peirce). There are: (i) the various and sundry concrete green things; (ii) the abstract property at issue (viz., the property or characteristic of being green); and (iii) the mediative conception or idea of greenness which is the thought-instrumentalility through which that abstract property comes to be imputed to those concrete

items that putatively manifest it. The medieval metaphysical dispute between nominalism, conceptualism, and platonism needs to be resolved *conjunctively*: all three are needed: a nominalism is required for concrete particulars, a conceptualism for particular-applicative concepts, and a platonism for abstractions (e.g., in prime mathematics). The situation is not one of either/or; we must endorse all those doctrinal positions -- each in its own place.

Yet how can temporalized thought deal in timeless information? How is it that particularized episodic thought can make episode-abstractive generalizations? The long and short of it is that that’s just how thought works. To puzzle about this is like puzzling about any of the world’s brute facts. And once those realities are taken in stride the problem has been left behind. One might as well ask "How is it that money can be used to buy things? or that words can be used for speaking?" No matter how much we may wonder at the phenomena we have to accept them as part of the world’s realities.

There indeed are fundamental problems that lie in the background here: how standardized exchange is possible or how verbal communication is possible. But once such fundamental background issues are resolved, the original question is dissolved as such: something that is not a medium of exchange would not be called *money*, nor would something that could not play a generalized role in verbal or written communication be called a *word*. Even so something would not be called thought if it could not function abstractly to convey general information transcending the episodic occurrences at issue.

8. Quantum Issues

As Whitehead’s own reaction shows, the rise of the quantum theory put money in process philosopher’s bank account. The classical conception of an atom was predicated on the principle that "by definition, atoms cannot be cut up or broken into smaller parts," so that "atom-splitting" was, from the traditional point of view, simply a *contradiction in terms*. Here the demise of classical atomism brought on by the dematerialization of physical matter in the wake of the quantum theory did much to bring aid and comfort to a process-oriented metaphysics. For quantum theory taught that, at the microlevel, what was usually deemed a physical *thing*, a stably perduring object, is itself no more than a statistical pattern -- a stability wave in a surging sea of process. Those so-called enduring "things" come about through the emergence of stabilities in statistical fluctuations.

The quantum view of the world is inherently probabilistic -- indeed it has trouble coming to terms with concrete definiteness (with the "collapse of the wave packet" problem). And this too is congenial to processists, seeing that process philosophy rejects a pervasive determinism of law-compulsion. Processists see the laws of nature as imposed from below rather than above -- as servants rather than masters of the world’s existents.

Twentieth century physics has thus turned the tables on classical atomism. Instead of very small *things* (atoms) combining to produce standard processes (windstorms and such), modern physics envisions very small processes (quantum phenomena) combining in their *modus operandi* to produce standard things (ordinary macro-objects). The quantum view of reality has accordingly led to the unravelling of that

classical atomism that has, from the start, been paradigmatic for substance metaphysics.

Process metaphysics envisions a limit to determinism that makes room for creative spontaneity and novelty in the world (be it by way of random mutations with naturalistic processists or purposeful innovation with those who incline to a theologically teleological position).

Moreover, process philosophers have reason to favor quantum physics over relativistic physics. For relativity sees space time as a block that encompasses all real events concurrently, leaving the time differentiation of earlier-later to be supplied from the subjective resources of observers relative to their own mode of emplacement within the grand scheme of things. Special relativity with its preoccupation with time-invariant relationships in effect suppresses time as a factor in physical reality and relegates it to the penumbral status of a subjective phenomenon. This serves to explain why Whitehead sought to provide a new theoretical basis to relativity theory and reconstrue space-time, as well as the conception of other physical objects, as being a construction made from "fragmentary individual experiences." Processes are not the machinations of stable things; things are the stability-patterns of variable processes. All such perspectives of modern physics at the level of fundamentals dovetail smoothly into the process approach.

9. Process Theology

The God of scholastic Christian theology, like the deity of Aristotle on whose model this conception was in part based, is an immaterial individual, located outside of time -- entirely external to the realm of change and process. By contrast, process theologians, however much they may disagree on other matters, take the radical (but surely not heretical) step of according God an active role also *within* the natural world's spatio-temporal frame. They envision a foothold for God within the overall processual order of the reality that is supposed to be his creation. After all, active participation in the world's processual commerce need not necessarily make God into a physical or material object. (While the world indeed contains various physical processes like the evolution of galaxies, it also contains immaterial processes such as the diffusion of knowledge or the emergence of order.)

For process theology, then, God does *not* constitute part of the world's making of physical processes, but nevertheless in some fashion or other *participates* in it. Clearly no ready analogy-model for this mode of participation (spectator, witness, judge, etc.) can begin to do full justice to the situation. But what matters first and foremost to the angle of process theology is *the fact that* God and his world are processually inter-connected -- the issue of *the manner how* is something secondary that can be left open for further reflection. So conceived, God is not exactly *of* the world of physical reality, but does indeed participate *in* it processually -- everywhere touching, affecting, and informing its operations. Thus while not emplaced in the world, the processists' God is nevertheless bound up with it in an experiential process of interaction with it. In general, process theists do not believe that God actually controls the world. The process God makes an impact persuasively, influencing but never unilaterally imposing the world's process.

Process theology accordingly invites us to think of God's relationship to the world in terms of a process

of influence like "the spread of Greek learning in medieval Islam." Greek learning did not become literally *internal* to the Islamic world, but exerted a substantial and extensive influence upon and within it. Analogously, God is not of the world but exerts and extends an all-pervasive influence upon and within it. After all, processes need not themselves be spatial to have an impact upon things in space (think of a price inflation on the economy of a country.) The idea of process provides a category for conceptualizing God's relation to the world that averts many of the difficulties and perplexities of the traditional substance paradigm.

Even apart from process philosophy, various influential theologians have in recent years urged the necessity and desirability of seeing God not through the lens of unchanging stability but with reference to movement, change, development, and process. But, the process theorists among theologians want to go beyond this. For them, God is not only to be related to the world's processes in a productive manner, but must himself be regarded in terms of process -- as encompassing processuality as a salient aspect of the divine nature.

To be sure, process theologians differ among themselves in various matters of emphasis. Whitehead sees God in cosmological terms as an "actual occasion" functioning within nature, reflective of "the eternal urge of desire" that works "strongly and quietly by love," to guide the course of things within the world into "the creative advance into novelty." For Hartshorne, by contrast, God is less an active force within the world's processual commerce than an intelligent being or mind that interacts with it. His God is less a force of some sort than a personal being who interacts with the other mind-endowed agents through personal contact and love. Hartshorne wants neither to separate God from the world too sharply nor yet to have him be pantheistically immanent in nature. He views God as an intelligent world-separated being who participates experientially in everything that occurs in nature and resonates with it in experiential participation.

Such differences of approach, however, are only of secondary importance. The crucial fact is that the stratagem of conceiving of God in terms of a *process* that is at work in and beyond the world makes it possible to overcome a whole host of substance-geared difficulties at one blow. For it now becomes far easier to understand how God can be and be operative. To be sure, the processual view of God involves a recourse to processes of a very special kind. But extraordinary (or even supra-natural) *processes* pose far fewer difficulties than extraordinary (let alone supra-natural) *substances*, seeing that process is an inherently more flexible conception. After all, many sorts of processes are in their own way unique -- or, at any rate, radically different from all others. Clearly, processes like the creation of a world or the inauguration of its lawful order are by their very nature bound to be unusual, but much the same can be said of any particular type of process. Moreover, through its recourse to the idea of a mega-process that embraces and encompasses a variety of subordinate processes, process theology is able to provide a conceptual rationale for reconciling the idea of an all-pervasive and omnitemporal mode of reality with that of a manifold of finitely temporalized constituents.

The processist view of nature as a spatiotemporal whole constituting one vast, all-embracing cosmic process unfolding under the directive aegis of a benign intelligence is in various ways in harmony with the Judeo-Christian view of things. For this tradition has always seen God as active within the historical

process which, in consequence, represents not only a causal but also a purposive order. After all, the only sort of God who can have meaning and significance for us is one who stands in some active interrelationship with ourselves and our world. (Think here of the Nicene creed's phraseology: "the maker of all things … who for us men and for our salvation …".) But of course such an "active interrelationship" is a matter of the processes that constitute the participation and entry of the divine into the world's scheme of things -- and conversely.

And of course not only is it feasible and potentially constructive for the relation of God to the world and its creatures to be conceived of in terms of processes, but it is so also with the relationship of people to God. Here too process theology sees such a relationship as thoroughly processual because it rests on a potentially interactive communion established in contemplation, worship, prayer, etc.

In particular, for processists there is little difficulty in conceiving God as a *person*. For once we have an account of personhood in general in process terms as a systemic complex of characteristic activities, it is no longer all that strange to see God in these terms as well. If we processify the human person, then we can more readily conceive of the divine person as the focal source of a creative intelligence that engenders and sustains the world and endows it with law, beauty (harmony and order), value and meaning.

Then too there is the problem of the Trinity with its mystery of fitting three persons into one being or substance, which has always been a stumbling block for the substantialism of the Church Fathers. A process approach makes it possible to bypass this perplexity. For processes can interact and interpenetrate one another. With the laying of a single branch a woodsman can be building a wall, erecting a house, and extending a village. One act, many processes; one mode of activity many sorts of agency.

For process theology, then, God is active in relation to the world, and the world's people can and should be active in relation to God. People's relationship to the divine is a two-way street, providing for a benevolent God's care for the world's creatures and allowing those intelligent beings capable of realizing this to establish contact with God through prayer, worship, and spiritual communion. Process theology accordingly contemplates a wider realm of processes that embrace both the natural and the spiritual realms and interconnect God with the vast community of worshippers in one communal state of macroprocess that encompasses and gives embodiment to such a comprehensive whole.

To be sure, process theologians usually see the divine as one power among others and view God's role in relation to the world as rather diffused and indirect and limited. But this seems to be more because a novel perspective appeals to those of theologically liberal and unorthodox orientation than to the inherent demands of a process appraisal. In theory a process theology could take a more theologically conservative form than has been the case.

10. A Question of Legitimacy

From the days of the Pyrrhonian sceptics of antiquity we are told again and again throughout the history

of philosophy that speculative systematization is inappropriate -- that such knowledge as we humans can actually obtain is limited to the realm of everyday life and/or its precisification through science. Repeated in every era, this stricture is also rejected by many within each. The impetus for big-picture understanding, for a coherent, and panoramic view of things that puts the variegated bits and pieces together, represents an irrepressible demand of the human intellect as a possession of "the rational animal." And process metaphysics affords one of the most promising and serious options for accommodating this demand.

11. Institutionalization

Process thought constitutes one (albeit only one) very prominent sector of the active philosophical scene in the USA at the present time. Apart from the proliferation of books and articles on the topic, it has achieved considerable institutionalization during the years after World War II. Indications of this phenomenon include the formation of the Society for Process Studies, as well as the prominence of process philosophizing within the aegis of the Society for American Philosophy and the American Metaphysical Society. Another clear token is the journal *Process Studies*, Published by the Center for Process Studies in Claremont CA, and founded in 1971 by Lewis S. Ford and John B. Cobb, Jr., a publication that has in recent years become a major vehicle for article-length discussions in the field. Representatives of process philosophy occupy influential posts in departments of philosophy and of religious studies in many of American universities and colleges, and some half-dozen doctoral dissertations are produced annually in this field. American philosophy is at this historic juncture an agglomeration of different cottage industries, and process philosophy is prominent among them.

Bibliography

- Browning, Douglas, *Philosophers of Process* (New York: Random House, 1965).
- Cobb, John B., *A Christian Natural Theology* (Philadelphia: Westminster Press, 1965).
- Cobb, John B. and David R. Griffin, *Process Theology: An Introductory Exposition* (Philadelphia, Westminster Press, 1976).
- -----, *Process Theology as Political Ecology* (Philadelphia, Westminster Press, 1982).
- Gray, James R., *Modern Process Thought* (Lanham, MD.: University of America, 1982).
- Hartshorne, Charles, "Contingency and the New Era in Metaphysic," *Journal of Philosophy*, vol. 29 91932), pp. 421-431 and 457-469.
- -----, *Creative Synthesis and philosophic Method* (La Salle, IL.: Open Court, 1970).
- -----, "The Development of Process Philosophy," in *Process Theology*, ed. Ewert H. Cousins (New York, Newman Press, 1971).
- -----, *The Divine Relativity: A Social Conception of God* (New Haven: Yale University Press, 1948).
- -----, *A Natural Theology for Our Time* (La Salle, IL.: Open Court, 1967).
- -----, *Whitehead's Philosophy: Selected Essays, 1935-1970* (Lincoln, NB.: University of Nebraska Press, 1972).

- Lucas, George R. Jr., *Two View of Freedom in Process Thought: A Study of Hegel and Whitehead* (Missoula, MN: Scholar's Press, 1979).
- -----, *The Genesis of Modern Process Thought* (Metuchen, NJ.: Scarecrow Press, 1983).
- -----, *Hegel and Whitehead: Contemporary Perspectives on Systematic Philosophy* (Albany: SUNY Press, 1986).
- -----, *The Rehabilitation of Whitehead: An Analytical and Historical Arsenal of Process Philosophy* (Albany, NY.: SUNY Press, 1989).
- Palter, Robert M., *Whitehead's Organic Philosophy of Science* (Albany, NY.: SUNY Press, 1979).
- Rescher, Nicholas, *Process Metaphysics: An Introduction to Process Philosophy* (New York: SUNY Press, 1996).
- -----, *Process Philosophy: A Survey of Basic issues* (Pittsburgh, Pa.: University of Pittsburgh Press, 2000).
- Seibt, Johanna, *Properties as Processes: A Synoptic Study of W. Sellars' Nominalism* (Reseda, CA.: Ridgeview, 1990).
- Strawson, P. F., *Individuals* (London: Methuen, 1959).
- Whitehead, A. N., *An Enquiry Concerning the Principles of Natural Knowledge* (Cambridge: Cambridge University Press, 1919; reprinted New York: Kraus Reprints, 1982).
- -----, *The Concept of Nature* (Cambridge: Cambridge University Press, 1920).
- -----, *The Principle or Relativity* (Cambridge: Cambridge University Press, 1922).
- -----, *Science and the Modern World* (New York: Macmillan, 1925).
- -----, *Religion in the Making* (New York: Macmillan, 1926).
- -----, *Process and Reality: An Essay in Cosmology* (New York: Macmillan, 1929). Critical edition by D. R. Griffin and D. W. Sherbourne (New York: Macmillan, 1978).
- -----, *Symbolism: Its Meaning and Effect* (New York: Macmillan, 1927; reprinted New York: G. P. Putnam's Sons, 1959).
- -----, *The Function of Reason* (Boston: Beacon Press, 1929).
- -----, *Adventures of Ideas* (New York: Macmillan, 1933).
- -----, *Nature and Life* (Cambridge: Cambridge University Press, 1934).
- -----, *Modes of Thought* (New York: Macmillan, 1938).
- -----, *Essays in Science and Philosophy* (New York: Philosophical Library, 1948).
- Whittemore, Robert C., (ed.), *Studies in Process Philosophy* (New Orleans: Tulane University Press, 1974).
- -----, *Studies in Process Philosophy, II* (New Orleans: Tulane University Press, 1976).
- -----, *Studies in Process Philosophy, III* (New Orleans: Tulane University Press, 1975).

Other Internet Resources

- Center for Process Studies
- Process Philosophy and the New Thought Movement (C. Alan Anderson, Emeritus, Curry College)
- Process Philosophy Bibliography (Ronald Tobey, History, UC/Riverside)

Related Entries

evolution | God, nature of | Hartshorne, Charles | process | quantum theory | Whitehead, Alfred North

Nicholas Rescher
rescher+@pitt.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 2, 2002
Content last modified: April 2, 2002

Charles Hartshorne

Charles Hartshorne is considered by many philosophers to be one of the most important philosophers of religion and metaphysicians of the twentieth century. Although Hartshorne often criticized the metaphysics of substance found in medieval philosophy, he was very much like medieval thinkers in developing a philosophy that was theocentric. Throughout his career he defended the rationality of theism and for several decades was almost alone in doing so among English-language philosophers. Hartshorne was also one of the thinkers responsible for the rediscovery of St. Anselm's ontological argument. But his most influential contribution to philosophical theism did not concern arguments for the *existence* of God, but rather was related to a theory of the *actuality* of God, i.e., *how* God exists. In traditional or classical theism, God was seen as the supreme, unchanging being, but in Hartshorne's process-based or neoclassical conception, God is seen as supreme becoming in which there is a factor of supreme being. That is, we humans become for a while, whereas God *always* becomes, Hartshorne maintains. The neoclassical view of Hartshorne has influenced the way many philosophers understand the concept of God. In fact, a small number of scholars--some philosophers and some theologians--think of him as the greatest metaphysician of the second half of the twentieth century, yet, with a few exceptions to be treated below, his work has not been very influential among analytic philosophers who are theists.

Life

Charles Hartshorne was born in the nineteenth century and lived to philosophize in the twenty-first. He was born in Kittanning, Pennsylvania (U.S.A.) on June 5, 1897. He was, like Alfred North Whitehead, the son of an Anglican minister, although many of his ancestors were Quakers. After attending Haverford College he served in World War One in France as a medic, taking a box of philosophy books with him to the front. After the war Hartshorne received his doctorate in philosophy at Harvard, and there he met Whitehead. Most of the major elements of Hartshorne's philosophy were already apparent by the time he

became familiar with Whitehead's thought, contrary to a popular misconception. From 1923-1925 a postdoctoral fellowship took him to Germany, where he had classes with both Husserl and Heidegger. But neither of these thinkers influenced his philosophy as much as C.S. Peirce, whose collected papers he edited with Paul Weiss. In addition to many visiting appointments, Hartshorne spent his teaching career at three institutions. From 1928-1955 he taught at the University of Chicago, where he was a dominant intellectual force in the School of Divinity, despite the fact that he was housed in the Philosophy Department, where he was not nearly as influential. He was at Emory University from 1955 until 1962, when he moved to the University of Texas at Austin. Hartshorne eventually became a long-term emeritus professor at Austin and lived there until his death on October 9, 2000. His wife, Dorothy, was as colorful as her husband and was mentioned often in his writings. Hartshorne never owned an automobile, nor did he smoke or drink alcohol or caffeine; he had a passion for birdsong and became an internationally known expert in the field.

Method

Three primary methodological devices or procedures are at work in Hartshorne's metaphysics. First, he very often uses a systematic exhaustion of theoretical options--or the development of position matrices, sometimes containing thirty-two alternatives(!)--in considering philosophical problems. This procedure is evident throughout his philosophy, but it is most apparent in his various treatments of the ontological argument. To take another example, he thinks it important to notice that regarding the mind-body problem there are three options available to us, not two, as is usually assumed: some form of dualism, some form of the materialist view that psyche is reducible to body, *and* some form of the panpsychist (or, as he terms it, psychicalist) view that body is in some way reducible to psyche if all concrete singulars (e.g., cells or electrons) in some way show signs of self-motion or activity. More recently, Thomas Nagel considers this third option, but Hartshorne actually defends it.

Second, Hartshorne frequently uses the history of philosophy to see which of the logically possible options made available by position matrices have been defended before so as to avail ourselves of the insights of others in the effort to examine in detail the consistency of these positions and to assess their consequences. Nonetheless, those logically possible options that have not historically found support should be analyzed both in terms of internal consistency and practical ramifications. It should be noted that Hartshorne's use of the history of philosophy often involves lesser known views of famous thinkers (like Plato's belief in God as the soul for the body of the whole natural world, or Leibniz's defense of panpsychism) as well as the thought of lesser known thinkers (such as Faustus Socinus, Nicholas Berdyaev or Jules Lecquier).

Third, after a careful reading of the history of philosophy has facilitated the conceptual and pragmatic examination of all the available options made explicit by a position matrix, the (Greek) principle of moderation is used by Hartshorne as a guide to negotiate the way between extreme views on either side. For example, regarding the issue of personal identity, the view of Hume (and of Bertrand Russell at one stage in his career) is that, strictly speaking, there is no personal identity in that each event in "a person's life" is externally related to the others. This is just as disastrous, Hartshorne thinks, as Leibniz's view that

all such events are internally related to the others, so that implicit in the fetus are all the experiences of the adult. (This Leibnizian view relies on the classical theistic, strong notion of omniscience, wherein God knows in minute detail and with absolute assurance what will happen in the future.) The Humean view fails to explain the continuity we experience in our lives and the Leibnizian view fails to explain the indeterminateness we experience when considering the future. The truth lies between these two extremes, Hartshorne thinks. His view of personal identity is based on a conception of time as asymmetrical in which later events in a person's life are internally related to former events, but they are externally related to those that follow, thus leading to a position that is at once partially deterministic and partially indeterministic. That is, the past supplies necessary but not sufficient conditions for human identity in the present, which always faces a partially indeterminate future.

Only the first of these methodological devices or procedures supports the widely held claim that Hartshorne is a rationalist. His overall method is a complex one that involves the other two methods or procedures, where he does borrow from the rationalists, but also from the pragmatists and the Greeks. It must be admitted, however, that Hartshorne was educated in a philosophic world still heavily influenced by late nineteenth and early twentieth century idealism.

The Existence and Actuality of God

Philosophers commonly use a metaphor that suggests that the chain of an argument, say for the existence of God, is only as strong as its weakest link. Hartshorne rejects this metaphor on Peircian grounds. He replaces it by suggesting that various arguments for the existence of God--ontological, cosmological, design, etc.--are like mutually reinforcing strands in a cable.

He argues that Hume's and Kant's criticisms of the ontological argument of St. Anselm are not directed at the strongest version of his argument found in *Proslogion*, chapter 3. Here, he thinks, there is a modal distinction implied between existing necessarily and existing contingently. Hartshorne's view is that existence alone might not be a real predicate, but existing necessarily certainly is. That is, contra Kant and others, Hartshorne believes that there are necessary truths concerning existence. To say that absolute nonexistence in some fashion exists is to contradict oneself; hence he thinks that absolute nonexistence is unintelligible. It is necessarily the case that *something* exists, he thinks, and, relying on the ontological argument, he also thinks it necessarily true that God exists.

On Hartshorne's view, metaphysics does not deal with realities beyond the physical, but rather with those features of reality that are ubiquitous or that would exist in any possible world. And he does not think that it is possible to think of a preeminent being that only existed contingently since if it did exist contingently rather than necessarily, it would not be preeminent. That is, God's existence is either impossible (positivism) or possible, and, if possible, then necessary (theism). He is assuming here that there are three alternatives for us to consider: (1) God is impossible; (2) God is possible, but may or may not exist; (3) God exists necessarily. The ontological argument shows that the second alternative makes no sense. Hence, he thinks that the prime task for the philosophical theist is to show that God is not impossible.

In addition, Hartshorne's detailed treatment of the argument from design is connected to his view of biology. It is hard to reconcile an omnipotent, classical theistic God with all of the monstrosities and chance mutations produced in nature, but the general orderliness of the natural world is just as difficult to reconcile with there being no Orderer or Persuader at all. Belief in God as omnipotent, he thinks, has three problems: (1) it is at odds with the disorderliness in nature; (2) it yields the acutest form of the theodicy problem; and (3) it conflicts with the notion from Plato's *Sophist*, defended by Hartshorne, that being *is* dynamic power (*dynamis*). An *omni*potent being would ultimately have all power over others, who would ultimately be powerless. But any being-in-becoming, according to Hartshorne, has *some* power to affect, or to be affected by, others; this power, however slight, provides counterevidence to a belief in divine omnipotence. In contrast, God is ideally powerful, on the Hartshornian view. That is, God is as powerful as it is possible to be, given the partial freedom and power of creatures.

Hartshorne's dispute with traditional or classical philosophical theism concerns not so much the *existence* of God, but rather its assumption that the *actuality* of God (i.e., *how* God exists) could be described in the same terms as the existence of God. A God who exists necessarily is not necessary or unchanging in every other respect (e.g., in terms of divine responsiveness to creaturely changes), he thinks. Although Hartshorne believes that the medieval thinkers were correct in trying to think through the logic of perfection, he also thinks that this logic has traditionally been misapplied in the effort to articulate the attributes of a being called "God," roughly defined as the greatest conceivable being. The traditional or classical theistic logic of perfection sees God as monopolar in that regarding various contrasts (permanence-change, one-many, activity-passivity, etc.) the traditional or classical philosophical theist has chosen one element in each pair as a divine attribute (the former element of each pair) and denigrated the other.

By way of contrast, Hartshorne's logic of perfection is dipolar. Within each element of these pairs there are good features that should be attributed in the preeminent sense to God (e.g., excellent permanence in the sense of steadfastness, excellent change in the sense of preeminent ability to respond to the sufferings of creatures). In each element in these pairs there are also invidious features (e.g., pigheaded stubbornness, fickleness). The task for the philosophical theist, he thinks, is to attribute the excellences of both elements of these pairs to God and to eschew the invidious aspects of both elements. However, it should be noted that *some* contrasts are not fit for dipolar analysis (e.g., good-evil) in that "good good" is a redundancy and "evil good" is a contradiction. The greatest conceivable being, he thinks, cannot be evil in any sense whatsoever.

Hartshorne does not claim to believe in two gods, nor does he wish to defend a cosmological dualism. In fact, we can see that the opposite is the case when we consider that, in addition to calling his theism *dipolar*, he refers to it as a type of *panentheism*, which literally means that all is *in* the one God by means of omniscience (as Hartshorne defines the term) and omnibenevolence. All creaturely feelings, especially feelings of suffering, are included in the divine life. God is seen by Hartshorne as the mind or soul for the whole body of the natural world (see above regarding Plato's World Soul), although he thinks of God as distinguishable from the creatures. Another way to categorize Hartshorne's theism is to see it as *neoclassical* in the sense that he relies on the classical or traditional theistic proofs for the existence of God and on the classical theistic metaphysics of being as *first steps* in the effort to think through properly

the logic of perfection. However, these efforts need to be supplemented, he thinks, by the efforts of those who see becoming as more inclusive than being. God is not outside of time, as in the Boethian view that is influential among traditional philosophical theists, but rather exists through all of time, on Hartshorne's view. On the neoclassical view, God's permanent "being" consists in steadfast benevolence exhibited through everlasting becoming.

God is omniscient, on Hartshorne's view, but "omniscience" here refers to the divine ability to know everything that is knowable: past actualities as already actualized; present realities to the extent that they are knowable according to the laws of physics (e.g., what is present epistemically may very well be the most recent past, given the speed of light); and future possibilities or probabilities *as possibilities or probabilities*. On the traditional or classical conception of omniscience, however, God has knowledge of future possibilities or probabilities as already actualized. According to Hartshorne, this is not an example of supreme knowledge, but is rather an example of ignorance of the (at least partially) indeterminate character of the future.

The asymmetrical view of time, common to process thinkers in general (e.g., Bergson, Whitehead, Hartshorne), in which the relationship between the present and the past is radically different from the relationship between the present and the future, also has implications for Hartshorne's theodicy. A plurality of partially free agents, including nonhuman ones, facing a future that is neither completely determined nor foreknown in detail, makes it not only possible, but likely, that these agents will get in each other's way, clash, and cause each other to suffer. On this view, God is the fellow sufferer who understands.

Axiology

Hartshorne views the cosmos as a "metaphysical monarchy," with God as the presiding, but not omnipotent, head, and he sees human society as a "metaphysical democracy," with each member as an equal. This makes him a liberal in politics if "liberalism" refers to the egalitarian belief that none of us is God. Although Hartshorne and Whitehead are both political liberals, Hartshorne is, despite his view of panpsychist reality as thoroughly social, more of a libertarian liberal and Whitehead more of a redistributive liberal. In axiology as well as in metaphysics/theodicy, freedom is crucial, on Hartshorne's view.

Hartshorne's panpsychism (or psychicalism) entails the belief that each active singular in nature, even those like electrons and plant cells that do not exhibit mentality, is nonetheless a center of intrinsic, and not merely instrumental, value. As a result, Hartshorne's metaphysics is meant to provide a basis for both an aesthetic appreciation of the value in nature, as well as for an environmental ethics where intrinsic and instrumental values in nature are weighed.

As a published expert on bird song, Hartshorne is the first philosopher since Aristotle to be an expert in both metaphysics and ornithology. He writes specifically of the aesthetic categories required to explain why birds sing outside of mating season and when territory is not threatened--two occasions for bird song

that are crucial to the behaviorists' account. Birds *like* to sing, he concludes.

Hartshorne's criticism of anthropocentrism is due not only to his concern for God, but also for beings-in-becoming who experience in a less sophisticated way than humans. To say that all active singulars feel--leaving out of the picture abstractions like "twoness" or insentient composites of active singulars that do not themselves feel as wholes--is not to say that they are self-conscious or that they think. As before, however, Hartshorne's axiology is ultimately theocentric in character.

Critical Evaluation

It seems fair to say that analytic philosophers in general, even analytic philosophers who are theists, have largely ignored Hartshorne's philosophy. This is in contrast to his wide influence among theologians, which is odd when it is considered that he is not a theologian and does not rely on sacred scripture or religious authority for his insights. Another oddity is the fact that Hartshorne's influence among theologians is due to the defense he offers of the *rationality* of belief in a neoclassical God.

There is one important philosopher whose work indicates the sort of debate that has occurred between Hartshorne and analytic theists, who tend to rely on traditional or classical versions of the concept of God. That is William Alston. There are two reasons why an evaluation of Hartshorne's philosophy is facilitated by a consideration of Alston's critique. First, Alston is as important a theist as any among analytic philosophers and his criticisms of Hartshorne's thought are like those of other analytic philosophers like Thomas Morris, Richard Creel, Michael Durrant, Colin Gunton, and others. And second, Alston is a former student of Hartshorne's and is thoroughly familiar with Hartshorne's arguments. Alston is a philosopher who is not scandalized by Hartshorne's panentheism, nor by his neoclassical theism. But Alston thinks that the contrast that Hartshorne draws between his neoclassical theism and the classical theism of Thomas Aquinas is too sharp.

Alston thinks that Hartshorne presents neoclassical theism and classical theism as complete packages, whereas it would be better to be able to pick and choose among individual items within these packages. Alston seeks some sort of rapprochement between Thomism and neoclassical theism, a rapprochement that Hartshorne himself would like to bring about to the extent that he is a neo*classical* thinker, but that is difficult to accomplish to the extent that he is *neo*classical.

A consideration of ten contrasting attributes will best facilitate an initial understanding of Hartshorne's view of God. Consider the first group of attributes treated by Alston.

CLASSICAL ATTRIBUTES	NEOCLASSICAL ATTRIBUTES
1. absoluteness (absence of internal relatedness)	1. relativity (God is internally related to creatures by way of knowledge of them and actions toward them)

2. pure actuality (there is no potentiality in God)	2. potentiality (not everything is actualized that is possible for God)
3. total necessity (every truth about God is necessarily true)	3. necessity and contingency (God exists necessarily, but various things are true of God contingently, e.g., God's knowledge of what is contingent)
4. absolute simplicity	4. complexity

Alston distinguishes two lines of argument regarding absoluteness and relativity, which he sees as the key contrast. Alston thinks that only one of these is successful. As indicated in the diagram above, what Hartshorne means by absoluteness is absence of internal relatedness. A relation is internal to a term of a relation just in case that term would not be exactly as it is if it were not in that relationship. Hartshorne's view is that God has internal relations to creatures by way of knowing and acting towards them.

On Alston's interpretation, Hartshorne's first line of argument is to say that if the relation of the absolute to the world really fell outside the absolute, then this relation would necessarily fall within some further and genuinely single entity that embraced both the absolute and the world and the relations between them. Thus, we must hold, according to Hartshorne, that the God-creature relation is internal to God; otherwise we will have to admit that there is something greater or more inclusive than God. Alston does not find this argument convincing because it includes the claim that God "contains" the world due to the internal relations God has with the world. Alston's view is that the entity to which a relation is internal contains the terms only in the sense that those terms enter into a description of the entity, but it does not follow from this that those terms are contained in that entity as marbles are in a box.

Divine inclusiveness, for Hartshorne, is sometimes like the inclusion of thoughts in a mind, but usually it is described as like the inclusion of cells within a living body. It is never like the inclusion of marbles in a box. The inorganic and insentient character of a box is inadequate as a model for divinity, he thinks, and divine inclusiveness is never like the inclusion of theorems in a set of axioms, as it might be for certain idealists. Divine inclusiveness in Hartshorne is *organic* inclusiveness.

Hartshorne's second argument against absoluteness fares much better, according to Alston. He agrees with Hartshorne's stance regarding the cognitive relation God has with the world; in any case of knowledge, the knowledge relation is internal to the subject, external to the object. When a human being knows something, the fact that she knows it is part of what makes her the concrete being that she is. If she recognizes a certain tree she is different from the being she might have been if she had not recognized the tree. But the tree is unaffected by her recognition. Likewise, according to Alston, one cannot maintain that God has perfect knowledge of everything knowable and still hold that God is not qualified to any degree by relations with other beings.

One might respond to Alston and Hartshorne on this point by saying that since creatures depend for their existence on God, their relations to God affect *them*, but not God. Creel seems to make this very point.

But even if beings other than God depend for their existence on God, it still remains true that if God had created a different world from the one that exists at present, then God would be somewhat different from the way God is at present: God's knowledge would have been of *that* world and not this one, according to both Alston and Hartshorne.

Alston's concessions to Hartshorne's concept of God extend to contrasts 2-4. The above argument for the internal relatedness of God (as cognitive subject) to the world presupposes that there are alternative possibilities for God, and if there are alternative possibilities for divine knowledge then this implies that there are unrealized potentialities for God. *Pure* actuality and *total* necessity cannot be defended as divine attributes, according to Alston and Hartshorne. Alston's version of Hartshorne's argument goes as follows:

> (1) (a) "God knows that W exists" entails (b) "W exists."
>
> (2) If (a) were necessary, (b) would be necessary.
>
> (3) But (b) is contingent.
>
> (4) Hence (a) is contingent.

We can totally exclude contingency from God only by denying God any knowledge of anything contingent, a step that not even traditional or classical theists wish to take. Contrast 4 must also be treated in a dipolar way in that the main support for a doctrine of pure divine simplicity is the absence of any unrealized potentialities in God.

In sum, Alston and Hartshorne agree on contrasts 1-4, except for the fact that Hartshorne's concept of divine inclusiveness, in contrast to Alston's, is organic in character.

Regarding a second group of attribues, however, Alston and other theists who are analytic philosophers diverge from Hartshorne rather significantly:

CLASSICAL ATTRIBUTES	NEOCLASSICAL ATTRIBUTES
5. creation *ex nihilo* by a free act of will; God could have refrained from creating anything	5. both God and the world of creatures exist necessarily, although the details are contingent
6. omnipotence (God has the power to do anything God wills to do that is logically consistent)	6. God has all the power one agent could have given metaphysical, in addition to logical, limitations
7. incorporeality	7. corporeality (the world is the body of God)

8. nontemporality (God does not live through a series of moments)	8. temporality (God lives through temporal succession, but everlastingly)
9. immutability (God cannot change because God is not temporally successive)	9. mutability (God is continually attaining richer syntheses of experience)
10. absolute perfection (God is eternally that than which no more perfect can be conceived)	10. relative perfection (at any moment God is more perfect than any other, but God is self-surpassing at a later stage of development)

Concerning contrast 5, Alston takes creation *ex nihilo* to be fundamental to theism because it has deep roots in religious experience. He thinks that to say that God has unrealized potentialities and contingent properties is not to say that God *must* be in relation with some world of entities other than God. Alston admits that Hartshorne legitimately points out some of the internal contradictions contained in the classical theistic version of creation *ex nihilo*, but he claims that there is no connection drawn by Hartshorne between divine creation and metaphysical principles regarding relativity, contingency, and potentiality. Alston's belief seems to be that those who accept creation *ex nihilo* are not saying that there is absolutely nothing at any stage: there is God. Rather, creation *ex nihilo* only means that there is nothing out of which God creates the universe. Here Alston seems to agree with Norman Kretzmann, Eleonore Stump, and most other theists who are analytic philosophers.

Alston's stance here is problematic for two reasons, from Hartshorne's point of view. First, although belief in *some* sort of divine creativity has deep roots in the history of religious experience, it is not clear that these roots have to tap into creation *ex nihilo*. For example, it is not clear that creation *ex nihilo* is the sort of creation described in Genesis, in that when the Bible starts with the statement that the spirit of God hovered above the waters, one gets the impression that both God and the aqueous muck had been around forever. If one believes in creation *ex nihilo*, however, as Alston does, one might nonetheless claim that creation *ex nihilo* does not necessarily mean a temporal beginning to the act of creation. But even on this hypothesis there are problems, and this would seem to be Hartshorne's second point. If Plato and Hartshorne are correct that being *is* dynamic power, then the sort of unlimited power implied by creation *ex nihilo* is impossible. Hartshorne would argue, contra Alston, that there is a connection between belief in creation *ex hyle* (as opposed to creation *ex nihilo*) and the metaphysical principle that being is dynamic power. Creation *ex nihilo*, Hartshorne thinks, is a convenient fiction invented in the first centuries B.C.E. and C.E. in order to exalt divine power, but it is not the only sort of creation that religious believers have defended, nor is it defensible if being is dynamic power.

Concerning contrast 6, Alston claims that belief in creation *ex nihilo* and belief in divine omnipotence are separate beliefs such that to argue against the former is not necessarily to argue against the latter. Hartshorne tries to do too much, he thinks, with the claim that being is power when he uses this claim to argue against divine omnipotence. According to Alston, God can have *unlimited* power, power to do anything that God wills to do, without having *all* power in that, if being is power, the creatures also have some power.

On Hartshorne's interpretation of Alston, however, God can have unlimited power, but not all power, because God delegates some power to others. Although God does not have all power, Hartshorne thinks that on Alston's view God *could* have all power. In effect, what Alston has done, according to Hartshorne, is reduce his stance regarding divine omnipotence to that regarding creation *ex nihilo* in that the claim that God could have all power is due to the prior belief that God brings everything into existence out of absolutely nothing, a belief that Alston thinks has to be the traditional one and in point of fact is intelligible. It is not quite clear to Hartshorne, however, that it is unquestionably the traditional one, nor is it clear to him that we can develop an intelligible concept of "absolutely nothing."

Hartshorne's Platonic or Bergsonian argument against creation *ex nihilo*, in simplified form, looks something like this: one can in fact imagine the nonexistence of this or that, or even of this or that class of things, a fact that gives some the confidence to (erroneously) think that this process can go on infinitely such that one could imagine a state in which there was "absolutely nothing." However, not every verbally possible statement is made conceptually cogent by even the most generous notion of "conceptual," according to Hartshorne. At the specific, ordinary, empirical level negative instances are possible, but at the generic, metaphysical level only positive instances are possible, on this view. The sheer absence of reality cannot conceivably be experienced, he thinks, for if it were experienced an existing experiencer would be presupposed.

Contrast 7 deals with divine embodiment. Alston is willing to grant that God is embodied in two senses: (1) God is aware, with maximal immediacy, of what goes on in the world; and (2) God can directly affect what happens in the world. That is, Alston defends a limited version of divine embodiment, similar to that defended by Richard Swinburne. However, Alston is sceptical regarding a stronger version of divine embodiment wherein the world exists by metaphysical necessity such that God *must* animate it. Alston is willing to accept the idea that God has a body, but *only if* having such a body is on God's terms. It seems that this weaker version of divine embodiment defended by Alston, as opposed to Hartshorne's stronger version wherein there is essential corporeality in God, stands or falls with the defense of creation *ex nihilo*. In fact, despite Alston's desire to examine each contrast individually, as opposed to Hartshorne's stark contrast between classical theistic attributes (all ten of them) and neoclassical attributes (all ten of them), he ends up linking his criticisms of Hartshorne regarding contrasts 5-7, at the very least. All three of these classical theistic attributes stand together only with a defensible version of creation *ex nihilo*.

Contrasts 8-9, concerning nontemporality and immutability, are also linked by Alston. He concedes that if God is temporal, Hartshorne has offered us the best version to date of what divine temporality and divine mutability would be like. Alston dismisses as idle the view that God could remain completely unchanged through a succession of temporal moments, but this admission still leaves us, he thinks, with the following conditional statement: "God undergoes change *if* God is in time." Alston's critique of Hartshorne's view also consists in a refusal to grant that contingency and temporality are coextensive in the way mutability and temporality are. Alston believes, contra Hartshorne, that God can be in some way contingent (that any relation in which God stands to the world might have been otherwise) and still be nontemporal.

Alston knows that the notion of a nontemporal God who is qualified by relation to temporal beings will strike Hartshorne as unintelligible. Alston's attempt to make his position intelligible rests on his own Thomistic-Whiteheadian stance, or better, on his Thomistic or Boethian interpretation of Whitehead (strange as this seems). We should not think of God as involved in process or becoming of any sort. The best temporal analogy, he thinks, for this conception is an unextended instant or an "eternal now." For Alston this does not commit one, however, as Hartshorne would allege, to a static deity frozen in immobility. On the contrary, according to Alston, God is eternally active in ways that do not require temporal succession. God's acts can be complete in an instant. Alston includes God's acts of knowledge, a stance that at least seems to conflict with one of the concessions he made to Hartshorne regarding the first group of attributes.

The Boethian-Thomistic notion of the specious present for God, on the analogue of a human being's perceiving some temporally extended stretch of a process in one temporally indivisible act, is also defended by Alston. For example, one can perceive the flight of a bee "all at once" without first perceiving the first half of the stretch of flight and then perceiving the second. One's perception can be without temporal succession even if the object of one's perception is, in fact, temporally successive. All we have to do, on Alston's view, is expand our specious present to cover all of time and we have a model for God's awareness of the world. This is a much more difficult project for Hartshorne to imagine than it is for Alston. Apparently Alston thinks that it is easy to conceptualize God "seeing" Neanderthal man (or Adam), Moses, Jesus, and Dorothy Day all at once in their immediacy. Here Alston has a view similar to that of William Mann.

But even if it were possible to have nonsuccessive *awareness* of a vast succession, which Hartshorne would deny, it is even more implausible, from Hartshorne's point of view, to claim, as does Alston, that God could have nonsuccessive *responses* to stages of that succession. It might make more sense for Alston to say "indesponses" or "presponses" rather than "responses," as Creel would urge.

It is correct of Alston to notice that there is no loss in God, but this is not incompatible with God's temporality, according to Hartshorne. There can be succession in God without there being loss or perishing due to the fact that God's inheritance of what happens in the world and God's memory are ideal. Hartshorne thinks that the future is incomplete and indeterminate for God as well as from our limited perspective. Alston, by way of contrast, wants to defend a God who is not strictly necessary in actuality, but is contingent, *despite the fact that* God does not undergo temporal change, nor is God fluent. Hartshorne's defenders, by way of contrast, think that one of the greatest virtues of process thinking is its effort to eliminate what they see as such self-contradiction in philosophical theology.

Alston's treatment of contrast 10, concerning absolute versus relative perfection, follows from what he has said regarding contrasts 8-9. Relative perfection in God, as opposed to absolute perfection, has a point only for a temporal being; hence God is absolutely perfect, according to Alston. A being that does not successively assume different states could not possibly surpass itself. Here, once again, Alston engages in linkage, thereby, at the very least, confirming Hartshorne's belief that we need both to consider the divine attributes together and to determine whether the classical theist's linkage or the neoclassical theist's linkage is more defensible. For the most past, Alston opts for classical theism. Or

more precisely, he thinks that the strongest concept of God is acquired when we take a modified version of the neoclassical attributes in contrasts 1-4 and combine them with the classical attributes in contrasts 5-10.

This rapproachement in Alston between classical theism and neoclassical theism is a step beyond James Ross's belief that these are two competing descriptions of God at an impasse. Hartshorne seems to agree with Ross. Neoclassical, dipolar theism *already* includes the best insights of classical theism, he thinks.

From Hartshorne's point of view the linkage of attributes *within* the first group and *within* the second group needs to be corrected by a greater concern for reticulating the attributes in these two groups. He thinks that an explanation is needed regarding how Alston can be committed to both monopolar and dipolar theism. For example, Alston ends up defending the view that God is changed by the objects God knows (pace the neoclassical, dipolar attributes), but these are not changes that occur in time (pace the classical, monopolar attributes). It is one thing, Hartshorne thinks, to say that God exists in a nontemporal specious present, and it is another to say that God is changed by temporal beings in a nontemporal specious present. The former view is at least problematic, he thinks, and the latter seems to be part of the traditional classical theistic view wherein, from a Hartshornian perspective, inconsistency goes in the guise of mystery.

Bibliography

Books by Hartshorne

- (1923) "An Outline and Defense of the Argument for the Unity of Being in the Absolute or Divine Good." Ph.D. Dissertation, Harvard University.
- (1934) *The Philosophy and Psychology of Sensation*. Chicago: University of Chicago Press.
- (1937) *Beyond Humanism*. Chicago: Willet, Clark, and Co.
- (1941) *Man's Vision of God*. N.Y.: Harper and Brothers.
- (1948) *The Divine Relativity*. New Haven: Yale University Press.
- (1953) *Reality as Social Process*. Boston: Beacon Press.
- (1953) *Philosophers Speak of God*. Chicago: University of Chicago Press.
- (1962) *The Logic of Perfection*. LaSalle, Il.: Open Court.
- (1967) *A Natural Theology for Our Time*. LaSalle, Il: Open Court.
- (1967) *Anselm's Discovery*. LaSalle, Il.: Open Court.
- 1970) *Creative Synthesis and Philosophic Method*. LaSalle, Il.: Open Court.
- (1972) *Whitehead's Philosophy*. Lincoln: University of Nebraska Press.
- (1973) *Born to Sing*. Bloomington: Indiana University Press.
- (1976) *Aquinas to Whitehead*. Milwaukee: Marquette University Press.
- (1983) *Insights and Oversights of Great Thinkers*. Albany: State University of New York Press.
- (1984) *Existence and Actuality: Conversations with Charles Hartshorne*. Chicago: University of Chicago Press.
- (1984) *Creativity in American Philosophy*. Albany: State University of New York Press.

- (1984) *Omnipotence and Other Theological Mistakes*. Albany: State University of New York Press.
- (1987) *Wisdom as Moderation*. Albany: State University of New York Press.
- (1990) *The Darkness and the Light*. Albany: State University of New York Press.
- (1991) *The Philosophy of Charles Hartshorne*. LaSalle, Il.: Open Court.
- (1997) *The Zero Fallacy and Other Essays in Neoclassical Metaphysics*. LaSalle, Il.: Open Court.

Secondary Sources

- Alston, William. (1964) "The Elucidation of Religious Statements." in *Process and Divinity: The Hartshorne Festschrift*. LaSalle, Il.: Open Court.
- ----- (1989) "Hartshorne and Aquinas: A Via Media." in *Divine Nature and Human Language*. Ithaca: Cornell University Press.
- Cobb, John and Franklin Gamwell, eds. (1984) *Existence and Actuality: Conversations with Charles Hartshorne*. Chicago: University of Chicago Press.
- Dombrowski, Daniel. (1988) *Hartshorne and the Metaphysics of Animal Rights*. Albany: State University of New York Press.
- ----- (1996) *Analytic Theism, Hartshorne, and the Concept of God*. Albany: State University of New York Press.
- Ford, Lewis, ed. (1973) *Two Process Philosophers*. Tallahassee, Fl.: American Academy of Religion.
- Gilroy, John. "Hartshorne and the Ultimate Issue in Metaphysics." *Process Studies* 18: 38-56.
- Griffin, David Ray. (2001) *Reenchantment without Supernaturalism*. Ithaca: Cornell University Press.
- Gunton, Colin. (1978) *Becoming and Being: The Doctrine of God in Charles Hartshorne and Karl Barth*. Oxford: Oxford University Press.
- Hahn, Lewis, ed. (1991) *The Philosophy of Charles Hartshorne*. LaSalle, Il.: Open Court.
- Morris, Randall. (1991) *Process Philosophy and Political Ideology*. Albany: State University of New York Press.
- Peters, Eugene. (1970) *Hartshorne and Neoclassical Metaphysics*. Lincoln: University of Nebraska Press.
- Ross, James. (1977) "An Impasse on Competing Descriptions of God." *International Journal for Philosophy of Religion* 8: 233-249.
- Shields, George, ed. (forthcoming) *Process and Analysis: Essays on Whitehead, Hartshorne, and the Analytic Tradition*.
- Sia, Santiago, ed. (1990) *Charles Hartshorne's Concept of God*. Boston: Kluwer.
- Tracy, David. (1985) "Analogy, Metaphor, and God-Language: Charles Hartshorne." *Modern Schoolman* 62: 249-265.
- Viney, Don. (1985) *Charles Hartshorne and the Existence of God.* Albany: State University of New York Press.
- Vitali, Theodore. (1977) "The Peircian Influence on Hartshorne's Subjectivism." *Process Studies* 7: 238-249.
- Whitney, Barry. (1985) *Evil and the Process God*. Toronto: Edwin Mellon Press.

Other Internet Resources

- The Center for Process Studies, at the Claremont School of Theology

Related Entries

Anselm, Saint [Anselm of Bec, Anselm of Canterbury] | cosmology: and theology | immutability | ontological arguments | panpsychism | Peirce, Charles Sanders | religion: epistemology of | time | Whitehead, Alfred North

Dan Dombrowski
ddombrow@seattleu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 23, 2001
Content last modified: July 23, 2001

Saint Anselm

Saint Anselm of Canterbury (1033-1109) was the outstanding Christian philosopher and theologian of the eleventh century. He is best known for the celebrated "ontological argument" for the existence of God in chapter two of the *Proslogion*, but his contributions to philosophical theology (and indeed to philosophy more generally) go well beyond the ontological argument. In what follows I examine Anselm's theistic proofs, his conception of the divine nature, and his account of human freedom, sin, and redemption.

Life and Works

Anselm was born in 1033 near Aosta, in those days a Burgundian town on the frontier with Lombardy. Little is known of his early life. He left home at twenty-three, and after three years of apparently aimless travelling through Burgundy and France, he came to Normandy in 1059. Once he was in Normandy, Anselm's interest was captured by the Benedictine abbey at Bec, whose famous school was under the direction of Lanfranc, the abbey's prior. Lanfranc was a scholar and teacher of wide reputation, and under his leadership the school at Bec had become an important center of learning, especially in dialectic. In 1060 Anselm entered the abbey as a novice. His intellectual and spiritual gifts brought him rapid advancement, and when Lanfranc was appointed abbot of Caen in 1063, Anselm was elected to succeed him as prior. He was elected abbot in 1078 upon the death of Herluin, the founder and first abbot of Bec. Under Anselm's leadership the reputation of Bec as an intellectual center grew, and Anselm managed to write a good deal of philosophy and theology in addition to his teaching, administrative duties, and extensive correspondence as an adviser and counsellor to rulers and nobles all over Europe and beyond. His works while at Bec include the *Monologion* (1076), the *Proslogion* (1077-78), and his four

philosophical dialogues: *De grammatico*, *De veritate*, and *De libertate arbitrii* (1080-85), and *De casu diaboli* (1085-1090).

In 1093 Anselm was enthroned as Archbishop of Canterbury. The previous Archbishop, Anselm's old master Lanfranc, had died four years earlier, but the King, William Rufus, had left the see vacant in order to plunder the archiepiscopal revenues. Anselm was understandably reluctant to undertake the primacy of the Church of England under a ruler as ruthless and venal as William, and his tenure as Archbishop proved to be as turbulent and vexatious as he must have feared. William was intent on maintaining royal authority over ecclesiastical affairs and would not be dictated to by Archbishop or Pope or anyone else. So, for example, when Anselm went to Rome in 1097 without the King's permission, William would not allow him to return. When William was killed in 1100, his successor, Henry I, invited Anselm to return to his see. But Henry was as intent as William had been on maintaining royal jurisdiction over the Church, and Anselm found himself in exile again from 1103 to 1107. Despite these distractions and troubles, Anselm continued to write. His works as Archbishop of Canterbury include the *Epistola de Incarnatione Verbi* (1092-94), *Cur Deus Homo* (1094-98), *De conceptu virginali* (1099-1100), *De processione Spiritus Sancti* (1102), the *Epistola de sacrificio azymi et fermentati* (1106-7), *De sacramentis ecclesiae* (1106-7), and *De concordia* (1107-8). Anselm died on 21 April 1109. He was canonized in 1494 and named a Doctor of the Church in 1720.

The Theistic Proofs

"Faith Seeking Understanding": The character and purpose of Anselm's theistic proofs

Anselm's motto is "faith seeking understanding" (*fides quaerens intellectum*). This motto lends itself to at least two misunderstandings. First, many philosophers have taken it to mean that Anselm hopes to *replace* faith with understanding. If one takes 'faith' to mean roughly 'belief on the basis of testimony' and 'understanding' to mean 'belief on the basis of philosophical insight', one is likely to regard faith as an epistemically substandard position; any self-respecting philosopher would surely want to leave faith behind as quickly as possible. The theistic proofs are then interpreted as the means by which we come to have philosophical insight into things we previously believed solely on testimony. But as I have argued elsewhere (Williams 1996, xiii-xiv), Anselm is not hoping to replace faith with understanding. Faith for Anselm is more a volitional state than an epistemic state: it is love for God and a drive to act as God wills. In fact, Anselm describes the sort of faith that "merely believes what it ought to believe" as "dead" (*M* 78). (For the abbreviations used in references, see the Bibliography below.) So "faith seeking understanding" means something like "an active love of God seeking a deeper knowledge of God."

Other philosophers have noted that "faith seeking understanding" begins with "faith," not with doubt or suspension of belief. Hence, they argue, the theistic arguments proposed by faith seeking understanding are not really meant to convince unbelievers; they are intended solely for the edification of those who already believe. This too is a misreading of Anselm's motto. For although the theistic proofs are borne of

an active love of God seeking a deeper knowledge of the beloved, the proofs themselves are intended to be convincing even to unbelievers. Thus Anselm opens the *Monologion* with these words:

> If anyone does not know, either because he has not heard or because he does not believe, that there is one nature, supreme among all existing things, who alone is self-sufficient in his eternal happiness, who through his omnipotent goodness grants and brings it about that all other things exist or have any sort of well-being, and a great many other things that we must believe about God or his creation, I think he could at least convince himself of most of these things by reason alone, if he is even moderately intelligent. (*M* 1)

And in the *Proslogion* Anselm sets out to convince "the fool," that is, the person who "has said in his heart, ‘There is no God’ " (Psalm 14:1; 53:1).

The arguments of the *Monologion*

Having clarified what Anselm takes himself to be doing in his theistic proofs, we can now examine the proofs themselves. In the first chapter of the *Monologion* Anselm argues that there must be some one thing that is supremely good, through which all good things have their goodness. For whenever we say that different things are *F* in different degrees, we must understand them as being *F* through *F*-ness; *F*-ness itself is the same in each of them. Thus, for example, all more or less just things "must be more or less just through justice, which is not different in diverse things" (*M* 1). Now we speak of things as being *good* in different degrees. So by the principle just stated, these things must be good through some one thing. Clearly that thing is itself a great good, since it is the source of the goodness of all other things. Moreover, that thing is good *through itself*; after all, if all good things are good through that thing, it follows trivially that that thing, being good, is good through itself. Things that are good through another (i.e., things whose goodness derives from something other than themselves) cannot be equal to or greater than the good thing that is good through itself, and so that which is good through itself is *supremely* good. Anselm concludes, "Now that which is supremely good is also supremely great. There is, therefore, some one thing that is supremely good and supremely great--in other words, supreme among all existing things" (*M* 1). In chapter 2 he applies the principle of chapter 1 in order to derive (again) the conclusion that there is something supremely great.

In chapter 3 Anselm argues that all existing things exist through some one thing. Every existing thing, he begins, exists either through something or through nothing. But of course nothing exists through nothing, so every existing thing exists through something. There is, then, either some one thing through which all existing things exist, or there is more than one such thing. If there is more than one, either (i) they all exist through some one thing, or (ii) each of them exists through itself, or (iii) they exist through each other. (iii) makes no sense. If (ii) is true, then "there is surely some one power or nature of self-existing that they have in order to exist through themselves" (*M* 3); in that case, "all things exist more truly through that one thing than through the several things that cannot exist without that one thing" (*M* 3). So (ii) collapses into (i), and there is some one thing through which all things exist. That one thing, of course, exists through itself, and so it is greater than all the other things. It is therefore "best and greatest

and supreme among all existing things" (*M* 3).

In chapter 4 Anselm begins with the premise that things "are not all of equal dignity; rather, some of them are on different and unequal levels" (*M* 4). For example, a horse is better than wood, and a human being is more excellent than a horse. Now it is absurd to think that there is no limit to how high these levels can go, "so that there is no level so high that an even higher level cannot be found" (*M* 4). The only question is how many beings occupy that highest level of all. Is there just one, or are there more than one? Suppose there are more than one. By hypothesis, they must all be equals. If they are equals, they are equals through the same thing. That thing is either identical with them or distinct from them. If it is identical with them, then they are not in fact many, but one, since they are all identical with some one thing. On the other hand, if that thing is distinct from them, then they do not occupy the highest level after all. Instead, that thing is greater than they are. Either way, there can be only one being occupying the highest level of all.

Anselm concludes the first four chapters by summarizing his results:

> Therefore, there is a certain nature or substance or essence who through himself is good and great and through himself is what he is; through whom exists whatever truly is good or great or anything at all; and who is the supreme good, the supreme great thing, the supreme being or subsistent, that is, supreme among all existing things. (*M* 4)

He then goes on (in chapters 5-65) to derive the attributes that must belong to the being who fits this description. But before we look at Anselm's understanding of the divine attributes, we should turn to the famous proof in the *Proslogion*.

The arguments of the *Proslogion*

Looking back on the sixty-five chapters of complicated argument in the *Monologion*, Anselm found himself wishing for a simpler way to establish all the conclusions he wanted to prove. As he tells us in the preface to the *Proslogion*, he wanted to find

> a single argument that needed nothing but itself alone for proof, that would by itself be enough to show that God really exists; that he is the supreme good, who depends on nothing else, but on whom all things depend for their being and for their well-being; and whatever we believe about the divine nature. (*P*, preface)

That "single argument" is the one that appears in chapter 2 of the *Proslogion*. (We owe the curiously unhelpful name "ontological argument" to Kant. The medievals simply called it "that argument of Anselm's" [*argumentum Anselmi*].)

The argument goes like this. God is "that than which nothing greater can be thought"; in other words, he is a being so great, so full of metaphysical oomph, that one cannot so much as conceive of a being who

would be greater than God. The Psalmist, however, tells us that "The fool has said in his heart, 'There is no God' " (Psalm 14:1; 53:1). Is it possible to convince the fool that he is wrong? It is. All we need is the definition of God as "that than which nothing greater can be thought." The fool does at least understand that definition. But whatever is understood exists in the understanding, just as the plan of a painting he has yet to execute already exists in the understanding of the painter. So that than which nothing greater can be thought exists in the understanding. But if it exists in the understanding, it must also exist in reality. For it is greater to exist in reality than to exist merely in the understanding. Therefore, if that than which nothing greater can be thought existed only in the understanding, it would be possible to think of something greater than it (namely, that same being existing in reality as well). It follows, then, that if that than which nothing greater can be thought existed only in the understanding, it would not be that than which nothing greater can be thought; and that, obviously, is a contradiction. So that than which nothing greater can be thought must exist in reality, not merely in the understanding.

Versions of this argument have been defended and criticized by a succession of philosophers from Anselm's time through the present day (see ontological arguments). Our concern here is with Anselm's own version, the criticism he encountered, and his response to that criticism. A monk named Gaunilo wrote a "Reply on Behalf of the Fool," contending that Anselm's argument gave the Psalmist's fool no good reason at all to believe that that than which nothing greater can be thought exists in reality. Gaunilo's most famous objection is an argument intended to be exactly parallel to Anselm's that generates an obviously absurd conclusion. Gaunilo proposes that instead of "that than which nothing greater can be thought" we consider "that island than which no greater can be thought." We understand what that expression means, so (following Anselm's reasoning) the greatest conceivable island exists in our understanding. But (again following Anselm's reasoning) that island must exist in reality as well; for if it did not, we could imagine a greater island--namely, one that existed in reality--and the greatest conceivable island would not be the greatest conceivable island after all. Surely, though, it is absurd to suppose that the greatest conceivable island actually exists in reality. Gaunilo concludes that Anselm's reasoning is fallacious.

In order to defend himself against Gaunilo's criticism, Anselm would have to show why Gaunilo's argument about the island is not in fact analogous to his own argument about that than which nothing greater can be thought. Surprisingly, he never does this. His long-winded and indeed somewhat intemperate "Reply to Gaunilo" *asserts* more than once that the island example fails, but he never explains *why* it fails. The usual reply given on Anselm's behalf (and indeed often attributed to Anselm himself) is that the notion of a greatest conceivable island is incoherent; however great an island might be, one could always conceive of a greater. This is a lame response, since it is open to Gaunilo to say exactly the same thing about the greatest conceivable *being*; it is therefore no wonder that Anselm did not say anything of the sort. (For a reading of the argument that endorses a response of this sort, see Klima 2000.) Indeed, Nicholas Wolterstorff argues convincingly that Anselm

> realized the 'tellingness' of [Gaunilo's] points. . . . The sign of his realization, however, is not concession; Anselm does not concede. The sign is rather bluster. . . . Anselm's glittering genius has made many reluctant to concede that Gaunilo made any telling points against him; his saintly reputation makes us all reluctant to concede that he concealed

when he should have conceded. (Wolterstorff 1993, 87)

The Divine Nature

Proving the divine attributes

Recall that Anselm's intention in the *Proslogion* was to offer a single argument that would establish not only the existence of God but also the various attributes that Christians believe God possesses. If the argument of chapter 2 proved only the existence of God, leaving the divine attributes to be established piecemeal as in the *Monologion*, Anselm would consider the *Proslogion* a failure. But in fact the concept of that than which nothing greater can be thought turns out to be marvelously fertile. God must, for example, be omnipotent. For if he were not, we could conceive of a being greater than he. But God is that than which no greater can be thought, so he must be omnipotent. Similarly, God must be just, self-existent, invulnerable to suffering, merciful, timelessly eternal, non-physical, non-composite, and so forth. For if he lacked any of these qualities, he would be less than the greatest conceivable being, which is impossible.

The ontological argument thus works as a sort of divine-attribute-generating machine. Admittedly, though, the appearance of theoretical simplicity is somewhat misleading. The "single argument" produces conclusions about the divine attributes only when conjoined with certain beliefs about what is greater or better. That is, the ontological argument tells us that God has whatever characteristics it is better or greater to have than to lack, but it does not tell us which characteristics those are. We must have some independent way of identifying them before we can plug them into the ontological argument and generate a full-blown conception of the divine nature. Anselm identifies these characteristics in part by appeal to intuitions about value, in part by independent argument. To illustrate Anselm's method, I shall examine his discussions of God's impassibility, timelessness, and simplicity.

According to the doctrine of divine impassibility, God is invulnerable to suffering. Nothing can act upon him; he is in no way passive. He therefore does not feel emotions, since emotions are states that one undergoes rather than actions one performs. Anselm does not find it necessary to *argue* that impassibility is a perfection; he thinks it is perfectly obvious that "it is better to be . . . impassible than not" (*P* 6), just as it is perfectly obvious that it is better to be just than not-just. His intuitions about value are shaped by the Platonic-Augustinian tradition of which he was a part. Augustine took from the Platonists the idea that the really real things, the greatest and best of beings, are stable, uniform, and unchanging. He says in *On Free Choice of the Will* 2.10, "And you surely could not deny that the uncorrupted is better than the corrupt, the eternal than the temporal, and the invulnerable than the vulnerable"; his interlocutor replies simply, "Could anyone?" Through Augustine (and others) these ideas, and the conception of God to which they naturally lead, became the common view of Christian theologians for well over a millennium. For Anselm, then, it is obvious that a being who is in no way passive, who cannot experience anything of which he is not himself the origin, is better and greater than any being who can be acted upon by something outside himself. So God, being that than which nothing greater can be thought, is wholly

active; he is impassible.

Notice that Augustine also found it obvious that the eternal is better than the temporal. According to Plato's *Timaeus*, time is a "moving image of eternity" (37d). It is a shifting and shadowy reflection of the really real. As later Platonists, including Augustine, develop this idea, temporal beings have their existence piecemeal; they exist only in this tiny sliver of a now, which is constantly flowing away from them and passing into nothingness. An eternal being, by contrast, is (to use my earlier description) stable, uniform, and unchanging. What it has, it always has; what it is, it always is; what it does, it always does. So it seems intuitively obvious to Anselm that if God is to be that than which nothing greater can be thought, he must be eternal. That is, he must be not merely everlasting, but outside time altogether.

In addition to this strong intuitive consideration, Anselm at least hints at a further argument for the claim that it is better to be eternal than temporal. He opens chapter 13 of the *Proslogion* by observing, "Everything that is at all enclosed in a place or time is less than that which is subject to no law of place or time" (*P* 13). His idea seems to be that if God were in time (or in a place), he would be bound by certain constraints inherent in the nature of time (or place). His discussion in *Monologion* 22 makes the problem clear:

> This, then, is the condition of place and time: whatever is enclosed within their boundaries does not escape being characterized by parts, whether the sort of parts its place receives with respect to size, or the sort its time suffers with respect to duration; nor can it in any way be contained as a whole all at once by different places or times. By contrast, if something is in no way constrained by confinement in a place or time, no law of places or times forces it into a multiplicity of parts or prevents it from being present as a whole all at once in several places or times. (*M* 22)

So at least part of the reason for holding that God is timeless is that the nature of time would impose constraints upon God, and of course it is better to be subject to no external constraints.

The other part of the reason, though, is that if God were in place or time he would have *parts*. But what is so bad about having parts? This question brings us naturally to the doctrine of divine simplicity, which is simply the doctrine that God has no parts of any kind. Even for an Augustinian like Anselm, the claim that it is better to lack parts than to have them is less than intuitively compelling, so Anselm offers further arguments for that claim. In the *Proslogion* he argues that "whatever is composed of parts is not completely one. It is in some sense a plurality and not identical with itself, and it can be broken up either in fact or at least in the understanding" (*P* 18). The argument in the *Monologion* goes somewhat differently. "Every composite," Anselm argues, "needs the things of which it is composed if it is to subsist, and it owes its existence to them, since whatever it is, it is through them, whereas those things are not through it what they are" (*M* 17). The argument in the *Proslogion*, then, seeks to relate simplicity to the intuitive considerations that identify what is greatest and best with what is stable, uniform, and unchanging; the argument in the *Monologion*, by contrast, seeks to show that simplicity is necessary if God is to be--as the theistic proofs have already established--the ultimate source of his own goodness and

existence.

The consistency of the divine attributes

Anselm's success in generating a whole host of divine attributes through the ontological argument does present him with a problem. He must show that the attributes are consistent with each other--in other words, that is possible for one and the same being to have all of them. For example, there seems at first glance to be a conflict between justice and omnipotence. If God is perfectly just, he cannot lie. But if God is omnipotent, how can there be something he cannot do? Anselm's solution is to explain that omnipotence does not mean the ability to do everything; instead, it means the possession of unlimited power. Now the so-called "ability" or "power" to lie is not really a power at all; it is a kind of weakness. Being omnipotent, God has no weakness. So it turns out that omnipotence actually *entails* the inability to lie.

Another apparent contradiction is between God's mercy and his justice. If God is just, he will surely punish the wicked as they deserve. But because he is merciful, he spares the wicked. Anselm tries to resolve this apparent contradiction by appeal to God's goodness. It is better, he says, for God "to be good both to the good and to the wicked than to be good only to the good, and it is better to be good to the wicked both in punishing and in sparing them than to be good only in punishing them" (*P* 9). So God's supreme goodness requires that he be both just and merciful. But Anselm is not content to resolve the apparent tension between justice and mercy by appealing to some other attribute, goodness, that entails both justice and mercy; he goes on to argue that justice itself requires mercy. Justice to sinners obviously requires that God punish them; but God's justice *to himself* requires that he exercise his supreme goodness in sparing the wicked. "Thus," Anselm says to God, "in saving us whom you might justly destroy . . . you are just, not because you give us our due, but because you do what is fitting for you who are supremely good" (*P* 10). In spite of these arguments, Anselm acknowledges that there is a residue of mystery here:

> Thus your mercy is born of your justice, since it is just for you to be so good that you are good even in sparing the wicked. And perhaps this is why the one who is supremely just can will good things for the wicked. But even if one can somehow grasp why you can will to save the wicked, certainly no reasoning can comprehend why, from those who are alike in wickedness, you save some rather than others through your supreme goodness and condemn some rather than others through your supreme justice. (*P* 11)

In other words, the philosopher can trace the conceptual relations among goodness, justice, and mercy, and show that God not only can but must have all three; but no human reasoning can hope to show why God displays his justice and mercy in precisely the ways in which he does.

Freedom, Sin, and Redemption

Truth in statements and in the will

In *On Freedom of Choice* (*De libertate arbitrii*) Anselm defines freedom of choice as "the power to preserve rectitude of will for its own sake" (*DLA* 3). He explores the notion of rectitude of will most thoroughly in *On Truth* (*De veritate*), so in order to understand the definition of freedom of choice, we must look first at Anselm's discussion of truth. Truth is a much broader notion for Anselm than for us; he speaks of truth not only in statements and opinions but also in the will, actions, the senses, and even the essences of things. In every case, he argues, truth consists in correctness or "rectitude." Rectitude, in turn, is understood teleologically; a thing is correct whenever it is or does whatever it ought, or was designed, to be or do. For example, statements are made for the purpose of "signifying that what-is is" (*DV* 2). A statement therefore is correct (has rectitude) when, and only when, it signifies that what-is is. So Anselm holds a correspondence theory of truth, but it is a somewhat unusual correspondence theory. Statements are true when they correspond to reality, but only because corresponding to reality is what statements are *for*. That is, statements (like anything else) are true when they do what they were designed to do; and what they were designed to do, as it happens, is to correspond to reality.

Truth in the will also turns out to be rectitude, again understood teleologically. Rectitude of will means willing what one ought to will or (in other words) willing that for the sake of which one was given a will. So, just as the truth or rectitude of a statement is the statement's doing what statements were made to do, the truth or rectitude of a will is the will's doing what wills were made to do. In *DV* 12 Anselm connects rectitude of will to both justice and moral evaluation. In a broad sense of ‘just’, whatever is as it ought to be is just. Thus, an animal is just when it blindly follows its appetites, because that is what animals were meant to do. But in the narrower sense of ‘just’, in which justice is what deserves moral approval and injustice is what deserves reproach, justice is best defined as "rectitude of will preserved for its own sake" (*DV* 12). Such rectitude requires that agents perceive the rectitude of their actions and will them for the sake of that rectitude. Anselm takes the second requirement to exclude both coercion and "being bribed by an extraneous reward" (*DV* 12). For an agent who is coerced into doing what is right is not willing rectitude for its own sake; and similarly, an agent who must be bribed to do what is right is willing rectitude for the sake of the bribe, not for the sake of rectitude.

Since, as we have already seen, Anselm will define freedom as "the power to preserve rectitude of will for its own sake," the arguments of *On Truth* imply that freedom is also the capacity for justice and the capacity for moral praiseworthiness. Now it is both necessary and sufficient for justice, and thus for praiseworthiness, that an agent wills what is right, knowing it to be right, because it is right. That an agent wills what is right because it is right entails that he is neither compelled nor bribed to perform the act. Freedom, then, must be neither more nor less than the power to perform acts of that sort.

Freedom and sin

Thus Anselm takes it to be obvious that freedom is a power *for* something: its purpose is to preserve rectitude of will for its own sake. God and the good angels cannot sin, but they are still free, because they can (and do) preserve rectitude of will for its own sake. In fact, they are freer than those who can sin:

"someone who has what is fitting and expedient in such a way that he cannot lose it is freer than someone who has it in such a way that he can lose it and be seduced into what is unfitting and inexpedient" (*DLA* 1). It obviously follows, as Anselm points out, that freedom of choice neither is nor entails the power to sin; God and the good angels have freedom of choice, but they are incapable of sinning.

But if free choice is the power to hold on to what is fitting and expedient, and it is not the power to sin, does it make any sense to say that the first human beings and the rebel angels sinned through free choice? Anselm's reply to this question is both subtle and plausible. In order to be able to preserve rectitude of will for its own sake, an agent must be able to perform an action that has its ultimate origin in the agent him- or herself rather than in some external source. (For convenience I will refer to that power as "the power for self-initiated action.") Any being that has freedom of choice, therefore, will thereby have the power for self-initiated action. The first human beings and the rebel angels sinned through an exercise of their power for self-initiated action, and so it is appropriate to say that they sinned through free choice. Nonetheless, free choice does not entail the power to sin. For free choice can be perfected by something else, as yet unspecified, that renders it incapable of sinning.

In *On the Fall of the Devil* (*De casu diaboli*) Anselm extends his account of freedom and sin by discussing the first sin of the angels. In order for the angels to have the power to preserve rectitude of will for its own sake, they had to have both a will for justice and a will for happiness. If God had given them only a will for happiness, they would have been necessitated to will whatever they thought would make them happy. Their willing of happiness would have had its ultimate origin in God and not in the angels themselves. So they would not have had the power for self-initiated action, which means that they would not have had free choice. The same thing would have been true, *mutatis mutandis*, if God had given them only the will for justice.

Since God gave them both wills, however, they had the power for self-initiated action. Whether they chose to subject their wills for happiness to the demands of justice or to ignore the demands of justice in the interest of happiness, that choice had its ultimate origin in the angels; it was not received from God. The rebel angels chose to abandon justice in an attempt to gain happiness for themselves, whereas the good angels chose to persevere in justice even if it meant less happiness. God punished the rebel angels by taking away their happiness; he rewarded the good angels by granting them all the happiness they could possibly want. For this reason, the good angels are no longer able to sin. Since there is no further happiness left for them to will, their will for happiness can no longer entice them to overstep the bounds of justice. Thus Anselm finally explains what it is that perfects free choice so that it becomes unable to sin.

Grace and redemption

Like the fallen angels, the first human beings willed happiness in preference to justice. By doing so they abandoned the will for justice and became unable to will justice for its own sake. Apart from divine grace, then, fallen human beings cannot help but sin. Anselm claims that we are still free, because we continue to be such that if we had rectitude of will, we could preserve it for its own sake; but we cannot

exercise our freedom, since we no longer have the rectitude of will to preserve. (Whether fallen human beings also retain the power for self-initiated action apart from divine grace is a tricky question, and one I do not propose to answer here.)

So the restoration of human beings to the justice they were intended to enjoy requires divine grace. But even more is needed than God's restoration of the will for justice. In *Cur Deus Homo* (*Why the God-Man?* or *Why God Became Man*) Anselm famously attempts to show on purely rational grounds that the debt incurred by human sin could be suitably discharged, and the affront to God's infinite dignity could be suitably rectified, only if one who was both fully divine and fully human took it upon himself to offer his own life on our behalf.

Bibliography

References in this article to Anselm's works use the following abbreviations:

DLA = *De libertate arbitrii*
DV = *De veritate*
M = *Monologion*
P = *Proslogion*

All translations are my own.

Critical Edition

- Schmitt, Franciscus Salesius (1968). *S. Anselmi Cantuariensis Archiepiscopi Opera Omnia*. Stuttgart-Bad Cannstatt: Friedrich Fromann Verlag, 1968.

Translations

- Davies, Brian, and G. R. Evans, ed. (1998). *Anselm of Canterbury: The Major Works*. Oxford: Oxford University Press, 1998.
- Williams, Thomas (1996). *Monologion and Proslogion*. Indianapolis: Hackett Publishing Company, 1996.
- Williams, Thomas (2002). *Three Philosophical Dialogues: On Truth, On Freedom of Choice, On the Fall of the Devil*. Indianapolis: Hackett Publishing Company, 2002.

Secondary Works

- Evans, G. R. (1978). *Anselm and Talking about God*. Oxford: Clarendon Press, 1978.

- Evans, G. R. (1984). *A Concordance to the Works of Saint Anselm*. Millwood, NY: Kraus International Publications, 1984.
- Evans, G. R. (1989). *Anselm*. London: G. Chapman, 1989.
- Henry, Desmond Paul (1967). *The Logic of Saint Anselm*. Oxford: Clarendon Press, 1967.
- Holopainen, Toivo (1996). *Dialectic and Theology in the Eleventh Century*. Leiden: E. J. Brill, 1996.
- Hopkins, Jasper (1972). *A Companion to the Study of St. Anselm*. Minneapolis: University of Minnesota Press, 1972.
- Klima, Gyula (2000). "Saint Anselm's Proof: A Problem of Reference, Intentional Identity and Mutual Understanding", in G. Hintikka (ed.), *Medieval Philosophy and Modern Times* (Proceedings of "Medieval and Modern Philosophy of Religion", Boston University, August 25-27, 1992), Dordrecht: Kluwer, pp. 69-88. [Preprint available online]
- Leftow, Brian (1997). "Anselm on the Cost of Salvation," *Medieval Philosophy and Theology* 6 (1997): 73-92.
- Oppenheimer, P., and Zalta, E. (1991). "On the Logic of the Ontological Argument", *Philosophical Perspectives* 5 (1991): 509-529; reprinted in *The Philosopher's Annual: 1991*, XIV (1993): 255-275
- Plantinga, Alvin, ed. (1965). *The Ontological Argument*. Garden City, NY: Anchor Books, 1965.
- Southern, R. W. (1990). *Saint Anselm: A Portrait in Landscape*. Cambridge: Cambridge University Press, 1990.
- Williams, Thomas, and Sandra Visser (2001). "Anselm's Account of Freedom," *Canadian Journal of Philosophy* 31 (2001): 221-244. [Preprint available online]
- Williams, Thomas (1997). Review of Holopainen (1996) in *History and Philosophy of Logic* 18 (1997): 55-59.
- Wolterstorff, Nicholas (1993). "In Defense of Gaunilo's Defense of the Fool," in C. Stephen Evans and Merold Westphal, ed., *Christian Perspectives on Religious Knowledge*. Grand Rapids, MI: William B. Eerdmans Publishing Company, 1993.

Other Internet Resources

- *Catholic Encyclopedia* article on Anselm

Related Entries

Augustine, Saint | Duns Scotus, John | free will | medieval philosophy | ontological arguments

Thomas Williams
thomas-williams@uiowa.edu

First published: May 18, 2000
Content last modified: March 5, 2002

Ontological Arguments

Ontological arguments are arguments, for the conclusion that God exists, from premises which are supposed to derive from some source other than observation of the world -- e.g., from reason alone. In other words, ontological arguments are arguments from nothing but analytic, *a priori* and necessary premises to the conclusion that God exists.

The first, and best-known, ontological argument was proposed by St. Anselm of Canterbury in the 11th. century A.D. In his *Proslogion*, St. Anselm claims to derive the existence of God from the concept of a *being than which no greater can be conceived*. St. Anselm reasoned that, if such a being fails to exist, then a greater being -- namely, a *being than which no greater can be conceived, and which exists* -- can be conceived. But this would be absurd: nothing can be greater than a being than which no greater can be conceived. So a being than which no greater can be conceived -- i.e., God -- exists.

In the seventeenth century, Rene Descartes defended a family of similar arguments. For instance, in the *Fifth Meditation*, Descartes claims to provide a proof demonstrating the existence of God from the idea of a supremely perfect being. Descartes argues that there is no less contradiction in conceiving a supremely perfect being who lacks existence than there is in conceiving a triangle whose interior angles do not sum to 180 degrees. Hence, he supposes, since we do conceive a supremely perfect being -- we do have the idea of a supremely perfect being -- we must conclude that a supremely perfect being exists.

In the early eighteenth century, Gottfried Leibniz attempted to fill what he took to be a shortcoming in Descartes' view. According to Leibniz, Descartes' arguments fail unless one first shows that the idea of a supremely perfect being is coherent, or that it is possible for there to be a supremely perfect being. Leibniz argued that, since perfections are unanalysable, it is impossible to demonstrate that perfections are incompatible -- and he concluded from this that all perfections can co-exist together in a single entity.

In more recent times, Kurt Gödel, Charles Hartshorne, Norman Malcolm and Alvin Plantinga have all presented much-discussed ontological arguments which bear interesting connections to the earlier arguments of St. Anselm, Descartes and Leibniz. Of these, the most interesting are those of Gödel and Plantinga; in these cases, however, it is unclear whether we should really say that these authors claim that the arguments are *proofs* of the existence of God.

Critiques of ontological arguments begin with Gaunilo, a contemporary of St. Anselm. Perhaps the best known criticisms of ontological arguments are due to Immanuel Kant, in his *Critique of Pure Reason*. Most famously, Kant claims that ontological arguments are vitiated by their reliance upon the implicit

assumption that "existence" is a predicate. However, as Bertrand Russell observed, it is much easier to be persuaded that ontological arguments are no good than it is to say exactly what is wrong with them. This helps to explain why ontological arguments have fascinated philosophers for almost a thousand years.

In various ways, the account provided to this point is rough, and susceptible of improvement. Sections 1 - 5 in what follows provide some of the requisite embellishments, though -- as is usually the case in philosophy -- there are many issues taken up here which could be pursued at much greater length. Sections 6 - 8 take up some of the central questions at a slightly more sophisticated level of discussion:

---

1. History of Ontological Arguments

1078: St. Anselm, *Proslogion*. Followed soon after by Gaunilo's critique *In Behalf of the Fool*.

1264: St. Thomas Aquinas, *Summa*. Criticises an argument which somehow descends from St. Anselm.

1637: Descartes, *Meditations*. The Objections -- particularly those of Caterus and Gassendi -- and the Replies contain much valuable discussion of the Cartesian arguments.

c1680: Spinoza, *Ethics*. Intimations of a defensible mereological ontological argument, albeit one whose conclusion is not (obviously) endowed with religious significance.

1709: Leibniz, *New Essays Concerning Human Understanding*. Contains Leibniz's attempt to complete the Cartesian argument by showing that the Cartesian conception of God is not inconsistent.

1776: Hume, *Dialogues Concerning Natural Religion*. Part IX is a general attack on *a priori* arguments (both analytic and synthetic). Includes a purported demonstration that no such arguments can be any good.

1787: Kant, *Critique of Pure Reason*. Contains famous attack on traditional theistic arguments. Three objections to "the ontological argument", including the famous objection based on the dictum that existence is not a predicate.

1831: Hegel, *Lectures of 1831*. Famous assertion -- uncontaminated by argument -- of the correctness of ontological arguments.

1884: Frege, *Foundations of Arithmetic*. Existence is a second-order predicate. First-order existence claims are meaningless. So ontological arguments -- whose conclusions are first-order existence claims -- are doomed.

1941: Hartshorne, *Man's Vision of God*. Defence of modal ontological arguments, allegedly derived from Proslogion 3.

1960: Malcolm, "Anselm's Ontological Argument". Defence of modal ontological arguments by a famous ordinary philosopher.

1970: Lewis, "Anselm and Actuality". The key critique of ontological arguments. All ontological arguments are either invalid or question-begging; moreover, in many cases, they have two closely related readings, one of which falls into each of the above categories.

1974: Plantinga, *The Nature of Necessity*. Plantinga's "victorious" modal ontological argument.

1995: Gödel, *Collected Works Volume III*. Gödel's ontological argument.

2. Taxonomy of Ontological Arguments

According to the taxonomy of Oppy (1995), there are seven major kinds of ontological arguments, viz:

1. definitional ontological arguments;
2. conceptual (or hyperintensional) ontological arguments;
3. modal ontological arguments;
4. Meinongian ontological arguments;
5. experiential ontological arguments;
6. mereological ontological arguments; and
7. ‘Hegelian’ ontological arguments.

Examples of each follow. These are mostly toy examples. But they serve to highlight the deficiencies which more complex examples also share.

> 1. God is a being which has every perfection. (This is true as a matter of definition.) Existence is a perfection. Hence God exists.

> 2. I conceive of a being than which no greater can be conceived. If a being than which no greater can be conceived does not exist, then I can conceive of a being greater than a being than which no greater can be conceived -- namely, a being than which no greater can be conceived that exists. I cannot conceive of a being greater than a being than which no greater can be conceived. Hence, a being than which no greater can be conceived exists.

> 3. It is possible that that God exists. God is not a contingent being, i.e., either it is not

possible that God exists, or it is necessary that God exists. Hence, it is necessary that God exists. Hence, God exists. (See Malcolm (1960), Hartshorne (1965), and Plantinga (1974) for closely related arguments.)

4. [It is analytic, necessary and a priori that] Each instance of the schema "The F G is F" expresses a truth. Hence the sentence "The existent perfect being is existent" expresses a truth. Hence, the existent perfect being is existent. Hence, God is existent, i.e. God exists. (The last step is justified by the observation that, as a matter of definition, if there is exactly one existent perfect being, then that being is God.)

5. The word ‘God’ has a meaning that is revealed in religious experience. The word ‘God’ has a meaning only if God exists. Hence, God exists. (See Rescher (1959) for a live version of this argument.)

6. I exist. Therefore something exists. Whenever a bunch of things exist, their mereological sum also exists. Therefore the sum of all things exists. Therefore God -- the sum of all things -- exists.

7. God must exist.

Of course, this taxonomy is not exclusive: an argument can belong to several categories at once. Moreover, an argument can be ambiguous between a range of readings, each of which belongs to different categories. This latter fact may help to explain part of the curious fascination of ontological arguments. Finally, the taxonomy can be further specialised: there are, for example, at least four importantly different kinds of modal ontological arguments which should be distinguished. (See, e.g., Ross (1969) for a rather different kind of modal ontological argument.)

3. Characterisation of Ontological Arguments

It is not easy to give a good characterisation of ontological arguments. The traditional characterisation involves the use of problematic notions -- analyticity, necessity, and *a priority* -- and also fails to apply to many arguments to which defenders have affixed the label "ontological". (Consider, for example, the claim that I conceive of a being than which no greater can be conceived. This claim is clearly not analytic (its truth doesn't follow immediately from the meanings of the words used to express it), nor necessary (I might never have entertained the concept), nor *a priori* (except perhaps in my own case, though even this is unclear -- perhaps even I don't know independently of experience that I have this concept.)) However, it is unclear how that traditional characterisation should be improved upon.

Perhaps one might resolve to use the label "ontological argument" for any argument which gets classified as "an ontological argument" by its proponent(s). This procedure would make good sense if one thought that there is a natural kind -- ontological arguments -- which our practice carves out, but for which is hard to specify defining conditions. Moreover, this procedure can be adapted as a *pro tem* stop gap: when there

is a better definition to hand, that definition will be adopted instead. On the other hand, it seems worthwhile to attempt a more informative definition.

Focus on the case of ontological arguments for the conclusion that God exists. One characteristic feature of these arguments is the use which they make of "referential vocabulary" -- names, definite descriptions, indefinite descriptions, quantified noun phrases, etc. -- whose ontological commitments -- for occurrences of this vocabulary in "referential position" -- non-theists do not accept.

Theists and non-theists alike (can) agree that there is spatio-temporal, or causal, or nomic, or modal structure to the world (the basis for cosmological arguments); and that there are certain kinds of complexity of organisation, structure and function in the world (the basis for teleological arguments); and so on. But theists and non-theists are in dispute about whether there are perfect beings, or beings than which no greater can be conceived, or ... ; thus, theists and non-theists are in dispute about the *indirect* subject matter of the premises of ontological arguments.

Of course, the premises of ontological arguments often do not deal directly with perfect beings, beings than which no greater can be conceived, etc.; rather, they deal with descriptions of, or ideas of, or concepts of, or the possibility of the existence of, these things. However, the basic point remains: ontological arguments require the use of vocabulary which non-theists should certainly find problematic when it is used in ontologically committing contexts (i.e not inside the scope of prophylactic operators -- such as "according to the story" or "by the lights of theists" or "by the definition" -- which can be taken to afford protection against unwanted commitments).

Note that this characterisation does not beg the question against the possibility of the construction of a successful ontological argument -- i.e., it does not lead immediately to the conclusion that all ontological arguments are question-begging (in virtue of the ontologically committing vocabulary which they employ). For it may be that the vocabulary in question only gets used in premises under the protection of prophylactic operators (which ward off the unwanted commitments.) Of course, there will then be questions about whether the resulting arguments can possibly be valid -- how could the commitments turn up in the conclusion if they are not there in the premises? -- but those are further questions, which would remain to be addressed.

4. Objections to Ontological Arguments

Objections to ontological arguments take many forms. Some objections are intended to apply only to particular ontological arguments, or particular forms of ontological arguments; other objections are intended to apply to all ontological arguments. It is a controversial question whether there are any successful general objections to ontological arguments.

One general criticism of ontological arguments which have appeared hitherto is this: none of them is *persuasive*, i.e., none of them provides those who do not already accept the conclusion that God exists -- and who are reasonable, reflective, well-informed, etc. -- with either a *pro tanto* reason or an all-things-

considered reason to accept that conclusion. Any reading of any ontological argument which has been produced so far which is sufficiently clearly stated to admit of evaluation yields a result which is invalid, or possesses a set of premises which it is clear in advance that no reasonable, reflective, well-informed, etc. non-theists will accept, or has a benign conclusion which has no religious significance, or else falls prey to more than one of the above failings.

For each of the families of arguments introduced in the earlier taxonomy, we can give general reasons why arguments of that family fall under the general criticism. In what follows, we shall apply these general considerations to the exemplar arguments introduced in section 2.

(1) Definitional arguments: These are arguments in which ontologically committing vocabulary is introduced solely via a definition. An obvious problem is that claims involving that vocabulary cannot then be non-question-beggingly detached from the scope of that definition. (The inference from ‘By definition, God is an existent being’ to ‘God exists’ is patently invalid; while the inference to ‘By definition, God exists’ is valid, but uninteresting. In the example given earlier, the premises licence the claim that, as a matter of definition, God possesses the perfection of existence. But, as just noted, there is no valid inference from this claim to the further claim that God exists.)

(2) Conceptual arguments: These are arguments in which ontologically committing vocabulary is introduced solely within the scope of hyperintensional operators (e.g. ‘believes that’, ‘conceives of’, etc.). Often, these operators have two readings, one of which can cancel ontological commitment, and the other of which cannot. On the reading which can give cancellation (as in the most likely reading of ‘John believes in Santa Claus’), the inference to a conclusion in which the ontological commitment is not cancelled will be invalid. On the reading which cannot cancel ontological commitment (as in that reading of ‘John thinks about God’ which can only be true if there is a God to think about), the premises are question-begging: they incur ontological commitments which non-theists reject. In our sample argument, the claim, that I conceive of an existent being than which no greater being can be conceived, admits of the two kinds of readings just distinguished. On the one hand, on the reading which gives cancellation, the inference to the conclusion that there is a being than which no greater can be conceived is plainly invalid. On the other hand, on the reading in which there is no cancellation, it is clear that this claim is one which no reasonable, etc. non-theist will accept: if you doubt that there is a being than which no greater can be conceived, then, of course, you doubt whether you can have thoughts *about* such a being.

(3) Modal arguments: These are arguments with premises which concern modal claims about God, i.e., claims about the possibility or necessity of God's attributes and existence. Suppose that we agree to think about possibility and necessity in terms of possible worlds: a claim is possibly true just in case it is true in at least one possible world; a claim is necessarily true just in case it is true in every possible world; and a claim is contingent just in case it is true in some possible worlds and false in others. Some theists hold that God is a necessarily existent being, i.e., that God exists in every possible world. Non-theists do not accept the claim that God exists in the actual world. Plainly enough, non-theists and necessitarian theists disagree about the layout of logical space, i.e., the space of possible worlds. The sample argument consists, in effect, of two premises: one which says that God exists in at least one possible world; and one which says that God exists in all possible worlds if God exists in any. It is perfectly obvious that no non-

theist can accept this pair of premises. Of course, a non-theist can allow -- if they wish -- that there are possible worlds in which there are contingent Gods. However, it is quite clear that no rational, reflective, etc. non-theist will accept the pair of premises in the sample argument.

(4) Meinongian arguments: These are arguments which depend somehow or other on Meinongian theories of objects. Consider the schema ‘The F G is F’. Naive Meinongians will suppose that if F is instantiated with any property, then the result is true (and, quite likely, necessary, analytic and a priori). So, for example, the round square is round; the bald current King of France is bald; and so on. However, more sophisticiated Meinongians will insist that there must be some restriction on the substitution instances for F, in order to allow one to draw the obvious and important ontological distinction between the following two groups: {Bill Clinton, the sun, the Eiffel Tower} and {Santa Claus, Mickey Mouse, the round square}. Choice of vocabulary here is controversial: Let us suppose (for the sake of example) that the right thing to say is that the former things exist and the latter do not. Under this supposition, ‘existent’ will not be a suitable substitution instance for F -- obviously, since we all agree that there is no existent round square. Of course, nothing hangs on the choice of ‘existent’ as the crucial vocabulary. The point is that non-theists are not prepared to include god(s) in the former group of objects -- and hence will be unpersuaded by any argument which tries to use whatever vocabulary is used to discriminate between the two classes as the basis for an argument that god(s) belong to the former group. (Cognoscenti will recognise that the crucial point is that Meinongian ontological arguments fail to respect the distinction between nuclear (assumptible, characterising) properties and non-nuclear (non-assumptible, non-characterising) properties. It should, of course, be noted that neither Meinong, nor any of his well-known modern supporters -- e.g. Terence Parsons, Richard Sylvan -- ever endorses a Meinongian ontological argument; and it should also be noted that most motivate the distinction between nuclear and non-nuclear properties in part by a need to avoid Meinongian ontological arguments. The reason for calling these arguments "Meinongian" is that they rely on quantification over -- or reference to -- non-existent objects; there is no perjorative intent in the use of this label.)

(5) Experiential arguments: These are arguments which try to make use of ‘externalist’ or ‘object-involving’ accounts of content. It should not be surprising that they fail. After all, those accounts of content need to have something to say about expressions which fail to refer (‘Santa Claus’, ‘phlogiston’, etc.). But, however the account goes, non-theists will insist that expressions which purport to refer to god(s) should be given exactly the same kind of treatment.

(6) Mereological arguments: Those who dislike mereology will not be impressed by these arguments. However, even those who accept principles of unrestricted composition -- i.e., who accept principles which claim, e.g., that, whenever there are some things, there is something which is the sum or fusion of all of those things -- need not be perturbed by them: for it is plausible to think that the conclusions of these arguments have no religious significance whatsoever -- they are merely arguments for, e.g., the existence of the physical universe.

(7) ‘Hegelian arguments’: Since these are not strictly speaking arguments -- but merely unsupported assertions -- there is nothing to refute.

Even if the forgoing analyses are correct, it is important to note that no argument has been given for the conclusion that no ontological argument *can* be successful. Even if all of the kinds of arguments produced to date are pretty clearly unsuccessful -- i.e., not such as ought to give non-theists reason to accept the conclusion that God exists -- it remains an open question whether there is some other kind of hitherto undiscovered ontological argument which does succeed. (Perhaps it is worth adding here that there is fairly widespread consensus, even amongst theists, that no known ontological arguments for the existence of God are persuasive. Most categories of ontological argument have some actual defenders; but none has a large following.)

Many other objections to (some) ontological arguments have been proposed. All of the following have been alleged to be the key to the explanation of the failure of (at least some) ontological arguments: (1) existence is not a predicate (see, e.g., Kant, Smart (1955), Alston (1960)); (2) the concept of god is meaningless/incoherent/ inconsistent (see, e.g., Findlay (1949)); (3) ontological arguments are ruled out by "the missing explanation argument" (see Johnston (1992); (4) ontological arguments all trade on mistaken uses of singular terms (see, e.g., Barnes (1972); (5) existence is not a perfection (see almost any textbook in philosophy of religion); (6) ontological arguments presuppose a Meinongian approach to ontology (see, e.g., Dummett (1993)); and (7) ontological arguments are question-begging, i.e., presuppose what they set out to prove (see, e.g., Rowe (1989)). There are many things to say about these objections: the most important point is that almost all of them require far more controversial assumptions than non-theists require in order to be able to reject ontological arguments with good conscience. Trying to support most of these claims merely in order to beat up on ontological arguments is like using a steamroller to crack a nut (in circumstances in which one is unsure that one can get the steamroller to move!).

Of course, all of the above discussion is directed merely to the claim that ontological arguments are not dialectically efficacious -- i.e., they give reasonable non-theists no reason to change their views. It might be wondered whether there is some other use which ontological arguments have -- e.g., as Plantinga claims, in establishing the reasonableness of theism. This seems unlikely. After all, at best these arguments show that certain sets of sentences (beliefs, etc.) are incompatible -- one cannot reject the conclusions of these arguments while accepting their premises. But the arguments themselves say nothing about the reasonableness of accepting the premisses. So the arguments themselves say nothing about the (unconditional) reasonableness of accepting the conclusions of these arguments. Those who are disposed to think that theism is irrational need find nothing in ontological arguments to make them change their minds (and those who are disposed to think that theism is true should take no comfort from them either).

5. Parodies of Ontological Arguments

Positive ontological arguments -- i.e., arguments FOR the existence of god(s) -- invariably admit of various kinds of parodies, i.e., parallel arguments which seem at least equally acceptable to non-theists, but which establish absurd or contradictory conclusions. For many positive ontological arguments, there are parodies which purport to establish the non-existence of god(s); and for many positive ontological arguments there are lots (usually a large infinity!) of similar arguments which purport to establish the

existence of lots (usally a large infinity) of distinct god-like beings. Here are some modest examples:

(1) By definition, God is a non-existent being who has every (other) perfection. Hence God does not exist.

(2) I conceive of a being than which no greater can be conceived except that it only ever creates N universes. If such a being does not exist, then we can conceive of a greater being -- namely, one exactly like it which does exist. But I cannot conceive of a being which is greater in this way. Hence, a being than which no greater can be conceived except that it only ever creates N universes exists.

(3) It is possible that God does not exist. God is not a contingent being, i.e., either it is not possible that God exists, or it is necessary that God exists. Hence it is not possible that God exists. Hence God does not exist.

(4) It is analytic, necessary, and a priori that the F G is F. Hence, the existent perfect being who creates exactly N universes is existent. Hence the perfect being who creates exactly N universes exists.

There are many kinds of parodies on Ontological Arguments. The aim is to construct arguments which non-theists can reasonably claim to have no more reason to accept than the original Ontological Arguments themselves. Of course, theists may well be able to hold that the originals are sound, and the parodies not -- but that is an entirely unrelated issue. (All theists -- and no non-theists -- should grant that the following argument is sound, given that the connectives are to be interpretted classically: "Either 2+2=5, or God exists. Not 2+2=5. Hence God exists." It should be completely obvious that this argument is useless.)

There are some very nice parodic discussions of Ontological Arguments in the literature. A particularly pretty one is due to Raymond Smullyan, in *5000 BC and Other Philosophical Fantasies*, in which the argument is attributed to "the unknown Dutch theologian van Dollard". A relatively recent addition to the genre is described in Grey (2000), though the date of its construction is uncertain. It is the work of Douglas Gasking, one time Professor of Philosophy at the University of Melbourne (with emendations by William Grey and Denis Robinson):

1. The creation of the world is the most marvellous achievement imaginable.
2. The merit of an achievement is the product of (a) its intrinsic quality, and (b) the ability of its creator.
3. The greater the disability or handicap of the creator, the more impressive the achievement.
4. The most formidable handicap for a creator would be non-existence.
5. Therefore, if we suppose that the universe is the product of an existent creator, we can conceive a greater being -- namely, one who created everything while not existing.
6. An existing God, therefore, would not be a being than which a greater cannot be conceived, because an even more formidable and incredible creator would be a God which did not exist.
7. (Hence) God does not exist.

This parody -- at least in its current state -- seems to me to be inferior to other parodies in the literature, including the early parodies of Gaunilo and Caterus. To mention but one difficulty, while we *might* suppose that it would be a greater achievement to create something if one did not exist than if one did exist, it doesn't follow from this that a non-existent creator is greater (*qua* being) than an existent creator. Perhaps it might be replied that this objection fails to take the first premise into account: if the creation of the world really is "the most marvellous achievement imaginable", then surely there is some plausibility to the claim that the creator must have been non-existent (since that would make the achievement more marvellous than it would otherwise have been). But what reason is there to believe that the creation of the world is "the most marvellous achievement imaginable", in the sense which is required for this argument? Surely it is quite easy to imagine even more marvellous achievements -- e.g., the creation of many worlds at least as good as this one! (Of course, one might also want to say that, in fact, one cannot conceive of a non-existent being's actually creating something: that is literally inconceivable. Etc.)

6. Gödel's Ontological Argument

There is a small, but steadily growing, literature on the ontological arguments which Gödel developed in his notebooks, but which did not appear in print until well after his death. These arguments have been discussed, annotated and amended by various leading logicians: the upshot is a family of arguments with impeccable *logical* credentials. (Interested readers are referred to Sobel (1987), Anderson (1990), Adams (1995b), and Hazen (1999) for the history of these arguments, and for the scholarly annotations and emendations.) Here, I shall give a brief presentation of the version of the argument which is developed by Anderson, and then make some comments on *that* version. This discussion follows the presentation and discussion in Oppy (1996)(2000).

> Definition 1: *x* is God-like iff *x* has as essential properties those and only those properties which are positive
>
> Definition 2: *A* is an essence of *x* iff for every property *B*, *x* has *B* necessarily iff *A* entails *B*
>
> Definition 3: *x* necessarily exists iff every essence of *x* is necessarily exemplified
>
> Axiom 1: If a property is positive, then its negation is not positive.
>
> Axiom 2: Any property entailed by -- i.e., strictly implied by -- a positive property is positive
>
> Axiom 3: The property of being God-like is positive
>
> Axiom 4: If a property is positive, then it is necessarily positive
>
> Axiom 5: Necessary existence is positive

> Axiom 6: For any property *P*, if *P* is positive, then being necessarily P is positive.
>
> Theorem 1: If a property is positive, then it is consistent, i.e., possibly exemplified.
>
> Corollary 1: The property of being God-like is consistent.
>
> Theorem 2: If something is God-like, then the property of being God-like is an essence of that thing.
>
> Theorem 3: Necessarily, the property of being God-like is exemplified.

Given a sufficiently generous conception of properties, and granted the acceptability of the underlying modal logic, the listed theorems do follow from the axioms. (This point was argued in detail by Dana Scott, in unpublished lecture notes which circulated for many years. It is also made by Sobel, Anderson, and Adams.) So, criticisms of the argument are bound to focus on the axioms, or on the other assumptions which are required in order to construct the proof.

Some philosophers have denied the acceptability of the underlying modal logic. And some philosophers have rejected generous conceptions of properties in favour of sparse conceptions according to which only some predicates express properties. But suppose that we adopt neither of these avenues of potential criticism of the proof. What else might we say against it?

One important point to note is that no *definition* of the notion of "positive property" is supplied with the proof. At most, the various axioms which involve this concept can be taken to provide a *partial* implicit definition. If we suppose that the "positive properties" form a set, then the axioms provide us with the following information about this set:

1. If a property belongs to the set, then its negation does not belong to the set.
2. The set is closed under entailment.
3. The property of having as essential properties just those properties which are in the set is itself a member of the set.
4. The set has exactly the same members in all possible worlds.
5. The property of necessary existence is in the set.
6. If a property is in the set, then the property of having that property necessarily is also in the set.

On Gödel's theoretical assumptions, we can show that *any* set which conforms to (1) - (6) is such that the property of having as essential properties just those properties which are in *that* set is exemplified. Gödel wants us to conclude that there is just one intuitive, theologically interesting set of properties which is such that the property of having as essential properties just the properties in that set is exemplified. But, on the one hand, what reason do we have to think that there is *any* theologically interesting set of properties which conforms to the Gödelian specification? And, on the other hand, what reason do we have

to deny that, if there is one set of theologically interesting set of properties which conforms to the Gödelian specification, then there are many theologically threatening sets of properties which also conform to that specification?

In particular, there is some reason to think that the Gödelian ontological argument goes through just as well -- or just as badly -- with respect to other sets of properties (and in ways which are damaging to the original argument). Suppose that there is some set of independent properties {*I*, *G*1, *G*2, ...} which can be used to generate the set of positive properties by closure under entailment and "necessitation". ("Independence" means: no one of the properties in the set is entailed by all the rest. "Necessitation" means: if *P* is in the set, then so is necessarily having *P*. *I* is the property of having as essential properties just those properties which are in the set. *G*1, *G*2, ... are further properties, of which we require at least two.) Consider any proper subset of the set {*G*1, *G*2, ...} -- {*H*1, *H*2, ...}, say -- and define a new generating set {*I**, *H*1, *H*2, ...}, which *I** is the property of having as essential properties just those properties which are in the newly generated set. A "proof" parallel to that offered by Gödel "establishes" that there is a being which has as essential properties just those properties in this new set. If there are as few as 7 independent properties in the original generating set, then we shall be able to establish the existence of 720 distinct"God-like" creatures by the kind of argument which Gödel offers. (The creatures are distinct because each has a different set of essential properties.)

Even if the above considerations are sufficient to cast doubt on the credentials of Gödel's "proof", they do not pinpoint where the "proof" goes wrong. If we accept that the role of Axioms 1, 2, 4, and 6 is really just to constrain the notion of "positive property" in the right way -- or, in other words, if we suppose that Axioms 1, 2, 4, and 6 are "analytic truths" about "positive properties" -- then there is good reason for opponents of the "proof" to be sceptical about Axioms 3 and 5. Kant would not have been happy with Axiom 5; and there is at least some reason to think that whether the property of being God-like is "positive" ought to depend upon whether or not there is a God-like being.

7. Plantinga's Ontological Argument

The "victorious" modal ontological argument of Plantinga (1974) goes roughly as follows: Say that an entity possesses "maximal excellence" iff it is omnipotent, omnscient, and morally perfect. Say, further, that an entity possesses "maximal greatness" iff it possesses maximal excellence in every possible world -- that is, iff it is necessarily existent and necessarily maximally excellent. Then consider the following argument:

1. There is a possible world in which there is an entity which possesses maximal greatness.
2. (Hence) There is an entity which possesses maximal greatness.

Under suitable assumptions about the nature of accessibility relations between possible worlds, this argument is valid: from it is possible that it is necessary that *p*, one can infer that it is necessary that *p*. Setting aside the possibility that one might challenge this widely accepted modal principle, it seems that opponents of the argument are bound to challenge the acceptability of the premise.

And, of course, they do. Let's just run the argument in reverse.

1. There is no entity which possesses maximal greatness.
2. (Hence) There is no possible world in which there is an entity which possesses maximal greatness.

Plainly enough, if you do not already accept the claim that there is an entity which possesses maximal greatness, then you won't agree that the first of these arguments is more acceptable than the second. So, as a proof of the existence of a being which posseses maximal greatness, Plantinga's argument seems to be a non-starter.

Perhaps somewhat surprisingly, Plantinga himself agrees: the "victorious" modal ontological argument is not a proof of the existence of a being which possesses maximal greatness. But how, then, is it "victorious"? Plantinga writes: "Our verdict on these reformulated versions of St. Anselm's argument must be as follows. They cannot, perhaps, be said to *prove* or *establish* their conclusion. But since it is rational to accept their central premise, they do show that it is rational to accept that conclusion." (Plantinga (1974:221)).

It is pretty clear that Plantinga's argument does not show what he claims that it shows. Consider, again, the argument: "Either God exists, or 2+2=5. It is not the case that 2+2=5. So God exists." It is just a mistake for a theist to say: "Since the premise is *true* (and the argument is valid), this argument *shows* that the conclusion of the argument is *true*". No-one thinks that that argument *shows* any such thing. Similarly, it is just a mistake for a theist to say: "Since it is *rational* to accept the premise (and the argument is valid), this argument *shows* that it is *rational* to accept the conclusion of the argument". Again, no one thinks that that argument *shows* any such thing. But why don't these arguments *show* the things in question? There is room for argument about this. But it is at least plausible to claim that, in each case, any even minimally rational person who has doubts about the claimed status of the conclusion of the argument will have exactly the same doubts about the claimed status of the premise. If, for example, I doubt that it is rational to accept the claim that God exists, then you can quite sure that I will doubt that it is rational to accept the claim that either 2+2=5 or God exists. But, of course, the very same point can be made about Plantinga's argument: anyone with even minimal rationality who understands the premise and the conclusion of the argument, and who has doubts about the claim that there is an entity which possesses maximal greatness, will have exactly the same doubts about the claim that there is a possible world in which there is an entity which possesses maximal greatness.

For further discussion of Plantinga's argument, see -- for example -- Adams (1988), Chandler (1993), Oppy (1995:70-78, 248-259), Tooley (1981), and van Inwagen (1977)).

8. St. Anselm's Ontological Argument

There is an enormous literature on the material in *Proslogion II-III*. Some commentators deny that St. Anselm tried to put forward any proofs of the existence of God. Even among commentators who agree

that St. Anselm intended to prove the existence of God, there is disagreement about where the proof is located. Some commentators claim that the main proof is in *Proslogion II*, and that the rest of the work draws out corollaries of that proof (see, e.g., Charlesworth (1965)). Other commentators claim that the main proof is in *Prologion III*, and that the proof in *Proslogion II* is merely an inferior first attempt (see, e.g., Malcolm (1960)). Yet other commentators claim that there is a single proof which spans at least *Proslogion II-III* -- see, e.g., Campbell (1976) and, perhaps, the entire work -- see, e.g., La Croix (1972). I shall ignore this aspect of the controversy about the *Proslogion*. Instead, I shall just focus on the question of the analysis of the material in *Proslogion II* on the assumption that there is an independent argument for the existence of God which is given therein.

Here is one translation of the crucial part of *Proslogion II* (due to William Mann (1972:260-1); alternative translations can be found in Barnes (1972), Campbell (1976), Charlesworth (1965), and elsewhere):

> Thus even the fool is convinced that something than which nothing greater can be conceived is in the understanding, since when he hears this, he understands it; and whatever is understood is in the understanding. And certainly that than which a greater cannot be conceived cannot be in the understanding alone. For if it is even in the understanding alone, it can be conceived to exist in reality also, which is greater. Thus if that than which a greater cannot be conceived is in the understanding alone, then that than which a greater cannot be conceived is itself that than which a greater can be conceived. But surely this cannot be. Thus without doubt something than which a greater cannot be conceived exists, both in the understanding and in reality.

There have been many ingenious attempts to find an argument which can be expressed in modern logical formalism, which is logically valid, and which might plausibly be claimed to be *the* argument which is expressed in this passage. To take a few prime examples, Adams (1971), Barnes (1972) and Oppenheimer and Zalta (1991) have all produced formally valid analyses of the argument in this passage. We begin with a brief presentation of each of these analyses, preceded by a presentation of the formulation of the argument given by Plantinga (1967), and including a presentation of some of the formulations of Lewis (1970). (Chambers (2000) works with the analysis of Adams (1971).)

Plantinga

1. God exists in the understanding but not in reality.	(Assumption for *reductio*)
2. Existence in reality is greater than existence in the understanding alone.	(Premise)
3. A being having all of God's properties plus existence in reality can be conceived.	(Premise)
4. A being having all of God's properties plus existence in reality is greater than God	(From (1) and (2).)
5. A being greater than God can be conceived.	(From (3) and (4).)

6. It is false that a being greater than God can be conceived. (From definition of "God".)
7. Hence, it is false that God exists in the understanding but not in reality. (From (1), (5), (6).)
8. God exists in the understanding. (Premise, to which even the Fool agrees.)
9. Hence God exists in reality. (From (7), (8).)

Barnes

1. The Fool understands the expression "the being than which no greater can be conceived". (Premise)
2. If a person understands an expression "*b*", then *b* is in that person's understanding. (Premise)
3. If a thing is in a person's understanding, then the person can conceive of that thing's existing in reality. (Premise)
4. Each thing which exists in reality is greater than any thing which exists only in the understanding. (Premise)
5. If a person can conceive of something, and that thing entails something else, then the person can also conceive of that other thing. (Premise)
6. If a person can conceive that a specified object has a given property, then that person can conceive that something or other has that property. (Premise)
7. Hence the being than which no greater can be conceived exists in reality. (From (1)-(6), by a complex series of steps here omitted.)

Adams

1. There is a thing *x*, and a magnitude *m*, such that *x* exists in the understanding, *m* is the magnitude of *x*, and it it not possible that there is a thing *y* and a magnitude *n* such that *n* is the magnitude of *y* and $n>m$. (Premise)
2. For any thing *x* and magnitude *m*, if *x* exists in the understanding, *m* is the magnitude of *x*, and it is not possible that there is a thing *y* and magnitude *n* such that *n* is the magnitude of *y* and $n>m$, then it is possible that *x* exists in reality. (Premise)
3. For any thing *x* and magnitude *m*, if *m* is the magnitude of *x*, and it it not possible that there is a thing *y* and a magnitude *n* such that *n* is the magnitude of *y* and $n>m$, and *x* does not exist in reality, then it is not possible that if *x* exists in reality then there is a magnitude *n* such that *n* is greater than *m* and *n* is the magnitude of *x*. (Premise)

4. (Hence) There is a thing *x* and a magnitude *m* such that *x* exist in the understanding, and *x* exists in reality, and *m* is the magnitude of *x*, and it it not possible that there is a thing *y* and a magnitude *n* such that *n* is the magnitude of *y* and *n*>*m*. (From 1, 2, 3)

Lewis

1. For any understandable being *x*, there is a world *w* such that *x* exists in *w*. (Premise)
2. For any understandable being *x*, and for any worlds *w* and *v*, if *x* exists in *w*, but *x* does not exist in *v*, then the greatness of *x* in *w* exceeds the greatness of *x* in *v*. (Premise)
3. There is an understandable being *x* such that for no world *w* and being *y* does the greatness of *y* in *w* exceed the greatness of *x* in the actual world. (Premise)
4. (Hence) There is a being *x* existing in the actual world such that for no world *w* and being *y* does the greatness of *y* in *w* exceed the greatness of *x* in the actual world. (From (1)-(3).)

Lewis also suggests an alternative to (3) which yields a valid argument:

> (3′) There is an understandable being *x* such that for no worlds *v* and *w* and being *y* does the greatness of *y* in w exceed the greatness of *x* in *v*.

and two alternatives to (3) -- not presented here -- which yield invalid arguments. (Of course, there further two alternatives are crucial to Lewis' overall analysis of the passage: essentially, Lewis suggests that Anselm equivocates between an invalid argument with plausible premises and a valid argument with question-begging premises. In this respect, Lewis' analysis is quite different from the other analyses currently under discussion.)

Oppenheimer and Zalta

1. There is (in the understanding) something than which there is no greater. (Premise)
2. (Hence) There is (in the understanding) a unique thing than which there is no greater. (From (1), assuming that the "greater-than" relation is connected.)
3. (Hence) There is (in the understanding) something which is the thing than which there is no greater. (From (2), by a theorem about descriptions.)
4. (Hence) There is (in the understanding) nothing which is greater than the thing than which there is no greater. (From (3), by another theorem about descriptions.)

5. If that thing than which there is no greater does not exist (in reality), then there is (in the understanding) something which is greater than that thing than which there is no greater. (Premise)
6. (Hence) That thing than which there is no greater exists (in reality). (From (4) and (5).)
7. (Hence) God exist. (From (6).)

Critical Appraisal

Considered as interpretations of the argument presented in the *Proslogion*, these formulations are subject to various kinds of criticisms.

First, the modal interpretations of Lewis (1970) and Adams (1971) don't square very well with the rest of the *Proslogion*: the claim that "being than which no greater can be conceived" should be read as "being than which no greater is possible" would have us render the claim of *Proslogion 15* to be that God is a being greater than any which is possible. And that is surely a bad result.

Second, the Meinongian interpretations of Barnes (1972), Adams (1971) and Oppenheimer and Zalta (1991) produce arguments which, given the principles involved, could easily be much simplified, and which are *obviously* vulnerable to Gaunilo-type objections.

Consider, for example, the case of Oppenheimer and Zalta. They have Anselm committed to the claim that if anyone can understand the phrase "that than which F", then there is something in the understanding such that F (see their footnote 25); and they also have him committed to the claim that if there is something which is the F-thing, then it -- i.e., the F-thing -- has the property F (see page 7). Plainly though, if Anselm is really committed to these principles, then he could hardly fail to be committed to the more general principles: (1) if anyone can understand the phrase "an F", then there is at least one F-thing in the understanding; and (2) if there are some things which are the F-things, then they -- i.e., the F-things -- must have the property F. (It would surely be absurd to claim that Anselm is only committed to the less general principles: what could possibly have justified the restrictions to the special cases?)

But, then, mark the consequences. We all understand the expression "an existent perfect being". So, by the first claim, there is at least one existent perfect being in the understanding. And, by the second claim, any existent perfect being is existent. So, from these two claims combined, there is -- in reality -- at least one existent perfect being.

This argument gives Anselm everything that he wants, and very much more briefly. (The *Proslogion* goes on and on, trying to establish the properties of that than which no greater can be conceived. How much easier if we can just explicitly build all of the properties which want to "derive" into the initial description.) So, if Anselm really were committed to the principles which Oppenheimer and Zalta appear

to attribute to him, it is hard to understand why he didn't give the simpler argument. And, of course, it is also hard to understand why he didn't take Gaunilo's criticism. After all, when it is set out in this way, it is obvious that the argument proves far too much.

Third, some of the arguments have Anselm committed to claims about greatness which do not seem to correspond with what he actually says. The natural reading of the text is that, if two beings are identical save that one exists only in the understanding and the other exists in reality as well, then the latter is greater than the former. But Barnes (1971), for example, has Anselm committed to the much stronger claim that any existing thing is greater than every non-existent thing.

Given these kinds of considerations, it is natural to wonder whether there are *better* interpretations of *Proslogion II* according to which the argument in question turns out *NOT* to be logically valid. Here is a modest attempt to provide such an analysis:

We start with the claim that the Fool understands the expression "being than which no greater can be conceived", i.e., even the Fool can entertain the idea or possess the concept of a being than which no greater can be conceived. Now, entertaining this idea or possessing this concept requires the entertainer or possessor to recognise certain relationships which hold between given properties and the idea or concept in question. For example, given that you possess the concept of, or entertain the idea of, a smallest really existent Martian, it follows that you must recognise some kind of connection between the properties of being a Martian, really existing, and being smaller than other really existing Martians, and the concept or idea in question.

Following Anselm, we might say that, since you understand the expression "smallest really existent Martian", there is, in your understanding, at least one smallest really existent Martian. (Or, apparently following Descartes, one might say that real existence is "part of" -- or "contained in" -- the idea of a smallest really existent Martian.) However, in saying this, it must be understood that we are not actually predicating properties of anything: we aren't supposing that there is something which possesses the properties of being a Martian, really existing, and being no larger than any other Martian. (After all, we can safely suppose, we don't think that any Martians really exist.) In other words, we must be able to have the concept of, or entertain the idea of, a smallest really existing Martian without believing that there really are any smallest Martians. Indeed, more strongly, we must be able to entertain the concept of a smallest really existent Martian -- and to recognise that the property of "really existing" is part of this concept -- while nonetheless maintaining that there are no smallest existent Martians.

It will be useful to introduce vocabulary to mark the point which is being made here. We could, for instance, distinguish between the properties which are *encoded* in an idea or concept, and the properties which are *attributed* in positive atomic beliefs which have that idea or concept as an ingredient. The idea "really existent Santa Claus" encodes the property of real existence; but it is perfectly possible to entertain this idea without attributing real existence to Santa Claus, i.e., without believing that Santa Claus really exists.

We can then apply this distinction to Anselm's argument. On the one hand, the idea "being than which no

greater can be conceived" encodes the property of real existence -- this is what the *reductio* argument establishes (if it establishes anything at all). On the other hand, it is perfectly possible to entertain the idea of a being than which no greater can be conceived -- and to recognise that this idea encodes the property of real existence -- without attributing real existence to a being than which no greater can be conceived, i.e., without believing that a being than which no greater can be conceived really exists.

Of course, the argument which Anselm actually presents pays no attention to this distinction between encoding and attributing -- i.e., between entertaining an idea and bolding a belief -- and nor does it pay attention to various other niceties. We begin from the point that the Fool entertains the idea of that than which no greater can be conceived (because the Fool understands the words "that than which no greater can be conceived"). From this, we move quickly to the claim that even the Fool is "convinced" -- i.e., believes -- that that than which no greater can be conceived possesses the property of existing in the understanding. And then the *reductio* argument is produced to establish that that than which no greater can be conceived cannot exist only in the understanding but must also possess the property of existing in reality as well (and all mention of the Fool, and what it is that the Fool believes, disappears).

As it stands, this is deeply problematic. How are we supposed to regiment the references to the Fool in the argument? Is the *reductio* argument supposed to tell us something about what even the Fool believes, or ought to believe? Are the earlier references to the Fool supposed to be inessential and eliminable? How are we so much as to understand the claim that even the Fool believes that that than which no greater can be conceived exists in the understanding? And how do we get from the Fool's *understanding* the words "that than which no greater can be conceived" to his *believing* that that than which no greater can be conceived possesses the property of existing in the understanding?

Following the earlier line of thought, it *seems* that the argument might go something like this:

1. (Even) the Fool has the concept of that than which no greater can be conceived.
2. (Hence) (Even) the Fool believes that that than which no greater can be conceived exists in the understanding.
3. No one who believes that that than which no greater can be conceived exists in the understanding can reasonably believe that that than which no greater can be conceived exists only in the understanding.
4. (Hence) (Even) the Fool cannot reasonably deny that that than which no greater can be conceived exists in reality
5. (Hence) That than which no greater can be conceived exists in reality.

While this argument does not look very compelling, it is plausible to claim that it would have seemed compelling to someone who failed to attend to the distinction which we have drawn between entertaining ideas and holding beliefs, and who was also a bit hazy on the distinction between the vehicles of belief and their contents. When the Fool entertains the concept of that than which no greater can be conceived he recognises that he is entertaining this concept (i.e., he believes that he is entertaining the concept of that than which no greater can be conceived -- or, as we might say, that the concept is in his understanding). Conflating the concept with its object, this gives us the belief that than which no greater

can be conceived possesses the property of existing in the understanding. Now, suppose as hypothesis for *reductio*, that we can reasonably believe that that than which no greater can be conceived possesses the property of existing only in the understanding. Ignoring the distinction between entertaining ideas and holding beliefs, this means that we when we entertain the idea of that than which no greater can be conceived, we entertain the idea of a being which exists only in the understanding. But that is absurd: when we entertain the idea of that than which no greater can be conceived, our idea encodes the property of existing in reality. So there is a contradiction, and we can conclude that, in order to be reasonable, we must believe that that than which no greater can be conceived exists in reality. But if any reasonable person must believe that that than which no greater can be conceived exists in reality, then surely it is the case that that than which no greater can be conceived exists in reality. And so we are done.

No doubt this suggestion about the interpretation of Anselm's argument is deficient in various ways. However, the point of including it is illustrative rather than dogmatic. In the literature, there has been great resistance to the idea that the argument which Anselm gives is one which modern logicians would not hesitate to pronounce invalid. But it is very hard to see why there should be this resistance. (Certainly, it is not something for which there is much argument in the literature.) The text of the *Proslogion* is so rough, and so much in need of polishing, that we should not be too quick to dismiss the suggestion that Anselm's argument is rather more like the argument most recently sketched than it is like the logically valid demonstrations provided by commentators such as Barnes, Adams, and Oppenheimer and Zalta.

Bibliography

Primary Texts

- Anselm, St., *Proslogion*, in *St. Anselm's Proslogion*, M. Charlesworth (ed.), Oxford: OUP, 1965 [Available online, in the Internet Medieval Sourcebook, Paul Halsall (ed.), Fordham University Center for Medieval Studies, translation by David Burr]
- Aquinas, T., *Summa Theologica*, 1272, literally translated by Fathers of the English Dominican Province, London: Burn, Oates & Washbourne, 1920 [Available online, in the Internet Medieval Sourcebook, Paul Halsall (ed.), Fordham University Center for Medieval Studies, translation by David Burr]
- Ayer, A., *Language, Truth and Logic* , second edition, London: Gollancz, 1948
- Descartes, R., *Discourse on Method and The Meditations*, translated with an introduction by F. Sutcliffe, Harmondsworth: Penguin, 1968 [Translation of *The Meditations*, by John Veitch, LL.D., available online]
- Frege, G., *Die Grundlagen der Arithmetik*, Bresnau: Koebner, 1884; translated as *The Foundations of Arithmetic*, J.L. Austin (trans), Oxford: Blackwell, 1974, 2nd rev edition; Original German text available online, (528 KB PDF file), maintained by Alain Blachair, Académie de Nancy-Metz
- Gaunilo, "On Behalf of the Fool", in *St. Anselm's Proslogion*, M. Charlesworth (ed.), Oxford: OUP, 1965

[Available online in the Internet Medieval Sourcebook, Paul Halsall (ed.), Fordham Universit\ y Center for Medieval Studies, translation by David Burr]

- Hegel, G., *The Ontological Proof According to the Lectures of 1831*, in P. Hodgson (ed.), *Lectures on the Philosophy of Religion, Vol. III* Berkeley: University of California Press, 1985, pp.351-8
- Hume, D., *Dialogues Concerning Natural Religion*, 1779, edited with an introduction by H. Aiken, London: Macmillan, 1948
  [Electronic version, edited by James Fieser, available online]
- Kant, I., *Critique of Pure Reason*, 1787, second edition, translated by N. Kemp-Smith, London: Macmillan, 1933
- Leibniz, G., *New Essay Concerning Human Understanding*, 1709, translated by A. Langley, New York: Macmillan, 1896
- Spinoza, B., *The Ethics*, 1677, translation of 1883 by R. Elwes, New York: Dover, 1955
  [Available online, prepared by R. Bombardi, for the Philosophy Web Works project, Middle Tennessee State Univerity]

Other Texts

- Adams, R., 1971, "The Logical Structure of Anselm's Argument", *Philosophical Review* **80**: 28-54
- -----, 1988, "Presumption and the Necessary Existence of God" *Nous* **22**: 19-34
- -----, 1995a, *Leibniz: Determinist, Theist, Idealist* Oxford: Oxford University Press
- -----, 1995b, "Introductory Note to *1970" in K. Gödel *Collected Works Volume III: Unpublished essays and lectures* (editor-in-chief Solomon Feferman), New York: Oxford University Press, pp.388-402
- Alston, W. 1960, "The Ontological Argument Revisited" *Philosophical Review* **69**: 452-74
- Anderson, C., 1990, "Some Emendations on Gödel's Ontological Proof" *Faith and Philosophy* **7**: 291-303
- Barnes, J., 1972, *The Ontological Argument* London:Macmillan
- Campbell, R., 1976, *From Belief to Understanding* Canberra: ANU Press
- Chambers, T., 2000, "On Behalf of the Devil: A Parody of St. Anselm Revisited" Proceedings of the Aristotelian Society", New Series--Volume C: 93-113
- Chandler, H., 1993, "Some Ontological Arguments" *Faith and Philosophy* **10**: 18-32
- Charlesworth, M., 1965, *Anselm's Proslogion* Oxford: Oxford University Press
- Dummett, M., 1993, "Existence". In *The Seas of Language* Oxford: Oxford University Press
- Findlay, J., 1949, "Can God's Existence Be Disproved?" *Mind* **57**: 176-83
- Grey, W., 2000, "Gasking's Proof" *Analysis* **60**:368-370
- Hartshorne, C., 1965, *Anselm's Discovery: A Re-Examination of the Ontological Proof for God's Existence* La Salle, Ill: Open Court
- Hazen, A., 1999, "On Gödel's Ontological Proof" *Australasian Journal of Philosophy* **76**: 361-377
- Henle, P., 1961, "Uses of the Ontological Argument" *Philosophical Review* **70**: 102-9
- Johnston, M., 1992, "Explanation, Response-Dependence, and Judgement-Dependence". In P. Menzies, ed., *Response-Dependent Concepts*. Working Papers in Philosophy, RSSS, ANU, 123-83
- La Croix, R., 1972, *Proslogion II and III: A Third Interpretation of Anselm's Argument* Leiden:

Brill
- Lewis, D., 1970, "Anselm and Actuality", *Nous* **4**: 175-88
- Malcolm, N., 1960, "Anselm's Ontological Arguments" *Philosophical Review* **69**: 41-62
- Mann, W., 1972, "The Ontological Presuppositions of the Ontological Argument", *Review of Metaphysics* **26**: 260-77
- Oppenheimer, P., and Zalta, E., 1991, "On the Logic of the Ontological Argument" in J. Tomberlin (ed.) *Philosophical Perspectives 5: The Philosophy of Religion* Atascadero: Ridgeview: 509-29 [Preprint available online]
- Oppy, G., 1995, *Ontological Arguments and Belief in God*, New York: Cambridge University Press
- -----, 1996, "Gödelian Ontological Arguments" *Analysis* **56**: 226-230
- -----, 2000, "Response to Gettings" *Analysis* **60**: 363-367
- Plantinga, A., 1967, *God and Other Minds* Ithaca: Cornell University Press
- -----, 1974, *The Nature of Necessity* Oxford: Oxford University Press
- Rescher, N., 1959, "The Ontological Proof Revisited" *Australasian Journal of Philosophy* **37**: 138-48
- Ross, J., 1969, *Philosophical Theology* New York: Bobbs-Merrill
- Rowe, W., 1989, "The Ontological Argument" in J. Feinberg (ed.) *Reason and Responsibility*, seventh edition, Belmont, Ca: Wadsworth, pp. 8-17
- Salmon, N., 1987, "Existence" in J. Tomberlin (ed.) *Philosophical Perspectives 1: Metaphysics* Atascadero, Ca: Ridgeview: 49-108
- Smart, J., 1955, "The Existence of God" in A. Flew and A. MacIntyre (eds.) *New Essays in Philosophical Theology* London: SCM Press: 500-509
- Sobel, J., 1987, "Gödel's Ontological Proof". In *On Being and Saying: Essays for Richard Cartwright*, ed. J. Thomson, Cambridge, Mass: MIT Press, pp. 241-61
- Tooley, M., 1981, "Plantinga's Defence of the Ontological Argument" *Mind* **90**: 422-7
- van Inwagen, P., 1977, "Ontological Arguments" *Nous* **11**: 375-395

Other Internet Resources

- Kurt Gödel's Ontological Argument (Christopher Small, University of Waterloo)
- Medieval Sourcebook: Philosophers' Criticisms of Anslem's Ontological Argument for the Being of God (Paul Halsell, Fordham University)
- Handout for a Talk on the Ontological Argument (J. R. Lucas, Oxford University)
- Ontological Argument Revisited by Two Ottoman Muslim Scholars (Umit Dericioglu)
- The Ontological Argument (Kenneth Himma, University of Washington)
- Anselm's Ontological Argument (Gideon Rosen, Princeton University)
- The E-God Argument vs. The Ontological Argument (Valerie McCool, Santa Barbara City College)
- Hegel and Kant on the Ontological Argument (Maria de Lourdes Borges, Federal University of Santa Catarina)

- The Ontological Argument (Sally Haslanger, MIT)
- Anselm's Ontological Argument (Allen Stairs, University of Maryland)

Related Entries

Anselm, Saint [Anselm of Bec, Anselm of Canterbury] | *a priori* justification and knowledge | Descartes, René | existence | God | Gödel, Kurt | Hegel, Georg Wilhelm Friedrich | Kant, Immanuel | logic: informal | logic: modal | Meinong, Alexius

Graham Oppy
Graham.Oppy@arts.monash.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 8, 1996
Content last modified: April 1, 2002

Existence

Like many philosophically interesting notions, existence is at once familiar and rather elusive. Although we have no more trouble with using the verb ‘exists’ than with the two-times table, there is more than a little difficulty in saying just what existence is. Existing seems to be at least as mundane as walking or being hungry. Yet, when we say ‘Tom is hungry’ or ‘Tom is walking’, it may be news to those not in Tom's vicinity, whereas ‘Tom exists’ would be news to no one who knew Tom, and merely puzzling to anyone who did not. Again, we know what it is like to be hungry or to walk, but what is it like to exist, what kind of experience is that? Is it perhaps the experience of being oneself, of being identical with oneself? Yet again, we can readily indicate what is meant by Tom's walking, but surely Tom's existing is not something we can indicate to anyone. On the face of it, there would seem to be no way at all in which we can explain what existing is.

It may be tempting to think that ‘Tom exists’ means merely ‘Tom is real’. In fact, this could be distinctly appealing, for ‘real’ is what has been called an ‘excluder’ predicate, meaning thereby that it attributes nothing positive to Tom, but operates in a purely negative fashion simply to exclude Tom from being imaginary, mythical, fictional, and the like. To say that ‘exists’ meant ‘is real’ would be to say inter alia that it attributed nothing positive to Tom; and that would do much to relieve our frustration at being so fluent in our use of ‘exists’ despite having no idea of its attributing anything positive to Tom. It would be a relief to discover that ‘exists’ attributes nothing positive to him at all.

Unfortunately, this won't do; for among all the negatives that ‘is real’ might be applying to Tom would be not only ‘not imaginary’, ‘not mythical’, etc., but also ‘not nonexistent’. Now, suppose a seer predicted that in two years that a son would be born to Bill and Mary, and that he would be called ‘Tom’. When the prediction was finally fulfilled, we might imagine the seer announcing triumphantly ‘At last Tom exists, exactly as I predicted he would’. If ‘exists’ were an excluder like ‘is real’, then the seer could only be understood as excluding something from Tom; and in this case it would be non-existence. As said by the seer, therefore, ‘At last Tom exists’ could only mean ‘At last Tom is not-nonexistent’. And if he really were to mean that, we should be entitled to ask him just when Tom could ever have been said to be nonexistent, i.e. never to have existed. In fact, before he existed Tom could never even have been referred to, and hence at that time nothing at all could have been attributed to him, not even the property of being nonexistent. Promising as it may have seemed, therefore, ‘Tom exists’ is not to be understood simply as ‘Tom is real’.

Of course, the failure of attempts to understand ‘exists’ as ‘is real’ leaves plenty of room for other suggestions, each proposing to substitute one or more terms for ‘exists’, and thereby to show why our

original disquiet about it and existence has been sadly misplaced. If one thinks that ‘exists’ is readily dispensable in favour of some other (less troublesome) expression, then there will be no difficulty in dismissing the thought of there being some such property or attribute as existence. Alternatively, if one thinks that ‘exists’ is not to be dispensed with in this way, then one might be inclined to continue pursuing the puzzle of just what existence is.

It is probably now reasonably clear that the question of existence is inextricably intertwined with the question of ‘exists’. In some languages, the predicate ‘is’ does duty for ‘exists’, and even in English there are archaic uses of ‘is’ in that role. In discussing existence, therefore, we shall be much concerned also with the predicates ‘is’ and ‘exists’. In this regard, the predominant view on existence among contemporary philosophers of an analytic persuasion might be summarized in two theses, the first of which is the Frege-Russell distinction between four different meanings of ‘is’ - the ‘is’ of existence, of identity, of predication, and of generic implication (inclusion), as illustrated below.

- ‘Socrates is’, rendered in regimented language as ‘$(\exists x)(\text{Socrates} = x)$’.
- ‘Cicero is Tully’, rendered as ‘Cicero = Tully’.
- ‘Socrates is wise’, rendered as ‘Wise(Socrates)’.
- ‘Man is an animal’, rendered as ‘$(x)(\text{Man}(x) \rightarrow \text{Animal}(x))$’.

On this view, the different uses of ‘is’ entail correspondingly different meanings, so different in fact as to have nothing whatever in common. That is to say, they are casually ambiguous rather than being merely systematically ambigous or analogical, which would have been the case had their meanings been inter-related though without being univocal in any way - not even partially.

The second thesis commonly, though not universally, held by analytic philosophers might be summed up in the familiar dictum, ‘Existence is not a predicate’. More accurately, it should be written either as ‘Existence is not a (first-level) property’ or as ‘"Exists" is not a (first-level) predicate’. Before discussing current views on this and the earlier thesis, it will therefore be useful to be reminded of what some earlier philosophers have had to say about existence and, correlatively, about ‘is’ and ‘exists’ as verbs of being.

Earlier Views

Aristotle (384-322 BC): It is reasonably common ground among commentators that the Frege-Russell distinction is not to be found in Aristotle either in regard to the uses or the senses of 'is'. In so far as this applies to the existential and predicative uses, some have explained the former as being merely elliptical for the latter. Thus, 'Socrates is' would be merely elliptical for 'Socrates is a something or other', where the permissible substitutions for 'something or other' are any of Socrates' essential predicates. On this view, therefore, 'is' would be unambiguous because its use was always predicative, either explicitly as in 'Socrates is a man' or merely implicitly as in 'Socrates is'. Not only was the one sense being used, it was being employed with the same force in each case, namely, predicative.

For the preceding view, G. E. L. Owens claims support from Aristotle's saying that to be is to be something or other. Hintikka, however, reminds us of several passsages that would seem to conflict with the ellipsis hypothesis, among them the following.

> For it is not the same thing not to be something and not to be simpliciter, though owing to the similarity of language to be something appears to differ only a little from to be, and not to be something from not to be. (*De Soph. El.* 167a4-6)

Having undermined the ellipsis hypothesis to account for the lack of the Frege-Russell ambiguity in Aristotle, Hintikka suggests that what distinguishes different uses of 'is' in Aristotle is not a difference in sense but merely a difference in *force* - predicative, existential, and identificatory, respectively.

So far as distinct uses are concerned, there is no difficulty in finding prima facie examples of those distinguished by Frege-Russell. Ostensibly, 'Homer is' would seem to be a clear cut existential use, 'Socrates is a man' and 'Socrates is pale' predicative uses, and 'A man is an animal' a use of generic implication. As Aristotle understands them, however, these uses are not nearly so unrelated as they may appear to us. So far as the existential use is concerned, the earlier quotation is pertinent, for Hintikka interprets it as showing that the existential and predicative occurrences of 'is' are indeed related to each other as absolute and relative uses of the same notion. Although being used with existential force in the former but predicative *force*in the latter, 'is' has the same *sense* in both.

Just as 'is' can sometimes be used with existential force and sometimes with predicative force, it can also be used sometimes with predicative force and sometimes with the force of identity. Among predications, Aristotle distinguished between essential and accidental, with 'Socrates is a man' being an example of the first and 'Socrates is pale' an example of the second. In regard to essential predication but not to accidental predication, however, Aristotle takes 'is' to express identity. The former is to be understood as saying that Socrates is identically what 'to be a man' signifies, whereas in the latter he is not identically what 'to be pale' signifies. Similarly, in 'a man is an animal', a man is taken to be identical with what 'to

be an animal' signifies. So, just as the existential and predicative uses are not unrelated, neither are the predicative, identity, and generic implication uses unrelated. Being related to each other, the Aristotelian uses do not correspond to the four uses distinguished by Frege, since these are presented as being totally unrelated to each other.

As a final point about 'is', we might ask whether Aristotle recognized it as having one or two existential uses. According to Owen, he nowhere distinguishes between the two that Geach has called the 'actuality' and the 'there-is' uses.

Turning now from Aristotle on 'is' to Aristotle on existence. His conclusion in the *Metaphysics* was that for any entity to be was for it to be what it is, i.e. what it essentially is. If Socrates is essentially a man, then for Socrates to be would be for him to be a man. So, the *immediate* explanation of the reality of Socrates would be in terms of his being man. One might then ask what would explain the reality of being a man, with the answer being that it would stem from being an animal, and so on. At each point, the explanation of Socrates' existence would be in terms of what it is essentially. The point would be reached, however, when the explanation would be in terms of the *category* to which he belongs, which is substance. But, this could not be the *ultimate* explanation of his reality, since the same question could be asked of substance.

Now, if being were itself a genus, then substance (and indeed all the other categories) might belong to it, and the ultimate explanation of Socrates' existence could be in terms of being (existence). This option, however, is not available to Aristotle, since he insisted that being is decidedly not a genus. So, although for Socrates to be is for him to be what he is essentially, we can pursue that lead as long as wish, but will never reach the point at which being (existence) is part of the essence of any genus - even the ultimate one - to which Socrates belongs. As Aristotle recognized, Socrates' reality would ultimately have to be demonstrated.

This, however, should not be taken to be a tacit recognition that existence is some kind of ontological element additional to Socrates, for Aristotle draws no distinction between the being (existence) of entities and the being (occurrence) of events or the being (obtaining) of facts, and no one would suggest that the obtaining of a fact was some kind of ontological element additional to the fact. Thus, Aristotle provides an excellent example of someone who, unlike Frege, recognized that 'is' could indeed be predicated of individuals, but did not feel bound to accept any ontological implications therefrom.

Avicenna (980-1037): With Averroes (1126-1196), Avicenna was one of the pre-eminent Arabian philosophers of the middle ages. I mention him because, unlike Aristotle, he was insistent on existence being an ontological constituent that was quite distinct from essence. Essences, he noted, can be present either in things or (intentionally) in intellect: in the former case they are engaged in the reality of things, in the latter they are conceived of by intellect. Considered in se, however, they are in neither; for if, in themselves, they were in things, they could never be in intellect, nor vice versa. Considered in themselves, therefore, they are in fact pure possibility which, in Avicenna's view, is far from saying that they are nothing at all. Rather, they can be said to have a certain kind of existence (*esse essentiae*), albeit one very much inferior to actual existence (*esse existentiae*). The notion of esse essentiae was to reappear

later in Henry of Ghent (d. 1293) among others. Although it might seem to be a highly fanciful notion, it is hardly more fanciful than some haecceity theories which employ the same distinction, nor perhaps than some possible worlds theories either.

Whereas the immediate explanation of the actuality of Aristotle's substances lay in what they were essentially, that was not the case with Avicenna's essences for their status was that of the merely possible. They entered the realm of the actual only if existence (*esse existentiae*) happened (*accidit*) to them. Here, therefore, he has introduced a new element into the ontological scheme of things, an innovation that was later to earn him the reproach of his compatriot Averroes. The new element is existence, which Avicenna regarded as an accident, a property of things. In his view, it was sharply to be distinguished from essence. Such views had been quite foreign to Aristotle's scheme of things.

Aquinas (1225-1274): It was one thing to recognize existence as an ontologcal constituent of any existent, as Avicenna did, but quite another to accept some of his ancillary notions. While agreeing with Avicenna that a substance was to be distinguished from its existence, Aquinas rejected his view of the relation between the two. In particular, he had no room for the notion of esse essentiae and hence no room either for a realm of the purely possible, as Avicenna conceived essences in se to be. That is to say, he did not think there was anything (such as a purely possible essence) to which existence might be said to accrue. For just that reason, existence could not be an accident of substance, since accidents do in fact accrue to something. And, if it were not an accident, it could not be related to substance in the same way as accidents are: it could not *inhere*in a substance. Having said this, however, Aquinas freely admitted that existence was indeed accidental to substance. This was not to contradict himself, since it meant merely that it was a contingent matter whether something existed or did not. It did, however, place the onus on Aquinas to explain just what the relation might be between an entity and its existence, if the latter were not an accident.

Moving now to the linguistic plane, it has to be said that Aquinas' views on the verb ‘is’ are scattered throughout his many works, but have fortunately been assembled and inter-related in an illuminating study by Hermann Weidemann. Like Aristotle, Aquinas had no difficulty in distinguishing the existential use of ‘is’ from the ‘is’ of identity and the ‘is’ of predication. Having said that, however, it has also to be said that he regarded the last two as closely related, with the predicative use being marked by an element of identity, and the identity use being marked by an element of predication. The predication ‘Socrates is pale’, for example, was understood to be proposing the identity of Socrates with something pale, and ‘Socrates = Socrates’ was understood to be proposing the inherence in Socrates of the property of being Socrates. Unlike the Frege-Russell thesis, Aquinas did not regard these three uses of ‘is’ as being totally unrelated.

A crucial difference between Aquinas and Avicenna, however, lies in the distinction he draws between two existential uses of ‘is’. In one of them, ‘is’ is taken to express the being of whatever falls under the Aristotelian categories, whether it be of substance or of any of the accidents. As used in this way, ‘is’ refers to that by which something is actual. In the second existential sense, however, it expresses the truth of a proposition. Following Geach, these two existential uses might be called the ‘actuality’ and the ‘there-is’ uses respectively. Interestingly, the actuality use is said to occur not only in such propositions as

'Socrates exists' but , surprisingly to modern ears, in such propositions as 'Socrates is a man'. Indeed, it occurs in any of those predicates that respond to the question '*Quid est .....?*' ('What is ....?'), and which Aquinas calls substantial. It is in this use that 'is' is taken to express the being of whatever falls under the Aristotelian categories.

In its there-is use, 'is' is said to express the truth of a proposition, and to answer the question '*An est...?*' ('Is there any such thing as ....?'), and which Aquinas calls accidental. In these cases 'is' has the dual function not only of linking subject and predicate, but also of expressing the truth claim that is being made thereby.

As for the ambiguity of 'is', Aquinas' position would seem to be:

1. There is systematic ambiguity between the uses of 'is' to express the actuality of what falls under any of the ten categories.
2. There is no ambiguity at all between the actuality use of 'is' and its use in substantial predication: they are the same.
3. There is no ambiguity at all between the there-is use of 'is' and its use in accidental predication: they are the same.
4. There is systematic ambiguity between the actuality and the there-is senses of 'is', and this is founded on the supposition that the truth of what we say is founded upon the actual existence of what we talk about.

Hume (1711-1776): Aquinas' distinction between essence and existence was not long unchallenged even among the Scholastics, being rejected early by Scotus and much later by Suarez. Descartes and Leibniz also denied it, and Hume took the same view, though for reasons peculiar to his own impression-based epistemology. Thus, he argued that 'the idea of existence must either be derived from a distinct impression, conjoined with every perception or object of our thought, or must be the very same with the idea of the perception or object'. (*Treatise of Human Nature,* Bk.I, Part II, sect. vi) There being nothing to indicate the presence of any impression at all that is 'conjoined with every perception or object of our thought', he concludes that there is no distinct impression from which the idea of existence is derived. Rather, it is 'the very same with the idea of what we conceive to be existent'. Any one dissenting from this, suggests Hume, has the task of indicating just what is the distinct idea from which the idea of existence derives.

Hume's contention that the idea of existence 'makes no addition' to the idea of any object was to be reaffirmed in Kant

Kant (1724-1804): The following familiar passage clearly indicates how closely Kant is aligned with Hume's conclusion about existence.

> By whatever and by however many predicates we may think a thing - even if we completely determine it - we do not make the least addition to the thing when we further

> declare that this thing is. ... If we think in a thing every feature of reality except one, the missing reality is not added by my saying that this defective thing exists. (Critique of Pure Reason, B628)

Earlier he had reminded us that ‘the real contains no more than the merely possible. A hundred real thalers do not contain the least coin more than a hundred possible thalers.’ Thus, he is able to claim that ‘"being" is not a real predicate’, though he does allow that it is a logical predicate, for ‘anything we please can be made to serve as a logical predicate’.

That being so, just what role does ‘is’ play in such propositions as ‘God is’ or ‘God is omnipotent’? In both, says Kant, its role is simply to posit (*setzen*) the subject. In the former, it posits the subject (God) ‘in itself with all its predicates’; in the latter it posits the predicate in relation to the subject (God). The ‘is’ in the former and the ‘is’ in the latter seem to be merely two uses of the one notion - in some respect reminiscent of the ‘ellipsis’ interpretation of Aristotle's uses of ‘is’. In any case, this position is clearly contrary to the ambiguity of ‘is’ that was later to be espoused by Frege.

Frege (1848-1925): Frege regarded existence as a second-level concept. In explanation of this, it has to be remembered that, in his terminology, concepts are not intentional entities, as one might have surmised, but are no less ontological items than are objects. Just as objects are the referents of singular terms, so are concepts the referents of predicates, which might also be called ‘concept expressions’. Some predicates are attached to singular terms to form a proposition and hence say something about the objects to which those terms refer, e.g. in ‘Socrates wise’. These are called first-level predicates or concept expressions, and refer to first-level concepts, e.g. being wise. Other predicates are attached to first-level predicates to form a proposition and hence say something about the concept to which the first-level predicate refers. They are second-level predicates, and refer to second-level concepts. A case in point would be ‘Wisdom (or being wise) is rare’, in which the predicate ‘is rare’ would refer to a second-level concept, being rare. There can of course be still higher-level predicates with their correlative concepts to which they refer.

In saying that existence was a second-level concept, Frege was also denying not only that it was a first-level concept but also that ‘exists’ was a first-level predicate, contrary to what might appear to be the case in propositions like ‘Socrates exists’. Interestingly, one of his reasons for opposing a first-level use was similar to the paradox that some contemporaries insist is generated by negative existential propositions like ‘Socrates does not exist’. The putative paradox is said to arise because ‘does not exist’ could be said of Socrates only if he did in fact exist. In his ‘Dialog mit Puenjer ueber Existenz’ Frege argued that, in the proposition ‘Leo Sachse is’, nothing is being attributed to Leo Sachse. Puenjer had proposed that the ‘is’ be understood as elliptical for ‘is something that can be experienced’, to which Frege replied that if ‘*A* is not’ were to be understood as ‘*A* is not an object of experience’, it would be self-contradictory. Why? Because on the one hand *A* certainly is an object of experience, whereas on the other hand *A* is being said not to be an object of experience: it both is and is not an object of experience. Since the same point could be made no matter what was substituted for ‘is’ in ‘Leo Sachse is’, the role of ‘is’ in that proposition could not be that of a first-level predicate.

Given that ‘is’ or ‘exists’ are precluded from being first-level predicates, just how are affirmations of

existence to be understood? According to Frege, ‘Affirmation of existence is in fact nothing but the denial of the number nought’ (*Die Grundlagen der Arithmetik*). This means, as he says elsewhere, that existence is ‘a property of a concept’: it is a second-level property (or concept). As for the predicate ‘exists’, it is therefore understood to be attached to a first-level predicate, and hence is itself not a first-level predicate but a second-level one. More concretely, ‘Leo Sachse is’ should, in Frege's view, be rendered as ‘(∃x)(x = Leo Sachse)’ or, in non-symbolic terms, as ‘There is at least one thing that is identical with Leo Sachse’. Thus understood, it is clear that nothing is being predicated of Leo Sachse. Rather, something is being said about the property (or concept) of being identical with Leo Sachse: we are being told how often that particular property is instantiated, namely, at least once. And, of course, ‘at least once’ is ‘the denial of nought’.

Issues Raised by the Historical Survey

The foregoing brief and selective historical survey enables us to distinguish a number of questions not only in regard to the verbs ‘is’ or ‘exists’ but, on the ontological level, in regard to existence. In regard to ‘is’, it has been generally agreed that it admits of various uses. Apart from the ‘ellipsis’ interpretation of Aristotle which admits of only one use, the differences have not been about whether there are many uses of ‘is’, but about just how many there are. The main difference, however, has been about senses rather than uses: is the one sense being used with different forces, or are there as many senses as there are uses? This is not an idle question, for it can never be assumed that different senses correspond to different uses. Moreover, even if the senses do differ, it cannot be assumed that they will be casually ambiguous, for they might be merely systematically ambiguous, as Aristotle conceives some of the predicative uses of ‘is’ to be.

Have we to choose between the Frege-Russell treatment of ‘is’, the Aristotelian treatment, and perhaps some other? According to Hintikka, we can both have our cake and eat it since, as he argues, there can be more than one correct way of dealing with ‘is’. He reminds us that the Frege-Russell ambiguity need have no part in game-theoretical semantics. He even argues that we can say not only ‘that Aristotle's procedure is free from any taint of fallacy; he may even have been a better semanticist of natural languages than Frege and Russell ever were.’ (‘The Varieties of Being in Aristotle’)

Not all questions about the use of ‘is’ can be resolved so eirenically, and the question of whether there is a first-level use is one of them. Although Hume and Kant accept predicative uses, it is not clear that they would accept existential ones. The three A's - Aristotle, Avicenna, and Aquinas - would seem to accept both. Frege, too, would recognize both. But, whereas Aquinas recognizes two existential uses, Frege accepts only the there-is use. For Frege, the existential use of ‘is’ or ‘exists‘ is invariably as a second-level predicate. For Aquinas, there is a first-level use in addition to the second-level one. Geach, too, argues for a first-level use of ‘exists’ as do Parsons and Zalta, though on different grounds.

If there were a straightforward a priori way of showing that Frege was right, we should be absolved from having to consider attempts to undermine his view. Unfortunately, things are not so simple for, although a case can be made out for two existential uses, its cogency is still much disputed. The case turns on the

possibility of showing not only that ‘is’ or ‘exists’ can sometimes be predicated of individuals but also why the there-is sense cannot substitute for it.

Where the discussion would go from here is something that depends on whether the case for two uses is thought to prevail. If it does, then a number of additional issues would have to be addressed, for if ‘is’ or ‘exists’ were a first-level predicate then, like all predicates, it would have a reference which in this case would be existence. And, if we describe a property in the wide sense as whatever would be attributable to something by a predicate, then existence would be a property of individuals, thereby undermining the widely held dictum ‘Existence is not a predicate’.

Now, if existence were a property of individuals, then it would seem legitimate to join Hume and Kant in asking just what it adds to individuals. Since their question is posed merely rhetorically, it would presumably be satisfied by existence being simply a Cambridge property, for the distinguishing mark of Cambridge properties is precisely that they add nothing to individuals. The immediate task confronting proponents of the actuality use of ‘is’ or ‘exists’ would be to determine whether the existence to which they were committed was in fact a real property or a Cambridge one.

If the contention were that it was real, the Hume-Kant question would pose a pressing problem, albeit not necessarily a decisive one. It would be decisive only if existence were an accident of individuals, only if its relation to individuals were what Aristotle would call one of inherence. According to the notion of a property which I introduced earlier, a property is whatever is attributed to something by a predicate. This notion, however, is neutral as to the kind of relation a property would have to the subject to which it was attributed. In particular, it does not imply that what is attributed should be an accident of the subject. Moreover, even in the Aristotelian ontology, not everything that was attributed to a subject had to be an accident. Hence, if defenders of existence as a first-level property wished to rebut the Hume-Kant objection, they would have to establish that, contrary to Avicenna, existence was not an accident of individuals. If that proved impossible, they might be forced to surrender their claims, and accept that Frege was right after all about the single existential use of ‘is’.

Arguments for the Fregean View of ‘Exists’

In the heyday of logical positivism Alfred Ayer confidently asserted that, if ‘exists’ were a predicate or existence a property, ‘it would follow that all existential propositions were tautologies, and all negative existential propositions were self-contradictory’. No need to distinguish between singular and general existential propositions, nor to make any distinction between kinds of general existential propositions. No matter where it occurred, ‘exists’ was not a predicate.

Few would take that view today. Many would hold that ‘exists’ is indeed a predicate, but would hasten to add some qualification - either that it was ‘peculiar’, had ‘special characteristics’, was ‘redundant’, was a second-level predicate, of that it was a metalinguistic predicate. The dominant view is that ‘exists’ is a second-level predicate, and it is grounded largely on the contention that to admit ‘exists’ as a property of individuals would lead on the one hand to the absurdity of regarding existence as a property, and on the

other hand to the paradox supposedly generated by negative existential propositions.

As to the first, presumably it is thought that if existence were a property, then non-existence would be one also. And that would seem to give rise to the ludicrous situation described by David Londey who invited us to ‘reflect on the absurdity of a farmer who daily inspected his flock with the aim of sorting the existing from the non-existent ones - searching for the stigmata of existence’. It would give rise also to C.J.F. Williams' enquiry as to whether, if told that blue buttercups did not exist, he would ‘have felt obliged to examine several specimens of blue buttercup before concluding that none of them exist, that as a variety blue buttercup lacks existence’. Hume and Kant would have nodded approvingly.

As to the paradox generated by negative existential propositions, it arises in this way. If ‘exists’ were a predicate, then its negation (‘does not exist’) should be a predicate also. But if ‘does not exist’ were a predicate, then in ‘Dragons do not exist’ it would be predicated of dragons only if dragons existed. And similarly for all negative existential propositions; paradoxically, if it is to be predicated at all, ‘does (do) not exist’ can be predicated only of what does exist.

Nor do singular propositions like ‘Socrates does not exist’ seem to fare any better than general ones like ‘Dragons do not exist’. Despite his demise, there are of course many true predications that can still be made of Socrates, e.g. that he was a philosopher, that he was Greek, and the like. But all of these were true of him while he was yet alive. If, however, ‘does not exist’ were true of him, it could be true, paradoxically, only after there was any Socrates for it to be true of. Again, as a variant formulation of the paradox, it might be argued that if ‘exists’ were a predicate of individuals, it would be true of everything. In that case, ‘does not exist’ would be true of nothing: neither ‘Socrates does not exist’ nor any other singular negative existential proposition could be true. In that case, however, ‘does not exist’ could not be a predicate. But, if ‘does not exist’ could not be a predicate, its negation (‘exists’) could not be a predicate. The assumption that ‘exists’ *is* a first-level predicate has therefore led to the paradoxical result that ‘exists’ *cannot* be a predicate. So, singular negative existential propositions are no less paradoxical than are general ones.

It is argued that the paradox could be avoided by treating ‘exists’ as something other than a first-level predicate. In doing so we should also remove perhaps the strongest ground for regarding existence as a property. And thus in one move we should avoid putative paradox and absurdity alike.

We have already seen how Frege succeeded in doing exactly that. Russell did the same with his similar view of existence as a property not of things but of propositional functions, a proposal made in connection with his treatment of definite descriptions. As such, its immediate application is only to general existential propositions, but not to propositions like ‘Socrates exists’. Since Russell regarded ‘Socrates’ as merely a disguised description, he himself would have rejected this restriction. But for those who do not share his view on proper names, the restriction could be overcome by adopting Quine's proposal that ‘Socrates exists’ be reparsed as ‘$(\exists x)(x = \text{Socrates})$’. Quine takes the further step of eliminating the predicable ‘= Socrates’ in favour of one that contains no proper name, viz. the predicable ‘socratizes’. Thus, ‘Socrates exists’ would be understood as ‘$(\exists x)(x$ socratizes$)$’, or ‘The property of socratizing is instantiated at least once’. If Quine is correct, then we have a means of handling existential

propositions that treats them neither as tautologies nor as contradictions, yet without the difficulties that would arise if ‘exists’ were a predicate. Further advantages are claimed for it. Basically, singular existential propositions would be treated in the same way as general ones. Moreover, there would be no need for multiple senses of ‘exists’. Indeed, ‘exists’ itself would be made redundant, being replaceable by the more general apparatus of quantifiers and identity. And, finally, there would be no need to recognize, as the early Russell did, entities that subsist in addition to those which exist. Thus, the theory seems not only to solve the problem, but to do so with an economy that enhances its appeal.

Criticisms of the Fregean View

Questions have been raised not only about the validity of the paradoxes and absurdities allegedly generated by accepting ‘exists’ as a first-level predicate, but also about the Quinean treatment of propositions like ‘Socrates exists’. As for the paradoxes and absurdities, it might be argued that they stem not from allowing existence to be a property, but from allowing non-existence to be one. Only by thinking that non-existence was some kind of real property would any sheep-farmer be led to the absurdity of inspecting his flock ‘with the aim of sorting the existing sheep from the non-existent ones’ or would anyone feel inclined to ‘examine several specimens of blue buttercup before concluding that none of them exist’. Only if non-existence were a real property (rather than merely a Cambridge one) would it seem paradoxical that ‘does not exist’ could be true of Socrates only after there was any Socrates for it to be true of.

It might seem strange, therefore, that the blame has been laid on treating existence as a real property of individuals, when it should surely have been laid on treating non-existence as one. Why deny that existence is a property, when it was necessary only to deny that non-existence was one? Perhaps the answer lies in the understandable but mistaken belief that the two denials are inseparable, and so there could be no denying non-existence to be a real property of individuals without denying existence to be one also. After all, if properties are what predicates stand for, how could it be said that ‘exists’ stood for a property, but that ‘does not exist’ did not? If we accept existence as a property, are we not bound also to accept non-existence as one? Clearly, these suggestions rest on two assumptions that need to be tested:

1. that ‘Socrates does not exist’ contains a negative existential predicate,
2. that a negative existential predicate stands for a real property.

In regard to (1), although ‘does not exist’ is a grammatical predicate in ‘Socrates does not exist’, it does not follow that it must also be a logical one. We need to recognize the possibility of construing the proposition as having the logical form of ‘It is not the case that (Socrates exists)’. In that case, what is predicated (though not asserted) of Socrates would be simply ‘exists’ (and not ‘does not exist’); and what would be asserted is that it is not the case that Socrates exists. On such an analysis of singular negative existential propositions, ‘does not exist’ would not be a predicate, and nor therefore need non-existence be a property of any kind.

The distinction being employed above is one between internal or predicate negation on the one hand and

external or propositional negation on the other. In both cases something is said about an individual, namely, Socrates; the former says that non-existence is had by Socrates, and the latter denies that existence is had by Socrates. It is true that first-order predicate logic is so constructed as to admit no such distinction, but that does not mean that there is no such distinction tout court. Consider the proposition '*a* is not moral' which may mean either of two things. It may mean that *a* has the 'property' of being non-moral; alternatively, it may simply be denying that *a* has the 'property' of being moral. Internal negation ('*a* (is not moral)') is being used in the first case, but external negation ('It is not the case that (*a* is moral)') in the second. If, therefore, the distinction between internal and external negation were one without a difference, those two renderings should mean the same. Yet, that is just what they do not mean; for the first is to be taken as '*a* is immoral', but the second as the quite different '*a* is either immoral or amoral'. The distinction cannot therefore be dismissed as a 'distinction without a difference'.

It is therefore not a matter of indifference whether 'Socrates does not exist' is rendered as '(Socrates) does not exist' or as 'It is not the case that (Socrates exists)'. Because it is the former but not the latter that gives rise to problems, the latter is clearly to be chosen. So, it remains true that 'Socrates does not exist' would not contain 'does not exist' as a logical predicate, and that existence could be recognized as a real property without the embarrassment of having to recognize non-existence as well.

It is worth noting that the paradox could be argued not to arise even if the negation in 'Socrates does not exist' were to be intemal rather than external. First, however, we need some criterion for deciding when individual *a* could lack some property *F* only by having another property non-*F* correlative to the one it lacks. Well, let us consider this not in regard to existence but in regard to the property red. The question therefore is whether the absence of redness from something which could be red must bespeak the presence of a property correlative to red. Certainly, if *a* were a piece of wood then it could lack redness only if it had some colour or colours other than red - be it brown, cream, fawn, or whatever - all of which are properties. That does not settle the question, however, since the result would be very different if *a* were not a piece of wood but a piece of glass.

Now, although glass is like wood in being something that could be red, it is also unlike wood in that its failure to be red does not mean that it is any colour at all: it may be quite colourless. To say that it is non-red, therefore, is not to say that it has any correlative property. Reflecting on this example, it is not difficult to see that lack of a property *F* bespeaks the presence of a correlative property non-*F* only if *F* and non-*F* are understood as determinates of a common determinable. Thus, if red and non-red were related as differentiae of the determinable property being coloured, red could be lacking in an *a* that was coloured only if some determinate of colour were present in *a*

The relevance to the discussion of non-existence is fairly clear. If lack of existence in *a* (which had existed) were to bespeak the presence in *a* of non-existence as a real property rather than as merely a Cambridge one, existence and non-existence should be related to some real property just as red and non-red would have had to be related to the property of being coloured. For convenience, let us call this determinable property '*E*'. Then, just as red and non-red would have had to be understood as coloured red and coloured other than red, so existence and non-existence would have to be understood as being *E* in an existential way and being *E* in a non-existential way.

Thus, whether *a* existed or did not exist, it would have some form of being: it would be *E*. But that is false. Hence, even if ‘Socrates does not exist’ were to contain the predicate ‘does not exist’, the property stood for by that predicate could at most be a Cambridge one; and Cambridge properties can be acquired even by individuals that do not exist at the time. Consequently, no paradox would arise from ‘Socrates does not exist’. This has already been demonstrated for the case where the negation was taken to be external to the proposition; it has now been demonstrated for internal negation as well.

To sum up. It has been argued that, no matter whether the distinction between internal and external negation in this context were accepted or rejected, the result would be the same. In neither case would we be committed to Socrates acquiring any property whose acquisition is conditional upon Socrates existing at that time. In neither case, therefore, would ‘Socrates does not exist’ generate the paradoxes or absurdities which would make it impossible to count ‘exists’ as a first-level predicate and existence as a real property of individuals.

Turning now to the questioning of Quine's construal of ‘Socrates exists’ not as ‘Socrates has (the property) existence’ but as ‘The property of socratizing is instantiated at least once’. This has the appearance of dispensing with the use of proper names and dispensing also with the recognition of what proper names typically refer to, viz. individuals. It can be argued, however, that this is simply an illusion, because the notion of instantiation in fact makes no sense except in relation to what is instantiated (a property) and what it is instantiated *in* (an individual). Or, more technically, instantiation is akin to a second-level function that Frege called the ‘application’ of a function to an object: a second-level function is inconceivable except in terms of what are stood for by the expressions filling its two gaps. So, individuals cannot be eliminated by introducing the notion of instantiation, for that very notion is itself intelligible only in terms of individuals. On this view, therefore, ‘Socrates exists’ is logically more basic than the proposition ‘$(\exists x)(x$ socratizes)’ that has been advanced to dispense with it. (It might be replied that this objection would fail to impinge on Quine's position, since he eschews properties in favour of classes. On the contrary, the classes which are substituted for properties are no more intelligible without the notion of an individual than are the properties they are supposed to replace.)

Arguments for the Two-Sense Use of ‘Exists’

To accept the Fregean view of ‘exists’ as a second-level predicate is to accept that ‘exists’ can in fact always be rendered by ‘instantiates’. Of course, there are many cases where it can indeed be understood in just that way. However, although ‘Elephants exist’ can be understood as ‘The property of being an elephant (or the species elephant) is instantiated at least once’, there are grave difficulties about regarding ‘Socrates exists’ as ‘Socrates is instantiated at least once’. The problem is that individuals are just not the kind of thing that ever could be instantiated. Rather than being themselves instantiable they are the kind of thing in which instantiations occur, e.g. wisdom is instantiated in Socrates, but Socrates himself cannot be instantiated in anything. Russell and Quine would certainly have recognized this, and each in his own way attempted to get round it, though with questionable success.

The two-sense theory, in effect, accepts that there is no way round it. Thus, while not denying that ‘exists’ does have a second-level use, it insists on there being a first-level use as well, one in which ‘exists’ is not to be understood in terms of ‘instantiates’. The theory's claim for different uses of ‘exists’ in ‘Socrates exists’ and ‘Elephants exist’ is in some respects like different uses for ‘disappear’ in ‘Jack the Ripper disappeared’ and ‘Dodos disappeared’.

Now, there are at least two ways of arguing for the theory, the first of which is explained briefly by Hintikka, as quoted below.

> If we take an individual in the actual world and assign to it a predicate which involves existence or nonexistence in some other world, surely we ought ... be able to take "a merely possible individual", i.e. a denizen of some other world, and attribute to it predicates definable in terms of its actual existence, maybe the "predicate of (actual) existence itself". Basically, it seems to me, that this argument is unanswerable. (‘Kant on Existence and Predication’, p.255)

His view of possible individuals is reminiscent of Avicenna's claim that possible essences do indeed have a certain kind of being, namely, esse existentiae. And, being aware that his argument might be criticized for supposing that there could be individuals that have never existed - merely possible individuals -, Hintikka attempts to forestall it by contending that it is ‘based on an unrealistically narrow view of how our language actually functions’. That might be a compelling consideration if the criticism concerned the use of proper names for these individuals, for in such circumstances the names would be no more exceptional than are names for fictional individuals. The objection, however, is not to the use of names, but to the apparent supposition that merely possible individuals have as much claim to be accepted as individuals as do actual ones. In other words, actual and merely possible individuals would be two kinds of individual, no less than Labradors and Beagles are two kinds of dog. On the contrary, merely possible individuals (and fictional individuals as well) have no more claim to be regarded as a kind of individual than rocking horses have to be regarded as a kind of horse.

A more promising argument is one that draws upon an insight of Peter Geach, and would run as follows:

- What can be predicated of a kind differs absolutely from what can be predicated of an individual.
- But ‘exists’ is predicated both of individuals and of kinds.
- Therefore, ‘exists’ has two senses, one as predicated of individuals, the other as predicated of kinds.

In regard to the first premiss, although there are two ways in which a predicate might be conceived of as being applicable to both kinds and individuals, it is not difficult to show that neither is tenable. One way would be for a second-level predicate to be said of both individuals and kinds, the other for a first-level predicate to be said of them. In regard to the first alternative, we must be clear as to precisely what kind of expression can be said of what the first-level predicate refers to (e.g. a kind, or Fregean concept). If we consider the proposition ‘$(\exists x)(x \text{ is } F)$’, the first level predicable is ‘......is *F*.’ The second-level predicate

attached to it is, however, not simply '(∃*x*)', but '(∃*x*)(*x*......)'. If we now ask whether the second-level predicate could equally well be attached not only to a first-level predicable but to a proper name, it is clear that it could not. The bound variable, which filled the gap in '......is *F*', has nowhere to go when '......is *F*' is replaced by a proper name. The expression that results from such a combination is therefore not even a closed sentence. Nor does anything better come of the second alternative mentioned above. If '......is *F*' and '......is *G*' are two first-level predicables, then the result of combining them would be '(......is *G*) is *F*' or '(......is *F*) is *G*'. Once again, neither combination would be even a closed sentence, as the gap would be filled neither by a bound variable nor by a proper name. Thus, it follows that no predicate, whether of first-level or of second-level, could be said both of individuals and of kinds. And that would establish the first premiss.

The second premiss can be argued for in two ways - either by contrasting singular existential propositions with one kind of general existential proposition, or by contrasting two kinds of general existential proposition. As an example of a singular proposition in which 'exists' is predicated of an individual, one might be tempted to suggest 'Socrates exists', were it not for oft-voiced protests of its not being 'usable outside philosophy'. Rather than resist that claim, therefore, it would be better to employ an example that unquestionably is usable outside philosophy, viz. 'Socrates no longer exists'. This should be understood as 'It is no longer the case that (Socrates exists)', which clearly involves use of the proposition 'Socrates exists'. T? ?e?ning now to the second-level uses of 'exists', they are both numerous and non-controversial. 'Men exist' is a case in point, for it may often be rendered as '(∃*x*)(*x* is a man)', thus showing it not to be about any individual but, rather, about the property of being a man; for it says that being a man is instantiated at least once. Hence, 'Socrates no longer exists' and 'Men exist' provide the evidence necessary for the premiss that 'exists' is predicable both of individuals and of kinds; for the only way of eliminating the difference between them is to reparse 'Socrates' as a predicable after the manner of Quine, and that has already been shown to be unacceptable.

There is, however, a second way of proving the minor premiss, and this even without recourse to any singular existential propositions. It can be done by showing that not even all general existential propositions are about kinds, but that some are about individuals. As an example of the two kinds of general existential proposition, consider the occurrences of 'elephants exist' below.

a. 'Elephants exist, but mermaids do not.'
b. 'Elephants exist, but dinosaurs do not.'

As will now be shown, 'exists' in (a) is being said of the property of being an elephant, not about individual elephants, whereas in (b) it is being said of individual elephants, not about the property of being an elephant. These claims can readily be substantiated. Since, in (a), 'Elephants exist' is being contrasted with 'Mermaids do not', the sense in which 'elephants' is being used will be the same as that in which 'mermaids' is being used. Now, 'Mermaids do not exist' makes sense only if it means that all predications of the form '*x* is a mermaid' are false. And it cannot mean that any proper name which turns '*x* is a mermaid' into a true statement will turn '*x* does not exist' into a true one, the simple reason being that there are no non-fictional proper names available for substitution in '*x* is a mermaid'. Hence, 'mermaids' is being used to refer to the property of being a mermaid, not to individual mermaids. Thus,

‘elephants’ in the accompanying clause must refer to the property of being an elephant. One might try to escape that conclusion by suggesting that a fictional name might well be substituted for ‘*x*’, as of course it might. That, however, would do nothing to alter our conclusion since fictional individuals are not concrete individuals any more than rocking horses are horses. So, there are no grounds for saying that non-fictional proper names can be substituted in ‘*x* is a mermaid’, and so no grounds for saying that ‘mermaids do not exist’ can be equally about kinds or about concrete individuals.

In (b), on the contrary, neither do ‘elephants’ refer to the property of being an elephant nor ‘dinosaurs’ to the property of being a dinosaur. If they did, the proposition would not only be false, but the ‘but’ would be quite misleading since there would be no point of contrast between the first and second clauses. The only way to retain that contrast is for ‘elephants’ and ‘dinosaurs’ to refer to individuals. So, in (b), ‘Elephants exist’ is a general existential proposition that is about individuals, as contrasted with the same clause in (a) which is not about individual elephants at all but merely about the property of being an elephant. It is of some interest to have noted that the first-level use of ‘exists’ is not restricted to singular propositions, as might have been supposed, but can occur in (some) general propositions as well.

The second premiss - that ‘exists’ is predicable of both kinds and of individuals - has therefore been supported in more than one way way. From it and the major it would follow that ‘exists’ has two senses, one as predicable of individuals, the other as predicable of kinds, and which have been called by Geach the actuality and there-is senses respectively.

Arguments for Distinguishing ‘Exists’ from ‘Is’:

The two-sense theory has been concerned with the the role of ‘exists’ as a predicate of concrete individuals among which it has certainly not included numbers, universal properties, propositions, and such intentional entities as Pegasus, Sherlock Holmes, unicorns, round squares, and golden mountains. The inclusion of precisely those entities has, however, been a feature of the work of those like Terence Parsons and Edward N. Zalta who draw some of their inspiration from Meinong. Their common aim is to provide a logic - an intensional logic - which will explain the apparent failure of one or other principles of non-intensional logic in certain contexts. Expressed otherwise, their goal is to ‘explain how commonsense non-existence claims of natural language mean what they seem to mean and have the truth-values, logical form, and entailments that they seem to have’. (Zalta, p.103) For our purposes, these views are notable for introducing a further consideration for recognizing the use of ‘exists’ as a first-level predicate. As we shall see, however, the price of achieving a logic that is more faithful to the ways in which we ordinarily talk is to embrace ontologies that are rather more distant from the ways in which we ordinarily think.

Parsons' system has three basic features:

- *The distinction between nuclear and extranuclear predicates and likewise between their corresponding properties.* As examples of nuclear predicates we are offered ‘is blue’, ‘is tall’, ‘kicked Socrates’, ‘was kicked by Socrates’, ‘kicked something’, ‘is golden’, ‘is a mountain’.

Examples of extranuclear predicates include ‘exists’, ‘is mythical’, ‘is fictional’, ‘is possible’, ‘is impossible’, ‘is thought about by Meinong’, ‘is worshipped by someone’, ‘is complete’. As critics have noted, the distinction is far from transparent.

- *The correlation between non-empty sets of nuclear properties and genuine objects.* Objects are correlated with (but not identified with) sets of nuclear properties - real objects with complete sets, and non-existent sets with incomplete sets. A set is complete if, for any nuclear property, an object has that property or its negation. Otherwise, a set is incomplete. Critics consider this distinction, too, to be less than obvious. Moreover, Parson's stipulation that no two objects have exactly the same nuclear properties might suggest that he adheres to the suspect Identity of Indiscernibles.
- *The distinction between objects that **are** and objects that **exist**.* The predicate ‘exists’ is said to be applicable to ‘all the ordinary [concrete] objects that we normally take to exist’. (p.11) Parsons recognizes objects that would not normally be regarded as concrete, namely, golden mountains, winged horses, round squares, Pegasus, and Sherlock Holmes. These objects are said not to exist, but merely *to be*. Thus, of objects like tables we can say both ‘Tables exist’ and ‘There exist tables’. Of other objects we can say merely that there *are* such things, e.g., ‘There are unicorns’. Such objects are said to be *non-existent*. In virtue of the distinction between existing and merely being it is then possible to say without any kind of contradiction, or even paradox, ‘There are unicorns, but they do not exist.’ This makes perfectly good sense because, as Parsons uses the terms ‘is’ and ‘exist’, what exists does not exhaust what there is. (p.5)

Obviously, this is directly at odds with the Frege-Russell-Quine one-sense theory, in which what exists is precisely what there is. Less obviously, perhaps, and despite its distinction between what exists and what there is, it sits uneasily with the two-sense theory as well, for that theory presupposes an absolute distinction between individuals and properties, a distinction which is blurred by Parsons. Even if we allow that his *existing* objects are not mere bundles or sets of properties, it is not obvious how the same can be said of his non-existent objects. He needs therefore to tell us in what sense the golden mountain is anything more than a set of properties, namely, of being golden, of being a mountain, and of being existent. If it is nothing but such a set, it is no more concrete than are properties; and these he has classed as abstract. Contrary to his declared aim therefore, his theory would seem not to be one about concrete objects after all. But, if it is not about concrete objects, ‘is existent’ cannot be a first-level predicate - unlike ‘is golden’ and ‘is a mountain’. But, then, what becomes of the claim that existence and being existent are two kinds of being? They can hardly be two kinds of being if the former is a first-level property while the latter is merely a second-level one.

Zalta, comparing his own theory with that of Parsons (p.134), draws attention to its ‘explanatory elegance’ in replying to Russell's three objections against Meinong. Whereas each of Parsons' replies employs a different strategy, each of Zalta's replies employs just the same two hypotheses, viz:

1. That ‘is’ can also be read as ‘encodes’.
2. That descriptions of the form ‘The $G_1, \ldots, G_n$,’ when said in a Meinongian way, have readings in which they denote A-objects that encode the properties $G_1, \ldots, G_n$.

The notions of A-objects (abstract objects) and of encoding are central to his theory, and need now to be explained.

One of the problems with intentional objects is that they themselves do not seem to have the properties that they represent - the object which represents blue is not itself blue. Although this was a point to which Brentano adverted in the last century, its history extends as far as the middle ages and beyond to Aristotle. The mediaevals resolved their difficulty by suggesting that the identical form may have two modes of existence, physical and intentional. For example, the identical form or forms had materially (physically) by Bucephalus would be had immaterially (intentionally) by the person perceiving him. All the forms (substantial and accidental alike) existing physically in Bucephalus would exist intentionally in whoever perceived him.

To deal with the same problem, Zalta distinguishes not between physical and intentional *existence* but between physical and intentional *objects*. Physical objects are said to be concrete and to *exemplify* various properties. Intentional objects, on the contrary, are abstract, a term which Zalta seems to regard as synonymous with ‘non-spatiotemporal’. These A-objects, as he calls them, fail to exemplify those properties at all but do, however, *encode* them. As he notes, the distinction between exemplifying and encoding properties is due to Ernst Mally. Abstract objects have content not by exemplifying properties but only by encoding them. Abstract objects that encode properties serve two purposes (p.15):

- They serve as intentional objects of states directed towards nonexistents.
- They also serve to characterize and reify the content of mental representations.

Zalta's reasoning for A-objects would seem to be:

- There must be some way or other in which the golden mountain *has* the properties of being golden and being a mountain.
- But, in the case of the golden mountain, ‘having those properties’ cannot be understood as ‘exemplifying those properties’.
- Therefore, there must be some way of having properties which neither is nor entails exemplifying them.
- Call this other way ‘encoding’. And call the object which encodes properties an *abstract* one. Let concrete objects be said to *exist*, and abstract objects be said not to exist, but merely to *be*.

Although it is clear what encoding is meant to achieve, it is not quite so clear what it actually is - no more clear, it would seem, than the mediaeval notion of intentional existence, nor any more clear than Parsons' nuclear/extranuclear distinction.

Why draw the ‘exists’/‘is’ distinction? Because it seems intuitively clear that ‘Pegasus’, ‘Zeus’, and ‘Hamlet’ are names of nonexistent, mythical, and fictional creatures. After all, ‘the logic of natural language seems to presuppose that it makes sense to refer to and talk about these creatures’ (p.103) Indeed, if ‘Quine were right when he says that "Hamlet doesn't exist" means "$\sim(\exists x)(x = h)$", then

"Hamlet" would fail to denote', and that would make quite mysterious the natural inference from 'John's paper is about Hamlet' to 'John's paper is about something' (p.104) Consequently, there is indeed a case for not dismissing Pegasus and others as complete nonentities, but for admitting that they do have some kind of entity. One might even be tempted to say that they exist, except that this could be taken to imply that they were no less real than Aristotle, Plato and Julius Caesar, which few would be prepared to accept. To do justice to these considerations, therefore, we are urged to allow that Pegasus and Zeus truly *are* (or have being), but not to go so far as to accept that they *exist* (or have existence). For Zalta, 'is' or 'being' is represented by '∃' and is called 'logical or metaphysical existence', while existing is called 'physical existence' and 'exists' is represented by 'E!'. It is with the aid of this distinction between existing and being that Zalta develops an intensional logic which handles intentional objects with a facility not to be found in non-intensional logics.

The relative positions of Parsons and Zalta on 'exists' and 'is' can be summarized as follows:

1. *Parsons*: Distinguishes 'exists' from 'is', existence from being. Both 'exists' and 'is' are first-level predicates, since each is said of *concrete* objects, the former of complete objects and the latter of incomplete. Indeed, Parsons explicitly denies that 'is' should be rendered as '∃'. (p.6)
2. *Zalta*: Likewise distinguishes 'exists' from 'is', existence from being. 'Exists' ('E!') is a first-level predicate. However, 'is' is predicated of abstract entities and, contrary to Parsons, is represented by '∃', which would indicate that it is a second-level predicate.

We have been attending to 'exists' and 'is' not for their own sake but purely as a prolegomenon to an ontological question, namely, that of existence. It is not for the present entry to discuss the relative merits of Frege-Russell logic (perhaps augmented by treatments of intentional propositions suggested by the likes of Howell and van Inwagen) vs intensional logics, nor the merits of Parsons' intensional proposals vs those of Zalta. The relevance of all those logics to our present purposes lies solely in their implications for the ontological status of existence, namely, that it is a first-level property. Whether it is real or Cambridge, however, will not be settled by logic but only by extra-logical considerations.

Criticisms of First-Level Use of 'Exists'

Critical responses to the suggestion that 'exists' is predicable of individuals are broadly of two kinds. One it to argue against it, the other to accept it but to restrict its consequences. Both have now to be outlined.

Argument against First-Level Use of 'Exists': C.J.F.Williams has challenged advocates of an actuality sense of 'exists' to give a genuine non-philosophical example in which 'exists' is predicated of an individual. 'Socrates no longer exists' and 'Socrates might never have existed' have been suggested, because it would seem that the former can be understood as 'It used to be the case that (Socrates exists), but it is not now the case that (Socrates exists)', and the latter as 'It might have been the case but was never the case that (Socrates exists)'. Thus understood, each proposition has 'Socrates exists' embedded in it, and hence could not make sense unless the embedded proposition also made sense, which it could

not do unless ‘exists’ were predicable of Socrates.

These putative counter-examples have failed to impress Williams (*Being, Identity, and Truth*, pp.28-33), who rejects them on two grounds:

1. Despite appearances to the contrary, ‘Socrates exists’ is *not* embedded in either proposition. Hence, they are not examples of ‘exists’ being predicated of an individual.
2. Despite its apparent absence, it is the there-is sense of ‘exists’ (and not any actuality sense) that is implicit in both propositions.

Williams agrees that a first-level use of ‘exists’ would require ‘Socrates no longer exists’ to be rendered as ‘It used to be the case that (Socrates exists), but is not now the case that (Socrates exists)’. However, because he denies any distinction between external and internal negation, he is able to say that the second clause - ‘it is not now the case that (Socrates exists)’ - should be understood as ‘Socrates does not exist’. ‘Socrates does not exist’ could never be true, however, since there could never be any fact about Socrates without Socrates himself being a constituent of it. Because a nonexistent Socrates cannot be a constituent of anything, ‘Socrates does not exist’ can make no sense. Moreover, no propositions of which it were a logical part could make any sense either. Hence, if it were a logical part of ‘Socrates no longer exists’, that proposition would make no sense. But, since it undoubtedly does make sense, neither ‘Socrates exists’ nor ‘Socrates does not exist’ can be embedded in it - contrary to what supporters of a first-level use of ‘exists’ affirm.

Having argued against any first-level use of ‘exists’ in ‘Socrates no longer exists’, Williams has to explain how that proposition is to be understood if it neither expresses any fact of which Socrates would be a constituent, nor is it *about Socrates*. He does so by reminding us that ‘when we are attempting to discover whether something is the same as something which possessed some property at an earlier time, we need predicables of reidentification’. As he explains, this means that we need two predicables, one being true of Socrates before he died and the other true of him after his death. With this in mind, he claims that ‘Socrates no longer exists’ is to be understood simply as the denial that there is any such pair of predicables. It has therefore to be understood as ‘There is *no* pair of predicables of reidentification such that one of them can be truly predicated of Socrates and the other truly predicated of someone at the present moment’.

The virtue of this interpretation, says Williams, is that it says something without, however, implying either that there is nothing for it to be about or that there is a fact of which Socrates would have to be a constituent. Although ‘Socrates no longer exists’ would be senseless if ‘exists’ were to be used in the actuality sense, it would make perfectly good sense if the there-is sense were employed.

Williams' strategy with ‘Socrates might not have existed ‘is much the same as with ‘Socrates no longer exists’. To begin with, he rejects understanding it as ‘It might have been the case that it was *not* always the case that (Socrates exists)’. Rather, the ‘not’ has to occur within the brackets, as in ‘It might have been the case that it was always the case that (Socrates does *not* exist)’. As in the earlier case, this relies

on denying any distinction between internal and external negation. Once the proposition is understood in this way, it is then open to criticism for attempting at one and the same time both to say something about a person and to imply that there is no such person to say anything about. In other words, it is an attempt to present a fact about Socrates without his being a constituent of that fact.

The problem is to find an interpretation of ‘Socrates might never have existed’ in which ‘exists’ would not even seem to be predicated of Socrates. Williams' candidate is, ‘There is a property which was an essential property of Socrates, and it might have been the case that nothing at all ever possessed this property.’ This does not imply that there is any fact having a nonexistent Socrates as one of its constituents. Consequently, propositions like ‘Socrates might never have existed’ do have a place in the language, but only if that they are understood as being about one of Socrates' properties rather than about Socrates himself.

If Williams is right, the two propositions which certainly do appear to be saying something about Socrates, are in fact saying nothing at all about him. Rather, they are merely saying something about the *instantiation of properties*. ‘Socrates might never have existed’ is saying that *there is* an essential property which belonged to Socrates and which might have belonged to no one. ‘Socrates no longer exists’ is saying that *there is* no pair of properties of reidentification one of which belongs to someone now, the other of which belonged to Socrates.

Advocates of the two-sense thesis would not feel threatened by these considerations. Rather, they would argue that it is important to realize that the arguments against ‘Socrates exists’ being embedded in the two propositions were based on the assumption that no property *at all* could be acquired by any individual who no longer existed. This assumption, however, is merely a half-truth. What is true is that no *real* property can be acquired by such an individual: Socrates cannot become wise, or tired, or angry, and so on after his death. He can have no more real properties after death than he had during his lifetime. About Cambridge properties, however, that is far from true, for he can become admired by antipoedeans, emulated by twentieth century students, reviled by twenthieth century totalitarian regimes, and so on and so on: the list could go on forever.

The question to be asked, therefore, about the predicate ‘does not exist’ is not simply whether it would, if admitted as a first-level predicate, attribute some property to a nonexistent individual, but whether that property would be a *real* one. One has only to ask what the putative property would be, in order to be assured that the answer is ‘no’; for the property in question would be non-existence, and non-existence would surely have to be the very paradigm of a Cambridge property. But, once that is accepted, there would be no reason to doubt that ‘Socrates exists’ is indeed embedded in both ‘Socrates no longer exists’ and in ‘Socrates might never have existed’.

What weight, then, should be attached to Williams' rendering of the two propositions in terms of the there-is sense of ‘exists’ rather than of the actuality sense? According to some critics, not much. The only thing to commend those renderings was the conviction that they would make no sense if understood to be employing the actuality sense. If, as the critics argue, that conviction is ill-founded, Williams' strained interpretations would have little to recommend them.

Two further objections to the first-level use of ‘exists’ are raised by Michael Dummett. One of them makes the valid point that it would follow from the two-sense thesis that the first- and second-level senses of ‘exists’ would be equivocal. Why? Because, as Dummett says, there could be no greater difference of sense than ‘one involving a difference of logical type, that between a quantifier and a first-level predicate’.(*Frege, Philosophy of Language*, p.386) The defender of the two-sense thesis would certainly agree that a difference of logical type is an absolute one, but would need to be convinced that this precludes there being any connection whatever between the two senses. As a counter-example to Dummett's claim, Geach has drawn our attention to two uses of ‘disappears’. The sense in ‘When the rescuers reached the site of the accident, the body had disappeared’ differs from that in ‘Dinosaurs have disappeared from the face of the earth’. Even though the former ia first-level use of ‘disappears’ and the latter a second-level one, however, there is at least some connection between their senses, albeit not one of even partial univocity. In Geach's view, there is no reason why the two senses of ‘exists’ should not be similarly related.

Dummett's second objection concerns the first-level use of ‘exists’ in propositions like ‘Cleopatra no longer exists’, about which he maintains that, if ‘exists’ were being said of Cleopatra, it would mean that she no longer had a certain property. But this is unacceptable for ‘existence, even when temporal, is not a property that may be first acquired and later lost’, and it makes no sense to say or imply that it is.(*Op.cit.*, p.387) Defenders of the two-sense thesis would agree. Indeed, that is precisely why they insist that propositions like ‘Cleopatra no longer exists’, ‘Cleopatra came to exist’, and ‘Cleopatra ceased to exist’ are to be understood in such a way as certainly not to imply any acquisition or loss of existence. Rather, they are to be understood respectively as ‘It is no longer the case that (Socrates exists)’, ‘It came to be that (Socrates exists)’, and ‘It ceased to be that (Socrates exists)’, none of which carries the unacceptable implication that Socrates acquired or lost any property.

Attempts to Restrict the Consequences of ‘Exists’ being a First-Level Predicate: While not wanting to deny that ‘exists’ could be predicated of individuals, some would be much concerned with what might be inferred from such a doctrine. In particular, they would want to preclude any possibility of inferring that the property referred to by ‘exists’ (existence) might not only be a real property, but also one that was irreducible to any *non*-existential properties. That possibility would be removed if, as has been argued, ‘exists’ were merely a formal predicate, an excluder, a predicate variable, or one that did duty for a disjunction of predicates. As we shall see, no one of these proposals has gone unchallenged.

‘Exists’ a Formal Predicate?: It is worth considering whether existence might not be what Wittgenstein called a formal concept, and whether ‘exists’ might not be the kind of predicate that expresses such a concept, even if only improperly. Examples of such predicates occur in ‘2 is a number’, ‘"2" is a numeral’, ‘Tom is an object’, ‘"Tom" is a name‘, ‘"The mother of Socrates" is a complex.’ Although all the predicates are first-level ones, they attribute no real property to what they are said of, but simply place them in some category. The propositions in which they occur are all quite uninformative; and, although like tautologies in that particular respect, they are unlike tautologies in that their denial is not self-contradictory. ‘Black stones are not black’ is self-contradictory, whereas ‘2 is not a number’ is not, even though it can never be true. In these respects they have much in common with ‘exists’. Hence, it might be

suggested that ‘exists’, too, is simply a formal predicate, for it is commonly claimed that ‘Socrates exists’ is uninformative and that ‘Socrates does not exist’ is not self-contradictory, notwithstanding that in certain circumstances it would be extremely odd to affirm it.

Now, an interesting feature of the propositions listed above is that, despite not being tautologies, each of them is necessarily true. 2 cannot cease to be a number, ‘2’ cannot cease to be a numeral, Tom cannot cease to be an object, nor can ‘Tom’ cease to be a proper name. It may not have been necessary that there be a 2, ‘2’, Tom, or ‘Tom’; but, given that we do have them, it can never be false to predicate the relevant formal predicates of them. It is no more true to say ‘Socrates is no longer an individual’ than to say ‘2 is no longer a number’. Some would say that it is just that characteristic of formal predicates which disqualifies ‘exists’ from being one of them. If ‘exists’ were a formal predicate, then, once ‘Socrates exists’ were true, it could never be false. Yet, although ‘Socrates exists’ was once true, it not only can be false but indeed is now false. Consequently, ‘exists’ seems not to be a formal predicate, attractive as it may have been to think otherwise.

‘Exists’ an Excluder?: To have ruled out ‘exists’ from being a formal predicate is not necessarily to have ruled out all possibility of its being a first-level predicate without existence being a real property. One other possibility is that it be what Roland Hall has called ‘an excluder’, and which he introduces as follows:

> Adjectives that (1) are attributive as opposed to predicative, (2) serve to rule out something without themselves adding anything, and (3) ambiguously rule out different things according to context, I call "excluders".

As examples, he suggests ‘ordinary’, ‘absolute’, ‘accidental’, ‘barbarian’, ‘base’, ‘civil’, ‘real’, amongst many others. The one most relevant to present purposes, however, is ‘real’.

In Hall's view, ‘real’ is the kind of adjective that merely rules out something without itself adding anything. According to context, it can rule out *a*'s being imaginary, or artificial, or counterfeit. In doing so, however, it attributes nothing positive to what it is said of: its contribution is purely negative. Moreover, what it excludes varies with context, and therein, says Hall, lies its ambiguity. Although Hall himself does not suggest that ‘exists’ is an excluder, others have done so, and still others have suggested that ‘exists’ means simply ‘is real’. On either suggestion existence would not be a real property, and that is why these views might commend themselves to anyone who was bothered by the possibility of existence being a real property of things.

The question is whether ‘exists’ really does have the three marks required of an excluder. Since it is not an adjective at all, it obviously cannot be an attributive one. However, that is of no account, for in his closing remarks Hall allows that excluders may be found not only among adjectives, but also among ‘nouns and other parts of speech.’ The questions we have to ask of ‘exists’ therefore are:

- Is it ambiguous?

- Not merely can it be defined negatively, but must it be so defined?

If ‘exists’ is to be an excluder, the answer has to be ‘yes’ not merely to one of these questions, but to both. If we accept that ‘exists’ is ambiguous, then the question of its being an excluder will turn on whether it satisfies (2), e.g., on whether ‘*a* exists’ must be understood simply in terms of what *a* was precluded from being. If so, then no matter how different the context, ‘*a* exists’ could only be understood negatively, e.g., as ‘*a* is not-fictional’, ‘*a* is not-dead’, ‘*a* is not illusory’, ‘*a* is not-mythical’, or ‘*a* is not-nonexistent’. The simplest case to consider is the last. We might envisage a seer predicting that in two years a son would be born to parents *b* and *c*, and that he would be called ‘Socrates’. When the prediction was fulfilled, we might imagine the seer announcing triumphantly ‘At last Socrates exists, just as I said he would.’ If ‘exists’ were an excluder, then the only way of understanding the seer would be as excluding some property from Socrates; and in this case the property excluded would be that of non-existence. As said by the seer, therefore, ‘At last Socrates exists’ could only mean ‘At last Socrates is not-nonexistent’. If he really were to mean that, he would have to explain just when Socrates could ever have been said to be nonexistent, i.e., never to have existed. In fact, Prior, Ryle, and others have maintained that before Socrates existed he could not even have been referred to, and hence at that time nothing at all could have been attributed to him, not even the property of being nonexistent. In that case, it would be impossible for ‘At last Socrates exists’ to mean ‘At last Socrates is no longer nonexistent’. If this is correct, ‘exists’ could not be an excluder, for there was never any property for it to exclude.

Still, in other contexts ‘Socrates exists’ might be proposed as meaning ‘Socrates is not-dead’ or ‘Socrates is non-fictional’, and ‘exists’ as excluding from Socrates the properties of being dead and being fictional respectively. The first case can scarcely be evidence for ‘exists’ being an excluder, for ‘is dead’ is itself to be understood as ‘is not-alive’. If it were evidence for anything, it would be for ‘exists’ as a synonym for ‘is alive’, except that ‘The Euston Arch no longer exists’ could hardly be understood as ‘The Euston Arch is no longer alive’.

As for ‘Socrates is non-fictional’ (‘Socrates is a non-fictional character’ would be better), it could support the claim that ‘exists’ is an excluder only if Socrates really could have been a fictional character. That, however, could have occurred only if a fictional character could ever be the same person as a real-life one, something which is extremely debatable to say the least, even though it is entirely possible that a real-life character should satisfy the same *description* as a fictional one. Precisely the same point can be made about any attempt to depict ‘exists’ as excluding ‘is mythical’. Since Socrates never could have been either a fictional or mythical character, there is nothing for ‘exists’ to exclude. That is not to say that there is anything wrong with saying ‘Socrates is not fictional’ or ‘Socrates is not mythical’, but only that it is misleading to construe those propositions as ‘Socrates is not-fictional’ and ‘Socrates is not-mythical’ rather than as ‘Not(Socrates is fictional)’ and ‘Not(Socrates is mythical)’. From all this it is clear enough that the answer to (ii) above is ‘No, "exists" need not be defined or understood purely negatively.’ Once again, it would seem that ‘exists’ cannot be an excluder.

‘Exists’ a Predicate Variable or a Disjunction of Predicates?: It has been suggested that ‘exists’ should be construed as a predicate variable or as a disjunction of predicates. In the former case ‘a exists’ would be rendered as ‘a has some property or other’ or ‘(∃*P*)(*P* is had by *a*)’, where *‘P’* is a predicate variable.

In the latter case '*a* exists' might be rendered as '*a* is *F* or *G* or *H*', where *'F'*, *'G'* and *'H'* are first-level predicates. In either case the result would be to disqualify existence from being an irreducible property of *a*. In the first case, although *a* might be allowed to have the properties referred to by whatever predicates are substituted for *'P'*, it would have no irreducibly existential property. Similarly in the second case; although *a* might have one of the properties *F*, *G*, and *H*, it would have no irreducible property of existence.

If the first suggestion were correct, 'Socrates does not exist' could be rendered as 'Socrates has no properties'. Likewise, if the second suggestion were correct, 'Socrates does not exist' could be rendered as 'Socrates has neither *F*, nor *G*, nor *H*', i.e., 'Socrates has no properties'. In either case, therefore, 'Socrates does not exist' could be understood as 'Socrates has no properties'. Defenders of the actuality sense of 'exists' would point out, however, that 'Socrates has no properties' could be true only if Socrates were a bare particular. On the contrary, 'Socrates does not exist' could be true irrespective of whether Socrates was or was not a bare particular. Yet, there should be no such difference, if 'Socrates does not exist' were to be understood as 'Socrates has no properties'. Thus, they argue that 'Socrates exists' can be rendered neither by 'Socrates has some property or other' nor by 'Socrates is either *F* or *G* or *H*'.

Ontological Implications of the Debate

Ontological Implications of 'Exists' being a Second-Level Predicate: Since, on this view, 'exists' is solely a second-level predicate, existence would ipso facto be a second-level property, a property of first-level properties. What it tells us about them is simply how often they are instantiated, namely, at least once. Moreover, since it is not a property of individuals, a fortiori it is incapable of adding anything to them, thus confirming the views of Hume, Kant, and many others.

If, as Russell, Quine, and Williams maintain, 'Socrates exists' is not about Socrates but is about various properties, this suggests the possibility that even *non*-existential propositions like 'Socrates is wise' could likewise be merely about properties rather than about Socrates. The result would be an ontology in which properties were ontologically primitive, with individuals being reducible to them. Russell and Goodman certainly accepted that view in their notion of individuals as mere 'bundles' of properties. There have indeed been a variety of bundle theories, differing according as the bundles' constituents and/or structures were different. For Russell and Goodman the constituents were universal properties, for Castaneda they were guises (constructed inter alia from properties), and for D.C.Williams and K.Campbell the constituents were not universal properties but *singular* ones known as 'tropes'. Closely related to these theories is one kind of haecceity theory, according to which individuals would be constructs of universal properties with the very important addition of one singular property (an haecceity). All such theories stand on its head the Aristotelian ontology, in which both individuals and properties are primitive. Properties, however, are ontologically posterior to individuals, for there can be no universals existing outside individuals, and their instances are individuated by the individual in which they are instantiated.

However, although the reducibility of individuals to various complexes of properties would entail a second-level view of existence, the converse is not true: the second-level view of 'exists' and existence

does not entail the reducibility of individuals, but is merely consistent with it and therefore congenial to those who have independent reasons for espousing it. Even though Williams, for example, holds to a second-level view of ‘exists’ in his interpretation of ‘Socrates no longer exists’, he could quite consistently hold to a first-level view of ‘is walking’ in his interpretation of ‘Socrates is no longer walking’.

Ontological Implications of ‘Exists’ being a First-Level Predicate:For the two-sense theorist, as well as for Parsons and Zalta, existence is a first-level property and, since ‘exists’ is arguably irreducible to other predicates, existence is arguably irreducible to other properties. The question, then, is whether it is a Cambridge property or a real one. An argument for its not being a Cambridge property is simply this:

- A Cambridge property is the referent either of a relational predicate or of a purely formal predicate.
- But, ‘exists’ is neither a relational nor a formal predicate. Not formal, for reasons given earlier. Not relational, because the totality of universes cannot be related to any other spatio-temporal entity, yet can quite properly be said to exist.
- Therefore, ‘exists’ is not a Cambridge property.
- But, if ‘exists’ is a property but not a Cambridge one, it is a real one.

A prime stumbling block to this conclusion is the objection raised by Hume, Kant, and many others that Socrates' existence could be a real property only if it added something to him, which it clearly does not. The obvious assumption here is that, if his existence were a real property, it would have to be like his other real properties such as his wisdom; and since they do add something to him, his existence should either do the same or forfeit all claim to being a real property. Underlying this suggestion is the further assumption that the relation between Socrates and his existence must duplicate that between him and his wisdom not merely in some respects but in all respects.

Now, a condition of wisdom and existence being properties of Socrates is that they each be individuated by him. That is to say, just as the wisdom-of Socrates must differ at least numerically from the wisdom-of-Plato, so too must the existence-of-Socrates differ at least numerically from the existence-of-Plato. (Although bundle theorists of all persuasions would deny this inference, their denials can be discounted, since it can be argued that each version of the theory entails the self-defeating conclusion that in some cases even individuals could themselves be instantiated in other individuals.) It is one thing, however, to claim (correctly) that existence and wisdom have both to be individuated by Socrates, but quite another to say that both individuations must have the same ground. In the case of his wisdom and many other properties, he individuates them in being their recipient. Yet, although he clearly could not be the recipient of his existence, that need not preclude there being some other way in which he does individuate it. If there were, his existence would be no less entitled to be called a property than is his wisdom.

Is there such a way? It has been argued that there is, for Socrates would individuate his existence if he were not its recipient but its *bound*. (Cf. B.Miller, *The Fullness of Being*, chapter 4) The notion of a bound is more than a merely a spatial one, for there are also bounds of thought, bounds of desire, artistic bounds, and so on. Socrates would be the bound of his existence in respect of all human areas. In being

bounded by him, his existence would be individuated and distinguished from Plato's no less than, in being *received* by Socrates, his wisdom is individuated and distinguished from Plato's. Moreover, since no bound can be real unless what it bounds is also real, the fact that Socrates is real would entail that what he bounds (his existence) must be real as well. The salient point, however, is that it makes no sense to speak of a thing that is bounded being added to by its bound. Thus, even though it makes no sense to say that Socrates' existence adds anything to him, that would not detract from its being not only a property, but a real one to boot.

From this conclusion follow two interesting consequences, one ontological and the other logical. The former is that it would remove the misconception of existence as the most impoverished of properties. This view derives from an understanding of existence as simply ‘that attribute which is common to mice and men, dust and angels’. (A.Kenny, *The Five Ways*, London, Routledge and Kegan Paul, 1969, p.92) If, however, instances of existence were both real and bounded by the individuals that had them, then not only would Kenny's inference be invalid, but his conclusion would be false. This is because the wealth or poverty of instances of existence would vary in direct relation to the constricting character of their bounds. An amoeba would be a more constricting bound than a gazelle, which in turn would be more constricting than a human. And, naturally, there would be variations from one amoeba to another, from one gazelle to another, and from one human to another. Thus, the richness or poverty of existence would be far from invariant across the whole range of individuals, as Kenny and others have supposed it to be.

This ontological point has a logical corollary. It has commonly been thought that a first-level role for ‘exists’ in ‘Socrates exists’ would be the same as the innumerable predicates like ‘is brown’ in ‘Socrates is brown’ and ‘This gazelle is brown’. The point about ‘is brown’ is that it can attribute exactly the same kind of property to individuals as diverse as Socrates, a gazelle, and even a plank of wood: it is possible that the brownness of each individual differ neither in kind nor degree from that had by any other. To think of ‘exists’ in this way would be to regard it as attributing to Socrates neither more nor less that it does to a grain of sand.

What the analogy ignores, however, is that not all predicates conform to the model of ‘is brown’. An everyday example of a quite different model is ‘is fast’ in ‘Simon (a snail) is fast’, ‘Socrates is fast’, ‘Gerry (a gazelle) is fast’ and ‘Fred (a fighter plane) is fast’. The relevance of these four propositions is that sameness of predicate in each of them does not entail sameness of speed in each of their subjects. On the contrary, the property of being fast is directly relative to whatever its subject may be. Similarly with ‘is big’, which can be said of a flea no less than of an elephant or a skyscraper without however attributing the same size to each. ‘Fast’ and ‘big’ are what Geach has called attributive adjectives; ‘brown’ is called a predicative adjective. If instances of existence vary in direct relation to the individuals that bound them, it is clear that ‘exists’ cannot be like ‘is brown’, but is in fact like ‘is fast’ and ‘is big’, for which reason we might call it not an attributive adjective (for it is not an adjective at all), but an attributive term.

Of course, if existence proved not to be a real property, then neither the ontological nor the logical point would ensue.

Bibliography

- Aquinas, *St.Thomas, Summa Theologiae* , Rome: Leonine ed., vols. 4-12, 1888-1906. English translation, London: Eyre and Spottiswoode, 1964-73; *Quaestiones disputatae de veritate*, Rome: Leonine ed., vol.22, 1970-76. English translation, by R.Schmidt, *The Disputed Questions on Truth*, Chicago: Regnery, 1954 ; *Quaestiones disputatae de potentiae*, Rome: Marietti, 1965. English translation by the Dominican Fathers, *On the power of God*; *Quaestiones disputatae de malo*, Rome: Marietti, 1965. English translation by J.Oesterle, *On Evil*, Notre Dame: University of Notre Dame Press, 1995; *Quaestiones quodlibetales*, Rome: Marietti, 1956
- Butchvarov, P.,*Being qua Being*, Bloomington: Indiana University Press, 1979
- Connell, D., *Essays in Metaphysics*, Dublin: Four Courts Press, 1996
- Dancy, R., ‘Aristotle and Existence’, in Knuuttilla & Hintikka (see below), pp.49-80
- Dummett, M., *Frege: Philosophy of Language*, London, Duckworth, second revised edition, 1973
- Frege, G., *Die Grundlagen der Arithmetik*. English translation: J.L. Austin, *The Foundations of Arithmetic*, Oxford: Blackwell, second revised edition 1974.
- Geach, P.T., ‘Form and Existence’, *Proceedings of the Aristotelian Society*, 55(1954-5); reprinted in his *God and the Soul*,London: Routledge & Kegan Paul, 1969, pp.42-64
- Geach, P.T., ‘What Actually Exists’, *Proceedings of the Aristotlian Society*, Supp. vol. 42(1968), pp. 7-16
- Gibson, Q., *The Existence Principle*, Dordrecht: Kluwer Academic Publishers, 1998
- Gilson, E., *Being and Some Philosophers*, Toronto: Pontifical Institute of Mediaeval Studies Press, 1952
- Haaparanta, L., ‘On Frege's Concept of Being’, in Knuuttila & Hintikka (see below), pp.269-290
- Hintikka, J., ‘The Varieties of Being in Aristotle’ in Knuuttila & Hintikka (see below), pp.81-114
- Hintikka, J., ‘Kant, Existence, Predication and the Ontological Argument’ in Knuuttila & Hintikka (see below), pp.249-268
- Hume, D., *A Treatise of Human Nature*, ed.L.A. Selby-Bigge, Oxford: O.U.P., 1951
- Kant, I., *Critique of Pure Reason*, , ed. by N. Kemp-Smith, London: Macmillan, 1929
- Knuuttila, S. & Hintikka, J., *The Logic of Being*, Dordrecht: Reidel, 1986
- Mackie, J., ‘The Riddle of Existence’, *Proceedings of the Aristotelian Society*, Supp. vol. 50(1976), pp.247-266
- Miller, B., *The Fullness of Being*, Notre Dame: University of Notre Dame Press, 2002
- Miller, B., ‘In Defence of the Predicate "Exists"’, *Mind*, 84(1975), pp.338-354
- Miller, B., ‘"Exists" and Existence’, *Review Of Metaphysics*, 40(1986-7), 237-270
- Miller, B., *A Most Unlikely God*, Notre Dame: University of Notre Dame Press, 1996, ch.3
- Munitz, M., *Existence and Logic*, New York: New York University Press, 1974
- Owen, G. E. L., ‘Aristotle on the Snares of Ontology’, in R.Bambrough (ed.), *New Essays on Plato and Aristotle*, London: Routledge and Kegan Paul, 1965, pp.69-96
- Parsons, T., *Nonexistent Objects*, New Haven: Yale University Press, 1980
- Pears, D., ‘Is Existence a Predicate?’, in P. Strawson, (ed.), *Philosophical Logic*, Oxford: O.U.P., 1967. pp.97-102
- Quine, W.V., ‘On What There Is’ in his *From a Logical Point of View*, New York: Harper and

Row, 1963, 1-19

- Quine, W.V., *Word and Object*, Cambridge MA, M.I.T. Press, 1960
- Russell, B., ‘On Denoting’, in R. C. Marsh (ed.), *Logic and Knowledge*, London: Allen and Unwin, 1968
- Weidemann, H., ‘The Logic of Being in Thomas Aquinas’ in Knuutilla & Hintikka (see above), pp.181-200
- Williams, C. J. F., *What is Existence?*, Oxford, O.U.P., 1981
- Williams, C. J. F., *Being, Identity, and Truth*, Oxford, O.U.P., 1992
- Zalta, E., *Intensional Logic and the Metaphysics of Intentionality*, Cambridge MA: MIT Press, 1988

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Anselm, Saint [Anselm of Bec, Anselm of Canterbury] | Aquinas, Saint Thomas | Aristotle | Avicenna | Frege, Gottlob | Hume, David | Kant, Immanuel

Barry Miller
brmiller@ihug.com.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 22, 1996
Content last modified: May 23, 2002

Saint Thomas Aquinas

Thomas Aquinas (1225-1274) lived at a critical juncture of western culture when the arrival of the Aristotelian *corpus* in Latin translation reopened the question of the relation between faith and reason, calling into question the *modus vivendi* that had obtained for centuries. This crisis flared up just as universities were being founded. Thomas, after early studies at Montecassino, moved on to the University of Naples, where he met members of the new Dominican Order. It was at Naples too that Thomas had his first extended contact with the new learning. When he joined the Dominican Order he went north to study with Albertus Magnus, author of a paraphrase of the Aristotelian *corpus*. Thomas completed his studies at the University of Paris, which had been formed out of the monastic schools on the Left Bank and the cathedral school at Notre Dame. In two stints as a regent master Thomas defended the mendicant orders and, of greater historical importance, countered both the Averroistic interpretations of Aristotle and the Franciscan tendency to reject Greek philosophy. The result was a new *modus vivendi* between faith and philosophy which survived until the rise of the new physics. Thomas's theological writings became regulative of the Catholic Church and his close textual commentaries on Aristotle represent a cultural resource which is now receiving increased recognition. The following account concentrates on Thomas the philosopher and presents him as fundamentally an Aristotelian.

Life and Works

Vita Brevis

Thomas was born in 1225 at Roccasecca, a hilltop castle from which the great Benedictine abbey of Montecassino is not quite visible, midway between Rome and Naples. At the age of five, he was entered at Montecassino where his studies began. When the monastery became a battle site -- not for the last time -- Thomas was transferred by his family to the University of Naples. It was here that he came into contact with the "new" Aristotle and with the Order of Preachers or Dominicans, a recently founded mendicant order. He became a Dominican over the protests of his family and eventually went north to study, perhaps first briefly at Paris, then at Cologne with Albert the Great, whose interest in Aristotle strengthened Thomas's own predilections. Returned to Paris, he completed his studies, became a Master and for three years occupied one of the Dominican chairs in the Faculty of Theology. The next ten years were spent in various places in Italy, with the mobile papal court, at various Dominican houses, eventually in Rome. From there he was called back to Paris to confront the hullabaloo variously called Latin Averroism and Heterodox Aristotelianism. After this second three year stint, he was assigned to Naples. In 1274, on his way to the Council of Lyon, he fell ill and died on March 7 in the Cistercian abbey at Fossanova, which is perhaps twenty kilometers from Roccasecca.

Education

Little is known of Thomas's studies at Montecassino, but much is known of the shape that the monastic schools had taken. They were one of the principal conduits of the liberal arts tradition which stretches back to Cassiodorus Senator in the 6^{th} century. The arts of the trivium (grammar, rhetoric, logic) and those of the quadrivium (arithmetic, geometry, music and astronomy) were fragments shored against the ruinous loss of classical knowledge. They constituted the secular education that complemented sacred doctrine as learned from the Bible. When Thomas transferred to Naples, his education in the arts continued. Here it would have been borne upon him that the liberal arts were no longer adequate categories of secular learning: the new translations of Aristotle spelled the end of the liberal arts tradition, although the universities effected a transition rather than a breach.

Taking Thomas's alma mater Paris as reference point, the Faculty of Arts provided the point of entry to teen-aged boys. With the attainment of the Master of Arts at about the age of 20, one could go on to study in a higher faculty, law, medicine or theology. The theological program Thomas entered in Paris was a grueling one, with the master's typically attained in the early thirties. Extensive and progressively more intensive study of the scriptures, Old and New Testament, and of the summary of Christian doctrine called the *Sentences* which was compiled by the twelfth century Bishop of Paris, Peter Lombard. These close textual studies were complemented by public disputations and the even more unruly quodlibetal questions. Modeled more or less on the guilds, the student served a long apprenticeship, established his competence in stages and eventually, after a public examination was named a master and then gave his inaugural lecture.

Writings

Thomas's writings by and large show their provenance in his teaching duties. His commentary on the *Sentences* put the seal on his student days and many of his very early commentaries on Scripture have come down to us. But from the very beginning Thomas produces writings which would not have emerged from the usual tasks of the theological master. *On Being and Essence* and *The Principles of Nature* date from his first stay at Paris, and unlike his commentaries on Boethius' *On the Trinity* and *De hebdomadibus*, are quite obviously philosophical works. Some of his disputed questions date from his first stint as regius master at Paris. When he returned to Italy his productivity increased. He finished the *Summa contra gentiles*, wrote various disputed questions and began the *Summa theologiae*. In 1268, at Rome, he began the work of commenting on Aristotle with *On the Soul*, and during the next five or six years commented on eleven more (not all of these are complete). During this time he was caught up in magisterial duties of unusual scope and was writing such polemical works as *On the Eternity of the World* and *On There Being Only One Intellect*.

At Naples, he was given the task of elevating the status of the Dominican House of Studies. His writing continued until he had a mystical experience which made him think of all he had done as "mere straw." At the time of his death in 1274 he was under a cloud in Paris and in 1277, 219 propositions were condemned by a commission appointed by the Bishop of Paris, among them tenets of Thomas. This was soon lifted, he was canonized and eventually was given the title of Common Doctor of the Church. But the subtle and delicate assimilation of Aristotle that characterized his work in both philosophy and theology did not survive his death, outside the Dominican Order, and has experienced ups and downs ever since.

Philosophy and Theology

Many contemporary philosophers are unsure how to read Thomas. As the above sketch makes clear, he was a professional theologian. Nonetheless, we find among his writings works anyone would recognize as philosophical and the dozen commentaries on Aristotle increasingly enjoy the respect and interest of Aristotelian scholars. But his best known work is the *Summa theologiae*, which is most often cited when Thomas's position on this or that is sought. How can a theological work provide grist for philosophical mills? How did Thomas distinguish between philosophy and theology?

Sometimes, Thomas puts the difference this way: "... the believer and the philosopher consider creatures differently. The philosopher considers what belongs to their proper natures, while the believer considers only what is true of creatures insofar as they are related to God, for example, that they are created by God and are subject to him, and the like." (*Summa contra gentiles*, bk II, chap. 4) Since the philosopher too, according to Thomas, considers things as they relate to God, this statement does not put the difference in a formal light.

The first and major formal difference between philosophy and theology is found in their principles, that

is, starting points. The assumptions of the philosopher, that to which his discussions and arguments are ultimately driven back, are in the public domain. They are things that everyone knows; they are where disagreement between us must come to an end. These principles are not themselves the products of proof -- which does not of course mean that they are immune to rational analysis and inquiry -- and thus they are said to be known by themselves (*per se*, as opposed to *per alia*). This is proportionately true of each of the sciences, where the most common principles just alluded to are in the background and the proper principles or starting points of the particular science function regionally as the common principles do across the whole terrain of thought and being.

By contrast, the discourse of the theologian is ultimately driven back to starting points or principles that are held to be true on the basis of faith, that is, the truths that are authoritatively conveyed by the Bible. Some believers reflect on these truths and see other truths implied by them, spell out their interrelations and defend them against the accusation of being nonsense. Theological discourse looks like any other discourse and is, needless to say, governed by the common principles of thought and being, but it is characterized formally by the fact that its arguments and analyses are truth-bearing only for one who accepts Scriptural revelation as true.

This provides a formal test for deciding whether a piece of discourse is philosophical or theological. If it relies only on truths anyone can be expected to know about the world, and if it offers to lead to new truths on the basis of such truths, and only on that basis, then it is philosophical discourse. On the other hand, discourse whose cogency -- not formal, but substantive -- depends upon our accepting as true such claims as that there are three persons in one divine nature, that our salvation was effected by the sacrifice of Jesus, that Jesus is one person but two natures, one human, one divine, and the like, is theological discourse. Any appeal to an authoritative scriptural source as the necessary nexus in an argument is thereby other than philosophical discourse.

More will be said of this contrast later, but this is the essential difference Thomas recognizes between philosophy and theology. I will conclude this paragraph with a passage in which Thomas summarizes his position. He is confronting an objection to there being any need for theological discourse. Whatever can be the object of inquiry will qualify as a being of one sort or another; but the philosophical disciplines seem to cover every kind of being, indeed there is even a part of it which Aristotle calls theology. So what need is there for discourse beyond philosophical discourse?

> ....it should be noted that different ways of knowing (*ratio cognoscibilis*) give us different sciences. The astronomer and the natural philosopher both conclude that the earth is round, but the astronomer does this through a mathematical middle that is abstracted from matter, whereas the natural philosopher considers a middle lodged in matter. Thus there is nothing to prevent another science from treating in the light of divine revelation what the philosophical disciplines treat as knowable in the light of human reason. (*Summa theologiae*, Ia, q. 1, a., ad 2)

Christian Philosophy

It will be observed that the formal distinction between philosophical and theological discourse leaves untouched what has often been the mark of one who is at once a believer and a philosopher. It is not simply that he might on one occasion produce an argument that is philosophical and at another time one that is theological; his religious beliefs are clearly not put in escrow but are very much in evidence when he functions as a philosopher. Many of the questions that can be raised philosophically are such that the believer already has answers to them -- from his religious faith. How then can he be thought to be ready to follow the argument whither it listeth, as the objector is unlikely to put it? Furthermore, the inquiries in which the believer who philosophizes engages will often betray his religious beliefs.

When such observations turn into objections, perhaps into the accusation that a believer cannot be a proper philosopher, there is often an unexamined notion of what a proper philosopher looks like. The proper philosopher may be thought to be someone -- perhaps merely some mind -- without antecedents or history who first comes to consciousness posing a philosophical question the answer to which is pursued without prejudice. But of course no human being and thus no philosopher is pure reason, mind alone, without previous history as he embarks on the task of philosophizing. One has necessarily knocked about in the world for a long time before he signs up for Philosophy 101. He has at hand or rattling around in his mind all kinds of ready responses to situations and questions. He very likely engaged in some kind of inquiry about whether or not to begin the formal study of philosophy in the first place. This may be acknowledged, but with the proviso that step one in the pursuit of philosophy is to rid the mind of all such antecedents. They must be put in the dock, put in brackets, placed in doubt, regarded with suspicion. Only after appropriate epistemological cleansing is the mind equipped to make its first warranted knowledge claim. Knowledge thus becomes a deliverance of philosophy, a product of philosophizing. Outside of philosophy there is no knowledge.

The preceding paragraph has been meant to capture the salient note of much modern philosophy since Descartes. Philosophy is first of all a search for defensible knowledge claims, and for the method according to which it will be found. As opposed to what?

As opposed to the view of philosophy described in paragraph 2, Thomas understands philosophizing to depend upon antecedent knowledge, to proceed from it, and to be unintelligible unless, in its sophisticated modes, it can be traced back to the common truths known to all. But this tracing back will pass through very different terrains, depending on the upbringing, culture and other vagaries and accidents of a given person's experience. The pre-philosophical -- I refer to the formal study of philosophy -- outlook of the believer will be characterizable in a given way, a way suggested above. It is more difficult to characterize the pre-philosophical attitudes and beliefs out of which the non-believer philosophizes. Let us imagine that he holds in a more or less unexamined way that all events, including thinking, are physical events. If he should, as a philosopher, take up the question of the immortality of the soul, he is going to regard with suspicion those classical proofs which rely on an analysis of thinking as a non-physical process. The Christian, on the other hand, will be well-disposed towards efforts to prove the immortality of the human soul and will accordingly approach descriptions of thinking as non-physical sympathetically. He is unlikely to view with equanimity any claim that for human beings death is the utter end.

The importance of this is that a believer runs the risk of accepting bad proofs of the immateriality of thinking and thus of the human soul. On the other hand, a committed materialist may be too quick to accept a bad proof that thinking is just a material process. Such antecedent stances are often the reason why philosophical agreement is so hard to reach. Does it make it impossible? Do such considerations destroy any hope of philosophical objectivity? Surely not, in principle. Believers and non-believers should be able to agree on what counts as a good proof in a given area even if they expect different results from such a proof. Thinking either is or is not merely a physical process and antecedent expectations do not settle the question, however they influence the pursuit of that objective resolution. But the important point is that antecedent dispositions and expectations are the common condition of philosophers, believers and unbelievers alike. Of course, believers hold that they have an advantage here, since the antecedents that influence them are revealed truths, not just hearsay, received opinion, the *zeitgeist* or prejudice.

Thomas and Aristotle

As a philosopher, Thomas is emphatically Aristotelian. His interest in and perceptive understanding of the Stagyrite is present from his earliest years and certainly did not await the period toward the end of his life when he wrote his close textual commentaries on Aristotle. When Thomas referred to Aristotle as the Philosopher, he was not merely adopting a *facon de parler* of the time. He adopted Aristotle's analysis of physical objects, his view of place, time and motion, his proof of the prime mover, his cosmology. He made his own Aristotle's account of sense perception and intellectual knowledge. His moral philosophy is closely based on what he learned from Aristotle and in his commentary on the *Metaphysics* he provides the most cogent and coherent account of what is going on in those difficult pages. But to acknowledge the primary role of Aristotle in Thomas's philosophy is not to deny other influences. Augustine is a massively important presence. Boethius, Pseudo-Dionysius and Proclus were conduits through which he learned Neo-platonism. There is nothing more obviously Aristotelian about Thomas than his assumption that there is something to be learned from any author, if only mistakes to be avoided. But he adopted many features from non-Aristotelian sources.

This has led some to suggest that what is called Thomistic philosophy is an eclectic hodgepodge, not a set of coherent disciplines. Others, struck by the prominence in Thomas of such Platonic notions as participation, have argued that his thought is fundamentally Platonic, not Aristotelian. There is also an undeniable anti-Aristotelian animus in some influential students of Thomas, for example, Etienne Gilson. For Gilson, there is a radically original Thomistic philosophy which cannot be characterized by anything it shares with earlier thinkers, particularly Aristotle.

The recognition that Thomas is fundamentally an Aristotelian is not equivalent to the claim that Aristotle is the only influence on him. It is the claim that whatever Thomas takes on from other sources is held to be compatible with what he already holds in common with Aristotle. And, of course, to draw attention to the sources of Thomas's philosophy is not to say that everything he holds philosophically can be parsed back into historical antecedents.

The Order of Philosophical Inquiry

Thomas takes "philosophy" to be an umbrella term which covers an ordered set of sciences. Philosophical thinking is characterized by its argumentative structure and a science is taken to be principally the discovery of the properties of kinds of things. But thinking is sometimes theoretical and sometimes practical. The practical use of the mind has as its object the guidance of some activity other than thinking -- choosing in the case of moral action, some product in the case of art. The theoretical use of the mind has truth as its object: it seeks not to change the world but to understand it. Like Aristotle, Thomas holds that there is a plurality of both theoretical and practical sciences. Ethics, economics and politics are the practical sciences, while physics, mathematics and metaphysics are the theoretical sciences.

That is one way to lay out the various philosophical disciplines. But there is another that has to do with the appropriate order in which they should be studied. That order of learning is as follows: logic, mathematics, natural philosophy, moral philosophy, metaphysics. The primacy of logic stems from the fact that we have to know what knowledge is so we will recognize that we have met its demands in a particular case. The study of mathematics comes early because little experience of the world is required to master it. But when we turn to knowledge of the physical world, there is an ever increasing dependence upon a wide and deep experience of things. Moral philosophy requires not only experience, but good upbringing and the banking of the passions. Metaphysics or wisdom, is the culminating and defining goal of philosophical inquiry: it is such knowledge as we can achieve of the divine, the first cause of all else.

Thomas commented on two logical works of Aristotle: *On Interpretation* (incomplete) and *Posterior Analytics*. On mathematics, there are only glancing allusions in Thomas's writings. Thomas describes logic as dealing with "second intentions," that is, with relations which attach to concepts expressive of the natures of existent things, first intentions. This means that logic rides piggy-back on direct knowledge of the world and thus incorporates the view that what is primary in our knowledge is the things of which we first form concepts. Mathematical entities are idealizations made by way of abstraction from our knowledge of sensible things. It is knowledge of sensible things which is primary and thus prior to the "order of learning" the philosophical sciences.

This epistemological primacy of knowledge of what we grasp by our senses is the basis for the primacy of the sensible in our language. Language is expressive of knowledge and thus what is first and most easily knowable by us will be what our language first expresses. That is the rule. It is interesting to see its application in the development of the philosophy of nature.

Composition of Physical Objects

The concern of natural science is of course natural things, physical objects, which may be described as "what come to be as the result of a change and undergo change." The first task of natural philosophy, accordingly, is to define and analyze physical objects.

The first thing to notice about this is the assumption that we begin our study of the natural world, not with the presumed ultimate alphabet with which macrocosmic things are spelled, but with a vague and comprehensive concept which encompasses whatever has come to be as the result of a change and undergoes change. The reader of Aquinas becomes familiar with this assumption. Thomas learned it from the beginning of Aristotle's *Physics*.

> The natural way of doing this is to start from the things which are more knowable and clear to us and to proceed towards those which are clearer and more knowable by nature; for the same things are not knowable relatively to us and knowable without qualification. So we must follow this method and advance from what is more obscure by nature, but clearer to us, towards what is more clear and more knowable by nature.
>
> Now what is to us plain and clear at first is rather confused masses, the elements and principles of which become known to us later by analysis. Thus we must advance from universals to particulars; for it is a whole that is more knowable to sense-perception, and a universal is a kind of whole, comprehending many things within it, like parts. Much the same thing happens in the relation of the name to the formula. A name, e.g. ‘Circle’, means vaguely a sort of whole: its definition analyses this into particulars. Similarly a child begins by calling all men father, and all women mother, but later on distinguishes each of them." (*Physics*, 1, 1.)

Thomas calls the movement from the more to the less general in a science the "order of determination" or specification of the subject matter. The first purchase on natural things is via "physical object" or "natural thing." The "order of demonstration" involves finding the properties of things as known through this general concept. Then, specifying the subject further, one seeks properties of things known through the less common concepts. For example, in plane geometry, one would begin with plane figure and discover what belongs to it as such. Then one would turn to, say, triangle and seek its properties, after which one would go on to scalene and isosceles. So one will, having determined what is true of things insofar as they are physical objects, go on to seek the properties of things which are physical objects of this kind or that, for example, living and non-living bodies.

Thomas emphasizes those passages in the Aristotelian natural writings which speak of the order of determination, that is, of what considerations come first and are presupposed to those that come later. In several places, Thomas takes great pains to array the Aristotelian natural writings according to this Aristotelian principle, most notably perhaps at the outset of his commentary on *Sense and sensibilia.* The *Physics* is the first step in the study of the natural world and exhibits the rule that what is first and most easily known by us are generalities. The language used to express knowledge of such generalities will have, as we shall emphasize, a long career in subsequent inquiries, both in natural philosophy and beyond. What is sometimes thought of as a technical vocabulary, perhaps even as Aristotelian jargon, is seen by Thomas Aquinas as exemplifying the rule that we name things as we know them and that we come to know more difficult things after the easier things and extend the language used to speak of the easier, adjusting it to an every expanding set of referents.

Matter and Form

Although natural things are first thought of and analyzed in the most general of terms, there are not of course any general physical objects, only particular ones. Thus, in seeking to discern what is true of anything that has come to be as a result of a change and is subject to change until it ceases to be, Aristotle had to begin with a particular example of change, one so obvious that we would not be distracted by any difficulties in accepting it as such. "A man becomes musical." Someone acquires a skill he did not previously have. Little Imogene learns to play the harmonica. Thomas pores over the analysis Aristotle provides of this instance of change and its product.

The change may be expressed in three ways:

1. Man becomes musical.
2. What is not-musical becomes musical.
3. A not-musical man becomes musical.

These are three different expressions of the same change and they all exhibit the form A becomes B. But change can also be expressed as From A, B comes to be. Could 1, 2 and 3 be restated in that second form? To say "From the not-musical the musical comes to be" and "From a not-musical man the musical comes to be" seem acceptable alternatives, but "From man musical comes to be" would give us pause. Why? Unlike "A becomes B" the form "From A, B comes to be" suggests that in order for B to emerge, A must cease to be. This grounds the distinction between the grammatical subject of the sentence expressing a change and the subject of the change. The definition of the subject of the change is "that to which the change is attributed and which survives the change." The grammatical subjects of 2 and 3 do not express the subject of the change, only in 1 is the grammatical subject expressive of the subject of the change.

This makes clear that the different expressions of the change involve two things other than the subject of the change: the characteristics of the subject before (not-musical) and after (musical) the change. These elements of the change get the names that stick from another example, whittling wood. The term for wood in Greek is *hyle* and the term for shape, the external contours of a thing, is *morphe*. In English, form, a synonym of shape, is used to express the characteristic that the subject acquires as the result of the change, e.g. musical. The characterization of the subject prior to the change as not having the form is called privation. Using this language as canonical, Aristotle speaks of the subject of the change as its *hyle* or matter, the character it gains as its *morphe* or form, and its prior lack of the form as its privation. Any change will involve these three elements: matter, form and privation. The product of a change involves two things: matter and form.

Change takes place in the categories of quality, quantity and place, but in all cases the terminology of matter, form and privation comes to be used. The terms bind together similar but different kinds of change -- a subject changing temperature is like a subject changing place or size.

Substantial Change

The analysis of change and the product of change begins with surface changes. Some enduring thing changes place or quality or quantity. But enduring things like men and trees and horses and the like have also come into being and are destined some day to cease to be. Such things are called substances. It is a given that there are substances and that they come to be and pass away. The question is: Can the analysis of surface change be adjusted and applied to substantial change? What would its subject be? Aristotle said that the subject of substantial change is known on an analogy with the subject of incidental or surface change. That is, if substances come to be as the result of a change, and if our analysis of change can apply, there must be a subject of the change. The subject of a surface or incidental change is a substance. The subject of a substantial change cannot be a substance; if it were, the result would be a modification of that substance, that is, an incidental change. But we are trying to understand how a substance itself comes into being as the result of a change. There must be a matter or subject but it cannot be matter in the sense of a substance. In order to signal this, we can call the matter *prime matter*, first matter. The form such a subject takes on as the result of the change cannot be an incidental form like size or location or temperature. Substances do not become or cease to be substances as a result of changes in these features. As the analysis of incidental change makes clear, the substance previously existed without the form it acquires in the change and it could lose it and still be itself. The form in a substantial change must be that which makes the substance to be what it is. Call it *substantial form*.

The thing to notice about this analysis is that substantial change is spoken of on an analogy with incidental change. The analysis of incidental change is presupposed and regulative. Moreover, the language used to speak of the elements of incidental change are extended to substantial change and altered in meaning so as to avoid equivocation. Thomas sees Aristotle's philosophical vocabulary arising out of analysis of what is most obvious to us and then progressively extended to more and more things insofar as the later is made known by appeal to the prior. Thus we can see that matter and form apply in a graded and connected way to the various kinds of incidental change and then to substantial change. It is this feature of Aristotle's language that Thomas adopts as his own. It both provides the lens through which Aristotle can be correctly seen and it provides a rule for knowing and naming that will characterize Thomas's use of Latin in philosophy and in theology as well.

Perception and Thought

When the discussion moves on from what may be said of all physical objects as such to an inquiry into living physical things, the analyses build upon those already completed. Thus, "soul" will be defined as the substantial form of living bodies. The peculiar activities of living things will be grouped under headings like nutrition and growth, sense perception and knowing and willing. Since a living thing sometimes manifests an instance of such activities and sometimes does not, they relate to it in the manner of the incidental forms of any physical object. And this provides Thomas, as he feels it had Aristotle, with a path of procedure. Let us skip to the cognitive activities of living things.

How can we best analyze perception -- that is, seeing, feeling, hearing, and the like? In continuity with what has gone before, the questions are put in this form: How best to analyze coming to see, coming to

feel, coming to hear, and the like. Seeing these on the analogy of change as already analyzed, we look for a subject, a privation and a form. The sensing subject is, say, the animal, but the proximate subjects to which they are attributed are the powers of sight, touch, hearing, and the like. An instance of seeing is describable as the power's moving from not seeing to seeing. Since the object of seeing is color, the change from not seeing to seeing issues in the power having the form of color.

We could give as an example of physical change, a substance acquiring a color. Now, while there are physical changes involved in sensation -- the organs are altered in the way physical bodies are -- that is not the change involved in perception as such. That is, in feeling a body my hand's own temperature is altered by the contact. But feeling cannot be just that, since any two physical bodies that come into contact would undergo a similar alteration of temperature. Feeling the temperature, becoming aware of it, is another sort of change, however much it depends on a contemporary physical change in the organs of sense. Having the color or temperature in this further sense is thus made known and named by reference to physical change. The fundamental difference between the two ways of acquiring a form is this. In a physical change of color, the change produces a new numerical instance of the color. In grasping or sensing a color, a numerically new instance of color does not result.

We have here the basis for talk of immateriality in perception. If the acquiring of a form by matter in physical change results in a new instance of the form and this is not the case with perception, we can make this point that acquiring the form in sensation is not identical to the acquiring of the form by matter *in the primary sense.* Thus, we both want to speak of the subject of sensation on an analogy with physical change *and* to distinguish the former from the latter. This is done by speaking of the immaterial reception of a form. Nonetheless, the sense power is a subject and thus matter in an analogous sense.

In his interpretation of Aristotle's *De anima* Thomas defends a view that was as contested in his own time as it is almost an orphan in our own. Among the tenets of so-called Latin Averroism was the view, first held by Averroes, that the move from perceptive acts to intellection is not one from a lower to a higher set of capacities or faculties of the human soul. When Aristotle contrasts intellection with perception and argues that the former does not employ a sense organ because it displays none of the characteristics of perception which does employ an organ, he is not, on the Latin-Averroistic view, referring to another capacity of the human soul, the intellect, but rather referring to a separate entity thanks to whose action human beings engage in what we call thinking. But the cause of this, the agent intellect, is not a faculty of the soul. (Aristotle distinguished two intellects, a passive and an active.) The proof for immortality which results from a wholly immaterial activity is therefore a statement about the incorruptibility of the separate entity, not a basis for arguing that each human soul is immortal because it has the capacity to perform immaterial activities. The Latin-Averroists consequently denied that Aristotle taught personal immortality.

Given this consequence, Thomas's adoption of the opposite interpretation -- viz. that the agent intellect is, like the passive intellect, a faculty of the human soul -- may seem merely an interested desire to enlist Aristotle's support for a position in harmony with Christian belief. Thomas is frequently said to have baptized Aristotle, which seems to mean that he fitted him to the Procrustean bed of Christian doctrine. Of course, the full Christian view is not simply that the soul survives death but that it will be reunited with

body, and Thomas nowhere suggests that there is any intimation of this in Aristotle. Oddly enough, it is often friends of St. Thomas who suggest that he *used* Aristotle and was not chiefly concerned with what Aristotle might actually have intended.

Surely this is libelous. It would be less of an accusation to say that Thomas got a passage wrong than that he pretended it meant something he knew it did not. But the important point, all these centuries later, is whether Thomas's reading is or is not supported by the text. When he commented on the *De anima*, he seems not to be concerned with the flare up in Paris over Latin Averroism. This is the basis for dating the commentary in 1268, before Thomas returned to Paris. The commentary, accordingly, cannot be read as though it were prompted by the controversy. Of course, Thomas might be said to have long term interests in taming Aristotle to behave in a Christian way. As it happens, during the second Parisian period, in the thick of the Latin-Averroist controversy, Thomas wrote an opusculum dedicated to the question: what did Aristotle actually teach? The work is called in the Latin, *De unitate intellectus contra averroistas*. I have translated it as, *On there being only one intellect*. This little work is absolutely essential for assessing the nature of Thomas's Aristotelianism. He provides us with an extended textual analysis to show that the rival interpretation cannot be sustained by the text and that the only coherent reading of the *De anima* must view the agent and passive intellects as faculties of the human soul. His interpretation may be right or wrong, but the matter must be decided on the basis of textual interpretation, not vague remarks about Thomas's intentions.

Beyond Physics

When Aristotle rejected the Platonic Ideas or Forms, accepting some of the arguments against them that Plato himself had devised in the *Parmenides*, he did not thereby reject the notion that the telos of philosophical enquiry is a wisdom which turns on what man can know of God. The magnificent panorama provided at the beginning of the *Metaphysics* as gloss on the claim that all men naturally desire to know rises to and culminates in the conception of wisdom as knowledge of all things in their ultimate or first causes.

For much of the twentieth century, Aristotelian studies have been conducted under the repressive pall of Werner Jaeger's evolutionary hypothesis. On this view, Aristotle began as an ardent Platonist for whom the really real lay beyond sensible reality. With maturity, however, came the sober Macedonian empiricism which trained its attention on the things of this world and eschewed all efforts to transcend it. As for the *Metaphysics*, Jaeger saw it as an amalgam of both theories. The passage just alluded to at the beginning of the work is ascribed to the Platonic phase. Other passages have a far more modest understanding of the range and point of a science over and above natural philosophy and mathematics. *Platonice loquendo*, there are entities which exist separately from sensible things and they constitute the object of the higher science. The more sober view finds a role for a science beyond natural philosophy and mathematics, but it will deal with things those particular sciences leave unattended, e.g. defense of the first principle of reasoning. But these tasks do not call for, and do not imply, a range of beings over and above sensible things.

Jaeger found both these conceptions of metaphysics clumsily juxtaposed in a crucial passage of Book Six.

> One might indeed raise the question whether first philosophy is universal, or deals with one genus, i.e. some one kind of being; for not even the mathematical sciences are all alike in this respect, -- geometry and astronomy deal with a certain particular kind of thing, while universal mathematics applies alike to all. We answer that if there is no substance other than those which are formed by nature, natural science will be the first science; but if there is an immovable substance, the science of this must be prior and must be first philosophy, and universal in this way, because it is first. And it will belong to this to consider being *qua* being -- both what it is and the attributes which belong to it *qua* being. (1025a24-33)

Jaeger invites us to see here a monument to a lost hope and an abiding reluctance to bid it a definitive farewell. Aristotle mentions the possibility of an immovable substance, something existing apart from the natural realm. Without such a separate substance, natural philosophy will be first philosophy. If there is such a substance, it will be a kind of being different from material being. The science that studies it will bear on a certain kind of being, immovable substance, immaterial being, not on being as being. It will be a special, not a universal, science. Jaeger sees Aristotle seeking to glue on to the special science the tasks that belong to a universal science, to make a theology into an ontology.

Jaeger's hypothesis, which cannot withstand half an hour's scrutiny, dominated interpretations of the *Metaphysics* until yesterday. Giovanni Reale's book had to await English translation before it could have any impact. By that time, people were turning from Jaeger and toward Aristotle neat, but this was only to weary of Jaeger, not to disprove him. Thomas's reading of the *Metaphysics* makes clear how mistaken Jaeger's claims are.

But let us first lay out Thomas's view of metaphysics. His question is Aristotle's: is there any science beyond natural science and mathematics? If to be and to be material are identical, then the science of being as being will be identical with the science of material being. That is what Aristotle rejects in the passage just quoted. It is in the course of doing natural philosophy that one gains certain knowledge that not everything that is is material. At the end of the *Physics*, Aristotle argues from the nature of moved movers that they require a first unmoved mover. If successful, this proof establishes that there is a first mover of all moved movers which is not itself material. Furthermore, the discussion of intellect in *On the Soul III* to which we alluded in the preceding paragraph, points beyond the material world. If the activity of intellect provides a basis for saying that, while the human soul is the substantial form of the body, it can exist apart from the body, that is, survive death, it is an immaterial existent. The Prime Mover and the immortal souls of human beings entail that to be and to be material are not identical. Since these are acquisitions at the limit of natural philosophy, they represent possible objects of inquiry in their own right. This is pre-eminently the case with the Prime Mover. It seems inevitable that there should be a discipline whose principal aim is to know more about the divine. How can it be described?

By common consent, Thomas's early discussion of the way theoretical sciences are distinguished from one another in the course of his exposition of the tractate of Boethius *On the Trinity* is masterful. The text speaks of three kinds of theoretical science, physics, mathematics and theology, and Thomas invokes the

methodology of the *Posterior Analytics*. A *scientia* is constituted by a demonstrative syllogism. From a formal point of view, a conclusion follows necessarily from the premises in a well-formed syllogism. Still the conclusion may state a merely contingent truth. What is needed in a demonstrative syllogism is not just the necessity of the consequence but a necessary consequent, and this requires that the premises express necessary truths. That which is necessary cannot be otherwise than as it is; it cannot change. Science thus requires that it bear on immobile things. There is another requirement of the object of speculative or theoretical knowledge which stems from intellection. The activity of the mind, as has been mentioned, is not a material event; it is immaterial. Since it is the mind that knows, science is a mode of its knowing, and will share its nature. Thomas thus states two essential characteristics of the object of speculation, the *speculabile*: it must be removed both from matter and from motion. If that is the case then insofar as there are formally different ways in which *speculabilia* can be removed from matter and motion, there will be formally different speculative sciences.

By this analysis, Thomas has provided the necessary background for understanding the text of Boethius but also more importantly that of Aristotle as it is developed in the chapter from which Werner Jaeger quoted in order to display the failure of the Aristotelian project. "Now we must not fail to notice the nature of the essence and of its formula, for, without this, inquiry is but idle. Of things defined, i.e. of essences, some are like snub, and some like concave. And these differ because snub is bound up with matter (for what is snub is a concave *nose*), while concavity is independent of perceptible matter." (1025a28-32) The objects of natural philosophy are defined like ‘snub’ and the objects of mathematics like ‘concave’. This makes it clear that the way in which natural things are separated from sensible matter is the way in which the definition common to many things abstracts from the singular characteristics of each. But it is the matter as singular that is the principle of change in things, so the common definition has the requisite necessity for science. This or that man comes to be, but what-it-is-to-be-a-man does not come to be or pass away.

Mathematical things, on the analogy of ‘concave’, do not have sensible matter in their definitions. Lines, points, numbers, triangles -- these do not have sensible qualities whether stated universally or singularly. The fact that we define mathematicals without sensible matter does not commit us to the view that mathematicals actually exist apart from sensible matter.

In the commentary on Boethius to which reference has been made, Thomas has early on recalled another fundamental aspect of Aristotle's thought. The objects of thought are either simple or complex, where complex means that one thing is affirmed or denied of another. Knowledge of simples is expressed in a definition, that of the complex in a proposition. Thinking of human nature without thinking of singular characters of this man or that is a matter of definition, not of assertion, as if one were denying that human nature is found in singular matter. So too defining mathematicals without sensible matter is not tantamount to the judgment that mathematicals exist apart from sensible matter. These are both instances of abstraction, where abstraction means to think apart what does not exist apart. Thus it is that the question of metaphysics turns on what Thomas calls *separatio*. To separate differs from abstraction in this that separation is expressed in a negative judgment, a proposition: this is not that, that *this* exists apart from *that*. The relevant separation for metaphysics is the negative judgment that to be and to be material are not the same. That is, there are things which exist apart from matter and motion -- not just are defined

without, but exist without matter and motion.

What then is the subject of metaphysics? "Subject" here means the subject of the conclusion of the demonstrative syllogism. The discussion of definition in effect bore on the middle terms of demonstrative syllogisms and the suggestion is that formally different modes of defining, with respect to removal from matter and motion, ground the formal difference between types of theoretical science. The subject of a demonstration in natural philosophy is defined without singular but with common or universal sensible matter; the subject of a mathematical demonstration is defined without any sensible matter. How can the subject of metaphysics be expressed? The possibility of the science depends on our knowing that some things exist apart from matter and motion. Mathematics does not presuppose the separate existence of its objects; metaphysics does. Why not then say that metaphysics deals with things separated from matter and motion, that is, as Jaeger notes, with a particular kind of being? But that is the not the subject ever assigned to this effort by Aristotle. The methodological reasons can be found in chapter 17 of Book Seven of the *Metaphysics*: the subject of a science must always be a complex entity. That is why the subject of this discipline is being as being.

Why should we say that, in our desire to learn more about separate substances, we should take as our subject all the things that are? The short answer is this: in order to be a theology, metaphysics must first be an ontology. Separate substance, divine being, is not directly accessible for our inspection or study. We come upon our first secure knowledge of God in the proof of the Prime Mover. Tantalizingly, once seen as a necessary requirement for there being any moved movers, the Prime Mover does not become a thematic object of inquiry in natural philosophy. One obvious reason for this is that such an entity is not an instance of the things which fall under the scope of the science. Knowledge of it comes about obliquely and indirectly. The same restriction is operative when the philosopher turns his culminating attention to the deity. How can he know more about the first cause of things? If the Prime Mover is known through moved movers as his effects, any further knowledge of him must be through his effects. It is by describing the effect as widely as possible that one seeks to come to a knowledge of the first cause unrestricted by the characteristics of mobile things. That characterization is being as being. The subject of metaphysics is being in all its amplitude in order to acquire a knowledge of the cause of being that will be correspondingly unbounded.

Philosophical and Scriptural Theology

Earlier we indicated the difference between philosophy and theology in the writings of St. Thomas. That distinction takes theology to mean discourse that takes its rise from the revealed truths of the Bible. But there is also a theology which constitutes the defining telos of philosophical inquiry. In the following passage, Thomas contrasts the two theologies in a way which throws light on what was said in the preceding paragraph.

> Thus it is that divine science or theology is of two kinds, one in which divine things are considered not as the subject of the science but as principles of the subject and this is the theology that the philosophers pursue, also called metaphysics. The other considers divine

> things in themselves as the subject of the science, and this is the theology which is treated in Sacred Scripture. They are both concerned with things which exist separately from matter and motion, but differently, insofar as they are two ways in which something can exist separately from matter and motion: first, such that it is of the definition of the things said to be separate, that they can never exist in matter in motion, as God and the angels are said to be separate from matter and motion; second, such that it is not part of their definition that they exist in matter and motion, because they can exist apart from matter and motion, although sometimes they are found in matter and motion, for example, substance, potency and act are separate from matter and motion because they do not require matter in order to exist as mathematicals do, although they can be understood without sensible matter. Philosophical theology treats of things separate in the second way as its subjects and of things separate in the first way as the principles of its subject. But the theology of Sacred Scripture treats of things separate in the first way as its subjects, although in it some things which exist in matter and motion are considered insofar as they are needed to make the divine manifest." (*Exposition of Boethius' On the Trinity*, q. 5, a. 4)

Philosophical theology is not some science distinct from metaphysics; it is simply the name that can be given to metaphysics because it appeals to God as the cause of its subject. This may make it seem that knowledge of God is merely a bonus, a tangential consideration; on the contrary, it is the chief aim of the science. But the divine can only be known indirectly, through its effects. For this reason, metaphysics can be viewed as an extended effort to examine substance in order to come to knowledge of the first cause. And given the principle that we name things as we know then, this can be regarded as a prolonged effort to devise a language with which to speak of God.

Analogous Names

Aristotle spoke of "things said in many ways", a notable instance of which is "being." One of the difficulties with assigning being or being as being as the subject of a science is that a subject must be univocally common to the things that fall under it and being is not univocal but has a plurality of meanings. Aristotle solved this problem with his account of "things said in many ways," by observing that while they have many meanings, these form an ordered set and one of the meanings is primary and regulative. Substance is being in the primary sense and that is why the science of being as being is effectively a science of substance. Thomas's term for such names is analogy: ‘being’ is an analogous term and its primary analogate is substance.

In the crucial middle books of the *Metaphysics* -- Seven and Eight -- we have an analysis of substance which takes off from material substance, which is a compound of matter and form, and arrives at a notion of substance as form alone. This definition does not fit material substance, of course, but it is devised in order to be able to apply the term substance to the immaterial things whose existence has been established in natural philosophy. This extension of names whose natural habitat is sensible things to God is another instance of analogous naming for Aquinas. Names common to God and creature bring out another feature of our knowing insisted upon by both Aristotle and Aquinas. If we ask what the primary analogate of names common to God and creature is, the answer is: the meaning of the term as it applies to creatures.

The word must be refined before it can be applied to God and this means the formation of an extended meaning which leans on the primary meaning for its intelligibility.

Consider the example of ‘wise.’ Both men and God are said to be wise. What can we mean when we say that God is wise? Not the same thing as when we say that Socrates is wise. Socrates became wise and wisdom is a trait which with age and forgetfulness he could lose. Thus to be Socrates and to be wise are not the same thing. But in the case of God, ‘wise’ does not signify some incidental property He might or might not have. This is captured by noting that while we say God is wise, we also say he *is* wisdom. This suffices to indicate the way in which the meaning of the term as applied to God involves negating features of its meaning as it applies to men.

If God is thus named secondarily by the common name, so that the creature is primarily named by it, nonetheless God's wisdom is the cause and source of human wisdom. Ontologically, God is primary and the creature secondary. Names analogously common to God and creature thus underscore the way in which what comes to be known first for us is not first in reality and what is first in reality is not first in our knowledge.

Essence and Existence

It is evident that material substances exist contingently. They come into being and they pass out of being and while they exist, existence is not what they are. Thomas accepts from Boethius that it is self-evident that what a thing is and for it to exist differ (*diversum est esse et id quod est*). Material things depend upon causes to exist, both to become and to be. There is no need to dwell on this except insofar as it provides a springboard to speak of immaterial substance. Only in God is it the case that what he is and that he is are identical: God is existence. The phrase Thomas uses to express this is *ipsum esse subsistens*. Of course this is paradoxical. Existence is the actuality of a substance, not itself something subsistent. This is true with material substances, but when we ask what we mean by saying that God exists, we have to negate aspects of material existence in order to avoid speaking of Him as if he were a contingent being.

The problem that Thomas now faces is how to speak of the immaterial substances which are less than God although superior to material substances, that is, angels. For a material thing to exist is for its form actually to inhere in its matter. But what is it for a pure form to exist? Since immaterial substances less than God are dependent on the divine causality in order to exist, existence cannot be what they are, of their essence. In short, in angels too there is a distinction of essence and existence. Thomas notes that a created separate substance is what it is and not another thing: that is, it has the perfection it has, but not unlimited perfection. It is a being of a kind, not being as such. Gabriel is perfect as to Gabrielitas, but he is not Raphael or Michael. Form thus operates as a restriction on existence as such. In God alone is there unrestricted existence; he is existence, *ipsum esse subsistens.*

Moral Doctrine

When Aristotle sought to isolate the human good, he employed the so-called function argument. If one know what a carpenter is or does he has the criteria for recognizing a good carpenter. So too with bank-tellers, golfers, brain surgeons and locksmiths. If then man as such has a function, we will have a basis for deciding whether someone is a good human being. But what could this function be? Just as we do not appraise carpenters on the basis of their golf game or golfers on the basis of their being able to pick locks, we will not want to appraise the human agent on an incidental basis. So too we do not appraise the carpenter in terms of his weight, the condition of his lungs or his taste buds. No more would we appraise a human being on the basis of activities he shares with non-humans. The activity that sets the human agent apart from all others is rational activity. The human agent acts knowingly and willingly. If this is the human function, the human being who performs it well will be a good person.

Many have come to this point, pulse quickened by the possibilities of the function-argument, only to be gripped with doubt at this final application of it. Rational activity seems too unmanageable a description to permit a function-analysis of it. Of course Aristotle agrees, having made the point himself. Rational activity is said in many ways or, as Thomas would put it, it is an analogous term. It covers an ordered set of instances. There is the activity of reason as such, there is the activity of reason in its directive or practical capacity, and there are bodily movements and the like which are rational insofar as they are directed by reason. If the virtue of a function is to perform it well, the analogy of "rational activity" makes clear that there is a plurality of virtues. Moral virtues are habits of appetite brought about by the direction of reason. Temperance is to seek pleasure rationally, courage is to react to the threat of harm rationally. The virtues of practical intellect are art and prudence; the virtues of theoretical intellect are insight, science and wisdom.

All this and much more enters into Thomas's moral teaching. Here as elsewhere the Aristotelian component looms large. Thomas will distinguish acts of a man from human acts, the former being activities truly found in human agents but also found in other non-human agents too. The human act is one which proceeds from knowledge and will. Since the human act by definition is the pursuit of a known good, the question arises as to the relationship between the objects of the myriad acts that human perform. Is there some over-all good sought by human agents? Is there an ultimate end of human action?

In commenting on chapter two of Book One of the *Nicomachean Ethics* where Aristotle argues for there being an ultimate end, Thomas points out that the argument is actually a series of *reductiones ad absurdum*. That is, the denial of an ultimate end of human action reduces to the claim that there is no end to human seeking, that it is pointless. This analysis has not gotten the attention it deserves: the implication is that it is self-evident that there is an ultimate end which is why denials of it must flounder in incoherence. The argument for ultimate end that Thomas puts forth in the *Summa theologiae* is somewhat different. Any action aims at some good. A particular good by definition shares in and is not identical with goodness itself. What binds together all the acts that humans perform is the overarching goodness they seek in this, that and the other thing. That over arching goodness, what Thomas calls the *ratio bonitatis*, is the ultimate end. It follows that anything a human agent does is done for the sake of the ultimate end.

This dissatisfies because we feel we are owed a richer account of goodness. After all, human agents differ

insofar as they have different notions of what goodness is. Fame, wealth, pleasure, power, and so on seem to function as the dominant purpose of different persons. Thomas could scarcely overlook this, let alone deny it. Can his earlier position on the unity of the ultimate end still stand? The fact that there are false or inadequate identifications of goodness does not mean that there is not a true and adequate account of what is perfecting or fulfilling of human agents. Everyone acts on the supposition that what he does will contribute to his overall good; one's overall good is the ultimate reason for doing anything. But not everything one does under this aegis actually contributes to one's overall good. Thus in one sense there is one and the same ultimate end for every human agent -- the integral human good -- and there are correct and mistaken notions of what actually constitutes this integral good.

This may seem like an empty claim, but it provides a basis on which to proceed. If indeed every human agent acts for the sake of his overall good, the discussion can turn to whether or not what he here and now pursues, or his general theory of what constitutes the overall good, can withstand scrutiny. It is not necessary to persuade anyone that he ought to pursue the ultimate end in the sense of his overall good. What else would he pursue? But if one is persuaded that what he pursues does not contribute to his overall good, he already has reasons for changing his ways.

Natural Law

Thomas's reading of Aristotle's argument for the ultimate end as a *reductio* and his own claim that in one sense of it everyone pursues the ultimate end since one chooses whatever he chooses *sub ratione boni* and as conducive to or a constituent of his fulfillment and perfection, tell us something important about Thomas's mode of procedure. We said earlier that philosophy begins from pre-philosophical principles already had by everyone. In the moral order, it is essential that one uncover the starting point, the latent assumption of any action, clarify it and proceed from there. This procedure is equally manifest in Thomas's treatment of what he calls natural law.

What is natural law? One description of it is: the peculiarly human participation in the eternal law, in providence. All creatures are ordered to an end, have natures whose fulfillment is what it is because of those natures. It is not peculiar to man that he is fashioned so as to find his good in the fulfillment of his nature. That is true of anything. But others things are ordered to ends of which they themselves are not conscious. It is peculiar to man that he becomes aware of the good and freely directs himself to it. Of course man is not free to choose the good -- any choice is a choice of the good. But as to what is really as opposed to only apparently his good, he is not free to make that what it is. He is free to direct himself or not to his true end, however.

A second description of natural law is: the first principles or starting points of practical reasoning. To indicate what he means by this, Thomas invokes the analogy of the starting points of reasoning as such. We have already mentioned the distinction between knowledge of the simple and knowledge of the complex. The former is a concept and is expressed in a definition or description. The latter is an affirmation or negation of one thing of another. There is something which is first in each of these orders. That is, Thomas holds that there is a conception which is prior to and presupposed by all other conceptions and a judgment that is prior to and presupposed by all other judgments. Since knowledge is

expressed by language, this seems to come down to the assertion that there is a first word that everyone utters and a first sentence that would appear in everyone's baby book on the appropriate page. But surely that is false. So what does Thomas mean?

He says that our first conception is of being, of that which is, and our first judgment is that you cannot affirm and deny the same thing in the same sense simultaneously. Since few if any humans first utter ‘being’ or its equivalent and no one fashions as his first enunciation the principle of contradiction, facts as known to Thomas as ourselves, his meaning must be more subtle. It is this. Whatever concept one first forms and expresses verbally -- Mama, hot, whatever -- is a specification or an instance of that which is. Aristotle has observed that children at first call all men father and all women mother. The terms then function as generic for any male or female. Even more basically, each presupposes that what is generically grasped is an instance of being. Being is prior not because it is grasped absolutely, without reference to this being or that. It is some particular being that is first of all grasped and however it is named it will mean minimally something that is.

So too with regard to the first judgment. Children express their recognition of this principle when they disagree over the location of some quite specific thing, say a baseball mitt. One accuses the other of taking it. You did. I didn't. You did. I didn't. A fundamental disagreement. But what they are agreed on is that if it were true that one did it could not simultaneously and in the same sense be true that he did not. The principle is latent in, implicit in, any concrete judgment just as being is involved in any other conception.

It is on an analogy with these starting points of thinking as such that Thomas develops what he means by natural law. In the practical order there is a first concept analogous to being in the theoretical order and it is the good. The good means what is sought as fulfilling of the seeker. The first practical judgment is: the good should be done and pursued and evil avoided. Any other practical judgment is a specification of this one and thus includes it. Natural Law consists of this first judgment and other most general ones that are beyond contest. These will be fashioned with reference to constituents of our complete good -- existence, food, drink, sex and family, society, desire to know. We have natural inclinations to such goods. Natural law precepts concerning them refer the objects of natural inclinations to our overall or integral good, which they specify.

Most moral judgments are true, if true, only by and large. They express means to achieve our overall good but because there is not a necessary connection between the means and end, they can hold only for the most part. Thus there are innumerable ways in which men lead their lives in keeping with the ultimate end. Not all means are necessarily related to the end. Moral philosophy reposes on natural law precepts as common presuppositions, but its advice will be true only in the main.

It might be noted that when Thomas, following Aristotle, says that man is by nature a social or political animal, he does not mean that each of us has a tendency to enter into social contracts or the like. The natural in this sense is what is not chosen, but given, and what is given about human life is that we are born into the community of the family and are dependent on it for years in order to survive. The moral consists in behaving well in this given setting.

Thomism

Thomas's teaching came under attack, largely by Franciscans, immediately after his death. Dominicans responded. This had the effect of making Dominicans Thomists and Franciscans non-Thomists -- Bonaventurians, Scotists, Ockhamists. The Jesuits were founded after the Reformation and they tended to be Thomists, often with a Suarezian twist.

When in 1879 Leo XIII issued the encyclical *Aeterni Patris* calling for the revival of the study of Thomas Aquinas, he was not directing his readers to one school as opposed to others. Thomas was put forward as the paladin of philosophy in its true sense, as over and against the vagaries of modern thought since Descartes. The response to Leo's call was global and sustained. New journals and learned societies were founded, curricula were reshaped to benefit from the thought of Thomas and this not simply in seminaries and pontifical universities but throughout the world in colleges and universities. Such giants as Jacques Maritain and Etienne Gilson may be taken to symbolize the best of this Thomistic revival.

Vatican II, the ecumenical council that met from 1962-1965 spelled the end of the Thomistic Revival. It was widely held that the Council had dethroned Thomas in favor of unnamed contemporary philosophers. (When they were named, quarrels began.) In the post-conciliar period, Catholics have managed to board every sinking ship in sight and now with the vogue of the notion that modernity has failed and the Enlightenment Project come a cropper, many, Catholics and others, are turning to Thomas as a spur or foil for their thinking. In 1998 John Paul II issued an encyclical called *Fides et Ratio* which may be regarded as the charter of the Thomism of the third millennium.

Bibliography

Life and Works

- Weisheipl, James A. *Thomas D'Aquino: His Life, Thought and Work.* Washington, Catholic University of America Press, 1974.
- Torrell, Jean-Pierre. *Initiation à saint Thomas d'Aquinas.* Paris: Editions Cerf, 1993. English translation, *Saint Thomas Aquinas*, Volume 1: *The Person and His Work*, by Robert Royal, Washington: Catholic University of America Press, 1996.
- Pieper, Josef. *Guide to Thomas Aquinas.* New York: Pantheon, 1962

Readers

- McDermott, Timothy. *Aquinas Selected Writings.* New York: Oxford University Press, 1993.
- McInerny, Ralph. *Thomas Aquinas Selected Writings.* London: Penguin Classics, 1998.

Introductions

- McInerny, Ralph. *A First Glance at Thomas Aquinas: Handbook for Peeping Thomists*. Notre Dame: University Press, 1990
- Copleston, F. C. *Aquinas*. London: Penguin Books, 1955.
- Bourke, Vernon J. *Aquinas Search for Wisdom*. Milwaukee: Bruce, 1965.
- Maritain, Jacques. *St. Thomas Aquinas*. New York: Meridian Books, 1964.

Other Internet Resources

- *Summa Theologiae* (English)
- *Summa Contra Gentiles* (English)
- Jacques Maritain's *St. Thomas Aquinas*
- *A Companion to the Summa* by Walter Farrell, O.P.
- *Catholic Encyclopedia* article on Saint Thomas Aquinas (1907).
- Thomas Instituut te Utrecht
- St. Thomas Aquinas Links
- Thomistic Philosophy Page
- Jacques Maritain Center at Notre Dame

Related Entries

Maritain, Jacques

Ralph McInerny
University of Notre Dame
McInerny.1@nd.edu

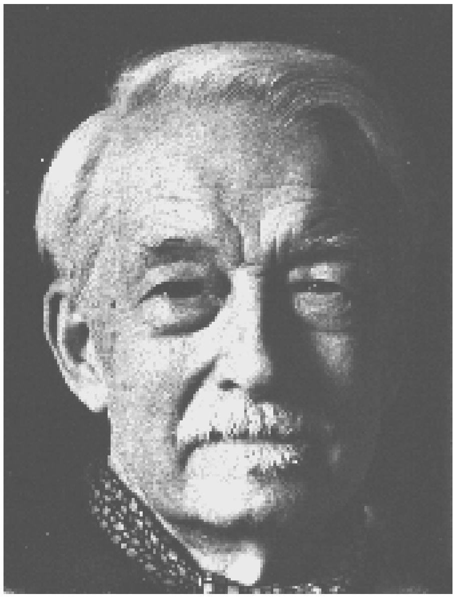
A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 12, 1999
Content last modified: July 12, 1999

Stanford Encyclopedia of Philosophy

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z



Jacques Maritain

Jacques Maritain (1882-1973), French philosopher and political thinker, was one of the principal exponents of Thomism in the twentieth century and an influential interpreter of the thought of St Thomas Aquinas.

Life

Jacques Maritain was born on November 18, 1882 in Paris. The son of Paul Maritain, a prominent lawyer, and Geneviève Favre, daughter of the French statesman, Jules Favre, Jacques Maritain studied at the Lycée Henri IV (1898-99) and at the Sorbonne, where he prepared a *licence* in philosophy (1900-1901) and in the natural sciences (1901-1902). He was initially attracted to the philosophy of Spinoza. Largely at the suggestion of his friend, the poet (and, later, religious thinker) Charles Péguy, he attended lectures by Henri Bergson at the Collège de France (1903-1904) and was briefly influenced by Bergson's work.

In 1901, Maritain met Raïssa Oumansoff, a fellow student at the Sorbonne and the daughter of Russian Jewish immigrants. Both were struck by the spiritual aridity of French intellectual life and made a vow to commit suicide within a year should they not find some answer to the apparent meaninglessness of life. Bergson's challenges to the then-dominant positivism sufficed to lead them to give up their thoughts of suicide, and Jacques and Raïssa married in 1904. Soon thereafter, through the influence of the writer Léon Bloy, both Maritains sought baptism in the Roman Catholic Church (1906).

Maritain received his *agrégation* in philosophy in 1905 and, late in 1906, Jacques and Raïssa left for Heidelberg, where Jacques continued his studies in the natural sciences. They returned to France in the summer of 1908, and it was at this time that the Maritains explicitly abandoned *bergsonisme* and Jacques began an intensive study of the writings of Thomas Aquinas.

In 1912, Maritain became professor of philosophy at the Lycée Stanislaus, though he undertook to give lectures at the Institut Catholique de Paris. He was named Assistant Professor at the Institut Catholique (attached to the Chair of the History of Modern Philosophy) in 1914. (He became full Professor in 1921 and, in 1928, was appointed to the Chair of Logic and Cosmology, which he held until 1939.)

In his early philosophical work (e.g., "La science moderne et la raison," 1910, and *La philosophie bergsonienne*, 1913), Maritain sought to defend Thomistic philosophy from its Bergsonian and secular opponents. Following brief service in the first world war, Maritain returned to teaching and research. The focus of his philosophical work continued to be the defense of Catholicism and Catholic thought (e.g., *Antimoderne* [1922], *Trois réformateurs -- Luther, Descartes, Rousseau* [1925], and *Clairvoyance de Rome par les auteurs du 'Pourquoi Rome a parlé' (J. Maritain et D. Lallement)* [1929]), but Maritain also prepared some introductory philosophical texts (e.g., *Éléments de philosophie* [2 volumes, 1921-23]) and his interests expanded to include aesthetics (e.g., *Art et scholastique*, 1921; 2nd ed. 1927).

By the late 1920s, Maritain's attention began to turn to social issues. Although he had some contact with the Catholic social action movement, Action Française, he abandoned it in 1926 when it was condemned by the Catholic Church for its nationalistic and anti-democratic tendencies. Still, encouraged through his friendships with the Russian philosopher Nicholas Berdiaev (beginning in 1924) and Emmanuel Mounier (from 1928), Maritain began to develop the principles of a liberal Christian humanism and defense of

natural rights.

Maritain's philosophical work during this time was eclectic, with the publication of books on Thomas Aquinas (1930), on religion and culture (1930), on Christian philosophy (1933), on Descartes (1932), on the philosophy of science and epistemology (*Distinguer pour unir ou les degrés du savoir*, 1932; 8th ed., 1963) and, perhaps most importantly, on political philosophy. Beginning in 1936, he produced a number of texts, including *Humanisme intégral* (1936), *De la justice politique* (1940), *Les droits de l'homme et la loi naturelle* (1942), *Christianisme et démocratie* (1943), *Principes d'une politique humaniste* (1944), *La personne et le bien commun* (1947), *Man and the State* (written in 1949, but published in 1951), and the posthumously published *La loi naturelle ou loi non-écrite* (lectures delivered in August 1950).

Maritain's ideas were especially influential in Latin America and, largely as a result of the ‘liberal’ character of his political philosophy, he increasingly came under attack from both the left and the right, in France and abroad. Lectures in Latin America in 1936 led to him being named as a corresponding member of the Brazilian Academy of Letters, but also to being the object of a campaign of vilification.

By the early 1930s Maritain was an established figure in Catholic thought. He was already a frequent visitor to North America and, since 1932, had come annually to the Institute of Mediaeval Studies in Toronto (Canada) to give courses of lectures. With the outbreak of war at the end of 1939, Maritain decided not to return to France. Following his lectures in Toronto at the beginning of 1940, he moved to the United States, teaching at Princeton University (1941-42) and Columbia (1941-44).

Maritain remained in the United States during the war, where he was active in the war effort (recording broadcasts destined for occupied France and contributing to the Voice of America). He also continued to lecture and publish on a wide range of subjects -- not only in political philosophy, but in aesthetics (e.g., *Art and Poetry*, 1943), philosophy of education, and metaphysics (*De Bergson à St Thomas d'Aquin*, 1944). Following the liberation of France in the summer of 1944, he was named French ambassador to the Vatican, serving until 1948, but was also actively involved in drafting the United Nations Universal Declaration of Human Rights (1948).

In the spring of 1948, Maritain returned to Princeton as Professor Emeritus, though he also lectured at a number of American universities (particularly at the University of Notre Dame and the University of Chicago), and frequently returned to France to give short courses in philosophy -- notably at ‘L'Eau vive,’ in the town of Soisy, near Paris. During this time, in addition to his work in political philosophy (cf. above, as well as *Le philosophe dans la cité*, 1960), Maritain published on aesthetics (*Creative Intuition in Art and Poetry*, 1953), religion (*Approches de Dieu*, 1953), moral philosophy (*Neuf leçons sur les notions premières de la philosophie morale*, 1951; *La philosophie morale*, 1960), and the philosophy of history (*On the Philosophy of History*, 1957).

In 1960, Maritain and his wife returned to France. Following Raïssa's death later that year, Maritain moved to Toulouse, where he decided to live with a religious order, the Little Brothers of Jesus. During this time he wrote a number of books, the best-known of which was *Le paysan de la Garonne* (a work

sharply critical of post-Vatican Council reforms), published in 1967. In 1970, he petitioned to join the order, and died in Toulouse on April 28, 1973. He is buried alongside Raïssa in Kolbsheim (Alsace) France.

General Background

Maritain saw himself as working in continuity with the thought of Thomas Aquinas, and his writings frequently contain quotations from and references to Thomas' texts. While his turn to Catholicism and his intellectual itinerary were largely due to personal reasons and to the influence of friends, his defense of Catholic thought and Thomistic philosophy were undoubtedly affected by events involving his adopted church.

One such event was the attack on (principally Catholic) religious organisations by secular and humanist forces within the French state, culminating in a number of laws affecting the taxation and ownership of church property and the place of religion in public affairs. At about the same time, there were tensions within Catholicism -- particularly in France -- in reaction to theological modernism. The writings of George Tyrell in England and Ernest Renan and Alfred Loisy in France were condemned for such ‘errors’ as claiming that conscience is the primary source of religious truth and that all knowledge -- including dogma -- has a historical and contingent character, and challenging the authoritative character of magisterial pronouncements. French philosophy itself was seen as incompatible with Catholic theology. The dominant views were the spiritualism or intuitionism of Bergson (which held that the emphasis in metaphysics on ‘being’ should be replaced by one on *durée* or pure change), the idealism of Léon Brunschvicg, the spiritualism of André Lalande, and the materialism of Edmond Goblot -- and each challenged claims that were seen as essential to Catholicism. The Catholic Church in France was, not surprisingly, in some turmoil, and a defense of religious orthodoxy was called for from several quarters.

Maritain's early writings, then, sought to address some of the concerns arising out of these events. Having been attracted initially to Spinoza's idealism and, later, to Bergson's vitalist intuitionism, he was able to come to the defense of Catholic thought with a knowledge of its critics that surpassed many of his contemporaries. Maritain rejected ‘modernity’ -- Cartesian and post-Cartesian thought -- for what he saw as its emphasis on epistemology over metaphysics, and sought to return to the ‘pre-modern’ views of Aquinas. Nevertheless, he saw that philosophy had to do more than merely repeat Thomas' views, and he took it upon himself to develop some aspects of Thomistic philosophy to address the problems of the contemporary world. Thus, though the most profound inspiration of many of Maritain's ideas was the work of St Thomas Aquinas, his epistemology and aesthetics show the influence of Christian mysticism, particularly that of St John of the Cross, and his social and political philosophy clearly reflects many of the ideals of European liberalism.

Principal Contributions

Epistemology

Maritain's primary work in epistemology is *Distinguer pour unir: ou, les degrès du savoir* [*Distinguish to Unite: or, The Degrees of Knowledge*] (1932), though one finds a number of important essays on the topic in *Raison et raisons, essais détachés* [*The Range of Reason*] (1948) and in *Quatre essais sur l'esprit* (1939). He largely follows the realist view of St Thomas Aquinas -- though he was also influenced by St John of the Cross and St Augustine, and the structure of *Les degrès du savoir* appears to reflect the procedure traced in the *Itinerarium mentis in Deum* [*The Journey of the Mind to God*] of St Bonaventure.

Against 'modern philosophy', Maritain insisted on the priority of metaphysics over epistemology -- in fact, he held that "the critique of knowledge is part of metaphysics" [*The Range of Reason*, p. 25] -- and also maintained that the structure and method of the various sciences were determined by the nature of the object to be known.

Maritain called his view critical realism, and argued particularly against the then-dominant rationalist and empiricist accounts of knowledge. He maintained that, despite the differences among them, Kantianism, idealism, pragmatism, and positivism all reflect the influence of nominalism -- that universal notions are creations of the human mind and have no foundation in reality. Maritain's critical realism holds that what the mind knows is identical with what exists. To know a thing is for its 'essence' to exist immaterially in the mind. This is not to say that the mind mirrors or copies that which it knows, but that, in virtue of the properties apprehended, it 'becomes' the things it knows. Maritain held that our knowledge of reality was through the 'concept' -- the *esse intentionale* -- which was immaterial and universal, though the concept itself was something known only by reflection. Thus, when it comes to knowledge of sensible objects, for example, the mind has both a passive role (receiving sense impressions) and an active one (constructing knowledge from these impressions).

Maritain's epistemology sought to explain not just the nature of knowledge found in science and philosophy, but religious faith and mysticism, and one of his aims was to show the different 'kinds' of knowledge and their relations to one another. He argued that there were different 'orders' of knowledge and, within them, different 'degrees' determined by the nature of the object to be known and the 'degree of abstraction' involved.

First, in the order of rational knowledge, one can speak of the knowledge of sensible nature (i.e., of the objects of experimental science), which is different from the knowledge of mathematics or of 'physico-mathematical' objects (which is limited because its objects do not have a direct relation to the actual), which is, in turn, distinct from the knowledge of trans-sensible or metaphysical nature.

These 'degrees of knowledge' are not, however, independent of one another, and they have in common the requirement that to know something is *to know why* it is -- "the mind is not satisfied when it merely attains a thing [...], but only when it grasps that upon which that datum is founded in being and intelligibility" (*Degrees of Knowledge*, p. 23). For example, natural science, which is based on sense perception, aims at formulating laws which reflect certain features of the objects perceived. The scientist, then, is primarily concerned with looking for regularities in nature and in pursuing the empiriological

method of engaging in observation, articulating a hypothesis, and then engaging in further testing; this Maritain calls *perinoetical* knowledge.

But for natural science to achieve the status of a science, it must presuppose natural philosophy -- that is, our capacity to know things apart from their particular individuating characteristics (though not apart from the existence of matter). Natural philosophy ‘gets behind’ phenomena in order to discover essential connections and causes. Thus, from that which is presented in sense perception, the mind constructs an object which is universal. (This is possible because, Maritain maintains, there are essences or natures of things.) This process of ‘thinking through’ to the nature of the thing is called by Maritain *dianoetical* knowledge. While natural science and natural philosophy both focus on the physical, the natural philosopher -- unlike the scientist -- is concerned with the essence of the object and its definition (or, at the very least, an account of its various properties). This, then, is knowledge at the level of the first ‘degree of abstraction.’

Physico-mathematical objects (e.g., quantity, number, and extension) stand at a second level of abstraction. While they cannot exist without the existence of material things, once known, they can be conceived of without any reference to such objects. Metaphysical or speculative knowledge deals with objects existing at a third level of abstraction (i.e., independently of matter), such as substance, quality, goodness, and the divine. Because of the nature of the objects of metaphysics, this latter kind of knowledge does not involve logical inference as much as reasoning by analogy or what Maritain calls *ananoetic* knowledge. Such knowledge (e.g., of the divine) is not through any direct apprehension, but indirectly, through creatures.

There is a hierarchy among these ‘degrees of knowledge.’ Those objects which are highest in intelligibility, immateriality, and potential to be known are the objects of the highest degree of knowledge. Maritain writes, "[t]he metaphysician considers an object of knowing of a specifically higher nature and intelligibility, and from it he acquires a proper knowledge, a scientific knowledge, by means that absolutely transcend those of the physicist or the mathematician" (*Degrees of Knowledge*, p. 37). Nevertheless, one should not conclude that there are different ‘knowledges.’

Maritain points out that philosophical demonstration is different from natural scientific or mathematical demonstrations: "philosophy is concerned with an objectively distinct field of knowledge and constitutes a really autonomous discipline, possessing its own adequate means of explaining this field of knowledge" (*Range of Reason*, p. 5). Specifically, Maritain writes, natural philosophy penetrates to the nature of its object. Metaphysics -- which is also a kind of philosophic knowing -- is concerned with purely intelligible being. Science, however, is at best ‘empiriological’ -- it does not lead us to being itself, but only to the observable and measurable. Thus, to employ a method of scientific demonstration to establish, or to criticize, claims about the object of metaphysical knowledge is, to use Ryle's classic term, a category mistake. And it is precisely because he holds that empiricist and Enlightenment epistemology do this that Maritain takes issue with them.

Just as there is an order of rational knowledge, with its ‘degrees,’ there are degrees of suprarational

knowledge -- of a higher wisdom -- that is beyond ‘natural knowledge.’ These are, on the one hand, ‘the science of revealed mysteries’ or ‘theological wisdom’ and, on the other, ‘mystical theology.’ Here, Maritain's debt to Augustine and John of the Cross is particularly evident. According to Maritain, in theological wisdom, the divine is known by drawing not just on reason but on faith. (This is distinct from metaphysical knowledge which, so to speak, approaches the divine from the ‘outside.’) Mystical knowledge stands one level higher still -- where there is no mediation by concepts -- and "consists in knowing [...] Deity as such -- *according to a mode that is suprahuman and supernatural*" (*The Degrees of Knowledge*, p. 253). This is a knowledge by connaturality, but also a knowledge that can be pursued through the practical discipline of ‘mystical contemplation.’ In this way, human beings acquire a kind of knowledge that makes them more loving and more spiritual.

A number of questions have been raised concerning Maritain's epistemology, particularly concerning his characterization of philosophical knowledge. For example, while Maritain suggests that there is a difference in method between the sciences and philosophy, it is not clear what exactly that difference is. For example, Maritain would follow Aquinas in holding that metaphysics uses *demonstratio quia* -- demonstration from effects. But it would seem that science also sometimes uses such a method of demonstration. Indeed, it is not clear what it is in the *method* (as distinct from the content of the premises) that differentiates a metaphysical proof (e.g., of God's existence) from a scientific argument establishing the existence of a cause of a natural object.

Second, Maritain holds that scientific knowledge is distinguished from philosophical knowledge in terms of their different methods and different objects. But if scientific knowledge and philosophical knowledge are, as it were, incommensurable, it is not clear how philosophy can judge, or be corrective of, the sciences.

Finally, it would seem that the model of demonstration that Maritain employs is foundationalist and, thus, has to answer to those criticisms that modern anti-foundationalism draws attention to -- e.g., that a foundationalist theory sets a standard for knowledge that is not only without justification, but is a standard that it cannot itself satisfy. Some recent defences of Thomistic epistemology (e.g., Henry Veatch in *Thomistic Papers*, Volume IV, 1990) suggest ways in which such concerns might be addressed.

Natural Theology and Philosophy of Religion

Like St Thomas Aquinas, Maritain held that there was no conflict between faith and true reason, that religious belief was open to rational discussion, and that the existence of God and certain fundamental religious beliefs could be philosophically demonstrated. Religious belief, then, was not an attitude or a matter of private opinion -- an option that could be embraced or not according to one's private preferences; it was a matter of ‘truth’. For Maritain, one must choose between "the true God or radical irrationality" (*Introduction to Philosophy*, p. 259).

Maritain held that philosophy was an *ancilla theologiae*, and that philosophy, under the rubric of

metaphysical knowledge, allows for the demonstration of a number of basic religious beliefs. And, like Aquinas, Maritain accepted the classical foundationalist position that these beliefs could be established by rational deduction from self-evident principles and constituted genuine knowledge. Specifically, he held that, by the use of natural reason, one can come to know certain truths about God, and that the ‘five ways’ of Thomas Aquinas provided sure knowledge of God's existence. But Maritain also argued that there could be other ‘proofs’ of the existence of the divine and, in *Approches de Dieu*, he developed what he called a "sixth way."

There is, Maritain writes, an intuition that is awakened in persons when they are engaged in thought -- that is, that it seems impossible that they, as thinking beings, should at some time have not been. As a thinking being, one seems to be free from the vicissitudes of time and space; there is no coming to be or ceasing to be -- I cannot think what it is not to be. Nevertheless, we all know very well that we were born -- we came into existence. We are confronted, then, with a contradiction -- not a logical contradiction, but a lived contradiction. The only solution to this is that one has always existed, but not through oneself, but within "a Being of transcendent personality" and from whom "the self which is thinking now proceeded into temporal existence" (*Approches de Dieu*, in *Oeuvres complètes*, p. 64). This being "must contain all things in itself in an eminent mode and be itself -- in an absolutely transcendent way -- being, thought and personality. This implies that the first existence is the infinite plenitude of being, separate by essence from all diversity of existents." (p. 66).

Maritain also acknowledges the possibility of a natural, pre-philosophical, but still rational knowledge of God (see *Approches de Dieu*, pp. 13-22). This is, Maritain claims, a ‘knowledge’ that is necessary to -- and, in fact leads to -- a philosophical demonstration of God's existence. (In this way, then, one can know that some religious beliefs are true, even without being able to demonstrate them.) Maritain's argument, which resembles the Thomistic argument from contingent being, is that, in one's intuition of being, one is aware, first, of a reality separate from oneself, second, of oneself as finite and limited, and, third, of the necessity that there is something "completely free from nothingness and death" (*Approches de Dieu*, p. 15). This is concurrent with a "spontaneous reasoning" that follows the same course to the conclusion that there is "another Whole [...] another Being, transcendent and self-sufficient and unknown in itself and activating all beings [...] that is, self-subsisting Being, Being existing through itself" (*Approches de Dieu*, p. 16). This ‘knowledge’ of God, Maritain admits, is not demonstrative but is, nevertheless, "rich in certitude" (*Approches de Dieu*, p. 19) and is both presupposed by, and is the underlying force for, philosophical demonstrations of God's existence.

The difference between the pre-philosophical and the philosophical ‘knowledge’ of God is that the latter is one which is based on a "scientific demonstration" (*Approches de Dieu*, p. 19) -- on empirical facts -- and involving analogy, from which we have terms that can be properly predicated of the divine. On the other hand, ‘pre-philosophical’ knowledge is "intuition" -- an *approach* to knowledge, though not a "way," (*Approches de Dieu*, p. 20) a proof or a demonstration. This knowledge is based on a natural reasoning which cannot be expressed in words. Yet, it is important also to realize that while Maritain allows that certain ‘truths’ "are grasped by the common sense before being the object of philosophical concern" (*Approches de Dieu*, p. 24), philosophical proofs of the existence of God "are not only established and justified philosophically at the level of philosophy itself, but are already valid and

efficacious at the level of this incohative and spontaneous philosophy," (*Approches de Dieu*, p. 24) and that what one arrives at by means of such an ‘approach’ is (as it is in philosophical demonstrations) knowledge of the truth of a proposition.

It has been argued, however, that there are some difficulties with Maritain's position here. For example, even if it is true that people may ‘naturally’ affirm the proposition that there is a God, it is not obvious how they can claim that they *know* it. In other words, even if the proposition is true, it is not clear how we can say that we know or believe it to be true. What Maritain seems to give us here is an explanation of how one arrives at a certain proposition and of one's certainty, but nothing more. But, since the state of certainty of an individual is not the same as the assertion that that person knows something to be true, it is not clear that the pre-philosophical approach provides one with an adequate basis to say that a religious belief is true, only that one is convinced of it. And, one might argue, parallel conclusions can be drawn if one examines the other ways that Maritain suggests will lead to a putative ‘knowledge’ of God.

(Interestingly, Maritain was a critic of a number of arguments proposed in defence of religious belief. He argued that such defenses fail because they do not recognise that there are different types of knowledge, that these different types are hierarchically arranged, and that the methods they employ are, by definition, unsuited to demonstrate certain things. Thus, Maritain holds that while ‘reason,’ as ‘intelligence moving in a progressive way towards its term, the real’, can attain knowledge of God by means of demonstration, if we take ‘reason’ to be a purely discursive method -- one which Maritain identifies with the "physical-mathematical sciences" and which he also calls "the ‘reason’ of rationalism" (*Antimoderne*, p. 64) -- it can know or say nothing at all about God. Because reason must be ordered to its object, reason (in this second sense) can neither demonstrate nor even encounter revealed truths.)

In addition to the possibility of the demonstration of the existence of God and of the coherence of the divine attributes, Maritain allows that there are a number of other ways in which one might come to ‘know’ religious beliefs to be true. Besides knowing God ‘naturally,’ there is a ‘non-conscious knowledge of God’ in the first act of human freedom (see *Range of Reason*, pp. 69-71), ‘connatural knowledge’ (which is typical of mystical experience), ‘abstract intuition’ (by which one knows ‘primary principles’ such as the laws of identity and of non-contradiction and the principle of causality), the "ways of the practical intellect" (*Approches de Dieu*, ch. IV) (i.e., through moral or aesthetic experience -- though these do not provide a strict demonstration) and, of course, divine revelation.

Nevertheless, Maritain also held that it was reasonable to believe even in the absence of such arguments or evidence. (To say that one can attain, by reason, some knowledge of God is not to say that everyone can do this.) Moreover, Maritain argues that even if one holds that a belief is capable of a rational demonstration, it does not follow that one must be able to provide it. For a religious belief to be ‘reasonable,’ it must not contradict the results of ‘true reason’ but, for one's knowledge of ‘revealed truths’ to be reasonable, Maritain (like Aquinas) would never claim that one must be able to produce a philosophical demonstration of them. In fact, Maritain allows that theology can "reject as false any philosophic affirmation which contradicts a theological truth" (*Introduction to Philosophy*, p. 126).

Maritain writes that there can also be knowledge of the divine attributes. As with all natural knowledge of the divine, this is basically analogical, and it follows a *via negativa*. Thus, he insists that we can say that we know some things about God. We can know that God is (*quia est*), though not what God is in himself (*quid est*) (*Degrees of Knowledge*, Appendix III, p. 423). In fact, against Sertillanges and Etienne Gilson, Maritain maintains that we can have affirmative knowledge about God -- know in a more or less imperfect but, nevertheless, true way what God is. Besides, Maritain holds, knowledge through negation presupposes positive knowledge. Mary Daly notes, however, that Maritain is not clear about the extent to which our affirmative knowledge of God is arrived at by means of philosophical argument (Daly, p. 53). Nevertheless, Maritain acknowledges that the knowledge of God that philosophy provides us with is incomplete and imperfect. The *analogical* knowledge that we have of God falls short of a complete description of what God is.

It is not clear, however, that Maritain avoids many of the concerns expressed by critics of 'analogical knowledge.' For example, if the term 'cause' is used analogously when applied to God, then when one utters the proposition 'God is the cause of the universe' after examining Aquinas's 'second way,' it would seem that one has to be using this term in exactly the same sense as it has been used throughout the preceding argument. If it is not being used in exactly the same sense, then how can one claim that Aquinas has *demonstrated* this conclusion? The problem is not whether analogical predication is possible but, first, whether one can understand the analogical predicate and, second, whether one can employ such a predicate in a demonstration without committing the fallacy of equivocation.

Given Maritain's account of faith and of suprarational knowledge, it would seem that he would see religious beliefs as 'trusts' and, hence, as having more than a purely cognitive character. He would, no doubt, follow Aquinas who spoke of religion as a 'disposition.' A disposition or *habitus* is, of course, not merely the product of action, but itself is ordered to action. Thus, to say that religious beliefs are propositional in form is not to say that their function is only descriptive. Nevertheless, Maritain's account of religious belief and its relation to argument and proof is not complete. Moreover, given that he does employ 'foundationalism' as a standard of sufficient evidence for claiming that some propositions expressing religious belief are true, it is not clear that it can directly address the challenges of recent critics -- particularly those raised by some 'postmodern' philosophers concerning the epistemology underlying his view.

Moral and Political Philosophy and Philosophy of Law

Maritain's moral and political philosophy lies within what may be called the Aristotelian-Thomistic natural law tradition. Maritain held, however, that Aristotelian ethics, by itself, was inadequate because it lacked knowledge of humanity's ultimate end. The Thomistic view -- that there is a law in human nature that is derivative of (though knowable separately from) a divine or eternal law and that humanity's 'end' goes beyond anything attainable in this life -- was, Maritain thought, a significant advance on what Aristotle had provided.

Following Aquinas, Maritain maintained that there is a natural law that is 'unwritten' but immanent in

nature. Specifically, given that nature has a teleological character, one can know what a thing ‘should’ do or how it ‘should’ be used by examining its ‘end’ and the ‘normality of its functioning.’ Maritain therefore defines ‘natural law’ as "an order or a disposition that the human reason may discover and according to which the human will must act to accord itself with the necessary ends of the human being" (*La loi naturelle*, p. 21; see *Man and the State*, p. 86). This law "prescribes our most fundamental duties" (*Man and the State*, p. 95) and is coextensive with morality.

There is, Maritain holds, a single natural law governing all beings with a human nature. The first principles of this law are known *connaturally*, not rationally or through concepts -- by an activity that Maritain, following Aquinas, called ‘synderesis.’ Thus, ‘natural ‘law’ is ‘natural’ because it not only reflects human nature, but is known naturally. Maritain acknowledges, however, that knowledge of the natural law varies throughout humanity and according to individuals’ capacities and abilities, and he speaks of growth in an individual's or a collectivity's moral awareness. This allows him to reply to the challenge that there cannot be any universal, natural law because no such law is known or respected universally. Again, though this law is progressively known, it is never known completely, and so the natural law is never exhausted in any particular articulation of it. This recognition of the historical element in human consciousness did not, however, prevent Maritain from holding that this law is objective and binding. (Critics have argued, however, that to speak of ‘connatural knowledge’ is obscure; it is quite unlike what we ordinarily call ‘knowledge’ and is, therefore, inadequate as a basis for knowledge of law.)

A key notion in Maritain's moral philosophy is that of human freedom. He says that the ‘end’ of humanity is to be free but, by ‘freedom,’ he does not mean license or pure rational autonomy, but the realisation of the human person in accord with his or her nature -- specifically, the achievement of moral and spiritual perfection. Maritain's moral philosophy, then, cannot be considered independently of his analysis of human nature. Maritain distinguishes between the human being as an individual and as a person. Human beings are ‘individuals’ who are related to a common, social order of which they are parts. But they are also persons. The person is a ‘whole’, is an object of dignity, "must be treated as an end" (*Les droits de l'homme*, p. 84) and has a transcendent destiny. In both the material and the spiritual order, however, human beings participate in a ‘common good.’ Thus, one is an individual in virtue of being a material being; one is a person so far as one is capable of intellectual activity and freedom. Still, while distinct, both elements are equally necessary to being a human being. It is in virtue of their individuality that human beings have obligations to the social order, but it is in virtue of their personality that they cannot be subordinated to that order. Maritain's emphasis on the value of the human person has been described as a form of *personalism*, which he saw as a *via media* between individualism and socialism.

Maritain's political philosophy and his philosophy of law are clearly related to his moral philosophy. The position that he defended was described by him in one of his earliest political works as ‘integral Christian Humanism’ -- ‘integral’, because it considers the human Being, an entity that has both material and spiritual dimensions, as a unified whole and because it sees human beings in society as participants in a common good. The object of Maritain's political philosophy was to outline the conditions necessary to making the individual more fully human in all respects. His integral humanism, then, seeks to bring the

different dimensions of the human person together, without ignoring or diminishing the value of either. While one's private good as an individual is subordinate to the (temporal) common good of the community, as a person with a supernatural end, one's ‘spiritual good’ is superior to society -- and this is something that all political communities should recognize.

For Maritain, the best political order is one which recognizes the sovereignty of God. He rejects, therefore, not only fascism and communism, but all secular humanisms. He objects that such views -- particularly fascism and communism -- are not only secular religions, but dehumanizing and, while he was a defender of American-style democracy, he is clearly not interested in combining his attachment to Christianity with capitalism. A theocentric humanism, Maritain would argue, has its philosophical foundation in the recognition of the nature of the human person as a spiritual and material being -- a being that has a relation to God -- and morality and social and political institutions must therefore reflect this.

Maritain envisages a political society under the rule of law -- and he distinguishes four types of law: the eternal, the natural, the ‘common law of civilisation’ (*droit des gens* or *jus gentium*), and the positive (*droit positif*).

The natural law is "universal and invariable" and deals with "the rights and duties which follow [necessarily] from the first principle" (see *Man and the State*, pp. 97-98) or precept of law -- that good is to be done and evil avoided. Nevertheless, while the natural law is "self-evident" (see *Man and the State*, p. 90) and consistent with and confirmed by experience -- something which many critics have challenged -- Maritain holds that it is not founded on human nature. It is rooted in divine reason and in a transcendent order (i.e., in the eternal law), and is ‘written into’ human nature by God. At times, Maritain appears to hold that natural law acquires its obligatory character only because of its relation to the eternal law; he writes that "natural law is law only because it is participation in Eternal Law" (see *Man and the State*, p. 96). (Some have concluded, then, that such a theory must be ultimately theological.)

The *droit des gens* or ‘common law of civilisation’ is an extension of the natural law to the circumstances of life in society, and thus it is concerned with human beings as social beings (e.g., as citizens or as members of families). The ‘positive law’ is the system of rules and regulations involved in assuring general order within a particular society. It varies according to the stage of social or economic development within that community and according to the specific activities of individuals within it. Neither the positive law nor the *droit des gens* is, however, deducible from the natural law alone. Neither is known connaturally and, therefore, is not part of the natural law. Nevertheless, it is in virtue of their relation to natural law that they "have the force of law and impose themselves on conscience" (*Les droits de l'homme*, pp. 90-91). When a positive law acts against the natural law, it is, strictly speaking, not a law. Thus, Maritain clearly rejects legal positivism.

The term ‘natural law’ and its relations both to ‘eternal law’ and to positive law have, however, been the focus of much controversy. Maritain's account of natural law both presupposes a metaphysical view of the nature of human beings and a realistic epistemology, and has a number of tensions or inconsistencies

internal to it. Some of the principal criticisms of this account are i) that it is inconsistent because it sets forth a naturalistic theory of what is good and bad and yet claims that only a supernatural sanction will serve to explain moral obligation, ii) that connatural knowledge is not only inadequate for what we normally count as knowledge, but it is, in fact, also incapable of establishing that something is a natural moral law, iii) that the first principle of moral law is vacuous, and iv) that Maritain glosses over the fact/value distinction.

Maritain held that natural law theory entailed an account of human rights. Since the natural end of each person is to achieve moral and spiritual perfection, it is necessary to have the means to do so, i.e., to have rights which, since they serve to realise his or her nature, are called ‘natural’. This respects the Aristotelian-Thomistic principle of justice, that we should distribute to each ‘what is truly his or hers’. Maritain replies to the criticism that there are no such rights, since they are not universally recognised, by reminding his readers that, just as the natural law comes to be recognised gradually and over time, so also is there a gradual recognition of rights. Indeed, Maritain held that certain basic natural rights can be recognised by all, without there having to be agreement on their foundation and, as an illustration of this, he pointed to the general agreement on those rights found in the 1948 United Nations Declaration of Human Rights.

Maritain held that natural rights are fundamental and inalienable, and antecedent in nature, and superior, to society. Still, they should not be understood as ‘antecedent’ in a temporal sense and do not form the basis of the state or of the civil law. Rights are grounded in the natural law, and specifically in relation to the common good. It is this good, and not individual rights, that is the basis of the state, and it is because of this that Maritain held that there can be a hierarchical ordering of these rights (*Man and the State*, p. 106-107).

One consequence of his natural law and natural rights theory is that Maritain favoured a democratic and liberal view of the state, and argued for a political society that is both personalist, pluralist, and Christianly-inspired. He held that the authority to rule derives from the people -- for people have a natural right to govern themselves. Still, this is consistent with a commitment to Christianity, Maritain thought, because the ideals of democracy are themselves inspired by a belief in God's rule, and that the primary source of all authority is God (*Man and the State*, p, 127).

Maritain also favoured a number of liberal ideals, and the list of rights that he recognises extends significantly beyond that found in many liberal theories, and includes the rights of workers as well as those of the human and the civic person.

Furthermore, the ideal of freedom or liberty to be found in the state is close to that which is now generally called ‘positive freedom’ -- i.e., one that reflects a view of the person as sharing in a common good. As a polity that attempts to provide the conditions for the realisation of the human person as an individual who is, thereby, a member of a temporal community, it recognises that the use of goods by individuals must serve the good of all (*Integral Humanism*, p. 184), and that individuals can be required to serve the community. Moreover, in such a state political leaders would be more than just

spokespersons for the people (*Man and the State*, p. 140), and Maritain recognises that they can represent the ‘hidden will’ of the people. Their aim -- and the aim of the state as a whole -- is, however, always the common good. Since minorities may themselves reflect this ‘hidden will,’ Maritain also recognised the important role to be played by dissenting minorities.

(Maritain does not discuss in any detail how his model ‘Christian’ polity might be realised, but suggests that it is the only one that takes account of each person's spiritual worth and that recognises the importance of providing the means to foster one's growth as a person. It recognises differences of religious conscience and is, in this way, fundamentally pluralistic.)

In such an ideal polity, Maritain imagines that a leadership role would be played by a multiplicity of ‘civic fraternities,’ founded on freedom, inspired by the virtues of Christianity, reflecting a moral and spiritual discipline, and which are fundamentally democratic. While such groups would not necessarily exercise political power, the society as a whole would reflect Christian values -- not just because these values are part of a privileged religion or faith (a matter that Maritain would be wary of), but because these are necessary to the temporal community. In such a polity one would, of course, find a church and a state, though Maritain would see them as cooperative entities, with the state occupying itself with those matters that, while focusing on temporal concerns, addressed the needs of the whole of the human person, and with the church focussing on spiritual matters.

It is, perhaps, evident that such a polity could not survive within a single nation state that existed among a plurality of states with different ideals, and so Maritain supported the ideal of a world federation of political societies. While the realisation of such an ideal was something that lay in a distant future, Maritain nevertheless thought that such a federation was possible, providing that the individual states retained a fair degree of autonomy and that persons could be found from each state who would voluntarily distance themselves from the particular interests of their home country.

General Assessment

At the time of his death, Maritain was arguably the best known Catholic philosopher in the world. The breadth of his philosophical work, his influence in the social philosophy of the Catholic Church, and his ardent defenses of human rights made him one of the central figures of his times.

Maritain's philosophical work has been translated into some twenty languages. As is evident from the preceding remarks, it covers a wide range of areas -- though much of it was written for a general, rather than a professional academic, audience. Still, some of Maritain's writings are polemical and, because much of his concern (especially in the history of philosophy) was to address very specific philosophical and theological issues of his time, they often have a rather dated character.

Maritain's most enduring legacy is undoubtedly his moral and political philosophy, and the influence of his work on human rights can be seen, not only in the United Nations Declaration of 1948 but, it has been

claimed, in a number of national declarations, such as the Canadian Charter of Rights and Freedoms and the preamble to the Constitution of the Fourth French Republic (1946) -- this last was likely a reflection of Maritain's lengthy correspondence with the French war hero and, later, President, General Charles DeGaulle. Maritain's Christian humanism and personalism have also had a significant influence in the social encyclicals of Pope Paul VI and in the thought of Pope John Paul II. Interestingly, since the end of the Cold War, there has been a revival of Maritain's political ideas in Central and Eastern Europe.

Two other areas in which Maritain's thought has been influential are his aesthetics and his philosophy of education. Although no longer as strong as they once were, they were particularly significant in Latin America and French-speaking Africa from the 1930s until recent years. Maritain's work in epistemology, though clearly essential to his political and religious thought and to his aesthetics, has not, however, had the reception Maritain would have held it deserved.

It is, in short, not easy to place Maritain's work within the history of philosophy in the 20th century. Clearly, his influence was strongest in those countries where Thomistic philosophy had pride of place. While his political philosophy led him, at least in his time, to be considered a liberal and even a social democrat, he eschewed socialism and, in *Le paysan de la Garonne*, was an early critic of many of the religious reforms that followed the Second Vatican Council. One can say, then, that he would be considered by present-day liberals as too conservative, and by many conservatives as too liberal. Again, though generally considered to be a Thomist, the extent to which he was is a matter of some debate. Indeed, according to Etienne Gilson, Maritain's ‘Thomism’ was really an epistemology and, hence, not a real Thomism at all. There is, not surprisingly, no generally shared view of the precise character of Maritain's philosophy.

Maritain's work nevertheless remains influential. Since 1958 there has been a Jacques Maritain Center at the University of Notre Dame in the United States, there are journals devoted to his work, such as *Études maritainiennes / Maritain Studies*, *Notes et documents*, and the *Cahiers Jacques et Raïssa Maritain*, and there are currently some twenty national associations which meet regularly, in addition to the *Institut International Jacques Maritain*. The continuity of interest in his thought in the English-speaking world has recently led the University of Notre Dame Press to undertake the publication of a *Collected Works* of the English-language editions of Maritain's writings.

Maritain's Principal Works

A French-language edition of the works of Maritain is available under the title *Oeuvres complètes de Jacques et Raïssa Maritain*, 15 vols., Fribourg (Switzerland): Éditions universitaires, 1982-. The publication of a 20 volume set, in English, of *The Collected Works of Jacques Maritain* (under the general editorship of Ralph McInerny) is currently underway under auspices of the University of Notre Dame Press.

Maritain's principal works are listed below in chronological order: (Except where indicated otherwise, place of publication is Paris. English-language editions are noted as well.)

- *La Philosophie bergsonienne: études critiques*. Marcel Rivière et Cie., 1914. [*Bergsonian Philosophy and Thomism*. New York: Philosophical Library, 1955.]
- *Art et scolastique*. Librairie de l'Art Catholique, 1920. (The 1927 edition contains "Frontières de la poésie" and important notes.) [*Art and Scholasticism and The Frontiers of Poetry*. Tr. Joseph W. Evans. New York: Charles Scribner's Sons, 1962.]
- *Éléments de Philosophie I: Introduction générale à la philosophie*. Téqui, 1920. [*An Introduction to Philosophy*. Tr. E. I. Watkin. London: Sheed and Ward, 1944.]
- *Théonas ou les entretiens d'un sage et deux philosophes sur diverses matières inégalement actuelles*. First publication in the *Revue Universelle*, April 1920 to April 1921, 1st edition, Nouvelle Librairie Nationale, 1921; 2nd edition, corrected, 1925. [*Theonas: Conversations of a Sage*. Tr. F.J. Sheed. London: Sheed and Ward, 1933.]
- *Antimoderne*. Éditions de la Revue des Jeunes, 1922.
- *De la vie d'oraison*. 1st edition privately printed, Saint-Maurice d'Augaune, 1922; 2nd edition revised, l'Art Catholique, 1925. [with Raïssa Maritain] [*Prayer and intelligence*. New York: P. J. Kennedy, 1928.]
- *Éléments de philosophie II: L'ordre des concepts, I - Petite logique (Logique formelle*). Téqui, 1923. [*An Introduction to Logic*. New York: Sheed and Ward, 1937; *Formal Logic*. New York: Sheed and Ward, 1937.]
- *Réflexions sur l'intelligence et sur sa vie propre*. Bibliothèque français de philosophie, Nouvelle Librairie Nationale, 1924, 1926, 1930; Desclée, 1938.
- *Trois Réformateurs: Luther, Descartes, Rousseau*. Librairie Plon, 1925. [*Three Reformers: Luther, Descartes, Rousseau*. New York: Charles Scribner's Sons, 1929.]
- *Georges Rouault, peintre et lithographe*. Éditions Polyglotte, Frapier, 1926. [*George Roualt*. New York: Harry N. Abrams, Inc., in association with Pocket Books, Inc., 1954.]
- *Reponse à Jean Cocteau*. Librairie Stock, 1926 [*Art and Faith: Letters between Jacques Maritain and Jean Cocteau*. New York: Philosophical Library, 1948.]
- *Une opinion sur Charles Maurras et le devoir des catholiques*. Plon, 1926.
- *Primauté du spirituel*. Plon, 1927. [*The Things That Are Not Caesar's*. Tr. J.F. Scanlan. New York: Charles Scribner's Sons, 1930.]
- *Quelques pages sur Léon Bloy*. *Cahiers de la Quinzaine*, 10e de la 18 serie, á l'Action du Livre, 1927.
- *Clairvoyance de Rome par les auteurs du 'Pourquoi Rome a parlé.' (J. Maritain et D. Lallement)*, Ed. Spes, 1929.
- *Le Docteur angelique*. Desclée de Brouwer, 1930. [*St. Thomas Aquinas*. Tr. F.J. Scanlan, London: Sheed and Ward, 1931; Tr. revised Peter O' Reilly and Joseph W. Evans. New York: Meridian Books, 1958.]
- *Religion et culture. dition originale: premier numéro de la collection des questions disputées*. Desclée de Brouwer, 1930. 2nd edition, with a preface, 1946. [*Religion and Culture*. London: Sheed and Ward, 1931.]
- *Distinguer pour unir: ou, les degrès du savoir*. Desclée de Brouwer, 1932. [*Distinguish to Unite: or, The Degrees of Knowledge*. Tr. under the supervision of G.B. Phelan. New York: Charles Scribner's Sons, 1959.]

- *Le songe de Descartes*. Correa, 1932. [*The Dream of Descartes*. Tr. Mabelle L. Andison, New York: Philosophical Library, 1944.]
- *Du regime temporel et de la liberté*. Desclée de Brouwer, 1933. [*Freedom in the modern world*. Tr. Richard O'Sullivan. London: Sheed & Ward, 1935.]
- *De la philosophie chrétienne*. ["Questions disputées," Vol. IX], Desclée de Brouwer, 1933.[*An essay on Christian philosophy*. Tr. Edward H. Flannery. New York, Philosophical Library, 1955.]
- *Sept leçons sur l'être et les premiers principes de la raison spéculative.* Téqui, 1934. [*A Preface to Metaphysics: Seven Lectures on Being*. New York and London: Sheed and Ward, 1939.]
- *Frontières de la poésie*. Louis Rouart et Fils, 1935. [*Art and Poetry.* Tr. E. de P. Matthews. New York: Philosophical Library, 1943.]
- *Science et sagesse, suivi d'éclaircissements sur ses frontières et son objet.* "Cours et documents de Philosophie." Téqui, 1935. [*Science and Wisdom.* New York: Charles Scribner's Sons, 1940.]
- *Lettre sur l'indépendance.* Desclée de Brouwer, 1935.
- *La philosophie de la nature, essai critique sur ses frontières et son objet.* Téqui, 1935. [*Philosophy of Nature.* Tr. Imelda C. Byrne. New York: Philosophical Library, 1951]
- *Humanisme intégral: problemes temporels et spirituals d'une nouvelle chrétienté.* Fernand Aubier, 1936. Two translations: *True humanism*. Tr. M.R. Adamson. London: Bles, 1938; *Integral Humanism: Temporal and Spiritual Problems of a New Christendom*. Tr. Joseph W. Evans New York: Charles Scribner's Sons, 1968.]
- *Situation de la poésie*. Desclée de Brouwer, 1938. [with Raïssa Maritain] [*The Situation of Poetry*. Tr. Marshall Suther. New York: Philosophical Library, 1955.]
- *Les Juifs parmi les nations.* Éditions du Cerf, 1938. [*A Christian Look at the Jewish Question*. New York: Longmans, Green, 1939.]
- *Questions de conscience.* ["Questions disputées," Vol. XXI] 2nd edition. Desclée de Brouwer, 1938.
- *Quatre essais sur l'esprit dans sa condition charnelle.* Bibliothèque français de philosophie, Desclée de Brouwer, 1939. Nouvelle Édition revue, Alsatia, 1956.
- *Antisemitism.* London, G. Bles, 1939.
- *De la justice politique, notes sur la presente guerre.* Collection "Presences," Plon, 1940.
- *Scholasticism and Politics*. New York, The Macmillan Company, 1940. [with Mortimer Jerome Adler]
- *A travers le désastre*. New York: Éditions de la Maison française, 1941 [*France, My Country, Through the Disaster*. New York: Longmans, Green, 1941.]
- *La pensée de Saint Paul, textes choisis et présentés.* New York: Éditions de la Maison française, 1941. [*The Living Thoughts of Saint Paul*. Tr. Harry Lorin Binsse. New York: Longmans, Green, 1941.]
- *Ransoming the Time*. New York: Scribner's, 1941. Tr. Harry Lorin Binsse. [A translation of essays written in French, but not collected in a single volume.]
- *Confession de foi,* New York: Éditions de la Maison française, 1941. [French translation of an essay published in *I Believe*. Ed. Clifton Fadiman. London: Allen and Unwin, 1940.]
- *Le Crépuscule de la civilisation.* Montréal: Éditions de l'Arbre, 1941. [*The Twilight of Civilization.* London: Sheed and Ward, 1946.]
- *Les droits de l'homme et la loi naturelle*. New York: Éditions de la Maison française, 1942. [*The*

Rights of Man and Natural Law. Tr. Doris C. Anson. New York: Charles Scribner's Sons, 1943.]

- *Saint Thomas and the problem of evil*. Tr. Mrs. Gordon Andison. Milwaukee: Marquette University Press, 1942.
- *Christianisme et Démocratie*. New York: Éditions de la Maison française, 1943. [*Christianity and Democracy*. Tr. Doris C. Anson. New York: Charles Scribner's Sons, 1944.]
- *Sort de l'homme*, Neuchâtel: Éditions de La Baconnière, 1943.
- *Education at the Crossroads.* New Haven: Yale University Press, 1943. [*L'éducation à la croisée des chemins*. Egloff, 1947; republished, with additional material, as *Pour une philosophie de l'education.* Arthème Fayard, 1959. Nouvelle édition, 1969.]
- *Principes d'une politique humaniste*. New York: Éditions de la maison française, 1944.
- *A travers la victoire*. Hartmann, 1945.
- *Messages 1941-1944*. New York: Éditions de la Maison française; Hartmann, 1945.
- *Pour la justice*. New York: Éditions de la Maison française, 1945.
- *Court traité de l'existence et de l'existant.* Hartmann, 1947. [*Existence and the Existent*. Tr. Lewis Galantière and Gerald B. Phelan. New York: Pantheon Books, 1948.]
- *La personne et le bien commun*. Desclée de Brouwer, 1947. [*The Person and the Common Good.* Tr. John J. Fitzgerald. New York: Charles Scribner's Sons, 1947.]
- *De Bergson à Thomas d'Aquin, essais de métaphysique et de morale.* New York: Éditions de la Maison française, 1944. Hartmann, 1947.
- *Art and faith*. New York, Philosophical Library, 1948. [see *Réponse a Jean Cocteau*, 1933.]
- *Raison et raisons, essais détachés.* Egloff, 1948. [*The Range of Reason*. New York: Charles Scribner's Sons, 1952.]
- *La signification de l'athéisme contemporain.* Collection "Courier de Iles." Desclée de Brouwer, 1949.
- *Man's destiny in eternity*. Boston: Beacon Pr., 1949. [with Arthur H. Compton, Maude Boyden et al.]
- *Etienne Gilson, philosophe de la chrétienté*. Cerf, 1949.
- *Man and the State*. Chicago: University of Chicago Press, 1951. [*L'Homme et L'État.* Tr. into French by Robert and France Duval. Presses Universitaires de France, 1953.]
- *Lettres inédites sur l'inquiétude moderne*. Éditions universelles, 1951. [with René Schwob]
- *Neuf leçons sur les notions premières de la philosophie morale.* "Cours et documents de philosophie." Téqui, 1951. [*An Introduction to Basic Problems of Moral Philosophy.* Albany, N.Y.: Magi Books, 1990.]
- *Approches de Dieu.* Collection "Sagesse et cultures." Alsatia, 1953. [*Approaches to God.* New York: Harper and Brothers, 1954.]
- *Creative Intuition in Art and Poetry.* New York: Pantheon Books, 1953. [*L'Intuition créatrice dans l'art et dans la poésie.* Desclée de Brouwer, 1966.]
- *On the Philosophy of History*. New York: Charles Scribner's Sons, 1957. [*Pour une philosophie de l'histoire*. Tr. Mgr Charles Journet. Éditions du Seuil, 1959.]
- *Reflections on America.* New York: Charles Scribner's Sons, 1958. [*Réflexions sur l'Amérique.* Fayard, 1959.]
- *Pour une philosophie de l'éducation*. Fayard, 1959.
- *Liturgy and contemplation*. London: G. Chapman, 1960. [with Raïssa Maritain]

- *The Responsibility of the Artist.* New York: Charles Scribner's Sons,1960. [*La responsabilité de l'artiste*. Tr. Georges and Christianne Brazzola, Fayard, 1961.]
- *Le philosophe dans la cité.* Alsatia, 1960.
- *La philosophie morale. I. Examen historique et critique des grands systemes.* Gallimard, Bibliothèque de Idées, 1960. [*Moral Philosophy*. Ed. Joseph W. Evans. London: G. Bles, 1964.]
- *On the use of philosophy; three essays.* Princeton, N.J.: Princeton University Press, 1961.
- *The Education of Man.* Ed. Donald and Idella Gallagher. New York: Doubleday and Co., 1962.
- *Dieu et la permission du mal.* Desclée de Brouwer, 1963. [*God and the Permission of Evil.* Milwaukee: The Bruce Publishing Co., 1966.]
- *Carnet de notes.* Desclée de Brouwer, 1965. [*Notebooks*. Tr. Joseph W. Evans. Albany: Magi Books/ Notre Dame, IN: University of Notre Dame Press, 1984.]
- *Le mystère d'Israël et autres essais*. Desclée de Brouwer, 1965.
- *Le paysan de la Garonne: Une vieux laïc s'interroge à propos du temps présent.* Desclée de Brouwer, 1967. [*The Peasant of the Garonne: An Old Layman Questions Himself about the Present Time.* Tr. Micheal Cuddihy and Elizabeth Hughes. New York: Holt, Rinehart, and Winston, 1968.]
- *De la grâce et de l'humanité de Jésus*, Bruges: Desclée de Brouwer, 1967. [*On the Grace and Humanity of Jesus*. Tr. Joseph W. Evans. New York: Herder and Herder, 1969.]
- *De l'église du Christ, la personne de l'Église et son personnel.* Desclée de Brouwer, 1970. [*On the Church of Christ: The Person of the Church and Her Personnel.* Tr. Joseph W. Evans. Notre Dame, IN: University of Notre Dame Press, 1973.]
- *Approches sans entraves.* Librairie Arthème Fayard, 1973.
- *Jacques Maritain, Emmanuel Mounier (1929-1939): [Correspondance]*. Desclée de Brouwer, 1973.
- *Une grande amitié: correspondance, 1926-1972 / Julien Green et Jacques Maritain.* Présentée et annotée par Jean-Pierre Piriou. Précédée de *Jacques Maritain vivant* de Julien Green. Plon, 1979. [*The story of two souls: the correspondence of Jacques Maritain and Julien Green.* Ed. Henry Bars and Eric Jourdan. Tr. with an introduction and revised notes by Bernard Doering. New York: Fordham University Press, 1988.]
- *La loi naturelle ou loi non écrite: texte inédit, établi par Georges Brazzola*. Fribourg, Suisse: Éditions universitaires, 1986. [*Lectures on Natural Law*. Tr. William Sweet. In *The Collected Works of Jacques Maritain*, Vol. VI, Notre Dame, IN: University of Notre Dame Press, (forthcoming).]
- *Exiles and fugitives: the letters of Jacques and Raïssa Maritain, Allen Tate, and Caroline Gordon.* Ed. John M. Dunaway. Baton Rouge: Louisiana State University Press, 1992.
- *L'Europe et l'idée fédérale:* textes publiés par le Cercle d'études Jacques et Raïssa Maritain. Mame, 1993.

Bibliography

The most comprehensive list of studies of Maritain's work is found in Jean-Louis Allard and Pierre Germain, *Répertoire bibliographique sur la vie et l'oeuvre de Jacques et Raïssa Maritain*. Ottawa, 1994,

232 p.

Major Sources

- Allard, Jean-Louis. *L'éducation à la liberté ou la philosophie de l'éducation de Jacques Maritain.* Ottawa: Éditions de l'Université d'Ottawa, 1978.
- Allard, Jean-Louis. *Education for Freedom: The Philosophy of Education of Jacques Maritain.* Notre Dame, IN: University of Notre Dame Press, 1982.
- Allard, Jean-Louis. *Jacques Maritain, Philosophe dans la Cité/ A Philosopher in the World.* Ottawa: University of Ottawa Press, 1985.
- American Maritain Association. *Selected Papers from the Conference-Seminar on Jacques Maritain's The Degrees of Knowledge*. St. Louis, MO: American Maritain Association, 1981.
- Bars, Henry. *Maritain en notre temps.* Paris: Bernard Grasset, 1959.
- Croteau, Jacques. *Les fondements thomistes du personnalisme de Maritain.* Ottawa: Éditions de l'Université d'Ottawa, 1950.
- Daly, Mary F. *Natural Knowledge of God in the Philosophy of Jacques Maritain*. Rome: Officium Libri Catholici-Catholic Book Agency, 1966.
- Daujat, Jean. *Maritain: Un maître pour notre temps*. Paris: Téqui, 1978.
- DiJoseph, John. *Jacques Maritain and the Moral Foundation of Democracy*. Lanham, MD: Rowman & Littlefield, 1996.
- Doering, Bernard. *Jacques Maritain and the French Catholic Intellectuals.* Notre Dame, IN: University of Notre Dame Press, 1983.
- Dunaway, John M. *Jacques Maritain.* Boston: Twayne Publishers, 1978.
- Eco, Umberto. *Storiografia Medievale ed Estetica Teorica Appunti Metodologici su Jacques Maritain.* Turin: Edizioni di "Filosofia," 1961.
- Evans, Joseph W. *Jacques Maritain (1882-1973): A Biographical Memoir*. Washington, D.C.: National Academy of Education, 1973.
- Evans, Joseph W., ed. *Jacques Maritain: The Man and His Achievement*. New York: Sheed and Ward, 1963.
- Fecher, Charles A. *The Philosophy of Jacques Maritain.* Westminster, MD: Newman Press, 1953.
- Floucat, Yves. *Pour une philosophie chrétienne: éléments d'un débat fondamental*. Paris: Téqui, 1983.
- Gallagher, Donald and Idella. *The Achievement of Jacques and Raissa Maritain: A Bibliography.* New York: Doubleday and Co., 1962.
- Hubert, Bernard and Yves Floucat, eds. *Jacques Maritain et ses contemporains.* Paris: Desclée, 1991.
- Hudson, Deal W. and Matthew J. Mancini, eds. *Understanding Maritain: Philosopher and Friend.* Macon, GA: Mercer Univ. Press, 1987.
- Institut International Jacques Maritain. International Jacques Maritain Institute. *Droits des peuples, Droits de l'homme.* Paris: Éditions du Centurion, 1984.
- Jimenez Berguecio, Julio, S.J. *La ortodoxia de Jacques Maritain, ante un ataque recente.* Talca, Chile: Libreria Cervantes, 1948.
- Jung, Hwa Yol. *The Foundation of Jacques Maritain's Political Philosophy.* Gainesville, FL:

University of Florida Press, 1960.

- Knasas, John F. X., ed. *Jacques Maritain: The Man and His Metaphysics.* [Volume IV of *Études maritainiennes/Maritain Studies*.] Mishawaka, IN: American Maritain Association, 1988.
- McInerny, Ralph. *Art and Prudence: Studies in the Thought of Jacques Maritain.* Notre Dame, IN: University of Notre Dame Press, 1988.
- Michener, Norah Willis. *Maritain on the Nature of Man in a Christian Democracy*. Hull (Canada): Éditions "L'Eclair," 1955.
- Minkiel, Stephen J., C.M., ed. *Jacques Maritain: The Man for Our Times.* Erie, PA: Gannon University Press, 1981.
- National Academy of Education. *Proceedings of the National Academy of Education*. "Jacques Maritain, A Biographical Memoir," Joseph W. Evans. Washington, D.C.: National Academy of Education, 1978, pp. 92-127.
- Nielsen, Kai. "An Examination of the Thomistic Theory of Natural [Moral] Law." In *Natural Law Forum* 4 (1959), 44-71. Reprinted in his *God and the Grounding of Morality*. Ottawa, ON: University of Ottawa Press, 1991, ch. 3, pp. 41-68.
- Nottingham, William J. *Christian Faith and Secular Action: An Introduction to the Life and Thought of Jacques Maritain.* St. Louis, MO: The Bethany Press, 1968.
- Papini, Roberto, ed. *L'Apporto del Personalismo alla Costruzione dell' Europa.* Milan: Massimo, 1981.
- Papini, Roberto, ed. *Jacques Maritain e la Società Contemporanea.* Milan: Massimo, 1978.
- Possenti, Vittorio. *Una Filosofia per la Transizione.* Milan: Massimo, 1984.
- Possenti, Vittorio, ed. *Jacques Maritain: Oggi.* Milan: Vita e Pensiero, 1983.
- Possenti, Vittorio, ed. *Maritain e Marx*. Milan: Massimo, 1978.
- Possenti, Vittorio. "Philosophie du droit et loi naturelle selon Jacques Maritain." In *Jacques Maritain: philosophe dans la cité / a philosopher in the world*. Ed. Jean-Louis Allard. Ottawa: University of Ottawa Press, 1985, pp. 313-326.
- Prouvost, Géry. *Catholicité de l'intelligence métaphysique: La philosophie dans la foi selon Jacques Maritain.* Paris: Pierre Téqui, 1991.
- Prouvost, Géry. *Étienne Gilson-Jacques Maritain: Correspondance 1923-1971*. Paris: Librairie Philosophique J. Vrin, 1991.
- Ramsey, Paul. *Nine Modern Moralists.* Englewood Cliffs, NJ: Prentice-Hall, 1962.
- Redpath, Peter A., ed. *From Twilight to Dawn: The Cultural Vision of Jacques Maritain.* Mishawaka, IN: American Maritain Association, 1990.
- Torre, Michael D., ed. *Freedom in the Modern World: Jacques Maritain, Yves R. Simon, Mortimer J. Adler*. Mishawaka, IN: American Maritain Association, 1990.

Collections

- *Jacques Maritain: Son oeuvre philosophique.* Paris: Bibliothèque de la Revue Thomiste, 1948.
- *The Review of Politics*. "Maritain Centenary." Vol. 44, No. 4, October, 1982.
- The Maritain Volume of *The Thomist*. New York: Sheed and Ward, 1943.

Additional studies of Maritain's work are available in such journals as *Études maritainiennes-Maritain Studies*, *Cahiers Jacques et Raïssa Maritain*, and *Notes et documents: pour une recherche personnaliste*.

Other Internet Resources

- The Jacques Maritain Center, University of Notre Dame
- Resources on Jacques Maritain (Malsapina College, British Columbia)
- The Canadian Jacques Maritain Association

Related Entries

Aquinas, Saint Thomas | God | liberalism | religion: philosophy of | theism

William Sweet
wsweet@stfx.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 5, 1997
Content last modified: September 13, 1998

Liberalism

Liberalism can be understood as (1) a political tradition (2) a political philosophy and (3) a general philosophical theory, encompassing a theory of value, a conception of the person and a moral theory as well as a political philosophy. As a *political tradition* liberalism has varied in different countries. In England --- in many ways the birthplace of liberalism --- the liberal tradition in politics has centred on religious toleration, government by consent, personal and, especially, economic freedom. In France liberalism has been more closely associated with secularism and democracy. In the United States liberals often combine a devotion to personal liberty with an antipathy to capitalism, while the liberalism of Australia tends to be much more sympathetic to capitalism but often less enthusiastic about civil liberties. To understand this diversity in political traditions, we need to examine liberalism as a *political theory* and as a *general philosophy*. These latter two are the concerns of this essay.

Liberalism as a Political Theory

Liberty

'By definition', Maurice Cranston rightly pointed out, 'a liberal is a man who believes in liberty' (Cranston, 459). In two different ways, liberals accord liberty primacy as a political value. First, liberals have typically maintained that humans are naturally in 'a *State of perfect Freedom* to order their Actions...as they think fit...without asking leave, or depending on the Will of any other Man' (Locke, 1960 [1689]: 287). Mill too argued that '[T]he burthen of proof is supposed to ith those who are against liberty; who contend for any restriction or prohibition.... The *a priori* assumption is in favour of freedom...'(Mill, 1991 [1859]: 472). This might be called the **Fundamental Liberal Principle** (Gaus, 1996: 162-166): freedom is normatively basic, and so the onus of justification is on those who would limit freedom. It follows from this that political authority and law must be justified, as they limit the liberty of

citizens. Consequently, a central question of liberal political theory is whether political authority can be justified, and if so, how. It is for this reason that social contract theory, as developed Thomas Hobbes (1948 [1651]), John Locke (1960 [1689]), Jean-Jacques Rousseau (1973 [1762]) and Immanuel Kant (1965 [1797]), is usually viewed as liberal even though the actual political prescriptions of, say, Hobbes and Rousseau, have distinctly illiberal features. Insofar as they take as their starting point a state of nature in which humans are free and equal, and so argue that any limitation of this freedom and equality stands in need of justification (i.e., by the social contract), the contractual tradition expresses the Fundamental Liberal Principle.

The Fundamental Liberal Principle holds that restrictions on liberty must be justified, and because he accepts this, we can understand Hobbes as espousing a liberal political theory. But Hobbes is at best a qualified liberal, for he also argues that drastic limitations on liberty *can be* justified. Paradigmatic liberals such as Locke not only advocate the Fundamental Liberal Principle, but also maintain that justified limitations on liberty are fairly modest. Only a limited government can be justified; indeed, the basic task of government is to protect the equal liberty of citizens. Thus John Rawls's first principle of justice: ‘Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system for all’ (Rawls, 1971: 302).

Negative and Positive Liberty

Liberals disagree, however, about the concept of liberty, and as a result the liberal ideal of protecting individual liberty can lead to very different conceptions of the task of government. As is well-known, Isaiah Berlin has advocated a negative conception of liberty:

> I am normally said to be free to the degree to which no man or body of men interferes with my activity. Political liberty in this sense is simply the area within which a man can act unobstructed by others. If I am prevented by others from doing what I could otherwise do, I am to that degree unfree; and if this area is contracted by other men beyond a certain minimum, I can be described as being coerced, or, it may be, enslaved. Coercion is not, however, a term that covers every form of inability. If I say that I am unable to jump more than ten feet in the air, or cannot read because I am blind...it would be eccentric to say that I am to that degree enslaved or coerced. Coercion implies the deliberate interference of other human beings within the area in which I could otherwise act. You lack political liberty or freedom only if you are prevented from attaining a goal by other human beings (Berlin, 1969: 122).

For Berlin and those who follow him, then, the heart of liberty is the absence of coercion by others; consequently, the liberal state's commitment to protecting liberty is, essentially, the job of ensuring that citizens do not coerce each other without compelling justification. However, despite the powerful case for negative liberty, many liberals have been attracted to more ‘positive’ conceptions of liberty. Although Rousseau (1973 [1762]) seemed to advocate a positive conception of liberty, according to which one was free when one acted according to one's true will (the general will), the positive conception was best

developed by the British neo-Hegelians of the late nineteenth and early twentieth centuries, such as Thomas Green and Bernard Bosanquet (1923). Green acknowledged that '...it must be of course admitted that every usage of the term [i.e., freedom] to express anything but a social and political relation of one man to other involves a metaphor...It always implies...some exemption from compulsion by another...(1986 [1895]: 229). Nevertheless, Green went on to claim that a person can be unfree if he is subject to an impulse or craving that cannot be controlled. Such a person, Green argued, is '...in the condition of a bondsman who is carrying out the will of another, not his own' (1986 [1895]: 228). Just as a slave is not doing what he *really* wants to do, one who is, say, an alcoholic is being led by a craving to look for satisfaction where it cannot, ultimately, be found.

For Green, a person is free only if she is self-directed or autonomous. Running throughout liberal political theory is an ideal of a free person as one whose actions are in some sense her *own*. Such a person is not subject to compulsions, critically reflects on her ideals and so does not unreflectively follow custom and does not ignore her long-term interests for short term pleasures. This ideal of freedom as autonomy has its roots not only in Rousseau's and Kant's political theory, but in John Stuart Mill's *On Liberty*. And today it is a dominant strain in liberalism, as witnessed by the work of S.I. Benn (1988), Gerald Dworkin (1988), and Joseph Raz (1986).

A continuing problem for liberal theory is whether these two conceptions of freedom can somehow co-exist (Benn, 1988) or whether they are, as Berlin argued, fundamentally at odds. The worry about positive freedom is that it seems to lay the basis for justifying paternalistic interferences *on the grounds that they are freedom-enhancing*. A paternalist imposes on a person for his own good, and typically this imposition would appear to limit that person's freedom. The paternalist stops him from drinking or taking drugs. But if drinking and so on hinders a person from acting autonomously, then this imposition may actually increase his freedom, since it increases his long-term propensity to act autonomously (Young, 1986). Thus, in Rousseau's chilling words, it seems that one can be 'forced to be free' (1973 [1762]: 177).

In any event, those who emphasise positive freedom and autonomy are apt to take a much more expansive view of the job of the liberal state. Benn, for instance, endorses 'rights to the conditions for autonomy' --- 'the rights to such conditions would presumably include the *Universal Declaration* rights to education, to leisure, and to participate in the cultural life of the community...' (1988: 251-52; but cf. Spector, 1992). This sort of welfarist-participatory state is precisely that which many advocates of negative liberty see as a threat to freedom. Thus argues Jan Narveson:

> The trouble with the enthusiast for 'positive liberty' is that to bring about the no doubt excellent goals he professes, he is willing to violate negative liberty. [And]...'positive liberty' might as well be replaced by 'welfare' or some such term. If we think people's liberty important, we should think it important enough not to violate it for the sake of promoting any such goal, even if it be called 'liberty' (1988: 33).

Property and the Market

Liberal political theory, then, fractures over the conception of liberty. But a more important division concerns the place of private property and the market order. For classical liberals liberty and private property are intimately related. From the eighteenth century right up to today, classical liberals have insisted that an economic system based on private property is uniquely consistent with individual liberty, allowing each to live her life --- including employing her labour and her capital --- as she sees fit. Indeed, classical liberals and libertarians have often asserted that in some way liberty and property are really the same thing; it has been argued, for example, that all rights, including liberty rights, are forms of property; others have maintained that property is itself a form of freedom (Gaus, 1994a; Steiner, 1994). A market order based on private property is thus seen an *embodiment* of freedom (Robbins, 1961: 104). Unless people are free to make contracts and to sell their labour, or unless they are free to save their incomes and then invest them as they see fit, or unless they are free to run enterprises when they have obtained the capital, they are not really free.

Classical liberals employ a second argument connecting liberty and private property. Rather than insisting that the freedom to obtain and employ private property is simply one aspect of people's liberty, this second argument insists that private property is the only effective means for the protection of liberty. Here the idea is that the dispersion of power that results from a free market economy based on private property protects the liberty of subjects against encroachments by the state. As F.A. Hayek argues, ‘There can be no freedom of press if the instruments of printing are under government control, no freedom of assembly if the needed rooms are so controlled, no freedom of movement if the means of transport are a government monopoly’ (1978: 149).

What has come to be known as ‘new’, ‘revisionist’, or ‘welfare state’ liberalism challenges this intimate connection between personal liberty and a private property based market order (Freeden, 1978; Gaus, 1983a; Macpherson, 1973: ch. 4). Three factors help explain the rise of this revisionist theory. First, the new liberalism arose in the late nineteenth and early twentieth centuries, a period in which the ability of a free market to sustain what Lord Beveridge (1944: 96) called a ‘prosperous equilibrium’ was being questioned (Gaus, 1983b). If a private property based market tended to be unstable, or could, as Keynes argued (1973 [1936]), get stuck in an equilibrium with high unemployment, new liberals came to doubt that it was an adequate foundation for a stable, free society. Here the second factor comes into play: just as the new liberals were losing faith in the market, their faith in government as a means of supervising economic life was increasing. This was partly due to the experiences of the First World War, in which government attempts at economic planning seemed to succeed (Dewey, 1929: 551-60); more importantly, this reevaluation of the state was spurred by the democratisation of western states, and the conviction that, for the first time, elected officials could truly be, in J.A. Hobson's phrase ‘representatives of the community’ (1922: 49). As D.G. Ritchie observed:

> be it observed that arguments used against ‘government’ action, where the government is entirely or mainly in the hands of a ruling class or caste, exercising wisely or unwisely a paternal or grandmotherly authority --- such arguments lose their force just in proportion as the government becomes more and more genuinely the government of the people by the people themselves (1896: 64).

The third factor underlying the development of the new liberalism was probably the most fundamental: a growing conviction that, so far from being ‘the guardian of every other right’ (Ely, 1992: 26), property rights generated an unjust inequality of power that led to a less-than-equal liberty (typically, ‘positive liberty’) for the working class. This theme is central to contemporary American liberalism, which combines strong endorsement of civil and personal liberties with, at best, an indifference, and often enough an antipathy, to private ownership. Once again, the seeds of this newer liberalism can be found in Mill's *On Liberty*. Although Mill insisted that the ‘so-called doctrine of Free Trade’ rested on ‘equally solid’ grounds as did the ‘principle of individual liberty’ (1991 [1859]: 105), he nevertheless insisted that the justifications of personal and economic liberty were entirely distinct. And in his *Principles of Political Economy* Mill consistently emphasises that it is an open question whether personal liberty can flourish without private property (1976 [1871]: 210), a position that Rawls was to reaffirm a century later (1971: 258).

Liberalism as a Philosophy

Although liberalism is, first and foremost, a political philosophy, ‘liberal’ has come to be employed to describe a group of comprehensive philosophies (Rawls, 1993), including a theories of ethics, value, the person and knowledge.

Liberal Ethics

Following Wilhelm von Humboldt (1993 [1854]), Mill's *On Liberty* based the case for the primacy of freedom on the goodness of developing individuality and the cultivating capacities:

> Individuality is the same thing with development, and...it is only the cultivation of individuality which produces, or can produce, well-developed human beings...what more can be said of any condition of human affairs, than that it brings human beings themselves nearer to the best thing they can be? or what worse can be said of any obstruction to good, than that it prevents this? (Mill, 1991 [1859]: 71)

This is not just a theory about politics: it is a substantive, perfectionist, moral theory about the good. And, on this view, the right thing to do is to promote development, and only a regime securing each individual extensive liberty can accomplish this. This moral ideal of human perfection and development dominated liberal thinking in the latter part of the nineteenth, and for most of the twentieth, century: not only Mill, but T.H. Green, L.T. Hobhouse, Bernard Bosanquet, John Dewey and even John Rawls show allegiance to variants of this perfectionist ethic and the claim that it provides the foundation for a regime of liberal rights. (Gaus, 1983a). And it is fundamental to the proponents of liberal autonomy discussed above as well as ‘liberal virtue’ theorists such as William Galston (1980). That the good life is necessarily a freely chosen one in which a person develops his unique capacities as part of a plan of life is probably the dominant liberal ethic of the past century.

This may seem a surprising claim given that, at least since the publication of Rawls's *Theory of Justice*, it

has generally been thought that the main moral dispute among liberals stems from the divide between utilitarians and rights theorists. This is, of course, a real divide, and in the last twenty years it has indeed come to dominate liberal debate. But interestingly, sometimes even those on opposite sides of this supposedly fundamental split advocate some version of liberal perfectionism. Thus the utilitarian-inspired J.S. Mill formulated the canonical version of liberal perfectionism while the apparently anti-utilitarian Rawls insists in his *Theory of Justice* that ‘human beings enjoy the exercise of their realized capacities (their innate and trained abilities), and this enjoyment increases the more the capacity is realized, or the greater its complexity’ (1971: 426; Gaus, 1981).

To say that liberal perfectionism has come to be a distinctly liberal ethic is not to merely assert that it is an ethic employed to defend liberal political positions. It is to make a stronger claim that liberals have come to understand the nature of moral rightness as founded on the pursuit of individuality and value of human development. In this light, it is less clear that utilitarianism constitutes a liberal ethic. To be sure, the notion of a ‘liberal utilitarianism’ is not an oxymoron; but neither is the term at all redundant. Interestingly, in his attempt to defend an explicitly *Liberal Utilitarianism*, Jonathan Riley advocates a social welfare function that restricts the domain of preferences to the ‘morally admissible’ or ‘ideal’, and these turn out to be those that reflect the sort of character ideal presented by Mill (1988: 83-92). Thus Riley liberalises utilitarianism by building in Mill's perfectionism. Given sufficient assumptions about human motivation, preferences, lack of knowledge and so on, a utilitarian ethic can endorse a liberal politics, but the relation between the two is highly contingent.

The main challenge to Millian perfectionism as a distinctly liberal ethic comes not from utilitarianism but from moral contractualism, which can be divided into what might very roughly be labeled ‘Kantian’ and ‘Hobbesian’ versions. According to Kantian contractualism, ‘society, being composed of a plurality of persons, each with his own aims, interests, and conceptions of the good, is best arranged when it is governed by principles that do not themselves presuppose any particular conception of the good...’(Sandel, 1982: 1-7). On this view, respect for the person of others demands that we refrain from imposing our view of the good life on them. Only principles that can be justified to all respect the personhood of each. Thus the tendency of recent liberal theory (Rawls, 1971; Reiman, 1990) to transform the social contract from an account of the state to an overall justification of morality, or at least a social morality. Basic to such ‘Kantian contractualism’ is the idea that individuals are motivated not by the pursuit of gain, but by a commitment or desire to publicly justify the claims they make on others (Reiman, 1990; Gaus, 1990; Scanlon, 1982). A moral code that could be the object of agreement among rational individuals is thus a publicly justified morality.

In contrast, the Hobbesian version of contractualism supposes only that individuals are self-interested, and correctly perceive that each person's ability to effectively pursue her interests is enhanced by a framework of norms that structure social life and divide the fruits of social cooperation (Gauthier, 1986; Kavka, 1986). Morality, then, is common framework that advances the self-interest of each. The claim of Hobbesian contractualism to be a distinctly liberal conception of morality stems from the importance of individual freedom and property in such a common framework: only systems of norms that allow each person great freedom to pursue her interests as she sees fit could, it is argued, be the object of consensus among self-interest agents. The continuing problem for Hobbesian contractualism is the apparent

rationality of free-riding: if everyone (or enough) comply with the terms of the contract, and so social order is achieved, it would seem rational to defect, and act immorally when one can gain by doing so. This is essentially the argument of Hobbes's ‘Foole’, and from Hobbes (1948 [1651]: 94ff) to Gauthier (1986: 160ff), Hobbesians have tried to reply to it.

Liberal Theories of Value

Turning from rightness to goodness, we can identify three main candidates for a liberal theory of value. We have already encountered the first: Millian perfectionism. Insofar as perfectionism is a theory of right action --- that rightness consists of promoting what Mill called ‘utility in the largest sense’, i.e., human development (1991 [1859]: 15) --- it can be understood as an account of morality. Obviously, however, it is an account of rightness that presuppose a theory of value or the good: the ultimate human value is developed personalities or an autonomous life. Competing with this objectivist theory of value are two other liberal accounts: pluralism and subjectivism.

In his famous defence of negative liberty, Isaiah Berlin insisted that values or ends are plural, and no interpersonally justifiable ranking among these many ends is to be had. More than that, Berlin maintained that the pursuit of one end necessarily implies that other ends will not be achieved. In this sense ends collide or, in the more prosaic terms of economics, the pursuit of one end necessarily entails opportunity costs in relation to others which cannot be impersonally shown to be less worthy. So there is no interpersonally justifiable way to rank the ends, and there is no way to achieve them all. The upshot is that each person must devote herself to some ends at the cost of ignoring others. For the pluralist, then, autonomy, perfection or development are not necessarily ranked higher than hedonistic pleasures, environmental preservation or economic equality. All compete for our allegiance, but because they are incommensurable, no choice can be interpersonally justified as correct.

The pluralist is not a subjectivist: that values are many, competing and incommensurable does not imply that they are somehow dependent on subjective experiences. But the claim that what a person values rests on experiences that vary from person to person has long been a part of the liberal tradition. To Hobbes, what one values depends on what one desires (1948 [1651]: 48). Locke advances a ‘taste theory of value’ (Gaus, 1986):

> The Mind has a different relish, as well as the Palate; and you will as fruitlessly endeavour to delight all Man with Riches or Glory, (which yet some Men place their Happiness in,) as you would satisfy all men's Hunger with Cheese or Lobsters; which, though very agreeable and delicious fare to some, are to others extremely nauseous and offensive: And many People would with reason preferr [sic] the griping of an hungry Belly, to those Dishes, which are a Feast to others. Hence it was, I think, that the Philosophers of old did in vain enquire, whether the *Summun bonum* consisted in Riches, or bodily Delights, or Virtue, or Contemplation: And they might have as reasonably disputed, whether the best Relish were to be found in Apples, Plumbs or Nuts; and have divided themselves into Sects upon it. For...pleasant Tastes depend not on the things themselves, but their agreeableness to this or that particulare Palate, wherein there is great variety...(1975 [1706]: 269).

The perfectionist, the pluralist and the subjectivist concur on the crucial point: the nature of value is such that people will pursue different ways of living. To the perfectionist, this is because each person has unique capacities, the development of which confers value on her life; to the pluralist, it is because values are many and conflicting, and no one life can include them all, or make the interpersonally correct choice among them; and to the subjectivist, it is because our ideas about what is valuable stem from our desires or tastes, and these differ from one individual to another. All three views, then, defend the basic liberal idea that people rationally follow very different ways of living. But in themselves, such notions of the good do not constitute a full-fledged liberal ethic, for an additional argument is required linking liberal value with norms of equal liberty. To be sure, Berlin seems to believe this is a very quick argument: the inherent plurality of ends points to the *political* preeminence of liberty (Kocis: 1980). Guaranteeing each a measure of negative liberty is, Berlin argues, the most humane ideal, as it recognises that 'human goals are many,' and no one can make a choice that is right for all people (1969: 171). But the move from diversity to equal liberty and individual rights seems a complicated one; it is here that both subjectivists and pluralists often rely on versions of moral contractualism. Those who insist that liberalism is ultimately a nihilistic theory can be interpreted as arguing that this transition cannot be made successfully: liberals, on their view, are stuck with a subjectivistic or pluralistic theory of value, and no account of the right emerges from it.

Liberal Epistemology

On the face of it, it may seem odd to think of a distinctively liberal theory of knowledge, but liberalism has always been closely associated with the Enlightenment and its defence of reason. Once again, we witness a split between two liberal camps. The rationalistic camp is best represented by Voltaire and the *philosophes* in whom it takes the form not only of a defence of science but an attack on superstition, custom and, importantly, religion. Thus the secular and anti-religious character of much liberal thought. This sort of militant, confident, rationalism is, however, also associated with great confidence in the ability of humans to understand nature and control their social world. According to Hayek, the flaw at the heart of utilitarianism stems from embracing this rationalism:

> The trouble with the whole utilitarian approach is that, as a theory professing to account for a phenomenon which consists of a body of rules, it completely eliminates the factors which makes rules necessary, namely our ignorance. It has indeed always amazed me how serious and intelligent men, as the utilitarians undoubtedly were, could have failed to take seriously this crucial fact of our necessary ignorance of most of the particular facts, and could have proposed a theory which supposes knowledge of the particular effects of our individual actions when in fact the whole existence of the phenomenon they set out to explain, namely a system of rules of conduct, was due to the impossibility of such knowledge (1976: 20).

Karl Popper (1945) made a similar charge against Plato, Hegel and Marx: *viz.* they failed to appreciate the limits of knowledge. Hayek and Popper, then, represent the other strain of liberal epistemology: an insistence that reason is limited, and our basic position is one of ignorance. This cautious, fallibilistic,

liberalism is less apt to be militantly secular than tolerant of religion; it is more likely to stress the incremental and experimental nature of social policy than to advocate grand social reconstructions. And it is more likely to appreciate the market, as a device for coping with our constitutional ignorance, and less likely to be enamored with state planning. To a significant extent, the marked differences between, on the one hand, British and Austrian liberalism and, on the other, French liberalism, stem from these different views of the power of reason in human affairs.

The Metaphysics of Liberalism

In his *Liberalism and the Limits of Justice* Michael Sandel charged that ‘Kantian liberals’ in general, and John Rawls in particular, are committed to a conception of the person according to which the self is in some way prior to its ends or substantive attachments. What has become known as the ‘communitarian critique of liberalism’ has insisted that this is implausible --- people are ‘constituted’ by their ends or values, and they cannot abstract from these particular ends and social commitments to deliberate on matters of justice from a ‘disembodied’ perspective. Although it is dubious that liberals are really committed to all *that*, it is surely the case that, overwhelmingly, liberals do believe that individual persons are ontologically prior to social groups and relations and, so, persons and their identities are distinct, and that central to personhood is a capacity to choose among alternative ways of living.

A perennial issue in liberal theory is the extent to which this basic individualism can be combined with a recognition of the social nature of humans, and the importance of one's social environment in the formation of personality. Stanley Benn is among those who insist that the liberal commitment to persons as choosers is in no way inconsistent with appreciating the importance of our social inheritance. The liberal individual, he insists, does not:

> conjure his *nomos* out of thin air, adopting it by a kind of random fancy, kicking aside the *nomoi* of his culture, its traditions, as so much clutter. One's reasons for engaging in an activity as worthwhile, for accepting the principles and standards that regulate it as constraining one's own performances, must already be built into one's conception of the world, which one must have received initially from those about one, as conceptual resources made available by the cluster of subcultures that combine to make one what one is --- or rather that provide the materials for what one can become (1988: 220-21).

It has been plausibly argued that the French, as opposed to the British liberal tradition, has taken more seriously these social influences on individuals and their lives (Seidentop, 1979). More generally, Will Kymlicka (1989) has argued that liberalism can indeed makes sense of ‘cultural membership’ and the way individual identities are dependent upon it. Yet it is unclear just how far liberalism's basic individualism permits accommodation with communitarian conceptions of the self, in which one's identity is bound up with a group identity (Gaus, 1983). Although the liberal can certainly acknowledge that we are both individual and social creatures, it seems doubtful that liberals can see individualised personalities as simply social artefacts of a particular, Western, culture: some sort of inherent individuation of personalities seems, as John Chapman argues (1977), a basic element of the liberal conception of human nature. Thus the worry that T.H. Green's or Bernard Bosanquet neo-Hegelian theory of the person is at

odds with their liberal politics. According Bosanquet's Absolute Idealism, individual persons are less real, because less complete and coherent, than the social whole (Gaus, 1994b). Moreover, Bosanquet insisted that ‘it is very hard to establish a difference in principle between the unity of what we call one mind and that of all the “minds” which enter into a single social experience’ (1923: 166).

The Return of a Purely Political Liberalism?

Prominent liberals have recently shied away from the conception of liberalism as a comprehensive philosophy, and have sought to return to its roots: as a purely political doctrine. This important development, aptly enough described as ‘political liberalism’, insists that liberalism as a comprehensive philosophy --- as including an ethical theory, an epistemology or a metaphysics of the person and society --- is just one more controversial or ‘sectarian’ doctrine in a society already filled with such doctrines. To John Rawls (1993: 5ff), the preeminent proponent of this view, such a ‘sectarian liberalism’ is open to rational dispute, and thus is not in the requisite sense publicly justified. If it is to serve as the basis for public reasoning in our diverse western societies, liberalism must be restricted to a core set of political principles that are, or can be, the subject of consensus among all reasonable citizens. Rawls's notion of a purely political conception seems in fact more austere than the traditional liberal political theories discussed above, being largely restricted to constitutional principles upholding basic civil liberties and the democratic process.

There are good grounds for doubting that liberalism can really rid itself of controversial metaphysical (Hampton, 1989) or epistemological (Raz, 1990; Gaus, 1996) commitments. As indicated above, Rawls seems to rest his case on the requirements of public justification, yet he seeks a distinctly political, non-epistemological, conception of justification (1993:44). And this, of course, because epistemological theories are controversial. We thus seem driven to the idea that a citizen could ‘politically justify’ a claim in a way that violates her epistemic standards of what constitutes a good reason. It is not at all clear, though, whether one could see oneself as having a politically justified claim on another while recognising that the argument for that claim depends on what, from one's epistemic perspective, are bad reasons (Gaus, 1996: 131ff).

Liberalism is, first and foremost, a political theory, yet it seems dubious that it can be a *purely* political theory. While no liberal need embrace every element of the wider liberal philosophy --- not every liberal must advance a liberal notion of the morally right, a liberal conception of value, a liberal epistemology and a liberal metaphysics of the person --- it is hard to see how any liberal political theory can avoid *all* of these. To be sure, no necessary principles mandate how political philosophy links up to the rest of philosophy. But neither is it an entirely autonomous field; hence the ‘comprehensive’ nature of all liberal theories.

Bibliography

- Benn, Stanley I. (1988). *A Theory of Freedom.* Cambridge: Cambridge University Press.

- Berlin, Isaiah (1969). ‘Two Concepts of Liberty’ in his *Four Essays on Liberty*. Oxford: Oxford University Press: 118-72.
- Beveridge, William (1944). *Full Employment in a Free Society*. London: Allen and Unwin.
- Bosanquet, Bernard (1923). *The Philosophical Theory of the State,* fourth edn. London: Macmillan.
- Chapman, John W. (1977). ‘Toward a General Theory of Human Nature and Dynamics’ in *NOMOS XVII: Human Nature in Politics,* J. Roland Pennock and John W. Chapman, eds. New York: New York University Press: 292-319.
- Cranston, Maurice (1967). ‘Liberalism’ in *The Encyclopedia of Philosophy*, Paul Edwards, ed. New York: Macmillan and the Free Press: 458-461.
- Dewey, John (1929). *Characters and Events,* Joseph Ratner, ed. New York: Henry Holt.
- Dworkin, Gerald (1988) *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.
- Ely, James W. Jr (1992). *The Guardian of Every Other Right: A Constitutional History of Property Rights.* New York: Oxford University Press.
- Freeden, Michael (1978). *The New Liberalism: An Ideology of Social Reform.* Oxford: Clarendon Press.
- Galston, William (1980). *Justice and the Human Good.* Chicago: University of Chicago Press.
- Gaus, Gerald F. (1981) ‘The Convergence of Rights and Utility: The Case of Rawls and Mill’, 92 *Ethics:* 57-72.
- Gaus, Gerald F. (1983a). *The Modern Liberal Theory of Man.* New York: St. Martin's Press.
- Gaus, Gerald F. (1983b). ‘Public and Private Interests in Liberal Political Economy, Old and New’ in *Public and Private in Social Life,* S.I. Benn and G.F. Gaus, eds. New York: St. Martin's Press: 183-221.
- Gaus, Gerald F. (1986). ‘Subjective Value and Justificatory Political Theory’ in *NOMOS XXVIII: Justification,* J. Roland Pennock and John W. Chapman, eds. New York: New York University Press: 241-69.
- Gaus, Gerald F. (1990). *Value and Justification.* Cambridge: Cambridge University Press.
- Gaus, Gerald F. (1994a). ‘Property, Rights and Freedom.’ 11 *Social Philosophy and Policy:* 209-40.
- Gaus, Gerald F. (1994b). ‘Green, Bosanquet and the Philosophy of Coherence’ in *The Routledge History of Philosophy*, vol VII: *The Nineteenth Century,* C.L. Ten, ed. London: Routledge: 408-33.
- Gaus, Gerald F (1996). *Justificatory Liberalism: An Essay on Epistemology and Political Theory.* New York: Oxford University Press.
- Gauthier, David (1986). *Morals By Agreement.* Oxford: Oxford University Press.
- Green, Thomas Hill (1986 [1895]). *Lectures on the Principles of Political Obligation and Other Essays*, Paul Harris and John Morrow, eds. Cambridge: Cambridge University Press.
- Hampton, Jean (1989) ‘Should Political Philosophy by Done without Metaphysics?’ 99 *Ethics:* 791-814.
- Hayek, F.A. (1976). *The Mirage of Social Justice.* Chicago: University of Chicago Press.
- Hayek, F.A. (1978). ‘Liberalism’ in his *New Studies in Philosophy, Politics, Economics and the History of Ideas.* London: Routledge and Kegan Paul.
- Hobbes, Thomas (1948 [1651]). *Leviathan,* Michael Oakeshott, ed. Oxford: Blackwell.

- Hobson, J.A. (1922). *The Economics of Unemployment*. London: Allen and Unwin.
- Kant, Immanuel, (1965 [1797]). *The Metaphysical Elements of Justice*, John Ladd, trans. Indianapolis: Bobbs-Merrill.
- Kavka, Gregory S. (1986). *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.
- Keynes, John Maynard (1973 [1936]). *The General Theory of Employment, Interest and Money*. London and Cambridge: Macmillan and Cambridge University Press.
- Kocis, Robert A. 'Reason, Development and the Conflict of Human Ends: Sir Isaiah Berlin's Vision of Politics.' *American Political Science Review*, 74 (March 1980): 38-52.
- Kymlicka, Will (1989). *Liberalism, Community and Culture*. Oxford: Clarendon Press.
- Locke, John (1960 [1689]). *The Second Treatise of Government* in *Two Treatises of Government*, Peter Laslett, ed. Cambridge: Cambridge University Press: 283-446.
- Locke, John (1975 [1706]). *An Essay Concerning Human Understanding*, Peter H. Nidditch, ed., Oxford: Clarendon Press.
- Macpherson, C.B. (1973). *Democratic Theory: Essays in Retrieval*. Oxford: Clarendon Press.
- Mill, John Stuart (1976 [1871]). *Principles of Political Economy*. Fairfield: Augustus M. Kelley.
- Mill, John Stuart (1991 [1859]). *On Liberty and Other Essays*, John Gray, ed. New York: Oxford University Press: 471-582.
- Narveson, Jan (1988). *The Libertarian Idea*. Philadelphia: Temple University Press.
- Popper, Karl (1945) *The Open Society and its Enemies*. London: Routledge.
- Rawls, John (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Rawls, John (1993). *Political Liberalism*. New York: Columbia University Press.
- Raz, Joseph (1986). *The Morality of Freedom*. Oxford: Clarendon Press.
- Raz, Joseph (1990). 'Facing Diversity: The Case of Epistemic Abstinence,' 19 *Philosophy & Public Affairs:*3-46.
- Reiman, Jeffrey (1990). *Justice and Modern Moral Philosophy*. New Haven, CT: Yale University Press.
- Riley, Jonathan (1988). *Liberal Utilitarianism*. Cambridge: Cambridge University Press.
- Ritchie, D.G. (1896). *Principles of State Interference*, 2nd, edn. London: Swan Sonnenschein.
- Robbins, L. (1961). *The Theory of Economic Policy in English Classical Political Economy*. London: Macmillan.
- Rousseau, Jean-Jacques (1973 [1762]). *The Social Contract and Discourses*, G.D.H. Cole, trans. New York: Dutton.
- Sandel, Michael. (1982) *Liberalism and the Limits of Justice*. Cambridge: Cambridge University Press.
- Scanlon, Thomas (1982) 'Contractualism and Utilitarianism' in *Utilitarianism and Beyond*, Amartya Sen and Bernard Williams, eds. Cambridge: Cambridge University Press: 103-28.
- Seidentop, Larry (1979). 'Two Liberal Traditions' in *The Idea of Freedom*, Alan Ryan, ed. Oxford: Oxford University Press: 153-74.
- Spector, Horacio (1992). *Autonomy and Rights: The Moral Foundations of Liberalism*. Oxford: Clarendon.
- Steiner, Hillel (1994). *An Essay on Rights*. Oxford: Basil Blackwell.
- von Humboldt, Wilhelm (1993 [1854]). *The Limits of State Action*. Indianapolis: Liberty Press.

- Young, Robert (1986). *Personal Autonomy: Beyond Negative and Positive Freedom.* London: Croom-Helm.

Related Entries

consequentialism | Hobbes, Thomas | justification, political: public | Kant, Immanuel | liberty | Locke, John | social contract: contemporary approaches to

Gerald F. Gaus
ggaus@tulane.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 28, 1996
Content last modified: November 29, 1996

Public Justification

The idea of public justification is the key idea in contemporary liberal-democratic political theory. The idea is, roughly, that no regime is *legitimate* unless it is reasonable from every individual's point of view. (As will be seen, this formula conceals considerable ambiguity.) Although John Rawls is the foremost exponent of the idea of public justification, its importance is also marked in the work of Jürgen Habermas, David Gauthier, and others. There is considerable current interest in the question of how the ideal of public justification is to be properly articulated. Some theorists, such as Rawls, seem to read 'reasonable from every individual's point of view' more or less 'empirically', whereas others, most notably Jerry Gaus, claim that this crucial phrase must be given a normatively-loaded reading. On the first account, something like actual (or only mildly idealized) agreement is required for legitimacy, whereas, on the second account, 'reasonable' means 'supported by good reasons'. If the first account is accepted, there is some danger that regimes will be judged legitimate which are supported only or mainly by 'bad' reasons--i.e. which depend for their 'legitimacy' on mistaken beliefs or on morally inadmissible desires and preferences. If the second account is accepted, demonstrations of legitimacy may not be practically efficacious--i.e. they may need to be supplemented by forceful impositions of requirements which, while supported by 'good reasons', are not actually accepted by the individuals concerned. Much work in the area is concerned with the degree to which these competing demands--of 'practical efficacy' and 'morality'--can be balanced to yield some *public conception* of public justification. (For a good overview of work in this area, see D'Agostino and Gaus, 1998.)

Situating the Idea of Public Justification

It may seem innocuous to suggest that the measure of legitimacy of regimes is fundamentally tied to discursive justifiability. This view is nevertheless controversial and, indeed, may be unsupportable.

On the first point, methodological conservatives such as F.A. Hayek and Michael Oakeshott would certainly be sceptical about the idea that legitimacy could be established discursively.

According to Hayek, this kind of approach embodies the ‘synoptic delusion’, or, in other words, the idea that it is possible to survey adequately the full range of information that might be relevant to determinations of legitimacy, to manipulate this information as needed, and to arrive, via reasoning, at some conclusion one way or another. Furthermore, any such attempt would fail, as other approaches might not, to take adequate account of the ‘dispersed knowledge’ of a large community. On Hayek's account, issues of legitimacy are largely responsible to processes of social evolution--that is, to make an assessment of legitimacy, you need to consider facts about the course of development of the institutions.

According to Oakeshott, a fully discursive treatment of complex political issues associated with legitimacy will fail on account of its inability to take adequate notice of the ‘tacit knowledge’ of practitioners. There is, on this account, much more that is relevant to determining legitimacy than could be expressly articulated, and much more, hence, than could ever be made available as ‘input’ to a justificatory course of reasoning.

These are, if you like, methodologically *traditionalist* alternatives to the rational, discursive, and intellectualized approach of the advocates of public justification.

Another point of contrast is with those approaches to politics which think that it is mistaken--perhaps corrupt--to seek the kind of stasis (however temporary) that seems to be implied in the idea that some determinate regime might be publicly justifiable and hence legitimate. On this account, the point of politics is not to resolve existing differences--which seems, at least in some sense, to be the point of attempting public justification; it is, rather, to keep alive those differences which are (i) fertile of edifying forms of controversy, and (ii) expressive of the irresolvable plurality of views, frameworks, and ‘life-styles’ in any decently ‘free’ and moderately diverse community.

Jean-Francois Lyotard seems to be an advocate of this position, and resistance to reduction to consensus is also implied by Michel Foucault's views on ‘normalization’; he says, for instance, that "The search for a form of morality acceptable to everyone ... seems catastrophic". On this account, the main mechanism for facilitating the achievement of public justification would have to be coercive; whether actually or from some normatively-loaded perspective that corrects for ‘ideological distortion’, people's interests are irreconcilable and hence can be brought only by coercion into sufficient alignment to support a claim to public justification. The very demand for public justification is hence itself implicitly *authoritarian*.

A final point of contrast is with those approaches which see in Rawlsian, Habermasian, and other mechanisms of public justification some attempt to *substitute* allegedly but not actually politically-innocent notions of rationality for self-admittedly political negotiation of conflict. Benjamin Barber is representative of this school, which wants to seek legitimacy where it can be found, not in the theorist's discourse on justification, but, instead, in institutionally-embodied, public, ‘real-time’ explorations of disputed issues. ‘We, the People’ must do the work of legitimation ourselves, on this account; no theorist's claims to have discovered some ‘overlapping consensus’ implicit in our public political culture can save us from this work. As with Lyotard and others of his ilk, members of this school see the theorist

of public justification as a closet authoritarian.

Contrasting Conceptions of Public Justification

As mentioned earlier, the phrase ‘reasonable from the point of view of every individual’ conceals numerous ambiguities. At least three are of special importance in understanding the great variety of different schemes for articulating the concept of legitimacy as justifiability.

Empircal/Normative

On the empirical reading, some proposal is reasonable from some particular person's point of view only if that person has beliefs and desires which, according to h/er own scheme of reasoning, support that proposal to the degree, which, by h/er standards, is required. There are multiple concessions to the individual's actual perspective embodied in this idea. First of all, we deal with the individual's actual beliefs and desires, not those that s/he would have if s/he were better informed, more committed to believing what is supported by evidence, less selfish, etc. We take the individual, *qua* believer and desirer, as we find h/er. Secondly, we ‘respect’ the individual's actual ways of reasoning, however ‘defective’ they might be from the point of view of formal theories of decision-making and inference. Thirdly, we accept as given the particular level of evidential and inferential adequacy that the individual h/erself sets; we don't require h/er to meet standards that we think are better from the point of view of formal rationality.

On the normative reading, all this is reversed, at least at the limit. Some proposal is reasonable for a particular individual, on this reading, if it conforms to the beliefs and desires that individual *would* have *if*___; if it is supported to the degree that s/he *would* demand *if*___; if it was developed in accordance with the inferential procedures she *would* use *if*___. In each case, ‘if’ s/he were ideally rational, at least up to the point at which, given fallibility and finitude, it makes sense to imagine h/er being.

A normative reading of the demands of public justification seems to recommend itself. Certainly, we don't have to embrace Marxist notions of *ideology* to realize that a purely empirical approach to legitimation is likely to fail in two opposed ways. (i) There are (pre-theoretically) illegitimate regimes that would be judged legitimate, on the empirical standard, on account of the population's believing and desiring what they shouldn't, and wouldn't if their thinking hadn't been corrupted--precisely by their membership of this community. (ii) There are (pre-theoretically) legitimate regimes that would be judged illegitimate, on the empirical standard, again on account of their population's believing and desiring what they shouldn't. Surely, a ‘normative’ approach must be preferred in view of these possibilities.

On the other hand, a ‘normative’ approach seems to be vulnerable on two counts. *First*, it presupposes an accessibly univocal reading of what it is reasonable to believe and desire and to infer from one's beliefs and desires with respect to public political arrangements. If there is ‘reasonable’ disagreement *about the demands of reason itself*, then potentially--and if so reasonably--groups within a community may have

different understandings of what is publicly justified within that community, and hence different notions of what institutions might be legitimate for that community. And, of course, it *is* arguable, at least, that reason does not pronounce univocally in disciplining our empirical beliefs, desires, and inferences, at least in the realm of human social arrangements. This, if you like, is the lesson of much philosophy of science and epistemology inspired by the work of Thomas Kuhn, who noted very clearly the possibilities (i) that different people might apply the standards of reason in different ways, and (ii) that their doing so was an important factor in the longer-term development of scientific understanding.

Secondly, a normative approach seems to abandon an important guiding principle of justificationist accounts of legitimacy--to wit, their responsiveness to broadly ‘voluntaristic’ considerations. As Thomas Nagel has pointed out, a regime which is publicly justified, although inevitably employing coercion to enforce its primary rules of association, has, on account of its being reasonable from every point of view, some of the advantages of a purely voluntary association. A ‘normative’ approach also seems to sever the historically important link between liberalism and anti-paternalistic thinking. If social arrangements are legitimately enforced against the actual ‘reasons’ of the populace, then this is, within a ‘normative’ framework, on account of the superiority of the theorist's understanding of the reasons for those people to embrace this regime. Obviously, this is a form of ‘patronage’.

Consensus/Convergence

A second basic ambiguity also can be detected in the phrase ‘reasonable from every point of view’. On the one hand, this might be read as invoking the notion of a *consensus* of grounding reasons. If **A** and **B** are members of a community for which a regime **S** is publicly justified, then, on this reading, there must be some reason **r**, which **A** has and **B** has too, which is their common (consensual) reason in relation to the regime **S**. On the other hand, this phrase might be read as invoking the notion of a *convergence* on a conclusion from a variety of distinct (sets of) premises. If **S** is reasonable from **A**'s point of view on account of **r**(**A**), then it may well be that it is reasonable from **B**'s point of view on account of an **r**(**B**) which is distinct from **r**(**A**).

Rawls's position on this matter is somewhat obscure. On the one hand, his preferred phrase, ‘overlapping consensus’, might suggest the first, consensual reading. On the other hand, other statements of his seem to suggest that, for instance, the Kantian's legitimating reasons for some regime might differ quite markedly from the utilitarian's or the Christian's, thus suggesting that the convergence interpretation is more faithful to his intentions. Nevertheless, the consensual reading might be the right one. In his book *Political Liberalism*, Rawls appears to suggest that the overlapping consensus which he has in mind is one of ‘reasonable’ doctrines. Roughly, these are doctrines which acknowledge or are at least compatible with the acknowledgement of (i) the ‘burdens of judgment’ (roughly, the reasonableness of disagreement over high-level issues in ethico-political and scientific investigations), and (ii) an obligation not to ‘free ride’ on working and mutually beneficial social arrangements. In fact, it is arguable that it is these constituents of ‘reasonableness’ that do the work in the Rawlsian derivation of principles of justice, and, hence, that what makes a regime embodying those principles legitimate is its conformity with an individual's commitment to being ‘reasonable’. In this case, the mode of justification *is* consensual, as

initial appearances would suggest.

Maximizing/Universalizing

A third basic ambiguity trades on different modalities of reason. Brian Barry and David Gauthier, for instance, both distinguish *maximizing* and *universalizing* conceptions of reason. Considering matters from the point of view of the maximizing conception, a regime might be legitimate, for an individual, for the reason that it (maximally) advances that individual's interests. Considering matters from the point of view of the universalizing conception, a regime might be legitimate, for an individual, for the reason that it fairly advances the interests of all as seen from that individual's point of view. Obviously, these two perspectives are distinct, at least superficially. On the one hand, the maximizing conception of public justification suggests a *negotiation* between parties in search of a stable equilibrium of opposed forces--each has been assigned, by a regime, as much as s/he can be assigned consistently with the need to assign enough to every party to secure h/er compliance. On the other hand, the universalizing conception suggests a *discourse* between the parties in search of an ideal that each can endorse as a public ideal for the community of which they are members. In the first case, we have individuals thinking as private agents about their individual welfare; in the second, we have individuals thinking as public deliberators about the common good. The first conception produces the public out of the private, much in the manner of Bernard Mandeville and Adam Smith, whereas the second conception presupposes (as in the case of Rawls's ‘reasonableness’) some prior public-spiritedness on the part of individuals. Jon Elster notes this distinction and labels it that between the market and the forum.

These three ambiguities provide the basis for a typology of styles of public justification, and, certainly, different approaches to public justification can be observed. Rawls and Gauthier are both proponents of public justification, but their approaches could hardly be more different. The typology developed here helps foreground the ways in which they are different--and to provide pigeon-holes for the accommodation of numerous other perspectives.

Of course, there is a problem implicit in this variety of different approaches to public justification. Suppose that we wonder whether a regime **S** is legitimate. We want to know whether we should give our willing commitment to its demands or, instead, hold ourselves in readiness to oppose these demands when circumstances permit this. Suppose we accept that public justification--reasonableness from every perspective--is the basis for legitimacy. Suppose someone suggests that **S** is legitimate because there is an actual empirical convergence of maximizing reasons in favor of **S**. Do we know now whether to conform or resist? Not necessarily. Someone else might come along with the information that there is no hypothetical (i.e. normatively informed) consensus of universalizing reasons in favor of **S**. The question of legitimacy is--so far, anyway--*indeterminate* because the notion on which it depends is ambiguous.

Prospects for Public Justification

The notion of legitimacy is a *practical* notion in the sense that it is meant to inform our decision to

acquiesce in or (try to) resist the demands of the political regime of which we are citizens. It is an important desideratum for such a notion that it facilitate the formation of determinate judgments--e.g. that the regime **S** *determinately* is (or isn't) a legitimate regime. As indicated, the notion of public justification may not satisfy this desideratum. Crudely, it is ambiguous, at least *prima facie* what is in accordance with reason in relation to social arrangements.

Some mechanisms developed by Rawls and put by him to other uses may be relevant in resolving this ambiguity. In *A Theory of Justice*, Rawls considers a related problem--posed by the fact that there is, in our society and societies like it, no *public conception of justice*. Of course, there is a shared *concept* of justice, but this is too vague and abstract to serve as a practical guide. And while there are *conceptions* of justice which are sufficiently specific to have concrete practical implications, there is no agreement about which of these is the ‘right’ one for the group as a whole. So far the analogy is obvious with the issue of public justification. There is widespread agreement on the concept, but a plurality of interpretations of that concept. What does Rawls suggest in this case? To determine which conception of **X** is best fitted to play the role of public conception of **X**, we ask which conception best embodies the *ideal* of **X**, where that is given by *functional analysis* of the concept and its role in collective life. (This method is most obvious at section 23 of *A Theory of Justice*, where Rawls sets out the "formal conditions that it seems reasonable to impose on the conceptions of justice".) A similar approach might be satisfactory in the case of public justification.

A functional analysis in relation to public justification reveals many different desiderata for any adequate public conception of that concept. In particular, it is reasonable, given the role of the idea of public justification, to expect that any public justification will capture the ‘double-sidedness’ of that concept. On the one hand, public justification does seek some distance from people's actual ‘reasons’, in search of genuine justification rather than specious rationalization. On the other hand, publicly justificatory arguments are meant to have an impact on the individuals which they target, and, in particular, to give them motives as well as reasons for conformity (or otherwise) to the demands they are subject to. (Stephen Macedo makes these points especially forcefully.) Unfortunately, there is *prima facie* incompatibility between these dual demands. To the extent that a given course of reasoning satisfies the demands for ‘normative distance’, to that extent is it likely to fail to meet the demands for motivational impact. In fact, there is a catalogue of such incompatibilities. (See D'Agostino, *Free Public Reason*, chapter 6.)

Of course, *prima facie* means ‘at first glance’. Perhaps there is some *canonical* way of balancing these competing demands and hence some public conception of what's required for public justification to be achieved. For reasons which Kuhn has suggested in another context, and which have already been alluded to, and which perhaps reflect elements of Lyotard's preferred ‘heterology’, it is hard to see how a ‘canonical trade-off’ amongst the various desiderata can in fact be identified. (See D'Agostino, *Free Public Reason*, chapter 7.) This means that the *prospects* are poor for public justification--and hence for discursive redemption of legitimacy claims--in societies like ours. Perhaps the Foucauldians and post-modernists are right in claiming that notions of legitimacy are inherently and inescapably themselves instruments of power, rather than ‘rational’ alternatives to force. Certainly, if there is no public conception of public justification, any regime is ‘legitimate’ only *given* a conception of legitimacy that is

itself controversial, and hence can be imposed only by force--not by the inducements of ‘reason’.

Bibliography

- Barber, Benjamin, *The Conquest of Politics*, Princeton University Press, 1988.
- Barry, Brian, *Theories of Justice*, Harvester/Wheatsheaf, 1989.
- D'Agostino, Fred, *Free Public Reason*, Oxford University Press, 1996.
- D'Agostino, Fred and Gaus, Gerald F., eds., Public Reason, Ashgate, 1998.
- Elster, Jon, "The Market and the Forum", in Jon Elster and Aanund Hylland, eds., *Foundations of Social Choice Theory*, Cambridge University Press, 1986.
- Gaus, Gerald, *Justificatory Liberalism*, Oxford University Press, 1996.
- Gauthier, David, *Morals by Agreement*, Clarendon Press, 1986.
- Hayek, F.A., *Law, Legislation and Liberty*, vol. I: *Rules and Order*, University of Chicago Press, 1973, especially chapter 1.
- Hoy, David, *Foucault: A Critical Reader*, Blackwell, 1986.
- Kuhn, Thomas, *The Structure of Scientific Revolutions*, University of Chicago Press, 1970.
- Lyotard, Jean-Francois, *Just Gaming*, University of Minnesota Press, 1985.
- Macedo, Stephen, *Liberal Virtues*, Clarendon Press, 1991.
- Nagel, Thomas, *Equality and Partiality*, Oxford University of Press, 1991.
- Oakeshott, Michael, *Rationalism in Politics*, Methuen, 1962.
- Rawls, John, *A Theory of Justice*, Oxford University Press, 1973.
- Rawls, John, *Political Liberalism*, Columbia University Press, 1993, especially Lecture VI.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

liberalism | social contract: contemporary approaches to

Fred D'Agostino
fdagosti@metz.une.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 27, 1996
Content last modified: July 28, 1997

Contemporary Approaches to the Social Contract

The idea of the social contract goes back, in a recognizably modern form, to Thomas Hobbes, but is most notably embodied, in our times, in the work of John Rawls. The basic idea is a simple one. What makes some particular system of collectively enforced social arrangements *legitimate* is that it is the object of an agreement for the people who are subject to it. (As in the case of *public justification*, the key phrase, 'the object of an agreement', is multiply ambiguous.) In the case of a literal contract--say for an exchange of goods--each of the parties has reason to honor the terms of the contract either in the (bare) fact of having agreed to its terms (under certain circumstances) or in the fact of its terms being agreeable ones. Similarly, in the case of a social contract in the manner of Hobbes or Rawls, each of the parties has reason to honor h/er responsibilities under the terms of the contract--e.g. to pay taxes, conform to laws, participate in decision-making, etc.--either on account of h/er agreement to do so, or, perhaps, on account of its being reasonable that s/he do so. (These are what Michael Lessnoff calls the *voluntaristic* and *rationalistic* readings of the contract.) In its modern guises, contractarianism is not intended as an account of the historical origins of current social arrangements, but, instead, as an answer to, or framework for answering, questions about legitimacy and political obligation. Important issues associated with contractarianism include the binding force of hypothetical agreements, the reduction (or not) of ethico-political to instrumental reasoning, and the compatibility of contractarianism with fairness and liberty.

Hypothetical Contracts?

Contemporary contractarianism is, characteristically, *doubly* hypothetical. Certainly, no prominent

theorist thinks that questions of legitimacy and obligation are settled by an actual survey of attitudes towards existing social arrangements, and are not settled until such a survey has been carried out. The question, then, is NOT "Are these arrangements the object of an actual agreement amongst ‘stakeholders’?" (If this were the question, the answer would typically be "No".) The question, rather, is "*Would* these arrangements be the object of an agreement if ‘stakeholders’ were surveyed?" Although both of the questions are, in some sense, susceptible to an empirical ‘reading’, only the latter is in play in present-day theorizing. The contract nowadays is always hypothetical in at least this first sense.

There is a reading of the (first-order) hypothetical question "Would the arrangements be the object of agreement if___" which, as indicated, is still resolutely empirical in some sense. This is the reading where what's required of the theorist is that s/he try to determine what an actual survey of actual ‘stakeholders’ would reveal about their actual attitudes towards their system of social arrangements. (We don't really perform the survey, but we do perform it ‘in imagination’.) But there is another reading that is more widely accepted in the contemporary context. On this reading, the question, really, is no longer a hypothetical question about actual reactions; it is, rather, a hypothetical question about *hypothetical* reactions--it is, as I have said, *doubly* hypothetical. *Framing* the question is the first hypothetical element: "Would it be the object of agreement if they were surveyed?" *Framed by* this question is the second hypothetical element, one which involves the so-called ‘stakeholders’, who are no longer treated ‘empirically’, i.e. taken as given, but are, instead, themselves considered from a hypothetical point of view--as they would be if (typically) they were better informed or more impartial, etc. The question for most contemporary contractarians, then, is, roughly, this. "If we surveyed the ‘idealized surrogates’ of the actual ‘stakeholders’ in this polity, would current social arrangements be the object of an agreement for them?" A "Yes" answer confers legitimacy and imposes obligations; a "No" answer signals illegitimacy and relieves us of or shows the purely ‘historical’ status of obligations that we might now submit to.

Of course, questions arise--and have been raised most notably by Ronald Dworkin--about how a (doubly) hypothetical agreement can bind any actual person. The point of second-stage hypotheticalizing is, *inter alia*, that, as I actually am, I would *not* agree to be bound by some system of social arrangements **S**. Suppose that it could be shown, however, that my ‘surrogate’ (a better informed, more impartial version of me) would agree to be bound by **S**. What has that to do with me? Where this second-stage hypotheticalization is employed, it seems to be proposed that *I* can be bound by agreements that *others*, different from me, would have made. It is like saying that I ought to be bound to respect **S** on account of *your* having agreed to be bound by **S**. While it might (though it needn't) be reasonable to suppose that I can be bound by agreements that I would myself have entered into if given the opportunity, it is just crazy to think that I can be bound by agreements that, demonstrably, I wouldn't have made *even if* I had been asked.

Rawls's solution to this problem reflects the complexity of his *original position argumentation* and the idea of *reflective equilibrium* which it depends on. In effect, Rawls identifies *two contracts*, one framing the other. The ‘first’ contract is one that, as we actually are, each of us makes with the ‘surrogate’ who is to represent us in second-stage contractual reasoning. As I am, I agree that the question is NOT "Do I agree as I actually am to **S**?" but, instead, "Would I agree if I were ___ to **S**?", or, in other words, "Will I

be bound by agreements that will be made in respect of **S** by my ‘idealized surrogate’ (or better self)?" Once I have answered "Yes" (of course hypothetically; there is no actual survey) to the first, framing question, I will be bound to the demands of **S** so long as my idealized surrogate--the subject of the second, framed (and still hypothetical) contractual question--says "Yes" to the system **S** of social arrangements. (This is what Rawls means when he characterizes the parties to the original position as ‘trustees’ for the interests of ‘you and me’.) Crudely, the reasoning runs as follows. I agree to be represented by **X** for certain purposes; **X** agrees that the system **S** is a legitimate one; hence I am bound by this system, for my ‘trustee’ has agreed to it on my behalf--this is one of the purposes for which s/he was to represent me. As Rawls says (*A Theory of Justice*, p.587): "Finally, we may remind ourselves that the hypothetical nature of the original position invites the question: why should we take any interest in it, moral or otherwise. Recall the answer: the conditions embodied in the description of this situation are ones that we do in fact accept."

Reductionism?

Contemporary contractarians tend to be *constructivists* in the sense that they recognize no independent and determinate external standard of legitimacy that the contractual device is intended to approximate, but, rather, make the truth-maker for "**S** is legitimate" that **S** was the object of an agreement (for ‘stakeholders’ or their surrogates). Crudely, being agreed to makes a regime legitimate; it is not that being agreed to is evidence for legitimacy otherwise conceived.

Within this constructivist framework, there are two main schools of contractarian thinking, which reflect differences which were already apparent between Hobbes's approach and John Locke's.

On the one hand, we have those contractarians, such as David Gauthier and James Buchanan, who think that legitimacy of regimes is determined by their prudential acceptability from the diverse points of view that are represented in relevant communities. On this account, the *basis* for an individual's agreement, and hence for h/er judgment that the regime is a legitimate one is that enforcement of the regime's demands contributes to the realization of h/er aspirations. On this account, to say "**S** is legitimate" is to say, more or less, that **S** is good for its various members. This is a *prudential* account of legitimacy and, if we think that prudence is a more ‘basic’ idea than the ideas of ‘morality’, then this approach is *reductionistic* in the sense that it derives ethico-political notions like ‘legitimacy’ and ‘obligation’ from non-ethico-political notions such as acceptance-grounded-on-prudence. Insofar as there is some problem in understanding how genuinely ethico-political reasons can also function as *motives* (alleged to be a problem on cognitivist interpretations), such a reductionistic strategy may be appealing; there is alleged to be little trouble understanding how purely prudential reasons can serve as motives--though, of course, this is a common assumption, rather than a demonstrated conclusion.

On the other hand, some contractarians, most notably Rawls, already build ethico-political assumptions into their particular approach to hypotheticalization. The kinds of ‘surrogates’ which ‘you and I’ commission to act as our agents in reasoning about legitimacy are, on Rawls's account, already so situated that their deliberations will be framed by ethico-political considerations. (See the article

"Original Position". The agents' deliberations are carried out in purely prudential terms, but they are subject to the ‘veil of ignorance’, which itself embodies important ethico-political notions.) Indeed, it can fairly be said that *any* approach to contractarian thinking that substitutes ‘impartial’ surrogates for concrete individuals is anti-reductionist in this way. (As Rawls himself points out, some broadly utilitarian approaches to legitimacy can be conceptualized as involving a hypothetical contract. The assumptions that John Harsanyi, for instance, makes in developing his version of the social contract make his approach an anti-reductionistic one.)

Challenges

Bruce Ackerman has mounted a profound challenge to contractarian thinking. It works, crudely, on the idea that the premises of a course of contractarian reasoning can be manipulated so as to yield (more or less) any conclusion that the theorist has some antecedent interest in producing. The argument goes as follows. Each version of contractarianism must specify three features: (i) the chooser **c**, (ii) the situation of choice **C**, and (iii) the alternatives **A** from which a choice is made. If **c** chooses **S** from **A** in **C**, then this establishes the legitimacy of **S** only *relative to* the specification of chooser, situation, and choice-set, and hence could be overturned, absent further constraints on the situation, by alternative specifications. Most obviously, if the choice-set is arbitrarily confined to **S** and various alternatives that are chosen by the theorist for their utterly unacceptable character, then any selection of **S** from **A** is meaningless as a legitimator. (The trick could be worked as easily, though not as transparently perhaps, by restricting the characterization of **c** or of **C**.) Perhaps the solution to this problem is simply to lift all restrictions on **C**, **A**, and **c**, hence banishing all arbitrary elements. If *all* the alternatives are canvassed and if the chooser isn't biased for or against any of them, then, surely, the fact that h/er choice was **S** does indeed confer legitimacy. However, as Ackerman points out, this won't work either, for, with no limits on **A**, and with no biasing--and hence distinguishing--characteristics built into **c**, there can be no choice at all.

Ackerman only considers two possibilities: (i) that contractarian thinking is unrestricted--but, if so, utterly empty; (ii) that contractarian thinking is ‘rigged’--but, if so, utterly devoid of normative force. However, there may be a non-arbitrary, non-empty form of such reasoning. This, anyway, is what Rawls thinks. He certainly doesn't permit ‘unrestricted’ reasoning about legitimacy; his deliberators in the original position are (surrogate) maximizers of social primary goods, reasoning in accordance with certain ethico-politically significant constraints--a veil of ignorance embodying a concept of justice. On the other hand, although the course of reasoning does make concrete assumptions about choosers, choice-situations, and choice-sets, the assumptions it makes, or so Rawls claims, are ones that we share. So while they might be arbitrary *sub specie aeternitatis*, they are not arbitrary from the point of view of the concrete individuals on whose behalf they are made. As these individuals see it, these particular assumptions are ‘privileged’ at least in the sense that they represent current community attitudes about justice. A demonstration of legitimacy is *relative to* these assumptions, but, since they are the best the community can do, there is nothing untoward about this. (This raises, but does not solve, the problem of ‘ideology’--i.e. that the community's very notion of justice is already corrupted. Notice that, within a constructivist framework, this judgment can itself be made only from the point of view of another community's concept of justice--a concept that might itself also be ‘corrupted’.)

Robert Nozick also mounts a powerful challenge to contractarian thinking, at least of the kinds that Rawls represents. His argument proceeds in three stages. (1) Nozick distinguishes ‘end-state principles’ of social life from ‘historical principles’. An end-state principle says, in effect, that a situation **S** is legitimate only if it approximates a particular canonical state of affairs **S***. An historical principle, on the other hand, says that any state is legitimate, whatever it might be substantively, so long as it has been reached from previous states by processes which are themselves legitimate. (2) Nozick notes that end-state judgments of legitimacy are very volatile in changing circumstances, and, hence, that ensuring continued legitimacy could, on this account, involve such arbitrarily large restrictions of liberty as to be unacceptable. (3) Nozick then notes that contractarian thinking, at least of the Rawlsian kind, delivers up end-state principles of legitimacy, and hence must be rejected--on account of point (2).

Nozick's argument is most vulnerable on point (2). It is true that Rawls's particular argument does yield an end-state principle in the form of the ‘difference principle’. And it certainly seems likely that continued legitimacy, by this standard, might require large impositions on individuals' liberties. But these considerations only establish that liberty and fairness may need to be traded off against one another in uncongenial ways, as Isaiah Berline had, of course, long ago pointed out; they do not establish a decisive objection to the contractarian approach.

Bibliography

- Ackerman, Bruce, *Social Justice in the Liberal State*, Yale University Press, 1980, especially section 66.
- Berlin, Isaiah, *Two Concepts of Liberty*, Clarendon Press, 1958.
- Buchanan, James and Tullock, Gordon, *The Calculus of Consent*, University of Michigan Press, 1965.
- Dworkin, Ronald, "The Original Position", in Norman Daniels, ed., *Reading Rawls*, Basil Blackwell, 1975.
- Gauthier, David, *Morals by Agreement*, Clarendon Press, 1986.
- Harsanyi, John, *Essays on Ethics, Social Behaviour and Scientific Explanation*, Reidel, 1977.
- Lessnoff, Michael, *Social Contract*, Macmillan, 1986.
- Nozick, Robert, *Anarchy, State, and Utopia*, Basil Blackwell, 1974.
- Rawls, John, *A Theory of Justice*, Oxford University Press, 1973.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

ethics: deontological | justification, political: public | liberalism | original position

Fred D'Agostino
fdagosti@metz.une.edu.au

First published: March 3, 1996
Content last modified: July 28, 1997

Stanford Encyclopedia of Philosophy

Original Position

The idea of the original position is perhaps the most lasting contribution of John Rawls to our theorizing about social justice. The original position is a hypothetical situation in which rational calculators, acting as agents or trustees for the interests of concrete individuals, are pictured as choosing those principles of social relations under which their principals would do best. Their choices are subject to certain constraints, however, and it is these constraints which embody the specifically moral elements of original position argumentation. Crudely, the rational calculators do not know facts about their principals which would be morally irrelevant to the choice of principles of justice. This restriction on their reasoning is embodied, picturesquely, in Rawls's so-called *veil of ignorance*, which occludes information, for instance, about principals' age, sex, religious beliefs, etc. Once this information about principals is unavailable to their agents, the plurality of interested parties disappears, and the problem of choice is rendered determinate. According to Rawls, agents so situated would choose two principles of justice, lexically ordered, affirming the equality of basic rights and an approach to social inequalities governed by the ‘difference principle’, according to which inequalities are unjust unless removing them would worsen the situations of the worst-off members of society. Original position argumentation is an example of contemporary contractarianism, involves a pure-proceduralist approach to the determination of moral principles, and is framed by reflective equilibration with widely agreed principles of public morality.

Reflective Equilibrium

There are *epistemological* and *political* interpretations of the original position.

On the epistemological reading, the original position is a methodological device for ridding the ethico-political ‘observer’ of hindrances to h/er clear and distinct perception of ethico-political facts. Just as it may be necessary to employ prosthetic sensory devices to make observations of distant or minute objects

or to use techniques of controlled experimentation to eliminate the influence of ‘noise’ and theoretically irrelevant confounding variables, so too, on this reading, might it be necessary, in order clearly to observe the ethico-political facts, to use some such device as the original position. Indeed, the original position might be well adapted to such a task. Eliminating knowledge of personal characteristics eliminates the possibility of bias in favor of those characteristics and thus enforces the kind of *impartiality* or disinterestedness that is held to be integral to a moral perspective. (In this regard, as Rawls himself recognizes, the original position device resembles that of ideal spectator theorists.)

This epistemological reading is nevertheless not the interpretation of the original position favored by Rawls himself. Although he has not repudiated the kind of ethico-political ‘realism’ that is presupposed by this reading, he believes that, since ‘realism’ in this sense is a reasonably disputed doctrine, a *practical* approach to the task of ethico-political justification must ‘prescind’ from the realism/non-realism debate within ethico-political meta-theory. Since there is reasonable disagreement about realism, we cannot presuppose it in the context of *public* political disputation. (On account of the ‘burdens of judgment’, we cannot expect to resolve the debate about realism to the (reasonable) satisfaction of every reasonable person; this doctrine therefore cannot provide a basis for *political* theorizing.)

On the political reading, the original position is a ‘device of representation’. Specifically, it represents, in the veil of ignorance, widely accepted principles for the choice of principles of justice. More concretely, the veil of ignorance embodies the *concept* of justice--i.e. the idea that distributions should not be based on morally irrelevant features. The information occluded by the veil of ignorance is, precisely, the community's understanding of what features are morally *ir*relevant to the choice of principles of justice. Although members of a given community may disagree about many matters relevant to issues of justice, they share--or are alleged or assumed to share--an understanding of justice that, while insufficiently concrete or detailed to provide on its own a workable *conception* of justice, *is* adequate to the task of framing the choice of such a conception. Working within the framework defined by the veil of ignorance and derived from this widely shared concept of justice, rational calculators choose principles of justice on the basis of their fiduciary duty to the concrete individuals whom they represent. Their choice is not of an ‘objectively correct’ conception of justice; it is, rather, of that conception which best expresses a shared understanding of justice in the community whose members they represent.

Rawls's idea of reflective equilibrium expresses this political understanding of justification. How are we to justify the claim that some particular conception of justice is the appropriate one? We are to do so, according to Rawls, by finding that conception which can be brought into reflective equilibrium with the *considered judgments of justice* which are current in a particular community. Of course, the process of reflective equilibration is dialectical. The main moments of the process are these.

- We articulate the concept of justice which is widely accepted within a given community.
- We so devise the veil of ignorance that it embodies this concept.
- We consider what implications about concrete and specific matters of justice rational calculators standing in a trustee relation would reach subject to the particular restrictions on their calculations represented by *this* veil of ignorance.
- We compare these implications with individuals' considered judgments of justice about these more

concrete and specific issues.
- Where there is divergence between implications and judgments, we consider whether individuals might be willing to alter their judgments to bring them into line with principles which, after all, already express their own more abstract views about the concept of justice.
- If there is residual divergence, we modify the veil of ignorance to minimize this divergence.

These operations are repeated until eliminable divergence is at a minimum; this is the state of reflective equilibrium. Individuals' concrete and specific judgments about justice are in equilibrium with those of other individuals, and all individuals in the community share both an abstract concept of justice (embodied in the veil) and a workable *public conception* of justice.

Early discussants assumed that the method of reflective equilibration was to be understood epistemologically. Even in *A Theory of Justice*, there was much textual support for the alternative political reading, but, whatever the situation in the early 1970s, it soon became clear that Rawls's preferred reading was indeed the political one. There are two stories about the development of Rawls's thinking. On the one hand, some commentators believe that Rawls had adopted an epistemological, specifically Kantian, approach to ethico-political justification in his earlier work, at least up to *A Theory of Justice*, which he then abandoned under the pressure of communitarian, specifically Hegelian, criticism at the hands, in particular, of Michael Sandel. On the other hand, some commentators believe that Rawls's position, at least since *A Theory of Justice*, has remained resolutely political, and that any genuine development of his thought has been prompted by considerations internal to his own perspective. (Rawls seems, in *Political Liberalism*, to endorse this latter reading of the history.)

For the mature Rawls (and perhaps too for the Rawls of *A Theory of Justice*), all ethico-political justification, in public contexts, is unavoidably politically rather than epistemologically based. It is based, in other words, on a convergence--or as Rawls calls it, an *overlapping consensus* of the main substantive ethico-political doctrines current in a community. Absent such a basis for convergence, there is no possibility of discovering, via reflective equilibration, principles of justice which can effectively regulate interactions between and distributions to the members of the community. And since such disagreement would make improbable any uncoerced acceptance of some epistemologically sanctioned set of principles, no voluntaristic basis for social justice could be found in this community--even if an ‘objective’ basis could be.

Pure Proceduralism

The method of original position argumentation is an example of pure proceduralism in ethico-political theorizing. This aspect of Rawls's work seems not to have been adequately conceptualized, but it is crucial for understanding larger issues.

Imagine that there is for a particular community a public conception of the good. In this case, it might be possible to develop rules for the distribution of goods and services on a broadly *teleological* basis. That is right (whether action or distribution or institution) whose implementation maximizes the realization of

the good. Of course, the availability of a public conception of the good is not, perhaps, a sufficient condition for the viability of such a teleological approach. Even given such a conception, a teleological approach may still be insufficiently sensitive to distributional issues. And, indeed, this is *one* reason why Rawls rejects a teleological approach to ethico-political justification. But Rawls also argues on other grounds against a teleological approach. In particular, he thinks that no such approach is viable (i) because the availability of a public conception of the good *is* a necessary condition for the viability of such an approach, and (ii) because there is no such public conception of the good in our society and in societies like it.

If we cannot develop ethico-political principles of right and justice on a teleological basis, then how can we do so? According to Rawls, we can do so via original position argumentation, framed with considerations of reflective equilibrium. That is right and just which would be acknowledged as such from the point of view of the original position. And what makes being acknowledged from this point of view the ‘right-maker’ for principles of justice? Because this point of view is the appropriate one for determining principles of justice, on account of its reflecting the community's existing concept of justice--on account of its reflecting their overlapping consensus of views about justice.

Notice that there is no teleological reasoning at work here. The right-maker for principles of justice is not defined in terms of the consequences for the realization of the good of conformity with those principles. The right-maker is (hypothetical) acceptance from a particular point of view. The right-maker for principles, in other words, is their being the ‘output’ of a particular procedure, in particular the procedure of original position argumentation. The reasoners in the original position are not trying, through their deliberations, to ensure an outcome that meets some already existing standard of justice for institutions. Why not? Because there is no such standard until it is constructed by their deliberation. And there is no such standard because there is pre-existing consensus within the community on *neither* a conception of the good--which, were it to exist, might permit a perfectly or imperfectly procedural approach to determining the principles of right, *nor* a full-blown conception of justice--which, were it to exist, would render any further reasoning otiose.

Veil of Ignorance

Far and away the most striking feature of Rawls's original position idea is the veil of ignorance. As Rawls points out, the idea of an *initial situation* of choice for ethico-political principles is common to other approaches, and represents a hypotheticalization of familiar reasoning within the *social contract* tradition. What is particularly interesting about Rawls's approach is that he proposes to restrict the basis for reasoning rather than expanding it, which is, for instance, the approach taken within the *ideal spectator* framework.

Crudely, ideal-spectator theorists make two theoretical ‘moves’ which Rawls more or less reverses. Recognizing that ethico-political thinking ought to be conducted from an impartial perspective, ideal-spectator theorists capture this notion of impartiality by amalgamating ethically-relevant information about *all* relevant parties--e.g. all the members of some community, and by assuming that the spectator in

whom this information is lodged makes h/er determination of principles on an equitable basis--e.g. in assigning equal weights to information about the preferences of individuals. There are various reasons for wondering whether this procedure is really a coherent one. Most notably, assumptions about the spectator's ability to store and synthesize information and calculate on its basis are wildly unrealistic. (See Cherniak 1986.) Furthermore, the spectator's calculations not only permit, they force h/er to reckon inter-personal gains and losses in the same way that a purely prudential self-interested reasoner would reckon *intra*-personal gains and losses. This is problematic for two reasons, one of which Rawls himself emphasizes. First of all, and this is Rawls's primary objection, such a procedure forces the spectator to sacrifice one individual's interests to those of others, *theoretically without limit*, whenever doing so would result in the maximization of the total the spectator is calculating. Secondly, the idea is suspect, to say the least, that there is some basis for the commensuration of individuals' diverse ways of valuing that would permit the determination of some socially valid aggregate for each of the various states of affairs which are being evaluated.

Crudely, Rawls hopes to avoid these difficulties by reversing the ‘moves’ of the spectator theorist. Instead of augmenting the information available to choosers, Rawls deliberately impoverishes it. Instead of requiring choosers to be impartial, he requires them to be purely self-interested--though, of course, in an extended sense; his choosers act to advance the interests of their principals. And by requiring unanimity among the various trustees or agents, Rawls ensures that individuals' interests are not sacrificed to that of ‘the collective’; each individual can veto, through h/er agent/trustee, any social settlement that isn't adequately respectful of h/er individuality. The veil of ignorance is of importance in this context. It ensures impartiality, despite the self-interestedness of the choosers, by preventing them, through lack of knowledge, from choosing in accordance with partial perspectives that might be favored by their principals. My agent **A** cannot hold out for some social settlement that favors people with those characteristics; s/he doesn't know what they are. S/he will therefore have to protect *my* interests, as s/he must as their trustee, only by holding out for a social settlement in which *no one's* interests are given short shrift. H/er impartiality is a product of h/er self-interestedness plus h/er ignorance. And the latter, crucial to this procedure, is a product of the veil of ignorance.

This account of matters also enables us to clear up a confusion that was often voiced in the first few years after the publication of *A Theory of Justice*. It was said that Rawls had sought--as others such as David Gauthier do seek--to reduce ethico-political principles of right to principles of prudence. This on account of the purely self-interested deliberations of choosers in the original position. What this suggestion ignores is that, though the choosers *reason* in a purely prudential way, their reasoning is *constrained* by their ignorance, and their ignorance is *expressive* of the moral demand for impartiality.

Bibliography

- Cherniak, Christopher, *Minimal Rationality*, M.I.T. Pressm 1986.
- Daniels, Norman, "Wide Reflective Equilibrium and Theory Acceptance in Ethics", *Journal of Philosophy* **76** (1979): 256-282.
- Dworkin, Ronald, "The Original Position", in Norman Daniels, ed., *Reading Rawls*, Basil

Blackwell, 1975.
- Rawls, John, *A Theory of Justice*, Oxford University Press, 1973, especially chapter III.
- Sandel, Michael, *Liberalism and the Limits of Justice*, Cambridge University Press, 1982, especially chapter 3.

Other Internet Resources

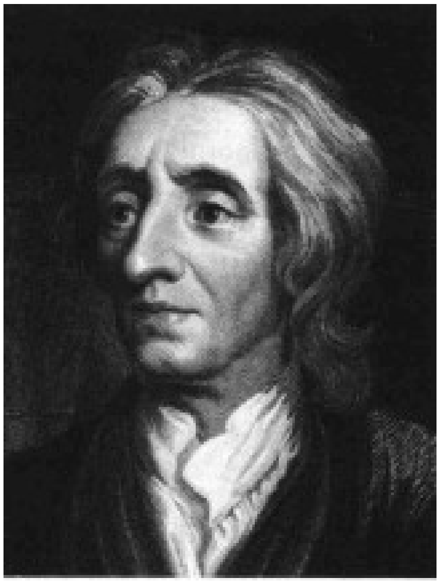
[Please contact the author with suggestions.]

Related Entries

ethics: deontological | justification, political: public | liberalism | social contract: contemporary approaches to

Fred D'Agostino
fdagosti@metz.une.edu.au

First published: February 27, 1996
Content last modified: July 28, 1997



John Locke

John Locke (b. 1632, d. 1704) was a British philosopher, Oxford academic and medical researcher, whose association with Anthony Ashley Cooper (later the First Earl of Shaftesbury) led him to become successively a government official charged with collecting information about trade and colonies, economic writer, opposition political activist, and finally a revolutionary whose cause ultimately triumphed in the Glorious Revolution of 1688. Much of Locke's work is characterized by opposition to authoritarianism. This opposition is both on the level of the individual person and on the level of institutions such as government and church. For the individual, Locke wants each of us to use reason to search after truth rather than simply accept the opinion of authorities or be subject to superstition. He wants us to proportion assent to propositions to the evidence for them. On the level of institutions it becomes important to distinguish the legitimate from the illegitimate functions of institutions and to make the corresponding distinction for the uses of force by these institutions. The positive side of Locke's anti-authoritarianism is that he believes that using reason to try to grasp the truth, and determining the legitimate functions of institutions will optimize human flourishing for the individual and society both in respect to its material and spiritual welfare. This in turn, amounts to following natural law and the fulfillment of the divine purpose for humanity. Locke's monumental *An Essay Concerning Human Understanding* concerns itself with determining the limits of human understanding in respect to God, the self, natural kinds and artifacts, as well as a variety of different kinds of ideas. It thus tells us in some detail what one can legitimately claim to know and what one cannot. Locke also wrote a variety of important political, religious and educational works including the *Two Treatises of Civil Government*, the *Letters Concerning Toleration*, *The Reasonableness of Christianity* and *Some Thoughts Concerning Education*.

1. Historical Background and Locke's Life

John Locke (1632-1704) was one of the greatest philosophers in Europe at the end of the seventeenth century. Locke grew up and lived through one of the most extraordinary centuries of English political and intellectual history. It was a century in which conflicts between Crown and Parliament and the overlapping conflicts between Protestants, Anglicans and Catholics swirled into civil war in the 1640s. With the defeat and death of Charles I, there began a great experiment in governmental institutions including the abolishment of the monarchy, the House of Lords and the Anglican church, and the establishment of Oliver Cromwell's Protectorate in the 1650s. The collapse of the Protectorate after the death of Cromwell was followed by the Restoration of Charles II -- the return of the monarchy, the House of Lords and the Anglican Church. This period lasted from 1660 to 1688. It was marked by continued conflicts between King and Parliament and debates over religious toleration for Protestant dissenters and Catholics. This period ends with the Glorious Revolution of 1688 in which James II was driven from England and replaced by William of Orange and his wife Mary. The final period during which Locke lived involved the consolidation of power by William and Mary, and the beginning of William's efforts to oppose the domination of Europe by the France of Louis XIV, which later culminated in the military victories of the John Churchill -- the Duke of Marlborough.

1.1 Locke's Life up to His Meeting with Lord Ashley in 1666

Locke was born in Wrington to Puritan parents of modest means. His father was a country lawyer who served in a cavalry company on the Puritan side in the early stages of the English civil war. His father's commander, Alexander Popham, became the local MP, and it was his patronage which allowed the young John Locke to gain an excellent education. In 1647 Locke went to Westminster School in London. The importance of Westminster school in the intellectual life of the seventeenth century can scarcely be exaggerated. Locke was a King's Scholar. The King's Scholars were a small group of special boys who had the privilege of living in the school and who received a stipend for two or three years before standing

for election for either Christ Church, Oxford or Trinity College Cambridge. While the "major elections" were probably political, the "minor elections or "challenges" were among the most genuinely competitive admissions processes in English schools of the period. Locke did not succeed in the challenge until 1650.

From Westminster school he went to Christ Church, Oxford, in the autumn of 1652 at the age of twenty. As Westminster school was the most important English school, so Christ Church was the most important Oxford college. Education at Oxford was medieval. Reform came, but not in Locke's time there. The three and a half years devoted to getting a B.A. was mainly given to logic and metaphysics and the classical languages. Conversations with tutors, even between undergraduates in the Hall were in Latin. Locke, like Hobbes before him, found the Aristotelian philosophy he was taught at Oxford of little use. There was, however, more at Oxford than Aristotle. The new experimental philosophy had arrived. John Wilkins, Cromwell's brother in law, had become Warden of Wadham College. The group around Wilkins was the nucleus of what was to become the English Royal Society. The Society grew out of informal meetings and discussion groups and moved to London after the Restoration and became a formal institution in the 1660s with charters from Charles II. The Society saw its aims in contrast with the Scholastic/Aristotelian traditions that dominated the universities. The program was to study nature rather than books.[1] Many of Wilkins associates were people interested in pursuing medicine by observation rather than the reading of classic texts. Bacon's interest in careful experimentation and the systematic collection of facts from which generalizations could be made was characteristic of this group. One of Locke's friends from Westminster school, Richard Lower, introduced Locke to medicine and the experimental philosophy being pursued by the virtuosi at Wadham.

Locke received his B.A. in February 1656. His career at Oxford, however, continued beyond his undergraduate days. In June of 1658 Locke qualified as a Master of Arts and was elected a Senior Student of Christ Church College. The rank was equivalent to a Fellow at any of the other colleges, but was not permanent. Locke had yet to determine what his career was to be. Locke was elected Lecturer in Greek at Christ Church in December of 1660 and he was elected Lecturer in Rhetoric in 1663. At this point, Locke needed to make a decision. The statutes of Christ Church laid it down that fifty five of the senior studentships should be reserved for men in orders or reading for orders. Only five could be held by others, two in medicine, two in law and one in moral philosophy. Thus, there was good reason for Locke to become a clergyman. Locke decided to become a doctor.

John Wilkins had left Oxford with the Restoration of Charles II. The new leader of the Oxford scientific group was Robert Boyle. He was also Locke's scientific mentor. Boyle (with the help of his astonishing assistant Robert Hooke) built an air pump which led to the formulation of Boyle's law and devised a barometer as a weather indicator. Boyle was, however, most influential as a theorist. He was a mechanical philosopher who treated the world as reducible to matter in motion. Locke read Boyle before he read Descartes. When he did read Descartes, he saw the great French philosopher as providing a viable alternative to the sterile Aristotelianism he had been taught at Oxford. In writing *An Essay Concerning Human Understanding* Locke adopted Descartes' ‘way of ideas’; though it is transformed so as to become an organic part of Locke's philosophy. Still, while admiring Descartes, Locke's involvement with the Oxford scientists gave him a perspective which made him critical of the rationalist elements in Descartes' philosophy.

In the Epistle to the Reader at the beginning of the *Essay* Locke remarks:

> The commonwealth of learning is not at this time without master-builders, whose mighty designs, in advancing the sciences, will leave lasting monuments to the admiration of posterity: but every one must not hope to be a Boyle or a Sydenham; and in an age that produces such masters as the great Huygenius and the incomparable Mr. Newton, with some others of that strain, it is ambition enough to be employed as an under-labourer in clearing the ground a little, and removing some of the rubbish that lies in the way to knowledge ... (pp. 9-10. All quotations are from the Nidditch edition of *An Essay Concerning Human Understanding*.)

Locke knew all of these men and their work. Locke, Boyle and Newton were all founding or early members of the English Royal Society. It is from Boyle that Locke learned about atomism (or the corpuscular hypothesis) and it is from Boyle's book *The Origin of Forms and Qualities* that Locke took the language of primary and secondary qualities. Sydenham was one of the most famous English physicians of the 17th century and Locke did medical research with him. Locke read Newton's *Principia Mathematica Philsophiae Naturalis* while in exile in Holland, and consulted Huygens as to the soundness of its mathematics. Locke and Newton became friends after Locke's return from Holland in 1688. It may be that in referring to himself as an ‘under-labourer’, Locke is not only displaying a certain literary modesty, he is contrasting the positive discoveries of these men, with his own attempt to show the inadequacies of the Aristotelian and Scholastic and to some degree the Cartesian philosophies. There are, however, many aspects of Locke's project to which this image of an under-labourer does not do justice. (See Jolley 1999, pp. 15-17) While the corpuscular philosophy and Newton's discoveries clearly influenced Locke, it is the Baconian program of producing natural histories that Locke makes reference to when he talks about the *Essay* in the Introduction. He writes:

> It shall suffice to my present Purpose, to consider the discerning Faculties of a Man, as they are employ'd about the Objects, which they have to do with: and I shall imagine that I have not wholly misimploy'd my self in the Thoughts I shall have on this Occasion, if in this Historical, Plain Method, I can give any Account of the Ways, whereby our Understanding comes to attain those Notions of Things, and can set down any Measure of the Certainty of our Knowledge... (I. 1. 2., pp. 43-4 -- the three numbers, are book, chapter and section numbers respectively, followed by the page number in the Nidditch edition.)

The ‘Historical, Plain Method’ is apparently to give a genetic account of how we come by our ideas. Presumably this will reveal the degree of certainty of the knowledge based on such ideas. Locke's own active involvement with the scientific movement was largely through his informal studies of medicine. Dr. David Thomas was his friend and collaborator. Locke and Thomas had a laboratory in Oxford which was very likely, in effect, a pharmacy. In 1666 Locke had a fateful meeting with Lord Ashley as a result of his friendship with Thomas. Ashley, one of the richest men in England, came to Oxford. He proposed to drink some medicinal waters there. He had asked Dr. Thomas to provide them. Thomas had to be out

of town and asked Locke to see that the water was delivered. Locke met Ashley and they liked one another. As a result of this encounter, Ashley invited Locke to come to London as his personal physician. In 1667 Locke did move to London becoming not only Lord Ashley's personal physician, but secretary, researcher, political operative and friend. Living with him Locke found himself at the very heart of English politics in the 1670s and 1680s.

1.2 Locke and Lord Shaftesbury 1666 to 1688

Locke's chief work while living at Lord Ashley's residence, Exeter House, in 1668 was his work as secretary of the Board of Trade and Plantations and Secretary to the Lords Proprietors of the Carolinas. Lord Ashley was one of the advocates of the view that England would prosper through trade and that colonies could play an important role in promoting trade. Ashley persuaded Charles II to create a Board of Trade and Plantations to collect information about trade and colonies, and Locke became its secretary. In his capacity as the secretary of the Board of Trade Locke was the collection point for information from around the globe about trade and colonies for the English government. Among Ashley's commercial projects was an effort to found colonies in the Carolinas. In his capacity as the secretary to the Lords Proprietors, Locke was involved in the writing of the fundamental constitution of the Carolinas. There is some controversy about the extent of Locke's role in writing the constitution.[2] In addition to issues about trade and colonies, Locke was involved through Shaftesbury in other controversies about public policy. There was a monetary crisis in England involving the value of money, and the clipping of coins. Locke wrote papers for Lord Ashley on economic matters, including the coinage crisis.

While living in London at Exeter House, Locke continued to be involved in philosophical discussions. He tells us that:

> Were it fit to trouble thee with the history of this Essay, I should tell thee, that five or six friends meeting at my chamber, and discoursing on a subject very remote from this, found themselves quickly at a stand, by the difficulties that rose on every side. After we had awhile puzzled ourselves, without coming any nearer a resolution of those doubts which perplexed us, it came into my thoughts that we took a wrong course; and that before we set ourselves upon inquiries of that nature, it was necessary to examine our own abilities, and see what objects our understandings were, or were not, fitted to deal with. This I proposed to the company, who all readily assented; and thereupon it was agreed that this should be our first inquiry. Some hasty and undigested thoughts, on a subject I had never before considered, which I set down against our next meeting, gave the first entrance into this Discourse; which having been thus begun by chance, was continued by intreaty; written by incoherent parcels; and after long intervals of neglect, resumed again, as my humour or occasions permitted; and at last, in a retirement where an attendance on my health gave me leisure, it was brought into that order thou now seest it. (Epistle to the Reader, p. 7)

James Tyrrell, one of Locke's friends was at that meeting. He recalls the discussion being about the principles of morality and revealed religion. (Cranston, 1957, pp. 140-1) Thus the Oxford scholar and

medical researcher came to begin the work which was to occupy him off and on over the next twenty years.

In 1674 after Shaftesbury had left the government, Locke went back to Oxford, where he acquired the degree Bachelor of medicine, and a license to practice medicine, and then went to France. (Cranston, 1957. p. 160) In France Locke went from Calais to Paris, Lyons and on to Montpellier, where he spent the next fifteen months. Much of Locke's time was spent learning about Protestantism in France. The Edict of Nantes was in force, and so there was a degree of religious toleration in France. Louis XIV was to revoke the edict in 1685 and French Protestants were then killed or forced into exile.

While Locke was in France, Shaftesbury's fortunes fluctuated. In 1676 Shaftesbury was imprisoned in the tower. His imprisonment lasted for a year. In 1678, after the mysterious murder of a London judge, informers (most notably Titus Oates) started coming forward to reveal a supposed Catholic conspiracy to assassinate the King and put his brother on the throne. This whipped up public anti-Catholic frenzy and gave Shaftesbury a wide base of public support for excluding James, Duke of York from the throne. Though Shaftesbury had not fabricated the conspiracy story, nor did he prompt Oates to come forward, he did exploit the situation to the advantage of his party. In the public chaos surrounding the sensational revelations, Shaftesbury organized an extensive party network, exercised great control over elections, and built up a large parliamentary majority. His strategy was to secure the passage of an Exclusion bill that would prevent Charles II's Catholic brother from becoming King. Although the Exclusion bill passed in the Commons it was rejected in the House of Lords because of the King's strong opposition to it. As the panic over the Popish plot receded, Shaftesbury was left without a following or a cause. Shaftesbury was seized on July 21, 1681 and again put in the tower. He was tried on trumped-up charges of treason but acquitted by a London grand jury (filled with his supporters) in November.

At this point some of the Country Party leaders began plotting an armed insurrection which, had it come off, would have begun with the assassination of Charles and his brother on their way back to London from the races at Newmarket. The chances of such a rising occurring were not as good as the plotters supposed. Memories of the turmoil of the civil war were still relatively fresh. Eventually Shaftesbury, who was moving from safe house to safe house, gave up and fled to Holland in November 1682. He died there in January 1683. Locke stayed in England until the Rye House Plot (named after the house from which the plotters were to fire upon the King and his brother) was discovered. He took ship for Holland that very week.

While in exile Locke finished *An Essay Concerning Human Understanding* and published a fifty page advanced notice of it in French. (This was to provide the intellectual world on the continent with most of their information about the *Essay* until Pierre Coste's French translation appeared.) He also wrote and published his *Epistola de Tolerentia* in Latin. Recent scholarship suggests that while in Holland Locke was not only finishing *An Essay Concerning Human Understanding* and nursing his health, he was closely associated with the English revolutionaries in exile. The English government was much concerned with this group. They tried to get a number of them, including Locke, extradited to England. Locke's studentship at Oxford was taken away from him. In the meanwhile, the English intelligence service infiltrated the rebel group in Holland and effectively thwarted their efforts -- at least for a while.

While Locke was living in exile in Holland, Charles II died on Feb. 6, 1685 and was succeeded by his brother -- who became James II of England. Soon after this the rebels in Holland sent a force of soldiers under the Duke of Monmouth to England to try to overthrow James II. Because of the excellent work of the Stuart spies, the government knew where the force was going to land before the troops on the ships did. The revolt was crushed, Monmouth captured and executed.(Ashcraft, 1986)

Ultimately, however, the rebels were successful. James II alienated most of his supporters and William of Orange was invited to bring a Dutch force to England. After William's army landed, James II realizing that he could not mount an effective resistance, fled the country to exile in France. This became known as the Glorious Revolution of 1688. It is an watershed in English history. For it marks the point at which the balance of power in the English government passed from the King to the Parliament. Locke returned to England in 1688 on board the royal yacht, accompanying Princess Mary on her voyage to join her husband.

1.3 The End of Locke's Life 1689-1704

After his return from exile, Locke published *An Essay Concerning Human Understanding* and *The Two Treatises of Government*. In addition, Popple's translation of Locke's *A Letter Concerning Toleration* was also published. It is worth noting that the *Two Treatises* and the *Letter Concerning Toleration* were published anonymously. Locke took up residence in the country at Oates in Essex, the home of Sir Francis and Lady Masham (Damaris Cudworth). Locke had met Damaris Cudworth in 1682 and became involved intellectually and romantically with her. She was the daughter of Ralph Cudworth, the Cambridge Platonist, and a philosopher in her own right. After Locke went into exile in Holland in 1683, she married Sir Francis Masham. Locke and Lady Masham remained good friends and intellectual companions to the end of Locke's life. During the remaining years of his life Locke oversaw four more editions of the *Essay* and engaged in controversies over the Essay most notably in a series of published letters with Edward Stillingfleet, Bishop of Worcester. In a similar way, Locke defended the *Letter Concerning Toleration* against a series of attacks. He wrote *The Reasonableness of Christianity* and *Some Thoughts on Education* during this period as well.

Nor was Locke finished with public affairs. In 1696 the Board of Trade was revived. Locke played an important part in its revival and served as the most influential member on it until 1700. The Board of Trade was, in Peter Laslett's phrase "... the body which administered the United States before the American revolution." (Laslett in Yolton 1990 p. 127) The board was, in fact, concerned with a wide range of issues, from the Irish wool trade and the suppression of piracy, to the governance of the colonies and the treatment of the poor in England. During these last eight years of his life, Locke was asthmatic, and he suffered so much from it that he could only bear the smoke of London during the four warmer months of the year. Locke plainly engaged in the activities of the Board out of a strong sense of patriotic duty. After his retirement from the Board of Trade in 1700, Locke remained in retirement at Oates until his death on Sunday 28 October 1704.

2. The Limits of Human Understanding

Locke is often classified as the first of the great English empiricists (ignoring the claims of Bacon and Hobbes). This reputation rests on Locke's greatest work, the monumental *An Essay Concerning Human Understanding.* Locke explains his project in several places. Perhaps the most important of his goals is to determine the limits of human understanding. Locke writes:

> For I thought that the first Step towards satisfying the several Enquiries, the Mind of Man was apt to run into, was, to take a Survey of our own Understandings, examine our own Powers, and see to what Things they were adapted. Till that was done, I suspected that we began at the wrong end, and in vain sought for Satisfaction in a quiet and secure Possession of Truths, that most concern'd us whilst we let loose our Thoughts into the vast Ocean of *Being,*as if all the boundless Extent, were the natural and undoubted Possessions of our Understandings, wherein there was nothing that escaped its Decisions, or that escaped its Comprehension. Thus Men, extending their Enquiries beyond their Capacities, and letting their Thoughts wander into those depths where they can find no sure Footing; 'tis no Wonder, that they raise Questions and multiply Disputes, which never coming to any clear Resolution, are proper to only continue and increase their Doubts, and to confirm them at last in a perfect Skepticism. Wheras were the Capacities of our Understanding well considered, the Extent of our Knowledge once discovered, and the Horizon found, which sets the boundary between the enlightened and the dark Parts of Things; between what is and what is not comprehensible by us, Men would perhaps with less scruple acquiesce in the avow'd Ignorance of the one; and employ their Thoughts and Discourse, with more Advantage and Satisfaction in the other. (I.1.7., p. 47)

Some philosophers before Locke had suggested that it would be good to find the limits of the Understanding, but what Locke does is to carry out this project in detail. In the four books of the *Essay* Locke considers the sources and nature of human knowledge. Book I argues that we have no innate knowledge. (In this he resembles Berkeley and Hume, and differs from Descartes and Leibniz.) So, at birth, the human mind is a sort of blank slate on which experience writes. In Book II Locke claims that ideas are the materials of knowledge and all ideas come from experience. The term ‘idea,’ Locke tells us "...stands for whatsoever is the Object of the Understanding, when a man thinks." (Essay I, 1, 8, p. 47) Experience is of two kinds, sensation and reflection. One of these -- sensation -- tells us about things and processes in the external world. The other -- reflection -- tells us about the operations of our own minds. Reflection is a sort of internal sense that makes us conscious of the mental processes we are engaged in. Some ideas we get only from sensation, some only from reflection and some from both.

Locke has an atomic or perhaps more accurately a corpuscular theory of ideas.[3] There is, that is to say, an analogy between the way atoms or corpuscles combine into complexes to form physical objects and the way ideas combine. Ideas are either simple or complex. We cannot create simple ideas, we can only get them from experience. In this respect the mind is passive. Once the mind has a store of simple ideas, it can combine them into complex ideas of a variety of kinds. In this respect the mind is active. Thus, Locke subscribes to a version of the empiricist axiom that there is nothing in the intellect that was not

previously in the senses -- where the senses are broadened to include reflection. Book III deals with the nature of language, its connections with ideas and its role in knowledge. Book IV, the culmination of the previous reflections, explains the nature and limits of knowledge, probability, and the relation of reason and faith. Let us now consider the *Essay* in some detail.

2.1 Book I

At the beginning of *An Essay Concerning Human Understanding* Locke says that since his purpose is "to enquire into the Original, Certainty and Extant of human knowledge, together with the grounds and degrees of Belief, Opinion and Assent" he is going to begin with ideas -- the materials out of which knowledge is constructed. His first task is to "enquire into the Original of these Ideas...and the ways whereby the Understanding comes to be furnished with them." (I. 1. 3. p. 44) The role of Book I of the *Essay* is to make the case that being innate is not a way in which the understanding is furnished with principles and ideas. Locke treats innateness as an empirical hypothesis and argues that there is no good evidence to support it.

Locke describes innate ideas as "some primary notions...Characters as it were stamped upon the Mind of Man, which the Soul receives in its very first Being; and brings into the world with it." (I. 2. 1. p. 48) In pursuing this enquiry, Locke rejects the claim that there are speculative innate principles (I. Chapter 2), practical innate moral principles (I. Chapter 3) or that we have innate ideas of God, identity or impossibility (I. Chapter 4). Locke rejects arguments from universal assent and attacks dispositional accounts of innate principles. Thus, in considering what would count as evidence from universal assent to such propositions as "What is, is" or "It is impossible for the same thing to be and not to be" he holds that children and idiots should be aware of such truths if they were innate but that they "have not the least apprehension or thought of them." Why should children and idiots be aware of and able to articulate such propositions? Locke says: "It seems to me a near Contradiction to say that there are truths imprinted on the Soul, which it perceives or understands not; imprinting if it signify anything, being nothing else but the making certain Truths to be perceived." (I. 2. 5., p. 49). So, Locke's first point is that if propositions were innate they should be immediately perceived -- by infants and idiots (and indeed everyone else) -- but there is no evidence that they are. Locke then proceeds to attack dispositional accounts that say, roughly, that innate propositions are capable of being perceived under certain circumstances. Until these circumstances come about the propositions remain unperceived in the mind. With the advent of these conditions, the propositions are then perceived. Locke gives the following argument against innate propositions being dispositional:

> For if any one [proposition] may [be in the mind but not be known]; then, by the same Reason, all Propositions that are true, and the Mind is ever capable of assenting to, may be said to be in the Mind, and to be imprinted: since if any one can be said to be in the Mind, which it never yet knew, it must be only because it is capable of knowing it; and so the Mind is of all Truths it ever shall know. (I. 2. 5., p. 50)

The essence of this argument and many of Locke's other arguments against dispositional accounts of

innate propositions is that such dispositional accounts do not provide an adequate criterion for distinguishing innate propositions from other propositions that the mind may come to discover. Thus, even if some criterion is proposed, it will turn out not to do the work it is supposed to do. For example Locke considers the claim that innate propositions are discovered and assented to when people "come to the use of Reason. (I. 2. 6., p. 51) Locke considers two possible meanings of this phrase. One is that we use reason to discover these innate propositions. Here he argues that the criterion is inadequate because it would not distinguish axioms from theorems in mathematics. Presumably the theorems are not innate while the axioms should be. But if both need to be discovered by reason, then there is no distinction between them. Nor will it do to say that one class (the axioms) are assented to as soon as perceived while the others are not. To be assented to as soon as perceived is a mark of certainty, but not of innateness. Locke also objects that truths that need to be discovered by reason could never be thought to be innate. The second possible meaning of "come to the use of reason" is that we discover these ideas at the time we come to use reason, but that we do not use reason to do so. He argues that this claim simply is not true. We know that children acquire such propositions before they acquire the use of reason, while others who are reasonable never acquire them.

When Locke turns from speculative principles to the question of whether there are innate practical moral principles, many of the arguments against innate speculative principles continue to apply, but there are some additional considerations. Practical principles, such as the Golden Rule, are not self-evident in the way such speculative principles as "What is, is" are. Thus, one can clearly and sensibly ask reasons for why one should hold the Golden Rule true or obey it. (I, 3. 4. p. 68) There are substantial differences between people over the content of practical principles. Thus, they are even less likely candidates to be innate propositions or to meet the criterion of universal assent. In the fourth chapter of Book I, Locke raises similar points about the ideas which compose both speculative and practical principles. The point is that if the ideas that are constitutive of the principles are not innate, this gives us even more reason to hold that the principles are not innate. He examines the ideas of identity, impossibility and God to make these points.

John Yolton has persuasively argued (Yolton, 1956) that the view that innate ideas and principles were necessary for the stability of religion, morality and natural law was widespread in England in the seventeenth century, and that in attacking both the naive account of innate ideas, Locke is attacking positions which were widely held and continued to be held after the publication of the *Essay*. Thus, the charge that Locke's account of innate ideas is made of straw, is not a just criticism. Whether the views of more important philosophers, Descartes before Locke or Leibniz after him escape the criticisms of innate ideas that Locke proposes lies beyond the scope of this article.

2.2 Book II

In Book II of the *Essay*, Locke gives his positive account of how we acquire the materials of knowledge. Locke distinguishes a variety of different kinds of ideas in Book II. Locke holds that the mind is a *tabula rasa* or blank sheet until experience in the form of sensation and reflection provide the basic materials -- simple ideas -- out of which most of our more complex knowledge is constructed. While the mind may

be a blank slate in regard to content, it is plain that Locke thinks we are born with a variety of faculties to receive and abilities to manipulate or process the content once we acquire it. Thus, for example, the mind can engage in three different types of action in putting simple ideas together. The first of these kinds of action is to combine them into complex ideas. Complex ideas are of two kinds, ideas of substances and ideas of modes. Substances are independent existences. Beings that count as substances include God, angels, humans, animals, plants and a variety of constructed things. Modes, are dependent existences. These include mathematical and moral ideas, and all the conventional language of religion, politics and culture. The second action which the mind performs is the bringing of two ideas, whether simple or complex, by one another so as to take a view of them at once, without uniting them This gives us our ideas of relations. (II. xii. 1., p. 163) The third act of the mind is the production of our general ideas by abstraction from particulars, leaving out the particular circumstances of time and place, which would limit the application of an idea to a particular individual. In addition to these abilities, there are such faculties as memory which allow for the storing of ideas.

Having set forth the general machinery of how simple and complex ideas of substances, modes, relations and so forth are derived from sensation and reflection Locke also explains how a variety of particular kinds of ideas, such as the ideas of solidity, number, space, time, power, identity, and moral relations arise from sensation and reflection. Several of these are of particular interest. Locke's chapter on power giving rise to a discussion of free will, voluntary action, and so forth, is of considerable interest. Some of these topics will be discussed in separate Encyclopedia entries. I have provided an account of Locke's views on personal identity and the immateriality of the soul in supplementary document:

> [Supplementary Document: The Immateriality of the Soul and Personal Identity]

In what follows, I focus on some central issues in Locke's account of physical objects.

Locke offers an account of physical objects based in the mechanical philosophy and the corpuscular hypothesis. The adherents of the mechanical philosophy held that all material phenomena can be explained by matter in motion and the impact of one body on another. They viewed matter as passive. They rejected the "occult qualities" and "causation at a distance" of the Aristotelian and Scholastic philosophy. The corpuscular hypothesis is that all matter is composed of particles. In the material world, all that exists are particles and the void or empty space in which the particles move. Some corupscularians held that corpuscles could be further divided. Atomists, on the other hand, held that there were indivisible or atomic particles.

Atoms have properties. They are extended, they are solid, they have a particular shape and they are in motion or rest. They combine together to produce the familiar stuff and physical objects, the gold and the wood, the horses and violets, the tables and chairs of our world. These familiar things also have properties. They are extended, solid, have a particular shape and are in motion and at rest. In addition to these properties that they share with the atoms that compose them, they have other properties such as colors, smells, tastes that they get by standing in relation to perceivers. The distinction between these two kinds of properties goes back to the Greek atomists. It is articulated by Galileo and Descartes as well as

Locke's mentor Robert Boyle.

Locke makes this distinction early in Book II of the Essay and using Boyle's terminology calls the two different classes of properties the primary and secondary qualities of an object. This distinction is made by both of the main branches of the mechanical philosophy of the seventeenth and early eighteenth century. Both the Cartesian plenum theorists, who held that the world was full of matter and that there was no void space, and the atomists such as Gassendi, who held that there were atoms and void space in which the atoms move, made the distinction between these two classes of properties. Locke accepted the corpuscular hypothesis as the most likely hypothesis. Thus, in the Chapter on Solidity Locke rejects the Cartesian definition of body as simply extended and argues that bodies are both extended and impenetrable or solid. The primary qualities of an object are properties which the object possesses independent of us -- such as occupying space, being either in motion or at rest, having texture. The secondary qualities are powers in bodies to produce in us ideas like color, taste, smell and so on that are caused by the interaction of our particular perceptual apparatus with these powers of the primary qualities of the object. Our ideas of primary qualities resemble the qualities in the object, while our ideas of secondary qualities do not resemble the powers that cause them. Locke also distinguishes tertiary properties that are the powers that one substance has to effect another, e.g. the power of a fire to melt a piece of wax.

There has been considerable scholarly debate concerning the details of Locke's account of the distinction. Among the issues are which qualities Locke assigns to each of the two categories. Locke gives several lists. Another issue is what the criterion is for putting a quality in one list rather than another. Does Locke hold that all the ideas of secondary qualities come to us by one sense while the ideas of primary qualities come to us through two or is Locke not making the distinction in this way? Another issue is whether on Locke's view primary qualities are perceptible at all. And while Locke claims our ideas of primary qualities resemble the primary qualities in objects, while the ideas of secondary qualities do not resemble their causes in the object, what does 'resemble' mean in this context? Related to this issue is how we are supposed to know about particles that we cannot sense. Maurice Mandlebaum called this process 'transdiction.' It seems clear that Locke holds that there are certain analogies between the middle sized macroscopic objects we encounter in the world, e.g. porphyry and manna for example, and the particles that compose these things. These analogies allow us to say certain things about the nature of particles and primary qualities, but may not get us very far in grasping the necessary connections between qualities in nature. Yet another issue is whether Locke sees the distinction as reductionistic -- that is whether only the primary qualities are real.

Locke probably holds some version of the representational theory of perception, though some scholars dispute even this. On such a theory what the mind immediately perceives are ideas, and the ideas are caused by and represent the objects which cause them. Thus perception is a triadic relation, rather than simply being a dyadic relation between an object and a perceiver. Such a dyadic relational theory is often called naive realism because it suggests that the perceiver is directly perceiving the object, and naive because this view is open to a variety of serious objections. Some versions of the representational theory are open to serious objections as well. If, for example, one makes ideas into things, then one can imagine that because one sees ideas, the ideas actually block one from seeing things in the external world. The

idea would be like a picture or painting. The picture would copy the original object in the external world, but because our immediate object of perception is the picture we would be prevented from seeing the original just as standing in front of a painting on an easel might prevent us from seeing the person being painted. Thus, this is sometimes called the picture/original theory of perception. Alternatively, Jonathan Bennett called it "the veil of perception". to emphasize that ’seeing’ the ideas prevents us from seeing the external world. One philosopher who arguably held such a view was Nicholas Malebranche, a follower of Descartes. Antoine Arnauld, by contrast, while believing in the representative character of ideas, is a direct realist about perception. Arnauld engaged in a lengthy controversy with Malebranche, and criticized Malebranche's account of ideas. Locke follows Arnauld in his criticism of Malebranche on this point. Yet Berkeley attributed the veil of perception interpretation of the representational theory of perception to Locke as have many later commentators including Bennett. A.D. Woozley puts the difficulty of doing this succinctly: "...it is scarcely credible both that Locke should be able to see and state so clearly the fundamental objection to the picture-original theory of sense perception, and that he should have held the same theory himself." Just what Locke's account of perception involves, is still a matter of scholarly debate.

Another issue that has been a matter of controversy since the first publication of the *Essay* is what Locke means by the term ‘substance’. The primary/secondary quality distinction gets us a certain ways in understanding physical objects, but Locke is puzzled about what underlies or supports the primary qualities themselves. He is also puzzled about what material and immaterial substances might have in common that would lead us to apply the same word to both. These kinds of reflections led him to the relative and obscure idea of substance in general. This is that "I know not what" which is the support of qualities which cannot subsist by themselves. We experience properties appearing in regular clumps, but we must infer that there is something that supports or perhaps ‘holds together’ those qualities. For we have no experience of that supporting substance. I think it is clear that Locke sees no alternative to the claim that there are substances supporting qualities. He does not, for example, have a theory of tropes (tropes are properties that can exist independently of substances) which he might use to dispense with the notion of substance. (In fact, he may be rejecting something like a theory of tropes when he rejects the Aristotelian doctrine of real qualities and insists on the need for substances.) He is thus not at all a skeptic about ‘substance’ in the way that Hume is. But, it is also quite clear that he is regularly insistent about the limitations of our ideas of substances. Bishop Stillingfleet accused Locke of putting substance out of the reasonable part of the world. But Locke is not doing that.

Since Berkeley, Locke's doctrine of the substratum or substance in general has been attacked as incoherent. It seems to imply that we have a particular without any properties, and this seems like a notion that is inconsistent with empiricism. In order to avoid this problem, Michael Ayers has proposed that we must understand the notions of ‘substratum’ and ‘substance in general’ in terms of Locke's doctrine of real essences developed in Book III of the Essay rather than as a separate problem from that of knowing real essences. The real essence of a material thing is its atomic constitution. This atomic constitution is the causal basis of all the observable properties of the thing. Were the real essence known, all the observable properties could be deduced from it. Locke claims that the real essences of material things are quite unknown to us. Locke's concept of substance in general is also a ‘something I know not what.’ Thus, on Ayers' interpretation ‘substance in general’ means something like ’whatever it is that

supports qualities' while the real essence means 'this particular atomic constitution that explains this set of observable qualities'. Thus, Ayers wants to treat the unknown substratum as picking out the same thing as the real essence -- thus eliminating the need for particulars without properties. This proposed way of interpreting Locke has been criticized by scholars both because of a lack of textural support, and on the stronger grounds that it conflicts with some things that Locke does say. (See Jolley 1999 pp. 71-3) As we have reached one of the important concepts in Book III, let us turn to that Book and Locke's discussion of language.

2.3 Book III

Locke devotes Book III of *An Essay Concerning Human Understanding* to language. This is a strong indication that Locke thinks issues about language were of considerable importance in attaining knowledge. At the beginning of the Book he notes the importance of abstract general ideas to knowledge. These serve as sorts under which we rank all the vast multitude of particular existences. Thus, abstract ideas and classification are of central importance in Locke's discussion of language.

There is a clear connection between Book II and III in that Locke claims that words stand for ideas. In his discussion of language Locke distinguishes words according to the categories of ideas established in Book II of the Essay. So there are ideas of substances, simple modes, mixed modes, relations and so on. It is in this context that Locke makes the distinction between real and nominal essences noted above. Perhaps because of his focus on the role that kind terms play in classification, Locke pays vastly more attention to nouns than to verbs. Locke recognizes that not all words relate to ideas. There are the many particles, words that "...signify the connexion that the Mind gives to Ideas, or Propositions, one with another. (II., 7. 1. p. 471) Still, it is the relation of words and ideas that gets most of Locke's attention in Book III.

Norman Kretzmann calls the claim that *'words in their primary or immediate signification signify nothing but the ideas in the mind of him that uses them'* Locke's main semantic thesis. (See Norman Kretzmann, ""The Main Thesis of Locke's Semantic Theory" in Tipton, 1977. pp. 123-140) This thesis has often been criticized as a classic blunder in semantic theory. Thus Mill, for example, wrote, "When I say, "the sun is the cause of the day," I do not mean that my idea of the sun causes or excites in me the idea of day." This criticism of Locke's account of language parallels the "veil of perception" critique of his account of perception and suggests that Locke is not distinguishing the meaning of a word from its reference. Kretzmann, however, argues persuasively that Locke distinguishes between meaning and reference and that ideas provide the meaning but not the reference of words. Thus, the line of criticism represented by the quotation from Mill is ill founded.

In addition to the kinds of ideas noted above, there are also particular and abstract ideas. Particular ideas have in them the ideas of particular places and times which limit the application of the idea to a single individual, while abstract general ideas leave out the ideas of particular times and places in order to allow the idea to apply to other similar qualities or things. There has been considerable philosophical and scholarly debate about the nature of the process of abstraction and Locke's account of it. Berkeley argued

that the process as Locke conceives it is incoherent. In part this is because Berkeley is an imagist -- that is he believes that all ideas are images. If one is an imagist it becomes impossible to imagine what idea could include both the ideas of a right and equilateral triangle. Michael Ayers has recently argued that Locke too was an imagist. This would make Berkeley's criticism of Locke very much to the point. Ayers' claim, however, has been disputed. The process of abstraction is of considerable importance to human knowledge. Locke thinks most words we use are general. (III, I. 1. p., 409) Clearly, it is only general or sortal ideas that can serve in a classificatory scheme.

In his discussion of names of substances and in the contrast between names of substances and names of modes, a number of interesting features of Locke's views about language and knowledge emerge. Physical substances are atoms and things made up of atoms. But we have no experience of the atomic structure of horses and tables. We know horses and tables mainly by secondary qualities such as color, taste and smell and so on and primary qualities such as shape and extension. So, since the real essence (the atomic constitution) of a horse is unknown to us, our word ‘horse’ cannot get its meaning from that real essence. What the general word signifies is the complex of ideas we have decided are parts of the idea of that sort of thing. These ideas we get from experience. Locke calls such a general idea that picks out a sort, the nominal essence of that sort.

One of the central issues in Book III has to do with classification. On what basis do we divide things into kinds and organize those kinds into a system of species and genera? In the Aristotelian and Scholastic tradition that Locke rejects, necessary properties are those that an individual must have in order to exist and continue to exist. These contrast with accidental properties. Accidental properties are those that an individual can gain and lose and yet continue in existence. If a set of necessary properties is shared by a number of individuals, that set of properties constitutes the essence of a natural kind. The aim of Aristotelian science is to discover the essences of natural kinds. Kinds can then be organized hierarchically into a classificatory system of species and genera. This classification of the world by natural kinds will be unique and privileged because it alone corresponds to the structure of the world. This doctrine of essences and kinds is often called Aristotelian essentialism. Locke rejects a variety of aspects of this doctrine. He rejects the notion that an individual has an essence apart from being treated as belonging to a kind. He also rejects the claim that there is a single classification of things in nature that the natural philosopher should seek to discover. He holds that there are many possible ways to classify the world each of which might be particularly useful depending on one's purposes.

Locke's pragmatic account of language and the distinction between nominal and real essences constitute an anti-essentialist alternative to this Aristotelian essentialism and its correlative account of the classification of natural kinds. He claims that there are no fixed boundaries in nature to be discovered -- that is there are no clear demarcation points between species. There are always borderline cases. There is scholarly debate over whether Locke's view is that this lack of fixed boundaries is true on both the level of appearances and nominal essences, and atomic constitutions and real essences, or on the level of nominal essences alone. The first view is that Locke holds that there are no natural kinds on either the level of appearance or atomic reality while the second view holds that Locke thinks there are real natural kinds on the atomic level, it is simply that we cannot get at them or know what they are. On either of these interpretations, the real essence cannot provide the meaning to names of substances.

By contrast, the ideas that we use to make up our nominal essences come to us from experience. Locke claims that the mind is active in making our ideas of sorts and that there are so many properties to choose among that it is possible for different people to make quite different ideas of the essence of a certain substance. This has given some commentators the impression that the making of sorts is utterly arbitrary and conventional for Locke and that there is no basis for criticizing a particular nominal essence. Sometimes Locke says things that might suggest this. But he also points out that the making of nominal essences is constrained both by usage (where words standing for ideas that are already in use) and by the fact that substance words are supposed to copy the properties of the substances they refer to.

Let us begin with the usage of words first. It is important that in a community of language users that words be used with the same meaning. If this condition is met it facilitates the chief end of language which is communication. If one fails to use words with the meaning that most people attach to them, one will fail to communicate effectively with others. Thus one would defeat the main purpose of language. It should also be noted that traditions of usage for Locke can be modified. Otherwise we would not be able to improve our knowledge and understanding by getting more clear and determinate ideas.

In the making of the names of substances there is a period of discovery as the abstract general idea is put together (e.g. the discovery of violets or gold) and then the naming of that idea and then its introduction into language. Language itself is viewed as an instrument for carrying out the mainly prosaic purposes and practices of every day life. Ordinary people are the chief makers of language.

> Vulgar Notions suit vulgar Discourses; and both though confused enough, yet serve pretty well for the Market and the Wake. Merchants and Lovers, Cooks and Taylors, have Words wherewith to dispatch their ordinary affairs; and so, I think, might Philosophers and Disputants too, if they had a mind to understand and to be clearly understood. (III. Xi. 10. p. 514)

These ordinary people use a few apparent qualities, mainly ideas of secondary qualities to make ideas and words that will serve their purposes.

Scientists come along later to try to determine if the connections between properties which the ordinary folk have put together in a particular idea in fact holds in nature. Scientists are seeking to find the necessary connections between properties. Still, even scientists, in Locke's view, are restricted to using observable (and mainly secondary) qualities to categorize things in nature. Sometimes, the scientists may find that the ordinary folk have erred, as when they called whales ‘fish’. A whale is not a fish, as it turns out, but a mammal. There is a characteristic group of qualities which fish have which whales do not have. There is a characteristic group of qualities which mammals have which whales also have. To classify a whale as a fish therefore is a mistake. Similarly, we might make an idea of gold that only included being a soft metal and gold color. If so, we would be unable to distinguish between gold and fool's gold. Thus, since it is the mind that makes complex ideas (they are ‘the workmanship of the understanding’), one is free to put together any combination of ideas one wishes and call it what one will.

But the product of such work is open to criticism, either on the grounds that it does not conform to already current usage, or that it inadequately represents the archetypes that it is supposed to copy in the world. We engage in such criticism in order to improve human understanding of the material world and thus the human condition.

The distinction between modes and substances is surely one of the most important in Locke's philosophy. In contrast with substances modes are dependent existences -- they can be thought of as the ordering of substances. These are technical terms for Locke, so we should see how they are defined. Locke writes: "First, *Modes* I call such complex *Ideas,* which however compounded, contain not in themselves the supposition of subsisting by themselves; such are the words signified by the Words *Triangle, Gratitude, Murther, etc."* (II. xii.4, p. 165) Locke goes on to distinguish between simple and mixed modes. He writes:

> Of these *Modes,* there are two sorts, which deserve distinct consideration. First, there are some that are only variations, or different combinations of the same simple *Idea,* without the mixture of any other, as a dozen or score; which are nothing but the *ideas* of so many distinct unities being added together, and these I call as being contained within the bounds of one simple *Idea.* Secondly, There are others, compounded of *Ideas* of several kinds, put together to make one complex one; v.g. *Beauty,* consisting of a certain combination of Colour and Figure, causing Delight to the Beholder; *Theft,* which being the concealed change of the Possession of any thing, without the consent of the Proprietor, contains, as is visible, a combination of several *Ideas* of several kinds; and these I call *Mixed Modes.* (II, xii. 5., p. 165)

When we make ideas of modes, the mind is again active, but the archetype is in our mind. The question becomes whether things in the world fit our ideas, and not whether our ideas correspond to the nature of things in the world. Our ideas are adequate. Thus we define ‘bachelor’ as an unmarried, adult, male human being. If we find that someone does not fit this definition, this does not reflect badly on our definition, it simply means that that individual does not belong to the class of bachelors. Modes give us the ideas of mathematics, of morality, of religion and politics and indeed of human conventions in general. Since these modal ideas are not only made by us but serve as standards that things in the world either fit or do not fit and thus belong or do not belong to that sort, ideas of modes are clear and distinct, adequate and complete. Thus in modes, we get the real and nominal essences combined. One can give precise definitions of mathematical terms (that is, give necessary and sufficient conditions) and one can give deductive demonstrations of mathematical truths. Locke sometimes says that morality too is capable of deductive demonstration. Though pressed by his friend William Molyneux to produce such a demonstrative morality, Locke never did so. The terms of political discourse have some of the same features for Locke. When Locke defines the states of nature, slavery and war in the *Second Treatise of Government,* for example, we are presumably getting precise modal definitions from which one can deduce consequences. It is possible, however, that with politics we are getting a study which requires both experience as well as the deductive modal aspect.

The extant of the influence that Locke’ account of language has had over the centuries is a matter of

scholarly debate. Norman Kretzmann holds that Locke's views, while not original had a powerful influence on the Enlightenment view of the connection of words and ideas. Noam Chomsky in *Cartesian Linguistics* traces the important ideas in linguistics back to Descartes and the school at Port Royal rather than Locke. This is largely a matter of the importance of the innate in Chomsky's thought. Hans Aarsleff, on the other hand, believes that Locke stands at the beginning of the developments that produced contemporary linguistics and Aarsleff argues that Chomsky's account is more polemical than historical.

2.4 Book IV

In the fourth book of *An Essay Concerning Human Understanding* Locke tells us what knowledge is and what humans can know and what they cannot (not simply what they do and do not happen to know). Locke defines knowledge as "the perception of the connexion and agreement or disagreement and repugnancy of any of our Ideas" (IV. I. 1. p. 525). This definition of knowledge contrasts with the Cartesian definition of knowledge as any ideas that are clear and distinct. Locke's account of knowledge allows him to say that we can know substances in spite of the fact that our ideas of them always include the obscure and relative idea of substance in general. Still, Locke's definition of knowledge raises in this domain a problem analogous to those we have seen with perception and language. If knowledge is the perception of the agreement or disagreement of any of our Ideas" -- are we not trapped in the circle of our own ideas? what about knowing the real existence of things? Locke is plainly aware of this problem, but how adequate his solution to it is is more doubtful.

What then can we know and with what degree of certainty? We can know that God exists with the second highest degree of assurance, that of demonstration. We also know that we exist with the highest degree of certainty. The truths of morality and mathematics we can know with certainty as well, because these are modal ideas whose adequacy is guaranteed by the fact that we make such ideas as ideal models which other things must fit, rather than trying to copy some external archetype which we can only grasp inadequately. On the other hand, our efforts to grasp the nature of external objects is limited largely to the connection between their apparent qualities. The real essence of elephants and gold is hidden from us: though in general we suppose them to be some distinct combination of atoms which cause the grouping of apparent qualities which leads us to see elephants and violets, gold and lead as distinct kinds. Our knowledge of material things is probabilistic and thus opinion rather than knowledge. Thus our "knowledge" of external objects is inferior to our knowledge of mathematics and morality, of ourselves, and of God. While Locke holds that we only have knowledge of a limited number of things, he thinks we can judge the truth or falsity of many propositions in addition to those we can legitimately claim to know. This brings us to a discussion of probability.

2.5 Knowledge and Probability

Knowledge involves the seeing of the agreement or disagreement of our ideas. What then is probability and how does it relate to knowledge? Locke writes:

> The Understanding Faculties being given to Man, not barely for Speculation, but also for

> the Conduct of his Life, Man would be at a great loss, if he had nothing to direct him, but what has the Certainty of true *Knowledge*... Therefore, as God has set some Things in broad day-light; as he has given us some certain Knowledge...So in the greater part of our Concernment, he has afforded us only the twilight, as I may say so, of Probability, suitable, I presume, to that State of Mediocrity and Probationership, he has been pleased to place us in here, wherein to check our over-confidence and presumption, we might by every day's Experience be made sensible of our short sightedness and liableness to Error...(IV, xiv, 1-2., p. 652)

So, apart from the few important things that we can know for certain, e.g. the existence of ourselves and God, the nature of mathematics and morality broadly construed, for the most part we must lead our lives without knowledge. What then is probability? Locke writes:

> As Demonstration is the shewing of the agreement or disagreement of two Ideas, by the intervention of one or more Proofs, which have a constant, immutable, and visible connexion one with another: so Probability is nothing but the appearance of such an Agreement or Disagreement, by the intervention of Proofs, whose connection is not constant and immutable, or at least is not perceived to be so, but is or appears, for the most part to be so, and is enough to induce the Mind to judge the Proposition to be true, or false, rather than the contrary. (IV., xv, 1., p. 654)

Probable reasoning, on this account, is an argument, similar in certain ways to the demonstrative reasoning that produces knowledge but different also in certain crucial respects. It is an argument that provides evidence that leads the mind to judge a proposition true or false but without a guarantee that the judgment is correct. This kind of probable judgment comes in degrees, ranging from near demonstrations and certainty to unlikeliness and improbability to near the vicinity of impossibility. It is correlated with degrees of assent ranging from full assurance down to conjecture, doubt and distrust.

The new science of mathematical probability had come into being on the continent just around the time that Locke was writing the *Essay*. His account of probability, however, shows little or no awareness of mathematical probability. Rather it reflects an older tradition that treated testimony as probable reasoning. Given that Locke's aim, above all, is to discuss what degree of assent we should give to various religious propositions, the older conception of probability very likely serves his purposes best. Thus, when Locke comes to describe the grounds for probability he cites the conformity of the proposition to our knowledge, observation and experience, and the testimony of others who are reporting their observation and experience. Concerning the latter we must consider the number of witnesses, their integrity, their skill in observation, counter testimony and so on. In judging rationally how much to assent to a probable proposition, these are the relevant considerations that the mind should review. We should, Locke also suggests, be tolerant of differing opinions as we have more reason to retain the opinions we have than to give them up to strangers or adversaries who may well have some interest in our doing so.

Locke distinguishes two sorts of probable propositions. The first of these have to do with particular existences or matters of fact, and the second that are beyond the testimony of the senses. Matters of fact are open to observation and experience, and so all of the tests noted above for determining rational assent to propositions about them are available to us. Things are quite otherwise with matters that are beyond the testimony of the senses. These include the knowledge of finite immaterial spirits such as angels or things such as atoms that are too small to be sensed, or the plants, animals or inhabitants of other planets that are beyond our range of sensation because of their distance from us. Concerning this latter category, Locke says we must depend on analogy as the only help for our reasoning. He writes: "Thus the observing that the bare rubbing of two bodies violently one upon the other, produce heat, and very often fire it self, we have reason to think, that what we call Heat and Fire consist of the violent agitation of the imperceptible minute parts of the burning matter..." (IV. xvi. 12. Pp 665-6) We reason about angels by considering the Great Chain of Being; figuring that while we have no experience of angels, the ranks of species above us is likely as numerous as that below of which we do have experience. This reasoning is, however, only probable.

2.6 Reason, Faith and Enthusiasm

The relative merits of the senses, reason and faith for attaining truth and the guidance of life were a significant issue during this period. As noted above James Tyrrell recalled that the original impetus for the writing of *An Essay Concerning Human Understanding* was a discussion about the principles of morality and revealed religion. In Book IV Chapters XVII, XVIII and XIX Locke deals with the nature of reason, the relation of reason to faith and the nature of enthusiasm. Locke remarks that all sects make use of reason as far as they can. It is only when this fails them that they have recourse to faith and claim that what is revealed is above reason. But he adds: "And I do not see how they can argue with anyone or even convince a gainsayer who uses the same plea, without setting down strict boundaries between faith and reason." (IV. xviii. 2. p. 689) Locke then defines reason as "the discovery of the certainty or probability of such propositions or truths, which the mind arrives at by deduction made from such ideas, as it has got by the use of its natural faculties; viz, by the use of sensation or reflection." (IV. xviii. ii. p. 689) Faith, on the other hand, is assent to any proposition "...upon the credit of the proposer, as coming from God, in some extraordinary way of communication." That is we have faith in what is disclosed by revelation and which cannot be discovered by reason. Locke also distinguishes between the *original revelation* by God to some person, and *traditional revelation* which is the original revelation "...delivered over to others in Words, and the ordinary ways of our conveying our Conceptions one to another. (IV. xviii, 3 p. 690)

Locke makes the point that some things could be discovered both by reason and by revelation -- God could reveal the propositions of Euclid's geometry, or they could be discovered by reason. In such cases there would be little use for faith. Traditional revelation can never produce as much certainty as the contemplation of the agreement or disagreement of our own ideas. Similarly revelations about matters of fact do not produce as much certainty as having the experience one self. Revelation, then cannot contradict what we know to be true. If it could, it would undermine the trustworthiness of all of our faculties. This would be a disastrous result. Where revelation comes into its own is where reason cannot

reach. Where we have few or no ideas for reason to contradict or confirm, this is the proper matters for faith. "...that Part of the Angels rebelled against GOD, and thereby lost their first happy state: and that the dead shall rise, and live again: These and the like, being Beyond the Discovery of Reason, are purely matters of Faith; with which Reason has nothing to do." (IV. xviii. 8. p. 694) Still, reason does have a crucial role to play in respect to revelation. Locke writes:

> Because the Mind, not being certain of the Truth of that it evidently does not know, but only yielding to the Probability that appears to it, is bound to give up its assent to such Testimony, which, it is satisfied, comes from one who cannot err, and will not deceive. But yet, it still belongs to Reason, to judge of the truth of its being a Revelation, and of the significance of the Words, wherein it is delivered. (IV. 18. 8., p. 694)

So, in respect to the crucial question of how we are to know whether a revelation is genuine, we are supposed to use reason and the canons of probability to judge. Locke claims that if the boundaries between faith and reason are not clearly marked, then there will be no place for reason in religion and one then gets all the "extravagant Opinions and Ceremonies, that are to be found in the religions of the world..." (IV. 18. 11. p. 696)

Should one accept revelation without using reason to judge whether it is genuine revelation or not, one gets what Locke calls a third principle of assent besides reason and revelation, namely enthusiasm. Enthusiasm is a vain or unfounded confidence in divine favor or communication. It implies that there is no need to use reason to judge whether such favor or communication is genuine or not. Clearly when such communications are not genuine they are 'the ungrounded Fancies of a Man's own Brain.'(IV. xix. 2. p. 698) This kind of enthusiasm was characteristic of Protestant extremists going back to the era of the Civil War. Locke was not alone in rejecting enthusiasm, but he rejects it in the strongest terms. Enthusiasm violates the fundamental principle by which the understanding operates -- that assent be proportioned to the evidence. To abandon that fundamental principle would be catastrophic. This is a point that Locke also makes in *The Conduct of the Understanding*,and *The Reasonableness of Christianity*. Locke wants each of us to use our understanding to search after truth. Of enthusiasts, those who would abandon reason and claim to know on the basis of faith alone, Locke writes: "...he that takes away Reason to make way for Revelation, puts out the Light of both, and does much what the same, as if he would perswade a Man to put out his eyes, the better to receive the remote Light of an invisible Star by a Telescope." (IV. xix. 4. p. 698) Rather than engage in the tedious labor required to reason correctly, enthusiasts persuade themselves that they are possessed of immediate revelation, without having to use reason to judge of the genuineness of their revelation. This leads to "odd Opinions and extravagant actions" that are characteristic of enthusiasm and which should warn that this is a wrong principle. Thus, Locke strongly rejects any attempt to make inward persuasion not judged by reason a legitimate principle.

I turn now to a consideration of Locke's political works.

3. The *Two Treatises Of Government*

The introduction of the work was written latter than the main text, and gave people the impression that the book was written in 1688 to justify the Glorious Revolution. We now know that the *Two Treatises of Government* were written during the Exclusion crisis and were probably intended to justify the general armed rising which the Country Party leaders were planning. It was a truly revolutionary work. Supposing that the *Two Treatises* may have been intended to explain and defend the revolutionary plot against Charles II and his brother, how does it do this?

The First Treatise of Government is a polemical work aimed at refuting the patriarchal version of the Divine Right of Kings doctrine put forth by Sir Robert Filmer. Locke singles out Filmer's contention that men are not "naturally free" as the key issue, for that is the "ground" or premise on which Filmer erects his argument for the claim that all "legitimate" government is "absolute monarchy." -- kings being descended from the first man, Adam. Early in the First Treatise Locke denies that either scripture or reason supports Filmer's premise or arguments. In what follows, Locke minutely examines key Biblical passages.

The Second Treatise of Government provides Locke's positive theory of government - he explicitly says that he must do this "lest men fall into the dangerous belief that all government in the world is merely the product of force and violence." Locke's account involves several devices which were common in seventeenth and eighteenth century political philosophy -- natural rights theory and the social contract. Natural rights are those rights which we are supposed to have as human beings before ever government comes into being. We might suppose, that like other animals, we have a natural right to struggle for our survival. Locke will argue that we have a right to the means to survive. When Locke comes to explain how government comes into being, he uses the idea that people agree that their condition in the state of nature is unsatisfactory, and so agree to transfer some of their rights to a central government, while retaining others. This is the theory of the social contract. There are many versions of natural rights theory and the social contract in seventeenth and eighteenth century European political philosophy, some conservative and some radical. Locke's version belongs on the radical side of the spectrum. These radical natural right theories influenced the ideologies of the American and French revolutions.

3.1 *The Second Treatise of Government*

Here is the subject matter of the various chapters of the *Second Treatise:*

- Chapter 1 Book I: the definition of Political power
- Chapter II-VII: the bases of government, states of nature, war, slavery, the nature of property
- Chapters VIII-XIV: the nature of political power and legitimate civil government
- Chapter XV: recapitulates the fundamental distinctions between paternal. political and despotic power.
- Chapter XVI-XVIII: elaborates the nature of illegitimate civil government. It specifies three forms of such illegitimacy: 1. an unjust foreign conquest, 2. internal usurpation of political rule and 3. tyrannical extension of power by those who were originally legitimately in power.

- Chapter XIX: gives the conditions under which legitimate revolution may occur.

Figuring out what the proper or legitimate role of civil government is would be a difficult task indeed if one were to examine the vast complexity of existing civil governments. How should one proceed? One strategy is to consider what life is like in the absence of civil government. Presumably this is a simpler state, one which may be easier to understand. Then one might see what role civil government ought to play. This is the strategy which Locke pursues, following Hobbes and others. So, in the first chapter of the *Second Treatise* Locke defines political power.

> *Political power*, then, I take to be a *right* of making laws with penalties of death, and consequently all less penalties, for the regulating and preserving of property, and of employing the force of the community, in the execution of such laws, and in the defence of the common-wealth from foreign injury; and all this only for the public good

In the second chapter of *The Second Treatise* Locke describes the state in which there is no government with real political power. This is the state of nature. It is sometimes assumed that the state of nature is a state in which there is no government at all. This is only partially true. It is possible to have in the state of nature either no government, illegitimate government, or legitimate government with less than full political power.

If we consider the state of nature before there was government, it is a state of political equality in which there is no natural superior or inferior. From this equality flows the obligation to mutual love and the duties that people owe one another, and the great maxims of justice and charity. Was there ever such a state? There has been considerable debate about this. Still, it is plain that both Hobbes and Locke would answer this question affirmatively. Whenever people have not agreed to establish a common political authority, they remain in the state of nature. It's like saying that people are in the state of being naturally single until they are married. Locke clearly thinks one can find the state of nature in his time at least in the inland vacant parts of America and in the relations between different peoples. Perhaps the historical development of states also went though the stages of a state of nature. An alternative possibility is that the state of nature is not a real historical state, but rather a theoretical construct, intended to help determine the proper function of government. If one rejects the historicity of states of nature, one may still find them a useful analytical device. For Locke, it is very likely both.

3.2 Human Nature and God's Purposes

According to Locke, God created man and we are, in effect, God's property. The chief end set us by our creator as a species and as individuals is survival. A wise and omnipotent God, having made people and sent them into this world:

> ...by his order and about his business, they are his property whose workmanship they are, made to last during his, not one another's pleasure: and being furnished with like faculties, sharing all in one community of nature, there cannot be supposed any subordination among

> us, that may authorize us to destroy one another, as if we were made for one another's uses, as the inferior ranks of creatures are for our's.

It follows immediately that "he has no liberty to destroy himself, or so much as any creature in his possession, yet when some nobler use than its bare possession calls for it." (II. ii. 5) So, murder and suicide violate the divine purpose.

If one takes survival as the end, then we may ask what are the means necessary to that end. On Locke's account, these turn out to be life, liberty, health and property. Since the end is set by God, on Locke's view we have a right to the means to that end. So we have rights to life, liberty, health and property. These are natural rights, that is they are rights that we have in a state of nature before the introduction of civil government, and all people have these rights equally.

If God's purpose for me on earth is my survival and that of my species, and the means to that survival are my life, health, liberty and property -- then clearly I don't want anyone to violate my rights to these things. Equally, considering other people, who are my natural equals, I should conclude that I should not violate their rights to life, liberty, health and property. This is the law of nature. It is the Golden Rule, interpreted in terms of natural rights. Thus Locke writes: "The state of nature has a law of nature to govern it, which obliges everyone: and reason which is that law, teaches all mankind who will but consult it, that being all equal and independent, no one ought to harm another in his life, health, liberty or possessions..." (II, 6) Locke tells us that the law of nature is revealed by reason. Locke makes the point about the law that it commands what is best for us. If it did not, he says, the law would vanish for it would not be obeyed. It is in this sense, I think, that Locke means that reason reveals the law. If you reflect on what is best for yourself and others, given the goal of survival and our natural equality, you will come to this conclusion.

Locke does not intend his account of the state of nature as a sort of utopia. Rather it serves as an analytical device that explains why it becomes necessary to introduce civil government and what the legitimate function of civil government is. Thus, as Locke conceives it, there are problems with life in the state of nature. The law of nature, like civil laws can be violated. There are no police, prosecutors or judges in the state of nature as these are all representatives of a government with full political power. The victims, then, must enforce the law of nature in the state of nature. In addition to our other rights in the state of nature, we have the rights to enforce the law and to judge on our own behalf. We may, Locke tells us, help one another. We may intervene in cases where our own interests are not directly under threat to help enforce the law of nature. Still, the person who is most likely to enforce the law under these circumstances is the person who has been wronged. The basic principle of justice is that the punishment should be proportionate to the crime. But when the victims are judging the seriousness of the crime, they are more likely to judge it of greater severity than might an impartial judge. As a result, there will be regular miscarriages of justice. This is perhaps the most important problem with the state of nature.

In Chapters 3 and 4, Locke defines the states of war and slavery. The state of war is a state in which someone has a sedate and settled intention of violating someone's right to life. Such a person puts

themselves into a state of war with the person whose life they intend to take. In such a war the person who intends to violate someone's right to life is an unjust aggressor. This is not the normal relationship between people enjoined by the law of nature in the state of nature. Locke is distancing himself from Hobbes who had made the state of nature and the state of war equivalent terms. For Locke, the state of nature is ordinarily one in which we follow the Golden Rule interpreted in terms of natural rights, and thus love our fellow human creatures. The state of war only comes about when someone proposes to violate someone else's rights. Thus, on Locke's theory of war, there will always be an innocent victim on on side and an unjust aggressor on the other.

Slavery is the state of being in the absolute or arbitrary power of another. On Locke's definition of slavery there is only one rather remarkable way to become a legitimate slave. In order to do so one must be an unjust aggressor defeated in war. The just victor then has the option to either kill the aggressor or enslave them. Locke tells us that the state of slavery is the continuation of the state of war between a lawful conqueror and a captive, in which the conqueror delays to take the life of the captive, and instead makes use of him. This is a continued war because if conqueror and captive make some compact for obedience on the one side and limited power on the other, the state of slavery ceases. The reason that slavery ceases with the compact is that "no man, can, by agreement pass over to another that which he hath not in himself, a power over his own life." (II. 4, 24) Legitimate slavery is an important concept in Locke's political philosophy largely because it tells us what the legitimate extant of despotic power is and defines and illuminates by contrast the nature of illegitimate slavery. Illegitimate slavery is that state in which someone possesses absolute or despotic power over someone else without just cause. Locke holds that it is this illegitimate state of slavery which absolute monarchs wish to impose upon their subjects. It is very likely for this reason that legitimate slavery is so narrowly defined.

There have been a steady stream of articles over the last forty years arguing that given Locke's involvement with trade and colonial government, the theory of slavery in the *Second Treatise* was intended to justify the institutions and practices of Afro-American slavery. This seems quite unlikely. Had he intended to do so, Locke would have done much better with a vastly more inclusive definition of legitimate slavery than the one he gives. It is sometimes suggested that Locke's account of "just war" is so vague that it could easily be twisted to justify the institutions and practices of Afro-American slavery. This, however, is also not the case. In the Chapter "Of Conquest" Locke explicitly lists the limits of the legitimate power of conquerors. These limits on who can become a legitimate slave and what the powers of a just conqueror are ensure that this theory of conquest and slavery would condemn the institutions and practices of Afro-American slavery in the 17th, 18th and 19th centuries.

"Of Property" is one of the most famous, influential and important chapters in the *Second Treatise of Government.* Indeed, some of the most controversial issues about the *Second Treatise* come from varying interpretations of it. In this chapter Locke, in effect, describes the evolution of the state of nature to the point where it becomes expedient for those in it to found a civil government. So, it is not only an account of the nature and origin of private property, but leads up to the explanation of why civil government replaces the state of nature.

In discussing the origin of private property Locke begins by noting that God gave the earth to all men in

common. Thus there is a question about how private property comes to be. Locke finds it a serious difficulty. He points out, however, that we are supposed to make use of the earth "for the best advantage of life and convenience." (II. 5, 25) What then is the means to appropriate property from the common store? Locke argues that private property does not come about by universal consent. If one had to go about and ask everyone if one could eat these berries, one would starve to death before getting everyone's agreement. Locke holds that we have a property in our own person. And the labor of our body and the work of our hands properly belong to us. So, when one picks up acorns or berries, they thereby belong to the person who picked them up.

One might think that one could then acquire as much as one wished, but this is not the case. Locke introduces at least two important qualifications on how much property can be acquired. The first qualification has to do with waste. Locke writes: "As much as anyone can make use of to any advantage of life before it spoils, so much by his labor he may fix a property in; whatever is beyond this, is more than his share, and belongs to others." (II. v. 31) Since originally, populations were small and resources great, living within the bounds set by reason, there would be little quarrel or contention over property, for a single man could make use of only a very small part of what was available.

Note that Locke has, thus far, been talking about hunting and gathering, and the kinds of limitations which reason imposes on the kind of property that hunters and gatherers hold. In the next section he turns to agriculture and the ownership of land and the kinds of limitations there are on that kind of property. In effect, we see the evolution of the state of nature from a hunter/gatherer kind of society to that of a farming and agricultural society. Once again it is labor which imposes limitations upon how much land can be enclosed. It is only as much as one can work. But there is an additional qualification. Locke says:

> Nor was this *appropriation* of any parcel of *land*, by improving it, any prejudice to any other man, since there was still enough, and as good left; and more than the as yet unprovided could use. So that, in effect, there was never the less for others because of his inclosure for himself: for he that leaves as much as another can make use of, does as good as take nothing at all. No body could consider himself injured by the drinking of another man, though he took a good draught, who had a whole river of the same water left to quench his thirst: and the case of land and water, where there is enough, is perfectly the same. (II. v. 33)

The next stage in the evolution of the state of nature involves the introduction of money. Locke remarks that:

> . ... before the desire of having more than one needed had altered the intrinsic value of things, which depends only on their usefulness to the life of man; or had agreed, that a little piece of yellow metal, which would keep without wasting or decay, should be worth a great piece of flesh, or a whole heap of corn; though men had a right to appropriate by their labor, each one of himself, as much of the things of nature, as he could use; yet this could not be much, nor to the prejudice of others, where the same plenty was left to those

> who would use the same industry. (II. 5. 37.)

So, before the introduction of money, there was a degree of economic equality imposed on mankind both by reason and the barter system. And men were largely confined to the satisfaction of their needs and conveniences. Most of the necessities of life are relatively short lived -- berries, plums, venison and so forth. One could reasonably barter one's berries for nuts which would last not weeks but perhaps a whole year. And says Locke:

> ...if he would give his nuts for a piece of metal, pleased with its color, or exchange his sheep for shells, or wool for a sparkling pebble or diamond, and keep those by him all his life, he invaded not the right of others, he might heap up as much of these durable things as he pleased; the exceeding of the bounds of his property not lying in the largeness of his possessions, but the perishing of anything uselessly in it. (II. 5. 146.)

The introduction of money is necessary for the differential increase in property, with resulting economic inequality. Without money there would be no point in going beyond the economic equality of the earlier stage. In a money economy, different degrees of industry could give men vastly different proportions. "This partage of things in an inequality of private possessions, men have made practicable out of the bounds of society, and without compact, only by putting a value on gold and silver, and tacitly agreeing to the use of money: for in governments, the laws regulate the rights of property, and the possession of land is determined by positive constitutions." (II. 5. 50) The implication is that it is the introduction of money, which causes inequality, which in turn causes quarrels and contentions and increased numbers of violations of the law of nature. This leads to the decision to create a civil government. Before turning to the institution of civil government, however, we should ask what happens to the qualifications on the acquisition of property after the advent of money? One answer proposed by C. B. Macpherson is that the qualifications are completely set aside, and we now have a system for the unlimited acquisition of private property. This does not seem to be correct. It seems plain, rather, that at least the non-spoilage qualification is satisfied, because money does not spoil. The other qualifications may be rendered somewhat irrelevant by the advent of the conventions about property adopted in civil society. This leaves open the question of whether Locke approved of these changes. Macpherson, who takes Locke to be a spokesman for a proto-capitalist system, sees Locke as advocating the unlimited acquisition of wealth. According to James Tully, on the other side, Locke sees the new conditions, the change in values and the economic inequality which arise as a result of the advent of money, as the fall of man. Tully sees Locke as a persistent and powerful critic of self-interest. This remarkable difference in interpretation has been a significant topic for debates among scholars over the last forty years. Let us then, turn to the institution of civil government.

The institution of civil government comes about because of the difficulties in the state of nature. Rather clearly, on Locke's view, these difficulties increase with the increase in population, the decrease in available resources, and the advent of economic inequality which results from the introduction of money. These conditions lead to an increase in the number of violations of the natural law. Thus, the inconvenience of having to redress such grievances on one's own behalf become much more acute, since there are significantly more of them. These lead to the introduction of civil government.

3.3 The Social Contract Theory

Just as natural rights and natural law theory had a florescence in the 17th and 18th century, so did the social contract theory. Why is Locke a social contract theorist? Is it merely that this was one prevailing way of thinking about government at the time which Locke blindly adopted? I think the answer is that there is something about Locke's project which pushes him strongly in the direction of the social contract. One might hold that governments were originally instituted by force, and that no agreement was involved. Where Locke to adopt this view, he would be forced to go back on many of the things which are at the heart of his project in the *Second Treatise*. Remember that *The Second Treatise* provides Locke's positive theory of government, and that he explicitly says that he must do this "lest men fall into the dangerous belief that "all government in the world is merely the product of force and violence." So, while Locke might admit that some governments come about through force or violence, he would be destroying the most central and vital distinction, that between legitimate and illegitimate civil government, if he admitted that legitimate civil government can come about in this way. So, for Locke, legitimate civil government is instituted by the explicit consent of those governed. Those who make this agreement transfer to the civil government their right of executing the law of nature and judging their own case. These are the powers which they give to the central government, and this is what makes the justice system of civil governments a legitimate function of such governments.

Ruth Grant has persuasively argued that the establishment of civil government is in effect a two step process. Universal consent is necessary to form a political community. Consent to join a community once given is binding and cannot be withdrawn. This makes political communities stable. Grant writes: "Having established that the membership in a community entails the obligation to abide by the will of the community, the question remains: Who rules?" (Grant, 1987 p. 115) The answer to this question is determined by majority rule. The point is that universal consent is necessary to establish a political community, majority consent to answer the question who is to rule such a community. Universal consent and majority consent are thus different in kind, not just in degree. Grant writes:

> Locke's argument for the right of the majority is the theoretical ground for the distinction between duty to society and duty to government, the distinction that permits an argument for resistance without anarchy. When the designated government dissolves, men remain obligated to society acting through majority rule.

It is entirely possible for the majority to confer the rule of the community on a king and his heirs, or a group of oligarchs or on a democratic assembly. Thus, the social contract is not inextricably linked to democracy. Still, a government of any kind must perform the legitimate function of a civil government.

3.4 The Function Of Civil Government

Locke is now in a position to explain the function of a legitimate civil government and distinguish it from illegitimate civil government. The aim of such a legitimate civil government is to preserve, so far as

possible, the rights to life, liberty, health and property of its citizens, and to prosecute and punish those of its citizens who violate the rights of others others and to pursue the public good even where this may conflict with the rights of individuals. In doing this it provides something unavailable in the state of nature, an impartial judge to determine the severity of the crime, and to set a punishment proportionate to the crime. This is one of the main reasons why civil society is an improvement on the state of nature. An illegitimate civil government will fail to protect the rights to life, liberty, health and property of its subjects, and in the worst cases, such an illegitimate government will claim to be able to violate the rights of its subjects, that is it will claim to have despotic power over its subjects. Since Locke is arguing against the position of Sir Robert Filmer who held that patriarchal power and political power are the same, and that in effect these amount to despotic power, Locke is at pains to distinguish these three forms of power, and to show that they are not equivalent. Thus at the beginning of Chapter XV Of Paternal, Political and Despotic power considered together he writes: "THOUGH I have had occasion to speak of these before, yet the great mistakes of late about government, having as I suppose arisen from confounding these distinct powers one with another, it may not be amiss, to consider them together." Chapters VI and VII give Locke's account of paternal and political power respectively. Paternal power is limited. It lasts only through the minority of children, and has other limitations. Political power, derived as it is from the transfer of the power of individuals to enforce the law of nature, has with it the right to kill in the interest of preserving the rights of the citizens or otherwise supporting the public good. Despotic power, by contrast, implies the right to take the life, liberty, health and at least some of the property of any person subject to such a power.

3.5 Rebellion and Regicide

At the end of the Second Treatise we learn about the nature of illegitimate civil governments and the conditions under which rebellion and regicide are legitimate and appropriate. As noted above, scholars now hold that the book was written during the Exclusion crisis, and may have been written to justify a general insurrection and the assassination of the king of England and his brother. The argument for legitimate revolution follows from making the distinction between legitimate and illegitimate civil government. A legitimate civil government seeks to preserve the life, health, liberty and property of its subjects, insofar as this is compatible with the public good. Because it does this it deserves obedience. An illegitimate civil government seeks to systematically violate the natural rights of its subjects. It seeks to make them illegitimate slaves. Because an illegitimate civil government does this, it puts itself in a state of nature and a state of war with its subjects. The magistrate or king of such a state violates the law of nature and so makes himself into a dangerous beast of prey who operates on the principle that might makes right, or that the strongest carries it. In such circumstances, rebellion is legitimate as is the killing of such a dangerous beast of prey. Thus Locke justifies rebellion and regicide (regarded by many during this period as the most heinous of crimes) under certain circumstances. Presumably this was the justification that was going to be offered for the killing of the King of England and his brother had the Rye House Plot succeeded.

4. Locke and Religious Toleration

In England itself, religious conflict dominated the 17th century, contributing in important respects to the coming of the English civil war, and the abolishing of the Anglican Church during the Protectorate. After the Restoration of Charles II, Anglicans in parliament passed laws which repressed both Catholics and Protestant sects such as Presbyterians, Baptists, Quakers and Unitarians who did not agree with the doctrines or practices of the state Church. Of these various dissenting sects, some were closer to the Anglicans, others more remote. One reason among others why King Charles may have found Shaftesbury useful was that they were both concerned about religious toleration. They parted when it became clear that the King was mainly interested in toleration for Catholics, and Shaftesbury of Protestant dissenters.

One widely discussed strategy for reducing religious conflict in England was called comprehension. The idea was to reduce the doctrines and practices of the Anglican church to a minimum so that most, if not all, of the dissenting sects would be included in the state church. For those which even this measure would not serve, there was to be toleration. Toleration we may define as a lack of state persecution. Neither of these strategies made much progress during the course of the Restoration.

What were Locke's religious views and where did he fit into the debates about religious toleration? This is a quite difficult question to answer. Religious persecution creates a situation where it might well be wise to conceal one's actual views. It is clear that Newton did this and did it in collaboration with Locke (McLachlan, 1941). Given this, what can we say about Locke? Religion and Christianity in particular is perhaps the most important influence on the shape of Locke's philosophy. But what kind of Christian was Locke? Locke's family were Puritans. At Oxford, Locke avoided becoming an Anglican priest. Still, Locke's nineteenth century biographer Fox Bourne thought that Locke was an Anglican. Others have identified him with the Latitudinarians -- a movement among Anglicans to argue for a reasonable Christianity that dissenters ought to accept. This would seem to comport well with Locke's attempt in the *The Reasonableness of Christianity* to reduce the doctrines and practices required to be a Christian to a bare minimum. Still, there are some reasons to think that Locke was neither an Anglican or a Latitudinarian. Locke got Isaac Newton to write Newton's most powerful anti-Trinitarian tract. Locke arranged to have the work published anonymously in Holland (though in the end Newton decided not to publish). This strongly suggests that Locke too was a unitarian. Given that one main theme of Locke's Letter on Toleration is that there should be a separation between Church and State, this does not seem like the view of a man devoted to a state religion. It might appear that Locke's writing *The Reasonableness of Christianity* in which he argues that the basic doctrines of Christianity are few and compatible with reason make him a Latitudinarian. Yet Richard Ashcraft has argued that that comprehension for the Anglicans meant conforming to the existing practices of the Anglican Church; that is, the abandonment of religious dissent. Ashcraft also suggests that Latitudinarians were thus not a moderate middle ground between contending extremes but part of one of the extremes -- "the acceptable face of the persecution of religious dissent." (Ashcraft in Kroll, Ashcraft and Zagorin 1992 p. 155) Ashcraft holds that while the Latitudinarians may have represented the ‘rational theology’ of the Anglican church, there was a competing dissenting ‘rational theology’ Thus, while it is true that Locke had Latitudinarian friends, given Ashcraft's distinction between Anglican and dissenting "rational theologies", it is entirely possible that *The Reasonableness of Christianity* is a work of dissenting "rational theology."

Locke had been thinking, talking and writing about religious toleration since 1659. He and Shaftesbury had instituted religious toleration in the *Fundamental Constitution of the Carolinas.* He wrote the *Epistola de Tolerentia* in Latin in 1685 while in exile in Holland. He very likely was seeing Protestant refugees pouring over the borders from France where Louis XIV had just revoked the Edict of Nantes. Holland itself was a Calvinist theocracy with significant problems with religious toleration. But Locke's Letter does not confine itself to the issues of the time. Locke gives a principled account of religious toleration, though this is mixed in with arguments which apply only to Christians, and perhaps in some cases only to Protestants. He gives his general defense of religious toleration while continuing the anti-Papist rhetoric of the Country party which sought to exclude James II from the throne.

Locke's arguments for religious toleration connect nicely to his account of civil government. Locke defines life, liberty, health and property as our civil interests. These are the proper concern of a magistrate or civil government. The magistrate can use force and violence where this is necessary to preserve civil interests against attack. This is the central function of the state. One's religious concerns with salvation, however, are not within the domain of civil interests, and so lie outside of the legitimate concern of the magistrate or the civil government. In effect, Locke adds an additional right to the natural rights of life, liberty, health and property -- the right of freedom to choose one's own road to salvation.

Locke holds that the use of force by the state to get people to hold certain beliefs or engage in certain ceremonies or practices is illegitimate. The chief means which the magistrate has at her disposal is force, but force is not an effective means for changing or maintaining belief. Suppose then, that the magistrate uses force so as to make people profess that they believe. Locke writes:

> A sweet religion, indeed, that obliges men to dissemble, and tell lies to both God and man, for the salvation of their souls! If the magistrate thinks to save men thus, he seems to understand little of the way of salvation; and if he does it not in order to save them, why is he so solitious of the articles of faith as to enact them by a law. (Mendus, 1991. p. 41)

So, religious persecution by the state is inappropriate. Locke holds that "Whatever is lawful in the commonwealth cannot be prohibited by the magistrate in the church." This means that the use of bread and wine, or even the sacrificing of a calf could not be prohibited by the magistrate.

If there are competing churches, one might ask which one should have the power? The answer is clearly that power should go to the true church and not to the heretical church. But Locke claims, this amounts to saying nothing. For, every church believes itself to be the true church, and there is no judge but God who can determine which of these claims is correct. Thus, skepticism about the possibility of religious knowledge is central to Locke' argument for religious toleration.

Bibliography

Locke's Works

Oxford University Press is in the process of producing a new edition of all of Locke's works. This will supersede *The Works of John Locke* of which the 1823 edition is probably the most standard. The new Clarendon editions began with Peter Nidditch's edition of *An Essay Concerning Human Understanding* in 1975. The Oxford Clarendon editions contain much of the material of the Lovelace collection, purchased and donated to Oxford by Paul Mellon. This treasure trove of Locke's works and letters, which includes early drafts of the *Essay* and much other material, comes down from Peter King, Locke's nephew, who inherited Locke's papers. Access to these papers has given scholars in the twentieth century a much better view of Locke's philosophical development and provided a window into the details of his activities which is truly remarkable. Hence the new edition of Locke's works will very likely be definitive.

In addition to the Oxford Press edition, there are a few editions of some of Locke's works which are worth noting.

- Laslett, Peter *Locke's Two Treatises of Government*, Cambridge, Cambridge University Press.
- Richard Aschcraft, The Two Treatises of Civil Government
- Abrams, Phillip, *John Locke, Two Tracts of Government*, Cambridge University Press.
- Gough, (1968) J.W, and Klibansky, '*Epistola de Tolerentia*', *A Letter on Toleration*, Oxford, Oxford University Press.
- Aaron, R. and Gibb, J. eds. (1936) *An Early Draft of Locke's Essay*

Biographies

- King, Peter Lord (1991) *The Life of John Locke: with extracts from his correspondence, journals, and common-place books*, Bristol, England, Thoemmes
- Fox Bourne, H.R. (1876) *Life of John Locke* 2 volumes, reprinted Scientia Aalen, 1969.
- Maurice Cranston, (1957) *John Locke, A Biography*, reprinted Oxford University Press, 1985.

Bibliographies

- Hall, Roland, Woolhouse, Roger (1983) *80 years of Locke scholarship: a bibliographical guide*, Edinburgh, University Press.

Newsletter

- *The Locke Newsletter*, edited by Roland Hall, University of York, Heslington, York, UK <rhl@vaxa.york.ac.uk>.

Selected Books

- Aarsleff, Hans, (1982) *From Locke to Saussure: Essays on the Study of Language and Intellectual History*, Minneapolis, University of Minnesota Press
- Alexander, Peter (1985) *Ideas Qualities and Corpuscles*, Cambridge, Cambridge University Press.
- Arneil, Barbara, (1996) *John Locke and America*, Oxford, Clarendon Press
- Aaron, Richard, (1937) *John Locke*, Oxford, Oxford University Press
- Ashcraft, Richard, (1986) *Revolutionary Politics and Locke's Two Treatises of Civil Government*, Princeton, Princeton University Press.
- Ayers, Michael (1991) *Locke: Epistemology and Ontology*, 2 volumes, London Routledge.
- Bennett, Jonathan, (1971) *Locke, Berkeley, Hume: Central Themes*, Oxford, Oxford University Press.
- Brandt, Reinhard, ed. (1981) *John Locke: Symposium Wolfenbuttel 1979*, Berlin, de Gruyter.
- Chappell, Vere (1992) *Essays on Early Modern Philosophy, John Locke -- Theory of Knowledge*, London, Garland Publishing, Inc.
- Chappell, Vere (1994) *The Cambridge Companion to Locke*, Cambridge, Cambridge University Press.
- Dunn, John (1969) *The Political Thought of John Locke*, Cambridge University Press.
- Fox, Christopher, (1988) *Locke and the Scriblerians*, Berkeley, University of California Press.
- Gibson, James, (1968) *Locke's Theory of Knowledge and its Historical Relations*, Cambridge, Cambridge University Press
- Grant, Ruth, (1987) *John Locke's Liberalism*, Chicago, University of Chicago Press.
- Kroll, Peter; Ashcraft, Richard; Zagorin, Peter, (1992) *Philosophy, Science and Religion in England 1640-1700*, Cambridge, Cambridge University Press.
- Jolley, Nicholas, (1984) *Leibniz and Locke*, Oxford, Oxford University Press.
- Jolley, Nicholas, (1999) *Locke, His Philosophical Thought*, Oxford, Oxford University Press.
- Lott, Tommy, (1998) *Subjugation and Bondage: Critical Essays on Slavery and Social Philosophy*, New York, Rowman and Littlefield Publishers Inc..
- Lowe, E.J., (1995) *Locke on Human Understanding*, London, Routledge Publishing Co..
- Mackie, J. L. (1976) *Problems from Locke*, Oxford, Clarendon Press
- Macpherson, C.B. (1962) *The Political Theory of Possessive Individualism*, Oxford, Oxford University Press.
- Mandelbaum, Maurice, *Philosophy, Science and Sense Perception: Historical and Critical Studies*, Baltimore, The John Hopkins University Press.
- Martin, C. B. and D. M. Armstrong, eds. (1968) *Locke and Berkeley: A Collection of Critical Essays*, New York, Anchor Books.
- McLachlan, Hugh, (1941) *Religious Opinions of Milton, Locke and Newton,* Manchester, Manchester University Press.
- Mendus, Susan, (1991) *Locke on Toleration in Focus*, London, Routledge.
- Schouls, Peter, (1992) *Reasoned Freedom: John Locke and the Enlightenment*, Ithaca, NY, Cornell University Press
- Simmons, A. John, (1992) *The Lockean Theory of Rights*, Princeton, Princeton University Press.
- Tarcov, Nathan, (1984) *Locke's Education for Liberty*, Chicago, The University of Chicago Press.
- Tipton, I.C., (1977) *Locke on Human Understanding: Selected Essays*, Oxford, Oxford University

Press

- Tully, James, (1980) *A Discourse on Property*, Cambridge, Cambridge University Press
- Tully, James, (1993) *An Approach to Political Philosophy: Locke in Contexts*, Cambridge, Cambridge University Press.
- Wood, Neal, (1983) *The Politics of Locke's Philosophy*, Berkeley, University of California Press.
- Woolhouse, R.S., (1971) *Locke's Philosophy of Science and Knowledge* New York, Barnes and Noble.
- Woolhouse, R.S., (1983) *Locke*, Minneapolis, University of Minnesota Press.
- Woolhouse, R.S., (1988) *The Empiricists*, Oxford, Oxford University Press.
- Yaffe, Gideon, (2000) *Liberty Worth the Name: Locke on Free Agency*, Princeton, Princeton University Press.
- Yolton, Jean, (1990) *A Locke Miscellany*, Bristol, Thommes Antiquarian Books.
- Yolton, John, (1956) *John Locke and the Way of Ideas* Oxford, Oxford University Press, Thoemmes Press reprint 1996.
- Yolton, John (1969) *John Locke: Problems and Perspectives*, Cambridge, Cambridge University Press.
- Yolton, John (1970) *John Locke and the Compass of Human Understanding* Cambridge, Cambridge University Press
- Yolton, John (1984) *Perceptual Acquaintance: From Descartes to Reid* Minneapolis, University of Minnesota Press
- Yolton, John (1984) *Thinking Matter: Materialism in Eighteenth Century Britain*, Minneapolis, University of Minnesota Press

Other Internet Resources

- The Episteme Links Locke page
  (Keeps an up-to-date listing of links to Locke sites on the web.)
- John Locke
  (The Internet Encyclopedia of Philosophy entry on Locke)
- The Locke Page
  (The Great Voyages web site.)
- Images of Locke
  (National Portrait Gallery, Great Britain)

Related Entries

Berkeley, George | Hume, David | Leibniz, Gottfried Wilhelm | liberalism | Masham, Lady Damaris | personal identity | substance | tropes

William Uzgalis
WUzgalis@orst.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 2, 2001
Content last modified: September 26, 2001

Stanford Encyclopedia of Philosophy
Notes to John Locke

Notes

1. The scope of the activities engaged in by members of the Royal Society was much broader than what we recongize as modern science. The very idea of science was emerging during this period. Thus, only a minority of the early members were what we would call scientists. Similar societies were being founded in other European countries during this period. The Society still exists today and is one of the pillars of tthe orthodox scientific establishment in England. To be a member of it today implies that one has made a substantial contribution to experimental or theoretical science and membership thus confers the highest prestige. For more information about the Royal Society, one might consult Chapter 5 in Woolhouse (1988).

2. This controversy is of some interest for a variety of reasons. The constitution, for example, has quite liberal views concerning the formation of churches. Consider provision CIX for example. It reads: "No person whatsoever shall distu rb, molest, or persecute another for his speculative opinions in religion, or his way of worship." Was Locke responsible for this provision? On the other hand, Sir Leslie Stephan charged Locke with personal racism for inserting section CX: "Every freeman of Carolina shall have absolute power and authority over his negro slaves, of what opinion or religion soever." There is some evidence to suggest that Locke did play a part in formulating the sections on religion -- though it is possible this may have b e en at the bidding of Lord Ashley. Sir John Colleton is a much more likely candidate for the authorship of the sentence about negro slaves. Colleton, the real originator of the Carolinas project and one of the proprietors, was a Barbados planter who owned slaves. Part of the plan for the Carolinas was that people were going to emigrate from the overcrowed Barbados taking their slaves with them. They might well worry about whether this move might endanger the power they held over their slaves. It would be natural for Colleton to propose such a clause to allay their fears. Thus the inclusion of this clause may well say little or nothing about Locke's views.

3. Some commentators distinguish between the corpuscular and the atomic hypotheses on the grounds that corpuscles may be only relatively simple while atoms are supposed to be genuinely indivisible. (See Jolley, 1999) Thus it is possible to be committed to the existence of corpuscles but not atoms. While this distinction has its uses (especially in taking about the simplicity of simple ideas) I am convinced that in respect to physical corpuscles or atoms that Locke treats the two hypotheses as equivalent. One would expect that if the distinction were important for Locke, if he did believe in the existence of corpuscles but not atoms, he would not regularly use the terms ‘atom’ or ‘atoms’, but he does.

William Uzgalis
WUzgalis@orst.edu

First published: September 2, 2001
Content last modified: September 2, 2001

Stanford Encyclopedia of Philosophy
Supplement to John Locke

The Immateriality of the Soul and Personal Identity

Both in his discussion of personal identity and in his discussion of the immateriality of the soul in Book IV of the *Essay* Locke is agnostic about the immateriality of the soul. In Book IV he suggests that immateriality is not needed for the great ends of religion, and in Book II he crafts a theory of personal identity that does not require (though it is not inconsistent with) the immateriality of the soul.

The Immateriality of the Soul

In giving us his estimate of the limits of human understanding, Locke made some claims which surprised his contemporaries. In IV 3, 6 he suggests that given our ignorance of substances, it was possible that God could make matter fitly disposed think. He suggested that it was no farther beyond our comprehension motions of the body could give rise to pleasure and pain than that an immaterial soul could feel pain after the occurrence of some motions in the body. He suggested that the immateriality of the soul was not particularly important. In a passage from Book IV, Chapter 2, section 6, Locke writes:

> All the great ends of Morality and Religion, are well enough secured without the philosophical Proofs of the Soul's Immateriality; since it is evident that he who, at first made us beings to subsist here, sensible intelligent Beings, and for several years continued us in such a state, can and will restore us to a like state of Sensibility in another World, and make us there capable to receive the Retribution he has designed to men, according to the doings in this life. And therefore tis not a mighty necessity to determine one way or t'other, as some overzealous for or against the Immateriality of the Soul, have been foreward to make the World believe.

These suggestions were often taken as stronger than intended. Many of Locke's critics were suspicious that Locke had materialist tendencies. Instead of the skeptical conclusions about immaterial versus material substance which Locke is clearly arguing for, his remarks were sometimes treated as proposing that matter can and does think. It hardly matters however. Samuel Clarke, for example, a student of Newton's and an orthodox Anglican theologian, engaged in a debate by correspondence or rather public pamphlet with Anthony Collins over this issue between 1706 and 1708. Clarke sought to show that from our ideas alone it would be possible to show that matter thinking would involve a contradiction. If Clarke is right, Locke (even on the weaker interpretation I am proposing) would be wrong. There was an explosion of refutations of the claim that for all we know matter can think and the discussion of this issue lasted at least three quarters of the way through the eighteenth century.

Personal Identity

Locke added his Chapter "Of Identity and Diversity" (II. xxvii) which gives his account of identity and personal identity to the second edition of the *Essay*. Locke's account of personal identity turned out to be revolutionary. His account of personal identity is embedded in a general account of identity. In this general account of identity Locke distinguishes between the identity of atoms, masses of atoms and living things. Each individual atom is the same at a time, and stays the same over time. So, there is no problem about the identity of atoms. Masses of atoms are individuated by their constituent atoms without regard to the way in which they are organized. Living things, by contrast, are individuated by their functional organization. This organization is instantiated at any time by a collection of atoms. But the organization can persist through changes in the particles which make it up -- at least gradual change which continues the functions which the organization performs. Clearly the most important of these functions is the continuation of the same life. It is the continuation of the same functional organization and thus the same life which is the criterion of identity for sameness of living thing, be it an oak or a horse. Locke holds that man is an animal and is thus individuated just like other living things. So ‘man’ refers to a living body of a particular shape. Locke is perfectly aware that the definition of man is not really settled, and that there are a variety of competing definitions. He argues for his own definition, which involves distinguishing between ‘man’ and ‘person’ by using a variety of thought experiments and deducing unacceptable consequences from competing definitions. He points our, for example, that while those who individuate man solely in terms of the possession of a soul can explain the sameness of man from infancy to old age, if they accept some doctrine of reincarnation, their definition requires that the same soul in different bodies be the same man as much as infant and old man. If the doctrine of reincarnation allows the soul of a man to be reborn in the body of an animal, such as a hog, if we knew that the soul of a man was in one of our hogs, it would require us to call the hog a man. Locke pairs the examples of a rational talking parrot with a creature that has the shape of a man but cannot engage in rational discourse as a thought experiment which demonstrates that rational discourse is neither a necessary or sufficient condition for being a man. If man is a living body, an animal of a certain shape, then what is a person? A person is an intelligent thinking being that can know itself as itself the same thinking thing in different times and places.

Why does Locke make this distinction between ‘man’ and ‘person?’ One answer is that the distinction solves the problem of the resurrection of the dead. What is this problem? The problem begins with Biblical texts asserting that we will have the same body at the Resurrection as we did in this life. The issue is in what sense this is true. Clearly there are problems with the supposition that one will. Robert Boyle, in his essay, "Some Physico-Theological Considerations About the Possibility of the Resurrection" had raised some of these puzzles. Boyle writes:

> When a man is once really dead, divers of the parts of his body will, according to the course of nature, resolve themselves into multitudes of steams that wander to and fro in the air; and the remaining parts, that are either liquid or soft, undergo so great a corruption and change, that it is not possible so many scattered parts should be again brought together, and reunited after the same manner, wherein the existed in a human body whilst it was yet

> alive. And much more impossible it is to effect this reunion, if the body have been, as it often happens, devoured by wild beasts or fishes; since in this case, though the scattered parts of the cadaver might be recovered as particles of matter, yet already having passed into the substance of other animals, they are quite transmuted, as being informed by the new form of the beast or fish that devoured them and of which they now make a substantial part. (Robert Boyle, *Selected Philosophical Papers of Robert Boyle,* ed. M.A. Stewart, Manchester University Press, New York, 1979. p. 198.

These difficulties with putting bodies back together are obviously considerable, though not perhaps beyond the powers of Omnipotence. The culminating problem, however, is what happens to the man whose body is eaten by cannibals? Boyle continues:

> And yet far more impossible will this reintegration be, if we put the case that the dead man was devoured by cannibals; for then, the same flesh belonging successively to two different persons, it is impossible that both should have it restored to them at once, or that any footsteps should remain of the relation it had to the first possessor.

These problems I suspect represent the kinds of difficulties which faced the scientists of the Royal Society, and with which Boyle was particularly concerned, in integrating the kinds of explanations of natural phenomena in terms of particles and matter in motion, with the truths of religion.

Locke explicitly tells us that the case of the prince and the cobbler shows us the resolution of the problem of the resurrection. The case is one in which the soul of the prince with all of its princely thoughts is transferred from the body of the prince to the body of the cobbler, the cobbler's soul having departed. The result of this exchange, is that the prince still consider himself the prince, even though he finds himself in an altogether new body. Locke's distinction between man and person makes it possible for the same person to show up in a different body at the resurrection and yet still be the same person. Locke focuses on the prince with all his princely thoughts because, on his view, it is consciousness which is crucial to the reward and punishment which is to be meted out at the Last Judgment. In this chapter on identity, Locke is also making a distinction between consciousness and the soul, but that distinction is not crucial to the resolution of the kinds of problems that Boyle considered in his essay on the resurrection. Let us turn then, to the distinction between soul and consciousness.

Though the distinction between man and person is controversial, Locke's distinction between the soul or the thing which thinks in us and consciousness is even more radical. Locke holds that consciousness can be transferred from one soul to another, and that personal identity goes with consciousness. In section 12 of the Chapter of Identity and Diversity he raises the question: "...if the same Substance which thinks be changed, it can be the same person, or remaining the same, it can be a different person." Locke's answer to both of these questions is affirmative. Consciousness can be transferred from one substance to another and thus while the soul is changed, consciousness remains the same and thus personal identity is preserved through the change. And on the other hand, consciousness can be lost as in utter forgetfulness while the soul or thinking substance remains the same. Under these conditions there is the same soul but

a different person. These affirmations amount to the claim that the same soul or thinking substance is neither necessary not sufficient for personal identity over time. The arguments are developed by analogy with the functional organization of animals which is preserved through the gradual changes in the atoms which instantiate that organization at any given time. So, at any given time there must be a soul or thinking substance, but over time there is no necessity that one have the same soul to preserve personal identity.

Why does Locke make this distinction between soul and consciousness? This distinction has little bearing on problems about the same body at the resurrection. Still, the ressurection is important. I would suggest that the answer lies in Locke's interest in justice at the final judgment. Locke is skeptical about our ability to reidentify the same soul over time. He claims that if we were always awake, we could be certain that we had the same soul. But consciousness has natural gaps in it, such as periods during which we are asleep. Locke claims that there is no way of knowing that one soul has not been substituted for another during this period of absence of consciousness.

I would argue that the whole force of Locke's definition of person as a thinking intelligent being that can know itself as the same thinking thing in different times and places is designed to account for the fact that we are creatures who are capable of operating the machinery of the law. When contemplating an action we can think that in the future we will be the same being who will be punished or rewarded for the course of action which we choose. When being punished we can look back and see that we are the same being who committed the act for which we are being punished. Locke holds that consciousness is essential for justice to be done. If one is punished for doing something which one does not remember doing, it is equivalent to being created miserable. So, since consciousness plays the most important role in our being punished or rewarded at the last judgment for our actions, and consciousness can be transferred from one soul to another, and we have no mechanism to reidentify souls over time, it becomes clear why consciousness is Locke's choice for the bearer of personal identity, and why he makes the distinction between the substance which thinks in us and consciousness. I think this account explains a variety of oddities and difficulties in Locke's account. On his account, for example, memory must be completely accurate -- at least in the respects relevant for divine judicial purposes. Evidence which others might produce about one's identity has no role to play and so forth. Locke's account of freedom of action is also connected with his view of the forensic nature of personal identity. Freedom to review the decisions one has made about how to act are clearly of great importance in being able to operate the law. If one could not pause to consider, and change one•s mind about what one was going to do, it might well be said that one could not do otherwise.

William Uzgalis
wuzgalis@orst.edu

Return to John Locke

First published: September 2, 2001
Content last modified: September 2, 2001

David Hume

Generally regarded as the most important philosopher ever to write in English, David Hume (1711-1776) -- the last of the great triumvirate of "British empiricists" -- was also noted as an historian and essayist. A master stylist in any genre, Hume's major philosophical works -- *A Treatise of Human Nature* (1739-1740), the *Enquiries concerning Human Understanding* (1748) and *concerning the Principles of Morals* (1751), as well as the posthumously published *Dialogues concerning Natural Religion* (1779) -- remain widely and deeply influential, despite their being denounced by many of his contemporaries as works of scepticism and atheism. While Hume's influence is evident in the moral philosophy and economic writings of his close friend Adam Smith, he also awakened Immanuel Kant from his "dogmatic slumbers" and "caused the scales to fall" from Jeremy Bentham's eyes. Charles Darwin counted Hume as a central influence, as did "Darwin's bulldog," Thomas Henry Huxley. The diverse directions in which these writers took what they gleaned from reading Hume reflect not only the richness of their sources but also the wide range of Hume's empiricism. Comtemporary philosophers recognize Hume as one of the most thoroughgoing exponents of philosophical naturalism.

Life and Works

Born in Edinburgh, Hume spent his childhood at Ninewells, the family's modest estate on the Whitadder River in the border lowlands near Berwick. His father died just after David's second birthday, "leaving me, with an elder brother and a sister under the care of our Mother, a woman of singular Merit, who, though young and handsome, devoted herself to the rearing and educating of her Children." (All quotations in this section from Hume's autobiographical essay, "My Own life", reprinted in HL.)

Katherine Falconer Home realized that young David was "uncommonly wake-minded" -- precocious, in her lowland dialect -- so when his brother went up to Edinburgh University, David, not yet twelve, joined him. He studied mathematics and contemporary science, and read widely in history, literature, and ancient and modern philosophy.

Hume's family thought him suited for a career in the law, but he preferred reading classical authors, especially Cicero, whose *Offices* became his secular substitute for *The Whole Duty of Man* and his family's strict Calvinism. Pursuing the goal of becoming "a Scholar & Philosopher," he followed a rigorous program of reading and reflection for three years until "there seem'd to be open'd up to me a New Scene of Thought."

The intensity of developing this philosophical vision precipitated a psychological crisis in the isolated scholar. Believing that "a more active scene of life" might improve his condition, Hume made "a very feeble trial" in the world of commerce, as a clerk for a Bristol sugar importer. The crisis passed and he remained intent on articulating his "new scene of thought." He moved to France, where he could live frugally, and settled in La Flèche, a sleepy village in Anjou best known for its Jesuit college. Here, where Descartes and Mersenne studied a century before, Hume read French and other continental authors, especially Malebranche, Dubos, and Bayle; he occasionally baited the Jesuits with iconoclastic arguments; and, between 1734 and 1737, he drafted *A Treatise of Human Nature*.

Hume returned to England in 1737 to ready his *Treatise* for the press. To curry favor with Bishop Butler, he "castrated" his manuscript, deleting his controversial discussion of miracles, along with other "nobler parts." Book I (*Of the Understanding*) and Book II (*Of the Passions*) was published anonymously in 1739. Book III (*Of Morals*) appeared in 1740, as well as an anonymous *Abstract* of the first two books of the *Treatise*. Although other candidates, especially Adam Smith, have occasionally been proposed as the *Abstract*'s author, scholars now agree that it is Hume's work. The *Abstract* features a clear, succinct account of "one simple argument" concerning causation and the formation of belief. Hume's elegant summary presages his "recasting" of that argument in the first *Enquiry*.

The *Treatise* was no literary sensation but it didn't "fall dead-born from the press," as Hume disappointedly described its reception. Despite his surgical deletions, the *Treatise* attracted enough of a "murmour among the zealots" to fuel his life-long reputation as an atheist and a sceptic.

Back at Ninewells, Hume published two modestly successful volumes of *Essays, Moral and Political* in 1741 and 1742. When the Chair of Ethics and Pneumatical ("Mental") Philosophy at Edinburgh became vacant in 1745, Hume hoped to fill it, but his reputation provoked vocal and ultimately successful

opposition. Six years later, he stood for the Chair of Logic at Glasgow, only to be turned down again. Hume never held an academic post.

In the wake of the Edinburgh debacle, Hume made the unfortunate decision to accept a position as tutor to the Marquess of Annandale, only to find that the young Marquess was insane and his estate manager dishonest. Hume managed to extricate himself from this situation, and accepted the invitation of his cousin, Lieutenant-General James St. Clair, to be his Secretary ("I wore the uniform of an officer.") on a military expedition against the French in Quebec. Contrary winds delayed St. Clair's fleet until the Ministry canceled the plan, only to spawn a new expedition that ended as an abortive raid on the coastal town of L'Orient in Brittany.

Hume also accompanied St. Clair on an extended diplomatic mission to Vienna and Turin in 1748. While he was in Italy, the *Philosophical Essays concerning Human Understanding* appeared. A recasting of the central ideas of Book I of the *Treatise*, the *Philosophical Essays* were read and reprinted, eventually becoming part of Hume's *Essays and Treatises* under the title by which they are known today, *An Enquiry concerning Human Understanding*. In 1751, this *Enquiry* was joined by a second, *An Enquiry concerning the Principles of Morals*. Hume described the second *Enquiry*, a substantially rewritten version of Book III of the *Treatise*, as "incomparably the best" of all his works. More essays, the *Political Discourses*, appeared in 1752, and Hume's correspondence also reveals that a draft of the *Dialogues concerning Natural Religion* was underway at this time.

An offer to serve as Librarian to the Edinburgh Faculty of Advocates gave Hume the opportunity to work steadily on another project, a *History of England*, which was published in six volumes in 1754, 1756, 1759, and 1762. His *History* became a best-seller, finally giving him the financial independence he had long sought.

But even as a librarian, Hume managed to arouse the ire of the "zealots." In 1754, his order for several "indecent Books unworthy of a place in a learned Library" prompted a move for his dismissal, and in 1756, an unsuccessful attempt to excommunicate him. The Library's Trustees canceled his order for the offending volumes, which Hume regarded as a personal insult. Since he needed the Library's resources for his *History*, Hume did not resign his post; he did turn over his salary to Thomas Blacklock, a blind poet he befriended and sponsored. When research for the *History* was done in 1757, Hume quickly resigned to make the position available for Adam Ferguson.

Hume's publication of *Four Dissertations* (1757) was also surrounded by controversy. In 1755, he was ready to publish a volume that included "Of Suicide" and "Of the Immortality of the Soul." He suppressed the controversial essays when his publisher, Andrew Millar, was threatened with legal action, due largely to the machinations of the minor theologian William Warburton. Hume added "Of Tragedy" and "Of the Standard of Taste" to round out the volume, which also included *The Natural History of Religion* and *A Dissertation on the Passions*.

In 1763, Hume accepted an invitation from Lord Hertford, the Ambassador to France, to serve as his

Private Secretary. During his three years in Paris, Hume became Secretary to the Embassy and eventually its Chargè d'Affaires. He also become the rage of the Parisian salons, enjoying the conversation and company of Diderot, D'Alembert, and d'Holbach, as well as the attentions and affections of the *salonnières*, especially the Comtesse de Boufflers.

Hume returned to England in 1766, accompanied by Jean-Jacques Rousseau, who was then fleeing persecution in Switzerland. Their friendship ended quickly and miserably when the paranoid Rousseau became convinced that Hume was masterminding an international conspiracy against him.

After a year (1767-68) as an Under-Secretary of State, Hume returned to Edinburgh to stay. His autumnal years were spent quietly and comfortably, dining and conversing with friends, and revising his works for new editions of his *Essays and Treatises,* which contained his collected essays, the two *Enquiries*, *A Dissertation on the Passions*, and *The Natural History of Religion*. In 1775, he added an "Advertisement" to these volumes, in which he appeared to disavow the *Treatise*. Though he regarded this note as "a compleat Answer" to his critics, especially "Dr. Reid and that biggotted, silly fellow, Beattie," subsequent readers have wisely chosen to ignore Hume's admonition to ignore his greatest philosophical work.

Upon finding that he had intestinal cancer, Hume prepared for his death with the same peaceful cheer that characterized his life. He arranged for the posthumous publication of his most controversial work, the *Dialogues concerning Natural Religion*; it was seen through the press by his nephew and namesake in 1779, three years after his uncle's death.

The *Treatise* and the *Enquiries*

Hume's apparent disavowal of the *Treatise* raises a question as to how we should read his works. Should we take his "Advertisement" *literally* and let the *Enquiries* represent his considered view? Or should we take him *seriously* and conclude -- whatever *he* may have said or thought -- that the *Treatise* is the best statement of his position?

Both responses presuppose that there are substantial enough differences between the works to warrant our reading them as disjoint. This is highly dubious. Even in the "Advertisement," Hume says that "most of the principles, and reasonings, contained in this volume, were published" in the *Treatise*, and that he has "cast the whole anew in the following pieces, where some negligences in his former reasoning and more in the expression, are...corrected" (EHU, "Advertisement"). Despite his protests, this hardly sounds like the claims of one who has genuinely repudiated his earlier work.

Hume reinforced this perspective when he wrote Gilbert Elliot of Minto that "the philosophical principles are the same in both...by shortening and simplifying the questions, I really render them much more complete" (HL, I:158). And in "My Own Life," he opined that the *Treatise's* lack of success "proceeded more from the manner than the matter." Hume's "recasting" of the *Treatise* was probably designed primarily to address this point. This brief overview of Hume's central views on method, epistemology, and ethics therefore follows the structure -- "the manner" -- of the *Enquiries* and emphasizes "the matter"

they have in common with the *Treatise*.

Method

In his Introduction to the *Treatise,* Hume bemoans the sorry state of philosophy, evident even to "the rabble without doors," which has given rise to "that common prejudice against metaphysical reasonings of all kinds" (T, xiv). He hopes to correct this miserable situation by introducing "the experimental method of reasoning into moral subjects," establishing "a science of human nature" that will put philosophy on a "solid foundation" of "experience and observation" (T, Introduction).

Hume's positive, naturalistic project has much in common with contemporary cognitive science. Recent readers have paid more attention to these aspects of his philosophy than his earlier critics apparently did. As a result, no contemporary Hume scholar entirely accepts the traditional view that Hume was solely a negative philosopher whose goal was to make manifest the sceptical consequences of the views of his empiricist predecessors. But there remains considerable disagreement about the role and extent of scepticism in his philosophy, and disagreement about its relation to the naturalistic elements of his system. What Hume says about his aims and method helps clarify these issues.

In *An Enquiry concerning the Principles of Morals,* Hume says that he will "follow a very simple method," which will nonetheless bring about "a reformation in moral disquisitions" like that already accomplished in natural philosophy, where we have been cured of "a common source of illusion and mistake" -- our "passion for hypotheses and systems." To make parallel progress in the moral sciences, we should "reject every system...however subtle or ingenious, which is not founded on fact and observation," and "hearken to no arguments but those which are derived from experience" (EPM, 173-175).

The "hypotheses and systems" Hume rejects cover a wide range of philosophical and theological views. These theories were too entrenched, too influential, and too different from his proposed science of human nature to permit him just to present his "new scene of thought" as their replacement. He needed to show why we should reject these theories, so that he might have space to develop his own.

Hume outlines this strategy in the first section of *An Enquiry concerning Human Understanding*. He considers two prominent types of "false metaphysics" (EHU, 12). Though each type has as its basis an appealing human characteristic, both views extend their accounts of these characteristics beyond their basis in experience, and so beyond the bounds of cognitive content.

The first view looks at humans as active creatures, driven by desires and feelings. It paints a flattering picture of human nature, easy to understand and even easier to accept. These philosophers make us *feel* what they *say* about our feelings, and what they say is so useful and agreeable that ordinary people who encounter these views are readily inclined to accept them. This view might be called *sentimentalism*. It is a generic characterization of the position defended in Hume's time by Shaftesbury and Francis Hutcheson.

The other view downgrades sentiment to concentrate on *rationality*, which it treats as the distinctive human characteristic. This view glorifies the reasonable aspects of our natures and appeals to them in its emphasis on rarefied speculation and abstract argument. The systems of Descartes and other rationalist philosophers fit this general description. Given its emphasis on the role of the intellect, this view might be called *intellectualism.*

Intellectualism and sentimentalism seem to be exhaustive alternatives, ways of characterizing the ancient debate as to whether reason or passion is, or should be, the dominant force in human life. Hume saw that *both* approaches capture important aspects of human nature, but that *neither* tells the whole story. We are active *and* reasonable creatures. A view that mixes both styles of philosophy will be best, so long as it gets the mixture right.

But getting the mixture right, Hume realized, is no easy task. Intellectualism is too abstract, too remote from ordinary life to have any practical application. It can indulge the worst excesses of human vanity, especially when it treats matters that are beyond the limits of human understanding. It can be co-opted by popular superstitions, peddling religious fears and prejudices cloaked in profound-sounding but meaningless metaphysical jargon.

It is tempting to react to these features of intellectualism by arguing that we should abandon metaphysics altogether. But ordinary life doesn't equip us to do good metaphysics, and without some measure of accurate metaphysical description, sentimentalism can't be as precise as it should be. Delicate sentiment requires just reasoning, and an adequate account of just reasoning requires an accurate and precise metaphysics. The only way to correct sentiment and to avoid the sources of error and uncertainty rooted in intellectualism, is to do more metaphysics -- but of the right kind. We must pursue *true metaphysics* if we want to jettison these false and deceptive views.

Hume's insight was to see that getting the correct mixture requires a two-fold task, with negative and positive aspects. To develop a science of human nature, it is first necessary to undermine the foundations of all forms of false and misleading metaphysics. When we are rid of these sources of superstition, prejudice, and error, the stage will be clear for the kind of mental geography that constitutes true metaphysics. Accurate, just reasoning about human nature -- the descriptive project of true metaphysics -- requires us to examine the scope and limits of our cognitive capacities, so that we may at last obtain an exact picture of the powers and limitations of human understanding.

The negative phase of Hume's project scrutinizes the central arguments of the dominant philosophical and theological views of his day and exposes the lack of cognitive content in their key notions. Hume's sceptical arguments are an important part of this negative phase. Since these arguments are among the most prominent and powerful Hume has to offer, it is not surprising that they are often mistaken for his final view. But these arguments function as *reductios* of theories he rejects, not as parts of the positive position he offers in their place. They point up the poverty of false metaphysics to rid us of the temptation of doing metaphysics this way. Only then will we be ready for the positive phase -- true metaphysics,

which will replace the old incoherent metaphysics with the careful accurate description that is the proper goal of philosophy.

Empiricism

This combination of negative and positive aims is a distinguishing feature of Hume's particular brand of empiricism, and the strategy he devised to achieve these aims is revelatory of his philosophical genius. For Hume, all the materials of thinking -- *perceptions* -- are derived either from *sensation* ("outward sentiment") or from *reflection* ("inward sentiment") (EHU, 19). He divides perceptions into two categories, distinguished by their different degrees of force and vivacity. Our "more feeble" perceptions, *ideas*, are ultimately derived from our livelier *impressions*(EHU, Section II).

Although we permute and combine ideas in imagination to form complex ideas of things we haven't experienced, our creative powers extend no farther than "the materials afforded us by the senses and experience." *Complex ideas* are composed of *simple ideas*, which are fainter copies of the simple impressions from which they are ultimately derived, to which they correspond and exactly resemble. Hume offers this "general proposition" as his "first principle...in the science of human nature" (T, 7). Usually called the "Copy Principle," Hume's distinctive brand of empiricism is often identified with his commitment to it.

Hume presents the Copy Principle as an empirical thesis. He emphasizes this point by offering, in both the *Treatise* and the first *Enquiry*, as an empirical counterexample to the principle, "one contradictory phenomenon" (T, 5-6; EHU, 20-21) -- the infamous missing shade of blue. Hume asks us to consider "a person to have enjoyed his sight for thirty years, and to have become perfectly well acquainted with colours of all kinds, excepting one particular shade of blue..."(T, 6). Then

"Let all the different shades of that colour, except that single one, be plac'd before him, descending gradually from the deepest to the lightest; 'tis plain, that he will perceive a blank, where that shade is wanting, and will be sensible, that there is a greater distance in that place betwixt the contiguous colours, than in any other. Now I ask, whether 'tis possible for him, from his own imagination, to supply this deficiency, and raise up to himself the idea of that particular shade, tho' it had never been conveyed to him by his senses? I believe there are few but will be of the opinion that he can; and this may serve as a proof, that the simple ideas are not always derived from the correspondent impressions; tho' the instance is so particular and singular, that 'tis scarce worth our observing, and does not merit that for it alone we should alter our general maxim" (T 6). Hume's critics have objected that, in offering this counterexample, he either unwittingly destroys the generality of the Copy Principle, which he needs, given the uses to which he will put it, or else his dismissive attitude toward the counterexample reflects his disingenuous willingness to apply the Copy Principle arbitrarily, while pretending that it really possesses the generality his uses of it require.

Hume's defenders, on the other hand, maintain either that he should have granted that the imaginative construction of the missing shade really produces a *complex* idea, or that he should have insisted that

such counterexamples are exceedingly rare, and that the contentious metaphysical ideas, the cognitive content of which he uses the Copy Principle to critique, are not possibly ideas that could be generated by the imagination in the way the missing shade is supposedly generatred.

These defenses have their attractive points, but there is a far more satisfying resolution of the issue the missing shade raises available to Hume. In Book II of the *Treatise*, he describes a similar remarkably similar phenomenon that occurs with certain passions:

"Ideas may be compar'd to the extension and solidity of matter, and impressions, especially reflective ones, to colours, tastes, smells and other sensible qualities. Ideas never admit of a total union, but are endow'd with a kind of impenetrability, by which they exclude each other, and are capable of forming a compound by their conjunction, not by their mixture. On the other hand, impressions and passions are susceptible of an entire union; and like colours, may be blended so perfectly together, that each of them may lose itself, and contribute only to vary that uniform impression, which arises from the whole. Some of the most curious phaenomena of the human mind are deriv'd from this property of the passions" (T 366).

In these cases of "impressions and passions," both of which are simples for Hume, two impressions or two passions are *blended* to form a third, which is also a simple impression or passion. It seems plausible to think, and Hume's language in this passage certainly suggests as much, that one's ideas of two shades of (say) blue could also be blended to produce a third simple idea -- an idea of the missing shade.

While Hume's empiricism is usually identified with the Copy Principle, it is actually his use of its *reverse* in his account of *definition* that is really the most distinctive element of his empiricism.

Believing that "the chief obstacle...to our improvement in the moral or metaphysical sciences is the obscurity of the ideas, and ambiguity of the terms" (EHU, 61), Hume argued that conventional definitions -- defining terms in terms of other terms -- replicate philosophical confusions by substituting synonyms for the original and thus never break out of a narrow "definitional circle." Determining the cognitive content of an idea or term requires something else.

Hume supplied what was required with his account of definition, which offers a simple series of tests to determine cognitive content. First, find the idea to which a term is annexed. If none can be found, then the term has no content, however prominently it may figure in philosophy or theology. If the idea is complex, break it up into the simple ideas of which it is composed. Then trace the simple ideas back to their original impressions: "These impressions are all strong and sensible. They admit not of ambiguity. They are not only placed in a full light themselves, but may throw light on their correspondent ideas, which lie in obscurity" (EHU, 62).

If the process fails at any point, the idea in question lacks cognitive content. When carried out successfully, it yields a full account -- a "just definition" -- of the troublesome idea or term; a Humean definition gives us its exact cognitive content. So, whenever we are suspicious that a "philosophical term

is employed without any meaning or idea (as is too frequent), we need but enquire, *from what impression is that supposed idea derived*? And if it be impossible to assign any, this will serve to confirm our suspicion. By bringing ideas into so clear a light we may reasonably hope to remove all dispute, which may arise, concerning their nature and reality" (EHU, 22).

Hume's account of definition is not only the most distinctive feature of his empiricism, it is also a brilliant strategic device. He regards it as "a new microscope or species of optics, by which, in the moral sciences, the most minute, and most simple ideas may be so enlarged as to fall readily under our apprehension, and be equally known with the grossest and most sensible ideas, that can be the object of our enquiry" (EHU, 62).

Association

The Copy Principle accounts for the *origins* of our ideas. But our ideas are also regularly *connected*. As Hume put the point in his "Abstract" of the *Treatise*, "there is a secret tie or union among particular ideas, which causes the mind to conjoin them more frequently together, and makes the one, upon its appearance, introduce the other" (T, 662).

A science of human nature should account for these connections. Otherwise, we are stuck with an *eidetic atomism* -- a set of discrete, independent ideas, unified only in that they are the contents of a particular mind. Eidetic atomism thus fails to explain how ideas are "bound together," and its inadequacy in this regard encourages us, as Hume thought it encouraged Locke, to postulate theoretical notions -- power and substance being the most notorious -- to account for the connections we find among our ideas. Eidetic atomism is thus a prime source of the philosophical "hypotheses" Hume aims to eliminate.

The principles required for connecting our ideas aren't theoretical and rational; they are natural operations of the mind, *associations* we experience in "internal sensation." Hume's introduction of these "principles of association" is the other distinctive feature of his empiricism, so distinctive that in the *Abstract* he advertises it as his most original contribution: "If any thing can intitle the author to so glorious a name as that of an inventor, 'tis the use he makes of the principle of the association of ideas" (T, 661-662).

Hume locates "three principles of connexion" or association: resemblance, contiguity, and cause and effect. Of the three, causation is the only principle that takes us "beyond the evidence of our memory and senses." It establishes a link or connection between past and present experiences with events that we predict or explain, so that "all reasonings concerning matter of fact seem to be founded on the relation of cause and effect." But causation and the ideas closely related to it also raise serious metaphysical problems: "there are no ideas, which occur in metaphysics, more obscure and uncertain, than those of power, force, energy or necessary connexion" (EHU, 61-62).

Hume wants to "fix, if possible, the precise meaning of these terms, and thereby remove some part of that obscurity, which is so much complained of in this species of philosophy" (EHU, 62). This project provides a crucial experiment for Hume's metaphysical microscope, one designed to prove the worth of

his method, to provide a paradigm for investigating problematic philosophical and theological notions, and to supply valuable material for these inquiries.

Causation: The Negative Phase

Hume's strategy dictates that he first show that alternative accounts of our "causal reasonings" are inadequate. This negative project directs his metaphysical microscope toward the intellectualist view that causal connections are made on the basis of the operations of the understanding. Hume proceeds by examining all of the possible ways in which our "causal reasonings" might be based on reason.

Reasoning concerns either *relations of ideas* or *matters of fact*. Hume quickly establishes that, whatever assures us that a causal relation obtains, it is not reasoning concerning relations between ideas. Effects are distinct events from their causes: we can always conceive of one such event occurring and the other not. So causal reasoning can't be *a priori* reasoning.

Causes and effects are discovered, not by reason but through experience, when we find that particular objects are constantly conjoined with one another. We tend to overlook this because most ordinary causal judgments are so familiar; we've made them so many times that our judgment seems immediate. But when we consider the matter, we realize that "an (absolutely) unexperienced reasoner could be no reasoner at all" (EHU, 45n). Even in applied mathematics, where we use abstract reasoning and geometrical methods to apply principles we regard as laws to particular cases in order to derive further principles as consequences of these laws, the discovery of the original law itself was due to experience and observation, not to *a priori* reasoning.

Even after we have experience of causal connections, our conclusions from those experiences aren't based on any reasoning or on any other process of the understanding. They are based on our past experiences of similar cases, without which we could draw no conclusions at all.

But this leaves us without any link between the past and the future. How can we justify extending our conclusions from past observation and experience to the future? The connection between a proposition that summarizes past experience and one that predicts what will occur at some future time is surely not an intuitive connection; it needs to be established by reasoning or argument. The reasoning involved must either be *demonstrative,* concerning relations of ideas, or *probable,* concerning matters of fact and existence.

There is no room for demonstrative reasoning here. We can always conceive of a change in the course of nature. However unlikely it may seem, such a supposition is intelligible and can be distinctly conceived. It therefore implies no contradiction, so it can't be proven false by *a priori* demonstrative reasoning.

Probable reasoning can't establish the connection, either, since it is based on the relation of cause and effect. What we understand of that relation is based on experience and any inference from experience is

based on the supposition that nature is uniform -- that the future will be like the past.

The connection could be established by adding a premise stating that nature is uniform. But how could we justify such a claim? Appeal to experience will either be circular or question-begging. For any such appeal must be founded on some version of the uniformity principle itself -- the very principle we need to justify.

This argument exhausts the ways reason might establish a connection between cause and effect, and so completes the negative phase of Hume's project. The explanatory model of human nature which makes reason prominent and dominant in thought and action is indefensible. Scepticism about it is well-founded: the model must go.

Hume insists that he offers his "sceptical doubts about the operations of the understanding," not as "discouragement, but rather an incitement...to attempt something more full and satisfactory" (EHU, 26). Having cleared a space for his own account, Hume is now ready to do just that.

Causation: The Positive Phase

Hume's negative argument showed that our causal expectations aren't formed on the basis of reason. But we do form them, and "if the mind be not engaged by argument...it must be induced by some other principle of equal weight and authority" (EHU, 41).

This principle can't be some "intricate or profound" metaphysical argument Hume overlooked. For all of us -- ordinary people, infants, even animals -- "improve by experience," forming causal expectations and refining them in the light of experience. Hume's "sceptical solution" limits our inquiries to common life, where no sophisticated metaphysical arguments are available and none are required.

When we examine experience to see how expectations are actually produced, we discover that they arise after we have experienced "the constant conjunction of two objects;" only then do we "expect the one from the appearance of the other." But when "repetition of any particular act or operation produces a propensity to renew the same act or operation...we always say, that this propensity is the effect of *Custom*" (EHU, 43).

So the process that produces our causal expectations is itself causal. Custom or habit "determines the mind...to suppose the future conformable to the past." But if this background of experienced constant conjunctions was all that was involved, then our "reasonings" would be merely hypothetical. Expecting that fire will warm, however, isn't just *conceiving o*f its warming, it is *believing that* it will warm. Belief requires that there also be some fact present to the senses or memory, which gives "strength and solidity to the related idea." In these circumstances, belief is as unavoidable as is the feeling of a passion; it is "a species of natural instinct," "the necessary result of placing the mind" in this situation.

Belief is "a peculiar sentiment, or lively conception produced by habit" that results from the *manner* in which ideas are conceived, and "in their feeling to the mind." It is "nothing but a more vivid, lively, forcible, firm, steady conception of an object, than what the imagination alone is ever able to attain" (EHU, 49). Belief is thus "more an act of the sensitive, than of the cogitative part of our natures" (T, 183), so that "all probable reasoning is nothing but a species of sensation" (T, 103). This should not be surprising, given that belief is "so essential to the subsistence of all human creatures." "It is more conformable to the ordinary wisdom of nature to secure so necessary an act of the mind, by some instinct or mechanical tendency" than to trust it "to the fallacious deductions of our reason" (EHU, 55). Hume's "sceptical solution" thus gives a descriptive alternative, appropriately "independent of all the laboured deductions of the understanding," to philosophers' attempts to account for our causal "reasonings" by appeal to reason and argument. For the other notions in the definitional circle, "either we have no idea of force or energy, and these words are altogether insignificant, or they can mean nothing but that determination of the thought, acquir'd by habit, to pass from the cause to its usual effect" (T, 657).

Necessary Connection and the Definition of Cause

It remains only for Hume to "confirm and illustrate" his positive account by providing a precise definition of our idea of causation. In doing so, he accounts in his own terms for the *necessary connection* so many philosophers have taken to be an essential component of the idea of causation.

As we should expect from the preceding discussion, when we examine a single case of two events we regard as causally related, our impressions are only of their *conjunction*; the single case, taken by itself, yields no notion of their *connection. W*hen we go beyond the single case to examine the background of experienced constant conjunctions of similar pairs of events, we find little to add, for "there is nothing in a number of instances, different from every single instance, which is supposed to be exactly similar" (EHU, 75). How can the mere repetition of *conjunctions* produce a *connection*?

While there is indeed nothing added to our *external* senses by this exercise, something does happen: "after a repetition of similar instances, the mind is carried by habit, upon the appearance of one event, to expect its usual attendant, and to believe that it will exist." We *feel* this transition as an impression of *reflection*, or i*nternal sensation*, and it is *this feeling of determination* that is "the sentiment or impression from which we form the idea of power or necessary connexion. Nothing farther is in the case" (EHU, 75).

Although the impression of reflection -- the internal sensation -- is the source of our idea of the connection, that experience wouldn't have occurred if we hadn't had the requisite impressions of sensation -- the external impressions -- of the current situation, together with the background of memories of our past impressions of relevant similar instances.

All the impressions involved are relevant to a complete account of the origin of the idea, even though they seem, strictly speaking, to be "drawn from objects foreign to the cause."
Hume sums up all of the relevant impressions in not one but two definitions of *cause*.

The relation -- or the lack of it -- between these definitions has been a matter of considerable controversy. If we follow his account of definition, however, the first definition, which defines a cause as "*an object, followed by another, and where all objects similar to the first are followed by objects similar to the second*" (EHU, 76), accounts for all the external impressions involved in the case. His second definition, which defines a cause as "*an object followed by another, and whose appearance always conveys the thought to that other*" (EHU, 77) captures the internal sensation -- the feeling of determination -- involved. *Both* are definitions, by Hume's account, but the "just definition" of *cause* he claims to provide is expressed only by the conjunction of the two: only together do the definitions capture all the relevant impressions involved.

Hume's account of causation provides a paradigm of how philosophy, as he conceives it, should be done. He goes on to apply his method to other thorny traditional problems of philosophy and theology: liberty and necessity, miracles, design. In each case, the moral is that *a priori* reasoning and argument gets us nowhere: "it is only experience which teaches us the nature and bounds of cause and effect, and enables us to infer the existence of one object from that of another. Such is the foundation of moral reasoning, which forms the greater part of human knowledge, and is the source of all human action and behaviour" (EHU, 164). Since we all have limited experience, our conclusions should always be tentative, modest, reserved, cautious. This conservative, fallibilist position, which Hume calls *mitigated scepticism*, is the proper epistemic attitude for anyone "sensible of the strange infirmities of human understanding" (EHU, 161).

Moral Philosophy

The cautious attitude Hume recommends is noticeably absent in moral philosophy, where "systems and hypotheses" have also "perverted our natural understanding," the most prominent being the views of the moral rationalists -- Samuel Clarke, Locke, and William Wollaston, the theories of "the selfish schools" -- Hobbes and Mandeville -- and the pernicious theological ethics of "the schools," whose promotion of the dismal "monkish virtues" frame a catalogue of virtues diametrically opposed to Hume's. Although he offers arguments against the "systems" he opposes, Hume thinks the strongest case against them is to be made descriptively: all these theories offer accounts of human nature that experience and observation prove false.

Against the moral rationalists -- the intellectualists of moral philosophy -- who hold that moral judgments are based on reason, Hume maintains that it is difficult even to make their hypothesis intelligible (T, 455-470; EPM, Appendix I). Reason, Hume argues, judges either of matters of fact or of relations. Morality never consists in any single matter of fact that could be immediately perceived, intuited, or grasped by reason alone; morality for rationalists must therefore involve the perception of relations. But inanimate objects and animals can bear the same relations to one another that humans can, though we don't draw the same moral conclusions from determining that objects or animals are in a given relation as we do when humans are in that same relation. Distinguishing these cases requires more than reason alone can provide. Even if we could determine an appropriate subject-matter for the moral rationalist, it would still be the case that, after determining that a matter of fact or a relation obtains, the understanding has no more room

to operate, so the praise or blame that follows can't be the work of reason.

Reason, Hume maintains, can at most inform us of the tendencies of actions. It can recommend means for attaining a given end, but it can't recommend ultimate ends. Reason can provide no motive to action, for reason alone is insufficient to produce moral blame or approbation. We need sentiment to give a preference to the useful tendencies of actions.

Finally, the moral rationalists' account of justice fares no better. Justice can't be determined by examining a single case, since the advantage to society of a rule of justice depends on how it works in general under the circumstances in which it is introduced.

Thus the views of the moral rationalists on the role of reason in ethics, even if they can be made coherent, are false.

Hume then turns to the claims of "the selfish schools," that morality is either altogether illusory (Mandeville) or can be reduced to considerations of self-interest (Hobbes). He argues that an accurate description of the social virtues, benevolence and justice, will show that their views are false.

There has been much discussion over the differences between Hume's presentation of these arguments in the *Treatise* and the second *Enquiry*. "Sympathy" is the key term in the *Treatise*, while *benevolence* does the work in the *Enquiry*. But this need not reflect any substantial shift in doctrine. If we look closely, we see that benevolence plays much the same functional role in the *Enquiry* that sympathy plays in the *Treatise*. Hume sometimes describes benevolence as a manifestation of our "natural" or "social sympathy." In both texts, Hume's central point is that we experience this "feeling for humanity" in ourselves and observe it in others, so "the selfish hypothesis" is "contrary both to common feeling and to our most unprejudiced notions" (EPM, 298).

Borrowing from Butler and Hutcheson, Hume argues that, however prominent considerations of self-interest may be, we do find cases where, when self-interest is not at stake, we respond with benevolence, not indifference. We approve of benevolence in others, even when their benevolence is not, and never will be, directed toward us. We even observe benevolence in animals. Haggling over how much benevolence is found in human nature is pointless; that there is any benevolence at all refutes the selfish hypothesis.

Against Hobbes, Hume argues that our benevolent sentiments can't be reduced to self-interest. It is true that, when we desire the happiness of others, and try to make them happy, we may enjoy doing so. But benevolence is necessary for our self-enjoyment, and although we may act from the combined motives of benevolence and enjoyment, our benevolent sentiments aren't identical with our self-enjoyment.

We approve of benevolence in large part because it is useful. Benevolent acts tend to promote social welfare, and those who are benevolent are motivated to cultivate the other social virtue, justice. But while benevolence is an original principle in human nature, justice is not. Our need for rules of justice isn't

universal; it arises only under conditions of relative scarcity, where property must be regulated to preserve order in society.

The need for rules of justice is also a function of a society's size. In very small societies, where the members are more of an extended family, there may be no need for rules of *justice*, because there is no need for regulating *property* -- no need, indeed, for *our* notion of property at all. Only when society becomes extensive enough that it is impossible for everyone in it to be part of one's "narrow circle" does the need for rules of justice arise.

The rules of justice in a given society are "the product of artifice and contrivance." They are constructed by the society to solve the problem of how to regulate property; other rules might do just as well. The real need is for some set of "general inflexible rules...adopted as best to serve public utility" (EPM, 305).

Hobbesians try to reduce justice to self-interest, because everyone recognizes that it is in their interest that there be rules regulating property. But even here, the benefits for each individual result from the whole scheme or system being in place, not from the fact that each just act benefits each individual directly. As with benevolence, Hume argues that we approve of the system itself even where our self-interest isn't at stake. We can see this not only from cases in our own society, but also when we consider societies distant in space and time.

Hume's social virtues are related. Sentiments of benevolence draw us to society, allow us to perceive its advantages, provide a source of approval for just acts, and motivate us to do just acts ourselves. We approve of both virtues because we recognize their role in promoting the happiness and prosperity of society. Their functional roles are, nonetheless, distinct. Hume compares the benefits of benevolence to "a wall, built by many hands, which still rises by every stone that is heaped upon it, and receives increase proportional to the diligence and care of each workman," while the happiness justice produces is like the results of building "a vault, where each individual stone would, of itself, fall to the ground" (EPM, 305).

"Daily observation" confirms that we recognize and approve of the utility of acts of benevolence and justice. While much of the agreeableness of the utility we find in these acts may be due to the fact that they promote our self-interest, it is also true that, in approving of useful acts, we don't restrict ourselves to those that serve our particular interests. Similarly, our private interests often differ from the public interest, but, despite our sentiments in favor of our self-interest, we often also retain our sentiment in favor of the public interest. Where these interests concur, we observe a sensible increase of the sentiment, so it must be the case that the interests of society are not entirely indifferent to us.

With that final nail in Hobbes' coffin, Hume turns to develop his account of the sources of morality. Though we often approve or disapprove of the actions of those remote from us in space and time, it is nonetheless true that, in considering the acts of (say) an Athenian statesman, the good he produced "affects us with a less lively sympathy," even though we judge their "merit to be equally great" as the similar acts of our contemporaries. In such cases our judgment "corrects the inequalities of our internal emotions and perceptions; in like manner, as it preserves us from error, in the several variations of

images, presented to our external senses" (EPM, 227). Adjustment and correction is necessary in both cases if we are to think and talk consistently and coherently.

"The intercourse of sentiments" that conversation produces is the vehicle for these adjustments, for it takes us out of our own peculiar positions. We begin to employ general language which, since it is formed for general use, "must be moulded on some general views ... ." In so doing, we take up a "general" or "common point of view," detached from our self-interested perspectives, to form "some general unalterable standard, by which we may approve or disapprove of characters and manners." We begin to "speak another language" -- the language of morals, which "implies some sentiment common to all mankind, which recommends the same object to general approbation, and makes every man, or most men, agree in the same opinion or decision concerning it. It also implies some sentiment, so universal and comprehensive as to extend to all mankind, and render the actions and conduct, even of the persons the most remote, an object of applause or censure, according as they agree or disagree with that rule of right which is established. These two requisite circumstances belong alone to the sentiment of humanity here insisted on" (EPM, 272). It is the *extended* or *extensive* sentiment of humanity -- benevolence or sympathy -- that for Hume is ultimately "the foundation of morals."

But even if the *social* virtues move us from a perspective of self-interest to one more universal and extensive, it might appear that the *individual* virtues do not. But since these virtues also receive our approbation because of their usefulness, and since "these advantages are enjoyed by the person possessed of the character, it can never be self-love which renders the prospect of them agreeable to us, the spectators, and prompts our esteem and approbation" (EPM, 234).

Just as we make judgments about others, we are aware, from infancy, that others make judgments about us. We desire their approval and modify our behavior in response to their judgments. This *love of fame* gives rise to the habit of reflectively evaluating our own actions and character traits. We first see ourselves as others see us, but eventually we develop our own standards of evaluation, keeping "alive all the sentiments of right and wrong," which "begats, in noble natures, a certain reverence" for ourselves as well as others, "which is the surest guardian of every virtue" (EPM, 276). The general character of moral language, produced and promoted by our social sympathies, permits us to judge ourselves and others from the general point of view, the proper perspective of morality. For Hume, that is "...the most perfect morality with which we are acquainted" (EPM, 276).

Hume summarizes his account in this definition of *virtue*, or *Personal Merit*: "every quality of the mind, which is *useful* or *agreeable* to the *person himself* or to *others,* communicates a pleasure to the spectator, engages his esteem, and is admitted under the honourable denomination of virtue or merit" (EPM, 277). That is, as observers -- of ourselves as well as others -- to the extent that we regard certain acts as manifestations of certain character traits, we consider the usual tendencies of acts done from those traits, and find them useful or agreeable, to the agent or to others, and approve or disapprove of them accordingly. A striking feature of this definition is its precise parallel to the two definitions of *cause* that Hume gave as the conclusion of his central argument in the first *Enquiry*. Both definitions pick out features of events, and both record a spectator's reaction or response to those events.

Politics, Criticism, History, and Religion

Hume's "Advertisement" for the first two books of the *Treatise* promised subsequent works on morals, politics and criticism, but his *Political Discourses*, "Of Tragedy," and "Of the Standard of Taste" are our only hints as to what he might have said about those topics.

Hume's political essays range widely, covering not only the constitutional issues one might expect, but also venturing into what we now call economics, dealing with issues of commerce, luxury, and their implications for society. His treatments of these scattered topics exhibit a unity of purpose and method that makes the essays much more than the sum of their parts, and links them, not only with his more narrowly philosophical concerns, but also with his earlier moral and literary essays.

Adopting a causal, descriptive approach to the problems he discusses, Hume stresses that current events and concerns are best understood by tracing them historically to their origins. This approach contrasts sharply with contemporary discussions, which treated these events as the products of chance, or -- worse -- of providence. Hume substitutes a concern for the "moral causes" -- the human choices and actions -- of the events, conditions, or institutions he considers. This thoroughly secular approach is accentuated by his willingness to point out the bad effects of superstition and enthusiasm on society, government, and political and social life.

"Of the Standard of Taste" is a rich contribution to the then-emerging discipline of what we now call *aesthetics*. This complex essay contains a lucid statement of Hume's views on what constitutes "just criticism," but it is not *just* about criticism, as some readers are beginning to realize. Though Hume's account of aesthetic judgment precisely parallels his account of causal and moral judgment, the essay also contains a discussion of how a naturalistic theory might deal with questions of normativity, and so is important, not just as a significant contribution to Hume's overall view, but also for its immediate relevance for problems in contemporary empirical naturalism.

Hume's *History of England*, published in six volumes over as many years in the 1750s, recalls his characterization, in the first *Enquiry*, of history as "so many collections of experiments." Hume not surprisingly rejects the theoretical commitments of both Tory and Whig accounts of British history, and offers what he believes is an impartial account that looks at political institutions as historical developments responsive to Britons' experience of changing conditions, evaluating political decisions in the contexts in which they were made, instead of second-guessing them in the light of subsequent developments.

The Natural History of Religion is also a *history* in a sense, though it has been described as "philosophical" or "conjectural" history. It is an account of the origins and development of religious beliefs, with the thinly-disguised agenda of making clear not only the nonrational origins of religion, but also of exposing and describing the pathology of its current forms. Religion began in the postulation, by primitive peoples, of "invisible intelligences" to account for frightening, uncontrollable natural phenomena, such as disease and earthquakes. In its original forms, it was polytheistic, which Hume

regards as relatively harmless because of its tolerance of diversity. But polytheism eventually gives way to monotheism, when the followers of one deity hold sway over the others. Monotheism is dogmatic and intolerant; worse, it gives rise to theological systems which spread absurdity and intolerance, but which use reason to corrupt philosophical thought. But since religion is not universal in the way that our nonrational beliefs in causation or physical objects are, perhaps it can eventually be dislodged from human thinking altogether.

Hume's *Natural History* cemented his reputation as a religious sceptic and an atheist, even before its publication. Prompted by his own prudence, as well as the pleas of his friends, he resisted publishing the *Dialogues concerning Natural Religion*, which he had worked on since the early 1750s, though he continued revising the manuscript until his death. An expansion and dramatic revision of the argument previewed in Section XI of the first *Enquiry*, the *Dialogues* are so riddled with irony that controversy still rages as to what character, if any, speaks for Hume. But his devastating critique of the argument from design leaves no doubt that -- scholarly details about its enigmatic final section aside -- the conclusions philosophers and theologians have drawn from that argument go far beyond any evidence the argument itself provides.

A fitting conclusion to a philosophical life, the posthumously published *Dialogues* would alone insure the philosophical and literary immortality of their author. In this magnificent work, Hume demonstrates his mastery of the dialogue form, while producing *the* preeminent work in the philosophy of religion.

Bibliography

Hume's Works

The abbreviations and texts cited above are as follows:

[T] *A Treatise of Human Nature*, edited by L. A. Selby-Bigge, 2nd ed. revised by P.H. Nidditch, Oxford: Clarendon Press, 1975. [Page references above are to this edition.]

A Treatise of Human Nature, edited by David Fate Norton and Mary J. Norton, Oxford/New York: Oxford University Press, 2000

[EHU] *Enquiry concerning Human Understanding*, in *Enquiries concerning Human Understanding and concerning the Principles of Morals*, edited by L. A. Selby-Bigge, 3rd edition revised by P. H. Nidditch, Oxford: Clarendon Press, 1975. [Page references above are to this edition.]

An Enquiry concerning Human Understanding, edited by Tom L. Beauchamp, Oxford/New York: Oxford University Press, 1999

[EPM] *Enquiry concerning the Principles of Morals*, edited by L. A. Selby-Bigge, 3rd edition revised by P. H. Nidditch, Oxford: Clarendon Press, 1975. [Page references above are to this edition.]

Enquiry concerning the Principles of Morals, edited by Tom L. Beauchamp, Oxford/New York: Oxford University Press, 1998

[HL] *The Letters of David Hume*, edited by J.Y.T. Greig, 2 volumes, Oxford: Clarendon Press, 1932. [This edition also contains Hume's autobiographical essay, "My Own Life" (HL, I:1-7).]

Other works by Hume and editions of Hume's writings are:

- *Dialogues concerning Natural Religion*, edited by Norman Kemp Smith, Oxford: Oxford University Press, 1935
- *The Natural History of Religion*, edited by H. E. Root, Stanford: Stanford University Press, 1967
- *Essays, Moral, Political, Literary*, edited by Eugene F. Miller, Indianapolis: Liberty Classics, 1985
- *The History of England*, edited by William B. Todd, Indianapolis: Liberty Classics, 1983

In addition to the letters found in [HL], Hume's correspondence may be found in:

- *New Letters of David Hume*, edited by Raymond Klibansky and Ernest C. Mossner, Oxford: Clarendon Press, 1954

Finally, the closest thing at present to a complete edition remains that of Green and Grose:

- *The Philosophical Works of David Hume*, edited by T. H. Green and T. H. Grose. 4 volumes, London: Longman, Green, 1874-75

Bibliographical Studies

A useful bibliography of work on Hume is:

- Hall, Roland. *Fifty Years of Hume Scholarship: A Bibliographical Guide*, Edinburgh: Edinburgh University Press, 1978
- Hall also prepared annual bibliographies of the Hume literature for *Hume Studies*, a journal specializing in work on Hume, for the years 1977-1986; these bibliographies appeared in the November issues of that journal from 1978 to 1988
- *Hume Studies* revived the practice of including bibliographies with its November 1994 issue, which contained a comprehensive bibliography of the Hume literature from 1986-1993 by William Edward Morris. Subsequent volumes contain annual supplements to this bibiliography, also by Morris

Works on Hume

- Árdal, Páll S. *Passion and Value in Hume's* Treatise, Edinburgh: Edinburgh University Press, 1966; 2nd edition, revised, 1989
- Baier, Annette C. *A Progress of Sentiments: Reflections on Hume's* Treatise, Cambridge: Harvard University Press, 1991
- Beauchamp, Tom L. and Alexander Rosenberg. *Hume and the Problem of Causation*, New York: Oxford University Press, 1981
- Bennett, Jonathan. *Locke, Berkeley, Hume: Central Themes*, Oxford: Oxford University Press, 1973
- Bricke, John. *Hume's Philosophy of Mind*, Princeton: Princeton University Press, 1980
- Box, Mark A. *The Suasive Art of David Hume*, Princeton: Princeton University Press, 1990
- Capaldi, Nicholas. *Hume's Place in Moral Philosophy*, New York: Peter Lang, 1989
- Fogelin, Robert. *Hume's Scepticism in the* Treatise of Human Nature, London: Routledge and Kegan Paul, 1985
- Garrett, Don. *Cognition and Commitment in Hume's Philosophy*, Oxford/New York: Oxford University Press, 1996
- Jones, Peter. *Hume's Sentiments*, Edinburgh: Edinburgh University Press, 1982
- Livingston, Donald W. *Hume's Philosophy of Common Life*, Chicago: University of Chicago Press, 1984
- Livingston, Donald W. *Philosophical Melancholy and Delirium: Hume's Pathology of Philosophy*, Chicago: University of Chicago Press, 1998
- Mossner, Ernest Campbell. *The Life of David Hume*, London: Nelson, 1954
- Norton, David Fate. *David Hume, Common Sense Moralist, Sceptical Metaphysician*, Princeton: Princeton University Press, 1982
- Norton, David Fate (ed.) *The Cambridge Companion to Hume*, Cambridge: Cambridge University Press, 1993
- Noxon, James. *Hume's Philosophical Development*, Oxford: Oxford University Press, 1973
- Owen, David. *Hume's Reason*, Oxford: Oxford University Press, 2000.
- Passmore, John. *Hume's Intentions*, Cambridge: Cambridge University Press, 1952
- Pears, David. *Hume's System*, Oxford: Oxford University Press, 1990
- Penelhum, Terence. *Hume*, London: Macmillan, 1975
- Russell, Paul. F*reedom and Moral Sentiment*, New York: Oxford University Press, 1995
- Smith, Norman Kemp. *The Philosophy of David Hume*, London: Macmillan, 1941
- Stewart, John B. *Opinion and Reform in Hume's Political Philosophy*, Princeton: Princeton University Press, 1992
- Stewart, M. A. and John P. Wright. *Hume and Hume's Connexions*, Edinburgh: Edinburgh University Press, 1994
- Strawson, Galen. *The Secret Connexion: Causation, Realism and David Hume*, Oxford: Oxford University Press, 1989
- Stroud, Barry. *Hume*, London: Routledge and Kegan Paul, 1977
- Wright, John P. *The Sceptical Realism of David Hume*, Minneapolis: University of Minnesota

Press, 1983

Other Internet Resources

- The Leeds Hume Project, University of Leeds
- The Hume Society, based at the Philosophy Department, University of Iceland
- David Hume page, by Bill Uzgalis (Philosophy/Oregon State University), including links to texts of the *Enquiry*
- Annotated Selections from Hume's Writings, by Jonathan Bennett (Syracuse University)
- Bibliography on Hume, by Adam Potkay (English, William and Mary)
- Ty's Hume Homepage, maintained by D. Tycerium Lightner
- Entries on Hume in the *Internet Encyclopedia of Philosophy*, by James Fieser, U. Tennessee/Martin
    - Hume's Life and Writings
    - Hume's Metaphysical and Epistemological Theories
    - Hume's Moral Theories
    - Hume's Writing's on Religion
    - Hume's *Essays*: Moral, Political, and Literary

Related Entries

Berkeley, George | Hobbes, Thomas | Locke, John | miracles

William Edward Morris
Illinois Wesleyan University
wmorris@titan.iwu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 26, 2001
Content last modified: February 26, 2001

Miracles

Aquinas (*Summa Contra Gentiles*, III) says "those things are properly called miracles which are done by divine agency beyond the order commonly observed in nature *(praeter ordinem communiter observatum in rebus*)." A miracle, philosophically speaking, is never a mere coincidence no matter how extraordinary or significant. (If you miss a plane and the plane crashes, that is not a miracle unless God intervened in the natural course of events causing you to miss the flight.) A miracle is a supernaturally (divinely) caused event - an event (ordinarily) different from what would have occurred in the normal ("natural") course of events. It is a divine overriding of, or interference with, the natural order. As such, it need not be extraordinary, marvelous or significant, and it must be something other than a coincidence, no matter how remarkable - unless the "coincidence" itself is caused by divine intervention (i.e., not really a coincidence at all). Miracles, however, are ordinarily understood to be not just products of divine intervention in the natural order, but extraordinary, marvelous and significant as well. Thus, Aquinas says a miracle is "beyond the order commonly observed;" and Dr. Eric Mascall says that the word "miracle" "signifies in Christian theology a striking interposition of divine power by which the operations of the ordinary course of nature are overruled, suspended, or modified" (*Chamber's Encyclopaedia*).

The *locus classicus* for modern and contemporary philosophical discussion of miracles is Chapter X ("*Of Miracles*") of David Hume's *Enquiries Concerning Human Understanding*, first published in 1748. He says "A miracle may accurately be defined, a transgression of a law of nature by a particular volition of the deity, or by the interposition of some invisible agent" (*Enquiries*, p. 115n). His slightly different definition of a miracle as "a violation of the laws of nature" appears to be central to his argument against justified belief in miracles. "A miracle is a violation of the laws of nature; and as a firm and unalterable experience has established these laws, the proof against a miracle, from the very nature of the fact, is as entire as any argument from experience can possibly be imagined" (*Enquiries*, p. 114).

Miracles and Laws of Nature

Hume's argument against justified belief in miracles, as well as much subsequent discussion, appears to depend heavily upon the premise that "a miracle is a violation of the laws of nature." However, the actual role such a premise plays in Hume's argument, and whether Hume meant to define a miracle as a violation of a law of nature, or merely to characterise a miracle as, in some epistemologically relevant sense, "contrary" to the ordinary course of nature is controversial. It is clear, however, that on most commonsense, philosophical, *or* "scientific" accounts of what a law of nature is, technically speaking miracles are not violations of such laws but instead are positive instances of those laws. This is because laws of nature do not, and are not meant to, account for or describe events with supernatural causes - but only those with natural causes. Once some event is assumed to have a supernatural cause it is, by that very fact, outside the scope of laws of nature altogether and so cannot violate them. Only if one disregards the possibility of supernatural causes can known exceptions to laws possibly be regarded as violations of laws. However, in such a case there might be better reason to suppose that the exception simply shows that what was taken to be a law is not really a law, rather than that the exception is a violation of a genuine law of nature.

If the explanation I offer of Hume's argument against justified belief in miracles is correct (see the section below "Hume's Argument Against Justified Belief In Miracles"), then the premise that "a miracle is a violation of a law of nature" plays no significant role in his argument. The premise is a gloss for the underlying supposition that one cannot have an "impression" of a supernatural event. Because no such impression can be had, any allegedly miraculous event, simply because it is allegedly miraculous, cannot *ex hypothesis* be judged relevantly similar to any other event in experience. And *any* event that cannot be judged relevantly similar to others in our collective experience, cannot justifiably be believed to have occurred in accordance with Hume's principles of *a posteriori* reasoning. (Nor, can one justifiably believe that such an event will occur with any degree of probability whatsoever.) Before examining Hume's argument it is worth examining in detail the view that a miracle is a violation of a law of nature - an issue that far too much has been written about.

Hume is the pre-eminent proponent of the "regularity theory of causation." In characterizing regularity theories of causation, Tom Beauchamp (1974: 36) says,

> The modern claim is that universality and not objective necessity is that which is central to the concept of cause and also that which is implicit in any use *of* causal terminology. The philosophical problem of causation has thus largely come to be interpreted in this regularity tradition as the problem of the proper analysis of causal laws. Regularity exponents analyze laws as true, contingent, universal generalizations which are omnispatially and omnitemporally unrestricted in scope. Purported necessary connections between the antecedent and consequent events described in the law are regarded as gratuitous.

In fact, Beauchamp's view of the "modern claim" is problematic and the problem concerning a proper analysis of causal laws has not been resolved. But let us suppose that on Hume's view the conditions Beauchamp cites are sufficient for a statement to be a law of nature. That is: A statement L is a law of nature only if (i) L is contingent, (ii) L is general, and (iii) L is true. Given these constraints, consider the following two questions.

(1) Is there a law such that there might have been an event contrary to it?
(2) Can there be a law L that is violated while a law?

Given the above criteria for laws of nature, the answer to question 2 is "no." If an event occurred that fell within the scope of the laws of nature (i.e. was covered by those laws), but conflicted with the statement of the law, then either L would not be true, or else L would not be "general" - and therefore not a law of nature. The statement that the event occurred would be logically incompatible with the statement of the law of nature. Specifically it would be incompatible with the law whose status as a "law" is undermined because its truth or generality requirements are not met. If the event occurred, and we could know that it occurred and was "natural" (i.e. *within the scope of the law),* then we could no longer accept L as a genuine law. If laws of nature were descriptive of the scope and substance of everything that could logically happen, instead of their scope being limited to what can happen naturally (i.e., apart from supernatural interference), then of course miracles would not be possible. But that which is "physically impossible" - impossible within the constraints of the laws of nature - has a narrower scope than that which is logically possible. Apart from an argument to the contrary one need not assume that the logically and physically impossible are coextensive. To regard an event a natural is to regard it as falling within the scope of laws of nature, and anything that is covered by laws of nature cannot, *ex hypothesis*, violate them. Suppose an event assumed to be natural occurred, it really was natural and one could know that it did violate some alleged law of nature - no mistake was being made. Then this would show that the law it allegedly violated was in need of revision and was therefore not a genuine law at all.

However, suppose the laws of nature are regarded as non-universal or incomplete in the sense that while they cover natural events, they do not cover, and are not intended to cover, nonnatural events such as supernaturally caused events if there are or could be any. Then there is no contradiction in supposing that a physically impossible event could occur. A physically impossible occurrence would not violate a law of nature because it would not be covered by (i.e., within the scope of) such a law. So while the answer to question 2 is "no," this does not rule out the possibility of supernatural interference with the natural - perhaps as Robert Young (1972) suggests, as one causal condition among many necessary for an event's occurrence. What it does rule out is understanding this interference as a violation of the laws of nature in a technical sense. But this does not undermine the possibility of a miracle since the crucial element in the notion of a miracle - a "a supernatural interference with the natural order" - is not ruled out in showing that a miracle cannot really (strictly) be a violation of a law of nature.

If a miracle is not a violation of a law of nature, then how is it to be defined in relation to laws of nature? Question 1 above suggests a solution. A miracle can be defined as an event *contrary* to, but not a *violation* of, a law of nature. If "violation" is not being used in a technical sense, then a miracle can still be described as a violation of a law of nature - where "violation" would mean something like "contrary to

what could have happened had nature been the only force operative. " An event may be contrary to a law of nature without thereby invalidating it if it is caused by nonnatural forces - or in epistemic terms, if its occurrence can only be correctly explained in terms of nonnatural forces. A positive answer to question 1 follows from the fact that laws of nature do not describe, nor are they intended to describe, the logically possible. They only describe the physically possible. There is also a sense in which the positive answer to question 1 follows from the contingency of laws of nature. But his is not the sense that interests us. Even if the laws of nature were logically necessary, there could be events contrary to those laws if it is assumed that the scope of those laws is limited.

A violation of a law of nature by natural means is what one wants, normatively, to hold as a contradiction in terms - assuming insistence on generality (i.e., nonlocal empirical terms) in the statement of the law. One does not want to hold the occurrence of an event contrary to a law of nature due to nonnatural means as a contradiction in terms - at least not on the basis of an analysis of laws of nature. To hold this position, an analysis of laws would have to be combined with an argument against the possibility of nonnaturally caused events. (This is more or less what occurs in Hume's argument. Hume's empiricism and his theory of meaning are the basis of at least an implicit argument, employed by Hume, against the possibility of the supernatural in his discussion of miracles.) To say that miracles are impossible *because* violations of laws of nature are impossible is to improperly assume either (1) that a miracle must involve a violation of a law; or (2) that nothing contrary to a law of nature can occur because laws of nature circumscribe the logically possible and not merely the physically possible. But apart from distinct arguments to the contrary neither assumption appears to be warranted - at least not *prima facie* warranted.

To say, "an event is physically impossible and a violation of laws of nature if a statement of its occurrence is logically incompatible with a statement of the laws of nature," and *then* to assume that laws of nature circumscribe that which is logically, and not merely physically, impossible is to rule out the occurrence of the physically impossible on ill-conceived logical grounds. It is to deal with the possibility of miracles in the most superficial of ways by defining them out of existence using either an indefensible concept of a law of nature, or supposing a suppressed argument against the possibility of nonnatural interference.

A law of nature cannot be violated by natural forces. It can only be undermined as a genuine law. This happens if something natural occurs that the law was supposed to account but in fact could not. But neither can a law of nature be violated by a nonnatural force. Nor can it be undermined, assuming we can distinguish natural from nonnatural occurrences. A law of nature is, whatever else it may be, a true description of both the physically and logically possible occurrences within its scope, in the actual world *only* if it is assumed that no nonnatural forces could exist or interfere. Otherwise, a law describes only what can happen as a matter of physical possibility. Its presupposed scope is limited to what can happen given only natural forces. It allows for the possibility that the physically impossible remains logically possible, assuming the possibility of nonnatural forces capable of interaction in the actual world. Thus, nonnatural interventions are not, strictly speaking, violations of laws of nature. An intervention is, however, physically impossible, because (or so long as) that which is physically possible is defined in terms of the scope of laws of nature. An interference that is outside the scope of the laws of nature does

not violate any laws of nature by doing that which is physically impossible - that is - in doing that which is not possible given only natural forces.

Regularity theorists sometimes say that causal statements entail implicit or explicit reference to causal laws (e.g. laws of nature) and are instances of those laws. This appears to be false, or at least suspect, for a variety of different types of singular causal statements, including those in which a sufficient condition of X causing Y is dependent upon some subjective (nonphysical) factor. "The joke I was told caused me to ..." would not ordinarily be thought of as referring, either explicitly or implicitly, to a law of nature, or a general causal law, unless one were a strict determinist or maintained a strong form of a "covering law model" as essential to all forms of explanation. Even if there are some psychophysical laws, it is counterintuitive to argue that the *meaning* of causal statements, like the one above, implies a causal generalization. (Similarly, even if there are covering-law models that imply historical explanation is generalizable, it does not seem to be part of the *meaning* of historical explanation in causal form that it be generalizable.) Even supposing that the causal statement has counterfactual force (e.g. "If I had not been told the joke I would not have . . ."), one would not intuitively argue that this statement is an instance of a causal generalization by virtue of its meaning.

Leaving these controversial cases aside, a statement that a miracle occurred, usually - as in the case of many of the biblical miracles - refers to God as causing something that is not the sort of occurrence that one would expect to be explainable in terms of laws of nature, if it could be explained at all. I am here supposing supernatural explanation to be a viable alternative and the one that might plausibly be chosen in a case like the Red Sea parting as depicted in the movie "The Ten Commandments" (i.e., not simply a low tide). If causal statements did require reference to laws of nature, then this would appear to rule out the possibility of miracles since a miracle refers to a type of causal statement whose nature rules out reference to laws of nature taken as generalized cases of which they are instances. (John Locke (1706) denies that miracles are not instances of laws. They are not, however, instances of laws of nature according to Locke. He thinks that to say they are not instances of any laws whatsoever (e.g. not even of supernatural laws) is to say that they are random occurrences, and he thinks that this is absurd.) Miracles are contrary to laws of nature, *not* "violations" of them and *not* instances of them. (Actually, miracles are *vacuous* instances of true laws of nature as I explain below.) Note that it is not simply a miracle's *uniqueness* that rules out such reference to laws of nature. It cannot be uniqueness since even miracles that are supposed to be repeatable, such as raising one from the dead, cannot in principle refer to laws of nature for a complete explanation of their occurrence. Presumably they must also refer to divine intervention.

A miracle's uniqueness presents only a *preface* difficulty for supposing miracles to be supernaturally "caused." It is not difficult to show that causal terminology is applicable to statements about miracles. A regularity theory should be understood as requiring reference to laws of nature only when the causal statement is about natural events. More generally, a regularity theory requires reference to causal generalizations, but not necessarily to generalizations in terms of laws of nature. There is no reason to suppose that a miracle's uniqueness, if it is unique, cannot or does not carry with it implicit reference to a causal generalization. The counterfactual force that is constitutive of the meaning of some causal statements that specify necessary and sufficient material conditions for some event to occur may indicate

the presence of an implicit generalization in the causal statement about a miracle. Generally, if we say "X caused Y" we mean, in part, that if X had not occurred, then Y would not have occurred in the circumstances. But also implicit in the meaning of this is that if X occurred again, in relevantly similar circumstances, then Y would also occur again. To say that God caused X is to say that X would not have come about apart from God's activity and also that X would again come about if God acted similarly in a relevantly similar situation. If there is a supernature, then it is reasonable to suppose, as Locke did, that there are "laws of supernature" and that singular causal statements concerning the supernatural may be understood as implicitly assuming the generalizability of such singular causal statements in terms of those laws.

Consider the following objection to the characterization of a miracle as being contrary to a law of nature and outside its scope. Suppose, as I have, that true laws of nature do not have the form:

> (1) Whenever an event of type C occurs, an event of type E occurs.

Assume instead that they are of the form:

> (2) If an event of type C occurs, and there is no supernatural intervention, then an event of type E occurs.

Or, schematically:

> (3) (C & N) → E

Now consider a case where an event of type C occurs, there *is* supernatural intervention, and no event of type E occurs. From the truth table of the conditional function it follows that this case will be a positive instance of a true law, where such laws are of the form (C & N) → E. ("P → Q" will be true if the first component is false or if the second component is true. In the case under consideration, the first component will be false if "N" is false - this is, if there is supernatural intervention as hypothesized.) Thus, a miracle is not contrary to, or a violation of, a law of nature, and it is not outside the scope of such a law.

My response to this objection is as follows. I agree that the case considered above (i.e., a miraculous event occurs due to supernatural intervention) is a positive instance of a true law of nature where such laws are schematically of the form (C & N) → E. Miracles do not violate true laws of nature because such laws contain the supposition, either explicit or implicit, that laws describe what will happen given the presence of only natural forces. However, once there is supernatural interference, then no matter what follows C (whether or not E occurs), (C & N) → E will be trivially true just because N is false. (It would be true even if C is false.)

While it is true that a miraculous occurrence would be a positive instance of a true law, because true laws of the form (C & N) → E are never false if N is false (N is false if there is supernatural interference), I

want to call attention to the fact that while miracles do not violate true laws (i.e., they are positive instances of them), they should not be thought of as "within the scope of the laws of nature." This is because laws of nature are meant to account for, or describe, what occurs and what could possibly occur only apart from supernatural intervention. Laws describe what is naturally or physically possible. Because a positive instance of a true law of nature will be trivially true in cases of ~N, it will not explain why E does not occur even though C does occur. It is the assertion of ~N that does the explaining. Whether or not C occurred, or E occurs, (C & N) → E will be true when N is false. But if N is false (i.e., if there is supernatural interference), then the law of nature will not be able to explain either E or ~E in terms of natural forces. Yet this is what one normally expects a law of nature to do.

By saying that cases of supernatural interference are outside the scope of laws of nature, one is thereby refusing to consider cases of (C & N) → E, when N is false, to be *significant* instantiations of laws of nature, even though they are formally expressible in terms of laws of nature. While laws of nature can, and do, formally account for such cases, there is no explanation of E's nonoccurrence in terms of the natural forces that it is usually assumed to be the concern of laws of nature to describe.

To think of miracles as positive instances of laws of nature is to trivialize what is interesting about them viz. their relationship to laws of nature, where such laws are understood as describing what will and can happen given the presence of only natural forces. Speaking of cases in which there is supernatural intervention as outside the scope of laws of nature is clearly truer to our concept of such laws as descriptive only of those things that occur due to natural forces alone. That is their scope. Therefore, even though miracles can formally be accounted for by laws of nature, materially speaking this inadequate. It is inadequate because this "accounting for" is really done by the supposition of the supernatural interference and not with the miraculous event being a positive instance of the true law (i.e., because ~N results in (C & N) → E being true) as it would in cases where there was no supernatural intervention. Formally, even a positive instance of a true law can be "contrary" to a law of nature of which it is a positive instance. This will be the case in all instances of which an occurrence being a positive instance of a true law is due to supernatural intervention - that is, in all cases which make (C & N) → E trivially true in supposing ~N. This is formally unobjectionable but awkward. In keeping with ordinary usage it is therefore preferable to consider such positive instances of true laws as outside the scope of laws of nature, and to consider only positive instances of laws to be within the scope of laws if (C & N) → E is not true because of the falsity of N.

Hume's Argument Against Justified Belief In Miracles

Remarkably, the discussion of Hume on miracles has not been confined to, or even principally concerned with, whether or not Hume was correct in his argument against justified belief in miracles - and/or the possibility of justified belief in miracles. Instead, philosophical discussion has focused on exegetical issues concerning exactly what Hume was arguing. There is, for example, still no generally accepted view on the fundamental points of whether his argument (Part I of his essay) against the justified belief in

miracles on the basis of testimony is (i) meant as an *a priori* or *a posteriori* argument; (ii) if that argument can be, or is meant to be, generalized to include first-hand experience of an allegedly miraculous event; or indeed, (iii) if his argument, whether regarded as *a priori* or *a posteriori*, is meant to establish that one can never be justified in believing in a miracle on the basis of testimony. Hume does not appear to claim that miracles are impossible - only that justified belief in a miracle on the basis of testimony (may be) impossible. His argument is basically epistemological. There are, however, grounds for supposing that a miracle is not even possible on Hume's account - at least not given his wider empiricist views.

Hume's position on miracles cannot be properly understood apart from his analysis of causation, *a posteriori* reasoning, and indeed the most fundamental element of his empiricism - his analysis of "impressions" and "ideas" (Book I, *A Treatise of Human Nature*, pp. 1-7). In fact, Hume's position on miracles has never been properly understood because its connection to his views on causation has never been adequately examined. There is considerable controversy over what Hume's position actually was - let alone what his argument for that position is. I offer one highly abbreviated interpretation. (For a more complete account of this interpretation see Levine: 1989: 1-52.) The bibliography contains citations to other interpretations completely at odds with this one and with each other.

To understand Hume on miracles the following question must be answered. Why did Hume think that one could justifiably believe that an extraordinary event had occurred, under certain circumstances, but that one could never justifiably believe a miracle had occurred? The proposed interpretation of Hume's analysis of miracles in relation to his analysis of causation and his wider empiricism yields the only plausible answer to this question that I know of. This interpretation also shows why it makes no substantial difference whether we interpret Hume's argument in Part I "Of Miracles" against the possibility of justified belief in testimony to the miraculous as an *a priori* argument or an *a posteriori* argument since the arguments essentially coalesce.

Hume (*Enquiries*, p. 128) gives the following example of an extraordinary event that he thinks could be rendered credible on the basis of testimony.

> ...suppose, all authors, in all languages, agree, that from the first day of January 1600, there was a total darkness over the whole earth for eight days, suppose that the tradition of this extraordinary event is still strong and lively among the people: that all travelers, who return from foreign countries, bring us accounts of the same tradition, without the least variation or contradiction: it is evident, that our present philosophers, instead of doubting the fact, ought to receive it as certain, and ought to search for the causes whence it might be derived. The decay, corruption, and dissolution of nature, is an event rendered probable by so many analogies, that any phenomenon, which seems to have a tendency towards that catastrophe comes within the reach of human testimony, if that testimony be very extensive and uniform.

In this case not only is the testimony to the alleged event very extensive and uniform, but Hume also

thinks it necessary that our past experience does not render the event completely unlikely. He argues that the eight day darkness can be "rendered probable by so many analogies," assuming it is testified to extensively and uniformly. In such a case Hume assumes that the event is natural and that "we ought to search for the causes." Hume compares this with another imaginary case (*Enquiries*, p. 128).

> ...suppose, that all historians who treat of England, should agree, that, on the first of January 1600, Queen Elizabeth died...and that, after being interred a month, she again appeared, resumed the throne, and governed England for three years: I must confess that I should be surprised at the concurrence of so many odd circumstances, but should not have the least inclination to believe so miraculous event.

Since both events are assumed to be equally well testified to, the reason that Hume thinks the former can be judged credible but not the latter is that in the former case the "event is rendered probable by so many analogies." One can object and say that this appears to be nothing more than a subjective judgement on the part of Hume. His experience suggests analogies for the former type of event but not the latter. The eight day darkness "sufficiently resembles" events that Hume has experienced, or believes in on the basis of experience, to warrant belief in the eight day darkness given that the event is extraordinarily well attested to. In the latter case Hume can find no analogies to draw upon from experience. Given the similarity, in relevant respects, of most peoples' experience (i.e., the experience of Scots at the time of Hume), Hume thinks that if people base their judgments on their experience (in accordance with the principles of *a posteriori* reasoning (Levine 1989: 5-12) extrapolated from his analysis of causation) they will agree that the former (extraordinary) event can be judged credible but not the (miraculous) latter. Hume would agree that if an individual's experience were very different from his own in relevant respects, than that individual could justifiably believe many things that he himself could not.

So despite Hume's *a priori* arguments against justified belief in miracles he argues that under certain circumstances the "evidence" may justify belief in the occurrence of an extraordinary event as long as we have experienced events analogous in type. However, an extraordinary event is not necessarily a miraculous one. In the case of extraordinary events that are well attested to and for which we have suitable experiential analogies, Hume thinks that the most we are justified in believing is that the event did occur - not that the event is a miracle. We are to "search for the [natural] causes whence it might be derived." Such cases may even require us to reassess, to some extent, our estimation of what nature is capable of doing on her own, so to speak. Sometimes statements of laws of nature must be reassessed and altered in light of new experience. Also, we must be careful not to extend our judgments as to what to believe or expect of nature to situations in which all of the relevant circumstances are not the same. This requires explanation.

Hume relates the case of the Indian who refused to believe that water turned to ice. According to Hume, the Indian "reasoned justly" on the basis of his past experience. He refused, at first, to believe that water turned to ice, despite the fact that it was well attested to, because the event not only had the Indian's constant and uniform experience to count against it, but also because the event "bore so little analogy" to that experience (*Enquiries*, pp. 11-15). The Indian "reasoned justly" but he extended his judgments about the properties of water to cases where all the circumstances were not the same (i.e., the relevant

circumstance here being temperature). In certain situations in which we hear testimony to extraordinary events we may be in situations similar to that of the Indian. Indeed, according to Hume, if we justifiably believe that an extraordinary event did occur, then we *should* assume that we are in a situation just like that of the Indian. We should assume this because, as I shall show, there are logically compelling reasons why the consistent Humean, in accordance with the principles of *a posteriori* reasoning based on Hume's analysis of causation and his empiricism, can do nothing else. The extraordinary event should be judged "[not] contrary to uniform experience of the course of nature in cases where all the circumstances are the same" *(Enquiries*, p. 114n).

Why should we judge our situation to be like that of the Indian's? Are there logically compelling reasons for doing so? Hume does not explicitly say why, but it must be because our experience has shown us that situations like the Indian's do arise. On the basis of experience, when we are justified in believing in the occurrence of an extraordinary event, we should liken ourselves to the Indian. That is why, in a case like the eight days of darkness, "we ought to search for the **[natural]** causes whence it might be derived." Experience demands it. It seems then, that according to Hume, when an extraordinary event is extraordinarily well attested to we have only two options. One is to accept the testimony and look for the event's natural causes. The other is to reject the testimony on the grounds that the event testified to bears no *significant* analogy to events we have experienced. Hume thinks that testimony, no matter how reliable, *can never* establish the occurrence of a miraculous event, in accordance with the principles of *a posteriori* reasoning - reasoning that is a type of causal reasoning according to Hume. He says (*Enquiries*, pp. 111-112),

> It being a general maxim, that no objects have any discoverable connection together, and that all the inferences, which we can draw from one to another, are founded merely on our experience of their constant and regular conjunction; it is evident that we ought not to make an exception to this maxim in favor of human testimony.... This species of reasoning, perhaps, one may deny to be founded on the relation of cause and effect. I shall not dispute about a word.

Thus, Hume thinks that if we justifiably accept testimony to an extraordinary event, then on the basis of past experience, we must liken ourselves to the Indian and search for natural causes of which we are unaware. This would be for us the equivalent of the Indian moving north to "Muscovy during the winter" (*Enquiries*, p. 114n). (Think about the last astonishing thing you learned that nature could accomplish as a matter of course and you have a basic part of Hume's argument.)

Contrary to Hume one might try to argue as follows:

> Is it inconceivable that we experience events for which no explanation like that suitable for the Indian has been forthcoming? It may be true that in some situations a seemingly naturally inexplicable event was later learned to have natural causes, but it is at least conceivable that there may be other inexplicable events for which no natural causes can be found. If experience can show that we are unable to find natural causes for certain events -

though these events are every bit as well attested to as other events only some of which we have discovered natural causes for - then why must we liken ourselves to the Indian in cases where we justifiably believe in the occurrence of an extraordinary event? Why does experience demand that we either reject belief in the event's occurrence or believe it but posit natural causes for the event? Justified belief does not entail belief in a natural cause. Experientially you have not shown that it does. Moreover, if we had independent reasons for thinking that no cause of some extraordinary event could be found (e g., on the basis of prophecy), then it is conceivable that we could be justified in believing that an extraordinary event occurred without thereby likening ourselves to the Indian. The grounds on which we might reject the supposition of a natural cause could themselves be experiential (e.g., a prophet's track record). It does seem to be the case that we can always posit a natural explanation for an extraordinary event and base that supposition on experience. On the other hand, we may reject such a supposition, not only on the basis of a priori arguments of natural theology, but also on the basis of experience. For example, suppose that an extraordinary event that had some religious significance was prophesied, testimony justified belief in the event's occurrence, the prophet had been right about certain predictions made in the past, and no immediate natural explanation for the event that had the least bit of plausibility was forthcoming. The option of positing a natural explanation remains open, but experience does not necessarily demand that we avail ourselves of that option. Hume thinks that the most that testimony can establish is that an extraordinary event has occurred, not that a miracle has occurred. To support this one must establish the suppressed premise that we can have no good reasons on the basis of experience, for identifying an event as miraculous. Though Hume employs this premise he does not support it, and the example just given suggests a reason for believing the premise to be invalid.

Hume has not specified adequate criteria for determining when an event can be judged suitably analogous to past experience as to warrant belief when adequately testified to. (This is probably because he thought no such criteria could be given - each extraordinary case having to be considered on its own merits.) Experientially, there are no clear cut criteria enabling us to determine, with any degree of assurance, that an eight day darkness is analogous to past experience while a resurrection does not, in the least, bear any resemblance to aspects of our past experience that could make it at least as likely an event to be believed in as the eight day darkness. Could not a resurrection be found analogous to past experience in precisely the same way that an eight day darkness could (i.e., experience of the "decay, corruption, and dissolution of nature")? In the absence of such criteria there is no logically compelling reason, and not even necessarily compelling experiential reasons, for assuming that the extraordinary event occurred (naturally) but a resurrection did not occur (miraculously). If the darkness can be justifiably believed in then so too can a resurrection. Furthermore, under the appropriate circumstances, not only could the resurrection be judged miraculous and not merely extraordinary, but so could the eight day darkness. The thing that would determine whether or not the event was to be judged miraculous would be whether we had reason to believe that God or God's agents caused

> the event - better reasons than for thinking that the event was caused naturally. It is conceivable that a judgment that God caused a particular event can be experientially warranted. Again, imagine a prophet who is known to predict future events accurately. The prophet has a track record of empirically verifiable prophecies concerning events of a most extraordinary nature. Or, imagine a case in which every time a "holy-person" pointed at someone that person lay down dead. An explanation of such goings-on can be sought in terms of natural (e.g., parapsychical) causes and abilities. However, would experience necessitate the acceptance of this explanation over the supernatural one? Hume has not shown that it would.

Hume would reject this argument - and therein lies the entire tale of his argument against justified belief in miracles - whether on the basis of testimony or first-hand experience. He would insist that his principles of reasoning about empirical matters, and his philosophical empiricism (i.e., his theory of "impressions" and "ideas") show that supernatural explanation cannot be justified experientially. In the case of a prophet accurately predicting events, or the "holy-person" pointing their finger and people falling dead, Hume would say that experience justifies us in believing that the event prophesied will come to pass and that if the holy-person lifts their finger in our direction we are justified in running away - and foolish if we do not. But we are not justified in believing such events to be miracles.

We need to ask "What is it about experience, in the sense of expectations about future events or judgments about past events, that could justify the positing of a supernatural cause?" For positing such a cause is necessary if one is to justifiably believe some event to be a miracle. Hume would say that positing such a cause is speculative. It can have no basis in experience. Even if some event really were a miracle, whether it be a resurrection, or "the raising of a feather, when the wind wants ever so little of a force requisite for that purpose" (*Enquiries*, p. 115n), we would not be justified in believing that it was anything more than an extraordinary event. Extraordinary events are at the limits of (our) experience, the supernatural is beyond it. Hume (*Enquiries*, p. 129) says:

> Though the Being to whom the miracle is ascribed, be Almighty, it [the miracle] does not, upon that account, become a whit more probable, since it is impossible for us to know the attributes or actions of such a Being, otherwise than from the experience of his productions, in the usual course of nature. This still reduces us to past observations, and obliges us to compare the instances of the violation of truth in the testimony of men, with those of the violations of the laws of nature by miracles, in order to judge which of them is most likely and probable.

For Hume, a "cause," insofar as it can be used as an item in reasoning from experience, can only be something that we can have an "impression" of. The cause of a miracle would have to be identified as something we could perceive, even if we were to posit some metaphysical "power" of this cause and attribute it speculatively to God. The "cause" of Lazarus's coming forth from the grave would have to be identified with Christ's beckoning - either his voice or some physical gesture - both of which we have "impressions" of and both of which are events "in the usual course of nature."

If a resurrection were well enough attested to that it warranted belief, then that event could still only be assigned status as an extraordinary event with a natural explanation. Hume is thus constrained by his empiricism. He is constrained in such a way that had he been at the shore of the Red Sea with Moses when they were being chased (as in the movie version); and had Moses raised his staff and the Red Sea split up the middle (i.e., no low tide but raging waters on both sides); and had the Red Sea crashed to a close the moment the last Israelite was safe - killing those in pursuit; and had Hume himself lacked grounds for assuming he was hallucinating or perceiving events in any way other than as they were actually happening - Hume would still be constrained by his principles to deny that what he was witnessing was a miracle. This example suffices to show the unacceptability of Hume's argument. Indeed, assuming Hume would have agreed that had he been there with Moses, and had events transpired in a manner suitably similar to the way they are depicted in the film, he would have (readily) agreed that he was justified in believing that a miracle occurred; then his argument against justified belief in miracles can be used as a *reductio ad absurdum*. Flew (1967: 349) is mistaken in his claim that "it be neither arbitrary nor irrational to insist on a definition of a ‘law of nature’ such that the idea of a miracle as an exception to a law of nature is ruled out as self-contradictory."

A resurrection could only be well enough attested to to be justifiably believed if it could be judged as somehow analogous with something in our past experience. If it is, then it must be considered a natural event because, for Hume, anything analogous to our experience is at least analogous in the sense of suggesting that it too has a natural cause. We experience only that which occurs in nature and judgments based on that experience will not warrant positing causes outside of that experience. Suppose that some event actually was supernaturally caused. (Let us suppose Hume recognizes this as a logical possibility in his essay, though I do not think it is given his analysis of causation and his empiricism.) Hume would say that we could not, on the basis of experience, attribute a supernatural cause to the event because we experience only natural causes (i.e., events occurring in the usual course of nature). If an event were supernaturally caused we could legitimately say that we "experienced" some supernatural event, but the sense of experience used here would be an equivocation on Hume's usage This "cause," being transcendent, and not discernible by means of "sense impressions," "internal impressions," or "impressions of reflexion" could not be an item of experience at all as Hume sees it. Thus, because Hume thinks that every cause must be regarded as natural, he is committed to the view that one *could* justifiably believe that an extraordinary event had occurred, but *never* a miracle.

Hume's *a priori* argument against justified belief in miracles actually coalesces with his *a posteriori* argument against such justified belief. On *a posteriori* grounds we could never justifiably believe testimony to the miraculous because we could never judge the occurrence of such an event to be similar, in relevant respects, to anything we have experienced. However, that a miraculous occurrence could never be judged relevantly similar to anything in experience (i.e., that there must be "a firm and unalterable experience" counting against belief in it) is something that we can know *a priori,* since *a priori* we can know that we cannot have an " impression" of a supernatural cause. It follows from this that on *a priori* grounds we can also rule out the possibility of justified belief in testimony to the miraculous.

It follows from what has been said that unless one accepts Hume's analysis of *a posteriori* reasoning as a

type of causal reasoning, and also accepts his analysis of causation, which ultimately rests on his theory of impressions and ideas - a theory that even staunch empiricists should reject as simplistic - then there is no reason to accept *his* argument against the possibility of justified belief in miracles.

Of course, nothing in this critique of Hume's argument should be taken to suggest, in any way, that miracles have ever occurred, or that we are justified in believing that any have occurred. But it would be most surprising if some people at some time and in certain circumstances have not been, and will not again be, justified in believing in the occurrence of a miracle. However, nothing I have said suggests that that the evidence available for the occurrence of any alleged miracle warrants justified belief in miracles for most people - including those who really do believe in them.

Bayesian Analyses of Hume's Argument Concerning Miracles

There are various versions of Bayes's theorem. For example, John Earman (1993:307n4) employs the following:

$$\Pr(H/E\&K) = \frac{\Pr(H/K) \times \Pr(E/H\&K)}{\Pr(E/K)}$$

"The reader is invited to think of *H* as a hypothesis at issue; *K* as the background knowledge; and *E* as the additional evidence. Pr(*H/E&K*) is called the posterior probability of *H*. Pr(*H/K*) and Pr(*E/H&K*) are respectively called the prior probability of *H* and the (posterior) likelihood of *E*."

Bayesian analyses are prominent among the several recent and allegedly novel interpretations of Hume's argument against the justified belief in miracles. However, since there is no consensus on just what Hume's argument is, or exactly what he is trying to establish, it is impossible that any Bayesian analysis, let alone a "Bayesian proof" of that argument, or a recasting of the argument in terms of some version of Bayes's theorem, will not beg crucial issues of interpretation. In so doing, such analyses, in and of themselves, will also beg fundamental epistemological issues concerning, for example, evidence. Furthermore, it is difficult to see how recasting Hume's argument in a Bayesian form can clarify the structure or substance of the argument without presupposing what the argument is.

On the interpretation of Hume's argument given above, a Bayesian analysis sheds no light whatsoever on the structure or substance of the argument, and can do nothing by way of either supporting or refuting the argument. Indeed, any Bayesian analysis of the question of justified belief in miracles must be otiose until the difficult and essential questions concerning "evidence" in relation to an allegedly miraculous occurrence are resolved - at which point any Bayesian analysis will add little except the technical complexity of a formal apparatus that may or may not "clarify" the structure of Hume's argument.

The balancing of probabilities is of no use until it is decided what goes into the balance - that is, what constitutes the evidence that is to be subject to the balancing of probabilities. The point is this; apart from independent philosophical arguments - arguments that would in effect undermine the relevance of a Bayesian analysis to the question of the credibility of reports of the miraculous - no such analysis can, in principle, prove that no testimony can (or cannot) establish the credibility of a miracle. So-called Bayesian analyses of Hume's argument are not analyses of Hume's argument at all - but superfluous representations of it.

Are Miracles Religiously Significant?

Some contemporary theologians have claimed that the issue of miracles has been largely misunderstood and co-opted by philosophers and critics of religion for their own purposes - specifically in order to deny that certain central events the Bible alleges to have occur did occur. Thus, David Strauss (1835) claims that reports of miracles can be rejected "as simply impossible and irreconcilable with the known and universal laws which govern the course of events." Theologians sometimes claim, for example, that there is no word for "miracle," in the Old or New Testament. What are described there as "prodigies" or "wonders" or "effects of powers," are interpreted by philosophers, but not by the Biblical writers, as "miracles." Anthony Flew (1967:347) says that according to Spinoza, as well as some contemporary theologians, "conventional interpreters of the Bible read far more miracles into it than it contains, because they constantly read poetic Hebrew idioms literally." This may be true but it is also inconsequential. No matter how events such as a parting of the sea or a resurrection are described, whether as "wonders" or as "miracles," it is clear that they were understood in Biblical times, as well as in contemporary times, to be in some remarkable sense "contrary to the normal course of nature." People living at the time of Moses or Christ knew just as well as we do that seas do not "normally" part and that people are not, in the normal course of nature, resurrected. What is required for a notion of the miraculous is not some sophisticated notion of what a law of nature is, but just a strong sense of what constitutes the normal, natural course of events. And this is something the ancients had just as much as those living in a scientific age have.

Many people regard the philosophical dimension, the epistemological one in particular, of the issue of miracles as insignificant. Indeed, philosophy of religion, or at least natural theology and the analytic philosophy of religion is regarded as inconsequential by believers as well as some involved in the academic study of religion.

Nevertheless, I think that the philosopher's concerns are more closely allied with those of religious people than, for example, are those of the social scientist. This does not make them more important - just more closely allied. The philosopher is interested in the truth about religious truth-claims and (though many would disagree on philosophical grounds) can pursue that interest more or less independent of dogma, tradition, and what socio-scientific study tells us about the various functions of religion. Philosophical issues in religion are fundamental in a way that other areas of investigation are not. A system of beliefs may serve a variety of functions, personally and socially, but for the religious person these are consequences of the system of beliefs itself. They believe their systems of belief, their religion,

to be more or less coherent and true. It makes a great difference to most believers who are traditional theists whether or not miracles could occur, and whether one could justifiably believe they did occur. One does not have to be a fundamentalist Christian, for example, to believe with Paul that the Christian faith is, in an important sense, a vain pursuit if the central miracles associated with Christianity did not occur. Indeed, one can assume that David Hume, Bertrand Russell and Richard Swinburne would concur. The same point can of course be made, *mutatis mutandis*, for Islam and Judaism or any tradition fundamentally connected to miraculous claims.

Apart from belief in miracles, one is left with a system of beliefs that has had and will continue to have enormous significance - good and bad - for people's lives. However, for the majority of persons for whom these beliefs have that significance, religion could no longer function in the way it does if they became convinced of the falsity of their beliefs. That seems to be verifiable. There is a sense, albeit perhaps not a very important one, in which one is involved in pretense if one practices a system of beliefs whose central tenets one denies. What one is practicing may be similar in significant respects to the religious tradition in question, but one will not be practicing that religion, nor will one properly be regarded as a believer.

Having said that, it should also be said that the issue of miracles is regarded as overly important by contemporary analytic philosophers of religion. Some 19th and 20th century philosophical theologians, as well as those in various disciplines within the academic study of religion, including the philosophy of religion, no longer regard the issue of miracles as central or crucial, either to various religious traditions, to the religious life per se, and definitely not to the more fundamental questions of God and meaning in the traditional domain of philosophical theology. Philosophers of religion, even if sophisticated in terms of their analyses, naively attribute an importance to the issue that may not be altogether warranted.

Bibliography

- Adams, Robert, M. "Miracles, Laws of Nature and Causation - II." *The Aristotelian Society*, Supplementary Volume, 60 (1992), pp. 207-224.
- Armstrong, Benjamin, F. "Hume on Miracles: Begging Questions Against." *History of Philosophy Quarterly*, 9 (1992), pp. 319-328.
- Armstrong Jr, Benjamin, F. "Hume's Actual Argument Against Belief in Miracles." *History of Philosophy Quarterly*, 12 (1995), pp. 65-76.
- Augustine. *De Civitate Dei*, xxi 6-8 and xxii 8-10.
- Ahern, Dennis. "Hume on the Evidential Impossibility of Miracles." In *Studies In Epistemology*. Ed. Nicholas Rescher. Oxford: Blackwell, 1975, pp. 1-32.
- Ahern, Dennis. "Miracles and Physical Impossibility." *Canadian Journal of Philosophy*, VII (1977), pp. 71-79.
- Aquinas, Thomas. *Summa Contra Gentiles*, III, chapters 98-103.
- Backhaus, Wilfried, K. "Advantageous Falsehood: The Person Moved by Faith Strikes Back." *Philosophy and Theology*, 7 (1993), pp. 289-310.
- Basinger, David. "Miracles as Violations: Some Clarifications." *Southern Journal of Philosophy,*

22 (1984), pp. 1-8.

- Bahlul, Raja. "Miracles and Ghazali's First Theory of Causation." *Philosophy and Theology*, 15 (1990), pp. 137-150.
- Basinger, David. "Flew, Miracles and History." *Sophia*, 22 (1983) pp. 5-22.
- Basinger, David. "Miracles as Evidence for Theism." *Sophia*, 29 (1990), pp. 56-59.
- Basinger, David. *Philosophy and Miracle: The Contemporary Debate*. New York: E. Mellen Press, 1986.
- Beauchamp, Tom L., ed. *Philosophical Problems of Causation*. Encino: Dickinson Publ.ishers, 1974.
- Beauchamp, Tom L. and Rosenberg, Alexander. *Hume and The Problem of Causation*. Oxford: Oxford University Press, 1981.
- Beckwith, Francis. *David Hume's Argument Against Miracles: A Critical Analysis*. Lanham: University Press of America, 1989.
- Beckwith, Francis, J. "Hume's Evidential/Testimonial Epistemology, Probability, and Miracles." *Logos*, 12 (1991), pp. 87-104.
- Boden, Margaret. "Miracles and Scientific Explanation." *Ratio*, 11 (1969), pp. 137-44.
- Brand, Myles, ed. *The Nature of Causation*. Urbana: University of Illinois Press, 1976. This contains an extended annotated bibliography on causation.
- Brand, Myles, ed. "Causality." In *Current Research in Philosophy of Science*. Ed. P. Asquith and H. Kyburg. East Lansing: Philosophy of Science Association, 1979, pp. 252-281.
- Broad, C.D. "Hume's Theory of the Credibility of Miracles." *Proceedings of the Aristotelian Society*, NS XVIII (1916-1917), pp.77-94.
- Brown, Gregory. "Miracles in the Best of All Possible Worlds: Leibniz's Dilemma and Leibniz's Razor." *History of Philosophy Quarterly*, 12 (1995), pp. 19-39.
- Burheen, Herbert. "Attributing Miracles to Agents - A Reply to George Chryssides." *Religious Studies*, 4 (1977), pp. 485-488.
- Burns, R.M. *The Great Debate on Miracles*. Lewisburg: Bucknell University Press, 1981. This contains an extended and wide ranging bibliography, beginning with the seventeenth century, of works relevant to the problem of miracles and Hume's essay. It is especially useful for the problem in its historical setting.
- Butler, Joseph. *The Analogy of Religion*. London, 1736. New York: Frederick Ungar, 1961.
- Campbell, George. *A Dissertation on Miracles*. Edinburgh, 1762. Reprinted, New York: Garland Publishing, 1983.
- Cherry, Christopher. "On Characterizing the Extraordinary." *Ratio*, 17 (1975), pp. 52-64.
- Chryssides, George. "Miracles and Agents." *Religious Studies*, 11 (1975), pp. 319-237.
- Collier, John. "Against Miracles." *Dialogue*, 25 (1986), pp. 349-352.
- Collwell, Gary. "Miracles and History." *Sophia*, 22 (1983), pp. 9-14.
- Daston, Lorraine. "Marvellous Facts and Miraculous Evidence in Early Modern Europe." *Critical Inquiry*, 18 (1991), pp. 93-124.
- Dawid, Philip and Gillies, Donald. "A Bayesian Analysis of Hume's Argument Concerning Miracles." *Philosophical Quarterly*, 39 (1989), pp. 57-65.
- Diamond, Malcolm. "Miracles." *Religious Studies*, 9 (1973), pp. 307-324.
- Dietl, Paul. "On Miracles." *American Philosophical Quarterly*, 5 (1968), pp. 130-134.

- Ducasse, C.J. *Causation and The Types of Necessity*. University of Washington Publications in the Social Sciences. Vol.1, 1924, pp. 70-200. Reprinted by Dover Publications.
- Earman, John. "Bayes, Hume, and Miracles." *Faith and Philosophy*, 10 (1993), pp. 293-310.
- Eaton, Jeffrey. "The Problem of Miracles and The Paradox of Double Agency." *Modern Theology*, 1 (1985), pp. 211-222.
- Ellin, Joseph, S. "Again: Hume on Miracles." *Hume Studies*, 19 (1993), pp. 203-212.
- Erlandson, Douglas. "A New Look at Miracles." *Religious Studies*, 13 (1977), pp. 417-428.
- Evans, C., Stephen. "Critical Historical Judgment and Biblical Faith." *Faith and Philosophy*, 11 (1994), pp. 184-206.
- Ferguson, Kenneth. "An Intervention into the Flew/Fogelin Debate." *Hume Studies*, 18 (1992), pp.105-112.
- Fern, Richard, L. "Hume's' Critique of Miracles: An Irrelevant Triumph," *Religious Studies*, 18 (1982), pp. 337-354.
- Fitzgerald, Paul. "Miracles." *Philosophical Forum*, 17 (1985), pp. 48-64.
- Flew, Anthony. "Miracles." *Encyclopedia of Philosophy*. New York: Macmillan and Free Press, 1967, vol. 5, pp. 346-353.
- Flew, Antony. "Fogelin on Hume on Miracles." *Hume Studies*, 15 (1990), pp. 141-144.
- Flew, Anthony. *Hume's Philosophy of Belief*. London: RKP, 1961.
- Flew, Anthony. *God and Philosophy*. New York: Harcourt, Brace and World, 1966.
- Fogelin, Robert, J. "What Hume Actually Said about Miracles." *Hume Studies*, 16 (1990), pp. 81-86.
- Force, James, E. "The Breakdown of the Newtonian Synthesis of Science and Religion: Hume, Newton and the Royal Society." In *Essays on the Context, Nature and Influence of Isaac Newton's Theology*. Ed. J. Force and R. Popkin. Dordrecht: Kluwer Academic Publishers, 1990.
- Foster, Stephen, P. "Edward Gibbon and the Anti-Miracle Man: Hume's 'Of Miracles' at Work in the 'Decline and Fall of the Roman Empire'." *The Modern Schoolman*, 71 (1994), pp. 223-245.
- Gaskin, J.C.A. *Hume's Philosophy of Religion*. London: Macmillan, 1978.
- Gaskin, J.C.A. "Contrary Miracles Concluded." *Hume Studies*, (1985), pp.1-14.
- Gillies, Donald. "A Bayesian Proof of a Humean Principle." *British Journal for the Philosophy of Science*, 42 (1991), pp. 255-256.
- Gilman, James, E. "Reconceiving Miracles." R*eligion Studies*, 25 (1989), pp. 477-487.
- Goggans, Phillip. "Do the Closest Counterfactual Worlds Contain Miracles?" *Pacific Philosophical Quarterly*, 73 (1992), pp. 137-149.
- Gower, Barry. "David Hume and the Probability of Miracles." *Hume Studies*, 16 (1990), pp.17-31.
- Hájek, Alan. "In Defense of Hume's Balancing of Probabilities in the Miracle Argument." *Southwest Philosophy Review*, 11 (1995), pp. 111-118.
- Halpin, John, F. "The Miraculous Conception of Counterfactuals." *Philosophical Studies*, 63 (1991), pp.271-290.
- Hambourger, Robert. "Belief in Miracles and Hume's Essay." *Nous*, 14 (1980), pp. 587-604.
- Hambourger, Robert. "Need Miracles Be Extraordinary?" *Philosophy and Phenomenological Research*, XLVII (1987), pp. 435-449.
- Harre, Rom and Madden, Edward. *Causal Powers: A Theory of Natural Necessity*. Oxford: Basil Blackwell, 1975.

- Hoffman, Joshua. "Comments on 'Miracles and The Laws of Nature.'" *Faith and Philosophy*, 2 (1985), pp. 347-352.
- Holland, R.F. "The Miraculous." *American Philosophical Quarterly*, 2 (1965), pp. 43-51.
- Houston, J. *Reported Miracles: A Critique of Hume*. New York: Cambridge University Press, 1994.
- Hughes, Christopher. "Miracles, Laws of Nature and Causation - I.." *The Aristotelian Society: Supplementary Volume*, 66 (1992), pp. 179-205.
- Hume, David. *David Hume: Writings on Religion*. Ed. A. Flew. Peru: Open Court, 1992.
- Hume, David. *Enquiries Concerning Human Understanding*. Ed. L.A. Selby-Bigge. 3rd ed. Oxford: Oxford University Press, 1975.
- Hume, David. *A Treatise of Human Nature*. Ed. L.A. Selby-Bigge. 2nd ed. Oxford: *Clarendon Press*, 1975.
- Hume, David. *Dialogues Concerning Natural Religion*. Ed. N. Kemp-Smith. New York: Bobbs-Merrill, 1947.
- Kellenberger, J. "Miracles." *International Journal For Philosophy of Religion*, X (1979), pp. 145-162.
- Landrum, George. "What A Miracle Is." *Religious Studies*, 12 (1976), pp. 49-57.
- Langford, M.J. "The Problem of The Meaning of 'Miracle.'" *Religious Studies*, 7 (1971), pp. 43-52.
- Langtry, Bruce. "Hume, Probability, Lotteries and Miracles." *Hume Studies*, 16 (1990), pp. 67-74.
- Larmer, Robert, A. H. "Miracles and Natural Explanations: A Rejoinder." *Sophia*, 28 (1989), pp. 7-12.
- Larmer, Robert, A. H. "Miracles and Conservation Laws: A Reply to Professor MacGill." *Sophia*, 31 (1992), pp. 89-95.
- Larmer, Robert, A.H. "Miracles, Evidence and Theism: A Further Apologia." *Sophia*, 33 (1994), pp. 51-57.
- Langtry, Bruce. "Miracles and Principles of Relative Likelihood." *International Journal for Philosophy of Religion*, 18 (1985) pp. 123-131.
- Larmer, Robert A.H. *Water Into Wine? An Investigation of the Concept of Miracle*. Montreal: McGill-Queen's University Press, 1986.
- Larmer, Robert A.H. "Miracles and Criteria." *Sophia*, 23 (1984), pp. 5-12.
- Larmer, Robert A.H. "Miracles and Laws of Nature." *Dialogue*, 24, pp. 227-235.
- Levine, Michael, P. *Hume and the Problem of Miracles: A Solution*. Dordrecht: Kluwer Publishers, 1989.
- Lewis, C.S. *Miracles*. New York: Macmillan, 1947.
- Locke, John. *A Discourse on Miracles*. Published posthumously, 1706. Reprinted and edited by I.T. Ramsey. London: . and C. Black, 1958. Lowe, E.J. "Miracles and Laws of Nature." *Religious Studies*, 23 (1987), pp. 263-278.
- MacGill, Neil, W. "Miracles and Conservation Laws." *Sophia*, 31 (1992), pp. 79-87.
- Mackie, J.L. *The Cement of The Universe*. Oxford: Oxford University Press, 1974.
- Mackie, J.L. *The Miracle of Theism*. Oxford: Oxford University Press, 1982. Millican, Peter. "Hume's Theorem Concerning Miracles." *Philosophical Quarterly*, 43 (1993), pp. 489-495.
- Madden, Edward. "Causality and the Notion of Necessity." In *Boston Studies in the Philosophy of*

Science, vol. IV. Ed. R. Cohen and M. Wartofsky. Dordrecht: D. Reidel, 1969.

- Madden, Edwards. "A Third View of Causality." *Review of Metaphysics*, 23 (1969), pp. 67-84.
- Madden, Edward. "Nonlogical Necessity." *Idealistic Studies*, 5 (1975), pp. 17-19.
- Madden, Edward and Hare, P.H. "The Powers That Be." *Dialogue*, 10 (1971), pp. 12-31.
- Madden, Edward. "Natural Necessity." *New Scholasticism*, 47 (1973), pp.214-227.
- Mavrodes, George. "Miracles and the Laws of Nature." *Faith and Philosophy*, 2 (1985), pp. 333-346.
- McKinnon, Alastair. "‘Miracle’ and ‘Paradox.’" *American Philosophical Quarterly*, 4 (1967), pp. 308-314.
- Merrill, Kenneth, R. "Hume's ‘Of Miracles,’ Peirce, and the Balancing of Likelihoods." *Journal of the History of Philosophy*, 29 (1991), pp. 85-113.
- Nelson, John. "The Burial and Resurrection of Hume's Essay ‘Of Miracles.’" *Hume Studies*, 12 (1986), pp. 57-76.
- Nowell-Smith, P.K. "Miracles." In *New Essays in Philosophical Theology*. Ed. Anthony Flew and Alasdair MacIntyre. London: SCM, 1955, pp. 243-253.
- Odegard, Douglas. "Miracles and Good Evidence." *Religious Studies*, 18 (1982), pp. 37-46.
- Otte, Richard. "Schlesinger on Miracles." *Faith and Philosophy*, 10 (1993), pp. 93-98.
- Overall, Christine. "Miracles As Evidence Against The Existence of God." *Southern Journal of Philosophy*, 23 (1985), pp. 347-353.
- Owen, David. "Hume Versus Price On Miracles and Prior Probabilities: Testimony and The Bayesian Calculation." *Philosophical Quarterly*, 37 (1987), pp. 187-202.
- Paley, William. *A View of The Evidences of Christianity*. Published in 1794.
- Penelhum, Terrence. *Religion and Rationality*. New York: Harper and Row, 1971.
- Rein, Andrew. "Repeatable Miracles." *Analysis*, 46 (1986), pp. 109-112.
- Reppert, Victor. "Miracles and the Case for Theism." *International Journal for Philosophy of Religion*, 25 (1989), pp. 35-51.
- Robinson, Guy. "Miracles." *Ratio*, ix (1967), pp. 155-166.
- Root, Michael. "Miracles and the Uniformity of Nature." *American Philosophical Quarterly*, 26 (1989), pp. 333-342.
- Rutherford, Donald. "Natures, Laws, and Miracles: The Roots of Leibniz's Critique of Occasionalism." In *Causation in Early Modern Philosophy*. Ed. S. Nadler. University Park: Pennsylvania State University Press, 1993.
- Schlesinger, George. "Miracles and Probabilities." *Nous*, XXI (1987), pp. 219-232.
- Schlesinger, George, N. "The Credibility of Extraordinary Events." *Analysis*, 51 (1991), pp. 120-126.
- Schoen, Edward, L. "David Hume and the Mysterious Shroud of Turin." *Religious Studies*, 27 (1991), pp. 209-222.
- Schooman, A.P. *The Metaphysics of Religious Belief*. UK: Avebury Publishing, 1990.
- Slupik, Chris. "A New Interpretation of Hume's ‘Of Miracles.’" *Religious Studies*, 31 (1995), pp. 517-536.
- Smart, Ninian. *Philosophers and Religious Truth*. Chapter 2, "Miracles and David Hume." London: SCM Press, 1964.
- Sobel, Jordan H. "Hume's Theorem on Testimony Sufficient to Establish a Miracle."

Philosophical Quarterly, 41 (1991), pp.229-237.
- Sobel, Jordan. "On the Evidence of Testimony for Miracles: A Bayesian Interpretation of Hume's Analysis." *Philosophical Quarterly*, 37 (1987), pp. 166-186.
- Sorensen, Roy. "Hume's Skepticism Concerning Reports of Miracles." *Analysis*, 43 (1983), p. 60.
- Sosa, Ernest, and Tooley, Michael, eds. *Causation*. Oxford: Oxford University Press, 1993.
- Sosa, Ernest. ed. *Causation and Conditionals*. Oxford: Oxford University Press, 1975.
- Stewart, M.A. "Hume's Historical View of Miracles." In *Hume and Hume's Connexions*. Ed. M.A. Stewart. University Park: Pennsylvania State University Press, 1995.
- Stove, D.C. *Probability and Hume's Inductive Skepticism*. Oxford: Clarendon Press, 1973.
- Stove, D.C. "Hume Probability and Induction." In *Hume*. Ed. V.C. Chappel. Notre Dame: Notre Dame University Press, 1966.
- Strauss, David. *Das Leben Jesu*. 2 vols. Tbingen. Translated by Mary Ann Evans as *The Life of Jesus Critically Examined*. London, 1848.
- Swinburne, Richard. *The Concept of Miracle*. London: Macmillan, 1970. This contains a brief bibliography of the discussion of miracles in historical theological and philosophical perspective.
- Taylor, A.E. *David Hume and the Miraculous*. Cambridge: Cambridge University Press, 1927.
- Tennant, F.R. *Miracle*. Cambridge: Cambridge University Press, 1927.
- Thornton, J.C. "Miracles and God's Existence." *Philosophy*, 59 (1984), pp. 219-230.
- Tillich, Paul. *Systematic Theology*, vol.1. London: Nisbet, 1953.
- Tillotson, John. "Discourse Against Transubstantiation." In *Works*. Vol.2. Ed. T. Birch. London, 1820.
- Walker, Ian. "Miracles and Violations." *International Journal For Philosophy of Religion*, 13 (1982), pp.103-108.
- Walker, Keith. "Miracles and Coincidences." *Sophia*, 22 (1983), pp. 29-36.
- Ward, Keith. "Miracles and Testimony." *Religious Studies*, 21 (1985), pp. 131-145.
- Wei, Tan Tai. "Mr. Young on Miracles." *Religious Studies*, 10 (1974), pp. 333-337.
- Wei, Tan Tai. "Recent Discussions on Miracles." *Sophia*, 11 (1972), pp. 21-28.
- Wiebe, Phillip, H. "Authenticating Biblical Reports of Miracles." *Journal of Philosophical Research*, 18 (1993), pp. 309-325.
- Williams, T.C. *The Idea of the Miraculous: The Challenge to Science and Religion*. New York: St Martin's Press, 1991.
- Wilson, Fred. "The Logic of Probabilities in Hume's Argument Against Miracles." *Hume Studies*, 15 (1989), pp. 255-275.
- Yandell, Keith. "Religious Experience and Rational Appraisal." *Religious Studies*, 10 (1974), pp. 173-187.
- Young, Robert. "Miracles and Epistemology." *Religious Studies*, 8 (1972), pp.115-126.
- Young, Robert. "Miracles and Physical Impossibility." *Sophia*, 2 (1972), pp. 29-35.

Other Internet Resources

Related Entries

Bayes' Theorem | causation: probabilistic | causation: the metaphysics of | cause and effect | Hume, David | induction: problem of | religion: epistemology of

Michael P. Levine
mlevine@arts.uwa.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 4, 1996
Content last modified: August 15, 1997

Probabilistic Causation

"Probabilistic Causation" designates a group of philosophical theories that aim to characterize the relationship between cause and effect using the tools of probability theory. A primary motivation for the development of such theories is the desire for a theory of causation that does not presuppose physical determinism. The central idea behind these theories is that causes raise the probabilities of their effects, all else being equal. As we shall see, a great deal of the work that has been done in this area has been concerned with making the *ceteris paribus* clause more precise. Issues within, and objections to, probabilistic theories of causation will also be discussed.

1. Introduction

According to David Hume, causes are sufficient conditions for their effects: "We may define a cause to be *an object, followed by another, and where all the objects similar to the first, are followed by objects similar to the second.*" (1748, section VII.) Later writers refined Hume's theory, but still characterized the causal relation in terms of necessary and sufficient conditions. One of the best known approaches is Mackie's theory of *inus* conditions. An inus condition for some effect is an insufficient but non-redundant part of an unnecessary but sufficient condition. Suppose, for example, that a lit match causes a forest fire. The lighting of the match, by itself, is not sufficient; many matches are lit without ensuing forest fires. The lit match is, however, a part of some constellation of conditions that are jointly sufficient for the fire. Moreover, given that this set of conditions occurred, rather than some other set sufficient for fire, the lighting of the match was necessary: without it, the fire would not have occurred.

The necessity/sufficiency approach makes causation incompatible with indeterminism: if an event is not

determined to occur, then no event can be a part of a sufficient condition for that event. (An analogous point may be made about necessity.) The recent success of quantum mechanics -- and to a lesser extent, other theories employing probability -- has shaken our faith in determinism. Thus it has struck many philosophers as desirable to develop a theory of causation that does not presuppose determinism. The central idea behind probabilistic theories of causation is that causes *raise the probability* of their effects; an effect may still occur in the absence of a cause or fail to occur in its presence.

Suggested Readings: Hume (1748), especially section VII; Mackie (1974), especially chapter 3.

2. Is Probabilistic Causation an Oxymoron?

Many philosophers find the idea of indeterministic causation counterintuitive. Indeed, the word "causality" is sometimes used as a synonym for determinism. A strong case for indeterministic causation can be made by considering the epistemic warrant for causal claims. There is now very strong empirical evidence that smoking causes lung cancer. Yet the question of whether there is a deterministic relationship between smoking and lung cancer is wide open. The formation of cancer cells depends upon mutation, which is a strong candidate for being an indeterministic process. Moreover, whether an individual smoker develops lung cancer or not depends upon a host of additional factors, such as whether or not she is hit by a bus before cancer cells begin to form. Thus the price of preserving the intuition that causation presupposes determinism is agnosticism about even our best supported causal claims.

Suggested Readings: Suppes (1970) makes the case for a probabilistic theory of causation in his introduction. Humphreys (1989), contains a sensitive treatment of issues involving indeterminism and causation; see especially sections 10 and 11.

3. Main Developments

3.1 The Central Idea

The central idea that causes raise the probability of their effects can be expressed formally using the apparatus of conditional probability. Let A, B, C,... represent entities that potentially stand in causal relations. Depending upon the account, these may be particular events, such as the assassination of Archduke Ferdinand, or event types, such as exposure to ultraviolet radiation. We will discuss this issue at greater length in section 4.3 below. For now we will adopt the generic word "factor" to describe the relevant entities. Let P be a probability function, satisfying the normal rules of the probability calculus, such that P(A) represents the empirical probability that factor A occurs or is instantiated (and likewise for

the other factors). The issue of how empirical probability is to be interpreted will not be addressed here. (See the entry under probability). The probability of B, given A, is represented as a conditional probability:

$$P(B|A) = P(A \& B)/P(A).$$

One natural way of understanding the idea that A raises the probability of B is that $P(B|A) > P(B|\text{not-}A)$. Thus a first attempt at a probabilistic theory of causation would be:

> *PR:* A causes B if and only if $P(B|A) > P(B|\text{not-}A)$.

This formulation is labeled *PR* for "Probability-Raising".

There are two central problems with this theory. The first is that probability-raising is symmetric: if $P(B|A) > P(B|\text{not-}A)$, then $P(A|B) > P(A|\text{not-}B)$. The causal relation, however, is *asymmetric:* if A causes B, then typically B does not cause A. The problem of causal asymmetry arises for virtually every theory of causation, and probabilistic theories of causation are no exception.

The second problem concerns *spurious correlations*. If, for example, A and B are both caused by some third event, say C, then it may be that $P(B|A) > P(B|\text{not-}A)$ even though A does not cause B. For example, let A be an individual's having yellow-stained fingers, and B that individual's having lung cancer. Then we would expect that $P(B|A) > P(B|\text{not-}A)$. The reason that those with yellow-stained fingers are more likely to suffer from lung cancer is that smoking tends to produce both effects. Because individuals with yellow-stained fingers are more likely to be smokers, they are also more likely to suffer from lung cancer. Intuitively, the way to address this problem is to require that causes raise the probabilities of their effects *ceteris paribus.* The history of probabilistic causation is to a large extent a history of attempts to resolve these two central problems.

3.2 Spurious Correlations

Hans Reichenbach introduced the terminology of "screening off" to apply to a particular type of probabilistic relationship. If $P(B|A \& C) = P(B|C)$, then C is said to screen A off from B. Intuitively, C renders A probabilistically irrelevant to B. With this notion in hand, we can attempt to avoid the problem of spurious correlations by adding a 'no screening off' condition to the basic probability-raising condition:

> *NSO:* Factor A occurring at time t, is a cause of the later factor B if and only if:
>
> 1. $P(B|A) > P(B|\text{not-}A)$
> 2. There is no factor C, occurring earlier than or simultaneously with A, that screens A off from B.

We will call this the *NSO*, or 'No Screening Off' formulation. Suppose, as in our example above, that smoking (C) causes both yellow-stained fingers (A) and lung cancer (B). Then smoking will screen yellow-stained fingers off from lung cancer: given that an individual smokes, his yellow-stained fingers have no impact upon his probability of developing lung cancer.

The second condition of *NSO* does not suffice to resolve the problem of spurious correlations, however. This condition was added to eliminate cases where spurious correlations give rise to factors that raise the probability of other factors without causing them. Spurious correlations can also give rise to cases where a cause does not raise the probability of its effect. So genuine causes need not satisfy the *first* condition of *NSO*. Suppose, for example, that smoking is highly correlated with exercise: those who smoke are much more likely to exercise as well. Smoking is a cause of heart disease, but suppose that exercise is an even stronger preventative of heart disease. Then it may be that smokers are, over all, less likely to suffer from heart disease than non-smokers. That is, letting A represent smoking, C exercise, and B heart disease, P(B|A) < P(B|not-A). Note, however, that if we conditionalize on whether one exercises or not, this inequality is reversed: P(B|A & C) > P(B|not-A & C), and P(B|A & not-C) > P(B|not-A & not-C).

The next step is to replace conditions 1 and 2 with the requirement that causes must raise the probability of their effects in *test situations:*

> *TS:* A causes B if P(B|A & T) > P(B|not-A & T) for every test situation T.

A test situation is a conjunction of factors. When such a conjunction of factors is conditioned on, those factors are said to be "held fixed". To specify what the test situations will be, then, we must specify what factors are to be held fixed. In the previous example, we saw that the true causal relevance of smoking for lung cancer was revealed when we held exercise fixed, either positively (conditioning on C) or negatively (conditioning on not-C). This suggests that in evaluating the causal relevance of A for B, we need to hold fixed other causes of B, either positively or negatively. This suggestion is not entirely correct, however. Let A and B be smoking and lung cancer as above. Suppose C is a causal intermediary, say the presence of tar (and other carcinogens) in the lungs. If A causes B exclusively via C, then C will screen A off from B: given the presence (absence) of carcinogens in the lungs, the probability of lung cancer is not affected by whether those carcinogens got there by smoking (are absent despite smoking). Thus we will not want to hold fixed any causes of B *that are themselves caused by A*. Let us call the set of all factors that are causes of B, but are not caused by A, the set of *independent* causes of B. A test situation for A and B will then be a maximal conjunction, each of whose conjuncts is either an independent cause of B, or the negation of an independent cause of B.

Note that the specification of factors that need to be held fixed appeals to causal relations. This appears to rob the theory of its status as a *reductive analysis* of causation. We will see in section 4.4 below, however, that the issue is substantially more complex than that. In any event, even if there is no reduction of causation to probability, a theory detailing the systematic connections between causation and probability would be of great philosophical interest.

TS can be generalized in a number of ways. For example, one could define a 'negative cause' or 'preventer' or 'inhibitor' as a factor that lowers the probability of its 'effect' in all test situations, and a 'mixed' or 'interacting' cause as one that affects the probability of its 'effect' in different ways in different test situations. Or one could define causal relationships between variables that are non-binary, such as caloric intake and blood pressure. In principle, there are infinitely many ways in which one variable might depend probabilistically on another, even holding fixed some particular test situation, so this approach abandons any neat classification of causal factors into causes and preventers. These generalizations will also suggest revisions of the method for constructing test situations, since they suggest different sorts of factors to be held fixed.

An alternative approach to the problem of spurious correlations is through counterfactuals. According to a probabilistic counterfactual theory of causation *(PC),* A causes B if both occur and the probability that B would occur, at the time of A's occurrence, was much higher than it *would have been* at the corresponding time if A had not occurred. This counterfactual is to be understood in terms of possible worlds: it is true if, in the nearest possible world(s) where A does not occur, the probability of B is much lower than it was in the actual world. On this account, one does not compare conditional probabilities, but unconditional probabilities in different possible worlds. The test situation is not some specified conjunction of factors, but the sum total of all that remains unchanged in moving to the nearest possible world(s) where A does not occur. Obviously a great deal hinges here upon the account of what makes some worlds nearer than others; for more on this issue, see the entry under "causation, counterfactual theories."

Suggested Readings: This section more or less follows the main developments in the history of probabilistic theories of causation. Versions of the *NSO* theory are found in Reichenbach (1956, section 23), and Suppes (1970, chapter 2). Salmon (1980) is an influential critique of these theories. The first version of *TS* was presented in Cartwright (1979). Eells (1991, chapters 2, 3, and 4) and Hitchcock (1993) carry out the two generalizations of *TS* described. Lewis (1986) is the locus classicus for *PC*. Good (1961, 1962) is an early essay on probabilistic causation that is rich in insights, but has had surprisingly little influence on the formulation of later theories.

3.3 Asymmetry

The second major problem with the basic probability-raising idea was that the relationship of probability-raising is symmetrical. One way of cutting through the Gordian knot is to require that causes precede their effects in time. This has several systematic disadvantages. It rules out the possibility of backwards-in-time causation *a priori*, whereas many believe that it is only a contingent fact that causes precede their effects in time. This is less of a worry if one is not concerned to give a conceptual analysis of causation. Second, this approach rules out the possibility of developing a causal theory of temporal order (on pain of vicious circularity), a theory that has seemed attractive to some philosophers. Note also that while assigning temporal locations to particular events is entirely coherent, it is not so clear what it means to say that one property or event type occurs before another. For example, what does it mean to say that smoking precedes lung cancer? There have been many episodes of smoking, and many of lung cancer,

and not all of the former occurred prior to all of the latter. This will be a problem for those who are interested in providing a probabilistic theory of causal relations among properties or event types.

A more ambitious approach to the problem of causal asymmetry is to try to characterize that asymmetry in terms of probability relations alone. The best-known proposal of this sort is due to Hans Reichenbach. Suppose that factors A and B are positively correlated:

1. P(A & B) > P(A)P(B)

It is easy to see that this will hold exactly when A raises the probability of B and vice versa. Suppose, moreover, that there is some factor C having the following properties:

2. P(A & B|C) = P(A|C)P(B|C)

3. P(A & B|not-C) = P(A|not-C)P(B|not-C)

4. P(A|C) > P(A|not-C)

5. P(B|C) > P(B|not-C).

In this case, the trio ACB is said to form a *conjunctive fork*. Conditions 2 and 3 stipulate that C and not-C screen off A from B. As we have seen, this sometimes occurs when C is a common cause of A and B. Conditions 2 through 5 entail 1, so in some sense C explains the correlation between A and B. If C occurs earlier than A and B, and there is no event satisfying 2 through 5 that occurs later than A and B, then ACB is said to form a conjunctive fork *open to the future.* Analogously, if there is a future factor satisfying 2 through 5, but no past factor, we have a conjunctive fork open to the past. If a past factor C and a future factor D both satisfy 2 through 5, then ACBD forms a closed fork. Reichenbach's proposal was that the direction from cause to effect is the direction in which open forks predominate. In our world, there are many forks open to the future, few or none open to the past.

It is not clear, however, that this asymmetry between forks open to the past and open to the future will be as pervasive as this proposal seems to presuppose. In quantum mechanics, there are correlated effects that are believed to have no common cause that screens them off. Moreover, if ACB forms a conjunctive fork in which C precedes A and B, but C has a deterministic effect D which occurs after A and B, then ACBD will form a closed fork. A further difficulty with this proposal is that since it provides a global ordering of causes and effects, it seems to rule out a priori the possibility that some effects might precede their causes. More complex attempts to derive the direction of causation from probabilities have been offered; the issues here intersect with the problem of reduction, discussed in section 4.4 below.

Proponents of counterfactual theories of causation attempt to derive the asymmetry of causation from a corresponding asymmetry in the truth values of counterfactuals. For details see the entry for " causation, counterfactual theories."

Suggested Readings: Suppes (1970, chapter 2) and Eells (1991, chapter 5) define causal asymmetry in terms of temporal asymmetry. Reichenbach's proposal is presented in his (1956, chapter IV). Some difficulties with this proposal are discussed in Arntzenius (1993). Papineau (1993) is a good overall discussion of the problem of causal asymmetry within probabilistic theories.

4. Further Issues and Problems

4.1 Context-unanimity

According to *TS*, a cause must raise the probability of its effect in *every* test situation. This has been called the requirement of *context-unanimity*. This requirement is vulnerable to the following sort of counterexample. Suppose that there is a gene that has the following unusual effect: those that possess the gene have their chances of contracting lung cancer *lowered* when they smoke. This gene is very rare, let us imagine -- indeed, it need not exist at all in the human population, so long as humans have some non-zero probability of possessing this gene (perhaps as a result of a very improbable mutation). In this scenario, there would be test situations (those that hold fixed the presence of the gene) in which smoking lowers the probability of lung cancer: thus smoking would not be a cause of lung cancer according to the context-unanimity requirement. Nonetheless, it seems unlikely that the discovery of such a gene (or of the mere possibility of its occurrence) would lead us to abandon the claim that smoking causes lung cancer.

This line of objection is surely right about our ordinary use of causal language. It is nonetheless open to the defender of context-unanimity to respond that she is interested in supplying a precise concept to replace the vague notion of causation that corresponds to our everyday usage. In a population consisting of individuals lacking the gene, smoking causes lung cancer. In a population consisting entirely of individuals who possess the gene, smoking prevents lung cancer. In contexts where one desires causal information for purposes of deliberation (say concerning whether to smoke), it is this more precise type of information that is desired.

Suggested Readings: Dupré (1984) presents this challenge to the context-unanimity requirement, and offers an alternative. Eells (1991, chapters 1 and 2), defends context-unanimity using the idea that causal claims are made relative to a population.

4.2 Potential Counterexamples

Given the basic probability-raising idea, one would expect putative counterexamples to probabilistic theories of causation to be of two basic types: cases where causes fail to raise the probabilities of their effects, and cases where non-causes raise the probabilities of non-effects. The discussion in the literature has focused almost entirely on the first sort of example. Consider the following example, due to Deborah Rosen. A golfer badly slices a golf ball, which heads toward the rough, but then bounces off a tree and into the cup for a hole in one. The golfer's slice lowered the probability that the ball would wind up in the cup, yet nonetheless caused this result. One way of avoiding this problem is to attend to the probabilities that are being compared. If we label the slice A, not-A is a disjunction of several alternatives. One such alternative is a clean shot -- compared to this alternative, the slice lowered the probability of a hole-in-one. Another alternative is no shot at all, relative to which the slice increases the probability of a hole-in-one. By making the latter sort of comparison, we can recover our original intuitions about the example.

For an example of the second type, suppose that two gunmen shoot at a target. Each has a certain probability of hitting, and a certain probability of missing. Assume that none of the probabilities are one or zero. As a matter of fact, the first gunman hits, and the second gunman misses. Nonetheless, the second gunman did fire, and by firing, increased the probability that the target would be hit, which it was. While it is obviously wrong to say that the second gunman's shot caused the target to be hit, it would seem that a probabilistic theory of causation is committed to this consequence. A natural approach to this problem would be to try to strengthen the probabilistic theory of causation with a requirement of spatiotemporal connection between cause and effect (see the entry on "causation, causal processes"), but to date, no successful proposal along these lines has been proffered.

Suggested Readings: Salmon (1980) presents several examples of probability-lowering causes. Hitchcock (1995) presents a response. Woodward (1990) describes the structure that is instantiated in the example of the two gunmen. Humphreys (1989, section 14) responds. Menzies (1989, 1996) discusses examples involving causal pre-emption where non-causes raise the probabilities of non-effects.

4.3 Singular and General Causation

We make at least two different kinds of causal claim. *Singular* causal claims, such as "Jill's heavy smoking during the '80's caused her to develop lung cancer," relate particular events that have spatiotemporal locations. *General* causal claims, such "smoking causes lung cancer" relate event types or properties. With this distinction in mind, we may note that the counterexamples mentioned above are both formulated in terms of singular causation. The examples do not undermine the *General* causal claims that a probabilistic theory of causation would appear to license in these cases: slices prevent (are negative causes of) holes-in-one; shooting at targets causes them to be hit. So one possible reaction to the counterexamples of the previous section would be to maintain that the probabilistic theory of causation whose development was sketched in section 3 above is a theory of general causation only, and that singular causation requires a distinct philosophical theory. One consequence of this move is that there are (at least) two distinct species of causal relation, each requiring its own philosophical account--not an altogether happy predicament.

Suggested Readings: The need for distinct theories of singular and general causation is defended in Good (1961, 1962), Sober (1985), and Eells (1991, introduction and chapter 6). Eells (1991, chapter 6) offers a distinct probabilistic theory of singular causation in terms of the temporal evolution of probabilities. Carroll (1991) and Hitchcock (1995) offer two quite different lines of response.

4.4. Reduction and Circularity

Returning to the theories outlined in section 3, recall that theory *NSO* was an attempt at a *reductive analysis* of causation in terms of probabilities (and perhaps also temporal order). By contrast, *TS* defines causal relations in terms of probabilities conditional upon specifications of test conditions, which are themselves characterized in causal terms. Thus it appears that the latter theories cannot be analyses of causation, since causation appears in the analysans. Given that *TS* contains much needed improvements over *NSO*, it looks as though there can be no reduction of causation to probabilities. This may be giving up too soon, however. In order to determine whether a probabilistic reduction of causation is possible, the central issue is not whether the word ‘cause’ appears in both the analysandum and the analysans; rather, the key question should be whether, given an assignment of probabilities to a set of factors, there is a unique set of causal relations among those factors compatible with the probability assignment and the theory in question. Suppose that a set of factors, and a system of causal relations among those factors is given: call this the *causal structure CS*. Let *T* be a theory connecting causal relations among factors with probabilistic relations among factors. Then the causal structure *CS* will be *probabilistically distinguishable* relative to *T*, if for every assignment of probabilities to the factors in *CS* that is compatible with *CS* and *T*, *CS* is the unique causal structure compatible with *T* and those probabilities. (One could formulate a weaker sense of distinguishability by requiring that only some assignment of probabilities uniquely determines *CS*). Intuitively, *T* allows you to infer that the causal structure is in fact *CS* given the probability relations between factors. Given a probabilistic theory of causation *T*, it is possible to imagine many different properties it might have. Here are some possibilities:

1. All causal structures are probabilistically distinguishable relative to *T*

2. All causal structures having some interesting property are probabilistically distinguishable relative to *T*

3. Any causal structure can be embedded in a causal structure that is probabilistically distinguishable relative to *T*

4. The actual causal structure of the world (assuming there is such a thing) is probabilistically distinguishable relative to *T* .

It is not obvious which type of distinguishability properties a theory must have in order to constitute a reduction of causation to probabilities. This sort of approach to the question of probabilistic reduction is quite new, and currently an active area of investigation.

Suggested Readings: The most detailed treatment of probabilistic distinguishability is given in Spirtes, Glymour and Scheines (1993); see especially chapter 4. Spirtes, Glymour and Scheines prove (theorem 4.6) a result along the lines of 3 for a theory that they propose. This work is very technical. An accessible presentation is contained in Papineau (1993), which defends a position along the lines of 4.

Bibliography

- Arntzenius, Frank (1993) "The Common Cause Principle," in Hull, Forbes, and Okruhlik (1993), pp. 227 - 237.
- Carroll, John (1991) "Property-level Causation?" *Philosophical Studies* **63**: 245-70.
- Cartwright, Nancy (1979) "Causal Laws and Effective Strategies," *Noûs* **13**: 419-437.
- Dupré, John (1984) "Probabilistic Causality Emancipated," in Peter French, Theodore Uehling, Jr., and Howard Wettstein, eds., (1984) *Midwest Studies in Philosophy IX* (Minneapolis: University of Minnesota Press), pp. 169 - 175.
- Eells, Ellery (1991) *Probabilistic Causality*. Cambridge, U.K.: Cambridge University Press.
- Good, I. J. (1961) "A Causal Calculus I," *British Journal for the Philosophy of Science* **11**: 305-18.
- Good, I. J. (1962) "A Causal Calculus II," *British Journal for the Philosophy of Science* **12**: 43-51.
- Hitchcock, Christopher (1993) "A Generalized Probabilistic Theory of Causal Relevance," *Synthese* **97**: 335-364.
- Hitchcock, Christopher (1995) "The Mishap at Reichenbach Fall: Singular vs. General Causation," *Philosophical Studies* **78**: 257 - 291.
- Hull, David, Mickey Forbes, and Kathleen Okruhlik, eds. (1993) *PSA 1992, Volume Two* (East Lansing: Philosophy of Science Association).
- Hume, David (1748) *An Enquiry Concerning Human Understanding.*
- Humphreys, Paul (1989) *The Chances of Explanation: Causal Explanations in the Social, Medical, and Physical Sciences*, Princeton: Princeton University Press.
- Lewis, David (1986) "Causation" and "Postscripts to 'Causation'," in *Philosophical Papers, Volume II*, Oxford: Oxford University Press, pp. 172-213.
- Mackie, John (1974) *The Cement of the Universe.* Oxford: Clarendon Press.
- Menzies, Peter (1989) "Probabilistic Causation and Causal Processes: A Critique of Lewis," *Philosophy of Science* **56**: 642-63.
- Menzies, Peter (1996) "Probabilistic Causation and the Pre-emption Problem", *Mind* **105**: 85-117.
- Papineau, David (1993) "Can We Reduce Causal Direction to Probabilities?" in Hull, Forbes and Okruhlik (1993), pp. 238-252.
- Reichenbach, Hans (1956) *The Direction of Time.* Berkeley and Los Angeles: University of California Press.
- Salmon, Wesley (1980) "Probabilistic Causality," *Pacific Philosophical Quarterly* **61**: 50 - 74.
- Sober, Elliott (1985) "Two Concepts of Cause" in Peter Asquith and Philip Kitcher, eds., *PSA 1984, Vol. II* (East Lansing: Philosophy of Science Association), pp. 405-424.

- Spirtes, Peter, Clark Glymour, and Richard Scheines (1993) *Causation, Prediction and Search*. New York: Springer-Verlag.
- Suppes, Patrick (1970) *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company.
- Woodward, James (1990) "Supervenience and Singular Causal Claims," in Dudley Knowles, ed., *Explanation and its Limits* (Cambridge, U.K: Cambridge University Press), pp. 211 - 246.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

causation: causal processes | causation: counterfactual theories of | cause and effect | conditionals: counterfactual | determinism, causal | events | Hume, David | physics: Reichenbach's common cause principle | probability calculus: interpretations of | quantum mechanics | time

Christopher Hitchcock
cricky@hss.caltech.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 11, 1997
Content last modified: July 17, 1997

Causal Processes

Taking their point of departure from what science tells us about the world rather than from our everyday concept of a ‘process’, philosophers interested in analysing causal processes have tended to see the chief task to be to distinguish *causal* processes such as atoms decaying and billiard balls moving across the table from *pseudo* processes such as moving shadows and spots of light. These philosophers have found, in the notion of a causal process, a key to understanding causation in general.

1. Russell's Theory of Causal Lines

An important forerunner of contemporary notions of causal processes is Bertrand Russell's account of causal lines. This may be surprising to those who are more accustomed to associate the name ‘Bertrand Russell’ with scepticism about causation. Russell's 1912/13 paper, ‘On the Notion of Cause’, is famous for the quote,

> The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm. (Russell, 1913, p. 1).

In that paper Russell argued that the philosopher's concept of causation involving, as it does, the law of universal determinism that every event has a cause and the associated concept of causation as a relation between events, is "otiose" and in modern science is replaced by the concept of causal laws understood in

terms of functional relations, where these causal laws are not necessarily deterministic.

However, in a later book written in 1948, entitled *Human Knowledge* Bertrand Russell outlines a similar view but does so in language which is much more flattering to causation. He still holds that the philosophical idea of causation should be seen as a primitive version of the scientific idea of causal laws. Nevertheless, his emphasis now is on certain postulates of causation which he takes to be fundamental to scientific (inductive) inference, and Russell's aim is to show how scientific inference is possible.

The problem with thinking about causal laws as the underpinning of scientific inference is that the world is a complex place, and while causal laws might hold true, they often do not obtain because of preventing circumstances, and it is impractical to bring in innumerable ‘unless’ clauses. But, even though there is infinite complexity in the world, there are also causal lines of quasi-permanence, and these warrant our inferences.

Russell elaborates these ideas into five postulates which he says are necessary "to validate scientific method" (1948, p. 487). The first is ‘The Postulate of Quasi-permanence’ which states that there is a certain kind of persistence in the world, for generally things do not change discontinuously. The second postulate, ‘Of Separable Causal Lines’, allows that there is often long term persistence in things and processes. The third postulate, ‘Of Spatio-temporal Continuity’ denies action at a distance. Russell claims "when there is a causal connection between two events that are not contiguous, there must be intermediate links in the causal chain such that each is contiguous to the next, or (alternatively) such that there is a process which is continuous." (1948, p. 487). ‘The Structural Postulate’, the fourth, allows us to infer from structurally similar complex events ranged about a centre to an event of similar structure linked by causal lines to each event. The fifth postulate, ‘Of Analogy’ allows us to infer the existence of a causal effect when it is unobservable.

The key postulate concerns the idea of causal lines or, in our terminology, causal processes. Russell's 1948 view is that causal lines replace the primitive notion of causation in the scientific view of the world, and not only replace but also explain the extent to which the primitive notion, causation, is correct. He writes,

> The concept "cause", as it occurs in the works of most philosophers, is one which is apparently not used in any advanced science. But the concepts that are used have been developed from the primitive concept (which is that prevalent among philosophers), and the primitive concept, as I shall try to show, still has importance as the source of approximate generalisations and pre-scientific inductions, and as a concept which is valid when suitably limited. (1948, p. 471).

Russell also says, "When two events belong to one causal line the earlier may be said to "cause" the later. In this way laws of the form "A causes B" may preserve a certain validity." (1948, p. 334). So Russell can be seen, in his 1948 book, as proposing the view that within limits causal lines, or causal processes, may be taken to analyse causation. So what is a causal line? Russell writes,

> I call a series of events a "causal line" if, given some of them, we can infer something about the others without having to know anything about the environment. (1948, p. 333).
>
> A causal line may always be regarded as a persistence of something, a person, a table, a photon, or what not. Throughout a given causal line, there may be constancy of quality, constancy of structure, or gradual changes in either, but not sudden change of any considerable magnitude. (1948, pp. 475-7).

So the trajectory through time of something is a causal line if it doesn't change too much, and if it persists in isolation from other things. A series of events which display this kind of similarity display what Russell calls ‘quasi-permanence’.

> The concept of more or less permanent physical object in its common-sense form involves "substance", and when "substance" is rejected we have to find some other way of defining the identity of a physical object at different times. I think this must be done by means of the concept "causal line". (1948, p. 333).

Elsewhere Russell writes,

> The law of quasi-permanence as I intend it ... is designed to explain the success of the common-sense notion of "things" and the physical notion of "matter" (in classical physics). ... a "thing" or a piece of matter is not to be regarded as a single persistent substantial entity, but as a string of events having a certain kind of causal connection with each other. This kind is what I call "quasi-permanence". The causal law that I suggest may be enunciated as follows: "Given an event at a certain time, then at any slightly earlier or slightly later time there is, at some neighbouring place, a closely similar event". I do not assert that this happens always, but only that it happens very often- sufficiently often to give a high probability to an induction confirming it in a particular case. When "substance" is abandoned, the identity, for commonsense, of a thing or a person at different times must be explained as consisting in what may be called a "causal line". (1948, pp. 475-7).

This has relevence for the question of identity through time, and in *Human Knowledge* we find that Bertrand Russell sees that there is an important connection between causal process and identity, namely, that the concept of a causal line can be used to explain the identity through time of an object or a person.

So what we may call Russell's causal theory of identity (Dowe, forthcoming) asserts that the identity over time of an object or a person consists in the different temporal parts of that person being all part of the one causal line. This is the causal theory of identity (Armstrong, 1980) couched in terms of causal processes or lines. A causal line in turn is understood by way of an inference which is licensed by the law of quasi permanence.

2. Objections to Russell's Theory

Wesley Salmon has urged a number of objections against Russell's theory of causal lines. (1984, p. 140-5). The first objection is that Russell's theory is couched in epistemic terms rather than ontological terms, yet causation is itself an ontic matter not an epistemic matter. Russell's account is formulated in terms of how we make inferences. For example, Russell says

> A "causal line," as I wish to define the term, is a temporal series of events so related that, given some of them, something can be inferred about the others whatever may be happening elsewhere. (1948, p. 459).

Salmon's criticism of this is precisely that it is formulated in epistemic terms, "for the existence of the vast majority of causal processes in the history of the universe is quite independent of human knowers." (1984, p. 145). Salmon, as we shall see in the next section, develops *his* account of causal processes as an explicitly ‘ontic’, as opposed to an ‘epistemic’ account. (1984, ch. 1).

There is a further reason why Russell's epistemic approach is unacceptable. While it is true that causal processes *do* warrant inferences of the sort Russell has in mind, it is not the case that all rational inferences are warranted by the existence (‘postulation’, in Russell's thinking) of causal lines. There are other types of causal structures besides a causal line. Russell himself gives an example: two clouds of incandescent gas of a given element both emit the same spectral lines, but are not causally connected. (1948, p. 455). Yet we may rightly make inferences from one to the other. A pervasive type of case is where two events are not directly causally connected but have a common cause.

The second objection is that Russell's theory of a causal line does not enable the distinction between pseudo and causal processes to be made, yet to delineate causal from pseudo processes is a key issue which needs to be addressed by any theory of causal processes. As Reichenbach argued (1958, pp. 147-9), as he reflected on the implications of Einstein's special theory of relativity, science requires that we distinguish between causal and pseudo processes. Reichenbach noticed that the central principle that nothing travels faster than the speed of light is ‘violated’ by certain processes. For example, a spot of light moving along a wall is capable of moving faster than the speed of light. (One needs just a powerful enough light and a wall sufficiently large and sufficiently distant.) Other examples include shadows, and the point of intersection of two rulers (see Salmon's clear exposition in his 1984, pp. 141-4). Such pseudo processes, as we shall call them (Reichenbach called them "unreal sequences"; 1958, pp. 147-9), do not violate special relativity, Reichenbach argued, simply because they are not causal processes, and the principle that nothing travels faster than the speed of light applies only to causal processes. Thus special relativity demands a distinction between causal and pseudo processes. But Russell's theory doesn't explain this distinction, because both causal processes and pseudo processes display constancy of structure and quality; and both licence inferences of the sort Russell has in mind. For example, the phase velocity of a wave packet is a pseudo process but the group velocity is a causal process; yet both licence reliable predictions.

3. Salmon's Process Theory

In this section we consider Wesley Salmon's theory of causality as presented in his book *Scientific Explanation and the Causal Structure of the World* (1984). Although it draws on the work of Reichenbach and Russell, Salmon's theory is highly original and contains many innovative contributions. Salmon's broad objective is to offer a theory which is consistent with the following assumptions: (a) causality is an *objective* feature of the world; (b) causality is a *contingent* feature of the world; (c) a theory of causality must be consistent with the possibility of *indeterminism*; (d) the theory should be (in principle) *time-independent* so that it is consistent with a causal theory of time; (e) the theory should not violate Hume's strictures concerning ‘hidden powers’.

Salmon treats causality as primarily a characteristic of continuous processes rather than as a relation between events. His theory involves two elements, the *production* and the *propagation* of causal influence. (See, for example, 1984, p. 139.) The latter is achieved by causal processes. Salmon defines a process as anything that displays consistency of structure over time. (1984, p.144). To distinguish between causal and pseudo processes (which Reichenbach called "unreal sequences"; 1958, pp. 147-9). Salmon makes use of Reichenbach's ‘mark criterion’: a process is causal if it is capable of transmitting a local modification in structure (a ‘mark’) (1984, p. 147). Drawing on the work of Bertrand Russell, Salmon seeks to explicate the notion of ‘transmission’ by the ‘at-at theory’ of mark transmission. The principle of mark transmission (MT) states:

> MT: Let P be a process that, in the absence of interactions with other processes would remain uniform with respect to a characteristic Q, which it would manifest consistently over an interval that includes both of the space - time points A and B (A - B). Then, a mark (consisting of a modification of Q into Q*), which has been introduced into process P by means of a single local interaction at a point A, is transmitted to point B if [and only if] P manifests the modification Q* at B and at all stages of the process between A and B without additional interactions. (1984, p. 148).

Salmon himself omits the ‘only if’ condition. However, as suggested by Sober (1987, p. 253), this condition is essential because the principle is to be used to identify pseudo processes on the grounds that they do not transmit a mark (Dowe, 1992b, p. 198). Thus for Salmon a causal process is one which can transmit a mark, and it is these spatiotemporally continuous processes that propagate causal influence.

To accompany this theory of the propagation of causal influence, Salmon also analyses the production of causal processes. According to Salmon, causal production can be explained in terms of causal forks, whose main role is the part they play in the production of order and structure of causal processes. The causal forks are characterised by statistical forks; to Reichenbach's ‘conjunctive fork’ Salmon has added the ‘interactive’ and ‘perfect’ forks, each corresponding to a different type of common-cause.

Firstly there is the ‘conjunctive fork’, where two processes arise from a special set of background conditions often in a non-lawful fashion. (Salmon, 1984, p. 179). In such a case we get a statistical

correlation between the two processes which can be explained by appealing to the common cause, which ‘screens off’ the statistical connection. This is the principle of the common cause (due originally to Reichenbach (1956)) which, stated formally, is that if, for two events A and B,

$$(1)\ P(A.B) > P(A).P(B)$$

holds, then look for an event C such that

$$(2)\ P(A.B|C) = P(A|C).P(B|C)$$

The events A, B, and C form a conjunctive fork (For the full account see Salmon, 1984, ch. 6). In Salmon's theory of causality, conjunctive forks produce structure and order from ‘de-facto’ background conditions.(1984, p. 179).

Secondly, there is the ‘interactive fork’, where an intersection between two processes produces a modification in both (1984, p. 170) and an ensuing correlation between the two processes cannot be screened off by the common cause. Instead, the interaction is governed by conservation laws. For example, consider a pool table where the cue ball is placed in such a position relative to the eight ball that, if the eight ball is sunk in one pocket A , the cue ball will almost certainly drop into the other pocket B. There is a correlation between A and B such that equation (1) holds. But the common cause C, the striking of the cue ball, does not screen off this correlation. Salmon has suggested that the interactive fork can be characterised by the relation

$$(3)\ P(A.B|C) > P(A|C).P(B|C)$$

together with (1). (1978, p. 704, n. 31). Interactive forks are involved in the production of modifications in order and structure of causal processes. (1982, p. 265; 1984, p. 179). In this paper ‘interactive fork’ is used to mean precisely ‘a set of three events related according to equations (1) and (3)’.

The idea of a causal interaction is further analysed by Salmon in terms of the notion of mutual modification. The principle of causal interaction (CI) states:

> CI: Let P1 and P2 be two processes that intersect with one another at the space-time S, which belongs to the histories of both. Let Q be a characteristic of that process P1 would exhibit throughout an interval (which includes subintervals on both sides of S in the history of P1) if the intersection with P2 did not occur; let R be a characteristic that process P2 would exhibit throughout an interval (which includes subintervals on both sides of S in the history of P2) if the intersection with P1 did not occur. Then, the intersection of P1 and P2 at S constitutes a causal interaction if (1) P1 exhibits the characteristic Q before S, but it exhibits a modified characteristic Q* throughout an interval immediately following S; and (2) P2 exhibits R before S but it exhibits a modified characteristic R′ throughout an interval immediately following S. (1984, p. 171).

Thirdly, there is the perfect fork, which is the deterministic limit of both the conjunctive and interactive fork. It is included as a special case because in the deterministic limit the interactive fork is indistinguishable from the conjunctive fork. (1984, pp. 177-8). Thus, a perfect fork could be involved in either the production of order and structure, or the production of changes in order and structure of causal processes.

4. Objections to Salmon's Theory

The major objection against Samon's account of causal processes concerns the adequacy of the mark theory (Dowe, 1992a; 1992b; Kitcher, 1989). The mark transmission (MT) principle carries a considerable burden in Salmon's account, for it provides the criterion for distinguishing causal from pseudo processes. However, it has serious shortcomings in doing this. In fact, it fails on two counts: it excludes many causal processes; and it fails to exclude many pseudo processes. We shall consider each of these problems in turn.

1. MT *excludes causal processes*. Firstly, the principle requires that processes display *a degree of uniformity over a time period*. This distinguishes processes (causal and pseudo) from ‘spatiotemporal junk’, to use Kitcher's term. One problem with this is that it seems to exclude many causal effects which are short lived. For example, short lived subatomic particles play important causal roles, but they don't seem to qualify as causal processes. On any criterion there are causal processes which are ‘relatively short lived’. Also, the question concerning how long a regularity must persist raises philosophical difficulties about degrees which need answering before we have an adequate distinction between processes and spatiotemporal junk. However, if these were the only difficulties I think that the theory could be saved. Unfortunately, they are not.

More seriously, the MT principle requires that causal processes would remain uniform *in the absence of interactions* and that marks can be transmitted *in the absence of additional interventions*. However, in real situations processes are continuously involved in interactions of one sort or another.(Kitcher, 1989, p. 464). Even in the most idealised of situations interactions of sorts occur. For example, consider a universe that contains only one single moving particle. Not even this process moves in the absence of interactions, for the particle is forever intersecting with spatial regions. If we required that the interactions be causal (at the risk of circularity), then it is still true that in real cases there are many causal interactions continuously affecting processes. Even in carefully controlled scientific experiments there are many (admittedly irrelevant) causal interactions going on. Further, Salmon's central insight that causal processes are self propagating is not entirely well founded. For while some causal processes (light radiation, inertial motion) are self propagating, others are not. Falling bodies and electric currents are moved by their respective fields. (In particular there is no electric counterpart to inertia.) Sound waves are propagated within a medium, and simply do not exist ‘in the absence of interactions’. Such processes require a ‘causal background’, some can even be described as being a series of causal interactions. These causal processes *cannot* move in the absence of interactions. Thus there are a whole range of causal processes which are excluded by the requirement that they would remain uniform in the absence of any

interactions.

It seems desirable, therefore, to abandon the requirement that a causal process is one that is capable of transmitting a mark in the absence of further interactions. However, the requirement is there for a reason, and that is that without it the theory is open to the objection that certain pseudo processes will count as being capable of transmitting marks. Salmon considers a case where a moving spot is marked by a red filter held up close to the wall. If someone ran alongside the wall holding up the filter, then it seems that the modification to the process is transmitted beyond the space-time locality of the original marking interaction. Thus there are problems if the requirement is kept, and there are problems if it is omitted. So it is not clear how the theory can be saved from the problem that some causal processes can not move in the absence of further interactions.

2. MT *fails to exclude pseudo processes*. Salmon's explicit intention in employing the MT principle is to show how pseudo processes are different from causal processes. If MT fails here then it fails its major test. However, a strong case can be made for saying that it does indeed fail this test.

Firstly, there are cases where pseudo processes qualify as being capable of transmitting a mark, because of the *vagueness of the notion of a characteristic*. We have seen that Salmon's approach to causality is to give an informal characterisation of the concepts of ‘production’ and ‘propagation’. In these characterisations, the primitive notions include ‘characteristic’, but nothing precise is said about this notion. While Salmon is entitled to take this informal approach, in this case more needs to be said about a primitive notion such as ‘characteristic’, at least indicating the range of its application, because the vagueness renders the account open to counter-examples.

For example, in the early morning the top (leading) edge of the shadow of the Sydney Opera House has the characteristic of being closer to the Harbour Bridge than to the Opera House. But later in the day (at time *t* say), this characteristic changes. This characteristic qualifies as a mark by IV, since it is a change in a characteristic introduced by the local intersection of two processes, namely, the movement of the shadow across the ground, and the (stationary) patch of ground which represents the midpoint between the Opera House and the Harbour Bridge. By III this mark which the shadow displays continuously after time *t*, is transmitted by the process. Thus, by II, the shadow is a causal process. This is similar to Sober's counter-example of where a light spot ‘transmits’ the characteristic of occurring after a glass filter is bolted in place. (1987, p. 254).

So there are some restrictions that need to be placed on the type of property allowed as a characteristic. Having the property of "occurring after a certain time" (Sober, 1987, p. 254),or the property of "being the shadow of a scratched car" (Kitcher, 1989, p. 638) or the property of "being closer to the Harbour Bridge than to the Opera House" (Dowe, 1992b, sec. 2.2) can qualify a shadow to be a causal process. There is a need to specify what kinds of properties can count as the appropriate characteristics for marking. It is not sufficient to say that the mark has to be introduced by a single local interaction, for as the above discussion suggests it is always possible to identify a single local interaction.

The difficulty lies in the type of characteristic allowed. A less informal approach to the subject could have provided, for example, a restriction of ‘property’ to ‘non-relational property’, thereby avoiding this particular problem.

There are a number of possible ways to provide a more precise account of ‘characteristic’, either in philosophical terms such as ‘property’; or in terms of precise scientific notions such as ‘molecular structure’, ‘energy’ or ‘information’. In physics and chemistry description of the structure of a molecule, or larger solid body is given in terms of geometrical arrangement as well as the constituent particles and bonding forces. In biology the *structure* of a cell refers to its geometry as well as its constituents. Clearly a specific definition such as ‘chemical structure’ is not broad enough for Salmon's purposes. Although he uses examples such as the drug which causes a person to lose consciousness because it retains its ‘chemical structure’ as it is absorbed into the blood stream (1984, p. 155) it is nevertheless clear that the ‘structure’ of a car, a golf ball, a shadow, or a pulse of light is not simply ‘chemical structure.’ But perhaps this suggests a general characterisation in terms of constituent material, bonding forces and geometrical shape. I believe such an account has a lot of potential. For example, a chalk mark on a ball is a change in constituent material, a dent in a car is a change in geometrical shape, etc.

A different approach would be to link ‘characteristic’ to ‘property’ of which there are precise philosophical accounts available. (For example, (Armstrong, 1978)). Rogers takes this approach, defining the state of a process as the set of properties of the process at a given time. (Rogers, 1981, p. 203). A ‘law of non-interactive evolution’ gives the probability of the possible states at a later time, conditional on the actual state.

However, even if that approach were successful, there are further difficulties of a different kind. Firstly, there are cases of "derivative marks" (Kitcher, 1989, p. 463) where a pseudo process displays a modification in a characteristic on account of a change in the causal processes on which it depends. This change could either be in the source, or in the causal background. A change at the source would include cases where the spotlight spot is ‘marked’ by a coloured filter at the source (Salmon, 1984, p. 142) or a car's shadow is marked when a passenger's arm holds up a flag. (Kitcher, 1989, p. 463).

The clause ‘by means of a single *local* interaction’ is intended to exclude this type of example: but it is not clear that this works, for does not the shadow intersect with the modified sunlight pattern *locally*? It is true that the ‘modified sunlight pattern’ originated, or was caused by, the passenger raising his arm with the flag, but the fact that the marking interaction is the result of a chain of causes cannot be held to exclude those interactions, for genuine marking interactions are always the result of a chain of causal processes and interactions. (Kitcher, 1989, p. 464) Similarly, there is a local spacetime intersection of the spotlight spot and the red beam.

However, even if the ‘local’ requirement did exclude these cases, there are other cases where pseudo processes can be marked by changes in the causal background which are local. For example, imagine that there is a long stretch of road where the side of the road is flat, then farther along there is a long fence close to the road. The shadow of the car will change its shape abruptly as the car reaches the fence. We

can say that there is a local interaction between the shadow and the beginning of the fence which produces a permanent modification to the shadow. A similar case is where a person runs around an astrodome holding up a red filter which modifies the spot. The clause ‘by means of a single local interaction’ is intended to block these cases: but if this is employed too heavily it would exclude causal processes as well, such as Salmon's paradigm case where a red filter modifies the light beam. In this case the filter continues to act on the beam, just like the fence or the moving filter cases.

In any case it is possible to modify a pseudo process by a single interaction: take the case where a stationary car (a causal process) throws its shadow on a fence. Suddenly the fence falls over, producing a permanent modification in the shadow. Then the shadow has been marked by the single local action of the falling fence. Salmon's counterfactual requirement that the process would remain uniform (presumably in the absence of the marking interaction, all other things being equal) does not help in these cases: the shadow would have remained uniform had the fence not fallen. Indeed most of the above cases fulfil this counterfactual requirement. Further, these cases are not ruled out by attempts to restrict the kinds of admissible properties by admitting only those which can be detected by physically possible detectors, since the relevant property here is shape, which certainly is detectable by physical detectors. So there does not seem to be any obvious way of answering this difficulty. Thus there are two separate classes of pseudo processes (in Kitcher's terminology, derivative marks and pseudomarks) which qualify as causal according to the MT principle.

5. The Conserved Quantity Theory

The idea of appealing to conserved quantities has its forerunners in Aronson's and Fair's appeal to energy and momentum. (Aronson, 1971; Fair, 1979) But the first explicit formulation was given in a brief suggestion made by Skyrms in 1980, in his book *Causal Necessity* (1980, p. 111) and the first detailed conserved quantity theory by Dowe (1992a; 1992b). See also Salmon, 1994 and Dowe, 1995. The conserved quantity theory can be expressed in two propositions:

> CQ1. A *causal process* is a world line of an object which possesses a conserved quantity.
>
> CQ2. A *causal interaction* is an intersection of world lines which involves exchange of a conserved quantity.

A *process* is the world line of an object, regardless of whether or not it possesses any conserved quantities. A process can be either causal or non-causal (pseudo). A *world line* is the collection of points on a space-time (Minkowski) diagram which represents the history of an object. This means that processes are determinate regions, or ‘worms’, in space time. Such processes, or worms in space time, will normally be time-like; that is, every point on its world line lies in the future lightcone of the process' starting point.

An *object* is anything found in the ontology of science (such as particles, waves or fields), or common

sense (such as chairs, buildings, or people). This will include non-causal objects such as spots and shadows. It is important to appreciate the difference between an object and a process. Loosely speaking, a process is the development over time of an object. Processes are usually extended in time.

Worms in space time which are not processes Kitcher calls 'spatiotemporal junk' (1989). Thus a representation on a space time diagram represents either a process or a piece of spatiotemporal junk, and a process is either a causal or a pseudo process. In a sense what counts as an object is unimportant; any old gerrymandered thing qualifies (except time-wise gerrymanders) (Dowe, 1995). In the case of a causal process what matters is whether the object possesses the right type of quantity. A shadow is an object but it does not possess the right type of conserved quantities; for example, a shadow cannot possess energy or momentum. It has other properties, such as shape, velocity, and position but possesses no conserved quantities. (The theory could be formulated in terms of *objects*: there are causal objects and pseudo objects. Causal objects are those which possess conserved quantities, pseudo objects are those which do not. Then a causal process is the world line of a causal object.)

A *conserved quantity* is any quantity which is universally conserved, and current scientific theory is our best guide as to what these are. For example, we have good reason to believe that mass-energy, linear momentum, and charge are conserved quantities.

An *intersection* is simply the overlapping in space time of two or more processes. The intersection occurs at the location consisting of all the space time points which are common to both (or all) processes. An *exchange* occurs when at least one incoming, and at least one outgoing process undergoes a change in the value of the conserved quantity, where 'outgoing' and 'incoming' are delineated on the space-time diagram by the forward and backward light cones, but are essentially interchangeable. The exchange is governed by the conservation law, which guarantees that it is a genuine causal interaction. It follows that an interaction can be of the form X, Y, l, or of a more complicated form.

'Possesses' is to be understood in the sense of 'instantiates'. We suppose an object possesses energy if science attributes that quantity to that body. It does not matter whether that process transmits the quantity or not, nor whether the object keeps a constant amount of the quantity. It must simply be that the quantity may be truly predicated of the object.

6. Objections to the Conserved Quantity Theory

Salmon (1994, p. 308) has argued that the conserved quantity theory requires "transmits" rather than just "possesses": Consider a rotating spotlight spot moving around the wall of a large building. This is a classic case of a pseudo-process: in theory such a spot could move faster than the speed of light. But the spot manifests energy at each point along its trajectory. Therefore, Salmon's argument goes, we need more than just the regular appearance of energy to characterise causal processes; we need the notion of transmission. In this section I show how the CQ theory avoids this problem without appealing to the notion of transmission.

Dowe (1992a, p. 127) had argued as follows: a spot or moving patch of illumination does not possess conserved quantities. A moving spot has other properties: speed, size, shape etc; but not *conserved* quantities such as energy or momentum. What possesses the energy which is "regularly appearing" is not the spot but a series of different patches of the wall. The spot and the patch of wall are *not* the same object. The patch of wall does not move. It *does* possess conserved quantities, *its* world line does constitute a causal process, and *it* is not capable of moving faster than the speed of light. The spot does move, but does *not* possess energy and *is* capable of moving faster than the speed of light. Therefore "whether or not an object possesses a conserved quantity" is an adequate criterion for distinguishing causal from pseudo processes.

Salmon (1994, p. 308) provides an ingenious counterexample to this account. He asks us to consider "the worldline of the part of the wall surface that is absorbing energy as a result of being illuminated". (1994, p. 308). This "gerrymandered" object is the aggregate of all the patches of wall that are sequentially illuminated, taken only for the time that they are currently being illuminated. Salmon argues that this object does possess energy over the relevant interval, but does not *transmit* energy. The implication is that the world line of this object is not a causal process, yet the object possesses energy; therefore we need to invoke the notion of transmission- possession is not enough. Salmon himself proposes a new theory which combines elements of his earlier account with elements of the conserved quantity theory).

In response to Salmon's objection Dowe introduces a further condition: that an object wholly exists at a time (1995; forthcoming). This allows him to say that objects have a primitive identity over time. But, while this does rule out what Dowe calls 'timewise gerrymanders' (1995), it is unlikely to satisfy critics such as Salmon because it seems to introduce into the theory an element which is irreducibly hidden to empirical investigation.

Bibliography

- Armstrong, D. M. (1978). *Nominalism and Realism*. Cambridge: University Press.
- Armstrong, D. M. (1980). Identity Through Time. In P. van Inwagen (Ed.), *Time and Cause* (pp. 67-78). Dordrecht: Reidel.
- Aronson, J. (1971). On the Grammar of 'Cause'. *Synthese* 22: 414-430.
- Dowe, P. (1992a). An Empiricist Defence of the Causal Account of Explanation. *International Studies in the Philosophy of Science* 6: 123-128.
- Dowe, P. (1992b). Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory. *Philosophy of Science* 59: 195-216.
- Dowe, P. (1995). Causality and Conserved Quantities: A Reply to Salmon. *Philosophy of Science* 62: 321-333.
- Dowe, P. (forthcoming). Good Connections: Causation and Causal Processes. In H. Sankey (Ed.), *Causation and Laws of Nature* Dordrecht: Kluwer.
- Fair, D. (1979). Causation and the Flow of Energy. *Erkenntnis* 14: 219-250.
- Hanna, J. (1986). Book Review: Scientific Explanation and the Causal Structure of the World. *Review of Metaphysics* 39: 582.

- Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher & W. Salmon (Eds.), *Minnesota Studies in the Philosophy of Science Volume XIII* (pp. 410-505). Minneapolis: University of Minnesota Press.
- Reichenbach, H. (1956). *The Direction of Time.* Berkeley: University of California Press.
- Reichenbach, H. (1958). *The Philosophy of Space and Time.* New York: Dover.
- Rogers, B. (1981). Probabilistic Causality, Explanation, and Detection. *Synthese* 48: 201-223.
- Russell, B. (1913). On the Notion of Cause. *Proceedings of the Aristotelian Society* 13: 1-26.
- Russell, B. (1948). *Human Knowledge.* New York: Simon and Schuster.
- Salmon, W. (1978). Why ask, "Why?"? *Proceedings of the American Philosophical Association* 51: 683-705.
- Salmon, W. (1982). Further Reflections. In R. McLaughlin (Ed.), *What? Where? When? Why?* (pp. 231-280). Dordrecht: Reidel.
- Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.
- Salmon, W. (1994). Causality Without Counterfactuals. *Philosophy of Science* 61: 297-312.
- Skyrms, B. (1980). Causal Necessity. New Haven: Yale University Press.
- Sober, E. (1987). Explanation and Causation. *British Journal for the Philosophy of Science* 38: 243-257.

Other Internet Resources

- Bertrand Russell Archives at McMasters University, Canada.

Related Entries

causation: counterfactual theories of | cause and effect | physics: Reichenbach's common cause principle | Russell, Bertrand

Phil Dowe
phil.dowe@phil.utas.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 8, 1996

Content last modified: December 9, 1996

</html

Counterfactual Theories of Causation

The basic idea of counterfactual theories of causation is that the meaning of a singular causal claim of the form "Event *c* caused event *e*" can be explained in terms of counterfactual conditionals of the form "If *c* had not occurred, *e* would not have occurred". Analyses along these lines have become popular in the last quarter of the twentieth century, especially since the development in the 1970's of possible world semantics for counterfactuals. The best known counterfactual analysis of causation is David Lewis's (1973b) theory. However, intense discussion over twenty years has cast doubt on the adequacy of any simple analysis of singular causation in terms of counterfactuals. Recent years have seen a proliferation of different refinements of the basic idea to achieve a closer match with commonsense judgements about causation.

---

1. Early Counterfactual Theories

The first explicit definition of causation in terms of counterfactuals was, surprisingly enough, given by Hume, when he wrote: "We may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second*. Or, in other words, *where, if the first object had not been, the second never had existed*." (1748, Section VII). It is difficult to understand how Hume could have confused the first, regularity definition with the second, very different counterfactual definition.

At any rate, Hume never explored the alternative counterfactual approach to causation. In this, as in much else, he was followed by generations of empiricist philosophers. The chief obstacle in empiricists' minds to explaining causation in terms of counterfactuals was the obscurity of counterfactuals

themselves, owing chiefly to their reference to unactualised possibilities. Starting with J. S. Mill (1843), empiricists tried to analyse counterfactuals 'metalinguistically' in terms of implication relations between statements. The rough idea is that a counterfactual of the form "If it had been the case that A, it would have been the case that *C*" is true if and only if there is an auxiliary set *S* of true statements consistent with the antecedent *A*, such that the members of *S*, when conjoined with *A*, entail the consequent *C*. Much debate centred around the issue of the precise specification of the set *S*. (See N. Goodman (1983).) Most empiricists agreed that *S* would have to include statements of laws of nature, while some thought that it would have to include statements of singular causation. While the truth conditions of counterfactuals remained obscure in these ways, few empiricists thought it worthwhile to try to explain causation via counterfactuals.

Indeed, the first real attempts to present rigorous counterfactual analyses of causation came only in the late 1960's. (See A. Lyon's (1967).) Typical of these attempts was J. L. Mackie's counterfactual analysis in Chapter 2 of his seminal book *The Cement of the Universe* (1974). As well as offering a sophisticated regularity theory of causation 'in the objects', Mackie presented a counterfactual account of the concept of a cause as "what makes the difference in relation to some background or causal field" (1980, p.xi). Mackie's account of the concept of causation is rich in insights, especially concerning its relativity to a field of background conditions. However, his account never gained as much attention as his regularity theory of causation 'in the objects', no doubt because his view of counterfactuals (in his (1973)), as condensed arguments that do not have truth values, compounded empiricists' scepticism about counterfactuals.

The true potential of the counterfactual approach to causation did not become clear until counterfactuals became better understood through the development of possible world semantics in the early 1970's.

2. Lewis's 1973 Counterfactual Analysis

The best known and most thoroughly elaborated counterfactual theory of causation is David Lewis's theory in his (1973b), which was refined and extended in articles subsequently collected in his (1986a). In response to doubts about the theory's treatment of preemption, Lewis has recently proposed a fairly radical revision of the theory . (See his Whitehead Lectures published in his (2001a), and in a shortened form in his (2000).) In this section we shall confine our attention to the original 1973 theory, deferring for later consideration the recent changes he has proposed.

2.1 Counterfactuals and Causal Dependence

Like most contemporary counterfactual theories, Lewis's theory employs a possible world semantics for counterfactuals. Such a semantics states truth conditions for counterfactuals in terms of relations among possible worlds. Lewis famously espouses a realism about possible worlds, according to which non-actual possible worlds are real concrete entities on a par with the actual world. (See Lewis's defence of modal realism in his (1986e).) However, most contemporary philosophers would seek to deploy the explanatorily fruitful possible worlds framework while distancing themselves from full-blown realism about possible worlds themselves. For example, many would propose to understand possible worlds as maximally consistent sets of propositions; or even to treat them instrumentally as useful theoretical entities having no independent reality.

The central notion of a possible world semantics for counterfactuals is a relation of *comparative similarity* between worlds (Lewis (1973a)). One world is said to be *closer to actuality* than another if the first resembles the actual world more than the second does. Shortly we consider the respects of similarity that Lewis says are important for the counterfactuals linked to causation. For now we simply note two formal constraints he imposes on this similarity relation. First, the relation of similarity produces a weak ordering of worlds so that any two worlds can be ordered with respect to their closeness to the actual world, with allowance being made for ties in closeness. Secondly, the actual world is closest to actuality, resembling itself more than any other world resembles it.

In terms of this similarity relation, the truth condition for the counterfactual "If *A* were (or had been) the case, *C* would be (or have been) the case", symbolised as $A \Box\!\!\rightarrow C$, is stated as follows:

> (1) $A \Box\!\!\rightarrow C$ is *true* in the actual world if and only if (i) there are no possible *A*-worlds; or (ii) some *A*-world where *C* holds is closer to to the actual world than is any *A*-world where *C* does not hold.

We shall ignore the first case in which the counterfactual is vacuously true. The fundamental idea of this analysis is that the counterfactual $A \Box\!\!\rightarrow C$ is true just in case it takes less of a departure from actuality to make the antecedent true along with the consequent than to make the antecedent true without the consequent.

In terms of counterfactuals, Lewis defines a notion of causal dependence between events, which plays a central role in his theory (1973b).

> (2) Where *c* and *e* are two distinct possible events, *e* causally depends on *c* if and only if *c* occurs $\Box\!\!\rightarrow$ *e* occurs and *c* does not occur $\Box\!\!\rightarrow$ *e* does not occur.

This condition states that whether *e* occurs or not depends on whether *c* occurs or not. Where *c* and *e* are actual occurrent events, this truth condition can be simplified somewhat. For in this case it follows from the second formal condition on the comparative similarity relation that the counterfactual "*c* occurs $\Box\!\!\rightarrow$ *e* occurs" is automatically true: this formal condition implies that a counterfactual with true antecedent and

true consequent is itself true. Consequently, the truth condition for causal dependence becomes:

> (3) Where *c* and *e* are two distinct actual events, *e* causally depends on *c* if and only if *c* does not occur □→ *e* does not occur.

The right hand side of this condition is, of course, Hume's second definition of causation. Lewis's official definition of causation differs from it, for he defines causation not in terms of causal dependence directly, but in terms of chains of causal dependence.

There are two immediate things to note about the definition of causal dependence. First, it takes the primary relata of causal dependence to be *events*. Lewis's own theory of events (1986b) construes events as classes of possible spatiotemporal regions. However, very different conceptions of events are compatible with the basic definition. Indeed, it even seems possible to formulate it in terms of facts rather than events. (For instance, see Mellor (1996).) Secondly, the definition requires the causally dependent events to be *distinct* from each other. This qualification is important if spurious non-causal dependences are to be ruled out. (For this point see J. Kim (1973).) For it may be that your saying "Hello" loudly depends on your saying "Hello"; and your writing "Larry" depends on your writing "Lar". But neither dependence counts as a causal dependence since the paired events are not distinct from each other.

2.2 The Temporal Asymmetry of Causal Dependence

What constitutes the direction of the causal relation? Why is this direction typically aligned with the temporal direction from past to future? In answer to these questions, Lewis argues (1979) that the direction of causation is the direction of causal dependence; and it is typically true that events causally depend on earlier events but not on later events. He emphasises the contingency of the latter fact because he regards backwards or time-reversed causation as a conceptual possibility that cannot be ruled out *a priori*. Accordingly, he dismisses any analysis of counterfactuals that would deliver the temporal asymmetry by conceptual fiat.

Lewis's explanation of the temporal asymmetry of counterfactual dependence is based on a *de facto* asymmetry about the actual world. He defines a *determinant* for an event as any set of conditions jointly sufficient, given the laws of nature, for the event's occurrence. (Determinants of an event may be causes or traces of the event.) He observes it is contingently true that events typically have very few earlier determinants but very many later determinants. For example, a spherical wavefront expanding outwards from a point source is a process where each sample of the wave postdetermines what happens at the point at which the wave is emitted. The opposite process in which a spherical wave contracts inward with each sample of wave predetermining what happens at the point the wave is absorbed would obey the laws of nature, but seldom happens in actual fact.

Lewis combines this *de facto* asymmetry of overdetermination with his analysis of the comparative similarity relation (1979). According to this analysis, the most similar worlds are those in which the

actual laws of nature are never violated. But exact similarity with respect to particular matters of facts in some spatiotemporal region is also an important aspect of similarity if it can be achieved at the cost of a small, local miracle, but not at the cost of a big, diverse miracle. There is no built-in time bias in this account. That comes only when it is combined with the asymmetry of overdetermination.

To see how the two parts combine consider the famous example of Nixon and the Nuclear Holocaust. An early objection to Lewis's account of counterfactuals (K. Fine (1975)) was that, counterintuitively, it makes this counterfactual false:

> (4) If Nixon had pressed the button, there would have been a nuclear war.

The argument is that a world in which Nixon pressed the button, but some minute violation of the laws then prevented a nuclear war, is much more like the actual world than one in which Nixon pressed the button and a nuclear war took place. Lewis replies (1979) that this does not accord with his account of the similarity relation. On this account, a button-pressing world that diverges from the actual world by virtue of a miracle is more like the actual world than a button-pressing world that converges with the actual world by virtue of a miracle. For in view of the asymmetry of overdetermination, the divergence miracle that allows Nixon to press the button need only be a small, local miracle, but the convergence miracle required to wipe out the traces of Nixon's pressing the button must be a very big, diverse miracle. Of course, if the asymmetry of overdetermination went in the opposite temporal direction, the very same standards of similarity would dictate the opposite verdict.

In general, then, the symmetric analysis of similarity, combined with the *de facto* asymmetry of overdetermination, implies that it is easier to reconcile a hypothetical change in the actual course of events by preserving the past and allowing for a divergence miracle rather than shielding the future from change by having a convergence miracle. This fact in turn implies that, where the asymmetry of overdetermination obtains, the present counterfactually depends on the past, but not on the future.

2.3 Transitivity and Preemption

Lewis says that though causal dependence between actual events is sufficient for causation, it is not necessary (1973b). Counterfactual dependence is not transitive, so it can happen that three actual events *c*, *d* and *e* are such that *d* would not have occurred without *c*, and *e* would not have occurrred without *d*, but *e* would still have occurred without *c*. We consider an example shortly. Nonetheless, Lewis insists that causation is transitive so that *c* must be a cause of *e* if *c* is a cause of *d* and *d* is a cause of *e*.

To overcome this problem he extends causal dependence to a transitive relation in the usual way by taking its ancestral. He defines a *causal chain* as a finite sequence of actual events *c, d, e*,... where *d* depends causally on *c*, *e* on *d*, and so on throughout the sequence. Then causation is finally be defined in these terms:

(5) *c* is a cause of *e* if and only if there exists a causal chain leading from *c* to *e*.

This definition has the virtue of killing two birds with one stone. Not only does it ensure the transitivity of causation, but it also appears to solve an additional problem to do with preemption that is illustrated by the following example. Suppose that two crack marksmen conspire to assassinate a hated dictator, agreeing that one or other will shoot their victim on a public occasion. Acting side-by-side, assassins *A* and *B* find a good vantage point, and, when the dictator appears, both take aim. *A* pulls his trigger and fires a shot that hits its mark, but *B* desists from firing when he sees *A* pull his trigger. Here assassin *A*'s actions are the actual cause of the dictator's death, while *B*'s actions are a preempted potential cause. (Lewis distinguishes these cases from cases of symmetrical overdetermination in which two processes terminate in the effect, with neither process preempting the other. Lewis believes that these cases are not suitable test cases for a theory of causation since they do not elicit clear judgements.) The problem seems to be that both actions are on a par from the point of view of causal dependence: if neither *A* nor *B* acted, then the dictator would not have died; and if either had acted without the other, the dictator would have been killed.

However, given the definition of causation in terms of causal chains, Lewis is able to distinguish the preempting actual cause from the preempted potential cause. There is a causal chain running from *A*'s actions to the dictator's death. For consider the intermediary event occurring between *A*'s taking aim and the dictator's death: the bullet in mid-trajectory. The bullet's trajectory depends causally on *A*'s action and the dictator's death depends causally on the bullet's trajectory. Hence, we have a causal chain, and so causation. But no corresponding intermediary can be found between *B*'s actions and the dictator's death; and so for this reason *B*'s actions do not count as an actual cause of the death.

2.4 Chancy Causation

So far we have considered how the counterfactual theory of causation works under the assumption of determinism. But what about causation when determinism fails? Lewis (1986c) argues that chancy causation is a conceptual possibility that must be accommodated by a theory of causation. Indeed, contemporary physics tells us the actual world abounds with probabilistic processes that are causal in character. To take a familiar example (1986c): suppose that you mischievously hook up a bomb to a radioactive source and geiger counter in such a way that the bomb explodes when the counter registers a certain number of clicks. If it happens that the counter registers the required number of clicks and the bomb explodes, your act caused the explosion, even though there is no deterministic connection between them.

In order to accommodate chancy causation, Lewis (1986c) defines a more general notion of causal dependence in terms of chancy counterfactuals. These counterfactuals are of the form *A* o-> *Pr* (*C*) = *x*, where the counterfactual is an ordinary world-counterfactual, interpreted according to the semantics above, and the *Pr* operator is a probability operator with narrow scope confined to the consequent of the counterfactual. Lewis interprets the probabilities involved as temporally indexed single-case chances. (See his (1980) for the theory of single-case chance).

The more general notion of causal dependence reads:

> (6) Where *c* and *e* are distinct events, *e* causally depends on *c* if and only if, if *c* had not occurred, the chance of *e*'s ocurring would have been much less than it actually was (given that *c* occurred).

This definition covers cases of deterministic causation in which the chance of the effect with the cause is 1 and the chance of the effect without the cause is 0. But it also allows for cases of irreducible probabilistic causation where these chances can take non-extreme values. It is similar to the central notion of probabilistic relevance used in probabilistic theories of causation, except that it employs chancy counterfactuals rather than conditional probabilities. (See the discussion in Lewis (1986c) for the advantages of the counterfactual approach over the probabilistic one.)

The rest of the theory of chancy causation follows the outlines of the theory of deterministic causation. Causal dependence is extended to a transitive notion by taking its ancestral. As before, we have causation when we have one or more steps of causal dependence.

3. Problems for Lewis's Counterfactual Theory

In this section we consider the principal difficulties for Lewis's theory that have emerged in discussion over the last twenty years.

3.1 Context-sensitivity

One relatively overlooked aspect of the concept of causation is its sensitivity to contextual factors. In so far as Lewis's theory overlooks this context-sensitivity, it represents a problem for the theory.

The theory assumes that causation is an absolute relation whose nature does not vary from one context to another. (This follows from the way the counterfactuals that define the central notion of causal dependence are governed by a unique, context-invariant system of weighted respects of similarity.) According to the theory, any event but for which an effect would not have occurred is one of the effect's causes. But this generates some absurd results. For example, suppose a person develops lung cancer as a result of years of smoking. It is true that if he had not smoked he would not have got lung cancer. But it is also true that if he had not possessed lungs or had not been born, he would not have got lung cancer. Commonsense draws a distinction between causes and background conditions, ranking the person's

smoking among the former and the person's birth and possession of lungs among the latter.

In their seminal work *Causation in the Law* (1965; 2nd ed 1985), H. L. A. Hart and A. Honor argue that the distinction between causes and conditions is relative to context in at least two different ways. One form of relativity might be called relativity to the context of occurrence. If a building is destroyed by fire, the presence of oxygen would be cited as a mere condition of the building's destruction. On the other hand, if a fire breaks out in a laboratory where oxygen is deliberately excluded, it may be appropriate to cite the presence of oxygen as a cause of the fire. The second form of relativity might be called relativity to the context of enquiry. For example, the cause of a great famine in India may be identified by the Indian peasant as the drought, but the World Food Authority may identify the Indian government's failure to build up reserves as the cause, and the drought as a mere condition.

For the most part, Lewis ignores these subtle context-sensitive distinctions. In his view (1986d), every event has an objective causal history consisting of a vast structure of events ordered by causal dependence. The human mind may select parts of the causal history for attention, perhaps different parts for different purposes of enquiry. However, Lewis does not specify the 'principles of invidious selection' by which some parts of the causal history are selected for attention, except to mention the relevance of Grice's maxims of conversation. But Grice's maxims of conversation are not well suited to explaining the context-sensitive distinctions involved in our causal judgements. As many philosophers have pointed out (A. Garfinkel (1981); C. Hitchcock (1996); P. Lipton (1990); J. Woodward (1984); and B. Van Fraassen (1981)), the contextual principles behind our causal judgements seem to rely on considerations concerning which class of situations the effect is contrasted with. As general principles of rational information exchange, Grice's maxims miss out on these causation-specific principles. (For discussion of the relevance of contrastive explanation to the causes/conditions distinction see Menzies (2001).)

3.2 Temporal Asymmetry

There have been several important critical discussions of Lewis's explanation of the temporal asymmetry of causation. (See P. Horwich (1987, Chap. 10); H. Price (1992) and (1996, Chap. 6); and D. Hausman (1998, Chap. 6).)

One kind of criticism has focused on the psychological implausibility of Lewis's explanation. (See Horwich (1987).) Recall that the explanation appeals, on the one hand, to a system of weighted respects of similarity between possible worlds that is delivered by *a priori* conceptual analysis and, on the other hand, to an asymmetry of overdetermination that is established *a posteriori* as a contingent truth about the actual world. The two-part explanation is supposed to employ facts that are sufficiently well known to play a role in the explanation of our linguistic use of counterfactuals. However, it is psychologically implausible that the intricate system of weighted respects of similarity involving comparison of miracles of different sizes could capture the intuitive similarity relation used in counterfactual reasoning. Why should we have developed such a baroque notion of similarity? Moreover, the asymmetry of overdetermination is an esoteric scientific finding that is not common knowledge to everyone using counterfactuals. Long before the discovery of the asymmetry of overdetermination, our ancestors used

counterfactuals and assumed their temporal asymmetry. So it is very unlikely that this scientific finding could account for their mastery of this aspect of the use of counterfactuals. (For Lewis's reply to this criticism see Postscript E to "Counterfactual Dependence and Time's Arrow" in his (1986a, p. 66).)

Another criticism is that the *de facto* asymmetry of overdetermination is not sufficiently extensive to constitute the objective basis for the temporal asymmetry of causal dependence. Lewis concedes this asymmetry is a contingent feature of the actual world which may not obtain in other worlds. For example, in a simple world inhabited by a single atom, the asymmetry of overdetermination fails to obtain. Convergence to this world takes no more of a varied and widespread miracle than divergence from it, so that counterfactuals about what happens in this world are not temporally asymmetric. However, Lewis's use of the asymmetry of overdetermination has been criticised on the grounds that it fails not only in simple worlds of this kind but also in the actual world. (See Price (1992) and (1996).) This asymmetry, like the related fork asymmetry (correlated events typically have an earlier common cause but seldom a later common effect), is a product of thermodynamic asymmetries. The example of radiation that Lewis offers as a paradigm of the asymmetry of overdetermination depends on the fact that the sources of radiation stem from big disturbances in the initial conditions, but not in the final conditions of the system in question. If the system were a closed system in thermodynamic equilibrium, there would be no asymmetry. This means that the asymmetry of overdetermination, as a feature of thermodynamic disequilibrium, is a large-scale, statistical asymmetry, appearing at the macroscopic but not mircoscopic level. However, the commonplace judgement of physicists is that microphysical processes are causal and temporally asymmetric in character. In view of the failure of the asymmetry of overdetermination at the microscopic level, Lewis's theory is powerless to explain this fact.

3.3 Transitivity

As we have seen, Lewis builds transitivity into causation by defining it in terms of chains of causal dependence. However, a number of counter-examples have been presented which cast doubt on transitivity. (Lewis (2001a) presents a short catalogue of these counterexamples.).

Here is an example due to Michael McDermott (1995). *A* and *B* each have a switch in front of them, which they can move to the left or right. If both switches are thrown into the same position, a third person *C* receives a shock. *A* does not want to shock *C*. Seeing *B*'s switch in the left position, *A* moves her switch to the right. *B* does want to shock *C*. Seeing *A*'s switch thrown to the right, she now moves her switch to the right as well. *C* receives a shock. Clearly, *A*'s throwing her switch to the right causes *B* to throw her switch to the right, which in turn causes *C* to receive the shock. But *A* attempted to prevent the shock so that it seems unreasonable to say that *A*'s move causes *C* to be shocked.

Here is one more example, due to Ned Hall (2001). A person is walking along a mountain trail, when a boulder high above is dislodged and comes careering down the mountain slopes. The walker notices the boulder and ducks at the appropriate time. The careering boulder causes the walker to duck and this, in turn, causes his continued stride. (This second causal link involves *double prevention*: the duck prevents the collision between walker and boulder which, had it occurred, would have prevented the walker's

continued stride.) However, the careeering boulder is the sort of thing that would prevent the walker's continued stride and so it seems counterintuitive to say that it causes the stride.

Lewis has noted (2000) that all counterexamples to transitivity have a common structure. An event *c* occurs that initiates a process that threatens to prevent some later event *e*. However, *c* also causes another event *d* which cuts short the threatening process but also causes *e* to occur anyway. (Hall (2000), while noting this common structure, also observes that examples involving double prevention have distinctive features which make them easier to deflect as counterexamples to transitivity.)

The strategy Lewis adopts in his (2000) is to defend transitivity by diagnosing the sources of our inclination to accept them. For example, he points out that the examples typically involve a structure in which a *c*-type event generally prevents an *e*-type but in the particular case the *c*-event actually causes another event that causes the *e*-event. If we mix up questions of what is generally conducive to what, with questions about what caused what in this particular case, he says, we may think that it is reasonable to deny that *c* causes *e*. But if we keep the focus sharply on the particular case, we must insist that *c* does in fact cause *e*.

3.4 Preemption

As we have seen, Lewis employs his strategy of defining causation in terms of chains of causal dependence not only to make causation transitive, but also to deal with preemption examples. However, there are preemption examples that this strategy cannot deal with satisfactorily. Difficulties concerning preemption have proven to be the biggest bugbear for Lewis's theory.

In his (1986c), Lewis distinguishes cases of *early* and *late preemption*. In early preemption examples, the process running from the preempted alternative is cut short before the main process running from the preempting cause has gone to completion. The example of the two assassins, given above, is an example of this sort. The theory of causation in terms of chains of causal dependence can handle this sort of example. In contrast, cases of late preemption are ones in which the process running from the preempted cause is cut short only after the main process has gone to completion and brought about the effect. The following is an example of late preemption due to Hall (2001).

Billy and Suzy throw rocks at a bottle. Suzy throws first so that her rock arrives first and shatters the glass. Without Suzy's throw, Billy's throw would have shattered the bottle. However, Suzy's throw is the actual cause of the shattered bottle, while Billy's throw is merely a preempted potential cause. This is a case of late preemption because the alternative process (Billy's throw) is cut short after the main process (Suzy's throw) has actually brought about the effect.

Lewis's theory cannot explain the judgement that Suzy's throw was the actual cause of the shattering of the bottle. For there is no causal dependence between Suzy's throw and the shattering, since even if Suzy had not thrown her rock, the bottle would have shattered due to Billy's throw. Nor is there a chain of stepwise dependences running cause to effect, because there is no event intermediate between Suzy's

throw and the shattering that links up them into a chain of dependences. Take, for instance, Suzy's rock in mid-trajectory. Certainly, this event depends on Suzy's initial throw, but the problem is that the shattering of the bottle does not depend on it, because even without it the bottle would still have shattered because of Billy's throw.

To be sure, the bottle shattering that would have occurred without Suzy's throw would be very different from the bottle shattering that actually occurred with Suzy's throw. For a start, it would have occurred later. Following Lewis, let us call an event *fragile* to the extent that it could not have occurred at a different time, or in a different manner. One response to the problem of late preemption is to insist that the events involved should be construed as fragile events. Accordingly, it will be true rather than false that if Suzy had not thrown her rock, then the actual bottle shattering, taken as a fragile event with an essential time and manner of occurrence, would not have occurred. Lewis himself does not endorse this response on the grounds that a uniform policy of construing events as fragile would go against our usual practices, and would generate many spurious causal dependences. For example, suppose that a poison kills its victim more slowly and painfully when taken on a full stomach. Then, the victim's eating dinner before he drinks the poison would count as a cause of his death since the time and manner of the death depend on the eating of the dinner. (See his (1986c) for discussion.)

When we turn from preemption examples involving deterministic causation to those involving chancy causation, we see that the problems for Lewis's theory multiply. One particularly recalcitrant problem is described in Menzies (1989). (See also J. Woodward (1990).) Suppose that two systems can produce the same effect, perhaps at the same time and in the same manner. (It does not matter whether this is an example of early or late preemption.) However, one system is much more reliable than the other. The reliable system starts and, left to itself, will very probably produce the effect. But you do not leave it to itself. You throw a switch that shuts down the reliable system and turns on the unreliable one. As luck would have it, the unreliable system works and brings about the effect. This kind of example presents a problem for the probabilistic generalisation of the counterfactual theory because the preempting actual cause decreases the chance of the effect while the preempted potential cause increases its chance. In addition to the problem of explaining how the preempting cause qualifies as a cause when the effect does not causally depend on it, the probabilistic counterfactual theory faces the problem of explaining how the preempted cause is not really a cause when the effect does causally depend on it.

4. Recent Developments

In this section we shall consider some recent developments of the counterfactual approach to causation, which have been motivated by the desire to overcome the deficiencies in Lewis's 1973 theory, especially with respect to preemption.

4.1 Lewis's 1999 Theory

In an attempt to deal with the various problems facing his 1973 theory, Lewis has recently presented a revised version of the theory. The new version was first presented in his Whitehead Lectures at Harvard University in March 1999. (For the lectures see his (2001); a shortened version has appeared as his (2000).)

Counterfactuals play as central a role in the new theory as in the old. But the counterfactuals it employs do not simply state dependences of *whether* one event occurs on *whether* another event occurs. The counterfactuals state dependences of *whether*, *when*, and *how* one event occurs on *whether*, *when*, and *how* another event occurs. A key idea in the formulation of these counterfactuals is that of an *alteration* of an event. This is an actualised or unactualised event that occurs at a slightly different time or in a slightly different manner from the given event. An alteration is, by definition, a very fragile event that could not occur at a different time, or in a different manner without being a different event. Lewis stipulates that one alteration of an event is the very fragile version that actually occurs.

The central notion of the new theory is that of influence.

> (7) Where *c* and *e* are distinct events, *c influences e* if and only if there is a substantial range of *c*1, *c*2, ... of different not-too-distant alterations of *c* (including the actual alteration of *c*) and there is a range of *e*1, *e*2, ... of alterations of *e*, at least some of which differ, such that if *c*1 had occurred, *e*1 would have occurred, and if *c*2 had occurred, e2 would have occurred, and so on.

Where one event influences another, there is a pattern of counterfactual dependence of whether, when, and how upon whether, when, and how. As before, causation is defined as an ancestral relation.

> (8) *c causes e* if and only if there is a chain of stepwise influence from *c* to *e*.

One of the points Lewis advances in favour of this new theory is that it handles cases of late as well as early pre-emption. (The theory is restricted to deterministic causation and so does not address the example of probabilistic preemption described in section 3.4.) Reconsider, for instance, the example of late preemption involving Billy and Suzy throwing rocks at a bottle. The theory is supposed to explain why Suzy's throw, and not Billy's throw, is the cause of the shattering of the bottle. If we take an alteration in which Suzy's throw is slightly different (the rock is lighter, or she throws sooner), while holding fixed Billy's throw, we find that the shattering is different too. But if we make the same alterations to Billy's throw while holding Suzy's throw fixed, we find that the shattering is unchanged.

Another point in favour of the new theory is that it handles a type of preemption Lewis calls *trumping*. (Trumping was discovered by Jonathan Schaffer: see his (2000).) Lewis gives an example involving a major and a sergeant who are shouting orders at the soldiers. The major and sergeant simultaneously

shout "Advance"; the soldiers hear them both and advance. Since the soldiers obey the superior officer, they advance because the major orders them to, not because the sergeant does. So the major's command preempts or trumps the sergeant's. Where other theories have difficulty with trumping cases, Lewis's argues his new theory handles them with ease. Altering the major's command while holding fixed the sergeant's, the soldier's response would be correspondingly altered. In contrast, altering the sergeant's command, while holding fixed the major's, would make no difference at all.

There is, however, some reason for scepticism about whether the new theory handles the examples of late preemption and trumping completely satisfactorily. In the example of late preemption, Billy's throw has some degree of influence on the shattering of the bottle. For if Billy had thrown his rock earlier (so that it preceded Suzy's throw) and in a different manner, the bottle would have shattered earlier and in a different manner. Likewise, the sergeant's command has some degree of influence on the soldiers' advance in that if the sergeant had shouted earlier than the major with a different command, the soldiers would have obeyed his order. In response to these points, Lewis must say that these alterations of the events are too-distant to be considered relevant. But some metric of distance in alterations is required, since it seems that similar alterations of Suzy's throw and the major's command are relevant to their having causal influence.

It has also been argued that the new theory generates a great number of spurious instances of causation. (For discussion see J. Collins (2000); I. Kvart (2001); and P. Dowe (2001).) The theory implies that any event that influences another event to a certain degree counts as one of its causes. But commonsense is more discriminating about causes. To take an example of Jonathan Bennett (1987): rain in December delays a forest fire; if there had been no December rain, the forest would have caught fire in January rather than when it actually did in February. The rain influences the fire with respect to its timing, location, rapidity, and so forth. But commonsense denies that the rain was a cause of the *fire*, though it allows that it is a cause of the *delay* in the fire. Similarly, in the example of the poison victim discussed above, the victim's ingesting poison on a full stomach influences the time and manner of his death (making it a slow and painful death), but commonsense refuses to countenance his eating dinner as a cause of his death, though it may countenance it as a cause of its being a slow and painful death. *Pace* Lewis, commonsense does not take anything that affects the time and manner of an event to be a cause of the event *simpliciter*.

4.2 Causation as Intrinsic Relation

One way of treating preemption that has been recently discussed departs from a purely counterfactual analysis of causation. It has been argued that preemption examples highlight the intuitive idea that causation is an intrinsic relation between events, which is to say it is a local relation depending on the intrinsic properties of the events and what goes on between them, and nothing else. The proferred treatments of preemption marry this intuitive idea with a crucial deployment of counterfactuals.

At one time Lewis himself resorts to this way of treating late preemption examples when he invoked the notion of *quasi-dependence*. (See his (1986c).) To explain this notion consider a case that resembles the

case of Billy and Suzy throwing rocks at a bottle. Suzy throws a rock and shatters the bottle in exactly the same way in which she does in the original case. But in this case Billy and his rock are entirely absent. Lewis argued that since the process in the original case and the process in the comparison case are intrinsically alike (and also obey the same laws), both or neither must be causal. However, the comparison process is surely a causal process since, thanks to Billy's absence, it exhibits a causal dependence. Accordingly, the process in the original case must be a causal process too, even though it does not exhibit a causal dependence. In such examples Lewis has said that the actual process that does not exhibit causal dependence is, nonetheless, causal by courtesy: it exhibits *quasi-dependence* in virtue of its intrinsic resemblance to the causal process in the comparison case.

A related idea is pursued in Menzies (1996 and 1999). Menzies argues that there is an element in our concept of causation that resists capture in purely counterfactual terms. This element consists in the idea that causation is a structural relation that underlies and supports causal dependences. This idea can be captured by treating the concept of causation as the concept of a theoretical entity. Applying a standard treatment of theoretical concepts, he argues that causation should be defined as the unique occupant of a certain characteristic role given by the platitudes of the folk theory of causation. One platitude is that causation is an intrinsic relation between events. Another platitude is that it is *typically*, but not invariably, accompanied by causal dependence. Accordingly, causation is defined in the following way:

> (9) *c* causes *e* if only if the intrinsic relation that typically accompanies causal dependence holds between *c* and *e*.

On this account, causation is not constituted by causal dependence. It is, in fact, a distinct relation for which causal dependence is, at best, a defeasible marker. The relation may be identified *a posteriori* with some physically specificable relation such as energy-momentum transfer. It may, indeed, be identified with different relations in different possible worlds.

This definition is supposed to explain commonsense intuitions about preemption examples. For example, Suzy's throw, and not Billy's throw, caused the shattering of the bottle, because the intrinsic relation that typically accompanies causal dependence connects Suzy's throw, but not Billy's throw, with the shattering of the bottle.

Lewis has since rejected his approach to preemption via quasi-dependence in favour of his current theory in terms of influence. (See his (2001a; 2001b).) He now claims that theories of causation as an intrinsic relation do not do justice to the full range of our intuitions about causation. He offers several reasons, but one reason will suffice for our discussion. He notes that the intuition that causation is an intrinsic matter does not apply to cases of double prevention. Consider a case of double prevention due to Hall (2001). A fighter is escorting a bomber on a raid. The pilot of the fighter shoots down an interceptor that would otherwise have shot down the bomber. The fighter pilot's action counts as a cause of the successful bombing because the action prevented something that would have prevented the bombing. Lewis observes that the causation in such cases of double prevention is partly an extrinsic matter. If the interceptor had been about to receive a radio order to return to base without attacking the bomber, the

fighter pilot's action would not have been a cause of the bombing. Moreover, he notes that much of the spatiotemporal region between the pilot's shooting down of the interceptor and the successful bombing is simply empty so that there is no chain of events to serve as a connecting process between cause and effect. The intuition that causation is an intrinsic relation does not apply in this case. More generally, he argues that theories of causation as an intrinsic relation are overhasty generalisations of one specific kind of causation, and they fail to do justice to our intuitions about causation involving absences (as causes, effects or intermediaries).

Bibliography

- Bennett, J. (1987): "Event Causation: the Counterfactual Analysis", *Philosophical Perspectives*, 1, pp.367-86.
- Collins, J., Hall, E., and Paul, L. (2001): *Causation and Counterfactuals*. Cambridge, Mass: MIT Press.
- Collins, J. (2000): "Preemptive Preemption", *Journal of Philosophy*, 97, pp.223-34.
- Dowe, P. (2001): "Is Causation Influence?", forthcoming.
- Fine, K. (1975): "Review of Lewis (1973a)", *Mind*, 84, pp.451-8.
- Garfinkel, A. (1981). *Forms of Explanation*. New Haven: Yale University Press.
- Goodman, N. (1983): *Fact, Fiction, and Forecast*. Cambridge, Mass.: Harvard University Press.
- Hall, N. (2000): "Causation and the Price of Transitivity", *Journal of Philosophy*, 97, pp.198-222.
- ------- (2001): "Two Concepts of Causation", in Collins, Hall, and Paul (2001).
- Hart, H. L. and Honor, A. (1965): *Causation in the Law*. Oxford: Clarendon Press. Second edition 1985.
- Hausman, D. (1998): *Causal Asymmetries*. Cambridge: Cambridge University Press.
- Hitchcock, C. (1996): "The Role of Contrast in Causal and Explanatory Claims", *Synthese*, 107, pp.395-419.
- Horwich, P. (1987): *Asymmetries in Time*. Cambridge, Mass. : MIT Press.
- Hume, D. (1748): *An Enquiry concerning Human Understanding*.
- Kim, J. (1973): "Causes and Counterfactuals", *Journal of Philosophy*, 70, pp.570-72.
- Kvart, I. (2001): "Counterexamples to Lewis' ‘Causation as Influence’", *Australasian Journal of Philosophy*, forthcoming.
- Lewis, D. (1973a): *Counterfactuals*. Oxford: Blackwell.
- ------- (1973b): "Causation", *Journal of Philosophy*, 70, pp.556-67. Reprinted in his (1986a)
- ------- (1979): Counterfactual Dependence and Time's Arrow", *Nous*, 13, pp.455-76. Reprinted in his (1986a).
- ------- (1980): "A Subjectivist's Guide to Objective Chance", in R. Jeffrey, ed., *Studies in Inductive Logic and Probability: Volume II*, Reprinted in his (1986a).
- ------- (1986a): *Philosophical Papers: Volume II*. Oxford: Oxford University Press.
- ------- (1986b): "Events", in his (1986a).
- ------- (1986c): "Postscripts to ‘Causation’", in his (1986a).
- ------- (1986d): "Causal Explanation", in his (1986a).
- ------- (1986e): *The Plurality of Worlds*. Oxford: Blackwell.

- ------- (2000): "Causation as Influence", *Journal of Philosophy*, 97, pp.182-97.
- ------- (2001a): "Causation as Influence", in Collins, Hall, and Paul (2001).
- ------- (2001b): "Void and Object", in Collins, Hall, and Paul (2001).
- Lipton, P. (1991): *Inference to the Best Explanation*. London: Routledge.
- Lyons, A. (1967): "Causality", *British Journal for the Philosophy of Science*, 18, pp.1-20.
- Mill, J. S. (1843): *A System of Logic*.
- McDermott, M. (1995): "Redundant Causation", *British Journal for the Philosophy of Science*, 40, pp.523-544.
- Mackie, J. L. (1973): *Truth, Probability, and Paradox*. Oxford: Oxford University Press
- ------- (1974): *The Cement of the Universe*. Oxford: Oxford University Press. Second edition 1980.
- Mellor, D. H. (1995): *The Facts of Causation*. London: Routledge.
- Menzies, P. (1989): "Probabilistic Causation and Causal Processes: A Critique of Lewis", *Philosophy of Science*, 56, pp.642-663.
- ------- (1996): "Probabilistic Causation and the Pre-emption Problem", *Mind*, 105, pp.85-117.
- ------- (1999): "Intrinsic versus Extrinsic Conceptions of Causation", in H. Sankey, ed., *Causation and Laws of Nature*, Kluwer Academic Publishers, pp.313-29.
- ------- (2001): "Difference-Making in Context", in Collins, Hall, and Paul (2001).
- Price, H. (1992): "Agency and Causal Asymmetry", *Mind*, 101, pp.501-20.
- ------- (1996): *Time's Arrow and Archimedes' Point*. Oxford: Oxford University Press.
- Schaffer, J. (2000): "Trumping Preemption", *Journal of Philosophy*, 9, pp.165-81.
- van Fraassen, B. (1981): *The Scientific Image*. Oxford: Clarendon Press.
- Woodward, J. (1984): "A Theory of Singular Causal Explanation", *Erkenntnis*, 21, pp.31-62.
- ------- (1990): "Supervenience and Singular Causal Statements", in Dudley Knowles, ed., *Explanation and its Limits*, Cambridge: Cambridge University Press, pp.215-61.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

causation: causal processes | causation: probabilistic | conditionals: counterfactual | determinism, causal | events | facts | Hume, David | implicature | intrinsic vs. extrinsic properties | possible worlds | probability calculus: interpretations of | rationalism vs. empiricism | scientific explanation | time: thermodynamic asymmetry in

Peter Menzies
peter.menzies@mq.edu.au

First published: January 10, 2001
Content last modified: January 12, 2001

Events

Broadly understood, events are things that happen -- things such as births and deaths, thunder and lightening, explosions, weddings, hiccups and hand-waves, dances, smiles, walks. Whether such things form a genuine metaphysical category is a question that has attracted the sustained interest of philosophers, especially in the second half of the 20th century. But there is little question that human perception, action, language, and thought manifest at least a *prima facie* commitment to entities of this sort:

- Pre-linguistic infants appear to be able to discriminate and "count" events. The content of adult perception, especially in the auditory realm, endorses the discrimination and recognition *as events* of some aspects of the perceived scene.
- Humans (and arguably other animals) form the intention to plan and execute actions, and to bring about changes in the world.
- Dedicated linguistic devices (such as verb tenses and aspects, nominalization of some verbs, certain proper names) are tuned to events and event structures, as opposed to entities and structures of other sorts.
- Thinking about the temporal, causal, and intentional aspects of the world seems to require parsing those aspects in terms of events and their descriptions.

It is not clear to what extent such commitments are to be understood as an integrated phenomenon or as four separate, independent dispositions. However, there exist significant signs of convergence among the various commitments. For instance, the events that are perceived appear to be categorically homogeneous with those that are talked about or thought of in causal explanations [Zacks *et al*. 2001].

Because the *prima facie* commitments of human perception, action, language, and thought are not taken for granted in philosophy, various forms of non-realism about events have been defended. Historically, the main arguments *in favor* of a realist attitude towards events has arisen out of theories concerning the metaphysics of the scientific image of the world [Whitehead 1919] and the semantics of natural language [Davidson 1967a]. But even in these contexts events have been introduced as questionable categories, meant to be somewhat in competition with (if not alternative to) entities of other sorts. This suggests that in spite of the apparent simplicity of the examples in our initial list, the event category is not uncontroversial.

As a matter of fact, it is not even easy to give an uncontroversial *characterization* of events. (Their broad characterization as ‘things that happen’ is commonly found in dictionaries, but it clearly just shifts the

burden to the task of clarifying the meaning of 'happen'.) One useful approach is to set events against entities belonging to other, philosophically more familiar, metaphysical categories. In the following we review the main contrasts between events and those categories that have in the literature been put forward explicitly as their ontological competitors, or at least as categories exhibiting significant differences with the category of events. Along the way, we shall review also the main conceptual tools that metaphysicians and other philosophers have adopted in their attempts to cope with events, either from a realist or from a non-realist perspective.

---

1. Events and Other Categories

1.1 Events vs. Objects

Although not undisputed, some standard differences between events and physical objects are commonplace in the philosophical literature. First, there is a difference in mode of being: material objects such as stones and chairs are said to *exist*; events are said to *occur* or *happen* or *take place* [Hacker 1982a]. Second, there are differences in the way objects and events are said to relate to space and time. Ordinary objects have relatively clear spatial boundaries and unclear temporal boundaries; events have relatively unclear spatial boundaries and clear temporal boundaries. Objects are invidiously located in space -- they *occupy* their spatial location; events tolerate co-location [Quinton 1979, Hacker 1982b]. Objects can move; events cannot [Dretske 1967]. Objects are continuants -- they are *in* time and they persist through time by being wholly present at every time at which they exist; events are occurrents -- they *take up* time and they persist by having different parts (or "stages") at different times [Mellor 1980].

The last distinction is particularly controversial, as there are philosophers who conceive of objects as four-dimensional entities that extend across time just as they extend across space. Some such philosophers would in fact draw no metaphysically significant distinction between objects and events [Quine 1960]. Rather, they would regard the relevant distinction as one of degree: both objects and events would be species of the same "material inhabitant of space-time" genus (as opposed to the genus "immaterial inhabitant", such as the Equator); but whereas events appear to develop quickly in time, objects are relatively "firm and internally coherent" [Quine 1970].

If a metaphysical distinction between objects and events is granted, then a question arises as to the relation between entities in the two categories. Objects are prime actors in events; objectless events are uncommon. But so are eventless objects; events make up the lives of objects. In a radical mood, however, one can think of the entities in one category as being metaphysically dependent on entities in the other. For instance, it has been claimed that events supervene on their participants [Lombard 1986], or that objects depend on the events in which they partake [Parsons 1991]. In a more moderate way, one can grant objects and events an equal ontological status but maintain that either objects or events are primary in the order of thought. For instance, it has been claimed that a pure event-based ontology would not be sufficient for the success of our re-identifying practices, which require some stable frame of reference, adequately provided by objects [Strawson 1959]. A similar asymmetry between objects and events seems to be endorsed by natural language, which has expressions such as ‘the fall of the apple’ but not ‘the pomification of the fall’. However, these asymmetries may be attenuated to the extent that objects, too, may and sometimes must be identified via reference to events [Moravcsik 1965, Davidson 1969, Tiles 1981].

1.2 Events vs. Facts

No matter what their relationships, events are naturally contrasted with objects insofar as both are conceived of as *individuals*. Both appear to be concrete, temporally and spatially located entities organized into part-whole hierarchies. Both can be counted, compared, quantified over, referred to, and variously described and re-described. (It has been argued that our conceptions of these two categories are so closely tied as to be structurally complementary, in that any characterization of the concept *event* that only mentions spatial and temporal features yields a characterization of the concept *object* by a simple replacement of temporal with spatial predicates, and vice versa [Mayo 1961].) From this point of view, events are to be distinguished from facts, which are characterized by features of abstractness and a-temporality: the event of Caesar's death took place in Rome in 44 B.C., but that Caesar died is a fact here as in Rome, today as in 44 B.C. [Ramsey 1927]. One could indeed speculate that for every event there is a companion fact (the fact that the event took place), but the two would still be categorially distinct [Bennett 1988]. Some philosophers, however, have conceived of the link between events and facts as being much closer than this -- close enough to justify assimilating the two categories [Wilson 1974] or at least treating both as species of the same "state of affairs" genus [Chisholm 1970]. This has two main consequences. On the one hand, because facts corresponding to non-equivalent propositions are distinct, events conceived of as facts are fine-grained entities that cannot be freely re-described or re-identified under different conceptualizations: the fact that Caesar died violently is different from the fact that he

died, so the death of Caesar would be a different event from his violent death [Chisholm 1971]. On the other hand, because linguistic expressions of facts are semantically transparent, a Fregean line of argument could be concocted to show instead that events construed as facts are too coarse-grained, to the point of melting into a single "big" entity [Davidson 1967a]. (The argument is known as the "slingshot argument" since [Barwise & Perry 1981].) Either way, the assimilation of events to facts appears to give rise to difficulties.

Some authors have insisted on distinguishing events from facts but have given accounts that effectively amount to such an assimilation. This is the case with those theories that construe events as property exemplifications, i.e., exemplifications *of* properties *by* objects *at* times [Kim 1966, Martin 1969, Goldman 1970, Taylor 1985]. On such theories, events are individual entities. But because they have a structure, a difference in any constituent is sufficient to yield a different event. In particular, a difference in the relevant constitutive property is sufficient to distinguish events such as Caesar's death, construed as Caesar's exemplification of the property of dying, and Caesar's violent death, construed as his exemplification of the property of dying violently [Kim 1976]. This makes events virtually as fine-grained as facts. It bears emphasis, however, that this consequence is not intrinsic to the theory of events as property exemplifications. Both Caesar's death and his violent death could be construed as Caesar's exemplification of one and the same property *P*, describable both as a type of dying and -- with greater accuracy -- as a type of dying violently. Thus, even if construed as a structured complex, an event can be coarsely referred to insofar as its names need not be sensitive to this structure [Bennett 1988]. In this way the distinction between events and facts can be reinstated in terms of a firm distinction between semantic and metaphysical aspects of the theory of event-descriptions.

Similar considerations apply to those theories that treat events as situations, in the sense familiar from situation semantics [Barwise & Perry 1983]. On such theories, events are construed as sets of functions from spatiotemporal locations to "situation types" defined as sequences of objects standing or failing to stand in a certain relation. But while the formal machinery delivers a fine-grained account, the algorithm for applying the machinery to natural language sentences leaves room for flexibility.

1.3 Events vs. Properties

A third major metaphysical category with which events have sometimes been contrasted is that of properties. If events are individuals then they are not properties, for properties are normally construed as universals. Individuals occur whereas universals recur. However, some philosophers have taken very seriously the intuition that in some cases events may be said to recur, as when we say that the sun rises every morning. If so, then events are more similar to properties than to individuals, similar enough to justify treating them as a *kind* of property -- e.g., as properties of moments or intervals of time [Montague 1969], properties of cross-world classes of individuals [Lewis 1986], or properties of sets of world segments [von Kutschera 1993]. For instance, on the first of these accounts, the event of the sun's rising is the property of being an interval during which the sun rises. As a characterization of event types this would be uncontroversial and would allow one to construe particular events as tokens of the corresponding type. (One such construal would correspond to the above-mentioned conception of events as property exemplifications.) But to conceive of events as universal properties is to go beyond this

uncontroversial fact and to reject the existence of event tokens altogether, even when it comes to "particular" events such as the unique rising of the sun that we witnessed this morning. Rather than an instance of the universal *sun rising*, such an event would be a universal in its own right, albeit a universal of such a restricted sort and of such a degree of singularity as to be instantiated only once.

One possible view about properties is that they are not universals but rather particulars of a special sort -- viz. abstract particulars [Stout 1923] or *tropes* [Williams 1953]. According to this view, the redness of this apple is different from the redness of anything else, not because of its extreme singularity (other things could agree with the apple colorwise) but because it is the redness of *this* apple. It exists here and now, where and while the apple exists. Likewise, this morning's rising of the sun would be a different property than any other morning's rising of the sun. If so, then the view that events are properties becomes compatible with the view that they are spatiotemporally located. An event would just be a particularized property located at some region of space-time [Bennett 1996]. (Once again, this conception is closely related to the conception of events as property exemplifications, although the term ‘exemplification’ suggests a construal of properties as universals. Some authors actually identify the two conceptions [Bennett 1988]; others reject the identification on account of the difference between property instances and property exemplifications [Macdonald 1989].)

A variant of the trope conception construes events as trope sequences [Campbell 1981]. However, since tropes are particulars, a sequence of tropes at a place is itself a trope, hence this variant is best regarded as a specification of what sort of tropes events are.

1.4. Events vs. Times

The intuition that events are properties of times can also be fleshed out in terms of thinner metaphysical commitments, by construing events as times *cum description*, i.e., as temporal instants or intervals during which certain statements hold [Van Benthem 1983]. On this view, for example, this morning's rising of the sun is identified by an ordered pair $<i, \wp>$ where i is the relevant time period (corresponding to the descriptor ‘this morning’) and $\wp$ is the sentence ‘The sun rises’. Of course, this treatment does not do justice to some of the intuitions underlying the *prima facie* commitments to events -- for instance, events can be perceived but times cannot [Gibson 1975]. But because of the availability of fully developed theories of intervals along with fully developed interval-based semantics [Cresswell 1979, Dowty 1979], and because of equally well worked out traditional theories of instants and instant-based semantics [Prior 1967], such accounts are especially attractive from a reductionist perspective. (A more general account would construe events as *spatiotemporal* regions cum description, distinguishing e.g between this morning's rising of the sun in London and its rising in Paris.)

The link between events and times has also been explored in the opposite direction, though. If events are assumed as a primitive ontological category, then one can dispense with temporal instants or intervals and construe them as derived entities. The most classical treatment of this sort proceeds by construing temporal instants as maximal sets of pairwise simultaneous (or partially simultaneous) events [Russell 1914, Whitehead 1929, Walker 1947], but other treatments are possible. For example, it has been

suggested that the mathematical connection between the way events are perceived to be ordered and the underlying temporal dimension is essentially that of a free construction (in the category-theoretic sense) of linear orderings from event orderings, induced by the binary relation *x wholly precedes y* [Thomason 1989]. Treatments such as these provide a reduction of time in terms of relations among events and are therefore especially germane to a relational conception of time (and, more generally, of space-time). Modal variants [Forbes 1993] as well as mereological variants [Pianesi & Varzi 1996] of such views are also available.

2. Types of Events

2.1 Activities, Accomplishments, Achievements, and States

Philosophers who agree with a conception of events as particulars typically distinguish different sorts of such particulars. A classic typology distinguishes four sorts: activities, accomplishments, achievements, and states [Ryle 1949, Vendler 1957]. An *activity*, such as John's walking uphill, is a homogeneous event: its sub-events satisfy the same description as the activity itself and has no natural finishing point or culmination. An *accomplishment*, such as John's climbing the mountain, may have a culmination, but is never homogeneous. An *achievement*, such as John's reaching the top, *is* a culminating event (and is therefore always instantaneous). And a *state*, such as John's knowing the shortest way, is homogeneous and may extend over time, but it makes no sense to ask how long it took or whether it culminated. Sometimes accomplishments and achievements are grouped together into a single category of *performances* [Kenny 1963]. Sometimes achievements have also been called *events* tout court and all other events have been grouped together into a broadly understood category of temporally extended entities, called *processes* [Ingarden 1935]; the word '*eventuality*' may then be used as a label covering both categories [Bach 1986].

Some authors introduce aspectual considerations into the taxonomy, drawing on Aristotle's distinction between *Energeia* and *Kinêsis* [Ackrill 1965]. The idea is that different verbs describe different types of events: verbs with no continuous form ('know') correspond to states; verbs with continuous form for which the present continuous entails the past perfect ('John is walking uphill' entails 'John walked uphill') correspond to activities; and verbs for which the present continuous entails the negation of the past perfect ('John is climbing the mountain' entails 'John has not (yet) climbed the mountain', at least in the relevant context) correspond to performances [Mourelatos 1978]. Several authors have followed in these footsteps to develop linguistically sophisticated theories [Taylor 1977, Dowty 1979, Bach 1981, Galton 1984, Verkuyl 1989], but the legitimacy of drawing ontological categorizations from such linguistic distinctions has been questioned [Gill 1993].

2.2 Static and Dynamic Events

One may want to distinguish between *dynamic* events, such as John's walk, and *static* events, such as John's rest under a tree. According to some authors, the latter are not events proper because they do not

involve any change [Ducasse 1926]. In the most abstract construal, a change is an ordered pair of facts: the fact that obtains prior to the change and the fact that obtains after the change took place [von Wright 1963]. More substantial accounts of events as changes describe them as the exemplifications of dynamic properties, i.e., properties that an object has by virtue of a "movement" in some quality space [Lombard 1979]. However, the question of whether all events should be or involve changes of some sort has been found by most authors to be ultimately a matter of conceptual stipulation -- hence of little metaphysical import.

If static events are admitted, the question arises of whether they should be kept distinct from states [Parsons 1989]. One plausible assumption is that the distinction between the static and the dynamic aspects of the world is skew to the distinction between states and activities. As there may be static activities, so there may be dynamic states. Walking is a state of John's that is dynamic, as opposed to his state of resting, which is static. The walk itself is an activity of John's that is dynamic, as opposed to the rest John takes, which can be considered a static activity.

2.3 Actions and Bodily Movements

Prima facie, actions are categorized as a subclass of events, namely, animate events. Like all events, actions are said to occur or take place, not to exist, and their relation to time and space is event-like as well: they have relatively clear beginnings and endings but unclear spatial boundaries, they appear to tolerate co-location, and they cannot be said to move from one place to another or to endure from one time to another, but rather extend in space and time by having spatial as well as temporal parts [Thomson 1977]. Actions and events appear to be homogeneous in causal explanations, too: actions can be causes of which events are effects [Davidson 1967b]. Some authors, however, prefer to draw a distinction here and to treat actions as *relations* between agents and events, namely as instances of the relation of ‘bringing about’ that may hold between an agent and an event [von Wright 1963, Chisholm 1964, Bach 1980]. On such views, actions are not individuals unless relations are themselves construed as tropes.

Whether or not actions are treated as events, one might be tempted to distinguish between *actions* proper (such as John's raising his arm) and *bodily movements* (such as John's arm rising), or between *intentional* actions (John's walk) and *unintentional* ones (John's falling into a hole). For some authors this is necessary in order to explain important facts of human behavior [Brand 1984, Mele 1997]. However, it has been argued that such a distinction does not pertain to metaphysics but rather to the conceptual apparatus by means of which we describe the realm of things that happen. On this view, an arm raising is just an arm's rising under a mentalistic description [Anscombe 1957, 1979].

2.4 Mental and Physical Events

A similar story applies to the distinction between *mental* events (John's decision to wear boots) and *physical* or *physiological* events (such and such neurons firing). One may think that this distinction is real insofar as the latter events are expected to fall naturally into the nomological net of physical theory whereas the former seem to escape it. But one may also want to resist this line of thought and maintain

that the distinction between the mental and the physical concerns exclusively the vocabulary with which we describe what goes on. These options have important ramifications for various issues philosophy of mind -- e.g., issues of mental causation [Heil & Mele 1993]. If the distinction between mental and physical events is ontologically significant, then the question arises of how they causally interact with each other, leading to various forms of anomalous or nomological dualism [Foster 1991]. By contrast, the claim that the distinction is purely semantic is congenial to a monist position, whether nomological (e.g., reductive materialism) or anomalous [Macdonald 1989]. Anomalous monism has been popular especially among philosophers who accept a particularist conception of events as widely redescribable entities, for such a conception allows one to accept the materialist claim that all events are physical (regardless of whether one describes them in mentalistic terms) while rejecting the seeming consequence that mental goings-on can be given purely physical explanations (precisely because only a physicalistic vocabulary is suited to such an explanation, so that it is only under its physical description that a mental event can be seen to enter into causal relations) [Nagel 1965, Davidson 1970, 1993]. Some authors, however, have argued that this line of argument falls prey to the charge of epiphenomenalism [Honderich 1982, Kim 1993] and on such matters the debate is still quite open.

3. Existence, Identity, and Indeterminacy

As mentioned in the Introduction, one finds a *prima facie* commitment to events in various aspects of human perception, action, language, and thought. The main line of argument offered to back this commitment up, however, comes from considerations of logical form. Not only does ordinary talk involve explicit reference to and quantification over events, as when one says that *John's walk* was pleasant or that *two explosions* were heard last night. Ordinary talk also seems to advert implicitly to events through adverbial modification [Reichenbach 1947]. We say that Brutus stabbed Caesar with a knife. If this statement is taken to assert that a certain three-place relation obtains among Brutus, Caesar, and a knife, then it is hard to explain why our statement entails that Brutus stabbed Caesar (a statement that involves a different, two-place relation) [Kenny 1963]. By contrast, if we take our statement to assert that a certain event occurred (namely, a stabbing of Caesar by Brutus) *and* that it had a certain property (namely, of being done with a knife), then the entailment is straightforward [Davidson 1967a]. These reasons do not constitute a proof that there are such entities as events. But they are telling insofar as one is interested in an account of how it is that certain statements mean what they mean, where the meaning of a statement is at least in part determined by its logical relations to other statements. For another example, it has been argued that singular causal statements cannot be analyzed in terms of a causal connective (essentially for reasons having to do with the above mentioned slingshot argument) but rather require that causation be treated as a binary relation holding between individual events [Davidson 1967b]. Still a third example involves the semantics of perceptual reports with *naked infinitive* complements, as in ‘John saw Mary cry’, which is analyzed as ‘John saw an event which was a crying by Mary’ [Higginbotham 1983]. Many more such arguments have been offered, also by authors working within different programs in linguistics [Parsons 1990, Schein 1993, Rothstein 1998, Link 1998, Higginbotham *et al*. 2000, Tenny & Pustejovsky, 2000].

On the other hand, some philosophers have been dissatisfied with this sort of "existential proof" and have

argued instead that all talk that seems to involve explicit or implicit reference to or quantification over events can be paraphrased so as to avoid the commitment. For example, it has been argued that a term such as ‘John's walk’ goes proxy for the corresponding statement ‘John walked’ [Geach 1965], so to say that John's walk was pleasant is just to say that John walked pleasantly. Similar paraphrases have been offered to deal with the case of explicit quantifier-phrases such as ‘two explosions’ as well as with the implicit event quantification that lies behind adverb-dropping inferences [Clark 1970], singular causal statements [Horgan 1978], and so on. On the face of it, it appears that questions of logical form leave the existential issue undecided, at least insofar as an event-committing analysis automatically turns into an eliminativist paraphrase when read in the opposite direction (and vice versa).

Another issue that appears to be undecided is that of identity criteria, which has been the focus of an intense debate [Pfeifer 1989]. Is John's walk the same event as his pleasant walk? Was Brutus's stabbing of Caesar the same event as his killing of Caesar? Was it the same as the violent assassination of Caesar? Some philosophers take these to be metaphysical questions -- questions whose answers call for identity criteria, which must be provided before we are allowed to take our event talk seriously. Different conceptions of events tend to suggest different answers, and widely varying ones. At one extreme we find the radical "unifiers" (who take events to be as coarse-grained as objects [Quine 1950]), at the other the radical "multipliers" (who take events to be as fine-grained as facts [Kim 1966]). Other philosophers, however, regard questions of identity to be first and foremost semantic questions -- questions about the way we talk and about what we say. No metaphysical theory, it is said, can settle the semantics of ordinary event talk, hence there is no way of determining the truth or falsity of an event identity statement exclusively on the basis of one's metaphysical views. Which events a statement speaks of depends heavily (more heavily than with ordinary material objects) on local context and unprincipled intuitions [Bennett 1988]. If so, the whole identity issue is undecidable, since one is demanding metaphysical answers to questions that are in large part semantical.

4. Conclusion

One could take the massive indeterminacy that seems to surround the existence and identity issues to be evidence that systematic theorizing about events is impossible. On the other hand, it is not clear that the indeterminacy is any worse in the case of events than in the case of objects, and we seem able to theorize in a systematic fashion about them [Lombard 1998]. If that is right, then the indeterminacy in our event concept would seem to be no fatal hindrance to the development of systematic theorizing about events.

Bibliography

Collections

- Casati, R., and Varzi, A. C. (eds.), 1996, *Events*, Dartmouth, Aldershot [referred to below as *Events*]
- Casati, R., and Varzi, A. C., 1997, *Fifty Years of Events. An Annotated Bibliography 1947 to 1997*,

Bowling Green (OH), Philosophy Documentation Center.

References

- Ackrill, J. L.,1965, ‘Aristotle's Distinction Between *Energeia* and *Kinêsis*’, in R. Bambrough (ed.), *New Essays on Plato and Aristotle*, London: Routledge and Kegan Paul, pp. 121-41.
- Anscombe, G. E. M, 1957, *Intention*, Oxford: Blackwell (second edition 1963).
- Anscombe, G. E. M., 1979, ‘Under a Description’, *Noûs*, 13, 219-33; reprinted in *Events*, pp. 303-17.
- Bach, K., 1980, ‘Actions Are Not Events’, *Mind*, 89, 114-20; reprinted in *Events*, pp. 343-49.
- Bach, E., 1981, ‘On Time, Tense and Aspect: An Essay in English Metaphysics’ , in P. Cole (ed.), *Radical Pragmatics*, New York: Academic Press, 63-81.
- Bach, E., 1986, ‘The Algebra of Events’, *Linguistics and Philosophy*, 9, 5-16; reprinted *Events*, pp. 497-508.
- Barwise, K. J., and Perry, J., 1981, ‘Semantic Innocence and Uncompromising Situations’, in P. A. French, T. Uehling, and H. K. Wettstein (eds.), *Foundations of Analytic Philosophy* (Midwest Studies in Philosophy, Vol. 6), Minneapolis: University of Minnesota Press, 387-403.
- Barwise, K. J., and Perry, J., 1983, *Situations and Attitudes*, Cambridge (MA) and London: MIT Press.
- Bennett, J., 1988, *Events and Their Names*, Oxford: Clarendon Press.
- Bennett, J., 1996, ‘What Events Are’, in *Events*, pp 137-151.
- Brand, M., 1984, *Intending and Acting. Toward a Naturalized Action Theory*, Cambridge (MA): MIT Press.
- Campbell, K., 1981, ‘The Metaphysic of Abstract Particulars’, in P. A. French, T. Uehling, and H. K. Wettstein (eds.), *Foundations of Analytic Philosophy* (Midwest Studies in Philosophy, Vol. 6), Minneapolis: University of Minnesota Press, pp. 477-88.
- Chisholm, R. M., 1964, ‘The Descriptive Element in the Concept of Action’, *Journal of Philosophy*, 61, 613-24.
- Chisholm, R. M., 1970, ‘Events and Propositions’, *Noûs*, 4, 15-24; reprinted in *Events*, pp. 89-98.
- Chisholm, R. M., 1971, ‘States of Affairs Again’, *Noûs*, 5, 179-89.
- Clark, R., 1970, ‘Concerning the Logic of Predicate Modifiers’, *Noûs*, 4, 311-35.
- Cresswell, M. J., 1979, ‘Interval Semantics for Some Event Expressions’, in R. Bäuerle, U. Egli, and A. von Stechow (eds.), *Semantics from Different Points of View*, Berlin and Heidelberg: Springer-Verlag, pp. 90-116.
- Davidson, D., 1967a, ‘The Logical Form of Action Sentences’, in N. Rescher (ed.), *The Logic of Decision and Action*, Pittsburgh: University of Pittsburgh Press, pp. 81-95; reprinted in *Events*, pp. 3-17, and in Davidson 1980, pp. 105-22.
- Davidson, D., 1967b, ‘Causal Relations’, *Journal of Philosophy*, 64, 691-703; reprinted in *Events*, pp. 401-13, and in Davidson 1980, pp. 149-62.
- Davidson, D., 1969, ‘The Individuation of Events’, in N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*, Dordrecht: Reidel, pp. 216 -- 34; reprinted in *Events*, pp. 265-83.
- Davidson, D., 1970, ‘Mental Events’, in L. Foster and J. W. Swanson (eds.), *Experience and Theory*, Amherst: University of Massachusetts Press, pp. 79-101; reprinted in Davidson (1980),

pp. 207-27.

- Davidson, D., 1980, *Essays on Actions and Events*, Oxford: Clarendon Press.
- Davidson, D., 1993, ‘Thinking Causes’, in J. Heil and A. R. Mele (eds.), *Mental Causation*, Oxford: Clarendon Press, pp. 3-17.
- Dowty, D. R., 1979, *Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and Montague's PTQ*, Reidel: Dordrecht.
- Dretske, F., 1967, ‘Can Events Move?’, *Mind*, 76, 479-92; reprinted in *Events*, pp. 415-428.
- Ducasse, C. J., 1926, ‘On the Nature and the Observability of the Causal Relation’, *Journal of Philosophy*, 23, 57-68.
- Forbes, G., 1993, ‘Time, Events and Modality’, in R. Le Poidevin and M. MacBeath (eds.), *The Philosophy of Time*, Oxford: Oxford University Press, pp. 80-95.
- Foster, J., 1991, *The Immaterial Self*, London and New York: Routledge.
- Galton, A. P., 1984, *The Logic of Aspect. An Axiomatic Approach*, Oxford: Clarendon Press.
- Geach, P., 1965, ‘Some Problems about Time’, *Proceedings of the British Academy*, 51, 321-36.
- Gibson, J. J., 1975, ‘Events Are Perceivable but Time IKs Not’, in J. T. Fraser and N. Lawrence (eds.), *The Study of Time II. Proceedings of the Second Conference of the International Society for the Study of Time*, Berlin: Springer-Verlag, pp. 295-301.
- Gill, K., 1993, ‘On the Metaphysical Distinction Between Processes and Events’, *Canadian Journal of Philosophy*, 23, 365-84; reprinted in *Events*, pp. 477-96.
- Goldman, A. I., 1970, *A Theory of Human Action*, New York, Prentice-Hall.
- Hacker, P. M. S., 1982a, ‘Events, Ontology and Grammar’, *Philosophy*, 57, 477-86; reprinted in *Events*, pp. 79-88.
- Hacker, P. M. S., 1982b, ‘Events and Objects in Space and Time’, *Mind*, 91, 1-19; reprinted in *Events*, pp. 429-47.
- Heil, J., and Mele, A., 1993, *Mental Causation*, Oxford: Clarendon Press.
- Higginbotham, J., 1983, ‘The Logic of Perceptual Reports: An Extensional Alternative to Situation Semantics’, *Journal of Philosophy*, 80, 100-27; reprinted in *Events*, pp. 19-46.
- Higginbotham, J., Pianesi, F., and Varzi, A. C. (eds.), 2000, *Speaking of Events*, Oxford: Oxford University Press.
- Honderich, T., 1982, ‘The Argument for Anomalous Monism’, *Analysis*, 42, 59-64.
- Horgan, T., 1978, ‘The Case Against Events’, *Philosophical Review*, 87, 28-47; reprinted in *Events*, pp. 243-62.
- Ingarden, R., 1935, ‘Vom formalen Aufbau des individuellen Gegenstandes’ [The Formal Structure of Individual Objects], *Studia Philosophica*, 1, 29-106.
- Kenny, A., 1963, *Action, Emotion and Will*, London: Routledge and Kegan Paul.
- Kim, J., 1966, ‘On the Psycho-Physical Identity Theory’, *American Philosophical Quarterly*, 3, 277-85.
- Kim, J., 1976, ‘Events as Property Exemplifications’, in M. Brand and D. Walton (eds.), *Action Theory*, Dordrecht: Reidel, pp. 159-77; reprinted in *Events*, pp. 117-35.
- Kim, J., 1993, *Supervenience and Mind: Selected Philosophical Essays*, Cambridge: Cambridge University Press.
- Lewis, D. K., 1986, ‘Events’, in his *Philosophical Papers, Volume 2*, New York: Oxford University Press, pp. 241-69; reprinted in *Events*, pp. 213-41.

- Link, G., 1998, *Algebraic Semantics and in Language and Philosophy*, Stanford: CSLI Publications.
- Lombard, L. B., 1979, ‘Events’, *Canadian Journal of Philosophy*, 9, 425-60; reprinted in *Events*, pp. 177-212.
- Lombard, L. B., 1986, *Events: a Metaphysical Study*, London: Routledge and Kegan Paul.
- Lombard, L. B., 1998, ‘Ontologies of Events’, in S. Laurence and C. Macdonald (eds.), *Contemporary Readings in the Foundations of Metaphysics*, Oxford: Blackwell, pp. 277-94.
- Macdonald, C. A., 1989, *Mind-Body Identity Theories*, London and New York: Routledge.
- Mayo, B., 1961, ‘Objects, Events, and Complementarity’, *Mind*, 70, 340-361.
- Martin, R., 1969, ‘On Events and Event-Descriptions’, in J. Margolis (ed.), *Fact and Existence*, Oxford: Basil Blackwell, pp. 63-73, 97-109.
- Mele, A. R. (ed.), 1997, *The Philosophy of Action*, Oxford: Oxford University Press.
- Mellor, D. H., 1980, ‘Things and Causes in Spacetime’, *British Journal for the Philosophy of Science*, 31, 282-88.
- Montague, R., 1969, ‘On the Nature of Certain Philosophical Entities’, *The Monist* 53, 159-94.
- Moravcsik, J. M. E., 1968, ‘Strawson and Ontological Priority’, in R. J. Butler (ed.), *Analytical Philosophy*, Second Series, New York: Barnes and Noble, pp. 106-19.
- Mourelatos, A. P. D., 1978, ‘Events, Processes, and States’, *Linguistics and Philosophy*, 2, 415-34; reprinted in *Events*, pp. 457-76.
- Nagel, T., 1965, ‘Physicalism’, *The Philosophical Review*, 74, 339-56.
- Parsons, T., 1989, ‘The Progressive in English: Events, States and Processes’, *Linguistics and Philosophy*, 12, 213-41; reprinted in *Events*, pp. 47-76.
- Parsons, T., 1990, *Events in the Semantics of English. A Study in Subatomic Semantics*, Cambridge (MA) and London: MIT Press.
- Parsons, T., 1991, ‘Tropes and Supervenience’, *Philosophy and Phenomenological Research*, 51, 629-32.
- Pfeifer, K., 1989, *Actions and Other Events: The Unifier-Multiplier Controversy*, New York and Bern: Peter Lang.
- Pianesi, F., and Varzi, A. C., 1996, ‘Events, Topology, and Temporal Relations’, *The Monist*, 78, 89-116.
- Prior, A., 1967, *Past, Present, and Future*, Oxford: Oxford University Press.
- Quine, W. V. O., 1950, ‘Identity, Ostension and Hyposthasis’, *Journal of Philosophy*, 47, 621-33.
- Quine, W. V. O., 1960, *Word and Object*, Cambridge (MA): MIT Press.
- Quine, W. V. O., 1970, *Philosophy of Logic*, Englewood Cliffs (NJ): Prentice-Hall.
- Quinton, A., 1979, ‘Objects and Events’, *Mind*, 88, 197-214.
- Ramsey, F. P., 1927, ‘Facts and Propositions’, *Proceedings of the Aristotelian Society*, Suppl. Vol. 7, 153-70.
- Reichenbach, H., 1947, *Elements of Symbolic Logic*, New York: Macmillan.
- Rothstein, S. (ed.), 1998, *Events and Grammar*, Dordrecht, Kluwer.
- Russell, B. A. W., 1914, *Our Knowledge of the External World*, London: Allen and Unwin.
- Ryle, G., 1949, *The Concept of Mind*, London: Hutchinson.
- Schein, B., 1993, *Plurals and Events*, Cambridge (MA) and London: MIT Press.
- Stout, G. F., 1923, ‘Are the Characteristic of Things Universal or Particular?’, *Proceedings of the*

Aristotelian Society, Supp. Vol. 3, 114-22.
- Strawson, P. F., 1959, *Individuals: An Essay in Descriptive Metaphysics*, London: Methuen.
- Taylor, B., 1977, ‘Tense and Continuity’, *Linguistics and Philosophy*, 1, 119-220.
- Taylor, B., 1985, Modes of Occurrence: Verbs, Adverbs and Events, Oxford: Blackwell.
- Tenny, C., and Pustejovsky, J. (eds.), 2000, *Events as Grammatical Objects: The Converging Perspectives of Lexical Semantics, Logical Semantics and Syntax*, Stanford (CA): CSLI Publications.
- Thomason, S. K., 1989, ‘Free Construction of Time from Events’, *Journal of Philosophical Logic*, 18, 43-67.
- Thomson, J. J., 1977, *Acts and Other Events*, Ithaca (NY): Cornell University Press.
- Tiles, J. E., 1981, *Things That Happen*, Aberdeen: Aberdeen University Press.
- Van Benthem, J., 1983, *The Logic of Time*, Dordrecht: Kluwer.
- Vendler, Z., 1957, ‘Verbs and Times’, *Philosophical Review*, 66, 143-60.
- Verkuyl, H. J., 1989, ‘Aspectual Classes and Aspectual Composition’, *Linguistics and Philosophy*, 12, 39-94.
- Von Kutschera, F., 1993, ‘Sebastian's Strolls’, *Grazer Philosophische Studien*, 45, 75-88.
- Von Wright, G. H., 1963, *Norm and Action. A Logical Inquiry*, London: Routledge and Kegan Paul.
- Walker, A. G, 1947, ‘Durées et instants’ [Durations and Instants], *Revue Scientifique*, 85, 131-34.
- Whitehead, A. N., 1919, *An Enquiry Concerning the Principles of Human Knowledge*, Cambridge: Cambridge University Press.
- Whitehead, A. N., 1929, *Process and Reality*. *An Essay in Cosmology*, New York: Macmillan.
- Williams, D. C., 1953 ‘On the Elements of Being’, *Review of Metaphysics*, 7, 3-18 (Part I), 171-92 (Part II).
- Wilson, N. L., 1974, ‘Facts, Events, and Their Identity Conditions’, *Philosophical Studies* 25, 303-21.
- Zacks, J., Tversky, B., and Iyer, G., 2001, ‘Perceiving, Remembering, and Communicating Structure in Events’, *Journal of Experimental Psychology: General*, 130, 29-58.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

action | anomalous monism | causation: the metaphysics of | Davidson, Donald | dualism | epiphenomenalism | facts | identity theory of mind | logical form | mental causation | supervenience | tropes

casati@ehess.fr
and
Achille C. Varzi
achille.varzi@columbia.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 22, 2002
Content last modified: August 2, 2002

Action

If a person's head moves, she may or may not have moved her head, and, if she did move it, she may have actively performed the movement of her head or merely, by doing something else, caused a passive movement. And, if she performed the movement, she might have done so intentionally or not. This short array of contrasts (and others like them) has motivated questions about the nature, variety, and identity of action. Beyond the matter of her moving, when the person moves her head, she may be indicating agreement or shaking an insect off her ear. Should we think of the consequences, conventional or causal, of physical behavior as constituents of an action distinct from but 'generated by' the movement? Or should we think that there is a single action describable in a host of ways? Also, actions, in even the most minimal sense, seem to be essentially 'active'. But, how can we explain what this property amounts to and defend our wavering intuitions about which events fall in the category of the 'active' and which do not?

Donald Davidson [1980, essay 3] asserted that an action, in some basic sense, is something an agent does that was 'intentional under some description,' and many other philosophers have agreed with him that there is a conceptual tie between genuine action, on the one hand, and intention, on the other. However, it is tricky to explicate the purported tie between the two concepts. First, the concept of 'intention' has various conceptual inflections whose connections to one another are not at all easy to delineate, and there have been many attempts to map the relations between intentions for the future, acting intentionally, and acting with a certain intention. Second, the notion that human behavior is often intentional under one description but not under another is itself hard to pin down. For example, as Davidson pointed out, an agent may intentionally cause himself to trip, and the activity that caused the tripping may have been intentional under that description while, presumably, the foreseen but involuntary tripping behavior that it caused is not supposed to be intentional under any heading. Nevertheless, both the tripping and its active cause are required to make it true that the agent intentionally caused himself to trip. Both occurrences fall equally, in that sense, 'under' the operative description. So further clarification is called for.

There has been a notable or notorious debate about whether the agent's reasons in acting are causes of the action -- a longstanding debate about the character of our common sense explanations of actions. Some philosophers have maintained that we explain why an agent acted as he did when we explicate how the agent's normative reasons rendered the action intelligible in his eyes. Others have stressed that the concept of 'an intention with which a person acted' has a teleological dimension that does not, in their view, reduce to the concept of 'causal guidance by the agent's reasons.' But the view that reason explanations are somehow causal explanations remains the dominant position. Finally, recent discussions have raised important new questions about the force of normative reasons for action in the context of the

agent's practical deliberation and related questions about the rational role these reasons have in moving him to act.

1. The Nature of Action and Agency

It has been common to motivate a central question about the nature of action by invoking an intuitive distinction between the things that merely *happen* to people -- the events they undergo -- and the various things they genuinely *do*. The latter events, the doings, are the *acts* or *actions* of the agent, and the problem about the nature of action is supposed to be: what distinguishes an action from a mere happening or occurrence? For some time now, however, there has been a better appreciation of the vagaries of the verb 'to do' and a livelier sense that the question is not well framed. For instance, a person may cough, sneeze, blink, blush, and thrash about in a seizure, and these are all things the person has, in some minimal sense, 'done,' although in the usual cases, the agent will have been altogether passive throughout these 'doings.' It is natural to protest that this is not the sense of "do" the canny philosopher of action originally had in mind, but it is also not so easy to say just what sense that is. Moreover, as Harry Frankfurt [1978] has pointed out, the purposeful behavior of animals constitutes a low-level type of 'active' doing. When a spider walks across the table, the spider *directly controls* the movements of his legs, and they are directed at taking him from one location to another. Those very movements have an aim or purpose for the spider, and hence they are subject to a kind of teleological explanation. Similarly, the idle, unnoticed movements of my fingers may have the goal of releasing the candy wrapper from my grasp. All this behavioral *activity* is 'action' in some fairly weak sense.

Nevertheless, a great deal of human action has a richer psychological structure than this. An agent performs activity that is directed at a goal, and commonly it is a goal the agent has adopted on the basis of an overall practical assessment of his options and opportunities. Moreover, it is immediately available to the agent's awareness both that he is performing the activity in question and that the activity is aimed by him at such-and-such a chosen end. At a still more sophisticated conceptual level, Frankfurt [1988, 1999] has also argued that basic issues concerning freedom of action presuppose and give weight to a concept of 'acting on a desire with which the agent *identifies*.' Under Frankfurt's influence on this point, a good deal has been written to elucidate the nature of 'full-blooded' human agency, whether the notion is finally delineated either in Frankfurt's way or along different but related lines [see Velleman 2000, essay 6, Bratman 1999, essay 10]. Thus, there are different levels of action to be distinguished, and these include at least the following: unconscious and/or involuntary behavior, purposeful or goal directed activity (of Frankfurt's spider, for instance), intentional action, and the autonomous acts or actions of self-

consciously active human agents. Each of the key concepts in these characterizations raises some hard puzzles.

1.1 Knowledge of one's own actions.

It is frequently noted that the agent has some sort of *immediate awareness* of his physical activity and of the goals that the activity is aimed at realizing. In this connection, Elizabeth Anscombe [1963] spoke of 'knowledge without observation.' The agent knows 'without observation' that he is performing certain bodily movements (perhaps under some rough but non-negligible description), and he knows 'without observation' what purpose(s) the behavior is meant to serve [see also Falvey 2000]. Anscombe's discussion of her claim is rich and suggestive, but her conception of 'knowledge through observation' is problematic. Surely, one wants to say, propioception and kinesthetic sensation play some role in informing the agent of the positions and movements of his body, and it is uncertain why these informational roles should fail to count as modes of inner 'observation' of the agent's own overt physical behavior. What Anscombe explicitly denies is that agents generally know of the positions or movements of their own bodies by means of 'separably describable sensations' that serve as criteria for their judgements about the narrowly physical performance of their bodies. However, when a person sees that there is a goldfinch in front of him, his knowledge is not derived as an inference from the 'separably describable' visual impressions he has in seeing the goldfinch, but this is an instance of knowledge through observation nonetheless.

In a related vein, David Velleman [1989] described this knowledge as 'spontaneous,' i.e., as knowledge that the agent has achieved without deriving it from evidence adequate to warrant it. However, it is not so plain that an agent's knowledge that certain of his movements have been guided by him toward an objective *O* are not derived from prior evidence, reached by him on the basis of a simple causal inference. That is, he knows, in an immediate, first person way, that he is committed to objective *O* as his goal. In addition, he knows, also immediately, that those movements are caused -- causally guided, as it were -- by the state of having *O* as his goal then. If these points are correct, it may be that an agent knows his present goals and intentions without inner or outer evidence, but it may also be that this same non-observational, non-inferential knowledge itself serves as evidence for his further belief that his current behavior is directed at a such-and-such goals. In the same way, an agent can often identify right off, apparently without consulting evidence at all, what action it is that he will perform *next*. Again, it may be that all that the individual agent really knows immediately is that he has an intention to do so-and-so next, and the knowledge that he is actually on the verge of doing so-and-so is grounded for him in an inference that takes his intention for the near future as its primary epistemic ground [see Wilson 2000, Moran 2001].

These considerations, if right, would mean that one's knowledge of what one is presently doing and one's knowledge of what one is about to do are not spontaneous, in Velleman's suggested sense. And yet, at this juncture, the issues are intertwined with difficult questions about the nature of intentions and their relations to first-person beliefs about one's forthcoming actions. Velleman and others reject the picture of evidential support sketched just above, maintaining that the agent's belief that he will shortly *F* is contemporaneous with and embodied in his intention to *F*. It cannot be, therefore, that his knowledge of

his intention to *F* provides the grounds from which his expectation of imminent *F*ing has been derived. The tangle of issues here merits additional unweaving in future research.

1.2 Governance of one's own actions.

It is also important to the concept of 'goal directed action' that agents normally implement a kind of *direct* control or guidance over their own behavior. An agent may guide her paralyzed left arm along a certain path by using her active right arm to shove it through the relevant trajectory. The moving of her right arm, activated as it is by the normal exercise of her system of motor control, is a genuine action, but the movement of her left arm is not. That movement is merely the causal upshot of her guiding action, just as the onset of illumination in the light bulb is the mere effect of her action when she turned on the light. The agent has direct control over the movement of the right arm, but not over the movement of the left. And yet it is hardly clear what 'direct control of behavior' can amount to here. It does not simply mean that behavior *A*, constituting a successful or attempted *F*ing, was initiated and causally guided throughout its course by a present-directed intention to be *F*ing then. Even the externally guided movement of the paralyzed left arm would seem to satisfy a condition of this weak sort. Alfred Mele [1992] has suggested that the intuitive 'directness' of the guidance of action *A* can partially be captured by stipulating that the action-guiding intention must trigger and sustain *A proximally*. In other words, it is stipulated that the agent's present-directed intention to be *F*ing should govern action *A*, but not by producing some other prior or concurrent action *A** that causally controls *A* in turn. But the proposal is dubious. On certain assumptions, most ordinary physical actions are liable to flunk this strengthened requirement. The normal voluntary movements of an agent's limbs are caused by complicated contractions of suitable muscles, and the muscle contractions, since they are aimed at causing the agent's limbs to move, may themselves count as causally prior human actions. For instance, on Davidson's account of action they will since the agent's muscle contracting is intentional under the description 'doing something that causes the arm to move' [see Davidson 1980, essay 2]. Thus, the overt arm movement, in a normal act of voluntary arm moving, *will* have been causally guided by a prior action, the muscle contracting, and consequently the causal guidance of the arm's movement will fail to be an instance of 'proximal' causation at all [see Sehon 1998].

As one might imagine, this conclusion depends upon how an act of moving a part of one's body is to be conceived. Some philosophers maintain that the movements of an agent's body are never actions. It is only the agent's direct *moving* of, say, his leg that constitutes a physical action; the leg movement is merely caused by and/or incorporated as a part of the act of moving [see Hornsby 1980]. This thesis re-opens the possibility that the causal guidance of the moving of the agent's leg by the pertinent intention *is* proximal after all. The intention proximally governs the moving, if not the movement, where the act of moving is now thought to start at the earliest, inner stage of act initiation. Still, this proposal is also controversial. For instance, J.L. Austin [1962] held that the statement

> (1) The agent moved his leg

is ambiguous between (roughly)

(1′) The agent caused his leg to move

and the more specific

(1″) The agent performed a movement with his leg.

If Austin is right about this, then the nominalization “the agent’s moving of his leg” should be correspondingly ambiguous, with a second reading that denotes a certain leg movement, a movement the agent has *performed*. Thus, no simple appeal to a putative distinction between ‘movement’ and ‘moving’ will easily patch up the conception of ‘direct control of action’ under present scrutiny.

In any event, there is another well-known reason for doubting that the ‘directness’ of an agent’s governance of his own actions involves the condition of causal proximality -- that an action is not to be controlled by still another action of the same agent. Some philosophers believe that the agent’s moving his leg is triggered and sustained by the agent’s *trying* to move his leg in just that way, and that the efficacious trying is itself an action [see Hornsby 1980, Ginet 1990, and O’Shaughnessy 1973, 1980]. If, in addition, the agent’s act of leg moving is distinct from the trying, then, again, the moving of the leg has not been caused proximally by the intention. The truth or falsity of this third assumption is linked with a wider issue about the individuation of action that has also been the subject of elaborate discussion.

Donald Davidson [1980, essay 1], concurring with Anscombe, held that

(2) If a person *F*s *by G*ing, then her act of *F*ing = her act of *G*ing.

In Davidson’s famous example, someone alerts a burglar by illuminating a room, which he does by turning on a light, which he does in turn by flipping the appropriate switch. According to the Davidson/Anscombe thesis above, the alerting of the burglar = the illuminating of the room = the turning on of the light = the flipping of the switch. And this is so despite the fact that the alerting of the burglar was unintentional while the flipping of the switch, the turning on of the light, and the illuminating of the room were intentional. Suppose now that it is also true that the agent moved his leg *by* trying to move his leg in just that matter. Combined with the Davidson/Anscombe thesis about act identification, this implies that the agent’s act of moving his leg = his act of trying to move that leg. So, perhaps the act of trying to move the leg doesn’t cause the act of moving after all, since they are just the same.

The questions involved in these debates are potentially quite confusing. First, it is important to distinguish between phrases like

(a) the agent’s turning on the light

and gerundive phrases such as

(b) the agent's turning on of the light.

Very roughly, the expression (a) operates more like a 'that' clause, viz.

(a′) that the agent turned on the light,

while the latter phrase appears to be a definite description, i.e.,

(b′) the turning on of the light [performed] by the agent.

What is more, even when this distinction has been drawn, the denotations of the gerundive phrases often remain ambiguous, especially when the verbs whose nominalizations appear in these phrases are causatives. No one denies that there is an internally complex process that is initiated by the agent's switch-flipping hand movement and that is terminated by the light's coming on as a result. This process includes, but is not identical with, the act that initiates it and the event that is its culminating upshot. Nevertheless, in a suitable conversational setting, the phrases (b) and (b′) can be properly used to designate any of the three events: the act that turned on the light, the onset of illumination in the light, and whole process whereby the light has come to be turned on. [For further discussion, see Parsons 1990, Pietrofsky 2000, and Higgenbotham 2000].

Now, the Davidson-Anscombe thesis plainly is concerned with the relation between the agent's *act* of turning on the light, his *act* of flipping the switch, etc. But which configuration of events, either prior to or contained within the extended causal process of turning on the light, really constitutes the agent's action? Some philosophers have favored the overt arm movement the agent performs, some favor the extended causal process he initiates, and some prefer the relevant event of trying that precedes and 'generates' the rest. It has proved difficult to argue for one choice over another without simply begging the question against competing positions. As noted before, Hornsby and other authors have pointed to the intuitive truth of

(3) The agent moved his arm *by* trying to move his arm,

and they appeal to the Davidson-Anscombe thesis to argue that the act of moving the arm = the *act* of trying to move the arm. On this view, the act of trying -- which *is* the act of moving -- causes a movement of the arm in much the same way that an act of moving the arm causes the onset of illumination in the light. Both the onset of illumination and the overt arm movement are simply causal consequences of the act itself, the act of trying to move his arm in just this way. Further, in light of the apparent immediacy and strong first person authority of agents' judgements that they have tried to do a certain thing, it appears that acts of trying are intrinsically mental acts. So, a distinctive type of mental act stands as the causal source of the bodily behavior that validates various physical re-descriptions of the act.

And yet none of this seems inevitable. It is arguable that

(4) The agent tried to turn on the light

simply means, as a first approximation at least, that

(4′) The agent *did something* that was directed at turning on the light.

Moreover, when (4) or (4′) is true, then the something the agent did that was directed at turning on the light will have been some other causally prior action, the act of flipping the switch, for example. If this is true of trying to perform basic acts (e.g., moving one's own arm) as well as non-basic, instrumental acts, then trying to move one's arm may be nothing more than doing something directed at making one's arm move. In this case, the something which was done may simply consist in the contracting of the agent's muscles. Or, perhaps, if we focus on the classic case of the person whose arm, unknown to her, is paralyzed, then the trying in that case (and perhaps in all) may be nothing more than the activation of certain neural systems in the brain. Of course, most agents are not aware *that* they are initiating appropriate neural activity, but they are aware *of* doing something that is meant to make their arms move. And, in point of fact, it may well be that the something of which they are aware as a causing of the arm movement just is the neural activity in the brain. From this perspective, ‘trying to *F*’ does not name a natural kind of mental act that ordinarily sets off a train of fitting physical responses. Rather, it gives us a way of describing actions in terms of a goal aimed at in the behavior without committing us as to whether the goal was realized or not. It also carries no commitment,

i. concerning the intrinsic character of the behavior that was aimed at *F*ing,
ii. whether one or several acts were performed in the course of trying, and
iii. whether any further bodily effects of the trying were themselves additional physical actions [see Cleveland 1997].

By contrast, it is a familiar doctrine that what the agent does, in the first instance, in order to cause his arm to move is to form a distinctive mental occurrence whose intrinsic psychological nature and content is immediately available to introspection. The agent *wills* his arm to move or produces a *volition* that his arm is to move, and it is this mental willing or volition that is aimed at causing his arm to move. Just as an attempt to turn on the light may be constituted by the agent's flipping of the switch, so also, in standard cases, trying to move his arm is constituted by the agent's willing his arm to move. For traditional ‘volitionalism,’ willings, volitions, basic tryings are, in Brian O'Shaughnessy's apt formulation, ‘primitive elements of animal consciousness.’[1] They are elements of consciousness in which the agent has played an active role, and occurrences that normally have the power of producing the bodily movements they represent. Nevertheless, it is one thing to grant that, in trying to move one's body, there is some ‘inner’ activity that is meant to initiate an envisaged bodily movement. It is quite another matter to argue successfully that the initiating activity has the particular mentalistic attributes that volitionism has characteristically ascribed to acts of willing.

It is also a further question whether there is only a *single* action, bodily or otherwise, that is performed along the causal route that begins with trying to move and terminates with a movement of the chosen

type. One possibility, adverted to above, is that there is a whole causal chain of *actions* that is implicated in the performance of even the simplest physical act of moving a part of one's body. If, for example, 'action' is goal-directed behavior, then the initiating neural activity, the resulting muscle contractions, and the overt movement of the arm may *all* be actions on their own, with each member in the line-up causing every subsequent member, and with all of these actions causing an eventual switch flipping somewhere further down the causal chain. On this approach, there may be nothing which is *the* act of flipping the switch or of turning on the light, because each causal link is now an act which flipped the switch and (thereby) turned on the light [see Wilson 1989]. Nevertheless, there still will be a single *overt* action that made the switch flip, the light turn on, and the burglar become alert, i.e., the overt movement of the agent's hand and arm. In this sense, the proposal supports a modified version of the Davidson/Anscombe thesis.

However, all of this discussion suppresses a basic metaphysical mystery. In the preceding two paragraphs, it has been proposed that the neural activity, the muscle contractions, and the overt hand movements may all be actions, while the switch's flipping on, the light's coming on, and the burglar's becoming alert are simply happenings outside the agent, the mere effects of the agent's overt action. As we have seen, there is plenty of disagreement about where basic agency starts and stops, whether within the agent's body or somewhere on its surface. There is less disagreement that the effects of bodily movement beyond the body, e.g., the switch's flipping on, the onset of illumination in the room, and so on, are not, by themselves at least, purposeful actions. Still, what could conceivably rationalize *any* set of discriminations between action and non-action as one traces along the pertinent complex causal chains from the initial mind or brain activity, through the bodily behavior, to the occurrences produced in the agent's wider environment?

Perhaps, one wants to say, as suggested above, that the agent has a certain kind of direct (motor) control over the goal-seeking behavior of his own body. In virtue of that fundamental biological capacity, his bodily activity, both inner and overt, is governed by him and directed at relevant objectives. Inner physical activity causes and is aimed at causing the overt arm movements and, in turn, those movements cause and are aimed at causing the switch to flip, the light to go on, and the room to become illuminated. Emphasizing considerations of this sort, one might urge that they validate the restriction of action to events in or at the agent's body. And yet, the stubborn fact remains that the agent also does have a certain 'control' over what happens to the switch, the light, and even over the burglar's state of mind. It is a goal for the agent of the switch's flipping on that it turn on the light, a goal for the agent of the onset of illumination in the room that it render the room space visible, etc. Hence, the basis of any discrimination between minimal agency and non-active consequences within the extended causal chains will have to rest on some special feature of the person's guidance: the supposed 'directness' of the motor control, the immediacy or relative certainty of the agent's expectations about actions vs. results, or facts concerning the special status the agent's living body. The earlier remarks in this section hint at the serious difficulty of seeing how any such routes are likely to provide a rationale for grounding the requisite metaphysical distinction(s).

2. Intentional Action and Intention

Anscombe opened her monograph *Intention* by noting that the concept of 'intention' figures in each of the constructions:

> (5) The agent intends to *G*;
> (6) The agent *G*'d intentionally; and
> (7) The agent *F*'d with the intention of *G*ing,

For that matter, one could add

> (7′) In *F*ing (by *F*ing), the agent intended to *G*.

Although (7) and (7′) are closely related, they seem not to say quite the same thing. For example, although it may be true that

> (8) Veronica mopped the kitchen then with the intention of feeding her flamingo afterwards,

it normally won't be true that

> (8′) In (by) mopping the kitchen, Veronica intended to feed her flamingo afterwards.

Despite the differences between them, I will call instances of (7) and (7′) ascriptions of *intention in action*.[2] These sentential forms represent familiar, succinct ways of *explaining* action. A specification of the intention with which an agent acted or the intention that the agent had in acting provides a common type of explanation of why the agent acted as he did. This observation will be examined at some length in Section 3.

Statements of form (5) are ascriptions of *intention for the future*, although, as a special case, they include ascriptions of *present-directed intentions*, i.e., the agent's intention to be *G*ing *now*. Statements of form 6), ascriptions of *acting intentionally*, bear close connections to corresponding instances of (7). As a first approximation at least, it is plausible that (6) is true just in case

> (6′) The agent *G*'d with the intention of (thereby) *G*ing.

However, several authors have questioned whether such a simple equivalence captures the special complexities of what it is to *G* intentionally.[3] Here is an example adapted from Davidson [1980, essay 4]. Suppose that Betty kills Jughead, and she does so with the intention of killing him. And yet suppose also that her intention is realized only by a wholly unexpected accident. The bullet she fires misses Jughead by a mile, but it dislodges a tree branch above his head and releases a swarm of hornets that attack him and sting him until he dies. In this case, it is at least dubious that, in this manner, Betty has killed Jughead *intentionally*. (It is equally doubtful that Betty killed him *unintentionally* either.) Or

suppose that Reggie wins the lottery, and having bizarre illusions about his ability to control which ticket will win, he enters the lottery and wins it with the intention of winning it [Mele 1997]. The first example suggests that there needs to be some condition added to (6′) that says the agent succeeded in *G*ing in a manner sufficiently in accordance with whatever plan she had for *G*ing as she acted. The second suggests that the agent's success in *G*ing must result from her competent exercise of the relevant skills, and it must not depend too much on sheer luck, whether the luck has been foreseen or not. Various other examples have prompted additional emendations and qualifications [see Harman 1976].

There are still more fundamental issues about intentions in action and how they are related to intentions directed at the present and the immediate future. In "Actions, Reasons, and Causes," Davidson seemed to suppose that ascriptions of intention in action reduce to something like the following.

> (7*) The agent *F*'ed, and at that time he had a pro-attitude toward *G*ing and believed that by *F*ing he would or might promote *G*ing, and the pro-attitude in conjunction with the means-end belief caused his *F*ing, and together they caused it 'in the right way.'

(In Davidson's widely used phrase, the pro-attitude and associated means-end belief constitute a *primary reason* for the agent to *F*.) In this account of 'acting with an intention' there is, by design, no mention of a distinctive state of intending. Davidson, at the time of this early paper, seemed to favor a reductive treatment of intentions, including intentions for the future, in terms of pro-attitudes, associated beliefs, and other potential mental causes of action. In any case, Davidson's approach to intention in action was distinctly at odds with the view Anscombe had adopted in *Intention*. She stressed the fact that constructions like (7) and (7′) supply commonsense explanations of why the agent *F*'d, and she insisted that the explanations in question do not cite the agent's reasons as causes of the action. Thus, she implicitly rejected anything like (7*), the causal analysis of 'acting with a certain intention' that Davidson apparently endorsed. On the other hand, it was less than clear from her discussion how it is that intentions give rise to an alternative mode of action explanation.

Davidson's causal analysis is modified in his later article "Intending" [1980, essay 5]. By the time of this essay, he dropped the view that there is no primitive state of intending. Intentions are now accepted as irreducible, and the category of intentions is distinguished from the broad, diverse category that includes the various pro-attitudes. In particular, he identifies intentions for the future with the agent's all-out judgments (evaluations) of what she is to do. Although there is some lack of clarity about the specific character of these practical 'all-out' judgements, they play an important role in Davidson's overall theory of action, particularly in his striking account of weakness of will [1980, essay 2]. Despite his altered outlook on intentions, however, Davidson does not give up the chief lines of his causal account of intentions in action -- of what it is to act with a certain intention. In the modified version,

> (7**) The agent's primary reason for *G*ing must cause her, in the right way, to intend to *G*, and her intending to *G* must itself cause, again in the right way, the agent's particular act of *F*ing.[4]

The interpolated, albeit vague, conditions that require causation in ‘the right way’ are meant to cover well-known counterexamples that depend upon deviant causal chains occurring either in the course of the agent’s practical reasoning or in the execution of his intentions. Here is one familiar type of example. A waiter intends to startle his boss by knocking over a stack of glasses in their vicinity, but the imminent prospect of alarming his irascible employer unsettles the waiter so badly that he involuntarily staggers into the stack and knocks the glasses over. Despite the causal role of the waiter's intention to knock over the glass, he doesn't do this intentionally. In this example, where the deviant causation occurs as part of the performance of the physical behavior itself, we have what is known as ‘primary causal deviance.’ When the deviant causation occurs on the path between the behavior and its intended further effects -- as in the example of Betty and Jughead above -- the deviance is said to be ‘secondary.’ There have been many attempts by proponents of a causal analysis of intention in action (‘causalists,’ in the terminology of von Wright 1971) to spell out what ‘the right kind(s)’ of causation might be, but with little agreement about their success [see Bishop 1989, Mele 1997]. Some other causalists, including Davidson, maintain that no armchair analysis of this matter is either possible or required. However, most causalists agree with Davidson’s later view that the concept of ‘present directed intention’ is needed in any plausible causal account of intention in action and acting intentionally. It is, after all, the present directed intention that is supposed to guide causally the ongoing activity of the agent [see also Searle 1983].

The simplest version of such an account depends on what Michael Bratman has dubbed “the Simple View.” This is the thesis that proposition (6) above, [The agent *G*’d intentionally] and, correspondingly, proposition (7) [The agent *F*ed with the intention of *G*ing] entail that, at the time of action, the agent intended *to G*. Surely, from the causalist point of view, the most natural account of *G*ing intentionally is that the action of *G*ing is governed by a present directed intention whose content for the agent is, “I am *G*ing now.’ So the causalist’s natural account presupposes the Simple View, but Bratman [1984, 1987] has presented a well-known example to show that the Simple View is false. He describes a type of case in which the agent wants either to Φ or to Θ, without having any significant preference between the two alternatives. The agent does know, however, that it is flatly impossible, in the given circumstances, for him to *both* Φ and Θ although, in these same circumstances, it *is* open to him to try to Φ and try to Θ concurrently. (Perhaps, in trying to Φ, he does something with one hand, and, in trying to Θ, he does something with the other.) Believing that such a two-pronged strategy of trying to achieve each goal maximizes his chances of achieving his actual goal of either Φing or Θing, the agent actively aims at both of the subordinate ends, trying to accomplish one or the other. The example can be spelled out in such a way that it is clear that the agent is wholly rational, in his actions and attitudes, as he knowingly pursues this bifurcated attack on his disjunctive goal. Suppose now that the agent actually succeeds in, say, Φing and that he succeeds in virtue of his skill and insight, and not through some silly accident. So, the agent Φ’s intentionally. It follows from the Simple View that the agent intended to Φ. And yet, the agent was also doing something with the intention of Θing and had this attempt succeeded instead (without the intervention of too much luck), then the agent would have Θ’d intentionally. By a second application of the Simple View, it follows that he also intended to Θ. And yet, just as it is irrational to intend to Φ while believing that it is flatly impossible for him to Φ, so also does it seem irrational to have an intention to Φ *and* an intention to Θ, while believing that it is flatly impossible to do the two things together. So the agent here should be open to criticisms of irrationality in his endeavor to Φ or Θ. Nevertheless, we observed at the outset that he is not. The only way out is to block the conclusion that, in

trying to Φ and trying to Θ in these circumstances, the agent has the contextually irrational pair of intentions, and rejecting the Simple View is the most direct manner of blocking that conclusion.

Even if Bratman's argument defeats the Simple View [see McCann 1986], it doesn't rule out some type of causal analysis of acting intentionally; it doesn't even rule out such an analysis that takes the crucial controlling cause to be an intention in every instance. One might suppose, for example, that (i) in a Bratman case, the agent merely intends to *try* to Φ and intends to *try* to Θ, and that (ii) it is these intentions that drive the agent's actions [Mele 1997]. The analysis in (7**) would be modified accordingly. However, the project of finding a workable and non-circular emendation of (7**) remains an open question.

The conceptual situation is complicated by the fact that Bratman holds that (7) [The agent *F*'d with the intention of *G*ing] is ambiguous between

> The agent *F*'d with the aim or goal of *G*ing

and

> The agent *F*'d as part of a plan that incorporated an intention to *G*.

(8) above is an especially clear example in which the second reading is required. The second reading does entail that the agent intends to *F*, and it is only the first that, according to Bratman's argument, does not. Therefore, Bratman thinks that we need to distinguish intention as an aim or goal of actions and intention as a distinctive state of commitment to future action, a state that results from and subsequently constrains our practical endeavors as planning agents. It can be rational to aim at a pair of ends one knows to be jointly unrealizable, because aiming at both may be the best way to realize one or the other. However, it is not rational to plan on accomplishing both of two objectives, known to be incompatible, since intentions that figure in rational planning should agglomerate, i.e., should fit together in a coherent larger plan. Bratman's example and the various critical discussions of it have promoted important topics concerning the very idea of the *rationality* of actions and intentions, measured against the backdrop of the agent's beliefs and suppositions.

It has been mentioned earlier that Davidson came to identify intentions for the future with all out judgements about what the agent is to be doing now or should do in the relevant future. Velleman [1989], by contrast, identifies an intention with the agent's spontaneous belief, derived from practical reflection, which says that he is presently doing a certain act (or that he will do such an act in the future), and that his act is (or will be) performed precisely as a consequence of his acceptance of this self-referential belief. Paul Grice [1971] favored a closely related view in which intention consists in the agent's willing that certain results ensue, combined with the belief that they will ensue as a consequence of the particular willing in question. Hector-Neri Castañeda [1975], influenced by Sellars [1966] maintained that intentions are a special species of internal self-command, which he calls "practitions." Bratman [1987] develops a functionalist account of intention: it is the psychological state that plays a certain kind of

characteristic causal role in our practical reasoning, in our planning for the future, and in the carrying out of our actions. This causal role, he argues, is distinct from the characteristic causal or functional roles of expectations, desires, hopes, and other attitudes about the agent's future actions.

3. The Explanation of Action

For many years, the most intensely debated topic in the philosophy of action concerned the explanation of intentional actions in terms of the agent's reasons for acting. As stated previously, Davidson and other action theorists defended the position that reason explanations are causal explanations -- explanations that cite the agent's desires, intentions, and means-end beliefs as causes of the action [see Goldman 1970]. These causalists about the explanation of action were reacting against a neo-Wittgensteinian outlook that claimed otherwise. In retrospect, the very terms in which the debate was conducted were flawed. First, for the most part, the non-causalist position relied chiefly on negative arguments that purported to show that, for conceptual reasons, motivating reasons could not be causes of action. Davidson did a great deal to rebut these arguments. It was difficult, moreover, to find a reasonably clear account of what sort of non-causal explanation the neo-Wittgensteinians had in mind. Charles Taylor, in his book *The Explanation of Action* [1964], wound up claiming that reason explanations are grounded in a kind of 'non-causal bringing about,' but neither Taylor nor anyone else ever explained how any bringing about of an event could fail to be causal. Second, the circumstances of the debate were not improved by the loose behavior of the ordinary concept of 'a cause.' When someone says that John has cause to be offended by Jane's truculent behavior, then "cause" in this setting just means 'reason,' and the statement, "John was caused to seek revenge by his anger," may means nothing more than, "John's anger was among the reasons for which he sought revenge." If so, then presumably no one denies that reasons are *in some sense* causes. In the pertinent literature, it has been common to fall back on the qualified claim that reasons are not 'efficient' or 'Humean' or 'producing' causes of action. Unfortunately, the import of these qualifications has been less than perspicuous.

George Wilson [1989] and Carl Ginet [1990] follow Anscombe in holding that reason explanations are distinctively grounded in an agent's intentions in action. Both authors hold that ascriptions of intention in action have the force of propositions that say *of* a particular act of *F*ing that it was intended by its agent to *G* (by means of *F*ing), and they claim that such *de re* propositions constitute non-causal reason explanations of why the agent *F*ed on the designated occasion. Wilson goes beyond Ginet in claiming that statements of intention in action have the meaning of

> (9) The agent's act of *F*ing was directed by him at [the objective] of *G*ing,

In this analyzed form, the teleological character of ascriptions of intention in action is made explicit. Given the goal-directed nature of action, one can provide a familiar kind of teleological explanation of the relevant behavior by mentioning a goal or purpose of the behavior for the agent at the time, and this is the information (9) conveys. Or, alternatively, when a speaker explains that

> (10) The agent *F*'d because he wanted to *G*,

the agent's desire to *G* is cited in the explanation, not as a cause of the *F*ing, but rather as indicating a desired goal or end at which the act of *F*ing came to be directed.

Most causalists will allow that reason explanations of action are teleological but contend that teleological explanations in terms of goals -- purposive explanations in other words -- are themselves analyzable as causal explanations in which the agent's primary reason(s) for *F*ing are specified as guiding causes of the act of *F*ing. Therefore, just as there are causalist analyses of what it is to do something intentionally, so there are similar counterpart analyses of teleological explanations of goal directed and, more narrowly, intentional action. The causalist about teleological explanation maintains that the goal of the behavior for the agent just is a goal the agent had at the time, one that caused the behavior and, of course, one that caused it in the right way [for criticism, see Sehon 1998].

It has not been easy to see how these disagreements are to be adjudicated. The claim that purposive explanations do or do not *reduce* to suitable counterpart causal explanations is surprisingly elusive. It is not clear, in the first place, what it is for one form of explanation to reduce to another. Moreover, as indicated above, Davidson himself has insisted that it is not possible to give an explicit, reductive account of what 'the right kind of causing' is supposed to be and that none is needed. Naturally, he may simply be right about this, but others have felt that causalism about reason explanations is illicitly protected by endemic fuzziness in the concept of 'causation of the right kind.' Some causalists who otherwise agree with Davidson have accepted the demand for a more detailed and explicit account, and some of the proposed accounts get extremely complicated. Without better agreement about the concept of 'cause' itself, the prospects for a resolution of the debate do not appear cheerful. Finally, Abraham Roth [2000] has pointed out that reasons explanations might both be irreducibly teleological *and* also cite primary reasons as efficient causes at the same time. It is arguable that similar explanations, having both causal and teleological force, figure already in specifically homeostatic (feedback) explanations of certain biological phenomena. When we explain that the organism *V*ed *because* it needed *W*, we may well be explaining both that the goal of the *V*ing was to satisfy the need for *W and* that it was the need for *W* that triggered the *V*ing.

One of the principal arguments that were used to show that reason explanations of action could not be causal was the following. If the agent's explaining reasons *R* were among the causes of his action *A*, then there must be some universal causal law which nomologically links the psychological factors in *R* (together with other relevant conditions) to the *A*-type action that they rationalize. However, it was argued, there simply are no such psychological laws; there are no strict laws and co-ordinate conditions that ensure that a suitable action will be the invariant product of the combined presence of pertinent pro-attitudes, beliefs, and other psychological states. Therefore, reasons can't be causes. In "Actions, Reasons, and, Causes," Davidson first pointed out that the thesis that there are no reason-to-action laws is crucially ambiguous between a stronger and a weaker reading, and he observes that it is the stronger version that is required for the non-causalist conclusion. The weaker reading says that there are no reason-to-action laws in which the antecedent is formulated in terms of the 'belief/desire/intention' vocabulary of commonsense psychology and the consequent is stated in terms of goal directed and intentional action. Davidson accepted that the thesis, on this reading, is correct, and he has continued to accept it ever since.

The stronger reading says that there are no reason-to-action laws in any guise, including laws in which the psychological states and events are re-described in narrowly physical terms and the actions are re-described as bare movement. Davidson affirms that there *are* laws of this second variety, whether we have discovered them or not.[5]

Many have felt that this position only lands Davidson (*qua* causalist) in deeper trouble. It is not simply that we suppose that states of having certain pro-attitudes and of having corresponding means-end beliefs are among the causes of our actions. We suppose further that the agent did what he did because the having of the pro-attitude and belief were states with (respectively) a conative and a cognitive nature, and even more importantly, they are psychological states with certain propositional contents. The specific character of the causation of the action depended crucially on the fact that these psychological states had 'the direction of fit' and the propositional contents that they did. The agent *F*'ed at a given time, we think, because, at that time, he had a desire that represented *F*ing, and not some other act, as worthwhile or otherwise attractive to him.

Fred Dretske [1988] gave a famous example in this connection. When the soprano's singing of the aria shatters the glass, it will have been facts about the acoustic properties of the singing that were relevant to the breaking. The breaking does *not* depend upon the fact that she was singing lyrics and that those lyrics expressed such-and-such a content. We therefore expect that it will be the acoustic properties, and not the 'content' properties that figure in the pertinent explanatory laws. In the case of action, by contrast, we believe that the contents of the agent's attitudes *are* causally relevant to behavior. The contents of the agent's desires and beliefs not only help justify the action that is performed but, according to causalists at least, they play a causal role in determining the actions the agent was motivated to attempt. It has been difficult to see how Davidson, rejecting laws of mental content as he does, is in any position to accommodate the intuitive counterfactual dependence of action on the content of the agent's motivating reasons. His theory seems to offer no explication whatsoever of the fundamental role of mental content in reason explanations. Nevertheless, it should be admitted that no one really has a very good theory of how mental content plays its role. An enormous amount of research has been conducted to explicate what it is for propositional attitudes, realized as states of the nervous system, to express propositional contents at all. Without some better consensus on this enormous topic, we are not likely to get far on the question of mental causation, and solid progress on the attribution of content may still leave it murky how the contents of attitudes can be among the causal factors that produce behavior.

In a fairly early phase of the debate over the causal status of reasons for action, Norman Malcolm [1968] and Charles Taylor [1964] defended the thesis that ordinary reason explanations stand in potential rivalry with the explanations of human and animal behavior the neural sciences can be expected to provide. More recently, Jaegwon Kim [1989] has revived this issue in a more general way, seeing the two modes of explanation as joint instances of a Principle of Explanatory Exclusion. That Principle tells us that, if there exist two 'complete' and 'independent' explanations of the same event or phenomenon, then one or the other of these alternative explanations must be wrong. Influenced by Davidson, many philosophers reject more than just reason-to-action laws. They believe, more generally, that there are no laws that connect the reason-giving attitudes with *any* material states, events, and processes, under purely physical descriptions. As a consequence, commonsense psychology is not strictly reducible to the neural sciences, and this

means that reason explanations of action and corresponding neural explanations are, in the intended sense, 'independent' of one another. But, detailed causal explanations of behavior in terms of neural factors should also be, again in the intended sense, 'complete.' Hence, Explanatory Exclusion affirms that either the reason explanations or the prospective neural explanations must be abandoned as incorrect. Since we are not likely to renege upon our best, most worked-out scientific accounts, it is the ultimate viability of the reason explanations from commonsense 'vernacular' psychology that appear to be threatened. The issues here are complicated and controversial -- particularly issues about the proper understanding of 'theoretical reduction.' However, if Explanatory Exclusion applies to reason explanations of action, construed as causal, we have a very general incentive for searching for a workable philosophical account of reason explanations that construes them as non-causal. Just as certain function explanations in biology may not reduce to, but also certainly do not compete with, related causal explanations in molecular biology, so also non-causal reason explanations could be expected to co-exist with neural analyses of the causes of behavior.

In the foregoing, reference has been made to explanations of actions in terms of *reasons*, but recent work on agency has questioned whether contemporary frameworks for the philosophy of action have really articulated the way in which an agent's desires and other pro-attitudes have the distinctive force of reasons in the setting of these ordinary explanations [see Frankfurt 1988, 1999, Smith 1994]. Of course, it is widely recognized that reason explanations both tell us what motivated the agent's action and elucidate the justification that the action had, at least from the agent's own standpoint. However, the motivating role of 'reasons' can come to be separated from their role in providing an apparent justification. Compare the following two cases. In the first case, Smith hears some malicious gossip about the past career of Jones. Smith takes Jones to be a person with an absolutely impeccable character, and knows the rumor she has heard to be untrue. But, Smith's character is not so good. For a long time, she has felt a stifling envy of Jones, and, on this occasion, she has an irresistible, spiteful urge to repeat the defamatory false gossip and, thereby, damage Jones' exemplary reputation. Smith knows her desire for what it is -- a powerful but thoroughly unworthy urge to injure Jones. And she knows that it gives her no justifying reason, no justification whatsoever, for repeating the nasty gossip. In this case, however, Smith gives in to her jealous inclination, and passes the misinformation along. Now, when she tells the false story, Smith's behavior certainly does have a goal or purpose for her, and we can cite that goal or purpose in explaining why she acted as she did. But, as has been stipulated already, there is an important sense in which even Smith herself does not regard her desire as constituting any full-blooded ground or reason for what she does.

The contrasting case is much the same, but, in this case, Smith still has her envious impulses, but she is not subject to their control. Further, there is a central, new dimension to the character of Smith's practical reflections. She thinks that damaging Jones' unspotted reputation may do something to undermine the standing of a certain organization to which Jones belongs, and Smith seriously feels that there are grave political objections to this organization. She therefore believes that there would be real value in discrediting it. Smith may have her doubts about whether the envisaged end (discrediting the organization) justifies the choice of means (harming the innocent Smith). In this variant example, however, it seems that Smith can rightly think that her desire to hurt Jones' reputation constitutes a genuine reason -- constitutes at least some legitimate *prima facie* justification -- for injuring Jones' good

name. Now, most ordinary explanations of action in terms of reasons are more like the second case. The agent regards her potentially motivating pro-attitudes as providing intelligible grounds for action, and it is a key task of her practical reasoning to sort out the relative reason-giving force of the competing considerations.

The theory of action should be able to explain the differences between Smith's reasons and the import they have for her in the two contrasting cases. It should provide an account of why some teleological explanations of action are also explanations in terms of genuine normative reasons for acting and other purposive explanations are not. In this area, two fundamental questions seem to be linked closely with one another. What is it for a person or some other organism to be an agent of autonomous action? And, how do we explicate the special force of reasons for autonomous action in practical reasoning -- a 'force' quite different from the motivating impact of an overmastering desire? A good deal of attention has recently been devoted to these important problems and the associated issues they engender [see Korsgaard 1996, Bratman 1999, Velleman 2000, and Moran 2001]. Progress on such area questions may eventually reconfigure and improve the more established debates about how reason explanations do their work, including, one hopes, the venerable debate about whether reasons can be causes.

Bibliography

- Anscombe, G.E.M. (Elizabeth), 1963, *Intention*, 2nd. ed., Cornell University Press, Ithaca, NY
- Austin, J.L., 1962, *How to do Things with Words*, Harvard University Press, Cambridge, MA
- Austin, J.L., 1970, *Philosophical Essays*, J.O. Urmson and G.J. Warnock (ed.), Oxford University Press, Oxford
- Bishop, John, 1989, *Natural Agency*, Cambridge University Press, Cambridge
- Bratman, Michael, 1984, 'Two Faces of Intention', *Philosophical Review* 93: 375-405 [reprinted in Mele 1997]
- Bratman, Michael, 1987, *Intention,Plans, and Practical Reasoning*, Harvard University Press, Cambridge, MA
- Bratman, Michael, 1999, *Faces of Intention: Selected Essays on Intention and Agency*, Cambridge University Press, Cambridge
- Castañeda, Hector-Neri, 1975, *Thinking and Doing*, Reidel, Dordrecht
- Cleveland, Timothy, 1997, *Trying Without Willing*, Ashgate Publishing, Aldershot, England
- Davidson, Donald, 1980, *Essays on Actions and Events*, Oxford University Press, Oxford
- Dretske, Fred, 1988, *Explaining Behavior*, MIT Press, Cambridge, MA
- Falvey, Kevin, 2000, 'Knowledge in Intention', *Philosophical Studies*, 99: 21-44.
- Farrell, Dan, 1989, Intention, Reason, and Action, *American Philosophical Quarterly* 26: 283-95
- Fodor, Jerry, 1990, *A Theory of Content and Other Essays*, MIT Press, Cambridge, MA
- Frankfurt, Harry, 1978 'The Problem of Action', *American Philosophical Quarterly* 15: 157-62 [reprinted in Mele 1997]
- Frankfurt, Harry, 1988, *The Importance of What We Care About*, Cambridge University Press, Cambridge
- Frankfurt, Harry, 1999, *Volition, Necessity, and Love*, Cambridge, Cambridge University Press

- Ginet, Carl, 1990, *On Action*, Cambridge, Cambridge University Press
- Goldman, Alvin, 1970, *A Theory of Human Action*, Prentice-Hall, Englewood Cliffs, NJ
- Grice, H.P., 1971, 'Intention and Certainty', *Proceedings of the British Academy* 57: 263-79
- Harman, Gilbert, 'Practical Reasoning', *Review of Metaphysics* 79: 431-63 [reprinted in Mele 1997]
- Harman, Gilbert, 1986, *Change in View*, MIT Press, Cambridge, MA
- Higginbotham, James (ed.), 2000, *Speaking of Events*, Oxford University Press, New York
- Hornsby, Jennifer, 1980, *Actions*, Routledge & Kegan Paul, London
- Hornsby, Jennifer, 1997, *Simple-Mindedness: In Defense of Naïve Naturalism in the Philosophy of Mind*, Harvard University Press, Cambridge, MA
- Kim, Jaegwon, 1989, 'Mechanism, Purpose, and Explanatory Exclusion', *Philosophical Perspectives* 3: 77-108 [reprinted in Mele 1997]
- Korsgaard, Christine, 1996, *The Sources of Normativity*, Cambridge University Press, Cambridge
- Malcolm, Norman, 1968, 'The Conceivability of Mechanism', *Philosophical Review* 77: 45-72.
- McCann, Hugh, 1986, 'Rationality and the Range of Intention', *Midwest Studies in Philosophy* 10: 191-211.
- Mele, Alfred, 1992, *The Springs of Action*, Oxford University Press, New York
- Mele, Alfred (ed.), 1997, *The Philosophy of Action*, Oxford University Press, Oxford
- Millikan, Ruth, 1993, *White Queen Psychology and other Essays for Alice* , MIT Press, Cambridge, MA
- Moran, Richard, 2001, *Authority and Estrangement: An Essay on Self-Knowledge* Princeton University Press, Princeton
- O'Shaughnessy, Brian, 1973, 'Trying (as the Mental 'Pineal Gland') *Journal of Philosophy* 70: 365-86 [reprinted in Mele 1997]
- O'Shaughnessy, Brian, 1980, *The Will* (2 volumes), Cambridge University Press, Cambridge
- Parsons, Terence, 1990, *Events in the Semantics of English*, MIT Press, Cambridge, MA
- Pietroski, Paul, 2000, *Causing Actions*, Oxford University Press, New York
- Roth, Abraham, 'Reasons Explanation of Actions: Causal, Singular, and Situational', *Philosophy and Phenomenological Research* 59: 839-74
- Searle, John, 1983, *Intentionality*, Cambridge University Press, Cambridge
- Sehon, Scott, 1994, 'Teleology and the Nature of Mental States', *American Philosophical Quarterly*, 31: 63-72
- Sehon, Scott, 1998, 'Deviant Causal Chains and the Irreducibility of Teleological Explanation', *Pacific Philosophical Quarterly* 78: 195-213
- Sellars, Wilfrid, 1966, 'Thought and Action', in Keith Lehrer (ed.) *Freedom and Determinism*, Random House, New York
- Smith, Michael, 1987, 'The Humean Theory of Motivation', *Mind* 96: 36-61
- Smith, Michael, 1994, *The Moral Problem*, Blackwell, Oxford
- Stich, Stephen and Warfield (eds), Ted, 1994, *Mental Representation: a Reader*, Blackwell, Oxford
- Taylor, Charles, 1964, *The Explanation of Behavior*, Routledge & Kegan Paul, London
- Tuomela, R., 1977, *Human Action and its Explanation*, Reidel, Dordrecht
- Velleman, David, 1989, *Practical Reflection*, Princeton University Press, Princeton

- Velleman, David, 2000, *The Possibility of Practical Reason*, Oxford University Press, Oxford
- Vermazen, Bruce and Hintikka, Merrill (eds), 1985, *Essays on Davidson: Actions and Events*, MIT Press. Cambridge, MA
- von Wright, Georg, 1971, *Explanation and Understanding*, Cornell University Press, Ithaca, NY
- Wilson, George, 1989, *The Intentionality of Human Action*, Stanford University Press, Palo Alto, CA
- Wilson, George, 2000, 'Proximal Practical Foresight', *Philosophical Studies* 99: 3-19

Other Internet Resources

- Action Theory page (Andrei Buckareff, University of Rochester)
- Action Theory (Élisabeth Pacherie, Institut Jean-Nicod, CNRS)

Related Entries

Davidson, Donald | intention | reasons: justification vs. explanation | self-knowledge

George Wilson
gmwilson@ucdavis.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 17, 2002
Content last modified: April 2, 2002

Stanford Encyclopedia of Philosophy
Notes to Action

Notes

1. O'Shaughnessy 1973, p. 67. In O'Shaughnessy 1980, the author significantly modifies his position on this and related matters. This latter work constitutes the best, most extended investigation of fundamental metaphysical questions about action. Cleveland 1997 provides an instructive critical discussion.

2. The phrase originates with Anscombe 1963, but it has been picked up by various authors, not always with clear agreement about its use. For a causalist interpretation, treating the content of intentions as self-referential, see, for example, Searle 1983.

3. It is more usual in the literature to evaluate, following a suggestion of Davidson's, the purported equivalence of 'The agent intentionally *G*'d' and 'The agent *G*'d for some reason.' Whatever differences there might be between this proposal and the one in the text do not affect the present discussion.

4. Davidson gives the most explicit endorsement of (7**), or some minor variant thereof, on p. 221 of his "Reply to Vermazen" in Vermazen and Hintikka 1985.

5. In his paper, "Mental Events" [1980, essay 11], his denial that there are reason to action laws framed in the psychological vocabulary of ordinary discourse takes on special emphasis, and it plays a key role in an original argument he constructs for the token/token identity of mental and physical states.

George Wilson
gmwilson@ucdavis.edu

First published: March 17, 2002
Content last modified: March 17, 2002

Stanford Encyclopedia of Philosophy

Donald Davidson

Donald Davidson is one of the most important philosophers of the latter half of the twentieth century. His ideas, presented in a series of essays from the 1960's onwards, have been influential across a range of areas from semantic theory through to epistemology and ethics. Davidson's work exhibits a breadth of approach, as well as a unitary and systematic character, which is unusual within twentieth century analytic philosophy. Thus, although he acknowledges an important debt to W. V. O. Quine, Davidson's thought amalgamates influences (though these are not always explicit) from a variety of sources, including Quine, C. I. Lewis, Frank Ramsey, Immanuel Kant and the later Wittgenstein. And while often developed separately, Davidson's ideas nevertheless combine in such a way as to provide a single integrated approach to the problems of knowledge, action, language and mind. The breadth and unity of his thought, in combination with the sometimes-terse character of his prose, means that Davidson is not an easy writer to approach. Yet however demanding his work might sometimes appear, this in no way detracts from either the significance of that work or the influence it has exercised and will undoubtedly continue to exercise. Indeed, in the hands of Richard Rorty and others, and through the widespread translation of his writings, Davidson's ideas have reached an audience that extends far beyond the confines of English-speaking analytic philosophy. Of contemporary American philosophers, perhaps only Quine has had a similar reception and influence.

Biographical Sketch

Born on March 6th, 1917, in Springfield, Massachusetts, USA, Donald Herbert Davidson completed his undergraduate study at Harvard, graduating in 1939. His early interests were in literature and classics and, as an undergraduate, Davidson was strongly influenced by A. N. Whitehead. After starting graduate

work in classical philosophy (completing a Master's degree in 1941), Davidson's studies were interrupted by service with the US Navy in the Mediterranean from 1942-45. He continued work in classical philosophy after the war, graduating from Harvard in 1949 with a dissertation on Plato's 'Philebus'. By this time, however, the direction of Davidson's thinking had already, under Quine's influence, changed quite dramatically (the two having first met at Harvard in 1939-40) and he had begun to move away from the largely literary and historical concerns that had preoccupied him as an undergraduate towards a more strongly analytical approach.

While his first position was at Queen's College in New York, Davidson spent much of the early part of his career (1951-1967) at Stanford University. He has subsequently held positions at Princeton (1967-1970), Rockefeller (1970-1976), and the University of Chicago (1976-1981). Since 1981 he has taught at the University of California, Berkeley. Davidson has also been the recipient of a number of awards and fellowships and, in 1970, was the John Locke Lecturer at the University of Oxford.

Action and Mind

Reasons as Causes

Much of Davidson's early work was in decision theory (see *Decision-Making: An Experimental Approach*, with P. Suppes and S. Siegel [1957]), and it was not until the early 1960's that the work for which he is best known began to appear in print. Indeed, Davidson's first major philosophical publication was the seminal paper 'Actions, Reasons and Causes' (1963). In that paper Davidson sets out to defend the view that the explanation of action by reference to reasons (something we do, for instance, when we refer to an agent's intentions or motives in acting) is also a form of causal explanation. Indeed, he argues that reasons explain actions just inasmuch as they are the causes of those actions. This approach was in clear opposition to the Wittgensteinian orthodoxy of the time. On this latter account causal explanation was viewed as essentially a matter of showing the event to be explained as an instance of some law-like regularity (as we might explain the whistling of a kettle by reference to certain laws involving, among other things, the behaviour of gases under pressure). Since rational explanation was held, in general, not to involve any such reference to laws, but rather required showing how the action fitted into some larger pattern of rational behaviour, explanation by reference to reasons was held to be distinct from and independent of explanation by reference to causes.

Although directed against the Wittgensteinian-inspired view that reasons cannot be causes, Davidson's argument nevertheless effectively redeploys a number of Wittgensteinian notions. Two ideas play an especially significant role in the Davidsonian account -- ideas that are also, in one form or another,

important in Davidson's thinking elsewhere. The first of these ideas is the notion of a 'primary reason' -- the pairing of a belief and a desire (or 'pro-attitude') in the light of which an action is explained. Thus, my action of flipping the light switch can be explained by reference to my having the *belief* that flipping the switch turns on the light in combination with my having the *desire* to turn on the light (for most explanations explicit reference to both the belief and the desire is unnecessary). An action is thus rendered intelligible through being embedded in a broader system of attitudes attributable to the agent -- through being embedded, that is, in a broader framework of *rationality*. The second idea is that of action 'under a description' (a phrase originally appearing in G. E. M. Anscombe's *Intention* [1959]). As with the concept of a primary reason the idea here is simple enough: one and the same action is always amenable to more than one correct description. This idea is especially important, however, as it provides a means by which the same item of behaviour can be understood as intentional under some descriptions but not under others. Thus my action of flipping the light switch can be redescribed as the act of turning on the light (under which it is intentional) and also as the act of alerting the prowler who, unbeknown to me, is lurking in the bushes outside (under which it is unintentional). Generalising this point we can say that the same event can be referred to under quite disparate descriptions: the event of alerting the prowler is the same event as my flipping the light switch which is the same event as my moving of my body (or a part of my body) in a certain way.

Davidson treats the connection between reason and action (where the reason is indeed *the* reason for the action) as a connection that obtains between two events (the agent's believing and desiring on the one hand and her acting on the other) that can be variously described. The connection is both rational, inasmuch as the belief-desire pair (the 'primary reason') specifies the reason for the action, but it is also causal, inasmuch as the one event causes the other if it is indeed the reason for it. It is precisely because the reason is causally related to the action that the action can be explained by reference to the reason. Indeed, where an agent has a number of reasons for acting, and yet acts on the basis of one reason in particular, there is no way to pick out just that reason on which the agent acts other than by saying that it is the reason that *caused* her action.

Understood as rational the connection between reason and action cannot be described in terms of any strict law. Yet inasmuch as the connection is also a causal connection, so there must exist some law-like regularity, though not describable in the language of rationality, under which the events in question fall (an explanation can be causal, then, even though it does not specify any strict law). Davidson is thus able to maintain that rational explanation need not involve explicit reference to any law-like regularity, while nevertheless also holding that there must be some such regularity that underlies the rational connection just inasmuch as it is causal. Moreover, since Davidson resists the idea that rational explanations can be formulated in the terms of a predictive science, so he seems committed to denying that there can be any reduction of rational to non-rational explanation.

The Anomalism of the Mental

The more developed argument for this latter claim, and for the more general position in the philosophy of mind, of which it forms a part, appears at a number of places in Davidson's work. The first and best-

known presentation is that of ‘Mental Events’ (1970) in which Davidson argues for the compatibility of three principles (all three of which are adumbrated in various ways in the argument of ‘Actions, Reasons and Causes’): (i) that at least some mental events interact causally with physical events -- *’The Principle of Causal Interaction’*; (ii) that events related as cause and effect fall under strict laws (that is, laws that are ‘precise, explicit and as exceptionless as possible’) -- *’The Principle of the Nomological Character of Causality’*; and (iii) that there are no strict laws (as opposed to mere generalisations) relating mental and physical events -- *’The Anomalism of the Mental’*. Of these principles the first two would ordinarily be held to be incompatible with the third, and to imply, not the ‘anomalism’ of the mental, but rather, in the case of mental and physical events related as cause and effect, the existence of strict laws relating those events. To argue, as does Davidson, for the compatibility of the original principles is thus also to argue for the truth of the third, that is, for the truth of anomalism monism.

Davidson holds that events are particulars such that the same event can be referred to under more than one description. He also holds that events that are causally related must be related under some strict law. However, since Davidson takes laws to be linguistic entities, so they can relate events only as those events are given under specific descriptions. Thus, as was already evident in Davidson’s approach to the theory of action, the same pair of events may instantiate a law under one description, but not under others. There is, for example, no strict law that relates, under just those descriptions, the formation of ice on the surface of a road to the skidding of a car on that road, and yet, under a different description (a description that will employ a completely different set of concepts), the events at issue will indeed be covered by some strict law or set of laws. But while nomological relations between events (relations involving laws) depend on the descriptions under which the events are given, relations of causality and identity obtain irrespective of descriptions -- if the icing-up of the road did indeed cause the skid, then it did so no matter how the events at issue are described. (The form of description -- whether mental or physical -- is thus irrelevant to the fact that a particular causal relation obtains). It follows that the same pair of events may be related causally, and yet, under certain descriptions (though not under all), there be no strict law under which those events fall. In particular, it is possible that a mental event -- an event given under some mental description -- will be causally related to some physical event -- an event given under a physical description -- and yet there will be no strict law covering those events *under just those descriptions*. My wanting to read Tolstoy, for instance, leads me to take *War and Peace* from the shelf, and so my wanting causes a change in the physical arrangement of a certain region of space-time, but there is no strict law that relates my wanting to the physical change. Similarly, while any mental event will be identical with some physical event -- it will indeed be one and the same event under two descriptions -- it is possible that there will be no strict law relating the event as described in mentalistic terms with the event as physically described. In fact, Davidson is explicit in claiming that there can be *no* strict laws that relate the mental and the physical in this way -- there is no strict law that relates, for instance, wanting to read with a particular kind of brain activity.

Davidson’s denial of the existence of any strict ‘psycho-physical’ laws follows from his view of the mental as constrained by quite general principles of rationality that do not apply, at least not in the same way, to physical descriptions: normative considerations of overall consistency and coherence, for instance, constrain our own thinking about events as physically described, but they have no purchase on physical events as such. This does not mean, of course, that there are no correlations whatsoever to be

discerned between the mental and the physical, but it does mean that the correlations that can be discerned cannot be rendered in the precise, explicit and exceptionless form -- in the form, that is, of strict laws -- that would be required in order to achieve any reduction of mental to physical descriptions. The lack of strict laws covering events under mental descriptions is thus an insuperable barrier to any attempt to bring the mental within the framework of unified physical science. However, while the mental is not reducible to the physical, every mental event can be paired with some physical event -- that is, every mental description of an event can be paired with a physical description of the very same event. This leads Davidson to speak of the mental as 'supervening' on the physical in a way that implies a certain dependence of mental predicates on physical predicates: predicate *p* supervenes on a set of predicates *S* 'if and only if *p* does not distinguish any entities that cannot be distinguished by *S*' (see 'Thinking Causes' [1993]). Put more simply, events that cannot be distinguished under some physical description cannot be distinguished under a mental description either.

On the face of it, anomalous monism appears a highly attractive way to think about the relation between the mental and the physical - inasmuch as it combines 'monism' with 'anomalism' so it seems to preserve what is important about physicalism while nevertheless retaining the ordinary language of so-called 'folk-psychology' (the language of beliefs and desires, actions and reasons). In fact anomalous monism has proved to be a highly contentious position drawing criticism from both physicalists and non-physicalists alike. The nomological conception of causality (the second of the three principles defended in 'Mental Events') has often been seen as something for which Davidson fails to supply any real argument (a criticism he has attempted to address in 'Laws and Cause' [1995]); the Davidsonian account of supervenience has been viewed as incompatible with other aspects of his position and sometimes as simply mistaken or confused; and, perhaps the most serious and widespread criticism, anomalous monism has been seen as making the mental causally inert. These criticisms have not, however, gone unanswered (see especially 'Thinking Causes'), and while Davidson has modified aspects of his position, he has continued to hold to, and to defend, the basic theses first made explicit in 'Mental Events'.

Problems of Irrationality

Davidson's commitment to the rationality of the mental as one of the cornerstones of anomalous monism (as well as to the account of 'radical interpretation' [see 'Meaning and Truth' below]) leads him to take a special interest in the problem of apparently irrational belief and action -- something first addressed in 'How is Weakness of the Will Possible?'(1970). While Davidson treats irrationality as a real feature of our mental lives, he offers a way of dealing with it that aims at preserving, in some sense, the overall rationality of the mind (see especially 'Two Paradoxes of Irrationality' [1982]). A belief or desire in the mind of one person can cause a belief or desire in the mind of another without this compromising the rationality of the mental. (Davidson's example is my growing of a beautiful flower because I desire you to enter my garden -- you develop a craving to see the flower as a result of my desire and my desire has thereby caused, without being a reason for, your craving) Davidson suggests that we should view the same sort of relation as sometimes holding within a single mind. To this end we should view the mind as weakly 'partitioned' so that different attitudes may be located within different 'territories' and need not, therefore, be taken to come into direct conflict.

Ontology and Logical Form

Davidson's accounts of action and of mind call upon a well-developed set of analyses concerning psychological concepts such as belief, desire and intention -- concepts whose analysis is taken further in a number of papers that follow on from, and develop or modify, the ideas first set out in 'Actions, Reasons and Causes' (papers such as 'Agency' (1971) and 'Intending' [1978]) as well as in Davidson's discussions of epistemological and semantic issues (see below). But Davidson 's work in this area is also dependent on his account of the notions of cause, event and law and, in particular, on his defence of the view that events are particulars and so constitute a fundamental ontological category. If events are indeed particulars then an important question concerns the conditions of identity for events. In 'The Individuation of Events' [1969] Davidson argues that events are identical if and only if they have exactly the same causes and effects. In 'Reply to Quine on Events' [1985] he abandons this criterion in favour of the Quinean suggestion that events are identical if and only if they occupy exactly the same location in space and time.

A characteristic feature of Davidson's approach to such ontological questions has been to focus on the logical structure of sentences about the entities at issue rather than on those entities as such. Davidson's approach to events, for instance, is grounded in an analysis of the underlying logical form of sentences about events; in the case of causal relations, in an analysis of the logical form of sentences that express such relations (see 'Causal Relations' [1967]); and in his approach to action also, Davidson's approach involves an analysis of the logical form of sentences about actions (see 'The Logical Form of Action Sentences' [1967]). This reflects a more general commitment on Davidson's part to the inseparability of questions of ontology from questions of logic. This commitment is spelt out explicitly in 'The Method of Truth in Metaphysics' (1977) and it provides a further point of connection between Davidson's work in the philosophy of action, event and mind and his work on questions of meaning and language.

Meaning and Truth

The Structure of a Semantic Theory

Although Davidson has written on a wide range of topics, a great deal of his work, particularly during the late 1960s and early 1970s, has focussed on the problem of developing an approach to the theory of meaning that will be adequate to natural language. The characteristic feature of Davidson's approach to this problem is his proposal that meaning is best understood via the concept of truth, and, more particularly, that the basic structure for any adequate theory of meaning is that given in a formal theory

of truth.

Davidson's thinking about semantic theory is developed on the basis of a holistic conception of linguistic understanding (see 'Truth and Meaning' [1967]). Providing a theory of meaning for a language is thus a matter of developing a theory that will enable us to generate, for every actual and potential sentence of the language in question, a theorem that specifies what each sentence means. On this basis a theory of meaning for German that was given in English might be expected to generate theorems that would explicate the German sentence 'Schnee ist weiss' as meaning that snow is white. Since the number of potential sentences in any natural language is infinite, a theory of meaning for a language that is to be of use to creatures with finite powers such as ourselves, must be a theory that can generate an infinity of theorems (one for each sentence) on the basis of a finite set of axioms. Indeed, any language that is to be learnable by creatures such as ourselves must possess a structure that is amenable to such an approach. Consequently, the commitment to *holism* also entails a commitment to a *compositional* approach according to which the meanings of sentences are seen to depend upon the meanings of their parts, that is, upon the meanings of the words that form the finite base of the language and out of which sentences are composed. Compositionality does not compromise holism, since not only does it follow from it, but, on the Davidsonian approach, it is only as they play a role in whole sentences that individual words can be viewed as meaningful. It is sentences, and not words, that are thus the primary focus for a Davidsonian theory of meaning. Developing a theory for a language is a matter of developing a systematic account of the finite structure of the language that enables the user of the theory to understand any and every sentence of the language.

A Davidsonian theory of meaning explicates the meanings of expressions holistically through the interconnection that obtains among expressions within the structure of the language as a whole. Consequently, although it is indeed a theory *of* meaning, a theory of the sort Davidson proposes will have no use for a concept of meaning understood as some discrete entity (whether a determinate mental state or an abstract 'idea') to which meaningful expressions refer. One important implication of this is that the theorems that are generated by such a theory of meaning cannot be understood as theorems that relate expressions and 'meanings'. Instead such theorems will relate sentences to other sentences. More particularly, they will relate sentences in the language to which the theory applies (the 'object-language') to sentences in the language in which the theory of meaning is itself couched (the 'meta-language') in such a way that the latter effectively 'give the meanings of' or translate the former. It might be thought that the way to arrive at theorems of this sort is to take as the general form of such theorems '*s* means that *p*' where *s* names an object-language sentence and *p* is a sentence in the meta-language. But this would be already to assume that we could give a formal account of the connecting phrase 'means that', and not only does this seem unlikely, but it also appears to assume a concept of meaning when it is precisely that concept (at least as it applies within a particular language) that the theory aims to elucidate. It is at this point that Davidson turns to the concept of truth. Truth, he argues, is a less opaque concept than that of meaning. Moreover, to specify the conditions under which a sentence is true is also a way of specifying the meaning of a sentence. Thus, instead of '*s* means that *p*', Davidson proposes, as the model for theorems of an adequate theory of meaning, '*s* is true if and only if *p*' (the use of the biconditional 'if and only if' is crucial here as it ensures the truth-functional equivalence of the sentences *s* and *p*, that is, it ensures they will have identical truth-values). The theorems of a Davidsonian theory of meaning for

German couched in English would thus take the form of sentences such as "'Schnee ist weiss' is true if and only if snow is white."

Tarski and 'Convention T'

One of the great advantages of this proposal is that it enables Davidson to connect his account of a theory of meaning with an already existing approach to the theory of truth, namely that developed by Alfred Tarski (see Tarski, 'The Concept of Truth in Formalised Languages', first published in German in 1936, reprinted in Tarski, *Logic, Semantics and Metamathematics* [1956]). Tarski's theory of truth was originally intended, not as a general account of the nature of truth, but rather as a way of defining the truth-predicate as it applies within a formal language. Tarski suggests that we arrive at a formal definition of the predicate 'is true' by providing, for every sentence *s* in the object language, a matching sentence *p* in the meta-language that is a translation of *s*. The resulting 'T-sentences' will have the form '*s* is true (in language *L*) if and only if *p*'. That an adequate theory should indeed be capable of generating a T-sentence for every sentence in the object-language is the essence of Tarski's 'Convention T' -- a requirement that clearly matches the holistic requirement Davidson also specifies for an adequate theory of meaning. And just as a Davidsonian theory of meaning treats the meaning of whole sentences as dependent on the components of those sentences, so a Tarskian theory of truth defines truth *recursively* in that it treats the truth of complex expressions as depending on the truth of more primitive expressions. In the case of those primitive expressions -- such as predicates and other terms -- to which the concept of truth does not attach, Tarski makes use of a technical notion of *satisfaction* which stands to predicates and other primitive expressions as truth stands to whole sentences: primitive expressions are understood as being satisfied or not satisfied by certain sequences of objects (just as sentences are true or false according to the obtaining or not of certain conditions). On the Tarskian account truth-conditions turn out to be definable in terms of satisfaction.

The formal structure that Tarski articulates here is identical to that which Davidson explicates as the basis for a theory of meaning: a Tarskian truth theory can generate, for every sentence of the object-language, a T-sentence that specifies the meaning of each sentence in the sense of specifying the conditions under which it is true. What Davidson's work shows, then, is that satisfaction of Tarski's Convention T can be seen as the basic requirement for an adequate theory of meaning.

A Tarskian truth theory defines truth on the basis of logical resources that are no more than those available within first-order quantificational logic. Moreover, it also defines truth 'extensionally', that is, it defines truth in terms of the objects that satisfy expressions -- in terms, we might say, of the objects that fall under those expressions -- rather than in terms of meanings, descriptions or other 'intensional' entities. Both these features represent important advantages for a Davidsonian approach (Davidson's rejection of determinate meanings as having a significant role to play in a theory of meaning already involves a commitment to an extensional approach to language). However, these features also present certain problems. Davidson wishes to apply the Tarskian model as the basis for a theory of meaning for natural languages, but such languages are far richer than the well-defined formal systems for which Tarski originally developed his approach. In particular natural languages contain features that seem to

require resources beyond those of first-order logic or of any purely extensional analysis. Examples of such features include indirect or reported speech ('Galileo said that the earth moves'), adverbial expressions ('Flora swam slowly ' where 'slowly' modifies 'Flora swam') and non-indicative sentences such as imperatives ('Eat your eggplant!'). An important part of Davidson's work in the philosophy of language has been to show how such apparently recalcitrant features of natural language can indeed be analysed so as to make them amenable to a Tarskian treatment. In 'On Saying That' (1968) and 'Quotation' (1979) he addresses the question of indirect speech; in 'Moods and Performances' (1979) he deals with non-indicative utterances; and in 'Adverbs of Action' (1985) he takes up the problem of adverbial modification. As in Davidson's analysis of actions and events, the notion of logical form plays an important part in his approach here -- the problem of how to apply a Tarskian truth theory to natural language is shown to depend on providing an analysis of the underlying logical form of natural language expressions which renders them in such a way that they fall under the scope of a purely extensional approach employing only the resources of first-order quantificational logic.

There is, however, another more general problem that affects Davidson's appropriation of Tarski. While Tarski uses the notion of sameness of meaning, through the notion of translation, as the means to provide a definition of truth -- one of the requirements of Convention T is that the sentence on the right hand side of a Tarskian T-sentence be a translation of the sentence on the left - Davidson aims to use truth to provide an account of meaning. But in that case it seems that he needs some other way to constrain the formation of T-sentences so as to ensure that they do indeed deliver correct specifications of what sentences mean. This problem is readily illustrated by the question of how we are we to rule out T-sentences of the form "'Schnee ist weiss' is true if and only if grass is green." Since the biconditional 'if and only if' ensures only that the sentence named on the left will have the same truth value as the sentence on the right, so it would seem to allow us to make any substitution of sentences on the right so long as their truth value is identical to that on the left. In one respect this problem is met by simply insisting on the way in which T-sentences must be seen as theorems generated by a theory of meaning that is adequate to the language in question as a whole (see 'Truth and Meaning'). Since the meaning of particular expressions will not be independent of the meaning of other expressions (in virtue of the commitment to compositionality the meanings of all sentences must be generated on the same finite base), so a theory that generates problematic results in respect of one expression can be expected to generate problematic results elsewhere, and, in particular, to also generate results that do not meet the requirements of Convention T. This problem can also be seen, however, as closely related to another important point of difference between a Tarskian truth theory and a Davidsonian theory of meaning: a theory of meaning for a natural language must be an empirical theory -- it is, indeed, a theory that ought to apply to actual linguistic behaviour -- and as such it ought to be empirically verifiable. Satisfaction of the requirement that a theory of meaning be adequate as an empirical theory, and so that it be adequate to the actual behaviour of speakers, will also ensure tighter constraints (if such are needed) on the formation of T-sentences. Indeed, Davidson is not only quite explicit in emphasising the empirical character of a theory of meaning, but he also offers a detailed account that both explains how such a theory might be developed and specifies the nature of the evidence on which it must be based.

Radical Interpretation

Davidson’s strategy is to embed the formal structure for a theory of meaning (the structure he finds in a Tarskian truth theory) within a more general theory of interpretation the broad outlines of which he draws from Quine (see Quine, *Word and Object* [1960]). ‘Radical translation’ is intended by Quine as an idealisation of the project of translation that will exhibit that project in its purest form. Normally the task of the translator is aided by prior linguistic knowledge -- either of the actual language to be translated or of some related language. Quine envisages a case in which translation of a language must proceed without any prior linguistic knowledge and solely on the basis of the observed behaviour of the speakers of the language in conjunction with observation of the basic perceptual stimulations that give rise to that behaviour. Davidson has a broader conception of the behavioural evidence available than does Quine (he allows that we may, for instance, identify speakers as having the attitude of ‘holding true’ with respect to sentences) and, in addition, rejects the Quinean insistence on a special role being given to simple perceptual stimulations. Moreover, since Davidson’s interest is more properly semantic than Quine’s (Quine sees radical translation as part of a primarily epistemological inquiry), while Davidson also views a theory of translation alone as insufficient to ensure understanding of the language it translates (the translation may be into a language we do not understand), so the notion of ‘translation’ is replaced in the Davidsonian account with that of ‘interpretation’. *Radical interpretation* is a matter of interpreting the linguistic behaviour of a speaker ‘from scratch’ and so without reliance on any prior knowledge either of the speaker’s beliefs or the meanings of the speaker’s utterances. It is intended to lay bare the knowledge that is required if linguistic understanding is to be possible, but it involves no claims about the possible instantiation of that knowledge in the minds of interpreters (Davidson thus makes no commitments about the underlying psychological reality of the knowledge that a theory of interpretation makes explicit).

The basic problem that radical interpretation must address is that one cannot assign meanings to a speaker’s utterances without knowing what the speaker believes, while one cannot identify beliefs without knowing what the speaker’s utterances mean. It seems that we must provide both a theory of belief and a theory of meaning at one and the same time. Davidson claims that the way to achieve this is through the application of the so-called ‘principle of charity’ (Davidson has also referred to it as the principle of ‘rational accommodation’) a version of which is also to be found in Quine. In Davidson’s work this principle, which admits of various formulations and cannot be rendered in any completely precise form, often appears in terms of the injunction to optimise agreement between ourselves and those we interpret, that is, it counsels us to interpret speakers as holding true beliefs (true by our lights at least) wherever it is plausible to do (see ‘Radical Interpretation’ [1973]). In fact the principle can be seen as combining two notions: a holistic assumption of rationality in belief (‘coherence’) and an assumption of causal relatedness between beliefs -- especially perceptual beliefs -- and the objects of belief (‘correspondence’) (see ‘Three Varieties of Knowledge [1991]). The process of interpretation turns out to depend on both aspects of the principle. Attributions of belief and assignments of meaning must be consistent with one another and with the speaker’s overall behaviour; they must also be consistent with the evidence afforded by our knowledge of the speaker’s environment, since it is the worldly causes of beliefs that must, in the ‘most basic cases’, be taken to be the objects of belief (see ‘A Coherence Theory of Truth and Knowledge’ [1983]). Inasmuch as charity is taken to generate particular attributions of belief, so those attributions are, of course, always defeasible. The principle itself is not so, however, since it remains, on the Davidsonian account, a presupposition of any interpretation whatsoever. Charity is, in this respect, both a constraint and an enabling principle in all interpretation -- it is more than just a

heuristic device to be employed in the opening stages of interpretative engagement.

If we assume that the speaker's beliefs, at least in the simplest and most basic cases, are largely in agreement with our own, and so, by our account, are largely true, then we can use our own beliefs about the world as a guide to the speaker's beliefs. And, provided that we can identify simple assertoric utterances on the part of a speaker (that is, provided we can identify the attitude of holding true), then the interconnection between belief and meaning enables us to use our *beliefs* as a guide to the *meanings* of the speaker's utterances -- we get the basis for both a rudimentary theory of belief and a rudimentary account of meaning. So, for example, when the speaker with whom we are engaged uses a certain sequence of sounds repeatedly in the presence of what we believe to be a rabbit, we can, as a preliminary hypothesis, interpret those sounds as utterances about rabbits or about some particular rabbit. Once we have arrived at a preliminary assignment of meanings for a significant body of utterances, we can test our assignments against further linguistic behaviour on the part of the speaker, modifying those assignments in accordance with the results. Using our developing theory of meaning we are then able to test the initial attributions of belief that were generated through the application of charity, and, where necessary, modify those attributions also. This enables us, in turn, to further adjust our assignments of meaning, which enables further adjustment in the attribution of beliefs, ... and so the process continues until some sort of equilibrium is reached. The development of a more finely tuned theory of belief thus allows us to better adjust our theory of meaning, while the adjustment of our theory of meaning in turn enables us to better tune our theory of belief. Through balancing attributions of belief against assignments of meaning, we are able to move towards an overall theory of behaviour for a speaker or speakers that combines both a theory of meaning and of belief within a single theory of interpretation.

Holism and Indeterminacy

Since it is indeed a single, combined theory that is the aim here, so the adequacy of any such theory must be measured in terms of the extent to which the theory does indeed provide a unified view of the totality of behavioural evidence available to us (taken in conjunction with our own beliefs about the world) rather than by reference to any single item of behaviour. This can be viewed as a more general version of the same requirement, made in relation to a formal theory of meaning, that a theory of meaning for a language address the totality of utterances for that language, although, in the context of radical interpretation, this requirement must be understood as also closely tied to the the need to attend to normative considerations of overall rationality. A direct consequence of this holistic approach is that there will always be more than one theory of interpretation that will be adequate to any particular body of evidence since theories may differ in particular attributions of belief or assignments of meaning while nevertheless providing an equally satisfactory account of the speaker's overall behaviour. It is this failure of uniqueness that Davidson terms the 'indeterminacy' of interpretation and which provides a counterpart to the 'indeterminacy of translation' that also appears, though it has a more limited application, in Quine. On the Davidsonian account, while such indeterminacy often goes unnoticed and is indeed rather less for Davidson than than for Quine (partly as a consequence of Davidson's employment of Tarski and so of the need to read the structure of first-order logic into the language interpreted), it nevertheless remains an ineliminable feature of all interpretation. Moreover, indeterminacy is not to be viewed merely as

reflecting some epistemological limitation on interpretation, but rather reflects the holistic character of meaning and of belief. Such concepts refer us to overall patterns in the behaviour of speakers rather than to discrete, entities to which interpretation must somehow gain access. Indeed, holism of this sort applies, not only to meanings and beliefs, but also to the so-called 'propositional attitudes' in general. The latter are most simply characterised as attitudes specifiable by reference to a proposition (believing that there is eggplant for dinner is a matter of holding true the proposition that there is eggplant for dinner; desiring that there be eggplant for dinner is a matter of wanting it to be true that there be eggplant for dinner) and so the *contents* of attitudes of this sort are always *propositional*. Davidsonian holism is thus a holism that applies to meanings, to attitudes, and also, thereby, to the content of attitudes. Indeed, we can speak of the Davidsonian account of interpretation as providing a quite general account of how mental content is determined (such content being understood as the content of propositional mental states such as belief): through the causal relation between speakers and objects in the world and through the rational integration of speakers' behaviour. Thus, as Davidson's approach to the theory of meaning turns out to imply a more general theory of interpretation, so his holistic view of meaning implies a holistic view of the mental, and of mental content, in general.

Davidson's commitment to the indeterminacy that follows from his holistic approach has lead some to view his position as involving a form of anti-realism about the mind and about beliefs, desires and so forth. Davidson argues, however, that the indeterminacy of interpretation should be understood analogously with the indeterminacy that attaches to measurement. Such theories assign numerical values to objects on the basis of empirically observable phenomena and in accordance with certain formal theoretical constraints. Where there exist different theories that address the same phenomena, each theory may assign different numerical values to the objects at issue (as do Celsius and Fahrenheit in the measurement of temperature), and yet there need be no difference in the empirical adequacy of those theories, since what is significant is the overall pattern of assignments rather than the value assigned in any particular case. Similarly in interpretation, it is the overall pattern that a theory finds in behaviour that is significant and that remains invariant between different, but equally adequate, theories. An account of meaning for a language is an account of just this pattern.

Although the indeterminacy thesis has sometimes been a focus for objections to Davidson's approach, it is the more basic thesis of holism as developed in its full-blown form in the account of radical interpretation (and particularly as it relates to meaning) that has often attracted the most direct and trenchant criticism. Michael Dummett has been one of the most important critics of the Davidsonian position (see especially 'What is a Theory of Meaning', in S. Guttenplan [ed.], *Mind and Language* [1975]). Dummett argues that Davidson's commitment to holism not only gives rise to problems concerning, for instance, how a language can be learnt (since it seems to require that one come to understand the whole of the language at one go, whereas learning is always piecemeal), but that it also restricts Davidson from being able to give what Dummett views as a properly full-blooded account of the nature of linguistic understanding (since it means that Davidson cannot provide an account that explicates the semantic in terms of the non-semantic). More recent criticisms have come from Jerry Fodor, amongst others, whose opposition to holism (not only in Davidson, but in Quine, Dennett and elsewhere) is largely motivated by a desire to defend the possibility of a certain scientific approach to the mind (see especially, Jerry Fodor and Ernest LePore, *Holism: A Shopper's Guide* [1992]).

Language and Convention

The heart of a Davidsonian theory of interpretation is, of course, a Tarskian truth theory. But a truth theory provides only the formal structure on which linguistic interpretation is based: such a theory needs to be embedded within a broader approach that looks to the interconnections between utterances, other behaviour and attitudes; in addition, the application of such a theory to actual linguistic behaviour must also take account of the dynamic and shifting character of such behaviour. This latter point is easily overlooked, but it leads Davidson to some important conclusions. Ordinary speech is full of ungrammatical constructions (constructions that may even be acknowledged to be ungrammatical by the speaker herself), incomplete sentences or phrases, metaphors, neologisms, jokes, puns and all manner of phenomena that cannot be met simply by the application to utterances of a pre-existing theory for the language being spoken. Linguistic understanding cannot, then, be a matter simply of the mechanical application of a Tarski-like theory (although this is just what Davidson might be taken to suggest in the early essays). In papers such as ‘A Nice Derangement of Epitaphs’ (1986), Davidson addresses just this point, arguing that while linguistic understanding does indeed depend upon a grasp of the formal structure of a language, that structure always stands in need of modification in the light of actual linguistic behaviour. Understanding a language is a matter of continually adjusting interpretative presuppositions (presuppositions that are often not explicit) in accord with the utterances to be interpreted. Furthermore, this calls upon skills and knowledge (imagination, attentiveness to the attitudes and behaviour of others, knowledge of the world) that are not specifically linguistic and that are part of a more general ability to get on in the world and in relation to others -- an ability that also resists any formal explication. In ‘A Nice Derangment of Epitaphs’, Davidson puts this point, in provocative fashion, by claiming that ‘there is no such thing as a language’ (adding the immediate qualification ‘not if a language is anything like what many philosophers and linguists have supposed’). Put less provocatively, the essential point is that linguistic conventions (and in particular linguistic conventions that take the form of agreement over the employment of shared syntactic and semantic rules), while they may well facilitate understanding, cannot be the basis for such understanding.

Davidson’s denial of rule-based conventions as having a founding role in linguistic understanding, together with his emphasis on the way in which the capacity for linguistic understanding must be seen as part as part of a more general set of capacities for getting on in the world, underlie Davidson’s much-discussed account of metaphor and related features of language (see ‘What Metaphors Mean’ [1978]). Davidson rejects the idea that metaphorical language can be explained by reference to any set of rules that govern such meaning. Instead it depends on using sentences with their ‘literal’ or standard meanings in ways that give rise to new or unexpected insights -- and just as there are no rules by which we can work out what a speaker means when she utters an ungrammatical sentence, makes a pun or otherwise uses language in a way that diverges from the norm, so there are no rules that govern the grasp of metaphor.

Knowledge and Belief

‘Three Varieties of Knowledge’

In Davidson’s work the question ‘what is meaning?’ is replaced by the question ‘What would a speaker need to know to understand the utterances of another?’ The result is an account that treats the theory of meaning as necessarily part of a much broader theory of interpretation and, indeed, of a much broader approach to the mental as such. This account is holistic inasmuch as it requires that any adequate theory must address linguistic and non-linguistic behaviour in its entirety. As we have already seen, this means that a theory of interpretation must adopt a compositional approach to the analysis of meaning; it must recognise the interconnected character of attitudes and of attitudes and behaviour; and it must also attribute attitudes and interpret behaviour in a way constrained by normative principles of rationality. Rationality is not, however, the only principle on which Davidson’s account of radical interpretation depends. It involves, in fact, a marriage of both holistic and ‘externalist’ considerations: considerations concerning the dependence of attitudinal content on the rational connections between attitudes (‘holism’) and concerning the dependence of such content on the causal connections between attitudes and objects in the world (‘externalism’). Indeed, this marriage is evident, as we saw earlier, in the principle of charity itself and its combination of considerations of both ‘coherence’ and ‘correspondence’. Davidson holds, in fact, that attitudes can be attributed, and so attitudinal content determined, only on the basis of a triangular structure that requires interaction between at least two creatures as well as interaction between each creature and a set of common objects in the world.

Identifying the content of attitudes is a matter of identifying the objects of those attitudes, and, in the most basic cases, the objects of attitudes are identical with the causes of those same attitudes (as the cause of my belief that there is a bird outside my window is the bird outside my window). Identifying beliefs involves a process analogous to that of ‘triangulation’ whereby the position of an object is determined by taking a line from each of two already known locations to the object in question -- the intersection of the lines fixes the position of the object (this idea first appears in ‘Rational Animals [1982]). Similarly, the objects of propositional attitudes are fixed by looking to find objects that are the common causes, and so the common objects, of the attitudes of two or more speakers who are capable of observing and responding to one another’s behaviour. In ‘Three Varieties of Knowledge’, Davidson develops the metaphor of triangulation into the idea of a three-way conceptual interdependence between knowledge of oneself, knowledge of others and knowledge of the world. Just as knowledge of language cannot be separated from our more general knowledge of the world, so Davidson argues that knowledge of oneself, knowledge of other persons and knowledge of a common, ‘objective’ world form an interdependent set of concepts no one of which is possible in the absence of the others.

Against Relativism and Scepticism

The inseparability of these ‘varieties’ of knowledge has a number of important implications. Since our knowledge of our own minds is not independent of our knowledge of the world nor of our knowledge of others, so we cannot treat self-knowledge as a matter of our having access to some set of private ‘mental’ objects. Our knowledge of ourselves arises only in relation to our involvement with others and with respect to a publically accessible world. Even so, we retain a certain authority over our own attitudes and utterances simply in virtue of the fact that those attitudes and utterances are indeed our own (see ‘First-Person Authority’, [1984]). Since knowledge of the world is inseparable from other forms of knowledge, so global epistemological scepticism -- the view that all or most of our beliefs about the world could be false -- turns out to be committed to much more than is usually supposed. Should it indeed turn out that our beliefs about the world were all, or for the most part, false, then this would not only imply the falsity of most of our beliefs about others, but it would also have the peculiar consequence of making false most of our beliefs about ourselves -- including the supposition that we do indeed hold those particular false beliefs. Although this may fall short of demonstrating the falsity of such scepticism, it surely demonstrates it to be deeply problematic.

The way in which the Davidsonian rejection of scepticism does indeed derive quite directly from Davidson’s adoption of a holistic, externalist approach to knowledge, and to attitudinal content in general, has sometimes been obscured by Davidson’s presentation of his argument against scepticism through the employment (for the first time in ‘Thought and Talk’) of the rather problematic notion of an ‘omniscient interpreter’. Such an interpreter would attribute beliefs to others and assign meanings to their utterances, but would nevertheless do so on the basis of his own, true, beliefs. The omniscient interpreter would therefore have to find a large amount of agreement between his own beliefs and the beliefs of those he interprets -- and what was agreed would also, by hypothesis, be true. This argument disappears from Davidson’s later discussions in which the idea of triangulation comes much more to the fore.

A feature of both the triangulation argument, and the Davidsonian account of radical interpretation, is that the attribution of attitudes must always proceed in tandem with the interpretation of utterances -- identifying content, whether of utterances or of attitudes, is indeed a single project. An inability to interpret utterances (that is, an inability to assign meanings to instances of putative linguistic behaviour) will thereby imply an inability to attribute attitudes (and vice versa). A creature that we cannot interpret as capable of meaningful speech will thus also be a creature that we cannot interpret as capable of possessing contentful attitudes. Such considerations lead Davidson to deny that non-linguistic animals are capable of thought -- where thought involves the possession of propositional attitudes such as beliefs or desires (see especially ‘Thought and Talk’ [1975]). This does not mean that such animals have no mental life at all, nor does it mean that we cannot usefully use mental concepts in explaining and predicting the behaviour of such creatures. What it does mean, however, is that the extent to which we can think of such creatures as having attitudes and a mental life like our own is measured by the extent to which we can assign determinate propositional content to the attitudes we would ascribe to those creatures. A further consequence of this view is that the idea of an untranslatable language -- an idea often found in association with the thesis of conceptual relativism -- cannot be given any coherent formulation. Inability to translate counts as evidence, not of the existence of an untranslatable language, but of the absence of a language of any sort (see ‘On the Very Idea of a Conceptual Scheme’ [1974]).

The 'Third Dogma' of Empiricism

Davidson's rejection of the idea of an untranslatable language (and the associated idea, also common to many forms of conceptual relativism, of a radically different, and so 'incommensurable' system of belief) is part of a more general argument that he advances (notably in 'On the Very Idea of a Conceptual Scheme') against the so-called 'third dogma' of empiricism. The first two dogmas are those famously identified by Quine in 'Two Dogmas of Empiricism' (in *From a Logical Point of View* [1953]). The first is that of reductionism (the idea that, for any meaningful statement, it can be recast in the language of pure sensory experience, or, at least, in terms of a set of confirmatory instances), while the second is the analytic-synthetic distinction (the idea that, with respect to all meaningful statements, one can distinguish between statements that are true in virtue of their meaning and those that are true in virtue of both their meanings and some fact or facts about the world). The rejection of both these dogmas can be seen as an important element throughout Davidson's thinking. The third dogma, which Davidson claims can still be discerned in Quine's work (and so can survive the rejection even of the analytic-synthetic distinction), consists in the idea that one can distinguish within knowledge or experience between a conceptual component (the 'conceptual scheme') and an empirical component (the 'empirical content') -- the former is often taken to derive from language and the later from experience, nature or some form of 'sensory input'. While there are difficulties in even arriving at a clear formulation of this distinction (particularly so far as the nature of the relation between the two components is concerned), such a distinction depends on being able to distinguish, at some basic level, between a 'subjective' contribution to knowledge that comes from ourselves and an 'objective' contribution that comes from the world. What the Davidsonian account of knowledge and interpretation demonstrates, however, is that no such distinction can be drawn. Attitudes are already interconnected -- causally, semantically and epistemically -- with objects and events in the world; while knowledge of self and others already presupposes knowledge of the world. The very idea of a conceptual scheme is thus rejected by Davidson along with the idea of any strong form of conceptual relativism. To possess attitudes and be capable of speech is already to be capable of interpreting others and to be open to interpretation by them.

Realism, Anti-Realism and Theories of Truth

Davidson emphasises the holistic character of the mental (both in terms of the interdependence that obtains between various forms of knowledge as well as the interconnected character of attitudes and of attitudes and behaviour). He has, at times, also referred to his position as involving a 'coherence' theory of truth and of knowledge (in 'A Coherence Theory of Truth and Knowledge' [1983]). Nevertheless, Davidson is not a coherentist, in any standard sense, about either truth or knowledge. Nor, for all that he adopts a Tarskian approach to meaning, does he espouse a correspondence theory of truth. Davidson eschews any attempt to provide an account of the nature of truth, maintaining that truth is an absolutely central concept that cannot be reduced to or replaced by any other notion (see 'The Structure and Content of Truth' [1990]). His employment of the notion of coherence is best seen as reflecting his commitment to the fundamentally rational and holistic character of the mind. It can also be seen to be tied to Davidson's rejection of those forms of epistemological foundationalism that would attempt to ground knowledge or belief in the sensory causes of belief -- beliefs, as one might expect given Davidson's

holistic approach, can find evidential support only in other beliefs. Similarly, Davidson's sometime employment of the notion of correspondence is best understood, not as providing, any direct elucidation of the nature of truth, but rather as deriving from his externalist commitment to the idea that the content of belief is dependent upon the worldly causes of belief. In 'True to the Facts' (1969) Davidson does defend what he there presents as a form of correspondence theory of truth. However, not only has Davidson since relinquished the claim that his is a 'correspondence' view of truth (see 'The Structure and Content of Truth'), but the account set out in 'True to the Facts' is, in any case, far removed from what is usually taken to be involved in any correspondence theory.

Since Davidson rejects both sceptical and relativist positions, while nevertheless insisting of the indispensability of an irreducibly basic concept of objective truth, Davidson cannot be easily situated with respect to the realist/anti-realist controversy that, until quite recently, was a major concern of many anglo-american philosophers. The Davidsonian position has, nevertheless, been variously assimilated, at different times and by different critics, to both the realist and the anti-realist camp. Yet realism and anti-realism are equally unsatisfactory from a Davidsonian point of view, since neither is compatible with the holistic and externalist character of knowledge and belief. Realism makes truth inaccessible (inasmuch as it admits the sceptical possibility that even our best-confirmed theories about the world could all be false), while anti-realism makes truth too epistemic (inasmuch as it rejects the idea of truth as objective). In this respect, and as he himself makes clear (see 'The Structure and Content of Truth'), Davidson does not merely reject the specific premises that underlie the realist and anti-realist positions, but views the very dispute between them as essentially misconceived. This reflects a characteristic feature of Davidson's thinking in general (and not just as it relates to realism and anti-realism), namely its resistance to any simple classification using the standard philosophical categories of the day.

Bibliography

[Note: The dates given in brackets after the titles of publications mentioned in the text above are the dates of first publication. For the most part articles by Davidson that appear in the collections listed below were previously published elsewhere. As a consequence, where those articles are mentioned in the text, they carry a date of publication earlier than that of the collection in which they also appear. In addition to the three volumes of collected papers listed here, a further two volumes, titled *Problems of Rationality* and *Truth*, *Lamguage and History*, are scheduled to appear in the near future]

- Davidson, Donald, *Essays on Actions and Events* (Oxford: Clarendon Press, 2nd edn, 2002), includes (amongst others): 'Actions, Reasons and Causes', 'Agency', 'Intending', 'The Logical Form of Action Sentences', 'Causal Relations', 'The Individuation of Events', 'Mental Events', 'Adverbs of Action'.
- Davidson Donald, *Inquiries into Truth and Interpretation*(Oxford: Clarendon Press, 2nd edn, 2001), includes (amongst others): 'Truth and Meaning', 'Quotation', 'True to the Facts', 'On Saying That', 'Radical Interpretation', 'Thought and Talk', 'On the Very Idea of a Conceptual Scheme'.
- Davidson, Donald, *Subjective, Intersubjective, Objective*(Oxford: Clarendon Press, 2001),

includes (amongst others): ‘First Person Authority’, ‘Rational Animnals’, ‘A Coherence Theory of Truth and Knowledge’, ‘Three Varieties of Knowledge’.
- Davidson, Donald, ‘Two Paradoxes of Irrationality’, in R. Wollheim and J. Hopkins (eds.) *Philosophical Essays on Freud* (Cambridge: Cambridge University Press, 1982): 289-305.
- Davidson, Donald, ‘The Structure and Content of Truth’ (The Dewey Lectures 1989), *Journal of Philosophy* 87 (1990): 279-328.
- Davidson, Donald, ‘Thinking Causes’ in John Heil and Alfred Mele (eds.), *Mental Causation* (Oxford: Clarendon Press, 1993): 3-18.
- Davidson, Donald, ‘Laws and Causes,’ *Dialectica* 49 (1995): 263-278.
- Evnine, Simon, *Donald Davidson* (Cambridge: Polity Press, 1991).
- Hahn, Lewis Edwin (ed.), *The Philosophy of Donald Davidson, Library of Living Philosophers XXVII*(Chicago: Open Court, 1999) - contains an extensive bibliography, compiled by Davidson himself, of primary and secondary material, plus an autobiographical essay.
- Kotatko, Petr, Pagin, Peter and Segal, Gabriel (eds.), *Interpreting Davidson*(Stanford: CSLI Publications, 2001).
- LePore, Ernest (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson* (Oxford: Basil Blackwell, 1986) - includes ’A Nice Derangement of Epitaphs’.
- LePore, Ernest and McLaughlin, Brian, (eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson* (Oxford: Basil Blackwell, 1985).
- Malpas, J. E., *Donald Davidson and the Mirror of Meaning* (Cambridge: Cambridge University Press, 1992)
- Ramberg, Bjorn, *Donald Davidson’s Philosophy of Language: An Introduction* (Oxford: Basil Blackwell, 1989)
- Stoecker, Ralf (ed.), *Reflecting Davidson* (Berlin: W. de Gruyter, 1993).

Other Internet Resources

- Kalugin, V., "Donald Davidson", *Internet Encyclopedia of Philosophy*, J. Fieser (ed.), U. Tennessee/Martin

Related Entries

action | causation: the metaphysics of | events | knowledge: analysis of | meaning | meaning holism | mental content: externalist theories of | mind: philosophy of | Quine, Willard van Orman | rationality | semantics | supervenience | Tarski, Alfred | truth

Jeff Malpas
Jeff.Malpas@utas.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 29, 1996
Content last modified: February 1, 2002

The Analysis of Knowledge

The objective of the analysis of knowledge is to state the conditions that are individually necessary and jointly sufficient for propositional knowledge: knowledge that such-and-such is the case. Propositional knowledge must be distinguished from two other kinds of knowledge that fall outside the scope of the analysis: knowing a place or a person, and knowing how to do something. The concept to be analyzed -- the analysandum -- is commonly expressed using the schema "*S* knows that *p*", where "*S*" refers to the knowing subject, and "*p*" to the proposition that is known. A proposed analysis consists of a statement of the following form: *S* knows that *p* if and only if -- . The blank is to be replaced by the analysans: a list of conditions that are individually necessary and jointly sufficient. To test whether a proposed analysis is correct, we must ask (a) whether every possible case in which the conditions listed in the analysans are met is a case in which *S* knows that *p*, and (b) whether every possible case in which *S* knows that *p* is a case in which each of these conditions is met. When we ask (a), we wish to find out whether the proposed analysans is sufficient for *S*'s knowing that *p*; when we ask (b), we wish to determine whether each of the conditions listed in the analysans is necessary.

1. Knowledge as Justified True Belief

According to the following analysis, which is usually referred to as the "JTB" account, knowledge is

justified true belief.

> **The JTB Analysis of Knowledge**:
> *S* knows that *p* iff
> (i) *p* is true;
> (ii) *S* believes that *p*;
> (iii) *S* is justified in believing that *p*.

Condition (i), the truth condition, enjoys nearly universal assent, and thus has not generated any significant degree of discussion. It is overwhelmingly clear that what is false cannot be known. For example, it is false that G. E. Moore is the author of *Sense and Sensibilia.* Since it is false, it is not the sort of thing anybody can know.

1.1 The Belief Condition

Unlike the truth condition, condition (ii), the belief condition, has generated at least some discussion. Although initially it might seems obvious that knowing that *p* requires believing that *p*, some philosophers have argued that knowledge without belief is indeed possible. Suppose Walter comes home after work to find out that his house has burned down. He utters the words "I don't believe it." Critics of the belief condition might argue that Walter knows that his house has burned down (he sees that it has), but, as his words indicate, he does not believe it. Therefore, there is knowledge without belief. To this objection, there is an effective reply. What Walter wishes to convey by saying "I don't believe it" is not that he really does not believe what he sees with his own eyes, but rather that he finds it hard to come to terms with what he sees.

A more serious counterexample has been suggested by Colin Radford.[1] Suppose Albert is quizzed on English history. One of the questions is: When did Queen Elizabeth die?" Albert doesn't think he knows, but answers the question correctly. Moreover, he gives correct answers to many other question to which he didn't think he knew the answer. Let us focus on Albert's answer to the question about Elizabeth:

> (E) Elizabeth died in 1603.

Radford makes the following two claims about this example:

(a) Albert does not believe (E). Reason: He thinks he doesn't know the answer to the question. He doesn't trust his answer because he takes it to be a mere guess.

(b) Albert knows (E). Reason: His answer is not at all just a lucky guess. The fact that he answers most of the questions correctly indicates that he has actually learned, and never forgotten, the basic facts of English history.

Since he takes (a) and (b) to be true, Radford would argue that knowledge without belief is indeed possible. How would an advocate of the JTB account respond to Radford's proposed counterexample? Their response would be, in short, that this is not a case of knowledge without belief because it isn't a case of knowledge to begin with. Albert doesn't know (E) because he has no justification for believing (E). If he were to believe (E), his belief would be unjustified. This reply anticipates what we have not yet discussed: the necessity of the justification condition. Let us first discuss why friends of JTB hold that knowledge requires justification, and then discuss in greater detail why they would not accept Radford's alleged counterexample.

1.2 The Justification Condition

Why is condition (iii) necessary? Why not say that knowledge is true belief? The standard answer is that to identify knowledge with true belief would be implausible because a belief that is true just because of luck does not qualify as knowledge. Beliefs that are lacking justification are false more often than not. However, on occasion, such beliefs happen to be true. Suppose William takes a medication that has the following side effect: it causes him to be overcome with irrational fears. One of his fears is that he has cancer. This fear is so powerful that he starts believing it. Suppose further that, by sheer coincidence, he does have cancer. So his belief is true. Clearly, though, his belief does not amount to knowledge. But why not? Most epistemologists would agree that William does not know because his belief's truth is due to luck (bad luck, in this case). Let us refer to a belief's turning out to be true because of mere luck as *epistemic luck*.It is uncontroversial that knowledge is incompatible with epistemic luck. What, though, is needed to rule out epistemic luck? Advocates of the JTB account would say that what is needed is justification. A true belief, if an instance of knowledge and thus not true because of epistemic luck, must be justified.[2] But what is it for a belief to be justified?

Among the philosophers who favor the JTB approach, we find bewildering disagreement on how this question is to be answered. According to one prominent view, typically referred to as "evidentialism", a belief is justified if, and only if, it fits the subject's evidence.[3] Evidentialists, then, would say that the reason why knowledge is not the same as true belief is that knowledge requires evidence. Opponents of evidentialism would say that evidentialist justification (i.e., having adequate evidence) is not needed to rule out epistemic luck. They would argue that what is needed instead is a suitable relation between the belief and the mental process that brought it about. What we are looking at here is an important disagreement about the nature of knowledge, which will be our main focus further below. In the meantime, we will continue our examination of the JTB analysis.

Let us return to Radford's counterexample to the belief condition, which we considered above. We are now in a position to discuss further the reply to it. Recall that Albert does not take himself to know the answer to the question about the date of Elizabeth's death. He does not because he has does not remember having learned the basic facts of British history. Now, it is of course true that he did learn these facts, and is indeed able to recall them. But is this by itself sufficient for knowing them? Philosophers who think that knowledge requires evidence would say that it is not. Albert needs to have evidence for believing that he learned those facts. Until he is quizzed, he has no such evidence. *After* the

quiz, when he is told that most of his answers were correct, he does have the requisite evidence. For once he comes to know that he is able to produce consistently correct answers to the questions he is asked, he has acquired evidence for believing that he must have learned this subject matter at school. This evidence is also evidence for the answers he has given. So at that point, the justification condition is met, and thus (since the other conditions of knowledge are also met) he knows (again) that Elizabeth died in 1603. However, he did not know this before he finds out that he must have learned those facts, for at that point his answer to the question lacked justification, and thus did not add up to knowledge. Evidentialists would deny, therefore, that Radford has supplied us with a counterexample to the belief condition.[4]

2. The Gettier Problem

In his short 1963 paper, "Is Justified True Belief Knowledge?", Edmund Gettier presented two effective counterexamples to the JTB analysis.[5] The second of these goes as follows. Suppose Smith has good evidence for the false proposition

> (1) Jones owns a Ford.[6]

Suppose further Smith infers from (1) the following three disjunctions:

> (2) Either Jones owns a Ford or Brown is in Boston.
> (3) Either Jones owns a Ford or Brown is in Barcelona.
> (4) Either Jones owns a Ford or Brown is in Brest-Litovsk.

Since (1) entails each of the propositions (2) through (4), and since Smith recognizes these entailments, he is justified in believing each of propositions (2)-(4). Now suppose that, by sheer coincidence, Brown is indeed in Barcelona. Given these assumptions, in believing (3), Smith holds a justified true belief. However, is it an instance of knowledge? Since Smith has no evidence whatever as to Brown's whereabouts, and believes what is true only because of luck, the answer would have to be ‘no’. Consequently, the three conditions of the JTB account -- truth, belief, and justification -- are not sufficient for knowledge.[7] How must the analysis of knowledge be modified to make it immune to cases like the one we just considered? This is what is commonly referred to as the "Gettier problem".

Epistemologists who think that the JTB approach is basically on the right track must choose between two different strategies for solving the Gettier problem. The first is to strengthen the justification condition. This was attempted by Roderick Chisholm.[8] The second strategy is to search for a suitable further condition, a condition that would, so to speak, "degettierize" justified true belief. Let us focus on this second strategy. According to one suggestion, the following fourth condition would do the trick:

> (iv) *S*'s belief that *p* is not inferred from any falsehood.[9]

Unfortunately, this proposal is unsuccessful. Since Gettier cases need not involve any inference, there are

possible cases of justified true belief in which the subject fails to have knowledge although condition (iv) is met. Suppose, for example, that James, who is relaxing on a bench in a park, observes a dog that, about 8 yards away from him, is chewing on a bone. So he believes

(5) There is a dog over there.

Suppose further that what he takes to be a dog is actually a robot dog so perfect that, by vision alone, it could not be distinguished from an actual dog. James does not know that such robot dogs exist. But in fact a Japanese toy manufacturer has recently developed them, and what James sees is a prototype that is used for testing the public's response. Given these assumptions, (5) is of course false. But suppose further that just a few feet away from the robot dog, there is a real dog. Sitting behind a bush, he is concealed from James's view. Given this further assumption, James's belief is true. So once again, what we have before us is a justified true belief that doesn't qualify as an instance of knowledge. Arguably, this belief is directly justified by a visual experience; it is not inferred from any falsehood. But if (5) is indeed a non-inferential belief, then the JTB account, even if supplemented with (iv), gives us the wrong result that James knows (5).

Another case illustrating that clause (iv) won't do the job is the well-known Barn County case.[10] Suppose there is a county in the Midwest with the following peculiar feature. The landscape next to the road leading through that county is peppered with barn-facades: structures that from the road look exactly like barns. Observation from any other viewpoint would immediately reveal these structures to be fakes: devices erected for the purpose of fooling unsuspecting motorists into believing in the presence of barns. Suppose Henry is driving along the road that leads through Barn County. Naturally, he will on numerous occasions form a false belief in the presence of a barn. Since Henry has no reason to suspect that he is the victim of organized deception, these beliefs are justified. Now suppose further that, on one of those occasions when he believes there is a barn over there, he happens to be looking at the one and only real barn in the county. This time, his belief is justified and true. But its truth is the result of luck, and thus his belief is not an instance of knowledge. Yet condition (iv) is met in this case. His belief is clearly not the result of any inference from a falsehood. Once again, we see that (iv) does not succeed as a solution to the Gettier problem.

Above, we noted that the role of the justification condition is to ensure that the analysans does not mistakenly identify as knowledge a belief that is true because of epistemic luck. The lesson to be learned from the Gettier problem is that the justification condition by itself cannot ensure this. Even a justified belief, understood as a belief based on good evidence, can be true because of luck. Thus if a JTB analysis of knowledge is to rule out the full range of cases of epistemic luck, it must be amended with a suitable fourth condition, a condition that succeeds in preventing justified true belief from being "gettiered." We will refer to an analysis of this type as a "JTB+" conception of knowledge.

3. An Alternative Approach: Reliabilism

The analysis of knowledge may be approached by asking the following question: What turns a true belief

into knowledge? An uncontroversial answer to this question would be: the sort of thing that effectively prevents a belief from being true as a result of epistemic luck. Controversy begins as soon as this formula is turned into a substantive proposal. According to evidentialism, which endorses the JTB+ conception of knowledge, the combination of two things accomplishes this goal: evidentialist justification plus degettierization (a condition that prevents a true and justified belief from being "gettiered"). However, according to an alternative approach that has in the last three decades become increasingly popular, what stands in the way of epistemic luck -- what turns a true belief into knowledge -- is the reliability of the cognitive process that produced the belief. Consider how we acquire knowledge of our physical environment: we do so through sense experience. Sense experiential processes are, at least under normal conditions, highly reliable. There is nothing accidental about the truth of the beliefs these processes produce. Thus beliefs produced by sense experience, if true, should qualify as instances of knowledge. An analogous point could be made for other reliable cognitive processes, such as introspection, memory, and rational intuition. We might, therefore, say that what turns true belief into knowledge is the reliability of our cognitive processes.

This approach -- reliabilism, as it is usually called -- can be carried out in two different ways. First, there is reliabilism as a theory of justification (J- reliabilism).[11] Here the idea is that while justification is indeed necessary for knowledge, its nature is not evidentialist but reliabilist. The most basic version of this view -- let's call it "simple" J-reliabilism -- goes as follows:

> **Simple J-Reliabilism**:
> *S* is justified in believing that *p* if, and only if, *S*'s belief that *p* was produced by a reliable cognitive process.

Second, there is reliabilism as a theory of knowledge (K-reliabilism).[12] According to this approach, knowledge does not require justification. Rather, what it requires (in addition to truth) is reliable belief formation. Fred Dretske defends this view as follows:

> Those who think knowledge requires something *other than*, or at least *more than*, reliably produced true belief, something (usually) in the way of justification for the belief that one's reliably produced beliefs *are* being reliably produced, have, it seems to me, an obligation to say what benefits this justification is supposed to confer . . . Who needs it, and why? If an animal inherits a perfectly reliable belief-generating mechanism, and it also inherits a disposition, everything being equal, to *act* on the basis of the belief so generated, what additional benefits are conferred by a justification that the beliefs *are* being produced in some reliable way? If there are no additional benefits, what good is this justification? Why should we insist that no one can have knowledge without it?[13]

Further below we will discuss how advocates of the JTB approach might answer Dretske's question. In the meantime, let us focus a bit more on Dretske's account of knowledge. According to Dretske, reliable cognitive processes convey information, and thus endow not only humans, but (nonhuman) animals as well, with knowledge. He writes:

> I wanted a characterization that would at least allow for the possibility that animals (a frog, rat, ape, or my dog) could know things without my having to suppose them capable of the more sophisticated intellectual operations involved in traditional analyses of knowledge.[14]

Attributing knowledge to animals is certainly in accord with our ordinary practiceof using the word ‘knowledge’. Dretske seems right, therefore, when he views the result that animals have knowledge as a desideratum.

A second advantage of his theory is, so Dretske claims, that it avoids Gettier problems. He says:

> Gettier difficulties . . . arise for any account of knowledge that makes knowledge a product of some justificatory relationship (having good evidence, excellent reasons, etc.) that *could* relate one to something false . . . This is [a] problem for justificational accounts. The problem is evaded in the information-theoretic model, because one can get into an appropriate justificational relationship to something false, but one cannot get into an appropriate informational relationship to something false.[15]

Solving the Gettier-problem is, however, a bit more complex than this passage suggests. Consider again the case of Henry in Barn County. He sees a real barn in front of him, yet does not know that there is a barn near-by. Exactly how can Dretske's theory explain Henry's failure to know? After all, he perceives an actual barn, and so does not stand in any informational relationship to something false. So if perception, on account of its reliability, normally conveys information, it should do so in this case as well. Alas, it doesn't. Clearly, if a theory like Dretske's is to handle this case and others like it, it must be supplemented with a clause that makes it immune to the case of the fake barns, and other examples like it.[16]

4. Internalism and Externalism

Evidentialists reject both J-reliabilism and K-reliabilism. They reject J-reliabilism because they advocate internalism: they take justification to be something that is "internal" to the subject. J-reliabilists disagree; they take justification to be something that is "external" to the subject.[17]

In order to pin down what the "internality" of justification is supposed to be, let us turn to Roderick Chisholm, one of the chief advocates of internalism. In the third edition of *The Theory of Knowledge*, Chisholm says the following:

> If a person *S* is internally justified in believing a certain thing, then this may be something he can know just by reflecting upon his own state of mind.[18]

In the second edition of this book, he characterizes internalism in a somewhat different way:

> We presuppose . . . that the things we know are justified for us in the following sense: we can know what it is, on any occasion, that constitutes our grounds, or reasons, or evidence for thinking that we know.[19]

These passages differ in the following respect: in the first Chisholm is concerned with the property of justification (a belief's being justified); in the second, with justifiers: the things that make justified beliefs justified. What is common to both passages is the constraint Chisholm imposes. In the first passage, Chisholm characterizes justification as something that is recognizable *on reflection*, and in the second as the sort of thing that can be known *on any occasion.*Arguably, this is just a terminological difference. It would not be implausible to claim that what can be recognized through reflection is something that can be recognized on any occasion, and what can be recognized on any occasion is something that can be recognized through reflection. Although this point deserves further examination, let us here simply assume that recognizability on reflection and recognizability on any occasion amount to the same thing. In what follows, we will refer to it as *direct recognizability*.

As already noted, in the first passage Chisholm imposes the direct recognizability constraint on justification, in the second on justifiers. Does this amount to a substantive difference? If the direct recognizability of justifiers implies the direct recognizability of justification, and vice versa, then the two passages we considered would indeed just be alternative ways of stating the same point. Whether they really are is debatable, but here we will simply assume that it makes no difference whether internalism is characterized in terms of the direct recognizability of justification, or that of justifiers.

Chisholm, then, defines internalism in terms of how justification (justifiers) is (are) knowable, that is, in terms of direct recognizability, or epistemic accessibility. This type of internalism may therefore be called *accessibility internalism*. Alternatively, internalism can be defined in terms of limiting justifiers to mental states. According to this second way of defining internalism, justifiers must be internal to the mind, i.e., must be mental events or states. Internalism thus defined could be referred to as *mental state internalism*.[20] Whether accessibility internalism and mental state internalism are genuine alternatives depends on whether mental states (and events) are directly recognizable. If they are, what appear to be genuine alternatives might in fact not be.[21] Since here we cannot go into the details of this issue, we will cut this matter short and simply define internalism, as suggested by Chisholm, in terms of direct recognizability, while acknowledging that it might be preferable to define it by restricting justifies to mental states. We will refer to internalism as defined here as "J-internalism," since it imposes the direct recognizability constraint on not knowledge, but justification.

> **J-Internalism**:
> Justification is directly recognizable. At any time *t* at which *S* holds a justified belief *B*, *S* is in a position to know at *t* that *B* is justified.[22]

J-internalism is to be contrasted with J-externalism, which is simply its negation.

> **J-Externalism**:
> Justification is not directly recognizable. It is not the case that at any time *t* at which *S* holds a justified belief B, *S* is in a position to know at *t* that *B* is justified. (There are times at which *S* holds a justified belief *B* but is not in a position to know that *B* is justified.)

Next, we will discuss what consequences we can derive from J-internalism. To begin with, we can derive the result that simple J-reliabilism is an externalist theory. According to Simple J-Reliabilism, reliability by itself -- without the subject's having any evidence indicating its presence -- is sufficient for justification. So simple J-reliabilism allows for possible cases of the following kind:

- (i) the subject holds a reliably formed belief;
- (ii) the subject has no evidence whatever indicating that this belief was reliably formed, nor any other evidence in its support;
- (iii) the belief in question is justified.

To illustrate this point, let us consider a familiar example due to Laurence BonJour. Suppose Norman is a perfectly reliable clairvoyant. At time t, his clairvoyance causes Norman to form the belief that the president is presently in New York. However, Norman has no evidence whatever indicating that he is clairvoyant. Nor has he at *t* any way of recognizing that his belief was caused by his clairvoyance. Norman, then, cannot at *t* recognize that his belief is justified. So Simple J-reliabilism implies that Norman's belief is justified at *t* although Norman cannot recognize at *t* that that his belief is justified. Simple J-Reliabilism, therefore, is a version of J-externalism.

Second, J-internalism allows us to derive the consequence -- as it should -- that evidentialism is an internalist theory. The question of what a person's evidence consists of is of course not uncontroversial.[23] Nor is it uncontroversial what kind of cognitive access a subject has to her evidence.[24] However, it would certainly not be without a good deal of initial plausibility, at least if one looks at the matter from the point of view of the evidentialist, to make the following two assumptions. First, a subject's evidence consists of both her beliefs and experiential states (such as sensory, introspective, memorial, and intuitional states). Second, a subject's beliefs and experiential states are directly recognizable to her. And if we now add the further assumption (mentioned above) that the direct recognizability of justifiers implies the direct recognizability of justification, then we get the result that evidentialism is a form of J-internalism. Let us display the argument in detail:

> **Why Evidentialism Is A Version of J-Internalism**:
> (1) According to evidentialism, justifiers consist of a person's evidence.
> (2) A person's evidence is directly recognizable to that person.
> Therefore:
> (3) According to evidentialism, a person's justifiers are directly recognizable to that person.

(4) If the justifiers that make a person's justified beliefs justified are directly recognizable to that person, then the justification of that person's justified beliefs is directly recognizable to that person.

Therefore:

(5) According to evidentialism, the justification of a person's justified beliefs is directly recognizable to that person.

The crucial premises in this argument are (2) and (4). Surely, evidentialists would be reluctant to call "evidence" something that is not directly recognizable to a subject.[25] So (2) would appear to be a premise that evidentialists are likely to endorse. And (4) expresses no more than one part of what we already assumed: that the direct recognizability of justifiers implies the direct recognizability of justification, and vice versa. Of course, this assumption might be challenged. What seems safe to say, therefore, is the conditional point that, if (2) and (4) capture what is essential to evidentialism, then evidentialism implies internalism about justification.

As mentioned at the beginning of this section, evidentialists also reject K-reliabilism. They do so because, pace Dretske, they think that internal justification -- justification in the form of having adequate evidence -- is necessary for knowledge. In other words, they deny that a belief's origin in a reliable cognitive process is sufficient for the belief's being an instance ofknowledge. Let us refer to this position as internalism about knowledge, or K-internalism, and let us define it using the concept of *internal justification*: the kind of justification that meets the direct recognizability constraint.

> **K-Internalism**:
> Internal justification is a necessary condition of knowledge. A belief's origin in a reliable cognitive process is *not* sufficient for its being an instance of knowledge.

K-externalism is simply the negation of internalism:

> **K-Externalism**:
> Internal justification is not a necessary condition of knowledge. A belief's origin in a reliable cognitive process *is* sufficient for its being an instance of knowledge. Consequently, there are cases of knowledge without internal justification.

In this section, we have merely concerned ourselves with what internalists and externalists disagree about with regard to both justification and knowledge. In the next two sections, we will examine what reasons internalists and externalists can cite in support of their respective views.

5. Why Internalism?

To begin with, one straightforward argument for J-internalism proceeds from evidentialism as a premise. For as we have seen above, there is a plausible construal of evidentialism that proceeds from the direct

recognizability of a person's evidence to the direct recognizability of justification. So philosophers who are attracted to evidentialism are likely to be attracted to J-internalism as well. Furthermore, as was already mentioned at the end of the previous section, evidentialism is not only a view about the nature of justification, but also a view about the nature of knowledge. And what evidentialists would say about the nature of knowledge is this: having justification -- in the form of having adequate evidence -- is a necessary condition of knowledge.[26] But such justification is plausibly construed as internal justification, as satisfying the direct recognizability constraint that J-internalism imposes. It would appear, therefore, that evidentialists take internal justification to be necessary for knowledge, and thus hold the view we have labeled "K-internalism".

Second, there is an argument for internalism that starts with what is referred to as the deontological conception of epistemic justification:

> **Deontological Justification**:
> *S* is justified in believing that *p* iff in believing that *p*, *S* does not violate his epistemic duty.

The concept of duty employed here must not be confused with ethical duty, or prudential duty. The type of duty in question is specifically epistemic.[27] Exactly what epistemic duties are, however, is a matter of controversy. The basic idea is that epistemic duties are those that arise in the pursuit of truth.[28] Thus we might express (1) alternatively as follows: *S* is justified in believing that *p* iff in believing that *p*, *S* does not fail to do what he ought to do in the pursuit of truth. Of course, this way of putting things leads us directly to a further question: in the pursuit of truth, exactly what is it that one ought to do? Evidentialists would say: it is to believe what, and only what, one has evidence for.[29] Now if that is one's epistemic duty, then those who take justification to be deontological can employ the argument considered above (which proceeds from evidentialism to J-internalism) to derive the conclusion that deontological justification is internal justification. So the combination of deontology about justification with evidentialism allows for a pretty straightforward derivation of J-internalism.

It has also been suggested that there is a more direct argument from deontology to J-internalism, an argument that does not depend on evidentialism as a premise.[30] It derives the direct recognizability of justification from the premise that what determines epistemic duty is directly recognizable.

> **From Deontology to Internalism**:
> (1) Justification is a matter of epistemic duty fulfillment.
> Therefore:
> (2) What determines justification is identical to what determines epistemic duty.
> (3) What determines epistemic duty is directly recognizable.
> Therefore:
> (4) What determines justification is directly recognizable.

(5) If what determines justification is directly recognizable, then justification itself is directly recognizable.

Therefore:

(6) Justification is directly recognizable.

(2) follows directly from the deontological conception of justification. (5) is nothing new; we have assumed it above already. The argument's main premise is of course (3).[31] Certainly (3) is not obviously implausible. Nevertheless, it is open to criticism, as is (5), which we merely assumed. Obviously, then, the argument is not uncontroversial. Nevertheless, it seems fair to say that it represents a straightforward and defensible derivation of internalism from deontology.

Third, internalism (J or K) can be defended indirectly on the basis of objections to particular externalist accounts of justification or knowledge. Since reliabilism is the dominant externalist approach, let us briefly consider a couple of internalist objections to reliabilism. First, recall BonJour's example of Norman: a subject who unwittingly possesses a reliable faculty of clairvoyance.[32] This faculty produces the belief that the president is in New York, a belief that is reliably produced, and thus according to simple J-reliabilism justified. But is that belief really justified? Internalists would say that Norman's belief is actually unjustified, and thus not an instance of knowledge. They would say, therefore, that a belief's being reliably produced is not sufficient for making it justified, and that a true belief's being reliably produced is not sufficient for making it an instance of knowledge.

Second, internalists would say that reliable belief production is not even necessary for knowledge. Suppose you are a victim of Descartes's evil demon. You believe that you have a body and that there is a world of physical things, but in fact neither of these beliefs is true. There is no physical world at all. Since your perceptual beliefs are not reliably produced under these circumstances, simple J-reliabilism implies that they are unjustified. To internalists, this is an intuitively implausible result. They would take your beliefs to be (by and large) justified because they are (by and large) based on adequate evidence or good reasons.[33] Hence they would reject the claim that being produced by reliable faculties is a necessary condition of epistemic justification.[34]

6. Why Externalism?

One reason for externalism lies in the attraction of "philosophical naturalism." According to Gilbert Harman, this view, when applied to ethics, "is the doctrine that moral facts are facts of nature. Naturalism as a general view is the sensible thesis that *all* facts are facts of nature."[35] What naturalists in ethics want, according to Harman,

> is to be able to locate value, justice, right, wrong, and so forth in the world in the way that tables, colors, genes, temperatures, and so on can be located in the world.[36]

According to this conception of naturalism, a naturalist in epistemology wants to be able to locate such things as knowledge, certainty, epistemic justification, and probability "in the world in the way that tables, colors, genes, temperatures, and so on can be located in the world." How, though, are naturalists to accomplish this? According to one answer to this question, they can accomplish this by identifying the non-epistemic grounds on which epistemic phenomena supervene. Alvin Goldman describes this desideratum as follows:

> The term "justified," I presume, is an evaluative term, a term of appraisal. Any correct definition or synonym of it would also feature evaluative terms. I assume that such definitions or synonyms might be given, but I am not interested in them. I want a set of *substantive* conditions that specify when a belief is justified . . . I want a theory of justified belief to specify in non-epistemic terms when a belief is justified."[37]

However, internalists need not deny that epistemic phenomena supervene on non-epistemic grounds, and that it is the task of epistemology to reveal these grounds. That is, internalists might as well agree that what a theory of justification ought to accomplish is an account of the substantive conditions of justification that is carried out in non-epistemic terms. It is doubtful, therefore, that the goal of locating epistemic value in the natural world establishes a link between philosophical naturalism and externalism.[38]

According to a second answer to the question of how epistemic value can be located in the natural world, the way to do that is to employ the methods of the natural sciences.[39] Appealing to this methodological constraint, externalists might argue that, because the study of justification and knowledge is an empirical study, justification and knowledge cannot be what internalists take it to be, but rather must be identified with reliable belief production: a phenomenon that can be studied empirically. It is far from clear, however, that the fundamental questions of epistemology can be answered by employing the methods of natural science. If they cannot be answered that way, then epistemology cannot be done without employing, at least to some extent, the a priori methods of the armchair philosopher. But then the universal scope of the methodological constraint in question remains unmotivated, and no compelling reason remains to think that justification and knowledge are the sort of thing that can only be studied empirically, and thus cannot be what internalist take them to be.

A second reason for externalism (more specifically, J-externalism) has to do with the connection between justification and truth. Internalists conceive of a justified belief as a belief that, relative to the subject's evidence or reasons, is likely to be true. However, such likelihood of truth is compatible with the belief's actual falsity. Indeed, such likelihood of truth is compatible with the evil demon scenario in which the vast majority of your empirical beliefs, although justified, is in fact false. Externalists consider this connection between justification and truth too thin, and thus demand a stronger kind of likelihood of truth.[40] Reliability is usually taken to fill the bill.[41] William Alston, for example, would concur that, without a reliability constraint, the connection between justification and truth becomes too tenuous.[42] He argues that only reliably formed beliefs can be justified, and defines a reliable belief-producing mechanism as one that "would yield mostly true beliefs in a sufficiently large and varied run of

employments in situations of the sorts we typically encounter."[43] Suppose we endorse this conception of justification. Let's suppose further that most of our beliefs are justified. It then follows that most of the beliefs we form in ordinary circumstances would have to be true most of the time. Such a belief system could still be brought about by an evil demon. However, it would not be a belief-system consisting of mostly false beliefs, and thus the evil demon responsible for it wouldn't be quite as evil as he could be. So what Alston-type justification rules out is this: a belief system of mostly justified beliefs that is generated by an evil demon who sees to it that most of our beliefs are false. This, then, is the benefit we can secure when, as externalists suggest, we make reliability a necessary element of justification.

Internalists would object that a strong link between justification and truth runs afoul of the rather forceful intuition that the beliefs of an evil demon victim are justified although they are mostly false. In response, externalists might concede that the sort of justification internalists have in mind and attribute to evil demon victims is a legitimate concept, but question the epistemological relevance of that concept. Of what *epistemic* value (of what value to the acquisition of knowledge), they might ask, is internal justification if it is the sort of thing an evil demon victim can enjoy, a person whose belief system is massively marred by falsehood? Internalists would reply that internal justification should not be expected to supply us with a guarantee of truth, and that its value derives from the fact that internal justification is necessary for knowledge.

A third reason for externalism has to do with Dretske's question about justification: "Who needs it, and why?" Dretske would say, of course, that nobody needs it (for the acquisition of knowledge, that is) because reliable belief production is sufficient for turning true belief into knowledge. With this, internalists disagree.[44] They take the existence of examples like BonJour's clairvoyant Norman as a decisive reason to reject this sufficiency claim. According to them, Norman's belief about the whereabouts of the president, although reliably formed, is clearly unjustified, and thus not an instance of knowledge. Internalists, therefore, would answer Dretske's question thus: Those who wish to enjoy knowledge need justification, and they need it because one does not know that *p* unless one has adequate evidence or undefeated reasons for believing that *p*.

In reply to this, Dretske might repeat a point -- a point that amounts to a fourth reason for externalism -- from the passage we considered above: he takes animals such as frogs, rats, apes, and dogs to have knowledge. This is surely in line with the way we ordinarily use the concept of knowledge. The owner of a pet who does not attribute knowledge to it would be hard to find. But are animals capable of the sophisticated mental operations required by beings who enjoy the sort of justification internalists have in mind? It would seem not.[45] At this point, the disagreement between internalists and externalists appears unresolvable. On the one hand, there are examples like BonJour's clairvoyant Norman, examples that strongly suggest that internal justification *is* necessary for knowledge. On the other hand, there is Dretske's point that knowledge is enjoyed by not only humans but animals as well. And this strongly suggests that internal justification is *not* necessary for knowledge.

7. Two Analyses of Knowledge

K-internalism and K-externalism, then, are supported by conflicting intuitions. On the one hand, there is the thought that in order to know, one must have justification in the form of having adequate evidence or reasons. On the other hand, there is the thought that knowledge, resulting from reliable cognitive faculties, is not reserved to humans only. Both of these thoughts are inherently plausible. However, if it is indeed true that animals are not the sort of beings that can have internally justified or unjustified beliefs, these intuitions cannot be reconciled. If they cannot, then we get as a result of this irreconcilability two alternative and competing analyses of knowledge: one internalist, the other externalist. Let us state a gloss of the respective analyses. In these analyses, the term "internal justification" stands for the kind of concept internalists have in mind, and the term "external justification" for the kind of concept externalists employ.

> **External Knowledge (EK)**:
> *S* knows that *p* iff
>
> (i) *p* is true;
>
> (ii) *S* believes that *p*;
>
> (iii) *S* is externally justified in believing that *p*.

> **Internal Knowledge (IK)**:
> *S* knows that *p* iff
>
> (i) *p* is true;
>
> (ii) *S* believes that *p*;
>
> (iii) *S* is internally justified in believing that *p*;
>
> (iv) *S*'s belief that *p* is degettiered.

EK and IK agree and differ in the following respects:

(a) According to both EK and IK, knowledge requires true belief. The question each of these analyses is intended to answer is: what do we need to add to true belief to get knowledge?

(b) According to both, external justification is necessary for knowledge. K-internalists acknowledge the necessity of external justification at least indirectly when they make degettierization a necessary condition of knowledge. The explanation of this point has to do with the nature of degettierization. What sort of thing would achieve it? Let us venture a hypothesis: what achieves degettierization is the sort of thing that produces external justification. If that is correct, then it follows that K-internalists are in agreement with K-externalists about the necessity of external justification.

(c) IK requires internal justification, EK does not. That is the one condition where the two analyses differ. As a result of this difference, EK includes within the scope of knowledge animals, but fails to accommodate the intuition underlying BonJour's case of clairvoyant Norman, and other cases like that. IK, on the other hand, does accommodate this intuition, but -- counter-intuitively, as K-externalists would say -- excludes animals from the range of subjects that can have knowledge.

If the internalism/externalism controversy is seen as essentially a controversy over the nature of knowledge, the debate over J-internalism vs. J-externalism would appear to be a case of talking past each other. J-internalists and J-externalists simply intend justification to achieve different things. They each operate with a different concept of justification. J-externalists take justification to be the sort of thing that turns true belief into knowledge, and view the Gettier problem merely as the problem of adding the right sort of bells and whistles to the justification-condition. J-internalists, on the other hand, cannot view degettierization as something that can, in the form of a suitable clause, be tacked on to the justification condition, for degettierization is an external matter.[46] Rather, internalists take justification to be the sort of thing that turns true *and* degettiered belief into knowledge. Since J-internalists and J-externalists assign different roles to justification, what they ultimately disagree about is not the nature of justification, but the sort of thing in relation to which the theoretical role of epistemic justification is fixed: knowledge. Internalists assign justification the role of turning true and degettiered belief into knowledge because they take internal justification to be necessary for knowledge. In contrast, externalists assign a different role -- that of turning true belief into knowledge -- to justification because they think that internal justification is *not* necessary for knowledge. It is this difference in their respective views on the nature of knowledge that leads to different views on the nature of justification.

Thus we are confronted with a fundamental disagreement about the nature of knowledge. Externalists such as Dretske would say that the desideratum of making knowledge a natural phenomenon that is instantiated equally by humans and animals must trump the demand that knowledge require the possession of justification in the form of adequate evidence. They would have to say, therefore, that Norman, the unwitting clairvoyant, has knowledge just as much as a mouse that knows where to look for the cheese. Internalists would argue the other way around. To them, Norman-type cases establish the necessity of adequate evidence or undefeated reasons. And so they would say that, just as Norman's reliable clairvoyance (by itself, in the absence of evidence) does not give him knowledge, a mouse's reliable cognitive mechanisms do not give it knowledge of where to look for the cheese. Externalists would say that it merely seems to us that Norman lacks knowledge when in fact he has it. Internalists would say that it merely seems to us that animals know when in fact they do not.

Who is right about the nature of knowledge: internalists or externalist? It might be a mistake to expect that there is a decisive argument that settles the dispute one way or the other.[47] Most likely, one reason why the nature of knowledge is a subject matter of philosophy is that in the end its nature remains enigmatic. Nevertheless, the common ground shared by IK and EK should not be overlooked. Both require true belief and external justification. What is contentious is merely the further question of whether knowledge requires internal justification as well.[48]

Bibliography

- Almeder, Robert. 1998. *Harmless Naturalism. The Limits of Science and the Nature of Philosophy*. Chicago and La Salle: Open Court.
- Alston, William. 1989. *Epistemic Justification. Essays in the Theory of Knowledge*. Ithaca: Cornell University Press.
- -------. 1991. *Perceiving God. The Epistemology of Religious Experience*. Ithaca: Cornell University Press.
- -------. 1993. *The Reliability of Sense Perception*. Ithaca: Cornell University Press.
- -------. 1996. *A Realist Conception of Truth*. Ithaca: Cornell University Press.
- Armstrong, D.M. 1973. *Belief, Truth, and Knowledge*. Cambrdidge: Cambridge University Press.
- BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge: Harvard University Press.
- Chisholm, Roderick. 1989. *Theory of Knowledge*, 3rd. ed., Englewood Cliffs: Prentice Hall.
- -------. 1977.*Theory of Knowledge*, 2nd ed., Englewood Cliffs: Prentice Hall.
- Clark, Michael. 1963. "Knowledge and Grounds. A Comment on Mr. Gettier's Paper. *Analysis* 24, pp. 46-48.
- Cohen, Stewart. 1984. "Justification and Truth," *Philosophical Studies* 46, pp. 279-95.
- Conee, Earl and Feldman, Richard. 1985. "Evidentialism." *Philosophical Studies* 48.
- -------. Forthcoming. "Internalism Defended."
- David, Marian. 2001. "Truth and the Epistemic Goal." In: Steup 2001a.
- DePaul, Michael. 2001. "Value Monism in Epistemology." In: Steup, 2001a.
- DeRose, Keith. 1999. "Contextualism: An Explanation and Defense?" In: Greco 1999, pp. 187.
- -------. 2000. "Ought We to Follow Our Evidence?" *Philosophy and Phenomenological Research* 60, pp. 697-706.
- Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge: MIT Press.
- -------. 1985. "Precis of Knowledge and the Flow of Information." In: Hilary Kornblith, ed., *Naturalizing Epistemology*. Cambridge: MIT Press.
- -------. 1989. "The Need to Know." In: Marjorie Clay and Keith Lehrer, eds., *Knowledge and Skepticism*. Boulder: Westview Press.
- Feldman, Richard. 1988a. "Epistemic Obligations," in J.E. Tomberlin, ed., *Philosophical Perspectives* 2. Atascadero: Ridgeview, pp. 235-56.
- -------. 1988b. "Having Evidence." In: D. Austin (ed.), *Essays Presented to Edmund Gettier*. Dordrecht: Reidel.
- -------. 1992. "Evidence." In: Jonathan Dancy and Ernest Sosa. *A Companion to Epistemology*. Oxford: Blackwell, pp. 119-122.
- -------.1999. "Methodological Naturalism in Epistemology." In: Greco 1999.
- Fumerton, Richard. 1995. *Metaepistemology and Skepticism*. Lanham: Rowman and Littlefield.
- Goldman, Alvin. 1976. "Discrimination and Perceptual Knowledge." *The Journal of Philosophy* 73, pp. 771-791.
- -------. 1979. "What is Justified Belief?" In: *Justification and Knowledge*, ed. George S. Pappas. Dordrecht: Reidel.

- -------. 1986. *Epistemology and Cognition*. Cambridge: Harvard University Press.
- -------.1991. "Epistemic Folkways and Scientific Epistemology." In: *Liaisons: Philosophy Meets the Cognitive and Social Sciences*. (Cambridge: MIT Press.)
- -------. 1999. "Internalism Exposed." *The Journal of Philosophy* 96, pp. 271-293.
- Greco, John. 1993. "Virtues and Vices of Virtue Epistemology," *Canadian Journal of Philosophy* 23.
- Greco, John, and Sosa, Ernest (eds.). 1999. *The Blackwell Guide to Epistemology*. Oxford: Blackwell.
- Harman, Gilbert. 1977. *The Nature of Morality*. Oxford: Oxford University Press.
- -------.1984. "Is There a Single True Morality." In: David Copp and David Zimmerman (eds.). *Morality, Reason and Truth. New Essays on the Foundation of Ethics*. Totowa: Rowman and Allenheld, pp. 27-48.
- Kornblith, Hilary. 1999. "In Defense of a Naturalized Epistemology." In: Greco 1999.
- Lehrer, Keith. 1990. *Theory of Knowledge*. Boulder: Westview Press.
- Nozick, Robert. 1981. *Philosophical Explanations*. Cambridge: Harvard University Press.
- Plantinga, Alvin. 1993. *Warrant: The Current Debate*. Oxford: Oxford University Press.
- -------. 1993b. *Warrant and Proper Function*. Oxford: Oxford University Press.
- -------. 1996. "Respondeo." In: Jonathan L. Kvanvig. *Warrant in Contemporary Epistemology. Essays in Honor of Plantinga's Theory of Knowledge*. Lanham: Rowman and Littlefield.
- Pollock, John. 1986. *Contemporary Theories of Knowledge*. Totowa: Rowman and Littlefield.
- Radford, Colin. 1966. "Knowledge---By Examples," *Analysis* 27, pp. 1-11.
- Russell, Bruce. 2001 "Epistemic and Moral Duty." In: Steup 2001 a.
- Shope, Robert K. 1983. *The Analysis of Knowing. A Decade of Research*. Princeton: Princeton University Press.
- Sosa, Ernest. 1991. *Knowledge in Perspective. Selected Essays in Epistemology*. Cambridge: Cambridge University Press.
- -------. 1999. "Skepticism and the Internal/External Divide." In: Greco 1999, pp. 145-157.
- Steup, Matthias. 1996. *An Introduction to Contemporary Epistemology*. Upper Saddle River: Prentice Hall.
- -------. 1999. "A Defense of Internalism." In: Louis P. Pojman (ed.). *The Theory of Knowledge. Classical and Contemporary Readings*. Belmont: Wadsworth, pp. 373-384.
- -------. 2001a. *Knowledge, Truth, and Duty. Essays on Epistemic Justification, Responsibility, and Virtue*. Oxford: Oxford University Press.
- -------. 2001 b. "Epistemic Duty, Evidence, and Internality." In: Steup 2001a.
- Swain, Marshall.1981. *Reasons and Knowledge*. Ithaca: Cornell University Press.
- Zagzebski, Linda Trinkaus. 1996. *Virtues of the Mind. An Inquiry Into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press.
- -------. 1999. "What is Knowledge?" In: Greco 1999, pp. 92-116.

Other Internet Resources

- Keith DeRose's Epistemology Page

Related Entries

epistemology: naturalized | epistemology: social | epistemology: virtue | justification, epistemic: coherentist theories of | justification, epistemic: contextualist theories of | justification, epistemic: foundationalist theories of | justification, epistemic: internalist vs. externalist conceptions of

Acknowledgements

I wish to thank Laurence BonJour and Michael Bergman for helpful comments and criticisms.

Matthias Steup
steup@stcloudstate.edu

First published: February 5, 2001
Content last modified: February 5, 2001

Stanford Encyclopedia of Philosophy
Notes to The Analysis of Knowledge

Notes

1. See Radford 1966.

2. It is important not to misunderstand the justification condition. What it requires for knowledge is not that the subject have engaged in the *activity* of justifying the belief, or performed the act of showing (or trying to show) that the belief is justified. Rather, what the justification condition requires is merely that the belief that qualifies as knowledge have the *property* of being justified. It can have that property even if the subject has not in fact shown to anyone that the belief is justified. Consider an ordinary person's belief that five and five is ten. Most people have never attempted to justify this belief, and probably would be at a loss as to how to go about justifying it. But for most people, that belief would qualify as an instance of knowledge. The importance of the distinction between the activity of justifying and a belief's property of being justified is emphasized by William Alston in the following passage: "To turn to justification, the first point is that I will be working with the concept of a subject S's *being justified in believing that p*, rather than with the concept of S's *justifying* a belief. That is, I will be concerned with the state or conditionof being justified, rather than with the *activity* of justifying a belief. It is amazing how often these concepts are conflated in the literature. The crucial difference between them is that while to justify a belief is to marshal considerations in its support, in order for me to be justified in believing that *p* it is not necessary that I have done anything by way of an argument for *p* or for my epistemic situation vis-à-vis *p*. Unless I am justified in many beliefs without arguing for them, there is precious little I justifiably believe." Alston 1991, p. 71. For an alternative view, see Almeder 1999, pp. 90 and 123. Almeder defends the view that, "as a matter of ordinary discourse, 'being justified' is not something we can always separate from the activity of giving, or being able to give, reasons when the question 'How do you know?' is appropriately asked." Ibid, p. 92.

3. For a defense of evidentialism, see Feldman and Conee 1985. For criticisms of evidentialism, see De Rose 2000, and Plantinga 1996, pp. 358-361.

4. For a good discussion of objections to the belief condition, see Lehrer 1990, chapter 2.

5. Gettier 1963.

6. Gettier states explicitly the assumption that a justified belief can be false.

7. There are, therefore, two reasons why knowledge is not to be analyzed as true belief. First, a true belief

might fail to be justified. Second, a true belief might be justified but fail to be knowledge because the subject is in a Gettier-type situation.

8. See Chisholm 1977, chapter 6. Chisholm's strategy of building a degettierization clause into the justification condition is difficult to understand, given Chisholm's deep commitment to internalism. Since degettierization is an external matter, this strategy makes justification an external property. (See note 46) Thus, for internalists the place to take care of the Gettier problem is clearly a fourth condition.

9. See, for example, Armstrong 1973, p. 152, and Clark 1963. For further references, see Shope 1983, p. 24. This monograph provides a comprehensive discussion of the Gettier literature up to 1980. For a shorter but excellent discussion of the Gettier problem, see the Appendix in Pollock 1986.

10. See Goldman 1976.

For an example of a reliability condition amended so as to solve the Gettier problem, see Goldman 1976. See also Goldman 1986, pp. 46-7.

11. For examples of J-reliabilism, see: Goldman 1979, 1986, and Swain 1981.

12. For examples of K-reliabilism, see: Armstrong 1973, Dretske 1981, and Nozick 1981.

13. Dretske, 1989, p. 95.

14. Dretske 1985, p. 177.

15. Ibid, p. 179.

16. We might, therefore, distinguish between reliabilism, and reliabilism+, where the latter, unlike the former, involves a suitable degettierization clause. For an example of a reliability condition amended so as to solve the Gettier problem, see Goldman 1976. See also Goldman 1986, pp. 46-7. It should be noted, however, that the "+" in reliabilism+ need not take the form of a fourth condition. Rather, the "+" can simply be a suitable clause within the justification condition, where that condition is, of course, formulated in terms of reliability. (For an example of a reliability condition amended so as to solve the Gettier problem, see Goldman 1976. See also Goldman 1986, pp. 46-7.) But within an *internalist* JTB+ account of knowledge (to be discussed in the next section), the"+" demands a separate and fourth condition, since internal justification and degettierization differ in a crucial respect: unlike the former, the latter is an external affair. See note 46. Consequently, if we think of the "+" in JTB+ as a separate condition that cannot be mixed in with the justification condition, then the need to account for Gettier cases does not require of us to move from reliabilism to reliabilism+ in the same way as it requires of us to move from JTB to JTB+.

17. For literature discussing the internalism/externalism debate, see Alston 1989, pp. 185-226, BonJour 1985, chapter 3, Conee and Feldman, forthcoming, Goldman 1999, Fumerton 1995, Sosa 1999, Steup 1998, and Steup, 2001b.

18. Chisholm 1989, p. 7.

19. Chisholm 1977, p. 17.

20. Conee and Feldman (forthcoming) argue that internalism should be characterized in terms of mental states and events.

21. On the other hand, if mental states/events are not directly recognizable, the mental state criterion might not give internalists what they want. The problem here is that internalists would not want to count, for example, the state of being reliably clairvoyant as a mental state. If they did, they would not be in a position to deny that (unwittingly)reliable clairvoyance is a source of justification. Thus the need arises to differentiate between neurophysiological states that are, and those that are not, mental states. It might be that the best way to achieve this is to say that only directly recognizable neurophysiological states are mental states. But if the direct recognizability criterion is needed to define mental states, then mental state internalism does not appreciably differ from accessibility internalism.

22. If *S* is knocked out, or perhaps only sleeping, is he in a position to know the justificational status of any of her beliefs? Those who would say ‘no’ should reformulate the definition of J-internalism as follows: At any time *t* at which *S* is capable of thinking and holds a justified belief *B*, *S* is in a position to know at *t* that *B* is justified.

23. For a brief article on the concept of evidence, see Feldman 1992. See also Feldman 1988b.

24. For discussion of this issue, see Goldman 1999, and Steup 2001b.

25. The reason for this is that the kind of evidence that is relevant to a belief's justificational status must be evidence the subject possesses. But what brings evidence into the subject's possession? According to a strict view, only if it comes in the form of propositions that the subject believes. The problem with this view is its narrow scope. It excludes sensory experiences and memories from the kind of evidence that a subject possesses. Arguably, the best way to include them is to say this: evidence the subject possesses consists of the sorts of things that she can recognize on reflection, that is, that are directly recognizable to her.

26. It could be objected that an evidentialist would not have to be a K-internalist. The two theories are logically independent. Perhaps they are. It seems to me, however, that an evidentialist who doesn't take the possession of evidence to be a necessary condition of knowledge would be a strange bird indeed.

27. This view was held, for example, by Chisholm. In the second edition of his *Theory of Knowledge*, he wrote:

> We may assume that every person is subject to a purely intellectual requirement---that of trying his best to bring it about that, for every proposition *h* that he considers, he accepts *h* if and only if *h* is true. (1977, p. 14)

Laurence BonJour is another epistemologist who endorses a deontological conception of epistemic justification. He expresses it thus:

> The distinguishing characteristic of epistemic justification is thus its essential or internal relation to the cognitive goal of truth. It follows that one's cognitive endeavors are epistemically justified only if and to the extent that they are aimed at this goal, which means very roughly that one accepts all and only those beliefs which one has good reason to think are true. To accept a belief in the absence of such a reason, however appealing or even mandatory such acceptance might be from some other standpoint, is to neglect the pursuit of truth; such acceptance is, one might say, epistemically irresponsible. My contention here is that the idea of avoiding such irresponsibility, of being epistemically responsible in one's believings, is the core of the notion of epistemic justification. (1985, p. 8)

For further literature on the deontological conception of justification, see Alston 1989, pp. 81-152, Feldman 1988, and Steup 1996, chapter four.

28. For literature on the role of truth in epistemology, see Alston 1996, chapter 8, David 2001, and De Paul 2001.

29. Alternatively, evidentialists could drop the first part of this conjunction. Is it really my epistemic duty to believe *everything* that my evidence supports? It could be objected that it is hard to understand why it should be my epistemic duty to clutter up my belief system with trivial logical consequences of what my evidence supports. The reply to that would be that the consideration of clutter is not an epistemic, but a practical consideration. For discussion of this issue, see Feldman 1988.

30. For example, Goldman (1999) takes the rationale for internalism to rest on the deontological conception of justification as its main premise. For discussion of this issue, see Feldman forthcoming, and Steup 2001b.

31. The argument for (3) would go as follows: The concept of a duty has built into it an epistemic aspect. That by virtue of which a subject has a duty must, in the very least, be readily knowable, if not directly recognizable, to the subject. For otherwise, the could be such a thing as an unrecognizable duty, which is a conceptual impossibility. Critics of the argument displayed in the main text could argue that ready

knowability is less than direct recognizability, but certainly enough to put the concept of duty on a solid footing. Hard-boiled internalists would insist that ready knowability either amounts to direct recognizability, or isn't quite enough.

32. See the critique of externalism in chapter 3 of BonJour 1985.

33. I add "good reasons" to allow for the possibilities of internalists who are not evidentialists, and would characterize internal justification not in terms of not evidence but reasons.

34. The evil demon objection to reliabilism can be found in Cohen 1984, pp. 280-82, Foley 1985, Ginet 1985, and Lehrer 1990, p. 166.

35. Harman 1977, p. 17.

36. Harman 1984, p. 33. Harman 1984, p. 33.

37. At the beginning of his 1977, Goldman explicitly states his analytic goal: to give necessary and sufficient conditions of epistemic justification in non-epistemic terms.

38. See Chisholm 1989, p. 61f, where Chisholm makes it abundantly clear that he intends his "formal epistemic principles" to state what epistemic justification supervenes on. See also Steup 1996, chapter 9, where I argue that, as far as the analytic goal is concerned, there is surprising overlap between the internalist Chisholm and the externalist Goldman.

39. For recent literature on the naturalization of epistemology, see the Kornblith 1999 and Feldman 1999.

40. Thus Chisholm writes: "According to [the] traditional conception of "internal" epistemic justification, there is no *logical* connection between epistemic justification and truth. A belief may be internally justified and yet be *false*. This consequence is not acceptable to the externalist. He feels that an adequate account of epistemic justification should exhibit *some* logical connection between epistemic justification and truth." Chisholm 1989, p. 76f. See also Fumerton 1995, pp. 200-203.

41. An externalist alternative to reliabilism is Plantinga's proper functionalism. See Plantinga 1993a and 1993b.

42. It should be mentioned, however, that Alston also makes an effort to appreciate the appeal of internalism. See pp. 227-248 in his 1989.

43. Alston 1993, p. 9.

44. For a discussion of Dretske's question, see Almeder 1998, p. 132-136.

45. Alston, for example, endorses the view that animals and small children cannot have justified beliefs. He writes: "Lower animals, very small children, and idiots acquire and utilize much perceptual knowledge concerning the immediate environment; otherwise they would not be able to move around in it successfully. But they are not capable of acting in the light rules. So [justification as a normative property] is at best a necessary condition for the knowledge possessed by the likes of normal mature human beings." (1989, p. 173) I am inclined to concur with Alston. Although animals do of course have sensory experiences, I do not think that these experiences constitute *evidence* for them, for I take the concept of evidence to have a deontological aspect. A subject doesn't have evidence for *p* simply by virtue of being, for example, in a sensory state of the right sort. Rather, such a sensory state is evidence only if it *entitles* the subject, or makes it *permissible* for her, to believe that *p*. But animals are not subject to entitlements or permissions of that sort. I would, therefore, reject the suggestion that animals and little children can have evidence. For a contrary view, see Russell 2001.

46. Degettierization is an external matter because gettierization is an external matter. And gettierization is an external matter because a subject who is in a Gettier-type situation cannot tell on reflection (recognize directly) that she is in such a situation. On reflection, such a subject would have to come to the conclusion that she has knowledge.

47. I think it would be fair to say, however, that the subject matter of traditional epistemology has been IK rather than EK. Of course, this is not by itself a reason to prefer IK, since it might be a reflection of the shortcomings of traditional epistemology instead of an indication of the superiority of IK.

48. There are alternative approaches to the analysis of knowledge which I did not discuss in this article. First of all, I should mention Plantinga's proper functionalism, as represented by his 1993a and 1993b. I should also mention Linda Zagzebski's recent defense of virtue epistemology. Zagzebski 1996 and 1999. The virtue approach is also advocated in Goldman 1991, various papers in Sosa 1991, and Greco 1993.

Matthias Steup
steup@stcloudstate.edu

First published: February 5, 2001
Content last modified: February 5, 2001

Naturalized Epistemology

Naturalized epistemology is best seen as a cluster of views according to which epistemology is closely connected to natural science. Some advocates of naturalized epistemology emphasize methodological issues, arguing that epistemologists must make use of results from the sciences that study human reasoning in pursuing epistemological questions. The most extreme view along these lines recommends replacing traditional epistemology with the psychological study of how we reason. A more modest view recommends that philosophers make use of results from sciences studying cognition to resolve epistemological issues. A rather different form of naturalized epistemology is about the content of paradigmatically epistemological statements. Advocates of this kind of naturalized epistemology propose accounts of these statements entirely in terms of scientifically respectable objects and properties. In this they seem to contrast with more traditional epistemologists whose accounts make free use of evaluative terms such as "good reasons" and "adequate evidence". The significance of the claims of advocates of naturalized epistemology can best be appreciated by seeing them as a reaction to the methods and views that have been prominent in much of the twentieth century.

---

1. Background

A great deal of work in epistemology is devoted to two projects: replying to arguments for skepticism and analyzing key epistemological concepts. Although the discussion of skepticism goes back to ancient times, much of the contemporary thought on the topic is a reaction to Descartes. According to the Cartesian picture, if we have knowledge of the world around us, this knowledge is ultimately traceable to our knowledge of our own experiences. We know how we are experiencing the world, and somehow we reason from this knowledge out to the world. There is, however, a possibility of error in such reasoning.

Our experiences might result from dreams, hallucinations, or the manipulations of an evil demon or his modern counterpart, an evil neuroscientist artificially stimulating a bodiless brain in a vat. Arguments beginning with premises involving hypotheses such as these yield skeptical conclusions. Other skeptical arguments begin with premises about more routine sorts of errors and confusions. These arguments call into question apparent knowledge gained by ordinary people in ordinary circumstances as well as apparent knowledge gained through scientific inquiry. Epistemologists continue to try to identify precisely the structure of these skeptical arguments and the assumptions about knowledge and the world that they rely on. Traditional epistemologists, that is, those who do not advocate naturalizing epistemology, typically carry out this activity in their armchairs. They often see the scientific study of our cognitive systems as at best only distantly related to their effort to reply to skeptical arguments.

Arguments for skepticism rely, implicitly or explicitly, on assumptions about what is required for knowledge. Hence, consideration of those arguments invites analysis of what is necessary for knowledge. Widely shared views imply that for a person to know a proposition to be true, the person must have a well justified belief in the proposition and the proposition must in fact be true. Independent interest in what differentiates knowledge from mere true belief and shared opinion also generates interest in questions about the analysis of knowledge, as well as its key components. These analyses are intended to state necessary and sufficient conditions for the application of the concept. The details of the many analyses of knowledge epistemologists have offered are not crucial for present purposes, nor are the details of the varying accounts of justification that have occupied center stage in the epistemological literature. Section 4 will review the general differences between the sorts of analyses some naturalists propose and more traditional purportedly non-natural analyses. The central point to be reviewed in Sections 2 and 3 turns on the fact that the discussion of these analyses has typically gone on by means of reflection on possible cases. Epistemologists describe possible cases, consult their intuitions about whether the cases would be cases of knowledge or not, and decide on that basis whether or not the cases show the proposed analysis to fail. Once again, the task is carried out in the epistemologist's armchair, without the aid of science. It is this method of inquiry to which some naturalists object.

2. Replacement Naturalism

The source of much of the recent interest in Naturalized Epistemology is W.V.O Quine's celebrated essay, "Epistemology Naturalized" (Quine, 1969). Quine begins this essay by saying that "Epistemology is concerned with the foundations of science." In an effort to show that science has an adequate foundation, epistemologists attempted to derive statements about the world around us from statements about our own sensations. Given that we are certain about our own sensations, if we could strictly derive our beliefs about the world from our beliefs about sensations, we could then be certain of the derived truths about the world as well. We would then have a firm foundation for both everyday knowledge and scientific knowledge. Quine argues that such efforts to ground our beliefs about the world have failed. The proposed derivations just don't work. Virtually all contemporary philosophers agree. Quine concludes that the traditional effort to respond to skepticism is a failure and recommends what on the surface seems to be the abandonment of epistemology altogether. He apparently thinks that the failure of this sort of foundationalism shows that epistemology is impossible. He writes:

> The stimulation of his sensory receptors is all the evidence anybody has had to go on, ultimately, in arriving at his picture of the world. Why not just see how this construction really proceeds? Why not settle for psychology? (Quine, 1969: 75)

Quine seems to be recommending that we abandon the effort to show that we do in fact have knowledge and that we instead study the ways in which we form beliefs. His proposal is that we study the psychological processes that take us from sensory stimulations to beliefs about the world. He elaborates on this idea in a widely quoted passage:

> Epistemology, or something like it, simply falls into place as a chapter of psychology and hence of natural science. It studies a natural phenomenon, viz., a physical human subject. This human subject is accorded a certain experimentally controlled input -- certain patterns of irradiation in assorted frequencies, for instance -- and in the fullness of time the subject delivers as output a description of the three-dimensional external world and its history. The relation between the meager input and the torrential output is a relation that we are prompted to study for somewhat the same reasons that always prompted epistemology: namely, in order to see how evidence relates to theory, and in what ways one's theory of nature transcends any available evidence...But a conspicuous difference between old epistemology and the epistemological enterprise in this new psychological setting is that we can now make free use of empirical psychology. (Quine, 1969: 82-3)

As Jaegwon Kim points out in a widely cited critical discussion of Quine, another conspicuous difference between traditional epistemology and what Quine recommends is that they study strikingly different topics (Kim, 1988: 390). The old epistemology was interested in questions about rationality, justification, and knowledge. The central questions concerned whether an epistemic support relation--a justifying relation--holds between our basic evidence and our beliefs about the world. Analysis of some of the arguments for skepticism reveals that they rely on the view that our evidence supports our beliefs only if our beliefs are deducible from that evidence. Seeing that they are not, many epistemologists are drawn to investigate other accounts of the epistemic support relation, accounts that allow for the possibility that our beliefs about the world are well supported by our sensory evidence, even if they are not strictly derivable from that evidence. As Kim sees it, Quine has proposed ignoring these questions about epistemic support and investigating instead the causal connections between our sensory evidence and our beliefs about the world. Thus, if we follow the Quinean recommendation, we will study the same relata--our basic evidence and our beliefs about the world. However, we will study a different relation. In the original case, we looked to see if there was an epistemic support relation between the data and the beliefs. In the new case, we look to see the nature of the causal connection between them.

The Quinean view that we should abandon epistemology for psychology is not widely accepted by contemporary naturalists in epistemology. (See Almeder, 1998; BonJour, 1994; Foley, 1994; Fumerton, 1994.) Even Quine's later views were more moderate (Quine, 1990). Perhaps this is because questions about the quality of our reasons for our beliefs about the world strike even naturalists as perfectly good questions, questions deserving of investigation and analysis. Perhaps it is because new views about the

nature of knowledge and justification hold that they require the use of processes and methods that reliably lead to truth rather than recognizably good reasons. Whether we actually use such processes and methods seems to be a perfectly good question. In any case, Quinean Replacement Naturalism finds relatively few supporters.

One recent author who does defend a view close to Quine's is Hilary Kornblith. Kornblith contends that once traditional epistemologists admit that the Cartesian program of deriving beliefs about the world from certain foundations fails, they end up endorsing as legitimate whatever principles enable them to ratify the beliefs they started with. He writes,

> Of course knowledge is possible if we weaken the standards for knowledge far enough, in particular if we weaken them until we can show that many of our beliefs then pass the standards. But this seems to be nothing more than an exercise in self-congratulation. Why should we care about knowledge so defined? (Kornblith, 1999: 160)

He goes on to say,

> But if our standards for knowledge are merely designed to allow us to attach the epithet 'knowledge' to whatever it is we pretheoretically believe, then ... the result is an uncritical endorsement of the epistemological status quo. (Kornblith, 1999: 160)

Somewhat similar sentiments, though perhaps for different reasons, can be found in the writings of Stephen Stich. (Stich, 1990)

There is, however, a difference between the view that our pretheoretical beliefs are justified (or amount to knowledge), no matter what else is true of them, and the search for plausible general principles about knowledge and justification that have the result that those beliefs are justified. The principles philosophers put forward, including inference to the best explanation, principles about coherence and the conservation of belief, and so on, can be assessed and criticized. Few armchair epistemologists say that "whatever it is we pretheoretically believe" amounts to knowledge. Virtually all epistemologists agree that many everyday beliefs fall short of what's needed for justification and knowledge. Some conclude that knowledge is less common than one would initially think. By reflecting carefully on what they take to be realistic examples, they attempt to identify what is good about possible ways of reasoning. By calling our attention to the reasoning that withstands scrutiny and reflection, they can contribute to an effort to help us improve. So many defenders of traditional epistemology would deny that they are simply endorsing the status quo. (See Feldman, 1999)

3. Cooperative Naturalism

Although Quine's Replacement Naturalism is not widely accepted, a more modest descendant of his view is extremely popular. This view, Cooperative Naturalism, holds that, while there are evaluative questions

to pursue, empirical results from psychology concerning how we actually think and reason are essential or useful for making progress in addressing evaluative questions. A representative claim of this sort can be found in Susan Haack's *Evidence and Inquiry*, " ... the results from the sciences of cognition may be relevant to, and may be legitimately used in the resolution of traditional epistemological problems" (Haack, 1993: 118). Many philosophers who are more favorably disposed to naturalism than Haack is have voiced similar sentiments. (For example, Goldman, 1992; Stich and Nisbett, 1980: 118; Harman, 1986: 7; Kornblith, 1994: 7.)

What role empirical results can play in epistemology depends in large part on what counts as epistemology. Philip Kitcher's 1992 article, "The Naturalists Return," is a long and comprehensive study of naturalism in epistemology, arguing in part that the apsychologistic tendencies of the 20th century are in fact departures from what was standard in philosophy. Kitcher asks, "How could our psychological and biological capacities and limitations *fail* to be relevant to the study of human knowledge?" (Kitcher, 1992: 58) Obviously, empirical work is relevant to "the study of human knowledge." But this shows its relevance to epistemology only if epistemology is itself as broad as the study of human knowledge. If, however, there are specifically philosophical questions about knowledge, and epistemology is the study of those questions, then the relevance of empirical work about such things as our psychological and biological capacities remains open for debate. Disputes about which questions count as philosophical are likely to be futile. However, it is possible to examine fruitfully the merits of Cooperative Naturalism with respect to some of the issues typically addressed by epistemologists.

In thinking about the role of empirical information in epistemology it is helpful to keep in mind the fact that there are at least three possible views about the potential sources of information for epistemological theorizing, rather than the two that are sometimes identified. Some philosophers seem to be *a priorists*, in that they think that only what can be known *a priori* is relevant to epistemological questions. Other philosophers are *armchair epistemologists*, since they are willing to rely on common sense empirical knowledge -- what can be known from one's armchair, as well as *a priori* information. And *scientific epistemologists* proclaim the value of (or need for) the results from empirical studies for epistemology. This three way classification complicates the discussion of Cooperative Naturalism. If Cooperative Naturalism is the view that empirical information is important for resolving epistemological issues, then armchair epistemologists can accept it. However, if Cooperative Naturalism is the view that detailed information from the empirical sciences is important for epistemology, then armchair epistemologists are likely not to agree. There is no point in arguing about what Cooperative Naturalism "really" is. It will be enough to notice these different possibilities.

Traditional epistemologists often attempt to reply to arguments for skepticism without appeal to information available from the sciences. The arguments they consider typically include premises of two sorts. Premises of one sort specify some necessary condition for knowledge. Premises of the other sort say that people's beliefs never, or rarely, satisfy that necessary condition, or perhaps that they can't satisfy that condition. To the extent that an evaluation of the skeptical argument focuses on a premise of the first sort, *a priorists* will be often be in a position to carry out the task. A good analysis of knowledge will enable them to determine whether knowledge really does have the necessary condition the argument describes. Premises of this sort are in fact the focus of a good deal of the traditional debate. Thus, there

are debates about whether knowledge and justification require conclusive reasons or merely very strong reasons, whether they require reasons at all rather than mere reliability or causal connectedness, whether the fact that a belief provides a good explanation of some data, or is natural, or is widely accepted provide epistemic support for that belief. It is not clear that empirical studies of how we actually reason help with these debates. They can be carried out in the armchair.

Traditionalists often turn their attention to our knowledge in specific domains. They wonder whether, given the sort of basic evidence we have, can we know about other minds, about right and wrong, or about religious matters. They might assume, for example, that our evidence for our beliefs about the mental states of others consists primarily in our observations of their behavior. The key question is then whether that sort of evidence is good evidence for the sort of conclusion we tend to draw. Traditionalists seem to assume the general character of evidence we have for beliefs about other minds (or any of the other categories listed) and then to ask whether that sort of evidence is good enough to justify beliefs of the kind in question. It is not clear that detailed information from empirical studies of human reasoning is needed here. The key issue concerns what is needed for knowledge and whether the general sort of evidence we have - something we can identify from our armchairs - meets the relevant necessary condition for knowledge. Given what they have set out to do, it seems sensible for traditionalists to proceed without scientific input. But it is important to realize that even if traditionalists succeed in refuting arguments for skepticism, they will not thereby show that we do have knowledge. They will only have refuted arguments for the denial of that claim.

It is possible that naturalists tend to focus on different questions: Can we show that we do have knowledge in one area or another? In the various areas do we in fact tend to draw the right conclusions from the evidence we do have? Are the processes we actually use reliable ones? Quine seemed to take the question to be whether our actual scientific beliefs had a firm foundation. The goal was to "reconstruct" our knowledge. There is no doubt that answering these questions requires empirical input. While we can from our armchairs have some ideas about the sorts of inferences we make and the reliability of the processes that occur in us, detailed scientific information is needed to have a clear picture of our actual practices. Furthermore, information about the sorts of errors and mistakes we are apt to make about specific topics is vital to assessing the merits of our actual beliefs about those topics.

Claims to the effect that actual people know actual facts about the world are contingent propositions about the world. They cannot be known a priori. Perhaps such things can be known from our armchairs, since we can know quite a bit from our armchairs. It is difficult to see why we cannot know some things about knowledge from our armchairs. Still, information from empirical sciences cannot be irrelevant to issues about what people actually know. We might plausibly judge from our armchairs that we have knowledge in some range of cases. It is possible that cognitive science will discover that in some or all of these cases our beliefs result from bizarre, thoroughly unreliable, deviant causal chains. If that were to happen, we might learn that we lack knowledge in cases where we thought we had it. Though the possibility that we will learn such things in a variety of familiar cases is extremely small, it is not zero. So empirical results could overturn our judgments about these cases. Thus, if it is epistemology's business to make judgments about whether actual people have knowledge in actual cases, and Cooperative Naturalism is the view that empirical information from the natural sciences is potentially

relevant to those judgments, then Cooperative Naturalism is unquestionably true. What would be remarkable about this is that anyone ever denied it.

Some traditional epistemologists give the impression that they simply assume that we do have knowledge and that no empirical information can overturn that judgment. The claim, or assumption, they make is that we know pretty much what we think we know. A representative formulation of this approach can be found in John Pollock's *Contemporary Theories of Knowledge*. Pollock writes:

> In typical skeptical arguments, we invariably find that we are more certain of the of the knowledge seemingly denied us than we are of some of the premises. Thus it is not reasonable to adopt the skeptical conclusion that we do not have that knowledge. The rational stance is instead to deny one or more of the premises. (Pollock, 1986: 6)

Pollock goes on to say that contemporary epistemologists largely see figuring out what knowledge is as their goal, rather than refuting skepticism. As he puts it, we "need not refute the skeptic - we already know that the skeptic is wrong" (Pollock, 1986: 6). This view, of course, has its similarities to views endorsed by G. E. Moore, and it is by no means uncommon. More generally, many epistemologists proceed on the assumption that we do know pretty much what we think we know, and thus on the assumption that skepticism is false. It may seem that this position, or this attitude, rules out the possibility that empirical results will overturn their philosophical starting point. (For discussion, see Kornblith, 1988.)

It is not clear, however, that the philosophers who take this position really do need to say anything quite so strong. What they may be committed to, in fact, is the more modest claim that no abstract philosophical claim is initially more plausible than the claim that we do have knowledge in a typical range of actual cases. They need not be committed to the more extreme, and quite implausible, claim that no empirical results could possibly show that we lack knowledge in particular actual cases. Surely empirical results could show that any contingent claim to knowledge that we make is false, either by showing that the thing we claim to know is false or by showing that our belief in that fact arose in some untoward way. The Moorean assumption is better taken to be a position toward highly abstract philosophical arguments for skepticism, typically arguments that imply that we lack knowledge no matter what the facts in the world are. It may be a mistake to reject these arguments out of hand. But whatever we say about such arguments, it would be mistake to attribute to Pollock, Moore, and others who share their views the idea that specific claims about what we know cannot be overthrown on the basis of empirical information. Nothing about their philosophical positions requires them to maintain that extreme view.

On a related point, Philip Kitcher takes tradtionalists to hold that "our favored logical principles are prescriptions for thought" (Kitcher, 1992: 63). But he thinks that the mere fact that we favor certain logical principles is of no value in establishing that they are meritorious principles. Our principles are good ones only if they actually do enable us to attain our epistemic ends. "Simply asserting that [certain rules] unfold our conception of rationality will be beside the crucial point" (Kitcher, 1992: 63). What

needs to be determined is whether our principles actually get us to the truth, and empirical inquiry is needed to find that out. So, as he sees it, traditionalists simply endorse the principles we favor whereas naturalists are willing to put those principles to empirical tests.

Kitcher illustrates his point by means of Hume's problem of induction. Hume famously asked whether we have any good reason to believe the conclusions of our inductive arguments. We notice that all observed instances of some sort of object have had a certain property and we infer that the next object of that kind will have that property. Our premise does not entail our conclusion and it turns out that it is extraordinarily difficult to justify these inductive inferences in a way that does not illicitly rely on induction itself. One solution to the problem, associated with Peter Strawson (1952), is that "adopting the inductive practices and principles that we do is constitutive of our concept of rationality." But, Kitcher asks, "why should we treat our current concept of rationality as privileged?" (Kitcher, 1992: 63). After all, rival societies might have rival conceptions. Anti-inductivists could proclaim their practices rational because they are constitutive of their conception of rationality. As Stephen Stich asks, "Why should we care one whit whether the cognitive processes we use are sanctioned by [our] evaluative concepts?" (Stich, 1990: 92).

Once again, it is possible that the debate between naturalists and traditionalists results in part from an emphasis on different questions. One way to view the issue about induction that Hume raised makes it a very general issue about whether it is ever reasonable to use past, or unobserved, cases as the basis for beliefs about future, or observed cases. The Strawsonian view he mentions says that it is, and defends this on the basis of claims about our concept of rationality. Kitcher takes this to be a defense of our adoption of "the inductive practices and principles" that we actually use. If the principles to which he refers are more specific principles licensing particular inductive inferences, then it may be that they are contingent principles that cannot be justified by *a priori* or even armchair methods. Kitcher is right to object to philosophers who contend that these lower level principles are necessary consequences of our concept of rationality may be mistaken. Still, Strawson may have been right about the more general issue about the rationality of using past cases as the basis for belief about future cases. Analogously, there may be some general issues about the rationality of reliance on testimony, but claims about whether particular sources are trustworthy require empirical defense.

4. Substantive Naturalism

Paradigmatic epistemological facts are expressed by sentences using terms such as those on the following list provided by Alvin Goldman:

> justified,
> warranted,
> has (good) grounds,
> has reason (to believe),
> knows that,
> sees that,

> apprehends that,
> is probable (in an epistemic or inductive sense),
> shows that,
> establishes that
> ascertains that.
> (Goldman, 1979: 1-2)

The crucial thing about sentences using these terms are that they seem to do more than merely describe how things are. They say or imply how something is to be evaluated from an epistemological perspective. Traditional epistemologists take these evaluative epistemological sentences to be objectively true or false, and thus they are committed to there being epistemological facts. The status and nature of these facts constitutes a second major issue falling under the heading of epistemological naturalism.

In a discussion of naturalism in ethics Gilbert Harman writes:

> [Ethical naturalism] is the doctrine that moral facts are facts of nature. Naturalism as a general view is the sensible thesis that *all* facts are facts of nature. (Harman, 1977: 17)

Substantive epistemological naturalism is the view that all epistemic facts are natural facts.

This is not informative unless it is supplemented with some account of what counts as a natural fact. One view is that the natural facts include all the facts that a complete science will acknowledge. Another way to characterize the natural facts is to provide a list of representative examples of the sorts of things that count as natural, with the hope that we have at least a reasonably good idea of what else might be included. The two approaches are not incompatible, since the examples might be a list of the sorts of facts science acknowledges. One such list was produced by Alvin Goldman in his classic paper, "What is Justified Belief?" (Goldman, 1979). Though Goldman was not explicitly discussing naturalism in this paper, the things he mentions will serve present purposes quite well. Goldman suggests that the following are non-epistemic terms:

> believes that,
> is true,
> causes,
> it is necessary that,
> implies,
> is deducible from
> is probable (either in the frequency sense or the propensity sense)
> (Goldman, 1979: 2)

He says:

> In general, (purely) doxastic, metaphysical, modal, semantic, or syntactic expressions are

not epistemic. (Goldman, 1979: 2)

Naturalized epistemologists contend that if epistemic terms make sense at all, they must be understood in terms of items such as those on this list. Their view seems to contrast with that of traditional epistemologists, who are content to formulate their analyses using evaluative epistemic terms.

It is possible for naturalists to deny that epistemic sentences report facts at all. Defenders of such a view might contend that the sentences are sheer nonsense. More plausibly, they might argue that they are meaningful but non-factual, perhaps because they are expressions of approval or disapproval of the beliefs and believers or the acts and actors mentioned. There have been some defenses of this view, though more frequently in ethics than in epistemology. (See Gibbard, 1990)

Another approach for naturalists is to argue that epistemic terms can be given naturalistic definitions. For present purposes it will suffice to take a definition of a term to be a statement of logically necessary and sufficient conditions for its application. Familiar accounts of epistemic terms seem to be divisible into those that employ only clearly naturalistic terms and those that do not. Many traditional definitions of epistemic justification make essential use of other evaluative epistemic terms. Thus, it is common to define justification in terms of good reasons, adequate evidence, strong grounds, the right to be sure, and the like. For example, evidentialism holds that a person is justified in believing a proposition at a time if and only if the evidence the person has at that time supports believing that proposition (rather than disbelieving it or suspending judgment about it). This definition appeals to two crucial elements: the evidence a person has at a time and the support relation that can hold between a body of evidence and an attitude toward a proposition.

There is no reason for naturalists to worry about the naturalistic credentials of the idea of evidence possessed. This is not to say that the concept is entirely clear. But there is nothing metaphysically spooky, or even evaluative, about it. The evidence one possesses is some combination or other of the experiences one is having, the memories one has, and the other beliefs one has. It may be that other beliefs count as evidence only if they themselves are justified, but this at most shows that the account of justification will turn out to be recursive. It does not show that this element of the evidentialist definition invokes anything non-natural.

The second component of the evidentialist definition of justification is the idea of evidential support. It is easy enough to give examples of cases in which this relation is supposed to obtain and other cases in which it does not obtain. There may also be some cases that are difficult to assess. Despite this, its naturalistic credentials are in doubt. Evidential support seems to be precisely the sort of relation whose connection to the scientifically identifiable objects and properties is difficult to ascertain. Similar worries apply to accounts of epistemic justification in terms of epistemic duties or rights or in terms of epistemically responsible behavior.

In contrast to evidentialism are causal and reliabilist accounts. The simplest version of the causal theory says that a belief that p is justified when the fact that p is causally connected to the belief that p. This

theory invokes facts, beliefs, and causal connections, all terms acceptable to naturalists. Reliabilism also seems to invoke only naturalistically respectable terms. In its simplest form it holds that a belief is justified provided it is produced by a belief-forming process that reliably leads to true beliefs. More complex forms of reliabilism also appear to be naturalistically acceptable.

These considerations suggest that one aspect of the debate between naturalists and non-naturalists is best understood as a debate about whether knowledge and justification can be understood in naturalistically acceptable causal and reliabilist terms or must be understood in terms of naturalistically suspect evaluative terms. What remains unclear, however, is whether philosophers who defend traditional accounts of justification, such as evidentialism, are committed to any troubling form of non-naturalism.

Defenders of evidentialist accounts are not (or at least need not be) committed to the idea that no naturalistic definitions of the terms they employ in their definitions are possible. It may be that they just have not produced such definitions yet. And even if the terms are not strictly definable, it does not follow that they are not themselves perfectly good naturalistic terms. This is because it remains possible that evaluative epistemic facts *supervene* on naturalistic ones. To say that the evaluative facts supervene on the natural ones is to say that in any two worlds in which all the natural facts are alike, all the evaluative facts are also alike. Alternatively, one might say that facts about the epistemic status of beliefs supervene on natural facts provided that, necessarily, if two believers share all the same natural properties, then same beliefs are justified for them. If this supervenience thesis is true, then, arguably, the epistemic facts depend on the natural facts and, perhaps, they are therefore natural facts.

The supervenience thesis is endorsed by a great many philosophers. A representative assertion of it is:

> ...if a belief is justified, that must be so *because* it has certain factual, non-epistemic properties...That it is a justified belief cannot be a brute fundamental fact... [it] must be grounded in the factual descriptive properties of that particular belief. (Kim, 1988: 399)

This quotation comes from Kim's critical discussion of Quine's "Naturalized Epistemology." One can find similar passages in the writings of philosophers, including many who would not typically be regarded as naturalists. (Chisholm, 1982: 12; Van Cleve, 1985: 97-101). Indeed, there are very few who deny the supervenience thesis. One who does is Keith Lehrer (Lehrer, 1997).

So, the supervenience thesis is widely accepted. There is, of course, considerable disagreement about which facts are central to the supervenience base for epistemic facts. *Evidentialism* holds that the key natural facts that determine whether a belief is justified are facts about the evidence the person has for that belief. The evidence one has is some combination or other of the experiences one is having, the memories one has, and the other beliefs one has. All of these are unquestionably natural facts about a person. And evidentialism holds that necessarily, people who have the same evidence are justified in believing the same things. In other words, the theory is that natural facts about evidence possessed determine epistemic facts. *Reliabilism* holds that the crucial facts in the supervenience base of epistemic facts are facts about the reliability of the causal produce producing or sustaining the belief. These too are

unquestionably natural facts.

If supervening on natural facts is sufficient for making a fact natural, then it follows that it is widely agreed that epistemic facts are natural facts. It difficult to determine whether supervening on what is natural is sufficient for naturalness, but things participants to the debate say can help to clarify the issues. In an extensive review of naturalism in epistemology, James Maffie writes that a key claim of naturalism is that:

> epistemic value is anchored to descriptive fact, no longer entering the world autonomously as brute, fundamental fact...(Maffie, 1990: 284)

One can find comparable claims in the writings of many others. (See Steup, 1996: 185-6; Lycan, 1988: 122) If epistemic facts supervene on unquestionably natural facts, then they do not float free, they are not autonomous, they are not brute facts, they are anchored in the natural world. That seems like a good reason to conclude that they are natural facts.

Of course, some naturalists may contend that a substantive naturalist view must treat epistemic facts as supervening on causal rather than logical facts. (See Kitcher, 1992) Perhaps some support for rejecting the view that anything that supervenes on what's natural is itself natural comes from the fact that this thesis yields the surprising result that the famous ethical non-naturalists were actually naturalists. For example, G. E. Moore would have endorsed the supervenience thesis. However, Kim says that we use the term "naturalism" ambiguously in "ethical naturalism" and "epistemological naturalism". The former requires definitions in natural terms. The latter requires only supervenience. (Kim, 1988: 398) So Kim's view seems to be that, so far as the debate about naturalistic epistemology goes, epistemic facts are natural facts if they supervene on unquestionably natural facts. His view, like nearly all participants to the debate, is that they do.

Even if supervenience of the sort just discussed shows that many epistemic facts are natural facts, a question remains for some traditionalists. In addition to facts about particular people being justified in believing particular propositions, they are committed to the existence of epistemic facts about what beliefs are supported by a particular body of evidence. It remains unclear whether these are natural facts. Traditionalists often regard these facts as necessary truths, and it is their necessity that enables evidentialists to endorse the supervenience thesis. The sentences expressing these epistemic relations express facts - call them *epistemic support* facts. It is legitimate to ask whether they count as natural facts.

If the epistemic support facts are not natural facts, then not all epistemic facts are natural facts and, according to traditionalists, substantive epistemological naturalism is false. If the epistemic support facts are natural facts, and justification is defined in terms of evidence possessed and epistemic support, then justification is defined in entirely natural terms. In that case, evidentialists do not need to rely on supervenience to defend naturalism. (If reliabilists admit that there are epistemic support facts, then they too must account for their status.)

Whether epistemic support facts are natural facts is a difficult and unexplored question. Epistemic support facts are, on many views, necessary truths. On standard definitions of supervenience, it is trivial that necessary truths supervene on natural facts. (They supervene on anything.) Given that supervening on what is natural is sufficient for being natural, it follows that they are natural facts. However, this response seems to avoid the central questions. This is because some traditionalists, notably Roderick Chisholm, assert the existence of a special class of epistemic support facts whose naturalistic status is questionable. Chisholm contends that "there are principles of evidence other than the formal principles of deductive logic and inductive logic" (Chisholm, 1977: 67). For example, Chisholm held that there were special principles about perceptual evidence along the lines of:

> (**R**). Being in the state of seeming to see something red (being appeared to redly) is evidence for the proposition that one really does see something red.

The evidential support described in (R) is, of course, defeasible. One could have evidence that one is not really seeing something red in spite of being appeared to redly. The key thing about (R) is that it is not, or at least not obviously, an instance of some general deductive or probabilistic principle. It is not a principle of logic. It is, instead, a fundamental principle of epistemology. Thus, at least some evidentialists seem to be committed to there being fundamental epistemic principles, or evidential support relations, that differ from any deductive or probabilistic relations and cannot be defined in terms of any complex of psychological relations and familiar logical relations. Naturalists are apt to find this all highly suspect.

Some traditionalists go further. They hold that to be justified in a belief not only must we have evidence that supports the belief, but we must "grasp" the connection between the evidence and the proposition believed (Fumerton, 1995: 183-224) This grasping of evidential relations may strike some naturalists as exceedingly dubious. However, it is best not to see the issue here as one about naturalism itself. If traditionalists say that there is a special sort of grasping or understanding or acquaintance with epistemic support facts and naturalists deny that there is any such thing, it is seriously misleading to attribute some sort of non-naturalism to the traditionalists. They are making a claim about what natural relations there actually are. Their view is that acquaintance is a particularly fundamental sort of psychological, hence natural, relation. Possibly some naturalists will deny that any such grasping of connections occurs, but the issue here is something like an issue that arises in discussions of extra-sensory perception. Clear-headed defenders of the view that we are capable of perceiving things that ordinary psychology would declare us incapable of perceiving differ with their opponents over what natural relations (or abilities) we have. While some may try to cast their view in mystical language, there is no reason to invoke non-naturalism here. If a hard-headed naturalist became convinced that some people did have some abilities previously thought to be beyond us, the response should be a change in views about what nature is like, not a renunciation of naturalism.

Still, there remains the question of the status of facts like (R). As noted, Chisholm regarded it as a special epistemic principle. But not all traditionalists must follow suit. They might argue that these principles are special cases of naturalistically acceptable general principles. For example, one could argue that (R)

follows from some general principle about best explanations. Or perhaps it really does have inductive support. Support from appeals to conservatism (what we already believe), what we naturally believe, what is generally accepted, and other factors are possible. Philosophers who endorse any such view can proudly proclaim themselves substantive naturalists.

Chisholm, of course, would not accept this. He thought that there were fundamental epistemic principles that could not be explained in any such terms. Perhaps this counts as non-naturalism. Even here, however, the issues are less than fully clear. Chisholm apparently thought that, in addition to deductive and probabilistic connections, there was another species of connection between propositions (or between experiences and propositions). His view was that these relations are part of the real, or natural, world. Some may deny that there are any such relations. This seems, once again, to be a dispute about what there is, not a dispute about whether there is something beyond what is natural. In other words, if Chisholm is right, it is quite unclear why terms such as "supports" and other epistemic terms do not belong on the list of naturalistically acceptable terms in the first place. If they do, then even Chisholm can plausibly maintain that epistemic support facts are natural facts. If so, then almost all epistemologists are substantive naturalists.

Bibliography

- Almeder, Robert (1998) *Harmless Naturalism: The Limits of Science and the Nature of Philosophy*, Peru, Illinois: Open Court.
- BonJour, Laurence (1994) "Against Naturalized Epistemology," *Midwest Studies in Philosophy*, XIX: 283-300.
- Chisholm, Roderick (1966)*Theory of Knowledge*, Englewood Cliffs, NJ: Prentice-Hall.
- Chisholm, Roderick (1982) *The Foundations of Knowing*, Minneapolis: University of Minnesota Press.
- Chisholm, Roderick (1989)*Theory of Knowledge*, 3rd ed., Englewood Cliffs, NJ: Prentice-Hall.
- Feldman, Richard (1999), "Methodological Naturalism in Epistemology," in *The Blackwell Guide to Epistemology*, edited by John Greco and Ernest Sosa, Malden, Ma: Blackwell, pp. 170-186.
- Foley, Richard (1994) "Quine and Naturalized Epistemology," *Midwest Studies in Philosophy*, XIX: 243-260.
- Fumerton, Richard (1994) "Skepticism and Naturalistic Epistemology," *Midwest Studies in Philosophy*, XIX: 321-340.
- Fumerton, Richard (1995) *Metaepistemology and Skepticism*, Lanham, MD: Rowman and Littlefield.
- Gibbard, Allan (1990) *Wise Feelings, Apt Choices*, Cambridge: Harvard University Press.
- Goldman, Alvin (1979) "What is Justified Belief?," in G. Pappas, ed., *Justification and Knowledge: New Studies in Epistemology*, Dordrecht, Reidel: 1-23.
- Goldman, Alvin (1992), *Liaisons: Philosophy Meets the Cognitive and Social Sciences*, Cambridge: MIT Press.
- Haack, Susan (1993) *Evidence and Inquiry: Towards Reconstruction in Epistemology*, Oxford: Blackwell.

- Harman, Gilbert (1977) *Thought*, Princeton: Princeton University Press.
- Kim, Jaegwon (1988) "What is Naturalized Epistemology?" *Philosophical Perspectives* 2 edited by James E. Tomberlin, Asascadero, CA: Ridgeview Publishing Co: 381-406.
- Kitcher, Philip (1992) "The Naturalists Return," *Philosophical Review*, 101: 53-114.
- Kornblith, Hilary (1994) *Naturalizing Epistemology* 2nd Edition, Cambridge: MIT Press.
- Kornblith, Hilary (1999) "In Defense of a Naturalized Epistemology" in *The Blackwell Guide to Epistemology*, edited by John Greco and Ernest Sosa, Malden, Ma: Blackwell, pp. 158-169.
- Kornblith, Hilary (1988) "How Internal Can You Get?," *Synthese*, 74: 313-327.
- Lehrer, Keith (1997) *Self-Trust: A study of Reason, Knowledge and Autonomy*, Oxford: Clarendon Press.
- Lycan, William (1988) *Judgement and Justification*, Cambridge: Cambridge University Press.
- Mafffie, James (1990) "Recent Work on Naturalizing Epistemology," *American Philosophical Quarterly* 27: 281-293.
- Pollock, John (1986) *Contemporary Theories of Knowledge*, Totawa, NJ: Rowman and Littlefield.
- Quine, W.V.O. (1969) *Ontological Relativity and Other Essays*, New York: Columbia University Press.
- Quine, W.V.O. (1990) "Norms and Aims" in *The Pursuit of Truth*, Cambridge: Harvard University Press.
- Steup, Matthias, *An Introduction to Contemporary Epistemology*, Prentice-Hall, 1996.
- Stich, Stephen and Richard Nisbett (1980), "Justification and the Psychology of Human Reasoning," *Philosophy of Science* 47: 188-202.
- Stich, Stephen (1990) *The Fragmentation of Reason*, Cambridge, MA: MIT Press.
- Strawson, Peter (1952) *Introduction to Logical Theory*, New York: Wiley.
- van Cleve, James (1985) "Epistemic Supervenience and the Circle of Belief" *Monist* 68: 90-104.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

epistemology: social | justification, epistemic: internalist vs. externalist conceptions of

Richard Feldman
feldman@philosophy.rochester.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 5, 2001
Content last modified: July 5, 2001

Stanford Encyclopedia of Philosophy

Social Epistemology

Social epistemology is the study of the social dimensions of knowledge or information. There is little consensus, however, on what the term "knowledge" comprehends, what is the scope of the "social", or what the style or purpose of the study should be. According to some writers, social epistemology should retain the same general mission as classical epistemology, revamped in the recognition that classical epistemology was too individualistic. According to other writers, social epistemology should be a more radical departure from classical epistemology, a successor discipline that would replace epistemology as traditionally conceived. The classical approach could be realized in at least two forms. One would emphasize the traditional epistemic goal of acquiring true beliefs. It would study social practices in terms of their impact on the truth-values of agents' beliefs. A second version of the classical approach would focus on the epistemic goal of having justified or rational beliefs. Applied to the social realm, it might concentrate, for example, on when a cognitive agent is justified or warranted in accepting the statements and opinions of others. Proponents of the anti-classical approach have little or no use for concepts like truth and justification. In addressing the social dimensions of knowledge, they understand "knowledge" as simply what is believed, or what beliefs are "institutionalized" in this or that community, culture, or context. They seek to identify the social forces and influences responsible for knowledge production so conceived. Social epistemology is theoretically significant because of the central role of society in the knowledge-forming process. It also has practical importance because of its possible role in the redesign of information-related social institutions.

1. History of Social Epistemology

The phrase "social epistemology" does not have a long history of systematic use. It is not difficult, however, to find historical philosophers who made at least brief forays into the social dimensions of knowledge or rational belief. In his dialogue *Charmides*, Plato posed the question of how a layperson can determine whether someone who purports to be an expert in an area really is one. Since dependence on experts or authorities is a problem within the scope of social epistemology, this was a mini-exploration of the subject. The seventeenth and eighteenth century British philosophers John Locke, David Hume, and Thomas Reid devoted portions of their epistemologies -- often just scattered remarks -- to the problem of "testimony": When should cognitive agents rely on the opinions and reports of others? What must a hearer know about a speaker to be entitled to trust his assertions? Locke so emphasized the importance of intellectual self-reliance that he expressed strong doubts about giving authority to the opinions of others (1959, I. iii. 23). Hume took it for granted that we regularly rely on the factual statements of others, but insisted that it is reasonable to do so only to the extent that we have adequate reasons for thinking that these sources are reliable. Hume's empiricism led him to require that these reasons be based on personal observations that establish the veracity of human testimony (Hume 1975, X, 111). Reid, by contrast, claimed that our natural attitude of trusting others is reasonable even if we know little if anything about their reliability. Testimony, at least sincere testimony, is always *prima facie* credible (Reid 1975, VI, xxiv). All of these positions, of course, are epistemological positions. However, they were generally part of an epistemological enterprise that was basically egocentric in orientation, so they are perhaps not ideal or pure paradigms of social epistemology. Nonetheless, they are clear examples of early epistemologies that examined social dimensions of epistemic justification.

A different tradition focused on aspects of knowledge that are "social" in a more sociological or political sense, though members of this tradition less frequently aligned their work to core issues in epistemology. Karl Marx's theory of ideology could well be considered a type of social epistemology. On one interpretation of Marx's conception of "ideology", an ideology is a set of beliefs, a world-view, or a form of consciousness that is in some fashion false or delusive. The cause of these beliefs, and perhaps of their delusiveness, is the social situation and interests of the believers. Since the theory of ideology, so described, is concerned with the truth and falsity of beliefs, it might even be considered a form of *classical* social epistemology.

Karl Mannheim (1936) extended Marx's theory of ideology into a sociology of knowledge. He classed forms of consciousness as ideological when the thoughts of a social group can be traced to the group's social situation or "life conditions" (1936: 78). The descriptive enterprise of tracing these thoughts to the social situation might be construed as social epistemology. The further enterprise of critiquing and dissolving ideological delusions --"Ideologiekritik" -- is surely a form of social epistemology. The critical theory of the Frankfurt School was one attempt, or a family of attempts, to develop this idea. Critical theory aims at emancipation and enlightenment by making agents aware of hidden coercion in their environment, enabling them to determine where their true interests lie (Geuss 1981: 54). In a variant of critical theory, Jurgen Habermas introduced the idea of an "ideal speech situation", a hypothetical situation of absolutely uncoerced and unlimited discussion between completely free and equal human agents (Habermas 1973; Geuss 1981: 65). In some writings Habermas uses the ideal speech situation as a transcendental criterion of truth. Beliefs that agents would agree upon in the ideal speech situation are ipso facto true beliefs (Habermas and Luhmann 1971: 139, 224). Here a social communicational device

is treated as a type of epistemic standard.

Subsequent developments in the sociology of knowledge, and especially in the sociology of science, can also be considered forms of social epistemology. Since science is widely considered the paradigmatic knowledge-producing enterprise, and since epistemology is centrally concerned with knowledge, any endeavor that seeks to identify social determinants of science might plausibly be categorized as a form of social epistemology. Both Mannheim and the sociologist of science Robert Merton (1973) exempted (natural) science from the influence of societal or "existential" factors of the types that influence other categories of beliefs. Science was viewed as a society unto itself, largely autonomous from the rest of society. But later sociologists of science have declined to offer the same exemption. The Edinburgh School contends that all scientific beliefs are on a par with other beliefs in terms of their causes. Barry Barnes and David Bloor formulated a "symmetry" or "equivalence" postulate, according to which all beliefs are on a par with respect to the causes of their credibility (1982). Many historical case studies conducted in this tradition have tried to show how scientists too are swayed by class interests, political interests, and other factors usually considered "external" to pure science (Forman 1971; Shapin 1975; Mackenzie 1981). Thomas Kuhn (1962/1970) is thought to have shown that purely objective considerations can never settle disputes between competing scientific theories or paradigms, and hence scientific beliefs must be influenced by "social factors". Kuhn's descriptions of the practices of scientific research communities, especially descriptions of the inculcation and preservation of paradigms during periods of "normal" science, were clear and influential examples of a social analysis of science, especially when contrasted with the positivist tradition of analysis. Michel Foucault developed a radically political view of knowledge and science, arguing that practices of so-called knowledge-seeking, especially in the modern world, really serve the aims of power and social domination (1977, 1980). All of these writers may be considered "social epistemologists", although they themselves do not employ this phrase.

Perhaps the first use of the phrase "social epistemology" appears in the writings of a library scientist, Jesse Shera, who in turn credits his associate Margaret Egan. "[S]ocial epistemology," says Shera, "is the study of knowledge in society.... The focus of this discipline should be upon the production, flow, integration, and consumption of all forms of communicated thought throughout the entire social fabric" (1970: 86). Shera was particularly interested in the affinity between social epistemology and librarianship. He did not, however, construct a conception of social epistemology with very definite philosophical or social-scientific contours. What might such contours be?

2. Classical Approaches

Classical epistemology has been concerned with the pursuit of truth. How can an individual engage in cognitive activity so as to arrive at true belief and avoid false belief? This was the task Rene Descartes set for himself in his *Discourse on the Method of Rightly Conducting the Reason and Seeking for Truth in the Sciences* (1637/1955) and in his *Meditations on First Philosophy* (1641/1955). Classical epistemology has equally been concerned with rationality or epistemic justification, as suggested by part of the title of the *Discourse*. A person might rightly conduct her reason in the search for truth but not

succeed in getting the truth. However, as long as she forms a belief by a proper use of reason -- and perhaps by proper use of other faculties like perception and memory -- then her belief is rationally warranted or justified. Classical epistemologists all regard this as one sort of epistemic desideratum. Furthermore, according to the standard account of knowledge in classical epistemology, for a person to *know* a proposition, she must believe it, it must be true, and the belief in it must be justified or rationally warranted. So if epistemology is the study of knowledge, and more specifically the study of how knowledge can be attained, it must also be the study of how true and justified belief can be attained. Epistemological projects restricted to just one of these dimensions -- truth or justification -- would also fit the classical mold.

The foregoing remarks apply to classical epistemology in its "individualist" guise. What type of epistemology does one get if one tries to "socialize" classical epistemology? One gets some sort of social angle on the pursuit of true belief and/or the pursuit of justified belief. Some projects in social epistemology have adopted precisely these themes.

Perhaps the first formulation of a truth-oriented social epistemology is found in writings by Alvin Goldman from the late 1970s through the mid-1980s (Goldman 1978, 1986, 1987). Goldman there proposes to divide epistemology into two branches: individual epistemology and social epistemology (or "epistemics"). Both branches would seek to identify and assess processes, methods or practices in terms of their contributions -- positive or negative -- to the production of true belief. Individual epistemology would identify and evaluate psychological processes that occur within the epistemic subject. Social epistemology would identify and evaluate social processes by which epistemic subjects interact with other agents who exert causal influence on their beliefs. The communicational acts of other agents and the institutional structures that guide or frame such communicational acts would be prime examples of social-epistemic practices that would be studied within social epistemology. In Goldman's subsequent book, *Knowledge in a Social World* (1999), this conception of social epistemology is developed in detail. It is argued that, both in everyday life and in specialized arenas such as science, law, and education, a certain value is placed on having true beliefs rather than false beliefs or no opinion (uncertainty). This type of value is called "veritistic value", and a measure of veritistic value is proposed. The rest of the book examines types of social practices that make positive or negative contributions toward increasing veritistic value. Types of practices examined include speech practices of reporting and arguing, market and non-market mechanisms that regulate the flow of speech, types of information technologies, assigning scientific credit and guiding scientific inquiries with an eye to credit, trial procedures or legal adjudication systems, and systems that disseminate political information about electoral candidates.

The veritistic approach to social epistemology aims to be evaluative or normative rather than purely descriptive or explanatory. It seeks to evaluate actual and prospective practices in terms of their impacts on true versus false beliefs. Although truth may have no explanatory role to play in the social studies of knowledge, it can play a regulative role. How can truth play a regulative role, it may be asked, unless we already have ways of deciding what is true? How can the social epistemologist assess the truth-propensity of a practice unless she already has a method of determining whether the beliefs caused by the practice are true or false? But if she has such a method of determination, why bother with social epistemology? In answer to these questions, it is sometimes possible to demonstrate mathematically that

a certain practice would have certain veritistic properties. For example, Goldman indicates that a particular (difficult to instantiate) practice of Bayesian inference has a general propensity, on average, to increase the veritistic properties of one's beliefs (Goldman 1999: 115-123). Similarly, it can be shown mathematically that a certain mode of amalgamating expert opinions in a group yields greater group accuracy than other modes of amalgamation (Shapley and Grofman 1984; Goldman 1999: 81-82). Finally, a practice can sometimes be judged veritistically unsatisfactory when later and better evidence shows that many judgments issued under its aegis were false. The medieval practice of trial by ordeal was abandoned in part because it was shown that the ordeal had produced numerous erroneous judgments of guilt. This emerged when voluntary confessions were later obtained from different people, or new eye-witnesses came forward.

Philip Kitcher has also developed the social epistemology of science from a truth-oriented perspective. One of his chief concerns has been the division of cognitive labor (Kitcher 1990, 1993: chap. 8). The progress of science will be optimized, says Kitcher, when there is an optimal distribution of effort within the scientific community. It may be better for a scientific community to attack a given problem by encouraging some members to pursue one strategy and others to pursue another, rather than all pursue the single most promising strategy. In saying that progress will be "optimized", it is meant that it will be optimized in terms of getting true answers to significant scientific questions. In *The Advancement of Science* (1993) Kitcher constructs the notion of a "consensus practice", a social practice built up from individual practices consisting of an individual's beliefs, the informants he regards as credible, the methodology of scientific reasoning he accepts, and so forth. A "core" consensus practice consists of the elements of individual practices common to all members of the community. A "virtual" consensus practice is a practice generated by taking into account the statements, methodologies, etc. that members accept "indirectly" by deferring to other scientists as authorities. Kitcher then constructs a family of notions of scientific "progress", and he characterizes progress in terms of improvements of consensus practices in getting significant truth and achieving explanatory success.

Feminist epistemologists often embrace the idea of social epistemology. However, many of them strongly criticize traditional epistemology and view it as a poor model for feminist epistemology. At least a few feminist epistemologists, however, take a fundamentally truth-oriented position. Elizabeth Anderson explicitly views feminist epistemology as a branch of social epistemology (1995: 54). Furthermore, when she proceeds to explain the aim of social epistemology, she identifies it as the aim of promoting our reliable, i.e., truth-conducive, processes of belief formation and checking or canceling out our unreliable belief-forming processes (1995: 55). Thus, the fundamental aim is the classical one of seeking true beliefs and avoiding false ones.

Thus far our examples of classically-oriented social epistemology center on the truth aim. What about the aim of epistemic justification or rationality? As indicated earlier, the problem of testimony is a problem about justification: What makes a hearer *justified* in accepting a report or other factual statement by a speaker? In the last two decades, testimony has become an active area of epistemological investigation. Although testimony theorists do not generally use the phrase "social epistemology" to describe their inquiry, that seems to be an appropriate label (see Schmitt 1994a).

According to "reductionism" about testimony, a hearer H is justified or warranted in accepting a speaker's report or factual statement only if H is justified is believing that the speaker is reliable and sincere, where the latter justification rests on sources other than testimony itself. In other words, testimony can only be a derivative source of epistemic warrant, not a "basic" source like perception, memory, or inductive inference. This reductionist view was held by Hume. In contrast to reductionism, there is the view that testimony is itself a "basic" source of warrant. A hearer has default, or *prima facie*, warrant in believing what a speaker says, no matter how little he knows about the speaker's reliability and sincerity. Of course, evidence of the speaker's unreliability or insincerity may defeat or override his *prima facie* warrant for acceptance. But that does not conflict with the claim that testimony *per se* is a basic source of evidence. C. A. J. Coady (1992) argues that reductionism would lead to widespread skepticism, because people do not generally have enough testimony-free evidence of testifiers' trustworthiness to confer the sort of justification demanded by reductionism. This argument is amplified and qualified by Elizabeth Fricker (1995). Others such as Tyler Burge (1993) and Richard Foley (1994) have argued that reductionism is in any case inadequately motivated.

Another form of justificationist social epistemology that is broadly classical in conception is that of Helen Longino (1990, 1993). In Longino's conception, a scientific belief is justified to the extent that it results from the application of "objective" methods. "To say that a theory or hypothesis was accepted on the basis of objective methods does not guarantee that it is true, but it does--if anything does--justify us in asserting that it is true" (1993: 268). The social character of Longino's approach is reflected in her insistence that objectivity is a characteristic of a community's scientific practice rather than a characteristic of an individual scientist, for objectivity refers to the avoidance of individual subjectivity or bias. Longino develops an account of objectivity that is tied to critical discourse. Objectivity in a scientific community, she says, would require four features: recognized avenues for criticism, responsiveness of beliefs to critical discussion, shared standards of responsiveness to criticism, and equality of intellectual authority.

3. Anti-Classical Approaches

Many researchers in the social studies of knowledge reject or ignore such classical concerns of epistemology as truth, justification, and rationality. It is acknowledged, of course, that various communities and cultures speak the language of truth, justification, or rationality, but the researchers in question do not find such concepts legitimate or useful for their own purposes. They seek to describe and understand a selected community's norms of rationality, like anthropologists describing the norms or mores of an alien culture. But they reject the notion that there are any universal or "objective" norms of rationality, or criteria of truth, that they themselves could appropriately invoke. As Barry Barnes and David Bloor put it, "there are no context-free or super-cultural norms of rationality" (1982: 27). So they are not prepared to decree that certain practices are more rational or more truth-conducive than others. In other words, they officially decline to make any judgments about the epistemic properties of various belief-forming practices (though the debunking connotations of their work, discussed below, may belie this stance). They indicate that such judgments would have no culture-free basis or foundation.

They are, nonetheless, clearly interested in belief-forming practices. If we use the term "knowledge" for any sort of belief (or at least for "institutionalized" belief), whether true or false, justified or unjustified, then they can be said to be investigators of knowledge. Since they are specifically interested in social influences on knowledge (so understood), they plausibly qualify as social epistemologists. They do not typically apply this label to themselves, perhaps in recognition that what was traditionally called "epistemology" had different purposes or aspirations. But if the old aspirations must be abandoned -- as Richard Rorty (1979) explicitly argued -- why not use the old label for the new type of project? For this reason, researchers in the social studies of science, or science and technology studies, will here be considered social epistemologists. There is, however, an additional reason why some of these writers might be called social epistemologists. Some claim to derive epistemologically significant conclusions (in the classical sense of "epistemology") from their sociological or anthropological investigations. Two examples are cases in point. First, as indicated earlier, historical case studies undertaken by members of the Edinburgh School attempt to show that scientists are heavily influenced by social factors "external" to the proper business of science. Other social analyses of science try to show how the game of scientific persuasion is really driven by factors resembling battles or political contests, where the outcome depends on the number or strength of one's allies *as contrasted* with, say, genuine epistemic worth. If either of these claims were right, the epistemic status of science as an objective and authoritative source of information would be greatly reduced. This claim, if true, seems to have genuine epistemological significance. Second, some sociologists of science claim to show that scientific "facts" are not "out-there" entities, which obtain independently of the human social interactions, but are mere "fabrications" resulting from those social interactions. This is an epistemological thesis, or at least a metaphysical thesis, of some philosophical significance. So some of these writers seem to have philosophical aspirations, not merely social science aspirations.

Let us begin with the first type of thrust, i.e., attempts to debunk the epistemic authority of science. The debunking of science's epistemic authority, at least by sociologists or historians of science, would have to be accomplished by empirical means, for example, by showing how scientific beliefs were actually produced in this or that socio-historical episode. This is precisely what various historians and sociologists of science purport to accomplish. One challenge to this would be a straightforwardly empirical challenge: Do these historical accounts get matters right? Many debunking efforts by members of the "Strong Programme" in the sociology of science have been disputed by others. In addition, there is an obvious, theoretically more interesting, response. How can these studies establish the debunking conclusions unless the studies themselves have epistemic authority? Yet the studies themselves use some of the very empirical, scientific procedures they purport to debunk. If such procedures are epistemically questionable, the studies' own results should be in question. There is, in other words, a problem of "reflexivity" facing this type of debunking challenge.

Not all sociological approaches are linked to historical case studies. Some offer a more theoretical analysis of how scientists persuade one another of this or that conclusion. For example, Bruno Latour sketches an account of how persuasion is effected in science by marshalling "allies" of substantial reputation on one's own side of a controversy (1987: chap. 1). Can this ostensibly non-epistemic account of science support a successful debunking of its epistemic pretensions? A first point to notice is that any successful debunking of epistemic authority, if explicitly spelled out, must address epistemic issues. It

must be shown that the procedures used by scientists have poor epistemic qualities. But this presupposes that there are objective, *bona fide* epistemic categories, which sociologists of science of Latour's persuasion tend to doubt or deny. If such categories are admitted, the further question arises as to whether persuasion by reference to the numbers of concurring "allies" is really an epistemically bad procedure. Although Latour's military/political vocabulary provides an amusing contrast with conventional characterizations of science, it isn't clear that the practices described are epistemically bad, or sub-rational, practices.

Let us turn now to the social construction of scientific facts. Again there is a question of how this sort of thesis could be established by sociologists. How could any scrutinizing of the activities of human scientists have determinate implications as to whether certain chemical substances, for example, exist independently of these scientists' interactions? Yet this is exactly what Latour and Steve Woolgar imply in their book *Laboratory Life: The [Social] Construction of Scientific Facts* (1979/1986). Latour and Woolgar claim that the "reality [of a scientific entity or fact] is formed as a consequence of [the] stabilization [of a controversy]" (1986: 180). In other words, the reality does not exist prior to the social event of stabilization, but is the result of such stabilization. How can they determine this without being biochemists as opposed to sociologists? How can the study of macro-events of a social nature establish that there do or do not exist certain biochemical substances independently of those macro-events?

In discussing social constructivism, it is essential to distinguish between weak and strong versions. Weak social constructivism is the view that human *representations* of reality -- either linguistic or mental representations -- are social constructs. For example, to say that gender is socially constructed, in this weak version of social constructivism, is to say that people's representations or conceptions of gender are socially constructed. Strong social constructivism claims not only that representations are socially constructed, but that the *entities themselves* to which these representations refer are socially constructed. In other words, not only are scientific representations of certain biochemical substances socially constructed, but the substances themselves are socially constructed. The weak version of social constructivism is quite innocuous, at least in the present context. Only the thesis of strong social constructivism is metaphysically (and, by implication, epistemologically) interesting. It is this sort of metaphysical thesis that Latour and Woolgar seem to endorse.

But there are many problems with this metaphysical thesis. One question is whether social constructivists like Latour and Woolgar mean to be "causal" constructivists or "constitutive" constructivists, in the terminology of Andre Kukla (2000). Causal constructivism is the view that human activity causes and sustains facts about the world, including scientific facts, whereas constitutive constructivism is the view that what we call "facts about the world" are really just facts about human activity (Kukla 2000: 21). Although Latour and Woolgar use the language of causal constructivism, it seems more likely that the doctrine they intend is constitutive constructivism. There are, however, severe philosophical difficulties for constitutive social constructivism as a general metaphysical doctrine, as Kukla explains.

Not all researchers within the social studies of science think of social epistemology as confined to the description and explanation of science. Steve Fuller (1987, 1988, 1999), who champions social epistemology in that very phrase, sees the enterprise as normative: How should the institution of science

be organized and run? What is the best (scientific) means to knowledge production? However, Fuller does not construe "knowledge" in a truth-entailing fashion, and so parts company with classical epistemology. What does he take the end of knowledge production to be? In one place he says that it's a matter of empirical determination what that end is (1987: 177). But if we don't now know the end, how can we try to direct science toward it? And how can one determine science's end empirically? Science might be found to have many different results. Which of them is its "end"?

4. Conceptions of the Social

In what sense is social epistemology "social"? Different writers have different conceptions of the social, and this inevitably leads to different conceptions of social epistemology. In the Marxian tradition and in early forms of the sociology of knowledge, "social factors" referred primarily to various types of "interests": class interests, political interests, or anything else pertaining to the "existential" world of power and politics. Under this conception of the social, it is natural to see social factors as antithetical to "reason". If science is infiltrated by social factors, in this sense, how can it be a successful instrument for getting at truth? Thinking of the relationship between the rational and the social as one of opposition, it is not surprising to find Larry Laudan proposing an "arationality principle": "[T]he sociology of knowledge may step in to explain beliefs if and only if those beliefs cannot be explained in terms of their rational merits" (Laudan 1977: 202).

Can the opposition between the rational and the social be eliminated, or at least relaxed? A first possible move is to allow "interests" to include the private or professional interests of scientists. It seems undeniable that scientists are at least partly driven by a desire for "credit" from their peers (Hull 1988). But won't private and professional interests deflect scientists from reason and truth as much as class or political interests? Several writers argue to the contrary. There is no necessary conflict between professional interest and successful pursuit of truth. Kitcher (1990) argues that the optimal division of labor in scientific research may be attained not by "pure", altruistic scientists but by scientists with "grubby" and epistemically "sullied" motives. Similarly, Goldman and Shaked (1991) show that, given certain assumptions about credit-giving practices and experimental choices, there will be little difference between choices of truth-motivated scientists and choices of credit-motivated scientists. Hence, there will be little difference in expected success of moving the community toward truth. Credit-driven interests need not be inimical to truth-promotion.

A further proposal is to expand the "social" beyond politics and interests altogether. The most inclusive sense of the social is simply any relationship among two or more individuals. There is no reason why social epistemology cannot be social in this broad sense. Any interaction among individuals affecting the credal states of some of them might be considered a social-epistemic relationship. So understood, a wide range of communicative interactions would be fit subjects for social epistemology. For example, many knowledge-seeking enterprises are collaborative in nature, including scientific enterprises involving research teams. An interesting task for social epistemology is to identify the types of collaboration that would be optimal in terms of some epistemically relevant measure (Thagard 1997).

Can the "social" be fully captured by inter-individual relationships? Some theorists would argue in the negative, pointing specifically to collective entities such as corporations, committees, juries, and teams. We often attribute mental or mental-like states, including beliefs, to such collective entities (Gilbert 1989; Tuomela 1995; Searle 1995). We might say, for example, that a jury was convinced that the defendant intended such-and-such, or that the jury doubted that a certain alleged conversation really took place. Collective entities are obviously "social" in an important way; and if it is granted that such entities are bearers of beliefs and other doxastic states, shouldn't these collective states be an important target for social epistemology? Precisely this is suggested by Lynn Hankinson Nelson (1993), who goes even further in proposing that the *only* real knowers are communities. For purposes of social epistemology it may not be sufficient that the notion of collective belief be legitimized. What must also be legitimized is the thesis that collective beliefs have epistemic properties, such as "rational" and "irrational", or "justified" and "unjustified". Perhaps this thesis is defensible, however. Frederick Schmitt (1994b) has argued that sense can be made of justification for group beliefs. So the door seems open to this widening of the concept of the "social" for the purposes of social epistemology.

5. The Scope and Methods of Social Epistemology

Epistemology has traditionally focused on the processes or activities of an inquiring and believing agent, an agent who gathers evidence and forms (or withholds) belief based on this evidence. From this point of view, it is natural to restrict epistemology to individual epistemology. But it is also possible to consider the activities of other agents, groups of agents, and institutions whose activities impinge on the quantity and quality of evidence available to a doxastic agent. What truths known by one agent are communicated in some form to additional agents? What falsehoods are similarly disseminated, whether from sincere conviction or deceptive intent? Ascending a level, what norms or practices are in place, or might be put in place, to influence the contents of communications in ways that encourage greater truth acquisition? If epistemology is approached from this perspective, many social dimensions come into view. Additional dimensions are introduced by considering agents and institutions that assist the gathering of evidence by others. Experimentation is the core of scientific evidence gathering, but experimentation requires resources that individual scientific agents cannot marshall on their own. Other scientists and society at large must make choices about which experimental projects to support.

The foregoing questions open up a wide territory for social epistemology to occupy. A great deal of social epistemology has thus far been rather "local" in scope, centering heavily on the domain of science. This is understandable, since science is the most visible, organized, and influential enterprise of knowledge seeking, or knowledge production. But science isn't the only enterprise of knowledge acquisition or knowledge dissemination. People acquire everyday factual information (e.g., the location of the nearest hardware store, the stance that political candidate X has taken on issue Y) by personal observation, and by hearing or reading the reports of others. These routes to knowledge are not strictly scientific. Moreover, most members of the general public never conduct scientific inquiries, but at best rely on first- or second-hand reports about them. Since specialists themselves often disagree in interpreting scientific results, laypersons must choose among them. They must decide what to believe without personally applying scientific methods. Finally, some contexts do not afford opportunities for

(full) application of scientific methods. Jurors have no opportunity to apply all the tools of science to the factual questions before them, although they may, of course, hear expert witnesses who have applied scientific methods to related matters. Thus, while social epistemology has every reason to be concerned with science, it should also target other domains and knowledge-related practices.

For reasons such as these, social epistemology might be divided into topics that examine either global or local practices. Global practices cut across subject-matters and specific institutions. Local practices are tied to particular domains or types of institutions. Styles of rhetoric and interpersonal argumentation, for example, are found in every speech community and speech context. They are examples of global practices; and norms that guide such practices can be highly relevant to knowledge acquisition. Similarly, communication technologies and communicational institutions can massively affect the diffusion of all categories of information or misinformation. The printing press, the computer, and the Internet are particularly salient examples of pertinent communicational technologies. But we should not forget the scholarly library, or print, radio, and television journalism. Finally, systems of public and private education, at all levels, play crucial roles in disseminating knowledge. The form, structure, and policies associated with these communicational entities are highly relevant to the distribution of societal knowledge (Goldman 1999).

Local practices are those associated with particular domains or institutions, such as the adjudicative aspect of the law. Different legal traditions provide different sorts of structures for judging disputes. In the Anglo-American adversarial system, lawyers for the contending parties take the initiative in discovering and presenting evidence for their side, and lay jurors often play the role of fact-finders. In the Continental system, investigations and trials are led by judges rather than lawyers, and the same judges are the principal fact-finders. These systems presumably have the same central aim, viz., rendering correct or accurate judgments about questions under dispute, but one system might be better than the other at achieving that aim..

In all of these domains, social epistemology can ask questions about knowledge-enhancing practices and policies. Which journalistic practices, which Website-designing practices, which rules for intellectual property are best from a knowledge-promoting standpoint (Herman and Chomsky 1988; Fallis 2000; Lessig 1999)? Which system of legal adjudication -- the Continental system or the Anglo-American system -- is optimal in epistemic terms (Langbein 1985; Goldman 1999, chap. 9)? Looking at the details of the Anglo-American system, which specific rules concerning evidence admission or "discovery" of evidence would be epistemically optimal (Damaska 1997; Talbott and Goldman 1998)?

Depending on the specific area of social epistemology one chooses, different methodologies or research paradigms may be appropriate. The choice of methodology will also depend, of course, on whether a descriptive or normative approach is taken. In general, however, it seems likely that methodologies from many different disciplines are needed for a comprehensive treatment of social epistemology. Here is a sampling of methods that have already been applied in some of these terrains.

In the history and sociology of science, case studies and "field" studies (of laboratories) are prevalent.

Rhetorical theory provides another approach to scientific discourse (McCloskey 1985; Fuller 1993). Yet another analytical tool for social epistemology is probability theory. For example, it can be used to prescribe rational changes in an agent's degree of belief, given credibility weights assigned to other agents and their degrees of belief (Lehrer and Wagner 1981). Various techniques of economic analysis, including game theory, can be helpful in social epistemology. These have been utilized in at least two domains: journalism (Cox and Goldman 1994) and the market for speech (Goldman and Cox 1996). Economists William Brock and Steven Durlauf (1999) have created a formal model of theory choice in science, using a neoclassical style of economic analysis. General algebraic techniques are used by Philip Kitcher (1993, chap. 8) to analyze social aspects of science. Clearly, the multiple problems and approaches of social epistemology invite varied research tools borrowed from many disciplines, and initial applications of these tools have been made. It is equally clear, however, that the work of the field lies more in the future than in the past.

Bibliography

- Anderson, Elizabeth (1995) "Feminist Epistemology: An Interpretation and a Defense," *Hypatia* 10 (3): 50-84.
- Barnes, Barry and Bloor, David (1982) "Relativism, Rationalism, and the Sociology of Knowledge," in *Rationality and Relativism*, ed. M. Hollis and S. Lukes, Cambridge: MIT Press.
- Brock, William and Durlauf, Steven (1999) "A Formal Model of Theory Choice in Science," *Economic Theory* 14: 113-130.
- Burge, Tyler (1993) "Content Preservation," *The Philosophical Review* 102: 457-488.
- Coady, C. A. J. (1992) *Testimony*, Oxford: Oxford University Press.
- Cox, James and Goldman, Alvin (1994) "Accuracy in Journalism: An Economic Approach," in *Socializing Epistemology*, ed. F. Schmitt, Lanham, MD: Rowman and Littlefield.
- Damaska, Mirjan (197) *Evidence Law Adrift*, New Haven: Yale University Press.
- Descartes, Rene (1637/1955). *Discourse on the Method of Rightly Conducting the Reason and Seeking for Truth in the Sciences*, trans. E. Haldane and G. Ross, *The Philosophical Works of Descartes*, vol. 1, New York: Dover.
- ----- (1641/1955). *Meditations on First Philosophy*, trans. E. Haldane and G. Ross, *The Philosophical Works of Descartes*, vol. 1, New York: Dover.
- Fallis, Don (2000) "Veritistic Social Epistemology and Information Science," *Social Epistemology* 14 (4): 305-316.
- Foley, Richard (1994) "Egoism in Epistemology," in *Socializing Epistemology*, ed. F. Schmitt, Lanham, MD: Rowman and Littlefield.
- Forman, Paul (1971) "Weimar Culture, Causality and Quantum Theory, 1918-1927: Adaptation by German Physicists and Mathematicians to a Hostile Intellectual Environment," in *Historical Studies in the Physical Sciences* 3, ed. R. McCormmach, Philadelphia: University of Pennsylvania Press.
- Foucault, Michel (1977) *Discipline and Punish*, trans. A. Sheridan, New York: Random House.
- ----- (1980) *Power/Knowledge*, New York: Pantheon.
- Fricker, Elizabeth (1995) "Telling and Trusting: Reductionism and Anti-Reductionism in the

Epistemology of Testimony," *Mind* 104: 393-411.

- Fuller, Steve (1987) "On Regulating What is Known: A Way to Social Epistemology," *Synthese* 73: 145-183
- ----- (1988) *Social Epistemology*, Bloomington: Indiana University Press.
- ----- (1993) *Philosophy, Rhetoric, and the End of Knowledge*, Madison: University of Wisconsin Press.
- ----- (1999) *The Governance of Science: Ideology and the Future of the Open Society*, London: Open University Press.
- Geuss, Raymond (1981) *The Idea of a Critical Theory: Habermas and the Frankfurt School*, Cambridge: Cambridge University Press.
- Gilbert, Margaret (1989) *On Social Facts*, London: Routledge.
- Goldman, Alvin (1978) "Epistemics: The Regulative Theory of Cognition," *The Journal of Philosophy* 75: 509-523.
- ----- (1986) *Epistemology and Cognition*, Cambridge: Harvard University Press.
- ----- (1987) "Foundations of Social Epistemics," *Synthese* 73: 109-144.
- ----- (1999) *Knowledge in a Social World*, Oxford: Oxford University Press.
- Goldman, Alvin and Cox, James (1996) "Speech, Truth, and the Free Market for Ideas," *Legal Theory* 2: 1-32.
- Goldman, Alvin and Shaked, Moshe (1991) "An Economic Model of Scientific Activity and Truth Acquisition," *Philosophical Studies* 63: 31-55.
- Habermas, Jurgen (1973) "Wahrheitstheorien," in *Wirklichkeit und Reflexion: Festschrift fur Walter Schulz*, Pfullingen: Neske.
- Habermas, Jurgen and Luhmann, Niklas (1971) *Theorie der Gesellschaft oder Sozialtechnologie -- Was Leistet die Systemforschung?* Frankfurt: Suhrkamp.
- Herman, Edward and Chomsky, Noam (1988) *Manufacturing Consent: The Political Economy of the Mass Media*, New York: Pantheon Books.
- Hull, David (1988) *Science as a Process*, Chicago: University of Chicago Press.
- ----- (1975) *An Enquiry Concerning Human Understanding*, in *Hume's Enquiries*, ed. P. H. Nidditch and L. A. Selby-Bigge, Oxford: Oxford University Press.
- Kitcher, Philip (1990) "The Division of Cognitive Labor," *The Journal of Philosophy* 87: 5-22.
- ----- (1993) *The Advancement of Science*, New York: Oxford University Press.
- Kuhn, Thomas (1962/1970) *The Structure of Scientific Revolutions*, 2nd ed., Chicago: University of Chicago Press.
- Kukla, Andre (2000) *Social Construction and the Philosophy of Science*, London: Routledge.
- Langbein, John (1985) "The German Advantage in Civil Procedure," *University of Chicago Law Review* 52: 823-866.
- Latour, Bruno (1987) *Science in Action*, Cambridge: Harvard University Press.
- Latour, Bruno and Woolgar, Steve (1979/1986) *Laboratory Life: The [Social] Construction of Scientific Facts*, Princeton: Princeton University Press.
- Laudan, Larry (1977) *Progress and Its Problems*, Berkeley: University of California Press.
- Lehrer, Keith and Wagner, Carl (1981) *Rational Consensus in Science and Society*, Dordrecht: Reidel.
- Lessig, Lawrence (1999) *Code and Other Laws of Cyberspace*, New York: Basic Books.

- Locke, John (1959) *An Essay Concerning Human Understanding*, 2 volumes, ed. A.C. Fraser, New York: Dover.
- Longino, Helen (1990) *Science as Social Knowledge*, Princeton: Princeton University Press.
- ----- (1993) "Essential Tensions--Phase Two: Feminist, Philosophical, and Social Studies of Science," in *A Mind of One's Own*, ed. L. Antony and C. Witt, Boulder, Co: Westview Press.
- Mackenzie, Donald (1981) *Statistics in Britain: 1865-1930, The Social Construction of Scientific Knowledge*, Edinburgh: Edinburgh University Press.
- Mannheim, Karl (1936) *Ideology and Utopia*, trans. L. Wirth and E. Shils, New York: Harcourt, Brace and World.
- McCloskey, Donald (1985) *The Rhetoric of Economics*, Madison: University of Wisconsin Press.
- Merton, Robert (1973) *The Sociology of Science*, Chicago: University of Chicago Press.
- Nelson, Lynn Hankinson (1993) "Epistemological Communities," in *Feminist Epistemologies*, ed. L. Alcoff and E. Potter, New York: Routledge.
- Reid, Thomas (1975) *An Inquiry into the Human Mind on the Principles of Common Sense*, in *Thomas Reid's Inquiry and Essays*, ed. R. Beanblossom and K. Lehrer, Indianapolis: Bobbs-Merrill.
- Rorty, Richard (1979) *Philosophy and the Mirror of Nature*, Princeton: Princeton University Press.
- Shapin, Steven (1975) "Phrenological Knowledge and the Social Structure of Early Nineteenth-Century Edinburgh," *Annals of Science* 32: 219-243.
- Shapley, Lloyd and Grofman, Bernard (1984) "Optimizing Group Judgmental Accuracy in the Presence of Interdependence," *Public Choice* 43: 329-343.
- Shera, Jesse (1970) *Sociological Foundations of Librarianship*, New York: Asia Publishing House.
- Schmitt, Frederick (1994a) "Socializing Epistemology: An Introduction through Two Sample Issues," in *Socializing Epistemology*, ed. F. Schmitt, Lanham, MD: Rowman and Littlefield.
- ----- (1994b) "The Justification of Group Beliefs," in *Socializing Epistemology*, ed. F. Schmitt, Lanham, MD: Rowman and Littlefield.
- Searle, John (1995) *The Construction of Social Reality*, New York: Free Press.
- Talbott, William and Goldman, Alvin (1998) "Games Lawyers Play: Legal Discovery and Social Epistemology," *Legal Theory* 4: 93-163.
- Thagard, Paul (1997) "Collaborative Knowledge," *Nous* 31: 242-261.
- Tuomela, Raimo (1995) *The Importance of Us: A Philosophical Study of Basic Social Notions*, Stanford: Stanford University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

testimony: epistemological problems of

Alvin I. Goldman
goldman@u.arizona.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 26, 2001
Content last modified: February 26, 2001

Virtue Epistemology

An approach in epistemology that applies the resources of virtue theory to problems in the theory of knowledge. It is argued that by doing so it is possible to give informative accounts of knowledge, evidence, and other important epistemic concepts, while solving a wide range of problems that have plagued other approaches in the theory of knowledge.

Introduction

Virtue epistemology begins with the assumption that epistemology is a normative discipline. The main idea of virtue epistemology is to understand the kind of normativity involved on the model of virtue theories in ethics. This main idea is best understood in terms of a thesis about the direction of analysis. Just as virtue theories in ethics try to understand the normative properties of actions in terms of the normative properties of moral agents, virtue epistemology tries to understand the normative properties of beliefs in terms of the normative properties of cognitive agents. Hence virtue theories in ethics have been described as person-based rather than act-based, and virtue epistemology has been described as person-based rather than belief-based.

For example, non-virtue theories might try to understand the epistemic justification of belief in terms of doing one's epistemic duty, believing according to the evidence, or using a reliable method. In each case

the account of justified belief makes no reference to any normative properties of persons. On the contrary, it would be natural on such views to think of epistemic virtues as dispositions to believe in the ways in question. Virtue epistemology changes this direction of analysis by understanding justified belief in terms of epistemic virtues. For example, Ernest Sosa has argued that justified belief is belief that is grounded in epistemic virtue. Similarly, Linda Zagzebski has argued that knowledge is true belief arising out of acts of intellectual virtue. Of course the next task is to give an informative account of the cognitive virtues involved in such definitions. Depending on how this is done, we get further versions of virtue epistemology.

A number of claims have been made on behalf of virtue epistemology. We have already seen that virtue epistemologists promise to define a range of important epistemic concepts by drawing on the resources of virtue theory. Beyond this, it has been claimed that virtue epistemology can overcome the debate between internalist and externalist conceptions of justification, that it can solve problems pertaining to skepticism, that it can solve Gettier problems, and that it can contribute to a unified theory of value across epistemology and ethics. Recent interest in virtue epistemology began with a paper by Ernest Sosa, where he claimed that a turn to virtue theory would allow a solution to the impass between foundationalist theories of justification and coherentist theories of justification. One way to organize the literature is to begin with Sosa's paper and the development of his own version of virtue epistemology. We may then look at various reactions to Sosa's seminal work. As we shall see, these may be divided into two categories. While some critics have responded with objections to the idea that we should turn to virtue theory in epistemology, another group has responded with objections that Sosa does not go far enough in exploiting the various resources of virtue theory.

Virtue Perspectivism

In "The Raft and the Pyramid: Coherence versus Foundations in the Theory of Knowledge," Sosa suggested that virtue epistemology would allow a solution to the foundationalismÐcoherentism problematic in epistemology. We may think of foundationalism on the metaphor of a pyramid: there is a structure to knowledge involving a nonsymmetrical relation of support among levels, with one level having the special status of a foundation which supports all the rest. In the most plausible versions of foundationalism sensory experience plays an important role in the foundation, providing a ground for observational knowledge from which further knowledge can be inferred higher up in the structure. Coherentism counters this account of knowledge with its metaphor of the raft: knowledge is a structure that floats free of any secure anchor or tie. No part of knowledge is more fundamental than the rest to the overall structure, all of the parts being held together by the ties of logical relations.

According to Sosa, both these accounts of knowledge have fatal flaws. The problem with coherentism is that it cannot account for knowledge at the periphery of a system of beliefs. This is because coherentism makes justification entirely a function of the logical relations among beliefs in the system, but perceptual beliefs have very few logical ties to the remainder. This makes it possible to generate counterexamples to coherentism by means of the following recipe. First, take a perfectly coherent system of beliefs that seems to provide good examples of justified belief and knowledge. Second, replace one perceptual belief

in the system with its negation, while also making any other slight changes that are necessary to preserve coherence. This will have very little effect on the overall coherence of the system, since that is a function only of the logical relations among the system's beliefs. Accordingly, it will turn out to be the case that the new "perceptual belief" is as coherent as the old one, and is therefore, according to coherentism, equally well justified. This result is counter-intuitive, however, since the person's sensory experience has remained the same. Surely she is not justified in believing that she is not standing in front of a tree, for example, if her sensory experience is as if she is standing in front of a tree. Examples like this one suggest that justification is a function of more than the relations among beliefs. Specifically, it is partly a function of one's sensory experience.

This gives the advantage to foundationalism, which allows a role for sensory experience in justified belief and knowledge. But an equally problematic dilemma arises for foundationalism, depending on how one thinks of foundationalism's epistemic principles. Suppose we agree that there is some true epistemic principle relating (i) a relevant sensory experience to (ii) one's justified belief that one is standing in front of a tree. Is this to be understood as a fundmental principle about epistemic justification, or is it to be understood as an instance of some more general principle? If we say the former, then the foundationalist is faced with a seemingly infinite multitude of fundamental principles with no unifying ground. There would be different fundamental principles for visual and auditory experience, for example, as well as possible principles for beings not like us at all, but capable of having their own kind of sensory knowledge. The more attractive alternative is to think of the foundationalist's principles as derived, but then we need an account of some deeper, unifying ground.

This is the context in which Sosa suggests that virtue epistemology will do the trick. Suppose we think of virtues in general as excellences of character. A virtue is a stable and successful disposition: an innate ability or an acquired habit, that allows one to reliably achieve some good. An intellectual virtue will then be a cognitive excellence: an innate ability or acquired habit that allows one to reliably achieve some intellectual good, such as truth in a relevant matter. We may now think of justified belief as belief that is appropriately grounded in one's intellectual virtues, and we may think of knowledge as true belief that is so grounded. By adopting this position, we can see the foundationalist's epistemic principles as instances of this more general account of justified belief and knowledge. The idea is that human beings possess intellectual virtues that involve sensory experience; i.e. stable and reliable dispositions for forming beliefs about the environment on the basis of experiential inputs. Such dispositions involve various sensory modalities such as vision and hearing. Other cognitive beings might be possessed of analogous dispositions, involving kinds of sensory experience unknown to humans. Accordingly, Sosa argues, virtue epistemology provides the unified account that was needed.

The same idea accounts for the truth involved in coherentism as well. Namely, coherence gives rise to justified belief and knowledge precisely because it is the manifestation of intellectual virtue. In our world, and for beings like us, coherence increases reliability, and therefore constitutes a kind of intellectual virtue in its own right. Moreover, coherence of a certain sort allows for reflective knowledge as opposed to mere animal knowledge. According to Sosa, we rise to a different and superior kind of justification and knowledge when we are able to see our beliefs as deriving from intellectual virtues. This perspective on our virtues must itself derive from a second-order intellectual virtue, one that allows us to

reliably monitor and adjust our first-level cognitive dispositions.

Notice that the above ideas involve the direction of analysis thesis discussed above. Traditional foundationalism and coherentism try to account for justified belief and knowledge solely by reference to the properties of beliefs; i.e. their logical relations (coherentism) or their logical relations plus their relations to sensory experiences (traditional foundationalism). Sosa's version of virtue epistemology accounts for various kinds of justified belief and knowledge by first defining the notion of an intellectual virtue, and then by defining various normative properties of beliefs in terms of this more fundamental property of persons.

Responses to Sosa have focussed on various objections to the position outlined above, including the general claim that a turn to virtue theory would be a fruitful approach in epistemology. A second group of critics has endorsed Sosa's call for a turn to virtue theory in epistemology, but have argued that he does not go far enough in exploiting the various resources that virtue theory offers.

Responsibilism I

One early response to Sosa along this second line is by Lorraine Code, who argues for the centrality of epistemic responsibility in epistemology. Code agrees with Sosa's direction of analysis thesis, endorsing the idea that primary justification is best understood as attaching to stable dispositions to act in certain ways, while secondary justification accrues to particular acts because of their sources in virtues. This approach, she argues, appropriately focuses epistemology on persons, their cognitive activities, and their membership in a community defined by social practices of enquiry. The individual knower is now recognized as part of a community, with all the moral and intellectual obligations that this entails. However, Code argues, redirecting epistemology in this way gives the notion of epistemic responsibility central importance. Characterizing Sosa's position as a version of reliabilism, she argues that her own "responsibilism" constitutes a more adequate development of Sosa's initial insights. This is because, in part, the notion of responsibility emphasizes the active nature of the knower, as well as the element of choice involved in the knower's activity. Whereas a merely passive recorder of experience can be described as reliable, only an active, creative agent can be assessed as responsible or irresponsible, as having fulfilled her obligations to fellow enquirers, etc. According to Code, then, Sosa is correct to call for a focus on intellectual virtues in epistemology, with the focus on agency and community that this implies. But the natural way to develop this insight is to understand the intellectual virtues in terms of epistemic responsibility. Code goes so far as to say that epistemic responsibility is the central intellectual virtue, from which all other intellectual virtues radiate.

Another interesting feature of Code's view concerns some theses about the prospects for epistemology. Placing emphasis on virtue and responsibility, she argues, has consequences for both how epistemology should be conducted and the kind of epistemological insights we should hope for. First, emphasizing the contextual and social dimensions of knowledge introduces complexity into theorizing, and in such a way that shows the usual examples and counter-examples in epistemology to be inadequate. Such examples under-describe the relevant epistemic circumstances, leaving out such relevant considerations as history,

social role, conflicting obligations, etc. To show how such factors are indeed relevant, it is necessary to replace these thin examples with thickly descriptive narrative. Only stories that tie a whole life together provide an adequate context for epistemic evaluations, precisely because the factors that govern such evaluations are that rich and complex.

Moreover, Code argues, thick narratives are essential for understanding the very nature of intellectual virtue. Echoing a point by Alasdair MacIntyre, Code argues that an adequate understanding of what it is to be virtuous requires placing virtuous selves in the unity of a narrative. A consequence of this is that we should not expect to describe tidy conditions for justification and knowledge. The relevant criteria for epistemic evaluation are too varied and complex for that, and so any simple theory of knowledge will distort rather than adequately capture those criteria. This does not mean, however, that insight into the nature and conditions of justification and knowledge is impossible. Rather, such insight is to be gained by narrative history rather than theory construction of the traditional sort.

Responsibilism II

Following Sosa, Code tends to think of intellectual virtues as broad cognitive faculties or abilities pertaining to some subject matter. In this respect both authors are following Aristotle, who names intuition, science, wisdom and prudence as intellectual virtues. For example, for Aristotle intuition is the ability to know first principles, while science is the ability to deduce further truths from these. James Montmarquet has developed the notion of an intellectual virtue in a different direction, conceiving them on the model of Aristotle's moral virtues. Rather than thinking of intellectual virtues as cognitive faculties or abilities, he conceives them as personality traits, such as impartiality and intellectual courage. In sum, intellectual virtues are personality traits that a person who desires the truth would want to have.

Like Code, Montmarquet criticizes Sosa's position for not sufficiently exploiting the resources of virtue theory in ethics. Also like Code, he criticizes Sosa's emphasis on the reliability of intellectual virtues, and wants to replace this with an emphasis on responsibility and other concepts related to agency. According to Montmarquet, it is a mistake to characterize the intellectual virtues as reliable in the sense of truth-conducive. This is because we can imagine possible worlds, such as Descartes' demon world, where the beliefs of epistemically virtuous people are almost entirely false. Alternatively, we can imagine worlds where the intellectually lazy and careless have mostly true beliefs. Suppose we were to somehow discover that ours was such a world. Would we then revise our opinions about which traits count as intellectual virtues and which as vices? Montmarquet argues that we would not. Traits like intellectual courage and carefulness are virtues even if we are unfortunate enough to be the victims of a Cartesian deceiver, and traits like laziness and carelessness are vices even if, contrary to appearances, they turn out to be reliable. But then reliability cannot be a distinctive mark of the intellectual virtues.

A different approach is to characterize the virtues in terms of a desire for truth. According to Montmarquet, the central intellectual virtue is epistemic conscientiousness. To be conscientious in this sense is to be motivated to arrive at truth and to avoid error; it is to have an appropriate desire for the truth. Here there is a parallel with moral conscientiousness, where a morally conscientious person is

someone who tries her best to do what is right. This notion of epistemic conscientiousness is closely related to that of epistemic responsibility, or perhaps identical with it. Hence with Code, Montmarquet makes epistemic responsibility rather than reliability central to his understanding of intellectual virtue.

According to Montmarquet, then, epistemic conscientiousness is the central intellectual virtue. However, intellectual virtue cannot be understood solely in terms of a desire for truth, since one's desire for truth must be appropriately regulated. We must therefore countenance additional regulative virtues, which constitute ways of being conscientious. Montmarquet classifies these under three main categories. "Virtues of impartiality" include such personality traits as openness to the ideas of others, willingness to exchange ideas, and a lively sense of one's own fallibility. "Virtues of intellectual sobriety" oppose the excitement and rashness of the overly enthusiastic. Finally, "virtues of intellectual courage" include a willingness to conceive and examine alternatives to popular ideas, perseverance in the face of opposition from others, and determination to see an inquiry through to the end.

Montmarquet suggests that we can use the above account of intellectual virtue to define an important sense of subjective justification. Specifically,

> *S is subjectively justified in believing p insofar as S is epistemically virtuous in believing p.*

This is not the kind of justification that turns true belief into knowledge. This is because Gettier cases show that a person can be justified in believing something in this sense, but still lack the kind of objective relation to the truth required for knowledge. Nevertheless, Montmarquet argues, the above sense of justification is important regarding a different issue. Namely, Montmarquet is concerned with the problem of morally evaluating actions. More specifically, he is concerned with the problem of blaming persons for actions which, from their own point of view, are morally justified. Often enough, the morally outrageous actions of tyrants, racists and terrorists seem perfectly reasonable, even necessary, in the context of their distorted belief system. In order to find the actions blameworthy in such cases, it would seem that we have to find the beliefs blameworthy as well. In other words, we need some account of "doxastic responsibility," or the kind of responsibility for belief that can ground responsibility for actions. The above account of subjective justification, Montmarquet argues, provides what we are looking for. Precisely because it understands justification in terms of intellectually virtuous behavior, the account allows a plausible sense in which justified (and unjustified) belief is under a person's control. This, in turn, makes the relevant beliefs to be appropriate objects of blame and praise.

One objection to this sort of view is that judgements of responsibility are inappropriate in the cognitive domain. The idea is that judgements of praise and blame presuppose voluntary control, and that we lack such control over our beliefs. Montmarquet responds to this objection by distinguishing between a weak and a strong sense of voluntary control. Roughly, a belief is voluntary in the weak sense if it is formed in circumstances which do not interfere with virtuous belief formation. This kind of voluntariness amounts to freedom from interference or coercion. A belief is voluntary in the strong sense (again roughly) if it is subject to one's will. Montmarquet's strategy is to concede that responsibility requires weak voluntary control, but to argue that we often have this kind of control over our beliefs. Second, he concedes that we

do not typically have strong voluntary control over our beliefs, but argues that responsibility does not require it.

The analogy with action is instructive. One can be appropriately blamed for negligent actions and inadvertent actions, and even in cases where there is no actual choice regarding the action in question. In cases of action as well as belief, strong voluntary control is not necessary for responsibility. On the other hand, praise or blame would be inappropriate in cases where action is coerced. However, many of our beliefs satisfy the relevant "no coercion condition," and so are weakly voluntary in that sense.

A Mixed Theory

Greco has argued that intellectual virtue is closely tied to epistemic responsibility, but without rejecting Sosa's position that the virtues are reliable, or truth-conducive. The main idea is that an adequate account of knowledge ought to contain both a responsibility condition and a reliability condition. Moreover, a virtue account can explain how the two are tied together. In cases of knowledge, objective reliability is grounded in epistemically responsible action.

The way this works is as follows. First, we can give an account of subjective justification in terms of epistemic responsibility:

> *S is subjectively justified in believing p if and only if S's believing p is epistemically responsible.*

The notion of responsibility, in turn, can be understood in terms of the dispositions S manifests when S is thinking conscientiously, or is motivated to believe the truth. Such motivation need not be self-conscious, or even univocal. Rather, it is meant to specify the kind of default position that people are usually in, and to oppose this to the alternative motivations involved in such things as wishful thinking, pig-headedness and attention grabbing. This suggests the following account of subjective justification.

> *S is subjectively justified in believing p if and only if S's believing p results from the dispositions that S manifests when S is motivated to believe the truth.*

Finally, this kind of subjective justification gives rise to objective reliability when things go well:

> *S knows p only in cases where (a) S is subjectively justified in believing p, and (b) as a result of this S is objectively reliable in believing p.*

One feature of the above account is that it understands both justified belief and knowledge in terms of the dispositions that make up S's cognitive character. In other words, it makes the notion of virtuous character primary, and then gives accounts of justified belief and knowledge in terms of this. Accordingly, we can define virtuous character in terms of proper motivation and reliability as these

notions are understood above, and then given the following (partial) account of knowledge.

> *S knows p only in cases where S's believing p results from a virtuous cognitive character.*

A Social/Genetic Approach

Jonathan Kvanvig has argued for a more radical departure from traditional epistemological concerns. According to Kvanvig, traditional epistemology is dominated by an "individualistic" and "synchronic" conception of knowledge. Accordingly, one of the most important tasks from the traditional perspective is to specify the conditions under which an individual S knows a proposition p at a particular time t. Kvanvig argues that this perspective should be abandoned in favor of a new social/genetic approach. Whereas the traditional perspective focuses on questions about justified belief and knowledge of individuals at particular times, a new genetic epistemology would focus on the cognitive life of the mind as it develops within a social context. From the new perspective, questions concerning individuals are replaced with questions concerning the group, and questions concerning knowledge at a particular time are abandoned for questions about cognitive development and learning. Kvanvig argues that there are at least two ways in which the virtues would be central within the new perspective. First, the virtues are essential to understanding the cognitive life of the mind, particularly the development and learning which takes place over time through mimicking and imitation of virtuous agents. Second, in a social/genetic approach the virtues would play a central role in the characterization of cognitive ideals. For example, what makes a certain structuring of information superior, Kvanvig argues, is that it is the kind of structuring that a person of intellectual virtue would come to possess in the appropriate circumstances.

A Neo-Aristotelian Theory

We have seen that both Code and Montmarquet argue for a closer affinity between virtue epistemology and Aristotle's theory of the moral virtues. For example, Montmarquet thinks of the intellectual virtues as epistemically relevant personality traits, and both authors emphasize the close connection between virtue, agency and responsibility. The most detailed and systematic presentation of a neo-Aristotelian view, however, is due to Linda Zagzebski. She argues for a unified account of the intellectual and moral virtues, modeled on Aristotle's account of the moral virtues. Her view should be characterize as "neo-Aristotelian" rather than "Aristotelian," because Aristotle did not hold that the moral and intellectual virtues are unified in this way.

First, Zagzebski endorses the "direction of analysis thesis" characterized above. The distinctive feature of a virtue theory in ethics, she argues, is that it analyzes right action in terms of virtuous character, rather than the other way around.

> "By a pure virtue theory I mean a theory that makes the concept of a right act derivative from the concept of a virtue or some inner state of a person that is a component of virtue. This is a point both about conceptual priority and about moral ontology. In a pure virtue

> theory the concept of a right act is defined in terms of the concept of a virtue or a component of virtue such as motivation. Furthermore, the property of rightness is something that emerges from the inner traits of persons." (Zagzebski 1996, p. 79)

An epistemology modeled on this kind of ethical theory, then, would analyze justification and other important normative properties of belief in terms of intellectual virtue. Moreover, Zagzebski argues, we can give a unified account of moral and intellectual virtue based on an Aristotelian model of the moral virtues. In fact, she argues, intellectual virtues are best understood as a subset of the moral virtues.

According to Aristotle, the moral virtues are acquired traits of character that involve both a motivational component and a reliable success component. For example, moral courage is the virtue according to which a person is characteristically motivated to risk danger when something of value is at stake, and is reliably successful at doing so. Likewise, we can understand benevolence as the virtue according to which a person is motivated to bring about the well-being of others, and is reliably successful at doing so. Intellectual virtues have an analogous structure, Zagzebski argues. Just as all moral virtues can be understood in terms of a general motivation for the good, all intellectual virtues may be understood in terms of a general motivation for knowledge and other kinds of high-quality cognitive contact with reality. Individual intellectual virtues can then be specified in terms of more specific motivations that are related to the general motivation for knowledge. For example, open-mindedness is the virtue according to which a person is motivated to be receptive to new ideas, and is reliably successful at achieving the end of this motivation. Intellectual courage is the virtue according to which a person is motivated to be perservering in her own ideas, and is reliably successful at doing this.

Understanding the intellectual virtues this way, we can go on to define a number of important deontic properties of belief. Each definition, Zagzebski argues, is parallel to a definition for an analogous deontic property of actions.

> *A justified belief is what a person who is motivated by intellectual virtue, and who has the understanding of his cognitive situation a virtuous person would have, might believe in like circumstances.*
>
> *An unjustified belief is what a person who is motivated by intellectual virtue, and who has the understanding of his cognitive situation a virtuous person would have, would not believe in like circumstances.*
>
> *A belief of epistemic duty is what a person who is motivated by intellectual virtue, and who has the understanding of his cognitive situation a virtuous person would have, would believe in like circumstances.*

As with the moral virtues, it is possible for a conflict among the intellectual virtues to arise. Thus the intellectually courageous thing to do might conflict with the intellectually humble thing to do. This problem is solved by introducing the mediating virtue of phronesis, or practical wisdom. The practically

wise person is able to weigh the demands of all the relevant virtues is a given situation, so as to direct her cognitive activity appropriately. Accordingly we get the following definitions of "all things considered" justification.

> *A justified belief, all things considered, is what a person with phronesis might believe in like circumstances.*

> *An unjustified belief, all things considered, is what a person with phronesis would not believe in like circumstances.*

> *A belief is a duty, all things considered, just in case it is what a person with phronesis would believe in like circumstances.*

Finally, Zagzebski argues that we can give a definition of knowledge by first defining an "act of intellectual virtue".

> *An act of intellectual virtue A is an act that arises from the motivational component of A, is something a person with virtue A would (probably) do in the circumstances, is successful in achieving the end of the A motivation, and is such that the agent acquires a true belief (cognitive contact with reality) through these features of the act.*

We may then define knowledge as follows:

> *Knowledge is a state of true belief (cognitive contact with reality) arising out of acts of intellectual virtue.*

Since the truth condition is redundant, we may say alternatively:

> *Knowledge is a state of belief arising out of acts of intellectual virtue.*

The Scope of Virtue Epistemology

We have seen that a number of authors endorse a turn to virtue theory in epistemology. Moreover, these same authors have variously invoked Aristotle, Aquinas, Reid, Dewey and Peirce as early adherents of a virtue approach. This gives rise to the question of the scope of virtue epistemology. In fact, there has been some controversy about this in the contemporary literature.

Kvanvig has argued that early versions of reliabilism, including those of David Armstrong, Alvin Goldman and Robert Nozick, are best understood as versions of virtue epistemology. Although these views make explicit reference to reliable processes and reliable methods, Kvanvig argues that they are most charitably understood as concerned with reliable cognitive character. As such, they are implicit

versions of virtue epistemology. Sosa has made similar arguments regarding Alvin Plantinga's view, and Greco has argued that any view that makes justification or knowledge a function of agent reliability thereby counts as a version of virtue epistemology.

Against this, Code and Zagzebski argue that reliabilist views fail to exploit the most valuable resources of virtue theory. Moreover, Plantinga has explicitly rejected the virtue label, arguing that the foundational concept for his view is proper function rather than intellectual virtue. These disputes are not merely semantic. Rather, they reflect disagreements over what is truly of value in virtue theory. Put another way, they are disputes about what aspects of virtue theory, if any, are doing valuable work in various accounts of justification, knowledge and other important epistemic notions.

However such disputes are resolved, we have seen that there is one way of characterizing virtue epistemology so that a broad range of views count as versions of the position. Specifically, we may understand virtue epistemology primarily as a thesis about the direction of analysis: that the normative properties of beliefs are to be defined in terms of the normative properties of agents, rather than the other way around. If we understand the position this way, then a broad range of views will count as versions of virtue epistemology. We may then understand further disputes among them as concerning the nature of the intellectual virtues. In other words, different versions of virtue epistemology disagree over what kind of agent character is essentially involved in justification, knowledge and other important epistemic notions.

Bibliography

- Audi, Robert, "Epistemic Virtue and Justified Beliefs," in Fairweather and Zagzebski (2001).
- Axtell, Guy, "Epistemic-Virtue Talk: The Reemergence of American Axiology?," *The Journal of Speculative Philosophy* 10, 3 (1996):172-198.
- Axtell, Guy, "Recent Work in Virtue Epistemology," *American Philosophical Quarterly* 34, 1 (1997): 410-430.
- Axtell, Guy, "The Role of the Intellectual Virtues in the Reunification of Epistemology," *The Monist* 81, 3 (1998): 488-508.
- Axtell, Guy (ed), *Knowledge, Belief and Character: Readings in Virtue Epistemology* (Lanham: Rowman and Littlefield, 2000).
- Axtell, Guy, "Epistemic Luck in Light of the Virtues," in Fairweather and Zagzebski (2001).
- Battaly, Heather D. "Thin Concepts to the Rescue: Thinning the Concepts of Epistemic Justification and Intellectual Virtues," in Fairweather and Zagzebski (2001).
- Battaly, Heather D. "What is Virtue Epistemology?", 20th World Congress of Philosophy. [Available online]
- Code, Lorraine, "Toward a 'Responsibilist' Epistemology," *Philosophy and Phenomenological Research* XVL, I (1984): 29-50.
- Code, Lorraine, *Epistemic Responsibility* (Hanover: University Press of New England and Brown University Press, 1987).
- Dalmiya, Vrinda, "Knowing People," in Steup (2001).

- DePaul, Micheal and Linda Zagzebski (eds) *Intellectual Virtue: Perspectives from Ethics and Epistemology*. (Oxford: Oxford UP 2003).
- Driver, Julia, "Moral and Epistemic Virtue", in Axtell (2000).
- Fairweather, Abrol, "Epistemic Motivation", in Fairweather and Zagzebski (2001).
- Fairweather, Abrol and Linda Zagzebski (eds), *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility* (New York: Oxford UP, 2001).
- Goldman, Alvin, "Epistemic Folkways and Scientific Epistemology," in Goldman (1992).
- Goldman, Alvin, *Liaisons: Philosophy Meets the Cognitive and Social Sciences* (Cambridge, MA: MIT Press, 1992).
- Goldman, Alvin, "The Unity of the Epistemic Virtues" in Fairweather and Zagzebski (2001).
- Greco, John, "Virtues and Vices of Virtue Epistemology," *Canadian Journal of Philosophy* 23 (1993): 413-432.
- Greco, John, "Virtue Epistemology and the Relevant Sense of "Relevant Possibility’," *Southern Journal of Philosophy* XXXII,1 (1994): 61-77.
- Greco, John, "Agent Reliabilism," in James Tomberlin, ed., *Philosophical Perspectives, 13, Epistemology* (Atascadero, California: Ridgeview Publishing Co., 1999).
- Greco, John, *Putting Skeptics in Their Place* (New York: Cambridge University Press, 2000).
- Greco,John, "Virtue and Rules in Epistemology", in Fairweather and Zagzebski (2001).
- Greco, John, "Virtues in Epistemology" in Paul Moser (ed), *Oxford Handbook of Epistemology* (New York: Oxford UP, 2002).
- Greco, John, "Knowledge as Credit for True Beliefs", in DePaul and Zagzebski (2003).
- Greco, John (ed), *Sosa and his Critics*, (Oxford: Blackwell, 2003).
- Grimm, Stephen, "Ernest Sosa, Knowledge and Understanding" *Philosophical Studies* 106, 3 (2001): 171-191.
- Hibbs, Thomas S., "Aquinas, Virtues and Recent Epistemology," *Review of Metaphysics* 52,3 (2001): 171-191.
- Hookway, Christopher, "Mimicking Foundationalism: on Sentiment and Self-control," *European Journal of Philosophy* 1,2 (1993): 156-174.
- Hookway, Christopher, "Cognitive Virtues and Epistemic Evaluations," *International Journal of Philosophical Studies* 2,2 (1994): 211-227.
- Hookway, Christopher, "Regulating Inquiry: Virtue, Doubt and Sentiment," in Axtell (2000).
- Hookway, Christopher, "Epistemic *Akrasia* and Epistemic Virtue," in Fairweather and Zagzebski (2001).
- Kvanvig, Jonathan, *The Intellectual Virtues and the Life of the Mind* (Savage, Maryland: Rowman and Littlefield, 1992).
- Kvanvig, Jonathan (ed),*Warrant in Contemporary Epistemology: Essays in Honor of Plantiga's Theory of Knowledge*. (Lanham: Rowmann and Littlefield, 1996).
- Montmarquet, James, "Epistemic Virtue," *Mind* 96 (1987): 482-497.
- Montmarquet, James, *Epistemic Virtue and Doxastic Responsibility* (Lanham, MD: Rowman and Littlefield, 1993).
- Montmarquet, James, "An 'Internalist' Conception of Epistemic Virtue," in Axtell (2000).
- Plantinga, Alvin, *Warrant and Proper Function* (Oxford: Oxford University Press, 1993).
- Plantinga, Alvin, "Why We Need Proper Function," *Nous* 27, 1 (1993): 66-82

- Riggs, Wayne, "Reliability and the Value of Knowledge," *Philosophy and Phenomenological Research* LXIV, I (2002): 79-96.
- Sosa, Ernest, "The Raft and the Pyramid: Coherence versus Foundations in the Theory of Knowledge," *Midwest Studies in Philosophy* V (1980): 3-25. Reprinted in Sosa (1991).
- Sosa, Ernest, "The Coherence of Virtue and the Virtue of Coherence: Justification in Epistemology," *Synthese* 64 (1985): 3-28. Reprinted in Sosa (1991).
- Sosa, Ernest, *Knowledge in Perspective (Cambridge: Cambridge University Press, 1991).*
- *Sosa, Ernest, "Proper Functionalism and Virtue Epistemology," Nous 27, 1 (1993): 51-65.*
- *Sosa, Ernest, "Virtue Perspectivism: A Response to Foley and Fumerton" in Enrique Villanueva (ed), Truth and Rationality (Ridgeview: Atascadero, 1994).*
- *Sosa, Ernest, "How must Knowledge be Modally Related to what is Known?" in Philosophical Topics 26, 1 & 2 (1999): 373-384.*
- *Sosa, Ernest, "For the Love of Truth?", in Fairweather and Zagzebski (2001).*
- *Sosa, Ernest, "Virtue Epistemology", in Ernest Sosa and Lawrence BonJour, Epistemology: Internalism versus Externalism (Oxford: Blackwell, Forthcoming).*
- *Steup, Matthais (ed), Knowledge, Truth and Duty: Essays on Epistemic Justification, Responsibility and Virtue. (New York: Oxford UP, 2001).*
- *Taliaferro, Charles, "The Virtues of Embodiment," Philosophy 76 (2001):111-125.*
- *Woods, Jay. Epistemology: Becoming Intellectually Virtuous (Grand Rapids: Intervarsity Press, 1998).*
- *Zagzebski, Linda, Virtues of the Mind (Cambridge: Cambridge University Press, 1996).*
- *Zagzebski, Linda, "Virtue in Ethics and Epistemology," American Catholic Philosophical Quarterly 71 (Supp) (1997): 1-17.*
- *Zagzebski, Linda, "Virtue Epistemology" in Encyclopedia of Philosophy (New York: Routledge, 1998).*
- *Zagzebski, Linda, "What is Knowledge?", in Greco, John and Sosa, Ernest, eds., The Blackwell Guide to Epistemology (Oxford: Blackwell, 1999).*
- *Zagzebski, Linda, "From Reliablilism to Virtue Epistemology", in Axtell (2000).*
- *Zagzebski, Linda, "Must Knowers be Agents?", in Fairweather and Zagzebski (2001).*
- *Zagzebski, Linda, "Recovering Understanding", in Steup (2001).*

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

justification, epistemic: coherentist theories of | justification, epistemic: foundationalist theories of | justification, epistemic: internalist vs. externalist conceptions of | skepticism

John Greco
greco@fordham.edu

A | *B* | *C* | *D* | *E* | *F* | *G* | *H* | *I* | *J* | *K* | *L* | *M* | *N* | *O* | *P* | *Q* | *R* | *S* | *T* | *U* | *V* | *W* | *X* | *Y* | *Z*

Table of Contents

First published: July 9, 1999
Content last modified: July 10, 2002

Foundationalist Theories of Epistemic Justification

Foundationalism is a view about the structure of justification or knowledge. The foundationalist's thesis in short is that all knowledge and justified belief rest ultimately on a foundation of noninferential knowledge or justified belief.

A little reflection suggests that the vast majority of the propositions we know or justifiably believe have that status only because we know or justifiably believe other different propositions. So, for example, I know or justifiably believe that Caesar was an assassinated Roman leader, but only because I know or justifiably believe (among other things) that various historical texts describe the event. Arguably, my knowledge (justified belief) about Caesar's death also depends on my knowing (justifiably believing) that the texts in question are reliable guides to the past. Foundationalists want to contrast my inferential knowledge (justified belief) about Caesar with a kind of knowledge (justified belief) that doesn't involve the having of other knowledge (justified belief). There is no standard terminology for what we shall henceforth refer to as noninferential knowledge or justification.[1]

For convenience, in what follows we will concentrate on foundationalism about justification. Everything said about justified belief will apply mutatis mutandis to foundationalist views about knowledge. On the "classical" analysis of knowledge, the core of the concept of knowledge *is* justified true belief and the foundational structure of knowledge simply derives from the foundational structure or justification.

It is surely fair to suggest that for literally thousands of years the foundationalist's thesis was taken to be almost trivially true. When an argument was implicitly or explicitly offered for the view it was most often the now famous regress argument. It is important, however, to distinguish two quite different regress arguments for foundationalism--the epistemic regress argument and the conceptual regress argument.

---

1. The Regress Arguments for Foundationalism

Suppose I claim to be justified in believing that Fred will die shortly and offer as my evidence that Fred has an untreatable and serious form of cancer. Concerned, you ask me how I discovered that Fred has the cancer and I respond that it is just a hunch on my part. As soon as you discover that I have no reason at all to suppose that Fred has the cancer, you will immediately conclude that my whimsical belief about Fred's condition gives me no justification for believing that Fred will soon die. Generalizing, one might suggest the following principle:

> To be justified in believing P on the basis of E one must be justified in believing E.

Now consider another example. Suppose I claim to be justified in believing that Fred will die shortly and offer as my justification that a certain line across his palm (his infamous "lifeline") is short. Rightly skeptical, you wonder this time what reason I have for believing that palm lines have anything whatsoever to do with length of life. As soon as you become satisfied that I have no justification for supposing that there is any kind of probabilistic connection between the character of this line and Fred's life you will again reject my claim to have a justified belief about Fred's impending demise. That suggests that we might expand our Principle of Inferential Justification (PIJ) to include a second clause:

> **Principle of Inferential Justification:**
> To be justified in believing P on the basis of E one must not only be (1) justified in believing E, but also (2) justified in believing that E makes probable P.

With PIJ one can present a relatively straightforward *epistemic* regress argument for foundationalism. If all justification were inferential then for someone S to be justified in believing some proposition P, S must be in a position to legitimately infer it from some other proposition E1. But E1 could justify S in believing P only if S were justified in believing E1, and if all justification were inferential the only way for S to do that would be to infer it from some other proposition justifiably believed, E2, a proposition which in turn would have to be inferred from some other proposition E3 which is justifiably believed, and so on, ad infinitum. But finite beings cannot complete an infinitely long chain of reasoning and so if all justification were inferential no-one would be justified in believing anything at all to any extent whatsoever. This most radical of all skepticisms is absurd (it entails that one couldn't even be justified in believing it) and so there must be a kind of justification which is not inferential, i.e. there must be noninferentially justified beliefs which terminate regresses of justification.

If we accept the more controversial second clause of PIJ, the looming regresses proliferate. Not only must S above be justified in believing E1, S must also be justified in believing that E1 makes likely P, a proposition he would have to infer (if there are no foundations) from some other proposition F1, which he would have to infer from F2, which he would have to infer from F3, and so on ad infinitum. But S would also need to be justified in believing that F1 does in fact make likely that E1 makes likely P, a proposition he would need to infer from some other proposition G1, which he would need to infer from some other proposition G2... . And he would need to infer that G1 does indeed make likely that F1 makes likely that E1 makes likely P... . Without noninferentially justified beliefs, it would seem that we would need to complete an infinite number of infinitely long chains of reasoning in order to be justified in believing anything!

The above argument relies on the unacceptability of a vicious *epistemic* regress. But one might also argue, more fundamentally, that without a *concept* of noninferential justification, one faces a vicious *conceptual* regress. What precisely is our *understanding* of inferential justification? What makes PIJ true? It is at least tempting to answer that PIJ is an analytic truth. Part of what it *means* to claim that someone has inferential justification for believing some proposition P is that his justification consists in his ability to infer P from some other proposition E1 that is justifiably believed. But if anything like this is a plausible analysis of the concept of inferential justification, we face a potential vicious conceptual regress. The analysis of inferential justification presupposes an understanding of justified belief. We need to introduce a concept of noninferential justification in terms of which we can then recursively define inferential justification.

Consider an analogy. Suppose a philosopher introduces the notion of instrumental goodness (something's being good as a means). That philosopher offers the following crude analysis of what it is for something to be instrumentally good. X is instrumentally good when X leads to something Y which is good. Even if we were to accept this analysis of instrumental goodness, it is clear that we haven't yet located the conceptual source of goodness. Our analysis of instrumental goodness presupposes an understanding of what it is for something to be good. In short we can't understand what it is for something to be instrumentally good until we have some prior (and more fundamental) understanding of what it is for something to be intrinsically good. The conceptual regress argument for foundationalism puts forth the thesis that inferential justification stands to noninferential justification as instrumental goodness stands to intrinsic goodness.

2. The Analysis of Noninferential Justification

If foundationalists are united in their conviction that there must be a kind of justification that does not depend on the having of other justified beliefs, they nevertheless disagree radically among themselves as to how to understand noninferential justification. In the latter part of this century, the rise of externalist epistemologies has generated even more fundamentally different versions of foundationalism. It will not be possible to survey all of the strikingly different analyses that have been offered of noninferential justification. In what follows we will examine a few of the more prominent versions of classical and contemporary externalist foundationalisms.

2.1 Noninferential Justification as Infallible Belief

Descartes is often taken to be the paradigm of a classical foundationalist. Determined to build knowledge on appropriate and secure foundations he seemed to want to identify foundational knowledge with infallible belief. Implicitly or explicitly others seemed to follow his lead by restricting noninferentially justified beliefs to beliefs that cannot be mistaken. Thus, for example, when Price (1950) introduced the notion of sense data, knowledge of which would be included in his foundations of empirical knowledge, he contrasted sense data and their nonrelational properties with other sorts of things about which one could be mistaken, implying again that the way to find the correct foundations of knowledge is to eliminate from one's beliefs system all those beliefs that could be false. Following Lehrer (1974, p. 81)) we might formulate the following definition of infallible belief:

> S's belief that P at t is infallible if S's believing P at t entails[2] that P is true.

As Lehrer and others have pointed, it is far from clear that this concept of infallible belief has much relevance to an attempt to understand the epistemic concept of noninferential justification. The first and most striking problem involves necessary truths. Every necessary truth is entailed by every proposition, and thus if I happen to believe a necessary truth, P, that I believe P will entail that P is true. Thus by the above definition my belief that P will be infallible whenever P is a necessary truth even if P is far too complicated for me to prove and I believe it solely on a whim.

Furthermore, a foundation of knowledge and justified belief restricted to infallible beliefs (as defined above) would arguably be far too flimsy to support any sort of substantial epistemic edifice. There are a few contingent propositions that are trivially entailed by the fact that they are believed. My belief that I exist entails that I exist, that I have at least one belief, that someone has beliefs, that experience (broadly construed) exists, etc. But once we get past these sorts of "self-referential" propositions, propositions whose very subject matter encompasses the fact that they are believed, it is hard to come up with uncontroversial examples of infallible beliefs. Ayer (1956, p. 19) argues that as long as the belief that P is one state of affairs and P's being the case is an entirely different state of affairs (not including as a constituent the former) there can be no logical absurdity in the supposition that the former could occur without the latter.

Although it doesn't add much to the logical force of the argument, one might employ our hunches about how the brain might work to rhetorically bolster the argument. Consider a standard candidate for an infallible empirical belief, my belief that I am in pain now. It is surely possible that the region of the brain causally responsible for producing the belief that I am in pain is entirely different from the region of the brain causally responsible for producing the pain. There may be a causal connection between the occurrence of the "pain" brain event and the occurrence of the "belief" brain event, or vice versa, but even if the causal connection holds it will be a contingent fact that it does. It hardly seems that the neurophysiologist could discover these (or any other) causal connections purely a priori. But if the brain state responsible for my belief that I am in pain is wholly different from the brain state responsible for

the pain, and if the connections between them are merely nomological, then it is in principle possible to produce the one without the other. The belief will not entail the truth of what is believed.

2.2 Noninferential Justification as Infallible Justification

It may be that classical foundationalists start off on the wrong foot if they seek foundations in logical relations between the mere fact that someone believes some proposition and the proposition's being true. Noninferential justification is, after all, a kind of justification and if the impossibility of error is essential to noninferential justification, it may be more plausible to locate the source of infallibility in a special kind of justification available in support of a belief. Let us say that S's belief is infallibly justified at t when S's justification for believing P at t relevantly entails the truth of P. We need to qualify the entailment as relevant to circumvent the problems discussed earlier. Whenever I have any justification at all for believing a proposition that turns out to be necessarily true, that justification will entail the necessary truth. But we do not want just any sort of justification to yield infallibly justified belief even if the object of that belief is a necessary truth.

What is the difference between relevant and irrelevant entailment? This is a question notoriously difficult to answer, but intuitively it should have something to do with the fact that would make true the proposition entailed and the fact that would make true the proposition that entails it. More specifically, we could say that P relevantly entails Q only if the fact that would make P true is at least a constituent of the fact that would make Q true. This suggestion can be considered at best only preliminary since we will obviously need a more detailed account of facts and their constituents. That I have grey hair entails that someone has grey hair, but is my having grey hair a constituent of the fact that is someone's having grey hair? There is certainly a sense in which it is something one can point to in answer to the question "What makes it true that someone has grey hair?" One cannot appropriately point to my having grey hair as something that makes it true that two plus two equals four.

Consider again my belief that I'm in pain (when I am). If such a belief is noninferentially justified, in what does the justification for that belief consist. Surely not in the mere fact that the proposition is believed. What is it that distinguishes this belief from my belief about Caesar's assassination. Some foundationalists want to locate the noninferential justification in the truth-maker for the proposition believed. What justifies me in believing that I'm in pain when I am is the mere fact that I'm in pain. But again, what is it about my being in pain as opposed to Caesar's being assassinated which makes it appropriate to claim that my being in pain justifies me in believing that I'm in pain while Caesar's having been assassinated doesn't justify me in believing that Caesar was assassinated.

It is tempting to think that the foundationalist is better off appealing to some special *relation* that I have to my pain which makes it unnecessary to look to other beliefs in order to justify my belief that I'm in pain. It is the fact that I have a kind of *access* to my pain that no-one else has that makes my belief noninferentially justified while others must rely on inference in order to discover that I am in this state. This takes us to another classical version of foundationalism, the acquaintance theory. Perhaps the best known proponent of an acquaintance theory is Bertrand Russell,[3] but it takes little imagination to read

the view into most of the British empiricists. Roughly the view is that what justifies S in believing that he is in pain when he does is the fact that S is directly and immediately acquainted with his pain in a way in which he is not directly and immediately acquainted with any contingent facts about Caesar, the physical world, the future, and so on. On a correspondence conception of truth, one might add that to be fully justified in believing a proposition to be true one must be acquainted not only with the fact that makes the proposition true but the relation of correspondence that holds between the proposition and the fact.

In one of the most influential arguments against foundationalism, Wilfrid Sellars (1963, 131-32) argued that the idea of the given in traditional epistemology contains irreconcilable tensions. On the one hand, to ensure that something's being given does not involve any other beliefs, proponents of the view want the given to be untainted by the application of *concepts*. On the other hand, the whole doctrine of the given is designed to end the regress of justification, to give us secure foundations for the rest of what we justifiably infer from the given. But to make sense of making inferences from the given the given must have a truth value. The kind of thing that has a truth value involves the application of concepts or thought, a capacity not possessed (we may presume) by at least lower-order animals.

If there is a solution to the dilemma presented by Sellars (and others) it is to emphasize that acquaintance is not by itself an epistemic relation. Acquaintance is a relation that other animals might bear to properties and even facts, but it also probably does not give these animals any kind of justification for *believing* anything, precisely because these other animals probably do not have beliefs. Without thought or propositions entertained there is no truth, and without a bearer of truth value in the picture there is nothing to be justified or unjustified. The acquaintance theorist can argue that one has a noninferentially justified belief that P only when one has the thought that P and one is acquainted with both the fact that P, the thought that P, and the relation of correspondence holding between the thought that P and the fact that P. On such a view no single act of acquaintance yields knowledge or justified belief, but when one has the relevant thought (entertains the relevant proposition), the three acts together constitute noninferential justification. When everything that is constitutive of a thought or a proposition's being true is immediately before consciousness, there is nothing more that one could want or need to justify a belief. The state that constitutes noninferential justification is a state that contains as constituents both the bearer of truth-value and the truth-maker.[4]

When an acquaintance with the fact that P is part of what constitutes my noninferential justification for believing P, there is a trivial sense in which my noninferential justification is infallible. I can't be directly acquainted with the fact that P while I believe P falsely. There is, however, nothing to prevent an acquaintance theorist from allowing that one can be noninferentially justified in believing P by virtue of being directly acquainted with a fact very similar to, but ultimately different from the fact that P (the fact that makes P true). Such an acquaintance theory could allow for the possibility of noninferentially justified but false belief that P.[5]

3. Objections to Classical Foundationalism

Once the received view, classical foundationalism has come under considerable attack in the last few

decades. We have already considered the very influential objection raised by Sellars to the idea of there being a "given" element in experience. It is crucial that the foundationalist discover a kind of *truth* that can be known without inference. But there can be no bearers of truth value without judgment and judgment involves the application of concepts. But to apply a concept is to make a judgment about class membership, and to make a judgment about class membership always involves relating the thing about which the judgment is made to other paradigm members of the class. These judgments of relevant similarity will minimally involve beliefs about the past, and thus be inferential in character (assuming that we can have no "direct" access to facts about the past). A reply to this objection would take us far afield indeed. Perhaps it will suffice to observe that the objection relies on a number of highly controversial claims about the nature of judgment, most of which the classical foundationalist should and would reject.

The direct acquaintance theorist does presuppose the intelligibility of a world that has "structure" independent of any structure imposed by the mind. Without nonlinguistic facts that are independent of the thoughts and judgments that represent them, one could not make sense of a relation of acquaintance between a person and a fact, a relation that grounds noninferential justification. More radical contemporary rejections of foundationalism may well involve dissastification with the foundationalist's implicit commitment to a strong realistic correspondence conception of truth. Since Kant there has always been a strong undercurrent of anti-realism running through philosophy. The metaphor is that of the mind imposing structure on reality. And there is an intuitively plausible sense in which one can genuinely wonder whether it makes sense to ask about the number of colors that are exemplified in the world independently of some framework provided by color *concepts*. But despite the periodic popularity of extreme anti-realism, it is surely absurd to suppose that it is even in principle possible for a mind to force a structure on a literally unstructured world. There are indefinitely many ways to sort the books in a library and some are just as useful as others, but there would be no way to begin sorting books were books undifferentiated. If a rejection of foundationalism relies on an extreme form of anti-realism so much the worse for the anti-foundationalist.

Just as some anti-foundationalists reject the conception of truth underlying classical foundationalist accounts of noninferential justification, so others profess to be bewildered by some of the fundamental concepts employed in defining noninferential justification. The acquaintance theorist tends to have relatively little to say by way of analyzing what direct acquaintance is. To be sure one can try to give someone a feel for what one is talking about by contrasting one's awareness of pain with the temporary distraction caused by an engrossing conversation. It is tempting to suppose that for a short time the pain was still present but the person with the pain was no longer aware of the fact that the pain exists. This awareness, the acquaintance theorist will argue, is obviously something over and above mere belief in the existence of the pain, as one can believe that one is in a mental state (say a subconscious mental state) without being aware of that state. Like most theories foundationalism will, however, ultimately rest its intelligibility on an appeal to a sui generis concept that defies further analysis. Just as one needs to end epistemic regresses with foundational justification, the foundationalist will argue, so one needs to end conceptual regresses with concepts one grasps without further definition.

Laurence BonJour (1985) raised another highly influential objection to all forms of classical

foundationalism (an objection raised before he joined the ranks of foundationalists). The objection presupposed a strong form of what we might call access internalism. Put very superficially the access internalist argues that a feature of a belief or epistemic situation that makes a belief noninferentially justified must be a feature to which we have actual or potential access. Moreover, we must have access to the fact that the feature in question is probabilistically related to the truth of what we believe. So suppose some foundationalist offers an account of noninferential justification according to which a belief is noninferentially justified if it has some characteristic X. BonJour then argues that the mere fact that the belief has X could not, even in principle, justify the believer in holding the belief. The believer would also need access to (justified belief that!) the belief in question has X and that beliefs of this sort (X beliefs) are likely to be true. At least one of these propositions could only be known through inference, and thus the putative noninferential justification is destroyed.

BonJour presented the objection on the way to developing a coherence theory of empirical justification. But it ultimately became obvious that the objection to foundationalism, if good, was too strong. Given the structure of the argument it should become evident that the coherence theory (and any other theory) would be equally vulnerable to the argument. Just replace "X" with some complicated description of beliefs cohering with each other. That might suggest to the classical foundationalist that strong access internalism is a view to be avoided.

The Principle of Inferential Justification used to generate the regress argument for foundationalism is itself controversial. It is important to note that either clause of the principle can be used by itself to generate the allegedly vicious epistemic and conceptual regress for the philosopher who rejects foundations. It is the two clauses combined that are supposed to present the anti-foundationalist with an infinite number of vicious regresses. A number of philosophers (among them foundationalists) would argue that the second clause of PIJ confuses levels of epistemic questions. It is far too strong to require someone to have a justified belief in a probabilistic connection between available evidence and the conclusion reached on the basis of that evidence. Such a requirement is at best plausible for having second-level justification for believing that one has an inferentially justified belief. In responding to a challenge presented to one's having an inferentially justified belief in P on the basis of E one might find oneself searching for justification to support the claim that E makes probable P, but that is only because in the context of the challenge one is trying to make good (i.e. justify) the claim that one has a justified belief. A similar claim might be made with respect to clause 1) of the principle, although it is more difficult to generate the supporting intuition.

In any event, the careful foundationalist is certainly not *confused* about level-distinctions. The foundationalist who supports PIJ is claiming that a necessary condition for someone's having an inferentially justified belief in P based on E is that the person have both a justified belief in E and a justified belief in the proposition that E makes P probable. It is simply not enough that E is true or that E does in fact make probable P. Our original examples used to support PIJ would seem to reinforce that conclusion. Even if there happened to be some bizarre connection between palm lines and length of life, for example, the person who has no reason to believe that such a connection exists has no justification for conclusions reached about length of life based on this anatomical feature of people.

There are, of course, other responses to the charge of vicious regress facing anti-foundationalists. The coherence theorist rejects the foundationalist's presupposition that justification is linear. Each belief is justified by virtue of its coherence with the rest of what one believes but one avoids the appearance of vicious circularity by insisting that one needn't *first* have justification for believing the other propositions in one's belief system. The coherence theorist's response to the argument for foundationalism is, of course, only as plausible as the coherence theory of justification (See coherence theories of justification).

Peter Klein (1998) may be the lone supporter of a view he calls infinitism. The infinitist accepts the need to be *able* to supply non-circular justification for believing what we do, but argues that given the complexity of the human mind and its capacity to entertain and justifiably believe an infinite number of propositions, there is nothing vicious about the relevant regresses we face. There is no reason to suppose that we would be unable to justify every proposition we believe by appeal to some other different proposition which we justifiably believe. Infinitism is a view that should be seriously considered, particularly once one realizes that one not only can but does have an infinite number of justified beliefs (e.g., that 2 is greater than 1, that 3 is greater that 1, and so on.). It is not clear, however, that even if the infinitist can cope with the epistemic regress argument foundationalism, he has a response to the conceptual regress argument discussed earlier.

Although anti-foundationalists are not always eager to admit it, I suspect that the primary dissatisfaction with classical foundationalism lies with the difficulty the view has avoiding radical skepticism. On infallible belief, infallible justification, or direct acquaintance theories of foundational justification, there is precious little included in the foundations of knowledge. Most classical foundationalists reject the idea that one can have noninferentially justified beliefs about the past, but the present disappears into the past in the blink of an eye. How can one even hope to get back the vast body of knowledge one pre-philosophically supposes one has, if one's epistemic base is so impoverished. If the second clause of the Principle of Inferential Justification were accepted, the problem is even more serious. One might be able to convince oneself that one can know noninferentially the principles of deductive reasoning, but deduction will not take one usefully beyond the foundations of knowledge and justified belief. As Mill (1906, p. 126) argued, there is a very real sense in which one doesn't advance one's knowledge significantly employing a form of reasoning that takes one only to conclusions that were implicitly contained in the conjunction of one's premises. To advance beyond foundations we will inevitably need to employ non-deductive reasoning and according to PIJ that will ultimately require us to have noninferential (direct) knowledge of propositions describing probability connections between evidence and conclusions. It is not absurd on the face of it to suppose that one can have noninferential a priori knowledge of probabilistic connections, but it is perhaps an understatement to suppose that the view is not popular.[6]

4. Externalist Versions of Foundationalism

The epistemic landscape has changed dramatically in the last quarter of a century with the rise of externalist epistemologies. It is notoriously difficult to define clearly the controversy between internalists and externalists in epistemology.[7] It is sometimes taken to be a controversy over whether or not one can

identify epistemic properties with "internal" states of believers. Others seem to think that the controversy centers over the question of whether one requires certain sorts of access (or potential access) to the states or properties that constitute having justification. Certainly, paradigm externalists would reject the second clause of the principle of inferential justification. According to virtually all externalists, one can arrive at a justified belief in P by inferring it from E without being aware of any sort of evidential connection between E and P.

While the externalist defends radically different views than those of classical foundationalists, the structure of knowledge and justification that emerges from such theories is still often a foundationalist structure. We might first illustrate the point by examining the view defended by the most prominent of the externalists, Alvin Goldman's reliabilism.[8]

The fundamental idea behind reliabilism is strikingly simple. Justified beliefs are reliably produced beliefs. Justified beliefs are worth having because justified beliefs are probably true. Goldman initially distinguished, however, two importantly different sorts of justified beliefs--those that result from belief-independent processes and those that result from belief-dependent processes. The former are beliefs that are produced by "software" of the brain that takes as its "input" stimuli other than beliefs; the latter are beliefs produced by processes that take as their input at least some other beliefs.. So, for example, it is possible that we have evolved in such a way that when prompted with certain sensory input we immediately and unreflectively reach conclusions about external objects. And we may live in a world in which beliefs about the external world produced in this way are usually true (or would usually be true if enough of them were generated).[9] Such beliefs will be justified by virtue of being the product of reliable belief-independent processes. They can in turn be taken as input for reliable belief-dependent processes in order to generate still more justified beliefs. A belief-dependent process is reliable if its out beliefs are usually (or would usually) be true if the relevant input beliefs are true, and the output beliefs of reliable belief-independent processes are justified provided that the input beliefs are justified.[10]

The above is but the crudest sketch of Goldman's early reliabilism--he later modified it to deal with a number of objections. But the sketch is enough to bring out the foundationalist structure inherent in a reliabilist account. The reliabilist actually accepts the first clause of PIJ, but avoids both the epistemic and conceptual regresses by embracing a kind of justified belief that does not owe its justification to the having of other different justified beliefs. That the reliabilist is concerned with avoiding the conceptual regress is clear from the fact that the analysis offered is explicitly recursive. The base clause of the recursive analysis in effect captures the concept of a noninferentially justified belief.

I have illustrated the way in which an externalist account of justified belief can exemplify a foundationalist structure by examining one of the most prominent versions of externalism, reliabilism. But other versions of externalism are also implicitly or explicitly committed to a version of foundationalism, or, at the very least, give an account of justification that would enable one to distinguish noninferential from inferential justification, direct from indirect knowledge. Consider, for example, a crude version of the so-called causal theory of knowledge according to which one knows a proposition when one believes it and the belief is caused (in the "right" way) by the very fact that makes true what is

believed. Obviously, on such an account one can distinguish causal chains leading to the belief in question that involve intermediate beliefs from those that do not, and using this distinction one can again define a distinction between direct and indirect knowledge.[11]

Externalist versions of foundationalism are probably attractive to many because they often allow at least the possibility of a much expanded foundational base of justified beliefs. The reliabilist's noninferentially justified beliefs, for example, might be produced by processes that are not even very reliable. Unlike the Cartesian, the reliabilist's distinction between noninferentially and inferentially justified belief has nothing to do with how probable it is that the belief in question is true. If nature has been co-operative enough to insure the evolution of cognitive agents who respond to their environmental stimuli with mostly true beliefs then there might be an enormous store of foundational knowledge upon which we can draw in arriving at inferentially justified conclusions. On most externalist accounts of noninferentially justified belief there are literally no *a priori* constraints on what might end up being noninferentially justified.

A full evaluation of externalist versions of foundationalism is far beyond the scope of this article. The very ease with which the externalist can potentially broaden the foundational base of noninferentially justified belief is, ironically, one of the primary concerns of those philosophers unhappy with externalist epistemology. Many internalists are convinced that externalists are simply re-defining epistemic terms in such a way that they lose the kind of meaning that the philosopher wants them to have in order to ask the kind of penetrating philosophical questions that are the peculiar product of a kind of philosophical curiosity. When a philosopher starts looking for justification in support of a belief, the internalist will argue, the philosopher is interested in achieving a state in which a kind of philosophical curiosity is satisfied. That philosopher wants epistemic justification to provide a kind of assurance of truth. If I'm wondering whether or not I have justification to believe that God exists, I'm hardly going to think that my question has been answered when I'm told by the reliabilist that I might have a reliably produced belief that God exists or when I'm told by the causal theorist that my belief that God exists might be caused by the very fact that God exists. As far as satisfying intellectual curiosity, exemplifying reliably-produced belief or belief caused by the right fact is no more useful than having true belief. If I were to stipulate a technical sense of foundational knowledge* according to which I foundationally know that P when I believe truly that P and my belief isn't caused by any other belief, there may well be all sorts of truths I "know", but will having such knowledge do me any good as far as putting me in a state that satisfies my philosophical curiosity?

Bibliography

- Armstrong, David. 1973. *Belief, Truth and Knowledge*. London: Cambridge University Press.
- Ayer, A. J. 1956. *The Problem of Knowledge*. London: Cambridge University Press.
- BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge: Harvard University Press.
- ______. 2000. "Toward a Defense of Empirical Foundationalism." In *Resuurecting Old-Fashioned Foundationalism*, ed. Michael DePaul. Lanham, Ma.: Rowman and Littlefield.

- Fumerton, Richard. 1995. *Metaepistemology and Skepticism*. Lanham, Ma.: Rowman and Littlefield.
- Goldman, Alvin. 1979. "What is Justified Belief?" Pp. 1-23 in *Justification and Knowledg*, ed. George Pappas. Dordrecht: Reidel.
- ______. 1986. *Epistemology and Cognition*. Cambridge, Mass.: Harvard University Press.
- ______. 1988. "Strong and Weak Justification." Pp. 51-69 in *Philosophical Perspectives 2: Epistemology*, ed. James Tomberlin. Atascadero, Calif.: Ridgeview Publishing Co.
- Klein, Peter. 1998. "Foundationalism and the Infinite Regress of Reasons." *Philosophy and Phenomenological Research* LVIII: 919-26.
- Lehrer, Keith. 1974. *Knowledge*. Oxford: Clarendon Press.
- Mill, John Stuart. 1906. *A System of Logic*. London: Longmans, Green, and Co.
- Nozick, Robert. 1981. *Philosophical Explanations*. Cambridge: Harvard University Press.
- Price, H. H. 1950. *Perception*. London: Methuen.
- Russell, Bertrand. 1910-11. "Knowledge by Acquaintance and Knowledge by Description." *The Proceedings of the Aristotelian Society*, Vol. 11: 209-32.
- ______. *Theory of Knowledge: The 1913 Manuscript*. Ed. by Elizabeth Eames. London: Allen and Unwin Ltd.
- ______. 1948. *Human Knowledge: Its Scope and Limits*. New York: Simon and Schuster.
- Sellars, Wilfrid. 1963. *Science Perception and Reality*. London: Routledge and Kegan Paul.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

justification, epistemic: coherentist theories of | justification, epistemic: contextualist theories of | justification, epistemic: internalist vs. externalist conceptions of

Richard Fumerton
fumerton@blue.weeg.uiowa.edu

First published: February 21, 2000

Content last modified: February 21, 2000

Stanford Encyclopedia of Philosophy
Notes to Foundationalist Theories of Epistemic Justification

Notes

1. Foundational knowledge or justified belief has also been called by foundationalists direct knowledge (justification), immediate knowledge, intuitive knowledge (justification); and the truths known have been referred to as self-evident truths, directly evident truths, self-presenting truths, and the given. This last locution "the given" is, however, ambiguous as between *truths* that are said to be known directly and *facts* or features of the world that are said to be immediately "before" consciousness.

2. For present purposes let us construe entailment broadly so that P may be said to entail Q if P formally, analytically or synthetically entails Q.

3. See, for example, Russell (1910-11) and 1913.

4. For a more detailed account and defense of an acquaintance theory of noninferential justification see Fumerton (1995). BonJour, once one of the leading coherence theorists of empirical justification, has recently moved to a version of the acquaintance theory of justification. See BonJour (2000).

5. For a further elaboration of such a view see Fumerton (1995).

6. For an excellent discussion of this issue see Russell (1948).

7. For a detailed discussion of alternative ways of defining the internalism/externalism controversy, see Fumerton (1995), Chapters 3 and 4.

8. Most of what I say here is based on the early seminal paper "What is Justified Belief." Goldman's view changed quite dramatically in his book *Epistemology and Cognition*, but shortly after publishing the book he returned to the earlier account for at least one conception of justification (strong justification). See Goldman (1988).

9. We shall not concern ourselves with the difficulties that reliabilists face defining the relevant notion of reliability--as these few remarks might indicate, reliabilists will inevitably move beyond actual frequencies and turn to propensities or counterfactuals in defining the concept of a reliable belief-producing process.

10. For technical reasons, it might, perhaps, be better to require that the conjunction of the propositions

believed must itself be the object of a justified belief.

11. See, for example, Armstrong's (1973) account of direct knowledge. Though more complicated that a causal theory of knowledge, Nozick's (1981) "tracking" account of knowledge also allows a distinction between beliefs which noninferentially track facts and beliefs which inferentially track facts.

Robert Fumerton
fumerton@blue.weeg.uiowa.edu

First published: February 21, 2000
Content last modified: February 21, 2000

Skepticism

Much of epistemology has arisen in defense or in opposition to various forms of skepticism. Indeed, one could classify various theories of knowledge by their responses to skepticism. For example, rationalists could be viewed as skeptical about the possibility of empirical knowledge while not being skeptical with regard to a priori knowledge and empiricists could be seen as skeptical about the possibility of a priori knowledge but not so with regard to empirical knowledge. In addition, many traditional problems, for example the problem of other minds or the problem of our knowledge of God's existence, can be seen as *restricted* forms of skepticism which hold that we cannot have knowledge of any propositions in some particular domain thought to be within our ken. Although this essay will comment briefly about some restricted forms of skepticism, it will focus on the *general* forms of skepticism which question our knowledge in many, if not all, domains in which we ordinarily think knowledge is possible. Since this essay is not primarily devoted to a discussion of the history of philosophical skepticism, the general forms of skepticism to be discussed are those which contemporary philosophers still find the most interesting.

1. Philosophical Skepticism vs. Ordinary Incredulity

Even before examining the various forms of skepticism, it is crucial that we distinguish between philosophical skepticism and ordinary doubt because doing so will help to explain why philosophical skepticism is so intriguing. Consider an ordinary case in which we think someone fails to have knowledge. Suppose Anne claims that she knows that the bird she is looking at is a robin and that I believe that if Anne were to look carefully, she would see that its coloration is not quite that of a robin. Its breast is too orange. Further, I believe that it flies somewhat differently than robins do. This bird seems to flitter more than a typical robin.

Thus, there are two grounds for doubting that Anne knows that this is a robin:

a. The color of this bird isn't typical of robins;
b. The flight pattern of this bird is not typical of robins.

Now, what makes this a case of ordinary doubt is that there are, in principle, two ways of removing the basis for doubt:

i. Anne could show that the alleged grounds for doubt are false; or
ii. Anne could show that the grounds for doubt, though true, can be neutralized.[1]

Taking alternative (i), Anne could show that there are many robins with the coloration of the bird in question by citing the *Audubon Field Guide for Birds* in which many of the pictured robins have very orange breasts. In other words, Anne could show that (a) is false.

But in order to remove grounds for doubt, it is not necessary that Anne show that the alleged grounds are false. Consider ground (b). It could be granted that the bird in question flies in a way that is not at all typical of robins. But suppose that on closer inspection we see that some of its tail feathers have been damaged in a way that could account for the unusual flight pattern. Because the bird has difficulty gliding and flying in a straight line, it flaps its wings much more rapidly than is typical of robins. Thus, although we can grant that (b) is true, we would have explained away, or neutralized, the grounds for doubt.

The point here is that in this case, and in all *ordinary* cases of incredulity, the grounds for the doubt can, in principle, be removed. As Wittgenstein would say, doubt occurs within the context of things undoubted. If something is doubted, something else must be held fast because doubt presupposes that there are means of removing the doubt.[2] We doubt that the bird is a robin because, at least in part, we think we know how robins typically fly and what their typical coloration is. That is, we think our general picture of the world is right -- or right enough -- so that it does provide us with both the grounds for doubt and the means for potentially removing the doubt. Thus, ordinary incredulity, say about some feature of the world, occurs against a background of sequestered beliefs about the world. We are not doubting that we have any knowledge of the world. Far from it, we are presupposing that we do know some things about the world. To quote Wittgenstein, "A doubt without an end is not even a doubt" (Wittgenstein 1969, ¶ 625).

In contrast, philosophical skepticism attempts to render doubtful *every* member of a class of propositions that we think falls within our ken. One member of the class is not pitted against another. The grounds for either withholding assent to the claim that we can have such knowledge or denying that we can have such knowledge are such that there is no possible way to either answer them or neutralize them by appealing to another member of the class. Thus, philosophic doubt or philosophical skepticism, as opposed to ordinary incredulity, does not, in principle, come to an end. Or so the philosophic skeptic will claim!

To clarify the distinction, let us consider a restricted form of ordinary incredulity and a restricted form of philosophical skepticism. Suppose that I claim that although you might think you know that John is unconcerned about the doctor's report he just received, John is really just putting up a brave front so as not to alarm his family. My basis for doubt could be removed, or at least somewhat mitigated, if he acts in the same fashion when he thinks no one is observing him. Both my basis for doubting that he was unconcerned and the manner of removing the doubt presupposes that we can know the contents of other minds.

Now, contrast that case with one in which a basis is proposed for philosophical skepticism about our purported knowledge that John is unconcerned about the report. Here the grounds are that in order to know the contents of someone else's mind, my reasoning must employ an argument from analogy that is too weak to establish its conclusion. The form of the argument the skeptic is referring to could be put as follows:

1. Typically, when I behave in some way, *b*, I am in some mental state, *m*.	[by observation and introspection]
2. When someone else, say *S*, is behaving in way *b*, S is in mental state *m*.	[by analogy from 1]
3. *S* is behaving in way *b*.	[by observation]
4. *S* is in mental state *m*.	[from 2, 3]

The skeptic could point out that the analogy in 2 is illegitimately employed because arguments from analogy must be based upon more than one example. Since there can never be more than one case in which I note the correlation between mental states and behavior (namely, my own), I could never legitimately employ such an argument.

Now whether this restricted form of philosophical skepticism is plausible is not the issue here. The point is to make clear the distinction between ordinary incredulity and philosophical skepticism. Ordinary incredulity arises within the context of other propositions of a similar sort taken to be known, and it can be removed by discovering some further proposition of the relevant type. On the other hand, philosophical skepticism about a proposition of a certain type derives from considerations that are such that they cannot be removed by appealing to additional propositions of that type -- or so the skeptic claims.

This case illustrates one other fundamental feature of the philosophical arguments for skepticism, namely, that the debate between the skeptics and their opponents takes place within the evidentialist account of knowledge which holds that knowledge is at least true, sufficiently justified belief. The debate is over whether the grounds are such that they can make a belief sufficiently justified so that a responsible epistemic agent is entitled to assent to the proposition.[3] The basic issue at stake is whether the justification condition can be fulfilled. A corollary of this is that strictly reliabilist or externalist responses to philosophic skepticism constitute a change of subject. A belief could be reliably produced, i.e., its causal pedigree could be such that anything having that causal etiology is sufficiently likely to be true, but the reasons available for it could fail to satisfy the standards agreed upon by both the skeptics and their opponents.

2. Two Basic Forms of Philosophical Skepticism

Consider some proposition, *p*. There are just three possible propositional attitudes one can have with regard to *p*'s truth when considering whether *p* is true. One can either assent to *p*, or assent to ~*p* or withhold assenting both to *p* and to ~*p*. Of course, there are other attitudes one could have toward *p*. One could just be uninterested that *p* or be excited or depressed that *p*. But those attitudes are either ones we have when we are *not* considering whether *p* is true or they are attitudes that result from our believing, denying or withholding *p*. For example, I might be happy or sorry that *p* is true when I come to believe that it is.

I just spoke of "assent" and I mean to be using it to depict the pro-attitude, whatever it is, toward a proposition that is required for knowing that proposition. Philosophers have differed about what that attitude is. Some take it to be something akin to being certain that *p* or guaranteeing that *p* (Malcolm 1963, 58-72). Others have taken it not to be a form of belief at all because, for example, one can know that *p* without believing it as in cases in which I might in fact remember that Queen Victoria died in 1901 but not believe that I remember it and hence might be said not to believe it (Radford 1966). For the purposes of this essay we need not attempt to pin down precisely the nature of the pro-attitude toward *p* that is necessary for knowing that *p*. It is sufficient for our purposes to stipulate that assent is the pro-attitude toward *p* required to know that *p*.

Let us use "EI-type" propositions to refer to **epistemically interesting** types of propositions. Such types of propositions contain tokens some of which are generally thought to be known given what we ordinarily take knowledge to be. Thus, it would *not* be epistemically interesting if we did not know exactly what the rainfall will be on March 3 ten years from now. That kind of thing (a fine grained distant future state) is not generally thought to be known given what we ordinarily take knowledge to be. But it would be epistemically interesting if we cannot know anything about the future, or anything about the contents of someone else's mind, or anything about the past, or anything at all about the "external world." We think we know many propositions about those types of things.

Now, consider the (meta) proposition concerning the scope of our knowledge, namely: *We can have knowledge of EI-type propositions.* Given that there are just three stances we can have toward any

proposition when considering whether to assent to it, we can:

i. Assent that we can have knowledge of EI-type propositions.
ii. Assent that we cannot have knowledge of EI-type propositions.
iii. Withhold assent to both that we can and that we cannot have knowledge of EI-type propositions.

Let us call someone with the attitude depicted in (i) an "Epistemist."[4] Such a person assents to the claim that we can have knowledge of EI-type propositions.

The attitude portrayed in (ii) has gone under many names. I will follow the terminology suggested by Sextus Empiricus. He used the term "Academics" to refer to the leaders of the Academy (founded by Plato) during the 3rd to 1st century B.C. According to Sextus, they assented to the claim that we cannot have knowledge of what I have called EI-type propositions -- although it is far from clear that this was an accurate description of their views. (See the entry on ancient skepticism.) Perhaps the prime example was Carneades (214-129 B.C.). Other philosophers will refer to this view as "Cartesian skepticism" because of the arguments investigated by Descartes and his critics in the mid-17th century. And still others will refer to it as "switched world skepticism" or "possible world skepticism" because the arguments for it typically involve imagining oneself to be in some possible world that is both vastly different from the actual world and at the same time absolutely indistinguishable (at least by us) from the actual world. What underlies this form of skepticism is assent to the proposition that we cannot know EI-type propositions because our evidence is inadequate.

Those assenting neither to the proposition that knowledge of EI-type propositions is possible nor to the proposition that such knowledge is not possible can be called "Pyrrhonian Skeptics" after Pyrrho who lived between ca 365 - ca 275 B.C. The primary source of Pyrrhonian Skepticism is the writing of Sextus Empiricus who lived at the end of the second century AD. The Pyrrhonians withheld assent to every non-evident proposition. That is, they withheld assent to all propositions about which genuine dispute was possible, and they took that class of propositions to include the (meta) proposition that we can have knowledge of EI-type propositions. Indeed, they sometimes classified the Epistemists and the Academic Skeptics together as dogmatists because the Epistemists assented to the proposition that we can have knowledge, while the Academic Skeptics denied that we can have knowledge.[5]

Another difference between Academic and Pyrrhonian Skepticism is closely related to the charge by the latter that the former is really a disguised type of dogmatism. The Academic Skeptic thinks that her view can be shown to be the correct one by an argument (or by arguments). The Pyrrhonian would point out that the Academic Skeptic maintains confidence in the ability of reason to settle matters -- at least with regard to the extent of our knowledge of propositions in the EI-class. One way of understanding the so-called problem of the "Cartesian circle" illustrates the Pyrrhonian point: Descartes is relying throughout the *Meditations* on his power of reasoning to remove the skeptical doubts that he raises, but to do so means that he has exempted the faculty of reasoning from the doubts that he raised in the "First Meditation" about the epistemic reliability of our faculties. A Cartesian reply could be as simple as paraphrasing Luther: Here I stand, as a philosopher with confidence in reason, and as such I can do no

other.[6] Regardless of the adequacy of that kind of response, the point here is that the Pyrrhonians did not think that they had a convincing argument whose conclusion was that withholding assent to non-evident propositions was the appropriate epistemic attitude to have.

I think it is fair to say that Academic Skepticism is usually what is meant when most contemporary philosophers write about skepticism. Thus, it is that form of skepticism to which we will now turn and it is that form that will be the primary focus of this essay, although we will discuss some aspects of Pyrrhonism later.

3. Academic Skepticism

A way to motivate Academic Skepticism and to clearly distinguish it from ordinary incredulity is to trace the way in which Descartes expanded the realm of what was doubtful (and hence not worthy of assent) in the "First Meditation."[7] Descartes begins by noting that the senses have deceived him on some occasions and, in the voice of his skeptical interlocutor, he conjectures that it is never prudent to trust what occasionally misleads. So, we don't have "certain" knowledge of the external world based upon the testimony of our senses. However, in the voice of the non-skeptical interlocutor, he replies that even though the senses have misled him, he can neutralize that purported basis for doubt by pointing out that we are able to determine when our senses are not trustworthy. Thus, this is a case of ordinary incredulity because he appeals to some knowledge of the world gained through our senses to neutralize this basis for doubt. For example, in looking at a straight stick in water, even though it appears bent, we know not to accept the testimony of our senses at face value. We can neutralize the potentially knowledge-robbing proposition *that my senses have deceived me on some occasions* by adding to it another proposition to which we assent. Some propositions in the EI-type (propositions about the external world) can be used to rebuff the grounds for ordinary incredulity. Thus, no basis for (philosophical) Academic Skepticism has been located.

Descartes next considers dreaming. What if he were dreaming at that very moment? Would he still have some knowledge of the external world? Yes; because in dreams and in waking life there are some common general features. So, if he were dreaming, he would not know in particular what is going on about him at that moment, but that does not imply that he fails to have any knowledge at that moment. For example, he still could know that there are hands. More importantly, even more simple things about nature "in general" are not thereby made doubtful. We have not found any reason for doubting that there are material objects or that they have a spatial location, or are in motion or at rest, or can exist for a long or short period of time. Again, no basis for Academic Skepticism has been established.

But then Descartes thinks of a grounds for doubt for which he says he "certainly has no reply." He puts it this way:

> … In whatever way [it is supposed] that I have arrived at the state of being that I have reached -- whether [it is attributed] to fate or to accident, or [made] out that it is by a continual succession of antecedents, or by some other method -- since to err and deceive

> oneself is a defect, it is clear that the greater will be the probability of my being so imperfect as to deceive myself ever, as is the Author of my being to whom [is assigned] my origin the less powerful. (*Meditations*, 147)

In other words, at this point in the *Meditations*, since he lacks an argument for the claim that whatever is causally responsible for his "state of being" is capable of making his being such that to err would not be natural for it, assenting to propositions arrived at by his "state of being" is not legitimate. Thus, Descartes believes that he has located a basis for doubting all of his supposed former knowledge of the external world that cannot be repulsed by locating another such proposition to which he is entitled. He has found a proposition that, if true, would (by itself) defeat the justification he has for his assenting to propositions about the external world and which is such that (1) he does not (at least at this point in the *Meditations*) have a way to reject it and such that (2) he has no way to neutralize its effect. Thus, a basis for philosophical skepticism has been found because an entire class of EI-type propositions -- propositions that his "nature" has led him to assent to -- is now thrown into doubt because he cannot use one member of the class to reject or neutralize the basis for doubting another member of the class.

Descartes apparently thinks that something is worthy of assent only if it is immune to genuine doubt. Something, *d*, is a grounds for genuine doubt of *p* for *S* iff:

1. *d* added to *S*'s beliefs makes assent to *p* no longer adequately justified;
2. *S* is not justified in rejecting *d*;[8]
3. *S* has no way to neutralize *d*.[9]

The final step is to say that some proposition is not worthy of assent if there are genuine grounds for doubting it. Indeed, Descartes grants that even after *d* is located, *p* might still be more reasonable to believe than to deny (*Meditations*, 148). The point is that the pro-attitude should not rise to the level required for knowledge because there is a genuine ground for doubt. Thus, a crucial feature of the Cartesian-style argument for Academic Skepticism is that it employs a very stringent requirement on the type of evidence required for knowledge. It must make a proposition immune to genuine doubt.

To make that clear, let us state the epistemic principle, which we can call the "Eliminate All Doubt Principle," that apparently informs the Cartesian-style argument:

> **Eliminate All Doubt Principle** [**EAD**]:
> For all propositions *x* and *d*, if *d* provides a basis for genuine doubt that *x*, then if assenting to *x* is adequately justified for *S*, then *S* is adequately justified in eliminating *d*.

In more contemporary terminology, the ground for doubt proposed by Descartes can be put like this:

> **U**: My epistemic equipment is untrustworthy.

The *Cartesian-style argument for Academic Skepticism* can now be put like this:

1. If I know that *p*, then there are no genuine grounds for doubting that *p*.
2. **U** is a genuine ground for doubting that *p*.
3. Therefore, I do not know that *p*.

The Cartesian-style argument does not readily lend itself to the objection that by employing it the skeptic is contradicting herself on the grounds that the argument purportedly shows that she fails to know because her epistemic equipment is untrustworthy while at the same time she is employing the very equipment that, were the argument sound, would be unreliable. The reason is that she is neither asserting that her equipment is untrustworthy nor claiming that there is an argument which shows that her equipment is untrustworthy. She is merely claiming that **U** is a genuine ground for doubt. Thus she neither is holding contradictory beliefs nor is her practice somehow incompatible with her beliefs.

The Cartesian-style argument for Academic Skepticism should be contrasted with what many contemporary philosophers take to be the canonical argument for Academic Skepticism which employs the Closure Principle (CP).[10] Letting "*h*" stand for an EI-type proposition, for example, G. E. Moore's famous "here's a hand" and letting "*sk*" stand for "I am in a switched-world in which there are no hands, but it appears just as though there were hands," we can state the canonical *CP-style argument for Academic Skepticism* as follows:

> CP1. If I am justified in believing that *h*, then I am justified in believing that ~*sk*.
>
> CP2. I am not justified in believing that ~*sk*.
>
> Therefore, I am not justified in believing that *h*.

This argument appeals to a form of the Closure Principle in Premise 1. Letting "*Jsx*" stand for "*S* is justified in having some pro-attitude, *J*, regarding *x*," that principle can be stated as:

> **Closure Principle** [**CP**]:
> For all propositions *x* and *y*, if *x* entails *y*, and *Jsx*, then *Jsy*.

(In the CP-style argument: $x = h$ and $y = {\sim}sk$.)

A crucial feature of CP is that it does not depend upon employing a stringent notion of justification. Suppose that (positive) justification comes in degrees, where the lowest degree is something like mere plausibility and the highest degree is absolute certainty. CP could be recast as follows:

> **CP*:** For all propositions, *x* and *y*, if *x* entails *y*, and *Jsx* to degree *u*, then *Jsy* to degree *v* (where $u \leq v$).

Thus, when the Academic Skeptic employs CP (or CP*), she need not be employing a very stringent

notion of justification. That is the primary difference between the CP-style and the Cartesian-style argument for Academic Skepticism.

Another difference is that the Cartesian-style argument concerns knowledge, whereas the CP-style argument concerns justification (to whatever degree). Nevertheless, that difference is insignificant because the debate about the merits of skepticism takes place within the evidentialist account of knowledge. Knowledge is taken to entail adequately justified assent and, hence, "knowledge" could be replaced by "adequately justified assent" in the Cartesian style argument.

Let us return to the central difference between Cartesian and CP-style arguments, namely the former employs EAD while the later employs CP (or CP*). EAD requires that we eliminate any genuine grounds for doubt and those include more than mere contraries. In addition, recall that according to the Cartesian to be adequately justified in eliminating d as a ground for doubt for *x*, either *S* is adequately justified in denying *d* (assenting to ~*d*) or *S* is adequately justified in assenting to some neutralizing proposition, *n*, such that adding (*n* & *d*) to *S*'s beliefs fails to make it the case that *x* is no longer adequately justified.[11] Thus, since every contrary of some proposition is a potential genuine ground for doubt, EAD entails CP but CP does not entail EAD.[12] To see that, consider any contrary, say *c*, of a proposition, say *h*. The proposition, *c*, would be a potential genuine ground for doubting *h* since if *c* were added to *S*'s beliefs, h would no longer be adequately justified because *S*'s beliefs would then contain a proposition, *c*, that entailed the denial of *h*. Furthermore, the only way *S* could eliminate *c* as a ground for doubt would be by denying it, since nothing could neutralize it. Thus, EAD has the consequence that *if* S *is justified in assenting to h, then S is justified in denying every contrary of h*. But that is just an instance of CP, since (by hypothesis) *h* entails ~*c*. That CP does not entail EAD should be clear since there are grounds for doubting *h* that are not contraries of *h*. For example, the proposition, **U**, considered above is a grounds for doubting *h*, but *h* and **U** could both be true.

Thus, there are two basic forms of Academic Skepticism: The Cartesian-style argument that employs the strong EAD principle and the CP-style that employs the weaker CP. Since the CP-style skeptic employs the weaker epistemic principle, it will be best to begin by focusing on it because any criticisms of it are likely to redound to the stronger form.

4. The Argument for Academic Skepticism Employing the Closure Principle

There *appear* to be only three ways that one can respond to the CP-style skeptical argument: deny at least one premise, deny that the argument is valid, or reluctantly accept the conclusion -- if neither of the first two alternatives succeeds. I say "appear" because I will mention later a fourth alternative that is available to the Pyrrhonian Skeptic. The second alternative -- denying the validity of the argument -- has not been taken because it would lead to embracing an extremely severe form of skepticism. If one were to deny that *modus tollens* is a valid form of inference, one would also have to deny the validity of (i) disjunctive syllogism or (ii) *modus ponens* and contraposition, since it is easy to transform *modus tollens* arguments

into ones employing the other forms of inference. Hence, if this alternative were chosen, reasoning would apparently come to a complete standstill. That, presumably, is why no one has ever seriously considered this alternative.

So, if we are not to reluctantly embrace the conclusion, it *appears* as though we must reject either the first premise -- an instantiation of closure -- or the second premise.

Consideration of CP1

Let us begin an examination of CP1 and the general closure principle of which it is an instantiation. The basic issue is this: Does closure hold for justified belief?

Closure certainly does hold for some properties, for example, truth. If *p* is true and it strictly implies *q*, then *q* is true. It just as clearly does not hold for other properties. If *p* is a belief of mine, and *p* strictly implies *q*, it does not follow that *q* is a belief of mine. I might fail to see the implication or I might be "wired" incorrectly (from birth or as the result of an injury) or I simply might be epistemically perverse. I might, for example, believe all of the axioms of Euclidean plane geometry, but fail to believe (or perhaps even refuse to believe) that the exterior angle of a triangle is equivalent to the sum of the two opposite interior angles.

What about justified belief? It is easy to see that, as stated above, CP (or CP*) is clearly false. Every necessary truth is entailed by every proposition, and we can be justified in believing a false proposition. But one surely does not want to claim that *S* is justified in believing *every* necessary truth whenever *S* has some justified belief in a false proposition. In addition, some entailments might be beyond *S*'s capacity to grasp. Finally, there might even be some contingent propositions that are beyond S's capacity to grasp which are entailed by some propositions that *S* does, indeed, grasp. And it might be thought that *S* is not entitled to believe anything that *S* cannot grasp.

But it appears that CP can easily be repaired. We can stipulate (i) that the domain of the propositions in the generalization of CP includes only contingent propositions that are within *S*'s capacity to grasp and (ii) that the entailment is "obvious" to *S*. The skeptic can agree to those restrictions because the skeptical scenarios are posited in such a way as to render it obvious that our ordinary beliefs are false in those scenarios, and it is taken to be a contingent claim that *S* is in the actual circumstances as described in the antecedent. There is one other required clarification of the restricted version of CP. "Justified belief" is ambiguous. It could be used to refer to a species of actually held beliefs -- namely, those actually held beliefs of *S* that are justified. Or it could refer to propositions that *S* is entitled to hold -- regardless of whether *S* does indeed hold them. If CP is to be acceptable, "justified belief" must be used so as to mean the latter for a reason already cited, i.e., belief does not transmit through entailment.

We are now in a position to ask: Does the restricted form of closure hold regarding what we are *entitled* to believe -- even if we don't, in fact, believe it?

There appears to be a perfectly general argument for the restricted version. Let *p* entail *q*, and *let us suppose* that *S* is entitled to believe that *p* iff *S* has (non-overridden) grounds that make *p* sufficiently likely to be true:[13]

1. If *S* is entitled to believe that *p*, then *S* has (non-overridden) grounds that make *p* sufficiently likely to be true. [by the supposition]
2. If *S* has (non-overridden) grounds that make *p* sufficiently likely to be true, then *S* has (non-overridden) grounds making *q* sufficiently likely to be true. [because *p* entails *q*]
3. If *S* is entitled to believe that *p*, then *S* has (non-overridden) grounds making *q* sufficiently likely to be true. [from 1,2]
4. If *S* has (non-overridden) grounds making *q* sufficiently likely to be true, then *S* is entitled to believe that *q*. [by the supposition]
5. Therefore, if *S* is entitled to believe that *p*, *S* is entitled to believe that *q*. [from 2,3]

The supposition mentioned above seems plausible given that the debate over the merits of Academic Skepticism employs an evidentialist account of justification. That is, the debate between the Academic Skeptic and the Epistemist is over whether *S* has adequate grounds for EI-type propositions such that those grounds make *p* sufficiently likely to be true. Premise 2 contains the key claim. In spite of the fact that the probabilities (whether subjective or objective) transmit through entailment, it has been challenged. Fred Dretske and others have produced cases in which they believe CP fails and fails precisely because Premise 2 in the general argument for CP is false.[14] Dretske writes:

> … something's being a zebra implies that it is not a mule … cleverly disguised by the zoo authorities to look like a zebra. Do you know that these animals are not mules cleverly disguised? If you are tempted to say "Yes" to this question, think a moment about what reasons you have, what evidence you can produce in favor of this claim. The evidence you *had* for thinking them zebras has been effectively neutralized, since it does not count toward their *not* being mules cleverly disguised to look like zebras. (Dretske 1970, 1015-1016)

Dretske is speaking of "knowledge" rather than beliefs to which one is entitled, but that seems irrelevant since the issue concerns the supposed lack of a sufficient source of evidence or reasons for the claim that the animal is not a cleverly disguised mule. In other words, Dretske grants that *S* has (non-overridden) grounds that make it sufficiently likely that the animals are zebras, but he holds that S does not have (non-overridden) grounds making it sufficiently likely that the animals are not cleverly disguised mules because S's evidence for the former has been "effectively neutralized."

The crucial thing to note about this proposed counterexample is that it works only if the Closure Principle entails that the *very same* source of evidence that justifies *S* in believing that the animals are zebras *must* justify *S* in believing that they are not cleverly disguised mules. Since the "evidence" for the former has been "effectively neutralized," it is not available for the latter. Now, in response one could claim that once the question of whether the animals are disguised mules has been raised, the evidence is "effectively neutralized" for *both* the former and the latter, and *S* is no longer justified in believing that the animals are

zebras. Thus, it could be held that this example could actually be used to *support* CP.

Nevertheless, let us grant that *S*'s evidence for the claim that the animals are zebras cannot be used to show that they are not cleverly disguised mules. It could be argued that this would not force giving up Premise 2 in the general argument for CP.

Such an argument could begin by recalling that Premise 2 claimed merely that whenever *S* had (non-overridden) grounds that make *p* sufficiently likely to be true, then *S* has (non-overridden) grounds for making *q* sufficiently likely to be true. It did *not* require that it was the very same grounds in both cases. Dretske's purported counterexample seems to require that CP implies that the adequate source of evidence is the same for both propositions. Thus, letting "*xRy*" mean that *x* provides an adequate evidence for *y*, the counter example depends upon assuming that if closure holds between *p* and *q*, then the evidence "path" must look like this:

Pattern 1

```
              ...Rp
             /
            /
...Re
            \
             \
              ...Rq
```

Evidence paths specify what propositions serve as good enough reasons, *ceteris paribus*, for believing other propositions. Dretske is supposing that the very same evidence, *e*, that I have for *p* must be adequate for *q* whenever *p* entails *q*.

No doubt this constraint sometimes correctly portrays the relevant evidential relationships when some proposition, *p*, entails some other proposition, *q*. For example, suppose I have adequate evidence for the claim that Anne has two brothers, then it would seem that the very same evidence would be adequate for believing that Anne has at least one brother. But the defender of CP, and more particularly the Academic Skeptic, could point out that closure need not require that type of evidence path in all cases in which one proposition entails another.

Two are two other possibilities for instantiating closure that are captured by Premise 2 that can be depicted as follows:

Pattern 2

… **ReRp**… **Rq**

Pattern 3

… **Re**(where *e* includes *q*)… **Rp**

In Pattern 2 cases there is some adequate evidence, *e*, for *p*; and *p*, itself, is the adequate evidence for *q*, since *p* strictly implies *q*. For example, if I have adequate evidence for believing that 2 is a prime number, I can use that proposition as an adequate reason for believing that there is at least one even prime. Indeed, consider any belief arrived at as a result of deductive inference. In such a case, we legitimately infer the entailed proposition from the conjunction of the premises that entails it. The plausibility of the famous Gettier cases depends upon Pattern 2 type cases in which closure holds. Gettier says:

> … for any proposition *p*, if *S* is justified in believing *p*, and *p* entails *q*, and *S* deduces *q* from *p* and accepts *q* as a result of this deduction, then *S* is justified in believing *q*. (Gettier 1963, 122)[15]

In Pattern 3 cases the order of the evidence is reversed because *q* serves as part of the evidence for *p*. For example, I am justified in believing that water is present if I am justified in believing that there is present a clear, odorless, watery-tasting and watery-looking fluid at standard temperature and pressure. This pattern is typical of abductive inferences. In addition, there are cases in which some contraries of *h* need to be eliminated prior to *h*'s being justified. For example, in the zebra-in-the-zoo case, if I had some reason to think that the animals were cleverly disguised mules, then it could be argued that such a contrary would need to be eliminated *before* I would be justified in believing that the animals were zebras.

The crucial point for the discussion here is that granting that there is no Pattern 1 type evidence path available to *S* in the zebra-in-the-zoo case does not require relinquishing premise 2 in the general argument for CP. The reason is simply that CP does not entail that there is Pattern 1 type evidence available in every case in which *p* entails *q*. Indeed, it could be suggested that the animals looking like zebras in a pen marked "zebras" is, *ceteris paribus*, adequate evidence to justify the claim that they are zebras; and once *S* is entitled to believe that the animals are zebras, *S* can, using the principle stated by Gettier, justifiably deduce that they are not cleverly disguised mules. That is, *S* can employ an evidence path like that depicted in Pattern 2. (See Klein 1981, 1995, and 2000a.) Further, if *S* had *some* reason to think that the animals were cleverly disguised mules, then *S* might have to eliminate that possibility before she could justifiably believe that they are zebras. In other words, *S* might have to employ an evidence path like the one depicted in Pattern 3. The point is that the Dretske-like counterexamples appear to depend upon the false claim that if Premise 2 in the general argument for CP is true, then the evidential relationship between the entailing and the entailed proposition is always correctly depicted by Pattern 1.

In addition to the purported counterexamples to closure that we have just examined, there are some

general theories of knowledge in which closure fails. Robert Nozick's account of knowledge is the best such example. Roughly his account is this (Nozick 1981, 172-187):

> *S* knows that *p iff*:
>
> 1. *S* believes *p*;
> 2. *p* is true;
> 3. if *p* were true, *S* would believe *p*;
> 4. if *p* were not true, *S* would not believe *p*.

This account is often referred to as a tracking account of knowledge because whenever *S* knows that *p*, *S*'s beliefs track *p*. Think of a guided missile *tracking* its target. If the target moves left, the missile moves left. If the target moves right, the missile moves to the right. According to the tracking account of knowledge our beliefs must track the truth as a guided missile tracks its target.

There is one important clarification of conditions 3 and 4 discussed by Nozick, namely, that the method by which *S* acquires the belief must be held constant from the actual world to the possible world. A doting grandmother might know that her grandchild is not a thief on the basis of very good evidence, but would still believe that he wasn't a thief, even if he were, because she loves him. So, we must require that the grandmother use the same method in both the actual and the near possible worlds, for otherwise condition 4 would exclude cases of knowledge. This is not the place to provide a full examination of Nozick's account of knowledge.[16] What is crucial for our discussion is that it is easy to see that closure will fail for knowledge in just the kind of case that the Academic Skeptic is putting forward because of condition 4. Suppose *S* knows that there is a chair before her. Would she know that she is not in a skeptical scenario in which it merely appears that there is a chair? If the fourth condition were true, she would not know that because if she were in such a scenario, she would be fooled into thinking that she wasn't. Thus, either condition 4 is too strong or CP fails.

There are some reasons for thinking that condition 4 is too strong. Consider a relatively simple case in which *S* seems to have knowledge but condition 4 does not obtain. *S* looks at a thermometer that is displaying the temperature as 72 degrees. The thermometer is working perfectly and *S* comes to believe that the temperature is 72 degrees by reading the thermometer and coming to believe what it says. But if the temperature were not 72, suppose that something would affect the thermometer in a way that made it read "72," so that by employing the same method (looking at the thermometer and coming to believe what it reads) *S* would still believe that it was 72. (One could imagine all kinds of circumstances that would have that causal result. A comical one: Imagine a lizard that is now sleeping on the thermometer that would stir were the temperature to rise, thus dislodging a small rock that hits the thermometer breaking the mercury column in a way that makes the thermometer still read 72.)

Or consider this case in the literature: You put a glass of ice-cold lemonade on a picnic table in your back yard. You go inside and get a telephone call from a friend and talk for half an hour. When you hang up you remember that you had left the ice-cold lemonade outside exposed to the hot sun and come to believe

that it isn't ice-cold anymore. It would seem that you could know that even if in some near world a friend of yours who just happened to be walking by noticed the glass and happening to have a cooler full of ice with him put the glass of lemonade in the cooler to keep it ice-cold for you. Thus, if the lemonade were still ice-cold, you would believe that it wasn't. (See Vogel 1987, 206.)

The moral of these cases seems to be that *S* can know that *p* even if there are some near possible worlds in which *p* is false but *S* still believes that *p* (employing the same method of belief formation). Indeed, it could plausibly be maintained that what is required for knowledge is that the method of belief formation work in this world -- exactly as it is -- even if the method would fail were there to be some slight variation in the actual world.

Further Clarification of Closure

In order to clarify CP further, it would be useful to contrast it with a stronger principle. I have already pointed out that in some cases some contraries of *h* need to be eliminated before *h* becomes justified. Suppose, however, that the skeptic requires that all contraries to *h* be eliminated before *h* is justified. That is *much* stronger than CP because CP is compatible with Pattern 1 and Pattern 2 type evidential relationships. In neither is every contrary to h eliminated prior to *h*. In Pattern 2, the contrary of h is eliminated after *h*; in Pattern 1, *h* is arrived at and its contrary is eliminated simultaneously. Keith Lehrer *might* be appealing to the stronger principle when he writes:

> … generally arguments about where the burden of proof lies are unproductive. It is more reasonable to suppose that such questions are best left to courts of law where they have suitable application. In *philosophy* [emphasis added] a different principle of agnoiology [the study of ignorance] is appropriate, to wit, that no hypothesis should be rejected as unjustified without argument against it. Consequently, if the sceptic puts forth a hypothesis inconsistent with the hypothesis of common sense, then there is no burden of proof on either side … . (Lehrer 1971, 53)

The passage is a bit ambiguous, but it will serve to illustrate my point, namely that there is a very strong principle -- call it the "**Eliminate All Contraries First** Principle" [EACF Principle] -- which requires that all evidence paths exhibit Pattern 3.

If EACF were accepted, there is a really easy route to Academic Skepticism. Consider any two contraries, c_1 and c_2. In order to be justified in believing c_1, *S* would first have to eliminate c_2. And in order to be justified in believing that c_2, *S* would first have to eliminate c_1. So, of course, *S* could never either be justified in believing c_1 or be justified in believing c_2. It could be plausibly argued that this is a too quick and too dirty argument for skepticism because in so far as skepticism remains an interesting philosophical position, the skeptic cannot impose such an outrageous departure from our ordinary epistemic practices.

There is a related point worth mentioning. If it were required that the evidence, *e*, for some hypothesis, *h*, must contain the denials of all the contraries of *h*, it is clear that *e* would have to entail *h*. To see that, note

that (~*h* & *p*) as well as (~*h* & ~*p*) are contraries of *h*, and that {~(~*h* & *p*), ~(~*h* & ~*p*)} entails *h*. Thus, if the skeptic were to adopt EACF, the evidence for *h* would have to entail *h*. (See Klein 1981, 100-104.) That requirement seems to be too strong for many, if not most, empirically contingent propositions.

Note that even EAD, although requiring that we be able to reject or neutralize every ground for doubt, does not require what EACF does. EAD does not require that we eliminate all of the grounds for doubt (including contraries) *before* we are justified in believing a hypothesis. Indeed, EAD allows for the possibility that we could use *h*, itself, or something that *h* justifies as the basis for rejecting or neutralizing some grounds for doubt.

Consideration of CP2

Now, with those clarifications of CP (and EAD) in mind, we can turn to CP2. It claims that we are not justified in denying the skeptical hypothesis -- in other words that we are not justified in believing that we are not being deceived. What arguments can be given for CP2? It is tempting to suggest something like this: The skeptical scenarios are developed in such a way that it is supposed that we *could not* tell that we were being deceived. For example, we are asked to consider that there is an Evil Genius "so powerful" that it could (1) make me believe that there were hands when there were none and (2) make it such that I *could* not detect the illusion. But the skeptic must be very careful here. She cannot require that in order for *S* to know (or be justified in assenting to) something, say *x*, that if *x* were false, she would not still assent to *x*. We have just seen (while examining Nozick's account of knowledge) that this requirement is too strong. So the mere fact that there could be skeptical scenarios in which *S* still believes that she is not in such a scenario cannot provide the skeptic with a basis for thinking that she fails to know that she is not (actually) in a skeptical scenario. But *even more importantly*, were that a requirement of knowledge (or justification), then we have seen that closure would fail and, consequently, the basis for the first premise in the CP-style argument for Academic Skepticism would be forfeited.[17]

In addition, we have also seen that if CP is true, and it did seem to be true, then there is one evidence pattern between entailing and entailed propositions that might prove useful to the Epistemist at this point in the discussion. If *S* could be justified in believing some proposition that entailed the denial of the skeptical hypothesis, then *S* could be justified in denying that hypothesis by employing evidence Pattern 2. Indeed, as G. E Moore suggested (1962, 242), what is to prevent the Epistemist from claiming that *S* is justified in denying that she is in a skeptical scenario because *S* is justified in believing that she has hands and CP is true? A plausible answer to Moore seems to be something like this: The issue that is under dispute is whether *S* is justified in assenting to (or knows that) she has hands. Thus, the Epistemist cannot reject CP2 by assuming the denial of the conclusion of the skeptical argument. All well and good. But the same sauce cooks the gander, and the skeptic cannot claim as the reason for CP2 that *since S* is not justified in believing that she has hands, she cannot avail herself of that as her reason for being justified in believing that she is not in a skeptical scenario.

So, what reason can the skeptic give for CP2? I do not know of one that has been offered that is consistent with the defense of CP and that does not beg the question. That is not to say that CP2 is false. Far from it.

Perhaps it is true. The issue here is whether we are justified in accepting or rejecting it. It seems that in order to accept it and CP, the skeptic would have to assert that *S* is not justified in believing that she has hands because evidence Pattern 2 depicts one way in which *S* could be justified in denying the skeptical scenario.[18]

I had mentioned earlier that there seemed to be only three responses available when confronting the CP-style argument for Academic Skepticism -- accept the conclusion, reject one or both of the premises, or deny the validity of the argument. The fourth alternative is simply to point out that given the required defense of CP1, there might be no good argument for CP2.

Of course, there is the possibility that there is also no good argument to the conclusion that we do have knowledge of EI-type propositions. Some might think that the Academic Skeptic wins in such a stand-off. But recall that what distinguishes the Academic Skeptic from the Pyrrhonian Skeptic is that only the Academic Skeptic assents to the claim that we cannot have knowledge. The Pyrrhonian Skeptic withholds judgement regarding whether we can have knowledge. And in a stand-off, the Pyrrhonian would seem to have appropriate attitude.

This concludes the discussion of CP-style skepticism. I would now like to briefly consider the second form of Academic Skepticism, namely the Cartesian-style that employs the Eliminate All Doubt Principle. Then, before we conclude our discussion of Academic Skepticism, I would like to consider one quite popular response to it -- contextualism.

5. The Cartesian-style Argument for Academic Skepticism Employing the Eliminate All Doubt Principle

This section can be brief because we can apply the lessons learned in the discussion of CP-style arguments to an evaluation of the Cartesian-style arguments that employ EAD. First, it should be clear that the general argument for the Closure Principle, considered earlier, cannot be used as a model for a general argument for EAD. That argument depended crucially on the fact that *h* entailed ~*sk*. (That is what provided the basis for premise 2 in the general argument for CP.) As we saw, the negation of a genuine ground for doubt need not be entailed by *h*. So, the skeptic has a much harder task in motivating EAD.

Nevertheless, let us grant that some argument could be provided that makes EAD plausible. The same dialectical issues that we have considered in discussing potential counterexamples to CP will recur regarding EAD. Reconsider the zebra-in-the-zoo case. But this time instead of the (contrary) proposition "the animals (I am seeing) are cleverly disguised mules," consider a potential ground for doubt, i.e., "there are cleverly disguised mules within my perceptual field," which according to EAD would have to be rejected or neutralized. Now, if the evidence I had for believing that the animals are zebras wasn't

adequate to deny the former, it is certainly not adequate for the denying the later. So the EAD skeptic will have to appeal to the analogs of Pattern 2 and Pattern 3 type cases in order to save the principle from a Dretske-like counterexample. Thus, the skeptic employing EAD would be put in the same dialectical situation as the CP-style skeptic because she must provide a basis for the second premise in her argument for Academic Skepticism that (1) is compatible with her required defense of EAD against Dretske-like objections and (2) does not beg the question or appeal to a requirement that all grounds for doubt must be eliminated prior to a proposition being justified.

To sum up: The Cartesian-style skeptic employing EAD is in a worse dialectical position than the skeptic employing CP. Whatever problems are associated with CP skepticism transfer to EAD skepticism and, in addition, there appears to be no plausible general argument for EAD while there was one for CP.

6. Contextualism[19]

Examining the contextualist diagnosis of Academic Skepticism and its suggested solution will allow us to explore a question that remains concerning CP and EAD. It could be held that such skeptics need not employ CP or EAD in general, but rather more restricted versions, namely merely their instantiations as they appear in their respective arguments. The skeptic could maintain that there is something quite special about the skeptical hypothesis such that even though closure might not hold in general between any entailing proposition and every proposition it entails, it does hold between such propositions as "here's a hand" and "it does not merely appear that here is a hand." Even more strongly, the skeptic could maintain that

only the Pattern 3 type evidence path correctly depicts the evidential relationships between those propositions. Hence, in order to be justified in believing the former I must *first* eliminate the latter. The requirement that we eliminate all contraries to some proposition, h, before we are entitled to believe that h is too stringent for ordinary contexts, for the reasons cited, but perhaps when engaged in doing philosophy we have to show that the skeptical hypothesis is false before the propositions of common sense are justified. That is essentially what the contextualists claim. They hold that in some contexts -- philosophical ones -- more stringent standards of evidence obtain than obtain in ordinary contexts.

There are two questions we should consider: Is contextualism about knowledge (or justified belief) the correct view to hold? If so, will it shed light on Academic Skepticism?

In answering the first question, it could be argued that contextualism with regard to the attribution of virtually any property is true. (Perhaps it doesn't apply to highly technical ones that only occur in one type of context.) For example, suppose that Mr. Lax says that Sam is happy. We discover that Lax is using "happy" to mean that a person is happy just in case he/she has had more happy moments than unhappy moments during a lifetime. Mr. Stringent demurs. For him, a person is happy only if he/she hardly ever experiences an unhappy moment.

Who is right about whether Sam is happy? Contextualists would say that they both could be because they are not using "happy" with the same criteria in mind. But it is crucial to note that given that each person recognizes that the other is applying different standards, Mr. Lax and Mr. Stringent can agree that, *given what Lax means*, Sam is happy and that, *given what Stringent means*, Sam is not happy.

Now, of course, we cannot employ *any* standards we please and still be speaking a common language. For

example, Mr. Lax cannot legitimately lower the standards so as to make it the case that Sam is happy simply because he *once* was happy for a very short period and, similarly, Mr. Stringent cannot require that Sam is happy only if it is *logically impossible* that Sam experience an unhappy moment. There is a limited range, albeit rather wide, of appropriate standards for the application of a term.

The predicates "having knowledge," "having adequate evidence," "being justified," and the like, do appear to be similar to most other predicates in this respect: Within a wide but non-arbitrary range of standards, speakers can legitimately demand that *S* have more or less of the relevant evidence for *p* before they will agree that "*S* knows that *p*" or "*S* has adequate evidence for *p*." So, the answer to the first question about the truth of contextualism seems to be: Contextualism about knowledge attributions is correct. It is just one instance of the general truth that standards for the application of a term vary within a wide but non-arbitrary range as determined by various features of the context.

Let us turn to the second and *much* more philosophically interesting question: Does the truth of this version of contextualism shed much, if any, light on Academic Skepticism? If it did, then the correct way to diagnose the dispute between the Academic Skeptic and the Epistemist would be to note that the Epistemist is using a lax standard and the skeptic a more stringent one. Having one's ordinary cake is compatible with eating one's skeptical cake because in the ordinary context we do have knowledge, but as standards rise to those employed by the skeptics, we do not have knowledge.

In response, it might be objected that this is not the proper diagnosis of the disagreement between the Academic Skeptic and the Epistemist. What the Academic Skeptic seems to be claiming is that we do not know what we ordinarily claim to know. We don't know EI-type propositions. That is, the Academic Skeptic claims that our ordinary knowledge claims are false. If she is merely claiming that on *her* standards we don't know, the skeptic's claims -- like those of Mr. Stringent -- can be granted and then promptly ignored because nothing that we formerly believed that we knew turns out to be not known. The scope of our knowledge or justified beliefs in the ordinary context is left intact.

Thus, the parallel with the case of Sam's putative happiness seems to break down. In that case, Mr. Stringent would grant that Mr. Lax is correct *given what Lax meant by "happy."* But the Academic Skeptic will not grant that the Epistemist is correct when he asserts that he has knowledge. The skeptic reasons that the Epistemist doesn't know that *h*, *even given what the Epistemist means by "know,"* because the Epistemist's justification for *h* isn't good enough. Indeed, the Academic Skeptic employing CP (or the stronger EAD) thinks that since there *cannot* be any evidence for ~*sk*, *h* could not be known.

The issue seems to boil down to this: In the ordinary context is it true -- as the Academic Skeptic claims -- that in order to know that there are hands, we must first eliminate the skeptical hypothesis?

The Epistemist could argue that this is not required. Suppose we are looking at Dretske's zebras and someone asks whether we have eliminated the possibility that those are cleverly disguised aliens from some planet thousands of light years from our solar system? Or that they are not super-robots newly invented by some very clever third graders in Mrs. Johnson's English class? Or that they are not members of the lost tribe of Israel disguised as zebras who have been hiding out from the Assyrians since the 8th

century BC. (They've had lots of time to perfect the disguise!)

Those are so far-fetched, the Epistemist could claim, that even if someone advancing those alternatives happens to believe them, there appears to be no reason why one should have to rise to the bait and eliminate those alternatives prior to being justified in believing that the animals are zebras. The Epistemist could continue by claiming that the skeptical hypothesis -- that we are not in the actual world but rather in one which seems identical to it -- is just as, or possibly even more, farfetched.

Thus the Epistemist could argue that try as she might, the Academic Skeptic cannot impose the burden of eliminating a farfetched hypothesis merely by raising it, even were she to believe it. On the other hand, the Epistemist could agree that in Dretske's zebra-in-the zoo case, if there really were some evidence, however slight, for the claim that the animals are painted mules, then Mr. Stringent might be able to legitimately require that *S* rule out that possibility prior to being justified in believing that the animals are zebras. But absent any evidence of that sort, the skeptic's requirements will fall on deaf ears. In parallel fashion, if there really were some evidence, however slight, that there is an evil genius making it merely appear that there are hands, then, perhaps the Academic Skeptic could legitimately require that *S* eliminate that possibility prior to being justified in believing that there are hands.

Put another way: The Epistemist can claim that the range of relevant alternatives is bounded by those propositions for which there is some, even minimal, evidence. The Epistemist will claim that it is a context-invariant feature of knowledge attributions that the relevant evidence does not include the denial of contraries for which there is no evidence whatsoever. The issue seems to be whether our ordinary knowledge claims are true -- not whether they *would* be true in some context with requirements more stringent that those ordinarily applied.

7. Pyrrhonism

As mentioned at the beginning of this essay, what distinguishes Pyrrhonian Skepticism from Academic Skepticism is that the former does not deny that we can have knowledge of what I have called EI-type propositions. They also would not assent to the Epistemist's claim that we can have such knowledge. Let us see how they arrived at that position.

To deny something is merely to assent to its negation. Since the Pyrrhonians took assent, i.e., the pro-attitude required for knowledge, to involve a kind of certainty that the matter had been finally and fully resolved, they did not assent to what they took to be non-evident propositions.

In distinguishing Pyrrhonism from the Academic Skeptics (in particular, Carneades and Cleitomachus), Sextus writes in *Outlines of Pyrrhonism*, [***PH***]:

> … although both the Academics and the [Pyrrhonian] Skeptics say that they believe some things, yet here too the difference between the two philosophies is quite plain. For the word "believe" has different meanings; it means not to resist but simply to follow without any

> strong impulse or inclination, as the boy is said to believe his tutor; but sometimes it means to assent to a thing of deliberate choice and with a kind of sympathy due to strong desire, as when the incontinent man believes him who approves of an extravagant mode of life. Since, therefore, Carneades and Cleitomachus declare that a strong inclination accompanies their credence … while we say that our belief is a matter of simply yielding without any consent, here too there must be difference between us and them. (***PH*** I:230)

So, the Pyrrhonians would not assent to non-evident propositions. Of course, a crucial issue concerns the scope of the non-evident. To try to resolve that is beyond the scope of this essay (but see Burnyeat & Frede 1997). For our discussion we can suppose that a sufficient condition for some proposition being non-evident obtains whenever there can legitimately be disagreement about it. And, taking the cue from our discussion of Academic Skepticism, I think we can also safely stipulate that there can be legitimate disagreement about some proposition if there is some evidence for it and some evidence against it. So, the question is whether the proposition *S can have knowledge of EI-type propositions* can be the subject of legitimate disagreement.

Putting the matter that way seems to make the answer obvious. There are arguments for Academic Skepticism which have some plausibility, and some plausible objections to those arguments that support the Epistemist's view. Plausible arguments for something constitute some evidence for it. So, we can safely conjecture that it is not evident that we can have knowledge of EI-type propositions and it is not evident that such propositions necessarily fall outside our cognizance. Thus, the primary question then becomes this: What prompted the Pyrrhonian to withhold assent to all non-evident propositions?

The answer is that they found over and over again that neither experience nor reason was able to settle disputes about the non-evident. But the Pyrrhonians did not eschew what they called "appearances" or reasoning. Quite the contrary, the Greek for "skeptic" is closely related to the verb "*sképtomai*" which means "to inquire." Thus, calling oneself a Pyrrhonian Skeptic did not imply a disregard for inquiry or reasoning. Indeed, the modes, to be discussed later, were not designed to inhibit reasoning. Rather, they were designed to assist the Pyrrhonian in continuing to inquire by shielding her from the disquieting state of dogmatism.

Pyrrhonian skepticism was thus a way of life without assent. As such, it has been ridiculed. The Pyrrhonian was likened to someone with Alzheimer's -- surviving only if someone else were around to save him from all sorts of perils: falling into pits, being attacked by a dog or run over by a chariot. That caricature seems to miss the point that the Pyrrhonian only withheld assent with regard to the non-evident propositions.[20] Assent to what was evident (i.e., what appears to be) or a weaker pro-attitude toward the non-evident were commonplace.

As mentioned above, the Pyrrhonians would practice what they called the "modes" in order to try to assure that they were not "perturbed" by assenting. Like piano exercises for the fingers that would result in semi-automatic responses to the printed notes on a sheet of music, the modes were mental exercises that would result in semi-automatic responses to claims being made by the dogmatists -- those who

assented to the non-evident.

The Pyrrhonians believed that (but would not have assented to the claim that) there were two potential sources of knowledge: perception and reasoning. When the results of perception were introduced to settle a non-evident matter -- say the *actual* color of an object (as opposed to how it appeared to someone), they would point out some or all of the following (Sextus Empiricus, ***PH*** I:40-128):

1. Members of *different* species of animals probably perceive colors quite differently because their eyes are constructed differently.
2. Members of the *same* species would have different perceptions of the color depending upon such things as the condition of their eyes, the nature of the medium of perception (varying light conditions for example), and the order in which objects were perceived.

Being reminded of the relativity of perception could incline a person to refrain from assenting to judgements of perception, when those judgements were about the "real" properties of the objects. As Sextus wrote:

> … When we question whether the underlying object is such as it appears, we grant the fact that it appears, and our doubt does not concern the appearance itself, but the account given of the appearance. (***PH*** I:19-20)

Now, perhaps a careful analysis of what is meant by "real" properties coupled with a Cartesian-like answer to some of the doubts raised earlier in the *Meditations* would suffice to respond to the Pyrrhonian concerning the relativity of our senses. For example, if we took the "real" color of objects to be that property (or state) of the object, whatever it is, that produces perceptions of a certain sort in humans under "normal" circumstances and if we could distinguish (as Descartes suggested) normal from abnormal circumstances, then we might have a basis for resisting the Pyrrhonian modes concerning perception. But be that as it may, whether we can have knowledge of EI-type propositions is not a matter that is potentially resolvable by direct appeal to our senses. It will only be resolved if either the Epistemist or the Academic Skeptic has a compelling argument. Thus, the issue here is whether reasoning can settle matters. The Pyrrhonians thought that there were modes which could induce withholding assent to the results of reasoning. It is to those modes that we should turn.

Perhaps the most influential passage in the corpus of the Pyrrhonian literature is a section from ***PH*** entitled "Five Modes of Agrippa." Although the chapter title mentions five modes, two of them repeat those found elsewhere and are similar to the ones just discussed concerning perception. They are the modes of discrepancy and relativity and are important because they provide the background for understanding the description of the three modes concerning reasoning. Specifically, it is presumed that the relevant object of inquiry is subject to legitimate dispute and that reasoning is employed to resolve the dispute. The issue before us then is whether reasoning can legitimately lead to assent. Sextus writes:

> The Mode based upon regress *ad infinitum* is that whereby we assert that the thing adduced

> as a proof of the matter proposed needs a further proof, and this again another, and so on *ad infinitum*, so that the consequence is suspension [of assent], as we possess no starting-point for our argument … We have the Mode based upon hypothesis when the Dogmatists, being forced to recede *ad infinitum*, take as their starting-point something which they do not establish but claim to assume as granted simply and without demonstration. The Mode of circular reasoning is the form used when the proof itself which ought to establish the matter of inquiry requires confirmation derived from the matter; in this case, being unable to assume either in order to establish the other, we suspend judgement about both. (***PH*** I:166-169)

The question is this: Supposing that the dogmatist assents to something, say *p*, on the basis of a reason, say *q*, and gives *r* as his reason for *q*, etc., how should the Pyrrhonian react in order to avoid the snares of dogmatism? The suggestion in this passage appears to be to force the dogmatist into either an apparently never ending regress or an arbitrary assertion or begging the question.

This strategy seems to be based upon the claim that there are (only) three possible patterns which any instance of reasoning can take. I will call the first pattern "infinitism." Today we commonly refer to the second account as "foundationalism." Finally, I will refer to the third possibility as "coherentism."

The so-called "regress problem," can be stated briefly in this way: There are only three possible patterns of reasoning. Either the process of producing reasons stops at a purported foundational proposition or it doesn't. If it does, then the reasoner is employing a foundationalist pattern. If it doesn't, then either the reasoning is circular, or it is infinite and non-repeating. There are no other significant possibilities.[21] Thus, if none of these forms of reasoning can properly lead to assent, then no form can.

So, we must look briefly at the reasons that a Pyrrhonian might have for thinking that foundationalism, coherentism and infinitism are inherently incapable of providing an adequate basis for assent.[22]

8. The Mode to Respond to the Foundationalist

The Pyrrhonian is not (and cannot consistently be) assenting to the claim that foundationalism is false. Rather, a Pyrrhonian employing this mode would be attempting to reassure herself (and perhaps show the Epistemist) that the so-called foundational proposition stands in need of further support. In other words, the Pyrrhonian believes that a foundationalist cannot rationally practice his foundationalism because it inevitably leads to arbitrariness -- i. e., assenting to a proposition which can legitimately be questioned but is, nevertheless, assented to without rational support.

So, how could the Pyrrhonian proceed? To begin to answer that question it is important to note that foundationalism comes in many forms. But all forms hold that the set of propositions can be partitioned into basic and non-basic propositions. *Basic propositions* have some autonomous bit of warrant that does not depend (at all) upon the warrant of any other proposition.[23] *Non-basic* propositions depend (directly

or indirectly) upon basic propositions for all of their warrant.

Suppose that an inquirer, say Fred D'Foundationalist, has given some reasons for his beliefs. Fred offers *q* (where *q* could be a conjunction) for his belief that *p*, and he offers *r* (which could also be a conjunction) as his reason for *q*. Etc. Now, being a foundationalist, Fred finally offers some basic proposition, say *b*, as his reason for the immediately preceding belief. Sally D'Pyrrhonian asks Fred why he believes that *b* is true. Sally adds the "is true" to make clear to Fred that she is not asking what causes Fred to believe that *b*. She wants to know why Fred thinks that *b* is true. Now, Fred could respond by giving some reason for thinking that *b* is true even if *b* is basic, because basic propositions could have some non-autonomous warrant that depends upon the warrant of other propositions. But that is merely a delaying tactic since Fred is not a coherentist. In other words, he might be able to appeal to the conjunction of some other basic propositions and the non-basic propositions that they warrant as a reason for thinking that *b* is true. But Sally D'Pyrrhonian will ask whether he has any reason that does not appeal to another member in the set of basic propositions for thinking that each member in the set is true. If he says that he has none, then he has forfeited his foundationalism because he is really a closet coherentist. Being true to his foundationalism, he must think that there is some warrant that each basic proposition has that does not depend upon the warrant possessed by any other proposition.

The crucial point to note here is that Sally can grant that the proposition has autonomous warrant but continue to press the issue because she can ask Fred whether the possession of autonomous warrant is at all truth conducive. That is, she can ask whether a proposition with autonomous warrant is, *ipso facto*, at all likely to be true. If Fred says "yes," then the regress will have continued. For he has this reason for thinking that *b* is true: "*b* has autonomous warrant and propositions with autonomous warrant are somewhat likely to be true." If he says "no" then Sally can point out that he is being arbitrary since she has asked why he thinks b is true and he has not been able to provide an answer.

Let us look at an example. Often it is held that first-person introspective reports are basic because they have some "privileged" status. My basic reason for thinking that there is an "external" object of a certain sort is that I am having an experience of a certain sort. Now, what Sally should ask is this: "Why do you think you are having an experience of that sort? Or, again to emphasize that she is not asking for an explanation of the etiology of Fred's belief that he is having an experience of that sort, she could ask: "Why do you think that the proposition 'I am having an experience of a certain sort' is true?"

The dilemma is that either Fred has a reason for thinking that proposition is true or he doesn't. If he does, then the regress has not stopped -- *in practice*. If he doesn't, then he is being arbitrary -- *in practice*.

Once again, it is crucial to recall that Pyrrhonians are *not* claiming that foundationalism is false. They could grant that some propositions do have autonomous warrant which is truth-conducive and that all other propositions depend for some of their warrant upon those basic propositions. What lies at the heart of their view is that there is a deep irrationality in being a practicing self-conscious foundationalist. The question to Fred can be put this way: On the assumption that you cannot appeal to any other proposition, do you have any reason for thinking that *b* is true? Fred not only won't have any such reason for thinking *b* is true, given that assumption, he *cannot* have one (if he remains true to his foundationalism).

Arbitrariness seems inevitable. Of course, foundationalists typically realize this and, in order to avoid arbitrariness, tell some story (for example, about privileged access) that, if true, would provide a reason for thinking basic propositions are at least somewhat likely to be true. But then, the regress of reasons has continued.

9. The Mode to Respond to the Coherentist

At its base, coherentism holds that there are no propositions with autonomous warrant. But it is important to note that coherentism comes in two forms. What I choose to call the "warrant-transfer form" responds to the regress problem by suggesting that the propositions are arranged in a circle and that warrant is transferred within the circle -- just as basketball players standing in a circle pass the ball from one player to another. (See Sosa 1980, and BonJour 1978.) I could, for example, reason that it rained last night by calling forth my belief that there is water on the grass and I could reason that there is water (as opposed to some other liquid, say glycerin, that looks like water) on the grass by calling forth my belief that it rained last night.

Long ago, Aristotle pointed out that this process of reasoning could not resolve matters. As he put it: This is a "simple way of proving anything" (*Posterior Analytics*, I, iii, 73a5). The propositions in the circle might be mutually probability enhancing, but the point is that we could just as well have circular reasoning to the conclusion that it did not rain last night because the liquid is not water and the liquid is not water because it did not rain last night. In this fashion anything could be justified -- too simply! It is ultimately arbitrary which set of mutually probability enhancing propositions we believe because there is no basis for preferring one over the other.

The warrant-transfer coherentist could reply to this objection by claiming that there is some property, *P*, in one of the two competing circles that is not present in the other and the presence of that property makes the propositions in one and only one of the circles worthy of assent. For example, in one and only one of the circles are there propositions that we actually believe, or perhaps believe spontaneously.[24] But, then, it seems clear that the warrant-transfer coherentist has adopted a form of foundationalism since he is now claiming that all and only the propositions in circles with *P* have some autonomous bit of warrant. And, all that we have said about the dilemma facing the foundationalist transfers immediately. Is the possession of *P* truth conducive or not? If it is … You can see how that would go.

So much for the warrant-transfer version of coherentism. The second form of coherentism, what we can call the "warrant-emergent form" does not imagine the circle as consisting of propositions that transfer their warrant from one proposition to another. Rather warrant for each proposition in the circle obtains because it is a member of a set of mutually probability enhancing propositions. Coherence itself is the property in virtue of which each member of the set of propositions has warrant. Warrant emerges all at once, so to speak, from the web-like structure of the set of propositions. The coherentist can then argue that the fact that the propositions cohere provides each of them with some *prima facie* credibility.

This might initially seem to be a more plausible view since it avoids the circularity charge. But, aside

from the problem that are too many competing circles that are coherent, the coherentist has, once again, embraced foundationalism. The coherentist is now explicitly assigning some initial positive warrant to all of the individual propositions in a set of coherent propositions that does not depend upon the warrant of any other proposition in the set. In other words, he is assigning to them what we have called the autonomous bit of warrant and, once again, the dilemma facing the foundationalist returns.

10. The Mode to Respond to the Infinitist

The third mode is designed to provide the Pyrrhonian with a way of responding to a dogmatist who assents to some EI-type proposition, *x*, and ceaselessly provides new answers to the question "What reason do you have for *x*?" Since there is always another reason, one that has not already been employed, that needs to be given for any offered reason, assenting to x would be inappropriate. Since the Pyrrhonian (or the Epistemist) does not know either whether there is such an infinite set of reasons available or whether there is no such set available, withholding assent to the proposition that knowledge of EI-type propositions is possible seems appropriate.

For those reasons, infinitism (as far as I can tell) has never been seriously considered as a model of reasoning suitable for the dogmatist because it is obvious that it cannot provide a model of reasoning that could lead to assent. A disputed proposition could never be fully justified for whenever a reason is provided the infinitist is committed to thinking that in order to settle the matter another, as yet "unused," reason must be provided. Since that process can never be completed, infinitism cannot provide the dogmatist with a model that will settle matters.[25]

11. The Overall Effect of the Modes

It appears that the Pyrrhonian has a viable strategy for resisting dogmatism because no process of reasoning is such that assent -- that is, holding that the matter in question has been settled -- is the appropriate attitude to have toward any non-evident proposition.

Bibliography

- Audi, R., 1988, *Belief, Justification and Knowledge*, Belmont, California: Wadsworth.
- BonJour, L., 1978, "Can Empirical Knowledge Have a Foundation?" *American Philosophical Quarterly*, 15/1, 1-13.
- ------, 1985, *The Structure of Empirical Knowledge*, Cambridge, Mass: Harvard University Press.
- Burnyeat, M., and Frede, M., 1997, *The Original Sceptics: A Controversy*, Indianapolis/Cambridge: Hackett Publishing Co.
- DeRose, K., and Warfield, T. (eds.), 1999, *Skepticism: A Contemporary Reader*, New York and Oxford: Oxford University Press.
- Descartes, R., *Meditations on First Philosophy*, in E. Haldane and G. Ross (eds.), *Philosophical*

Works of Descartes, Volume 1, Dover Publications, 1931.

- Dretske, F., 1970, "Epistemic Operators," *Journal of Philosophy*, 67, 1007-1023.
- Gettier, E., 1963, "Is Knowledge True Justified Belief?" *Analysis*, 23, 121-123.
- Klein, P., 2002a, "Skepticism" in *The Oxford Handbook of Epistemology*, P. Moser (ed.), Oxford: Oxford University Press, forthcoming.
- -----, 2002b, "How a Pyrrhonian Skeptic Might Respond to Academic Skepticism" in *The Skeptics: Contemporary Essays*, Steven Luper (ed.), Ashgate Press, forthcoming.
- -----, 2000a, "Contextualism and the Real Nature of Academic Skepticism," *Philosophical Issues*, 10, 108-116.
- -----, 2000b, "Why Not Infinitism?" *Epistemology: Proceedings of the Twentieth World Congress in Philosophy*, R. Cobb-Stevens (ed.), 2000, vol 5, 199-208.
- -----, 1999, "Human Knowledge and the Infinite Regress of Reasons," *Philosophical Perspectives*, 13, J. Tomberlin (ed.), Atascadero: Ridgeview Press, 297-325.
- -----, 1995, "Skepticism and Closure: Why the Evil Genius Argument Fails," *Philosophical Topics*, 23/1 (Spring): 213-236.
- -----, 1981, *Certainty: A Refutation of Scepticism*, Minneapolis: University of Minnesota Press.
- Lehrer, K., 2000, *Theory of Knowledge*, Boulder, Colorado: Westview Press, second edition.
- -----, K., 1971, "Why Not Skepticism?"*The Philosophical Forum*, 2/3, 283-298. (Page reference is to the reprint in *The Theory of Knowledge*, L. Pojman (ed.), Belmont, CA: Wadsworth Publishing Company, 1993.)
- -----, 1997, *Self-Trust: A Study of Reason, Knowledge and Autonomy*, Oxford: Clarendon Press, Oxford University Press.
- Luper-Foy, S., (ed.), 1987, *The Possibility of Knowledge*, Totowa, NJ: Rowman & Littlefield.
- Malcolm, N., 1963, *Knowledge and Certainty*, Englewood Cliffs, NJ: Prentice-Hall.
- Moore, G.E., 1962, "Certainty" in *Philosophical Papers*, New York, NY: Collier Books.
- Nozick, R., 1981, *Philosophical Explanations*, Cambridge, Massachusetts: Harvard University Press.
- Radford, C., 1966, "Knowledge -- By Example," *Analysis*, 27, 1-11.
- Sextus Empiricus, *Outlines of Pyrrhonism*[PH], R. G. Bury, trans., Cambridge, Massachusetts: Harvard University Press, 1967.
- Sextus Empiricus, *Against the Logicians*, R. G. Bury, trans., Cambridge, Massachusetts: Harvard University Press, 1967.
- Sosa, E., 1980, "The Raft and the Pyramid," *Midwest Studies in Philosophy*, 5, 3-25.
- Stroud, B., 1984, *The Significance of Philosophical Scepticism*, Oxford: Clarendon Press.
- Unger, P., 1975, *Ignorance: A Case for Scepticism*, Oxford: Oxford University Press.
- Vogel, J., 1987, "Tracking, Closure and Inductive Knowledge" in Luper-Foy 1987.
- Wittgenstein, L., 1969, *On Certainty*, New York: Harper Torchbooks.

Other Internet Resources

- Links to papers on Skepticism, in the Epistemology Research Guide, maintained by Keith Korcz (U. Louisiana/Fayetteville)

- Entry on Scepticism, *Oxford Companion to Philosophy*

Related Entries

Descartes, René: epistemology | justification, epistemic: coherentist theories of | justification, epistemic: contextualist theories of | justification, epistemic: foundationalist theories of | justification, epistemic: internalist vs. externalist conceptions of | skepticism: ancient

Acknowledgements

I wish to thank Anne Ashbaugh and Laurence BonJour for their help with this entry. I should also note that some parts of the entry rely upon and, in some cases, significantly repeat sections of Klein 2002a. I also relied on parts of Klein 1995, 1999, 2000a, and 2002b.

Peter Klein
pdklein@rci.rutgers.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 8, 2001
Content last modified: December 8, 2001

Stanford Encyclopedia of Philosophy
Notes to Skepticism

Notes

1. For an interesting discussion of rejecting or neutralizing the skeptic's objections, see Lehrer 2000, 131-136.

2. Wittgenstein 1969. In paragraph 519, he writes:

> Doubt itself rests only on what is beyond doubt. . . But since a language-game is something that consists in the recurrent procedures of the game in time, it seems impossible to say in any *individual* case that such-and-such must be beyond doubt if there is to be a language-game -- though it is right enough to say that *as a rule* some empirical judgment or other must be beyond doubt.

3. Peter Unger's view is an exception. He directs the attack on the psychological condition which he takes to be required for knowledge, namely certainty. He thinks very little, if anything, is such that we are certain of it since "certainty" is an absolute term. Further, he claims that since "certainty" is an absolute term, if we are more (nearly) certain of *x* than we are of *y*, *y* cannot be something about which we are certain. And virtually everything is such that we are more (nearly) certain of something else. See Unger 1975.

4. There is no term readily available. A natural one would be "cognitivist" but that term already has a very specific application in ethics. "Cognitist" just strikes my ears as cacophonic.

5. Sextus Empiricus, *Outlines of Pyrrhonism*, I:226. Also see his discussion of Carneades in *Against the Logicians*, I:159-190. The "logicians" were considered to be dogmatists by Sextus. (In fact, the essays against the logicians were part of his larger work called "Against the Dogmatists." See the preface by Bury, *ibid.*, vii.)

6. Indeed, Keith Lehrer 1997 seems to be giving that response.

7. For a contrasting discussion of the realm of the doubtful, see Stroud 1984, Chapter 1.

8. In the "First Meditation," Descartes does not suggest a potential ground for doubt that he rejects unless, perhaps, that he is mad (insane). He asks whether he could be mad and like people who imagine that they are kings when they are in reality poor, that they are clothed when they are naked or that they

are pumpkins or made of glass. His answer is "But they are mad, and I should not be any the less insane were I to follow examples so extravagant" (*ibid*., 145). That is a puzzling response. Is the evil genius hypothesis less "extravagant?" Or from his point of view, is the possibility that his creator was something other than a perfect god any less extravagant? Nevertheless, he at least seems to be giving reasons for rejecting that grounds for doubt. Of course, later in the *Meditations* he rejects the claim that his maker might have been less than perfect.

9. Strictly speaking, condition (1) entails (3), and hence, (3) is not needed. For if *S* did have a way of neutralizing the effect of *d*, then adding *d* to *S*'s beliefs would not have the effect of making assenting to *p* no longer adequately justified. (That is, ~(3) entails ~(1).) I chose to include (3) to make clear the distinction between denying the potential ground for doubt and neutralizing it.

10. See, for example, DeRose and Warfield 1999. In that volume most of the authors take the CP-style argument to be the primary one. There is an excellent discussion of Academic Skepticism in the Introduction to that volume.

11. For the sake of clarity, it is important to point out that the restoring proposition could itself have a genuine ground for doubt, so that even if (*r* & *d*) did not reduce the warrant for *x*, [(*r* & *d*) & d_1] could defeat the justification for *x* since d_1 would defeat the restoring effect of *r*. But then, (*d* & d_1) would be a new grounds for doubt. So, we need not include this epicycle.

12. Two propositions, *x*, *y* are contraries just in case *x* entails ~*y*, but ~*x* does not entail *y*. Here are some examples: The ball is red all over, the ball is yellow all over; *X* is an aunt, *X* is an uncle. More to the point, *h* and *sk* are contraries since *h* entails ~*sk*, but ~*h* does not entail *sk*. For example, it could be the case that there is no hand before me and I am not in a switched-world (or it doesn't appear that there is a hand before me).

13. The probability could be either subjective or objective. The reason for including "non-overridden" in the supposition is that it would not be sufficient for *S* to be entitled to believe something if *S* only had good enough grounds to render a proposition sufficiently likely to be true because *S* might also have counter evidence that overrides those positive grounds.

14. For another, similar, proposed counterexample, see Audi 1988, 77.

15. For the sake of employing consistent terminology, I have changed "*P*" to "*p*" and "*Q*" to "*q*."

16. For a full discussion of Nozick's account of knowledge, see Luper-Foy 1987.

17. It is crucial to note that the truth of CP does not depend upon the antecedent being fulfilled.

18. The claim here is not that the evidential relationship between *h* and ~*sk* is such that *S must* use Pattern 2. The claim is merely that such a path is available.

19. Much of the material in this section recapitulates Klein 2000a.

20. The caricature is mentioned by Diogenes Laertius, but is also somewhat mitigated by pointing out that the suspension of assent was only "his philosophy" and that Pyrrho lived to be nearly ninety. Here is what he says:

> He [Pyrrho] led a life consistent with his doctrine, going out of his way for nothing, taking no precaution, but facing all risks as they came, whether carts, precipices, dogs or what not, and, generally leaving nothing to the arbitrament of the senses; but he was kept out of harm's way by his friends who . . . used to follow close after him. But Aenesidemus says that it was only his philosophy that was based upon suspension of judgement, and that he did not lack foresight in his everyday acts. He lived to be nearly ninety. (Diogenes Laertius, *Lives of Eminent Philosophers*, Bk IX, 62)

21. Strictly speaking, there is a fourth possibility, namely that there are foundational propositions *and* that there are an infinite number of propositions between a foundational proposition and the one for which reasons are initially being sought. Interestingly, such a hybrid view might be indistinguishable in practice from infinitism and, hence, I think for our purposes we can treat this as a form of infinitism.

22. These are my own glosses on what I take to be the best arguments. I do not claim that the Pyrrhonians gave these very arguments.

23. I put it that way in order to make clear that foundationalism can embrace some aspects of coherentism. Basic propositions with only minimal justification, if coherent, can gain additional credibility. Thus, this account of foundationalism includes both weak and strong foundationalism as characterized in BonJour 1978.

24. This is the suggestion put forward in BonJour 1985.

25. I should note that nothing said here implies that infinitism is an inappropriate model for reasoning that is not meant to result in assent. For a defense of infinitism, see Klein 1999, 2000b, and 2002b.

Peter Klein
pdklein@rci.rutgers.edu

First published: December 8, 2001

Content last modified: December 8, 2001

Ancient Skepticism

Used in its most specific sense, the expression "ancient skepticism" refers to two movements in ancient philosophy. One is Pyrrhonism, which claims Pyrrho of Elis (4th-3rd c. B.C.) as its founder but was especially prominent during and after the 1st c. B.C. The other is Academic Skepticism, which encompasses a skeptical phase in the history of Plato's Academy (3rd to early 1st c. B.C.).

Used more broadly and more loosely, the term "skepticism" is sometimes used in conjunction with a great many ancient thinkers who are not tied to these two movements, but are characterized by significant skeptical inclinations. The most important of these are Protagoras and Socrates, but one might also include Gorgias, Democritus, Aristippus and Diogenes of Sinope (the "Cynic"). While the views of these figures are sometimes mentioned in the present article, it focuses on the narrow notion of "ancient skepticism" and the figures and schools that it encompasses.

An Overview

Following Sextus Empiricus, we can say that the ancient "skeptic" (from a Greek verb meaning "to examine carefully") was an "investigator." He was someone who investigated the questions of philosophy

but "suspended judgment" because he was unable to resolve the contrary attitudes, opinions and arguments that characterized the debated topics of philosophy, hence unable to arrive at a definitive position of his own on any of them. Instead of adhering to some standard philosophical position, the skeptic therefore described himself as someone who continues to investigate -- a "zetetic."

Sextus (end 2nd c. A.D.) describes Pyrrhonian skepticism's relationship to other ancient philosophies in the opening passage of his *Outlines of Pyrrhonism* (*PH*).

> When people search for something, the likely outcome is that either they find it or, not finding it, they accept that it cannot be found, or they continue to search. So also in the case of what is sought in philosophy, I think, some people have claimed to have found the truth, others have asserted that it cannot be apprehended, and others are still searching. Those who think that they have found it are the Dogmatists, properly so called -- for example, the followers of Aristotle and Epicurus, the Stoics, and certain others. The followers of Clitomachus and Carneades, as well as other Academics, have asserted that it cannot be apprehended. The Skeptics [*skeptikoi*] continue to search [i.e., investigate]. (*PH* 1.1-3, Mates)

Two aspects of these remarks warrant special comment. One is Sextus' suggestion that the Pyrrhonian leaves open the possibility of apprehending truth. While consistency forces him to speak this way -- for continuing to investigate makes no sense unless it might conceivably lead to the discovery of some definitive solution to the problems investigated -- one might reasonably wonder whether the ancient skeptic was genuinely open to this possibility. Certainly it must be said that the Pyrrhonian stance one finds in Sextus is an overwhelmingly negative one which functions primarily as a negative critique of any attempt to establish truth.

Sextus' comments on Carneades, Clitomachus and other Academic skeptics are also controversial. His suggestion that they maintain that truth cannot be apprehended (which might imply that they inconsistently maintain they have apprehended that this is true) is most plausibly interpreted as an attempt to drive a wedge between Pyrrhonism and what at Sextus' time was recognized as a competing school of skepticism. In the present context, it is enough to say that these Academics would be skeptics in the sense in which we use the term even if they did adopt the negative dogmatism which Sextus ascribes to them, for it still constitutes a comprehensive rejection of any claim to have apprehended what is true.

The conviction (or deep-seated suspicion) that philosophical claims to apprehend some truth are inherently uncertain is, therefore, the heart of ancient skepticism. The ancient skeptics propounded and defended this conviction by opposing any and all positions with contrary positions, each of which is said to demonstrate the other's uncertainty. Sextus describes the "method of antithesis" this implies when he explains later Pyrrhonism's practice of *epoche* (suspending judgment):

> Broadly speaking, this [suspension of judgement about all things] comes about because of the setting of things in opposition. We oppose either appearances to appearances, or ideas to ideas, or appearances to ideas. We oppose appearances to appearances when we say "The

> same tower seems round from a distance but square from near by." We oppose ideas to ideas when someone establishes the existence of providence from the orderliness of the things in the heavens and we oppose to this the frequency with which the good fare badly and the bad prosper, thereby deducing the non-existence of providence. We oppose ideas to appearances in the way in which Anaxagoras opposed to snow's being white the consideration: snow is water, and water is black, therefore snow is black too. On a different scheme, we oppose sometimes present things to present things, but sometimes present things to past and future things... (*PH* 1.31-5, Long & Sedley)

Diogenes Laertius (1st half of 3rd c. A.D.) associates a similar method of antithesis with Academic skepticism when he writes that Arcesilaus (mid-3rd c. B.C.) "was the originator of the Middle Academy, being the first to suspend his assertions owing to the contrarieties of arguments, and the first to argue pro and contra" (4.28-44, Long & Sedley). The most famous of the Academic skeptics, Carneades (mid-2nd c. B.C.), is reported to have demonstrated his ability to argue for opposing views on a famous trip to Rome, where he is said to have argued impressively for justice on one day and on the next to have argued with equal force against it (Lactantius, *Div. Ins.*, 5.16, 6.6). Judging by Cicero's account in *De Finibus*, the attitude of opposition which this reflects played an integral role in lessons in the skeptical Academy, where the teacher proceeded by opposing a thesis enunciated by a student (e.g., "The Chief Good in my opinion is pleasure"). In a more technical way, antithesis is evident in the Academics' argument against the Stoics' "cataleptic" impressions, which paired alleged examples of such impressions with equally forceful, but indistinguishable impressions which are false.

We might recognize the emphasis on opposition and antithesis which characterized ancient skepticism by describing it as a rejection of our ability to apprehend truth which was founded on the attempt to oppose other philosophies, both by opposing their arguments and positions, and by devising general strategies of opposition.

The Historical Context

It is sometimes said that skeptical doubts characterize times of social upheaval (not only in ancient times but also in the fourteenth century and in contemporary philosophy). Whether these kinds of considerations can help explain the rise of ancient skepticism is very difficult to say, in part because the role which social influences play in determining any philosophical position is inherently complex and obscure. In view of this, it can best be said that ancient skepticism is a natural extension of many of the trends and movements that characterize mainstream ancient philosophy.

Skepticism's affinity to other ancient philosophies is most evident in the kinds of considerations that convince the skeptic that he should suspend judgment on the truth of any philosophical claim. For though skeptical *conclusions* (that on given issues the truth is uncertain) are at odds with the "dogmatist" philosophies the skeptics criticize, these philosophies were frequently founded on a similar concern with opposition, antithesis, and opposing points of view. One might, for example, easily compare the Pyrrhonian conviction that there are equally convincing arguments for and against any claim to the

Protagorean conviction that one can argue convincingly on both sides of any question. This similarity reflects similar philosophical concerns even though Protagoras' conclusion (at least as it is reported by Plato in the *Theaetetus*) that opposing points of view are true is diametrically at odds with the skeptic's rejection of all claims to truth.

The situation is similar in many other cases. Opposing points of view played an important role in the development of Greek atomism, which can be seen as an attempt to explain such opposition by hypothesizing atoms which impact on different kinds of bodies in different kinds of ways. Opposites which include opposing points of view are also emphasized in Heracleitean and Platonic metaphysics. Even Aristotle recognized the possibility of arguing for conflicting points of view in his work on rhetoric.

In other cases, ancient philosophers anticipated skepticism by stressing the difficulties inherent in the search for truth. Xenophanes was known for his claim that no one can know clear truth. Democritus maintained that "bastard" knowledge gained through our senses exists only by convention. Plato rejected everyday opinions, comparing them to shadows in a cave. Diogenes of Sinope, Epictetus and similar moralists dismiss philosophical speculation on the grounds that practical demonstration is what matters. Such philosophers did not endorse a full fledged skepticism, but their views clearly added impetus to the skeptics' moves in this direction.

Much more generally, ancient skepticism flourished in an intellectual climate which was naturally conducive to skeptical conclusions. In marked contrast with modern science, ancient science did not, for example, boast the kinds of practical and theoretical successes we now take for granted. In part because of this, a bewildering array of opposing philosophical perspectives characterized ancient philosophical inquiry and important philosophers were famous for their ability to construct dazzling arguments for paradoxical conclusions (that motion is impossible, that nothing exists, that time is an illusion, etc.). An interest in foreign cultures drew attention to opposing customs and traditions, mysticism and irrationalism flourished as powerful cultural forces, and opposing interests and perspectives were manifest in war, political rivalries and a religion and mythology which pitted god against god, man against man and even god against man. In the midst of the opposing views that this implies, it cannot be judged surprising that radical skepticism in a variety of forms became a prominent philosophical perspective.

Pyrrho and Equanimity

The movements that make up ancient skepticism begin with Pyrrho (ca. 365-ca. 275 B.C.). In marked contrast to modern skeptics, he proposed skepticism as a way of life which functioned as a route to equanimity and contentment. He left no writings, and except for what has survived of his pupil Timon's exegetical writings ancient reports about him are heavily colored by anachronism drawn from the philosophical outlook of the Pyrrhonian "revival" in the 1st c. B.C. (see below, "The Arguments for Later Pyrrhonism"). Sextus is, for example, noticeably reticent when he declares that "Pyrrho appears to us to have applied himself to Skepticism [i.e. Pyrrhonian skepticism as Sextus knew it] more thoroughly and more conspicuously than his predecessors" (*PH*1.7, Bury). We find a less hedged view of Pyrrho and his views in the following fragment of Aristocles (Peripatetic of uncertain date, perhaps 1st c. B.C.-A.D.,

perhaps 2nd c. A.D.).

> He [Pyrrho] himself has left nothing in writing, but this pupil Timon says that whoever wants to be happy must consider these three questions: first, how are things by nature? Secondly, what attitude should we adopt towards them? Thirdly, what will be the outcome for those who have such an attitude? According to Timon, Pyrrho declared that things are equally indifferent, unmeasurable and inarbitrable. For this reason neither our sensations nor our opinions tell us truths or falsehoods. Therefore for this reason we should not put our trust in them one bit, but should be unopinionated, uncommitted and unwavering, saying concerning each individual thing that it no more is than is not, or both is and is not, or neither is nor is not. The outcome for those who actually adopt this attitude, says Timon, will be first speechlessness [*aphasia*], and then freedom from disturbance; and Aenesidemus says pleasure. (Eusebius, *Prep. Ev*. 14.18.2-5, Long & Sedley)

According to Diogenes Laertius (9.76), Timon explained Pyrrho's formula *ou mallon* ("no more is than is not") as a way of expressing a decision to suspend judgment and determine nothing. The practical result of the indifference to opinions and sensations which results is Pyrrho's "peace of mind" (D.L. 9.65).

The *Life of Pyrrho* which Diogenes Laertius includes in his *Lives of Eminent Philosophers* suggests that Pyrrho lived a life in accord with his own emphasis on equanimity and indifference. Among other things, he lived like a recluse, did not "so much as frown" when treated with disinfectants, surgery and cautery, voluntarily adopted a life of piety and poverty, and performed menial tasks to show his indifference. According to one anecdote, he was criticized when he failed to maintain his composure when a cur rushed at him and terrified him (Pyrrho answered that it is difficult to strip oneself of human nature). The citizens of his native Elis rewarded him with honors, making him a high priest, raising a statue in his honor (Pausanias 6.24.5), on his account passing a law which exempted philosophers from taxes (D.L. 9.64).

Flintoff locates the origins of Pyrrho's philosophy in India, where Pyrrho travelled with the court of Alexander the Great and was in this way exposed to Indian ascetics and their commitment to an enlightened state of mind. Certainly it is likely that Pyrrho was impressed with the indifference of India's gymno-sophists (the "Naked Philosophers"). This much being granted, it can still be said that his skepticism has Greek origins which are plausibly located in Democritean atomism.

Looked at from the point of view of earlier Greek philosophy, Pyrrho's skepticism is a natural evolution of Democritus' doubts about ordinary opinions, which Democritus rejected as purely "conventional" on the grounds that they are contradictory and truth resides in atoms and the void. It is in keeping with this that Pyrrho's teacher is the Democritean Anaxarchus (whom he followed to India); his formula *ou mallon* is borrowed from atomism (DeLacy); his goal of equanimity reflects Democritean practical ideals; and he is said to have admired Democritus above all others (D.L. 9.67). But Pyrrho takes skeptical inclinations one step further than Democritus and rejects atomism as well ordinary opinions, in the process giving up on philosophy and on all attempts to establish what is true (D.L. 9.69,65; cf. Sextus, *PH* 1.28-29; *AM* [*Adversus Mathematicos*] 11.1). As Aristocles puts it, "if we are so constituted that we know nothing, then there is no need to continue enquiry into other things.... Pyrrho of Elis was ... a powerful spokesman

of such a position" (Eusebius, 14.18.1-2, Long & Sedley).

Sextus describes Pyrrhonism's ties to equanimity (*ataraxia*) with an anecdote which probably relates events which occurred during Pyrrho's time with Alexander's court. It tells how Apelles, Alexander's court painter, was frustrated by his inability to paint the froth on a horse's mouth and in exasperation threw a sponge at his painting, accidentally producing the effect he wanted. "So, too, the Skeptics were hoping to achieve *ataraxia* by resolving the anomaly of phenomena and noumena, and, being unable to do this, they suspended judgment. But then, by chance as it were, when they were suspending judgment the *ataraxia* followed, as a shadow follows the body." (*PH* 1.29, Mates)

Apparently, equanimity accompanies skepticism "like a shadow" for two reasons. First, because it eliminates the anxiety that accompanied the study of philosophy in the hope of arriving at an apprehension of the truth about reality and what is good and bad in human life. Second, it promotes indifference to the misfortunes and calamities that disturb our peace of mind, for the skeptic concludes that misfortunes and calamities can't be known to be bad. In the context of his actual life, Pyrrho probably maintained his attitude of calm composure by using the method of antithesis outlined in the following fragment of Democritus:

> [In order to achieve cheerfulness]... one must keep one's mind on what is attainable, and be content with what one has, paying little heed to things envied and admired, and not dwelling on them in one's mind. Rather must you consider the lives of those in distress, reflecting on their intense sufferings, in order that your own possessions and condition may seem great and enviable, and you may, by ceasing to desire more, cease to suffer in your soul... One must... [compare] one's own life with that of those in worse cases, and must consider oneself fortunate, reflecting on their sufferings, on being so much better off than they. If you keep to this way of thinking, you will live more serenely (fr. 191, cf. fr. 3; Kirk, Raven and Schofield).

The exercises here proposed allow one to be content by continually opposing one's misfortunes with comparisons that make one seem well off. The relativity of value judgments -- a natural component of skepticism -- can in this way provide a psychological basis for peace of mind. As the old saw goes, "I was upset about my lack of shoes until I met a man with no feet."

Pyrrho's own use of such tactics is implied by the report that he was fond of Homer's lines (*Il.* 21.106-7): "Ay friend, die thou; why thus thy fate deplore? Patroclus, thy better, is no more" -- lines that combat upset with one's own fate with the thought that one does not deserve anything better, since the great warrior Patroclus has suffered the same. Oppositions of this sort are probably implied when it is said that Pyrrho "talked to himself" when he trained himself to be good (D.L. 9.64, cf. 69). The same method and ideals are reflected in an incident in which his teacher Anaxarchus cures Alexander's despondency after he has killed a friend (Plutarch, *Alex.*, 52), and in Anaxarchus' own fame as "the happy one," which is in part founded on (or perhaps the basis for) the story that he was unflappable even when he sufferred a horrible death at the hands of the tyrant Nicocreon (D.L. 9.59-60).

Appearances

Pyrrho's philosophy raises a number of issues which reverbate throughout the history of skepticism. Questions about the consistency of the skeptical perspective are particularly significant. As Aristocles says, "in admonishing us to have no opinion, they [the skeptics] at the same time bid us to form an opinion, and in saying that men ought to make no statement they make a statement themselves: and though they require you to agree with no one, they command you to believe themselves..." (Eus. *Prep. Ev*. 14.18, Gifford).

Other commentators ask how Pyrrho survived the pitfalls of day to day life -- much less achieved supreme contentment -- if he refused to believe the truth of his sense impressions. According to one ancient report, this was a practical as well as a theoretical issue, for Pyrrho accepted skepticism "in his actual way of life, avoiding nothing and taking no precautions, facing everything as it came, wagons, precipices, dogs, and entrusting nothing whatsoever to his sensations. But he was looked after... by his disciples, who accompanied him" (D.L. 9.62, Long & Sedley).

Though the consistency of skepticism is open to debate, not much is to be made of this account of Pyrrho's actions, which can be grouped with many other unbelievable stories which Diogenes Laertius reports -- that Pythagoras descended into Hades, that Apollo appeared to Plato's father, that Zeno of Elea (and, again, Aristarchus) bit off his tongue and spat it at a tyrant who was persecuting him, and so on. Laertius has a penchant for such stories and is happy to stretch himself to include them -- in this case he does so by citing as his authority "those around" Antigonus of Carystus (3rd c. B.C. author of *Lives of Philosophers*), making this account of Pyrrho little more than a rumor.

As Hallie says, we can usefully contrast the claim that Pyrrho rejected the senses with Posidonius' account (1st c. B.C.) of his actions when he was caught in a wild storm at sea (D.L. 9.68). Confronted with other passengers wailing and cringing with horror, Pyrrho is said to have remained calm and pointed to a small pig which was calmly eating on the deck, saying that its attitude demonstrated the unperturbed state of the wise man. Even though Timon included "sensations" as well as "opinions" within the scope of Pyrrho's skepticism, this suggests that it is human fears and frailties, not sense impressions, which Pyrrho was concerned to expunge by skeptical inquiry.

It can still be asked how Pyrrho could consistently embrace his senses and his skeptical conclusions. Timon answers that the Pyrrhonian guides himself by "appearances" (*phainomena* -- what "appears to be the case"). This suggests that Pyrrho rejected claims to truth and viewed his skepticism and his day to day beliefs as a mere acceptance of appearances that stops short of claims to truth. As Diogenes Laertius puts it:

> ...the dogmatists say that they [the skeptics] abolish life, in the sense that they throw out everything that goes to make up a life. But the skeptics say that these charges are false. For they do not abolish, say, sight, but only hold that we are ignorant of its explanation.... We do sense that fire burns, but we suspend judgement as to whether it is fire's nature to burn....

> "We only object," they say, "to the non-evident things added on to the phenomena [the appearances].... For this reason, Timon in his *Pytho* says that he has not diverged from what is customary. And in his *Likenesses* he says, "But the apparent utterly dominates wherever it goes." And in his work *On the Senses* he says, "That honey is sweet I do not posit; that it appears so I concede." (D.L. 9.104-5, Inwood & Gerson)

Such claims suggest that we should interpret early Pyrrhonian claims -- and even Pyrrho's claim that things "are" indifferent, unmeasurable and inarbitrable -- as claims about what appears to be the case. Whether such moves can, in the end, save the skeptic from the charge of inconsistency is a matter of much debate (for two sides of this debate, see Frede and Burnyeat).

Arcesilaus in the Academy

Pyrrho's impact on his immediate contemporaries seems quite limited. Timon is his only student of repute and ancient skepticism's next phase is not Pyrrhonian but Academic. No doubt the Academy became a school of skepticism by exploiting the skeptical aspects of Plato's philosophical writings -- Socrates' heroic skepticism in the early dialogues; the questioning of the forms in the *Parmenides*; Plato's pessimism about "ordinary" knowledge; and the indeterminate nature of his dialogues which are intrinsically open to many interpretations. Cicero, who defends a late version of Academic skepticism, says Plato is a skeptic because he is always arguing pro and contra, states nothing positively, inquires into everything, and makes no certain statements (*Ac* 1.46).

The first of the Academic skeptics is Arcesilaus (316/315-242/241 B.C.), the head of what Diogenes Laertius calls the "Middle" Academy. He was influenced by Plato, Pyrrho and Diodorus Cronus (a dialectician of impressive skill). Ariston (3rd c. Stoic philosopher) described him as "Plato in front, Pyrrho behind, Diodorus in the middle" (D.L. 4.33). According to Sextus, his skepticism is "virtually identical" with Pyrrhonism (*PH* 1.232). While Arcesilaus was no ascetic (see, e.g., D.L. 4.37-42), he still held that skepticism aims at happiness (*AM* 7.158) and some of the anecdotes we find in Plutarch suggest that he, like Pyrrho, believed we should deal with misfortune and unhappiness by finding opposing ways of looking at trying situations (see "On Controlling Anger," 461E and "On Tranquillity of Mind," 470A-B).

Arcesilaus' arguments focus primarily on Stoic epistemology. According to Couissin, he has no views of his own on the epistemological topics he disputes with the Stoics, and offers his arguments merely as a reductio ad absurdum of the Stoic point of view. It is difficult (perhaps impossible) to judge whether this is so in the context of scanty evidence almost 2000 years later, especially as it is never easy to tell how a philosopher intends a particular argument or position (Caton has even argued that Descartes is not committed to the *cogito* in the *Meditations*).

However one interprets it, the crux of Arcesilaus' attack on Stoic epistemology is his attack on the "cataleptic" impression (the *kataleptike phantasia*). According to the Stoics, such an impression is clear and distinct and -- in virtue of its clearness and distinctness -- reveals certain truth. As such, it becomes

the criterion or foundational guarantee of truth. According to Arcesilaus, there is no such impression (and no impression can be a guarantee of truth), for any allegedly cataleptic impression can be paired with an impression which is equally clear and distinct but nonetheless mistaken -- because it is experienced in dreaming, hallucinating, etc. (*Ac*. 2.77; *AM* 7.252).

According to Sextus, Arcesilaus combines his skeptical arguments with a commitment to "the reasonable" (the *eulogon*) which he propounds as a practical criterion in day to day affairs.

> ... since it was necessary ... to inquire into the conduct of life which naturally cannot be directed without a criterion, upon which happiness too, that is, the goal of life depends for its reliability, Arcesilaus says that he who suspends judgment about everything regulates choices and avoidances and, generally, actions by reasonableness, and, proceeding according to this criterion, will act correctly. For happiness arises because of prudence, and prudence resides in correct actions, and a correct action is that which, having been done, has a reasonable defence. Therefore, he who adheres to reasonableness will act correctly and will be happy. (*AM* 7.158, Inwood & Gerson)

Those who, like Couissin, see Arcesilaus as a purely negative dialectician do not believe that he actually endorsed such views, at any rate not on the basis of an independent examination of the issues (as against what would reasonably follow from Stoic assumptions). Sextus, however, does not frame his report of Arcesilaus' views in this way, and seems to claim that the use of the "reasonable" as a criterion of choice was Arcesilaus' own philosophical position, which he proposed as an alternative to a reliance on the Stoics' "cataleptic impression." In any case, some such commitment makes good philosophical sense (see Hankinson, 86-91), especially in a historical context in which philosophy is expected to provide a practical guide to life.

One might, of course, still debate whether Arcesilaus' skepticism was consistent with an acceptance of the "reasonable" or, much more fundamentally, the actions which daily life requires (for one might argue that eating, drinking, moving, etc. require beliefs that skepticism undermines). That this was a heated issue already in ancient times is evident in Plutarch's *Against Colotes* (2nd c. A.D.), which takes Colotes (3rd c. B.C.) to task for his attack on Arcesilaus and other philosophers in a book entitled *On the fact that the doctrines of the other philosophers make it impossible even to live*. Colotes' book also attacked Democritus, Aristotle, Parmenides, and Socrates -- indeed, virtually everyone but his mentor, Epicurus -- so he was not concerned especially with Arcesilaus. But Plutarch includes a notable defense of Arcesilaus in his response. It argues that the soul has three movements: sensation, impulse, and assent, and that Arcesilaus allows us to accept sensation and impulse so long as we stop short of assent and opinion (*Mor*. 1122C-D). According to an angry Plutarch, it follows that Arcesilaus' views in this way provide a basis for action and get from Colotes the kind of unappreciative attention that a performance on the lyre gets from an ass (*Mor*. 1122B).

Carneades in the Academy

After Arcesilaus, the leadership of the Academy passed to Lacydes, to Telecles and Evander, and then to Hegesinus. Little is known about their views, but it seems that they preserved Arcesilaus' skepticism. The next phase in the history of ancient skepticism begins with Carneades (214/213-129/128 B.C.), who Diogenes Laertius describes as the head of the "New" Academy. Though he wrote nothing, he appears to have been a remarkably successful philosopher. So much so that Numenius (2nd c. A.D. Platonist) says, in a fragment in Eusebius (a 3rd-4th c. A.D. Christian bishop), that he was victorious on every issue. According to Diogenes Laertius, he became so famous attacking Stoic arguments that he said, "if Chrysippus had not existed neither would I," mimicking the Stoic maxim, "if Chrysippus had not existed, neither would the Stoa" (D.L. 4.62, cf. 7.183).

One finds an account of two of Carneades' central arguments against the "criterion" of knowledge (including especially the Stoic "cataleptic impression") in Sextus' work, "Against the Logicians" (*AM*7.159-165). According to Sextus' account, they were addressed against all of Carneades' (dogmatic) predecessors. The first maintained that there can be no criterion of certain truth because reason, the senses, and any other supposed criterion can play us false. The second argued that the impressions (or "presentations") that inform our judgments are not purely objective, but reflect also their own subjective nature -- as light shows both itself and the things it illuminates. It appears that the subjectivity of impressions which this second argument emphasizes was underscored by an appeal to the by now standard argument that any impression which appears true can be paired with (and opposed by) an indistinguishably similar impression which is false.

Though Carneades' cleverness in argument (rather than the moral austerity we associate with Pyrrho) is the most notable feature of extant evidence about him, Cicero implies that he used antithesis to promote equanimity when he says that Carneades criticized Chrysippus for approving of a passage in which Euripides recounts the pain of life. According to Carneades, Chrysippus was promoting depression whereas the passage should instead be used to bring comfort to the ill-disposed by reciting the misfortunes of others (*Tusc. Disp.* 3.59-60). A similar concern with equanimity is evident in Carneades' claim that we should oppose the expected with the unexpected -- health with the possibility of sickness, safety with the possibility of accident, etc. -- because the unexpected causes us grief when it catches us off guard (Plutarch, *Tranq.* 474F-75A). In a famous speech Carneades demonstrated how to use opposing arguments as a means of promoting peace of mind by arguing, for the sake of Clitomachus in the wake of the destruction of his native Carthage, that the wise man is not distressed even at the loss of his native city (Cicero, *Tusc. Disp.* 3.54).

Despite his arguments against all criteria of certain truth, Eusebius says that Carneades did not suspend judgment on all matters (*Prep. Ev.* 14.7.15), but distinguished between things that are "non-evident" (non-apparent) and those that are "non-apprehensible." According to this account, he held that everything is non-apprehensible but that some things are not non-evident. It is tempting to compare this alleged commitment to "evident" things with the Pyrrhonian commitment to appearances, but this is difficult given that Carneades (unlike the Pyrrhonians) is said to rank different kinds of impressions as more and less persuasive.

In "Against the Logicians," Sextus, in conjunction with his report of Carneades' attacks on the Stoic

theory of cataleptic impressions, says that Carneades adopted the *pithanon* (the "plausible") as a practical criterion and distinguished between impressions which are: (i) implausible; (ii) plausible (i.e. appear true "to an intense degree"); (iii) irreversible (i.e. plausible and confirmed by other impressions); and (iv) tested (i.e. irreversible and tested by the scrutiny of surrounding circumstances). One might argue that this is an improvement on Stoic epistemology, insofar as it suggests that they should propose as their criterion, not the merely "clear and distinct" impressions, but those that are irreversible, and tested as well. According to Sextus' account of Carneades' views, he added some further sophistication to his criterion of choice by holding that different levels of plausibility are appropriate in different kinds of circumstances. While he proposed plausible impressions as a guide in matters of no importance, for example, he is said to hold that weighty matters call for impressions which are irreversible and tested (*AM* 7.184).

Sextus illustrates Carneadean plausibility with an illuminating example:

> On seeing a coil of rope in an unlighted room a man jumps over it, conceiving it for the moment to be a snake [i.e. judging this to be plausible], but turning back afterwards he inquires into the truth, and on finding it motionless he is already inclined to think that it is not a snake [for this impression seems reversible], but as he reckons, all the same, that snakes too are motionless at times when numbed by winter's frost, he prods at the coiled mass with a stick, and then, after thus testing the impression received, he assents to the fact that it is false to suppose that the body presented to him is a snake. (*AM* 7.187-88, Bury)

Judging by Sextus and some of our other ancient sources, Carneades tried to make the *pithanon* compatible with his skepticism by emphasizing that plausibility is inherently subjective and a criterion of choice but not a measure of objective probability or truth. Clitomachus thus writes that "The Academic school holds that there are dissimilarities between things of such a nature that some of them seem plausible and others the contrary; but this is not an adequate ground for saying that some things can be apprehended [or grasped as true] and others cannot, because many false objects are plausible..." (Cicero *Ac*. 2.103, Rackham, tr. altered; cf. 104 and *AM* 7.169). This makes Carneadean assent to something's plausibility consciously subjective and, in view of this, more constrained than the assent which seems to be implied by claims to truth.

Though Carneades may in this way have avoided claims to truth, his account of plausible and implausible impressions still drives an important wedge between his views and those of the Pyrrhonians, for the Pyrrhonians attempt to accept appearances with a minimum (one might say ascetic) inclination that seems incompatible with the conviction that some of the things assented to are highly plausible. As Sextus puts it, "[A]lthough both the [later] Academics and the Skeptics say that they are persuaded of certain things, here too the difference of the philosophies is very evident. For ‘to be persuaded’ has different senses: on the one hand, it means not to resist but simply to follow without much proclivity or strong pro feeling, as the child is said to be obedient to his teacher; but sometimes it means assent to something by choice and with a kind of sympathy due to strong desire, as when a profligate man is persuaded by one who approves of living extravagantly. Since, therefore, the followers of Carneades and Clitomachus say both that they are strongly persuaded and that things are strongly persuasive [i.e. plausible, *pithanon*], whereas we say

that we simply make a concession without any strong feeling, we would differ from them in this respect, too." (*PH* 1.230, Mates). This difference highlights the much more significant role that ascetic indifference plays in Pyrrhonian -- as opposed to Carneadean -- skepticism.

Carneades as Dialectician

Some commentators on ancient skepticism argue that Carneades did not endorse the positive philosophy implied in the suggestion that we should follow the "plausible" impression, which Sextus seems to ascribe to him. According to this reading, Carneades proposed the plausible merely "for the sake of argument" -- to show that alternatives to dogmatic epistemology as a basis for living an active life are in principle possible (Striker's views in this regard are notable). On this interpretation, Carneades was a dialectician, and a skeptic only in the sense that he never committed himself to any of the premisses from which he argued, or to any of the conclusions he drew from them. It follows that he was not a full fledged skeptic, in the sense of one who believed that no certain knowledge was possible, or who (given that) advocated a skeptical way of life in dependence on mere "belief" in "plausible" ("irreversible," or "tested") impressions. On this view his achievement was not a skeptical philosophy but a dialectical ability to argue for (and primarily against) any point of view.

The issue is a thorny one, as any philosopher is likely to act as a dialectician at some time or other, and dialectical argument is an integral part of ordinary skepticism, which continually propounds particular points of view "for the sake of argument." Sextus is a case in point, for he spends very little time expounding his own philosophy and instead propounds a huge catalogue of arguments with conclusions to which he is not, in the final analysis, committed. If we had lost only a few pages of his extant works, we could easily have been left with texts which were completely dialectical.

In this context, it may be useful to consider Cicero, *Academica* 2.78, where Philo (of Larissa, Cicero's Academic teacher) and Metrodorus (a pupil of Carneades') are said to attribute to Carneades a skepticism which holds that the wise man cannot apprehend anything (grasp it as true) but may accept an opinion nonetheless. Cicero says he prefers the view of Clitomachus, who holds that Carneades "did not so much accept this view as advance it in argument." This clearly suggests that Carneades offered such a view only for dialectical purposes (as an account of what, given premisses they would accept, one should say about the "wise man" of the Stoics and other philosophers), but it provides limited evidence for the dialectical interpretation, for it does not show that this is Carneades' only mode of argument (cf. Hankinson, 94).

The most important textual evidence in favor of the dialectical interpretation is found at *Academica* 2.139, where Cicero says that Clitomachus used to declare that he had never been able to understand what Carneades did accept (see Striker, 55; Hankinson, 94; Inwood & Gerson, 165; Long & Sedley, Vol 1, 455). This is not the claim that Carneades was a dialectician, however, and it is compatible with the possibility that Clitomachus believed that Carneades accepted some claims, but he was not sure which. More importantly perhaps, Cicero's report is embedded in a discussion of the good, in which Carneades is said to have defended Calliphon's view that the good is pleasure with such zeal "that he was thought

actually to accept it (although Clitomachus used to declare that he had never been able to understand what Carneades did accept)" (tr. Rackham). Taken in this context, the parenthetical comment about Clitomachus' view of Carneades can be interpreted as the claim that Clitomachus did not understand what Carneades held *in this regard*. It does not, therefore, provide definitive evidence for the claim that Carneades was a dialectician and did not have definite views of his own on disputed questions of philosophy.

The dialectical interpretation of Carneades does have the advantage that it saves Carneades from inconsistency (for if he has no positive philosophical views, he need not render anything he says in one argument consistent with what he may say in another), though this advantage gained in this regard is earned by turning Carneades' philosophy into a purely negative philosophy which provides no basis for action. One might therefore try to excuse Carneades from inconsistency without abandoning the claim that he is a skeptic (in the sense that he believes certain knowledge impossible), by emphasizing (as Sextus suggests) the qualified and subjective nature of the assent that he endorsed. So understood, his commitment to persuasiveness ("plausibility") and the assent that this implies is, in virtue of its subjectivity, an attempt to formulate a conception of belief which is compatible with a rejection of claims to objective truth.

The Arguments for Later Pyrrhonism

Carneades' successor as head of the Academy was Clitomachus (d. 110/9 B.C.), the author of exegetical writings (now lost) reporting and explaining Carneades' arguments and his skepticism -- writings referred to in Cicero's *Academica*. He was succeeded by Philo of Larissa (d. 84/3 B.C.). The latter taught, on the basis of Carneades' notion of "plausible" impressions, an epistemology which allowed one to adopt whatever position on a disputed philosophical question seemed to oneself most persuasive, after thorough examination of arguments on all sides -- provided that one carefully refrained from claiming to have established the truth on the matter in question with certainty. As we can see from his pupil Cicero's philosophical writings, this meant in practice the adoption (in this tentative spirit) of many Stoic positions.

The next important ancient skeptic was Aenesidemus, who defected from Philo's Academy and revived Pyrrhonism in the early years of the first century B.C. "The Academics," he said, "especially the ones now, sometimes agree with Stoic opinions and, to tell the truth, appear to be just Stoics in conflict with Stoics" (Photius, *Bibl.* 212, Inwood & Gerson). In response, his eight books of *Pyrrhonian Arguments* propounded the view that "the Pyrrhonist determines nothing, not even this, that he determines nothing" (ibid.). It is perhaps ironic that he himself is reported to have given up on Pyrrhonism, and to have finished his career as a Heracleitean, apparently on the grounds that skeptical antithesis should be seen as a road leading to the realization that reality is full of opposites (*PH* 1.210, compare *AM* 7.349, 9.336-67, 10.216, and Tertullian, *De Anima* 9.5, 14.5).

Though Aenesidemus' books on Pyrrhonism do not survive, they are summarized by Photius (9th c. A.D.), whose account suggests that they systematized Pyrrhonism by establishing standard argumentative

strategies and collecting an array of arguments, puzzles and conundrums borrowed from the whole of Greek philosophy.

We know of later Pyrrhonism primarily through three surviving works of Sextus Empiricus (ca. 200 A.D.): *The Outlines of Pyrrhonism*; a second work *Against the Dogmatists*, consisting of "Against the Logicians" (2 books), "Against the Physicists" (2 books), and one book "Against the Ethicists;" and a third work called *Against the Learned* (*Adversus Mathematicos*), combining the latter five books with six further ones attacking the epistemological pretensions of mathematicians, grammarians, etc. The relations between these books are complex and not yet well explored (in Sextus 1997, Bett argues for a reading of *Against the Ethicists* which would make it propound a very different skepticism than the *Outlines of Pyrrhonism*).

Aenesidemus' most important arguments are the ten modes (or "tropes") which Sextus attributes to "the older skeptics" at *PH* 1.35-163. They create antitheses and promote *epoche* by contrasting:

> (i) the opposing perceptions and views of the world which characterize different species: "For how could one say, with regard to touch [for example], that animals are similarly affected whether their surfaces consist of shell, flesh, needles, feathers or scales? And, as regards hearing, how could one say that perceptions are alike in animals with a very narrow auditory canal and in those with a very wide one, or in those with hairy ears and those with ears that are hairless... [P]erfume seems very pleasant to human beings but intolerable to dung beetles and bees, and the application of olive oil is beneficial to human beings but kills wasps and bees." (*PH* 1.50, 55, Mates)
>
> (ii) the opposing perceptions and views of the world which characterize different individuals: "...the greatest indication of the vast and limitless difference in the intellect of human beings is the inconsistency of the various statements of the Dogmatists concerning what may be appropriately chosen, what avoided, and so on." (*PH* 1.85-86, Mates)
>
> (iii) the opposing perceptions and views of the world which characterize different sense organs: "Pictures seem to the sense of sight to have concavities and convexities," for example, "but not to the touch," and "Let us imagine someone who from birth has ...lacked hearing and sight. He will start out believing in the existence of nothing visible or audible, but only of the three kinds of quality he can register. It is therefore a possibility that we too, having only our five senses, only register from the qualities belonging to the apple those which we are capable of registering. But it may be that there objectively exist other qualities" (*PH* 1.92, 96-7, Mates).
>
> (iv) the opposing perceptions and views of the world which characterize different circumstances: "Thus, things affect us in dissimilar ways depending on whether we are in a natural or unnatural condition, as when people who are delirious or possessed by a god seem to hear spirits but we do not.... And the same water that seems to us to be lukewarm seems boiling hot when poured on an inflamed place.... Further, if someone says that an

intermingling of certain humors produces, in persons who are in an unnatural condition, odd *phantasiai* [impressions] of the external objects, it must be replied that since healthy people, too, have intermingled humors, it is possible that the external objects are in nature such as they appear to those persons who are said to be in an unnatural state, but that these humors are making the external objects appear to the healthy in a natural people other than they are. (*PH* 1.101-2, Mates).

(v) the opposing perceptions and views of the world that characterize different positions and distances and places: for example, "lamplight appears dim in sunlight but bright in the dark. The same oar appears bent in water but straight when out of it" (*PH* 1.119, Mates).

(vi) the opposing perceptions and views of the world that characterize mixtures: "[W]e deduce that since no object strikes us entirely by itself, but along with something, it may perhaps be possible to say what the mixture compounded out of the external object and the thing perceived with it is like, but we would not be able to say what the external object is like by itself... The same sound appears one way when accompanied by a rarefied atmosphere, another way when accompanied by a dense atmosphere" (*PH* 1.124, 125, Mates).

(vii) the opposing perceptions and views of the world due to different quantities and structures: "[I]ndividual filings of a piece of silver appear black, but when united with the whole they affect us as white... And wine, when drunk in moderation, strengthens us, but when taken in excess, disables the body..." (*PH* 1.129, 131, Mates).

(viii) the opposing views possible because of the relativity of all things: "...since all things are relative, we will suspend judgment about what things exist absolutely and in nature... This has two senses. One is in relation to the judging subject [different subjects perceiving differently]... The other in relation to the conceptions perceived with it..." (*PH* 1.135, Mates).

(ix) the opposing perceptions and views of the world due to constancy or rarity of occurrence: "The sun is certainly a much more marvelous thing than a comet. But since we see the sun all the time but the comet only infrequently, we marvel at the comet so much as even to suppose it a divine portent, but we do nothing like that for the sun. If, however, we thought of the sun as appearing infrequently and setting infrequently, and as illuminating everything all at once and then suddenly being eclipsed, we sould find much to marvel at in the matter." (*PH* 1.141, Mates). And

(x) the opposing perceptions and views of what is right and wrong which characterize different ways of life, laws, myths and "dogmatic suppositions": "among the Persians sodomy is customary but among the Romans it is prohibited by law; and with us adultery is prohibited, but among the Massagetae it is by custom treated as a matter of indifference, as

> Eudoxus of Cnidos reports... and with us it is forbidden to have intercourse with one's mother, whereas with the Persians this sort of marriage is very much the custom. And among the Egyptians men marry their sisters, which for us is prohibited by law. (*PH* 1.152, Mates).

Later Pyrrhonian modes more clearly isolate the basic epistemological issues which are raised by the traditional ten modes. The five modes of Agrippa (date unknown; later than Aenesidemus), discussed at *PH* 1.164-77 (which are analyzed in detail by Barnes) promote the suspension of judgment by invoking:

> -- *disagreement*, for among philosophers and ordinary people there is interminable disagreement;
>
> -- *regress ad infinitum*, for the skeptic asks for a proof of a claim, a proof of the reliability of this proof, and so on ad infinitum;
>
> -- *relativity*, for things are relative to both one's subjective nature and the concepts one employs in judging them;
>
> -- *hypothesis*, for the skeptic does not allow us to take as our starting point something which is taken for granted;
>
> -- *circular reasoning*, for the skeptic rejects proofs that are circular, as when sense impressions are used to establish the veracity of the senses.

The standard modes are reduced even further in a basic set of two modes propounded in the following section of the *Outlines of Pyrrhonism* (1.178-79). There it is argued that everything which is apprehended (as true) must be apprehended through itself or some other thing. But according to the Pyrrhonians, the first alternative is undermined by the "controversy among philosophers" and the second by a demand for justification which entails a regress *ad infinitum* which can be stopped only by claiming that something is apprehended as true in virtue of itself (a possibility undermined by the first mode).

The various sets of Pyrrhonian modes systematize ancient arguments against dogmatic philosophical positions, but we should not exaggerate the role they played in ancient skepticism. Judging by Sextus, they are usually backed -- and very frequently supplanted -- by an enormous catalogue of other, specific arguments which were used to argue for *epoche* on whatever topic happens to be at hand (space, time, the good, the gods, fate, the meaningfulness of standard conceptions of human nature, and so on and so forth). No encyclopedia article can fully convey the spirit of the seemingly endless assortment of claims and counter claims that Sextus is ready to marshal on any topic.

The Practical Criterion

In the midst of Sextus' attack on other philosophers, it is easy to forget that he, like Pyrrho, proposed skepticism as a way of life (an *agoge*). Its practical merits are said to include its alleged ability to undermine useless and unfounded speculation which is claimed to characterize dogmatist philosophy. Like Hume, the later Pyrrhonians in this way attempt to supplant philosophical speculation with mundane matters of practical concern. The spirit of this rejection is well captured at *PH* 2.241-44, where Sextus condemns the convoluted arguments and conundrums of ancient dialectic:

> As regards sophisms the exposure of which is useful, the dialectician will not have a word to say, but will propound such arguments as these -- "If it is not so that you both have fair horns and have horns, you have horns; but it is not so that you have fair horns and have horns, therefore you have horns." "If a thing moves, it moves either in the spot where it is or where it is not; but it neither moves in the spot where it is (for it is at rest) nor in that where it is not (for how could a thing be active in a spot where it does not so much as exist?); therefore nothing moves." "Either the existent becomes or the non-existent; now the existent does not become (for it exists); nor yet does the non-existent (for the becoming is passive but the non-existent is not passive); therefore nothing becomes." "Snow is frozen water; but water is black; therefore snow is black."
>
> And when he has made a collection of such trash he draws his eyebrows together, and expounds Dialectic and endeavours very solemnly to establish for us by syllogistic proofs that a thing becomes, a thing moves, snow is white, and we do not have horns, although it is probably sufficient to confront the trash with the plain facts, smashing up their positive affirmation by means of equally weighty contradictory evidence derived from the appearances. (*PH* 2.241-44, Bury, revised, cf. Timon's attitude reported in D.L. 9.111, 2.107)

Appealing to a precedent which was set by early Pyrrhonism, later Pyrrhonians propose that we replace philosophical attempts to establish what is true with an acceptance of appearances which provides a basis for ordinary actions and skeptical assertions. As Diogenes Laertius writes:

> Aenesidemus too in the first book of his *Pyrrhonian Arguments* says that Pyrrho determines nothing dogmatically because of the existence of contradictory arguments, but rather follows appearances. He says the same thing in *Against Wisdom* and *On Investigation*. And Zeuxis, an associate of Aenesidemus, in *On Twofold Arguments* and Antiochus of Laodicea and Apellas in his *Agrippa* posit the phenomena alone. Therefore, according to the skeptics, the appearance is a criterion, as Aenesidemus too says. (D.L. 9.106, Inwood & Gerson)

According to Sextus, "when we question whether the external object is such as it appears, we grant that it does appear, and we are not raising a question about the appearance but rather about what is said about the appearance; this is different from raising a question about the appearance itself. For example, the honey appears to us to be sweet. This we grant, for we sense the sweetness... And even when we do present arguments in oppostion to the appearances, we do not put these forward with the intention of

denying the appearances but by way of pointing out the precipitancy of the Dogmatists..." (*PH* 1.19, Mates).

The later Pyrrhonian commitment to appearances is consolidated in a "Practical Criterion" which was used to establish a "standard of action" which allows the Pyrrhonian to "perform some actions and abstain from others" while not adopting any beliefs in support of so choosing and acting.

> Holding to the appearances, then, we live without beliefs but in accord with the ordinary regimen of life, since we cannot be wholly inactive. And this regimen of life seems to be fourfold: one part has to do with the guidance of nature (*physis*), another with the compulsion of the *pathe* [feelings, affections of the soul], another with the handing down of laws and customs, and a fourth with instruction in arts and crafts (*techne*). Nature's guidance is that by which we are naturally capable of sensation and thought; compulsion of the *pathe* is that by which hunger drives us to food and thirst makes us drink; the handing down of customs and laws is that by which we accept that piety in the conduct of life is good and impiety bad; and instruction in arts and crafts is that by which we are not inactive in whichever of these we acquire. (*PH* 1.23-4, Mates)

Like the early Pyrrhonians, the later Pyrrhonians claimed that skeptical arguments and the Pyrrhonian acceptance of appearances could provide the basis for a happy life characterized by peace of mind. As Diogenes Laertius puts it, "The skeptics say the goal is suspension of judgement, upon which freedom from anxiety follows like a shadow, as Timon and Aenesidemus and their followers put it." (D.L. 9.107, Inwood & Gerson, cf. *PH* 1.29). According to Sextus, the *telos* of skepticism is tranquillity of mind (*ataraxia*) and "moderate" feeling "in respect of things unavoidable." (*PH* 1.26)

> We do not... take Sceptics to be undisturbed in every way -- we say that they are disturbed by things which are forced upon them; for we agree that at times they shiver and are thirsty and have other feelings of this kind. But in these cases ordinary people are afflicted by two sets of circumstances: by the feelings themselves, and no less by believing that these circumstances are bad by nature. Sceptics, who shed the additional opinion that each of these things is bad in its nature, come off more moderately even in these cases. (*PH* 1.29-30, Annas & Barnes)

Mates has criticized this aspect of Pyrrhonism, writing that "It is hard to find much plausibility in the general claim that the person who, on a given occasion, thinks ‘this *appears* to me to be very, very bad’ will be any less upset than if he thought ‘this *is* very, very bad’" (63). One might answer that the Pyrrhonian acceptance of appearances is more constrained than this suggests, for it takes place within the context of equally convincing arguments for and against the view that things are as they appear (the equal force of opposing arguments -- *isostheneia* -- thus plays a central role in Pyrrhonian thinking). In this way the Pyrrhonians purposely try to eliminate thoughts like "This appears very, very bad," trying to substitute in their place thoughts like "This appears bad, but I have equally convincing reasons for thinking it may not be so" (this is clearly manifest in Sextus' rejection of Carneadean plausibility). It is hard to say whether this suffices but the qualifications which Pyrrhonism thus introduces do in this way provide a

more substantial psychological basis for the detached and distant "following" of appearances which is supposed to nurture Pyrrhonian equanimity.

Given the practical goals of Pyrrhonism, the psychological force of Pyrrhonian arguments is in some ways as important as their logical force, for it functioned as a way to constrain the extent of the Pyrrhonian's conviction when he followed his appearances. This highlights an important difference between ancient and modern skeptical arguments. For though the former were employed as logical devices that establish epistemological conclusions, they were also used as psychological tools which were designed to break down attachment to belief and in this way foster *ataraxia*. In explaining why the skeptic's collection of arguments includes some which are weak, Sextus therefore says that the skeptic uses arguments of different strengths "just as doctors have remedies of different strengths for bodily ailments and for those suffering excessively employs the strong ones and for those suffering mildly the mild ones" (*PH* 3.280, Inwood & Gerson).

The Logic of Ancient Skepticism

How radical is ancient skepticism?

Though Sextus makes much of the skeptic's open-minded attitude to the possibility of apprehending truth, it is clear that the arguments that he and other skeptics employed can be used to raise questions about any claim to have established certain truth. At one point Sextus says that the skeptic will not, for example, assent even if he can find no fault with a position. For "[W]hen someone propounds to us a theory which we are unable to refute, we say to him in reply ‘Just as, before the birth of the founder of the School to which you belong, the theory it holds was not as yet apparent as a sound theory... so likewise it is possible that the opposite theory to that which you now propound is... not yet apparent to us, so that we ought not as yet yield assent to this theory which at the moment seems to be valid.’" (*PH* 1.33-34, Bury)

The extent of the ancient skeptic's concerns is also evident in the modes of skepticism (and especially the later modes), which are universally applicable and can in principle be used to question all of our beliefs. Even the skeptics' constant appeals to other ancient philosophies allowed for far more radical doubt than we normally engage, for ancient philosophy contained many extreme points of view. An example is Gorgias' argument for the conclusions that nothing exists, that if it did we could not know so, and if we knew so we could not communicate it. In his work, Sextus takes a special interest in this argument (and preserves one version of our most important fragment) precisely because it raises radical doubts about all things. In a similar vein, we find him exploiting for skeptical ends the opinions of obscure thinkers like Xeniades of Corinth -- who, he says, maintained that every impression and opinion is false (*AM* 7.53, cf. 48: a disconcerting view but arguably no more so than the Protagorean view that every opinion is true).

Mates has underscored the radical nature of the questions that the ancient skeptics, emphasizing that Sextus will not even grant that we have coherent concepts of the external world, soul, body, sense impressions, etc. As he puts it in discussing the Sextus' attitude to the external world, "His own deep skepticism leaves him in a state of *epoche*, not only as to whether there are any such things as ‘external

objects,' but even as to whether these terms of the Dogmatists have any intelligible meaning at all." (55)

How relevant is ancient skepticism?

The fact that ancient skepticism raises radical questions about all opinions and beliefs does not prove that it is relevant to modern and contemporary philosophy. A positive answer to the question whether the questions raised by the skeptics remain relevant must instead be founded on a recognition that they raise doubts that are still taken seriously in mainstream philosophical inquiry.

In this regard it can be said that ancient skepticism contains many arguments which remain of central importance, even though these arguments are frequently obscured by foreign philosophical terminology and ancient ways of speaking. In view of this, many commentators have explored and demonstrated the significance of ancient skepticism in the context of modern philosophy (see, e.g., Popkin, Schmitt, Jardine, Groarke, Fosl). Though the ancient skeptics do not as clearly anticipate modern and contemporary responses to skeptical concerns (skepticism's apparent tie to liberal political concerns, for example), it can be said that they achieved a very clear understanding of the basic epistemological issues raised by the attempt to build a rational basis for belief. The problem of the criterion and the later modes in particular ask pointed questions about our ability to establish a basis for justified belief which still resonate with us today.

One answer to skepticism which appears unique to contemporary philosophy is the suggestion that it can in some way be linguistically dissolved. Wittgenstein, Putnam and many others thus argue that skeptical claims in some way violate the norms that govern meaningful language and in view of this can be rejected as nonsensical. In ancient times, Aristocles wrote that skepticism is inconsistent with the assumption that the skeptic understands language (Eus., *Prep. Ev.* 14.18) but there is no close analogue of this linguistic answer to skepticism within ancient thought. How serious an omission this is depends on whether one believes that attempts to undermine skepticism in this way are plausible or successful (for a negative assessment, see Mates, 68-85).

Is ancient skepticism consistent?

Arguably the most significant question which needs to be asked about skepticism is a recurring feature of the skeptical/anti-skeptical debate. It is the question whether ancient skepticism is consistent. Or is it untenable because inconsistent? More specifically, we might ask how the skeptic's suspension of judgment allows him to come to the *conclusion* that we have no certain knowledge, or that certain knowledge is unattainable. Judging by the extant evidence we have, the skeptics themselves answered that their views are consistent because they accept skepticism in some "undogmatic" way which does not contradict their rejection of claims to truth (see Frede) -- by endorsing appearances, the *eulogon*, the *pithanon*, the Pyrrhonian practical criterion, and so on.

This is not the place for a detailed discussion of such issues, but one cautionary comment is in order. It is in this regard important to remember that the ancient skeptical attack on truth assumes a particular

conception of truth. Burnyeat in particular emphasizes that this ancient conception is thoroughly "realist." It suggests that a claim is true if it corresponds to a real objective world that is not subjective, but exists, as we might now put it, from "a god's eye point of view." As Burnyeat writes:

> In the controversy between the skeptic and the dogmatists over whether any truth exists at all, the issue is whether any proposition of a class of propositions can be accepted as true of a real objective world as distinct from mere appearance. For "true" in these discussions means "true of a real objective world"; the true, if there is such a thing is what conforms with the real, an association traditional to the word *alethes* since the earliest period of Greek philosophy (cf. *AM* XI 221).
>
> Now clearly, if truth is restricted to matters pertaining to real existence, as contrasted with appearance, the same will apply [to related skeptical conceptions]... The notions involved, consistency and conflict, undecidability, *isostheneia*, *epoche*, *ataraxia*, since they are defined in terms of truth, will all relate, via truth to real existence rather than appearance. (Burnyeat, "Can the Sceptic Live His Scepticism," p. 121)

Burnyeat's point should play a central role in contemporary attempts to assess ancient skepticism, for it makes such skepticism an attack on realist truth which has affinities to modern and contemporary anti-realism. In the context of questions about consistency, it provides a possible answer to the charge that the skeptics were inconsistent. For in attempting to understand the skeptics, we must recognize that belief, at least in contemporary philosophical parlance, need not mean "accepting something as true" in the realist sense. It follows that the ancient skeptic's decision to suspend judgment on claims to (realist) truth in principle leaves room for anti-realist forms of belief and assent which are now commonplace in epistemological discussion. Rather than eschew all belief (i.e. belief in *our* sense), this suggests that the ancient skeptic rejects a particular kind of belief to which contemporary epistemology offers a variety of alternatives (founded on coherence accounts of truth, etc.). Unlike the contemporary anti-realist, the ancient skeptic retained a realist conception of "truth" and "belief" and therefore expressed his position as the rejection of belief and the adoption of a weaker following of appearances, subjective impressions, and so on. This difference notwithstanding, the move away from realist conceptions of belief is similar in both cases.

The extent to which the analogy between ancient skepticism and contemporary anti-realism can be carried is open to debate, but it is an important comparison, for it suggests both that skepticism is not fatally inconsistent (for it rejects realist truth and endorses an anti-realist conception of belief) and that the positive account of belief that it proposes is, like many of its arguments against claims to truth, relevant to modern and contemporary philosophical concerns.

Bibliography

- Annas, Julia & Jonathan Barnes. *The Modes of Scepticism: Ancient Texts and Modern Interpretations*. New York: Cambridge University Press, 1985.

- Barnes, Jonathan. *The Toils of Scepticism*. New York: Cambridge University Press, 1990.
- Burnyeat, Myles, ed. *The Skeptical Tradition*. Berkeley: University of California Press, 1983.
- Burnyeat, Myles. "Can the Sceptic Live His Scepticism?" In Schofield, et. al., 1981; Reprinted in Burnyeat, 1983; and Burnyeat & Frede, 1997.
- Burnyeat, Myles & Michael Frede, eds. *The Original Sceptics: A Controversy*. Indianapolis: Hackett Publishing Inc., 1998.
- Carroll, Lewis. "What the Tortoise Said to Achilles." Copi & Gould, *Readings on Logic*. New York: Macmillan, 1972.
- Caton, Hiram. *The Origin of Subjectivity: An Essay on Descartes*. New Haven: Yale University Press, 1973.
- Cicero. *Academica*. H. Rackham, tr. Cambridge: Harvard University Press, Loeb Classical Library, 1961.
- Couissin, P. "The Stoicism of the New Academy." In Burnyeat, ed.
- DeLacy, Phillip. "*Ou Mallon* and the Antecedents of Ancient Scepticism." *Phronesis* **3** (1958).
- Diogenes Laertius. *Lives of Eminent Philosophers*. 2 Vols. R.D. Hicks. Loeb Classical Library. Cambridge: Harvard University Press, 1925.
- Eusebius. *Preparation for the Gospel*. Edwin Hamilton Gifford. 2 vols. Grand Rapids: Baker Book House, 1981 (reprint of the 1903 Clarendon edition).
- Fogelin, R.J. *Pyrrhonian Reflections on Knowledge and Justification*. Oxford: Oxford University Press, 1994.
- Fosl, Peter S. "The Bibliographic Bases of Hume's Understanding of Sextus Empiricus and Pyrrhonism," *Journal of the History of Philosophy*. Vol. 36, No. 2, 1998.
- Flintoff, Everard. "Pyrrho and India." *Phronesis*. **25** (1980) 88-108.
- Frede, Michael. "The Sceptic's Beliefs." In Burnyeat & Frede, eds.
- Frede, Michael. "The Sceptic's Two Kinds of Assent." In Burnyeat & Frede, eds.
- Groarke, Leo. *Greek Scepticism: Anti-Realist Trends in Ancient Thought*. Montreal: McGill-Queen's University Press, 1990.
- Groarke, Leo. "Descartes' First Meditation: Something Old, Something New, Something Borrowed." *Journal of the History of Philosophy,* Vol 22, No. 2, 1984.
- Hallie, Philip P., Sanford G. Etheridge, Donald R. Morrison. *Sextus Empiricus: Selections from the Major Writings on Scepticism, Man & God*. Indianapolis: Hackett Publishing, Co., 1985.
- Hankinson, R.J. *The Sceptics*. New York: Routledge, 1995.
- Hiley, David R. *Philosophy In Question: Essays on a Pyrrhonian Theme*. Chicago: University of Chicago Press, 1988.
- Inwood, Brad & L.P. Gerson. *Hellenistic Philosophy*. Indianapolis: Hackett Publishing Co., 1988; 2nd ed. 1997.
- Jardine, Lisa. "Lorenzo Valla: Academic Skepticism and the New Humanist Dialectic." In Burnyeat, ed.
- Long, A.A. & D. N. Sedley. *The Hellenistic Philosophers*. 2 Vols. New York: Cambridge University Press, 1987.
- Mates, Benson. *The Skeptic Way: Sextus Empricus's Outlines of Pyrrhonism*. New York: Oxford University Press, 1996.
- Popkin, Richard H. *The High Road to Pyrrhonism*. Ed. by Richard A. Watson & James E. Force.

San Diego: Austin Hill Press, 1980.

- Popkin, Richard H. *The History of Scepticism: from Erasmus to Spinoza*. Berkeley: University of California Press, 1979.
- Putnam, Hilary. *Reason, Truth and History*. Cambridge: Cambridge University Press, 1981.
- Robin, Leon. *Pyrrhon et le Scepticisme Grec*. Paris: Presses Universitaires de France, 1944 (Reprinted by Garland Publishing, New York, 1980).
- Schmitt, Charles B. *Cicero Scepticus: A Study of the Influence of the 'Academica' in the Renaissance*. The Hague: Martinus Nijhoff, 1974.
- Schmitt, Charles B. "The Rediscovery of Ancient Skepticism in Modern Times." In Burnyeat, ed.
- Schofield, Malcolm, Myles Burnyeat, Jonathan Barnes. *Doubt and Dogmatism: Studies in Hellenistic Epistemology*. Oxford: Clarendon Press, 1980.
- Sextus Empiricus. 4 Vols. tr. by R. G. Bury. Loeb Classical Library. Cambridge: Harvard University Press, 1933-1949.
- Sextus Empiricus. *Against the Ethicists (Adversus Mathematicos XI)*. tr. with a commentary by Richard Bett. Oxford: Clarendon Press, 1997.
- Sextus Empiricus. *Against the Grammarians (Adversus Mathematicos I)*. tr. with a commentary by D.L. Blank. New York: Oxford, 1998.
- Sextus Empiricus. *Against the Musicians (Adversus Musicos)*. tr. with an introduction by Denise Davidson Greaves. Lincoln: University of Nebraska Press, 1986.
- Sextus Empiricus. *Outlines of Scepticism*. tr. Julia Annas & Jonathan Barnes. New York: Cambridge University Press, 1994.
- Stough, C.L. *Greek Scepticism: A Study in Epistemology* Berkeley: UCLA Press, 1969.
- Wittgenstein, Ludwig. *On Certainty*. G.E.M. Anscombe & G.H. von Wright, eds. Evanston: Harper & Row, 1969.

Other Internet Resources

- Perseus Project

Related Entries

appearance vs. reality | Plato | Sextus Empiricus | skepticism | Socrates | Stoicism | Wittgenstein, Ludwig

Leo Groarke
Wilfrid Laurier University
lgroarke@wlu.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 4, 1997
Content last modified: September 2, 1998

Stoicism

Stoicism was one of the new philosophical movements of the Hellenistic period. The name derives from the porch (*stoa poikilê*) in the Agora at Athens decorated with mural paintings, where the members of the school congregated, and their lectures were held. Unlike 'epicurean,' the sense of the English adjective 'stoical' is not utterly misleading with regard to its philosophical origins. The Stoics did, in fact, hold that emotions like fear or envy (or impassioned sexual attachments, or passionate love of anything whatsoever) either were, or arose from, false judgements and that the sage--a person who had attained moral and intellectual perfection--would not undergo them. The later Stoics of Roman Imperial times, Seneca and Epictetus, emphasise the doctrines (already central to the early Stoics' teachings) that the sage is utterly immune to misfortune and that virtue is sufficient for happiness. Our phrase 'stoic calm' perhaps encapsulates the general drift of these claims. It does not, however, hint at the even more radical ethical views which the Stoics defended, e.g. that only the sage is free while all others are slaves, or that all those who are morally vicious are equally so. Though it seems clear that some Stoics took a kind of perverse joy in advocating views which seem so at odds with common sense, they did not do so simply to shock. Stoic ethics achieves a certain plausibility within the context of their physical theory and psychology, and within the framework of Greek ethical theory as that was handed down to them from Plato and Aristotle. It seems that they were well aware of the mutually interdependent nature of their philosophical views, likening philosophy itself to a living animal in which logic is bones and sinews; ethics and physics, the flesh and the soul respectively (another version reverses this assignment, making ethics the soul). Their views in logic and physics are no less distinctive and interesting than those in ethics itself.

Sources of our information on the Stoics

Since the Stoics stress the systematic nature of their philosophy, the ideal way to evaluate the Stoics' distinctive ethical views would be to study them within the context of a full exposition of their philosophy. Here, however, we meet with the problem about the sources of our knowledge about Stoicism. We do not possess a single complete work by any of the first three heads of the Stoic school: the 'founder,' Zeno of Citium in Cyprus (344-262 BC), Cleanthes (d. 232 BC) or Chrysippus (d. ca. 206 BC). Chrysippus was particularly prolific, composing over 165 works, but we have only fragments of his works. The only complete works by Stoic philosophers that we possess are those by writers of Imperial times, Seneca (4 BC-65 AD), Epictetus (c. 55-135) and the Emperor Marcus Aurelius (121-180) and these works are principally focused on ethics. They tend to be long on moral exhortation but give only clues to the theoretical bases of the moral system. For detailed information about the Old Stoa (i.e. the first three heads of the school and their pupils and associates) we have to depend on either doxographies, like pseudo-Plutarch *Philosophers' Opinions on Nature*, Diogenes Laertius' *Lives of Eminent Philosophers* (3rd c. AD), and Stobaeus' *Excerpts* (5th c. AD)--and their sources Aetius (ca. 1st c. AD) and Arius Didymus (1st c. B.C.-AD)--or other philosophers (or Christian apologists) who discuss the Stoics for their own purposes. Nearly all of the latter group are hostile witnesses. Among them are the Aristotelian commentator Alexander of Aphrodisias (late 2nd c. AD) who criticises the Stoics in *On Mixture* and *On Fate*, among other works; the Platonist Plutarch of Chaeronea (1st-2nd c. AD) who authored works such as *On Stoic Self-Contradictions* and *Against the Stoics on Common Conceptions*; the medical writer Galen (2nd c. AD), whose outlook is roughly Platonist; the Pyrrhonian skeptic, Sextus Empiricus (2nd c. AD); Plotinus (3rd c. AD); the Christian bishops Eusebius (3rd-4th c. A.D.) and Nemesius (ca. 400 AD); and the sixth-century neoplatonist commentator on Aristotle, Simplicius. Another important source is Cicero (1st c. BC). Though his own philosophical position derives from that of his teacher Philo of Larissa and the New Academy, he is not without sympathy for what he sees as the high moral tone of Stoicism. In works like his *Academic Books*, *On the Nature of the Gods*, and *On Ends* he provides summaries in Latin, with critical discussion, of the views of the major Hellenistic schools of thought.

From these sources, scholars have attempted to piece together a picture of the content, and in some cases, the development of Stoic doctrine. In some areas, there is a fair bit of consensus about what the Stoics thought and we can even attach names to some particular innovations. However, in other areas the proper interpretation of our meagre evidence is hotly contested. Until recently, non-specialists have been largely excluded from the debate because many important sources were not translated into modern languages. Fragments of Stoic works and testimonia in their original Greek and Latin were collected into a three-volume set in 1903-5 by H. von Arnim, *Stoicorum Veterum Fragmenta*. In writings on the 'old' Stoics, fragments and testimonia are often referred to by von Arnim's volume numbers and text numeration; e.g. SVF I.345=Diogenes Laertius, *Lives* 4.40. In 1987, A. A. Long and David Sedley brought out *The Hellenistic Philosophers* (*LS*) which contains in its first volume English translations and commentary of many important texts on Stoics, Epicureans and Skeptics. Unless otherwise specifically noted, I refer in what follows to texts by or about Stoics using the author's name followed by Long and Sedley's notation for the text, e.g. 47G=section 47 of their work, text G (unless otherwise noted, I use their translation,

sometimes slightly altered).

Philosophy and Life

When considering the doctrines of the Stoics, it is important to remember that they think of philosophy not as an interesting pastime or even a particular body of knowledge, but as a way of life. They define philosophy as a kind of practice or exercise (*askêsis*) in the expertise concerning what is beneficial (Aetius, 26A). Once we come to know what we and the world around us are really like, and especially the nature of value, we will be utterly transformed. This soteriological element is common to their main competitors, the Epicureans, and perhaps helps to explain why both were eventually eclipsed by Christianity. The *Meditations* of Marcus Aurelius provide a fascinating picture of a would-be Stoic sage at work on himself. The book, also called *To Himself*, is the emperor's diary. In it, he not only reminds himself of the content of important Stoic teaching but also reproaches himself when he realises that he has failed to incorporate this teaching into his life in some particular instance. For the influence of Stoic philosophy on a life in our times, see Admiral James Stockdale's account of his use of the philosophy of Epictetus as a prisoner of war in Vietnam.

Physical Theory

An examination of Stoic ontology might profitably begin with a passage from Plato's *Sophist*. There (247d-e), Plato asks for a mark or indication of what is real or what has being. One answer which is mooted is that the capacity to act or be acted upon is the distinctive mark of real existence or ‘that which is.’ The Stoics accept this criterion and add the rider that only bodies can act or be acted upon. Thus, only bodies exist. However, they allow that there are other ways of being part of nature than by virtue of existing. Incorporeal things like time, place or sayables (*lekta*, see below) are ‘subsistent’ (*huphestos*, Galen 27G)--as are imaginary things like centaurs. Moreover, all existent things are particular. The Stoics call universals ‘figments of the mind’ and seem to offer a conceptualist treatment akin to Locke's, treating an apparent predication like "man is a rational, mortal animal" as the disguised conditional, "if something is a man, then it is a rational mortal animal" (Sextus Empiricus, 30I).

In accord with this ontology, the Stoics, like the Epicureans, make God material. But while the Epicureans think the gods are too busy being blessed and happy to be bothered with the governance of the universe, the Stoic God is immanent throughout the whole of creation and directs its development down to the smallest detail. God is identical with one of the two ungenerated and indestructible first principles (*archai*) of the universe. One principle is matter which they regard as utterly unqualified and inert. It is that which is acted upon. God is identified with an eternal reason (*logos*, Diog. Laert. 44B) or intelligent designing fire (Aetius, 46A) which structures matter in accordance with Its plan. This plan is enacted time and time again, beginning from a state in which all is fire, through the generation of the elements, to the creation of the world we are familiar with, and eventually back to fire in a cycle of endless recurrence. The designing fire of the conflagration is likened to a sperm which contains the principles or stories of all the things which will subsequently develop (Aristocles in Eusebius, 46G).

Under this guise, God is also called ‘fate.’ It is important to realise that the Stoic God does not craft its world in accordance with its plan from the outside, as the demiurge in Plato's *Timaeus* is described as doing. Rather, the history of the universe is determined by God's activity internal to it, shaping it with its differentiated characteristics. The biological conception of God as a kind of living heat or seed from which things grow seems to be fully intended.

The first thing to develop from the conflagration are the elements. Of the four elements, the Stoics identify two as active (fire and air) and two as passive (water and earth). The active elements, or at least the principles of hot and cold, combine to form breath or pneuma. Pneuma, in turn, is the ‘sustaining cause’ (*causa continens*, *synektikon aition*) of all existing bodies and guides the growth and development of animate bodies. What is a sustaining cause? The Stoics think that the universe is a plenum. Like Aristotle, they reject the existence of empty space or void (except that the universe as a whole is surrounded by it). Thus, one might reasonably ask, ‘What marks any one object off from others surrounding it?’ or, ‘What keeps an object from constantly falling apart as it rubs elbows with other things in the crowd?’ The answer is: pneuma. Pneuma, by its nature, has a simultaneous movement inward and outward which constitutes its inherent ‘tensility.’ (Perhaps this was suggested by the expansion and contraction associated with heat and cold.) Pneuma passes through all (other) bodies; in its outward motion it gives them the qualities that they have, and in its inward motion makes them unified objects (Nemesius, 47J). In this respect, pneuma plays something of the role of substantial form in Aristotle for this too makes the thing of which it is the form both ‘some this,’ i.e. an individual, and ‘what it is’ (*Metaph.* VII, 17). Because pneuma acts, it must be a body and it appears that the Stoics stressed the fact that its blending with matter is ‘through and through’ (Galen 47H, Alex. Aph. 48C). Perhaps as a result of this, they developed a theory of mixture which allowed for two bodies to be in the same place at the same time. It should be noted, however, that some scholars (e.g. Richard Sorabji) think that the claim that pneuma is blended through the totality of matter is a conclusion that the Stoics' critics adversely drew about what some of their statements committed them to. Perhaps instead they proposed merely that pneuma is the matter of a body at a different level of description.

Pneuma comes in gradations and endows the bodies which it pervades with different qualities as a result. The pneuma which sustains an inanimate object is called (*LS*) a ‘tenor’ (*hexis*, lit. a holding). Pneuma in plants is, in addition, (*LS*) physique (*phusis*, lit. ‘nature’). In animals, pneuma gets called also soul (*psychê*) and in rational animals pneuma is, besides, the commanding faculty (*hêgemonikon*) (Diog. Laert. 47O, Philo 47P)--that responsible for thinking, planning, deciding. The Stoics assign to ‘physique’ or ‘nature’ all the purely physiological life functions of a human animal (such as digestion, breathing, growth etc.)--self-movement from place to place is due to soul. Their account of the human soul (mind) is strongly monistic. Though they speak of the soul's faculties, these are parts of the commanding faculty associated with the physical sense organs (Aetius, 53H). Unlike the Platonic tri-partite soul, all impulses or desires are direct functions of the rational, commanding faculty. This strongly monistic conception of the human soul has serious implications for Stoic epistemology and ethics. In the first case, our impressions of sense are affections of the commanding faculty. In mature rational animals, these impressions are thoughts, or representations with propositional content. Though a person may have no choice about whether she has a particular rational impression, there is another power of the commanding faculty which the Stoics call ‘assent’ and whether one assents to a rational impression is a matter of

volition. To assent to an impression is to take its content as true. To withhold assent is to suspend judgement about whether it is true. Because both impression and assent are part of one and the same commanding faculty, there can be no conflict between separate and distinct rational and nonrational elements within oneself--a fight which reason might lose. Compare this situation with Plato's description of the conflict between the inferior soul within us which is taken in by sensory illusions and the calculating part which is not (*Rep.* X, 602e). There is no reason to think that the calculating part can always win the epistemological civil war which Plato imagines to take place within us. But because the impression and assent are both aspects of one and the same commanding faculty according to the Stoics, they think that we can always avoid falling into error if only our reason is sufficiently disciplined. In a similar fashion, impulses or desires are movements of the soul toward something. In a rational creature, these are exercises of the rational faculty which do not arise without assent. Thus, a movement of the soul toward X is not automatically consequent upon the impression that X is desirable. This is what the Stoics' opponents, the Academic Skeptics, argue against them is possible (Plutarch, 69A.) The Stoics, however, claim that there will be no impulse toward X--much less an action--unless one assents to the impression (Plutarch, 53S). The upshot of this is that all desires are not only (at least potentially) *under the control* of reason, they *are* acts of reason. Thus there could be no gap between forming the decisive judgement that one ought to do X and an effective impulse to do X.

Since pneuma is a body, there is a sense in which the Stoics have a materialist theory of mind. The pneuma which is a person's soul is subject to generation and destruction (Plutarch 53 C, Eusebius 53W). Unlike for the Epicureans, however, it does not follow from this that my soul will be destroyed at the time at which my body dies. Chrysippus alleged that the souls of the wise would not perish until the next conflagration (Diog. Laert. 7.157=SVF 2.811, not in *LS*). Is this simply a failure of nerve on the part of an otherwise thorough-going materialist? Recall that the distinctive movement of pneuma is its simultaneous inward and outward motion. It is this which makes it tensile and capable of preserving, organising and, in some cases, animating the bodies which it interpenetrates. The Stoics equate virtue with wisdom and both with a kind of firmness or tensile strength within the commanding faculty of the soul (Arius Didymus 41H, Plutarch 61B, Galen 65T). Perhaps the thought was that the souls of the wise had a sufficient tensile strength that they could subsist as a distinct body on their own. Later Stoics like Panaetius (2nd c. BC) and Posidonius (first half 1st c. BC) may have abandoned this view of Chrysippus'.

Logic

For the Stoics, the scope of what they called ‘logic’ (*logikê*, i.e. knowledge of the functions of *logos* or reason) is very wide, including not only the analysis of argument forms, but also rhetoric, grammar, the theories of concepts, propositions, perception, and thought, and what we would call epistemology and philosophy of language. Formally, it was standardly divided into just two parts: rhetoric and dialectic (Diog. Laert., 31A). Much has been written about the Stoics' advances in logic (in our narrow sense of the word). In general, one may say that theirs is a logic of propositions rather than a logic of terms, like the Aristotelian syllogistic. One of the accounts they offer of validity is that an argument is valid if, through the use of certain ground rules (*themata*), it is possible to reduce it to one of the five

indemonstrable forms (Diog. Laert., 36A). These five indemonstrables are the familiar forms:

- if p then q; p; therefore q (modus ponens);
- if p then q; not q; therefore not-p (modus tollens);
- it is not the case that both p and q; p; therefore not-q;
- either p or q; p; therefore not-q;
- either p or q; not p; therefore q

Though these and other developments in logic are interesting in their own right, the Stoic treatment of certain problems about modality and bivalence are more significant for the shape of Stoicism as a whole. Chrysippus in particular was convinced that bivalence and the law of excluded middle apply even to contingent statements about particular future events or states of affairs. (The law of excluded middle says that for a proposition, p, and its contradictory, not-p, '(p or not-p)' is necessarily true, while bivalence insists that the truth table that defines a connective like 'or' contains only two values, true and false.) Aristotle's discussion in chapter 9 of *On Interpretation* of a hypothetical sea battle which either will or will not happen tomorrow has traditionally been taken to deny this. (The proper interpretation of Aristotle's position is disputed.) He presents the argument that if it is either true or false now that there will be a sea battle tomorrow (and let us suppose for the sake of argument that it is false), then our present deliberation about whether we should go out and fight tomorrow is pointless for it is already true now, whatever we decide, that we won't fight. Perhaps there are causal factors at work which will determine this, e.g. we may decide to fight but today's high temperatures will cause the wind to be against us tomorrow. On one reading, Aristotle's response to this is to deny the principle of bivalence for future contingent statements: it is now neither true nor false that there will be a sea battle tomorrow. Chrysippus apparently could not agree to making such an exception and he may have taken the price of consistency to be a strict causal determinism: all things happen through antecedent causes (Cicero, 38G). Above I noted that the Stoics thought that God or designing fire contained within itself the plan of all that is to happen between conflagrations and that it brings this plan to fruition in its action upon matter. Viewed in isolation from Stoic logic, this might have seemed arbitrary but clearly it was not.

The Stoics express their commitment to causal determinism in a potentially misleading way. They treat the claim that "all things happen through antecedent causes" as an alternative formulation of the claim that "all things happen through fate" (*kath heimarmenên*). But, in fact, the Stoics do not accept the doctrine that modern philosophers call fatalism. The matter is doubly confused, because the modern arguments for fatalism often emerge from the very considerations about bivalence that Aristotle discusses in *On Interpretation*. The classic example is Richard Taylor's argument. One way to see the difference between Taylor's fatalism and Chrysippus' causal determinism, is to ask, "What makes it the case that we won't have a sea battle tomorrow?" The Chrysippean causal determinist can say, "the lack of wind" or perhaps even "our decision not to go out and fight" and these things could all have been different, if only things had been different at some earlier time. So, though the present state of affairs determines that the future will only be one way, nonetheless there is a sense in which other things are possible (Alex. Aph., 38H). The fatalist responds that what makes it the case that we will not fight tomorrow is the fact that the proposition S, "There will not be a sea battle on such and such a date," has always been true. Much turns on what one says about the modal status of this truth. Is the proposition "It

is true that not-S" itself necessary? Diodorus Cronus, against whom Chrysippus argued, claimed that (1) truths about the past are necessary: it is not merely that they aren't other than they are--they can't be other than they are, for nothing has the power to change the past (Epictetus, 38A). He also claimed that (2) nothing impossible follows from what is possible. In the so-called Master Argument, he attempted to show that these two theses were incompatible with the claim that (3) there is something which is possible, but yet does not happen. The details of the Master Argument are a matter of much dispute. We know that it was alleged to show that these three propositions formed an inconsistent triad, but exactly how it did this remains uncertain. We also know that Diodorus' manner of resolving this inconsistency was to reject (3) and to define the possible as that which is or will be the case. Now consider our sea battle which will not take place tomorrow. If "there is a sea battle on such and such a date" is now false and will not be true, then by Diodorus' lights, it is impossible (Boethius, 38C)! Chrysippus felt the need to preserve the thesis that there are things which are possible but which do not happen. To this end, he rejected the proposition (2) that what is impossible does not follow from what is possible, using the following example: consider the conditional "if Dion is dead, then this one is dead" when ostensive reference is being made to Dion. The antecedent is possible, since Dion will one day be dead. Hence, let us suppose it true. Then, by modus ponens, it follows that "this one is dead." However, the proposition that "this one is dead" is impossible (necessarily false), since one cannot make the requisite ostensive reference to a dead man so as to make it true that "this one [i.e. the (living) thing I'm pointing to] is dead," for a dead person isn't the same thing as what was there previously (Alex. Aph., 38F). This may appear utterly *ad hoc*, and it is possibly wrong, given the Stoics' views about ‘sayables’ (*lekta*); but it is exactly the response that Chrysippus should make. It once again illustrates the systematic character of Stoic philosophy.

With respect to language, the Stoics distinguish between the signification, the signifier and the name-bearer. Two of these are bodies: the signifier which is the utterance and the name-bearer which gets signified. The signification, however, is an incorporeal thing called a *lekton*, or ‘sayable,’ and it, and neither of the other two, is what is true or false (Sextus Empiricus, 33B). They define a sayable as "that which subsists in accordance with a rational impression." Rational impressions are those alterations of the commanding faculty whose content can be exhibited in language. Presumably ‘*graphei Sôkratês*’ and ‘Socrates writes’ exhibit the contents of one and the same rational impression in different languages. At first glance, this looks very like a modern theory of propositions. But propositions (*axiômata*) are only one subspecies of sayables. Sayables also include questions and commands on the one hand, and, in a category of sayables called ‘incomplete,’ the Stoics include predicates and incomplete expressions like ‘*graphei*’ (he or she writes) (Diog. Laert., 33F). An incomplete sayable like ‘writes’ gets transformed into a proposition by being attached to a nominative case (*ptôsis*, Diog. Laert., 33G). Here a ‘nominative case’ seems to mean the *signification* of the inflected word, *Sôkratês*’ or ‘*ho anthrôpos*’--the latter being the nominative case (as *we* would say) of the Greek *word* ‘man’--not that inflected word itself. The Stoic doctrine of case is one of those areas where there is as yet little consensus. Stoic propositions are unlike propositions in contemporary theories in another way too: Stoic sayables are not timelessly true or false. If it is now daytime, the *lekton* corresponding to an utterance of ‘it is day’ is true. Tonight, however, it will be false (cf. Alex. Aph. in Simplicius, 37K). Finally, the Stoic theory gives a certain kind of priority to propositions involving demonstratives. ‘This one is writing’ is definite, while ‘someone is writing’ is indefinite. Strikingly, ‘Socrates is writing’ is said to be intermediate between these two. When there is a

failure of reference, the Stoics say that the *lekton* is destroyed and this is supposed to provide the reason why ‘this one is dead’ (spoken in relation to Dion) is impossible (necessarily false).

Perhaps the most famous topic considered under the Stoic heading of logic is that of the criterion of truth and the Stoics' disputes with the skeptical New Academy about it. According to Chrysippus, the criterion of truth is the ‘cognitive impression’ (*phantasia katalêptikê*, lit. an impression that firmly grasps its object) (Diog. Laert., 40A). A criterion or canon of truth is an instrument for definitely determining that something is true, and the Hellenistic schools all provide some view on how it is that we are to measure or evaluate whether something is true or not. The Stoics' cognitive impression is an impression which (according to Zeno's definition, cf. Cicero, SVF I.59) "arises from that which is; is stamped and impressed in accordance with that very thing; and of such a kind as could not arise from what is not" (Sextus Empiricus, 40E). Recall that among the powers of the commanding faculty is the capacity to assent or withhold assent to impressions. The fact that it is always within our power to withhold assent means that if we are sufficiently disciplined, we are capable of avoiding error. In itself, it does not mean that we are capable of attaining knowledge, for there might not be any impressions that one can be confident in assenting to. The cognitive impression was supposed to fill that role: when you experience one of these, provided that you recognize it as such, you can, on its basis, assert definitely that the matter in question is true. It was initially supposed that such an impression commanded one's assent by its very nature: it "all but seizes us by the hair" and drags us to assent. But this optimistic assessment seems to have been qualified in the face of criticism by members of the Skeptical Academy--perhaps, even if there are such impressions, it is not so easy to be sure when one is experiencing one.

However, the Stoics do not maintain that the mere having of a cognitive impression constitutes knowledge (*epistêmê*). Indeed, not even assent to such an impression amounts to knowledge: this is only cognition or grasp (*katalêpsis*) of some individual fact. Real knowledge (*epistêmê*) requires cognition which is secure, firm and unchangeable by reason (Sextus Empiricus, 41C)--and, furthermore, worked into a systematic whole with other such cognitions (Arius Didymus, 41H). Weak and changeable assent to a cognitive impression is only an act of ignorance. It is not entirely clear where opinion or belief in general (*doxa*) stands in this categorization. Most Stoic sources define it as ‘assent to the incognitive’ (i.e. to an impression that does not firmly grasp its object) (see Sextus Empiricus, 41E) but some suggest that changeable assent to a cognitive impression might still count as opinion. There is a potential for serious confusion when we try to assimilate the Stoic view to contemporary epistemology. Modern definitions of knowledge make the agent's belief that P a necessary but not sufficient condition for knowing that P. For the Stoics, *doxa* (involving ‘weak’ assent) and knowledge are incompatible. In any event, there is an absolute distinction between the wise and the ignorant. Only the Stoic sage's assent to cognitive impressions clearly counts as knowledge for only a sage has the proper discipline always to avoid withdrawing assent, or assenting to things that one shouldn't. The Stoics call this epistemic virtue ‘non-precipitancy’ (*aproptôsia*) and it underlies their claim that the Stoic sage never makes mistakes (41D).

The Skeptics responded by denying the existence of cognitive impressions. According to Arcesilaus, "no impression arising from something true is such that an impression arising from something false could not also be just like it" (Cicero, 40D). So Arcesilaus denies that the third conjunct of the Stoic definition of

the cognitive impression is ever satisfied. We can distinguish two specific tactics for denying this. First, the Skeptics point to cases of insanity. In his madness, Heracles had the impression that his children were, in fact, the children of his enemy Eurystheus and killed them. Since the impression must have been utterly convincing to him at the time at which he had it (judging by his subsequent action), it is clear from this that there can be false impressions which are indistinguishable from ones that are allegedly "stamped and impressed in accordance with that very thing" (Sextus Empiricus, 41H). Their second line of attack was to draw attention to objects which are so similar as to be indistinguishable (so that a completely accurate impression from one would be indistinguishable from one from the other). The story is related (Diog. Laert., 40F) that the Stoic philosopher Sphaerus (a student of Zeno's) was tricked into thinking that wax pomegranates were real. This was again supposed to show thay there could be impressions arising from what is not [sc. a pomegranate] which are indistinguishable from a cognitive impression.

The Stoics met these arguments by first pointing out that Heracles' inability to distinguish cognitive from incognitive impressions in his madness says nothing about the capacities of normal human beings. It is no part of their thesis that *just anyone* can distinguish between cognitive and incognitive impressions. Their response to the second line of attack was two-fold. The first is a metaphysically motivated answer: if any two objects really were indistinguishable, they would be identical. This doctrine has come to be known as the identity of indiscernibles. They also replied that the Stoic sage would withhold assent in cases where things are too similar to be confident that one had it right (Cicero, 40I)--Sphaerus' response to his predicament was to say that he only assented to the proposition that it was ‘reasonable’ that what he was presented with were pomegranates (and that was true!).

In some ways, the Stoics have an easier time with Skeptical objections than contemporary non-skeptics do. At bottom what the Stoics are committed to is the two-fold view that it is within our power to avoid falling into error and that there is a kind of impression which reveals to us the world as it really is and which is different from those impressions which might not so reveal the world. They are manifestly *not* committed to defending our ordinary intuitions about the range of knowledge: that most people in fact know most of the things that they and everyone else thinks that they know. The only person we can be sure has any knowledge is the Stoic sage and sages are as rare as the phoenix (Alex. Aph., 61N). Everyone else is equally ignorant. This absolute distinction between the wise and the ignorant is a consequence of the Stoic definition of knowledge as the "cognition which is secure and unchangeable by reason" (Arius Didymus, 41H). Either one's cognition is like this or it is not. By making opinion a kind of ignorance (contrast Plato, *Rep* V. 474a ff), they do not allow room for an intermediate state between the wise man and all the rest of us.

But if we leave aside the question of whether we in fact know anything, there are some serious puzzles about the cognitive impression. The Stoics insist that the cognitive impression not only "arises from what is and is stamped and impressed in accordance" with the thing from which it arises, but also that it is "such as could not arise from that which is not." But it seems that we can imagine all kinds of situations in which we might be in a position where the sense impressions that we have are indistinguishable from ones that misrepresent the world. Thus, consider Descartes' evil demon hypothesis or its modern counterpart, the brain in a vat scenario. In the latter example it is stipulated that electrical stimulation of

your brain by incredibly clever but unscrupulous scientists produces sense impressions that are indistinguishable from the ones that you are presently having. Surely here we have a demonstration that there could not be a true impression which is such that it *could not* arise from what is not. No sane person thinks that these skeptical hypotheses are actually true. The point is rather that if one of them were true, our sense experience would be indistinguishable from what (we take to be) our true and accurate sense impressions of real tables, chairs and fireplaces. Doesn't this show that there is no such thing as a cognitive impression?

One thing to note in passing is that skeptical scenarios like the evil demon or the brain in the vat did not seem to figure in the debate between the Stoics and Skeptics. The Skeptics press the point that *at the time* the dream may be completely convincing to the dreamer, even if she does not believe that the events actually transpired when she awakes (Cicero, *Lucullus* or *Academica* II, 88). They do not consider thought experiments in which *all* our sense experience is systematically misleading. But if we set this aside, there will still be one important difference between a clear and distinct impression that arises from a real fireplace and one that arises from the manipulation of my neurons by unscrupulous brain scientists. The first is *caused by* a fireplace, while the second is caused by some other means. When the Stoics say that a cognitive impression is "of such a sort as could not arise from what is not," they can be interpreted to mean that the true clear and distinct impression will be *different* from a false one. They might deny that the difference between the two is always something that can be *discerned* from the subject's point of view. If this is so, then the Stoics' position would be somewhat akin to externalist theories of knowledge or justification. Externalists insist that an agent might know a proposition or be justified in believing a proposition even when, nonetheless, the evidence for that belief is not subjectively available to the person. So, on one early externalist theory of knowledge, it was suggested that an agent might know a certain sort of proposition (e.g. that there is a fireplace here) if their belief that there is a fireplace here was caused by a reliable causal process (e.g. a normal visual system)--and not, e.g., by the interventions of wicked scientists.

So where does this leave the matter? *If* this is the right way to understand the definition of the Stoic cognitive impression, then it would seem that they win their argument with the Skeptics. Examples of false impressions that are subjectively indiscernible from clear and distinct, true, ones do not show that there are no cognitive impressions. However, the admission that a cognitive impression might be subjectively indistinguishable from a false impression does alter the sense in which the cognitive impression can serve as a criterion of truth. Assent to a cognitive impression will guarantee that what you assent to is true. But, because cognitive impressions can be indistinguishable from the subject's point of view from false ones, the Stoics can no longer say that even the sage can be confident that what seems to be a cognitive impression actually is one. Thus instead of automatically commanding assent, the cognitive impression (according to later Stoics) commands assent "if there is no impediment" (Sextus Empiricus, 40K), and if it has been successfully "tested" and is "irreversible" (cf. Sextus Empiricus, 69E). This means that I should only assent to what seems to me to be a cognitive impression if I have reason to believe that I'm not in a context where deceptive but convincing impressions are possible. But the Stoic sage never errs. So when will I have absolutely compelling reasons to believe that I'm not presented with a convincing but deceptive impression? For these reasons, the Pyrrhonian skeptic Sextus Empiricus argues that the Stoic sage will never assent to any impression. In practice, he will suspend

judgement, just like the Skeptic does (41C). Another suggestion is that the Stoic sage hedges his bets by assenting only to the impression that it is *reasonable* that there is fireplace here (as Sphaerus did about the pomegranates, 40F). In this case it will also be hard to see how he differs from a skeptic who takes ‘the reasonable’ as his criterion (Sextus Empiricus, 69B).

Ethics

In many ways, Aristotle's ethics provides the form for the adumbration of the ethical teaching of the Hellenistic schools. One must first provide a specification of the goal or end (*telos*) of living. This may have been thought to provide something like the dust jacket blurb or course description for the competing philosophical systems--which differed radically over how to give the required specification.

A bit of reflection tells us that the goal that we all have is happiness or flourishing (*eudaimonia*). But what is happiness? The Epicureans' answer was deceptively straightforward: the happy life is the one which is most pleasant. (But their account of what the highest pleasure consists in was not at all straightforward.) Zeno's answer was "a good flow of life" (Arius Didymus, 63A) or "living in agreement," and Cleanthes clarified that with the formulation that the end was "living in agreement with nature" (Arius Didymus, 63B). Chrysippus amplified this to (among other formulations) "living in accordance with experience of what happens by nature;" later Stoics inadvisably, in response to Academic attacks, substituted such formulations as "the rational selection of the primary things according to nature." The Stoics' specification of what happiness consists in cannot be adequately understood apart from their views about value and human psychology.

The best way into the thicket of Stoic ethics is through the question of what is good, for all parties agree that possession of what is genuinely good secures a person's happiness. The Stoics claim that whatever is good must benefit its possessor under all circumstances. But there are situations in which it is not to my benefit to be healthy or wealthy. (We may imagine that if I had money I would spend it on heroin which would not benefit me.) Thus, things like money are simply not good, in spite of how nearly everyone speaks, and the Stoics call them ‘indifferents’ (Diog. Laert., 58A)--i.e., neither good nor bad. The only things that are good are the characteristic excellences or virtues of human beings (or of human minds): prudence or wisdom, justice, courage and moderation, and other related qualities. These are the first two of the ‘Stoic paradoxes’ discussed by Cicero in his short work of that title: that only what is noble or fine or morally good (*kalon*) is good at all, and that the possession (and exercise) of the virtues is both necessary and sufficient for happiness. But the Stoics are not such lovers of paradox that they are willing to say that my preference for wealth over poverty in most circumstances is utterly groundless. They draw a distinction between what is good and things which have value (*axia*). Some indifferent things, like health or wealth, have value and therefore are to be preferred, even if they are not good, because they are *typically* appropriate, fitting or suitable (*oikeion*) for us.

Impulse, as noted above, is a movement of the soul toward an object. Though these movements are subject to the capacity for assent in fully rational creatures, impulse is present in all animate (self-moving) things from the moment of birth. The Stoics argue that the original impulse of ensouled

creatures is toward what is appropriate for them, or aids in their self-preservation, and not toward what is pleasurable, as the Epicureans contend. Because the whole of the world is identical with the fully rational creature which is God, each part of it is naturally constituted so that it seeks what is appropriate or suitable to it, just as our own body parts are so constituted as to preserve both themselves and the whole of which they are parts. The Stoic doctrine of the natural attachment to what is appropriate (*oikeiôsis*) thus provides a foundation in nature for an objective ordering of preferences, at least on a prima facie basis. Other things being equal, it is objectively preferable to have health rather than sickness. The Stoics call things whose preferability is overridden only in very rare circumstances "things according to nature." As we mature, we discover new things which are according to our natures. As infants perhaps we only recognised that food and warmth are appropriate to us, but since humans are rational, more than these basic necessities are appropriate to us. The Greek term ‘*oikeion*’ can mean not only what is suitable, but also what is akin to oneself, standing in a natural relation of affection. Thus, my blood relatives are--or least ought to be--*oikeioi*. It is partly in this sense that we eventually come to the recognition--or at least ought to--that other people, insofar as they are rational, are appropriate to us. Cicero's quotation of Terence's line ‘nothing human is alien to me’ in the context of *On Duties* I.30 echoes this thought. More generally, the unfolding of God's providential plan is rational (and therefore beneficial) through and through, so that in some sense what will in fact happen to me in accordance with that plan must be appropriate to me.

When we take the rationality of the world order into consideration, we can begin to understand the Stoic formulations of the goal or end. "Living in agreement with nature" is meant to work at a variety of levels. Since *my* nature is such that health and wealth are appropriate to me (according to my nature), other things being equal, I ought to choose them. Hence the formulations of the end by later Stoics stress the idea that happiness consists in the rational selection of the things according to nature. But, we must bear in mind an important caveat here. Health and wealth are not the only things which are appropriate to me. So are other rational beings and it would be irrational to choose one thing which is appropriate to me without due consideration of the effect of that choice on other things which are also appropriate to me. This is why the later formulations stress that happiness consists in the *rational* selection of the things according to nature. But if I am faced with a choice between increasing my wealth (something which is prima facie appropriate to my nature) and preserving someone else's health (which is something appropriate to something which is appropriate to me, i.e. another rational being), which course of action is the rational one? The Stoic response is that it is the one which is ultimately both natural and rational: that is, the one that, so far as I can tell from my experience with what happens in the course of nature (see Chrysippus' formula for the end cited above, 63B), is most in agreement with the unfolding of nature's rational and providential plan. Living in agreement with nature in this sense can even demand that I select things which are not typically appropriate to my nature at all--when that nature is considered in isolation from these particular circumstances. Here Chrysippus' remark about what his foot would will if it were conscious is apposite.

> As long as the future is uncertain to me I always hold to those things which are better adapted to obtaining the things in accordance with nature; for God himself has made me disposed to select these. But if I actually knew that I was fated now to be ill, I would even have an impulse to be ill. For my foot too, if it had intelligence, would have an impulse to

get muddy. (Epictetus, 58J)

We too, as rational parts of rational nature, ought to choose in accordance with what will in fact happen (*provided* we can know what that will be, which we rarely can--we are not gods; outcomes are uncertain to us) since this is wholly good and rational: when we cannot know the outcome, we ought to choose in accordance with what is typically or usually nature's purpose, as we can see from experience of what usually does happen in the course of nature. In extreme circumstances, however, a choice, for example, to end our lives by suicide can be in agreement with nature.

So far the emphasis has been on just one component of the Stoic formulation of the goal or end of life: it is the "rational selection of the things according to nature." The other thing that needs to be stressed is that it is rational *selection*--not the attainment of--these things which constitutes happiness. (The Stoics mark the distinction between the way we ought to opt for health as opposed to virtue by saying that I select (*eklegomai*) the preferred indifferent but I choose (*hairoûmai*) the virtuous action.) Even though the things according to nature have a kind of value (*axia*) which grounds the rationality of preferring them (other things being equal), this kind of value is still not goodness. From the point of view of happiness, the things according to nature are still indifferent. What matters for our happiness is whether we select them rationally and, as it turns out, this means selecting them in accordance with the virtuous way of regarding them (and virtuous action itself). Surely one motive for this is the rejection of even the limited role that external goods and fortune play in Aristotelian ethics. According to the Peripatetics, the happy life is one in which one exercises one's moral and theoretical virtues. But one can't exercise a moral virtue like liberality (*Nic. Eth.* IV.1) without having some, even considerable, money. The Stoics, by contrast, claim that so long as I order (and express) my preferences in accordance with my nature and universal nature, I will be virtuous and happy, even if I do not actually get the things I prefer. Though these things are typically appropriate to me, rational choice is even more appropriate or akin to me, and so long as I have that, then I have perfected my nature. The perfection of one's rational nature is the condition of being virtuous and it is exercising this, and this alone, which is good. Since possession of that which is good is sufficient for happiness, virtuous agents are happy even if they do not attain the preferred indifferents they select.

One is tempted to think that this is simply a misuse of the word ‘happiness’ (or would be, if the Stoics had been speaking English). We are inclined to think (and a Greek talking about *eudaimonia* would arguably be similarly inclined) that happiness has something to do with getting what you want and not merely ordering one's wants rationally regardless of whether they are satisfied. People are also frequently tempted to assimilate the Stoics' position to one (increasingly contested) interpretation of Kant's moral philosophy. On this reading, acting with the right motive is the only thing that is good--but being good in this sense has nothing whatsoever to do with happiness.

With respect to the first point, the Stoic sage typically selects the preferred indifferents and selects them in light of her knowledge of how the world works. There will be times when the circumstances make it rational for her to select something that is (generally speaking) contrary to her nature (e.g. cutting off one's own hand in order to thwart a tyrant). But these circumstances will be rare and the sage will not be oppressed by the additional false beliefs that this act of self-mutilation is a genuinely bad thing: only vice

is genuinely bad. For the most part, her knowledge of nature and other people will mean that she attains the things that she selects. Her conditional positive attitude toward them will mean that when circumstances do conspire to bring it about that the object of her selection is not secured, she doesn't care. She only preferred to be wealthy if it was fated for her to be wealthy. These reflections illustrate the way in which the virtuous person is self-sufficient (*autarkês*) and this seems to be an important component of our intuitive idea of happiness. The person who is genuinely happy lacks nothing and enjoys a kind of independence from the vagaries of fortune. To this extent at least, the Stoics are not just using the word ‘happiness’ for a condition that has *nothing at all* to do with what we typically mean by it. With respect to the second point, the Stoic sage will never find herself in a situation where she acts contrary to what Kant calls inclination or desire. The only thing she unconditionally wants is to live virtuously. Anything that she conditionally prefers is always subordinate to her conception of the genuine good. Thus, there is no room for a conflict between duty and happiness where the latter is thought of solely in terms of the satisfaction of our desires. Cicero provides an engaging, if not altogether rigorous, discussion of the question of whether virtue is sufficient for happiness in *Tusculan Disputations*, book V.

How do these general considerations about the goal of living translate into an evaluation of actions? When I perform an action that accords with my nature and for which a good reason can be given, then I perform what the Stoics call (LS) a ‘proper function’ (*kathêkon*, Arius Didymus, 59B)--something that it "falls to me" to do. It is important to note that non-rational animals and plants perform proper functions as well (Diog. Laert., 59C). This shows how much importance is placed upon the idea of what accords with one's nature or, in another formulation, "activity which is consequential upon a thing's nature." It also shows the gap between proper functions and morally right actions, for the Stoics, like most contemporary philosophers, think that animals cannot act morally or immorally--let alone plants.

Most proper functions are directed toward securing things which are appropriate to nature. Thus, if I take good care of my body, then this is a proper function. The Stoics divide proper functions into those which do not depend upon circumstances and those that do. Taking care of one's health is among the former, while mutilating oneself is among the latter (Diog. Laert., 59E). It appears that this is an attempt to work out a set of prima facie duties based upon our natures. Other things being equal, looking after one's health is a course of action which accords with one's nature and thus is one for which a good reason can be given. However, there are circumstances in which a better reason can be given for mutilating oneself--for instance, if this is the only way you can prevent Fagin from compelling you to steal for him.

Since both ordinary people and Stoic wise men look after their health except in very extraordinary circumstances, both the sage and the ordinary person perform proper functions. A proper function becomes a fully correct action (*katorthôma*) only when it is perfected as an action of the specific kind to which it belongs, and so is done virtuously. In the tradition of Socratic moral theory, the Stoics regard virtues like courage and justice, and so on, as knowledge or science within the soul about how to live. Thus a specific virtue like moderation is defined as "the science (*epistêmê*) of what is to be chosen and what is to be avoided and what is neither of these" (Arius Didymus, 61H). More broadly, virtue is "an expertise (*technê*) concerned with the whole of life" (Arius Didymus, 61G). Like other forms of knowledge, virtues are characters of the soul's commanding faculty which are firm and unchangeable. The other similarity with Socratic ethics is that the Stoics think that the virtues are really just one state of

soul (Plutarch , 61B, C; Arius Didymus, 61D). No one can be moderate without also being just, courageous and prudent as well--moreover, "anyone who does any action in accordance with one does so in accordance with them all" (Plutarch, 61F). When someone who has any virtue, and therefore all the virtues, performs any proper function, he performs it in accordance with virtue or virtuously (i.e. with all the virtues) and this transforms it into a right action or a perfect function. The connection here between a perfect function and a virtuous one is almost analytic in Greek ethical theorizing. Virtues just are those features which make a thing a good thing of its kind or allow it to perform its function well. So, actions done in accordance with virtue are actions which are done well. The Stoics draw the conclusion from this that the wise (and therefore virtuous) person does everything within the scope of moral action well (Arius Didymus, 61G). This makes it seem far less strange than it might at first appear to say that virtue is sufficient for happiness. Furthermore, because virtue is a kind of knowledge and there is no cognitive state between knowledge and ignorance, those who are not wise do everything equally badly. Strictly speaking, there is no such thing as moral progress for the Stoics (if that means progress *within* morality), and they give the charming illustration of drowning to make their point: a person an arm's length from the surface is drowning every bit as surely as one who is five hundred fathoms down (Plutarch, 61T). Of course, as the analogy also suggests, it is possible to be closer or farther *from* finally being able to perform proper functions in this perfected way. In that sense, progress is possible.

We are finally in a position to understand and evaluate the Stoic view on emotions, since it is a consequence of their views on the soul and the good. It is perhaps more accurate to call it the Stoic view of the passions, though this is a somewhat dated term. The passions or *pathê* are literally ‘things which one undergoes’ and are to be contrasted with actions or things that one does. Thus, the view that one should be ‘apathetic,’ in its original Hellenistic sense, is not the view that you shouldn't care about anything, but rather the view that you should not be psychologically subject to anything--manipulated and moved by *it*, rather than yourself being actively and positively in command of your reactions and responses to things as they occur or are in prospect. It connotes a kind of complete self-sufficiency. The Stoics distinguish two primary passions: appetite and fear. These arise in relation to what appears to us to be good or bad. They are associated with two other passions: pleasure and distress. These result when we get or fail to avoid the objects of the first two passions. What distinguishes these states of soul from normal impulses is that they are "excessive impulses which are disobedient to reason" (Arius Didymus, 65A). Part of what this means is that one's fear of dogs may not go away with the rational recognition that this blind, 16 year old, 3 legged Yorkshire terrier poses no threat to you. But this is not all. The Stoics call a passion like distress a *fresh* opinion that something bad is present (Andronicus, 65B): you may have been excitedly delighted when you first saw you'd won the race, but after a while, when the impression of the victory is no longer fresh, you may calm down. Recall that opinion is assent to a false impression. Given the Stoics' view about good and bad, as against merely indifferent things, the only time that one should assent to the impression that something bad is present is when there is something which might threaten one's virtue, for this and this alone is good. Thus all passions involve an element of false value-judgement. But these are false judgements which are inseparable from physiological changes in the pneuma which constitutes one's commanding faculty. The Stoics describe these changes as shrinkings (like fear) or swellings (like delight), and part of the reason that they locate the commanding faculty in the heart (rather than the head, as Plato in the *Timaeus* and many medical writers did) is that this seems to be where the physical sensations which accompany passions like fear are manifested.

Taking note of this point of physiology is surely necessary to give their theory any plausibility. From the inside a value-judgement--even one like "this impending dog bite will be bad"--might often just not feel like such an emotional state as fear. But when the judgement is vivid and so the commanding faculty is undergoing such a change, one can readily enough see that the characteristic sensations might inexorably accompany the judgement.

Another obvious objection to the Stoic theory is that someone who fears, say pigeons, may not *think* that they are dangerous. We say that she knows rationally that pigeons are harmless but that she has an irrational fear. It might be thought that in such a case, the judgement which the Stoics think is essential to the passion is missing. Here they resort to the idea that a passion is a fluttering of the commanding faculty. At one instant my commanding faculty judges (rightly) that this pigeon is not dangerous, but an instant later assents to the impression that it is and from this assent flows the excessive impulse away from the pigeon which is my fear. This switch of assent occurs repeatedly and rapidly so that it appears that one has the fear without the requisite judgement but in fact you are making it and taking it back during the time you undergo the passion (Plutarch, 65G).

It is important to bear in mind that the Stoics do not think that all impulses are to be done away with. What distinguishes normal impulses or desires from passions is the idea that the latter are excessive and irrational. Galen provides a nice illustration of the difference (65J). Suppose I want to run, or, in Stoic terminology, I have an impulse to run. If I go running down a sharp incline I may be unable to stop or change direction in response to a new impulse. My running is excessive in relation to my initial impulse. Passions are distinguished from normal impulses in much the same way: they have a kind of momentum which carries one beyond the dictates of reason. If, for instance, you are consumed with lust (a passion falling under appetite), you might not do what under other circumstances you yourself would judge to be the sensible thing.

Even in antiquity the Stoics were ridiculed for their views on the passions. Some critics called them the men of stone. But this is not entirely fair, for the Stoics allow that the sage will experience what they call the good feelings (*eupatheiai*, Diog. Laert. 65F). These include joy, watchfulness and wishing and are distinguished from their negative counterparts (pleasure, fear and appetite) in being well-reasoned and not excessive. Naturally there is no positive counterpart to distress. The species under wishing include kindness, generosity and warmth. A good feeling like kindness is a moderate and reasonable stretching or expansion of the soul presumably prompted by the correct judgement that other rational beings are appropriate to oneself.

Criticisms of the Stoic theory of the passions in antiquity focused on two issues. The first was whether the passions were, in fact, activities of the rational soul. The medical writer and philosopher Galen defended the Platonic account of emotions as a product of an irrational part of the soul. Posidonius, a 1st c. BC Stoic, also criticised Chrysippus on the psychology of emotions, and developed a position that recognized the influence in the mind of something like Plato's irrational soul-parts. The other opposition to the Stoic doctrine came from philosophers in the Aristotelian tradition. They, like the Stoics, made judgement a component in emotions. But they argued that the happy life required the *moderation* of the passions, not their complete extinction. Cicero's *Tusculan Disputations*, books III and IV take up the

question of whether it is possible and desirable to rid oneself of the emotions.

Influence

The influence of Stoicism on Greek and Roman culture was enormous. Zeno, the first head of the school, had a statue raised to him in Athens at public expense. The inscription read, in part:

> Whereas Zeno of Citium, son of Mnaseas, has for many years been devoted to philosophy in the city and has continued to be a man of worth in all other respects, exhorting to virtue and temperance those of the youth who came to him to be taught, directing them to what is best, affording to all in his own conduct a pattern for imitation in perfect consistency with his teaching ... (Diog. Laert. 7.10-11, tr. Hicks)

Of course the citizens of Athens couldn't have honoured Zeno for a life lived in consistency with his philosophical principles unless the content of those principles was known to the general public. Since the Stoics gathered, discussed and taught philosophy in a public place, the general import of their philosophy was widely known. Stoicism became a "popular philosophy" in a way that neither Platonism nor Aristotelianism ever did. In part this is because Stoicism, like its rival Epicureanism, addressed the questions that most people are concerned with in very direct and practical ways. It tells you how you should regard death, suffering, great wealth, poverty, power over others and slavery. In the political and social context of the Hellenistic period (where a person could move between these extremes in very short order) Stoicism provided a psychological fortress which was secure from bad fortune. Historians of philosophy earlier in this century regarded this as a mark against Hellenistic philosophy generally. The notion was that philosophy peaked with Plato and Aristotle and then degenerated into the popular "feel good" philosophy of the Hellenistic period and did not approach its earlier glory again until Plotinus. It may be true that the lack of political autonomy in the Greek city states made the ideal of the self-sufficient Stoic sage appear more relevant and desirable. But even if the philosophy suited the times, the Stoics and Epicureans provided arguments for their view which still have interest for us who live in a social and political context that is quite different. (Or perhaps not so different. I suppose it depends on how pessimistic you are about the possibilities for self-determination for individuals and small communities in the age of globalization.)

At the political level, the Antigonid dynasty (which ruled Greece and Macedon from shortly after the death of Alexander to 168 BC) had connections with the Stoic philosophers. Antigonus Gonatas was alleged to have been a pupil of Zeno of Citium. He requested that Zeno serve as the tutor to his son, Demetrius, but Zeno excused himself on the ground that he was too old for the job. The man he sent instead, Persaeus, was deeply involved in affairs at court and, according to some sources, died in battle at Corinth in the service of Antigonus. Another Hellenistic strong-man, Cleomenes of Sparta, had the Stoic philosopher Sphaerus as one of his advisors. The reforms instituted in Sparta (including the extension of citizenship to foreigners and the redistribution of land) were seen by some as a Stoic social reform, though it is less clear that it was anything other than an instrument of power for Cleomenes. (See Plutarch's *Life of Cleomenes*.) Peter Green takes a rather more cynical view of the matter (p. 248 ff), but

Green is perhaps unduly hard on Stoicism generally.

In 155 BC Athens sent a delegation of three philosophers (Stoic, Academic skeptic, and Peripatetic) on an embassy to Rome--no Epicurean was included, perhaps because Epicureans refused on principle to participate in public affairs. Their teachings caused a sensation among the educated. The Skeptic Carneades addressed a crowd of thousands on one day and argued that justice was a genuine good in its own right. The next day he argued against the proposition that it was in an agent's interest to be just in terms every bit as convincing. This dazzling display of dialectical skill, together with the deep seated suspicion of philosophical culture, generated a conservative backlash against all Greek philosophers led by Marcus Porcius Cato (the Censor). By 86, however, Rome was ready to receive Greek philosophy with open arms. It was natural that an ambitious and well off Roman like Cicero should go and study at the philosophical schools in Athens and return to popularise Greek philosophy for his less cosmopolitan countrymen. Epicureanism tended to be favored in the ranks in Rome's military, while Stoicism appealed more to members of the Senate and other political movers and shakers. Many Roman politicians at least adopted the high moral tone of Stoicism according to which only virtue is a genuine good, while money, health and even life itself are simply preferred indifferents. Roman political figures associated with Stoicism include Cato the Younger and Scipio Aemilianus (though some of the claims made in earlier scholarship about Greek philosophy and culture and the Scipionic Circle are now regarded with some suspicion). Marcus Brutus (the friend of Cicero who took part in the murder of Julius Caesar) professed Stoicism but was not above engaging in loan-sharking (hence the joke that he was a man of high principles and even higher interest). Pompey thought it sufficiently important to look in on the Stoic philosopher Panaetius of Rhodes in his comings and goings. Octavian (who became Augustus) had a Stoic tutor. Among the Roman emperors, the Stoic philosopher Seneca was the advisor of Nero. Helvidius Priscus advertised himself as a Stoic. When he rather unwisely criticised the Emperor Vespasian in the Senate, he was executed and all the philosophers were excluded from Rome as trouble makers. Under Domitian, they were banished from all of Italy. Clearly the worst of the Roman emperors had no use for people who did not regard death as the greatest of evils! The hostility of the Empire did not last long. Hadrian (117-138) was a friend of Greek philosophy and saw to it that his relative and his successor Antoninus' heir, Marcus Aurelius, had an education which included it. The latter's *Meditations* are still a good read, even if you know nothing about Stoicism but especially if you do. Marcus atoned in effect for Rome's sins against philosophy by establishing ‘professorships’ in the four schools of philosophy in Athens and other cities in 176. In spite of this, Stoicism as a philosophical movement in its own right nearly disappears after the second century.

The influence of Stoicism on the subsequent history of philosophical and religious thought is hard to evaluate directly. The tradition of theories of natural law in ethics seems to stem directly from Stoicism. (Compare Cicero, *de Legibus* I, 18 with later writers like Aquinas in *Summa Theologica* II, 2, q. 94.) Christian theologians were certainly receptive to some of the elements of Stoicism. There was even a forged correspondence between St Paul and Seneca. (That it could ever have been thought genuine tells us something.) Augustine, alas, chose to follow the Stoics rather than the Platonists (his usual allies among the philosophers) on the question of animals' membership in the moral community (*City of God* 1.20). Medieval and Renaissance philosophers were acquainted with Stoicism chiefly through the writings of Seneca and Cicero. The influence of Stoicism on Medieval thought has been considered by

Verbeke and Colish. I know of no study on the impact of Stoicism on Renaissance and early modern thought.

Select Bibliography

Collections of primary texts

- A. A. Long and D. N. Sedley, *The Hellenistic Philosophers* 2 vols (Cambridge, 1987) [Vol. 2 contains an extensive bibliography of scholarly books and articles.]
- B. Inwood and L. Gerson, *Hellenistic Philosophy* 2nd ed. (Indianapolis, 1997) [This volume is cheaper than Long and Sedley, but it lacks the valuable commentary that LS provide. On the other hand, Inwood and Gerson give you more texts on Pyrrhonism.]
- Hans von Arnim, *Stoicorum Veterum Fragmenta* (Leipzig, 1903- 5; vol. 4 indexes, 1924)

Introductions to Stoicism

- F. H. Sandbach, *The Stoics* 2nd ed. (London, 1994)
- A. A. Long, *Hellenistic Philosophy: Stoics, Epicureans, Skeptics* 2nd edition (London, 1986)
- S. Sambursky, *The Physics of the Stoics* (London, 1959) [An interesting book insofar as it attempts to connect aspects of the Stoics physical theory to many contemporary scientific notions. It might be best to read it alongside the review by Wasserstein in *Journal of Hellenic Studies* 83 (1963) before you make up your mind]
- J. M. Rist, *Stoic Philosophy* (Cambridge, Eng., 1969) [Includes a nice discussion of the Stoic views on suicide.]
- R. W. Sharples, *Stoics, Epicureans and Skeptics* (London, 1996) [A thematic treatment of the competing Hellenistic schools.]
- M. Nussbaum, *The Therapy of Desire* (Princeton, 1994) [Not really an introduction, but a splendid book accessible to a wide readership. Considers the important therapeutic element in Hellenistic philosophy.]

Books and Collections of Essays (mostly for advanced students and professional philosophers/classicists)

- J. Annas, *The Morality of Happiness* (New York and Oxford, 1993)
- J. Brunschwig, *Papers in Hellenistic Philosophy* (Cambridge, Eng., 1994)
- J. Brunschwig and M. Nussbaum, *Passions and Perceptions* (Cambridge, Eng., 1993)
- A. A. Long (ed.), *Problems in Stoicism* (London, 1971)
- J. M. Rist, *The Stoics* (Berkeley, 1978)
- M. Schofield, M. Burnyeat and J. Barnes (eds), *Doubt and Dogmatism: Studies in Hellenistic Epistemology* (Oxford, 1980)
- M. Schofield and G. Striker (eds), *The Norms of Nature* (Cambridge, Eng., 1986)

- G. Striker, *Essays on Hellenistic Epistemology and Ethics* (Cambridge, Eng., 1996)
- R. Taylor,*Metaphysics* (2nd edition, Englewood Cliffs, NJ, 1974)
- James Stockdale, *In Love and War* (New York, 1984)

History of the Hellenistic period and subsequent influence of Stoicism

- P. Green, *Alexander to Actium* (Berkeley, 1990)
- G. Verbeke, *The Presence of Stoicism in Medieval Thought* (Washington, 1983)
- M. Colish, *The Stoic Tradition from Antiquity to the Early Middle Ages: 2 vols.* (Leiden, 1985)

Other Internet Resources

- The works two of the later Roman Stoics are available as e-texts:
  - The *Meditations* of Marcus Aurelius
  - Epictetus, *Discourses* and *Enchiridion*
- Ataktos: a dialogue on Stoic ethics (Dirk Baltzly)
  This is a dialogue on the relative merits of the Stoic, Aristotelian and Epicurean conceptions of happiness. It was written for first year students of the subject on morality and objectivity. (Please do not alter the text if you use this for teaching purposes.)
- Table of Contents for the Study Guide (Monash Universtiy: Open Learning Program)
- Stoicism on the Web (Douglas Moore)
  This site contains links to lots of interesting sites as well as biographical material on Zeno, Chrysippus, etc.

Related Entries

Aristotle | Epicurus | skepticism

Dirk Baltzly
dirk.baltzly@arts.monash.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 15, 1996

Content last modified: June 23, 2000

Descartes' Epistemology

René Descartes' approach to the theory of knowledge plays a prominent role in shaping the agenda of early modern philosophy. It continues to effect (some would say "infect") the way problems in epistemology are conceived today. Students of philosophy (in his own day, and in the history since) have found the distinctive features of his epistemology to be at once attractive and troubling; features such as the emphasis on method, the role of epistemic foundations, the conception of the doubtful as contrasting with the warranted, the sceptical arguments of the First Meditation, and the *cogito ergo sum*--to mention just a few that we shall consider.

Depending on context, Descartes thinks that different standards of warrant are appropriate. The context for which he is most famous, and on which the present treatment will focus, is that of investigating First Philosophy. The *first*-ness of First Philosophy is (as Descartes conceives it) one of epistemic priority, referring to the matters one must "first" confront if one is to succeed in acquiring systematic and expansive knowledge.

Section links:

---

1. Knowledge as normative, internalist, and

methodist

1.1. Cartesian knowledge is *normative*.

In recent epistemologies, it is increasingly fashionable to conceive of knowledge in terms that leave little (or no) room for epistemic responsibility. The processes leading to knowledge *happen to* the subject via organs and faculties, much as do circulation of blood, digestion of food, and the like. Such processes (cognitive or otherwise) may occur more or less reliably. Notions as praise and blame, however, need be no more apropos to the reliability of cognitive processes than to that of non-cognitive processes.

In contrast, Descartes regards the categories of praise and blame as quite appropriate to epistemology. Where the acquisition of knowledge is concerned--the philosopher's knowledge, at any rate--one is obliged to withhold assent except when warranted. Judgments grounded in shoddy evidence are conducive to error and thus merit blame, even when the judged matter is true. Among the aims of epistemology, accordingly, is to identify epistemic duties the following of which will ensure error-free judging.

Granting that *ought* implies *can*, normative accounts presuppose some degree of voluntary control over those actions we're duty-bound to perform (or to restrain from performing). In keeping with this, Descartes holds that epistemic warrant or justification (unlike circulation and digestion) is the willful result of a dutiful, reflective mind; specifically, warrant occurs only where there is a systematic withholding of assent, except when prescribed, and a careful directing of attention towards prescribed matters. In the course of the *Meditations*, Descartes purports to reveal both prescriptions.

1.2. Cartesian knowledge is *internalist*.

Where warrant is understood as a normative concept, the would-be knower needs cognitive access to the factors by which warrant is rendered. The jargon of "internalism" and "externalism" is often applied here, but an adequate characterization of the internalism-externalism distinction is difficult and beyond our aims. According to a construal that will serve our purposes, a theory of justification is internalist insofar as it requires that the justifying factors grounding knowledge claims are accessible to the knower's consciousness. On this construal, Descartes is a *thorough-going* internalist.

Descartes thinks it a straightforward consequence of 17th century mechanist doctrine that requisite cognitive access can come only by means of ideas (conscious states of mind). To use a metaphor Descartes is fond of, evidence of the sort relevant to the internalist is directly present to the mind's eye--it is "intuited" with mental vision. To the extent that we have cognitive access to that which is directly present to our literal eyes (and other organs of physiology), it is by means of ideas. As Descartes conceives the task of acquiring knowledge, the would-be knower must begin with the data of consciousness (the *internal* world) and then somehow build up to (hopefully expansive) knowledge wholly on the basis of such data. The problem of so expanding our knowledge to include even the

external world is among the central epistemological problems of 17th and 18th century philosophy.

1.3. Cartesian knowledge is *methodist*.

How is the epistemologist to proceed in identifying candidates for knowledge? A distinction will prove useful--that between particularist and methodist responses to this question. The *particularist* is apt to presuppose the credibility of our prima facie intuitions regarding particular items of knowledge, and then use these intuitions to help identify more general epistemic duties. The *methodist*, on the other hand, is apt to hold that our prima facie intuitions are unstable and unreliable; that we ought to begin with a method to help us arrive at settled, reflective intuitions as to which particular knowledge claims are credible.

Famously, Descartes is in the methodist camp. Those who haphazardly "direct their minds down untrodden paths" are sometimes "lucky enough in their wanderings to hit upon some truth," but "it is far better," writes Descartes, "never to contemplate investigating the truth about any matter than to do so without a method" (*Rules* 4, AT 10:371). Though it's prima facie palpable that the earth is unmoved, and that ordinary objects (as tables and chairs) *are* just as they *seem*, the newly emerging mechanist doctrines of the 17th century have it that such matters are in fact false. Such cases underscore the unreliability of our prima facie intuitions and the need for a method to help us distinguish these seductive though false matters from genuine truth.

In the dialectic of the First Meditation, Descartes confronts common sense particularism in an effort to show (among other things) that our prima facie intuitions are in need of revision. (As the narrative unfolds, the common sense perspective is often intermingled with that of Aristotelian scholasticism, the prevailing philosophical system in the schools--a system that was common sense oriented in terms of its credulous trust in sense perception.) At the first level of the dialectic, while speaking on behalf of common sense, the meditator-spokesperson (hereafter referred to as the "meditator") appeals to the seeming obviousness of claims as "that I am here, sitting by the fire, wearing a winter dressing-gown, holding this piece of paper in my hands, and so on"--particular matters "about which doubt is quite impossible," or so it would seem (AT 7:18). In response (and at each level of the dialectic), Descartes invokes his own methods to show that the prima facie obviousness of such particular claims is insufficient to meet the burden of proof to which we're epistemically obliged.

Further reading: On normativity, see the Fourth Meditation and *Prin*. 1.36ff ; see also Alston (1989). On belief voluntarism, see Audi (1999). On the internalism-externalism distinction, see Alston (1989) and Plantinga (1993). On the methodism-particularism distinction, see Chisholm (1982).

[Return to Section links]

2. Descartes' methods: foundationalism and doubt

Of his own methods, Descartes writes:

> Throughout my writings I have made it clear that my method imitates that of the architect. When an architect wants to build a house which is stable on ground where there is a sandy topsoil over underlying rock, or clay, or some other firm base, he begins by digging out a set of trenches from which he removes the sand, and anything resting on or mixed in with the sand, so that he can lay his foundations on firm soil. In the same way, I began by taking everything that was doubtful and throwing it out, like sand ... (Replies 7, AT 7:537)

As suggested by these remarks, Descartes' two principal methods--foundationalism and doubt--are integrated. Since he holds that matters that are prima facie obvious, or self-evident, often turn out false, and since the method of foundationalism depends on an ability to avoid such error when identifying the foundations, a complement method is required if the search for unshakable foundations is to succeed. In the present Section, we consider both of Descartes' methods, making liberal use of his architectural metaphor.

2.1. Foundationalism.

The central insight of foundationalism is that a system of epistemic justification can be fashioned after the manner of a structurally sound house. Such a house might owe its unshakability to two features: a firm *foundation* and a well-anchored *superstructure* consisting of tightly linked support beams firmly grounded into the foundation. Likewise, where warrant is construed in terms of unshakable certainty, a system of knowledge might emerge from two parallel features: a foundation of unshakably certain first principles, and a superstructure of further claims anchored into the foundation by means of unshakably certain inference. Among its considerable alleged benefits, foundationalism provides the means for a potentially indefinite expansion of one's stock of knowledge, from relatively meager beginnings. This is especially significant where there are few first principles.

(It's worth noting that foundationalism, as here characterized, presupposes a conception of truth that involves more than coherence. If relations of coherence exhausted the notion of truth, judgments as to the truth of first principles could not be justified (qua first principles), nor could inference be justification *preserving* as opposed to justification *conferring*.)

A paradigmatic example of a foundationalist system is Euclid's geometry. Euclid begins with a foundation of first principles--his definitions, postulates, and axioms or common notions--on which he then bases a superstructure of further propositions. Descartes' own designs for metaphysical knowledge are inspired by Euclid's system: "Those long chains composed of very simple and easy reasoning, which geometers customarily use to arrive at their most difficult demonstrations, had given me occasion to suppose that all the things which can fall under human knowledge are interconnected in the same way" (*Discourse* 2, AT 6:19). Though it would be misleading to characterize the constructive component of the *Meditations* as straightforwardly fashioned after the manner of the geometers, Descartes does think it *can* be reconstructed as such, and he expressly does so at end of the Second Replies--providing a

"geometrical" exposition of his central constructive steps, under the following headings: *definitions*, *postulates*, *axioms or common notions*, and *propositions* (AT 7:160ff).

As alluded to above, the *Meditations* contains a destructive component that Descartes likens to the architect's preparations for laying a foundation. Though the component finds no analogue in the method of the geometers, Descartes looks to hold that it is needed because of unique problems associated with the discovery of first principles in metaphysical inquiry. The discovery of Euclid's first principles (some of them, at any rate) is comparatively unproblematic: claims as that *things which are equal to the same thing are also equal to one another* (Euclid's first axiom) wear their truth on their sleeve. In metaphysics, however, our analyses might reveal first principles that are in conflict with preconceived opinions deriving from the senses.

> The difference is that the primary notions which are presupposed for the demonstration of geometrical truths are readily accepted by anyone, since they accord with the use of our senses. Hence there is no difficulty there, except in the proper deduction of the consequences, which can be done even by the less attentive, provided they remember what has gone before. ... In metaphysics by contrast there is nothing which causes so much effort as making our perception of the primary notions clear and distinct. Admittedly, they are by their nature as evident as, or even more evident than, the primary notions which the geometers study; but they conflict with many preconceived opinions derived from the senses which we have got into the habit of holding from our earliest years, and so only those who really concentrate and meditate and withdraw their minds from corporeal things, so far as possible, will achieve perfect knowledge of them. (Replies 2, AT 7:156-57)

(Note: Descartes uses ‘perceive’/‘perception’ (*percipio*/*perceptio*) with much wider scope than is the current practice in philosophy: for Descartes, to perceive *X* is, roughly, to be aware of *X*.)

Among Descartes' persistent themes is that such preconceived opinions can have the effect of obscuring our mental vision; that where there are disputes about first principles, it is not "because one man's faculty of knowledge extends more widely than another's, but because the common notions are in conflict with the preconceived opinions of some people who, as a result, cannot easily grasp them"; whereas, "we cannot fail to know them when the occasion for thinking about them arises, provided that we are not blinded by preconceived opinions" (*Prin*. 1.49-50, AT 8a:24). These "preconceived opinions" must be "set aside," says Descartes, since doing so "is wholly necessary in order to lay the first foundations of philosophy" (May 1643, AT 8b:37). Unless they are set aside, we're apt to regard, as bona fide first principles, the mistaken (though prima facie obvious) sensory claims that particularists find attractive. Again, what could be more palpable than that the earth is unmoved? (Indeed, defenders of common sense have even proposed that "the consent of ages and nations, of the learned and unlearned, ought to have great authority with regard to first principles" (Reid 1785, 6.4), perhaps unwittingly making the case that the scholastic doctrine that the earth is unmoved be regarded as a first principle.)

Not only do preconceived opinions interfere with the search for first principles, Descartes thinks they hinder our ability to generate foundationalist superstructure:

> [A]lthough the proofs I employ here are in my view as certain and evident as the proofs of geometry, if not more so, it will, I fear, be impossible for many people to achieve an adequate perception of them, both because they are rather long and some depend on others, and also, above all, because they require a mind which is completely free from preconceived opinions and which can easily detach itself from involvement with the senses. (*Dedicatory Epistle* for *Meditations*, AT 7:4)

Though the method of foundationalism brilliantly allows for the expansion of knowledge from meager beginnings, the method is incomplete if prejudices thwart our ability to identify the foundations. To help "set aside" these preconceived opinions, Descartes devises a second method--the method of doubt.

(We should add that there has been a history of controversy surrounding Euclid's fifth postulate--the parallel postulate. Some have taken the worries, here, as suggesting intractable problems for any endeavor to discover unshakably certain foundations.)

2.2. Methodological doubt.

Descartes opens the First Meditation asserting the need "to demolish everything completely and start again right from the foundations" (AT 7:17). Whereas, in the architectural metaphor, we can think of tractors and bulldozers as the tools of demolition, the tool of epistemic demolition is sceptical doubt. Tractors undermine literal ground; doubt undermines epistemic ground. Descartes' aim, however, is not that of "the sceptics, who doubt only for the sake of doubting," but instead "to reach certainty--to cast aside the loose earth and sand so as to come upon rock or clay" (*Discourse* 3, AT 6:28-29). In our consideration of the method of doubt, we begin by addressing a few misunderstandings; we then focus on how the *universality* and *hyperbole* of doubt are intended to impede the corrupting influence of preconceived opinion.

Misconceptions about the method abound. Two of these are illustrated in a passage from the pragmatist Peirce:

> We cannot begin with complete doubt. We must begin with all the prejudices which we actually have when we enter upon the study of philosophy. These prejudices are not to be dispelled by a maxim [viz. the maxim that the philosopher "must begin with universal doubt"], for they are things which it does not occur to us *can* be questioned. Hence this initial skepticism will be a mere self-deception, and not real doubt ... A person may, it is true, in the course of his studies, find reason to doubt what he began by believing; but in that case he doubts because he has a positive reason for it, and not on account of the Cartesian maxim. Let us not pretend to doubt in philosophy what we do not doubt in our hearts. (1955, 228f)

One misconception is in supposing that universal doubt is intended to result from the mere effort to

adhere to the maxim--as if by sheer effort of will. Quite to the contrary, Descartes introduces sceptical arguments precisely in acknowledgement that, where it does not occur to us that our judgments can be questioned, we need some prompting; we need to be provided with reasons for doubt on which to reflect. Descartes writes:

> I did say that there was some difficulty in expelling from our belief everything we have previously accepted. One reason for this is that before we can decide to doubt, we need some reason for doubting; and that is why in my First Meditation I put forward the principal reasons for doubt. (Replies 5, appendix, AT 9a:204)

A further misconception has it that the intended attitude of doubt need involve the kind of sincerity that Peirce characterizes as a "doubt in our hearts." Distinguish (what I'll call) a *hypothetical* from a *sincere* doubt. Merely hypothetical doubt involves recognition that one's warrant is undermined by a sceptical hypothesis that one regards as extravagant (but not unthinkable). As such, a merely hypothetical doubt need not undermine one's inclination to assent. In contrast, sincere doubt involves recognition that one's warrant is undermined by a sceptical hypothesis that one regards as plausible. Sincere doubt does undermine one's inclination to assent, and only it involves "doubt in our hearts" (as it were). The misconception is in supposing that Descartes' method requires a sincere doubt as opposed to a merely hypothetical doubt. Hypothetical doubt is in fact sufficient to induce recognition that one's confidence is not unshakably firm as is required for the foundations of knowledge. Moreover, it is clear that Descartes regards his hyperbolic doubts as merely hypothetical. He concludes the Synopsis of the *Meditations* by noting that "no sane person has ever seriously doubted" many of the matters that are called into question, such as the existence of the external world. (Rather, he says the primary aim is to establish issues of epistemic priority.) And subsequent to introducing his most extravagant doubt, he concedes that the preceding opinions it undermines remain "much more reasonable to believe than to deny" (AT 7:22).

A related misconception has Descartes calling not merely for doubt, but for disbelief or dissent. Gassendi, e.g., misreads Descartes as urging us to "consider everything as false" (Objs. 5, AT 7:257-58). But surely the spirit of Descartes' invocation to doubt (even if not always its letter) is that, due to the importance of unshakable foundations, we are to "hold back [our] assent from opinions which are not completely certain and indubitable just as carefully as [we] do from those which are patently false" (Med. 1, AT 7:18).

Yet another misconception has it that the universality of doubt renders inert Descartes' own sceptical hypotheses, since they are dubious in every case. Yet since the motivation for the method is to avoid building on dubious/shakable foundations, the scope of universal doubt need not encompass the sceptical hypotheses themselves, as opposed merely to *every* candidate for knowledge (without exception). And since the doubt required by the method is merely hypothetical, the sceptical hypotheses need not seem plausible in order to serve as tools of demolition--"there may be reasons which are strong enough to compel us to doubt, even though these reasons are themselves doubtful, and hence are not to be retained later on" (Replies 7, AT 7:473-74).

These are by no means the only misconceptions about the method of doubt. Even barring such misconception, Descartes admits that "the usefulness of such extensive doubt is not apparent at first

sight," adding that "its greatest benefit lies in freeing us from all our preconceived opinions" (Synopsis, AT 7:12). Further appeal to the architectural metaphor will help to elucidate how the *universality* and *hyperbole* of Descartes' doubt (two features that he consistently emphasizes) are supposed to help us to "set aside" the preconceived opinions apt to obscure the search for first principles.

That doubt is to be *universal* has two aspects: it is to be applied *without exception*, and it is to be applied *collectively rather than piecemeal*. The latter deserves further comment. The requirement that doubt be collective rather than piecemeal is intended in part to prevent a pseudo-firmness that might otherwise result from the supportive influence that judgments have on one another. Suppose an architect attempted to secure the structural stability of a house by renovating it in piecemeal fashion, rather than by means of a universal demolition followed by a complete rebuilding. A potential problem looms in that, roughly in proportion with the number of boards one binds together (other things equal), a structure will seem increasingly sturdy--via a mass-coherence-induced-pseudo-firmness. Imagine our architect moving from one support beam to another, checking each one for sturdiness even while the whole structure remains intact. This piecemeal procedure is apt to lead her to confuse the stability of mass-coherence with that of bedrock-grounding. Indeed, such a procedure might well exhibit an aircraft carrier as resting on unshakable foundations, even though it is (in the relevant sense) nothing more than a massive raft--grounded in nothing, perhaps drifting aimlessly. Likewise, where one's web of preconceived opinion is sufficiently massive and coherent, a procedure of attempting to shake individual such opinions, in piecemeal fashion, while allowing the rest of one's web to remain intact, is apt to exhibit a great many falsehoods as seeming to be quite firm/sure. Universal doubt is intended to avoid this kind of *pseudo*-firmness. If one first disassembles the various support beams from one another before testing their shakability, the difference between mass-coherence-induced-pseudo-stability and bedrock-grounded-stability is more perspicuous. Descartes is supposing that the same benefit is realized when building epistemic structures: by first razing the existing structure in its entirety, one is more likely to hit on first principles whose firmness is not mass-coherence-induced.

In the Seventh Replies, Descartes offers yet another metaphor to illustrate the benefit of the universality requirement in preventing the corrupting influence that judgments can have on one another.

> Suppose [an inquirer] had a basket full of apples and, being worried that some of the apples were rotten, wanted to take out the rotten ones to prevent the rot spreading. How would he proceed? Would he not begin by tipping the whole lot out of the basket? And would not the next step be to cast his eye over each apple in turn, and pick up and put back in the basket only those he saw to be sound, leaving the others? In just the same way, those who have never philosophized correctly have various opinions in their minds which they have begun to store up since childhood, and which they therefore have reason to believe may in many cases be false. They then attempt to separate the false beliefs from the others, so as to prevent their contaminating the rest and making the whole lot uncertain. Now the best way they can accomplish this is to reject all their beliefs together in one go, as if they were all uncertain and false. They can then go over each belief in turn and re-adopt only those which they recognize to be true and indubitable. (Replies 7, AT 7:481)

Turning our attention to Descartes' emphasis on a *hyperbolic* doubt, again the requirement is intended to prevent a pseudo-firmness that might otherwise result when attempting to identify first principles. Remarks by Gassendi make him a useful foil:

> There is just one point I am not clear about, namely why you did not make a simple and brief statement to the effect that you were regarding your previous knowledge as uncertain so that you could later single out what you found to be true. ... This strategy made it necessary for you to convince yourself by imagining a deceiving God or some evil demon who tricks us, whereas it would surely have been sufficient to cite the darkness of the human mind or the weakness of our nature. (Objs. 5, AT 7:257-58)

The architectural metaphor is again helpful. Suppose our architect is vigilant in meeting the first requirement--a universal demolition. Even so, she is apt to confound firm dirt with genuine bedrock if she is not using heavy-duty demolition tools. Indeed, she's apt to rebuild on the very same ground that, prima facie, seemed immovable when the original edifice was built. Descartes thinks there's a nearly irresistible urge to regard as immov*able* the comfortable ground which has, for a lifetime, been unmov*ed*. In the absence of powerful demolition equipment, Descartes' own project thus threatens to amount to no more than an exercise in foundationalist rebuilding on a given (preconceived), unquestioned foundation--a serious concern, in view of the proclivity of his scholastic readers to regard the questioning of ancient and divine authorities as tantamount to heresy. As Descartes writes, in response to Gassendi:

> Is it really so easy to free ourselves from all the errors which we have soaked up since our infancy? Can we really be too careful in carrying out a project which everyone agrees should be performed? ... most people, although verbally admitting that we should escape from preconceived opinions, never do so in fact, because ... they reckon that nothing they have once accepted as true should be regarded as a preconceived opinion. (Replies 5, AT 7:348)

Just as light-duty gardening tools are incapable of uncovering bedrock (*as* bedrock), Descartes needs to provide his readers with heavy-duty demolition gear (i.e. hyperbolic sceptical arguments) if the method is to succeed in preventing the confounding of cherished prejudices with unshakable foundations.

Hence the importance of a universal, hyperbolic doubt to the larger effort to apply foundationalism to metaphysical inquiry. Indeed, Descartes regards the sceptical/destructive component as of such importance to the success of his larger project that he recommends we "devote several months" to the First Meditation alone (Replies 2, AT 7:130). Of course, there is no guarantee, at the outset, that a careful procedure of doubt will achieve the intended result. There are a number of plausible, failed outcomes: (i) perhaps there are no unshakable truths of the sort the method is intended to reveal; or, even granting that there are, (ii) perhaps Descartes' sceptical hypotheses are, though hyperbolic, nonetheless too weak to clear away the prejudices obscuring bedrock; or, even granting that they are sufficiently heavy-duty, (iii) perhaps the meditator will look in all the wrong places for bedrock. On the other hand, if the method does in fact reveal some unshakably certain matters--even in the face of the most hyperbolic doubt that we're capable of contriving--then Descartes thinks that such matters are the stuff of knowledge if anything is:

where our "conviction is so firm that it is impossible for us ever to have any reason for doubting what we are convinced of, then there are no further questions for us to ask: we have everything that we could reasonably want" (Replies 2, AT 7:144-45).

The propriety of Descartes' (oft-criticized as being too high) burden of proof is elucidated by analogy to criminal law. Consider two desiderata of our criminal court system: to convict *all* the guilty, and to convict *only* the guilty. As one raises the burden of proof imposed on the prosecution, the prospects of achieving the first desideratum are decreased; the prospects of achieving the second desideratum are increased. The burden can thus be raised or lowered in accordance with the desideratum one wishes to emphasize. The implications are clear for epistemologists who share the following two desiderata: to warrant *all* the true, and to warrant *only* the true. Descartes places a premium on achieving the second desideratum at all costs--a good method is such that following it ensures that "one will never take what is false to be true" (*Rules* 4, AT 10:371-72). Given this priority, it is appropriate to impose a maximally high burden of proof on the inquirer--we "cannot possibly go too far" in the application of doubt (Med. 1, AT 7:22).

In view of this high burden of proof, it emerges that a requirement of knowledge (as Descartes conceives it) is *full* indefeasibility (i.e. full immunity to doubt): one's conviction must be "so firm that it is impossible for us ever to have any reason for doubting" (Replies 2, AT 7:144); "so strong that it can never be shaken" (24 May 1640, AT 3:65), not even in the face of the most hyperbolic of doubts which might be contrived by the sceptic. The effect may be to limit the eventual stock of knowledge, but the brand of knowledge worthy of the philosopher, as opposed to that appropriate for the mundane affairs of daily life, calls for high standards (cf. Replies 2, AT 7:149). Hereafter, I refer to this rigorous brand of knowledge that Descartes seeks (what he calls *scientia*) as Knowledge (uppercase ‘K’).

Further reading: On foundationalism: for Descartes' treatment, see *Discourse*, First Meditation, and Seventh Objections and Replies; for its treatment by ancients, see Euclid (1956) and Aristotle (*Posterior Analytics*); by interpreters of Descartes, see Sosa (1997a) and Van Cleve (1979). On Cartesian inference (there are disputes as to what "certain inference" (my term) comes to, for Descartes), see Descartes' *Rules* (bearing in mind that the he never finished this work, much less published it, and some of the doctrines there are at odds with his published writings); by commentators, see Gaukroger (1989) and Hacking (1980). On methodological doubt: for Descartes' treatment, see *Rules*,*Discourse*, First Meditation, and Seventh Replies; by commentators, see Frankfurt (1970), Garber (1986), Williams (1983), and Wilson (1978). For a contemporary application interestingly similar to Descartes' treatment of doubt vis-à-vis first principles, see Rawls (1971) on reflective equilibrium. A discussion of the relation between the methods of doubt and foundationalism is somewhat incomplete without consideration of another distinction of method, that between analysis and synthesis. On the analysis-synthesis distinction: see the Second Replies (AT 7:155ff); see also Galileo (1967, 50f), Arnauld (*L'Art de Penser*, 4.2-3), Curley (1986), and Hintikka (1978). On the indefeasibility (and other conditions) of Knowledge/*scientia*, see Newman and Nelson (1999).

[Return to Section links]

3. First Meditation sceptical arguments

Of his sceptical arguments, Descartes says he "could not have left them out, any more than a medical writer can leave out the description of a disease when he wants to explain how it can be cured" (Replies 3, AT 7:172). Among the aims of an accurate sceptical diagnosis is to reveal the comparative shakability of various kinds of judgments--so as to expose which kinds are more (and less) suitable as materials for building enduring Knowledge. Towards this aim, Descartes' diagnosis unfolds from the least hyperbolic doubt (that which undermines the fewest kinds of judgments) to the most hyperbolic doubt. In what follows, we first consider his treatment of doubts motivated by reflection on dreaming. We then consider his most powerful doubt, a doubt motivated by reflection on the veracity of our cognitive faculties.

3.1. Dreaming Doubts.

Historically, there are two distinct dream-related skeptical doubts. The one doubt undermines the judgment that one is *presently* awake--call this the Now Dreaming Doubt. The other doubt undermines the judgment that one is *ever* awake (i.e. in the way normally supposed)--call this the Always Dreaming Doubt. A textual case can be made on behalf of either formulation (or both) being raised in the *Meditations*. We'll not attempt to settle this interpretive dispute here, though we will discuss both.

Both kinds of dream doubt appeal to some version of the thesis that the experiences we take as dreams are (at their best) qualitatively similar to the experiences we take as waking--call this the Similarity Thesis. The Similarity Thesis may be formulated in a variety of strengths. Indeed, disputes as to *how similar* the experiences of the dreaming-kind are to those of the waking-kind have raged for more than two millennia. The tone of the debate suggests that the degree of qualitative similarity may vary across individuals (or, at least, across their *recollections* of dreams). Granting such variation, dreaming doubts that appeal to weaker versions of the Similarity Theses are apt (all other things equal) to persuade more people. Without attempting, here, to settle the interpretive issue of how strong a Similarity Thesis Descartes intends to advance, let's consider a textually defensible formulation that errs on the weak side. The relatively weak thesis I have in mind is this: that the qualitative features of the experiences we take as dreams are sufficiently similar to those of the experiences we take as waking, as to render it *not unthinkable* that waking-quality features should be reproduced in a dream: "every sensory experience I have ever thought I was having while awake I can also think of myself as sometimes having while asleep" (Med. 6, AT 7:77). This version of the Similarity Thesis is endorsable by those who never recollect dreams that seem (on hindsight) qualitatively indistinguishable from the experiences they take as waking; indeed, it's endorsable even by those who simply do not remember their dreams to any significant degree. Descartes' own meditator, however, appears to be someone who remembers dreams of sufficient qualitative similarity to have fooled him while having them: "As if I did not remember other occasions when I have been tricked by exactly similar thoughts while asleep" (Med. 1, AT 7:19).

This weak Similarity Thesis is sufficient for Descartes' purposes. Recall that his method only requires a hypothetical doubt. As long as it is *not unthinkable* that waking-quality experiences should be reproduced

in a dream, then Descartes thinks we're unable to meet the burden of proof requisite for Knowledge--"there are never any [*unshakably*] *sure signs* by means of which being awake can be distinguished from being asleep" (Med. 1, AT 7:19; italics added). Even if one is *very confident* of being able to distinguish dreaming from waking, Descartes thinks this level of confidence is not *so* high as to constitute warrant.

(Much to-do is often made over whether dreaming arguments are self-refuting. The worry is that Similarity Theses, on which such arguments are based, must presuppose that we already know which experiences are dreams and which are waking, in order to assess their comparative quality; this, however, is precisely the distinction that dreaming arguments purport to undermine. Descartes preempts this debate by couching his claims in terms of experiences we *think* of as dreams, those we *think* of as waking, those we *remember* as dreams, and so on.)

We're now in position to generate straightaway the Now Dreaming Doubt: since there are no unshakably sure signs of being awake, for all I Know I'm now dreaming. It's *not unthinkable*, even if I don't believe it. The sceptical consequences of this admission are not insignificant. The further judgments that are called into doubt include all those in which my confidence is shaken on the hypothesis that I'm *now dreaming*--e.g., the judgment that I'm now "holding this piece of paper in my hands." It's not unthinkable that, in a few moments, I'll awaken to the reality that I have no hands. Descartes is not denying the truth of such claims, as to what one is holding in one's hands. He's not even denying that we know them (small 'k', the sense appropriate for practical matters). He's questioning whether we have an unshakable certainty as is requisite for the foundations of Knowledge.

Descartes appears to hold that, from the very same Similarity Thesis, together with a further assumption, the more powerful sceptical result can be generated, namely the Always Dreaming Doubt. The further assumption is that, for all we Know, the processes producing the experiences we take as waking are no more veracious than those producing the experiences we take as dreams. As Descartes writes:

> [E]very sensory experience I have ever thought I was having while awake I can also think of myself as sometimes having while asleep; and since I do not believe that what I seem to perceive in sleep comes from things located outside me, I did not see why I should be any more inclined to believe this of what I think I perceive while awake. (Med. 6, AT 7:77)

The aim of the Always Dreaming Doubt is not to undermine our confidence that we're now awake; nor to question whether there are significant differences between the experiences we categorize as waking and those we categorize as dreaming. Instead, the aim is to undermine our confidence that experiences in the waking category need be produced by external objects. In both categories of experience, our (internalist relevant) cognitive access extends only to the productive *result*, but not to the productive *process*. On what basis, then, do we suppose that external objects play any more a role in producing the experiences in the waking category than those in the dreaming category? It is not unthinkable that "there may be," says the meditator, "some other faculty not yet fully known to me, which produces these ideas without any assistance from external things; this is, after all, just how I have always thought ideas are produced in me when I am dreaming" (Med. 3, AT 7:39). Granting, then, that we are not *unshakably certain* that our "waking" experiences are produced by external objects, *all* of our experiences might be dreams *of a sort*.

The sceptical consequences of this Always Dreaming Doubt are considerably more potent: if we are not unshakably certain that external objects contribute to the production of our "waking" ideas of them, then (thinks Descartes) we are not unshakably certain that there even *are* any such external objects. Our best evidence for the existence of an external world of tables and chairs, and the like, comes from our preconceived opinions as to the role of external objects in producing our ideas of them.

> All these considerations are enough to establish that it is not reliable judgement but merely some blind impulse that has made me believe up till now that there exist things distinct from myself which transmit to me ideas or images of themselves through the sense organs or in some other way. (Med. 3, AT 7:39-40)

The two dreaming doubts are parasitic on the same Similarity Thesis, though their sceptical consequences differ. The Now Dreaming Doubt invokes the *universal conceivability of delusion*: for any one of my sensory experiences, it's not unthinkable that the experience is delusive. The Always Dreaming Doubt invokes the *conceivability of universal delusion*: it's not unthinkable that all my sensory experiences are delusions (say, from a God's-eye perspective).

3.2. Meta-Cognitive Doubts.

Though dreaming doubts undermine a significant number of Knowledge claims, Descartes regards them as only moderately hyperbolic. His most hyperbolic doubt undermines everything the weaker doubts do, and more. Some readers might have thought to ask, "What's left to undermine?" The meditator proposes a response: "whether I am awake or asleep, two and three added together are five, and a square has no more than four sides. It seems impossible that such transparent truths should incur any suspicion of being false." (Med. 1, AT 7:20) Descartes' next doubt is intended to defeat even these claims.

(It's tempting to read Descartes as holding that only *a priori* judgments (i.e. in the post-Kantian sense) survive dreaming doubts--that the eventual foundations of Knowledge will be exclusively *a priori*. It emerges in the Second Meditation, however, that *a posteriori* judgments (as, e.g., concerning the present contents of consciousness) can be among our epistemic best (cf. §5.3 below). It would be more accurate to say that Descartes regards dreaming doubts as having their most damaging undermining impact on judgments of external sense.)

Suppose that our cognitive equipment is flawed and our intellectual faculties are no more reliable than a broken calculator--that we're "wired" (as it were) to compute even simple matters in error. In that case, even those matters that seem most evident (as e.g. those surviving dreaming doubts) might well be false notwithstanding their apparent status as supremely evident. According to Descartes, "the most serious doubt [arises] from our ignorance about whether our nature might not be such as to make us go wrong even in matters which seemed to us utterly evident" (*Prin*. 1.30, AT 8a:16). If we do not have unshakable certainty as to the perfect reliability of our cognitive equipment, observes Descartes, then there is nothing so evident as to escape doubt. Call this doubt about the reliability of our cognitive faculties Meta-

Cognitive Doubt (MCD).

(Note: If the very credibility of our cognitive equipment is in doubt, how *could* we ever gain Knowledge by means of such equipment? Descartes thinks this is what makes hyperbolic doubt so insidious, and why we should regard his eventual solution as so brilliant.)

Lest the reader suppose that even (mere) hypothetical doubt should not be pressed *this* far, Descartes presents a dilemma intended to motivate MCD. Either the design of our cognitive faculties ultimately traces to an all-powerful creator, or it does not. Assuming it does, then for all we Know this omnipotent creator might be a "malign genius" who intentionally designed us with cognitive flaw; a creator this powerful "could have given me a nature such that I was deceived even in matters which seemed most evident" (AT 7:36). On the other hand, given the opposite assumption (as to the omnipotence of the creator), we're no better off--as Descartes explains in the First Meditation:

> Perhaps there may be some who would prefer to deny the existence of so powerful a God rather than believe that everything else is uncertain. ... yet since deception and error seem to be imperfections, the less powerful they make my original cause, the more likely it is that I am so imperfect as to be deceived all the time. (AT 7:21).

Today's reader is likely to try blocking the second horn of dilemma--appealing to a natural selection account, perhaps even forgiving Descartes for not having the benefit of Darwinian theory. But natural selection accounts guarantee, at best, that an organism is *as fit* as the local competition--a result that does not preclude our having cognitive flaw.

The typical reader in Descartes' own day would have instead tried to block the first horn of the dilemma--appealing to then-standard doctrines about the nature of God, as being not only omnipotent but also omnibenevolent: "God would not have allowed me to be deceived in this way, since he is said to be supremely good" (Med. 1, AT 7:21). Anticipating this move, Descartes poses the so-called problem of evil, as applied to error: "if it were inconsistent with [God's] goodness to have created me such that I am deceived all the time, it would seem equally foreign to his goodness to allow me to be deceived even occasionally; yet this last assertion cannot be made" (ibid.). This reductio ad absurdum is supposed to raise doubts as to whether the existence of an omnipotent and omnibenevolent creator is compatible with the undisputed occurrence of (at least) occasional error. And so, this kind of effort to block MCD fails, that is unless the reader already has an unshakably certain solution to the problem of evil.

(A common interpretive mistake is to suppose that, according to Descartes, it would take an omnipotent evil genius in order to undermine our confidence in our epistemic best. Descartes' handling of the above dilemma, along with a perusal of the various texts in which he discusses his radical doubt (cf. AT 3:64-65, 7:21, 7:36, 7:70, 7:77, 8a:6, 8a:9-10, 8a:16), shows that the evil genius plays an ancillary role to the more fundamental worry about cognitive flaw.)

Having reflected on the dilemma, Descartes' meditator states: "I have no answer to these arguments, but

am finally compelled to admit that there is not one of my former beliefs about which a doubt may not properly be raised; and this is not a flippant or ill-considered conclusion, but is based on powerful and well thought-out reasons" (Med. 1, AT 7:21).

Doubt is now both *hyperbolic* and *universal*. Since Descartes regards the refutation of MCD as essential to epistemic warrant, it emerges that a requirement of Knowledge is that the would-be Knower have an unshakable confidence in the perfect reliability of his cognitive equipment. In §5.2, we return to a consideration of MCD and the *indirect* manner in which it calls matters into doubt.

Further reading: On Descartes' sceptical arguments, see Bouwsma (1949), Curley (1978), Newman (1994), Newman and Nelson (1999), Williams (1986), and Wilson (1978). For a more general philosophical treatment of dreaming arguments, see Dunlap (1977).

[Return to Section links]

4. Cogito ergo sum

Famously, Descartes holds that the occurrence of thought guarantees the existence of a thinker. A version of this insight appears in every published work in which he treats scepticism--unlike the canonical slogan, '*cogito ergo sum*' ('*je pense, donc je suis*', 'I think therefore I am'). As illustrated early in the Second Meditation, the purported insight has it that though the existence of my body is subject to doubt, the existence of me--qua *thinker*--looks to withstand even the most hyperbolic of doubts: let the evil genius "deceive me as much as he can, he will never bring it about that I am nothing so long as I think that I am something" (AT 7:25). The very attempt to doubt one's own existence turns out self-stultifying: every such effort *is* an occurrence of thought; in turn, the occurrence of thought requires a thinker, albeit only a minimally construed thinker. Descartes regards the *cogito* (as I shall refer to it) as the "first and most certain of all to occur to anyone who philosophizes in an orderly way" (*Prin*. 1.7, AT 8a:7).

The *cogito* has generated an enormous literature--both pro and con. Though the relevant issues are far too numerous and difficult to address here in a systematic way, a few points of clarification might help to preclude some of the philosophic and interpretive mistakes that have ensnared many a critic. I'll restrict my observations concerning the *cogito* to its treatment by Descartes in contexts of systematic doubt.

First, a first-person formulation is essential to the success of the *cogito*. Third-person claims, such as "Icarus thinks," or even "Descartes thinks," are not unshakably certain--not for me, at any rate; only the occurrence of my own thought has a chance of resisting hyperbolic doubt. There are a number of passages in which Descartes refers to a third-person version of the *cogito*, but none of these occur in the context of trying to establish categorically the existence of a particular thinker (as opposed merely to the conditional existence of whatever thinks).

Second, the *cogito* is intended to involve the occurrence of *cogitatio* (i.e. cogitation or thinking, or

consciousness more generally). Any mode of thinking will do: doubt, understanding, affirmation, denial, volition, imagination, sensation, or the like (cf. Med. 2, AT 7:28). Non-cogitative occurrences, on the other hand, will not work. For instance, it won't suffice to reason that "I exist since I am walking," because hyperbolic doubt calls into question the existence of my legs. A simple revision, as "I exist since it *seems* I'm walking," restores the anti-sceptical potency (cf. Replies 5, AT 7:352).

A caveat is in order. That Descartes rejects the indubitability of such formulations, as those that presuppose the existence of body, commits him to nothing more than an epistemological distinction, but not yet an ontological distinction (not to mention a real, or substantial distinction), between mind and body. Indeed, on the heels of the *cogito*, Descartes has his meditator say:

> And yet may it not perhaps be the case that these very things which I am supposing to be nothing [e.g., "that structure of limbs which is called a human body"], because they are unknown to me, are in reality identical with the "I" of which I am aware? I do not know, and for the moment I shall not argue the point, since I can make judgements only about things which are known to me. (Med. 2, AT 7:27)

Third, and related to this last quotation, is that Descartes' reference to an "I" (*ego*), in the "I think" (*cogito*), is not intended to presuppose the existence of a *substantial self*. Indeed, in the very next sentence following the initial statement of the *cogito*, the meditator says: "But I do not yet have a sufficient understanding of what this ‘I’ is, that now necessarily exists" (Med. 2, AT 7:25). What function, then, does the "I" serve in the evidential claim "I think"? Many a critic has complained that in referring to the "I" Descartes begs the question, since he presupposes what he intends to establish in "I exist." Bertrand Russell objects that "the word ‘I’ is really illegitimate"; that Descartes should have, instead, stated "his ultimate premiss in the form ‘there are thoughts’." As Russell adds, "the word ‘I’ is grammatically convenient, but does not describe a datum." (1945, 567) Accordingly, "there is pain" and "I am in pain" have different contents, and Descartes is entitled only to the former.

On behalf of Descartes, it seems that such objections fail to consider fully the *subjective* character of experience. There surely *is* something more to the experience of pain than what "there is pain" conveys. The additional feature is the *subject* of the pain, a subjective character that minimally includes a *point-of-view*. If we take Descartes to be using ‘I’ to designate this subjective character of consciousness--but not *yet* designating anything like a substantial self (as Descartes' express qualifications indicate he intends)--then there is no question begging: rather than being smuggled in, the "I"-ness of consciousness turns out to be (contra Russell) a primary datum of experience. Taking the "I" in this minimalist, subjective sense, introspection reveals (as Descartes writes) that the various modes of thinking all seem to have "one and the same ‘I’," as their subject (Med. 2, AT 7:28). Of course, for all Descartes' meditator Knows (at this early stage of the inquiry), the "I" might well turn out to be indicative of nothing more than a Humean bundle, or perhaps even a committee of substantial selves. But whatever the eventual outcome, vis-à-vis the ontological status of the self, the "I" can be read as a placeholder for what *is* a primary datum of experience.

Fourth, the necessity of Descartes' claim, that "*I am, I exist*" is true "*whenever it is put forward* by me or

conceived in my mind," can be read as having a performatory aspect. There is no formal inconsistency in the suggestion that I might never have existed. But there is a conceptual repugnance, or what Hintikka calls an "existential inconsistency," in actually having or performing the thought that I--qua subject of my thoughts--do not exist. As such, the certainty of the *cogito* endures only as long as the performance: "for it could be," says the meditator, "that were I totally to cease from thinking, I should totally cease to exist" (Med. 2, AT 7:27).

Fifth, much of the debate over whether the *cogito* is intended to involve inference or simple intuition (roughly self-evidence), rests on one of two kinds of mistake--one interpretive, the other philosophic. The interpretive mistake is in focusing on the absence of an express '*ergo*' ('therefore') in the *Meditations* account: this is surely a mistake, since this is the one place (of his various published treatments of the *cogito*) where Descartes does his reader the favor of elaborating the line of inferential reflection that is left implicit in the comparatively elliptical accounts that include an express '*ergo*' (or '*donc*'). The philosophic mistake is in supposing that the *cogito* must either involve inference or intuition, but not both. There is no inconsistency in the view that the meditator comes to appreciate the persuasive force of the *cogito* by means of inferential reflection, while also holding that his eventual conviction is not grounded in inference. (A common theme among rationalists is that the genesis of assent need not serve as its ground.) Moreover, *what* one intuits might well include an inference (as is widely held of *modus ponens*).

Finally, insofar as the *cogito* does involve inference, many readers of Descartes have worried that the inference is invalid. It is far from clear, however, that a worry this strong is well-founded. In view of its unique (aforementioned) features, the *cogito* is not amenable to a straightforward characterization in first-order logic. Presumably, *formal* validity is a special case of a more general conception of validity (perhaps such, that the denial of a conclusion is inconceivable on the condition of a premise set); otherwise we'd have no means of identifying which forms were valid. That having been said, an inference may be valid even if it is not formally valid.

A final observation about the relation between the *cogito* and Descartes' search for unshakable foundations. We earlier observed (cf. §2) that the method of doubt is intended to complement the method of foundationalism by making possible the identification of bona fide first principles. This observation suggests that the foundations of Descartes' foundationalism need not be prima facie self-evident, a suggestion confirmed by his treatment of the *cogito*. If anything in the *Meditations* is prima facie self-evident, the *cogito* is. But, as the third paragraph of the Second Meditation reveals, Descartes intends that the *cogito* emerges from inferential reflection.

Further reading: See the second and third sets of Objections and Replies. See also Beyssade (1993), Hintikka (1962), and Markie (1992).

[Return to Section links]

5. Epistemic privilege

The method of doubt promises only to avoid error. Among Descartes' further desiderata is a *positive* criterion--a set of internal marks/criteria the occurrence of which guarantee truth. The *cogito* provides grounds for optimism, in that further reflection on its impressiveness might yield this desideratum.

In the present Section, we first consider Descartes' candidate truth criterion. We then consider the indirect sense in which even the privileged judgments grounded in this candidate criterion are defeasible by means of MCD. Finally, we confront the well discussed doctrine that judgments about the mind are epistemically privileged compared with those about body.

5.1. Our epistemic best: clear and distinct perception.

Descartes opens the Third Meditation by observing that the impressiveness of the *cogito* is owed to its being clearly and distinctly perceived (AT 7:35). Later texts indicate both an epistemic and a psychological benefit of such perception. While in the midst of clear and distinct attention, our evidence is *complete* (i.e. by the lights of our cognitive nature); the perceived matter is *understood* as true (cf. AT 7:36, 7:38, 7:56ff). Moreover, on the occasion of clear and distinct understanding, our assent is compelled--we "cannot but assent to these things, at least so long as" we continue to so perceive them (Med. 5, AT 7:65; cf. 7:36, 7:69). Elsewhere, Descartes defines clarity and distinctness:

> A perception which can serve as the basis for a certain and indubitable judgement needs to be not merely clear but also distinct. I call a perception "clear" when it is present and accessible to the attentive mind--just as we say that we see something clearly when it is present to the eye's gaze and stimulates it with a sufficient degree of strength and accessibility. I call a perception "distinct" if, as well as being clear, it is so sharply separated from all other perceptions that it contains within itself only what is clear. (*Prin.* 1.45, AT 8a:21-22)

He adds that "a concept is not any more distinct because we include less in it; its distinctness simply depends on our carefully distinguishing what we do include in it from everything else" (*Prin.* 1.63, AT 8a:31).

Since our evidentially best ground is clarity and distinctness, if anything will issue as the mark of truth, clarity and distinctness will. It thus emerges as Descartes' candidate truth criterion: "I now seem [*videor*] to be able to lay it down as a general rule that whatever I perceive very clearly and distinctly is true" (Med. 3, AT 7:35). I shall call this general criterion the C&D Rule. The announcement of the candidate criterion is carefully tinged with caution (*videor*), as the C&D Rule has yet to be subjected to hyperbolic doubt. Should it turn out that clarity and distinctness--as ground--is shakable, then, writes Descartes, being grounded in clarity and distinctness "would not be enough to make me certain of the truth of the matter" (ibid.).

5.2. The defeasibility of even our epistemic best.

Notwithstanding its evidential impressiveness, Descartes thinks that matters grounded in clarity and distinctness are defeasible and thus are not (yet) sufficiently warranted for Knowledge. Earlier, we introduced Meta-Cognitive Doubt (MCD). In the Fourth paragraph of the Third Meditation (a passage widely discussed by commentators), the doubt resistance of judgments grounded in clarity and distinctness is measured against MCD. Even these privileged judgments are defeasible, though only in an *indirect* way. During moments of clear and distinct attention, we're incapable of doubt; upon turning our attention away, we're able to entertain doubt by noticing the shakability of our confidence in the reliability of our cognitive equipment.

> But what about when I was considering something very simple and straightforward in arithmetic or geometry, for example that two and three added together make five, and so on? Did I not see at least these things clearly enough to affirm their truth? Indeed, the only reason for my *later* judgement that they were open to doubt was that it occurred to me that perhaps some God could have *given me a nature such that I was deceived even in matters which seemed most evident*. And whenever my preconceived belief in the supreme power of God comes to mind, I cannot but admit that it would be easy for him, if he so desired, to bring it about that I go wrong even in those matters which I think I see utterly clearly with my mind's eye. Yet *when I turn to the things themselves* which I think I perceive very clearly, I am so convinced by them that I spontaneously declare: let whoever can do so deceive me, he will never bring it about that I am nothing, so long as I continue to think I am something; or make it true at some future time that I have never existed, since it is now true that I exist; or bring it about that two and three added together are more or less than five, or anything of this kind in which I see a manifest contradiction. (Med. 3, AT 7:36; italics added)

On a very natural reading of this notoriously difficult passage, Descartes appears to be subjecting even the *cogito* to the indirect doubt invoked by MCD. The plausibility of this reading is bolstered by the very character of MCD: whether our epistemic best states are described as being "directly intuited by the mind's eye," or "revealed by the lights of our cognitive nature" (the *lumen naturale*), or "clearly and distinctly perceived," or the like, such cognitive states *are cognitive*; if the hypothesis of cognitive flaw renders even our cognitive best states subject to hyperbolic doubt, it would seem that the *cogito* is no exception. There is, however, considerable debate about this amongst commentators; indeed, it is widely held that the *cogito* is intended to be immune to *any* form of doubt.

Putting aside the debate about the *cogito*, there is little dispute that Descartes intends that *some* clearly and distinctly perceived matters are subject to MCD. The *indirect* manner of their defeasibility deserves further comment. There are two modes of doubting invoked by Descartes. In some cases, the perceiver may entertain doubt *while attending* to the dubious matter; e.g., I may doubt whether the tower in the distance is square or round, even while attending very carefully to my perception of it (cf. Med. 6, AT 7:76). Call this a *direct* doubt. In other cases, the perceiver is capable of doubt only when *re*directing attention *away* from the perceived matter; e.g., in order to undermine such matters as 2+3=5, the meditator (in the above passage) must turn his attention away from them--a redirecting of attention that

enables him to focus on his lack of confidence in the reliability of his cognitive faculties. Call this an *indirect* doubt.

At the end of §2, we noted that Descartes conceives Knowledge as fully indefeasible. This is to say that he conceives Knowledge as being immune to both direct and indirect doubts. *While* clearly and distinctly attending to a matter, it is not subject to direct doubt; such matters are, however, subject to indirect doubt, so long as one can perform Descartes' most hyperbolic doubt--MCD. Even though such matters are so evidentially impressive that we "see a manifest contradiction" in denying them, this is not yet sufficient warrant for Knowledge, if we're able to doubt the credibility of the cognitive equipment by means of which we apprehend such evidence. In closing the above passage (in the last block quote), Descartes concedes that his indirect doubt (MCD) is "a very slight and, so to speak, metaphysical" doubt, but he adds nonetheless that so long as we're uncertain as to whether a veracious God exists, as opposed to one who endows us with cognitive flaw, we "can never be quite certain about anything else" (AT 7:36).

What next? How are we to make epistemic progress if even our *epistemic best* is subject to hyperbolic doubt? Perhaps the one, seemingly hopeful note is that we're able--even in the face of hyperbolic doubt--to construct anti-sceptical arguments that are psychologically compelling for so long as they are clearly and distinctly perceived (they resist direct doubt). But these same anti-sceptical arguments are then subject to indirect doubt, by MCD, as soon as we turn our attention away from them. On one plausible reading of the anti-sceptical arguments of the Third and Fourth Meditations, Descartes intends that his meditator is "progressing" (by means of psychological compulsion) by attending clearly and distinctly to these arguments--attempting to demonstrate the existence of an all-perfect creator who guarantees the C&D Rule. On this interpretation, however, the anti-sceptical effort would appear to be both Sisyphean (since subject to later, indirect doubts) and circular (since relying on that which he attempts to establish, namely the criterion of clarity and distinctness). In §6 we return to a more detailed consideration of such problems. Before doing so, let's stray a bit from the project of philosophizing "in an orderly way" and consider a Cartesian doctrine that has received much attention in its subsequent history.

5.3. The epistemic privilege of judgments about the mind.

In our natural, pre-reflective condition, Descartes thinks we're apt to confound mind and body in a manner that obstructs our ability to perceive with clarity and distinctness. The confusion is clearly expressed (Descartes would say) in a 1925 essay written by G. E. Moore, "A Defence of Common Sense":

> I begin, then, with my list of truisms, every one of which (in my own opinion) I *know*, with certainty, to be true. ... There exists at present a living human body, which is *my* body. This body was born at a certain time in the past, and has existed continuously ever since ... But the earth had existed also for many years before my body was born ... (1962, 32-33)

In contrast, Descartes writes:

> [I]f I judge that the earth exists from the fact that I touch it or see it, this very fact undoubtedly gives even greater support for the judgement that my mind exists. For it may perhaps be the case that I judge that I am touching the earth even though the earth does not exist at all; but it cannot be that, when I make this judgement, my mind which is making the judgement does not exist. (*Prin*. 1.11, AT 8a:8-9)

The method of doubt is intended to help us appreciate the folly of the commonsensical position--assisting us in recognizing that the perception of our own minds is "prior to" and "more evident" than that of our own bodies. As the *cogito* is supposed to show, Knowledge claims (as Moore's) are subject to doubts to which claims about the existence of our own cogitation is immune. "Disagreement on this point," adds Descartes, comes from "those who have not done their philosophizing in an orderly way"; from those who, while properly acknowledging the "certainty of their own existence," mistakenly "take ‘themselves’ to mean only their bodies"--failing to "realize that they should have taken ‘themselves’ in this context to mean their minds alone" (*Prin*. 1.12, AT 8a:9). Reflection on reasons for doubt helps us ascertain that our own bodies are part of the external world--that our direct cognitive access reaches no further than states of consciousness.

We observed earlier (cf. §3) that a full sceptical diagnosis is intended to reveal the comparative shakability of various kinds of judgments, thereby exposing which kinds are most suitable as the building materials of Knowledge. By way of example, suppose there are two elevators available to you. You have grounds for doubting the integrity of the electrical wiring in the elevator on the left (but not in the elevator on the right); both elevators, however, have a cable of dubious integrity. Under these circumstances, the elevator on the right is clearly more conducive to safety (other things equal). An analogous point holds when comparing judgments about the body and those about the mind: the array of First Meditation doubts reveals that, though both kinds of judgments are subject to doubt at some level, judgments about the mind are subject to less kinds of doubt than those about the body. Our very best judgments of external sense are subject to direct doubt (by dreaming doubts) and to indirect doubt (by MCD); our very best judgments about the mind are subject only to indirect doubt.

This mind-is-better-known-than-body doctrine seems, for Descartes, to be motivated in part by his understanding of the new mechanistic science. As one of its pioneers, he gleaned epistemic consequences for our judgments about body from the mechanical causal story involved in sensory perception, as well as from the new mechanical conception of body. According to Descartes' theory of sensation, our physiological organs and nerves serve as mediating links that stand (spatially and causally) between the tables and limbs (and the like) that we perceive and the brain events that occasion our perceptual awareness of them (cf. *Prin*. 4.196). He thinks mechanism helps explain how non-veridical perception--e.g. dreams of colored tables, and pains in phantom limbs--can be occasioned by physiological processes largely similar to those that occasion veridical perception; indeed, he thinks the Similarity Thesis (cf. §3.1) has scientific support.

> [I]t is the soul which sees, and not the eye; and it does not see directly, but only by means of the brain. That is why madmen and those who are asleep often see, or think they see, various objects which are nevertheless not before their eyes: namely, certain vapours

> disturb their brain and arrange those of its parts normally engaged in vision exactly as they would be if these objects were present. (*Optics*, AT 6:141; cf. Med. 6., AT 7:83ff; *Passions* 26)

Moreover, according to the new science, bodies have no properties that resemble the qualitative features of our conscious experience of them (specifically, what would later be called secondary qualities, as e.g. the *color* "of" a table, and the *pain* "in" one's foot). The famous wax passage (at the end of the Second Meditation) is supposed to help show that our clear and distinct ideas of body include "none of the features" that we arrive at "by means of the senses" (AT 7:30). Sofar as internalist-relevant evidence is concerned, all that sensory experience avails us of is a glimpse of our own minds, a glimpse that largely misrepresents the real properties of external things.

Notwithstanding the undisputed role our nervous systems play in mediating our perception of tables and chairs (and even bracketing that sensory images of tables and chairs bear little resemblance to the supposed scientific image), there is considerable philosophical controversy concerning in what respects sense perception should be regarded as direct/immediate as opposed to indirect/mediated. Since, when looking through a window, say, at a tree, one's perception is *of* or *about* the tree rather than of/about the window, it seems one's perception is *intentionally* direct, and might remain direct even upon adding additional mediating panes of glass. Moreover, the same perception is likely to be *inferentially* direct, in that one's judgments about the tree do not become inferentially (more) complex in virtue of adding mediating panes of glass. Indeed, it would be misleading to characterize the judgment forming process as follows: "Aha, I see that the additional panes of glass are clean, and not tinted, and are thus not corrupting my perception of the tree." In view of such considerations, why would one hold that the mediation of our nervous systems is of epistemological importance?

One kind of reply open to Descartes lies in the evidentially relevant difference between the mediation of our nervous systems and that in our example of adding panes of glass: the former case, unlike the latter, is not such that we *could* compare our sensory perception of a tree both with and without such mediation in order to check for a corrupting influence. This means that the mediation of our nervous systems renders our sensory apprehension of tables and chairs as--essentially--*evidentially incomplete*. By means of our senses, we could never have Knowledge as to whether our sensory ideas accurately represent the external objects they purport to be of. The evidential problem thus far characterized holds even for materialist ontologies and is intensified with dualism. Some version or other of the mind-is-better-known-than-body doctrine is widely embraced in the 17th century.

On a slightly different note, it is generally overlooked that the mind-is-*better*-known-than-body doctrine is intended to convey a *comparative* rather than a *superlative* thesis. For Descartes, the only superlative state is that of clarity and distinctness: only it is correctly characterized as "our cognitive best"; and only it is regarded as a promising candidate for an infallible criterion of truth. Though the better-known doctrine entails that introspective judgments (i.e. those concerning the present contents of consciousness) are privileged, Descartes regards them as nonetheless subject to error. Even introspective perception--e.g. of occurrent pains and other sensations--must be rendered clear and distinct to be among our cognitive best. Such matters are clearly and distinctly perceivable, writes Descartes,

> provided we take great care in our judgements concerning them to include no more than what is strictly contained in our perception--no more than that of which we have inner awareness. But this is a very difficult rule to observe, at least with regard to sensations. (*Prin*. 1.66, AT 8a:32; cf. 1.68)

Indeed, we do "frequently make mistakes, even in our judgements concerning pain" (*Prin*. 1.67), since "people commonly confuse this perception [of pain] with an obscure judgement they make concerning the nature of something which they think exists in the painful spot and which they suppose to resemble the sensation of pain" (*Prin*. 1.46, AT 8a:22). Not only are introspective judgments error prone, introspective *reports* provide an additional source of error, since we may "attach our concepts to words which do not precisely correspond to real things" (*Prin*. 1.74, AT 8a:37).

Though Descartes is quite clear (as these texts show) as to the fallibility of introspective judgments, contemporary thinkers widely attribute to him a variety of related doctrines that he rejects. Compare the doctrines of the *infallibility* of the mental--roughly, the doctrine that sincere introspective judgments are always true; the *indubitability* of the mental--roughly, that sincere introspective judgments are indefeasible; and *omniscience* with respect to the mental--roughly, that one has Knowledge of every true proposition about one's own present contents of consciousness. (There is some variation in the way these doctrines are formulated in the literature.)

The widespread attribution of such doctrines to Descartes appears to stem from two *Meditations* texts in which he says, of examples of introspective claims, that they *cannot be false* strictly speaking. In each case, however, the larger context suggests a reconciliation of these claims with the fallibility conceded in the above cited passages. The first text (Med. 2) follows immediately on the heels of a reflective procedure of analysis intended to strip away every feature of introspection except that which is clearly and distinctly perceived--the very same procedure to which Descartes alludes, in *Principles* 1.66 (cf. the last block quote). Where the analysis has been run *successfully* (and in context we're to assume that the meditator has been successful), Descartes thinks that there is no room left for an appearance-reality gap: that we "*seem* to see, to hear, and to be warmed" (again, in the context of a successful stripping away) is what "having a sensory perception" just *is,* in the "restricted sense" (i.e. the *mental part* of sense perception) (AT 7:29). The second text that has misled interpreters (Med. 3) can be read as alluding to the very same stripping away procedure, and Descartes carefully adds the following qualification (as to our perception of such ideas): "provided they are considered solely in themselves and I do not refer them to anything else" (AT 7:37). So, in both texts the claimed epistemic privilege is conditional on successfully performing the requisite introspection. Not just any sincere effort at introspection will suffice: the internalist-relevant criterion of success is clarity and distinctness, not sincerity.

One final thought on the subject of epistemic privilege. As has been stressed, clear and distinct perception is *the* superlative epistemic state of privilege. But how are we to be sure that we're in a perceptual state of bona fide clarity and distinctness? As Gassendi complains, Descartes owes us a "method to guide us" as to "when we are mistaken and when not," when we suppose we "clearly and distinctly perceive something" (Objs. 5, AT 7:279). In his reply, Descartes acknowledges the need and adds: "I carefully

provided such a method in the appropriate place, where I first eliminated all preconceived opinions and afterwards listed all my principal ideas, distinguishing those which were clear from those which were obscure or confused" (AT 7:362).

Further reading: On discussions of *truth criteria* in the 16th and 17th centuries, see Popkin (1979). On Descartes' doctrine of ideas, see Chappell (1986), Hoffman (1996), Jolley (1990), and Nelson (1997). On the indirect doubt of matters perceived clearly and distinctly (including the *cogito*), see Newman and Nelson (1999). On contemporary treatments of infallibility, indubitability, and omniscience, see Alston (1989) and Audi (1993). For a historical anticipator of Descartes' mind-is-better-known-than-body doctrine, see Augustine's *Contra Academicos*.

[Return to Section links]

6. Cartesian Circle

Returning to the project of philosophizing "in an orderly way," we left the meditator (at the end of §5.2) seemingly unable to make epistemic progress in the face of hyperbolic doubt. He can construct compelling, anti-sceptical arguments, but they are subject to later, indirect doubt by means of MCD. In his *Principles* treatment, Descartes summarizes the problem:

> The mind, then, knowing itself, but still in doubt about all other things, looks around in all directions in order to extend its knowledge further. First of all, it finds within itself ideas of many things; and so long as it merely contemplates these ideas and does not affirm or deny the existence outside itself of anything resembling them, it cannot be mistaken. Next, it finds certain common notions from which it constructs various proofs; and, for as long as it attends to them, it is completely convinced of their truth. ... But it cannot attend to them all the time; and subsequently, when it happens that it remembers a conclusion without attending to the sequence which enables it to be demonstrated, recalling that it is still ignorant as to whether it may have been created with the kind of nature that makes it go wrong even in matters which appear most evident, the mind sees that it has just cause to doubt such conclusions, and that the possession of certain knowledge [*scientiam*] will not be possible until it has come to know the author of its being. (*Prin.* 1.13, AT 8a:9-10)

(Note: Translations typically obscure the fact that Descartes is careful to use different terms when referring to weaker forms of knowledge/knowing than when referring to the strong form which we're signaling with an uppercase ‘K’ (i.e. *scientia*). In the above passage, for instance, only in that occurrence of "knowledge"-talk that has been expressly marked does Descartes refer to *scientia* as opposed to some weaker concept as *cognitio*.)

Descartes' meditator thus constructs arguments in an effort to show that the "author of his being" did not endow him with a flawed cognitive nature--a seemingly hopeless effort, since these same proofs are

subject to indirect doubt by MCD. Included in the effort is Descartes' Third Meditation proof of the existence of God, and a further proof intended to establish that God guarantees the perfect reliability of the C&D Rule (Med. 4). Both of these proofs present difficult interpretive and philosophic problems that we'll not here consider. In the present treatment, we'll instead focus our attention on the broader line of reasoning which, quite famously, appears to involve circularity--the so-called Cartesian Circle. The apparent circle is defined by two arcs:

> (1) I am certain that God exists only because *I am certain of whatever I clearly and distinctly perceive*.

> (2) *I am certain of whatever I clearly and distinctly perceive* only because I am certain that God exists.

There is wide agreement amongst commentators that Descartes' procedure is not viciously circular in the manner suggested by a bald reading of (1) and (2). There is less agreement, however, both as to how Descartes avoids circularity, and as to what extent his procedure provides a satisfactory solution to the sceptical problem (this, even when granting the exceedingly permissive assumption that the component Third and Fourth Meditation proofs are themselves sound). It would be useful to isolate the issues about which there's fairly general agreement amongst commentators, from those about which there's not. Towards this end, let's first try to show why (1) and (2) misrepresent Descartes' procedure, by exposing an ambiguity in the italicized phrase--some such diagnosis is either explicit or implicit in much of the secondary literature. We'll then be in position to clarify the remaining interpretive problems on which commentators are divided, problems that arise in connection with the Third and Fourth Meditation proofs being subject to indirect doubt.

Both (1) and (2) involve an expression of confidence as to the credibility of clarity and distinctness: *I am certain of whatever I clearly and distinctly perceive*. Call this the Certainty Thesis. The Certainty Thesis turns out to be ambiguous between the following two claims:

> (3) I am certain that *p*, if I clearly and distinctly attend to *p* (and its proof, if any).

> (4) I am certain that the C&D Rule is perfectly reliable, if I clearly and distinctly attend to it and its proof.

The certainty expressed in (3) stems from the evidential impressiveness of clearly and distinctly understood matters (cf. §5.1): "my nature is such that so long as I perceive something very clearly and distinctly I cannot but believe it to be true" (Med. 5, AT 7:69). As such, this certainty does not depend on having established a divine guarantee of the perfect reliability of the C&D Rule; indeed, the certainty of atheist geometers stems from (3).

The claim in (4) is a substitution instance of (3). The consequent of (4) expresses a confidence in the C&D Rule as would result from carefully attending to a supporting demonstration--a demonstration that

Descartes claims an atheist could not produce. Such certainty is derivative of that expressed in (3). Accepting (3), however, does not presuppose a commitment to there being any such demonstration as would provide for the confidence expressed in (the consequent of) (4).

In the ambiguity of the Certainty Thesis lies the key to *dis*solving the Cartesian Circle. Descartes' procedure might at first appear to define the arcs of (1) and (2), but it actually unfolds (in part) as a subtle effort to use the certainty of (3) as a fulcrum on which to lever compelling anti-sceptical arguments in the pursuit of (4). To the extent that the two "arcs" (better: "stages") properly characterize the project, the express statement of the Certainty Thesis has a different sense in each stage. One *can* characterize the first stage by means of (1), so long as the Certainty Thesis is understood as referring to (3). So understood, (1) might be rewritten (though somewhat awkwardly):

> (1′) I am certain that God exists only because (I am certain that *p*, if I clearly and distinctly attend to *p* and its proof).

As for the second stage, one *can* characterize it by means of (2), so long as the express statement of the Certainty Thesis, there, is understood as referring to (4). So understood, (2) might be rewritten (again, somewhat awkwardly):

> (2′) (I am certain that the C&D Rule is perfectly reliable, if I clearly and distinctly attend to it and its proof) only because I am certain that God exists.

Since, according to Descartes, the C&D Rule rests on a divine guarantee, the meditator's eventual, demonstrative confidence in its perfect reliability is parasitic on earlier steps (in connection with (1′)) establishing the existence of God. What prevents circularity is that the proof of the C&D Rule (in connection with (2′)) is not intended to provide for (atheist-available) confidence stemming from (3), nor does the appeal to God (in (2′)) rest on the divinely guaranteed C&D Rule. Naturally, a preferred characterization of the two stages would not use numerically the same express language to signify two, importantly different certainty theses. Making the point by means of (1′) and (2′) amounts to an exercise showing that it is possible (even though misleading) to characterize Descartes' project by means of the express Certainty Thesis in (1) and (2): the two stages, as expressed in (1) and (2), would form a bona fide circle only if the Certainty Thesis had the same sense in both cases; their correct sense is represented by (1′) and (2′), but these do not define a circle.

Granting that Descartes' project is not circular in the manner of (1) and (2), significant problems remain. Since, as already noted, the anti-sceptical proofs (even if non-circular) remain subject to indirect doubt by means of MCD, it is unclear how these proofs are supposed to contribute to a final solution to the sceptical problem. Indeed, it is unclear how they are supposed to provide even for epistemic *progress*: prior to reflection on the proofs of God, clearly and distinctly perceived matters resist direct doubt while being vulnerable to indirect doubt; subsequent to reflection on these same proofs (and granting they are recognized as sound), it appears to remain the case that clearly and distinctly perceived matters resist direct doubt while being vulnerable to indirect doubt. To put the interpretive puzzle in terms of a

comparison that Descartes is fond of, it is unclear in what respect a theist geometer (one who's successfully worked through the *Meditations*) is epistemically advantaged over an atheist geometer. As we've thus far set-up the interpretive issues, it would appear that the theist has no epistemic advantage over the atheist insofar as the best proofs of each are vulnerable to indirect doubt. Yet Descartes is adamant that the atheist does not have Knowledge of the matters he clearly and distinctly perceives, precisely "since no act of awareness that can be rendered doubtful seems fit to be called knowledge [*scientia*]" (Replies 2, AT 7:141).

An obvious desideratum of an adequate interpretation is to avoid this unsavory result. What all existing interpretations share in common is the (implicit) assumption that unless the undermining effects of Descartes' indirect doubt (MCD) are somehow mitigated, there is in principle no way to have the fully indefeasible Knowledge he purports to attain; and no way to account for the alleged epistemic advantage enjoyed by the theist. A useful way of understanding the disagreement amongst interpreters is in terms of how/where they think the mitigation is to occur. The interpretive camps that have emerged rally around two kinds of strategies for mitigating the effects of MCD. Both kinds of strategy achieve a mitigation of MCD by *exempting* a category of judgments from its sceptical reach. The one strategy involves antecedent exemption, the other involves subsequent exemption.

(Note: There are many possible taxonomies of the views in the voluminous literature. I pick this particular taxonomy, since it best clarifies that none of the interpretive efforts thus far succeeds in attributing to Descartes what he should regard as an adequate response to the sceptic. For another recent taxonomy, see Loeb (1992).)

According to interpretations involving *antecedent exemption*, Descartes preempts at the outset the problem of indirect doubt by exempting from MCD a class of first principles. He then constructs his anti-sceptical demonstrations out of steps consisting of items from the exempted class. That each step in a demonstration is exempted from MCD is evidently thought to ensure that the resulting conclusion is also immune, thus allowing Descartes to use a foundationalist architecture to build up to Knowledge. Among the texts that are often cited on behalf of antecedent exemption is a Second Replies passage in which Descartes says, of very "simple" and "transparently clear" matters, including the *cogito* and other axioms, that "we cannot doubt them without at the same time believing they are true; that is, we can never doubt them" (AT 7:145-46)--suggesting that he regards these matters as immune even to indirect doubt.

According to interpretations involving *subsequent exemption*, Descartes initially allows that hyperbolic doubt *is* universal such that even simple axioms are subject to indirect doubt; the exemption then kicks-in later, at the level of superstructure. In short, subsequent to careful reflection on the theistic proofs, further contemplation of MCD no longer moves us to doubt. The claim is not that clear and distinct perception is now immune to indirect doubt. Rather, the interpretation seems to have it that hyperbolic worries (as MCD) are no longer regarded as being of epistemic significance. Interpreters in this camp are impressed by the following, late Fifth Meditation passage:

> Now, however, I have perceived that God exists, and at the same time I have understood that everything else depends on him, and that he is no deceiver; and I have drawn the

> conclusion that everything which I clearly and distinctly perceive is of necessity true. Accordingly, even if I am no longer attending to the arguments which led me to judge that this is true, as long as I remember that I clearly and distinctly perceived it, there are no counter-arguments which can be adduced to make me doubt it, but on the contrary I have true and certain knowledge of it. And I have knowledge not just of this matter, but of all the matters which I remember ever having demonstrated, in geometry and so on. (Med. 5, AT 7:70)

On one recent subsequent exemption account, Descartes is interested in anti-sceptical *reproducibility*: so long as the meditator retains an on-demand ability to reproduce the demonstration of the C&D Rule, it provides him all the anti-sceptical oomph that Descartes thinks one needs for Knowledge. On another subsequent exemption account, Descartes is interested in establishing a mere *reflective coherence* sufficient to induce a general confidence in the reliability of our cognitive equipment: prior to reflecting on the existence and nature of God, we lack this reflective coherence; subsequent to such reflection, we no longer need be troubled by the hyperbolic doubts of the First Meditation.

Significant problems face all exemption strategies. Such declarations of exemption (whether antecedent or subsequent) seem hopelessly arbitrary: there appears to be no principled basis on which to exclude a class of cognitive states (whether axioms or theorems) from a doubt so hyperbolic that it undermines the reliability of our cognitive equipment. Moreover, exemption strategies seem irreconcilable with Descartes' systematic insistence that there can be no atheistic Knowledge: antecedent exemption opens the door for an atheist meditator to build up to Knowledge from a foundation of exempted first principles (since these are equally available to him); subsequent exemption, in fudging on issues of indirect doubt, sets a precedent that arguably allows the atheist meditator to similarly fudge. Finally, exemption strategies are difficult to reconcile with Descartes' somewhat strident claims to being the first philosopher to have decisively refuted the sceptic on the sceptic's own terms--terms that Descartes clarifies as requiring immunity to indirect doubts.

Further reading: (Note: The foregoing discussion is in large part excerpted from Newman and Nelson (1999).) See the Fourth Replies for Descartes' express response to charges of circularity; see the Fifth Meditation, Second Replies, and the letter to Regius (24 May 1640), for texts on his intended, final solution to doubt. For examples of antecedent exemption accounts, see Kenny (1968), (Morris 1973), and Wilson (1978). For examples of subsequent exemption accounts, see DeRose (1992) and Loeb (1992), both of whom offer reproducibility accounts; Frankfurt (1970) and Sosa (1997a and 1997b), both of whom offer reflective coherence accounts; and Curley (1978 and 1993). For an interpretation that involves neither antecedent nor subsequent exemption, see Newman and Nelson (1999). For an anthology devoted largely to the Cartesian Circle, see Doney (1987). For a treatment of the Fourth Meditation contribution to the demonstration of the divine guarantee of the C&D Rule, see Newman (1999).

[Return to Section links]

7. Proving the existence of the external, material

world

Granting that Descartes has laid a theistic foundation for Knowledge--specifically, the divine guarantee of the C&D Rule--he attempts to expand his clear and distinct perception to encompass the existence of the external, material world. The attempt builds on a familiar strategy in the history of philosophy, an appeal to the involuntariness of sensory experience. Let's first consider the familiar strategy and its defect in the face of First Meditation doubt, and then consider Descartes' effort to repair it.

The familiar strategy turns on the following prima facie plausible inference: some of our sensory experiences come to us involuntarily; therefore, they are caused by things external to us.

> I know by experience that these ideas do not depend on my will, and hence that they do not depend simply on me. Frequently I notice them even when I do not want to: now, for example, I feel the heat whether I want to or not, and this is why I think that this sensation or idea of heat comes to me from something other than myself, namely the heat of the fire by which I am sitting. (Med. 3, AT 7:38)

In spite of its seeming plausibility, the inference falls prey to First Meditation doubt.

> Then again, although these [apparently adventitious] ideas do not depend on my will, it does not follow that they must come from things located outside me. Just as the impulses which I was speaking of a moment ago seem opposed to my will even though they are within me, so *there may be some other faculty not yet fully known to me, which produces these ideas without any assistance from external things;* this is, after all, just how *I have always thought ideas are produced in me when I am dreaming*. (Med. 3, AT 7:39; italics added)

The familiar inference presupposes exactly what is in dispute--namely, that our seemingly involuntary ideas are not caused by some hidden faculty of our minds, some unconscious mental component. But according to the Always Dreaming Doubt, what we take as normal, waking experiences might well be another kind of dreaming, whereby the conscious images of "waking" life are fictitious inventions of our minds; such images might well be produced by some unknown, immaterial (internal) faculty. (This unknown faculty hypothesis can also be motivated by the supposition of an evil genius who endowed us with cognitive equipment by which we're inevitably misled.) Thus, as a first step in repairing the familiar argument, Descartes thinks we need to refute the sceptical hypothesis that our involuntary sensory experiences are produced by an unknown, internal faculty.

By the time of the Sixth Meditation, Descartes has new premises at his disposal, premises which he thinks block the unknown faculty scenario. Among the metaphysical theses developed throughout the *Meditations* is that mind and body have distinct essences; that the essence of thinking substance is pure thought/consciousness/awareness, while the essence of body is pure extension. Writes Descartes, "nothing can be in me, that is to say, in my mind, of which I am not aware," and this "follows from the

fact that the soul is distinct from the body and that its essence is to think" (31 Dec. 1640, AT 3:273). This result makes possible the following line of reasoning. Since the essence of mind is awareness, it follows that there can be no hidden mental occurrences--no unconscious happenings of the mind. As such, if the cause of my seemingly involuntary sensory ideas were a faculty in me, I would be aware of this faculty on the occasion of its operations. But I am not aware of the operations of any such faculty. Thus, the cause of the ideas is not in me qua thinking thing. "I proved the existence of material things," says Descartes, "not from the fact that we have ideas of them but from the fact that these ideas come to us in such a way as to make us aware that they are not produced by ourselves" (August 1641, AT 3:428-29). The cause of such ideas

> cannot be in me, since clearly it presupposes no intellectual act [viz. a volition] on my part, and the ideas in question are produced without my cooperation and often even against my will. So the only alternative is that it is in another substance distinct from me ... (Med. 6, AT 7:79)

Granting Descartes his moves thus far, it still remains to be shown that this external cause of our involuntary sensory ideas is a body. There are three possible options. The external cause is (a) corporeal/material substance (just as it seems), (b) some other non-material creature, or (c) God.

> But since God is not a deceiver, it is quite clear that he does not transmit the ideas to me either directly from himself, or indirectly, via some creature ... For God has given me no faculty at all for recognizing any such source for these ideas; on the contrary, he has given me a great propensity to believe that they are produced by corporeal things. It follows that corporeal things exist. (Med. 6, AT 7:79-80)

The inference here is quite troubling to the interpreter: in view of the epistemic duty (emerging from the Fourth Meditation) that we withhold assent except when our perception is clear and distinct, Descartes would appear to be straying from those rigorous standards earlier enforced. The worry is apt. The above line of reasoning does not have the meditator clearly and distinctly perceiving straightaway that option (a) is correct. Rather, he appears to be drawing a roundabout conclusion on the basis of a perception with credentials significantly less impressive than clarity and distinctness; the perception doing the evidential work is said to yield a (mere) "great propensity" to judge that (a) is the correct option.

According to one interpretation (that we'll not here elaborate), Descartes is not in fact straying from his earlier standards. Rather, the Fourth Meditation demonstration of the C&D Rule proceeds by means of steps that license the inference that he is here making. On this reading, in the Fourth Meditation Descartes establishes a divine guarantee of the infallible truth of a judgment, that *p*, whenever (as the above text indicates):

> (i) I am positively inclined to assent to *p*; and
> (ii) I have no faculty/capacity for correction by which I could ascertain that not-*p*.

(There are hard questions, e.g. as to the nature on this inclination, that find no straightforward answers in the texts. As will emerge, Descartes again calls on this same inferential move in his effort to prove that he is not dreaming.)

Even granting this interpretation, potential problems loom. Since the procedure allows that mere propensity/inclination producing perceptions can result in Knowledge, it threatens to prove *too much*. Not only are we "inclined" to judge that corporeal things (in general) exist, we're also inclined to form particular existential judgments, as that *this* table and *that* chair exist; indeed, we're inclined to judge that such bodies have qualitative properties (as color, sound, taste, etc). But Descartes' own mechanist principles entail that bodies have only quantitative properties (primary qualities, as size, shape, motion, etc.). And on the assumption that I'm now dreaming (Descartes has yet to rebut the Now Dreaming Doubt) there might not even *be* any particular bodies quite like those now appearing to me. How, then, does Descartes prevent his argument from establishing these further results that he regards as unwarranted?

In the continuation of the last quoted passage, Descartes qualifies the conclusion of his argument:

> They [bodies] may not all exist in a way that exactly corresponds with my sensory grasp of them, for in many cases the grasp of the senses is very obscure and confused. But at least they possess all the properties which I clearly and distinctly understand, that is, all those which, viewed in general terms, are comprised within the subject-matter of pure mathematics. (Med. 6, AT 7:80)

Issues of interpretation are difficult here. Descartes may hold that the second of the two conditions above--(ii), concerning whether we have a faculty/capacity for correction--is what gives him the proper result without proving too much: since he thinks we do have the ability to ascertain that bodies aren't colored, our "inclination" to suppose they are colored is not conclusive; and since we do have the ability to ascertain whether we are awake (as he later discusses), we'd need to do so before forming Knowledge-worthy judgments as to the *particular* objects around us.

Further reading: See Descartes' *Prin.* 2.1 for a variation of his Sixth Meditation argument. On Descartes' treatment of the problem of the existence of the external, material world, see Friedman (1997), Garber (1992), and Newman (1994). On the respects in which the Sixth Meditation inference draws on Fourth Meditation work, see Newman (1999).

[Return to Section links]

8. Proving that one is not dreaming

As one nears the end of the *Meditations*, Descartes purports to have made astonishing anti-sceptical advances. None of these foregoing results, however, entails that the meditator is awake. It remains to be

shown whether Knowledge extends this far. Descartes confronts the Now Dreaming Doubt in the closing paragraph of the *Meditations*.

It is tempting to read him as offering a *wholly* naturalistic solution to the problem, in the form of a Continuity Test: since continuity with past experiences holds only for one's waking experiences but not for one's dream experiences, checking for the requisite continuity reveals whether one is awake. The following passage can be read as suggesting this litmus test:

> I now notice that there is a vast difference between [being asleep and being awake], in that dreams are never linked by memory with all the other actions of life as waking experiences are. ... But when I distinctly see where things come from and where and when they come to me, and when I can connect my perceptions of them with the whole of the rest of my life without a break, then I am quite certain that when I encounter these things I am not asleep but awake. (Med. 6, AT 7:89-90)

This "solution" prompts two obvious criticisms (both of which were raised by Hobbes, in the Third Objections). First, the solution runs contrary to Descartes' no-atheist-Knowledge thesis: since the Continuity Test involves no appeal to God, it appears, as Hobbes notes, "that someone *can* know he is awake without knowledge of the true God" (AT 7:196). Second, adds Hobbes, it seems one could *dream* the requisite continuity; i.e., one could "dream that his dream fits in with his ideas of a long series of past events," thus undermining the credibility of the Continuity Test (AT 7:195). Let's consider Descartes' reply to both objections.

As to the first, it turns out on closer inspection that the Continuity Test is used in conjunction with a theistic appeal. We saw earlier, in the proof of the external, material world, that Descartes invokes the following (divinely guaranteed) truth rule, namely that *p* is true whenever:

> (i) I am positively inclined to assent to *p*; and
> (ii) I have no faculty/capacity for correction by which I could ascertain that not-*p*.

In the final paragraph of the Sixth Meditation, while confronting the Now Dreaming Doubt, Descartes again appears to invoke this rule. The larger passage opens with the meditator discussing his various faculties for correcting sensory error.

> I can almost always make use of more than one sense to investigate the same thing; and in addition, I can use both my memory, which connects present experiences with preceding ones, and my intellect, which has by now examined all the causes of error. Accordingly, I should not have any further fears about the falsity of what my senses tell me every day; on the contrary, the exaggerated doubts of the last few days should be dismissed as laughable. This applies especially to ... my inability to distinguish between being asleep and being awake. (Med. 6, AT 7:89)

Referring to the worry (that he's dreaming) as exaggerated suggests that condition (i) is met--that he's "positively inclined" to judge that he is awake. As such, he need only to establish condition (ii) and he'll have a divine guarantee of being awake. Thus, says the meditator (speaking of sensory appearances),

> when I distinctly see where things come from and where and when they come to me, and when I can connect my perceptions of them with the whole of the rest of my life without a break, then I am quite certain that when I encounter these things I am not asleep but awake. And I ought not to have even the slightest doubt of their reality if, *after calling upon all the senses as well as my memory and my intellect in order to check them, I receive no conflicting reports from any of these sources*. For *from the fact that God is not a deceiver* it follows that in *cases like these* I am completely free from error. (Med. 6, AT 7:90; italics added)

Central to the inference is the meditator's effort to ascertain the correctness of the judgment towards which he is inclined, by means of his various faculties: the inclination is sufficient to warrant the judgment that he is awake, provided his faculties do not enable him to ascertain that he is instead dreaming. (Elsewhere Descartes remarks that, when "we are asleep and are aware that we are dreaming, we need imagination in order to dream, but to be aware that we are dreaming we need only the intellect"; Replies 5, AT 7:358-59.) The *cases like these* to which Descartes refers look to be those where conditions (i) and (ii) are both satisfied.

There are, then, multiple internalist criteria that Descartes discusses in connection with the Now Dreaming Doubt. Initially (in the First Meditation), the meditator draws the provisional, sceptical conclusion that "there are never any sure [internal] signs by means of which being awake can be distinguished from being asleep" (AT 7:19), a result of reflecting on the qualitative similarity between waking and dreaming (cf. the Similarity Thesis from §3.1). By the end of the Sixth Meditation, following an extended inquiry into dropsy-type error, he notices that we have an ability to cross-check various of our cognitive faculties in order to correct sensory errors that would have an adverse effect on the well-being of the body. This observation leads to the disclosure of the Continuity Test. Though the Continuity Test *is* available to an atheist meditator, this naturalistic criterion--alone--is insufficient to warrant that one is awake. As Descartes says to Hobbes, "an atheist can infer that he is awake on the basis of memory of his past life" (via the Continuity Test), but "he cannot know that this criterion is sufficient to give him the certainty that he is not mistaken, if he does not know that he was created by a non-deceiving God" (Replies 3, AT 7:196). The further internalist criteria to which Descartes thus appeals looks to be (i) and (ii); when combined with the Continuity Test, he thinks this further (divinely guaranteed) truth rule provides justification requisite for Knowledge.

True to form, then, Descartes' procedure does preclude atheistic Knowledge. There are, however, further questions as to how the procedure is supposed to work for the theist. This brings us to the second objection. As Descartes correctly notes, "a dreamer cannot really connect his dreams with the ideas of past events" in the successful manner required by the Continuity Test; yet, as Descartes concedes, the dreamer "may dream that he does" (Replies 3, AT 7:196). This reminds us of the more general problem that plagues internalist epistemologies: if it can *seem* to me that I'm in the requisite criterial state (e.g., the

state of recognizing that I'm passing the Continuity Test), even on occasions when I'm not actually in the requisite state (e.g., I'm merely *dreaming that* I am), how can I rely on how things seem as a guide to truth? (Cf. the objection from Gassendi that we considered at the very end of §5.3.) Unfortunately, Descartes does not elaborate in his reply to Hobbes, and it is not at all clear how an adequate response would go. It is clear, however, that Descartes recognizes that even theists face difficulties in implementing his procedure for determining that one is awake, and he adds the following caveat as his closing remark in the *Meditations*:

> But since the pressure of things to be done does not always allow us to *stop and make such a meticulous check*, it must be admitted that in this human life we are often liable to make mistakes about particular things, and we must acknowledge the weakness of our nature. (Med. 6, AT 7:90; italics added)

Further reading: See Newman (1999), Williams (1978), and Wilson (1978).

[Return to Section links]

Bibliography

Descartes works cited

- Abbreviations Used:

Rules	=	*Rules for the Direction of our Native Intelligence*
Discourse	=	*Discourse on Method*
Synopsis	=	Synopsis of the *Meditations*
Meditations	=	*Meditations on First Philosophy*
Med.	=	any one of the six *Meditations*
Objs./Replies	=	any of the seven sets of objections/replies that Descartes published along with the *Meditations*
Prin.	=	*Principles of Philosophy*
Passions	=	*The Passions of the Soul*
AT	=	*Oeuvres de Descartes*, edited by Adam and Tannery (see below)

  Dates in parentheses indicate a reference to Descartes' correspondence.

- Adam, Charles, and Paul Tannery, eds. 1904. *Oeuvres de Descartes*. Paris: J. Vrin. References are to volume number and page.
- All quoted texts are from the translation by John Cottingham, Robert Stoothoff, and Dugald Murdoch. 1984. *The Philosophical Writings of Descartes*. Cambridge: Cambridge University

Press.
- For full bibliographic information on Descartes' writings, see the entry on Descartes.

Other works cited.

- Alston, William. 1989. *Epistemic Justification*. Ithaca: Cornell University Press.
- Audi, Robert. 1993. *The Structure of Justification*. Cambridge: Cambridge University Press.
- Audi, Robert. 1999. "Doxastic Voluntarism and the Ethics of Belief." *Facta Philosophica* 1.
- Beyssade, Michelle. 1993. "Privileged Truth or Exemplary Truth?" In *Essays on the Philosophy and Science of René Descartes*, ed. Stephen Voss. Oxford: Oxford University Press.
- Bouwsma, O. K. 1949. "Descartes' Evil Genius." *Philosophical Review* 58:141-151.
- Chappell, Vere. 1986. "The Theory of Ideas." In *Essays on Descartes' Meditations*, ed. Amélie Oksenberg Rorty. Berkeley: University of California Press.
- Chisholm, Roderick M. 1982. *The Foundations of Knowing*. Minneapolis: University of Minnesota Press.
- Curley, E. M. 1978. *Descartes Against the Sceptics*. Cambridge, MA: Harvard University Press.
- Curley, E. M. 1986. "Analysis in the *Meditations*: The Quest for Clear and Distinct Ideas." In *Essays on Descartes' Meditations*, ed. Amélie Oksenberg Rorty. Berkeley: University of California Press.
- Curley, E. M. 1993. "Certainty: Psychological, Moral, and Metaphysical." In *Essays on the Philosophy and Science of René Descartes*, ed. Stephen Voss. Oxford: Oxford University Press.
- DeRose, Keith. 1992. "Descartes, Epistemic Principles, Epistemic Circularity, and *Scientia*." *Pacific Philosophical Quarterly* 73:220-38.
- Doney, Willis, ed. 1987. *Eternal Truths and the Cartesian Circle*. New York: Garland Publishing.
- Frankfurt, Harry G. 1970. *Demons, Dreamers, and Madmen*. Indianapolis: The Bobbs-Merrill Company.
- Friedman, Michael. 1997. "Descartes on the Real Existence of Matter." *Topoi* 16:153-162.
- Dunlop, Charles E. M, ed. 1977. *Philosophical Essays on Dreaming*. Ithaca: Cornell University Press.
- Euclid. 1956. *The Thirteen Books of Euclid's Elements*, ed. Thomas L. Heath. New York: Dover Publications.
- Galileo. 1967. *Dialogue Concerning the Two Chief World Systems*, tran. Stillman Drake. Berkeley: University of California Press.
- Garber, Daniel. 1986. "*Semel in vita*: The Scientific Background to Descartes' *Meditations*." In *Essays on Descartes' Meditations*, ed. Amélie Oksenberg Rorty. Berkeley: University of California Press.
- Garber, Daniel. 1992. *Descartes' Metaphysical Physics*. Chicago: University of Chicago Press.
- Gaukroger, Stephen. 1989. *Cartesian Logic: An Essay on Descartes's Conception of Inference*. Oxford: Clarendon Press.
- Hacking, Ian. 1980. "Proof and Eternal Truths: Descartes and Leibniz." In Descartes: *Philosophy, Mathematics and Physics*, ed. Stephen Gaukroger. Sussex: The Harvester Press.
- Hintikka, Jaakko. 1962. "*Cogito ergo sum*: Inference or Performance?" *Philosophical Review*, 71:3-32.

- Hintikka, Jaakko. 1978. "A Discourse on Descartes's Method." In *Descartes: Critical and Interpretive Essays*, ed. Michael Hooker. Baltimore: Johns Hopkins University Press.
- Hoffman, Paul. 1996. "Descartes on Misrepresentation." *Journal of the History of Philosophy*, 34 (July):357-381.
- Jolley, Nicholas. 1990. *The Light of the Soul: Theories of Ideas in Leibniz, Malebranche, and Descartes*. Oxford: Clarendon Press.
- Kenny, Anthony. 1968. *Descartes: A Study of His Philosophy*. New York: Random House.
- Loeb, Louis E. 1992. "The Cartesian Circle." In *The Cambridge Companion to Descartes*, ed. John Cottingham. Cambridge: Cambridge University Press.
- Markie, Peter. 1992. "The Cogito and Its Importance." In *The Cambridge Companion to Descartes*, ed. John Cottingham. Cambridge: Cambridge University Press.
- Moore, G. E. 1962. *Philosophical Papers*. New York: Collier Books.
- Morris, John. 1973. "Descartes' Natural Light." *Journal of the History of Philosophy*, 11:169-187.
- Nelson, Alan. 1997. "Descartes's Ontology of Thought." *Topoi* 16:163-178.
- Newman, Lex, and Alan Nelson. 1999 "Circumventing Cartesian Circles." *Noûs* 33:370-404.
- Newman, Lex. 1994. "Descartes on Unknown Faculties and Our Knowledge of the External World." *Philosophical Review* 103 (July):489-531.
- Newman, Lex. 1999 "The Fourth Meditation." *Philosophy & Phenomenological Research* 59 (September):559-591.
- Peirce, Charles. 1955. *Philosophical Writings of Peirce*, ed. Justus Buchler. New York: Dover Publications.
- Plantinga, Alvin. 1993. *Warrant: The Current Debate*. Oxford: Oxford University Press.
- Popkin, Richard H. 1979. *The History of Scepticism from Erasmus to Spinoza*. Berkeley: University of California Press.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge, Mass.: The Belknap Press.
- Reid, Thomas. 1785. *Essays on the Intellectual Powers of Man*.
- Russell, Bertrand. 1945. *A History of Western Philosophy*. New York: Simon and Schuster.
- Sosa, Ernest. 1997a. "How to Resolve the Pyrrhonian Problematic: A Lesson from Descartes." *Philosophical Studies* 85:229-49.
- Sosa, Ernest. 1997b. "Reflective Knowledge in the Best Circles." *Journal of Philosophy* 94 (August):410-430.
- Van Cleve, James. 1979. "Foundationalism, Epistemic Principles, and the Cartesian Circle." *Philosophical Review* 88 (January):55-91.
- Williams, Bernard. 1978. *Descartes: The Project of Pure Enquiry*. New Jersey: Humanities Press.
- Williams, Bernard. 1983. "Descartes's Use of Skepticism." In *The Skeptical Tradition*, ed. Myles Burnyeat. Berkeley: University of California Press.
- Williams, Michael. 1986. "Descartes and the Metaphysics of Doubt." In *Essays on Descartes' Meditations*, ed. Amélie Oksenberg Rorty. Berkeley: University of California Press.
- Wilson, Margaret Dauler. 1978. *Descartes*. London: Routledge & Kegan Paul.

Other Internet Resources

- Latin text of the *Meditations* (original version, 1641)
- French translation of the *Meditations* (by Duc de Luynes, 1647)
- English translation of the *Meditations* (by John Veitch, 1901)
- English translation of the *Discourse on Method* (by John Veitch, 1901)

Related Entries

appearance vs. reality | *a priori* justification and knowledge | certainty | Descartes, René | Descartes, René: modal metaphysics | Descartes, René: theory of sensation | idealism | idealism: British | ideas | innate ideas | justification, epistemic: foundationalist theories of | knowledge: analysis of | moral particularism | original position | primary and secondary qualities | rationalism vs. empiricism | realism | reasoning: defeasible | sense-data | skepticism | truth

Acknowledgements

I am grateful to Robert Audi, for commenting on an earlier version of this article, and to Alan Nelson for discussions on many of the ideas here and for permission to excerpt from our joint work, in 1999.

Lex Newman
lnewman@philosophy.utah.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 3, 1997
Content last modified: August 31, 1999

Descartes' Modal Metaphysics

Descartes sometimes speaks of things that are possible. He also speaks of eternal and necessary truths that are created by God. One of the interpretive projects that these claims inspire is the construction of a general Cartesian theory of modality. Any such theory of course needs to be sensitive to what Descartes says about possibility and necessity. However, what Descartes says in these instances sometimes appears to be in conflict with pillars of his larger system. For example, Descartes' dualistic ontology is very economical, and any entities that he posits must respect this economy. If he posits possibilities or necessities in a way that his parsimonious system does not allow, then he is helping himself to things to which he is not entitled. Also, Descartes holds that God is omnipotent. If this means that God is so powerful that He can make eternal truths false, then it is not at all clear how they could be necessary. There is an important interpretive issue about how we are to proceed in cases where Descartes makes claims that appear to conflict with his larger systematic commitments. Very generally, if Descartes makes claims about *X*, and if his claims about *X* appear to conflict with his systematic commitment to *Y*, we have a few options. One is that we can attempt to secure a reading of Descartes' claims about *X* that has them squaring with his commitment to *Y*. Another (presumably less attractive) option is that we can conclude that Descartes is not really committed to *Y* or that he does not mean what he says about *X*. Another is that we can interpret Descartes' commitment to *Y* as squaring with his claims about *X*. Descartes' views on modality are touched by many of his other views. Existing interpretations of Descartes on modality turn on how his claims about possibility and necessity are to be understood in light of his larger system.

1. The Analytic Method and Descartes' Views on

Modality

Though Descartes makes a number of claims about possibilities and necessities, not all of them are to be considered in determining his views on modality if we take seriously his view that when doing metaphysics we ought not affirm what we do not clearly and distinctly perceive. Descartes reveals his commitment to this view in a number of places:

> I should like you to remember here that, in matters which may be embraced by the will, I made a very careful distinction between the conduct of life and the contemplation of the truth. As far as the conduct of life is concerned, I am very far from thinking that we should assent only to what is clearly perceived. …But when we are dealing solely with the contemplation of the truth, surely no one has ever denied that we should refrain from giving assent to matters which we do not perceive with sufficient distinctness.[1]

If Descartes holds that when doing strict metaphysics we ought not speak of what we do not clearly and distinctly perceive, a general Cartesian theory of modality should not be sensitive to claims that Descartes makes about possibilities or necessities that are confused.

For reasons having to do with his method, Descartes still makes a number of such claims throughout his corpus. For example, in the First Meditation Descartes entertains a number of possibilities that suggest that our minds might be mistaken about results that seem perfectly evident to us. One such possibility is that we have been created by a supremely good God but that, for reasons unbeknownst to us, our nature is such that we "go wrong every time [we] add two and three or count the sides of a square, or in some even simpler matter" (AT 7:21, CSM 2:14). Another is that we have not been created by God but have "arrived at [our] present state by fate or chance or a continuous chain of events" (AT 7:21, CSM 2:14). A third is that we have been created by a malicious demon (AT 7:22, CSM 2:15). Descartes mentions all of these possibilities to set up his Third Meditation point that until we know what made our minds, we are not in a position to trust them and so not in a position to know anything at all (AT 7:36, CSM 2:25). However, none of these possibilities is a possibility that is a constituent of Descartes' ontology. In the Fifth Meditation, Descartes reveals that a person who has meditated to a sufficient grasp of God's nature understands self-evidently that God exists and did not create our minds defectively:

> …as regards God, if I were not overwhelmed by preconceived opinions, and if the images of things perceived by the senses did not besiege my thought on every side, I would certainly acknowledge him sooner and more easily than anything else. For what is more self-evident than the fact that the supreme being exists, or that God, to whose essence alone existence belongs, exists (AT 7:69, CSM 2:47).

For Descartes, a person who thinks that God is a deceiver or even that there is no such being as God is very confused.

It might seem surprising that Descartes flirts with these confusions at all. When he does so he is just employing his *analytic* method -- what he calls the "best and truest method of instruction…" (Second Replies, AT 7:156, CSM 2:111). Descartes does not hide that he sometimes makes claims early in the *Meditations* that from a later and more sophisticated point of view he will retract. He says,

> The analytic style of writing that I adopted there [in the *Meditations*] allows us from time to time to make certain assumptions that have not yet been thoroughly examined; and this comes out in the First Meditation where I made many assumptions which I proceeded to refute in subsequent Meditations.[2]

Descartes makes these claims because he is trying to teach his metaphysics and because he thinks that his readers will not be in a position to grasp that metaphysics if he only makes claims that are true.[3] Descartes holds that

> All of our ideas of what belongs to the mind have up till now been very confused and mixed up with the ideas of things that can be perceived by the senses. This is the first and most important reason for our inability to understand with sufficient clarity the customary assertions about the soul and God.[4]

In order to maneuver his readers into a position in which they *are* able to understand his metaphysics, Descartes needs to help them to clear up their ideas. If he simply tells us his view, we will hear it in terms of the confused ideas that (on his view) it is imperative we discard. The view that we would walk away with would not be Descartes' view but something else. Descartes therefore opts for a special strategy for presenting his metaphysics. If he is not going package his view by putting it forward unadorned, he will have to embellish it. If we object (as does Gassendi) to this kind of maneuver, Descartes insists that under the circumstances it is only appropriate. He says,

> A philosopher would be no more surprised at such suppositions of falsity than he would be if, in order to straighten out a curved stick, we bent it round in the opposite direction. The philosopher knows that it is often useful to assume falsehoods instead of truths in this way in order to shed light on the truth… (Fifth Replies, AT 7:349-350, CSM 2:242).

The falsehoods that Descartes assumes early in the *Meditations* include his claims about the possibility of God's deception and the possibility of His non-existence. But these are not real possibilities in Descartes' ontology. When Descartes talks about these alleged possibilities, he is not interested in truth. Rather, he is interested in helping his reader to abandon misconceptions. Descartes holds that when doing metaphysics, we ought refrain from making judgments about what we do not clearly and distinctly perceive. Along with anything else that we do not clearly and distinctly perceive, the alleged possibilities of God's deception and His non-existence do not belong in Descartes' ontology.

2. Unactualized Possibles

Descartes is unambiguous that there are some things that we clearly and distinctly perceive to be possible. To Mersenne he speaks of beings whose existence we clearly perceive to be possible (To Mersenne, March 1642, AT 3:544-545, CSMK 211); he also says that God "can bring about everything that I clearly and distinctly recognize as possible" (Fourth Replies, AT 7:219, CSM 2:154). In First Replies he says that

> …we must distinguish between possible and necessary existence. It must be noted that possible existence is contained in the concept or idea of everything that we clearly and distinctly understand; but in no case is necessary existence so contained except in the case of the idea of God.[5]

In addition, he says to Mesland that "our mind is finite and so created as to be able to perceive as possible things which God has wished to be in fact possible."[6] The interpretive issue here is what Descartes is talking about when he speaks in such terms. One interpretation that immediately suggests itself is that Descartes holds that there are things or states of affairs that, though not actual, are counterfactually possible.[7] A number of considerations speak in favor of such a reading. One is that Descartes' view that mind and body are really distinct appears to be the view that minds and bodies that are in fact united can exist apart.[8] Another is that Descartes says that we have clear and distinct perceptions of possible existence and of what is possible, thus making it appear that he holds that God's creatures include not only actuals but also unactualized possibles.[9] Finally, this sort of reading fits Descartes within a long tradition of figures like Scotus and Brandwardine who posit possible being to secure the meaning and reference of claims about things that could be but aren't.[10] If Descartes wants to make claims about unactualized possibles and if he does not want those claims to be non-sensical, unactualized possibles need to have some kind of ontological status in his system.

Descartes may in fact be committed to attributing reality to unactualized possibles, but a few interpretive problems arise if he does. One is that Descartes says elsewhere that unactualized being has no ontological status. In the Third Meditation, he argues that the objective reality of the idea of God cannot have been caused by potential perfections towards which a finite being might be evolving because potentialities are nothing and so have no causal power. He says,

> ... I perceive that the objective being of an idea cannot be produced merely by potential being, which strictly speaking is nothing, but only by actual or formal being. (AT 7:47, CSM 2:32)

If Descartes holds that potential being is nothing, then it is difficult to see how he can include unactualized possibles in his ontology.

Another interpretive problem is that it is difficult to see where possible reality would fit into Descartes' parsimonious dualistic ontology. For Descartes, possibles would be creatures, yet Descartes holds that the only creatures are finite minds and bodies and their modes:

> I recognize only two ultimate classes of things: first, intellectual of thinking things, i.e. those which pertain to mind or thinking substance; and secondly, material things, i.e. those which pertain to extended substance or body. Perception, volition and all the modes both of perceiving and willing are referred to thinking substance; while to extended substance belong size (that is, extension in length, breadth and depth), shape, motion, position, divisibility of component parts and the like.[11]

One category of being in Descartes' ontology is thinking substance (and its modes), and another is extended substance (and its modes). If Descartes' possibles are just created thinking or extended substances, then presumably they are actuals and not possibles. Of course, one might suggest that Descartes' dualism entails that there are two kinds of things -- thinking things and material things -- but that in each of these classes there are substances with actual existence and also substances with possible existence. On such a view, the class of thinking substances (for example) includes thinking substances with actual existence and thinking substances with possible existence. A problem with this view, though, is that Descartes cannot adhere to it if he also adheres to his theory of the conceptual distinction between a substance and its attributes. Descartes holds that all created substances have possible existence and that the existence of a substance is only conceptually distinct from that substance. He says,

> we do not sufficiently distinguish between things existing outside our thought and the ideas of things, which are in our thought. Thus, when I think of the essence of a triangle, and of the existence of the same triangle, these two thoughts, as thoughts, even taken objectively differ modally in the strict sense of the term 'mode'; but the case is not the same with the triangle existing outside thought, in which it seems to me manifest that essence and existence are in no way distinct. The same is the case with all universals. Thus, when I say that Peter is a man, the thought by which I think of Peter differs modally from the thought by which I think of man, but in Peter himself being a man is nothing other than being Peter. (To ***, 1645 or 1646, AT 7:350, CSMK 280-281)

Descartes holds that *in re* a thing's existence is identical to the thing itself. If all creatures have possible existence, then a creature that actually exists has possible existence. If the thing's existence is just identical to that thing itself, then the thing's possible existence is identical to the thing and the thing's actual existence is identical to the thing. That is, a thing's possible existence just is its actual existence.

If Descartes wants to distinguish between possible existence and actual existence, he must abandon his theory of conceptual distinction. If he retains this theory, he has to say that "possible existence" is just another name for actual existence. Descartes actually suggests this equation in a few places. To describe the kind of existence had by creatures, Descartes uses 'possible existence' and 'contingent existence' interchangeably: he sometimes speaks of the beings that depend on God for their existence as having "possible or contingent existence" (Second Replies, AT 7:166, CSM 2:117; *Notae*, AT 8B:361, CSM 1:306), and sometimes he speaks of them as having just "contingent existence" (*Principles* 15, AT 8A:10, CSM 1:198). When he says that creatures have possible *or* contingent existence, he identifies the two kinds of existence: "possible or [*vel*] contingent existence." If contingent existence is just the kind of

existence had by beings that depend for their existence on God's will, then the fact that a thing has possible existence in Descartes' ontology does not suggest that the thing does not actually exist. Such a thing exists, but in a way that has it wholly dependent on God. Descartes suggests exactly this definition of "possible existence" when he contrasts necessary existence with the kind of existence had by creatures in First Replies: unlike necessary existence, the existence had by a creature is marked by the fact that it "has no power to create itself or maintain itself in existence" (AT 7:118, CSM 2:84). It might be that like some of his predecessors Descartes is using 'possible existence' to describe a kind of being had only by actually existing things.[12]

A final worry is whether or not unactualized possibles are allowed by Descartes' commitment to divine simplicity. Descartes embraces the view that a perfectly simple God would have no distinct parts and concludes that

> In God willing and knowing are a single thing in such a way that by the very fact of willing something he knows it and it is only for this reason that such a thing is true. (To Mersenne, 6 May 1630, AT 1:149, CSMK 24)

A philosopher like Leibniz will insist on a distinction between God's understanding and will so as to secure the existence of things in God's understanding that God does not actually create. For Descartes, however, there is no such distinction, and whatever is the object of God's understanding is also the object of his will.[13] Descartes says,

> …in God, willing, understanding, and creating are all the same thing without one being prior to the other even conceptually. (To Mersenne, 27 May 1630, AT 1:152, CSMK 25-26)

What is *not* the object of God's understanding is nothing at all, and what is the object of God's understanding is created and made actual. If Descartes is seriously committed to the identity of God's intellect and will, it is difficult to see how he can also be committed to the existence of unactualized possibles.

There are a number of passages in which Descartes speaks of what is possible. To fix an interpretation of these passages, we can look to a number of different places. One is common-sense. We might argue that any view is crazy that does not admit that there are things that could happen or exist but that do not. Since Descartes is not crazy, what he must mean when he speaks of the possible is unactualized being.[14] Or, we might argue that Descartes is continuing the tradition of thinkers who clearly *do* posit unactualized possibles. If these figures include unactualized being within their ontologies, and if Descartes is building on their work, then (again) Descartes' claims about the possible are about unactualized being. Or, we might attempt to isolate parts of Descartes' system that have a bearing on what "possible" might mean in his system. Parts of this system entail that 'possible existence' is just the actual existence of dependent beings. If by "possible existence" Descartes just means the dependent existence of actually existing creatures, then passages in which Descartes speaks of a thing as being

possible or having possible existence are not evidence that Descartes holds that there are things that could be but are not.[15] Of course, it might just be the case that Descartes has reason to help himself to entities that the rest of his system shuts out.

If Descartes does hold that there are things that could be but are not, his view still demands an important qualification. Descartes of course realizes that in everyday discourse we speak of things that can happen but don't. However, if our understanding of these things is not clear and distinct, and if our understanding of them as possible is not clear and distinct, then Descartes will not introduce them as possibilities. Descartes appreciates that according to common ways of speaking, all kinds of things are possible. He considers this concept of 'possible' after Mersenne introduces it in Second Objections:

> If by 'possible' you mean what everyone commonly means, namely 'whatever does not conflict with our human concepts', then it is manifest that the nature of God, as I have described it, is possible in this sense…. (Second Replies, AT 7:150, CSM 2:107)

Here Descartes might seem to be offering a theory of possibility according to which what it means for something to be possible is just for it to be conceivable.[16] However, this cannot be Descartes' view. Descartes holds that whatever we clearly and distinctly perceive is true and that truth is "the conformity of thought with its object" (To Mersenne, 16 October 1639; AT 2:597, CSMK 139). If a possibility that we are considering is clearly and distinctly perceived, then our clear and distinct perception conforms to reality, and the possibility that we are conceiving is not merely conceptual. Instead, there is also an object to which the clear and distinct perception conforms -- the sort of thing posited by commentators who argue that Descartes holds that God's creation consists not only of actuals but of unactualized possibles. Thus, for any possibility of which we have a conception, if there is no object to which that conception conforms -- that is, if the possibility exists *only* in thought -- the possibility is not clearly and distinctly perceived. Possibilities which exist *only* in thought are not part of Descartes' ontology and so on Descartes' view are not possibilities at all. Descartes' remarks to Mersenne actually bear this out. Descartes is indeed considering the view of possibility as conceivability, but in doing so he is merely acknowledging what "everyone commonly means" by 'possible'. Descartes sometimes speaks of the possible as clearly and distinctly perceived. It is on these passages that any interpretation of Descartes' views on possibility must be built.

3. The Eternal Truths

Descartes is infamous for his doctrine of the divine creation of the eternal truths:

> You ask me by what kind of causality God established the eternal truths. I reply: by the same kind of causality as he created all things, that is to say, as their efficient and total cause. (To [Mersenne], 27 May 1630, AT 1:152, CSMK 25)

On the surface the position is baffling, especially when considered in conjunction with Descartes' view

that God is omnipotent. The author of Fifth Objections, Pierre Gassendi, complained that the view is very difficult to conceive. Descartes' reply is interesting:

> You say that you think it is 'very hard' to propose that there is anything immutable and eternal apart from God. You would be right to think this if I was talking about existing things, or if I was proposing something as immutable in the sense that its immutability was independent of God. But just as the poets suppose that the Fates were originally established by Jupiter, but that after they were established he bound himself to abide by them, so I do not think that the essences of things, and the mathematical truths which we can know concerning them, are independent of God. Nevertheless I do think that they are immutable and eternal, since the will and decree of God willed and decreed that they should be so. Whether you think this is hard or easy to accept, it is enough for me that it is true. (Fifth Replies, AT 7:380, CSM 2:261)

Descartes holds that each and every thing depends on God for its existence and that, as things, eternal truths depend on God as well. One of Gassendi's worries is that if God can do anything and thus can alter any item that He creates, nothing that He creates is immutable.

But this is not the only worry that arises with respect to Descartes' view on eternal truths. There is also a question about whether or not Descartes can account for the necessity that he attributes to them. In at least one place Descartes identifies eternal truths as necessary: he says that "the necessity of these truths does not surpass our knowledge" (To Mersenne, 6 May 1630; AT 1:150, CSMK 25). The worry here is that if they are necessary then it should not be the case that they could have been otherwise. Yet Descartes' commitment to divine omnipotence appears to commit him to this view:

> You ask what necessitated God to create these truths; and I reply that he was free to make it not true that all the radii of the circle are equal -- just as free as he was not to create the world. And it is certain that these truths are no more necessarily attached to his essence than are other created things. (To [Mersenne], 27 May 1630; AT 1:152, CSMK 25)

It appears that something in Descartes' comments about the eternal truths has to give. It might be that, since Descartes is clearly not prepared to adjust his commitment to divine omnipotence, he instead abandons his view that they are necessary.[17] On this view, all eternal truths are inherently contingent because they could have been false, and they could have been false because God could have made their contradictories true. This does not just follow from Descartes' commitment to divine omnipotence; there are also some texts:

> … God cannot have been determined to make it true that contradictories cannot be true together, and therefore… he could have done the opposite.[18]

> I do not think that we should ever say of anything that it cannot be brought about by God. For since every basis of truth and goodness depends on his omnipotence, I would not dare

> to say that God cannot make a mountain without a valley, or bring it about that 1 and 2 are not 3.[19]

On this reading, then Descartes does not really hold that the eternal truths are necessary. If to our rational faculties they appear to be necessary, this is just a function of the makeup of our rational faculties and not of the necessity of the truths themselves.[20]

If Descartes holds that eternal truths are necessary in any robust sense, the latter view has an obvious drawback. A second view is that Descartes holds that eternal truths are necessary, but not necessarily so. On this reading, Descartes' view involves iterated modalities: a number of truths are possibly necessary, but God chooses only some of these possibilities to be the actual necessary truths.[21] One of the passages that supports such a reading is from the already-cited letter to Mesland:

> And even if God has willed that some truths should be necessary, this does not mean that he willed them necessarily; for it is one thing to will that they be necessary, and quite another to will this necessarily, or to be necessitated to will it.[22]

On this reading, Descartes' view is that eternal truths are necessary, but they are not necessarily necessary.

An alternative reading of Descartes' comments on the eternal truths suggests still another interpretation. Jonathan Bennett has pointed out that in some of the key passages in question Descartes does not say that God can make contradictories true but that we should not say that God cannot make contradictories true (Bennett 1994, 653-655). Here Descartes is not saying anything about God's power but about us and what we ought not say. Presumably, Descartes is just invoking his Fourth Meditation rule for judging in these passages.[23] He is clear that when doing metaphysics we ought not affirm what we do not clearly and distinctly understand. Since we do not come close to a clear and distinct understanding of what it would be for one and two to add up to something other than three, and since we do not understand God's being unable to do something, the prospect that God cannot make 1 and 2 not add to 3 is hopelessly confused. Accordingly, we ought not speak of it. This analysis applies also to the important passage in the Mesland letter. Immediately after saying that God can make contradictories true together, Descartes takes it back: "… even if this be true, we should not try to comprehend it, since our nature is incapable of doing so." Descartes' claim that God can make contradictories true is something that we ought not affirm when doing metaphysics and thus something that should have no bearing on our interpretation of Descartes' system. Here Descartes may just be speaking in the language of faith and devotion in an attempt to gesture at God's perfection.

This last view handles very easily the passages in which Descartes says that we ought not say of impossibilities that God cannot bring them about. The view also squares nicely with the view that Descartes is an actualist. However, there are still passages in which Descartes says that God can bring about impossibilities (and not just that we ought not say that He cannot). For example, there is his claim to Mersenne that God was free to not make the radii of a circle equal. On the actualist reading, the

freedom of Descartes' God to not create eternal truths would have to reduce to His independence from all things.[24] Although Descartes does not come forward and actually state the Spinozistic view, it is interesting that he is reported as having stated it in *Conversation with Burman*:

> Concerning ethics and religion,… the opinion has prevailed that God can be altered, because of the prayers of mankind; for no one would have prayed to God if he knew, or had convinced himself, that God was unalterable…. From the metaphysical point of view, however, it is quite unintelligible that God should be anything but completely unalterable. It is irrelevant that the decrees could have been separated from God; indeed, this should not really be asserted. For although God is completely indifferent with respect to all things, he necessarily made the decrees he did, since he necessarily willed what was best, even though it was of his own will that he did what was best. We should not make a separation here between the necessity and the indifference that apply to God's decrees; although his actions were completely indifferent, they were also completely necessary. Then again, although we may conceive that the decrees could have been separated from God, this is merely a token procedure of our own reasoning: the distinction thus introduced between God himself and his decrees is a mental, not a real one. In reality the decrees could not have been separated from God: he is not prior to them or distinct from them, nor could he have existed without them.

If Descartes does hold that there are possible eternal truths that God does not actualize, it is difficult to see where they would fit in Descartes' ontology. It is also difficult to see how they square with some important pillars of Descartes' system. Descartes might be an actualist, and if he is he might be revealing this to Burman. Alternatively, Descartes might include unactualized eternal truths in his ontology. If he does posit such things, that might be evidence that he is not really so committed to his dualism or to the tenets that entail that potential being is strictly speaking nothing.

Thus far we have considered the interpretive issue of whether or not Descartes' eternal truths could have been otherwise. A question that still remains to be considered concerns the ontological status of Descartes' eternal truths, regardless of whether or not they could have been otherwise.

One view is that since Descartes' eternal truths are neither finite mental things, finite physical things, nor God, they must be something akin to Platonic forms.[25] A problem with this view, of course, is that it does violence to Descartes' parsimonious dualism. A second view is that eternal truths are to be located in God.[26] One of the merits of this view is that it provides for eternal truths to be eternal in a very robust sense. Since God is eternal, eternal truths are eternal presumably only if they are in God. Still, this view conflicts with the fact that Descartes holds that eternal truths are creatures.

A third reading of Descartes on eternal truths is that they are true ideas that conform to God's creation. In *Principles* I:48, Descartes says that eternal truths are beings which "have no existence outside our thought" (AT 8A:23, CSM 1:208). If eternal truths are truths that have no existence outside of our thought, one interpretive possibility is that they are simply true ideas. Since for Descartes truth is the

conformity of thought with its object, like any other true ideas eternal truths conform to God or His creation. A problem with this interpretation is that, although it is easy to see how it allows eternal truths to be true, it is difficult to see how it allows them to be eternal.[27] Descartes does allow that things can properly be called 'eternal' when they "are always the same" (Fifth Replies, AT 7:381, CSM 2:262). However, Descartes says to Mersenne that "from all eternity [God] willed them [eternal truths] to be, and by that very fact he created them" (To [Mersenne], 27 May 1630, AT 1:152, CSMK 25). Here Descartes appears to attribute to eternal truths an eternity that they cannot have if they have no existence outside of our thought.[28]

4. Real Distinction

Any interpretation of Descartes' views on modality needs to be sensitive to his view that mind and body are really distinct. The conclusion of his Sixth Meditation argument for substance dualism is that "I am really distinct from my body, and can exist without it" (AT 7:78, CSM 2:54). A natural reading of Descartes' conclusion has him saying that, for any minds and bodies that are united, it is counterfactually possible that they be separated.[29] His argument would be as follows:

1. I clearly and distinctly understand mind apart from body and body apart from mind.
2. God can bring about whatever I clearly and distinctly perceive.
3. God can bring about that mind is apart from body and body is apart from mind.
4. If God can bring about that mind is apart from body and body is apart from mind, then mind and body can exist apart.
5. Mind and body can exist apart.

Such a reading is sensitive not only to Descartes' Sixth Meditation comments but also to his further remarks on real distinction in Fourth Replies. There Descartes says that for "establishing a real distinction it is sufficient that two things can be understood as 'complete' and that each one can be understood apart from the other" (AT 7:221, CSM 2:156). A thing is complete, for Descartes, when it is a substance:

> … [B]y a 'complete thing' I simply mean a substance endowed with the forms or attributes which enable me to recognize that it is a substance. (AT 7:222, CSM 2:156)

Since a Cartesian substance is a thing that is ontologically independent (*Principles* I:51-52), a complete thing is an ontologically independent thing. When we clearly and distinctly perceive mind and body to be complete, we know that they are substances. When we still clearly and distinctly perceive them to be substances after clearly and distinctly perceiving them apart from each other, we know that they are not the same substance under different descriptions. On this view, Descartes holds that mind and body are ontologically independent substances, and their distinctness consists in their ability to continue to exist even after God separates them.[30]

An alternative interpretation of Descartes on the real distinction between mind and body reads the distinction as consisting in the the ontological independence of mind and body, but not in their separability.[31] Descartes holds that a sufficient condition for establishing a real distinction between two things is clearly and distinctly perceiving them to be non-identical substances (AT 7:13, CSM 2:9; AT 7:221-223, CSM 2:156-156). He therefore holds that the substantiality of two non-identical substances does not consist in their being separable but is just an indication of their separability.[32] On this view, mind and body are separable for Descartes; it's just that their separability is a consequence of the (different) fact that they are really distinct.[33]

One of the puzzles that the latter view addresses but that the earlier view does not is that Descartes holds that God brings about what we clearly and distinctly perceive but says (in his proof of real distinction) that God *can* bring about whatever we clearly and distinctly perceive. Presumably, if Descartes holds that our clear and distinct perceptions are veridical, a clear and distinct perception of mind apart from body should not tell us that mind and body *can* exist apart from each other. The latter view allows for the veridicality of our clear and distinct perception by understanding our clear and distinct perception of the apartness of mind and body as entailing their ontological independence. However, the view does not resolve the problem of why Descartes says that God can bring about whatever we clearly and distinctly perceive. How we proceed here is a function of the extent to which we appeal to the rest of Descartes' system in determining his views on a particular issue. The rest of Descartes' system entails that God has made things as we clearly and distinctly perceive them. Even more puzzling is that in Fourth Replies Descartes says that his reason for mentioning God's power in the Sixth Meditation proof of real distinction is to remind his reader that our clear and distinct perceptions are veridical (AT 7:226, CSM 2:159).

The interpretive issue of reading Descartes' particular views in light of his systematic commitments is especially pressing when it comes to his view that mind and body are really distinct. If Descartes holds that potential being is strictly speaking nothing, and if his parsimonious ontology and his commitment to divine simplicity bar possibilities from his system, it is not clear what to make of the possibility of mind and body existing in separation. One view might be that Descartes is committed to the reality of this possibility and that, with his commitment to other possibilities, this commitment is part of a pattern of Descartes' helping himself to entities that his system does not allow. That is, it might be the case that Descartes wants his system to be rich enough to posit possibilities and that, when it isn't, he posits them anyway. It also might be the case that Descartes' conclusion that mind and body can exist apart just reflects that when he draws the conclusion he has not yet proven that anything material exists. If he is adhering to his rule of affirming only what he clearly and distinctly perceives, he should not say that mind and body are actually apart -- that is, are actually substances -- until he has proven that body actually exists. The claim that mind and body can exist apart might be tracking just this.[34] His claim that God can make mind and matter into separate substances might just be a reminder to us that God has enough power to have done this even when it might seem impossible to us for such different substances to be united.

5. The System and its Pillars

It is presumably an adequacy condition on the interpretation of the work of any systematic philosopher that the work be interpreted in light of the central tenets of that philosopher's system. The interpretive problem of course is that for almost any such philosopher there are controversies about what these central tenets are. It is uncontroversial that Descartes sometimes speaks of possibility and necessity. What is not so uncontroversial is what he is talking about when he talks about these. In particular, if Descartes is committed to the view that there is unactualized possible being, then either (1) the alleged parts of his system that seem to disallow such being are not parts of his system, (2) the parts of his system that seem to disallow such being do not really disallow it, or (3) he is not a systematic philosopher.

Bibliography

- Adam, C. and Tannery, P., *Ouevres de Descartes*, Volumes I-XII, Paris: Vrin (1996)
- Alanen, L., "Descartes, Conceivability and Logical Modality," in Tamara Horowitz and Gerald J. Massey (eds.), *Thought Experiments in Science and Philosophy*, Savage, MD: Rowman and Littlefield (1991), 65-84.
- Bennett, J., "Descartes's Theory of Modality," *The Philosophical Review* 103 (1994).
- Chappell, V., "Descartes's Ontology," *Topoi* 16 (1997), 111-127.
- Cottingham, J., *The Rationalists* , Oxford and New York: Oxford University Press (1988).
- Cottingham, J., Stoothoof, R., and Murdoch, D., *The Philosophical Writings of Descartes, Volume II* , Cambridge: Cambridge University Press (1984).
- -----, *The Philosophical Writings of Descartes, Volume I* , Cambridge: Cambridge University Press (1985).
- Cottingham, J., Stoothoff, R., Murdoch, D., and Kenny, A., *The Philosophical Writings of Descartes, Volume III*, Cambridge: Cambridge University Press (1991).
- Cunning, D., "Systematic Divergences in Malebranche and Cudworth," *Journal of the History of Philosophy* (forthcoming 2002).
- -----, "True and Immutable Natures and Epistemic Progress in Descartes' *Meditations*," *British Journal for the History of Philosophy* (forthcoming 2003).
- -----, "Descartes on the Immutability of the Divine Will," *Religious Studies* (forthcoming 2003).
- Curley, E., *Descartes Against the Skeptics*, Cambridge, Harvard University Press (1978).
- ------, 1984, "Descartes on the Creation of the Eternal Truths," *The Philosophical Review* 93 (1984), 569-597.
- -----, "Analysis in the Meditations: The Quest for Clear and Distinct Ideas," in Amelie Oksenberg Rorty, ed., *Essays on Descartes' Meditations*, Berkeley: University of California Press (1986), 153-176.
- Frankfurt, H., "Descartes on the Creation of the Eternal Truths," *The Philosophical Review* 86 (1977), 36-57.
- Garber, D., " *Semel in Vita*: The Scientific Background to Descartes' *Meditations* ," in Rorty (1986), 81-116.
- Kenny, A., "The Cartesian Circle and the Eternal Truths," *Journal of Philosophy* 67 (1970) , 685-700.

- Gueroult, M., *Descartes' Philosophy Interpreted According to the Order of Reasons, Volume I*, Minneapolis: University of Minnesota Press (1984).
- Mattern, R., "Descartes: 'All Things Which I Conceive Clearly and Distinctly in Corporeal Objects are in Them," in Rorty (1986), 473-489.
- Nelson, A., "Cartesian Actualism in the Leibniz-Arnauld Correspondence," *Canadian Journal of Philosophy* 23 (1993), 675-694.
- Nelson, A., and Cunning, D., "Cognition and Modality in Descartes," *Acta Philosophica Fennica* 64 (1999), 137-153.
- Normore, C., "Meaning and Objective Being: Descartes and His Sources," in Rorty (1986), 223-241.
- -----, "Descartes's Possibilities," in Georges J. D. Moyal, *Rene Descartes: Critical Assessments , Volume III*, London and New York, Routledge (1991), 68-83.
- Rozemond, M., *Descartes's Dualism*, Cambridge: Harvard University Press (1998).
- Schmaltz, T., "Platonismand Descartes' View of Immutable Essences," *Archiv fur Geschichte der Philosophie* 73 (1991), 123-170.
- Wilson, M., *Descartes* , London and New York: Routledge (1978).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

actualism | Descartes, René: epistemology | Descartes, René: ontological argument | Leibniz, Gottfried Wilhelm: modal metaphysics | modality: medieval theories of | possible worlds

David Cunning
dcunning@niu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 15, 2002
Content last modified: April 22, 2002

Stanford Encyclopedia of Philosophy
Notes to Descartes' Modal Metaphysics

Notes

1. Second Replies, AT 7:149, CSM 2:106. This is of course just a restatement of Descartes' Fourth Meditation truth rule. See also Fifth Replies, AT 7:350-351, CSM 2:243. I use "CSM" to refer to the pagination in Cottingham, Stoothoff, and Murdoch 1984 and Cottingham, Stoothoff, and Murdoch 1985. I use "CSMK" to refer to the pagination in Cottingham, Stoothoff, Murdoch, and Kenny 1991. I use "AT" to refer to the pagination in Adam and Tannery 1996.

2. Fourth Replies, AT 7:249, CSM 2:173. See also Curley 1986; Garber 1986, 81-97; Cottingham 1988, 43-46; Cunning 2002, Cunning 2003a, and Cunning 2003b.

3. See Cunning 2003b.

4. Second Replies, AT 7:130-131, CSM 2:94. See also *Principles of Philosophy* I:73, AT 8A:37, CSM 1:220. Descartes says that because "there is nothing whose true nature we perceive by the senses alone, it turns out that most people have nothing but confused perceptions throughout their entire lives."

5. There are parallel passages in Second Replies at AT 7:163-164, CSM 2:115, and AT 7:166, CSM 2:117.

6. To [Mesland], 2 May 1644; AT 4:118, CSMK 235. See also "Conversation with Burman, 16 April 1648," AT 5:160, CSMK 343. The passages in the latter are tricky because they are Burman's reports of one of his discussions with Descartes.

7. See also Mattern 1984, 475, 486-487; and Alanen (1991).

8. Wilson 1978, 185-198. A discussion of Wilson's view is in section four.

9. Normore 1991, 68.

10. Normore 1991, 69-71, and Normore 1986, 224 and 231-234.

11. *Principles* I:48, AT 8A:23, CSM 1:208. Descartes mentions in the first line of this section of *Principles* that "eternal truths" are part of his ontology as well. These are beings which, as he puts it,

"have no existence outside our thought." Whatever they are, they fit nicely into Descartes' dualistic ontology as features internal to created minds. (See Cunning 2003b.) A discussion of eternal truths is in section three below.

12. See Nelson and Cunning 1999, 141-143, and Cunning 2003a.

13. See Nelson 1993, 685-688, and Nelson and Cunning 1999.

14. It should be noted that in the Early Modern period this view was not so unusual. Spinoza embraced a full-blown actualism, and it might be that in doing so he was just following Descartes. (See Cunning 2003b.)

15. It is interesting to note that after Descartes says to Mesland that "our mind is finite and so created as to be able to perceive as possible things which God has wished to be in fact possible," he immediately retreats. After flirting with this and other ideas concerning what God has made possible, Descartes says,

> … if we would know the immensity of his power we should not put these thoughts before our minds, nor should we conceive any precedence or priority between his intellect and his will; for the idea which we have of God teaches us that there is in him only a single activity, entirely simple and entirely pure. (AT 4:119, CSMK 235)

Here Descartes appears to be applying his actualism and saying that since God wills whatever he understands, the thought of an unactualized possible is incoherent and thus something of which we ought not speak when doing metaphysics.

16. See Bennett 1994, 647-649.

17. See Frankfurt 1977.

18. To [Mesland], 2 May 1644; AT 4:118, CSMK 235; Frankfurt 1977, 43.

19. For [Arnauld], 29 July 1648; AT 5:224, CSMK 358-359; Frankfurt 1977, 49.

20. Frankfurt 1977, 50-51.

21. Curley 1984, 579-503.

22. AT 4:118-119, CSMK 235; Curley 1984, 582.

23. See Nelson and Cunning 1999, 144-146.

24. See Nelson and Cunning 1999, 144-145, and Cunning 2003b.

25. Kenny 1970, 692-700.

26. Schmaltz 1991, 135.

27. Vere Chappell points this out in Chappell 1997, 123-27.

28. It may be that Descartes equivocates in his use of eternal. He holds that God exists from eternity and that He wills everything from eternity, yet he holds that eternal truths have no existence outside of our thought. His view might just be that eternal truths exist in non-eternal finite minds and are eternal in the sense that their truth values do not change. This reading squares with the passages (AT 1:145, CSMK 23; and AT 2:138, CSMK 103) in which Descartes distances himself from the view that eternal truths are eternal in any robust sense. See Cunning 2003b.

29. Wilson 1978, 185-198, and Curley 1978, 193-206. See also Gueroult 1984, 47-57.

30. Wilson 1978, 196-198.

31. Rozemond 1998, 28-37.

32. Rozemond 1998, 3-8.

33. Rozemond 1998, 28-37.

34. See Nelson and Cunning 1999, 148.

David Cunning
dcunning@niu.edu

First published: April 22, 2002
Content last modified: April 22, 2002

Descartes' Ontological Argument

Descartes' ontological (or *a priori*) argument is both one of the most fascinating and poorly understood aspects of his philosophy. Fascination with the argument stems from the effort to prove God's existence from simple but powerful premises. Existence is derived immediately from the clear and distinct idea of a supremely perfect being. Ironically, the simplicity of the argument has also produced several misreadings, exacerbated in part by Descartes' failure to formulate a single version.

The main statement of the argument appears in the Fifth Meditation. This comes on the heels of an earlier causal argument for God's existence in the Third Meditation, raising questions about the order and relation between these two distinct proofs. Descartes repeats the ontological argument in a few other central texts including the *Principles of Philosophy*. He also defends it in the First, Second, and Fifth Replies against scathing objections by some of the leading intellectuals of his day.

Descartes was not the first philosopher to formulate an ontological argument. An earlier version of the argument had been vigorously defended by St. Anselm in the eleventh century, and then criticized by a monk named Gaunilo (Anselm's contemporary) and later by St. Thomas Aquinas (though his remarks were directed against yet another version of the argument). Aquinas' critique was regarded as so devastating that the ontological argument died out for several centuries. It thus came as a surprise to Descartes' contemporaries that he should attempt to resurrect it. Although he claims not to be familiar with Anselm's version of the proof, Descartes appears to craft his own argument so as to block traditional objections.

Despite similarities, Descartes' version of the argument differs from Anselm's in important ways. The latter's version is thought to proceed from the meaning of the word "God," by definition, God is a being a greater than which cannot be conceived. Descartes' argument, in contrast, is grounded in two central tenets of his philosophy -- the theory of innate ideas and the doctrine of clear and distinct perception. He purports not to rely on an arbitrary definition of God but rather on an innate idea whose content is "given." Descartes' version is also extremely simple. God's existence is inferred directly from the fact that necessary existence is contained in the clear and distinct idea of a supremely perfect being. Indeed, on some occasions he suggests that the so-called ontological "argument" is not a formal proof at all but a self-evident axiom grasped intuitively by a mind free of philosophical prejudice.

Descartes often compares the ontological argument to a geometric demonstration, arguing that necessary existence cannot be excluded from idea of God anymore than the fact that its angles equal two right angles, for example, can be excluded from the idea of a triangle. The analogy underscores once again the

argument's supreme simplicity. God's existence is purported to be as obvious and self-evident as the most basic mathematical truth. It also attempts to show how the "logic" of the demonstration is rooted in our ordinary reasoning practices.

In the same context, Descartes also characterizes the ontological argument as a proof from the "essence" or "nature" of God, arguing that necessary existence cannot be separated from the essence of a supremely perfect being without contradiction. In casting the argument in these terms, he is implicitly relying on a traditional medieval distinction between a thing's essence and its existence. According to this tradition, one can determine what something is (i.e. its essence), independently of knowing whether it exists. This distinction appears useful to Descartes' aims, some have thought, because it allows him to specify God's essence without begging the question of his existence.

---

1. The Simplicity of the "Argument"

One of the hallmarks of Descartes' version of the ontological argument is its simplicity. Indeed, it reads more like the report of an intuition than a formal proof. Descartes underscores the simplicity of his demonstration by comparing it to the way we ordinarily establish very basic truths in arithmetic and geometry, such as that the number two is even or that the sum of the angles of a triangle is equal to the sum of two right angles. We intuit such truths directly by inspecting our clear and distinct ideas of the number two and of a triangle. So, likewise, we are able to attain knowledge of God's existence simply by apprehending that necessary existence is included in the clear and distinct idea of a supremely perfect being. As Descartes writes in the Fifth Meditation:

> [1] But if the mere fact that I can produce from my thought the idea of something entails that everything which I clearly and distinctly perceive to belong to that thing really does belong to it, is not this a possible basis for another argument to prove the existence of God? Certainly, the idea of God, or a supremely perfect being, is one that I find within me just as surely as the idea of any shape or number. And my understanding that it belongs to his nature that he always exists is no less clear and distinct than is the case when I prove of any shape or number that some property belongs to its nature (AT 7:65; CSM 2:45).

One is easily misled by the analogy between the ontological argument and a geometric demonstration,

and by the language of "proof" in this passage and others like it. Descartes does not conceive the ontological argument on the model of an Euclidean or axiomatic proof, in which theorems are derived from epistemically prior axioms and definitions. On the contrary, he is drawing our attention to another method of establishing truths that informs our ordinary practices and is non-discursive. This method employs intuition or, what is the same for Descartes, clear and distinct perception. It consists in unveiling the contents of our clear and distinct ideas. The basis for this method is the rule for truth, which was previously established in the Fourth Meditation. According to the version of this rule invoked in the Fifth Meditation, whatever I clearly and distinctly perceive to be contained in the idea of something is true of that thing. So if I clearly and distinctly perceive that necessary existence pertains to the idea of a supremely perfect being, then such a being truly exists.

Although Descartes maintains that God's existence is ultimately known through intuition, he is not averse to presenting formal versions of the ontological argument. He never forgets that he is writing for a seventeenth-century audience, steeped in scholastic logic, that would have expected to be engaged at the level of the Aristotelian syllogism. Descartes satisfies such expectations, presenting not one but at least two separate versions of the ontological argument. These proofs, however, are stunningly brief and betray his true intentions. One version of the argument simply codifies the psychological process by which one intuits God's existence, in the manner described above:

Version A:

1. Whatever I clearly and distinctly perceive to be contained in the idea of something is true of that thing.
2. I clearly and distinctly perceive that necessary existence is contained in the idea of God.
3. Therefore, God exists.

The rule for truth appears here in the guise of the first premise, but it is more naturally read as a statement of Descartes' own alternative method of "demonstration" via clear and distinct perception or intuition. In effect, the first "premise" is designed to instruct the meditator on how to apply this method, the same role that the analogy with a geometric demonstration serves in passage [1].

When presenting this version of the argument in the First Replies, Descartes sets aside this first premise and focuses our attention on the second. In so doing, he is indicating the relative unimportance of the proof itself. Having learned how to apply Descartes' alternative method of reasoning, one need only perceive that necessary existence pertains to the idea of a supremely perfect being. Once one attains this perception, formal arguments are no longer required; God's existence will be self-evident (Second Replies, Fifth Postulate; AT 7:163-4; CSM 2:115).

Descartes sometimes uses traditional arguments as heuristic devices, not merely to appease a scholastically trained audience but to help induce clear and distinct perceptions. This is evident for example in the version of the ontological argument standardly associated with his name:

Version B:

1. I have an idea of supremely perfect being, i.e. a being having all perfections.
2. Necessary existence is a perfection.
3. Therefore, a supremely perfect being exists.

While this set of sentences has the surface structure of a formal argument, its persuasive force lies at a different level. A meditator who is having trouble perceiving that necessary existence is contained in the idea of a supreme perfect being can attain this perception indirectly by first recognizing that this idea includes every perfection. Indeed, the idea of a supremely perfect being just is the idea of a being having all perfections. To attempt to exclude any or all perfections from the idea of a supremely being, Descartes observes, involves one in a contradiction and is akin to conceiving a mountain without a valley (or, better, an up-slope without a down-slope). Having formed this perception, one need only intuit that necessary existence is itself a perfection. It will then be clear that necessary existence is one of the attributes included in the idea of a supremely perfect being.

While such considerations might suffice to induce the requisite clear and distinct perception in the meditator, Descartes is aiming a deeper point, namely that there is a conceptual link between necessary existence and each of the other divine perfections. It is important to recall that in the Third Meditation, in the midst of the causal argument for the existence of God, the meditator already discovered many of these perfections -- omnipotence, omniscience, immutability, eternality, simplicity, etc. Because our mind is finite, we normally think of the divine perfections separately and "hence may not immediately notice the necessity of their being joined together" (First Replies, AT 7:119; CSM 2:85). But if we attend carefully to "whether existence belongs to a supremely perfect being, and what sort of existence it is" we shall discover that we cannot conceive any one of the other attributes while excluding necessary existence from it (ibid.).

To illustrate this point Descartes appeals to divine omnipotence. He thinks that we cannot conceive an omnipotent being except as existing. Descartes' illustration presupposes the traditional, medieval understanding of "necessary existence." When speaking of this divine attribute, he sometimes uses the term "existence" *simpliciter* as shorthand. But in his more careful pronouncements he always insists on the phrase "necessary and eternal existence," which resonates with tradition. Medieval, scholastic philosophers often spoke of God as the sole "necessary being," by which they meant a being who depends only on himself for his existence (*a se esse*). This is the notion of "aseity" or self-existence. Since such a being does not depend on anything else for its existence, he has neither a beginning nor an end, but is eternal. Returning to the discussion in the First Replies, one can see how omnipotence is linked conceptually to necessary existence in this traditional sense. An omnipotent or all-powerful being does not depend ontologically on anything (for if it did then it would not be omnipotent). It exists by its own power:

> [2] when we attend to immense power of this being, we shall be unable to think of its existence as possible without also recognizing that it can exist by its own power; and we

> shall infer from this that this being does really exist and has existed from eternity, since it is quite evident by the natural light that what can exist by its own power always exists. So we shall come to understand that necessary existence is contained in the idea of a supremely perfect being ... . (ibid.)

Some readers have thought that Descartes offers yet a third version of the ontological argument in this passage (Wilson, 1978, 174-76), but whether or not that was his intention is unimportant, since his primary aim, as indicated in the last line, is to enable his meditator to intuit that necessary existence is included in the idea of God. Since there is a conceptual link between the divine attributes, a clear and distinct perception of one provides a cognitive route to any of the others.

Although Descartes sometimes uses formal versions of the ontological argument to achieve his aims, he consistently affirms that God's existence is ultimately known through clear and distinct perception. The formal versions of the argument are merely heuristic devices, to be jettisoned once has attained the requisite intuition of a supremely perfect being. Descartes stresses this point explicitly in the Fifth Meditation, immediately after presenting the two versions of the argument considered above:

> [3] whatever method of proof I use, I am always brought back to the fact that it is only what I clearly and distinctly perceive that completely convinces me. Some of the things I clearly and distinctly perceive are obvious to everyone, while others are discovered only by those who look more closely and investigate more carefully; but once they have been discovered, the latter are judged to be just as certain as the former. In the case of a right-angled triangle, for example, the fact that the square on the hypotenuse is equal to the square on the other two sides is not so readily apparent as the fact that the hypotenuse subtends the largest angle; but once one has seen it, one believes it just as strongly. But as regards God, if I were not overwhelmed by philosophical prejudices, and if the images of things perceived by the senses did not besiege my thought on every side, I would certainly acknowledge him sooner and more easily than anything else. For what is more manifest than the fact that the supreme being exists, or that God, to whose essence alone existence belongs, exists? (AT 7:68-69; CSM 2:47)

Here Descartes develops his earlier analogy between the (so-called) ontological argument and a geometric demonstration. He suggests that there are some meditators for whom God's existence is immediately manifest; for them God's existence is akin to an axiom or definition in geometry, such as that the hypotenuse of a right triangle subtends its largest angle. But other meditators, whose minds are confused and mired in sensory images, must work much harder, and might even require a proof to attain the requisite clear and distinct perception. For them, God's existence is akin to the Pythagorean Theorem. The important point is that both kinds of meditators ultimately attain knowledge of God's by clearly and distinctly perceiving that necessary existence is contained in the idea of supremely perfect being. Once one has achieved this perception, God's existence will be manifest or, as Descartes says elsewhere, "self-evident" (*per se notam*) (Second Replies, Fifth Postulate; AT 7: 164; CSM 2:115).

Descartes' contemporaries would have been surprised by this last remark. While reviewing an earlier version of the ontological argument, Aquinas had rejected the claim that God's existence is self-evident, at least with respect to us. He argued that what is self-evident cannot be denied without contradiction, but God's existence can be denied. Indeed, the proverbial fool says in his heart "There is no God" (Psalm 53.1).

When confronted with this criticism by a contemporary objector, Descartes tries to find common ground: "St. Thomas asks whether existence is self-evident as far as we are concerned, that is, whether it is obvious to everyone; and he answers, correctly, that it is not" (First Replies, AT 7:115; CSM 2:82). Descartes interprets Aquinas to be claiming that God's existence is not self-evident *to everyone*, which is something with which he can agree. Descartes does not hold that God's existence is immediately self-evident, or self-evident to everyone, but that it can become self-evident to some careful and industrious meditators.

[Return to Section links]

2. The Distinction between Essence and Existence

In the Fifth Meditation and elsewhere Descartes says that God's existence follows from the fact that existence is contained in the "true and immutable essence, nature, or form" of a supremely perfect being, just as it follows from the essence of a triangle that its angles equal two right angles. This way of putting the *a priori* argument has puzzled commentators and has led to a lively debate about the ontological status of Cartesian essences and the objects which are purported to "have" them. Some commentators have thought that Descartes is committed to a species of Platonic realism. According to this view, some objects that fall short of actual existence nevertheless subsist as abstract, logical entities outside the mind and beyond the physical world (Kenny, 1968; Wilson, 1978). Another commentator places Cartesian essences in God (Schmaltz 1991), while two recent revisionist interpretations (Chappell, 1997; Nolan, 1997) read Descartes as a conceptualist who takes essences to be ideas in human minds.

Descartes' reference to "essences" raises another important issue more directly related to the ontological argument. In claiming that necessary existence cannot be excluded from the essence of God, Descartes is drawing on the traditional medieval distinction between essence and existence. According to this distinction, one can say *what* something is (i.e. its essence), prior to knowing whether it exists. So, for example, one can define what a horse is -- enumerating all of its essential properties -- before knowing whether there are any horses in the world. The only exception to this distinction was thought to be God himself, whose essence just is to exist. It is easy to see how this traditional distinction could be exploited by a defender of the ontological argument. Existence is included in the essence of a supremely perfect being, but not in the essence of any finite thing. Thus it follows solely from the essence of the former that such a being actually exists. At times, Descartes appears to support this interpretation of the ontological argument. In the Fifth Replies, for example, he writes that "the existence of a triangle should not be compared with the existence of God, since the relation between existence and essence is manifestly quite different in the case of God from what it is in the case of the triangle. God is his own existence, but this

is not true of the triangle" (AT 7:383; CSM 2:263). But Descartes' complete view is subtler and more sophisticated than these remarks first suggest. Understanding this view requires a more careful investigation of the distinction between essence and existence as it appears in medieval sources. Although one often speaks of the "traditional" distinction, the exact nature of the relation between essence and existence in finite things was the subject of a fierce debate among medieval philosophers. Seeing where Descartes' position fits within this debate will provide a deeper understanding of his version of the ontological argument.

The distinction between essence and existence can be traced back as far as Boethius in the fifth century. It was later developed by Islamic thinkers such as Avicenna. But the issue did not become a major philosophical problem until it was taken up by Aquinas in the thirteenth century. The issue arose not as part of an effort to establish God's existence on *a priori* grounds (as mentioned above, Aquinas was one of the staunchest critics of the ontological argument), but out of concern to distinguish God from finite spiritual entities such as angels. Like many scholastic philosophers, Aquinas believed that God is perfectly simple and that created beings, in contrast, have a composite character that accounts for their finitude and imperfection. In the case of purely spiritual or nonmaterial creatures, he attempted to locate this character in the composition of essence and existence.

Some of the details of Aquinas' account will emerge from our discussion below. The primary interest of his theory for our purposes, however, is that it led to a lively debate among his successors both as to how to interpret the master and about the true nature of the relation between essence and existence in created things. This debate produced three main positions:

1. The Theory of Real Distinction
2. The Intermediate Position
3. The Theory of Rational Distinction

Proponents of the first view conceived the distinction between essence and existence as obtaining between two separate things. In the eyes of many Thomists, this view was considered to be quite radical, especially as an interpretation of Aquinas' original position. The latter is sometimes expressed by saying that essence and existence are "principles of being" rather than beings themselves. One problem with the theory of real distinction then was that it reified essence and existence, treating them as real beings in addition to the created entity that they compose.

The theory of real distinction was also considered objectionable for philosophical reasons. Following Aquinas, many participants in the debate urged that essence and existence are related to each other as potency and act, so that existence can be said to "actualize" essence. On the theory of real distinction, this view leads to an infinite regress. If an essence becomes actual only in virtue of something else -- *viz.* existence -- being superadded to it, then what gives existence its reality, and so on *ad infinitum*? (Wippel, 1982, 393f).

In response to these difficulties some scholastic philosophers developed a position at the polar extreme

from the theory of real distinction. This was the view that there is merely a rational distinction or a "distinction of reason" between essence and existence in created beings. As the term suggests, this theory held that essence and existence of a creature are identical in reality and distinguished only within our thought by means of reason. Needless to say, proponents of this theory were forced to distinguish purely spiritual entities from God on grounds other than real composition.

Giving up the doctrine of real composition seemed too much for another group of thinkers who were also critical of the theory of real distinction. This led to the development of a number of intermediate positions. One such position was that essence and existence are modally or formally distinct, such that existence constitutes a mode or property of a thing's essence.

Like Francisco Suárez, his most immediate scholastic predecessor, Descartes sides with the proponents of a rational distinction between essence and existence. His position is unique, however, insofar as it springs from a more general theory of "attributes". Articulating this theory in an important passage in the *Principles of Philosophy*, Descartes claims that there is merely a distinction of reason between a substance and any one of its attributes or between any two attributes of a single substance (1:62, AT 8A:30; CSM 1:214). For Descartes' purposes, the most significant instance of a rational distinction is that which obtains between a substance and its essence -- or what he sometimes refers to as its "principal attribute" (1:53, AT 8A:25; CSM 1:210). Since thought and extension constitute the essence of mind and body, respectively, a mind is merely rationally distinct from its thinking and a body is merely rationally distinct from its extension (1:63, AT 8A:31; CSM 1:215). But Descartes insists that a rational distinction also obtains between any two attributes of a substance. Since existence qualifies as an attribute in this technical sense, the essence and existence of a substance are also distinct merely by reason (1:56, AT 8A:26; CSM 1:211). Descartes reaffirms this conclusion in a letter intended to elucidate his account of the relation between essence and existence:

> [4]... existence, duration, size, number and all universals are not, it seems to me, modes in the strict sense ... . They are referred to by a broader term and called attributes ... because we do indeed understand the essence of a thing in one way when we consider it in abstraction from whether it exists or not, and in a different way when we consider it as existing; but the thing itself cannot be outside our thought without its existence ... . Accordingly I say that shape and other similar modes are strictly speaking modally distinct from the substance whose modes they are; but there is a lesser distinction between the other attributes ... . I call it a rational distinction ... . (To an unknown correspondent, AT 4:349; CSMK 3:280)

Indications are given here as to how a rational distinction is produced in our thought. Descartes explains that we regard a single thing in different abstract ways. Case in point, we can regard a thing as existing, or we can abstract from its existence and attend to its other aspects. In so doing, we have distinguished the existence of a substance from its essence within our thought. Like scholastic proponents of the theory of rational distinction, however, Descartes is keen to emphasize that this distinction is purely conceptual. Indeed, he goes on to explain that the essence and existence of a substance are "in no way distinct" outside thought (AT 4:350; CSMK 3:280). In reality they are identical.

While borrowing much from scholasticism, Descartes' account is distinguished by its scope of application. He extends the theory of rational distinction from created substances to God. In general, the essence and the existence of a substance are merely rationally distinct, and hence identical in reality.

This result appears to wreak havoc on Descartes' ontological argument. One of the most important objections to the argument is that if it were valid, one could proliferate such arguments for all sorts of things, including beings whose existence is merely contingent. By supposing that there is merely a rational distinction between essence and existence abroad in all things, Descartes seems to confirm this objection. In general, a substance is to be identified with its existence, whether it is God or a finite created thing.

The problem with this objection, in this instance, is that it assumes that Descartes locates the difference between God and creatures in the relation each of these things bears to its existence. This is not the case. In a few important passages, Descartes affirms that existence is contained in the clear and distinct idea of every single thing, but he also insists that there are different grades of existence:

> [5] Existence is contained in the idea or concept of every single thing, since we cannot conceive of anything except as existing. Possible or contingent existence is contained in the concept of a limited thing, whereas necessary and perfect existence is contained in the concept of a supremely perfect being (Axiom 10, Second Replies; AT 7:166; CSM 2:117).

In light of this passage and others like it, we can refine the theory of rational distinction. What one should say, strictly speaking, is that God is merely rationally distinct from his *necessary* existence, while every finite created thing is merely rationally distinct from its *possible* or *contingent* existence. The distinction between possible or contingent existence on the one hand, and necessary existence on the other, allows Descartes to account for the theological difference between God and his creatures.

Now, when Descartes says that a substance (be it finite or infinite) is merely rationally distinct from its existence, he always means an actually existing substance. So how are we to understand the claim that a finite substance is merely rationally distinct from its *possible* existence? What is meant by "possible (or contingent) existence"? It is tempting to suppose that this term means non-actual existence. But as we saw already with the case of necessary existence, Descartes does not intend these terms in their logical or modal senses. If "necessary existence" means ontologically independent existence, then "possible existence" means something like *dependent* existence. After all, Descartes contrasts possible existence not with actual existence but with necessary existence in the traditional sense. This account is also suggested by the term "contingent." Created things are contingent in the sense that they depend for their existence on God, the sole independent being.

This result explains why Descartes believes that we cannot proliferate ontological arguments for created substances. It is not that the relation between essence and existence is any different in God than it is in finite things. In both cases there is merely a rational distinction. The difference is in the grade of

existence that attaches to each. Whereas the concept of an independent being entails that such a being exists, the concept of a finite thing entails only that it has dependent existence.

Looking back at the problematic passage cited above from the Fifth Replies, it becomes clear that Descartes intended something along these lines even there. He says that "the existence of a triangle should not be compared with the existence of God", reinforcing the point that it is the kind of existence involved that makes God unique. And just before this statement, he writes, "in the case of God necessary existence...applies to him alone and forms a part of his essence as it does of no other thing". Later he adds: "I do not ... deny that possible existence is a perfection in the idea of a triangle, just as necessary existence is perfection in the idea of God" (AT 7:383; CSM 2:263). Descartes' final position then is that essence and existence are identical in all things. What distinguishes God from creatures is his grade of existence. We can produce an ontological argument for God, and not for finite substances, because the idea of a supremely perfect being uniquely contains necessary -- or ontologically independent -- existence.

[Return to Section links]

3. Objections and Replies

Because of its simplicity, Descartes' version of the ontological argument is commonly thought to be cruder and more obviously fallacious than the one put forward by Anselm in the eleventh century. But when the complete apparatus of the Cartesian system is brought forth, the argument proves itself to be quite resilient, at least on its own terms. Indeed, Descartes' version is superior to his predecessor's insofar as it is grounded in a theory of innate ideas and the doctrine of clear and distinct perception. These two doctrines inoculate Descartes from the charge made against Anselm, for example, that the ontological argument attempts to define God into existence by arbitrarily building existence into the concept of a supremely perfect being. In the Third Meditation, the meditator discovers that her idea of God is not a fiction that she has conveniently invented but something native to the mind. As we shall see below, these two doctrines provide the resources for answering other objections as well.

Given our earlier discussion concerning the non-logical status of the ontological argument, it may seem surprising that Descartes would take objections to it seriously. He should be able to dismiss most objections in one neat trick by insisting on the non-logical nature of the demonstration. This is especially true of objection that the ontological argument begs the question. If God's existence is ultimately self-evident and known by a simple intuition of the mind, then there are no questions to be begged. Unfortunately, not all of the objections to the ontological argument can be dismissed so handily, for the simple reason that they do not all depend on the assumption that we are dealing with a formal proof.

Although it is often overlooked, many of the best known criticisms of the ontological argument were put to Descartes by official objectors to the *Meditations*. He in turn responded to these objections -- sometimes in lengthy replies -- though many contemporary readers have found his responses opaque and unsatisfying. We can better understand his replies and, in some cases, improve upon them by appealing

to discussions from previous sections.

One classical objection to the ontological argument, which was first leveled by Gaunilo against Anselm's version of the proof, is that it makes an illicit logical leap from the mental world of concepts to the real world of things. The claim is that even if we were to concede that necessary existence is inseparable from the idea of God (in Kant's terms, even if necessary existence were analytic of the concept "God"), nothing follows from this about what does or does not exist in the actual world. Johannes Caterus, the author of the First Set of Objections to the *Meditations*, puts the point as follows:

> [6] Even if it is granted that a supremely perfect being carries the implication of existence in virtue of its very title, it still does not follow that the existence in question is anything actual in the real world; all that follows is that the concept of existence is inseparably linked to the concept of a supreme being. So you cannot infer that the existence of God is anything actual unless you suppose that the supreme being actually exists; for then it will actually contain all perfections, including the perfection of real existence (AT 7:99; CSM 2:72).

To meet this challenge, Descartes must explain how he "bridges" the inferential gap between thought and reality. The principle of clear and distinct perception is intended to do just that. According to this principle, for which he argues in the Fourth Meditation, whatever one clearly and distinctly perceives or understands is true -- true not just of ideas but of things in the real world represented by those ideas. Thus, Descartes' commitment to the principle of clear and distinct perception allows him to elude another objection that had haunted Anselm's version of the argument.

The previous objection is related to another difficulty raised by Caterus. In order to illustrate that the inference from the mental to the extra-mental commits a logical error, critics have observed that if such inferences were legitimate then we could proliferate ontological arguments for supremely perfect islands, existing lions, and all sorts of things which either do not exist or whose existence is contingent and thus should not follow *a priori* from their concept. The trick is simply to build existence into the concept. So while existence does not follow from the concept of lion *per se*, it does follow from the concept of an "existing lion."

Descartes' actual reply to this objection, which he took very seriously, is highly complex and couched in terms of a theory of "true and immutable natures." We can simplify matters by focusing on its key elements. One of his first moves is to introduce a point that we discussed earlier **(**see passage [5] in section 2), namely that existence is contained in the idea of every thing that we clearly and distinctly perceive: possible (or dependent) existence is contained in our clear and distinct idea of every finite thing and necessary (or independent) existence is uniquely contained in the idea of God (AT 7:117; CSM 2:83). So for Descartes one does not have to build existence into the idea of something if that idea is clear and distinct; existence is already included in every clear and distinct idea. But it does not follow that the thing represented by such an idea actually exists, except in the case of God. We cannot produce ontological arguments for finite things for the simple reason that the clear and distinct ideas of them

contain merely dependent existence. Actual existence is demanded only by the idea of God, which uniquely contains independent existence.

A natural rejoinder to this reply would be to ask about the idea of a lion having not possible but wholly *necessary* existence. If Descartes' method of reasoning were valid, it would seem to follow from this idea that such a creature exists. This formulation of the objection requires Descartes' second and deeper point, which is only hinted at in his official reply. This is that the idea of a lion -- let alone the idea of a lion having necessary existence -- is hopelessly obscure and confused. As Descartes says, the nature of a lion is "not transparently clear to us" (Axiom 10, Second Replies; AT 7:117; CSM 2:84). Since this idea is not clear and distinct, the method of demonstration employed in the ontological argument does not apply to it. Recall that the geometrical method of demonstration is grounded in the principle of clear and distinct perception and consists in drawing out the contents of our clear and distinct ideas. If an idea is not clear and distinct then we cannot draw any conclusions from it about things outside thought.

The key difference then between the idea of God on the one hand and the idea of a necessarily existing lion is that the former can be clearly and distinctly perceived. For Descartes, it is just a brute fact that certain ideas can be clearly and distinctly perceived and others cannot. Some critics have charged him with dogmatism in this regard. Why should Descartes be allowed to legislate the scope of our clear and distinct perceptions? Perhaps we can clearly and distinctly perceive something that he could not.

Descartes cannot be saved entirely from this charge, but two important points can be made in his defense. First, he has principled reasons for thinking that everyone has the same set of innate or clear and distinct ideas. When the meditator first proved God's existence in the Third Meditation, she also established that God is supremely good and hence no deceiver. One consequence of God's perfect benevolence is that he implanted the same set of innate ideas in all finite minds. Thus, Descartes feels justified in concluding that the limits of his capacity for clear and distinct perception will be shared by everyone.

Second, when responding to objections to the ontological argument such as the ones considered above, Descartes typically does more than insist dogmatically on a unique set of clear and distinct ideas. He also tries to dispel the confusion which he thinks is at the root of the objection. Since the ontological argument ultimately reduces to an axiom, the source of an objection according to Descartes' diagnosis is the failure of the objector to perceive this axiom clearly and distinctly. Thus, Descartes devotes the bulk of his efforts to trying to remove those philosophical prejudices which are hindering his objector from intuiting the axiom. These efforts are not always obvious, however. Descartes is good at maintaining the pretense of answering criticisms to a formal proof. But his replies to Caterus' objections to the ontological argument are best read as an extended effort to dispel prejudice and confusion, so as to enable his reader to intuit God's existence for himself.

Perhaps the most interesting objection to the ontological argument, and the one that has received the most attention since being formulated by Immanuel Kant, is that existence is not a property or predicate. This objection enjoys the status of a slogan known by every undergraduate philosophy major worth her salt. In claiming that existence is included in the idea of a supremely perfect being, along with all the

other divine attributes, Descartes' version of the argument appears to succumb to this objection.

It is not obvious of course that existence is *not* a predicate. To convince us of this point, Kant observes that there is no intrinsic difference between the concept of a hundred real thalers (coins common in Kant's time) and the concept of a hundred possible thalers. Whenever we think of anything, we regard it *as* existing, even if the thing in question does not actually exist. Thus, existence does not add anything to the concept of a thing. What then is existence if not a predicate? Kant's answer is that existence is "merely the positing of a thing" or "the copula of a judgment," the point being that when we say "God exists" we are simply affirming that there is an object answering to the concept of God. We are not ascribing any new predicates to God, but merely judging that there is a subject, with all its predicates, in the world (CPR:B626-27).

Kant's formulation of the objection was later refined by Bertrand Russell in his famous theory of descriptions. He argues that existential statements such as "God exists" are misleading as to their logical form. While serving grammatically as a predicate, the term "exists" in this sentence has a much different logical function, which is revealed only by analysis. Properly analyzed, "God exists" means "there is one (and only one) x such that ‘x is omnipotent, omniscient, etc.’ is true." Russell thinks this translation shows that, appearances to the contrary, the statement "God exists" is not ascribing existence to a subject, but asserting that a certain description (in single quotes) applies to something in reality. Russell's view is reflected in the standard modern logical treatment of existence as a quantifier rather than a predicate.

It is widely believed that Descartes did not have a response to this objection, indeed that he blithely assumed that existence is a property without ever considering the matter carefully. But this is not the case. The seventeenth-century empiricist Pierre Gassendi confronted Descartes with this criticism in the Fifth Set of Objections (and probably deserves credit for being the first to enunciate it): "existence is not a perfection either in God or in anything else; it is that without which no perfections can be present" (AT 7:323; CSM 2:224). As with most of his replies to Gassendi (whom he regarded as a loathsome materialist and quibbler), Descartes responded somewhat curtly. But it is clear from the discussion in section 2 that he had the resources for addressing this objection in a systematic manner.

Before examining how Descartes might defend himself, it is important to note that the question at issue is typically framed in non-Cartesian terms and thus often misses its target. Both Kant and Russell for example are interested in the logical issue of whether existence is a *predicate*. Descartes, in contrast, was not a logician and disparaged the standard subject-predicate logic inherited from Aristotle. Although, as discussed above, he sometimes presents formal versions of the ontological arguments as heuristic devices, Descartes thought that God's existence is ultimately known through intuition. This intuitive process is psychological in character. It is not a matter of assigning predicates to subjects but of determining whether the idea of a supremely perfect being can be clearly and distinctly perceived while excluding necessary existence from it through a purely intellectual operation. To be sure, Descartes was interested in the *ontological* question of whether existence is a "property" of substances. For him, however, the analogues of properties are clear and distinct ideas and ways of regarding them, not predicates.

Having said that, Descartes' best strategy for answering the ontological version of the objection is to concede it, or at least certain aspects of it. Descartes explicitly affirms Kant's point that existence does not add anything to the idea of something (provided that the terms "idea" and "concept" are regarded as psychological items). Once again we should recall passage [4] from the Second Replies: "Existence is contained in the idea or concept of every single thing, since we cannot conceive of anything except as existing" (Axiom 10, AT 7:166; CSM 2:117). So Descartes agrees with Kant that there is no conceptual difference between conceiving a given substance as actually existing and conceiving it as merely possible. In the first instance one is attending to the existence that is contained on every clear and distinct idea, and in the other instance one is ignoring the thing's existence without actively excluding it. He would, however, stress another conceptual difference that Kant and other critics do not address, namely that between the two grades of existence -- contingent and necessary. The clear and distinct ideas of all finite things contain merely contingent or dependent existence, whereas the clear and distinct idea of God uniquely contains necessary or wholly independent existence (ibid.). As discussed previously, the ontological argument hinges on this distinction.

Another intuition underlying the claim that existence is not a property is that there is more intimate connection between an individual and its existence than the traditional one between a substance and a property, especially if the property in question is conceived as something accidental. If existence were accidental, then a thing could be without its existence, which seems absurd. It seems no less absurd to say that existence is a property among other properties (accidental or essential), for how can a thing even have properties if it does not exist? Descartes shares this intuition. He does not think that existence is a property in the traditional sense or is even distinct from the substance that is said to bear it. Recall the view discussed in section 2 that there is merely a rational distinction between a substance and its existence, or between the essence and existence of a substance. This means that the distinction between a substance and its existence is confined to thought or reason. Human beings, in their efforts to understand things using their finite intellects, draw distinctions in thought that do not obtain in reality. In reality, a substance (whether it be created or divine) just is its existence.

The purpose of this defense of Descartes is not to render a verdict as whether he has the correct account of existence, but to show that he had a rather sophisticated and systematic treatment of what has been one of the great bugbears in the history of philosophy. He did not simply make the ad hoc assumption that existence is an attribute in order to serve the needs of the ontological argument. Indeed, on Descartes' view, existence is not a property in the traditional sense, nor can one conceive something without regarding it as existing. Descartes' critics might not be convinced by his account of existence, but then they have the burden of providing a better account. The focus of the debate will then be shifted to the question of who has the correct ontology, rather than whether the ontological argument *per se* is sound.

[Return to Section links]

Bibliography

Primary Texts

- Adam, Charles, and Paul Tannery. 1964-1976. *Oeuvres de Descartes*, vols. I-XII, revised edition. Paris: J. Vrin/C.N.R.S. [references to this work (abbreviated as "AT") are by volume and page, separated by a colon.]
- Cottingham, John, Robert Stoothoff, Dugald Murdoch, and (for vol. 3) Anthony Kenny, eds. and trans. 1984. *The Philosophical Writings of Descartes*, vols. 1-3. Cambridge: Cambridge University Press. [All quotations are taken from this edition (abbreviated as "CSM"); any deviations from it are the author's own. References to this work are by volume and page, separated by a colon.]
- Kant, Immanuel. 1990. *Critique of Pure Reason*, trans. Norman Kemp Smith. London: Macmillan Education Ltd. [abbreviated as CPR]

Secondary Texts

- Barnes, Jonathan. 1972. *The Ontological Argument*. London: Macmillan.
- Chappell, Vere. 1997. "Descartes' Ontology," *Topoi* 16, 111-127.
- Cottingham, John. 1986. *Descartes*. Oxford: Blackwell.
- Doney, Willis. 1993. "Did Caterus Misunderstand Descartes's Ontological Proof?" in *Essays on The Philosophy and Science of Rene Descartes*, ed. Stephen Voss. New York: Oxford University Press.
- Doney, Willis. 1978. "The Geometrical Presentation of Descartes's A Priori Proof," in *Descartes: Critical and Interpretive Essays*, ed. Michael Hooker. Baltimore: Johns Hopkins University Press, 1-25.
- Edelberg, Walter. 1990. "The Fifth Meditation," *Philosophical Review* XCIX: 493-533.
- Gueroult, Martial. 1984. *Descartes's Philosophy Interpreted According to the Order of Reasons*, vol. 1. Trans. Roger Ariew. Minneapolis: University of Minnesota Press.
- Hick, John and McGill, Arthur C. 1967. *The Many-Faced Argument*. New York: The Macmillan Company.
- Kenny, Anthony. 1997. "Descartes' Ontological Argument," in *Descartes' Meditations: Critical Essays*, ed. Vere Chappell. New York: Rowman & Littlefield Publishers, Inc, 177-194.
- Kenny, Anthony. 1995. *Descartes: A Study of His Philosophy*. Bristol: Thoemmes Press, 146-171.
- Kenny, Anthony. 1970. "The Cartesian Circle and the Eternal Truths," *Journal of Philosophy* LXVII, 685-700.
- Newman, Lex, and Alan Nelson. 1999. "Circumventing Cartesian Circles," *Noûs* 33, 370-404.
- Nolan, Lawrence. 1998. "Descartes' Theory of Universals," *Philosophical Studies* 89, Nos. 2-3, 161-180.
- Nolan, Lawrence. 1997. "The Ontological Status of Cartesian Natures," *Pacific Philosophical Quarterly* 78, 169-194.
- Oppenheimer, Paul, and Zalta, Edward. 1991. "On the Logic of the Ontological Argument" in *Philosophical Perspectives 5: The Philosophy of Religion*, ed. J. Tomberlin. Atascadero: Ridgeview, 509-29.

- Schmaltz, Tad. 1991. "Platonism and Descartes' View of Immutable Essences," *Archiv fur Geschichte der Philosophie* 73, 129-70.
- Wippel, John. 1982. "Essence and Existence," in *The Cambridge History of Later Medieval Philosophy*, eds. Norman Kretzmann, Anthony Kenny and Jan Pinborg. New York: Cambridge University Press, 385-410.
- Wilson, Margaret. 1978. *Descartes.* New York: Routledge and Kegan Paul.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Anselm, Saint [Anselm of Bec, Anselm of Canterbury] | Aquinas, Saint Thomas | Descartes, René: epistemology | Descartes, René: life and works | existence | Kant, Immanuel | ontological arguments | Russell, Bertrand

Lawrence Nolan
lpnolan@csulb.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 18, 2001
Content last modified: June 18, 2001

Stanford Encyclopedia of Philosophy

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

René Descartes' Life and Works

Descartes has been heralded the first modern philosopher. He is famous for having made an important connection between geometry and algebra, which allowed for the solving of geometrical problems by way of algebraic equations. He is also famous for having promoted a new conception of matter, which allowed for the accounting of physical phenomena by way of mechanical explanations. However, he is most famous for having written a relatively short work, *Meditationes de Prima Philosophia* (*Meditations On First Philosophy*), published in 1641, in which he provides a philosophical groundwork for the possibility of the sciences.

Early Years

Descartes was born in La Haye on March 31, 1596 of Joachim Descartes and Jeanne Brochard. He was one of a number of surviving children (two siblings and two half-siblings). His father was a lawyer and magistrate, which apparently left little time for family. Descartes's mother died in May of the year following his birth, and he, his full brother and sister, Pierre and Jeanne, were left to be raised by their grandmother in La Haye. At around ten years of age, in 1606, he was sent to the Jesuit college of La Fleche. He studied there until 1614, and in 1615 entered the University of Poitiers, where a year later he received his Baccalaureate and License in Canon & Civil Law. For the history and the text of his thesis, see the following supplementary document:

Descartes' Law Thesis

In 1618, at the age of twenty-two, he enlisted in the army of Prince Maurice of Nassau. It is not known what his duties were exactly, though Baillet suggests that he would have very likely been drawn to what would now be called the Corps of Engineers (Baillet, Livre 1, Chapitre 9, p. 41). This division would have engaged in applied mathematics, designing a variety of structures and machines aimed at protecting and assisting soldiers in battle. Sorell, on the other hand, notes that in Breda, where Descartes was stationed, the army "doubled as military academy for young noblemen on the Continent" (Sorell, p. 6). And, Gaukroger notes that the education of the young noblemen was structured around the educational model of Lipsius (1547-1606), a highly respected Dutch political theorist who received a Jesuit education at Cologne (Gaukroger, pp. 65-6). It is likely that the military environment (that is, the academy) at Breda would have reminded Descartes of La Fleche. Though there are reasons for thinking that he may have been a soldier, the majority of biographers argue that it is more likely that his duties were oriented more towards education or engineering.

While stationed at Breda, Descartes met the mathematician Isaac Beeckman (1588-1637). Notes that Descartes kept related to his correspondence reveal that he and Beeckman had become more than simple acquaintances-their relationship was more one of teacher and student (Descartes being the latter). This relationship would rekindle in Descartes an intense interest in the sciences. In addition to discussions about a wide variety of topics in natural science, a direct result of certain questions posed by Beeckman compelled Descartes to write the *Compendium Musicae*. Among other things, the *Compendium* attempted to work out a theory of harmony, rooted in the concepts of proportion or ratio, which (along the lines of the ancients) attempted to express the notion of harmony in mathematical terms. It would not be published during Descartes's lifetime. As for Beeckman, Descartes would later downplay his influence.

The World and *Discourse*

After Descartes left the army, in 1619, his whereabouts for the next few years are unknown. Based on what he says in the *Discours de la Methode* (*Discourse on the Method*), published in 1637, there is speculation that he spent time near Ulm (Descartes apparently attended the coronation of Ferdinand II in

Frankfurt in 1619). There is some evidence that would suggest that he was in France in 1622, for it was at this time that property he had inherited was sold-the proceeds of which would provide him a simple income for many years. There is some speculation that between 1623 and 1625 he visited Italy. Descartes emerges in 1625 in Paris, his notes revealing that he was in contact with Father Marin Mersenne (1588-1648), a member of the Order of Minims. This relationship would prompt Descartes to make public his thoughts on natural philosophy (science). It is by way of Mersenne that Descartes's work would find its way into the hands of some of the best minds living in Paris-for instance, Antoine Arnauld (1612-1694), Pierre Gassendi (1592-1655), and Thomas Hobbes (1588-1679).

In 1628 Descartes left Paris. At this time he seems to have been working on the *Regulae ad Directionem Ingenii* (*Rules for the Direction of the Mind*), a work that he would abandon, some speculating around the time of the move from Paris. In 1630 he moved to Amsterdam. There he worked on drafts of the *Dioptrique* (the *Optics*) and the *Meteors* (the *Meteorology*), which were very likely intended to be a part of a larger work, *Le Monde* (*The World*). In 1632 he moved again, this time to Deventer, to apparently teach Henry Reneri (1593-1639) his physics. It is also during his stay in Deventer that Descartes probably worked on a final draft of the *Traite de l'homme* (*Treatise on Man*), which in connection to the *Optics* and the *Meteorology* was probably originally intended to be a part of *The World*.

When *The World* had become ready for publication in 1633, upon hearing of the Church's condemnation of Galileo (1564-1642) in the same year, Descartes decided against its publication. For, the world system he had adopted in the book assumed, as did Galileo's, the heliocentric Copernican model. In a letter to Mersenne, dated November 1633, Descartes expresses his fear that were he to publish *The World*, the same fate that befell Galileo would befall him. And, this is something that he understandably wanted to avoid. *The World* appears to have been constituted of several smaller, but related, works: a treatise on physics, a treatise on mechanics (machines), a treatise on animals, and a treatise on man. Although much of *The World* has been lost, some of it seems to have survived in the form of essays attached to the *Discourse* which, as was mentioned earlier, would be published four years later, in 1637. And, some of it was published posthumously. Arguably, Constantijn Huygens (1596-1687) received what Descartes refers to as "three sheets" of *The World*, along with a letter dated 5 October 1637. These "sheets" deal primarily with mechanics.

Around 1635, Reneri began to teach "Cartesian" physics. Also during this year, a domestic servant by the name of Helene gave birth to a baby girl, Francine. According to a baptismal record, dated 28 July 1635, Descartes is named the father (AT I 395n). Gaukroger gives the baptismal date as 7 August 1635 (Gaukroger, p. 294). And, Genevieve Rodis-Lewis gives Francine's date of birth as 19 June 1635 (Rodis-Lewis, p. 40). In 1636 Reneri acquired an official chair in Philosophy at the University of Utrecht, and continued to build a following of students interested in Cartesian science. Around March of 1636, at the age of forty, Descartes moved to Leiden to work out the publishing of the *Discourse*. And, in 1637 it is published. With the *Discourse* out and a following of students building in Utrecht, Descartes seems to have turned his attention from career to family. In a letter dated 30 August 1637 we find him apparently working out an arrangement for Francine, but strangely refers to her as his "niece"-which suggests that he did not want certain people to know that he was a father. Gaukroger suggests that despite this apparent denial of paternity, Descartes not only corresponds with Francine, but in 1637 brings her and Helene to

his new home at Santpoort or Egmond-Binnen (Gaukroger, pp. 294, 332).

The *Discourse* is the first published work of Descartes's, coming some four years after his abandonment of the publishing of *The World*. This work is important for many reasons. For instance, it tells us what Descartes himself seems to have thought of his early education, and in particular, his early exposure to mathematics. Roger Ariew suggests that these refletions are not so much those of the historical Descartes, as much as they are those of a personna Descartes adopts in telling the story of the *Discourse* (Ariew, pp. 58-63). Uncontested, however, is the view that the *Discourse* sketches out the metaphysical underpinnings of the Cartesian system. And, as a bonus, it has three works that are attached to it that are apparently added so as to exemplify the method of inquiry it develops (though admittedly it is unclear how the method is applied in these essays). The attached essays are the *Optics*, the *Meteorology*, and *Le Geometrie* (the *Geometry*). As was suggested earlier, the *Optics* and *Meteorology* were very likely versions of works originally intended for *The World*.

It should be stressed that the three attached essays are important independent of the *Discourse*, for they contain much worth studying. In the *Optics*, for example, Descartes works out his laws of refraction, and within this context, what would later be called Snell's Law (which Descartes seems to have worked out as early as 1632). Further, although the Geometry would seem to have come out of nowhere, there is evidence in Descartes's notes to himself, from which Clerselier reconstructed some of Descartes's correspondence, that he had been working on some version of it as early as 1619. In a letter to Beeckman, dated 26 March 1619, for example, Descartes discusses the subject matter that is found in the *Geometry*, and in a letter dated 23 April 1619, he explicitly mentions the book's title. It is in this work that Descartes shows how certain geometrical problems can be solved by way of algebraic equations.

The significance of the sort of connection that Descartes made between geometry and algebra was great indeed, for without it the mathematization of the physics and the development of the calculus might not have happened when they did-a generation later via Sir Isaac Newton (1642-1727). It should be noted, however, that as groundbreaking as this work may be, contrary to the claims of many, nowhere in the *Geometry* is a "Cartesian Coordinate System" ever developed (that is, the *x*-*y* coordinate system taught to today's students of algebra), nor is he the originator of other mathematical concepts that bear his name, for example, the "Cartesian Product". Carl Boyer notes that various concepts that lead to analytic geometry are found for the first time in the *Geometry*, and that the *Geometry's* mathematical notation is still used today. But, he argues, although Cartesian geometry is taken by many to be synonymous with analytic geometry, the fact is that the fundamental aim of Descartes's system is quite different from that of contemporary analytic geometry (Boyer, pp. 370-1). And so, the claim that Descartes is the originator of analytic geometry, at least as we understand it today, overstates the case. As Boyer rightly points out, however, this does not diminish the importance of the work in the history of mathematics.

The *Meditations*

In 1639 Descartes began writing the *Meditations*. And, in 1640 he returned to Leiden to help work out its publication. During the year, Francine died. She was only five years old. Understandably, Descartes was

deeply saddened by her death. There is evidence to suggest that he was called away from Leiden around the time of her death, returning soon after. Some have speculated that he left Leiden to be at her side. Also during this year, Descartes's father and sister died. Descartes's relationship with his father (and brother) was of the sort that Pierre, his brother, failed to even bother him with the news of their father's death. Rather, it seems to have been in a letter from Mersenne that Descartes first learns of it. In a follow up letter to Mersenne, dated 3 December 1640, Descartes expresses regret in not having been able to see his father before his death. But, unlike any speculation that might be made concerning his having been with Francine at her death, it is clear that he refused to return to Paris (where his father had died) upon hearing of his father's death. Instead, he says, he will stay (in Leiden) to complete the publishing of the *Meditations*. Some have suggested that this in part demonstrates a profound strain that existed in the relationship between Descartes and his father.

Today, the *Meditations* is by far Descartes's most popular work-though this would not have been the case in Descartes's day. This work is important to today's scholar for many reasons, the least of which is its including as an attached text written objections from some of the best minds living in Paris. Mersenne sent the *Meditations* to philosophers and theologians for criticism. The list of critics includes: Caterus, Hobbes, Arnauld, Gassendi, and Mersenne himself, with several other unnamed readers who raised their objections through Mersenne. A later edition would include Bordin. Descartes replied to each critic, and the result was an appended text referred to as "The Objections and Replies." The second edition contains seven sets in all.

The *Meditations* opens by developing skeptical questions concerning the possibility of knowledge. Through a series of several carefully thought out meditations, the reader establishes (along with the author) the groundwork for the possibility of knowledge (scientia). Descartes is not a skeptic, as some have insisted, but uses skepticism as a vehicle to motivate his reader to "discover" by way of philosophical investigation what constitutes this ground. In the Second Replies, Descartes refers to this style of presentation as the "analytic" style. There were two styles of presentation: analytic and synthetic. It is important not to confuse these terms with those, say, used by Kant. For Descartes the analytic style of presentation (and inquiry) proceeds by beginning with what is commonly taken to be known and discovering what is necessary for such knowledge. Thus, the inquiry moves from what is commonly known to first principles. The "discovery" moves in such a way that each discovery is based on what was discovered before. By contrast, the synthetic style of presentation begins by asserting first principles and then to determining what follows. Prompted by Mersenne, Descartes sketches out in the Second Replies a synthetic rendering of the *Meditations*.

In establishing the ground for science, Descartes was at the same time overthrowing a system of natural philosophy that had been established for centuries-a qualitative, Aristotelian physics. In a letter to Mersenne, dated 28 January 1641, Descartes says "these six meditations contain all the foundations of my physics. But please do not tell people, for that might make it harder for supporters of Aristotle to approve them. I hope that readers will gradually get used to my principles, and recognize their truth, before they notice that they destroy the principles of Aristotle." Unlike his earlier work, *The World*, the *Meditations* parts ways with the "old" science without explicitly forwarding controversial views, like that of the Copernican heliocentric model of the solar system. Specifically, the Cartesian view denies that physics is

grounded in hot, cold, wet, and dry. It argues that contrary to Aristotle's view, such "qualities" are not properties of bodies at all. Rather, the only properties of bodies with which the physicist can concern him or herself are size, shape, motion, position, and so on-those modifications that conceptually (or logically) entail extension in length, breadth, and depth. In contrast to Aristotle's "qualities," the properties (or modes) of bodies dealt with in Cartesian physics are measurable specifically on ratio scales (as opposed to intensive scales), and hence are subject in all the right ways to mathematics (Buroker, pp. 596-7). This conception of matter, conjoined with the sort of mathematics found in the *Geometry*, allies itself with the work of such Italian natural philosophers as Tartaglia, Ubaldo, and Galileo, and helps further the movement of early thinkers in their attempts to establish a mathematical physics.

Descartes's letter to the "learned and distinguished men" of the Sorbonne, which is appended to the *Meditations*, suggests that he was trying to pitch the *Meditations* as a textbook for the university. Though the endorsement of the Learned Men would not have guaranteed that the *Meditations* would be accepted or used as a textbook, it could certainly be viewed as an important step to getting it accepted. Unlike today's notion of a textbook, in Descartes's day "textbooks" were intended mostly for teachers, not students. Typically, at the close of a teacher's career, his notes would be published for the benefit of those who would go on to teach such course material. The awkwardness of Descartes's seeking the acceptance and use of his *Meditations* by teachers is amplified by the fact that he was not a teacher himself. Consequently, his seeking "textbook" status would have very likely been viewed by those Learned Men as being a bit pretentious. He was, it could be said, a freelancer with no academic or political ties to the university (outside of his connection to Mersenne). And, he certainly lacked the credentials and reputation of someone like a Eustachius, whose widely used textbook of the period is of the sort the *Meditations* was in all likelihood aimed at replacing. Although the *Meditations* seems to have been endorsed by the Sorbonne, it was never adopted as a text for the university.

The *Principles*

Soon after the Sorbonne passed on the *Meditations*, Descartes's public life was further complicated by the Dutch theologian, Gisbert Voetius (1588-1676). Voetius had attacked Regius, a Dutch physician who taught medicine at the University of Utrecht, for his having taught certain "Cartesian" ideas that conflicted with traditional theological doctrine. Regius was friend to both Reneri and Descartes, and was a strong adherent to Descartes's philosophical views. Voetius tried to have Regius removed from his position as professor, while at the same time attacking Descartes's work and character. In his defense Descartes entered into the debate. The controversy would leave Regius confined to teaching medicine, forbidden to teach anything Cartesian, and his published defense of Cartesian thought would be officially condemned by Voetius, who in five years time would rise to the position of University rector. At the end of the debate, which off and on lasted about five years, the situation became desperate for Descartes. He feared being expelled from the country and of seeing his books burned. He would even seek protection by asking the Prince of Orange to intervene and quell Voetius' attack.

In 1643, at the age of forty-seven, Descartes moved to Egmond du Hoef. With the Voetius controversy seemingly behind him (though, as mentioned above, it would again raise its head and climax five years

down the road), Descartes and Princess Elizabeth of Bohemia began to correspond. In this exchange, Princess Elizabeth probed Descartes on the implications of his commitment to mind-body dualism. During this time, he completed a final draft of a new textbook, which he had begun three years earlier, the *Principia Philosophiae* (*Principles of Philosophy*), and in 1644 it was published. He dedicated it to Princess Elizabeth.

The *Principles* is an important text. The work is divided into four Parts, with five hundred and four articles. Part One develops Descartes's metaphysics. Although it would appear to be a quick run through of the *Meditations*, there are a number of dissimilarities. For example, the order of presentation of the proofs for God's existence, which some have argued is significant, found in the Third and Fifth Meditations, is reversed in the *Principles*. The principles introduced in Part Two are based on the metaphysics of Part One. And, the subsequent physics developed in Parts Three and Four is based upon the principles of Part Two. Although the physics turns out to be unsound, the *Principles* nevertheless inspired such great thinkers as Robert Boyle (1627-1691), Edmond Halley (1656-1742), and Isaac Newton. As an important side note, it must be stressed that even though Descartes had throughout his career put a great deal of emphasis on mathematics, the physics developed in the *Principles* is not a mathematical physics. Rather, it is a conceptual project with only a hint of empirical overtones-a physics rooted entirely in metaphysics. Two parts, never completed, were originally intended to deal with plants, animals, and man. In light of this and what Descartes says in a 31 January 1642 letter to the mathematician Constantijn Huygens, it is plausible to think that the *Principles* would have looked something like *The World* had it been completed as planned.

One of the more controversial positions the *Principles* forwarded, at least according to Newton, was that a vacuum was impossible. This followed from Descartes's commitment to the view that the essence of body was extension. Supposing that a vacuum was taken to be a "gap" between bodies-that is, it is taken to be an utter absence of body (matter)-if it turned out (as indeed it does) that this gap was extended in length, breadth, and depth, then it would not be an absence of body, but would by definition *be* a body-as much a body as the two bodies between which it is taken to be a gap. The corporeal universe was thus a plenum, individual bodies separated only by their surfaces. Newton argued in his *De Gravitatione* and *Principia* that the concept of motion becomes problematic if the universe is taken to be a plenum. Another controversial position was Descartes's insistence that matter is infinitely divisible. Gassendi, and later Cordemoy, argued that there must be a bottom, a "substance," to the physical universe upon which the being of all corporeal things depend. In line with the ancient atomist Epicurus, they argued that if matter was infinitely divisible, so dividing it would show that there was no bottom-and so, corporeality would not be substantial. So, if corporeality is substantial, as Descartes himself had claimed, there must be a minimum measure of extension that could not be divided (by natural means, anyway). And so, there are atoms. But, this conclusion was something that Descartes explicitly rejected in the *Principles*.

The *Passions*

In 1646, as a result of the probings of Princess Elizabeth, Descartes completed a working draft of *Passions de l'ame* (*Passions of the Soul*). During this year another prominent political figure began to

correspond with Descartes, Queen Christina of Sweden. And, Regius published what he took to be a new and improved version of Cartesian science, which as we now know would draw the wrath of Voetius. But Regius did not stop there, for he seemed to have found important differences between his "Cartesian" view and that of Descartes's, and attempted to separate the two, publishing a broadsheet that listed twenty-one anti-Cartesian theses (which his version of "Cartesian" science rejected). In response to this, Descartes wrote a single-page printed defense that was posted on public kiosks for all to read. Published in 1648, the *Notae in Programma Quoddam* (*Notes on a Program*-also referred to as *Comments on a Certain Broadsheet*) is Descartes's public defense. However, as mentioned earlier, tensions mounted as a result of the public exchange and Descartes felt his way of life in the Netherlands to be threatened. As luck would have it, an admirer and friend of Descartes's-Chanut, who worked for Queen Christina's court-and Queen Christina herself began probing Descartes about the possibility of coming to Sweden. And, after a not too lengthy correspondence, Queen Christina offered Descartes a position in her court. For many reasons, which would certanily include those related to his concerns about Voetius, Descartes accepted the offer. And, in 1649 he left for Sweden.

Queen Christina at first required very little from Descartes. However, according to Gaukroger, this would change. For, after he had some time to settle in, she ordered him to do two things: first, to put all of his papers in order, and secondly, to put together designs for an academy (Gaukroger, p. 415). Arguably, Descartes had some idea of how the latter might be done by way of his experience in Breda. In January of 1650 Queen Christina began to require Descartes to give her lessons in philosophy. These apparently would begin at five in the morning and would last for about five hours. They were given three days a week (Gaukroger, p. 415). During this time Descartes published the *Passions*, the work having emerged primarily from his correspondence with Princess Elizabeth (to whom he had dedicated the *Principles*). One aim of the *Passions* was to explain how the emotional (and thus moral) life of a human being was connected to the soul's being essentially united to a body. Simply put, a ’passion of the soul’ is a mental state (or thought) that arises as a direct result of brain activity. Such passions can move us to action. Since this is so, Descartes suggests that one needs to learn to control one's passions, for they can move one to perform vicious acts. Critics of Descartes, including Elizabeth, argued that Descartes's metaphysical commitments put real pressure on the view expounded in the *Passions*. For, according to Descartes's metaphysics, the nature of mind is to think and the nature of body is to be extended in length, breadth, and depth. One view concerning causation, a view that Descartes's critics seemed to have attributed to him, is that one thing causes another to move, for example, by way of contact. Contact,in this context, seems to be possible only by way of surfaces. Now, bodies, since they are extended and thus have surfaces, can come into contact with one another and thus can cause one another to move. However, if minds are not extended, they lack surfaces. And, if they lack surfaces, there is no way in principle for bodies to come into contact with them. Thus, there is no way in principle for bodies to move minds, and visa versa. That is, minds and bodies cannot in principle causally interact. And so, if the view expounded in the *Passions* requires that bodies and minds be capable of causal interaction, and Descartes's metaphysical commitments make such interaction impossible, Descartes's metaphysics puts a great deal of pressure on the view expounded in the *Passions*.

Although things seemed to be moving forward, they were not going as well as one would have hoped. In a letter to Bregy, for instance, dated 15 January 1650, Descartes expresses reservations about his decision

to come to Sweden. He sees himself to be "out of his element," the winter so harsh that "men's thoughts are frozen here, like the water" (AT V 467; CSMK III 383). Given the sentiment expressed in the letter, this remark was probably intended to be as much Descartes's take on the intellectual climate as it was about the weather. In early February, less than a month after writing Bregy, Descartes fell ill. His illness quickly turned into a serious respiratory infection. And, although at the end of a week he appeared to have made some movement towards recovery, things took a turn for the worse and he died in the early morning of 11 February 1650. He was fifty-three years old.

Bibliography

NOTE: In what follows, the Adam and Tannery volumes, *Oeuvres De Descartes*, 11 vols., are cited. Such citations are abbreviated as "AT," followed by the appropriate volume and page numbers. I have whenever possible used the Cottingham, Stoothoff, and Murdoch translation, *The Philosophical Writings Of Descartes*, 3 vols. Volume 3 includes Anthony Kenny as a translator. This has been abbreviated as "CSMK," followed by the appropriate volume and page numbers. The AT and CSMK numbers are cited, side by side, separated by a semicolon.]

References

- Ariew, Roger. "Descartes and Scholasticism: the intellectual background to Descartes' thought," in *The Cambridge Companion to Descartes*, edited by John Cottingham (Cambridge: Cambridge University Press, 1992), pp. 58-90.
- Baillet, Adrien. *La Vie de M. Descartes* (2 vols.) Paris (1691).
- Boyer, Carl B. *A History of Mathematics* (Princeton: Princeton University Press, 1985).
- Buroker, Jill. "Descartes On Sensible Qualities," *Journal Of The History Of Philosophy*, October 1991, Vol. XXIX, No. 4, pp. 585-611.
- Descartes, René. *Oeuvres De Descartes*, 11 vols., edited by Charles Adam and Paul Tannery (Paris: Librairie Philosophique J. Vrin, 1983).
- Descartes, René. *The Philosophical Writings Of Descartes*, 3 vols., translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch, volume 3 including Anthony Kenny (Cambridge: Cambridge University Press, 1988).
- Gaukroger, Stephen. *Descartes: An Intellectual Biography* (Oxford: Clarendon Press, 1995).
- Rodis-Lewis, Genevieve. "Descartes' life and the development of his philosophy," in *The Cambridge Companion to Descartes*, edited by John Cottingham (Cambridge: Cambridge University Press, 1992), pp. 21-57.
- Sorell, Tom. *Descartes* (Oxford: Oxford University Press, 1987).

Other English Translations

- Descartes, René. *The Philosophical Writings Of Descartes*, 3 vols., translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch, volume 3 including Anthony Kenny (Cambridge:

Cambridge University Press, 1988).
- -------. *Meditations on First Philosophy*, translated by John Cottingham (Cambridge: Cambridge University Press, 1996).
- -------. *Principles of Philosophy*, translated by V.R. Miller and R.P. Miller (Dordrecht: D. Reidel, 1983).
- -------. *The Geometry of René Descartes*, translated by David Eugene Smith and Marcia L. Lantham (New York: Dover Publications, 1954).
- -------. *The Passions of the Soul*, translated by Stephen H. Voss (Indianapolis: Hackett Publishing Company, 1989).

Helpful Sources

- *Essays on Descartes'* Meditations, edited by Emelie Oksenberg Rorty (Berkeley: University of California Press, 1986).
- *Descartes's* Meditations: *Critical Essays*, edited by Vere Chappell (Lanhan: Rowan & Littlefield Publishers, Inc., 1997).
- *Descartes'* Meditations: *Background Source Materials*, edited by Roger Ariew, John Cottingham, and Tom Sorell (Cambridge: Cambridge University Press, 1998).
- *Descartes*, edited by John Cottingham (Oxford: Oxford University Press, 1998).
- *Reason, Will and Sensation: Studies in Descartes' Metaphysics*, edited by John Cottingham (Oxford: Clarendon Press, 1994).
- *Studies in Seventeenth-Century European Philosophy*, Volume 2 of *Oxford Studies in the History of Philosophy*, edited by M.A. Stewart (OxfordL Clarendon Press, 1997).
- *The Cambridge History of Seventeenth-Century Philosophy*, 2 vols., edited by Daniel Garber and Michael Ayers (Cambridge: Cambridge University Press, 1998).
- Garber, Daniel. *Descartes' Metaphysical Physics* (Chicago: University of Chicago Press, 1992).
- Gaukroger, Stephen. *Descartes: An Intellectual Biography* (Oxford: Clarendon Press, 1995).
- Gueroult, Martial. *Descartes' Philosophy Interpreted According to the Order of Reasons*, 2 vols., translated by Roger Ariew (Minneapolis: University of Minnesota Press, 1984).
- Kenny, Anthony. *Descartes: A Study of His Philosophy* (Bristol: Thoemmes Press, 1968).
- Rodis-Lewis, Genevieve. *Descartes: His Life and Thought*, translated by Jane Marie Todd (Ithaca: Cornell University Press, 1999).
- Rozemond, Marleen. *Descartes's Dualism* (Cambridge: Harvard University Press, 1998).
- Sorell, Tom. *Descartes* (Oxford: Oxford University Press, 1987).
- Williams, Bernard. *Descartes: The Project of Pure Enquiry* (London: Penguin Books, 1978).
- Wilson, Margaret. *Descartes* (London: Routledge & Kegan Paul plc, 1978).

Other Internet Resources

- Descartes E Il Seicento, maintained by Giulia Belgioioso (Director, Centro Interdipartmentali Di Studi Su Descartes E Il Seicento), Jean-Robert Armogathe (Centre d'Etudes cartésiennes), and their colleagues

A superb website on Descartes

Related Entries

Descartes, René: epistemology | Descartes, René: modal metaphysics | Descartes, René: ontological argument | Descartes, René: theory of sensation | Desgabets, Robert

Acknowledgements

I am indebted to the NEH for allowing me the opportunity to participate in the 2000 NEH Summer Seminar, "Descartes and His Contemporaries," held at Virgina Tech. In addition to learning a great deal from the Seminar's leaders, Roger Ariew and Daniel Garber, I learned a great deal from my fellow participants. I am also indebted to Bloomsburg University of Pennsylvania for providing me with a Faculty Development Grant, summer 2001. Lastly, I am indebted to Alan Nelson and Roger Ariew for comments on earlier versions of this article.

Kurt Smith
ksmit4@bloomu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 9, 2001
Content last modified: March 25, 2002

Stanford Encyclopedia of Philosophy
Supplement to René Descartes' Life and Works

Descartes' Law Thesis

Traditionally, scholars have paid little attention to Descartes's moral views, and even less to his legal and political views--this even though it is well known that he had earned a degree and license in Canon and Civil Law at Poitiers. Of course, an explanation for this is that Descartes wrote very little on such things. And so, the lack of attention is the direct result of a lack of text. However, in 1981, while reframing a seventeenth-century engraving that had been hanging in a museum restaurant, a curator for the Sainte-Croix Museum discovered a document that could change all of this. Stuffed in the back of the engraving was a small broadsheet, which was used by the original framer to secure the engraving in its frame. As was typical, such broadsheets were used to announce upcoming defenses and the theses of the candidates, the latter presented as a list of conclusions to be defended. After the defense dates had passed, the posters were discarded, and some found their way into the hands of framers who on occasion would use them as stuffing, as was the case here. The discovered 1616 broadsheet announced an upcoming (oral) thesis defense at the law school at Poitiers -- the defense of the law thesis of one Rene Descartes.

It wasn't until October of 1986 that the discovery would be officially documented in the *Archives departementales de la Vienne*, at Poitiers. And even though the existence of the document has been known for several years, very little has been done to bring the thesis and its contents to the attention of Descartes scholars. It is not included in the Adam and Tannery volumes. And, it is not included in any of the French or English translations of Descartes's work -- this, even though Jean-Robert Armogathe, Vincent Carraud, and Robert Feenstra have verified its authenticity. In fact, only one article appears that deals with the thesis in its entirety: an article published in 1988 in *Nouvelles De La Republique Des Lettres*, co-written by the three scholars just mentioned. The broadsheet (hereafter referred to as the "1616 Law Thesis") contains a title, an introductory discussion, and a list of forty conclusions to be defended. And, although the document is only a single page, it appears to be fruitful enough to support new and exciting projects in Descartes scholarship.

The *1616 Law Thesis* deals specifically with the legal concept of inheritance. Descartes's view concerning the transference of property from one generation to the next seems to foreshadow a view expounded in works to come -- for example, in the *Rules for the Direction of the Mind* (1628) and in the *Discourse on the Method* (1637). "Last Wills can be generally defined" he says, "as the ultimate rule by way of which inheritance is handed over" (*1616 Law Thesis*, my translation). This rule, he argues, is grounded in both Civil and Natural Law (Ibid.). The idea is that the property of one generation can be justifiably passed to the next by way of this rule. This, even though the Crown may have a claim to such property. The rule keeps the Crown from legally interfering with the transfer. In the *Rules* and *Discourse* we can find a parallel. Like the rule of inheritance, a rule (or a set of rules -- i.e., a method) must be used

to secure the transference of knowledge from one generation to the next, the rule (or rules) keeping authority from interfering with the transfer.

We learn in the *Discourse*, and later in the *Meditations*, that Descartes was dissatisfied with the 'knowledge' he had acquired as a child (at La Fleche). This dissatisfaction resulted in his search for a method that would guarantee knowledge. Traditionally, Descartes's insight into the need for a method is dated around 1619, and is said to have found expression in print for the first time, some eighteen years later, in the *Discourse*. However, the insight is found lurking about in the introductory paragraph of the *1616 Law Thesis*:

> I thought that while even a tender sprout that I was especially inquisitive, for while nearly all the youngsters cried as they departed their youth, I was devoted to the fountains, the wet milk of my step-mother -- the nectar of the liberal arts -- dripping from my lips. At first in fact, I, wonderfully delighted, muttered a noisy stream of flattery, strongly desiring to drink the stuff of the honey-flowing poet. But soon, my admiration for their [my teachers'] heavy clammer and voices produced in me a torrent of images in which I took refuge, in turn hiding from me the eloquent waters for which I had [originally] thirsted. And, not only did these things produce in me more of a thirst for knowledge than they could quench, but none of them ever really satisfied me. The [resulting] desolation of knowledge eventually leveled me, and I, at that point, began to search with great zeal for any streams that were more abundant than this other, and that flowed [perhaps] in a different direction. [At first I thought that] carrying out something so ambitious was certainly not insane, since I did not think that such an undertaking would wear me down, or judge that a single petty stream would wear me out. My hope was that I could root out some [of the streams] from the rest, the sweet drops of the former at last calming my nature. And, my hope was that all [of this] would be settled by way of proof.

The reference to streams in this passage is connected to a central metaphor of the *1616 Law Thesis*: the metaphor of the fountain. We find this metaphor used by Descartes in later works: *The World*, *Treatise on Man*, *Discourse*, and *Optics*, to name a few. It is used, for example, in explanations of light, force, motion, dynamic and mechanical processes, and animal behavior. It would seem that this metaphor, in addition to its being central to the *1616 Law Thesis*, is central to understanding Cartesian views on the dynamic processes of both mind and body. This suggests the use of a central 'picture' that runs throughout the entire Cartesian corpus, a picture that is first introduced in his law thesis at Poitiers.

Further Documentation

Descartes' 1616 Law Thesis -- Copy of Original Document

Descartes' 1616 Law Thesis -- Latin Transcription

Descartes' 1616 Law Thesis -- English Translation (Work in Progress)

Acknowledgement

Thanks to Roger Ariew and Helen Hattab for their advice on this project, and to Daniel Garber for providing me with a copy of the original document.

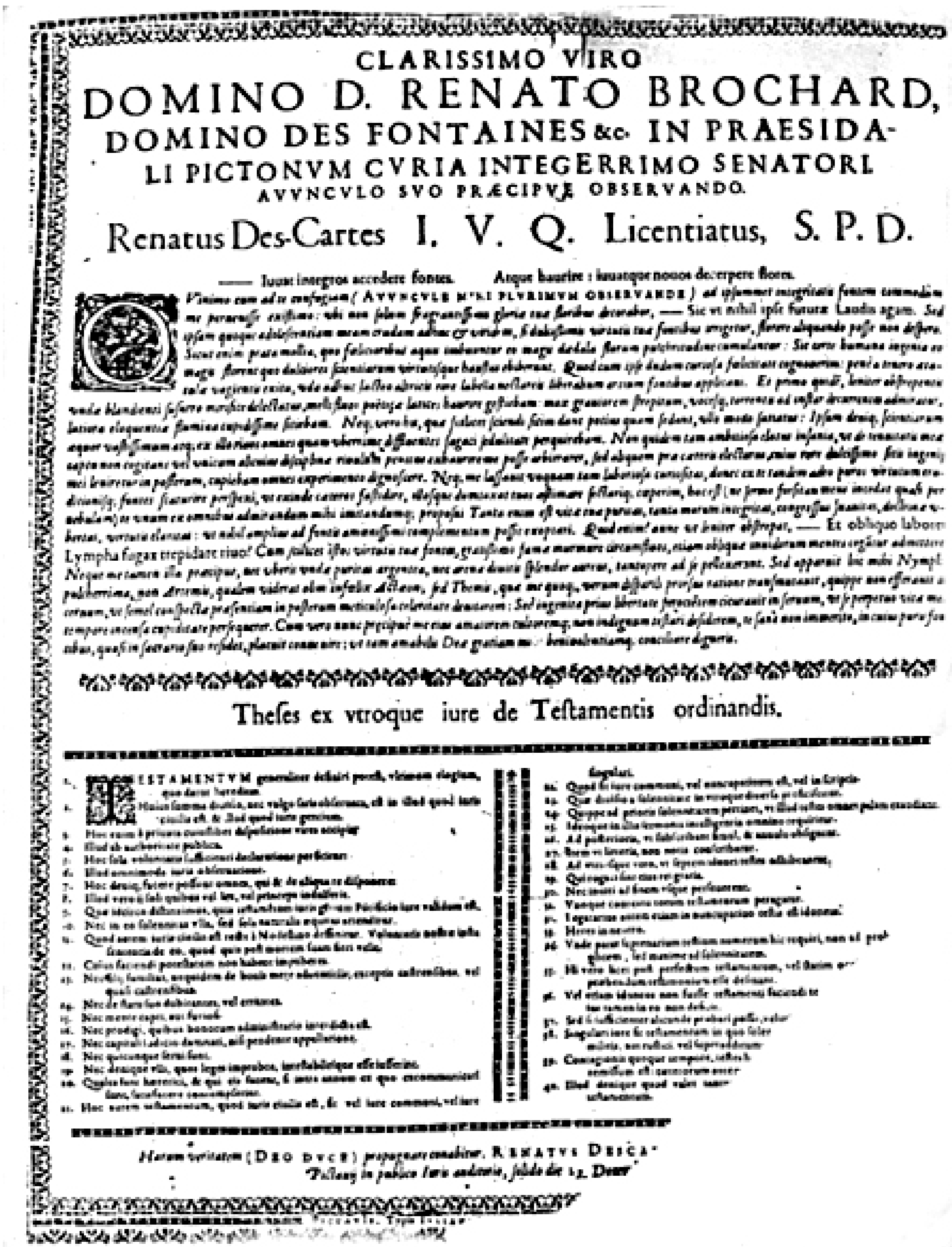
Kurt Smith
ksmit4@bloomu.edu

Return to René Descartes' Life and Works

First published: January 2, 2002
Content last modified: January 2, 2002

Stanford Encyclopedia of Philosophy
Supplement to René Descartes' Life and Works

Descartes' 1616 Law Thesis -- Copy of the Original Document



CLARISSIMO VIRO

DOMINO D. RENATO BROCHARD,

DOMINO DES FONTAINES &c. IN PRAESIDALI PICTONVM CVRIA INTEGERRIMO SENATORI

AVVNCVLO SVO PRÆCIPVÈ OBSERVANDO.

Renatus Des-Cartes I. V. Q. Licentiatus, S. P. D.

Theses ex vtroque iure de Testamentis ordinandis.

Further Documentation

Descartes' 1616 Law Thesis -- Latin Transcription

Descartes' 1616 Law Thesis -- English Translation (Work in Progress)

Acknowledgements

Thanks to Roger Ariew and Helen Hattab for their advice on this project, and to Daniel Garber for providing me with a copy of the original document.

Kurt Smith
ksmit4@bloomu.edu

Return to Supplementary Document: Descartes' Law Thesis

First published: January 2, 2002
Content last modified: January 2, 2002

Stanford Encyclopedia of Philosophy
Supplement to René Descartes' Life and Works

Descartes' 1616 Law Thesis -- Latin Translation

CLARISSIMO VIRO

DOMINO D. RENATO BROCHARD,

DOMINO DES FONTAINES &c. IN PRAESIDA-

LI PICTONUM CURIA INTEGERRIMO SENATORI,

AVVUNCULO SUO PRAECIPUAE OBSERVANDO.

Renatus Des-Cartes I.V.Q. Licentiatus, S.P.D.

--Iuuat integros accedere fontes. Atque haurire: iuuatque nouos decerpere flores.

Quinimo cum ad te confugiam (AVUNCULE MIHI PLURIMUM OBSERVANDE) ad ipsummet integitatis fontem commodum me pervenisse existimo: ubi non solum fragrantissimis gloriae tuae floribus decorabor, --Sic ut nihil ipse futurae laudis agam. Sed ipsam quoque adolescentiam meam crudam adhuc & viridem, si dulcissimis virtutis tuae fontibus irrigetur, florere aliquando posse non despero. Sicut enim prata mollia, quo foelicioribus aquis imbuuntur eo magis daedala florum pulchritudine cumulantur: Sic certe humana ingenia eo magis florent quo dulciores scientiarum virtutisque haustus ebiberunt. Quod cum ipse dudum curiosa foelicitate cognouerim: pene a tenero aetatulae vaientis exitu, uda adhuc lacteo altricis rore balella nectareis liberalium artium fontibus applicaui. Et primo quide, leniter obstrepentis undae blandienti susurro mirifice delectatus, melliflous poeticae latices haurire gestiebam: mox graviorem strepitum, voceque torrentis ad instar decurrentes admiratus, latiroa eloquentiae flumina cupidissime sitiebam. Neque vero his, quae scilicet sciendi sitim dant potius quam sedant ullo vodo satiatus: Ipsum denique scientiarum aequor vastissimum atque ex illo rivos omnes quam uberrime diffluentes sagaci sedulitate perquirebam. Non quidem tam ambitiosa elatus insania, ut de tenuitatis meae captu non cogitans vel unicum alicuius disciplinae rivulum pentitus eshaurire me posse arbitrarer, sed aliqum prae caeteris electurus, cuius rore dulcissimo sitis ingenii mei leniretur in posterum, cupiebam onmes experimeno dignoscere. Neque me lassauit unquam tam laoriosa curiositas, donec ex te tandem adeo puos virtutm erudiionisuque fontes scaturire perspexi, ut exinde caeteros fastidiere, illosque dumtaxat tuos aestimare sectarique coeperim, hoc est (ne sermo forsitan meus incedat quasi per nebulam) te unum ex omnibus admirandum mihi imitandumque proposui. Tanta enim est vitae tuae puritas, tanta morum integritas, congressus suavitas, doctrinae ubertas, virtutis

claritas: ut nihil amplius ad fontis amoenissimi complementum possit exoptari. Quid enim? Anne ut leniter obstepat, --Et obliquo laboret Lympha fugax trepidare rivo? Cum scilicet istos virtutis tuae fontes, gratissimo famae murmure circumfluos, etiam nec uberis undae puritas argentae, nec arenae divitis splendor aureus, tantopere ad se pellexerunt. Sed apparuit hic mihi Nympha pulcherrima, non Artemis, qualem viderat olim infoelix Actaeon; se Themis, quae me quoque verum disparili prorsus ratione transmutauit, quippe non efferauit in ceruum, ut semel conspectae praesentiam in posterum meticulosa celeritate deuitarem: Sed ingenita prius liberate feroietem cicurauit in seruum, ut se perpetuo vitae meae tempore incensa cupiditate persequerer. Cum vero nunc preciue me eius amatorem cultoremque non indignum testari desiderem, te sane non immerito, in cuius pris fontibus, quasi in sacrario suo residet, placuit conuenire; ut tam amabilis Deae gratiam mihi benivolentiamque conciliare digneris.

Theses ex utroque iure de Testamentis ordinandis

1. TESTAMENTUM generaliter definiri potest, ultimum elogium, quo datur hereditas.
2. Huius summa divisio, nec vulgo satis observata, est in illud quod iuris civilis est, & illud quod iuris gentium.
3. Hoc enim a privata cuiuslibet dispositione vires accipit;
4. Illud ab authoritate publica.
5. Hoc sola voluntatis sufficienti declaratione perficitur:
6. Illud ominmoda iuris observatione.
7. Hoc denique facere possunt omnes, qui & de aliqua re disponere:
8. Illud vero ii soli quibus vel lex, vel princeps indulserit.
9. Quae idcirco distinximus, quia testamentum iuris gentium Potificio iure validum est.
10. Nec in eo solenitas ulla, sed sola naturalis aequitas attenditur.
11. Quod atuem iuris civilis est recte a Modestino deffinitur Voluntatis nostrae iusta sententia de eo, quod quis post mortem suam fiere velit.
12. Cuius faciendi potestatem non habent impuberes.
13. Nec filii familias, nequidem de bonis mere adventitiis; exceptis castrensibus, vel quasi castrensibus.
14. Nec de statu suo dubitantes, vel errantes.
15. Nec mente capti, aut furiosi.
16. Nec prodigi, quibus bonorum administratio interdicta est.
17. Nec capitali iudicio damnati, nisi pendente appellatione.
18. Nec quicunque serui sunt.
19. Nec denque ulli, quos leges improbos, intestabilesque esse iusserint.
20. Quales sunt haeretici, & qui eis fauent, si intra annum ex quo excommunicati sunt, satisfacere contempserint.
21. Hoc autem testamentum, quod iuris civilis est, fit vel iure communi, vel iure singulari.
22. Quod fit iure communi, vel nuncupatiuum est, vel in scriptis.
23. Quae divisio a solennitate in utroque diversa proficiscitur.
24. Quippe ad prioris solennitatem pertinet, ut illud testes omnes palam exaudiant.
25. Ideoque in illis sermonis intelligentia omnino requiritur.
26. Ad posterioris, ut subscribant simul, & annulo absignent.

27. Item ut litteris, non notis conscribatur.
28. Ad utriusque vero, ut septem idonei testes adhibeantur.
29. Qui rogati sint eius rei gratia.
30. Nec inuiti ad finem usque perseuerent.
31. Unoque contextu totum testamentum peragatur.
32. Legatarius autem etiam in nuncupativo testis est idoneus.
33. Here in neutro.
34. Unde patet septenarium testium numerum hic requiri, non ad pro[bationem sim]plicem, sed maxime ad solennitatem.
35. Hi vero licet post perfectum testamentum, vel statim occi [MISSING TEXT] praebendum testimonium esse desinant.
36. Vel etiam idoneos non fuisse testamenti faciendi te[MISSING TEXT] ius tame in eo non deficit.
37. Sed si sufficenter alicunde probari possit, veluti [MISSING TEXT].
38. Singulari iure fit testamentum in quo solen [MISSING TEXT] militis, aut rustici, vel superadditum [MISSING TEXT].
39. Contagionis quoque tempore, testes h[MISSING TEXT] remissum est: caeterorum autem [MISSING TEXT].
40. Illud denique quod valet tantum [MISSING TEXT] testamentum.

Harum veritatem (DEO DUCE) propugnare conabitur, RENATUS DESCAR[MISSING TEXT]

Pictauii in publico Iuris auditorio, solido die 21. Decem[MISSING TEXT]

PICTAVIS, Typis IVLIA[MISSING TEXT]

Further Documentation

Descartes' 1616 Law Thesis -- English Translation (Work in Progress)

Descartes' 161 Law Thesis -- Copy of Original Document

Acknowledgements

I am indebted to Daniel Garber for providing me with a copy of the original Latin Law Thesis.

Kurt Smith
ksmit4@bloomu.edu

Return to Supplementary Document: Descartes's Law Thesis

First published: January 2, 2002
Content last modified: January 2, 2002

Stanford Encyclopedia of Philosophy
Supplement to René Descartes' Life and Works

Descartes' 1616 Law Thesis -- English Translation

WITH THE MOST ILLUSTRIOUS MAN

LORD D. RENE BROCHARD,

LORD OF THE FOUNTAINS, ETC. A DEFENSE

AT THE POITIERS SENATE HOUSE WITH ALL THE SENATE,

HIS UNCLE IN SPECIAL ATTENDANCE.

Rene Descartes: Canon and Civil Law License [I.V.Q.], Gives salutations and peace [S.P.D.]

--*It pleases one to approach the pure fountains, and to drink, and to gather new flowers.*

Indeed, in taking refuge in you (my uncle, having seen much) I think that I have reached the pleasant fountain of integrity itself. Not only will I be adorned with the glorious fragrance of your flowers--*Thus I do nothing to bring about my coming glory*--but also my adolescence, still fresh and green, irrigated by your fountain of sweetest virtue, will blossom hereafter, and I cannot despair. Indeed, just as the gentle meadows whose foliage is drenched with water, on which a more variegated beauty of flowers is heaped, certainly human nature will the more flourish which drinks the sweet drink of knowledge and virtue.

I thought that while even a tender sprout that I was especially inquisitive, for while nearly all the youngsters cried as they departed their youth, I was devoted to the fountains, the wet milk of my step-mother--the nectar of the liberal arts--dripping from my lips. At first in fact, I, wonderfully delighted, muttered a noisy stream of flattery, strongly desiring to drink the stuff of the honey-flowing poet. But soon, my admiration for their heavy clamor and voices produced in me a torrent of images in which I took refuge, in turn hiding from me the eloquent waters for which I had [originally] thirsted. And, not only did these things produce in me more of a thirst for knowledge than they could quench, but none of them ever really satisfied me.

The [resulting] desolation of knowledge eventually leveled me, and I, at that point, began to search with great zeal for any streams that were more abundant than this other, and that flowed [perhaps even] in a different direction. [At first I thought that] carrying out something so ambitious was certainly not insane, since I did not think that such an undertaking would wear me down, or judge that a single petty stream

would wear me out. My hope was that I could root out some [of the streams] from the rest, the sweet drops of the former at last calming my nature. And, my hope was that all would be distinguished by way of [proof of] experience.

[I did not realize, however, how much] my ever so laborious inquisitiveness had worn me out until I had arrived at a [certain] place and had learned of pure virtue, having seen it gushing from the fountain. And, ever since then I disliked these other [streams], and I began to value and follow only that [stream] of yours, [Uncle], this (and not my speech that now advances us as if through a thick fog) I propose is the one thing itself from all others that I now admire and imitate. For great is the purity or your life, great the integrity of [your] character, great the sweetness of [your] social intercourse, great the abundance of [your] teaching, great the strength of [your] virtue: as nothing more of this most pleasant fountain can be longed for but to be filled up [by it].

For what? In order to gently make a noise--For am I not turned awry as it works, the speeding water that hurries along with the river? Certainly with respect to your fountains of virtue, the greatest gratitude overflows with the roar of public opinion, [and even] the minds of the envious allow it to be thought. Not only have these distinguished things so greatly attracted me, but the abundant surge [of water] is [surely] pure silver, and the richness of the sandy banks brilliant gold.

It is thus that my Maiden appears to be the most beautiful, not Artemis, who Actaeon saw naked, but Themis, who has by way of truly unequal straightforward reason changed me, not into a wild stag, where following I [like Acteaon] quickly avoid appearing visibly fearful. [Rather,] she tames through freedom the wildness implanted previously in the brute, without interruption, and at this the appropriate time has excited [in me] the desire to pursue my own life. And, since indeed it is not unfit for me to desire to testify concerning my lover and sustainer, [and that] you are not undeserving of that which is in the pure fountains--it residing as though in your sacred place--it is now especially pleasing [for us] to gather, and it is even more lovely to bring about your deeming worthy for the Goddess [her] favor and for me [my] kindness.

A THESIS FROM TWO SIDES OF LAW CONCERNING THE REGULATING OF LAST WILLS

1. Last Wills can be generally defined as the ultimate rule by way of which inheritance is handed over.
2. Of this final distribution [of property], not commonly enough observed, the Last Will is an ultimate rule because it is [contained in] both *ius civilis* (Civil Law) and *ius gentium* (Law of the Nations).
3. The latter (*ius gentium*) accepts the power [of a man] to dispose of private things.
4. The former (*ius civilis*) [accepts the same] from public authority.
5. The latter (*ius gentium*) grounds the execution of [one's last] will by way of a declaration.
6. The former (*ius civilis*) [concerning distribution by way of declaration] is in observance of the law in every way.
7. The latter (*ius gentium*), finally, makes it possible for all men to dispense of a thing in some

manner.

8. The former (*ius civilis*) indeed [supports the dispensing of] those [things] for which the law or a Prince allows.
9. Because a Last Will of *ius gentium* through Pontifical Law is valid, we therefore distinguish [those things] which [are so dispensed].
10. It is not any custom, but only an innate fairness [that] is attended [to].
11. NOT YET TRANSLATED
12. NOT YET TRANSLATED
13. NOT YET TRANSLATED
14. Concerning one's own position it is neither doubtful nor in error.
15. It is neither insane nor mad.
16. It is not extravagant, for which an administrating of good things is prohibited.
17. NOT YET TRANSLATED
18. NOT YET TRANSLATED
19. NOT YET TRANSLATED
20. NOT YET TRANSLATED
21. NOT YET TRANSLATED
22. NOT YET TRANSLATED
23. NOT YET TRANSLATED
24. NOT YET TRANSLATED
25. NOT YET TRANSLATED
26. NOT YET TRANSLATED
27. NOT YET TRANSLATED
28. NOT YET TRANSLATED
29. NOT YET TRANSLATED
30. NOT YET TRANSLATED
31. NOT YET TRANSLATED
32. NOT YET TRANSLATED
33. NOT YET TRANSLATED
34. NOT YET TRANSLATED
35. NOT YET TRANSLATED
36. NOT YET TRANSLATED
37. NOT YET TRANSLATED
38. NOT YET TRANSLATED
39. NOT YET TRANSLATED
40. NOT YET TRANSLATED

NOT YET TRANSLATED

Further Documentation

Descartes' 1616 Law Thesis -- Latin Translation

Descartes' 161 Law Thesis -- Copy of Original Document

Acknowledgements

I am indebted to Daniel Garber for providing me with a copy of the original Latin Law Thesis.

Kurt Smith
ksmit4@bloomu.edu

Return to Supplementary Document: Descartes's Law Thesis

First published: January 2, 2002
Content last modified: January 2, 2002

Robert Desgabets

Dom Robert Desgabets (1610-1678) was an earler defender and teacher of the Cartesian philosophy at St. Maur in the region of Lorraine, France. He was born in Ancemont and in 1636 became a monk in the Benedictine order. He taught theology at Saint-Evre at Toul between 1635-1655, and served as Procurer General of Mihiel to Paris during 1648-49. Although he is little-known today, he played an important role in the development and transmission of the Cartesian philosophy, especially in Paris and Toulouse. His major philosophical writings only appeared in print in 1983.[1] His contributions include pioneering work in the study of blood transfusion and mechanics, and his defence and developement of the Cartesian philosophy. His unusual marriage of Cartesianism and empiricism challenges many standard views of Descartes and the Cartesian philosophy.

Life and Writings

Although Robert Desgabets (1610-1678) is perhaps the most original of the Cartesian thinkers, even lauded by his most famous student, Pierre-Sylvain Régis, as "one of the greatest metaphysicians of our century," (Regis, 1704, p. 328.) only one book and two small works were published during his lifetime.[2] His correspondence indicates that he was interested in mechanics before 1644, well before being acquainted with Descartes; and for a metaphysical thinker like Desgabets, it was Descartes, not Galileo or Bacon, who offered a new and complete philosophic system. In Desgabets's estimation, the only legitimate rival system to Descartes's was the one developed by Pierre Gassendi, but in the final analysis, new scientific discoveries weighed decisively in Descartes's favor.

In 1658 Desgabets participated in the Cartesian conferences held at M. de Montmort's, where he

reportedly participated in discussions with Rohault, Clerselier, and Cordemoy. Desgabets delivered a lecture which outlined his invention of an apparatus and procedure for blood transfusion, but he seems to have abandoned its study shortly thereafter. In 1667, after a controversy erupted between the English and the French over who first invented the procedure, a physician by the name of Jean Denis was spurred to publish the written version of Desgabets's lecture. This appeared in 1668, four years after the physicians Clark and Henshaw of England had attempted the operation on animals without success. (Desgabets, 1668) Prior to this research, in 1656, Christopher Wren successfully injected medications directly into the veins of animals, after which in 1665, another Englishman and physician, Lower, a teacher and friend of John Locke, successfully injected blood into animals using this same method. Rodis-Lewis has sorted out many of the details of this history, and shows that the two procedures created by Lower and Desgabets are so different as to confirm the independence of their inventions. (Rodis-Lewis, 1974) But whether and when Desgabets experimented with his procedure has not been firmly established. What is evident, however, is that Desgabets, like Wren, was inspired by Harvey's discovery of the circulation of blood. Once Harvey had shown how the circulation of blood is best understood as a mechanism operating according to lawful movements, he opened the way for blood transfusion to be understood along similar lines, as a species of the communication of movement. Where evidence of Desgabets's actual experimentation with blood transfusion is lacking, his descriptions of it show that he was aware of the possibility of shock if the quantities transferred were too great for the subject. (Rodis-Lewis, 1974)

Throughout his lifetime he engaged in many theological and philosophical controversies with such distinguished seventeenth-century thinkers as Mabillon, Rapin, Foucher, Malebranche, Cordemoy, Arnauld and Poisson. One of his more famous interlocutors was the Cartesian Gérauld de Cordemoy (1626-1684). Despite his admiration for Cordemoy, Desgabets was shocked by the atomism in the *Discernement du corps et de l'âme* (1666), a copy of which had been sent to him by Clerselier in the year of its publication. This resulted in a letter written by Desgabets to Clerselier, opposing certain developments favourable to the existence of the void and against the infinite divisibility of extension. According to Desgabets, the marriage of Cartesian and anti-Cartesian elements in this work formed an irreconcilable schism in the Cartesian philosophy. Although Desgabets himself was not one to adopt Cartesianism in its entirety, in his view his own criticisms perfected and maintained the integrity of the Cartesian principles, while Cordemoy's adoption of atoms and the void was a direct affront to the Cartesian metaphysics.[3]

Another, much more scandalous interchange took place between Desgabets and Thomas Le Géant between 1671-1672 over Desgabets's thoughts on the Eucharist, which he stated in an anonymous work, the second of his publications, *Considérations sur l'état présent de la controverse touchant le T. S. Sacrement de l'autel* (1671). Lemaire credited this work with having been the primary cause of the persecution of Cartesianism in France, since it brought to light the incompatibility of the Cartesian philosophy with the official Church doctrine on the mystery of the Eucharist.(Lemaire, 1901, p. 124) Desgabets had entered this on-going debate in 1654 at Clerselier's request. Desgabets's intention was to defend the Cartesian doctrine of material substance against the Peripatetic doctrine of substantial forms in his explication of transubstantiation. Clerselier and Rohault had defended Descartes's ideas on the subject along similar lines, but no one had been willing, either privately or publicly, to argue as Desgabets eventually did, that the body of Christ is really extended in the host. In addition, Desgabets's

persistence and perhaps even imprudence, pushed the issue into the open. For, it was shortly after the publication of the anonymous *Considérations* in 1671, that Desgabets sent additional writings on the topic to Abbey Le Roi, who communicated them to Nicole and Arnauld. The latter found Desgabets's views dangerous and completely against tradition. It was through his acquaintance with Nicole and Arnauld, that the non-Cartesian Le Géant learned of the document and the identity of its author. Le Géant alerted the Procurer General of the Congregation of Benedictines, who ordered Desgabets to report to his superiors concerning the matter. This led to an interrogation and the subsequent issuance of an order on the 15th of December 1672, which demanded that Desgabets renounce his views on the Eucharist.(Armogathe, 1977, pp. 25-26) Desgabets promised to obey, and retreated to a monastery at Breuil. Fortunately this did not spell the end of his philosophical career, since the controversy attracted the attention of Cardinal de Retz, who was known for his radical spirit of reform among conservative ecclesiastics in France. Cardinal de Retz, who was a partisan of the new Cartesian philosophy, provided protection for Desgabets and invited him to the Cartesian conferences held at le Chateau de Commercy. It was here that Desgabets criticized and corrected what he saw as the errors of Descartes, and completed his "indefectibility thesis," which he had started in 1653-1654.

Desgabets's last published work, *Critique de la critique de la Recherche de la vérité* (1675), was intended as a defence of Malebranche against the sceptic Simon Foucher. However, Malebranche did not share this assessment, since he responded that though he was pleased with the person he found in Desgabets, he was not extremely pleased with the contents of his book; and that if one were to take on the defence of another, one ought to better know the other's thoughts. Desgabets himself was never to see Malebranche's scolding response which was published with the third edition of the *Search After Truth*, since Desgabets died at Breuil on March 13, 1678, just fourteen days before it appeared.

Desgabets's most important philosophical work, *Supplément à la philosophie de M. Descartes* (1675), was intended as a supplement to Descartes's *Meditations*. In this work, Desgabets examines many of Descartes's important doctrines and arguments. Desgabets defends the Cartesian doctrines of sensible qualities, matter, mind-body dualism, mind-body union and interaction in man while criticizing Descartes's argument for the *cogito* as the first principle of knowledge, the claim that there is such a process as pure intellection, that there are innate ideas, and that ideas have objective reality (pure possibility). What is particularly interesting is the central importance that Desgabets gives to the role of sensation in knowledge, and his development of Descartes's treatment of truth as both eternal and immutable but in some sense (one much contested in the literature) contingent. Desgabets more than once remarks in this work that "M. Descartes is not always a good Cartesian," which typifies his conviction that Cartesianism is more than the sum of the particulars set down by Descartes himself.

Viewing Desgabets's work as a whole, there is no doubt that what he viewed as a revision or perfection of the Cartesian philosophy others have viewed as a fundamental departure. In his favour, he never strayed from the Cartesian metaphysics, i.e., its substance dualism of mind and matter, substance-mode ontology, mind-body union and interaction, and the view that extension is the essence of matter and thought the essence of mind; and he remained loyal to the Cartesian physics against that of the atomists. However, heretically to some, he strongly rejected the rationalist epistemology which often dominates in Descartes, and argued that Descartes's own principles favour a form of empiricism.

Metaphysics

Despite the originality of many of Desgabets's ideas, he was orthodox in his Cartesianism. (Watson, 1966) Desgabets adhered strictly to Descartes's doctrine of matter as body extended in three dimensions. Physics, as Desgabets conceived it, was both mathematically and metaphysically grounded in the solid of the geometer. He also maintained the substance dualism of mind and matter while retaining their causal interaction. And he, like Descartes, insisted upon the substantial unity of man and the specificity of his nature and defended the Cartesian interpretation of ideas against such non-orthodox Cartesians as Malebranche. Although he claimed that the nature of the mind-body union is "the most impenetrable thing in the world," he had much to say about its sensory basis and its operations.

Desgabets subscribed completely to the Cartesian doctrine of matter which holds that matter is a substance extended in length, width and breadth. It was posited against the prevailing Peripatetic view which made reference to such things as substantial forms and prime matter in order to explain the permanence underlying the continuous change that matter undergoes. On Descartes's view, the essence of matter is its extension in three dimensions, and all of its changing attributes, properties and modes, such as movement, rest, figure, situation and composition of parts, are completely dependent upon and follow from this unified extension. (OPD 2: 27) Matter, or corporeal substance, is clearly known when it is viewed as the geometrical object of mathematicians, as a magnitude extended in length, width and breadth. Likewise, corporeal bodies are best viewed as the movements, rest, figures, arrangements and sizes of which corporeal substance is capable, and which, in its various diversifications and assemblages, "... now pass for the form of all particular bodies of which the world is composed." (OPD 1: 3) And, in all of this, "... there is nothing which is not governed by the laws of mechanics." (OPD 2: 4)

Furthermore, Desgabets viewed bodies as portions of matter, which in the case of animals are highly delicate and organized machines composed of an infinite number of parts capable of an infinite diversity of movements. He held this view against that of the Peripatetics who believed that there must be an internal principle of thought in beasts because of the intricacy and apparent human-like intelligence of their movements. According to Desgabets, the kind of thinking that attributes such an internal principle to beasts is the same kind of thinking that results in the attribution of intelligence to clocks, "However, the same affront to reason and philosophy is committed by the Americans and the Barbarians of the Orient who not being able to understand the mechanical reasons for the movement of clocks, or the true causes of natural effects, attribute souls and intelligence to machines, and likewise to fire, lakes etc., and in doing so expose themselves to the mockery of Europeans." (OPD 4: 132-133) In other words, no appeal to internal principles, or final causes is required to explain the movements of beasts, any more than it is needed to explain the movements of clocks. The complexity of movements, whether involved in the operations of clocks or those of animate bodies, is explicable in purely mechanistic terms.

Likewise, all motions communicated within or between material bodies are effected by physical contact and proceed according to the laws of local motion. Desgabets was careful to stress that the ultimate source of movement is not matter but God, who imparted it to the universe in its creation. For Desgabets,

it is this fact that grounds the laws of nature: "It is the constant and uniform manner of the action of God which founds these laws, by means of which He forms and maintains this beautiful harmony in the world which is one of the greatest objects of our sciences." (OPD 1: 13) As Desgabets saw it, the operation of these laws of nature and the rules of the communication of movement were the true and unique foundations of the new physics. In this respect, it can be seen how Descartes's account of mechanism and his distinction between principles of matter and principles of thought comport well with Desgabets's world view.

Desgabets also subscribed to the Cartesian conception of mind as an immaterial substance whose essence is thought. The distinction between mind and matter is thus a real and substantial one, impossible not to perceive: "Never has an infant asked for lies or truths for breakfast, nor has he imagined that the stones encountered along his path were the gross thoughts of some countryman." (OPD 5: 197) Minds, or immaterial substances, are of three kinds: uncreated, which is God; mind detached from body, which is an angel; and mind united with organized body, which is a reasonable soul. The second of these, angels, are the only pure minds in the created universe and they have no corporeal extension, no local presence or correspondence to time—they are simple and indivisible existence. Such minds or spirits cannot be perfected by any substantial union since their specific spiritual being does not require anything corporeal to carry out its functions. However, they are capable of participating in the movements of the visible world, though in doing so, they must undergo a kind of degradation and punishment in order to receive pain. The third of these spiritual beings is man, who consists of a cross between purely intellectual and purely corporeal things. In this Desgabets appears to have strayed slightly from the official Cartesian doctrine of dual substances, since he claimed that outside of God there are three sorts of simple created substances, matter or body, angel, and one which is composed of body and soul, which is man. However, this tripartite division of substances is supported by a more fundamental bipartite division of material and spiritual substances, so that man is best understood as a state of being which emerges out of the conjunction of two substances, rather than as a substance in the primary sense. This is not so unlike Descartes's account of man as a "composite entity" possessing two principle attributes, namely extension and thought, a being which itself is not a simple substance.

It was an important point for Desgabets that man, unlike angels, is a being composed of body *and* soul, who continually experiences the union of the two substances by the endless impressions he finds in himself. The relation of mind and body in man is an essential one which must not be regarded as a penal state of the soul, but as the *accomplishment* of its natural perfection. Human thought has duration, succession, a beginning and an end—qualities that depend on the movements of the corporeal organs and which follow the rules of local movement. Although Descartes and Malebranche after him claimed that the human mind is capable of detachment from the body, Degabets rejected this, since the human mind is not like an angel's but requires continual commerce with the senses for all its operations.

One further issue concerning Descartes's general account of the nature of the mind-body union, which Desgabets raised in order to dispel, concerns the common complaint that the Cartesian view fails to account for the interaction of material and immaterial substances. Desgabets did not regard this as a true problem because he thought that the question demanded the impossible—an explanation of how the organized body and reasonable soul, which are in fact made for each other, *can* exercise a mutual

commerce. He likened this to asking an artisan to explain how the convex surface of a peg *can possibly* fill that of a concave hole. Neither Descartes, nor the artisan could have an explanation for these states in terms of the four causes because they lay at the very foundation of metaphysics and relate directly to what Descartes identified as primitive notions. It is akin to asking why extension is the essence of body, or why thought is the essence of mind.

What could be understood, according to Desgabets, is the way in which the mind and body are united and dependent upon one another. He found an analogue in the nature of the body-body relation: two bodies are united when their superficies touch and their movements take on a mutual dependence. Similarly, two minds are united when their thoughts and wills agree and depend on one another. The union of mind and body is not by touch, nor by agreement of thoughts, but by the dependence that exists between certain thoughts and certain movements such that one in fact *follows from* the other. The union of the mind and body in man is so strict that it founds a species of communication for their commerce, in virtue of which our thoughts are said to have movement, duration, succession, etc., without themselves being corporeal. This, Desgabets insisted, is no more difficult to understand than how two surfaces, one concave and the other convex can be strictly united. Experience tells us that the mind and body depend upon one another in a mutual commerce, and their dependence is the essence of the nature of the mind-body union, and cannot be further reduced in explanation. In short, the essence of man, who is composed of mind and body, *is* the union of mind and body, and the only legitimate questions which concern this union relate to how, *given* the union, the mind determines the body and the body determines the mind, which is an empirical question. These hows must conform to the laws of local movement which apply equally to all created substances. As we will see, according to Desgabets, the mind depends on the movements of the sensory organs for its operation, from which result the modes of our thought consisting in duration, succession, and continuation; and the body depends upon the movements of the will, from which results the course of the voluntary movements of the body. Once the primacy of the mind-body union is recognized, it becomes clear that mind and body interact in virtue of the union. In fact, neither the mind nor the body move anything *per se*, but rather, they can only determine the course of movements in each other. For, recall that movement itself is imparted to the universe by God who is the only true author and "unique motor" of movement.

The metaphysics of human beings as modal beings whose nature depends upon the substantial union of mind and matter have some interesting epistemological consquences. For, it precludes the possibility of any detachement of the mind from the body, such as in pure intellection, since it would mean the destruction of the mind-body union, without which no process of thought would be possible. By man's nature, as a composite of two substances, the human mind depends upon the body for its operation just as the operation of the body depends upon the direction of the will.

Epistemology

In Desgabets's estimation, Descartes's truly great discovery was his identification of the true nature of sensible qualities. (OPD 5: 164) As Desgabets understood it, sensible qualities "... are nothing else in

objects but the local dispositions of the small parts from which result the sensations that we call heat, sound, light, etc."(OPD 2: 17) Moreover, sensible qualities *qua* modes of the mind have no resemblance or similarity to the modes or accidents of matter. Sensible qualities are states or modes of the human mind having only a causal relation to the specific local movements of our sensory organs, which in turn, are the effects of the local movements produced by the arrangements and local dispositions of parts of matter. Unfortunately, the ontological status of these qualities becomes less clear when they are considered as qualities in objects. In this context, for both Descartes and Desgabets, sensible qualities are "various dispositions" in objects. Since Descartes clearly held that sensible qualties are not the forms of material things but rather modes of the mind, the only sense in which sensible qualites could be "in" the various dispositions of objects is in the causal sense, that is, as their effects. While the status and nature of these dispositions in Descartes's account has been a subject of much debate, it is clear that he distinguished between such qualities as light, colour, smell, taste, sound and touch, which are sensed or perceived qualities, and other qualities, such as size, shape and motion, which are found in all bodies. Descartes wrote that the former qualities are dispositions which depend on size, shape and motion. So, it would seem that for Descartes, secondary qualities are sensible qualities bearing no resemblance to qualities of matter, while primary qualities are qualities belonging to physical objects.

The importance of this discovery and its interpretation in Desgabets's thought is paramount, since he believed that because of it, the way had been opened at last to lay the foundations of a true philosophy. Based on Descartes's conception of matter, and his consequent discovery of the true nature of sensible qualities, Desgabets concluded that our perception of sensible qualities and sensible objects does *not* constitute knowledge of the true state of exterior things. In fact, sensible qualities and objects considered in relation to their being as modes of the mind are not the true objects of knowledge or of science, since they are only modal beings subject to change. In this, Descartes would concur. But Desgabets went on to argue that all (true) knowledge depends on the senses, and hence on our perception of these sensible qualities and objects; this is a point of great contention in Descartes scholarship, and one of great significance in understanding Desgabets's marriage of Cartesianism and empiricism. The story is a somewhat complicated, yet intriguing one.

Desgabets claimed that the corporeal form of all particular bodies *results from* an *assemblage* of the local dispositions of matter, and that these dispositions of matter in turn come from extended matter. Thus, in the same way that sensible qualities are nothing outside us but the local dispositions of matter, sensible objects are nothing outside us but *assemblages* of local dispositions of matter. If sensible qualities such as heat, colour and light are really modes of the mind which have no resemblance to the modes or accidents of matter that cause them then, for the same reason, sensible bodies such as earth, water and animals must also be modes of the mind that have no resemblance to the assemblages of local dispositions of matter that cause them. This is a simple but perspicuous consequence of the Cartesian doctrine of sensible qualities and its conception of matter. It treats particular bodies as assemblages of local dispositions of matter, and it treats sensible bodies as our grasp of those assemblages. Hence, sensible bodies are beings of the mind which are merely caused by assemblages of local dispositions of matter. The analogy between sensible qualities such as heat, light etc., and sensible bodies such as earth, water, animals etc., is meant to be an exact one. Just as sensible qualities are the immediate effects of local dispositions of matter acting upon the senses, sensible bodies are the immediate effects of

assemblages of local dispositions of matter acting upon the senses.

Another important consequence of the true nature of sensible qualities is that the senses are not the source of error. This is a theme found in Descartes that depends on an important distinction between the scope of the two basic operations of the mind—that of the intellect and that of the will. The functions of assenting and dissenting are performed only by the will, while those of perception and conception are performed by the senses and intellect. For Desgabets, as for Descartes, simple conception is always true and conforms to its object, while error is a product of precipitous judgement. But unlike Descartes, Desgabets concluded that *everything* that is conceived of is a simple conception, not just God, soul and body. For Desgabets, a thing truly conceived, actually exists. Errors and so-called beings of reason which are said not to exist outside our conceptions, are not really conceptions, and are "... nothing but a false judgement which extends itself beyond perception. Likewise the failure that is often attributed to the senses is equally nothing but a precipitous judgement by which one says that the senses do not know." (OPD 6: 227) According to Desgabets, after having distinguished simple ideas from judgement in the *Third Meditation*, Descartes erred gravely in placing chimeras and other such beings of reason among the number of objects of simple conception. It was this that led to Descartes's fundamental error—the reversal of the basic truth that the first operation of the mind has only *real* things for its object. By including chimerical beings among the objects of conception Descartes opened the door to the possibility of thoughts without true objects. But, judgement or the extension of the will beyond the domain of what is conceived is the true source of chimeras and error. (OPD 4: 103)

All conceptions conform to their object, and error consists in the will forming a judgement beyond conception. On this point, Desgabets's account of error looks very much like that found in Descartes. However, in Desgabets, the domain of simple conception is less restricted, as is the proper function of the senses. The senses not only serve to determine what is good or harmful in life, but they also serve in knowing the nature of things. However, we often form false judgements of the true nature of the movements experienced by the senses because of their role in preserving life. (OPD 6: 228) On Desgabets's account, it is because the senses serve to promote the preservation of our life by telling us immediately which particular bodies are harmful or profitable that they can not also provide *immediate knowledge* of the true nature of these movements that impinge upon us. If the senses were such that they gave us immediate knowledge of the true nature of the movements impinging upon them, then they could not tell us *at the same time* what particular bodies are good or harmful to us. Thoughts are successive, and so the senses, in order to preserve life by warding off immediate and unexpected dangers, must be able to respond first and foremost to the preservation value of a particular object.

This does not mean that the senses cannot penetrate the true nature of things as well. In fact, for Desgabets, since all knowledge comes from the senses, the function of knowing the essences of things is a necessary role served by the senses. Recall that the true nature of particular objects consists in the local dispositions of parts of matter which produce the course of the movements we experience. These movements are the first and sole contact we have with the material world, and so the senses present, in some form, not only the existence of a given object, but first and foremost, the value of that object in relation to our survival. We only reach knowledge of the true nature of material objects by considering their being in relation to their substance. In order to achieve this, we must abstract from all relations of

time, and only then see it as it is in itself. What the senses offer us is knowledge of matter in its various divisions, shapes, sizes, etc., that is, as matter exists at a certain time and place. Thus, what Desgabets seemed to have assumed was that the reason we so often err with regard to the true nature of these movements is that we judge precipitously, or too quickly, before the mind has had a chance to conceive of material substance as it is in itself, independent of any temporal or spatial ties. Worth noting is that Desgabets's explanation for the mind's natural tendency to judge precipitously derives from the other function of the senses, namely, to immediately grasp the preservation value of a given object.

But what on Desgabets's account is the difference between a simple conception and a precipitous judgement? What is the criterion for distinguishing one from the other? In his *Supplément*, he offers the example of imaginary space. When we speak of imaginary space, many are persuaded that it does not include any judgement on their part, but if one looks closely, one sees that the simple object of conception here is space and extension with its perceived dimensions *from which* we form a judgement concerning its imaginary nature. Similarly, when someone approaches fire and has the sensation of heat, pain or pleasure, he may be persuaded that the heat he feels is in the fire just as he perceives it, since this sensation is known to him clearly. But, he has erred because of a tacit judgement. His conception of the fire and of the heat, as well as of his intervening judgement are all the immediate and true objects of his thought. These all qualify as simple conceptions in so far as they relate to their true object. But, that the heat is in the fire by which it is claimed that the heat in us resembles that of the fire, is a tacit judgement which is false. In this sense, the illusions that are attributed to the senses always involve false judgement. It is a genuine question to ask about a baton partially submerged in water whether it is straight or curved, but whether the baton, straightness, curvature etc., are real things, is not. Simple conceptions are the mind's grasp of things as they are in themselves, which is to say, in relation to their essence as extended substances, or as thinking minds. Space, as an attribute of matter, is extended in three dimensions, and sensations, as attributes of the mind, are perceptions not qualities of material objects.

This explains what Desgabets meant by his claim that all simple conceptions always have an existent object outside the understanding, since simple conceptions are, in virtue of their relation to substances, *of* things which actually exist. It is only when one judges of things outside one's conception that one is erroneously said to conceive something which does not exist. The task in the search after truth, then, is to separate precipitous judgements from judgements based only on what is *simply*, and hence, actually conceived.

Desgabets drew yet another, final consequence from Descartes's discovery of the true nature of sensible qualities (and bodies). This consquence bears directly on our ability to see the right path to a true physics which was hitherto impossible. The true physics is one that recognizes that "everything that happens in matter by the different movements and modes of its parts, belongs to mathematics and mechanics, which have all of this for their object." (OPD 5: 166) The object of the true physics is not the substantial forms of the Scholastics, but matter whose essence is extension, and its modes which are nothing else but various divisions and groupings of its parts. Thus, the natural object of the physcists is the same as the solid of the mathematician. The solid of the mathematicians, according to Desgabets, consists in a magnitude that has three dimensions, i.e., length, width, and thickness while the natural body of the physicist consists in a solid substance extended in three dimensions. This is the crowning statement of

Desgabets's metaphysical view that the physical world is *really* a single object or substance whose parts, under various divisions, shapes and arrangements, form all the *appearances* in the "grand theatre of nature." (OPD 5: 166)

This, of course, does not make the world, or its particular inhabitants any less real (or any more phenomenal) in Desgabets's view. Quite to the contrary, Desgabets thought that the foundation he had laid out supports the natural realist in each of us. The individuation of matter into sensible objects is a mode of thought in so far as the mind gives particular objects (local assemblages of dispositions of matter) their extrinsic form, but that such conceived or known objects actually exist outside the understanding by an intrinsic form, as modes or states of matter.

In Part 1, chapter three of the *Supplément*, Desgabets explicitly examines Descartes's rejection of the empiricist motto, that all knowledge comes from the senses. While Desgabets rejects the Scholastic rendering of the motto, he is equally critical of Descartes's precipitous evaluation of it. According to Desgabets, the proper sense of the empiricist motto is that all thoughts *originate* by the senses [*a sensu*] rather than in the senses [*in sensu*], and what reaches the intellect is not what is found in the senses. The soul must always be in commerce with the senses, and though our thoughts depend on the corporeal traces in the brain for their source or origin, it does not follow that our ideas must be corporeal, or even similar to corporeal things. The very fact that all of our thoughts have a beginning, duration, cessation, and sucession proves that they depend on motion, and motion is only communicated through the senses. This is why he modifies the empiricist motto from: *nihil est in intellectu quin prius fuerit in sensu*, to *a sensu*.

It is not surprising then, that Desgabets argued against Descartes's "pretention" that he could detach himself from all commerce with the senses, which sets the rationalist tone of the *Meditations*. Of course, Descartes saw this detachment as desirable because he thought it was in virtue of it that he could defend the certainty of human knowledge. He thought that the soul is known more clearly than the body because we, as thinking things, are intimately tied to the immaterial, thinking substance, which is our soul. The body, on the other hand, is part of material not immaterial substance, and so is not known immediately. It is because of this that Descartes erroneously took the *cogito* to be the foundation of certainty in human knowledge, and our knowledge of body to be less clear and less immediate. But had Descartes reflected more closely on the intimate union of the mind and body in man, he would have concluded with Desgabets, that our ideas of body and mind are equally clear, and equally evident. For, the soul of man is not an immaterial substance, but a result of the union of mind and body, and all our ideas, even of the soul, equally depend upon the operation of the senses. The new foundation of certainty in Desgabets's scheme is the principle of intentionality, from which Desgabets claimed to have demonstrably derived the principle of clarity and distinctness.

Despite Descartes's errors on this point, Desgabets claimed to have drawn the empiricist thesis out of Descartes's own principles. For, though Descartes had erroneously persuaded himself that the body had no part in metaphysical reasoning, Desgabets insisted that this does not mean that Descartes advocated the extravagant opinion that man has a pure mind. In other words, although Descartes held the mistaken view that the mind is capable of pure intellection, he did not hold its metaphysical counterpart—that man

is a pure mind. Descartes rightly viewed the metaphysical nature of man to be an essential mind-body union. In fact, Descartes himself proves that the soul is united to the body in the course of the *Meditations*. It is because Descartes and Malebranche did not attend closely enough to this latter truth concerning the mutual dependence of the soul and the body in their consideration of the nature of metaphysical reasoning that they erred with respect to the senses.

According to Desgabets, both Descartes and Malebranche rejected the empiricist thesis because they erroneously thought that it would commit them to the materialist thesis that thoughts and the soul are material. This led them to adopt the intellectualist thesis, or what we would call the rationalist thesis, regarding the mind's perception of metaphysical essences. However, Desgabets saw a third option, one which is founded on the mind-body union and the sensory basis of all knowledge, and one which founds the proper sense of the empiricist motto, that all ideas originate by the senses.

According to Desgabets, it is undeniable that man is a being who reasons, draws consequences, does not see things indivisibly, who has thoughts in succession which begin, continue, and finish, and who often experiences doubt and conjecture. Such doubting, discursive reasoning and succession of thought prove that all thought is tied to the body since duration and successive extension are nothing but the local movements of the body. (OPD 7: 299) A pure thought, the kind that Descartes (and Malebranche after him) envisioned for metaphysical reasoning, would have no beginning, duration, end, or succession. In short, such a thought would be indivisible, and hence unthinkable by the human mind.

Desgabets, having declared that all thoughts depend on movement, was especially concerned to defend his empiricist thesis against charges of materialism. Since movement is a mode which belongs to body, many tended to conclude that if minds depend on movement, then minds must have something corporeal in them. How could it be that our thoughts, which have something corporeal in them, namely, movement, are not themselves corporeal? And yet, this is exactly what Desgabets argued. He believed that thoughts have something corporeal in their being without themselves being corporeal—in the same way that every object in our thought has a beginning, continuation, and end without these objects having duration in themselves. While things really have duration, it is only *extrinsically* and by thought, "the same way a pole is divided into ten feet when one imagines the ten feet." (OPD 7: 299)

At first glance, Desgabets's analogy between attributing a corporeal nature to thoughts and attributing a duration or division to things serves more to confuse than clarify. But if we draw on his earlier point about the true nature of individual bodies, we can make some sense of it. Recall that individual bodies are real in that they result from the causal effect of assemblages of local dispositions of parts of matter acting on our sensory organs which in turn act on the mind. But the *individuation* of matter into *sensible* bodies is essentially an operation performed by thought, even though it has its foundation in the local dispositions of matter itself. For example, our perception of a ten foot pole is the effect of a specific assemblage of local dispositions of matter acting upon our sensory organs at a certain time *and* the imagination's division of that local disposition into ten feet. The local disposition itself is not intrinsically a pole, or ten feet long, but it is so extrinsically, that is, by thought. Similarly, thoughts themselves are not intrinsically extended or subject to movement, but they are so extrinsically, that is, as an effect of the operations of the sensory organs of the body. In other words, just as the local movements of matter

individuate thought extrinsically giving the mind individual thoughts, the thought of immaterial substance individuates matter extrinsically thereby giving it individual bodies. The mind-body union and the essential intentionality which results from it create the mutual dependence of the operations of mind and matter without requiring that mind be material or matter be immaterial.

This is at the heart of Desgabets's "fundamental truth," that for the proper use of reason we must recognize that all our ideas or simple conceptions have a real object outside of it, which is in itself what is represented by thought, and which actually contains the degree of being that one sees there. If individual and sensible objects are modal beings in the sense claimed by Desgabets then there is no cleavage between object and object known because they are one and the same thing. What is represented to our thought is the real object of knowledge, namely, material and spiritual substance, but what we sense are these substances as they exist in time, or in other words, as they exist at a given time *by thought* in virtue of the particular local dispositions of matter. It explains the sense in which objects thought of are objects actually existing outside the mind, since they are nothing more than the direct and immediate effect of actions produced in us by assemblages of local dispositions of matter.

Truth

According to Desgabets, Descartes's fine doctrine concerning the creation of the eternal truths, is the foundation of the true (Cartesian) philosophy. He claimed that had Descartes attended consistently to this doctrine, he would have avoided all error. Although there is a great deal of controversy surrounding the interpretation and significance of this doctrine, as Desgabets understood it, it requires that God be equally the author of all created things, in their being and their essence. It requires that there be no essence without existence, which is to say, that there can be no purely possible beings either in the human mind as in Descartes, or in God's mind, as in Malebranche. The notion that such truths are in some way prior to God's creation of the world, existing as purely possible essences separate from actual existence, involves a separation of essence from existence, and supposes that essence is something actually separable and conceivable without existence.

But, as Descartes recognized, the distinction between essence and existence is a distinction of reason; essence is the nature of a thing as it is contained in its definition, and existence is the perfection by which this essence actually exists which is possessed by the action of the Creator. According to Desgabets's account of the creation doctrine, "... objects precede truth in the order of nature," such that, "... it is impossible that the whole be greater than its part if there is no whole or parts, which is to say in a word, that our principle is found yet more true, and it is impossible to think of nothing." (OPD 6: 232) The principle Desgabets refered to here is the principle of intentionality, which states that every idea is an idea of some thing which exists as it is represented by thought. Becauses objects precede truth in the order of nature, to think of an object is to think of something that really exists. The correspondence of our thoughts to reality, that is the truth of our ideas, depends on something that actually exists and not on something independent of existence.

Truth, which exists only by the relation of conformity of thought to its object, is contingent in that it

depends on God's will in his free creation of the universe. Truth is eternal in its independence from time or temporal variations, and it is immutable in that once God wills the world, things never change in respect to their substance but only in respect to their modes of being. Although we must wait to know what things God actually created, we are guaranteed of the truth of our ideas since it is the object itself that determines what we perceive. Truth is necessary in that God gives all objects He creates an irrevocable being, but it is contingent in relation to His absolute and unlimited power. God is the equal author of the essence and the existence of things He created, He "... gave them their essence and their existence which are equally contingent, and which once received, are nevertheless possessed by them irrevocably." (OPD 6: 249) Desgabets's voluntarism, then, is a qualified one, for though the eternal truths depend upon God for their existence as their primary cause, they are no less indefectible, that is, they are unchangeable in their substantial being.

Bibliography

Primary Texts

- Cordemoy, Gérauld de. (1968) *Oeuvres philosophiques*, eds. Pierre Clair & François Girbal, Paris: Presses Universitaires de France. [This work includes letter written by Desgabets (1666) in which he argued against the atomist thesis of Cordemoy.]
- --------. (1666) *Discernement du corps et de l'âme*, Paris.
- Desgabets, Robert. (1668) "Discourse de la communication ou transfusion du sang," published with, "Lettre ècrite à M. Sorbière," by J. B. Denis, Paris. [This is a scientific piece in which Desgabets described an apparatus and procedure for blood transfusion of his own invention.]
- --------. (1671) *Considérations sur l'état présent de la controverse touchant le Très Saint-Sacrement de l'autel*, published anonymously, Holland. [This is the work that stirred a great deal of controversy for Cartesians, which is not surprising given the theologically sensitive nature of the thesis that the body of Christ is actually present (extended) in the host.]
- --------. (1675) *Critique de la Critique de la Recherche de la vérité*, Paris.
- --------. (1983) *Oeuvres philosophiques inédites*, Analecta Cartesiana 2, ed., J. Beaude with introduction by G. Rodis-Lewis, Amsterdam: Quadratures.
- Régis, Pierre-Sylvain. (1704) *L'usage de la raison et de la foi*, Paris.

Selected Studies and Critical Discussions

- Armogathe, J.-R. (1977) *Theologia Cartesiana: L'Explication physique de l'Euchariste chez Descartes et dom Desgabets*, The Hague: Martinus Nijhoff.
- Beaude, Joseph. (1974) "Desgabets et son oeuvre," *Revue de sythèse* 95:7-17.
- ---------. (1979) "Cartésianisme et anticartésianisme de Desgabets," *Studia Cartesiana 1*, Amsterdam: Quadratures, pp. 1-24.
- ---------. (1980) "Le Guide de la raison naturelle dans l'oeuvre de Desgabets," *Recherches sur le XVIIe siècle IV*, Paris: Centre National de la Recherche scientifique.

- Cousin, Victor, (1852) *Fragments de philosophie cartésienne*, Paris: Didier; reprinted (1970) Geneva: Slatkine Reprints. [Volume III includes selections from unpublished manuscripts which contain discussions by Retz, Malebranche, and Corbinelli of Desgabets's revision and extension of Descartes's philosophy.]
- Lemaire, Paul, (1901) *Le Cartésianisme chez les Bénédictins: Dom Robert Desgabets son système, son influence et son école,* Paris: Alcan.
- Lennon, Thomas M, and Easton, Patricia A, (1992) *The Cartesian Empiricism of François Bayle*, New York: Garland, 1992.
- Lennon, Thomas, M. 199? “The Cartesian Dialectic of Creation,” in: *The Cambridge History of Seventeenth Century Philosophy*, eds., M. Ayers and D. Garber, Cambridge: Cambridge University Press.
- Prost, Jean. (1907) *Essai sur l'atomisme et l'occasionalisme dans la philosophie cartésienne,* Paris. [This work includes useful material on Desgabets's critique of Cordemoy's atomism, and his rejection of occasionalism.]
- Robinet, André, (1974) “Dom Robert Desgabets, le conflit philosophique avec Malebranche et son l'oeuvre métaphysique,” Journée Desgabets, Revue de synthèse 95:65-83. ("Dom Robert Desgabets, his Philosophical Conflict with Malebranche and His Metaphysical Work.") (This artical examines Desgabets's influence on Malebranche which is found in Malebranche's views on occasionalism, the non-materiality of thoughts, and the nature of the eternal truths.)
- Rodis-Lewis, Géneviève, (1993) "Der Cartesianismus in Frankreich," Grundriss der Geschichte der Philosophie, Die Philosophie des 17. Jahrhunderts, Bond II, Basel/Struttgart, pp. 398-445.
- ----------. (1981) “Polémiques sur la création des possibles et sur l'impossible dans l'école cartésienne." *Studia Cartesiana 2*, Amsterdam: Quadratures, pp. 105-123.
- ----------. (1974) “L'écrit de Desgabets sur la transfusion du sang et sa place dans les polémiques contemporaines,” *Journée de Desgabets Revue de synthèse* 95: 31-64.
- Watson, Richard A., (1982) “Transubstantiation among the Cartesians,” *Problems of Cartesianism*, eds., T. M. Lennon, J. M. Nicholas, and J. W. Davis, Kingston and Montreal: McGill-Queens University Press, pp. 127-148.
- ----------. (1966) *The Downfall of Cartesianism 1673-1712*, The Hague: Martinus Nijhoff.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Descartes, René | Malebranche, Nicolas

Patricia Easton
patricia.easton@cgu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 22, 2001
Content last modified: March 22, 2001

Stanford Encyclopedia of Philosophy
Notes to Robert Desgabets

Notes

1. Dom Robert Desgabets (1983), *Oeuvres philosophiques inédites*, Analecta Cartesiana 2, ed., J. Beaude with introduction by G. Rodis-Lewis, Amsterdam: Quadratures. This essential but not widely available collection of Desgabets's philosophical writings comes in 7 opuscules; it includes works that previously existed only in manuscript, *Traité de l'indéfectibilité des créatures* (c. 1653-74), *Supplément à la philosophie de M. Descartes* (1675), *Guide de la raison naturelle* (1671), and *Traité de l'union de l'âme et du corps*. This work will be referred to by abbreviated title, OPD, opuscule number, and page number. The page numbering is continuous from opuscule to opuscule. For example, OPD 5: 152, refers to *Oeuvres philosophiques inédites*, opuscule 5, page 152.

2. All three published works were published anonymously: Dom R. Desgabets *Critique de la critique de la Recherche de la vérité,* (Paris, 1675); *Discours de la communication ou transfusion du sang* (Paris, 1668); and *Considérations sur l'état présent de la controverse touchant le T. S. Sacrement de l'autel* (Paris, 1671).

3. This letter, "Une lettre de D. Robert Desgabets à D. Jean Mabillon sur la question des azymes" is printed in: *Oeuvres des Mabillon et de D. Thierri-Ruinart* (Paris, 1724).

Patricia Easton
patricia.easton@cgu.edu

First published: March 22, 2001
Content last modified: March 22, 2001

Nicolas Malebranche

The French Cartesian Nicolas Malebranche was hailed by his contemporary, Pierre Bayle, as "the premier philosopher of our age." Over the course of his philosophical career, Malebranche published major works on metaphysics, theology, and ethics, as well as studies of optics, the laws of motion and the nature of color. He is known principally for offering a highly original synthesis of the views of his intellectual heroes, St. Augustine and René Descartes. Two distinctive results of this synthesis are Malebranche's doctrine that we see bodies through ideas in God and his occasionalist conclusion that God is the only real cause.

1. Life and Works

Malebranche was born in Paris on August 6, 1638, one month prior to Louis XIV, and died, also in Paris, on October 13, 1715, six weeks after the great French monarch. Malebranche was one of many children born to his mother, Catherine de Lauzon, the sister of a Viceroy of Canada, and his father, also Nicolas Malebranche, a secretary to Louis XIII. As in the case of Descartes and Pascal, Malebranche was born in frail health. His particular affliction was a severe malformation of the spine, and due to this condition as well as his weak lungs he needed to be tutored at home until the age of sixteen. Subsequently he was a student at the Collège de la Marche, and after graduating he went to study theology at the Sorbonne. His education left him with a distaste for a scholasticism that focused on the work of Aristotle. Thus, in 1660 he decided to leave the universities and to enter the Oratory, a religious congregation founded in 1611 by the Augustinian theologian Pierre Bérulle (1575-1629). At the Oratory in Paris, Malebranche studied

ecclesiastical history, linguistics, and the Bible, and with his fellow students also immersed himself in the work of Augustine. He was ordained a priest on September 14, 1664.

The same year he was ordained, Malebranche happened in a Paris bookstall upon a posthumous edition of Descartes's *Traité de l'homme* (*Treatise on Man*) which provides a sketch of a mechanistic account of the physiology of the human body. Malebranche's early biographer, Father Yves André, reported that he was so “ecstatic” on reading this account that he experienced “such violent palpitations of the heart that he was obliged to leave his book at frequent intervals, and to interrupt his reading of it in order to breathe more easily” (André 1970, 11-12). While André does not indicate why Malebranche was so moved, one can speculate that he had discovered in this text a way to investigate the natural world without relying on a stagnant Aristotelian scholasticism. In any case, after his encounter with *L'homme* Malebranche devoted himself to a decade-long study of the Cartesian method and its results in mathematics and natural philosophy.

The fruit of this study is a two-volume work bearing the title, *De la recherche de la vérité. Où l'on traitte de la nature de l'esprit de l'homme, et de l'usage qu'il en doit faire pour eviter l'erreur dans les sciences* (*Search after Truth. In which is treated the nature of the human mind and the use that must be made of it to avoid error in the sciences*) (1674-75). It is primarily this text which provides the basis for Malebranche's reputation in the modern period. As its full title indicates, the *Recherche* focuses on the principal sources of human error and on the method for avoiding those errors and for finding the truth. The first five books enumerate the various errors deriving from the senses, imagination, pure understanding, inclinations and passions, respectively, and a sixth book is devoted to the Cartesian method of avoiding such errors through attention to clear and distinct ideas. The centerpiece of the third book, on pure understanding, is a defense of the claim that the ideas through which we perceive bodies exist in God. Tucked away in the final book, on method, is a critique of “the most dangerous error of the ancients,” namely, the Aristotelian position that there are secondary causes in nature distinct from God.

The first volume of the *Recherche*, containing the first three books, drew an immediate response in 1675 from the abbé Simon Foucher (1644-1696), canon of Sainte Chapelle of Dijon. Foucher was an “academic skeptic” who attacked the assumption that ideas in us can represent objects distinct from us (see Foucher 1969). The Cartesian Benedictine Robert Desgabets (1610-1678) replied to Foucher by insisting that the Cartesian rule that clear and distinct ideas are true presupposes that our thoughts correspond to real external objects. In brief prefaces added to the second volume of the *Recherche*, Malebranche chastised both thinkers for failing to read the work they were discussing, noting in particular that he had explicitly argued in the *Recherche* that the ideas we perceive exist in God rather than in us.

Malebranche solicited written responses to the *Recherche* modeled on the sets of objections published with Descartes's *Meditations*. Perhaps put off by Malebranche's harsh treatment of Foucher and Desgabets, his critics offered instead only informal objections channeled through mutual friends. In 1678, Malebranche appended to the *Recherche* a set of sixteen *Eclaircissements*, or clarifications, that respond to these objections. Among the more important objections addressed are those that concern Malebranche's assertion that we have a freedom to “consent” to certain motives for action

(“Eclaircissement I”), his claim that reason provides no conclusive demonstration of the existence of the material world (“Eclaircissement VI”), his doctrine of the vision of ideas in God (“Eclaircissement X”), his conclusion that we know our own soul through a confused consciousness rather than through a clear idea of its nature (“Eclaircissement XI”), and his occasionalist thesis that God is the only true cause (“Eclaircissement XV”). In the 1678 edition there is a final Eclaircissement that defends the importance “not only for knowledge of nature but also for knowledge of religion and morals” of the view, only hinted at in the *Recherche*, that God acts for the most part through “general volitions” (*volontez générales*) and acts though “particular volitions” (*volontez particulières*) only in the exceptional case of miracles.

Malebranche developed this last point in his 1680 *Traité de la nature et de la grâce* (*Treatise on Nature and Grace*). He published this work over the objections of the Jansenist theologian and Cartesian philosopher Antoine Arnauld (1612-1694), who was disturbed by what he saw as Malebranche's denial of the assertion in the Scriptures and the tradition of God's attention to particular details in matters of grace. Arnauld's responded to the publication of *Nature et grâce* by engaging in open combat, and the ensuing battle became one of the major intellectual events of the day. His opening salvo was the 1683 *Des vraies et des fausses idées* (*On True and False Ideas*), which attacks not *Nature et grâce* but rather the *Recherche* (see Arnauld 1990). Arnauld's strategy here was to undermine Malebranche's influence in theological matters by revealing the inadequacy of his philosophical views. In particular, he attacked Malebranche's assumption that ideas are “representative beings” distinct from our perceptions, offering instead the position, which he plausibly ascribed to Descartes, that ideas are simply a feature of the perceptual modifications of our soul. This argument reflects a sympathy for Descartes' views that dates back to Arnauld's set of comments on the *Meditations*.

The same year that Arnauld presented his initial critique, Malebranche published the *Méditations chretiennes et métaphysiques* (*Christian and Metaphysical Meditations*), where “the Word” (i.e., the Second Person of the Trinity) offers a summary of his system that highlights the central role that God plays in both metaphysics and morality. This work was in some ways a follow up to his *Conversations chrétiennes* (*Christian Conversations*), published in 1677. In that earlier text, Malebranche presented a defense of the Christian religion that emphasizes the Augustinian theme of our dependence on God for knowledge and happiness. In 1684, Malebranche further developed his views on moral theory in the *Traité de morale* (*Treatise on Ethics*), in which he argued that moral virtue requires a love of the “immutable order” that God reveals to those who seek to know it.

In 1684, Malebranche also responded to Arnauld's *Idées*, and after a further exchange on the topic of the nature of ideas the debate turned to the religious issues of divine providence, grace and miracles. The battle became increasingly bitter, and as a result of a campaign on the part of Arnauld and his supporters, Malebranche's *Nature et grâce* was put on the Catholic *Index librorum prohibitorum* (*Index of Prohibited Books*) in 1690 (the *Recherche* was added in 1709). The Malebranche-Arnauld polemic continued even after Arnauld's death in 1694, with the posthumous publication of two letters from Arnauld in 1699 and of Malebranche's responses to those letters in 1704.

In 1688, Malebranche published his *Entretiens sur la métaphysique et la religion* (*Dialogues on*

Metaphysics and Religion), a concise summary of his main metaphysical doctrines of the vision in God and occasionalism that also addresses the problem of evil. In 1696, he appended to this text the *Entretiens sur la mort* (*Dialogues on Death*), which he composed after a life-threatening illness.

In 1692, Malebranche published a short study, the *Lois de la communication des mouvements* (*Laws of the Communication of Motions*), in which he endorsed Descartes's law of the conservation of the quantity of motion but offered rules governing collision that, unlike Descartes's own rules, involve no appeal to a force in bodies to remain at rest. In correspondence with Malebranche, Gottfried Wilhelm Leibniz (1646-1716) emphasized difficulties with Descartes's conservation law, and that correspondence led Malebranche to insert into a 1700 edition of the *Lois* that experience reveals that this law does not hold.

In 1693, Malebranche responded to the criticisms of the *Recherche* in the 1690 *Systême de philosophie* (*System of Philosophy*) of the French Cartesian Pierre-Sylvain Régis (1632-1707). Régis had defended an account of ideas similar to the one that Arnauld had defended against Malebranche during the 1680s, and Arnauld used the Régis-Malebranche exchange as an occasion to return to the issue of ideas during the last year of his life. Despite their dispute, Malebranche and Régis were both appointed as honorary members of this organization when it was reorganized in 1699. Malebranche presented an inaugural lecture to the Académie that defends against Descartes an account of color in terms of the frequency of vibrations of light. In later published versions of the lecture, Malebranche revised his discussion to take into account the theory of the nature of color in the work of Sir Isaac Newton.

In 1699, Malebranche also published the *Traité de l'amour de Dieu* (*Treatise on the Love of God*) with *Trois lettres à Lamy* (*Three Letters to Lamy*), in which he rejected the claim in the Benedictine François Lamy (1636-1711) that the *Traité de morale* supports the quietist position that moral action derives from a disinterested “pure love of God.” This rejection of Lamy's quietism provided the basis for Malebranche's reconciliation with the famous French cleric, Jacques-Bénigne Bossuet (1627-1704). Bossuet had earlier enlisted the aid of François de Fénelon (1651-1715) in writing against Malebranche's occasionalism and his appeals to God's “general will,” but later became a bitter enemy of Fénelon's quietism.

With the support of the apostolic vicar in China, Malebranche published in 1708 an *Entretien d'un philosophe chrétien et d'un philosophe chinois, sur l'existence et la nature de Dieu* (*Dialogue between a Christian Philosopher and a Chinese Philosopher on the Existence and Nature of God*). A sixth and last edition of the *Recherche* appeared in 1712, and in 1715 Malebranche published his final work, *Réflexions sur la prémotion physique* (*Reflections on Physical Premotion*), in which he responded to the claim of the abbé Laurent-François Boursier (1679-1749) that occasionalism leads naturally to the Thomistic position that God determines our action by means of a “physical premotion.” In his response, Malebranche defended the claim, present from the first edition of the *Recherche*, that our free action involves a “consent” that God does not determine.

Malebranche (1958-84), which consists of 20 volumes, is the standard critical edition of Malebranche's writings. The increasing popularity of Malebranche in the English language literature is indicated by the

presence of recent English translations of his writings; see Malebranche (1980a), Malebranche (1980b), Malebranche (1993), Malebranche (1997a), and Malebranche (1997b). Easton, Lennon and Sebba (1991) is a comprehensive bibliography of work on Malebranche in various languages. This work supersedes the bibliography in volume 20 of Malebranche (1958-84), which had superseded Sebba (1959).

2. Ideas and the Vision in God

In a section of the third book of the *Recherche* devoted to “the nature of ideas,” Malebranche argued for his famous doctrine of the vision in God. More precisely, the thesis in this section is that we see external objects by means of ideas in God. The argument for this thesis begins with the claim at the beginning of this section that “everyone agrees that we do not perceive objects external to us by themselves” since it can hardly be the case that “the soul should leave the body to stroll about the heavens to see the objects present there” (Malebranche 1958-84, 1:413). Arnauld later took exception to this starting point, countering that “ideas, taken in the sense of representative beings, distinct from perceptions, are not needed by our soul in order to see bodies” (Arnauld 1990, 18). His main objection is that Malebranche stacked the deck in favor of his doctrine that we see ideas of bodies in God by assuming from the start that these ideas are distinct from our own perceptions.

In developing his own position, Arnauld appealed to Descartes' distinction in “Meditation III” between the formal reality of an idea as a perceptual modification of mind and its objective reality as something that represents an object. Arnauld insisted that a representative idea is simply the objective reality of a perception, and thus not something distinct from that perception. However, it is important to note that Malebranche's definition of an idea does not rule out such a position from the start. As he himself insisted to Arnauld, the claim that we must perceive external objects through ideas leaves open the question of whether an idea is “*a modality of the soul*, according to the opinion of M. Arnauld; an *express species*, according to certain philosophers, or an *entity created with the soul*, according to others; or finally *intelligible extension rendered sensible by color or light*, according to my opinion” (Malebranche 1958-84, 6:95).

Malebranche's description of his own opinion goes beyond what can be found in the original edition of the *Recherche*. However, his description of the other alternatives is drawn directly from this text. In particular, Malebranche had argued there that there are only four alternatives to the conclusion that we see bodies through ideas in God: (1) Bodies transmit resembling species to the soul; (2) Our soul has the power to produce ideas when triggered by non-resembling bodily impression; (3) Ideas are created with the soul or produced in it successively by God; and (4) Our soul sees both the essence and the existence of bodies by considering its own perfections. Malebranche told Arnauld that since this list constitutes “an exact division … of all the ways in which we can see objects” and since each of the alternative accounts yields “manifest contradictions,” his argument from elimination serves to demonstrate the doctrine of the vision in God (Malebranche 1958-84, 6:198f).

It is difficult to determine from the *Recherche* the precise source of the enumeration. However, Connell (1967) has established that Malebranche's argument was drawn from the account of angelic knowledge in

the work of the Spanish scholastic, Francisco Suárez (1548-1617). Particularly crucial for Malebranche's enumeration is Suárez's claim that angels must know material objects through species that God adds to their mind given that God alone can know them through His own substance. In light of this claim, we can take Malebranche's first three hypotheses to cover the various ways in which we could perceive bodies through immaterial species "superadded" to our soul, and his fourth hypothesis to cover the possibility that we perceive bodies in the perfections of our soul. In arguing against the last hypothesis, Malebranche noted that since a finite being can see in itself neither the infinite nor an infinite number of beings (as Suárez had argued in the case of angels), and since we in fact perceive both the infinite and infinity in external objects, it must be that we see these objects by means of perfections contained in the only being that can possess an infinity of ideas, namely, God Himself.

Malebranche took the conclusion here to confirm the view in "an infinity of passages" in Augustine that "we see God" in knowing eternal truths. This appeal to the Augustinian theory of divine illumination provides the basis for an argument for the vision in God that bypasses the unusual enumeration in the *Recherche*. This more direct argument is introduced in "Eclaircissement X," where Malebranche urged that the ideas we perceive must exist in an "immutable and necessary Reason" since they are themselves immutable and necessary (Malebranche 1958-84, 3:129f). Malebranche emphasized that the Augustinian view that eternal truths derive from uncreated features of the divine intellect conflicts directly with the voluntarist conclusion in Descartes that these truths derive rather from God's free and indifferent will. Particularly in his exchanges with Arnauld, Malebranche attempted to present his doctrine of the vision in God as a natural consequence of Descartes' account of ideas. However, his Augustinian argument serves to show that Descartes himself could not have accepted this doctrine. Moreover, such an argument reveals the most fundamental reason for Malebranche's rejection of Arnauld's Cartesian identification of ideas with our own perceptions. Since Malebranche identified these ideas with necessary and immutable essences, and since he held that these ideas derive their necessity and immutability from the divine intellect, he concluded that Arnauld's position can only result in a radical subjectivism that renders impossible any sort of a priori knowledge of the material world.

"Eclaircissement X" also introduces the notion of "intelligibile extension" mentioned in Malebranche's claim to Arnauld quoted above concerning his own opinion. According to this Eclaircissement, God has a single ideal extension that serves to represent particular bodies to Him. Arnauld objected that this position involves a retraction of the claim in the *Recherche* that we perceive bodies by means of distinct ideas in God. In response, Malebranche insisted that his view all along was that God represents particular bodies by means of His own simple "absolute being." For Arnauld, however, the view that God contains extension in this way is objectionable since it is connected to the heretical view in the work of the Dutch thinker Benedict Spinoza that God is extended substance. The charge of Spinozism reappears in Malebranche's 1713-14 correspondence with one of his former students, J.J. Dortous de Mairan (1678-1771), who later became the Secretary of the Paris Académie des sciences. As in the case of Arnauld, so in this correspondence Malebranche vigorously denied this charge. In both cases, he responded by emphasizing that the infinite and indivisible ideal extension that exists in God differs from the finite and divisible extension in the material world.

A final feature of Malebranche's doctrine of the vision in God is connected to the notion in his writings

of the “efficacious idea” (*idée efficace*). As first noted in Robinet (1965), this notion became entrenched in Malebranche's system around 1695, after his encounter with his Cartesian critic Régis. In his *Systême de philosophie,* Régis had challenged the claim in the preface to the *Recherche* that our mind is united to God in a manner that “raises the mind above all things” and is the source of “its life, its light, and its entire felicity” (Malebranche 1958-84, 1:9). While he granted the commonplace claim that God must create and conserve our soul, Régis denied that we are enlightened by means of a union with ideas of bodies in God. Rather, he insisted that God conserves in us ideas that derive directly from the bodies they represent. In the 1693 *Réponse à Régis* (*Response to Régis*), Malebranche emphasized his Augustinian position that we can be instructed as to the nature of bodies only through a union with God. However, he put a new spin on this position when he noted that the union with God involves an “affecting” or “touching” of our mind by God's idea of extension.

Already in the 1688 *Entretiens sur la métaphysique* Malebranche had suggested that the union with God can be explicated in terms of a causal relation between God's ideas and our mind. After 1695, he developed this suggestion by introducing the notion of “pure” or non-sensory intellectual perceptions that are produced by God's efficacious idea of extension. Yet he also stressed in this later period that such an idea is the causal source of our sensations. One advantage of this extension of the doctrine of efficacious ideas to sensations is that it yields a fairly clear explanation of Malebranche's claim to Arnauld that an idea is “intelligible extension rendered sensible by color or light.” Prior to 1695, Malebranche explained how intelligible extension is so rendered by appealing somewhat obscurely to the fact that the soul “attaches” colors to a non-sensory idea. However, the theory of efficacious ideas allowed him to say that this idea is rendered sensible by causing in us the appropriate sensations of color and light. The claim that we see ideas in God is thus transformed into the claim that our soul has intellectual and sensory perceptions that yield an understanding of the truth concerning bodies in virtue of their causal relation to God's idea of extension. Drawing on Robinet's results, one scholar has concluded that while Malebranche started with the vision *in* God, he ended with a vision *by* God (Alquié 1974, 209).

3. Cartesian Dualism

Malebranche told Arnauld that it was Augustine's authority “which has given me the desire to put forth *the new philosophy of ideas*” (Malebranche 1958-84, 6:80). By contrast, he emphasized in the preface of the *Recherche* that Augustine had failed to see that sensible qualities “are not clearly contained in the idea we have of matter,” adding that “the difference between mind and body has been known with sufficient clarity for only a few years” (Malebranche 1958-84, 1:20). The allusion here is to Descartes's recent discovery of an idea of matter that reveals that its nature consists in extension alone. This idea dictates that sensible qualities such as colors, tastes and odors that are not reducible to modes of extension cannot exist external to mind. But since these qualities exist in the mind, and in particular in the mind's perception of the qualities, the mind itself must be distinguished from body. In this way the Cartesian idea of matter reveals “the difference between mind and body.”

In the initial book of the *Recherche,* on the senses, Malebranche proposed that the erroneous belief in the Aristotelians as well as in Augustine that sensible qualities exist in bodies has its source in a misuse of

"natural judgments" that help in the conservation of the human body. Here he was following Descartes' account in "Meditation VI" of the "teachings of nature," and in particular the claim there that the purpose of sensations is not to teach us about the nature of bodies but simply to inform us of what is beneficial or harmful to the human composite. Just as Descartes had urged that erroneous beliefs about the nature of body can be avoided by attending to the clear and distinct perceptions of the intellect, so Malebranche counseled that we avoid error by attending to what the clear idea of matter reveals to us about the nature of body. As we have seen, Malebranche had Augustinian reasons for saying that the idea that so instructs us exists in God. By his own admission, however, the conclusion that the idea that instructs us is an idea of *extension* derives from the recent discoveries of Descartes.

Malebranche emphasized that the clear idea of extension must be distinguished from our confused sensations. One point he wanted to make is that the idea exists in God while the sensations are only modifications of our mind. However, his emphasis on the fact that this idea is "pure" or non-sensory indicates that our experience of the material world has an intellectual component. We have seen that his late doctrine of the efficacious idea involved the position that we have pure intellectual perceptions produced by God's intellectual idea of extension. But his mature position that this idea is also the cause of our sensations allows for the claim that our most basic sensory contact with the material world has an intellectual component.

We know that Malebranche's doctrine of the vision in God conflicts with Descartes' doctrine of the creation of the eternal truths. However, there are further departures from orthodox Cartesianism that are linked to two qualifications of this doctrine. The first qualification is that God's idea of extension can reveal only the nature of bodies and not their existence. This qualification is not explicit in the initial edition of the *Recherche*, which says only that the existence of properties of bodies external to us is "very difficult to prove" (Malebranche 1958-84, 1:122). Foucher had objected that Malebranche has no good reason to affirm that the external existence of these properties. In "Eclaircissement VI," Malebranche urged that the idea of extension does reveal the possible existence of the material world, and that Descartes has shown that we have a probable argument for its actual existence deriving from our natural propensity to believe that there are bodies. However, he conceded in this text—without crediting Foucher—that neither he nor Descartes can provide an argument from reason that demonstrates "with evidence" or "with geometric rigor" that this belief is true. His conclusion is that such an argument must appeal to faith in the veracity of the report in the Scriptures that God has created the heavens and the earth.

According to the second qualification of the vision in God—which is found in the original edition of the *Recherche*—we perceive the nature of our soul not through a clear idea in God, but only through a confused "consciousness or inner sensation" (*conscience ou sentiment intérieur*). Malebranche accepted the Cartesian commonplace that consciousness reveals immediately the existence of the soul. He allowed that we know the nature of our soul to consist in thought, moreover, and he embraced the Cartesian conclusion that the soul as a thinking thing is distinct from body as an extended thing. Yet he insisted that we know that the soul is distinct from the body not by means of any direct insight into the nature of thought, but rather by seeing that thought is not contained in the idea of matter. More generally, Malebranche claimed that our lack of access to a clear idea of the soul is evident from the fact that we do

not have knowledge of thought that matches our knowledge of the mathematical features of bodies. This last point turns on its head Descartes' own conclusion in “Meditation II” that the nature of the human mind is “better known” than the nature of body; for Malebranche, it is the nature of body that is better known than the nature of mind.

In “Eclaircissement XI,” Malebranche attempted to counter “the authority of Descartes” by arguing that the Cartesians themselves must admit that they have only a confused awareness of the nature of the sensory modifications of the soul. He noted that while the intellectual idea allows the various modes of extension to be related in a precise manner, there is no clear scale on which we can order our sensations of different shades of the same color, not to mention our sensations of sensible qualities of different kinds. Malebranche took the confusion in the sensations to reveal a confusion in our perception of the nature of the soul. He added that Cartesians can discern that sensible qualities are modifications of an immaterial soul only by seeing that they are “not clearly contained in the idea we have of matter.”

4. Occasionalism

Malebranche is known for his occasionalism, that is, his doctrine that God is the only causal agent, and that creatures merely provide the “occasion” for divine action. On the old textbook account, occasionalism was an *ad hoc* response to the purported problem in Descartes of how substances as distinct in nature as mind and body are can causally interact. According to this account, Malebranche was driven by this problem with Cartesian dualism to propose that it is God who brings it about that our sensations and volitions are correlated with motions in our body.

However, occasionalism was already an old doctrine at the time that Thomas Aquinas (1225-1274) wrote against it. Thomas indicated that the primary concern of the occasionalists was to strengthen the assertion of God's omnipotence. Though he allowed that God must “concur” with creatures in producing effects, Thomas also claimed that there is reason to conclude that creatures are true secondary causes. For instance, he urged that it is more in accord with divine greatness to say that God communicates His power to creatures. Moreover, he claimed that it is simply evident to the senses that creatures have the power to bring about effects. Thomas also argued that if there were no natures in creatures that explain effects, there could be no true scientific explanation of effects through their natural causes.

Malebranche was concerned to respond to all of these arguments against occasionalism, particularly as they were developed in the work of scholastics such as Suárez. Against the first point that God's greatness requires the communication of His power, he countered that it is in fact idolatrous to attribute divine power to creatures. Malebranche's argument that God alone can produce effects relies on the assumption that “a true cause … is one such that the mind perceives a necessary connection [*liaison nécessaire*] between it and its effects” (Malebranche 1958-84, 2:316). He claimed that there is such a connection neither among bodily states, nor between bodily and mental states, nor among mental states. In all of these cases, one can deny the connections without contradiction. There can be a necessary causal connection in only one case, namely, the connection between the volitions of an omnipotent agent and its upshots. Thus, only such an agent, namely, God, can be a true cause.

In the *Entretiens sur la métaphysique*, Malebranche offered a different argument based on Descartes's suggestion in "Meditation III" that God conserves the world by continuously creating it. The argument begins with the claim that God must create bodies in some particular place and in determinate relations of distance to other bodies. If God conserves a body by creating it in the same place from moment to moment, that body remains at rest, and if he conserves it by creating it in different places from moment to moment, it is in motion. We cannot even create motion in our own bodies. Rather, it is God who must produce it on the occasion of volitional states. Moreover, it is not motions in our brain that cause our sensory states, but God who produces them on the occasion of the presence of such motions. Finally, I have indicated the view in the *Entretiens* that God produces our intellectual states through the union of our mind with His "intelligible extension." While the argument from the necessity of the causal connection yields the result that only an omnipotent being can be a cause, the argument here is that only that being which creates/conserves the world can cause various bodily and mental states. However, both arguments converge on the conclusion, which Malebranche claimed to find in Augustine, that all creatures depend entirely on God.

The second scholastic argument against occasionalism appealed to the purported fact that it is evident to the senses that creatures have causal power. For Malebranche, however, this argument is no more persuasive than the argument that bodies must have colors and tastes since our senses tell us that they do. As indicated above, Malebranche offered Cartesian grounds for thinking that the purpose of our sensations is not to reveal the true nature of the material world, but rather to indicate what is helpful or harmful to our body. Malebranche held that our attribution of causal powers to bodies manifests in particular an attachment to the body that is an effect of original sin. Due to this attachment, we take objects in the material world to be a cause of our happiness rather than God.

In "Eclaircissement XV," Malebranche responded to the scholastic point that occasionalism renders scientific explanation impossible by appealing to the fact that God is not an arbitrary agent, but acts in accord with His wisdom. This wisdom dictates that He act "almost always" by means of a "general and efficacious will." Such a will produces effects that are perfectly law-like. For instance, God acts by a general will in producing changes in bodies in accord with the law of the communication of motion. Malebranche did allow that God can produce miracles by "particular volitions" that are not law-like. However, he emphasized that there are relatively few such volitions in God. Thus, we can offer scientific explanations that appeal to the laws of motion that reflect the nature of God's general will.

Malebranche was not the first Cartesian to endorse occasionalism. There were followers of Descartes such as Louis de la Forge (1632-1666) and Claude Clerselier (1614-1684) who stressed that God must be the cause of the communication of motion in bodily collisions given the passivity of Cartesian matter. These Cartesians attempted to preserve some room for the action of finite minds on body, but the Cartesian Geraud de Cordemoy (1626-1684) went further in claiming that only God can cause changes in the material world. However, none of these thinkers went as far as Malebranche in asserting that God must produce all real changes in nature. Moreover, Malebranche is distinctive in providing an explanation of God's action that distinguishes His general will from His particular volitions.

5. Theodicy

The presence of various evils in the world is problematic for anyone who claims that this world was created by a God who has infinite power, knowledge and goodness. However, the problem is particularly acute for an occasionalist such as Malebranche who holds that God is the only true cause of effects in nature. Malebranche offered a theodicy that addresses the problem of evil by emphasizing that in the "order of nature" God acts for the most part through His general will. In *Nature et grâce*, he started by admitting that God could have acted by particular volitions to prevent natural evils such as malformed offspring (a fitting example given his own malformed spine), and thus could have produced a more perfect world than He actually did create. However, he urged that God could have done so only by departing from simple laws, thereby sacrificing the simplicity and uniformity of action that is a supreme mark of His wisdom. God produces the natural evils that follow from simple laws not because He wills those particular effects, but because He wills a world that best reflects His wisdom by possessing the most effects governed by the fewest laws.

In his *Réflexions* on Malebranche's *Nature et grâce,* Arnauld objected to what he took to be the suggestion in his target text that God has concern only for general features of the world and does not will the details of His effects. For Arnauld, divine providence requires that God intend all of the particularities of the world He creates. There is some controversy over whether Arnauld's critique is based on a proper interpretation of Malebranche. Certain commentators follow Arnauld in thinking that Malebranche's claim in *Nature et grâce* that God acts by relatively few general volitions involves a rejection of the position that He has volitions for each particular effect. Others have insisted that this claim says only that God has volitions in accord with general laws, and that the doctrine of God's continual creation in the *Entretiens* in fact requires distinct volitions for distinct effects. Some evidence for the former view is provided by the fact that Malebranche emphasized that the laws themselves are "efficacious," and that God employs relatively few volitions in producing effects in the order of nature.

Malebranche insisted that God's general will is operative not only in the order of nature, but also in the "order of grace." However, he noted that the production of effects in the latter order also involves human action that is free in the strong sense of not being determined by anything external to the agent. His appeal to this sort of freedom is in fact central to his solution to the problem of moral evil, that is, the compatibility of sin with God's goodness. According to Malebranche, God is not responsible for sinful action since such action derives not from Him but from sinful agents. Arnauld objected that this solution is "more pelagian than anything in Pelagius," and that one must side with Augustine, who declared Pelagianism a heresy. Malebranche responded that he did not follow Pelagius in denying the importance of grace, and that Augustine himself had emphasized our freedom in action.

Malebranche also held that it is obvious by "inner sensation" that we are genuinely free. However, there is some question whether this introspective report is compatible with Malebranche's occasionalist claim that God is the only real cause. Malebranche did hold that God alone is the cause of our inclination to love "the good in general." However, he insisted that we are free to "consent" to the stopping of that inclination at a particular object other than God. Such consent results in an "absolute and intrinsic" love

of that object that is sinful given that this love is worthy only of God. The consent is free because one is always able to suspend consent and to search for objects more worthy of our love. Malebranche claimed that our freedom to consent or suspend consent does not conflict with occasionalism since these acts produce no “real” or “physical” change in our minds. The real question for Malebranche is whether it is intelligible to say that we can act freely without bringing about some real change in ourselves.

6. Moral Theory

The theocentrism that is evident in Malebranche's doctrines of the vision in God and occasionalism would lead us to expect that God plays a central role in his moral theory. This expectation is borne out by his discussion in the *Traité de morale*. Indeed, Malebranche's two doctrines are present in that work. The vision in God is reflected in the insistence there that moral duties are dictated by “relations of perfection” revealed in God's wisdom. As in the case of necessary truths concerning body, so in the case of moral truths Malebranche unequivocally rejected Cartesian voluntarism. The doctrine of occasionalism is reflected in Malebranche's insistence that God is our greatest good since He alone can cause our happiness. This point indicates that Malebranche took moral action to require a consideration not only of abstract relations of perfection, but also the happiness of the self.

Malebranche starts from the Augustinian position that morality concerns the proper ordering of our love. Given the importance of human freedom for his theodicy, it is not surprising that Malebranche insisted that the love required for moral action involve the free exercise of the will. His version of the “good will” is one that freely strives to be guided in action by objective relations of perfection that hold among the various objects of love. God is the most perfect being, and hence the most worthy of our love, while human beings are more perfect than mere material beings, and thus more worthy of our love. When the intensity of our love matches the order among perfections, we have a right love that provides the basis for virtue, that is, a habitual inclination to love objects according to their perfections.

Malebranche held that due to original sin, we are inclined not to right love directed by our perception of relations of perfection in God's wisdom, but rather to a disordered love directed by bodily pleasures deriving from the soul-body union. This is the counterpart to the disordered inclination of our will to make judgments about the nature of the material world that are based on sensations deriving from the union. For Malebranche, a corrective to both of these disorders of the will is to attend to clear ideas that exist in God.

Malebranche sometimes suggested that disordered love of bodily pleasure derives from self-love. Encouraged by this suggestion, one of his followers, François Lamy, claimed that his position leads to the quietist view in Fénélon that moral conduct requires a “pure love of God” that involves no concern for the self or its pleasure. This position, which Lamy himself endorsed, was later condemned by the Catholic Church, due in large part to a campaign against Fénélon directed by his critic, Bossuet. But Malebranche insisted that such a position directly conflicts with his own view that pleasure itself is a good that is required as a motive for action. When critics such as Arnauld and Régis charged that this view results in hedonism, Malebranche responded that it is only ordered pleasures that bring the greatest

good. This response is reflected in his claim to Lamy that a disordered love of self is to be contrasted not with pure love of God, but rather with an ordered love that seeks happiness in the contemplation of the greatest good, God. In emphasizing the need for this sort of love of God, Malebranche was returning to his view in the preface to the *Recherche* that it is through a union with God that the mind "receives its life, its light, and its entire felicity."

Bibliography

Malebranche's Works

- Malebranche, N. (1958-84). *Œuvres complètes de Malebranche* (20 vols.) (A. Robinet, Ed.). Paris: J. Vrin.
- Malebranche, N. (1980a). *Dialogues on Metaphysics*. (W. Doney, Trans.). New York: Abaris Books.
- Malebranche, N. (1980b). *Dialogue between a Christian Philosopher and a Chinese Philosopher on the Existence and Nature of God*. (D. A. Iorio, Trans.). Washington, DC: Catholic University Press.
- Malebranche, N. (1993). *Treatise on Ethics*. (C. Walton, Trans.). Dordrecht: Kluwer.
- Malebranche, N. (1997a). *Dialogues on Metaphysics and on Religion*. (N. Jolley and D. Scott, Trans.). Cambridge: Cambridge University Press.
- Malebranche, N. (1997b). *The Search after Truth*. (T. M. Lennon and P. J. Olscamp, Trans.). Columbus: Ohio State University Press, 1980; Cambridge: Cambridge University Press.

Related Early Modern Works

- André, Y. M. (1970). *La vie du R. P. Malebranche*. Paris: Ingold, 1886; Geneva: Slatkin Reprints.
- Arnauld, A. (1990). *On True and False Ideas*. (E. J. Kremer, trans.). Lewiston: The Edwin Mellen Press.
- Foucher, S. *Critique de la recherche de la verité*. (R. A. Watson, Ed.) New York: Johnson Reprint.

Bibliographical Sources

- Easton, P., Lennon, T. M., and Sebba, G. (1992). *Bibliographia Malebranchiana: A Critical Guide to the Malebranche Literature into 1989*. Carbondale and Edwardsville: Southern Illinois Press.
- Sebba, G. (1959). *Nicolas Malebranche, 1638-1715: A Preliminary Bibliography*. Athens: University of Georgia Press.

Recommended Secondary Literature

- Alquié, F. (1974). *Le cartésianisme de Malebranche*. Paris: J. Vrin.
- Bardout, J.-C. (1999). *Malebranche et la métaphysique*. Paris: Presses Universitaires de France.
- Brown, S. (1991). *Nicolas Malebranche: His Philosophical Critics and Successors*. Assen: Van Gorcum.
- Chappell, V. (1992) (Ed.). *Nicholas Malebranche*. New York: Garland.
- Connell, D. (1967). *The Vision in God: Malebranche's Scholastic Sources*. Louvain: Nauwelaerts.
- Gueroult, M. (1955-59). *Malebranche*(3 vols.). Paris: Aubier.
- Jolley, N. (1990). *The Light of the Soul: Theories of Ideas in Leibniz, Malebranche, and Descartes* . Oxford: Clarendon Press.
- McCracken, C. (1983). *Malebranche and British Philosophy*. Oxford: Clarendon Press.
- Moreau, D. (1999). *Deux cartésiens: La polemique entre Antoine Arnauld et Nicolas Malebranche*. Paris: J. Vrin.
- Nadler, S. (1992). *Malebranche and Ideas*. New York: Oxford University Press.
- Nadler, S. (2000) (Ed.). *The Cambridge Companion to Malebranche*. Cambridge: Cambridge University Press.
- Radner, D. (1978). *Malebranche: A Study of a Cartesian System*. Assen: Van Gorcum.
- Robinet, A. (1965). *Système et existence dans l'œuvre de Malebranche*. Paris: J. Vrin.
- Rodis-Lewis, G. (1963). *Nicolas Malebranche*. Paris: Presses Universitaires de France.
- Schmaltz, T. M. (1996). *Malebranche's Theory of the Soul: A Cartesian Interpretation*. New York: Oxford University Press.
- Walton, C. (1972). *De la Recherche du Bien: A Study of Malebranche's Science of Ethics*. The Hague: Martinus Nijhoff.

Other Internet Resources

- Great Voyages
- Galileo Project
- Internet Encyclopedia of Philosophy
- Catholic Encyclopedia
- The Cartesian School

[Please contact the author with further suggestions]

Related Entries

Descartes, René | Desgabets, Robert | Leibniz, Gottfried Wilhelm | Spinoza, Baruch [Benedict]

Tad M. Schmaltz
tad.schmaltz@duke.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 24, 2002 Content last modified: May 24, 2002

Stanford Encyclopedia of Philosophy

Baruch Spinoza

Baruch (or Benedictus) Spinoza is one of the most important philosophers -- and certainly the most radical -- of the early modern period. His thought combines a commitment to Cartesian metaphysical and epistemological principles with elements from ancient Stoicism and medieval Jewish rationalism into a nonetheless highly original system. His extremely naturalistic views on God, the world, the human being and knowledge serve to ground a moral philosophy centered on the control of the passions leading to virtue and happiness. They also lay the foundations for a strongly democratic political thought and a deep critique of the pretensions of Scripture and sectarian religion. Of all the philosophers of the seventeenth-century, perhaps none have more relevance today than Spinoza.

1. Biography

Baruch Spinoza was born in 1632 in Amsterdam. He was the middle son in a prominent family of moderate means in Amsterdam's Portuguese-Jewish community. As a boy, he had undoubtedly been one of the star pupils in the congregation's Talmud Torah school. He was intellectually gifted, and this could not have gone unremarked by the congregation's rabbis. It is possible that Spinoza, as he made progress through his studies, was being groomed for a career as a rabbi. But he never made it into the upper levels

of the curriculum, those which included advanced study of Talmud. At the age of seventeen, he was forced to cut short his formal studies to help run the family's importing business.

And then, on July 27, 1656, Spinoza was issued the harshest writ of *cherem*, or excommunication, ever pronounced by the Sephardic community of Amsterdam; it was never rescinded. We do not know for certain what Spinoza's "monstrous deeds" and "abominable heresies" were alleged to have been, but an educated guess comes quite easy. No doubt he was giving utterance to just those ideas that would soon appear in his philosophical treatises. In those works, Spinoza denies the immortality of the soul; strongly rejects the notion of a providential God -- the God of Abraham, Isaac and Jacob; and claims that the Law was neither literally given by God nor any longer binding on Jews. Can there be any mystery as to why one of history's boldest and most radical thinkers was sanctioned by an orthodox Jewish community?

To all appearances, Spinoza was content finally to have an excuse for departing from the community and leaving Judaism behind; his faith and religious commitment were, by this point, gone. Within a few years, he left Amsterdam altogether. By the time his extant correspondence begins, in 1661, he is living in Rijnsburg, not far from Leiden. While in Rijnsburg, he worked on the *Treatise on the Emendation of the Intellect*, an essay on philosophical method, and the *Short Treatise on God, Man and His Well-Being*, an initial but aborted effort to lay out his metaphysical, epistemological and moral views. His critical exposition of Descartes's *Principles of Philosophy*, the only work he published under his own name in his lifetime, was completed in 1663, after he had moved to Voorburg, outside The Hague. By this time, he was also working on what would eventually be called the *Ethics*, his philosophical masterpiece. However, when he saw the principles of toleration in Holland being threatened by reactionary forces, he put it aside to complete his "scandalous" *Theological-Political Treatise*, published anonymously and to great alarm in 1670. When Spinoza died in 1677, in The Hague, he was still at work on his *Political Treatise*; this was soon published by his friends along with his other unpublished writings, including a *Compendium to Hebrew Grammar*.

2. *Ethics*

The *Ethics* is an ambitious and multifaceted work. It is also bold to the point of audacity, as one would expect of a systematic and unforgiving critique of the traditional philosophical conceptions of God, the human being and the universe, and, above all, of the religions and the theological and moral beliefs grounded thereupon. What Spinoza intends to demonstrate (in the strongest sense of that word) is the truth about God, nature and especially ourselves; and the highest principles of society, religion and the good life. Despite the great deal of metaphysics, physics, anthropology and psychology that take up Parts One through Three, Spinoza took the crucial message of the work to be ethical in nature. It consists in showing that our happiness and well-being lie not in a life enslaved to the passions and to the transitory goods we ordinarily pursue; nor in the related unreflective attachment to the superstitions that pass as religion, but rather in the life of reason. To clarify and support these broadly ethical conclusions, however, Spinoza must first demystify the universe and show it for what it really is. This requires laying out some metaphysical foundations, the project of Part One.

God or Nature

"On God" begins with some deceptively simple definitions of terms that would be familiar to any seventeenth century philosopher. "By substance I understand what is in itself and is conceived through itself"; "By attribute I understand what the intellect perceives of a substance, as constituting its essence"; "By God I understand a being absolutely infinite, i.e., a substance consisting of an infinity of attributes, of which each one expresses an eternal and infinite essence." The definitions of Part One are, in effect, simply clear concepts that ground the rest of his system. They are followed by a number of axioms that, he assumes, will be regarded as obvious and unproblematic by the philosophically informed ("Whatever is, is either in itself or in another"; "From a given determinate cause the effect follows necessarily"). From these, the first proposition necessarily follows, and every subsequent proposition can be demonstrated using only what precedes it. (References to the *Ethics* will be by part (I-V), proposition (p), definition (d), scholium (s) and corollary (c).)

In propositions one through fifteen of Part One, Spinoza presents the basic elements of his picture of God. God is the infinite, necessarily existing (that is, uncaused), unique substance of the universe. There is only one substance in the universe; it is God; and everything else that is, is in God.

> Proposition 1: A substance is prior in nature to its affections.
>
> Proposition 2: Two substances having different attributes have nothing in common with one another. (In other words, if two substances differ in nature, then they have nothing in common).
>
> Proposition 3: If things have nothing in common with one another, one of them cannot be the cause of the other.
>
> Proposition 4: Two or more distinct things are distinguished from one another, either by a difference in the attributes [i.e., the natures or essences] of the substances or by a difference in their affections [i.e., their accidental properties].
>
> Proposition 5: In nature, there cannot be two or more substances of the same nature or attribute.
>
> Proposition 6: One substance cannot be produced by another substance.
>
> Proposition 7: It pertains to the nature of a substance to exist.
>
> Proposition 8: Every substance is necessarily infinite.
>
> Proposition 9: The more reality or being each thing has, the more attributes belong to it.

> Proposition 10: Each attribute of a substance must be conceived through itself.
>
> Proposition 11: God, or a substance consisting of infinite attributes, each of which expresses eternal and infinite essence, necessarily exists. (The proof of this proposition consists simply in the classic "ontological proof for God's existence". Spinoza writes that "if you deny this, conceive, if you can, that God does not exist. Therefore, by axiom 7 [‘If a thing can be conceived as not existing, its essence does not involve existence’], his essence does not involve existence. But this, by proposition 7, is absurd. Therefore, God necessarily exists, q.e.d.")
>
> Proposition 12: No attribute of a substance can be truly conceived from which it follows that the substance can be divided.
>
> Proposition 13: A substance which is absolutely infinite is indivisible.
>
> Proposition 14: Except God, no substance can be or be conceived.

This proof that God -- an infinite, necessary and uncaused, indivisible being -- is the only substance of the universe proceeds in three simple steps. First, establish that no two substances can share an attribute or essence (Ip5). Then, prove that there is a substance with infinite attributes (i.e., God) (Ip11). It follows, in conclusion, that the existence of that infinite substance precludes the existence of any other substance. For if there *were* to be a second substance, it would have to have *some* attribute or essence. But since God has *all* possible attributes, then the attribute to be possessed by this second substance would be one of the attributes already possessed by God. But it has already been established that no two substances can have the same attribute. Therefore, there can be, besides God, no such second substance.

If God is the only substance, and (by axiom 1) whatever is, is either a substance or *in* a substance, then everything else must be in God. "Whatever is, is in God, and nothing can be or be conceived without God" (Ip15).

As soon as this preliminary conclusion has been established, Spinoza immediately reveals the objective of his attack. His definition of God -- condemned since his excommunication from the Jewish community as a "God existing in only a philosophical sense" -- is meant to preclude any anthropomorphizing of the divine being. In the scholium to proposition fifteen, he writes against "those who feign a God, like man, consisting of a body and a mind, and subject to passions. But how far they wander from the true knowledge of God, is sufficiently established by what has already been demonstrated." Besides being false, such an anthropomorphic conception of God can have only deleterious effects on human freedom and activity.

Much of the technical language of Part One is, to all appearances, right out of Descartes. But even the most devoted Cartesian would have had a hard time understanding the full import of propositions one through fifteen. What does it mean to say that God is substance and that everything else is "in" God? Is

Spinoza saying that rocks, tables, chairs, birds, mountains, rivers and human beings are all *properties* of God, and hence can be predicated of God (just as one would say that the table "is red")? It seems very odd to think that objects and individuals -- what we ordinarily think of as independent "things" -- are, in fact, merely properties of a thing. Spinoza was sensitive to the strangeness of this kind of talk, not to mention the philosophical problems to which it gives rise. When a person feels pain, does it follow that the pain is ultimately just a *property* of God, and thus that God feels pain? Conundrums such as this may explain why, as of Proposition Sixteen, there is a subtle but important shift in Spinoza's language. God is now described not so much as the underlying substance of all things, but as the universal, immanent and sustaining cause of all that exists: "From the necessity of the divine nature there must follow infinitely many things in infinitely many modes, (i.e., everything that can fall under an infinite intellect)".

According to the traditional Judeo-Christian conception of divinity, God is a transcendent creator, a being who causes a world distinct from himself to come into being by creating it out of nothing. God produces that world by a spontaneous act of free will, and could just as easily have not created anything outside himself. By contrast, Spinoza's God is the cause of all things because all things follow causally and necessarily from the divine nature. Or, as he puts it, from God's infinite power or nature "all things have necessarily flowed, or always followed, by the same necessity and in the same way as from the nature of a triangle it follows, from eternity and to eternity, that its three angles are equal to two right angles" (Ip17s1). The existence of the world is, thus, mathematically necessary. It is impossible that God should exist but not the world. This does not mean that God does not cause the world to come into being freely, since nothing *outside* of God constrains him to bring it into existence. But Spinoza does deny that God creates the world by some arbitrary and undetermined act of free will. God could not have done otherwise. There are no possible alternatives to the actual world, and absolutely no contingency or spontaneity within that world. Everything is absolutely and necessarily determined.

> (Ip29): In nature there is nothing contingent, but all things have been determined from the necessity of the divine nature to exist and produce an effect in a certain way.
>
> (Ip33): Things could have been produced by God in no other way, and in no other order than they have been produced.

There are, however, differences in the way things depend on God. Some features of the universe follow necessarily from God -- or, more precisely, from the absolute nature of one of God's attributes -- in a direct and unmediated manner. These are the universal and eternal aspects of the world, and they do not come into or go out of being. They include the most general laws of the universe, together governing all things in all ways. From the attribute of extension there follow the principles governing all extended objects (the truths of geometry) and laws governing the motion and rest of bodies (the laws of physics); from the attribute of thought, there follow laws of thought (understood by commentators to be either the laws of logic or the laws of psychology). Particular and individual things are causally more remote from God. They are nothing but "affections of God's attributes, or modes by which God's attributes are expressed in a certain and determinate way" (Ip25c).

There are two causal orders or dimensions governing the production and actions of particular things. On

the one hand, they are determined by the general laws of the universe that follow immediately from God's natures. On the other hand, each particular thing is determined to act and to be acted upon by other particular things. Thus, the actual behavior of a body in motion is a function not just of the universal laws of motion, but also of the other bodies in motion and rest surrounding it and with which it comes into contact.

Spinoza's metaphysics of God is neatly summed up in a phrase that occurs in the Latin (but not the Dutch) edition of the *Ethics*: "God, or Nature", *Deus, sive Natura*: "That eternal and infinite being we call God, or Nature, acts from the same necessity from which he exists" (Part IV, Preface). It is an ambiguous phrase, since Spinoza could be read as trying either to divinize nature or to naturalize God. But for the careful reader there is no mistaking Spinoza's intention. The friends who, after his death, published his writings must have left out the "or Nature" clause from the more widely accessible Dutch version out of fear of the reaction that this identification would, predictably, arouse among a vernacular audience.

There are, Spinoza insists, two sides of Nature. First, there is the active, productive aspect of the universe -- God and his attributes, from which all else follows. This is what Spinoza, employing the same terms he used in the *Short Treatise*, calls *Natura naturans*, "naturing Nature". Strictly speaking, this is identical with God. The other aspect of the universe is that which is produced and sustained by the active aspect, *Natura naturata*, "natured Nature".

> By *Natura naturata* I understand whatever follows from the necessity of God's nature, or from any of God's attributes, i.e., all the modes of God's attributes insofar as they are considered as things that are in God, and can neither be nor be conceived without God. (Ip29s).

Spinoza's fundamental insight in Book One is that Nature is an indivisible, uncaused, substantial whole -- in fact, it is the *only* substantial whole. Outside of Nature, there is nothing, and everything that exists is a part of Nature and is brought into being by Nature with a deterministic necessity. This unified, unique, productive, necessary being just *is* what is meant by ‘God’. Because of the necessity inherent in Nature, there is no teleology in the universe. Nature does not act for any ends, and things do not exist for any set purposes. There are no "final causes" (to use the common Aristotelian phrase). God does not "do" things for the sake of anything else. The order of things just follows from God's essences with an inviolable determinism. All talk of God's purposes, intentions, goals, preferences or aims is just an anthropomorphizing fiction.

> All the prejudices I here undertake to expose depend on this one: that men commonly suppose that all natural things act, as men do, on account of an end; indeed, they maintain as certain that God himself directs all things to some certain end, for they say that God has made all things for man, and man that he might worship God. (I, Appendix)

God is not some goal-oriented planner who then judges things by how well they conform to his purposes. Things happen only because of Nature and its laws. "Nature has no end set before it . . . All things proceed by a certain eternal necessity of nature." To believe otherwise is to fall prey to the same

superstitions that lie at the heart of the organized religions.

> [People] find -- both in themselves and outside themselves -- many means that are very helpful in seeking their own advantage, e.g., eyes for seeing, teeth for chewing, plants and animals for food, the sun for light, the sea for supporting fish . . . Hence, they consider all natural things as means to their own advantage. And knowing that they had found these means, not provided them for themselves, they had reason to believe that there was someone else who had prepared those means for their use. For after they considered things as means, they could not believe that the things had made themselves; but from the means they were accustomed to prepare for themselves, they had to infer that there was a ruler, or a number of rulers of nature, endowed with human freedom, who had taken care of all things for them, and made all things for their use.
>
> And since they had never heard anything about the temperament of these rulers, they had to judge it from their own. Hence, they maintained that the Gods direct all things for the use of men in order to bind men to them and be held by men in the highest honor. So it has happened that each of them has thought up from his own temperament different ways of worshipping God, so that God might love them above all the rest, and direct the whole of Nature according to the needs of their blind desire and insatiable greed. Thus this prejudice was changed into superstition, and struck deep roots in their minds. (I, Appendix)

A judging God who has plans and acts purposively is a God to be obeyed and placated. Opportunistic preachers are then able to play on our hopes and fears in the face of such a God. They prescribe ways of acting that are calculated to avoid being punished by that God and earn his rewards. But, Spinoza insists, to see God or Nature as acting for the sake of ends -- to find purpose in Nature -- is to misconstrue Nature and "turn it upside down" by putting the effect (the end result) before the true cause.

Nor does God perform miracles, since there are no departures whatsoever from the necessary course of nature. The belief in miracles is due only to ignorance of the true causes of phenomena.

> If a stone has fallen from a room onto someone's head and killed him, they will show, in the following way, that the stone fell in order to kill the man. For if it did not fall to that end, God willing it, how could so many circumstances have concurred by chance (for often many circumstances do concur at once)? Perhaps you will answer that it happened because the wind was blowing hard and the man was walking that way. But they will persist: why was the wind blowing hard at that time? why was the man walking that way at that time? If you answer again that the wind arose then because on the preceding day, while the weather was still calm, the sea began to toss, and that the man had been invited by a friend, they will press on -- for there is no end to the questions which can be asked: but why was the sea tossing? why was the man invited at just that time? And so they will not stop asking for the causes of causes until you take refuge in the will of God, i.e., the sanctuary of ignorance. (I, Appendix)

This is strong language, and Spinoza is clearly not unaware of the risks of his position. The same preachers who take advantage of our credulity will fulminate against anyone who tries to pull aside the curtain and reveal the truths of Nature. "One who seeks the true causes of miracles, and is eager, like an educated man, to understand natural things, not to wonder at them, like a fool, is generally considered and denounced as an impious heretic by whose whom the people honor as interpreters of nature and the Gods. For they know that if ignorance is taken away, then foolish wonder, the only means they have of arguing and defending their authority is also taken away."

The Human Being

In Part Two, Spinoza turns to the origin and nature of the human being. The two attributes of God of which we have cognizance are extension and thought. This, in itself, involves what would have been an astounding thesis in the eyes of his contemporaries, one that was usually misunderstood and always vilified. When Spinoza claims in Proposition Two that "Extension is an attribute of God, or God is an extended thing", he was almost universally -- but erroneously -- interpreted as saying that God is literally corporeal. For just this reason, "Spinozism" became, for his critics, synonymous with atheistic materialism.

According to one interpretation, God is indeed material, even matter itself, but this does not imply that God has a body. Another interpretation, however, one which will be adopted here, is that what is in God is not matter per se, but extension as an essence. And extension and thought are two distinct essences that have absolutely nothing in common. The modes or expressions of extension are physical bodies; the modes of thought are ideas. Because extension and thought have nothing in common, the two realms of matter and mind are causally closed systems. Everything that is extended follows from the attribute of extension alone. Every bodily event is part of an infinite causal series of bodily events and is determined only by the nature of extension and its laws, in conjunction with its relations to other extended bodies. Similarly, every idea follows only from the attribute of thought. Any idea is an integral part of an infinite series of ideas and is determined by the nature of thought and its laws, along with its relations to other ideas. There is, in other words, no causal interaction between bodies and ideas, between the physical and the mental. There is, however, a thoroughgoing correlation and parallelism between the two series. For every mode in extension that is a relatively stable collection of matter, there is a corresponding mode in thought. In fact, he insists, "a mode of extension and the idea of that mode are one and the same thing, but expressed in two ways". Because of the fundamental and underlying unity of Nature, or of Substance, Thought and Extension are just two different ways of "comprehending" one and the same Nature. Every material thing thus has its own particular idea -- a kind of Platonic concept -- that expresses or represents it. Since that idea is just a mode of one of God's attributes -- Thought -- it is in God, and the infinite series of ideas constitutes God's mind. As he explains,

> A circle existing in nature and the idea of the existing circle, which is also in God, are one and the same thing, which is explained through different attributes. Therefore, whether we conceive nature under the attribute of Extension, or under the attribute of Thought, or under any other attribute, we shall find one and the same order, or one and the same connection of causes, i.e., that the same things follow one another.

It follows from this, he argues, that the causal relations between bodies is mirrored in the logical relations between God's ideas. Or, as Spinoza notes in Proposition Seven, "the order and connection of ideas is the same as the order and connection of things".

One kind of extended body, however, is significantly more complex than any others in its composition and in its dispositions to act and be acted upon. That complexity is reflected in its corresponding idea. The body in question is the human body; and its corresponding idea is the human mind or soul. The mind, then, like any other idea, is simply one particular mode of God's attribute, Thought. Whatever happens in the body is reflected or expressed in the mind. In this way, the mind perceives, more or less obscurely, what is taking place in its body. And through its body's interactions with other bodies, the mind is aware of what is happening in the physical world around it. But the human mind no more interacts with its body than any mode of Thought interacts with a mode of Extension.

One of the pressing questions in seventeenth century philosophy, and perhaps the most celebrated legacy of Descartes's dualism, is the problem of how two radically different substances such as mind and body enter into a union in a human being and cause effects in each other. How can the extended body causally engage the unextended mind, which is incapable of contact or motion, and "move" it, that is, cause mental effects such as pains, sensations and perceptions. Spinoza, in effect, denies that the human being is a union of two *substances*. The human mind and the human body are two different expressions -- under Thought and under Extension -- of one and the same thing: the person. And because there is no causal interaction between the mind and the body, the so-called mind-body problem does not, technically speaking, arise.

Knowledge

The human mind, like God, contains ideas. Some of these ideas -- sensory images, qualitative "feels" (like pains and pleasures), perceptual data -- are imprecise qualitative phenomena, being the expression in thought of states of the body as it is affected by the bodies surrounding it. Such ideas do not convey adequate and true knowledge of the world, but only a relative, partial and subjective picture of how things presently seem to be to the perceiver. There is no systematic order to these perceptions, nor any critical oversight by reason. "As long as the human Mind perceives things from the common order of nature, it does not have an adequate, but only a confused and mutilated knowledge of itself, of its own Body, and of external bodies" (Iip29c). Under such circumstances, we are simply determined in our ideas by our fortuitous and haphazard encounter with things in the external world. This superficial acquaintance will never provide us with knowledge of the essences of those things. In fact, it is an invariable source of falsehood and error. This "knowledge from random experience" is also the origin of great delusions, since we -- thinking ourselves free -- are, in our ignorance, unaware of just how we *are* determined by causes.

Adequate ideas, on the other hand, are formed in a rational and orderly manner, and are necessarily true and revelatory of the essences of things. "Reason", the second kind of knowledge (after "random experience"), is the apprehension of the essence of a thing through a discursive, inferential procedure. "A

true idea means nothing other than knowing a thing perfectly, or in the best way." It involves grasping a thing's causal connections not just to other objects but, more importantly, to the attributes of God and the infinite modes (the laws of nature) that follow immediately from them. The adequate idea of a thing clearly and distinctly situates its object in all of its causal nexuses and shows not just *that* it is, but *how* and *why* it is. The person who truly knows a thing sees the reasons why the thing was determined to be and could not have been otherwise. "It is of the nature of Reason to regard things as necessary, not as contingent" (IIp44). The belief that some thing is accidental or spontaneous can be based only on an inadequate grasp of the thing's causal explanation, on a partial and "mutilated" familiarity with it. To perceive by way of adequate ideas is to perceive the necessity inherent in Nature.

Sense experience alone could never provide the information conveyed by an adequate idea. The senses present things only as they appear from a given perspective at a given moment in time. An adequate idea, on the other hand, by showing how a thing follows necessarily from one or another of God's attributes, presents it in its "eternal" aspects -- *sub specie aeternitatis*, as Spinoza puts it -- without any relation to time. "It is of the nature of Reason to regard things as necessary and not as contingent. And Reason perceives this necessity of things truly, i.e., as it is in itself. But this necessity of things is the very necessity of God's eternal nature. Therefore, it is of the nature of Reason to regard things under this species of eternity". The third kind of knowledge, intuition, takes what is known by Reason and grasps it in a single act of the mind.

Spinoza's conception of adequate knowledge reveals an unrivaled optimism in the cognitive powers of the human being. Not even Descartes believed that we could know all of Nature and its innermost secrets with the degree of depth and certainty that Spinoza thought possible. Most remarkably, because Spinoza thought that the adequate knowledge of any object, and of Nature as a whole, involves a thorough knowledge of God and of how things related to God and his attributes, he also had no scruples about claiming that we can, at least in principle, know God perfectly and adequately. "The knowledge of God's eternal and infinite essence that each idea involves is adequate and perfect" (IIp46). "The human Mind has an adequate knowledge of God's eternal and infinite essence" (Iip47). No other philosopher in history has been willing to make this claim. But, then again, no other philosopher identified God with Nature.

Passion and Action

Spinoza engages in such a detailed analysis of the composition of the human being because it is essential to his goal of showing how the human being is a part of Nature, existing within the same causal nexuses as other extended and mental beings. This has serious ethical implications. First, it implies that a human being is not endowed with freedom, at least in the ordinary sense of that term. Because our minds and the events in our minds are simply ideas that exist within the causal series of ideas that follows from God's attribute Thought, our actions and volitions are as necessarily determined as any other natural events. "In the Mind there is no absolute, or free, will, but the Mind is determined to will this or that by a cause that is also determined by another, and this again by another, and so to infinity."

What is true of the will (and, of course, of our bodies) is true of all the phenomena of our psychological

lives. Spinoza believes that this is something that has not been sufficiently understood by previous thinkers, who seem to have wanted to place the human being on a pedestal outside of (or above) nature.

> Most of those who have written about the Affects, and men's way of living, seem to treat, not of natural things, which follow the common laws of nature, but of things that are outside nature. Indeed they seem to conceive man in nature as a dominion within a dominion. For they believe that man disturbs, rather than follows, the order of nature, that he has absolute power over his actions, and that he is determined only by himself. (III, Preface)

Descartes, for example, believed that if the freedom of the human being is to be preserved, the soul must be exempt from the kind of deterministic laws that rule over the material universe.

Spinoza's aim in Parts Three and Four is, as he says in his Preface to Part Three, to restore the human being and his volitional and emotional life into their proper place in nature. For nothing stands outside of nature, not even the human mind.

> Nature is always the same, and its virtue and power of acting are everywhere one and the same, i.e., the laws and rules of nature, according to which all things happen, and change from one form to another, are always and everywhere the same. So the way of understanding the nature of anything, of whatever kind, must also be the same, viz. through the universal laws and rules of nature.

Our affects -- our love, anger, hate, envy, pride, jealousy, etc. -- "follow from the same necessity and force of nature as the other singular things". Spinoza, therefore, explains these emotions -- as determined in their occurrence as are a body in motion and the properties of a mathematical figure -- just as he would explain any other things in nature. "I shall treat the nature and power of the Affects, and the power of the Mind over them, by the same Method by which, in the preceding parts, I treated God and the Mind, and I shall consider human actions and appetites just as if it were a Question of lines, planes, and bodies."

Our affects are divided into actions and passions. When the cause of an event lies in our own nature -- more particularly, our knowledge or adequate ideas -- then it is a case of the mind acting. On the other hand, when something happens in us the cause of which lies outside of our nature, then we are passive and being acted upon. Usually what takes place, both when we are acting and when we are being acted upon, is some change in our mental or physical capacities, what Spinoza calls "an increase or decrease in our power of acting" or in our "power to persevere in being". All beings are naturally endowed with such a power or striving. This *conatus*, a kind of existential inertia, constitutes the "essence" of any being. "Each thing, as far as it can by its own power, strives to persevere in its being." An affect just *is* any change in this power, for better or for worse. Affects that are actions are changes in this power that have their source (or "adequate cause") in our nature alone; affects that are passions are those changes in this power that originate outside of us.

What we should strive for is to be free from the passions -- or, since this is not absolutely possible, at least

to learn how to moderate and restrain them -- and become active, autonomous beings. If we can acheive this, then we will be "free" to the extent that whatever happens to us will result not from our relations with things outside us, but from our own nature (as that follows from, and is ultimately and necessarily determined by the attributes of God of which our minds and bodies are modes). We will, consequently, be truly liberated from the troublesome emotional ups and downs of this life. The way to bring this about is to increase our knowledge, our store of adequate ideas, and eliminate as far as possible our inadequate ideas, which follow not from the nature of the mind alone but from its being an expresssion of how our body is affected by other bodies. In other words, we need to free ourselves from a reliance on the senses and the imagination, since a life of the senses and images is a life being affected and led by the objects around us, and rely as much as we can only on our rational faculties.

Because of our innate striving to persevere -- which, in the human being, is called "will" or "appetite" -- we naturally pursue those things that we believe will benefit us by increasing our power of acting and shun or flee those things that we believe will harm us by decreasing our power of acting. This provides Spinoza with a foundation for cataloguing the human passions. For the passions are all functions of the ways in which external things affect our powers or capacities. Joy [*Laaetitiae*, sometimes translated as "pleasure"], for example, is simply the movement or passage to a greater capacity for action. "By Joy . . . I shall understand that passion by which the Mind passes to a greater perfection" (IIIp11s). Being a passion, joy is always brought about by some external object. Sadness [*Tristitiae*, or "pain"], on the other hand, is the passage to a lesser state of perfection, also occasioned by a thing outside us. Love is simply Joy accompanied by an awareness of the external cause that brings about the passage to a greater perfection. We love that object that benefits us and causes us joy. Hate is nothing but "Sadness with the accompanying idea of an external cause". Hope is simply "an inconstant Joy which has arisen from the image of a future or past thing whose outcome we doubt". We hope for a thing whose presence, as yet uncertain, will bring about joy. We fear, however, a thing whose presence, equally uncertain, will bring about sadness. When that whose outcome was doubtful becomes certain, hope is changed into confidence, while fear is changed into despair.

All of the human emotions, in so far as they are passions, are constantly directed outward, towards things and their capacities to affect us one way or another. Aroused by our passions and desires, we seek or flee those things that we believe cause joy or sadness. "We strive to further the occurrence of whatever we imagine will lead to Joy, and to avert or destroy what we imagine is contrary to it, or will lead to Sadness." Our hopes and fears fluctuate depending on whether we regard the objects of our desires or aversions as remote, near, necessary, possible or unlikely. But the objects of our passions, being external to us, are completely beyond our control. Thus, the more we allow ourselves to be controlled by *them*, the more we are subject to passions and the less active and free we are. The upshot is a fairly pathetic picture of a life mired in the passions and pursuing and fleeing the changeable and fleeting objects that occasion them: "We are driven about in many ways by external causes, and . . . like waves on the sea, driven by contrary winds, we toss about, not knowing our outcome and fate" (IIIp59s). The title for Part Four of the *Ethics* reveals with perfect clarity Spinoza's evaluation of such a life for a human being: "On Human Bondage, or the Powers of the Affects". He explains that the human being's "lack of power to moderate and restrain the affects I call Bondage. For the man who is subject to affects is under the control, not of himself, but of fortune, in whose power he so greatly is that often, though he sees the better for himself,

he is still forced to follow the worse". It is, he says, a kind of "sickness of the mind" to suffer too much love for a thing "that is liable to many variations and that we can never fully possess."

Virtue and Happiness

The solution to this predicament is an ancient one. Since we cannot control the objects that we tend to value and that we allow to influence our well-being, we ought instead to try to control our evaluations themselves and thereby minimize the sway that external objects and the passions have over us. We can never eliminate the passive affects entirely. We are essentially a part of nature, and can never fully remove outselves from the causal series that link us to external things. But we can, ultimately, counteract the passions, control them, and achieve a certain degree of relief from their turmoil.

The path to restraining and moderating the affects is through virtue. Spinoza is a psychological and ethical egoist. All beings naturally seek their own advantage -- to preserve their own being -- and it is right for them do so. This is what virtue consists in. Since we are thinking beings, endowed with intelligence and reason, what is to our greatest advantage is knowledge. Our virtue, therefore, consists in the pursuit of knowledge and understanding, of adequate ideas. The best kind of knowledge is a purely intellectual intuition of the essences of things. This "third kind of knowledge" -- beyond both random experience and ratiocination -- sees things not in their temporal dimension, not in their duration and in relation to other particular things, but under the aspect of eternity, that is, abstracted from all considerations of time and place and situated in their relationship to God and his attributes. They are apprehended, that is, in their conceptual and causal relationship to the universal essences (thought and extension) and the eternal laws of nature.

> We conceive things as actual in two ways: either insofar as we conceive them to exist in relation to a certain time and place, or insofar as we conceive them to be contained in God and to follow from the necessity of the divine nature. But the things we conceive in this second way as true, or real, we conceive under a species of eternity, and to that extent they involve the eternal and infinite essence of God. (Vp39s)

But this is just to say that, ultimately, we strive for a knowledge of God. The concept of any body involves the concept of extension; and the concept of any idea or mind involves the concept of thought. But thought and extension just are God's attributes. So the proper and adequate conception of any body or mind necessarily involves the concept or knowledge of God. "The third kind of knowledge proceeds from an adquate idea of certain attributes of God to an adequate knowledge of the essence of things, and the more we understand things in this way, the more we understand God." Knowledge of God is, thus, the Mind's greatest good and its greatest virtue.

What we see when we understand things through the third kind of knowledge, under the aspect of eternity and in relation to God, is the deterministic necessity of all things. We see that all bodies and their states follow necessarily from the essence of matter and the universal laws of physics; and we see that all ideas, including all the properties of minds, follow necessarily from the essence of thought and its universal

laws. This insight can only weaken the power that the passions have over us. We are no longer hopeful or fearful of what shall come to pass, and no longer anxious or despondent over our possessions. We regard all things with equanimity, and we are not inordinately and irrationally affected in different ways by past, present or future events. The result is self-control and a calmness of mind.

> The more this knowledge that things are necessary is concerned with singular things, which we imagine more distinctly and vividly, the greater is this power of the Mind over the affects, as experience itself also testifies. For we see that Sadness over some good which has perished is lessened as soon as the man who has lost it realizes that this good could not, in any way, have been kept. Similarly, we see that [because we regard infancy as a natural and necessary thing], no one pities infants because of their inability to speak, to walk, or to reason, or because they live so many years, as it were, unconscious of themselves. (Vp6s)

Our affects themselves can be understood in this way, which further diminishes their power over us.

Spinoza's ethical theory is, to a certain degree, Stoic, and recalls the doctrines of thinkers such as Cicero and Seneca:

> We do not have an absolute power to adapt things outside us to our use. Nevertheless, we shall bear calmly those things that happen to us contrary to what the principle of our advantage demands, if we are conscious that we have done our duty, that the power we have could not have extended itself to the point where we could have avoided those things, and that we are a part of the whole of nature, whose order we follow. If we understand this clearly and distinctly, that part of us which is defined by understanding, i.e., the better part of us, will be entirely satisfied with this, and will strive to persevere in that satisfaction. For insofar as we understand, we can want nothing except what is necessary, nor absolutely be satisfied with anything except what is true. (IV, Appendix)

What, in the end, replaces the passionate love for ephemeral "goods" is an intellectual love for an eternal, immutable good that we can fully and stably possess, God. The third kind of knowledge generates a love for its object, and in this love consists not joy, a passion, but blessedness itself. Taking his cue from Maimonides's view of human *eudaimonia*, Spinoza argues that the mind's intellectual love of God *is* our understanding of the universe, our virtue, our happiness, our well-being and our "salvation". It is also our freedom and autonomy, as we approach the condition wherein what happens to us follows from our nature (as a determinate and determined mode of one of God's attributes) alone and not as a result of the ways external things affect us. Spinoza's "free person" is one who bears the gifts and losses of fortune with equanimity, does only those things that he believes to be "the most important in life", takes care for the well-being of others (doing what he can to insure that they, too, achieve some relief from the disturbances of the passions through understanding), and is not anxious about death. The free person neither hopes for any eternal, otherworldly rewards nor fears any eternal punishments. He knows that the soul is not immortal in any personal sense, but is endowed only with a certain kind of eternity. The more the mind consists of true and adequate ideas (which are eternal), the more of it remains -- within God's attribute of Thought -- after the death of the body and the disappearance of that part of the mind that corresponds to

the body's duration. This understanding of his place in the natural scheme of things brings to the free individual true peace of mind.

There are a number of social and political ramifications that follow from Spinoza's ethical doctrines of human action and well-being. Because disagreement and discord between human beings is always the result of our different and changeable passions, "free" individuals -- who all share the same nature and act on the same principles -- will naturally and effortlessly form a harmonious society. "Insofar as men are torn by affects that are passions, they can be contrary to one another . . .[But] insofar as men live according to the guidance of reason, they must do only those things that are good for human nature, and hence, for each man, i.e., those things that agree with the nature of each man. Hence, insofar as men live according to the guidance of reason, they must always agree among themselves" (IVp34-35). Free human beings will be mutually beneficial and useful, and will be tolerant of the opinions and even the errors of others. However, human beings do not generally live under the guidance of reason. The state or sovereign, therefore, is required in order to insure -- not by reason, but by the threat of force -- that individuals are protected from the unrestrained pursuit of self-interest on the part of other individuals. The transition from a state of nature, where each seeks his own advantage without limitation, to a civil state involves the universal renunciation of certain natural rights -- such as "the right everyone has of avenging himself, and of judging good and evil" -- and the investment of those prerogatives in a central authority. As long as human beings are guided by their passions, the state is necessary to bring it about that they "live harmoniously and be of assistance to one another".

3. *Theological-Political Treatise*

The ostensive aim of the *Theological-Political Treatise*, widely vilified in its time, is to show that the freedom to philosophize can not only be granted without injury to piety and the peace of the Commonwealth, but that the peace of the Commonwealth and Piety are endangered by the suppression of this freedom". But Spinoza's ultimate intention is reveal the truth about Scripture and religion, and thereby to undercut the political power exercised in modern states by religious authorities. He also defends, at least as a political ideal, the tolerant, secular, and democratic polity.

On Religion and Scripture

Spinoza begins the treatise by alerting his readers, through a kind of "natural history of religion", to just those superstitious beliefs and behaviors that clergy, by playing on ordinary human emotions, encourage in their followers. A person guided by fear and hope, the main emotions in a life devoted to the pursuit of temporal advantages, turns, in the face of the vagaries of fortune, to behaviors calculated to secure the goods he desires. Thus, we pray, worship, make votive offerings, sacrifice and engage in all the various rituals of popular religion. But the emotions are as fleeting as the objects that occasion them, and thus the superstitions grounded in those emotions subject to fluctuations. Ambitious and self-serving clergy do their best to stabilize this situation and give some permanence to those beliefs and behaviors. "Immense efforts have been made to invest religion, true or false, with such pomp and ceremony that it can sustain any shock and constantly evoke the deepest reverence in all its worshippers." Religious leaders are

generally abetted in their purposes by the civil authority, which threatens to punish all deviations from theological orthodoxy as "sedition". The result is a state religion that has no rational foundations, a mere "respect for ecclesiastics" that involves adulation and mysteries but no true worship of God.

The solution to this state of affairs, Spinoza believes, is to examine the Bible anew and find the doctrines of the "true religion". Only then will we be able to delimit exactly what we need to do to show proper respect for God and obtain blessedness. This will reduce the sway that religious authorities have over our emotional, intellectual and physical lives, and reinstate a proper and healthy relationship between the state and religion. A close analysis of the Bible is particularly important for any argument that the freedom of philosophizing -- essentially, freedom of thought and speech -- is not prejudicial to piety. If it can be demonstrated that Scripture is not a source of "natural truth", but the bearer of only a simple moral message ("Love your neighbor"), then people will see that "faith is something separate from philosophy". Spinoza intends to show that in that moral message alone -- and not in Scripture's words or history -- lies the sacredness of what is otherwise merely a human document. The Bible teaches only "obedience [to God]", not knowledge. Thus, philosophy and religion, reason and faith, inhabit two distinct and exclusive spheres, and neither should tread in the domain of the other. The freedom to philosophize and speculate can therefore be granted without any harm to true religion. In fact, such freedom is essential to public peace and piety, since most civil disturbances arise from sectarian disputes. The real danger to the Republic comes from those who would worship not God, but some words on a page: "It will be said that, although God's law is inscribed in our hearts, Scripture is nevertheless the Word of God, and it is no more permissable to say of Scripture that it is mutilated and contaminated than to say this of God's Word. In reply, I have to say that such objectors are carrying their piety too far, and are turning religion into superstition; indeed, instead of God's Word they are beginning to worship likenesses and images, that is, paper and ink."

From a proper and informed reading of Scripture, a number of things become clear. First, the prophets were not men of exceptional intellectual talents -- they were not, that is, naturally gifted philosophers -- but simply very pious, even morally superior individuals endowed with vivid imaginations. They were able to perceive God's revelation through their imaginative faculties via words or real or imaginary figures. This is what allowed them to apprehend that which lies beyond the boundary of the intellect. Moreover, the content of a prophecy varied according to the physical temperament, imaginative powers, and particular opinions or prejudices of the prophet. It follows that prophecy, while it has its origins in the power of God -- and in this respect it is, in Spinoza's metaphysical scheme, no different from any other natural event -- does not provide privileged knowledge of natural or spiritual phenomena. The prophets are not necessarily to be trusted when it comes to matters of the intellect, on questions of philosophy, history or science; and their pronouncements set no parameters on what should or should not be believed about the natural world on the basis of our rational faculties.

Spinoza provides an equally deflationary account of God's election, or the "vocation", of the Hebrews. It is "childish", he insists, for anyone to base their happiness on the uniqueness of their gifts; in the case of the Jews, it would be the uniqueness of their being chosen among all people. The ancient Hebrews, in fact, did not surpass other nations in their wisdom or in their proximity to God. They were neither intellectually nor morally superior to other peoples. They were "chosen" only with respect to their social

organization and political good fortune. God (or Nature) gave them a set of laws and they obeyed those laws, with the natural result that their society was well-ordered and their autonomous government persisted for a long time. Their election was thus a temporal and conditional one, and their kingdom is now long gone. Thus, "at the present time there is nothing whatsoever that the Jews can arrogate to themselves above other nations." Spinoza thereby rejects the particularism that many -- including Amsterdam's Sephardic rabbis -- insisted was essential to Judaism. True piety and blessedness are universal in their scope and accesssible to anyone, regardless of their confessional creed.

Central to Spinoza's analysis of the Jewish religion -- although it is applicable to any religion whatsoever -- is the distinction between the divine law and the ceremonial law. The law of God commands only the knowledge and love of God and the actions required for attaining that condition. Such love must arise not from fear of possible penalties or hope for any rewards, but solely from the goodness of its object. The divine law does *not* demand any particular rites or ceremonies such as sacrifices or dietary restrictions or festival observances. The six hundred and thirteen precepts of the Torah have nothing to do with blessedness or virtue. They were directed only at the Hebrews so that they might govern themselves in an autonomous state. The ceremonial laws helped preserve their kingdom and insure its prosperity, but were valid only as long as that political entity lasted. They are not binding on all Jews under all circumstances. They were, in fact, instituted by Moses for a purely practical reason: so that people might do their duty and not go their own way. This is true not just of the rites and practices of Judaism, but of the outer ceremonies of all religions. None of these activities have anything to do with true happiness or piety. They serve only to control people's behavior and preserve a particular society.

A similar practical function is served by stories of miracles. Scripture speaks in a language suited to affect the imagination of ordinary people and compel their obedience. Rather than appealing to the natural and real causes of all events, its authors sometimes narrate things in a way calculated to move people -- particularly uneducated people -- to devotion. "If Scripture were to describe the downfall of an empire in the style adopted by political historians, the common people would not be stirred . . ." Strictly speaking, however, miracles -- understood as divinely caused departures from the ordinary course of nature -- are impossible. Every event, no matter how extraordinary, has a natural cause and explanation. "Nothing happens in nature that does not follow from her laws." This is simply a consequence of Spinoza's metaphysical doctrines. Miracles as traditionally conceived require a distinction between God and nature, something that Spinoza's philosophy rules out in principle. Moreover, nature's order is inviolable in so far as the sequence of events in nature is a necessary consequence of God's attributes. There certainly are "miracles" in the sense of events whose natural causes are unknown to us, and which we therefore attribute to the powers of a supernatural God. But this is, once again, to retreat to superstition, "the bitter enemy of all true knowledge and true morality".

By analyzing prophecy in terms of vividness of imagination, Jewish election as political fortune, the ceremonial law as a kind of social and political expediency, and the belief in miracles as an ignorance of nature's necessary causal operations, Spinoza naturalizes (and, consequently, demystifies) some of the fundamental elements of Judaism and other religions and undermines the foundations of their external, superstitious rites. At the same time, he thereby reduces the fundamental doctrine of piety to a simple and universal formula, naturalistic in itself, involving love and knowledge. This process of naturalization

achieves its stunning climax when Spinoza turns to consider the authorship and interpretation of the Bible itself. Spinoza's views on Scripture constitute, without question, the most radical theses of the *Treatise*, and explain why he was attacked with such vitriol by his contemporaries. Others before Spinoza had suggested that Moses was not the author of the entire Pentateuch. But no one had taken that claim to the extreme limit that Spinoza did, arguing for it with such boldness and at such length. Nor had anyone before Spinoza been willing to draw from it the conclusions about the status, meaning and interpretation of Scripture that Spinoza drew.

Spinoza denied that Moses wrote all, or even most of the Torah. The references in the Pentateuch to Moses in the third person; the narration of his death and, particularly, of events following his death; and the fact that some places are called by names that they did not bear in the time of Moses all "make it clear beyond a shadow of doubt" that the writings commonly referred to as "the Five Books of Moses" were, in fact, written by someone who lived many generations after Moses. Moses did, to be sure, compose some books of history and of law; and remnants of those long lost books can be found in the Pentateuch. But the Torah as we have it, as well as as other books of the Hebrew Bible (such as Joshua, Judges, Samuel and Kings) were written neither by the individuals whose names they bear nor by any person appearing in them. Spinoza believes that these were, in fact, all composed by a single historian living many generations after the events narrated, and that this was most likely Ezra. It was the post-exilic leader who took the many writings that had come down to him and began weaving them into a single (but not seamless) narrative. Ezra's work was later completed and supplemented by the editorial labors of others. What we now possess, then, is nothing but a compilation, and a rather mismanaged, haphazard and "mutilated" one at that.

As for the books of the Prophets, they are of even later provenance, compiled (or "heaped together", in Spinoza's view) by a chronicler or scribe perhaps as late as the Second Temple period. Canonization into Scripture occurred only in the second century BCE, when the Pharisees selected a number of texts from a multitude of others. Because the process of transmission was a historical one, involving the conveyence of writings of human origin over a long period of time through numerous scribes, and because the decision to include some books but not others was made by fallible human beings, there are good reasons for believing that a significant portion of the text of the "Old Testament" is corrupt.

Now in 1670 there was nothing novel in claiming that Moses did not write all of the Torah. Spinoza's most radical and innovative claim, in fact, was to argue that this holds great significance for how Scripture is to be read and interpreted. He was dismayed by the way in which Scripture itself was worshipped, by the reverence accorded to the words on the page rather than to the message they conveyed. If the Bible is an historical (i.e., natural) document, then it should be treated like any other work of nature. The study of Scripture, or Biblical hermeneutics, should therefore proceed as the study of nature, or natural science proceeds: by gathering and evaluating empirical data, that is, by examining the "book" itself -- along with the contextual conditions of its composition -- for its general principles.

> I hold that the method of interpreting Scripture is no different from the method of interpreting Nature, and is in fact in complete accord with it. For the method of interpreting Nature consists essentially in composing a detailed study of Nature from which, as being

> the source of our assured data, we can deduce the definitions of the things of Nature. Now in exactly the same way the task of Scriptural interpretation requires us to make a straightforward study of Scripture, and from this, as the source of our fixed data and principles, to deduce by logical inference the meaning of the authors of Scripture. In this way -- that is, by allowing no other principles or data for the interpretation of Scripture and study of its contents except those that can be gathered only from Scripture itself and from a historical study of Scripture -- steady progress can be made without any danger of error, and one can deal with matters that surpass our understanding with no less confidence than those matters that are known to us by the natural light of reason.

Just as the knowledge of nature must be sought from nature alone, so must the knowledge of Scripture -- an apprehension of its intended meaning -- be sought from Scripture alone and through the appropriate exercise of rational inquiry.

When properly interpreted, the universal message conveyed by Scripture is a simple moral one: "To know and love God, and to love one's neighbor as oneself". This is the *real* word of God and the foundation of true piety, and it lies uncorrupted in a faulty, tampered and corrupt text. The lesson involves no metaphysical doctrines about God or nature, and requires no sophisticated training in philosophy. The object of Scripture is not to impart knowledge, but to compel obedience and regulate our conduct. "Scriptural doctrine contains not abstruse speculation or philosophic reasoning, but very simple matters able to be understood by the most sluggish mind." Spinoza claims, in fact, that a familiarity with Scripture is not even necessary for piety and blessedness, since its message can be known by our rational faculties alone, although with great difficulty for most people. "He who, while unacquainted with these writings, nevertheless knows by the natural light that there is a God having the attributes we have recounted, and who also pursues a true way of life, is altogether blessed."

It follows that the only practical commandments that properly belong to religion are those that are necessary to carry out the moral precept and "confirm in our hearts the love of our neighbor". "A catholic faith should therefore contain only those dogmas which obedience to God absolutely demands, and without which such obedience is absolutely impossible . . . these must all be directed to this one end: that there is a Supreme Being who loves justice and charity, whom all must obey in order to be saved, and must worship by practicing justice and charity to their neighbor." As for other dogmas, "every person should embrace those that he, being the best judge of himself, feels will do most to strengthen in him love of justice".

This is the heart of Spinoza's case for toleration, for freedom of philosophizing and freedom of religious expression. By reducing the central message of Scripture -- and the essential content of piety -- to a simple moral maxim, one that is free of any superfluous speculative doctrines or ceremonial practices; and by freeing Scripture of the burden of having to communicate specific philosophical truths or of prescribing (or proscribing) a multitude of required behaviors, he has demonstrated both that philosophy is independent from religion and that the liberty of each individual to interpret religion as he wishes can be upheld without any detriment to piety.

> As to the question of what God, the exemplar of true life, really is, whether he is fire, or spirit, or light, or thought, or something else, this is irrelevant to faith. And so likewise is the question as to why he is the exemplar of true life, whether this is because he has a just and merciful disposition, or because all things exist and act through him and consequently we, too, understand through him, and through him we see what is true, just and good. On these questions it matters not what beliefs a man holds. Nor, again, does it matter for faith whether one believes that God is omnipresent in essence or in potency, whether he directs everything from free will or from the necessity of his nature, whether he lays down laws as a rule or teaches them as being eternal truths, whether man obeys God from free will or from the necessity of the divine decree, whether the rewarding of the good and the punishing of the wicked is natural or supernatural. The view one takes on these and similar questions has no bearing on faith, provided that such a belief does not lead to the assumption of greater license to sin, or hinders submission to God. Indeed . . . every person is in duty bound to adapt these religious dogmas to his own understanding and to interpret them for himself in whatever way makes him feel that he can the more readily accept them with full confidence and conviction.

Faith and piety belong not to the person who has the most rational argument for the existence of God or the most thorough philosophical understanding of his attributes, but to the person "who best displays works of justice and charity".

The State

Spinoza's account of religion has clear political ramifications. There had always been a quasi-political agenda behind his decision to write the *Treatise*, since his attack was directed at political meddling by religious authorities. But he also took the opportunity to give a more detailed and thorough presentation of a general theory of the state that is only sketchily present in the *Ethics*. Such an examination of the true nature of political society is particularly important to his argument for intellectual and religious freedom, since he must show that such freedom is not only compatible with political well-being, but essential to it.

The individual egoism of the *Ethics* plays itself out in a pre-political context -- the so-called "state of nature", a universal condition where there is no law or religion or moral right and wrong -- as the right of every individual to do whatever he can to preserve himself. "Whatever every person, whenever he is considered as solely under the dominion of Nature, believes to be to his advantage, whether under the guidance of sound reason or under passion's sway, he may by sovereign natural right seek and get for himself by any means, by force, deceit, entreaty, or in any other way he best can, and he may consequently regard as his enemy anyone who tries to hinder him from getting what he wants." Naturally, this is a rather insecure and dangerous condition under which to live. In Hobbes' celebrated phrase -- and Spinoza was clearly influenced by his reading of that British thinker -- life in the state of nature is "solitary, poor, nasty, brutish and short". As rational creatures, we soon realize that we would be better off, still from a thoroughly egoistic perspective, coming to an agreement among ourselves to restrain our opposing desires and the unbounded pursuit of self-interest -- in sum, that it would be in our greater self-interest to live under the law of reason rather than the law of nature. We thus agree to hand over to a

sovereign our natural right and power to do whatever we can to satisfy our interests. That sovereign -- whether it be an individual (in which case the resulting state is a monarchy), a small group of individuals (an oligarchy) or the body-politic as a whole (a democracy) -- will be absolute and unrestrained in the scope of its powers. It will be charged with keeping all the members of society to the agreement, mostly by playing on their fear of the consequences of breaking the "social contract".

Obedience to the sovereign does not infringe upon our autonomy, since in following the commands of the sovereign we are following an authority whom we have freely authorized and whose commands have no other object than our own rational self- interest. The type of government most likely to respect and preserve that autonomy, issue laws based on sound reason and to serve the ends for which government is instituted is democracy. It is the "most natural" form of governing arising out of a social contract -- since in a democracy the people obey only laws that issue from the general will of the body politic -- and the least subject to various abuses of power. In a democracy, the rationality of the sovereign's commands is practically secured, since it is unlikely that a majority of a large number of people will agree to an irrational design. Monarchy, on the other hand, is the least stable form of government and the one most likely to degenerate into tyranny.

Since the outward practices of religion impinge upon the comportment and relations of citizens, they fall under "state business" and, thus, within the sphere of the sovereign's power. The sovereign should have complete dominion in all public matters secular and spiritual. There should be no church separate from the religion instituted and regulated by the state. This will prevent sectarianism and the multiplication of religious disputes. All questions concerning external religious rites and ceremonies are in the hands of the sovereign. This is in the best interest of everyone, since the sovereign will, ideally and in conformity with its "contractual" duty, insure that such practices are in accord with public peace and safety and social well-being. The sovereign should rule in such a way that his commands enforce God's law. Justice and charity thereby acquire the force of civil law, backed by the power of the sovereign.

On the other hand, dominion over the "inward worship of God" and the beliefs accompanying it -- in other words, inner piety -- belongs exclusively to the individual. This is a matter of inalienable, private right, and it cannot be legislated, not even by the sovereign. No one can limit or control another person's thoughts anyway, and it would be foolhardy and destructive to the polity for a sovereign to attempt to do so. Nor can speech ever truly and effectively be controlled, since people will always say want they want, at least in private. "Everyone is by absolute natural right the master of his own thoughts, and thus utter failure will attend any attempt in a commonwealth to force men to speak only as prescribed by the sovereign despite their different and opposing opinions." There must, Spinoza grants, be *some* limits to speech and teaching. Seditious discourse that encourages individuals to nullify the social contract should not be tolerated. But the best government will err on the side of leniency and allow the freedom of philosophical speculation and the freedom of religious belief. Certain "inconveniences" will, no doubt, sometimes result from such an extensive liberty. But the attempt to regulate everything by law is "more likely to arouse vices than to reform them". In a passage that foreshadows John Stuart Mill's utilitarian defense of liberty nearly two centuries later, Spinoza adds that "this freedom is of the first importance in fostering the sciences and the arts, for only those whose judgment is free and unbiased can attain success in these fields."

It is hard to imagine a more passionate and reasoned defense of freedom and toleration than that offered by Spinoza.

Bibliography

Spinoza's Works

- *Spinoza Opera*, edited by Carl Gebhardt, 5 volumes (Heidelberg: Carl Winters, 1925, 1972 [volume 5, 1987]).
- Spinoza, *Ethics*, in Edwin Curley, translator, *The Collected Writings of Spinoza* (Princeton: Princeton University Press, 1985), volume 1.
- Spinoza, *Theological-Political Treatise*, Samuel Shirley, translator (Leiden: Brill, 1989).

Recommended Secondary Literature

- Allison, Henry (1987), *Benedict de Spinoza: An Introduction* (New Haven: Yale University Press).
- Bennett, Jonathan (1984), *A Study of Spinoza's Ethics* (Indianapolis: Hackett Publishing).
- Curley, Edwin (1988), *Behind the Geometric Method* (Princeton: Princeton University Press).
- Donagan, Alan (1988), *Spinoza* (Chicago: University of Chicago Press).
- Garrett, Don, ed. (1996), *The Cambridge Companion to Spinoza* (Cambridge and New York: Cambridge University Press).
- Nadler, Steven (1999), *Spinoza: A Life* (Cambridge and New York: Cambridge University Press).
- Smith, Steven B. (1997), *Spinoza, Liberalism and the Question of Jewish Identity* (New Haven: Yale University Press).
- Wolfson, Harry (1934), *The Philosophy of Spinoza*, 2 vols. (Cambridge, MA: Harvard University Press).
- Yovel, Yirmiyahu (1989), *Spinoza and Other Heretics*, 2 vols. (Princeton: Princeton University Press).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Descartes, René | Leibniz, Gottfried Wilhelm

smnadler@facstaff.wisc.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 29, 2001
Content last modified: June 29, 2001

Medieval Theories of Modality

There are four modal paradigms in ancient philosophy: the ‘statistical’ or ‘temporal frequency interpretation’ of modality, the model of possibility as a potency, the model of antecedent necessities and possibilities with respect to a certain moment of time (diachronic modalities), and the model of possibility as non-contradictoriness. None of these conceptions, which were well known to early medieval thinkers through the works of Boethius, was associated with the idea of modality as involving reference to simultaneous alternatives. This new paradigm was introduced into Western thought in early twelfth-century discussions influenced by Augustine's theological conception of God as acting by choice between alternative histories. Ancient habits of thinking continued to play an important role in scholasticism, however, and the theoretical significance of the new conception was not fully realized before the works of John Duns Scotus and some other early fourteenth century thinkers.

Aspects of Ancient Modal Paradigms

In speaking about the general features of the universe, ancient philosophers were inclined to think that all generic possibilities will be actualized, a habit of thinking called the principle of plenitude by Arthur O. Lovejoy (1936). Correspondingly, it was natural for them to think that the types of things which never occur are impossible and that the invariant structures of reality are necessary. This line of thought is found, e.g., in Plato's doctrine of ideas which are exhaustively imitated in the Receptacle, in Aristotle's theory of the priority of actuality over potentiality, in the Stoic doctrine of God, the world-order, and the eternal cosmic cycle, and in Plotinus's metaphysics of emanation (Knuuttila 1993).

In these approaches to the constituents of the universe, modal terms are used in accordance with the so-called ‘statistical’ or ‘temporal frequency’ model of modality where the meaning of modal terms is

spelled out extensionally as follows: what is necessary is always actual, what is impossible is never actual and what is possible is at least sometimes actual. The term 'statistical interpretation of modality' was introduced into the modern discussion by Oscar Becker (1952), and it has been applied since in descriptions of certain ways of thinking in the history of philosophy as well, particularly by Jaakko Hintikka (1973).

Even though Aristotle did not define modal terms with the help of extensional notions, this model can be found in his discussion of eternal beings, the natures of things, the types of events, or generic statements about such things. Modal terms refer to the one and only world of ours and classify the types of things and events on the basis of their occurrence. This paradigm suggests that actualization is the general criterion of the genuineness of possibilities, but the deterministic implications of this view compelled Aristotle to seek ways of speaking about unrealized singular possibilities. Diodorus Chronus (fl. 300 B.C.) was a determinist who found no problem in this way of thinking. (For different interpretations and evaluations of the role of this model in Aristotle, see Hintikka 1973, Sorabji 1980, Seel 1982, Waterlow 1982a, White 1985, van Rijen 1989, Gaskin 1995.) In *Posterior Analytics* I.6 Aristotle seems to imply that certain predicates may belong to their subjects at all times without belonging to them necessarily. Averroes and his followers took this to mean that invariant connections are necessary in a strict sense (*per se*) only if they are essential; otherwise they are accidentally necessary (Lagerlund 1999). This is one of the texts some modern scholars have referred to in arguing that Aristotle's views showing similarities to the statistical model are not based on the meaning of modal terms but on some special metaphysical and ontological doctrines (van Rijen 1989; cf. Waterlow 1982a).

Another Aristotelian modal paradigm was that of possibility as potency. In *Met*. V.12 and IX.1 potency is said to be the principle of motion or change either as the activator or as the receptor of a relevant influence. (For agent and patient in Aristotle's natural philosophy in general, see Waterlow 1982b.) The types of potency-based possibilities belonging to a species are recognized as possibilities because of their actualization - no natural potency type remains eternally frustrated. Aristotle says that when the agent and the patient come together as being capable, the one must act and the other must be acted on (*Met*. IX.5). I shall return to this formulation.

In *De Caelo* I.12 Aristotle supposes, *per impossibile*, that a thing has contrary potencies, one of which is always actualized. He argues that the alleged unactualized potencies cannot be real, because one cannot assume them to be realized at any time without contradiction. Aristotle applies here the model of possibility as non-contradictoriness which is defined in *Prior Analytics* I.13 as follows: when a possibility is assumed to be realized, it results in nothing impossible. In speaking about the assumed non-contradictory actualization of a possibility Aristotle thinks that it is realized in our one and only history. The argument in *De caelo* excludes from the set of genuine possibilities those which remain eternally unrealized. It also shows how strongly Aristotle's modal thought was influenced by the absence of the idea of synchronic alternatives. (See also *Met*. IX.4.)

Aristotle heavily criticized some of his contemporaries who claimed that only that which takes place is possible (*Met*. IX.3). His problem was that the assumptions of his modal thinking pushed him towards a very similar position with respect to singular possibilities. In Chapter 9 of *De interpretatione* Aristotle

says that what is necessarily is when it is, but he then qualifies this necessity of the present with the remark that it does not follow that what is actual is necessary without qualification. If he meant that the temporal necessity of a present event does not imply that such an event necessarily takes place in circumstances of that type, this is an unsatisfactory attempt to avoid the problem that changeability as a criterion of contingency makes all temporally definite singular events necessary (Hintikka 1973). Another possibility is that Aristotle wanted to show that the necessity of an event at a certain time does not imply that it would have been antecedently necessary (von Wright 1984). The remark on the necessity of the present is included in Aristotle's discussion of future contingent statements which is one of the most controversial themes of Aristotle's philosophy. The model of possibility as potency *prima facie* allowed Aristotle to speak about all kinds of unrealized singular possibilities by referring to passive or active potencies, but taken separately they represent partial possibilities which do not guarantee that their actualization can take place. More is required for a real singular possibility, but when the further requirements are added, such as a contact between the active and passive factor and the absence of an external hindrance, the potency model suggests that the potency can really be actualized only when a change towards its end is initiated (*Met*. IX.5, *Phys*. VIII.1). Some scholars have referred to the diachronic idea that the conditions which at t_1 ($t_1 < t_2$) are necessary for the obtaining of ‘p’ at t_2 are not necessarily sufficient for the obtaining of ‘*p*’ though they are sufficient for the possibility (at t_1) that ‘*p*’ obtains at t_2 (Gaskin 1995). Unfortunately Aristotle seldom was that explicit. Some others have paid attention to Aristotle's definition of the time-taking process (*kinêsis*) as the actuality of the potentiality (of the end) *qua* potentiality (*Phys*. III.1), but this did not help him more than offering a place for full singular possibilities of what will be (Hintikka et al. 1977).

Aristotle's conceptual difficulties can be seen from his various attempts to characterize the possibilities based on dispositional properties such as heatable, separable, or countable. Analogous discussions were not unusual in later ancient philosophy. In Philo's definition of possibility (ca. 300 B.C.), the existence of a passive potency was regarded as a sufficient ground for speaking about a singular possibility. The Stoics revised this definition by adding the condition of the absence of external hindrance, thinking that otherwise the alleged possibility could not be realized. They did not add that an activator is needed as well, because then the difference between potentiality and actuality would disappear. According to the deterministic Stoic world view, fate as a kind of active potency necessitates everything, but the number of passive potencies with respect to a definite future instant of time (t_1) is greater than what will be realized. As long as these possibilities are not prevented from being realized by other things which will be actual at t_1, they in some sense represent open possibilities. When t_1 is present, all unrealized possibilities are prevented from being actualized by other things. (The Stoics did not accept the Master Argument of Diodorus Cronus against possibilities which will not be realized.) Passive potencies as alternative prospective possibilities show what might happen at a certain moment, but because everything is determined, the alternatives seem to be only epistemic possibilities relative to our ignorance. (For different interpretations of the Stoic and Megarian conceptions of modality, see Vuillemin 1984, White 1985, Bobzien 1986, Engberg-Pedersen 1990, Bobzien 1993, Gaskin 1995.)

Alexander of Aphrodisias claimed in *De fato* that the Peripatetics, as distinct from the Stoics, thought that there are genuine prospective alternatives which remain open options until the moment of time to which

they refer. It was the Stoic doctrine of future alternatives which led Alexander to consider diachronic modalities which he then tried to interpret in a different way (Sharples 1983). Aristotle sometimes referred to diachronic modalities of this kind (*Met*. VI.3), but he did not elaborate this idea, which might have been his most promising attempt to formulate a theory of unrealized singular possibilities. (The importance of this model is particularly stressed in Waterlow 1982a; see also Weidemann 1986, Gaskin 1995.) Neither Aristotle nor his followers had any conception of synchronic alternatives. They thought that what is necessarily is when it is, and that the alternative possibilities disappear when the future is fixed. The Peripatetic theory of alternative prospective possibilities could be called the model of diachronic modalities without synchronic alternatives: there are transient singular alternative possibilities, but those which will not be realized disappear instead of remaining unrealized.

Early Medieval Developments

The early medieval thinkers were well acquainted with ancient modal conceptions through Boethius's works. One of the Aristotelian modal paradigms occurring in Boethius is that of possibility as potency (*potestas*, *potentia*). According to Boethius, when the term ‘possibility’ (*possibilitas*) is applied in the sense of ‘potency’, it refers to real powers or tendencies, the ends of which are either actual or non-actual at the moment of utterance. Some potencies are never unrealized. They are said to be necessarily actual. When potencies are not actualized, their ends are said to exist potentially (*In Periherm*. II.453-455). Necessarily actual potencies leave no room for the potencies of their contraries, for they would remain unrealized forever and the constitution of nature cannot include elements which would be in vain (*In Periherm*. II. 236, 243). The potencies of non-necessary features of being do not exclude contrary potencies. They are not always and universally actualized, but as potency-types even these potencies are taken to satisfy the actualization criterion of genuineness (*In Periherm*. I.120-1, II.237).

Boethius's view that the types of potencies and potency based possibilities are sometimes actualized is in agreement with the Aristotelian statistical interpretation of modality. This is another Boethian conception of necessity and possibility. He thought that modal notions can be regarded as tools for expressing temporal or generic frequencies. According to the temporal version, what always is is by necessity, and what never is is impossible. Possibility is interpreted as expressing what is at least sometimes actual. Correspondingly, a generic property of a species is possible only if it is exemplified at least in one member of that species (*In Periherm*. I.120-1, 200-201, II.237, 239).

Like Aristotle, Boethius often treated statement-making utterances as temporally indeterminate sentences. The same sentence can be uttered at different times, and many of these temporally indeterminate sentences may sometimes be true and sometimes false, depending on the circumstances at the moment of utterance. If the state of affairs the actuality of which makes the sentence true is omnitemporally actual, the sentence is true whenever it is uttered. In this case, it is necessarily true. If the state of affairs associated with an assertoric sentence is always non-actual, the sentence is always false and therefore impossible. A sentence is possible only if what is asserted is not always non-actual (I.124-125). The statistical interpretation of modal terms is also employed in Ammonius's Greek commentary on Aristotle's *De interpretatione* which shares some sources with Boethius's work (88.12-28).

In dealing with Chapter 9 of Aristotle's *De interpretatione* Boethius argues (II.241) that because

> (1) $M(p_t \ \& \ \neg p_t)$
> (1) It is possible that p obtains at t and not-p obtains at t.

is not acceptable, one should also deny

> (2) $p_t \ \& \ M_t \neg p_t$
> (2) p obtains at t and it is possible at t that not-p obtains at t.

The denial of (2) is equivalent to

> (3) $p_t \rightarrow L_t p_t$
> (3) If p obtains at t then it is necessary at t that p obtains at t.

This line of thought is natural only when possibilities are treated without any idea of synchronic alternatives. (2) was generally denied in ancient philosophy and its denial was taken as an axiom by Boethius as well. Correspondingly, (3) shows how the necessity of the present was understood in ancient thought. Boethius thought that the temporal necessity of *p* can be qualified by shifting attention from temporally definite cases or statements to their temporally indeterminate counterparts (I.121-122, II.242-3). The same statistical idea occurs in Ammonius (153.24-6). This was one of Boethius's interpretations of the Aristotelian distinction between necessity now and necessity without qualification. But he also made use of the diachronic model according to which the necessity of ‘*p*’ at *t* does not imply that, before *t*, it is necessary that ‘*p*’ obtains at *t*.

Boethius developed the diachronic ideas as part of his criticism of Stoic determinism. If it is not true that everything is causally necessitated, there must be genuine alternatives in the course of events. Free choice was the source of contingency in which Boethius was mainly interested, but he thought in addition that according to the Peripatetic doctrine there is a real factor of indeterminacy in the causal nexus of nature. When Boethius refers to chance, free choice, and possibility in this context, his examples include temporalized modal notions which refer to diachronic prospective possibilities at a given moment of time. A temporally determinate prospective possibility is unrealized before the time to which it refers, and it may be not realized even when the time is present. This means, however, that the possibility no longer exists. Boethius did not develop the idea of simultaneous synchronic possibilities which would remain intact even when diachronic possibilities had vanished. On the contrary, he insisted that only what is actual at a certain time is at that time possible at that time (cf. (3) above). But he also thought that there are objective singular contingencies, so that the result of some prospective possibilities is indefinite and uncertain ‘not only to us who are ignorant, but to nature’ (*In Periherm*. I.106, 120, II.190-192, 197-198, 203, 207). Boethius's modal paradigms are discussed in Kretzmann 1985, Mignucci 1989, and Knuuttila 1993.

As for future contingent statements, Boethius seems to think the principle of bivalence is universally valid, but statements about future contingents, unlike those about past and present things, do not obey the stronger principle that each affirmative or negative statement is either determinately true or determinately false. A true statement is indeterminately true as long as the conditions which make it true are not yet fixed (*In Periherm*. I.125, II.208; Mignucci 1989). Boethius's formulations are somewhat ambiguous and it is possible that ‘indeterminate truth’ sometimes means that a statement will be either true or false (Kretzmann 1987). The first alternative became the standard medieval view, but there were different opinions of whether Aristotle abandoned bivalence for future contingent statements. (See Normore 1982, Lewis 1987, Normore 1993.) Boethius, Thomas Aquinas, and many others thought that God can know future contingents only because the flux of time is present to divine eternity. Many late medieval thinkers defended God's ability to foreknow free acts. This led to the so called middle knowledge theory of the counterfactuals of freedom (Craig 1988, Freddoso 1988).

From the point of view of the history of modal thought, interesting things took place in theology in the eleventh and twelfth centuries. Augustine had already criticized the application of the statistical model of possibility to divine power; for him, God has freely chosen the actual world and its providential plan from alternatives which he could have realized but did not will to do (*potuit sed noluit*). This way of thinking differs from ancient philosophical modal paradigms, because the metaphysical basis is now the eternal domain of alternative possible histories instead of the idea of one necessary world order (*De spiritu et littera* 1-2, *De civitate Dei* 12.19, 21.5-10, 22.4, 11, *Contra Faustum* 29.4). The idea of a discrepancy between the Catholic doctrine of God's freedom and power and the philosophical modal conceptions was brought into the scope of discussion by Peter Damian's *De divina omnipotentia* (Holopainen 1996) and was developed in a more sophisticated way by Peter Abelard, Gilbert of Poitiers and some other twelfth century authors. This is how the new modal paradigm based on the idea of synchronic alternatives became a part of Western theology, and it was particularly applied in the discussions of the distinction between God's absolute and ordained power and between divine and natural possibilities.

The modal paradigm based on synchronic alternatives was very different from the traditional ones, but there were few people in the twelfth and thirteenth centuries to realize its general philosophical significance. It was more usual to consider it a specially theological matter which did not affect the use of traditional modal paradigms in other disciplines. Abelard and Gilbert were inclined to think in this way, and this attitude was supported by the general reception of Aristotle's philosophy which clearly contributed to the frequent use of the Aristotelian modal paradigms in thirteenth century logical treatises on modalities, in metaphysical theories of the principles of being, and in the discussions of causes and effects in natural philosophy (Knuuttila 1993; for Jewish and Arabic discussions, see also Rescher 1974, Manekin 1992, Bäck 1992).

Modalities in Thirteenth Century Logical Treatises

The anonymous *Dialectica Monacensis* (ca. 1200) is one of the numerous works representing the new terminist approach to logic and can be used as an example of how modalities were treated in it. (A collection of late twelfth and early thirteenth century logical texts is edited in de Rijk 1962-67.) In

discussing the quantity (universal, particular, singular) and quality (affirmative, negative) of the modals, the author states that modal terms may be adverbial or nominal. The modal adverb qualifies the copula, and the structure of the sentence can be described as follows:

> (4) quantity/subject/modalized copula/predicate (Some A's are necessarily B)

In this form, the negation can be located in different places, either

> (5) quantity/subject/copula modalized by a negated mode/predicate (Some A's are-not-necessarily B)

or

> (6) quantity/subject/modalized negative copula/predicate (Some A's are-necessarily-not B)

If sentences with a negation sign are read in accordance with (5), then the mode is denied in them; if they are read in accordance with (6), the modal adverb qualifies a negated predication (De Rijk 1967, II-2, 479.35-480.3).

As for the modal sentences with nominal modes, the author says that they can be read in two ways. One can apply an adverbial type of reading to them, which is said to be how Aristotle treats modal sentences in the *Prior Analytics*. The quality and quantity of such a *de re* modal sentence is determined by the corresponding non-modal sentence. In a *de dicto* modal sentence that which is asserted in a non-modal sentence is considered as the subject about which the mode is predicated. When modal sentences are understood in this way, they are always singular, their form being:

> (7) subject/copula/mode. (That some A's are B is necessary.)

This reading is said to be the one which Aristotle presented in *De interpretatione* (480.3-26). The idea of the systematic distinction between the readings *de dicto* (*in sensu composito*) and *de re* (*in sensu diviso*) of modally qualified statements was introduced into medieval discussions in Abelard's investigations of modal statements (*Super Periherm*. 3-47, *Dialectica* 191.1-210.19), and was often mentioned, as in the *Dialectica Monacensis*, in discussions of the composition-division ambiguity of sentences.

The author of the *Dialectica Monacensis* says that the matter of an assertoric sentence may be natural, remote, or contingent. True affirmative sentences about a natural matter maintain the existence of natural compounds which cannot be otherwise; these sentences as well as the natural compounds are called necessary. False affirmative sentences about a remote matter maintain the existence of compounds which are necessarily non-existent; they are called impossible. Sentences about a contingent matter are about compounds which can be actual and which can be non-actual (472.9-473.22). The theory of the modal matter was popular in early medieval logic and was also dealt with in mid-thirteenth century handbooks. Another often discussed theme was the distinction between modalities *per se* and *per accidens* which was

based on the idea that the modal status of a temporally indefinite sentence may be changeable or not - for example ‘You have not been in Paris’ may begin to be impossible, whereas ‘You either have or have not been in Paris’ may not. Another distinction between sentences necessary *per se* and *per accidens* was based on Aristotle's theory of *per se* predication in *Posterior Analytics* I.4. A sentence was said to be accidentally necessary when it was unchangeably true but there was no conceptual connection between subject and predicate.

One example of the prevalence of the traditional use of modal notions can be found in the early medieval *de dicto/de re* analysis of examples such as ‘A standing man can sit’. It was commonly stated that the composite (*de dicto*) sense is ‘It is possible that a man sits and stands at the same time’ and that on this reading the sentence is false. The divided (*de re*) sense is ‘A man who is now standing can sit’ and on this reading the sentence is true. Many authors formulated the divided possibility as follows: ‘A standing man can sit at another time’. It was assumed that a possibility refers to an actualization in the one and only world history and that it cannot refer to the present moment because of the necessity of the present understood in the Aristotelian sense formulated in (2) and (3) above. When authors referred to another time, some of them thought in accordance with the statistical model that the possibility would be realized at that time. But the Boethian idea of diachronic prospective alternatives was also often used in thirteenth century logical treatises, and some authors thought that the divided possibility refers to the future even though it may remain unrealized. A third group of authors made use of the modern idea of synchronic alternatives in this connection. The composite reading refers then to one and the same state of affairs and the divided reading to alternative states of affairs (at the same time). This analysis was also applied to the question of whether God's knowledge of things makes them necessary. (There are textual references for all these themes in Knuuttila 1993. See also Maierù 1972, Jacobi 1980.)

The variety of intuitions about the meanings of modal notions may be one of the reasons for the fact that the logical analysis of *de re* modalities remained sketchy in early terminist logic. Modifying Boethius's systematization of Aristotle's remarks in *De interpretatione* 12 and 13, the logicians often presented the equipollences and other relations between unanalysed modals with the help of a square of opposition. Abelard's attempt to extend this analysis to quantified *de re* modals was badly confused (*Super Periherm*. 26.8-15) and the later progress in this area was slow. It was only in the fourteenth century that they were analysed in a satisfactory manner. (See Hughes 1989 and his description of Buridan's octagon of modal opposites and equipollences.)

Dialectica Monacensis involves a brief summary of Aristotle's modal syllogistic. (Its first Latin discussion is found in an anonymous twelfth century commentary on the *Prior Analytics*; Ebbesen 1981.) The first thorough commentary was Robert Kilwardby's *In libros Priorum Analyticorum expositio* (ca. 1240). Albert the Great's comments on modal syllogisms in his commentary on the *Prior Analytics* were mainly derived from Kilwardby's work. Abelard, who did not deal with Aristotle's modal syllogistics, said that the modals in mixed syllogisms with both modal and assertoric premises should be read *de re* (*Super Periherm*. 10.22-11.16) This became a common view before Aristotle's modal syllogistic was studied in any detail. Although the reception of the *Prior Analytics* with this interpretation strengthened the interest in quantified modal sentences, Aristotle's work was an ambiguous guidebook. The remarks on the structure of the premises are scattered and even if it is natural to think that the presupposition of the

mixed moods is a *de re* reading of modally qualified premises, it creates serious difficulties when it is applied to the conversion rules, most of which are unproblematic only if understood as rules for modals *de dicto*.

There are several recent works on Aristotle's modal syllogistic but no generally accepted historical construction which would make it a coherent theory. (For recent attempts of reconstruction, see van Rijen 1989, Patterson 1995, Thom 1996, Nortmann 1996.) Robert Kilwardby and Albert the Great thought that Aristotle's modal syllogistic was a consistent theory and that the difficulties of understanding some parts disappear when certain philosophical presuppositions are explicated. Instead of employing the *de dicto/de re* distinction they suggested that the notion of necessity in syllogistic premises refers to the Aristotelian *per se* necessity of an essential predication and they also applied many *ad hoc* restrictions pertaining to the notion of contingency in order to give a uniform reading of Aristotelian moods and conversion rules. This approach, which shows similarities to some modern reconstructions of Aristotle's theory, was partially influenced by Averroes's works. While Kilwardby's interpretation of modal statements and modal syllogisms was influential in the thirteenth century, in the early fourteenth century it gave way to a quite different theory (Lagerlund 1999).

The principles of propositional modal logic were generally expressed as follows: if the antecedent of a valid consequence is possible/necessary, the consequent is possible/necessary (Abelard, *Dialectica* 202.6-8). A great deal of Abelard's logical works consisted of discussions of topics, consequences and conditionals. Like Boethius, Abelard thought that true conditionals express necessary conceptual connections between the antecedents and the consequents. Some twelfth century masters regarded the principle that the antecedent is not true without the consequent as a sufficient condition for the truth of a conditional and accepted the so-called paradoxes of implication (Martin 1987). The question of the nature of conditionals and consequences remained a popular theme in medieval logic (Broadie 1993).

Fourteenth Century Discussions

John Duns Scotus's modal theory can be regarded as the first systematic exposition of the new intensional theory of modality, some elements of which were put forward in the twelfth century. In criticizing Henry of Ghent's theory of theological modalities, Scotus sketched the famous and sometimes misunderstood model of ‘divine psychology’ in which certain relations between theological, metaphysical, and modal notions are defined. According to Scotus, we can suppose that everything which can be understood eternally receives intelligible being (*esse intelligibile*) in the intellect of God. Because the intelligible being comprises the individual concept of each thing that can be known, all thinkable realizable individuals receive possible being (*esse possibile*) as the intentional correlates of divine power and choice. The possibilities are thought of as partitioned into classes on the basis of relations of compossibility. Impossibility means incompossibility between possible properties or states of affairs. One of the compossible sets into which the possibilities are partitioned is chosen by divine will to be actualized through divine power (*Ord*. I.43; other relevant texts are *Ord*. I.36 and *Lect*. I.39). Scotus wants to make it clear that, contrary to what Henry of Ghent did, one should not think that the domain of possibility is posterior to God's power. An infinite power can realize whatever is realizable, and it can be

decided which things are realizable without any reference to that power. When Scotus says that God produces things in intelligible being, he means that they are given an ontological status, weaker than existence, as intentional correlates of divine thought. Although God's intellect in this sense eternally produces its objects, it does not create them by choice. God's infinite intellect comprehends all that can be thought without contradiction and understands them necessarily and not freely. The intentional actuality of the domain of possibility in God's mind is caused by divine intellect, but possibilities and impossibilities themselves are what they are independently of whether anything exists. They would remain the a priori conditions of being and thinking even if there were no God, though they would then not have any kind of existence. Scotus thus gave up the view of Thomas Aquinas, Bonaventura, Henry of Ghent and many other thirteenth century theologians that absolute possibilities have an ontological foundation in God's essence. He thought that logical possibilities must be regarded as non-existing preconditions of any beings and their acts. Scotus heavily criticized extensional modal theories in which the actualization at some moment of time was explicitly or implicitly used as the criterion of the genuineness of possibility. He redefined a contingent event as follows: ‘I do not call something contingent because it is not always or necessarily the case, but because its opposite could be actual at the very moment when it occurs’ (*Ord.* I, *d.* 2, *p.* 1, *q.* 1-2, n. 86). The criterion of what is possible is that we can imagine it taking place without contradiction in at least one alternative state of affairs. This is what ‘logical possibility’ means. The domain of possibility as such consists of all thinkable individuals, their properties, and their mutual relations. Because many possibilities are mutually exclusive, this domain must be structured into sets on the basis of compossibility relations. Scotus's theory can be characterized as an intuitive predecessor of possible worlds semantics, though he did not use the term ‘possible world’ in a technical sense. Scotus took it for granted that the same individual can occur in alternative states of affairs. In this respect his theory of possible individuals was different from Leibniz's view that individuals are world-bound. (For Scotus's modal theory, see Knuuttila 1993, 1996, Vos et al. 1994, Lewis 1996, Marrone 1996. Normore 1996 argues for a revised version of the traditional interpretation that modalities are primarily dependent on God.)

Scotus's model theoretical approach to modalities brought some new themes into philosophical discussion. One of these was the idea of the domain of possibility as a non-existent objective precondition of all being and thinking. This was well known in the seventeenth century as well through Suárez's works (Honnefelder 1990). In his discussion of eternal truths, Descartes criticized the classical view of the ontological foundation of modality as well as the Scotist theory of modality and conceivability. He seems to have thought that the domain of conceivability is freely set by God and that it could therefore be different from what it is. (There are different views of Descartes's theory and its connections to late medieval views; see Alanen and Knuuttila 1988, Alanen 1990, Normore 1991.)

Another influential idea was the new distinction between logical and natural necessities and possibilities. In Scotus's theory, logically necessary attributes and relations are attached to things in all those sets of compossibilities in which they occur. Against this background one could ask which of the natural invariances treated as necessities in earlier natural philosophy were necessary in this strong sense of necessity, and which of them were merely empirical generalizations without being logically necessary. The distinction between logical and natural necessity is crucial to the works of William Ockham and John Buridan.

One important branch of medieval logic developed in treatises called *De obligationibus* dealt, roughly speaking, with how an increasing set of true and false propositions might remain coherent. According to the thirteenth century rules, false present tense statements could be accepted only if they were taken to refer to a moment of time different from the actual one. Scotus deleted this rule, based on the Aristotelian axiom of the necessity of the present, and later theories accepted the Scotist revision. In this new form, obligations logic could be regarded as a theory of how to describe logically possible states of affairs and their mutual relationships. These discussions influenced the philosophical theory of counterfactual conditionals (Yrjönsuuri 1994.)

With the new modal semantics, William Ockham, John Buridan and some other fourteenth century authors could formulate the principles of modal logic much more completely and satisfactorily than did their predecessors. Questions of modal logic were discussed separately with respect to modal propositions *de dicto* and *de re*; modal propositions *de re* were further divided into two groups depending on whether the subject terms refer to actual or possible beings. It was thought that logicians should also analyse the relationships between these readings and, furthermore, the consequences having various types of modal sentences as their parts. Richard of Campsall played an interesting role in the development of medieval modal syllogistics. He introduced the habit of treating the *de dicto* and *de re* moods separately, but he was also dependent on Kilwardby's interpretation. The new modal logic of William Ockham, John Buridan and Pseudo-Scotus was among the most remarkable achievements of medieval logic. In its light Aristotle's modal syllogistic was regarded as a fragmentary theory in which the distinctions between different types of fine structures were not explicated. These authors did not try to reconstruct it into a uniform system; they believed, like some modern commentators, that such a reconstruction is not possible (Lagerlund 1999). Buridan's modal logic was dominant in late medieval times. It was embraced by Marsilius of Inghen, Albert of Saxony, and Jodocus Trutfetter. (For the later influence of medieval modal theories, see Coombs 1990, Roncaglia 1996.) The rise of the new modal logic was accompanied by theories of epistemic logic (Boh 1993) and deontic logic (Knuuttila 1993) which also belong among the remarkable achievements of late medieval philosophy.

Bibliography

- Abelard, Peter (1970), *Dialectica*, ed. L.M. de Rijk, Assen: Van Gorcum.
- Abelard, Peter (1958), *Super Periermenias XII-XIV*, ed. L. Minio-Paluello in *Twelfth Century Logic. Text and Studies II. Abaelardiana inedita*, Rome: Edizioni di Storia e Litteratura.
- Alanen, L. (1990), ‘Descartes, Conceivability and Logical Modality’ in T. Horowitz and G.J. Massey (eds.), *Thought Experiments in Science and Philosophy*, Savage: Rowman & Littlefield, 65-84.
- Alanen, L. and Knuuttila, S. (1988), ‘The Foundations of Modality and Conceivability in Descartes and His Predecessors’ in S. Knuuttila (ed.), *Modern Modalities. Studies of the History of Modal Theories from Medieval Nominalism to Logical Positivism* (Synthese Historical Library 33), Dordrecht: Kluwer, 1-69.
- Albert the Great (1890), *Liber I Priorum Analyticorum* in *Opera omnia I*, ed. A Borgnet, Paris:

Vivès.

- Ammonius (1897), *In Aristotelis De interpretatione commentarius*, ed. A. Busse, Berlin: Reimer.
- Bäck, A. (1992), ‘Avicenna's Conception of the Modalities’, *Vivarium* 30, 217-255.
- Becker, O. (1952), *Untersuchungen über den Modalkalkül*, Meisenheim am Glan: Anton Hain.
- Bobzien, S. (1986), *Die stoische Modallogik*, Würzburg: Königshausen & Neumann.
- Bobzien, S. (1993), ‘Chrysippus's Modal Logic and Its Relation to Philo and Diodorus’ in K. Döring and T. Ebert (eds.), *Dialektiker und Stoiker*, Stuttgart: Franz Steiner, 63-84.
- Boethius, A.M.S. (1877-80), *Commentarii in librum Aristotelis Perihermeneias I-II*, ed. C. Meiser, Leipzig: Teubner.
- Boh, I. (1993), *Epistemic Logic in the Later Middle Ages*, London: Routledge.
- Broadie, A. (1993), *Introduction to Medieval Logic*, 2nd edition, Oxford: Clarendon Press.
- >Buridan, John (1976) *Tractatus de consequentiis*, ed. H. Hubien, Louvain: Publications Universitaires.
- Coombs, J.C. (1990), *The Truth and Falsity of Modal Propositions in Renaissance Nominalism*, Ph. D. diss., University of Texas at Austin.
- Craig, W.L. (1988), *The Problem of Divine Foreknowledge and Future Contingents from Aristotle to Suárez*, Leiden: Brill.
- Damian, Peter (1972), *De divina omnipotentia* (Sources chrétiennes 191), ed. A. Cantin, Paris: Cerf.
- De Rijk, L.M. (1962-67), *Logica modernorum I-II.1-2*, Assen: van Gorcum.
- Duns Scotus, John (1950-), *Opera omnia*, studio et cura Commissionis Scotisticae, Vatican City: Vatican Press.
- Ebbesen, S. (1981), ‘Analyzing Syllogisms or Anonymus Aurelianensis III - the (presumably) Earliest Extant Latin commentary on the Prior Analytics, and its Greek Model’, *Cahiers de l'Institut Moyen-Âge Grec et Latin, Université de Copenhague* 37, 1-20.
- Engberg-Pedersen, T. (1990), *The Stoic Theory of Oikeiosis*, Aarhus: Aarhus University Press.
- Freddoso, A.J. (1988), *Louis de Molina: On Divine Foreknowledge, Part IV of the Concordia* (translation with introduction and texts), Ithaca: Cornell University Press.
- Gaskin, R. (1995), *The Sea Battle and the Master Argument*, Berlin: de Gruyter.
- Hintikka, J. (1973), *Time and Necessity: Studies in Aristotle's Theory of Modality*, Oxford: Oxford University Press.
- Hintikka, J., in collaboration with Remes, U. and Knuuttila, S. (1977), *Aristotle on Modality and Determinism* (Acta Philosophica Fennica 29, 1), Amsterdam: North-Holland.
- Holopainen, T. (1996), *Dialectic and Theology in the Eleventh Century*, Leiden: Brill.
- Honnefelder, L. (1990), *Scientia transcendens. Die formale Bestimmung der Seiendheit und Realität in der Metaphysik des Mittelalters und der Neuzeit*, Hamburg: Felix Meiner.
- Honnefelder, L., Wood, R., Dreyer M. (eds., 1996), *John Duns Scotus: Metaphysics and Ethics*, Brill: Leiden.
- Hughes, G.E. (1989), ‘The Modal Logic of John Buridan’, in *Atti del Convegno internazionale di storia della logica: le teorie delle modalità*, Bologna: CLUEB, 93-111.
- Jacobi, K. (1980), *Die Modalbegriffe in den logischen Schriften des Wilhelm von Shyreswood und in anderen Kompendien des 12. und 13. Jahrhunderts: Funktionsbestimmung und Gebrauch in der logischen Analyse*, Cologne: Brill.

- Kilwardby, Robert (1968), *In libros Priorum Analyticorum expositio*, Venice 1516 (under the name Aegidius Romanus), reprint Frankfurt am Main: Minerva.
- King, P. (1985), *Jean Buridan's Logic. The Treatise on Supposition. The Treatise on Consequences* (translation with introduction and commentary), (Synthese Historical Library 27), Dordrecht: Reidel.
- Knuuttila, S. (1993), *Modalities in Medieval Philosophy*, London, New York: Routledge.
- Knuuttila, S. (1996), 'Duns Scotus and the Foundations of Logical Modalities' in L. Honnefelder, R. Wood, M. Dreyer (eds.), 127-143.
- Kretzmann, N. (1985), 'Nos ipsi principia sumus: Boethius and the Basis of Contingency' in T. Rudavsky (ed.), *Divine Omniscience and Omnipotency in Medieval Philosophy* (Synthese Historical Library 25), Dordrecht: Reidel 23-50.
- Kretzmann, N. (1987), 'Boethius and the Truth about Tomorrow's Sea Battle' in L.M. de Rijk and H.A.G. Braakhuis (eds.), *Logos and Pragma. Essays on the Philosophy of Language in Honour of Professor Gabriel Nuchelmans* (Artistarium Supplementa 3), Nijmegen: Ingenium Publishers, 63-97.
- Lagerlund, H. (1999), *Modal Syllogistics in the Middle Ages*, Ph. D. diss., Uppsala University 1999.
- Lewis, N.T. (1987), 'Determinate Truth in Abelard', *Vivarium* 25, 81-109.
- Lewis, N.T. (1996), 'Power and Contingency in Robert Grosseteste and Duns Scotus' in L. Honnefelder, R. Wood, M. Dreyer (eds.), 205-225.
- Lovejoy, A. (1936), *The Great Chain of Being: A Study of the History of an Idea*, Cambridge, Mass.: Harvard University Press.
- Maierù, A. (1972), *Terminologia logica della tarda scolastica*, Rome: Edizioni dell' Ateneo.
- Manekin, C.H. (1992), *The Logic of Gersonides* (translation with introduction and commentary), (The New Synthese Historical Library 40), Dordrecht: Kluwer 1992.
- Marrone, S.P. (1996), 'Revisiting Duns Scotus and Henry of Ghent on Modality' in L. Honnefelder, R. Wood & M. Dreyer (eds.), 175-189.
- Martin, C.J. (1987), 'Embarrassing Arguments and Surprising Conclusions in the Development of Theories of the Conditional in the Twelfth Century' in J. Jolivet and A. de Libera (eds.), *Gilbert de Poitiers et ses contemporains: aux origines de la* Logica Modernorum, Naples: Bibliopolis, 377-400.
- Mignucci, M. (1989), 'Truth and Modality in Late Antiquity: Boethius on Future Contingent Propositions', G. Corsi, C. Mangione and M. Mugnai (eds.), in *Atti del Convegno internazionale di storia della logica: le teorie delle modalità*, Bologna: CLUEB, 47-78.
- Normore, C. (1982), 'Future Contingents' in N. Kretzmann, A. Kenny, J. Pinborg (eds.), *The Cambridge History of Later Medieval Philosophy*, Cambridge: Cambridge University Press, 358-81.
- Normore, C. (1991), 'Descartes's Possibilities' in G.J.D. Moyal (ed.), *René Descartes: Critical Assessments*, vol. III, London, Routledge, 68-83.
- Normore, C. (1993), 'Petrus Aureoli and His Contemporaries on Future Contingents and Excluded Middle', *Synthese* 96, 83-92.
- Normore, C. (1996), 'Scotus, Modality, Instants of Nature, and the Contingency of the Present' in L. Honnefelder, R. Wood, M. Dreyer (eds.), 161-174.

- Nortmann, U. (1996), *Modale Syllogismen, mögliche Welten, Essentialismus. Eine Analyse der Aristotelischen Modallogik*, Berlin: de Gruyter.
- Ockham, William (1974), *Summa logicae*, ed. Ph. Boehner, G. Gál, S. Brown, St Bonaventure: Franciscan Institute of St Bonaventure University. (English translation of part I by M.J. Loux, Notre Dame University Press 1974, part II by A.J. Freddoso and H. Schuurman, Notre Dame University Press 1980.)
- Patterson, R. (1995), *Aristotle's Modal Logic: Essence and Entailment in the Organon*, Cambridge: Cambridge University Press.
- Rescher, N. (1974), 'The Theory of Modal Syllogistic in Medieval Arabic Philosophy', in N. Rescher et al., *Studies in Modality* (American Philosophical Quarterly Monograph Series 8), Oxford: Blackwell, 17-56.
- Roncaglia, G. (1996), *Palestra rationis. Discussioni su natura della copula e modalità nella filosofia 'scolastica' tedesca del XVII secolo*, Florence: Leo S. Olschki.
- Seel, G. (1982), *Die Aristotelische Modaltheorie*, Berlin: Walter de Gruyter.
- Sharples, R.W. (1983), *Alexander of Aphrodisias' On Fate* (text, translation and commentary), London: Duckworth.
- Sorabji, R. (1980), *Necessity, Cause, and Blame: Perspectives on Aristotle's Theory*, Ithaca, NY: Cornell University Press.
- Thom, P. (1996), *The Logic of Essentialism. An Interpretation of Aristotle's Modal Syllogistic* (The New Synthese Historical Library 43), Dordrecht: Kluwer.
- van Rijen, J. (1989), *Aspects of Aristotle's Logic of Modalities* (Synthese Historical Library 35), Dordrecht: Kluwer.
- von Wright, G.H. (1984), Truth, Knowledge, and Modality, Oxford: Blackwell.
- Vos, A., Veldhuis, H., Looman-Graanskamp, A.H., Dekker, E., den Bok, N.W.(1994), *John Duns Scotus: Contingency and Freedom. Lectura I 39* (introduction, translation and commentary), Dordrecht: Kluwer.
- Vuillemin, J. (1984), *Nécessité ou contingence: l'aporie de Diodore et les systèmes philosophiques*, Paris: Minuit.
- Waterlow, S. (1982a), *Passage and Possibility: A Study of Aristotle's Modal Concepts*, Oxford: Clarendon Press.
- Waterlow, S. (1982b) *Nature, Change, and Agency in Aristotle's* Physics, Oxford: Clarendon Press.
- Weidemann, H.(1986), 'Aristoteles und das Problem des kausalen Determinismus (*Met*. E 3)', *Phronesis* 31, 27-50.
- White, M.J. (1985), *Agency and Integrality. Philosophical Themes in the Ancient Discussions of Determinism and Responsibility* (Philosophical Studies Series in Philosophy 32), Dordrecht: Reidel.
- Yrjönsuuri, M. (1994), *Obligationes: Fourteenth Century Logic of Disputational Duties* (Acta Philosophica Fennica 55), Helsinki: The Philosophical Society of Finland.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle | Boethius, Anicius Manlius Severinus | Buridan, John [Jean] | Duns Scotus, John | Kilwardby, Robert | logic: modal | medieval philosophy | Ockham [Occam], William | possible worlds | Stoicism | truth: necessary vs. contingent

Simo Knuuttila
simo.knuuttila@helsinki.fi

First published: June 30, 1999
Content last modified: June 30, 1999

John Buridan

Although he was one of the most famous and influential philosophers of the later Middle Ages, John Buridan is today among the least well known. He spent his entire career as a master in the arts faculty at the University of Paris, lecturing on logic and the works of Aristotle, and producing many commentaries and independent treatises on logic, metaphysics, natural philosophy, and ethics. His logic textbook, the *Summulae de dialectica* (Compendium of Dialectic), is a work of astonishing breadth and originality which redeems the older medieval tradition of Aristotelian logic using the newer, terminist logic of ‘moderns’ such as Peter of Spain and William of Ockham. Buridan applied these analytical techniques so successfully in his metaphysics, physics, and ethics that, for many of his successors, they came to be identified with the very method of philosophy, understood as a *secular* practice.

1. Life

John Buridan was born sometime before 1300 at or near the town of Béthune in Picardy, France. He was educated in Paris, first at the Collège Lemoine, where he was awarded a benefice or stipend for needy students, and then at the University of Paris, where he received his Master of Arts degree and formal license to teach by the mid-1320s. He enjoyed a long and illustrious career as an arts master at Paris, serving twice (in 1328 and 1340) as university rector and supporting himself with numerous benefices. He last appears alive and well in a document of 1358, which mentions him adjudicating a territorial dispute between the English and Picard nations (the Parisian student body at the time was organized

according to one's place of origin). He must have died shortly thereafter because in 1361, one of his benefices was awarded to another person.[1]

Such is the historical record -- a handful of relatively minor details.[2] Buridan's fame as a teacher and philosopher, however, quickly turned his life into the stuff of legend. There are stories that he met his end when the King of France had him thrown into the Seine River in a sack because of a scandalous affair with the Queen, that he went on to found the University of Vienna after being expelled from Paris for his nominalist teachings, and even that he hit the future Pope Clement VI over the head with a shoe while competing for the affections of the wife of a German shoemaker (the blow apparently caused the prodigious memory for which Clement became known). But none of these legends can be independently verified, and most are inconsistent with what we do know about him.[3] Nevertheless, they illustrate what the French scholar Edmond Faral called the "*bruits de ville*" or 'buzz' surrounding Buridan's name in Parisian circles, which continued for some time after his death.[4]

Buridan's academic career was unusual in two respects, both of which help to explain his philosophy. First, he spent his entire career in the faculty of arts, without ever moving on to study for a doctoral degree in one of the higher faculties of law, medicine, or theology, which would have been the more typical academic career path of the period. Most of the figures we think of as medieval philosophers were trained as theologians, including Thomas Aquinas, Duns Scotus, and even William of Ockham (although he did not finish his degree). Since university statutes forbade arts masters from teaching or writing about theology, Buridan produced no theological works and no commentary on the *Sentences* of Peter Lombard, one of the principal genres of philosophical writing in the 14th century. Why did he remain in the faculty of arts? It is unlikely that someone of his talents would have spent his whole life teaching without being recognized. Nor can the reason be lack of means, for despite his impoverished background he was from the moment he arrived in Paris a magnet for bursaries and stipends, and is even named in a document of 1349 as being among those masters capable of supporting themselves without financial assistance from the University. The only plausible answer is that he chose to remain one of the 'artists [*artistae*]'. We can only speculate about his reasons, but it has been suggested that Buridan was committed to a vision of philosophy as a secular enterprise beginning from what is evident to the senses and intellect, as opposed to theology, which begins from non-evident truths revealed in scripture and doctrine.[5]

Buridan is also different in that he remained for his entire life a secular cleric rather than joining a religious order such as the Dominicans or Franciscans. A papal letter of 1329 refers to him as simply, "*clericus Atrebatensis diocoesis, magister in artibus* [a cleric from the Diocese of Arras and Master of Arts]".[6] By contrast, the best-known names of the period all had religious and intellectual affiliations beyond the university: Thomas Aquinas (Dominican), Duns Scotus (Franciscan), William of Ockham (Franciscan), Gregory of Rimini (Augustinian), to name a few. Among other things, this freed Buridan from the obligation to enter into the ongoing doctrinal disputes that arose between religious orders, or between an order and the church hierarchy -- such as the dispute between the papacy and the Franciscans over apostolic poverty. Thus, whereas Ockham spent his entire later career on various political crusades, away from philosophy, and was finally excommunicated for his troubles, Buridan was bound only by the

intellectual and pedagogical traditions of his university. His confessional independence also meant that he could help himself to philosophical insights from a variety of sources. This emerges in the sometimes eclectic character of his work, as we shall see below.

2. Writings

Most of Buridan's works are in the form of commentaries on Aristotle. He wrote both *expositiones* (expositions), or literal commentaries which feature detailed, line-by-line explanations of the meaning of Aristotle's remarks, and *quaestiones* (questions), or longer, critical studies of the philosophical issues raised by them, usually centered on a specific lemma from the text. Both genres originated in the classroom, a fact that becomes clear in the number of references to student queries and student concerns which survive in the written versions. Buridan often lectured more than once on the same text over the course of his career, with the result that we sometimes have different written versions of his commentary on the same work. For example, there are three versions of his *Quaestiones* on Aristotle's *De anima*, the last of which identifies itself as the "third or final lecture [*tertia sive ultima lectura*]". Where there are multiple versions of the same commentary, their relationship is generally one of increasing length and sophistication over time.

Buridan commented on virtually all of the major works of Aristotle. In addition to the entire *Organon*, there are commentaries on Aristotle's *Physics, On the Heavens, On Generation and Corruption, De Anima*, *Parva Naturalia*, *Metaphysics*, *Nicomachean Ethics*, and *Rhetoric*.[7] He also wrote a number of shorter, independent treatises on current philosophical issues, such as the *Tractatus de relationibus* [Treatise on Relations], *Tractatus de universalibus* [Treatise on Universals], *Tractatus de consequentiis* [Treatise on Consequences], and *Quaestio de puncto* [Question on <the Nature of> Points]. He was a very prolific author.

But Buridan's masterwork is the *Summulae de dialectica* [Compendium of Dialectic], a comprehensive logic textbook that started out as a commentary on the *Summulae logicales* or logical compendium of the 13th-century dialectician Peter of Spain,[8] but soon evolved into an independent work of astonishing breadth and originality. In it, Buridan redeems the older medieval tradition of Aristotelian logic through the *via moderna* [modern way] -- the newer, terminist logic that had gradually replaced it. Because the work was accessible to master and student alike, it became extremely popular at Paris and in the newly founded universities of Heidelberg, Prague, and Vienna.

Buridan's other works were almost as widely read as the *Summulae*. Handwritten copies and early printed editions were carried by his students and followers throughout Europe, where they were often used as primary texts in university courses on logic and Aristotelian philosophy. The *via Buridani* continued to shape European thought well into the Renaissance.

Like other medieval philosophers, Buridan has not been fully appreciated because of the lack of modern editions and translations of his work (see bibliography). The situation has improved recently with the

appearance of Gyula Klima's mammoth translation of the entire *Summulae* from the (now almost complete) Latin critical edition of the text. But knowledge of Latin and the ability to read medieval manuscripts are still essential if one wishes to study Buridan's thought first-hand.

3. Language

In the medieval university, arts masters provided students with their basic education in grammar, logic or dialectic, and Aristotelian philosophy, subjects which together embodied the medieval conceptions of literacy and wisdom. What they taught was the language of rational inquiry, without which further study in law, medicine, or theology would have been impossible. Students were required to learn, in ascending order, the exposition and interpretation of authoritative texts (grammar), the structure and modes of reasoning in conventional discourse (logic), and finally, the analysis and systematic disclosure of the order of nature (Aristotelian philosophy). Although he did not write grammatical treatises,[9] Buridan asserts in the very first section of the *Summulae* that "positive grammar [*grammatica positiva*] has to be learned first, by means of which the master is able to communicate with the disciple, whether it be in Latin, French, Greek, or Hebrew, or whatever else" (S 1.1.1: 6). The disciple's knowledge of logic and Aristotelian philosophy is built upon this grammatical foundation.

The importance of language in Buridan's philosophy emerges on many levels, all of which are driven by pedagogy. In logic, grammatical rules are explicitly subsumed as necessary conditions of the science (*scientia*) of logic, so that although the logician's notion of truth and the grammarian's notion of congruence are separable in theory, in practice the complete significance of a piece of discourse cannot be determined without both. Context is crucial to interpretation:

> We should also note that some might ask whether it is the composite or the divided sense that is properly expressed by 'Every man or (a) donkey runs', that is to say, whether only the term 'man' or the whole subject is distributed. And I say that we have to respond differently, in accordance with the different manners of speaking and writing. For if immediately after 'man' there is a sign of division, namely, a pause or a period, then the proposition will be called divided, and only 'man' will be distributed, but if not, then it will be called 'composite', and the whole subject will be distributed. (S 4.2.6: 250; cf. S 9.4, 15th sophism: 912-13).

Accordingly, logic is not about some conceptually ideal or canonical language but the practical art of interpreting discourse:

> an utterance [*vox*] does not have any proper import [*virtus propria*] in signifying and suppositing, except from ourselves. So by an agreement of the disputing parties, as in obligational disputes, we can impose on it a new signification and not use it according to its common signification. We can also speak figuratively [*transsumptive*] and ironically, according to a different signification. But we call a locution 'proper' when we use it according to the signification commonly and principally given to it, and we call a locution

> ‘improper’ when we use it otherwise, although we can legitimately use it otherwise. So it is absurd to say that a proposition of an author is false, absolutely speaking, if he puts it forth incorporating an improper locution, according to which it is true. Instead, we ought to say that it is true, since it is enunciated according to the sense in which it is true … So it absolutely seems to me that wherever it is evident that an author puts forward a proposition in a true sense, although not as a proper locution, then to deny that proposition without qualification would be cantankerous and insolent [*dyscolum et protervum*]. But to avoid error, it should be properly pointed out that the proposition is not true in the proper sense, or by virtue of its proper meaning, and then it has to be shown in which sense it is true. (S 4.3.2: 256; cf. QIP 5: 144-145, ll. 800-829)

In Buridan's view, the logician cannot expound the meaning of a proposition without carefully attending to its internal features, i.e., the sense of the particular locutions incorporated in it, as well as to its external features, i.e., the discourse conditions which surround it.

This willingness to take human language as it is found, with all of its ambiguities and rough edges, marks an important difference between Buridan and Ockham, the 14th-century philosopher with whom he is most often compared. Both authors make the traditional assumption that propositions, be they spoken, written, or mental, are the bearers of truth and falsity. Ockham, however, tends to see mental propositions as logically ideal or, in modern parlance, ‘canonical’.[10] The problem with spoken and written propositions is that because they depend on the meaning conventions of fallible users, they fail to be universal and logically perspicuous. But these shortcomings can be filtered out metalinguistically once we realize that the meanings of their constituent terms depend on their corresponding mental concepts, which naturally signify the same for everyone. So in Ockham's logic, the semantic relation between these concepts and what they signify outside the mind is of paramount concern; spoken and written terms have semantic properties too, of course, but in an entirely derivative way. By contrast, Buridan never privileges conceptual discourse or suggests that the logician might use it to systematically reform spoken or written language. He holds that spoken and written utterances -- sometimes he uses the term ‘utterance [*vox*]’ where Ockham has ‘term [*terminus*]’ -- signify concepts primarily: “the capability of speaking was given to us in order that we could signify our concepts to others and also the capacity of hearing was given to us in order that the concepts of speakers could be signified to us” (S 4.1.2: 222). Accordingly, “utterances are imposed to signify things only through the mediation of the concepts by which those things are conceived” (QC 1: 4, ll. 45-6). Concepts are just the medium of signification for Buridan, the cognitive or psychological aspect of the signification of a word.

This difference helps to explain why Buridan uses paradoxes of self-reference to test the functionality of his logic, whereas Ockham avoids them by claiming that a term -- or at least a term in those circumstances -- cannot refer to itself.[11] Buridan takes seriously the fact that people can and do utter self-referential propositions, and thinks the logician should say what is going on when they do. This is really a difference of perspective. As Joël Biard has pointed out, we can divide medieval logicians into those who try to restrict the possibilities of human discourse in the direction of what is logically ideal, and those who are willing to accept a proposition because it is grammatical and because the person who

utters it intends to signify something by it. Ockham, William of Sherwood, and Walter Burley belong to the former group; Buridan and Thomas Bradwardine to the latter.[12] It also explains why Buridan tells us that it is not possible to analyze contradictory propositions unless they "have the same subject and predicate in utterance and also in intention" (S 9.7, 2nd sophism: 943), and his reminding us that, when testing a proposition for contradictoriness, "it is [sometimes] necessary to add other utterances when contradicting it…. For one should primarily attend to the intention, for we use words only to express the intention" (S 9.8, 11th sophism: 979). The logician must above all be a skilled interpreter of human discourse.

For Buridan, linguistic confusion is the source of many of the traditional problems of metaphysics and natural philosophy. His approach is broadly nominalistic, but Buridan's nominalism is more of a parsimonious way of doing philosophy than a doctrine about universals. For example, when a cause is understood as being actual rather than merely potential, does our conception of it *qua* cause change in any way? Aristotle leaves this ambiguous in *Metaphysics* V.2, but some medieval philosophers thought it necessary to posit an additional state of affairs to explain the dynamic aspect of causality, i.e., the fact that a contingent state of affairs needs to be brought about by some agent. Thus, if we think of God as the cause of Socrates, there must be something else, God's-being-the-cause-of-Socrates (*deum esse causam Sortis*), distinct from both God and Socrates, to account for his existence. This 'something else' then becomes not only what is signified by the proposition 'God is the cause of Socrates' (*complexe significabile*), but also the proper object of our knowledge that God is the cause of Socrates. Buridan replies by arguing that philosophers who think this way do not know how to interpret human language. They take everything too literally. But we should not be misled by what a proposition literally says, or seems to say, into thinking that there must be some new kind of entity corresponding to God's being the cause of Socrates, especially since such reifying moves do not help us to understand what is happening when a cause acts (QM V.7-8: 30va-33ra).

The same sensitivity to questions of interpretation is evident in his treatment of propositions in natural philosophy, where he argues that we can meaningfully use propositions containing terms such as 'infinite [*infinitum*]' and 'point [*punctum*]' to express actual states of affairs without committing ourselves to the existence of infinite magnitudes or indivisible points. The key is to realize that propositions do not always wear their meanings on their sleeves. Their component terms can be differently construed (e.g., categorematically or syncategorematically), and can even perform different modal functions, which, depending on how the proposition is read (e.g., in a composite or divided sense) can change its truth conditions.[13] There will be ambiguities, of course, but Buridan is steadfast in his assumption that a rational person will always be able to sort them out, given the right dialectical tools.

4. Logic

Following Aristotle and Porphyry, Buridan's logic is based on two distinct, but complementary, conceptions of its purpose: theoretical or pedagogical (*logica docens*) and practical (*logica utens*).[14] The former, he says, is so called because "it teaches [*docet*] us how, and from what [materials], arguments should be constructed, whether those arguments be demonstrative, dialectical, or of some

other kind". The latter takes its name from the fact that "it uses [*utitur*] arguments in order to prove whether some conclusion is evident, regardless of the subject matter of the conclusion" (QIP 1: 126-7, ll.176-80). But since the teaching of logic is ordained to its use, Buridan holds that logic is ultimately a practical rather than a speculative discipline.

Historians of logic usually classify Buridan as one of the terminists or 'moderns', a diverse group of 13th and 14th-century logicians who regarded the semantic properties of terms (literally, the 'ends [*termini*]', or subjects and predicates, of propositions) as the primary unit of logical analysis. As we saw above, in addition to commenting on Aristotle's *Organon*, he wrote a logical compendium, the *Summulae de dialectica*, ostensibly as a commentary on Peter of Spain's *Summulae logicales*, an influential terminist textbook written a century earlier. But Buridan's *Summulae* was essentially a new work, more that ten times longer than the original and featuring many new and completely rewritten sections. Buridan moves his students and readers through an orderly progression of teachings, beginning with propositions (Treatise I), shifting down to the signification and referential function of their component terms (II-IV), then back up to terms and propositions insofar as they figure in more complex patterns of reasoning: syllogisms (V), topics (VI), fallacies (VII), and finally, demonstrations (VIII). The work concludes with a kind of exercise book on paradoxical and otherwise puzzling propositions, showing how they can be resolved using the techniques of the previous eight treatises.

No encyclopedia article can do justice to the richness and variety of Buridan's logic, but there are several moments from the text that illustrate how he practiced the art of dialectic.

First, Buridan did much to streamline and better articulate the methods of terminist logic. The most important analytical tool in the *Summulae* is the doctrine of *suppositio* [supposition], which had been a feature of terminist logic for several generations by the time Buridan arrived in Paris. Terms were thought to possess two general semantic properties: signification, which refers to what a term 'makes known' in the mind of the person who sees it or hears it or conceives of it, whether mediately or immediately (thus, the written term 'Socrates' brings to mind the concept of Socrates, which in turn signifies the actual person); and supposition, which refers to the capacity of certain substantive terms to stand for or 'pick out' something in a particular context, such as in a proposition. Supposition roughly corresponds to what we would call the theory of reference.[15] Traditional accounts divided supposition into proper supposition, where a term is used with its typical or standard meaning, and improper supposition, where a term is used in some metaphorical or figurative sense. Most logicians went on to distinguish three kinds of proper supposition: personal, where a term stands for what it signifies (e.g., 'Socrates' in 'Socrates is a man'); material, where it stands for itself (e.g., 'man' in 'Man has three letters'); and simple, where it stands for a common nature or concept (e.g., 'man' in 'Man is a species'). Simple supposition appears to have been a vestige of early terminist logic,[16] whose realist practitioners needed to distinguish between referring to a universal thing and referring to a particular thing. But by the 14th century, the whole notion of universals had become more controversial, and nominalist logicians in particular were not about to accept any special device for referring to them through common terms such as 'man'. So simple supposition was readapted to model reference to common concepts or intentions. Thus, Ockham holds that a term exhibits simple supposition when it supposits "for a concept in the mind

[*pro conceptu mentis*]," and is not being used significatively.[17] What united Ockham and his terminist predecessors was their realization that if the proposition 'Man is a species' is to be true, then the term 'man' cannot supposit personally for any of the individual men it ultimately signifies, since it cannot be said of any of them that he is a species (Socrates is a man, not a species). Accordingly, the reference of 'man' must be to a common nature or concept.

But Buridan contends that there are only two ways a term can stand for something in a proposition, personally and materially:

> Of the first [section on the divisions of supposition], we should realize that some people have posited also a third member, which they call 'simple supposition'. For they [e.g., Peter of Spain] held that universal natures are distinct from the singulars outside of the soul. And so they said that a term supposits personally when it supposits for the singulars themselves, that it supposits simply when it supposits for that material nature, and materially when it supposits for itself. But I hold that Aristotle correctly refuted that opinion in the seventh book of the *Metaphysics* [VII.3.1038b1-1039a23] and so this kind of supposition has to be eliminated, at least according to this interpretation. In another manner, others [e.g., Ockham] call supposition 'simple' when an utterance supposits for the concept according to which it is imposed and material when it supposits for itself or another similar to itself. And this can be permitted, but I do not care [about this usage], for I call both 'material supposition'. (S 4.3.2: 253)

Buridan sees that it is misleading to assign a special logical sense to terms being used to refer to themselves or to the concepts they express, as if this were any different from figurative or metaphorical uses, since only terms that refer to things existing *per se* are being used in their proper sense. Thus, terms can stand either for the things they ordinarily signify, in which case they supposit personally, or for something else, in which case they supposit materially. Material supposition applies whenever a term is used in some way that departs from the meaning imposed upon it by the linguistic community:

> an utterance does not have any proper import [*virtus propria*] in signifying and suppositing, except from ourselves. So by an agreement of the disputing parties, as in obligational disputes, we can impose on it a new signification, and not use it according to its common signification. We can also speak figuratively [*transsumptive*] and ironically, according to a different signification. But we call a locution 'proper' when we use it according to the signification properly and principally given to it, and we call a locution 'improper' when we use it otherwise, although we legitimately can use it otherwise. So it is absurd to say that a proposition of an author is false, absolutely speaking, if he puts it forth incorporating an improper locution, according to which it is true. Instead, we ought to say that it is true, since it is enunciated according to the sense in which it is true. (S 4.3.2: 256)

Notice that the default interpretation of a term is its proper sense, defined as "the signification properly

and principally given to it". The proper signification of a term must be based on the fact that "utterances were primarily and principally imposed to signify so as to stand for their ultimate *significata*, and not for themselves" (S 4.3.2: 256);[18] that is to say, just as concepts (at least in the first instance) naturally signify those extra-mental things which just as naturally give rise to them, so spoken and written terms (at least in the first instance) are imposed to signify, via their corresponding concepts, the same ultimate *significata*.[19] For Buridan, this capacity is part of our nature as cognitive creatures. It is why he insists that determining the nature of concepts pertains not to logic but to psychology or metaphysics, speculative sciences whose conclusions cannot be otherwise (QDI I.3:6, ll. 4-10).[20] So even if we say that the proposition, 'Man is a species' is true insofar as it is put forward in the context of Aristotle's *Categories*, it is not literally true, or true according to "the signification properly and principally given to it", because of no *per se* existing man is it true to say that he is a species. What happens is that when we work on the *Categories*, we follow Aristotle's lead and depart from conventional usage in such a way that the term 'man' supposits not for individual men but for the universal concept according to which it was imposed to signify, in which case the proposition is true because "species and genera are universals according to predication" (S 4.3.2: 254). So it can be true that man is a species without it being *literally* true that man is a species.

The second way in which Buridan changed the dialectical landscape was to extend the range of traditional logic. There are numerous examples of this, not all of them uncontroversial because it is often hard to tell what Buridan intended to achieve by a given innovation (this goes for other medieval logicians as well).[21] But a fairly uncontroversial example can be found in his use of supposition to examine the structure of certain complex terms that would remain unanalyzed on the traditional account of syllogistic inferences. Of particular importance here is the doctrine of ampliation. Thus, although the syllogism, 'Nothing dead is an animal, some man is dead; therefore, some man is not an animal', is an acceptable fourth-mode syllogism of the first figure (*Ferio*), Buridan denies that the consequence is "formally valid". The reason is that in this syllogism, 'man' is an ampliative term, and "from an ampliated nondistributed term the same term does not follow nonampliated" that is, "in the minor proposition the term 'man' was ampliated to past [things], whereas in the conclusion it was not ampliated," making the premises true but the conclusion false (S 5.3.2: 326; cf. QAnPr I.14). Similarly, terms referring to the divine persons sometimes generate counterexamples to the traditionally accepted modes. Thus, "the following syllogism in *Barbara* is invalid: 'Every God is the son, every divine Father is God; therefore every divine Father is the Son", for the transitivity of identity fails in cases "where the most simple unity is a trinity of really distinct persons" (S 5.3.2: 327). Buridan also urges the reader to be wary of modal contexts introduced by verbs of knowing and believing because "the verb 'know' ampliates the subject to supposit not only for present things, but also for future and past ones". This means that without suitable qualification, "although I know that every man is an animal, nevertheless, it does not follow that every man is known by me to be an animal; for then it would follow that every man, whether alive or dead or yet to be born, would be known by me to be an animal, which is false" (S 5.6.8: 348). What is noteworthy in these and other examples is Buridan's use of the doctrine of supposition to extend the range of 'truth-makers' for modal inferences, e.g., in his assumption that "the presence of a modal copula -- *any* modal copula -- in a proposition ampliates the subject to stand for not only the actual things but also the possible things that fall under that term".[22] Because it makes merely possible objects relevant to the evaluation of modal inferences, ampliation can be seen as a kind of Buridanian equivalent

of possible worlds semantics, though it would be a mistake to regard it as a remarkable anticipation of that 20th-century theory. Buridan's remarks on its theoretical significance are few,[23] and, despite the degree of technical sophistication involved, he probably did not see it as a radical innovation. He probably understood it as part of his ongoing effort to make existing schemes for checking inferences more practicable.

Third, and finally, Buridan made major contributions to certain forms of logical inquiry that originated in the medieval period. Most modern logicians know of his solutions to alethic paradoxes such as the Liar, which are addressed in the eighth and final chapter of ninth treatise of the *Summulae*, which belongs to the medieval literature of *sophismata* or *insolubilia*.[24] The 7th sophism Buridan considers is ‘Every proposition is false’. The case posits “that all true propositions are annihilated while the false ones remain [in existence], and then Socrates propounds only this [proposition]: ‘Every proposition is false’” (S 9.8, 7th sophism: 965). The question is then asked whether Socrates's proposition is true or false. The arguments on each side of the question illustrate the difficulties one faces if ‘true’ and ‘false’ are interpreted strictly. The argument that it is false assumes that “it is impossible for the same proposition to be true and false when propounded in the same language and understood in the same way by everyone hearing it”, and proceeds to argue that the sophism is false because any proposition which entails its own contradictory is impossible, and therefore false. The opposite side begins by focusing on the logical form of the sophism as a universal affirmative that has no counter-instance in the case at hand, which stated that all true propositions have been annihilated with only the false ones remaining. Second, the sophism must be true because the subject and predicate terms supposit for the same things: if every proposition is false, then each and every propositional significate of the term ‘every proposition’ must be false, as it indeed is, according to the case. Finally, the sophism must be true because “it signifies only that every proposition is false; and this is how things are [*ita est*]”, according to the case (S 9.8, 7th sophism: 965).

Buridan writes as if this particular sophism enjoyed some notoriety among logic teachers at Paris, although all but one of the alternative solutions he mentions were discussed and criticized from the very beginning of the *insolubilia* literature. These involve various *ad hoc* proposals that either build new assumptions into the case or else make up new rules about how the terms of the sophism are to be interpreted.[25] Into the first category falls a solution known as the ‘*transcasus*’, which involves the bizarre suggestion that the time Socrates utters his proposition and the time referred to by the verb of the proposition are not the same. This would allow us to say that if there are no true propositions during the first hour of a certain day, Socrates could utter his proposition at the end of this hour and it would be true, where he is understood as referring “not to the time at which he speaks but to the time of that first hour”. But this is of no help if we stick to the case and assume that the times are the same. Alternatively, in a solution advocated by the ‘*restringentes* [restrictors]’ -- so-called because they avoided self-reference by restricting what a term can supposit for -- we could make the proposition non-reflexive by stipulating that “terms that are apt to supposit for propositions are not put in propositions to supposit for the propositions in which they are put, but for others”. But Buridan rightly rejects this second strategy as failing to take seriously our conventional understanding of terms, for when one uses the term ‘proposition’, he says, “one understands indifferently all propositions, indeed, present, past, and future ones, his own as well as those of another person”. A moment's reflection should make it obvious that

“this solution is worth nothing: for what one understands, of that he can speak [*quod aliquis intelligit, de hoc potest loqui*]” (S 9.8, 7th sophism: 966).[26]

Buridan's quick answer to the sophism is that Socrates's proposition is false in the case at hand. But before moving on to his final answer, he first discusses a solution described as being held by some people, including himself.[27] This is that there is another condition, in addition to the requirement that its terms stand for the same thing or things, which a proposition must meet if it is to be true. A proposition must also signify or assert itself to be true (S 9.8, 7th sophism: 967).[28] In her detailed analysis of this sophism, Fabienne Pironet has shown that the text in which Buridan defends this earlier view is his question commentary on Aristotle's *Posterior Analytics*, where it is expressed in terms of the traditional formula that “howsoever [a proposition] signifies, so it is [*qualitercumque significat, ita est*]” (QAnPo I.10).[29] Now Buridan holds that all propositions satisfy this condition trivially: “every proposition by its form signifies or asserts itself to be true” (S 9.8, 7th sophism: 967). The problem with self-referential paradoxes is that they also seem to signify that they are false. Thus, although the proposition ‘I say what is false [*ego dico falsum*]’ “signifies itself to be true in some fashion, nevertheless this is not so entirely, or howsoever it signifies [*licet aliqualiter sic significat, non tamen totaliter vel qualitercumque ita est*]. Therefore it is false” (QAnPo I.10).

Unfortunately, this looks no less *ad hoc* than the *transcasus* and restriction solutions he has just criticized. Why shouldn't other propositions, besides the paradoxical ones, be able to signify that they are false? Buridan does not say in his commentary on the *Posterior Analytics*, and in the *Summulae* he rejects his earlier view for the rather different reason that it is false “that every proposition signifies or asserts itself to be true” (S 9.8, 7th sophism: 968). His argument is not exactly clear, but the problem appears to be semantic: he cannot find an interpretation of the phrase ‘itself to be true [*se esse veram*]’ in the supplementary condition that will permit it to function as a general principle. Consider the proposition ‘A man is an animal [*homo est animal*]’. If we understand it materially, i.e., as standing for a proposition, then it will signify ‘The proposition “A man is an animal” is true’, which is false because it refers to second intentions (concepts or signs by means of which we conceive of other concepts or signs as such), and the original proposition refers to things (human beings and animals), not concepts. But what if we say that a proposition signifies itself to be true if it is taken significatively for the things or first intentions, rather than materially? This will not work either, argues Buridan, because then the affirmative proposition ‘A man is a donkey [*homo est asinus*]’ would signify that a man is a donkey, which is false because the subject term ‘man’ does not supposit for anything (no human beings are donkeys).[30] Accordingly, we cannot base our solution to self-referential paradoxes on the idea that every proposition signifies or asserts itself to be true.[31]

The solution Buridan finally settles on receives the somewhat tepid endorsement of being “closer to the truth” than the previous solution -- a reflection, perhaps, of his awareness of the imperfectability of any formal system that stays close to the fact of human language. Here, the idea that a proposition formally signifies itself to be true is replaced by the notion of implication from the doctrine of consequences. “Every proposition,” he says, “virtually implies another proposition in which the predicate ‘true’ is affirmed of the subject that supposits for [the original proposition]” (S 9.8, 7th sophism: 969).[32] Unlike

the old solution, in which the second proposition is signified by the first and hence part of its meaning, the new solution assumes only that the second proposition follows logically from the first, so that its meaning can be expounded separately. In this way, for the truth of any proposition P, it is required not only (1) that the subject and predicate terms of P stand for the same thing or things,[33] but also (2) that P implies another proposition, ‘P is true’, which must also be true. Otherwise, we would have a true antecedent and a false consequent, violating Buridan's fifth theorem regarding assertoric consequences, which states, “it is impossible for what is false to follow from what is true [*impossibile est ex veris sequi falsum*]” (TC I.8: 34, l. 97). Applying this to the 7th sophism, the constituent terms in the proposition uttered by Socrates -- ‘Every proposition’ and ‘false’ -- stand for the same things, since in the posited case, “all true propositions are annihilated and the false ones remain, and then Socrates propounds only this: ‘Every proposition is false’”. So the first condition is satisfied. But the implied proposition, ‘P is true’ (where P is the name of ‘Every proposition is false’), is false because its constituent terms, ‘Every proposition is false’ and ‘true’, do not stand for the same thing, since *ex hypothesi*, P stands for the antecedent proposition ‘Every proposition is false’, not for things that are true. But this gives us a true antecedent and a false consequent, and so the consequence does not hold. Therefore, the sophism is false.

5. Metaphysics

Buridan's metaphysics is thoroughly informed by his logic. He tries wherever possible to apply the *Summulae*'s analytical techniques to solve problems in speculative philosophy. His approach is critical in that it tends to view traditional questions in metaphysics as based on confusions of logic or language. For the same reason, his solutions are not original in the modern sense of being without precedent, although they have few equals in terms of their elegance and economy of expression. Buridan is a master craftsman of philosophical argumentation.

Like most arts masters in 14th-century Paris, Buridan is careful to dissociate the metaphysical questions he asks as a philosopher from similar questions that might be asked by a theologian. Jurisdictional wrangling forced Parisian masters to be quite clear about how they were approaching a problem,[34] although what we now identify as medieval philosophy was practiced by both sides. As far as Buridan is concerned, much of the confusion over the proper domains of metaphysics and theology stems from Aristotle himself, who identifies three kinds of speculative science: physics, mathematics, and theology.[35] In his question commentary on the *Metaphysics*, he provides the standard medieval interpretation of this passage, reading ‘theology’ as ‘metaphysics’ and differentiating the three sciences in terms of how they treat their respective subjects: whereas the natural philosopher considers things insofar as they are qualified by motion, and the mathematician insofar as they are quantified by number, the metaphysician considers them only insofar as they pertain to the “concept of being [*ad rationem essendi*]” (QM VI.2: 34ra).[36] For the difference between metaphysics and theology, however, we need to go to the very beginning of the commentary, where he offers the following gloss of Aristotle's remarks:

> It should also be noted that [when we ask whether metaphysics is the same as wisdom,] we are not comparing metaphysics to theology, which proceeds from beliefs that are not

> known, because although these beliefs are not known *per se* and most evident, we hold without doubt that theology is the more principal discipline and that it is wisdom most properly speaking. In this question, however, we are merely asking about intellectual habits based on human reason, [i.e.,] those discovered by the process of reasoning, which are deduced from what is evident to us. For it is in this sense that Aristotle calls metaphysics 'theology' and 'the divine science'. Accordingly, metaphysics differs from theology by the fact that although each considers God and those things that pertain to divinity, metaphysics only considers them as regards what can be proved and implied, or inductively inferred, by demonstrative reason. But theology has for its principles articles [of faith], which are believed quite apart from their evidentness, and further, considers whatever can be deduced from articles of this kind. (QM I.2: 4ra-rb)

The difference is that theologians take their principles ('*principia*' = (lit.) 'starting points') from articles of faith rather than, as the philosophers do, from what is evident to the senses and intellect. So it is possible for the same question -- e.g., about the limits of divine omnipotence -- to be raised in both domains, though it will have a more creaturely orientation for the philosopher.[37] Buridan concedes *de jure* place of privilege to theology because of its subject matter, but treats philosophy and theology as *de facto* equivalent in the speculative realm. Metaphysics, or philosophical wisdom, cannot be ordained by theology because its methods, which are rooted in its principles, are different. Philosophy is accordingly not inferior to theology, just different.[38]

With regard to the problem of universals,[39] Buridan does not so much create a new theory as show how our theoretical commitments can be expressed with a minimum of ambiguity and fuss. Like Ockham, he is a nominalist, although this term must be used with caution in later medieval philosophy because of the modern tendency to identify it simply with the denial of real universals. Most 14th-century philosophers were nominalists in this sense because they associated the contrary doctrine with Plato, with whom they were familiar only secondarily as one of the authors thoroughly discredited by Aristotle in Book I of the *Metaphysics*.[40] But medieval nominalism involved much more than rejecting Platonic universals. Its history can be traced to 12th-century disputes over the reading of sacred texts, in which the techniques of logicians such as Abelard were pitted against those of grammarians such as Peter Helias.[41] As these disputes matured, nominalism was gradually absorbed into the teaching of philosophers working in the faculty of arts, so that by Buridan's time, it is better to think of nominalism as a practice, or way of doing philosophy, and not as a piece of doctrine.

When Buridan considers questions such as "whether universals actually exist outside the soul" (TDUI: 137), his remarks are almost always aimed at clarifying the meaning of the term 'universal' with respect to other terms such as 'individual', 'particular', and 'singular'. His rejection of realism is expressed in the claim that universal terms have no ultimate significate, i.e., nothing outside the soul they can 'make known' as such.[42] Hence, an account is needed of what such terms mean. Here, it is almost as if Buridan thinks there is something ill-formed about propositions where the term 'universal' occurs in the subject position, for when confronted with them his first move is always to tell us how the term 'universal' should be understood (QIP 3: 136, ll. 477-488).[43] He argues further that the primary

signification of ‘universal’ is ‘predicable of many’, which makes it a term of second intention, or a term of terms, since only terms are predicable (QIP 3: 135-136; 4: 139; TDUI: 148).[44] The second-intentional status of the term ‘universal’ is also evident in propositions, where it does not signify a ‘what’ but a ‘how’, i.e., how we conceive of something -- in this case how the term so designated is “indifferent to many supposits,” or individuals (TDUI:59).[45] As we saw above, logic is the study of terms such as ‘proposition’ taken materially, signifying actual tokens of the type (QDI I.1: 6). Moving from propositions to arguments, Buridan insists that terms in the premises and conclusions of demonstrative arguments must be taken as standing materially, i.e., for themselves in the particular discourse conditions which surround them, rather than personally, for their extra-mental significates (QIP 1: 128, ll. 223-237). Likewise, the proximate object of scientific knowledge is the actual demonstrated or demonstrable conclusion rather than the state of affairs it signifies, although Buridan is willing to concede that “the terms of those demonstrable conclusions, or even the things signified by those conclusions” might be considered “remote”, or secondary, objects of knowledge (QIP 1: 127,l. 208-209).[46] Careful and systematic analysis is the best antidote for metaphysical perplexity because the trouble usually begins with untutored persons who don’t know what some word or concept means.

It should be noted that Buridan's methods do not always produce parsimonious results. Like Ockham, he has only substances and qualities in his basic ontology, but he is much more willing than Ockham to expand this in the direction of modes or ways of being when confronted with recalcitrant phenomena. Thus, he argues that we must treat the question of *how* something is as distinct from *what* it is if we are to have a coherent understanding of motion, especially since the Ockhamist view is forced to posit an infinite succession of spatial qualities (QM V.9: 32va; QP II.3:1ra-rb; QP IV.11: 77va-78rb).[47] Likewise, in a famous passage, Buridan is driven by his own experience to reject Ockham's explanation of condensation and rarefaction as kinds of locomotion. Why, he asks, does he find himself unable to compress further the air in a bellows which has been stopped up at one end? Not because of its matter, because more matter could exist in a much smaller space; nor because of the substantial form of the air, which would fill a much smaller space once it has cooled; nor even because of the heat it possesses, since more heat could exist in a much smaller space, such as at the end of a red-hot poker. No, the air must have a distinct quantitative form or magnitude preventing it from being moved. Against those who would do away with this distinction, Buridan argues that “a magnitude of this sort has not been posited in vain, for we have been compelled to posit it by arguments that make it seem as useful or even more useful to natural philosophy than [the qualities of] whiteness or blackness” (QP.8: 11vb).[48]

6. Natural Philosophy

Medieval natural philosophy included both the rich commentary tradition on Aristotle's Physics, as well as treatises and commentaries on Aristotle's other writings about the natural world: *On the Heavens*, *On Generation and Corruption*, *De anima*, the short treatises on animate powers known as the ‘*Parva Naturalia*’, and the works on the history, parts, and generation of animals. Buridan wrote commentaries on nearly all of these texts. Like his contemporaries, he understood the speculative sub-disciplines of physics as an orderly progression of learning. Thus, a demonstration in psychology was thought to borrow its principles or starting points from physics, the higher science to which it is subordinated, with

its conclusions in turn supplying principles for demonstrations in the more specialized studies of sensation and animal locomotion. The ordering metaphor is never far from Buridan's mind. He remarks to readers of his commentary on Aristotle's *De motibus animalium* that in this book, “we descend to the different species of motion in particular, e.g., to the fact that some animals fly, others swim, and so on” (DMA I: 535)

Besides his contributions to logic, natural philosophy is the field where Buridan has enjoyed some recognition. This is due to the efforts of the pioneering historians of science Pierre Duhem and Anneliese Maier, who saw that Buridan had played a key role in the demise of the Aristotelian view of the cosmos.[49] Buridan's major contribution here was to develop and popularize the theory of impetus, or impressed force, to explain projectile motion. Rejecting the discredited Aristotelian idea of antiperistasis, according to which the tendency of a moving projectile to continue moving is due to a proximate but external moving cause (such as the air surrounding it),[50] he argues that only an internal motive force, transmitted from the mover to the projectile, could explain its continued motion. The theory of impetus probably did not originate with Buridan,[51] but his account differs from the others in that he entertains the possibility that it might not be self-dissipating: “after leaving the arm of the thrower, the projectile would be moved by an impetus given to it by the thrower,” he says, “and would continue to be moved as long as the impetus remained stronger than the resistance, and would be of infinite duration were it not diminished and corrupted by a contrary force resisting it or by something inclining it to a contrary motion” (QM XII.9: 73ra). He also contends that impetus is a variable quality whose force is determined by the speed and quantity of the matter in the subject, so that the acceleration of a falling body can be understood in terms of its gradual accumulation of units of impetus. But despite its revolutionary implications, Buridan did not use the concept of impetus to transform the science of mechanics. He was not, as Duhem argued, a forerunner of Galileo. He remained unapologetically Aristotelian in too many other respects, continuing to hold that motion and rest are contrary states of bodies and that the world is finite in extent. Buridan seems to have been a philosopher who, though he was well aware of the shortcomings of the Aristotelian natural philosophy, tried to reshape as much of it as he could in the face of an increasingly mechanistic worldview.[52] His revolution was in the details. The big, decisive break was left to his successors.

Buridan's account of motion is in keeping with his approach to natural science, which is empirical in the sense that it emphasizes the evidentness of appearances, the reliability of a posteriori modes of reasoning, and the application of naturalistic tropes or models of explanation (such as the concept of impetus) to a variety of phenomena. He is inclined to dismiss purely theological assumptions as irrelevant to the practice of philosophy: “one might assume that there are many more separate substances than there are celestial spheres and celestial motions, viz., great legions of angels [*magnae legiones angelorum*], but this cannot be proved by demonstrative arguments originating from sense perception” (QMII.9: 73ra).[53] But there are some theological considerations that must be taken seriously. He concedes that divine omnipotence is such that it is always possible for God to deceive us in ways we could never detect, although this is tempered by his confidence, for which he cites empirical evidence, that our ordinary powers of perception and inductive inference are sufficiently reliable to make “the comprehension of truth with certitude possible for us” (QM II.1: 9ra). He has little patience for skeptical arguments

questioning the possibility of scientific knowledge, such as those he believed were advanced by Nicholas of Autrecourt, arguing that it is absurd to demand that all knowledge be demonstrable by reduction to the principle of non-contradiction. Natural science is about what happens for the most part, i.e., assuming the common course of nature. For the same reason, explanatory models in one special science can be helpfully applied to others. Thus, the concept of impetus recurs in Buridan's psychology to explain the difference between occurrent and dispositional thinking, as well as in his ethics, to account for the relative ease with which virtuous persons are able to do the right thing (they have acquired a kind of impetus to act virtuously).

In psychology, the study of moving beings *qua* animate, Buridan changed the Aristotelian paradigm in important ways. In contrast with Thomas Aquinas, who wanted to attribute metaphysically more robust qualities, such as *per se* subsistence, to the human soul, Buridan does not think that psychology is in a position to reveal anything about the inherent nature of its subject, and refuses to speculate about it. He sees the science of psychology as concerned not with some quidditative concept of the soul arrived at by *a priori* reasoning, but with specifying the relation between animate qualities and the soul as their proper subject: “the natural philosopher only studies substances in relation to their motion and operations. And since natural forms require for their operations a determinate matter made suitable for them by qualitative and quantitative dispositions, natural scientists must define forms through their proper matter. Therefore, in its natural definition, the soul must be defined by means of a physical, organic body” (QDA II.3: 34). This led to an even more attenuated conception of the soul in the *De anima* commentaries of Buridan's Parisian successors, Nicole Oresme and Peter of Ailly, for whom the soul functions as a kind of placeholder whose nature is not even relevant to psychology. Consideration of the soul's ultimate nature moved instead to the faculty of theology, where it was considered along with the same Augustinian dreaming arguments that undoubtedly influenced Descartes.[54]

Buridan's resolute naturalism is even more evident in his account of human knowledge. Here he is a representationalist, but mainly because he can find no evidence to support the contrary view, defended by both Duns Scotus and William of Ockham, that humans also possess an intuitive mode of knowing which gives them a direct and unmediated awareness of the existence of some object.[55] Buridan holds that cognition can occur only when the intellect apprehends an object by means of a particular species or concept representing it. How we use the concept in our act of apprehension determines whether we cognize singularly or universally: singularly, if its object appears to us “in the manner of something existing in the presence of the person cognizing it” universally, if we focus on certain features of the concept to the exclusion of others, such as when we cognize “all human beings indifferently” via the humanity that is in the concept of Socrates (QDA III.8: 79-80; QP I.7: 8va). How do we know whether our concept of Socrates is really a concept of Socrates? As noted above, Buridan tends to regard skeptical worries as tiresome and pedantic. Like most pre-Cartesian epistemologists (if it makes sense to use that term before Descartes), he is more interested in explaining the process by which we come to have knowledge than he is in justifying knowledge claims.

7. Ethics

Buridan's commentary on Aristotle's *Nicomachean Ethics* is one of his most influential works, though it is one of the least studied. It contains significant discussions of the structure of the will and its relation to the intellect, the nature of human freedom, the phenomenon of *akrasia* or weakness of will, practical reason, and the unity of the virtues.[56]

In moral psychology, Buridan appears to effect a compromise between two rival views of the relation between the will and the intellect: the intellectualist or naturalist tradition associated with Aristotle and Thomas Aquinas, according to which the will is always subordinate to the intellect, and the voluntarist tradition of Augustine and Franciscan thinkers such as Duns Scotus and William of Ockham, which held that the will is sometimes capable of autonomous activity. Buridan's apparent compromise is to argue (with the intellectualists) that human happiness ultimately consists in an intellectual act, “the perfect apprehension of God”, rather than in a volitional act such as perfectly willing or perfectly loving God (QNE X.5: 213rb), although he emphasizes (with the voluntarists) the role of the will as a self-determining power in achieving that end. The compromise turns on Buridan's innovative conception of free choice, which develops Albert the Great's notion that certainty admits of degrees.[57] The idea is that even if the will lacks the power to choose evil as such, it is still able to defer its choice and do nothing if the goodness of one of the alternatives presented to it is unclear or uncertain. Of course, given our poor epistemic position in this life, it will almost always have this power because it is almost always possible in practice to doubt the goodness of a proposed course of action. The compromise is only apparent, however, because it turns out that the will's act of deferment is possible only if “the intellect would judge it to be good to consider the matter further” (QNE III.5: 44va).[58] This claim, together with his assumptions that the will can only choose non-optimally through ignorance or impediment (QNE III.3-5; 9), places deferment squarely under the jurisdiction of the intellect, which must weigh the relative goodness of different possible courses of action, including deferment. It is not the case that the will can choose to defer regardless of what the intellect decides, or that the will can choose to defer even if the intellect has judged deferment to be less good than some other course of action. Accordingly, if this is a compromise between the intellectualists and the voluntarists, it is a disingenuous one. It is more likely that Buridan simply appropriated voluntarist terminology to express what is otherwise a straightforwardly intellectualist account of the will, perhaps to dispel the cloud of heterodoxy which had surrounded intellectualist moral psychology since the Condemnation of 1277.[59]

It is also in Buridan's moral psychology that we find the most plausible explanation of the example that has come down to us known as ‘Buridan's Ass’, in which a donkey starves to death because it has no reason to choose between two equidistant and equally tempting piles of hay. This particular example is nowhere to be found in Buridan's writings, although there are versions of it going back at least to Aristotle (see *De Caelo* 295b32).[60] The best explanation of its association with Buridan is that it originated as a parody of his account of free choice by later critics, who found absurd the idea that the will's freedom could consist in inaction, i.e., in its ability to defer or ‘send back’ for further consideration any practical judgment that is not absolutely certain.

Bibliography

Early Printed Editions

- Buridan, John: 1509, *Subtilissimae Quaestiones super octo Physicorum libros Aristotelis*, Paris. Rpr. 1964, as *Kommentar zur Aristotelischen Physik*, Minerva, Frankfurt a. M. [QM]
- Buridan, John: 1513, *Quaestiones super decem libros Ethicorum Aristotelis ad Nicomachum*, Paris. Rpr. 1968, as *Super decem libros Ethicorum*, Minerva, Frankfurt a. M. [QNE]
- Buridan, John: 1588 (actually 1518), *In Metaphysicen Aristotelis Questiones argutissimae*, Paris. Rpr. 1964, as *Kommentar zur Aristotelischen Metaphysik*, Minerva, Frankfurt a. M. [QP]

Modern Editions

- Bos, E. P. (ed.): 1994, *Johannes Buridanus, Summulae: In Praedicamenta*, Artistarium 10-3, Ingenium, Nijmegen.
- Hubien, Hubert (ed.): 1976, *Iohannis Buridani Tractatus de consequentiis*, Philosophes Médiévaux XVI, Publications universitaires, Louvain. [TC]
- Hubien, Hubert (ed.): *Iohannis Buridani Quaestiones in duos libros Aristotelis Posteriorum Analyticorum*, unpublished typescript. [QAnPo]
- Hubien, Hubert (ed.): *Iohannis Buridani Quaestiones in duos libros Aristotelis Priorum Analyticorum*, unpublished typescript. [QAnPr]
- Kilcullen, R. J. (ed.): 1996: “Buridan, On Aristotle's Ethics, Book X.” [Available online]
- van der Lecq, Ria (ed.): 1983, *Johannes Buridanus, Questiones longe super librum Perihermeneias*, Artistarium 4, Ingenium, Nijmegen. [QDI]
- van der Lecq, Ria, and H. A. G. Braakhuis, (ed.): 1994, *Johannes Buridanus, Questiones Elencorum*, Artistarium 9, Ingenium, Nijmegen.
- Moody, E. A. (ed.): 1942, *Iohannis Buridani Quaestiones super libris quattuor De caelo et mundo*, Medieval Academy of America, Cambridge (Mass.).
- Patar, Benoît (ed.): 1991, *Le Traité de l’âme de Jean Buridan [De prima lectura]*, Philosophes Médiévaux, Tome 29, Éditions de l’Institut Supérieur de Philosophie-Éditions du Préambule, Louvain-Longueuil (Québec).
- Rijk, L. M. de (ed.): 1995, *Johannes Buridanus, Summulae: De Praedicabilibus*, Artistarium 10-2, Ingenium, Nijmegen.
- Schneider, Johannes (ed.): 1983, *Iohannes Buridanus Quaestiones in Praedicamenta*, Beck, München. [QC]
- Scott, Frederick, and Herman Shapiro (ed.): 1967, “John Buridan's De motibus animalium,” *Isis* 58, 533-552. [DMA]
- Sobol, Peter Gordon (ed.): 1984, “John Buridan on the Soul and Sensation: An Edition of Book II of His Commentary on Aristotle's Book of the Soul, with an Introduction and a Translation of Question 18 on Sensible Species,” Ph.D. dissertation, Indiana University. [QDA]
- Szyller, Slawomir (ed.): 1987, “Johannis Buridani, Tractatus de differentia universalis ad individuum,” *Przeglad Tomistyczny* 3, 137-178. [TDUI]
- Tatarzynski, Ryszard (ed.): 1986, “Jan Buridan, Kommentarz do Isagogi Porfiriusza,” *Przeglad Tomistyczny* 2, 111-195. [QIP]

- Thijssen, J. M. M. H. (ed.): 1991, *John Buridan's 'Tractatus de Infinito'*, Artistarium Supplementa 6, Ingenium, Nijmegen.
- Zoubov, Vassili (ed.): 1961, "Jean Buridan et les concepts du point au quatorzième siècle," *Medieval and Renaissance Studies* 5, 63-95.
- Zupko, John Alexander (ed. & tr.): 1989, "John Buridan's Philosophy of Mind: An Edition and Translation of Book III of his 'Questions on Aristotle's De anima' (Third Redaction), with Commentary and Critical and Interpretative Essays," Ph.D. dissertation, Cornell University. [QDA III]

Translations

- Hughes, G. E. (ed. & tr.): 1982, *John Buridan on Self-Reference*: Chapter Eight of Buridan's 'Sophismata', An Edition and Translation with an Introduction and Philosophical Commentary, Cambridge University Press, Cambridge-London-New York. (Cambridge also published a paperbound edition of this book, though without the facing Latin text and hence with different pagination. It also has a slightly different subtitle, "Chapter Eight of Buridan's 'Sophismata', translated with an Introduction and a philosophical Commentary", which is useful for identifying it in online searches.)
- King, Peter (tr.): 1985, *John Buridan's Logic*: The Treatise on Supposition; The Treatise on Consequences, Translation from the Latin with a Philosophical Introduction, Reidel, Dordrecht-Boston-Lancaster.
- Klima, Gyula (tr.): 2001, *John Buridan: 'Summulae de Dialectica'*, Yale Library of Medieval Philosophy, Yale University Press, New Haven-London. [S]

Secondary Sources

- Adams, Marilyn McCord: 1987, *William Ockham*, 2 vols., University of Notre Dame Press, Notre Dame, IN.
- Biard, Joël: 1989, *Logique et théorie du signe au XIVe siècle*, Vrin, Paris.
- Courtenay, William J.: 1999, *Parisian Scholars in the Early Fourteenth Century: A social portrait*, Cambridge University Press, Cambridge-New York.
- Des Chene, Dennis: 1996, *Physiologia*: Natural Philosophy in Late Aristotelian and Cartesian Thought, Cornell University Press, Ithaca-London.
- Duhem, Pierre: 1906-13, *Études sur Léonard de Vinci*, 3 vols., Hermann, Paris.
- Faral, Edmond: 1950, *Jean Buridan: Maître ès Arts de l'Université de Paris*, Extrait de l'Histoire littéraire de la France, Tome 28, 2e partie, Imprimerie Nationale, Paris.
- Flüeler, Christoph: 1992, *Rezeption und Interpretation der Aristotelischen "Politica" im späten Mittelalter*, 2 vols., B. R. Grüner, Amsterdam-Philadelphia.
- Grant, Edward: 1977, *Physical Science in the Middle Ages*, Cambridge University Press, New York.
- Hughes, G. E.: 1989, "The Modal Logic of John Buridan," in *Atti del Convegno internazionale di storia della logica: la teorie delle modalita*, ed. G. Corsi, C. Mangione, and M. Mugnani,

CLUEB, Bologna, 93-111.

- Knuuttila, Simo: 1991, “Buridan and Aristotle's Modal Syllogistic,” *Historia Philosophiae Medii Aevi. Studien zur Geschichte der Philosophie des Mittelalters*, ed. Burkhard Mojsisch and Olaf Pluta, Band 1, B. R. Grüner, Amsterdam-Philadelphia, 477-488.
- Krieger, Gerhard: 1986, *Der Begriff der praktischen Vernunft nach Johannes Buridanus*, Aschendorff, Münster.
- Lagerlund, Henrik: 2000, *Modal Syllogistics in the Middle Ages*, Brill, Leiden-Boston-Köln.
- Michael, Bernd: 1985, “Johannes Buridan: Studien zu seinem Leben, seinen Werken und zu Rezeption seiner Theorien im Europa des späten Mittelalters”, 2 Teile, Ph.D. Dissertation, University of Berlin.
- Maier, Anneliese: 1955, *Metaphysische Hintergründe der spätscholastischen Naturphilosophie*, Studien zur Naturphilosophie der Spätscholastik, Bd. IV, Storia e Letteratura, Roma.
- Normore, Calvin: 1985, “Buridan's Ontology,” ed. James Bogen, and James E. McGuire, *How Things Are: Studies in Predication and the History and Philosophy of Science*, Reidel, Dordrecht-Boston-Lancaster, 189-203.
- Pironet, Fabienne: 1993, “John Buridan on the Liar Paradox: Study of an Opinion and Chronology of the Texts,” in *Argumentationstheorie: Scholastische Forschungen zu den logischen und semantischen Regeln korrekten Folgerns*, ed. Klaus Jacobi, E. J. Brill, Leiden-New York-Köln, 293-300.
- Rijk, L. M. de: 1992, “John Buridan on Universals,” *Revue de Métaphysique et de Morale* 97, 35-59.
- Saarinen, Risto: 1994, *Weakness of the Will in Medieval Thought, From Augustine to Buridan*, E. J. Brill, Leiden-New York-Köln.
- Schönberger, Rolf: 1994, *Relation als Vergleich: die Relationstheorie des Johannes Buridan im Kontext seines Denken und der Scholastik*, E. J. Brill, Leiden-New York-Köln.
- Thijssen, J. M. M. H.: 1998, *Censure and Heresy at the University of Paris, 1200-1400*, University of Pennsylvania Press, Philadelphia.
- Thijssen, J. M. M. H., and Jack Zupko (ed.): 2001, *The Metaphysics and Natural Philosophy of John Buridan*, Brill, Leiden-Boston-Köln.
- Walsh, James J.: 1986, “Buridan on the Connection of the Virtues,” *Journal of the History of Philosophy* 24: 453-482.
- William of Ockham: 1974, *Summa Logicae, Opera Philosophica I*, ed. Philotheus Boehner, Gedeon Gál, and Stephen F. Brown, St. Bonaventure University Press, St. Bonaventure, NY.
- Zupko, Jack: 1993a, “Buridan and Skepticism,” *Journal of the History of Philosophy* 31, 191-221.
- Zupko, Jack: 1993b: “Nominalism Meets Indivisibilism,” *Medieval Philosophy and Theology* 3, 158-185.
- Zupko, Jack: 1995, “Freedom of Choice in Buridan's Moral Psychology,” *Mediaeval Studies* 57, 75-99.
- Zupko, Jack: 1997, “What Is the Science of the Soul? A Case Study in the Evolution of Late Medieval Natural Philosophy,” *Synthese* 110.2, 297-334.
- Zupko, Jack: 2001, “John Buridan on the Immateriality of the Intellect,” *Proceedings of the Society for Medieval Logic and Metaphysics* 1 (2001): 4-18. [Preprint available online]
- Zupko, Jack: forthcoming, *John Buridan: Portrait of a 14th-Century Arts Master*, University of

Notre Dame Press, Notre Dame, IN.

Other Internet Resources

- Special Bibliography on Jean Buridan (in French) (maintained by Fabienne Pironet, U. Montreal)
- Medieval Logic and Philosophy (maintained by Paul Vincent Spade, Indiana University)
- Late Medieval and Early Modern Intellectual History (maintained by R. J. Kilcullen, Macquarie University)

Related Entries

insolubles [= *insolubilia*] | medieval philosophy: literary forms of | modality: medieval theories of | Nicholas of Autrecourt | Peter of Spain [= Petrus Hispanus] | *sophismata* [= sophisms] | terms, properties of: medieval theories of | universals: the medieval problem of

Jack Zupko
jzupko@emory.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 13, 2002
Content last modified: May 13, 2002

Stanford Encyclopedia of Philosophy
Notes to John Buridan

Notes

1. Michael 1985: 79-238; 399-404. The best bibliographical resource on Buridan is by Fabienne Pironet, available at her website (see address under *Other Internet Resources* below).

2. However, the significance of such details is beginning to be understood more fully by historians through recent prosopographical research. See Courtenay 1999.

3. For a discussion and careful debunking of each of these stories, see Faral 1950: 9-33.

4. Faral 1950: 16.

5. See Zupko, *John Buridan: Portrait of a 14th-Century Arts Master*, Chapter 10.

6. Faral 1950: 11.

7. There is also a commentary on Aristotle's *Politics* that has been mistakenly attributed to Buridan, though it is actually the work of Nicholas of Vaudemont, a late 14th-century Parisian Arts Master who was much influenced by Buridan. See Flüeler 1992: vol. 1, 132-68.

8. For discussion of Peter of Spain and the teachings of the *Summulae logicales*, see Joke Spruyt's article, "Peter of Spain (Petrus Hispanus)" in this *Encyclopedia*.

9. By the beginning of the 14th century, arts masters were no longer teaching grammar at the University of Paris. The last arts master to compose a treatise on grammar was Radulphus Brito (d. 1320). What appears to have happened is that the teaching of grammar was gradually taken over by quasi-independent colleges and other schools that grew up on the periphery of the university.

10. See John Trentman, "Ockham on Mental," *Mind* N.S. 79 (1970): 586-90, and Claude Panaccio, "Semantics and Mental Language," in *The Cambridge Companion to Ockham*, ed. Paul Vincent Spade, (Cambridge: Cambridge University Press, 1999): 53. See also Ockham, *Summa logicae* I.1-3 (OPh I: 7-14).

11. *Summa logicae* III.3 (OPh: 46). As Paul Spade puts it, "[for Ockham,] self-reference is to be allowed

except where it would lead to paradox — in short, it is licit except where it is illicit" ("Ockham on Self-Reference," *Notre Dame Journal of Formal Logic* 15 (1974): 299).

12. I owe this insight to Joël Biard (1989: 196).

13. For Buridan's treatment of the concept of infinity, see Thijssen 1991. For his conception of points, see Zupko 1993b.

14. For discussion of some 14th-century views on this distinction, see Sten Ebbesen, "Is Logic Theoretical or Practical Knowledge?," in *Itinéraires d'Albert de Saxe: Paris-Vienne au XIVe siècle*, ed. Joël Biard (Paris: Vrin, 1991): 267-283.

15. For discussion, see Paul Vincent Spade, "The Semantics of Terms," in *The Cambridge History of Later Medieval Philosophy*, ed. Norman Kretzmann, Anthony Kenny, and Jan Pinborg (Cambridge: Cambridge University Press, 1982): 192-3, and E. J. Ashworth, "Logic, Medieval," in *The Routledge Encyclopedia of Philosophy*, vol. 5, ed. Edward Craig (London: Routledge, 1999): 753-4.

16. Thus, Peter of Spain: "Simple supposition is the taking [*acceptio*] of a common term for the universal thing signified by it" (Peter of Spain: Tractatus (called afterwards '*Summulae logicales*'), ed. L. M. de Rijk (Assen: van Gorcum, 1972): 81). Early terminist logicians, it should be mentioned, did not all agree about the divisions of supposition. Peter, for example, does not mention material supposition, though William of Sherwood includes it in his *Introductiones in logicam*, 5.2 (Norman Kretzmann, *William of Sherwood's 'Introduction to Logic'*, Minneapolis: University of Minnesota Press, 1966): 107). See also Joke Spruyt, "Peter of Spain (Petrus Hispanus)" in this *Encyclopedia*.

17. William of Ockham, *Summa logicae* I.68; OPh I: 207. For nominalists such as Ockham, personal supposition offers a guide to ontology, since a term can supposit personally only for what exists *per se*.

18. Notice also that this severs the traditional connection between personal and material supposition as varieties of proper supposition. For Buridan, the only proper supposition is personal supposition; all of the others are strictly speaking improper. Stephen Read has linked Buridan's position on material supposition as improper with the breakthrough doctrine—first fully realized in the work of Buridan's student, Marsilius of Inghen—that material supposition is possible only if materially suppositing terms are significative, or stand for what they signify. See Read, "How is Material Supposition Possible?" in *Medieval Philosophy and Theology* 8 (1999): 18 and S 7.3.4: 522, where Buridan considers '*Homo est species*' as an instance of the fallacy of equivocation.

19. The qualification, 'at least in the first instance', is intended to cover the conventionality of signification beyond these primary acts of imposition, which are in Buridan's view naturally determined. He conveys this idea by saying that such concepts are acquired 'immediately [*statim*]', i.e., without deliberation: "from the singular visual cognition there immediately arises the universal intellectual

cognition, and so when we see this man, we immediately think of [a] man" (S 4.5.3: 296). Nevertheless, it is clear that even the signification commonly and principally given to the term 'man' could be changed after the fact if everyone agreed to use it in a different way. The case is somewhat more complicated at the conceptual level, since it does not seem open to any individual language-user to change the significance of his/her concepts at will. But even these can be changed indirectly, as a result of the dialectical relationship Buridan takes to hold between concepts and spoken or written languages. Thus, someone who learns from a book that kangaroos are marsupials does not acquire a new concept, but augments or modifies the concept he already has.

20. Not surprisingly, terms in the propositions of logic are said to occur in material supposition, since logic concerns the conventional classification of significant utterances and patterns of reasoning and persuasion. Its objects are the immediate, rather than the ultimate, significates of terms. See S 4.3.2: 257-8.

21. See Paul Vincent Spade, "Why Don't Mediaeval Logicians Ever Tell Us What They're Doing? Or, What Is This, A Conspiracy?", available online at www.pvspade.com/Logic.

22. Hughes 1989: 97. See also Simo Knuuttila, "Medieval Theories of Modality," in this *Encylopedia*.

23. See Knuuttila 1991: 487; Lagerlund 2000: 162-4. See also Knuuttila, "Modal Logic," in the *Cambridge History of Later Medieval Philosophy*: 355-7, and "Medieval Theories of Modality," in this *Encylopedia*.

24. See Eileen Sweeney, "Literary Forms of Medieval Philosophy," in this *Encyclopedia*.

25. A typical list of positions can be found, e.g., in the *Insolubilia* of Thomas Bradwardine, which was written in the 1320s. See Paul Vincent Spade, "Insolubles," in this *Encyclopedia*.

26. Another solution, which concedes that Socrates's proposition "is true and false at the same time", is rejected as sacrificing too much. The problem with this theory is that it makes it impossible to give the contradictory of Socrates's proposition, which means that it has no proper coordinates in Aristotelian logical space.

27. This is effectively Thomas Bradwardine's solution to the Liar (see Paul Vincent Spade, "Insolubles," in this *Encyclopedia*). But it was not without precedent. Among the others who defended Buridan's solution would have been Bonaventure, who, in the course of discussing one of Augustine's arguments for the existence of God (*Soliloquies* I.15) – i.e., that if no truth exists, then some truth exists; and if some truth exists, a First Truth exists – records the objection that the first inference fails because no proposition can entail its own contradictory. Bonaventure agrees, but adds the following qualification: "one must understand that an affirmative proposition makes a two-fold assertion, one which asserts the predicate of the subject, and the other which asserts that the proposition is true … Contradiction is

concerned with the first type of assertion, not the second. So when it is said that no truth exists [*nulla veritas est*], this proposition, insofar as it denies the predicate of the subject, does not imply its opposite, which is that some truth exists. But insofar as it asserts itself to be true, it implies that some truth exists [*infert aliquam veritatem esse*]" (*Quaestiones disputatae de mysterio Trinitatis*, q. 1, a. 1, ad 5; Latin text excerpted in Spade 1975: 53). For other advocates of this solution, see Paul Vincent Spade, "*Insolubilia*" in the *Cambridge History of Later Medieval Philosophy*: 249.

28. Cf. Buridan's earlier claim that contradiction requires not only the logical form of contradiction, but the speaker's intention (S 9.7, 2nd sophism: 943).

29. Pironet 1993: 294-5. Cf. Hughes 1982: 167-9.

30. Buridan's presentation of this alternative is complicated somewhat by the doctrinal claim that the 'nothing' signified by 'A man is a donkey' is not any kind of proposition, but 'that a man is a donkey [*hominem esse asinum*]', which is the *dictum* or sentential nominalization of that proposition (expressed in Latin by the accusative + infinitive construction). In Buridanian semantics, such a construction supposits for whatever both the subject and predicate terms of its corresponding proposition supposit for, provided the proposition is true; otherwise it supposits for nothing. Gyula Klima remarks that this is how "Buridan manages to assign some credible semantic function to such sentential nominalizations without having to subscribe to a dubious ontology of eternal or quasi-eternal *enuntiabilia*, or *complexe significabilia*, distinct from ordinary substances and accidents" (Klima 2001: 844, n. 28; for his reaction to those who did seem to subscribe to such an ontology, see section 3 above). Buridan's sensitivity to the ontological dimensions of the problem emerges when he says that "'that a man is a donkey' is nothing, because a man cannot be a donkey [*hominem esse asinum nihil est, eo quod homo non potest esse asinus*]" – which also suffices for its falsity.

31. A further problem has been drawn to my attention by Paul Spade, in his comments on an earlier draft of this article: "if every proposition signifies *se esse veram*, and we're construing the infinitival expression personally, we've got a problem. For it the proposition is *false*, the infinitival expression has nothing it can signify, so that the proposition really *doesn't* have any additional signification at all, contrary to the whole point of the theory". If Buridan was aware of this as an additional problem for the first solution, he does not mention it.

32. As Hughes 1982 has suggested, the force of 'virtually' in 'virtually implies' is that the second proposition would be implied by the first only if the first is actually formulated (169). This emerges in a closing comment on the sophism in which Buridan remarks, "perfecting this solution, we have to say that every proposition, adding *that it exists*, implies that it is true" (S 9.8, 7th sophism: 970).

33. Of course, this holds only for affirmative propositions. Negative propositions are true if their subject and predicate terms *do not* stand for the same thing or things.

34. For a helpful overview, see Thijssen 1998: "In many official documents and other texts, philosophers and theologians were exhorted not to cross the boundaries of their own fields—a reference to *Proverbs* 22:28—and not to become theologizing philosophers and philosophizing theologians" (2).

35. Aristotle, *Metaphysics* VI.1.1026a18; XI.7.1064b1.

36. Precedents for reading 'theology' as 'metaphysics' in this context—which is somewhat obscured by the fact that the incunabular edition of Buridan's *Metaphysics* commentary erroneously gives '*metaphysica totalis*' for '*mathematica totalis*' and '*metaphysicus*' for '*mathematicus*' in the 23rd and 25th lines of folio 34ra—can be found in Robert Kilwardby, *De Ortu Scientiarum* LXVI.655 and Thomas Aquinas, *In De trin*. V, a. 4, as well as in Albert the Great.

37. Thomas puts this same distinction in terms of the *ratio* or concept under which the subject is considered: "although philosophy considers all existing things according to concepts [*rationes*] taken from creatures, there must be another science, which considers existing things according to concepts [*rationes*] taken from the inspiration of the divine light" (*In I Sent*., Prol., q.1, a.1, ad 1; cf. *In De trin*., q.5, a.1-4).

38. For this reason as well, Buridan never tries to compare the methods of philosophy and theology, let alone to suggest how the former might be subsumed by the latter. He is aware of the possibility of rapprochement between the two sides, if only by its absence from the Parisian scene: "it seems to me that this question [about whether it is possible for demonstration to cross disciplinary lines] is difficult first because there has been exceedingly little discussion between the philosophers and the doctors [of theology], and second because it touches on the means of distinguishing the sciences, and it is even more difficult to assign whence, and in what way, the sciences originally received their distinction" (*QAnPo* I.23).

39. See also Gyula Klima's article, "Medieval Theories of Universals" in this *Encyclopedia*, especially sections 8-10.

40. Buridan's allusions to Plato on universals are fairly typical: see, e.g., QM VII.15: 50va-vb, and de Rijk 1992. On other topics, however, he is sometimes suspicious about whether a position handed down as Plato's is in fact Plato's. For example, after disposing of an argument introduced "on the authority of Plato" for the role of separate substances in the generation of living things, he concludes, "and so in this way Plato's opinion is destroyed – if he in fact had an opinion of the sort we have ascribed to him" (QM VII.9: 47ra). For discussion of Buridan and Plato, see Schönberger 1994: 292-95.

41. For the meaning of 12th-century nominalism, see the special issue of *Vivarium* (30.1, May 1992) devoted to this topic.

42. See Peter King, "John Buridan's Solution to the Problem of Universals" in Thijssen and Zupko 2001:

1-27. King argues that Buridan's nominalism has three interrelated aspects: (1) the ontological thesis that there are no non-individual entities in the world; (2) the psychological thesis that some concepts, though metaphysically particular, can represent more than one individual thing; and (3) the semantic thesis that such concepts also function as common names in Mental Language.

43. Cf. QIP 4: 138, ll. 565-566. Buridan's treatment of transcendental terms is similar: "the subject of metaphysics is being, that is, the term 'being'" (QIP 3: 135, ll. 449-450).

44. Buridan even regards it as conventional that we treat universals as second-intentional names (TDUI: 145-146). Universals are substances in the second mode of substance only, i.e., "they are terms in the category of substance" (QIP 4: 140, ll. 635-636); likewise, "'universal' is a transcendent name", and such names occur on one level only (TDUI: 147); and as a form, a universal is a second intention (TDUI: 148)

45. Note also that the differences in universal names do not correspond to any real diversity in the things signified by those names, "but in the medium through which we arrive at the concepts by which those names are imposed" (QIP 11: 173, ll. 1853-1854). The "medium" is the more common or general concept, e.g., of sensing, from which the common concept of everything sensitive is formed, and names such as 'animal' are imposed (QIP 11: 173, ll. 1847-1850).

46. That the object of knowledge is a *complexum*, or proposition, rather than an *incomplexum*, or a term, follows from the fact that we can believe or know only what can be true or false, and only propositions can be true or false.

47. For discussion, see Normore 1985.

48. For discussion, see Maier 1955: 214-15 and Adams 1987: 184-5.

49. See Duhem 1906-13 and Maier 1955. For discussion, see Grant 1977.

50. Thus, for Aristotle, the thrown javelin does not continue to move because of any force inside the javelin (which is, after all, an inanimate object), but because its movement through the air creates a vacuum behind it which the surrounding air rushes in to fill, thereby pushing the javelin forward.

51. We do not know precisely where Buridan got the idea of impetus, but a less sophisticated notion of impressed forced can be found in Avicenna's doctrine of *mayl* (inclination). In this he was possibly influenced by Philoponus, who was developing the Stoic notion of *hormé* (impulse). For discussion, see Zupko 1997.

52. We can also see this in his defense of the sufficiency of efficient causality in explanations of natural phenomena, which eventually led to the eclipse of final causality several centuries later. For discussion, see Des Chene 1996: 186-7.

53. It would not be like Buridan to rule out theological considerations completely. Indeed, he sometimes confronts theological issues with what can only be described as intellectual playfulness. Edith Sylla nicely describes his method here: “Buridan does not exclude theology from physics, along the lines of Boethius of Dacia, nor does he overwhelm physics with theology, along the lines of today’s Creationists. Rather, in a moderate way, Buridan introduces theological truths into the body of Aristotelian physics and then shows, plausibly, that to draw inferences from physics plus theology, it is necessary to add other hypotheses [e.g., to assume a reference frame for extra-cosmic motion, or a ‘time’ to measure duration before creation]—to beg the human intellect” (Sylla, “*Ideo quasi mendicare oportet intellectum humanum*: The Role of Theology in John Buridan’s Natural Philosophy,” in Thijssen and Zupko 2000: 244-5).

54. See Zupko 1999 and Zupko 2001.

55. See Zupko 1993a.

56. For further discussion, see Krieger 1986 and Walsh 1986.

57. See Saarinen 1993: 161-93.

58. The page number is mistakenly given as “lxiiii” in the Paris 1513 edition.

59. For further discussion, see Zupko 1995 and Fabienne Pironet, “The Notion of ‘*Non Velle*’ in Buridan’s Ethics,” in Thijssen and Zupko 2001: 199-219.

60. The earliest association of the example with Buridan appears to be in Spinoza, *Ethica* II, scholium to Proposition 49.

Jack Zupko
jzupko@emory.edu

First published: June 27, 2002
Content last modified: June 27, 2002

Insolubles

The medieval name for paradoxes like the famous Liar Paradox ("This proposition is false") was "insolubles" or *insolubilia*.[1] From the late-twelfth century to the end of the Middle Ages and beyond, such paradoxes were discussed at length by an enormous number of authors. Yet, unlike twentieth century interest in the paradoxes, medieval interest seems not to have been prompted by any sense of theoretical "crisis."

The history of the medieval discussions can be divided into three main periods: (a) an early stage, from the late-twelfth century to the 1320s; (b) a period of especially intense and original work, during roughly the second quarter of the fourteenth century; (c) a late period, from about 1350 on. The discussion in this article will be organized as follows:

- Related Entries

---

1. Origins of the Medieval Discussion

The Liar Paradox was well known to antiquity. Its discovery is often credited to Eubulides the Megarian (4th century BCE), on the basis of a remark by Diogenes Laertius (*Lives of the Philosophers* II.108), although in fact Diogenes says only that Eubulides discussed the paradox, not that he discovered it.[2] A little later, the poet and grammarian Philetus (or Philitas) of Cos (c. 330-c. 270 BCE), if we are to believe the story in Athenaeus of Naucratis's *Deipnosophists* IX.401e, worried so much over the Liar that he wasted away and died of insomnia, as, according to Athenaeus, his epitaph recorded:

> Philetus of Cos am I
> 'Twas The Liar who made me die,
> And the bad nights caused thereby.[3]

Diogenes Laertius also reports (VII.196-98) that the Stoic logician Chrysippus (c. 279-206 BCE) wrote:

- *Introduction to the Liar;*
- *Liar Propositions: An Introduction;*
- six books on the Liar itself;
- *Reply to Those Who Think There Are Propositions That Are Both True and False;*
- *Reply to Those Who Solve the Liar Proposition by Division;*
- *On the Solution to the Liar* (in three books);
- *Reply to Those Who Say The Liar Argument Has False Premises.*

1.1 Unlikely Ancient Sources

But it does not appear that medieval interest in insolubles was derived directly from these or any other known ancient sources that discuss the Liar. Many of the relevant works were lost (e.g., the works of Chrysippus), while others were never translated into Latin and so were effectively unavailable to the Middle Ages. Indeed, it is not at all clear just *what* it was that prompted medieval interest. One might have supposed that, even if particular *theories* about the Liar were not transmitted to the Middle Ages from antiquity, at least *formulations* of Liar-type paradoxes must have been known and available to stimulate the medieval discussions. In fact, however, there are strikingly few possibilities.

Seneca (*Epistle* 45.10), for instance, mentions the Liar paradox by its Greek name *pseudomenon*, but does not actually formulate it. Again, St. Augustine perhaps has the Liar in mind in his *Against the Academicians* (III.13.29), where he refers to the "most lying calumny, 'if it is true [it is] false, if it is false

it is true'." But neither passage would likely be sufficient by itself to suggest the special problems of the Liar to anyone not already familiar with them.

Somewhat more explicit is Aulus Gellius's (2nd century CE) *Attic Nights* (XVIII.ii.10), "When I lie and say I am lying, am I lying or saying the truth?" But Gellius was not widely read in the Middle Ages, and no known medieval author cites him in the context of insolubles.[4] Again, Cicero's *Academica priora,* II.xxix.95-xxx.97, contains a fairly clear formulation: "If you lie and speak that truth, are you lying or speaking the truth? ... If you say you lie, and you speak the truth, you lie; but you say you lie, and you speak the truth; therefore, you lie." But this passage is never cited in the *insolubilia*-literature. Moreover Cicero, who wrote in Latin and so did not have to be translated to be available to the Middle Ages, calls such paradoxes "inexplicables" (*inexplicabilia*). If he was the catalyst for the medieval discussions, we would have expected to find that term in the *insolubilia*-literature, and we do not; the unanimous medieval term is 'insolubles'.

1.2 St. Paul's Reference to Epimenides

One initially plausible stimulus for the medieval discussions would appear to be the Epistle to Titus 1:12: "One of themselves, even a prophet of their own, said, The Cretians [= Cretans] are always liars, evil beasts, slow bellies." The Cretan in question is traditionally said to have been Epimenides, himself a Cretan. For this reason, the Liar Paradox is nowadays sometimes referred to as the "Epimenides." Yet, blatant as the paradox is here, and authoritative as the Epistle was taken to be, not a single medieval author is known to have discussed or even acknowledged the logical and semantic problems this text poses. When medieval authors discuss the passage at all, for instance in Scriptural commentaries, they seem to be concerned only with why St. Paul should be quoting pagan sources.[5] It is not known who was the first to link this text with the Liar Paradox.

1.3 Aristotle's *Sophistic Refutations*

By contrast with these passages, none of which was cited in the *insolubilia*-literature, there is a text from Aristotle's *Sophistic Refutations* 25, 180a27-b7, that, from almost the very beginning of the *insolubilia*-literature to the end of the Middle Ages, served as the framework for discussing insolubles. It occurs in Aristotle's discussion of the fallacy of confusing things said "in a certain respect" (*secundum quid*) with things said "absolutely" or "on the whole" (*simpliciter*). In this context, Aristotle supposes a man who takes an oath that he will become an oath-breaker, and then does so. Absolutely or on the whole, Aristotle says, such a man is an oath-breaker, even though with respect to the *particular* oath to become an oath-breaker he is an oath-keeper. Then Aristotle adds the intriguing remark, "The argument is similar too concerning the same man's lying and speaking the truth at the same time" (180b2-3). It was this sentence that many medieval authors took to be a reference to the Liar Paradox, which therefore, on the authority of Aristotle, could be solved as fallacy *secundum quid et simpliciter*.

The widespread appeal to this passage throughout the history of the *insolubilia*-literature indicates that the text did play some role in prompting medieval interest in insolubles. This suggestion is reinforced by

the fact that the earliest known medieval statement of the Liar occurs in 1132, around the time the *Sophistic Refutations* first began to circulate in Western Europe in Latin translation. (See Section 2 below.)

Nevertheless, it is hard to see how Aristotle's remarks can be made to fit the Liar Paradox. The oath-breaker, as the example was generally interpreted, takes *two* oaths: one, which he keeps, that he will commit perjury, and a second (it does not matter what it is) that he breaks, thereby fulfilling the first oath. The man is an oath-breaker and an oath-fulfiller, but with respect to *different* oaths; by breaking his second oath, rendering it false, he fulfills the first oath, making it true. It is a long way from that to the Liar Paradox, in which the *same* proposition is (it seems) both false and true. Despite what was interpreted as Aristotle's suggestion that the latter case fits the pattern of the former, it does not. Unless one *already* knew about the Liar, therefore, it is hard to see how this passage from Aristotle would have suggested it to anyone.[6]

In short, it seems clear that the *Sophistic Refutations* was instrumental in prompting medieval interest in insolubles. But more must have been involved too. Before medieval logicians could formulate genuine Liar-type paradoxes, they first had to go well beyond anything found in Aristotle's text. At present we cannot say whether they did this on the basis of some still unidentified ancient source or whether it was through their own intellectual power and logical insight.

2. Early Developments to the 1320s

In 1132, Adam of Balsham, the founder of the important logical school of the "*Parvipontani*" (so called because they gathered at the Petit Pont in Paris), wrote an *Art of Discussing* (*Ars disserendi*), in which he treats, among other things, various kinds of yes/no questions, including "whether he speaks truly who says he lies" and "whether he who says nothing but that he lies is saying the truth." (Adam of Balsham 1956, p. 107.)

The importance of this passage should not be exaggerated. It is true that it gives us the earliest known explicit medieval formulation of the Liar.[7] But Adam gives no attempt to *solve* the paradox, does not say it was a current topic of discussion in his day, and in fact does not even indicate he recognized its paradoxicalness. He simply offers it as an *example* of one kind of yes/no question.

It is not until later in the twelfth century that one finds an explicit statement of the special problems raised by insolubles. In his *On the Natures of Things* (*De naturis rerum*), of unknown date but apparently well known by the end of the century, Alexander Neckham 1967, p. 289, says[8]:

> Again, if Socrates says he lies, and says nothing else, he says some proposition. Therefore, either a true one or a false one. Therefore, if Socrates says only that he lies, he says what is true or what is false.

> But if (1) Socrates says only the proposition that Socrates lies, and he says what is *true*, then it is true that Socrates lies. And if it is true that Socrates lies, Socrates says what is false. Therefore, if Socrates says only the proposition that Socrates lies, and he says what is true, he says what is false.
>
> But if (2) Socrates says only the proposition that Socrates lies, and he says what is *false*, then it is false that Socrates says what is false. And if it is false that Socrates says what is false, Socrates does not say what is false. But if Socrates says only that he lies, he says either what is true or what is false. Therefore, if Socrates says he lies, he says what is true. Therefore, if Socrates says only that he lies, and he says what is false, then he says what is true.
>
> But if Socrates says only that he lies, he says what is true or false. Therefore, if Socrates says only that he lies, he says what is true *and* says what is false.

Nevertheless, although clearly Neckham was fully aware of what is paradoxical about the Liar, he makes no attempt to *solve* the paradox. He presents it only as an example of the "vanities" logic deals with. This suggests that by his day others *were* trying to solve the paradox, and in fact in the discussion of the fallacy *secundum quid et simpliciter* contained in the so called *Munich Dialectic* (= *Dialectica Monacensis*) from sometime in the second half of the century, we find the remark: "But how this fallacy arises in uttering the insoluble ‘I am saying a falsehood’, that is a matter discussed in the treatise on insolubles."[9]

The first text we have that actually tries to solve the paradox is an anonymous treatise from the very end of the twelfth or the very early thirteenth century (De Rijk 1966). From then on, there are a great number of surviving treatments. (See Spade 1975.) In the early 1320s, Thomas Bradwardine, in a preliminary section of his own treatise on insolubles, lists nine views in circulation in his day. (See Spade 1975, pp. 106-08; Spade 1987, pp. 43-46.) Some of these views can no longer be identified in the texts that survive from the period before Bradwardine, but among the surviving views, we can distinguish five broad approaches to "solving" the paradox.[10] (Sometimes these approaches are combined in a single author.)

2.1 Insolubles as Fallacies *secundum quid et simpliciter*

As might be expected in view of Section 1.3 above, many of these early theories attempted to analyze insolubles as fallacies *secundum quid et simpliciter*. Later in the *insolubilia*-literature, discussions often continued to be cast in terms of this fallacy, even though their real focus was generally on entirely different theoretical issues; Bradwardine is a good example.[11] The role of the fallacy thus became purely "honorary," preserving the authority of Aristotle.

In the early period, however, many but by no means all authors actually tried to *solve* insolubles as fallacies *secundum quid et simpliciter*. But for reasons described in part in Section 1.3 above, such

attempts were not very satisfactory. They often ended up adopting Aristotle's terminology, but using it in ways quite different from what he intended.

Aristotle had suggested (180b5-7) that insolubles are false *simpliciter* (absolutely/on the whole), but true *secundum quid* (in a certain respect). Some authors in the early medieval literature, however, argued that insolubles are absolutely neither true nor false, but only true in a certain respect and false in a certain respect.[12] Others used the terminology of *simpliciter* and *secundum quid*, but applied it to reference (*suppositio*) rather than to truth and falsehood, so that in insolubles certain terms did not refer "absolutely" to their referents, but only "in a certain respect." These views are in effect a kind of restriction on self-reference.[13]

2.2 The Theory of *Transcasus*

The theory of *transcasus* has nothing to do with the fallacy *secundum quid et simpliciter,* although it too seems to have had its origins in antiquity. The word *transcasus* is not a usual Latin word. It seems to be a literal translation of Greek *metaptosis.* In Stoic logic, propositions that change their true value over time were called *metapiptonta* (from the same root). Walter Burley in fact used the word exactly this way in 1302 in two short logical works. (Spade 1987, pp. 33-34.)

Nevertheless, while the term *transcasus* in the context of insolubles does have an association with time, it does not imply any *change* of truth value over time. Rather the theory of *transcasus* held that in the proposition ‘This statement is false’, the term ‘false’ refers *not* to the proposition in which it occurs, but rather to some proposition uttered earlier. Thus, when the liar says "I am lying," what he really means is "What I said just a moment ago was a lie." If the speaker did not in fact say anything earlier, then his present statement is simply false and no paradox arises.[14]

This odd view, like the last of those discussed in Section 2.1 above, amounts in practice to a restriction on self-reference. But it is not clear exactly what motivated it. In any event, the theory of *transcasus* appears to have disappeared as a theory actually held by anyone after the early period, although it continued to be mentioned in later authors' surveys of earlier views.[15]

2.3 Exercised Act vs. Signified (or Conceived) Act

A third theory from this early period distinguishes the "exercised" act from the "signified" or "conceived" act. The details of this theory are not yet well understood, but the basic strategy is to distinguish what the liar *says* he is doing (namely, lying) from he really *is* doing. John Duns Scotus, who held a version of this theory in his *Questiones on the Sophistic Refutations* (Scotus 1958), thought that what the liar is *really* doing (his "exercised act") is speaking the truth. In order to avoid the paradox, this theory would seem to be committed to saying that the exercised act and the signified act are *two* distinct acts, so that the theory, like the theory of *transcasus* (Section 2.2 above), is committed to some kind of restriction on self-reference.[16]

2.4 The Theory of Restriction

Even when not combined with *transcasus* or the theory that distinguishes exercised act from signified act, a very popular approach throughout the *insolubilia*-literature, in the early as well as the later period (and for that matter even in our own day), was to deny or restrict the possibility of self-reference. Such theories were called "restriction," and their proponents were called "restricters" (*restringentes*). All such theories maintained that in some or all cases, terms in propositions could not "supposit for" (stand for, refer to) the propositions in which they occur.

Some theories of restriction went further and also ruled out other patterns of reference. For example:

- Proposition *a* = ‘*b* is true’, and *b* = ‘*a* is false’. Here *a* refers to *b* and *b* refers back to *a*. But reference is not a transitive relation, so that there is no real *self*-reference here. Nevertheless, the situation is paradoxical, and as a result some authors ruled out all referential "loops."
- Proposition *a* is a certain token of the form ‘*a* is false’, and *b* is a second token of the same type. Token *a* is self-referential, but token *b* is not, since it refers to *a*, not to itself. Yet some authors thought the two tokens should be treated semantically alike, so that not only the subject of *a* could not refer to *a* itself, neither could the subject of *b*.
- Proposition *a* = ‘*b* is true’, and proposition *b* = ‘*b* is false’. Here, *b* is self-referential, but *a* is not. Nevertheless, *b* is the contradictory of *a*. Hence, by saying its contradictory is true, *a* is in effect saying that it itself is false. Thus, although it is not self-referential, *a* is nevertheless paradoxical. Some authors prevented such cases by maintaining that not only were terms unable to refer to the propositions in which they occurred, they also could not refer to the *contradictories* of the propositions in which they occurred.

As a general theory, restriction is open to an obvious objection: it rules out innocuous forms of reference along with pathological ones. The sentence ‘This sentence has five words’ is not paradoxical, after all, even though it is self-referential; in fact, it seems obviously true. Yet the theory of restriction would disallow it.

Medieval authors sometimes raised this objection. As a result, we find *two* kinds of restriction-theories in the medieval literature: (a) general or strong theories that rule out self-reference, and perhaps other patterns of reference too, in innocuous as well as pathological cases; and (b) more specialized or weaker theories that rule out certain forms of reference only when they result in paradox. Walter Burley and William of Ockham, for example, held the latter form of restriction (Spade, 1974).

If general or strong theories of restriction are open to the objection stated above, the weaker theories are open to a different objection: they are vacuous. The proponents of weaker theories did not have any independent way of identifying paradoxical cases. In practice, their theories amounted to saying "all forms of reference are allowed, except for paradoxical ones, which are not allowed." This is no doubt true, but it is also a tautology.[17]

2.5 Cassation

Unlike restriction, which remained (and remains) a popular view, the theory of "cassation" disappeared very early. It is maintained in the earliest known treatise on insolubles (De Rijk 1966) and in one other anonymous text (Spade 1975, pp. 43-44), but died out after about 1225, although it continued to be mentioned in later authors' surveys of previous views, no doubt because of its inclusion in Bradwardine's own survey.

‘Cassation’ is now an archaic word, but merely means "making null and void, canceling." In effect, this theory holds that one who utters an insoluble proposition "isn't saying anything." The second of the two texts just cited even gives a curious "ordinary language" argument, appealing to the *rusticus* (the man-on-the-street), who, if you were to say to him "What I am saying is false," would reply *"Nil dicis"* ("You are saying nothing").

The treatise in De Rijk 1966 presents more of a theory. Much of it is obscure to modern scholars, but it seems to appeal to a distinction between a mental act of asserting and a vocal act of uttering a proposition. "Saying" requires both acts; it is "an assertion with utterance." In the case of the liar who says "What I am saying is false," the mental act of asserting is present, and for that matter so is the physical act of uttering the words. But somehow (this is the obscure part) there is no "saying."

It is tempting to interpret this view as an appeal to a kind of fallacy of composition; just as someone who is both good and an author is not necessarily a good author, so too something that is both mentally asserted and vocally uttered is not necessarily "said" (asserted with utterance). It is tempting, yes, but highly speculative. Nevertheless, whatever the correct interpretation, it appears that the distinction between asserting and uttering drawn by this theory escapes the facile "refutation" of it used as early as the mid-thirteenth century, that it "plainly contradicts sensation that is not deceived."[18]

3. The Second Quarter of the Fourteenth Century

The preceding theories represent the earliest stage of the *insolubilia*-literature. Although these theories are sometimes mentioned in the later literature, and in the case of "restriction" often *accepted* in the later literature, much more sophisticated treatments began to emerge in the second quarter of the fourteenth century. The turning point is Thomas Bradwardine, whose own theory was enormously influential on later authors. Shortly after Bradwardine, two other authors from this middle period are also important: Roger Swyneshed, and William Heytesbury. A little later, Gregory of Rimini appears to have made important contributions to the discussion as well.

3.1 Thomas Bradwardine

Thomas Bradwardine (c. 1295-1349) wrote his *Insolubles* sometime between 1321 and 1324. It became

one of the most important works on the topic in the Middle Ages. In fact, sometime in the third quarter of the fourteenth century, Ralph Strode, in his own treatise on the topic, surveys the earlier views (quoting Bradwardine's own survey almost verbatim), and then says (Spade 1981, p. 116):

> For the opinions mentioned above were those of the old [logicians], who understood little or nothing about insolubles. After them there arose the prince of modern philosophers of nature, namely Master Thomas Bradwardine. He was the first one who discovered something worthwhile about insolubles.

Bradwardine's theory is built around a theory of truth. He adopts what has been called an "adverbial" theory of propositional signification. (Spade 1996, pp. 178-85.) By virtue of their constituent terms, propositions signify things; but, in addition, the proposition as a whole signifies *that such-and-such is the case.* It is this latter kind of signification that is the basis for Bradwardine's theory of truth.

For him, a proposition is true if and only if it signifies *only* as is the case (*tantum sicut est*), and false if and only if it signifies otherwise than is the case (*aliter quam est*). Note the absence of the ‘only’ in the criterion for falsehood. Truth therefore, is more demanding than falsehood. In order for a proposition to be true, *all* of what it signifies to be the case must in fact be the case; if *any* of what it signifies to be the case *fails* to be the case, the proposition is false.

Furthermore, like many authors, Bradwardine held that what propositions signify *follows* from them. In addition, he seems to have been the first to maintain the converse claim, what has been called "The Bradwardine Principle" (Spade 1981, p. 119-20): whatever follows from a proposition is signified by it. (That is, in more recent terminology, signification is "closed" under the consequence relation.) Combined, these two theses, together with the account of truth given in the preceding paragraph, amount to saying that a proposition is true if and only if whatever follows from it is the case.

This general theory of truth provides a solution to insolubles. For it follows from Bradwardine's semantics as just outlined that *every* proposition signifies that it itself is true.[19] Given this, consider the insoluble case where *a* = ‘*a* is false’. Now *a* signifies that *a* itself is false. We also know that it cannot signify *just* that, but must also signify that *a* is *true,* and in general whatever else follows from *a*.

Proposition *a* cannot be true. If it were, then it would signify only as is the case, and so, since it signifies that it itself is false, it would have to be *false,* not true. But if *a* is false, there is no way to argue, in the other direction, that *a* is true after all. All that follows is that *a* signifies *somehow* otherwise than is the case. And it certainly does, since it signifies that *a* is *true.* The paradox is broken.

Insolubles, then, are simply false. They are false not because of what they signify on the face of it, because that much of what they signify holds. Rather, they are false because *in addition* they also signify that they are true, and that does *not* hold.

This ingenious theory became enormously popular and widespread.[20] But it raises serious and puzzling

questions. For example, does Bradwardine accept "semantic ascent" or not? (Where ‘P’ names the proposition replacing ’*p*’, does he accept ‘*p* ⥽ P is true’ or does he not?) On one hand, since for Bradwardine every proposition signifies that it is true, and since propositions signify just whatever follows from them, it seems that he does accept it. On the other hand, we just saw that where *a* = ‘*a* is false’, we *cannot* infer that if *a* is false then *a* is true, so that it seems he does not accept it in all cases. Such questions lead into Bradwardine's theory of contradiction and his propositional logic, which are beyond the scope of this article.[21]

3.2 Roger Swyneshed

Sometime between roughly 1330 and 1335, the English Benedictine Roger Swyneshed adopted a theory in some respects reminiscent of Bradwardine's, but with interesting features of its own. Like Bradwardine, Swyneshed held that for a proposition to be true, it is not enough that it "signify as is the case." But whereas Bradwardine maintained that in addition the proposition must not signify *otherwise* than is the case (that is, it must signify *only* as is the case), Swyneshed said that in addition the proposition must not "falsify itself." Insolubles do falsify themselves, and so are false for that reason, even though they signify as is the case. Propositions that falsify themselves are said to be those that are "relevant (*pertinens*) to inferring that they are false."

The notions of "relevance," "self-falsification," and "signifying as is the case" (or "otherwise than is the case") are mysterious ones in Swyneshed's theory and not yet well understood by scholars.[22] But the main historical interest of his theory does not lie there. Rather, it lies in three famous and controversial conclusions he drew from his principles:

- Some false propositions signify as is the case. Insolubles do.[23] Thus, where *a* is the insoluble ‘*a* is false’, *a* is self-falsifying and so false. But it signifies as is the case (namely, that it is false).
- In some valid formal inferences, falsehoods follow from truths. For consider the formally valid inference "The conclusion of this inference is false; therefore, the conclusion of this inference is false." The premise and the conclusion of this inference are two tokens of the same type. But while the conclusion is a self-falsifying insoluble, and so is false, the premise is *not* self-falsifying, and is in fact *true.* (The conclusion of the inference *is* false, on Swyneshed's account.) Here then, a falsehood validly follows from a truth.
- In the case of insolubles, two mutually contradictory propositions are false at the same time. Where *a* = ‘*a* is false’, *a* is insoluble and false. But its contradictory, ‘*a* is *not* false’, Swyneshed claims, is *not* insoluble and is not self-falsifying. Nevertheless, it is false because it signifies otherwise than is the case. The insoluble *a* really *is* false.[24]

Many authors found these conclusions ridiculous, especially the second and third ones. But they had their defenders as well.[25]

Two other features of Swyneshed's theory should be at least mentioned, although our understanding of his view does not yet allow a thorough treatment of them. First, he explicitly holds that while valid

inference does not always preserve truth, it does preserve the property of signifying as is the case. Second, Swyneshed explicitly considers a situation where *a* = ‘*a* does not signify as is the case’, and says that that *a* is neither true nor false in that situation. This is the only known case of a medieval author's actually allowing failure of bivalence for insolubles, even though several authors refer to (and reject) such theories; even Swyneshed allows failure of bivalence only in some instances.[26]

3.3 William Heytesbury

In 1335, the Mertonian logician and philosopher of nature William Heytesbury wrote an important treatise *Rules for Solving Sophisms* (*Regulae solvendi sophismata*).[27] The first of its six chapters is on insolubles. The *Rules* as a whole, and this first chapter in particular, were widely read and commented on, particularly in Italy in the late-fourteenth and fifteenth centuries. Indeed, Heytesbury's theory is a competitor to Bradwardine's as the most influential theory of insolubles in the whole of the Middle Ages.[28]

Heytesbury treated insolubles as paradoxical only *with respect to* certain assumed circumstances (what he calls the *casus* or "case"). For example, the proposition ‘Socrates is saying a falsehood’ is not paradoxical in the abstract, all by itself, but only in contexts where, say, it is Socrates who utters that proposition, the proposition is the *only* proposition Socrates utters (it is not an embedded quotation, for instance, part of some larger statement he is making), and where his proposition signifies just as it normally does. Spoken and written language are thoroughly conventional, for medieval authors, so that the vocal sequence or inscription ‘Socrates is saying a falsehood’ could theoretically signify any way you want. It might, for example, signify that 2 + 2 = 4, in which case it would not be insoluble at all but straightforwardly true.

It is the last condition that is the focal point for Heytesbury's attack. He holds that in the *casus* where Socrates himself says just ‘Socrates is saying a falsehood’ and nothing else, his proposition *cannot*, on pain of contradiction, signify *just* as it normally does ("precisely as its words pretend," as he puts it). If it does signify as it normally does, it must signify some other way as well.

How else might it signify? Heytesbury did not think it was his duty to answer that question. The proposition's additional signification cannot be predicted, given the conventionality of spoken and written language. Depending on what else it signifies, different verdicts about the proposition are appropriate. In short, Heytesbury's strategy is to say, "You tell me *exactly* what Socrates's statement signifies, and I'll tell you first of all whether the case you describe is possible, and if it is, I'll tell you whether his statement is true or false."

This "shift the burden" strategy is a consequence of the fact that Heytesbury views the question of insolubles in the context of the *obligationes*, a highly formalized medieval disputation context that is still not fully understood.[29] But many later authors felt that Heytesbury had simply sidestepped the real theoretical issue, and went on to stipulate what Heytesbury would not: an insoluble's "additional" signification. They held that, in circumstances that make it insoluble, a proposition not only signifies as it

normally does; it also signifies *that is it true*. This "adjustment" to Heytesbury's theory has the effect of combining it with the tradition stemming from Bradwardine.[30] It proved to be an appealing combination.

3.4 Gregory of Rimini

Gregory of Rimini's main writing was done in the 1340s. Although today we know of no text or passage of his that discusses insolubles, there must have been one, because in 1372 Peter of Ailly cites Gregory's theory in some detail and uses it in writing his own treatise on insolubles. (Peter of Ailly 1980.)

Gregory's view relied on the traditional medieval notion (going back to Aristotle's *On Interpretation* 1, 16a3-5) of "mental language," the "language of thought" that underlies and is expressed in spoken and written language.[31] Unlike spoken and written languages, where the signification of words and propositions is thoroughly a matter of convention, signification in mental language is fixed by nature once and for all, the same for everyone. It follows that propositions in mental language can never signify otherwise than they "normally" do. Thus Heytesbury's analysis, according to which insolubles do signify otherwise than they normally do, cannot be applied to propositions formed in mental language. Although Heytesbury himself did not draw this conclusion, it follows from his theory that insolubles cannot be formulated in mental language.

In the absence of any text by Gregory on the topic, we cannot be sure that he reasoned like this from Heytesbury's position. But for whatever reason, he apparently did confine insolubles to spoken and written language; for Gregory there are no insolubles in mental language. An insoluble proposition in spoken or written language corresponds to and expresses *not* the mental proposition one would normally expect on the basis of the usual linguistic conventions, but to a complex and *non*-paradoxical mental proposition.

For example, where *a* is the spoken or written proposition '*a* is false', it corresponds to and expresses the *conjunction* of two mental propositions. The first conjunct signifies that *a* is false. Note that this is not the insoluble *a*, since that was in spoken or written language whereas this proposition is mental. Unlike *a*, this proposition is not self-referential; it refers instead to *a*.

The second conjunct signifies that the *first* conjunct is false. Since the first conjunct signifies that *a* is false, this means that the second conjunct amounts to saying that *a* is *not* false, but rather true.

One way, therefore, of viewing Gregory's theory is to say that he adopted the hybrid view described at the end of Section 3.3 above, the view that combines Heytesbury with Bradwardine, but then moved that whole analysis into mental language. Just as for Heytesbury's theory, insolubles for Gregory do not signify "precisely as their words pretend" (they do not express the mental proposition one would expect from the normal linguistic conventions). Just as for Bradwardine's theory, insolubles for Gregory do signify "as their words pretend" (through the first conjunct of the mental proposition). But they do not signify *precisely* that way; they also signify that they are *true* (through the second conjunct of the mental

proposition).

Given our present knowledge of Gregory's views, this reconstruction must remain speculative.

4. The Late Period

The period of greatest innovation and sophistication in the medieval *insolubilia*-literature was the second quarter of the fourteenth century. After about 1350, little original work was done. Insolubles continued to be discussed, but it seems that for the most part the theories adopted were variations or elaborations of the ones already seen. This period is not yet well researched, however, so it is too early for a final verdict.

4.1 John Wyclif

At any event, it is clear that one of the main (and one of the few genuinely new) theories to emerge from this late period is that of John Wyclif, who wrote a *Summa of Insolubles* (*Summa insolubilium*), probably in the early 1360s,[32] and included another discussion of insolubles in his *Continuation of the Logic* (*Logicae continuatio*), III.8. The theory is essentially the same in the two treatments.

For Wyclif, the key to resolving insolubles is to recognize various senses in which propositions can be true or false. There are three main senses of ‘true’, and accordingly of ‘false’:

- In the transcendental sense, truth is convertible with being, so that *any* proposition is true in this sense, no matter what it signifies. This sense can be disregarded in discussing insolubles. Nothing (that is, no being) is false in the sense of failing to be true in this sense.
- In a second sense, a proposition is true if and only if what it "primarily signifies" exists. These "primary significates" are neither substances nor accidents, but rather "beings of reason." It is perhaps plausible to interpret an existing primary significate as analogous to a "fact" in the modern philosophical sense. A proposition is false in this second sense if and only if its primary significate fails to exist.
- In a third sense, a proposition is true if and only if what it primarily signifies exists and is *independent* of the proposition itself. It is false in this third sense if and only if its primary significate either fails to exist or else exists but depends on the proposition itself.

The "independence" required by the third kind of truth is an obscure and difficult matter, not yet well understood. But here is how it applies to insolubles:

Where *a* = ‘*a* is false’, its primary significate either exists or does not. If it does, then in any event it is not independent of *a* in the sense required by the third kind of truth. In either case, then, *a* will be *false* in the third sense. If the word ‘false’ in *a* is taken in the third sense, therefore, *a*'s primary significate does exist, since it is a fact that *a* is false in the third sense. In short, the insoluble is true in the second sense,

but false in the third sense.

Our present understanding of Wyclif's theory does not go much beyond this. Many questions and problems remain. For instance, if the word ‘false’ in *a* is not taken in the third sense but in the *second,* the paradox seems to emerge all over again in a form that cannot be handled by this theory.

Whatever its virtues or defects, Wyclif's theory had some influence on later authors. Robert Alyngton's own *Insolubilia,* for instance, from around 1380, explicitly appeals to Wyclif's theory. Its influence can also be seen in an anonymous late treatise preserved in a Prague manuscript. (See Wyclif 1984, pp. xxiv-xxv.)

4.2 Peter of Ailly

As already mentioned (Section 3.4 above), in 1372 the Frenchman Peter of Ailly (Petrus de Alliaco) wrote an *Insolubilia* that preserves all we know of Gregory of Rimini's theory. Peter's theory looks much like Gregory's. Nevertheless, he did not accept Gregory's view entirely. Whereas for Gregory, an insoluble in spoken or written language corresponds to or expresses a *conjunction* of two propositions in mental language, for Peter it corresponds to or expresses *two distinct* mental propositions, not their conjunction. (The two distinct mental propositions are the same two that Gregory had conjoined.)

In medieval semantics, propositions that correspond to two distinct mental propositions are ambiguous or equivocal. (Indeed, that is the medieval *account* of equivocation.) Thus, for Peter, insolubles in spoken or written language are strictly equivocal and do not have a single signification. In one sense (answering to the first of Gregory's conjuncts), they are true; in another sense (answering to Gregory's second conjunct), they are false. By contrast, for Gregory, insolubles are just false, not ambiguous at all; they correspond to a *single* false conjunction, one conjunct of which is true and the other false.

Peter's theory has the phenomenological advantage that it accounts for the psychological "flip-flop" sense we have when thinking about insolubles. When we look at them one way they seem true; when we look at them another way, they seem false. No other medieval theory seems to account for this psychological fact.

5. Observations

Several instructive observations can be made about the medieval *insolubilia*-literature.

First, although this article has focused on Liar-type paradoxes, and although the medieval literature did too, it also included other kinds of puzzles. For example, where *a* = ‘*b* is false’ and *b* = ‘*a* is false’, no Liar-type paradox arises; contradiction can be avoided by simply taking one of the two propositions as true and the other as false. But medieval logicians regarded such cases as problematic because they require us to assign different truth values to propositions that are semantically exactly alike; there is no

reason to pick *a* as the true proposition rather than *b* or conversely. Cases like this, which violate only a kind of semantic "principle of sufficient reason," were often included under the heading "insolubles." (For example, Buridan, *Sophismata* VIII.8.) A variety of epistemic and pragmatic puzzles were often included as well.[33] There is no attempt, as there sometimes is in present-day literature on the paradoxes, to ignore all the inessentials and focus in on a single paradigmatic case that gets at the kernel of the issue. For medieval authors, the issue was a broad one. They did not attempt to give any precise and rigorous characterization of what it takes to be an insoluble. The definitions they did give are quite general and include much more than Liar-type paradoxes. For instance, Bradwardine defines an insoluble as "a difficult paralogism *secundum quid et simpliciter*[34] arising from some [speech-] act's reflection on itself with a privative determination." (Spade 1975, p. 106.) Still, their discussions always tended to focus on Liar-type cases.

Second, medieval authors did not have any sense of theoretical "crisis" over insolubles, as modern discussions of the paradoxes often do. The medievals did not regard the paradoxes as threatening the very foundations of reasoning. On the contrary, most authors seem to have regarded them as merely argumentative nuisances, and their main concern was to come up with a way of dealing with them when they arise in disputation. No doubt this difference is due to the different contexts in which the discussions emerged. Modern logic is a formalized, axiomatic (or at any rate systematized) discipline, closely tied to the foundations of mathematics; medieval logic, by contrast, was much looser and informal (which of course is not to say it lacked insight), much more tied to the give and take of live academic disputation.

Third, and related to the second point, most medieval authors thought it was entirely possible to find a completely satisfactory "solution" to insolubles. There was no deep "lesson" to be learned about the nature of language or thought, about the limits of expressibility. Insolubles were thought of as resting on a straightforward but pernicious fallacy, although authors disagreed over just what the fallacy is. William of Ockham, for instance, writes, "As for insolubles, you should know it is not because they can in no way be solved that some sophisms are called insolubles, but because they are solved *with difficulty*." (Ockham, *Summa logicae* III-3, 46.)

The only medieval author who is known to have departed from this confident view is William Heytesbury, who raises objections against his own view, and then remarks (Heytesbury 1979, p. 45):

> Many objections of this sort can be raised against this view, which it would be difficult *or impossible* to answer to complete satisfaction.

Again, about his own view he says (p. 21):

> I do not claim that it or any [opinion] is altogether satisfactory, *because I do not see that this is possible.* Nevertheless I rate this one among all of them to be nearer the truth.

Richard Lavenham, an English contemporary of Wyclif, perhaps put the prevailing optimism best (Spade 1975, p. 93; Heytesbury 1979, p. 8):

> Just as the bond of love is sometimes called insoluble, not because it can in no way be untied (*sit solubilis*) but because it can be untied [only] with difficulty, so a proposition is sometimes called insoluble, not because it is not solvable but because it is solvable [only] with difficulty.

Bibliography

Primary Literature in Translation

- Anonymous (1964). *Treatise on Insolubles.* In *Peter of Spain: Tractatus syncategorematum and Selected Anonymous Treatises.* Joseph P. Mullally, trans. "Mediaeval Philosophical Texts in Translation," vol. 13. Milwaukee, Wis.: Marquette University Press, 1964. Among the "anonymous treatises" translated at the end of this volume is a late (probably fifteenth century) treatise on insolubles.
- Athenaeus of Naucratis (1927-41). *The Deipnosophists, with an English translation by Charles Burton Gulick.* "The Loeb Classical Library." Cambridge, Mass.: Harvard University Press, 1927-41.
- Buridan, John (1966). *Sophisms on Meaning and Truth.* Theodore Kermit Scott, trans. New York: Appleton-Century-Crofts, 1966. A translation of the ninth treatise of Buridan's *Sophismata.* Chap. 8 is on insolubles.
- Buridan, John (1982). *John Buridan on Self-Reference.* G. E. Hughes, ed. and trans. Cambridge: Cambridge University Press, 1982. A translation, with philosophical commentary, of Chap. 8 of Buridan's *Sophismata*, on insolubles. (The *Sophismata* is the ninth treatise in Buridan's *Summulae de Dialectica.*) [Note: There are two versions of this book, with different pagination. The paperbound publication has the subtitle: *Chapter Eight of Buridan's 'Sophismata', translated with an Introduction and a philosophical Commentary.* The hardbound publication includes a Latin edition, and has the slightly different subtitle: *Chapter Eight of Buridan's 'Sophismata', with a Translation, an Introduction, and a philosophical Commentary.*]
- Buridan, John (forthcoming). *Summulae de Dialectica.* Gyula Klima, trans. "Yale Library of Medieval Philosophy." New Haven, Conn.: Yale University Press, forthcoming.
- Heytesbury, William (1979). *On "Insoluble" Sentences: Chapter One of His Rules for Solving Sophisms.* Paul Vincent Spade, trans. "Mediaeval Sources in Translation," vol. 21. Toronto: Pontifical Institute of Mediaeval Studies, 1979.
- Peter of Ailly (1980). *Concepts and Insolubles: An Annotated Translation.* Paul Vincent Spade, trans. "Synthese Historical Library," vol. 19. Dordrecht: D. Reidel, 1980.

Primary Literature in the Original Languages

- Adam of Balsham (1956). *Adam Balsamiensis Parvipontani Ars Disserendi (Dialectica Alexandri).* Lorenzo Minio-Paluello, ed. "Twelfth Century Logic: Texts and Studies," vol. 1.

Rome: Edizioni di storia e letteratura, 1956.

- Anonymous (1971). *Insolubilia.* In Paul Vincent Spade, "An Anonymous Tract on *Insolubilia* from Ms Vat. lat. 674: An Edition and Analysis of the Text," *Vivarium* 9 (1971), pp. 1-18. The text comes from 1368.
- Augustine (1970). *Contra academicos, De beata vita, De ordine, De magistro, De libero arbitrio,* W. M. Green and K.-D. Daur, ed. "Corpus Christianorum." Turnholt: Brepols, 1970.
- Bricot, Thomas (1986). *Tractatus insolubilium.* E. J. Ashworth, ed. "Artistarium," vol. 6: Nijmegen: Ingenium, 1986. Bricot was a very late fifteenth century author.
- Buridan, John (1977). *Sophismata.* T. K. Scott, ed., "Grammatica Speculativa," vol. 1: Stutgart-Bad Cannstatt, 1977. The Latin edition that formed the basis for Scott's translation in Buridan (1966). A new critical edition is being prepared by Fabienne Pironet.
- Cicero (1966). *Academica.* James S. Reid, ed. Hildesheim: Georg Olms, 1966. (Photoreprint of the London 1885 edition.)
- Duns Scotus, John (1958). *Quaestiones super libro elenchorum.* In his *Opera omnia.* Luke Wadding, ed. Hildesheim: Georg Olms, 1958. (photoreprint of the edition Lyon: Laurentius Durand, 1639), vol. 1, pp. 268-69.
- Fland, Robert (1978). *Insolubilia.* In Paul Vincent Spade, "Robert Fland's *Insolubilia:* An Edition, with Comments on the Dating of Fland's Works," *Mediaeval Studies* 40 (1978), pp. 56-80. Fland wrote between 1335 and about 1370.
- Gellius, Aulus (1968). *Noctes Atticae.* P. K. Marshall, ed. 3 vols. Oxford: Clarendon Press, 1968.
- John of Holland (1985). *Four Tracts on Logic: Suppositiones, Fallacie, Obligationes, Insolubilia.* E. P. Bos, ed. "Artistarium," vol. 5. Nijmegen: Ingenium, 1985. Includes a Latin edition of a widely read text on insolubles.
- Laertius, Diogenes (1964). *Vitae philosophorum.* H. S. Long, ed. 2 vols. "Oxford Classical Tests." Oxford: Clarendon Press, 1964.
- Neckham, Alexander (1967). *De naturis rerum libri duo, with the Poem of the Same Author, De laudibus divinae sapientiae.* Thomas Wright, ed. "Rerum Britannicarum Medii Aevi Scriptores, or Chronicles and Memorials of Great Britain and Ireland during the Middle Ages (Rolls Series)," no. 34. London: Longman, Green, 1863. (Reprint Kraus reprints, 1967.)
- Ockham, William of (1974). *Summa logicae.* Gedeon Gál *et al.*, ed. St. Bonaventure, NY: The Franciscan Institute, 1974.
- Roure, Marie Louise Roure (1970). "La problématique des propositions insolubles au XIIIe siècle et au début du XIVe, suivie de l'édition des traités de W. Shyreswood, W. Burleigh et Th. Bradwardine." *Archives d'histoire doctrinale et littéraire du moyen age* 37 (1970), pp. 205-326. Latin editions of the treatises by Walter Burley and Thomas Bradwardine, as well as of a thirteenth century treatise sometimes attributed to William of Sherwood.
- Seneca (1965). *L. Annaei Senecae Ad Lucilium epistulae morales.* L. D. Reynolds, ed. 2 vols. Oxford: Clarendon Press, 1965.
- Swyneshed, Roger (1979). *Insolubilia.* In Paul Vincent Spade, ""Roger Swyneshed's *Insolubilia:* Edition and Comments," *Archives d'histoire doctrinale et littéraire du moyen age* 46 (1979), 177-220. Reprinted in Spade (1988).
- Wyclif, John (1893-99). *Tractatus de logica.* Michael Henry Dziewicki, ed. 3 vols. London: Trübner & Co., for the Wyclif Society, 1893-99. Includes his *Logicae continuatio.*

- Wyclif, John (1984). *Summa insolubilium.* Paul Vincent Spade and Gordon Anthony Wilson, ed. "Medieval & Renaissance Texts & Studies," vol. 41. Binghamton, N.Y.: Medieval & Renaissance Texts & Studies, 1984.

Secondary Literature

- Bottin, Francesco (1976). *Le antinomie semantiche nella logica medievale.* Padua: Antonore, 1976.
- Bottin, Francesco (1983). "The Mertonians' Metalinguistic Science and the *Insolubilia.*" In P. Osmund Lewry, ed., *The Rise of British Logic.* "Papers in Mediaeval Studies," vol. 7. Toronto: Pontifical Institute of Mediaeval Studies, 1983, pp. 235-48.
- De Rijk, L. M. (1962-67). *Logica Modernorum: A Contribution to the History of Early Terminist Logic.* Vol. 1: *On the Twelfth Century Theories of Fallacy.* Vol. 2: *The Origin and Early Development of the Theory of Supposition.* Assen: Van Gorcum, 1962-67. (Also contains many Latin editions of primary texts.)
- De Rijk, L. M. (1966). "Some Notes on the Mediaeval Tract *De insolubilibus,* with the Edition of a Tract Dating from the End of the Twelfth Century." *Vivarium* 4 (1966), pp. 83-115.
- Martin, Christopher J. (1993). "Obligations and Liars." In Stephen Read, ed., *Sophisms in Medieval Logic and Grammar.* "Nijhoff International Philosophy Series," vol. 48. Dordrecht: Kluwer, 1993, pp. 357-81.
- Mates, Benson (1961). *Stoic Logic.* Berkeley/Los Angeles, Cal.: University of California Press, 1961. (Originally published in 1953 as vol. 26 of the series "University of California Publications in Philosophy.")
- Spade, Paul Vincent (1973). "The Origins of the Mediaeval *Insolubilia*-Literature." *Franciscan Studies* 33 (1973), pp. 292-309. Reprinted in Spade (1988).
- ------ (1974). "Ockham on Self-Reference." *Notre Dame Journal of Formal Logic* 15 (1974), pp. 298-300. Reprinted in Spade (1988).
- ------ (1975). *The Mediaeval Liar: A Catalogue of the Insolubilia-Literature.* "Subsidia Mediaevalia," vol. 5. Toronto: Pontifical Institute of Mediaeval Studies, 1975.
- ------ (1976). "William Heutesbury's Position on ‘Insolubles’: One Possible Source," *Vivarium* 14 (1976), pp. 114-20. Reprinted in Spade (1988).
- ------ (1978). "John Buridan on the Liar: A Study and Reconstruction," *Notre Dame Journal of Formal Logic* 19 (1978), pp. 579-90. Reprinted in spade (1988).
- ------ (1981). "*Insolubilia* and Bradwardine's Theory of Signification," *Medioevo: Revista di storia della filosofia medievale* 7 (1981), pp. 115-34. Reprinted in Spade (1988).
- ------ (1982a). *"Insolubilia."* In Norman Kretzmann, *et al.*, ed. *The Cambridge History of Later Medieval Philosophy*. Cambridge: Cambridge University Press, 1982. Chap. 12, pp. 246-53.
- ------ (1982b.) "Three Theories of *Obligationes:* Burley, Kilvington and Swyneshed on Counterfactual Reasoning," *History and Philosophy of Logic* 3 (1982), pp. 1-32.
- ------ (1983). "Roger Swyneshed's Theory of *Insolubilia:* A Study of Some of his Preliminary Semantic Notions." *History of Semiotics.* Achim Eschbach and Jürgen Trabant, ed. "Foundations of Semiotics," vol. 7. Amsterdam: John Benjamins, 1983. Reprinted in Spade (1988).
- ------ (1987). "Five Early Theories in the Mediaeval *Insolubilia*-Literature." *Vivarium* 25 (1987),

pp. 24-46.

- ------ (1988). *Lies, Language and Logic in the Later Middle Ages.* "Variorum Collected Studies Series"; London: Variorum Reprints, 1988. A collection of seventeen previously published papers, seven of them in insolubles.
- ------ (1991). "Richard Brinkley's *De insolubilibus:* A Preliminary Assessment," *Rivista di storia della filosofia* 46 (1991), pp. 245-56. Brinkley was a contemporary of Wyclif.
- Stock, St. George (1908). *Stoicism.* London: Archibald Constable, 1908.

Other Internet Resources

- Spade, Paul Vincent. *Thoughts, Words and Things: An Introduction to Late Mediaeval Logic and Semantic Theory*, Version 1.0 (July 1, 1996), (Adobe PDF format).
- Spade, Paul Vincent. "Three Questions by John of Wesel on *Obligationes* and *Insolubilia.*" (Adobe PDF format.) Includes a Latin edition of the first of his five questions on insolubles, which discusses Swyneshed's second conclusion.

Related Entries

Alyngton, Robert | Buridan, John [Jean] | Burley [Burleigh], Walter | dialetheism [dialethism] | Duns Scotus, John | fallacies: medieval theories of | Gregory of Rimini | Heytesbury, William | logic: paraconsistent | medieval philosophy | Ockham [Occam], William | Paul of Venice | *sophismata* [= sophisms] | truth: revision theory of | Wyclif, John

Paul Vincent Spade
spade@indiana.edu

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Notes to Insolubles

Notes

1. Singular *insolubile,* with the stress on the antepenult.

2. "Eubulides the Milesian belongs to Euclides's school. He raised many arguments in dialectic: the Liar, the Unnoticed, Electra, the [Man] in a Veil, the Heap, the Horns, the Bald Head."

3. This delightful although rather free translation comes from Stock, 1908, p. 36. It is quoted in Mates 1961, p. 42.

4. Gellius's wording makes it likely that he intended a situation more like the one Aristotle discusses in *Sophistic Refutations* 25 (where someone first tells a lie, but then takes it back by revealing that his first statement was a lie) than a real Liar-type paradox. (For the diffeence, see Spade 1973 and section 1.3 below.) Still, Gellius's words might have been enough to suggest the Liar paradox to a medieval logician, if only they had been read.

5. Spade 1973, p. 296 n. 24. It is sometimes observed that Epimenides's statement is not really a paradox of the Liar type at all. If he did say "The Cretans [i.e., *all* Cretans] are *always* liars, evil beasts, slow bellies," then his statement is true only if it is false, since he himself is a Cretan and so always lying. Thus it cannot be true. But if it is false, all that follows is that *some* Cretans are not *always* liars and evil beasts and slow bellies. It does *not* follow that Epimenides's own remark is not a lie, and so it does not follow that it is not false. Hence there is no real paradox at all; Epimenides's statement is just false. This much is correct about what *Epimenides* said. But *St. Paul* (or whoever wrote the epistle) goes on to say in the very next verse (Titus 1:13) "This witness is *true*." Hence there was ample opportunity for a reader of this text to be introduced to the Liar Paradox, even if he did not already know about it.

6. For details of the points made in this paragraph, see Spade 1973.

7. Particularly the second proposition, with the important words ‘nothing but’. These words indicate a proposition that cannot be cast in the mold of Aristotle's oath-breaker.

8. It is worth noting that Neckham had studied at the Petit Pont, at the logical school founded by Adam of Balsham.

9. De Rijk 1962-67, II.2, p. 594:30-31. The *Munich Dialectic's* "treatise on insolubles" is now lost. This

occurrence of the word ‘insolubles’ is the first known use of the term in its technical sense.

10. Thus, the third view in Bradwardine's survey tries to fit the paradox into the Aristotelian fallacy of "false cause." The seventh view, interestingly enough from a present-day perspective, denies bivalence for insolubles; they are neither true nor false. Certain authors after Bradwardine likewise refer to the latter view, although no text has been found, from before or after Bradwardine, that actually maintains it as a general claim about insolubles. (But see n. 12 below. In the first half of the 1330s, Roger Swyneshed did allow failure of bivalence in certain special cases of insolubles. See Spade 1983, and section 3.2 below.)

11. For references, and for another example, see Spade 1987, p. 32 n. 46.

12. It is possible that these are the authors Bradwardine refers to as denying bivalence. (See n. 10 above.) But this is uncertain.

13. For details of claims in this paragraph, see Spade 1987, pp. 32-33.

14. Medieval logicians typically held that affirmative propositions with non-denoting subject terms are false, not meaningless or without truth value, as is sometimes held nowadays.

15. For more on this theory, and for some very tentative speculations on its motivation, see Spade 1987, pp. 33-36.

16 . Not surprisingly, given that the discussion occurs in a set of questions on the *Sophistic Refutations*, Scotus applies this distinction within the context of the Aristotelian fallacy *secundum quid et simpliciter*. Bradwardine interprets the theory in a way that commits it to *transcasus* as well, although Scotus's own text does not seem committed to that. It may be that Bradwardine is not thinking of Scotus in particular when he mentions this theory, and in any event Scotus himself makes it plain that the theory is not original with him. For more on the theory, see Spade 1987, pp. 36-38.

17. This objection does not seem to have been raised in the Middle Ages, even by authors like Bradwardine who opposed all forms of restriction. (See Spade 1975, p. 106.) This is probably to be explained by the fact that medieval authors on the whole do not seem especially concerned with a theoretical understanding of the paradoxes; they are much more concerned with knowing what to do with them when they arise in argumentation. See section 5 below. On the theory of restriction, see Spade 1987, pp. 38-42.

18. Later authors express this objection more clearly, if no more successfully. Bradwardine, for example, remarks that Socrates says letters (i.e., phonemes), syllables, words and a sentence, and so does not say "nothing." The objection, therefore, confuses "saying" with "uttering." (See Spade 1975, item LXIV, p. 107.) On the theory of cassation, see Spade 1987, pp. 43-45.

19. Bradwardine himself does not state this general claim, but it is easy enough to show that it follows from his semantics. (See the proof in the supplementary document A Proof Concerning Bradwardine's Theory.) The general claim greatly simplifies his own complicted presentation of his theory. See Spade 1981, especially pp. 123-25. A great many later authors did explicitly maintain the general claim.

20. Variations on Bradwardine's theory, sometimes combined with elements from other views, were held by several anonymous authors (see Spade 1975, items IV, VIII and XII), as well as by Albert of Saxony, Henry Hopton, John Buridan, John of Holland, John Huntman (*or* Venator), Paul of Pergula, Ralph Strode, Richard Lavenham, and Robert Fland. See Spade 1975, and Spade 1981, p. 123 n. 32.

21. For a discussion of these matters, see Spade 1981, especially pp. 125-34.

22. For a discussion of these notions and some of the complications with them, see Spade 1983.

23. Bradwardine would have agreed with this conclusion as stated. But whereas for Bradwardine insolubles *also* signify otherwise than is the case, Swyneshed does not seem to have a notion of any *additional* signification of a proposition. Thus, his first conclusion amounts to saying that some false propositions signify *only* as is the case. Bradwardine certainly would not have agreed with *that*.

24. Note that Swyneshed is here taking the notion of "contradictories" as a syntactical notion, so that it in effect means "a proposition and its negation." Other authors had an Aristotelian, semantic notion of contradictories, according to which contradictories are propositions that cannot be either true together or false together but must have opposite truth values. But Swyneshed's third conclusion is striking, whether put in terms of "contradictories" or in terms of a proposition and its negation.

25. For some of the parties in the controversy, see Spade 1975, the anonymous item III, as well as the entries there for Anthony de Monte, John of Wesel, Paul of Pergula, the *Logica magna* attributed to Paul of Venice, Robert Fland, Roger Roseth, and William Heytesbury.

26. If we push the point, and ask not whether *a* in this situation is true or false, but whether it signifies as is the case or does not, Swyneshed explicitly says it does not. Nevertheless, it is not clear what prevents our paradoxically inferring from this that, yes, *a* does signify as is the case after all. In other words, what is to prevent our *reconstructing*, in terms of the notion of "signifying as is the case" alone, the paradoxes that arise about truth when truth is identified with signifying as is the case? Swyneshed does discuss the issue, but his reasoning is obscure. (See Spade 1983, pp. 107-08.)

27. ‘Sophisms’ here does not mean "sophistry" in the modern, pejorative sense, but rather puzzle-propositions the study of which illustrates various logical points. Good modern examples might be Frege's "The morning star is the evening star," or Russell's "The present king of France is bald."

28. Heytesbury may win the competition. See Spade 1975, the anonymous items V, VII, VIII, XII, XIII, XXIII, and the entires there for Angelo of Fossombrone, Gaetano di Thiene, John of Constance, John Dumbleton, John of Holland, John Hunter (Huntman, Venator), John of Wesel, John Wyclif, Paul of Pergula, the *Logica magna* attributed to Paul of Venice, Ralph Strode, and Robert Fland.

29. For a discussion of some of the possibilities, see Spade 1982b.

30. See Spade 1975, the anonymous items VII and XII, and the entries there for John of Holland and John Hunter (Huntman, Venator).

31. On the notion of "mental language," see Spade, *Thoughts, Words and Things*.

32. There is some dispute over the authenticity of this work. Still, the theory is genuinely Wyclif's, even if this text is not. See the discussion in Wyclif 1984, xxiii-xxviii.

33. For example, Buridan, *Sophismata* VIII.13, concerns the ingenious proposition ‘Socrates knows the proposition written on the wall to be doubtful to him’, where it is supposed that this proposition is the only proposition written on the wall, that Socrates sees it and is in a state of doubt about its truth, and furthermore *knows* he is in that state of doubt. Is the proposition true or not? Since it is stipulated that Socrates does know that he doubts the proposition, it would seem to be true. But how can Socrates simultaneously know and doubt the same proposition? Again (VIII.18), Socrates wants to eat if and only if Plato wants to eat, since Socrates likes company at meals. But Plato is angry at Socrates and, out of perverse spite, wants to eat if and only if Socrates does *not* want to eat. Does Socrates want to eat or not? The literature abounds in such delightful examples.

34. Despite this lip-service to the Aristotleian fallacy, we saw in section 3.1 that the core of Bradwardine's solution is quite different.

Paul Vincent Spade
spade@indiana.edu

First published: August 27, 2001
Content last modified: August 27, 2001

Robert Alyngton

Robert Alyngton was one of the most important authors of the generation after John Wyclif. He was deeply influenced by Walter Burley's logico-ontological system and Wyclif's metaphysics. (His major extant work, a commentary on the *Categories,* heavily depends on Burley's last commentary on the *Categories* and Wyclif's *De ente praedicamentali.*) Yet he was able to develop new logical and semantic theories as well as the general strategy adopted by the Oxford Realists, as he methodically substituted reference to external objective realities for reference to linguistic and/or mental activities.

Life and Works

Not a great deal is known of Robert Alyngton's life. Most of the information about him comes from Emden 1957-59. From 1379 until 1386, he was fellow of Queens College (the same Oxonian college where Wyclif started his theological studies in 1363 and Johannes Sharpe taught in the 1390s); he became *Magister Artium* and, by 1393, doctor of theology. He was chancellor of the University in 1393 and 1395. In 1382 he preached Wyclif's religious and political ideas in Hampshire (McHardy 1987). He was rector of Long Whatton, Leicestershire, where he died by September 1398.

According to Emden 1957-59 and Ashworth & Spade 1992, Alyngton was of considerable repute as a logician. Among his extant works, the following can be mentioned (the most complete list of his writings is found in Bale 1557-59 [pp. 519-20]):

- A commentary on Aristotle's *Categories* (*Litteralis sententia super Praedicamenta Aristotelis*

[henceforward *In Cat*.]), partially edited in Conti 1993 (pp. 242-306). (All references are to the pages of this edition or, for the unedited portions, to the ms. London, Lambeth Palace 393.)

- A treatise on the supposition of terms (*Tractatus de suppositionibus terminorum*).
- A commentary on the *Liber sex principiorum*.
- A treatise on the genera of being (*Tractatus generum*).

Being and Categories

Like Burley, Alyngton affirms that (i) the division into categories is first of all a division of things existing outside the mind, and only secondarily of the mental concepts and spoken or written terms which signify them, and (ii) things belonging to one categorial field are really distinct from those belonging to another -- for instance, substances are really distinct from quantities, qualities, and relations, quantities are really distinct from substances, qualities, and relations, and so on (*In Cat*., Conti pp. 251, 252-53).

As far as the problem of the relationship of the ten categories to being is concerned, Alyngton does not follow Burley but Wyclif, since he hypostatises the notion of being and considers equivocity, analogy, and univocity not only as semantic relations between terms and things, but also as real relations between extrametal objects (*In Cat*., ms. London, Lambeth Palace 393, ff. 69v-70r). According to the common interpretation of the opening passage of the *Categories*, equivocal terms are correlated with more than one concept and refer to a multiplicity of things sharing different natures, whereas univocal terms are correlated with only one concept and refer to a multiplicity of things sharing one and the same nature. Within Alyngton's system, what differentiates analogy from univocity is the way in which a certain nature (or property) is shared by a set of things: analogous things share it according to different degrees (*secundum magis et minus*, or *secundum prius et posterius*), univocal things share it all in the same manner and to the same degree (*In Cat*., pp. 255-256). Alyngton admits four main types of equivocity: by chance, deliberate, analogical, and generic. Equivocals by chance are those things to which it just happens that they have the same name, but with different meanings and/or reasons for imposing the name. Those things are deliberate equivocals which have distinct natures but the same name, and are subordinated to different but correlated concepts. Those things are analogical which share the nature signified by their common name in various degrees and/or ways. Generic equivocals are those things which share the same generic nature in the same way, but have distinct specific natures of different absolute value (*In Cat*., f. 70r). According to this account, being is a sort of basic component of the metaphysical structure of each reality, which posses it in accordance with its own nature, value, and position in the hierarchy of created beings.

Universals and Predication

Alyngton recognizes three main kinds of universals:

1. *ante rem* or ideal universals -- that is, the ideas in God, the archetypes of all that is;

2. *in re* or formal universals -- that is, the common natures shared by individual things; and
3. *post rem* or intentional universals -- that is, mental signs of the formal universals.

The ideas in God are the causes of formal universals, and formal universals are the causes of intentional universals. Furthermore, like Burley and Wyclif, Alyngton holds that formal universals actually exist (*in actu*) outside our minds, and not potentially only (*in potentia*) as moderate realists thought (*In Cat*., p. 279) -- even if, unlike Burley (the *Doctor Planus et Perspicuus*), he maintains they are really identical with their individuals, for otherwise it would be impossible to explain, against the Nominalists, why and how individual substances show different and more or less close kinds of similarity among themselves (*In Cat*., pp. 267-68).

According to Alyngton, who depends here on Avicenna and Wyclif, formal universals are common natures in virtue of which the individuals that share them are exactly what they are -- as the human species is the form by which every man formally is a man. *Qua* natures, they are prior, and so indifferent, to any division into universals and individuals. Universality (*universalitas* or *communicabilitas*) is as it were their inseparable property, but not a *constitutive* mark of the nature itself (*In Cat*., f. 101v). As a consequence, formal universals can be conceived of in two different manners: as first intentions or as second intentions. In the former case, they are natures of a certain kind and are identical with their individuals (for example, man is the same thing as Socrates). In the latter case, they are properly universals (that is, something that can exist in many things and can be shared by them), and are distinct from their individuals considered *qua* individuals, because of opposite constitutive principles (*In Cat*., p. 268). Therefore, universals are really (*realiter*) identical to, but formally (*formaliter*) distinct from their individuals. In fact, universals are formal causes in relation to their individuals, and individuals are material causes in relation to their universals. Thus three different kinds of entities can be qualified as formal universals:

1. the common natures instantiated by individuals -- which are things of first intention;
2. the form itself of universality, which belongs to a certain common nature when seen in its relation to the individuals -- which is a thing of second intention; and
3. the thinkability proper to the common nature, by which it is a possible object of our mind -- that is, the real principle that connects formal universals with mental universals (*In Cat*., p. 277).

Alyngton accepts the traditional realistic account of the relationship between formal universals and individuals, and, like Wyclif, improves it by defining its logical structure more accurately. Alyngton thought that a universal of the category of substance could directly receive only the predications of substantial forms more common than itself. On the other hand, accidental forms inhering in substantial individuals could be predicated only indirectly (*essentialiter*) of the substantial form itself that those individuals instantiate, predicated indirectly through and in virtue of the individuals of that substantial form. So his description of the logical structure of the relationship between universals and individuals demanded a redefinition of predication. Alyngton was probably the first to ameliorate Wyclif's theory of predication by dividing predication into formal predication (*praedicatio formalis*) and remote inherence (*inhaerentia remota*) or predication by essence (*praedicatio secundum essentiam*). Remote inherence is grounded in a partial identity between subject and predicate, which share some but not all metaphysical

constituents, and does not demand that the form signified by the predicate term be directly present in the entity signified by the subject term. On the other hand, such a direct presence is needed by formal predication. "Man is an animal" and "Socrates is white" are instances of formal predication; "(What is) singular is (what is) common" (*singulare est commune*) and "Humanity is running" (*humanitas est currens*) are instances of remote inherence, as according to Alyngton it is possible to attribute the property of being running to the form of humanity if at least one man is running. However, he makes sure to use as a predicate term a substantival adjective in its neuter form, because only in this way can it be made apparent that the form signified by the predicate term is not directly present in the subject, but is indirectly attributed to it, through its individuals (*In Cat*., pp. 288-90).

The Theory of Relations

Aristotle's treatment of relations in the *Categories* and in the *Metaphysics* is opaque and incomplete. Because of this fact, in Late Antiquity and the Middle Ages many authors tried to reformulate the doctrine of relatives. Alyngton's attempt is the most interesting among those of the Late Middle Ages, as he was the only one able to work out a concept of relation conceived of as an accidental form which is in both the relatives at once, even if in different ways (*In Cat*., p. 296). Consequently his notion of relation can be considered the ontological equivalent to our modern functions with two variables, or two-place predicates, whereas all other authors of the Middle Ages had thought of relations in terms of monadic functions.

According to Alyngton, whose account partially differs from those of Burley and Wyclif, in the act of relating one substance to another four distinct constitutive elements can be singled out:

1. the relation itself -- for instance, the form of paternity;
2. the substrate of the relation, that is, the substance that denominatively receives the name of the relation -- for instance, the (substance that is the) father;
3. the object of the relation, that is, the substance the substrate of the relation is connected with -- for instance, the (substance that is the) son; and
4. the foundation (*fundamentum*) of the relation, that is, the absolute entity in virtue of which the relation inheres in the substrate and in the object (*In Cat*., p. 299).

The foundation is the main component, since it (i) joins the relation to the underlying substances, (ii) lets the relation link the substrate to the object, and (iii) transmits some of its properties to the relation. Unlike Burley and Wyclif, Alyngton affirms that not only qualities and quantities, but substances too can be the foundation of a relation (*In Cat*., p. 291).

On this basis, Alyngton can define relations of reason while eliminating from their description any reference to our mind and using objective criteria only, based on the framework of reality itself. In fact, he maintains that what characterizes relations of reason is the fulfillment of at least one of these conditions: (i) either the rrelation's subject of inherence or its object is not a substance; (ii) the object is not an actual entity; (iii) the foundation of the relation is not an absolute being -- that is. a substance, a

quantity, or a quality (*In Cat*., pp. 291-92, and 294-95).

The Semantics of Second Intentions

Not until the end of the fourteenth century did anyone claim extramental reality for second intentions, not even Walter Burley. According to him, second intentions are concepts that have a foundation in the extramental world, but are not "things" in the proper sense of the term. This account implied that the keystone of medieval realism, the principle of one-to-one correspondence between language and the world, has to suffer an exception, since no common nature matches second intention terms. It was just in order to do away with this exception that Alyngton (then followed by William Penbygull, Roger Whelpdale, and John Tarteys) hypostatized second intentions, heavily modifying the standard theory of the status of second intentions. In fact, Alyngton not only considers second intentions as objective, but clearly hypostatizes them, speaking of them in terms of real determinations joined to the modes of being of extramental things and directly inhering in them (*In Cat*., pp. 268-69). As a consequence, he conceives of logic as an analysis of the general framework of reality, since according to him logic turns on structural forms (aimed at building up semantic contents), which are, as forms, independent of both such contents and of the mental acts by which they are learned. It is through these forms that the network connecting the basic constituents of the world (individuals and universals, substances and accidents) is disclosed to us (*In Cat*., pp. 278-79). The strategy that supports this choice is evident: as in the case of relations of reason, Alyngton is trying to substitute references to external reality for references to mental activity. In other words, he seeks to reduce epistemology to ontology. From a logical point of view, this means that the same interpretative pattern is employed in order to account for both the semantic power of proper names and common terms (that is, those expressions that refer to a class of individuals), and first and second intentions. Like proper names, common terms also primarily signify and label a unique object -- that is, a common nature. But unlike the object signified by a proper name, the reality of the common nature is distributed among many individuals as their main metaphysical constituent, since it determines the typical features of the individuals themselves. By associating common terms with such objects as their main referent, Alyngton thinks he can explain the fact that a common term can stand for and label many individuals at once. Only in this way does he believe we can grant the value of our knowledge, which otherwise lacks an adequate foundation (*In Cat*., ff. 101v-102v).

Still, this procedure, so strong and powerful, leads to a paradox when applied to terms of second intention by which we speak of singular objects considered as such -- that is, terms (or expressions) like ‘first substance’ (*substantia prima*), ‘individual’ (*individuum*), and so on. In fact, according to Alyngton (and many other Realists of that period), a common term is always matched by a common nature really existing in the world. Therefore, as the term ‘individual’ appears to be common, since it can stand for a multiplicity of things, it should signify an extramental common nature shared by them. As a consequence, we would have to admit the existence of an *individual common nature*, that is a (self-contradictory) entity present in all the individuals as the cause of their being individuals. Alyngton, who would not give up the principle of the one-to-one relation between philosophical language and the world, could remove this paradox only by classifying this kind of term among atomic (*discreti*) terms -- that is, terms or nominal syntagms, like ‘Socrates’ or ‘this man’ (*hic homo*), that refer to individuals and not to

sets of individuals. According to Alyngton, there are three main kinds of atomic terms:

1. personal pronouns, which identify a singular definite referent by means of an ostensive definition (*a demonstratione*);
2. proper names; and
3. range-narrowed expressions (*a limitatione intellectus*) -- that is, expressions, like ‘this man’, that identify a singular referent as a member of a given (manifested) set of individuals. Also expressions like ‘first substance’ and ‘individual’ belong to this third type, since they presuppose a general concept (substance and being, in the example), the range of which is narrowed to a unique object among substances and beings by an act of our intellect -- to one object that is not common (*In Cat*., pp. 270-71).

The rule that terms can be listed as common ones if and only if they signify a common nature is safe, but at the cost of a counterintuitive categorization of their semantic power. In fact, according to Alyngton's account, saying that Socrates and Plato are first substances simply means that (i) each one is what he is, and that (ii) what each one is is a non-universal substance. This is a solution that entails that to be an individual is not a positive state of affairs, but a negative one, and therefore connects Alyngton's ontology with Henry of Ghent's theory of individuation.

Bibliography

Edited works

- *Litteralis sententia super Praedicamenta Aristotelis*, in A.D. Conti, "Linguaggio e realtà nel commento alle *Categorie* di Robert Alyngton," *Documenti e studi sulla tradizione filosofica medievale* 4 (1993), pp. 179-306, at pp. 242-306. Partial edition: only chaps. 2, 3, 4, and 7, and the first section of chap. 5.

Secondary literature

- E.J. Ashworth & P.V. Spade, "Logic in Late Medieval Oxford," in *The History of the University of Oxford*, J.I. Catto & R. Evans eds., vol. 2, Oxford: Clarendon Press, 1992, pp. 50-62.
- J. Bale, *Scriptorum illustrium Maioris Britanniae* ..., 2 vols. in 1, Basle: Joannes Oporinus, 1557-59.
- A.D. Conti, ed., *Johannes Sharpe, Quaestio super universalia*, Firenze: Olschki, 1990. See the "Studio storico-critico," pp. 309-15.
- A.D. Conti, "Linguaggio e realtà nel commento alle *Categorie* di Robert Alyngton," (see "Edited works," above), pp. 179-241.
- A. de Libera, *La querelle des universaux. De Platon à la fin du Moyen Age*, Paris: Éditions du Seuil, 1996. See pp. 403-28.
- L.M. de Rijk, "*Logica oxoniensis*: an Attempt to Reconstruct a 15th Century Manual of Logic,"

Medioevo 3 (1977), pp. 125-55.

- A.B. Emden, *A Biographical Register of the University of Oxford to AD 1500,* 3 vols., Oxford: Clarendon Press, 1957-59, at vol. 1, pp. 30-31.
- Ch.H. Lohr, "Medieval Latin Aristotle Commentaries. Authors: Robertus-Wilgelmus," *Traditio* 29 (1973), pp. 96-97.
- A.K. McHardy, "The Dissemination of Wyclif's Ideas," in A. Hudson & M. Wilks eds., *From Ockham to Wyclif*, Oxford: Blackwell, 1987, pp. 361-62.
- P.V. Spade & G.A. Wilson, eds. *Johannis Wyclif Summa insolubilium*, Binghamton, N.Y.: Medieval & Renaissance Texts & Studies, 1986. See the "Introduction," especially pp. xxii-xlvii.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Burley [Burleigh], Walter | Wyclif, John

Alessandro D. Conti
a.conti@tiscalinet.it

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 25, 2001
Content last modified: July 25, 2001

John Wyclif

John Wyclif (ca. 1330-84) was one of the most important and authoritative thinkers of the Middle Ages. His activity is set in the very crucial period of late Scholasticism, when the new ideas and doctrines there propounded accelerated the transition to the modern way of thought. On the one hand, he led a movement of opposition to the medieval Church and to some of its dogmas and institutions, and was a forerunner of the Reformation; on the other, he was also the most prominent English philosopher of the second half of the 14th century. His logical and ontological theories are, at the same time, the final result of the preceding realistic tradition of thought and the starting-point of the new forms of realism at the end of the Middle Ages, since many authors active during the last decades of the 14th and/or the first decades of the 15th centuries (Robert Alyngton, William Penbygull, Johannes Sharpe, William Milverley, Roger Whelpdale, John Tarteys, and Paul of Venice), were heavily influenced by his metaphysics and largely used his logical apparatus. However, his philosophical system, rigorous in its general design, contains unclear and aporetic points that his followers attempted to remove. So, although an influential thinker, Wyclif *pointed to* the strategy the Realists at the end of the Middle Ages were to adopt, rather than fully developed it.

1. Life and Works

1.1 Life

John Wyclif was born near Richmond (Yorkshire) before 1330 and ordained in 1351. He spent the greater part of his life in the schools at Oxford: he was fellow of Merton in 1356, master of arts at Balliol

in 1360, and doctor of divinity in 1372. He definitely left Oxford in 1381 for Lutterworth (Leicestershire), where he died on 31 December, 1384. It was not until 1374 (when he went on a diplomatic mission to Bruges) that Wyclif entered the royal service, but his connection with John of Gaunt, Duke of Lancaster, probably dates back to 1371. His ideas on lordship and church wealth, expressed in *De civili dominio* (*On Civil Dominion*), caused in 1377 his first official condemnation by the Pope (Gregory XI), who censured nineteen articles. As has been pointed out (Leff 1967), in 1377-78 Wyclif made a swift progression from unqualified fundamentalism to a heretical view of the Church and its Sacraments. He clearly claimed the supremacy of the king over the priesthood (see for instance his *De ecclesia* [*On the Church*], between early 1378 and early 1379), and the simultaneous presence in the Eucharist of the substance of the bread and the body of Christ (*De eucharistia* [*On the Eucharist*], and *De apostasia* [*On Apostasy*], both ca. 1380). His theses would influence Jan Hus and Jerome of Prague in the 15th century. So long as he limited his attack to abuses and the wealth of the Church, he could rely on the support of a (more or less extended) part of the clergy and aristocracy, but once he dismissed the traditional doctrine of transubstantiation, his (unorthodox) theses could not be defended any more. Thus in 1382 Archbishop Courtenay had twenty-four propositions that were attributed to Wyclif condemned by a council of theologians, and could force Wyclif's followers at Oxford University to retract their views or flee. The Council of Constance (1414-18) condemned Wyclif's writings and ordered his books burned and his body removed from consecrated ground. This last order, confirmed by Pope Martin V, was carried out in 1428.

The most complete biographical study of Wyclif is still the monograph of Workman 1926, but the best analysis of his intellectual development and of the philosophical and thelogical context of his ideas is Robson 1961.

1.2 Works

Wyclif produced a very large body of work, both in Latin and English, a great portion of which has been edited by the Wyclif Society between the end of the 19th and the beginning of the 20th centuries, even though some of his most important books are still unpublished -- for instance, his treatises on time (*De tempore*) and on divine ideas (*De ideis*). W. R. Thomson 1983 wrote a full bibliography of Wyclif's Latin writings, among which the following can be mentioned: *De logica* (*On Logic* -- ca. 1360); *De ente in communi* (*On Universal Being* -- ca. 1365); *De ente primo in communi* (*On Primary Being* -- ca. 1365); *Purgans errores circa universalia in communi* (*Amending Errors about Universals* -- between 1366 and 1368); *De ente praedicamentali* (*On Categorial Being* -- ca. 1369); *Tractatus de universalibus* (*Treatise on Universals* -- ca. 1368-69 according to Thomson 1983, but between 1373 and 1374 according to Mueller 1985); *De materia et forma* (*On matter and form* -- between late 1370 and early 1372 according to Thomson 1983, but about 1374-75 according to Mueller 1985). Many of these treatises were later arranged as a *Summa*, called *Summa de ente* (*Summa on Being*), in two books, containing seven and six treatises respectively. (On the genesis, nature, structure, and tasks of this work see Robson 1961, pp. 115-40.)

2. Logic

2.1 Some preliminary remarks

Late medieval Nominalists, like Ockham and his followers, drew a distinction between things as they exist in the extra-mental world and the schemata by means of which we think of and talk about them. While the world consists only of two genera of individuals, substances and qualities, the concepts by which they are grasped and expressed are universal and of ten different types. Nor do the relations through which we connect our notions in a proposition analytically correspond to the real links that join individuals in a state of affairs. Thus, our conceptual forms do not coincide with the elements and structures of reality, and our knowledge does not reproduce its objects but merely *regards* them.

Wyclif maintained that such an approach to philosophical questions was misleading and deleterious. Many times in his works he expressed the deepest hostility to such a tendency. He thought that only on the basis of a close isomorphism between language and the world could the signifying power of terms and statements, the possibility of definitions, and finally the validity and universality of our knowledge be explained and ensured. So the nucleus of his metaphysics lies in his trust in the scheme *object-label* as *the* general interpretative key of every logico-epistemological problem. He firmly believed that language was an ordered collection of signs, each referring to one of the constitutive elements of reality, and that true (linguistic) propositions were like pictures of those elements' inner structures or/and mutual relationships. From this point of view, universals are conceived of as the real essences common to many individual things, which are necessary conditions for our language to be significant. Wyclif thought that by associating common terms with such universal realities the fact could be accounted for that each common term can stand for many things at once and can label all of them in the same way.

This conviction explains the main characteristic of his philosophical style, to which all his contributions can be traced back: a strong propensity towards hypostatisation. Wyclif methodically replaces logical and epistemological rules with ontological criteria and references. He thought of logic as turning on structural forms, independent of both their semantic contents and the mental acts by which they are grasped. It is through these forms that the network connecting the basic constituents of the world (individuals and universals, substances and accidents, concrete properties, like being-white, and abstract forms, like whiteness) is disclosed to us. His peculiar analysis of predication and his own formulation of the Scotistic formal distinction are logically necessary requirements of this philosophical approach. They are two absolute novelties in late medieval philosophy, and certainly the most important of Wyclif's contributions to the thought of his times.

Wyclif's last formulation of the theory of difference and his theory of universals and predication are linked together, and rest upon a sort of componential analysis where things substitute for lexemes and ontological properties for semantic features. Within Wyclif's world, difference (or distinction) is defined in terms of partial identity, and is the main kind of transcendental relation holding among the world's objects, since in virtue of its metaphysical composition everything is at the same time partially identical to and different from any other. When the objects at issue are categorial items, and among what differentiates them is their own individual being, the objects differ *essentially*. If the objects share the

same individual being and what differentiates them is (at least) one of their *concrete* metaphysical components (or features), then the objects differ *really*, whereas if what differentiates them is one of their *abstract* metaphysical components, then they differ *formally*. Formal distinction is therefore the tool by means of which the dialectic of one-many internal to the world's objects is regulated. It explains why one and the same thing is at the same time an atomic state of affairs and how many different beings can constitute just one thing.

2.2 The formal distinction

Wyclif explains the notion of formal distinction (or difference) in the *Purgans errores circa universalia in communi* (chap. 4, p. 38) and in the later *Tractatus de universalibus*. (On Wyclif's formulation of the formal distinction see Spade 1985, pp. xx-xxxi, and Conti 1997, pp. 158-63.) The two versions differ from each other on some important points, and are both unsatisfactory, since Wyclif's definitions of the different types of distinction are rather ambiguous.

In the *Tractatus de universalibus* (chap. 4, pp. 90-92), Wyclif lists three main kinds of differences (or distinctions):

- real-and-essential;
- real-but-not-essential; and
- formal (or notional).

He does not define the real-and-essential difference, but identifies it through a rough account of its three sub-types. The things that differ really-and-essentially are those that differ from each other either (i) in genus, like man and quantity, or (ii) in species, like man and donkey, or (iii) in number, like two human beings.

The real-but-not-essential difference is more subtle than the first kind, since it holds between things that are the same single essence but really differ from each other nevertheless -- like memory, reason, and will, which are one and the same soul, and the three Persons of the Holy Trinity, who are the one and same God.

The third main kind of difference is the formal one. It is described as the difference by which things differ from each other even though they are constitutive elements of the same single essence or supposit. According to Wyclif, this is the case for:

1. the concrete accidents inhering in the same substance, since they coincide in the same particular subject but differ from each other because of their own natures;
2. the matter and substantial form of the same individual substance;
3. what is more common in relation to what is less common, like (*a*) the divine nature and the three Persons, (*b*) the world and this world; and, (*c*) among the categorial items belonging to the same category, a superior item and one of its inferiors.

This account of the various kinds of distinctions is more detailed than that of the *Purgans errores circa universalia in communi*, but not more clear. What is the difference, for instance, between the definition of the real-but-not-essential distinction and the definition of the formal distinction? What feature do all the kinds of formal distinction agree in? Some points are obvious, however:

1. The real-and-essential distinction matches the traditional real difference.
2. The real-but-not-essential distinction and the first sub-type of the formal distinction (that is, the distinction that holds between two or more concrete accidents belonging to the same individual substance) are two slightly different versions of the Scotistic formal distinction as defined in Scotus' *Lectura* (book I, d. 2, p. 2, qq. 1-4, ed. Vaticana, vol. xvi, p. 216) and *Ordinatio* (book I, d. 2, p. 2, qq. 1-4, ed. Vaticana, vol. ii, pp. 356-57; book II, d. 3, p. 1, q. 6, ed. Vaticana, vol. vii, pp. 483-84).
3. The third sub-type of the formal distinction is a reformulation of the Scotistic formal distinction as described in Scotus' *Reportata Parisiensia* (book I, d. 33, qq. 2-3, and d. 34, q. 1, ed. Vivès, vol. xxii, pp. 402-8, 410).

The main apparent dissimilarities between the analyses proposed in the *Tractatus de universalibus* and in the *Purgans errores circa universalia in communi* are the following:

1. There are three main kinds of differences instead of two.
2. Notwithstanding the presence of the qualification ‘real’, the real-but-not-essential difference in the *Tractatus de universalibus* is closer to the formal difference than is the corresponding kind of difference in the *Purgans errores circa universalia in communi,* since in the former the term ‘essence’ has the technical meaning of real entity with a given nature, and so is equivalent to ‘thing’.
3. The difference between the matter and the substantial form of the same individual substance is seen as a sub-type of real difference in the *Purgans errores circa universalia in communi* and as a sub-type of formal distinction in the *Tractatus de universalibus.*

2.3 The analysis of predication

Wyclif presents his opinion on universals as intermediate between those ones of St. Thomas (and Giles of Rome) and Walter Burley. Like Giles, whom he quotes by name, Wyclif recognizes three main kinds of universals:

1. *ante rem*, or ideal universals; that is, the ideas in God, archetypes of all that there is;
2. *in re*, or formal universals; that is, the common natures shared by individual things; and
3. *post rem*, or intentional universals; that is, mental signs by which we refer to the universals *in re.*

The ideas in God are the causes of the formal universals, and the formal universals are the causes of the intentional universals. On the other hand, like Burley, Wyclif holds that formal universals exist *in actu*

outside our minds, not *in potentia* as moderate Realists thought -- even though, unlike Burley, he maintains they are really identical with their own individuals. So Wyclif accepts the traditional realistic account of the relationship between universals and individuals, but translates it into the terms of his own system. According to him, universals and individuals are *really* the same, but *formally* distinct, since they share the same empirical reality (that of individuals) but, considered as universals and individuals, they have opposite constituent principles. On the logical side, this means that, notwithstandig real identity, not all that is predicated of individuals can be *directly* predicated of universals or *vice versa*, though an indirect predication is always possible. Hence Wyclif's description of the logical structure of the relationship between universals and individuals demanded the introduction of a new kind of predication, unknown to Aristotle, to cover cases, admitted by the theory, of indirect inherence of an accidental form in a substantial universal and of one second intention in another.

Therefore Wyclif distinguished three main types of predication, which he conceived as a real relation that holds between metaphysical entities. (On Wyclif's theory of predication, see Spade 1985, pp. xxxi-xli, and Conti 1997, pp. 150-58.)

In the *Purgans errores circa universalia in communi* (chap. 2), the three main types of predication are the following: formal predication, essential predication, and causal predication. In the *Tractatus de universalibus* (chap. 1, pp. 28-37), causal predication has been replaced by habitudinal predication -- a kind of predication that Wyclif had already recognized in the *Purgans errores circa universalia in communi,* but whose position within the main division of types of predication was not clear. In the *Tractatus de universalibus*, formal predication, essential predication, and habitudinal predication are described as three non-exclusive ways of predicating, each more general than the preceding. We speak of causal predication when the form designated by the predicate term is not present in the entity signified by the subject term, but is something caused by that entity. No instances of this kind of predication are given by Wyclif. Formal predication, essential predication, and habitudinal predication are defined in almost the same way in the *Purgans errores circa universalia* and in the *Tractatus de universalibus.*

Formal predication is that in which the form designated by the predicate term is directly present in the entity signified by the subject term. This happens whenever an item in the categorial line is predicated of something inferior, or an accident is predicated of its subject of inherence. In fact, in both cases, the subject term and the predicate term refer to the same reality in virtue of the form connoted by the predicate term itself.

To speak of essential predication, it is sufficient that the same empirical reality is both the real subject and the predicate, even though the formal principle connoted by the predicate term differs from that connoted by the subject term. ‘God is man’ and ‘The universal is particular’ are instances of this kind of predication. In fact, the same empirical reality (or essence) that is a universal is also an individual, but the forms connoted by the subject term and by the predicate term differ from each other.

Finally we speak of habitudinal predication when the form connoted by the predicate term does not inhere, either directly or indirectly, in the essence designated by the subject, but simply implies a relation to it, so that the same predicate may be at different times truly or falsely spoken of its subject without

there being any change in the subject itself. According to Wyclif, we use such a kind of predication mainly when we want to express theological truths, like: God is known and loved by many creatures, and brings about, as efficient, exemplar, and final cause, many good effects. It is evident that habitudinal predication does not require any kind of identity between the entity signified by the subject term and the entity signified by the predicate term, but formal predication and essential predication do. So the ontological presuppositions of the most general type of predication, implied by the other types, are completely different from those of the other two.

The final result of Wyclif's revolution is therefore a not fully developed system of intensional logic, which he superimposes on the standard extensional system inherited from Aristotle. As a result, the copula of the philosophical propositions that are dealt with cannot be extensionally interpreted, since it does not properly mean that a given object is a member of a certain set or that a given set is included in another; rather it means degrees of identity. Only in virtue of renouncing any extensional approach to the matter were Wyclif's followers able to give a logically satisfactory solution of the problem of the relationship between universals and individuals, which had always been the most difficult issue for medieval Realists.

Metaphysics

3.1 Being and analogy

The point of departure for Wyclif's metaphysics is the notion of being, since it occupies the central place in his ontology. After Duns Scotus, the real issue for metaphysics was the relationship between being and, on the other side, God and creatures, as Scotus' theory of the univocity of the concept of being was an absolute novelty, full of important consequences for the development of later medieval philosophy. Wyclif takes many aspects from Scotus' explanation, but strongly stresses the ontological implications of the doctrine. Wyclif, like Scotus, claims that the notion of being is the most general one, a notion entailed by all others, but he also states that an extra-mental reality corresponds to the concept of being-in-general (*ens in communi*). This extra-mental reality is predicated of everything (God and creatures, substances and accidents, universal and individual essences) according to different degrees, since God *is* in the proper sense of the term and any other entity is (something real) only insofar as it shares the being of God (*De ente in communi*, chap. 1, pp. 1-2; chap. 2, p. 29; *De ente praedicamentali*, chap. 1, p. 13; chap. 4, p. 30; *Tractatus de universalibus*, chap. 4, p. 89; chap. 7, p. 130; chap. 12, p. 279; *De materia et forma*, chap. 6, p. 213).

If being is a reality, it is then clear that it is impossible to affirm its univocity. The *Doctor Subtilis* thought of being as simply a concept, and therefore could describe it as univocal in a broad sense (one name -- one concept -- many natures). Wyclif, on the contrary, is convinced that the being-in-general is an extra-mental reality, so he works out his theory at a different level than does Scotus: no more at the intensional level (the meaning connected with the univocal sign, or *univocum univocans*), but at the extensional one (the thing signified by the mental sign, considered as shared by different entities

according to different degrees). For that reason, he cannot use Aristotelian univocation, which hides these differences in sharing. Thus he denies the univocity of being and prefers to use one of the traditional notions of analogy (*De ente praedicamentali*, chap. 3, pp. 25, 27), since the being of God is the measure of the being of other things, which are drawn up on a scale with the separated spiritual substances at the top and prime matter at the bottom. Therefore he qualifies being as an ambiguous genus (*ibidem*, p. 29), borrowing an expression already used by Grosseteste in his commentary on Aristotle's *Posterior Analytics*. The analogy of being does not entail a multiplicity of correlated meanings, however, as in Thomas Aquinas. Since Wyclif hypostatizes the notion of being and considers equivocity, analogy, and univocity as real relations between things, not as semantic relations between terms and things, his analogy is partially equivalent to the standard Aristotelian univocity, since what differentiates analogy from univocity is the way a certain nature (or property) is shared by a set of things: analogous things (*analoga*) share it according to different degrees (*secundum magis et minus*, or *secundum prius et posterius*), while univocal things (*univoca*) share it all in the same manner and at the same degree. This is the true sense of his distinction between ambiguous genera, like being and accident (*accidens*), and logical genera, like substance (*De ente praedicamentali*, chap. 4, pp. 30, 32). Hence, according to this account, being in general is the basic component of the metaphysical structure of each reality, which posses it in accordance with its own nature, value, and position in the hierarchy of created beings.

Unfortunately, this theory is weak in an important point, since Wyclif does not clarify the relation between being-in-general and God. On the one hand, being is a creature, the first of all the creatures; on the other hand, God should share it, as being-in-general is the most common reality, predicated of all, and according to him to-be-predicated-of something means to-be-shared-by it. As a consequence, a creature would be in some respect superordinated to God -- a theological puzzle that Wyclif failed to acknowledge.

3.2 Being and truth

According to Wyclif, the constitutive property of each kind of being is the capacity to be the object of a complex act of signifying (*De ente in communi*, chap. 3, p. 36; *De ente primo in communi*, chap. 1, p. 70). This choice implies a revolution in the standard medieval theory of transcendentals, since Wyclif actually replaces being (*ens*) with true (*verum*). According to the common belief, among the transcendentals (being, thing, one, something, true, good) being was the primitive notion, from which all the others stemmed by adding a specific connotation in relation to something else, or by adding some new determination. So true (*verum*) was nothing but being (*ens*) itself considered in relation to an intellect, no matter whether divine or human. In Wyclif's view, on the contrary, being is no longer the main transcendental and its notion is not the first and simplest; rather there is something more basic to which being can be reduced: truth (*veritas* or *verum*). According to the English philosopher, only what can be signified by a complex expression is a being, and whatever is the proper object of an act of signifying is a truth. Truth is therefore the true name of being itself (*Tractatus de universalibus*, chap. 7, p. 139). Thus everything that is is a truth, and every truth is something not simple but complex. Absolute simplicity is unknown within Wyclif's metaphysical world. From the semantic point of view, this means the collapsing of the fundamental distinction in the common Aristotelian theory of meaning, the one between simple signs (like nouns) and compound signs (like propositions). From the ontological point of

view, this entails the uniqueness in type of what is signified by every class of categorematic expressions (*Logica*, chap. 5, p. 14). Within Wyclif's world, it is the same (kind of) object that both concrete terms and propositions refer to, as individual substances have to be regarded as (atomic) states of affairs. According to him, from the metaphysical point of view a singular man is nothing but a real proposition (*propositio realis*), where actual existence in time as an individual plays the role of the subject, the common nature (i.e., human nature) plays the role of the predicate, and the singular essence (i.e., that by means of which this individual is a man) plays the role of the copula (*ibid.*, pp. 14-15).

Despite appearances, Wyclif's opinion on this subject is not just a new formulation of the theory of the *complexe significabile*. According to the supporters of the *complexe significabile* theory, the same things that are signified by simple concrete terms are signified by complex expressions (or propositions). In Wyclif's thought, on the contrary, there are no simple things in the world that correspond to simple concrete terms; rather, simple concrete terms designate real propositions, that is, (atomic) states of affairs. Wyclif's real proposition is everything that is, as everything save God is composed at least of potency and act (*De ente praedicamentali*, chap. 5, pp. 38-39), and therefore can be conceived of and signified both in a complex (*complexe*) and in a non-complex manner (*incomplexe*) (*Tractatus de universalibus*, chap. 2, pp. 55-56; chap. 3, pp. 70, 74, and 84; chap. 6, pp. 118-19). When we conceive of a thing in a complex manner, we think of that thing considered according to its metaphysical structure, and so according to its many levels of being and kinds of essence. As a consequence, Wyclif's metaphysical world, like his physical world, consists of atomic objects, that is, single essences belonging to the ten different types or categories. But these metaphysical atoms are not simple but rather composite, because they are reducible to something else, belonging to a different rank of reality and unable to exist by itself: being and essence, potency and act, matter and form, abstract genera, species and differences. For that reason, everything one can speak about or think of is both a thing and an atomic state of affairs, while every true sentence expresses a molecular state of affairs, that is, the union (if the sentence is affirmative) or the separation (if the sentence is negative) of two (or more) atomic objects.

3.3 Being and essence

Among the many kinds of beings Wyclif lists, the most important set is that consisting of categorial beings. They are characterized by the double fact of having a nature and of being the constitutive elements of finite corporeal beings or atomic states of affairs. These categorial items, conceived of as instances of a certain kind of being, are called by Wyclif ‘essences’ (‘*essentiae*’). An essence therefore is a being that has a well defined nature, even if the name ’essence’ does not make this nature known (*De ente primo in communi*, chap. 3, pp. 88-89; *De ente praedicamentali*, chap. 5, p. 43; *Tractatus de universalibus*, chap. 7, pp. 128-29; *De materia et forma*, chap. 4, pp. 185-86). So the term ‘essence’ (‘*essentia*’) is less general than ‘being’ (‘*ens*’), but more general than ‘quiddity’ (‘*quidditas*’), since (i) every essence is a being, and not every being is an essence, and (ii) every quiddity is an essence, and not every essence is a quiddity, as individual things are essences but are not quiddities (see Kenny 1985, pp. 21 ff.; and Conti 1993, pp. 171-81).

According to Wyclif, being is the stuff that the ten categories modulate according to their own nature, so

that everything is immediately something that is (*De ente praedicamentali*, chap. 4, p. 30; *Tractatus de universalibus*, chap. 7, p. 130); therefore, he maintains no real distinction between essence and being. The essences of creatures do not precede their beings, not even causally, since every thing is (identical with) its essence. The being of a thing is brought into existence by God at the same instant as its essence, since essence without being and being without essence would be two self-contradictory states of affairs. In fact, essence without being would imply that an individual could be something of a given type without being real in any way, and being without essence would imply that there could be the existence of a thing without the thing itself (*Tractatus de universalibus*, chap. 6, pp. 122-23). As a consequence, the *pars destruens* of his theory of being and essence is a strong refutation of the twin opinions of St. Thomas and Giles of Rome. Although Wyclif does not name either the Dominican master or the Augustinian one, it is nevertheless clear from the context that their conceptions are the object of his criticisms (*ibid.*, pp. 120-22).

On the other hand, it is evident that while from the extensional point of view the being and essence of creatures are equipollent, since every being is an essence and *vice versa*, from the intensional point of view there is a difference, because the being of a thing *logically* presupposes its essence and not *vice versa* (*De materia et forma*, chap. 4, pp. 184-85). Moreover, in Wyclif's opinion, every creature has two different kinds of essence and four levels of being. Indeed, he clearly distinguishes between singular essence and universal essence (*essentia quidditativa speciei vel generis*) -- that is, the traditional *forma partis* and *forma totius*. The singular essence is the form that in union with the matter brings about the substantial composite. The universal essence is the type that the former instantiates; it is present in the singular substance as a constitutive part of its nature, and it discloses the inner metaphysical structure of the substantial composite (*Tractatus de universalibus*, chap. 6, pp. 116-18). Furthermore, he speaks of four-fold level of reality (*esse*):

1. First, the eternal mental being (*esse ideale*) that every creature has in God, as an object of His mind.
2. Second, the potential being everything has in its causes, both universal (genus, species) and particular. This is closely connected with the nature of the individual substance on which the finite corporeal being is founded, and is independent of its actual existence. It is called ‘*esse essentiae*’ or ‘*esse in genere*’.
3. Third, the actual existence in time as an earthly object.
4. Fourth, the accidental being (*modus essendi accidentalis substantiae*) caused in a substance by the inhering in it of its appropriate accidental forms (*Tractatus de universalibus*, chap. 7, pp. 126-28).

Thus the identity between essence and being cannot be complete. Consequently Wyclif speaks of a formal difference (*distinctio* or *differentia formalis*) -- which he also calls a ‘difference of reason’ (*distinctio rationis*) -- between essence and being. More precisely, he holds that:

1. The *esse ideale* is formally distinct from the singular essence;
2. The actual existence is formally distinct from the universal essence; and
3. The singular essence is formally distinct from the actual existence.

In this way, Wyclif establishes a close connection between singular essence and essential being, on the one hand, and a real identity between universal and individual (that is, between universal essence and singular essence), on the other hand. Essential being is the level of being that matches singular essence, while actual existence is in a certain way accidental to the singular essence, as the latter is nothing else but the universal essence considered as informing matter.

4. Theology

4.1 Divine ideas

According to St. Thomas (see *In I Sent*., d. 19, q. 5; d. 36, qq. 1-2; and *STh*. I, qq. 14-15) -- whose doctrine of divine ideas can be considered a perfect background for a better understanding of Wyclif's theory -- divine ideas are really the same as the divine essence, but distinct in reason from it. Everything produced by God has a certain similarity with the divine essence, since divine ideas are the ways in which God views His essence as capable of being imitated by a possible creature. When a given possible creature is brought into existence by the divine volition, the divine idea that is its corresponding paradigm also serves as a principle of divine creation and becomes therefore an *exemplar* in the strict sense of the term. As a consequence, according to Aquinas, there is a difference between a divine idea as a mere principle of knowledge (*ratio*) by means of which God thinks of a given possible and a divine idea as an exemplar by means of which God produces a certain set of individuals. This difference prevents Aquinas' system from being a form of necessitarism, as the two spheres of existent and possible do not coincide, since the existent is a sub-set of the possible.

Wyclif defines ideas as the divine nature in action, since they are the means by which God creates all that is outside Himself. In this way, any distinction between the ideas as *rationes* and the ideas as *exemplaria* is abolished. Furthermore, ideas are the constitutive principles of the divine nature, essentially identical with it. Thus divine ideas become as necessary as the divine nature itself (*Tractatus de universalibus*, chap. 15, pp. 371-74). On the other side, ideas are the first level of being proper to creatures (see above Section 3.3). As a consequence, everything that is is necessary and a necessary object of God's volition; the three spheres of possible, existent, and necessary totally coincide.

4.2 Divine omnipotence

This doctrine of divine ideas and the connected theory of being had a significant result also for the notion of divine omnipotence. In the Middle Ages, one of the most important features of divine omnipotence was the capacity of annihilating, which was viewed as the necessary counterpart of the divine capacity of creating. Wyclif denies the thesis of an opposition between creation and annihilation, and openly denies that God can annihilate creatures. He argues that nothing is contrary to creation, since the act of creating is peculiar to God, and nothing is opposite or contrary to God. In fact, *absolute* non-being (the only "thing" that could be considered opposite to God) is something self-contradictory, and therefore logically impossible. Accordingly, there cannot be any action opposite to creation. The only possible kind of non-

being admitted by Wyclif is corruption (*corruptio*), that is, the natural destruction of the actual existence in time of an object in the world (*Tractatus de universalibus*, chap. 13, pp. 302-3).

On the other hand, according to Wyclif, annihilation, if possible, would be equivalent to the total destruction of all of a creature's levels of being (*ibid.*, p. 307), and thus would imply the following absurdities:

1. God could not annihilate any creature without destroying the whole world at once, since universal-being is the basic constitutive element of the second level of being (the *esse essentiae* or *esse in genere*) of each creature (*ibid.*, pp. 307-8).
2. Since annihilation would be nothing but an accident, and more precisely an action, it would be really different from both the acting subject (i.e. God) and the object of the action (i.e., the thing that would be annihilated). But any accident requires a substrate of inherence. In this case, it cannot be God. Thus, it must be the object of annihilation. Yet, because of its particular nature, if there is annihilation, its substrate of inherence cannot be, and therefore the annihilation itself cannot be, since no accident can exist without any substrate of inherence -- an apparently self-contradictory state of affairs (*ibid.*, pp. 310-11).
3. God could not annihilate any creature without annihilating Himself at the same time, because the first and most basic level of being of every creature is rooted in the divine essence itself (*ibid.*, pp. 313-14).

The image of God Wyclif draws here is not the Christian image of the Lord of the universe, who freely creates by an act of His will and has absolute power and control over everything, but a variation of the Neoplatonic notion of the One. Wyclif's God is simply the supreme principle of the universe from which everything necessarily flows. Within Wyclif's system, creation is a form of emanation, as each creature is necessarily connected with the divine essence itself by means of its *esse ideale*. God has been deprived of the power of revocation (*ibid.*, pp. 304-5), and the only action He can, or rather has to, perform is creation. Because of the necessary links between (i) the divine essence and the eternal mental being that every creature has in God and (ii) this first level of being of creatures and the remaining three, in God to think of creatures is already to create them. But God cannot help thinking of creatures, since to think of Himself is to think of His constitutive principles, that is, of the ideas of creatures. Therefore, God cannot help creating. Indeed, He could not help creating just this universe.

Wyclif's rejection of the possibility of annihilation and the subsequent new notion of divine onnipotence shed light on his theory of universals, as they help us to appreciate the difference between his thesis of the identity between universals and individuals and the analogous thesis of moderate Realists. For these latter, this identiy meant that *the* individuals are *in potentia* universal; for Wyclif it means that *the* individuals are *the* universals *qua* existing *in actu* -- that is, the individuals are the outcome of a process of production that is inscribed into the nature of general essences themselves, and through which general essences change from an incomplete type of subsistence as forms to a full existence as individuals. This position is consistent with (i) his theory of substance, where the main and basic composition of every substance, both individual and universal, is not the hylemorphic one, but the composition of potency and act (*De ente praedicamentali*, chap. 5, pp. 38-39), and (ii) a Neoplatonic reading of Aristotelian

metaphysics, where universal substances, and not individual ones as the Stagirite had taught, are the main and fundamental kind of being.

4.3 The Eucharist

Wyclif's heretical theses concerning the Eucharist are the logical consequence of the application of this philosophical apparatus to the problem of the real presence of the body of Christ in the consecrated host. According to Catholic doctrine, after consecration the body of Christ is really present in the host instead of the substance of the host itself, while the accidents of the host are the same as before. St. Thomas's explanation of this process, called ‘transubstantiation’, was that the substance of the bread (and wine) was changed into the body (and blood) of Christ, whereas its quantity, through which the substance of the bread received physical extension and the other accidental forms, was now the entity that kept the other accidental forms physically in being. Duns Scotus and Ockham, on the contrary, had claimed that after consecration the substance of the bread (and wine) was annihilated by God, while the accidents of the bread (and wine) remained the same as before because of an intervention of divine omnipotence.

Wyclif rejects both solutions as well as the Catholic formulation of the dogma, since he could not accept the ideas of the destruction of a substance by God and of the existence of the accidents of a given singular substance without and apart from that singular substance itself -- two evident absurdities within the metaphyisical framework of his system of thought. As a consequence, Wyclif affirms the simultaneous presence in the Eucharist of the body of Crhist and of the substance of the bread (and wine), which continues to exist even after the consecration. According to him, transubstantiation is therefore a twofold process, natural and supernatural. There is natural transubstantiation when a substitution of one substantial form for another takes place, but the subject-matter remains the same. This is the case with water that becomes wine. There is supernatural transubstantiation when a miraculous transformation of the substantial entity at issue takes place. This was the case, for instance, with the incarnation of the second person of the Trinity, who is God and became man (*De apostasia*, p. 170). The Eucharist implies this second kind of transubstantiation, since the Eucharist, like Christ, has a dual nature: earthly and divine. According to its earthly nature the Eucharist is bread (and wine), but according to its divine nature it is the body of Christ, which is present in the host spiritually or habitudinnaly, since it is in virtue and by means of faith only that it could be received (*De apostasia,* pp. 180 and 210; *De eucharistia,* pp. 17, 19, 51-52, and 230; for a description of the habitudinal presence, see the definition of the habitudinal predication above, Section 2.3).

5. Religious and Political Thought

5.1 The Bible and the Church

Wyclif conceives Sacred Scripture as a direct emanation from God himself, and therefore as a timeless, unchanging, and archetypal truth independent of the present world and of the concrete material text by means of which it is manifested. As a consequence, in his *De veritate Sacrae Scripturae* (*On the Truth of*

Sacred Scripture -- between late 1377 and the end of 1378) he tries to show that, despite appearences, the Bible is free from error and contradictions. The exegetic principle he adopts is the following: since the authority of Scripture is greater than our capacity of understanding, therefore if some error and/or inconsistencies are found in the Bible, there is something wrong with our interpretation. The Bible contains the whole truth and nothing but the truth, so that nothing can be added to it or subtracted from it. Every part of it has to be taken absolutely and without qualification (*De veritate Sacrae Scripturae,* vol. 1, pp. 1-2, 395, 399; vol. 2, pp. 99, 181-84).

In attributing inerrancy to the Bible, Wyclif was following the traditional attitude towards it, but the way he viewed the book detached him from Catholic tradition, as he thought that his own metaphysical system was the necessary interpretative key for the correct understanding of Biblical truth. In fact, in the *Trialogus* (*Trialogue* -- between late 1382 and early 1383), where Wyclif gives us the conditions for achieving the true meaning of the Bible, they are the following:

1. knowledge of the nature and ontological status of universals;
2. knowledge of the peculiar nature of accidents as dependent in existence on their substantial substrates;
3. knowledge of past and future states of affairs (*praeteritiones* and *futuritiones*) as real in the present as past and future truths, not as things (*res*) that have been real in the past and will be real in the future (a thesis of his already claimed in the *De ente praedicamentali,* chap. 1, pp. 2 and 5; *Purgans errores circa veritates in communi,* chap. 1, pp. 1-2; chap. 3, pp. 10-11);
4. knowledge of the eternal existence of creatures in God at the level of intelligible being really identical with the divine essence itself;
5. knowledge of the perpetual existence of material essences (*Trialogus,* book 3, chap. 31, pp. 242-43).

Only on the basis of this logical and metaphyisical machinery is it possible to grasp the five different levels of reality of the Bible, which is at the same time:

1. the book of life mentioned in the *Apocalypse*;
2. the ideal being proper to the truths written in the book of life;
3. the truths that are to be believed as they are written in the book of life;
4. the truths that are to believed as they are written in the natural books that are men's souls;
5. all the artificial signs of the truth (*De veritate Sacrae Scripturae,* vol. 1, p. 109).

This same approach, when applied to the Church, led Wyclif to fight against it in its contemporary state. (On Wyclif's ecclesiology see Leff 1967, pp. 516-46.) The starting point of Wyclif's reflection on the Church is the distinction between the heavenly and the earthly cities that St. Augustine draws in his *De civitate Dei.* In St. Augustine such a division is metaphorical, but Wyclif made it literal. So he claims that the Holy Catholic Church is the mystical and indivisible community of the saved, eternally bound together by the grace of predestination, while the foreknown, i.e. the damned, are eternally excluded from it (*De civili dominio,* vol. 1, p. 11). This community of the elect is really distinct from the various particular earthly churches (*ibid.*, p. 381). It is timeless and outside space, and therefore is not a physical

entity; its being, like the actual being of any other universal, is wherever any of its members is (*De ecclesia*, p. 99). All its members always remain in grace, even if temporally in mortal sin (*ibid.*, p. 409), as conversely the damned remain in mortal sin, even if temporally in grace (*ibid.*, p. 139). The true Church is presently divided into three parts: the triumphant Church in heaven; the sleeping Church in purgatory; and the militant Church on earth (*ibid.*, p. 8). But the militant Church on earth cannot be identified with the visible church and its hierarchy. Even more, since we cannot know who are the elect, there is no reason for consenting to recognize and obey the authority of the visible church (see *De civili dominio*, vol. 1, p. 409; *De ecclesia*, pp. 71-2). Authority and dominion rely on God's law manifested by Sacred Scripture. As a consequence, obedience to any member of the hierarchy is to be subordinated to his fidelity to the precepts of the Bible (*De civili dominio*, vol. 2, p. 243; *De potestate papae* [*On the Power of the Pope* -- ca. 1379], p. 149; De *ecclesia*, p. 465). Faithfulness to the true Church can entail the necessity of rebelling against the visible church and its members, when their requests are in conflict with the teaching of Christ (*De civili dominio*, vol. 1, pp. 384, 392).

In conclusion, since the visible church cannot help the believers gain salvation, which is fixed from eternity, and its authority depends on its fidelity to divine revelation, it cannot perform any of the functions traditionally attributed to it, and it therefore has no reason for its own existence. To be ordained a priest offers no certainty of divine approval and authority (*De ecclesia*, p. 577). Orthodoxy can only result from the application of right reason to the faith of the Bible (*De veritate Sacrae Scripturae*, vol. 1, p. 249). The Pope, bishops, abbots, and priests are expected to prove their really belonging to the Holy Catholic Church through their exemplary behavior; they should be poor and free from worldly concerns, and they should spend their time preaching and praying (*De ecclesia*, pp. 41, 89, 129). In particular, the Pope should not interfere in worldly matters, but should be an example of holiness. Believers are always allowed to doubt the clergy's legitimacy, which can be evaluated only on the basis of its consistency with the Evangelic rules (*ibid.*, pp. 43, 456). Unworthy priests forfeit their right to exercise authority and to hold property, and lay lords might deprive them of their benefices (*De civili dominio*, vol. 1, p. 353; vol. 3, pp. 326, 413; *De ecclesia*, p. 257).

5.2 Dominion

As Leff remarked (Leff 1967, p. 546), the importance of Wyclif's teaching on dominion and grace has been exaggerated. His doctrine depends on Richard Fitzralph's theory, according to which the original lordship is independent of natural and civil circumstances (on Fitzralph's conception see Robson 1961, pp. 70-96), and is only a particular application of Wyclif's general view on election and damnation. In fact, the three main theses of the first book of his *De civili dominio* are the following:

1. a man in sin has no right to dominion;
2. a man who is in a state of grace possesses all the goods of the world;
3. as a consequence, there can be no dominion without grace as its formal cause (*De civili dominio*, vol. 1, p 1).

Wyclif defines dominion as the right to exercise authority and, indirectly, to hold property. According to

him, there are three kinds of possession: natural, civil, and evangelical. Natural possession is the simple possession of goods without any legal title. Civil possession is the possession of goods on the basis of some civil law. Evangelical possession requires, beyond civil possession, a state of grace in the legal owner. Thus God alone can confer evangelical possession (*ibid.,* p. 45). On the other hand, a man in a state of grace is lord of the visible universe, but on the condition that he shares his lordship with all the other men who are in a state of grace, as all men in a state of grace have the same rights. This ultimately means that all the goods of God should be in common, just as they were before the Fall. Private property was introduced as a result of sin. From this point of view it is also evident that Aristotle's criticisms against Plato are unsound, since Platonic communism is correct in essence (*ibid.,* pp. 96 ff.). The purpose of civil law is to preserve the necessities of life (*ibid.,* pp. 128-29). The best form of government is monarchy. Kings must be obeyed and have taxes paid to them, even if they become tyrants, since they are God's vicars that He alone can depose -- so that only secular lordship is justified in the world (*ibid.,* p. 201).

Bibliography

Primary Literature

- *De civili dominio,* R. L. Poole, J. Loserth & F. D. Matthew, ed., 4 vols., London: Trübner for the Wyclif Society, 1895-1904.
- *De ecclesia,* J. Loserth & F. D. Matthew, ed., London: Trübner for the Wyclif Society, 1886.
- *De ente in communi* and *De ente primo in communi,* in *Johannis Wyclif Summa de ente, libri primi tractatus primus et secundus,* S. H. Thomson, ed., Oxford: Clarendon Press 1930.
- *De ente praedicamentali,* R. Beer, ed., London: Trübner for the Wyclif Society, 1891.
- *De eucharistia,* J. Loserth, ed., London: Trübner for the Wyclif Society, 1892.
- *De materia et forma*, in *Johannis Wyclif miscellanea philosophica*, vol. 1, M. H. Dziewicki, ed., London: Trübner for the Wyclif Society, London 1902, at pp. 163-242.
- *De officio regis,* R. A. Pollard & C. E. Sayle, ed., London: Trübner for the Wyclif Society, 1887.
- *De potestate papae,* J. Loserth, ed., London: Trübner for the Wyclif Society, 1907.
- *Purgans errores circa universalia in communi,* chap. 2, in S. H. Thomson, "A ‘Lost’ Chapter of Wyclif's *Summa de ente*," *Speculum* 4 (1929), pp. 339-46. The ms. Cambridge, Trinity College, B.16.2, used by Dziewicki for his edition of the work, lacks the second chapter and the first section of the third chapter. S. H. Thomson integrated the text on the basis of the ms. Wien, Österreichische Nationalbibliothek, 4307.
- *Purgans errores circa veritates in communi, Purgans errores circa universalia in communi, De intelleccione Dei,* and *De volucione Dei*, in *De ente librorum duorum excerpta,* M. H. Dziewicki, ed., London: C. K. Paul & Co. for the Wyclif Society, 1909.
- *Summa insolubilium,* P. V. Spade & G. A. Wilson, ed., Binghamton, N.Y.: Medieval and Renaissance Texts and Studies, 1986.
- *Tractatus de logica,* M. H. Dziewicki, ed., 3 vols., London: Trübner for the Wyclif Society, 1893-99.
- *Tractatus de universalibus,* I. J. Mueller, ed., Oxford: Clarendon Press, 1985.

- *De Trinitate*, A. du Pont Breck, ed., Boulder, Col.: University of Colorado Press, 1962.
- *On Universals*, A. Kenny, trans., with an Introduction by P. V. Spade, Oxford: Clarendon Press, 1985. (English translation of *Tractatus de universalibus*.)
- *De veritate Sacrae Scripturae*, R. Buddensieg, ed., 3 vols., London: Trübner for the Wyclif Society, 1905-07.
- *Trialogus*, G. Lechler, ed., Oxford: Clarendon Press, 1869.

Secondary Literature

- E. J. Ashworth & P. V. Spade, "Logic in Late Medieval Oxford," in J. I. Catto & R. Evans, ed., *The History of the University of Oxford*, Oxford: Clarendon Press, 1992, vol. 2, pp. 35-64.
- J. I. Catto, "Wyclif and Wycliffism in Oxford, 1356-1430," in Catto & Evans, *The History of the University of Oxford* (see above), pp. 175-261.
- A. D. Conti, "Essenza ed essere nel pensiero della tarda scolastica (Burley, Wyclif, Paolo Veneto)," *Medioevo* 15 (1989), pp. 235-67.
- A. D. Conti, "Logica intensionale e metafisica dell'essenza in John Wyclif," *Bullettino dell'Istituto Storico Italiano per il Medioevo e Archivio muratoriano* 99.1 (1993), pp. 159-219.
- A. D. Conti, "Analogy and Formal Distinction: On the Logical Basis of Wyclif's Metaphysics," *Medieval Philosophy and Theology* 6.2 (1997), pp. 133-65.
- L. J. Daly, *The Political Theory of John Wyclif*, Chicago: Loyola University Press, 1962.
- A. de Libera, *La querelle des universaux: De Platon à la fin du Moyen Age*, Paris: Éditions du Seuil, 1996, at pp. 403-28.
- W. Farr, *John Wyclif as Legal Reformer*, Leiden: E. J. Brill, 1974.
- N. W. Gilbert, "Ockham, Wyclif and the *via moderna*," in A. Zimmermann, ed., *Antiqui und Moderni: Traditionsbewuβtsein und Fortschrittsbewuβtsein im späten Mittelalter*, Berlin: Miscellanea Mediaevalia, 1974, pp. 85-125.
- A. Hudson & M. Wilks, ed., *From Ockham to Wyclif*, Oxford: Blackwell, 1987, at pp. 165-77, 185-215, and 217-32.
- A. Hudson, *The Premature Reformation: Wycliffite Texts and Lollard History*, Oxford: Clarendon Press, 1988.
- A. Kenny, *Wyclif*, Oxford: Clarendon Press, 1985.
- A. Kenny, ed., *Wyclif in his Times*, Oxford: Clarendon Press, 1986.
- G. Leff, *Heresy in the Later Middel Ages*, 2 vols., Manchester: Manchester University Press, 1967, vol. 2, at pp. 494-573.
- K. B. McFarlane, *John Wycliffe and the Beginnings of English Non-Conformity*, London: English Universities Press, 1952.
- J. A. Robson, *Wyclif and the Oxford Schools*, Cambridge: Cambridge University Press, 1961.
- S. Simonetta, "John Wyclif between Utopia and Plan," in *Société et Église: Textes et discussions dans les universités de l'Europe centrale pendant le moyen âge tardif*, Turnhout: Brepols, 1995, pp. 65-86.
- S. Simonetta, "La maturazione del progetto riformatore di Giovanni Wyclif: dal *De civili dominio* al *De officio regis*," *Medioevo* 22 (1996), pp. 225-58.
- P. V. Spade, "Introduction," in J. Wyclif, *On Universals*, pp. vii-l (see "Primary literature" above).

- P. V. Spade & G. A. Wilson, "Introduction," in J. Wyclif, *Summa insolubilium* (see "Primary literature" above).
- S. H. Thomson, "The Philosophical Basis of Wyclif's Theology," *Journal of Religion*, 11 (1931), pp. 86-116.
- W. R. Thomson, *The Latin Writings of John Wyclif: An Annotated Catalog*, Toronto: Pontfical Institute of Mediaeval Studies, 1983.
- K. Walsh & D. Wood, ed., *The Bible in the Medieval World*, Oxford: Blackwell, 1985, at pp. 269-315.
- M. Wilks, "Predestination, Property and Power: Wyclif's Theory of Dominion and Grace," *Studies in Church History*, 2 (1965), pp. 220-36.
- M. Wilks, "*Reformatio regni*: Wyclif and Hus: Leaders of Religious Protest Movements," *Studies in Church History* 9 (1972), pp. 109-30.
- H. B. Workman, *John Wyclif: A Study of the English Medieval Church*, 2 vols., Oxford: Clarendon Press, 1926.

Other Internet Resources

- Religious Movements in the Fourteenth Century, by Rev. J. P. Whitney, B.D., (King's College), in the *Cambridge History of English and American Literature*. See especially sections 5-8 on Wyclif.

[Please contact the author with other suggestions.]

Related Entries

Alyngton, Robert | analogy: medieval theories of | Aquinas, Saint Thomas | Burley [Burleigh], Walter | Duns Scotus, John | Giles of Rome | Paul of Venice | Penbygull, William | Sharpe, Johannes | universals: the medieval problem of

Alessandro D. Conti
a.conti@tiscalinet.it

First published: September 18, 2001
Content last modified: December 12, 2001

Medieval Theories of Analogy

Medieval theories of analogy were a response to problems in three areas: logic, theology, and metaphysics. Logicians were concerned with the use of words having more than one sense, whether completely different, or related in some way. Theologians were concerned with language about God. How can we speak about a transcendent, totally simple spiritual being without altering the sense of the words we use? Metaphysicians were concerned with talk about reality. How can we say that both substances (e.g., Socrates) and accidents (e.g., the beardedness of Socrates) exist when one is dependent on the other; how can we say that both God and creatures exist, when one is created by the other? Medieval thinkers reacted to these three problems by developing a theory which divided words into three sorts, independently of context. Some were univocal (always used with the same sense), some were purely equivocal (used with quite different senses), and some were analogical (used with related senses). Analogical terms were thought to be particularly useful in metaphysics and theology, but they were routinely discussed in commentaries on Aristotle's logic and in logic textbooks. The background to the discussion was given by what is often called the analogy of being, the doctrine that reality is divided horizontally into the very different realities of substances and accidents, and vertically into the very different realities of God and creatures. Nonetheless, the phrase "medieval theories of analogy" refers not to ontology but to a set of linguistic and logical doctrines supplemented, at least from the fourteenth century on, by doctrines about the nature of human concepts.

There were three main types of analogy. In the original Greek sense, analogy involved a comparison of two proportions or relations. Thus ‘principle’ was said to be an analogical term when said of a point and a spring of water because a point is related to a line as a spring is related to a river. This type of analogy came to be called the analogy of proportionality. In the second sense, analogy involved a relation between two things, of which one is primary and the other secondary. Thus ‘healthy’ was said to be an analogical term when said of a dog and its food because while the dog has health in the primary sense, its food is healthy only secondarily as contributing to or causing the health of the dog. This second type of analogy became known as the analogy of attribution, and its special mark was being said in a prior and a posterior sense (*per prius et posterius*). A third type of analogy, sometimes appealed to by theologians, appealed to a relation of likeness between God and creatures. Creatures are called good or just because their goodness or justice imitates or reflects the goodness or justice of God. This type of analogy was called the analogy of imitation or participation. Of the three types, it is the analogy of attribution that is central to medieval discussions.

From the fourteenth century on discussions of analogy focused not so much on linguistic usages as on the nature of the concepts that corresponded to the words used. Is there just one concept that corresponds to

an analogical term, or is there a sequence of concepts? If the latter, how are the members of the sequence ordered and related to each other? Moreover, how far should we distinguish between so-called formal concepts (or acts of mind) and objective concepts (whatever it is that is the object of the act of understanding)? These discussions were still influential at the time of Descartes.

---

1. Medieval Theories of Language

Medieval logicians and philosophers of language were principally concerned with the relationship between utterances, concepts, and things. Written language was only of secondary importance. They agreed that spoken language was conventional, having its origin in imposition, or the decision to correlate certain sounds with certain objects. Concepts, however, were natural, in the sense that all human beings with similar experiences had the same concepts, without any decision being involved. The key semantic notion was signification, rather than meaning, though translated sources tend to obscure this by translating ‘significatio’ as ‘meaning’. For a term to signify is for it to function as a sign, to represent or make known something beyond itself. A typical spoken term, such as ‘horse’ or ‘dog’, signifies in two ways. It signifies or makes known the concept with which it has to be correlated in order to function significatively at all, and it also signifies or makes known something external to and independent of the mind. Modifications were made to this simple scheme to cover the cases of special terms, including syncategorematic terms, such as ‘every’ and ‘not’, fictional terms such as ‘chimera’, and privative terms such as ‘blindness’; and modifications were also made to cover the case of special predicates, such as "is a genus", or "is thought about", but we need not concern ourselves with these modifications here.

Theories of signification were complicated by the metaphysical problem of common natures. If we say that words signify things external to and independent of the mind do we mean that ‘a human being’ and ‘beard’ signify special common objects such as humanity or beardedness, or do we mean that they

signify Socrates and his beard? For some thinkers, the primary significate of a common noun was the common nature, and the secondary significate was the thing having that nature. For Aquinas, who did not want to give common natures any kind of intermediary existence independent of both concepts and actual things, the significate (*significatum*) of a term was the intellect's conception (whether simple or definitional) of the thing signified; the thing signified (*res significata*) was the property or the nature characterizing individual external objects; and the referent (*suppositum*) was the individual external object itself. In the fourteenth century, further developments took place. On the one hand, there was a new focus on the notion of a mental language superior to spoken language, and concepts, as parts of this mental language, were themselves regarded as having signification. On the other hand, William Ockham and his followers not only denied the existence of common natures but insisted that spoken words did not signify concepts. As a result, both spoken words and the concepts to which spoken words are subordinated have the same significates, namely individual things and their individual properties (e.g., the beardedness of Socrates).

In addition to having signification, terms were also said to have modes of signifying (*modi significandi*). These modes of signifying were related to the term's logical and grammatical functions, and include such essential features as being a noun, verb, or adjective, and such accidental features as time (which includes tense without being limited to it), gender, and case. More generally, they included being abstract (e.g., justice) and concrete (e.g. just). They also include modes of predication, related to Aristotle's ten categories, such as substantial (e.g. horse), qualitative (e.g., brown), quantitative (e.g. square), relative and so on. The notion of modes of signifying was developed from the early twelfth century on, though it was specially emphasized by the speculative grammarians of the late thirteenth century.

It is important to recognize that words were thought to be endowed as units both with their signification and with nearly all their modes of signifying in advance of the role they would subsequently play in propositions. Moreover, the doctrine of common natures suggested that terms, at least those terms which seem to fall within Aristotle's ten categories (substance, quality, quantity and so on), each correspond to a common nature and so have a signification which is fixed and precise. This meant that questions of use and context, though explored by medieval logicians, for instance through supposition theory, were not thought to be crucial to the determination of a term as equivocal, analogical or univocal. It also meant that terms which did not fit into Aristotle's categorial framework needed a special account. This problem relates especially to theology, because God was thought to transcend the categories in the sense that none of them apply to him, and also to metaphysics, because of the so-called transcendental terms, ‘being’, ‘one’, ‘good’. These transcend the categories in the sense that they belong no more to one category than to another, and they do not correspond to common natures.

2. Problems in Logic, Theology, and Metaphysics

In order to understand how the theory of analogy arose we have to bear in mind the history of education in the Latin-speaking western part of Europe. During the so-called dark ages, learning was largely confined to monasteries, and people had access to very few texts from the ancient world. This situation had changed dramatically by the beginning of the thirteenth century. The first universities (Bologna,

Paris, Oxford) had been established, and the recovery of the writings of Aristotle supplemented by the works of Islamic philosophers was well under way.

One source for the theory of analogy is the doctrine of equivocal terms found in logic texts. Until the early twelfth century, the only parts of Aristotle's logic to be available in Latin were the *Categories* and *On Interpretation*, supplemented by a few other works including the monographs and commentaries of Boethius. The *Categories* opens with a brief characterization of terms used equivocally, such as ‘animal’ used of real human beings and pictured human beings, and terms used univocally, such as ‘animal’ used of human beings and oxen. In the first case, the spoken term is the same but there are two distinct significates or intellectual conceptions; in the second case, both the spoken term and the significate are the same. We should note that equivocal terms include homonyms (two words with the same form but different senses, e.g., ‘pen’), polysemous words (one word with two or more senses), and, for medieval thinkers, proper names shared by different people. By the mid-twelfth century the rest of Aristotle's logic had been recovered, including the *Sophistical Refutations* in which Aristotle discusses three types of equivocation and how these contribute to fallacies in logic. For writers throughout the later middle ages, the discussion of analogical terms was fitted into the framework of univocal and equivocal terms provided by Aristotle and his commentators. We will return to this below.

Twelfth-century theology is another important source for the doctrine of analogy, for twelfth-century theologians such as Gilbert of Poitiers and Alan of Lille explored the problem of divine language in depth. Their work initially sprang from works on the Trinity by Augustine and Boethius. These authors insisted that God is absolutely simple, so that no distinctions can be made between God's essence and his existence, or between one perfection, such as goodness, and another, such as wisdom, or, more generally, between God and his properties. New attention was also paid to Greek theologians, especially Pseudo-Dionysius. These theologians insisted on God's absolute transcendence, and on what came to be called negative theology. We cannot affirm anything positive about God, because no affirmation can be appropriate to a transcendent being. It is better to deny properties of God, saying for instance that he is not good (i.e., in the human sense), and still better to say that God is not existent but super-existent, not substance but super-substantial, not good but super-good. These theological doctrines raised the general problem of how we can speak meaningfully of God at all, but they also raised a number of particular problems. Must we say that "God is justice" means the same as "God is just"? Must we say that "God is just" means the same as "God is good"? Can we say that God is just and that Peter is just as well? For our purposes, this last question is the most important, for it raises the question of one word used of two different realities.

The third source for doctrines of analogy is metaphysics. The first part of Aristotle's *Metaphysics* had been translated by the mid-twelfth century, though the full text was recovered only gradually. One crucial text is found in *Metaphysics* 4.2 (1003a33-35): "There are many senses (*multis modis*) in which being (*ens*) can be said, but they are related to one central point (*ad unum*), one definite kind of thing, and are not equivocal. Everything which is healthy is related to health.... and everything which is medical to medicine...." In this text, Aristotle raises the general problem of the word ‘being’ and its different senses, and he also introduces what is known as *pros hen* equivocation or focal meaning, the idea that different senses may be unified through a relationship to one central sense. Another foundational text is

from Avicenna's *Metaphysics*, also translated into Latin during the twelfth century, where he writes that being (*ens*) is neither a genus nor a predicate predicated equally of all its subordinates, but rather a notion (*intentio*) in which they agree according to the prior and the posterior. As we shall see below, this reference to the prior and the posterior is particularly important.

3. History of the Word ‘Analogy’

The Latin term ‘*analogia*’ had various senses. In scriptural exegesis, according to Aquinas, analogy was the method of showing that one part of scripture did not conflict with another. In rhetoric and grammar, analogy was the method of settling a doubt about a word's form by appeal to a similar and more certain case. Several twelfth-century theologians use the word in this sense. In translations of Pseudo-Dionysius, analogy refers to an angel's cognitive capacity in relation to lower or higher beings. In logic, authors were aware that the Greek word ‘*analogia*’, sometimes called ‘*analogia*’ in Latin, but often translated as ‘*proportio*’ or ‘*proportionalitas*’, referred to the comparison between two proportions. However, by the 1220s theologians had begun to use ‘analogy’ in the new sense of a word said in a prior and a posterior sense, and by the 1250s this new use was embedded in the logical texts. The phrase "in a prior and a posterior sense" came into use before the word ‘analogy’ was employed, and seems to have been derived from Arabic philosophy. H. A. Wolfson has presented evidence for Aristotle's recognition of a type of term intermediate between equivocal and univocal terms, some instances of which were characterized by their use according to priority and posteriority. He showed that Alexander of Aphrodisias called this type of term ‘ambiguous’ and that the Arabic philosophers, starting with Alfarabi, made being said in a prior and a posterior sense the main characteristic of all ambiguous terms.

So far as the medieval Latin west is concerned, the main sources for the notion of an ambiguous term said in a prior and a posterior sense are Algazel and Avicenna, both of whom became known in the second half of the twelfth century, and both of whom used the notion to explain uses of the word ‘being’. The word ‘analogical’ soon replaced the word ‘ambiguous’ in Latin authors. Alexander of Hales in his *Glossa* which dates from the 1220s links being said in a prior and a posterior sense with ambiguity and (in one manuscript) with analogy. In the writings of Philip the Chancellor from the same decade, being said in a prior and a posterior sense is called ‘analogy’. In logic, the word ‘analogy’ in the new sense appears in the *Summe metenses*, once dated around 1220, but now thought to be by Nicholas of Paris, writing between 1240 and 1260. The new use of ‘analogy’ rapidly became standard in both logicians and theologians.

4. Divisions of Equivocation

In order to understand the way in which theories of analogy developed, we need to consider the divisions of equivocation found in medieval authors. In his commentary on the *Categories*, Boethius presented a series of divisions which he took from Greek authors. The first division was into chance equivocals and deliberate equivocals. In the first case, the occurrences of the equivocal term were totally unconnected, as when a barking animal, a marine animal, and a constellation were all called ‘*canis*’ (dog). Chance

equivocation was also called pure equivocation, and it was carefully distinguished from analogy by later writers. In the second case, that of deliberate equivocation, some intention on the part of the speakers was involved, and the occurrences of the equivocal term could be related in various ways. Boethius himself gave four subdivisions. These are found in various later sources, including Ockham's commentary on the *Categories*, but as we shall see, other subdivisions became more popular.

The first of Boethius's four subdivisions was similitude, used of the case of the noun ‘animal’ said of both real human beings and pictured human beings. Medieval logicians seem to have been totally unaware of the fact that the Greek word used by Aristotle was genuinely polysemous, meaning both animal and image, and they explained the extended use of ‘animal’ in terms of a likeness between the two referents -- a likeness which had nothing to do with the significate of the term ‘animal’, which picks out a certain kind of nature, but which was nonetheless more than metaphorical in that the external shape of the pictured object does correspond to that of the living object. Those medieval authors whose discussion of equivocation was very brief tended to use this example, and they often claimed that Aristotle introduced it in order to accommodate analogy as a kind of equivocation. On the other hand, authors whose discussion was more extensive tended to drop both the example and the subdivision of similitude.

Boethius's second type of equivocation is ‘*analogia*’ in the Greek sense, and the standard example was the word ‘*principium*’ (principle or origin), which was said to apply to unity with respect to number and to point with respect to a line, or to both the source of a river and the heart of an animal. ‘*Principium*’ is a noun and, as such, might be expected to pick out a common nature, but although a unity, a point, a source and a heart can all be called ‘*principium*’ with equal propriety, there is no common nature involved. Mathematical objects, rivers, and hearts, represent not merely different natural kinds, but different categories, in that mathematical objects fall under the category of quantity, and hearts at least under the category of substance. What allows these disparate things to be grouped together is a similarity of relations: a source is to a river as a heart is to an animal -- or so it was claimed. While theologians, including Aquinas himself in De veritate, and the fourteenth-century Dominican Thomas Sutton, occasionally make use of this type of analogy, most logicians do not even mention it.

The last two subdivisions found in Boethius are ‘of one origin’ (*ab uno*), with the example of the word ‘medical’, and ‘in relation to one’ (*ad unum*), with the example of the word ‘healthy’. These subdivisions correspond to Aristotle's *pros hen* equivocation. The example ‘healthy’ (*sanum*) as said of animals, their diet, and their urine is particularly important here. ‘Sanum’, like other adjectives, was classified as a concrete accidental term. As such, it did not fall within an Aristotelian category, since its primary signification had two elements whose combination was variously explained. On the one hand, some kind of reference is made to the abstract entity health, which belongs to the category of quality; on the other hand, some kind of reference is made to an external object which belongs to the category of substance. This dual reference precludes the term from picking out a natural kind, though in the case of other adjectives, such as ‘brown’, no problem is caused thereby. Brown things may not form a natural kind, but at least they are all physical objects, and ‘brown’ is used in the same sense of each one. ‘Healthy’, however, is more complicated. To say that Rover is healthy is to say that Rover is a thing having health, and obviously this analysis can't be applied to diet, which is called healthy only because it causes health

in an animal, or to urine, which is called healthy only when it is the sign of health in an animal. Whatever the properties which characterize urine and food, they are different from those characterizing the animal.

5. Divisions of Analogy

Boethius's subdivisions had one major failure: they did not seem to accommodate the different uses of the word ‘being’ (*ens*). As a result, many authors used a new threefold division which included Boethius's last two subdivisions and one more. They presented the division as a division of deliberate equivocals, and they identified deliberate equivocals with analogical terms. This threefold division of analogy was established in the thirteenth-century, in response to a remark by Averroes in his commentary on the *Metaphysics* to the effect that Aristotle had classified ‘healthy’ as a case of relationship to one thing as an end, ‘medical’ as a case of relationship to one thing as an agent, and ‘being’ (*ens*) as a case of relationship to one subject. It is found in Thomas Aquinas's own commentary on the *Metaphysics*, as well as in his fifteenth-century follower Capreolus. An analogical term is now seen as one which is said of two things in a prior and a posterior sense, and it is grounded in various kinds of attribution or relationship to the primary object: food is healthy as a cause of a healthy animal, a procedure is medical when applied by a medical agent, a quality has being by virtue of the existent substance that it characterizes.

A second threefold division of analogy arose from reflection on the relationship between equivocal and analogical terms. Analogical terms were said to be intermediaries between equivocal and univocal terms, and the standard view was that analogical terms were intermediary between chance equivocals and univocals, and hence that they were to be identified with deliberate equivocals. The notion of an intermediary term, however, is open to more than one interpretation, and some authors went further in suggesting that at least some analogical terms were intermediary between univocals and deliberate equivocals, so that they were not equivocal in any of the normal senses at all. Towards the end of the thirteenth century, an anonymous commentator on the *Sophistical Refutations* gives the following classification. First, there are analogical terms which are univocal in a broad sense of ‘univocal’. Here reference was made to genus terms such as ‘animal’. Human beings and donkeys participate equally in the common nature animal, but are not themselves equal, since human beings are more perfect than donkeys. This type of analogy was routinely discussed in response to a remark Aristotle had made in *Physics* VII (249a22-25) which, in Latin translation, asserted that many equivocations are hidden in a genus. Medieval logicians felt obliged to fit this claim into the framework of equivocation and analogy, even if the consensus was that in the end the use of genus terms was univocal. Second, there are those analogical terms such as ‘being’ (*ens*) which are not equivocal, because only one concept or nature (*ratio*) seems to be involved, and which are not univocal either, because things participate this one ratio unequally, in a prior and a posterior way. It is these terms which are the genuine intermediaries. Third, there are those analogical terms which are deliberate equivocals, because there are two concepts or natures (*rationes*) which are participated in a prior and a posterior way. The example here was ‘healthy’. This second threefold division was much discussed. Duns Scotus bitterly criticized it in his earlier logical writings. Walter Burley claimed that both the first and the second kinds of analogical term could properly be regarded as univocal in a wide sense. The division was popular in the fifteenth century with such

Thomists as Capreolus, who realized its closeness to the account given by Aquinas in his *Sentences* commentary.

6. Thomas Aquinas

Despite the vast modern literature devoted to Aquinas's theory of analogy, he has very little to say about analogy as such. He uses a general division into equivocal, univocal, and analogical uses of terms, and he presents both of the threefold divisions of analogy mentioned in the previous section, but he offers no prolonged discussion, and he writes as if he is simply using the divisions, definitions, and examples with which everyone is familiar. His importance lies in the way he used this standard material to present an account of the divine names, or how it is we can meaningfully use such words as ‘good’ and ‘wise’ of God.

The background to this account has to be understood in terms of Aquinas's theology and metaphysics. Three doctrines are particularly important. First, there is the distinction between being existent, good, wise, and so on, essentially, and being existent, good, wise, and so on, by participation. God is whatever he is essentially, and as a result he is existence itself, goodness itself, wisdom itself. Creatures are existent, good, wise, only by sharing in God's existence, goodness, and wisdom, and this sharing has three features. It involves a separation between the creature and what the creature has; it involves a deficient similarity to God; and it is based on a causal relation. What is essentially existent or good is the cause of what has existence or goodness by participation. Second, there is the general doctrine of causality according to which every agent produces something like itself. Agent causality and similarity cannot be separated. Third, there is Aquinas's belief that we are indeed entitled to claim that God is existent, good, wise, and so on, even though we cannot know his essence.

Against this background, Aquinas asks how we are to interpret the divine names. He argues that they cannot be purely equivocal, for we could not then make intelligible claims about God. Nor can they be purely univocal, for God's manner of existence and his relationship to his properties are sufficiently different from ours that the words must be used in somewhat different senses. Hence, the words we use of God must be analogical, used in different but related senses. To be more precise, it seems that such words as ‘good’ and ‘wise’ must involve a relationship to one prior reality, and they must be predicated in a prior and a posterior sense, for these are the marks of analogical terms.

Nonetheless, the divine names do not function exactly like ordinary analogical terms such as ‘healthy’. We need to begin by making use of the distinction between the thing signified (a nature or property) and the mode of signifying. All the words we use have a creaturely mode of signifying in that they imply time and composition, neither of which can pertain to God. When speaking of God, we must recognize this fact, and attempt to discount it. To say "God is good" is not to imply that God has a separable property, goodness, and that he has it in a temporally limited way. God is eternally identical to goodness itself. But even when we have discounted the creaturely mode of signifying, we are left with the fact that God's goodness is not like our goodness. This is where the analogy of attribution enters the picture.

In his early writings, Aquinas questioned whether the standard account of the analogy of attribution was sufficient for his purposes. In his commentary on the *Sentences*, he suggests that there is one kind of analogy in which the analogical term is used in a prior and a posterior sense, and another kind of analogy, the analogy of imitation, which applies to God and creatures. In his *De veritate*, he argues that the analogy of attribution involves a determinate relation, which cannot hold between God and creatures, and that the analogy of proportionality must be used for the divine names. We must compare the relation between God and his properties to the relation between creatures and their properties. This solution was deeply flawed, given that the problem of divine names arises precisely because the relationship of God to his properties is so radically different from our relation to our properties. Accordingly, in his later discussions of the divine names, notably in the *Summa contra gentiles* and the *Summa theologiae*, Aquinas returns to the analogy of attribution, but links it much more closely with his doctrines of causal similitude. As Montagnes has pointed out, he came to place much greater emphasis on agent causation, the active transmission of properties from God to creatures, than on exemplar causality, the creature's passive reflection or imitation of God's properties. In this context, Aquinas makes considerable use of his ontological distinction between univocal causes, whose effects are fully like them, and non-univocal causes, whose effects are not fully like them. God is an analogical cause, and this is the reality that underlies our use of analogical language.

Aquinas's views about agent causality explain his insistence on definitional inclusion He says explicitly that the term said in a prior way must be included in the definition of the posterior, just as the definition of healthy food includes a reference to the health of the animal. In the divine case, the reference is not direct or explicit, but is a function of our recognition that when humans are said to be good, this means that they have a participated goodness which must be caused by that which is goodness itself. The nature of this causal relation between God and creature also helps to explain the sense in which terms are said in a prior way of God. So far as imposition is concerned, terms are given their signification on the basis of creaturely effects, and, before we learn about God, we may think that their prior use is to refer to creaturely perfections. However, when we come to know God as the first cause and fully perfect being, we recognize that their prior application is to God. Finally, Aquinas's causal doctrines help us to explain his insistence on the distinction between the analogy of many-to-one and the analogy of one-to-another. In the first case, both food and medicine are said to be healthy because each is related to something else, the health of an animal. In the second case, food is said to be healthy because of its relation to the health of an animal. Only the second kind of analogy applies to the divine names, for no non-metaphorical name we apply to God can ever be explained in terms of something other than God. Our use of divine names has to reflect God's absolute priority.

7. John Duns Scotus and the Role of Concepts

One of the issues that Aquinas touched on but did not settle was that of the number of *rationes* an analogical term was associated with. This issue stemmed from Aristotle's *Categories*. As translated by Boethius, Aristotle introduced the distinction between univocal and equivocal terms by claiming that whereas univocal terms were subordinated to one *substantiae ratio*, equivocal terms were subordinated to more than one *substantiae ratio*. The word ‘*ratio*’ here is capable of various interpretations, including

‘definition or description’, ‘analysis’, or ‘concept’, but by the early fourteenth century logicians and theologians came to agree that the appropriate interpretation was ‘concept’. The second threefold division of analogy given above suggests the importance of a focus on concepts; and the question of how many concepts an analogical term was subordinated to came to be central. The nominalists held that analogical terms were straightforwardly equivocal terms subordinated to two distinct concepts but the Thomists were split. Analogical terms could be viewed as subordinated to an ordered cluster of concepts (possibly but not necessarily described as a disjunction of concepts); or they could be subordinated to a single concept which represents in a prior and a posterior manner (*per prius et posterius*).

The issue was considerably complicated by the influence of John Duns Scotus and his argument that ‘being’ was not analogical but univocal. Scotus believed that without a unified conception of being, theology as a science would be impossible, and we would have no natural knowledge of God. Accordingly, he rejected the view that for a term to be univocal it had to be a strictly categorial term, picking out some natural kind or other. He argued that it was sufficient for univocity that contradiction would arise when the term was affirmed and denied of the same thing. He then argued that ‘being’ (*ens*) was a univocal term subordinated to a single univocal concept. Even for those within the Thomistic tradition, Scotus' arguments about the univocity of ‘being’ had to be taken seriously. On the one hand, the word does not seem to be straightforwardly equivocal, in the sense of being subordinated to more than one concept, for we at least have the illusion of being able to grasp ‘being’ as a general term. As Scotus pointed out, in an argument reproduced by all who considered the issue, we can grasp that something is a being while doubting whether it is a substance or an accident, and this surely involves having a relatively simple concept of being at our disposal. On the other hand, there does not seem to be any common nature involved, and in the absence of a common nature, Thomists thought that to call the term ‘univocal’ was inappropriate. What was needed was a way of allowing the concept to enjoy some kind of unity, while allowing the word to have a significate that was not a simple common nature. For many thinkers from the early fourteenth century onward, the distinction between formal and objective concepts provided the answer.

The formal concept was the act of mind or conception that represented an object, and the objective concept was the object represented. If the spoken word ‘being’ corresponds to just one formal concept (a point on which there were some differences of opinion), the focus of discussion shifts to the status of the objective concept. Is it the actual thing in the world which is thought about; is it a common nature or some other kind of intermediary entity which is distinct from the external object without being mind-dependent; or is it a special kind of mind-dependent object which has only objective being, the being of being thought (*esse objective*, *esse cognitum*)? In the case of ‘being’, since we are clearly not talking about either an individual thing or a common nature, we get back to the original set of questions about analogical concepts, now posed at a different level. That is, are we talking about a special ordering intrinsic to a single objective concept, or are we talking about an ordered sequence of objective concepts which corresponds to the one formal concept?

8. Cardinal Cajetan: A New Approach

In 1498 Thomas de Vio, Cardinal Cajetan, wrote a little book called *On the Analogy of Names* which he intended to supplement his commentary on Aristotle's *Categories*. The book rapidly became very popular, and it had a significant effect on subsequent discussions of analogy. Part of the work is devoted to formal and objective concepts and ways in which the latter can be ordered, but Cajetan also offered a new account of types of analogy. He began by presenting the second threefold division. He called the first type of analogy, the case of genus terms, the analogy of inequality, and dismissed it as unimportant, indeed, not properly analogy at all. He called the second type the analogy of attribution, and here he made two changes. First, he gave a new account of its subdivisions by adding Boethius's subdivision, similitude, to the first threefold division involving attribution to one efficient cause, one end, and one subject. He described the resulting four subdivisions in terms of Aristotle's four causes. Second, he claimed that attribution involved only extrinsic denomination. That is, in each case of attribution, only the prior object is intrinsically characterized by the property in question, e.g., health.

He called the third type of analogy the analogy of proportionality. It included metaphor and what he called proper proportionality. The latter, he said, is analogy in the Greek sense of the word, and is the only true kind of analogy. Moreover, it involves only intrinsic denomination: both the primary and the secondary object referred to are characterized by the property in question. While the word ‘being’ can be used in accordance with attribution, Cajetan claimed that it, and all other metaphysically significant analogical terms, principally belonged in this last division. Both in his insistence on the priority of the analogy of proper proportionality and in his use of the distinction between intrinsic and extrinsic denomination, Cajetan departed from earlier medieval discussions of analogy. Unfortunately, many later commentators have been misled into taking his account as a typical one, and, even more unfortunately, as a useful summary of the doctrines of Aquinas.

Bibliography

- Ashworth, E. J., 1991, "Signification and Modes of Signifying in Thirteenth-Century Logic: A Preface to Aquinas on Analogy", *Medieval Philosophy and Theology* **1**: 39-67.
- ---------, 1992, "Analogy and Equivocation in Thirteenth-Century Logic: Aquinas in Context", *Mediaeval Studies* **54**: 94-135.
- ---------, 1992, "Analogical Concepts: The Fourteenth-Century Background to Cajetan", *Dialogue* **31**: 399-413.
- ---------, 1995, "Suárez on the Analogy of Being: Some Historical Background", *Vivarium* **33**: 50-75.
- ---------, 1996, "Analogy, Univocation, and Equivocation in some Early Fourteenth-Century Authors", in *Aristotle in Britain during the Middle Ages*, John Marenbon (ed.), Belgium: Brepols, pp. 233-247.
- ---------, 1997, "Analogy and Equivocation in Thomas Sutton, O.P." in *Vestigia, Imagines, Verba: Semiotics and Logic in Medieval Theological Texts (XIIth-XIVth Century)*, Costantino Marmo (ed.), Turnhout: Brepols, pp. 289-303.
- Boulnois, Olivier. 1996, "Duns Scot, théoricien de l'analogie de l'être" in *John Duns Scotus: Metaphysics and Ethics*, edited by Ludger Honnefelder, Rega Wood, and Mechthild Dreyer,

Studien und Texte zur Geistesgeschichte des Mittelalters, **53**, Leiden, New York, Cologne: E. J. Brill, pp. 293-315.

- Libera, Alain de, 1989, "Les sources gréco-arabes de la théorie médiévale de l'analogie de l'être", *Les études philosophiques* **3/4**: 319-345.
- McInerny, Ralph, 1986, "The Analogy of Names is a Logical Doctrine" in idem, *Being and Predication: Thomistic Interpretations*, Washington, D.C.: The Catholic University of America Press, pp. 279-286.
- McInerny, Ralph, 1992, "Aquinas and Analogy: Where Cajetan Went Wrong", *Philosophical Topics* **20**/2: 103-124, [*Medieval Philosophy*, Sandra Edwards (ed.)]
- Montagnes, Bernard, 1963, "La doctrine de l'analogie de l'être d'après Saint Thomas d'Aquin", *Philosophes médiévaux* **6**, Louvain: Publications Universitaires; Paris: Béatrice-Nauwelaerts.
- Rosier, Irène, 1995, "Res significata et modus significandi: Les implications d'une distinction médiévale" in *Sprachtheorien in Spätantike und Mittelalter*, Sten Ebbesen (ed.), Tübingen: Gunter Narr Verlag, pp. 135-168.
- Wolfson, H. A., 1938, "The Amphibolous Terms in Aristotle, Arabic Philosophy and Maimonides", *Harvard Theological Review* **31**: 151-173; reprinted in idem, *Studies in the History of Philosophy and Religion*, I. Twersky and G. H. Williams (eds.), Cambridge, MA and London: Harvard University Press, 1973, vol.1, pp. 455-477.

Other Internet Resources

- Paul Vincent Spade's Mediaeval Logic and Philosophy pages
- Studies in Medieval and Early Modern Intellectual History
- Thomas Aquinas: List of available texts

Related Entries

Aquinas, Saint Thomas | Duns Scotus, John

E. Jennifer Ashworth
ejashwor@uwaterloo.ca

First published: November 29, 1999
Content last modified: November 29, 1999

John Duns Scotus

John Duns Scotus (1265/66-1308) was one of the most important and influential philosopher-theologians of the High Middle Ages. His brilliantly complex and nuanced thought, which earned him the nickname "the Subtle Doctor," left a mark on discussions of such disparate topics as the semantics of religious language, the problem of universals, divine illumination, and the nature of human freedom. This essay first lays out what is known about Scotus's life and the dating of his works. It then offers an overview of some of his key positions in four main areas of philosophy: natural theology, metaphysics, the theory of knowledge, and ethics and moral psychology.

1. Life and Works

1.1 The life of John Duns the Scot

‘Scotus’ is a nickname: it identifies Scotus as a Scot. His family name was Duns, which was also the name of the Scottish village in which he was born, just a few miles from the English border. We do not know the precise date of his birth, but we do know that Scotus was ordained to the priesthood in the Order of Friars Minor -- the Franciscans -- at Saint Andrew's Priory in Northampton, England, on 17 March 1291. The minimum age for ordination was twenty-five, so we can conclude that Scotus was born before 17 March 1266. But how much before? The conjecture, plausible but by no means certain, is that Scotus would have been ordained as early as canonically permitted. Since the Bishop of Lincoln (the diocese that included Oxford, where Scotus was studying, as well as St Andrew's Priory) had ordained priests in Wycombe on 23 December 1290, we can place Scotus's birth between 23 December 1265 and 17 March

1266.

It appears that Scotus began his formal studies at Oxford in October 1288 and concluded them in June 1301. In the academic year 1298-99 he commented on the *Sentences* of Peter Lombard. We know that by the fall of 1302 Scotus was lecturing on the *Sentences* in Paris. In June 1303 Scotus was expelled from France along with eighty other friars for taking the Pope's side in a dispute with the king. They were allowed to return in April 1304; it appears that Scotus completed his lectures on the *Sentences* not long thereafter. On 18 November 1304 Scotus was appointed the Franciscan regent master in theology at Paris. For reasons no one quite understands, Scotus was transferred to the Franciscan *studium* at Cologne, probably beginning his duties as lector in October 1307. He died there in 1308; the date of his death is traditionally given as 8 November.

1.2 Scotus's works

It is generally agreed that Scotus's earliest works were his "*parva logicalia*" (little logical works): questions on Porphyry's *Isagoge* and Aristotle's *Categories*, *Peri hermeneias*, and *De sophisticis elenchis*. These probably date to around 1295; the possibly inauthentic *Quaestiones super De anima* is also very likely an early work. Scotus's other Aristotelian commentary, the *Quaestiones subtilissimae super Metaphysicam Aristotelis*, seems to have been started early; but Book 9 is probably late, and it is possible that Books 6 through 9 are all late or were at least revised later in Scotus's career.

Things really get complicated when we come to Scotus's commentaries on the four books of *Sentences* of Peter Lombard, since it appears that he commented on the *Sentences* on several occasions, and the relations among the various versions are not always clear. Certainly the *Lectura* presents us with Scotus's notes for his Oxford lectures on Books 1 and 2 of the *Sentences* in 1298-99 (or possibly 1300-01). There is a *Reportatio* (i.e., a transcript based on student notes) of lectures at Cambridge; this probably dates to some time between 1297 and 1300. (It has never been critically edited and exists in only three manuscripts.) There is an *Ordinatio* (i.e., a version prepared for publication by the author himself) of lectures at Oxford, based in part on the *Lectura* and on material from his lectures in Paris. The *Ordinatio* is generally taken to be Scotus's premier work, but unfortunately the critical edition is nowhere near complete (at this time it extends through Book 2, distinction 3). Scotus seems to have been revising it up to his death. Finally, Scotus lectured on the *Sentences* several times at Paris, and there are various *Reportationes* of these lectures, all dating from the period 1302-1307. Easily the most important is the *Reportatio examinata* of Book 1; the designation *examinata* indicates that it was examined and corrected by Scotus himself. Unfortunately it has never been critically edited.

In addition to these works, we have 46 short disputations called *Collationes* dating from 1300-1305, a late work in natural theology called *De primo principio*, and *Quaestiones Quodlibetales* from Scotus's days as regent master (either Advent 1306 or Lent 1307). Finally, there is a work of dubious authenticity called *Theoremata*.

2. Natural Theology

2.1 Some methodological preliminaries

Natural theology is, roughly, the effort to establish the existence and nature of God by arguments that in no way depend on the contents of a purported revelation. But is it even *possible* for human beings to come to know God apart from revelation? Scotus certainly thinks so. Like any good Aristotelian, he thinks all our knowledge begins in some way with our experience of sensible things. But he is confident that even from such humble beginnings we can come to grasp God.

Scotus agrees with Thomas Aquinas that all our knowledge of God starts from creatures, and that as a result we can only prove the existence and nature of God by what the medievals call an argument *quia* (reasoning from effect to cause), not by an argument *propter quid* (reasoning from essence to characteristic). Aquinas and Scotus further agree that, for that same reason, we cannot know the essence of God in this life. The main difference between the two authors is that Scotus believes we can apply certain predicates univocally -- with exactly the same meaning -- to God and creatures, whereas Aquinas insists that this is impossible, and that we can only use analogical predication, in which a word as applied to God has a meaning different from, although related to, the meaning of that same word as applied to creatures. (See medieval theories of analogy for details.)

Scotus has a number of arguments for univocal predication and against the doctrine of analogy (*Ordinatio* 1, d. 3, pars 1, q. 1-2, nn. 26-55). One of the most compelling uses Aquinas's own view against him. Aquinas had said that all our concepts come from creatures. Scotus says, very well, where will that analogous concept come from? It can't come from anywhere. If all our concepts come from creatures (and Scotus doesn't deny this), then the concepts we apply to God will also come from creatures. They won't just be *like* the concepts that come from creatures, as in analogous predication; they will have to be *the very same* concepts that come from creatures, as in univocal predication. Those are the only concepts we can have -- the only concepts we can possibly get. So if we can't use the concepts we get from creatures, we can't use any concepts at all, and so we can't talk about God -- which is false.

Another argument for univocal predication is based on an argument from Anselm. Consider all predicates, Anselm says. Now get rid of the ones that are merely relatives, since no relative expresses the nature of a thing as it is in itself. (So we're not talking about such predicates as "supreme being" or "Creator," since even though those properly apply to God, they don't tell us anything about what God is in himself, only about how he is related to other things.) Now take the predicates that are left. Here's the test. Let *F* be our predicate-variable. For any *F*, either

> (a) It is in every respect better to be *F* than not to be *F*.
>
> ~or~
>
> (b) It is in some respect better to be not-*F* than *F*.

A predicate will fall into the second category if and only if it implies some sort of limitation or deficiency. Anselm's argument is that we can (indeed must) predicate of God every predicate that falls into the first category, and that we cannot predicate of God any predicate that falls into the second (except metaphorically, perhaps). Scotus agrees with Anselm on this point (as did Aquinas: see *SCG* I.30). Scotus has his own terminology for whatever it is in every respect better to be than not to be. He calls such things "pure perfections" (*perfectiones simpliciter*). A pure perfection is any predicate that does not imply limitation.

So Scotus claims that pure perfections can be predicated of God. But he takes this a step further than Anselm. He says that they have to be predicated *univocally* of God; otherwise the whole business of pure perfections won't make any sense. Here's the argument. If we are going to use Anselm's test, we must first come up with our concept -- say, of good. Then we check out the concept to see whether it is in every respect better to be good than not-good. We realize that it is, and so we predicate 'good' of God. That test obviously won't work unless it's the same concept that we're applying in both cases.

One can see this more clearly by considering the two possible ways in which one might deny that the same concept is applied to both God and creatures. One might say that the concept of the pure perfection applies only to creatures, and the concept we apply to God has to be something different; or one might try it the other way around and say that the concept of the pure perfection applies only to God, and the concept we apply to creatures has to be something different. Take the first possibility. If we come up with the idea of a pure perfection from creatures and don't apply the same concept to God, we're saying that we can come up with something that is in every respect better to be than not to be, but it doesn't apply to God. Such a view would destroy the idea that God is the greatest and most perfect being. So then one might try the second possibility: the concept of the pure perfection really applies only to God. Scotus points out that that can't be right either. For then the perfection we apply to creatures won't be the pure perfection any more, and so the creature wouldn't be better off for having this pseudo-perfection. But the whole way in which we came up with the idea of the pure perfection in the first place was by considering perfections in creatures -- in other words, by considering what features made creatures better in every respect. So this possibility gets the test backwards: it says that we have to start with knowing what features God has and thereby determining what is a pure perfection, but in fact we first figure out what the pure perfections are and thereby know what features God has.

Not only can we come up with concepts that apply univocally to God and creatures, we can even come up with a *proper* (distinctive) concept of God. Now in one sense we can't have a proper concept of God in this life, since we can't know his essence as a particular thing. We know God in the way that we know, say, a person we have heard about but have never met. That is, we know him through general concepts that can apply both to him and to other things. In another sense, though, we can have a proper concept of God, that is, one that applies only to God. If we take any of the pure perfections to the highest degree, they will be predicable of God alone. Better yet, we can describe God more completely by taking all the pure perfections in the highest degree and attributing them all to him.

But these are all composite concepts; they all involve putting two quite different notions together: 'highest' with 'good', 'first' with 'cause', and so on. Scotus says that we can come up with a relatively

simple concept that is proper to God alone, the concept of "infinite being." Now that concept might seem to be every bit as composite as "highest good" or "first cause," but it's really not. For "infinite being" is a concept of something essentially one: a being that has infinity (unlimitedness) as its intrinsic way of existing. I will return to the crucial role of the concept of infinite being in Scotus's natural theology after I examine his proof of the existence of God.

2.2 Proof of the existence of God

Scotus's argument for the existence of God is rightly regarded as one of the most outstanding contributions ever made to natural theology. The argument is enormously complex, with several sub-arguments for almost every important conclusion, and I can only sketch it here. (Different versions of the proof are given at *Lectura* 1, d. 2, q. 1, nn. 38-135; *Ordinatio* 1, d. 2, q. 1, nn. 39-190; *Reportatio* 1, d. 2, q. 1; and *De primo principio*.)

Scotus begins by arguing that there is a first agent (a being that is first in efficient causality). Consider first the distinction between essentially ordered causes and accidentally ordered causes. In an accidentally ordered series, the fact that a given member of that series is itself caused is accidental to that member's own causal activity. For example, Grandpa A generates a son, Dad B, who in turn generates a son of his own, Grandson C. B's generating C in no way depends on A -- A could be long dead by the time B starts having children. The fact that B was caused by A is irrelevant to B's own causal activity. That's how an accidentally ordered series of causes works.

In an essentially ordered series, by contrast, the causal activity of later members of the series depends essentially on the causal activity of earlier members. For example, my shoulders move my arms, which in turn move my golf club. My arms are capable of moving the golf club only because they are being moved by my shoulders.

With that distinction in mind, we can examine Scotus's argument for the existence of a first efficient cause:

(1) No effect can produce itself.
(2) No effect can be produced by just nothing at all.
(3) A circle of causes is impossible.
(4) Therefore, an effect must be produced by something else. (from 1, 2, and 3)
(5) There is no infinite regress in an essentially ordered series of causes.

(5a) It is not necessarily the case that a being possessing a causal power C possesses C in an imperfect way.
(5b) Therefore, it is possible that C is possessed without imperfection by some item.

- (5c) If it is not possible for any item to possess C without dependence on some prior item, then it is not possible that there is any item that possesses C without imperfection (since dependence is a kind of imperfection).
- (5d) Therefore, it is possible that some item possesses C without dependence on some prior item. (from 5b and 5c by modus tollens)
- (5e) Any item possessing C without dependence on some prior item is a first agent (i.e., an agent that is not subsequent to any prior causes in an essentially ordered series).
- (5f) Therefore, it is possible that something is a first agent. (from 5d and 5e)
- (5g) If it is possible that something is a first agent, something is a first agent. (For, by definition, if there were no first agent, there would be no cause that could bring it about, so it would not in fact be possible for there to be a first agent.)
- (5h) Therefore, something is a first agent (i.e., an agent that is not subsequent to any prior causes in an essentially ordered series -- Scotus still has to prove that there is an agent that is not subsequent to any prior causes in an accidentally ordered series either. That's what he does in step (6) below). (from 5f and 5g)

(6) It is not possible for there to be an accidentally ordered series of causes unless there is an essentially ordered series.

- (6a) In an accidentally ordered series, each member of the series (except the first, if there is a first) comes into existence as a result of the causal activity of a prior member of the series.
- (6b) That causal activity is exercised in virtue of a certain form.
- (6c) Therefore, each member of the series depends on that form for its causal activity.
- (6d) The form is not itself a member of the series.
- (6e) Therefore, the accidentally ordered series is essentially dependent on a higher-order cause.

(7) Therefore, there is a first agent. (from 4, 5, and 6)

Scotus then goes on to argue that there is an ultimate goal of activity (a being that is first in final causality), and a maximally excellent being (a being that is first in what Scotus calls "pre-eminence").

Thus he has proved what he calls the "triple primacy": there is a being that is first in efficient causality, in

final causality, and in pre-eminence. Scotus next proves that the three primacies are coextensive: that is, any being that is first in one of these three ways will also be first in the other two ways. Scotus then argues that a being enjoying the triple primacy is endowed with intellect and will, and that any such being is infinite. Finally, he argues that there can be only one such being.

2.3 Divine infinity and the doctrine of univocity

In laying out Scotus's proof of the existence of God, I passed rather quickly over the claim that God is infinite. But the divine infinity deserves more detailed treatment. As we have already seen, the concept of "infinite being" has a privileged role in Scotus's natural theology. As a first approximation, we can say that divine infinity is for Scotus what divine simplicity is for Aquinas. It's the central divine-attribute generator. But there are some important differences between the role of simplicity in Aquinas and the role of infinity in Scotus. The most important, I think, is that in Aquinas simplicity acts as an ontological spoilsport for theological semantics. Simplicity is in some sense the key thing about God, metaphysically speaking, but it seriously complicates our language about God. God is supposed to be a subsistent simple, but because our language is all derived from creatures, which are all either subsistent but complex or simple but non-subsistent, we don't have any way to apply our language straightforwardly to God. The divine nature systematically resists being captured in language.

For Scotus, though, infinity is not only what's ontologically central about God, it's the key component of our best available concept of God and a guarantor of the success of theological language. That is, our best ontology, far from fighting with our theological semantics, both supports and is supported by our theological semantics. The doctrine of univocity rests in part on the claim that "[t]he difference between God and creatures, at least with regard to God's possession of the pure perfections, is ultimately one of degree" (Cross [1999], 39). Remember one of Scotus's arguments for univocity. If we are to follow Anselm in ascribing to God every pure perfection, we have to affirm that we are ascribing to God *the very same thing* that we ascribe to creatures: God has it infinitely, creatures in a limited way. One could hardly ask for a more harmonious cooperation between ontology (what God is) and semantics (how we can think and talk about him).

Scotus ascribes to Aquinas the following argument for the divine infinity: If a form is limited by matter, it is finite. God, being simple, is not limited by matter. Therefore, God is not finite. This, as Scotus points out, is a fallacious argument. (It's an instance of denying the antecedent.) But even apart from the fallacy, simplicity is not going to get us infinity. As Scotus puts it: "if an entity is finite or infinite, it is so not by reason of something accidental to itself, but because it has its own intrinsic degree of finite or infinite perfection" (*Ordinatio* 1, d. 1, pars 1, q. 1-2, n. 142). So simplicity does not entail infinity, because finitude is not the result of composition. To look at it another way, Aquinas's conception of infinity is *negative* and *relational*. The infinite is that which is *not bounded by something else*. But Scotus thinks we can have a positive conception of infinity, according to which infinity is not a negative, relational property, but instead a positive, intrinsic property. It is an "intrinsic degree of perfection."

How do we acquire that conception of positive, intrinsic infinity? The story goes like this. We begin with

"the potentially infinite in quantity." According to Aristotle, you can never have an actual quantitative infinity, since no matter how great a quantity you have, you can always have more. What you can have (and in fact do have, Aristotle thinks) is a quantitative infinity by successive parts. The next step is to imagine that all the parts of that quantitative infinity remained in existence simultaneously. That is, we imagine an actual quantitative infinity. Scotus then asks us to shift from thinking about an actual quantitative infinity to thinking about an actual *qualitative* infinity. Think of some quality (say, goodness) as existing infinitely: so that there is, as it were, no more goodness that you could add to that goodness to make it any greater. That's infinite goodness. But notice that you can't think of infinite goodness as in some way composed of little goodness-bits (just an infinite number of them). If I say that an angel is better than a human being, I can't mean that a human being has a certain number of goodness-bits while the angel has that many plus some extras. Rather, the specific degree of goodness of a thing is just an intrinsic, non-quantitative feature of that thing. Infinite being is just like that. Scotus describes it as "a measure of intrinsic excellence that is not finite." This is why the concept of "infinite being" is the simplest concept available to us for understanding God. Infinity is not some sort of accidental addition to being, but an intrinsic mode of being. Of course, if this is right, then the concepts of ‘infinite goodness’, ‘infinite power’, and so forth, are every bit as simple as the concept of ‘infinite being’. So why does Scotus make such a big deal about ‘infinite being’? Because ‘infinite being’ "virtually contains" all the other infinite perfections of God. That is, we can deduce the other infinite perfections from infinite being. So besides being the next best thing to a simple concept, it's the most theoretically fruitful concept we can have of God in this life.

3. Metaphysics

3.1 The subject matter of metaphysics

Metaphysics, according to Scotus, is a "real theoretical science": it is real in that it treats things rather than concepts, theoretical in that it is pursued for its own sake rather than as a guide for doing or making things, and a science in that it proceeds from self-evident principles to conclusions that follow deductively from them. The various real theoretical sciences are distinguished by their subject matter, and Scotus devotes considerable attention to determining what the distinctive subject matter of metaphysics is. His conclusion is that metaphysics concerns "being *qua* being" (*ens inquantum ens*). That is, the metaphysician studies being simply as such, rather than studying, say, material being as material.

The study of being *qua* being includes, first of all, the study of the transcendentals, so called because they transcend the division of being into finite and infinite, and the further division of finite being into the ten Aristotelian categories. Being itself is a transcendental, and so are the "proper attributes" of being -- one, true, and good -- which are coextensive with being. Scotus also identifies an indefinite number of disjunctions that are coextensive with being and therefore count as transcendentals, such as infinite-or-finite and necessary-or-contingent. Finally, all the pure perfections (see above) are transcendentals, since they transcend the division of being into finite and infinite. Unlike the proper attributes of being and the disjunctive transcendentals, however, they are not coextensive with being. For God is wise and Socrates is wise, but earthworms -- though they are certainly beings -- are not wise.

The study of the Aristotelian categories also belongs to metaphysics insofar as the categories, or the things falling under them, are studied as beings. (If they are studied as concepts, they belong instead to the logician.) There are exactly ten categories, Scotus argues. The first and most important is the category of substance. Substances are beings in the most robust sense, since they have an independent existence: that is, they do not exist *in* something else. Beings in any of the other nine categories, called accidents, exist in substances. The nine categories of accidents are quantity, quality, relation, action, passion, place, time, position, and state (*habitus*).

3.2 Matter and form, body and soul

Now imagine some particular substance, say, me. Suppose I go from being pale to being tan. Now it is still I who exist both before and after the sun has had its characteristic effect on me. This illustrates an important feature of substances: they can successively have contrary accidents and yet retain their numerical identity. This sort of change is known, appropriately enough, as accidental change. In an accidental change, a substance persists through the change, having first one accident and then another. But clearly not all changes are accidental changes. There was once a time when I did not exist, and then I came into existence. We can't analyze this change as an accidental change, since there doesn't seem to be any substance that persists through the change. Instead, a substance is precisely what comes into being; this is not an accidental but a *substantial* change. And yet there must be something that persists even through substantial change, since otherwise we wouldn't have change at all; substances would come to exist from nothing and disappear into nothing. Scotus follows Aristotle in identifying *matter* as what persists through substantial change and *substantial form* as what makes a given parcel of matter the definite, unique, individual substance that it is. (There are also accidental forms, which are a substance's accidental qualities.)

Thus far Scotus is simply repeating Aristotelian orthodoxy, and none of his contemporaries or immediate predecessors would have found any of this at all strange. But as Scotus elaborates his views on form and matter, he espouses three important theses that mark him off from some other philosophers of his day: he holds that there exists matter that has no form whatsoever, that not all created substances are composites of form and matter, and that one and the same substance can have more than one substantial form. Let us examine each of these theses in turn.

First, Scotus argues that there is matter that is entirely devoid of form, or what is known as "prime matter" (*Quaestiones in Metaphysicam* 7, q. 5; *Lectura* 2, d. 12, q. un.). Scholars debate now (just as they debated in Scotus's day) whether Aristotle himself really believed that there is prime matter or merely introduced it as a theoretical substratum for substantial change, believing instead that in actual fact matter always has at least some minimal form (the form of the elements being the most minimal of all). Aquinas denied both that Aristotle intended to posit it and that it could exist on its own. For something totally devoid of form would be utterly featureless; it would be pure potentiality, but not actually anything. Scotus, by contrast, argues that prime matter not only can but does exist as such: "it is one and the same stuff that underlies every substantial change" (King [forthcoming]).

Second, Scotus denies "universal hylemorphism," the view that all created substances are composites of form and matter (*Lectura* 2, d. 12, q. un., n. 55). Universal hylemorphism (from the Greek *hyle*, meaning ‘matter’, and *morphe*, meaning ‘form’) had been the predominant view among Franciscans before Scotus. Saint Bonaventure, for example, had argued that even angels could not be altogether immaterial; they must be compounds of form and "spiritual matter." For matter is potentiality and form is actuality, so if the angels were altogether immaterial, they would be pure actuality without any admixture of potentiality. But only God is pure actuality. But as we have already seen in his affirmation of the existence of prime matter, Scotus simply denies the unqualified equation of matter with potentiality and form with actuality. Prime matter, though entirely without form, is actual; and a purely immaterial being is not automatically bereft of potentiality.

Third, Scotus holds that some substances have more than one substantial form (*Ordinatio* 4, d. 11, q. 3, n. 54). This doctrine of the plurality of substantial forms was commonly held among the Franciscans but vigorously disputed by others. We can very easily see the motivation for the view by recalling that a substantial form is supposed to be what makes a given parcel of matter the definite, unique, individual substance that it is. Now suppose, as many medieval thinkers (including Aquinas) did, that the soul is the one and only substantial form of the human being. It would then follow that when a human being dies, and the soul ceases to inform that parcel of matter, what is left is not the same body that existed just before death. For what made it that very body was its substantial form, which (*ex hypothesi*) is no longer there. When the soul is separated from the body, then, what is left is not a body, but just a parcel of matter arranged corpse-wise. To Scotus and many of his fellow Franciscans it seemed obvious that the corpse of a person is the very same body that existed before death. Moreover, they argued, if the only thing responsible for informing the matter of a human being is the soul, it would seem that (what used to be) the body should immediately dissipate when a person dies. Accordingly, Scotus argues that the human being has at least two substantial forms. There is the "form of the body" (*forma corporeitatis*) that makes a given parcel of matter to be a definite, unique, individual human body, and the "animating form" or soul, which makes that human body alive. At death, the animating soul ceases to vivify the body, but numerically the same body remains, and the form of the body keeps the matter organized, at least for a while. Since the form of the body is too weak on its own to keep the body in existence indefinitely, however, it gradually decomposes.

While Scotus's account of form and matter has clear implications for what happens to the *body* at death, it is less forthcoming about what happens to the *soul*. Can the animating soul survive the death of the body it informs? Scotus considers a number of arguments for the incorruptibility of the human soul, but he finds none of them persuasive. This is not to say that he denies the immortality of the soul, of course, but that he does not think it can be proved by human reason unaided by revelation.

Note that the general tendency of Scotus's theories of form and matter is to allow a high degree of independence to form and matter. In positing the existence of prime matter, Scotus envisions matter as existing without any form; in denying universal hylpemorphism, he envision form as existing without any matter. And the doctrine of the plurality of substantial forms strongly suggests that the human soul is an identifiable individual in its own right. So everything Scotus says in this connection seems to make room for the *possibility* that the soul survives the death of the body and continues to exist as an immaterial

substance in its own right. That this possibility is in fact realized, however, is something we can know only through faith.

3.3 Universals and individuation

The problem of universals may be thought of as the question of what, if anything, is the metaphysical basis of our using the same predicate for more than one distinct individual. Socrates is human and Plato is human. Does this mean that there must be some one universal reality -- humanity -- that is somehow *repeatable*, in which Socrates and Plato both share? Or is there nothing metaphysically common to them at all? Those who think there is some actual universal existing outside the mind are called realists; those who deny extra-mental universals are called nominalists. Scotus was a realist about universals, and like all realists he had to give an account of what exactly those universals are: what their status is, what sort of existence they have outside the mind. So, in the case of Socrates and Plato, the question is "What sort of item is this humanity that both Socrates and Plato exemplify?" A related question that realists have to face is the problem of individuation. Given that there is some extra-mental reality common to Socrates and Plato, we also need to know what it is in each of them that makes them *distinct* exemplifications of that extra-mental reality.

Scotus calls the extra-mental universal the "common nature" (*natura communis*) and the principle of individuation the "haecceity" (*haecceitas*). The common nature is common in that it is "indifferent" to existing in any number of individuals. But it has extra-mental existence only *in* the particular things in which it exists, and in them it is always "contracted" by the haecceity. So the common nature *humanity* exists in both Socrates and Plato, although in Socrates it is made individual by Socrates's *haecceitas* and in Plato by Plato's *haecceitas*. The humanity-of-Socrates is individual and non-repeatable, as is the humanity-of-Plato; yet humanity itself is common and repeatable, and it is ontologically prior to any particular exemplification of it (*Ordinatio* 2, d. 3, pars 1, qq. 1-6, translated in Spade [1994], 57-113).

4. Theory of Knowledge

4.1 Sensation and abstraction

Scotus adopts the standard medieval Aristotelian view that human beings, alone among the animals, have two different sorts of cognitive powers: senses and intellect. The senses differ from the intellect in that they have physical organs; the intellect is immaterial. In order for the intellect to make use of sensory information, therefore, it must somehow take the raw material provided by the senses in the form of material images and make them into suitable objects for understanding. This process is known as abstraction, from the Latin *abstrahere*, which is literally "to drag out." The intellect pulls out the universal, as it were, from the material singular in which it is embedded. This activity is performed by the active or agent intellect, which takes the "phantasms" derived from sense experience and turns them into "intelligible species." Those species are actualized in the possible or receptive intellect, whose function is to receive and then store the intelligible species provided by the active intellect. Scotus denies that the

active and passive intellect are really distinct. Rather, there is one intellect that has these two distinct functions or powers.

Phantasms do not, however, become irrelevant once the intelligible species has been abstracted. Scotus holds (just as Aquinas had held) that the human intellect never understands anything without turning towards phantasms (*Lectura* 2, d. 3, pars 2, q. 1, n. 255). That is, in order to deploy a concept that has already been acquired, one must make some use of sensory data -- although the phantasms employed in using a concept already acquired need not be anything like the phantasms from which that concept was abstracted in the first place. I acquired the intelligible species of dog from phantasms of dogs, but I can make use of that concept now not only by calling up an image of a dog but also by (say) imagining the sound of the Latin word for dog. Scotus's point is simply that there must be some sensory context for any act of intellectual cognition.

And even that point is not quite as general as my unqualified statement suggests. For one thing, Scotus believes that our intellect's need for phantasms is a temporary state. It is only in this present life that the intellect must turn to phantasms; in the next life we will be able to do without them. For another thing, Scotus may have thought that even in this life we enjoy a kind of intellectual cognition that bypasses phantasms. He called it "intuitive cognition."

4.2 Intuitive cognition

Scotus understands intuitive cognition by way of contrast with abstractive cognition. The latter, as we have seen, involves the universal; and a universal as such need not be exemplified. That is, my intelligible species of dog only tells me what it is to be a dog; it doesn't tell me whether any particular dog actually exists. Intuitive cognition, by contrast, "yields information about how things are right now" (Pasnau [forthcoming]). Sensory cognition, as Scotus explicitly acknowledges, counts as intuitive cognition on this account. It is, after all, quite uncontroversial that my seeing or hearing a dog gives me information about some particular dog as it exists when I see or hear it. Scotus's much bolder claim concerns *intellectual* intuitive cognition, by which the intellect cognizes a particular thing as existing at that very moment. Intellectual intuitive cognition does not require phantasms; the cognized object somehow just causes the intellectual act by which its existence is made present to the intellect. As Robert Pasnau rightly notes, intellectual intuitive cognition is in effect a "form of extra-sensory perception" (Pasnau [forthcoming]).

In some places Scotus seems to think of this sort of intuitive cognition as a mere theoretical possibility, but in others he argues vigorously for the reality of intellectual intuitive cognition. Indeed, in the latter sorts of passages it becomes clear that intuitive cognition is quite pervasive in human thought. (For three different takes on what to make of Scotus's apparently conflicting signals on this matter, see Day [1947], Pasnau [forthcoming], and Wolter [1990a].) He argues, for example, that since the intellect engages in reasoning that makes reference to the actual existence of particular sensible objects, it must know that they exist. Abstractive cognition, of course, cannot provide such knowledge. Moreover, without intuitive cognition I could never know about my own intellectual states. Abstractive cognition could provide me

with an abstract concept of *thinking about Scotus*, but I need intuitive cognition to know that I am in fact exemplifying that concept right this minute.

If these arguments represent Scotus's considered views on intuitive cognition, then Scotus is making a bold exception to the general rule that in this life the intellect acquires knowledge only by turning to phantasms. It would seem that he has little choice, given the importance he attaches to our intuitive self-knowledge (as I discuss in the next section). For our intellect is immaterial, as are its acts, and it is difficult to see how an immaterial act can be captured in a sensory phantasm. Even so, Scotus is enough of an Aristotelian about the functioning of our intellect on this side of heaven to insist that even though our brute *acquaintance* with those acts is independent of phantasms, the *descriptions* under which we know those acts must be capable of being captured in a phantasm. And our intuitive cognition of extra-mental singulars extends only to *material* singulars, i.e., those that are capable of being captured in a phantasm. Scotus consistently denies that we can have intuitive cognition of non-sensible objects (such as angels) or universals in this life.

4.3 The attack on skepticism and illuminationism

Scotus argues that the human intellect is capable of achieving certainty in its knowledge of the truth simply by the exercise of its own natural powers, with no special divine help. He therefore opposes both skepticism, which denies the possibility of certain knowledge, and illuminationism, which insists that we need special divine illumination in order to attain certainty. He works out his attack on both doctrines in the course of a reply to Henry of Ghent in *Ordinatio* 1, d. 3, pars 1, q. 4. (For the text and translation, see Wolter [1987], 96-132.)

According to Henry, truth involves a relation to an "exemplar." (We can think of this relation as akin to the relation of correspondence appealed to by certain theories of truth, and the exemplar itself as the mental item that is one of the relata of the correspondence-relation. The other relatum, of course, is "the way things really are.") Now there are two exemplars: the created exemplar, which is the species of the universal caused by the thing known, and the uncreated exemplar, which is an idea in the divine mind. Henry argues that the created exemplar cannot provide us with certain and infallible knowledge of a thing. For, first, the object from which the exemplar is abstracted is itself mutable and therefore cannot be the cause of something immutable. And how can there be certain knowledge apart from some immutable basis for that knowledge? Second, the soul itself is mutable and subject to error, and it can be preserved from error only by something less mutable than itself. But the created exemplar is even more mutable than the soul. Third, the created exemplar by itself does not allow us to distinguish between reality and dreaming, since the content of the exemplar is the same in either case. Henry therefore concludes that if we are to have certainty, we must look to the uncreated exemplar. And since we cannot look to the uncreated exemplar by our natural powers, certainty is impossible apart from some special divine illumination.

Scotus argues that if Henry is right about the limitations of our natural powers, even divine illumination is not enough to save us from pervasive uncertainty. To Henry's first argument he replies that there is no

certainty to be had by knowing a mutable object as immutable. To the second he replies that anything in the soul -- including the very act of understanding that Henry thinks is achieved through illumination -- is mutable. So by Henry's argument it would be impossible for anything whatever to preserve the soul from error. And to the third argument he replies that if the created exemplar is such as to preclude certainty, adding extra exemplars will not solve the problem: "When something incompatible with certainty concurs, certainty cannot be attained" (*Ordinatio* 1, d. 3, pars 1, q. 4, n. 221).

So Henry's arguments, far from showing that certainty is possible through divine illumination, actually lead to a pervasive skepticism. Scotus counters that we can show that skepticism is false. We can in fact attain certainty, and we can do so by the unaided exercise of our natural intellectual powers. There are four types of knowledge in which infallible certainty is possible. First, knowledge of first principles is certain because the intellect has only to form such judgments to see that they are true. (And since the validity of proper syllogistic inference can be known in just this way, it follows that anything that is seen to be properly derived from first principles by syllogistic inference is also known with certainty.) Second, we have certainty with respect to quite a lot of causal judgments derived from experience. Third, Scotus says that many of our own acts are as certain as first principles. It is no objection to point out that our acts are contingent, since some contingent propositions must be known immediately (that is, without needing to be derived from some other proposition). For otherwise, either some contingent proposition would follow from a necessary proposition (which is impossible), or there would be an infinite regress in contingent propositions (in which case no contingent proposition would ever be known). Fourth, certain propositions about present sense experience are also known with certainty if they are properly vetted by the intellect in light of the causal judgments derived from experience.

5. Ethics and Moral Psychology

5.1 The natural law

For Scotus the natural law in the strict sense contains only those moral propositions that are *per se notae ex terminis* along with whatever propositions can be derived from them deductively (*Ordinatio* 3, d. 37, q. un.). *Per se notae* means that they are self-evident; *ex terminis* adds that they are self-evident in virtue of being analytically true. Now one important fact about propositions that are self-evident and analytically true is that God himself can't make them false. They are necessary truths. So the natural law in the strict sense does not depend on God's will. This means that even if (as I believe) Scotus is some sort of divine-command theorist, he is not whole-hog in his divine command theory. Some moral truths are necessary truths, and even God can't change those. They would be true no matter what God willed.

Which ones are those? Scotus's basic answer is that they are the commandments of the first tablet of the Decalogue (Ten Commandments). The Decalogue has often been thought of as involving two tablets. The first covers our obligations to God and consists of the first three commandments: *You shall have no other gods before me*, *You shall not take the name of the Lord your God in vain*, and *Remember the Sabbath day to keep it holy*. (Note that many Protestants divide them up differently.) The second tablet spells out our obligations toward others: *Honor your father and mother*, *You shall not kill*, *You shall not commit*

adultery, *You shall not steal*, *You shall not bear false witness against your neighbor*, and two commandments against coveting. The commandments of the first tablet are part of the natural law in the strict sense because they have to do with God himself, and with the way in which God is to be treated. For Scotus says that the following proposition is *per se nota ex terminis*: "If God exists, then he is to be loved as God, and nothing else is to be worshiped as God, and no irreverence is to be done to him." Given the very definition of God, it follows that if there is such a being, he is to be loved and worshiped, and no irreverence should be shown to him. Because these commandments are self-evident and analytic, they are necessary truths. Not even God himself could make them false.

But even the first three commandments, once we start looking at them, are not obviously part of the natural law in the strict sense. In particular, the third commandment, the one about the Sabbath day, is a little tricky. Obviously, the proposition "God is to be worshiped on Saturday" is not self-evident or analytic. In fact, Scotus says it's not even true any more, since Christians are to worship on Sunday, not Saturday. So, Scotus asks, what about the proposition "God is to be worshiped at some time or other"? Even that is not self-evident or analytic. The best one can do is "God is not to be hated." Now that's self-evident and analytic, since by definition God is the being most worthy of love and there is nothing in him worthy of hate. But obviously that's far weaker than any positive commandment about whether and when we should worship God.

So by the time Scotus completes his analysis, we are left with nothing in the natural law in the strict sense except for negative propositions: God is not to be hated, no other gods are to be worshiped, no irreverence is to be done to him. Everything else in the Decalogue belongs to the natural law in a weaker or looser sense. These are propositions that are not *per se notae ex terminis* and do not follow from such propositions, but are "highly consonant" with such propositions. Now the important point for Scotus is this: since these propositions are contingent, they are completely up to God's discretion. Any contingent truth whatsoever depends on God's will.

According to Scotus, God of course is aware of all contingent propositions. Now God gets to assign the truth values to those propositions. For example, "Unicorns exist" is a contingent proposition. Therefore, it is up to God's will whether that proposition will be true or false. The same goes for contingent moral propositions. Take any such proposition and call it *L*, and call the opposite of *L*, not-*L*. Both *L* and not-*L* are contingent propositions. God can make either of them true, but he can't make both of them true, since they are contradictories. Suppose that God wills *L*. *L* is now part of the moral law. How do we explain why God willed *L* rather than not-*L*? Scotus says we can't. God's will with respect to contingent propositions is unqualifiedly free. So while there might be some reasons why God chose the laws he chose, there is no fully adequate reason, no total explanation. If there were a total explanation other than God's will itself, those propositions wouldn't be contingent at all. They would be necessary. So at bottom there is simply the sheer fact that God willed one law rather than another.

Scotus intends this claim to be exactly parallel to the way we think about contingent beings. Why are there elephants but no unicorns? As everyone would agree, it's because God willed for there to be elephants but no unicorns. And why did he will that? He just did. That's part of what we mean by saying that God was free in creating. There was nothing constraining him or forcing him to create one thing

rather than another. The same is true about the moral law. Why is there an obligation to honor one's parents but no such obligation toward cousins? Because God willed that there be an obligation to honor one's parents, and he did not will that there be any such obligation toward one's cousins. He could have willed both of these obligations, and he could have willed neither. What explains the way that he did in fact will? Nothing whatsoever except the sheer fact that he did will that way.

5.2 The will, freedom, and morality

Scotus quite self-consciously puts forward his understanding of freedom as an alternative to Aquinas's. According to Aquinas, freedom comes in simply because the will is intellectual appetite rather than mere sense appetite. Intellectual appetite is aimed at objects as presented by the intellect and sense appetite at objects as presented by the senses. Sense appetite is not free because the senses provide only particulars as objects of appetite. But intellectual appetite is free because the intellect deals with universals, not particulars. Since universals by definition include many particulars, intellectual appetite will have a variety of objects. Consider goodness as an example. The will is not aimed at this good thing or that good thing, but at goodness in general. Since that universal, goodness, contains many different particular things, intellectual appetite has many different options.

But Scotus insists that mere intellectual appetite is not enough to guarantee freedom in the sense needed for morality. The basic difference comes down to this. When Aquinas argues that intellectual appetite has different options, he seems to be thinking of this over a span of time. Right now the intellect presents *x* as good, so I will *x*; but later on the intellect presents *y* as good, so then I will *y*. But Scotus thinks of freedom as involving multiple options at the very moment of choice. It's not enough to say that now I will *x*, but later I can will *y*. We have to say that at the very moment at which I will *x*, I also am able to will *y*. Aquinas's arguments don't show that intellectual appetite is free in this stronger sense. So as far as Scotus is concerned, Aquinas hasn't made room for the right kind of freedom.

This is where Scotus brings in his well-known doctrine of the two affections of the will (see especially *Ordinatio* 2, d. 6, q. 2; 2, d. 39, q. 2; 3, d. 17, q. un.; and 3, d. 26, q. un.). The two affections are fundamental inclinations in the will: the *affectio commodi*, or affection for the advantageous, and the *affectio iustitiae*, or affection for justice. Scotus identifies the *affectio commodi* with intellectual appetite. Notice how important that is. For Aquinas intellectual appetite is the same thing as will, whereas for Scotus intellectual appetite is only part of what the will is. Intellectual appetite is just one of the two fundamental inclinations in the will. Why does Scotus make this crucial change? For the reason we've already discussed. He doesn't see how intellectual appetite could be genuinely free. Now he can't deny that the will involves intellectual appetite. Intellectual appetite is aimed at happiness, and surely happiness does have some role to play in our moral psychology. But the will has to include something more than intellectual appetite if it's going to be free. That something more is the *affectio iustitiae*. But one can't fully understand what the *affectio iustitiae* is until Aquinas and Scotus are compared on a further point.

For Aquinas the norms of morality are defined in terms of their relationship to human happiness. We have a natural inclination toward our good, which is happiness, and it is that good that determines the content

of morality. So like Aristotle, Aquinas holds a eudaimonistic theory of ethics: the point of the moral life is happiness. That's why Aquinas can understand the will as an intellectual appetite for happiness. All of our choosing is aimed at the human good (or at least, it's aimed at the human good as we conceive it). And choices are good -- and, indeed, fully intelligible -- only when they are aimed at the ultimate end, which is happiness. So Aquinas just defines the will as the capacity to choose in accordance with a conception of the human good -- in other words, as intellectual appetite.

When Scotus rejects the idea that will is merely intellectual appetite, he is saying that there is something fundamentally wrong with eudaimonistic ethics. Morality is not tied to human flourishing at all. For it is Scotus's fundamental conviction that morality is impossible without libertarian freedom, and since he sees no way for there to be libertarian freedom on Aquinas's eudaimonistic understanding of ethics, Aquinas's understanding must be rejected. And just as Aquinas's conception of the will was tailor-made to suit his eudaimonistic conception of morality, Scotus's conception of the will is tailor-made to suit his anti-eudaimonistic conception of morality. It's not merely that he thinks there can be no genuine freedom in mere intellectual appetite. It's also that he rejects the idea that moral norms are intimately bound up with human nature and human happiness. The fact that God creates human beings with a certain kind of nature does not require God to command or forbid the actions that he in fact commanded or forbade. The actions he commands are not necessary for our happiness, and the actions he forbids are not incompatible with our happiness. Now if the will were merely intellectual appetite -- that is, if it were aimed solely at happiness -- we would not be able to choose in accordance with the moral law, since the moral law itself is not determined by any considerations about human happiness. So Scotus relegates concerns about happiness to the *affectio commodi* and assigns whatever is properly moral to the other affection, the *affectio iustitiae*.

Bibliography

Primary texts in Latin

- *Cuestiones Cuodlibetales*. In *Obras del Doctor Sutil, Juan Duns Escoto*. Ed. Felix Alluntis. Madrid: Biblioteca de Autores Cristianos, 1963.
- *Opera Omnia*. ("The Wadding edition") Lyon, 1639; reprinted Hildesheim: Georg Olms Verlagsbuchhandlung, 1968. This is the best source for material not yet available in the critical editions. It does include some material now known to be inauthentic, and it prints as Book 1 of the *Reportatio* what is actually the *Additiones magnae* compiled and edited by Scotus's student and secretary, William of Alnwick. An Adobe Actobat (PDF) version of the entire Wadding edition is available through the French National Library: see Other Internet Resources below.
- *Opera Omnia*. ("The Vatican edition") Civitas Vaticana: Typis Polyglottis Vaticanis, 1950-. So far includes the *Ordinatio* up through Book 2, distinction 3 (vols. I-VI) and Books 1 and 2 of the *Lectura* (vols. XVI-XIX).
- *Opera Theologica*. St. Bonaventure, NY: The Franciscan Institute, 1997-. So far includes the question-commentaries on Porphyry's *Isagoge* and Aristotle's *Categories* (vol. I) and the *Quaestiones super libros Metaphysicorum Aristotelis* (vols. III-IV).

Primary texts in English translation

- Spade, Paul Vincent. (1994). *Five Texts on the Mediaeval Problem of Universals*. Indianapolis: Hackett Publishing Company, 1994.
- Wolter, Allan B., OFM, and Felix Alluntis. (1975). *John Duns Scotus, God and Creatures. The Quodlibetal Questions*. Washington, D.C.: The Catholic University of America Press, 1975.
- Wolter, Allan B., OFM. (1986). *Duns Scotus on the Will and Morality*. Washington, DC: The Catholic University of America Press, 1986.
- Wolter, Allan B., OFM. (1987). *Duns Scotus: Philosophical Writings*. Indianapolis: Hackett Publishing Company, 1987.

Secondary literature

- Cross, Richard. (1999). *Duns Scotus*. Oxford: Oxford University Press, 1999.
- Cross, Richard. (forthcoming). "Philosophy of Mind." In Williams (forthcoming).
- Day, Sebastian. (1947). *Intuitive Cognition: A Key to the Significance of the Later Scholastics*. St Bonaventure, NY: The Franciscan Institute, 1947.
- Frank, William A. and Allan B. Wolter, OFM. (1995). *Duns Scotus: Metaphysician*. Lafayette, IN: Purdue University Press, 1995.
- King, Peter (forthcoming). "Scotus on Metaphysics." In Williams (forthcoming).
- Pasnau, Robert (forthcoming). "Cognition." In Williams (forthcoming).
- Williams, Thomas. (1995). "How Scotus Separates Morality from Happiness," *American Catholic Philosophical Quarterly* 69 (1995): 425-445. [Preprint available online.]
- Williams, Thomas. (1998). "The Unmitigated Scotus," *Archiv für Geschichte der Philosophie* 80 (1998): 162-181. [Preprint available online.]
- Williams, Thomas. (2000). "A Most Methodical Lover: On Scotus's Arbitrary Creator," *Journal of the History of Philosophy* 38 (2000): 169-202. [Preprint available online.]
- Williams, Thomas (forthcoming). *The Cambridge Companion to Duns Scotus*. New York: Cambridge University Press, forthcoming.
- Wolter, Allan B., OFM. (1990a). "Duns Scotus on Intuition, Memory, and Our Knowledge of Individuals." In Wolter (1990b).
- Wolter, Allan B., OFM. (1990b). *The Philosophical Theology of John Duns Scotus*. Ed. Marilyn McCord Adams. Ithaca, NY: Cornell University Press, 1990.

Other Internet Resources

- Bibliotheque Nationale de France. The complete Wadding edition in Adobe Acrobat (PDF) format. For access, type "Duns" in the "auteur" field and then click on "Rechercher."

Related Entries

analogy: medieval theories of | Anselm, Saint [Anselm of Bec, Anselm of Canterbury] | Aquinas, Saint Thomas | conscience: medieval theories of | divine illumination | free will | future contingents: medieval theories of | haecceity: medieval theories of | intentionality: medieval theories of | medieval philosophy | modality: medieval theories of | Ockham [Occam], William | practical reason: medieval theories of | relations: medieval theories of | universals: the medieval problem of

Thomas Williams
thomas-williams@uiowa.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 30, 2001
Content last modified: May 30, 2001

Medieval Theories of Conscience

Through conscience and its related notion, synderesis, human beings discern what is right and wrong. While there are many medieval views about the nature of conscience, most views regard human beings as capable of knowing in general what ought to be done and applying this knowledge through conscience to particular decisions about action. The ability to act on the determinations of conscience is, moreover, tied to the development of the moral virtues, which in turn refines the functions of conscience.

Section 1: Background

Late medieval discussions of conscience derive from Peter Lombard's presentation of the concepts of conscience and synderesis in his *Sentences*. Lombard cites a passage from St. Jerome, interpreting Ezekiel's vision of four living creatures coming out of a cloud. Each creature was shaped like a man, but each had four faces: the front face was human; the right was that of a lion; the left was that of an ox; and the back was that of an eagle (Ezekiel 1.4-14). Jerome identifies the human face as representing the rational part of man, the lion as the emotional, the ox as the appetitive, and the eagle as that "which the Greeks call synteresis: that spark of conscience which was not even extinguished in the breast of Cain after he was turned out of paradise, and by which we discern that we sin, when we are overcome by pleasures or frenzy and meanwhile are misled by an imitation of reason." Jerome's comment that synteresis (alternatively, synderesis) is never extinguished in human beings and his remarks elsewhere to the effect that wicked people do cease to have any conscience led Lombard and subsequent thinkers to distinguish synderesis from conscience. While it is unclear that Jerome meant to distinguish the two, the distinction plays a major role in late medieval discussions of conscience.

In these discussions, constant reference was made to certain works by Plato and Aristotle. Neither Plato nor Aristotle explicitly mention conscience, however. It is their discussions of the virtues, practical wisdom, and weakness of will that form the critical backdrop to medieval discussions of conscience. These discussions were heavily influenced by Augustine's modification of these classical authors. For example, Augustine championed Plato's notion of the unity of the virtues, but he argued that love of God provided the unity to them. Moreover, he claimed that what pagan authors regarded as virtues were in fact vices unless they were developed for the love of God.

Two distinct views about the relationship between conscience and synderesis emerged in the late Middle Ages. The first view, a voluntaristic one, can be identified with Franciscan thinkers like Bonaventure. The second, most clearly expounded by Aquinas, is an intellectualistic view. Both seem to derive from Philip the Chancellor's treatise on conscience. In his treatise, Philip chiefly discusses synderesis, and at times he describes it as an unerring intellectual dispositional potentiality that provides general truths to conscience for specific application. At other times, he describes synderesis as the desire for the good, and it is equated with emotional reactions when one follows evil instead of good. This latter description fits well with Bonaventure's views on synderesis and conscience.

Section 2: Bonaventure

Bonaventure discusses both in his *Commentary on the Sentences*, Book II, distinction 39. He places conscience squarely within the rational faculty, specifying that it is part of practical reason since it is connected to the performance of actions. It is thus also connected to the will as well as the emotions. On the other hand, he places synderesis in the affective part of human beings, for he regards synderesis as that which stimulates us to the good.

Conscience is divided into two general parts by Bonaventure. The first part seems to be a power for discovering the truth of very general practical principles like "obey God," "honor your parents," and "do not harm your neighbors." This part of conscience is innate and unerring; it cannot be lost to any person, no matter how morally corrupt that person may become. The second part of conscience involves the application of the very general principles to situations that may be either general or particular. This second part is also innate, but it can be mistaken since the very general principles of the first part may be misapplied through ignorance or faulty reasoning. The misapplication explains, to a certain extent, how conscience, oriented to good, can be involved in the performance of evil actions. The distinction between the two parts of conscience also opens up the possibility for developing, through experience, practical principles of behavior not directly entailed by the content of the synderesis. By generalizing on activities performed in accordance with the principles of the synderesis, one can formulate new general principles not contained in the synderesis that can guide behavior in a number of contexts. Conscience thus appears to be a dynamic faculty for Bonaventure.

Bonaventure calls synderesis the "spark of conscience," and he sees it as resting in the affective part of human beings. It is the spark because, as the general drive to do good, synderesis provides the movement that conscience needs to operate. In general, Bonaventure regards conscience and synderesis as

interpenetrating one another. The formation of ethical rules by conscience is seen by him as an implementation of a human being's desire for good (the synderesis). He also sees the following of these principles as another aspect of the desire for good. Because we naturally have a desire for the good, we also desire the means to that goal. The principles of conscience are such means, and so we are naturally disposed to carry out the principles of conscience. Similarly, the emotional reaction to doing evil (guilt or remorse) is a reaction to the frustration of the desire for good caused when one fails to adhere to what the conscience has determined will lead to good. Bonaventure, while placing synderesis and conscience in different parts of a human being, does not isolate them. On the contrary, he views conscience as driven by synderesis and at the same time directing synderesis.

Section 3: Aquinas

Thomas Aquinas, the principal advocate of the intellectualistic view of the relationship of conscience and synderesis, explicitly defines ‘conscience’ as the "application of knowledge to activity" (*Summa Theologiae*, I-II, I) The knowledge he has in mind here comes from the synderesis, which he regards as the natural disposition of the human mind by which we apprehend without inquiry the basic principles of behavior. For Aquinas, then, the conscience applies the first principles of the synderesis to particular situations. The principles of synderesis are rather general in form. Examples are "Do good and avoid evil" and "Obey God." To be helpful in human activity, conscience requires principles that contain much more content. One can call these "secondary principles" and Aquinas discusses them in several places and suggests that they are derived from experience and instruction through the virtue of prudence. Thus, the function of conscience for Aquinas is to apply the general principles of synderesis and the more content-laden secondary principles developed from prudence to particular circumstances. Prudence is involved in the application to particular circumstances, according to Aquinas, because it is connected to the correct perception of individual circumstances. And this aspect of prudence connects both conscience and prudence to the problem of weakness of will.

In Aquinas's presentation of Aristotle's discussion of weakness of will in his Commentary on the *Nicomachean Ethics*, the fourth position offered in Book 7, Chapter 3 of the *Nicomachean Ethics* is emphasized. According to this position, the incontinent man knows the appropriate general principles of behavior concerning what should be done, e.g., one should not fornicate. If the incontinent man sees a particular action as falling under this general principle, e.g., a man sees that having intercourse with an unmarried woman is a case of fornication, he will not perform the action. However, the incontinent man also holds the general rule that pleasures should be enjoyed. If the incontinent man, driven by his particular desire for a particular unmarried woman, sees the proposed sexual liaison as a case of pleasure, he subsumes it under the general rule about pursuing pleasure and pursues the relationship. The desire he has, as it were, blinds him to the general principle about fornication he still possesses, but only habitually. The actual knowledge he possesses is that the proposed liaison is a case of pleasure to be pursued. He thus has (habitually) the knowledge that he should avoid fornication, but he fornicates nonetheless because he actually sees the fornication as an act of pleasure to be pursued. As a general comment on Aristotle's analysis, Aquinas remarks: "It is not the knowledge of the universal but only the evaluation of the sensible, which is not so excellent, that is dragged about by passion." (*Commentary on*

the Nicomachean Ethics, Book 7, lecture 3, paragraph 1352) The point Aquinas is making is that the incontinent man possesses the knowledge of what he should do, but he is driven by the passion he has for a particular; this passion leads him to act contrary to what the knows (habitually) should not be done. The incontinent man so fails because he has failed to cultivate the appropriate virtues that would enable him to size up the situation correctly (*synesis*) and deliberate well about it (*eubulia*). Aquinas's linking of conscience with prudence and the virtues in general through his concern with weakness of will is innovative and undoubtedly connected with his interest in the *Nicomachean Ethics*. Duns Scotus and William of Ockham follow his lead in linking conscience with issues surrounding development of the virtues.

Section 4: Scotus and Ockham

Scotus offers very little explicit discussion of either conscience or synderesis. Yet, from his discussion of issues chiefly concerned with development of the virtues, it is apparent that his view of conscience and synderesis seems to draw from both Bonaventure and Aquinas. Following Aquinas, Scotus thinks that both synderesis and conscience are to be placed in the intellectual order. In agreement with Bonaventure, Scotus gives conscience much more of a dynamic role in the human personality than a mechanical application of general principles. Scotus's close linking of conscience and the development of the virtues allows him to combine the two sources.

According to the virtue tradition, in order to perform a virtuous action, one must have the right dictates associated with the relevant virtue. Yet, one must perform appropriate virtuous actions to develop the habit of the virtue and to know the relevant right dictates. The obvious circularity seems vicious enough to undermine any attempt to cultivate virtues. Scotus regards conscience as offering a way into the circle. Whenever a person formulates what is to be done in some circumstance, this is an exercise of conscience, which has determined proper action from the principles of synderesis. On the basis of the dictates of conscience, a person can perform an action that will provide the basis for the development of the relevant virtues. For the performance of these acts from conscience leads to the type of habit that Scotus thinks of as a virtue. Ideally, the moral virtues are unified since a perfect, virtuous person should possess all virtues. In fact, Scotus's perfect, virtuous person seems very similar to Aristotle's man of practical wisdom. This is the person who has, through long experience, developed the moral virtues and is able to deliberate so well about all moral situations that in Aristotle's view to be moral is to do what a man of practical wisdom would do. Scotus's perfect, virtuous person, like the man of practical wisdom, is skilled at determining what should be done in given circumstances; he takes delight in acting in accord with his virtues, and he possesses all of the moral virtues by developing them through experience.

Ockham's discussion of conscience, prudence, and the virtues indicate that he follows Scotus's turn towards discussing conscience in relation to the virtues. He agrees with Scotus that conscience can provide the entry into the seeming circularity of performing virtuous actions in order to develop intentions that seem to be required for performing the virtuous actions in the first place. Nevertheless, he criticizes Scotus for failing to make a number of necessary distinctions about degrees of virtues and the relationship of conscience to prudence. He never mentions synderesis in his writings and emphasizes the

fact that only internal acts have moral worth. According to him, external acts are morally significant only by extrinsic denomination from internal acts. Particularly in these last two claims, Ockham exercised considerable influence on Reformation thinkers like Luther and Calvin in their discussions of conscience.

Bibliography

- Baylor, Michael G. *Action and Person: Conscience in Late Scholasticism and the Young Luther*. Studies in Medieval and Reformation Thought, Volume XX. E. J. Brill, Leiden, 1977.
- D'Arcy, Eric. *Conscience and Its Right to Freedom*. Sheed and Ward, New York and London, 1961.
- Dolan, Joseph V. "Conscience in the Catholic Theological Tradition." In Bier, William C., editor. *Conscience: Its Freedom and Limitations* (Fordham University Press, N. Y., 1971)
- Holopainen, Taina M. *William Ockham's Theory of the Foundations of Ethics*. Luther-Agricola-Society, Helsinki, 1991.
- Kent, Bonnie. "Transitory Vice: Thomas Aquinas on Incontinence." *The Journal of the History of Philosophy*, Volume 27, 19 .
- Kent, Bonnie. *Virtues of the Will. The Transformation of Ethics in the Late Thirteenth Century*. Catholic University of America Press, Washington, D. C., 1995.
- Lottin, O. *Psychologie et morale aux XIIe et XIIIe siecles*. Volumes I (second edition) and II. J. Duculot, Gembloux, 1957; 1948.
- Nelson, Dan. *The Priority of Prudence*. Penn State Press, College Park, 1988.
- Potts Timothy C. *Conscience in Medieval Philosophy*. Cambridge University Press, Cambridge; 1980.
- Potts, Timothy C. "Conscience." In Kretzmann, N.; Kenny, A.; Pinborg, J., editors. *The Cambridge History of Later Medieval Philosophy*. Cambridge University Press, Cambridge, 1982.
- Saarinen, Risto. *Weakness of the Will in Medieval Thought From Augustine to Buridan*. E. J. Brill, Leiden, New York, 1994.
- Zachman, Randall C. *The Assurance of Faith. Conscience in the Theology of Martin Luther and John Calvin*. Augusburg Fortress Press, Minneapolis, 1993.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aquinas, Saint Thomas | Aristotle: ethics | Augustine, Saint | Bonaventure, Saint | Duns Scotus, John | Ockham [Occam], William | Plato | practical reason: medieval theories of | virtue: medieval theories of

Doug Langston
New College of the University of South Florida
langston@sar.usf.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 23, 1998
Content last modified: February 25, 2000

Aristotle's Ethics

Aristotle conceives of ethical theory as a field distinct from the theoretical sciences. Its methodology must match its subject matter - good action - and must respect the fact that in this field many generalizations hold only for the most part. We study ethics in order to improve our lives, and therefore its principal concern is the nature of human well-being. Aristotle follows Socrates and Plato in taking the virtues to be central to a well-lived life. Like Plato, he regards the ethical virtues (justice, courage, temperance and so on) as complex rational, emotional and social skills. But he rejects Plato's idea that a training in the sciences and metaphysics are a necessary prerequisite for a full understanding of our good. What we need, in order to live well, is a proper appreciation of the way in which such goods as friendship, pleasure, virtue, honor and wealth fit together as a whole. In order to apply that general understanding to particular cases, we must acquire, through proper upbringing and habits, the ability to see, on each occasion, which course of action is best supported by reasons. Therefore practical wisdom, as he conceives it, cannot be acquired solely by learning general rules. We also must also acquire, through practice, those deliberative, emotional, and social skills that enable us to put our general understanding of well-being into practice in ways that are suitable to each occasion.

---

1. Preliminaries

Aristotle wrote two ethical treatises: the *Nicomachean Ethics* and the *Eudemian Ethics*. He does not himself use either of these titles, although in the *Politics* (1295a36) he refers back to one of them -- probably the *Eudemian Ethics* -- as "*ta êthika*" -- his writings about character. The words "*Eudemian*" and "*Nicomachean*" were added later, perhaps because the former was edited by his friend, Eudemus, and the latter by his son, Nicomachus. In any case, these two works cover more or less the same ground: they begin with a discussion of *eudaimonia* ("happiness," "flourishing"), and turn to an examination of the nature of *aretê* ("virtue," "excellence") and the character traits that human beings need in order to live life at its best. Both treatises examine the conditions in which praise or blame are appropriate, and the nature of pleasure and friendship; near the end of each work, we find a brief discussion of the proper relationship between human beings and the divine.

Though the general point of view expressed in each work is the same, there are many subtle differences in organization and content as well. Clearly, one is a re-working of the other, and although no single piece of evidence shows conclusively what their order is, it is widely assumed that the *Nicomachean Ethics* is a later and improved version of the *Eudemian Ethics*. (Not all of the *Eudemian Ethics* was revised: its Books IV, V, and VI re-appear as V, VI, VII of the *Nicomachean Ethics*.) Perhaps the most telling indication of this ordering is that in several instances the *Nicomachean Ethics* develops a theme about which its *Eudemian* cousin is silent. Only the *Nicomachean Ethics* discusses the close relationship between ethical inquiry and politics; only the *Nicomachean Ethics* examines Solon's dictum that no man should be counted happy until he is dead; and only the *Nicomachean Ethics* gives a series of arguments for the superiority of the philosophical life to the political life. The remainder of this article will therefore focus on this work. [**Note**: Page and line numbers shall henceforth refer to this treatise.]

Although Aristotle is deeply indebted to Plato's moral philosophy, particularly Plato's central insight that moral thinking must be integrated with our emotions and appetites, and that the preparation for such unity of character should begin with childhood education, the systematic character of Aristotle's discussion of these themes was a remarkable innovation. No one had written ethical treatises before Aristotle. Plato's *Republic*, for example, does not treat ethics as a distinct subject matter; nor does it offer a systematic examination of the nature of happiness, virtue, voluntariness, pleasure, or friendship. To be sure, we can find in Plato's works important discussions of these phenomena, but they are not brought together and unified as they are in Aristotle's ethical writings.

2. The Human Good and the Function Argument

The principal idea with which Aristotle begins is that there are differences of opinion about what is best for human beings, and that to profit from ethical inquiry we must resolve this disagreement. He insists that ethics is not a theoretical discipline: we are asking what the good for human beings is not simply because we want to have knowledge, but because we will be better able to achieve our good if we develop a fuller understanding of what it is to flourish. In raising this question -- what is the good? -- Aristotle is not looking for a list of items that are good. He assumes that such a list can be compiled rather easily; most would agree, for example, that it is good to have friends, to experience pleasure, to be healthy, to be honored, and to have such virtues as courage at least to some degree. The difficult and controversial question arises when we ask whether certain of these goods are more desirable than others. Aristotle's search for *the* good is a search for the *highest* good, and he assumes that the highest good, whatever it turns out to be, has three characteristics: it is desirable for itself, it is not desirable for the sake of some other good, and all other goods are desirable for its sake.

Aristotle thinks everyone will agree that the terms "*eudaimonia*" ("happiness") and "*eu zên*" ("living well") designate such an end. The Greek term "*eudaimon*" is composed of two parts: "*eu*" means "well" and "*daimon*" means "divinity" or "spirit." To be *eudaimon* is therefore to be living in a way that is well-favored by a god. But Aristotle never calls attention to this etymology, and it seems to have little influence on his thinking. He regards "*eudaimon*" as a mere substitute for *eu zên* ("living well"). These terms play an evaluative role, and are not simply descriptions of someone's state of mind.

No one tries to live well for the sake of some further goal; rather, being *eudaimon is* the highest end, and all subordinate goals -- health, wealth, and other such resources -- are sought because they promote well-being, not because they are what well-being consists in. But unless we can determine which good or goods happiness consists in, it is of little use to acknowledge that it is the highest end. To resolve this issue, Aristotle asks what the *ergon* ("function," "task," "work") of a human being is, and argues that it consists in activity of the rational part of the soul in accordance with virtue (1097b22-1098a20). One important component of this argument is expressed in terms of distinctions he makes in his psychological and biological works. The soul is analyzed into a connected series of capacities: the nutritive soul is responsible for growth and reproduction, the locomotive soul for motion, the perceptive soul for perception, and so on. The biological fact Aristotle makes use of is that human beings are the only species that has not only these lower capacities but a rational soul as well. The good of a human being must have something to do with being human; and what sets humanity off from other species, giving us the potential to live a better life, is our capacity to guide ourselves by using reason. If we use reason well, we live well as human beings; or, to be more precise, using reason well over the course of a full life is what happiness consists in. Doing anything well requires virtue or excellence, and therefore living well consists in activities caused by the rational soul in accordance with virtue or excellence.

Aristotle's conclusion about the nature of happiness is in a sense uniquely his own. No other writer or thinker had said precisely what he says about what it is to live well. But at the same time his view is not too distant from a common idea. As he himself points out, one traditional conception of happiness identifies it with virtue (1098b30-1). Aristotle's theory should be construed as a refinement of this position. He says, not that happiness is virtue, but that it is virtuous *activity*. Living well consists in doing something, not just being in a certain state or condition. It consists in those lifelong activities that

actualize the virtues of the rational part of the soul.

At the same time, Aristotle makes it clear that in order to be happy one must possess others goods as well -- such goods as friends, wealth, and power. And one's happiness is endangered if one is severely lacking in certain advantages -- if, for example, one is extremely ugly, or has lost children or good friends through death (1098a31-b6). But why so? If one's ultimate end should simply be virtuous activity, then why should it make any difference to one's happiness whether one has or lacks these other types of good? Aristotle's reply is that one's virtuous activity will be to some extent diminished or defective, if one lacks an adequate supply of other goods (1153b17-19). Someone who is friendless, childless, powerless, weak, and ugly will simply not be able to find many opportunities for virtuous activity over a long period of time, and what little he can accomplish will not be of great merit. To some extent, then, living well requires good fortune; happenstance can rob even the most excellent human beings of happiness. Nonetheless, Aristotle insists, the highest good, virtuous activity, is not something that comes to us by chance. Although we must be fortunate enough to have parents and fellow citizens who help us become virtuous, we ourselves share much of the responsibility for acquiring and exercising the virtues.

3. Methodology

A. Traditional Virtues and the Skeptic.

A common complaint about Aristotle's attempt to defend his conception of happiness is that his argument is too general to show that it is in one's interest to possess any of the particular virtues as they are traditionally conceived. Suppose we grant, at least for the sake of argument, that doing anything well, including living well, consists in exercising certain skills; and let us call these skills, whatever they turn out to be, virtues. Even so, that point does not by itself allow us to infer that such qualities as temperance, justice, courage, as they are normally understood, are virtues. They should be counted as virtues only if it can be shown that actualizing precisely these skills is what happiness consists in. What Aristotle owes us, then, is an account of these traditional qualities that explains why they must play a central role in any well-lived life.

But perhaps Aristotle disagrees, and refuses to accept this argumentative burden. In one of several important methodological remarks he makes near the beginning of the *Nicomachean Ethics*, he says that in order to profit from the sort of study he is undertaking, one must already have been brought up in good habits (1095b4-6). The audience he is addressing, in other words, consists of people who already just, courageous, and generous; or, at any rate, they are well on their way to possessing these virtues. Why such a restricted audience? Why does he not address those who have serious doubts about the value of these traditional qualities, and who therefore have not yet decided to cultivate and embrace them?

Addressing the moral skeptic, after all, is the project Plato undertook in the *Republic*: in Book I he rehearses an argument to show that justice is not really a virtue, and the remainder of this work is an attempt to rebut this thesis. Aristotle's project seems, at least on the surface, to be quite different. He does not appear to be addressing someone who has genuine doubts about the value of justice or kindred

qualities. Perhaps, then, he realizes how little can be accomplished, in the study of ethics, to provide it with a rational foundation. Perhaps he thinks that no reason can be given for being just, generous, and courageous. These are qualities one learns to love when one is a child, and having been properly habituated, one no longer looks for or needs a reason to exercise them. One can show, as a general point, that happiness consists in exercising some skills or other, but that the moral skills of a virtuous person are what one needs is not a proposition that can be established on the basis of argument.

This is not the only way of reading the *Ethics*, however. For surely we cannot expect Aristotle to show what it is about the traditional virtues that makes them so worthwhile until he has fully discussed the nature of those virtues. He himself warns us that his initial statement of what happiness is should be treated as a rough outline whose details are to be filled in later (1098a20-22). His intention in Book I of the *Ethics* is to indicate in a general way why the virtues are important; why particular virtues -- courage, justice, and the like -- are components of happiness is something we should be able to better understand only at a later point.

In any case, Aristotle's assertion that his audience must already have begun to cultivate the virtues need not be taken to mean that no reasons can be found for being courageous, just, and generous. His point, rather, may be that in ethics, as in any other study, we cannot make progress towards understanding why things are as they are unless we begin with certain assumptions about what is the case. Neither theoretical nor practical inquiry starts from scratch. Someone who has made no observations of astronomical or biological phenomena is not yet equipped with sufficient data to develop an understanding of these sciences. The parallel point in ethics is that to make progress in this sphere we must already have come to enjoy doing what is just, courageous, generous and the like. We must experience these activities not as burdensome constraints, but as noble, worthwhile, and enjoyable in themselves. Then, when we engage in ethical inquiry, we can ask what it is about these activities that makes them worthwhile. We can also compare these goods with other things that are desirable in themselves -- pleasure, friendship, honor, and so on -- and ask whether any of them is more desirable than the others. We approach ethical theory with a disorganized bundle of likes and dislikes based on habit and experience; such disorder is an inevitable feature of childhood. But what is not inevitable is that our early experience will be rich enough to provide an adequate basis for worthwhile ethical reflection; that is why we need to have been brought up well. Yet such an upbringing can take us only so far. We seek a deeper understanding of the objects of our childhood enthusiasms, and we must systematize our goals so that as adults we have a coherent plan of life. We need to engage in ethical theory, and to reason well in this field, if we are to move beyond the low-grade form of virtue we acquired as children.

B. Differences and Affinities with Plato

Read in this way, Aristotle is engaged in a project similar in some respects to the one Plato carried out in the *Republic*. One of Plato's central points is that it is a great advantage to establish a hierarchical ordering of the elements in one's soul; and he shows how the traditional virtues can be interpreted to foster or express the proper relation between reason and less rational elements of the psyche. Aristotle's approach is similar: his "function argument" shows in a general way that our good lies in the dominance of reason, and the detailed studies of the particular virtues reveal how each of them involves the right kind of

ordering of the soul. Aristotle's goal is to arrive at conclusions similar to Plato's, but without relying on the Platonic metaphysics that plays a central role in the argument of the *Republic*. He rejects the existence of Plato's forms in general and the form of the good in particular; and he rejects the idea that in order to become fully virtuous one must study mathematics and the sciences, and see all branches of knowledge as a unified whole. Even though Aristotle's ethical theory sometimes relies on philosophical distinctions that are more fully developed in his other works, he never proposes that students of ethics need to engage in a specialized study of the natural world, or mathematics, or eternal and changing objects. His project is to make ethics an autonomous field, and to show why a full understanding of what is good does not require expertise in any other field.

There is another contrast with Plato that should be emphasized: In Book II of the *Republic*, we are told that the best type of good is one that is desirable both in itself and for the sake of its results (357d-358a). Plato argues that justice should be placed in this category, but since it is generally agreed that it is desirable for its consequences, he devotes most of his time to establishing his more controversial point -- that justice is to be sought for its own sake. By contrast, Aristotle assumes that if *A* is desirable for the sake of *B*, then *B* is better than *A* (1094a14-16); therefore, the highest kind of good must be one that is not desirable for the sake of anything else. To show that *A* deserves to be our ultimate end, one must show that all other goods are best thought of as instruments that promote *A* in some way or other. Accordingly, it would not serve Aristotle's purpose to consider virtuous activity in isolation from all other goods. He needs to discuss honor, wealth, pleasure, and friendship in order to show how these goods, properly understood, can be seen as resources that serve the higher goal of virtuous activity. He vindicates the centrality of virtue in a well-lived life by showing that in the normal course of things a virtuous person will not live a life devoid of friends, honor, wealth, pleasure, and the like. Virtuous activity makes a life happy not by guaranteeing happiness in all circumstances, but by serving as the goal for the sake of which lesser goods are to be pursued. Aristotle's methodology in ethics therefore pays more attention than does Plato's to the connections that normally obtain between virtue and other goods. That is why he stresses that in this sort of study one must be satisfied with conclusions that hold only for the most part (1094b11-22). Poverty, isolation, and dishonor are normally, impediments to the exercise of virtue and therefore to happiness, although there may be special circumstances in which they are not. The possibility of exceptions does not undermine the point that, as a rule, to live well is to have sufficient resources for the pursuit of virtue over the course of a lifetime.

4. Virtues and Deficiencies, Continence and Incontinence

Aristotle distinguishes two kinds of virtue (1103a1-10): those that pertain to the part of the soul that engages in reasoning (virtues of mind or intellect), and those that pertain to the part of the soul that cannot itself reason but is nonetheless capable of following reason (ethical virtues, virtues of character). Intellectual virtues are in turn divided into two sorts: those that pertain to theoretical reasoning, and those that pertain to practical thinking 1139a3-8). He organizes his material by first studying ethical virtue in general, then moving to a discussion of particular ethical virtues (temperance, courage, and so on), and finally completing his survey by considering the intellectual virtues (practical wisdom, theoretical

wisdom, etc.).

All free males are born with the potential to become ethically virtuous and practically wise, but to achieve these goals they must go through two stages: during their childhood, they must develop the proper habits; and then, when their reason is fully developed, they must acquire practical wisdom (*phronêsis*). This does not mean that first we fully acquire the ethical virtues, and then, at a later stage, add on practical wisdom. Ethical virtue is fully developed only when it is combined with practical wisdom (1144b14-17). A low-grade form of ethical virtue emerges in us during childhood as we are repeatedly placed in situations that call for appropriate actions and emotions; but as we rely less on others and become capable of doing more of our own thinking, we learn to develop a larger picture of human life, our deliberative skills improve, and our emotional responses are perfected. Like anyone who has developed a skill in performing a complex and difficult activity, the virtuous person takes pleasure in exercising his intellectual skills. Furthermore, when he has decided what to do, he does not have to contend with internal pressures to act otherwise. He does not long to do something that he regards as shameful; and he is not greatly distressed at having to give up a pleasure that he realizes he should forego.

Aristotle places those who suffer from such internal disorders into one of three categories: (A) Some agents, having reached a decision about what to do on a particular occasion, experience some counter-pressure brought on by an appetite for pleasure, or anger, or some other emotion; and this countervailing influence is not completely under the control of reason. (1) Within this category, some are typically better able to resist these counter-rational pressures than is the average person. Such people are not virtuous, although they generally do what a virtuous person does. Aristotle calls them "continent" (*enkratês*). But (2) others are less successful than the average person in resisting these counter-pressures. They are "incontinent" (*akratês*). (The explanation of *akrasia* is a topic to which we will return in section 8.) In addition, (B) there is a type of agent who refuses even to try to do what an ethically virtuous agent would do, because he has become convinced that justice, temperance, generosity and the like are of little or no value. Such people Aristotle calls evil (*kakos*, *phaulos*). He assumes that evil people are driven by desires for domination and luxury, and although they are single-minded in their pursuit of these goals, he portrays them as deeply divided, because their *pleonexia* -- their desire for more and more -- leaves them dissatisfied and full of self-hatred.

It should be noticed that all three of these deficiencies -- continence, incontinence, vice -- involve some lack of internal harmony. (Here Aristotle's debt to Plato is particularly evident, for one of the central ideas of the *Republic* is that the life of a good person is harmonious, and all other lives deviate to some degree from this ideal.) The evil person may wholeheartedly endorse some evil plan of action at a particular moment, but over the course of time, Aristotle supposes, he will regret his decision, because whatever he does will prove inadequate for the achievement of his goals (1166b5-29). Aristotle assumes that that when someone systematically makes bad decisions about how to live his life, his failures are caused by psychological forces that are less than fully rational. His desires for pleasure, power or some other external goal have become so strong that they make him care too little or not at all about acting ethically. To keep such destructive inner forces at bay, we need to develop the proper habits and emotional responses when we are children, and to reflect intelligently on our aims when we are adults. But some vulnerability to these disruptive forces is present even in more-or-less virtuous people; that is why even a

good political community needs laws and the threat of punishment. Clear thinking about the best goals of human life and the proper way to put them into practice is a rare achievement, because the human psyche is not a hospitable environment for the development of these insights.

5. The Doctrine of the Mean

A. Ethical Virtue as Disposition

Aristotle describes ethical virtue as a "*hexis*" ("state" "condition" "disposition") -- a tendency or disposition, induced by our habits, to have appropriate feelings (1105b25-6). Defective states of character are *hexeis* (plural of *hexis*) as well, but they are tendencies to have inappropriate feelings. The significance of Aristotle's characterization of these states as *hexeis* is his decisive rejection of the thesis, found throughout Plato's early dialogues, that virtue is nothing but a kind of knowledge and vice nothing but a lack of knowledge. Although Aristotle frequently draws analogies between the crafts and the virtues (and similarly between physical health and *eudaimonia*), he insists that the virtues differ from the crafts and all branches of knowledge in that the former involve appropriate emotional responses and are not purely intellectual conditions.

Furthermore, every ethical virtue is a condition intermediate between two other states, one involving excess, and the other deficiency (1106a26-b28). In this respect, Aristotle says, the virtues are no different from technical skills: every skilled worker knows how to avoid excess and deficiency, and is in a condition intermediate between two extremes. The courageous person, for example, judges that some dangers are worth facing and others not, and experiences fear to a degree that is appropriate to his circumstances. He lies between the coward, who flees every danger and experiences excessive fear, and the rash person, who judges every danger worth facing and experiences little or no fear. Aristotle holds that this same topography applies to every ethical virtue: all are located on a map that places the virtues between states of excess and deficiency. He is careful to add, however, that the mean is to be determined in a way that takes into account the particular circumstances of the individual (1106a36-b7). The arithmetic mean between 10 and 2 is 6, and this is so invariably, whatever is being counted. But the intermediate point that is chosen by an expert in any of the crafts will vary from one situation to another. There is no universal rule, for example, about how much food an athlete should eat, and it would be absurd to infer from the fact that 10 lbs. is too much and 2 lbs. too little for me that I should eat 6 lbs. Finding the mean in any given situation is not a mechanical or thoughtless procedure, but requires a full and detailed acquaintance with the circumstances.

It should be evident that Aristotle's treatment of virtues as mean states endorses the idea that we should sometimes have strong feelings -- when such feelings are called for by our situation. Sometimes a only a small degree of anger is appropriate; but at other times, circumstances call for great anger. The right amount is not some quantity between zero and the highest possible level, but rather the amount, whatever it happens to be, that is proportionate to the seriousness of the situation. Of course, Aristotle is committed to saying that anger should never reach the point at which it undermines reason; and this means that our passion should always fall short of the extreme point at which we would lose control. But it is possible to

be very angry without going to this extreme, and Aristotle does not intend to deny this.

The theory of the mean is open to several objections, but before considering them, we should recognize that in fact there are two distinct theses each of which might be called a doctrine of the mean. First, there is the thesis that every virtue is a state that lies between two vices, one of excess and the other of deficiency. Second, there is the idea that whenever a virtuous person chooses to perform a virtuous act, he can be described as aiming at an act that is in some way or other intermediate between alternatives that he rejects. It is this second thesis that is most likely to be found objectionable. A critic might concede that in some cases virtuous acts can be described in Aristotle's terms. If, for example, one is trying to decide how much to spend on a wedding present, one is looking for an amount that is neither excessive nor deficient. But surely many other problems that confront a virtuous agent are not susceptible to this quantitative analysis. If one must decide whether to attend a wedding or respect a competing obligation instead, it would not be illuminating to describe this as a search for a mean between extremes -- unless "aiming at the mean" simply becomes another phrase for trying to make the right decision. The objection, then, is that Aristotle's doctrine of the mean, taken as a doctrine about what the ethical agent does when he deliberates, is in many cases inapplicable or unilluminating.

A defense of Aristotle would have to say that the virtuous person does after all aim at a mean, if we allow for a broad enough notion of what sort of aiming is involved. For example, consider a juror who must determine whether a defendant is guilty as charged. He does not have before his mind a quantitative question; he is trying to decide whether the accused committed the crime, and is not looking for some quantity of action intermediate between extremes. Nonetheless, an excellent juror can be described as someone who, in trying to arrive at the correct decision, seeks to express the right degree of concern for all relevant considerations. He searches for the verdict that results from a deliberative process that is neither overly credulous or unduly skeptical. Similarly, in facing situations that arouse anger, a virtuous agent must determine what action (if any) to take in response to an insult, and although this is not itself a quantitative question, his attempt to answer it properly requires him to have the right degree of concern for his standing as a member of the community. He aims at a mean in the sense that he looks for a response that avoids too much or too little attention to factors that must be taken into account in making a wise decision.

Perhaps a greater difficulty can be raised if we ask how Aristotle determines which emotions are governed by the doctrine of the mean. Consider someone who loves to wrestle, for example. Is this passion something that must be felt by every human being at appropriate times and to the right degree? Surely someone who never felt this emotion to any degree could still live a perfectly happy life. Why then should we not say the same about at least some of the emotions that Aristotle builds into his analysis of the ethically virtuous agent? Why should we experience anger at all, or fear, or the degree of concern for wealth and honor that Aristotle commends? These are precisely the questions that were asked in antiquity by the Stoics, and they came to the conclusion that such common emotions as anger and fear are always inappropriate. Aristotle assumes, on the contrary, not simply that these common passions are sometimes appropriate, but that it is essential that every human being learn how to master them and experience them in the right way at the right times. A defense of his position would have to show that the emotions that figure in his account of the virtues are valuable components of any well-lived human life, when they are

experienced properly. Perhaps such a project could be carried out, but Aristotle himself does not attempt to do so.

He often says, in the course of his discussion, that when the good person chooses to act virtuously, he does so for the sake of the "*kalon*" -- a word that can mean "beautiful," "noble," or "fine." (See for example 1120a23-4.) This term indicates that Aristotle sees in ethical activity an attraction that is comparable to the beauty of well-crafted artifacts, including such artifacts as poetry, music, and drama. He draws this analogy in his discussion of the mean, when he says that every craft tries to produce a work from which nothing should be taken away and to which nothing further should be added (1106b5-14). A craft product, when well designed and produced by a good craftsman, is not merely useful, but also has such elements as balance, proportion and harmony -- for these are properties that help make it useful. Similarly, Aristotle holds that a well-executed project that expresses the ethical virtues will not merely be advantageous but *kalon* as well -- for the balance it strikes is part of what makes it advantageous. The young person learning to acquire the virtues must develop a love of doing what is *kalon* and a strong aversion to its opposite -- the *aischron*, the shameful and ugly. Determining what is *kalon* is difficult (1106b28-33, 1109a24-30, and the normal human aversion to embracing difficulties helps account for the scarcity of virtue (1104b10-11).

B. Ethical Theory Does Not Offer a Decision Procedure

It should be clear that neither the thesis that virtues lie between extremes nor the thesis that the good person aims at what is intermediate is intended as a procedure for making decisions. These doctrines of the mean help show what is attractive about the virtues, and they also help systematize our understanding of which qualities are virtues. Once we see that temperance, courage, and other generally recognized characteristics are mean states, we are in a position to generalize and to identify other mean states as virtues, even though they are not qualities for which we have a name. Aristotle remarks, for example, that the mean state with respect to anger has no name in Greek (1125b26-7). Though he is guided to some degree by distinctions captured by ordinary terms, his methodology allows him to recognize states for which no names exist.

So far from offering a decision procedure, Aristotle insists that this is something that no ethical theory can do. His theory elucidates the nature of virtue, but what must be done on any particular occasion by a virtuous agent depends on the circumstances, and these vary so much from one occasion to another that there is no possibility of stating a series of rules, however complicated, that collectively solve every practical problem. This feature of ethical theory is not unique; Aristotle thinks it applies to many crafts, such as medicine and navigation (1104a7-10). He says that the virtuous person "sees the truth in each case, being as it were a standard and measure of them" (1113a32-3); but this appeal to the good person's vision should not be taken to mean that he has an inarticulate and incommunicable insight into the truth. Aristotle thinks of the good person as someone who is good at deliberation, and he describes deliberation as a process of rational inquiry. The intermediate point that the good person tries to find is "determined by *logos* ("reason," "account") and in the way that the person of practical reason would determine it" (1107a1-2). To say that such a person "sees" what to do is simply a way of registering the point that the good person's reasoning does succeed in discovering what is best in each situation. He is "as it were a standard

and measure" in the sense that his views should be regarded as authoritative by other members of the community. A standard or measure is something that settles disputes; and because good people are so skilled at discovering the mean in difficult cases, their advice must be sought and heeded.

Although there is no possibility of writing a book of rules, however long, that will serve as a complete guide to wise decision-making, it would be a mistake to attribute to Aristotle the opposite position, namely that every purported rule admits of exceptions, so that even a small rule-book that applies to a limited number of situations is an impossibility. He makes it clear that certain emotions (spite, shamelessness, envy) and actions (adultery, theft, murder) are always wrong, regardless of the circumstances (1107a8-12). Although he says that the names of these emotions and actions convey their wrongness, he should not be taken to mean that their wrongness derives from linguistic usage. He defends the family as a social institution against the criticisms of Plato (*Politics* II.3-4), and so when he says that adultery is always wrong, he is prepared to argue for his point by explaining why marriage is a valuable custom and why extra-marital intercourse undermines the relationship between husband and wife. He is not making the tautological claim that wrongful sexual activity is wrong, but the more specific and contentious point that marriages ought to be governed by a rule of strict fidelity. Similarly, when he says that murder and theft are always wrong, he does not mean that wrongful killing and taking are wrong, but that the current system of laws regarding these matters ought to be strictly enforced. So, although Aristotle holds that ethics cannot be reduced to a system of rules, however complex, he insists that some rules are inviolable.

C. The Starting Point for Practical Reasoning

We have seen that the decisions of a practically wise person are not mere intuitions, but can be justified by a chain of reasoning. (This is why Aristotle often talks in term of a practical syllogism, with a major premise that identifies some good to be achieved, and a minor premise that locates the good in some present-to-hand situation.) At the same time, he is acutely aware of the fact that reasoning can always be traced back to a starting point that is not itself justified by further reasoning. Neither good theoretical reasoning nor good practical reasoning moves in a circle; true thinking always presupposes and progresses in linear fashion from proper starting points. And that leads him to ask for an account of how the proper starting points of reasoning are to be determined. Practical reasoning always presupposes that one has some end, some goal one is trying to achieve; and the task of reasoning is determine how that goal is to be accomplished. (This need not be means-end reasoning in the conventional sense; if, for example, our goal is the just resolution of a conflict, we must determine what constitutes justice in these particular circumstances. Here we are engaged in ethical inquiry, and are not asking a purely instrumental question.) But if practical reasoning is correct only if it begins from a correct premise, what is it that insures the correctness of its starting point?

Aristotle replies: "Virtue makes the goal right, practical wisdom the things leading to it" (1144a7-8). By this he cannot mean that there is no room for reasoning about our ultimate end. For as we have seen, he gives a reasoned defense of his conception of happiness as virtuous activity. What he must have in mind, when he says that virtue makes the goal right, is that deliberation typically proceeds from a goal that is far more specific than the goal of attaining happiness by acting virtuously. To be sure, there may be

occasions when a good person approaches an ethical problem by beginning with the premise that happiness consists in virtuous activity. But more often what happens is that a concrete goal presents itself as his starting point -- helping a friend in need, or supporting a worthwhile civic project. Which specific project we set for ourselves is determined by our character. A good person starts from worthwhile concrete ends because his habits and emotional orientation have given him the ability to recognize that such goals are within reach, here and now. Those who are defective in character may have the rational skill needed to achieve their ends -- the skill Aristotle calls cleverness (1144a23-8) -- but often the ends they seek are worthless. The cause of this deficiency lies not in some impairment in their capacity to reason -- for we are assuming that they are normal in this respect -- but in the training of their passions.

6. Intellectual Virtues

Since Aristotle often calls attention to the imprecision of ethical theory (see e.g. 1104a1-7), it comes as a surprise to many readers of the *Ethics* that he begins Book VI with the admission that his earlier statements about the mean need supplementation because they are not yet clear (*saphes*). In every practical discipline, the expert aims at a mark and uses right reason to avoid the twin extremes of excess and deficiency. But what is this right reason, and by what standard (*horos*) is it to be determined? Aristotle says that unless we answer that question, we will be none the wiser -- just as a student of medicine will have failed to master his subject if he can only say that the right medicines to administer are the ones that are prescribed by medical expertise, but has no standard other than this (1128b18-34).

It is not easy to understand the point Aristotle is making here. Has he not already told us that there can be no complete theoretical guide to ethics, that the best one can hope for is that in particular situations one's ethical habits and practical wisdom will help one to determine what to do? Furthermore, Aristotle nowhere announces, in the remainder of Book VI, that we have achieved the greater degree of accuracy that he seems to be looking for. The rest of this Book is a discussion of the various kinds of intellectual virtues: theoretical wisdom, science (*epistêmê*), intuitive understanding (*nous*), practical wisdom, and craft expertise. Aristotle explains what each of these states of mind is, draws various contrasts among them, and takes up various questions that can be raised about their usefulness. At no point does he explicitly return to the question he raised at the beginning of Book VI; he never says, "and now we have the standard of right reason that we were looking for." Nor is it easy to see how his discussion of these five intellectual virtues can bring greater precision to the doctrine of the mean.

We can make some progress towards solving this problem if we remind ourselves that at the beginning of the *Ethics*, Aristotle describes his inquiry as an attempt to develop a better understanding of what our ultimate aim should be. The sketchy answer he gives in Book I is that happiness consists in virtuous activity. In Books II through V, he describes the virtues of the part of the soul that is rational in that it can be attentive to reason, even though it is not capable of deliberating. But precisely because these virtues are rational only in this derivative way, they are a less important component of our ultimate end than is the intellectual virtue -- practical wisdom -- with which they are integrated. If what we know about virtue is only what is said in Books II through V, then our grasp of our ultimate end is radically incomplete, because we still have not studied the intellectual virtue that enables us to reason well in any given

situation. One of the things, at least, towards which Aristotle is gesturing, as he begins Book VI, is practical wisdom. This state of mind has not yet been analyzed, and that is one reason why he complains that his account of our ultimate end is not yet clear enough.

But is practical wisdom the only ingredient of our ultimate end that has not yet been sufficiently discussed? Book VI discusses five intellectual virtues, not just practical wisdom, but it is clear that at least one of these -- craft knowledge -- is considered only in order to provide a contrast with the others. Aristotle is not recommending that his readers make this intellectual virtue part of their ultimate aim. But what of the remaining three: science, intuitive understanding, and the virtue that combines them, theoretical wisdom? Are these present in Book VI only in order to provide a contrast with practical wisdom, or is Aristotle saying that these too must be components of our goal? He does not fully address this issue, but it is evident from several of his remarks in Book VI that he takes theoretical wisdom to be a more valuable state of mind than practical wisdom. "It is strange if someone thinks that politics or practical wisdom is the most excellent kind of knowledge, unless man is the best thing in the cosmos" (1141a20-22). He says that theoretical wisdom produces happiness by being a part of virtue (1144a3-6), and that practical wisdom looks to the development of theoretical wisdom, and issues commands for its sake (1145a8-11). So it is clear that exercising theoretical wisdom is a more important component of our ultimate goal than practical wisdom.

Even so, it may still seem perplexing that these two intellectual virtues, either separately or collectively, should somehow fill a gap in the doctrine of the mean. Having read Book VI and completed our study of what these two forms of wisdom are, how are we better able to succeed in finding the mean in particular situations?

The answer to this question may be that Aristotle does not intend Book VI to provide a full answer to that question, but rather to serve as a prolegomenon to an answer. For it is only near the end of Book X that he presents a full discussion of the relative merits of these two kinds of intellectual virtue, and comments on the different degrees to which each needs to be provided with resources. In X.7-8, he argues that the happiest kind of life is that of a philosopher -- someone who exercises, over a long period of time, the virtue of theoretical wisdom, and has sufficient resources for doing so. (We will discuss these chapters more fully in section 10 below.) One of his reasons for thinking that such a life is superior to the second-best kind of life -- that of a political leader, someone who devotes himself to the exercise of practical rather than theoretical wisdom -- is that it requires less external equipment (1178a23-b7). Aristotle has already made it clear in his discussion of the ethical virtues that someone who is greatly honored by his community and commands large financial resources is in a position to exercise a higher order of ethical virtue than is someone who receives few honors and has little property. The virtue of magnificence is superior to mere liberality, and similarly greatness of soul is a higher excellence than the ordinary virtue that has to do with honor. (These qualities are discussed in IV.1-4.) The grandest expression of ethical virtue requires great political power, because it is the political leader who is in a position to do the greatest amount of good for the community. The person who chooses to lead a political life, and who aims at the fullest expression of practical wisdom, has a standard for deciding what level of resources he needs: he should have friends, property, and honors in sufficient quantities to allow his practical wisdom to express itself without impediment. But if one chooses instead the life of a philosopher, then one will look to a

different standard -- the fullest expression of theoretical wisdom -- and one will need a smaller supply of these resources.

This enables us to see how Aristotle's treatment of the intellectual virtues does give greater content and precision to the doctrine of the mean. The best standard is the one adopted by the philosopher; the second-best is the one adopted by the political leader. In either case, it is the exercise of an intellectual virtue that provides a guideline for making important quantitative decisions. This supplement to the doctrine of the mean is fully compatible with Aristotle's thesis that no set of rules, no matter how long and detailed, obviates the need for deliberative and ethical virtue. If one chooses the life of a philosopher, one should keep the level of one's resources high enough to secure the leisure necessary for such a life, but not so high that one's external equipment becomes a burden and a distraction rather than an aid to living well. That gives one a firmer idea of how to hit the mean, but it still leaves the details to be worked out. The philosopher will need to determine, in particular situations, where justice lies, what generosity requires, when courage requires meeting a danger, and so on. All of the normal difficulties of ethical life remain, and they can be solved only by means of a detailed understanding of the particulars of each situation. Having philosophy as one's ultimate aim does not put an end to the need for developing and exercising practical wisdom and the ethical virtues.

7. *Akrasia*

In VII.1-10 Aristotle investigates character traits -- continence and incontinence -- that are not as blameworthy as the vices but not as praiseworthy as the virtues. (We began our discussion of these qualities in section 4.) The Greek terms are *akrasia* ("incontinence"; literally: "lack of mastery") and *enkrateia* ("continence"; literally "mastery"). An akratic person goes against reason as a result of some *pathos* ("emotion," "feeling"). Like the akratic, an enkratic person experiences a feeling that is contrary to reason; but unlike the akratic, he acts in accordance with reason. His defect consists solely in the fact that, more than most people, he experiences passions that conflict with his rational choice. The akratic person has not only this defect, but has the further flaw that he gives in to feeling rather than reason more often than the average person.

Aristotle distinguishes two kinds of akrasia: impetuosity (*propeteia*) and weakness (*astheneia*). The person who is weak goes through a process of deliberation and makes a choice; but rather than act in accordance with his reasoned choice, he acts under the influence of a passion. By contrast, the impetuous person does not go through a process of deliberation and does not make a reasoned choice; he simply acts under the influence of a passion. At the time of action, the impetuous person experiences no internal conflict. But once his act has been completed, he regrets what he has done. One could say that he deliberates, if deliberation were something that post-dated rather than preceded action; but the thought process he goes through after he acts comes too late to save him from error.

It is important to bear in mind that when Aristotle talks about impetuosity and weakness, he is discussing chronic conditions. The impetuous person is someone who acts emotionally and fails to deliberate not just once or twice but with some frequency; he makes this error more than most people do. Because of this

pattern in his actions, we would be justified in saying of the impetuous person that had his passions not prevented him from doing so, he would have deliberated and chosen an action different from the one he did perform.

The two kinds of passions that Aristotle focuses on, in his treatment of *akrasia*, are the appetite for pleasure and anger. Either can lead to impetuosity and weakness. But Aristotle gives pride of place to the appetite for pleasure as the passion that undermines reason. He calls the kind of *akrasia* caused by an appetite for pleasure "unqualified *akrasia*" -- or, as we might say, *akrasia* "full stop"; *akrasia* caused by anger he considers a qualified form of *akrasia* and calls it *akrasia* "with respect to anger". We thus have these four forms of *akrasia*: (A) impetuosity caused by pleasure, (B) impetuosity caused by anger, (C) weakness caused by pleasure (D) weakness caused by anger. It should be noticed that Aristotle's treatment of *akrasia* is heavily influenced by Plato's tripartite division of the soul in the *Republic*. Plato holds that either the spirited part (which houses anger, as well as other emotions) or the appetitive part (which houses the desire for physical pleasures) can disrupt the dictates of reason and result in action contrary to reason. The same threefold division of the soul can be seen in Aristotle's approach to this topic.

Although Aristotle characterizes *akrasia* and *enkrateia* in terms of a conflict between reason and feeling, his detailed analysis of these states of mind shows that what takes place is best described in a more complicated way. For the feeling that undermines reason contains some thought, which may be implicitly general. As Aristotle says, anger "reasoning as it were that one must fight against such a thing, is immediately provoked"(1149a33-4). And although in the next sentence he denies that our appetite for pleasure works in this way, he earlier had said that there can be a syllogism that favors pursuing enjoyment: "Everything sweet is pleasant, and this is sweet" leads to the pursuit of a particular pleasure (1147a31-30). Perhaps what he has in mind is that pleasure can operate in either way: it can prompt action unmediated by a general premise, or it can prompt us to act on such a syllogism. By contrast, anger always moves us by presenting itself as a bit of general, although hasty, reasoning.

But of course Aristotle does not mean that a conflicted person has more than one faculty of reason. Rather his idea seems to be that in addition to our full-fledged reasoning capacity, we also have psychological mechanisms that are capable of a limited range of reasoning. When feeling conflicts with reason, what occurs is better described as a fight between feeling-allied-with-limited-reasoning and full-fledged reason. Part of us -- reason -- can remove itself from the distorting influence of feeling and consider all relevant factors, positive and negative. But another part of us -- feeling or emotion -- has a more limited field of reasoning -- and sometimes it does not even make use of it.

Although "passion" is sometimes used as a translation of Aristotle's word *pathos* (other alternatives are "emotion" and "feeling"), it is important to bear in mind that his term does not necessarily designate a strong psychological force. Anger is a *pathos* whether it is weak or strong; so too is the appetite for bodily pleasures. And he clearly indicates that it is possible for an akratic person to be defeated by a weak *pathos* -- the kind that most people would easily be able to control (1150a9-b16). So the general explanation for the occurrence of *akrasia* cannot be that the strength of a passion overwhelms reason. Aristotle should therefore be acquitted of an accusation made against him by J.L. Austin in a well-known footnote to his paper, "A Plea For Excuses." Plato and Aristotle, he says, collapsed all succumbing to temptation into

losing control of ourselves -- a mistake illustrated by this example: "I am very partial to ice cream, and a bombe is served divided into segments corresponding one to one with the persons at High Table: I am tempted to help myself to two segments and do so, thus succumbing to temptation and even conceivably (but why necessarily?) going against my principles. But do I lose control of myself? Do I raven, do I snatch the morsels from the dish and wolf them down, impervious to the consternation of my colleagues? Not a bit of it. We often succumb to temptation with calm and even with finesse." (*Philosophical Papers*, 1961, p. 146.) With this, Aristotle can agree: the *pathos* for the bombe can be a weak one, and in some people that will be enough to get them to act in a way that is disapproved by their reason at the very time of action.

What is most remarkable about Aristotle's discussion of *akrasia* is that he defends a position close to that of Socrates. When he first introduces the topic of *akrasia*, and surveys some of the problems involved in understanding this phenomenon, he says (1145b25-8) that Socrates held that there is no *akrasia*, and he describes this as a thesis that clearly conflicts with the appearances (*phainomena*). Since he says that his goal is to preserve as many of the appearances as possible (1145b2-7), it may come as a surprise that when he analyzes the conflict between reason and feeling, he arrives at the conclusion that in a way Socrates was right after all (1147b13-17). For, he says, the person who acts against reason does not have what is thought to be unqualified knowledge; in a way he has knowledge, but in a way does not.

Aristotle explains what he has in mind by comparing *akrasia* to the condition of other people who might be described as knowing in a way, but not in an unqualified way. His examples are people who are asleep, mad, or drunk; he also treats the akratic as someone like a student who has just begun to learn a subject, or an actor on the stage (1147a10-24). All of these people, he says, can utter the very words used by those who have knowledge; but their talk does not prove that they really have knowledge, strictly speaking.

These analogies can be taken to mean that the form of *akrasia* that Aristotle calls weakness rather than impetuosity always results from some diminution of cognitive or intellectual acuity at the moment of action. The akratic says, at the time of action, that he ought not to indulge in this particular pleasure at this time. But does he know or even believe that he should refrain? Aristotle might be taken to reply: yes and no. He has some degree of recognition that he must not do this now, but not full recognition. His feeling, even if it is weak, has to some degree prevented him from completely grasping or affirming the point that he should not do this. And so in a way Socrates was right. When reason remains unimpaired and unclouded, its dictates will carry us all the way to action, so long as we are able to act.

But Aristotle's agreement with Socrates is only partial, because he insists on the power of the emotions to rival, weaken or bypass reason. Emotion challenges reason in all three of these ways. In both the akratic and the enkratic, it competes with reason for control over action; even when reason wins, it faces the difficult task of having to struggle with an internal rival. Second, in the akratic, it temporarily robs reason of its full acuity, thus handicapping it as a competitor. It is not merely a rival force, in these cases; it is a force that keeps reason from fully exercising its power. And third, passion can make someone impetuous; here its victory over reason is so powerful that the latter does not even enter into the arena of conscious reflection until it is too late to influence action.

[Supplementary Document: Alternate Readings of Aristotle on *Akrasia*]

8. Pleasure

Aristotle frequently emphasizes the importance of pleasure to human life and therefore to his study of how we should live (see for example 1099a7-20 and 1104b3-1105a16), but his full-scale examination of the nature and value of pleasure is found in two places: VII.11-14 and X.1-5. It is odd that pleasure receives two lengthy treatments; no other topic in the *Ethics* is revisited in this way. Book VII of the *Nicomachean Ethics* is identical to Book VI of the *Eudemian Ethics*; for unknown reasons, the editor of the former decided to include within it both the treatment of pleasure that is unique to that work (X.1-5) and the study that is common to both treatises (VII.11-14). The two accounts are broadly similar. They agree about the value of pleasure, defend a theory about its nature, and oppose competing theories. Aristotle holds that a happy life must include pleasure, and he therefore opposes those who argue that pleasure is by its nature bad. He insists that there are other pleasures besides those of the senses, and that the best pleasures are the ones experienced by virtuous people who have sufficient resources for excellent activity.

Book VII offers a brief account of what pleasure is and is not. It is not a process but an unimpeded activity of a natural state (1153a7-17). Aristotle does not elaborate on what a natural state is, but he obviously has in mind the healthy condition of the body, especially its sense faculties, and the virtuous condition of the soul. Little is said about what it is for an activity to be unimpeded, but Aristotle does remind us that virtuous activity is impeded by the absence of a sufficient supply of external goods (1153b17-19). One might object that people who are sick or who have moral deficiencies can experience pleasure, even though Aristotle does not take them to be in a natural state. He has two strategies for responding. First, when a sick person experiences some degree of pleasure as he is being restored to health, the pleasure he is feeling is caused by the fact that he is no longer completely ill. Some small part of him is in a natural state and is acting without impediment (1152b35-6). Second, Aristotle is willing to say that what seems pleasant to some people may in fact not be pleasant (1152b31-2), just as what tastes bitter to an unhealthy palate may not be bitter. To call something a pleasure is not only to report a state of mind but also to endorse it to others. Aristotle's analysis of the nature of pleasure is not meant to apply to every case in which something seems pleasant to someone, but only to activities that really are pleasures. All of these are unimpeded activities of a natural state.

It follows from this conception of pleasure that every instance of pleasure must be good to some extent. For how could an unimpeded activity of a natural state be bad or a matter of indifference? On the other hand, Aristotle does not mean to imply that every pleasure should be chosen. He briefly mentions the point that pleasures compete with each other, so that the enjoyment of one kind of activity impedes other activities that cannot be carried out at the same time (1153a20-22). His point is simply that although some pleasures may be good, they are not worth choosing when they interfere with other activities that are far better. This point is developed more fully in *Ethics* X.5.

Furthermore, Aristotle's analysis allows him to speak of certain pleasures as "bad without qualification"

(1152b26-33), even though pleasure is the unimpeded activity of a natural state. To call a pleasure "bad without qualification" is to insist that it should be avoided, but allow that nonetheless it should be chosen in constraining circumstances. The pleasure of recovering from an illness, for example, is bad without qualification -- meaning that it is not one of the pleasures one would ideally choose, if one could completely control one's circumstances. Although it really is a pleasure and so something can be said in its favor, it is so inferior to other goods that ideally one ought to forego it. Nonetheless, it is a pleasure worth having -- if one adds the qualification that it is only worth having in undesirable circumstances. The pleasure of recovering from an illness is good, because some small part of oneself is in a natural state and is acting without impediment; but it can also be called bad, if what one means by this is that one should avoid getting into a situation in which one experiences that pleasure.

Aristotle indicates several times in VII.11-14 that merely to say that pleasure is *a* good does not do it enough justice; he also wants to say that the highest good is a pleasure. Here he is influenced by an idea expressed in the opening line of the *Ethics*: the good is that at which all things aim. In VII.13, he hints at the idea that all living things imitate the contemplative activity of god (1153b31-2). Plants and non-human animals seek to reproduce themselves because that is their way of participating in an unending series, and this is the closest they can come to the ceaseless thinking of the unmoved mover. Aristotle makes this point in several of his works (see for example *De Anima* 415a23-b7), and in *Ethics* X.7-8 he gives a full defense of the idea that the happiest human life resembles the life of a divine being. He conceives of god as a being who continually enjoy a "single and simple pleasure" (1154b26) -- the pleasure of pure thought -- whereas human beings, because of their complexity, grow weary of whatever they do. He will elaborate on these points in X.8; in VII.11-14, he appeals to his conception of divine activity only in order to defend the thesis that our highest good consists in a certain kind of pleasure. Human happiness does not consist in every kind of pleasure, but it does consist in one kind of pleasure -- the pleasure felt by a human being who engages in theoretical activity and thereby imitates the pleasurable thinking of god.

Book X offers a much more elaborate account of what pleasure is and what it is not. It is not a process, because processes go through developmental stages: building a temple is a process because the temple is not present all at once, but only comes into being through stages that unfold over time. By contrast, pleasure, like seeing and many other activities, is not something that comes into existence through a developmental process. If I am enjoying a conversation, for example, I do not need to wait until it is finished in order to feel pleased; I take pleasure in the activity all along the way. The defining nature of pleasure is that it is an activity that accompanies other activities, and in some sense brings them to completion. Pleasure occurs when something within us, having been brought into good condition, is activated in relation to an external object that is also in good condition. The pleasure of drawing, for example, requires both the development of drawing ability and an object of attention that is worth drawing.

The conception of pleasure that Aristotle develops in Book X is obviously closely related to the analysis he gives in Book VII. But the theory proposed in the later Book brings out a point that had received too little attention earlier: pleasure is by its nature something that accompanies something else. It is not enough to say that it is what happens when we are in good condition and are active in unimpeded circumstances; one must add to that point the further idea that pleasure plays a certain role in

complementing something other than itself. Drawing well and the pleasure of drawing well always occur together, and so they are easy to confuse, but Aristotle's analysis in Book X emphasizes the importance of making this distinction.

He says that pleasure completes the activity that it accompanies, but then adds, mysteriously, that it completes the activity in the manner of an end that is added on. In the translation of W.D. Ross, it "supervenes as the bloom of youth does on those in the flower of their age" (1174b33). It is unclear what thought is being expressed here, but perhaps Aristotle is merely trying to avoid a possible misunderstanding: when he says that pleasure completes an activity, he does not mean that the activity it accompanies is in some way defective, and that the pleasure improves the activity by removing this defect. Aristotle's language is open to that misinterpretation because the verb that is translated "complete" (*teleein*) can also mean "perfect." The latter might be taken to mean that the activity accompanied by pleasure has not yet reached a sufficiently high level of excellence, and that the role of pleasure is to bring it to the point of perfection. Aristotle does not deny that when we take pleasure in an activity we get better at it, but when he says that pleasure completes an activity by supervening on it, like the bloom that accompanies those who have achieved the highest point of physical beauty, his point is that the activity complemented by pleasure is already perfect, and the pleasure that accompanies it is a bonus that serves no further purpose. Taking pleasure in an activity does help us improve at it, but enjoyment does not cease when perfection is achieved -- on the contrary, that is when pleasure is at its peak. That is when it reveals most fully what it is: an added bonus that crowns our achievement.

It is clear, at any rate, that in Book X Aristotle gives a fuller account of what pleasure is than he had in Book VII. We should take note of a further difference between these two discussions: In Book X, he makes the point that pleasure is *a* good but not *the* good. He cites and endorses an argument given by Plato in the *Philebus*: If we imagine a life filled with pleasure and then mentally add wisdom to it, the result is made more desirable. But the good is something that cannot be improved upon in this way. Therefore pleasure is not the good (1172b23-35). By contrast, in Book VII Aristotle strongly implies that the pleasure of contemplation *is* the good, because in one way or another all living beings aim at this sort of pleasure. Aristotle observes in Book X that what all things aim at is good (1172b35-1173a1); significantly, he falls short of endorsing the argument that since all aim at pleasure, it must be *the* good.

Book VII makes the point that pleasures interfere with each other, and so even if all kinds of pleasures are good, it does not follow that all of them are worth choosing. One must make a selection among pleasures by determining which are better. But how is one to make this choice? Book VII does not say, but in Book X, Aristotle holds that the selection of pleasures is not to be made with reference to pleasure itself, but with reference to the activities they accompany. "Since activities differ with respect to goodness and badness, some being worth choosing, others worth avoiding, and others neither, the same is true of pleasures as well" (1175b24-6). Aristotle's statement implies that in order to determine whether (for example) the pleasure of virtuous activity is more desirable than that of eating, we are not to attend to the pleasures themselves but to the activities with which we are pleased. A pleasure's goodness derives from the goodness of its associated activity. And surely the reason why pleasure is not the criterion to which we should look in making these decisions is that it is not the good. The standard we should use in making comparisons between rival options is virtuous activity, because that has been shown to be identical to

happiness.

That is why Aristotle says that what is judged pleasant by a good man really is pleasant, because the good man is the measure of things (1176a15-19). He does not mean that the way to lead our lives is to search for a good man and continually rely on him to tell us what is pleasurable. Rather, his point is that there is no way of telling what is genuinely pleasurable (and therefore what is most pleasurable) unless we already have some other standard of value. Aristotle's discussion of pleasure thus helps confirm his initial hypothesis that to live our lives well we must focus on one sort of good above all others: virtuous activity. It is the good in terms of which all other goods must be understood. Aristotle's analysis of friendship supports the same conclusion.

9. Friendship

The topic of Books VIII and IX of the *Ethics* is friendship. Although it is difficult to avoid the term "friendship" as a translation of "*philia*," and this is an accurate term for the kind of relationship he is most interested in, we should bear in mind that he is discussing a wider range of phenomena than this translation might lead us to expect, for the Greeks use the term, "*philia*," to name the relationship that holds among family members, and do not reserve it for voluntary relationships. Although Aristotle is interested in classifying the different forms that friendship takes, his main theme in Books VIII and IX is to show the close relationship between virtuous activity and friendship. He is vindicating his conception of happiness as virtuous activity by showing how satisfying are the relationships that a virtuous person can normally expect to have.

His taxonomy begins with the premise that there are three main reasons why one person might like someone else. (The verb, "*philein*," which is cognate to the noun "*philia*," can sometimes be translated "like" or even "love" -- though in other cases *philia* involves very little in the way of feeling.) One might like someone because he is good, or because he is useful, or because he is pleasant. And so there are three bases for friendships, depending on which of these qualities binds friends together. When two individuals recognize that the other person is someone of good character, and they spend time with each other, engaged in activities that exercise their virtues, then they form one kind of friendship. If they are equally virtuous, their friendship is perfect. If, however, there is a large gap in their moral development (as between a parent and a small child, or between a husband and a wife), then although their relationship may be based on the other person's good character, it will be imperfect precisely because of their inequality.

The imperfect friendships that Aristotle focuses on, however, are not unequal relationships based on good character. Rather, they are relationships held together because each individual regards the other as the source of some advantage to himself or some pleasure he receives. When Aristotle calls these relationships "imperfect," he is tacitly relying on widely accepted assumptions about what makes a relationship satisfying. These friendships are defective, and have a smaller claim to be called "friendships," because the individuals involved have little trust in each other, quarrel frequently, and are ready to break off their association abruptly. Aristotle does not mean to suggest that unequal relations

based on the mutual recognition of good character are defective in these same ways. Rather, when he says that unequal relationships based on character are imperfect, his point is that people are friends in the fullest sense when they gladly spend their days together in shared activities, and this close and constant interaction is less available to those who are not equal in their moral development.

When Aristotle begins his discussion of friendship, he introduces a notion that is central to his understanding of this phenomenon: a genuine friend is someone who loves or likes another person for the sake of that other person. Wanting what is good for the sake of another he calls "good will" (*eunoia*), and friendship is reciprocal good will, provided that each recognizes the presence of this attitude in the other. Does such good will exist in all three kinds of friendship, or is it confined to relationships based on virtue? At first, Aristotle leaves open the first of these two possibilities. He says: "it is necessary that friends bear good will to each other and wish good things for each other, without this escaping their notice, because of one of the reasons mentioned" (1156a4-5). The reasons mentioned are goodness, pleasure, and advantage; and so it seems that Aristotle is leaving room for the idea that in all three kinds of friendships, even those based on advantage and pleasure alone, the individuals wish each other well for the sake of the other.

But in fact, as Aristotle continues to develop his taxonomy, he does not choose to exploit this possibility. He speaks as though it is only in friendships based on character that one finds a desire to benefit the other person for the sake of the other person. "Those who wish good things to their friends for the sake of the latter are friends most of all, because they do so because of their friends themselves, and not coincidentally" (1156b9-11). When one benefits someone not because of the kind of person he is, but only because of the advantages to oneself, then, Aristotle says, one is not a friend towards the other person, but only towards the profit that comes one's way (1157a15-16).

In such statements as these, Aristotle comes rather close to saying that relationships based on profit or pleasure should not be called friendships at all. But he decides to stay close to common parlance and to use the term "friend" loosely. Friendships based on character are the ones in which each person benefits the other for the sake of other; and these are friendships most of all. Because each party benefits the other, it is advantageous to form such friendships. And since each enjoys the trust and companionship of the other, there is considerable pleasure in these relationships as well. Because these perfect friendships produce advantages and pleasures for each of the parties, there is some basis for going along with common usage and calling any relationship entered into for the sake of just one of these goods a friendship. Friendships based on advantage alone or pleasure alone deserve to be called friendships because in full-fledged friendships these two properties, advantage and pleasure, are present. It is striking that in the *Ethics* Aristotle never thinks of saying that the uniting factor in all friendships is the desire each friend has for the good of the other.

Aristotle does not raise questions about what it is to desire good for the sake of another person. He treats this as an easily understood phenomenon, and has no doubts about its existence. But it is also clear that he takes this motive to be compatible with a love of one's own good and a desire for one's own happiness. Someone who has practical wisdom will recognize that he needs friends and other resources in order to exercise his virtues over a long period of time. When he makes friends, and benefits friends he has made,

he will be aware of the fact that such a relationship is good for him. And yet to have a friend is to want to benefit someone for that other person's sake; it is not a merely self-interested strategy. Aristotle sees no difficulty here, and rightly so. For there is no reason why acts of friendship should not be undertaken partly for the good of one's friend and partly for one's own good. Acting for the sake of another does not in itself demand self-sacrifice. It requires caring about someone other than oneself, but does not demand some loss of care for oneself. For when we know how to benefit a friend for his sake, we exercise the ethical virtues, and this is precisely what our happiness consists in.

Aristotle makes it clear that the number of people with whom one can sustain the kind of relationship he calls a perfect friendship is quite small (IX.10). Even if one lived in a city populated entirely by perfectly virtuous citizens, the number with whom one could carry on a friendship of the perfect type would be at most a handful. For he thinks that this kind of friendship can exist only when one spends a great deal of time with the other person, participating in joint activities and engaging in mutually beneficial behavior; and one cannot cooperate on these close terms with every member of the political community. One may well ask why this kind of close friendship is necessary for happiness. If one lived in a community filled with good people, and cooperated on an occasional basis with each of them, in a spirit of good will and admiration, would that not provide sufficient scope for virtuous activity and a well-lived life? Admittedly, close friends are often in a better position to benefit each other than are fellow citizens, who generally have little knowledge of one's individual circumstances. But this only shows that it is advantageous to be on the receiving end of a friend's help. The more important question for Aristotle is why one needs to be on the giving end of this relationship. And obviously the answer cannot be that one needs to give in order to receive; that would turn active love for one's friend into a mere means to the benefits received.

Aristotle attempts to answer this question in IX.11, but his treatment is disappointing. His fullest argument depends crucially on the notion that a friend is "another self," someone, in other words, with whom one has a relationship very similar to the relationship one has with oneself. A virtuous person loves the recognition of himself as virtuous; to have a close friend is to possess yet another person, besides oneself, whose virtue one can recognize at extremely close quarters; and so, it must be desirable to have someone very much like oneself whose virtuous activity one can perceive. The argument is unconvincing because it does not explain why the perception of virtuous activity in fellow citizens would not be an adequate substitute for the perception of virtue in one's friends.

Aristotle would be on stronger grounds if he could show that in the absence of close friends one would be severely restricted in the kinds of virtuous activities one could undertake. But he cannot present such an argument, because he does not believe it. He says that it is "finer and more godlike" to bring about the well being of a whole city than to sustain the happiness of just one person (1094b7-10). He refuses to regard private life -- the realm of the household and the small circle of one's friends -- as the best or most favorable location for the exercise of virtue. He is convinced that the loss of this private sphere would greatly detract from a well-lived life, but he is hard put to explain why. He might have done better to focus on the benefits of being the object of a close friend's solicitude. Just as property is ill cared for when it owned by all, and just as a child would be poorly nurtured were he to receive no special parental care -- points Aristotle makes in *Politics* II.2-5 -- so in the absence of friendship we would lose a benefit that could not be replaced by the care of the larger community. But Aristotle is not looking for a defense of

this sort, because he conceives of friendship as lying primarily in activity rather than receptivity. It is difficult, within his framework, to show that virtuous activity towards a friend is a uniquely important good.

Since Aristotle thinks that the pursuit of one's own happiness, properly understood, requires ethically virtuous activity and will therefore be of great value not only to one's friends but to the larger political community as well, he argues that self-love is an entirely proper emotion -- provided it is expressed in the love of virtue (IX.8). Self-love is rightly condemned when it consists in the pursuit of as large a share of external goods -- particularly wealth and power -- as one can acquire, because such self-love inevitably brings one into conflict with others and undermines the stability of the political community. It may be tempting to cast Aristotle's defense of self-love into modern terms by calling him an egoist, and "egoism" is a broad enough term so that, properly defined, it can be made to fit Aristotle's ethical outlook. If egoism is the thesis that one will always act rightly if one consults one's self-interest, properly understood, then nothing would be amiss in identifying him as an egoist.

But egoism is sometimes understood in a stronger sense. Just as consequentialism is the thesis that one should maximize the general good, whatever the good turns out to be, so egoism can be defined as the parallel thesis that one should maximize one's own good, whatever the good turns out to be. Egoism, in other words, can be treated as a purely formal thesis: it holds that whether the good is pleasure, or virtue, or the satisfaction of desires, one should not attempt to maximize the total amount of good in the world, but only one's own. When egoism takes this abstract form, it is an expression of the idea that the claims of others are never worth attending to, unless in some way or other their good can be shown to serve one's own. The only underived reason for action is self-interest; that an act helps another does not by itself provide a reason for performing it, unless some connection can be made between the good of that other and one's own.

There is no reason to attribute this extreme form of egoism to Aristotle. On the contrary, his defense of self-love makes it clear that he is not willing to defend the bare idea that one ought to love oneself alone or above others; he defends self-love only when this emotion is tied to the correct theory of where one's good lies, for it is only in this way that he can show that self-love need not be a destructive passion. He takes it for granted that self-love is properly condemned whenever it can be shown to be harmful to the community. It is praiseworthy only if it can be shown that a self-lover will be an admirable citizen. In making this assumption, Aristotle reveals that he thinks that the claims of other members of the community to proper treatment are intrinsically valid. This is precisely what a strong form of egoism cannot accept.

We should also keep in mind Aristotle's statement in the *Politics* that the political community is prior to the individual citizen -- just as the whole body is prior to any of its parts (1253a18-29). Aristotle makes use of this claim when he proposes that in the ideal community each child should receive the same education, and that the responsibility for providing such an education should be taken out of the hands of private individuals and made a matter of common concern (1337a21-7). No citizen, he says, belongs to himself; all belong to the city (1337a28-9). What he means is that when it comes to such matters as education, which affect the good of all, each individual should be guided by the collective decisions of the

whole community. An individual citizen does not belong to himself, in the sense that it is not up to him alone to determine how he should act; he should subordinate his individual decision-making powers to those of the whole. The strong form of egoism we have been discussing cannot accept Aristotle's doctrine of the priority of the city to the individual. It tells the individual that the good of others has, in itself, no valid claim on him, but that he should serve other members of the community only to the extent that he can connect their interests to his own. Such a doctrine leaves no room for the thought that the individual citizen does not belong to himself but to the whole.

10. Three Lives Compared

In Book I Aristotle says that three kinds of lives are thought to be especially attractive: one is devoted to pleasure, a second to politics, and a third to knowledge and understanding (1095b17-19). In X.6-9 Aristotle returns to these three alternatives, and explores them more fully than he had in Book I. The life of pleasure is construed in Book I as a life devoted to physical pleasure, and is quickly dismissed because of its vulgarity. In X.6, Aristotle concedes that physical pleasures, and more generally, amusements of all sorts, are desirable in themselves, and therefore have some claim to be our ultimate end. But his discussion of happiness in Book X does not start from scratch; he builds on his thesis that pleasure cannot be our ultimate target, because what counts as pleasant must be judged by some standard other than pleasure itself, namely the judgment of the virtuous person. Amusements will not be absent from a happy life, since everyone needs relaxation, and amusements fill this need. But they play a subordinate role, because we seek relaxation in order to return to more important activities.

Aristotle turns therefore, in X.7-8, to the two remaining alternatives -- politics and philosophy -- and presents a series of arguments to show that the philosophical life, a life devoted to *theoria* (contemplation, study), is best. *Theoria* is not the process of learning that leads to understanding; that process is not a candidate for our ultimate end, because it is undertaken for the sake of a further goal. What Aristotle has in mind when he talks about *theoria* is the activity of someone who has already achieved theoretical wisdom. The happiest life is lived by someone who has a full understanding of the basic causal principles that govern the operation of the universe, and who has the resources needed for living a life devoted to the exercise of that understanding. Evidently Aristotle believes that his own life and that of his philosophical friends was the best available to a human being. He compares it to the life of a god: god thinks without interruption and endlessly, and a philosopher enjoys something similar for a limited period of time.

It may seem odd that after devoting so much attention to the practical virtues, Aristotle should conclude his treatise with the thesis that the best activity of the best life is not ethical. In fact, some scholars have held that X.7-8 are deeply at odds with the rest of the *Ethics*; they take Aristotle to be saying that we should be prepared to act unethically, if need be, in order to devote ourselves as much as possible to contemplation. But it is difficult to believe that he intends to reverse himself so abruptly, and there are many indications that he intends the arguments of X.7-8 to be continuous with the themes he emphasizes throughout the rest of the *Ethics*. The best way to understand him is to take him to be assuming that one will need the ethical virtues in order to live the life of a philosopher, even though exercising those virtues is not the philosopher's ultimate end. To be adequately equipped to live a life of thought and discussion,

one will need practical wisdom, temperance, justice, and the other ethical virtues. To say that there is something better even than ethical activity, and that ethical activity promotes this higher goal, is entirely compatible with everything else that we find in the *Ethics*.

Although Aristotle's principal goal in X.7-8 is to show the superiority of philosophy to politics, he does not deny that a political life is happy. Perfect happiness, he says, consists in contemplation; but he indicates that the life devoted to practical thought and ethical virtue is happy in a secondary way. He thinks of this second-best life as that of a political leader, because he assumes that the person who most fully exercises such qualities as justice and greatness of soul is the man who has the large resources needed to promote the common good of the city. The political life has a major defect, despite the fact that it consists in fully exercising the ethical virtues, because it is a life devoid of philosophical understanding and activity. Were someone to combine both careers, practicing politics at certain times and engaged in philosophical discussion at other times (as Plato's philosopher-kings do), he would lead a life better than that of Aristotle's politician, but worse than that of Aristotle's philosopher.

But his complaint about the political life is not simply that it is devoid of philosophical activity. The points he makes against it reveal drawbacks inherent in ethical and political activity. Perhaps the most telling of these defects is that the life of the political leader is in a certain sense unleisurely (1177b4-15). What Aristotle has in mind when he makes this complaint is that ethical activities are remedial: they are needed when something has gone wrong, or threatens to do so. Courage, for example, is exercised in war, and war remedies an evil; it is not something we should not wish for. Aristotle implies that all other political activities have the same feature, although perhaps to a smaller degree. Corrective justice would provide him with further evidence for his thesis -- but what of justice in the distribution of goods? Perhaps Aristotle would reply that in existing political communities a virtuous person must accommodate himself to the least bad method of distribution, because, human nature being what it is, a certain amount of injustice must be tolerated. As the courageous person cannot be completely satisfied with his courageous action, no matter how much self-mastery it shows, because he is a peace-lover and not a killer, so the just person living in the real world must experience some degree of dissatisfaction with his attempts to give each person his due. The pleasures of exercising the ethical virtues are, in normal circumstances, mixed with pain. Unalloyed pleasure is available to us only when we remove ourselves from the all-too-human world and contemplate the rational order of the cosmos. No human life can consist solely in these pure pleasures; and in certain circumstances one may owe it to one's community to forego a philosophical life and devote oneself to the good of the city. But the paradigms of human happiness are those people who are lucky enough to devote much of their time to the study of a world more orderly than the human world we inhabit.

Although Aristotle argues for the superiority of the philosophical life in X.7-8, he says in X.9, the final chapter of the *Ethics*, that his project is not yet complete, because we can make human beings virtuous, or good even to some small degree, only if we undertake a study of the art of legislation. The final section of the *Ethics* is therefore intended as an prolegomenon to Aristotle's political writings. We must investigate the kinds of political systems exhibited by existing Greek cities, the forces that destroy or preserve cities, and the best sort of political order. Although the study of virtue Aristotle has just completed is meant to be helpful to all human beings who have been brought up well -- even those who have no intention of

pursuing a political career -- it is also designed to serve a larger purpose. Human beings cannot achieve happiness, or even something that approximates happiness, unless they live in communities that foster good habits and provide the basic equipment of a well-lived life.

The study of the human good has therefore led to two conclusions: The best life is not to be found in the practice of politics. But the well being of whole communities depends on the willingness of some to lead a second-best life -- a life devoted to the study and practice of the art of politics, and to the expression of those qualities of thought and passion that exhibit our rational self-mastery.

Glossary

- appearances: *phainomena*
- beautiful: *kalon*
- clear: *saphes*
- complete (verb, also: to perfect): *teleein*
- condition: *hexis*
- continence (literally: mastery): *enkrateia*
- continent: *enkratês*
- disposition: *hexis*
- emotion: *pathos*
- evil: *kakos*, *phaulos*
- excellence: *aretê*
- feeling: *pathos*
- fine: *kalon*
- flourishing: *eudaimonia*
- friendship: *philia*; *philein* (the verb cognate to the noun "*philia*," can sometimes be translated "like" or even "love")
- function: *ergon*
- good will: *eunoia*
- happiness: *eudaimonia*
- happy: *eudaimon*
- impetuosity: *propeteia*
- incontinence (literally: lack of mastery): *akrasia*
- incontinent: *akratês*
- intuitive understanding: *nous*
- live well: *eu zên*
- practical wisdom: *phronêsis*
- science: *epistêmê*
- standard: *horos*
- state: *hexis*
- task: *ergon*
- virtue: *aretê*

- weakness: *astheneia*
- work: *ergon*

Bibliography

A. Translations

1. *Nicomachean Ethics*:

- Crisp, Roger. Cambridge: Cambridge University Press, 2000.
- Irwin, T.H. Indianapolis: Hackett Publishing Co. With Introduction, Notes, and Glossary. Second edn., 1999.
- Ross, W.D., revised by J.O. Urmson. in *The Complete Works of Aristotle*, The Revised Oxford Translation, vol. 2, Jonathan Barnes, ed., Princeton: Princeton University Press, 1984.

2. *Eudemian Ethics*:

- J. Solomon, in *The Complete Works of Aristotle*, The Revised Oxford Translation, vol. 2, Jonathan Barnes, ed., Princeton: Princeton University Press, 1984.
- H. Rackham, in the Loeb Classical Library, Aristotle, vol. 20. Cambridge, Mass.: Harvard University Press, 1952.
- Woods, M.J., *Aristotle's Eudemian Ethics: Books I, II, and VIII*. Second edn. Oxford: Clarendon Press, 1992.

B. Single-Authored Overviews

- Broadie, Sarah. *Ethics with Aristotle*. New York: Oxford University Press, 1991.
- Gauthier, R.A. & J.Y. Jolif. *Aristote: L'Ethique à Nicomaque*. 3 vols. Louvain: Publications Universitaires de Louvain, 1958-9.
- Hardie. W.F.R. *Aristotle's Ethical Theory*, 2nd edn. Oxford: Clarendon Press, 2nd edn. 1980.
- Urmson, J.O. *Aristotle's Ethics*. Oxford: Basil Blackwell, 1987.

C. Anthologies

- Anton, J.P. & A. Preus (eds.). *Aristotle's Ethics: Essays in Ancient Greek Philosophy*, vol. 5. Albany: The State University of New York Press, 1991.
- Barnes, Jonathan, Malcolm Schofield, Richard Sorabji (eds.). *Articles on Aristotle*, vol. 2, *Ethics and Politics*. London; Duckworth, 1977.
- Bartlett, Robert C. & Susan D. Collins (eds.). *Action and Contemplation*. Albany: State University of New York Press, 1999.
- Engstrom, Stephen and Jennifer Whiting. *Aristotle, Kant, and the Stoics*. Cambridge: Cambridge

University Press, 1996.
- Heinaman, Robert (ed.). *Aristotle and Moral Realism*. Boulder: Westview, 1995.
- Roche, Timothy (ed.). *Aristotle's Ethics: The Southern Journal of Philosophy*, Spindel Conference, Supplement 27 (1988).
- Rorty, Amélie O. (ed.). *Essays on Aristotle's Ethics*. Berkeley: University of California Press, 1980.
- Sherman, Nancy (ed.). *Aristotle's Ethics: Critical Essays*. Lanham, Maryland: Rowman & Littlefied, 1999.
- Sim, May (ed.). *The Crossroads of Norm and Nature*. Lanham, Maryland: Rowman & Littlefield, 1995.

D. Studies of Particular Topics

1. The Chronological Order of Aristotle's Ethical Treatises

- Kenny, Anthony. *The Aristotelian Ethics*: *A Study of The Relationship between the Eudemian and Nicomachean Ethics of Aristotle*. Oxford: Clarendon Press, 1978.
- -------. *Aristotle's Theory of the Will*. New Haven: Yale University Press, 1979.
- -------. *Aristotle on the Perfect Life*. Oxford: Clarendon Press, 1992.
- Rowe, C.J. *The Eudemian and Nicomachean Ethics* -- *a Study in the Development of Aristotle's Thought*. Cambridge: Proceedings of the Cambridge Philological Society, suppl. no. 3, 1971.

2. The Methodology and Metaphysics of Ethical Theory

- Barnes, Jonathan. "Aristotle and the methods of ethics." *Revue Internationale de la Philosophie* 34 (1981), pp. 490-511.
- Cooper, John M. *Reason and Emotion*. Princeton: Princeton University Press, 1999. Chapter 12.
- Heinaman, Robert. *Aristotle and Moral Realism*. Boulder: Westview Press, 1995.
- Irwin, T. Aristotle's First Principles. Oxford: Clarendon Press, 1988.
- Kraut, Richard. "Aristotle on Method and Moral Education." In Jyl Gentzler (ed.) *Method in Ancient Philosophy*. Oxford: Oxford University Press, 271-90.
- Nussbaum, Martha C. *The Fragility of Goodness*. Cambridge: Cambridge University Press, 1986. Chapters 8-9.
- -----. "The Discernment of Perception: An Aristotelian Conception of Private and Public Rationality." In Nussbaum, *Love's Knowledge*. New York: Oxford University Press, 1990, pp. 54-105.
- Reeve, C.D.C. *Practices of Reason*. Oxford: Oxford University Press, 1992. Chapter 1.
- Roche, Timothy. "On the Alleged Metaphysical Foundation of Aristotle's Ethics." *Ancient Philosophy* 8 (1988), pp. 49-62.
- -----. "In Defense of an Alternative View of the Foundation of Aristotle's Moral Theory," *Phronesis* 37 (1992), pp. 46-84.

3. The Human Good and the Human Function

- Annas, Julia. *The Morality of Happiness*. New York: Oxford University Press, 1993. Chapter 18.
- Charles, David and Scott Dominic. "Aristotle on Well-Being and Intellectual Contemplation." *Proceedings of the Aristotelian Society*, Supplementary Volume 73, pp. 205-42.
- Clark, Stephen R.L. *Aristotle's Man*. Oxford: Oxford University Press, 1975. Pp. 14-27, 145-63.
- Cooper, John M. *Reason and Human Good in Aristotle*. Indianapolis: Hackett, 1986. Chapters 1, 3.
- -----. *Reason and Emotion*. Princeton: Princeton University Press, 1999. Chapters 9, 13.
- Curzer, Howard J. "The Supremely Happy Life in Aristotle's *Nicomachean Ethics*." *Apeiron* 24 (1991), pp. 47-69.
- Gadamer, Hans-Georg. *The Idea of the Good in Platonic-Aristotelian Philosophy*. New Haven: Yale University Press, 1986.
- Gomez-Lobo, Alfonso. "The *Ergon* Inference." *Phronesis* 34 (1989), pp. 170-84.
- Keyt, David. "Intellectualism in Aristotle." *Paideia* 7 1978, pp. 138-57.
- Korsgaard, Christine. "Aristotle on Function and Virtue." *History of Philosophy Quarterly* 3 (1986), pp. 259-79.
- -----. "Aristotle and Kant on the Source of Value," *Ethics* 96 (1986), pp. 486-505.
- Kraut, Richard. "Two Conceptions of Happiness." *Philosophical Review* 88 (1979), pp. 167-197.
- -----. "The Peculiar Function of Human Beings." *Canadian Journal of Philosophy* 9 (1979), pp. 53-62
- -----. *Aristotle on the Human Good*. Princeton: Princeton University Press, 1989.
- Lawrence, Gavin. "Aristotle and the Ideal Life," *Philosophical Review* 102 (1993), pp. 1-34.
- MacDonald, Scott. "Aristotle and the Homonymy of the Good." *Archiv Für Geschichte der Philosophie* 71 (1989), pp. 150-74.
- Nussbaum, Martha C. *The Fragility of Goodness*. Cambridge: Cambridge University Press, 1986. Chapters 11 and 12.
- Reeve, C.D.C. *Practices of Reason*. Oxford: Oxford University Press, 1992. Chapters 3, 4.
- Roche, Timothy. "*Ergon* and *Eudaimonia* in *Nicomachean Ethics* I: Reconsidering the Intellectualist Interpretation," *Journal of the History of Philosophy* 26 (1988), pp. 175-94.
- Suits, Bernard. "Aristotle on the Function of Man." *Canadian Journal of Philosophy* 4 (1974), pp. 23-40.
- Wedin, Michael. "Aristotle on the Good for Man." *Mind* 90 (1981), pp. 243-62.
- White, Stephen. A. *Sovereign Virtue*. Stanford: Stanford University Press, 1992.
- Whiting, Jennifer. "Human Nature and Intellectualism in Aristotle," *Archiv für Geschichte der Philosophie* 68 (1986), pp. 70-95.
- Whiting, Jennifer. "Aristotle's Function Argument: A Defense." *Ancient Philosophy* 8 (1988), pp. 33-48.
- Williams, Bernard. *Ethics and the Limits of Philosophy*. Cambridge, Mass. Harvard University Press, 1985, Chapter 3.

4. The Nature of Virtue and Accounts of Particular Virtues

- Brunschwig, Jacques. "The Aristotelian Theory of Equity." In Michael Frede & Gisela Striker

(eds.), *Rationality in Greek Thought*. Oxford: Clarendon Press, 1996, pp. 115-155.

- Clark, Stephen R.L. *Aristotle's Man*. Oxford: Oxford University Press, 1975. Pp. 84-97.
- Cooper, Neil. "Aristotle's Crowning Virtue," *Apeiron* 22 (1989), pp. 191-205.
- Curzer, Howard J. "A Great Philosopher's Not So Great Account of Great Virtue: Aristotle's Treatment of 'Greatness of Soul'," *Canadian Journal of Philosophy* 20 (1990), pp. 517-37.
- -----. "Aristotle's Account of the Virtue of Justice." *Apeiron* 28 (1995), pp. 207-38.
- -----. "Aristotle's Account of the Virtue of Temperance in *Nicomachean Ethics* III 10-11," *Journal of the History of Philosophy* 35 (1997), pp. 5-25.
- -----. "A Defense of Aristotle's Doctrine of the Mean," *Ancient Philosophy* 16 (1996), pp. 129-38.
- Gottlieb, Paula. "Aristotle and Protagoras: The Good Human Being as the Measure of Goods," *Apeiron* 24 (1991), pp. 25-45.
- -----. "Aristotle's 'Nameless' Virtues," *Apeiron* 27 (1994), pp. 1-15.
- -----. "Aristotle on Dividing the Soul and Uniting the Virtues," *Phronesis* 39 (1994), pp. 275-90.
- -----. "Aristotle's Ethical Egoism," *Pacific Philosophical Quarterly* 77 (1996), pp. 1-18.
- Hardie, W.F.R. "Magnanimity in Aristotle's Ethics," *Phronesis* 78 (1978), pp. 63-79.
- Hursthouse, Rosalind. "Moral Habituation." *Oxford Studies in Ancient Philosophy* 6 (1988), pp. 201-19.
- Hutchinson, D.S. *The Virtues of Aristotle*. London: Routledge & Kegan Paul, 1986.
- Irwin, T.H. "Disunity in the Aristotelian Virtues." *Oxford Studies in Ancient Philosophy*, Supplementary Volume (1988), pp. 61-78.
- Peterson, Sandra. "*'Horos'* (limit) in Aristotle's *Nicomachean Ethics*," *Phronesis* 33 (1988), pp. 233-50.
- Theodore Scaltsas, "Reciprocal Justice in Aristotle's *Nicomachean Ethics*." *Archiv für Geschichte der Philosophie* 77 (1995), pp. 248-62.
- Schütrumpf, Eckart. "Magnanimity, *Megalopsuchia*, and the System of Aristotle's *Nicomachean Ethics*," *Archiv für Geschichte der Philosophie* 71 (1989), pp. 10-22.
- Sherman, Nancy. *The Fabric of Character*. Oxford: Clarendon Press, 1989.
- -------. *Making a Virtue of Necessity*. Cambridge: Cambridge University Press, 1997.
- Telfer, Elizabeth. "The Unity of Moral Virtues in Aristotle's *Nicomachean Ethics*," *Proceedings of the Aristotelian Society* 91 (1989-90), pp. 35-48.
- Young, Charles. "Aristotle on Temperance" *Philosophical Review* 97 (1988), pp. 521-42.

5. Practical Reasoning, Moral Psychology, and Action

- Broadie, Sarah. "Interpreting Aristotle's Directions." In Jyl Gentzler (ed.) *Method in Ancient Philosophy*. Oxford: Oxford University Press, pp. 291-306.
- Charles, David. *Aristotle's Philosophy of Action*. London: Duckworth, 1984.
- Cooper, John. *Reason and Human Good in Aristotle*. Indianapolis: Hackett, 1986. Chapter 1.
- -----. *Reason and Emotion*. Princeton: Princeton University Press, 1999. Chapters 10, 11, 19.
- Dahl, Norman O., *Practical Reason, Aristotle, and Weakness of Will*. Minneapolis: University of Minnesota Press, 1984.
- Engberg-Pedersen, Troels. *Aristotle's Theory of Moral Insight*. Oxford: Clarendon Press, 1983.
- Fortenbaugh, W.W. *Aristotle on Emotion*. London: Duckworth, 1975.

- Hursthouse, Rosalind. "Acting and Feeling in Character: *Nicomachean Ethics* 3.1," *Phronesis* s29 (1984), pp. 252-66.
- Meyer, Susan Sauvé. *Aristotle on Moral Responsibility*. Oxford: 1993.
- Milo, R.D. *Aristotle on Practical Knowledge and Weakness of Will*. The Hague, 1966.
- Nussbaum, Martha C. *The Fragility of Goodness*. Cambridge: Cambridge University Press, 1986. Chapter 10.
- Reeve, C.D.C. *Practices of Reason*. Oxford: Oxford University Press, 1992. Chapter 2.
- Walsh, James. *Aristotle's Conception of Moral Weakness*. New York: Columbia University Press, 1963.

6. Pleasure

- Gottlieb, Paula. "Aristotle's Measure Doctrine and Pleasure." *Archiv für Geschichte der Philosophie* 75 (1993), pp. 31-46.
- Gosling, J.C.B. & C.C.W. Taylor. *The Greeks on Pleasure*. Oxford, 1982. Chapters 11-17.
- Owen, G.E.L. "Aristotelian Pleasures." In Martha Nussbaum (ed.), *Logic, Science and Dialectic*. Ithaca, N.Y.: Cornell University Press, pp. 334-46.
- Rorty, Amélie O. "The Place of Pleasure in Aristotle's Ethics," *Mind* 83 (1974), PP. 481-93.
- Urmson, J.O. Aristotle on Pleasure." In J.M.E. Moravcsik (ed.), *Aristotle: A Collection of Critical Essays*. Garden City, N.Y.: Anchor Books, 1967, pp. 323-33.

7. Friendship

- Annas, Julia. "Plato and Aristotle on Friendship and Altruism," *Mind* 86 (1977), pp. 532-54.
- -----. *The Morality of Happiness*. New York: Oxford University Press, 1993. Chapter 12
- Cooper, John M. *Reason and Emotion*. Princeton: Princeton University Press, 1999. Chapters 14, 15.
- Kahn, Charles H. "Aristotle and Altruism," *Mind* 90 (1981), pp. 20-40.
- Pakaluk, *Aristotle Nicomachean Ethics Books VIII and IX*. Oxford: Clarendon Press, 1998.
- Price, A. W. *Love and Friendship in Plato and Aristotle*. New York: Oxford University Press, 1989. Chapters xx.
- Schollmeier, Paul. *Other Selves: Aristotle on Personal and Political Friendship*. Albany: State University of New York Press, 1994.
- Stern-Gillet, Suzanne. *Aristotle's Philosophy of Friendship*. Albany: State University of New York Press, 1995.
- Whiting, Jennifer. "Impersonal Friends," *Monist* 75 (1991), pp. 3-29.

8. Feminism and Aristotle

- Freeland, Cynthia, ed. *Feminist Interpretations of Aristotle*. University Park: The Pennsylvania State University Press, 1998.
- Modrak, Deborah. "Aristotle: Women, Deliberation, and Nature." In Bat-Ami Bar On (ed.), *Engendering Origins: Critical Feminist Readings in Plato and Aristotle*." Albany: State University

of New York Press, 1994, pp. 207-22.
- Ward, Julie K., ed. *Feminism and Ancient Philosophy*. New York: Routledge, 1996.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

egoism | ethics: virtue

Richard Kraut
rkraut1@northwestern.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 1, 2001
Content last modified: May 1, 2001

Stanford Encyclopedia of Philosophy
Supplement to Aristotle's Ethics

Alternate Readings of Aristotle on *Akrasia*

That, at any rate, is one way of interpreting Aristotle's statements. But it must be admitted that his remarks are obscure and leave room for alternative readings. It is possible that when he denies that the akratic has knowledge in the strict sense, he is simply insisting on the point that no one should be classified as having practical knowledge unless he actually acts in accordance with it. A practical knower is not someone who merely has knowledge of general premises; he must also have knowledge of particulars, and he must actually draw the conclusion of the syllogism. Perhaps drawing such a conclusion consists in nothing less than performing the action called for by the major and minor premises. Since this is something the akratic does not do, he lacks knowledge; his ignorance is constituted by his error in action. On this reading, there is no basis for attributing to Aristotle the thesis that the kind of *akrasia* he calls weakness is caused by a diminution of intellectual acuity. His explanation of *akrasia* is simply that *pathos* is sometimes a stronger motivational force than full-fledged reason.

This is a difficult reading to defend, however, for Aristotle says that after someone experiences a bout of *akrasia* his ignorance is dissolved and he becomes a knower again (1147b6-7). In context, that appears to be a remark about the form of *akrasia* Aristotle calls weakness rather than impetuosity. If so, he is saying that when an akratic person is subject to two conflicting influences -- full-fledged reason versus the minimal rationality of emotion -- his state of knowledge is somehow temporarily undone but is later restored. Here, knowledge cannot be constituted by the performance of an act, because that is not the sort of thing that can be restored at a later time. What can be restored is one's full recognition or affirmation of the fact that this act has a certain undesirable feature, or that it should not be performed. Aristotle's analysis seems to be that both forms of *akrasia* -- weakness and impetuosity -- share a common structure: in each case, one's full affirmation or grasp of what one should do comes too late. The difference is that in the case of weakness but not impetuosity, the akratic act is preceded by a full-fledged rational cognition of what one should do right now. That recognition is briefly and temporarily diminished by the onset of a less than fully rational affect.

There is one other way in which Aristotle's treatment of *akrasia* is close to the Socratic thesis that what people call *akrasia* is really ignorance. Aristotle holds that if one is in the special mental condition that he calls practical wisdom, then one cannot be, nor will one ever become, an akratic person (1152a6-7). For practical wisdom is present only in those who also possess the ethical virtues, and these qualities require complete emotional mastery. Anger and appetite are fully in harmony with reason, if one is practically wise, and so this intellectual virtue is incompatible with the sort of inner conflict experienced by the akratic person. Furthermore, one is called practically wise not merely on the basis of what one believes or knows, but also on the basis of what one does. Therefore, the sort of knowledge that is lost

and regained during a bout of *akrasia* cannot be called practical wisdom. It is knowledge only in a loose sense. The ordinary person's low-level grasp of what to do is precisely the sort of thing that can lose its acuity and motivating power, because it was never much of an intellectual accomplishment to begin with. That is what Aristotle is getting at when he compares it with the utterances of actors, students, sleepers, drunks, and madmen.

Richard Kraut
rkraut1@northwestern.edu

Return to Aristotle's Ethics

First published: May 1, 2001
Content last modified: May 1, 2001

Saint Augustine

Aurelius Augustinus [more commonly "St. Augustine of Hippo," often simply "Augustine"] (354-430 C.E.): *rhetor*, Christian Neoplatonist, North African Bishop, Doctor of the Roman Catholic Church. One of the decisive developments in the western philosophical tradition was the eventually widespread merging of the Greek philosophical tradition and the Judeo-Christian religious and scriptural traditions. Augustine is one of the main figures through and by whom this merging was accomplished. He is, as well, one of the towering figures of medieval philosophy whose authority and thought came to exert a pervasive and enduring influence well into the modern period (e.g. Descartes and especially Malebranche), and even up to the present day, especially among those sympathetic to the religious tradition which he helped to shape (e.g. Plantinga 1992; Adams 1999). But even for those who do not share this sympathy, there is much in Augustine's thought that is worthy of serious philosophical attention. Augustine is not only one of the major sources whereby classical philosophy in general and Neoplatonism in particular enter into the mainstream of early and subsequent medieval philosophy, but there are significant contributions of his own that emerge from his modification of that Greco-Roman inheritance, e.g., his subtle accounts of belief and authority, his account of knowledge and illumination, his emphasis upon the importance and centrality of the will, and his focus upon a new way of conceptualizing the phenomena of human history, just to cite a few of the more conspicuous examples.

Context

Only four of his seventy-five years were spent outside Northern Africa, and fifty-seven of the remaining seventy-one were in such relatively out of the way places as Thagaste and Hippo Regius, both belonging to Roman provinces, neither notable for either cultural or commercial prominence. However, the few years Augustine spent away from Northern Africa exerted an incalculable influence upon his thought, and his geographical distance from the major intellectual and political capitals of the Later Roman Empire should not obscure the tremendous influence he came to exert even in his own lifetime. Here, as elsewhere, one is confronted by a figure both strikingly liminal and, at times, intriguingly ambivalent. He was, as already noted, a long time resident and, eventually, Bishop in Northern Africa whose thought was transformed and redirected during the four brief years he spent in Rome and Milan, far away from the provincial context where he was born and died and spent almost all of the years in between; he was a man who tells us that he never thought of himself as not being in some sense a Christian [*Confessions* III.iv.8], yet he composed a spiritual autobiography containing one of the most celebrated conversion accounts in all of Christian literature; he was a classically trained rhetorician who used his skills to eloquently proclaim at length the superiority of Christian culture over Greco-Roman culture, and he also served as one of the central figures by whom the latter was transformed and transmitted to the former. Perhaps most striking of all, Augustine bequeathed to the Latin West a voluminous body of work that contains at its chronological extremes two quite dissimilar portraits of the human condition. In the beginning, there is a largely Hellenistic portrait, one that is notable for the optimism that a sufficiently rational and disciplined life can safely escape the ever-threatening circumstantial adversity that seems to surround us. Nearer the end, however, there emerges a considerably grimmer portrait, one that emphasizes the impotence of the unaided human will, and the later Augustine presents a moral landscape populated largely by the *massa damnata* [*De Civitate Dei* XXI.12], the overwhelming majority who are justly predestined to eternal punishment by an omnipotent God, intermingled with a small minority whom God, with unmerited mercy, has predestined to be saved. The sheer quantity of the writing that unites these two extremes, much of which survives, is truly staggering. There are well over 100 titles [listed at Fitzgerald 1999, pp. xxxv-il], many of which are themselves voluminous and composed over lengthy periods of time, not to mention over 200 letters [listed at Fitzgerald 1999, pp. 299-305] and close to 400 sermons [listed at Fitzgerald 1999, pp. 774-789]. It is arguably impossible to construct any moderate sized and manageable list of his major philosophical works that would not occasion some controversy in terms of what is omitted, but surely any list would have to include *Contra Academicos* [*Against the Academicians*, 386-387 C.E.], *De Libero Arbitrio* [*On Free Choice of the Will*, Book I, 387/9 C.E.; Books II & III, circa 391-395 C.E.], *De Magistro* [*On The Teacher*, 389 C.E.], *Confessiones* [*Confessions*, 397-401 C.E.], *De Trinitate* [*On The Trinity*, 399-422 C.E.], *De Genesi ad Litteram* [*On The Literal Meaning of Genesis*, 401-415 C.E.], *De Civitate Dei* [*On The City of God*, 413-427C.E.], and *Retractationes* [*Reconsiderations*, 426-427 C.E.].

Born in 354 C.E. in Thagaste (in what is now Algeria), he was educated in Thagaste, Madauros, and Carthage, and sometime around 370 he began a sixteen-year, monogamous relationship with the mother of his son, Adeodatus (born 372). He subsequently taught rhetoric in Thagaste and Carthage, and in 383 he made the risk-laden journey from Northern Africa to Rome, seeking the better sort of students that was rumored to be there. Disappointed by the moral quality of those students (academically superior to his previous students, they nonetheless had an annoying tendency to disappear without paying their fees), he successfully applied for a professorship of rhetoric in Milan. Augustine's professional ambitions pointed in the direction of an arranged marriage, and this in turn entailed a separation from his long-time

companion and mother of his son. After this separation, however, Augustine abruptly resigned his professorship in 386 claiming ill health, renounced his professional ambitions, and was baptized by Bishop Ambrose of Milan on Easter Sunday, 387, after spending four months at Cassiciacum where he composed his earliest extant works. Shortly thereafter, Augustine began his return to Northern Africa, but not before his mother died at Ostia, a seaport outside Rome, while awaiting the voyage across the Mediterranean. Not too long after this, Augustine, now back in Thagaste, also lost his son (389). The remainder of his years would be spent immersed in the affairs and controversies of the Church into which he had been recently baptized, a Church that henceforth provided for Augustine the crucial nexus of relations that his family and friends had once been. In 391, Augustine was reluctantly ordained as a priest by the congregation of Hippo Regius (a not uncommon practice in Northern Africa), in 395 he was made Bishop, and he died August 430 in Hippo, thirty-five years later, as the Vandals were besieging the gates of the city. However, when Augustine himself recounts his first thirty-two years in his *Confessions*, he makes clear that many of the decisive events of his early life were, to use his own imagery, of a considerably more internal nature than the relatively external facts cited above.

From his own account, he was a precocious and able student, much enamored of the Latin classics, Virgil in particular [*Confessions* I.xiii.20]. However, at age nineteen, he happened upon Cicero's *Hortensius*, now lost except for fragments [see Straume-Zimmermann 1990], and he found himself suddenly imbued with a passion for philosophy [*Confessions* III.iv.7-8]. It is clear from his account of Cicero's effect upon him that his passion was not for philosophy as often understood today, i.e. an academic, largely argument-oriented conceptual discipline, but rather as the paradigmatically Hellenistic pursuit of a wisdom that transcended and blurred the boundaries of what are now viewed as the separate spheres of philosophy, religion, and psychology. In particular, philosophy for Augustine was centered on what is sometimes misleadingly referred to as "the problem of evil." This problem, needless to say, was not the sort of analytic, largely logical problem of theodicy that later came to preoccupy philosophers of religion. For Augustine, the problem was of a more general and visceral sort: it was the concern with the issue of how to make sense of and live within a world that seemed so adversarial and fraught with danger, a world in which so much of what matters most to us is so easily lost [see e.g. *Confessions* IV.x.15]. In this sense, the wisdom that Augustine sought was a common denominator uniting the conflicting views of such Hellenistic philosophical sects as the Epicureans, Stoics, Skeptics, and Neoplatonists (though this is a later title) such as Plotinus and Porphyry, as well as many Christians of varying degrees of orthodoxy, including very unorthodox gnostic sects such as the Manicheans.

Augustine himself comes to spend nine years as a hearer among the Manicheans [see Brown 1967, pp. 46-60], and while there are no extant writings from this period of his life, the Manicheans are clearly the target of many of the writings he would compose after his conversion to the more orthodox, if Neoplatonizing, Christianity he encountered under Bishop Ambrose of Milan. The Manicheans proposed a powerful, if somewhat mythical and philosophically awkward explanation of the problem of evil: there is a perpetual struggle between co-eternal principles of Light and Darkness (good and evil, respectively), and our souls are particles of Light which have become trapped in the Darkness of the physical world. By means of sufficient insight and a sufficiently ascetic life, however, one could eventually, over the course of several lives, come to liberate the Light within from the surrounding Darkness, thus rejoining the larger Light of which the soul is but a fragmented and isolated part.

As Augustine recounts it in the *Confessions* [see *Confessions* V.3.5 and V.7.13] and elsewhere [e.g. *De Moribus Ecclesiae Catholicae* 1], he became disenchanted with the inability of the Manichean elect to provide sufficiently detailed and rigorous explanations of their cosmology. As a result, he began to drift away from the sect during his sojourn in Rome, flirting for awhile with academic skepticism [*Confessions* V.xiv.25] before finally coming upon the Platonizing influence of Ambrose and the "books of the Platonists" [*Confessions* VII.9.13]. When Augustine eventually comes to write about the Manicheans, there are three features upon which he will focus: their implicit materialism (a widespread feature of Hellenistic thought, the Neoplatonists being a notable exception); their substantive dualism whereby Darkness, and hence, evil, is granted a co-eternal, substantial existence opposed to the Light; and their identification of the human soul as a fragmented particle of the Light. According to Augustine, this latter identification not only serves to render the human soul divine, thereby obliterating the crucial distinction between creator and creature, but it also raises doubts about the extent to which the individual human soul can be held responsible for morally bad actions, responsibility instead being attributed to the body in which the soul (itself quasi material) is trapped. Although Augustine is vehement and at times merciless in his repudiation of the Manicheans, questions can still be asked about the influence the Manichean world-view continued to exert upon his understanding and presentation of Neoplatonic and Christian themes [see "Philosophical Anthropology" below].

The single most decisive event, however, in Augustine's philosophical development has to be his encounter with those unnamed books of the Platonists in Milan in 384. While there are other important influences, it was his encounter with the Platonism ambient in Ambrose's Milan that provided the major turning point, reorienting his thought along basic themes that would persist until his death forty-six years later. There has been controversy regarding just which books of the Platonists Augustine encountered [O'Connell 1968, pp. 6-10; O'Donnell 1992, vol. II, pp. 421-423; Beatrice, 1989], but we know from his own account that they were translated by Marius Victorinus [*Confessions* VIII.2.3], and there is widespread agreement that they were texts by Plotinus and Porphyry, although there is again controversy regarding how much influence is to be attributed to each [O'Connell 1968, pp. 20-26; O'Donnell 1992, vol II, pp. 423-4]. These uncertainties notwithstanding, Augustine himself makes it clear that it was his encounter with the the books of the Platonists that made it possible for him to view both the Church and its scriptural tradition as having an intellectually satisfying and, indeed, resourceful content.

As decisive as this encounter was, however, it would be a mistake simply to view Augustine's writings as the uncritical application of a Neoplatonic framework to a static body of Christian doctrine. In his earliest writings [e.g. *Contra Academicos*, 386 C.E.], Augustine is amazingly confident with regard to the compatibility of the two traditions [see *Contra Academicos* 3.10.43]. But by the time he composes the *Confessions* (397-401C.E.), he is already aware that there are significant points of divergence [*Confessions* VII.20.26], and by the time he composes Book VIII of *De Civitate Dei* (circa 416 C.E.), he still has laudatory things to say about the Platonic tradition, but it is clear that the points of divergence have become more important to him and that he regards the Roman Catholic Church as having sufficient internal resources to address whatever difficulties confront it. Part of this gradual change of attitude is attributable to his detailed study of scriptural texts (especially the Pauline letters), as well as his immersion in both the daily affairs of his monastic community and the rather focused sorts of controversies that confronted the Church in the fourth and fifth centuries. Beyond his already noted,

protracted battle with Manicheanism, there is also his involvement in the North African Donatist controversy [see Brown 1967, pp. 212-225], a controversy concerning the validity of sacraments administered in the wake of the persecution of 304-305, and most especially the Pelagian controversy which engaged him from about 411 until his death in 430 [see Brown 1967, pp. 340-52 and the section on "Will" below]. In this latter case, serious issues arose regarding the role of grace and the efficacy of the unaided human will, issues that, as we will see, played an important role in shaping his views on human freedom and predestination.

These important qualifications notwithstanding, the fact remains that this Platonism also provided Augustine with a philosophical framework far more pliable and enduring than he himself is willing to admit in his later works. Moreover, this framework itself forms an important part of the philosophical legacy that Augustine bequeathed to both the medieval and modern periods.

Perhaps a good place to begin an examination of this legacy is to begin with that upon which Augustine himself focuses when he recounts the impact that the books of the Platonists had upon him, i.e. his ontology, and the eudaimonism it is intended to support.

Ontology and Eudaimonism

In the *Confessions*, where Augustine gives his most extensive discussion of the impact of the books of the Platonists, he makes clear that he regards his previous thinking as having been dominated by a common-sense materialism [*Confessons* IV.xv.24; VII.i.1]. It was the books of the Platonists that first made it possible for him to conceive the possibility of a non-physical substance [*Confessions* VII.x.16], providing him with a non-Manichean solution to the problem of the origin of evil. In addition, the books of the Platonists provided him with a metaphysical framework of extraordinary depth and subtlety, a richly textured tableau upon which the human condition can be plotted. It can both account for the obvious difficulties with which life confronts us, while presenting them them within a much larger context that provides grounds for a eudaimonism notable for the depth of its moral optimism. In this respect, the ontology that Augustine acquired from the books of the Platonists is, in terms of its intent, not all that different from the materialism of the Epicureans, Stoics, and even the Manicheans. What sets the Neoplatonic ontology apart, however, is both the resoluteness of its promise and the architectonic grandeur with which it complements the world of visible appearances.

In the books of the Platonists, Augustine encountered an ontology in which there is a fundamental divide between the sensible/physical and the intelligible/spiritual [*Confessions* VII.x.16]. In spite of the dualistic implications, this is clearly not intended to be a dualistic alternative to the moral dualism of the Manicheans and other gnostics [see, e.g. Plotinus, *Enneads* II.9]. Instead, the divide is situated within what is supposed to be a larger, unified hierarchy that begins with absolute unity and progressively unfolds through various stages of increasing plurality and multiplicity, culminating in the lowest realm of isolated and fragmented material objects observed with the senses [see Bussanich 1996, pp. 38-65; O'Meara 1996, pp. 66-81]. Thus, for Augustine, God is regarded as the ultimate source and point of origin for all that comes below. Equated with Being [*Confessions* VII.x.16], Goodness [e.g. *De Trinitate* VIII.5],

and Truth [*Confessions* X.xxiii.33; *De Libero Arbitrio* III.16], God is the unchanging point which unifies all that comes after and below within an abiding and providentially ordained rational hierarchy.

Augustine, especially in his earlier works, focuses upon the contrast between the intelligible and the sensible, enjoining his reader to realize that the former alone holds out what we seek in the latter: the world of the senses is intractably private and isolated, whereas the intelligible realm is truly public and simultaneously open to all [*De Libero Arbitrio* II.7] ; the sensible world is one of transitory objects, whereas the intelligible realm contains abiding realities [*De Libero Arbitrio* II.6]; the sensible world is subject to the consumptive effects of temporality, whereas the intelligible realm is characterized by an atemporal eternity wherein we are safely removed from the eviscerating prospect of losing what and whom we love [*Confessions* XI.xxxix.39; see also *Confessions* IV.xii.18]. Indeed, in the vision at Ostia at *Confessions* IX.x.23-25, Augustine even seems to suggest that the intelligible realm holds out the prospect of fulfilling our desire for the unity that we seek in friendship and love, a unity that can never really be achieved as long as we are immersed in the sensible world and separated by physical bodies subject to inevitable dissolution [see Mendelson 2000]. The intelligible realm, with God as its source, promises the only lasting relief from the anxiety prompted by the transitory nature of the sensible realm.

Despite its dualistic overtones, the overall unity of the picture is central to its ability to provide a resolution of the problem of evil. The sensible world, for example, is not evil, nor is embodiment itself to be regarded as straightforwardly bad. The problem that plagues our condition is not that we are trapped in the visible world (as it is for the Manicheans); rather, it is a more subtle problem of perception and will: we are prone to view things materialistically and hence unaware that the sensible world is but a tiny portion of what is real [*Confessions* IV.xv.24], an error Augustine increasingly attributes to original sin [*De Libero Arbitrio* III.20; *De Civitate Dei* XIII.14-15]. Thus, we have a tendency to focus only upon the sensible, viewing it as a self-contained arena within which all questions of moral concern are to be resolved. Because we fail to perceive the larger unity of which the sensible world is itself a part, it easily becomes for us (though not in itself) a realm of moral danger, one wherein our will attaches itself to transitory objects that cannot but lead to anxiety [*Confessions* VII.xi.17-18]. Given the essentially rational nature of the human soul and the rational nature of the Neoplatonic ontology, there is nonetheless room for optimism. The human soul has the capacity to perceive its own liminal status as a being embodied partly in the sensible world while connected to the intelligible realm, and there is thus the possibility of reorienting one's moral relation to the sensible world, appreciating it for the goodness it manifests, but seeing it as an instrument for directing one's attention to what is above it [see *Confessions* VII.x.16 and VII.xvii.23]. Augustine's employment of this Neoplatonic hierarchy is thus central to his Hellenistic eudaimonism [see O'Connell 1972, pp. 39-40; Rist 1994, pp. 48-53; Kirwan 1999, pp. 183-4] which would redeem appearances by means of situating them within a more primary, if often unacknowledged context.

With respect to questions about specific instances of natural and moral evil, this ontology is even more subtle. Natural evils are attributed to the partiality of our perspective, a perspective that is often the result of our myopic materialism and tendency to focus upon our own self-interest. Understood within the larger context -- both the underlying order of the appearances and the providentially governed moral drama within which they appear -- natural evils are not evil at all [e.g. *Confessions* VII.xiii.19 and *De Civitate*

Dei XI.22]. With respect to the moral evil which is the product of human agency, these are the culpable products of a will that has become attached to lower goods, treating them as if they were higher. Moral evil is, strictly speaking, not a thing, but only the will's turning away from God and attaching itself to inferior goods as if they were higher [ibid.]. In *De Civitate Dei*, Augustine emphasizes the privative nature of evil by referring to the will's pursuit of inferior goods as being a deficient rather than efficient cause [*De Civitate Dei*, XII.7]. The inherent difficulty of this notion aside [see Rist 1994, pp. 106-8], the point behind it is clear enough: Augustine is using the resources of Neoplatonism to account for the phenomena we label evil while stressing human responsibility, thus avoiding either substantializing evil (as the Manicheans do) or making it the result of God's creative activity.

For all that Augustine takes from the books of the Platonists, there are two points where he conspicuously departs from their ontology. Frequently, Plotinus asserts that the ultimate principle, The One, is itself of such absolute unity and transcendence that, strictly speaking, it defies all predication and is itself beyond Being and Goodness [see, for example, Plotinus, *Enneads*, VI.9.3]. Augustine himself does not comment upon this feature of Plotinus' thought, and thus one can only conjecture as to his reason for resisting it, but given his repeated emphasis upon the soul's relation to God [e.g. *Soliloquia* 1.2.7 and *De Ordine* 2.18.47], the Plotinian picture may have seemed to him as positing too great a distance between the two, thus raising doubts about the ability of reason to take us towards our desired destination [see Mendelson 1995, pp. 244-45]. The other departure from Neoplatonism moves in the opposite direction. Rather than the danger of making the spiritual distance between God and the soul too great, there is as well in Neoplatonism a tendency to bridge that gap in a manner troubling to someone like Augustine, for whom the creator/creature distinction is fundamental. In Plotinus and other Neoplatonists, the relation of the ultimate principle to all that comes below is usually presented in terms of a sempiternal process of necessary emanations whereby lower stages constantly flow from the higher [see Plotinus, *Enneads IV.8.6*]. Augustine, not surprisingly, resists this aspect of the Neoplatonic ontology, always insisting upon the fundamentally volitional nature of God's activity [e.g. *De Genesi ad Litteram* 6.15.26]. Nor should it be surprising that Augustine should find himself obliged to depart in important respects from the Neoplatonic tradition. He is, after all, not merely taking over a Neoplatonic ontology, but he is attempting to combine it with a scriptural tradition of a rather different sort, one wherein the divine attributes most prized in the Greek tradition (e.g. necessity, immutability, and atemporal eternity) must somehow be combined with the personal attributes (e.g. will, justice, and historical purpose) of the God of Abraham, Isaac, and Jacob.

For all the changes that affected Augustine between his initial encounter with the books of the Platonists in 384-386 and his death in 430, he never abandoned this Neoplatonic ontology's distinction between the physical/sensible and the spiritual/intelligible and its hierarchy within which these realms are unified. However, these commitments still leave much room for development as well as for tension and uncertainty. In particular, Augustine's views on original sin and the necessity of grace in the face of the Pelagian controversy raised serious questions about the efficacy of the human will. Complicating the matter further is the question of the soul's origin, a question that has a significant impact on Augustine's philosophical anthropology.

Philosophical Anthropology

With respect to Augustine's desire to find a viable alternative to the awkward and intractable moral dualism of the Manicheans, there can be little question that his embracing of Neoplatonism is a positive development. Not only does it allow him to account for evil without substantializing it, but it also provides him with a unified account of the moral drama that constitutes the human condition. Even so, this metaphysical architectonic is prone to tensions of its own, some of which lend themselves to a kind of moral dualism not altogether unlike that of the Manicheans.

For Augustine, the individual human being is a body-soul composite, but in keeping with his Neoplatonism, there is an asymmetry between soul and body. As a spiritual entity, the soul is superior to the body, and it is the province of the soul to rule the body [e.g. *De Animae Quantitate* 13.22; *De Genesi contra Manicheos* II.11]. This presents a fairly positive conception of the soul-body relation, one that clearly runs counter to the Manichean picture of the soul's entrapment. Matters are somewhat less clear, however, when we turn to the question of how the soul comes to be embodied.

With respect to the soul's "origin," as Augustine frames the question, there is a strand of uncertainty that runs unbroken from his earliest completed post-conversion work [*De Beata Vita*, 386 C.E.] to the *Retractationes* of 427 C.E. In both works, Augustine professes to be puzzled about the soul's origin [*De Beata Vita* 1.5 and *Rectractationes* 1.1 and 2.45/71], but his uncertainty is clearly evolving, and the absence of certainty on the issue should not be interpreted as neutrality or indifference.

It is also important to note that, for Augustine, this evolving uncertainty is itself to be understood against the backdrop of other points about which he never seems to waver after 386. He became adamant, for example, that the soul is to be identified with neither the substance of God, nor with the body, nor with any other material entity [*Letters 143* and *166*.3-4]. In addition to the status of the soul as both created and immaterial (both points contrasting with the Manicheans), he also insists upon the mutability of the human soul, a feature that not only serves to distinguish it from its creator but one that he views as necessary to explain the possibility of moral change, be it for better or worse [*Letter 166*.3; *Confessions* IV.xv.26].

In *De Libero Arbitrio* III.20 & 21 (circa 395 C.E.), when Augustine first attends to the question of the soul's origin in a manner that focuses upon particular possibilities, he does so as part of an anti-Manichean theodicy intended to show that it is the human soul rather than God that is responsible for the presence of moral evil in the world. Thus, as he later points out in *Letter 143* (circa 412 C.E.), he is not concerned to adjudicate between these competing hypotheses, but merely to show that each is consistent with a non-Manichean, Neoplatonizing account of moral evil. Nonetheless, the four hypotheses he does advance are important evidence about how he understands the conceptual landscape [O'Daly 1987, pp. 15-20; Mendelson 1998, pp. 30-44], and the anti-Manichean polemic notwithstanding, it is instructive that he makes no attempt to choose between or even to offer a tentative ranking of them.

Interestingly enough, two of the four hypotheses require the soul's existence prior to embodiment. On the

first, the soul is sent by God to administer the body (henceforth the "sent" hypothesis); on the second, the soul comes to inhabit the body by its own choice (henceforth the "voluntarist" hypothesis). In later presentations of these hypotheses (though not in *De Libero Arbitrio* III), Augustine treats the voluntarist hypothesis as involving both a sin on the soul's part and a cyclical process whereby the soul is subject to multiple incarnations [*Letter 166*.27]. The other two hypotheses, the "traducianist" and the "creationist," do not involve pre-existence, but there is nonetheless a significant contrast between them. On the traducianist account, all souls are propagated from Adam's soul in a manner analogous to that of the body, thus linking each soul to all previous ones by a kind of genealogical chain. On the creationist hypothesis, however, God creates a new soul for each body, thus creating a kind of vertical link between God and each individual soul.

These hypotheses do not exhaust the logical possibilities, but they were the main contenders in Augustine's time. There remains controversy over the extent to which Augustine himself was inclined towards either of the hypotheses that required pre-existence [O'Connell 1968, O'Daly 1987, pp. 15-20; O'Donnell 1992 II.34-5], but there are passages in the *Confessions* [see *Confessions* I.6-8] and elsewhere [e.g. *De Genesi Contra Manicheos* 2.8 (circa 388-9 C.E.) and *De Genesi ad Literam Imperfectus Liber* 1.3 (circa 393 C.E.)] that have led some to regard it as a possibility he takes very seriously indeed, perhaps even preferring it, at least until the early part of the fifth century [O'Connell 1968; Teske 1991]. Moreover, given the Neoplatonic architectonic of the *Confessions*, this would not be all that surprising, for the notion that the preexistent soul falls into the body is a conspicuous feature of Plotinus' thought as well as of Neoplatonism in general [e.g. Plotnius, *Enneads* IV.8; Origen, *On First Principles* 1.4.4]. In this regard, it is also not surprising that Augustine should have come to identify the hypothesis of the soul's voluntary descent into the body as involving both sin and cyclicism. Not only are these features reminiscent of what he eventually came to learn of Origen's view, but given the Neoplatonic framework underlying his conception of the soul's origin, it is difficult to construe the soul's choice of embodiment in positive terms.

There is a puzzle at the heart of Augustine's philosophical anthropology, however, that raises serious questions about how we are to construe the human condition. Depending on which of the four hypotheses one were to choose, our condition can be regarded as a divinely ordained exile and trial (the sent hypothesis), the consequence of sin conjoined with an almost immediately self-inflicted punishment (the voluntarist hypothesis), or as some kind of relatively natural habitat (the traducianist and creationist hypotheses). In the latter case, there remain questions about how to construe the soul's creation in relation to God's activity (mediated in traducianism, direct in creationism) as well as about how at home the soul is in the realm of nature.

By the time Augustine comes to write *Letter 166* to Jerome in 415, there have been significant developments in his thinking on this issue. While he does not here sharply distinguish between the two hypotheses involving pre-existence, he is clearly bothered by the cyclicism he has increasingly come to associate with pre-existence, especially as it raises the prospect of a moral landscape wherein pre-incarnate and post-mortem sins are a genuine possibility, for this would entail that that there can be no security even for those who die in a state of grace [*Letter 166*.27]. Moreover, by the time he writes Book 10 of *De Genesi ad Litteram*, (circa 415-16 C.E.) he has a further objection to the notion of pre-incarnate

sin: this possibility, he writes, is ruled out by Romans 9:11 where we are told that the souls of the unborn have done neither good nor evil [*De Genesi ad Litteram* 10.15.27]. Whether or not this poses a decisive objection pre-existence is an obscure matter. In the discussion of *De Genesi ad Litteram* 10, a version of the sent hypothesis does appear as a serious contender, but it is abruptly dropped without explanation, leaving open the question of what lies behind the sudden omission [O'Connell 1987, pp. 227-9; Mendelson 1995, pp. 242-7]. Whatever the reasons may be, the fact is that henceforward, in this text and elsewhere [e.g. *De Anima et eius Origine*, circa 419/20 C.E.], Augustine writes as if there are only two competing hypotheses of the soul's origin, the traducianist and the creationist.

Matters are further complicated by the fact that in *Letter 166* and *De Genesi ad Litteram* [see especially *Letter 166*.27], Augustine makes clear his antipathy to the traducianist hypothesis, an antipathy that, while unexplained, seems to go beyond the materialism in which Tertullian had originally cast it. Creationism, however, hardly offers an unproblematic alternative. Both *Letter 166* and *De Genesi ad Litteram* reveal concern over the question of the acquisition of original sin, an issue that becomes all the more pressing when one considers the plight of the infant who dies unbaptized [*Letter 166*.16 and *De Genesi ad Litteram* 10.11-16]. The Pelagian controversy had by this time brought to the fore the issues of grace and moral autonomy, and Augustine is now adamant in insisting upon the necessity of grace and infant baptism in the face of what he regards as Pelagian challenges to these views. In this context, the case of the infant who dies prior to baptism seems to present the hardest case of all, and the creationist hypothesis, with its direct account of the soul's relation to God's creative activity, seems singularly at a loss to address it. Augustine feels obliged to confirm, contra the Pelagians, the condemnation of the unbaptized infant, but on a creationist reading of the soul's origin, this is hard to reconcile with divine justice, especially given the notion that the unborn have done neither good nor evil. Not surprisingly, the Pelagians themselves favor the creationist hypothesis, for it seems to fit best with their views on the individual's ability to fulfill the moral obligations of the Christian life [TeSelle 1972, pg. 67; Bonner 1972 pp. 23 & 30].

It is thus, again, not surprising that there is an unofficial fifth hypothesis that can be found elsewhere in Augustine's works. In *De Civitate Dei*, for example, Augustine suggests that God created only one soul, that of Adam, and subsequent human souls are not merely genealogical offshoots (as in traducianism) of that original soul, but they are actually identical to Adam's soul prior to assuming their own individual, particularized lives [*De Civitate Dei*, 13.14]. Not only does this avoid the mediation of the traducianist hypothesis, but it also manages to provide a theologically satisfying account of the universality of original sin without falling into the difficulties of God's placing an innocent soul into a sin-laden body, as would be the case in a general creationism. To what extent this constitutes a serious contender for Augustine's attention remains a matter of controversy [O'Connell 1987, esp. pp. 11-16; Rist 1989; Rist 1994, pp 121-9; Teske 1999 pg. 810]. As noted earlier, when Augustine writes of the soul's origin in the *Retractationes* near the end of his life, he still asserts the obscurity and difficulty of the issue, and he is clearly reluctant to take a decisive stand on it. Although he sometimes downplays the seriousness of this uncertainty [e.g. *De Libero Arbitrio* III.21.59 and *De Genesi ad Litteram*, 10.20], there is no getting around the fact that it leaves a significant lacuna at the heart of his philosophical anthropology, one which leaves unanswered crucial questions about how we are to understand the embodied status of the human soul. His Neoplatonic framework commits him to the view that the physical/sensible realm is an arena of temptation and moral danger, one wherein the human soul needs to be wary about becoming too attached to lower goods.

However, Augustine's enduring ambivalence on the the question of the soul leaves open the possibility that the physical/sensible realm is more than an arena of danger and that it is in fact a fundamentally alien context, not altogether different from the Manichean view of embodiment as a kind of entrapment. The ontological unity of the Neoplatonic hierarchy notwithstanding, there appears to be room in it for a moral dualism that may be as troubling in the end as that of the Manicheans.

Psychology and Epistemology

While Augustine remains vague about how we are to understand our embodied status, there is never any question that human life is to be conceived in terms of the categories of body and soul and that an adequate understanding of the soul is necessary for an appreciation of our place within the moral landscape around us. Here Augustine is once again best understood in light of the Greek philosophical tradition [see O'Daly 1987, pp. 11-15], in which "soul" need not have any spiritual connotations. It is, instead, the principle that accounts for the intuitively obvious distinction between things that are living and things that are not. To be alive is to have a soul, and death involves a process leading to the absence of this principle. Thus, not only do human beings have souls, but so do plants and other animals [e.g. *De Libero Arbitrio* I.8; *De Quantitate Animae*, 70; *De Civitate Dei* V.10]. Augustine's view is not unlike what one finds, for example, in Plato's *Timaeus* [e.g. 89d-92c] or Aristotle's *De Anima* [e.g. 414b-415a] where different levels of soul are discussed in terms of ascending degrees of complexity in their capacities, e.g., souls capable only of reproduction and nutrition, or of sensation and locomotion as well, or finally, of rational thinking. As noted in the previous section, there is an asymmetry in these functional capacities, and reason is seen as higher than the others.

As the history of Classical Greek philosophy shows, this schema leaves open a number of possibilities in terms of the relation of soul and body (dualism, hylomorphism, and materialism, to cite some of the more obvious examples), as well as room for disagreement concerning the soul's prospect for continued existence upon the dissolution of the body (Aristotelians tended towards and Epicureans actually embraced a mortalist position, whereas Platonists and Stoics were somewhat more optimistic). For Augustine, however, it is virtually axiomatic that the human soul is both immaterial and immortal. It is worth noting in this connection that while the Christian scriptural tradition clearly alludes to the idea of post-mortem existence, the issue of the soul's immateriality is another matter. It is not obvious that the scriptural tradition requires this, and Tertullian (160-230 C.E.) is a prime example of an early Christian thinker who felt comfortable with a materialist ontology [e.g. Tertullian, *De Anima* 37.6-7]. Thus, while the immortality of the soul is arguably a point of happy convergence of these two traditions, Augustine's emphasis upon the soul's immateriality, an emphasis that comes to have enormous historical importance, seems largely a contribution of his Neoplatonism. As we have seen, he insists upon the soul's mutability as being necessary to account for moral progress and deterioration; however, it is also clear that there must be limits to this mutability, and a material soul would not only run counter to Neoplatonic ontology, but it would also impose upon the soul a degree of vulnerability that would destroy the eudaimonistic promise that made the Neoplatonic ontology so attractive in the first place.

In keeping with the intellectualism of the Greek philosophical tradition, Augustine's psychology focuses

upon the asymmetrical and dominant relation that reason is supposed to exert over other capacities. Unlike post-Humean and post-Freudian views wherein considerable attention is focused upon the role of the non-rational influences that govern our thought, Augustine takes over the ancient Greek confidence in the superiority of the rational over the non-rational. As we will see in the next section, Augustine's views on the will tend to complicate things by qualifying the extent of his intellectualism, but certainly in epistemic contexts his intellectualism tends to hold sway. In this regard, the psychological hierarchy elaborated in *De Libero Arbitrio* II [II.3-II.15] and elsewhere [e.g. *Confessions* VII.x.16 and VII.xvi.21] is a useful illustration of his view.

In the psychology that emerges in *De Libero Arbitrio* II, Augustine posits a three-fold hierarchy of things that merely exist, things that exist and live, and things that exist, live, and possess understanding [*De Libero Arbitrio* II.3]. While he elsewhere allows that plants have souls, his primary interest is in souls capable of understanding, and here, as elsewhere, he is less concerned with a neutral description of the structure of nature than with showing how the soul may find happiness by extricating itself from an overly immersed relation to nature. This being the case, Augustine's psychology tends to focus upon cognitive capacities, beginning with sense perception and working up to reason. The criteria governing the hierarchy are the relative publicity of the object of the cognitive capacity [*De Libero Arbitrio* II.7 & 14], the reliability of the capacity and its object [*De Libero Arbitrio* II.8 & 12], and, corresponding to both of these, the relative degree of immateriality and immutability of the object [*De Libero Arbitrio* II.8 & 14]. Relying upon the criterion of relative publicity, Augustine begins by noting that even among the senses there is a hierarchy of sorts, for vision and hearing seem considerably less private than both smell and taste, wherein part of the object must actually be taken into one's body and consumed during the process [*De Libero Arbitrio* II.7]. Likewise, it seems possible to see or hear the same object at the same time. In between these two extremes is the sense of touch, since two individuals can touch the same part of an object, but not at the same time. Augustine also emphasizes the fact that even in sight and hearing, the most public of the senses, one's relation to the object is always perspectival. For example, one's visual or aural relation to the object imposes limits upon how many others can have a similar relation, as well as the nature of the relation they can have. Thus, sense experience, in addition to relating to objects that are material, mutable, and hence ultimately unreliable, is also intractably private, this latter point being of considerable importance, as we will see, with respect to Augustine's theory of illumination.

The senses are coordinated by what Augustine refers to as the "inner sense" [*De Libero Arbitrio* II.3], a faculty that bears some affinities to Aristotle's common sense [see Aristotle, *De Anima* II.6]. The inner sense for Augustine makes us aware that the disparate information converging upon us from our various senses comes from a common external source (e.g., the smell and taste belong to the same object one is looking at while holding it in one's hand). The inner sense also makes us aware when one of the senses is not functioning properly. In both of these respects, the inner sense bears an organizational and criterial relation to the senses, not only combining the information of the senses, but passing judgment on the results of this synthesis. It is for this reason regarded as being above the other senses [*De Libero Arbitrio* II.5]. At this point, however, we are still at a level shared with non-rational beings. It is only when we go above the inner sense and turn to reason that we reach what is distinctively human.

As with most thinkers influenced by the Greek philosophical tradition, Augustine conceives of reason

rather austerely, focusing upon the mind's ability to engage in deductive reasoning, where logical necessity is the criterion of adequacy. The point is an important one, for it helps explain the belief that reason is distinctively human (intuitively, we may want to attribute instrumental reasoning to other species, but there is still reluctance to attribute mathematical reasoning to them), as well as our tendency to place such enormous significance upon the fact that humans are capable of reasoning. Understood in this austere sense, i.e. in terms of the mind's ability to recognize logical necessity, reason is not merely one instrument among many; instead, it becomes the means whereby the human soul comes into contact with truths that are devoid of the mutability afflicting the objects of the senses. For Augustine, reason is the cognitive apex of the human soul, not only because it distinguishes us from other creatures, but more importantly for the way it distinguishes us: it gives us access to truths that are of an absolutely reliable sort [*De Libero Arbitrio* II.8].

It is also important to note that the necessity revealed by reason is not merely logical and certainly not merely psychological. Augustine, like other thinkers influenced by the Greek tradition, saw an ontological dimension in the truths of reason, i.e., an isomorphism between the necessity that governs our thinking and the necessity that governs the structure of that about which we are thinking. It is at this point that we come upon the intersection of Augustine's psychology and epistemology, for even if we assume a kind of isomorphism between the truths of reason and the structure of being, there is an enduring historical controversy regarding what structure reason reveals as well as how the truths of reason relate to the other cognitive capacities such as sense perception and imagination.

As we have seen, from 384 onwards Augustine accepted a Neoplatonic account of the ontological and moral condition in which we find ourselves. Moreover, the psychology sketched in *De Libero Arbitrio* II and elsewhere reflects an ascending hierarchy of capacities (sense perception, inner sense, and reason), providing a psychological analogue to the ontological hierarchy. Not surprisingly, Augustine's epistemology reflects these strongly Neoplatonic tendencies, but here, as elsewhere, it would be a mistake to view Augustine's thought as an uncritical application of an inherited framework; as is often the case in other areas, Augustine's approach to epistemology is conditioned by his own religious and philosophically eudaimonistic concerns.

In particular, Augustine's epistemology seeks to exploit the psychological hierarchy with the aim of showing the reader how to navigate through the corresponding ontological hierarchy, thereby enabling us to reap the moral benefits of his Christianized Neoplatonism. This point is important, for it helps to explain why Augustine can seem, at times, so overtly indifferent towards questions that are central from the perspective of later (especially post-Cartesian) epistemology. A case in point is Augustine's treatment of Academic skepticism. As already noted, Augustine flirted with Academic skepticism, and one of his first extant works, *Contra Academicos* (circa 386 C.E.) is a focused, if at times idiosyncratic argument against Academic skepticism. Leaving aside Augustine's claim that the Academic skeptics were really Platonic realists attempting to conceal their view from those too simple to grasp its subtlety [e.g. *Contra Academicos*, 3.17.37 and *Letter 1*.1], the overall argumentative thrust of the text is nonetheless instructive [see also Kirwan, 1983].

In the *Contra Academicos*, as elsewhere, Augustine attacks skepticism as an obstacle on the road to a

eudaimonistically-construed happiness. Thus he is content to show that there are problems in the skeptic's claim to live by the likeness of truth (how can one know the likeness of x if one professes not to know x itself?) [*Contra Academicos* 2.7.16-2.8.20], and to offer a set of examples where we do have certainty regarding the truth [*Contra Academicos* 3.10.23 and 3.11.25]. What Augustine does not do is to engage in any kind of foundationalist construction of basic beliefs, nor does he attempt any kind of systematic defense of our ordinary epistemic practices so as to vindicate them in the face of skeptical attack. Even when he offers his version of what later becomes known as the Cartesian *cogito* [e.g. *De Civitate Dei* XI.26; *De Trinitate* 10.14; see also *De Libero Arbitrio* II.3 and Rist 1994, pp. 63-7], he shows no interest in using it to epistemically ground other beliefs [see Markus 1967, pp. 363-4]. Here, as elsewhere, Augustine is content to attack skepticism on a piecemeal basis [see Matthews 1972; O'Daly 1987, pg. 171; and Rist 1994, pg. 53].

Another, related, feature of Augustine's epistemology is his willingness to accept that much of our belief about the world must as a matter of practical necessity rest upon trust and authority. As he tells us in *De Magistro*, we cannot hope to verify all our beliefs about history and even many beliefs about the present are a matter of trust [*De Magistro* 11.37]. Here as elsewhere, he emphasizes the role of belief as opposed to understanding, pointing out not only that we must believe many things that we cannot understand but also that belief is a necessary condition of understanding [see *Contra Academicos* 3.20.43; *De Libero Arbitrio* II.2; and Rist 1994, pp. 56-63]. From a Cartesian foundationalist perspective, this can seem a troublingly circular view. However, we are again obliged to note that Augustine's epistemological concerns do not lie in vindicating our beliefs about the sensible world in the face of skeptical doubt, but in utilizing our non-skeptical intuitions about the sensible world to construct an accessible and rhetorically compelling account of our relation to the intelligible realm, the latter serving as the haven towards which his eudaimonism consistently points. It is worth noting, moreover, that even among those who do not share Augustine's enthusiasm for the transcendental, there are many philosophers in this century who would applaud his indifference towards Cartesian foundationalist concerns. Certainly, his views on the relation of belief, authority, and understanding are worthy of contemporary attention. But for Augustine himself, the primary concern is to lay the groundwork for what many regard as the least compelling if nonetheless most conspicuous element of his epistemology, the doctrine of divine illumination [see Markus 1967, pp. 363-73; Nash 1969; O'Daly 1987, pp. 199-207; and Rist 1994, pp. 73-9].

Augustine presents our grasp of the sensible world as grounded in a relatively unproblematic relation of direct acquaintance [e.g *De Magistro* 12.39. See also Burnyeat 1987], although there are places where his view is complicated by his Neoplatonizing conviction that the higher (e.g. the mind) cannot be affected by the lower (e.g. the body) [e.g. *De Genesi ad Litteram* XII.16 circa 415 C.E.]. In fact, he will in places explicate the mind's relation to sensible objects by means of its focusing its attention and noticing what is presented to it by the body without being causally affected by the body; in the case of physical vision, he will even go so far as to adopt the extramissionist view that a visual ray extends from the eye to the object as opposed to an intromissionist view whereby the eye passively receives something from the sensible object [e.g. *De Quantitate Animae* 23.43, circa 388 C.E.]. Even so, direct acquaintance is at some level still a necessary condition for the formation of beliefs about the external world, and the relation of the senses to sensible objects is regarded as largely unproblematic. In *De Magistro*, for example, Augustine argues that the efficacy of language is ultimately dependent upon direct acquaintance with the external

world, and even our ability to learn from others presupposes that what they tell us can be reduced to elements with which one has had some prior acquaintance [*De Magistro* 11.37]. For Augustine, as for many classical thinkers, language is a kind of third realm entity. Belonging neither to the world nor to mind, it is an instrument used by minds to communicate about the world outside them, and direct acquaintance is what explains its ability to do so. Thus, learning from others is a matter of being reminded of prior acts with which we have been directly acquainted [*De Magistro* 11.36], although this reminding can occur in such a way as to reconfigure elements from those prior acts, thus accounting for the fact that our knowledge of the world seems to be extended by such descriptions.

However odd such a model might seem, it is important to note the plausibility of some of the assumptions that underlie it: (a) language is an instrument that mediates our relation to the world and to other minds; (b) there is a distinction between signs and what they signify; and (c) our relation to the sensible world is based on direct experience. Each of these assumptions is subject to serious objections, and the past two centuries have produced ample reasons to be cautious about them. Nevertheless, they still have considerable pre-reflective currency, and for all its oddness, Augustine's suggestion that learning is a matter of being reminded of prior acts of direct acquaintance rests upon a set of common sense assumptions. This in itself is an important point, for as noted above, much of Augustine's strategy in presenting his epistemology is to exploit the relatively unproblematic nature of our relation to the sensible world, and then to reason analogously regarding our relation to the more secure, public world of intelligible objects. The question we are supposed to ponder is: given that learning is really a matter of being reminded, and given that all such occasions of being reminded depend upon acts of direct acquaintance wherein we are taught by the things themselves [*De Magistro* 12.40], what does this imply about our relation to those truths that cannot be accounted for by sense perception? In other words, if we accept this as a viable model of our epistemic relation to the external world, how do we proceed from it to explain our access to those truths whose certainty goes beyond what can be experienced in sensible objects? The traditional example here is mathematics [e.g. *De Libero Arbitrio* II.8], and in *De Libero Arbitrio* II, Augustine even argues that our ability to count presupposes a notion of unity that is empirically underdetermined [ibid]. There are, of course, other examples for Augustine besides mathematical and logical truths. Of equal importance are such truths as the awareness that all seek a happiness that goes beyond anything we have experienced in this life, that good is to be sought and evil avoided, and the awareness that there is something above and more reliable than the human mind [see *De Libero Arbitrio* II.9 and 12]. These are the kinds of examples that Augustine regards as obliging us to reject the notion that our relation to the sensible world is sufficient to account for all our beliefs and to believe that there must be more, so tospeak, to complete the picture.

That something more is provided by the doctrine of illumination, the thesis that God plays an active role in human cognition by somehow illuminating the individual's mind so that it can perceive the intelligible realities which God simultaneously presents to it. Augustine is notoriously vague as to the precise details and mechanics of this divine illumination [see, e.g. Nash 1969, pp. 94-124], and it is therefore easy to read it in an uncharitable light. Viewed without sufficient attention to the few details he provides, it can appear as if Augustine has made human cognition into a special act of divine revelation, thus making the human mind into a merely passive receptacle and God into a kind of epistemic puppeteer. For all its attendant vagueness, however, the doctrine is rather more sophisticated than it might first appear.

In the account of illumination in *De Magistro*, Augustine uses an analogy as old as Plato [see *Republic* VI.508a ff.] according to which the mind's relation to intelligible objects is like the relation of the senses to sensible objects [see *De Magistro* 12.39; see also *Soliloquia* 1.12 and O'Daly 1987, pg. 204]. In both cases, there is a need for an adventitious object to be presented to the relevant capacity, as well as the need for an environment that is conducive to the successful exercise of the relevant capacity. In the case of vision, for example, this would be light; in the case of the mind's discernment of intelligible objects, Augustine characterizes this, relying upon Platonic imagery of which Plotinus is also fond [see Plotinus, *Enneads* V.3.8 and Schroeder 1996, pp. 341-3], as an intellectual illumination that occurs within us by that which is above us. In both cases, the criterion of success is the discernment of the actual details of the object itself. Perhaps most important of all, both cases clearly allow for and rely upon acts of direct acquaintance, since illumination is, above all, meant to be an account of the conditions necessary for the mind to have direct acquaintance with intelligible objects.

Seen in this light, Augustine's view hardly seems to reduce human cognition to special acts of divine revelation [see O'Daly 1987, pp. 206-7]. Illumination is instead something that is available to all rational minds, the atheistic mathematician as well as the pious farmer measuring a field [see Rist 1994, pg. 77]. Nor does it detract from the mind's own activity and acuity, any more than a world of adventitious sensible objects detracts from the activity and acuity of the senses. In both sensory and intellectual perception, one can require a considerable degree of activity and acuity on the part of the perceiver, and in both cases one can treat failed perception as a function both of the extent to which the capacity is possessed by the perceiver and the perceiver's efforts to employ it. What sets illumination apart from more familiar cases of sense perception is that it enables us to do two related things that cannot be done by sense perception alone. First and foremost, it explains how our knowledge can have the kind of necessity that understanding (as opposed to mere belief) requires, a necessity that is always, it seems, empirically underdetermined [see, e.g. *De Libero Arbitrio* II.8 and O'Daly 1987, pp. 180-1]. In this regard, Augustine's illuminationism is a worthy contender among more familiar attempts to make intellectual cognition epistemically secure and reliable. Though it has its own difficulties, it is not clear that Augustinian illumination is all that more extravagant than Platonic recollection of a pre-incarnate existence [e.g. Plotinus, *Enneads* V.5], Aristotelian induction of particulars that somehow leads to necessary and universal truths [e.g. Aristotle, *Posterior Analytics* II.19], psychologically private Cartesian innate ideas [*Meditations*, "Third Meditation"], or Kantian transcendental idealism, wherein we are obliged to sacrifice the isomorophism of reality and thought that made necessity so attractive in the first place [e.g. *Critique of Pure Reason*, "Preface" to the First and Second Editions]. Indeed, viewed in this regard, it is not all that surprising that Augustinian illuminationism came to have the historical influence that it did, nor that Malebranche, writing some twelve hundred years later, would, in his concern with the psychologistic implications of Cartesian innate ideas, turn to Augustinian illuminationism as a model for his vision in God [see, e.g. *The Search After Truth*, Bk. II, Part Two, Chapter Six].

The second way in which illumination enables us to surpass what we are able to accomplish by means of sense perception alone is even more tightly connected to Augustine's Neoplatonizing eudaimonism. For souls which have become immersed in the sensible world and which are thereby separated from other souls by bodies, illumination is crucial to our attempt to recapture our lost unity. Unlike the perspectival and private realm of sense perception, illumination holds out the prospect of fulfilling the yearning to

which Augustine's eudaimonism gives such prominence, the yearning to find a realm wherein we can overcome the vulnerability that besets us and the moral distance that divides us from one another. Both Augustine's *Confessions* and *De Civitate Dei* in their own ways portray this sort of philosophical and spiritual pilgrimage, and one would be hard pressed to find a better example than the vision at Ostia at *Confessions* IX.10.23-25 [see "Ontology and Eudaimonism" above]. There, Augustine and his mother Monica manage, albeit fleetingly, to find themselves in a place that is clearly not in space, united in a way that overcomes the distance imposed by their mortal bodies. This unification is for Augustine the eudaimonistic conclusion through which the pursuit of knowledge is vindicated and to which it is, ultimately, to be subordinated.

Will

As already noted, a conspicuous feature of the Greek philosophical tradition is its intellectualism. Not only is nature seen as governed by patterns that are accessible to the human mind, but human agency is conceived in terms that stress the role played by reason in a life that is in keeping with the larger order [see Markus 1967 pg. 387]. Reason is an instrument that is not only capable of acts of theoretical representation, but its exercise is also regarded as being of enormous practical significance. There are, to be sure, important and powerful non-rational factors that are relevant to our actions (e.g. appetite and desire), but in a well-ordered life they are to be constrained by the dictates of reason [see e.g. Plato, *Republic* IV.441e-4441 and Aristotle, *Nicomachean Ethics* X.7.1177a10-X.9.1179a33].

As we have seen above [e.g."Ontology and Eudaimonism" & "Psychology and Epistemology"], Augustine is deeply affected by Greek intellectualism, and his own Neoplatonizing Christianity is imbued with a hierarchical structure that emphasizes the reliability of the intelligible in contrast to all that is sensible and physical. However, as Augustine's views on human agency develop, this picture is complicated by an increasing emphasis upon non-rational factors that influence our behavior and by a tendency to regard intellectualism as insufficient to explain the dynamics of human agency. Early in Augustine's career [e.g. *De Libero Arbitrio* I, circa 387/8 C.E.], there is a conspicuous emphasis on the will, and it is here that one encounters some of the most difficult and obscure aspects of his thought [see Djuth 1999, pg. 881]. Nevertheless, it marks both a significant divergence from the Greek philosophical tradition and the intersection of the philosophical and religious dimensions of his thought. Moreover, the more Augustine immersed himself in theological questions, the more prominence the nature and role of the will came to have in his writings, and his reflection upon the limited powers of the unaided will has much to do with the pessimism of his later writings.

An example of Augustine's increasing emphasis upon the will can be found in his account of his intellectual and moral transformation in *Confessions* VII-VIII. As we have seen ["Context" and "Ontology and Eudaimonism"], he credits the books of the Platonists with making it possible for him to conceive of a non-physical, spiritual reality [*Confessions* IV.xv.24; VII.i.1]. Likewise, they removed the intellectual stumbling blocks that had made it so difficult for him to accept the non-Manichean form of Christianity he found in Ambrose's Milan. However, when Augustine tells the story of his conversion in *Confessions* VII and VIII, he makes clear that although he ceased to have any genuine intellectual reservations

regarding the Church [*Confessions* VII.xxi.27 and VIII.i.1], he remained unable to commit himself to the path he could see to be the right one [see *Confessions* VII.xx.26, VII.xxi.27, and VIII.i.1]. Throughout his discussion, Augustine indicates that certainty is not the issue; he regards his predicament as falling outside the scope of intellectual assent. The ensuing discussion of his struggle is surely one of the most famous in Christian literature [*Confessions* VIII in toto, esp. VIII.viii.19-VIII.xii.30], and it is marked by a subtlety of introspective analysis that defies any easy explication. Leaving aside the question of the accuracy of his account [O'Connell 1969, pp. 4-9 and 101-104; O'Donnell 1992, vol. 3, pp. 3-4 and 55-71], it is clear that Augustine is providing a dramatic account of moral transformation, one that stresses the role of intellectual discernment while at the same time highlighting his conviction that no amount of discernment is sufficient to account for what we might refer to, for want of a better phrase, as the phenomenology of internal moral conflict. In terms of this agonistic inner turmoil, the will as both present and emergent [*Confessions* VIII.v.11 and VIII.x.22] is on an equal footing with our powers of rational discernment.

There are three distinct features that explain why the will comes to have such prominence in Augustine's thinking. In Book I of *De Libero Arbitrio*, Augustine endeavors to construct an anti-Manichean theodicy [*De Libero Arbitrio* I.2], one that accounts for the presence of moral evil in the world without either substantializing it or finding its source in divine activity. In this regard, the will is what makes an action one's own, placing the burden of responsibility on the one performing the action [*De Libero Arbitrio* I.11]. By the time he composed Book III of *De Libero Arbitrio*, however, Augustine had come to conceive of the human condition in terms of the ignorance and difficulty that attend it [*De Libero Arbitrio* III.18], and these features tend to complicate the libertarian optimism of Book I by raising questions about whether it is even possible for us to overcome the ignorance and difficulty. But even here, the will is intended to serve as the fulcrum of moral responsibility [e.g. *De Libero Arbitrio* III.22].

Though closely related, the concern with moral responsibility needs to be distinguished from the points raised in the above discussion of *Confessions* VII-VIII. In that context, Augustine is still engaged in constructing an anti-Manichean portrait of the human condition, but he is equally concerned with the aspect of agency that falls outside the scope of a purely rational or intellectual analysis. This aspect of the discussion is heightened by the fact that the choice involves a fundamental moral reorientation running contrary to habits which have acquired a necessity all their own [*Confessions* VIII.v.10], but Augustine's discussion of the example suggests that he sees it as more than an idiosyncratic or isolated incident. Rather, it is intended to draw our attention to an introspectively accessible range of phenomena that forces us to acknowledge a fundamentally non-rational component of human volition.

There is, however, a third factor at work here. The problem of evil received a rather different treatment in the non-Hellenic religious and scriptural traditions than in the Greek tradition, a contrast that was not completely lost on Augustine as he increased his familiarity with the former [e.g. *Ad Simplicianum*, circa 396 C.E. and *Confessions* VII.ix.14]. Here, one finds less emphasis upon rational analysis and logical argumentation than upon pledged community membership, trans-generational authority, obedience to divinely-sanctioned standards, and, in some cases, an overt suspicion of intellectualism together with an emphasis upon the necessity of divine aid for moral transformation. This part of Augustine's inheritance helped to divert his attention away from the strictly rational features of human agency, and to invite him to think about rationality in new ways.

While it is no doubt a mistake to compartmentalize the religious and philosophical aspects of Augustine's classical inheritance, it is often helpful to view his thought as presenting a gradual movement away from a Greek intellectualism towards a voluntarism emphasizing the profound ignorance and difficulty of the human condition, as well as the need for divine aid to overcome the ignorance and difficulty. At the heart of this shift of emphasis are Augustine's developing views on the will. Not surprisingly, this development often has to be understood against the backdrop of the philosophical and theological difficulties that come to occupy him over the years.

One of these difficulties is the relation of human free will to divine foreknowledge. While it is tempting to view this as a conflict between Athens and Jerusalem, the problem initially arises within the Greco-Roman tradition itself [see Rist 1994, pg. 268]. Although Augustine's initial treatment of the problem at *De Libero Arbitrio* III.2-4 seems innocent of this fact, his later treatment at *De Civitate Dei* V.9-10 shows that he was aware of Cicero's discussion of the problem in *De Divinatione* and *De Fato*. It is also worth noting that in later medieval philosophy, we see the mirror-image of this problem in terms of the relation of divine freedom and power versus the extent of human knowledge [see, e.g. The Condemnation of 1277; Henry of Ghent, *Quodlibet* VIII, qu.9; John Duns Scotus, *Ordinatio* I, dist. 42]. In both cases, the problem is attributable to the notion of necessity which underlies the Greek conception of knowledge. In this particular case, the problem is how to reconcile the absolute necessity that attends God's knowledge (i.e. if God genuinely knows that x is going to happen, it is impossible for x not to take place --see *De Libero Arbitrio* III.4 and *De Civitate Dei* V.9) with the idea that there can be no moral responsibility unless it is in my power to choose to do other than I in fact do [e.g. *De Libero Arbitrio* III.3]. On the surface, freedom to do otherwise seems to rule out the possibility of foreknowledge, and conversely, foreknowledge seems to rule out the possibility of freedom to do otherwise. In both *De Libero Arbitrio* and *De Civitate Dei*, Augustine's treatment of this problem is complex and at times exceedingly obscure [see Rowe 1964 and Kirwan 1989,pp. 95-103], but his aim is clear enough. Augustine is anxious, contra the Manicheans and Cicero, to defend the compatibility of divine foreknowledge and human freedom by arguing that the free exercise of the will is among the events foreknown by God and that such foreknowledge in no way detracts from our culpability for our acts of willing [e.g. *De Libero Arbitrio* III.3 & 4; *De Civitate Dei* V.9]. The obscurity of the details notwithstanding, Augustine leaves no doubt that he wants to maintain both that God does have foreknowledge of our actions and that we are morally responsible for them.

Augustine's view becomes even more complicated, however, due to theological and doctrinal concerns. While the issue of predestination is not invoked in the discussion of divine foreknowledge and human freedom at *De Civitate Dei* V.9-10 [see Rist 1994, pp. 268-9], significant developments take place between the time Augustine composes *De Libero Arbitrio* III (circa 395 C.E.) and *De Civitate Dei* V (circa 415 C.E.). In particular, there are two events that have a momentous impact upon Augustine's work in the late 390's until his death in 430. The first is his increasing familiarity with scripture and the resulting modification of his earlier, Neoplatonizing views in light of what he finds in those texts. Pivotal in this regard is *Ad Simplicianum* (396 C.E.), wherein he focuses on a number of scriptural passages and begins to formulate his views on the universality of original sin and the necessity of grace to overcome its effects [see Bonner 1972, pp. 15-18 and Babcock 1979, pp. 65-67]. The second set of events center on his

involvement in the Pelagian controversy, which occupied him from roughly 411 until his death in 430. Under the pressures of this controversy and in conjunction with his interpretation of scriptural and especially Pauline views on original sin and grace, the intellectualistic optimism of his earlier work was gradually transformed into an exceedingly grim view of the human moral landscape.

Pelagius himself is an obscure figure, as is his relation to the view that has come to bear his name (Bonner 1972, 31-35), but at the heart of the Pelagian position seems to be an emphatic insistence upon the principle that "ought implies can," i.e. that it is unacceptable to require individuals to perform actions that they cannot in fact perform [Pelagius, *Ad Demetriadem* 2, op. cit. at Brown 1967, pg. 342; see also Bonner 1972, pg. 34]. The Pelagian insistence upon preserving the kind of autonomy that seems required by the moral ideals of Christianity set in motion a fierce controversy about the nature of original sin and the role of grace in overcoming it [Brown 1967, pp.340-364]. In general, Pelagians tended to deny the kind of insuperable original sin that Augustine believed he had found in scripture, and they proposed a milder view of grace as being an aid to a will disposed to a Christian life, as opposed to being a necessary condition for such a disposition in the first place [TeSelle 1999, pg. 635]. As is often the case with disputes that have a deep moral urgency, the controversy acquired a ferocity that can seem, from a modern perspective, out of keeping with the subtlety of the points made in it, but it is precisely the sort of dispute that cannot but have lasting effects upon its participants, and Augustine was one of the main participants during the last two decades of his life.

By the time Augustine completed *De Civitate Dei* in 427 C.E., he came even more emphatically to insist upon the conclusion to which his discussion in *Ad Simplicianum* had led him, i.e., that original sin is both universally debilitating and insuperable without the aid of unmerited grace [*De Civitate Dei* XIV.1]. Furthermore, there is a predestination at work that is as rigorous as the foreknowledge by which God knows its results [*De Civitate Dei* XIV.11]. Here too Augustine insists that we are morally culpable for the sinful choices that the will makes [*De Civitate Dei* XIV.3], but under the pressures of the Pelagian controversy -- a controversy in which he will find his earlier words being cited against him [see *Retractationes* I.9.3-6] -- he presents these views in a manner that is austere and uncompromising. So damaging are the effects of the original sin that the human will is free only to sin [*De Correptione et Gratia* 1.2; 11.31; Rist 1972, pg. 223]. Thus, the human race is comprised of a *massa damnata* [*De Dono Perseverantiae* 35; see also *De Civitate Dei* XXI.12], out of which God, in a manner inscrutable to us [*De Civitate Dei* XII.28], has predestined a small number to be saved [*De Civitate Dei* XXI.12], and to whom he has extended a grace without which it is impossible for the will not to sin. While there is some controversy over whether this grace is sufficient for redemption and whether it can be resisted [Rist, 1972, pp. 228ff.], Augustine makes clear that it is as much a necessary condition as it is unmerited and inscrutable. The ignorance and difficulty that afflict our condition in *De Libero Arbitrio* III have become more than obstacles to be overcome by means of our will [*De Libero Arbitrio* III.22]; they are now impassible barriers we have inherited from Adam, and without unmerited grace we are utterly incapable of initiating even the smallest movement away from sin and towards God. In *De Libero Arbitrio* I, Augustine suggests that the will is confronted by a rational choice between a life spent in the pursuit of what is temporal, changing, and perishable, and a life spent in the pursuit of what is eternal, immutable, and incapable of being lost [*De Libero Arbitrio* I.7]. By the time he comes to write *De Gratia et Libero Arbitrio* in 426 C.E., in the midst of the Pelagian controversy, we find a vastly different picture. Here too

the will is central, and here too we are culpable for our sins, but gone is the earlier optimism. The post-Adamic will is no longer in a position to initiate any choice of lives; the fact that we have any choice at all is entirely a product of unmerited grace [see, e.g. *De Gratia et Libero Arbitrio* xx and xxi], a grace that will be given to only a small number whom God has predestined to be saved out of the vast number who are eternally lost.

Being more a matter of theology than philosophy, it can be tempting for those interested in Augustine as a philosopher to turn away from his later thinking on the will, but one has to be careful in doing so. To begin with, the boundary between the philosophical and the theological is not as clear in Augustine as it is in later philosophers, and part of what makes Augustine such a fascinating thinker is his refusal to compartmentalize his thought in ways that are now taken for granted. Second, the development of Augustine's thinking on the will, as unsettling as the resulting moral landscape may be, does oblige one to confront questions about what a viable concept of the will should involve as well as questions about how to determine moral culpability in the face of external determination -- questions that are as easy to overlook as they are difficult to address. Finally, Augustine's reflections on the will had considerable influence upon those who inherited his vast legacy and on his own account of how we are to understand the drama of human history.

History and Eschatology

It is an irony that the man who bequeathed a Neoplatonic world view to the West also gave us a way of conceptualizing human history that is at odds with some of its most basic contours. In the Greco-Roman world in general and in Neoplatonism in particular, the importance of history is largely in the cyclical patterns that forge the past, present, and future into a continuous whole, emphasizing what is repeated and common over what is idiosyncratic and unique. In Augustine, we find a conception of human history that in effect reverses this schema by providing a linear account which presents history as the dramatic unfolding of a morally decisive set of non-repeatable events.

For the present day reader, it is easy to overlook both the plausibility of the cyclical view and the sorts of considerations that might stand in the way of the linear model with which we have become more familiar. Not only are there the obvious patterns of the seasons and the regularities discernible in astronomical phenomena, but, at a deeper level, there is the indispensable role that regularity and the recognition of common features play in our efforts to make the world intelligible. Moreover, the emphasis upon the common-qua-universal is a conspicuous feature of the Greek philosophical tradition. Thus, it is also hardly surprising that we find Aristotle telling us that poetry is more philosophical than history because it is more clearly concerned with universals, whereas history tends to be more concerned with particulars [Aristotle, *Poetics* 9.1451b1-7]; nor is it surprising that Thucydides presents his account of the Peloponnesian War as providing a pattern of events that will be repeated in the future [Thucydides, *History of the Peloponnesian War*, I.22]; or that Plutarch recounts past lives in a manner clearly designed to draw the reader's attention to patterns of virtue and vice rather than to faithfully recount particular facts [see, e.g. Plutarch, *Life of Pericles* 1.1-2]; or, for that matter, that Augustine himself would tell the tale of his first thirty-two years in the way that he does, more concerned to capture the Neoplatonic drama of the

soul's immersion and extraction from the sensible/physical world than with providing a factual account of dates, names, and places.

Approached from this angle, what wants an explanation is why one would subordinate indispensable patterns and regularities in order to emphasize what is idiosyncratic and unique . Here, as in the case of the will, it is important to understand that Augustine is bringing together two quite disparate traditions, and here again one needs to take note of his efforts to capture the data of revelation he sees embedded in Judeo-Christian scripture. If one approaches these latter texts as presenting a Christian drama of the soul's salvation, one cannot help but focus upon the unique, non-repeatable events that define the drama, e.g., the fall recounted in the early chapters of Genesis, the incarnation, passion, and resurrection of Christ in the synoptic and Johannine gospels, and the final judgement foretold in Revelations. One must, however, exercise some caution here. The cyclical and linear approaches are matters of emphasis rather than mutually exclusive alternatives, and the scriptural traditions upon which Augustine relies are certainly not devoid of cyclical motifs [e.g. Ecclesiastes 3.1-8], nor does Augustine himself embrace one approach wholly to the exclusion of the other, as even a cursory reading of his *Confessions* reveals. And, of course, the historically unique life of Christ becomes a pattern for the Christian life in general [e.g. *De Civitate Dei* XXII.5]. These points notwithstanding, there can be little question that Augustine provides an account of human history that is at times resolutely linear, a tendency which can be traced to the Judeo-Christian scriptural tradition.

Already in *De Magistro* (389 C.E.) Augustine is keenly aware that much of what we need to believe falls outside the austere standards of his Platonic conception of knowledge and understanding. Among the most prominent of these are beliefs based on scripture [*De Magistro* 11.37; cf.12.39]. In the *Confessions* as well, even when Augustine is especially laudatory of the Platonists, he is emphatic that there is much that these books leave out. They cannot, for example, speak about those historical truths definitive to the Christian view of redemption through the incarnation and passion of Christ [*Confessions* VII.ix.13-14; see Bittner 1999, pg. 346]. Augustine is acutely aware that scripture has an historical dimension, and he is sensitive as well to the tensions between the scriptural tradition and the Neoplatonic framework upon which he is relying, a tension that comes to eclipse much of the intellectualistic optimism we find in his earliest completed post-conversion works, e.g. the *Contra Academicos* of 386 C.E. [see *Contra Academicos* 3.20.43 and "Context" above].

As we have seen, Augustine's increasing familiarity with the contents of scripture leads him to focus more and more upon the historical dimension of this tradition, a dimension alien to the intellectualism of the books of the Platonists. We have already seen this development reflected in his interest in the fall and the subsequent necessity of grace set forth in the *Ad Simplicianum* of 396 C.E. But it is in Augustine's sprawling *City of God* [*De Civitate Dei*, 413-427 C.E.] that one finds his most extensive and focused treatment of human history [see Rist 1994, pp. 203-255]. It is important to bear in mind, however, that Augustine does not provide a philosophy of history of the sort that one might find in a Vico, Hegel, or Marx; his concern is not with articulating a notion of history that views its progress as intelligible, or that sees it as developing according to immanent processes that are themselves accessible and worthy of study. Human history, for Augustine, is subsumed by the larger context of an eschatology wherein history is the temporal playing out of a divine justice in which the end is as fixed as the beginning [see Bittner 1999, pg.

348]. While it is not for us to know all the details of the plot or its conclusion [*De Civitate Dei* XX.2], we can nonetheless discern the general direction of the drama, as well as the juridical nature of the conclusion at which aims.

The drama is, for the most part, a hauntingly somber one. Due to the universal contagion of original sin wherein all have sinned in Adam, humanity has become a mass of the deservedly damned [*De Civitate Dei* XXI.12] who have turned away from God and towards the rule of self [see *De Civitate Dei* XIII.14; XIV.3 & 13]. By means of an utterly unmerited grace, God has chosen a small minority out of this mass -- the smallness of the number is itself a means whereby God makes apparent what all in fact deserve [*De Civitate Dei* XXI.12] -- and thus human history is composed of the progress of two cities, the city of God and the city of Man [e.g. *De Civitate Dei* XIV.28; XV.1 & 21; see Cranz 1972]: those who by means of grace renounce the self and turn towards God, as opposed to the vast majority who have renounced God and turned towards the self [*De Civitate Dei* XIV.28]. In this life, we can never be sure of which individuals belong to which city [e.g. *De Civitate Dei* XX.27], and thus they are intermingled in a way that thwarts any moral complacency. While the visible church bears a special relation to the city of God, membership in the Church is no guarantee of salvation [e.g. *De Civitate Dei* XX.9], and the history that is visible to us is merely a vestige of the moral drama that takes place behind the scenes, defying the scrutiny of our weak and often presumptuous reason [*De Civitate Dei* XX.21 & 22]. Whatis certain is that the linear movement of human history aims at the eventual separation of the two cities [e.g. *De Civitate Dei* XX.21 & 28], in which the members of each city are united with their resurrected bodies [e.g *De Civitate Dei* XXI.1 & 3 and XXII.21] and given their respective just rewards: for the small minority saved by unmerited grace, there is the vision of God, a joy we can only dimly discern at the moment [*De Civitate Dei* XXII.29]. For the overwhelming mass of humanity, there is the second death wherein their resurrected bodies will be subject to eternal torment by flames that will inflict pain without consuming the body [*De Civitate Dei* XXI.2-4], the degree of torment proportional to the extent of sin [*De Civitate Dei* XXI.16], although the duration is equal in all cases: they must suffer without end, for to suffer any less would be to contradict scripture and undermine our confidence in the eternal blessedness of the small number God has saved [*De Civitate Dei* XXI.23].

In *De Civitate Dei* as in the earlier *Contra Academicos*, Augustine is a eudaimonist who enjoins us to seek a happiness understood in terms of our objective relation to an hierarchical structure [e.g. *De Civitate Dei* XIV.25 and XX.21], and he still invokes philosophy, rightly understood, as an instrument that can help us move towards this end [*De Civitate Dei* XXII.22]. Moreover, he still views the world we experience as only a small part of reality, and here too Augustine sees our earthly lives as perfected in a realm that is outside the flux of history as we know and experience it [*De Civitate Dei* XXI.26]. Much, however, has obviously changed. Gone is the confidence that the "harbor of philosophy" [e.g. *Contra Academicos* 2.1.1] is the haven wherein we can find the rest that we seek, and gone is the idea that the rational life will lead us to our eudaimonistic end; gone as well is the breathless excitement with which Augustine would enjoin others to pursue the life of rational enquiry [e.g. *Contra Academicos* 2.2.5]. In place of all this is a moral landscape that seems even sadder and more unsettling than the sense of loss it was originally intended to relieve. And yet, even at the very end of *De Civitate Dei*, Augustine makes clear that he still regards this as a landscape which holds out the prospect of an incomparable vision and rest from all anxiety, a renewed condition that defies all mortal estimation [*De Civitate Dei* XXII.30; see also XX.21].

Now the aging Bishop of Hippo, Augustine still shows a trait he first exhibited as a youthful convert at Cassiciacum: a keen sense of the moral darkness that surrounds us and a philosophical penchant for the unexpected turn of thought by which he would have us escape it.

Legacy

In the long and difficult controversy with the Pelagians, Augustine found his own earlier writings on the will cited by his opponents as evidence that he himself once advocated the view he came so vehemently to oppose [see *Retractationes* I.9.3-6]. What is more, he dies just as the Vandals are besieging the gates of Hippo, leaving unfinished yet another work against Julian of Eclanum, a Pelagian opponent of considerable intellectual resources who had, among other things, accused Augustine of holding views indistinguishable from those of the Manicheans whom Augustine had opposed so many years before [Bonner, 1999]. And here, perhaps, is an irony as cruel as it is intriguing: eleven centuries later, when the Church to which Augustine had devoted the last four and a half decades of his life was to split in a manner that still shows no signs of reconciliation, both sides would appeal to Augustine as an authority on questions of doctrine [Muller 1999; Grossi 1999].

Leaving aside the relative merits of these accusations and appeals, their mere existence is only possible because of the diversity and astonishing range of Augustine's thought over the course of his lifetime. Augustine's movement from a largely Hellenistic eudaimonism to the increasingly somber eschatology of his later works is much more than a mere shift of position. It is the emergent product of a mind continually immersed in controversy and ever obliged to rethink old positions in light of new exigencies, obliged to turn yet again the stone turned so many times before.

First and foremost in Augustine's legacy is the voluminous body of work that encompasses this movement, revealing a range of thought only a handful of philosophers have managed to achieve. The diversity contained in this body of work defies any easy or succinct synopsis, and anyone who approaches it will find a range of ideas that can alternately intrigue, surprise, and sometimes even disarm and shock. One will also find a range of genres and styles, ranging from texts crafted with great rhetorical subtlety to texts that seem to "jangle" with the "music" [O'Connell 1987, pg. 203] of one who is thinking aloud as he writes. For those who want arguments and evidential support, it is there to be had, sometimes in repetitive abundance; for those sensitive to and appreciative of the power of poetic imagery, that too is abundantly in place. Indeed, as Robert O'Connell says, "Augustine constructed more through a play of his teeming imagination than by the highly abstract processes of strict metaphysical thinking" [O'Connell 1986, pg. 3].

But if that vast, multifaceted corpus is the basis of Augustine's legacy, it is also the ultimate obstacle to any attempt at neatly packaging or compartmentalizing it within some "ism" that can be neatly taxonomized. This is, of course, true of most major philosophers, but it seems incontestably true of Augustine. In place of tidy boundaries, there is instead the "jangle" of the corpus itself and the enormous influence it comes to have. This influence is to be found, for example, throughout early medieval philosophy (e.g. Boethius and John Scotus Eriugena), and in Anselm of Canterbury, including in what later came to be known as the ontological argument [*Proslogion*, Chapters I-IV]. Augustine's influence is

plainly discernible in Bonaventure [e.g. *Itinerarium Mentis in Deum*] and others in the thirteenth century who sought an alternative to the Aristotelianism then gaining currency (e.g. John Peckham and Henry of Ghent). Even Thomas Aquinas, a pivotal figure in the rise of Aristotelianism, takes care to address and to accommodate Augustine's view on illumination among many other issues. In the modern period, the echoes in Descartes are conspicuous, both in the *cogito* [Matthews 1992] and elsewhere [Matthews 1999b]. And, of course, few philosophers have invoked Augustine as explicitly and as frequently as Malebranche [see, e.g. "Preface" to *The Search After Truth*]. More recently, one of the most influential works of twentieth century philosophy, Wittgenstein's *Philosophical Investigations*, opens with a lengthy quotation from Augustine's *Confessions* and a discussion of the picture of language that Wittgenstein sees invoked in it [Wittgenstein, *Philosophical Investigations*, Part I, pars 1-3 & 32]. And if this selective historical sampling were not enough, there is an enormous body of secondary literature devoted to Augustine ranging across disciplinary boundaries and across divisions within the philosophical community itself. In 1999 alone, there appeared, among numerous other works, a 900 page encyclopedia devoted to Augustine as a religious and philosophical figure [Fitzgerald, 1999] and a volume of essays by several prominent philosophers in the analytic tradition exploring Augustine's relation to a variety of topics including consequentialism, Kantian moral philosophy, and just war theory (an important issue which unfortunately falls outside the scope of the present discussion) [Matthews 1999]. If one examines the diverse interests of those influenced by Augustine together with the enormous body of secondary literature on Augustine, one finds again what one cannot fail to discern in the Augustinian corpus itself: a diversity as amazing as it is broad, one that defies any attempt at neat summary or tidy explication, a diversity as rich as it is discordant. It is unlikely that this is the legacy that Augustine would have wanted to leave behind, but it is a legacy of a sort that only a handful of philosophers have managed to achieve. The obvious irony notwithstanding, the discordance and diversity are both measures of, and testimony to, an intellectual depth and range seldom equalled in the history of western philosophy.

Bibliography

Selected Latin Texts and Critical Editions

The most common and most complete (but uncritical) edition of Augustine in Latin is the seventeenth century Maurist edition of Augustine's *Opera Omnia* which is reprinted in volumes 32-47 of J.P. Migne's *Patrologiae Cursus Completus*, *Series Latina* (Paris 1844-64), referred to below as **PL**. More critical texts are gradually emerging in four main series:

- *Corpus Scriptorum Ecclesiasticorum Latinorum*, Vienna: Tempsky, 1865- [**CSEL**]
- *Corpus Christianorum*, *Series Latina*, Turnhout: Brepolis, 1953- [**CCL**]

- *Bibliotheque Augustinenne, Oervres de Saint Augustin*, Paris: Desclee De Brouwer, 1949- [**BA**]
- *Nuova Biblioteca Agostiniana*, *Opera de S. Agostino, edizione latino-italiana*, Rome: Citta Nuova 1965- [**NBA**]

Given the voluminous number of Augustine's texts, the following list is confined to those especially relevant to the present article. In what follows, the Migne volume [**PL**] will be provided as well as those of any of the other above editions that have appeared. For information on Augustine texts not listed here, the reader is referred to Fitzgerald 1999, pp. xxxv-xlii, and the reader can also feel free to contact the author via the email address listed at the end of this article.

- *De Beata Vita* (*On The Happy Life*), circa 386/7 C.E.: **PL**32; **CSEL**63 (1922); **CCL**29 (1986); **NBA**3 (1970).
- *Contra Academicos* (*Against the Skeptics*), circa 386/7 C.E.: **PL**32; **CSEL**63 (1922); **CCL**29 (1970); **BA**4 (1939); **NBA**3 (1970).
- *Soliloquia* (*Soliloquies*) circa 386 C.E.: **PL**32; **CSEL**89 (1986); **BA**5 (1939); **NBA**3 (1970).
- *De Libero Arbitrio* (*On Free Will*) Book I circa 386/8 C.E., Books II-III, circa 391-5: **PL**32; **CSEL**74 (1956); **CCL**29 (1970); **BA**6 (1952); **NBA**3/2 (1976).
- *De Magistro* (*On The Teacher*) circa 389 C.E.: **PL**32; **CSEL**77 (1961); **CCL**29 (1970).
- *Ad Simplicianum* (*To Simplicianus*) circa 396 C.E.: **PL** 40; **CCL**44 (1970).
- *Confessiones* (*Confessions)* circa 397-401 C.E.: **PL**32; **CSEL** (1896); **CCL**27 (1981). See also O'Donnell 1992, volume 1 in "Selected Secondary Works" below.
- *De Trinitate* (*On The Trinity*) circa 399-422/6 C.E.: **PL** 42; **CCL** 50/50A.
- *De Genesi ad Litteram* (*On The Literal Meaning of Genesis*) circa 401-415 C.E.: **PL**42; **CSEL**28/1.
- *De Civitate Dei* (*On The City of God*) circa 413-427 C.E.: **PL**41; **CSEL**40; **CCL**47-8.
- *Retractationes* (*Retractations*) circa 426/7 C.E.: **PL**32; **CSEL**36 (1902); **CCL**57 (1984); **BA**12 (1950); **NBA** 2 (1994).
- *Epistulae* (*Letters*) circa 386-430 C.E.: **PL**33; *Ep*. 1-30: **CSEL**34/1 (1895); *Ep*. 31-123: **CSEL** 34/2 (1898); *Ep*. 124-84A: **CSEL**44 (1904); *Ep*. 185-270: **CSEL** 57 (1923); Recently discovered *Ep*.: 1*-29* **BA**46B (1987).

Selected English Translations

The following list is of standard and available English translations of the works cited above. Again, there is no attempt to be exhaustive, and readers seeking information for titles not listed should consult the relevant entry in Fitzgerald 1999 or contact the author via the email address at the end of this article.

- *De Beata Vita* is translated in *The Works of Saint Augustine: A Translation for the 21st Century*, vol 1.3, New City Press, 1990-.
- *Contra Academicos* is translated in *Against the Academicians and The Teacher*, translated by Peter King, Hackett Publishing Company, 1995
- *Soliloquia* is translated in *Soliloquies*, Library of Christian Classics, volume 6, 1953.
- *De Magistro* is translated in *Against the Academicians and The Teacher*, translated by Peter King,

Hackett Publishing Company, 1995

- *Ad Simplicianum* is translated in *The Works of Saint Augustine: A Translation for the 21st Century*, vol. 1.12, New City ress 1990-
- *Confessiones* are translated in *Confessions*, translated by Henry Chadwick, Oxford University Press, 1991.
- *De Trinitate* is translated in *The Works of Saint Augustine: A Translation for the 21st Century*, vol. I.5, New City Press 1990-
- *De Genesi ad Litteram* is translated in *St. Augustine: The Literal Meaning of Genesis*, translated by John H. Taylor, Ancient Christian Writers, vol 41-2, Newman Press 1982.
- *De Civitate Dei* is translated in *The City of God Against the Pagans*, translated by R.W. Dyson, *Cambridge Texts in the History of Political Thought*, Cambridge University Press 1998.
- *Retractationes* is translated in *The Works of Saint Augustine: A Translation for the 21st Century*, vol. I.2, New City Press 1990-
- *Epistulae* are translated by W. Parsons in the *Fathers of the Church* series: Letters 1-82, vol 12; Letters 83-130, vol. 18; Letters 131-64, vol. 20; Letters 165-203, vol. 30; Letters 204-70, vol. 32; recently discovered Letters *1-*29 are translated by R. Eno in vol. 81.

Selected General Studies

The following is a list of works that can be helpful as introductions, guides, or general studies of Augustine's thought. The list represents a variety of viewpoints and approaches to Augustine, but it makes no attempt at being exhaustive. Interested readers should also consult Markus 1967 in "Select Secondary Works" below. The author welcomes suggestions for further additions.

- Bonner, Gerald (1986): *Augustine of Hippo: Life and Controversies*, Canterbury Press 1986.
- Brown, Peter (1967): *Augustine of Hippo: A Biography*, University of California Press 1967.
- Chadwick, Henry (1986): *Augustine*, Past Masters Series, Oxford University Press 1986.
- Clark, Mary T.(1994): *Augustine*, Georgetown University Press 1994.
- Fitzgerald, Allan D. (ed.) (1999): *Augustine Through the Ages: An Encyclopedia*, William B. Eerdmans Publishing Company, 1999.
- Gilson, Etienne (1967): *The Christian Philosophy of Saint Augustine*, translated by L.E.M. Lynch, Random House 1967.
- Kirwan, Christopher (1989): *Augustine*, The Arguments of the Philosophers, Routledge, 1989.
- O'Donnell, James (1985): *Augustine*, Twayne's World Author Series, Twayne Publishers 1985.
- O'Meara, John J. (1954): *The Young Augustine: The Growth of St. Augustine's Mind Up to His Conversion*, Longmans, Green & Co. 1954.
- Rist, John (1994): *Augustine: Ancient Thought Baptized*, Cambridge University Press, 1994.
- Wills, Gary *Saint Augustine*, Viking (Peguin Lives Series), 1999.

Selected Secondary Works

The following provides a list of works relevant to topics covered in the present article, and most of the

works listed are referred to at some point in the body of the article. The author welcomes suggestions for further additions. Interested readers should also note that there is an annual bibliographical survey of literature on Augustine in the *Revue des Etudes Augustininnes*.

- Adams, Marilyn McCord (1999): "Romancing the Good: God and the Self according to St. Anselm of Canterbury" in Matthews 1999, pp. 91-109.
- Armstong, A.H. ed. (1967), *The Cambridge History of Later Greek & Early Medieval Philosophy*, Cambridge University Press, 1967.
- Babcock, William S. (1979): "Augustine's Interpretation of Romans (A.D. 394-3960)," *Augustinian Studies* 10 (1979), pp. 55-74.
- Beatrice, P.F (1989): "*Quosdam platonicorum libros*: The Platonic Readings of Augustine in Milan," *Vigiliae Christianae* 43 (1989) 248-281.
- Bittner, Rudiger (1999): "Augustine's Philosophy of History" in Matthews 1999, pp. 345-360.
- Bonner, Gerald (1972): *Augustine and Modern Research on Pelagianism, The Saint Augustine Lecture Series*, Villanova University Press, 1972.
- Bonner, Gerald (1999): "*Julianum opus imerfectum, Contra*" in Fitzgerald 1999, pp. 480-481.
- Bourke, Vernon J (1963): *Augustine's View of Reality: The Saint Augustine Lecture 1963*, Villanova University Press, 1963.
- Bubacz, Bruce (1981): *St. Augustine's Theory of Knowledge: A Contemporary Analysis*, Edwin Mellin 1981.
- Burnyeat, M.F. (1983): *The Skeptical Tradition*, University of California Press 1983.
- Burnyeat, M.F. (1987): "Wittgenstein and Augustine De Magistro," *Proceedings of the Aristotelian Society*, Supplementary Volume 61 (1987), pp. 1-24, reprinted in Matthews 1999, pp. 286-303.
- Bussanich, John (1996): "Plotinus' Metaphysics of the One" in Gerson 1996 pp.38-65.
- Cranz, Edward F. (1972): "*De Civitate Dei*, XV,2, and Augustine's Idea of Christian Society" in Markus 1972.
- Djuth, Marianne (1999): "Will" in Fitzgerald 1999, pp. 881-885.
- Evans, G.R. (1982): *Augustine On Evil*, Cambridge University Press, 1982.
- Gerson, Lloyd P. (ed.) (1996), *The Cambridge Companion to Plotinus*, Cambridge University Press 1996.
- Grossi, Vittorino (1999): "Council of Trent" in Fitzgerald 1999, pp. 843-845.
- Holscher, Ludger (1986): *The Reality of the Mind: Augustine's Philosophical Arguments for the Human Soul as A Spiritual Substance*, Routledge & Kegan Paul 1986.
- Kirwan, Christopher (1983): "Augustine against the Skeptics" in Burnyeat 1983, pp. 205-223.
- Kirwan, Christopher (1999): "Avoiding Sin: Augustine against Consequentialism," in Matthews 1999, pp. 183-194.
- Markus, R.A. (1967), "Marius Victorinus and Augustine," in Armstrong 1967, pp. 331-419.
- Markus, R.A. (ed.) (1972): *Augustine: A Collection of Critical Essays*, Anchor Books 1972.
- Matthews, Gareth B.(1972): "*Si Fallor, Sum*," in Markus 1972, pp. 151-167.
- Matthews, Gareth B. (1992): *Thought's Ego in Augustine and Descartes*, Cornell University Press, 1992.
- Matthews, Gareth B. (ed.) (1999): *The Augustinian Tradition*, University of California Press 1999.

- Matthews, Gareth B. (1999b): "Augustine and Descartes on Minds and Bodies" in Matthews 1999, pp. 222-232.
- Mendelson, Michael (1995): "The Dangling Thread: Augustine's Three Hypotheses of the Soul's Origin in the *De Genesi ad Litteram*," *British Journal of the History of Philosophy*, vol. 3, no. 2 (1995), pp. 219-247.
- Mendelson, Michael (1998): "The Business of Those Absent: The Origin of the Soul in Augustine's *De Genesi ad Litteram* 10.6-26," *Augustinian Studies* 29:1 (1998), pp. 25-81.
- Mendelson, Michael (2000): "*venter animi/distentio animi*: Memory and Temporality in Augustine's *Confessions*," *Augustinian Studies* 31:2 (2000), pp. 137-163.
- Miles, M.E. (1979): *Augustine on the Body*, Scholars Press 1979.
- Muller, Richard, "Augustinianism in the Reformation" in Fitzgerald 1999, pp. 705-707.
- Nash, Ronald H. (1969): *The Light of the Mind: St. Augustine's Theory of Knowledge*, The University Press of Kentucky, 1969.
- O'Connell, Robert J. (1968): *St. Augustine's Early Theory of Man*, Harvard University Press 1968.
- O'Connell, Robert J. (1969): *St. Augustine's Confessions: The Odyssey of Soul*, Harvard University Press, 1969.
- O'Connell, Robert J. (1972): "Action and Contemplation" in Markus 1972, pp. 38-58.
- O'Connell, Robert J. (1986): Imagination and Metaphysics in St. Augustine, Marquette University Press, 1986.
- O'Connell, Robert J. (1987): *The Origin of the Soul in St. Augustine's Later Works*, Fordham University Press, 1987.
- O'Connell, Robert J. (1993): "The *De Genesi contra Manichaeos* and the Origin of the Soul," *Revue des Etudes Augustinennes* 39 (1993), pp. 129-41
- O'Connell, Robert J. (1994): *Soundings in St. Augustine's Imagination*, Fordham University Press, 1994.
- O'Daly, Gerard, (1987): Augustine's Philosophy of Mind, University of California Press, 1987.
- O'Donnell, James J. (1992): *Augustine: Confessions. Text and Commentary* in 3 volumes, Oxford University Press, 1992.
- O'Meara, Dominic J. (1996): "The Hierarchical Ordering of Reality in Plotinus" in Gerson 1996, pp. 66-81.
- Plantinga, Alvin (1992): "Augustinian Christian Philosophy," *Monist* 75, no. 3 (1992), pp. 291-320, reprinted in Mathews 1999, pp. 1-26.
- Plotinus, *Enneads*, translated by A.H. Armstrong, 7 vols. Loeb Classical Library, Harvard University Press, 1966-1984.
- Rowe, William (1964): "Augustine on Foreknowledge and Free Will," *Review of Metaphysics* 18 (1964), pp. 356-63, reprinted in Markus 1972, pp. 209-17.
- Rist, John (1972): "Augustine on Free Will and Predestination" in Markus 1972, pp.218-252.
- Rist, John (1989): Review of O'Connell (1987) in *International Philosophical Quarterly* 1989.
- Schroeder, Frekeric M. (1996): "Plotinus and Language" in Gerson 1996, pp. 336-355.
- Straume-Zimmermann, L., F. Broemser, and O. Gigon, eds. and trans. (1990): *Marcus Tullius Cicero: Hortensius, Lucullus, Academici libri* Artemis 1990.
- TeSelle, Eugene (1972): "Rufinus the Syrian, Caelestius, Pelagius: Explorations in the Prehistory of the Pelagian Controversy," *Augustinian Studies* 3 (1972), pp. 61-95.

- TeSelle, Eugene (1999): "Pelagius, Pelagianism" in Fitzgerald 1999, pp. 633-640.
- Teske, Roland J. (1991): "St. Augustine's View of the Original Human Condition in *De Genesi contra Manichaeos*," *Augustinian Studies* 22 (1991), pp. 141-55.
- Teske, Roland J. (1999): "Soul" in Fitzgerald 1999, pp. 807-812.
- Wetzel, James (1992): *Augustine and the Limits of Virtue*, Cambridge University Press 1992.

Other Internet Resources

- Augustine Site, by James J. O'Donnell
- *Augustinian Studies* Site , by Allan Fitzgerald, O.S.A.
- A Bibliography of Augustine and the Latin West. by Fr. William Harmless, S.J.
- Charts of the Works of Augustine, by Fr. William Harmless S.J.

Related Entries

divine illumination | Hellenistic Philosophy | Manicheism | medieval philosophy | Neoplatonism | Plotinus | skepticism: ancient

In Memoriam:

Robert J. O'Connell, S.J.

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 24, 2000
Content last modified: October 2, 2000

Divine Illumination

Divine illumination is the oldest and most influential alternative to naturalism in the areas of mind and knowledge. The doctrine holds that human beings require a special divine assistance in their ordinary cognitive activities. Although most closely associated with Augustine and his scholastic followers, the doctrine has its origins in the ancient period and would reappear, transformed, in the early modern era.

Orientation

The theory of divine illumination is generally conceived of as distinctively Christian, distinctively medieval, and distinctively Augustinian. There is some justification for this, of course, inasmuch as Christian medieval philosophers gave the theory serious and sustained discussion, and inasmuch as Augustine gave illumination a very prominent role in his theory of knowledge. Still, it is better to think of the theory in a wider context. Divine illumination played a prominent part in ancient Greek philosophy, in the later Greek commentary tradition, in neo-Platonism, and in medieval Islamic philosophy. Moreover, it was Christian medieval philosophers, near the end of the thirteenth century, who were ultimately responsible for decisively refuting the theory. I will suggest that we view this last development as the first great turning point in the history of cognitive theory.

I understand a theory of divine illumination to be a theory on which the human mind regularly relies on

some kind of special supernatural assistance in order to complete (some part of) its ordinary cognitive activity. The assistance must be *supernatural*, of course, or it will not count as *divine* illumination. It must be *special*, in the sense that it must be something more than the divine creation and ongoing conservation of the human mind. (If the latter were to count as illumination, then all theists would be committed to the theory of divine illumination.) The mind must *regularly* rely on this assistance, in order to complete its *ordinary* cognitive activity: otherwise, an occasional mystical experience might suffice to confirm a theory of divine illumination. But a defender of the theory need hold only that we require this assistance for *some part of* our ordinary cognitive activities: hardly anyone has supposed that every form of human cognition requires divine illuimination.

It is useful to think of divine illumination as analogous to grace. Just as a proponent of grace postulates a special divine role on the volitional side, so a proponent of divine illumination postulates a special divine role on the cognitive side. Grace is intended as an explanation not of all human desires and motivations, nor even of all virtuous desires and motivations. Rather, the proponent of grace holds that there is a certain class of volitional states, crucial to human well-being, that we can achieve only with special divine assistance. Likewise, the theory of divine illumination is intended as an explanation not of all belief, nor even of all knowledge. Rather, the theory holds that there are certain kinds of knowledge, crucial to cognitive development, that we can achieve only with special divine assistance. It is an odd fact that, despite the close analogy, grace is regarded not as a philosophical question, but as a theological one. It is an equally odd fact that, whereas divine illumination hasn't generally been regarded as plausible since the thirteenth century, grace continues to be taken seriously by many theologians. Perhaps both of these facts can be accounted for by motivational psychology's relative obscurity in comparison to cognitive psychology.

For most people today it is hard to take divine illumination seriously, hard to view it as anything other than a quaint relic. A first step toward developing a proper perspective on the theory is to see it in its broader context, not as peculiarly Christian or medieval, but as an assumption shared by most premodern philosophers. A second step in the same direction is to identify and to take seriously the philosophical problem that drives illumination theory. In large part, the theory has been invoked to explain rational insight -- that is, a priori knowledge. Recent philosophers, preoccupied with empirical knowledge, have not had much interest in this topic. (A recent notable exception is Bonjour 1998.) But to see how something like divine illumination could have ever seemed at all plausible, one has to see how deeply puzzling the phenomenon of rational insight actually is. One way of seeing this, and of seeing how little we understand rational insight, is to look at cases where something goes wrong. A recent biography of the Nobel-prize winning mathematician John Nash describes his long period of mental illness, during which time he held various odd beliefs such as that extraterrestrials were recruiting him to save the world. How could he believe this, a friend asked during a hospital visit, given his devotion to reason and logic?

> "Because," Nash said slowly in his soft, reasonable southern drawl, as if talking to himself, "the ideas I had about supernatural beings came to me the same way that my mathematical ideas did. So I took them seriously" (Nasar 1998, p.11).

In a case such as this we don't know what to do, because we are accustomed to give unhesitating trust to the deliverances of pure reason. But why should we trust reason in this way? Why should we have confidence that others can come to share our insights? Where does it come from? The theory of divine illumination attempts to answer such questions.

The Ancient Background

What follows is by no means a comprehensive survey. Still, it seems useful to begin near the beginning, with this remark by Socrates from the *Apology*:

> I have a divine or spiritual sign which Meletus has ridiculed in his deposition. This began when I was a child. It is a voice, and whenever it speaks it turns me away from something I am about to do, but it never encourages me to do anything (31d, tr. Grube).

This appears to be an entirely straightforward expression of a theory of divine illumination. It is not clear who is providing the illumination. Apuleius (fl. 150 CE) would later identify the source as a certain kind of friendly demon, and argue that it was only fitting for Socrates, the most perfect of all human beings, to receive such illumination (*De deo Socratis*, XVII-XIX). (This idea that illumination comes to those who deserve it would be proposed by Augustine as well [e.g., *De magistro* 11.38], but abandoned in his later writings as untenable.) As one might expect from Socrates, his illumination seems restricted to the moral sphere. Its form is unclear: is it propositional? (see *Phaedrus* 242bc) Is it merely the pang of conscience? (For discussion of these and other questions, see McPherran 1996.) Whatever the details, Socrates is explicitly describing a kind of cognitive guidance that has a "divine or spiritual" source. The passage may be an embarrassment to classicists, but it surely belongs in the same tradition as later medieval endorsements of illumination.

Not all appeals to the divine involve this sort of direct communication. (Indeed, Socrates's reference to a "voice" is quite extraordinary.) But the leading figures in ancient Greek philosophy were equally committed to some kind of divine role in cognition. Plato's theory of recollection presupposes that the human mind somehow has built into it a grasp of the Forms, suggesting that at some point the soul must have received some kind of illumination. Indeed, Plato's arguments for recollection anticipate the two lines along which medieval views would later develop. The *Meno* focuses on a priori, rational insight, as illustrated by the slave's ability to see for himself the validity of a geometrical proof. The *Phaedo* in contrast focuses on universal properties -- Equality, for instance, as compared with two equal sticks (74a) -- contrasting the changeable imperfection of the physical world with the exemplary perfection of the Forms. Medieval philosophers from Augustine on, although largely lacking in firsthand knowledge of Plato, would argue for illumination along both of these lines.

Aristotle too seems to invoke the divine. He describes the active intellect in this way:

> This intellect is separate, unaffected, and unmixed, being in essence activity.... It is not the

> case that it sometimes thinks and at other times not. In separation it is just what it is, and this alone is immortal and eternal (*De anima* III 5, 430a17-23).

There is of course unending controversy over the meaning of this text. One very common reading, in ancient times and our own, is that this active intellect is something divine, not a human faculty at all. If one then makes the further, natural assumption that the active intellect participates in ordinary human cognition, then Aristotle would clearly be committed to a version of divine illumination. (One should, however, be cautious in supposing that Aristotle's *nous poietikos* plays anything like the role played by the scholastic *intellectus agens*. See Haldane 1992.) Not everyone has been persuaded that this active intellect is something literally separate and divine. But even if one supposes that the active intellect is a part of the human soul, it is nevertheless difficult to avoid the suspicion that some sort of special divine influence is at work. Everything in the passage cries out for some sort of supernatural element in human cognition.

Alexander of Aphrodisias (fl. 200 CE) was influential in pushing the separate-and-divine reading of *De anima* III.5, and Islamic philosophers (most notably Avicenna [980-1037] and Averroes [c.1126-c.1198]) would later follow suit. Themistius (4th century CE), in contrast, championed the part-of-soul reading, and Thomas Aquinas (1225-1274) would second this interpretation. A full treatment of this topic would cover the later Aristotelian and Platonic traditions, Greek and Islamic. But here I will focus exclusively on the Latin West, where it was Augustine who played the decisive role in formulating the doctrine of illumination. (For a useful survey of ancient and medieval Aristotelian accounts of agent intellect, see Brentano 1992. On the later Platonic tradition, see Gersh 1978. The Islamic tradition deserves a full essay in its own right, covering not just names well known in the Latin West, but also figures like Suhrawardi and Mulla Sadra, central figures in the Islamic illuminationist tradition [see Fakhry 1983].)

Augustine

Throughout his long literary career, Augustine (354-430) stresses the role of divine illumination in human thought. One could choose almost any work to illustrate this point; here I will focus on the most familiar of all, the *Confessions*, where Augustine invokes divine illumination constantly, and makes bold claims for its global necessity:

> The mind needs to be enlightened by light from outside itself, so that it can participate in truth, because it is not itself the nature of truth. You will light my lamp, Lord (IV.xv.25).
>
> None other than you is teacher of the truth, wherever and from whatever source it is manifest (V.vi.10).
>
> You hear nothing true from me which you have not first told me (X.ii.2).
>
> Truth, when did you ever fail to walk with me, teaching me what to avoid and what to

> seek.... Without you I could discern none of these things (X.xl.65).

Even during the Middle Ages, Augustine's readers disagreed on the precise nature of his theory. One thing that seems clear from these passages alone is that divine illumination is an influence that we receive in an ongoing way throughout our lives. Thomas Aquinas would later understand illumination as an infusion all at once at the start (see below), but this seems untenable as an interpretation of Augustine. The mind needs to be enligtened "from outside itself"; "it is not itself the nature of truth"; "you *will* light my lamp, Lord" (he has not done so already); truth "walk[s] with me," rather than merely setting me in motion at the start.

To speak of this influence as an illumination is of course to use a metaphor, one not likely to be unpacked fully. Our own minds present enough of a puzzle to us: when we try to understand how the divine mind might influence our own, we must inevitably fall back on metaphor. Still, there are a variety of ways in which we might seek some clarification. In particular, it is helpful to distinguish two ways in which God might provide illumination. First, he might simply give us information of certain kinds, telling us how things are. This is how illumination is most often understood, at least implicitly. But a second possibility is that God would provide not the information itself, but the insight into the truth of the information. On this second model, we would frame beliefs on our own, and God would illuminate our minds so that we could see the truth. In other words, God would supply the justification. It is clear that sometimes illumination takes the first form: much of Biblical revelation just is illumination in this sense. But Augustine's theory of illumination seems largely to be of the second kind. Consider this famous passage from the *Confessions*:

> If we both see that what you say is true, and we both see that what I say is true, then where do we see that? Not I in you, nor you in me, but both of us in that unalterable truth that is above our minds (XII.xxv.35).

At issue here is Biblical interpretation. When a reading is advanced that seems clearly correct, how is it that everyone listening grasps the truth of that reading? It is not that God gives us the interpretation itself, but that he allows us to see that the interpretation is true.

This understanding of illumination is particularly apparent in the *De magistro*, where Augustine argues that only God can teach us. Of course, other people can tell us things, and can thereby communicate ideas to us. And we can believe what others tell us: indeed, our lives would be impoverished if we didn't regularly accept what others tell us. But all of this stays at the level of mere belief. It is not knowledge unless we grasp with our minds the truth of what we are hearing:

> When I speak the truth, I do not teach someone who sees these truths. For he is taught not by my words but by the things themselves made manifest within when God discloses them (12.40).

The speaker's role is not irrelevant in this process. My words give the listener an idea that he can then

verify for himself in light of God's illumination. Illumination is what allows us to go from mere true belief to knowledge. Illumination provides justification.

This account is most attractive in cases of a priori knowledge or pure reasoning, where we grasp through the mind alone that an argument is valid or that a conclusion is necessary. But there is another strain of thought running through Augustine, one that focuses on the mind's ability to transcend the untrustworthy senses and grasp the truth that lies beyond mere appearances. The following passage is very often cited in this connection:

> Everything that the bodily senses attain, that which is also called sensible, is incessantly changing.... But what is not constant cannot be perceived; for that is perceived which is comprehended in knowledge. But something that is incessantly changing cannot be comprehended. Therefore we should not expect pure truth from the bodily senses (*De diversis quaestionibus octoginta tribus*, q.9).

In one stroke, this argument rules out the physical world as an object of pure truth, and rules out the senses as a source for that truth. It must be the mind, then, with which we attain truth, and that truth must be something beyond the sensible world. Plainly, the mind cannot rely on the senses. But what else is there? The conclusion we are invited to reach is that the mind must rely on God.

One might try assimilating this line of thought to those passages where Augustine has in mind necessary a priori truths. But it is more natural to take this in a different way, as an account of how the mind goes beyond the sensible data to a grasp of the real essences of things. Accordingly, the theory of divine illumination would be put to two very different sorts of work in the later Middle Ages: as an account of a priori knowledge, and as an account of concept formation. Each account raises its own set of issues. Taken in the first way, divine illumination has to compete against the claim that the mind is naturally capable of grasping such truth. Taken in the second way, questions immediately arise about the nature of conceptual knowledge. Do essences (or properties in general) exist in the physical world? Do they exist only in the divine mind? Do the senses play any role in the process of concept formation? Later medieval philosophers would handle these issues in interestingly different ways.

Thirteenth-century Franciscans

Augustine's position would remain ascendent among Christian philosophers for most of the Middle Ages. Thirteenth-century Franciscans, led by figures such as Bonaventure (c.1217-1274) and Matthew of Aquasparta (c.1237-1302), gave the theory a detailed and systematic defense, focusing on the changeability and hence uncertainty of the human mind and the sensory world (see Rohmer 1928). Bonaventure characterically argues,

> Things have existence in the mind, in their own nature (*proprio genere*), and in the eternal art. So the truth of things as they are in the mind or in their own nature -- given that both

> are changeable -- is sufficient for the soul to have certain knowledge only if the soul somehow reaches things as they are in the eternal art (*De scientia Christi*, q.4 resp.).

Certain knowledge requires steadfast unchangeability. Since that can be found only in the divine mind, and since we have access to the divine mind only through illumination, certain knowledge requires illumination.

This line of argument came to seem increasingly old-fashioned as the thirteenth century progressed. The growing influence of Aristotle's theory of cognition, as developed in particular by the Dominican friars Albert the Great (c.1200-1280) and his student, Thomas Aquinas (c.1225-1274), offered an impressive picture of how human beings might be able to achieve certain knowledge despite the changeability of mind and matter. These developments struck many Franciscans as a betrayal of Christianity. John Peckham (c.1225-1292), in a letter dating from 1285, writes

> I do not in any way disapprove of philosophical studies, insofar as they serve theological mysteries, but I do disapprove of irreverent innovations in language, introduced within the last twenty years into the depths of theology against philosophical truth and to the detriment of the Fathers, whose positions are disdained and openly held in contempt.

Continuing, Peckham criticizes the doctrine

> which fills the entire world with wordy quarrels, weakening and destroying with all its strength what Augustine teaches concerning the eternal rules and the unchangeable light.... (quoted in Gilson 1955, p.359).

At roughly the same time, Roger Marston (c.1235-1303) writes of those who, "drunk on the nectar of philosophy... twisted toward their own sense all of Augustine's authoritative texts on the unchanging light and the eternal rules" (*Quaestiones disputatae de anima* 3 ad 30).

Marston's view is particularly interesting because he proposes a synthesis of Augustine and Aristotle. On his view,

> It is necessary to posit in our mind, beyond the phantasms or abstracted species, something by which we to some degree attain the unchanging truths. I believe this to be no different than the influence of the eternal light.... For the eternal light, irradiating the human mind, makes a certain active impression on it, from which a certain passive impression is left in it, which is the formal principle of cognizing the unchanging truths (*De anima* 3, p.263).

Rather than dismiss the agent intellect as superfluous, Marston follows Alexander of Aphrodisias et al. in treating the agent intellect as separate and divine -- indeed, as God himself. Earlier in the thirteenth century, William of Auvergne (c.1180-1249) had likewise identified the agent intellect with God (*Tractatus de anima* 7.6; cf. Gilson 1926-27, pp.67-72). Such cases illustrate how the various medieval

disputes over whether human beings might share a single intellect -- so absurd on their face -- are in fact simply alternative formulations of the dispute over divine illumination. But there were subtle differences among the various approaches. So whereas Auvergne largely turns his back on Aristotle, Marston is more accommodating. In addition to the divine agent intellect, he allows that each human being possesses its own agent intellect.

On Marston's account, Aristotle and Augustine turn out to be entirely in harmony: each uses his own terminology to defend the same theory of divine illumination (*De anima* 3, p.258). Étienne Gilson (1933) has characterized Marston's position as Augustinianism gone Avicennian (*Augustinisme avicennisant*). But this label unfairly prejudices the case in favor of Aquinas's perspective: it assumes that Marston has been seduced by an Islamic misreading of Aristotle, and it closes off the possibility that Augustine and Aristotle might have more in common than is typically allowed.

The case of Peter John Olivi (1247/8-1298) demonstrates how precarious a position the illumination theory held by the 1280s. Olivi, a Franciscan whose work would eventually be condemned by his own order, presents a compelling critique of the Augustinian illumination theory (*I Sent*., q.2), presenting his comments in the form of "cautions" (*cavenda*). The theory, he notes, is often very vague with respect to the actual process of illumination. And in running through the various possible accounts of the process available to a defender of the theory, Olivi appears to rule out every one. The eternal reasons cannot represent things distinctly and specifically, because then we would have no need of any sensory input. But if the eternal reasons give us information only of a general and indistinct sort, then at what level of generality? Does it supply us with information about species, or genera? If divine illumination is efficacious at any level, why do we seem to need the senses for all of our concepts? Olivi's questions and cautions go on and on. But after laying an entire minefield of this sort for anyone who would defend divine illumination -- at ad 6 he rejects the Augustinian argument set out above by Bonaventure -- he nevertheless comes to the surprising conclusion that he accepts the theory:

> These things, since I don't know how to analyze them fully, I set out only as cautions. For although the stated position is in itself venerable (*sollemnis*) and sensible, it could nevertheless be quite dangerous to those who are not carefully supervised. And so I hold the stated position as it is, because it belongs to men who are highly venerable. Nevertheless I leave an exposition of the above to their wisdom (q.2, pp.512-13).

This was a favorite strategy of Olivi's: to criticize a theory fiercely, exposing seemingly devastating difficulties, and then to embrace the theory anyway, as a pious gesture of respect. It's hard to resist reading between the lines, and concluding that dusk was fast approaching for the theory of divine illumination.

Thomas Aquinas

Thomas Aquinas is often thought of as the figure most responsible for putting an end to the theory of divine illumination. Although there is some truth to this view, as we will see, it seems more accurate to

regard Aquinas as one of the last defenders of the theory, as a proponent of innate Aristotelian illumination.

A vivid example of the way Aquinas moves from an Augustinian to an Aristotelian framework occurs in his Treatise on Human Nature (*Summa theologiae* 1a 75-89), where he considers the Augustinian claim that "pure truth should not be looked for from the senses of the body" (84.6 obj.1). In reply, Aquinas invokes the Aristotelian agent intellect:

> From those words of Augustine we are given to understand that truth is not *entirely* to be looked for from the senses. For we require the light of agent intellect, through which we unchangeably cognize the truth in changeable things, and we distinguish the things themselves from the likenesses of things (ad 1).

It is not at all clear, here or elsewhere, how the agent intellect carries out the two tasks he describes. (If Aquinas had given us a satisfactory account of this, he would have thereby solved the leading problems of epistemology.) But for present purposes it is enough to notice how Aquinas seems to replace Augustinian illumination with the Aristotelian agent intellect. Thus the traditional verdict has been that Aquinas replaced Augustine with Aristotle, and exchanged illumination for abstraction.

There is more to the story. In the immediately preceding article, Aquinas explicitly discusses Augustinian divine illumination, and reaches the affirmative conclusion that "the intellective soul does cognize all true things in the eternal reasons" (84.5sc). Often this affirmative conclusion gets treated as little more than lip-service to the authority of Augustine, and the article as a whole gets taken as a backhanded repudiation of illumination theory: affirming the theory in form but denying it in substance. This is a misreading. Aquinas sees something important in Augustine's theory, something worth preserving.

Aquinas does reject certain conceptions of divine illumination. He denies that human beings in this life have the divine ideas as an *object* of cognition. And he denies that divine illumination is sufficient on its own, without the senses. Neither of these claims was controversial. What Aquinas further denies, and what was controversial, was the claim that there is a special ongoing divine influence, constantly required for the intellect's operation. Aquinas instead argues that human beings possess a sufficient capacity for thought on their own, without the need for any "new illumination added onto their natural illumination" (*Summa theol*. 1a2ae 109.1c). From one perspective this makes for an important difference between Aquinas and his Franciscan contemporaries. But from another perspective the difference seems slight, because Aquinas is by no means removing God from the picture. Here is how he expresses his endorsement of illumination theory:

> It is necessary to say that the human soul cognizes all things in the eternal reasons, through participating in which we cognize all things. For the intellectual light that is in us is nothing other than a certain likeness of the uncreated light, obtained through participation, in which the eternal reasons are contained. Thus it is said in Psalm 4, *Many say, Who shows us good things?* To this question the Psalmist replies, saying *The light of your face,*

> *Lord, is imprinted upon us*. This is as if to say, through that seal of the divine light on us, all things are demonstrated to us (*Summa theol*. 1a 84.5c).

There is some temptation to take all of this simply as an expression of Aquinas's more general view that God is the first cause of all things. He writes, for instance,

> *All* active created powers operate in virtue of being directed and moved by the Creator. So it is, then, that in all cognition of the truth, the human mind needs the divine operation. But in the case of things cognized naturally it does not need any new light, but only divine movement and direction (*In de trinitate* pro. 1.1c).

Here there seems to be nothing special about the intellect's need for illumination. The intellect, like all of nature, needs God as its first mover. If you like, think of this as divine illumination. But viewed under this aspect, it is no wonder the theory was controversial. While his Franciscan contemporaries were insisting on a special role for God in human cognition, Aquinas seems to move as far in the opposite direction as his theism would permit.

But passages of this last kind are misleading, because Aquinas does see something especially mysterious about human cognition, and he appeals to God as a way of solving this mystery. The agent intellect, on Aquinas's view, accounts for our capacity to grasp self-evident truths. We have an immediate and direct grasp of the truth of first principles, such as the principle of noncontradiction (see, e.g., *Summa contra Gentiles* II.83.1678). We do not infer the truth of this principle, we do not discover that it is true through any kind of induction. Instead we simply see its truth, as soon as we are confronted with an instance where it applies. This is not innate knowledge; we are not born knowing these principles. What we are born with is the capacity to recognize their truth as soon as we are confronted with instances of them. These first natural conceptions are "the seeds of all the things that are subsequently cognized" (*De veritate* 11.1 ad 5). In this sense, Aquinas is even willing to speak of the soul's having a prior knowledge of everything that it knows:

> The soul forms in itself likenesses of things inasmuch as, through the light of agent intellect, forms abstracted from sensible objects are made actually intelligible, so as to be received in the possible intellect. *And so, in a way, all knowledge is imparted to us at the start*, in the light of agent intellect, mediated by the universal concepts that are cognized at once by the light of agent intellect. Through these concepts, as through universal principles, we make judgments about other things, and in these universal concepts we have a prior cognition of those others. In this connection there is truth in the view that the things we learn, we already had knowledge of (*De veritate* 10.6c).

Because all of what we know can be traced back to these fundamental principles, there is a sense in which everything we learn, we already knew. An innate grasp of certain basic truths, recognized by the light of agent intellect, plays a crucial, foundational role.

The light of agent intellect is of course given to us from God -- "a certain likeness of the uncreated light, obtained through participation" (1a 84.5c). Without appealing to God, Aquinas sees no way of explaining how we recognize the truth of first principles. Neither deductive nor inductive reasoning can account for the way in which we immediately *see* that such principles are true. This insight, then, is simply something we are given:

> The light of this kind of reason, by which principles of this kind are known to us, is imparted to us from God. It is like a likeness of the uncreated truth reflecting in us. So, since no human teaching can be effective except in virtue of that light, it is clear that it is God alone who internally and principally teaches us (*De veritate* 11.1c).

The light of agent intellect, a likeness of the divine ideas, is the essential starting-point for all knowledge. This is the epistemological context for Aquinas's famous words, *Non nisi te*.

Aquinas agrees with his Franciscan contemporaries that intellective cognition is incomplete without some sort of supernaturally infused insight. The only difference is that Aquinas wants that insight to be given all at once, from the start -- bottled up within agent intellect, as we might think of it. His opponents, in contrast, think of illumination as an ongoing process, as necessary as the air we breathe. It is easy to see how, at the time, this difference might have seemed important. But from our present perspective the differences seem rather slight: they seem to be arguing simply over the means of transmission. Aquinas conceives of illumination as a deep well within us, whereas the Franciscans conceived of it as raining down in drops.

Henry of Ghent

Although in a sense Thomas Aquinas defends a version of divine illumination, he in another sense clearly weakens the theory by giving it the status of an innate gift rather than ongoing patronage. In making for the agent intellect a central place in his theory of cognition, Aquinas has less room for illumination. As the thirteenth century progressed, philosophers and theologians were increasingly willing to make this tradeoff. While the Aristotelian theory of cognition waxed, the Augustinian theory of divine illumination waned. To combine the two seemed, in the words of Étienne Gilson (1930), "unproductive and even, in a sense, contradictory."

It is this seemingly contradictory task that Henry of Ghent (c.1217-1293) took upon himself in the years immediately after Aquinas's death. Ghent was neither Dominican nor Franciscan, but a so-called "secular" master at the University of Paris. His project was to defend an Aristotelian theory of cognition while at the same time reviving divine illumination in its traditional Augustinian form. To those, like Aquinas, who were arguing for the self-sufficiency of the human cognitive powers, Ghent replies,

> this is true for natural things, as regards knowing what is true of the thing.... Pure truth, however, or any truth that must be cognized supernaturally, or perhaps any truth at all,

> cannot be known without God himself doing the teaching (*Summa* 1.7 ad 1; 17rM).

On Ghent's terminology, to know what is *true* of a thing is simply to have a veridical impression of it: to represent a thing as it is. To grasp the *truth* of a thing, in contrast, is to grasp its nature. Only this latter sort of cognition counts as knowledge in the strict sense, because only here are we getting at the unchanging reality of the material world.

> For there is no knowledge of things insofar as they are external in effect, but insofar as their nature and quiddity is grasped by the mind (*Summa* 2.2 ad 1; 24rG).

For knowledge of this kind, divine illumination is necessary.

Ghent's argument is interestingly different from that of Augustine and his Franciscan followers. Whereas they had dismissed the physical world as too changeable to be a fit subject for human knowledge, Ghent believes that pure truth and certain knowledge can be had of the physical world, provided we manage to grasp the real essences of things. Since we cannot do so on our own, we need divine illumination to go beyond sensory appearances, to have genuine insight into the nature of reality. At its most basic level, Ghent is offering a critique of the agent intellect. Although he accepts the doctrine of agent intellect, he refuses to give that faculty the kind of efficacy that it has for Aquinas and other medieval Aristotelians. Not surprisingly, Ghent proposes reviving the earlier thirteenth-century tradition of referring to God himself as a kind of agent intellect (*Quodlibet* 9.15). But Ghent is no longer proposing to synthesize Augustine and Aristotle; he wants to supplement Aristotle's incomplete account with the necessary Augustinian illumination. (This is a summary of the account in Pasnau 1995.)

John Duns Scotus

It was the Franciscan John Duns Scotus, more than anyone else, who put an end to the theory of divine illumination. (John Boler remarks [in correspondence] that "as with Nixon's trip to China, it probably could only be done by a Franciscan.") Scotus criticizes Ghent's argument in detail (*Ordinatio* I.3.1.4), arguing against Ghent's own arguments, against the skeptical consequences that would allegedly come from giving up divine illumination, and against the viability of such illumination in its own right. With respect to the last point, Scotus argues that if human cognition were fallible in the way Ghent argues, then outside illumination could not, even in principle, ensure "certain and pure knowledge." On Ghent's account, the human mind cooperates with the divine light in achieving such knowledge. Scotus replies:

> When one of what comes together is incompatible with certainty, then certainty cannot be achieved. For just as from one premise that is necessary and one that is contingent nothing follows but a contingent conclusion, so from something certain and something uncertain, coming together in some cognition, no cognition that is certain follows (*Ordinatio* I.3.1.4 n.221).

If one part of a system is fallible, then that fallibility infects the process of a whole. Scotus's startling claim is that if the human mind were intrinsically incapable of achieving certain knowledge, then not even divine illumination could save it.

Scotus's own view is that the human mind is capable of such knowledge on its own. If by "certain and pure truth" Ghent means "infallible truth, without doubt and deception," then Scotus thinks he has established that human beings "can achieve this, by purely natural means" (*Ord*. I.3.1.4 n.258). How *can* such a thing be established? How can the skeptic be refuted, without appealing to divine illumination? Scotus distinguishes four kinds of knowledge:

- self-evident (*principia per se nota*)
- inductive (*cognita per experientiam*)
- introspective (*cognoscibilia de actibus nostris*)
- sensory (*ea quae subsunt actibus sensus*)

The general strategy is to show that sensory knowledge rests on inductive knowledge, that inductive knowledge rests on self-evident knowledge, and that introspective knowledge can be defended as analogous to self-evident knowledge. Scotus's implicit aim is to shift as much weight as possible onto the broad shoulders of self-evident knowledge.

For Scotus, the self-evident is the bedrock on which other sorts of knowledge rest, and so he doesn't attempt to locate some further set of even more basic truths. Instead, he argues that our self-evident knowledge is foolproof because of certain psychological facts. When one considers a proposition like *Every whole is greater than its part*, one immediately grasps that the terms are related in such a way that the proposition must be true:

> There can be in the intellect no apprehension of the terms or composition of those terms without the conformity of that composition to the terms emerging (*quin stet conformitas*), just as two white things cannot arise without their likeness emerging (*Ord*. I.3.1.4 n.230).

When we see two white objects we immediately grasp, "without doubt and deception," their similarity to one another. Likewise, when we grasp a self-evident truth in our mind, we immediately grasp its truth. Of course, we won't grasp the truth of the proposition if we don't understand the meaning of the terms, but in that case we won't have truly formed the proposition in our mind. And in contrast to the analogous case of recognizing similarity, there is no room for sensory error here. The senses help us acquire certain concepts, but once we have those concepts, the senses drop out of the picture -- sensory reliability becomes irrelevant. Scotus offers the example of a blind man miraculously shown in his dreams an image of black and white. Once he acquires these concepts, he can recognize as truly and infallibly as anyone -- his blindness notwithstanding -- that white is not black (*Ord*. I.3.1.4 n.234).

Scotus is unwilling to discard Augustinian illumination entirely, and so he articulates four senses in which the human intellect sees infallible truths in the divine light. In each sense, the divine light acts not

on us but on the objects of our understanding. By giving objects their intelligibility (*esse intelligibile*), the divine intellect "is that in virtue of which secondarily the objects produced move the intellect in actuality" (*Ord*. I.3.1.4 n.267). When the human mind grasps a self-evident truth, it does so immediately and infallibly not because the mind has received any special illumination, but because the terms of the proposition are themselves intelligible: our grasp of a proposition "seems to follow necessarily from the character of the terms, which character they derive from the divine intellect's causing those terms to have intelligible being naturally" (*Ord*. I.3.1.4 n.268). It is not that we are illuminated by the divine light, but that the truth we grasp is illuminated.

This marks a turning-point in the history of philosophy, the first great victory for naturalism as a research strategy in cognitive theory. On Scotus's account, when we grasp some conceptual truth, nothing miraculous or divine happens within us: "the terms, once apprehended and put together, are naturally suited (*sunt nati naturaliter*) to cause an awareness of the composition's conformity with its terms" (*Ord*. I.3.1.4 n.269). It is of course God that gives the world its intelligibility, just as it is God that creates our cognitive powers. But what's new in Scotus is the idea that the mind is not a special case. From this point forward, divine illumination would cease to be a serious philosophical possibility.

Epilogue

It is easy to miss the significance of what Scotus brought about: in part because it now seems so inevitable, in part because Scotus comes at the end of a gradual trend toward naturalism, and in part because it's generally supposed that nothing of much philosophical importance happened between Aristotle and Descartes. Yet if one looks at the big picture of our evolving philosophical/scientific understanding of the mind, then it is clear that something important happened at the end of the thirteenth century.

Still, Scotus's impact shouldn't be overstated. Although divine illumination, so called, would no longer have prominent supporters, the tendency toward supernatural explanations of cognitive phenomena would survive well beyond the Renaissance. Descartes, to take just one prominent example, can speak of "certain seeds of truth which are naturally in our souls" (*Discourse on Method* 6, AT 64), and of "ideas implanted in the intellect by nature" (*Principles of Philosophy* 2.3, AT 42). In its details, the view is striking similar to Aquinas's: Descartes identifies these ideas as the basis of our knowledge of first principles; he holds that the ideas themselves are formed only in virtue of sensory impressions; he identifies God as the source of these ideas. Innate ideas are the modern successor to divine illumination. But by the seventeenth century the philosophical context has changed so dramatically that this must be regarded as a different topic.

Bibliography

Primary Literature

- Apuleius (1973). *De deo Socratis*, in J. Beaujeu (ed.) *Opuscules Philosophiques et Fragments* (Paris: Société d'Édition "Les Belles Lettres").
- Aristotle (1993). *De anima*, tr. D.W. Hamlyn (Oxford: Clarendon Press).
- Augustine (1982). *Eighty-three different questions* [=*De diversis quaestionibus octoginta tribus*], tr. D.L. Mosher (Washington, D.C.: Catholic University of America Press).
- -----, (1991). *Confessions*, tr. H. Chadwick (Oxford: Oxford University Press).
- -----, (1995). *Against the Academicians and The Teacher* [=*De magistro*] (1995), tr. P. King (Indianapolis: Hackett).
- Bonaventure (1992). *Disputed Questions on the Knowledge of Christ* [=*De scientia Christi*], tr. Z. Hayes (St. Bonaventure: Franciscan Institute Press).
- Descartes, René (1984). *The Philosophical Writings of Descartes*, tr. John Cottingham, Robert Stoothoff, Dugald Murdach (Cambridge: Cambridge University Press).
- Henry of Ghent (1520). *Summa quaestionum ordinariarum*, Paris (reprinted 1953, St. Bonaventure). (Relevant portions are translated in Pasnau forthcoming.)
- -----, (1979-). *Quodlibeta*, in *Henrici de Gandavo Opera Omnia* (Leiden: Brill).
- John Duns Scotus (1987). *Ordinatio* I.3.1.4, in Allan Wolter (tr.), *Philosophical Writings* (Indianapolis: Hackett).
- Matthew of Aquasparta (1930). *Quaestiones de fide et de cognitione* q.2, in R. McKeon (tr.). *Selections from Medieval Philosophers Vol.II: From Roger Bacon to William of Ockham* (New York: Scribners).
- Pasnau, Robert, tr. (forthcoming). *Cambridge Translations of Medieval Philosophical Texts. Volume 3: Mind and Knowledge* (Cambridge: Cambridge University Press).
- Peter John Olivi (1926). *Quaestiones in primum librum Sententiarum*, in *Quaestiones in secundum librum Sententiarum*, vol.3 appendix (Quaracchi: Collegium S. Bonaventurae).
- Plato (1997). *Complete Works*, ed. J.M. Cooper (Indianapolis: Hackett).
- Roger Marston (1932). *Quaestiones disputatae* (Florence: Collegium S. Bonaventurae).
- Thomas Aquinas (1975). *Summa contra gentiles*, tr. A.C. Pegis et al. (Notre Dame, IN: University of Notre Dame Press).
- ----- (1947-48). *Summa theologiae* tr. L. Shapcote (New York: Benzinger).
- ----- (1954). *Truth* [=*De veritate*], tr. R.W. Mulligan et al. (Chicago: Henry Regnery).
- ----- (1987). *Faith, reason and theology* [=*In de trinitate* 1-4], tr. A. Maurer (Toronto: Pontifical Institute).
- William of Auvergne (1674). *Tractatus de anima*, in *Opera omnia* (Paris; rpt. Frankfurt a.M.: Minerva, 1963).

Secondary Literature

- Bonjour, Laurence (1998). *In Defense of Pure Reason* (Cambridge: Cambridge University Press).
- Brentano, Franz (1992). "Nous Poiêtikos: Survey of Earlier Interpretations," in M. Nussbaum and A. Rorty (eds.) *Essays on Aristotle's De Anima* (Oxford: Clarendon Press) 313-41.
- Fakhry, Majid (1983). *A History of Islamic Philosophy* (New York: Columbia University Press).
- Gilson, Étienne (1926-27). "Pourquoi Saint Thomas a critiqué Saint Augustin," *Archives*

D'Histoire Doctrinale et Littéraire du Moyen Age 1: 5-127.

- ----- (1930). "Réflexions sur la controverse: S. Thomas - S. Augustin" in *Mélanges Mandonnet* v.1 (Paris: Vrin).
- ----- (1933). "Roger Marston: Un cas d'Augustinisme Avicennisant," *Archives d'Histoire Doctrinale et Littéraire du Moyen Age* 8: 37-42.
- ----- (1955). *History of Christian Philosophy in the Middle Ages* (New York: Random House).
- Gersh, Stephen (1978). *From Iamblichus to Eriugena: An Investigation of the Prehistory and Evolution of the Pseudo-Dionysian Tradition* (Leiden: Brill).
- Haldane, John (1992). "Aquinas and the Active Intellect," *Philosophy* 67: 199-210.
- McPherran, Mark (1996). *The Religion of Socrates* (University Park, Pa.: Pennsylvania State University Press).
- Nasar, Sylvia (1998). *A Beautiful Mind* (New York: Simon and Schuster).
- Pasnau, Robert (1995). "Henry of Ghent and the Twilight of Divine Illumination," *Review of Metaphysics* 49: 49-75.
- Rohmer, Jean (1928). "La théorie de l'abstraction dans l'école franciscaine, de Alexandre de Halès à Jean Peckam," *Archives d'Histoire Doctrinale et Littéraire du Moyen Age* 3: 105-84.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aquinas, Saint Thomas | Olivi, Peter John

Acknowledgements

This entry has benefitted greatly from suggestions by Marilyn Adams, John Boler, Neil Lewis, Tim Noone, Sarah Pessin, Chris Shields, and John Wippel.

Robert Pasnau
pasnau@colorado.edu

First published: November 2, 1999
Content last modified: November 2, 1999

Peter John Olivi

Peter John Olivi must be considered one of the most original and interesting philosophers of the later Middle Ages. Although not as clear and systematic as Thomas Aquinas, and not as brilliantly analytical as John Duns Scotus, Olivi's ideas are equally original and provocative, and scarcely known even to specialists in medieval philosophy.

1. Life and Work

Olivi (1248 - 1298) was born in Sérignan, in the Languedoc region of southern France. He entered the Franciscan order as a youth, studied in Paris without becoming a master of theology, and spent his life teaching at various Franciscan houses of study in southern France. (For biographical details, see Burr 1976, 1989.) Olivi's outspoken originality led him into conflict with religious authorities: his writings were condemned and burned twice, and this obviously limited the influence he would have on posterity. Although his philosophical views were controversial, what drew the most attention was his attacks on Church authorities. As an early leader of the so-called "Spiritual" reform movement within the Franciscan order, Olivi attracted both fervent followers and implacable enemies (see Burr 1989, 1993; Douie 1932).

Olivi produced a large and wide-ranging body of work, most of which has survived. Almost all of his philosophical views are contained in his (as yet untranslated) question-commentary on the second book of Peter Lombard's *Sentences*. This lengthy, carefully composed treatise takes up a wide range of

philosophical problems, such as the nature of matter, form, substance, quantity, the soul, causation, motion, and the problems of universals and individuation. In this article I will focus on just a few of Olivi's interesting philosophical views, summarizing material I have published elsewhere (see Pasnau 1993, 1997a, 1997b, 1999).

2. Human Freedom

Olivi devotes several extended questions to the topic of human freedom, beginning with the question of whether human beings even have free will (*liberum arbitrium*). Olivi's own argument for the affirmative begins by listing seven pairs of attitudes (*affectus*), each of which testifies to the existence of free will (Q57, p. 317):

1. Zeal and mercy
2. Friendship and hostility
3. Shame and glory
4. Gratitude and ingratitude
5. Subjugation and domination
6. Hope and distrust
7. Providence and negligence

Each of these attitudes, Olivi claims, is intelligible only given the existence of free will. More specifically, they are "its distinctive products, or its distinctive acts and habits" (ibid.). As he runs through the list, explaining how each attitude entails free will, it becomes clear that many of these claims are familiar ones. Zeal, for instance, is an angry reaction to bad deeds, motivated "only against the bad that one judges to have been done voluntarily, and thus which could have been freely avoided" (p. 318). Without free will, this attitude is based on an assumption that is "thoroughly false and grounded on a thoroughly false object" (p. 317). As zeal goes, so do the related phonemona of accusations, excuses, blame, and guilt. Generally, "a human being could no more be accused of some vice than he could be accused of death, for he could avoid the one as little as the other" (p. 336). Providence and negligence, the last pair on the list, likewise become meaningless : "For it is foolish to be careful about things that will occur necessarily" (p. 323). It becomes pointless to be careful about deliberation, for instance, "because the deliberation itself will or will not happen necessarily, and even one's carefulness will or will not occur necessarily" (p. 323).

For Olivi, these and other data stand as unshakeable evidence for the existence and nature of free will. He makes this clear from the beginning of his reply, when he introduces two premisses that "no one of sane mind ought to doubt" (p. 317). First, it is "impossible" for all of the attitudes of one's rational nature to be "thoroughly false and perverse and grounded on a thoroughly false and perverse object." Since Olivi thinks that the attitudes that distinguish us as rational creatures are founded on free will, giving up on free will would be to abandon most of what makes us human. We would become, he later says, "intellectual beasts" (p. 338). Second, it is impossible for attitudes to be entirely illusory when human beings improve and perfect themselves by assuming those attitudes (p. 317). If the practices of zeal,

deliberation, friendship, love, etc. were all founded on a false assumption, then surely these practices would not be so crucial to human well-being. Thus "no one of sane mind will believe that something could be the truth which so sharply puts an end to all good things and brings on so many bad things" (p. 338). In the face of these implications, we should reject whatever stands in the way of free will, whether that be the authority of Aristotle or some abstruse principle of metaphyics. "Even if there were no other argument establishing that [the denial of free will] is false, this alone ought to be sufficiently persuasive" (p. 338). Moreover, as he explicitly notes, we should be persuaded not just of our own free will, but of the free will of all human beings, since these arguments are based not on private experience, but on our relationships with others.

This top-down approach, beginning with the ethical and experiential data, leads Olivi to some provocative conclusions about the nature of will. He does not merely conclude that the will's choices are not necessitated; the further conclusion Olivi reaches is that the will, until it makes a choice, is entirely undetermined one way or another, and that it determines itself in the direction it chooses. This is something "every human being senses with complete certainty within himself" (p. 327). In arguing that the will determines itself, he means that it is a first mover, in need of no efficient cause other than itself. "Its free power is the cause of its motion, when it is moved, and the cause of its rest, when it rests" (ad 5, pp. 341-42). If the will did not have this capacity for self movement, then the will would have to be determined by something else, hence it would not be making its own choices. But this violates the unshakeable assumptions from which Olivi begins, because it would then turn out that the will is not autonomous, not making its own choices, hence not a suitable object of one's zeal or friendship, among other things.

Olivi is well aware that the lack of autonomy does not entirely preclude a sort of pseudo-zeal or pseudo-friendship. One might be angry with someone, for instance, not out of the conviction that the bad action was that person's fault, but simply in an effort to change that person's ways. But this line of thought does violence to our conceptions of ourselves and our fellow human beings. We want people to do the right thing not because they have been effectively manipulated, but "solely and purely because of the love of justice" (ad 22, p. 368). Further, when we urge a person to do the right thing, "we do not intend simply to move someone toward what is good, but rather to make it that he voluntarily moves himself toward the good" (p. 369).

Olivi's view obviously belongs within the libertarian camp in the free will debate. Indeed, it is arguable that Olivi deserves credit as the founder of this view. Although John Duns Scotus is better known as an early proponent of libertarian freedom, Scotus's views seem heavily indebted to Olivi (see Dumont 1995). An English translation of Olivi's writings on human freedom would be of enormous value.

3. Soul and Body

With the rediscovery of Arisotle's metaphysical and ethical works, thirteenth-century theologians devoted an increasing portion of their time to interpreting and developing Aristotelian accounts of human nature. Olivi was very far from a slavish admirer of Aristotle's -- indeed, he was if anything rather hostile to the

Philosopher's pervasive influence, once remarking that "without reason he is believed, as the god of this age" (Q58 ad 14, p. 482; see Burr 1971). This hostility manifests itself in many ways, one of the most notable being his attempt to rework Aristotle's account of the soul-body relationship.

Olivi questions a central strand of the Aristotelian account, arguing that it is "not only contrary to reason but also dangerous to the faith" to hold that "the [soul's] intellective and free part is the form of the body per se and considered as such" (Q51; p. 104). Others had questioned the extent to which soul and body could be analyzed in terms of form and matter. But Olivi goes farther because he explicitly denies that one part of the soul, the rational part, can be understood as the form of the body. As a consequence, he was condemned by the Council of Vienne in 1312. Pope Clement V, in the bull *Fidei catholicae fundamento,* declared it a heresy to hold that "the rational or intellective soul is not per se and essentially the form of the human body" (Denzinger 1965, n. 902).

Olivi denies that "the [soul's] intellective and free part is the form of the body per se and considered as such" (Q51; p. 104). This matches fairly closely with the doctrine that was condemned. But it is easy to misunderstand what Olivi is saying. First, he is not denying that the rational part of the soul is a form, or even that it is the form of a human being. He in fact believes that the rational part, intellect and will, is the form of a human being's spiritual matter, and for that reason he thinks it acceptable to speak of intellect as the form of a human being (see Q51 appendix, p. 146). But he contrasts spiritual matter with corporeal matter, and as a result he denies that the rational part is the form of the body.

Second, Olivi is not denying that the soul is the form of the body. What he denies is that the rational part of the soul ("the intellective and free part") is the form of the body. Another part of the soul, the sensory part, is the form of the body, and for that reason it is acceptable to say that the whole soul is the form of the body:

> It is said that the whole rational soul, rather than the sensory part, is the form of the body, even though it is informed by the whole only insofar as it is informed by the soul's sensory and nutritive part (Q51 app., p. 146).

We should say that the whole soul is the form of the body, in much the same way that we say a person talks, not a tongue (p. 144). But if we direct our attention to the various parts of the soul, then it is wrong to say that the rational part, "per se and considered as such," is the form of the body. The soul is the form of the body only with respect to its sensory and nutritive part.

Despite his anti-Aristotelian invective, Olivi might plausibly be said to be agreeing with Aristotle, who explicitly leaves room for parts of the soul that "are the actuality of no body" (*De an.* II 1, 413a7). Presumably, Aristotle is thinking of the intellect. But it's not at all obvious how that remark should be interpreted. Aquinas, for instance, holds without qualification that "the intellect... is the form of the human body" (*Summa theologiae* 1a 76.1c). So what is it that compels Olivi to drop intellect from the hylomorphic scheme?

Olivi writes that to identify the rational part -- "per se and considered as such" -- as the form of the body is "not only contrary to reason but also dangerous to the faith" (as above). More specifically, as he writes in a letter defending his views, he believes that that claim holds "the danger of destroying the soul's immortality, its liberty, and its intellectual nature" (*Epistola* n. 7). Each of these three consequences is based on one overarching assumption: that to make the soul's rational part the form of the body is to attribute to the body the distinctive capacities of the rational soul. Here is how Olivi puts that claim:

> If the intellective part is the form of the body then, since all matter is actualized by its form, it follows that just as a human body is truly sensory and living through the sensory soul, so that body will be truly intellective and free through the intellective part (Q51; pp. 104-5).

If the intellect is the form of the body, then the body must have the capacities for intellectual thought and free decision. Olivi is of course going to reject that as absurd. Notice the form of this argument. First, Olivi asserts that to be the form of something is to impart actuality to that thing. This seems uncontroversial. Second, Olivi argues by analogy. Just as the sensory soul actualizes a body by giving it life and the capacity for sensation, so the intellective part -- if it is the form of the body -- should actualize the body by making it intellective and free. Here too, I think, Olivi's claim seems plausible. If one accepts the first step of the argument, that to be the form of something is to impart actuality to that thing, then the rational part must be giving *something* to the body. Olivi says, "every form imparts to its matter some operation, and some power for operating" (Q51; p. 109). So, if the rational part does not give the body the capacity for intellective thought, we have to provide some sort of account of what the rational part does give the body. But what else could the rational part of the soul do for the body, if not endow it with the power to be rational?

We might view this argument as posing a dilemma. If the rational part is the form of the body, then one must either understand this formal relationship in the ordinary way, in terms of actualizing the body, or one must concede that the rational part is not the form of the body in any ordinary sense. The first horn of the dilemma leads in the direction of materialism, because it forces one to claim that the powers of the rational soul are instantiated in the body. The second horn of the dilemma leads one toward retracting the original assertion: that the rational part, the intellect, is the form of the body. For it is not at all clear what that means, if intellect is not in any way actualizing the body.

4. Cognitive Activity and Attention

One of the most interesting and original aspects of Olivi's philosophy is his critique of the standard Aristotelian model of cognition. The starting point of this critique is his insistence that sensation and intellection are active, not passive. On the conventional medieval view, a cognitive power simply receives impressions from the world, in the form of sensible or intelligible species. Olivi argues that such an account leaves out a crucial element, the focusing of the cognitive power's attention on the object to be cognized.

> However much the cognitive power is informed through a disposition and a species differing from the cognitive action, it cannot advance to a cognitive action unless before this it actually tends toward the object, so that the attention of its intention should be actually turned and directed to the object (Q72, p. 9).

Olivi gives the kinds of examples that one would expect. The ears of someone sleeping, for instance, receive the same impressions as the ears of someone awake, but the sleeper does not sense these impressions. Even when we are awake, we sometimes don't perceive objects right in front of us when we are intently focused on something else (Q73, pp. 89-90).

Olivi argues that this kind of cognitive atttention requires a "virtual extension" toward the object. One striking consequence of this claim is that the object itself needn't exert any causal influence, not on the cognitive faculties nor even on the physical sense organs. The external object need only be close enough to be apprehended by the cognizer's spiritual attention. In the cases of both sensation and intellection, the efficient agent is the cognitive power. The external object is merely a kind of final cause or, more precisely, a "terminative cause." (Q72, p. 36; *Epistola*, n. 12). It is merely by being the object of the cognitive power's attention that the external object plays a role in cognition. Though Olivi accepts the traditional theory of *species in medio*, sensible qualities that fill the air between the senses and their objects, he denies that these species are the efficient cause of cognition.

Is a virtual extension, as Olivi describes it, some special (perhaps nonphysical) but perfectly real kind of extension or extromission to external objects? The bulk of the evidence seems to show that Olivi means ‘virtual’ and ‘virtually’ to contrast with ‘real.’ He explicitly denies, for instance, that this virtual extension involves "any real emission of its essence" (Q73, p. 61). Elsewhere, considering the claim that "our mind is where it fixes its intention," he says that "these words are metaphorical. For we are not there really or substantially, but only virtually or intentionally" (Q37 ad 13, p. 672).

Olivi treats virtual attention not as a sui generis activity of the mind, but as a general kind of causal relationship that can be applied to physical agents just as much as to mental ones. For Olivi, every natural physical agent has a virtual attention of this sort that extends as far as its causal force does (Q23, pp. 424-25). One authority comments that Olivi's virtual attention is "in fact equivalent to action at a distance" (Jansen 1921, p. 118), a characterization that seems just.

Olivi allows that the object itself, through *species in medio*, can indirectly act on our spiritual faculties, through what he calls the *via colligantiae* (way of connection). A flash of lightning will make a physical impression on our eyes, and this physical impression can, through the *via colligantiae*, affect the spiritual sensory powers. But, crucially, this connection is not what brings about sensation. We *see* this flash, as opposed to receiving merely a physical impression from it, when we direct our spiritual attention toward it (*Quodlibet* I.4). This *via colligantiae* plays an important role across Olivi's philosophy of mind and the will, being his general method of explaining the vexed connection between mind and body (see Q59, pp. 546-54, and Jansen 1921, pp. 76-90).

5. Direct Realism

Olivi's direct realism is central to his thinking about cognition. If he were willing to say that the object of our spiritual attention is not the external object but an internal species of the object, then he could reformulate his theory of cognitive attention in a more plausible way, as a matter of grasping an internal impression from the object. But Olivi works very hard to avoid falling into any kind of position that might be called representationalist -- that is, a view on which the immediate objects of cognition are internal. It is this direct realism, above all else, that leads Olivi to reject the standard scholastic account of sensible and intelligible species. On that standard account, species serve as forms that provide the intentional content of sensation and thought. Although these forms were standardly described as merely the means by which we grasp external things, Olivi argued that in fact the proponent of species was committed to representationalism.

Olivi argues against the species theory by advancing through a series of ever-more-serious charges. First, the theory is committed to taking species as the objects of cognition:

> A species will never actually represent an object to the cognitive power unless the power attends to the species in such a way that it turns and fixes its attention on the species. But that to which the power's attention is turned has the character of an object, and that to which it is first turned has the character of a first object. Therefore these species will have the character of an object more than the character of an intermediate or representative source (Q58 ad 14, p. 469; cf. Q74, p. 123).

His argument for this conclusion turns on the first sentence of the passage, in which he claims that a species could not represent an object to a cognizer unless the cognizer attends to the species. Olivi takes this attention to the object to be both a necessary and a sufficient condition for a cognition of that object. So if we do have to focus our attention in this way on species, he infers that those species will be the object of cognition, not merely causal intermediaries.

Next, Olivi argues that species would have to be the first object of cognition. To turn toward a species in the way that we must if that species is to represent the external world "is the same as to attend to it as a first object" (Q74, p. 123). Elsewhere, "we would always cognize the species before the thing itself that is the object" (Q58 ad 14, p. 469). The point Olivi wants to make is one more often made by denying that the world is seen *directly* or *immediately*. If we see the external world at all, we see it only at second hand, indirectly.

The argument goes one final step. Someone who wants to claim that our internal sensations are themselves perceived has to choose whether or not to claim that the external world is also perceived. Olivi takes it that it is not; on the species account, we would not perceive the external world at all, only images of it:

> The attention will tend toward the species either in such a way that it would not pass beyond so as to attend to the object, or in such a way that it would pass beyond. If in the first way, then the thing will not be seen in itself but only its image will be seen as if it were the thing itself (Q74, p. 123; cf. Q58 ad 14, pp. 469-70, 487-88).

The argument is based on a dilemma. Granting that cognizers must attend to species, there either will or will not be a separate and further attention to the object itself. It would of course be quite odd to say that there is such a further attention. This would entail, as Olivi goes on to say, that one "considers the object in two ways -- first through a species, second in itself" (Q74, p. 123). This seems too much at odds with the phenomenal feel of perception to be a serious possibility. The obvious way out of the dilemma, then, is to say that there will not be any further attention: one apprehends the external world, if one does at all, in virtue of attending to the species themselves. This is what the representationalist will likely say. But if this is the case, Olivi argues, then we won't be seeing the things in themselves but only their images. Memorably, he remarks that a species "would veil the thing and impede its being attended to in itself as if present, rather than aid in its being attended to" (Q58 ad 14, p. 469).

In place of the species theory, Olivi offers an interesting alternative. Rather than treat mental representations as something separate from an act of cognition, Olivi proposes identifying the two. On his view, an act of cognition is itself a representation of the object perceived. There is no need to postulate any further representation beyond the act itself: that inevitably results in the mediation that Olivi wants to avoid. This act theory would prove influential on later scholastics, most notably William Ockham. And in our own era it has been reinvented and renamed, as the adverbial theory of thought and perception.

6. Word and Concept

Olivi extends his critique of species to the mental word (*verbum*), which was standardly postulated as the product of intellectual thought. His treatment of the *verbum* raises different issues from those associated with species. Here the issue is not direct realism, precisely, but rather the nature of concept formation. Near the start of his commentary on the Gospel of John, Olivi describes the standard view as follows: "Our mental word is something following an act of thought... and formed by that thought.... After it has been formed... the [extra-mental] object is clearly understood or viewed in that word as if in a mirror" (*Tractatus de verbo* 6.1). This word, moreover, "is that which is first cognized by intellect and is its first object;" the extra-mental object is cognized secondarily. This description closely matches a characterization Olivi gives in his later *Sentences* commentary:

> Some maintain that a kind of concept, or word, is formed through an abstractive, investigative, or inventive consideration, in which real objects are intellectively cognized as in a mirror. For they call this the first thing understood, and the immediate object; it is a kind of intention, concept, and defining notion of things (Q74, pp. 120-21).

This view has two characteristic features. First, its postulates a mental representation -- a concept or word

-- that is the product of intellectual activity. Second, it supposes that we understand the world through these representations, in such a way that we get at the world indirectly, or secondarily, "as if in a mirror." Call this an object theory of the *verbum*.

Olivi's own view is that the *verbum* should be identified with a particular act of thought: "our mental word is our actual thought" (*Tractatus* 6.2.1). When we engage in abstract intellectual cognition, Olivi says, "nothing serving as an object is really abstracted or formed that differs from the act of consideration already mentioned" (6.2.3). The *Sentences* commentary offers a concise characterization:

> This [sort of intervening concept] ought not to be called a *verbum*, nor can [such a concept] be anything other than the act of consideration itself or a memory species formed through that act (Q74, p. 121).

There are, then, acts of intellect, but there are no separate inner concepts that are the objects of those acts. Call this an *act theory* of the *verbum*.

Why is this act theory superior to an object theory? One line of argument holds that the object theory "contains in itself obvious absurdities and thus contradicts sound reason" (*Tractatus* 6.2.2). This claim is argued in different ways, with the following dilemma often playing a crucial role: On one hand the *verbum* is said to be the product of intellectual cognition. But on the other hand the *verbum* is said to be required for cognition as the "first thing understood." How can it be both? Olivi thinks his opponents will have to maintain that in some way the *verbum* is the product of one act of intellect and the object of a second. This leads him to argue that his opponents are treating the *verbum* as merely a memory. But Olivi is happy to countenance representations of this sort. Thus the object theory collapses into the act theory.

The second line of attack holds that the theory lacks support because "there is no necessity or utility in postulating such a *verbum*" (6.2.3). Here Olivi considers two parallel lines of argument that a proponent of the mental word might make against this charge of superfluity.

> First,... we experience in ourselves that we form in our mind new concepts of many propositions and conclusions. These concepts remain in us later and we return to them when we want to remember such propositions.... Second,... from individuals seen or imagined by us we abstract and form defining characterizations of their universal features,... and we come back to these when we wish to view such universal features (6.2.3).

Each argument appeals to our experience of forming within ourselves abstract ideas: in the first case propositional ideas, in the second universals. Intellect in each case is said to form a *verbum*. Olivi replies that no such inner word is necessary. In each case we have an act of conceptual thought, but no object is formed in intellect over and above the act of thinking itself. Indeed, if anything, such an object "would be an impediment" (6.2.3) -- alluding to the epistemological difficulties discussed in the previous section.

By eliminating the representations that might intervene between intellect and external reality, Olivi gives us what we might be tempted to think of as a direct realist theory of intellectual cognition. Yet direct realism faces a serious problem at the intellectual level, a problem that Olivi's discussion fails to acknowledge. Direct realism is attractive as a theory of sensation because it seems clear what the objects of sensation are. But what are we directly in touch with when our intellect thinks abstractly or propositionally? One answer to this question is Platonism: universals and/or propositions have some kind of abstract mode of existence, independently of the human mind. Like almost all the scholastics, Olivi firmly rejects this kind of account (Q13). Another kind of answer, sometimes called conceptualism, treats universals and/or propositions as mental constructs. Defenders of the object theory can take this approach. They can hold that although there are no universals or propositions *in re*, there are universals and propositions *in mente*. The *verbum*, serving as universal or as proposition, will (in some cautiously described sense) be the object of thought.

Olivi's act theory would seem to rule out this kind of conceptualism. But what then will Olivi put in its place? He speaks of intellect's "attending to and considering the real character of a common or specific nature" (376-379), as if he has an unproblematic account of intellect's relationship to the outside world. Yet he says nothing to clarify the status of this relationship. He simply does not seem to have recognized the problem of abstract knowledge as a fundamental metaphysical motivation for the object theory. In this respect his overall account, although conceptually innovative, remains fundamentally incomplete.

Bibliography

Texts

- *Quodlibeta* (Venice, 1509).
- *Quaestiones in secundum librum Sententiarum* (Bibliotheca Franciscana Scholastica 4-6), ed. B. Jansen (Quaracchi: Collegium S. Bonaventurae, 1922-26).
- *Quaestiones logicales*, ed. S. Brown, *Traditio* 42 (1986) 335-88.
- "Petri Iohannis Olivi Tractatus de verbo," ed. R. Pasnau, *Franciscan Studies (Essays in Honor of Fr. Gedeon Gál)* 53 (1993) 121-53. [Excerpt from the Commentary on the Gospel of John. Published in 1997.]
- "Petrus Ioannis Olivi. Epistola ad fratrem R.," ed. C. Kilmer and E. Marmursztejn, *Archivum Franciscanum Historicum* 91 (1998) 33-64.
- "The Mental Word" (= *Tractatus de verbo*), in R. Pasnau (tr.) *Cambridge Translations of Medieval Philosophical Texts. Volume 3: Mind and Knowledge* (Cambridge: Cambridge University Press, 2002), 136-51.

Secondary Literature

- Bettoni, Efrem (1959). *Le dottrina filosofiche di Pier di Giovanni Olivi* (Milan: Vita e Pensiero).

- Burr, David (1971). "Peter John Olivi and the Philosophers," *Franciscan Studies* 31 41-71.
- -----, (1976). "The Persecution of Peter Olivi," *Transactions of the American Philosophical Society* 66:3-98.
- -----, (1989). *Olivi and Franciscan Poverty: The Origins of the Usus Pauper Controversy* (Philadelphia: University of Pennsylvania Press).
- -----, (1993). *Olivi's Peaceable Kingdom: A Reading of the Apocalypse Commentary* (Philadelphia: University of Pennsylvania Press).
- Denzinger, H. (1965). *Enchiridion symbolorum* (Herder: Freiburg).
- Douie, Decima (1932). *The Nature and Effect of the Heresy of the Fraticelli* (Manchester: Manchester University Press).
- Dumont, Stephen (1995). "The Origin of Scotus's Theory of Synchronic Contingency," *Modern Schoolman* 72: 149-67.
- Gieben, S. (1968). "Bibliographia Oliviana (1885-1967)," *Collectanea Franciscana* 38:167-95.
- Jansen, Bernhard (1921). *Die Erkenntnislehre Olivis* (Berlin: Duemmlers).
- Pasnau, Robert (1993). "Petri Iohannis Olivi Tractatus de verbo" in *Franciscan Studies (Essays in Honor of Fr. Gedeon Gál)* 53: 121-53. [Published in 1997].
- ----- (1997a). *Theories of Cognition in the Later Middle Ages* (Cambridge: Cambridge University Press).
- -----, (1997b). "Olivi on the Metaphysics of Soul," *Medieval Philosophy and Theology* 6: 109-32.
- -----, (1999). "Olivi on Human Freedom" in *Pierre De Jean Olivi (1248-1298)* (Paris: Vrin) 15-25.
- Putallaz, François-Xavier (1991). *La connaissance de soi au XIIIe siècle. De Matthieu d'Aquasparta à Thierry de Freiberg* (Paris: Vrin).
- -----, (1995). *Insolente liberté. Controverses et condamnations au XIIIe siècle* (Paris: Cerf).

Other Internet Resources

- David Burr's Olivi page
- Translation of II Sent. Q72, concerning cognitive attention and the active nature of sensation.
- Translation of II Sent. Q74, against sensible and intelligible species.

Related Entries

divine illumination

Robert Pasnau
pasnau@colorado.edu

Table of Contents

First published: November 2, 1999
Content last modified: June 12, 2002

Medieval Theories of Practical Reason

Medieval theories of moral reasoning have their origins in the moral theology of St. Augustine and the rational ethics of Aristotle. Until the thirteenth century Augustine's responses to questions concerning free will, predestination, the nature of goodness and divine freedom dominated moral speculation in the Latin West. For Augustine morality demands the human will's conformity to the prescriptions of the immutable, necessary and eternal law. Augustine argues in his work on free will that the eternal law "is called supreme reason, which must always be obeyed, and through it the evil deserve an unhappy life and the good a blessed life; and through this law we have derived temporal laws rightly constructed and correctly emended." The ideals of eternal law are universally imprinted upon human intellects and are the immutable standards by which human actions may be judged.

The Thomistic Doctrine of Practical Reason

When one chooses through one's free will to live ‘honorably and rightly’ in accordance with divine law, one can reasonably be thought to live a moral life. Despite the human ability to reason according to divine principles, the human condition does not permit the attainment of moral perfection through natural means alone. Augustine asserts that only through grace, sent freely by God to assist the human will can one achieve true moral goodness. Prudence, which is the ability to choose good and avoid evil, intellectual contemplation, moral and political virtue, friendship, education and character (all essential elements in Aristotle's ethics) are subsumed in Augustine's moral theology under the command to love God. For Augustine the complexity of Greek moral thought can be reduced to the simple rule of conformity to divine law.

Despite the contributions of Anselm in the eleventh century on questions concerning free choice, divine foreknowledge and predestination, and Peter Abelard's startling assertion in the twelfth century that morality arises from the agent's intention alone, it is not until the thirteenth century that a scientific

approach to human moral reasoning takes shape. Aristotle's Nicomachean Ethics, newly translated into Latin, provide a philosophical basis on which a fresh examination of Augustine's doctrines could be based. The first great medieval commentary on Aristotle's ethics, which was the result of Albert the Great's teaching activity at the Dominican House of Studies in Cologne, marks the beginning of 'moral science' in the Middle Ages. Albert's careful exposition of Aristotle's text and his clarification of the concepts of natural law, moral reasoning and human virtue, and its influence on his most famous pupil, Thomas Aquinas, led directly to a consideration of the question of practical reasoning in the Middle Ages.

The notion of medieval practical reason can be investigated in two ways: 1) in light of the distinction between practical and theoretical sciences in the writings of the medieval university masters in the thirteenth and fourteenth centuries; 2) in comparison to the modern understanding of practical reason as described by Immanuel Kant. The first method allows for a strict adherence to the texts themselves, wherein the concept of 'ratio practica' is strictly limited to a type of philosophical reasoning. The second approach permits a deeper philosophical consideration of the parallels between the medieval understanding of the nature of moral goodness and the modern view of the will as practical reason. Both approaches are helpful in gaining a fuller understanding of the meaning of practical reason in the Middle Ages.

In its strictest sense the term 'practical reason' (*ratio practica*) refers specifically to a type of reasoning, which is analogous to the deductions of speculative or theoretical science. When discussing human knowledge of the precepts of natural law Thomas Aquinas argues that these commands are related to practical reasoning as the first principles of demonstration are related to speculative reasoning; in each science there are certain principles of demonstration which are known in themselves (*principia per se nota*). These principles, which comprise the universal laws of moral behavior direct all subsequent moral reasoning. Thomas bases his theory of correct reasoning on the human ability to discover an underlying order in any field of inquiry. The order of reasoning which determines metaphysical knowledge is derived from a recognition of the principle of non-contradiction. Reasoning from the notion that being and non-being are contradictory terms, a metaphysician argues for certain conclusions about the nature of being. In practical reason one begins with the principle that all human acts are directed to an end and then comes to recognize the fundamental element of ethics to be 'do good and avoid evil':

> Therefore the first principle in practical reason is that which is based on the nature of the good which is: the good is that which all things seek. This therefore is the principle of law: that good must be done and evil avoided. And on this <precept> all other precepts of natural law are based so that everything which is to be done or avoided pertains to the precepts of natural law. Practical reason naturally understands these precepts to be human goods. (S. th. I-II, 94, 2).

Although Albert sees a closer connection between prudence and practical reason than Thomas ("prudence and practical reason have the same acts in that reason gives the act while prudence informs the act by reason of justice, expediency and honesty." De bono # 443), Albert essentially agrees with Thomas' description of the method of practical reasoning (Super Ethica, VI, 7, pp. 436-437).

Thomas argues from the basic principle that what is a good always has the nature of the end to the conclusion that human beings seek to discover particular good acts as consequences of the determined end. Since practical reason mimics the deductive process of theoretical reasoning, the term, '*ratio practica*', primarily refers to a type of human knowledge. There is, however, a fundamental difference between the conclusions of theoretical and practical science:

> Because speculative reason is especially concerned with what is necessary and cannot be otherwise, the truth found in its conclusions is without flaw, just <the truth> in its general principles. But practical reason concerns what is contingent whose domain is human acts; and so even if there is some necessity in its general <principles>, the more one descends to its proper conclusions the more one finds a defect <in truth>. In speculative reasoning the truth is the same for all, both in principles and in conclusions. In operative reasoning there is not the same truth or practical rectitude according to its proper <conclusions>, but only according to its common principles. (S. th. I-II, 94, 4; Super ethica, VI, 7, p. 441).

Since the goal of practical reason is action not knowledge, the truth attained by the intellect must be caused by its conformity to right desire. There can be no necessary science of practical reason, since virtuous activity allows for variety and derivation from the universal rule in particular instances. Moral matters, which are within the domain of practical reason, are varied and inadequate (*deformis*), and cannot therefore provide the certainty which is expected in theoretical reasoning. If we are to have any science of practical reason at all, we must be content to apply the principles to various conclusions and proceed from rough arguments which demonstrate truth in a general way.

The variety and differences among the will's acts lead Thomas Aquinas to use the term 'ratio practica' specifically to distinguish the method of moral reasoning from that of strictly scientific knowledge:

> Therefore one finds something in practical reason which is related to operations just as a proposition in speculative reason is related to conclusions. (S. th. I-II, 90, 1 ad 2).

Albert argues that despite the similarities in method between the two types of reasoning, the force of a practical conclusion depends more on a particular desire than a universal principle. Since human desires differ so greatly, moral arguments are merely general and imperfectly formulated. (Super ethica, VI, 16, p. 491). These distinctions which are derived from Aristotle's analysis of the nature of human knowledge, do not approach the understanding of science found in the philosophy of Kant. For Kant, speculation does not, and cannot, give us knowledge of being, nor should we begin with being as the object of science, as is seen in the medieval principle of existential non-contradiction. Kant restricts speculative science to a consideration of the laws of appearance. Practical science, or practical reason, in Kant's philosophy, however, is concerned with freedom rather the natural good or the human end. Kant's practical reason attains the kind of rigor, based on pure *a priori* principles, which would be impossible in the ethics of Aristotle, Albert and Thomas who are content with a science of practical reason whose domain is contingent and changeable acts. The secure foundation which Kant sought for every moral

choice would be considered beyond the scope of practical reason as formulated by the medieval commentators on Aristotle's ethics.

Despite their enthusiastic reception of Aristotle's works, the medieval moralists did not grant him complete authority in moral reasoning. The flexibility of Aristotle's ethics, which Thomas himself acknowledges, did not lead medieval writers to construct a moral theory which, like Aristotle's, is based on societal norms, tradition and human actions. The medieval moralists sought a more secure foundation for determining ethical action than Aristotle's appeal to the man of practical wisdom (*phronimos*). A modern author recognizes the tensions that exist in medieval ethical theory when he writes of Thomas Aquinas: ‘How is it that Aquinas can seem so Aristotelian in his description of human action and yet be so Augustinian in his insistence on the need for conformity to the eternal law?’ (Westberg, 34). The answer to this question lies the account of the genesis of the human moral act and the conditions for its rectitude. Practical reason requires a foundation more secure than the accepted practice of human actions. In the search for that basis, Thomas and his contemporaries construct a theory of practical reason far more complex than the mere designation of a type of intellectual reasoning; it becomes an account of the nature of moral goodness itself.

The designation of the will as rational appetite, its end as goodness and its relationship to the intellect are well-known features of medieval followers of Aristotle and need not be treated at length here. What is of more interest in a discussion of practical reason is the analysis of those first principles which regulate moral reasoning. Thomas in developing the ideas of his former teacher, Albert, insists that the will must be moved by an end that is perceived as good. (S. th. I-II, 6, 1 & 8, 1; Super ethica, VI, 7, pp. 436-437). More specifically he argues that:

> ... the good in common, which has the nature of the end, is the object of the will. Therefore because of this element the will moves the other powers of the soul to their acts. (S. th. I-II, 9, 1)

Even though the intellect moves the will by presenting the object to be desired, the will itself has a natural inclination toward the good. The ultimate human end, or supreme good, is beatitude, the perfect, all-encompassing human good. Such a supreme good could never be perceived by practical reason as evil. (S. th. I-II, 13, 6 & 10, 2 ad 3; Super ethica, VI, 17, p. 497). The determination of the human end as beatitude says very little about the first principles of practical reason upon which specific moral judgments should be based. When the will desires, it seems to want more specific objects than the vague longing for beatitude. Thomas and Albert specify their theory of volition by means of the doctrines of natural law, synderesis, and prudence.

The source of the first principles in any science is a critical element in determining the validity and nature of that science. In moral reasoning the origin of the principles of actions not only reveals the understanding of the nature of goodness, but directs also all subsequent analysis as well. In Thomas' theory of practical reason the first principle of human actions is that of law:

> Just as reason is the first principle of human acts, so too in this reason something is the principle with respect to every other act. Thus it is necessary for law to pertain principally and most extensively. The first principle in actions, for which there is practical reason, is the ultimate end. The ultimate end of human life is happiness or beatitude. It is necessary for law to reflect in the highest degree that order which leads to beatitude. (S.th. I-II, 90, 2)

Albert formulates this position succinctly when he says, "prudence is regulated by divine and human law." (Super ethica, VI, 4, p. 417). While no other law than the dictates of practical reason guides human choices, the eternal law (*lex aeterna*) primarily and principally orders a human being to the end and determines the corresponding means. As a result, acts that are at odds with the eternal law must always be considered as contrary to the dictates of practical reason. (S. th. I-II, 71, 6 ad 3).

The insistence upon the binding force of eternal law and the natural human inclination towards, and participation in, it (which is called natural law) marks a decisive step away from Aristotle's ethics of *phronesis*. The ground of moral action is no longer thought to be the conformity of conduct to that of an outstanding person (*phronimos*); it is found in an external and universally binding source (eternal law). The process of correct practical reasoning is governed, and measured, by the conformity of acts to the precepts of this law:

> A similar process is found in practical and speculative reason… just as in speculative reason conclusions of diverse sciences are produced from indemonstrable principles naturally known… so too from the precepts of natural law as if from certain common and indemonstrable principles human reason proceeds necessarily to those things to which it should be more particularly disposed. (S. th. I-II, 91, 3)

The recognition of the principles of natural law allows practical reason to demonstrate how a human being naturally participates in eternal law according to its common principles.

Developing ideas found in Albert's work, Thomas argues that the will must be determined by its acceptance of the dictates of practical reason, which necessarily conform to the precepts of natural law:

> that human reason rules the human will, by which its goodness is measured, comes from the eternal law, which is divine reason. Thus the goodness of the human will clearly depends much more on eternal law than on human reason, and where human reason is deficient it is necessary to turn to eternal reason. (S. th. I-II, 19,5; Super ethica, VI, 4, p.417)

The content of the precepts, the way they are known and their influence on volition are the final elements in the natural law theory of practical reason. The assertion that natural law reflects eternal law gives little indication as to its specific precepts, but it does indicate that according to the order of natural inclinations an order of the precepts of natural law exists. There is in a human being, as in all other types of being, a primary inclination to self-preservation. It is, however, in the discussions of synderesis that Thomas and

Albert most clearly identify those principles which are the foundation of human moral reasoning. Synderesis was introduced into Latin by Jerome, perhaps as a variant of the Greek term, '*syneidesis*' (insight), and has no meaning whatsoever in Greek. While Albert compares the moral principles of synderesis to the innate natural seeds of law (*seminaria iuris*), Thomas defines synderesis as "the law of our intellect insofar as it is a *habitus* containing the precepts of natural law which are the principles of human acts." (S.th. I-II, 94, 1, ad 2). In an earlier work Thomas explains synderesis either as a natural habit similar to the habit of principles, or as the power (*potentia*) of reason with such a habit. He sees little difference in these two designations, since each describes the universal natural ability of reason to recognize the first principles of morality. The parallels to speculative reasoning that marked practical reasoning are also a feature of deriving conclusions from the dictates of synderesis. Thomas' claim that the function of synderesis is to recognize universal moral laws leaves little doubt that he understood the first principles of practical reason, natural law and synderesis to be the same:

> Just as there is a certain natural habit of the soul whereby it knows the principles of speculative science, which we call the understanding of the principles, so too in the soul is there a certain natural habit of the first principles of actions, which are the natural principles of natural law; and this habit pertains to synderesis and exists in no other power than reason. (De veritate, q. 16, a. 1)

The specific dictate of synderesis which refers to the eternal law is that one must obey God; the primary imperative with respect to the natural law is that one should avoid evil and seek the good. These principles are obviously not mutually exclusive, 'but the dictate of reason to pursue good is rationally and necessarily derived from the command to obey God.' Synderesis is the ability of reason that never errs in the recognition of those universal rules of moral action, since a denial of their universal validity contravenes human reason. Reason can, however, err in the application of the universal principle to a particular action. Moral wrong is then the result of an imperfect, or false, deduction from the principle. So properly speaking, error is not ascribed to universal principles (synderesis), but rather to conscience which may incorrectly apply a universal judgment. (De veritate, q. 16, a. 2, ad 2; Super ethica VI, 7, p. 441).

The ability to apply correctly the principles of practical reason to specific acts in particular circumstances is the function of the intellectual virtue of prudence, defined succinctly as *recta ratio agibilium*. Prudence "represents the agent's ability to deliberate, decide and properly to order the process of practical reason to action." (Westberg, p. 187). Prudence does not, however, direct the will infallibly to right conclusions. Since it merely directs choices, but does not determine them, the will can be said to remain free. Thomas argues that the will can choose freely in three ways, although it could never express a desire contrary to the primary moral rule of pursuing good. The will can be mistaken 1) with respect to its own act in that it can either will or not will; 2) with respect to its object in that it can either want or not want a particular thing; 3) with respect to what is ordered to the end insofar as it wills a particular good or evil act. (De veritate, q. 16, a. 2, ad 2 & q. 16, a. 1). Thomas' description of the will's freedom seems at times to be overwhelmed by his insistence upon the will's determination by the human intellect. This theory of freedom seems to consist merely in the human tendency for faulty reason, since both Albert and Thomas think it unlikely, or even impossible, for a human being to choose contrary to knowledge of the first

principles and their application to particular circumstances:

> It should be said that the root of liberty is the will as subject, but reason as its cause. The will therefore can be freely drawn to diverse things, since reason can be drawn to diverse things, since reason can have various conceptions of the good. (S. th. I-II, 17, 1)

For Albert the delight that arises from intemperate desire does not corrupt the natural habit of prudence, but rather its rule, when prudence fails in drawing the proper moral conclusion. In other words, no one can act contrary to the universal principles of morality, but only in their particular application. (Super ethica, VI, 7, p. 441).

According to these views human freedom could never lead a human being to act contrary to his own interests. To choose against the principles of natural law would not constitute freedom, but rather foolishness. Although the will's natural inclination to pursue the good presented by the intellect does not compel the will to act, the moralists of the natural law theory think it psychologically impossible to choose something incompatible with a properly deduced conclusion of practical reason:

> A second necessity can be imposed on the will, namely the will must necessarily chose x, if x must be pursued as good or x is to be avoided as evil. (De veritate, q. 17, a. 3)

The Franciscan Critique

The ethics of natural law, with its impressive union of the Augustinian theory of eternal principles and the Aristotelian method of moral reasoning, did not remain long unchallenged. Franciscan theologians, John Duns Scotus and William of Ockham, were especially critical of a theory which they considered too restrictive of human freedom. Although Scotus and Ockham never produced treatises specifically devoted to moral theory, the main lines of their critiques may be sketched from their theological works.

Scotus' main concern in his arguments against the intellectualism of Albert's and Thomas' ethics is his doctrine of the supremacy and freedom of the human will. He claims that there could be no basis for judging an action right or wrong if the will were not free to choose against the dictates of the intellect. The will, even if it should act ‘with reason’ still is able to choose between opposite courses of action that lie within its power; the intellect however, has no power of self-determination, since it must assent to what it recognizes as true. Only the will acts freely, for it has the power of self-determination. Scotus argues that a theory of practical reason in which judgments about actions were restricted to the type of reasoning characteristic of speculation not only restricts freedom, but also removes any basis for merit or blame. If the moral agent must act in accordance with intellectual deduction, then he can only be praised for his intellectual prowess, and not for his moral goodness.

Scotus is influenced by Aristotle's assertion that the end of practical knowledge is truth in agreement with right desire. For Scotus this philosophical expression of the natural law leads him to assert one

fundamental universal moral principle: ‘God should be loved’. This law is so deeply rooted in human reason that even God's power cannot release a human being from its obligations. This principle allows Scotus to view his moral theology as consistent with Aristotle's ethics, since reason leads man to obey God's commands. The will necessarily and perpetually seeks happiness and the will naturally desires its own perfection. The primary universal command informs the will's natural desire for perfection, and so particular actions, regardless of circumstances, are judged in accordance with the will's conformity to the precept to love God.

Scotus' primary consideration as a moral theologian is the nature of freely determined volitional choices. Only secondarily does he consider the goodness of the desired end of the action. When the will freely chooses in accordance with right reason only then can the act be considered morally good. Scotus does have some difficulty explaining the relationship between the will and the intellect. If a human being realizes intellectually that the most desirable goal is union with God, it seems that human intellectual reason would compel him to pursue such an end. If the will does not act in agreement with the rationally derived first principle then it must be necessarily wrong. It is thereby difficult to see how the will's absolute freedom can be maintained.

Faced with such a dilemma, Scotus argues that although the will pursues an object rationally determined by the intellect, this does not mean that the will is conditioned by ‘natural necessity’. The apprehension of a possible action is offered to the will as something neutral, while the will remains always free. (Ordinatio, IV, d. 46). In Scotus' view the natural law is comprised of self-evident *a priori* principles, whose validity the intellect immediately recognizes from the coherence of terms. The will then is naturally inclined to assent to their dictates, but is not compelled to do so. For Scotus the clearest expression of natural law is the decalogue, which directs all human actions towards the attainment of beatitude. The commands of natural law are not good merely because they are commanded, but are commanded because they are good. (Ordinatio IV, 17) Scotus considers the first two commandments, that God must be worshipped and revered, to be absolutely unalterable. God himself could never negate such moral principles and human beings are morally bound to their adherence.

Despite his unrelenting criticisms of many of Scotus' positions, William of Ockham's moral theology develops, rather than dismisses, the main lines of Scotus' ethical deliberations. Ockham too is concerned primarily with the preservation of volitional freedom, both divine and human. Ockham specifically rejects the theory of natural law for determining human acts invariably toward an intellectually determined end. His insistence upon the dignity of human nature and the absolute power of the will for self-determination leads Ockham to reject his predecessors' morality of natural law. Ockham's critique of the metaphysics of common nature was not limited to logical and metaphysical speculation; it pertains also to his moral doctrine, wherein the will must be free even to choose ‘evil which is neither really or apparently good.’

The autonomy of the will is so great that it can absolutely refuse to pursue beatitude even when it is presented either as a general or particular idea. Even after death (*in patria*) the will can refuse to desire its own perfection. Volitional freedom is absolute in Ockham's moral theology; it can be defined as a natural inclination to an end, only insofar as it is an observed general human tendency. The human will

can just as easily reject its end, as it can pursue it. (Ordinatio I, d. 1, q. 6).

The natural foundation of morality, so essential to Albert and Thomas and still an important element of Scotus' thought, is rejected by Ockham in favor of a more complete notion of volitional freedom. Still Ockham does not advocate an ethics of relativism. The basis for human moral judgments lies in the will's conformity to divine commands. Impressed by Scotus' dictum, ‘Deus nullius est debitor’ (God is indebted to no one), Ockham extends the power of God to reformulate all moral laws. Not only can the commandments that regulate human interactions be altered, but also those that determine the relationship between God and man. God could command human beings to hate him and such a precept must be considered as morally binding. Ockham's use of the more common language of medieval moral theory does not prevent him from emphasizing the contingency of human morality. His belief in the power of human reason to discern the rational principles of an ordered life cannot overcome his desire to preserve the unlimited power of God, on whose will all moral principles depends and are subject to change. Both Scotus and Ockham construct a moral theory of volitional freedom, rather than one that they believe to be the moral determinism of Aristotle and his followers.

The final question remains: can the medieval explanations of practical reason be aligned with Kant's description of practical reason? Kant's well known definition of practical reason is:

> Everything in nature works according to laws. Only a rational being has the ability to act in accordance with the concept of laws, that is according to principles; in other words, only a rational being has a will. Since reason demands the derivation of actions from laws, the will is nothing other than practical reason. (Grundlegung zur Metaphysik der Sitten, II, 37)

For Kant the aim of moral philosophy is to discover ideas and principles that would constitute as pure a concept of will as possible and not to determine the actions and conditions of willing, which are mainly the concerns of psychology (Grundlegung, Intro., XII). The good will is good not through its result or its capacity to attain a predetermined end, but only by means of the willing itself. (Grundlegung, I, 3). The will's ability for self-determination according to universally binding laws conveys the objectivity demanded by the proper concept of reason. Since a rational nature declares itself by the self-imposition of the end, morality is the relationship of actions to the autonomy of the will; and this autonomy is the relation of the will to the most universal mandating (*Gesetzgebung*) possible. (Grundlegung, II, 83 & 85-86). The will purified of any inclination and desire for an ulterior end is the proper matter for practical reason.

The differences between Kantian and medieval theories of practical reason are obvious even from such a short description of their elements. The adherence of Albert and Thomas to the Greek view of human nature led them to their conviction of the will's natural inclination toward the good. Such a notion permits Thomas to advise us to seek naturally what is useful to us. (De veritate, q. 24, a. 8). What is truly advantageous to human beings must always be conducive to attaining the human end, beatitude; utility in this sense will always be a reliable guide to correct moral choice. Thomas claims that every thing to which man has a natural inclination human reason apprehends as good. For Kant human beings cannot be

permitted to pursue their inclinations which are to be viewed as subjective desires. Such desires lead to moral error at least as often as they produce right action. The objectivity of the pure concept of the human will precludes the recognition of human inclination as a measure of human goodness.

Despite their insistence upon volitional freedom, which in some aspects anticipates Kant's thought, Scotus and Ockham argue that the true measure of morality lies in the conformity of the human will with the commands of God. Such precepts may change if God so desires and so Scotus and Ockham produce an ethics quite different in spirit from that of Kant's theory of universal imperatives. Ockham in his rejection of the intellectual basis for morality argues that an act in total conformity with right reason may not be virtuous, since God could possibly create such an act without human volitional consent. The act would be completely rational and in conformity with divine commands, but would lack any merit or virtue. For Ockham the goodness of any action lies completely in the will's desire to obey divine commands (De connexione virtutum, III, 11). When Kant describes practical reason as a necessary law for all rational beings, whose actions are always to be judged according to maxims completely bound to the concept of the will of a rational being, one may see parallels to the Franciscan doctrine of the autonomy of the will. As Thomas and Albert describe the genesis of moral action as a process in which practical reason has the ability to accept, understand and obey the principles of a universal law consonant with human nature, they approach Kant's description of the dignity of the will more closely than those who insist upon the necessity for adherence to the possibly arbitrary precepts of God. Practical reason as the desire for the good as good, as expressed in the first principle of natural law or obedience to divine commands, and the conformity of the will through free choice to the universally binding principles of practical reason indicate that in some important elements of their moral theories medieval philosophers approach the spirit of Kant as they move away from the legacy of Aristotle.

Bibliography

- Albertus Magnus, De bono, ed. H. Kühle et al. in Opera omnia 28 (Münster in W., 1951).
- Albertus Magnus, Quaestiones, edd. A. Fries et al. in Opera omnia 25, 2 (Münster in Westphalia, 1993).
- Albertus Magnus, Super ethica commentum et quaestiones, ed. W. Kubel in Opera omnia 14, 1&2, (Münster in W., 1968 & 1987).
- I. Kant, Grundlegung zur Metaphysik der Sitten, ed. W. Weischedel (Frankfurt a. M., 1994).
- Thomas Aquinas, Sententia libri ethicorum, ed. R.-A. Gauthier, in Opera omnia, 47, 1 (Rome, 1969).
- Thomas Aquinas, Summa theologiae, ed. Piana (Ottawa, 1914).
- D. Westberg, Right Practical Reason: Aristotle, Action, Prudence in Aquinas (Oxford, 1994).
- A Wolter, Duns Scotus on the Will and Morality (Catholic Univ. Press, Washington, 1986), contains Latin texts and English translations.
- William of Ockham, Scriptum in librum sententiarum ordinatio, edd. G. Gal and S. Brown in Opera theologica I and II (St. Bonaventure, NY, 1967 & 1970).
- William of Ockham Circa virtutes et vitia, in Quaestiones variae, edd G. Etzkorn et al.in Opera theologica VII (St. Bonaventure, NY, 1984)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aquinas, Saint Thomas | Augustine, Saint | Ockham [Occam], William

Anthony Celano
Stonehill College
acelano@stonehill.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 8, 1999
Content last modified: October 11, 1999

Stanford Encyclopedia of Philosophy

Free Will

"Free Will" is largely a philosophical term of art for a particular sort of capacity of rational agents to choose a course of action from among various alternatives. Which sort is the free will sort is what all the fuss is about. (And what a fuss it has been: philosophers have debated this question for over two millenia, and just about every major philosopher has had something to say about it.) Most philosophers suppose that the concept of free will is very closely connected to the concept of moral responsibility. Acting with free will, on such views, is just to satisfy the metaphysical requirement on being responsible for one's action. (Clearly, there will also be epistemic conditions on responsibility as well, such as being aware -- or failing that, being culpably unaware -- of relevant alternatives to one's action and of the alternatives' moral significance.) But the significance of free will is not exhausted by its connection to moral responsibility. Free will also appears to be a condition on desert for one's accomplishments (why sustained effort and creative work are praiseworthy); on the autonomy and dignity of persons; and on the value we accord to love and friendship. (See Kane, 1996, 81ff.)

Philosophers who distinguish freedom of action and freedom of will do so because our success in carrying out our ends depends in part on factors wholly beyond our control. Furthermore, there are always external constraints on the range of options we can meaningfully try to undertake. As the presence or absence of these conditions and constraints are not (usually) our responsibility, it is plausible that the central loci of our responsibility are our choices, or "willings."

I have implied that free willings are but a subset of willings, at least as a conceptual matter. But not every philosopher accepts this. Rene Descartes, for example, identifies the faculty of will with freedom of choice, "the ability to do or not do something" (Meditation IV), and even goes so far as to declare that "the will is by its nature so free that it can never be constrained" (Passions of the Soul, I, art.41). In taking this strong polar position on the nature of will, Descartes is reflecting a tradition running through certain late Scholastics (most prominently, Suarez) back to John Duns Scotus.

The majority view, however, is that we can readily conceive willings that are not free. Indeed, much of the debate about free will centers around whether we human beings *have* it, yet virtually no one doubts that we will to do this and that. The main perceived threats to our freedom of will are various alleged determinisms: physical/causal; psychological; biological; theological. For each such variety of determinism, there are philosophers who (i) deny its reality, either because of the existence of free will or on independent grounds; (ii) accept its reality but argue for its compatibility with free will; (iii) accept its reality and deny its compatibility with free will. (See the entries on compatibilism; causal determinism; fatalism; and arguments for incompatibilism.) There are also a few who say the truth of any variety of

determinism is irrelevant because free will is simply impossible.

If there is such a thing as free will, it has many dimensions. In what follows, I will sketch the freedom-conferring characteristics that have attracted most of the attention. The reader is warned, however, that while many philosophers emphasize a single such characteristic, perhaps in response to the views of their immediate audience, it is probable that most would recognize the significance of many of the other features discussed here.

---

1. Rational Deliberation

1.1 Free Will as choosing on the basis of one’s desires

The minimalist account of free will is as the ability to select a course of action as a means of fulfilling some desire. David Hume, for example, defines liberty as "a power of acting or of not acting, according to the determination of the will." (1748, sect.viii, part 1). And we find in Edwards a similar account of free willings as those which proceed from one’s own desires.

One reason to deem this insufficient is that it is consistent with the goal-directed behavior of some animals whom we do not suppose to be morally responsible agents. Such animals lack not only an awareness of the moral implications of their actions but also any capacity to reflect on their alternatives and their long-term consequences. Indeed, it is plausible that they have little by way of a self-conception as an agent with a past and with projects and purposes for the future. (See Baker 2000 on the ‘first-person perspective.’)

1.2 Free Will as deliberative choosing on the basis of desires and values

A natural suggestion, then, is to modify the minimalist thesis by taking account of (what may be) distinctively human capacities and self-conception. And indeed, philosophers since Plato have commonly distinguished the ‘animal’ and ‘rational’ parts of our nature, with the latter implying a great deal more

psychological complexity. Our rational nature includes our ability to judge some ends as ‘good’ or worth pursuing and value them even though satisfying them may result in considerable unpleasantness for ourselves. (Note that such judgments need not be based in moral value.) We might say that we act with free will when we act upon our considered judgments/valuings about what is good for us, whether or not our doing so conflicts with an ‘animal’ desire. (See Watson 1982 for a subtle development of this sort of view.) But this would seem unduly restrictive, since we clearly hold many people responsible for actions proceeding from ‘animal’ desires that conflict with their own assessment of what would be best in the circumstances. More plausible is the suggestion that one acts with free will when one’s deliberation is *sensitive to* one’s own judgments concerning what is best in the circumstances, whether or not one acts upon such a judgment.

Here we are clearly in the neighborhood of the ‘rational appetite’ accounts of will one finds in the medieval Aristotelians. The most elaborate medieval treatment is Thomas Aquinas’s.[1] His account involves identifying several distinct varieties of willings. Here I note only a few of his basic claims. Aquinas thinks our nature determines us to will certain general ends ordered to the most general goal of goodness. These we will of necessity, not freely. Freedom enters the picture when we consider various means to these ends, none of which appear to us either as unqualifiedly good or as uniquely satisfying the end we wish to fulfill. There is, then, free choice of means to our ends, along with a more basic freedom not to consider something, thereby perhaps avoiding willing it unavoidably once we recognized its value. Free choice is an activity that involves both our intellectual and volitional capacities, as it consists in both judgment and active commitment. A thorny question for this view is whether will or intellect is the ultimate determinant of free choices. How we understand Aquinas on this point will go a long ways towards determining whether or not he is a sort of compatibilist about freedom and determinism. (See below. Good expositions of Aquinas’ account are Donagan, 1985, and Stump, 1997.)

There are two general worries about theories of free will that principally rely on the capacity to deliberate about possible actions in the light of one’s conception of the good. First, there are agents who deliberately choose to act as they do but who are motivated to do so by a compulsive, controlling sort of desire. (And there seems to be no principled bar to a compulsive desire’s being informing a considered judgment of the agent about what the good is for him.) Such agents are not willing freely. Secondly, we can imagine a person’s psychology being externally manipulated by another agent (via neurophysiological implant, say), such that the agent is caused to deliberate and come to desire strongly a particular action which he previously was not disposed to choose. The deliberative process could be perfectly normal, reflective, and rational, but seemingly not freely made. The agent’s freedom seems undermined or at least greatly diminished by such psychological tampering.

1.3 Self-mastery, rightly-ordered appetite

Some theorists are much impressed by cases of inner, psychological compulsion and define freedom of will in contrast to this phenomenon. For such thinkers, true freedom of the will involves liberation from the tyranny of base desires and acquisition of desires for the Good. Plato, for example, posits rational, spirited, and appetitive aspects to the soul and holds that willings issue from the higher, rational part

alone. In other cases, one is dominated by the irrational desires of the two lower parts.[2] This is particularly characteristic of those working in a theological context -- for example, the New Testament writer St. Paul, speaking of Christian freedom (Romans vi-viii; Galatians v), and those influenced by him on this point, such as Augustine. (The latter, in both early and later writings, allows for a freedom of will that is not ordered to the good, but maintains that it is of less value than the rightly-ordered freedom. See, for example, the discussion in Books II-III of *On Free Choice*.) More recently, Susan Wolf (1990) defends an asymmetry thesis concerning freedom and responsibility. On her view, an agent acts freely only if he had the ability to choose the True and the Good. For an agent who does so choose, the requisite ability is automatically implied. But those who reject the Good choose freely only if they could have acted differently. This is a further substantive condition on freedom, making freedom of will a more demanding condition in cases of bad choices.

In considering such rightly-ordered-appetites views of freedom, I again focus on abstract features common to all. It explicitly handles the inner-compulsion worry facing the simple deliberation-based accounts. The other, external manipulation problem could perhaps be handled through the addition of an historical requirement: agents will freely only if their willings are not in part explicable by episodes of external manipulation which bypass their critical and deliberative faculties. But another problem suggests itself: an agent who was a ‘natural saint’, always and effortlessly choosing the good with no contrary inclination, would not have freedom of will among his virtues. Doubtless we would greatly admire such a person, but would it be an admiration suffused with moral praise of the person or would it, rather, be restricted to the goodness of the person’s qualities? (Cf. Kant, 1788.) The appropriate response to such a person, it seems, is on an analogy with aesthetic appreciation of natural beauty, in contrast to the admiration of the person who chooses the good in the face of real temptation to act selfishly. Since this view of freedom of will as orientation to the good sometimes results from theological reflections, it is worth noting that other theologian-philosophers emphasize the importance for human beings of being *able* to reject divine love: love of God that is not freely given -- given in the face of a significant possibility of one’s having not done so -- would be a sham, all the more so since, were it inevitable, it would find its ultimate and complete explanation in God Himself.

2. Ownership

Harry Frankfurt (1982) presents an insightful and original way of thinking about free will. He suggests that a central difference between human and merely animal activity is our capacity to reflect on our desires and beliefs and form desires and judgments concerning them. I may want to eat a candy bar (first-order desire), but I also may want not to want this (second-order desire) because of the connection between habitual candy eating and poor health. This difference, he argued, provides the key to understanding both free action and free will. (These are quite different, in Frankfurt’s view, with free will being the more demanding notion. Moreover, moral responsibility for an action requires only that the agent acted freely, not that the action proceeded from a free will.)

On Frankfurt’s analysis, I *act* freely when the desire on which I act is one that I desire to be effective. This second-order desire is one with which I *identify*: it reflects my true self. (Compare the addict:

typically, the addict acts out of a desire which he does not want to act upon. His will is divided, and his actions proceed from desires with which he does not reflectively identify. Hence, he is not acting freely.) My *will* is free when I am *able* to make any of my first-order desires the one upon which I act. As it happens, I will to eat the candy bar, but I could have willed to refrain from doing so.

With Frankfurt's account of free will, much hangs on what being able to will otherwise comes to, and on this Frankfurt is officially neutral. (See the related discussion below on ability to do otherwise.) But as he connects moral responsibility only to his weaker notion of free action, it is fitting to consider its adequacy here. The central objection that commentators have raised is this: what's so special about higher-order willings or desires? (See in particular Watson 1982a.) Why suppose that they inevitably reflect my true self, as against first-order desires? Frankfurt is explicit that higher-order desires need not be rooted in a person's moral or even settled outlook (89, n.6). So it seems that, in some cases, a first-order desire may be much more reflective of my true self (more "internal to me," in Frankfurt's terminology) than a weak, faint desire to be the sort of person who wills differently.

In later writings, Frankfurt responds to this worry first by appealing to "decisions made without reservations" ("Identification and Externality" and "Identification and Wholeheartedness" in Frankfurt, 1988) and then by appealing to higher-order desires with which one is "satisfied," such that one has no inclination to make changes to them (1992). But the absence of an inclination to change the desire does not obviously amount to the condition of freedom-conferring identification. It seems that such a negative state of satisfaction can be one that I just find myself with, one that I *neither* approve nor disapprove (Pettit, 2001, 56).

Furthermore, we can again imagine external manipulation consistent with Frankfurt's account of freedom but inconsistent with freedom itself. Armed with the appropriate neurophysiology-tampering technology of the late 21st century, one might discreetly induce a second-order desire in me to be moved by a first-order desire -- a higher-order desire with which I am satisfied -- and then let me deliberate as normal. Clearly, this desire should be deemed "external" to me, and the action that flows from it unfree.

3. Causation and Control

Our survey of several themes in philosophical accounts of free will suggests that a -- perhaps *the* -- root issue is that of *control*. Clearly, our capacity for deliberation and the potential sophistication of some of our our practical reflections are important conditions on freedom of will. But any proposed analysis of free will must also ensure that the process it describes is one that was up to, or controlled by, the agent.

Fantastic scenarios of external manipulation and less fantastic cases of hypnosis are not the only, or even primary, ones to give philosophers pause. It is consistent with my deliberating and choosing 'in the normal way' that my developing psychology and choices over time are part of an ineluctable system of causes necessitating effects. It might be, that is, that underlying the phenomena of purpose and will in human persons is an all-encompassing, mechanistic world-system of 'blind' cause and effect. Many accounts of free will are constructed against the backdrop possibility (whether accepted as actual or not)

that each stage of the world is determined by what preceded it by impersonal natural law. As always, there are optimists and pessimists.

3.1 Free Will as guidance control

John Martin Fischer (1994) distinguishes two sorts of control over one's actions: guidance and regulative. A person exerts guidance control over his own actions insofar as they proceed from a 'weakly' reasons-responsive (deliberative) mechanism. This obtains just in case there is some *possible* scenario where the agent is presented with a sufficient reason to do otherwise and the mechanism that led to the actual choice is operative and it issues in a different choice, one appropriate to the imagined reason. In Fischer and Ravizza (1998), the account is elaborated and refined. They require, more strongly, that the mechanism be the person's own mechanism (ruling out external manipulation) and that it be 'moderately' responsive to reasons: one that is "regularly *receptive* to reasons, some of which are moral reasons, and at least weakly *reactive* to reason" (82, emphasis added). Receptivity is evinced through an understandable pattern of reasons recognition -- beliefs of the agent about what would constitute a sufficient reason for undertaking various actions. (See 69-73 for details.)

None of this, importantly, requires 'regulative' control: a control involving the ability of the agent to choose and act differently in the actual circumstances. Regulative control requires *alternative* possibilities open to the agent, whereas guidance control is determined by characteristics of the *actual* sequence issuing in one's choice. Fischer allows that there is a notion of freedom that requires regulative control but does not believe that this kind of freedom is required for moral responsibility. (In this, he is persuaded by a form of argument originated by Harry Frankfurt. See Frankfurt 1969 and Fischer 1994, Ch.7.)

3.2 Free Will as ultimate origination (ability to do otherwise)

Many do not follow Fischer here, however, and maintain the traditional view that the sort of freedom required for moral responsibility does indeed require that the agent could have acted differently. As Aristotle put it, "...when the origin of the actions is in him, it is also up to him to do them or not to do them" (1985, Book III).[3]

A flood of ink has been spilled, especially in the modern era, on how to understand the concept of being able to do otherwise. On one side are those who give it a deflationary reading, on which it is consistent with my being able to do otherwise that the past (including my character and present beliefs and desires) and the basic laws of nature logically entail that I do what I actually do. These are the 'compatibilists,' holding that freedom and causal determinism are compatible. (For discussion, see O'Connor, 2000, Ch.1; *compatibilism*; and *incompatibilism: arguments for*.) Conditional analyses of ability to do otherwise have been popular among compatibilists. The general idea here is that to say that I am able to do otherwise is to say that I *would* do otherwise if it were the case that ... , where the ellipsis is filled by some elaboration of "I had an appropriately strong desire to do so, or I had different beliefs about the best available means to satisfy my goal, or ... ." In short: something about my prevailing character or present psychological

states would have differed, and so would have brought about a different outcome in my deliberation.

Incompatibilists think that something stronger is required: for me to act with free will requires that there are a plurality of futures open to me consistent with the past (and laws of nature) *being just as they were*. I could have chosen differently even without some further, non-actual consideration's occurring to me and 'tipping the scales of the balance' in another direction. Indeed, from their point of view, the whole scale-of-weights analogy is wrongheaded: free agents are not mechanisms that respond invariably to specified 'motive forces.' They are capable of acting upon any of a plurality of motives making attractive more than one course of action. Ultimately, the agent must determine *himself* this way or that.

We may distinguish two broad families of 'incompatibilist' or 'indeterminist' self-determination accounts. The more radical group holds that the agent who determines his own will is not causally influenced by external causal factors, including his own character. Descartes, in the midst of exploring the scope and influence of 'the passions,' declares that "the will is by its nature so free that it can never be constrained" (1984, v.I, 343). And as we've seen, he believed that such freedom is present on every occasion when we make a conscious choice -- even, he writes, "when a very evident reason moves us in one direction...." (1984, v.III, 245). More recently, John Paul Sartre notoriously held that human beings have 'absolute freedom': "No limits to my freedom can be found except freedom itself, or, if you prefer, we are not free to cease being free." (567) His views on freedom flowed from his radical conception of human beings as lacking any kind of positive nature. Instead, we are 'non-beings' whose being, moment to moment, is simply to choose:

> For human reality, to be is to choose oneself; nothing comes to it either from the outside or from within which it can receive or accept....it is entirely abandoned to the intolerable necessity of making itself be, down to the slightest details. Thus freedom...is the being of man, i.e., his nothingness of being. (568-9)

Scotus and, more recently, C.A. Campbell, appear to agree with Descartes and Sartre on the lack of direct causal influence on the activity of free choice while allowing that the scope of possibilities for what I might thus will may be more or less constricted. So while Scotus holds that "nothing other than the will is the total cause" of its activity, he grants (with Aquinas and other medieval Aristotelians) that we are not capable of willing something in which we see no good, nor of positively repudiating something which appears to us as unqualifiedly good. Contrary to Sartre, we come with a 'nature' that circumscribes what we might conceivably choose, and our past choices and environmental influences also shape the possibilities for us at any particular time. But if we are presented with what we recognize as an unqualified good, we still can choose to *refrain from willing* it. And while Campbell holds that character cannot explain a free choice, he supposes that "[t]here is one experiential situation, and *one only*, ... in which there is any possibility of the act of will not being in accordance with character; viz. the situation in which the course which formed character prescribes is a course in conflict with the agent's moral ideal: in other words, the situation of moral temptation" (1967, 46). (Van Inwagen 1994 and 1995 is another proponent of the idea that free will is exercised in but a small subset of our choices, although his position is less extreme on this point than Campbell's. Fischer and Ravizza 1992 criticize van Inwagen's argument for this position, as does O'Connor 2000, Ch.5.)

A more moderate grouping within the self-determination approach to free will allows that beliefs, desires, and external factors all can causally influence the act of free choice itself. But theorists within this camp differ sharply on the metaphysical nature of those choices and of the causal role of reasons. We may distinguish three varieties. I will discuss them only briefly, as they are explored at length in *incompatibilist (nondeterministic) theories of free will*.

First is a noncausal (or ownership) account (Ginet 1990 and McCann 1998). According to this view, I control my choice simply in virtue of its being mine -- its occurring on[in] me. I do not exert any special kind of causality in bringing it about. While there may be causal influences upon my choice, there need not be, and any such causal influences wholly irrelevant to understanding why it occurs. Reasons provide an autonomous, non-causal form of explanation. Provided my choice is not wholly determined by prior factors, it is free and under my control simply in virtue of being mine.

Proponents of the event-causal account (e.g. Nozick 1995 and Ekstrom 2001) would say that uncaused events of any kind would be random and uncontrolled by anyone, and so could hardly count as choices that an agent *made*. They hold that reasons influence choices precisely by causing them. Choices are *free* insofar as they are not deterministically caused, and so might not have occurred in just the circumstances in which they did occur. A special case of the event-causal account of self-determination is Kane (1996). Kane believes that the free choices of greatest significance to an agent’s autonomy are ones that are preceded by efforts of will within the process of deliberation. These are cases where one’s will is conflicted, as when one’s duty or long-term self-interest compete with a strong desire for a short-term good. As one struggles to sort out and prioritize one’s own values, the possible outcomes are not merely undetermined, but also *indeterminate*: at each stage of the struggle, the possible outcomes have no specific objective probability of occurring. This indeterminacy, Kane believes, is essential to freedom of will.

Finally, there are those who believe freedom of will consists in a distinctively personal form of causality, commonly referred to as "agent causation." The agent himself causes his choice or action, and this is not to be reductively analyzed as an event within the agent causing the choice. (Compare our ready restatement of "the rock broke the window" into the more precise "the rock’s being in momentum M at the point of contact with the window caused the window’s subsequent shattering.") This view is given clear articulation by Thomas Reid:

> I grant, then, that an effect uncaused is a contradiction, and that an event uncaused is an absurdity. The question that remains is whether a volition, undetermined by motives, is an event uncaused. This I deny. The cause of the volition is the man that willed it. (Letter to James Gregory, in 1967, 88)

Roderick Chisholm advocated this view of free will in numerous writings (e.g., 1982 and 1976). And recently it has been developed in different forms by Randolph Clarke (1993, 1996) and O’Connor (2000). Nowadays, many philosophers view this account as of doubtful coherence (e.g., Dennett 1984).

For some, this very idea of causation by a substance just as such is perplexing (Ginet 1997). Others see it as difficult to reconcile with the causal role of reasons in explaining choices. (Clarke and O'Connor devote considerable effort to addressing this concern.) And yet others hold that, coherent or not, it is inconsistent with seeing human beings as part of the natural world of cause and effect (Pereboom 2001).

A recent trend is to suppose that agent causation accounts capture, as well as possible, our prereflective idea of responsible, free action. But the failure of philosophers to work the account out in a fully satisfactory and intelligible form reveals that the very idea of free will (and so of responsibility) is incoherent (Strawson 1986) or at least inconsistent with a world very much like our own (Pereboom 2001). Smilansky (2000) takes a more complicated position, on which there are two 'levels' on which we may assess freedom, 'compatibilist' and 'ultimate'. On the ultimate level of evaluation, free will is indeed incoherent.

4. Theological Wrinkles

A large portion of Western philosophical writing on free will was and is written within an overarching theological framework, according to which God is the ultimate source *and sustainer* of all else. Some of these thinkers draw the conclusion that God must be a sufficient, wholly determining cause for everything that happens; all suppose that every creaturely act necessarily depends on the explanatorily prior, cooperative activity of God. It is also presumed that human beings are free and responsible (on pain of attributing evil in the world to God alone, and so impugning His perfect goodness). Hence, those who believe that God is omni-determining typically are compatibilists with respect to freedom and (in this case) theological determinism. Edwards (1957) is a good example. But those who suppose that God's sustaining activity (and special activity of conferring grace) is only a necessary condition on the outcome of human free choices need to tell a more subtle story, on which omnipotent God's cooperative activity can be (explanatorily) prior to a human choice and yet the outcome of that choice be settled only by the choice itself. (For important medieval discussions -- the period of the apex of treatments of philosophical/theological matters -- see the relevant portions of Aquinas 1945 and Scotus 1994.)

Another issue concerns the impact on human freedom of knowledge of God, the ultimate Good. Many philosophers, especially the medieval Aristotelians, were drawn to the idea that human beings cannot but will that which they take to be an unqualified good. (Duns Scotus appears to be an important exception to this consensus.) Hence, in the afterlife, when humans 'see God face to face,' they will inevitably be drawn to Him. Murray (1993) is a contemporary development of the argument that a good God would choose to make His existence and character less than certain for human beings, for the sake of their freedom. (He will do so, the argument goes, at least for a period of time in which human beings participate in their own character formation.) If it is a good for human beings that they freely choose to respond in love to God and to act in obedience to His will, then God must maintain an 'epistemic distance' from them lest they be overwhelmed by His goodness and respond out of necessity, rather than freedom.

Finally, there is the question of the freedom of God himself. Perfect goodness is an essential, not

acquired, attribute of God. God cannot lie or be in any way immoral in His dealings with His creatures. Unless we take the minority position on which this is a trivial claim, since whatever God does *definitionally* counts as good, this appears to be a significant, inner constraint on God's freedom. Did we not contemplate immediately above that human freedom would be curtailed by our having an unmistakable awareness of what is in fact the Good? And yet is it not passing strange to suppose that God should be less than perfectly free?

One suggested solution to this puzzle begins by reconsidering the relationship of two strands in (much) thinking about freedom of will: being able to do otherwise and being the ultimate source of one's will. Contemporary discussions of free will often emphasize the importance of being able to do otherwise. Yet it's plausible (Kane 1996) that the core metaphysical feature of freedom is being the ultimate source, or originator, of one's choices, and that being able to do otherwise is closely connected to this feature. For human beings or any created persons who owe their existence to factors outside themselves, the only way their acts of will could find their ultimate origin in themselves is for such acts not to be determined by their character and circumstances. For if all my willings were wholly determined, then if we were to trace my causal history back far enough, we would ultimately arrive at external factors that gave rise to me, with my particular genetic dispositions. My motives at the time would not be the ultimate source of my willings, only the most proximate ones. Only by there being less than deterministic connections between external influences and choices, then, is it be possible for me to be an ultimate source of my activity, concerning which I may truly say, "the buck stops here."

As is generally the case, things are different on this point in the case of God. Even if God's character absolutely precludes His performing certain actions in certain contexts, this will not imply that some external factor is in any way a partial origin of His willings and refrainings from willing. Indeed, this would not be so even if he were determined by character to will everything which He wills. For God's nature owes its existence to nothing. So God would be the sole and ultimate source of His will even if He couldn't will otherwise.

Well, then, might God have willed otherwise in any respect? The majority view in the history of philosophical theology is that He indeed could have. He might have chosen not to create anything at all. And given that He did create, He might have created any number of alternatives to what we observe. But there have been noteworthy thinkers who argued the contrary position, along with others who clearly felt the pull of the contrary position even while resisting it. The most famous such thinker is Leibniz (1985), who argued that God, being both perfectly good and perfectly powerful, cannot fail to will the best possible world. Leibniz insisted that this is consistent with saying that God is able to will otherwise, although his defense of this last claim is notoriously difficult to make out satisfactorily. Many read Leibniz, *malgre lui*, as one whose basic commitments imply that God could not have willed other than He does in any respect.

On might challenge Leibniz's reasoning on this point by questioning the assumption that there is a uniquely best possible Creation (Adams 1987). One way this could be is if there is no well-ordering of worlds: some worlds are sufficiently different in kind that they are incommensurate with each other (neither is better than the other, nor are they equal). Another way this could be is if there is no upper limit

on goodness of worlds: for every possible world God might have created, there are others (infinitely many, in fact) which are better. If such is the case, one might argue, it is reasonable for God to arbitrarily choose which world to create from among those worlds exceeding some threshhold value of overall goodness.

However, William Rowe (1993) has countered that the thesis that there is no upper limit on goodness of worlds has a very different consequence: it shows that there could not be a morally perfect Creator! For suppose our world has an on-balance moral value of n and that God chose to create it despite being aware of possibilities having values higher than n that He was able to create. It seems we can now imagine a morally better Creator: one having the same options who chooses to create a better world. For a critical reply to Rowe, see the Howard-Snyders (1994) and Wainwright (1996).

Finally, Norman Kretzmann (1997, 220-25) has argued in the context of Aquinas's theological system that there is strong pressure to say that God must have created something or other, though it may well have been open to Him to create any of a number of contingent orders. The reason is that there is no plausible account of how an absolutely perfect God might have a *resistible* motivation -- one consideration among other, competing considerations -- for creating something rather than nothing. (It obviously cannot have to do with any sort of utility, for example.) The best general understanding of God's being motivated to create at all -- one which in places Aquinas himself comes very close to endorsing -- is to see it as reflecting the fact that God's very being, which is goodness, necessarily diffuses itself. Perfect goodness will naturally communicate itself outwardly; God who is perfect goodness will naturally create, generating a dependent reality that imperfectly reflects that goodness. (Wainwright (1996) is a careful discussion of a somewhat similar line of thought in Jonathan Edwards.)

Further Reading. Pereboom (1997) provides nice selections from a number of important historical writers on free will. Bourke (1964) and Dilman (1999) provide critical overviews of many such writers. For thematic treatments, see Kane (1996), esp. Ch.1-2; 5-6; Ekstrom (2001); and the lengthy survey articles in Kane (forthcoming).

Bibliography

- Adams, Robert (1987). "Must God Create the Best?," in *The Virtue of Faith and Other Essays in Philosophical Theology*. New York: Oxford University Press, 51-64.
- Aquinas, Thomas (1945). *Basic Writings of Saint Thomas Aquinas* (2 vol.). New York: Random House.
- Aquinas, Thomas (1993). *Selected Philosophical Writings*, ed. T. McDermott. Oxford: Oxford University Press.
- Aristotle (1985). *Nichomachean Ethics*, translated by Terence Irwin. Indianapolis: Hackett Publishing.
- Augustine (1993). *On the Free Choice of the Will*, tr. Thomas Williams. Indianapolis: Hackett Publishing.
- Ayer, A.J. (1982). "Freedom and Necessity," in Watson (1982b), ed., 15-23.

- Baker, Lynne (2000). *Persons and Bodies: A Constitution View*. Cambridge: Cambridge University Press.
- Bourke, Vernon (1964). *Will in Western Thought*. New York: Sheed and Ward.
- Campbell, C.A. (1967). *In Defence of Free Will & other essays*. London: Allen & Unwin Ltd.
- Chisholm, Roderick (1982). "Human Freedom and the Self," in Watson (1982b), 24-35.
- Chisholm, Roderick (1976). *Person and Object*. LaSalle: Open Court.
- Clarke, Randolph (1993). "Toward a Credible Agent-Causal Account of Free Will," in O'Connor (1995), ed., 201-15.
- Clarke, Randolph (1995). "Indeterminism and Control," *American Philosophical Quarterly* 32, 125-138.
- Clarke, Randolph (1996). "Agent Causation and Event Causation in the Production of Free Action," *Philosophical Topics* 24 (Fall), 19-48.
- Dennett, Daniel (1984). *Elbow Room: The Varieties of Free Will Worth Having*. Cambridge. MA: MIT Press.
- Descartes, Rene (1984). *Meditations on First Philosophy* [1641] and *Passions of the Soul* [1649], in *The Philosophical Writings of Descartes*, vol. I-III, translated by Cottingham, J., Stoothoff, R., & Murdoch, D.. Cambridge: Cambridge University Press.
- Donagan, Alan (1985). *Human Ends and Human Actions: An Exploration in St. Thomas's Treatment*. Milwaukee: Marquette University Press.
- Dilman, Ilham (1999). *Free Will: An Historical and Philosophical Introduction*. London: Routledge.
- Double, Richard (1991). *The Non-Reality of Free Will*. New York: Oxford University Press.
- Edwards, Jonathan (1957) [1754]. *Freedom of Will*, ed. P. Ramsey. New Haven: Yale University Press.
- Ekstrom, Laura (2000). *Free Will: A Philosophical Study*. Boulder, CO: Westview Press.
- Farrer, Austin (1958). *The Freedom of the Will*. London: Adam & Charles Black.
- Fischer, John Martin (1994). The Metaphysics of Free Will. Oxford: Blackwell.
- Fischer, John Martin (1999). "Recent Work on Moral Responsibility," *Ethics* 110, 93-139.
- Fischer, John Martin and Ravizza, Mark. (1992). "When the Will is Free," in O'Connor (1995), ed., 239-269.
- Fischer, John Martin (1998) *Responsibility and Control*. Cambridge: Cambridge University Press.
- Frankfurt, Harry (1969). "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66, 829-39.
- Frankfurt, Harry (1982). "Freedom of the Will and the Concept of a Person," in Watson (1982), ed., 81-95.
- Frankfurt, Harry (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.
- Frankfurt, Harry (1992). "The Faintest Passion," *Proceedings and Addresses of the American Philosophical Association* 66, 5-16.
- Ginet, Carl (1990). *On Action*. Cambridge: Cambridge University Press.
- Ginet, Carl (1997) "Freedom, Responsibility, and Agency," *The Journal of Ethics* 1, 85-98.
- Hobbes, Thomas and Bramhall, John (1999) [1655-1658]. *Hobbes and Bramhall on Liberty and Necessity*, ed. V. Chappell. Cambridge: Cambridge University Press.

- Honderich, Ted (1988). *A Theory of Determinism*. Oxford: Oxford University Press.
- Howard-Snyder, Daniel and Frances (1994). "How an Unsurpassable Being Can Create a Surpassable World," *Faith and Philosophy* 11, 260-8.
- Hume, David (1977) [1748]. *An Enquiry Concerning Human Understanding*. Indianapolis: Hackett Publishing.
- Kane, Robert (1995). "Two Kinds of Incompatibilism," in O'Connor (1995), ed., 115-150.
- Kane, Robert (1996). *The Significance of Free Will*. New York: Oxford University Press.
- Kane, Robert, ed., (forthcoming). *Oxford Handbook on Free Will*. New York: Oxford University Press.
- Kant, Immanuel (1993) [1788]. *Critique of Practical Reason*, tr. by Lewis White Beck. Upper Saddle River, NJ: Prentice-Hall Inc.
- Kretzmann, Norman (1997). *The Metaphysics of Theism: Aquinas's Natural Theology in Summa Contra Gentiles I*. Oxford: Clarendon Press.
- Leibniz, Gottfried (1985) [1710]. *Theodicy*. LaSalle, IL: Open Court.
- Magill, Kevin (1997). *Freedom and Experience*. London: MacMillan.
- McCann, Hugh (1998). *The Works of Agency: On Human Action, Will, and Freedom*. Ithaca: Cornell University Press.
- Mele, Alfred (1995). Autonomous Agents (New York: Oxford University Press).
- Morris, Thomas (1993). "Perfection and Creation," in E. Stump. (1993), ed., 234-47
- Murray, Michael (1993). "Coercion and the Hiddenness of God," *American Philosophical Quarterly* 30, 27-38.
- Nozick, Robert (1995). "Choice and Indeterminism," in O'Connor (1995), ed., 101-14.
- O'Connor, Timothy (1993). "Indeterminism and Free Agency: Three Recent Views," *Philosophy and Phenomenological Research*, 53, 499-526.
- O'Connor, Timothy, ed., (1995). *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. New York: Oxford University Press.
- O'Connor, Timothy (2000). *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.
- Pettit, Philip (2001). *A Theory of Freedom*. Oxford: Oxford University Press.
- Pereboom, Derk (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, Derk, ed., (1997). *Free Will*. Indianapolis: Hackett Publishing.
- Plato (1997). *Complete Works*, ed. J. Cooper. Indianapolis: Hackett Publishing.
- Reid, Thomas (1969). *Essays on the Active Powers of the Human Mind*, ed. B. Brody. Cambridge: MIT Press.
- Rowe, William (1993). "The Problem of Divine Perfection and Freedom," in E. Stump (1993), ed., 223-33.
- Rowe, William (1995). "Two Concepts of Freedom," in O'Connor (1995), ed. 151-71.
- Sartre, John Paul (1956). *Being and Nothingness*. New York: Washington Square Press.
- Schopenhauer, Arthur (1999) [1839]. *Prize Essay on the Freedom of the Will*, ed. G. Zoller. Cambridge: Cambridge University Press.
- Scotus, John Duns (1986). "Questions on Aristotle's Metaphysics IX, Q.15" in *Duns Scotus on the Will and Morality* [selected and translated by Allan B. Wolter, O.F.M.]. Washington: Catholic University of America Press.

- Scotus, John Duns (1994) [1297-99]. *Contingency and Freedom: Lectura I 39*, tr. Vos Jaczn *et al*. Dordrecht: Kluwer Academic Publishers.
- Shatz, David (1986). "Free Will and the Structure of Motivation," *Midwest Studies in Philosophy* 10, 451-482.
- Smilansky, Saul (2000). *Free Will and Illusion*. Oxford: Oxford University Press.
- Strawson, Galen (1986). *Freedom and Belief*. Oxford: Clarendon Press.
- Strawson, Peter (1982). "Freedom and Resentment," in Watson (1982), ed., 59-80.
- Stump, Eleonore (1996). "Persons: Identification and Freedom," Philosophical Topics 24, 183-214.
- (1997). "Aquinas's Account of Freedom: Intellect and Will," *The Monist* 80, 576-597.
- Stump, Eleonore (1993). *Reasoned Faith*. Ithaca: Cornell University Press.
- van Inwagen, Peter (1983). *An Essay on Free Will*. Oxford: Oxford University Press.
- van Inwagen, Peter (1994). "When the Will is Not Free," *Philosophical Studies*, 75, 95-113.
- van Inwagen, Peter (1995). "When Is the Will Free?" in O'Connor (1995), ed., 219-238.
- Wainwright, William (1996). "Jonathan Edwards, William Rowe, and the Necessity of Creation," in J. Jordan and D. Howard-Snyder, eds., *Faith Freedom, and Rationality*. Lanham: Rowman and Littlefield, 119-133.
- Watson, Gary (1982a). "Free Agency," in Watson 1982b, 96-110.
- Watson, Gary (1982b) *Free Will*. Oxford: Oxford University Press.
- Watson, Gary (1987). "Free Action and Free Will," *Mind* 94, 145-72.
- Wolf, Susan (1990). *Freedom Within Reason*. Oxford: Oxford University Press.

Other Internet Resources

- The Determinism and Freedom Philosophy Website, edited by Ted Honderich (University College London)
- Bibliography on Ffree will, maintained by David Chalmers (U. Arizona)

[Please contact the author other further suggestions.]

Related Entries

action | compatibilism | determinism, causal | fatalism | incompatibilism: (nondeterministic) theories of free will | incompatibilism: arguments for | moral responsibility

Timothy O'Connor
toconnor@indiana.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 7, 2002
Content last modified: January 7, 2002

Under Construction

Under Construction

Stanford Encyclopedia of Philosophy
Notes to Free Will

Notes

1. His view is discussed in several places. See, for example, *Summa Theologiae* I, q.82 (1945, vol.I) and *Questions on Evil*, q.6 (1993).

2. Plato's views on the soul and its powers are set in numerous places. See, e.g., *The Republic,* Book IV; *Phaedrus,* 237e-238e and 246-248; *Gorgias,* 466. All are found in (1997).

3. Note that Aristotle here sees origination in the agent and ability to do otherwise as closely related. Robert Kane (1996) suggests that while some form of ability to do otherwise is indeed necessary for moral responsibility, this condition is but an indicator of something deeper to free will: the willing's finding its *ultimate origin* in the agent.

Timothy O'Connor
toconnor@indiana.edu

First published: January 7, 2002
Content last modified: January 7, 2002

Incompatibilist (Nondeterministic) Theories of Free Will

To act with free will is to exercise a certain type of control over one's behavior; what you do, when you act freely, is up to you. Incompatibilists hold that we exercise this control only if determinism is false. Many incompatibilists say little more about what, besides the falsehood of determinism, free will requires. And, indeed, the task of providing an incompatibilist account is not an easy one. For, if the truth of determinism would preclude free will, it is far from obvious how indeterminism would help.

Incompatibilist accounts that have been offered are of three main types, differing with respect to which form of indeterminism (uncaused events, nondeterministically caused events, agent- (or substance-) caused events) they require. Further variations among accounts concern where in the processes leading to actions they require indeterminism and what other conditions besides indeterminism they require. The first three sections below examine recent versions of each of the three main types of incompatibilist view. The fourth section considers the evidence regarding whether in fact there does exist what any of these accounts characterizes.

---

1. Noncausal Accounts

Some nondeterministic accounts require that a free decision or other free action have no cause at all; some require that it either have no cause or be only nondeterministically caused by prior events. Since both such views place no positive causal requirements on free action, we may call them "noncausal accounts."

Proponents of noncausal views generally hold that every action is or begins with a mental action. A decision or a choice is typically held to be an example of an action that is just such a mental action. An overt bodily action, such as raising one's arm, is held to be a complex action consisting of a mental action's bringing about a certain motion of one's body. The mental action here is often called a volition, which is said to be the agent's willing, trying, or endeavoring to move a certain part of her body in a certain way.

Ginet (1990) and McCann (1998) have set out the most fully developed recent noncausal views of free will. Other recent views of this type are advanced by Goetz (1988 and 1997) and McCall (1994: ch. 9).

Both Ginet and McCann hold that some event involving an agent is a basic action in virtue of its possessing some intrinsic feature. On Ginet's view, the feature in question is an "actish phenomenal quality," which he describes as its seeming to the agent as if she is directly producing the event that has this quality. McCann holds that the feature that characterizes these basic actions is intrinsic intentionality. For example, when one makes a decision, he holds, intrinsic to the decision is one's intending to make that very decision. (E.g., when one decides to A, one intends to decide to A.) One's so intending is not a matter of the content of the intention that is formed in deciding, nor is it a matter of one's having any further intention in addition to that formed in making the decision. Rather, McCann holds, it is a matter of a decision's being, by its very nature, an exercise of agency.

As is characteristic of proponents of noncausal accounts, neither Ginet nor McCann place any additional positive requirements on free action; the further requirements are instead that certain conditions be absent. Both require that the action not be causally determined. Ginet requires, further, that in performing the action, the agent not be subject to irresistible compulsion.

Two main problems arise for noncausal accounts of free will; both are problems, in the first instance, for noncausal accounts of action. The first concerns control. Performing an action, even acting unfreely, is exercising some variety of active control over one's behavior. Acting freely is exercising an especially valuable variety of active control. An account of free decision or other free action ought to say what this latter variety of control is or in what it consists. A common objection is that noncausal accounts fail to meet this requirement.

The second main problem concerns rationality. Acting freely is acting with a capacity for rational self-governance and determining, oneself, whether and how one exercises that capacity on a given occasion. Hence it must be possible for a free action to be an action performed for a certain reason, an action for which there is a rational explanation. Again, it is often objected that noncausal views cannot provide adequate accounts of these phenomena.

1.1 Control

An obvious candidate for an account of active control is that an agent's exercising such control consists

in her action's being caused, in an appropriate way, by her, or by certain events involving here, such as her having certain beliefs and desires and a certain intention. Noncausal views reject such an account; let us consider some alternatives.

On Ginet's view, the single positive characteristic present in all actions is the actish phenomenal quality, its seeming to the agent as if she is directly making happen the event that is her action. But it cannot be true, he holds, that we are agent-causes of our actions, nor need it be the case that any events involving us cause them. Ginet stresses that his description of this quality is metaphorical; the experience does not literally represent to the agent that she is bringing about the event in question.

Whatever the correct characterization of this phenomenal quality, it may be objected that the mere feel of a mental event, although it may be a (more or less reliable) sign of an individual's active control over that event, cannot itself *constitute* such control. The objection is reinforced by the fact that, on Ginet's view, an event with this quality could be brought about by external brain stimulation, in the absence of any relevant desire or intention in the "agent."

McCann holds that an agent's exercise of control has two aspects: it is a spontaneous, creative undertaking on the part of the agent, and it is intrinsically intentional. Causal theorists of action deny that the intentionality of, say, one's acquisition of an intention when one makes a decision can be intrinsic to that event. McCann observes that it is impossible for a decision to be unintentional, and he concludes from that observation that the intentionality must then be intrinsic to the decision. However, as Mele (1997) points out, an alternative explanation of this impossibility is available. For something to count as a decision, a causalist may hold, one's acquisition of an intention (to pursue some more or less specific course of action) must be caused, in an appropriate way, by one's having an intention to make up one's mind about what to do. Given the requirement of such a causal history, the causalist alternative continues, necessarily any event that is a decision is intentional. But this says nothing about the intrinsic features of decisions (any more than the fact that nothing counts as sunburn unless it is caused by the sun says anything about the intrinsic features of the burns that are so-caused).

The spontaneity of actions, McCann holds, "has a certain *sui generis* character that renders it incapable of being reduced to anything else" (p. 185). Such a claim suggests that no explication of active control is possible. This might be so if active control is a metaphysically simple phenomenon, a fundamental feature of the world. Causalists may object here that active control is comprehensible only as a causal phenomenon. Further, if active control is a simple phenomenon, then it would be emergent, a higher-level phenomenon appearing only in mentally sophisticated beings and yet not constituted by any lower-level phenomena. The question then arises whether there is evidence that there exist any such emergent, higher-level, metaphysically basic phenomena.

1.2 Rational Explanation

Turning now to acting for certain reasons and to rational explanation, again obvious candidates for accounts of the phenomena invoke causation: an agent acts for a certain reason only if the agent's having

that reason causes, in an appropriate way, the agent's behavior, and citing a reason contributes to a rational explanation of an action only if the agent's having that reason caused, in an appropriate way, the action. Let us consider alternatives offered by noncausalists.

Suppose that *S* wants her glasses, which she has left in her friend *R*'s room, where he is now sleeping. *S* also wants to wake *R*, because she desires his company, but she know that *R* needs some sleep, and hence she desires, too, not to wake him. *S* enters *R*'s room and retrieves her glasses, knowing as she does so that her action will satisfy her desire to wake *R*. What further facts about the situation could make it the case that *S* acts on her desire to get her glasses, and that citing that desire provides a true rational explanation of her action, while she does not act on her desire to wake *R*, and citing this latter desire does not give us a true rational explanation of what she did? Ginet maintains that the following conditions suffice for the truth of the explanation that cites *S*'s desire to get her glasses:

> (a) prior to entering the room, *S* had a desire to get her glasses, and
>
> (b) concurrently with her entering the room, *S* remembered that prior desire and intended of her entering the room that it satisfy (or contribute to satisfying) that desire.

Citing *S*'s desire to wake *R* will fail to give us a true rational explanation, Ginet holds, just in case *S* did not intend of her action that it satisfy (or contribute to satisfying) that desire.

Several objections may be raised against this account. Suppose that *S*'s desire to get her glasses played no role at all in bringing about her action, while her desire to wake *R*, of which she was fully aware when she acted, did play such a role. Causalists (see, e.g., Mele (1992: ch. 13)) will then deny that *S* really acted on her desire to get her glasses and that citing it truly explains her action. Moreover, Ginet's memory condition may not be necessary. If *S*'s desire to get her glasses is retained and remains fully conscious while she acts, there may be no need for her to remember it in order for her to be acting on it. Further problems attend the concurrent intention that is required. (McCann (1998: 162-63) raises some of these difficulties.) First, the required intention is a second-order attitude, an attitude about (among other things) another of the agent's own attitudes (a certain desire of hers). But it seems that *S* might act on her desire to get her glasses even if her only intention when she enters the room is to retrieve her glasses. Second, the concurrent intention to which Ginet appeals is said to make direct (or demonstrative) reference to the action, and Ginet holds that such direct reference requires that the intention be caused by the action (or at least by some initial segment of it). If that is so, then at least some part of the action has occurred before the concurrent intention is acquired. It does not then seem that the acquisition of that intention can figure in an account of what, if anything, rationally explains that part of the action. Finally, intention-acquisitions themselves can be explained by citing reasons. Since Ginet's account of the rational explanation of an action appeals to an intention, the question arises what can be said about the rational explanation of the acquisition of that intention. Repeating the same sort of account here would generate a regress, and Ginet offers no other sort.

On McCann's view, an agent decides for a certain reason, and citing that reason rationally explains the

decision, just in case, in cognizance of that reason, and in an intrinsically intentional act of intention formation, the agent forms an intention the content of which reflects the very goals presented in that reason. When *S* decides to enter *R*'s room, for example, she decides for the reason of getting her glasses only if the intention that she forms in making that decision is an intention to enter for the sake of getting her glasses. But such correspondence between the reasons for which one decides and the content of the decision may be too much to require. *S* may want her glasses so that she can finish reading a certain novel, which she may want to do so that she can contribute to the discussion in her book club tomorrow. Finishing the novel and getting ready for tomorrow's discussion may be among the reasons for which she decides to enter the room even if they are not included in the content of her decision. Finally, there will again be a clash of intuitions here between causalists and noncausalists, with the former maintaining that if *S*'s desire to get her glasses plays no role at all in bringing about her decision, then even if the content of her decision is to enter for the sake of getting her glasses, she does not really decide for that reason and citing it does not truly explain her decision. (One does not, the objection goes, make it so just by intending it to be so, no even by intrinsically intentionally intending it to be so.)

We have considered problems for noncausal accounts of *action*. However, since the noncausal views examined here place no positive requirements on *free action* beyond those that are placed on action, if they fail as adequate accounts of action, then *a fortiori* they fail as adequate accounts of free action.

2. Event-Causal Accounts

Compatibilist accounts of free action are typically event-causal accounts, invoking event-causal accounts of action. The simplest event-causal incompatibilist view takes the requirements of a good compatibilist account and adds that certain agent-involving events that cause the action must nondeterministically cause it. When these conditions are satisfied, it is held, the agent exercises in performing her action a certain variety of active control (which is said to consist in the action's being caused by those agent-involving events), the action is performed for reasons, and there was a chance of the agent's not performing that action. It is thus said to have been open to the agent to do otherwise, even given that (it is claimed) its being so open is incompatible with the truth of determinism.

One common objection against such a view is that the indeterminism that it requires is destructive, that it would diminish the control with which an agent acts. A second common objection is that the required indeterminism is gratuitous, that it adds nothing of value. We shall examine these objections below. First, let us consider a type of event-causal nondeterministic account that is advocated by writers who accept a qualified version of the first of these objections.

2.1 Deliberative Indeterminism

Some writers accept that indeterminism located in the immediate causation of a decision or other action would diminish control but hold that indeterminism confined to earlier stages in the processes leading to decision need not do so. Ekstrom (2000) and Mele (1995, 1996, and 1999b) have advanced the most fully

developed recent nondeterministic accounts of this sort. Such views have also been sketched by Dennett (1978) and Fischer (1995).

Overt action is sometimes preceded by a decision, and decision is sometimes preceded by a deliberative process in which the agent considers reasons for and against alternatives and makes an evaluative judgment concerning which alternative is best (or better or good enough). Focusing on decisions that follow such deliberation, Mele advances a view that allows (but does not require) the deterministic causation of the decision by the making of the judgment, and of the overt action by the decision. Indeterminism is required only at an earlier stage of the deliberative process. For example, the account is satisfied when it is undetermined which of a certain subset of the agent's nonoccurrent beliefs come to mind in the process of deliberation, where their coming to mind combines with other events to bring about the agent's evaluative judgment. (The subset in question consists of "beliefs whose coming or not coming to mind is not something that one would control even if determinism were true" (1995: 216).)

Mele argues that indeterminism of the sort required here does not diminish, at least not to any significant extent, "proximal control," a variety of control that is compatible with determinism. The required indeterminism nevertheless suffices, he holds, to provide the agent with "ultimate control" over her decision, which an agent has only if at no time prior to the decision is there any minimally causally sufficient condition for the agent's making that decision that includes no event or state internal to the agent.

In Ekstrom's account, the notion of preference, rather than that of an evaluative judgment, plays a prominent role. A preference, as she understands it, is a desire "formed by a process of critical evaluation with respect to one's conception of the good" (p. 106). The formation of a preference, she maintains, is an action. She requires indeterminism only in the production of these preferences. A decision or subsequent action is free, on her view, just in case it is brought about, in an appropriate way, by an active formation of a preference (favoring that decision or action), which preference-formation is in turn the result of an uncoerced exercise of the agent's evaluative faculty, the inputs to which (the considerations taken up in deliberation) nondeterministically cause the preference-formation.

Ekstrom holds that an agent *is* her preferences and acceptances (reflectively held beliefs), together with her faculty of forming these by reflective evaluation. When the formation of a preference is nondeterministically caused and it deterministically causes a decision and subsequent action, then, a preference that partly constitutes the agent, that is generated by an evaluative faculty that partly constitutes the agent, and that the agent could have prevented (by not forming that preference) causally determines the decision and subsequent action. What the agent does is then, Ekstrom holds, up to her.

Mele's and Ekstrom's views both allow that a free decision or other free action may be causally determined by events none of which are free actions. Indeed, on pain of regress, both must allow that a free decision or other free action may be causally determined by events none of which are free actions and to none of which has the agent contributed by her performance of any free action. (A familiar regress would be generated if, in order for a decision or other action to be free, it were required that it result from

the making of an evaluative judgment or the formation of a preference that was either itself a free action or to which the agent had contributed by her performance of some other free action.) Incompatibilists will typically not allow such a thing. Hence, an incompatibilist who finds either of these views acceptable will have something other than the typical incompatibilist's reasons for rejecting compatibilism.

2.2 Efforts of Will

Event-causal accounts of a more typical sort require that, for at least some free actions, there be nondeterministic causation of the free actions themselves. The most sophisticated recent account of this sort is that advanced by Kane (1996, 1999a, 1999b, and 2000). Other views of this type are sketched by Nozick (1981: 294-316), Sorabji (1980: ch. 2), van Inwagen (1983: 137-50), and Wiggins (1973).

A free decision or other free action, Kane holds, is one for which the agent is "ultimately responsible." Ultimate responsibility for an action requires either that the action not be causally determined or, if the action is causally determined, that some (nonredundant) part of any determining cause of it either be or result from some action by that agent that was not causally determined (and for which the agent was ultimately responsible). Thus, on Kane's view, an agent may be ultimately responsible for a decision that is causally determined by factors that include her having certain character traits. But somewhere among the events that contributed (however indirectly) to those traits, and thus to her decision, there must have been some actions by her that were not causally determined. Kane calls such "regress-stopping" actions "self-forming actions." All self-forming actions, he argues, are acts of will; they are mental actions. He thus calls them "self-forming willings," or SFWs. Kane identifies six different types of SFWs, giving the most detailed treatment to what he calls moral choices or decisions and prudential choices or decisions. We shall focus here on the former; the two are sufficiently similar that the points made here can be easily transferred to the latter.

In a case of moral choice, there is a motivational conflict within the agent. She believes that a certain type of thing morally ought to be done (and she is motivated to do that), but she also has a self-interested desire to perform an action of a type that is, in the circumstances, incompatible with her doing what she believes she ought to do. Given her commitment to her moral belief, she makes an effort of will to resist temptation, an effort "to get [her] ends or purposes sorted out" (1996: 126). If the choice is to be a SFW, then it is required that the strength of this effort be indeterminate; Kane likens its indeterminacy to that of the position or momentum of a microphysical particle. And the effort's indeterminacy is held to be the source of the required indeterminism in the causal production of the choice. Again an analogy is drawn with microphysics. Just as whether a particle will penetrate a barrier may be undetermined because the particle's position and momentum are not both determinate, so "[t]he choice one way or the other is *undetermined* because the process preceding and potentially terminating in it (i.e., the effort of will to overcome temptation) is *indeterminate*" (1996: 128).

Kane further requires that any choice that is a SFW satisfy three plurality conditions. These require that the choice be made for reasons (which Kane takes to consist partly in the choice's being caused by the

agent's having those reasons) and that it not be a result of coercion or compulsion. Each plurality condition also requires that, when the agent makes the choice, she wants more to act on the reasons for which she makes that choice than she wants to act on any competing reasons. An agent wants more to act on certain reasons, he holds, when her desire to act on those reasons has greater motivational strength than have any desires she has to act on competing reasons, and when it is settled in the agent's mind that those reasons, rather than her reasons for doing otherwise, are the ones that she will now and in the future act on. This wanting more to act on certain reasons is, on Kane's view, brought about by the choice in question. Finally, the plurality conditions require that, whichever choice is made, there have been at least one alternative choice that the agent was able to make such that, had she made it, it too would have satisfied the previously stated conditions.

In a situation of moral conflict, Kane maintains, the requirements for being a SFW may be satisfied by either choice that is made--the choice to do what one believes one ought to do or the choice to do what one is tempted to do. Where this is so, whichever choice the agent makes, she has chosen for the reasons that she wants more to act on, free from coercion and compulsion. If she has chosen to do what she believes she ought to do, then her choice is the result of her effort. If she has chosen to do what she was tempted to do, then she has not allowed her effort to succeed. Whichever choice she has made, she could have made the other. She is then ultimately responsible for the choice she has made.

Kane's requirement that the causation of a choice that is a SFW be nondeterministic has drawn the objection that indeterminism located here would diminish the agent's control over the making of the choice. The objection is often couched in terms of a problem of luck. (It is so developed by Haji (1999), Mele (1998, 1999a, and 1999b), and Strawson (1994).) If the agent's effort of will nondeterministically causes her choice, then, whichever choice the agent makes, there was, until the occurrence of that choice, a chance that it would not occur. If the agent's effort to chose in accord with her moral judgment happens to succeed, the objection goes, then her choice is at least partly due to good luck. In another possible world, with exactly the same laws of nature and exactly the same in its course of events up until the occurrence of the choice, the agent's (or her counterpart's) effort fails; there, but for good luck, goes she. And analogously, if, in the actual world, the agent's effort fails, then her choice is at least partly due to bad luck. Either way, the choice is to some degree due to luck. And to that degree, the objection concludes, the control that the agent exercises in making the choice is diminished.

Kane offers a complex reply to this objection. First, he counters that with indeterminate events, exact sameness is not defined. If an agent's effort of will was indeterminate, then it cannot be that she and her counterpart made exactly the same effort, and one got lucky while the other did not. An objection that assumes that such exact sameness is possible, he holds, does not apply to his view. Kane infers from this point that free will requires a form of indeterminism in which there is chance as well as indeterminacy, with the former stemming from the latter. (He calls worlds with such indeterminism "non-Epicurean.") The chance in an Epicurean world (an indeterministic world without indeterminacy), he implies, would constitute control-diminishing luck.

Kane's claim that indeterminacy precludes exact sameness has been contested (see Clarke (1999) and O'Connor (1996)). And Haji (1999) and Mele (1999a and 1999b) contend that the argument from luck is

just as effective if we consider an agent and her counterpart who is as similar as can be, given the indeterminacy of their efforts. Indeed, the argument might be advanced without any appeal to other worlds or counterparts: given that there is a chance that the effort will fail, the agent is lucky, it may be said, if it succeeds.

A further reply from Kane to the argument from luck appeals to the active nature of efforts of will. When an agent makes an effort to choose to do what she believes she ought to do, she actively tries to bring about a certain choice. When the agent makes that choice, she succeeds, despite the indeterminism, at doing what she was (actively) trying to do. And Kane points out that typically, when this is so, the indeterminism does not undermine responsibility (and hence it does not so diminish active control that there is not enough for responsibility). He describes a case (1999b: 227) in which a man hits a glass table top attempting to shatter it. Even if it is undetermined whether his effort will succeed, Kane notes, if the man does succeed, he may well be responsible for breaking the table top.

If left here, the reply would fail to address the problem of luck in a case where the agent chooses to do what she is tempted to do rather than what she believes she ought to do. In response to this shortcoming, Kane has recently proposed a "doubling" of effort in cases of moral conflict. In such a case, he now holds, the agent makes two, simultaneous efforts of will, both indeterminate in strength. The agent tries to make the moral choice, and at the same time she tries to make the self-interested choice. Whichever choice she makes, then, she succeeds, despite the indeterminism, at doing something that she was actively trying to do.

This doubling of efforts of will introduces a troubling incoherence into case of moral conflict. If an agent is actively trying, at one time, to do two obviously incompatible things, that fact raises a serious question about the agent's rationality.

A further question concerns the efficacy of the appeal to the active nature of these efforts. In the case of the man who breaks the table top, his breaking the table top is free (if it is) *not* just because it results from an active effort to break the table top, but because it results (we are to presume) from a *free* effort to break the table top. A successful effort to make a certain choice can contribute in an analogous way to the choice's being free, then, only if the effort itself is free. What is needed, then, is an account of the freedom-level active control with which the agent acts in making these efforts of will.

Kane maintains that, although the effort of will that precedes a choice that is a SFW must be an action for which the agent is ultimately responsible, the effort need not itself be a SFW; it is allowed that the effort be causally determined. However, if the agent is ultimately responsible for the effort and it is causally determined, then the agent must, on his view, have contributed (however indirectly), by her performance of at least one earlier SFW, to the effort in question. Since, on Kane's account, all SFWs either are efforts of one sort or another or must be preceded by efforts, the task of providing an account of the freedom of an effort cannot be avoided.

Kane faces the following dilemma in providing such an account. If the account of the freedom of an

effort of will requires that the effort itself result from a prior free effort, then a vicious regress looms. On the other hand, if the account of the freedom of an effort of will need not appeal to any prior free efforts of will, then it would seem that the account of a regress-stopping free choice could likewise dispense with such an appeal.

2.3 A Simpler View

Kane's appeal to indeterminate efforts of will, and the appeal thereby to non-Epicurean indeterminism, do not appear to help meet the objection that indeterminism located in the causation of decisions or other actions themselves diminishes control. (Neither does it appear that help comes from his requirement that, in making a choice that is a SFW, the agent come to want more to act on the reasons for which she makes that choice. For, on Kane's view, this wanting more is brought about by the choice. And, if an event-causal view is on the right track, the agent's control over the making of the choice is a matter of the production of the choice, not of what the choice produces.) A far simpler event-causal incompatibilist account, then, may fare as well against the problem of control. How badly would it fare?

Suppose that Elena is considering whether to *A* or to *B*. She recognizes what she regards as a fairly strong reason to *A*, and she recognizes what she regards as a somewhat weaker reason to *B*. At a certain time, *t*, she decides to *A*. Consider a view on which the prior deliberative events nondeterministically cause that decision. The view need not require, though it may allow, that there was a chance that at *t* Elena would decide to *B*. Incompatibilism requires here that something other than deciding at *t* to *A* have been causally open, and that requirement will be met if there was a chance that at *t* Elena would continue deliberating, seeking further reasons for or against the two alternatives she is considering.

It certainly is not clear that the chance required here constitutes control-diminishing luck. However, the open alternatives required may not be sufficiently robust to satisfy many incompatibilists, who may want to require that, at least sometimes, when an agent makes a decision, there was a chance that at that time she instead make an alternative decision, even if any such alternative would have been contrary to her assessment of the reasons that she had recognized. Suppose the view requires that, in Elena's case, there have been a chance that she would at *t* decide to *B*. Does this sort of chanciness constitute control-diminishing luck?

Note that Elena's case will still differ from the following sort. Suppose that Lucas throws a dart, attempting to hit a target, which he succeeds in doing. Due to certain properties of the dart and the air, the process leading from his releasing the dart and ending in the dart's hitting the target was nondeterministic, and there was a chance of the dart's missing the target. Lucas exercises (indirect) active control over the dart's hitting the target only by way of his prior action of throwing the dart. The indeterminism in this case (in comparison with a case in which his throwing the dart causally determines its hitting the target) diminishes the chance of his succeeding at bringing about a nonactive result that he is actively trying to bring about. It is for this reason that indeterminism here constitutes control-diminishing luck.

But the indeterminism in Elena's case is located differently; it is located in the causal connection between certain nonactive events--Elena's recognizing and assessing certain reasons--and her performance of a basic action--her making a decision. She exercises active control over the making of the decision not by performing any prior action, but by making the decision. If indeterminism, in the form of nondeterministic causation, located here diminishes control, the explanation of why it does so will have to be different from the explanation of why the indeterminism in Lucas' case diminishes control. It is not obvious what this alternative explanation would be.

Even if the indeterminism required by an event-causal incompatibilist account does not diminish control, it does not appear to increase it, and a second objection may be raised, charging that the requirement is then gratuitous. In order to assess this second objection, it may prove helpful to reflect on why free will is important to us.

We value a freedom in acting that grounds dignity and responsibility, in the exercise of which we make a difference to the way the world goes, and one that accords with the appearance of openness that we find in deliberating. We can distinguish two aspects of this freedom: a kind of leeway or openness of alternatives, and a type of control that is exercised in action. The freedom in which we are interested for some of the above things may involve one but not the other of these aspects. (For example, Frankfurt (1969) has argued that an agent may be responsible for what she has done even if she could not have done otherwise.) In a similar fashion, it may be that what is gained with the indeterminism required by an event-causal incompatibilist account has to do with one but not the other of these aspects.

An agent's exercise of control in acting is her exercise of a positive power to determine what she does. If event-causal views are correct, this is a matter of the action's being caused, in an appropriate way, by certain events involving the agent, such has her having certain reasons and a certain intention. An event-causal incompatibilist account adds no new causes to those that can be required by a compatibilist account, and hence the former appears to add nothing to the agent's positive power to determine what she does. As far as this aspect of freedom is concerned, the requirement of indeterminism may well be (at best) superfluous.

But not so when it comes to the other aspect, the openness of more than one course of action. If incompatibilists are correct, there is never any such openness if the world is deterministic. The indeterminism required by an event-causal incompatibilist account suffices to secure this leeway or openness, and this may be important to us for several reasons. Some individuals, at least, may find that when they deliberate, they cannot help but presume that more than one course of action is genuinely open to them. If the world is in fact deterministic (and if incompatibilism is true), these individuals are subject to an unavoidable illusion (since we cannot avoid deliberating). And they may reasonably judge that it would be, for this reason, better if things are as presented in an event-causal incompatibilist view. Similarly, some individuals may reasonably judge that if things are as presented in this view, that is better with regard to our decisions' and other actions' making a difference to how the world goes. Of course, even if the world is deterministic, there is a way in which our decisions and other actions generally make a difference: had we not made those decisions and performed those actions, things would have gone differently. If things are as presented in an event-causal incompatibilist account, our decisions

and other actions still generally make a difference in this way. But they may make a difference in a second way as well: they may be branch points in a probabilistic unfolding of history, branch points over which we exercise active control. There may have been a real chance of things' not going a certain way, and these decisions and other actions may be the events that set things going that way. One may reasonably judge that it is better to be making a difference in this second as well as in the first way with one's decisions and other actions. Since we cannot be making a difference in this second way if the world is deterministic, some individuals may have reason to find that the indeterminism required by an event-causal incompatibilist view is not superfluous but adds something of value.

It is less clear that anything is to be gained with respect to responsibility. As was suggested above, it does not appear that on an event-causal incompatibilist account, agents exercise any greater positive powers of control than they could in a deterministic world. If Frankfurt is right, and the openness provided by such an account is not required for responsibility, then whether the account secures responsibility would appear to depend just on what positive powers of control it offers. Hence, if responsibility is not compatible with the truth of determinism, it may not be compatible with the truth of an event-causal incompatibilist account, either.

3. Agent-Causal Accounts

If, on an event-causal incompatibilist view, agents do not exercise any greater positive powers of control than they can on a compatibilist account, what type of incompatibilist view would secure greater control? A number of incompatibilists have maintained that such a view must hold that a free decision or other free action (or some component of it), while not causally determined by events, is caused by the agent, and that causation by an agent does not consist in causation by events (such as the agent's having certain reasons). An agent, it is said, is a continuant or substance, and hence not the kind of thing that can itself be an effect (though various events in its life can be). On these agent-causal accounts, then, an agent is in a strict and literal sense an originator of her free actions, an uncaused cause of her behavior. This combination of indeterminism and origination is thought to capture best the kind of freedom we desire with respect to dignity, responsibility, difference-making, and the appearance of openness.

An early advocate of this type of view was Reid (1969 [1788]). In recent years, agent-causal accounts have been advanced by Chisholm (1966), (1971), (1976a), (1976b), and (1978), Clarke (1993) and (1996), Donagan (1987), O'Connor (1995), (1996), and (2000), Rowe (1991), Taylor (1966) and (1992), Thorp (1980), and Zimmerman (1984).

Two main problems confront defenders of agent-causal accounts, one concerning the notion of agent causation and the other concerning the rational explicability of free decisions or other actions on such views.

All theorists who accept a causal construal of agents' control over what they do--and this includes most compatibilists as well as many incompatibilists--hold that, in a sense, agents cause their free actions. However, most hold that causation by an agent is just causation by certain events involving the agent,

such as the agent's having certain reasons and a certain intention. But, as we have seen, the agent causation posited by agent-causal accounts is held not to be this at all. It is said by most agent-causal theorists to be fundamentally different from event causation. And this raises the question whether any intelligible account of it can be given. Even some proponents of agent-causal views (e.g., Taylor and Thorp) seem doubtful about this, declaring agent causation to be strange or even mysterious.

Moreover, even if the notion of agent causation can be made intelligible, the question remains whether the thing itself--causation by a substance or continuant--is possible. An argument deriving from Broad (1952: 215) suggests that it is not. Each event, including each action, it is said, occurs at a certain time. And if an action is caused, the argument continues, then some part of that action's total cause must be an event, something that itself occurs at a certain time. Otherwise there would be no way to account for the action's occurring when it did. Hence, if an agent causes an action, there must be something the agent does, or some change the agent undergoes, that causes that action. Since either something the agent does or some change the agent undergoes would be an event, it is concluded, it cannot be the case, as most agent-causal accounts maintain, that free actions are caused by agents and not by any events.

The second main problem for agent-causal views stems from the fact that free actions can be performed for reasons and can be rationally explicable. We saw earlier that serious difficulties confront noncausal accounts of acting for certain reasons and of rational explanation. In denying, then, as most agent-causalists do, that free actions have any event causes, these theorists may rule out rational free action.

In response to this second problem, Clarke (1993 and 1996) proposes an agent-causal account on which a free action is caused by the agent *and* nondeterministically caused by certain agent-involving events, such as the agent's having certain reasons and a certain intention. Given this appeal to reasons-causation, the view can provide the same accounts of acting for reasons and of rational explanation as can event-causal views. And since the event causation that is posited is required to be nondeterministic, the view secures the openness of alternatives, even on the assumption that this openness is incompatible with determinism. Finally, the agent causation itself is still held to be distinct from causation by any events, and so this view secures the origination of free actions that seemed an appealing feature of more traditional agent-causal accounts.

Even though, on this type of agent-causal view, a free action is nondeterministically caused by events involving the agent, since the agent makes a further causal contribution to what she does in addition to the contribution made by those events, it would seem that she exercises greater positive powers of control than what could be exercised if all causes were events. Hence this type of account may have a stronger defense against the problem of control than have event-causal incompatibilist accounts.

This modification of traditional agent-causal views also addresses Broad's objection concerning the possibility of agent causation. That objection concludes that it cannot be the case that free actions are caused by agents and not by any events; if an agent causes an action, it is said, then some event involving that agent must cause the action and account for the action's occurring when it does. On the proposed view, some events involving the agent do cause each free action and account for the action's occurring

when it does.

O'Connor rejects this combination of agent and event causation. A free action, on his view, consists in an agent's causing a certain event, and he maintains that an agent's causing an event cannot be caused. O'Connor defends an account of rational explanation similar to that of Ginet, with an added requirement (for explanation that cites a certain desire) that the action begin with the agent's causing her acquisition of an intention to act so as to satisfy that desire.

Turning to the intelligibility of agent causation, O'Connor (1995, 1996, and 2000) and Clarke (1993 and 1996), though differing on details, have both suggested that agent causation might be characterized along the same lines as event causation, if the latter is given a nonreductive account. Familiar reductive accounts characterize event causation in terms of constant conjunction or counterfactual dependence or probability increase, and if event causation is so characterizable, then certainly agent causation would have to be fundamentally different. But if causation is a basic, irreducible feature of the world, then we might with equal intelligibility be able to think of substances as well as events as causes.

Even if we can understand the idea of agent causation, and even if the argument for its impossibility considered earlier is not effective, there remain reasons to doubt that it is possible for a substance to cause something. To give just one example, even if causation cannot be reduced to probability increase, it seems plausible that any cause must be the kind of thing that can affect the probability of its effect prior to the occurrence of that effect, even when the cause directly brings about that effect. Events are the sort of thing that can so affect probabilities, and this is due, it seems, to the fact that they occur at times. Substances do not occur (events involving them do), and they do not appear to be the sort of thing that can affect probabilities in the indicated way. This consideration, although not decisive, seems to count against the possibility of causation by a substance.

4. The Evidence

Our assessment of incompatibilist accounts so far has primarily focused on whether they satisfactorily characterize what free will would be, if there is such a thing. However, even if one or another of these views characterizes well the freedom that we value, and even if what that account characterizes is something that is possible, the question remains whether there is good evidence that what is posited by that account actually exists.

Incompatibilist accounts require, first, that determinism be false. But more than this, they require that there be indeterminism of a certain sort (e.g., with some events entirely uncaused, or nondeterministically caused, or caused by agents and not deterministically caused by events) and that this indeterminism be located in specific places (generally, in the occurrence of decisions and other actions). What is our evidence with regard to these requirements' being satisfied?

It is sometimes claimed (e.g., by Campbell (1957: 168-70) and O'Connor (1995: 196-97)) that our

experience when we make decisions and act constitutes evidence that there is indeterminism of the required sort in the required place. We can distinguish two parts of this claim: one, that in deciding and acting, things appear to us to be the way that one or another incompatibilist account says they are, and two, that this appearance is evidence that things are in fact that way. Some compatibilists (e.g., Mele (1995: 135-37)) deny the first part. But even if this first part is correct, the second part seems dubious. If things are to be the way they are said to be by some incompatibilist account, then the laws of nature--laws of physics, chemistry, and biology--must be a certain way. (This is so for overt, bodily actions regardless of the relation between mind and body, and it is so for decisions and other mental actions barring a complete independence of mental events from physical, chemical, and biological events.) And it is incredible that how things seem to us in making decisions and acting gives us insight into the laws of nature. Our evidence for the required indeterminism, then, will have to come from the study of nature, from natural science.

The scientific evidence for quantum mechanics is sometimes said to show that determinism is false. Quantum theory is indeed very well confirmed. However, there is nothing approaching a consensus on how to interpret it, on what it shows us with respect to how things are in the world. Indeterministic as well as deterministic interpretations have been developed, but it is far from clear whether any of the existing interpretations is correct. Perhaps the best that can be said here is that, given the demise of classical mechanics and electromagnetic theory, there is no good evidence that determinism is true.

The evidence is even less decisive with respect to whether there is the kind of indeterminism located in exactly the places required by one or another incompatibilist account. Unless there is a complete independence of mental events from physical events, then even for free decisions there has to be indeterminism of a specific sort at specific junctures in certain brain processes. There are some interesting speculations in the works of some incompatibilists about how this might be so (see, e.g., Kane (1996: 128-30 and 137-42) and the sources cited there); but our current understanding of the brain gives us no evidence one way or the other about whether it is in fact so. At best, it seems we must remain, for the time being, agnostic about this matter.

If incompatibilist free will requires agent causation, and if such a thing is possible, that is another requirement about which we lack evidence. Indeed, it is not clear that there could be any empirical evidence for or against this aspect of agent-causal views.

In sum, we do not have good evidence that any incompatibilist account is true. Some incompatibilists (e.g., van Inwagen (1983: 204-13)) hold that we nevertheless have good reason to believe that some such view is correct, since we have good reason to believe that we are morally responsible, that moral responsibility requires free will, and that free will requires indeterminism. However, lacking empirical evidence for the required indeterminism, if we justifiably believe the last two of the just mentioned propositions, then we have a strong moral reason not to treat each other as morally responsible. For if we are not responsible, then whenever we treat someone as responsible, we do that individual an injustice. And if indeterminism of a certain sort and in a certain location is required for responsibility and we lack evidence for the required indeterminism, then we risk this injustice whenever we treat someone as responsible. That is a strong moral reason (for incompatibilists) not to do so.

Bibliography

- Broad, C. D. 1952. *Ethics and the History of Philosophy*. London: Routledge & Kegan Paul.
- Campbell, C. A. 1957. *On Selfhood and Godhood*. London: George Allen & Unwin.
- Chisholm, Roderick M. 1966. "Freedom and Action." In Keith Lehrer, ed., *Freedom and Determinism*. New York: Random House. Pp. 11-44.
- Chisholm, Roderick M. 1971. "reflections on Human Agency." *Idealistic Studies* 1: 33-46.
- Chisholm, Roderick M. 1976a. "The Agent as Cause." In Myles Brand and Douglas Walton, eds., *Action Theory*. Dordrecht: D. Reidel. Pp. 199-211.
- Chisholm, Roderick M. 1976b. *Person and Object: A Metaphysical Study*. La Salle: Open Court.
- Chisholm, Roderick M. 1978. "Comments and Replies." *Philosophia* 7: 597-636.
- Clarke, Randolph. 1993. "Toward a Credible Agent-Causal Account of Free Will." *Nous* 27: 191-203.
- Clarke, Randolph. 1996. "Agent Causation and Event Causation in the Production of Free Action." *Philosophical Topics* 24, No. 2: 19-48.
- Clarke, Randolph. 1999. "Free Choice, Effort, and Wanting More." *Philosophical Explorations* 2: 20-41.
- Clarke, Randolph. forthcoming-a. "Freedom of the Will." In Stephen Stich and Ted A. Warfield, eds., *The Blackwell Guide to Philosophy of Mind*. Blackwell.
- Clarke, Randolph. forthcoming-b. "Libertarian Views (II): Critical Survey of Noncausal and Event-Causal Theories. In Robert Kane, ed., *The Free Will Handbook*. Oxford University Press.
- Dennett, Daniel C. 1978. "On Giving Libertarians What They Say They Want." In *Brainstorms: Philosophical Essays on Mind and Psychology*. Montgomery, Vermont: Bradford Books. Pp. 286-99.
- Donagan, Alan. 1987. *Choice: The Essential Element in Human Action*. London: Routledge & Kegan Paul.
- Ekstrom, Laura Waddell. 2000. *Free Will: A Philosophical Study*. Boulder: Westview Press.
- Fischer, John Martin. 1995. "Libertarianism and Avoidability: A Reply to Widerker." *Faith and Philosophy* 12: 119-25.
- Frankfurt, Harry G. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66: 828-39.
- Ginet, Carl. 1990. *On Action*. Cambridge: Cambridge University Press.
- Goetz, Stewart. 1988. "A Noncausal Theory of Agency." *Philosophy and Phenomenological Research* 49: 303-16.
- Goetz, Stewart. 1997. "Libertarian Choice." *Faith and Philosophy* 14: 195-211.
- Haji, Ishtiyaque. 1999. "Indeterminism and Frankfurt-type Examples." *Philosophical Explorations* 2: 42-58.
- Kane, Robert. 1996. *The Significance of Free Will*. New York: Oxford University Press.
- Kane, Robert. 1999a. "On Free Will, Responsibility and Indeterminism." *Philosophical Explorations* 2: 105-21.
- Kane, Robert. 1999b. "Responsibility, Luck, and Chance: Reflections on Free Will and

Indeterminism." *Journal of Philosophy* 96: 217-40.

- Kane, Robert. 2000. "Responses to Bernard Berofsky, John Martin Fischer and Galen Strawson." *Philosophy and Phenomenological Research* 60: 157-67.
- McCall, Storrs. 1994. *A Model of the Universe*. Oxford: Clarendon Press.
- McCann, Hugh J. 1998. *The Works of Agency: On Human Action, Will, and Freedom*. Ithaca: Cornell University Press.
- Mele, Alfred R. 1992. *Springs of Action: Understanding Intentional Behavior*. New York: Oxford University Press.
- Mele, Alfred R. 1995. *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.
- Mele, Alfred R. 1996. "Soft Libertarianism and Frankfurt-Style Scenarios." *Philosophical Topics* 24, No. 2: 123-41.
- Mele, Alfred R. 1997. "Agency and Mental Action." *Philosophical Perspectives* 11: 231-49.
- Mele, Alfred R. 1998. Review of *The Significance of Free Will*. *Journal of Philosophy* 95: 581-84.
- Mele, Alfred R. 1999a. "Kane, Luck, and the Significance of Free Will." *Philosophical Explorations* 2: 96-104.
- Mele, Alfred R. 1999b. "Ultimate Responsibility and Dumb Luck." *Social Philosophy & Policy* 16: 274-93.
- Nozick, Robert. 1981. *Philosophical Explanations*. Cambridge, Massachusetts: Belknap Press.
- O'Connor, Timothy. 1995. "Agent Causation." In O'Connor, ed., *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. New York: Oxford University Press. Pp. 173-200.
- O'Connor, Timothy. 1996. "Why Agent Causation?" *Philosophical Topics* 24, No. 2: 143-58.
- O'Connor, Timothy. 2000. *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.
- Reid, Thomas. 1969 [1788]. *Essays on the Active Powers of the Human Mind*. Cambridge, Massachusetts: MIT Press.
- Rowe, William L. 1991. *Thomas Reid on Freedom and Morality*. Ithaca: Cornell University Press.
- Sorabji, Richard. 1980. *Necessity, Cause, and Blame: Perspectives on Aristotle's Theory*. Ithaca: Cornell University Press.
- Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75: 5-24.
- Taylor, Richard. 1966. *Action and Purpose*. Englewood Cliffs: Prentice-Hall.
- Taylor, Richard. 1992. *Metaphysics*, 4th edition. Englewood Cliffs: Prentice-Hall.
- Thorp, John. 1980. *Free Will: A Defence Against Neurophysiological Determinism*. London: Routledge & Kegan Paul.
- Van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.
- Wiggins, David. 1973. "Towards a Reasonable Libertarianism." In Ted Honderich, ed., *Essays on Freedom of Action*. London: Routledge & Kegan Paul. Pp. 33-61.
- Zimmerman, Michael J. 1984. *An Essay on Human Action*. New York: Peter Lang.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

action | causation: probabilistic | compatibilism | determinism, causal | free will | incompatibilism: arguments for | moral responsibility | quantum theory: and free will

Randolph Clarke
rclarke@arches.uga.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 16, 2000
Content last modified: August 16, 2000

Moral Responsibility

When a person performs or fails to perform a morally significant action, we sometimes think that he or she is deserving of a particular kind of response. Praise and blame are perhaps the most common forms this reaction takes. For example, one who encounters a car accident may be worthy of praise for having saved a child from inside the burning car, or alternatively, one may be worthy of blame for not having at least used one's cell phone to call for help. To regard such agents as worthy of one of these reactions is to ascribe moral responsibility to them on the basis of what they have done or left undone. (These are examples of other-directed ascriptions of responsibility. The reaction might also be self-directed, e.g., one can recognize oneself to be blameworthy). Thus, to be morally responsible for something, say an action, is to be worthy of a particular kind of reaction -- praise, blame, or something akin to these -- for having performed it.[1]

Though further elaboration and qualification of the above characterization of moral responsibility is called for and will be provided below, this is enough to distinguish concern about this form of responsibility from some others commonly referred to through use of the terms ‘responsibility’ or ‘responsible.’ To illustrate, we might say that higher than normal rainfall in the spring is responsible for an increase in the amount of vegetation or that it is the judge's responsibility to give instructions to the jury before they begin deliberating. In the first case, we mean to identify a causal connection between the earlier amount of rain and the later increased vegetation. In the second, we mean to say that when one assumes the role of judge, certain duties, or obligations, follow. Although these concepts are connected with the concept of moral responsibility, they are not the same, for in neither case are we directly concerned about whether it would be appropriate to react to some candidate (here, the rainfall or a particular judge) with something like praise or blame.[2]

Philosophical reflection on moral responsibility has a long history. One reason for this persistent interest is the way the topic seems connected with a widely shared conception of ourselves as members of an importantly distinct class of individuals -- call them ‘persons.’[3] Persons are thought to be qualitatively different from those of other known living species, despite their numerous similarities. Many have held that one distinct feature of persons is their status as morally responsible agents, a status resting, perhaps, on a special kind of control only they can exercise. Many who view persons in this way have wondered whether their special status is threatened if certain other claims about our universe are true. For example, can a person be morally responsible for her behavior if that behavior can be explained solely by reference to physical states of the universe and the laws governing changes in those physical states, or solely by reference to the existence of a sovereign God who guides the world along a divinely ordained path? Concerns like these have often motivated individuals to theorize about moral responsibility.

A comprehensive theory of moral responsibility would elucidate the following: (1) the concept, or idea, of moral responsibility itself; (2) the criteria for being a moral agent, i.e., one who qualifies generally as an agent open to responsibility ascriptions (e.g., only beings possessing the general capacity to evaluate reasons for acting can be moral agents); (3) the conditions under which the concept of moral responsibility is properly applied, i.e., those conditions under which a moral agent is responsible for a particular something (e.g., a moral agent can be responsible for an action she has performed only if she performed it freely, where acting freely entails the ability to have done otherwise at the time of action); and finally 4) possible objects of responsibility ascriptions (e.g., actions, omissions, consequences, character traits, etc.). Although each of these will be touched upon in the discussion below (see, e.g., the brief sketch of Aristotle's account in the next section), the primary focus of this entry is on the first component --i.e., the concept of moral responsibility. The first section below is a discussion of the origin and history of Western reflection on moral responsibility. This is followed by a sketch of recent work on the concept of moral responsibility. For further discussion of issues associated with moral responsibility, see the related entries below.

Some Historical Background

What follows in this section is a brief outline of the origins and trajectory of reflection on moral responsibility in the Western philosophical tradition. Against this background, a distinction will be drawn between two conceptions of moral responsibility that have exerted considerable influence on subsequent thinkers.

An understanding of the concept of moral responsibility and its application is present implicitly in some of the earliest surviving Greek texts, i.e., the Homeric epics (circa 8th century BCE but no doubt informed by a much earlier oral tradition).[4] In these texts, both human and superhuman agents are often regarded as fair targets of praise and blame on the basis of how they have behaved, and at other times, an agent's behavior is excused because of the presence of some factor that has undermined his/her control (Irwin, 1999, p. 225). Reflection on these factors gave rise to *fatalism* -- the view that one's future or some aspect of it is predetermined, e.g., by the gods, or the stars, or simply some facts about truth and time -- in such a way as to make one's particular deliberations, choices and actions irrelevant to whether that particular future is realized (recall, e.g., the plight of Oedipus). If some particular outcome is fated, then it seems that the agent concerned could not be morally responsible for that outcome. Likewise, if fatalism were true with respect to all human futures, then it would seem that no human agent could be morally

responsible for anything. Though fatalism has sometimes exerted significant historical influence, most philosophers have rejected it on the grounds that there is no good reason to think that our futures are fated in the sense that they will unfold no matter what particular deliberations we engage in, choices we make, or actions we perform.

Aristotle (384-323 BCE) seems to have been the first to construct explicitly a theory of moral responsibility.[5] In the course of discussing human virtues and their corresponding vices, Aristotle pauses in *Nicomachean Ethics* III.1-5 to explore their underpinnings. He begins with a brief statement of the concept of moral responsibility -- that it is sometimes appropriate to respond to an agent with praise or blame on the basis of her actions and/or dispositional traits of character (1109b30-35). A bit later, he clarifies that only a certain kind of agent qualifies as a moral agent and is thus properly subject to ascriptions of responsibility, namely, one who possess a capacity for decision. For Aristotle, a decision is a particular kind of desire resulting from deliberation, one that expresses the agent's conception of what is good (1111b5-1113b3). The remainder of Aristotle's discussion is devoted to spelling out the conditions under which it is appropriate to hold a moral agent blameworthy or praiseworthy for some particular action or trait. His general proposal is that one is an apt candidate for praise or blame if and only if the action and/or disposition is voluntary. According to Aristotle, a voluntary action or trait has two distinctive features. First, there is a control condition: the action or trait must have its origin in the agent. That is, it must be up to the agent whether to perform that action or possess the trait -- it cannot be compelled externally. Second, Aristotle proposes an epistemic condition: the agent must be aware of what it is she is doing or bringing about (1110a-1111b4).[6]

There is an instructive ambiguity in Aristotle's account of responsibility, an ambiguity that has led to competing interpretations of his view. Aristotle aims to identify the conditions under which it is appropriate to praise or blame an agent, but it is not entirely clear how to understand the pivotal notion of appropriateness in his conception of responsibility. There are at least two possibilities: a) praise or blame is appropriate in the sense that the agent *deserves* such a response, given his behavior and/or traits of character; or b) praise or blame is appropriate in the sense that such a reaction is likely to bring about a desired consequence, namely an improvement in the agent's behavior and/or character. These two possibilities may be characterized in terms of two competing interpretations of the concept of moral responsibility: 1) the *merit-based view*, according to which praise or blame would be an appropriate reaction toward the candidate if and only if she merits -- in the sense of ‘deserves’ -- such a reaction; vs. 2) the *consequentialist view*, according to which praise or blame would be appropriate if and only if a reaction of this sort would likely lead to a desired change in the agent and/or her behavior.[7]

Scholars disagree about which of the above views Aristotle endorsed, but the importance of distinguishing between them grew as philosophers began to focus on a newly conceived threat to moral responsibility. While Aristotle argued against a version of fatalism (*On Interpretation*, ch. 9), he may not have recognized the difference between it and the related possible threat of *causal determinism* (contra Sorabji). Causal determinism is the view that everything that happens or exists is caused by sufficient antecedent conditions, making it impossible for anything to happen or be other than it does or is. One variety of causal determinism, *scientific determinism*, identifies the relevant antecedent conditions as a combination of prior states of the universe and the laws of nature. Another, *theological determinism*,

identifies those conditions as being the nature and will of God. It seems likely that theological determinism evolved out of the shift, both in Greek religion and in Ancient Mesopotamian religions, from polytheism to belief in one sovereign God, or at least one god who reigned over all others. The doctrine of scientific determinism can be traced back as far as the Presocratic Atomists (5th cent. BCE), but the difference between it and the earlier fatalistic view seems not to be clearly recognized until the development of Stoic philosophy (3rd. cent. BCE). Though fatalism, like causal determinism, might seem to threaten moral responsibility by threatening an agent's control, the two differ on the significance of human deliberation, choice, and action. If fatalism is true, then human deliberation, choice, and action are completely otiose, for what is fated will transpire no matter what one chooses to do. According to causal determinism, however, one's deliberations, choices, and actions will often be necessary links in the causal chain that brings something about. In other words, even though our deliberations, choices, and actions are themselves determined like everything else, it is still the case, according to causal determinism, that the occurrence or existence of yet other things depends upon our deliberating, choosing and acting in a certain way (Irwin,1999, pp. 243-249; Meyer, 1998, pp. 225-227, and Pereboom).

Since the Stoics, the thesis of causal determinism and its ramifications, if true, have taken center stage in theorizing about moral responsibility. During the Medieval period, especially in the work of Augustine (354-430) and Aquinas (1225-1274), reflection on freedom and responsibility was often generated by questions concerning versions of theological determinism, including most prominently: a) Does God's sovereignty entail that God is responsible for evil?; and b) Does God's foreknowledge entail that we are not free and morally responsible since it would seem that we cannot do anything other than what God foreknows we will do? During the Modern period, there was renewed interest in scientific determinism -- a change attributable to the development of increasingly sophisticated mechanistic models of the universe culminating in the success of Newtonian physics. The possibility of giving a comprehensive explanation of every aspect of the universe -- including human action -- in terms of physical causes now seemed much more plausible. Many thought that persons could not be free and morally responsible if such an explanation of human action were possible. Others argued that freedom and responsibility would not be threatened should scientific determinism be true. In keeping with this focus on the ramifications of causal determinism for moral responsibility, thinkers may be classified as being one of two types: 1) an *incompatibilist* about causal determinism and moral responsibility -- one who maintains that if causal determinism is true, then there is nothing for which one can be morally responsible; or 2) a *compatibilist* -- one who holds that a person can be morally responsible for some things, even if both who she is and what she does is causally determined.[8] In Ancient Greece, these positions were exemplified in the thought of Epicurus (341-270 BCE) and the Stoics, respectively.

Above, an ambiguity in Aristotle's conception of moral responsibility was highlighted -- that it was not clear whether he endorsed a merit-based vs. a consequentialist conception of moral responsibility. The history of reflection on moral responsibility demonstrates that how one interprets the concept of moral responsibility strongly influences one's overall account of moral responsibility. For example, those who accept the merit-based conception of moral responsibility have tended to be incompatibilists. That is, most have thought that if an agent were to genuinely merit praise or blame for something, then he would need to exercise a special form of control over that thing (e.g., the ability at the time of action to both perform or not perform the action) that is incompatible with one's being causally determined. In addition

to Epicurus, we can cite early Augustine, Thomas Reid (1710-1796), and Immanuel Kant (1724-1804) as historical examples here. Those accepting the consequentialist conception of moral responsibility, on the other hand, have traditionally contended that determinism poses no threat to moral responsibility since praising and blaming could still be an effective means of influencing another's behavior, even in a deterministic world. Thomas Hobbes (1588-1679), David Hume (1711-1776), and John Stuart Mill (1806-1873) are, along with the Stoics, representatives of this view. This general trend of linking the consequentialist conception of moral responsibility with compatibilism about causal determinism and moral responsibility and the merit-based conception with incompatibilism continued to persist through the first half of the twentieth century.

Recent Work on the Concept of Responsibility

The issue of how best to understand the concept of moral responsibility is important, for it can strongly influence one's view of what, if any, philosophical problems might be associated with the notion, and further, if there are problems, what might count as a solution. As discussed above, philosophical reflection on moral responsibility has historically relied upon one of two broad interpretations of the concept: 1) the *merit-based view*, according to which praise or blame would be an appropriate reaction toward the candidate if and only if she merits -- in the sense of 'deserves' -- such a reaction; or 2) the *consequentialist view*, according to which praise or blame would be appropriate if and only if a reaction of this sort would likely lead to a desired change in the agent and/or her behavior. Though the consequentialist view has continued to garner support (see e.g., Schlick, 1966, Brandt 1992, Frankena 1963, ch. 4, Dennett 1984, ch. 7, and Kupperman 1991, ch. 3), work in the last 50 years on the concept of moral responsibility has increasingly focused on: a) offering alternatives to the two traditional interpretations of moral responsibility; and b) alternative versions of the merit-based view.

Increased attention to the practice of *holding* persons morally responsible has generated much of the recent work on the concept of moral responsibility. All theorists have recognized features of this practice -- inner attitudes and emotions, their outward expression in censure or praise, and the imposition of corresponding sanctions or rewards. However, most understood the inner attitudes and emotions involved to rest on a more fundamental theoretical judgment about the agent's *being* responsible. In other words, it was typically assumed that blame and praise depended upon a judgment, or belief (pre-reflective in most cases), that the agent in question had satisfied the objective conditions on being responsible. These judgments were presumed to be independent of the inner attitudinal/emotive states involved in holding responsible in the sense that reaching such judgments and evaluating them required no essential reference to the attitudes and emotions of the one making the judgment. For the holder of the consequentialist view, this is a judgment that the agent exercised a form of control that could be influenced through outward expressions of praise and blame in order to curb or promote certain behaviors. For those holding the merit view, it is a judgment that the agent has exercised the requisite form of metaphysical control, e.g., that she could have done otherwise at the time of action (Watson 1987, p. 258).

If holding responsible is best understood as resting on an independent judgment about being responsible, then it is legitimate to inquire whether such underlying judgments and their associated outward

expressions can be justified, as a whole, in the face of our understanding of the world, e.g., in the face of evidence that our world is possibly deterministic.[9] According to incompatibilists, a judgment that someone is morally responsible could never be true if the world were deterministic; thus praising and blaming in the merit-based sense would be beside the point. Compatibilists, on the other hand, contend that the truth of determinism would not undermine the relevant underlying judgments concerning the efficacy of praising and blaming practices, thereby leaving the rationale of such practices intact.

In his landmark essay, ‘Freedom and Resentment,’ P. F. Strawson (1962) sets out to adjudicate the dispute between those compatibilists who hold a consequentialist view of responsibility and those incompatibilists who hold the merit-based view.[10] Both are wrong, Strawson believes, because they distort the concept of moral responsibility by sharing the prevailing view sketched above -- the view of those who understand the practice of holding persons responsible to rest upon a theoretical judgment of their being responsible. According to Strawson, the attitudes expressed in holding persons morally responsible are varieties of a wide range of attitudes deriving from our participation in personal relationships, e.g., resentment, indignation, hurt feelings, anger, gratitude, reciprocal love, and forgiveness. The function of these attitudes is to express ". . . how much we actually mind, how much it matters to us, whether the actions of other people -- and particularly *some* other people -- reflect attitudes towards us of good will, affection, or esteem on the one hand or contempt, indifference, or malevolence on the other." (p. 5, author's emphasis) These attitudes are thus *participant reactive attitudes*, because they are: a) natural attitudinal reactions to the perception of another's good will, ill will, or indifference (pp. 4-6), and b) expressed from the stance of one who is immersed in interpersonal relationships and who regards the candidate held responsible as a participant in such relationships as well (p. 10).[11]

The reactive attitudes can be suspended or modified in at least two kinds of circumstances, corresponding to the two features just mentioned. In the first, one might conclude that, contrary to first appearances, the candidate did not violate the demand for a reasonable degree of good will. For example, a person's behavior may be *excused* when one determines that it was an accident, or one may determine that the behavior was *justified*, say, in the case of an emergency when some greater good is being pursued. In the second kind of circumstance, one may abandon the participant perspective in relation to the candidate. In these cases, one adopts the *objective standpoint*, one from which one ceases to regard the individual as capable of participating in genuine personal relations (either for some limited time or permanently). Instead, one regards the individual as psychologically/morally abnormal or undeveloped and thereby a candidate, not for the full range of reactive attitudes, but primarily for those objective attitudes associated with treatment or simply instrumental control. Such individuals lie, in some sense or to some varying extent, outside the boundaries of the moral community. For example, we may regard a very young child as initially exempt from the reactive attitudes (but increasingly less so in cases of normal development) or adopt the objective standpoint in relation to an individual we determine to be suffering from severe mental illness (P. F. Strawson, pp. 6-10, Bennett, p. 40, Watson 1987, p. 259-260, R. Jay Wallace, chs. 5-6).

The central criticism Strawson directs at both consequentialist and traditional merit views is that both have over-intellectualized the issue of moral responsibility, and it is this criticism that has rendered his view so influential in subsequent work.[12] The charge of over intellectualization stems from the

traditional tendency to presume that the rationality of holding a person responsible depends upon a judgment that the person in question has satisfied some set of objective requirements on being responsible (conditions on efficacy or metaphysical freedom) and that these requirements themselves are justifiable. Strawson, by contrast, maintains that the reactive attitudes are a natural expression of an essential feature of our form of life, in particular, the interpersonal nature of our way of life. The practice, then, of holding responsible -- embedded as it is in our way of life -- "neither calls for nor permits, an external 'rational' justification (p. 23)." Though judgments about the appropriateness of particular responses may arise (i.e., answers to questions like: Was the candidate's behavior really an expression of ill will?; or Is the candidate involved a genuine participant in the moral sphere of human relations?), these judgments are based on principles internal to the practice. That is, their justification refers back to an account of the reactive attitudes and their role in personal relationships, *not to some independent theoretical account of the conditions on being responsible*.

Given the above, Strawson contends that it is pointless to ask whether the practice of holding responsible can be rationally justified if determinism is true. This is either because it is not psychologically possible to divest ourselves of these reactions and so continually inhabit the objective standpoint, or even if that were possible, because it is not clear that rationality could ever demand that we give up the reactive attitudes, given the loss in quality of life should we do so. In sum, Strawson attempts to turn the traditional debate on its head, for now *judgments about being responsible* are understood in relation to the role reactive attitudes play in the *practice of holding responsible*, rather than the other way around. Whereas judgments are true or false and thereby can generate the need for justification, the desire for good will and those attitudes generated by it possess no truth value themselves, thereby eliminating any need for an external justification (Magill, p. 21, Double, 1996b, p. 848).

Strawson's concept of moral responsibility yields a compatibilist account of being responsible but one that departs significantly from earlier such accounts in two respects. First, Strawson's is a compatibilist view by default only. That is, on Strawson's view, the problem of determinism and freedom/responsibility is not so much *resolved* by showing that the objective conditions on being responsible are consistent with one's being determined but rather *dissolved* by showing that the practice of holding people responsible relies on no such conditions and therefore needs no external justification in the face of determinism. Second, Strawson's is a merit-based form of compatibilism. That is, unlike most former consequentialist forms of compatibilism, it helps to explain why we feel that some agents deserve our censure or merit our praise. They do so because they have violated, met, or exceeded our demand for a reasonable degree of good will. A number of compatibilists have followed Strawson in this regard, incorporating the reactive attitudes into their theories of moral responsibility in order to provide an account that reflects a merit-based conception of moral responsibility.

Most agree that Strawson's discussion of the reactive attitudes is a valuable contribution to our understanding of the practice of holding responsible, but many have taken issue with Strawson's contentions about the insular nature of that practice, namely a) that since propriety judgments about the reactive attitudes are strictly internal to the practice (i.e., being responsible is defined in relation to the practice of holding responsible), their justification cannot be considered from a standpoint outside that practice; and b) that since the reactive attitudes are natural responses deriving from our psychological

constitution, they cannot be dislodged by theoretical considerations. Responding to the first of these, some have argued that it does seem possible to critique existing practices of holding responsible from standpoints outside them. For example, one might judge that either one's own existing community practice or some other community's practice of holding responsible ought to be modified (Fischer and Ravizza, 1993, p. 18). If such evaluations are legitimate, then, contrary to what Strawson suggests, the justification of an existing practice can be questioned from a standpoint external to it. In other words, being responsible cannot be explicated strictly in terms of an existing practice of holding responsible. This, then, would suggest a possible role to be played by theoretical conditions on being responsible, whether they are compatibilist or incompatibilist in nature.

Replying to the second of the above contentions, some have argued that incompatibilist intuitions are embedded in the reactive attitudes themselves so that these attitudes cannot persist unless some justification can be given of them, or more weakly, that they cannot but be disturbed if something like determinism is true. Here, cases are often cited where negative reactive attitudes seem to be dispelled or mitigated upon learning that an agent's past includes severe deprivation and/or abuse. There is a strong pull to think that our reactive attitudes are altered in such cases because we perceive such a background to be deterministic. If this is the proper interpretation of the phenomenon, then it is evidence that theoretical considerations, like the truth of determinism, could in fact dislodge the reactive attitudes (Nagel, p. 125; Kane, 84-89; Galen Strawson, 1986, p. 88, Honderich, 1988, vol. 2: ch. 1; and replies by Watson, 1987, 1996, p. 240; and McKenna, 1998).

Finally, some have taken this last point -- that incompatibilist intuitions may be embedded in or necessarily linked to the reactive attitudes -- further, challenging Strawson's and others' assumption that the reactive attitudes form a unitary class with respect to issues of freedom and responsibility. Instead, this group argues that the reactive attitudes are linked to bifurcated and contradictory ideas about freedom and moral responsibility, suggesting some level of incoherence in the concepts themselves (Nagel, G. Strawson, 1986,105-117, 307-317; Honderich, 1988, vol. 2: ch. 1; Double, 1991; Double 1996a, chs. 6-7). Some, for example Ted Honderich, have argued that our idea of a free and responsible agent as *an originator* of action is incompatibilistic, while our idea of free and responsible action as *voluntary* is compatibilistic.[13] Reflection on causal determinism thus generates conflicting attitudes. Honderich believes that a compromise, of sorts, can be reached between these opposing ideas and associated attitudes given the truth of determinism -- that in spite of this one can train oneself to adopt a coherent and meaningful conception of both freedom and responsibility (Honderich, 1988, vol. 2: ch. 1, Honderich 1996). Others, including Richard Double, contend that the incongruity of our ideas and attitudes concerning freedom and moral responsibility are radically irreconcilable, which entails that our conceptions of freedom and responsibility are incoherent. On this view, there is no correct view about or attitudinal response to determinism, if the latter is descriptive of the nature of our world (Double, 1991, Double 1996a-b).

The future direction of reflection on moral responsibility is uncertain. On the one hand, there has been a resurgence of interest in metaphysical treatments of freedom and moral responsibility in recent years, a sign that many philosophers in this area have not been persuaded by Strawson's central critique of such treatments. On the other hand, the interpretations and further development of Strawson's work have

increased in both their significance and influence. What is clear is that the long-standing interest in understanding the concept of moral responsibility and its application shows no sign of abating.

Bibliography

- Adams, Robert Merrihew. "Involuntary Sins." *Philosophical Review* 94 (1985): 3-31.
- Aquinas, Thomas. *Basic Writings of St. Thomas Aquinas*, ed. A. C. Pegis (Indianapolis: Hackett Publishing Co., 1997).
- Aristotle. *The Nicomachean Ethics*, trans. by Terence Irwin. (Indianapolis: Hackett Publishing Co., 1985).
- ________. *The Complete Works of Aristotle: The Revised Oxford Translation, ed. Jonathan Barnes, 2 Vols. (Princeton: Princeton University Press, 1984).*
- Augustine. *On Free Choice of the Will* (Indianapolis: Hackett Publishing Co., 1993).
- Austin, J.L. "A Plea for Excuses" in *Philosophical Papers*, J.O. Urmson and G.J. Warnock, eds. (New York: Oxford University Press, 1979).
- Ayer, A.J. "Free Will and Rationality" in van Straatan.
- Bair, Annette. *A Progress of Sentiments: A Reflection on Hume's Treatise*. (Cambridge, MA: Harvard University Press, 1991).
- Benson, Paul. "The Moral Importance of Free Action." *Southern Journal of Philosophy* 28 (1990): 1-18.
- Berofsky, Bernard, ed. *Free Will and Determinism*. (New York: Harper and Row, Pub., 1966).
- Bennett, Jonathan. "Accountability" in *Philosophical Subjects*, Zak Van Straaten, ed. (Oxford: Clarendon Press, 1980).
- Brandt, Richard. "A Utilitarian Theory of Excuses" *The Philosophical Review* 78 (1969):337-361. Reprinted in *Morality, Utility, and Rights*. (New York: Cambridge University Press, 1992).
- ________. *Ethical Theory*. (Englewood Cliffs, NJ: Prentice Hall, Inc., 1959).
- ________. "Blameworthiness and Obligation" in Meldon 1958.
- Broadie, Sarah. *Ethics with Aristotle*. (New York: Oxford University Press, 1991).
- Curren, Randall. "The Contribution of *Nicomachean Ethics* iii.5 to Aristotle's Theory of Responsibility." *History of Philosophy Quarterly* 6(1989): 261-277.
- Dennett, Daniel. *Elbow Room: The Varieties of Free Will Worth Wanting*. (Cambridge, MA: MIT Press, 1984).
- Double, Richard. *Metaphilosophy and Free Will*. (New York: Oxford University Press, 1996a).
- ________. "Honderich on the Consequences of Determinism." *Philosophy and Phenomenological Research* 66 (December, 1996b): 847-854.
- ________. *The Non-reality of Free Will*. (New York: Oxford University Press, 1991).
- Everson, Stephen, ed. *Companions to Ancient Thought 4: Ethics*. (New York: Cambridge University Press, 1998).
- ________. "Aristotle's Compatibilism in the *Nicomachean Ethics*." *Ancient Philosophy* 10 (1990):81-103.
- Feinberg, Joel. *Doing and Deserving: Essays in the Theory of Responsibility* (Princeton: Princeton University Press, 1970).

- Fingarette, Herbert. *On Responsibility*. (New York: Basic Books, Inc., 1967).
- Fischer, John Martin. "Recent Work on Moral Responsibility" *Ethics* 110 (October 1999): 93-139.
- ________. *The Metaphysics of Free Will: An Essay on Control*. (Cambridge, MA: Blackwell Pub., 1994).
- ________, ed. *Moral Responsibility* (Ithaca: Cornell University Press, 1986).
- Fischer, John Martin and Ravizza, Mark. *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge University Press, 1998).
- ________, eds. *Perspectives on Moral Responsibility* (Cornell University Press, 1993).
- Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66 (1969): 828-839.
- Gibbard, Allan. *Wise Choices, Apt Feelings*: *A Theory of Normative Judgment* (Cambridge, MA: Harvard University Press, 1990).
- Glover, Jonathan. *Responsibility* (New York: Humanities Press, 1970).
- Haji, Ishtiyaque. *Moral Appraisability*. (New York: Oxford University Press, 1998).
- Hart, H. L. *Punishment and Responsibility*. (New York: Oxford University Press, 1968).
- Honderich, Ted. "Compatibilism, Incompatibilism, and the Smart Aleck." *Philosophy and Phenomenological Research* 66 (December, 1996): 855-862.
- ________. *A Theory of Determinism: The Mind, Neuroscience, and Life Hopes*. 2 Vols. (Oxford: Clarendon Press, 1988)
- Hume, David. *A Treatise of Human Nature*, 2nd ed., ed. by L.A. Selby-Bigge and P.H. Nidditch. (New York: Oxford University Press, 1978).
- Irwin, Terrance, ed. *Classical Philosophy*. (New York: Oxford University Press, 1999).
- ________. "Reason and Responsibility in Aristotle." in Rorty 1980.
- Kane, Robert. *The Significance of Free Will*. (New York: Oxford University Press, 1996).
- Kant, Immanuel. *The Critique of Practical Reason*, trans. by Lewis White Beck, 3rd. ed. (Englewood Cliffs, NJ: Macmillan Publishing Co., 1993).
- Kupperman, Joel. *Character*. (New York: Oxford University Press, 1991).
- Magill, Kevin. *Freedom and Experience: Self-Determination without Illusions*. (New York: St. Martins Press, 1997).
- McKenna, Michael. "The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism." *Journal of Ethics.*
- Meldon, A.I., ed. *Essays in Moral Philosophy*. (Seattle: University of Washington Press, 1958).
- Meyer, Susan Suave. "Moral Responsibility: Aristotle and After." in Everson 1998.
- ________. *Aristotle on Moral Responsibility*. (Cambridge, MA: Blackwell Pub., 1993).
- Mill, John Stuart. *A System of Logic*, 8th ed. (New York: Harper and Brothers, 1884).
- Milo, Ronald D. *Immorality* (Princeton, NJ: Princeton University Press, 1984).
- Nagel, Thomas. *The View From Nowhere.* (New York: Oxford University Press, 1986).
- Nozick, Robert. *Philosophical Explanations*. (Cambridge, MA: Harvard University Press, 1981).
- Oshana, Marina. "Ascriptions of Responsibility." *American Philosophical Quarterly* 34 (1997): 71-83.
- Pereboom, Derk, ed. *Free Will*. (Indianapolis: Hackett Publishing Co., 1997).
- Roberts, Jean. "Aristotle on Responsibility for Action and Character." *Ancient Philosophy* 9 (1984): 23-36.

- Rorty, Amelie Oksenberg, ed. *Essays on Aristotle's Ethics.* (Los Angeles: University of California Press, 1980).
- Russell, Paul. *Freedom and Moral Sentiment: Hume's Way of Naturalizing Responsibility*. (New York: Oxford University Press, 1995).
- ________. "Strawson's Way of Naturalizing Responsibility." *Ethics* 102 (1992): 287-302.
- Schlick, Moritz. "When is a Man Responsible," in Berofsky, 1966.
- Schoeman, Ferdinand, ed. *Responsibility, Character, and the Emotions*. (New York: Cambridge University Press, 1987)
- Sorabji, Richard. *Necessity, Cause, and Blame* (Ithaca: Cornell University Press, 1980).
- Stern, Lawrence. "Freedom, Blame, and the Moral Community." *The Journal of Philosophy* 71 (1974): 72-84.
- Strawson, Galen. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75 (1994): 5-24.
- ________. *Freedom and Belief*. (New York: Oxford University Press, 1986).
- Strawson, P. F. "Freedom and Resentment." *Proceedings of the British Academy* 48 (1962):1-25. Reprinted in Fischer and Ravizza, 1993.
- van Inwagen, Peter. *An Essay on Free Will*. (New York: Oxford University Press, 1978).
- van Stratten, Z., ed. *Philosophical Subjects: Essays Presented to P.F. Strawson* (New York: Oxford University Press, 1980).
- Wallace, James. "Excellences and Merit." *Philosophical Review* 83 (1974): 182-199.
- Wallace, R. J. *Responsibility and the Moral Sentiments*. (Cambridge, MA: Harvard University Press, 1994).
- Watson, Gary. "Two Faces of Responsibility." *Philosophical Topics* 24 (1996): 227-248.
- ________. "Responsibility and the Limits of Evil." in Schoeman, 1986.
- Williams, Bernard. *Shame and Necessity*. (Los Angeles: University of California Press, 1993).
- Wolf, Susan. "The Importance of Free Will." *Mind* 90 (1981): 386-405.
- Zimmerman, Michael. *An Essay on Moral Responsibility*. (Totowa, NJ: Roman and Littlefield, 1988).

Other Internet Resources

- P. S. Greenspan's Recent Research

[Please contact the author with further suggestions.]

Related Entries

Aristotle: ethics | compatibilism | determinism, causal | fatalism | free will | incompatibilism: (nondeterministic) theories of free will | luck: moral | responsibility: collective

Andrew S. Eshleman
aseshleman@ualr.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 5, 2001
Content last modified: January 5, 2001

Stanford Encyclopedia of Philosophy
Notes to Moral Responsibility

Notes

[1.] In an effort to streamline the following discussion, I have chosen to restrict my focus to morally significant actions (and possibly other items--e.g., traits--subject to moral evaluation) and have assumed that moral responsibility involves both positive and negative reactions like praise and blame. In doing so, I set aside two important ways of understanding the notion of responsibility, advocates of which might object to these introductory remarks. First, some think that the scope of responsibility is not restricted to actions (and other items) subject to moral evaluation but in principle applies to all intentional actions (as well as other items that may be linked to such actions. See e.g., Fischer and Ravizza, 1988, ch. 1; and Haji, ch. 1). Second, some contend that the idea of moral responsibility, properly explicated, involves reference to forms of blame alone (see e.g., R. J. Wallace, p. 12).

[2.] For further discussion of various kinds of responsibility, see Hart, ch. 9; and Feinberg, pp. 130-139.

[3.] The term, ‘person,’ is here being used as a technical term. This is important to realize because it is an open question whether the class of persons is co-extensive with the class of human beings. This may be either because there are (or someday may be) persons who are not human beings or because not all human beings qualify as persons in the relevant restricted sense.

[4.] For discussion of a challenge to this claim, see Williams, ch. 2-3.

[5.] Curren and Roberts have challenged the traditional view that Aristotle was discussing a conception of moral responsibility similar to our own. For a noteworthy defense of the traditional view in the face of such challenges, see Meyer, 1993, chs. 1-2.

[6.] This way of dividing control and epistemic conditions on responsibility continues to be influential. For excellent treatments of Aristotle's account, see Broadie, ch.3; Everson, 1990; Irwin, 1980; Meyer, 1993; and Roberts.

[7.] The ambiguity displayed in Aristotle's conception of moral responsibility is linked to another—how to interpret the condition that the action or trait be *up to the agent*. Some believe that the action (or trait) needs to be up to the agent in the sense that she could have, at the time of action (or development of the trait), either performed it or not performed it (or developed it vs. not developed it). Others believe the action or trait need be up to the agent only in the sense that it follows from the agent's desires/emotions/beliefs in such a way that had she decided otherwise, she would have done or been

otherwise. Those who adopt the former reading of Aristotle's account adopt the merit reading of Aristotle's conception of responsibility while the latter often adopt the consequentialist reading.

[8.] Until recently, this classification was parasitic upon and parallel to a more fundamental distinction between 1) those who believed that it could not both be true that persons sometimes acted freely and that persons were causally determined; and 2) those who believed both could be true. That is, acting freely—in the sense of being able to do otherwise--was assumed to be a precondition of being morally responsible for an action, so one's view of the compatibility of freedom and determinism was thought to entail one's view about the compatibility of moral responsibility and determinism. These assumptions have been called into question in recent years (see Frankfurt), opening up the possibility of views which are incompatibilist in one sense but compatibilist in another (see Fischer, 1994, 178-183). This recent wrinkle is ignored in the text for ease of exposition.

[9.] A contemporary example of this general way of thinking is a version of the merit view that has come to be dubbed, the ‘ledger view’ because of its reliance on the metaphor of a ledger, or report card. According to the ledger view, credit and debit entries are made on one's ledger on the basis of how one conducts one's life. Holding a person responsible is understood centrally in terms of a judgment that he or she has such a credit or debit. The associated attitudes, feelings, and their outward expression are byproducts of this judgment. These judgments are justified—i.e., the agent is praiseworthy or blameworthy for the existence of the credit or debit—when and only when the objective conditions on being responsible are met. On this view, then, particular judgments about an agent's responsibility can be justified only if there is reason to think that agents, in general, can meet such objective conditions. In other words, particular judgments rest upon the judgment that the practice of holding responsible can be rationally defended against possible defeaters (e.g., causal determinism). For examples of the ledger view, see Glover 1970, p. 64; Feinberg 1970, pp. 30-31; and Zimmerman 1988, pp. 38-39. For helpful discussions of the view, see Watson 1987, p. 262; Fischer and Ravizza 1993, pp. 16-17; and Fischer and Ravizza, 1998, pp. 8-10, nt. 12.

[10.] For helpful discussions of Strawson's view, see Bennett; Watson, 1987; Russell; Fischer and Ravizza, 1993, pp. 14-22; R. Jay Wallace; Magill, pp.19-22; and McKenna.

[11.] Note how this addresses one of the concerns motivating reflection on moral responsibility (see introduction). If Strawson is correct, then his view helps to explain one reason why persons are unique--namely, only they can be proper recipients of the reactive attitudes.

[12.] A more particular criticism of Strawson's aimed at consequentialist interpretations of the concept of responsibility has also been influential. This criticism follows from his account of the kind of attitudes involved in holding someone morally responsible and is nicely captured by Jonathan Bennett: "Displays of indignation or of gratitude often produce good results: but such feelings cannot be motivated by the desire to produce good results, nor, it seems, are we able closely to control them by thoughts of what will bring best results (p. 22)." In other words, Strawson denies that our practice of holding responsible is being driven by a desire to bring about good results and a judgment that our reacting in certain ways will

bring about that result (and/or that we could comprehensively adjust our reactions in accordance with such a goal), as the traditional consequentialist view seems to suggest. To be preoccupied with the goal of achieving the best results is to regard the candidate as an object of manipulation or treatment, a stance which precludes the participant reactive attitudes altogether. According to Strawson, any view that fails to acknowledge the essential role of the reactive attitudes within the participant perspective can no longer claim to be an account of *moral* responsibility, as we know it. (P. F. Strawson, 20-21). Others, building on Strawson's account, have argued that the consequentialist account also fails to capture the inherent backward-looking focus of the reactive attitudes (e.g., they are often reactions to what someone has *done*), focusing as they do on the achievement of some future goal (see Bennett, p. 37 and R. Jay Wallace, pp. 56-58). To criticisms of this kind, some consequentialists have argued that one must not confuse the intentions of the person engaged in the practice of holding responsible with the function of the overall practice. On this view, the consequentialist account is aimed at describing and justifying the function of holding a person responsible, not the intentions involved in doing so (see e.g., Kupperman, 60-64).

[13.] There are interesting parallels here with Gary Watson's "Two Faces of Responsibility", wherein Watson argues that there are two aspects to our concept of moral responsibility—one concerned with whether an action is attributable to an agent and another focusing on whether the agent is accountable for the action. But Watson does not go so far as to suggest that either of these aspects entails incompatibilism or compatibilism.

Andrew S. Eshleman
aseshleman@ualr.edu

First published: January 5, 2001
Content last modified: January 5, 2001

Medieval Theories of Relations

The purpose of this entry is to provide a systematic introduction to medieval views concerning the nature and ontological status of relations. Given the current state of our knowledge of medieval philosophy, especially with regard to relations, it is not possible to provide a detailed survey of the views of even the most important medieval philosophers. In what follows, therefore, I shall restrict my discussion to identifying and describing (a) the main types of position that were developed during the Middle Ages, and (b) the most important historical and dialectical considerations that shaped their development. Along the way, however, I will have occasion to examine in detail certain aspects of the views developed by important representatives of all the main medieval positions, including Peter Abelard (1079-1142), Gilbert of Poitiers (1085-1154), Albert the Great, (1200-1280), Thomas Aquinas (1225-1274), John Duns Scotus (1265-1308), Henry Harclay (1270-1317), Peter Aureoli (1280-1322), and William Ockham (1285-1347).

Introduction

It will be useful to begin this entry with a general introduction to the medieval discussion of relations, highlighting some of the main historical developments to be treated at more length in subsequent sections.

All theorizing about relations in the Middle Ages begins with Aristotle's short treatise, the *Categories*. Due to historical circumstances, this Aristotelian text was one of the very few pieces of ancient Greek philosophy available in the Latin west between the seventh and twelfth centuries, and the only one to contain a systematic philosophical treatment of relations. In the *Categories*, Aristotle identifies relations as one of the ten highest kinds or categories, one of the so-called *summa genera*, and he devotes an entire chapter -- the seventh chapter of the treatise -- to distinguishing the members of this category from those of the other categories (most notably, substance, quantity, and quality).[1] On the standard medieval interpretation of this chapter, Aristotle is attempting to distinguish relations at least partly on the basis of semantic or logical considerations -- that is, on the basis of the differences between statements containing relational (or 'relative') terms and those containing only non-relational (or 'absolute') terms. This approach to characterizing relations helps to shape the way medieval philosophers understand them, and, as we shall see, it gives rise to a common medieval distinction between things that qualify as relations merely according to speech (*relationes secundum dici*) and things that qualify as relations in accordance with their nature or being (*relationes secundum esse*).

In addition to discussing the proper characterization of relations, Aristotle also suggests in the *Categories* a general model or paradigm for analyzing relational situations. According to this analysis, whenever two substances are related, what relates them is a pair of monadic properties or accidents. Thus, if Socrates is similar to Theaetetus (i.e., resembles him with respect to some quality)[2], this is not to be explained by an entity to which Socrates and Theaetetus are somehow jointly attached (namely, the dyadic or two-place property, *being-similar-to*). Rather, it is to be explained by a pair of monadic properties or accidents, one of which inheres in Socrates and relates him to Theaetetus, and the other of which inheres in Theaetetus and relates him to Socrates.[3]

This Aristotelian analysis exercises enormous influence during the Middle Ages, and until at least the fourteenth century, medieval philosophers develop their own analysis of relational situations in terms of it. In the context of this analysis the most important debate that arises concerns the precise nature of the properties or accidents that relate particular substances. Some philosophers, such as Peter Abelard and William Ockham, adopt a form of reductivism, according to which the properties in question are accidents falling under categories other than relation. Thus, according to Abelard and Ockham, when Socrates is similar to Theaetetus, this is to be explained in terms of particular qualities, say, their respective colors. Other philosophers, however, such as Albert the Great and John Duns Scotus, reject this form of reductivism, maintaining instead that relations are accidents of a *sui generis* type. Thus, according to Albert and Scotus, when Socrates is similar to Theaetetus, this is to be explained not by their respective colors, but by a pair of accidents whose members are distinct from, and irreducible to, these colors.

Although the analysis of relational situations that Aristotle suggests in the *Categories* appears to be perfectly general, it is clear from his later writings that he does not think all relational situations conform to it. Thus, in the *Metaphysics* he claims that there are relational situations, such as Simmias's thinking about Socrates, in which two substances are related not by virtue of a *pair* of accidents, but in virtue of a single accident possessed by only one of them. Thus, he says "the object of thought [say, Socrates] is said to be related because something else [say, Simmias] is related to it."[4] Some relational situations, in other words, are grounded in a single property or accident of a single *relatum*.

As is well known, Aristotle's *Metaphysics* did not become available to philosophers in the Latin west until the mid-twelfth century, and it did not circulate widely before the thirteenth. Nonetheless, medieval philosophers were independently motivated -- largely on the basis of theological considerations -- to allow for relational situations that fail to conform to the paradigm suggested in the *Categories*. Thus, on the basis of considerations having to do with the doctrine of creation, they were led to acknowledge the existence of situations in which two substances are related by a single property or accident. Similarly, on the basis of considerations having to do with the Christian doctrine of the Trinity, they were led to admit relational situations in which substances themselves qualify as relations.[5] Prior to the fourteenth century, however, relational situations such as these are regarded by most philosophers as special cases or exceptions to the general rule. Indeed, philosophers from this period work hard to make even these sorts of cases conform, at least to some extent, to the paradigm of the *Categories*. This effort, as we shall see, helps to explain the pervasiveness in medieval philosophy of the distinction between the so-called real relations (*relationes reales*) and relations of reason (*relationes rationis*).

The fourteenth century marks an important turning point in the medieval discussion of relations. After this point, there is a noticeable shift away from the Aristotelian paradigm, with the result that many philosophers come to regard those situations in which substances are related by virtue of their accidents as the exception rather than the rule. Part of the explanation is due, perhaps, to a gradual waning of Aristotle's influence, but part of it is also due to substantive doctrinal changes. By the fourteenth century, for example, it is common to deny the existence of any real distinction between a substance and most of accidents, and consequently, philosophers are increasingly willing to say that substances themselves provide the ontological grounding for relational situations. Again, due to important semantic innovations around the same time, and the subsequent emergence of late-medieval nominalism, philosophers begin to depart also from the traditional Aristotelian characterization of relations. Thus, instead of thinking of relations as the items responsible for relating two or more substances, they now begin to think of them as items existing only in the mind -- that is, as mere beings of reason or concepts. With respect to both of these developments, as we shall see, Ockham is an important transitional figure.

2. Aristotelian Background

Although Aristotle discusses relations at various places in his work, his discussion in the *Categories* is the most important for understanding the development of medieval views about relations. Part of the reason, as I have already indicated, is that until the twelfth century, the *Categories* is medieval philosophers' sole direct source for Aristotle's views about relations. But part of the reason is also theoretical in nature: in the

Categories Aristotle develops a basic account of relations that is presupposed by his other discussions. Thus, even when some of these other discussions -- such as those in the *Physics* and *Metaphysics* -- become available, medieval philosophers interpret them in light of their understanding of the *Categories*.

Aristotle's discussion in the *Categories* influences the development of medieval views about relations in several ways. First of all, it introduces the basic framework and terminology that comes to dominate all medieval theorizing about relations. Second, and perhaps most importantly, it articulates the main claims in response to which medieval philosophers develop their own views. Indeed, as we shall see, medieval views about relations can be distinguished largely by the extent to which they agree with three claims commonly associated with Aristotle's discussion in the *Categories*: (1) relations are the items that relate substances, (2) the items that relate substances are accidents, and (3) no substance is a relation. Finally, the details of Aristotle's discussion in the *Categories*, especially in chapter 7, give rise to a number of important medieval debates, as well as to the traditional distinction between relations merely according to speech (*secundum dici*) vs. relations according to being (*secundum esse*). In order to understand the development of medieval views about relations, therefore, we must begin with some account of Aristotle's discussion in the *Categories*.

2.1 Framework and Terminology of the *Categories*

Throughout the *Categories* Aristotle assumes that relations comprise one of the accidental categories, and hence that they must be understood as items *inhering* in particular substances. This helps to explain why, unlike contemporary philosophers who speak of relations as holding *between* two or more things, Aristotle prefers to say that relations inhere in one thing and somehow point toward (*pros*) another. Indeed, Aristotle's preferred name for relations just is “things toward something” (*ta pros ti*).

During the Middle Ages, it is customary for philosophers to refer to relations using not only Aristotle's term, “toward something” (or *ad aliquid*, the verbatim Latin equivalent of *pros ti*), but two others as well: “relative” (*relativum*) and “relation” (*relatio*). In the early Middle Ages, philosophers often move freely among these terms without paying much attention to their various senses.[6] Over time, however, they come to emphasize their differences, sometimes giving elaborate explanations as to why one of the three terms is more appropriate than the others. To give just one example, consider the following passage by Albert the Great:

> Now the most general genus in the arrangement of this predicable is toward something (*ad aliquid*), or relative (*relativum*), or less properly speaking, relation (*relatio*), as some people say. But it must be recognized that the most general genus is signified most clearly by the name “toward something”, which is a preposition together with [the term] “something” in [the accusative] case. For this name conveys the two things that are in a relative, namely: [i] diversity, which the preposition indicates through its taking an object (*transitionem*), and [ii] the direction of the comparison, which the accusative case indicates when something is called *toward something*.[7]

Whatever their disagreements about the appropriateness of these terms, it was generally agreed that the terms “toward something” and “relative” are related to the term “relation” in the way that a concrete term (such as “white”) is related to its abstract counterpart (such as “whiteness”).

Since “relation” is the term most familiar to us, in what follows I shall rely on it whenever possible. It is important to recognize, however, that medieval philosophers introduce a number of other abstract terms more or less synonymous with “relation” (*relatio*) whose connotations they often regard as more informative than any of those discussed so far. A partial list of such terms includes: “comparison” (*comparatio*), which has psychological overtones, and so is often (though not exclusively) used by conceptualists; “outward-looking-ness” (*respectus*), which draws on a visual metaphor to suggest that relations are that in virtue of which a substance “looks out toward something” (*respicit ad aliquid*); “disposition” (*habitudo*) or “relative disposition” (*habitudo relativa*), which suggests that relations account for the way a thing “holds itself toward something” (*se habere ad aliquid*); and “order” (*ordo*), “ordering” (*ordinatio*), and “directionality” or “toward-ness” (*aditas*), which are used to indicate that relations account for the order or structure we find in the world.

Although Aristotle makes scattered remarks about relations throughout the *Categories*, it is only in chapter 7 that he singles them out for detailed independent consideration. Because the discussion in this chapter is important for understanding the development of the medieval debate, it will be useful to consider how it was commonly understood by philosophers from this period.

2.2 Relations in *Categories* 7

Aristotle's discussion in *Categories* 7 opens rather abruptly with the statement of a particular definition or account of relations. According to this definition, which medieval philosophers attribute to Plato, relations are things spoken of in a certain way. Whether or not this definition is really Plato's, it certainly appears to be well entrenched by Aristotle's time -- as is indicated by the fact that Aristotle says we are at least prepared to *call* something a relation just in case it satisfies the conditions specified by this definition:

> We call the following sorts of things toward something: all those things said to be just what they are *of* or *than* something, or *toward* something in some other way (any other way whatsoever). Thus, what is larger is said to be what it is *than* another (it is said to be larger than something); and a double is said to be just what it is *of* another (it is said to be double of something); similarly with all other such cases.[8]

As the medievals often interpret it, this definition identifies relations in terms of the predicates by means of which they are signified or “spoken of”. Medievals refer to the predicates in question as “relative terms” (*ad aliquid* or *relativa*)[9], and understand them, roughly speaking, as those terms whose true predication requires a comparison to something other than the subject of which they are predicated. Thus, “taller” (*maius*) is a relative term, they say, because when we assert of something that it is “taller” -- that is, when we predicate the term “taller” of it -- we necessarily do so in comparison to something else. We don't say merely that Simmias is taller; we say that he is taller *than* Socrates, or Theaetetus, or the average

man. Similar remarks apply, they say, to “double”. Borrowing on the medievals' behalf the notation of first-order logic, we can make this characterization of relative terms precise by saying that a term *F* is relative just in case a predication of the form “*Fx*” is more perspicuously represented as a predication of the form “*Rxy*”. Medievals refer to a term as *absolute* just in case it is not relative.

As the medievals understand it, therefore, the preliminary definition characterizes relations as the items signified by relative terms -- or what we would nowadays call polyadic or many-place predicates. This characterization seems initially promising, and medieval philosophers often employ it in their writings. In the end, however, they think it must be rejected -- or at least modified. For according to them, not everything we are prepared to *call* a relation actually *is* a relation. In rejecting the first or Platonic definition, the medievals take themselves to be following Aristotle, who rejects the definition on the grounds that it allows certain substances (namely, heads and hands) to qualify as relations. As Aristotle himself says:

> If the [first] definition of things toward something was adequately assigned, then it is exceedingly difficult, or impossible, to reach the conclusion that no substance is toward something. But if, on the other hand, it was not adequately assigned, and things toward something are rather [defined as] *those things for which this is their very being: to be toward another in a certain way*, then perhaps something may be said about the problem [of heads and hands].[10]

In rejecting the first definition of relations, medieval philosophers take Aristotle to be calling our attention to the fact that some relative terms do not signify relations. Terms such as “head” and “hand” clearly count as relative, since statements of the form “*x* is a head” and “*x* is a hand” are more perspicuously represented as of the form “*x* is a head of *y*” and “*x* is a hand of *y*”. Nonetheless, what these terms signify, the medievals say, are not relations but parts of substances. On the basis of these and other such examples, therefore, they conclude that relations must be identified not with the items signified by relative terms, but with a proper subset of them.

Now at this point we are still left with the question of how to identify the members of the relevant subset. Here medieval philosophers think that Aristotle's second definition -- which I have italicized in the passage quoted above -- can be of some help. The purpose of this definition, as it is commonly interpreted, is to provide a characterization of relations in terms of their metaphysical function or role. On this interpretation, relations are the items that actually serve to relate two (or more) things -- that is, the items *in virtue of which* things are related. Read in this way, Aristotle's definition provides a clear explanation for why heads and hands fail to qualify as relations. The items signified by the terms “head” and “hand” do not relate anything. On the contrary, they are things standing in relations.

The distinction between what actually serves to relate (or has the disposition to relate) and what merely stands in some relation is often used by medieval philosophers to explain the difference between the two definitions in *Categories* 7. This understanding of the definitions, moreover, gives rise to the common medieval distinction between things that qualify as relations merely according to speech (*relationes secundum dici*) and things that qualify as relations in accordance with their nature or being (*relationes*

secundum esse), as the following passage from Aquinas's *Summa Theologiae* nicely illustrates:

> Some relative terms -- such as “master” and “slave”, “father” and “son” -- are imposed to signify relative dispositions themselves (*ipsas habitudines relativas*); these terms express things relative *secundum esse*. But other relative terms -- such as “mover” and “moved”, “head” and “headed”, and terms of this sort -- are imposed to signify things on which certain relations follow; these terms express things relative *secundum dici*.[11]

The intuition behind this common distinction is that every relative term must somehow be associated with a relation, otherwise there would be no basis for the comparison involved in its predication. Nonetheless, since only certain relative terms such as “father” and “son” actually signify relations, others such as “head” and “hand” must signify non-relational things *via* their relations (such as being a part or being a whole).

Although medieval philosophers typically agree that there is a distinction to be drawn between relations *secundum dici* and relations *secundum esse*, and that heads and hands and other body parts should be included in the former class, there is often disagreement about which class certain other examples (most notably, knowledge and perception) should be included in. As the medievals recognize, terms such as “knowledge” and “perception” are relative (for statements of the form “*x* is knowledge” or “*x* is a perception” are more perspicuously represented by statements of the form “*x* is knowledge of *y*” and “*x* is a perception of *y*”). They disagree among themselves, however, about whether the items signified by these terms are relations *secundum dici* (i.e., things on which certain relations follow) or relations *secundum esse* (i.e., things which actually serve to relate two or more things). Near the end of *Categories* 7, Aristotle suggests an epistemic criterion for deciding in particular cases whether something qualifies as a genuine relation. But because this criterion is extremely difficult to understand, it does not provide any clear resolution to the original disagreement. Indeed, debate about how to distinguish between relations *secundum esse* and *dici* continues throughout the Middle Ages.

Aristotle ends *Categories* 7 almost as abruptly as he began it. Having indicated his own position, he leaves us with the following cautionary remarks:

> It is perhaps hard to make firm statements on such questions without having examined them many times. Still, to have gone through the various difficulties is not unprofitable.[12]

Medieval philosophers take Aristotle's reticence here as an invitation to refine and improve on his own account of relations and to develop it in their own ways. As Boethius, whose Latin translation is responsible for transmitting these words to the Middle Ages, says in his commentary on this passage of Aristotle:

> He would never have said this if he were not prompting us to further reflection and to even greater exercise of subtlety. Because of his exhortation, we shall not hesitate in the least to raise [further] questions and offer [our own] solutions to them in other places.[13]

3. Relations in Medieval Philosophy

Aristotle's discussion in *Categories* 7, at least as it is understood by his medieval commentators, presupposes a certain view about the nature and ontological status of relations, and hence is not primarily metaphysical in nature but definitional. Thus, Aristotle's discussion begins with the assumption that relations are a certain type of accident, and then proceeds to inquire after their proper characterization, eventually arriving at the view that all and only those accidents that are both signified by a relative term and actually serve to relate their subjects qualify as relations. Medieval philosophers, by contrast, tend to adopt Aristotle's characterization of relations, but not his assumptions about their nature and ontological status. That is to say, they begin with the view that relations are a proper subset of the items signified by relative terms -- namely, all and only those items that actually serve to relate things -- and then proceed to ask about the nature and ontological status of these items, in particular whether in every case they must be identified with an Aristotelian accident.

There are two reasons why medievals think they are entitled to adopt Aristotle's final characterization of relations without also adopting his views about the nature of the entities so characterized. First of all, Aristotle's preferred characterization of relations, as they see it, is ontologically neutral. Although it may be tempting, within a substance-accident framework, to assume that only accidents can both be signified by relative terms and relate substances, they think there is no reason in principle why a substance cannot also play these roles. Secondly, and perhaps more importantly, the medievals think there are at least some cases in which relations cannot be identified with accidents. The most common examples are, of course, theological in nature. On the Christian view of the Trinity, for example, the divine persons stand in various relations, both to one another and to things distinct from God. Nonetheless, the medievals argue, there are no accidents in God. In order to accommodate cases such as these, therefore, the medievals think that Aristotle's account must be extended to allow for relations that are not accidents.

3.1 The Rejection of Polyadic Properties

Most medieval philosophers, therefore, think they can accept Aristotle's second definition of relations without begging any substantive metaphysical questions as to their nature or ontological status. And this in turn allows for the possibility of identifying relations in different situations with entities of different ontological types. Before turning to the different types of entity with which medieval philosophers actually identify relations, it is worth noting that there is at least one type of entity with which they think relations can never be identifed, namely, the type we would nowadays refer to as *polyadic* or *many-place property*. Despite their differences, medieval philosophers appear to be unanimously agreed that no entities of this type exist in extramental reality.

Historians of philosophy sometimes speak as if medievals lacked the very concept of a polyadic property -- indeed as if this concept only became possible in the nineteenth and twentieth centuries with the advent of a formal logic of relations and multiple quantification.[14] But this is surely mistaken. What advances in formal logic have made possible is not the concept of a polyadic property, but merely its representation

within a formal system. And in any case, we have already seen that the notion of a polyadic predicate, from which our own concept of a polyadic property is derived, has a direct analogue in the medieval notion of a relative term.

There is, moreover, explicit textual evidence to support the claim that the medievals conceived of relations as polyadic properties. Medieval philosophers often compare relations to a road (*via*) that runs between two cities, or to a palisade running between two watchtowers (*inter-vallum*).[15] In drawing these sorts of analogies, medieval philosophers take themselves merely to be developing a suggestion of Aristotle's in the *Physics*.[16] By the time of Peter Aureoli (d. 1322), however, reference to Aristotle (or indeed any authority) for conceiving of relations in this way comes to seem almost otiose. As Aureoli himself says at one point:

> In the third book of his commentary on the *Physics*, comment twenty, the Commentator [= Averroes] says that a relation is a disposition (*dispositio*) existing between two things. But even apart from him it is clear that fatherhood is conceived of as if it were a kind of medium connecting a father with his son. And the same is true of other relations.[17]

This passage makes it clear that the standard way of conceiving of relations is as polyadic entities.

Despite the fact that medieval philosophers conceive of relations in this way, they explicitly deny that anything in extramental reality exactly corresponds to this conception. Returning to the passage from Aureoli, we can see that, immediately after pointing out that relations are conceived of as a sort of 'interval' (*intervallum*), he goes on to deny the existence of any such intervals in extramental reality:

> It appears that a single thing that must be imagined as some sort of interval (*intervallum*) existing between two things cannot exist in extramental reality, but only in the intellect. [This appears to be the case] not only because nature does not produce such intervals, but also because a medium or interval of this sort does not appear to be in either of the two things [it relates] as in a subject, but rather between them where it is clear that there is nothing which can serve as its subject.[18]

As this passage from Aureoli makes clear, the root of the medievals' objection to polyadic properties is ontological in nature, and has to do with the relationship of subjects to their attributes. From Aristotle, the medievals inherit a division of real beings into substances and accidents, and they typically take this division to be both exclusive and exhaustive (since it is given in terms of a contradictory pair of properties, namely, being in a subject vs. not being in a subject).[19] Like Aristotle, moreover, the medievals think of individual substances as the primary subjects of predication and divide their attributes into two main kinds, essential and accidental. Now in the case of individual substances, it is perhaps clear that they are not polyadic in nature. But medievals think the same is also true of their attributes. For attributes, as they see it, are to be conceived of as constituents, or in a broad sense as 'parts' of substances, and hence as incapable of belonging to more than one substance at a time. In the case of essential attributes, they think this is so obvious as to barely merit mentioning in the context of relations.[20] But

they are fond of pointing this out in the case of accidents. To take just one particularly striking example, consider how Aquinas responds in his *Sentences* commentary to the question whether it is possible for "numerically one and the same relation" to belong to belong to two subjects at a time: "This cannot be, for one accident cannot belong to two subjects".[21]

Although the medieval objection to polyadic properties stems in large part from ontological considerations, medieval philosophers also offer other reasons for rejecting them. Sometimes the reasons are grounded in an appeal to theoretical parsimony: all other things being equal, the ontology that postulates the fewest types of entity is to be preferred. Thus, as Ockham is so fond of saying: "It is futile to do with more [entities] what can be done with fewer".[22] Again, the reasons are sometimes grounded in phenomenological or epistemological considerations: we are not presented in experience with anything but individuals and their so-called absolute attributes -- that is, substances and their quantities and qualities. Albert the Great, for example, mentions this sort of objection and traces it to certain Muslim philosophers: "Some fairly recent philosophers such as Avicenna and Alfarabi . . . say that no form that is a being (*ens*) belongs to a thing unless it is absolute as far the being (*esse*) it has in itself is concerned -- as is clear from looking at cases of what is hot, cold, white, black, and *all other things*."[23]

3.2 Anti-Realism about Relations

The rejection of polyadic properties might seem to commit one to a form of anti-realism about relations. This was, in fact, the position of Peter Aureoli. If there are no polyadic properties or 'intervals', he argues, then there is nothing in extramental reality to correspond to our relational concepts. Reasoning from the common medieval assumption that relations are items corresponding to our relational concepts, he concludes that "a relation cannot be posited except in apprehension alone".[24] Thus, relations do exist, on his view, but only as the contents of certain concepts or apprehensions -- or as what he elsewhere describes as beings of reason (*entia rationis*). Strictly speaking, therefore, Aureoli does not want to deny that things can be related, but only that they can be related apart from the activity of the mind.

Given the current state of medieval research, it is difficult to know how widespread this (or any other) form of anti-realism was during the Middle Ages. Shortly after Aureoli's death, certain aspects of his conceptualism appear to have met with considerable success. Ockham and his followers, for example, accept Aureoli's view that relations are concepts or beings of reason, though unlike Aureoli, as we shall see, they think this is compatible with saying that things are related independently of the activity of the mind. And one can find echoes of the same sorts of view throughout the writings of early modern philosophers.[25] Prior to the fourteenth century, however, it is difficult to identify any unambiguous representatives of anti-realism, at least in the Latin west. Albert the Great claims to find something akin to Aureoli's argument in the writings of Alfarabi (d. 950) and Avicenna (d. 1037), but he does not say whether either of these figures endorses the argument or merely presents it for consideration.[26] Again, Aquinas suggests that a similar view can be found in the writings of Gilbert of Poitiers and his followers, the so-called Porretani, but the suggestion is controversial, and in any case Aquinas reports that Gilbert later retracted the view in the face of theological controversy.[27] Unless further research alters the current picture, therefore, it would appear that, although anti-realism may have had something of a foothold in the

Arabic speaking world, most notably among the members of a group of orthodox Muslim theologians known as the Mutakallim•n, in its more radical forms it was always a minority position in the Latin west.[28]

3.3 Realism without Polyadic Properties

Thus, if we take anti-realism to be the view that (a) nothing in extramental reality corresponds to our relational concepts and (b) nothing is related independently of the activity of the mind, it appears that most medieval philosophers would reject it. Nor is it hard to see why. Like most of us, the medievals recognize the implausibility of saying that facts like Simmias's being taller than Socrates are somehow dependent on the activity of the mind.[29] Indeed, they often argue that such an anti-realist view about relations cannot account for even such basic facts as the real structure of the universe, the existence of real composition, causality, spatial proximity, or even the objectivity of mathematical knowledge. Thus, Ockham speaks for the majority when he asserts in his *Ordinatio*:

> The intellect does nothing to bring it about that the universe is one, or that a whole is composed [of its parts], or that causes in spatial proximity [to their effects actually] cause [their effects], or that a triangle has three [angles], etc. . . .any more than [the intellect] brings it about that Socrates is white or that fire is hot or water cold.[30]

Again, realism about relations seems to be implied by a standard medieval interpretation of the Aristotelian categories. "For nothing is placed in a category" says Aquinas "unless it is something existing outside the soul".[31] Thus, if realism is false, there should not even be a category of relation

In addition to philosophical objections to anti-realism, medieval philosophers also had theological grounds for opposing it. Christian doctrines such as the Trinity were often thought to require realism about certain theological relations. Indeed, the existence of real relations in God is affirmed at the Council of Rheims in 1148, and shortly thereafter it becomes customary to say that the denial of this entails a form of heresy known as Sabellianism.[32]

But if there is general consensus among medieval philosophers that anti-realism ought to be rejected in favor of what we might call *realism without polyadic properties*, what is to be said about arguments such Aureoli's, which in effect deny the possibility of relations without polyadic properties? Here I think it is useful to consider the reply given by Albert the Great. Although he writes a generation before Aureoli, he explicitly addresses this sort of argument, and the reply he offers appears to be perfectly standard. According to Albert, the problem with arguments of this type is that they rely on the assumption that *if there are no polyadic forms or properties, then there is nothing in extramental reality corresponding to our relational concepts*. This assumption would be true, he suggests, only if our conceptual framework displayed an exact isomorphism to the structure of the world. But as Albert sees it, there is no reason in principle why a polyadic concept should not have something non-polyadic in extramental reality corresponding to it.[33] In making this sort of reply, Albert aligns himself with a medieval tradition of rejecting the view that *for every distinct type of concept there is a distinct type of entity*. This view is often

associated in medieval philosophy with Plato and his followers, and by the time of Ockham it is regarded, perhaps somewhat hyperbolically, as the source or root (*radix*) of the greatest errors in philosophy. As Ockham himself says at one point: "to multiply beings according to the multiplicity of terms . . . is erroneous and leads far away from the truth".[34]

For most medieval philosophers, therefore, the question is not *whether* there are any things in extramental reality corresponding to our relational concepts, but *what* these things are in themselves. But having ruled out the possibility of their being polyadic in nature, the medievals are not left with many options. Because they accept a form of realism, they are committed to saying that, at least in many cases, there is something in extramental reality grounding these concepts. But because they also accept a form of realism without polyadic properties, and are committed to a broadly Aristotelian ontology, they have no choice but to identify these extramental grounds with either individual substances or their monadic properties (whether they be accidental or essential).

Throughout the medieval period philosophers are attracted to the view that Aristotle suggests in the *Categories*, namely, that relations (or the items grounding our relational concepts) are accidents. For reasons I have already mentioned, no medieval can allow this account to hold for all relations, since there are cases, especially theological ones, in which involve relations but no accidents. Nonetheless, until the fourteenth century, this account is taken to apply so widely that there emerges a kind of paradigmatic analysis of relational situations. According to this analysis, if a judgment of the form '*aRb*' is true, what makes it true is a pair of individuals, *a* and *b*, and a pair of accidents, *F* and *G*. Thus, if it is true that Simmias is taller than Socrates, this is to be explained by a pair of monadic properties, one of which inheres in Simmias and relates him to Socrates as taller and the other of which inheres in Socrates and relates him to Simmias as shorter. According to this analysis, moreover, it is the particular accidents that correspond to, and thus ground, relational concepts such as 'taller than' and 'shorter than'.

As we shall see in the next section, it is within the context of this paradigmatic analysis of relational situations that one of the most hotly disputed and intractable debates of all the Middle Ages arises. In order to facilitate my discussion of this debate, I shall hereafter refer to situations that conform to this analysis as *paradigmatic relational situations* and to the pairs of properties or accidents involved in them as *paradigmatic relations*.

4. Paradigmatic Relational Situations

Before turning to the details of this debate, it is worth emphasizing that most medieval philosophers, at least those living prior to the fourteenth century, regard what I have been calling the paradigmatic analysis of relational situations as the analysis provided by Aristotle's discussion in the *Categories*. Aristotle does not explicitly speak of relations there as items corresponding to relational concepts, though this way of speaking would seem to be justified on the basis of his semantic characterization. Nonetheless, he does explicitly identify relations in the *Categories* with accidents and speaks of them as if they always come in pairs. Indeed, as Boethius points out in his commentary, whenever Aristotle refers to the category of relations he refers to it in the plural ("things toward something"), thereby departing from his usual

practice of referring to categories in the singular (“substance”, “quantity”, “quality”, etc.). According to Boethius, Aristotle speaks of the category in this way because its members are unique among categorial beings in that they cannot be understood to exist by themselves:

> Things toward something [i.e., relations] cannot be grasped by the intellect by themselves or individually, so that we could say that things toward something exist individually. Whatever is known regarding the nature of a relation must be considered together with something else. For example, when I speak of a master, this by itself means nothing if there is no slave. The naming of one relative immediately brings with it another thing toward something.[35]

Following Boethius, medieval commentators often take Aristotle's use of the plural to indicate that the category of relations is comprised by pairs of correlatives or converse relational accidents.[36] Simmias cannot be taller than Socrates unless Socrates is shorter than Simmias. Thus, if Simmias is taller than Socrates, it appears that a pair of accidents is needed to explain the situation, one of which is identified with Simmias's tallness (and hence corresponds to the concept ‘taller than’) and the other of which is identified with Socrates's shortness (and hence corresponds to the concept ‘shorter than’).

As the medievals recognize, there are a number of questions that arise about the precise nature and ontological status of these paradigmatic relations. What sort of monadic properties or accidents are they, and how are we to think of them? As the medievals see it, there are two main positions one can take. For convenience, I shall label them, respectively, *reductive* and *non-reductive realism (without polyadic properties*).

According to reductive realism, which is the simplest or most ontologically parsimonious form of realism without polyadic properties, paradigmatic relations are to be identified with ordinary, non-relational accidents -- that is to say, with accidents that fall under Aristotelian categories other than relation. Thus, if Simmias is taller than Socrates, the reductive realist will explain this by appealing to their respective heights, which fall under the category of quantity. Again, if Simmias is similar to Socrates (i.e., resembles him in color, say), the reductive realist will explain this by appealing to Simmias's and Socrates's particular colors, which fall under the category of quality. And so on for other paradigmatic relations.

According to non-reductive realism, by contrast, these relations are to be identified not with ordinary, non-relational accidents, but rather with accidents of a *sui generis* type. Thus, if Simmias is taller than Socrates, this is to be explained by appealing to a pair of *sui generis* accidents that are distinct from, but nonetheless necessitated by, Simmias's and Socrates's heights. Again, the accidents relating Simmias and Socrates as similar are not their colors, according to the non-reductive realist, but a pair of *sui generis* accidents necessitated by them. And so on for the relations involved in other paradigmatic relational situations.

In the medieval discussion of relations, it is this difference -- the difference between reductive and non-reductive realism -- that constitutes the greatest divide among philosophers, and representatives of both

positions can be found throughout the medieval period. Peter Abelard (d. 1142) and Albert the Great (d. 1280) are, perhaps, the best representatives of reductive and non-reductive realism (respectively) in the early and high Middle Ages, whereas William Ockham (d. 1347) and John Duns Scotus (d. 1308) are among the best known examples of these positions in the later Middle Ages. Again, some philosophers appear to have held different positions at different stages of their career. Henry Harclay (d. 1317), for example, began his career as a staunch defender of Scotus's non-reductive realism, but by the end of it gravitated towards a position that is much closer to, and in many ways anticipates, Ockham's specific form of reductive realism.[37]

Beginning in the mid-thirteenth century, the debate between reductive and non-reductive realists is often carried out in the context of the discussion whether relations are identical to their foundations. (This is the context, for example, in which Scotus and Ockham develop their positions.) By the mid-thirteenth century, moreover, it is customary to admit at least a conceptual distinction between relations (such as Simmias's tallness and Socrates's shortness) and the non-relational accidents in virtue of which these relations hold (such as Simmias's and Socrates's heights). The relevant non-relational accidents, in turn, are referred to as foundations or grounds (*fundamenta*) of relations, since their possession is thought to necessitate the holding of relations. In this context, therefore, the important question is whether the distinction between relations and their foundations is *merely* conceptual.

We might expect that reductive realists would always answer this question in the affirmative, whereas non-reductive realists would always answer it in the negative. It turns out, however, this is not the case. For reasons to be explained below, there are some reductive realists who, in spite of rejecting that relations comprise a *sui generis* type of monadic property, nonetheless maintain that relations are in an important sense distinct from their foundations. In light of this complication, I shall continue to cast the medieval debate about the nature paradigmatic relations in terms of reductive and non-reductive realism rather than in terms of the identity or distinctness of relations and their foundations.

4.1 Reductive Realism without Polyadic Properties

As I mentioned above, on the simplest or most ontologically parsimonious form of realism without polyadic properties, what I am calling *reductive realism*, paradigmatic relations are identified with ordinary, non-relational monadic properties or accidents. Now we might expect this form of realism to appeal to anyone committed to realism without polyadic properties. After all, failure to reduce such relations to such accidents threatens to make them mysterious. If paradigmatic relations are monadic properties, but not ordinary, non-relational accidents, then how are we to understand them? On this count, however, the position of the reductive realist is perfectly intelligible. According to it, if Simmias is taller than Socrates, the holding of the relation will be explained by the possession of a pair of ordinary heights -- say, being-six-feet-tall in the case of Simmias and being-five-feet-ten-inches-tall in the case of Socrates.

These sorts of considerations certainly play a role in the debate between medieval reductive and non-reductive realists. As we shall see, non-reductive realists typically recognize that the chief difficulty for their view lies in providing a principled motivation for, and then explaining the precise nature of, the *sui*

generis monadic properties that they identify with relations. Reductive realists, by contrast, often rely on considerations of theoretical parsimony to motivate their own views. This is certainly true in the case of Abelard and Ockham, whom I earlier identified as reductive realists. In the case of Abelard, these considerations are not explicitly formulated, though they are part-and-parcel of his general approach to metaphysics and philosophy of language.[38] In the case of Ockham, the appeal to parsimony is much more explicit, often taking the form of what has come to be known as Ockham's razor: “Plurality should not be assumed without necessity”.[39]

Another consideration that plays an important role in the debate between reductive and non-reductive realists has to do with the proper interpretation of authoritative texts. Since medieval philosophers develop their theories of relations in the course of reflecting on Aristotle's *Categories*, they often present their views as the one actually suggested by the text itself. This is especially true in the early Middle Ages. Abelard, for example, suggests that his reductive theory is a direct consequence of Aristotle's second definition of relations. Thus, when Aristotle characterizes relations in *Categories* 7 as “those things for which this is their very being: being toward something in a certain way”, Abelard argues that this should be interpreted to mean that relations are items that *make other things to be relative or related* -- that is to say, that they are what we might call relative-making characteristics.[40] Textual considerations and direct appeals to Aristotle's authority play a much bigger role in early medieval debates than they do in the high and later Middle Ages, when literal commentary was no longer the dominant form of philosophical literature. Nonetheless, these sorts of considerations continue to be important throughout the Middle Ages. Even Ockham, in his *Summa logicae*, addresses the issue of whether Aristotle should be interpreted as holding some form of non-reductive realism. “There are many theologians of this opinion,” he says “and at one time even I believed it to have been Aristotle's opinion. Now, however, it seems to me that the contrary opinion follows from his principles.”[41]

One final argument used by reductive realists to motivate their position is worth mentioning here. This argument has to do with the nature of relational change. Almost all medievals accept the intuitively plausible view that things can acquire (and lose) relations without undergoing any real (as opposed to what is nowadays called a merely Cambridge) change. Thus, when Simmias comes to be taller than Socrates, Socrates acquires the relation of being shorter than Simmias. But it would seem that Socrates can acquire this relation without undergoing any real change in himself (say, if he acquires it solely in virtue of a change in Simmias's height).[42] As reductive realists such as Ockham often point out, this fact about relational change admits of a ready explanation on their position.[43] For according to the reductive realist, Socrates's relation of being shorter is nothing over and above Simmias's and Socrates's heights, the latter of which remains unaltered. The same fact, however, appears to pose a serious challenge for the non-reductive realist. For if Socrates's relation is something distinct from Simmias's and Socrates's heights, then apparently Socrates acquires something when Simmias's height increases. But then, contrary to the original intuition, the non-reductive realist must say that Socrates undergoes a real change after all.

These considerations are not the only ones that reductive realists appeal to, but they are among the most important for understanding their position, and I shall rely on them in what follows to motivate and clarify the alternative position of the non-reductive realists.[44]

Before leaving the topic of reductive realism, however, I need to take account of a slight complication. In my discussion to this point, I have allowed myself to speak as if reductive realists intend to identify individual paradigmatic relations such as tallness or shortness with pairs of monadic properties *taken jointly*. I have done this because, at least initially, this seems to be the most natural way to make sense of their position. After all, it is not Simmias's height taken by itself that necessitates his being taller than anyone. On the contrary, it is only when his height is taken together with that of another, say Socrates's, that the relation comes to hold of him. But even if this seems initially to be the most natural way to construe reductive realism, it is not the way medieval philosophers typically understand it.[45] As I indicated earlier, medieval philosophers typically identify relations such as tallness or shortness, not with pairs of monadic properties taken jointly, but with the individual members of such pairs. Thus, even if a pair of monadic properties is required to explain a relational situation such as Simmias's being taller than Socrates, medieval philosophers typically want to say that only one of these properties is to be identified with Simmias's tallness (and hence to correspond to the concept 'taller than'), whereas the other is to be identified with Socrates's shortness (and hence to correspond to the concept 'shorter than'). But how is this possible, especially if the properties in question are supposed to be just ordinary heights?

To begin, let us note that medievals do recognize the worry associated with identifying relations with ordinary monadic properties taken individually. In fact, this recognition sometimes leads them to consider developing reductive realism in just the way suggested above. Thus, in a particularly striking passage from one of Henry Harclay's later works it is suggested that, although the similarity between two white things cannot plausibly be identified with the whiteness of either one, it might nonetheless be identified with the pair of whitenesses taken jointly. Indeed, he even draws on common medieval views about number to make the suggestion plausible:

> A relation is clearly a reality distinct from one foundation, but not a reality distinct from both foundations. For when one white thing [such as Simmias] is posited and then another white thing [such as Socrates] is posited, there is a relation beginning at that point. Thus, the similarity is not a [single] whiteness, but two whitenesses existing at once. Indeed, the two whitenesses can be sufficient to constitute a single species in the genus of relation. Just as two discrete unities of whatever magnitude are sufficient for producing a single species of number, so too the same thing holds for the case at hand. Again, just as plurality or multitude is not a reality different from the constituents of the plurality or multitude, neither is the similarity [of the white things] different [from their whitenesses].[46]

As this passage clearly indicates, medieval philosophers do consider the possibility of identifying relations with pairs of monadic properties or accidents. Having considered this sort of identification, however, they do not typically endorse it. Thus Harclay, who certainly seems to be attracted to a form of reductive realism, proceeds immediately to reject such an identification on the grounds that it would lead to certain absurdities. Citing Avicenna as his authority, he says:

> Avicenna argues against this on the basis of relations involving inferiority and superiority, for it is clearer in those cases than it is for other relations, such as those involving equals.

> For fatherhood is in the father alone and not in the son . . . and the same thing holds of sonship. Therefore one must hold that there are two relations [here]. And in this case it is clear that the relation is not the same reality as [the two] foundations because the relation is not in everything in which the foundation of the *relatum* is. This is the case, too, for relations involving equals, even if it is not as clear.[47]

In this passage, Harclay calls our attention to one of the consequences of identifying relations such as fatherhood and sonship with pairs of properties or accidents. If a relation such as fatherhood just is a pair of accidents, then it would follow that fatherhood is partly in the father and partly in the son (since the pair is such that one of its members inheres in the father, the other in the son). But this, says Harclay, is absurd. Fatherhood is an asymmetrical relation (or as Harclay prefers to put it, a relation involving superiority and inferiority) -- that is to say, if an individual *a* has fatherhood in respect of another individual, *b*, *b* cannot have fatherhood in respect of *a*. Harclay takes this to show that “fatherhood is in the father alone and not in the son”. And the same is true, he thinks, of its converse, sonship, as well as for all other relations, including symmetrical relations or “relations involving equals, even if it is not as clear”.

In this passage, Harclay makes explicit views that appear to be taken for granted by medieval philosophers generally. But if relations are not to be identified with pairs of accidents, how can reductive realists think of them in terms of ordinary, non-relational accidents? After all, the only other option seems to be identifying them with ordinary accidents taken individually -- which as we have seen is problematic. For it is not an ordinary accident such as Simmias's height that necessitates Simmias's being taller, but only this height taken together with that of another. Indeed, it was precisely this point that led us to conclude that “relative-making” is a description best reserved for pairs of ordinary properties taken jointly.

At this point, it seems to me that the reductive realist should respond by saying that, although Simmias's height is not by itself relative-making, it is nonetheless *potentially* relative-making. For if Simmias's height is by itself potentially relative-making, then apparently we are entitled to say that in the presence of another height, such as Socrates's, it becomes *actually* relative-making -- that is to say, it actually comes to relate Simmias to Socrates. This suggests a third option for reductive realists: to maintain that relations are to be identified with ordinary, non-relational accidents *in certain circumstances*.

In fact, it is this last sort of view that I think most reductive realists during the medieval period actually hold. Thus, as I interpret Harclay, this is precisely the view he goes on to develop in the discussion we have been following.[48] Simmias's being taller than Socrates is to be identified, on his view, not with Simmias's height *tout court*, but with Simmias's height in certain circumstances -- including the circumstance that Socrates is five-feet-ten.[49] (One is put in mind here of attempts by certain contemporary philosophers to identify dispositions, such as brittleness or solubility, with their categorical basis in certain circumstances -- namely, the circumstance that the appropriate laws of nature obtain.[50]) To put Harclay's view in a slightly different way, one which will help to bring out its semantic consequences, we might say that in a world in which Simmias exists by himself a judgment of the form ‘Simmias is taller than Socrates’ will be false, and hence nothing will correspond to the relational concept

'taller than'. But in a world in which Simmias is six-feet-tall while Socrates is five-feet-ten, a judgment of the same form will be true, and in this world there will be something corresponding to 'taller', namely, Simmias's height. In such a world, medieval philosophers would say that 'taller' primarily signifies Simmias's height, but indirectly signifies or connotes the height of Socrates.[51]

Now as it turns out, it is not only the reductive realists, but the non-reductive realists as well that speak of relations as accidents of single subjects -- that is, of Simmias's tallness as something belonging only to him, of Socrates's shortness as something belonging only to him, and so on for all other particular relations. In light of what has just been said, therefore, we can clarify our understanding of the paradigmatic analysis of relational situations generally. As indicated earlier, this analysis requires that when a judgment of the form '*aRb*' is true, what makes it true is a pair of individuals, *a* and *b*, and a pair of monadic properties or accidents, *F* and *G*. Reductive realists, as we have just seen, identify *F* and *G* with ordinary categorial accidents, whereas non-reductive realists, as we shall see in the next section, identify *F* and *G* with accidents of a *sui generis* type. But however they construe the nature of these accidents, they both deny that it is the pair of accidents -- that is, *F* and *G* taken jointly -- that corresponds to our relational concepts. On the contrary, they say, *F* directly corresponds to one of our relational concepts, whereas *G* directly corresponds to its converse. Thus, when Simmias is taller than Socrates, reductivists and non-reductivists are agreed that it is an accident of Simmias that directly corresponds to (or as they would prefer to say, is primarily signified by) the concept 'taller than', and an accident of Socrates that is primarily signified by the concept 'shorter than'. To the extent they disagree, therefore, their disagreement concerns only whether the accidents primarily signified by relational concepts can also be signified by ordinary non-relational concepts (and if so, precisely how and under what circumstances).

With these clarifications in mind, I now turn to considerations that lead some medieval philosophers to reject reductive realism in favor of a form of non-reductivism.

4.2 Non-Reductive Realism without Polyadic Properties

Of all the considerations favoring non-reductive realism, perhaps none is more compelling than the intuition that relations have a different nature or 'quiddity' from that of ordinary categorial accidents. As we have seen, medieval philosophers recognize that predications involving relative terms (such as "*x* is taller") are incomplete in a way that monadic or absolute predications (such as "*x* is white") are not. They also recognize, moreover, that unlike absolute terms relative terms come in pairs or sets. Every relative term has a correlative, and the meaning of the one (say, "taller") cannot be understood in isolation from the meaning of the other ("shorter"). These facts about relative terms and predication are often thought to show that the items signified by them have a distinct nature or quiddity from that of their foundations -- the ordinary, absolute accidents. Here again it is useful to quote Harclay:

> Avicenna (*Metaphysics* III, the chapter on relation) says that a relation has its own quiddity distinct from the quiddity of its foundation. Therefore it is a distinct reality. Moreover, in the *Categories* Aristotle says that the being associated with relatives is being-toward-something-else (*ad aliud se habere*). But it is not the case that the being of a foundation is

> being-toward-something-else. Therefore they are not the same."[52]

A reductive realist can, of course, explain the uniqueness of relative terms and predications without introducing relations over and above their foundations. For such a realist can always assert that relative terms are associated with concepts whose content is distinct from that of any non-relational or absolute concepts, and insist that this is all that is required to explain the peculiar nature of relative terms and predications. Even so, there is a question that remains for the reductive realist. Why, it must be asked, if Simmias's being taller than Socrates is nothing ontologically over and above Simmias's and Socrates's heights, do we represent this relation *as if* it were? At this point, it would seem that the reductive realist can only appeal to our psychological make-up. We simply do (or at lest can) represent one and the same situation in two very different ways. As Ockham says in one of his *Quodlibetal* questions, using a slightly different example:

> Socrates is similar to Plato by the very fact that Socrates is white and Plato is white . . . Yet, despite this, the intellect can express these many absolute things by means of concepts in diverse ways: in one way, by means of an absolute concept, as when one says simply 'Socrates is white' or 'Plato is white'; in a second way, by means of a relative concept, as when one says 'Socrates is similar to Plato with respect to whiteness'.[53]

Even if one does not find this sort of appeal to psychology implausible, the non-reductive realist would seem to have a more satisfying reply to the original question -- namely, "Why do we represent Simmias's being taller than Socrates as distinct from Simmias's being six-feet-tall while Socrates is five-feet-ten?" For according to the non-reductive realist, we represent these two situations as distinct because they *are* distinct. Indeed, the non-reductive realist can say that the logical incompleteness of predicates such as "taller" (and its correlative, "shorter") calls our attention to precisely what makes these situations distinct -- namely, the *sui generis* accidents that are possessed by Simmias and Socrates in addition to their respective heights.

Now as most non-reductive realists recognize, there is a difficulty posed by their position -- namely, that of giving a perspicuous account of the nature of paradigmatic relations or relational accidents. They often attribute this difficulty, however, to the fact that the nature in question is *sui generis*. As Albert the Great says, when he turns to the discussion of relations in his commentary on Aristotle's *Metaphysics*: "It is difficult for us to speak about [the category of] *toward something* or relation because it has a nature and being altogether different from the genera of being which have been considered so far [namely, substance, quantity, and quality]."[54] Given that non-reductivists construe the nature of paradigmatic relations as *sui generis*, it is not surprising that they feel the need to resort to metaphors to describe it. Albert himself appeals most often to a visual metaphor of outward-looking-ness (*respectus*), and describes individual relations as that in virtue of which a subject 'looks out toward' (*respicit ad*) another. Other philosophers, however, rely on other metaphors and variously describe the nature of relations as 'directionality' or 'toward-ness' (*aditas*), as a certain kind of disposition or way of holding oneself (*habitudo* or *relativa habitudo*), or again as the principle of 'structure' or 'order' (*ordinatio*).[55]

Of all these metaphors, the ones involving directionality -- or intentionality -- are likely to be the most helpful to us. For there are some contemporary philosophers who characterize intentionality, not in terms of a polyadic or many-place property, but as a *sui generis* type of monadic property.[56] According to these philosophers, intentionality is a property whose intrinsic nature is such that, when it is exemplified by a subject in appropriate circumstances, which include the presence of an appropriate object, it relates its subject to the object in question -- in particular, it relates the subject to it as thinker to object thought. This analogy is useful, I think, because non-reductive realists typically regard intentionality as a special case (or type) of relation. According to non-reductivists, all paradigmatic relations (or at least all those that qualify as *sui generis* accidents) are properties whose intrinsic nature is such that their exemplification in the appropriate circumstances will relate their subjects to something else. It is just that in certain cases, these properties relate their subjects specifically as thinkers to objects thought.

In the end, therefore, it would appear that the non-reductive realists do have something to say in response to the reductivists' worries about the mysterious nature of relations and their appeals to considerations of theoretical parsimony. But what about the phenomenon of relational change? As we have seen, the reductive realists have a natural explanation of it. But is there anything that the non-reductive realists can say about it?

As I said earlier, medieval philosophers share the intuition that the acquisition (or loss) of relations by a subject can occur without any real change taking place on the part of this subject -- that is, solely in virtue of a so-called Cambridge change. To explain this intuition, some non-reductivists think it is enough to say that a relation can be acquired (or lost) without its subject undergoing any real change *with respect to its absolute accidents*. Thus, the example of Socrates's coming to be shorter than Simmias is sometimes glossed by saying that, with respect to absolute accidents, it is true that Socrates comes to be shorter than Simmias solely in virtue of a real change in Simmias. Nonetheless, if we consider Socrates's relative accidents as well, such non-reductivists will say that Socrates has, in fact, undergone a real 'relative' change -- that is, a real change with respect to one of his relations. [57]

But non-reductivists need not take this line. Some, more sympathetic to the reductive realist position on relational change, take cases like that involving Simmias and Socrates as an opportunity to further clarify the nature of relations. Albert the Great, for example, argues that relational accidents are so closely tied to their foundations that they are acquired along with them.[58] Thus, if Socrates comes into existence in a world all by himself at a height of five-feet-ten-inches-tall, he will not only possess a certain quantity (namely, his height), but in virtue of possessing it, he will also possess a *sui generis* relational accident. Now, in such a world, the relational accident will not actually relate Socrates to anyone or anything. In virtue of possessing it, however, he will be *potentially* related. Suppose, however, that Simmias now comes into existence at six-feet-tall. Socrates's relational accident will come actually to relate him to Simmias as shorter. In doing so, however, Albert insists that Socrates will not undergo any real change whatsoever, absolute or otherwise. Not surprisingly this sort of view leads Albert to distinguish sharply between relational accidents and relations (which are just relational accidents in certain circumstances), and to suggests that strictly speaking we should think of the relevant category as comprised, not of relations, but of relational accidents. This, too, helps to explain why in referring to the members of this category Albert prefers the concrete terms "toward something" (*ad aliquid*) and "relative" (*relativum*) to

the abstract term “relation” (*relatio*) -- for relational accidents are ‘toward something’ or outwardly directed even if they are not actually relating their subject to anything.[59]

Like Albert, Aquinas too suggests that there is a close connection between relational accidents and their foundations. Following Albert's terminology, he suggests that in virtue of possessing a specific height or quantity, an individual such as Socrates is *potentially* equal to all those who have the same height, and *potentially* unequal to -- that is, potentially shorter or taller than -- all those with a different height.[60] Like Albert, moreover, Aquinas says that such an individual can come to be *actually* equal (or unequal) to another solely in virtue of a change in that other, and concludes from this that relations must be in their foundations as in a root (*in radice*). [61] On some interpretations, this is to be explained by saying that although relations and their foundations differ formally -- that is, involve different accidental forms or properties -- nonetheless their act of being (*esse*) is identical. Thus, when Socrates comes to be shorter than Simmias as a result of Simmias's growth, Socrates does not undergo a change properly speaking, since he does not acquire any new act of being. Rather, the ‘old’ act of being of his height (sometimes referred to as the ‘*esse-in*’ of the relation) merely acquires a new determination (sometimes referred to as the ‘*esse-ad*’ of the relation), in this case to Simmias.[62]

We have seen enough, I think, to appreciate the main ontological differences that divide reductive and non-reductive realists, and even some of the differences that divide non-reductivists among themselves. As I indicated earlier, in the later Middle Ages these differences tend to emerge in the context of the debate over whether relations are identical with their foundations. As I also indicated, however, we have to be careful not to assume that how a given philosopher answers this question is an infallible guide to his position. For although reductivists typically affirm that relations are identical to their foundations, and non-reductivists typically deny it, there are some reductivists who side, at least verbally, with the non-reductivists on this question. Thus, the later Harclay, as I interpret him, is a reductive realist -- for he not only rejects the existence of *sui generis* relational properties, but also maintains that in paradigmatic relational situations, substances are related by their ordinary, non-relational accidents. Nonetheless, Harclay denies that relations are identical with their foundations, since he thinks this would amount to saying that relations can be straightforwardly identified either with ordinary accidents taken by themselves or with pairs of such accidents taken jointly. Again, as I shall point out in the next section, a number of later medieval reductivists, including Ockham, eventually come to reject the traditional Aristotelian characterization of relations as items that relate substances in favor of the view that relations are items existing only in the mind (as concepts). As we shall see, this does not mean that Ockham and others deny that there are extramental grounds for our relational concepts or even that things can be related by their foundations independently of the mind. On the contrary, it means only that, unlike their predecessors, they refuse to call anything that grounds a relational concept a relation. The fact that these reductivists regard relations as concepts, however, explains why they too are willing to deny the identity of relations and their foundations: for the relevant foundations, according to these philosophers, are ordinary, extramental accidents, and obviously no concept (or act of the mind) could be identical with them.

Finally, it must be noted that even the notions of identity and real distinction come to be the subject of controversy during the high and later Middle Ages, and this too has the result of complicating the debate

over whether relations are identical to their foundations.[63] For all these reasons, therefore, we must be careful not to identify too closely the debate between reductive and non-reductive realists with the debate over whether relations are identical to their foundations.

5. Non-Paradigmatic Relational Situations

So far we have been focusing only on the medieval discussion of paradigmatic relational situations -- that is, relational situations conforming to the analysis suggested in Aristotle's *Categories*. According to this analysis, as we have seen, when a judgment of the form '*aRb*' is true, what makes it true is nothing but *a* and *b* and a pair of accidents inhering in them, *F* and *G*, which correspond, respectively, to the concepts '*R*' and, its converse, 'R^{-1}'. Although this analysis is suggested by Aristotle's discussion in the *Categories*, it is clear from some of his later works, most notably the *Metaphysics*, that Aristotle does not think all relational situations can be made to conform to it. Thus, in *Metaphysics* V, he suggests that there are some relational situations in which substances are related, not by a pair of accidents, but by a single accident belonging to just one of them. Here he cites the example of intentional relations. Thus, if it is true that Simmias is thinking about Socrates, what makes this true is nothing but Simmias, Socrates, and an accident of Simmias. For the sake of convenience, let us hereafter refer to relational situations that do not conform to the paradigm of the *Categories* as *non-paradigmatic relational situations.*

Although medieval philosophers did not have direct access to Aristotle's *Metaphysics* until the mid-twelfth century, they did feel pressure from other, largely theological sources to admit something like his non-paradigmatic relational situations. On the standard medieval conception of deity, God is an absolutely perfect being, where this is taken to imply that God not does not have any accidents, is immutable, and hence cannot undergo any real change. On this conception of deity, however, it is difficult to explain God's relational features. As Augustine points out in book V of his *De Trinitate*, the difficulty can be expressed in a particularly acute form with respect to the Christian doctrine of creation, according to which God freely created the universe.[64] This doctrine suggests that at one time God lacks, and at a later time acquires, a contingent or accidental relation -- namely, that of being creator. But then does it not also require that in creating the universe God underwent a real change? And what about the claim that there are no accidents in God?

Augustine's solution, which is taken up and developed by medieval philosophers, is to say that when God acquires a new relation, this is not to be explained in terms of any properties or accidents of him, but is rather to be explained in terms of properties or accidents of the things to which he is related. Thus, in the case of creation Augustine says:

> Even though [God's substance] begins to be [truly] spoken of [as related to a creature] at some time, still nothing is to be understood to have happened to the divine substance itself, but only to the creature in relation to whom it is spoken of.[65]

On the basis of theological considerations, therefore, Augustine is led to something like Aristotle's non-paradigmatic relational situations. For according to him what makes it true that God is related to his

creatures is nothing but God, the creatures, and a monadic property or accident of the latter.

Interestingly, Augustine does not think that the case of God is unique, but also suggests that there are non-theological cases in which things are related solely in virtue of the properties or accidents of other things. Thus, a coin, he says can increase or decrease in value solely in virtue of the intentional states of human beings.[66] Boethius, in a treatise also known as *De Trinitate*, discusses these same sorts of issues and in the course of doing so adds yet another non-theological example -- a variation of which comes to be the standard medieval example of a non-paradigmatic relational situation. Consider a man who walks up beside a stationary column. In this sort of case, medieval philosophers like to say, the column comes to be to the right of the man, but solely in virtue of a property of the man.[67]

Thus, even before medieval philosophers had direct access to Aristotle's *Metaphysics* they were led to acknowledge the existence of the type of relational situations mentioned there. Of course, once the relevant texts of Aristotle became available, the medievals worked hard to connect their discussions with these texts. Indeed, when the issue of God's relation to his creatures is taken up in the thirteenth century -- an issue that becomes one of the foci of medieval discussions of relations generally -- medieval philosophers often made direct appeals to Aristotle's discussion in *Metaphysics* V.[68]

5.1 Relations of Reason

So far, then, we have seen that medieval philosophers are committed to recognizing at least one type of non-paradigmatic relational situation -- namely, those involving a pair of substances and a single accident inhering in just one them. As we might expect, medieval philosophers disagree about the precise nature of the accidents involved in these situations. Reductive realists, of course, identify them with ordinary, non-relational accidents, whereas non-reductive realists identify them with accidents of a *sui generis* type. Thus, when Simmias is thinking of Socrates, the reductive realist will appeal to nothing more than a quality of Simmias, whereas the non-reductive realist will also postulate an accident distinct from, but necessitated by Simmias's quality.

For our purposes, however, what is most interesting about the type of relational situation in question is the complication it presents for both reductive and non-reductive realists. In paradigmatic relational situations, as we have seen, there is always a distinct property or accident corresponding to each member of the relevant pair of converse relational concepts. Thus, when Simmias is taller than Socrates, there is one accident corresponding to the concept ‘taller than’ (namely, an accident of Simmias) and one accident corresponding to the concept ‘shorter than’ (namely, an accident of Socrates). But now consider a situation such as Simmias's thinking about Socrates, or God's being related to a creature -- that is, a situation where two substances are related by a property or accident inhering in only one of the *relata*. What is to be said about the relationship between the relevant property and concepts in situations of this sort? Obviously, this is a question that arises regardless of how one analyzes the nature of the accident itself.

Initially, we might expect medievals to respond to this by saying that, in such situations, there is a single

accident corresponding to both members of the relevant pair of relational concepts. Thus, if Simmias is thinking about Socrates, one and the same thing corresponds to both the concepts 'thinking of' and 'thought about', namely, an accident of Simmias. The problem with this suggestion, however, is that it leads to the same sort of difficulties we encountered earlier for identifying relations with pairs of accidents taken jointly. Socrates's being thought about appears to be in Socrates and Socrates alone. But if we identify this relation with an accident of Simmias, then we shall have to say that Socrates's being thought about is not in Socrates after all, but only in Simmias. Although medievals do allow for cases of what they call *extrinsic denomination* -- that is, cases where a property or accident of one thing is predicated of something else -- they typically do not allow for this in the case of relations. That is to say, relations are typically regarded as intrinsic to the subjects of which their corresponding terms or concepts are predicated.[69] Evidently, therefore, the relevant accident of Simmias can correspond to the concept 'thinking of', but not to the concept 'thought about'. But, then, we still need to know what corresponds to this latter concept.

Now perhaps it will be suggested that, in the absence of any property or accident of Socrates to correspond to the concept 'thought about', Socrates himself can serve as the correspondent. In that case, however, we should have to admit that an individual substance is a relation (for according to the standard medieval characterization of relations, whatever corresponds to a relational concept is a relation). But such an admission goes against deep-seated intuitions deriving from the *Categories*. As we have seen, Aristotle not only identifies relations with accidents in the *Categories*, but intentionally characterizes relations in such a way as to exclude substances from this category. But then we still lack a solution to our problem. If there is no property or accident of Socrates to correspond to the concept 'thought about', and neither Socrates nor a property of anything else can correspond to it, then what, if anything, can?

It is at this point that the medieval notion of a relation of reason (*relatio rationis*) becomes important. Medieval philosophers often say that even if there are relational situations involving only a single accident, nonetheless we must still *conceive* of these situations *as if* they involved a pair of accidents, one belonging to each of the related things.[70] Like Boethius, the medievals accept the view that relations cannot be understood to exist by themselves or apart from their correlatives. Relations, they say, always come in pairs. Perhaps this is because they think our understanding of relational situations in general is based on our understanding of the paradigmatic cases, which always do involve pairs of accidents. In any case, medieval philosophers take the fact that relations always come in pairs to show that even if there is no real or extramental property in Socrates that accounts for his being thought about by Simmias, we must nonetheless *conceive* of this situation as if there were one and, as it were, *project* this property onto Socrates. Since such projections depend for their existence on the activity of the mind, medieval philosophers refer to them as beings of reason (*entia rationis*). And their suggestion is that we take these beings of reason to be the items corresponding to or signified by concepts such as 'thought about'.

The notion of a being of reason, or more specifically the notion of a relation of reason, does not appear to have been invoked in the Latin west before the thirteenth century, when Aristotle's *Metaphysics* and certain Muslim philosophical commentaries and treatises derived from it began to circulate widely. The notion of a relation of reason is not to be found explicitly in the works of Aristotle, though in the *Metaphysics* he does distinguish real beings from beings of reason.[71] Relations of reason are, however,

explicitly invoked by certain Muslim philosophers, most notably Avicenna (d. 1037), and it may well be that the distinction between real relations and relations of reason makes its way into the Latin west because of them.

However that may be, once the distinction is introduced in the Latin west, it becomes pervasive -- so pervasive, in fact, that even philosophers, such as Ockham, who complain that such a distinction is "not to be found in the writings of Aristotle" and that "'relation of reason' is not a philosophical term", nevertheless feel compelled to give some account of it in order to preserve common usage.[72] The pervasiveness of this distinction is explained at least partly by the fact that medieval philosophers think it can be used to clarify and explain a number of troublesome non-paradigmatic relational situations. Thus, by the end of thirteenth century, most philosophers explain the doctrine of creation in terms of this distinction, saying that creatures are related to God by a real relation, but God is related to them by a mere relation of reason.[73] Again, they often use the distinction to clarify certain cases of relational change. Thus, when a substance acquires a new relation without undergoing any real change this is often explained by saying that the substance acquired a mere relation of reason.[74] Finally, some medieval philosophers use relations of reason to identify a sense in which God can have accidents after all. Since relations of reason are mere projections (i.e., properties a thing has by virtue of the activity of some mind), it is possible to conceive of them as accidents in a broad sense -- that is, as properties or features that a thing can both acquire and lose. But since the acquisition or loss of these properties does not require a subject to undergo any real change, it is sometimes said that there is no reason in principle why even God should not have accidents of this sort.[75]

5.2 A Shifting Conception of Relations

The distinction between real relations and relations of reason has a number of important consequences for the development of medieval discussions of relations. For one thing, it enables philosophers to develop a number of further refinements and distinctions within the category of relations. By the mid-thirteenth century, for example, it becomes common to say that the category of relations is unique in allowing mere beings of reason among the things signified by its terms. To quote from Aquinas who is representative in this regard:

> The other genera, by virtue of what they are, posit something in extramental reality. Thus, quantity by virtue of the fact that it is quantity posits something. But relation alone is such that, by virtue of what it is, it does not posit anything in extramental reality, for it does not predicate *something* but *toward something*. Hence we find certain relations that do not posit anything in extramental reality, but only in reason.[76]

Again, Aquinas and others appeal to this special feature of the category of relations (namely, that the things signified by its terms can either be real beings or mere beings of reason) to provide a systematic division of relations or pairs of correlatives into three different types, depending on whether the members of these pairs are both real, both conceptual, or mixed (i.e., one member is real, the other conceptual). To quote again from Aquinas:

> It must be known that, since a relation requires two *relata*, there are three ways in which it can be something real or conceptual:
>
> [1] Sometimes it is a mere being of reason on the part of both *relata*, namely, when the order or [relative] disposition cannot exist between things except in virtue of the apprehension of reason alone. For example, when we say that something is identical to itself. For in virtue of the fact that reason apprehends one thing twice, it regards it as two; and in this way it apprehends a certain [relative] disposition of a thing to itself. And the same thing is true of all relations between being and non-being, which reason forms insofar as it apprehends a non-being as a certain *relatum*. Again, the same is true of all relations that follow upon the activity of reason, such as genus and species, and the like.
>
> [2] Now there are other relations that are real as regards both *relata*, namely, whenever there is a [relative] disposition between two things in virtue of something really belonging to each of them -- as is clear from all relations that follow on quantity, such as large and small, double and half, and things of this sort. For there is a quantity in both *relata*. And the same is true of relations that follow on action and passion, such as mover and movable, father and son, and the like.
>
> [3] Sometimes, however, a relation is something real in one of the *relata* and a mere being of reason in the other. And this happens whenever the two *relata* do not belong to a single order. For example, sense perception and knowledge are related to things that are sensible and intelligible. But insofar as the latter are things existing in extramental reality, they are outside the order of sensible and intelligible being. And so there is a real relation in the knowledge and sense perception in virtue of the fact that they are [really] ordered to things that can be known or sensed. However, things [that can be known or sensed] are outside this sort of order [when] considered in themselves, and hence there is not really a relation in them to knowledge or sense perception. On the contrary, there is only a relation in them in accordance with reason, insofar as the intellect apprehends them as terms of the relations of knowledge and sense perception. This is why the Philosopher says in *Metaphysics* V that they are spoken of relatively, not because they are related to other things, but because other things are related to them. Similarly, being to the right is not said of a column unless it is placed to the right of some animal. Hence a relation of this sort is not really in the column but in the animal.[77]

In this passage, Aquinas contrasts the relations involved in paradigmatic relational situations -- namely, relations of the second type -- with the relations involved in two different sorts of non-paradigmatic relational situation. We are already familiar with relations of the third or ‘mixed’ type -- which are comprised by pairs of accidents only one of whose members is real -- from our discussion of creation and intentional relations. Moreover, we can see that Aquinas's discussion of these relations follows the common practice of connecting them with both Aristotle's discussion in the *Metaphysics* and the Boethius-inspired example of the column. We have yet, however, to encounter relations of the first type. These are

relations comprised by pairs of properties or accidents both of whose members are beings of reason. As an example of this type of relation, Aquinas gives self-identity. And his view is that when we conceive of a situation involving this sort of relation -- say Socrates's being identical to himself -- we conceive of it both as if it involved two things ("a relation requires two *relata*" and hence "reason apprehends the one thing twice"), and as if the two things were ordered to each other by a pair of properties (or "[relative] dispositions"). Obviously, however, there are not two distinct things in extramental reality to serve as the *relata* of the relation of self-identity, much less two properties by which such *relata* are related. Like many other medievals, therefore, Aquinas concludes that in this case the relations (or relative dispositions) are not real, but mere beings of reason.[78]

Initially the claim that self-identity is a relation of reason might seem worrisome. For insofar as relations of reason depend for their existence on the activity of the mind, it would seem to follow that something's being self-identical is dependent on the activity of the mind. But that seems absurd. As medieval philosophers recognize, this sort of worry is perfectly generalizable, and so can be raised not only for situations involving pairs of relations of reason, but also for situations involving a single relation of reason. Thus, as Aquinas points out in one of his disputed questions, it might even lead one to doubt whether God's relation to his creatures can be considered a relation of reason:

> For if there were no created intellect in existence, God would still be Lord and Creator. But if there were no created intellect in existence, there would not be any beings of reason. Hence "Creator", "Lord", and terms of this sort, do not express relations of reason.[79]

This sort of worry about relations of reason helps to explain what is perhaps their most significant effect on the medieval discussion of relations -- namely, a gradual shift away from the traditional Aristotelian characterization of relations. On this characterization, as we saw earlier, relations are identified in terms of their metaphysical function -- that is to say, they are characterized as the items that actually serve to relate things. But in order to maintain that things can be self-identical apart from the activity of any mind, while at the same time maintaining that self-identity is a relation of reason, medieval philosophers have little choice but to move away from the traditional characterization. And of course the same thing is true in the case of God's relation to his creatures. Thus, as Aquinas says in reply to abovementioned doubt:

> A man is really (and not merely conceptually) identical to himself, even though his relation [of self-identity] is a being of reason. And the explanation for this is that the cause of his relation is real -- namely, the unity of his substance, which our intellect considers under the aspect of a relation. In the same way, the power to compel subjects is really in God, and our intellect considers this power as ordered to the subjects because of the subjects' order to God. It is for this reason that he is really said to be Lord, even though his relation is a mere being of reason. And for the same reason it is evident that he would be Lord [Creator, etc.] even if there were no created intellect in existence.[80]

In this passage, Aquinas makes it clear that in cases involving relations such as self-identity or God's relation to the world the *relata* are related, not by their relations (since these are mere beings of reason and hence dependent on the activity of the mind), but by what he refers to here as the *cause of* their

relations.[81] Now, in the case of a man's being self-identical, Aquinas says the cause is just “the unity of his substance”, where by this he seems to mean that what makes a man identical to himself is just the man himself. Again, in the case of God's being Lord he says that the cause is “the power to compel subjects”. In the case of God, however, Aquinas does not think the power to compel subjects is distinct from its subject, namely, the divine nature. Hence, he maintains that what makes it true that God is Lord is nothing but God, his creatures, and some property or attribute of the creatures, namely, their dependence for their existence on God.

In effect, therefore, reflection on relations of reason brings about a shift away from the conception of relations as items that relate, and thus forces medieval philosophers to fall back on what they might otherwise have thought of as an equivalent characterization, namely, the view that relations are items corresponding to or signified by our relational concepts. Thus, even if self-identity or God's relation to the world is a mere being of reason, and hence does not actually relate its subject to anything, nonetheless it can still be regarded as a relation on the grounds that it is signified by a relational concept. Now obviously this shift away from the traditional Aristotelian conception has the awkward consequence that things can be related even if their relations do not exist. Thus, Socrates can be identical to himself even if there is no self-identity, God can be Lord of creation even if his relation of Lordship does not exist, and more generally, a judgment of the form ‘*aRb*’ can be true, even when no judgment of form ‘*R*-ness exists’ is true. Of course, there is nothing ultimately incoherent about this consequence, provided we keep in mind that the relations in such cases are mere beings of reason. Nonetheless, accepting this consequence does force medieval philosophers to deny what at least initially appears to be a truth of reason, and at any rate is part of common sense -- namely, that things are related by their relations.[82]

Some philosophers, such as Aquinas, assume that the departure from the traditional Aristotelian characterization of relations is required only in non-paradigmatic relational situations. By the end of the Middle Ages, however, this sort of departure is so common and familiar that philosophers no longer feel the need to regard it as exceptional. Thus, Ockham eventually adopts a view according to which all relations depend for their existence on the activity of the mind. Indeed, things are so changed by the time of Ockham that he feels free not only to reject the traditional Aristotelian characterization of relations, but also to modify the standard medieval alternative to it. Thus, on his preferred characterization, relations are the items corresponding, not to all of our relational concepts, but to just one of them -- namely, the concept ‘relation’. As Ockham sees it, moreover, the concept ‘relation’ is a term of second intention -- that is to say, a term to which only concepts correspond. Thus, even though Ockham insists that many judgments of the form ‘*aRb*’ are true independently of the mind, he nonetheless maintains that properly speaking relations exist only in the mind as concepts.[83] This helps to explain why he often expresses his view using formulas such as: “This white thing really is similar, even though similarity is not really in this white thing.”[84]

5.3 Relations as Substances

Although relations of reason force a shift away from the Aristotelian characterization of relations, it is important to recall that they were originally invoked to preserve the deep-seated Aristotelian conviction

that no substance is a relation. Thus, as we saw, in order to avoid saying that Socrates is signified by the concept ‘thought about’, when he is being thought about by Simmias, medieval philosophers invoke the notion of a relation of reason: it is not Socrates, but a relation of reason, they say, that corresponds to the concept in question, and in this way avoid the consequence that a substance such as Socrates is a relation. And of course the same sort of invocation is needed to avoid saying that Socrates corresponds to the concept of ‘self-identical’ or that God corresponds to the concepts ‘Lord’ or ‘Creator’. Given the lengths to which medieval philosophers are willing to go, in situations such as these, to preserve the thesis that no substance is a relation, it is all the more interesting that there is at least one case -- namely, the Christian doctrine of the Trinity -- in which they are forced to admit that even this Aristotelian thesis cannot be upheld.

According to the Christian doctrine of the Trinity, God exists in three persons: Father, Son, and Holy Spirit. As this doctrine was typically understood during the Middle Ages, it implies not only that God possesses certain relations -- such as fatherhood and sonship -- but also that he possess them independently of the activity of any mind. As Aquinas says in his *Summa Theologiae*:

> Someone is said to be a father only by virtue of his fatherhood, and someone is said to be a son only by virtue of his sonship. Therefore, if [the relations of] fatherhood and sonship are not really in God, it follows that God is not a Father or Son really, but merely according to a concept of the mind -- which is the Sabellian heresy.[85]

Now when the claim that there are real relations in God is combined with another doctrine that was ubiquitous in the Middle Ages, namely the doctrine of divine simplicity, the conclusion that God (and hence at least one substance) is a relation seems to follow necessarily. For as the doctrine of divine simplicity is typically understood, there is no real distinction to be drawn between God and any of his attributes. Thus, if God is good, he is identical to his goodness; if he is wise, he is identical to his wisdom. By parity of reasoning, therefore, if God is a father or son, he must be identical to his fatherhood and sonship. Again, Aquinas is perfectly representative in this regard: “Whatever is in God *is* his nature . . . It is thus clear that a relation really existing in God is identical to his nature according to reality, and does not differ from it except according to a concept of the mind.”[86]

The doctrine of the Trinity, therefore, brings us to what is perhaps the medievals' greatest departure from Aristotle. Some medievals, however, see the Trinity not as providing a counterexample to Aristotle's thesis that no substance is a relation, but rather as calling our attention to a restriction on its range of applicability. Aquinas, for example, appears to think that the thesis was specifically formulated to apply only to the case of creatures,[87] and to some extent this is plausible, since obviously Aristotle was not thinking about theological examples such as the Trinity when he formulated it. Interestingly, however, other philosophers think that the thesis does not hold even in the case of all creatures. Thus, Gilbert of Poitiers (d. 1154), in a discussion of Boethius's *De hebdomadibus*, suggests that creaturely goodness is a relation, indeed, just the relation of being created by God. As I read him, moreover, Gilbert takes this relation to be nothing over and above the creature itself, so that in the case of creatures it is individual substances that correspond to the concept ‘created by’ and hence qualify as relations. Like Aquinas, however, Gilbert appears to think of these sorts of cases, not as providing counterexamples to Aristotle's

thesis, but rather as telling us something about the scope of its applicability. According to Gilbert, there is a distinction to be drawn between *natural philosophy*, which he says deals with natural things, and other areas of intellectual inquiry, including *theology* and *ethics*, which he says deal with a broader scope of things.[88] And Gilbert's suggestion is that if we restrict our attention to natural philosophy, as no doubt Aristotle did, then like him we will be led to the conclusion that no substance is a relation.[89]

Although Gilbert and Aquinas work hard to preserve, at least in non-theological contexts, something like the Aristotelian thesis that no substance is a relation, not all medieval philosophers feel the need to do so. Indeed, as Middle Ages progress, there appears to be a gradual shift, even in non-theological contexts, towards allowing substances to be relations, or at least towards allowing them to be the primary *significata* of our relational concepts. Here again Ockham appears to be an important transitional figure. Thus, he breaks with tradition in allowing that even self-identity is a real relation, or at least that it is “real . . . in the same way that similarity and equality are”.[90] Indeed, on Ockham's view, which becomes influential in the generations following him, it turns out that substances are signified by most of our relational concepts. For apart from some species of quality, Ockham thinks there is no real distinction to be drawn between substances and any of there accidents. According to him, therefore, relational situations do not typically involve anything more than individual substances.[91]

6. A Shift in Paradigms

We are now in a position to appreciate, I think, the main types of views that medievals developed concerning the nature and ontological status of relations, as well as the main historical and dialectical considerations that helped to shape them. As we have seen, with the exception of thinkers such as Peter Aureoli, medievals appear to have been drawn (almost to a person) to a form of realism about relations, one according to which at least some judgments of the form ‘*aRb*’ are true independently of the mind. There is some disagreement as to the precise analysis of the situations that makes these sorts of judgments true, but even here the medievals work out their views from within a common framework provided by Aristotle's *Categories*. Thus, it is generally agreed that relational situations do not include anything corresponding to the notion of a polyadic property, but instead include only substances and their monadic properties or accidents. The main disagreements, therefore, are best characterized as disagreements about the extent to which the proper analysis ought to conform to the paradigm suggested by the *Categories*. Prior to the fourteenth century, as we have seen, medievals tended to follow Aristotle in claiming that things are related by their accidents, and hence that relational situations typically involve not only pairs of substances, but pairs of monadic properties or accidents as well. As a result, one of the most pressing questions during this period concerns the precise nature of the accidents involved in relational situations. Reductive realists, as we have seen, identify them with ordinary, non-relational accidents such as quantities or qualities, whereas non-reductive realists identify them with monadic properties of a *sui generis* type.

With the advent of the fourteenth century, however, important changes begin to take place in the medieval discussion of relations. Philosophers and theologians continue, of course, to allow for situations in which substances are related by their accidents, and hence to worry about the precise nature of the accidents

involved in these situations. But at this point there is a decided shift toward regarding such situations as exceptional. As we have seen, there was always strong theological pressure to allow for at least some departures from the Aristotelian paradigm. But what appears to have happened over time is that these departures come to seem less and less peculiar to medievals, and eventually the departures themselves provide the basis for a new analysis of relational situations -- one according to which substances are the items responsible for relating. Around this same time, moreover, Ockham and his followers, most notably John Buridan, institute another sort of change, namely, a shift away from a standard medieval semantic characterization of relations. Prior to the fourteenth century, philosophers and theologians typically assume that the term "relation" signifies whatever it is in a relational situation that does the relating -- though here again theological considerations force them to allow for certain exceptions. By the time of Ockham, however, we get a complete break with this standard characterization. Thus, whereas earlier philosophers and theologians would allow the term "relation" to signify a being of reason in certain cases, such as creation, Ockham and his followers maintain that "relation" is a term of second intention, and so strictly speaking always signifies beings of reason. With Ockham, therefore, we have not only a complete severing of the connection between relations and those items in relational situations that actually do the relating, but also the advent of a new -- and, I might add, fairly harmless -- form of anti-realism.[92]

In the end, therefore, I think it is fair to say that the fourteenth century marks a shifting of medieval paradigms both with respect to the proper analysis of relational situations and with respect to the proper characterization of relations. I hasten to add, however, that these shifts cannot be fully explained in terms of developments within the medieval discussion of relations, but are instead part-and-parcel of broader theoretical shifts in medieval accounts of the relationship between mind, language, and reality -- shifts which are closely associated with the rise of late-medieval nominalism generally.[93]

Bibliography

[Note: This bibliography contains only items referred to in the notes and a few other selected items of interest. Further discussion of medieval theories of relations, as well as further bibliographical references, can be found in Henninger 1989, Olson 1987, and Weinberg 1965. For a contemporary defense of the view I have been calling "realism without polyadic properties", see Campbell 1990 and Fisk 1973.]

Primary Literature

- Albert the Great, *Metaphysica*, B. Geyer (ed.), *Alberti Magni Opera omnia edenda curavit*, vol. XVI, 1960-64.
- Albert the Great, *Liber de praedicamentis*, A. Borgnet (ed.), *Alberti Magni Opera omnia*, Vivès: Paris, 1890.
- Aristotle, *The Complete Works of Aristotle*, J. Barnes (ed.), Princeton: Princeton University Press, 1984.
- Augustine, *De Trinitate Libri XV*, W. J. Mountain (ed.), *Libri XV Corpus Christianorum*, Series Latin, 50, Turnhout: Brepols, 1968.
- Avicenna, *Liber de Philosophia Prima, sive Scientia Divina I-IV*, S. Van Riet (ed.), E. Peters:

Louvain/Leiden, 1977.

- Boethius, *In Categorias Aristotelis*, in *Patrologiae Latinae Cursus Completus*, J. P. Migne (ed.), Vivès: Paris, 1860, vol. 64.
- Freddoso, A. J., and F. E. Kelley, *William of Ockham: Quodlibetal Questions, Volumes 1 and 2, Quodlibets 1-7*, Yale University Press: New Haven/London, 1991.
- Henry Harclay, "Utrum Dei ad creaturam sit relatio realis," in "Henry of Harclay's Question on Relations," M. G. Henninger (ed.), *Mediaeval Studies* 49 (1987), pp. 76-123.
- Hume, D., *A Treatise of Human Nature*, L. A. Selby-Bigge (ed.), Clarendon Press: Oxford, 1888.
- John Buridan, *Summulae de Dialectica*, an annotated translation with an introduction, G. Klima, Yale Library of Medieval Philosophy, Yale University Press: New Haven, forthcoming in 2001.
- John Buridan, *Summulae: In Praedicamenta*, introduction, critical edition, and appendices, E.P. Bos, *Artistarium* 10-3, Nijmegen: Ingenium Publishers, 1994.
- John Duns Scotus, *Ioannis Duns Scoti Ordinis Fratrum Minorum Opera Omnia*, C. Bali• *et al.* (eds.), Typis Polyglottis Vaticanis: Vatican, 1950-.
- Leibniz, G. W., *Die Philosophische Schriften von Gottfried Wilhelm Leibniz*, C. I. Gerhardt (ed.), Georg Olm Verlag: Hildesheim, Germany, 1965.
- Locke, J., *An Essay Concerning Human Understanding*, Oxford University Press: Oxford, 1975.
- Loux, M. J. (tr.), *Ockham's Theory of Terms: Part I of Ockham's Summa Logicae*, University of Notre Dame Press: Notre Dame, 1974.
- MacDonald, S. (ed. and tr.), *Cambridge Translations of Medieval Philosophical Texts: Metaphysics*, Cambridge University Press: Cambridge, forthcoming.
- Peter Abelard, *Logica 'ingredientibus'* in *Peter Abaelards Philosophische Schriften* I, vol. 21, B. Geyer (ed.), Aschendorff: Münster, 1933.
- Peter Aurieoli, *Scriptum super Primum Sententiarum*, Vatican Library MS, Borghese 329, fols. 1-519.
- Spade, P. V. (tr.), *Five Texts on the Mediaeval Problem of Universals: Porphyry, Boethius, Abelard, Duns Scotus, Ockham*, Hackett: Indianapolis/Cambridge, 1994.
- Stewart, H. F, E. K. Rand, and S. J. Tester (edd. And tr.), *The Theological Tractates and the Consolation of Philosophy: Text and Translations*, Harvard University Press: Cambridge, Mass., 1978.
- Thomas Aquinas, *Opera Omnia*, R. Busa (ed.), Frommann-Holzboog: Stuttgart-Bad Canstatt, 1980.
- Velecky, C., (tr.), *St. Thomas Aquinas, Summa Theologiae, Volume 6: The Trinity (Ia. 27-32)*, Blackfriars: London/New York, 1965.
- William Ockham, *Opera Philosophica*, Ph. Boehner, *et al*. (eds.), 10 vols., The Franciscan Institute, St. Bonaventure, NY, 1967-86.
- William Ockham, *Opera Theologica*, Ph. Boehner, *et al*. (eds.), 7 vols., The Franciscan Institute, St. Bonaventure, NY, 1974-88.

Secondary Literature

- Adams, M. M., 1987, *William Ockham*, 2 vols., University of Notre Dame Press: Notre Dame, IN.
- Addis, L., 1989, *Natural Signs: A Theory of Intentionality*, Temple University Press: Philadelphia.

- Ashworth, E. J., 1974, *Language and Logic in the Post-Medieval Period*, Synthese Historical Library, vol. 12, Dordrecht: D. Reidel.
- Boler, J., 1985, “Ockham's Cleaver,” *Franciscan Studies* 45 (1985), pp. 119-144.
- Brower, J. E., 2001, “Relations without Polyadic Properties: Albert the Great on the Nature and Ontological Status of Relations,” *Archiv fur Geschichte der Philosophie* (forthcoming).
- Brower, J. E., 1998, “Abelard's Theory of Relations: Reductionism and the Aristotelian Tradition,” *The Review of Metaphysics*, 51 (1998), pp. 605-631.
- Campbell, K., 1990, *Abstract Particulars*, Basil Blackwell: Oxford.
- Cover, J. and J. O'Leary-Hawthorne, 1999, *Substance and Individuation in Leibniz*, Cambridge University Press: Cambridge.
- Crane, T (ed.), 1996, *Dispositions: A Debate*, Routledge: London.
- Fisk, M., 1973, *Necessity: An Essay in Physical Ontology*, Indiana University Press: Bloomington.
- Gracia, J., 1988, *Introduction to the Problem of Individuation in the Early Middle Ages*, Analytica Series, München and Washington, D.C.: Philosophia Verlag and Catholic University of America Press.
- Henninger, M., 1989, *Relations: Medieval Theories 1250-1325*, Clarendon Press: Oxford.
- Klima, G., 2000, “The Medieval Problem of Universals” in *The Stanford Encyclopedia of Philosophy*, Spring 2001 Edition, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/entries/universals-medieval/>.
- Klima, G., 1999, “Buridan's Logic and the Ontology of Modes," in S. Ebbesen (ed.), *Medieval Analyses in Language and Cognition*, The Royal Danish Academy of Sciences and Letters: Copenhagen, pp. 473-495.
- Klima, G., 1993, “The Changing Role of *Entia Rationis* in Medieval Philosophy: A Comparative Study with a Reconstruction,” *Synthese* 96 (1993), pp. 25-59.
- Klima, G., 1991, “Ontological Alternatives vs. Alternative Semantics in Medieval Philosophy,” *S-European Journal for Semiotic Studies* 3 (1991), pp. 587-618.
- Krempel, A., 1952, *La doctrine de la relation chez saint Thomas*, J. Vrin: Paris.
- Kretzmann, N., Kenny, A., and Pinborg, J. (eds.), 1982, *The Cambridge History of Later Medieval Philosophy*, Cambridge University Press: Cambridge.
- MacDonald, S., 1999, "Gilbert of Poitiers' Metaphysics of Goodness," *Recherches de Théologie et Philosophie médiévales* (1999), pp. 57-77.
- MacDonald, S., 1991, *Being and Goodness: The Concept of the Good in Metaphysics and Philosophical Theology*, Cornell University Press: Ithaca.
- Marinozzi, C., 1964, “La relazione trascendentale in S. Alberto M.,” *Laurentianum* 5 (1964), pp. 71-113.
- Marinozzi, C., 1965, “La realtà delle relazioni secondo S. Alberto Magno,” *Laurentianum* 6 (1965), pp. 31-72.
- Marmura, M. E., 1975, “Avicenna's Chapter, ‘On the Relative’, in the Metaphysics of the Shif•,” in G. F. Hourani (ed.), *Essays on Islamic Philosophy and Science*, State University of New York Press: Albany.
- Menn, S., 1997, “Suárez, Nominalism, and Modes,” in *Hispanic Philosophy in the Age of Discovery*, K. White (ed.), The Catholic University of America Press: Washington, D.C.
- Olson, K. R., 1987, *An Essay on Facts*, CSLI: Stanford.

- Schmidt, R. W., 1986, *The Domain of Logic According to Saint Thomas Aquinas*, Martinus Nijhoff: The Hague.
- Schönberger, R., 1994, *Relation als Vergleich: Die Relationstheorie des Johannes Buridan im Kontext seines Denkens und der Scholastik*, Studien und Texte zur Geistesgeschichte des Mittelalters, Band 43, E. J. Brill: Leiden.
- Spade, P. V., 1999, *The Cambridge Companion to Ockham*, Cambridge University Press: Cambridge.
- Weinberg, J., 1965, *Abstraction, Relation, and Induction: Three Essays in the History of Thought*, University of Wisconsin Press: Madison/Milwaukee.
- Wippel, J. F., 1987, "Thomas Aquinas's Derivation of the Aristotelian Categories (Predicaments)," *Journal of the History of Philosophy*, 25 (1987), pp. 13-34.

Other Internet Resources

- John Kilcullen's Medieval Philosophy Teaching Materials
- Klima, G., 1997, Comments on Peter King, "The Failure of Ockham's Nominalism", (Paper read at the Central Division Meeting of the American Philosophical Association, Pittsburgh, PA, April 26, 1997).
- Klima, G., 1998, Comments on Jack Zupko, "Philosophy Among the Artistae : A Late-Medieval Picture of the Limits of Rational Inquiry", (Paper read at Notre Dame Philosophy Colloquium, Notre Dame, IN, March 27, 1998).
- Paul Spade's Medieval Logic and Philosophy site
- Spade, P. V., 1996a, *Thoughts, Words and Things: An Introduction to Late Mediaeval Logic and Semantic Theory,* Version 1.0, 1996. (PDF file)
- Spade, P. V., 1985, *A Survey of Medieval Philosophy*, Version 2.0, 1985. (PDF file)

Related Entries

categories: medieval theories of | medieval philosophy | nominalism: medieval versions of | properties | realism | substance | tropes | universals: the medieval problem of

Acknowledgements

I am grateful to Gyula Klima, Paul Studtmann, Jack Zupko, and especially Susan Brower-Toland for detailed comments and suggestions on earlier versions of this article.

Jeffrey E. Brower
Purdue University

brower@purdue.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 29, 2001
Content last modified: May 29, 2001

Stanford Encyclopedia of Philosophy
Notes to Medieval Theories of Relations

Notes

1. For the record, the other categories are action, passion, time, place, position, and habit.

2. In medieval philosophy, "similar to" (or "*similis*" followed by the dative case) just means *same in quality as*.

3. In *Categories* 7, Aristotle makes it clear that he thinks the relating properties must be accidents, for according to him no substance is a relation. See *Categories* 7, 8a14f.

4. *Metaphysics* V, 15, 1021a30. All translations of Aristotle are taken, with slight modification, from *The Complete Works of Aristotle*.

5. Augustine discusses both types of theological consideration in *De Trinitate* V, especially pp. 208-215. This text exercises enormous influence on the subsequent medieval treatment of relational situations.

6. Boethius sets the precedent here. In his commentary on the *Categories*, which introduced medieval philosophers to all three terms, he not only alternates among them but explicitly denies that there is any difference in meaning between two of them, namely, "toward something" and "relative": "Whether we say *toward something* or *relatives* makes no difference." Boethius, *In Categorias Aristotelis*, p. 217.

7. *Liber de praedicamentis*, p. 225b.

8. *Categories* 7, 6a36-6b.

9. Medievals rely on context (rather than quotation marks) to distinguish use and mention, and hence employ the same terms to refer both to relations and to the predicates by means of which they are introduced.

10. *Categories* 7, 8a29-35.

11. ST I, q. 13, a. 7, ad 1.

12. *Categories* 7, 8b22-24.

13. *In Categorias Aristotelis*, p. 238.

14. Cf. Weinberg 1965, esp. pp. 61-63.

15. Cf. Albert the Great, *Liber de praedicamentis*, p. 241a-241b; Aquinas ST I, q. 28, a. 4, obj. 5; and the texts of Harclay cited below.

16. See *Physics* III, 202b11-15.

17. *Scriptum super Primum Sententiarum*, fols. 318^{v}a-b. See Henninger 1989, pp. 153-4, n. 12, for the Latin text.

18. *Scriptum super Primum Sententiarum*, fols. 318^{v}a-b.

19. Cf. *Categories* 1, 1a20-1b6.

20. Though following Boethius, medievals argue vigorously for this claim in the context of their discussions of universals. For Boethius's defense of this claim and consequent rejection of realism about substance universals, see Spade 1994, pp. 21-22. For a reconstruction of Boethius's argument and discussion of its historical significance, see Klima 2000 and the texts cited in his notes.

21. 1SN, d. 27, q. 1, a. 1, ad 2.

22. Cf. *Tractatus de quantitate* q. 3 (in *Opera Theologica* x, p. 104). There is some debate about how exactly Ockham's ‘razor-type’ claims are to be interpreted. For relevant discussion, see Boler 1985 and the notes cited in his article; cf. also Spade 1999, pp. 100-117.

23. *Liber de praedicamentis*, 222b. For further discussion, see Brower 2001.

24. *Scriptum super Primum Sententiarum*, fols. 318^{v}a-b. For useful discussion of Aureoli's views, see Henninger 1989, pp. 150-173.

25. In the early modern period, philosophers habitually speak of relations as the “products of comparison” or “results of thought”, while at the same time allowing for things to be related apart from the activity of any mind. See, e.g., Locke, *An Essay Concerning Human Understanding*, bk. 2, chap. 2; Hume, *A Treatise of Human Nature*, bk. 1, pt. 1, sec. 5; and Leibniz, *Philosophische Schriften*, vol. 2, p. 486. Cf. also Henninger 1989, pp. 184-186; Olson 1987, pp.37-43; and Cover and O'Leary-Hawthorne 1999, pp. 58-86.

26. See *Liber de praedicamentis*, 222b-223a.

27. Cf. ST I, q. 28, a. 2, corpus; cf. also Veckley 1965, pp. 28-9.

28. In his commentary on Aristotle's *Metaphysics*, Avicenna suggests that most of his contemporaries endorse a form of anti-realism. For a translation and discussion of the relevant texts, see Marmura 1975. On the Mutakallim•n, cf. Weinberg 1965, 89-91.

29. Part of Aureoli's innovation consists in the way he tries to accommodate this intuition. See Henninger, 1989, pp. 166-68.

30. *Ordinatio* I, d. 30, q. 1 (in *Opera Theologica* iv, pp. 316-317). Cf. also Scotus, *Ordinatio* II , d. 1, q. 5, n. 224 (in *Opera Omnia*); Aquinas, *De potentia*, q. 7, a. 9.

31. *De potentia*, q. 7, a. 9.

32. Cf., e.g., Aquinas, ST I, q. 28, a. 1 and *De potentia* q. 8, a.1.

33. This summary provides what I take to be the point of Albert's difficult discussion in *Liber de praedicamentis*, 224a-224b. For a detailed defense of this interpretation, and its contemporary relevance, see Brower 2001.

34. *Summa Logica* I, c. 51 (in *Opera Philosophica* i, p. 171); cf. Loux, 1974, p. 171.

35. *In Categorias Aristotelis*, p. 217.

36. See, e.g., Abelard, *Logica 'ingredientibus'*, pp. 80-95; and Albert the Great, *Liber de praedicamentis*, p. 22.

37. For discussion of Harclay's two views, and their relationship to Scotus's and Ockham's, see Henninger 1989, pp. 98-118.,

38. For a good example of this approach, see Abelard's discussion of universals in his *Logica 'ingredientibus'*, pp. 7-32. For an English translation of this discussion, see Spade 1994, pp. 26-56.

39. *Ordinatio* I, d. 30, q. 2 (in *Opera Theologica* iv, p. 322). Cf. also n. 20 above.

40. Abelard justifies this interpretation on the basis of an analogy to good-making characteristics: just as certain attributes (say, being red, juicy, and sweet) make certain things good, so too, he suggests, certain attributes (such as being-six-feet-tall and five-feet-ten) make certain things to be relative (or related). See

Logica 'ingredientibus', pp. 216-17. For further discussion, see Brower 1998.

41. *Summa logicae* I, c. 49 (in *Opera Philosophica* i, p. 154). Cf. Henninger 1989, p. 120.

42. During the thirteenth and fourteenth centuries, this intuitive view of relational change was often discussed in a connection with an authoritative passage from the *Physics*, where Aristotle says "There is no motion in respect of relation: for it may happen that when one correlative changes, the other can truly be said not to change at all, so that in these cases the motion is accidental". Cf. *Physics* V, 2, 225b11-13.

43. Ockham, *Ordinatio* I, d. 30, q. 3 (in *Opera Theologica* iv, p. 347); cf. also the other texts cited in Henninger 1989, p. 129, n. 31.

44. In passing, however, it should be mentioned that reductive realists also appealed to various forms of an infinite regress argument to support their position. This sort of argument is now typically associated with the absolute idealist, F. H. Bradley, but it was known to philosophers during the Middle Ages and taken by many (including, Ockham and the later Harclay) to show that relations cannot be really distinct from their foundations. Cf. Henninger 1989, pp. 110-112, 121-122 and Adams 1987, vol. 1, pp. 215-250. Again, reductive realists also relied on various forms of what is sometimes called the separation argument. After certain events in the late thirteenth century, including the condemnations of 1277, it was generally agreed that a real distinction between two or more items implies that either can exist in separation from the other—at least by God's power. As the reductive realists were quick to point out, however, this tells against non-reductive realism. For if relations are really distinct from their foundations, then it follows that God can create the foundations without the corresponding relations—and hence (absurdly) that, say, two white things could exist without being similar, or two quantified things without being either equal or unequal. Cf. Harclay, "Utrum Dei ad creaturam sit relatio realis" (hereafter abbreviated "Utrum Dei"), n. 18f. For further references and discussion, cf. also Menn 1997. For a variation of the separation argument, sometimes used against reductive realism, see Klima's discussion of Psuedo-Campsall vs. Ockham in Spade 1999, esp. pp. 123-127.

45. Ockham, however, leans toward this sort of view in several of his *Quodlibetal* questions, where he entertains the idea that abstract relative terms, such as 'similitude', function as collective names, and hence like other such terms ('people', 'army', 'crowd', 'company') refer to several things taken jointly (*conjunctim*). Cf. *Quodl.* VI, q. 8 (in *Opera Theologica* ix, pp. 616-617) and q. 25 (pp. 681-682).

46. "Utrum Dei", n. 46. All translations of Harclay are adopted, with slight modification, from MacDonald (forthcoming).

47. "Utrum Dei", n. 47.

48. Harclay's discussion is complicated and my interpretation is by no means uncontroversial. For an alternative interpretation of this discussion, see Henninger 1989, pp. 112-117.

49. Again, the idea here is that, by itself, Simmias's height is merely *potentially* relative-making, and it only becomes *actually* relative-making in the presence of another height, including Socrates's.

50. Cf. Armstrong's description of his own view in Crane 1996, esp. pp. 39-41.

51. For a formal reconstruction of this sort of view, see Klima 1991.

52. “Utrum Dei,” n. 43.

53. *Quodl*. VI, q. 25 (in *Opera Theologica* ix, p. 679).

54. M*etaphysica*, p. 266a.

55. Harclay uses the term “toward-ness” (*aditas*) in his question on relations (see “Utrum Dei,” n. 50), whereas Aquinas often uses “disposition” (*habitudo*) or “relative disposition” (*relativa habitudo*), and eventually comes to prefer “order” (*ordinatio*). For a helpful discussion of some of the terms used by medieval philosophers to signify relations, see Schmidt 1986, esp. pp. 133-40.

56. Especially worth mentioning in this context is Addis 1989 (see esp. pp. 27-46). Addis identifies several other contemporary or near-contemporary philosophers who hold this sort of view (including Meinong, Husserl, Bergmann, and Searle), but anyone holding a form of the so-called adverbial theory of consciousness would also appear to be a candidate.

57. Harclay explicitly addresses this sort of non-reductivist line in “Utrum Dei,” n. 10f.

58. Cf. *Metaphysica*, 266b-267a, ad 1.

59. *Metaphysica*, 267a, ad. 3.

60. *In V Phys*., lect. 3, n. 8.

61. *In V Phys*., lect. 3, n. 8.

62. For this interpretation of Aquinas, see Henninger 1989, pp. 13-28.

63. Thus, according to Henninger, the proper description of Aquinas's view about relations and their foundations will vary depending on which criteria one uses for identity and real distinction. See Henninger 1989, 29-31.

64. Augustine, *De Trinitate*, pp. 224-27.

65. *De Trinitate*, p. 226. ("…quamvis temporaliter incipiat dici, non tamen ipsi substantiae dei accidisse intelligatur sed illi creaturae ad quam dicitur")

66. *De Trinitate*, p. 226. Augustine introduces the example of the coin in the course of defending his views about of creation. Thus, he says, if a coin can change its relations solely in virtue of changes in the properties of something else, "how much easier ought we to accept this of the unchangeable substance of God?" (ibid).

67. Boethius's example actually involves a man walking up beside another, stationary human being. See *De Trinitate* 5. Medieval philosophers, however, prefer to use the example of a man and an immobile column, and over time come to speak as if this were Boethius's own example. Thus, as Harclay says at one point: "Boethius offers the example of a column's being to the right at the end of *De Trinitate*." See "Utrum Dei," n. 72.

68. Peter Lombard (c. 1095-1160) discusses the relationship between God and creatures in the first book of his *Sentences*, an influential summary of Christian doctrine. From the mid-thirteenth through the mid-fourteenth century, every student who earned a baccalaureate in theology was required to lecture and comment on Peter's text. As a result, commentaries on the first book of this text (around distinction 30) become the *locus classicus* for medieval discussions of relations. Cf. Henninger 1989, p. 8.

69. This way of thinking about relations derives from a late 12th-century work of unknown authorship (but traditionally ascribed to Gilbert of Poitiers) called *Liber sex principiorum*. This work distinguishes the first four Aristotelian categories (substance, quantity, quality, and relation) as intrinsic and the last six categories as extrinsic. For a representative treatment of the categories which follows this division, see Aquinas's discussion in *In V Met*., lect. 9, n. 892; cf. also the commentary in Wippel 1987. Even Aquinas, however, sometimes allows for extrinsic denomination in the case of relations. Cf. 1SN d. 40, q. 1, a. 1.

70. Cf., e.g., Aquinas, *De potentia* q. 1, a. 1, ad 10.

71. Cf. *Metaphysics* IV, 1 (esp. 1003a32-b11) and V, 7 (esp. 1017a31-35).

72. *Ordinatio* I, d. 30, q. 5 (in *Opera Theologica* iv, pp. 385-86). For a discussion of Ockham's distinction between real relations and relations of reason, see Henninger 1989, 136-40.

73. Cf. Henninger 1989.

74. In one of his early works, *Super Dionysium De divinis nominibus*, Albert the Great identifies all

relations acquired as a result of mere Cambridge changes as relations of reason—apparently overlooking the possibility of acquiring *real* relations in the same way (as in the case of Socrates's becoming shorter than Simmias as result of Simmias's growth). For the relevant texts of Albert, and helpful discussion, see MacDonald 1991, pp. 31-55 (esp. 42-47). No other medieval philosopher I know of makes this sort of identification, and Albert himself comes to reject it in some of his later works. Thus, as I indicated in section 4b above, in his *Metaphysica* Albert explicitly allows for cases in which a substance can acquire real relations without undergoing any real change, and uses the distinction between relations and relational accidents to explain how this is possible.

75. In *Monologium* c. 25, Anselm distinguishes those accidents which require a change in their subject from those that do not, and suggests that even God may have accidents of the latter sort (though as he goes on to explain, 'accidents' of this latter sort are accidents only according to an improper way of speaking). In "Utrum Dei," Harclay develops this distinction between two types of accident at greater length, and attempts to connect it not only with Anselm but also with Augustine and Boethius (cf. esp. nn. 110-120).

76. *De veritate*, q. 1, a. 5, ad 15.

77. ST I, q. 13, a. 7, corpus.

78. Interestingly, this same line of reasoning leads some philosophers to deny that self-identity is a relation at all. Consider, for example, the following passage from Harclay's "Utrum Dei," n. 32: "real identity is not a real relation, because it lacks one condition that is necessary for being a relation (and which it is impossible for something that is identical and that to which it is identical to satisfy), namely, real distinctness of *relata*. This condition is satisfied in the case of distinct things but not in the case of something that is one and the same [as itself]. For that reason, distinctness is a relation but identity is not."

79. *De potentia*, q. 7, a. 11, obj. 4.

80. *De Potentia* q. 7, a. 11, ad 3-5.

81. What Aquinas here refers to as 'cause' he elsewhere refers to as the 'foundation in reality (*fundamentum in re*)' of a being of reason (cf., e.g., SN1 d. 19, q. 5, a. 1). In the case at hand, therefore, the relevant causes are just the extramental foundations of the relations.

82. Medieval philosophers often apply this consequence to cases involving privations, such as blindness. For privations involve negation, which is typically regarded as a being of reason. Thus, it is commonly said that Homer would lack sight, even if Homer's blindness did not exist. For the truth of the latter claim depends for its existence on the activity of the mind. For further discussion and references, see Klima 1993.

83. Ockham does, however, want to maintain (with Aristotle) that relation is a real category or *genus generalissimum*. He does this by identifying a loose sense in which relations can be said to exist outside the mind, and indeed interprets Aristotle's own way of speaking along these lines: “according to the Philosopher's view, ‘relation’ is a category of the real not in the sense that it signifies things outside [the mind] … but rather in the sense that its species [i.e., specific relational concepts] signify such things [outside the mind].” See *Quodl.* VI, q. 22 (in *Opera Theologica* ix, p. 669); cf. Freddoso and Kelly 1991, p. 564 and Henninger 1989, p. 133.

84. *Quodl.* VI, q. 22 (in *Opera Theologica* ix, p. 669).

85. ST I, q. 28, a. 1, sed contra. Although this passage occurs in the *sed contra*, it is clear from other works that it represents Aquinas's own views. Cf. *De potentia*, q. 8, a. 1.

86. ST I, q. 28, a. 2, corpus.

87. Cf. ST I, q. 28, a. 2, corpus.

88. *The Commentaries on Boethius by Gilbert of Poitiers*, p. 79. Gilbert's division of the intellectual disciplines follows Boethius's discussion in *De Trinitate* 2.

89. Cf. *The Commentaries on Boethius by Gilbert of Poitiers*, pp. 193, 223, and 227. For an excellent discussion of the relevant texts, to which my own discussion is indebted, see MacDonald 1999.

90. *Quodl*. V, q. 27 (in *Opera Theologica* ix, p. 685).

91. One of Ockham's most influential followers, John Buridan (d. 1358/61) even goes so far as to say that everything is a relation (*relativum*, though not *relatio*, which like Ockham he construes as a second intention). For everything can be conceived of in terms of a relative concept, since everything is identical with itself and distinct from everything else. Cf. *Summulae de Dialectica*, 3.4.1.

92. I say “harmless” because, like other medieval realists, Ockham and his followers think many judgments of the form ‘*aRb*’ are true independently of the mind. Nonetheless, because they deny that the term “relation” refers to anything in extramental reality, and hence that strictly speaking relations are concepts or beings of reason, there is a sense in which their view qualifies as anti-realist.

93. For an excellent account of these broader historical developments, see Ashworth 1974; cf. also Klima 2000, and the exchange between Jack Zupko and Gyula Klima (in Klima 1998).

Jeffrey E. Brower
Purdue University
brower@purdue.edu

First published: May 29, 2001
Content last modified: May 29, 2001

The Medieval Problem of Universals

“The problem of universals” in general is a historically variable bundle of several closely related, yet in different conceptual frameworks rather differently articulated metaphysical, logical, and epistemological questions, ultimately all connected to the issue of how universal cognition of singular things is possible. How do we know, for example, that the Pythagorean theorem holds *universally*, for *all* possible right triangles? Indeed, how can we have any awareness of a potential infinity of all possible right triangles, given that we could only see a finite number of actual ones? How can we universally indicate all possible right triangles with the phrase ‘right triangle’? Is there something common to them all signified by this phrase? If so, what is it, and how is it related to the particular right triangles? The *medieval* problem of universals is a logical, and historical, continuation of the *ancient* problem generated by Plato's (428-348 B.C.) theory answering such a bundle of questions, namely, his theory of Ideas or Forms.

1. Introduction

The inherent problems with Plato's original theory were recognized already by Plato himself. In his

Parmenides Plato famously raised a number of difficulties, for which he apparently did not provide satisfactory answers. Aristotle (384-322 B.C.), with all due reverence to his teacher, consistently rejected Plato's theory, and heavily criticized it throughout his own work. (Hence the famous saying, *amicus Plato sed magis amica veritas*).[1] Nevertheless, despite this explicit doctrinal conflict, Neo-Platonic philosophers, pagans (such as Plotinus ca. 204-270, and Porphyry, ca. 234-305) and Christians (such as Augustine, 354-430, and Boethius, ca. 480-524) alike, observed a basic concordance between Plato's and Aristotle's approach, crediting Aristotle with an explanation of how the human mind acquires its universal concepts of particular things from experience, and Plato with providing an explanation of how the universal features of particular things are established by being modeled after their universal archetypes.[2] In any case, it was this general attitude toward the problem in late antiquity that set the stage for the ever more sophisticated medieval discussions.[3] In these discussions, the concepts of the human mind, therefore, were regarded as posterior to the particular things represented by these concepts, and hence they were referred to as *universalia post rem* ('universals after the thing'). The universal features of singular things, inherent in these things themselves, were referred to as *universalia in re* ('universals in the thing'), answering the universal exemplars in the divine mind, the *universalia ante rem* ('universals before the thing').[4] All these, universal concepts, universal features of singular things, and their exemplars, are expressed and signified by means of some obviously universal signs, the universal (or common) terms of human languages. For example, the term 'man', in English is a universal term, because it is truly predicable of all men in one and the same sense, as opposed to the singular term 'Socrates', which in the same sense, i.e., when not used equivocally, is only predicable of one man.

Depending on which of these items (universal features of singular things, their universal concepts, or their universal names) they regarded as the primary, really existing universals, it is customary to classify medieval authors as being *realists*, *conceptualists*, or *nominalists*. The *realists* are supposed to be those who assert the existence of real universals *in* and/or *before* particular things, the *conceptualists* those who allow universals only, or primarily, as concepts of the mind, whereas *nominalists* would be those who would acknowledge only, or primarily, universal words. But this rather crude classification does not adequately reflect the genuine, much more subtle differences of opinion between medieval thinkers. (No wonder one often finds in the secondary literature distinctions between, "moderate" and "extreme" versions of these crudely defined positions.) In the first place, nearly *all* medieval thinkers agreed on the existence of universals *before* things in the form of divine ideas existing in the divine mind,[5] but all of them denied their existence in the form of mind-independent eternal entities originally posited by Plato. Furthermore, medieval thinkers also agreed that particular things have certain features which the human mind is able to comprehend in a universal fashion, and signify by means of universal terms. As we shall see, their disagreements rather concerned the types of the relationships that hold between the particular things, their individual, yet universally comprehensible features, the universal concepts of the mind, and the universal terms of our languages, as well as the ontological status of, and distinctions between, the individualized features of the things and the universal concepts of the mind. Nevertheless, the distinction between "realism" and "nominalism", especially, when it is used to refer to the distinction between the radically different ways of doing philosophy and theology in late-medieval times, is quite justifiable, provided we clarify what *really* separated these ways, as I hope to do in the later sections of this article.

In this brief summary account I will survey the problem both from a systematic and from a historical point of view. In the next section I will first motivate the problem by showing how naturally the questions

concerning universals emerge if we consider how we come to know a universal claim, i.e., one that concerns a potentially infinite number of particulars of a given kind, in a simple geometrical demonstration. I will also briefly indicate why a naïve Platonic answer to these questions in terms of the theory of perfect Forms, however plausible it may seem at first, is inadequate. In the third section I will briefly discuss how the specific medieval questions concerning universals emerged, especially in the context of answering Porphyry's famous questions in his introduction to Aristotle's *Categories*, which will naturally lead us to a discussion of Boethius' Aristotelian answers to these questions in his second commentary on Porphyry in the fourth section. However, Boethius' Aristotelian answers anticipated only one side of the medieval discussions: the mundane, philosophical theory of universals, in terms of Aristotelian abstractionism. But the other important, Neo-Platonic, theological side of the issue provided by Boethius, and, most importantly, by St. Augustine, was for medieval thinkers the theory of ontologically primary universals as the creative archetypes of the divine mind, the Divine Ideas. Therefore, the fifth section is going to deal with the main ontological and epistemological problems generated by this theory, namely, the apparent conflict between divine simplicity and the multiplicity of divine ideas, on the one hand, and the tension between the Augustinian theory of divine illumination and Aristotelian abstractionism, on the other. Some details of the early medieval Boethian-Aristotelian approach to the problem and its combination with the Neo-Platonic Augustinian tradition *before* the influx of the newly recovered logical, metaphysical, and physical writings of Aristotle and their Arabic commentaries in the second half of the 12th century will be taken up in the sixth section, in connection with Abelard's (1079-1142) discussion of Porphyry's questions. The seventh section will discuss some details of the characteristic metaphysical approach to the problem in the 13th century, especially as it was shaped by the influence of Avicenna's (980-1037) doctrine of common nature. The eighth section outlines the most general features of the logical conceptual framework that served as the common background for the metaphysical disagreements among the authors of this period. I will argue that it is precisely this common logical-semantical framework that allows the grouping together of authors who endorse sometimes radically different metaphysics and epistemologies (not only in this period, but also much later, well into the early modern period) as belonging to what in later medieval philosophy came to be known as the “realist” *via antiqua*, the “old way” of doing philosophy and theology. By contrast, it was precisely the radically different logical-semantical approach initiated by William Ockham (ca. 1280-1350), and articulated and systematized most powerfully by Jean Buridan (ca. 1300-1358), that distinguished the “nominalist” *via moderna*, the “modern way” of doing philosophy and theology from the second half of the 14th century. The general, distinctive characteristics of this “modern way” will be the discussed in the ninth section. Finally, the concluding tenth section will briefly indicate how the separation of the two *viae*, in addition to a number of extrinsic social factors, contributed to the disintegration of scholastic discourse, and thereby to the disappearance of the characteristically medieval problem of universals, as well as to the re-mergence of recognizably the same problem in different guises in early modern philosophy.

2. The Emergence of the Problem

It is easy to see how the problem of universals emerges, if we consider a geometrical demonstration, for example, the demonstration of Thales' theorem. According to the theorem, any triangle inscribed in a semicircle is a right triangle, as is shown in the following diagram:

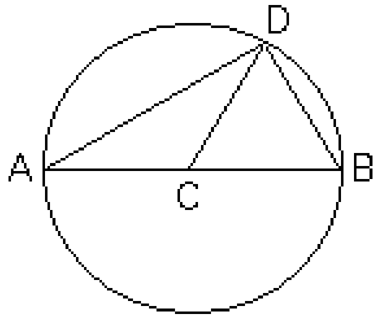


Figure 1. Thales' theorem

Looking at this diagram, we can see that all we need to prove is that the angle at vertex D of triangle ABD is a right angle. The proof is easy once we realize that since lines AC, DC, and BC are the radii of a circle, the triangles ACD and DCB are isosceles triangles, whence their base angles are equal. For then, if we denote the angles of ABD by their vertices, this fact entails that D=A + B; and so, since A + B + D=180º, it follows that 2A + 2B=180º; therefore, A + B=90º, that is, D=90º, **q. e. d.**

Of course, from our point of view, the important thing about this demonstration is not so much the *truth* of its conclusion as *the way* it proves this conclusion. For the conclusion is a universal theorem, which has to concern all possible triangles inscribed in any possible semicircle whatsoever, not just the one inscribed in the semicircle in the figure above. Yet, apparently, in the demonstration above we were talking only about that triangle. So, how can we claim that whatever we managed to prove concerning that particular triangle will hold for all possible triangles?

If we take a closer look at the diagram, we can easily see the appeal of the Platonic answer to this question. For upon a closer look it is clear that, despite appearances to the contrary, this demonstration *cannot* be about the triangle in this diagram. Indeed, in the demonstration we assumed that the lines AC, DC, and BC were all perfectly equal, straight lines. However, if we zoom in on the figure, we can clearly see that these lines are far from being equal; in fact, they are not even straight lines:

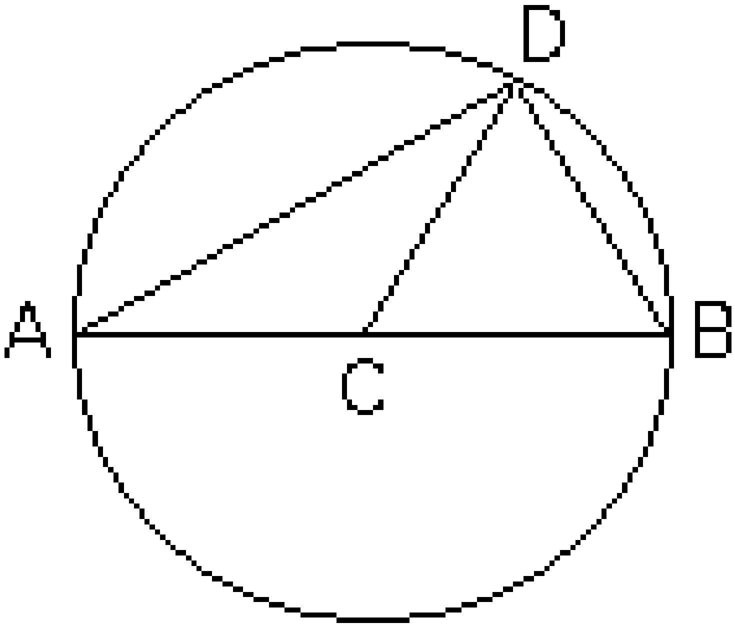


Figure 2. The result of zooming in on Figure 1.

The demonstration was certainly not about the collection of jagged black surfaces that we can see here. Rather, the demonstration concerned something we did not see with our bodily eyes, but what we had in mind all along, understanding it to be a triangle, with perfectly straight edges, touching a perfect circle in three unextended points, which are all perfectly equidistant from the center of the circle. The figure we could see was only a convenient "reminder" of what we are supposed to have in mind when we want to prove that a certain property, namely, that it is a right triangle, has to belong to the object in our mind in virtue of what it is, namely, a triangle inscribed in a semicircle. Obviously, the conclusion applies perfectly only to the perfect triangle we had in mind, whereas it holds for the visible figure only insofar as, and to the extent that, this figure resembles the object we had in mind. But this figure fails to have this property precisely insofar as, and to the extent that, it falls short of the object in our mind.

However, on the basis of this point it should also be clear that the conclusion *does* apply to this figure, and every other visible triangle inscribed in a semicircle as well, insofar as, and to the extent that, it manages to imitate the properties of the perfect object in our mind. Therefore, the Platonic answer to the question of what this demonstration was about, namely, that it was about a perfect, ideal triangle, which is invisible to the eyes, but is graspable by our understanding, at once provides us with an explanation of the possibility of universal, necessary knowledge. By knowing the properties of the Form or Idea, we know all its particulars, i.e., all the things that imitate it, insofar as they imitate or participate in it. So, the Form itself is a universal entity, a universal model of all its particulars; and since it is the knowledge of this universal entity that can enable us to know at once all its particulars, it is absolutely vital for us to know *what* it is, *what* it is *like*, and exactly *how* it is *related to* its particulars. However, obviously, all these questions presuppose that *it is* at all, namely, that such a universal entity *exists*.

But the existence of such an entity seems to be rather precarious. Consider, for instance, the perfect triangle we were supposed to have in mind during the demonstration of Thales' theorem. If it is a perfect triangle, it obviously has to have three sides, since a perfect triangle has to be a triangle, and nothing can be a triangle unless it has three sides. But of those three sides either at least two are equal or none, that is to say, the

triangle in question has to be either isosceles or scalene (taking ‘isosceles’ broadly, including even equilateral triangles, for the sake of simplicity). However, since it is supposed to be the universal model of *all* triangles, and not only of isosceles triangles, this perfect triangle cannot be an isosceles, and for the same reason it cannot be a scalene triangle either. Therefore, such a universal triangle would have to have inconsistent properties, namely, *both* that it is either isosceles or scalene *and* that it is neither isosceles nor scalene. However, obviously nothing can have these properties at the same time, so nothing can be a universal triangle any more than a round square. So, apparently, no universal triangle can exist. But then, what was our demonstration about? Just a little while ago, we concluded that it could not be directly about any particular triangle (for it was not about the triangle in the figure, and it was even less about any other particular triangle not in the figure), and now we had to conclude that it could not be about a universal triangle either. But are there any further alternatives? It seems obvious that by this demonstration we do gain universal knowledge concerning all particulars. Yet it is also clear that we do not, indeed, we cannot gain this knowledge by examining all particulars, both because they are potentially infinite and because none of them perfectly satisfies the conditions stated in the demonstration. So there must have been something wrong in our characterization of the universal, which compelled us to conclude that, in accordance with that characterization, universals could not exist. Therefore, we are left with a whole bundle of questions concerning the nature and characteristics of universals, questions that cannot be left unanswered if we want to know how universal, necessary knowledge is possible, if at all.

3. The Origin of the Specifically Medieval Problem of Universals

What we may justifiably call the first formulation of “the *medieval* problem of universals” (distinguishing it from the both logically and historically related ancient problems of Plato's Theory of Forms) was precisely such a bundle of questions famously raised by Porphyry in his *Isagoge*, that is, his *Introduction to Aristotle's Categories*. As he wrote:

> (1) Since, Chrysaorius, to teach about Aristotle's *Categories* it is necessary to know what genus and difference are, as well as species, property, and accident, and since reflection on these things is useful for giving definitions, and in general for matters pertaining to division and demonstration, therefore I shall give you a brief account and shall try in a few words, as in the manner of an introduction, to go over what our elders said about these things. I shall abstain from deeper enquiries and aim, as appropriate, at the simpler ones.
>
> (2) For example, I shall beg off saying anything about (a) whether genera and species are real or are situated in bare thoughts alone, (b) whether as real they are bodies or incorporeals, and (c) whether they are separated or in sensibles and have their reality in connection with them. Such business is profound, and requires another, greater investigation. Instead I shall now try to show how the ancients, the Peripatetics among them most of all, interpreted genus and species and the other matters before us in a more logical fashion.[6]

Even though in this way, by relegating them to a “greater investigation”, Porphyry left these questions

unanswered, they certainly proved to be irresistible for his medieval Latin commentators, beginning with Boethius, who produced not just one, but two commentaries on Porphyry's text; the first based on Marius Victorinus's (*fl*. 4th c.) translation, and the second on his own.[7]

In the course of his argument, Boethius makes it quite clear what sort of entity a universal would have to be.

> A universal has to be common to several particulars
>
> 1. in its entirety, and not only in part
> 2. simultaneously, and not in a temporal succession, and
> 3. it should constitute the substance of its particulars.[8]

However, as Boethius argues, nothing in real existence can satisfy these conditions. The main points of his argument can be reconstructed as follows.

Anything that is common to many things in the required manner has to be simultaneously, and as a whole, in the substance of these many things. But these many things are several beings precisely because they are distinct from one another in their being, that is to say, the act of being of the one is distinct from the act of being of the other. However, if the universal constitutes the substance of a particular, then it has to have the same act of being as the particular, because constituting the substance of something means precisely this, namely, sharing the act of being of the thing in question, as the thing's substantial part. But the universal is supposed to constitute the substance of all of its distinct particulars, as a whole, at the same time. Therefore, the one act of being of the universal entity would have to be identical with all the distinct acts of being of its several particulars at the same time, which is impossible.[9]

This argument, therefore, establishes that no one thing can be a universal in its being, that is to say, nothing can be both one being and common to many beings in such a manner that it shares its act of being with those many beings, constituting their substance.

This can easily be visualized in the following diagram, where the tiny lightning bolts indicate the acts of being of the entities involved, namely, a woman, a man, and their universal humanity (the larger dotted figure).

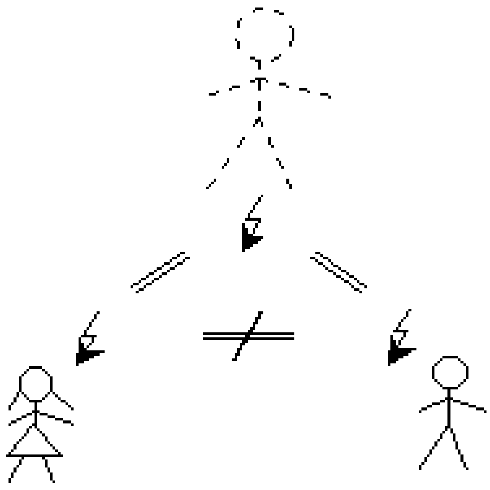


Figure 3. Illustration of the first part of Boethius' argument

But then, Boethius goes on, we should perhaps say that the universal is not one being, but rather many beings, that is, [the collection of][10] those constituents of the individual essences of its particulars on account of which they all fall under the same universal predicable. For example, on this conception, the genus 'animal' would not be some one entity, a universal animality over and above the individual animals, yet somehow sharing its being with them all (since, as we have just seen, that is impossible), but rather [the collection of] the individual animalities of all animals.

Boethius rejects this suggestion on the ground that whenever there are several generically similar entities, they have to have a genus; therefore, just as the individual animals had to have a genus, so too, their individual animalities would have to have another one. However, since the genus of animalities cannot be one entity, some 'super-animality' (for the same reason that the genus of animals could not be one entity, on the basis of the previous argument), it seems that the genus of animalities would have to be a number of further 'super-animalities'. But then again, the same line of reasoning should apply to these 'super-animalities', giving rise to a number of 'super-super-animalities', and so on to infinity, which is absurd. Therefore, we cannot regard the genus as some real being even in the form of [a collection of] several distinct entities. Since similar reasonings would apply to the other Porphyrian predicables as well, no universal can exist in this way.

Now, a universal either exists in reality independently of a mind conceiving of it, or it only exists in the mind. If it exists in reality, then it either has to be one being or several beings. But since it cannot exist in reality in either of these two ways, Boethius concludes that it can only exist in the mind.[11]

However, to complicate matters, it appears that a universal cannot exist in the mind either. For, as Boethius says, the universal existing in the mind is some universal understanding of some thing outside the mind. But then this universal understanding is either disposed in the same way as the thing is, or differently. If it is disposed in the same way, then the thing also has to be universal, and then we end up with the previous problem of a really existing universal. On the other hand, if it is disposed differently, then it is false, for "what is understood otherwise than the thing is is false".[12] But then, all universals in the understanding would have to be false representations of their things; therefore, no universal knowledge would be possible, whereas our considerations started out precisely from the existence of such knowledge, as seems to be clear,

e.g., in the case of geometrical knowledge.

4. Boethius' Aristotelian Solution

Boethius' solution of the problem stated in this form consists in the rejection of this last argument, by pointing out the ambiguity of the principle that “what is understood otherwise than the thing is is false”. For in one sense this principle states the obvious, namely, that an act of understanding that represents a thing *to be* otherwise than the thing is is false. This is precisely the reading of this principle that renders it plausible. However, in another sense this principle would state that an act of understanding which represents the thing in a manner which is different from the manner in which the thing exists is false. In this sense, then, the principle would state that if the mode of representation of the act of understanding is different from the mode of being of the thing, then the act of understanding is false. But this is far from plausible. In general, it is simply not true that a representation can be true or faithful only if the mode of representation matches the mode of being of the thing represented. For example, a written sentence is a true and faithful representation of a spoken sentence, although the written sentence is a visible, spatial sequence of characters, whereas the spoken sentence is an audible, temporal pattern of articulated sounds. So, what exists as an audible pattern of sounds is represented visually, that is, the mode of existence of the thing represented is radically different from the mode of its representation. In the same way, when particular things are represented by a universal act of thought, the things exist in a particular manner, while they are represented in a universal manner, still, this need not imply that the representation is false. But this is precisely the sense of the principle that the objection exploited. Therefore, since in this sense the principle can be rejected, the objection is not conclusive.[13]

However, it still needs to be shown that in the particular case of universal representation the mismatch between the mode of its representation and the mode of being of the thing represented does in fact not entail the falsity of the representation. This can easily be seen if we consider the fact that the falsity of an act of understanding consists in representing something *to be* in a way it is not. That is to say, properly speaking, it is only an act of *judgment* that can be false, by which we think something *to be* somehow. But a *simple* act of understanding, by which we simply understand something without thinking it *to be* somehow, that is, without attributing anything to it, cannot be false. For example, I can be mistaken if I form in my mind the judgment that a man *is* running, whereby I conceive a man *to be* somehow, but if I simply think of a man without attributing either running or not running to him, I certainly cannot make a mistake as to how he *is*.[14] In the same way, I would be mistaken if I were to think that a triangle is neither isosceles nor scalene, but I am certainly not in error if I simply think of a triangle without thinking either that it is isosceles or that it is scalene. Indeed, it is precisely this possibility that allows me to form the universal mental representation, that is, the universal concept of all particular triangles, regardless of whether they are isosceles or scalene. For when I think of a triangle in general, then I certainly do not think of something that is a triangle and is neither isosceles nor scalene, for that is impossible, but I simply think of a triangle, not thinking that it is an isosceles and not thinking that it is a scalene triangle. This is how the mind is able to separate in thought what are inseparable in real existence. Being either isosceles or scalene is inseparable from a triangle in real existence. For it is impossible for something *to be* a triangle, and yet *not to be* an isosceles and *not to be* a scalene triangle either. Still, it is not impossible for something to be *thought to be* a triangle and *not* to be *thought to be* an isosceles and *not* to be *thought to be* a scalene triangle either

(although of course, it still has to be thought to be either isosceles-or-scalene). This separation in thought of those things that cannot be separated in reality is the process of *abstraction*.[15] In general, by means of the process of abstraction, our mind (in particular, the faculty of our mind Aristotle calls *active intellect* (*nous poietikos*, in Greek, *intellectus agens*, in Latin) is able to form universal representations of particular objects by disregarding what distinguishes them, and conceiving of them only in terms of those of their features in respect of which they do not differ from one another.

In this way, therefore, if universals are regarded as universal mental representations existing in the mind, then the contradictions emerging from the Platonic conception no longer pose a threat. On this Aristotelian conception, universals need not be thought of as somehow sharing their being with all their distinct particulars, for their being simply consists in their being thought of, or rather, the particulars' being thought of in a universal manner. This is what Boethius expresses by saying in his final replies to Porphyry's questions the following:

> … genera and species subsist in one way, but are understood in an another. They are incorporeal, but subsist in sensibles, joined to sensibles. They are understood, however, as subsisting by themselves, and as not having their being in others.[16]

But then, if in this way, by positing universals in the mind, the most obvious inconsistencies of Plato's doctrine can be avoided, no wonder that Plato's “original” universals, the universal models which particulars try to imitate by their features, found their place, in accordance with the long-standing Neo-Platonic tradition, in the divine mind.[17] It is this tradition that explains Boethius' cautious formulation of his conclusion concerning Aristotelianism pure and simple, as not providing us with the whole story. As he writes:

> … Plato thinks that genera and species and the rest are not only understood as universals, but also exist and subsist apart from bodies. Aristotle, however, thinks that they are understood as incorporeal and universal, but subsist in sensibles.
>
> I did not regard it as appropriate to decide between their views. For that belongs to a higher philosophy. But we have carefully followed out Aristotle's view here, not because we would recommend it the most, but because this book, [the *Isagoge*], is written about the *Categories*, of which Aristotle is the author.[18]

5. Platonic Forms as Divine Ideas

Besides Boethius, the most important mediator between the Neo-Platonic philosophical tradition and the Christianity of the Medieval Latin West, pointing out also its theological implications, was St. Augustine. In a passage often quoted by medieval authors in their discussions of divine ideas, he writes as follows:

> … in Latin we can call the Ideas “forms” or “species”, in order to appear to translate word for word. But if we call them “reasons”, we depart to be sure from a proper translation - for

> reasons are called “logoi” in Greek, not Ideas - but nevertheless, whoever wants to use this word will not be in conflict with the fact. For Ideas are certain principal, stable and immutable forms or reasons of things. They are not themselves formed, and hence they are eternal and always stand in the same relations, and they are contained in the divine understanding".[19]

As we could see from Boethius' solution, in this way, if Platonic Forms are not universal beings existing in a universal manner, but their universality is due to a universal manner of understanding, we can avoid the contradictions arising from the “naïve” Platonic conception. Nevertheless, placing universal ideas in the divine mind as the archetypes of creation, this conception can still do justice to the Platonic intuition that what accounts for the necessary, universal features of the ephemeral particulars of the visible world is the presence of some universal exemplars in the source of their being. It is precisely in virtue of having some insight into these exemplars themselves that we can have the basis of universal knowledge Plato was looking for. As St. Augustine continues:

> And although they neither arise nor perish, nevertheless everything that is able to arise and perish, and everything that does arise and perish, is said to be formed in accordance with them. Now it is denied that the soul can look upon them, unless it is a rational one, [and even then it can do so] only by that part of itself by which it surpasses [other things] - that is, by its mind and reason, as if by a certain “face”, or by an inner and intelligible “eye”. To be sure, not each and every rational soul in itself, but [only] the one that is holy and pure, that [is the one that] is claimed to be fit for such a vision, that is, the one that keeps that very eye, by which these things are seen, healthy and pure and fair and like the things it means to see. What devout man imbued with true religion, even though he is not yet able to see these things, nevertheless dares to deny, or for that matter fails to profess, that all things that exist, that is, whatever things are contained in their own genus with a certain nature of their own, so that that they might exist, are begotten by God their author, and that by that same author everything that lives is alive, and that the entire safe preservation and the very order of things, by which changing things repeat their temporal courses according to a fixed regimen, are held together and governed by the laws of a supreme God? If this is established and granted, who dares to say that God has set up all things in an irrational manner? Now if it is not correct to say or believe this, it remains that all things are set up by reason, and a man not by the same reason as a horse - for that is absurd to suppose. Therefore, single things are created with their own reasons. But where are we to think these reasons exist, if not in the mind of the creator? For he did not look outside himself, to anything placed [there], in order to set up what he set up. To think that is sacrilege. But if these reasons of all things to be created and [already] created are contained in the divine mind, and [if] there cannot be anything in the divine mind that is not eternal and unchangeable, and [if] Plato calls these principal reasons of things “Ideas”, [then] not only are there Ideas but they are true, because they are eternal and [always] stay the same way, and [are] unchangeable. And whatever exists comes to exist, however it exists, by participation in them. But among the things set up by God, the rational soul surpasses all [others], and is closest to God when it is pure. And to the extent that it clings to God in charity, to that extent, drenched in a certain way and lit up by that intelligible light, it discerns these reasons, not by bodily eyes but by that principal [part] of it by which it surpasses [everything else], that is, by its intelligence. By this vision it becomes most blessed.

> These reasons, as was said, whether it is right to call them Ideas or forms or species or reasons, many are permitted to call [them] whatever they want, but [only] to a very few [is it permitted] to see what is true."[20]

Augustine's conception, then, saves Plato's original intuitions, yet without their inconsistencies, while it also combines his philosophical insights with Christianity. But, as a rule, a really intriguing solution of a philosophical problem usually gives rise to a number of further problems. This solution of the original problem with Plato's Forms is no exception.

Divine Ideas and Divine Simplicity

First of all, it generates a particular ontological/theological problem concerning the relationship between God and His Ideas. For according to the traditional philosophical conception of divine perfection, God's perfection demands that He is absolutely simple, without any composition of any sort of parts.[21] So God and the divine mind are not related to one another as a man and his mind, namely as a substance to one of its several powers, but whatever powers God *has* He *is*. Furthermore, the Divine Ideas themselves cannot be regarded as being somehow the eternal products of the divine mind distinct from the divine mind, and thus from God Himself, for the only eternal being is God, and everything else is His creature. Now, since the Ideas are not creatures, but the archetypes of creatures in God's mind, they cannot be distinct from God. However, as is clear from the passage above, there are several Ideas, and there is only one God. So how can these several Ideas possibly be one and the same God?

Augustine never explicitly raised the problem, but for example Aquinas, who (among others) did, provided the following rather intuitive solution for it.[22] The Divine Ideas are in the Divine Mind as its objects, i.e., as the things understood. But the diversity of the objects of an act of understanding need not diversify the act itself (as when understanding the Pythagorean theorem we understand both squares and triangles). Therefore, it is possible for the self-thinking divine essence to understand itself in a single act of understanding so perfectly that this act of understanding not only understands the divine essence as it is in itself, but also in respect of all possible ways in which it can be imperfectly participated by any finite creature. The cognition of the diversity of these diverse ways of participation accounts for the plurality of divine ideas. But since all these diverse ways are understood in a single eternal act of understanding, which is nothing but the act of divine being, and which in turn is again the divine essence itself, the multiplicity of ideas does not entail any corresponding multiplicity of the divine essence. To be sure, this solution may still give rise to the further questions as to what these diverse ways are, exactly how they are related to the divine essence, and how their diversity is compatible with the unity and simplicity of the ultimate object of divine thought, namely, divine essence itself. In fact, these are questions that were raised and discussed in detail by authors such as Henry of Ghent (c. 1217-1293), Thomas of Sutton (ca. 1250-1315), Duns Scotus (c. 1266-1308) and others.[23]

Illuminationism vs. Abstractionism

Another major issue connected to the doctrine of divine ideas, as should also be clear from the previously quoted passage, was the bundle of epistemological questions involved in Augustine's doctrine of divine

illumination. The doctrine -- according to which the human soul, especially “one that is holy and pure”, obtains a specific supernatural aid in its acts of understanding, by gaining a direct insight into the Divine Ideas themselves -- received philosophical support in terms of a typically Platonic argument in Augustine's *De Libero Arbitrio*.[24] The argument can be reconstructed as follows.

> **The Augustinian Argument for Illumination**.
>
> 1. I can come to know from experience only something that can be found in experience [self-evident]
> 2. Absolute unity cannot be found in experience [assumed]
> 3. Therefore, I cannot come to know absolute unity from experience. [1,2]
> 4. Whatever I know, but I cannot come to know from experience, I came to know from a source that is not in this world of experiences. [self-evident]
> 5. I know absolute unity. [assumed]
> 6. Therefore, I came to know absolute unity from a source that is not in this world of experiences. [3,4,5]
>
> *Proof of 2*. Whatever can be found in experience is some material being, extended in space, and so it has to have a multitude of spatially distinct parts. Therefore, it is many in respect of those parts. But what is many in some respect is not one in that respect, and what is not one in some respect is not absolutely one. Therefore, nothing can be found in experience that is absolutely one, that is, nothing in experience is an absolute unity.
>
> *Proof of 5*. I know that whatever is given in experience has many parts (even if I may not be able to discern those parts by my senses), and so I know that it is not an absolute unity. But I can have this knowledge only if I know absolute unity, namely, something that is not many in any respect, not even in respect of its parts, for, in general, I can know that something is F in a certain respect, and not an F in some other respect, only if I know what it is for something to be an F without any qualification. (For example, I know that the two halves of a body, taken together, are not absolutely two, for taken one by one, they are not absolutely one, since they are also divisible into two halves, etc. But I can know this only because I know that for obtaining absolutely two things [and not just two multitudes of further things], I would have to have two things that in themselves are absolutely one.) Therefore, I know absolute unity.

It is important to notice here that this argument (crucially) assumes that the intellect is passive in acquiring its concepts. According to this assumption, the intellect merely receives the cognition of its objects as it finds them. By contrast, on the Aristotelian conception, the human mind actively processes the information it receives from experience through the senses. So by means of its faculty appropriately called the active or agent intellect, it is able to produce from a limited number of experiences a universal concept equally representing all possible particulars falling under that concept. In his commentary on Aristotle's *De Anima* Aquinas insightfully remarks:

> The reason why Aristotle came to postulate an active intellect was his rejection of Plato's

theory that the essences of sensible things existed apart from matter, in a state of actual intelligibility. For Plato there was clearly no need to posit an active intellect. But Aristotle, who regarded the essences of sensible things as existing in matter with only a potential intelligibility, had to invoke some abstractive principle in the mind itself to render these essences actually intelligible.[25]

On the basis of these and similar considerations, therefore, one may construct a rather plausible Aristotelian counterargument, which is designed to show that we need not necessarily gain our concept of absolute unity from a supernatural source, for it is possible for us to obtain it from experience by means of the active intellect. Of course, similar considerations should apply to other concepts as well.

An Aristotelian-Thomistic counterargument from abstraction.

1. I know from experience everything whose concept my active intellect is able to abstract from experience. [self-evident]
2. But my active intellect is able to abstract from experience the concept of unity, since we all experience each singular thing as being one, distinct from another. [self-evident, common experience][26]
3. Therefore, I know unity from experience by abstraction. [1,2]
4. Whenever I know something from experience by abstraction, I know both the thing whose concept is abstracted and its limiting conditions from which its concept is abstracted. [self-evident]
5. Therefore, I know both unity and its limiting conditions from which its concept is abstracted. [3,4]
6. But whenever I know something and its limiting conditions, and I can conceive of it without its limiting conditions (and this is precisely what happens in abstraction), I can conceive of its absolute, unlimited realization. [self-evident]
7. Therefore, I can conceive of the absolute, unlimited realization of unity, based on the concept of unity I acquired from experience by abstraction. [5,6]
8. Therefore, it is not necessary for me to have a preliminary knowledge of absolute unity before all experience, from a source other than this world of experiences. [7]

To be sure, we should notice here that this argument *does not falsify the doctrine* of illumination. Provided it works, it only *invalidates* the Augustinian-Platonic *argument* for illumination. Furthermore, this is obviously not a sweeping, knock-down refutation of the idea that at least some of our concepts perhaps could not so simply be derived from experience by abstraction; in fact, in the particular case of unity, and in general, in connection with our transcendental notions (i.e., notions that apply in each Aristotelian category, so they *transcend* the limits of each one of them, such as the notions of *being*, *unity*, *goodness*, *truth*, etc.), even the otherwise consistently Aristotelian Aquinas would have a more complicated story to tell.[27] Nevertheless, although Aquinas would still leave some room for illumination in his epistemology, he would provide for illumination an entirely naturalistic interpretation, as far as the acquisition of our intellectual concepts of material things is concerned, by simply identifying it with the “intellectual light in us”, that is, the active intellect, which enables us to acquire these concepts from experience by abstraction.[28] Duns

Scotus, who opposed Aquinas on so many other points, takes basically the same stance on this issue. Other medieval theologians, especially such prominent "Augustinians" as Bonaventure, Matthew of Aquasparta, or Henry of Ghent, would provide greater room for illumination in the form of a direct, specific, supernatural influence needed for human intellectual cognition in this life besides the general divine cooperation needed for the workings of our natural powers, in particular, the abstractive function of the active intellect.[29]

In general, illuminationism and abstractionism were never treated by medieval thinkers as mutually exclusive alternatives. They rather served as the two poles of a balancing act in judging the respective roles of nature and direct divine intervention in human intellectual cognition.[30]

Although Platonism definitely survived throughout the Middle Ages (and beyond), in the guise of the interconnected doctrines of divine ideas, participation, and illumination, there was a quite general Aristotelian consensus,[31] especially after Abelard's time, that the mundane universals of the species and genera of material beings exist as such *in the human mind*, as a result of the mind's abstracting from their individuating conditions. But consensus concerning this much by no means entailed a unanimous agreement on exactly what the universals thus abstracted are, what it is for them to exist in the mind, how they are related to their particulars, what their real foundation in those particulars is, what their role is in the constitution of our universal knowledge, and how they contribute to the encoding and communication of this knowledge in the various human languages. For although the general Aristotelian stance towards universals successfully handles the inconsistencies quite obviously generated by a naïve Platonist ontology, it gives rise precisely to these further problems of its own.

6. Universals According to Abelard's Aristotelian Conception

It was Abelard who first dealt with the problem of universals explicitly in this form. Having relatively easily disposed of putative universal forms as real entities corresponding to Boethius' definition, in his *Logica Ingredientibus* he concludes that given Aristotle's definition of universals in his *On Interpretation* as those things that can be predicated of several things, it is only universal *words* that can be regarded as really existing universals. However, since according to Aristotle's account in the same work, words are meaningful in virtue of signifying concepts in the mind, Abelard soon arrives at the following questions:

1. What is the *common cause* in accordance with which a common name is imposed?
2. What is the understanding's *common conception* of the likeness of things?
3. Is a word called "common" on account of the common cause things agree in, or on account of the common conception, or on account of both together?[32]

These questions open up a new chapter in the history of the problem of universals. For these questions add a new aspect to the bundle of the originally primarily ontological, epistemological, and theological questions constituting the problem, namely, they add a *semantic* aspect. On the Aristotelian conception of universals as universal *predicables*, there obviously *are* universals, namely, our universal words. But the universality

of our words is clearly not dependent on the physical qualities of our articulate sounds, or of the various written marks indicating them, but on their representative function. So, to give an account of the universality of our universal words, we have to be able to tell in virtue of what they have this universal representative function, that is to say, we have to be able to assign a *common cause* by the recognition of which in terms of a *common concept* we can give a *common name* to a *potential infinity of individuals* belonging to the same kind.

But this common cause certainly cannot be a *common thing* in the way Boethius described universal things, for, as we have seen, the assumption of the existence of such a common thing leads to contradictions. To be sure, Abelard also provides a number of further arguments, dealing with several refinements of Boethius' characterization of universals proposed by his contemporaries, such as William of Champeaux, Bernard of Chartres, Clarembald of Arras, Jocelin of Soissons, and Walter of Mortagne – but I cannot go into those details here.[33] The point is that he refutes and rejects all these suggestions to save real universals either as common things, having their own real unity, or as collections of several things, having a merely collective unity. The gist of his arguments against the former view is that the universal thing on that view would have to have its own numerical unity, and therefore, since it constitutes the substance of all its singulars, all these singulars would have to be substantially one and the same thing which would have to have all their contrary properties at the same time, which is impossible. The main thrust of his arguments against the collection-theory is that collections are arbitrary integral wholes of the individuals that make them up, so they simply do not fill the bill of the Porphyrian characterizations of the essential predicables such as genera and species.[34]

So, the common cause of the imposition of universal words cannot be any one thing, or a multitude of things; yet, being a common *cause*, it cannot be nothing. Therefore, this common cause, which Abelard calls the *status*[35] of those things to which it is common, is a cause, but it is a cause which is a non-thing. However strange this may sound, Abelard observes that sometimes we do assign *causes* which are not *things*. For example, when we say “The ship was wrecked because the pilot was absent”, the cause that we assign, namely, that the pilot was absent is not some *thing*, it is rather *how* things were, i.e., the *way* things were, which in this case we signify by the whole proposition “The pilot was absent”.[36] From the point of view of understanding what Abelard's *status* are, it is significant that he assimilates the causal role of *status* as the common cause of imposition to causes that are signified by whole propositions. These *significata* of whole propositions, which in English we may refer to by using the corresponding “that-clauses” (as I did above, referring to the cause of the ship's wreck by the phrase “that the pilot was absent”), and in Latin by an accusative-with-infinitive construction, are what Abelard calls the *dicta* of propositions. These *dicta*, not being identifiable with any single thing, yet, not being nothing, constitute an ontological realm that is completely different from that of ordinary things. But it is also in this realm that Abelard's *common causes of imposition* may find their place.

Abelard says that the common cause of imposition of a universal name has to be something in which things falling under that name agree. For example, the name ‘man’ (in the sense of ‘human being’, and not in the sense of ‘male human being’) is imposed on all humans on account of something in which all humans, as such, agree. But that in which all humans as such agree is that each one of them is a man, that is, each one agrees with all others in their *being a man*. So it is their being human [*esse hominem*] that is the common cause Abelard was looking for, and this is what he calls the *status* of man. The *status* of man is not a thing;

it is not any singular man, for obviously no singular man is common to all men, and it is not a universal man, for there is no such a thing. But *being a man* is common in the required manner (i.e., it is something in which all humans agree), yet it is clearly not a thing. For let us consider the singular propositions ‘Socrates is a man’ [*Socrates est homo*], ‘Plato is a man’ [*Plato est homo*], etc. These signify their *dicta*, namely, Socrates's being a man [*Socratem esse hominem*], and Plato's being a man [*Platonem esse hominem*], etc. But then it is clear that if we abstract from the singular subjects and retain what is common to them all, we can get precisely the *status* in which all these subjects agree, namely, being a man [*esse hominem*]. So the *status*, just like the *dicta* from which they can be obtained, constitute an ontological realm that is entirely different from that of ordinary things.

Still, despite the fact that it clearly has to do something with abstraction, an activity of the mind, Abelard insists that a *status* is not a concept of our mind. The reason for his insistence is that the *status*, being the *common cause* of imposition of a common name, has to be something real, the existence of which is not dependent on the activity of our minds. A *status* is there in the nature of things, regardless of whether we form a mental act whereby we recognize it or not. In fact, for Abelard, a *status* is an object of the divine mind, whereby God preconceives the state of his creation from eternity.[37] A concept, or mental image *of our mind*, however, exists as the object of our mind only insofar as our mind performs the mental act whereby it forms this object. But this object, again, is not a thing, indeed, not any more than any other fictitious object of our minds. However, what distinguishes the *universal concept* from a merely *fictitious object* of our mind is that the former corresponds to a *status* of really existing singular things, whereas the latter does not have anything corresponding to it.

To be sure, there are a number of points left in obscurity by Abelard's discussion concerning the relationships of the items distinguished here. For example, Abelard says that we cannot conceive of the *status*. However, it seems that we can only signify by our words whatever we can conceive. Yet, Abelard insists that besides our concepts, our words *must* signify the *status* themselves.[38] A solution to the problem is only hinted at in Abelard's remark that the names can signify *status*, because“their inventor *meant* to impose them in accordance with certain natures or characteristics of things, even if he did not know how to think out the nature or characteristic of the thing”.[39] So, we may assume that although the inventor of the name does not know the *status*, his vague, “senses-bound” conception, *from which* he takes his word's signification, is directed at the *status*, as to *that which* he *intends* to signify.[40] However, Abelard does not work out this suggestion in any further detail. Again, it is unclear how the *status* is related to the individualized natures of the things that agree in the *status*. If the *status* is what the divine mind conceives of the singulars in abstraction from them, why couldn't the nature itself be conceived in the same way? – after all, the abstract nature would not have to be a thing any more than a *status* is, for its existence would not be *real being*, but merely its *being conceived*. Furthermore, it seems quite plausible that Abelard's *status* could be derived by abstraction from singular *dicta* with the same predicate, as suggested above. But *dicta* are the quite ordinary *significata* of *our* propositions, which Abelard never treats as epistemologically problematic, so why would the *status*, which we could apparently abstract from them, be accessible only to the divine mind?

I'm not suggesting that Abelard could not provide acceptable and coherent answers to these and similar questions and problems.[41] But perhaps these problems also contributed to the fact that by the 13th century his doctrine of *status* was no longer in currency. Another historical factor that may have contributed to the

waning of Abelard's theory was probably the influence of the newly translated Aristotelian writings along with the Arabic commentaries that flooded the Latin West in the second half of the 12th century.

7. Universal Natures in Singular Beings and in Singular Minds

The most important influence in this period from our point of view came from Avicenna's doctrine distinguishing the absolute consideration of a universal nature from what applies to the same nature in the subject in which it exists. The distinction is neatly summarized in the following passage.

> Horsehood, to be sure, has a definition that does not demand universality. Rather it is that to which universality happens. Hence horsehood itself is nothing but horsehood only. For in itself it is neither many nor one, neither is it existent in these sensibles nor in the soul, neither is it any of these things potentially or actually in such a way that this is contained under the definition of horsehood. Rather [in itself it consists] of what is horsehood only.[42]

In his little treatise *On Being and Essence*,Aquinas explains the distinction in greater detail in the following words:

> A nature, however, or essence …can be considered in two ways. First, we can consider it according to its proper notion, and this is its absolute consideration; and in this way nothing is true of it except what pertains to it as such; whence if anything else is attributed to it, that will yield a false attribution. …In the other way [an essence] is considered as it exists in this or that [individual]; and in this way something is predicated of it *per accidens* [non-essentially], on account of that in which it exists, as when we say that a man is white because Socrates is white, although this does not pertain to man as such.
>
> A nature considered in this way, however, has two sorts of existence. It exists in singulars on the one hand, and in the soul on the other, and from each of these [sorts of existence] it acquires accidents. In the singulars, furthermore, the essence has several [acts of] existence according to the multiplicity of singulars. Nevertheless, if we consider the essence in the first, or absolute, sense, none of these pertain to it. For it is false to say that the essence of man, considered absolutely, has existence in this singular, because if existence in this singular pertained to man insofar as he is man, man would never exist, except as this singular. Similarly, if it pertained to man insofar as he is man not to exist in this singular, then the essence would never exist in the singular. But it is true to say that man, but not insofar as he is man, may be in this singular or in that one, or else in the soul. Therefore, the nature of man considered absolutely abstracts from every existence, though it does not exclude any. And the nature thus considered is what is predicated of each individual.[43]

So, a common nature or essence according to its absolute consideration abstracts from all existence, both in the singulars and in the mind. Yet, and this is the important point, it is *the same* nature that informs both the

singulars that have this nature and the minds conceiving of them in terms of this nature. To be sure, this sameness is not numerical sameness, and thus it does not yield numerically one nature. On the contrary, it is the sameness of several, numerically distinct realizations of the same information-content, just like the sameness of a book in its several copies. Just as there is no such a thing as a universal book over and above the singular copies of the same book, so there is no such a thing as a universal nature existing over and above the singular things of the same nature; still, just as it is true to say that the singular copies are the copies of *the same book*, so it is true to say that these singulars are of *the same nature*.

Indeed, this analogy also shows why this conception should be so appealing from the point of view of the original epistemological problem of the possibility of universal knowledge, without entailing the ontological problems of naïve Platonism. For just as we do not need to read all copies of the same book in order to know what we can find on the same page in the next copy (provided it is not a corrupt copy),[44] so we can know what may apply to all singulars of the same nature without having to experience them all. Still, we need not assume that we can have this knowledge only if we can get somehow in a mysterious contact with the universal nature over and above the singulars; all we need is to learn how “to read” the singulars in our experience to discern the “common message”, the universal nature, informing them all, uniformly, yet in their distinct singularity. (Note that “reading the singulars” is not a mere metaphor: this is precisely what geneticists are quite literally doing in the process of gene sequencing, for instance, in the human genome project.) Therefore, the *same nature* is not *the same* in the same way as the *same individual* having this nature is the same as long as it exists. For that *same nature*, insofar as it is regarded as *the same*, does not even exist at all; it is said to be the same only insofar as it is *recognizable as the same*, if we disregard everything that distinguishes its instances in several singulars. (Note here that whoever would want to deny such a *recognizable sameness* in and across several singulars would have to deny that he is able to recognize the same words or the same letters in various sentences; so such a person would not be able to read, write, or even to speak, or understand human speech. But then we shouldn't really worry about such a person in a philosophical debate.)

However, at this point some further questions emerge. If this common nature is *recognizably the same* on account of disregarding its individuating conditions in the singulars, then isn't it the result of abstraction; and if so, isn't it in the abstractive mind as its object? But if it is, then how can Aquinas say that it abstracts *both* from being in the singulars *and* from being in the mind?

Here we should carefully distinguish between what we can say about *the same nature as such*, and what we can say about *the same nature on account of its conditions* as it exists in this or that subject. Again, using our analogy, we can certainly consistently say that the same book in its first edition was 200 pages, whereas in the second only 100, because it was printed on larger pages, but the book itself, as such, is neither 200 nor 100 pages, although it can be either. In the same way, we can consistently say that *the same nature as such* is neither in the singulars nor in the mind, but of course it is only insofar as it is in the mind that it can be *recognizably the same*, on account of the mind's abstraction. Therefore, that it is abstract and is actually recognized as the same in its many instances is something that belongs to the same nature only on account of being conceived by the abstractive mind. This is the reason why the nature is called a *universal concept*, insofar as it is in the mind. Indeed, it is only under this aspect that it is properly called a universal. So, although *that which is predicable* of several singulars is nothing but the common nature as such, considered absolutely, still, *that it is predicable* pertains to the same nature only on account of being conceived by the

abstractive intellect, insofar as it is a concept of the mind.

At any rate, this is how Aquinas solves the paralogism that seems to arise from this account, according to which the true claims that Socrates is a man and man is a species would seem to entail the falsity that Socrates is a species. For if we say that in the proposition ‘Socrates is a man’ the predicate signifies human nature absolutely, but the same nature, on account of its abstract character, is a species, the false conclusion seems inevitable.[45]

However, since the common nature is not a species in its absolute consideration, but only insofar as it is in the mind, the conclusion does not follow. Indeed, this reasoning would be just as invalid as the one trying to prove that this book, pointing to the second edition which is actually 100 pages, is 200 pages, because the same book was 200 hundred pages in its first edition. For just as its being 200 pages belongs to the same book only in its first edition, so its being a species belongs to human nature only as it exists in the mind.

Nevertheless, even though this solution works, the emergence of the paralogism itself, and the complexities involved in explaining it away, show the inherent difficulties of this account. The main difficulty is the trouble of keeping track of what we are talking about, when it becomes crucial to know what pertains to what on account of what; in general, when the conditions of identity and distinction of the items we are talking about become variable and occasionally rather unclear.

Indeed, we can appreciate just how acute these difficulties may become if we survey the items that needed to be distinguished in what may be described as the common conceptual framework of the “realist” *via antiqua*, the “old way” of doing philosophy and theology, before the emergence of the “modern way”, the “nominalist” *via moderna* challenging some fundamental principles of the older framework, resulting mostly from the semantic innovations introduced by William Ockham. The survey of these items and the problems they generate will then allow us to see in greater detail the main motivation for Ockham's innovations.

8. Universals in the *Via Antiqua*

In this framework, we have first of all the universal or common terms of spoken and written languages, which are common on account of being imposed upon universal concepts of the human mind. The concepts themselves are universal on account of being obtained by the activity of the abstractive human mind from experiences of singulars. But the process of concept formation also involves various stages.

In the first place, the sensory information collected by the single senses is distinguished, synthesized, and collated by the higher sensory faculties of the common sense [*sensus communis*] and the so-called cogitative power [*vis cogitativa*], to be stored in sensory memory as *phantasms*, the sensory representations of singulars in their singularity. The active intellect [*intellectus agens*] uses this sensory information to extract its intelligible content and produce the intelligible species [*species intelligibiles*], the universal representations of several individuals in their various degrees of formal unity, disregarding their distinctive features and individuating conditions in the process of abstraction.

The intelligible species are stored in the intellectual memory of the potential intellect [*intellectus possibilis*], which can then use them to form the corresponding concept in an act of thought, for example, in forming a judgment. The intelligible species and the concepts themselves, being formed by individual human minds, are individual in their being, insofar as they pertain to this or that human mind. However, since they are the result of abstraction, in their information content they are universal.

Now insofar as this universal information content is common to all minds that form these concepts at all, and therefore it is a common intelligible content gained by these minds from their objects insofar as they are conceived by these minds in a universal manner, later scholastic thinkers refer to it as the objective concept [*conceptus obiectivus*], distinguishing it from the formal or subjective concepts [*conceptus formales seu subiectivi*], which are the individual acts of individual minds carrying this information (just as the individual copies of a book carry the information content of the book).[46] It is this objective concept that is identified as the universal of the human mind (distinguished from the universals of the divine mind), namely, a species, a genus, a difference, a property, or an accident. (Note that these are only the simple concepts. Complex concepts, such as those corresponding to complex terms and propositions are the products of the potential intellect using these concepts in its further operations.)

These universals, then, as the objective concepts of the mind, would be classified as beings of reason [*entia rationis*], the being of which consists in their being conceived.[47] To be sure, they are not merely fictitious objects, for they are grounded in the nature of things insofar as they carry the universal information content abstracted from the singulars. But then again, the universal information content of the objective concept itself, considered not insofar as it is in the mind as its object, but in itself, disregarding whatever may carry it, is distinguished from its carriers both in the mind and in the ultimate objects of the mind, the singular things, as the nature of these things in its absolute consideration.

However, the common nature as such cannot exist on its own any more than a book could exist without any copies of it or any minds conceiving of it. So, this common nature has real existence only in the singulars, informing them, and giving them their recognizably common characteristics. However, these common characteristics can be recognized as such only by a mind capable of abstracting the common nature from experiencing it in its really existing singular instances. But it is on account of the real existence of these individualized instances in the singulars that the common nature can truly be predicated of the singulars, as long as they are actually informed by these individualized instances.

The items thus distinguished and their interconnections can be represented by the following block-diagram. The dashed frames indicate that the items enclosed by them have a certain reduced ontological status, a "diminished" mode of being, while the boxes partly sharing a side indicate the (possible) partial identities of the items they enclose.[48] The arrows pointing from the common term to the singulars, their individualized natures and items in the mind on this diagram represent semantic relations, which I am going to explain later, in connection with Ockham's innovations. The rest of the arrows indicate the flow of information from experience of singulars through the sensory faculties to the abstractive mind, and to the application of the universal information abstracted by the mind to further singular experiences in acts of judgment.

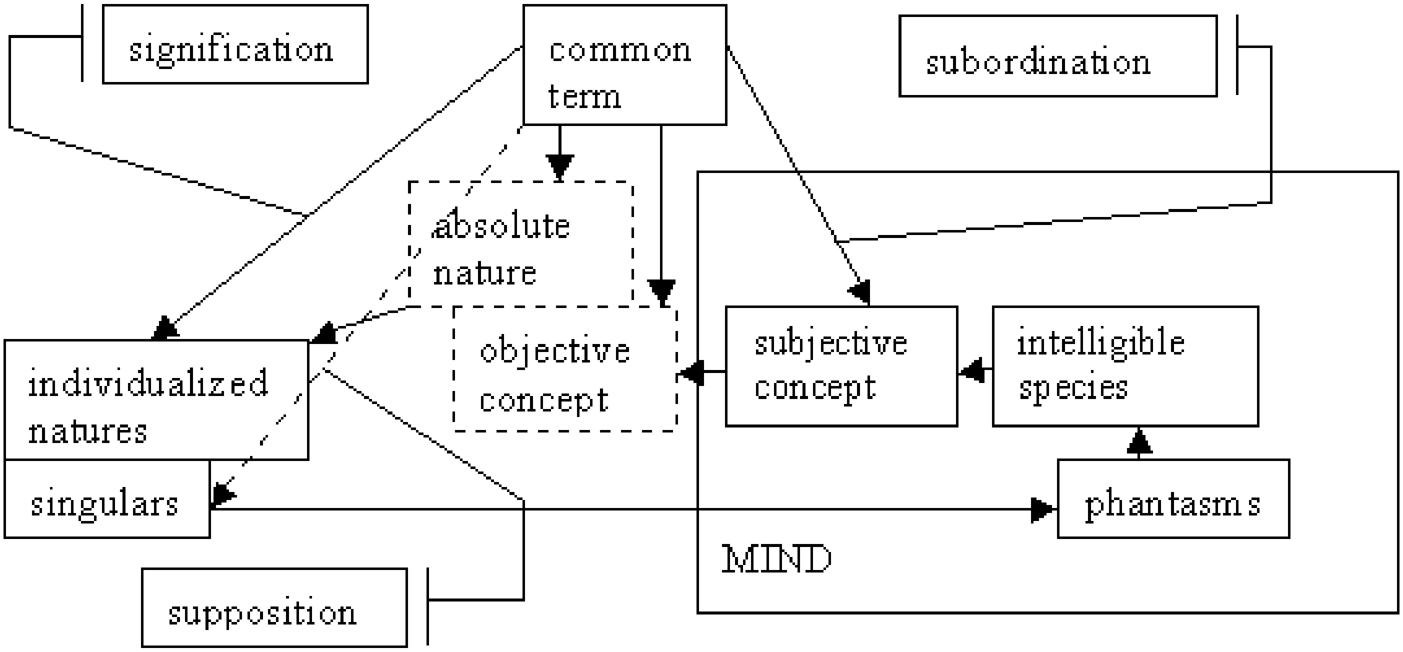


Figure 4. The *via antiqua* conception

Obviously, this is a rather complicated picture. However, its complexity itself should not be regarded as problematic or even surprising, for that matter. After all, this diagram merely summarizes, and distinguishes the main stages of, how the human mind processes the intelligible, universal information received from a multitude of singular experiences, and then again, how it applies this information in classifying further experiences. This process may reasonably be expected to be complex, and should not be expected to involve fewer stages than, e.g., setting up, and retrieving information from, a computer database.

What renders this picture more problematic is rather the difficulties involved in identifying and distinguishing these stages and the corresponding items. Further complications were also generated by the variations in terminology among several authors, and the various criteria of identity and distinctness applied by them in introducing various different notions of identity and distinctness. In fact, many of the great debates of the authors working within this framework can be characterized precisely as disputing the identity or distinctness of the items featured here, or the very criteria of identifying or distinguishing them.

For example, already Abelard raised the question whether the concept or mental image, which we may identify in the diagram as the objective concept of later authors, should be identified with the act of thought, which we may identify as the subjective concept, or perhaps a further act of the mind, called *formatio*, namely, the potential intellect's act of forming the concept, using the intelligible species as the principle of its action. Such distinctions were later on severely criticized by authors such as John Peter Olivi and others, who argued for the elimination of intelligible species, and, in general, of any intermediaries between an act of the intellect and its ultimate objects, the singulars conceived in a universal manner.[49]

Again, looking at the diagram on the side of the singulars, most 13th century authors agreed that what accounts for the specific unity of several individuals of the same species, namely, their specific nature, should be something other than what accounts for their numerical distinctness, namely, their principle of individuation. However, one singular entity in a species of several co-specific individuals has to contain

both the principle of the specific unity of these individuals and its own principle of individuation. Therefore, this singular entity, being a composite at least of its specific nature and its principle of individuation, has to be distinct from its specific nature. At any rate, this is the situation with material substances, whose principle of individuation was held to be their matter. However, based on this reasoning, immaterial substances, such as angels, could not be regarded as numerically distinct on account of their matter, but only on account of their form. But since form is the principle of specific unity, difference in form causes specific diversity. Therefore, on this basis, any two angels had to be regarded as different in species. This conclusion was explicitly drawn by Aquinas and others, but it was rejected by Augustinian theologians, and it was condemned in Paris in 1277.[50]

So, no wonder authors such as Henry of Ghent and Duns Scotus worked out alternative accounts of individuation, introducing not only different principles of individuation, such as the Scotists' famous (or infamous) *haecceity*, but also different criteria of distinctness and identity, such as those grounding Henry of Ghent's *intentional distinction*, or Scotus's *formal distinction*,[51] or even later Suarez' *modal distinction*.[52]

But even further problems arose from considering the identity or distinctness of the individualized natures signified by several common terms in one and the same individual. The metaphysical debate over the real distinction of essence and existence from this point of view is nothing but the issue whether the individualized common nature signified by the definition of a thing is the same as the act of being signified by the verb ‘is’ in the same thing. In fact, the famous problem of the plurality vs. unity of substantial forms may also be regarded as a dispute over whether the common natures signified by the substantial predicates on the Porphyrian tree in the category of substance are distinct or the same in the same individual.[53] Finally, and this appears to be the primary motivation for Ockham's innovations, there was the question whether one has to regard all individualized common natures signified in the same individual by several predicates in the ten Aristotelian categories as distinct from one another. For the affirmative answer would involve commitment to a virtually limitless multiplication of entities.

Indeed, according to Ockham, the *via antiqua* conception would entail that

> a column is to the right by to-the-rightness, God is creating by creation, is good by goodness, just by justice, mighty by might, an accident inheres by inherence, a subject is subjected by subjection, the apt is apt by aptitude, a chimera is nothing by nothingness, someone blind is blind by blindness, a body is mobile by mobility, and so on for other, innumerable cases.[54]

And this is nothing, but “multiplying beings according to the multiplicity of terms... which, however, is erroneous and leads far away from the truth”.[55]

9. Universals in the *Via Moderna*

To be sure, as the very debates within the *via antiqua* framework concerning the identity or non-identity of various items distinguished in that framework indicate, Ockham's charges are not quite justified.[56] After all, several *via antiqua* authors *did* allow the identification of the *significata* of terms belonging to various

categories, so their “multiplication of beings” did not necessarily match the multiplicity of terms. Furthermore, since *via antiqua* authors also distinguished between various modes or senses of being, allowing various sorts of “diminished” kinds of being, such as *beings of reason*, their ontological commitments were certainly not as unambiguous as Ockham would have us believe in this passage. However, if we contrast the diagram of the *via antiqua* framework above with the following schematic representation of the *via moderna* framework introduced by Ockham, we can immediately appreciate the point of Ockham's innovations.

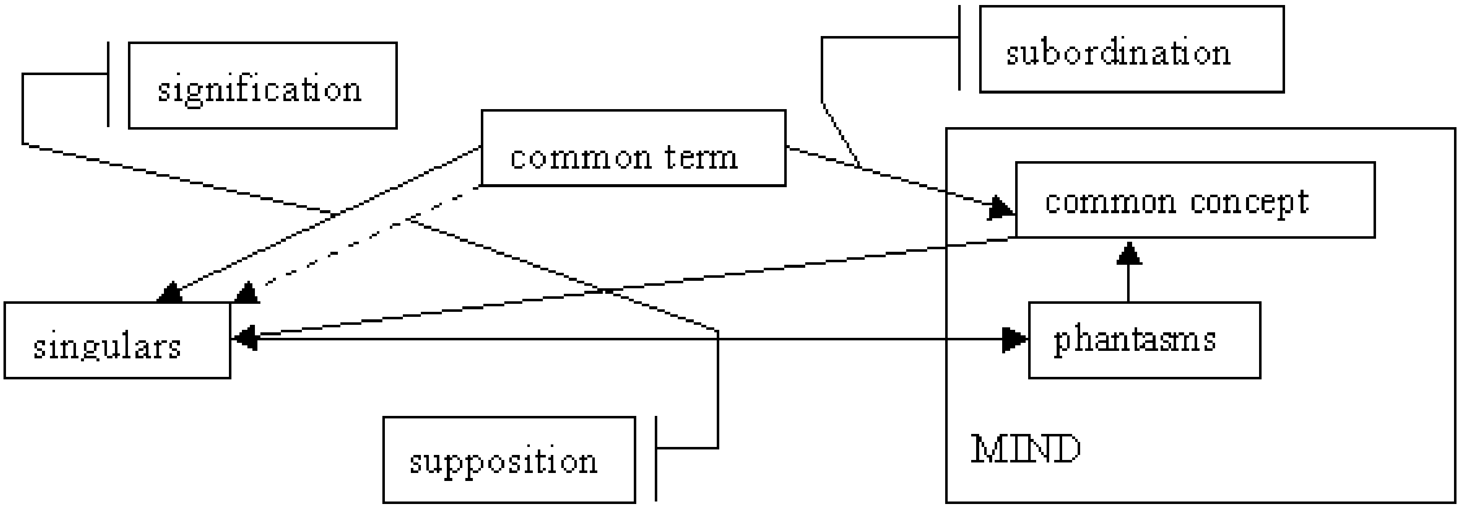


Figure 5. The *via moderna* conception

Without a doubt, it is the captivating simplicity of this picture, especially as compared with the complexity of the *via antiqua* picture, that was the major appeal of the Ockhamist approach. There are fewer items here, equally on the same ontological footing, distinguished from one another in terms of the same unambiguous distinction, the numerical distinction between individual real entities.

To be sure, there still are universals in this picture. But these universals are neither common natures “contracted” to individuals by some really or merely formally distinct principle of individuation, nor some universal objects of the mind, which exist in a “diminished” manner, as *beings of reason*. Ockham's universals, at least in his mature theory,[57] are just our common terms and our common concepts. Our common terms, which are just singular utterances or inscriptions, are common in virtue of being subordinated to our common concepts. Our common concepts, on the other hand, are just singular acts of our singular minds. Their universality consists simply in the universality of their representative function. For example, the common term ‘man’ is a spoken or written universal term of English, because it is subordinated to that concept of our minds by which we conceive of each man indifferently. It is this indifference in its representative function that enables the singular act of my mind to conceive of each man in a universal manner, and the same goes for the singular act of your mind. Accordingly, there is no need to assume that there is anything in the individual humans, distinct from these humans themselves, a common yet individualized nature waiting to be abstracted by the mind. All we need to assume is that two humans are more similar to each other than either of them to a brute animal, and all animals are more similar to each other than any of them to a plant, etc., and that the mind, being able to recognize this similarity, is able to represent the humans by means of a common specific concept, the animals by means of a common generic

concept, all living things by means of a more general generic concept, etc.[58] In this way, then, the common terms subordinated to these concepts need not signify some abstract common nature in the mind, and consequently its individualized instances in the singulars, for they directly signify the singulars themselves, just as they are directly conceived by the universally representative acts of the mind. So, what these common terms signify are just the singulars themselves, which are also the things referred to by these terms when they are used in propositions. Using the customary rendering of the medieval logical terminology, the things ultimately signified by a common term are its *significata*, while the things referred to by the same term when it is used in a proposition are their (personal) *supposita*.[59]

Now if we compare the two diagrams representing the respective conceptions of the two *viae*, we can see just how radically Ockham's innovations changed the character of the semantic relations connecting terms, concepts and things. In both *viae*, common terms are subordinated to common concepts, and it is in virtue of this subordination that they ultimately signify what their concepts represent. In the *via moderna*, a concept is just an act of the mind representing singulars in a more or less indifferent manner, yielding a more or less universal signification for the term. In the *via antiqua*, however, the act of the mind is just one item in a whole series of intermediary representations, distinguished in terms of their different functions in processing universal information, and connected by their common content, ultimately representing the common, yet individualized natures of their singulars.[60] Accordingly, a common term, expressing this common content, is primarily subordinated to the objective concept of the mind. But of course this objective concept is only the common content of the singular representative acts of singular minds, their subjective concepts, formed by means of the intelligible species, abstracted by their active intellects. On the other hand, the objective concept, abstracting from all individuating conditions, expresses only what is common to all singulars, namely, their nature considered absolutely. But this absolutely considered nature is only the common content of what informs each singular of the same nature in its actual real existence. So the term's ultimate *significata* will have to be the individualized natures of the singulars. But these ultimate *significata* may still not be the singulars themselves, namely, when the things informed by these *significata* are not metaphysically simple. In the *via moderna* conception, therefore, the ultimate *significata* of a term are nothing but those singular things that can be the term's *supposita* in various propositions, as a matter of semantics. By contrast, in the *via antiqua* conception, a term's ultimate *significata* may or may not be the same things as the term's (personal) *supposita*, depending on the constitution of these *supposita*, as a matter of metaphysics. The singulars will be the *supposita* of the term when it is used as the subject term of a proposition in which something is predicated about the things informed by these ultimate *significata* (in the case of metaphysically simple entities, the term's *significata* and *supposita* coincide).[61]

Nevertheless, despite the nominalists' charges to the contrary, the *via antiqua* framework, as far as its semantic considerations are concerned, was no more committed to the real distinction of the *significata* and *supposita* of its common terms than the *via moderna* framework was. For if the semantic theory in itself had precluded the identification of these semantic values, then the question of possible identity of these values could not have been meaningfully raised in the first place. Furthermore, in that case such identifications would have been precluded as meaningless even when talking about metaphysically simple entities, such as angels and God, whereas the metaphysical simplicity of these entities was expressed precisely in terms of such identifications. But also in the mundane cases of the *significata* and *supposita* of concrete and abstract universal terms in the nine accidental categories, several *via antiqua* authors argued for the identification of these semantic values both within and across categories. First of all there was Aristotle's authority for the

claim that action and passion are the same motion,[62] so the significata of terms in these two categories could not be regarded as really distinct entities. But several authors also argued for the identification of relations with their foundations, that is to say, for the identity of the significata of relative terms with the significata of terms in the categories quantity and quality. (For example, on this conception, my equality in height to you would be just my height, provided you were of the same height, and not a distinct "equality-thing" somehow attached to my height, caused by our equal heights.)[63]

By contrast, what makes the *via moderna* approach simpler is that it "automatically" achieves such identifications already on the basis of its semantic principles. Since in this approach the *significata* of concrete common terms are just the singulars directly represented by the corresponding concepts, the *significata* and (personal) *supposita* of terms are taken to be the same singulars from the beginning. So these common terms *signify* and *supposit* for the same things either absolutely, provided the term is *absolute*, or in relation to other singulars, provided the term is *connotative*. But even in the case of connotative terms, such as relative terms (in fact, all terms in the nine accidental categories, except for some abstract terms in the category quality, according to Ockham) we do not need to assume the existence of some mysterious relational entities informing singular substances. For example, the term 'father' need not be construed as signifying in me an inherent relation, my fatherhood, somehow connecting me to my son, and suppositing for me on that account in the context of a proposition; rather, it should merely be construed as signifying me in relation to my son, thereby suppositing for me in the context of a proposition, while connoting my son.

10. The Separation of the *viae*, and the Breakdown of Scholastic Discourse in Late-Medieval Philosophy

The appeal of the simplicity of the *via moderna* approach, especially as it was systematically articulated in the works of John Buridan and his students, had a tremendous impact on late-medieval philosophy and theology. To be sure, many late-medieval scholars, who were familiar with both ways, would have shared the sentiment expressed by the remark of Domingo Soto (1494-1560, describing himself as someone who was "born among nominalists and raised by realists")[64] to the effect that whereas the realist doctrine of the *via antiqua* was more difficult to understand, still, the nominalist doctrine of the *via moderna* was more difficult to believe.[65] Nevertheless, the overall simplicity and internal consistency of the nominalist approach were undeniable, gathering a strong following by the 15th century in all major universities of Europe, old and newly established alike.[66]

The resulting separation and the ensuing struggle of the medieval *viae* did not end with the victory of the one over the other. Instead, due to the primarily *semantic* nature of the separation, getting the parties embroiled in increasingly complicated ways of talking past each other, thereby generating an ever growing dissatisfaction, even contempt, in a new, lay, humanist intelligentsia,[67] it ended with the demise of the characteristically medieval conceptual frameworks of both *viae* in the late-medieval and early modern period.

These developments, therefore, also put an end to the specifically *medieval* problem of universals. However, the increasingly rarified late-medieval problem eventually vanished only to give way to several modern

variants of *recognizably the same* problem, which keeps recurring in one form or another in contemporary philosophy as well. Indeed, one may safely assert that as long as there is interest in the questions of how a human language obviously abounding in universal terms can be meaningfully mapped onto a world of singulars, there *is* a problem of universals, regardless of the details of the particular conceptual framework in which the relevant questions are articulated. Clearly, in this sense, the problem of universals is itself a universal, the universal problem of accounting for the relationships between mind, language, and reality.

Bibliography

[Note: This list contains only items referred to in the notes. Excellent, up to date, comprehensive surveys of the medieval problem of universals are provided in Spade 1985 [Other Internet Resources] and Libera 1996.]

Primary Literature

- Aquinas, *Opera Omnia*, R. Busa (ed.), Frommann-Holzboog: Stuttgart-Bad Canstatt, 1980.
- Aristotle, *The Complete Works of Aristotle*, J. Barnes (ed.), Princeton: Princeton University Press, 1984.
- Augustine, *On the Free Choice of the Will*, tr. T. Williams, Hackett: Indianapolis, 1993
- Augustine, *De diversis quaestionibus octoginta tribus*, A. Mutzenbecher (ed.), Corpus Christianorum, Series Latina, vol. 44a, Turnholt: Brepols, 1975
- Avicenna, *Metaphysica* V, 1, S. Van Riet (ed.), 2 vols., E. Peeters: Louvain, and E. J. Brill: Leiden, 1977, 1980.
- Berkeley, G., *A Treatise Concerning the Principles of Human Knowledge*, Hackett: Indianapolis, 1982.
- Bonaventure, *et al*., *De Humanae Cognitionis Ratione: anecdota quaedam Seraphici Doctoris Sancti Bonaventurae et nunnulorum eius discipulorum*, Ad Claras Aquas (Quaracchi): St. Bonaventure, 1883.
- Biel, G., *Collectorium*: Collectorium circa quattuor libros Sententiarum Gabrielis Biel; W. Werbeck, *et al*. (eds.), J.C.B. Mohr (Paul Siebeck): Tübingen, 1973.
- Cajetan, T., *Commentary on Being and Essence*, tr. L. J. Kendzierski and F. C. Wade, Marquette Univ. Press: Milwaukee, 1964.
- Giles of Rome, *In Primum Librum Sententiarum*, Minerva: Frankfurt/Main, 1968, (Venetiis, 1521).
- Henry of Ghent, *Summae quaestionum ordinariarum theologi recepto praeconio solennis Henrici a Gandavo*, (Parisiis In aedibus Jodoci Badii Ascensii 1520), Franciscan Institute: New York, 1953.
- John Duns Scotus, *B. Ioannis Duns Scoti Commentaria Oxoniensia ad IV libros Magistri Sententiarum*, novis curis edidit p. Marianus Fernandez Garcia, Ad Claras Aquas (Quaracchi) prope Florentiam, ex typographia Collegii s. Bonaventurae, 1912-1914.
- John of Salisbury, *Metalogicon*, Clarendon Press: Oxford, 1929.
- John Wyclif, *Tractatus de Universalibus*, I. J. Mueller (ed.), Clarendon Press: Oxford, 1985.
- Plato, *Collected Dialogues*, E. Hamilton and H. Cairns (eds.), Princeton: Princeton University Press, 1982.
- Seneca, *Ad Lucilium Epistulae Morales, Loeb classical library: Latin authors*, Harvard University

Press: Cambridge, Mass., 1962-1967.

- Soto, D., *In Porphyrii Isagogen, Aristotelis Categorias, librosque de Demonstratione, Commentaria*, Venice, ex officina Dominici Guarraei, et Io. Baptistae, fratrum, 1587, reprint, Minerva: Frankfurt, 1967.
- Spade, P. V., (tr.), *Five Texts on the Mediaeval Problem of Universals: Porphyry, Boethius, Abelard, Duns Scotus, Ockham*, Hackett: Indianapolis/Cambridge, 1994.
- Suarez, F., *Disputaciones Metafisicas*, Editorial Gredos: Madrid, 1960.
- Suarez, F., *On the Various Kinds of Distinctions* (*Disputationes metaphysicae, Disputatio VII, de variis distinctionum generibus*), tr. intro. C. Vollert, Marquette University Press: Milwaukee, 1947.
- Thomas of Sutton, *Quodlibeta*, Bayerische Akademie: München, 1969.
- Vincent Ferrer, *Tractatus de Suppositionibus*, Fromman Holzboog: Stuttgart-Bad Cannstatt, 1977.
- Vives, J. L., *Against the Pseudodialecticians*, tr. R. Guerlac, D. Reidel Publishing Company: Dordrecht-Boston-London, 1979.
- William Ockham, *Summa Logicae*, Ph. Boehner, *et al*. (eds.), *Opera Philosophica*, vol. I., The Franciscan Institute, St. Bonaventure, N.Y, 1974.
- William Ockham, *Ordinatio*:Guillelmi de Ockham *Scriptum in librum primum sententiarum ordinatio: distinctiones XIX-XLVIII*, G. I. Etzkorn and F. E. Kelley (eds.), *Opera Theologica*, vol. IV. The Franciscan Institute: St. Bonaventure, N.Y., 1979.

Secondary Literature

- Adams, M. M., 1987, *William Ockham*, 2 vols., University of Notre Dame Press: Notre Dame, IN.
- Ashworth, E. J., 1974, *Language and Logic in the Post-Medieval Period*, Synthese Historical Library, vol. 12, Dordrecht: D. Reidel.
- Callus, D. A., 1967, "Unicity and Plurality of Forms", in *New Catholic Encyclopedia*, McGraw-Hill: New York, 1967-79.
- Gracia, J., 1994, *Individuation in Scholasticism: The Later Middle Ages and the Counter Reformation* (1150-1650), Albany, NY: SUNY Press.
- Gracia, J., 1984, *Introduction to the Problem of Individuation in the Early Middle Ages*, Analytica Series, München and Washington, D.C.: Philosophia Verlag and Catholic University of America Press, 1984; 2nd ed. 1988.
- Henninger, M., 1989, *Relations: Medieval Theories 1250-1325*, Clarendon Press: Oxford.
- Hissette, R., 1977, *Enquête sur les 219 articles condamnés à Paris le 7 Mars 1277*. "Philosophes médiévaux," vol. 22; Louvain: Publications universitaires.
- Hyman, A., and Walsh, J. J., (eds.), 1973, *Philosophy of the Middle Ages*, Hackett: Indianapolis.
- Klima, G., 2000a, "Thomas of Sutton on the Nature of the Intellective Soul and the Thomistic Theory of Being", in J. Aertsen, *et al.* (eds.), *Nach der Verurteilung von 1277. Philosophie und Theologie an der Universität von Paris im letzten Viertel des 13. Jahrhunderts*, Studien und Texte (Miscellanea Mediaevalia 28), Berlin-New York, 2000, pp. 436-455.
- Klima, G., 2000b, "Aquinas on One and Many", *Documenti e Studi sulla Tradizione Filosofica Medievale (An International Journal on the Philosophical Tradition from Late Antiquity to the Late Middle Ages of the Società Internazionale per lo Studio del Medioevo Latino)*, 11 (2000), pp. 195-215.

- Klima, G., 1999, "Ockham's Semantics and Metaphysics of the Categories", in P. V. Spade (ed.), *The Cambridge Companion to Ockham*, Cambridge: Cambridge University Press, 1999, pp. 118–142.
- Klima, G., 1996a, "The Semantic Principles Underlying Saint Thomas Aquinas's Metaphysics of Being", *Medieval Philosophy and Theology*, 5 (1996), pp. 87-141.
- Klima, G., 1993a, "'Socrates est species': Logic, Metaphysics and Psychology in St. Thomas Aquinas' Treatment of a Paralogism", in K. Jacobi (ed.), *Argumentationstheorie: Scholastische Forschungen zu den logischen und semantischen Regeln korrekten Folgerns*, Brill: Leiden, the Netherlands, 1993, pp. 489-504.
- Klima, G., 1993b, "The Changing Role of *Entia Rationis* in Medieval Philosophy: A Comparative Study with a Reconstruction", *Synthese* 96 (1993), pp. 25-59.
- Kretzmann, N., Kenny, A., and Pinborg, J. (eds.), 1982, *The Cambridge History of Later Medieval Philosophy*, Cambridge University Press: Cambridge.
- Libera, A. de, 1996, *La querelle des universeaux: De Platon à la fin du Moyen Age*, Éditions du Seuil: Paris, 1996.
- Pasnau, R., 1999a, "Peter John Olivi", in *The Stanford Encyclopedia of Philosophy*. (Winter 1999 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/entries/olivi/>
- Pasnau, R., 1999b, "Divine Illumination", in *The Stanford Encyclopedia of Philosophy*. (Winter 1999 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/entries/illumination/>
- Pasnau, R., 1997, *Theories of Cognition in the Later Middle Ages*, Cambridge University Press: Cambridge, 1997.
- Read, S. L., 1977, "The Objective Being of Ockham's Ficta", *The Philosophical Quarterly*, 27, pp. 14-31.
- Schmidt, R. W., 1966, *The Domain of Logic according to Saint Thomas Aquinas*, Martinus Nijhoff: The Hague.
- Spade, P. V., 1982, "The Semantics of Terms", in Kretzmann, *et al*., 1982, pp. 188-196.
- Tweedale, M., 1982, "Abelard and the Culmination of Old Logic", in Kretzmann, *et al*., 1982, pp. 142-157.

Other Internet Resources

- John Kilcullen's Medieval Philosophy Teaching Materials
- Klima, G., 1997, Comments on Peter King, "The Failure of Ockham's Nominalism", (Paper read at the Central Division Meeting of the American Philosophical Association, Pittsburgh, PA, April 26, 1997.)
- Klima, G., 1996b "*Nulla virtus cognoscitiva circa proprium obiectum decipitur*" comments on Robert Pasnau, "The Identity of Knower and Known". (Paper read at the Central Division Meetings of the American Philosophical Association, Chicago, April 25, 1996.)
- Paul Spade's Medieval Logic and Philosophy site
- Spade, P. V., 1996a, *Thoughts, Words and Things: An Introduction to Late Mediaeval Logic and Semantic Theory,* Version 1.0, 1996. (PDF file)
- Spade, P. V., 1996b, "Boethius against Universals: The Arguments in the Second Commentary on

Porphyry" (PDF file)

- Spade, P. V., 1985, *A Survey of Medieval Philosophy*, Version 2.0, 1985. (PDF file)

Related Entries

language of thought hypothesis | mental representation | properties | tropes

Gyula Klima
klima@fordham.edu

First published: September 10, 2000
Content last modified: September 10, 2000

Stanford Encyclopedia of Philosophy
Notes to Medieval Theories of Universals

Notes

1. In a somewhat free translation: “I like Plato, but I like the truth even more.” Cf. Aristotle, *The Nicomachean Ethics*, 1096a15, in *The Complete Works of Aristotle*.

2. These are the opposing, yet complementary attitudes (universals come “from above”, for Plato, and “from below”, for Aristotle) that are famously immortalized in the gestures of the two philosophers in the center of Raphael's picture “The School of Athens”.

3. To be sure, there were medieval authors who were skeptical about the reconcilability of Plato and Aristotle on the issue of universals. Cf. the following remark by John of Salisbury (ca. 1120-1180): “Bernard of Chartres and his students worked hard to reconcile Aristotle and Plato. But in my judgment they arrived late, so they strove in vain to reconcile two dead men, who could have reconciled as long as they were alive, but disagreed.” John of Salisbury, *Metalogicon*, bk. 2, c. 17, 875d22-25.

4. Cf.: “In the second place it is clear what those professors mean who say that there are three sorts of universals, namely, before the individual thing, as an idea, in the thing, as a form communicated to many things, and after the thing, as the species or sign of the former.” John Wyclif, *Tractatus de Universalibus*, c. 2, p. 69. Cf. also: “To these two kinds of universals [Plato's and Aristotle's], a third is added, namely, the species in the understanding abstracted from the things is called universal, because it is related to many things, not because it is predicated of many, but because it is similar to many. … And perhaps this is whence the distinction originated that there are three kinds of universals: before the thing, in the thing, and after the thing. For a universal in the first way is before the thing, because it causes things. In the second way it is in the thing, because it is the same essence as the things. In the third way it is after the thing, because it is a species abstracted from the things and caused by them.” Giles of Rome, *In Primum Librum Sententiarum*, 1SN, d. 19, pars 2, q. 1. *Utrum in divinis sit totum universale?*. Concerning alleged “ontological extremities” in mediaeval philosophy see J. A. Trentman's *Introduction* to his edition of Vincent Ferrer's *Tractatus de Suppositionibus*, esp. pp. 20-30.

5. Notable exceptions would be Ockham, and his followers, such as Peter of Ailly, John Gerson, and Gabriel Biel. See Ockham, *Ordinatio* I, d. 35, q. 5, and Biel, *Collectorium* I, d. 35, q. 5. These authors argue that only singular creatures have ideas in the divine mind (which they identify with the singular creatures themselves as they are eternally pre-cognized by God), but neither their species, nor their genera do. To be sure, since created intellects do form universal concepts of singulars, and God does have cognition of those concepts in each singular created mind from eternity, in a roundabout way these authors still concede that there are universals in the divine mind. However, Ockham also claims that in

the divine mind these universals do not function as universal exemplars, as they do in the minds of created agents. But regardless of such details, the point is that even such hardheaded nominalists would not directly go against the authority of St. Augustine on this issue, for which see the section on divine ideas below.

6. Porphyry, *Isagoge*, in Spade 1994 (henceforth, *Five Texts*), p. 1.

7. In fact, by the 12th century the irresistible challenge posed by Porphyry's questions yielded a proliferation of answers that John of Salisbury found pedagogically completely unjustifiable, complaining in his *Metalogicon* that his contemporaries' endless discussions of universals in connection with what was supposed to be an introductory logic text placed “unbearable burden on the tender shoulders of their students” (*onera importabilia teneris auditorum humeris imponunt*). *op. cit.*, bk. 2, c. 19, 877a.

8. As Paul Spade has pointed out in his careful analysis of Boethius' argument, this characterization of what a universal is derives from Porphyry's commentary on Aristotle's *Categories.* See, Spade 1996b.

9. Although Boethius does not explicitly formulate his argument in terms of the distinctness and identity of the acts of being of the entities in question, in the conclusion of his argument, he quite explicitly alludes to the Aristotelian principle of the convertibility of unity and being. See Spade 1996b, p. 22 (12). In any case, this reformulation of Boethius' argument is just my attempt at clarifying what Boethius' may have meant by the requirement of metaphysical “constitutiveness” that accounts for the key motive in the argument which Paul Spade very aptly called the “contagiousness” of the plurality of the particulars.

10. To be sure, Boethius does not explicitly talk about a *collection* here, but this suggestion is certainly very similar to those later collection-theories which Abelard vehemently (and quite effectively) criticized. See *Five Texts*, pp. 35-37. In any case, in Boethius' argument nothing depends on whether the several items demanding a genus are grouped together into a collection, a quasi-entity with its own less-than-numerical quasi-unity. The point is that according to the assumption of this part of the argument, the term ‘genus’ has to stand for a number of distinct entities, no matter whether they are taken to form a collection or not.

11. This summary provides the overall structure of Boethius' argument. In the reconstruction above I did not follow the order of Boethius' actual presentation; I merely summarized what I take to be the gist of Boethius' reasoning. For an intriguing and thoroughgoing discussion of the interpretational possibilities as to the actual “fine-structure” of Boethius' argumentation, see again Spade 1996b.

12. *Five Texts*, p. 23 (21).

13. The same solution recurs, in ever more refined forms, e.g., in Abelard (*Five Texts*, p. 48), John of Salisbury (*Metalogicon*, bk. 2, c. 20, 877c7-878a9), and Aquinas (ST1, q. 85, a. 1, ad 1-um.).

14. To be sure, I may still be mistaken about *what* a man *is*. But that, again, would be a case of forming a judgment, as opposed to simply conceiving of man without any judgment whatsoever. The point is that by an act of simple apprehension I either do conceive of a certain (kind of) thing, or I don't. If I do, then in this act there cannot be falsity unless I make a judgment about the thing conceived; if I don't, then I literally have no idea of the thing, so I will certainly not make a judgment about it, let alone a false one.

15. Note that Berkeley's famous criticism of the Lockean conception of abstraction does not apply here, for abstraction in this sense, as selective intellectual attention, was admitted even by George Berkeley, in a significant insertion in the second edition of the *Introduction* to his *Principles*. See George Berkeley: *A Treatise Concerning the Principles of Human Knowledge*, p. 16.

16. *Five Texts*, p. 25.

17. The tradition goes back to Philo of Alexandria, and the Alexandrian school of Ammonius Saccas. See Spade 1985, p. 67. An excellent, detailed discussion of Greek and Arab Neo-Platonism is provided in Libera 1996. Interestingly, Aquinas, following Averroes, attributes this theory to Aristotle. Cf. 1SN, d. 36, q. 2, a. 1, ad 1-um. Domingo Soto, besides referring to the *locus classicus* from Augustine quoted below, indicates that he also found the same doctrine in the "through and through Platonic" (*ubique Platonicus*) Seneca. Cf. Soto, D., *In Porphyrii Isagogen, Aristotelis Categorias, librosque de Demonstratione, Commentaria*, p. 30 I. For Seneca, see, e.g., *Ad Lucilium Epistulae Morales*, lxx.7; lxv. 4-7. Seneca is also quoted by Ockham and Biel in the places referred to in n. 5 above.

18. *Five Texts*, p. 25.

19. *On Eighty-Three Different Questions*, q. 46, 2, translated by P. V. Spade (in Spade 1985, p. 383) from Augustine, *De diversis quaestionibus octoginta tribus*.

20. *Ibid*.

21. Quite significantly, Aquinas quotes Augustine as his authority for this position: "Augustine says (De Trin. iv, 6,7): 'God is truly and absolutely simple.'" ST1, q. 3, a. 7, sc. (See the whole question for Aquinas' detailed arguments for God's simplicity.)

22. ST1, q. 15, a. 2.

23. Cf.: Henry of Ghent, *Summa Quaestionum Ordinariarum*, 2, a. 65, q. 5; John Duns Scotus, *Op. Oxon*. lb. 1, d. 35, q. *unica*. Thomas of Sutton, *Quodlibeta*, IV, q. 5. For Aquinas' detailed discussion of the issue one should check 1SN, d. 36, q. 2, aa. 1-3.

24. *On the Free Choice of the Will*, II, 8; cf. *Phaedo*, 73c-75c, in Plato, *Collected Dialogues*.

25. *In De Anima*, bk. 3, lc. 10.

26. Note that this premise merely assumes the abstraction of the concept of some sort of unity, found, for example, in the experience of single bodies (which really *is* common experience), and not the formation of the concept of absolute unity, which would require (at least) the further steps indicated below. Of course, the premise also presumes the Aristotelian conception that the intellective soul has an active intellect, the active faculty enabling it to perform abstraction.

27. See Klima 2000b.

28. ST1, q. 84, a. 5.

29. An excellent collection of relevant texts can be found in Bonaventure, *et al*., *De Humanae Cognitionis Ratione: anecdota quaedam Seraphici Doctoris Sancti Bonaventurae et nunnulorum eius discipulorum*.

30. For more on this issue see Pasnau 1999b.

31. As John of Salisbury writes: "…everybody follows Aristotle" (…*omnes Aristotelem profitentur*). *Metalogicon*, bk. 2, c. 19, 877a18.

32. *Five Texts*, p. 41 (88).

33. Cf. Spade 1985, c. 40, where P. V. Spade carefully identifies Abelard's opponents on the basis of John of Salisbury's encyclopedic account of the current views on universals of his time.

34. No wonder that in modern philosophies of language, mostly inspired by the "collection-theorist" view of quantification theory, we have the persistent problem of providing a principled distinction between essential and non-essential predicates.

35. The term *status* (the plural form of which is also *status*, pronounced with a long *u*) designates just a state of a thing, the way the thing is. It is this simple, intuitive idea that gains some further, technical significance in the context of Abelard's theory.

36. Cf. Aristotle, *Physics*, 2, c. 3, 195 a13-14. Note that the *Physics* was not yet available in Abelard's time. Abelard's own examples include being flogged for not appearing in court, being hanged for stealing, dying because of not eating, and being damned for not acting rightly. In any case, Abelard's point clearly is that he takes the causal relation to hold between the *significata* of whole propositions, regardless of the particular content of these propositions, or the particular kind of causality involved.

37. This will immediately raise the problem whether the *status* of created things are there before their

creation, and if not (since nothing other than God can be coeternal with God), then whether God's providence of created things before their creation was empty. Abelard explicitly raises and answers this problem. Cf. *Five Texts*, pp. 49-50 (135)-(140).

38. *Five Texts*, p. 47 (123). Indeed, according to the immediately following paragraph (124), the *status* has a "greater force" in determining the community of universal names; so it may even be *primarily* signified by a common term.

39. *Five Texts*, p. 46 (116).

40. In fact, I think this may be an early anticipation of the distinction that appears in Aquinas, who distinguishes between that *from which* a name is imposed, and that *on which* the name is imposed, what is *intended* to be signified. For more on this distinction, see Klima 1996a.

41. For a more detailed discussion of the philosophical issues raised by Abelard's theory, see Tweedale 1982, and Spade 1985, c. 40.

42. Avicenna: *Metaphysica* V, 1, at II, p. 228 lines 31-36; 1508 ed., fol. 86va; tr. in Spade 1985, c. 21, p. 461.

43. *De ente et essentia*, c. 2 (in *Opera Omnia*)

44. In this connection the analogy perfectly applies to what Aristotle and Aquinas had to say about monstrous births, and what we know about genetic errors.

45. Cf. Klima 1993a.

46. "Human nature has in the intellect existence abstracted from all individuals, and thus it is related uniformly to all individuals that exist outside the soul, as it is equally similar to all of them, and it leads to knowledge of all insofar as they are men. Since the nature in the intellect has this relation to each individual, the intellect invents the notion of species and attributes it to itself. Hence, the Commentator, in *De Anima* I, com. 8, says, "The intellect is what makes universality in things," and Avicenna says the same in his *Metaphysicae* V, cap. 2. Although this nature understood in the intellect has the notion of a universal in relation to things outside the soul (because it is one likeness of them all), as the nature has existence in this intellect or in that one, it is a certain particular understood species." Aquinas, *De ente et essentia*, c. 3 (in *Opera Omnia*). Cf. Suarez, F., *Disputaciones Metafisicas*, pp. 360-361. For a somewhat different interpretation of the same distinction, however, cf. also Cajetan, *Commentary on Being and Essence*, pp. 67-71, 121-124.

47. Cf. Klima 1993b and Schmidt 1966.

48. For a more detailed, systematic discussion of the issue of different "degrees of being" and different "degrees of unity" this conception entails see Klima 1996a and 2000b.

49. See more on Olivi in Pasnau 1999a. For more discussion of the issue of intermediaries in cognition in medieval philosophy, especially in Aquinas, its relation to skepticism, and whether it should be construed in terms of the modern "direct realist" vs. "representationalist" distinction, see Klima 1996b. (See also *Appendix A* of Pasnau 1997).

50. Cf. For an excellent, brief "textbook-treatment" of, and relevant excerpts from, the 1277 condemnations see "The Condemnation of 1277", in A. Hyman and J. J. Walsh, (eds.), *Philosophy of the Middle Ages*, which also provides further scholarly references. For a thoroughgoing discussion, see Hissette 1977.

51. For Henry's view, as opposed and criticized by the Thomist theologian Thomas of Sutton, see Klima 2000a. For the same, as criticized by Duns Scotus, see *Five Texts*, pp. 69-71. For in-depth contemporary discussions of the problem of individuation in the Middle Ages, see Gracia 1994. For a comprehensive account of the early history of the problem, see Gracia 1984.

52. Cf. Suarez, F., *On the Various Kinds of Distinctions*.

53. Cf. Callus 1967.

54. Ockham, W. *Summa Logicae*, part 1, c. 51, p. 169.

55. *Ibid*., p. 171, where Ockham explicitly claims that this is the root (*radix*) of the errors of the moderns.

56. I argue for this claim in detail in Klima 1999.

57. For a recent summary of the development of Ockham's views on this matter see Pasnau 1997, pp. 277-289, esp. p. 278, n. 47, where the author provides further references to the existing secondary literature. A particularly illuminating account of Ockham's early *fictum* theory is provided by Read 1977.

58. To be sure, this conception of indifference of representation based on essential similarity of the represented things raises a number of further questions which seriously challenge the successfulness of Ockham's semantic program on the level of how the fundamental semantic relation between concepts and things is established. A detailed discussion of this issue can be found in Klima 1997 [Klima 1997]. Cf. also Adams 1987, I, c.3, pp. 121-133.

59. In medieval logic (regardless of many disagreements over the details between individual authors), *personal supposition* (when a term is used to refer in a proposition to something of which it can truly be

predicated with the mediation of the copula of the proposition in question) was commonly distinguished from *simple supposition* (when the term is used to refer to its *immediate significate*), and *material supposition* (when the term is used to refer to itself or other tokens of its type). For a brief, summary account of the medieval theories of the properties of terms, see Spade 1982. For *much* more, see Spade 1996a.

60. Again, this need not necessarily be regarded as a flaw of this theory. After all, in our age of satellite uplinks and computer networks we should not find the idea of *the same* information content (which is nothing but the *objective concept* in the scholastic jargon) being realized in several, physically radically different media. (Indeed, *Napster* would not be in trouble, if the recording companies did not think that the music downloaded is the same that they had recorded.)

61. For detailed, systematic comparisons of the different semantic principles of the two *viae*, see Klima 1993b and 1999.

62. Cf. Aristotle, *Physics*, III, 3, 202a15-202b29; *Metaphysics*, XI, 9, 1066a17-34.

63. An excellent survey of the medieval theories of relations is provided by Henninger 1989.

64. Soto, D., *In Porphyrii Isagogen, Aristotelis Categorias, librosque de Demonstratione, Commentaria*, p. 28H.

65. *Ibid*.

66. For an excellent historical survey of these developments see Ashworth 1974.

67. Documents of the typical humanistic reaction are provided with Rita Guerlac's excellent introduction in Vives, J. L., *Against the Pseudodialecticians*.

Gyula Klima
klima@murray.fordham.edu

First posted: September 10, 2000
Last modified: September 10, 2000

The Language of Thought Hypothesis

The Language of Thought Hypothesis (LOTH) postulates that thought and thinking take place in a mental language. This language consists of a system of representations that is physically realized in the brain of thinkers and has a combinatorial syntax (and semantics) such that operations on representations are causally sensitive only to the syntactic properties of representations. According to LOTH, thought is, roughly, the tokening of a representation that has a syntactic (constituent) structure with an appropriate semantics. Thinking thus consists in syntactic operations defined over such representations. Most of the arguments for LOTH derive their strength from their ability to explain certain empirical phenomena like productivity and systematicity of thought and thinking.

What is the Language of Thought Hypothesis?

LOTH is an empirical thesis about the nature of thought and thinking. According to LOTH, thought and thinking are done in a mental language, i.e. in a symbolic system physically realized in the brain of the relevant organisms. In formulating LOTH, philosophers have in mind primarily the variety of thoughts known as 'propositional attitudes'. Propositional attitudes are the thoughts described by such sentence forms as '*S* believes that *P*', '*S* hopes that *P*', '*S* desires that *P*', etc., where '*S*' refers to the subject of the attitude, '*P*' is any sentence, and 'that *P*' refers to the proposition that is the object of the attitude. If we let '*A*' stand for such attitude verbs as 'believe', 'desire', 'hope', 'intend', 'think', etc., then the propositional attitude statements all have the form: *S A*s that *P*.

LOTH can now be formulated more exactly as a hypothesis about the nature of propositional attitudes. It can be characterized as the conjunction of the following three theses (A), (B) and (C):

(A) Representational Theory of Mind (RTM): (cf. Field 1978: 37, Fodor 1987: 17)

> (1) Representational Theory of Thought: For each propositional attitude *A*, there is a unique and distinct (i.e. dedicated)[1] psychological relation *R*, and for all propositions *P* and subjects *S*, *S A*s that *P* if and only if there is a mental representation #*P*# such that
>
>> (a) *S* bears *R* to #*P*#, and
>> (b) #*P*# means that *P*.
>
> (2) Representational Theory of Thinking: Mental processes, thinking in particular, consists of causal sequences of tokenings of mental representations.

(B) Mental representations, which, as per (A1), constitute the direct "objects" of propositional attitudes, belong to a representational or symbolic *system* which is such that (cf. Fodor and Pylyshyn 1988:12-3)

> (1) representations of the system have a combinatorial syntax and semantics: structurally complex (molecular) representations are systematically built up out of structurally simple (atomic) constituents, and the semantic content of a molecular representation is a function of the semantic content of its atomic constituents together with its syntactic/formal structure, *and*
>
> (2) the operations on representations (constituting, as per (A2), the domain of mental processes, thinking) are causally sensitive to the syntactic/formal structure of representations defined by this combinatorial syntax.

(C) Functionalist Materialism. Mental representations so characterized are, at some suitable level, functionally characterizable entities that are realized by the physical properties of the subject having

propositional attitudes (if the subject is an organism, then the realizing properties are presumably the neurophysiological properties in the brain or the central nervous system of the organism).

The relation *R* in (A1), when RTM is combined with (B), is meant to be understood as a *computational/functional* relation. The idea is that each attitude is identified with a characteristic computational/functional role played by the mental sentence that is the direct "object" of that kind of attitude. (Scare quotes are necessary because it is more appropriate to reserve ‘object’ for a proposition as we have done above, but as long as we keep this in mind, it is harmless to use it in this way for LOT sentences.) For instance, what makes a certain mental sentence an (occurrent) belief might be that it is characteristically the output of perceptual output systems and input to an inferential system that interacts decision-theoretically with desires to produce further sentences or actions. Or equivalently, we may think of belief sentences as those that are accessible only to certain sorts of computational operations appropriate for beliefs, but not to others. Similarly, desire-sentences (and sentences for other attitudes) may be characterized by a different set of operations that define a characteristic computational role for them. In the literature it is customary to use the metaphor of a "belief-box" (cf. Schiffer 1981) as a blanket term to cover whatever specific computational role belief sentences turn out to have in the mental economy of their possessors. (Similarly for "desire-box", etc.)

The Language of Thought Hypothesis is so-called because of (B): token mental representations are like sentences in a language in that they have a syntactically and semantically regimented constituent structure. Put differently, mental representations that are the direct "objects" of attitudes are structurally complex symbols whose complexity lends itself to a syntactic and semantic analysis. This is also why the LOT is sometimes called *Mentalese*.

It is (B2) that makes LOTH a species of the so-called Computational Theory of Mind (CTM). This is why LOTH is sometimes called the Computational/Representational Theory of Mind or Thought (CRTM/CRTT) (cf. Rey 1991, 1997). Indeed, LOTH seems to be the most natural product when RTM is combined with a view that would treat mental processes or thinking as computational when computation is understood traditionally or *classically* (this is a recent term emphasizing the contrast with connectionist processing, which we will discuss later).

When someone believes that *P*, there is a trivial sense in which the immediate "object" of her belief, what she believes, can be said to be a complex symbol, according to LOTH, a sentence in her LOT physically realized in the neurophysiology of her brain, that has both syntactic structure and a semantic content, namely the proposition that *P*. So, contrary to the orthodox view that takes the belief relation as a dyadic relation between an agent and a proposition, LOTH takes it to be a triadic relation among an agent, a Mentalese sentence, and a proposition. The Mentalese sentence can then be said to have the proposition as its semantic/intentional content. It is only in this (perhaps indirect) sense can it be said that what is believed is a proposition, and thus the *object* of the attitude in more common philosophical parlance.

This triadic view seems to have an advantage over the orthodox dyadic view in that it is a puzzle in the dyadic view how what are thought to be purely physical organisms can stand in direct relation to abstract objects like propositions in such a way as to influence their causal powers. According to **folk psychology**,

it is because those states have the propositional content they do that they have the causal powers they do. LOTH makes this relatively non-mysterious by introducing a physical intermediary that is capable of having the relevant causal powers in virtue of its syntactic structure that encodes its semantic content. Another advantage of this is that the thought processes can be causally guided by the syntactic forms of the sentences in a way that respect their semantic contents. This is the virtue of (B) to which we'll come back below. Mainly because of these features, LOTH is said to scientifically vindicate folk psychology if it turns out to be true.

Status of LOTH

LOTH has primarily been advanced as an *empirical* thesis (although some have argued for the truth of LOTH on a priori or conceptual grounds, given the natural conceptual contours of folk psychology -- see Davies 1989, 1991; Lycan 1993; Rey 1995). It is not meant to be taken as an analysis of what the folk *mean* (or, for that matter, what the scientists ought to mean) when they talk about various propositional attitudes and their role in thinking. In this regard, LOT theorists typically view themselves as engaged in some sort of a proto-science, or at least in some empirical research program continuous with scientific psychology or more generally with empirical inquiry. Indeed, as we will see in more detail below, when Jerry Fodor first explicitly articulated and elaborated LOTH in some considerable detail in his (1975), he basically defended it on the ground that it was assumed by our best scientific theories or models in cognitive psychology and psycholinguistics. This empirical status accorded to LOTH should be kept firmly in mind when assessing its plausibility and especially its prospects in the light of new evidence and developments in scientific psychology.

When viewed this way, LOTH is not, strictly speaking, committed to preserving the folk taxonomy of the mental states in any very exact way. Notions like belief, desire, hope, fear, etc. are folk notions and, as such, it may not be utterly plausible to expect (eliminativist arguments aside) that a scientific psychology will preserve the exact contours of these concepts. On the contrary, there is every reason to believe that scientific counterparts of these notions will carve the mental space somewhat differently. For instance, it has been noted that the folk notion of belief harbors many distinctions. It is noted for example that it has both a dispositional and an occurrent sense. In the occurrent sense, it seems to mean something like consciously entertaining and accepting a thought (proposition) as true. There is quite a bit of literature and controversy on the dispositional sense.[2] Beliefs are also capable of being explicitly stored in long term memory as opposed to being merely dispositional or tacit. Compare, for instance: I believe that there was a big surprise party for my 24th birthday vs. I have always believed that lions don't eat their food with forks and knives, or that 13652/4=3413, even though until now these latter two thoughts had never occurred to me. There is furthermore the issue of degree of belief: while I may believe that George will come to dinner with his new girlfriend even though I wouldn't bet on it, you, thinking that you know him better than I do, may nevertheless go to the wall for it. It is unlikely that there will be one single construct of scientific psychology that will exactly correspond to the folk notion of belief in all these ways.

For LOTH to vindicate folk psychology it is sufficient that a scientific psychology with a LOT architecture come up with scientifically grounded psychological states that are recognizably like the

propositional attitudes of folk psychology, and that play more or less similar roles in psychological explanations.[3]

Scope of LOTH

LOTH is an hypothesis about the nature of thought and thinking with propositional content. As such, it may or may not be applicable to other aspects of mental life. Officially, it is silent about the nature of some mental phenomena such as experience, **qualia**,[4] sensory processes, mental images, visual and auditory imagination, sensory memory, perceptual pattern-recognition capacities, dreaming, hallucinating, etc. To be sure, many LOT theorists hold views about these aspects of mental life that make it seem that they are also to be explained by something similar to LOTH.[5]

For instance, Fodor seems to think that many modular input systems (Fodor 1983) have their own LOT to the extent to which they can be explained in representational and computational terms. Indeed, many contemporary psychological models treat perceptual input systems in just these terms.[6] There is indeed some evidence that this kind of treatment is appropriate for many perceptual processes. But it is to be kept in mind that a system may employ representations and be computational without necessarily satisfying any or both of the clauses in (B) above in any full-fledged way. Just think of finite automata theory where there are plenty of examples of a computational process defined over states or symbols which lack full-blown syntactic and/or semantic structural complexity. Whether sensory or perceptual processes are to be treated within the framework of full-blown LOTH is again an open empirical question. It may well be that the answer to this question is affirmative. If so, there may be more than one LOT realized in different subsystems or mechanisms in the mind/brain. So LOTH is not committed to there being a single representational system realized in the brain, nor is it committed to the claim that all mental representations are complex or language-like, nor would it be falsified if it turns out that most aspects of mental life other than the ones involving propositional attitudes don't require a LOT.

Similarly, there is strong evidence that the mind also exploits an image-like representational medium for certain kinds of mental tasks.[7] LOTH is non-committal about the existence of an image-like representational system for many mental tasks other than the ones involving propositional attitudes. But it *is* committed to the claim that propositional thought and thinking cannot be successfully accounted for in its entirety in purely imagistic terms. It claims that a combinatorial sentential syntax is necessary for propositional attitudes and a purely imagistic medium is not an adequate medium to capture that.[8]

There are in fact some interesting and difficult issues surrounding these claims. The adequacy of an imagistic system seems to turn on the nature of syntax at the *sentential* level. For instance, Fodor, in Chapter 4 of his (1975) book, allows that many lexical items in one's LOT may be image-like; he introduces the notion of a *mental image/picture under description* to avoid some obvious inadequacies of pictures (e.g., what makes a picture a picture of a fat woman rather than a pregnant one, or vice versa, etc.). This is an attempt to combine discursive and imagistic representational elements at the *lexical* level. There may even be a well defined sense in which pictures can be combined to produce structurally complex pictures (as in British Empiricism: image-like simple ideas are combined to produce complex

ideas, e.g., the idea of a unicorn) But what is absolutely essential for LOTH, and what Fodor insists on, is the claim that there is no adequate way in which a purely image-like system can capture what is involved in making judgments, i.e., in judging *propositions* to be true. This seems to require a discursive syntactic approach at the sentential level. The general problem here is the inadequacy of pictures or image-like representations to express propositions. I can judge that the blue box is on top of the red one without judging that the red box is under the blue one. I can judge that Mary kisses John without judging that John kisses Mary, and so on for indefinitely many such cases, concrete as well as abstract. It is hard to see how images or pictures can do that without using any syntactic structure or discursive elements, to say nothing of judging, e.g., conditionals, disjunctive or negative propositions, quantifications, negative existentials, etc.[9]

Moreover, there are difficulties with imagistic representations arising from demands on *processing* representations . As we will see below, (B2) turns out to provide the foundations for one of the most important arguments for LOTH: it makes it possible to mechanize thinking understood as a semantically coherent thought process, which, as per (A2), consists of a causal sequence of tokenings of mental representations. It is not clear, however, how an equivalent of (B2) could be provided for images or pictures in order to accommodate operations defined over them, even if something like an equivalent of (B1) could be given. On the other hand, there are truly promising attempts to *integrate* discursive symbolic theorem-proving with reasoning with image-like symbols. They achieve impressive efficiency in theorem-proving or in any deductive process defined over the expressions of such an integrated system. Such attempts, if they prove to be generalizable to psychological theorizing, are by no means threats to LOTH; on the contrary, such systems have every features to make them a species of a LOT system: they satisfy (B).[10]

Nativism and LOTH

In the book (1975) in which Fodor introduced the LOTH, he also argued that all concepts are innate. As a result, the connection between LOTH and an implausibly strong version of conceptual nativism looked very much internal. This historical coincidence has led some people to think that LOTH is essentially committed to a very strong version of nativism, so strong in fact that it seems to make a *reductio* of itself (see, for instance, P.S. Churchland 1986, Putnam 1988, Clark 1994). The gist of his argument was that since learning concepts is a form of hypothesis formation and confirmation, it requires a system of mental representations in which formation and confirmation of hypotheses are to be carried out, but then there is a non-trivial sense in which one already has (albeit potentially) the resources to express the extension of the concepts to be learned.

However, it should be emphasized that LOTH is not committed to such a strong version of nativism, especially about concepts. It is certainly plausible to assume that LOTH will turn out to have some empirically (as well as theoretically/a priori) motivated nativist commitments about the structural organization and dynamic management of the entire representational system. But this much is to be expected especially in the light of recent empirical findings and trends. This, however, does not constitutes a *reductio*. It is an open empirical question how much nativism is true about concepts, and

LOTH should be so taken as to be capable of accommodating whatever turns out to be true in this matter. LOTH, therefore, when properly conceived, is independent of any specific proposal about conceptual nativism.[11]

Naturalism and LOTH

One of the most attractive features of LOTH is that it is a central component of an ongoing research program in philosophy of psychology to naturalize the mind, to give a theoretical framework in which the mind could naturally be seen as part of the physical world without postulating irreducibly psychic entities, events, processes or properties. Fodor, the most ardent defender of LOTH, once identified the major mysteries in philosophy of mind thus:

> How could anything material have conscious states? How could anything material have semantical properties? How could anything material be rational? (where this means something like: how could the state transitions of a physical system preserve semantical properties?). (1991: 285, Reply to Devitt)

LOTH is a full-blown attempt to give a naturalist answer to the third question, an attempt to solve at least part of the problem underlying the second one, and is almost completely silent about the first.[12]

According to RTM, propositional attitudes are relations to meaningful mental representations whose tokenings constitute the domain of thinking. This much can, in principle, be granted by an intentional realist who would nevertheless reject LOTH. Indeed, there are plenty of theorists who accept RTM in some suitable form (and also happily accept (C) in many cases) but reject LOTH either by explicitly rejecting (B) or simply by remaining neutral about it. Among some of the prominent people who choose the former option are Searle (1984, 1990, 1992), Stalnaker (1984), Lewis (1972), Barwise and Perry (1983).[13] Some who opt for the latter include Loar (1982a, 1982b), Dretske (1981); Armstrong (1980), and many contemporary functionalists.[14]

But RTM per se doesn't so much propose a naturalistic solution to intentionality and mechanization of thinking as simply assert a framework to emphasize intentional realism and, perhaps, with (C), a declaration of a commitment to naturalism or physicalism at best. How, then, is the addition of (B) supposed to help? Let us first try to see in a bit more detail what the problem is supposed to be in the first place to which (B) is proposed as a solution. So let us start by reflecting on thinking and see what it is about thinking that makes it a mystery in Fodor's list. This will give rise to one of the most powerful (albeit still nondemonstrative) arguments for LOTH.

The Problem of Thinking

RTM's second clause (A2), in effect, says that thinking is at least the tokenings of states that are (a) intentional (i.e. have representational/propositional content) and (b) causally connected. But, surely,

thinking is more. There could be a causally connected series of intentional states that makes no sense at all. Thinking, therefore, is causally proceeding from states to states that would make semantic sense: the transitions among states must preserve some of their semantic properties to count as thinking. In the ideal case, this property would be the truth value of the states. But in most cases, any interesting intentional property like warrantedness, degree of confirmation, semantic coherence given a certain practical context like satisfaction of goals in a specific context, etc. would do. In general, it is hard to spell out what this requirement of "making sense" comes to. The intuitive idea, however, should be clear. Thinking is not proceeding from thoughts to thoughts in arbitrary fashion: thoughts that are causally connected are in some fashion semantically connected too. If this were not so, there would be little point and gain in thinking. Thinking couldn't serve any useful purpose. Call this general phenomenon, then, the *semantic coherence* of causally connected thought processes. LOTH is offered as a solution to this puzzle: how is thinking, conceived this way, physically possible? This is the problem of thinking, thus the problem of mechanization of rationality in Fodor's version. How does LOTH propose to solve this problem and bring us one big step closer to the naturalization of the mind?

Syntactic Engine Driving a Semantic Engine: Computation

The two most important achievements of 20th century that are at the foundations of LOTH as well as most of modern Artificial Intelligence (AI) research and the so-called information processing approaches to cognition (practically almost all of contemporary cognitive psychology) are (i) the developments in modern symbolic (formal) logic, and (ii) **Alan Turing**'s idea of a Turing Machine and Turing computability. It is putting these two ideas together that gives LOTH its enormous explanatory power within a naturalistic framework. Modern logic showed that most of deductive reasoning can be formalized, i.e. most semantic relations among symbols can be entirely captured by the symbols' formal/syntactic properties and the relations among them. And Turing showed, roughly, that if a process has a formally specifiable character then it can be mechanized. So we can appreciate the implications of (i) and (ii) for the philosophy of psychology in this way: if thinking consists in processing representations physically realized in the brain (in the way the internal data structures are realized in a computer) and these representations form a formal system, i.e. a language with its proper combinatorial syntax (and semantics) and a set of derivations rules formally defined over the syntactic features of those representations (allowing for specific but extremely powerful programs to be written in terms of them), then the problem of thinking, as I described it above, can in principle be solved in completely naturalistic terms, thus the mystery surrounding how a physical device can ever have semantically coherent state transitions (processes) can be removed. Thus, given the commitment to naturalism, the hypothesis that the brain is a kind of computer trafficking in representations in virtue of their syntactic properties is the basic idea of LOTH (and the AI vision of cognition).

Computers are environments in which symbols are manipulated in virtue of their formal features, but what is thus preserved are their semantic properties, hence the semantic coherence of symbolic processes. Slightly paraphrasing Haugeland (cf. 1985: 106), who puts the same point nicely in the form of a motto:

> **THE FORMALIST MOTTO**:
> If you take care of the syntax of a representational system, its semantics will take care of

itself.

This is in virtue of the mimicry or mirroring relation between the semantic and formal properties of symbols. As Dennett once put it in describing LOTH, we can view the thinking brain as a syntactically driven engine preserving semantic properties of its processes, i.e. driving a semantic engine. What is so nice about this picture is that if LOTH is true we have a naturalistically adequate causal treatment of *thinking* that respect the semantic properties of the *thoughts* involved: it is in virtue of the physically coded syntactic/formal features that thoughts cause each other while the coherence of their semantic properties is preserved precisely in virtue of this.

Whether or not LOTH actually turns out to be empirically true in the details or in its entire vision of rational thinking, this picture of a syntactic engine driving a semantic one can at least be taken to be an important *philosophical* demonstration of how Descartes' challenge can be met (cf. Rey 1997: chp.8). Descartes claimed that rationality in the sense of having the power "to act in all the contingencies of life in the way in which our reason makes us act" cannot possibly be possessed by a purely physical device: "The rational soul ... could not be in any way extracted from the power of matter ... but must ... be expressly created" (1637/1970: 117-18). Descartes was completely puzzled by just this rational character and semantic coherence of thought processes so much so that he failed to even imagine a possible mechanistic explication of it. He thus was forced to appeal to Divine creation. But we can now see/imagine at least a possible mechanistic/naturalistic scenario.[15]

Intentionality and LOTH

But where do the semantic properties of the mental representations come from in the first place? How can they mean anything? This is Brentano's challenge to a naturalist. Brentano's bafflement was with the **intentionality** of the human mind, its apparently mysterious power to represent things, events, properties in the world. He thought that nothing physical can have this property of intentionality: "The reference to something as an object is a distinguishing characteristic of all mental phenomena. No physical phenomenon exhibits anything similar" (Brentano 1874/1973: 97). This problem of intentionality is the second problem or mystery in Fodor's list that I quoted above. I said that LOTH officially offers only a partial solution to it and perhaps proposes a framework within which the remainder of the solution can be couched and elaborated in a naturalistically acceptable way.

As characterized at the beginning, RTM contains a clause (A1b) that says that the immediate object of a propositional attitude that *P* is a mental representation #*P*# that *means* that *P*. Again, (B1) attributes a compositional semantics to the syntactically complex symbols belonging to one's LOT that are, as per (C), physically realized in the brain of a thinking organism. According to LOTH, the semantic content of propositional attitudes is inherited from the semantic content of the mental symbols. So Brentano's questions for a LOT theorist becomes: how do the symbols in one's LOT get their meanings in the first place? There are two levels or stages at which this question can be raised and answered:

(1) At the level of *atomic* (simple) symbols: how do the atomic symbols represent what

they do?

(2) At the level of *molecular* (phrasal complexes or sentences) symbols: how do molecular symbols represent what they do?

There have been at least two major lines LOT theorists have taken regarding these questions. The one that is least committal might perhaps be usefully described as the official position regarding LOTH's treatment of intentionality. Most LOT theorists seem to have taken this line. The official line doesn't propose any theory about the first stage, but simply assumes that the first question can be answered in a naturalistically acceptable way. In other words, officially LOTH simply assumes that the atomic symbols/expressions in one's LOT have whatever meanings they have.[16]

But, the official line continues, LOTH has a lot to say about the second stage, the stage where the semantic contents are computed or assigned to complex (molecular) symbols on the basis of their combinatorial syntax or grammar together with whatever meanings atomic symbols are assumed to have in the first stage. This procedure is familiar from a Tarski-style[17] definition of truth conditions of *sentences*. The truth-value of complex sentences in propositional logic are completely determined by the truth-values of the atomic sentences they contain together with the rules fixed by the truth-tables of the connectives occurring in the complex sentences. Example: '*P* and *Q*' is true just in case both '*P*' and '*Q*' are true, but false otherwise. This process is similar but more complex in first-order languages, and even more so for natural languages -- in fact, we don't have a completely working compositional semantics for the latter at the moment. So, *if* we have a semantic interpretation of atomic symbols (*if* we have symbols whose reference and extension are fixed at the first stage by whatever naturalistic mechanism turns out to govern it), *then* the combinatorial syntax will take over and effectively determine the semantic interpretation (truth-conditions) of the complex sentences they are constituents of. So officially LOTH would only contribute to a complete naturalization project if there is a naturalistic story at the atomic level.

Early Fodor (1975, 1978, 1978a, 1980), for instance, envisaged a science of psychology which, among other things, would reasonably set for itself the goal of discovering the combinatorial syntactic principles of LOT and the computational rules governing its operations, without worrying much about semantic matters, especially about how to fix the semantics of atomic symbols (he probably thought that this was not a job for LOTH). Similarly, Field (1978) is very explicit about the combinatorial rules for assigning truth-conditions to the sentences of the internal code. In fact, Field's major argument for LOTH is that, given a naturalistic causal theory of reference for atomic symbols, about which he is optimistic (Field 1972), it is the only naturalistic theory that has a chance of solving Brentano's puzzle. For the moment, this is not much more than a hope, but, according to the LOT theorist, it is a well-founded hope based on a number of theoretical and empirical assumptions and data. Furthermore, it is a framework defining a naturalistic research program in which there have been promising successes.[18]

As I said, this official and, in a way, least committal line has been overall the more standard way of conceiving LOTH's role in the project of naturalizing intentionality. But some have gone beyond it and explored the ways in which the resources of LOTH can be exploited even in answering the first question

(1) about the semantics of atomic symbols.

Now, there is a weak version of an answer to (1) on the part of LOTH and a strong version. On the weak version, LOTH may be untendentiously viewed as inevitably providing *some* of the resources in giving the ultimate naturalistic theory in naturalizing the meaning of atomic symbols. The basic idea is that whatever the ultimate naturalistic theory turns out to be true about atomic expressions, computation as conceived by LOTH will be part of it. For instance, it may be that, as with nomic covariation theories of meaning (Fodor 1987, 1990a; Dretske 1981), the meaning of an atomic predicate may consist in its potential to get tokened in the presence of (or, in causal response to) something that instantiates the property the predicate is said to express. A natural way of explicating this potential may partly but ultimately rely on certain computational principles the symbol may be subjected to within a LOT framework, or principles that in some sense govern the "behavior" of the symbol. Insofar as computation is naturalistically understood in the way LOTH proposes, a complete answer to the first question about the semantics of atomic symbols may plausibly involve an explicatory appeal to computation within a system of symbols. This is the weak version because it doesn't see LOTH as proposing a complete solution to the first question (1) above, but only *helping* it.

A strong version would have it that LOTH provides a *complete* naturalistic solution to both questions: given the resources of LOTH we don't need to look any further to meet Brentano's challenge. The basic idea lies in so-called functional or conceptual role semantics, according to which a concept is the concept it is precisely in virtue of the particular causal/functional potential it has in interacting with other concepts. Each concept may be thought of as having a certain distinctive set of epistemic/semantic relations or liaisons to other concepts. We can conceive of this set as determining a certain "conceptual role" for each concept. We can then take these roles to determine the semantic identity of concepts: concepts are the concepts they are because they have the conceptual roles they have; that is to say, among other things, concepts represent whatever they do precisely in virtue of these roles. The idea then is to reduce each *conceptual* role to *causal/functional* role of atomic symbols (now conceived as primitive terms in LOTH), and then use the resources of LOTH to reduce it in turn to *computational* role. Since computation is naturalistically well-defined, the argument goes, and since causal interactions between thoughts and concepts can be understood completely in terms of computation, we can completely naturalize intentionality if we can successfully treat meanings as arising out of thoughts/concepts' internal interactions with each other. In other words, the strong version of LOTH would claim that atomic symbols in LOT have the content they do in virtue of their potential for causal interactions with other tokens, and cashing out this potential in mechanical/naturalistic terms is what, among other things, LOTH is for. LOTH then comes as a naturalistic rescuer for conceptual role semantics.

It is not clear whether any one holds this strong version of LOTH in this rather naive form. But certainly some people have elaborated the basic idea in quite subtle ways, for which Cummins (1989: chp.8) is perhaps the best example. (But also see Block 1986 and Field 1978.) But even in the best hands, the proposal turns out to be very problematic and full of difficulties nobody seems to know how to straighten out. In fact, some of the most ardent critics of taking LOTH as incorporating a functional role semantics turn out to be some of the most ardent defenders of LOTH understood in a weak, non-committal sense we have explored above -- see Fodor (1987: chp.3), Fodor and Lepore (1991), Fodor's attack (1978b) on AI's

way of doing procedural semantics is also relevant here. Haugeland (1981), Searle (1980, 1984), and Putnam (1988) quite explicitly take LOTH to involve a program for providing a complete semantic account of mental symbols, which they then attack accordingly.[19]

As indicated previously, LOTH is almost completely silent about consciousness and the problem of qualia, the third mystery in Fodor's list in the quote above. But the naturalist's hope is that this problem too will be solved, if not by LOTH, then by something else. On the other hand, it is important to emphasize that LOTH is neutral about the naturalizability of consciousness/qualia. If it turns out that qualia cannot be naturalized, this would by no means show that LOTH is false or defective in some way. In fact, there are people who *seem* to think that LOTH may well turn out to be true even though qualia can perhaps not be naturalized (e.g., Block 1980, Chalmers 1996, McGinn 1991).

Finally, it should be emphasized that LOTH has no particular commitment to every symbolic activity's being conscious. Conscious thoughts and thinking may be the tip of a computational iceberg. Nevertheless, there are ways in which LOTH can be helpful for an account of state consciousness that seeks to explain a thought's being conscious in terms of a higher order thought which is about the first order thought. So, to the extent to which thought and thinking are conscious, to that extent LOTH can perhaps be viewed as providing some of the necessary resources for a naturalistic account of state consciousness -- for elaboration see Rosenthal (1997) and Lycan (1997).

Arguments for LOTH

We have already seen two major arguments, perhaps the historically most important ones, for LOTH: First, we have noted that if LOTH is true then all the essential features of the common sense conception of propositional attitudes will be explicated in a naturalistic framework which is likely to be co-opted by scientific cognitive psychology, thus vindicating folk psychology. Second, we discussed that, if true, LOTH would solve one of the mysteries about thinking minds: how is thinking (as characterized above) possible? How is rationality mechanically possible? Then we have also seen a third argument that LOTH would partially contribute to the project of naturalizing intentionality by offering an account of how the semantic properties of whole attitudes are fixed on the basis of their atomic constituents. But there have been many other arguments for LOTH. In this section, I will try to describe only those arguments that have been historically more influential and controversial.

Argument from Contemporary Cognitive Psychology

When Fodor first formulated LOTH with significant elaboration in his (1975), he introduced his major argument for it along with its initial formulation in the first chapter. It was basically this: our best scientific theories and models of different aspects of higher cognition assume a framework that requires a computational/representational medium for them to be true. More specifically, he analyzes the basic form of the information processing models developed to account for three types of cognitive phenomena: *perception* as the fixation of perceptual beliefs, *concept learning* as hypothesis formation and confirmation, and *decision making* as a form of representing and evaluating the consequences of possible

actions carried out by the agent in a situation with a preordered set of preferences. He rightly points out that all these models treat mental processes as computational processes defined over representations. Then he draws what seems to be the obvious conclusion: if these models are right in at least treating mental processes as computational, even if not in detail, then there must be a LOT over which they are defined, hence LOTH.

In Fodor's (1975), the arguments for different aspects of LOTH are diffused and the emphasis, with the book's slogan "no computation without representation", is put on the RTM rather than on (B) or (C). But all the elements are surely there.

Argument from the Productivity of Thought

People seem to be capable of entertaining an infinite number of thoughts, at least in principle, although they in fact entertain only a finite number of them. Indeed adults who speak a natural language are capable of understanding sentences they have never heard uttered before. Here is one: there is a big lake of melted gold on the dark side of the moon. I bet that you never heard this sentence before, and yet, you have no difficulty in understanding it: it is one you're in fact likely to believe false. But this sentence was arbitrary, there are infinitely many such sentences I can in principle utter and you can in principle understand. But understanding a sentence is to entertain the thought/proposition it expresses. So there are in principle infinitely many thoughts you are capable of entertaining. This is sometimes expressed by saying that we have an unbounded *competence* in entertaining different thoughts, even though we have a bounded *performance*. But this unbounded capacity is to be achieved by finite means. For instance, storing an infinite number of representations in our heads is out of the question: we are finite beings. If human cognitive capacities (capacities to entertain an unbounded number of thoughts, or to have attitudes towards an unbounded number of propositions) are productive in this sense, how is this to be explained on the basis of finitary resources?

The explanation LOTH offers is straightforward: postulate a representational system that satisfies at least (B1). Indeed, recursion is the only known way to produce an infinite number of symbols from a finite base. In fact, given LOTH, productivity of thought as a competence mechanism seems to be guaranteed.[20]

Argument from the Systematicity and Compositionality of Thought

Systematicity of thought consists in the empirical fact that the ability to entertain certain thoughts is intrinsically connected to the ability to entertain certain others. Which ones? Thoughts that are related in a certain way. In what way? There is a certain initial difficulty in answering such questions. I think, partly because of this, Fodor (1987) and Fodor and Pylyshyn (1988), who are the original defenders of this kind of argument, first argue for the systematicity of language production and understanding: the ability to produce/understand certain sentences is intrinsically connected to the ability to produce/understand certain others. Given that a mature speaker is able to produce/understand a certain sentence in her native language, by psychological law, there always appear to be a cluster of other sentences that she is able to

produce/understand. For instance, you don't seem to find speakers who know how to express in their native language the fact that John loves the girl but not the fact that the girl loves John. This is apparently so, moreover, for expressions of any n-place relation.

Fodor and Pylyshyn bring out the force of this psychological fact by comparing learning languages the way we actually do with learning a language by memorizing a huge phrase book. In the phrase book model, there is nothing to prevent someone learning how to say ‘John loves the girl’ without learning how to say ‘the girl loves John.’ In fact, that is exactly the way some information booklets prepared for tourists help them to cope with their new social environment. You might, for example, learn from a phrase book how to say ‘I'd like to have a cup of coffee with sugar and milk’ in Turkish without knowing how to say/understand absolutely anything else in Turkish. In other words, the phrase book model of learning a language allows arbitrarily punctate linguistic capabilities. In contrast, a speaker's knowledge of her native language is not punctate, it is *systematic*. Accordingly, you do not find, by nomological necessity, native speakers whose linguistic capacities are punctate.

Now, how is this empirical truth (in fact, a law-like generalization) to be explained? Obviously if this is a general nomological fact, then learning one's native language cannot be modeled on the phrase book model. What is the alternative? The alternative is well known. Native speakers master the grammar and vocabulary of their language. But this is just to say that sentences are not atomic, but have syntactic constituent structure. If you have a vocabulary, the grammar tells you how to combine *systematically* the words into sentences. Hence, in this way, if you know how to construct a particular sentence out of certain words, you automatically know how to construct many others. If you view all sentences as atomic, then, as Fodor and Pylyshyn say, the systematicity of language production/understanding is a mystery, but if you acknowledge that sentences have syntactic constituent structure, systematicity of linguistic capacities is what you automatically get; it is guaranteed. This is the orthodox explanation of linguistic systematicity.

From here, according to Fodor and Pylyshyn, establishing the systematicity of thought as a nomological fact is one step away. If it is a law that the ability to understand a sentence is systematically connected to the ability to understand many others, then it is similarly a law that the ability to think a thought is systematically connected to the ability to think many others. For to understand a sentence is just to think the thought/proposition it expresses. Since, according to RTM, to think a certain thought is just to token a representation in the head that expresses the relevant proposition, the ability to token certain representations is systematically connected to the ability to token certain others. But then, this fact needs an adequate explanation too. The classical explanation LOTH offers is to postulate a system of representations with combinatorial syntax exactly as in the case of the explanation of the linguistic systematicity. This is what (B1) offers.[21] This seems to be the only explanation that does not make the systematicity of thought a miracle, and thus argues for the LOT hypothesis.

However, thought is not only systematic but also compositional: systematically connected thoughts are also always semantically related in such a way that the thoughts so related seem to be composed out of the same semantic elements. For instance, the ability to think ‘John loves the girl’ is connected to the ability to think ‘the girl loves John’ but not to, say, ‘protons are made up of quarks’ or to ‘2+2=4.’ Why is this

so? The answer LOTH gives is to postulate a combinatorial semantics in addition to a combinatorial syntax, where an atomic constituent of a mental sentence makes (approximately) the same semantic contribution to any complex mental expression in which it occurs. This is what Fodor and Pylyshyn call ‘the principle of compositionality’.[22]

In brief, it is an argument for LOTH that it offers a cogent and principled solution to the systematicity and compositionality of cognitive capacities by postulating a system of representations that has a combinatorial syntax *and* semantics, i.e., a system of representations that satisfies at least (B1).

Argument from the Systematicity of Thinking (Inferential Coherence)

Systematicity of thought does not seem to be restricted solely to the systematic ability to entertain certain *thoughts*. If the system of mental representations does have a combinatorial syntax, then there is a set of rules, syntactic formation rules, so to speak, that govern the construction of well-formed expressions in the system. It is this fact, (B1), that guarantees that if you can form a mental sentence on the basis of certain rules, then you can also form many others on the basis of the same rules. The rules of combinatorial syntax determine the syntactic or formal structure of complex mental representations. This is the *formative* (or, *formational*) aspect of systematicity. But inferential *thought processes* seem to be systematic too: the ability to make certain inferences is intrinsically connected to the ability to make certain many others. For instance, you do not find minds that can infer ‘*A*’ from ‘*A&B*’ but cannot infer ‘*C*’ from ‘*A&B&C*.’ It seems to be a psychological fact that inferential capacities come in clusters that are homogeneous in certain aspects. How is this fact (i.e., the *inferential* or *transformational* systematicity) to be explained?

As we have seen, the explanation LOTH offers depends on the exploitation of the notion of logical form or syntactic structure determined by the combinatorial syntax postulated for the representational system. The combinatorial syntax not only gives us a criterion of well-formedness for mental expressions, but it also defines the logical form or syntactic structure for each well-formed expression. The classical solution to inferential systematicity is to make the mental operations on representations sensitive to their form or structure, i.e. to insist on (B2). Since, from a syntactic view point, similarly formed expressions will have similar forms, it is possible to define a single operation which will apply to only certain expressions that have a certain form, say, only to conjunctions, or disjunctions. This allows the LOT theorist to give homogeneous explanations of what appear to be homogeneous classes of inferential capacities. This is one of the greatest virtues of LOTH, hence provides an argument for it.

The solution LOTH offers for what I called the problem of thinking, above, is connected to the argument here because the two phenomena are connected in a deep way. Thinking requires that the logico-semantic properties of a particular thought process (say, inferring that John is happy from knowing that if John is at the beach then John is happy and coming to realize that John is indeed at the beach) be somehow causally implicated in the process. The systematicity of inferential thought processes then is based on the observation that if the agent is capable of making *that* particular inference, then she is capable of making

many other somehow similarly organized inferences. But the idea of similar organization in this context seems to demand some sort of classification of thoughts independently of their *particular* content. But what can the basis of such a classification be? The only basis seems to be the logico-syntactic properties of thoughts, their form. Although it feels a little uneasy to talk about syntactic properties of thoughts common-sensically understood, it seems that they are forced upon us by the very attempt to understand their semantic properties: how, for instance, could we explain the semantic content of the thought that if John is at the beach then he is happy without somehow appealing to its being a *conditional*? This is the point of contact between the two phenomena. Especially when the demands of naturalism are added to this picture, inferring a LOT (= a representational system satisfying B) realized in the brain becomes almost irresistible. Indeed Rey (1995) doesn't resist and claims that, given the above observations, LOTH can be established on the basis of arguments that are not "merely empirical". I leave it to the reader to evaluate whether mere critical reflection on our concepts of thought and thinking could, all by itself, establish LOTH.[23]

Objections to LOTH

There have been numerous arguments against LOTH. Some of them are directed more specifically against the Representational Theory of Mind (A), some against functionalist materialism (C). Here I will concentrate only on those arguments specifically targeting (B) -- the most controversial component of LOTH.

Regress Arguments against LOTH

These arguments rely on the explanations offered by LOTH defenders for certain aspects of natural languages. In particular, many LOT theorists advert to LOTH to explain (1) how natural languages are learned, (2) how natural languages are understood, or (3) how the utterances in such languages can be meaningful. For instance, according to Fodor (1975), natural languages are learned by forming and confirming hypotheses about the translation of natural language sentences into Mentalese such as: ‘Snow is white’ is true in English if and only if *P*, where ‘*P*’ is a sentence in one's LOT. But to be able to do that, one needs a representational medium in which to form and confirm hypotheses. The LOT is such a medium. Again, natural languages are understood because, roughly, such an understanding consists in translating their sentences into one's Mentalese. Similarly, natural language utterances are meaningful in virtue of the meanings of corresponding Mentalese sentences.

The basic complaint is that in each of these cases, either the explanations generate a regress because the same sort of explanations ought to be given for how the LOT is learned, understood or can be meaningful, or else they are gratuitous because if a successful explanation can be given for LOT that does not generate a regress then it could and ought to be given for the natural language phenomena without introducing a LOT (see, e.g. Blackburn 1984). Fodor's response in (1975) is (1) that LOT is not learned, it's innate, (2) that it's understood in a different sense than the sense involved in natural language comprehension, and (3) that LOT sentences acquire their meanings not in virtue of another meaningful language but in a completely different way, perhaps by standing in some sort of causal relation to what they represent (see

above) or by having certain computational profiles. For many who have a Wittgensteinian bent, these replies are not likely to be very convincing. But here the issues tend to concern RTM rather than (B).

Laurence and Margolis (1997) point out that the regress arguments depend on the assumption that LOTH is introduced only to explain (1)-(3). If it can be shown that there are lots of other empirical phenomena for which the LOTH provides good explanations, then the regress arguments fail because LOTH then would not be gratuitous. In fact, as we have seen above, there are plenty of such phenomena. But still it is important to realize that the sort of explanations proposed for the understanding of one's LOT (computational use/activity of LOT sentences with certain meanings) and how LOT sentences can be meaningful (computational roles and/or nomic relations with the world) cannot be given for (1)-(3): it's unclear, for example, what it would be like to give a computational role and/or nomic relation account for the meanings of natural language utterances.

Propositional Attitudes without Explicit Representations

Dennett in his review of Fodor's (1975) has raised the following objection (cf. Fodor 1987: 21-3 for a similar discussion):

> In a recent conversation with the designer of a chess-playing program I heard the following criticism of a rival program: "it thinks it should get its queen out early." This ascribes a propositional attitude to the program in a very useful and predictive way, for as the designer went on to say, one can usefully count on chasing that queen around the board. But for all the many levels of explicit representation to be found in that program, nowhere is anything roughly synonymous with "I should get my queen out early" explicitly tokened. The level of analysis to which the designer's remark belongs describes features of the program that are, in an entirely innocent way, emergent properties of the computational processes that have "engineering reality." I see no reason to believe that the relation between belief-talk and psychological talk will be any more direct. (Dennett 1981: 107)

The objection, as Fodor (1987: 22) points out, isn't that the program has a *dispositional*, or *potential*, belief that it will get its queen out early. Rather, the program actually operates on this belief. There appear to be lots of other examples: e.g. in reasoning we pretty often follow certain inference rules like modus ponens, disjunctive syllogism, etc. without necessarily explicitly representing them.

The standard reply to such objections is to draw a distinction between rules on the basis of which Mentalese data-structures are manipulated, and the data-structures themselves (intuitively, the program/data distinction). LOTH is not committed to every rule's being explicitly represented. In fact, as a point of nomological fact, in a computational device not every rule can be explicitly represented: some *have to* be hard-wired and, thus, implicit in this sense. In other words, LOTH permits but doesn't require that rules be explicitly represented. On the other hand, data structures *have to* be explicitly represented: it is these that are manipulated formally by the rules. No causal manipulation is possible without explicit tokening of these structures. According to Fodor, if a propositional attitude is an actual episode in one's

reasoning that plays a causal role, then LOTH is committed to explicit representation of its content, which is as per (A2 and B2) causally implicated in the physical process realizing that reasoning. Dispositional propositional attitudes can then be accounted for in terms of an appropriate principle of inferential closure of explicitly represented propositional attitudes (cf. Lycan 1986).

Dennett's chess program certainly involves explicit representations of the chess board, the pieces, etc. and perhaps some of the rules. Which rules are implicit and which are explicit depend on the empirical details of the program. Pointing to the fact that there may be some rules that are emergent out of the implementation of explicit rules and data-structures does not suffice to undermine LOTH.

Explicit Representations without Propositional Attitudes

In any sufficiently complex computational system, there are bound to be many symbol manipulations with no obviously corresponding description at the level of propositional attitudes. For instance, when a multiplication program is run through a standard conventional computer, the steps of the program are translated into the computer's machine language and executed there, but at this level the operations apply to 1's and 0's with no obvious way to map them onto the original numbers to be multiplied or to the multiplication operation. So it seems that at the levels that, according to Dennett, have engineering reality there are plenty of explicit tokenings of representations with appropriate operations that don't correspond to anything like the propositional attitudes of folk psychology. In other words, there is plenty of symbolic activity which it would be wrong to say a *person* engages in. Rather, they are done by the person's subpersonal computational *components* as opposed to the person. How to rule out such cases? (cf. Fodor 1987: 23-6 for a similar discussion.)

They are ruled out by an appropriate reading of (A1) and (B1): (A1) says that the person herself must stand in an appropriate computational relation to a Mentalese sentence, which, as per (B1), has a suitable syntax and semantics. Only then, will the sentence constitute the person's having a propositional attitude. Not all explicit symbols in one's LOT will satisfy this. In other words, not every computational routine will correspond to a processing appropriately described as storage in, e.g., the "belief-box". Furthermore, as pointed out by Fodor (1987), LOTH would vindicate the common sense view of propositional attitudes if they turn out to be computational relations to Mentalese sentences. It may not be further required that every explicit representation correspond to a propositional attitude.

There have been many other objections to LOTH in recent years raised especially by connectionists: that LOT systems cannot handle certain cognitive tasks like perceptual pattern recognition, that they are too brittle and not sufficiently damage resistant, that they don't exhibit graceful degradation when physically damaged or as a response to noisy or degraded input, that they are too rigid, deterministic, so are not well-suited for modeling humans' capacity to satisfy multiple soft-constraints so gracefully, that they are not biologically realistic, and so on. For useful discussions of these and many similar objections, see Rumelhart, McClelland and the PDP Research Group (1986), Fodor and Pylyshyn (1988), Bechtel and Abrahamsen (1991), Horgan and Tienson (1996), and McLaughlin and Warfield (1994).

The Connectionism/Classicism Debate

When Jerry Fodor published his influential book, *The Language of Thought*, in (1975), he called LOTH "the only game in town." As we have seen, it was the philosophical articulation of the assumptions that underlay the new developments in "cognitive sciences" after the demise of behaviorism. Fodor argued for the truth of LOTH on the basis of the successes of the best scientific theories we had then. Indeed most of the scientific work in cognitive psychology, psycholinguistics, and AI assumed the framework of LOTH.

In the early 1980's, however, Fodor's claim that LOTH was the only game in town was beginning to be challenged by some people who were working on so-called connectionist networks. They claimed that **connectionism** offered a new and radically different alternative to classicism in modeling cognitive phenomena. The name 'classicism' has since then become to be applied to the LOTH framework. On the other hand, many classicists like Fodor thought that connectionism was nothing but a slightly more sophisticated way with which the old and long dead associationism, whose roots could be traced back to early British empiricists, was being revived. In 1988 Fodor and Pylyshyn (F&P) published a long article, "Connectionism and Cognitive Architecture: A Critical Analysis", in which they launched a formidable attack on connectionism, which largely set the terms for the ensuing debate between connectionists and classicists.

F&P's forceful criticism consists in posing a dilemma for connectionists: They either fail to explain the law-like cognitive regularities like systematicity and productivity in an adequate way or the connectionist models are nothing but mere implementation models of classical architectures; hence, they fail to provide a radically new paradigm as connectionists claim. This conclusion was also meant to be a challenge: Explain the cognitive regularities in question without postulating a LOT architecture.

First, let me present F&P's argument against connectionism in a somewhat reconstructed fashion. It will be helpful to characterize the debate by locating the issues according to the reactions many connectionists had to the premises of the argument.

F&P's Argument against Connectionism in their (1988):

> **(i)** Cognition essentially involves representational states and causal operations whose domain and range are these states; consequently, any scientifically adequate account of cognition should acknowledge such states and processes.
>
> **(ii)** Higher cognition (specifically, thought and thinking with propositional content) conceived in this way, has certain scientifically interesting properties: in particular, it is a law of nature that cognitive capacities are *productive*, *systematic,* and *inferentially coherent*.
>
> **(iii)** Accordingly, the architecture of any proposed cognitive model is scientifically adequate only if it guarantees that cognitive capacities are productive, systematic, etc. This

would amount to explaining, in the scientifically relevant and required sense, how it could be a law that cognition has these properties.

(iv) The only way (i.e. necessary condition) for a cognitive architecture to guarantee systematicity (etc.) is for it to involve a representational system for which (B) is true (see above). (Classical architectures necessarily satisfy (B).)

(v) Either the architecture of connectionist models does satisfy (B), or it does not.

(vi) If it does, then connectionist models are implementations of the classical LOT architecture and have little new to offer (i.e., they fail to compete with classicism, and thus connectionism does not constitute a radically new way of modeling cognition).

(vii) If it does not, then (since connectionism does not then guarantee systematicity, etc., in the required sense) connectionism is empirically false as a theory of the *cognitive* architecture.

(viii) Therefore, connectionism is either true as an implementation theory, or empirically false as a theory of cognitive architecture.

The notion of *cognitive architecture* assumes special importance in this debate. F&P's characterization of the notion goes as follows:

> The architecture of the cognitive system consists of the set of basic operations, resources, functions, principles, etc. (generally the sorts of properties that would be described in a "user's manual" for that architecture if it were available on a computer) whose domain and range are the *representational states* of the organism. (1988: 10)

Also, note that (B1) and (B2) are meta-architectural properties in that they are themselves conditions upon any proposed specific architecture's being classical. They define classicism per se, but not any particular way of being classical. Classicism as such simply claims that whatever the *particular* cognitive architecture of the brain might turn out to be (whatever the *specific* grammar of Mentalese turns out to be), (B) must be true of it. F&P claim that this is the only way an architecture can be said to guarantee the nomological necessity of cognitive regularities like systematicity, etc. This seems to be the relevant and required sense in which a scientific explanation of cognition is required to guarantee the regularities -- hence the third premise in their argument.

Connectionist responses have fallen into four classes:

(1) *Deny premise(i)*. The rejection of (i) commits connectionists to what is sometimes called *radical* or *eliminativist connectionism*. Premise (i), as F&P point out, draws a general line between eliminativism and representationalism (or, intentional realism). There has been

some controversy as to whether connectionism constitutes a serious challenge to the fundamental tenets of folk psychology.[24] It is too early to assess the potential of connectionism in this regard.[25] On the other hand, many connectionists do in fact advance their models as having causally efficacious representational states, and explicitly endorse F&P's first premise. So they seem to accept intentional realism.[26]

(2) *Accept the conclusion*. This group may be seen as more or less accepting the cogency of the entire argument, and characterizes itself as *implementationalist*: they hold that connectionist networks will *implement* a classical architecture or language of thought. According to this group, the appropriate niche for neural networks is closer to neuroscience than to cognitive psychology. They seem to view the importance of the program in terms of its prospects of closing the gap between the neurosciences and high-level cognitive theorizing. In this, many seem content to admit premise (vi).

(3) *Deny premise (ii) or (iv)*. Some connectionists reject (ii) or (iv),[27] holding that there are no lawlike cognitive regularities such as systematicity (etc.) to be explained, or that such regularities do not require a (B)-like architecture for their explanation. Those who question (ii) often question the empirical evidence for systematicity (etc.) and tend to ignore the challenge put forward by F&P. Those who question (iv) also often question (ii), or they argue that there can be very different sort of explanations for systematicity and the like (e.g. evolutionary explanations, see Braddon-Mitchell and Fitzpatrick 1990), or they question the very notion of explanation involved (e.g. Matthews 1994). There are indeed quite a number of different kinds of arguments in the literature against these premises.[28] For a sampling, see Aydede (1995) and McLaughlin (1993b), who partitions the debate similarly. (See also Aydede (1998), in the Other Internet Resources section of this entry.). The debate on the issues raised by these connectionists and their sympathizers is still very lively.

(4) *Deny premise (vi)*. The group of connectionists who have taken F&P's challenge most seriously has tended to reject premise (vi) in their argument, while accepting, on the face of it, the previous five premises (sometimes with reservations on the issue of productivity). They think that it is possible for connectionist representations to be syntactically structured in some sense without being classical. Prominent in this group are Smolensky (1990a, 1990b, 1995), van Gelder (1989, 1990, 1991), Chalmers (1990, 1993). Some connectionists whose models give support to this line include Elman (1989), Hinton (1990), Touretzky (1990), Pollack (1990).[29]

Much of the recent debate between connectionists and classicists has focused on this option, so let's see how it is possible to reject premise (vi), which seems true by definition of classicism. The connectionists' answer, roughly put, is that when you devise a representational system whose satisfaction of (B) relies on a *non-concatenative* realization of structural/syntactic complexity of representations, you have a non-classical system. (See especially Smolensky 1990a and van Gelder 1990.) Interestingly, some classicists like Fodor and McLaughlin (1990) (F&M) seem to agree. F&M stipulate that you have a classical system

only if the syntactic complexity of representations is realized *concatenatively*, or as it is sometimes put, *explicitly*:

> We ... stipulate that for a pair of expression types E1, E2, the first is a *Classical* constituent of the second *only if* the first is tokened whenever the second is tokened. (F&M 1990: 186)

The issues about how connectionists propose to obtain constituent sructure non-concatenatively tend to be complex and technical. But they propose to exploit so called *distributed representations* in certain novel ways. The essential idea behind most of them is to use vector (and tensor) algebra (involving superimposition, multiplication, etc. of vectors) in composing and decomposing connectionist representations which consist in coding patterns of activity across neuron-like units which can be modeled as vectors. The result of such techniques is the production of representations that have in some interesting sense a complexity whose constituent structure is largely implicit in that the constituents are not tokened explicitly when the representations are tokened, but can be recovered by further operations upon them. The interested reader should consult some of the pioneering work by Elman (1989), Hinton (1990), Smolensky (1989, 1990, 1995), Touretzky (1990), Pollack (1990).

F&M's criticism, more specifically stated, however, is this. Connectionists with such techniques only satisfy (B1) in some "extended sense", but they are incapable of satisfying (B2), precisely because their way of satisfying (B1) is committed to a non-concatenative realization of syntactic structures.

Some connectionists disagree (e.g. Chalmers 1993): they claim that you can have structure-sensitive transformations or operations defined over representations whose syntactic structure is non-concatenatively realized. So given the apparent agreement that non-concatenative realization is what makes a system non-classical, connectionists claim that they can and do perfectly satisfy (B) in its entirety with their connectionist models without implementing classical models.

The debate is still intense and there is a fast growing literature built around the many issues raised by it. Aydede (1997) offers an extensive analysis of the debate between classicists and this group of connectionists with special attention to the conceptual underpinnings of the debate. He argues that both parties are wrong in assuming that concatenative realization is relevant to the characterization of LOTH. He specifies what minimally needs to be the case for the LOTH to be true, and why.

Bibliography

- Aizawa, K. (forthcoming). "Representations without Rules, Connectionism and the Syntactic Argument."
- Aydede, Murat (1995). "Connectionism and Language of Thought", *CSLI Technical Report*, Stanford, CSLI-95-195. (This is an early version of Aydede 1997 but contains quite a lot of expository material not contained in 1997.)
- Aydede, Murat (1997). "Language of Thought: The Connectionist Contribution," *Minds and Machines*, Vol. 7, No. 1, pp. 57-101.

- Armstrong, D.M. (1980). *The Nature of Mind*, Ithaca, NY: Cornell University Press.
- Barsalou, L. W. (1993a). "Flexibility, Structure, and Linguistic Vagary in Concepts: Manifestations of a Compositional System of Perceptual Symbols" in *Theories of Memory*, edited by A. Collins, S. Gathercole, M. Conway and P. Morris, Hillsdale, NJ: Lawrence Erlbaum Associates.
- Barsalou, L. W., W. Yeh, B. J. Luka, K. L. Olseth, K. S. Mix, and L.-L. Wu. (1993b). "Concepts and Meaning", *Chicago Linguistics Society* 29.
- Barsalou, L. W., and J. J. Prinz. (1997). "Mundane Creativity in Perceptual Symbol Systems" in *Creative Thought: An Investigation of Conceptual Structures and Processes*, edited by T. B. Ward, S. M. Smith and J. Vaid, Washington, DC: American Psychological Association.
- Barwise, Jon and John Etchemendy (1995). *Hyperproof*, Stanford, Palo Alto: CSLI Publications.
- Barwise, J. and J. Perry (1983). *Situations and Attitudes*, Cambridge, Massachusetts: MIT Press.
- Bechtel, W. and A. Abrahamsen (1991). *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*, Oxford, UK: Basil Blackwell.
- Blacburn, S. (1984). *Spreading the Word*, Oxford, UK: Oxford University Press.
- Block, Ned (1980). "Troubles with Functionalism" in *Readings in Philosophy of Psychology*, N. Block (ed.), Vol.1, Cambridge, Massachusetts: Harvard University Press, 1980. (Originally appeared in *Perception and Cognition: Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science*, C.W. Savage (ed.), Minneapolis: The University of Minnesota Press, 1978.)
- Block, N. (ed.) (1981). *Imagery*. Cambridge, Massachusetts: MIT Press.
- Block, N. (1983a). "Mental Pictures and Cognitive Science," *Philosophical Review* 93: 499-542. (Reprinted in *Mind and Cognition*, W.G. Lycan (ed.), Oxford, UK: Basil Blackwell, 1990.)
- Block, N. (1983b). "The Photographic Fallacy in the Debate about Mental Imagery", *Nous* 17: 651-62.
- Block, Ned (1986). "Advertisement for a Semantics for Psychology" in *Studies in the Philosophy of Mind: Midwest Studies in Philosophy*, Vol.10, P. French, T. Euhling and H. Wettstein (eds.), Minneapolis: University of Minnesota Press.
- Braddon-Mitchell, David and John Fitzpatrick (1990). "Explanation and the Language of Thought," *Synthese* 83: 3-29.
- Brentano, Franz (1874/1973). *Psychology from an Empirical Standpoint*, A. Rancurello, D. Terrell and L. McAlister (trans.), London: Routledge and Kegan Paul.
- Butler, Keith (1991). "Towards a Connectionist Cognitive Architecture," *Mind and Language*, Vol. 6, No. 3, pp. 252-72.
- Chalmers, David J. (1990). "Syntactic Transformations on Distributed Representations," *Connection Science*, Vol. 2.
- Chalmers, David J. (1993). "Connectionism and Copositionality: Why Fodor and Pylyshyn Were Wrong" in *Philosophical Psychology* 6: 305-319.
- Chalmers, David J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*, Oxford, UK: Oxford University Press.
- Churchland, Patricia Smith (1986). *Neurophilosophy: Toward a Unified Science of Mind-Brain*, Cambridge, Massachusetts: MIT Press.
- Churchland, Patricia Smith (1987). "Epistemology in the Age of Neuroscience," *Journal of*

Philosophy, Vol. 84, No. 10, pp. 544-553.

- Churchland, Patricia S. and Terrence J. Sejnowski (1989). "Neural Representation and Neural Computation" in *Neural Connections, Neural Computation*, L. Nadel, L.A. Cooper, P. Culicover and R.M. Harnish (eds.), Cambridge, Massachusetts: MIT Press, 1989.
- Churchland, Paul M. (1990). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, Cambridge, Massachusetts: MIT Press.
- Churchland, Paul M. (1981). "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* 78: 67-90.
- Churchland, Paul M. and P.S. Churchland (1990). "Could a Machine Think?," *Scientific American*, Vol. 262, No. 1, pp. 32-37.
- Clark, Andy (1988). "Thoughts, Sentences and Cognitive Science," *Philosophical Psychology*, Vol. 1, No. 3, pp. 263-278.
- Clark, Andy (1989a). "Beyond Eliminativism," *Mind and Language,* Vol. 4, No. 4, pp. 251-279.
- Clark, Andy (1989b). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*, Cambridge, Massachusetts: MIT Press.
- Clark, Andy (1990). "Connectionism, Competence, and Explanation," *British Journal for Philosophy of Science*, 41: 195-222.
- Clark, Andy (1991). "Systematicity, Structured Representations and Cognitive Architecture: A Reply to Fodor and Pylyshyn" in *Connectionism and the Philosophy of Mind*, Terence Horgan and John Tienson (eds.), *Studies in Cognitive Systems* (Volume 9), Dordrecht: Kluwer Academic Publishers, 1991.
- Clark, Andy (1994). "Language of Thought (2)" in *A Companion to the Philosophy of Mind* edited by S. Guttenplan, Oxford, UK: Basil Blackwell, 1994.
- Cummins, Robert (1986). "Inexplicit Information" in The Representation of Knowledge and Belief, M. Brand and R.M. Harnish (eds.), Tucson, Arizona: Arizona University Press, 1986.
- Cummins, Robert (1989). *Meaning and Mental Representation*, Cambridge, Massachusetts: MIT Press.
- Cummins, Robert and Georg Schwarz (1987). "Radical Connectionism," *The Southern Journal of Philosophy*, Vol. XXVI, Supplement.
- Davidson, Donald (1984). *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press.
- Davies, Martin (1989). "Connectionism, Modularity, and Tacit Knowledge," *British Journal for the Philosophy of Science* 40: 541-555.
- Davies, Martin (1991). "Concepts, Connectionism, and the Language of Thought," in *Philosophy and Connectionist Theory*, W. Ramsey, S.P. Stich and D.E. Rumelhart (eds.), Hillsdale, NJ: Lawrence Erlbaum, 1991.
- Davies, M. (1995). "Two Notions of Implicit Rules," *Philosophical Perspectives* 9: 153-83.
- Dennett, D.C. (1978). "Two Approaches to Mental Images" in *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, Massachusetts: MIT Press, 1981.
- Dennett, D.C. (1981). "Cure for the Common Code" in *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, Massachusetts: MIT Press, 1981. (Originally appeared in *Mind*, April 1977.)
- Dennett, Daniel C. (1986). "The Logical Geography of Computational Approaches: A View from the East Pole" in *The Representation of Knowledge and Belief*, Myles Brand and Robert M.

Harnish (eds.), Tucson: The University of Arizona Press, 1986.

- Dennett, Daniel C. (1991a). "Real Patterns," *Journal of Philosophy*, Vol. LXXXVIII, No. 1, pp. 27-51.
- Dennett, Daniel C. (1991b). "Mother Nature Versus the Walking Encyclopedia: A Western Drama" in *Philosophy and Connectionist Theory*, W. Ramsey, S.P. Stich and D.E. Rumelhart (eds.), Lawrence Erlbaum Associates.
- Descartes, R. (1637/1970). "Discourse on the Method" in *The Philosophical Works of Descartes*, Vol.I, E.S. Haldane and G.R.T. Ross (trans.), Cambridge, UK: Cambridge University Press.
- Devitt, Michael (1990). "A Narrow Representational Theory of the Mind," *Mind and Cognition*, W.G. Lycan (ed.), Oxford, UK: Basil Blackwell, 1990.
- Devitt, Michael (1996). *Coming to our Senses: A Naturalistic Program for Semantic Localism*, Cambridge, UK: Cambridge University Press.
- Devitt, Michael and Sterelny, Kim (1987). *Language and Reality: An Introduction to the Philosophy of Language*, Cambridge, Massachusetts: MIT Press.
- Dretske, Fred (1981). *Knowledge and the Flow of Information*, Cambridge, Massachusetts: MIT Press.
- Dretske, Fred (1988). *Explaining Behavior*, Cambridge, Massachusetts: MIT Press.
- Elman, Jeffrey L. (1989). "Structured Representations and Connectionist Models", *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society*, Ann Arbor, Michigan, pp.17-23.
- Field, Hartry H. (1972). "Tarski's Theory of Truth", *Journal of Philosophy*, 69: 347-75.
- Field, Hartry H. (1978). "Mental Representation", *Erkenntnis* 13, 1, pp.9-61. (Also in *Mental Representation: A Reader*, S.P. Stich and T.A. Warfield (eds.), Oxford, UK: Basil Blackwell, 1994. References in the text are to this edition.)
- Fodor, Jerry A. (1975). *The Language of Thought*, Cambridge, Massachusetts: Harvard University Press.
- Fodor, Jerry A. (1978). "Propositional Attitudes" in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge, Massachusetts: MIT Press, 1981. (Originally appeared in *The Monist* 64, No.4, 1978.)
- Fodor, Jerry A. (1978a). "Computation and Reduction" in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge, Massachusetts: MIT Press. (Originally appeared in *Minnesota Studies in the Philosophy of Science: Perception and Cognition*, Vol. 9, W. Savage (ed.), 1978.)
- Fodor, Jerry A. (1978b). "Tom Swift and His Procedural Grandother," *Cognition*, Vol. 6. (Also in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge, Massachusetts: MIT Press, 1981.)
- Fodor, Jerry A. (1980). "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology", *Behaviorial and Brain Sciences* 3, 1, 1980. (Also in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge, Massachusetts: MIT Press, 1981. References in the text are to this edition.)
- Fodor, Jerry A. (1981a). *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, Cambridge, Massachusetts: MIT Press.
- Fodor, Jerry A. (1981b), "Introduction: Something on the State of the Art" in *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, J.A. Fodor, Cambridge,

Massachusetts: MIT Press, 1981.

- Fodor, Jerry A. (1983). *The Modularity of Mind*, Cambridge, Massachusetts: MIT Press.
- Fodor, Jerry A. (1985). "Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum", *Mind* 94, 1985, pp.76-100. (Also in *A Theory of Content and Other Essays*, J.A. Fodor, Cambridge, Massachusetts: MIT Press. References in the text are to this edition.)
- Fodor, Jerry A. (1986). "Banish DisContent" in *Language, Mind, and Logic*, J. Butterfield (ed.), Cambridge, UK: Cambridge University Press, 1986. (Also in *Mind and Cognition*, William Lycan (ed.), Oxford, UK: Basil Blackwell, 1990.)
- Fodor, Jerry A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, Massachusetts: MIT Press.
- Fodor, Jerry A. (1989). "Substitution Arguments and the Individuation of Belief" in *A Theory of Content and Other Essays*, J. Fodor, Cambridge, Massachusetts: MIT Press, 1990. (Originally appeared in *Method, Reason and Language*, G. Boolos (ed.), Cambridge, UK: Cambridge University Press, 1989.)
- Fodor, Jerry A. (1990). *A Theory of Content and Other Essays*, Cambridge, Massachusetts: MIT Press.
- Fodor, Jerry A. (1991). "Replies" (Ch.15) in *Meaning in Mind: Fodor and his Critics*, B. Loewer and G. Rey (eds.), Oxford, UK: Basil Blackwell, 1991.
- Fodor, Jerry A. and Ernest Lepore (1991). "Why Meaning (Probably) Isn't Conceptual Role?", *Mind and Language*, Vol. 6, No. 4, pp. 328-43.
- Fodor, Jerry A. and B. McLaughlin (1990). "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work," *Cognition* 35: 183-204.
- Fodor, Jerry A. and Zenon W. Pylyshyn (1988). "Connectionism and Cognitive Architecture: A Critical Analysis" in S. Pinker and J. Mehler, eds., *Connections and Symbols*, Cambridge, Massachusetts: MIT Press (A *Cognition* Special Issue).
- Garson, J. (forthcoming). "Systematicity and Classical Architecture".
- Grice, H.P. (1957). "Meaning", *Philosophical Review*, 66: 377-88.
- Hadley, R. F. (1995). "The 'Explicit-Implicit' Distinction," *Minds and Machines,* 5: 219-42.
- Haugeland, John (1981). "The Nature and Plausibility of Cognitivism," *Behavioral and Brain Sciences* I, 2: 215-60 (with peer commentary and replies).
- Haugeland, John (1985). *Artificial Intelligence: The Very Idea*, Cambridge, Massachusetts: MIT Press.
- Hinton, Geoffrey (1990). "Mapping Part-Whole Hierarchies into Connectionist Networks," *Artificial Intelligence*, Vol. 46, Nos. 1-2, (Special Issue on Connetionist Symbol Processing).
- Horgan, T. E. and J. Tienson (1996). *Connectionism and the Philosophy of Psychology*, Cambridge, Massachusetts: MIT Press.
- Kirsh, D. (1990). "When Is Information Explicitly Represented?" in *Information, Language and Cognition*. P. Hanson (ed.), University of British Columbia Press.
- Kosslyn, S.M. (1980). *Image and Mind*. Cambridge, Massachusetts: Harvard University Press.
- Kosslyn, S.M. (1981). "The Medium and the Message in Mental Imagery: A Theory" in *Imagery*, N. Block (ed.), Cambridge, Massachusetts: MIT Press, 1981.
- Kosslyn, S.M. (1994). *Image and Brain*, Cambridge, Massachusetts: MIT Press.
- Laurence, Stephen and Eric Margolis (1997). "Regress Arguments Against the Language of

Thought", *Analysis*, Vol. 57, No. 1.

- Lewis, David (1972). "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy*, 50(3):249-58. (Also in *Readings in Philosophy of Psychology*, Ned Block (ed.), Vols.1, Cambridge, Massachusetts: Harvard University Press, 1980.)
- Loar, Brian F. (1982a). *Mind and Meaning*, Cambridge, UK: Cambridge University Press.
- Loar, Brian F. (1982b). "Must Beliefs Be Sentences?" in *Proceedings of the Philosophy of Science Association for 1982*, Asquith, P. and T. Nickles (eds.), East Lansing, Michigan, 1983.
- Lycan, William G. (1981). "Toward a Homuncular Theory of Believing," *Cognition and Brain Theory* 4(2): 139-159.
- Lycan, W. G. (1986). "Tacit Belief" in *Belief: Form, Content, and Function*, R. Bogdan (ed.), Oxford, UK: Oxford University Press.
- Lycan, William (1993). "A Deductive Argument for the Representational Theory of Thinking," *Mind and Language*, Vol. 8, No. 3, pp. 404-22.
- Lycan, William (1997). "Consciousness as Internal Monitoring" in *The Nature of Consciousness: Philosophical Debates*, edited by N. Block, O. Flanagan and G. Güzeldere, Cambrigde, Massachusetts: MIT Press.
- Margolis, Eric (forthcoming). "How to Acquire a Concept?", *Mind and Language*.
- Marr, David (1982). *Vision*, San Francisco: W. H. Freeman.
- Matthew, Robert J. (1994). "Three-Concept Monte: Explanation, Implementation and Systematicity", *Synthese*, Vol. 101, No. 3, pp. 347-63.
- McGinn, Colin (1991). *The Problem of Consciousness*, Oxford, UK: Basil Blackwell.
- McLaughlin, B.P. (1993a). "The Connectionism/Classicism Battle to Win Souls," *Philosophical Studies* 71: 163-90.
- McLaughlin, B.P. (1993b). "Systematicity, Conceptual Truth, and Evolution," in *Philosophy and Cognitive Science,* C. Hookway and D. Peterson (eds.), Royal Institute of Philosophy, Supplement No. 34.
- McLaughlin, B.P. and Ted Warfield (1994). "The Allures of Connectionism Reaxamined", *Synthese* 101, pp. 365-400
- Millikan, Ruth Garrett (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*, Cambridge, Massachusetts: MIT Press.
- Milikan, Ruth Garrett (1993). *White Queen Psychology and Other Essays for Alice*, Cambridge, Massachusetts: MIT Press.
- Papineau, D. (1987). *Reality and Representation*, Oxford, UK: Basil Blackwell.
- Perry, John and David Israel (1991). "Fodor and Psychological Explanations" in *Meaning in Mind: Fodor and his Critics*, B. Loewer and G. Rey (eds.), Oxford, UK: Basil Blackwell, 1991.
- Pollack, J.B. (1990). "Recursive Distributed Representations," *Artificial Intelligence*, Vol.46, Nos.1-2, (Special Issue on Connetionist Symbol Processing).
- Prinz, Jesse J. (1997). *Perceptual Cognition*, Ph.D. Dissertation in philosophy, The Universiy of Chicago.
- Putnam, Hilary (1988), *Representation and Reality*, Cambridge, Massachusetts: MIT Press.
- Pylyshyn, Z.W. (1978). "Imagery and Artificial Intelligence" in *Perception and Cognition*. W. Savage (ed.), University of Minnesota Press. (Reprinted in *Readings in the Philosophy of Psychology*, N. Block (ed.), Cambridge, Massachusetts: MIT Press, 1980.)

- Pylyshyn, Zenon W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, Massachusetts: MIT Press.
- Ramsey, William, Stephen Stich and Joseph Garon (1991). "Connectionsim, Eliminativism and the Future of Folk Psychology," in *Philosophy and Connectionist Theory*, W. Ramsey, D. Rumelhart and Stephen Stich (eds.), Hillsdale, NJ: Lawrence Erlbaum.
- Rey, Georges (1981). "What are Mental Images?" in *Readings in the Philosophy of Psychology*, N. Block (ed.), Vol. 2, Cambridge, Massachusetts: Harvard University Press, 1981.
- Rey, Georges (1991). "An Explanatory Budget for Connectionism and Eliminativism" in *Connectionism and the Philosophy of Mind*, Terence Horgan and John Tienson (eds.), Studies in Cognitive Systems (Volume 9), Dordrecht: Kluwer Academic Publishers.
- Rey, Georges (1992). "Sensational Sentences Switched", *Philosophical Studies* 67: 73-103.
- Rey, Georges (1993). "Sensational Sentences" in Consciousness, M. Davies and G. Humphrey (eds.), Oxford, UK: Basil Blackwell, pp. 240-57.
- Rey, Georges (1995). "A Not 'Merely Empirical' Argument for a Language of Thought," in *Philosophical Perspectives* 9, J. Tomberlin (ed.), pp. 201-222.
- Rey, Georges (1997). *Contemporary Philosophy of Mind: A Contentiously Classical Approach*, Oxfor, UK: Basil Blackwell.
- Rosenthal, D.M. (1997). "A Theory of Consciousness" in *The Nature of Consciousness: Philosophical Debates*, edited by N. Block, O. Flanagan and G. Güzeldere, Cambridge, Massachusetts: MIT Press.
- Rumelhart, D.E. and J.L. McClelland (1986). "PDP Models and General Issues in Cognitive Science," in *Parallel Distributed Processing*, Vol.1, D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, Cambridge, Massachusetts: MIT Press, 1986.
- Rumelhart, D.E., J.L. McClelland, and the PDP Research Group (1986). *Parallel Distributed Processing*, (Vols. 1&2), Cambridge, Massachusetts: MIT Press.
- Schiffer, Stephen (1981). "Truth and the Theory of Content" in *Meaning and Understanding*, H. Parret and J. Bouvaresse (eds.), Berlin: Walter de Gruyter, 1981.
- Searle, John R. (1980). "Minds, Brains, and Programs" *Behavioral and Brain Sciences* III, 3: 417-24.
- Searle, John R. (1984). *Minds, Brains and Science*, Cambridge, Massachusetts: Harvard University Press.
- Searle, John R. (1990). "Is the Brain a Digital Computer?", *Proceedings and Addresses of the APA*, Vol. 64, No. 3, November 1990.
- Searle, John R. (1992). *The Rediscovery of Mind*, Cambridge, Massachusetts: MIT Press.
- Shepard, R. and Cooper, L. (1982). *Mental Images and their Transformations*. Cambridge, Massachusetts: MIT Press.
- Smolensky, Paul (1988). "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences* 11: 1-23.
- Smolensky, Paul (1990a). "Connectionism, Constituency, and the Language of Thought" in *Meaning in Mind: Fodor and His Critics*, B. Loewer and G. Rey (eds.), : Oxford, UK: Basil Blackwell, 1991.
- Smolensky, Paul (1990b). "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems," *Artificial Intelligence*, Vol. 46, Nos. 1-2, (Special Issue on

Connetionist Symbol Processing), November 1990.

- Smolensky, Paul (1995). "Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Cognitive Architecture" in *Connectionism: Debates on Psychological Explanation*, C. Macdonald and G. Macdonald (eds.), Oxford, UK: Basil Blackwell, 1995.
- Stalnaker, Robert C. (1984). *Inquiry*, Cambridge, Massachusetts: MIT Press.
- Sterelny, K. (1986). "The Imagery Debate", *Philosophy of Science* 53: 560-83. (Reprinted in *Mind and Cognition*, W. Lycan (ed.), Oxford, UK: Basil Blackwell, 1990.)
- Sterelny, Kim (1990). *The Representational Theory of Mind*, Cambridge, Massachusetts: MIT Press.
- Stich, Stephen (1983). *From Folk Psychology to Cognitive Science: The Case against Belief*, Cambridge, Massachusetts: MIT Press.
- Tarski, Alfred (1956). "The Concept of truth in Formalized Languages" in *Logic, Semantics and Metamathematics*, J.Woodger (trans.), Oxford, UK: Oxford University Press.
- Touretzky, D.S. (1990). "BoltzCONS: Dynamic Symbol Structures in a Connectionist Network," *Artificial Intelligence*, Vol. 46, Nos. 1-2, (Special Issue on Connetionist Symbol Processing).
- Tye, M. (1984). "The Debate about Mental Imagery", *Journal of Philosophy* 81: 678-91.
- Tye, M. (1991). *The Imagery Debate*, Cambridge, Massachusetts: MIT Press.
- van Gelder, Timothy (1989). "Compositionality and the Explanation of Cognitive Processes", *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society*, Ann Arbor, Michigan, pp. 34-41.
- van Gelder, Timothy (1990). "Compositionality: A Connectionist Variation on a Classical Theme," *Cognitive Science*, Vol. 14.
- van Gelder, Timothy (1991). "Classical Questions, Radical Answers: Connectionism and the Structure of Mental Representations" in *Connectionism and the Philosophy of Mind*, Terence Horgan and John Tienson (eds.), Studies in Cognitive Systems (Volume 9), Dordrecht: Kluwer Academic Publishers.
- Wallis, C. (forthcoming). "Nomic Necessity and Systematicity."

Other Internet Resources

- Aydede, Murat (1998), "LOTH: State of the Art"
  (this is a more detailed and comprehensive version of this entry)
- Chalmers, D., Bibliography on LOTH

Related Entries

artificial intelligence | belief | Church-Turing Thesis | cognitive science | concepts | connectionism | consciousness: representational theories of | folk psychology: as a theory | functionalism | intentionality | meaning | mental imagery | mind: computational models of | mind: philosophy of | naturalism | physicalism | propositional attitude reports | psychology, philosophy of | reasoning: automated | Turing, Alan

Murat Aydede
maydede@phil.ufl.edu

First published: May 28, 1998
Content last modified: July 11, 1999

Stanford Encyclopedia of Philosophy
Notes to The Language of Thought Hypothesis

Notes

1. This is to convey the basic idea that each type of attitude (e.g., believing) is realized by the same type of psychological relation (e.g., being inside the computationally defined B-Box) and by no others, for which see below. So the mapping from attitudes *A* into psychological relations *R* is meant to be injective.

2. Lycan (1986), Davies (1989, 1995), Cummins (1986), Hadley (1995). A parallel discussion is going on in AI: Kirsh (1990).

3. See Fodor (1985, 1986, 1987: chp.1), Devitt (1990).

4. But see Rey (1992, 1993) for an attempt to expand LOTH to sensations and qualia.

5. See for instance the controversy involved in the so-called imagery debate. The literature here is huge but the following sample may be useful: Block (1981, 1983b), Dennett (1978), Kosslyn (1980), Pylyshyn (1978), Rey (1981), Sterelny (1986), Tye (1991).

6. E.g., Marr (1982), or any textbook on vision or language comprehension and production.

7. E.g., Kosslyn (1980, 1994); Shepard and Cooper (1982). In fact, some theorists even go so far as to claim that *all* cognition is done in an image-like symbol system -- early British empiricists from Locke to Hume held something like this view, but more recently, see L. Barsalou and his colleagues who have been developing models to that effect (Barsalou 1993a, Barsalou et al 1993b, Barsalou and Prinz 1997).

8. The controversial issue here is not the absurdity of the claim that there are literally pictures or images in the brain. Probably no one believes this claim these days. Rather, postulating picture-like representations is to be cashed out in functionalist terms. Pictures as mental representations presumably bear some non-arbitrary isomorphisms to what they represent, although it is hard to make this sort of claim crystal-clear in purely functionalist terms. See, for instance, Kosslyn (1980, 1981), Block (1983a, 1983b), Tye (1984).

9. The issues here are too complex and difficult to go over here in any useful detail, but for a general criticism of pictures as mental representations, see the critical essays in Block (1981) and Rey (1981); for an attempt to overcome many such criticisms, see Barsalou and Prinz (1997) and Prinz (1997). The contemporary debate about the adequacy of a purely imagistic medium for capturing what is involved in

making a judgment and discursive thinking seem to parallel some of Kant's critique of British Empiricism in general and of Hume's associationism in particular, as indeed emphasized by many classicists like Fodor and Pylyshyn (1988), Rey (1997).

10. See, e.g., Barwise and Etchemendy's *Hyperproof* (1995).

11. For a non-nativist but otherwise quite Fodorian account of concept acquisition, see Margolis (forthcoming).

12. But see Rey (1992, 1993) for an attempt to extend LOTH in this direction.

13. Also, Hubert Dreyfus and John Haugeland's many writings indicate that they are realist about propositional attitudes but would reject LOTH nevertheless.

14. Almost all British empiricists might be put in this latter category too, but they were in fact closer to LOTH by having embraced something like (B1) in some imagistic version. But it looks like they could not be better than being associationist regarding *thought processes*: they could not exploit the clear implications of modern symbolic logic and the advancement of computers -- they did not have their Frege and Turing, though Hobbes came close. This rendering of RTM relies on a broad interpretation of the notion of mental representation, of course, which has not always been the intended interpretation of Fodor: there are many places where he defends RTM (by that name) meaning to include (B) by default (Fodor 1981b, 1985, 1987, 1998). This should cause no confusion. Here I have chosen to stick to the literal meaning of the phrase rather than to its historically more accurate use -- this has become necessary, at any rate, in the light of the recent classicism/connectionism debate to which we will return below.

15. For a powerful elaboration of this line of thought, see Rey (1997).

16. A number of proposals have been offered by contemporary theorists (who are not necessarily defenders of LOTH as opposed to being mere RTM theorists but whose proposals can be adapted by LOT theorists) about how exactly to pursue that project. See, for instance, Fodor (1987, 1990), Dretske (1981, 1988), Millikan (1984, 1993), Papineau (1987), Devitt (1996), Loar (1982a), Field (1972, 1978), Block (1986).

17. Tarski (1956), Field (1972), Davidson (1984).

18. Although I described the line above as official and presented it as requiring a compositional semantics, and although almost all the defenders of LOTH conceive of it in this way because they think that is what empirical facts about thought and language demand, nevertheless it is perhaps important to be pedantic about exactly what LOTH is minimally committed to. Minimally, it is *not* committed to regarding the internal code as having a compositional semantics, namely a semantics where the meaning

of complex sentences are determined by the meanings of its constituents together with their syntax; this, in effect, requires that the atomic expressions always make (approximately) the same semantic contributions to the whole of which they are constituents (idioms excepted). But strictly speaking LOTH can live without having a strictly compositional semantics if it turns out that there are other ways of explaining those empirical facts about the mind to which I will come below. Admittedly, in such a case LOTH would lose some portion of its appeal and interest. But even if this scenario turns out to be the case, there are still a lot of facts for LOTH to explain. Having said this, however, I will simply forget it in what follows.

19. For fairness I should add that Searle's and Haugeland's criticisms are directed against AI community at large, and there, it has been common to conceive the computational model of mind as potentially involving a complete solution to semantic worries among others. Thus, Haugeland termed his target ‘GOFAI’ (the Good Old Fashion Artifical Intelligence). Similarly, Searle's famous Chinese Room Argument was directed against what he called ‘Strong AI’.

20. See Fodor (1985, 1987), Fodor and Pylyshyn (1988) for an elaborate presentation of this argument for LOTH.

21. It should be noted however that (B1) is a meta-architectural condition that needs to be satisfied by any *particular* grammar for Mentalese, just as an analogue for (B1) is a condition upon the *specific* grammar of all systematic languages (see below).

22. It is somewhat confusing that Fodor and Pylyshyn called this *empirical cognitive regularity* "compositionality" of cognitive capacities. In particular, the empirical phenomenon -- i.e., the *fact* that systematically connected thoughts are also always semantically related or semantically close to each other -- that needs to be explained is explained by LOT theorists by what is also called semantic compositionality: namely, the semantic value of a complex expression is a function of the semantic value of its atomic constituents such that each atomic constituent makes approximately the same semantic contribution to the context in which it occurs. This is what the postulation of a combinatorial semantics in conjunction with a combinatorial syntax buys for LOT-theorists in adequately explaining the empirical regularity in question. See Fodor and Pylyshyn (1988: 41-5).

23. For a prioristic arguments of this sort, see also Lycan (1993) and Davies (1989, 1991).

24. For example, Patricia and Paul Churchland, who have been the champions of eliminativism, hope that connectionism is the long waited theory which will provide the scientific foundations of the elimination of folk psychological constructs in "psychology" (P.S. Churchland 1986, 1987; Churchland and Sejnowski 1989; P.M. Churchland 1990; P.S. Churchland and P.M. Churchland 1990). Ramsey, Stich and Garon (1991) have recently defended the claim that if certain sorts of connectionist models turn out to be right then the elimination of folk psychology will be inevitable. Dennett (1986), and Cummins and Schwartz (1987) have also pointed out the potential of connectionism in the elimination of at least certain aspects of folk psychology.

25.In fact, it is not clear at all, how connectionism can genuinely give support to intentional eliminativism as far as the units (or collections of units) in connectionist networks are treated as representing. If they are not treated as such, it is hard to see how they could be models of *cognitive* phenomena, and thus hard to see how they can present any eliminativist challenge. However, there appear to be two vague strands among eliminativists in this regard. One stems from the intuition that it is unlikely that there are really any concrete, isolable, and modularly identifiable symbol structures realized in the brain that would correspond to what Stich has called (1983: 237ff.) functionally discrete beliefs and desires of folk psychology, and connectionist networks, it is claimed, will vindicate this intuition. For similar remarks, among others, see Dennett (1986, 1991a), Clark (1988, 1989b). The second trend seems to be that connectionism will vindicate the claim that the explanation of mental phenomena doesn't require a full-blown semantics for such higher-order states as propositional attitudes. Rather, all that is needed is an account of some form of information processing at a much lower level, which, it is hoped, will be sufficient for the whole range of cognitive phenomena. Again, it is not clear what the proposals are. But see Paul Churchland (1990).

26. It seems clear from some of the so far proposed models that many connectionists have been developing their models ultimately with an eye to capture the generalizations in their respective psychological domain. To see this it is enough to look at some of the papers in the second PDP volume (Rumelhart, McClelland and the PDP Research Group, 1986) among which Rumelhart and McClelland's paper on modeling learning the past tenses of English verbs is particularly celebrated. At the end, it is of course an open empirical question whether connectionist models will ultimately be able to capture them, or whether the generalizations they come up with will be compatible with or be the ones implicitly recognized by the folk, just as it is an open question whether classical models will ultimately be successful in this respect. Whatever the final outcome might be, however, it is *prima facie* the case that many connectionists intend their models to be taken as contributions within the intentional realist tradition. Smolensky (1988) is the most articulated defense of something like this position. He calls his position "the Proper Treatment of Connectionism" (PTC) and clearly separates it from various eliminativist positions.

27. Premise (iii) is intimately connected to (ii) and (iv). So its rejection by itself does not mean much. Premise (iii), according to F&P, is there to prevent certain *ad hoc* solutions on the part of connectionists in the explanation of cognitive regularities mentioned in (ii). Premise (v) is close to being a tautology. So no one has any quarrel with it, although van Gelder (1991) comes very close to rejecting it on the ground that with every shift in scientific paradigms the conceptual apparatus of the previous and challenged paradigms becomes inadequate to correctly characterize the new and challenging paradigm.

28. Some people who object to (ii) or (iv) are Dennett (1991b), Sterelny (1990), Rumelhart and McClelland (1986), Clark (1989b, 1991), Braddon-Mitchell and Fitzpatrick (1990), Butler (1991), Matthew (1994), Aizawa (forthcoming), Garson (forthcoming) and Wallis (forthcoming).

29. Smolensky, for instance, is very explicit in his rejection of premise (vi): " ...distributed connectionist

architectures, without implementing the Classical architecture, can nonetheless provide structured mental representations and mental processes sensitive to that structure" (1990a: 215) .

Murat Aydede
m-aydede@uchicago.edu

First published: May 28, 1998
Content last modified: July 11, 1999

The Church-Turing Thesis

There are various equivalent formulations of the Church-Turing thesis. A common one is that every effective computation can be carried out by a Turing machine. The Church-Turing thesis is often misunderstood, particularly in recent writing in the philosophy of mind.

The Thesis and its History

The Church-Turing thesis concerns the notion of an *effective* or *mechanical* method in logic and mathematics. 'Effective' and its synonym 'mechanical' are terms of art in these disciplines: they do not carry their everyday meaning. A method, or procedure, M, for achieving some desired result is called 'effective' or 'mechanical' just in case

1. M is set out in terms of a finite number of exact instructions (each instruction being expressed by means of a finite number of symbols);
2. M will, if carried out without error, always produce the desired result in a finite number of steps;
3. M can (in practice or in principle) be carried out by a human being unaided by any machinery save paper and pencil;
4. M demands no insight or ingenuity on the part of the human being carrying it out.

A well-known example of an effective method is the truth table test for tautologousness. In practice, of course, this test is unworkable for formulae containing a large number of propositional variables, but in principle one could apply it successfully to any formula of the propositional calculus, given sufficient time, tenacity, paper, and pencils.

Statements that there is an effective method for achieving such-and-such a result are commonly expressed by saying that there is an effective method for obtaining the values of such-and-such a mathematical function. For example, that there is an effective method for determining whether or not any given formula of the propositional calculus is a tautology - e.g. the truth table method - is expressed in function-speak by saying that there is an effective method for obtaining the values of a function, call it T, whose domain is the set of formulae of the propositional calculus and whose value for any given formula x, written T(x), is 1 or 0 according to whether x is, or is not, a tautology.

The notion of an effective method is an informal one, and attempts to characterise effectiveness, such as the above, lack rigour, for the key requirement that the method demand no insight or ingenuity is left unexplicated. One of Turing's achievements in his paper of 1936 was to present a formally exact predicate with which the informal predicate 'can be calculated by means of an effective method' may be replaced. Church did the same (1936a). The replacement predicates that Turing and Church proposed were, on the face of it, very different from one another, but they turned out to be equivalent, in the sense that each picks out the same set of mathematical functions. The Church-Turing thesis is the assertion that this set contains every function whose values can be obtained by a method satisfying the above conditions for effectiveness. (Clearly, if there were functions of which the informal predicate, but not the formal predicate, were true, then the latter would be less general than the former and so could not reasonably be employed to replace it.) When the thesis is expressed in terms of the formal concept proposed by Turing, it is appropriate to refer to the thesis also as 'Turing's thesis'; and mutatis mutandis in the case of Church.

The formal concept proposed by Turing is that of *computability by Turing machine*. He argued for the claim (Turing's thesis) that whenever there is an effective method for obtaining the values of a mathematical function, the function can be computed by a Turing machine. The converse claim is easily established, for a Turing machine program is itself a specification of an effective method: a human being can work through the instructions in the program and carry out the operations called for without the exercise of any ingenuity or insight. If Turing's thesis is correct then talk about the existence and non-existence of effective methods can be replaced throughout mathematics and logic by talk about the existence or non-existence of Turing machine programs.

Turing stated his thesis in numerous places, with varying degrees of rigour. The following formulation is one of the most accessible.

> **Turing's thesis**: LCMs [logical computing machines: Turing's expression for Turing machines] can do anything that could be described as "rule of thumb" or "purely mechanical". (Turing 1948:7.)

He adds:

> This is sufficiently well established that it is now agreed amongst logicians that "calculable by means of an LCM" is the correct accurate rendering of such phrases. (Ibid.)

Turing introduced this thesis in the course of arguing that the Entscheidungsproblem, or decision problem, for the predicate calculus - posed by Hilbert (Hilbert and Ackermann 1928) - is unsolvable. Here is Church's account of the Entscheidungsproblem:

> By the Entscheidungsproblem of a system of symbolic logic is here understood the problem to find an effective method by which, given any expression Q in the notation of the system, it can be determined whether or not Q is provable in the system. (Church 1936b: 41.)

The truth table test is such a method for the propositional calculus. Turing showed that, given his thesis, there can be no such method for the predicate calculus. He proved formally that there is no Turing machine which can determine, in a finite number of steps, whether or not any given formula of the predicate calculus is a theorem of the calculus. So, given his thesis that if an effective method exists then it can be carried out by one of his machines, it follows that there is no such method to be found.

Church had arrived at the same negative result a few months earlier, employing the concept of lambda-definability in place of computability by Turing machine. Church and Turing discovered the result quite independently of one another. Turing's method of obtaining it is rather more satisfying than Church's, as Church himself acknowledged in a review of Turing's work:

> computability by a Turing machine ... has the advantage of making the identification with effectiveness in the ordinary (not explicitly defined) sense evident immediately. (1937a: 43.)

(Another aspect in which their approaches differ is that Turing's concerns were rather more general than Church's, in that the latter considered only functions of positive integers (see below), whereas Turing described his work as encompassing "computable functions of an integral variable or a real or computable variable, computable predicates, and so forth" (1936: 230). He intended to pursue the theory of computable functions of a real variable in a subsequent paper, but in fact did not do so.)

Church used the (informal) expression 'effectively calculable' to indicate that there is an effective method for calculating the values of the function. He proposed that we

> define the notion ... of an effectively calculable function of positive integers by identifying it with the notion of a recursive function of positive integers (or of a lambda-definable function of positive integers). (1936a: .)

The concept of a lambda-definable function is due to Church and Kleene (Church 1932, 1936a, 1941, Kleene 1935) and the concept of a recursive function to Gödel and Herbrand (Gödel 1934, Herbrand 1932). The class of lambda-definable functions and the class of recursive functions are identical. This was established in the case of functions of positive integers by Church and Kleene (Church 1936a, Kleene 1936). After learning of Church's proposal, Turing quickly established that the apparatus of lambda-definability and his own apparatus of computability are equivalent (1936: 263ff). Thus, in Church's

proposal, the words ‘recursive function of positive integers’ can be replaced by the words ‘function of positive integers computable by Turing machine’.

Post referred to Church’s identification of effective calculability with recursiveness as a "working hypothesis", and quite properly criticised Church for masking this hypothesis as a definition.

> [T]o mask this identification under a definition ... blinds us to the need of its continual verification. (Post 1936: 105.)

This, then, is the "working hypothesis" that, in effect, Church proposed:

> **Church’s thesis**: A function of positive integers is effectively calculable only if recursive.

The reverse implication, that every recursive function of positive integers is effectively calculable, is commonly referred to as *the converse of Church’s thesis* (although Church himself did not so distinguish, bundling both theses together in his ‘definition’). If attention is restricted to functions of positive integers then Church’s thesis and Turing’s thesis are equivalent, in view of the previously mentioned results by Church, Kleene and Turing.

The term ‘Church-Turing thesis’ seems to have been first introduced by Kleene, with a small flourish of bias in favour of Church:

> So Turing’s and Church’s theses are equivalent. We shall usually refer to them both as *Church’s thesis*, or in connection with that one of its ... versions which deals with ‘Turing machines’ as *the Church-Turing thesis*. (Kleene 1967: 232.)

Much evidence has been amassed for the ‘working hypothesis’ proposed by Church and Turing in 1936. Perhaps the fullest survey is to be found in chapters 12 and 13 of Kleene (1952). In summary: (1) Every effectively calculable function that has been investigated in this respect has turned out to be computable by Turing machine. (2) All known methods or operations for obtaining new effectively calculable functions from given effectively calculable functions are paralleled by methods for constructing new Turing machines from given Turing machines. (3) All attempts to give an exact analysis of the intuitive notion of an effectively calculable function have turned out to be equivalent in the sense that each analysis offered has been proved to pick out the same class of functions, namely those that are computable by Turing machine. Because of the diversity of the various analyses the latter is generally considered strong evidence. For example, apart from the analyses already mentioned in terms of lambda-definability and recursiveness, there are analyses in terms of register machines (Shepherdson and Sturgis 1963), Post’s canonical and normal systems (Post 1943, 1946), combinatory definability (Schönfinkel 1924, Curry 1929, 1930, 1932), Markov algorithms (Markov 1960), and Gödel’s notion of reckonability (Gödel 1936, Kleene 1952).

While there have from time to time been attempts to call the Church-Turing thesis into question (for

example by Kalmar (1959); Mendelson (1963) replies), the summary of the situation that Turing gave in 1948 is no less true today: "it is now agreed amongst logicians that ‘calculable by means of an LCM’ is the correct accurate rendering" (of the informal notion in question).

Misunderstandings of the Thesis

A myth has arisen concerning Turing’s paper of 1936, namely that he there gave a treatment of, and established fundamental results concerning, the limits of what can be computed by machine - a myth that has passed into the philosophy of mind, to wide and pernicious effect. For example, the *Oxford Companion to the Mind* states: "Turing showed that his very simple machine ... can specify the steps required for the solution of any problem that can be solved by instructions, explicitly stated rules, or procedures" (Gregory 1987: 784). Dennett maintains that "Turing had proven - and this is probably his greatest contribution - that his Universal Turing machine can compute any function that any computer, with any architecture, can compute" (1991: 215). Sterelny asserts "Astonishingly, Turing was able to show that any procedure that can be computed at all can be computed by a Turing machine. ... Despite their simple organisation, Turing machines are, in principle, as powerful as any other mode of organizing computing systems" (1990: 37, 238). In similar vein, Paul Churchland writes: "The interesting thing about a universal Turing machine is that, *for any well-defined computational procedure whatever, a universal Turing machine is capable of simulating a machine that will execute those procedures*. It does this by reproducing exactly the input/output behaviour of the machine being simulated" (1988:105). Also: Turing’s "results entail something remarkable, namely that a standard digital computer, given only the right program, a large enough memory and sufficient time, can compute *any* rule-governed input-output function. That is, it can display any systematic pattern of responses to the environment whatsoever" (Paul and Patricia Churchland 1990: 26). These various quotations are typical of current writing on the foundations of the computational theory of mind.

Turing did not show that his machines can solve any problem that can be solved "by instructions, explicitly stated rules, or procedures" and nor did he prove that a universal Turing machine "can compute any function that any computer, with any architecture, can compute". He proved that his universal machine can compute any function that any *Turing machine* can compute; and he put forward, and advanced philosophical arguments in support of, the thesis here called Turing’s thesis. But a thesis concerning the extent of effective methods - which is to say, concerning the extent of procedures of a certain sort that a *human being unaided by machinery* is capable of carrying out - carries no implication concerning the extent of the procedures that machines are capable of carrying out, even machines acting in accordance with ‘explicitly stated rules’. For among a machine’s repertoire of atomic operations there may be those that no human being unaided by machinery can perform.

The further proposition, very different from Turing’s own thesis, that a Turing machine can compute whatever can be computed *by any machine working on finite data in accordance with a finite program of instructions* is nowadays sometimes referred to as the Church-Turing thesis or as Church’s thesis. For example, Smolensky says:

> connectionist models ... may possibly even challenge the strong construal of Church's Thesis as the claim that the class of well-defined computations is exhausted by those of Turing machines. (Smolensky 1988: 3.)

This loosening of established terminology is unfortunate, for neither Church nor Turing endorsed, or even formulated, this further proposition. There are numerous examples of this extended usage in the literature. The following are typical.

> [T]he work of Church and Turing fundamentally connects computers and Turing machines. The limits of Turing machines, according to the Church-Turing thesis, also describe the theoretical limits of all computers. (McArthur 1991: 401.)

> [The] Church/ Turing thesis ... equates the mathematically precise notion of "solvable by a Turing machine" with the informal, intuitive notion of "solvable effectively", which alludes to all real computers and all programming languages, those that we know about at present as well as those that we do not. (Harel 1992: 233.)

> The Church-Turing thesis makes a bold claim about the theoretical limits to computation. (Cleland 1993: 283.)

Also (more distant still from anything that Church or Turing actually wrote):

> The first aspect that we examine of Church's Thesis ... [w]e can formulate, more precisely: The behaviour of any discrete physical system evolving according to local mechanical laws is recursive. (Odifreddi 1989: 107.)

> I can now state the physical version of the Church-Turing principle: "Every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means." This formulation is both better defined and more physical than Turing's own way of expressing it. (Deutsch 1985: 99.)

Gandy (1980) is one of the few writers to distinguish explicitly between Turing's thesis and the stronger proposition that whatever can be calculated by a machine can be calculated by a Turing machine. Borrowing Gandy's terminology, I will call the stronger proposition 'Thesis M'. I will use expressions such as 'the Church-Turing thesis properly so-called' for the proposition that Church and Turing themselves endorsed.

> **Thesis M**: Whatever can be calculated by a machine (working on finite data in accordance with a finite program of instructions) is Turing-machine-computable.

Thesis M itself admits of two interpretations, according to whether the phrase 'can be calculated by a machine' is taken in the narrow sense of 'can be calculated by a machine that conforms to the physical

laws (if not to the resource constraints) of the actual world', or in a wide sense that abstracts from the issue of whether or not the notional machine in question could exist in the actual world. The narrow version of thesis M is an empirical proposition whose truth-value is unknown. The wide version of thesis M is known to be false. Various notional machines have been described which can calculate functions that are not Turing-machine-computable (for example, Abramson (1971), da Costa and Doria (1991), (1994), Doyle (1982), Hogarth (1994), Pour-El and Richards (1979), (1981), Scarpellini (1963), Siegelmann and Sontag (1994), Stannett (1990), Stewart (1991); Copeland and Sylvan (1997) is a survey).

The literature on the computational theory of the mind contains numerous endorsements of propositions equivalent or similar to thesis M that are supported by nothing more than a reference to the work of Turing or Church (as is illustrated by a number of the quotations given earlier). Perhaps some writers are simply misled by the terminological practice that has grown up whereby a thesis concerning which there is little real doubt, the Church-Turing thesis properly so-called, and a different thesis of unknown truth-value, are referred to indiscriminately as Church's thesis or the Church-Turing thesis (albeit with accompanying hedges like 'strong form' and 'physical version'). Some - Dennett and Sterelny, for example - think themselves entitled to endorse the stronger proposition because they believe that, somehow, Turing proved it. Other writers maintain thesis M (or some equivalent or near equivalent) on the spurious ground that the various and prima facie very different attempts - by Turing, Church, Post, Markov, and others - to characterise in precise terms the informal notion of an effective procedure have turned out to be equivalent to one another. This is evidence concerning the extent of effective procedures, and not evidence concerning the extent of what can be calculated by machine or organ.

This simple error of confusing the Church-Turing thesis properly so-called with thesis M has led to some remarkable claims in the foundations of psychology. For example, Boden insists that "If a psychological science is possible at all, it must be capable of being expressed in computational terms " (Boden 1988: 259). This is presumably false. The possibility that psychological science will in the future find need to employ mathematical functions that are *not* Turing-machine-computable cannot be ruled out. Boden's reason for thinking her claim true is (she says) her belief that "Alan Turing ... *proved* that a language capable of defining 'effective procedures' suffices, in principle, to solve any computable problem" (ibid.;the italics are Boden's).

It is important to note that in the technical literature the word 'computable' is often tied *by definition* to effective calculability. Thus a function is said to be computable if and only if there is an effective procedure for determining its values. Accordingly, a common formulation of the Church-Turing thesis in the technical literature and in textbooks is:

> All computable functions are computable by Turing machine.

Corollaries such as the following are sometimes offered:

> certain functions are uncomputable in an absolute sense: uncomputable even by [Turing machine], and, therefore, uncomputable by any past, present, or future real machine.

> (Boolos and Jeffrey 1980: 55.)

Given this definition of ‘computable’, the Church-Turing thesis does entail that if a function f is not computable by Turing machine then it is not computable by any machine (for if f is not computable by Turing machine then, by the thesis, there is no effective procedure for determining f’s values, and so f is not computable). Of course, a terminological decision like this cannot settle the truth-value of thesis M; rather, those who abide by this decision are prevented from describing any machines that falsify thesis M as *computing*. Yet to a casual reader of the literature, statements like the one just quoted may appear to say more than they in fact do.

The word ‘mechanical’, too, in technical usage, is tied to effectiveness and, as already remarked, ‘mechanical’ and ‘effective’ are used interchangeably. (Gandy (1988) outlines the history of this usage of the word ‘mechanical’.) Thus statements like the following are to be found in the technical literature:

> Turing proposed that a certain class of abstract machines could perform any ‘mechanical’ computing procedure. (Mendelson 1964: 229.)

Understood correctly, this remark attributes to Turing not thesis M but the Church-Turing thesis. This usage of ‘mechanical’ tends to obscure the possibility that there may be machines, or biological organs, that calculate (or compute, in a broad sense) functions that are not Turing-machine-computable. For the question ‘Can a machine execute a procedure that is not mechanical?’ may appear self-answering, yet this is precisely what is asked if thesis M is questioned.

An error which, unfortunately, is common in modern writing on computability and the brain is to hold that Turing’s results somehow entail that the brain, and indeed any biological or physical system whatever, can be *simulated* by a Turing machine. For example, the entry on Turing in the recent *A Companion to the Philosophy of Mind* contains the following claims: "we can depend on there being a Turing machine that captures the functional relations of the brain", for so long as "these relations between input and output are functionally well-behaved enough to be describable by ... mathematical relationships ... we know that some specific version of a Turing machine will be able to mimic them" (Guttenplan 1994: 595). Searle writes in a similar fashion:

> Can the operations of the brain be simulated on a digital computer? ... The answer seems to me ... demonstrably ‘Yes’ ... That is, naturally interpreted, the question means: Is there some description of the brain such that under that description you could do a computational simulation of the operations of the brain. But given Church’s thesis that anything that can be given a precise enough characterization as a set of steps can be simulated on a digital computer, it follows trivially that the question has an affirmative answer. (Searle 1992: 200.)

So too Johnson-Laird, and the Churchlands:

> If you assume that [consciousness] is scientifically explicable ... [and] [g]ranted that the [Church-Turing] thesis is correct, then ... [i]f you believe [functionalism] to be false ... then ... you [should] hold that consciousness could be modelled in a computer program in the same way that, say, the weather can be modelled ... [and if] you accept functionalism ... you should believe that consciousness is a computational process. (Johnson-Laird 1987: 252.)

> Church's Thesis says that whatever is computable is Turing computable. Assuming, with some safety, that what the mind-brain does is computable, then it can in principle be simulated by a computer. (Churchland and Churchland 1983: 6.)

As previously mentioned, Churchland and Churchland believe, erroneously, that Turing's "results entail ... that a standard digital computer, given only the right program, a large enough memory and sufficient time, can ... display any systematic pattern of responses to the environment whatsoever" (1990: 26). This no doubt explains why they think they can assume "with some safety" that what the mind-brain does is computable, for on their understanding of matters this is to assume only that the mind-brain exhibits a systematic pattern of responses, or is characterised by a 'rule-governed' (1990: 26) input-output function.

The Church-Turing thesis does not entail that the brain (or the mind, or consciousness) can be modelled by a Turing machine program, not even in conjunction with the belief that the brain (or mind, etc.) is scientifically explicable, or exhibits a systematic pattern of responses to the environment, or is 'rule-governed' (etc.). Each of the authors quoted seems to be assuming the truth of a close cousin of thesis M, which I will call

> **Thesis S**: Any process that can be given a systematic mathematical description (or a 'precise enough characterization as a set of steps', or that is scientifically describable or scientifically explicable) can be simulated by a Turing machine.

As with thesis M, neither the Church-Turing thesis properly so-called nor any result proved by Turing or Church entails thesis S. This is so even when the thesis is taken narrowly, as concerning processes that conform to the physics of the real world. (Thesis S taken in the wide sense is known to be false; see the references given earlier re the wide version of thesis M.) Any device or organ whose internal processes can be described completely *by means of effectively calculable functions* can be simulated exactly by a Turing machine program (provided that the input into the device or organ is itself Turing-machine-computable, which is to say, is either finite or expressible as a computable number, in Turing's sense (which is explained below)); but any device or organ whose mathematical description involves functions that are not effectively calculable cannot be so simulated. As Turing showed, there are uncountably many such functions. (Examples from logic are Turing's famous halting function (described in the entry on Turing machines) and the function D whose domain is the set of well-formed formulae of the predicate calculus and whose values, D(x), are 1 or 0 according to whether x is, or is not, derivable from the Bernays-Hilbert-Ackermann axioms for predicate logic.) It is an open question whether a completed neuroscience will employ functions that are not effectively calculable.

Some Key Remarks by Turing

Turing introduces his machines with the intention of providing an idealised description of a certain human activity, the tedious one of *numerical computation*, which until the advent of automatic computing machines was the occupation of many thousands of people in commerce, government, and research establishments. He prefaces his first description of a Turing machine with the words:

> We may compare a man in the process of computing a ... number to a machine. (Turing 1936: 231.)

The Turing machine is a model, idealised in certain respects, of a human being engaged in computation. Wittgenstein put this point in a striking way:

> Turing's "Machines". These machines are *humans* who calculate. (Wittgenstein 1980, 1096.)

It is a point that Turing was to emphasise, in various forms, again and again. For example:

> A man provided with paper, pencil, and rubber, and subject to strict discipline, is in effect a universal machine. (Turing 1948: 9.)

The electronic stored-program digital computers for which the universal Turing machine was a blueprint are, each of them, computationally equivalent to a Turing machine with a finite tape, and so they too are, in a sense, models of human beings engaged in computation. Turing chose to emphasise this when explaining these electronic machines in a manner suitable for an audience of uninitiates:

> The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer. (Turing 1950: 436).

He makes the point a little more precisely in the technical document containing his preliminary design for the Automatic Computing Engine or ACE. (The ACE was an electronic stored-program computer built at the National Physical Laboratory, London. A pilot version first ran in 1950 and at the time was the fastest computer in the world. The commercial model was called the DEUCE.)

> The class of problems capable of solution by the machine [the ACE] can be defined fairly specifically. They are [a subset of] those problems which can be solved by human clerical labour, working to fixed rules, and without understanding. (Turing 1946: 38-9.)

(Turing went on to characterise the subset in terms of the amount of paper and time available to the human clerk.) It was presumably because he considered the point under discussion to be essential for

understanding the nature of the new electronic machines that he chose to begin the *Programmers' Handbook for the Manchester Computer* with this explanation:

> Electronic computers are intended to carry out any definite rule of thumb process which could have been done by a human operator working in a disciplined but unintelligent manner. (Turing 1951: 1.)

It was not some deficiency of imagination that led Turing to model his computing machines on what could be achieved by a human computer. The purpose for which the Turing machine was invented demanded it. The Entscheidungsproblem is the problem of finding a *humanly executable* procedure of a certain sort, and Turing's aim was precisely to show that there is no such procedure in the case of predicate logic. He proved that no Turing machine can compute the values of the function D that I described earlier, and he argued that his model of human computation is sufficiently general, in the sense that there are no intuitively computable (i.e. effectively calculable) functions that Turing machines are incapable of computing.

The latter claim is, of course, Turing's thesis. Here are two additional formulations of the thesis, from his paper of 1936.

> [T]he "computable numbers" [the numbers whose decimal representations can be generated progressively by a Turing machine] include all numbers which would naturally be regarded as computable. (Turing 1936: 249.)

> It is my contention that these operations [the primitive operations of a Turing machine] include all those which are used in the computation of a number. (Turing 1936: 232.)

(As Turing explains: "Although the subject of this paper is ostensibly the computable *numbers*, it is almost equally easy to define and investigate computable functions ... I have chosen the computable numbers for explicit treatment as involving the least cumbrous technique" (1936: 230).)

To understand these assertions as Turing intended them it is essential to keep in mind that when he uses the words 'computer', 'computable' and 'computation' he employs them not in their modern sense as pertaining to machines but as pertaining to human calculators. Many passages make this obvious.

> Computers always spend just as long in writing numbers down and deciding what to do next as they do in actual multiplications, and it is just the same with ACE ... [T]he ACE will do the work of about 10,000 computers ... Computers will still be employed on small calculations ... (Turing 1947: 116, 120.)

Thus when Turing maintains that every number or function that 'would naturally be regarded as computable' can be calculated by a Turing machine he is asserting not thesis M but a thesis concerning the extent of the effectively calculable numbers and functions. Similarly, when Church writes (in a review

of Post (1936)):

> To define effectiveness as computability by an arbitrary machine, subject to restrictions of finiteness, would seem to be an adequate representation of the ordinary notion (Church 1937b: 43),

he is to be understood not as entertaining some form of thesis M but as endorsing the identification of the effectively calculable functions with those functions that can be calculated by an arbitrary machine whose principles of operation are such as to mimic the actions of a human computer. (There is much that is ‘arbitrary’ about the machines described (independently, in the same year) by Turing and Post, for example the one-dimensional arrangement of the squares of the tape (or in Post’s case, of the ‘boxes’), the absence of a system of addresses for squares of the tape, the choice between a two-way and a one-way infinite tape, and, in Post’s case, the restriction that a square admit of only two possible conditions, blank or marked by a single vertical stroke.)

It is equally important to note that when Turing uses the word ‘machine’ he often means not machine-in-general but, as we would now say, Turing machine. At one point he explicitly draws attention to this idiosyncratic usage:

> The expression "machine process" of course means one which could be carried out by the type of machine I was considering [in Turing 1936]. (Turing 1947: 107).

Thus when, a few pages later, he asserts that "machine processes and rule of thumb processes are synonymous" (1947: 112), he is to be understood as advancing the Church-Turing thesis (and its converse), not a version of thesis M. Unless this idiosyncratic usage is borne in mind, misunderstanding is certain to ensue. Especially liable to mislead are statements like the following, which a casual reader, unaware of Turing’s idiosyncratic usage, might easily mistake for a formulation of thesis M:

> The importance of the universal machine is clear. We do not need to have an infinity of different machines doing different jobs. A single one will suffice. The engineering problem of producing various machines for various jobs is replaced by the office work of "programming" the universal machine to do these jobs. (Turing 1948: 7.)

In context it is perfectly clear that these remarks concern machines equivalent to Turing machines (the passage is embedded in a discussion of LCMs).

Whether or not Turing would, if queried, have assented to thesis M is unknown. There is certainly no textual evidence in favour of the ubiquitous belief that he did so assent.

Bibliography

- Abramson, F.G. 1971. 'Effective Computation over the Real Numbers'. *Twelfth Annual Symposium on Switching and Automata Theory*. Northridge, Calif.: Institute of Electrical and Electronics Engineers.
- Boden, M.A. 1988. *Computer Models of Mind*. Cambridge: Cambridge University Press.
- Boolos, G.S., Jeffrey, R.C. 1980. *Computability and Logic*. 2nd edition. Cambridge: Cambridge University Press.
- Church, A. 1932. 'A set of Postulates for the Foundation of Logic'. *Annals of Mathematics*, second series, 33, 346-366.
- Church, A. 1936a. 'An Unsolvable Problem of Elementary Number Theory'. *American Journal of Mathematics*, 58, 345-363.
- Church, A. 1936b. 'A Note on the Entscheidungsproblem'. *Journal of Symbolic Logic*, 1, 40-41.
- Church, A. 1937a. Review of Turing 1936. *Journal of Symbolic Logic*, 2, 42-43.
- Church, A. 1937b. Review of Post 1936. *Journal of Symbolic Logic*, 2, 43.
- Church, A. 1941. *The Calculi of Lambda-Conversion*. Princeton: Princeton University Press.
- Churchland, P.M. 1988. *Matter and Consciousness*. Cambridge, Mass.: MIT Press.
- Churchland, P.M., Churchland, P.S. 1983. 'Stalking the Wild Epistemic Engine'. *Nous*, 17, 5-18.
- Churchland, P.M., Churchland, P.S. 1990. 'Could a Machine Think?'. *Scientific American*, 262 (Jan.), 26-31.
- Cleland, C.E. 1993. 'Is the Church-Turing Thesis True?'. *Minds and Machines*, 3, 283-312.
- Copeland, B.J., Sylvan, R. 1997. 'Computability: A Heretical Approach'. Forthcoming.
- Curry, H.B. 1929. 'An Analysis of Logical Substitution'. *American Journal of Mathematics*, 51, 363-384.
- Curry, H.B. 1930. 'Grundlagen der kombinatorischen Logik'. *American Journal of Mathematics*, 52, 509-536, 789-834.
- Curry, H.B. 1932. 'Some Additions to the Theory of Combinators'. *American Journal of Mathematics*, 54, 551-558.
- da Costa, N.C.A., Doria, F.A. 1991. 'Classical Physics and Penrose's Thesis'. *Foundations of Physics Letters*, 4, 363-374.
- da Costa, N.C.A., Doria, F.A. 1994. 'Undecidable Hopf Bifurcation with Undecidable Fixed Point'. *International Journal of Theoretical Physics*, 33, 1913-1931.
- Dennett, D.C. 1991. *Consciousness Explained*. Boston: Little, Brown.
- Deutsch, D. 1985. 'Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer'. *Proceedings of the Royal Society*, Series A, 400, 97-117.
- Doyle, J. 1982. 'What is Church's Thesis? An Outline.' Laboratory for Computer Science, MIT.
- Gandy, R. 1980. 'Church's Thesis and Principles for Mechanisms'. In Barwise, J., Keisler, H.J., Kunen, K. (eds) 1980. *The Kleene Symposium*. Amsterdam: North-Holland.
- Gandy, R. 1988. 'The Confluence of Ideas in 1936'. In Herken, R. (ed.) 1988. *The Universal Turing Machine: A Half-Century Survey*. Oxford: Oxford University Press.
- Gödel, K. 1934. 'On Undecidable Propositions of Formal Mathematical Systems'. Lecture notes taken by Kleene and Rosser at the Institute for Advanced Study. Reprinted in Davis, M. (ed.) 1965. *The Undecidable*. New York: Raven.
- Gödel, K. 1936. 'Über die Lange von Beweisen'. *Ergebnisse eines mathematischen Kolloquiums*, 7, 23-24.

- Gregory, R.L. 1987. *The Oxford Companion to the Mind*. Oxford: Oxford University Press.
- Guttenplan, S. 1994. *A Companion to the Philosophy of Mind*. Oxford: Blackwell.
- Harel, D. 1992. *Algorithmics: The Spirit of Computing*. Reading, Mass.: Addison-Wesley.
- Herbrand, J. 1932. 'Sur la non-contradiction de l'arithmetique. *Journal fur die reine und angewandte Mathematik*, 166, 1-8.
- Hilbert, D., Ackermann, W. 1928. *Grundzuge der Theoretischen Logik*. Berlin: Springer.
- Hogarth, M.L. 1994. 'Non-Turing Computers and Non-Turing Computability'. *PSA 1994*, vol.1, 126-138.
- Johnson-Laird, P. 1987. 'How Could Consciousness Arise from the Computations of the Brain?'. In Blakemore, C., Greenfield, S. (eds) 1987. *Mindwaves*. Oxford: Basil Blackwell.
- Kalmar, L. 1959. 'An Argument Against the Plausibility of Church's Thesis'. In Heyting, A. (ed.) 1959. *Constructivity in Mathematics*. Amsterdam: North-Holland, pp.72-80.
- Kleene, S.C. 1935. 'A Theory of Positive Integers in Formal Logic'. *American Journal of Mathematics*, 57, 153-173, 219-244.
- Kleene, S.C. 1936. 'Lambda-Definability and Recursiveness'. *Duke Mathematical Journal*, 2, 340-353.
- Kleene, S.C. 1952. *Introduction to Metamathematics*. Amsterdam: North-Holland.
- Kleene, S.C. 1967. *Mathematical Logic*. New York: Wiley.
- Markov, A.A. 1960. 'The Theory of Algorithms'. *American Mathematical Society Translations*, series 2, 15, 1-14.
- McArthur, R.P. 1991. *From Logic to Computing*. Belmont, Calif.: Wadsworth.
- Mendelson, E. 1963. 'On Some Recent Criticism of Church's Thesis'. *Notre Dame Journal of Formal Logic*, 4, 201-205.
- Mendelson, E. 1964. *Introduction to Mathematical Logic*. New York: Van Nostrand.
- Odifreddi, P. 1989. *Classical Recursion Theory*. Amsterdam: North-Holland.
- Post, E.L. 1936. 'Finite Combinatory Processes - Formulation 1'. *Journal of Symbolic Logic*, 1, 103-105.
- Post, E.L. 1936. 'Finite Combinatory Processes - Formulation 1'. *Journal of Symbolic Logic*, 1, 103-5.
- Post, E.L. 1943. 'Formal Reductions of the General Combinatorial Decision Problem'. *American Journal of Mathematics*, 65, 197-215.
- Post, E.L. 1946. 'A Variant of a Recursively Unsolvable Problem'. *Bulletin of the American Mathematical Society*, 52, 264-268.
- Pour-El, M.B., Richards, I. 1979. 'A Computable Ordinary Differential Equation Which Possesses No Computable Solution'. *Annals of Mathematical Logic*, 17, 61-90.
- Pour-El, M.B., Richards, I. 1981. 'The Wave Equation with Computable Initial Data such that its Unique Solution is not Computable'. *Advances in Mathematics*, 39, 215-239.
- Scarpellini, B. 1963. 'Zwei Unentscheitbare Probleme der Analysis', *Zeitschrift fur mathematische Logik und Grundlagen der Mathematik*, 9, 265-289.
- Schönfinkel, M. 1924. 'Uber die Bausteine der mathematischen'. *Mathematische Annalen*, 92, 305-316.
- Searle, J. 1992. *The Rediscovery of the Mind*. Cambridge, Mass.: MIT Press.
- Shepherdson, J.C., Sturgis, H.E. 1963. 'Computability of Recursive Functions'. *Journal of the*

ACM, 10, 217-255.

- Siegelmann, H.T., Sontag, E.D. 1992. ‘On the Computational Power of Neural Nets’. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 440-449.
- Smolensky, P. 1988. ‘On the Proper Treatment of Connectionism’. *Behavioural and Brain Sciences*, 11, 1-23.
- Stannett, M. 1990. ‘X-Machines and the Halting Problem: Building a Super-Turing Machine’. *Formal Aspects of Computing*, 2, 331-341.
- Sterelny, K. 1990. *The Representational Theory of Mind*. Oxford: Basil Blackwell.
- Stewart, I. 1991. ‘Deciding the Undecidable’. *Nature*, 352, 664-5.
- Turing, A.M. 1936 .’On Computable Numbers, with an Application to the Entscheidungsproblem’. Proceedings of the London Mathematical Society, Series 2, 42 (1936-37), pp.230-265.
- Turing, A.M. 1946. ‘Proposal for Development in the Mathematics Division of an Automatic Computing Engine (ACE)’. In Carpenter, B.E., Doran, R.W. (eds) 1986. *A.M. Turing’s ACE Report of 1946 and Other Papers*. Cambridge, Mass.: MIT Press.
- Turing, A.M. 1947. ‘Lecture to the London Mathematical Society on 20 February 1947’. In Carpenter, B.E., Doran, R.W. (eds) 1986. *A.M. Turing’s ACE Report of 1946 and Other Papers*. Cambridge, Mass.: MIT Press.
- Turing, A.M. 1948. ‘Intelligent Machinery’. National Physical Laboratory Report. In Meltzer, B., Michie, D. (eds) 1969. Machine Intelligence 5. Edinburgh: Edinburgh University Press.
- Turing, A.M. 1950. ‘Computing Machinery and Intelligence’. *Mind* 59, 433-460.
- Turing, A.M. 1951. ‘Programmers’ Handbook for the Manchester Electronic Computer’. University of Manchester Computing Laboratory.
- Wittgenstein, L. 1980. *Remarks on the Philosophy of Psychology*. Vol.1. Oxford: Blackwell.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Church, Alonzo | computability theory | function: recursive | mind: philosophy of | Turing, Alan | Turing machine

B. Jack Copeland
bjcopeland@canterbury.ac.nz

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 8, 1997
Content last modified: January 8, 1997

Alan M. Turing

Alan Turing (1912-1954) never described himself as a philosopher, but his 1950 paper "Computing Machinery and Intelligence" is one of the most frequently cited in modern philosophical literature. It gave a fresh approach to the traditional mind-body problem, by relating it to the mathematical concept of computability he himself had introduced in his 1936-7 paper "On computable numbers, with an application to the Entscheidungproblem." His work can be regarded as the foundation of computer science and of the artificial intelligence program.

---

1. Outline of Life

Alan Turing's short and extraordinary life has attracted wide interest. It has inspired his mother's memoir (E. S. Turing 1959), a detailed biography (Hodges 1983), a play and television film (Whitemore 1986), and various other works of fiction and art.

There are many reasons for this interest, but one is that in every sphere of his life and work he made unexpected connections between apparently unrelated areas. His central contribution to science and philosophy came through his treating the subject of symbolic logic as a new branch of applied mathematics, giving it a physical and engineering content. Unwilling or unable to remain within any

standard role or department of thought, Alan Turing continued a life full of incongruity. Though a shy, boyish, man, he had a pivotal role in world history through his role in Second World War cryptology. Though the founder of the dominant technology of the twentieth century, he variously impressed, charmed or disturbed people with his unworldly innocence and his dislike of moral or intellectual compromise.

Alan Mathison Turing was born in London, 23 June 1912, to upper-middle-class British parents. His schooling was of a traditional kind, dominated by the British imperial system, but from earliest life his fascination with the scientific impulse -- expressed by him as finding the ‘commonest in nature’ -- found him at odds with authority. His scepticism, and disrespect for worldly values, were never tamed and became ever more confidently eccentric. His moody humour swung between gloom and vivacity. His life was also notable as that of a gay man with strong emotions and a growing insistence on his identity.

His first true home was at King’s College, Cambridge University, noted for its progressive intellectual life centred on J. M. Keynes. Turing studied mathematics with increasing distinction and was elected a Fellow of the college in 1935. This appointment was followed by a remarkable and sudden début in an area where he was an unknown figure: that of mathematical logic. The paper "On Computable Numbers..." (Turing 1936-7) was his first and perhaps greatest triumph. It gave a definition of computation and an absolute limitation on what computation could achieve, which makes it the founding work of modern computer science. It led him to Princeton for more advanced work in logic and other branches of mathematics. He had the opportunity to remain in the United States, but chose to return to Britain in 1938, and was immediately recruited for the British communications war.

From 1939 to 1945 Turing was almost totally engaged in the mastery of the German enciphering machine, Enigma, and other cryptological investigations at now-famous Bletchley Park, the British government’s wartime communications headquarters. Turing made a unique logical contribution to the decryption of the Enigma and became the chief scientific figure, with a particular responsibility for reading the U-boat communications. As such he became a top-level figure in Anglo-American liaison, and also gained exposure to the most advanced electronic technology of the day.

Combining his ideas from mathematical logic, his experience in cryptology, and some practical electronic knowledge, his ambition, at the end of the war in Europe, was to create an electronic computer in the full modern sense. His plans, commissioned by the National Physical Laboratory, London, were overshadowed by the more powerfully supported American projects. Turing also laboured under the disadvantage that his wartime achievements remained totally secret. His ideas led the field in 1946, but this was little recognised. Frustrated in his work, he emerged as a powerful marathon runner, and almost qualified for the British team in the 1948 Olympic games.

Turing’s motivations were scientific rather than industrial or commercial, and he soon returned to the theoretical limitations of computation, this time focussing on the comparison of the power of computation and the power of the human brain. His contention was that the computer, when properly programmed, could rival the brain. It founded the ‘Artifical Intelligence’ program of coming decades.

In 1948 he moved to Manchester University, where he partly fulfilled the expectations placed upon him to plan software for the pioneer computer development there, but still remained a free-ranging thinker. It was here that his famous 1950 paper, "Computing Machinery and Intelligence," (Turing 1950b) was written. In 1951 he was elected a Fellow of the Royal Society for his 1936 achievement, yet at the same time he was striking into entirely new territory with a mathematical theory of biological morphogenesis (Turing 1952).

This work was interrupted by Alan Turing’s arrest in February 1952 for his sexual affair with a young Manchester man, and he was obliged, to escape imprisonment, to undergo the injection of oestrogen intended to negate his sexual drive. He was disqualified from continuing secret cryptological work. His general libertarian attitude was enhanced rather than suppressed by the criminal trial, and his intellectual individuality also remained as lively as ever. While remaining formally a Reader in the Theory of Computing, he not only embarked on more ambitious applications of his biological theory, but advanced new ideas for fundamental physics.

For this reason his death, on 7 June 1954, at his home in Wilmslow, Cheshire, came as a general surprise. In hindsight it is obvious that Turing’s unique status in Anglo-American secret communication work meant that there were pressures on him of which his contemporaries were unaware; there was certainly another ‘security’ conflict with government in 1953 (Hodges 1983, p. 483). Some commentators, e.g. Dawson (1985), have argued that assassination should not be ruled out. But he had spoken of suicide, and his death, which was by cyanide poisoning, was most likely by his own hand, contrived so as to allow those who wished to do so to believe it a result of his penchant for chemistry experiments. The symbolism of its dramatic element -- a partly eaten apple -- has continued to haunt the intellectual Eden from which Alan Turing was expelled.

2. The Turing Machine and Computability

Alan Turing drew much between 1928 and 1933 from the work of the mathematical physicist and populariser A. S. Eddington, from J. von Neumann’s account of the foundations of quantum mechanics, and then from Bertrand Russell’s mathematical logic. Meanwhile, his lasting fascination with the problems of mind and matter was heightened by emotional elements in his own life (Hodges 1983, p. 63). In 1934 he graduated with an outstanding degree in mathematics from Cambridge University, followed by a successful dissertation in probability theory which won him a Fellowship of King’s College, Cambridge, in 1935. This was the background to his learning, also in 1935, of the problem which was to make his name.

It was from the lectures of the topologist M. H. A. (Max) Newman in that year that he learnt of Gödel’s 1931 proof of the formal incompleteness of logical systems rich enough to include arithmetic, and of the outstanding problem in the foundations of mathematics as posed by Hilbert: the "Entscheidungsproblem" (decision problem). Was there a method by which it could be decided, for any given mathematical proposition, whether or not it was provable?

The principal difficulty of this question lay in giving an unassailably correct and general definition of what was meant by such expressions as ‘definite method’ or ‘effective procedure.’ Turing worked on this alone for a year until April 1936; independence and isolation was to be both his strength, in formulating original ideas, and his weakness, when it came to promoting and implementing them.

The word ‘mechanical’ had often been used of the formalist approach lying behind Hilbert’s problem, and Turing seized on the concept of the *machine.*Turing’s solution lay in defining what was soon to be named the *Turing machine.* With this he defined the concept of ‘the mechanical’ in terms of simple atomic operations. The Turing machine formalism was modelled on the teleprinter, slightly enlarged in scope to allow a paper tape that could move in both directions and a ‘head’ that could read, erase and print new symbols, rather than only read and punch permanent holes.

The Turing machine is ‘theoretical,’ in the sense that it is not intended actually to be engineered (there being no point in doing so), although it is essential that its atomic components (the paper tape, movement to left and right, testing for the presence of a symbol) are such as *could* actually be implemented. The whole point of the formalism is to reduce the concept of ‘method’ to simple operations that can unquestionably be ‘effected.’

Nevertheless Turing’s purpose was to embody the most general mechanical process as carried out by a *human being.* His analysis began not with any existing computing machines, but with the picture of a child’s exercise book marked off in squares. From the beginning, the Turing machine concept aimed to capture what the *human mind* can do when carrying out a procedure.

In speaking of ‘the’ Turing machine it should be made clear that there are *infinitely many* Turing machines, each corresponding to a different method or procedure, by virtue of having a different ‘table of behaviour.’ Nowadays it is almost impossible to avoid imagery which did not exist in 1936: that of the computer. In modern terms, the ‘table of behaviour’ of a Turing machine is equivalent to a computer program.

If a Turing machine corresponds to a computer program, what is the analogy of the computer? It is what Turing described as a *universal* machine (Turing 1936-7, p. 241). Again, there are *infinitely many* universal Turing machines, forming a subset of Turing machines; they are those machines with ‘tables of behaviour’ complex enough to read the tables of other Turing machines, and then do what those machines would have done. If this seems strange, note the modern parallel that any computer can be simulated by software on another computer. The way that tables can read and simulate the effect of other tables is crucial to Turing’s theory, going far beyond Babbage’s ideas of a hundred years earlier. It also shows why Turing’s ideas go to the heart of the modern computer, in which it is essential that programs are themselves a form of data which can be manipulated by other programs. But the reader must always remember that in 1936 there were no such computers; indeed the modern computer arose *out of* the formulation of ‘behaving mechanically’ that Turing found in this work.

Turing’s machine formulation allowed the precise definition of the *computable:* namely, as what can be done by a Turing machine acting alone. More exactly, computable operations are those which can be

effected by what Turing called *automatic* machines. The crucial point here is that the action of an automatic Turing machine is totally determined by its ‘table of behaviour’. (Turing also allowed for ‘choice machines’ which call for human inputs, rather than being totally determined.) Turing then proposed that this definition of ‘computable’ captured precisely what was intended by such words as ‘definite method, procedure, mechanical process’ in stating the *Entscheidungsproblem.*

In applying his machine concept to the *Entscheidungsproblem,* Turing took the step of defining *computable numbers.* These are those real numbers, considered as infinite decimals, say, which it is possible for a Turing machine, starting with an empty tape, to print out. For example, the Turing machine which simply prints the digit 1 and moves to the right, then repeats that action for ever, can thereby compute the number .1111111... A more complicated Turing machine can compute the infinite decimal expansion of π.

Turing machines, like computer programs, are countable; indeed they can be ordered in a complete list by a kind of alphabetical ordering of their ‘tables of behaviour’. Turing did this by encoding the tables into ‘description numbers’ which can then be ordered in magnitude. Amongst this list, a subset of them (those with ‘satisfactory’ description numbers) are the machines which have the effect of printing out infinite decimals. It is readily shown, using a ‘diagonal’ argument first used by Cantor and familiar from the discoveries of Russell and Gödel, that there can be no Turing machine with the property of deciding whether a description number is satisfactory or not. The argument can be presented as follows. Suppose that such a Turing machine exists. Then it is possible to construct a new Turing machine which works out in turn the Nth digit from the Nth machine possessing a satisfactory description number. This new machine then prints an Nth digit differing from that digit. As the machine proceeds, it prints out an infinite decimal, and therefore has a ‘satisfactory’ description number. Yet this number must by construction differ from the outputs of every Turing machine with a satisfactory description number. This is a contradiction, so the hypothesis must be false (Turing 1936-7, p. 246). From this, Turing was able to answer Hilbert’s *Entscheidungsproblem* in the negative: there can be no such general method.

Turing’s proof can be recast in many ways, but the core idea depends on the *self-reference* involved in a machine operating on symbols, which is itself described by symbols and so can operate on its own description. Indeed, the self-referential aspect of the theory can be highlighted by a different form of the proof, which Turing preferred (Turing 1936-7, p. 247). Suppose that such a machine for deciding satisfactoriness does exist; then apply it to its own description number. A contradiction can readily be obtained. However, the ‘diagonal’ method has the advantage of bringing out the following: that a real number may be *defined* unambiguously, yet be *uncomputable.* It is a non-trivial discovery that whereas some infinite decimals (e.g. π) may be encapsulated in a finite table, other infinite decimals (in fact, almost all) cannot. Likewise there are decision problems such as ‘is this number prime?’ in which infinitely many answers are wrapped up in a finite recipe, while there are others (again, almost all) which are not, and must be regarded as requiring infinitely many different methods. ‘Is this a provable proposition?’ belongs to the latter category.

This is what Turing established, and into the bargain the remarkable fact that anything that *is* computable can in fact be computed by *one* machine, a universal Turing machine.

It was vital to Turing's work that he justified the definition by showing that it encompassed the most general idea of 'method'. For if it did not, the *Entscheidungproblem* remained open: there might be some more powerful type of method than was encompassed by Turing computability. One justification lay in showing that the definition included many processes a mathematician would consider to be natural in computation (Turing 1936-7, p. 254). Another argument involved a human calculator following written instruction notes. (Turing 1936-7, p. 253). But in a bolder argument, the one he placed first, he considered an 'intuitive' argument appealing to the *states of mind* of a human computer. (Turing 1936-7, p. 249). The entry of 'mind' into his argument was highly significant, but at this stage it was only a mind following a rule.

To summarise: Turing found, and justified on very general and far-reaching grounds, a precise mathematical formulation of the conception of a general process or method. His work, as presented to Newman in April 1936, argued that his formulation of 'computability' encompassed 'the possible processes which can be carried out in computing a number.' (Turing 1936-7, p. 232). This opened up new fields of discovery both in practical computation, and in the discussion of human mental processes. However, although Turing had worked as what Newman called 'a confirmed solitary' (Hodges 1983, p 113), he soon learned that he was not alone in what Gandy (1988) has called 'the confluence of ideas in 1936.'

The Princeton logician Alonzo Church had slightly outpaced Turing in finding a satisfactory definition of what he called 'effective calculability.' Church's definition required the logical formalism of the *lambda-calculus.* This meant that from the outset Turing's achievement merged with and superseded the formulation of *Church's Thesis,* namely the assertion that the lambda-calculus formalism correctly embodied the concept of effective process or method. Very rapidly it was shown that the mathematical scope of Turing computability coincided with Church's definition (and also with the scope of the *general recursive functions* defined by Gödel). Turing wrote his own statement (Turing 1939, p. 166) of the conclusions that had been reached in 1938; it is in the Ph.D. thesis that he wrote under Church's supervision, and so this statement is the nearest we have to a joint statement of the 'Church-Turing thesis':

> A function is said to be 'effectively calculable' if its values can be found by some purely mechanical process. Although it is fairly easy to get an intuitive grasp of this idea, it is nevertheless desirable to have some more definite, mathematically expressible definition. Such a definition was first given by Gödel at Princeton in 1934... These functions were described as 'general recursive' by Gödel... Another definition of effective calculability has been given by Church... who identifies it with lambda-definability. The author [i.e. Turing] has recently suggested a definition corresponding more closely to the intuitive idea... It was stated above that 'a function is effectively calculable if its values can be found by a purely mechanical process.' We may take this statement literally, understanding by a purely mechanical process one which could be carried out by a machine. It is possible to give a mathematical description, in a certain normal form, of the structures of these machines. The development of these ideas leads to the author's definition of a computable function, and to

> an identification of computability with effective calculability. It is not difficult, though somewhat laborious, to prove that these three definitions are equivalent.

Church accepted that Turing's definition gave a compelling, intuitive reason for why Church's thesis was true. The recent exposition by Davis (2000) emphasises that Gödel also was convinced by Turing's argument that an absolute concept had been identified (Gödel 1946). The situation has not changed since 1937. (For further comment, see the article on the Church-Turing Thesis.)

Turing himself did little to evangelise his formulation in the world of mathematical logic and early computer science. The textbooks of Davis (1958) and Minsky (1967) did more. Nowadays Turing computability is often reformulated (e.g. in terms of 'register machines'). However, computer simulations (e.g., Turing's World, from Stanford) have brought Turing's original imagery to life.

Turing's work also opened new areas for decidability questions within pure mathematics. From the 1970s, Turing machines also took on new life in the development of *complexity theory,* and as such underpin one of the most important research areas in computer science. This development exemplifies the lasting value of Turing's special quality of giving concrete illustration to abstract concepts.

3. The Logical and the Physical

As put by Gandy (1988), Turing's paper was 'a paradigm of philosophical analysis,' refining a vague notion into a precise definition. But it was more than being an analysis *within* the world of mathematical logic: in Turing's thought the question that constantly recurs both theoretically and practically is the relationship of the logical Turing machine to the physical world.

'Effective' means *doing,* not merely imagining or postulating. At this stage neither Turing nor any other logician made a serious investigation into the physics of such 'doing.' But Turing's image of a teleprinter-like machine does inescapably refer to something that could actually be physically 'done.' His concept is a distillation of the idea that one can only 'do' one simple action, or finite number of simple actions, at a time. How 'physical' a concept is it?

The tape never holds more than a finite number of marked squares at any point in a computation. Thus it can be thought of as being finite, but always capable of further extension as required. Obviously this unbounded extendibility is unphysical, but the definition is still of practical use: it means that anything done on a finite tape, however large, is computable. (Turing himself took such a finitistic approach when explaining the practical relevance of computability in his 1950 paper.) One aspect of Turing's formulation, however, involves absolute finiteness: the table of behaviour of a Turing machine must be finite, since Turing allows only a finite number of 'configurations' of a Turing machine, and only a finite repertoire of symbols which can be marked on the tape. This is essentially equivalent to allowing only computer programs with finite lengths of code.

'Calculable by finite means' was Turing's characterisation of computability, which he justified with the argument that 'the human memory is necessarily limited.' (Turing 1936-7, p. 231). The whole point of his definition lies in encoding infinite potential effects, (e.g. the printing of an infinite decimal) into finite 'tables of behaviour'. There would be no point in allowing machines with infinite 'tables of behaviour'. It is obvious, for instance, that any real number could be printed by such a 'machine', by letting the Nth configuration be 'programmed' to print the Nth digit, for example. Such a 'machine' could likewise store any countable number of statements about all possible mathematical expressions, and so make the *Entscheidungsproblem* trivial.

Church (1937), when reviewing Turing's paper while Turing was in Princeton under his supervision, actually gave a bolder characterisation of the Turing machine as an *arbitrary finite machine.*

> The author [i.e. Turing] proposes as a criterion that an infinite sequence of digits 0 and 1 be "computable" that it shall be possible to devise a computing machine, occupying a finite space and with working parts of finite size, which will write down the sequence to any desired number of terms if allowed to run for a sufficiently long time. As a matter of convenience, certain further restrictions are imposed on the character of the machine, but these are of such a nature as obviously to cause no loss of generality -- in particular, a human calculator, provided with pencil and paper and explicit instructions, can be regarded as a kind of Turing machine.

Church (1940) repeated this characterisation. Turing neither endorsed it nor said anything to contradict it, leaving the general concept of 'machine' itself undefined. The work of Gandy (1980) did more to justify this characterisation, by refining the statement of what is meant by 'a machine.' His results support Church's statement; they also argue strongly for the view that natural attempts to extend the notion of computability lead to trivialisation: if Gandy's conditions on a 'machine' are significantly weakened then every real number becomes calculable (Gandy 1980, p. 130ff.). (For a different interpretation of Church's statement, see the article on the Church-Turing Thesis.)

Turing did not explicitly discuss the question of the *speed* of his elementary actions. It is left implicit in his discussion, by his use of the word 'never,' that it is not possible for infinitely many steps to be performed in a finite time. Others have explored the effect of abandoning this restriction. Davies (2001), for instance, describes a 'machine' with an infinite number of parts, requiring components of arbitrarily small size, running at arbitrarily high speeds. Such a 'machine' could perform uncomputable tasks. Davies emphasises that such a machine cannot be built in our own physical world, but argues that it could be constructed in a universe with different physics. To the extent that it rules out such 'machines', the Church-Turing thesis must have at least some physical content.

True physics is quantum-mechanical, and this implies a different idea of matter and action from Turing's purely classical picture. It is perhaps odd that Turing did not point this out in this period, since he was well versed in quantum physics. Instead, the analysis and practical development of quantum computing was left to the 1980s. Quantum computation, using the evolution of wave-functions rather than classical machine states, is the most important way in which Turing machine model has been challenged. The

standard formulation of quantum computing (Deutsch 1985, following Feynman 1982) does not predict anything beyond computable effects, although within the realm of the computable, quantum computations may be very much more efficient than classical computations. It is possible that a deeper understanding of quantum mechanical physics may further change the picture of what can be physically 'done.'

4. The Uncomputable

Turing turned to the exploration of the *uncomputable* for his Princeton Ph.D. thesis (1938), which then appeared as *Systems of Logic based on Ordinals* (Turing 1939).

It is generally the view, as expressed by Feferman (1988), that this work was a diversion from the main thrust of his work. But from another angle, as expressed in (Hodges 1997), one can see Turing's development as turning naturally from considering the mind when following a rule, to the action of the mind when *not* following a rule. In particular this 1938 work considered the mind when seeing the truth of one of Gödel's true but formally unprovable propositions, and hence going beyond rules based on the axioms of the system. As Turing expressed it (Turing 1939, p. 198), there are 'formulae, seen intuitively to be correct, but which the Gödel theorem shows are unprovable in the original system.' Turing's theory of 'ordinal logics' was an attempt to 'avoid as far as possible the effects of Gödel's theorem' by studying the effect of adding Gödel sentences as new axioms to create stronger and stronger logics. It did not reach a definitive conclusion.

In his investigation, Turing introduced the idea of an 'oracle' capable of performing, as if by magic, an uncomputable operation. Turing's oracle cannot be considered as some 'black box' component of a new class of machines, to be put on a par with the primitive operations of reading single symbols, as has been suggested by (Copeland 1998). An oracle is *infinitely more powerful* than anything a modern computer can do, and nothing like an elementary component of a computer. Turing defined 'oracle-machines' as Turing machines with an additional configuration in which they 'call the oracle' so as to take an uncomputable step. But these oracle-machines are *not purely mechanical.* They are only partially mechanical, like Turing's choice-machines. Indeed the *whole point* of the oracle-machine is to explore the realm of what *cannot* be done by purely mechanical processes. Turing emphasised (Turing 1939, p. 173):

> We shall not go any further into the nature of this oracle apart from saying that it cannot be a machine.

Turing's oracle can be seen simply as a mathematical tool, useful for exploring the mathematics of the uncomputable. The idea of an oracle allows the formulation of questions of *relative* rather than absolute computability. Thus Turing opened new fields of investigation in mathematical logic. However, there is also a possible interpretation in terms of human cognitive capacity. On this interpretation, the oracle is related to the 'intuition' involved in seeing the truth of a Gödel statement. M. H. A. Newman, who introduced Turing to mathematical logic and continued to collaborate with him, wrote in (Newman 1955) that the oracle resembles a mathematician 'having an idea', as opposed to using a mechanical method. However, Turing's oracle cannot actually be *identified* with a human mental faculty. It is too powerful: it

immediately supplies the answer as to whether any given Turing machine is 'satisfactory,' something no human being could do. On the other hand, anyone hoping to see mental 'intuition' captured completely by an oracle, must face the difficulty that Turing showed how his argument for the incompleteness of Turing machines could be applied with equal force to oracle-machines (Turing 1939, p. 173). This point has been emphasised by Penrose (1994, p. 380). Newman's comment might better be taken to refer to the different oracle suggested later on (Turing 1939, p. 200), which has the property of recognising 'ordinal formulae.' One can only safely say that Turing's interest at this time in uncomputable operations appears in the *general setting* of studying the mental 'intuition' of truths which are not established by following mechanical processes (Turing 1939, p. 214ff.).

In Turing's presentation, intuition is in practice present in every part of a mathematician's thought, but when mathematical proof is formalised, intuition has an explicit manifestation in those steps where the mathematician sees the truth of a formally unprovable statement. Turing did not offer any suggestion as to what he considered the brain was physically doing in a moment of such 'intuition'; indeed the word 'brain' did not appear in his writing in this era. This question is of interest because of the views of Penrose (1989, 1990, 1994, 1996) on just this issue: Penrose holds that the ability of the mind to see formally unprovable truths shows that there must be uncomputable physical operations in the brain. It should be noted that there is widespread disagreement about whether the human mind is really seeing the truth of a Gödel sentence; see for instance the discussion in (Penrose 1990) and the reviews following it. However Turing's writing at this period accepted without criticism the concept of intuitive recognition of the truth.

It was also at this period that Turing met Wittgenstein, and there is a full record of their 1939 discussions on the foundations of mathematics in (Diamond 1976). To the disappointment of many, there is no record of any discussions between them, verbal or written, on the problem of Mind.

In 1939 Turing's various energetic investigations were broken off for war work. This did, however, have the positive feature of leading Turing to turn his universal machine into the practical form of the modern digital computer.

5. Building a Universal Machine

When apprised in 1936 of Turing's idea for a universal machine, Turing's contemporary and friend, the economist David Champernowne, reacted by saying that such a thing was impractical; it would need 'the Albert Hall.' If built from relays as then employed in telephone exchanges, that might indeed have been so, and Turing made no attempt at it. However, in 1937 Turing did work with relays on a smaller machine with a special cryptological function (Hodges 1983, p. 138). World history then led Turing to his unique role in the Enigma problem, to his becoming the chief figure in the mechanisation of logical procedures, and to his being introduced to ever faster and more ambitious technology as the war continued.

After 1942, Turing learnt that electronic components offered the speed, storage capacity and logical functions required to be effective as 'tapes' and instruction tables. So from 1945, Turing tried to use

electronics to turn his universal machine into practical reality. Turing rapidly composed a detailed plan for a modern stored-program computer: that is, a computer in which data and instructions are stored and manipulated alike. Turing's ideas led the field, although his report of 1946 postdated von Neumann's more famous EDVAC report (von Neumann 1945). It can however be argued, as does Davis (2000), that von Neumann gained his fundamental insight into the computer through his pre-war familiarity with Turing's logical work. At the time, however, these basic principles were not much discussed. The difficulty of engineering the electronic hardware dominated everything.

It therefore escaped observers that Turing was ahead of von Neumann and everyone else on the future of software, or as he called it, the 'construction of instruction tables.' Turing (1946) foresaw at once:

> Instruction tables will have to be made up by mathematicians with computing experiences and perhaps a certain puzzle-solving ability. There will probably be a great deal of work to be done, for every known process has got to be translated into instruction table form at some stage.
>
> The process of constructing instruction tables should be very fascinating. There need be no real danger of it ever becoming a drudge, for any processes that are quite mechanical may be turned over to the machine itself.

These remarks, reflecting the universality of the computer, and its ability to manipulate its own instructions, correctly described the future trajectory of the computer industry. However, Turing had in mind something greater: 'building a brain.'

6. Building a Brain

The provocative words 'building a brain' from the outset announced the relationship of Turing's technical computer engineering to a philosophy of Mind. Even in 1936, Turing had given an interpretation of computability in terms of 'states of mind'. His war work had shown the astounding power of the computable in mechanising expert human procedures and judgments. From 1941 onwards, Turing had also discussed the mechanisation of chess-playing and other 'intelligent' activities with his colleagues at Bletchley Park (Hodges 1983, p. 213). But more profoundly, it appears that Turing emerged in 1945 with a conviction that computable operations were sufficient to embrace *all* mental functions performed by the brain. As will become clear from the ensuing discussion, the uncomputable 'intuition' of 1938 disappeared from Turing's thought, and was replaced by new ideas all lying within the realm of the computable. This change shows even in the technical prospectus of (Turing 1946), where Turing referred to the possibility of making a machine calculate chess moves, and then continued:

> This ... raises the question 'Can a machine play chess?' It could fairly easily be made to play a rather bad game. It would be bad because chess requires intelligence. We stated ... that the machine should be treated as entirely without intelligence. There are indications however that it is possible to make the machine display intelligence at the risk of its making

> occasional serious mistakes. By following up this aspect the machine could probably be made to play very good chess.

The puzzling reference to 'mistakes' is made clear by a talk Turing gave a year later (Turing 1947), in which the issue of mistakes is linked to the issue of the significance of seeing the truth of formally unprovable statements.

> ...I would say that fair play must be given to the machine. Instead of it giving no answer we could arrange that it gives occasional wrong answers. But the human mathematician would likewise make blunders when trying out new techniques... In other words then, if a machine is expected to be infallible, it cannot also be intelligent. There are several mathematical theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretence at infallibility.

Turing's post-war view was that mathematicians make mistakes, and so do not in fact see the truth infallibly. Once the possibility of mistakes is admitted, Gödel's theorem become irrelevant. Mathematicians and computers alike apply computable processes to the problem of judging the correctness of assertions; both will therefore sometimes err, since seeing the truth is known not to be a computable operation, but there is no reason why the computer need do worse than the mathematician. This argument is still very much alive. For instance, Davis (2000) endorses Turing's view and attacks Penrose (1989, 1990, 1994, 1996) who argues against the significance of human error on the grounds of a Platonist account of mathematics.

Turing also pursued more constructively the question of how computers could be made to perform operations which did not appear to be 'mechanical' (to use common parlance). His guiding principle was that it should be possible to simulate the operation of human brains. In an unpublished report (Turing 1948), Turing explained that the question was that of how to simulate 'initiative' in addition to 'discipline' -- comparable to the need for 'intuition' as well as mechanical ingenuity expressed in his pre-war work. He announced ideas for how to achieve this: he thought 'initiative' could arise from systems where the algorithm applied is not consciously designed, but is arrived at by some other means. Thus, he now seemed to think that the mind when *not* actually following any conscious rule or plan, was nevertheless carrying out some computable process.

He suggested a range of ideas for systems which could be said to modify their own programs. These ideas included nets of logical components ('unorganised machines') whose properties could be 'trained' into a desired function. Thus, as expressed by (Ince 1989), he predicted neural networks. However, Turing's nets did not have the 'layered' structure of the neural networks that were to be developed from the 1950s onwards. By the expression 'genetical or evolutionary search', he also anticipated the 'genetic algorithms' which since the late 1980s have been developed as a less closely structured approach to self-modifying programs. Turing's proposals were not well developed in 1948, and at a time when electronic computers were only barely in operation, could not have been. Fresh attention to them has been drawn by Copeland and Proudfoot (1996), and they have now have been tried out (Teuscher 2001).

It is important to note that Turing identified his prototype neural networks and genetic algorithms as *computable*. This has to be emphasised since the word ‘nonalgorithmic’ is often now confusingly employed for computer operations that are not explicitly planned. Indeed, his ambition was explicit: he himself wanted to implement them as programs on a computer. Using the term Universal Practical Computing Machine for what is now called a digital computer, he wrote in (Turing 1948):

> It should be easy to make a model of any particular machine that one wishes to work on within such a UPCM instead of having to work with a paper machine as at present. If one also decided on quite definite ‘teaching policies’ these could also be programmed into the machine. One would then allow the whole system to run for an appreciable period, and then break in as a kind of ‘inspector of schools’ and see what progress had been made. One might also be able to make some progress with unorganised machines...

The upshot of this line of thought is that all mental operations are *computable* and hence realisable on a universal machine: the computer. Turing advanced this view with increasing confidence in the late 1940s, perfectly aware that it represented what he enjoyed calling ‘heresy’ to the believers in minds or souls beyond material description.

Turing was not a mechanical thinker, or a stickler for convention; far from it. Of all people, he knew the nature of originality and individual independence. Even in tackling the U-boat Enigma problem, for instance, he declared that he did so because no-one else was looking at it and he could have it to himself. Far from being trained or organised into this problem, he took it on despite the prevailing wisdom in 1939 that it was too difficult to attempt. His arrival at a thesis of ‘machine intelligence’ was not the outcome of some dull or restricted mentality, or a lack of appreciation of individual human creativity.

7. Machine Intelligence

Turing relished the paradox of ‘Machine Intelligence’: an apparent contradiction in terms. It is likely that he was already savouring this theme in 1941, when he read a theological book by the author Dorothy Sayers (Sayers 1941). In (Turing 1948) he quoted from this work to illustrate his full awareness that in common parlance ‘mechanical’ was used to to mean ‘devoid of intelligence.’ Giving a date which no doubt had his highly sophisticated Enigma-breaking machines secretly in mind, he wrote that ‘up to 1940’ only very limited machinery had been used, and this ‘encouraged the belief that machinery was necessarily limited to extremely straightforward, possibly even to repetitious, jobs.’ His object was to dispel these connotations.

In 1950, Turing wrote on the first page of his Manual for users of the Manchester University computer (Turing 1950a):

> Electronic computers are intended to carry out any definite rule of thumb process which could have been done by a human operator working in a disciplined but unintelligent manner.

This is, of course, just the 1936 universal Turing machine, now in electronic form. On the other hand, he also wrote in the more famous paper of that year (Turing 1950b, p. 460)

> We may hope that machines will eventually compete with men in all purely intellectual fields.

How could the *intelligent* arise from operations which were themselves totally *routine and mindless* -- 'entirely without intelligence'? This is the core of the problem Turing faced, and the same problem faces Artificial Intelligence research today. Turing's underlying argument was that the human brain must somehow be organised for intelligence, and that the organisation of the brain must be realisable as a finite discrete-state machine. The implications of this view were exposed to a wider circle in his famous paper, "Computing Machinery and Intelligence," which appeared in *Mind* in October 1950.

The appearance of this paper, Turing's first foray into a journal of philosophy, was stimulated by his discussions at Manchester University with Michael Polanyi. It also reflects the general sympathy of Gilbert Ryle, editor of *Mind*, with Turing's point of view.

Turing's 1950 paper was meant for a wide readership and should be read in its original; it has often been reprinted. Not surprisingly, the paper has attracted many critiques. Not all commentators note the careful explication of computability which opens the paper, with an emphasis on the concept of the universal machine. This explains why if mental function can be achieved by any finite discrete state machine, then the same effect can be achieved by programming a computer (Turing 1950b, p. 442). (Note, however, that Turing makes no claim that the nervous system should resemble a digital computer in its structure.) Turing's treatment has a severely finitistic flavour: his argument is that the relevant action of the brain is not only computable, but realisable as a totally finite machine, i.e. as a Turing machine that does not use any 'tape' at all. In his account, the full range of computable functions, defined in terms of Turing machines that use an infinite tape, only appears as being of 'special theoretical interest.' (Of uncomputable functions there is, *a fortiori,* no mention.) Turing uses the finiteness of the nervous system to give an estimate of about 10^9 bits of storage required for a limited simulation of intelligence (Turing 1950b, p. 455).

The wit and drama of Turing's 'imitation game' has attracted more fame than his careful groundwork. Turing's argument was designed to bypass discussions of the nature of thought, mind, and consciousness, and to give a criterion in terms of external observation alone. His justification for this was that one only judges that other human beings are thinking by external observation, and he applied a principle of 'fair play for machines' to argue that the same should hold for machine intelligence. He dramatised this viewpoint by a thought-experiment (which nowadays can readily be tried out). A human being and a programmed computer compete to convince an impartial judge, using textual messsages alone, as to which is the human being. If the computer wins, it must be credited with intelligence.

Turing introduced his 'game' confusingly with a poor analogy: a party game in which a man pretends to

be a woman. His loose wording (Turing 1950b, p. 434) has led some writers wrongly to suppose that Turing proposed an ‘imitation game’ in which a machine has to imitate a man imitating a woman. Others, like Lassègue (1998), place much weight on this game of gender pretence and its real or imaginary connotations. In fact, the whole point of the ‘test’ setting, with its remote text-message link, was to *separate* intelligence from other human faculties and properties. But it may fairly be said that this confusion reflects Turing’s richly ambitious concept of what is involved in human ‘intelligence’. It might also be said to illustrate his own human intelligence, in particular a delight in the Wildean reversal of roles, perhaps reflecting, as in Wilde, his homosexual identity. His friends knew an Alan Turing in whom intelligence, humour and sex were often intermingled.

Turing was in fact sensitive to the difficulty of separating ‘intelligence’ from other aspects of human senses and actions; he described ideas for robots with sensory attachments and raised questions as to whether they might enjoy strawberries and cream or feel racial kinship. In contrast, he paid scant attention to the questions of authenticity and deception implicit in his test, essentially because he wished to by-pass questions about the reality of consciousness. A subtle aspect of one of his imagined ‘intelligent’ conversations (Turing 1950b, p. 434) is where the computer imitates human intelligence by giving the *wrong answer* to a simple arithmetic problem. But in Turing’s setting we are not supposed to ask whether the computer ‘consciously’ deceives by giving the impression of innumerate humanity, nor why it should wish to do so. There is a certain lack of seriousness in this approach. Turing took on a second-rank target in countering the published views of the brain surgeon G. Jefferson, as regards the objectivity of consciousness. Wittgenstein’s views on Mind would have made a more serious point of departure.

Turing’s imitation principle perhaps also assumes (like ‘intelligence tests’ of that epoch) too much of a shared language and culture for his imagined interrogations. Neither does it address the possibility that there may be kinds of thought, by animals or extra-terrestrial intelligences, which are not amenable to communication.

A more positive feature of the paper lies in its constructive program for research, culminating in Turing’s ideas for ‘learning machines’ and educating ‘child’ machines (Turing 1950b, p. 454). It is generally thought (e.g. in Dreyfus and Dreyfus 1990) that there was always an antagonism between programming and the ‘connectionist’ approach of neural networks. But Turing never expressed such a dichotomy, writing that both approaches should be tried. Donald Michie, the British AI research pioneer profoundly influenced by early discussions with Turing, has called this suggestion ‘Alan Turing’s Buried Treasure’, in an allusion to a bizarre wartime episode in which Michie was himself involved (Hodges 1983, p. 345). The question is still highly pertinent.

It is also a commonly expressed view that Artifical Intelligence ideas only occurred to pioneers in the 1950s *after* the success of computers in large arithmetical calculations. It is hard to see why Turing’s work, which was rooted from the outset in the question of mechanising Mind, has been so much overlooked. But through his failure to publish and promote work such as that in (Turing 1948) he largely lost recognition and influence.

It is also curious that Turing’s best-known paper should appear in a journal of philosophy, for it may well

be said that Turing, always committed to materialist explanation, was not really a philosopher at all. Turing was a mathematician, and what he had to offer philosophy lay in illuminating its field with what had been discovered in mathematics and physics. In the 1950 paper this was surprisingly cursory, apart from his groundwork on the concept of computability. His emphasis on the sufficiency of the computable to explain the action of the mind was stated more as a hypothesis, even a manifesto, than argued in detail. Of his hypothesis he wrote (Turing 1950b, p. 442):

> ...I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. I believe further that no useful purpose is served by concealing these beliefs. The popular view that scientists proceed inexorably from established fact to established fact, never being influenced by any unproved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjecture, no harm can result. Conjectures are of great importance since they suggest useful lines of research.

Penrose (1994, p.21), probing into Turing's conjecture, has presented it as 'Turing's thesis' thus:

> It seems likely that he viewed physical action in general -- which would include the action of a human brain -- to be always reducible to some kind of Turing-machine action.

The statement that all physical action is in effect computable goes beyond Turing's explicit words, but is a fair characterisation of the implicit assumptions behind the 1950 paper. Turing's consideration of 'The Argument from Continuity in the Nervous System,' in particular, simply asserts that the physical system of the brain can be approximated as closely as is desired by a computer program (Turing 1950b, p. 451). Certainly there is nothing in Turing's work in the 1945-50 period to contradict Penrose's interpretation. The more technical precursor papers (Turing 1947, 1948) include wide-ranging comments on physical processes, but make no reference to the possibility of physical effects being uncomputable.

In particular, a section of (Turing 1948) is devoted to a general classification of 'machines.' The period between 1937 and 1948 had given Turing much more experience of actual machinery than he had in 1936, and his post-war remarks reflected this in a down-to-earth manner. Turing distinguished 'controlling' from 'active' machinery, the latter being illustrated by 'a bulldozer'. Naturally it is the former -- in modern terms 'information-based machinery' -- with which Turing's analysis is concerned. It is noteworthy that in 1948 as in 1936, despite his knowledge of physics, Turing made no mention of how quantum mechanics might affect the concept of 'controlling'. His concept of 'controlling' remained entirely within the classical framework of the Turing machine (which he called a Logical Computing Machine in this paper.)

The same section of (Turing 1948) also drew the distinction between *discrete* and *continuous* machinery, illustrating the latter with 'the telephone' as a continuous, controlling machine. He made light of the difficulty of reducing continuous physics to the discrete model of the Turing machine, and though citing 'the brain' as a continuous machine, stated that it could probably be treated as if discrete. He gave no indication that physical continuity threatened the paramount role of computability. In fact, his thrust in

(Turing 1947) was to promote the digital computer as *more powerful* than analog machines such as the differential analyser. When he discussed this comparison, he gave the following informal version of the Church-Turing thesis:

> One of my conclusions was that the idea of a ‘rule of thumb’ process and a ‘machine process’ were synonymous. The expression ‘machine process’ of course means one which could be carried out by the type of machine I was considering [i.e. Turing machines]

Turing gave no hint that the discreteness of the Turing machine consituted a real limitation, or that the non-discrete processes of analog machines might be of any deep significance.

Turing also introduced the idea of ‘random elements’ but his examples (using the digits of π) showed that he considered *pseudo-random* sequences (i.e. computable sequences with suitable ‘random’ properties) quite adequate for his discussion. He made no suggestion that randomness implied something uncomputable, and indeed gave no definition of the term ‘random’. This is perhaps surprising in view of the fact that his work in pure mathematics, logic and cryptography all gave him considerable motivation to approach this question at a serious level.

8. Unfinished Work

From 1950 Turing worked on a new mathematical theory of morphogenesis, based on showing the consequences of non-linear equations for chemical reaction and diffusion (Turing 1952). He was a pioneer in using a computer for such work. Some writers have referred to this theory as founding Artificial Life (A-life), but this is a misleading description, apt only to the extent that the theory was intended, as Turing saw it, to counter the Argument from Design. A-life since the 1980s has concerned itself with using computers to explore the logical consequences of evolutionary theory without worrying about specific physiological forms. Morphogenesis is complementary, being concerned to show which physiological pathways are feasible for evolution to exploit. Turing’s work was developed by others in the 1970s and is now regarded as central to this field.

It may well be that Turing’s interest in morphogenesis went back to a primordial childhood wonder at the appearance of plants and flowers. But in another late development, Turing went back to other stimuli of his youth. For in 1951 Turing did consider the problem, hitherto avoided, of setting computability in the context of quantum-mechanical physics. In a BBC radio talk of that year (Turing 1951) he discussed the basic groundwork of his 1950 paper, but this time dealing rather less certainly with the argument from Gödel’s theorem, and this time also referring to the quantum-mechanical physics underlying the brain. Turing described the universal machine property, applying it to the brain, but said that its applicability required that the machine whose behaviour is to be imitated

> ...should be of the sort whose behaviour is in principle predictable by calculation. We certainly do not know how any such calculation should be done, and it was even argued by Sir Arthur Eddington that on account of the indeterminacy principle in quantum mechanics

> no such prediction is even theoretically possible.

Copeland (1999) has rightly drawn attention to this sentence in his preface to his edition of the 1951 talk. However, Copeland's critical context suggests some connection with Turing's 'oracle.' There is is in fact no mention of oracles here (nor anywhere in Turing's post-war discussion of mind and machine.) Turing here is discussing the possiblility that, when seen as as a *quantum-mechanical machine* rather than a classical machine, the Turing machine model is inadequate. The correct connection to draw is not with Turing's 1938 work on ordinal logics, but with his knowledge of quantum mechanics from Eddington and von Neumann in his youth. Indeed, in an early speculation, influenced by Eddington, Turing had suggested that quantum mechanical physics could yield the basis of free-will (Hodges 1983, p. 63). Von Neumann's axioms of quantum mechanics involve two processes: unitary evolution of the wave function, which is predictable, and the measurement or reduction operation, which introduces unpredictability. Turing's reference to unpredictability must therefore refer to the reduction process. The essential difficulty is that still to this day there is no agreed or compelling theory of when or how reduction actually occurs. (It should be noted that 'quantum computing,' in the standard modern sense, is based on the predictability of the unitary evolution, and does not, as yet, go into the question of how reduction occurs.) It seems that this single sentence indicates the beginning of a new field of investigation for Turing, this time into the foundations of quantum mechanics. In 1953 Turing wrote to his friend and student Robin Gandy that he was 'trying to invent a new Quantum Mechanics but it won't really work.'

At Turing's death in June 1954, Gandy reported in a letter to Newman on what he knew of Turing's current work (Gandy 1954). He wrote of Turing having discussed a problem in understanding the reduction process, in the form of

> ...'the Turing Paradox'; it is easy to show using standard theory that if a system start in an eigenstate of some observable, and measurements are made of that observable N times a second, then, even if the state is not a stationary one, the probability that the system will be in the same state after, say, 1 second, tends to one as N tends to infinity; i.e. that continual observation will prevent motion. Alan and I tackled one or two theoretical physicists with this, and they rather pooh-poohed it by saying that continual observation is not possible. But there is nothing in the standard books (eg Dirac's) to this effect, so that at least the paradox shows up an inadequacy of Quantum Theory as usually presented.

Turing's investigations take on added significance in view of the assertion of Penrose (1989, 1990, 1994, 1996) that the reduction process must involve something uncomputable. Probably Turing was aiming at the opposite idea, of finding a theory of the reduction process that would be predictive and computable, and so plug the gap in his hypothesis that the action of the brain is computable. However Turing and Penrose are alike in seeing this as an important question affecting the assumption that all mental action is computable; in this they both differ from the mainstream view in which the question is accorded little significance.

Alan Turing's last postcards to Robin Gandy, in March 1954, headed 'Messages from the Unseen World' in allusion to Eddington, hinted at new ideas in the fundamental physics of relativity and particle physics

(Hodges 1983, p. 512). They illustrate the wealth of ideas with which he was concerned at that last point in his life, but which apart from these hints are entirely lost.

9. Alan Turing: the Unknown Mind

It is a pity that Turing did not write more about his ethical philosophy and world outlook. As a student he was an admirer of Bernard Shaw's plays of ideas, and to friends would openly voice both the hilarities and frustrations of his many difficult situations. Yet the nearest he came to serious personal writing, apart from occasional comments in private letters, was in penning a short story about his 1952 crisis (Hodges 1983, p. 448). His last two years were particularly full of Shavian drama and Wildean irony. In one letter (to his friend Norman Routledge; the letter is now in the Turing Archive at King's College, Cambridge) he wrote:

> Turing believes machines think
> Turing lies with men
> Therefore machines do not think

The syllogistic allusion to Socrates is unmistakeable, and his demise, with cyanide rather than hemlock, may have signalled something similar. A parallel figure in World War II, Robert Oppenheimer, suffered the loss of his reputation during the same week that Turing died. Both combined the purest scientific work and the most effective application of science in war. Alan Turing was even more directly on the receiving end of science, when his sexual mind was treated as a machine, against his protesting consciousness and will. But amidst all this human drama, he left little to say about what he really thought of himself and his relationship to the world of human events.

Alan Turing did not fit easily with any of the intellectual movements of his time, aesthetic, technocratic or marxist. In the 1950s, commentators struggled to find discreet words to categorise him: as 'a scientific Shelley,' as possessing great 'moral integrity'. But until the 1970s the reality of his life was unmentionable. He is still hard to place within twentieth-century thought. He exalted the science that existentialists held to have had robbed life of meaning. The most original figure, the most insistent on personal freedom, he held originality and will to be susceptible to mechanisation. The mind of Alan Turing remains an enigma.

Bibliography

- Boden. M. (ed.), 1990, *The philosophy of artificial intelligence*, Oxford: Oxford University Press
- Church, A., 1937, Review of Turing 1936-7, *Journal of Symbolic Logic*, 2: 42
- Church, A., 1940, 'On the concept of a random sequence', *Bull. Amer. Math. Soc.*, 46: 130-135
- Copeland B. J. and Proudfoot D., 1996, 'On Alan Turing's anticipation of connectionism', *Synthese*, 108: 361-377
- Copeland B. J., 1998, 'Turing's o-machines, Searle, Penrose and the brain', *Analysis* 58.2: 128-

138

- Copeland B. J., 1999, 'A lecture and two radio broadcasts on machine intelligence by Alan Turing', in *Machine Intelligence 15*, K. Furukawa, D. Michie, and S. Muggleton (eds.), Oxford: Oxford University Press
- Davies, E. B., 2001, 'Building infinite machines', *British Journal for the Philosophy of Science*, forthcoming. [Preprint available online]
- Davis, M. (ed.), 1958, *Computability and unsolvability*, New York: McGraw-Hill; new edition New York: Dover (1982)
- Davis, M. (ed.), 1965, *The undecidable*, New York: Raven
- Davis, M., 2000, *The universal computer*, New York: Norton
- Dawson, J. W., 1985, Review of Hodges (1983), *Journal of Symbolic Logic*, 50: 1065-1067
- Deutsch, D., 1985, 'Quantum theory, the Church-Turing principle and the universal quantum computer', *Proc. Roy. Soc.* A 400: 97-115
- Diamond, C. (ed.), 1976, *Wittgenstein's Lectures on the Foundations of Mathematics, Cambridge, 1939*, Hassocks: Harvester Press
- Dreyfus H. L. and Dreyfus S. E., 1990, 'Making a mind versus modelling the brain: artificial intelligence back at a branch-point', in (Boden 1990)
- Feferman, S., 1988, 'Turing in the Land of O(Z)', in (Herken 1988); also in (Gandy and Yates 2001)
- Feynman, R. P., 1982, 'Simulating physics with computers', *Int. Journal of Theoretical Physics* 21: 467-488
- Gandy, R. O., 1954, Letter to M. H. A. Newman, in (Gandy and Yates, 2001)
- Gandy, R. O., 1980, 'Principles of Mechanisms', in *The Kleene Symposium*, eds. J. Barwise, H. J. Keisler and K. Kunen, Amsterdam: North-Holland
- Gandy, R. O. 1988, 'The confluence of ideas in 1936', in (Herken 1988)
- Gandy, R. O. and Yates, C. E. M. (eds.), 2001, *The Collected Works of A M. Turing: Mathematical Logic*, Amsterdam: North-Holland
- Gödel, K., 1946, 'Remarks before the Princeton Bicentennial Conference on problems in mathematics', in (Davis 1965)
- Herken R. (ed.), 1988, *The universal Turing machine: a half-century survey*, Berlin: Kammerer und Unverzagt; Oxford: Oxford University Press
- Hodges, A., 1983, *Alan Turing: the Enigma*, London: Burnett; New York: Simon & Schuster; new editions London: Vintage (1992); New York: Walker (2000).
- Hodges, A., 1997, *Turing, a natural philosopher*, London: Phoenix; also New York: Routledge (1999). Included in *The great philosophers*, eds. R. Monk and F. Raphael, London: Weidenfeld and Nicolson (2000).
- Ince, D. C., 1989, Preface to (Turing 1948), in (Ince 1992)
- Ince, D. C. (ed.), 1992, *The Collected Works of A. M. Turing: Mechanical Intelligence*, Amsterdam: North-Holland
- Lassègue, J., 1998, *Turing*, Paris: les Belles Lettres
- Minsky, M. L., 1967, *Computation: finite and infinite machines*, Englewood Cliffs, N.J.: Prentice-Hall
- Newman, M. H. A., 1955, 'Alan Mathison Turing', *Biographical memoirs of the Royal Society*

(1955): 253-263
- Penrose, R., 1989, *The emperor's new mind*, Oxford: Oxford University Press
- Penrose, R., 1990, 'Précis of *The emperor's new mind...*,' *Behavioral and Brain Sciences* 13: 643-655
- Penrose, R., 1994, *Shadows of the mind*, Oxford: Oxford University Press
- Penrose, R., 1996, 'Beyond the doubting of a shadow: A Reply to Commentaries on *Shadows of the Mind*', in *Psyche: An Interdisciplinary Journal of Research on Consciousness*, Volume 2. [Available online]
- Sayers, D., 1941, *The mind of the maker*, London: Methuen
- Teuscher, C., 2001, *Turing's connectionism*, London: Springer-Verlag UK.
- Turing, A. M., 1936-7, 'On computable numbers, with an application to the Entscheidungsproblem', *Proc. London Maths. Soc., ser. 2*, 42: 230-265; also in (Davis 1965) and (Gandy and Yates 2001); [Available online]
- Turing A. M., 1939, 'Systems of logic defined by ordinals', *Proc. Lond. Math. Soc., ser. 2*, 45: 161-228; also in (Davis 1965) and in (Gandy and Yates 2001). This was Turing's Ph.D. thesis, Princeton University (1938).
- Turing, A. M., 1946, *Proposed electronic calculator*, report for National Physical Laboratory, Teddington; published in *A. M. Turing's ACE report of 1946 and other papers*, eds. B. E. Carpenter and R. W. Doran, Cambridge, Mass.: MIT Press (1986); also in (Ince 1992)
- Turing, A. M., 1947, 'Lecture to the London Mathematical Society on 20 February 1947', in *A. M. Turing's ACE report of 1946 and other papers*, eds. B. E. Carpenter and R. W. Doran, Cambridge, Mass.: MIT Press (1986); also in (Ince 1992).
- Turing, A. M., 1948, 'Intelligent machinery', report for National Physical Laboratory, in *Machine Intelligence 7*, eds. B. Meltzer and D. Michie (1969); also in (Ince 1992)
- Turing, A. M., 1950a, *Programmers' Handbook for the Manchester electronic computer*, Manchester University Computing Laboratory. [Available online]
- Turing, A. M., 1950b, 'Computing machinery and intelligence', *Mind* 50: 433-460; also in (Boden 1990), (Ince 1992), and [Available online]
- Turing, A. M., 1951, BBC radio talk, in *Machine Intelligence 15*, eds. K. Furukawa, D. Michie. and S. Muggleton, Oxford: Oxford University Press (1999)
- Turing, A. M., 1952, 'The chemical basis of morphogenesis', *Phil. Trans. R. Soc. London B* 237: 37-72; also in *The Collected Works of A. M. Turing: Morphogenesis*, ed. P. T. Saunders, Amsterdam: North-Holland (1992)
- Turing, E. S., 1959, *Alan M. Turing*, Cambridge: Heffers
- von Neumann, J., 1945, 'First draft of a report on the EDVAC', University of Pennsylvania; first printed in Stern, N., *From Eniac to Univac: an appraisal of the Eckert-Mauchly machines*, Bedford MA: Digital Press (1981)
- Whitemore, H., 1986, *Breaking the code*, London: S. French

Other Internet Resources

- Turing Digital Archive

- Alan Turing Home Page

Related Entries

Church-Turing Thesis | computing, modern history of | connectionism | quantum theory: measurement in | Turing machine

Andrew Hodges
andrew.hodges@wadh.ox.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 3, 2002
Content last modified: June 3, 2002

Turing Machine

A Turing machine is an abstract representation of a computing device. It consists of a read/write head that scans a (possibly infinite) one-dimensional (bi-directional) tape divided into squares, each of which is inscribed with a 0 or 1. Computation begins with the machine, in a given "state", scanning a square. It erases what it finds there, prints a 0 or 1, moves to an adjacent square, and goes into a new state. This behavior is completely determined by three parameters: (1) the state the machine is in, (2) the number on the square it is scanning, and (3) a table of instructions. The table of instructions specifies, for each state and binary input, what the machine should write, which direction it should move in, and which state it should go into. (E.g., "If in State 1 scanning a 0: print 1, move left, and go into State 3".) The table can list only finitely many states, each of which becomes implicitly defined by the role it plays in the table of instructions. These states are often referred to as the "functional states" of the machine.

A Turing machine, therefore, is more like a computer program (software) than a computer (hardware). Any given Turing machine can be realized or implemented on an infinite number of different physical computing devices. Computer scientists and logicians have shown that Turing machines -- given enough time and tape -- can compute any function that any conventional digital computers can compute. Also, a ‘probabilistic automaton’ can be defined as a Turing machine in which the transition from input and state to output and state takes place with a certain probability (E.g. "If in State 1 scanning a 0: (a) there is a 60% probability that the machine will print 1, move left, and go into State 3, and (b) there is a 40% probability that the machine will print 0, move left, and go into State 2".)

History

Turing machines were first proposed by Alan Turing, in an attempt to give a mathematically precise definition of "algorithm" or "mechanical procedure". Early work by Turing and Alonzo Church spawned the branch of mathematical logic now known as recursive function theory.

Later Developments

The concept of a Turing machine has played an important role in the recent philosophy of mind. The suggestion has been made that mental states just are functional states of a probabilistic automaton, in which binary inputs and outputs have been replaced by sensory inputs and motor outputs. This idea underlies the theory of mind known as "machine functionalism".

Bibliography

- Turing, A., "On Computable Numbers, With an Application to the Entscheidungsproblem", *Proceedings of the London Mathematical Soceity*, Series 2, Volume 42, 1936; reprinted in M. David (ed.), *The Undecidable*, Hewlett, NY: Raven Press, 1965
- Boolos, G. and Jeffrey, R., *Computability and Logic*, 2nd ed., Cambridge: Cambridge University Press, 1980.
- Putnam, H., "The Nature of Mental States", in *Mind, Language and Reality: Philosophical Papers II*, Cambridge: Cambridge University Press, 1975

Other Internet Resources

- Turing's World (computer software from CSLI)
- Andrew Hodge's Web Pages on Alan Turing
- Cristian Cheran's Visual Turing program (Downloadable on Windows systems)
- Paul Ming's Virtual Turing Machine website

Related Entries

artificial intelligence | Church-Turing Thesis | functionalism | Turing, Alan

Acknowledgement

The Editors would like to thank Stuart Shieber for pointing out that Turing machines need not have infinite, two-dimensional tapes, but that infinite, one-dimensional and bidirectional tapes suffice.

First published: September 14, 1995
Content last modified: March 9, 2002

The Modern History of Computing

Historically, computers were human clerks who calculated in accordance with effective methods. These human computers did the sorts of calculation nowadays carried out by electronic computers, and many thousands of them were employed in commerce, government, and research establishments. The term *computing machine*, used increasingly from the 1920s, refers to any machine that does the work of a human computer, i.e. any machine that calculates in accordance with effective methods. During the late 1940s and early 1950s, with the advent of electronic computing machines, the phrase 'computing machine' gradually gave way simply to 'computer', initially usually with the prefix 'electronic' or 'digital'. This entry surveys the history of these machines.

Babbage

Charles Babbage was Lucasian Professor of Mathematics at Cambridge University from 1828 to 1839 (a post formerly held by Isaac Newton). Babbage's proposed Difference Engine was a special-purpose digital computing machine for the automatic production of mathematical tables (such as logarithm tables, tide tables, and astronomical tables). The Difference Engine consisted entirely of mechanical components -

brass gear wheels, rods, ratchets, pinions, etc. Numbers were represented in the decimal system by the positions of 10-toothed metal wheels mounted in columns. Babbage exhibited a small working model in 1822. He never completed the full-scale machine that he had designed but did complete several fragments. The largest - one ninth of the complete calculator - is on display in the London Science Museum. Baabage used it to perform serious computational work, calculating various mathematical tables. In 1990, Babbage's Difference Engine No. 2 was finally built from Babbage's designs and is also on display at the London Science Museum.

The Swedes Georg and Edvard Scheutz (father and son) constructed a modified version of Babbage's Difference Engine. Three were made, a prototype and two commercial models, one of these being sold to an observatory in Albany, New York, and the other to the Registrar-General's office in London, where it calculated and printed actuarial tables.

Babbage's proposed Analytical Engine, considerably more ambitious than the Difference Engine, was to have been a general-purpose mechanical digital computer. The Analytical Engine was to have had a memory store and a central processing unit (or ‘mill’) and would have been able to select from among alternative actions consequent upon the outcome of its previous actions (a facility nowadays known as conditional branching). The behaviour of the Analytical Engine would have been controlled by a program of instructions contained on punched cards connected together with ribbons (an idea that Babbage had adopted from the Jacquard weaving loom). Babbage emphasised the generality of the Analytical Engine, saying ‘the conditions which enable a finite machine to make calculations of unlimited extent are fulfilled in the Analytical Engine’ (Babbage [1994], p. 97).

Babbage worked closely with Ada Lovelace, daughter of the poet Byron, after whom the modern programming language ADA is named. Lovelace foresaw the possibility of using the Analytical Engine for non-numeric computation, suggesting that the Engine might even be capable of composing elaborate pieces of music.

A large model of the Analytical Engine was under construction at the time of Babbage's death in 1871 but a full-scale version was never built. Babbage's idea of a general-purpose calculating engine was never forgotten, especially at Cambridge, and was on occasion a lively topic of mealtime discussion at the war-time headquarters of the Government Code and Cypher School, Bletchley Park, Buckinghamshire, birthplace of the electronic digital computer.

Analog computers

The earliest computing machines in wide use were not digital but analog. In analog representation, properties of the representational medium ape (or reflect or model) properties of the represented state-of-affairs. (In obvious contrast, the strings of binary digits employed in digital representation do *not* represent by means of possessing some physical property - such as length - whose magnitude varies in proportion to the magnitude of the property that is being represented.) Analog representations form a diverse class. Some examples: the longer a line on a motorway map, the longer the road that the line

represents; the greater the number of clear plastic squares in an architect's model, the greater the number of windows in the building represented; the higher the pitch of an acoustic depth meter, the shallower the water. In analog computers, numerical quantities are represented by, for example, the angle of rotation of a shaft or a difference in electrical potential. Thus the output voltage of the machine at a time might represent the momentary speed of the object being modelled.

As the case of the architect's model makes plain, analog representation may be *discrete* in nature (there is no such thing as a fractional number of windows). Among computer scientists, the term ‘analog’ is sometimes used narrowly, to indicate representation of one *continuously-valued* quantity by another (e.g. speed by voltage). As Brian Cantwell Smith has remarked:

> "Analog" should ... be a predicate on a representation whose structure corresponds to that of which it represents ... That continuous representations should historically have come to be called analog presumably betrays the recognition that, at the levels at which it matters to us, the world is more foundationally continuous than it is discrete. (Smith [1991], p. 271)

James Thomson, brother of Lord Kelvin, invented the mechanical wheel-and-disc integrator that became the foundation of analog computation (Thomson [1876]). The two brothers constructed a device for computing the integral of the product of two given functions, and Kelvin described (although did not construct) general-purpose analog machines for integrating linear differential equations of any order and for solving simultaneous linear equations. Kelvin's most successful analog computer was his tide predicting machine, which remained in use at the port of Liverpool until the 1960s. Mechanical analog devices based on the wheel-and-disc integrator were in use during World War I for gunnery calculations. Following the war, the design of the integrator was considerably improved by Hannibal Ford (Ford [1919]).

Stanley Fifer reports that the first semi-automatic mechanical analog computer was built in England by the Manchester firm of Metropolitan Vickers prior to 1930 (Fifer [1961], p. 29); however, I have so far been unable to verify this claim. In 1931, Vannevar Bush, working at MIT, built the differential analyser, the first large-scale automatic general-purpose mechanical analog computer. Bush's design was based on the wheel and disc integrator. Soon copies of his machine were in use around the world (including, at Cambridge and Manchester Universities in England, differential analysers built out of kit-set Meccano, the once popular engineering toy).

It required a skilled mechanic equipped with a lead hammer to set up Bush's mechanical differential analyser for each new job. Subsequently, Bush and his colleagues replaced the wheel-and-disc integrators and other mechanical components by electromechanical, and finally by electronic, devices.

A differential analyser may be conceptualised as a collection of ‘black boxes’ connected together in such a way as to allow considerable feedback. Each box performs a fundamental process, for example addition, multiplication of a variable by a constant, and integration. In setting up the machine for a given task, boxes are connected together so that the desired set of fundamental processes is executed. In the case of electrical machines, this was done typically by plugging wires into sockets on a patch panel (computing

machines whose function is determined in this way are referred to as ‘program-controlled’).

Since all the boxes work in parallel, an electronic differential analyser solves sets of equations very quickly. Against this has to be set the cost of massaging the problem to be solved into the form demanded by the analog machine, and of setting up the hardware to perform the desired computation. A major drawback of analog computation is the higher cost, relative to digital machines, of an increase in precision. During the 1960s and 1970s, there was considerable interest in ‘hybrid’ machines, where an analog section is controlled by and programmed via a digital section. However, such machines are now a rarity.

The Universal Turing Machine

In 1935, at Cambridge University, Turing invented the principle of the modern computer. He described an abstract digital computing machine consisting of a limitless memory and a scanner that moves back and forth through the memory, symbol by symbol, reading what it finds and writing further symbols (Turing [1936]). The actions of the scanner are dictated by a program of instructions that is stored in the memory in the form of symbols. This is Turing's stored-program concept, and implicit in it is the possibility of the machine operating on and modifying its own program. (In London in 1947, in the course of what was, so far as is known, the earliest public lecture to mention computer intelligence, Turing said, "What we want is a machine that can learn from experience", adding that the "possibility of letting the machine alter its own instructions provides the mechanism for this" (Turing [1947] p. 123). Turing's computing machine of 1935 is now known simply as the universal Turing machine. Cambridge mathematician Max Newman has remarked that right from the start Turing was interested in the possibility of actually building a computing machine of the sort that he had described (Newman in interview with Christopher Evans in Evans [1976].

During the Second World War, Turing was a leading cryptanalyst at the Government Code and Cypher School, Bletchley Park. Here he became familiar with Thomas Flowers' work involving large-scale high-speed electronic switching (described below). However, Turing could not turn to the project of building an electronic stored-program computing machine until the cessation of hostilities in Europe in 1945.

Turing did give considerable thought to the question of machine intelligence during the wartime years. Colleagues at Bletchley Park recall numerous off-duty discussions with him on the topic, and at one point Turing circulated a typewritten report (now lost) setting out some of his ideas. One of these colleagues, Donald Michie (who later founded the Department of Machine Intelligence and Perception at the University of Edinburgh), remembers Turing talking often about the possibility of computing machines (1) learning from experience and (2) solving problems by means of searching through the space of possible solutions, guided by rule-of-thumb principles (Michie in interview with Copeland, 1995 and 1998). The modern term for the latter idea is ‘heuristic search’, a heuristic being any rule-of-thumb principle that cuts down the amount of searching required in order to find a solution to a problem. At Bletchley Park Turing illustrated his ideas on machine intelligence by reference to chess. Michie recalls Turing experimenting with heuristics that later became common in chess programming (in particular minimax and best-first).

Electromechanical versus Electronic Computation

With some exceptions - including Babbage's purely mechanical engines, and the finger-powered National Accounting Machine - early digital computing machines were electromechanical. That is to say, their basic components were small, electrically-driven, mechanical switches called ‘relays’. These operate relatively slowly, whereas the basic components of an electronic computer -- originally vacuum tubes (valves) -- have no moving parts save electrons and so operate extremely fast. Electromechanical digital computing machines were built before and during the second world war by (among others) Howard Aiken at Harvard University, George Stibitz at Bell Telephone Laboratories, Turing at Princeton University and Bletchley Park, and Konrad Zuse in Berlin. To Zuse belongs the honour of having built the first working general-purpose program-controlled digital computer. This machine, later called the Z3, was functioning in 1941. (A program-controlled computer, as opposed to a stored-program computer, is set up for a new task by re-routing wires, by means of plugs etc.)

Relays were too slow and unreliable a medium to make practicable the construction of a large-scale general-purpose digital computer (notwithstanding valiant efforts in this direction by Aiken). It was the development of high-speed digital techniques using vacuum tubes that made the modern computer possible.

The earliest extensive use of vacuum tubes for digital data-processing appears to have been by the engineer Thomas Flowers, working in London at the British Post Office Research Station at Dollis Hill. (Material in this article concerning Flowers derives from personal communications from Flowers to Copeland (1996-8) and a tape-recorded interview between Flowers and Evans (see Evans [1976]). Electronic digital equipment designed by Flowers in 1934, for controlling the connections between telephone exchanges, went into operation in 1939, and involved between three and four thousand vacuum tubes running continuously. In 1938-1939 Flowers worked on an experimental high-speed electronic digital data-processing system, involving a data store. Flowers' aim, achieved after the war, was that such equipment should replace existing, less reliable, systems built from relays and used in telephone exchanges. Flowers did not investigate the idea of using electronic equipment for numerical calculation, but has remarked that at the outbreak of war with Germany in 1939 he was possibly the only person in Britain who realized that vacuum tubes could be used on a large scale for high-speed digital computation.

Atanasoff

The earliest comparable use of vacuum tubes in the U.S. seems to have been by John Atanasoff at what was then Iowa State College (now University). During the period 1937-1942 Atanasoff developed techniques for using vacuum tubes to perform numerical calculations digitally. In 1939, with the assistance of his student Clifford Berry, Atanasoff began building what is sometimes called the Atanasoff-Berry Computer, or ABC, a small-scale special-purpose electronic digital machine for the solution of systems of linear algebraic equations. The machine contained approximately 300 vacuum tubes. Although the electronic part of the machine functioned successfully, the computer as a whole never worked reliably,

errors being introduced by the unsatisfactory binary card-reader. Work was discontinued in 1942 when Atanasoff left Iowa State.

Colossus

The first fully functioning electronic digital computer was Colossus (1943), used by the Bletchley Park cryptanalysts from 1944.

From very early in the war the Government Code and Cypher School (GC&CS) was successfully deciphering German radio communications encoded by means of the Enigma system, and by early 1942 about 39,000 intercepted messages were being decoded each month, thanks to electromechanical machines known as ‘bombes’. These were designed by Turing and Gordon Welchman (building on earlier work by Polish cryptanalysts).

During the second half of 1940, messages encoded by means of a totally different method began to be intercepted. This new method of encryption, named ‘Fish’ by GC&CS, remained intractable until 1941 (the first major break-in occurring at the end of August 1941); current traffic was read for the first time in July 1942. Based on binary teleprinter code, Fish was used in preference to Morse-based Enigma for the encryption of high-level signals, for example messages from Hitler and other members of the German High Command.

The need to decipher this vital intelligence as rapidly as possible led Max Newman to propose in November 1942 (shortly after his recruitment to GC&CS from Cambridge University) that key parts of the decryption process be automated, by means of high-speed electronic counting devices. The first machine designed and built to Newman's specification, known as the Heath Robinson, was relay-based with electronic circuits for counting. (The electronic counters were designed by C.E. Wynn-Williams, who had been using thyratron tubes in counting circuits at the Cavendish Laboratory, Cambridge, since 1932 [Wynn-Williams 1932].) Installed in June 1943, Heath Robinson was unreliable and slow, and its high-speed paper tapes were continually breaking, but it proved the worth of Newman's method. Turing recommended that Newman approach Flowers - who had previously assisted with the design of a machine for use against Enigma - to improve the reliability of the Robinson. Flowers offered instead to design and build a fully electronic machine with a similar function to Heath Robinson. Flowers received little official encouragement from GC&CS but proceeded nonetheless, working independently at the Post Office Research Station at Dollis Hill. Colossus I was installed at Bletchley Park on 8 December 1943.

In all, ten Colossi were built. From a cryptanalytic viewpoint, a major difference between the prototype Colossus I and the later machines was the addition of the so-called Special Attachment, consequent upon a key discovery by cryptanalysts Donald Michie and Jack Good. This broadened the function of Colossus from ‘wheel setting’ - i.e. determining the settings of the encoding wheels of the German Lorenz cipher machine for a particular message, given the ‘patterns’ of the wheels - to ‘wheel breaking’, i.e. determining the wheel patterns themselves. The wheel patterns were eventually changed daily by the Germans on each of the numerous links between Berlin and strategically critical remote stations, notably the various Army

Group commanders in the field. By 1945 there were as many 30 links in total. About ten of these were broken and read regularly.

Colossus I contained approximately 1600 vacuum tubes and each of the subsequent machines approximately 2400 vacuum tubes. Like the smaller ABC, Colossus lacked two important features of modern computers. First, it had no internally stored programs. To set it up for a new task, the operator had to alter the machine's physical wiring, using plugs and switches. Second, Colossus was not a general-purpose machine, being designed for a specific cryptanalytic task involving counting and Boolean operations.

The magnificent working model presently on display at Bletchley Park, now a museum, is a mock-up of Colossus I.

F.H. Hinsley, official historian of GC&CS, has estimated that the war in Europe was shortened by at least two years as a result of the signals intelligence operation carried out at Bletchley Park, in which Colossus played a major role. Most of the Colossi were destroyed once hostilities ceased. Some of the electronic panels ended up at Newman's Computing Machine Laboratory in Manchester (see below), all trace of their original use having been removed, At least one Colossus was retained by GCHQ, the successor organisation to GC&CS. The last Colossus stopped running in 1960 (during its later years, it was used extensively for training).

Those who knew of Colossus were prohibited by the Official Secrets Act from sharing their knowledge. Until the 1970s, few had any idea that electronic computation had been used successfully during the second world war. In 1970 and 1975, respectively, Good and Michie published notes giving the barest outlines of Colossus. By 1983, Flowers had received clearance from the British Government to publish a full account of the hardware of Colossus I. Details of the later machines and of the Special Attachment, the uses to which the Colossi were put, and the cryptanalytic algorithms that they ran, were not declassified until 1996. Even today some documents remain classified.

For those acquainted with the universal Turing machine of 1935, and the associated stored-program concept, Flowers' racks of digital electronic equipment indicated the feasibility of using large numbers of vacuum tubes to implement a high-speed general-purpose stored-program digital computing machine. The war over, Newman lost no time in establishing the Royal Society Computing Machine Laboratory at Manchester University for precisely that purpose. A few months after his arrival at Manchester, Newman wrote as follows to the Princeton mathematician John von Neumann (February 1946):

> I am ... hoping to embark on a computing machine section here, having got very interested in electronic devices of this kind during the last two or three years. By about eighteen months ago I had decided to try my hand at starting up a machine unit when I got out. ... I am of course in close touch with Turing.

Turing's Automatic Computing Engine

Turing and Newman were thinking along similar lines. In 1945 Turing joined the National Physical Laboratory (NPL) in London, his brief to design and develop an electronic stored-program digital computer for scientific work. (Artificial Intelligence was not far from Turing's thoughts: he described himself as ‘building a brain’ and remarked in a letter that he was ‘more interested in the possibility of producing models of the action of the brain than in the practical applications to computing’.) John Womersley, Turing's immediate superior at NPL, christened Turing's proposed machine the Automatic Computing Engine, or ACE, in homage to Babbage's Difference Engine and Analytical Engine.

Turing's ‘Proposal for Development in the Mathematics Division of an Automatic Computing Engine (ACE)’ was the first relatively complete specification of an electronic stored-program general-purpose digital computer. An NPL file (now unfortunately destroyed) gave the date of Turing's proposal as 1945; Michael Woodger, Turing's assistant at NPL from May 1946, believes that the proposal was probably written between October and December 1945. (The proposal is reprinted in full in the collection (Carpenter and Doran [1986]). See also Copeland [1998].)

The first electronic stored-program digital computer to be proposed in the U.S. was the EDVAC (see below). The ‘First Draft of a Report on the EDVAC’ (May 1945), composed by von Neumann, contained little engineering detail, in particular concerning electronic hardware (owing to restrictions in the U.S.). Turing's proposal, on the other hand, supplied detailed circuit designs and specifications of hardware units, specimen programs in machine code, and even an estimate of the cost of building the machine (£11,200). ACE and EDVAC differed fundamentally from one another; for example, ACE employed distributed processing, while EDVAC had a centralised structure.

Turing saw that speed and memory were the keys to computing. (Turing's colleague at NPL, Jim Wilkinson, has observed that Turing ‘was obsessed with the idea of speed on the machine’ (in interview with Evans [1976]).) Turing's design had much in common with today's RISC architectures and it called for a high-speed memory of roughly the same capacity as an early Macintosh computer (enormous by the standards of his day). Had Turing's ACE been built as planned it would have been in a different league from the other early computers. However, progress on Turing's Automatic Computing Engine ran slowly, due to organisational difficulties at NPL, and in 1948 a ‘very fed up’ Turing (Robin Gandy's description, in interview with Copeland, 1995) left NPL for Newman's Computing Machine Laboratory at Manchester University. It was not until May 1950 that a small pilot model of the Automatic Computing Engine, built by Wilkinson, Edward Newman, Woodger, and others, first executed a program. With an operating speed of 1 MHz, the Pilot Model ACE was for some time the fastest computer in the world (Woodger in interview with Copeland, 1998).

Sales of DEUCE, the production version of the Pilot Model ACE, exceeded 30 (confounding a prediction by a top adviser to NPL that Britain's computing needs would be satisfied by a total of three digital computers (NPL archives)). The fundamentals of Turing's ACE design were employed by Harry Huskey (at Wayne State University, Detroit) in the Bendix G15 computer (Huskey in interview with Copeland, 1998). The G15 was arguably the first personal computer; over 400 were sold worldwide. DEUCE and the G15 remained in use until about 1970. Another computer deriving from Turing's ACE design, the

MOSAIC, played a role in Britain's air defences during the Cold War period; other derivatives include the Packard-Bell PB250 (1961).

The Manchester Machine

The earliest general-purpose stored-program electronic digital computer to work was built in the Royal Society Computing Machine Laboratory at Manchester University. The Manchester ‘Baby’, as it became known, was constructed by the engineers F.C. Williams and Tom Kilburn, and performed its first calculation on 21 June 1948. The tiny program, stored on the face of a cathode ray tube, was just seventeen instructions long. A much enlarged version of the machine, with a programming system designed by Turing, became the world's first commercially available computer, the Ferranti Mark I. The first to be completed was installed at Manchester University in February 1951; in all about ten were sold, in Britain, Canada, Holland and Italy.

The fundamental logico-mathematical contributions by Turing and Newman to the triumph at Manchester have been neglected, and the Manchester machine is nowadays remembered as the work of Williams and Kilburn. Indeed, Newman's role in the development of computers has never been sufficiently emphasised (due perhaps to his thoroughly self-effacing way of relating the relevant events).

It was Newman who, in a lecture in Cambridge in 1935, introduced Turing to the concept which led directly to the Turing machine: Newman defined a constructive process as one that a *machine* can carry out (Newman in interview with Evans, op. cit.). As a result of his acquaintance with Turing's work of 1935-36, Newman became interested in the possibilities of computing machinery in, as he put it, ‘a rather theoretical way’. It was not until Newman joined GC&CS in 1942 that his interest in computing machinery suddenly became practical, with his realisation that the attack on Fish could be mechanised. During the building of Colossus, Newman tried to interest Flowers in Turing's 1936 paper - birthplace of the stored-program concept - but Flowers (in his own words) ‘didn't really understand much of it’. There can be little doubt that by 1943, Newman had firmly in mind the idea of using electronic technology in order to construct a stored-program general-purpose digital computing machine.

In July of 1946 (the month in which the Royal Society approved Newman's application for funds to found the Computing Machine Laboratory), Freddie Williams, working at the Telecommunications Research Establishment, Malvern, began the series of experiments on cathode ray tube storage that was to lead to the Williams tube memory. Williams, until then a radar engineer, explains how it was that he came to be working on the problem of computer memory:

> [O]nce [the German Armies] collapsed ... nobody was going to care a toss about radar, and people like me ... were going to be in the soup unless we found something else to do. And computers were in the air. Knowing absolutely nothing about them I latched onto the problem of storage and tackled that. (Quoted in Bennett [unpublished].)

Newman learned of Williams' work, and there seems little doubt that Newman, with the able help of

Patrick Blackett, Langworthy Professor of Physics at Manchester and one of the most powerful figures in the University, was instrumental in the appointment of the 35 year old Williams to the recently vacated Chair of Electro-Technics at Manchester. (Newman himself was a member of the appointing committee (Tom Kilburn in interview with Copeland, 1997).) Williams immediately had Kilburn, his assistant at Malvern, seconded to Manchester. To take up the story in Williams' own words:

> [N]either Tom Kilburn nor I knew the first thing about computers when we arrived in Manchester University. We'd had enough explained to us to understand what the problem of storage was and what we wanted to store, and that we'd achieved, so the point now had been reached when we'd got to find out about computers ... Newman explained the whole business of how a computer works to us. (F.C. Williams in interview with Evans [1976])

Elsewhere Williams is explicit concerning Turing's role and gives something of the flavour of the explanation that he and Kilburn received:

> Tom Kilburn and I knew nothing about computers, but a lot about circuits. Professor Newman and Mr A.M. Turing ... knew a lot about computers and substantially nothing about electronics. [This is not entirely fair to Turing. BJC] They took us by the hand and explained how numbers could live in houses with addresses and how if they did they could be kept track of during a calculation. (Williams [1975], p. 328)

It seems that Newman must have used much the same words with Williams and Kilburn as he did in an address to the Royal Society on 4th March 1948:

> Professor Hartree ... has recalled that all the essential ideas of the general-purpose calculating machines now being made are to be found in Babbage's plans for his analytical engine. In modern times the idea of a universal calculating machine was independently introduced by Turing ... [T]he machines now being made in America and in this country ... [are] in certain general respects ... all similar. There is provision for storing numbers, say in the scale of 2, so that each number appears as a row of, say, forty 0's and 1's in certain places or "houses" in the machine. ... Certain of these numbers, or "words" are read, one after another, as orders. In one possible type of machine an order consists of four numbers, for example 11, 13, 27, 4. The number 4 signifies "add", and when control shifts to this word the "houses" H11 and H13 will be connected to the adder as inputs, and H27 as output. The numbers stored in H11 and H13 pass through the adder, are added, and the sum is passed on to H27. The control then shifts to the next order. In most real machines the process just described would be done by three separate orders, the first bringing [H11] (=content of H11) to a central accumulator, the second adding [H13] into the accumulator, and the third sending the result to H27; thus only one address would be required in each order. ... A machine with storage, with this automatic-telephone-exchange arrangement and with the necessary adders, subtractors and so on, is, in a sense, already a universal machine. (Newman [1948], pp. 271-272)

Newman goes on to explain program storage ('the orders shall be in a series of houses X1, X2, ...') and conditional branching. He then sums up:

> From this highly simplified account it emerges that the essential internal parts of the machine are, first, a storage for numbers (which may also be orders). ... Secondly, adders, multipliers, etc. Thirdly, an "automatic telephone exchange" for selecting "houses", connecting them to the arithmetic organ, and writing the answers in other prescribed houses. Finally, means of moving control at any stage to any chosen order, if a certain condition is satisfied, otherwise passing to the next order in the normal sequence. Besides these there must be ways of setting up the machine at the outset, and extracting the final answer in useable form. (Newman [1948], pp. 273-4)

There seems little doubt that the major credit for the Manchester machine belongs not only to Williams and Kilburn but also to Newman, and that the influence upon Newman of Turing's paper of 1936, which first set out the concept of the stored-program universal digital computer, was crucial.

The first working AI program, a draughts (checkers) player written by Christopher Strachey, ran on the Ferranti Mark I in the Manchester Computing Machine Laboratory. Strachey (at the time a teacher at Harrow School and an amateur programmer) wrote the program with Turing's encouragement and utilising the latter's recently completed Programmers' Handbook for the Ferranti. (Strachey later became Director of the Programming Research Group at Oxford University.) By the summer of 1952, the program could, Strachey reported, 'play a complete game of draughts at a reasonable speed'. (Strachey's program formed the basis for Arthur Samuel's well-known checkers program.) The first chess-playing program, also, was written for the Manchester Ferranti, by Dietrich Prinz; the program first ran in November 1951. Designed for solving simple problems of the mate-in-two variety, the program would examine every possible move until a solution was found. Turing started to program his 'Turochamp' chess-player on the Ferranti Mark I, but never completed the task. Unlike Prinz's program, the Turochamp could play a complete game (when hand-simulated) and operated not by exhaustive search but under the guidance of heuristics.

ENIAC and EDVAC

The first fully functioning electronic digital computer to be built in the U.S. was ENIAC, constructed at the Moore School of Electrical Engineering, University of Pennsylvania, for the Army Ordnance Department, by J. Presper Eckert and John Mauchly. Completed in 1945, ENIAC was somewhat similar to the earlier Colossus, but considerably larger and more flexible (although far from general-purpose). The primary function for which ENIAC was designed was the calculation of tables used in aiming artillery. ENIAC was not a stored-program computer, and setting it up for a new job involved reconfiguring the machine by means of plugs and switches. For many years, ENIAC was believed to have been the first functioning electronic digital computer, Colossus being unknown to all but a few.

In 1944, John von Neumann joined the ENIAC group. He had become 'intrigued' (Goldstine's word,

[1972], p. 275) with Turing's universal machine while Turing was at Princeton University during 1936-1938. At the Moore School, von Neumann emphasised the importance of the stored-program concept for electronic computing, including the possibility of allowing the machine to modify its own program in useful ways while running (for example, in order to control loops and branching). Turing's paper of 1936 (‘On Computable Numbers, with an Application to the Entscheidungsproblem’) was required reading for members of von Neumann's post-war computer project at the Institute for Advanced Study, Princeton University (Julian Bigelow in personal communication with William Aspray, reported in the Aspray [1990], pp. 178, 313). Eckert appears to have realised independently, and prior to von Neumann's joining the ENIAC group, that the way to take full advantage of the speed at which data is processed by electronic circuits is to place suitably encoded instructions for controlling the processing in the same high-speed storage devices that hold the data itself (Huskey in interview with Copeland, 1998). In 1945, while ENIAC was still under construction, von Neumann produced a draft report, mentioned previously, setting out the ENIAC group's ideas for an electronic stored-program general-purpose digital computer, the EDVAC (von Neuman [1945]). The EDVAC was completed six years later, but not by its originators, who left the Moore School to build computers elsewhere. Lectures held at the Moore School in 1946 on the proposed EDVAC were widely attended and contributed greatly to the dissemination of the new ideas.

Von Neumann was a prestigious figure and he made the concept of a high-speed stored-program digital computer widely known through his writings and public addresses. As a result of his high profile in the field, it became customary, although historically inappropriate, to refer to electronic stored-program digital computers as ‘von Neumann machines’.

The Los Alamos physicist Stanley Frankel, responsible with von Neumann and others for mechanising the large-scale calculations involved in the design of the atomic bomb, has described von Neumann's view of the importance of Turing's 1936 paper, in a letter to the historian Brian Randell:

> I know that in or about 1943 or ‘44 von Neumann was well aware of the fundamental importance of Turing's paper of 1936 ... Von Neumann introduced me to that paper and at his urging I studied it with care. Many people have acclaimed von Neumann as the "father of the computer" (in a modern sense of the term) but I am sure that he would never have made that mistake himself. He might well be called the midwife, perhaps, but he firmly emphasized to me, and to others I am sure, that the fundamental conception is owing to Turing, in so far as not anticipated by Babbage ... Both Turing and von Neumann, of course, also made substantial contributions to the "reduction to practice" of these concepts but I would not regard these as comparable in importance with the introduction and explication of the concept of a computer able to store in its memory its program of activities and of modifying that program in the course of these activities. (Quoted in Randell [1972], p. 10)

Other Notable Early Computers

Other notable early stored-program electronic digital computers were:

- EDSAC, 1949, built at Cambridge University by Maurice Wilkes
- BINAC, 1949, built by Eckert's and Mauchly's Electronic Control Co., Philadelphia (opinions differ over whether BINAC ever actually worked)
- Whirlwind I, 1949, Digital Computer Laboratory, Massachusetts Institute of Technology, Jay Forrester
- SEAC, 1950, US Bureau of Standards Eastern Division, Washington D.C., Samuel Alexander, Ralph Slutz
- SWAC, 1950, US Bureau of Standards Western Division, Institute for Numerical Analysis, University of California at Los Angeles, Harry Huskey
- UNIVAC, 1951, Eckert-Mauchly Computer Corporation, Philadelphia (the first computer to be available commercially in the U.S.)
- the IAS computer, 1952, Institute for Advanced Study, Princeton University, Julian Bigelow, Arthur Burks, Herman Goldstine, von Neumann, and others (thanks to von Neumann's publishing the specifications of the IAS machine, it became the model for a group of computers known as the Princeton Class machines; the IAS computer was also a strong influence on the IBM 701)
- IBM 701, 1952, International Business Machine's first mass-produced electronic stored-program computer.

High-Speed Memory

The EDVAC and ACE proposals both advocated the use of mercury-filled tubes, called ‘delay lines’, for high-speed internal memory. This form of memory is known as acoustic memory. Delay lines had initially been developed for echo cancellation in radar; the idea of using them as memory devices originated with Eckert at the Moore School. Here is Turing's description:

> It is proposed to build "delay line" units consisting of mercury ... tubes about 5′ long and 1″ in diameter in contact with a quartz crystal at each end. The velocity of sound in ... mercury ... is such that the delay will be 1.024 ms. The information to be stored may be considered to be a sequence of 1024 ‘digits’ (0 or 1) ... These digits will be represented by a corresponding sequence of pulses. The digit 0 ... will be represented by the absence of a pulse at the appropriate time, the digit 1 ... by its presence. This series of pulses is impressed on the end of the line by one piezo-crystal, it is transmitted down the line in the form of supersonic waves, and is reconverted into a varying voltage by the crystal at the far end. This voltage is amplified sufficiently to give an output of the order of 10 volts peak to peak and is used to gate a standard pulse generated by the clock. This pulse may be again fed into the line by means of the transmitting crystal, or we may feed in some altogether different signal. We also have the possibility of leading the gated pulse to some other part of the calculator, if we have need of that information at the time. Making use of the information does not of course preclude keeping it also. (Turing [1945], p. 24)

Mercury delay line memory was used in EDSAC, BINAC, SEAC, Pilot Model ACE, EDVAC, DEUCE,

and full-scale ACE (1958). The chief advantage of the delay line as a memory medium was, as Turing put it, that delay lines were "already a going concern" (Turing [1947], p. 108). The fundamental disadvantages of the delay line were that random access is impossible and, moreover, the time taken for an instruction, or number, to emerge from a delay line depends on where in the line it happens to be.

In order to minimize waiting-time, Turing arranged for instructions to be stored not in consecutive positions in the delay line, but in relative positions selected by the programmer in such a way that each instruction would emerge at exactly the time it was required, in so far as this was possible. Each instruction contained a specification of the location of the next. This system subsequently became known as ‘optimum coding’. It was an integral feature of every version of the ACE design. Optimum coding made for difficult and untidy programming, but the advantage in terms of speed was considerable. Thanks to optimum coding, the Pilot Model ACE was able to do a floating point multiplication in 3 milliseconds (Wilkes's EDSAC required 4.5 milliseconds to perform a single fixed point multiplication).

In the Williams tube or electrostatic memory, previously mentioned, a two-dimensional rectangular array of binary digits was stored on the face of a commercially-available cathode ray tube. Access to data was immediate. Williams tube memories were employed in the Manchester series of machines, SWAC, the IAS computer, and the IBM 701, and a modified form of Williams tube in Whirlwind I (until replacement by magnetic core in 1953).

Drum memories, in which data was stored magnetically on the surface of a metal cylinder, were developed on both sides of the Atlantic. The initial idea appears to have been Eckert's. The drum provided reasonably large quantities of medium-speed memory and was used to supplement a high-speed acoustic or electrostatic memory. In 1949, the Manchester computer was successfully equipped with a drum memory; this was constructed by the Manchester engineers on the model of a drum developed by Andrew Booth at Birkbeck College, London.

The final major event in the early history of electronic computation was the development of magnetic core memory. Jay Forrester realised that the hysteresis properties of magnetic core (normally used in transformers) lent themselves to the implementation of a three-dimensional solid array of randomly accessible storage points; in 1949, at Massachusetts Institute of Technology, he began to investigate this idea empirically. Forrester's early experiments with metallic core soon led him to develop the superior ferrite core memory (Forrester in interview with Evans, op. cit.). Digital Equipment Corporation undertook to build a computer similar to the Whirlwind I as a test vehicle for a ferrite core memory. The Memory Test Computer was completed in 1953. (This computer was used in 1954 for the first simulations of neural networks, by Belmont Farley and Wesley Clark of MIT's Lincoln Laboratory (see Copeland and Proudfoot [1996]).

Once the absolute reliability, relative cheapness, high capacity and permanent life of ferrite core memory became apparent, core soon replaced other forms of high-speed memory. The IBM 704 and 705 computers (announced in May and October 1954, respectively) brought core memory into wide use.

Bibliography

Works Cited

- Aspray, W., 1990, *John von Neumann and the Origins of Modern Computing*, Cambridge, Mass.: MIT Press
- Babbage, C. (ed. by Campbell-Kelly, M.), 1994, *Passages from the Life of a Philosopher*, New Brunswick: Rutgers University Press
- Bennett, S., unpublished, 'F.C. Williams: his contribution to the development of automatic control' (typescript based on interviews with Williams in 1976)
- Carpenter, B.E., and Doran, R.W. (eds), 1986, *A. M. Turing's ACE Report of 1946 and Other Papers* Cambridge, Mass.: MIT Press
- Copeland, B.J. (ed.), 1998, 'The Turing-Wilkinson Lecture Series on the Automatic Computing Engine', in Furukawa, K., Michie, D., Muggleton, S. (eds), *Machine Intelligence 15*, Oxford University Press (1998): 381-444
- Copeland, B.J., and Proudfoot, D., 1996, 'On Alan Turing's Anticipation of Connectionism' *Synthese*, **108**: 361-377
- Evans, C., 1976, 'The Pioneers of Computing: an Oral History of Computing', London Science Museum
- Fifer, S., 1961, *Analog Computation: Theory, Techniques, Applications* New York: McGraw-Hill
- Ford, H., 1919, 'Mechanical Movement', *Official Gazette of the United States Patent Office*, October 7, 1919: 48
- Goldstine, H., 1972, *The Computer from Pascal to von Neumann* Princeton University Press
- Newman, M.H.A., 1948, 'General Principles of the Design of All-Purpose Computing Machines' *Proceedings of the Royal Society of London*, series A, **195** (1948): 271-274
- Randell, B., 1972, 'On Alan Turing and the Origins of Digital Computers', in Meltzer, B., Michie, D. (eds), *Machine Intelligence 7*, Edinburgh: Edinburgh University Press, 1972
- Smith, B.C., 1991, 'The Owl and the Electric Encyclopaedia', *Artificial Intelligence*, **47**: 251-288
- Thomson, J., 1876, 'On an Integrating Machine Having a New Kinematic Principle' *Proceedings of the Royal Society of London*, **24**: 262-5
- Turing, A.M., 1936, 'On Computable Numbers, with an Application to the Entscheidungsproblem' *Proceedings of the London Mathematical Society*, Series 2, **42** (1936-37): 230-265
- Turing, A.M, 1945, 'Proposal for Development in the Mathematics Division of an Automatic Computing Engine (ACE)', in Carpenter and Doran [1986]
- Turing, A.M., 1947, 'Lecture to the London Mathematical Society on 20 February 1947', in Carpenter and Doran [1986]
- von Neumann, J., 1945, 'First Draft of a Report on the EDVAC', reprinted in full in Stern, N. *From ENIAC to UNIVAC: An Appraisal of the Eckert-Mauchly Computers* Bedford, Mass.: Digital Press (1981), pp. 181-246
- Williams, F.C., 1975, 'Early Computers at Manchester University' *The Radio and Electronic Engineer*, **45** (1975): 237-331
- Wynn-Williams, C.E., 1932, 'A Thyratron "Scale of Two" Automatic Counter' *Proceedings of the*

Royal Society of London, series A, **136**: 312-324

Further Reading

- Hinsley, H., and Stripp, A. (eds), 1993, 1994, *Codebreakers: The Inside Story of Bletchley Park* Oxford University Press
- Metropolis, N., Howlett, J., Rota, G.C. (eds), 1980, *A History of Computing in the Twentieth Century* New York: Academic Press
- Randell, B. (ed.), 1982, *The Origins of Digital Computers: Selected Papers* Berlin: Springer-Verlag

Other Internet Resources

- The Turing Archive for the History of Computing
- The Alan Turing Home Page
- Australian Computer Museum Society
- The Bletchley Park Home Page
- Charles Babbage Institute
- Computational Logic Group at St. Andrews
- The Computer Conservation Society (UK)
- CSIRAC (a.k.a. CSIR MARK I) Home Page
- Logic and Computation Group at Penn
- National Archive for the History of Computing
- National Cryptologic Museum
- The Virtual Museum of Computing

Related Entries

computability theory | function: recursive | Turing, Alan | Turing machine

B. Jack Copeland
bjcopeland@canterbury.ac.nz

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 17, 2000
Content last modified: December 17, 2000

Connectionism

Connectionism is a movement in cognitive science which hopes to explain human intellectual abilities using artificial neural networks (also known as ‘neural networks’ or ‘neural nets’). Neural networks are simplified models of the brain composed of large numbers of units (the analogs of neurons) together with weights that measure the strength of connections between the units. These weights model the effects of the synapses that link one neuron to another. Experiments on models of this kind have demonstrated an ability to learn such skills as face recognition, reading, and the detection of simple grammatical structure.

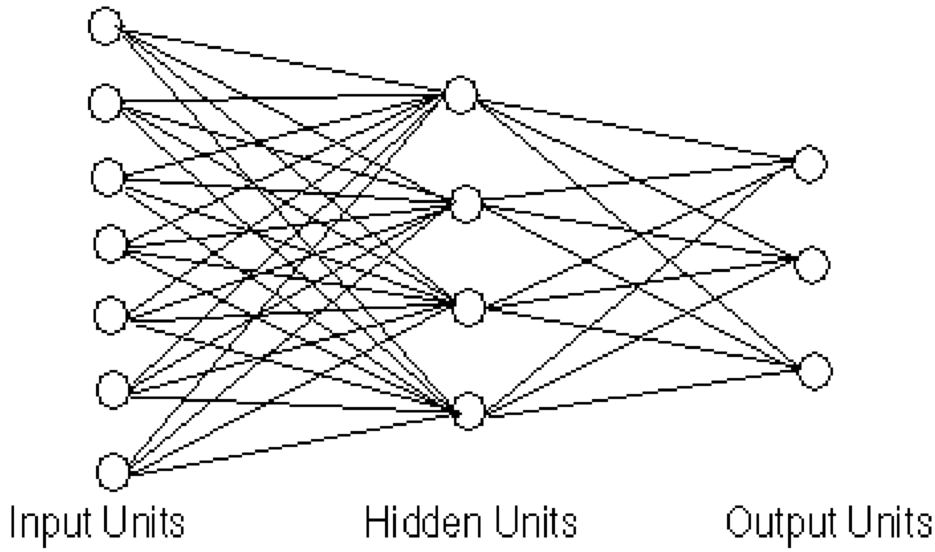
Philosophers have become interested in connectionism because it promises to provide an alternative to the classical theory of the mind: the widely held view that the mind is something akin to a digital computer processing a symbolic language. Exactly how and to what extent the connectionist paradigm constitutes a challenge to classicism has been a matter of hot debate in recent years.

A Description of Neural Networks

A neural network consists of large number of units joined together in a pattern of connections. Units in a net are usually segregated into three classes: input units, which receive information to be processed, output units where the results of the processing are found, and units in between called hidden units. If a neural net were to model the whole human nervous system, the input units would be analogous to the

sensory neurons, the output units to the motor neurons, and the hidden units to all other neurons.

Here is a simple illustration of a neural net:



Each input unit has an activation value that represents some feature external to the net. An input unit sends its activation value to each of the hidden units to which it is connected. Each of these hidden units calculates its own activation value depending on the activation values it receives from the input units. This signal is then passed on to output units or to another layer of hidden units. Those hidden units compute their activation values in the same way, and send them along to their neighbors. Eventually the signal at the input units propagates all the way through the net to determine the activation values at all the output units.

The pattern of activation set up by a net is determined by the weights, or strength of connections between the units. Weights may be both positive or negative. A negative weight represents the inhibition of the receiving unit by the activity of a sending unit. The activation value for each receiving unit is calculated according a simple activation function. Activation functions vary in detail, but they all conform to the same basic plan. The function sums together the contributions of all sending units, where the contribution of a unit is defined as the weight of the connection between the sending and receiving units times the sending unit's activation value. This sum is usually modified further, for example, by adjusting the activation sum to a value between 0 and 1 and/or by setting the activation to zero unless a threshold level for the sum is reached. Connectionists presume that cognitive functioning can be explained by collections of units that operate in this way. Since it is assumed that all the units calculate pretty much the same simple activation function, human intellectual accomplishments must depend primarily on the settings of the weights between the units.

Neural Network Learning and Backpropagation

Finding the right set of weights to accomplish a given task is the central goal in connectionist research.

Luckily, learning algorithms have been devised that can calculate the right weights for carrying out many tasks. (See Hinton (1992) for an accessible review.) One of the most widely used of these training methods is called backpropagation. To use this method one needs a training set consisting of many examples of inputs and their desired outputs for a given task. If, for example, the task is to distinguish male from female faces, the training set might contain pictures of faces together with an indication of the sex of the person depicted in each one. A net that can learn this task might have two output units (indicating "male" and "female") and many input units, one devoted to the brightness of each pixel (tiny area) in the picture. The weights of the net to be trained are initially set to random values, and then members of the training set are repeatedly exposed to the net. The values for the input of a member are placed on the input units and the output of the net is compared with the desired output for this member. Then all the weights in the net are adjusted slightly in the direction that would bring the net's output values closer to the values for the desired output. For example, when male's face is presented to the input units the weights are adjusted so that the value of the "male" output unit is increased and the value of the "female" output unit is decreased. After many repetitions of this process the net may learn to produce the desired output for each input in the training set. If the training goes well, the net many also have learned to generalize to the desired behavior for inputs and outputs that were not in the training set. For example, it may do a good job of distinguishing males from females in pictures that were never presented to it before.

Training nets to model aspects of human intelligence is a fine art. Success with backpropagation and other connectionist learning methods may depend on quite subtle adjustment of the algorithm and the training set. Training typically involves hundreds of thousands of rounds of weight adjustment. Given the limitations of computers presently available to connectionist researchers, training a net to perform an interesting task may take days or even weeks. Some of the difficulty may be resolved when parallel circuits specifically designed to run neural network models are widely available. But even here, some limitations to connectionist theories of learning will remain to be faced. Humans (and many less intelligent animals) display an ability to learn from single events; for example an animal that eats a food that later causes gastric distress will never try that food again. Connectionist learning techniques such as backpropagation are far from explaining this kind of "one shot" learning.

Samples of What Neural Networks Can Do

Connectionists have made significant progress in demonstrating the power of neural networks to master cognitive tasks. Here are three well-known experiments that have encouraged connectionists to believe that neural networks are good models of human intelligence. One of the most attractive of these efforts is Sejnowski and Rosenberg's (1987) work on a net that can read English text called NETtalk. The training set for NETtalk was a large data base consisting of English text coupled with its corresponding phonetic output, written in a code suitable for use with a speech synthesizer. Tapes of NETtalk's performance at different stages of its training are very interesting listening. At first the output is random noise. Later, the net sounds like it is babbling, and later still as though it is speaking English double-talk (speech that is formed of sounds that resemble English words). At the end of training, NETtalk does a fairly good job of pronouncing the text given to it. Furthermore, this ability generalizes fairly well to text that was not

presented in the training set.

Another influential early connectionist model was a net trained by Rumelhart and McClelland (1986) to predict the past tense of English verbs. The task is interesting because although most of the verbs in English (the regular verbs) form the past tense by adding the suffix ‘ed’, many of the most frequently verbs are irregular (is / was, come / came, go / went). The net was first trained on a set containing a large number of irregular verbs, and later on a set of 460 verbs containing mostly regulars. The net learned the past tenses of the 460 verbs in about 200 rounds of training, and it generalized fairly well to verbs not in the training set. It even showed a good appreciation of "regularities" to be found among the irregular verbs (send / sent, build / built; blow / blew, fly / flew). During learning, as the system was exposed to the training set containing more regular verbs, it had a tendency to overregularize, i.e. to combine both irregular and regular forms: (break / broked, instead of break / broke). This was corrected with more training. It is interesting to note that children are known to exhibit the same tendency to overregularize during language learning. However, there is hot debate over whether Rumelhart and McClelland’s is a good model of how humans actually learn and process verb endings. For example, (Pinker & Prince 1988) point out that the model does a poor job of generalizing to some novel regular verbs. They believe that this is a sign of a basic failing in connectionist models. Nets may be good at making associations and matching patterns, but they have fundamental limitations in mastering general rules such as the formation of the regular past tense. These complaints raise an important issue for connectionist modelers, namely whether nets can generalize properly to master cognitive tasks involving rules. Despite Pinker and Prince’s objections, many connectionists believe that generalization of the right kind is still possible (Niklasson and van Gelder, 1994).

Elman’s (1991) work on nets that can appreciate grammatical structure has important implications for the debate about whether neural networks can learn to master rules. Elman trained a neural network to predict the next word in a large corpus of English sentences. The sentences were formed from a simple vocabulary of 23 words using a subset of English grammar. The grammar, though simple, posed a hard test for linguistic awareness. It allowed unlimited formation of relative clauses while demanding agreement between the head noun and the verb. So for example, in the sentence

> Any **man** that chases dogs that chase cats .. run**s**.

the singular ‘**man**’ must agree with the verb ‘run**s**’ despite the intervening plural nouns (‘dogs’, ‘cats’) which might cause the selection of ‘run’. One of the important features of Elman’s model is the use of recurrent connections. The values at the hidden units are saved in a set of so called context units, to be sent back to the input level for the next round of processing. This looping back from hidden to input layers provides the net with a rudimentary form of memory of the sequence of words in the input sentence. Elman’s nets displayed an appreciation of the grammatical structure of sentences that were not in the training set. The net’s command of syntax was measured in the following way. Predicting the next word in an English sentence is, of course, and impossible task. However, these nets succeeded, at least by the following measure. At a given point in an input sentence, the output units for words that are grammatical continuations of the sentence at that point should be active and output units for all other words should be inactive. After intensive training, Elman was able to produce nets that displayed perfect

performance on this measure including sentences not in the training set. Although this performance is impressive, there is still a long way to go in training nets that can process language. Furthermore, doubts have been raised about the significance of Elman's results. For example, Marcus (to appear) argues that Elman's nets are not able to generalize this performance to sentences formed from a novel vocabulary. This, he claims, is a sign that connectionist models merely associate instances, and are unable to truly master abstract rules.

Strengths and Weaknesses of Neural Network Models

Philosophers are interested in neural networks because they may provide a new framework for understanding the nature of the mind and its relation to the brain (Rumelhart and McClelland, 1986, Chapter 1). Connectionist models seem particularly well matched to what we know about neurology. The brain is indeed a neural net, formed from massively many units (neurons) and their connections (synapses). Furthermore, several properties of neural network models suggest that connectionism may offer an especially faithful picture of the nature of cognitive processing. Neural networks exhibit robust flexibility in the face of the challenges posed by the real world. Noisy input or destruction of units causes graceful degradation of function. The net's response is still appropriate, though somewhat less accurate. In contrast, noise and loss of circuitry in classical computers typically result in catastrophic failure. Neural networks are also particularly well adapted for problems that require the resolution of many conflicting constraints in parallel. There is ample evidence from research in artificial intelligence that cognitive tasks such as object recognition, planning, and even coordinated motion present problems of this kind. Although classical systems are capable of multiple constraint satisfaction, connectionists argue that neural network models provide much more natural mechanisms for dealing with such problems.

Over the centuries, philosophers have struggled to understand how our concepts are defined. It is now widely acknowledged that trying to characterize ordinary notions with necessary and sufficient conditions is doomed to failure. Exceptions to almost any proposed definition are always waiting in the wings. For example, one might propose that a tiger is a large black and orange feline. But then what about albino tigers? Philosophers and cognitive psychologists have argued that categories are delimited in more flexible ways, for example via a notion of family resemblance or similarity to a prototype. Connectionist models seem especially well suited to accommodating graded notions of category membership of this kind. Nets can learn to appreciate subtle statistical patterns that would be very hard to express as hard and fast rules. Connectionism promises to explain flexibility and insight found in human intelligence using methods that cannot be easily expressed in the form of exception free principles (Horgan and Tienson, 1989, 1990), thus avoiding the brittleness that arises from standard forms of symbolic representation.

Despite these intriguing features, there are some weaknesses in connectionist models that bear mentioning. First, most neural network research abstracts away from many interesting and possibly important features of the brain. For example, connectionists usually do not attempt to explicitly model the variety of different kinds of brain neurons, nor the effects of neurotransmitters and hormones.

Furthermore, it is far from clear that the brain contains the kind of reverse connections that would be needed if the brain were to learn by a process like backpropagation, and the immense number of repetitions needed for such training methods seems far from realistic. Attention to these matters will probably be necessary if convincing connectionist models of human cognitive processing are to be constructed. A more serious objection must also be met. It is widely felt, especially among classicists, that neural networks are not particularly good at the kind of rule based processing that is thought to undergird language, reasoning, and higher forms of thought. We will discuss the matter further when we turn to the systematicity debate.

Connectionist Representation

Connectionist models provide a new paradigm for understanding how information might be represented in the brain. A seductive but naive idea is that single neurons (or tiny neural bundles) might be devoted to the representation of each thing the brain needs to record. For example, we may imagine that there is a grandmother neuron that fires when we think about our grandmother. However, such local representation is not likely. There is good evidence that our grandmother thought involves complex patterns of activity distributed across relatively large parts of cortex.

It is interesting to note that distributed, rather than local representations on the hidden units are the natural products of connectionist training methods. The activation patterns that appear on the hidden units while NETtalk processes text serve as an example. Analysis reveals that the net learned to represent such categories as consonants and vowels, not by creating one unit active for consonants and another for vowels, but rather in developing two different characteristic patterns of activity across all the hidden units.

Given the expectations formed from our experience with local representation on the printed page, distributed representation seems both novel and difficult to understand. But the technique exhibits important advantages. For example, distributed representations, (unlike symbols stored in separate fixed memory locations) remain relatively well preserved when parts of the model are destroyed or overloaded. More importantly, since representations are coded in patterns rather than firings of individual units, relationships between representations are coded in the similarities and differences between these patterns. So the internal properties of the representation carry information on what it is about (Clark 1993, p. 19). In contrast, local representation is conventional. No intrinsic properties of the representation (a unit's firing) determine its relationships to the other symbols. This self-reporting feature of distributed representations promises to resolve a philosophical conundrum about meaning. In a symbolic representational scheme, all representations are composed out of symbolic atoms (like words in a language). Meanings of complex symbol strings may be defined by the way they are built up out of their constituents, but what fixes the meanings of the atoms?

Connectionist representational schemes provide an end run around the puzzle by simply dispensing with atoms. Every distributed representation is a pattern of activity across all the units, so there is no principled way to distinguish between simple and complex representations. To be sure, representations are

composed out of the activities of the individual units. But none of these "atoms" codes for any symbol. The representations are sub-symbolic in the sense that analysis into their components leaves the symbolic level behind.

The sub-symbolic nature of distributed representation provides a novel way to conceive of information processing in the brain. If we model the activity of each neuron with a number, then the activity of the whole brain can be given by a giant vector (or list) of numbers, one for each neuron. Both the brain's input from sensory systems and its output to individual muscle neurons can also be treated as vectors of the same kind. So the brain amounts to a vector processor, and the problem of psychology is transformed into questions about which operations on vectors account for the different aspects of human cognition.

Sub-symbolic representation has interesting implications for the classical hypothesis that the brain must contain symbolic representations that are similar to sentences of a language. This idea, often referred to as the language of thought (or LOT) thesis may be challenged by the nature of connectionist representations. It is not easy to say exactly what the LOT thesis amounts to, but van Gelder (1990) offers an influential and widely accepted benchmark for determining when the brain should be said to contain sentence-like representations. It is that when a representation is tokened one thereby tokens the constituents of that representation. For example, if I write 'John loves Mary' I have thereby written the sentence's constituents: 'John' 'loves' and 'Mary'. Distributed representations for complex ideas like 'John loves Mary' can be constructed that do not contain any explicit representation of their parts (Smolensky 1991). The information about the constituents can be extracted from the representations, but neural network models do not need to explicitly extract this information themselves in order to process it correctly (Chalmers, 1990). This suggests that neural network models serve as counterexamples to the idea that the language of thought is a prerequisite for human cognition. However, the matter is still a topic of lively debate (Fodor, 1997).

The Shape of the Controversy between Connectionists and Classicists

The last thirty years have been dominated by the classical view that (at least higher) human cognition is analogous to symbolic computation in digital computers. On the classical account, information is represented by strings of symbols, just as we represent data in computer memory or on pieces of paper. The connectionist claims, on the other hand, that information is stored non-symbolically in the weights, or connection strengths, between the units of a neural net. The classicist believes that cognition resembles digital processing, where strings are produced in sequence according to the instructions of a (symbolic) program. The connectionist views mental processing as the dynamic and graded evolution of activity in a neural net, each unit's activation depending on the connection strengths and activity of its neighbors, according to the activation function.

On the face of it, these views seem very different. However many connectionists do not view their work as a challenge to classicism and some overtly support the classical picture. So-called implementational

connectionists seek an accommodation between the two paradigms. They hold that the brain's net implements a symbolic processor. True, the mind is a neural net; but it is also a symbolic processor at a higher and more abstract level of description. So the role for connectionist research according to the implementationalist is to discover how the machinery needed for symbolic processing can be forged from neural network materials, so that classical processing can be reduced to the neural network account.

However, many connectionists resist the implementational point of view. Such radical connectionists claim that symbolic processing was a bad guess about how the mind works. They complain that classical theory does a poor job of explaining graceful degradation of function, holistic representation of data, spontaneous generalization, appreciation of context, and many other features of human intelligence which are captured in their models. The failure of classical programming to match the flexibility and efficiency of human cognition is by their lights a symptom of the need for a new paradigm in cognitive science. So radical connectionists would eliminate symbolic processing from cognitive science forever.

The Systematicity Debate

The major points of controversy in the philosophical literature on connectionism have to do with whether connectionists provide a viable and novel paradigm for understanding the mind. One complaint is that connectionist models are only good at processing associations. But such tasks as language and reasoning cannot be accomplished by associative methods alone and so connectionists are unlikely to match the performance of classical models at explaining these higher-level cognitive abilities. However, it is a simple matter to prove that neural networks can do anything that symbolic processors can do since nets can be constructed that mimic a computer's circuits. So the objection can not be that connectionist models do not account for higher cognition; it is rather that they can do so only if they implement the classicist's symbolic processing tools. Implementational connectionism may succeed, but radical connectionists will never be able to account for the mind.

Fodor and Pylyshyn's often cited paper (1988) launches a debate of this kind. They identify a feature of human intelligence called systematicity which they feel connectionists cannot explain. The systematicity of language refers to the fact that the ability to produce/understand some sentences is intrinsically connected to the ability to produce/understand others of related structure. For example, no one with a command of English who understands 'John loves Mary' can fail to understand 'Mary loves John'. From the classical point of view, the connection between these two abilities can easily be explained by assuming that masters of English represent the constituents ('John', 'loves' and 'Mary') of , 'John loves Mary' and computes its meaning from the meanings of these constituents. If this is so, then understanding a novel sentence like 'Mary loves John' can be accounted for as another instance of the same symbolic process. In a similar way, symbolic processing would account for the systematicity of reasoning, learning and thought. It would explain why there are no people who are capable of concluding P from P&(Q&R), but incapable of concluding P from P&Q, why there are no people capable of learning to prefer red cube to green square who cannot learn to prefer a green cube to the red square, and why there isn't anyone who can think that John loves Mary who can't also think that Mary loves John.

Fodor and McLaughlin (1990) argue in detail that connectionists do not account for systematicity. Although connectionist models can be trained to be systematic, they can also be trained, for example, to recognize ‘John loves Mary’ without being able to recognize ‘Mary loves John’. Since connectionism does not guarantee systematicity, it does not explain why systematicity is found so pervasively in human cognition. Systematicity may exist in connectionist architectures, but where it exists, it is no more than a lucky accident. The classical solution is much better, because in classical models, pervasive systematicity comes for free.

The charge that connectionist nets cannot explain systematicity is initially quite plausible. However, careful analysis of the content of the claim is needed (Hadley, 1994). Furthermore, the view has been criticized lately by Aizawa (to appear), Garson (to appear), and Wallis, (to appear). One point common to these rebuttals is that symbolic processing models have exactly the same feature which was supposed to deny connectionists an ability to explain systematicity, for there are also classical models that can be programmed to accept ‘John loves Mary’ and reject ‘Mary loves John’.

Connectionism and the Elimination of Folk Psychology

Another important application of connectionist research to philosophical debate about the mind concerns the status of folk psychology. Folk psychology is the conceptual structure that we spontaneously apply to understanding and predicting human behavior. For example, knowing that John desires a beer and that he believes that there is one in the refrigerator allows us to explain why John just went into the kitchen. Such knowledge depends crucially on our ability to conceive of others as having desires and goals, plans for satisfying them, and beliefs to guide those plans. The idea that people have beliefs, plans and desires is a commonplace of ordinary life; but does it provide a faithful description of what is actually to be found in the brain?

Its defenders will argue that folk psychology is too good to be false (Fodor, 1988, Ch1). What more can we ask for the truth of a theory than that it provides an indispensable framework for successful negotiations with others? On the other hand, eliminativists will respond that the useful and widespread use of a conceptual scheme does not argue for its truth (Churchland 1989, Ch. 1). Ancient astronomers found the notion of celestial spheres useful (even essential) to the conduct of their discipline, but now we know that there are no celestial spheres. From the eliminativists point of view, an allegiance to folk psychology, like allegiance to folk (Aristotelian) physics, stands in the way of scientific progress. A viable psychology may require as radical a revolution in its conceptual foundations as is found in quantum mechanics.

Eliminativists are interested in connectionism because it promises to provide a conceptual foundation that would replace folk psychology. Simple cognitive tasks can be performed by neural networks that do not appear to contain any structures that could correspond to beliefs, desires and plans (Ramsey et. al., 1991). It is still an open question as to whether the complexities of human cognition can ever be captured by

such connectionist models. Furthermore, the whole issue of exactly what evidence about the brain would support the view that beliefs and desires are actively involved in the brain's processing is a cloudy one. The question is complicated further by disagreements about the nature of folk psychology. Many philosophers treat the beliefs and desires postulated by folk psychology as brain states with symbolic contents. For example, the belief that there is a beer in the refrigerator is thought to be a brain state that contains symbols corresponding to beer and a refrigerator. From this point of view, the fate of folk psychology is strongly tied to the symbolic processing hypothesis. On the other hand, some philosophers do not think folk psychology is essentially symbolic, and some would even challenge the idea that folk psychology is to be treated as a scientific theory in the first place. Under this conception, it is much more difficult to forge links between results in connectionist research and the rejection of folk psychology.

Bibliography

- Aizawa, K. "Representations without Rules, Connectionism and the Syntactic Argument," (to appear)
- Bechtel, W. "Connectionism and Rules and Representation Systems: Are They Compatible?," *Philosophical Psychology*, **1** (1988): 5-15
- Bechtel, W. "Connectionism and the Philosophy of Mind: an Overview," *The Southern Journal of Philosophy*, Supplement, (1987): 17-41
- Bechtel, W. and Abrahamsen, A. *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*, Cambridge, Mass.: Blackwell (1990)
- Butler, K."Towards a Connectionist Cognitive Architecture," *Mind and Language*, **6** (1991): 252-272
- Chalmers, D. "Syntactic Transformations on Distributed Representations," *Connection Science*, **2** (1990): 53-62
- Chalmers, D. "Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation," *Philosophical Psychology*, (1993): 305-319
- Christiansen, M and Chater, N. "Generalization and Connectionist Language Learning," *Mind and Language*, **9, #3** (1994): 273-287
- Churchland, P.M. *The Engine of Reason, the Seat of the Soul : a Philosophical Journey into the Brain*, Cambridge, Mass. : MIT Press (1995)
- Churchland, P. S. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, Cambridge, Mass. :MIT Press (1989)
- Clark, A. *Associative Engines*, Cambridge, Mass.: MIT Press (1993)
- Clark, A. *Microcognition*, Cambridge, Mass.: MIT Press (1989)
- Clark, A. and Lutz, R. (Eds.) *Connectionism in Context*, Springer, (1992)
- Cotrell G. and Small, S. "A Connectionist Scheme for Modelling Word Sense Disambiguation," *Cognition and Brain Theory*, **6** (1983): 89-120
- Cummins, R. "The Role of Representation in Connectionist Explanations of Cognitive Capacities," in Ramsey, Stich and Rumelhart (1991): 91-114
- Cummins, R. "Systematicity" *Journal of Philosophy*, **XCIII #22** (1996): 561-614
- Cummins, R. and Schwarz, G. "Connectionism, Computation, and Cognition," in Horgan and

Tienson (1991): 60-73

- Davies, M. "Connectionism, Modularity and Tacit Knowledge," *British Journal for the Philosophy of Science*, **40** (1989): 541-555
- Davies, M. "Concepts, Connectionism and the Language of Thought," in Ramsey et. al. (1991): 229-257
- Dinsmore, J. (ed.) *The Symbolic and Connectionist Paradigms: Closing the Gap*, Erlbaum (1992)
- Elman, J. L. "Distributed Representations, Simple Recurrent Networks, and Grammatical Structure," in Touretzky (1991): 91-122
- Fodor, J. "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Still Doesn't Work," *Cognition*, **62** (1997): 109-119
- Fodor, J. *Psychosemantics*, Cambridge, Mass.: MIT Press (1988)
- Fodor, J. and McLaughlin, B. "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work,"*Cognition*, **35** (1990): 183-204
- Fodor, J. and Pylyshyn, Z. "Connectionism and Cognitive Architecture: a Critical Analysis," *Cognition*, **28** (1988): 3-71
- Garfield, J. "Mentalese Not Spoken Here: Computation Cognition and Causation," (to appear)
- Garson, J. "What Connectionists Cannot Do: The Threat to Classical AI," in Horgan, T. and Tienson, J. (1991): 113-142
- Garson, J. "Cognition without Classical Architecture" *Synthese*, **100** (1994): 291-305
- Garson, J. "Syntax in a Dynamic Brain" *Synthese*, **110** (1997): 343-355
- Garson, J. "Systematicity and Classical Architecture" (to appear)
- Hadley, R. (1994) "Systematicity in Connectionist Language Learning," *Mind and Language*, vol 9, #3, pp. 247-271
- Hanson, J. and Kegl, J. (1987) "PARSNIP: A Connectionist Network that Learns Natural Language Grammar from Exposure to Natural Language Sentences," *Ninth Annual Conference of the Cognitive Science Society*, pp. 106-119
- Hatfield, G. "Representation in Perception and Cognition: Connectionist Affordances," in Ramsey et. al. (1991): 163-195
- Hatfield, G. "Representation and Rule-Instantiation in Connectionist Systems," in Horgan and Tienson (1991): 90-112
- Hawthorne, J. "On the Compatibility of Connectionist and Classical Models," *Philosophical Psychology*, **2** (1989): 5-15
- Hinton, G. "How Neural Networks Learn from Experience," *Scientific American* (September, 1992): 145-151
- Hinton, G., ed. *Connectionist Symbol Processing*, Cambridge, Mass.: MIT Press (1991)
- Hinton, G. "Mapping Part-Whole Hierarchies into Connectionist Networks," in Hinton (1991): 47-76
- Hinton, G., McClelland, and Rumelhart, D. "Distributed Representations," chapter 3 of Rumelhart, McClelland, et. al. (1986)
- Horgan, T. and Tienson, J. "Representations without Rules," *Philosophical Topics*, **17** (1989): 147-174
- Horgan, T. and Tienson, J. "Soft Laws," *Midwest Studies in Philosophy* **15** (1990): 256-279
- Horgan, T. and Tienson, J. *Connectionism and the Philosophy of Psychology*, Cambridge, Mass.:

MIT Press, (1996)

- Horgan, T. and Tienson, J. (eds.) *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer, (1991)
- Macdonald, C., ed. *Connectionism: Debates on Psychological Explanation,* Oxford: Blackwell (1995)
- Marcus, G. "Rethinking Eliminative Connectionism," (to appear)
- McClelland, J. and Elman, J. "The TRACE Model of Speech Perception," *Cognitive Psychology*, **18** (1986): 1-86
- McClelland, J., Rumelhart, D., et. al., *Parallel Distributed Processing*, vol. II, Cambridge, Mass.: MIT Press (1986)
- Miikkulainen. T. *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon and Memory*, Cambridge, Mass.: MIT Press (1993)
- Niklasson, L. and van Gelder, T. "On Being Systematically Connectionist," *Mind* and Language, **9**/3 (1994): 288-302
- Pinker, S. and Mehler, J., eds. *Connections and Symbols,* Cambridge, Mass.: MIT Press (1988)
- Pinker, S. and Prince, A. "On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition," *Cognition*, **23** (1988) 73-193
- Pollack, J. "Implications of Recursive Distributed Representations," in Touretzky (1989): 527-535.
- Pollack, J. "Induction of Dynamical Recognizers," in Touretzky (1991): 123-148
- Pollack, J. (1991) "Recursive Distributed Representation," in Hinton (1991): 77-106.
- Port, Robert, F. "Representation and Recognition of Temporal Patterns," *Connection Science*, **2** (1990): 151-176
- Port, R. and van Gelder, T. "Representing Aspects of Language," *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, Hillsdale, N.J.: Erlbaum (1991)
- Ramsey, W., Stich, S. and Rumelhart, D. *Philosophy and Connectionist Theory*, Hillsdale, N.J.: Erlbaum (1991)
- Ramsey, W., Stich, S. and Garon, J. "Connectionism, Eliminativism, and the Future of Folk Psychology," in Ramsey, Rumelhart and Stich (1991): 199-228
- Rumelhart, D. and McClelland, J. "On Learning the Past Tenses of English Verbs," in McClelland and Rumelhart et. al. Ch. 18, (1986): 216-271
- Rumelhart, D., McClelland, J., and the PDP Research Group, *Parallel Distributed Processing*, vol. I, Cambridge, Mass.: MIT Press (1986)
- Schwarz, G. "Connectionism, Processing, Memory," *Connection Science*, **4** (1992): 207-225.
- Sejnowski, T. and Rosenberg, C. "Parallel networks that Learn to Pronounce English Text," *Complex Systems*, **1** (1987): 145-168.
- Servan-Schreiber, D., Cleeremans, A. and McClelland, J. "Graded State Machines: The Representation of Temporal Contingencies in Simple Recurrent Networks," in Touretzky (1991): 57-89
- Shastri, L. and Ajjanagadde, V. "From Simple Associations to Systematic Reasoning: A Connectionist Representation of Rules, Variables, and Dynamic Bindings Using Temporal Synchrony" *Behavioral and Brain Sciences*, **16**/3 (1993): 417-494
- Smolensky, P. "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems," in Hinton (1991): 159-216

- Smolensky, P. "The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn," *The Southern Journal of Philosophy*, Supplement, **26** (1987): 137-161.
- Smolensky, P. "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences*, **11** (1988): 1-74.
- St. John, M. and McClelland, J. "Learning and Applying Contextual Constraints in Sentence Comprehension," in Hinton (1991): 217-257
- Touretzky, D. *Advances in Neural Information Processing Systems I*, San Mateo, CA: Kaufmann (1989)
- Touretzky, D. *Advances in Neural Information Processing Systems II*, San Mateo, CA: Kaufmann (1990)
- Touretzky, D. *Connectionist Approaches to Language Learning*, Dordrecht: Kluwer (1991)
- Touretzky, D., Hinton, G. and Sejnowski, T. *Proceedings of the 1988 Connectionist Models Summer School*, Kaufmann (1988)
- van Gelder, T. "Compositionality: A Connectionist Variation on a Classical Theme," *Cognitive Science*, **14** (1990): 355-384.
- van Gelder, T. "What is the 'D' in PDP?" in Ramsey et. al. (1991): 33-59.
- van Gelder, T and Port, R. "Beyond Symbolic: Prolegomena to a Kama-Sutra of Compositionality," in Honavar, V. and Uhr, L. (Eds.) *Symbol Processing and Connectionist Models in AI and Cognition: Steps Towards Integration*, Boston: Academic Press (1993)
- Wallis, C. "Nomic Necessity and Systematicity" (to appear).
- Waltz, D. and Pollack, J. "Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation," *Cognitive Science*, **9**, 51-74.

Other Internet Resources

- Bibliography on Connectionism (by David Chalmers)
- An Introduction to Neural Nets by Z Solutions
- Connectionism: A Short Reading List (by Ezra van Everbroeck)
- Gateway to Neural Networks

Related Entries

artificial intelligence | language of thought hypothesis | mental representation

James W. Garson
JGarson@uh.edu

Table of Contents

First published: May 18, 1997
Content last modified: July 29, 1997

Mental Representation

If a representation is an object with semantic properties, then a mental representation is a mental object with semantic properties. According to the Representational Theory of Mind (RTM), psychological states are to be understood as relations between agents and mental representations: for an agent to be in a psychological state Ψ with semantic property Φ is for that agent to be in a Ψ-appropriate relation to a mental representation of an appropriate kind with semantic property Φ.

Historically, RTM (which goes back at least to Aristotle) is a theory of commonsense psychological states, such as belief, desire (the *propositional attitudes*), and perception. According to RTM, to *believe that p*, for example, is, in part, to bear the belief-relation (whatever that may be) to a mental representation that *means that p*. To perceive that *a* is Φ is, in part (propositional attitudes may also be involved), to have a sensory experience of some kind which is appropriately related (however that may be) to *a*'s being Φ.

The leading contemporary version of RTM, the Computational Theory of Mind (CTM), makes the further claims that the brain is a kind of computer and that mental processes are computations on mental representations. According to CTM, cognitive states are constituted by *computational* relations to mental representations of various kinds, and cognitive processes are rule-governed sequences of such states.

CTM develops RTM by attempting to explain *all* psychological states and processes in terms of mental representation. In the course of constructing detailed empirical theories of human and animal cognition and developing models of cognitive processes implementable in artificial information processing systems, cognitive scientists have proposed a variety of types mental representations. While some of these may be suited to be mental relata of commonsense psychological states, some -- so-called "subpersonal" or "sub-doxastic" representations -- are not. Though many philosophers believe that CTM stands to provide the best scientific explanations of cognition and behavior, there is disagreement over whether or not such explanations will vindicate the commonsense psychological explanations (and representations) of prescientific RTM.

Mental representation has also been of interest to philosophers who hold that the semantic properties of expressions of natural language (and many non-linguistic symbols as well) are inherited from the mental states of their users. For these theorists, RTM is a component of a complete theory of linguistic meaning.

Propositional Attitudes

Intentional Realists such as Dretske (1988) and Fodor (1987) argue that explanations and predictions of human behavior must quantify over inner states that are both causally efficacious and contentful, and have their causal powers (their causal roles) in virtue of having the content they do (i.e., the generalizations that describe the interaction of such states apply to them in virtue of their content: the desire that *q* and the belief that not *q* unless *q* cause (*ceteris paribus*) the desire that *q*, for example, because they are the desire that *q* and the belief that not *q* unless *p*). The generalizations we apply in everyday life in explaining others' behavior (often collectively referred to as "folk psychology") are both remarkably successful and indispensable. What a person believes, doubts, desires, fears, etc. is a highly reliable indicator of what he will do; and we have no other way of making sense of his behavior than by ascribing such states to him and applying the relevant generalizations. We are thus committed to the generalizations of commonsense psychology and, hence, to the existence of the states they quantify over. Such states, the *propositional attitudes*, are individuated by their contents and their characteristic psychological roles: they are relations to mental representations. Some realists (such as Fodor) also hold that commonsense psychology will be vindicated by cognitive science, given that propositional attitudes can be construed as *computational* relations to mental representations.

Intentional Eliminativists, such as Churchland, Stich and (perhaps) Dennett argue that no such things as propositional attitudes (and their implicated representational states) are necessary to the explanation of our mental lives and behavior.

Churchland denies that the generalizations of commonsense propositional-attitude psychology are true. He (1981) argues that folk psychology is a *theory* of the mind with a long history of failure and decline, and that it resists incorporation into the framework of modern scientific theories (including cognitive psychology). As such, it is comparable to alchemy and phlogiston theory, and ought to suffer a comparable fate. Commonsense psychology is *false*, and the states (and representations) it quantifies over simply don't exist. (It should be noted that Churchland is not an eliminativist about representation *tout court*. See, e.g., Churchland 1989.)

Dennett (1987a) grants that the generalizations of commonsense psychology are true and indispensable, but denies that this is sufficient reason to believe in the entities they seem to quantify over. He argues that to give an intentional explanation of a system's behavior is merely to adopt the "intentional stance" toward it. If the strategy of assigning contentful states to a system and predicting and explaining its behavior (on the assumption that it is *rational* -- i.e., that it behaves as it should, given the propositional attitudes it should have in its environment) is successful, then the system is intentional, and the propositional-attitude generalizations we apply to it are true. But there is nothing more to having a propositional attitude than this. (See Dennett 1987a: 29.)

Though he has been taken to be thus claiming that intentional explanations should be construed *instrumentally*, Dennett (1991) insists that he is a "moderate" realist about propositional attitudes, since he believes that the patterns in the behavior and behavioral dispositions of a system on the basis of which we (truly) attribute intentional states to it are objectively real. In the event, however, that there are two or more explanatorily adequate but substantially different systems of intentional ascriptions to an individual, Dennett claims, there is no fact of the matter about what the system believes (1987b, 1991). This does suggest an irrealism at least with respect to the sorts of things Fodor and Dretske take beliefs to be, though it is not the view that there is simply *nothing* in the world that makes intentional explanations true.

(Davidson (1973, 1974) and Lewis (1974) also defend the view that what it is to have a propositional attitude is just to be interpretable in a particular way. It is, however, not completely clear whether they intend their views to imply irrealism about propositional attitudes.)

Stich (1983) argues that cognitive psychology does not (or, in any case, *should* not) taxonomize mental states by their semantic properties at all. The generalizations of a scientific psychology will not quantify over representational states *qua* representational, and commonsense psychology will not be vindicated by CTM (properly understood). Attribution of psychological states by content is, Stich believes, sensitive to factors that render it problematic in the context of a scientific psychology -- viz., relations between an organism and its environment, and relations among psychological states.

Cognitive psychology seeks systematic *causal* explanations of behavior and cognition, and the causal powers of a mental state are determined by its *intrinsic* "structural" or "syntactic" properties (Stich calls this "the principle of psychological autonomy." See Stich 1978.) The semantic properties of a mental state (paradigmatically, its *reference*), in contrast, are determined by *extrinsic* properties -- viz., its history and environmental relations. Thus, such properties cannot figure in causal explanations of behavior.

Moreover, to ascribe a psychological state to an individual *by content* is (roughly) to say that the individual is in a state content-identical to the sort of state that typically causes one's own utterances of the sentence appearing in the content clause of the attribution. But the appropriateness of such ascriptions depends on what *other* psychological states the ascribee is (or is disposed to be) in. Since attribution by content is thus *holistic*, content-based psychological explanation of subjects who differ substantially from the ascriber(s) in their total system of beliefs is precluded (the generalizations will not apply).

Finally, ascription of a psychological state by content is sensitive to what other states an individual is

disposed to be in *as a result of* being in that state (in particular, what *inferences* the subject is disposed to make). But this renders content-based theories unable to make sense of subjects whose states have inference potential substantially different from those of the theorist. (Stich makes these last two points using examples of individuals whose psychology is, from the point of view of the ascriber, *pathological*.)

Stich also rejects, for essentially the reasons just enumerated, what he calls the Weak Representational Theory of Mind, on which psychological states have content but psychological generalizations do not apply to them in virtue of it. He argues for a *syntactic* theory of the mind, on which the semantic properties of mental states play *no* explanatory role: only the syntactic properties of mental states are relevant to their computational profiles.

Computation and Cognition

According to Stich's Syntactic Theory of the Mind, computational theories of psychological states should concern themselves only with the *formal* properties of the objects those states are relations to. Commitment to the explanatory relevance of *content*, however, is for most cognitive scientists fundamental (Fodor 1981a, Pylyshyn 1984, Von Eckardt 1993). That mental processes are computations, that computations are rule-governed sequences *of semantically evaluable objects* (*symbols*), and that the rules apply to the symbols in virtue of their content, are central tenets of mainstream cognitive science.

Explanations in cognitive science appeal to a variety of types of mental representation, including, for example, the "mental models" of Johnson-Laird 1983, the "retinal arrays," "primal sketches" and "2½ -D sketches" of Marr 1982, the "sub-symbolic" structures of Smolensky 1989, the "quasi-pictures" of Kosslyn 1980, and the "interpreted symbol-filled arrays" of Tye 1991 -- in addition to representations that may be appropriate to the explanation of commonsense psychological states. Computational explanations have been offered of, among other mental phenomena, belief (Fodor 1975, Field 1978), visual perception (Marr 1982, Osherson, et al. 1990), rationality (Newell and Simon 1972, Fodor 1975, Johnson-Laird and Wason 1977), language learning and use (Chomsky 1965, Pinker 1989), and musical comprehension (Lerdahl and Jackendoff 1983).

There is, however, disagreement among proponents of CTM as to what kinds of representations the brain uses, what kinds of neural structures realize them, and what kinds of brain-processes realize computations -- in short, on what *kind* of computer the brain is. The central debate here is between proponents of *Classical Architectures* and proponents of *Connectionist Architectures*.

The Classicists (e.g., Turing 1950, Fodor 1975, Newell and Simon 1976, Marr 1982, Fodor and Pylyshyn 1988) hold that mental representations are symbolic structures, which typically have semantically evaluable constituents, and mental processes are rule-governed manipulations of them. The Connectionists (e.g., McCulloch and Pitts 1943, Rumelhart and McClelland 1986, Rumelhart 1989, Smolensky 1988) hold that mental representations are realized by patterns of activation in a network of simple processors ("nodes") and mental processes consist of the spreading activation of such patterns. The nodes themselves are, typically, not taken to be semantically evaluable; nor do the patterns have

semantically evaluable constituents. (Though there are versions of Connectionism -- "localist" versions -- on which individual nodes are taken to have semantic properties (Ballard 1986, Ballard and Hayes, 1984). It is arguable, however, that localist theories are neither definitive nor representative of the connectionist program (Smolensky 1988, 1991; Chalmers 1993).)

The Classicists are motivated (in part) by properties thought seems to share with language. Fodor's Language of Thought Hypothesis (LOTH) (Fodor 1975, 1987), according to which the system of mental symbols constituting the neural basis of thought is taken to be structured like a symbolic language, provides a well-worked-out version of the Classical approach as applied to commonsense psychology. (Cf. also Marr 1982 for an application of classical approach in scientific psychology.) According to the LOTH, the potential infinity of complex representational mental states isgenerated from a finite stock of primitive representational states, combined in accordance with recursive rules. This combinatorial structure accounts for the properties of *productivity* and *systematicity* of the system of mental representation. A representational system is productive if there are indefinitely many distinct representations that may be constructed in it; it is systematic if the constructability of some representations is intrinsically connected to the constructability of others. As in the case of symbolic languages, including natural languages (though Fodor does not suppose either that the LOTH explains only linguistic capacities or that only verbal creatures have this sort of cognitive architecture), these properties of thought are explained by appeal to the independent contentfulness of the representational units and their combinability into contentful complexes. That is, the semantics of both language and thought is *compositional*: the content of a complex representation is determined by the contents of its constituents and their structural configuration.

The Connectionists are motivated mainly by a consideration of the architecture of the brain, which apparently consists of layered networks of interconnected neurons. They argue that this sort of architecture is unsuited to carrying out classical serial computations. For one thing, processing in the brain is typically massively parallel. In addition, the elements whose manipulation drives computation in connectionist networks (principally, the connections between nodes) are neither semantically compositional nor semantically evaluable, as they are on the classical approach. This contrast with classical computationalism is often characterized by saying that representation is, with respect to computation, *distributed* as opposed to *local*: representation is local if it is computationally basic; and distributed if it is not. (Another way of putting this is to say that for the Classicist mental representations are computationally *atomic*, whereas for the Connectionist they are not.)

Moreover, connectionists argue that information processing as it occurs in connectionist networks more closely resembles actual human cognitive functioning. For example, whereas on the classical view learning involves something like hypothesis formation and testing (Fodor 1981c), on the connectionist model it is a matter of evolving distribution of weights on the connections between nodes, and typically does not involve the representation of identity conditions for the objects of knowledge. The connectionist network is "trained up" by repeated exposure to objects it is to learn to distinguish; and this seems to model at least one type of human learning quite well. Further, degradation in the performance of such networks in response to damage is a gradual, not sudden as in the case of a classical information processor, and hence more accurately models the loss of human cognitive function as it typically occurs in

response to brain damage. It is also sometimes claimed that connectionist systems show the kind of flexibility in response to novel situations typical of human cognition -- situations in which classical systems are relatively "brittle" or "fragile."

Some philosophers have maintained that Connectionism entails irrealism about propositional attitudes. Ramsey, Stich and Garon (1990) have argued that if Connectionist models of cognition are basically correct, then there are no discrete representational states as conceived in ordinary commonsense psychology and classical cognitive science. Others, however (e.g., Smolensky 1989), hold that certain types of higher-level patterns of activity in a neural network may be roughly identified with the representational states of commonsense psychology. Still others (e.g., Fodor and Pylyshyn 1988, Heil 1991, Horgan and Tienson 1996) argue that language-of-thought style representation is both necessary in general and realizable within connectionist architectures. (MacDonald and MacDonald 1995 collects the central contemporary papers in the Classicist/Connectionist debate, and provides useful introductory material as well. See also Von Eckardt forthcoming.)

Whereas Stich (1983) accepts that mental processes are computational, but denies that computations are sequences of mental representations, others accept the notion of mental representation, but deny that CTM provides the correct account of mental states and processes.

Van Gelder (1995), for example, denies that psychological processes are computational. He argues that cognitive systems are *dynamic*, and that cognitive states are not relations to mental *symbols*, but quantifiable states of a complex system consisting of (in the case of human beings) a nervous system, a body and the environment in which they are embedded. Cognitive processes are not rule-governed sequences of discrete symbolic states, but continuous, evolving total states of dynamic systems determined by continuous, simultaneous and mutually determining states of the systems' components. Representation in a dynamic system is essentially information-theoretic; though the bearers of information are not symbols, but state variables or parameters. (See also Port and Van Gelder 1995.)

Horst (1996), on the other hand, argues that though computational models may be useful in scientific psychology, they are of no help in achieving a philosophical understanding of the intentionality of commonsense mental states. CTM attempts to *reduce* the intentionality of such states to the intentionality of the mental symbols they are relations to. But, he claims, the relevant notion of symbolic content is essentially bound up with the notions of convention and intention. So CTM involves itself in a vicious circularity: the very properties that are supposed to be reduced are (tacitly) appealed to in the reduction.

Content Determination

Another important issue for proponents of RTM is how mental representations *come to have* their semantic properties. There are two basic types of theory of content determination, *informational* theories and *functional* theories. Though theories of these types were designed to account for the content of commonsense psychological states, they may, at least in broad outline, serve as theories of content determination for sub-personal representational states as well.

Informational theories (Dretske 1981, 1988) hold that the content of a mental representation is grounded in the information it carries about what does (Devitt 1996) or would (Fodor 1987, 1990) cause its tokening. (Roughly, a state S of an object O carries the information that an object O* is in state S* iff it O*'s being in S* *reliably causes* O to be in S.) Informational theorists agree that information alone is not sufficient for the kind of content appropriate to commonsense psychological states such as belief, though they disagree on what additional properties an informational state must have in order to be a representation of the appropriate kind. Information is taken to be insufficient for two reasons. First, there are objects whose states carry information about states of affairs (for example, ringing telephones, tree trunks and speedometers) which they cannot be said to represent in the sense relevant to psychological states. Second, there is the infamous "Disjunction Problem," which reveals that bare informational theories are unable to account for the fact that causal relations hold between mental/neural states and states of affairs they do *not* represent -- i.e., the fact that we can think *false* thoughts.

The main attempts to solve the Disjunction Problem are the *Asymmetric Dependency Theory* (Fodor 1987, 1990a, 1994) and the *Teleological Theory* (Fodor 1990b, Millikan 1984, Papineau 1987, Dretske 1988, 1994). According to the Asymmetric Dependency Theory, the causal relation that determines content is the one without which the others would not hold, but which would itself hold even if the others did not. For example, since we would not (or would not be disposed to) token a mental representation of a horse when confronted with a zebra (say, in non-optimal perceptual conditions) if we did not (or were not disposed to) token a mental representation of a horse when confronted with a horse, but not vice versa, the mental representation tokened in the presence of horses means *horse*, in spite of the fact that there is a causal(informational) relation between it and zebras.

According to the Teleological Theory, the relation that determines content is the one the representation-producing mechanism has the *selected* (by evolution or learning) *function* of subserving. (For example, zebra-caused horse-representations do not mean *zebra*, because the mechanism by which those tokens were produced has the selected function of indicating horses, not zebras. The horse-representation-producing mechanism that responds to zebras is *malfunctioning*.)

Functional theories (Harman 1973, Block 1986) hold that the content of a mental representation is grounded in its (causal, computational, inferential) relations to other mental representations. They differ on whether relata should include all other mental representations or only some of them, and on whether to include external states of affairs. The view that the content of a mental representation is determined by its inferential/computational relations with *all* other representations is *holism*; the view it is determined by relations to only *some* other mental states is *localism* (or *molecularism*). (The view that the content of a mental state depends on *none* of its relations to other mental states is *atomism*.) Functional theories which recognize no content-constitutive external relata have been called *solipsistic* (Harman 1987). Some theorists posit distinct roles for internal and external connections, the former determining semantic properties analogous to sense, the latter determining semantic properties analogous to reference (Sterelny 1989).

Internalism and Externalism

Generally, those who, like informational theorists, think relations to one's (natural or social) environment are at least partially determinative of the content of one's mental representations are *externalists* (e.g., Putnam 1975, Burge 1979, 1986), whereas those who, like some proponents of functional theories, think representational content is determined by intrinsic properties alone, are *internalists* (or *individualists*; cf. Putnam 1975, Fodor 1981b).

This issue is of central importance, since the explanations of cognitive science are causal (computational), and the representational states these explanations quantify over are supposed to be subsumed under psychological generalizations in virtue of their content. If, however (as stressed by Stich), a mental representation's having a particular content is due to factors *extrinsic* to it, it is unclear how its having that content could determine its causal powers, which, arguably, must be intrinsic (Stich 1983; see also Fodor, 1982, 1987, 1994). Some who accept the Putnam and Burge arguments for externalism have argued that internal factors determine a *component* of the content of a mental representation. They say that mental representations have both "narrow" content (determined by intrinsic factors) and "wide" content (determined by narrow content plus extrinsic factors). (The distinction may as well be applied to the sub-personal representations of cognitive science as to those of commonsense psychology. See Von Eckardt 1993: 189.)

Narrow content has been variously construed. Putnam (1975), Fodor (1982: 114; 1994: 39ff), and Block (1986: 627ff), for example, seem to understand it as something like *de dicto* content (i.e.,Fregean *sense*, or perhaps *character* à la Kaplan 1989). On this construal, narrow content is metaphysically context-independent and directly expressible. Fodor (1987) and Block (1986), however, have also characterized narrow content as metaphysically context-independent but radically *inexpressible*. On this construal, narrow content is a kind of proto-content, or content-determinant, and can be specified only indirectly, via specifications of context/wide-content pairings. On both construals, narrow contents are characterized as functions from context to (wide) content. The narrow content of a representation is determined by properties intrinsic to it or its possessor -- its syntactic structure or its intramental computational or inferential role, for example.

Fodor (1994, 1998) has more recently urged that cognitive science might not need narrow mental representations in order to supply naturalistic (causal) explanations of human cognition and action, since the sorts of cases they were introduced to handle, viz., Twin-Earth cases and Frege cases, are either nomologically impossible or dismissible as exceptions to non-strict psychological laws.

Phenomenal and Non-phenomenal Representation

It is a common (though by no means unchallenged -- see below) assumption among realists about mental representations -- at least those of the sort appropriate to commonsense psychological states -- that they come in two basic varieties (cf. Boghossian 1995), with, correspondingly, two different kinds of

representational content. There are those -- for example percepts, sensory experiences and (perhaps) images -- which have phenomenal properties, and there are those -- for example concepts and thoughts -- which do not. Those of the former type maybe said to have *non-phenomenal content*, those of the latter type, *phenomenal content*. On this taxonomy, mental representations may represent in the way expressions of natural languages do, but they may also represent in a way drawings, paintings or photographs do -- i.e., by *resembling* what they represent. (An analogous distinction may also be available for the sub-personal representations proposed by cognitive scientists.)

Disagreement over phenomenal representation concerns the existence and nature of phenomenal properties (Dennett 1988 argues that there simply *are* no such things as qualia as ordinarily conceived), and the role they play (if they exist) in determining the content of sensory, perceptual, and imagistic representations. If a *sensation* is the mere having of a qualitative experience (a *quale*: a pain or tickle, an experience of blue or smoothness), while *perception* is sensory experience *of* something (in the external world), then, if perceptions are constituted in part by sensations -- that is, if they have sensory phenomenal properties -- a crucial question is what, if anything, such properties have to do with their representationalcontent.

Some historical discussions of the representational properties of the mind (e.g., Aristotle 1984, Locke 1978, Hume 1978) assumed that phenonmenal representations -- viz., percepts ("impressions") and images ("ideas") -- are the *only* kinds of mental representations, and that the mind represents the world in virtue of being in states that resemble it. On such a view, mental representations have their content in virtue of their introspectable phenomenal features. Powerful arguments, however, focusing on the lack of generality (Berkeley 1975), ambiguity (Wittgenstein 1953),and non-compositionality (Fodor 1981c) of perceptual and imagistic representations, as well as their unsuitability to function as logical (Frege 1918, Geach 1957) or mathematical (Frege 1953) concepts, and the symmetry of resemblance (Goodman 1976), convinced philosophers that no theory of the mind can get by with *only* imagistic representations. Some contemporary philosophers (Harman 1990, Leeds 1993, Rey 1991, Tye 1995, 2000) have argued that theories of the representational mind can get by with *no* phenomenal properties, since *symbolic* representations can do all the representational work of perception and imagination. (Block 1996 calls such philosophers "Representationists.")

Others (Evans 1982; Peacocke 1983, 1989, 1992; Raffman 1995; Shoemaker 1990) argue that a satisfactory theory of the representational mind *must* acknowledge phenomenal representations. (Block 1996 calls such philosophers "Phenomenists.") They claim that phenomenal properties are (at least partly) responsible for the representational powers of (at least) perceptual experiences (they do not claim that symbolic representation plays *no* role in determining the content of experience). Peacocke (1983, 1992), and Raffman (1995), for example, argue that we are capable of mentally representing perceivable properties of our environment that we do not (or cannot) represent symbolically.

Peacocke 1992 develops the notion of a perceptual "scenario" (an assignment of phenomenal properties to coordinates of a three-dimensional egocentric space), whose content is "correct" (a semantic property) if in the corresponding "scene" (the portion of the external world with the same origin and axes as the scenario) properties are distributed as they are in the scenario. He claims that such scenarios are possible

in the absence of symbolic representations corresponding to the properties represented. (Cf. the distinction in Dretske 1969 between epistemic and non-epistemic perception.)

Still others, including Chalmers 1996, Flanagan 1992, Goldman 1993, Jackendoff 1987, Levine 1993, 1995, McGinn 1992, Searle 1990 and Strawson 1995, claim that purely symbolic (conscious) representational states themselves have a proprietary phenomenology. If this claim is correct, the question of what, if anything, these properties have to do with content rearises for symbolic representation. (A Representationist answer could not be eliminativist with respect to phenomenal content if this claim is correct.)

Imagery

In a series of psychological experiments done in the 1970s (summarized in Kosslyn 1980 and Shepard and Cooper 1982), subjects' response time in tasks involving mental manipulation and examination of presented figures was found to vary in proportion to the spatial properties (size, orientation, etc.) of the figures presented. The question of how these experimental results are best explained has kindled a lively debate on the question of imagery.

Kosslyn claims that the results suggest that the tasks were accomplished via the examination and manipulation of mental representations that themselves have spatial properties -- i.e., *pictorial* representations, or *images*. Others, principally Pylyshyn (1979, 1981a, 1981b), argue that the empirical facts can be explained in terms exclusively of *discursive*, or *propositional* representations and cognitive processes defined over them. (Pylyshyn takes such representations to be sentences in a language of thought.)

The idea that pictorial representations are literally *pictures* in the head is not taken seriously by proponents of the pictorial view of imagery (see, e.g., Kosslyn and Pomerantz 1977). The claim is, rather, that mental images represent in a way that is relevantly *like* the way pictures represent. (Attention has been focused on *visual* imagery -- hence the designation ‘pictorial’; though of course there may imagery in other modalities -- auditory, olfactory, etc. -- as well.)

The distinction between pictorial and discursive representation can be characterized in terms of the distinction between *analog* and *digital* representation (Goodman 1976). This distinction has itself been variously understood (Fodor and Pylyshyn 1981, Goodman 1976, Haugeland 1981, Lewis 1971, McGinn 1989), though a widely accepted construal is that analog representation is continuous (i.e., in virtue of continuously variable properties of the representation), while digital representation is discrete (i.e., in virtue of properties a representation either has or doesn't have) (Dretske 1981). (An analog/digital distinction may also be made with respect to cognitive *processes*. (Block 1983.))

On this understanding of the analog/digital distinction, phenomenal representations, which represent in virtue of properties that may vary continuously (such as being more or less bright, loud, vivid, etc.), would be analog, while non-phenomenal representations, whose properties do not vary continuously (a

thought cannot be more or less about Paris: either it is or it is not) would be digital. It seems clear, however, that commitment to pictorial representation is not *ipso facto* commitment to *phenomenal* representation, since representations may have non-phenomenal properties that vary continuously.

There are, moreover, other ways of understanding pictorial representation that presuppose *neither* phenomenality nor analogicity.

Kosslyn, (1980, 1983), for example, uses the metaphor of a spatial display on a CRT screen to characterize pictorial representation: images are generated on a screen by a computer using information stored (in discursive form) in memory, and the spatial properties of the screen-image (which is composed of illuminated and unilluminated pixels) may correspond to the spatial properties of the imaged object. This isomorphism may be achieved without *literal* spatial representation in the brain. According to Kosslyn, a mental representation is "quasi-pictorial" when every part of the representation corresponds to a part of the object represented, and relative distances between parts of the object represented are preserved among the parts of the representation. Distances between two locations within a representation may be defined by number of intervening positions. (Kosslyn 1982.)

Moreover, intervention need not be *spatial*: it may, for example, be *functional*. That is, the distance between two points on an object might be represented by *computational* distance between the representations of the points -- if, for example, the number of computational steps required to *combine* stored information about the positions of the particular points equals (or is proportional to) the number of spatially intermediate points on the object represented. (Cf. Rey 1981.)

Tye (1991) proposes a view of images on which they are *hybrid* representations, consisting both of pictorial and discursive elements. On Tye's account, images are "(labeled) interpreted symbol-filled arrays." The symbols represent discursively, while their arrangement in arrays has representational significance (the location of each "cell" in the array represents a specific viewer-centered 2-D location on the surface of the imagined object).

Thought and Language

To say that a mental object has semantic properties is, paradigmatically, to say that it may be *about*, or be true or false of, an object or objects, or that it may be true or false *simpliciter*. Suppose I think that ocelots take snuff. I am thinking about ocelots, and if what I think of them (that they take snuff) is true of them, then my thought is true. According to RTM such states are to be explained as relations between agents and mental representations. To think that ocelots take snuff is to token in some way a mental representation whose content is that ocelots take snuff. On this view, the semantic properties of mental states are the semantic properties of the representations they are relations to.

Linguistic acts seem to share such properties with mental states. Suppose I *say* that ocelots take snuff. I am talking about ocelots, and if what I say of them (that they take snuff) is true of them, then my utterance is true. Now, to say that ocelots take snuff is (in part) to utter a sentence that means that ocelots

take snuff. Many philosophers have thought that the semantic properties of linguistic expressions are inherited from the mental states they are conventionally used to express (Grice 1957, Fodor 1978, Schiffer 1988, Searle 1983). On this view, the semantic properties of linguistic expressions are the semantic properties of the representations that are the mental relata of the states they are conventionally used to express.

(Others, however, e.g., Davidson (1975, 1982) have suggested that the kind of thought human beings are capable of is not possible without language, so that the dependency might be reversed, or somehow mutual (see also Sellars 1956). (But see Martin 1987 for a defense of the claim that thought is possible without language. See also Chisholm and Sellars 1958.) Schiffer (1987) subsequently despaired of the success of what he calls "Intention Based Semantics.")

It is also widely held that in addition to having such properties as reference, truth-conditions and truth -- so-called *extensional* properties -- expressions of natural languages also have *intensional* properties, in virtue of expressing properties or propositions -- i.e., in virtue of having *meanings* or *senses,* where two expressions may have the same reference, truth-conditions or truth value, yet express different properties or propositions (Frege 1892). So, if the semantic properties of natural-language expressions are inherited from the thoughts and concepts they express (or vice versa -- or both), then an analogous distinction may be appropriate for mental representations.

This distinction, which is accepted by many philosophers of mind, can be made out in a number of different ways (for example, the *de dicto*/*de re* interpretation of the narrow/wide distinction mentioned above). In general, theories of mental representation that accept a distinction between intrinsic and extrinsic determinants of content are called "two-factor" theories (Field 1978, Loar 1981, McGinn 1982). Such components may or may not be taken to be analogous to the intension and extension of a linguistic expression (Sterelny 1989).

Bibliography

- Almog, J., Perry, J. and Wettstein, H., eds. *Themes from Kaplan*, New York: Oxford University Press (1989).
- Aristotle. *De Anima*, in *The Complete Works of Aristotle: The Revised Oxford Translation*, Oxford: Oxford University Press (1984).
- Ballard, D.H. "Cortical Connections and Parallel Processing: Structure and Function," *The Behavioral and Brain Sciences* 9 (1986): 67-120.
- Ballard, D.H and Hayes, P.J. "Parallel Logical Inference," *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, Rochester, NY. (1984).
- Beaney, M., ed. *The Frege Reader*, Oxford: Blackwell Publishers (1997)
- Berkeley, G. *Principles of Human Knowledge*, in M.R. Ayers, ed., *Berkeley: Philosophical Writings*, London: Dent (1975).
- Block, N. (ed.). *Readings in Philosophy of Psychology, Vol. 2,* Cambridge, Mass.: Harvard University Press (1981).

- Block, N. (ed.) *Imagery*, Cambridge, Mass.: The MIT Press (1982).
- Block, N. "Mental Pictures and Cognitive Science," *Philosophical Review* 93 (1983): 499-542.
- Block, N. "Advertisement for a Semantics for Psychology," in P.A. French, T.E. Uehling and H.K. Wettstein, eds., *MidwestStudies in Philosophy, Vol. X*, Minneapolis: University of Minnesota Press (1986): 615-678.
- Block, N. "Mental Paint and Mental Latex," in E. Villanueva, ed., *Philosophical Issues, 7: Perception* (1996): 19-49.
- Boghossian, P.A. "Content," in J. Kim and E. Sosa, eds., *A Companion to Metaphysics*, Oxford: Blackwell Publishers Ltd. (1995): 94-96.
- Burge, T. "Individualism and the Mental," in P.A. French, T.E. Uehling and H.K. Wettstein, eds., *Midwest Studies in Philosophy, Vol. IV*, Minneapolis: University of Minnesota Press (1986): 73-121.
- Chalmers, D. *The Conscious Mind*, New York: Oxford University Press (1996).
- Chalmers, D. "Connectionism and Compositionality: Why Fodor and Pylyshyn Were Wrong," *Philosophical Psychology* 6 (1993): 305-319.
- Chisholm, R. and Sellars, W. "The Chisholm-Sellars Correspondence on Intentionality," in H. Feigl, M. Scriven and G. Maxwell, eds., *Minnesota Studies in the Philosophy of Science, Vol. II*, Minneapolis : University of Minnesota Press (1958): 529-539.
- Chomsky, N. *Aspects of the Theory of Syntax*, Cambridge, Mass.: The MIT Press (1965).
- Churchland, P.M. "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* 78 (1981): 67-90.
- Churchland, P.M. "On the Nature of Theories: A Neurocomputational Perspective," in W. Savage, ed., *Scientific Theories: Minnesota Studies in the Philosophy of Science, Vol. 14*, Minneapolis: University of Minnesota Press (1989): 59-101.
- Davidson, D. "Radical Interpretation," *Dialectica* 27 (1973): 313-328.
- Davidson, D. "Belief and the Basis of Meaning," *Synthese* 27 (1974): 309-323.
- Davidson, D. "Thought and Talk," in S. Guttenplan, ed., *Mind and Language*, Oxford: Clarendon Press (1975): 7-23.
- Davidson, D. "Rational Animals," *Dialectica* 4 (1982): 317-327.
- Dennett, D. *Content and Consciousness*, London: Routledge and Kegan Paul (1969).
- Dennett, D. "The Nature of Images and the Introspective Trap," pages 132-141 of Dennett 1969, reprinted in Block 1981 (1981): 128-134.
- Dennett, D. *The Intentional Stance*, Cambridge, Mass.: The MIT Press (1987).
- Dennett, D. "True Believers: The Intentional Strategy and Why it Works," in Dennett 1987 (1987a): 13-35.
- Dennett, D. "Reflections: Real Patterns, Deeper Facts, and Empty Questions," in Dennett 1987 (1987b): 37-42.
- Dennett, D. "Quining Qualia," in A.J. Marcel and E. Bisiach, eds., *Consciousness in Contemporary Science*, Oxford: Clarendon Press (1988): 42-77.
- Dennett, D. "Real Patterns," *The Journal of Philosophy* LXXXVII (1991): 27-51.
- Devitt, M. *Coming to Our Senses: A Naturalistic Program for Semantic Localism*, Cambridge: Cambridge University Press (1996).
- Dretske, F. *Seeing and Knowing*, Chicago: The University of Chicago Press (1969).

- Dretske, F. *Knowledge and the Flow of Information*, Cambridge, Mass.: The MIT Press (1981).
- Dretske, F. *Explaining Behavior: Reasons in a World of Causes*, Cambridge, Mass.: The MIT Press (1988).
- Dretske, F. *Naturalizing the Mind*, Cambridge, Mass.: The MIT Press (1994).
- Evans, G. *The Varieties of Reference*, Oxford: Oxford University Press (1982).
- Field, H. "Mental representation," *Erkenntnis* 13 (1978): 9-61.
- Flanagan, O. *Consciousness Reconsidered*, Cambridge, Mass.: The MIT Press (1992).
- Fodor, J.A. *The Language of Thought*, Cambridge, Mass.: Harvard University Press (1975).
- Fodor, J.A. "Propositional Attitudes," *The Monist* 61 (1978): 501-523.
- Fodor, J.A. *Representations*, Cambridge, Mass.: The MIT Press (1981).
- Fodor, J.A. "Introduction," in Fodor 1981 (1981a): 1-31.
- Fodor, J.A. "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology," in Fodor 1981 (1981b): 225-253.
- Fodor, J.A. "The Present Status of the Innateness Controversy," in Fodor 1981 (1981c): 257-316.
- Fodor, J.A. "Cognitive Science and the Twin-Earth Problem," *Notre Dame Journal of Formal Logic* 23 (1982): 98-118.
- Fodor, J.A. *Psychosemantics*, Cambridge, Mass.: The MIT Press (1987).
- Fodor, J.A. *A Theory of Content and Other Essays*, Cambridge, Mass.: The MIT Press (1990a).
- Fodor, J.A. "Psychosemantics or: Where Do Truth Conditions Come From?" in W.G. Lycan, ed., *Mind and Cognition: A Reader*, Oxford: Blackwell Publishers (1990), (1990b): 312-337.
- Fodor, J.A. *The Elm and the Expert*, Cambridge, Mass.: The MIT Press (1994).
- Fodor, J.A. and Pylyshyn, Z. "How Direct is Visual Perception?: Some Reflections on Gibson's 'Ecological Approach'," *Cognition* 9 (1981): 207-246.
- Fodor, J.A. and Pylyshyn, Z. "Connectionism and Cognitive Architecture: A Critical Analysis," *Cognition* 28 (1988): 3-71.
- Frege, G. *The Foundations of Arithmetic*, trans. J.L. Austin, New York: Philosophical Library (1953).
- Frege, G. "On *Sinn* and *Bedeutung*" (1892) in Beany 1997: 151-171.
- Frege, G. "Thought" (1918) in Beany 1997: 325-345.
- Geach, P. *Mental Acts: Their Content and Their Objects*, London: Routledge & Kegan Paul (1957) .
- Goldman, A. "The Psychology of Folk Psychology," *Behavioral and Brian Sciences* 16 (1993): 15-28.
- Goodman, N. *Languages of Art* (2nd ed.), Indianapolis: Hackett (1976).
- Grice, H.P. "Meaning," *Philosophical Review*, 66 (1957): 377-388; reprinted in *Studies in the Way of Words*, Cambridge, Mass.: Harvard University Press (1989): 213-223.
- Harman, G. *Thought*, Princeton: Princeton University Press (1973).
- Harman, G. "(Non-Solipsistic) Conceptual Role Semantics," in E. Lepore, ed., *New Directions in Semantics*, London: Academic Press (1987): 55-81.
- Harman, G. "The Intrinsic Quality of Experience," in J. Tomberlin, ed., *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, Atascadero: Ridgeview Publishing Company (1990): 31-52.
- Haugeland, J. "Analog and analog," *Philosophical Topics* 12 (1981): 213-226.

- Heil, J. "Being Indiscrete," in J. Greenwood, ed., *The Future of Folk Psychology*, Cambridge: Cambridge University Press (1991): 120-134.
- Horgan, T. and Tienson, J. *Connectionism and the Philosophy of Psychology*, Cambridge, Mass: The MIT Press (1996).
- Horst, S. *Symbols, Computation, and Intentionality*, Berkeley: University of California Press (1996).
- Hume, D. *A Treatise of Human Nature*, L.A. Selby-Bigge, ed., revised P.H. Nidditch, Oxford: Oxford University Press (1978).
- Jackendoff, R. *Computation and Cognition*, Cambridge, Mass.: The MIT Press (1987).
- Johnson-Laird, P.N. *Mental Models*, Cambridge, Mass.: Harvard University Press (1983).
- Johnson-Laird, P.N. and Wason, P.C. *Thinking: Readings in Cognitive Science*, Cambridge University Press (1977).
- Kaplan, D. "Demonstratives," in Almog, Perry and Wettstein 1989 (1989): 481-614.
- Kosslyn, S.M. *Image and Mind*, Cambridge, Mass.: Harvard University Press (1980).
- Kosslyn, S.M. "The Medium and the Message in Mental Imagery," in Block 1982 (1982): 207-246.
- Kosslyn, S. *Ghosts in the Mind's Machine*, New York: W.W. Norton & Co. (1983).
- Kosslyn, S.M. and Pomerantz, J.R. "Imagery, Propositions, and the Form of Internal Representations," *Cognitive Psychology* 9 (1977): 52-76.
- Leeds, S. "Qualia, Awareness, Sellars," *Noûs* XXVII (1993): 303-329.
- Lerdahl, F. and Jackendoff, R. *A Generative Theory of Tonal Music*, Cambridge, Mass.: The MIT Press (1983).
- Levine, J. "On Leaving Out What It's Like," in M. Davies and G. Humphreys, eds., *Consciousness*, Oxford: Blackwell Publishers (1993): 121-136.
- Levine, J. "On What It Is Like to Grasp a Concept," in E. Villanueva, ed., *Philosophical Issues 6: Content*, Atascadero: Ridgeview Publishing Company (1995): 38-43.
- Lewis, D. "Analog and Digital," *Noûs* 5 (1971): 321-328.
- Lewis, D. "Radical Interpretation," *Synthese* 23 (1974): 331-344. (Reprinted, with Postscript, in Lewis 1983: 108-121.)
- Lewis, D. *Philosophical Papers, Vol. I*, New York: Oxford University Press (1983).
- Loar, B. *Mind and Meaning*, Cambridge: Cambridge University Press (1981).
- Locke, J. *An Essay Concerning Human Understanding*, P.H. Nidditch, ed., Oxford: Oxford University Press (1978).
- MacDonald, C. and MacDonald, G. *Connectionism: Debates on Psychological Explanation*, Oxford: Blackwell Publishers (1995).
- Marr, D. *Vision*, New York: W.H. Freeman and Company (1982).
- Martin, C.B. "Proto-Language," *Australasian Journal of Philosophy* 65 (1987): 277-289.
- McCulloch, W.S. and Pitts, W. "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* 5 (1943): 115-33.
- McGinn, C. "The Structure of Content," in A. Woodfield, ed., *Thought and Content,* Oxford: Oxford University Press (1982): 207-258.
- McGinn, C. *Mental Content*, Oxford: Blackwell Publishers (1989).
- McGinn, C. "Content and Consciousness," in C. McGinn, *The Problem of Consciousness*, Oxford:

Blackwell Publishers (1992): 23-43.

- Millikan, R. *Language, Thought and other Biological Categories*, Cambridge, Mass.: The MIT Press (1984).
- Newell, A. and Simon, H.A. *Human Problem Solving*, New York: Prentice-Hall (1972).
- Newell, A. and Simon, H.A. "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the Association for Computing Machinery* 19 (1976): 113-126.
- Osherson, D.N., Kosslyn, S.M. and Hollerbach, J.M. *Visual Cognition and Action: An Invitation to Cognitive Science, Vol. 2*, Cambridge, Mass.: The MIT Press (1990).
- Papineau, D. *Reality and Representation*, Oxford: Blackwell Publishers (1987).
- Peacocke, C. *Sense and Content*, Oxford: Clarendon Press (1983).
- Peacocke, C. "Perceptual Content," in Almog, Perry and Wettstein 1989 (1989): 297-329.
- Peacocke, C. "Scenarios, Concepts and Perception," in T. Crane, ed., *The Contents of Experience*, Cambridge: Cambridge University Press (1992): 105-35.
- Pinker, S. *Learnability and Cognition*, Cambridge, Mass.: The MIT Press (1989).
- Port, R. and Van Gelder, T. *Mind as Motion: Explorations in the Dynamics of Cognition*, Cambridge, Mass.: The MIT Press (1995).
- Putnam, H. "The Meaning of 'Meaning'," in *Philosophical Papers, Vol. 2*, Cambridge: Cambridge University Press (1975): 215-71.
- Pylyshyn, Z. "The Rate of 'Mental Rotation' of Images: A Test of a Holistic Analogue Hypothesis," *Memory and Cognition*, 7 (1979): 19-28.
- Pylyshyn, Z. "Imagery and Artificial Intelligence," in Block 1981 (1981a): 170-194.
- Pylyshyn, Z. "The Imagery Debate: Analog Media versus Tacit Knowledge,"*Psychological Review* 88 (1981b): 16-45.
- Pylyshyn, Z. *Computation and Cognition*, Cambridge, Mass.: The MIT Press (1984).
- Raffman, D. "The Persistence of Phenomenology," in T. Metzinger, ed., *Conscious Experience*, Paderborn: Schöningh/Imprint Academic (1995): 293-308.
- Ramsey, W., Stich, S. and Garon, J. "Connectionism, Eliminativism and the Future of Folk Psychology," *Philosophical Perspectives* 4 (1990): 499-533.
- Rey, G. "Introduction: What Are Mental Images?" in Block(1981): 117-127.
- Rey, G. "Sensations in a Language of Thought," in E. Villaneuva, ed., *Philosophical Issues 1: Consciousness*, Atascadero: Ridgeview Publishing Company (1991): 73-112.
- Rumelhart, D.E. "The Architecture of the Mind: A Connectionist Approach," in M.I. Posner, ed., *Foundations of Cognitive Science*, Cambridge, Mass.: The MIT Press (1989): 133-159.
- Rumelhart, D.E. and McCelland, J.L. *Parallel Distributed Processing, Vol. I*, Cambridge, Mass.: The MIT Press (1986).
- Schiffer, S. 1987. *Remnants of Meaning*, Cambridge, Mass.: The MIT Press (1987).
- Schiffer, S. "Introduction to the Paperback Edition," in *Meaning*, Oxford: Clarendon Press (1988): xi-xxix.
- Searle, J.R. *Intentionality*, Cambridge: Cambridge University Press (1983).
- Searle, J.R. *The Rediscovery of the Mind*, Cambridge, Mass.: The MIT Press (1990).
- Sellars, W. "Empiricism and the Philosophy of Mind," in K. Gunderson, ed., *Minnesota Studies in the Philosophy of Science, Vol. I*, Minneapolis: University of Minnesota Press (1956): 253-329.
- Shepard, R.N. and Cooper, L. *Mental Images and their Transformations*, Cambridge, Mass.: The

MIT Press (1982).

- Shoemaker, S. "Qualities and Qualia: What's in the Mind?" *Philosophy and Phenomenological Research* 50 (1990): 109-31.
- Smolensky, P. "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences*, 11 (1988): 1-74.
- Smolensky, P. "Connectionist Modeling: Neural Computation/Mental Connections," in L. Nadel, L.A. Cooper, P. Culicover and R.M. Harnish, eds., *Neural Connections, Mental Computation*: The MIT Press (1989): 49-67.
- Smolensky, P. "Connectionism and the Language of Thought," in B. Loewer and G. Rey, eds., *Meaning in Mind: Fodor and His Critics*, Oxford: Basil Blackwell Ltd (1991): 201-227.
- Sterelny, K. "Fodor's Nativism," *Philosophical Studies*55 (1989): 119-141.
- Stich, S. 1978. "Autonomous Psychology and the Belief-Desire Thesis," *The Monist* 61 (1978): 573-591.
- Stich, S. *From Folk Psychology to Cognitive Science*, Cambridge, Mass.: The MIT Press (1983).
- Strawson, G. *Mental Reality*, Cambridge, Mass.: The MIT Press (1994).
- Tye, M. *The Imagery Debate*, Cambridge, Mass.: The MIT Press (1991).
- Tye, M. *Ten Problems of Consciousness*, Cambridge, Mass.: The MIT Press (1995).
- Tye, M. *Consciousness, Color, and Content*, Cambridge, Mass.: The MIT Press (2000).
- Van Gelder, T. "What Might Cognition Be, if not Computation?", *Journal of Philosophy* XCI (1995): 345-381.
- Von Eckardt, B. 1993. *What Is Cognitive Science?*, Cambridge, Mass.: The MIT Press (1993).
- Von Eckardt, B. "Connectionism and the Propositional Attitudes," in C. Emeling and D.M. Johnson, eds., *The Mind as a Scientific Object: Between Brain and Culture* (forthcoming).
- Wittgenstein, L. *Philosophical Investigations*, trans. G.E.M. Anscombe, Oxford: Blackwell Publishers (1953).

Other Internet Resources

- A Field Guide to the Philosophy of Mind
- Dictionary of Philosophy of Mind
- Mind/Brain Resources

Related Entries

cognitive science | computing, modern history of | connectionism | consciousness: and intentionality | consciousness: representational theories of | folk psychology: as a theory | folk psychology: as mental simulation | information | language of thought hypothesis | mental content: causal theories of | mental imagery | neuroscience, philosophy of | perception | qualia | reference

Acknowledgements

Thanks to Brad Armour-Garb, Jim Garson, John Heil, Jeff Poland, Bill Robinson, Galen Strawson, Adam Vinueza and Barbara Von Eckardt for comments on earlier drafts of this entry.

David Pitt
CUNY/Brooklyn College
dalanpitt@yahoo.com

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 30, 2000
Content last modified: August 15, 2000

Cognitive Science

Cognitive science is the interdisciplinary study of mind and intelligence, embracing philosophy, psychology, artificial intelligence, neuroscience, linguistics, and anthropology. Its intellectual origins are in the mid-1950s when researchers in several fields began to develop theories of mind based on complex representations and computational procedures. Its organizational origins are in the mid-1970s when the Cognitive Science Society was formed and the journal Cognitive Science began. Since then, more than sixty universities in North America and Europe have established cognitive science programs and many others have instituted courses in cognitive science.

History

Attempts to understand the mind and its operation go back at least to the Ancient Greeks, when philosophers such as Plato and Aristotle tried to explain the nature of human knowledge. The study of mind remained the province of philosophy until the nineteenth century, when experimental psychology developed. Wilhelm Wundt and his students initiated laboratory methods for studying mental operations more systematically. Within a few decades, however, experimental psychology became dominated by behaviorism, a view that virtually denied the existence of mind. According to behaviorists such as J. B. Watson, psychology should restrict itself to examining the relation between observable stimuli and observable behavioral responses. Talk of consciousness and mental representations was banished from respectable scientific discussion. Especially in North America, behaviorism dominated the psychological scene through the 1950s. Around 1956, the intellectual landscape began to change dramatically. George Miller summarized numerous studies which showed that the capacity of human thinking is limited, with short-term memory, for example, limited to around seven items. He proposed that memory limitations can be overcome by recoding information into chunks, mental representations that require mental procedures for encoding and decoding the information. At this time, primitive computers had been

around for only a few years, but pioneers such as John McCarthy, Marvin Minsky, Allen Newell, and Herbert Simon were founding the field of artificial intelligence. In addition, Noam Chomsky rejected behaviorist assumptions about language as a learned habit and proposed instead to explain language comprehension in terms of mental grammars consisting of rules. The six thinkers mentioned in this paragraph can be viewed as the founders of cognitive science.

Representation and Computation

The central hypothesis of cognitive science is that thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures. While there is much disagreement about the nature of the representations and computations that constitute thinking, the central hypothesis is general enough to encompass the current range of thinking in cognitive science, including connectionist theories which model thinking using artificial neural networks.

Most work in cognitive science assumes that the mind has mental representations analogous to computer data structures, and computational procedures similar to computational algorithms. Cognitive theorists have proposed that the mind contains such mental representations as logical propositions, rules, concepts, images, and analogies, and that it uses mental procedures such as deduction, search, matching, rotating, and retrieval. The dominant mind-computer analogy in cognitive science has taken on a novel twist from the use of another analog, the brain. Connectionists have proposed novel ideas about representation and computation that use neurons and their connections as inspirations for data structures, and neuron firing and spreading activation as inspirations for algorithms. Cognitive science then works with a complex 3-way analogy among the mind, the brain, and computers. Mind, brain, and computation can each be used to suggest new ideas about the others. There is no single computational model of mind, since different kinds of computers and programming approaches suggest different ways in which the mind might work. The computers that most of us work with today are serial processors, performing one instruction at a time, but the brain and some recently developed computers are parallel processors, capable of doing many operations at once.

Philosophical Relevance

Philosophy, in particular philosophy of mind, is part of cognitive science. But the interdisciplinary field of cognitive science is relevant to philosophy in several ways. First, the psychological, computational, and other results of cognitive science investigations have important potential applications to traditional philosophical problems in epistemology, metaphysics, and ethics. Second, cognitive science can serve as an object of philosophical critique, particularly concerning the central assumption that thinking is representational and computational. Third and more constructively, cognitive science can taken as an object of investigation in the philosophy of science, generating reflections on the methodology and presuppositions of the enterprise.

1. Philosophical Applications

Much philosophical research today is naturalistic, treating philosophical investigations as continuous with empirical work in fields such as psychology. From a naturalistic perspective, philosophy of mind is closely allied with theoretical and experimental work in cognitive science. Metaphysical conclusions about the nature of mind are to be reached, not by a priori speculation, but by informed reflection on scientific developments in fields such as computer science and neuroscience. Similarly, epistemology is not a stand-alone conceptual exercise, but depends on and benefits from scientific findings concerning mental structures and learning procedures. Even ethics can benefit by using greater understanding of the psychology of moral thinking to bear on ethical questions such as the nature of deliberations concerning right and wrong. Goldman (1993) provides a concise review of applications of cognitive science to epistemology, philosophy of science, philosophy of mind, metaphysics, and ethics.

2. Critique of Cognitive Science

The claim that human minds work by representation and computation is an empirical conjecture and might be wrong. Although the computational-representational approach to cognitive science has been successful in explaining many aspects of human problem solving, learning, and language use, some philosophical critics such as Hubert Dreyfus and John Searle have claimed that this approach is fundamentally mistaken. Critics of cognitive science have offered such challenges as:

- 1. The emotion challenge: Cognitive science neglects the important role of emotions in human thinking.
- 2. The consciousness challenge: Cognitive science ignores the importance of consciousness in human thinking.
- 3. The world challenge: Cognitive science disregards the significant role of physical environments in human thinking.
- 4. The social challenge: Human thought is inherently social in ways that cognitive science ignores.
- 5. The dynamical systems challenge: The mind is a dynamical system, not a computational system.
- 6. The mathematics challenge: Mathematical results show that human thinking cannot be computational in the standard sense, so the brain must operate differently, perhaps as a quantum computer.

Thagard (1996) argues that all these challenges can best be met by expanding and supplementing the computational-representational approach, not by abandoning it.

3. Philosophy of Cognitive Science

Cognitive science raises many interesting methodological questions that are worthy of investigation by

philosophers of science. What is the nature of representation? What role do computational models play in the development of cognitive theories? What is the relation among apparently competing accounts of mind involving symbolic processing, neural networks, and dynamical systems? Von Eckardt (1993) provides a good discussion of some of the philosophical issues that arise in cognitive science.

Bibliography

- Goldman, A. (1993). *Philosophical Applications of Cognitive Science.* Boulder: Westview Press.
- Johnson-Laird, P., (1988). *The Computer and the Mind: An Introduction to Cognitive Science.* Cambridge, MA: Harvard University Press.
- Stillings, N., et al., (1995). *Cognitive Science.* Second edition. Cambridge, MA: MIT Press.
- Thagard, P., (1996). *Mind: Introduction to Cognitive Science*, Cambridge, MA: MIT Press.
- von Eckardt, B. (1993). *What is Cognitive Science?* Cambridge, MA: MIT Press.

Other Internet Resources

- Computational Epistemology Lab at the University of Waterloo
- Cognitive and Psychological Sciences at Stanford University
- Artificial intelligence
- Yahoo Cognitive Science Index

Related Entries

mind: philosophy of

Paul Thagard
pthagard@watarts.uwaterloo.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 23, 1996
Content last modified: September 23, 1996

Consciousness and Intentionality

To say one has an experience that is conscious (in the phenomenal sense) is to say that one is in a state of its *seeming* to one some way. In another formulation, to say experience is conscious is to say that there is *something it's like* for one to have it. Feeling pain and sensing colors are common illustrations of phenomenally conscious states. Consciousness has also been taken to consist in the monitoring of one's own states of mind (e.g., by forming thoughts about them, or by somehow "sensing" them), or else in the accessability of information to one's capacities for rational control or self-report. Intentionality has to do with the directedness or aboutness of mental states -- the fact that, for example, one's thinking is *of* or *about* something. Intentionality includes, and is sometimes taken to be equivalent to, what is called ‘mental representation.’

It can seem that consciousness and intentionality pervade mental life -- perhaps one or both somehow constitute what it is to have a mind. But achieving an articulate general understanding of either consciousness or intentionality presents an enormous challenge, part of which lies in figuring out how the two are related. Is one in some sense derived from or dependent on the other? Or are they perhaps quite independent and separate aspects of mind?

Sections (1) and (2) offer introductory accounts of what is meant by ‘consciousness’ and ‘intentionality,’ with sensitivity to the difficulties raised by their varying interpretation. Then, influential perspectives on intentionality that have emerged in both phenomenological (Section 3) and analytic (Section 4) philosophical traditions are sketched, so as to highlight basic issues about the relationship of consciousness and intentionality, and provide some of the background against which they have been understood. Sections (5) through (8) survey some contemporary views about consciousness, considering their implications for the connection between consciousness and intentionality. Section (9) distinguishes four broad options for understanding their relationship, and closes with some observations about the philosophical consequences of choosing among them.

---

1. Consciousness: Different Senses (or Kinds)?

On one understanding frequent among philosophers, consciousness is a certain feature shared by sense-experience and imagery, perhaps belonging also to a broad range of other mental phenomena (e.g., episodic thought, memory, and emotion). It is the feature that consists in its *seeming* some way to one to have experiences. To put it another way: conscious states are states of its seeming somehow to a subject.

For example, it seems to you some way to see red, and seems to you (some other way) to hear a crash, to visualize a triangle, and to suffer pain. The sense of ‘seems’ relevant here may be brought out by noting that, in the last example, we might just as well speak of the way it *feels* to be in pain. And -- some will want to say -- in the same sense, it seems to you some way to think through the answer to a math problem, or to recall where you parked the car, or to feel anger, shame, or elation. (Note however, that it is not simply to be assumed that saying it seems some way to you to *have* an experience is equivalent to saying that the experience *itself* seems or appears some way to you -- that it is an object of appearance. The point is just that the way something sounds to you, the way something looks to you, etc., all constitute ‘ways of seeming.’) States that are conscious in this sense are said to have some *phenomenal character* or other -- their phenomenal character being the specific way it seems to one to have a given experience. Sometimes this is called the ‘qualitative’ or ‘subjective’ character of experience.

Another oft-used means for trying to get at the relevant notion of consciousness, preferable to some, is to say that there is, in a certain sense, always ‘something it is like’ to be in a given conscious state -- something it's like *for* one who is in that state. Relating the two locutions we might say: there is something it is like for you to see red, to feel pain, etc., and the way it seems to you to have one of these experiences *is* what it is like for you to have it. The phenomenal character of an experience then, is what someone would inquire about by asking, e.g., ‘What is it like to experience orgasm?’ -- and it is what we speak of when we say that we know *what that is like*, even if we cannot convey this to one who *doesn't* know. And, if we want to speak of *persons*, or other creatures (as distinct from their *states*) being conscious, we will say that they are conscious just if there is something it is like for them to be the creatures they are -- for example, something it is like to be a bat.

The examples of conscious states given comprise a various lot. But some sense of their putative unity as instances of consciousness might be gained by contrasting them with what we are inclined to exclude, or can at least conceive of excluding, from their company. Much of what goes on we would ordinarily

believe is not (or at any rate, we may well suppose is not) conscious in the sense at issue. The leaf's fall from a tree branch, we may suppose, is not a conscious state of the leaf -- a state of its seeming somehow to the leaf. Nor, for that matter, is a *person's* fall off a branch a conscious state of that person -- rather, it is the *feeling* of falling that is conscious, if anything is. *Dreaming* of falling would also be a conscious experience in this sense. But, while we can in some way be said to sense the position of our limbs even while dreamlessly asleep, we may still suppose that this proprioception (though perhaps in some sense a mental or cognitive affair) is not conscious -- we may suppose that it does not then seem (or feel) any way to us sleepers to sense our limbs, as ordinarily it does when we are awake.

The ‘way of seeming’ or ‘what it's like’ conception of consciousness I have just invoked is sometimes marked by the term ‘phenomenal consciousness.’ But this qualifier ‘phenomenal’ suggests that there are other kinds of consciousness (or perhaps, other *senses* of ‘consciousness’). Indeed there are, at least, other *ways of introducing* notions of consciousness. And these may appear to pick out features or senses altogether distinct from that just presented. For example, it is said that some (but not all) that goes on in the mind is ‘accessible to consciousness.’ Of course this by itself does not so much specify a sense of ‘conscious’ as put one in use. (One will want to ask: And just what is this ‘consciousness’ that has ‘access’ to some mental goings-on but not others, and what could ‘access’ mean here, anyway?) However, some have evidently thought that, rather than speak of consciousness as what *has* access, we should understand consciousness as itself a certain kind of susceptibility to access. For example, Daniel Dennett (1969) once theorized that one's conscious states are just those whose contents are available to one's direct verbal report -- or, at least, to the ‘speech center’ responsible for generating such reports. And Ned Block (1995) has proposed that, on one understanding of ‘conscious,’ (to be found at work in many ‘cognitive’ theories of consciousness) a conscious state is just a ‘representation poised for free use in reasoning and other direct ‘rational’ control of action and speech.’ Block labels consciousness in this sense ‘access consciousness.’

Block would insist that we should distinguish phenomenal consciousness from ‘access consciousness,’ and he argues that a mental representation's being poised for use in reasoning and rational control of action is neither a necessary nor a sufficient condition for the state's being *phenomenally* conscious. Similarly he distinguishes phenomenal consciousness from what he calls ‘reflexive consciousness’ -- where this has to do with one's capacity to represent one's mind's to oneself -- to have, for example, thoughts about one's own thoughts, feelings, or desires. Such a conception of consciousness finds some support in a tendency to say that conscious states of mind are those one is ‘conscious of’ or ‘aware of’ being in, and to interpret this ‘of’ to indicate some kind of reflexivity is involved -- wherein one represents one's own mental representations (Rosenthal 1986, 1991, 1993). On one prominent variant of this conception, consciousness is taken to be a kind of scanning or perceiving of one's own psychological states or processes -- an ‘inner sense’ (Armstrong 1968, Lycan 1996).

Block's threefold division of phenomenal, access, and reflexive consciousness need not be taken to reflect clear and coherent distinctions already contained in our pre-theoretical use of the term ‘conscious.’ Block himself seems to think that (on the contrary) our initial, ordinary use of ‘conscious’ is too confused even to count as ambiguous. Thus in articulating an interpretation, or set of interpretations, of the term adequate to frame theoretical issues, we cannot simply describe how it is currently employed -- we must

assign it a more definite and coherent meaning than extant in common usage (Block 1995).

Whether or not this is correct, getting a firm footing here is not easy, and a number of theorists of consciousness would balk at proceeding on the basis of Block's proposed threefold distinction. Sometimes the difficulty may be merely terminological. John Searle, for example, would recognize phenomenal consciousness, but deny Block's other two candidates are proper senses of ‘conscious’ at all (Searle 1995). The reality of some sort of access and reflexivity is apparently not at issue -- just whether either captures a senseof ‘conscious’ (perhaps confusedly) woven into our use of the term. However, in contrast to both Block and Searle, there are also those who raise doubts that there is a properly *phenomenal* sense we can apply, distinct from both of the other two, for us to pick out with *any* term (see Dennett 1988, 1991, Lycan 1996, and Rey 1997). This is not just a dispute about words, but about what there is for us to talk about with them.

The substantive issues here are very much bound up with differences over the proper way to conceive of the relationship between consciousness and intentionality. If there are distinct senses in which states of mind could correctly said to be ‘conscious’ (answering perhaps to something like Block's three-fold distinction), then there will be distinct questions we can pose about the relation between consciousness and intentionality. But if one of Block's alleged senses is somehow fatally confused, or if he is wrong to distinguish itfrom the others, or if it is the sense of *no* term we can with warrant apply to ourselves or our states, then there will be no separate question in which it figures we should try to answer. Thus, trying to work out a reasoned view about what we are (or should be) talking about when we talk about consciousness is an unavoidable and non-trivial part of trying to understand the relation between consciousness and intentionality.

To clarify further the disputes about consciousness and their links to questions about its relation to intentionality, we need to get an initial grasp of the relevant way the terms ‘intentionality’ and ‘intentional’ are used in philosophy of mind.

2. Intentionality: Directedness; Conditions of Satisfaction; Content

The previous section gives some indication of why it is difficult to get a theory of consciousness started. While the term ‘conscious’ is not esoteric, its use is not easily characterized or rendered consistent in a manner providing some uncontentious framework for theoretical discussion. Where the term ‘intentional’ is concerned, we also face initially confusing and contentious usage. But here the difficulty lies partly in the fact that the relevant use of cognate terms is simply *not* that found in common speech (as when we speak of doing something ‘intentionally’). Though ‘intentionality,’ in the sense here at issue, does seem to attach to some real and fundamental (maybe even defining) aspect of mental phenomena, the relevant use of the term is tangled up with some rather involved philosophical history.

One way of explaining what is meant by ‘intentionality’ in the (more obscure) philosophical sense is this:

it is that aspect of mental states or events that consists in their being *of* or *about* things (as pertains to the questions, ‘What are you thinking of?’ and ‘What are you thinking about?’). Intentionality is the *aboutness* or *directedness* of mind (or states of mind) to things, objects, states of affairs, events. So if you are thinking about San Francisco, or about the increased cost of living there, or about your meeting someone there at Union Square -- your mind, your thinking, is directed toward San Francisco, or the increased cost of living, or the meeting in Union Square. To think at all is to think of or about something in this sense. This ‘directedness’ conception of intentionality plays a prominent role in the influential philosophical writings of Franz Brentano and those whose views developed in response to his (to be discussed further in Section 3).

But what kind of ‘aboutness’ or ‘of-ness’ or ‘directedness’ is this, and to what sorts of things does it apply? How do the relevant ‘intentionality-marking’ senses of these words (‘about,’ ‘of,’ ‘directed’) differ from: the sense in which the cat is wandering ‘about’ the room; the sense in which someone is a person ‘of’ high integrity; the sense in which the river's course is ‘directed’ towards the fields?

It has been said that the peculiarity of this kind of directedness/aboutness/of-ness lies in its capacity to relate thought or experience to objects that (unlike San Francisco) do not exist. One can think about a meeting that has not, or never will occur; one can think of Shangri La, or El Dorado, or the New Jerusalem; one may think of their shining streets, of their total lack of poverty, or of their citizens' peculiar garb. Thoughts, unlike roads, can lead to a city that is not there.

But to talk in this way only invites new perplexities. Is this to say (with apparent incoherence) that there *are* cities that *do not exist*? And what does it mean to say that, when a state of mind is in fact ‘directed toward’ something that does exist, that state nevertheless *could* be directed toward something that does not exist? It can well seem to be something very fundamental to the nature of mind that our thoughts, or states of mind more generally, can be of or about things or ‘point beyond themselves.’ But a coherent and satisfactory theoretical grasp of this phenomenon of ‘mental pointing’ in all its generality is difficult to achieve.

Another way of trying to get a grip on the topic asks us to note that the potential for a mental directedness towards the non-existent is evidently closely associated with the mind's potential for falsehood, error, inaccuracy, illusion, hallucination, and dissatisfaction. What makes it possible to believe (or even just suppose) something about Shangri La is that one can falsely believe (or suppose) that something exists. In the case of perception, what makes it possible to seem to see or hear what is not there is that one's experience may in various ways be inaccurate, nonveridical, subject to illusion, or hallucinatory. And, what makes it possible for one's desires and intentions to be directed toward what does not and never will exist is that one's desires and intentions can be unfulfilled or unsatisfied. This suggests another strategy for getting a theoretical hold on intentionality, employing a notion of *satisfaction*, stretched to encompass susceptibility to each of these modes of assessment, each of these ways in which something can either go right, or go wrong (true/false, veridical/nonveridical, fulfilled/unfulfilled), and speak of intentionality in terms of having ‘conditions of satisfaction.’ On John Searle's (1983) conception, intentional states are those having conditions of satisfaction. What are conditions of satisfaction? In the case of belief, these are the conditions under which the belief is true; in the case of perception, they are the conditions under

which sense-experience is veridical; in the case of intention, the conditions under which an intention is fulfilled or carried out.

However, while the conditions of satisfaction approach to the notion of intentionality may furnish an alternative to introducing this notion by talking of ‘directedness to objects,’ it is not clear that it can get us around the problems posed by the ‘directedness’ talk. For instance, what are we to say where thoughts are expressed using names of nonexistent deities or fictional characters? Will we do away with a troublesome directedness to the nonexistent by saying that the thoughts that Zeus is Poseidon's brother, and that Hamlet is a prince, are just false? This is problematic. Moreover, how will we state the conditions of satisfaction of such thoughts? Will this not also involve an apparent *reference* to the nonexistent? (For discussion of these issues, see, for example, Thomasson 1999.)

A third important way of conceiving of intentionality, one particularly central to the analytic tradition derived from the study of Frege and Russell (see Section 4), asks us to focus on the notion of mental (or intentional) *content*. Often, it is assumed: to have intentionality is to have content. And frequently mental content is otherwise described as *representational* or *informational* content -- and ‘intentionality’ (at least, as this applies to the mind) is seen as just another word for what is called ‘mental representation,’ or a certain way of bearing or carrying information.

But what is meant by ‘content’ here? As a start we may note: the content of *thought*, in this sense, is what is reported when answering the question, ‘What does she think?’ by something of the form, ‘She thinks that *p*.’ And the content of thought is what two people are said to share, when they are said to think the same thought. (Similarly, the content of belief is what two people share when they hold the same belief.) Content is also what may be shared in this way even while ‘psychological modes’ of states of mind may differ. For example: *believing* that I'll soon be bald and *fearing* that I'll soon be bald share the content: *that I'll soon be bald*.

Also, commonly, content is taken as not only that which is *shared* in the ways illustrated, but that which *differs* in a way revealed by considering certain logical features of sentences we use to talk about states of mind. Notably: the constituents of the sentence that fills in for ‘*p*’ when we say ‘*x* thinks that *p*’ or ‘*x* believes that *p*’ are often interpreted in such a way that they display ‘failures of substitutivity’ of (ordinarily) co-referential or co-extensional expressions, and this appears to reflect differences in mental content. For example: if George W. Bush is the eldest son of the vice-president under Ronald Reagan, and George W. Bush is the current U.S. President, then it can be validly inferred that the eldest son of Reagan's vice-president is the current U.S. President. However, we cannot always make the same sort of substitutions of terms when we use them to report what someone *believes*. From the fact that you *believe* that George W. Bush is the current U.S. President, we cannot validly infer that you believe that the eldest son of Reagan's vice-president is the current U.S. President. That last may still be false, even if George W. Bush is indeed the eldest son. These logical features of the sentences ‘*x* believes that George W. Bush is the current U.S. President’ and ‘*x* believes that George W. Bush is the eldest son of Reagan's vice-president’ seem to reflect the fact that the *beliefs* reported by their use have different *contents*: these sentences are used by someone to state *what is believed* (the belief content), and what is believed in each case is not just the same. Someone's belief may have the one content without having the other.

Similar observations can be made for other intentional states and the reports made of them -- especially when these reports contain an object clause beginning with ‘that’ and followed by a complete sentence (e.g., she thinks that *p*; he intends that *p*; she hopes that *p*; he fears that *p*; she sees that *p*). Sometimes it is said that the content of the states is ‘given’ by such a ‘that *p*’ clause when ‘*p*’ is replaced by a sentence -- the so-called ‘content clause.’

This ‘possession of content’ conception of intentionality may be coordinated with the ‘conditions of satisfaction’ conception roughly as follows. If states of mind contrast in respect of their satisfaction (say, one is true and the other false), they differ in content. (One and the same belief content cannot be both true and false -- at least not in the same context at the same time.) And if one says what the intentional *content* of a state of mind is, one says much or perhaps all of what *conditions must be met* if it is to be satisfied -- what its conditions of truth, or veridicality, or fulfillment, are. But one should be alert to how the notion of content employed in a given philosopher's views is heavily shaped by these views, and one should note how commonly it is held that the notion of content is in this or that way ambiguous or in need of refinement. (Consider, for example: Jerry Fodor's (1991) defense of a distinction between ‘narrow’ and ‘wide’ content; Edward Zalta's (1988) distinction between ‘cognitive’ and ‘objective’ content; and John Perry's (2001) distinction between ‘reflexive'and ‘subject-matter’ content.)

It is arguable that each of these gates of entry into the topic of intentionality (directedness; conditions of satisfaction; mental content) opens onto a unitary phenomenon. But evidently there is also considerable fragmentation in the conceptions of both consciousness and intentionality that are in the field. To get a better grasp of some of the ways the relationship between consciousness and intentionality can be viewed, without begging questions or trying to present a positive theory on the topic, it is useful to take a look at the recent history of thinking about intentionality, in a way that will bring several issues about its relationship with consciousness to the fore. Together with the preceding discussion, this should provide the background necessary for examining some of the differences that divide those who theorize about consciousness that are very intimately involved with views of the consciousness-intentionality relation.

If we are to acknowledge the extent to which the notion of intentionality is the creature of philosophical history, we have to come to terms with the divide in twentieth century western philosophy between so-called ‘analytic’ and ‘continental’ philosophical traditions. Both have been significantly concerned with intentionality. But differences in approach, vocabulary, and background assumptions have made dialogue between them difficult. It is almost inevitable, in a brief exposition, to give largely independent summaries of the two. We will start with the ‘continental’ side of the story -- more, specifically, with the phenomenological movement in continental philosophy. However, while these traditions have developed without a great deal of intercommunication, they do have common sources, and have come to focus on issues concerning the relationship of consciousness and intentionality that are recognizably similar.

3. Brentano, Husserl, and the Phenomenological Movement

A thorough look at the historical roots of controversies over consciousness and intentionality would take us farther into the past than it is feasible to go in this article. A relatively recent, convenient starting point would be in the philosophy of Franz Brentano. He more than any other single thinker is responsible for keeping the term ‘intentional’ alive in philosophical discussions of the last century or so, with something like its current use, and was much concerned to understand its relationship with consciousness (Brentano [1874] 1973). However, it is worth noting that Brentano himself was very aware of the deep historical background to his notion of intentionality: he looked back through scholastic discussions (crucial to the development of Descartes' immensely influential theory of ideas), and ultimately to Aristotle for his theme of intentionality (Brentano [1867] 1977). One may well go further back, to Plato's discussion (in the *Sophist*, and the *Theaetetus*) of difficulties in making sense of false belief, and yet further still, to the dawn of Western Philosophy, and Parmenides' attempt to draw momentous consequences from his alleged finding that it is not possible to think or speak of what is not.

In Brentano's treatment what seems crucial to intentionality is the mind's capacity to ‘refer’ or be ‘directed’ to objects existing solely in the mind -- what he called ‘mental or intentional inexistence.’ It is subject to interpretation just what Brentano meant by speaking of an object existing only in the mind and not outside of it, and what he meant by saying that such ‘immanent’ objects of thought are not ‘real.’ He complained that critics had misunderstood him here, and also appears to have revised his position significantly as his thought developed. But it is clear at least that his conception of intentionality is dominated by the first strand in thought about intentionality mentioned above -- intentionality as ‘directedness towards an object’ -- and whatever difficulties that brings in train.

Brentano's conception of the relation between consciousness and intentionality can be brought out partly by noting he held that every conscious mental phenomenon is both directed towards an object, and always (if only ‘secondarily’) directed towards itself. (That is, it includes a ‘presentation’ -- and ‘inner perception’ -- of itself). Since Brentano also denied the existence of unconscious mental phenomena, this amounts to the view that all mental phenomena are, in a sense ‘self-presentational.’

His lectures in the late nineteenth century attracted a diverse group of central European intellectuals (including that great promoter of the *unconscious*, Sigmund Freud) and the problems raised by Brentano's views were taken up by a number of prominent philosophers of the era, including Edmund Husserl, Alexius Meinong, and Kasimir Twardowski. Of these, it was Husserl's treatment of the Brentanian theme of intentionality that was to have the widest philosophical influence on the European Continent in the twentieth century -- both by means of its transformation in the hands of other prominent thinkers who worked under the aegis of ‘phenomenology’ -- such as Martin Heidegger, Jean-Paul Sartre, and Maurice Merleau-Ponty -- and through its rejection by those embracing the ‘deconstructionism’ of Jacques Derrida.

In responding to Brentano, Husserl also adopted his concern with properly understanding the way in which thought and experience are “directed towards objects.” Husserl ([1900] 1970, [1928] 1966) criticized Brentano's doctrine of ‘inner perception,’ and did not deny (even if he did not affirm) the reality of unconscious mentation. But Husserl retained Brentano's primary focus on describing conscious ‘mental acts.’ Also he believed that knowledge of one's own mental acts rests on an ‘intuitive’ apprehension of

their instances, and held that one is, in some sense, conscious *of* each of one's conscious experiences (though he denied this meant that every conscious experience is an object of an intentional act). Evidently Husserl wished to deny that all conscious acts are objects of inner perception, while also affirming that *some* kind of reflexivity -- one that is, however, neither judgment-like nor sense-like -- is essentially built into every conscious act. But the details of the view are not easy to make out. (A similar (and similarly elusive) view was expressed by Jean-Paul Sartre in the doctrine that “every consciousness is a non-positional consciousness of itself”([1937] 1957, [1942] 1953).) (For recent discussion of Brentano and Husserl and of the relation between the two on these topics, see Thomasson 2000 and Zahavi 1998.)

One of Husserl's principal points of departure in his early treatment of intentionality (in the *Logical Investigations* [1900] 1970) was his criticism of (what he took to be) Brentano's notion of the ‘mental inexistence’ of the objects of thought and perception. Husserl thought it a fundamental error to suppose that the object (the ‘intentional object’) of a thought, judgment, desire, etc. is always an object ‘in’ (or ‘immanent to’) the mind of the thinker, judger, or desirer. The objects of one's ‘mental acts’ of thinking, judging, etc. are often objects that ‘transcend,’ and exist independently of these acts (states of mind) that are directed towards them (that ‘intend’ them, in Husserl's terms). This is particularly striking, Husserl thought, if we focus on the intentionality of sense perception. The object of my visual experience is not something ‘in my mind,’ whose existence depends on the experience -- but something that goes beyond or ‘transcends’ any (necessarily perspectival) experience I may have of it. This view is phenomenologically based, for (Husserl says), the object is *experienced as* perspectivally given, hence as ‘transcendent’ in this sense.

In cases of hallucination, we should say, on Husserl's view, not that there is an object existing ‘in one's mind,’ but that the object intended does not exist at all. This does not do away with the ‘directedness’ of the experience, for that is properly understood (according to the *Logical Investigations*) as its having a certain ‘matter’ -- where the matter of a mental act is what may be common to different acts, when, for example, one believes that it will not rain tomorrow, and hopes that it will not rain tomorrow. The difference between the mental acts illustrated (between hoping and believing) Husserl would term a difference in their ‘quality.’ Husserl was to re-interpret his notions of act-matter and quality as components of what he called (in *Ideas* [1913] 1983) the ‘noema’ or ‘noematic structure’ that can be common to distinct particular acts. So intentional directedness is understood not as a relation to special (mental) objects towards which one is directed, but rather: as the possession by mental acts of matter/quality (or later, ‘noematic’) structure.

This unites Husserl's discussion with the ‘content’ conception of intentionality described above: he himself would accept that the matter of an act (later, its ‘noematic sense’) is the same as the *content* of judgment, belief, desire, etc., in one sense of the term (or rather, in one sense he found in the ambiguous German ‘*gehalt*’). However, it is not fully clear how Husserl would view the relationship between either act-matter or noematic sense quite generally and such semantic correlates of ordinary language sentences that some would identify as the contents of states of mind reported in them. This is a difficulty partly because of his later emphasis (e.g., in *Experience and Judgment* [1928] 1972) on the importance of what he called ‘pre-predicative’ experience. He believed that the sort of judgments we express in ordinary and scientific language are ‘founded on’ the intentionality of pre-predicative experience, and that it is a central

task of philosophy to clarify the way in which such experience of our surroundings and our own bodies underlies judgment, and the capacity it affords us to construct an 'objective' conception of the world. Pre-predicative experience is, paradigmatically, sense experience as it is given to us, independently of any active judging or predication. But did Husserl hold that what makes such experience pre-predicative is that it altogether lacks the content that is expressed linguistically in predicative judgment, or did he think that such judgment merely *renders explicit* a predicative content that even 'pre-predicative' experience already (implicitly) has? Just what does the 'pre-' in 'pre-predicative' entail?

Perhaps this is not clear. In any case, the theme of a type of intentionality more fundamental than that involved in predicative judgments that 'posit' objects, and to be found in everyday experience of our surroundings, was taken up, in different ways, by later phenomenologists, Heidegger and Merleau-Ponty. The former describes a type of 'directed' 'comportment' towards beings in which they 'show themselves' as 'ready-to-hand,' (or, in Dreyfus' (1991) interpretation, as 'available'). Heidegger thinks this characterizes our ordinary practical involvement with our surroundings, and regards it as distinct from, and somehow providing a basis for, entities showing themselves to us as 'present-at-hand' (or 'occurrent') -- as they do when we take a less context-bound, more theoretical stance towards the world (Heidegger [1927] 1962). Later, Merleau-Ponty ([1949] 1962), influenced by his study of Gestalt psychology and neurological case studies describing pathologies of perception and action, held that normal perception involves a consciousness of place tied essentially to one's capacities for exploratory and goal-directed movement, which is indeterminate relative to attempts to express or characterize it in terms of 'objective' representations -- though it makes such an objective conception of the world possible.

Whether Heidegger's and Merleau-Ponty's moves in these directions actually contradict Husserl, they clearly go beyond what he says. Another basic, exegetically complex, apparent difference between Husserl and the two later philosophers, pertinent to the relationship of consciousness and intentionality, lies in the controversy over Husserl's proposed 'phenomenological reduction.' Husserl claimed it is possible (and, indeed, essential to the practice of phenomenology) that one conduct an investigation into the structure of consciousness that carefully abstains from affirming the existence of anything in spatio-temporal reality. By this 'bracketing' of the natural world, by reducing the scope of one's assertions first to the subjective sphere of consciousness, then to its abstract (or 'ideal') atemporal structure, one is able to apprehend what consciousness and its various forms essentially are, in a way that supplies a foundation to the philosophical study of knowledge, meaning and value. Both Heidegger and Merleau-Ponty (along with a number of Husserl's other students) appear to have questioned whether it is possible to reduce one's commitments as thoroughly as Husserl appears to have prescribed through a 'mass abstention' from judgment about the world, and thus whether it is correct to regard one's intentional experience as a whole as essentially detachable from the world at which it is directed. Seemingly crucial to their doubts about Husserl's reduction is their belief that an essential part of intentionality consists in a distinctively practical involvement with the world that cannot be broken by any mere abstention from judgment.

The phenomenological themes just hinted at (the notion of a 'pre-predicative' type of intentionality; the (un)detachability of intentionality from the world) link up with issues regarding consciousness and intentionality as these are understood outside the phenomenological tradition -- in particular, the notion of non-conceptual content, and the internalism/externalism debate, to be considered in Section (4). But it is

by no means a straightforward matter to describe these links in detail. Part of the reason lies in the general difficulty in being clear about whether what one philosopher means by ‘consciousness’ (or its standard translations) is close enough to what another means for it to be correct to see them as speaking to the same issues. And while some of the phenomenological philosophers (Brentano, Husserl, Sartre) make thematically central use of terms cognate with ‘consciousness’ and ‘intentionality,’ and consider questions about intentionality first and foremost as questions about the intentionality of consciousness, they do not explicitly address much that (in the latter half of the twentieth century) came to seem problematic about consciousness and intentionality. Is *their* ‘consciousness’ the *phenomenal* kind? Would they reject theories of consciousness that reduce it to a species of access to content? If so, on what grounds? (Note: given their interest in the relation of consciousness, inner perception, and reflection, it may be easier to discern what their stance on reductive ‘higher order representation’ theories of consciousness would be.)

In some ways the situation is more difficult still in the cases of Merleau-Ponty and Heidegger. For the former, though he willingly enough uses words standardly translated as ‘consciousness’ and ‘intentionality,’ says little to explain how he understands such terms generally. And the latter deliberately avoids these terms in his central work, *Being and Time*, in order to forge a philosophical vocabulary free of errors in which they had, he thought, become enmeshed. However, it is not obvious how to articulate the precise difference between what Heidegger rejects, in rejecting the allegedly error-laden understanding of ‘consciousness’ and ‘intentionality’ (or their German translations), and what he accepts when he speaks of beings ‘showing’ or ‘disclosing’ themselves to us, and of our ‘comportment’ directed towards them.

Nevertheless, one can plausibly read Brentano's notion of ‘presentation’ as equivalent to the notion of phenomenally conscious experience, as this is understood in other writers. For Brentano ([1874] 1973) says, ‘We speak of presentation whenever something appears to us.’ And one may take *ways of appearing* as equivalent to *ways of seeming*, in the sense proper to phenomenal consciousness. Further, Brentano's attempt to state in a ‘descriptive’ or ‘phenomenological’ psychology, based on how intentional presentations present themselves, the fundamental kinds to which they belong and their necessary interrelationships, may plausibly be interpreted as an effort to articulate the philosophically salient, highly general phenomenal character of intentional states (or acts) of mind. And Husserl's attempts to delineate the structure of intentionality as it is ‘given’ in consciousness, as well as the phenomenological productions of Sartre, can arguably be seen as devoted to laying bare to thought the deepest and most general characteristics of phenomenal consciousness, as they are found in ‘directed’ perception, judgment, imagination, emotion and action. Also, one might reasonably regard Heideggerean disclosure of the ready-to-hand and Merleau-Ponty's ‘motor-intentional’ consciousness of place as forms of phenomenally conscious experience -- as long as one's conception of phenomenal consciousness is not tied to the notion that the subjective ‘sphere’ of consciousness is, in essence, independent of the world revealed through it.

In any event, to connect classic phenomenological writings with current discussions of consciousness and its relation to intentionality, more background is needed on aspects of the other main current of Western philosophy in the past century particularly relevant to the topic of intentionality -- broadly labeled ‘analytic.’

4. Frege, Russell, and the Analytic Tradition

It seems fair to say that recent work in philosophy of mind in the analytic tradition that has focussed on questions about the nature of intentionality (or ‘mental content’) has been most formed not by the writings of Brentano, Husserl and their direct intellectual descendants, but by the seminal discussions of logico-linguistic concerns found in Gottlob Frege's (1892) “On Sense and Reference,” and Bertrand Russell's “On Denoting” (1905). (Roderick Chisholm (1957), much influenced by his study of Brentano, is a notable exception.)

But Frege's and Russell's work comes from much the same era, and from much the same intellectual environment as Brentano's and the early Husserl's. And fairly clear points of contact have long been recognized, such as: Russell's criticism of Meinong's ‘theory of objects’; and the similarities between Husserl's meaning/object distinction (in *Logical Investigation* I) and Frege's (prior) sense/reference distinction. Indeed the case has been influentially made (by Follesdal 1969, 1990) that Husserl's ‘meaning/object’ distinction is borrowed from Frege (though with a change in terminology) and that Husserl's ‘noema’ is properly interpreted as having the characteristics of Fregean ‘sense.’

Nonetheless, a number of factors make comparison and integration of debates within the two traditions complicated and strenuous. Husserl's notion of noema (hence his notion of intentionality) is most fundamentally rooted, not in reflections on the logical features of language, but in a contrast between the object of an intentional act, and the object ‘as intended’ (the way in which it is intended), and in the idea that a structure would remain to perceptual experience, even if it were radically non-veridical. And what Husserl seeks is a ‘direct’ characterization of this (and other) kinds of experience from the point of view of the experiencer. On the other hand, Frege's and Russell's writings bearing on the topic of intentionality concentrate mainly and most explicitly on issues that grow from their own pioneering achievements in logic, and have given rise to ways of understanding mental states primarily through questions about the logic and semantics of the language used to speak of them.

Broadly speaking, logico-linguistic concerns have been methodologically and thematically dominant in the analytic Frege-Russell tradition, while the phenomenological Brentano-Husserl lineage is rooted in attempts to characterize experience as it is evident from the subject's point of view. For this reason perhaps, discussions of consciousness and intentionality are more obviously intertwined from the start in the phenomenological tradition than in the analytic one. The following sketch of relevant background in the latter case will, accordingly, most directly concern the treatment of intentionality. But by the end, the bearing of this on the treatment of consciousness in analytic philosophy of mind will have become more evident, and it will be clearer how similar issues concerning the consciousness-intentionality relationship arise in each tradition.

Central to Frege's legacy for discussions of mental or intentional content has been his distinction between ‘sense’ (*Sinn*) and ‘reference’ (*Bedeutung*), and his use of this distinction to cope with the apparent failures of substitutivity of (ordinarily) co-referential expressions in contexts created by psychological

verbs, of the sort mentioned above in exposition of the notion of mental content -- a task important to his development of logic. The need for a distinction between the sense and reference of an expression became evident to Frege, when he considered that, even if *a* is identical to *b*, and you understand both ‘*a*’ and ‘*b*,’ still, it can be for you a *discovery*, an addition to your knowledge, that *a* = *b*. This is intelligible, Frege thought, only if you have different *ways* of understanding the expressions ‘*a*’ and ‘*b*’ -- only if they involve for you distinct ‘modes of presentation’ of the self-same object to which they refer. In Frege's celebrated example: you may understand the expressions ‘The Morning Star’ and ‘The Evening Star’ and use them to refer to what is one and the same object -- the planet Venus. But this is not sufficient for you to know that the Morning Star is identical with the Evening Star. For the ways in which an object (‘the reference’) is ‘given’ to your mind when you employ these expressions (the senses or *Sinne* you ‘grasp’ when you use them) may differ in such a manner that ignorance of astronomy would prevent your realizing that they are but two ways in which the same object can be given.

The relevance of all this to intentionality becomes clearer, once we see how Frege applied the sense/reference distinction to whole sentences. The sentence, ‘The Evening Star = The Morning Star’ has a different sense than the sentence ‘The Evening Star = The Evening Star’, even if their reference (according to Frege, their truth value) is the same. The failure of substitutivity of co-referential expressions in ‘that *p*’ contexts created by psychological verbs can consequently be understood (Frege proposed) in this way: the reference of the terms *shifts* in these contexts, so that, for example, ‘the Evening Star’ no longer refers to its customary reference (the planet Venus), but to a sense that functions, for the subject of the verb (the person who thinks, judges, desires) as his or her mode of presentation of this object. The sentence occurring in this context no longer refers to its truth value, but to the sense in which the mode of presentation is embedded -- which might otherwise be called the ‘thought’ -- or, by other philosophers, the ‘content’ of the subject's state of mind. This thought or content is to be understood not as a mental image, or indeed as anything essentially private to the thinker's mind -- but as one and the same abstract entity that can be ‘grasped’ by two minds, and that must be so grasped if communication is to occur.

While on the surface this story may appear to be only about logic and semantics, and though Frege did not himself elaborate a general account of intentionality, what he says readily suggests the following picture. Intentional states of mind -- thinking about Venus, wishing to visit it -- involve some special relation (such as ‘mental grasping’) -- not to a Venus ‘in one's mind,’ nor to an image of Venus, but -- to an abstract entity, a thought, which also constitutes the sense of a linguistic expression that can be used to report one's state of mind, a sense which is grasped or understood by speakers who use it.

This style of account, together with the Fregean thesis that ‘sense determines reference,’ and the history of criticisms both have elicited, form much of the background of contemporary discussions of mental content. It is often assumed, with Frege, that we must recognize (as some thinkers in the empiricist tradition allegedly did not) that thoughts or contents cannot consist in images or essentially private ‘ideas.’ But philosophers have frequently criticized Frege's view of thought as some abstract entity ‘grasped’ or ‘present to’ the mind, and have wanted to replace Frege's unanalyzed ‘grasping’ with something more ‘naturalistic.’

Relatedly, it may be granted that the content of the thought reported is to be identified with the sense of the expression with which we report it. But then, it is argued, the identity of this content will not be determined individualistically, and may in some respects lie beyond the grasp (or not be fully ‘present to’ the mind of) the psychological subject. For what determines the reference of an expression may be a natural causal relation to the world -- as Saul Kripke (1972) and Hilary Putnam (1975) have influentially argued is true for proper names, like ‘Nixon’ and ‘Cicero,’ and ‘natural kind’ terms like ‘gold’ and ‘water.’ Or (as Tyler Burge (1979) has influentially argued) two speakers who, considered as individuals, are qualitatively the same, may nevertheless each assert something different simply because of differing relations they bear to their respective linguistic communities. (For example, what one speaker's utterance of ‘arthritis’ means is determined not by what is ‘in the head’ of that speaker, but by the medical experts in his or her community.) And, if reference and truth conditions of expressions by which one's thought is reported or expressed are not determined by what is in one's head, and the content of one's thought determines their reference and truth conditions, then the content of one's thought is also not determined individualistically. Rather it is necessarily bound up with one's causal relations to certain natural substances, and one's membership in a certain linguistic community. Both linguistic meaning and mental contents are ‘externally’ determined.

The development of such ‘externalist’ conceptions of intentionality informs the reception of Russell's legacy in contemporary philosophy of mind as well. Russell also helped to put in play a conception of the intentionality of mental states, according to which each such state is seen as involving the individual's ‘acquaintance with a proposition’ (counterpart to Fregean ‘grasping’) -- which proposition is at once both what is understood in understanding expressions by which the state of mind is reported, and the content of the individual's state of mind. Thus, intentional states are ‘propositional attitudes.’ Also importantly, Russell's famous analysis of definite descriptions into phrases employing existential quantifiers and general predicates underlay many subsequent philosophers' rejection of any conception of intentionality (like Meinong's) that sees in it a relation to non-existent objects. And, Russell's treatment drew attention to cases of what he called ‘logically proper names’ that apparently defy such analysis in descriptive terms (paradigmatically, the terms ‘this’ and ‘that’), and which (he thought) thus must refer ‘directly’ to objects. Reflection on such ‘demonstrative’ and ‘indexical’ (e.g., ‘I,’ ‘here,’ ‘now’) reference has led some (Kaplan 1979, Perry 1977) to maintain that the content of our states of mind cannot always be constituted by Fregean senses but must be seen as consisting partly of the very objects in the world outside our heads to which we refer, demonstratively, indexically -- another source of support for an ‘externalist’ view of mental content, hence, of intentionality.

Yet another important source of externalist proclivities in twentieth century philosophy lies in the thought that the meaningfulness of a speaker's utterances depends on its potential intelligibility to hearers: *language must be public* -- an idea that has found varying and influential expression in the work of Ludwig Wittgenstein, W.V.O. Quine, and Donald Davidson. This, coupled with the assumption that intentionality (or ‘thought’ in the broad (Cartesian) sense) must be expressible in language, has led some to conclude that what determines the content of one's mind must lie in the external conditions that enable others to attribute content.

However, the movement from Frege and Russell toward externalist views of intentionality should not

simply be accepted as yielding a fund of established results: it has been subject to powerful and detailed challenges (by Searle (1983), and Kirk Ludwig (1996b), for example). But without plunging into the details of the internalism/externalism debate about mental content, we can recognize, in the issues just raised, certain themes bearing particularly on the connection between consciousness and intentionality.

For example: it is sometimes assumed that, whatever may be true of content or intentionality, the phenomenal character of one's experience, at least, is ‘fixed internally’ -- i.e., it involves no necessary relations to the nature of particular substances in one's external environment or to one's linguistic community. But then the purported externalist finding that neither meaning nor content is ‘in the head’ may be read as showing the insufficiency of phenomenal consciousness to determine any intentionality or content. Something like this consequence is drawn by Putnam (1981), who takes the stream of consciousness to comprise nothing more than sensations and images, which (as Frege saw) should be sharply distinguished from thought and meaning. This interpretation of the import of externalist arguments may be reinforced by a tendency to tie (phenomenal) consciousness to non-intentional sensations, sensory qualities, or ‘raw feels,’ and hence to dissociate consciousness from intentionality (and allied notions of meaning and reference), a tendency that has been prominent in the analytic tradition (perhaps largely through the influence of that Ryle (1949) and Wittgenstein (1953) -- (see Section (7)).

But it is not at all evident that externalist theories of content require us to estrange consciousness from intentionality. One might argue (as do Martin Davies (1997) and Fred Dretske (1997)) that in certain relevant respects the phenomenal character of experience is also essentially determined by causal environmental connections. By contrast, one may argue (as do Ludwig (1996b) and Horgan and Tienson (2002)) that since it is conceivable that a subject have experience much like our own in phenomenal character, but radically different in external causes from what we take our own to be (in the extreme case, a mind bewitched by a Cartesian demon into massive hallucination), there must indeed be a realm of mental content that is *not* externally determined.

One other aspect of the Frege-Russell tradition of theorizing about content that impinges on the consciousness/intentionality connection is this. If ‘content’ is identified with the sense or the truth-condition determiners of the expressions used in the object-clause reporting intentional states of mind, it will seem natural to suppose that possession of mental content requires the possession of conceptual capacities of the sort involved in linguistic understanding -- ‘grasping senses.’ But then, to the extent the phenomenal character of experience is inadequate to endow a creature with such capacities, it may seem that phenomenal consciousness has little to do with intentionality.

But this raises large issues. One is this: it should not be granted without question that the phenomenal character of our experience could be as it is in the absence of the sorts of conceptual capacities sufficient for (at least some types of) intentionality. And this is tied to the issue of whether or not the phenomenal character of experience is (as some suppose) a purely sensory affair. Some would maintain, on the contrary, that thought (not just imagistic, but conceptual thought) has phenomenal character too. (See Flanagan 1992, Horgan and Tienson 2002, Siewert 1998, Strawson 1994.) If so, then it is very far from clear that phenomenal character can be divorced from whatever conceptual capacities are necessary for intentionality.

Moreover we may ask: are *concepts*, properly speaking, always necessary for intentionality anyway? Here another issue rears its head: is there not perhaps a form of *sensory* intentionality, which does not require anything as distinctively intellectual or conceptual as is needed for the grasping of linguistic senses or propositions? (This presumably would be a kind of intentionality had by the pre-linguistic (e.g., babies) or by non-linguistic creatures (e.g., dogs).) Suppose that there is, and that this type of intentionality is inseparable from the phenomenal character of perceptual experience. Then, even if one assumes that such phenomenal consciousness is insufficient to guarantee the possession of concepts, it would be wrong to say that it has little to do with intentionality. (Advocates of varying versions of the idea that there is a distinctively ‘non-conceptual’ kind of content include Bermudez 1998, Crane 1992, Evans 1982, Peacocke 1992, and Tye 1995 -- for a notable voice of opposition to this trend, see McDowell 1994.) A deep difficulty in assessing these debates lies in getting an acceptable conception of *concepts* to work with. We need to understand clearly what ‘having a concept of *F*’ does and does not require, before we can be clear about the content of and justification for the thesis of *non*-conceptual content.

These proposals about non-conceptual content bear some affinity with aspects of the phenomenological tradition alluded to earlier: Husserl's notion of ‘pre-predicative’ experience; Heidegger's treatment of the ‘ready-to-hand;’ and Merleau-Ponty's idea that in normal active perception we are conscious of place, not via a determinate ‘representation’ of it, but rather, relative to our capacities for goal-directed bodily behavior. Though to see the extent to which any of these are ‘non-conceptual’ in character would require not only more clarity about the conceptual/non-conceptual contrast, but considerable novel exegesis of these philosophers' works. (For one recent effort to draw a connection between phenomenological and analytic traditions on the issue of ‘non-conceptual content’ see Kelly 2001.)

Also, one may plausibly try to find an affinity between externalist views in analytic philosophy, and the later phenomenologists' rejection of Husserl's reduction, based on their doubt that we can prise consciousness off from the world at which it is directed, and study its ‘intentional essence’ in solipsistic isolation. But if externalism can be defined broadly enough to encompass Heidegger, Merleau-Ponty, Kripke, and Burge, still the comparison is strained when we take account of the different sources of ‘externalism’ in the phenomenologists. These have to do it seems (very roughly) with the idea that the way we are conscious of things (or at least, for Heidegger, the way they ‘show themselves’ to us) in our everyday activity cannot be quite generally separated from our actual engagement with entities of which we are thus conscious (which show themselves in this way). Also relevant is the idea that one's use of language (hence one's capacity for thought) requires gearing one's activity to a social world or cultural tradition, in which antecedently employed linguistic meaning is taken up and made one's own through one's relation with others. All this is supposed to make it infeasible to study the nature of intentionality by globally uprooting, in thought, the connection of experience with one's spatial surroundings (and -- crucially for Merleau-Ponty -- one's own body), and one's social environment. Whatever the merits of this line of thought, we should note: neither a causal connection with ‘natural kinds’ unmediated by reference-determining ‘modes of presentation,’ nor deference to the linguistic usage of specialists, nor belief in the need to reconstruct speaker's meaning from observed behavior, plays a role in the phenomenologists' doubts about the reduction.

The arduous exegesis required for a clearer and more detailed comparison of these views is not possible here. Nevertheless, following some of the main lines of thought in treatments of intentionality, descending on the one hand, primarily from Brentano and Husserl, and on the other, from Frege and Russell, certain fundamental issues concerning its relationship to consciousness have emerged. These include, first, the connection between consciousness and self-directed or self-reflexive intentionality. (It has already been seen that this topic preoccupied Brentano, Husserl and Sartre; its emergence as an important issue in analytic philosophy of mind will become more evident below, in Section 6.) Second, there is concern with the way in which (and the extent to which) mind is world-involving. (In the phenomenological tradition this can be seen in controversy over Husserl's phenomenological reduction; within the analytic tradition, in the critique of Fregean sense and the internalism/externalism debate.) Third, there is the putative distinction between conceptual or theoretical, and sensory or practical forms of intentionality. (In phenomenology this shows up in Husserl's contrast between judgment and pre-predicative experience, and related notions of his successors; in analytic philosophy this shows up in the (more recent) attention to the notion of ‘non-conceptual’ content.)

For more clarity regarding the consciousness-intentionality relationship and how these three topics figure prominently in views about it, it is necessary now to turn attention back to philosophical disagreements regarding consciousness that are much bound up with the distinctions mentioned in Section (1), among phenomenal consciousness, access consciousness, and reflexive consciousness.

5. Consciousness and Sensory Content

Consider the proposal that sense experience manifests a kind of intentionality distinct from and more basic than that involved in propositional thought and conceptual understanding. This might help form the basis for an account of consciousness. Perhaps conscious states of mind are distinguished partly by their possession of a type of content proper to the sensory subdivision of mind.

One source of the idea that a difference in type of content helps constitute a distinction between what is and is not phenomenally conscious, lies in the apparent distinction between sense experience and judgment. To have conscious *visual experience* of a stimulus -- for it to look some way to you -- is one thing. To make judgments about it is something else. (This seems evident in the persistence of a visual illusion, even once one has become convinced of the error.) However, on some accounts of consciousness, this distinction itself is doubtful, since conscious sense experience is taken to be nothing more than a form a judging. Such a view is expressed by Daniel Dennett (1991), who takes the relevant form of judging to consist in one's possession of information or mental content available to the appropriate sort of ‘probes’ -- the availability of content he calls ‘cerebral celebrity.’ For Dennett what distinguishes conscious states of mind is not their possession of a distinctive type of intentional content, but rather the richness of that content, plus its availability to the appropriate sort of cognitive operations. (Since the relevant class of operations is not sharply defined, neither, for Dennett, is the difference between which states of mind are conscious and which are not.)

Recent accounts of consciousness that, by contrast, give central place to a distinction between

(conceptual) judgment and (non-conceptual -- but still intentional) sense-experience include Michael Tye's (1995) theory, holding that it is (by metaphysical necessity) sufficient to have a conscious sense-perception that some representation of sensory stimuli is formed in one's head, ‘map-like’ in character, whose (‘non-conceptual’) content is ‘poised’ to affect one's (conceptual) beliefs. This form of mental representation Tye would contrast with the ‘sentential’ form proper to belief and judgment -- and in that way, he might preserve the judgment/experience contrast as Dennett does not. Consider also Fred Dretske's (1995) view, that phenomenally conscious sensory intentionality consists in a kind of mental representation whose content is bestowed through a naturally selected ‘function to indicate.’ Such natural (evolution-implanted) sensory representation can arise independently of learning (unlike the more conceptual, language dependent sort), and is found widely distributed among evolved life.

Both Tye's and Dretske's views of consciousness (unlike Dennett's) make crucial use of a contrast between the types of intentionality proper to sense-experience, and that proper to linguistically expressed judgment. On the other hand, there is also some similarity among the theories, which can be brought out by noting a criticism of Dennett's view, analogues of which arise for Tye's and Dretske's views as well.

Some might think Dennett's account concerns only some variety of what Block would call ‘access consciousness.’ For on Dennett's account, it seems, to speak of visual consciousness is to speak of nothing over and above the sort of availability of informational content that is evinced in unprompted verbal discriminations of visual stimuli. And this view has been criticized for neglecting phenomenal consciousness. (See Block 1995 and Siewert 1998.) It seems we may conceive of a capacity for spontaneous judgment triggered by and responsive to visual stimuli, which would occur in the absence of the judger's phenomenally conscious visual experience of the stimuli: the stimuli don't *look* any way to the subject, and yet they trigger accurate judgments about their presence. The notion of such a (hypothetical) form of ‘blindsight’ may be elaborated in such a way that we conceive of the judgment it affords as being at least as finely discriminatory (and as fine in informational content) as that enjoyed by those with extremely poor, blurry and un-acute conscious visual experience (as in the ‘legally blind’).But a view like Dennett's seems to make this scenario inconceivable.

However, this kind of criticism does not concern only those theories that would elide any experience/judgment distinction. For Tye's and Dretske's theories, though they depend on forms of that contrast (and are offered as theories of *phenomenal* consciousness), can raise similar concerns. For one might think that the hypothetical blindsighter would be as rightly regarded as having Tye's ‘poised’ maplike representations in her visual system as would someone with a comparable form of conscious vision. And one might find it unclear why we should think the visual system of such a blindsighter must be performing naturally endowed indicating functions more poorly than the visual system of a consciously sighted subject would.

Whatever the cogency of these concerns, one should note their distinctness from the issues about ‘kinds of intentionality’ that appear to separate both Tye and Dretske from Dennett. The notion that there is a fundamental distinction to be drawn in kinds of intentional content (separating the more intellectual from the more sensory departments of mind) sometimes forms the basis of an account of consciousness (as with Dretske's and Tye's, though not with Dennett's). But it is also important to recognize what unites

Dennett, Tye, and Dretske. Despite their differences, all propose to account for consciousness by starting with a general understanding of intentionality (or mental content or representation) to which consciousness is inessential. They then offer to explain consciousness as a special case of intentionality thus understood -- so, in terms of the operations the content is available for, or the form in which it is represented, or the nature of its external source. The blindsight-based objection to Dennett, and its possible extension to Dretske and Tye, helps bring this commonality to light.

6. Consciousness and Higher Order Representation

The last section showed how some theories purport to account for consciousness on the basis of intentionality, in a way that focuses attention on attempts to discern a distinctively sensory type of intentionality. A different strategy for explaining consciousness via intentionality highlights the importance of clarity regarding the connection between consciousness and reflexivity. On such a view (roughly): experiences or states of mind are conscious just insofar as the mind represents itself as having them.

In David Rosenthal's (1986, 1991, 1993) variant of this approach, a state is conscious just when it is a kind of (potentially non-conscious) mental state one has, which one (seemingly without inference) thinks that one is in. A theory of this sort starts with some way of classifying mental states that is supposed to apply to conscious and non-conscious states of mind alike. The proposal then is that such a state is conscious just when it belongs to one of those mental kinds, and the ('higher order') thought occurs to the person in that state that he or she is in a state of that kind. So, for example it is maintained that certain non-conscious states of mind can possess 'sensory qualities' of various sorts -- one may, in a sense, be *in* pain without feeling pain, one may have a red sensory quality, even when nothing looks red to one. The idea is that one has a conscious visual experience of red, or a conscious pain sensation, just when one has such a red sensory quality, or pain-quality, and the thought (itself also potentially non-conscious) occurs to one that one has a red sensory quality, or pain-quality.

This way of accounting for consciousness in terms of intentionality may, like theories mentioned in Section 5, provoke the concern that the distinctively phenomenal sense of consciousness has been slighted -- though this time, not in favor of some 'access' consciousness, but in favor of reflexive consciousness. One focus of such criticism lies in the idea that such higher-order thought requires the possession of concepts -- concepts of types of mental states -- that may be lacking in creatures with first order mentality. And it is unclear (in fact it seems false to say) these beings would therefore have no conscious sensory experience in the phenomenal sense. Mightn't there be a way the world looks to rabbits, dogs, monkeys, and human babies, and mightn't they feel pain, though they lack the conceptual wherewithal to think about their own experience?

One line of response to such concerns is simply to bite the bullet: dogs, babies and the like might altogether lack higher order thought, but that's no problem for the theory because, indeed, they also altogether lack *feelings* (Carruthers 1989). Rosenthal, for his part, takes a different line: lack of cognitive sophistication needn't instantly disqualify one for consciousness, since the possession of primitive

mentalistic concepts requires so little that practically any organism we would consider a serious candidate for sensory consciousness (certainly babies, dogs and bunnies) would obviously pass muster.

A number of additional worries have been raised about both the necessity and the sufficiency of ‘higher order thought’ for conscious sense experience. (See, for example, Dretske 1993, Guzeldere 1997, Seager 1999, and Siewert 1998.) In the face of such doubts, one may preserve the idea that consciousness consists in some kind of higher order representation -- the mind's ‘scanning’ itself -- by abandoning ‘higher order thought’ for some other form of representation: one which is not thought-like or conceptual, but somehow *sensory* in character (Lycan 1996). Maybe somewhat as we can distinguish between primitive sensory perception of things in our environment, and the more intellectual, conceptual operations based on them, so we can distinguish the thoughts we have about our own (‘inner’) mental goings-on from the (‘inner’) sensing of them. And, if we propose that consciousness consists in this latter sort of higher order representation, it seems we will escape the worries occasioned by the Rosenthalian variant of the ‘reflexivist’ doctrine. In considering such theories, two of the consciousness-themes earlier discerned come together, namely: reflexivity (or higher order representation), and the contrast between the conceptual and non-conceptual (or sensory).

Criticism of ‘inner sense’ theories is likely to focus not so much on the thought that such inner sensing can occur without phenomenal consciousness, or that the latter can occur without the former, as on the difficulty in understanding just what inner sensing (as distinct from higher order thought) is supposed to be, and why we should think we have it. It seems the inner sense theorists share with those who distinguish between conceptual and non-conceptual (or sensory) flavors of intentionality the challenge of clarifying and justifying some version of this distinction. But they bear the additional burden of showing how such a distinction can be applied not just to intentionality directed at tables and chairs, but at the "furniture of the mind" as well. One may grant that there are non-conceptual sensory experiences of objects in one's external environment while doubting one has anything analogous regarding the ‘inner’ landscape of mind.

It should be noted that, in spite of the difficulties faced by higher order representation theories, they draw on certain perennially influential sources of philosophical appeal. We do have some willingness to speak of conscious states of mind as states we are conscious or aware *of* being in. It is tempting to interpret this as indicating some kind of reflexivity. And the history of philosophy reveals many thinkers attracted to the idea that consciousness is inseparable from some kind of self-reflexivity of mind. As noted varying versions of this idea can be found in Brentano, Husserl, and Sartre. And we can go further back: Kant (1787) spoke explicitly of ‘inner sense,’ and Locke (1690) defined consciousness as the ‘perception of what passes in a man's mind.’ Brentano (controversially) interpreted Aristotle's enigmatic and terse discussion of “seeing that one sees” in *De Anima* III.2 as an anticipation of his own ‘inner perception’ view.

However, there is this critical difference between the thinkers just cited and contemporary purveyors of higher order representation theories. The former do not maintain, as do the latter, that consciousness consists in one's forming the right sort of higher order representation of a possibly non-conscious type of mental state. Even if they think that consciousness is inseparable from some sort of mental reflexivity,

they do not suggest that consciousness can, so to speak, be analyzed into mental parts, none of which themselves essentially require consciousness. (Some could not maintain this, since they explicitly deny mentality without consciousness.) There is a difference between saying that reflexivity is essential to consciousness and saying that consciousness just consists in or is reducible to a species of mental reflexivity. Advocacy of the former without advocacy of the latter is certainly possible. (See, for example, Smith 1986, Levine 2001, Ludwig 2001.)

7. Is Phenomenal Consciousness Intentional?

Suppose one holds that phenomenal consciousness is distinguishable both from ‘access’ and ‘reflexivity,’ and that it cannot be explained as a special case of intentionality. One might conclude from this that phenomenal consciousness and intentionality comprise two quite distinct realms of the mental, and embrace the idea that the phenomenal is a matter of non-intentional qualia or raw feels. One important current in the analytic tradition has evinced this attitude -- it is found, for example, in Wilfrid Sellars' (1956) distinction between ‘sentience’ (sensation) and ‘sapience.’ Whereas the qualities of *feelings* involved in the former -- mere sensations -- require no cognitive sophistication and are readily attributable to brutes, the latter -- involving awareness *of*, awareness *that* -- requires that one have the appropriate concepts, which cannot be guaranteed by just having sensations; one needs learning and inferential capacities of a sort Sellars believed possible only with language. “Awareness,” Sellars says, “is a linguistic affair.”

Thus we may arrive at a picture of mind that places sensation on one side, and thought, concepts, and ‘propositional attitudes’ on the other. If one recognizes a distinctively phenomenal consciousness not captured in ‘representationalist’ theories of the kinds just scouted, one may then want to say: that is because the phenomenal belongs to mere sentience, and the intentional to sapience. Other influential philosophers of mind have operated with a similar picture. Consider Gilbert Ryle's (1949) contention that the stream of consciousness contains nothing but sensations that provide “no possibility of deciding whether the creature that had these was an animal or a human being; an idiot, a lunatic, or a sane man” -- nothing of which it is appropriate to ask whether it is correct or incorrect, veridical or nonveridical. And Wittgenstein's (1953) influential criticisms of the notion of understanding as an ‘inner process,’ and of the idea of a language for private sensation divorced from public criteria, could be interpreted in ways that sever (phenomenal) consciousness from intentionality. (Such an interpretation would assume that if consciousness could secure understanding, understanding *would* be an ‘inner process,’ and if phenomenal character bore intentionality with it, private sensations *could* impart meaning to words.) Also recall Putnam's conviction that the (internal) stream of consciousness cannot furnish the (externally fixed) content of meaning and belief. A similar attitude is evident in Donald Davidson's (1983, 1986) distinction between sensation and thought (the former is nothing more than a causal condition of knowledge, while the latter can furnish reasons and justifications, but cannot occur without language). Richard Rorty (1979) makes a Sellarsian distinction between the phenomenal and the intentional key to his polemic against epistemological philosophy in general, and ‘foundationalism’ in particular (and takes a generally deflationary view of the phenomenal or ‘qualitative’ side of this divide).

But it is possible to reject attempts to subsume the phenomenal under the intentional as in the 'representationalist' accounts of consciousness variously exemplified in Dennett, Dretske, Lycan, Rosenthal,and Tye, without adopting this 'two separate realms' conception. We can believe that there is no conception of the intentional from which the phenomenal can be explanatorily derived that does not already include the phenomenal, but still believe also that the phenomenal character of experience cannot be separated from its intentionality, and that having experience of the right sort of phenomenal character is sufficient for having certain forms of intentionality (McGinn 1991, Horgan and Tienson 2002, Siewert 1998). Here one might leave open the question whether there is also some kind of phenomenal character (perhaps that involved in some kinds of bodily sensation or after-images) whose possession is *not* sufficient for intentionality. (Though if we say there is such non-intentional phenomenal character, this would give us a special reason for rejecting the representationalist explanations of phenomenal consciousness mentioned in Section 5.) If, on the other hand, we say phenomenal character always brings intentionality with it, that might be 'representationalism' of a sort. But its endorsement is consistent with a rejection of attempts to derive phenomenality from intentionality, or reduce the former to a species of the latter, which commonly attract the 'representationalist' label. We should distinguish the question of whether the phenomenal can be *explained by* the intentional from the question of whether the phenomenal is *separable from* the intentional.

Closer consideration of two of the three themes earlier identified as common to phenomenological and analytic traditions is needed to come to grips with the latter question. It is necessary to inquire: (a) whether an externalist conception of intentionality can justify separating phenomenal character from intentionality. And one needs to ask: (b) whether one's verdict on the 'separability' question stands or falls with acceptance of some version of a distinction between conceptual and non-conceptual (or distinctively sensory) forms of intentionality.

The dialectical situation regarding (a) is complex. One way it may seem plausible to answer question (a) in the affirmative, and restrict phenomenal character and intentionality to different sides of some internal/external divide, is to conduct a Cartesian thought experiment, in which one conceives of consciousness with all its subjective riches surviving the utter annihilation of the spatial realm of nature. (Similarly, but less radically, one may conceive of a 'brain in a vat' generating an extended history of sense experience indistinguishable in phenomenal character from that of an embodied subject.) If one is committed to an externalist view of intentionality -- but rejects the intentionalizing strategies of Sections (5) and (6) for dealing with consciousness -- one may conclude that phenomenal character is altogether separable from (and insufficient for) intentionality. However, one may draw rather different conclusions from the Cartesian thought experiment -- turning it *against externalism*. It may seem to one that, since the intentionality of experience would apparently survive along with its phenomenal character, one may then infer that the causal tie between the mind's content and the world of objects beyond it that (according to some versions of externalism) fixes content, is in reality and in at least some cases (or for some contents), no more than contingent. (This is the lesson Husserl, Ludwig, Horgan and Tienson would draw.)

Alternatively, whatever one relies on to argue that this or that relation of experience and world is essential to having any intentionality at all, one may well take this to show that phenomenal character is also externally determined in a way that renders the Cartesian scenario of consciousness totally unmoored

from the world an illusion. And, if Merleau-Ponty or Heidegger think that Husserl's phenomenological reduction to a sphere of ‘pure’ consciousness cannot be completed, and their reasons make them externalists of some sort, it hardly seems to establish that they are committed to a (Sellarsian or Rylean) realm of raw sensory phenomenal consciousness, devoid of intentionality. In fact their rejection of Husserl's notion of ‘uninterpreted’ sensory or ‘hyletic’ data in experience would seem to indicate they, at least, would strongly deny they held such views.

In this arena it is far from clear what we are entitled to regard as secure ground and what as ‘up for grabs.’ However, there do seem to be ways in which all would probably admit that the phenomenal character of experience and externally individuated content come apart, ways in which such content goes beyond anything phenomenal consciousness can supply. For the way it seems to me to experience this computer screen may be no different from the way it seems to my twin to experience some entirely distinct one. Thus where intentional contents are distinguished in such a way as to include the particular objects experienced or thought of, phenomenal character cannot determine the possession of content. Still, that does not show that no content of any sort is fixed by phenomenal character. Perhaps, as some would say, phenomenal character determines ‘narrow’ or ‘notional’ content, but not ‘wide’ (externally ‘fixed’) content. Nor is it even clear that we must judge the sufficiency of phenomenal character for intentionality by adopting some general account of content and its individuation (as ‘narrow’ or ‘wide’ for instance), and then ask whether one's possession of content so considered is entailed by the phenomenal character of one's experience. One may argue that the phenomenal character of one's experience suffices for intentionality as long as having it makes one assessable for truth, accuracy (or other sorts of ‘satisfaction’) without the addition of any interpretation, properly so-called, such as is involved in assessment of the truth or accuracy of sentences or pictures (Siewert 1998).

Even if one does not globally divide phenomenal character from intentionality along some inner/outer boundary line, to address questions of the sufficiency of phenomenal character for intentionality (and thus of the separability of the latter from the former), one still needs to look at question (b) above, and the potential relevance of distinctions that have been proposed between conceptual and non-conceptual forms of content or intentionality. Again the situation is complex. Suppose one regards the notion of non-conceptual intentionality or content as unacceptable on the grounds that all content is conceptual. But suppose one also thinks it is clear that phenomenal character is confined to sensory experience and imagery, and that this cannot bring with it the rational and inferential capacities required for genuine concept possession. Then one will have accepted the separability of phenomenal consciousness from intentionality. However, one may, by contrast, take the apparent susceptiblity of phenomenally conscious sense experience to assessment for accuracy, without need for additional, potentially absent interpretation, to show that the phenomenal character of experience is inherently intentional. Then one will say that the burden lies on anyone who claims conceptual powers are crucial to such assessability and can be detached from the possession of such experience: they must identify those powers and show that they are both crucial and detachable in this way. Additionally (as noted in Section 4), one may reasonably challenge the assumption that phenomenal consciousness is indeed confined to the sensory realm; one may say that conceptual thought also has phenomenal character. Even if one does not, one may still base one's confidence in the sufficiency of phenomenal character for intentionality on one's confidence that there is a kind of non-conceptual intentionality that clearly belongs essentially to sense experience. (Such

presumably would be Michael Tye's position.)

these considerations, we can see that it is critical to answer the following questions in order to decide whether or not phenomenal character is wholly or significantly separable from intentionality.

i. Does every sort of intentionality that belongs to thought and experience require external connections for which phenomenal character is insufficient?
ii. Does every sort of intentionality that belongs to sense-experience and sensory imagery require conceptual abilities for which phenomenal character is insufficient?
iii. Does every sort of intentionality that belongs to thought require conceptual capacities for which phenomenal character is insufficient?

Suppose one finds phenomenal character quite generally inadequate for the intentionality of thought and sense-experience by answering ‘yes’ either to (i), or to both (ii) and (iii). And suppose one makes the plausible (if non-trivial) assumption that what guarantees intentionality for neither sensory experience, nor imagery, nor conceptual thought, guarantees no intentionality that belongs to our minds (including that of emotion, desire and intention -- for these latter presuppose the former). Then one will find phenomenal character altogether separable from intentionality. Phenomenal character could be as it is, even if intentionality were completely taken away. There is no form of phenomenal consciousness, and no sort of intentionality, such that the first suffices for the second.

A more moderate view might merely answer only one of either (ii) or (iii) in the affirmative (and probably (iii) would be the choice). But still, in that case one recognizes some broad mental domain whose intentionality is in no repsect guaranteed by phenomenal character. And that too would mark a considerable limitation on the extent to which phenomenal consciousness brings intentionality with it.

On the other hand, suppose that one answers ‘no’ to (i), and to either (ii) or (iii). Now, external connections and conceptual capacities seem to be what we might most plausibly regard as conditions necessary for the intentionality of thought and experience that could be stripped away while phenomenal character remains constant. So if one thinks that actually neither are generally essential to intentionality and removable while phenomenal character persists unchanged, and one can think of nothing else that is essential for thought and experience to have any intentionality, but for which phenomenal character is insufficient, it seems reasonable to conclude that phenomenal character is indeed sufficient for intentionality of some sort. If one has gone this far, it seems unlikely that one will then think that actual differences in phenomenal character still leave massively underdetermined the different forms of intentionality we enjoy in perceiving and thinking. So, one will probably judge that some kind of phenomenal character suffices for, and is inseparable from, many significant forms of intentionality in at least one of these domains (sensory or cognitive): there are many differences in phenomenal character, and many in intentionality, such that you cannot have the former without the latter. If one also rejects both (ii) and (iii), then one will accept that (appropriate forms of) phenomenal consciousness are sufficient for a very broad and important range of human intentionality.

8. Is Phenomenality Essential to Intentionality?

Suppose one rejects both the view that consciousness is explanatorily derived from a more fundamental intentionality, as well as the view that phenomenal character is insufficient for intentionality because it is a matter of pure inward feel. It seems one might then press farther, and argue for what Flanagan (1992) calls ‘consciousness essentialism’ -- the view that the phenomenal character of experience is not only sufficient for various forms of intentionality, but necessary also.

This type of thesis needs careful formulation. It does not necessarily commit one to a Cartesian (or Brentanian or Sartrean) claim that all states of mind are conscious -- a total denial of the reality of the unconscious. A more qualified thesis does seem desirable. Freud's waning prestige has weakened tendencies to assume that he had somehow demonstrated the reality of unconscious intentionality, the rise of cognitive science has created a new climate of educated opinion that also takes elaborate non-conscious mental machinations for granted. Even if we do not acquiesce in this view, we do (and long have) appealed to explanations of human behavior that recognize some sort of intentional states other than phenomenally conscious experiences and thoughts.

One way of maintaining the necessity of consciousness to mind that can preserve some space for mind that is not conscious is Searle's (1990, 1992). He argues, roughly, that we should first distinguish between what he calls ‘intrinsic’ intentionality on the one hand, and merely ‘as if’ intentionality, and ‘interpreter relative’ intentionality, on the other. We sometimes may speak *as if* artifacts (like thermostats) had beliefs or desires -- but this isn't to be taken literally. And we may impose ‘conditions of satisfaction’ on our acts and creations (words, pictures, diagrams, etc.) by our *interpretation* of them -- but they have no intentionality independent of our interpretive practices. Intrinsic intentionality, on the other hand -- the kind that pertains to our beliefs, perception, and intentions -- is neither a mere ‘manner of speech,’ nor is our possession of it derived from others' interpretive stance towards us. But then, Searle asks, what accounts for the fact that some states of affairs in world have intrinsic intentionality -- that they are directed at objects under aspects -- and why they are directed under the aspects they are (why they have the content they do)? With conscious states of mind, Searle says, their phenomenal or subjective character determines their ‘aspectual shape.’ Where non-conscious states of mind are concerned, there is nothing to do the job, but their relationship to consciousness. The right relationship, he holds, is this: non-conscious states of mind must be ‘potentially conscious.’ If some psychological theories (of language, of vision) postulate an unconscious so deeply buried that its mental representations cannot even potentially become conscious, so much the worse for those theories.

Searle's views have aroused a number of criticisms. (See the peer commentary in response to Searle 1990.) Among the problem areas are these. First, how are we to spell out the requirement that intrinsically intentional states be ‘potentially conscious,’ without making it either too easy or to difficult to satisfy? Second, just why is it that the intrinsic intentionality of non-conscious states needs accounting for, while that of conscious states is somehow unproblematic? Third, it appears Searle's argument does not offer some general reason to rule out all efforts to give ‘naturalistic’ accounts of conditions sufficient to impose -- without the help of consciousness -- genuine and not merely interpreter relative intentionality.

Another approach is taken by Kirk Ludwig (1996a), who argues that there is nothing to determine *whose* state of mind a given non-conscious state of mind is, unless that state consists in a disposition to produce a conscious mental state of the right sort. Alleged mental processes that did not tend to produce someone's conscious states of mind appropriately would be no one's, which is to say that they would not be mental states at all. Roughly: consciousness is needed to provide that unity of mind without which there would be no mind. And Ludwig argues that it is therefore a mistake to attribute many of the unconscious inferences with which psychological theorists have long been wont to populate our minds.

The persuasiveness of Searle's and Ludwig's arguments depends heavily on demonstrating the failure of alternative accounts of the job that they enlist consciousness to do (such as secure ‘aspectual shape,’ or ownership). One may well grant (as does Colin McGinn 1991) that phenomenal character is inseparable from intentionality, but cannot be explained by it, while still maintaining that genuine intentionality (mental content) is quite adequately imposed on animal brains by their acquisition of natural functions of content-bearing -- in which consciousness evidently plays no essential role. Or one may (like Jerry Fodor 1987) maintain a robustly realist ‘representational theory of mind,’ proposing that the content of mental symbols is stamped on them by their being in the ‘right causal relation’ to the world -- while despairing of the prospects for a credible naturalistic theory of consciousness.

9. Four Views of the Consciousness-Intentionality Relationship

The preceding discussion has conveyed some of the complexities and potential ambiguities in talk of ‘consciousness’ (Section 1) and ‘intentionality’ (Section 2) that must be appreciated if one is to resolve questions about the relationship between consciousness and intentionality with any clarity. Brief surveys of relevant aspects of phenomenological (Section 3) and analytic (Section 4) traditions have brought out some shared areas of interest, namely: the relationship of consciousness to reflexivity and ‘self-directed’ intentionality; efforts to distinguish between conceptual and non-conceptual (or sensory) forms of intentionality; and a concern with the extent to which the character of either conscious experience or intentional states of mind is essentially ‘world-involving.’ These concerns were seen to bear on attempts to account for consciousness in terms of intentionality in Sections 5 and 6, and also on questions that arise even if those attempts are rejected -- questions regarding the separability of phenomenal consciousness and intentionality (Section 7). In Section 8 some attention was given to views that, in some sense, reverse the order of explanation proposed by intentionalizing views of consciousness, and take the facts of consciousness to explain the facts of intentionality. Now it is possible to step back and distinguish four general views of the consciousness-intentionality relationship discernable in the philosophical positions canvassed above, as follows.

a. Consciousness is explanatorily derived from intentionality.
b. Consciousness is underived and separable from intentionality.
c. Consciousness is underived but also inseparable from intentionality.
d. Consciousness is underived from, inseparable from, and essential to intentionality.

To adopt view (a) is to accept some intentionalizing strategy with respect to consciousness, such as is variously represented by Dennett, Dretske, Lycan, Rosenthal, and Tye. These views differ importantly among themselves, and their differences have much to do with how they treat consciousness-reflexivity issues and the conceptual/non-conceptual (or conceptual/sensory) contrast, and how they view the intersection between the two. But if we adopt (a), then our answer to the question of what consciousness has to do with intentionality will ultimately be given in some prior general account of content or intentionality. And there will be no special issue regarding the internal or external fixation of the phenomenal character of experience, over and above what arises for mental content generally.

On the other hand, suppose one rejects (a), and holds that experiences are conscious in a phenomenal sense that does not yield to an approach in which one conceives of intentionality (or content, or information bearing) independently of consciousness, and then, by adverting to special operations, or sources, or contents, tells us what consciousness is. At this point, one would face a choice between (b) and (c).

Adopting (b) yields the ‘raw feel’ conception of phenomenality seemingly implicit in Sellars and Ryle. If, on the other hand, we adopt (c), we endorse a much more intimate relationship between consciousness and intentionality. Without proposing to account for the former on the basis of the latter, we would hold that phenomenal character is sufficient for intentionality.

But adoption of (c) leaves open a further basic question. Consciousness (of the appropriate sort) may be sufficient for (but underived from) intentionality, and yet, intentionality does not require consciousness. Thus we come to ask whether having conscious experience of an appropriate sort is necessary to having either sensory or more-than-sensory (conceptual) intentionality.

If adopting thesis (d), we say ‘yes’ -- that such intentionality can come only with consciousness -- we will probably have gone as far in making consciousness fundamental to mind as one reasonably can. Again, this is not necessarily to deny the reality of non-conscious mental phenomena. But it could, in a broad way, be interpreted as siding with Husserl, Ludwig and Searle in thinking of consciousness as the irreplaceable source of intentionality and meaning.

This abstract list of four options might leave one without a sense of what is at stake in adopting this or that view. Perhaps the positions themselves will become a little clearer if we make explicit four broad areas of philosophical concern to which the choice among them is relevant.

First, they are relevant to the issue of how to conceive of the mind or the domain of psychology *as a whole*. Is there some *unity* to the concept of mind or psychological phenomenal? Is there something that deserves to be considered the essence of the mental? If consciousness can be thoroughly intentionalized (as (a) would have it), maybe (with suitable qualifications), we could uphold the thesis that intentionality is the "mark of the mental." If we reject (a) and embrace (c), seeing intentionality as inseparable from the phenomenal character of experience, then we still might maintain that both consciousness and intentionality are necessary for real minds -- at least, if we adopt (d) as well. But a unified view of the

mind seems difficult (if not impossible) to maintain if one segregates phenomenal character to non-intentional sensation -- as in (b). Even if one does not, one may lack a unifying conception of the mental domain, if one is not satisfied with arguments that show that phenomenal consciousness is essential to genuine (not merely “as if” or “interpreter derived”) intentionality. In any case, both consciousness and intentionality are broad enough psychological categories, that one's view of their extension and relationship will do much to draw one's map of psychology's terrain.

Second (and relatedly), views about the consciousness-intentionality relationship bear significantly on general questions about the *explanation* of mental phenomena. One may ask what kinds of things we might try to explain in the mental domain, what sorts of explanations we should seek, and what prospects of success we have in finding them. If we accept (a) and some intentionalizing account of consciousness, we will not suppose as do some (Chalmers 1996, Levine 2001, McGinn 1991, and Nagel 1974) that phenomenal consciousness poses some specially recalcitrant (maybe hopelessly unsolvable) problem for reductive physicalist or materialist explanations. Rather, we will see the basic challenge as that of giving a natural scientific account of intentionality or mental representation. And this indeed is a reason some are attracted to (a). One may believe that it offers us the only hope for a natural scientific understanding of consciousness. The underlying thought is that a science of consciousness must adopt this strategy: first conceive of intentionality (or content or mental representation) in a way that separates it from consciousness, and see intentionality as the outcome of familiar (and non-intentional) natural causal processes. Then, by further specifying the kind of intentionality involved (in terms of its use, its sources, its content), we can account for consciousness. In other words: ‘naturalize’ intentionality, then intentionalize consciousness, and mind has found its place in nature.

However, we should recognize a distinction between those whose envisioned naturalistic explanation would require underlying forms of necessity and impossibility stronger that that pertaining to laws of nature generally -- such as either conceptual or ‘metaphysical’ necessity -- and those who see the link between explanans and explanandum as simply one of natural scientific law. David Chalmers' (1996) proposals for ‘naturalistic dualism’ (unlike those of the aforementioned naturalizers) put him in the second group. He argues that phenomenal consciousness in its various forms supervenes (not conceptually or metaphysically but only as a matter of nature's laws) on functional organization, and that this permits us to envisage (‘non-reductive’) ways of explaining consciousness by appeal to such organization.

Those who reject attempts to explain the phenomenal consciousness via a theory of intentionality still may reasonably proclaim allegiance to ‘naturalism.’ One may take phenomenal consciousness to be, in a sense, *psychologically* basic (if all that is mental is either phenomenally conscious or intentional, and no intentionalizing account of phenomenal character is feasible). But one might still hold that some non-intentional neuropsychological (or other, recognizably physicalist) explanation of the phenomenal character of experience is to be had, either because the explanatory link here exhibits an appropriately strong (conceptual or metaphysical) necessity, or because nothing stronger than psychophysical laws of nature are needed to give us the prospect of a natural scientific account of consciousness.

However, if we not only reject intentionalizing accounts of phenomenal character, but also see it as

inseparable from intentionality (if we reject both (a) and (b) and accept (c)), then whatever problems attach to physicalist explanations of consciousness will also infect prospects for explaining intentionality -- to some extent at least. And this will hold, even if we remain aloof from (d), and do not claim that phenomenal consciousness is essential to intentionality. For if we think that much of the intentionality we have in perceiving, imagining, and thinking is integral to the phenomenal character of such experience, then without a reductive explanation of that phenomenal character, our possession of the intentionality it brings with it will not be reductively explained either.

Finally, it should be noted that if one holds (d), this may have important consequences for what forms of psychological explanation one finds acceptable. For Searle and Ludwig argue that one's mental processes must have the right relationship to one's conscious experiences to count as one's mental processes at all. If they are right, postulated processes that do *not* bear this relation to our experiential lives cannot be going on in our minds.

A third broad area of concern on which our choice among (a)-(d) bears is epistemological. If one adopts (b), and something like a Sellarsian or Davidsonian distinction between sensation and thought, putting phenomenal character exclusively on the ‘sensation’ side, and intentionality exclusively on the ‘thought’ side of this divide, the place of consciousness in a philosophical account of knowledge will likely be meager -- at most phenomenal character will be a causal condition, without a role to play in the warrant or justification of claims to knowledge. However, if one takes route (a) or (c) the situation will appear rather different. If one either intentionalizes consciousness, or else views intentionality as inseparable from phenomenal character, there will then be more room to view consciousness as central to accounts of the warrant involved in first-person (‘introspective’) knowledge of mind, and empirical or perceptual knowledge. Though just how one goes about this, and with what success, will depend on how (if one chooses (a)) one intentionalizes consciousness, and (if one chooses (a) or (c)), that will depend on what sort of intentionality or content one thinks phenomenal consciousness brings with it. The place of consciousness in one's understanding of introspective or empirical knowledge will be rather different, depending on how one resolves the issues regarding: reflexivity; the conceptual/non-conceptual distinction; and externalism.

A fourth area of philosophical concern we may indicate broadly, closely bound to our conception of the relation of consciousness and intentionality, has to do with *value*. How intimately is consciousness bound up with those features of our own and others' lives that give them intrinsic or non-instrumental value for us? We may think that the pleasure and suffering that demand our ethical concern are necessarily phenomenally conscious -- and that this evaluative significance remains even if phenomenal character is non-intentional. However, the more intentionality is seen as inherent to the phenomenal character of experience, the more the latter will be bound to manifestations of intelligence, emotion, and understanding that appear to give human (and perhaps at least some other animal life) its special importance for us.

It may seem that those opting for (c) share at least this much ground with their intentionalizing opponents who go for (a): they both (unlike those who adopt (b)) are in a position to claim consciousness is crucial to whatever special moral regard we think appropriate only towards those whose psychologies involve a

kind of intentionality for which possession of painful or pleasant experience is not sufficient. However, this needs qualification on two counts. First, if one's embrace of (a) includes an intentionalizing strategy that limits phenomenal character to the sensory realm, one will limit the moral significance of phenomenal consciousness accordingly. Second, to those who hold (c), it may well seem their opponents' intentionalizing theories remove from view those very qualities of experience that make life worth living, and so they will hardly seem like allies on the issue of value. Further, if the proponent of (c) balks at going so far as to take on (d) -- conscious essentialism -- those who make that additional commitment may well wonder how those who do not could ultimately accord the possession of consciousness much greater non-instrumental value than the possession of a sophisticated but totally non-conscious mind.

From this survey it seems fair to conclude that working out a detailed view of the relation between consciousness and intentionality is hardly a peripheral matter philosophically. Potentially it has far-reaching consequences for one's views concerning these four important, broad topics:

- *The unity of mental phenomena* (Do consciousness or intentionality (or both together) somehow unify the domain of the psychological?)
- *The explanation of mental phenomena* (Can consciousness and intentionality be explained separately? Is explaining the one key to explaining the other?)
- *Introspective and empirical knowledge* (What relation to intentionality would give consciousness a central epistemological role in either?)
- *The value of human and other animal life*. (What relation of consciousness and intentionality (if any) underlies the non-instrumental value we accord ourselves and others?)

Bibliography

- Armstrong, D. 1968. *A Materialist Theory of Mind*. Routledge.
- Bermudez, J. 1998. *The Paradox of Self-Consciousness*. MIT Press.
- Block, N. 1995. “A Confusion About a Function of Consciousness,” *Behavioral and Brain Sciences* 18: 227-47.
- Brentano, F. [1874] (1973) *Psychology from an Empirical Standpoint*. Trans. T. Rancurello, D. Terrell, and L. McAllister. Humanities Press.
- Brentano, F. [1867] (1976) *The Psychology of Aristotle*. Trans. R. George. University of California Press.
- Burge, T. 1979. “Individualism and the Mental.” *Midwest Studies in Philosophy*, vol. 4. University of Minnesota Press.
- Carruthers, P. 1989. “Brute Experience.” *Journal of Philosophy* 86:259-269.
- Chalmers, D. 1996. *The Conscious Mind*. Oxford University Press.
- Chisholm, R. 1957. *Perceiving: a Philosophical Study*. Cornell University Press.
- Crane, T. 1992. “The Nonconceptual Contents of Experience.” In T. Crane, ed. *The Contents of Experience: Essays on Perception*. Cambridge University Press.
- Davidson, D. 1983. “A Coherence Theory of Truth and Knowledge.” In E. LePore, ed. *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Basil Blackwell.

- Davidson, D. 1986. “Empirical Content.” In E. LePore, ed. *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Basil Blackwell.
- Davies, M. 1997. “Externalism and Experience.” In N. Block, O. Flanagan, G. Guzeldere, eds. *The Nature of Consciousness*. MIT Press.
- Dennett, D.C. 1969. *Content and Consciousness*. Routledge.
- Dennett, D.C. 1978. Toward a Cognitive Theory of Consciousness. In *Brainstorms*. MIT Press.
- Dennett, D.C. 1991. *Consciousness Explained*. Little, Brown.
- Dretske. F. 1993. Conscious Experience. In *Mind* 102: 406, 262-283.
- Dretske, F. 1995. *Naturalizing the Mind*. MIT Press.
- Dreyfus, H. 1991. *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I*. MIT Press.
- Evans, G. 1982. *Varieties of Reference*. Oxford University Press.
- Flanagan, O. 1992. *Consciousness Reconsidered*. MIT Press.
- Fodor, J. 1987. *Psychosemantics*. MIT Press.
- Fodor, J. 1991. “A Modal Argument for Narrow Content.” *Journal of Philosophy* 88:1, 5-26.
- Follesdal, D. 1969. “Husserl's Notion of Noema.” *Journal of Philosophy* 66.
- Follesdal, D. 1990. “Noema and Meaning in Husserl.” *Philosophy and Phenomenological Research* 50.
- Frege, G. (1892) “On Sense and Reference.” Trans. M. Black. In *Translations from the Philosophical Writings of Gottlob Frege*. P. Geach and M. Black, eds. Basil Blackwell.
- Guzeldere, G. 1997. “Is Consciousness the Perception of What Passes in One's Own Mind?” In N. Block, O. Flanagan, G. Guzeldere, eds. *The Nature of Consciousness*. MIT Press.
- Heidegger, M. [1927] 1962. *Being and Time*. Trans. J. Macquarrie and E. Robinson. Harper & Row.
- Horgan, T. and Tienson, J. 2002. “The Intentionality ofPhenomenology and the Phenomenology of Intentionality.” In D. Chalmers, ed., *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press.
- Husserl, E. [1900] 1970. *Logical Investigations*. Trans. J. Findlay. Routledge.
- Husserl, E. [1913] 1983. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy, First Book*. Trans. F. Kersten. Kluwer.
- Husserl, E. [1928] 1966. *The Phenomenology of Internal Time Consciousness*. Trans. J. Churchill. Indiana University Press.
- Husserl, E. [1928] 1973. *Experience and Judgment*. Trans. J. Churchill and K. Ameriks. Northwestern University Press.
- Kant, I. [1787] 1929. *Critique of Pure Reason*. Trans. N.K. Smith. Macmillan.
- Kaplan, D. 1979. “On the Logic of Demonstratives.” *Journal of Philosophical Logic* 8.
- Kripke, S. 1972. “Naming and Necessity.” In G. Harman and D. Davidson, eds., *Semantics of Natural Language*, Reidel.
- Kelly, S. 2001. “The Non-conceptual Content of Perceptual Experience: Situation Dependence and Fineness of Grain.” *Philosophy and Phenomenological Research*. 63:3.
- Levine, J. 2001. *Purple Haze: the Puzzle of Consciousness*. Oxford University Press.
- Locke, J [1690] 1970. *An Essay Concerning Human Understanding*. Oxford University Press.
- Ludwig, K. 1996a. “On Explaining Why Things Look the Way They Do.” In K.Akins. ed.

Perception. Oxford University Press.
- Ludwig, K. 1996b. “Singular Thought and the Cartesian Theory of Mind.” *Nous* 30:4.
- Ludwig, K. 2001. “Phenomenal Consciousness and Intentionality: Comments on The Significance of Consciousness.” In *Psyche* 7.
- Lycan, W. 1996. *Consciousness and Experience*. MIT Press.
- McDowell, J. 1994. *Mind and World*. Harvard University Press.
- McGinn, C. 1991. *The Problems of Consciousness: Essays Toward a Resolution*. Basil Blackwell.
- Merleau-Ponty, M. [1949] 1962. *Phenomenology of Perception*. Trans C. Smith. Routledge & Kegan Paul.
- Peacocke, C. 1992. *A Study of Concepts*. MIT Press.
- Perry, J. Frege on Demonstratives. *Philosophical Review* 79:2.
- Perry, J. 2001. *Knowledge, Possiblity, and Consciousness*. MIT Press.
- Putnam, H. 1975. “The Meaning of ‘Meaning.’” In *Mind, Language, and Reality: Philosophical Papers*, Vol. 2. Cambridge University Press.
- Putnam, H 1981. *Reason, Truth and History*. Cambridge University Press.
- Rey, G. 1997. “A Question about Consciousness.” In N. Block, O. Flanagan, G. Guzeldere, eds., *The Nature of Consciousness*. MIT Press.
- Rorty, R. 1979. *Philosophy and the Mirror of Nature*. Princeton University Press.
- Rosenthal, D. 1986. “Two Concepts of Consciousness.” Reprinted in D. Rosenthal, ed., *The Nature of Mind*. Oxford University Press, 1991.
- Rosenthal, D. 1991. “The Independence of Consciousness and Sensory Quality.” In E. Villanueva, ed. *Philosophical Issues*, vol. 1, Consciousness. Ridgeview Press.
- Rosenthal, D. 1993. “Thinking that One Thinks.” In M. Davies and G.W. Humphreys, eds. *Consciousness: Psychological and Philosophical Essays*. Basil Blackwell.
- Russell, B. 1905. “On Denoting.” *Mind* 14.
- Ryle, G. 1949. *The Concept of Mind*. University of Chicago Press.
- Sartre, J-P. [1937] 1957. *Transcendence of the Ego*. Trans. F. Williams and R. Kirkpatrick. Farrar, Straus and Giroux.
- Sartre, J-P. [1943] 1953. *Being and Nothingness*. Trans. H. Barnes. Washington Square Press.
- Seager, W. 1999. *Theories of Consciousness*. Routledge.
- Searle, J. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.
- Searle, J. 1990. “Consciousness, Explanatory Inversion, and Cognitive Science.” *Behavioral and Brain Science* 13: 585-642.
- Searle, J. 1992 *The Rediscovery of Mind*. MIT Press.
- Searle, J. 1995. “Comments on Block's ‘A Confusion About a Function of Consciousness.’” *Behavioral and Brain Sciences* 18.
- Sellars, W. 1956. *Empiricism and the Philosophy of Mind.*
- Siewert, C. 1998. *The Significance of Consciousness*. Princeton University Press.
- Smith, D.W. 1986. “The Structure of (Self-)Consciousness.” *Topoi* 5 (2): 149-56.
- Strawson, G. 1994. *Mental Reality*. MIT Press.
- Thomasson, A. 2000. “After Brentano: A One-Level Theory of Consciousness.” *European Journal of Philosophy* 8: 2, 190-209.
- Thomasson, A. 1999. *Fiction and Metaphysics*. Cambridge Univesity Press.

- Tye, M. 1995. *Ten Problems of Consciousness*. MIT Press.
- Wittgenstein, L. 1953. *Philosophical Investigations*. Trans. By G.E.M. Anscombe. Macmillan.
- Zalta, E. 1988. *Intensional Logic and the Metaphysics of Intentionality*. MIT Press.
- Zahavi, D. 1998. "Brentano and Husserl on Self-Awareness." *Études Phenomenologiques*. 14 (27-28): 127-168.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

consciousness | consciousness: higher-order theories | consciousness: representational theories of | consciousness: unity of

Charles Siewert
csiewert@miami.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 22, 2002
Content last modified: June 22, 2002

Higher-order Theories of Consciousness

Higher-order theories of consciousness try to explain the distinctive properties of consciousness in terms of some relation obtaining between the conscious state in question and a higher-order representation of some sort (either a higher-order experience of that state, or a higher-order thought or belief about it). The most challenging properties to explain are those involved in *phenomenal* consciousness -- the sort of state which has a *subjective* dimension, which has 'feel', or which it is *like something* to undergo. These properties will form the focus of this article.

---

1. Kinds of Consciousness

One of the advances made in recent years has been in distinguishing between different questions concerning consciousness (see particularly: Rosenthal, 1986; Dretske, 1993; Block, 1995; Lycan, 1996). Not everyone agrees on quite *which* distinctions need to be drawn. But all are agreed that we should distinguish *creature* consciousness from *mental-state* consciousness. It is one thing to say *of an individual person or organism* that it is conscious (either in general or of something in particular); and it is quite another thing to say *of one of the mental states* of a creature that it is conscious.

It is also agreed that within creature-consciousness itself we should distinguish between *intransitive* and *transitive* variants. To say of an organism that it is conscious *simpliciter* (intransitive) is to say just that it is awake, as opposed to asleep or comatose. There do not appear to be any deep philosophical difficulties lurking here (or at least, they are not difficulties specific to the topic of consciousness, as opposed to

mentality in general). But to say of an organism that it is conscious *of such-and-such* (transitive) is normally to say at least that it is *perceiving* such-and-such, or *aware of* such-and-such. So we say of the mouse that it is conscious of the cat outside its hole, in explaining why it does not come out; meaning that it *perceives* the cat's presence. To provide an account of transitive creature-consciousness would thus be to attempt a theory of perception.

There is a choice to be made concerning transitive creature-consciousness, failure to notice which may be a potential source of confusion. For we have to decide whether the perceptual state in virtue of which an organism may be said to be transitively-conscious of something must itself be a conscious one (state-conscious -- see below). If we say 'Yes' then we shall need to know more about the mouse than merely that it perceives the cat if we are to be assured that it is conscious of the cat -- we shall need to establish that its percept of the cat is itself conscious. If we say 'No', on the other hand, then the mouse's perception of the cat will be sufficient for the mouse to count as conscious of the cat; but we may have to say that although it is conscious of the cat, the mental state in virtue of which it is so conscious is not itself a conscious one! It may be best to by-pass any danger of confusion here by avoiding the language of transitive-creature-consciousness altogether. Nothing of importance would be lost to us by doing this. We can say simply that organism O *observes* or *perceives* X; and we can then assert explicitly, if we wish, that its percept is or is not conscious.

Turning now to the notion of *mental-state consciousness*, the major distinction here is between *phenomenal consciousness*, on the one hand -- which is a property of states which it is *like something* to be in, which have a distinctive 'feel' (Nagel, 1974) -- and various functionally-definable forms of *access consciousness*, on the other (Block, 1995). Most theorists believe that there are mental states -- such as occurrent thoughts or judgments -- which are access-conscious (in whatever is the correct functionally-definable sense), but which are not phenomenally conscious. In contrast, there is considerable dispute as to whether mental states can be phenomenally-conscious without also being conscious in the functionally-definable sense -- and even more dispute about whether phenomenal consciousness can be *reductively explained* in functional and/or representational terms.

It seems plain that there is nothing deeply problematic about functionally-definable notions of mental-state consciousness, from a naturalistic perspective. For mental functions and mental representations are the staple fare of naturalistic accounts of the mind. But this leaves plenty of room for dispute about the form that the correct functional account should take. Some claim that for a state to be conscious in the relevant sense is for it to be poised to have an impact on the organism's decision-making processes (Kirk, 1994; Dretske, 1995; Tye, 1995), perhaps also with the additional requirement that those processes should be distinctively *rational* ones (Block, 1995). Others think that the relevant requirement for access-consciousness is that the state should be suitably related to higher-order representations -- experiences and/or beliefs -- of that very state (Armstrong, 1968, 1984; Rosenthal, 1986, 1993; Dennett, 1991; Carruthers, 1996, 2000; Lycan, 1996).

What *is* often thought to be naturalistically problematic, in contrast, is phenomenal consciousness (Nagel, 1974, 1984; Jackson, 1982, 1986; McGinn, 1991; Block, 1995; Chalmers, 1996). And what is really and deeply controversial is whether phenomenal consciousness can be *explained* in terms of some or other

functionally-definable notion. *Cognitive* (or *representational*) theories maintain that it can. *Higher-order* cognitive theories maintain that phenomenal consciousness can be reductively explained in terms of representations (either experiences or beliefs) which are higher-order. It is such theories which concern us here.

2. The Motivation for a Higher-Order Approach

Higher-order theories, like cognitive / representational theories in general, assume that the right *level* at which to seek an explanation of phenomenal consciousness is a cognitive one, providing an explanation in terms of some combination of *causal role* and *intentional content*. All such theories claim that phenomenal consciousness consists in a certain kind of intentional or representational content (*analog* or 'fine-grained' in comparison with any concepts we may possess) figuring in a certain distinctive position in the causal architecture of the mind. They must therefore maintain that these latter sorts of mental property do not already implicate or presuppose phenomenal consciousness. In fact, all cognitive accounts are united in rejecting the thesis that the very properties of *mind* or *mentality* already presuppose phenomenal consciousness, as proposed by Searle (1992, 1997) for example.

The major divide amongst representational theories of phenomenal consciousness in general, is between accounts which are provided in purely first-order terms and those which implicate higher-order representations of one sort or another (see below). These higher-order theorists will allow that first-order accounts -- of the sort defended by Dretske (1995) and Tye (1995), for example -- can already make some progress with the problem of consciousness. According to first-order views, phenomenal consciousness consists in analog or fine-grained contents which are available to the first-order processes which guide thought and action. So a phenomenally-conscious percept of red, for example, consists in a state with the analog content *red* which is tokened in such a way as to feed into thoughts about red, or into actions which are in one way or another guided by redness. Now, the point to note in favor of such an account is that it can explain the natural temptation to think that phenomenal consciousness is in some sense *ineffable*, or *indescribable*. This will be because such states have fine-grained contents which can slip through the mesh of any conceptual net. We can always distinguish many more shades of red than we have concepts for, or could describe in language (other than indexically -- e.g., '*That* shade').

The main motivation behind higher-order theories of consciousness, in contrast, derives from the belief that all (or at least most) mental-state types admit of both conscious and non-conscious varieties. Almost everyone now accepts, for example, (post-Freud) that beliefs and desires can be activated non-consciously. (Think, here, of the way in which problems can apparently become resolved during sleep, or while one's attention is directed to other tasks. Notice, too, that appeals to non-conscious intentional states are now routine in cognitive science.) And then if we ask what makes the difference between a conscious and a non-conscious mental state, one natural answer is that conscious states are states we are *aware of*. And if *awareness* is thought to be a form of creature-consciousness (see section 1 above), then this will translate into the view that conscious states are *states of which the subject is aware*, or states of which the subject is creature-conscious. That is to say, these are states which are the objects of some sort of higher-order representation -- whether a higher-order perception or experience, or a higher-order belief or

thought.

One crucial question, then, is whether perceptual states as well as beliefs admit of both conscious and non-conscious varieties. Can there be, for example, such a thing as a non-conscious visual perceptual state? Higher-order theorists are united in thinking that there can. Armstrong (1968) uses the example of absent-minded driving to make the point. Most of us at some time have had the rather unnerving experience of ‘coming to’ after having been driving on ‘automatic pilot’ while our attention was directed elsewhere -- perhaps having been day-dreaming or engaged in intense conversation with a passenger. We were apparently not consciously aware of any of the route we have recently taken, nor of any of the obstacles we avoided on the way. Yet we must surely have been *seeing*, or we would have crashed the car. Others have used the example of blindsight (Carruthers, 1989, 1996). This is a condition in which subjects have had a portion of their primary visual cortex destroyed, and apparently become blind in a region of their visual field as a result. But it has now been known for some time that if subjects are asked to *guess* at the properties of their ‘blind’ field (e.g. whether it contains a horizontal or vertical grating, or whether it contains an ‘X’ or an ‘O’), they prove remarkably accurate. Subjects can also reach out and grasp objects in their ‘blind’ field with something like 80% or more of normal accuracy, and can catch a ball thrown from their ‘blind’ side, all without conscious awareness. (See Weiskrantz, 1986, 1997, for details and discussion.)

More recently, a powerful case for the existence of non-conscious visual experience has been generated by the *two-systems theory* of vision proposed and defended by Milner and Goodale (1995). They review a wide variety of kinds of neurological and neuro-psychological evidence for the substantial independence of two distinct visual systems, instantiated in the temporal and parietal lobes respectively. They conclude that the parietal lobes provide a set of specialized semi-independent modules for the on-line visual control of action; whereas the temporal lobes are primarily concerned with more off-line functions such as visual learning and object recognition. And only the experiences generated by the temporal-lobe system are phenomenally conscious, on their account.

(Note that this is not the familiar distinction between *what* and *where* visual systems, but is rather a successor to it. For the temporal-lobe system is supposed to have access both to property information and to spatial information. Instead, it is a distinction between a combined *what-where* system located in the temporal lobes and a *how-to* or action-guiding system located in the parietal lobes.)

To get the flavor of Milner and Goodale's hypothesis, consider just one strand from the wealth of evidence they provide. This is a neurological syndrome called *visual form agnosia*, which results from damage localized to both temporal lobes, leaving primary visual cortex and the parietal lobes intact. (Visual form agnosia is normally caused by carbon monoxide poisoning, for reasons which are little understood.) Such patients cannot recognize objects or shapes, and may be capable of little conscious visual experience; but their sensorimotor abilities remain largely intact.

One particular patient -- D.F. -- has now been examined in considerable detail. While D.F. is severely agnosic, she is not completely lacking in conscious visual experience. Her capacities to perceive colors and textures are almost completely preserved. (Why just these sub-modules in her temporal cortex should

have been spared is not known.) As a result, she can sometimes guess the identity of a presented object -- recognizing a banana, say, from its yellow color and the distinctive texture of its surface. But she is unable to perceive the shape of the banana (whether straight or curved, say); nor its orientation (upright or horizontal; pointing towards her or across). Yet many of her sensorimotor abilities are close to normal -- she would be able to reach out and grasp the banana, orienting her hand and wrist appropriately for its position and orientation, and using a normal and appropriate finger grip. Under experimental conditions it turns out that although D.F. is at chance in identifying the orientation of a broad line or letter-box, she is almost normal when posting a letter through a similarly-shaped slot oriented at random angles. In the same way, although she is at chance when trying to discriminate between rectangular blocks of very different sizes, her reaching and grasping behaviors when asked to pick up such a block are virtually indistinguishable from those of normal controls. It is very hard to make sense of this data without supposing that the sensorimotor perceptual system is functionally and anatomically distinct from the object-recognition/conscious system.

There is a powerful case, then, for thinking that there are non-conscious as well as conscious visual percepts. While the perceptions which ground your thoughts when you plan in relation to the perceived environment (‘I'll pick up *that* one’) may be conscious, and while you will continue to enjoy conscious perceptions of what you are doing while you act, the perceptual states which actually guide the details of your movements when you reach out and grab the object will *not* be conscious ones, if Milner and Goodale (1995) are correct.

But what implications does this have for phenomenal consciousness? Must these non-conscious percepts also be lacking in *phenomenal* properties? Most people think so. While it may be possible to get oneself to believe that the perceptions of the absent-minded car driver can remain phenomenally conscious (perhaps lying outside of the focus of attention, or being instantly forgotten), it is very hard to believe that either blindsight percepts or D.F.'s sensorimotor perceptual states might be phenomenally conscious ones. For these perceptions are ones to which the subjects of those states are *blind*, and of which they cannot be a*ware*. And the question, then, is what makes the relevant difference? What is it about a conscious perception which renders it *phenomenal*, which a blindsight perceptual state would correspondingly lack? Higher-order theorists are united in thinking that the relevant difference consists in the presence of something *higher-order* in the first case which is absent in the second. The core intuition is that a phenomenally conscious state will be a state *of which the subject is aware*.

What options does a first-order theorist have to resist this conclusion? One is to deny the data (as does Dretske, 1995). It can be said that the non-conscious states in question lack the kind of fineness of grain and richness of content necessary to count as genuinely *perceptual* states. On this view, the contrast discussed above isn't really a difference between conscious and non-conscious perceptions, but rather between conscious perceptions, on the one hand, and non-conscious belief-like states, on the other. Another option is to accept the distinction between conscious and non-conscious perceptions, and then to explain that distinction in first-order terms. It might be said, for example, that conscious perceptions are those which are available to *belief* and *thought*, whereas non-conscious ones are those which are available to guide *movement* (Kirk, 1994). A final option is to bite the bullet, and insist that blindsight and sensorimotor perceptual states are indeed phenomenally conscious while not being *access*-conscious. (See

Block, 1995; Tye, 1995; and Nelkin, 1996; all of whom defend versions of this view.) On this account, blindsight percepts are phenomenally conscious states to which the subjects of those states are *blind*. Higher-order theorists will argue, of course, that none of these alternatives is acceptable (see, e.g., Carruthers, 2000).

In general, then, higher-order theories of phenomenal consciousness claim the following:

> **Higher Order Theory (In General)**:
> A phenomenally conscious mental state is a mental state (of a certain sort -- see below) which either is, or is disposed to be, the object of a higher-order representation of a certain sort (see below).

Higher-order theorists will allow, of course, that mental states can be targets of higher-order representation without being phenomenally conscious. For example, a belief can give rise to a higher-order belief without thereby being phenomenally conscious. What is distinctive of phenomenal consciousness is that the states in question should be perceptual or quasi-perceptual ones (e.g. visual images as well as visual percepts). Moreover, most cognitive/representational theorists will maintain that these states must possess a certain kind of analog (fine-grained) or non-conceptual intentional content. What makes perceptual states, mental images, bodily sensations, and emotional feelings phenomenally conscious, on this approach, is that they are conscious states with analog or non-conceptual contents. So putting these points together, we get the view that phenomenally conscious states are those states *which possess fine-grained intentional contents of which the subject is aware*, being the target or potential target of some sort of higher-order representation.

There are then two main dimensions along which higher-order theorists disagree amongst themselves. One concerns whether the higher-order states in question are belief-like or perception-like. Those taking the former option are higher-order *thought* theorists, and those taking the latter are higher-order *experience* or ‘inner-sense’ theorists. The other disagreement is internal to higher-order thought approaches, and concerns whether the relevant relation between the first-order state and the higher-order thought is one of *availability* or not. That is, the question is whether a state is conscious by virtue of being *disposed* to give rise to a higher-order thought, or rather by virtue of being the *actual target* of such a thought. These are the options which will now concern us.

3. Inner-Sense Theory

According to this view, humans not only have first-order non-conceptual and/or analog perceptions of states of their environments and bodies, they also have second-order non-conceptual and/or analog perceptions of their first-order states of perception. Humans (and perhaps other animals) not only have sense-organs which scan the environment/body to produce fine-grained representations which can then serve to ground thoughts and action-planning, but they also have *inner* senses, charged with scanning the outputs of the first-order senses (i.e. perceptual experiences) to produce equally fine-grained, but higher-order, representations of those outputs (i.e. to produce higher-order experiences). A version of this view

was first proposed by the British Empiricist philosopher John Locke (1690). In our own time it has been defended especially by Armstrong (1968, 1984) and by Lycan (1996).

(A terminological point: this view is sometimes called a ‘higher-order experience (HOE) theory’ of phenomenal consciousness; but the term ‘inner-sense theory’ is more accurate. For as we shall see in section 5, there are versions of a higher-order thought (HOT) approach which also implicate higher-order perceptions, but without needing to appeal to any organs of inner sense.)

(Another terminological point: ‘*inner*-sense theory’ should more strictly be called ‘*higher-order*-sense theory’, since we of course have senses which are physically ‘inner’, such as pain-perception and internal touch-perception, which are not intended to fall under its scope. For these are first-order senses on a par with vision and hearing, differing only in that their purpose is to detect properties of the body rather than of the external world. According to the sort of higher-order theory under discussion in this section, these senses, too, will need to have their outputs scanned to produce higher-order analog contents in order for them to become phenomenally conscious. In what follows, however, the term ‘inner sense’ will be used to mean, more strictly, ‘higher-order sense’, since this terminology is now pretty firmly established.)

We therefore have the following proposal to consider:

> **Inner-Sense Theory**:
> A phenomenally conscious mental state is a state with analog/non-conceptual intentional content, which is in turn the target of a higher-order analog/non-conceptual intentional state, via the operations of a faculty of ‘inner sense’.

On this account, the difference between a phenomenally conscious percept of red and the sort of non-conscious percepts of red which guide the guesses of a blindsighter and the activity of sensorimotor system, is as follows. The former is scanned by our inner senses to produce a higher-order analog state with the content *experience of red* or *seems red*, whereas the latter states are not -- they remain *merely* first-order states with the analog content *red*; and in so remaining, they lack any dimension of *seeming* or *subjectivity*. According to inner-sense theory, it is our higher-order experiential contents produced by the operations of our inner-senses which make some mental states with analog contents, but not others, available to their subjects. And it is these same higher-order contents which constitute the subjective dimension or ‘feel’ of the former set of states, thus rendering them phenomenally conscious.

One of the main advantages of inner-sense theory is that it can explain how it is possible for us to acquire *purely recognitional concepts* of experience. For if we possess higher-order perceptual contents, then it should be possible for us to learn to recognize the occurrence of our own perceptual states immediately -- or ‘straight off’ -- grounded in those higher-order analog contents. And this should be possible without those recognitional concepts thereby having any conceptual connections with our beliefs about the nature or content of the states recognized, nor with any of our surrounding mental concepts. This is then how inner-sense theory will claim to explain the familiar philosophical thought-experiments concerning one's own experiences, which are supposed to cause such problems for physicalist/naturalistic accounts of the mind (Kripke, 1972; Chalmers, 1996).

For example, I can think, '*This* type of experience [as of red] might have occurred in me, or might normally occur in others, in the absence of any of its actual causes and effects.' So on any view of intentional content which sees content as tied to normal causes (i.e. to information carried) and/or to normal effects (i.e. to teleological or inferential role), *this* type of experience might occur without representing *red*. In the same sort of way, I shall be able to think, '*This* type of experience [pain] might have occurred in me, or might occur in others, in the absence of any of the usual causes and effects of pains. There could be someone in whom *these* experiences occur but who isn't bothered by them, and where those experiences are never caused by tissue damage or other forms of bodily insult. And conversely, there could be someone who behaves and acts just as I do when in pain, and in response to the same physical causes, but who is never subject to *this* type of experience.' If we possess purely recognitional concepts of experience, grounded in higher-order percepts of those experiences, then the thinkability of such thoughts is both readily explicable, and apparently unthreatening to a naturalistic approach to the mind.

Inner-sense theory does face a number of difficulties, however. One objection is as follows (see Dretske, 1995). If inner-sense theory were true, then how is it that there is no phenomenology distinctive of inner sense, in the way that there is a phenomenology associated with each outer sense? Since each of the outer senses gives rise to a distinctive set of phenomenological properties, you might expect that if there *were* such a thing as inner sense, then there would also be a phenomenology distinctive of its operation. But there doesn't appear to be any.

This point turns on the so-called 'transparency' of our perceptual experience (Harman, 1990). Concentrate as hard as you like on your 'outer' (first-order) experiences -- you will not find any *further* phenomenological properties arising out of the attention you pay to them, beyond those already belonging to the contents of the experiences themselves. Paying close attention to your experience of the color of the red rose, for example, just produces attention to the *redness* -- a property of the rose. But put like this, however, the objection just seems to beg the question in favor of first-order theories of phenomenal consciousness. It assumes that first-order -- 'outer' -- perceptions already have a phenomenology independently of their targeting by inner sense. But this is just what an inner-sense theorist will deny. And then in order to explain the absence of any kind of higher-order phenomenology, an inner-sense theorist only needs to maintain that our higher-order experiences are never themselves targeted by an inner-sense-organ which might produce *third*-order analog representations of them in turn.

Another objection to inner-sense theory is as follows (see Sturgeon, 2000). If there really were an organ of inner sense, then it ought to be possible for it to malfunction, just as our first-order senses sometimes do. And in that case, it ought to be possible for someone to have a first-order percept with the analog content *red* causing a higher-order percept with the analog content *seems-orange*. Someone in this situation would be disposed to judge, 'It's red', immediately and non-inferentially (i.e. not influenced by beliefs about the object's normal color or their own physical state). But at the same time they would be disposed to judge, 'It *seems* orange'. Not only does this sort of thing never apparently occur, but the idea that it might do so conflicts with a powerful intuition. This is that our awareness of our own experiences is *immediate*, in such a way that to *believe* that you are undergoing an experience of a certain sort *is* to be

undergoing an experience of that sort. But if inner-sense theory is correct, then it ought to be possible for someone to believe that they are in a state of *seeming-orange* when they are actually in a state of *seeming-red*.

A different sort of objection to inner-sense theory is developed by Carruthers (2000). It starts from the fact that the internal monitors postulated by such theories would need to have considerable computational complexity in order to generate the requisite higher-order experiences. In order to perceive an experience, the organism would need to have mechanisms to generate a set of internal representations with an analog or non-conceptual content representing the content of that experience, in all its richness and fine-grained detail. And notice that any inner scanner would have to be a physical device (just as the visual system itself is) which depends upon the detection of those *physical* events in the brain which are the outputs of the various sensory systems (just as the visual system is a physical device which depends upon detection of physical properties of surfaces via the reflection of light). For it is hard to see how any inner scanner could detect the presence of an experience *qua* experience. Rather, it would have to detect the physical *realizations* of experiences in the brain, and construct the requisite higher-order representation of the experiences which those physical events realize, on the basis of that physical-information input. This makes is seem inevitable that the scanning device which supposedly generates higher-order experiences of our first-order visual experience would have to be almost as sophisticated and complex as the visual system itself.

Now the problem which arises here is this. Given this complexity in the operations of our organs of inner sense, there had better be some plausible story to tell about the evolutionary pressures which led to their construction. For natural selection is the only theory which can explain the existence of organized functional complexity in nature (Pinker, 1994, 1997). But there would seem to be no such stories on the market. The most plausible suggestion is that inner-sense might have evolved to subserve our capacity to think about the mental states of conspecifics, thus enabling us to predict their actions and manipulate their responses. (This is the so-called 'Machiavellian hypothesis' to explain the evolution of intelligence in the great-ape lineage. See Byrne and Whiten, 1988, 1998.) But this suggestion presupposes that the organism must *already* have some capacity for higher-order *thought*, since it is such thoughts which inner sense is supposed to subserve. And yet as we shall see shortly (in section 5), some higher-order thought theories can claim all of the advantages of inner-sense theory as an explanation of phenomenal consciousness, but without the need to postulate any 'inner scanners'. At any rate, the 'computational complexity objection' to inner-sense theories remains as a challenge to be answered.

4. Higher-Order Thought Theory (1): Non-dispositionalist

Non-dispositionalist higher-order thought (HOT) theory is a proposal about the nature of state-consciousness in general, of which phenomenal consciousness is but one species. Its main proponent has been Rosenthal (1986, 1993, forthcoming). The proposal is this: a conscious mental state M, of mine, is a state which is actually causing an activated belief (generally a non-conscious one) that I have M, and

causing it non-inferentially. (The qualification concerning non-inferential causation is included to avoid one having to say that my non-conscious motives become conscious when I learn of them under psycho-analysis, or that my jealousy is conscious when I learn of it by interpreting my own behavior.) An account of phenomenal consciousness can then be generated by stipulating that the mental state M should have an analog content in order to count as an experience, and that when M is an experience (or a mental image, bodily sensation, or emotional feeling), it will be phenomenally conscious when (and only when) suitably targeted.

We therefore have the following proposal to consider:

> **Non-Dispositionalist Higher-Order Thought Theory**:
> A phenomenally conscious mental state is a state with analog/non-conceptual intentional content, which is the object of a higher-order thought, and which causes that thought non-inferentially.

This account avoids some of the difficulties inherent in inner-sense theory, while retaining the latter's ability to explain the distinction between conscious and non-conscious perceptions. (Conscious perceptions will be analog states which are targeted by a higher-order thought, whereas perceptions such as those involved in blindsight will be non-conscious by virtue of *not* being so targeted.) In particular, it is easy to see a function for higher-order thoughts, in general, and to tell a story about their likely evolution. A capacity to entertain higher-order thoughts about experiences would enable a creature to negotiate the is/seems distinction, perhaps learning not to trust its own experiences in certain circumstances, and also to induce appearances in others, by deceit. And a capacity to entertain higher-order thoughts about thoughts (beliefs and desires) would enable a creature to reflect on, and to alter, its own beliefs and patterns of reasoning, as well as to predict and manipulate the thoughts and behaviors of others. Indeed, it can plausibly be claimed that it is our capacity to target higher-order thoughts on our own mental states which underlies our status as rational agents (Burge, 1996; Sperber, 1996).

One well-known objection to this sort of higher-order thought theory is due to Dretske (1993). We are asked to imagine a case in which we carefully examine two line-drawings, say (or in Dretske's example, two patterns of differently-sized spots). These drawings are similar in almost all respects, but differ in just one aspect -- in Dretske's example, one of the pictures contains a black spot which the other lacks. It is surely plausible that, in the course of examining these two pictures, one will have enjoyed a conscious visual experience of the respect in which they differ -- e.g. of the offending spot. But, as is familiar, one can be in this position while not knowing *that* the two pictures are different, or in what *way* they are different. In which case, since one can have a conscious experience (e.g. of the spot) without being aware that one is having it, consciousness cannot require higher-order awareness.

Replies to this objection have been made by Seager (1994) and by Byrne (1997). They point out that it is one thing to have a conscious experience of the aspect which differentiates the two pictures, and quite another to consciously experience that the two pictures are differentiated by that aspect. That is, seeing the extra spot in one picture needn't mean seeing that this is the difference between the two pictures. So while scanning the two pictures one will enjoy conscious experience of the extra spot. A higher-order

thought theorist will say that this means undergoing a percept with the content *spot here* which forms the target of a higher-order belief that one is undergoing a perception with that content. But this can perfectly well be true without one undergoing a percept with the content *spot here in this picture but absent here in that one*. And it can also be true without one forming any higher-order belief to the effect that one is undergoing a perception with the content *spot here* when looking at a given picture but not when looking at the other. In which case the purported counter-example isn't really a counter-example.

A different sort of problem with the non-dispositionalist version of higher-order thought theory relates to the huge number of beliefs which would have to be caused by any given phenomenally conscious experience. (This is the analogue of the ‘computational complexity’ objection to inner-sense theory, sketched in section 3 above). Consider just how rich and detailed a conscious experience can be. It would seem that there can be an immense amount of which we can be consciously aware at any one time. Imagine looking down on a city from a window high up in a tower-block, for example. In such a case you can have phenomenally conscious percepts of a complex distribution of trees, roads, and buildings; colors on the ground and in the sky above; moving cars and pedestrians; and so on. And you can -- it seems -- be conscious of all of this simultaneously. According to non-dispositionalist higher-order thought theory, then, you would need to have a distinct activated higher-order belief for each distinct aspect of your experience -- either that, or just a few such beliefs with immensely complex contents. Either way, the objection is the same. For it seems implausible that all of this higher-order activity should be taking place (albeit non-consciously) every time someone is the subject of a complex conscious experience. For what would be the point? And think of the amount of cognitive space that these beliefs would take up!

This objection to non-dispositionalist forms of higher-order thought theory is considered at some length in Carruthers (2000), where a variety of possible replies are discussed and evaluated. Perhaps the most plausible and challenging such reply would be to deny the main premise lying behind the objection, concerning the rich and integrated nature of phenomenally conscious experience. Rather, the theory could align itself with Dennett's (1991) conception of consciousness as highly fragmented, with multiple streams of perceptual content being processed in parallel in different regions of the brain, and with no stage at which all of these contents are routinely integrated into a phenomenally conscious perceptual manifold. Rather, contents become conscious on a piecemeal basis, as a result of internal or external *probing* which gives rise to a higher-order belief about the content in question. (Dennett himself sees this process as essentially linguistic, with both probes and higher-order thoughts being formulated in natural language. This variant of the view, although important in its own right, is not relevant to our present concerns.) This serves to convey to us the mere *illusion* of riches, because wherever we direct our attention, there we find a conscious perceptual content.

It is doubtful whether this sort of ‘fragmentist’ account can really explain the phenomenology of our experience, however. For it still faces the objection that the objects of attention can be immensely rich and varied at any given moment, hence requiring there to be an equally rich and varied repertoire of higher-order thoughts tokened at the same time. Think of immersing yourself in the colors and textures of a Van Gogh painting, for example, or the scene as you look out at your garden -- it would seem that one can be phenomenally conscious of a *highly* complex set of properties, which one could not even begin to describe or conceptualize in any detail. However, since the issues here are large and controversial, it

cannot yet be concluded that non-dispositionalist forms of higher-order thought theory have been decisively refuted.

5. Higher-Order Thought Theory (2): Dispositionalist

According to all forms of dispositionalist higher-order thought theory, the conscious status of an experience consists in its *availability* to higher-order thought (Dennett, 1978; Carruthers, 1996, 2000). As with the non-dispositionalist version of the theory, in its simplest form we have here a quite general proposal concerning the conscious status of any type of occurrent mental state, which becomes an account of phenomenal consciousness when the states in question are experiences (or images, emotions, etc.) with analog content. The proposal is this: a conscious mental event M, of mine, is one which is disposed to cause an activated belief (generally a non-conscious one) that I have M, and to cause it non-inferentially.

The proposal before us is therefore as follows:

> **Dispositionalist Higher-Order Thought Theory**:
> A phenomenally conscious mental state is a state with analog/non-conceptual intentional content, which is held in a special-purpose short-term memory store in such a way as to be available to cause (non-inferentially) higher-order thoughts about any of the contents of that store.

In contrast with the non-dispositionalist form of theory, the higher-order thoughts which render a percept conscious are not necessarily actual, but potential, on this account. So the objection now disappears, that an unbelievable amount of cognitive space would have to be taken up with every conscious experience. (There need not *actually* be *any* higher-order thought occurring, in order for a given perceptual state to count as phenomenally conscious, on this view.) So we can retain our belief in the rich and integrated nature of phenomenally conscious experience -- we just have to suppose that all of the contents in question are simultaneously *available* to higher-order thought. Nor will there be any problem in explaining why our faculty of higher-order thought should have evolved, nor why it should have access to perceptual contents in the first place -- this can be the standard sort of story in terms of Machiavellian intelligence.

It might well be wondered how their mere *availability* to higher-order thoughts could confer on our perceptual states the positive properties distinctive of phenomenal consciousness -- that is, of states having a *subjective* dimension, or a distinctive subjective *feel*. The answer may lie in the theory of content. Suppose that one agrees with Millikan (1984) that the representational content of a state depends, in part, upon the powers of the systems which *consume* that state. That is, suppose one thinks that *what* a state represents will depend, in part, on the kinds of inferences which the cognitive system is prepared to make in the presence of that state, or on the kinds of behavioral control which it can exert. In which case the presence of first-order perceptual representations to a consumer-system which can deploy a ‘theory of mind’, and which is capable of recognitional applications of theoretically-embedded concepts of experience, may be sufficient to render those representations *at the same time* as higher-order ones. This

would be what confers on our phenomenally conscious experiences the dimension of subjectivity. Each experience would at the same time (while also representing some state of the world, or of our own bodies) be a representation that we are undergoing just such an experience, by virtue of the powers of the 'theory of mind' consumer-system. Each percept of green, for example, would at one and the same time be an analog representation of *green* and an analog representation of *seems green* or *experience of green*. In fact, the attachment of a 'theory of mind' faculty to our perceptual systems may completely transform the contents of the latter's outputs.

(Consumer semantics embraces not only a number of different varieties of *teleosemantics*, but also various forms of *inferential role semantics*. For the former, see Millikan, 1984, 1986, 1989; and Papineau, 1987, 1993. For the latter, see Loar, 1981, 1982; McGinn, 1982, 1989; Block, 1986; and Peacocke, 1986, 1992.)

This account might seem to achieve all of the benefits of inner-sense theory, but without the associated costs. (Some potential draw-backs will be noted in a moment.) In particular, we can endorse the claim that phenomenal consciousness consists in a set of higher-order perceptions. This enables us to explain, not only the difference between conscious and non-conscious perception, but also how analog states come to acquire a subjective dimension or 'feel'. And we can also explain how it can be possible for us to acquire some purely recognitional concepts of experience (thus explaining the standard philosophical thought-experiments). But we don't have to appeal to the existence of any 'inner scanners' or organs of inner sense (together with their associated problems) in order to do this. Moreover, it should also be obvious why there can be no question of our higher-order contents getting out of line with their first-order counterparts, in such a way that one might be disposed to make recognitional judgments of *red* and *seems orange* at the same time. This is because the content of the higher-order experience is parasitic on the content of the first-order one, being formed from it by virtue of the latter's availability to a 'theory of mind' system.

On the down-side, the account isn't neutral on questions of semantic theory. On the contrary, it requires us to reject any form of pure input-semantics, in favor of some sort of consumer-semantics. We cannot then accept that intentional content reduces to informational content, nor that it can be explicated purely in terms of causal co-variance relations to the environment. So anyone who finds such views attractive will think that the account is a hard one to swallow. (For discussion of various different versions of input-semantics, see Dretske, 1981, 1986; Fodor, 1987, 1990; and Loewer and Rey, 1991.)

What will no doubt be seen by most people as the *biggest* difficulty with dispositionalist higher-order thought theory, however, is that it may have to deny phenomenal consciousness to most species of non-human animal. This objection will be discussed, among others, in the section following, since it can arguably also be raised against *any* form of higher-order theory.

6. Objections to a Higher-Order Approach

There have been a whole host of objections raised against higher-order theories of phenomenal

consciousness. (See, e.g., Aquila, 1990; Jamieson and Bekoff, 1992; Dretske, 1993, 1995; Goldman, 1993; Güzeldere, 1995; Tye, 1995; Chalmers, 1996; Byrne, 1997; Siewert, 1998.) Unfortunately, many of these objections, although perhaps intended as objections to higher-order theories as such, are often framed in terms of one or another particular version of such a theory. One general moral to be taken away from the present discussion should then be this: the different versions of a higher-order theory of phenomenal consciousness need to be kept distinct from one another, and critics should take care to state which version of the approach is under attack, or to frame objections which turn merely on the *higher-order character* of all of these approaches.

One generic objection is that higher-order theories, when combined with plausible empirical claims about the representational powers of non-human animals, will conflict with our common-sense intuition that such animals enjoy phenomenally conscious experience (Jamieson and Bekoff, 1992; Dretske, 1995; Tye, 1995). This objection can be pressed most forcefully against higher-order *thought* theories, of either variety; but it is also faced by inner-sense theory (depending on what account can be offered of the evolutionary function of organs of inner sense). Since there is considerable dispute as to whether even chimpanzees have the kind of sophisticated 'theory of mind' which would enable them to entertain thoughts about experiential states as such (Byrne and Whiten, 1988, 1998; Povinelli, 2000), it seems most implausible that many other species of mammal (let alone reptiles, birds and fish) would qualify as phenomenally conscious, on these accounts. Yet the intuition that such creatures enjoy phenomenally conscious experiences is a powerful and deep-seated one, for many people. (Witness Nagel's classic 1974 paper, which argues that there must be something which it is like to be a bat.)

The grounds for this common-sense intuition can be challenged, however. (How, after all, are we supposed to *know* whether it is like something to be a bat?) And that intuition can perhaps be explained away as a mere by-product of imaginative identification with the animal. (Since our *images* of their experiences are phenomenally conscious, we may naturally assume that the experiences *imaged* are similarly conscious. See Carruthers, 1999, 2000.) But there is no doubt that one crux of resistance to higher-order theories will lie here, for many people.

Another generic objection is that higher-order approaches cannot really *explain* the distinctive properties of phenomenal consciousness (Chalmers, 1996; Siewert, 1998). Whereas the argument from animals is that higher-order representations aren't *necessary* for phenomenal consciousness, the argument here is that such representations aren't *sufficient*. It is claimed, for example, that we can easily conceive of creatures who enjoy the postulated kinds of higher-order representation, related in the right sort of way to their first-order perceptual states, but where those creatures are wholly *lacking* in phenomenal consciousness.

In response to this objection, higher-order theorists will join forces with first-order theorists and others in claiming that these objectors pitch the standards for explaining phenomenal consciousness too high (Block and Stalnaker, 1999; Tye, 1999; Carruthers, 2000; Lycan, 2001). We will insist that a reductive explanation of something -- and of phenomenal consciousness in particular -- doesn't have to be such that we cannot conceive of the *explanandum* (that which is being explained) in the absence of the *explanans* (that which does the explaining). Rather, we just need to have good reason to think that the explained

properties are *constituted by* the explaining ones, in such a way that nothing *else* needed to be added to the world once the explaining properties were present, in order for the world to contain the target phenomenon. But this is hotly contested territory. And it is on this ground that the battle for phenomenal consciousness may ultimately be won or lost.

Bibliography

- Aquila, R. 1990. Consciousness as higher-order thoughts: two objections. *American Philosophical Quarterly*, 27.
- Armstrong, D. 1968. *A Materialist Theory of the Mind*. Routledge.
- Armstrong, D. 1984. Consciousness and causality. In D. Armstrong and N. Malcolm, *Consciousness and Causality*, Blackwell.
- Block, N. 1986. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10.
- Block, N. 1995. A confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18.
- Block, N. and Stalnaker, R. 1999. Conceptual analysis, dualism and the explanatory gap. *Philosophical Review*, 108.
- Burge, T. 1996. Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society*, 96.
- Byrne, A. 1997. Some like it HOT: consciousness and higher-order thoughts. *Philosophical Studies*, 86.
- Byrne, R. and Whiten, A. eds. 1988. *Machiavellian Intelligence*. Oxford University Press.
- Byrne, R. and Whiten, A. eds. 1998. *Machiavellian Intelligence II: Evaluations and extensions*. Cambridge University Press.
- Carruthers, P. 1989. Brute experience. *Journal of Philosophy*, 86.
- Carruthers, P. 1996. *Language, Thought and Consciousness*. Cambridge University Press.
- Carruthers, P. 1999. Sympathy and subjectivity. *Australasian Journal of Philosophy*, 77.
- Carruthers, P. 2000. *Phenomenal Consciousness: a naturalistic theory*. Cambridge University Press.
- Chalmers, D. 1996. *The Conscious Mind*. Oxford University Press.
- Dennett, D. 1978. Toward a cognitive theory of consciousness. In C. Savage ed., *Minnesota Studies in the Philosophy of Science*, 9. (Reprinted in Dennett's *Brainstorms*, MIT Press, 1978.)
- Dennett, D. 1991. *Consciousness Explained*. Allen Lane.
- Dretske, F. 1981. *Knowledge and the Flow of Information*. MIT Press.
- Dretske, F. 1986. Misrepresentation. In R. Bogdan ed., *Belief*, Oxford University Press.
- Dretske, F. 1988. *Explaining Behavior* MIT Press.
- Dretske, F. 1993. Conscious experience. *Mind*, 102.
- Dretske, F. 1995. *Naturalizing the Mind*. MIT Press.
- Fodor, J. 1987. *Psychosemantics*. MIT Press.
- Fodor, J. 1990. *A Theory of Content and Other Essays*. MIT Press.
- Goldman, A. 1993. Consciousness, folk-psychology, and cognitive science. *Consciousness and Cognition*, 2.
- Güzeldere, G. 1995. Is consciousness perception of what passes in one's own mind? In T.

Metzinger, ed., *Conscious Experience*, Ferdinand Schoningh.

- Harman, G. 1990. The intrinsic quality of experience. *Philosophical Perspectives*, 4.
- Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly*, 32.
- Jackson, F. 1986. What Mary didn't know. *Journal of Philosophy*, 83.
- Jamieson, D. and Bekoff, M. 1992. Carruthers on non-conscious experience. *Analysis*, 52.
- Kirk, R. 1994. *Raw Feeling*. Oxford University Press.
- Kripke, S. 1972. Naming and necessity. In G. Harman and D. Davidson, eds., *Semantics of Natural Language*, Reidel. (Revised version printed in book form by Blackwell, 1980.)
- Locke, J. 1690. *An Essay Concerning Human Understanding*.
- Loewer, B. and Rey, G. eds. 1991. *Meaning in Mind: Fodor and his critics*. Blackwell.
- Lycan, W. 1996. *Consciousness and Experience*. MIT Press.
- Lycan, W. 2001. Have we neglected phenomenal consciousness? *Psyche*, 7.
- McGinn, C. 1982. The structure of content. In A. Woodfield, ed., *Thought and Object*, Oxford University Press.
- McGinn, C. 1989. *Mental Content*. Blackwell.
- McGinn, C. 1991. *The Problem of Consciousness*. Blackwell.
- Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. MIT Press.
- Millikan, R. 1986. Thoughts without laws: cognitive science with content. *Philosophical Review*, 95.
- Millikan, R. 1989. Biosemantics. *Journal of Philosophy*, 86.
- Milner, D. and Goodale, M. 1995. *The Visual Brain in Action*. Oxford University Press.
- Nagel, T. 1974. What is it like to be a bat? *Philosophical Review*, 83.
- Nagel, T. 1986. *The View from Nowhere*. Oxford University Press.
- Nelkin, N. 1996. *Consciousness and the Origins of Thought*. Cambridge University Press.
- Papineau, D. 1987. *Reality and Representation*. Blackwell.
- Papineau, D. 1993. *Philosophical Naturalism*. Blackwell.
- Peacocke, C. 1986. *Thoughts*. Blackwell.
- Peacocke, C. 1992. *A Study of Concepts*. MIT Press.
- Pinker, S. 1994. *The Language Instinct*. Penguin Press.
- Pinker, S. 1997. *How the Mind Works*. Penguin Press.
- Povinelli, D. 2000. *Folk Physics for Apes*. Oxford University Press.
- Rosenthal, D. 1986. Two concepts of consciousness. *Philosophical Studies*, 49.
- Rosenthal, D. 1993. Thinking that one thinks. In Davies and Humphreys, eds., 1993.
- Rosenthal, D. forthcoming. *Consciousness and the Mind*. Oxford University Press.
- Searle, J. 1992. *The Rediscovery of the Mind*. MIT Press.
- Searle, J. 1997. *The Mystery of Consciousness*. A New York Review Book.
- Seager, W. 1994. Dretske on HOT theories of consciousness. *Analysis*, 54.
- Siewert, C. 1998. *The Significance of Consciousness*. Princeton University Press.
- Sperber, D. 1996. *Explaining Culture*. Blackwell.
- Sturgeon, S. 2000. *Matters of Mind: consciousness, reason and nature*. Routledge.
- Tye, M. 1995. *Ten Problems of Consciousness*. MIT Press.
- Tye, M. 1999. Phenomenal consciousness: the explanatory gap as cognitive illusion. *Mind*, 108.
- Weiskrantz, L. 1986. *Blindsight*. Oxford University Press.

- Weiskrantz, L. 1997. *Consciousness Lost and Found*. Oxford University Press.

Other Internet Resources

- Cognitive Science E-print Archive
- Bibliography on Higher-order Thought Approaches to Consciousness, by David Chalmers (U. Arizona)
- *Psyche*, an interdisciplinary journal of research on consciousness

Related Entries

consciousness: and intentionality | consciousness: animal | consciousness: representational theories of

Peter Carruthers
peter_carruthers@umail.umd.edu

First published: April 3, 2001
Content last modified: April 3, 2001

Animal Consciousness

There are the many reasons for philosophical interest in nonhuman animal (hereafter "animal") consciousness. First, if philosophy often begins with questions about the place of humans in nature, one way humans have attempted to locate themselves is by comparison and contrast with those things in nature most similar to themselves, i.e., other animals. Second, the problem of determining whether animals are conscious stretches the limits of knowledge and scientific methodology (beyond breaking point, according to some). Third, the question of whether animals are conscious beings or "mere automata", as Cartesians would have it, is of considerable moral significance given the dependence of modern societies on mass farming and the use of animals for biomedical research. Fourth, while theories of consciousness are frequently developed without special regard to questions about animal consciousness, the plausibility of such theories has sometimes been assessed against the results of their application to animal consciousness.

Questions about animal consciousness are just one corner of a more general set of questions about animal cognition and mind. The so-called "cognitive revolution" that took place during the latter half of the 20th century has led to many innovative experiments by comparative psychologists and ethologists probing the cognitive capacities of animals. Despite all this work, the topic of consciousness per se in animals has remained controversial, even taboo, among scientists, even while it remains a matter of common sense to most people that many other animals do have conscious experiences.

Concepts of Consciousness

In discussions of animal consciousness there is no clearly agreed upon sense in which the term "consciousness" is used. Having origins in folk psychology, "consciousness" has a multitude of uses that may not be resolvable into a single, coherent concept (Wilkes 1984). Nevertheless, several useful distinctions among different notions of consciousness have been made, and with the help of these distinctions it is possible to gain some clarity on the important questions that remain about animal consciousness.

Two ordinary senses of consciousness which are not in dispute when applied to animals are the sense of consciousness involved when a creature is awake rather than asleep, or in a coma, and the sense of consciousness implicated in the basic ability of organisms to perceive and thereby respond to selected features of their environments, thus making them conscious or aware of of those features. Consciousness in both these senses is identifiable in organisms belong to a wide variety of taxonomic groups.

There are two remaining senses of consciousness that cause controversy when applied to animals: *phenomenal consciousness* and *self-consciousness*.

Phenomenal consciousness refers to the qualitative, subjective, experiential, or phenomenological aspects of conscious experience, sometimes identified with qualia. (In this article I also use the term "sentience" to refer to phenomenal consciousness.) To contemplate animal consciousness in this sense is to consider the possibility that, in Nagel's (1974) phrase, there might be "something it is like" to be a member of another species. Nagel disputes our capacity to know, imagine, or describe in scientific (objective) terms *what* it is like to be a bat, but he assumes that there *is* something it is like. There are those, however, who would challenge this assumption directly. Others would less directly challenge the possibility of scientifically investigating its truth. Nevertheless, there is broad commonsense agreement that phenomenal consciousness is more likely in mammals and birds than it is in invertebrates, such as insects, crustaceans or molluscs (with the possible exception of some cephalopods), while reptiles, amphibians, and fish constitute an enormous grey area.

Self-consciousness refers to an organism's capacity for second-order representation of the organism's own mental states. Because of its second-order character ("thought about thought") the capacity for self consciousness is closely related to questions about "theory of mind" in nonhuman animals -- whether any animals are capable of attributing mental states to others. Questions about self-consciousness and theory of mind in animals are a matter of active scientific controversy, with the most attention focused on chimpanzees and to a more limited extent on the other great apes. As attested by this controversy (and unlike questions about animal sentience) questions about self-consciousness in animals are commonly regarded as tractable by empirical means.

The remainder of this article deals primarily with the attribution of consciousness in its phenomenal sense to animals, although there will be some discussion of self-consciousness and theory of mind in animals, in connection with arguments by Carruthers (1998a,b, 2000) that theory of mind is required for phenomenal consciousness.

Basic Questions: Epistemological and Ontological

The topic of consciousness in nonhuman animals has been primarily of epistemological interest to philosophers of mind. Two central questions are:

1. Can we know which animals beside humans are conscious? (The Distribution Question)
2. Can we know what, if anything, the experiences of animals are like? (The Phenomenological Question)

In his seminal paper "What is it like to be a bat?" Thomas Nagel (1974) simply assumes that there *is* something that it is like to be a bat, and focuses his attention on what he argues is the scientifically intractable problem of knowing what it is like. Nagel's confidence in the existence of conscious bat experiences would generally be held to be the commonsense view, but as we shall see, it is subject to challenge, and there are those who would argue that the Distribution Question is just as intractable as the Phenomenological Question.

The two questions might be seen as special cases of the general skeptical "problem of other minds", which even if intractable is nevertheless generally ignored to good effect by psychologists. However it is often thought that knowledge of animal minds -- what Allen & Bekoff (1997) refer to as "the other species of mind problem" -- presents special methodological problems because we cannot interrogate animals directly about their experiences (but see Sober 2000 for discussion of tractability within an evolutionary framework). Although there have been attempts to teach human-like languages to members of other species, none has reached a level of conversational ability that would solve this problem directly. Furthermore, except for some language-related work with parrots and dolphins, such approaches are generally limited to those animals most like ourselves, particularly the great apes. But there is great interest in possible forms of consciousness in a much wider variety of species than are suitable for such research, both in connection with questions about the ethical treatment of animals (e.g. Singer 1975/1990; Regan 1983; Rollin 1989; Varner 1999), and in connection with questions about the natural history of consciousness (Griffin 1976, 1984, 1992; Bekoff et al. 2002).

Griffin's agenda for the discipline he labeled "cognitive ethology" features the topic of animal consciousness and advocates a methodology based in naturalistic observations of animal behavior. This agenda has been strongly criticized, with his methodological suggestions often dismissed as anthropomorphic (see Bekoff & Allen 1997 for a survey). But such criticisms may have overestimated the dangers of anthropomorphism (Fisher 1990) and many of the critics themselves rely on claims for which there is scant scientific data (e.g. Kennedy 1992, who claims that the "sin" of anthropomorphism may be programmed into humans genetically).

While epistemological and related methodological issues have been at the forefront of discussions about animal consciousness, the main wave of recent philosophical attention to consciousness has been focused on ontological questions about the nature of phenomenal consciousness. One might reasonably think that the question of what consciousness is should be settled prior to tackling the Distribution Question -- that

ontology should drive the epistemology. In an ideal world this order of proceeding might be the preferred one, but as we shall see in the next section, the current state of disarray among the ontological theories makes such an approach untenable.

Applying Ontological Theories

Non-reductive accounts

Whether because they are traditional dualists, or because they think that consciousness is an as-yet-undescribed fundamental constituent of the physical universe, some philosophers maintain that consciousness is not explainable in familiar scientific terms. Such non-reductive accounts of consciousness (with the possible exception of those based in anthropocentric theology) provide no principled ontological reasons, however, for doubting that animals are conscious.

Cartesian dualism is, of course, traditionally associated with the view that animals lack minds. But Descartes' argument for this view was not based on any ontological principles, but upon what he took to be the failure of animals to use language conversationally or reason generally. On this basis he claimed that nothing in animal behavior requires a non-mechanical (mental) explanation; hence he saw no reason to attribute possession of mind to animals.

There is, however, no ontological reason why animal bodies are any less suitable vehicles for embodying a Cartesian mind than are human bodies. Hence dualism itself does not preclude animal minds. Similarly, more recent non-reductive accounts of consciousness in terms of fundamental properties are quite compatible with the idea of animal consciousness. None of these accounts provides any constitutional reason why those fundamental properties should not be located in animals. Furthermore, given that none of these theories specify empirical means for detecting the right stuff for consciousness, and indeed dualist theories *cannot* do so, they seem forced to rely upon behavioral criteria rather than ontological criteria for the deciding the Distribution Question.

Reductive accounts

Other philosophers have tried to give reductive accounts of consciousness in terms either of the physical, biochemical, or neurological properties of nervous systems (*physicalist* accounts) or in terms of other cognitive processes (*functionalist* accounts).

Physicalist accounts of consciousness, which identify consciousness with physical or physiological properties of neurons, do not provide any particular obstacles to attributing consciousness to animals, given that animals and humans share the same basic biology. Of course there is no consensus about which physical or neurological properties are to be identified with consciousness. But if it could be determined that phenomenal consciousness was identical to a property such as quantum coherence in the microtubules of neurons, or brain waves of a specific frequency, then settling the Distribution Question

would be a straightforward matter of establishing whether or not members of other species possess the specified properties.

Functionalist reductive accounts have sought to explain consciousness in terms of other cognitive processes. Some of these accounts identify phenomenal consciousness with the (first-order) representational properties of mental states. Such accounts are generally quite friendly to attributions of consciousness to animals, for it is relatively uncontroversial that animals have internal states that have the requisite representational properties. Such a view underlies Dretske's (1995) claim that phenomenal consciousness is inseparable from a creature's capacity to perceive and respond to features of its environment, i.e., one of the uncontroversial senses of consciousness identified above. On Dretske's view, phenomenal consciousness is therefore very widespread in the animal kingdom. Likewise, Tye (2000) argues, based upon his first-order representational account of phenomenal consciousness, that it extends even to honeybees.

Functionalist theories of phenomenal consciousness that rely on more elaborately structured cognitive capacities can be less accommodating to the belief that animals do have conscious mental states. For example, some twentieth century philosophers, while rejecting Cartesian dualism, have turned his epistemological reliance upon language as an indicator of consciousness into an ontological point about the essential involvement of linguistic processing in human consciousness. Such insistence on the importance of language for consciousness underwrites the tendency of philosophers such as Dennett (1969, 1995, 1997) to deny that animals are conscious in anything like the same sense that humans are (see also Carruthers 1996).

Because Carruthers has explicitly applied his functionalist "higher order thought" theory of phenomenal consciousness to derive a negative conclusion about animal consciousness (Carruthers 1998a,b, 2000) this account deserves special attention here. According to Carruthers, a mental state is conscious for a subject just in case it is available to be thought about directly by that subject. Furthermore, according to Carruthers, such higher-order thoughts are not possible unless a creature has a "theory of mind" to provide it with the concepts necessary for thought about mental states. But, Carruthers argues, there is little, if any, scientific support for theory of mind in nonhuman animals, even among the great apes, so he concludes that there is little support either for the view that any animals possess phenomenological consciousness. The evaluation of this argument will be taken up further below, but it is worth noting here that since most developmental psychologists agree that young children before the age of 4 lack a theory of mind, Carruthers' view entails that they are not sentient either -- fear of needles notwithstanding! This is a bullet Carruthers bites, although for many it constitutes a *reductio* of his view (a response Carruthers would certainly regard as question-begging).

In contrast to Carruthers' higher-order *thought* account of sentience, other theorists such as Armstrong (1980), and Lycan (1996) have preferred a higher-order *experience* account, where consciousness is explained in terms of inner perception of mental states, a view that can be traced back to Aristotle, and also to John Locke. Because such models do not require the ability to conceptualize mental states, proponents of higher-order experience theories have been slightly more inclined than higher-order theorists to allow that such abilities may be found in other animals[1].

Phenomenal consciousness is just one feature (some would say the defining feature) of mental states or events. Any theory of animal consciousness must be understood, however, in the context of a larger investigation of animal cognition that (among philosophers) will also be concerned with issues such as intentionality (in the sense described by the 19th C. German psychologist Franz Brentano) and mental content (Dennett 1983, 1987; Allen 1992a,b, 1995, 1997).

Philosophical opinion divides over the relation of consciousness to intentionality with some philosophers maintaining that they are strictly independent, others (particularly proponents of the functionalist theories of consciousness described in this section) arguing that intentionality is necessary for consciousness, and still others arguing that consciousness is necessary for genuine intentionality (see Allen 1997 for discussion). Many behavioral scientists accept cognitivist explanations of animal behavior that attribute representational states to their subjects. Yet they remain hesitant to attribute consciousness. If the representations invoked within cognitive science are intentional in Brentano's sense, then these scientists seem committed to denying that consciousness is necessary for intentionality.

Limits of ontology

There remains great uncertainty about the ontological status of consciousness. It is beyond the scope of this article to survey the strong attacks that have been mounted against the various accounts of consciousness, but it is safe to say that none of them seems secure enough to hang a decisive endorsement or denial of animal consciousness upon it. Accounts of consciousness in terms of basic neurophysiological properties, the quantum-mechanical properties of neurons, or *sui generis* properties of the universe are just as insecure as the various functionalist accounts. And even those ontological accounts that are, in general outline, compatible with animal sentience are not specific enough to permit ready answers to the Distribution Question. Hence no firm conclusions about the distribution of consciousness can be drawn on the basis of the work to date by philosophers on the ontology of consciousness.

Where does this leave the epistemological questions about animal consciousness? While it may seem natural to think that we must have a theory of what consciousness is before we try to determine whether other animals have it, this may in fact be putting the conceptual cart before the empirical horse. In the early stages of the scientific investigation of any phenomenon, putative samples must be identified by rough rules of thumb (or working definitions) rather than complete theories. Early scientists identified gold by contingent characteristics rather than its atomic essence, knowledge of which had to await thorough investigation of many putative examples -- some of which turned out to be gold and some not. Likewise, at this stage of the game, perhaps the study of animal consciousness would benefit from the identification of animal traits worthy of further investigation, with no firm commitment to idea that all these examples will involve conscious experience.

Of course, as a part of this process some reasons must be given for identifying specific animal traits as "interesting" for the study of consciousness, and in a weak sense such reasons will constitute an argument

for attributing consciousness to the animals possessing those traits. These reasons can be evaluated even in the absence of an accepted ontology for consciousness. Furthermore, those who would bring animal consciousness into the scientific fold in this way must also explain how scientific methodology is adequate to the task in the face of various arguments that it is inadequate. These arguments, and the response to them, can also be evaluated in the absence of ontological certitude. Thus there is plenty to cover in the remaining sections of this encyclopedia entry.

Evaluation of Arguments Against Animal Consciousness

Dissimilarity arguments

Recall the Cartesian argument from the previous section against animal consciousness (or animal mind) on the grounds that animals do not use language conversationally or reason generally. This argument, based on the alleged failure of animals to display certain intellectual capacities, is illustrative of a general pattern of using certain *dissimilarities* between animals and humans to argue that animals lack consciousness.

A common refrain in response to such arguments is that, in situations of partial information, "absence of evidence is not evidence of absence". Descartes dismissed parrots vocalizing human words because he thought it was merely meaningless repetition. This judgement may have been appropriate for the few parrots he encountered, but it was not based on a systematic, scientific investigation of the capacities of parrots. Nowadays many would argue that Pepperberg's study of the African Grey parrot "Alex" (Pepperberg 1999) should lay the Cartesian prejudice to rest. This study, along with several on the acquisition of a degree of linguistic competence by chimpanzees (e.g., Gardner et al. 1989; Savage-Rumbaugh 1996) would seem to undermine Descartes' assertions about lack of conversational language use and general reasoning abilities in animals.

Cartesians respond by pointing out the limitations shown by animals in such studies (they can't play a good game of chess, after all), and they join linguists in protesting that the subjects of animal-language studies have not fully mastered the recursive syntax of natural human languages. But this kind of post hoc raising of the bar suggests to many scientists that the Cartesian position is not being held as a scientific hypothesis, but as a dogma to be defended by any means. Nevertheless, dissimilarity arguments are not entirely powerless to give some pause to defenders of animal sentience, for surely most would agree that, at some point, the dissimilarities between the capacities of humans and the members of another species (the common earthworm *Lumbricus terrestris*, for example) are so great that it is unlikely that such creatures are sentient. A grey area arises precisely because no one can say how much dissimilarity is enough to trigger the judgement that sentience is absent.

Similarity arguments

A different kind of strategy that has been used to deny animal consciousness is to focus on certain *similarities* between animal behaviors and behaviors which may be conducted *un*consciously by humans. Thus, for example, Carruthers (1989, 1992) argued that *all* animal behavior can be assimilated to the non-conscious activities of humans, such as driving while distracted ("on autopilot"), or to the capacities of "blindsight" patients whose damage to visual cortex leaves them phenomenologically blind in a portion of their visual fields (a "scotoma") but nonetheless able to identify things presented to the scotoma. (He refers to both of these as examples of "unconscious experiences".)

This comparison of animal behavior to the unconscious capacities of humans can be criticized on the grounds that, like Descartes' pronouncements on parrots, it is based only on unsystematic observation of animal behavior. There are grounds for thinking that careful investigation would reveal that there is not a very close analogy between animal behavior and human behaviors associated with these putative cases of unconscious experience. For instance, it is notable that the unconscious experiences of automatic driving are not remembered by their subjects, whereas there is no evidence that animals are similarly unable to recall their allegedly unconscious experiences. Likewise, blindsight subjects do not spontaneously respond to things presented to their scotomas, but must be trained to make responses using a forced-response paradigm. There is no evidence that such limitations are normal for animals, or that animals behave like blindsight victims with respect to their visual experiences (Jamieson & Bekoff 1991).

Carruthers' argument from the absence of self-consciousness and theory of mind

In his more recent publications, Carruthers (1998a,b, 2000) appears to have moved on from the similarity argument of the previous section, now placing more stock in the argument based on his higher order thought theory that was described above. Recall that according to this argument, phenomenal consciousness requires the capacity to think about, and therefore conceptualize, one's own thoughts.[2] Such conceptualization requires, according to Carruthers, a theory of mind. And, Carruthers maintains, there is little basis for thinking that any nonhuman animals have a theory of mind, with the possible exception of chimpanzees. This argument is, of course, no stronger than the higher-order thought account of consciousness upon which it is based. But setting that aside for the sake of argument, this challenge by Carruthers deserves further attention as perhaps the most empirically-detailed case against animal consciousness to have been made in the philosophical literature.

The systematic study of self-consciousness and theory of mind in nonhuman animals has its roots in an approach to the study of self-consciousness pioneered by Gallup (1970). It was long known that chimpanzees would use mirrors to inspect their images, but Gallup developed a protocol that appears to allow a scientific determination of whether it is merely the mirror image *per se* that is the object of interest to the animal inspecting it, or whether it is the image qua proxy for the animal itself that is the object of interest. Using chimpanzees with extensive prior familiarity with mirrors, Gallup anesthetized his subjects and marked their foreheads with a distinctive dye, or, in a control group, anesthetized them only. Upon waking, marked animals who were allowed to see themselves in a mirror touched their own foreheads in the region of the mark significantly more frequently than controls who were either unmarked

or not allowed to look into a mirror. Gallup's protocol has been repeated with other great apes and some monkey species, but besides chimpanzees only orang utans consistently "pass" the test. Using a modified version of Gallup's procedure, involving no anesthesia, Reiss & Marino (2001) have recently produced evidence of mirror self-recognition in bottlenose dolphins.

According to Gallup et al. (2002) "Mirror self-recognition is an indicator of self-awareness." Furthermore, he claims that "the ability to infer the existence of mental states in others (known as theory of mind, or mental state attribution) is a byproduct of being self-aware." He describes the connection between self-awareness and theory of mind thus: "If you are self-aware then you are in a position to use your experience to model the existence of comparable processes in others." The success of chimpanzees on the mirror self-recognition task thus may give some reason to maintain that they are phenomenally conscious on Carruthers' account, whereas the failure of most species that have been tested to pass the test might be taken as evidence against their sentience.

Carruthers neither endorses nor outright rejects the conclusion that chimpanzees are sentient. His suspicion that even chimpanzees might lack theory of mind, and therefore (on his view) phenomenal consciousness, is based on some ingenious laboratory studies by Povinelli (1996) showing that in interactions with human food providers, chimpanzees apparently fail to understand the role of eyes in providing visual information to the humans, despite their outwardly similar behavior to humans in attending to cues such as facial orientation. The interpretation of Povinelli's work remains controversial. Hare et al. (2000) conducted experiments in which dominant and subordinate animals competed with each other for food, and concluded that "at least in some situations chimpanzees know what conspecifics do and do not see and, furthermore, that they use this knowledge to formulate their behavioral strategies in food competition situations." They suggest that Povinelli's negative results may be due to the fact that his experiments involve less natural chimp-human interactions. Given the uncertainty, Carruthers is therefore well-advised in the tentative manner in which he puts forward his claims about chimpanzee sentience.

A full discussion of the controversy over theory of mind deserves an entry of its own (see also Heyes 1998), but it is worth remarking here that the theory of mind debate has origins in the hypothesis that primate intelligence in general, and human intelligence in particular, is specially adapted for social cognition (see Byrne & Whiten 1988, especially the first two chapters, by Jolly and Humphrey). Consequently, it has been argued that evidence for the ability to attribute mental states in a wide range of species might be better sought in natural activities such as social play, rather than in laboratory designed experiments which place the animals in artificial situations (Allen & Bekoff 1997; see esp. chapter 6; see also Hare et al. 2000, Hare et al. 2001, and Hare & Wrangham 2002). Furthermore, to reiterate the maxim that absence of evidence is not evidence of absence, it is quite possible that the mirror test is not an appropriate test for theory of mind in most species because of its specific dependence on the ability to match motor to visual information, a skill that may not have needed to evolve in a majority of species. Alternative approaches that have attempted to provide strong evidence of theory of mind in nonhuman animals under natural conditions have generally failed to produce such evidence (see, e.g., the conclusions about theory of mind in vervet monkeys by Cheney & Seyfarth 1990), although anecdotal evidence tantalizingly suggests that researchers still have not managed to devise the right experiments.

Methodological arguments

Many scientists remain convinced that even if questions about self-consciousness are empirically tractable, no amount of experimentation can provide access to phenomenal consciousness in nonhuman animals. This remains true even among those scientists who are willing to invoke cognitive explanations of animal behavior that advert to internal representations. Opposition to dealing with consciousness can be understood as a legacy of behavioristic psychology first because of the behaviorists' rejection of terms for unobservables unless they could be formally defined, and second because of the strong association in many behaviorists' minds between the use of mentalistic terms and the twin bugaboos of Cartesian dualism and introspectionist psychology (Bekoff & Allen 1997). In some cases these scientists are even dualists themselves, but they are strongly committed to denying the possibility of scientifically investigating consciousness, and remain skeptical of all attempts to bring it into the scientific mainstream.

It is worth remarking that there is often a considerable disconnect between philosophers and psychologists (or ethologists) on the topic of animal minds. Some of this can be explained by the failure of some psychologists to heed the philosophers' distinction between intentionality in its ordinary sense and intentionality in the technical sense derived from Brentano (with perhaps most of the blame being apportioned to philosophers for failing to give clear explanations of this distinction and its importance). Indeed, some pscyhologists, having conflated Brentano's notion with the ordinary sense of intentionality, and then identifying the ordinary sense of intentionality with "free will" and conscious deliberation, have literally gone on to substitute the term "consciousness" in their criticisms of philosophers who were discussing the intentionality of animal mental states and who were not explicitly concerned with consciousness at all (see, e.g., Blumberg & Wasserman 1995).

Because consciousness is assumed to be private or subjective, it is often taken to be beyond the reach of objective scientific methods (see Nagel 1974). This claim might be taken in either of two ways. On the one hand it might be taken to bear on the possibility of answering the Distribution Question, i.e. to reject the possibility of knowledge that a member of another taxonomic group (e.g. a bat) has conscious states. On the other hand it might be taken to bear on the possibility of answering the Phenomenological Question, i.e. to reject the possibility of knowledge of the phenomenological details of the mental states of a member of another taxonomic group. The difference between believing with justification *that* a bat is conscious and knowing *what* it is like to be a bat is important because, at best, the privacy of conscious experience supports a negative conclusion only about the latter. To support a negative conclusion about the former one must also assume that consciousness has absolutely no measurable effects on behavior, i.e. one must accept epiphenomenalism. But if one rejects epiphenomenalism and maintains that consciousness does have effects on behavior then a strategy of inference to the best explanation may be used to support its attribution. More will be said about this in the next section.

Evaluation of Arguments For Animal

Consciousness

Similarity arguments

Most people, if asked why they think familiar animals such as their pets are conscious would point to similarities between the behavior of those animals and human behavior. Similarity arguments for animal consciousness thus have roots in common sense observations. But they may also be bolstered by scientific investigations of behavior and neurology as well as considerations of evolutionary continuity (homology) between species. Nagel's own confidence in the existence of phenomenally conscious bat experiences is based on nothing more than this kind of reliance on shared mammalian traits (Nagel 1974).

Many judgements of the similarity between human and animal behavior are readily made by ordinary observers. The reactions of many animals, particularly other mammals, to bodily events that humans would report as painful are easily and automatically recognized by most people as pain responses. High-pitched vocalizations, fear responses, nursing of injuries, and learned avoidance are among the responses to noxious stimuli that are all part of the common mammalian heritage. Similar responses are also visible to some degree or other in organisms from other taxonomic groups.

Less accessible to casual observation, but still in the realm of behavioral evidence are scientific demonstrations that members of other species, even of other phyla, are susceptible to the same visual illusions as we are (e.g. Fujita et al. 1991) suggesting that their visual experiences are similar.

Neurological similarities between humans and other animals are also been taken to suggest commonality of conscious experience. All mammals share the same basic brain anatomy, and much is shared with vertebrates more generally. A large amount of scientific research that is of direct relevance to the treatment of human pain, including on the efficacy of analgesics and anesthetics, is conducted on rats and other animals. The validity of this research depends on the similar mechanisms involved[3] and to many it seems arbitrary to deny that injured rats, who respond well to opiates for example, feel pain[4]. Likewise, much of the basic research that is of direct relevance to understanding human visual consciousness has been conducted on the very similar visual systems of monkeys. Monkeys whose primary visual cortex is damaged even show impairments analogous to those of human blindsight patients (Stoerig & Cowey 1997) suggesting that the visual consciousness of intact monkeys is similar to that of intact humans.

Such similarity arguments are, of course, inherently weak for it is always open to critics to exploit some *dis*analogy between animals and humans to argue that the similarities don't entail the conclusion that both are sentient (Allen 1998). Even when bolstered by evolutionary considerations of continuity between the species, the arguments are vulnerable, for the mere fact that humans have a trait does not entail that our closest relatives must have that trait too. There is no inconsistency with evolutionary continuity to maintain that only humans have the capacity to learn to play chess. Likewise for consciousness. Povinelli & Giambrone (2000) also argue that the argument from analogy fails because superficial observation of quite similar behaviors even in closely related species does not guarantee that the underlying cognitive

principles are the same, a point that Povinelli believes is demonstrated by his research described in the previous section, into how chimpanzees use cues to track visual attention (Povinelli 1996).

Perhaps a combination of behavioral, physiological and morphological similarities with evolutionary theory amounts to a stronger overall case[5]. But in the absence of more specific theoretical grounds for attributing consciousness to animals, this composite argument -- which might be called "the argument from homology" -- despite its comportment with common sense, is unlikely to change the minds of those who are skeptical.

Inference to the best explanation

One way to get beyond the weaknesses in the similarity arguments is to try to articulate a theoretical basis for connecting the observable characteristics of animals (behavioral or neurological) to consciousness. As mentioned above, one approach to bringing consciousness into the scientific fold is to try identify behaviors for which it seems that an explanation in terms of mechanisms involving consciousness might be justified over unconscious mechanisms by a strategy of inference to the best explanation. This form of inference would be strengthened by a good understanding of the biological function or functions of consciousness. If one knew what phenomenal conscious is *for* then one could exploit that knowledge to infer its presence in cases where that function is fulfilled, so long as other kinds of explanations can be shown less satisfactory.

If phenomenal consciousness is completely epiphenomenal, as some philosophers believe, then a search for the functions of consciousness is doomed to futility. In fact, if consciousness is completely epiphenomenal then it cannot have evolved by natural selection. On the assumption that phenomenal consciousness is an evolved characteristic of human minds, at least, and therefore that epiphenomenalism is false, then an attempt to understand the biological functions of consciousness may provide the best chance of identifying its occurrence in different species.

Such an approach is nascent in Griffin's attempts to force ethologists to pay attention to questions about animal consciousness. (For the purposes of this discussion I assume that Griffin's proposals are intended to relate to phenomenal consciousness, as well, perhaps, to consciousness in its other senses.) In a series of books, Griffin (who made his scientific reputation by carefully detailing the physical and physiological characteristics of echolocation by bats) provides examples of communicative and problem-solving behavior by animals, particularly under natural conditions, and argues that these are prime places for ethologists to begin their investigations of animal consciousness (Griffin 1976, 1984, 1992).

Although he thinks that the intelligence displayed by these examples suggests conscious thought, many critics have been disappointed by the lack of systematic connection between Griffin's examples and the attribution of consciousness (see Alcock 1992; Bekoff & Allen 1996; Allen & Bekoff 1997). Griffin's main positive proposal in this respect has been the rather implausible suggestion that consciousness might have the function of compensating for limited neural machinery. Thus Griffin is motivated to suggest that consciousness may be more important to honey bees than to humans.

If compensating for small sets of neurons is not a plausible function for consciousness, what might be? The commonsensical answer would be that consciousness "tells" the organism about events in the environment, or, in the case of pain and other proprioceptive sensations, about the state of the body. But this answer begs the question against opponents of attributing conscious states to animals for it fails to respect the distinction between phenomenal consciousness and mere awareness (in the uncontroversial sense of detection) of environmental or bodily events. Opponents of attributing the phenomenal consciousness to animals are not committed to denying the more general kind of consciousness *of* various external and bodily events, so there is no logical entailment from awareness of things in the environment or the body to animal sentience.

Perhaps more sophisticated attempts to spell out the functions of consciousness are similarly doomed. But Allen & Bekoff (1997, ch. 8) suggest that progress might be made by investigating the capacities of animals to adjust to their own perceptual errors. Not all adjustments to error provide grounds for suspecting that consciousness is involved, but in cases where an organism can adjust to a perceptual error while retaining the capacity to exploit the content of the erroneous perception, then there may be a robust sense in which the animal internally distinguishes its own appearance states from other judgements about the world. (Humans, for instance, have conscious visual experiences that they know are misleading -- i.e., visual illusions -- yet they can exploit the erroneous content of these experiences for various purposes, such as deceiving others or answering questions about how things *appear* to them.) Given that there are theoretical grounds for identifying conscious experiences with "appearance states", attempts to discover whether animals have such capacities might be a good place to start looking for animal consciousness. It is important, however, to emphasize that such capacities are not themselves intended to be definitive or in any way criterial for consciousness.

Carruthers (2000) makes a similar suggestion about the function of consciousness, relating it to the general capacity for making an appearance-reality distinction; of course he continues to maintain that this capacity depends upon having conceptual resources that are beyond the grasp of nonhuman animals.

Summary

An article such as this perhaps raises more questions than it answers, but the topic would be of little philosophical interest if it were otherwise.

To philosophers interested in animal welfare or animal rights the issue of animal sentience is of utmost importance. This is due to wide, but by no means universal, acceptance of the biconditional statement [A]: animals deserve moral consideration if and only if they are sentient (especially possessing the capacity to feel pain). Some philosophers have defended the view that animals are not sentient and attempted to use one of [A]'s component conditionals for *modus tollens*. Indeed Carruthers (1989) even argued that given their lack of sentience, it would be immoral *not* to use animals for research and other experimentation if doing so would improve the lot of sentient creatures such as ourselves. He has more recently backed off this view (1998b), denying [A] by claiming that sentience is not the sole basis for

moral consideration, and that animals qualify for consideration on the basis of frustration of their *unconscious* desires. Varner (1999) disagrees with Carruthers by arguing for conscious desires throughout mammals and birds, but like Carruthers he also rejects [A], arguing for an even more inclusive criterion of moral considerability in terms of the biological "interests" that all living things have.

Others are inclined to use the other component conditional of [A] for *modus ponens*, taking for granted that animals are conscious, and regarding any theory of consciousness which denies this as defective. In this connection it is also sometimes argued that if there is uncertainty about whether other animals really are conscious, the morally safe position is to give them the benefit of the doubt.

The fact remains that for most philosophers of mind, the topic of animal consciousness is of peripheral interest to their main project of understanding the ontology of consciousness. Because of their focus on ontological rather than epistemological issues, there is often quite a disconnect between philosophers and scientists on these issues. But there are encouraging signs that interdisciplinary work between philosophers and behavioral scientists is beginning to lay the groundwork for addressing some questions about animal consciousness in a philosophically sophisticated yet empirically tractable way.

Bibliography

- Akins, K. A. (1993) "A bat without qualities." In *Consciousness*, ed. M. Davies and G. Humphreys. Oxford: Blackwell.
- Alcock, J. (1992) Review of Griffin 1992. *Natural History*, September 1992: 62-65.
- Allen, C. (1992a) "Mental content." *British Journal for the Philosophy of Science* 43: 537-553.
- Allen, C. (1992b) "Mental content and evolutionary explanation." *Biology and Philosophy* 7: 1-12.
- Allen, C. (1995) "Intentionality: Natural and artificial." In H. Roitblat and J.-A.Meyer (eds.) *Comparative Approaches to Cognitive Science.* Cambridge, MA: MIT Press.
- Allen, C. (1997) "Animal cognition and animal minds." In *Philosophy and the Sciences of the Mind: Pittsburgh-Konstanz Series in the Philosophy and History of Science vol. 4.* ed. P. Machamer & M. Carrier Pittsburgh. and Konstanz: Pittsburgh University Press and the Universitätsverlag Konstanz: pp. 227-243. [Preprint available online]
- Allen, C. (1998) "The discovery of animal consciousness: an optimistic assessment." *Journal of Agricultural and Environmental Ethics* 10: 217-225.
- Allen, C. & Bekoff, M. (1997) *Species of Mind*. Cambridge, MA: MIT Press. See especially ch. 8.
- Andrews, K. (1996) "The first step in the case for great ape equality: the argument for other minds." *Etica & Animali* 8/96 (Special issue devoted to The Great Ape Project): 131-141.
- Armstrong, D. A. (1980) *The Nature of Mind and Other Essays* Ithaca, NY: Cornell University Press.
- Bekoff, M. & Allen, C. (1997) "Cognitive ethology: slayers, skeptics, and proponents. In *Anthropomorphism, Anecdote, and Animals*, ed. R. Mitchell et al. New York: SUNY Press.
- Bekoff, M., Allen, C., & Burghardt, G.M. (eds.) (2002) *The Cognitive Animal*, Cambridge, MA:

The MIT Press.
- Blumberg, M. S. & Wasserman, E. A.(1995) "Animal mind and the argument from design." *Am. Psychologist* 50: 133-144
- Burkhardt, R. W. Jr. (1997) "The founders of ethology and the problem of animal subjective experience." In *Animal Consciousness and Animal Ethics: Perspectives from the Netherlands* ed. M. Dol, S. Kasanmoentalib, S. Lijmbach, E. Rivas & R. van den Bos. Assen, the Netherlands: van Gorcum: pp. 1-13.
- Byrne R. W. & Whiten, A. (1988) *Machiavellian Intelligence: social expertise and the evolution of intellect in monkeys, apes and humans.* Oxford: Oxford University Press.
- Carruthers, P. (1989) "Brute Experience." *J. Phil.* 86: 258-269.
- Carruthers, P. (1992) *The Animals Issue*. Cambridge: Cambridge University Press.
- Carruthers, P. (1996) *Language, Thought and Consciousness.* Cambridge: Cambridge University Press.
- Carruthers, P. (1998a) "Natural Theories of Consciousness" *European Journal of Philosophy* 6: 203-222.
- Carruthers, P. (1998b) "Animal Subjectivity", *Psyche*, Vol. 4/No. 3 (April 1998) [Available online]; see also "Replies to Critics: Explaining Subjectivity", (his response to commentators), in *Psyche*, Vol. 6/No. 3 (February 2000) [Available online].
- Carruthers, P. (2000) *Phenomenal Consciousness: A naturalistic theory*. Cambridge: Cambridge University Press.
- Cheney, D. L., and Seyfarth, R. M. (1990) *How Monkeys See the World: Inside the mind of another species.* University of Chicago Press.
- Davidson, D. (1975) "Thought and talk." In Guttenplan, S. (ed.) *Mind and Language.* Oxford: Oxford University Press.
- Dawkins, M.S. (1993) *Through Our Eyes Only? The Search for Animal Consciousness*. New York: W. H. Freeman.
- Dennett, D. C. (1969) *Content and Consciousness.* London: Routledge and Kegan Paul.
- Dennett, D. C. (1983) "Intentional systems in cognitive ethology: The 'Panglossian paradigm' defended." *Behavioral and Brain Sciences* 6: 343-390.
- Dennett, D. C. (1987) *The Intentional Stance.* Cambridge, MA: MIT Press.
- Dennett, D. C. (1995) "Animal consciousness and why it matters." *Social Research* 62: 691-710.
- Dennett, D. C. (1997) *Kinds of Minds: Towards an Understanding of Consciousness* New York: Basic Books (Science Masters Series).
- Dretske, F. (1995) *Naturalizing the Mind*. Cambridge, MA: MIT Press.
- Fisher, J. A. (1990) "The myth of anthropomorphism." Originally published in M. Bekoff & D. Jamieson (eds.) *Interpretation and explanation in the study of animal behavior: Vol. 1, Interpretation, intentionality, and communication*. Boulder: Westview Press. Reprinted in Bekoff, M. , & D. Jamieson (eds.) (1996) *Readings in Animal cognition.* Cambridge, MA: MIT Press.
- Fujita, K., Blough, D. S., & Blough, P. M. (1991) "Pigeons see the Ponzo illusion." *Animal Learning and Behavior*, 19, 283-293.
- Gallup, G. G., Jr. (1970) "Chimpanzees: Self-recognition." *Science* 167: 86-87.
- Gallup, G. G., Jr., Anderson, J. R., & Shillito, D. J. (2002) "The Mirror Test" in Bekoff, Allen, & Burghardt (eds.).

- Gardner, R. A., Gardner, B. T., & Van Cantfort, T. E. (1989) *Teaching sign language to chimpanzees.* Albany, NY: SUNY Press.
- Griffin, D. R. (1976) *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience*. New York: Rockefeller University Press. (second edition: 1981).
- Griffin, D. R. (1984) *Animal Thinking*. Cambridge, MA: Harvard University Press.
- Griffin, D. R. (1992) *Animal Minds*. Chicago: University of Chicago Press.
- Hare, B., Call, J., Agnetta, B. & Tomasello, M. (2000) "Chimpanzees know what conspecifics do and do not see." *Animal Behavior* 59: 771-785.
- Hare B., Call J., Tomasello M. (2001) "Do chimpanzees know what conspecifics know?" *Animal Behaviour* 63:139-151.
- Hare, B., & Wrangham, R. (2002) "Integrating two evolutionary models for the study of social cognition." in Bekoff, Allen, & Burghardt (2002).
- Heyes, C. (1998) "Theory of mind in nonhuman primates." *Behavioral and Brain Sciences* 21: 101-148.
- Jamieson, D. & Bekoff, M. (1992) "Carruthers on nonconscious experience." *Analysis* 52: 23-28.
- Kennedy, J. S. (1992) *The new anthropomorphism.* New York: Cambridge University Press.
- Lycan, W. (1996), *Consciousness and Experience*. Cambridge, MA: MIT Press.
- Nagel, T. (1974) "What is it like to be a bat?", *Philosophical Review* 83: 435-450.
- Pepperberg, I.M. (1999) *The Alex Studies: Cognitive and communicative abilities of Grey parrots.* Cambridge, MA: Harvard University Press.
- Povinelli, D.J. (1996) "Chimpanzee theory of mind?" In P. Carruthers and P. Smith (eds.) *Theories of Theories of Mind.* Cambridge: Cambridge University Press.
- Povinelli, D.J. and Giambrone, S.J (2000) "Inferring Other Minds: Failure of the Argument by Analogy." *Philosophical Topics* 27:161-201.
- Radner, D. (1994) "Heterophenomenology: learning about the birds and the bees. *J. Phil.* 91: 389-403.
- Radner, D. & Radner, M. (1986) *Animal Consciousness*. Amherst, New York: Prometheus Books,
- Regan, T. (1993) *The Case for Animal Rights.* Berkeley: University of California Press. (See especially chs. 1 and 2.)
- Reiss, D. & Marino, L. (2001) "Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence." *Proceedings of the National Academy of Science* 98:5937-5942.
- Rollin, B. E. (1989) *The Unheeded Cry: Animal Consciousness, Animal Pain and Science*. New York: Oxford University Press.
- Rosenthal, D. (1986) "Two concepts of consciousness." *Philosophical Studies* 49, 329-359.
- Rosenthal, D. (1993) "Thinking that one thinks." In M. Davies and G. Humphreys (eds.), *Consciousness.* Oxford: Blackwell; 197-223.
- Savage-Rumbaugh, S. (1996) *Kanzi: The ape at the brink of the human mind.* New York: John Wiley & Sons.
- Singer, P. (1975/1990) *Animal Liberation* (Revised Edition, 1990). New York: Avon Books.
- Sober, E. (2000) "Evolution and the problem of other minds." *Journal of Philosophy* 97: 365-386.
- Sorabji, R. (1993) *Animal Minds and Human Morals: the origins of the Western debate*, Ithaca, NY: Cornell University Press.
- Stoerig, P. & Cowey, A. (1997) "Blindsight in man and monkey." *Brain* 120: 535-559.

- Tye, M. (2000) *Consciousness, Color, and Content.* Cambridge, MA: MIT Press.
- Varner, G. (1999) *In Nature's Interests?* New York: Oxford University Press.
- Wilkes, K. (1984) "Is consciousness important?" *British Journal for the Philosophy of Science* 35: 223-243.
- Wilson, M.D. (1995) "Animal ideas", *Proceedings and Addresses of the APA*: 69: 7-25.

Other Internet Resources

- *Psyche* Special Symposium on Animal Consciousness: Target article by Peter Carruthers (1998a) with author's abstract, peer commentary, and author's response.
- *Field Guide to the Philosophy of Mind* entry on Philosophy of Cognitive Ethology by Colin Allen, with accompanying Annotated Bibliography.
- A combined Bibliography assembled by Profs. Donald Griffin, Colin Allen, and Marc Bekoff.
- The Animal Consciousness section from Prof. David Chalmer's bibliography on Contemporary Philosophy of Mind.

[Please contact the author with other suggestions.]

Related Entries

abduction | behaviorism | Brentano, Franz | consciousness: and intentionality | dualism | epiphenomenalism | folk psychology: as a theory | other minds | qualia | rights: of animals

Colin Allen
colin.allen@tamu.edu

First published: December 23, 1995
Content last modified: July 3, 2002

Stanford Encyclopedia of Philosophy
Notes to Animal Consciousness

Notes

1. The evidence for this is somewhat ambiguous, however. For example, Güven Güzeldere (1995) reports private correspondence from David Armstrong who wrote:

> ... following Locke and Kant, I think the introspective awareness is perception-like. For instance, it is very like proprioception, and seems not to involve any linguistic capacity. Perhaps chimpanzees and even dogs have such consciousness." [G. Güzeldere (1995, fn. 22): "Is consciousness the perception of what passes in one's own mind?" In *Conscious Experience* ed. Thomas Metzinger. Paderborn: Schöningh/Imprint Academic, 1995, fn. 22.]

And Lycan, commenting on Carruthers (1998a) in his contribution to the *Psyche* symposium on animal consciousness writes:

> Even if I am right about human beings, of course it does not follow that other animals exhibit a comparable degree of computational complexity. Perhaps some do and some do not; perhaps few if any do. I would continue to maintain that an animal has phenomenal-consciousness in the strong sense if and only if that animal has HOEs [Higher Order Experiences]. So I would at least provisionally conclude that if many animals (including very young human children) lack the computational complexity needed for HOEs, those same many animals lack phenomenal-consciousness in the strong sense. Carruthers would not disagree with that.

These are hardly ringing endorsements by higher-order experience theorists for animal consciousness.

2. This argument is very reminiscent of Davidson's (1975) argument against animal thought on the grounds that one must have the concept of thought to be able to think. I am not aware that Davidson was ever tempted to turn this into an argument against animal sentience.

3. See Varner (1999, pp. 51-54) for a discussion of the phylogenetic distribution of the mechanisms underlying pain. However, caution is recommended with respect to his table on p. 53 which suggests that the presence of specialized nociceptive neurons is limited to vertebrates. The existence of nociceptors seems to be well established in marine molluscs. (See E. T. Walters [1992] "Possible clues about the evolution of hyperalgesia from mechanisms of nociceptive sensitization in *aplysia*." In *Hyperalgesia and Allodynia* ed.W. D. Willis Jr. New York: Raven Press.)

4. While the locution "feel pain" suggests conscious experience to most readers, it does not do so for all. For instance, Carruthers (1998b) maintains that although animals do in fact feel pain, they do not feel pain *consciously* - or, as Carruthers somewhat paradoxically puts it, he denies that the feeling feels like anything. Many philosophers would, however, charge that the concept of an unconscious pain is conceptually incoherent, pain being *essentially* conscious.

5. Because of the importance of animal sentience to ethics, some of the most explicit statements of the evolution-reinforced similarity argument can be found in the animal rights and animal welfare literature (e.g., Singer 1975/1990; Regan 1983; Andrews 1996; Varner 1999).

Colin Allen
colin.allen@tamu.edu

First published: November 2, 2000
Content last modified: December 4, 2000

Behaviorism

It has sometimes been said that "behave is what organisms do." Behaviorism is built on this assumption, and its goal is to promote the scientific study of behavior.

In this entry I consider different types of behaviorism. I outline reasons for and against being a behaviorist. I consider contributions of behaviorism to the study of behavior. Special attention is given to the so-called "radical behaviorism" of B. F. Skinner (1904-90).

What is Behaviorism?

Behaviorism is a doctrine, or set of doctrines, about human and nonhuman animal behavior. Behaviorism is committed in its fullest and most complete sense to the truth of the following three sets of claims.

> (1) Psychology is the science of behavior. Psychology is not the science of mind.
>
> (2) Behavior can be described and explained without making reference to mental events or to internal psychological processes. The sources of behavior are external (in the environment), not internal (in the mind).

> (3) In the course of theory development in psychology, if, somehow, mental terms or concepts are deployed in describing or explaining behavior, then either (a) these terms or concepts should be eliminated and replaced by behavioral terms or (b) they can and should be translated or paraphrased into behavioral concepts.

The three sets of claims are logically distinct. Moreover, taken independently, each helps to form a type of behaviorism. "Methodological" behaviorism is committed to the truth of (1). "Psychological" behaviorism is committed to the truth of (2). "Analytical" behaviorism (also known as "philosophical" or "logical" behaviorism) is committed to the truth of the statement in (3) that mental terms or concepts can and should be translated into behavioral concepts.

Other nomenclature is sometimes used to classify behaviorisms. Georges Rey (1997, p. 96), for example, classifies behaviorisms as methodological, analytical, and radical, where "radical" is Rey's term for what I am classifying as psychological behaviorism. I reserve the term "radical" for the psychological behaviorism of B. F. Skinner. Skinner employs the expression "radical behaviorism" to describe his brand of behaviorism or his philosophy of behaviorism (see Skinner 1976, p. 18). In my classification scheme, radical behaviorism is a sub-type of psychological behaviorism, primarily, although it combines all three types of behaviorism (methodological, analytical, and psychological).

Three Types of Behaviorism

Methodological behaviorism is a normative theory about the scientific conduct of psychology. It claims that psychology should concern itself with the behavior of organisms (human and nonhuman animals). Psychology should not concern itself with mental states or events or with constructing internal information processing accounts of behavior. According to methodological behaviorism, reference to mental events (such as an animal's beliefs or desires) adds nothing to what psychology can and should understand about the sources of behavior. Mental events are private entities which, given the necessary publicity of science, do not form proper objects of empirical study. Methodological behaviorism is a dominant theme in the writings of John Watson (1878-1958).

Psychological behaviorism is a research program within psychology. It purports to explain human and animal behavior in terms of external physical stimuli, responses, learning histories, and (for certain types of behavior) reinforcements. Psychological behaviorism is present in the work of Ivan Pavlov (1849-1936), Edward Thorndike (1874-1949), as well as Watson. Its fullest and most influential expression is B. F. Skinner's (1904-90) work on schedules of reinforcement.

To illustrate, consider a food-deprived rat in an experimental chamber. If a particular movement, such as pressing a lever when a light is on, is followed by the presentation of food, then the likelihood of the rat's pressing the lever when hungry, again, and the light is on, is increased. Such presentations are reinforcements, such lights are (discriminative) stimuli, such lever pressings are responses, and such trials or associations are learning histories.

Analytical behaviorism is a theory within philosophy about the meaning or semantics of mental terms or concepts. It says that the very notion of a mental state or condition is the notion of a behavioral disposition or family of behavioral tendencies. When we attribute a belief to someone, for example, we are not saying that he or she is in a particular internal state or condition. Instead, we are characterizing the person in terms of what he or she might do in particular situations. Analytical behaviorism may be found in the work of Gilbert Ryle (1900-76) and the later work of Ludwig Wittgenstein (1889-51).

Roots of Behaviorism

Each of methodological, psychological, and analytical behaviorism has historical foundations. Analytical behaviorism traces its historical roots to the philosophical movement known as Logical Positivism (see Smith 1986). Logical positivism proposes that the meaning of sentences used in science be understood in terms of the experimental conditions or observations that verify their truth. This positivist doctrine is known as "verificationism". In psychology, verificationism grounds analytical behaviorism, namely, the claim that mental concepts refer to behavioral tendencies and so must be translated into behavioral terms.

Analytical behaviorism helps to avoid substance dualism. Substance dualism is the doctrine that mental states take place in a special, non-physical mental substance (the immaterial mind). By contrast, for analytical behaviorism the belief that I have in, for example, arriving on time for a 2pm dental appointment, believing that I have a 2pm appointment, is not the property of a mental substance. Believing is a family of tendencies of my body. We cannot identify the belief independently of my arrival or other members of the family. So, we cannot treat it as the cause of the arrival.

Psychological behaviorism's historical roots consist, in part, in the classical associationism of the British Empiricists, foremost John Locke (1632-1704) and David Hume (1711-76). According to classical associationism, intelligent behavior is the product of associative learning. As a result of associations or pairings between perceptual experiences or stimulations on the one hand, and ideas or thoughts on the other, persons and animals acquire knowledge of their environment and how to act. Associations enable creatures to discover the causal structure of the world. Association is most helpfully viewed as the acquisition of knowledge about relations between events. Intelligence in behavior is a mark of such knowledge.

Classical associationism relied on introspectible entities, such as perceptual experiences or stimulations as the first links in associations, and thoughts or ideas as the second links. Psychological behaviorism, motivated by experimental interests, claims that to understand the origins of behavior, reference to stimulations (experiences) should be replaced by reference to stimuli (physical events in the environment), and that reference to thoughts or ideas should be eliminated or displaced in favor of reference to responses (overt behavior). Psychological behaviorism is associationism without appeal to mental events.

Don't human beings talk of introspectible entities even if these are not recognized by behaviorism?

Psychological behaviorists regard the practice of talking about one's own states of mind, and of introspectively reporting those states, as potentially useful data in psychological experiments, but as not presupposing the metaphysical subjectivity or non-physical presence of those states. There are different sorts of causes behind introspective reports, and psychological behaviorists take these to be amenable to behavioral analysis. (See, by comparison, Dennett's method of heterophenomenology; Dennett 1991, pp. 72-81).

The task of psychological behaviorism is to specify types of association, understand how environmental events control behavior, discover and elucidate causal regularities or laws which govern the formation of associations, and predict how behavior will change as the environment changes. The word "conditioning" is commonly used to specify the process involved in acquiring new associations. Animals in so-called "operant" conditioning experiments are not learning to, for example, press levers. Instead, they are learning about the relationship between events in their environment, for example, that a particular behavior, pressing the lever, causes food to appear.

In its historical foundations, methodological behaviorism shares with analytical behaviorism the influence of positivism. One of the goals of positivism was to unify psychology with natural science. Watson wrote that "psychology as a behaviorist views it is a purely objective experimental branch of natural science. Its theoretical goal is . . . prediction and control" (1913, p. 158). Watson also wrote of the purpose of psychology as follows: "To predict, given the stimulus, what reaction will take place; or, given the reaction, state what the situation or stimulus is that has caused the reaction" (1930, p. 11).

Though logically distinct, methodological, psychological, and analytical behaviorisms often are found in one behaviorism. Skinner's radical behaviorism combines all three forms of behaviorism. It follows analytical strictures (at least loosely) in paraphrasing mental terms behaviorally, when or if they cannot be eliminated from explanatory discourse. In Verbal Behavior (1957) and elsewhere, Skinner tries to show how mental terms can be given behavioral interpretations. In About Behaviorism (1976) he says that when mental terminology cannot be eliminated it can be "translated into behavior" (p. 18, Skinner brackets the expression with his own double quotes).

Radical behaviorism is concerned with the behavior of organisms, not with internal processing. So, it is a form of methodological behaviorism. Finally, radical behaviorism understands behavior as a reflection of frequency effects among stimuli, which means that it is a form of psychological behaviorism.

Popularity of Behaviorism

Behaviorism of one sort or another was an immensely popular research program or methodological commitment among students of behavior from about the second decade of the twentieth century through its middle decade (see Bechtel, Abrahamsen, and Graham, 1998, pp. 15-17). In addition to Ryle and Wittgenstein, philosophers with sympathies for behaviorism included Carnap (1932-33), Hempel (1949), and Quine (1960). Quine, for example, took a behaviorist approach to the study of language. Quine claimed that the notion of psychological or mental activity has no place in a scientific account of either

the origins or the meaning of speech. To talk in a scientifically disciplined manner about the meaning of an utterance is to talk about stimuli for the utterance, its so-called "stimulus meaning". Hempel (1949) claimed that "all psychological statements that are meaningful . . . are translatable into statements that do not involve psychological concepts," but only concepts for physical behavior (p. 18).

Among psychologists behaviorism was even more popular than among philosophers. In addition to Pavlov, Skinner, Thorndike, and Watson, the list of behaviorists among psychologists included, among others, E. C. Tolman (1886-1959), C. L. Hull (1884-52), and E. R. Guthrie (1886-1959). Tolman, for example, wrote that "everything important in psychology . . . can be investigated in essence through the continued experimental and theoretical analysis of the determiners of rat behavior at a choice point in a maze" (1938, p. 34).

Behaviorists created journals, organized societies, and founded psychology graduate programs reflective of behaviorism. Behaviorists organized themselves into different types of research clusters, whose differences stemmed from such factors as varying approaches to conditioning and experimentation. Some clusters were named as follows: "the experimental analysis of behavior", "behavior analysis", "functional analysis", and, of course, "radical behaviorism". These labels sometimes were responsible for the titles of behaviorism's leading societies and journals, including the Society for the Advancement of Behavior Analysis (SABA), and the Journal of the Experimental Analysis of Behavior (begun in 1958) as well as the Journal of Applied Behavior Analysis (begun in 1968).

Behaviorism generated a type of therapy, known as behavior therapy (see Rimm and Masters 1974; Erwin 1978). It developed behavior management techniques for autistic children (see Lovaas and Newsom 1976) and token economies for the management of chronic schizophrenics (see Stahl and Leitenberg 1976). It fueled discussions of how best to understand the behavior of nonhuman animals, the relevance of laboratory study to the natural environmental occurrence of behavior, and whether there is built-in associative bias in learning (see Schwartz and Lacey 1982).

Behaviorism stumbled upon various critical difficulties with its commitments. One difficulty is confusion about the effects of reinforcement (see Gallistel 1990). In its original sense, a stimulus is a reinforcer only if its presentation increases the frequency of a response in a type of associative conditioning known as operant conditioning. A problem with this definition is that it defines reinforcers as stimuli that change behavior. The presentation of food, however, may have no observable effect on response frequency. It may, instead, be associated with an animal's ability to identify and remember temporal or spatial properties of the circumstances in which reinforcement occurs. This and other difficulties prompted changes in behaviorism's commitments and new directions of research. One recent and fresh direction has been the study of the role of short term memory in contributing to reinforcement effects on the so-called trajectory of behavior (see Killeen 1994).

Another stumbling block, in the case of analytical behaviorism, is the fact that the behavioral sentences that are intended to offer the behavioral paraphrases of mental terms almost always use mental terms themselves (see Chisholm 1957). In the example of my belief that I have a 2pm dental appointment, one

must also speak of my desire to arrive at 2pm, otherwise the behavior of arriving at 2pm could not count as believing that I have a 2pm appointment. The term "desire" is a mental term. Critics have charged that we can never escape from using mental terms in the characterization of the meaning of mental terms. This suggests that mental discourse cannot be displaced by behavioral discourse. At least it cannot be displaced term-by-term. Perhaps analytical behaviorists need to paraphrase a whole swarm of mental terms at once (see Rey 1997, p. 154-5).

Why be a Behaviorist

Why would anyone be a behaviorist? There are three main reasons (see also Zuriff 1985).

The first is epistemic. Warrant or evidence for saying that an animal or person is in a certain mental state, for example, possesses a certain belief, is grounded in behavior, understood as observable behavior. Moreover, the conceptual space between the claim that behavior warrants the attribution of belief and the claim that believing consists in behavior is a short and in some ways appealing step. If we look, for example, at how people are taught to use mental concepts and terms -- terms like "believe", "desire", and so on -- the conditions of use appear inseparably connected with behavioral tendencies in certain circumstances.

The second reason can be expressed as follows: One major difference between mentalistic (mental states in-the-head) and associationist or conditioning accounts of behavior is that mentalistic accounts tend to have a strong nativist bent. This is true even though there may be nothing inherently nativist about mentalistic accounts (see Cowie 1998).

Mentalistic accounts tend to assume, and sometimes even explicitly to embrace (see Fodor 1981), the hypothesis that the mind possesses at birth or innately a set of procedures or internally represented processing rules which are deployed when learning or acquiring new responses. Behaviorism, by contrast, is anti-nativist. Behaviorism, therefore, appeals to theorists who deny that there are innate rules by which organisms learn. To Skinner and Watson organisms learn without being innately or pre-experientially provided with explicit procedures by which to learn. Learning does not consist in rule-governed behavior. Learning is what organisms do in response to stimuli. A behaviorist organism learns, as it were, from its successes and mistakes. (See Dennett 1978).

Much contemporary work in cognitive science on the set of models known as connectionist or parallel distributed processing (PDP) models seems to share behaviorism's anti-nativism about learning. PDP takes an approach to learning which is response oriented rather than rule-governed and this is because, like behaviorism, it has roots in associationism (see Bechtel 1985; compare Graham 1991 with Maloney 1991). Whether PDP models ultimately are or must be anti-nativist depends upon what counts as native or innate rules (Bechtel and Abrahamsen 1991, pp. 103-105).

The third reason for behaviorism's appeal, popular at least historically, is related to its disdain for

reference to inner mental or information processing as a means to explain behavior. The disdain is most vigorously exemplified in the work of Skinner. Skinner's skepticism about explanatory references to mental innerness may be expressed as follows.

Behavior must be explained in terms which do not themselves presuppose the very thing that is explained. This is behavior. The outside (public) behavior of a person is not accounted for by referring to the inside (inner processing) behavior of the person (say, his or her internal problem solving or thinking) if, therein, the behavior of the person is unexplained. "The objection," wrote Skinner, "to inner states is not that they do not exist, but that they are not relevant in a functional analysis" (Skinner 1953, p. 35). ‘Not relevant’ means, for Skinner, explanatorily circular or regressive.

Skinner charges that since mental activity is a form of behavior (albeit inner), the only non-regressive, non-circular way to explain behavior is to appeal to something non-behavioral. This non-behavioral something is environmental stimuli and an organism's interactions with, and reinforcement from, the environment.

So, the third reason for behaviorism's appeal is that it tries to avoid circular, regressive explanations of behavior. It aims to refrain from accounting for one type of behavior (overt) in terms of another type of behavior (covert), all the while, in some sense, leaving behavior unexplained.

It should be noted that Skinner's views about explanation are particularly extreme (scientifically naive?), and that many who called themselves behaviorists including Guthrie, Tolman, and Hull, or continue to work within the tradition, broadly understood, including Killeen (1987) and Rescorla (1990), take exception to much that Skinner has said about explanatory references to innerness. It should also be noted that Skinner's derisive attitude towards explanatory references to mental innerness stems, in part, from his conviction that if the language of psychology is permitted to refer to internal processing in its explanations of behavior, this goes some way towards permitting talk of immaterial mental substances, agents endowed with contra-causal free will, and little persons (homunculi) within bodies, each of which Skinner takes to be incompatible with a scientific worldview (see Skinner 1971). Finally, it must be emphasized that Skinner's aversion to explanatory references to innerness is not an aversion to inner states or processes per se. Skinner countenances talk of inner events provided that they are treated in the same manner as public responses. "An adequate science of behavior," he wrote, "must consider events taking place within the skin of the organism . . . as part of behavior itself" (1984, p. 617).

This last point is worth additional discussion, since the failure to appreciate Skinner's willingness to talk of inner events has helped to produce confusion in understanding his attitude towards the mental. Skinner does not deny the existence of events such as thinking and perceiving or various other events which he sometimes classifies as mental or inner. True, it is hard to know what Skinner counts as inner, or mental, and why, especially given that he sometimes uses terms like ‘inner’ and ‘mental’ as epithets of dismissal. However, set against Skinner's derision of the mental is his persistent reminder that people think, perceive and therein respond beneath their skin. So, Skinner pictures inner events as follows.

Inner events are those about which we may make introspective reports (and thus they are private, observationally), but their causal explanatory force is idle. Because inner events are private observationally, their patterns of reinforcement are more elusive, less easy to deliberately regulate, than overt behavior. However, inner events are responses, ultimately, to environmental stimuli.

Skinner's Social Worldview

Skinner is the only major figure in the history of behaviorism to offer a socio-political world view based on his commitment to behaviorism. Skinner constructed a theory as well as narrative picture in Walden Two (1948) of what an ideal human society would be like if designed according to behaviorist principles (see also Skinner 1971). Skinner's social worldview illustrates both his aversion to free will, to homunculi, to dualism as well as his reasons for claiming that a person's history of environmental interactions controls his or her behavior.

One remarkable feature of human behavior which Skinner deliberately rejects is that people creatively make their own environments (see Chomsky 1971, Black 1973). The world is as it is, in part, because we make it that way. Skinner protests that "it is in the nature of an experimental analysis of human behavior that it should strip away the functions previously assigned to autonomous man and transfer them one by one to the controlling environment" (1971, p. 198).

Critics have raised several objections to the Skinnerian social picture. One of the most persuasive, and certainly one of the most frequent, adverts to Skinner's vision of the ideal human society. It is a question asked of the fictional founder of Walden Two, Frazier, by the philosopher Castle. It is the question of what is the best social mode of existence for a human being. Frazier's, and therein Skinner's, response to this question is both too general and incomplete. Frazier/Skinner speaks of the values of health, friendship, relaxation, rest, and so forth. However, these values are hardly the detailed basis of a social system.

There is a notorious problem in social theory of specifying the appropriate level of detail at which a blueprint for a new and ideal society must be presented (see Arnold 1990, pp. 4-10). Skinner identifies the behavioristic principles and learning incentives that he hopes will reduce systematic injustices in social systems. He also describes a few practices (concerning child rearing and the like) that are intended to contribute to human happiness. However he offers only the haziest descriptions of the daily lives of Walden Two citizens and no suggestions for how best to resolve disputes about alternative ways of life that are prima facie consistent with behaviorist principles (see Kane 1996, p. 203). He gives little or no serious attention to the crucial general problem of inter-personal conflict resolution and to the role of institutional arrangements in resolving conflicts.

In an essay which appeared in The Behavior Analyst (1985), nearly forty years after the publication of Walden Two, Skinner, in the guise of Frazier, tried to clarify his characterization of ideal human circumstances. He wrote that in the ideal human society "people just naturally do the things they need to do to maintain themselves . . . and treat each other well, and they just naturally do a hundred other things

they enjoy doing because they do not have to do them" (p. 9). However, of course, doing a hundred things humans enjoy doing means only that Walden Two is vaguely defined, not that its culturally instituted habits and the character of its institutions merit emulation.

The incompleteness of Skinner's description of the ideal human society or life is so widely acknowledged that one might wonder if actual experiments in Walden Two living could lend useful detail to his blueprint. At least two such experiments have been and are being conducted, one in Virginia, the other in Mexico. Both can be indirectly explored via the Internet (see Other Internet Resources).

Why be Anti-Behaviorist

Behaviorism is unpopular. It is dismissed by cognitive scientists developing intricate internal information processing models. It is neglected by cognitive ethologists and ecological psychologists convinced that its methods are irrelevant to studying how animals and persons behave in their natural and social environment. It is rejected by neuroscientists sure that direct study of the brain is the only way to understand the causes of behavior.

Remnants of behaviorism survive in both behavior therapy and laboratory-based animal learning theory. In the metaphysics of mind, too, behavioristic themes survive in the approach to mind known as functionalism. Functionalism defines states of mind as states that play particular causal-functional roles in animals or systems in which they occur. Paul Churchland writes of functionalism as follows: "The essential or defining feature of any type of mental states is the set of causal relations it bears to . . . bodily behavior (1984, p. 36). This functionalist notion is similiar to the behaviorist idea that reference to behavior and to stimulus/response relations enters centrally and essentially into any account of what it means for a creature to behave or to be subject, in the scheme of analytical behaviorism, to the attribution of mental states.

Remnants, however, are remnants. Behaviorism has lost strength and influence. Why?

The deepest and most complex reason for behaviorism's demise is its commitment to the thesis that behavior can be explained without reference to mental activity. Many philosophers and psychologists find this thesis hopelessly restrictive. They reject behaviorism because of it. At the lunch table, for instance, I recognize a situation in which I am presented with apples as a situation in which I am presented with apples and I form concepts of apples, sort apples into classes (e.g. ripe and unripe), and draw upon those classifications as the situation permits, eating a ripe apple and avoiding unripe apples. Recognizing, conceptually sorting, and drawing upon are information processing activities which take place inside my head -- in my mind. These events are not (overt) behavior, although they may be revealed or expressed in behavior and reference to them helps to explain behavior.

To illustrate the explanatory counter-intuitiveness of restricting psychology to outside-the-head environmental histories, suppose that one morning Mother Nature throws a lightening bolt at a swamp

and that, in the consequent chemical reaction, a creature appears in the swamp that is a molecule-for-molecule duplicate of some actual human being -- me, say. Should Swamp Me step out of the swamp, as I would step out of the swamp?

Since Swamp Me is an exact physical duplicate of me, whatever physical stimulus is applied, we may think, he should react to the stimulus in the same manner that I would, and produce exactly the same response. We should picture him as possessing my behavioral tendencies before his first behavioral response. Is this picture compatible with behaviorism?

At the very least, the expectation that Swamp Me behaves just as I would is prima facie incompatible with behaviorism. This is because Swamp Me has no environmental history. He has not been reinforced for anything.

Alternatively, one might reply that Swamp Me shares in my behavioral tendencies by virtue of being my physical duplicate. For an organism to have a behavioral disposition, it shouldn't be necessary for the organism to have its own learning history. What is at least equally sufficient is what is encoded or stored in its head. It must have a record of a learning history embedded in its brain or central nervous system. One need not insist that this is the organism's own history. Swamp Me may act just as I would act, however this is only because it ‘remembers’ certain behaviors and reinforcement for those behaviors.

This sort of reply seems pretty obvious. However, one might wonder whether it is compatible with behaviorism. Here's why.

The memory-in-the-head model of Swamp Me seems not be a behaviorist theory, as behaviorists have expressed their theories. Instead, it is close to a non-behaviorist account of the significance of having a reinforcement history, where what matters is the internal record, and not what occurs in the world (the history).

The critical question is what counts as a record of a reinforcement history. Many psychologists despair of describing what counts as a record without postulating internal memory states and internal processing over those states (see Roediger and Goff 1998). Because memory states serve to record past experiences and reinforcements, and serve referentially to stand for stimulus conditions which the organism remembers, they are commonly described as internal mental representations. Talk of internal mental representation, however, is a perspective from which behaviorism -- at least of Skinner and other traditional behaviorists -- has wished to depart.

One defining feature of traditional behaviorism is that it tried to free psychology from having to theorize about how animals and persons represent their environment. This was important, historically, because it seemed that behavior/environment connections are a lot clearer and more manageable experimentally than internal representations. Unfortunately, for behaviorism, it's hard to imagine a more restrictive rule for psychology than a rule which prohibits hypotheses about representational storage and processing. Stich, for example, complains against Skinner that "we now have an enormous collection of experimental

data which, it would seem, simply cannot be made sense of unless we postulate something like" information processing mechanisms in the heads of organisms (1998, p. 649).

A second reason for rejecting behaviorism is that some features of mentality -- some elements in the inner processing of persons -- have characteristic ‘feels’ or sensory or phenomenal qualities. To be in pain, for example, is not merely to produce appropriate pain behavior under the right environmental circumstances, it is to experience a ‘like-thisness’ to the pain (as something dull or sharp, perhaps). Behaviorist creatures may engage in pain behavior, including beneath the skin pain responses, yet completely lack whatever is qualitatively definitive of pain (its feel). (See also Graham 1998, pp. 47-51.; Graham and Horgan 2000).

Feels, or qualia, as they also are called, are difficult for behaviorism because feels subjectively are present in experience, but resist behavioral analysis or description. Indeed, it is tempting to postulate that feels affect non-qualitative elements of internal processing, and that they, for example, contribute to arousal, attention, and receptivity to associative conditioning.

The third reason for rejecting behaviorism is connected with Noam Chomsky. Chomsky has been one of behaviorism's most successful and damaging critics. In a review of Skinner's book on verbal behavior (see above), Chomsky (1959) argued that some behavior (linguistic behavior, in particular) has to be understood in terms of internally represented rules. These rules are not products of learned associations. They are part of our native psychological endowment as human beings. Chomsky charged that behaviorist models of language learning cannot explain various facts about language acquisition, such as the rapid acquisition of language by young children, which is sometimes referred to as the phenomenon of "lexical explosion". A child's linguistic abilities appear to be radically under-determined by the evidence of verbal behavior offered to the child in the short period in which he or she acquires those abilities. By the age of four or five (normal) children have an almost limitless capacity to understand and produce sentences which they have never heard before. The basic rules or principles of grammar, therefore, argues Chomsky, must be innate.

The problem to which Chomsky refers, which is the problem of behavioral capacities outstripping individual learning histories, seems to go beyond merely the issue of linguistic behavior in young children. It appears to be a fundamental fact about human beings that our sensitivities and behavioral capacities often surpass the limitations of our individual learning histories. Our history of reinforcement often is too impoverished to determine uniquely our behavior. Much learning, therefore, seems to require pre-existing or innate representational structures within which learning occurs. (See also Brewer 1974, but compare with Bates et al. 1998 and Cowie 1998).

Conclusion

In 1977 Willard Day, a behavioral psychologist and founding editor of the journal Behaviorism, published Skinner's "Why I am not a cognitive psychologist" (Skinner 1977). Skinner began the paper by stating that "the variables of which human behavior is a function lie in the environment" (p. 1). Skinner

ended by remarking that "cognitive constructs give . . . a misleading account of what" is inside a human being (p. 10)

More than a decade earlier, in 1966 Hempel announced his defection from behaviorism:

> In order to characterize . . . behavioral patterns, propensities, or capacities . . . we need not only a suitable behavioristic vocabulary, but psychological terms as well. (p. 110)

Hempel had come to believe that it is a mistake to imagine that human behavior can be understood exclusively in non-mental, behavioristic terms.

Contemporary philosophy and psychology largely share Hempel's conviction that the explanation of behavior cannot omit invoking a creature's representation of its world. Psychology must use psychological terms. Behavior without representation is blind. Psychological theorizing without reference to internal processing is explanatorily impaired. Behaviorism, not cognitive science or psychology, offers a misleading account of what is inside the head.

Bibliography

- Arnold, N. S. 1990. *Marx's Radical Critique of Capitalist Society*. Oxford: Oxford University Press.
- Bates, E., Ellman, J., Johnson, M., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. 1988. "Innateness and Emergentism". In W. Bechtel and G. Graham (eds.), *A Companion to Cognitive Science*. Oxford, UK: Blackwell.
- Bechtel, W. 1985. "Contemporary Connectionism: Are the New Parallel Distributed Processing Models Cognitive or Associationist?", *Behaviorism*, 13, 53-61. (See Skinner 1977.)
- Bechtel, W., and Abrahamsen, A. 1991. *Connectionism and the Mind*. Oxford, UK: Blackwell.
- Bechtel, W., Abrahamsen, A., and Graham, G. 1998. "The Life of Cognitive Science". In W. Bechtel and G. Graham (eds.). (See Bates 1988).
- Black, M. 1973. "Some Aversive Responses to a Would-be Reinforcer." In: H. Wheeler (ed.), *Beyond the Punitive Society* (pp. 125-34). San Francisco: W. H. Freeman.
- Brewer, W. F. 1974. "There is No Convincing Evidence for Operant or Classical Conditioning in Adult Humans". In: W. Weiner and D. Palermo (eds.), *Cognition and Symbolic Processes*. Hillsdale, N. J.: Earlbaum.
- Carnap, R. 1932/33. "Psychology in Physical Language," *Erkenntnis*, 3, 107-42.
- Chisholm, R. M. 1957. *Perceiving*. Ithaca: Cornell.
- Chomsky, N. 1959. "Review of Verbal Behavior," *Language*, 35, 26-58.
- Chomsky, N. 1971. "The Case Against B. F. Skinner," *New York Review of Books*, 30, 18-24.
- Chomsky, N. 1975. *Reflections on Language*. New York: Pantheon Books.
- Churchland, P. 1984. *Matter and Consciousness*. Cambridge, MA.: MIT Press/Bradford Books.
- Cowie, F. 1998. *What's Within: Nativism Reconsidered*. Oxford: Oxford.
- Dennett, D. 1978. "Why the Law of Effect Will Not Go Away." in: D. Dennett (ed.) *Brainstorms*

(pp. 71-89). Cambridge, MA.: MIT Press/Bradford Books.

- Dennett, D. 1991. *Consciousness Explained*. Boston: Little, Brown and Company.
- Erwin, E. 1978. *Behavior Therapy: Scientific, Philosophical, and Moral Foundations*. Cambridge: Cambridge University Press.
- Fodor, J. 1981. "The Present Status of the Innateness Controversy". In J. Fodor (ed.), *Representations* (pp. 257-316). Cambridge, MA.: MIT Press/Bradford Books.
- Gallistel, C. R. 1990. *The Organization of Learning*. Cambridge, MA.: MIT Press.
- Graham, G. 1991. "Connectionism in Pavlovian Harness". In T. Horgan and J. Tienson (eds.), *Connectionism and the Philosophy of Mind* (pp. 143-66). Dordrecht: Kluwer.
- Graham, G. 1998. *Philosophy of Mind: An Introduction*, 2nd edition. Oxford: Basil Blackwell.
- Graham, G. and Horgan, T. 2000. "Mary, Mary, Quite Contrary," *Philosophical Studies*, forthcoming.
- Hempel, C. 1949. "The Logical Analysis of Psychology". In H. Feigl and W. Sellars (eds.), *Readings in Philosophical Analysis* (pp. 373-84). New York: Appleton-Century-Crofts.
- Hempel, C. 1966. *Philosophy of Natural Science*. Englewood Cliffs, N. J.: Prentice-Hall.
- Honig, W. and J. G. Fetterman (eds), 1992. *Cognitive Aspects of Stimulus Control*. Hillsdale, N. J.: Erlbaum.
- Quine, W. 1960. *Word and Object*. Cambridge, MA.: MIT Press.
- Kane, R. 1996. *The Significance of Free Will*. Oxford: Oxford.
- Killeen, P. 1987. "Emergent Behaviorism". In S. Modgil and C. Modgil (eds.), *B. F. Skinner: Consensus and Controversy* (pp. 219-34). New York: Falmer
- Killeen, P. 1994. "Mathematical Principles of Reinforcement," *Behavioral and Brain Sciences*, 17, 105-172.
- Lovaas, O. I. and Newsom, C. D. 1976. "Behavior Modification with Psychotic Children". In H. Leiteberg (ed.), *Handbook of Behavior Modification and Behavior Therapy*. Englewoood Cliffs, N. J.: Prentice-Hall.
- O'Donnell, J. 1985. The Origins of Behaviorism: *American Psychology*, 1870-1920. New York: NYU Press.
- Mackenzie, B. 1977. *Behaviorism and the Limits of Scientific Method*. London: Routledge & Kegan Paul.
- Maloney, C. 1991. "Connectionism and Conditioning". In T. Horgan and J. Tienson (eds.), *Connectionism and the Philosophy of Mind* (pp. 167-95). Dordrecht: Kluwer.
- Rey, G. 1997. *Contemporary Philosophy of Mind: A Contentiously Classical Approach*. Oxford: Blackwell.
- Rescorla, R. A. 1990. "The Role of Information about the Response-Outcome Relationship in Instrumental Discrimination Learning," *Journal of Experimental Psychology*: Animal Behavior Processes, 16, 262-70.
- Rimm, D. C. and Masters, J. C. 1974. *Behavior Therapy: Techniques and Empirical Findings*. New York: Academic Press.
- Roediger, H. and Goff, L. 1998. "Memory". In: Bechtel, W. and Graham, G. (eds.) (See Bates 1998.)
- Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson. Schwartz, B. and Lacey, H. 1982. Behaviorism, Science, and Human Nature. New York: Norton.

- Skinner, B. F. 1948. *Walden Two*. New York: Macmillan.
- Skinner, B. F. 1953. *Science and Human Behavior*. New York: Macmillan.
- Skinner, B. F. 1971. *Beyond Freedom and Dignity*. New York: Knopf.
- Skinner, B. F. 1976. *About Behaviorism*. New York: Vintage.
- Skinner, B. F. 1977. "Why I am not a Cognitive Psychologist". *Behaviorism*, 5, 1-10. (This journal is now known as Behavior and Philosophy.)
- Skinner, B. F. 1998. "Behaviorism at Fifty". *Behavioral and Brain Sciences*. 7, 615-621. This paper originally appeared in Science (1983) 140,951-958. The BBS issue is a special issue devoted to the canonical papers of B. F. Skinner.
- Skinner, B. F. 1985. "News from Nowhere, 1984". *The Behavior Analyst*, 8, 5-14.
- Smith, L. 1986. *Behaviorism and Logical Positivism: A Reassessment of Their Alliance*. California: Stanford.
- Stahl, J. R. and Leitenberg, H. 1976. "Behavioral treatment of the chronic mental hospital patient". In: Leitenberg (ed.).
- Stich, S. 1984. "Is Behaviorism Vacuous?," *Behavioral and Brain Sciences*, 7, 647-649.
- Tolman, E. C. "The Determiners of Behavior at a Choice Point," *Psychological Review*, 45, 1-41.
- Watson, J. 1913. "Psychology as a Behaviorist Views It," *Psychological Review*, 20, 158-77.
- Watson, J. 1930. *Behaviorism*. Norton: New York.
- Wittgenstein, L. 1953/1968. *Philosophical Investigation*s, trans. G. E. M. Anscombe. Oxford: Basil Blackwell.
- Zuriff, G. 1985. *Behaviorism: A Conceptual Reconstruction*. New York: Columbia University Press.

Other Internet Resources

- Association for Behavior Analysis
- Behavior Analysis Homepage
- Walden Two, Twin Oaks Community, Virginia

Related Entries

cognitive science | connectionism | qualia

George Graham
ggraham@uab.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 26, 2000
Content last modified: December 11, 2000

Stanford Encyclopedia of Philosophy

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Qualia

Feelings and experiences vary widely. For example, I run my fingers over sandpaper, smell a skunk, feel a sharp pain in my finger, seem to see bright purple, become extremely angry. In each of these cases, I am the subject of a mental state with a very distinctive subjective character. There is something it is *like* for me to undergo each state, some phenomenology that it has. Philosophers often use the term ‘qualia’ (singular ‘quale’) to refer to the introspectively accessible, phenomenal aspects of our mental lives. In this standard, broad sense of the term, it is difficult to deny that there are qualia. Disagreement typically centers on which mental states have qualia, whether qualia are intrinsic qualities of their bearers, and how qualia relate to the physical world both inside and outside the head. The status of qualia is hotly debated in philosophy largely because it is central to a proper understanding of the nature of consciousness. Qualia are at the very heart of the mind-body problem.

The entry that follows is divided into eight sections. The first comments on other more restricted uses of the term ‘qualia’. The second addresses the question of which mental states have qualia. The third section brings out some of the main arguments for the view that qualia are irreducible and non-physical. The remaining sections focus on functionalism and qualia, the explanatory gap, qualia and introspection, representational theories of qualia, and finally the issue of qualia and simple minds.

I. Other Uses of the Term ‘Qualia’

Consider a painting of a dalmatian. Viewers of the painting can apprehend not only its content (i.e., its representing a dalmatian) but also the colors, shapes, and spatial relations obtaining among blobs of paint partly by virtue of which it has that content. It has sometimes been supposed that being aware or conscious of a visual experience is like viewing an inner, non-physical picture or sense-datum. So, for example, on this conception, if I see a dalmatian, I am subject to a mental picture-like representation of a dalmatian (a sense-datum), introspection of which reveals to me both its content and its intrinsic, non-intentional features partly by virtue of which it has that content. These intrinsic, non-intentional features have been taken by advocates of the sense-datum theory to be the sole determinants of what it is like for me to have the experience. In a second, more restricted sense of the term ‘qualia’, then, qualia are intrinsic, consciously accessible, non-intentional features of sense-data and other non-physical phenomenal objects that are responsible for their phenomenal character.

With the demise of the sense-datum theory, there are very few philosophers who believe that there are qualia, so conceived. Still, there is another established sense of the term ‘qualia’, which is similar to the one just given but which does not demand of qualia advocates that they endorse the now discredited sense-datum theory. However sensory experiences are ultimately analyzed -- whether, for example, they are taken to involve relations to sensory objects or they are identified with neural events or they are held to be physically irreducible events -- many philosophers suppose that they have intrinsic, consciously accessible features that are neither intentional nor intentionally determined and that are solely responsible for their phenomenal character. These features, whatever their ultimate nature, physical or non-physical, are often dubbed ‘qualia’.

In the case of visual experiences, for example, it is frequently supposed that there is a range of visual qualia, where these are taken to be intrinsic features that (a) are accessible to introspection, (b) can vary without any variation in the intentional contents of the experiences, (c) are mental counterparts to some directly visible properties of objects (e.g., color), and (d) are the sole determinants of the phenomenal character of the experiences. Philosophers who hold or have held this view include, for example, Nagel (1974), Peacocke (1983) and Block (1990).

Philosophers who deny that there are qualia often have in mind qualia, as the term is used in the senses specified in this section. Sometimes their target is qualia, conceived of as in the opening paragraph of the entry, but with the additional assumption (often not explicitly stated) that qualia are ineffable or nonphysical or ‘given’ to their subjects incorrigibly (without the possibility of error). Thus, announcements by philosophers who declare themselves opposed to qualia (e.g., Dennett 1987, 1991) need to be treated with some caution. One can agree that there are no qualia in the more restricted senses I have explained, and also agree that there are no ineffable or incorrigibly presented or non-physical qualities possessed by our mental states, while still endorsing qualia, in the standard broad sense.

In the rest of this entry, I shall use the term ‘qualia’ in the standard, broad way I did at the beginning of the entry. So, I shall take it for granted that there are qualia.

II. Which Mental States Possess Qualia?

The following would certainly be included on my own list. (1) Perceptual experiences, for example, experiences of the sort involved in seeing green, hearing loud trumpets, tasting liquorice, smelling the sea air, handling a piece of fur. (2) Bodily sensations, for example, feeling a twinge of pain, feeling an itch, feeling hungry, having a stomach ache, feeling hot, feeling dizzy. Think here also of experiences such as those present during orgasm or while running flat-out. (3) Felt reactions or passions or emotions, for example, feeling delight, lust, fear, love, feeling grief, jealousy, regret. (4) Felt moods, for example, feeling elated, depressed, calm, bored, tense, miserable. (For more here, see Haugeland 1985, pp. 230-235).

Should we include any other mental states on the list? Galen Strawson has recently claimed (1994) that there are such things as the *experience* of understanding a sentence, the *experience* of suddenly thinking of something, of suddenly remembering something, and so on. Moreover, in his view, experiences of these sorts are not reducible to associated sensory experiences and/or images. Strawson's position here seems to be that thought-experience is a distinctive experience in its own right. He says, for example: "Each sensory modality is an experiential modality, and thought experience (in which understanding-experience may be included) is an experiential modality to be reckoned alongside the other experiential modalities" (p. 196). On Strawson's view, then, some thoughts have qualia.

This view is controversial. One response is to claim that the phenomenal aspects of understanding derive largely from linguistic (or verbal) images, which have the phonological and syntactic structure of items in the subject's native language. These images frequently even come complete with details of stress and intonation. As we read, it is sometimes phenomenally as if we are speaking to ourselves. (Likewise when we consciously think about something without reading). We often "hear" an inner voice. Depending upon the content of the passage, we may also undergo a variety of emotions and feelings. We may feel tense, bored, excited, uneasy, angry. Once *all* these reactions are removed, together with the images of an inner voice and the visual sensations produced by reading, some would say (myself included) that no phenomenology remains.

In any event, images and sensations of the above sorts are not always present in thought. They are not *essential* to thought. Consider, for example, the thoughts involved in everyday visual recognition (or the thoughts of creatures without a natural language). Consider also deeply unconscious thoughts. So, certainly some thoughts lack qualia.

What about desires, for example, my desire for a week's holiday in Venice? It is certainly true that in some cases, there is an associated phenomenal character. Often when we strongly desire something, we experience a feeling of being "pulled" or "tugged". There may also be accompanying images in various modalities.

Should we include such propositional attitudes as feeling angry *that the house has been burgled* or seeing *that the computer is missing* on the list? These seem best treated as hybrid or complex states, one component of which is essentially a phenomenal state and the other (a judgment or belief) is not. Thus, in

both cases, there is a constituent experience that is the real bearer of the relevant quale or qualia.

III. Are Qualia Irreducible, Non-Physical Entities?

The literature on qualia is filled with thought-experiments of one sort or another. Perhaps the most famous of these is the case of Mary, the brilliant color scientist. Mary, so the story goes (Jackson 1982), is imprisoned in a black and white room. Never having been permitted to leave it, she acquires information about the world outside from the black and white books her captors have made available to her, from the black and white television sets attached to external cameras, and from the black and white monitor screens hooked up to banks of computers. As time passes, Mary acquires more and more information about the physical aspects of color and color vision. (For a real life case of a visual scientist (Knut Nordby) who is an achromotope, see Sacks 1996, Chapter 1.) Eventually, Mary becomes the world's leading authority on these matters. Indeed she comes to know *all* the physical facts pertinent to everyday colors and color vision.

Still, she wonders to herself: What do people in the outside world *experience* when they see the various colors? What is it *like* for them to see red or green? One day her captors release her. She is free at last to see things with their real colors (and free too to scrub off the awful black and white paint that covers her body). She steps outside her room into a garden full of flowers. "So, that is what it is like to experience red," she exclaims, as she sees a red rose. "And that," she adds, looking down at the grass, "is what it is like to experience green."

Mary here seems to make some important discoveries. She seems to find out things she did not know before. How can that be, if, as seems possible, at least in principle, she has all the physical information there is to have about color and color vision -- if she knows all the pertinent physical facts?

One popular explanation among philosophers (so-called ‘qualia freaks’) is that that there is a realm of subjective, phenomenal qualities associated with color, qualities the intrinsic nature of which Mary comes to discover upon her release, as she herself undergoes the various new color experiences. Before she left her room, she only knew the objective, physical basis of those subjective qualities, their causes and effects, and various relations of similarity and difference. She had no knowledge of the subjective qualities in themselves.

This explanation is not available to the physicalist. If what it is like for someone to experience red is one and the same as some physical quality, then Mary already knows *that* while in her room. Likewise, for experiences of the other colors. For Mary knows all the pertinent physical facts. What, then, can the physicalist say?

Some physicalists respond that knowing what it is like is know-how and nothing more. Mary acquires certain abilities, specifically in the case of red, the ability to recognize red things by sight alone, the ability to imagine a red expanse, the ability to remember the experience of red. She does *not* come to know any new information, any new facts about color, any new qualities. This is the view of David Lewis

(1990) and Lawrence Nemirow (1990).

The Ability Hypothesis, as it is often called, is more resilient than many philosophers suppose (see Tye forthcoming (a)). But it has difficulty in properly accounting for our knowledge of what it is like to undergo experiences of determinate hues while we are undergoing them. For example, I can know what it is like to experience red-17, as I stare at a rose of that color. Of course, I don't know the hue as red-17. My conception of it is likely just *that shade of red*. But I certainly know what it is like to experience the hue while it is present. Unfortunately, I lack the abilities Lewis cites and so does Mary even after she leaves her cell. She is not able to recognize things that are red-17 as red-17 by sight. Given the way human memory works and the limitations on it, she lacks the concept red-17. She has no mental template that is sufficiently fine-grained to permit her to identify the experience of red-17 when it comes again. Presented with two items, one red-17 and the other red-18, in a series of tests, she cannot say with any accuracy which experience her earlier experience of the rose matches. Sometimes she picks one; at other times she picks the other. Nor is she able afterwards to imagine things as having hue, red-17, or as having that very shade of red the rose had; and for precisely the same reason.

The Ability Hypothesis appears to be in trouble. An alternative physicalist proposal is that Mary in her room lacks certain *phenomenal concepts*, certain ways of thinking about or mentally representing color experiences and colors. Once she leaves the room, she acquires these new modes of thought as she experiences the various colors. Even so, the qualities the new concepts pick out are ones she knew in a different way in her room, for they are physical or functional qualities like all others.

One problem this approach faces is that it seems to imply that Mary does not really make a new discovery when she says, "So, that is what it is like to experience red." Upon reflection, however, it is far from obvious that this is really a consequence. For it is widely accepted that concepts or modes of presentation are involved in the individuation of thought-contents, given one sense of the term ‘content’ -- the sense in which thought-content is whatever information that-clauses provide that suffices for the purposes of even the most demanding rationalizing explanation. In this sense, what I think, when I think that Cicero was an orator, is not what I think when I think that Tully was an orator. This is precisely why it is possible to discover that Cicero is Tully. The thought that Cicero was an orator differs from the thought that Tully was an orator not at the level of truth-conditions -- the same singular proposition is partly constitutive of the content of both -- but at the level of concepts or mode of presentation. The one thought exercises the concept *Cicero*; the other the concept *Tully*. The concepts have the same reference, but they present the referent in different ways and thus the two thoughts can play different roles in rationalizing explanation.

So, there is no real difficulty in holding both that Mary comes to know some new things upon her release, while already knowing all the pertinent real-world physical facts, even though the new experiences she undergoes and their introspectible qualities are wholly physical. In an ordinary, everyday sense, Mary's knowledge increases. And that is all the physicalist needs to answer the Knowledge Argument. (The term ‘fact’, I should add, is itself ambiguous. Sometimes it is used to pick out real-world states of affairs alone; sometimes it is used for such states of affairs under certain conceptualizations. When I speak of the physical facts above, I should be taken to refer either to physical states of affairs alone or to those states of affairs under purely physical conceptualizations. For more on ‘fact’, see Tye 1995.)

Some philosophers insist that the difference between the old and the new concepts in this case is such that there must be a difference in the world between the properties these concepts stand for or denote (Jackson 1993, Chalmers 1996). Some of these properties Mary knew in her cell; others she becomes cognizant of only upon her release. The physicalist is committed to denying this claim.

The issues here are complex. If a necessary a posteriori identity claim about properties requires for its truth that one or the other of the rigid terms flanking the identity sign be a priori connected to a different property from the one it denotes (so that the mode of presentation is wholly distinct from the referent) then it may seem that physicalism about qualia is in trouble. For phenomenal terms are often viewed as picking out phenomenal qualities via those very qualities. (This point, or one very like it, is sometimes put within the framework of two-dimensional modal semantics by saying that phenomenal terms have the same primary and secondary intensions). But even granting this claim, the above view of necessary a posteriori identities entails that physicalism about qualia is false only if it is true that the physical term on the right hand side of the identity sign itself refers via a mode of presentation that is the same as the referent (so that its primary intension is the same as its secondary one). And, in general, that cannot be correct. To see this, consider a purely theoretical identity, that between H_2O and a certain quantum-mechanical system. The identity claim here is necessary a posteriori, assuming that 'H_2O' and the appropriate quantum-mechanical designator are both rigid. So, if a condition of a property identity being necessary a posteriori is that one or other of the designators express a concept whose mode of presentation is different from its referent, then we have, with this example, a case where the mode of presentation and the referent come apart for at least one of the theoretical terms (and arguably for both). Why not, then, for the rigid physical designator in the case of a phenomenal-physical identity?

Another famous anti-reductionist thought-experiment concerning qualia appeals to the possibility of zombies. A philosophical zombie is a molecule by molecule duplicate of a sentient creature, a normal human-being, for example, but who differs from that creature in lacking *any* phenomenal consciousness. For me, as I lie on the beach, happily drinking some wine and watching the waves, I undergo a variety of visual, olfactory, and gustatory experiences. But my zombie twin experiences nothing at all. He has no phenomenal consciousness. Since my twin is an exact physical duplicate of me, his inner psychological states will be *functionally* isomorphic with my own (assuming he is located in an identical environment). Whatever physical stimulus is applied, he will process the stimulus in the same way as I do, and produce exactly the same behavioral responses. Indeed, on the assumption that non-phenomenal psychological states are functional states (that is, states definable in terms of their role or function in mediating between stimuli and behavior), my zombie twin has just the same beliefs, thoughts, and desires as I do. He differs from me only with respect to experience. For him, there is nothing it is like to stare at the waves or to sip wine.

The hypothesis that there can be philosophical zombies is not normally the hypothesis that such zombies are *nomically* possible, that their existence is consistent with the actual laws of nature. Rather the suggestion is that zombie replicas of this sort are at least *imaginable* and hence metaphysically possible.

Philosophical zombies pose a serious threat to any sort of physicalist view of qualia. To begin with, if zombie replicas are metaphysically possible, then there is a simple argument that seems to show that phenomenal states are not identical with internal, objective, physical states. Suppose objective, physical state *P* can occur without phenomenal state *S* in some appropriate zombie replica (in the metaphysical sense of 'can' noted above). Intuitively *S* cannot occur without *S*. Pain, for example, cannot be felt without pain. So, *P* has a modal property *S* lacks, namely the property of *possibly* occurring without *S*. So, by Leibniz' Law (the law that for anything *x* and for anything *y*, if *x* is identical with *y* then *x* and *y* share *all* the same properties), *S* is not identical with *P*.

Secondly, if a person microphysically identical with me, located in an identical environment (both present and past), can lack *any* phenomenal experiences, then facts pertaining to experience and feeling, facts about qualia, are not necessarily fixed or determined by the objective microphysical facts. And this the physicalist cannot allow, even if she concedes that phenomenally conscious states are not strictly identical with internal, objective, physical states. For the physicalist, whatever her stripe, must at least believe that the microphysical facts determine all the facts, that any world that was exactly like ours in *all* microphysical respects (down to the smallest detail, to the position of every single boson, for example) would have to be like our world in all respects (having identical mountains, lakes, glaciers, trees, rocks, sentient creatures, cities, and so on).

One well-known physicalist reply to the case of zombies (Loar 1990) is to grant that they are conceptually possible, or at least that there is no *obvious* contradiction in the idea of a zombie, while denying that zombies are metaphysically possible. Since the anti-physicalist argument requires metaphysical possibility -- mere conceptual possibility will not suffice -- it now collapses. That conceptual possibility is too weak for the anti-physicalist's purposes (at least without further qualification and argument) is shown by the fact that it is conceptually possible that I am not Michael Tye (that I am an impostor or someone misinformed about his past) even though, given the actual facts, it is metaphysically impossible.

IV. Functionalism and Qualia

Functionalism is the view that individual qualia have functional natures, that the phenomenal character of, e.g., pain is one and the same as the property of playing such-and-such a causal or teleofunctional role in mediating between physical inputs (e.g., body damage) and physical outputs (e.g., withdrawal behavior). On this view (Lycan 1987), qualia are multiply physically realizable. Inner states that are physically very different may nonetheless feel the same. What is crucial to what it is like is functional role, not underlying hardware.

There are two famous objections to functionalist theories of qualia: the Inverted Spectrum and the Absent Qualia Hypothesis. The first move in the former objection consists in claiming that you might see red when I see green and vice-versa; likewise for the other colors so that our color experiences are phenomenally inverted. This does not suffice to create trouble for the functionalist yet. For you and I are surely representationally different here: for example, you have a visual experience that represents red

when I have one that represents green. And that representational difference brings with it a difference in our patterns of causal interactions with external things (and thereby a functional difference).

This reply can be handled by the advocate of inverted qualia by switching to a case in which we both have visual experiences with the same representational contents on the same occasions while still differing phenomenally. Whether such cases are really metaphysically possible is open to dispute, however. Certainly, those philosophers who are representationalists about qualia (see Section VII) would deny their possibility. Indeed, it is not even clear that such cases are conceptually possible (Harrison 1973, Hardin 1993, Tye 1995). But leaving this to one side, it is far from obvious that there would not have to be some salient fine-grained functional differences between us, notwithstanding our gross functional identity.

Consider a computational example. For any two numerical inputs, *M* and *N*, a given computer always produces as outputs the product of *M* and *N*. There is a second computer that does exactly the same thing. In this way, they are functionally identical. Does it follow that they are running exactly the same program? Of course, not! There are all sorts of programs that will multiply together two numbers. These programs can differ dramatically. At one gross level the machines are functionally identical, but at lower levels the machines can be functionally different.

In the case of you and me, then, the opponent of inverted qualia can claim that, even if we are functionally identical at a coarse level - we both call red things ‘red’, we both believe that those things are red on the basis of our experiences, we both are caused to undergo such experiences by viewing red things, etc. - there are necessarily fine-grained differences in our internal functional organization. And that is why our experiences are phenomenally different.

Some philosophers will no doubt respond that it is still imaginable that you and I are functionally identical in *all* relevant respects yet phenomenally different. But this claim presents a problem at least for those philosophers who oppose functionalism but who accept physicalism. For it is just as easy to imagine that there are inverted qualia in molecule-by-molecule duplicates (in the same external, physical settings) as it is to imagine inverted qualia in functional duplicates. If the former duplicates are really metaphysically impossible, as the physicalist is committed to claiming, why not the latter? Some further convincing argument needs to be given that the two cases are disanalogous. As yet, to my mind, no such argument has been presented. (Of course, this response does not apply to those philosophers who take the view that qualia are irreducible, non-physical entities. However, these philosophers have other severe problems of their own. In particular, they face the problem of phenomenal causation. Given the causal closure of the physical, how can qualia make any difference? For more here, see Tye 1995, Chalmers 1996).

The absent qualia hypothesis is the hypothesis that functional duplicates of sentient creatures are possible, duplicates that entirely lack qualia. For example, one writer (Block (1980)) asks us to suppose that a billion Chinese people are each given a two-way radio with which to communicate with one another and with an artificial (brainless) body. The movements of the body are controlled by the radio signals, and the signals themselves are made in accordance with instructions the Chinese people receive from a vast

display in the sky which is visible to all of them. The instructions are such that the participating Chinese people function like individual neurons, and the radio links like synapses, so that together the Chinese people duplicate the causal organization of a human brain. Whether or not this system, if it were ever actualized, would *actually* undergo any feelings and experiences, it seems coherent to suppose that it might not. But if this is a real metaphysical possibility, then qualia do not have functional essences.

One standard functionalist reply to cases like the China-body system is to bite the bullet and to argue that however strange it seems, the China-body system could not fail to undergo qualia. The oddness of this view derives, according to some functionalists (Lycan 1987), from our relative size. We are each so much smaller than the China-body system that we fail to see the forest for the trees. Just as a creature the size of a neuron trapped inside a human head might well be wrongly convinced that there could not be consciousness there, so we too draw the wrong conclusion as we contemplate the China-body system. It has also been argued (e.g., by Shoemaker 1975) that any system that was a full functional duplicate of one of us would have to be subject to all the same beliefs, including beliefs about its own internal states. Thus the China-Body system would have to believe that it experiences pain; and if it had beliefs of this sort, then it could not fail to be the subject of some experiences (and hence some states with phenomenal character). If this reply is successful, what it shows is that the property of having some phenomenal character or other has a functional essence. But it does not show that individual qualia are functional in nature. Thus one could accept that absent qualia are impossible while also holding that inverted spectra are possible (see, e.g., Shoemaker 1975).

V. Qualia and the Explanatory Gap

Our grasp of what it is like to undergo phenomenal states is supplied to us by introspection. We also have an admittedly incomplete grasp of what goes on objectively in the brain and the body. But there is, it seems, a vast chasm between the two. It is very hard to see how this chasm in our understanding could ever be bridged. For no matter how deeply we probe into the physical structure of neurons and the chemical transactions which occur when they fire, no matter how much objective information we come to acquire, we still seem to be left with something that we cannot explain, namely, why and how such-and-such objective, physical changes, whatever they might be, generate so-and-so subjective feeling, or any subjective feeling at all.

This is the famous "explanatory gap" for qualia (Levine 1983). Some say that the explanatory gap is unbridgeable and that the proper conclusion to draw from it is that there is a corresponding gap in the world. Experiences and feelings have irreducibly subjective, non-physical qualities (Jackson 1993, Chalmers 1996). Others take essentially the same position on the gap while insisting that this does not detract from a purely physicalist view of experiences and feelings. What it shows rather is that some physical qualities or states are irreducibly subjective entities (Searle 1992). Others hold that the explanatory gap may one day be bridged but we currently lack the concepts to bring the subjective and objective perspectives together. On this view, it may turn out that qualia are physical, but we currently have no clear conception as to how they could be (Nagel 1974). Still others adamantly insist that the explanatory gap is, in principle, bridgeable but not by us or by any creatures like us. Experiences and

feelings are as much a part of the physical, natural world as life, digestion, DNA, or lightning. It is just that with the concepts we have and the concepts we are capable of forming, we are cognitively closed to a full, bridging explanation by the very structure of our minds (McGinn 1991).

Another view that has been gaining adherents of late is that there is a real, unbridgeable gap, but it has no consequences for the nature of consciousness and physicalist or functionalist theories thereof. On this view, there is nothing in the gap that should lead us to any bifurcation *in the world* between experiences and feelings on the one hand and physical or functional phenomena on the other. There aren't two sorts of natural phenomena: the irreducibly subjective and the objective. The explanatory gap derives from the special character of phenomenal *concepts*. These concepts mislead us into thinking that the gap is deeper and more troublesome than it really is.

On one version of this view, phenomenal concepts are just indexical concepts applied to phenomenal states via introspection (see Lycan 1996). On an alternative version of the view, phenomenal concepts are very special, first-person concepts different in kind from all others (see Tye forthcoming (b)). This response to the explanatory gap obviously bears affinities to the second physicalist response I sketched in Section III to the Knowledge Argument.

There is no general agreement on how the gap is generated and what it shows.

VI. Qualia and Introspection

In the past, philosophers have often appealed directly to introspection on behalf of the view that qualia are intrinsic, non-intentional features of experiences. Recently, a number of philosophers have claimed that introspection reveals no such qualities (Harman 1990, Dretske 1995, Tye 1995). Suppose you are facing a white wall, on which you see a bright red, round patch of paint. Suppose you are attending closely to the color and shape of the patch as well as the background. Now turn your attention from what you see out there in the world before you to your visual experience. Focus upon *your awareness of* the patch as opposed to *the patch* of which you are aware. Do you find yourself suddenly acquainted with new qualities, qualities that are intrinsic to your visual experience in the way that redness and roundness are qualities intrinsic to the patch of paint? According to some philosophers, the answer to this question is a resounding ‘No’. As you look at the patch, you are aware of certain features out there in the world. When you turn your attention inwards to your experience of those features, you are aware of the *very same* features together with the fact that your mental state is representing them; no new features of your experience over and above its representing red, round, etc. are revealed. In this way, your visual experience is transparent or diaphanous. When you try to examine it, you see right though it, as it were, to the qualities you were experiencing all along in being a subject of the experience, qualities your experience is *of*.

This point holds good even if you are hallucinating and there is no real patch of paint on the wall before you. Still you have an experience *of* there being a patch of paint out there with a certain color and shape. It's just that this time your experience is a misrepresentation. And if you turn your attention inwards to

your experience, you will 'see' right through it again to those very same qualities.

These observations suggest that qualia, the immediately 'felt' qualities of experiences of which we are cognizant when we attend to them introspectively, are *representational* qualities -- qualities like representing red, representing round, representing a red, round shape, and so on. Not everyone agrees that this is what introspection shows (see Block 1991). Perhaps qualia are not presented to us in introspection *as* intrinsic, non-intentional features of our experiences. Still it does not follow from this that we are not introspectively acquainted with intrinsic qualia at all. For we do know on the basis of introspection what it is like to undergo a visual experience of blue, say. So, if what a state is like is a matter of which intrinsic, non-intentional features it tokens, then obviously we are introspectively aware of such features (in the de re sense of 'of'). On this view, whether qualia are intrinsic, non-intentional features of experiences is a theoretical matter. Introspection does not settle the matter one way or the other.

VII. Representational Theories of Qualia

Talk of the ways things look and feel is intensional. If I have a red after-image as a result of a flashbulb going off, the spot I 'see' in front of the photographer's face looks red, even though there is no such spot. If I live in a world in which all and only things that are purple are poisonous, it is still the case that an object that looks purple to me does not thereby look poisonous (in the phenomenal sense of 'looks'). If I feel a pain in a leg, I need not even have a leg. My pain might be a pain in a phantom limb. Facts such as these have been taken to provide further support for the contention that qualia are representational qualities.

If qualia are indeed representational, an important question arises: which aspects of the representational content of an experience are relevant to its phenomenal character, to what it is like to have the experience? Obviously not all aspects of content are phenomenally relevant. If you and I see a telescope from the same viewing angle, for example, then even if I do not recognize it as a telescope and you do (so that our experiences differ representationally at this level), the way the telescope *looks* to both of us is likely pretty much the same (in the phenomenal sense of 'looks'). Likewise, if a child is viewing the same item from the same vantage point, her experience will likely be pretty similar to yours and mine too. Phenomenally, our experiences are all very much alike, notwithstanding certain higher-level representational differences. This, according to some representationalists, is because we all have experiences that represent to us the same 3-D surfaces, edges, colors, and surface-shapes plus a myriad of other surface details.

The representation we share here has a content much like that of the 2 1/2-D sketch posited by David Marr in his famous theory of vision (1982) to which further shape and color information has been appended (for details, see Tye 1995). This content is plausibly viewed as nonconceptual. It forms the output of the early, largely modular sensory processing and the input to one or another system of higher-level cognitive processing. Representationalists sometimes claim that it is here at this level of content that qualia are to be found (see Dretske 1995, Tye 1995; for an opposing representational view, see McDowell 1994).

Representationalists about qualia are typically also externalists about representational content. On this view, what a given experience represents is metaphysically determined at least, in part, by factors in the external environment. Thus, it is usually held, microphysical twins can differ with respect to the representational contents of their experiences. If these differences in content are of the right sort then, according to the wide representationalist, microphysical twins cannot *fail* to differ with respect to the phenomenal character of their experiences. What makes for a difference in representational content in microphysical duplicates is some external difference, some connection between the subjects and items in their respective environments. The generic connection is sometimes called 'tracking', though there is no general agreement as to in what exactly tracking consists.

On wide representationalism, qualia (like meanings) ain't in the head. The classic, Cartesian-based picture of experience and its relation to the world is thus turned upside down. Qualia are not intrinsic qualities of inner ideas of which their subjects are directly aware, qualities that are necessarily shared by internal duplicates however different their environments may be. Rather, they are extrinsic qualities fixed by certain external relations between individuals and their environments.

Representationalism, as I have presented it so far, is an identity thesis with respect to qualia: qualia are supposedly one and the same as certain representational qualities. Sometimes a weaker supervenience thesis is adopted, according to which it is metaphysically necessary that experiences alike with respect to their representational contents are alike with respect to their qualia. Obviously, the supervenience thesis leaves open the further question as to the essential nature of qualia.

Objections to representationalism often take the form of putative counter-examples. One class of these consists of cases in which, it is claimed, experiences have the same representational content but different phenomenal character. Christopher Peacocke adduces examples of this sort in his 1983. According to some (e.g., Block 1990, Shoemaker forthcoming), the Inverted Spectrum also supplies an example that falls into this category. Another class is made up of problem cases in which allegedly experiences have different representational contents (of the relevant sort) but the same phenomenal character. Ned Block's Inverted Earth example (1990) is of this type. The latter cases only threaten strong representationalism, the former are intended to refute representationalism in both its strong and weaker forms. Counter-examples are also sometimes given in which supposedly experience of one sort or another is present but in which there is no state with representational content. Swampman (Davidson 1986) -- the molecule by molecule replica of one of us , formed accidentally by the chemical reaction that occurs in a swamp when a partially submerged log is hit by lightning -- is one such counter-example, according to some philosophers. But there are more mundane cases. Consider the exogenous feeling of depression. That, it may seem, has no representational content. Cases of the third sort, depending upon how they are elucidated further, can pose a challenge to either strong or weaker versions of representationalism.

I lack the space to go through all these objections. I shall discuss briefly just one: Inverted Earth. Inverted Earth is an imaginary planet, on which things have complementary colors to the colors of their counterparts on Earth. The sky is yellow, grass is red, ripe tomatoes are green, and so on. The inhabitants of Inverted Earth undergo psychological attitudes and experiences with inverted intentional contents

relative to those of people on Earth. They think that the sky is yellow, see that grass is red, etc. However, they call the sky ‘blue’, grass ‘green’, ripe tomatoes ‘red’, etc. just as we do. Indeed, in all respects consistent with the alterations just described, Inverted Earth is as much like Earth as possible.

In Block's original version of the tale, mad scientists insert color-inverting lenses in your eyes and take you to Inverted Earth, where you are substituted for your Inverted Earth twin or doppelganger. Upon awakening, you are aware of no difference, since the inverting lenses neutralize the inverted colors. You think that you are still where you were before. What it is like for you when you see the sky or anything else is just what it was like on earth. But after enough time has passed, after you have become sufficiently embedded in the language and physical environment of Inverted Earth, your intentional contents will come to match those of the other inhabitants. You will come to believe that the sky is yellow, for example, just as they do. Similarly, you will come to have a visual experience that represents the sky as yellow. For the experiential state you now undergo, as you view the sky, is the one that, in you, now normally tracks yellow things. So, the later you will come to be subject to inner states that are intentionally inverted relative to the inner states of the earlier you, while the phenomenal aspects of your experiences will remain unchanged. It follows that strong representationalism of the externalist sort is false.

Perhaps the simplest reply that the strong representationalist can make with respect to this objection is to deny that there really is any change in normal tracking with respect to color, at least as far as your experiences go. "Normal", after all, has both teleological and nonteleological senses. If what an experience normally tracks is what nature designed it to track, what it has as its biological purpose to track, then shifting environments from Earth to Inverted Earth will make no difference to normal tracking and hence no difference to the representational contents of your experiences. The sensory state that nature designed in your species to track blue in the setting in which your species evolved will continue to do just that even if through time, on Inverted Earth, in that alien environment, it is usually caused in you by looking at yellow things.

The suggestion that tracking is teleological in character, at least for the case of basic experiences, goes naturally with the plausible view that states like feeling pain or having a visual sensation of red are phylogenetically fixed (Dretske 1995). However, it encounters serious difficulties with respect to the Swampman case mentioned above. On a cladistic conception of species, Swampman is not human. Indeed, lacking any evolutionary history, he belongs to no species at all. His inner states play no teleological role. Nature did not design any of them to do anything. So, if phenomenal character is a certain sort of teleo-representational content, then Swampman has no experiences and no qualia.

There are alternative replies available to the strong representationalist (see Lycan 1996, Tye 1998) in connection with the Inverted Earth problem. These involve either denying that qualia do remain constant with the switch to Inverted Earth or arguing that a non-teleological (but still wide) account of sensory content may be elaborated, under which qualia stay the same.

VIII. Which Creatures Undergo States with Qualia?

Do frogs have qualia? Or fish? What about honey bees? Somewhere down the phylogenetic scale phenomenal consciousness ceases. But where? It is sometimes supposed that once we begin to reflect upon much simpler beings than ourselves -- snails, for example -- we are left with nothing physical or structural that we could plausibly take to help us determine whether they are phenomenally conscious (Papineau 1994). There is really *no* way of our knowing if spiders are subject to states with qualia, as they spin their webs, or if fish undergo any phenomenal experiences, as they swim about in the sea.

Representationalism has the beginnings of an answer to the above questions. If qualia are, by their very nature, qualities of states that carry information about certain features, internal or external, states that form the outputs of sensory modules and stand ready and available to make a direct difference to beliefs and desires, then creatures that are incapable of reasoning, of changing their behavior in light of assessments they make, based upon information provided to them by sensory stimulation of one sort or another, are not phenomenally conscious. Tropistic organisms, on this view, feel and experience nothing. They have no qualia. They are full-fledged unconscious automata or zombies, rather as blindsight subjects are restricted unconscious automata or partial zombies with respect to a range of visual stimuli.

Consider, for example, the case of plants. There are many different sorts of plant behavior. Some plants climb, others eat flies, still others catapult out seeds. Many plants close their leaves at night. The immediate cause of these activities is something internal to the plants. Seeds are ejected because of the hydration or dehydration of the cell walls in seed pods. Leaves are closed because of water movement in the stems and petioles of the leaves, itself induced by changes in the temperature and light. These inner events or states are surely not phenomenal. There is nothing it is like to be a Venus Fly Trap or a Morning-Glory.

The behavior of plants is inflexible. It is genetically determined and, therefore, not modifiable by learning. Natural selection has favored the behavior, since historically it has been beneficial to the plant species. But it need not be now. If, for example, flies start to carry on their wings some substance that sickens Venus Fly Traps for several days afterwards, this will not have any effect on the plant behavior with respect to flies. Each Venus Fly trap will continue to snap at flies as long as it has the strength to do so.

Plants do not learn from experience. They do not acquire beliefs and change them in light of things that happen to them. Nor do they have any desires. To be sure, we sometimes speak as if they do. We say that the wilting daffodils are just begging to be watered. But we recognize full well that this is a harmless facon de parler. What we mean is that the daffodils *need* water. There is here no goal-directed behavior, no purpose, nothing that is the result of any learning, no desire *for* water.

Plants, on the representational view, are not subject to any qualia. Nothing that goes on inside them is poised to make a direct difference to what they believe or desire, since they have no beliefs or desires.

Reasoning of the above sort can be used to make a case that even though qualia do not extend to plants and paramecia, qualia are very widely distributed in nature (see Tye 1997). Of course, such a case

requires decisions to be made about the attribution of beliefs and desires to much simpler creatures. And such decisions are likely to be controversial in some cases. Moreover, representationalism itself is a very controversial position. The general topic of the origins of qualia is not one on which philosophers have said a great deal.

Bibliography

- Block, N. 1980 "Troubles with Functionalism," in *Readings in the Philosophy of Psychology*, Volume 1, Ned Block, ed., Cambridge, Mass : Harvard University Press, 268-305.
- Block, N. 1990 "Inverted Earth," *Philosophical Perspectives*, 4, J. Tomberlin, ed., Northridge: Ridgeview Publishing Company.
- Block, N. 1996 "Mental Paint and Mental Latex," *Philosophical Issues*, 7, E. Villenueva, ed., Northridge: Ridgeview Publishing Company.
- Chalmers, D. 1996 *The Conscious Mind*, Oxford: Oxford University Press.
- Churchland, P. 1985 "Reduction, Qualia, and Direct Introspection of Brain States," *Journal of Philosophy*, 82, 8-28.
- Davies, M. and Humphreys, G. 1993 *Consciousness*, Blackwells: Oxford.
- Davidson, D. 1986 "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association*, 60, 441-458.
- DeBellis, M. 1991 "The Representational Content of Musical Experience," *Philosophy and Phenomenological Research*, 51, 303-324.
- Dennett, D. 1990 "Quining Qualia," in *Mind and Cognition*, W. Lycan, ed., Oxford : Blackwells, 519-548.
- Dennett, D. 1991 *Consciousness Explained*, Boston : Little, Brown and Company.
- Dretske, F. 1995 *Naturalizing the Mind*, Cambridge, Mass: The MIT Press, Bradford Books.
- Harman, G. 1990 "The Intrinsic Quality of Experience," in *Philosophical Perspectives*, 4, J. Tomberlin, ed., Northridge: Ridgeview Publishing Company.
- Hardin, C. 1993 *Color for Philosophers*, Cambridge : Hackett.
- Harrison, B. 1973 *Form and Content*, Oxford: Blackwells.
- Haugeland, J. 1985 *Artificial Intelligence: The Very Idea*, Cambridge, Mass : The MIT Press, Bradford Books.
- Hill, C. 1991 *Sensations : A Defense of Type Materialism*, Cambridge : Cambridge University Press.
- Horgan, T. 1984 "Jackson on Physical Information and Qualia," *Philosophical Quarterly*, 34, 147-83.
- Jackson, F. 1982 "Epiphenomenal Qualia," *Philosophical Quarterly*, 32, 127-136.
- Jackson, F. 1993 "Armchair Metaphysics," in *Philosophy of Mind*, ed. by J. O'Leary-Hawthorne and M. Michael, (Kluwer Books).
- Kripke, S. 1972 "Naming and Necessity," in *Semantics of Natural Language*, ed. by D. Davidson and G. Harman, Dordrecht, Holland: Reidel, 253-355.
- Levine, J. 1983 "Materialism and Qualia : The Explanatory Gap," *Pacific Philosophical Quarterly*, 64, 354-361.

- Lewis, D. 1990 "What Experience Teaches," in *Mind and Cognition: A Reader*, ed. by W. Lycan, (Oxford: Blackwells).
- Loar, B. 1990 "Phenomenal States," in *Philosophical Perspectives*, 4, J. Tomberlin, ed., Northridge : Ridgeview Publishing Company.
- Lycan, W. 1987 *Consciousness*, Cambridge, Mass : The MIT Press.
- Lycan, W. 1996 *Consciousness and Experience*, Cambridge, Mass: The MIT Press, Bradford Books.
- McDowell, J. 1994 "The Content of Perceptual Experience," *Philosophical Quarterly*.
- McGinn, C. 1991 *The Problem of Consciousness*.
- Marr, D. 1982 *Vision*, San Francisco: W.H. Freeman and Company.
- Moore, G. E. 1922 "The Refutation of Idealism," in his *Philosophical Studies*, London : Routledge and Kegan Paul.
- Nagel, T. 1974 "What is it like to be a Bat?" *Philosophical Review*, 83, 435-456.
- Nemirow, L. 1990 "Physicalism and the Cognitive Role of Acquaintance," in *Mind and Cognition: A Reader*, ed. by W. Lycan (Oxford: Blackwells).
- Papineau, D. 1994 *Philosophical Naturalism* (Oxford: Blackwells).
- Peacocke, C. 1983 *Sense and Content*, Oxford: Oxford University Press.
- Raffman, D. 1995 "On the Persistence of Phenomenology," in *Conscious Experience*, ed. by T. Metzinger, Schoningh-Verlag Publishers.
- Rey, G. 1992 "Sensational Sentences Switched," *Philosophical Studies*, 68, 289-319.
- Sacks O. 1996 *The Island of the Colorblind* (Alfred A. Knopf).
- Searle, J. 1992 *The Rediscovery of Mind* Cambridge, Mass : The MIT Press, Bradford Books.
- Shoemaker, S. 1975 "Functionalism and Qualia," *Philosophical Studies*, 27, 291-315.
- Shoemaker, S. 1982 "The Inverted Spectrum," *Journal of Philosophy*, 79, 357-381.
- Shoemaker, S. 1990 "Qualities and Qualia : What's in the Mind," *Philosophy and Phenomenological Research*, 50, Supplement, 109-131.
- Shoemaker, S. forthcoming "Two Cheers for Representationalism," *Philosophy and Phenomenological Research*.
- Strawson, G. 1994 *Mental Reality*, Cambridge, Mass: the MIT Press, Bradford Books.
- Tye, M. 1986 ‘The Subjective Qualities of Experience’, *Mind* 95, 1-17.
- Tye, M. 1995 *Ten Problems of Consciousness*, Cambridge, Mass: The MIT Press, Bradford Books.
- Tye, M. 1997 "The Problem of Simple Minds: Is There Anything it is Like to be a Honey-bee?", *Philosophical Studies*.
- Tye, M. 1998 "Inverted Earth, Swampman, and Representationism," *Philosophical Perspectives*, 12, J. Tomberlin, ed. Ridgeview Publishing Company.
- Tye, M. forthcoming (a) "Knowing What it is Like: the Ability Hypothesis and the Knowledge Argument," *Protosociologie*.
- Tye, M. forthcoming (b) "Phenomenal Consciousness: the Explanatory Gap."
- White, S. 1995 "Color and the Narrow Contents of Experience," *Philosophical Topics*.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

consciousness

Michael Tye
tye@vm.temple.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 20, 1997
Content last modified: November 1, 1997

Epiphenomenalism

Epiphenomenalism is the view that mental events are caused by physical events in the brain, but have no effects upon any physical events. Behavior is caused by muscles that contract upon receiving neural impulses, and neural impulses are generated by input from other neurons or from sense organs. On the epiphenomenalist view, mental events play no causal role in this process. They are like a steam whistle that contributes nothing to the work of a locomotive (Huxley, 1874). Mental events do not affect the brain activity that produces them "any more than a shadow reacts upon the steps of the traveller whom it accompanies" (James, 1879).

Epiphenomenalism arose in a 19th Century context in which a dualistic view of mental events was assumed to be correct. The first part of our discussion -- Traditional Arguments -- will be phrased in a style that reflects this dualistic presupposition. By contrast, many contemporary discussions work within a background assumption of the preferability of materialist monism. One might have supposed that this position would have put an end to the need to investigate epiphenomenalism; but, as we shall see under Arguments in the Age of Materialism, such a supposition is far from being the case. A brief outline of both discussions follows.

Many philosophers recognize a distinction between two kinds of mental events. (A) The first goes by many names, e.g., phenomenal experiences, occurrences of qualitative consciousness, the what-it-is-like of experience, qualia. Pains, colors of afterimages, timbres of auditory events, and the qualities of the experience of odors and tastes, can serve as examples. (B) Mental events of the second kind are occurrent propositional attitudes, e.g., (occurrent) beliefs and desires. Arguments about epiphenomenalism may concern either type of mental event, and it should not be assumed that an argument given for one type can be rephrased without loss for the other. The two types can often be connected, however, through beliefs that one has one's qualia. Thus, if it is held that pains have no physical effects, then one must say either (i) pains do not cause beliefs that one is in pain, or (ii) beliefs that one is in pain are epiphenomenal. For, if pains caused beliefs that one is in pain, and the latter had physical effects, then pains would, after all, have effects in the physical world (albeit indirectly). But epiphenomenalism says mental events have *no* effects in the physical world.[1]

Traditional Arguments (A) Pro

The central motivation for epiphenomenalism lies in the premise that all physical events have sufficient causes that lie within the class of physical events. If a mental event is something other than a physical event, then for it to make any causal contribution of its own in the physical world would require a violation of physical law. Descartes' (1649) interactionist model proposed that nonphysical events could cause small changes in the shape of the pineal gland. But such nonphysical effects, however slight, would mean that the physical account of motion is *false* -- for that account says that there will be no such change of shape unless there is a *physical* force that causes it.

One may try to rescue mental efficacy by supposing that whenever there is a mental effect in the physical world there is *also* a physical force that is a sufficient cause of the effect. This view, however, both offends Occamist principles and fails to satisfy the leading anti-epiphenomenalist intuition, namely, that the mental *makes a difference* to the physical, i.e., that it leads to behavior that would not have happened in absence of the mental. The view also leads to an epistemological problem: If there is always a sufficient physical cause for behavior, then one could never be in a position where one *needs* to suppose there is anything further. Thus, on the assumption of physical sufficiency, there could never be any reason to introduce mental causes into one's account of behavior.

Many contemporary thinkers would respond to the central motivation for epiphenomenalism by denying its dualistic presupposition, i.e., by holding that mental events are identical with physical events, and may therefore have physical effects. Discussion of this type of response will be given below in the Remarks on Kim's Way Out. Here it may be observed, however, that the argument stated in the previous two paragraphs is not supposed to be an argument for dualism, but only for adopting epiphenomenalism, once dualism is accepted.

Further support for epiphenomenalism can be derived from the fact, noted by Wilhelm Wundt (1912),

that "each simple sensation is joined to a very complicated combination of peripheral and central nerve processes", together with the fact that the causes of behavior are likewise complex neural events. This latter fact makes it natural to look for complex events throughout the causal chain leading to behavior; and these can be found in the neural events that are required for the occurrence of simple sensations. The sensations themselves could not contribute to behavior without first having neural effects that are more complex than themselves. Thus an anti-epiphenomenalist stance would require us to prefer the hypothesis that simple sensations cause (relatively) complex neural events over the hypothesis that complex neural events (that are required in any case for the causation of sensations) are adequate to cause the neural events required for the causation of behavior.

Traditional Arguments (B) Con (with Epiphenomenalists' Replies)

(1) Obvious Absurdity

Epiphenomenalism is absurd; it is just plain obvious that our pains, our thoughts, and our feelings make a difference to our (evidently physical) behavior; it is impossible to believe that all our behavior could be just as it is even if there were no pains, thoughts, or feelings. (Taylor, 1963 and subsequent editions, offers a representative statement.)

This argument is surely the briefest of those against epiphenomenalism, but it may have been more persuasive than any other. Epiphenomenalists, however, can make the following reply. First, it can never be obvious what causes what. Animated cartoons are full of causal illusions. Falling barometers are regularly followed by storms, but do not cause them. More generally, a regularity is causal only if it is not explained as a consequence of underlying regularities. It is part of epiphenomenalist theory, however, that the regularities that we observe to hold between mental events and actions can be explained by underlying regularities. Schematically, suppose physical event P1 causes both mental event M and physical successor P2, as in Figure 1.

```
 M
 |
 P1 --> P2 --> P3 --> ...

(Figure 1)
```

Suppose there is no other cause of M, and no other cause of P2. Then every M will be followed by P2, yet the cause of P2 will be adequately found in P1. It is true that, under the assumptions stated, the counterfactual, "If M had not occurred, then P2 would not have occurred" holds; but then, so may "if the barometer had not fallen, the storm would not have occurred." The moral to be drawn is that causation may imply that certain counterfactuals hold, but the holding of counterfactuals is not enough to show causation. So, the fact that under normal conditions, some of our actions would not have occurred unless

we had had certain mental events cannot show that those actions are caused by our mental events (rather than being caused by the physical causes of those mental events).

It is often said that pains cause withdrawals of affected parts of the body. In extreme cases, however -- for example in a case of touching a hot stove -- it can be observed that the affected part is withdrawn *before* the pain is felt. These cases cannot show that pain never causes withdrawals, but they do show that pain is not necessary as a cause of withdrawals. In less extreme cases, it is open to the epiphenomenalist to hold that the causal order is the same as in the extreme cases (i.e., some physical event, P1, causes both withdrawal and pain) but is not ordinarily recognized to be so.

(2) Natural Selection

The development of consciousness must be explainable through natural selection. But a property can be selected for only if it has an effect upon organisms' behavior. Therefore, consciousness (both qualia and intentional states) must have effects in behavior, i.e., epiphenomenalism is false. (Today, this argument is generally associated with Popper and Eccles, 1977. It is an old argument, however, and clear statements of it were offered by James (1879) and by Romanes in 1882 (see Romanes, 1896).)

According to the same biology that embraces natural selection, however, behavior has muscular causes, which in turn have neural causes. Barring neural events that are inexplicably in violation of biological constraints on their conditions of activation, there must be an adequate physical cause of every link in the causal chain leading to behavior. Thus, it is easily understood how certain kinds of neural events can be selected for. Epiphenomenalists hold that conscious events are effects of (certain) neural events. Thus, it fits well in their view that we have the conscious events we do because the neural causes of these events have been selected for. Indeed, if neural causes of behavior are selected for, and are sufficient causes, there *cannot* be any *further* effect attributed to natural selection.

William James (1879; 1890) offered an intriguing variant of the argument from natural selection. If pleasures and pains have no effects, there would seem to be no reason why we might not abhor the feelings that are caused by activities essential to life, or enjoy the feelings produced by what is detrimental. Thus, if epiphenomenalism (or, in James' own language, automaton-theory) were true, the felicitous alignment that generally holds between affective valuation of our feelings and the utility of the activities that generally produce them would require a special explanation. Yet on epiphenomenalist assumptions, this alignment could not receive a genuine explanation. The felicitous alignment could not be selected for, because if affective valuation had no behavioral effects, misalignment of affective valuation with utility of the causes of the evaluated feelings could not have any behavioral effects either. Epiphenomenalists would simply have to accept a brute and unscientific view of pre-established harmony of affective valuation of feelings and the utility of their causes.

Epiphenomenalists can meet James' argument, however, by supposing that both the pleasantness of pleasant feelings and the feelings themselves depend on neural causes (and analogously for painfulness and disliked qualities). So long as both types of neural events are efficacious in the production of

behavior, their combination can be selected for, and thus the felicitous alignment of feelings with evaluation can be explained. Moreover, the supposition that the neural causes of both evaluation and feeling qualities should have behavioral effects is independently plausible: on grounds of natural selection, there should be both a preference system for quick action and a system that fosters discriminability, for use in longer term planning; and these must, in general, work together in a successful organism.

(3) Knowledge of Other Minds

Our reason for believing in other minds is inference from behavioral effects to mental event causes. But epiphenomenalism denies such a causal connection. Therefore, epiphenomenalism implies the (exceedingly implausible) conclusion that we do not know that others have mental events. (Jackson, 1982, replies to this and several other arguments against epiphenomenalism. The argument is stated, and accepted, by Benecke, 1901.)

The first premise of this argument is a widely held dogma, but epiphenomenalists can deny it without evident absurdity. It is perfectly obvious to everyone that the bodies of human beings are very much alike in their construction, and it requires no sophisticated reasoning to infer that if others are made like me, they probably hurt when affected like me, e.g., when their bodies are stuck with pins, beaten, cut and so on. There is no principle that makes an inference from similar effects to similar causes more secure than an inference from similar causes to similar effects; on the contrary, the latter inference is more secure, because there can sometimes be different causes of undetectably similar effects. Thus, an inference to other minds that is allowed by epiphenomenalism must be at least as strong as the inferential route to other minds with which it is incompatible.

(4) Self-stultification

The most powerful reason for rejecting epiphenomenalism is the view that it is incompatible with knowledge of our own minds -- and thus, incompatible with knowing that epiphenomenalism is true. (A variant has it that we cannot even succeed in referring to our own minds, if epiphenomenalism is true.) If these destructive claims can be substantiated, then epiphenomenalists are, at the very least, caught in a practical contradiction, in which they must claim to know, or at least believe, a view which implies that they can have no reason to believe it. (Many authors offer this objection in one version or another. A full statement of this argument, and several others concerning epiphenomenalism, can be found in Chalmers, 1996.)

The argument that is given to support the destructive claims is that (i) knowledge of one's mental events requires that these events cause one's knowledge, but (ii) epiphenomenalism denies physical effects of mental events. So, either we cannot know our own mental events, or our knowledge of them cannot be what is causing the plainly physical event of our saying something about our mental events. Thus, suppose S is an epiphenomenalist, and that S utters "I am in terrible pain." S is committed to the view that the pain does not cause the utterance. But then, it seems, S would be making the same utterance whether

or not a pain were occurring. If this is so, then S's testimonies about S's own pains are worthless -- both to us and to S. They cannot be taken to represent any knowledge about pains on S's part (if S's epiphenomenalist view is true). In fact, on an epiphenomenalist view, all the arguments for epiphenomenalism and rebuttals to counterarguments we have reviewed might be given even if we were all zombies -- i.e., even if we were all possessed of physical causes of our utterances and completely devoid of any mental life whatsoever.

The argument that epiphenomenalism is self-stultifying in the way just described rests on the premise that knowledge of a mental event requires causation by that mental event. But epiphenomenalists may reject that premise without absurdity. One way of seeing how to do this involves considering the interactionist diagram in Fig.2, which shows P1 as directly causing M but not P2, and M directly causing P2. (Directly causing is an intransitive relation. Causation (when used without modifier) is transitive: events are causally related if there is a chain of direct causes, however long, that connects them.)

```
M
 \
| C
   \
P1  P2 --> P3 --> ....
```

(Figure 2)

Now consider P3, which is directly caused by P2 and which we will assume to cause (directly or indirectly) further behavior such as S's utterance of "I am in terrible pain". P3 is not directly caused by M. Does it convey knowledge of M? If we answer negatively, on the ground that P3 is not directly caused by M, we will be rejecting interactionism for virtually the same reason that epiphenomenalism is thought to be unacceptable. Since this is an extremely implausible stance, let us take it that P3 does convey knowledge of M. But what property does P3 actually have that makes it a case of conveying knowledge of M? Epiphenomenalists will wish to point out that P3 does not have any property that contains information as to how it was caused. Looking backward from P3, so to speak, one cannot tell whether it was indirectly caused by M (as in the interactionistic Fig. 2) or indirectly caused by M's cause (as in the epiphenomenalistic Fig. 1). There is, however, a property that P3 does have that is intuitively strongly connected to its conveying knowledge of M -- namely, that it would not be occurring unless M had recently occurred. But P3 has this property on epiphenomenalist and interactionist views alike. Thus, if not occurring unless M has recently occurred is the property that is responsible for P3's conveying knowledge of M, epiphenomenalists have as much right as anyone to claim that P3 conveys knowledge of M, and they are not debarred from knowing what they claim to know.

Critics of epiphenomenalism can of course point out that there is a property that interactionism, but not epiphenomenalism, assigns to P3 -- namely, the property of being indirectly caused by M. Epiphenomenalists, however, are likely to think that the intuitive connection between this property and knowledge is much weaker than that between knowledge of M and the fact that P3 would not be occurring unless M had recently occurred. In fact, they may hold that the relevance of indirect causation

is exhausted by its ensuring that P3 would not occur unless M had recently occurred. They can then reiterate that there is another way of ensuring that this condition holds, namely, the set of relations, diagrammed in Fig. 1, that is affirmed by epiphenomenalism.

The foregoing way of responding to the self-stultification argument is further explained and defended in Robinson (1982b). An alternative response can be found in Chalmers (1996). Chalmers' property-dualistic view holds that there is more to a person than just a brain and a body. It allows for persons to be directly acquainted with experiences, and it is this direct acquaintance, rather than any causal relation, that justifies our beliefs about experiences. On this view, experiences are partially constitutive of beliefs about experiences, and "the justification of my belief [about experiences] accrues not just in virtue of my physical features but in virtue of some of my nonphysical features -- namely the experiences themselves" (Chalmers, 1996, p. 198). In supplying non-causal relations to support the claim to knowledge of experiences, this view disconnects the knowledge question from the question of how things stand causally, and thus avoids the self-stultification argument.

Arguments in the Age of Materialism

One might have thought that if the mental and the physical are identical, there could be no room for epiphenomenalistic questions to arise. Behavior is caused by muscular events, and these are caused by neural events. Mental events will be identical with some of these neural events; so whatever effects these neural events have will be effects of mental events, and mental events will make a causal contribution to, i.e., will "make a difference" to our behavior.

Questions about epiphenomenalism, however, arise the moment any distinction is made between the mental properties and the physical properties of an event. There are several ways of doing this within a broadly materialist monism. The following discussion will briefly indicate two ways in which this distinction can be made, and the kind of epiphenomenalistic questions that ensue.

It should be noted that most recent writers take a somewhat dogmatic position against epiphenomenalism. They presume that epiphenomenalism is to be avoided, and they go to great lengths to try to show that they have avoided incurring that anathema, despite maintaining the sufficiency of physical causation in conjunction with some kind of distinction between the mental and the physical.

Two Routes to Puzzlement: Anomalous Monism and Externalism

(1) Donald Davidson's (1970) anomalous monism holds that (i) each mental event is identical with a physical event, but (ii) there are no psychophysical laws. To introduce a convenient example, it is plausible that each person's (occurrent) belief that the Mona Lisa is in the Louvre is identical with some physical event in that person's brain; but there is no one physical brain-event type such that all who hold that belief have that kind of brain event. Now suppose Jones believes that the Mona Lisa is in the Louvre, and is heading toward the Louvre. What causes Jones's motion? Let us first fix a background that consists

of neural events that are identical with Jones's desire to see the Mona Lisa, Jones's understanding of certain maps of Paris, and so forth. These neural events interact with the neural events that are identical with Jones's belief about Mona's location, with neural events that are identical with Jones's perceptions of sidewalks, obstacles, street signs, and so on, and these neural events eventually produce the muscular events that cause Jones's motion toward the Louvre. It seems that if we fill in this sketch, we will have a complete causal explanation of Jones's progress. But such a completely filled-in sketch will contain reference only to neural events, and the mental types with which they are identical will not be so much as mentioned. Moreover, the denial of psychophysical laws means that there is no law-like connection between any psychological type and the neural types that do the causal work. In short, despite the identity of mental and physical events, the mental character of Jones's brain events seems to have nothing to do with where Jones goes.

(2) The same point can be reached through a somewhat different route. Many philosophers hold an *externalist* view of intentionality, according to which intentionality requires representation, and representation depends on circumstances external to the body of the representing subject. (To illustrate with Putnam's (1975) famous example from Twin Earth, what a thought that S might express by "This is water" is actually about depends on what the transparent, tasteless liquid in S's environment actually is.) It seems, however, that the causal determinants of S's behavior can depend only on events occurring inside S's body. Thus, if externalism is right, what S does cannot depend on S's thoughts.

This conclusion is compatible with holding that a proper *description* of S's behavior should refer to circumstances external to S. For example, describing S as *reaching for a glass of water* may not be appropriate unless S believes the glass contains water; and that this is what S believes may depend on circumstances external to S. But then, it is at least tempting to conclude that it cannot be S's belief (at least not S's belief fully described in intentional terms) that is causing the extension of S's arm toward the glass. After all, S's twin-earth double has, in some sense, a different belief (one that refers to XYZ) but the internal bodily story of the causation of the double's arm-extension will be exactly the same as the story for S (up to substitution of XYZ in the double's body wherever S has an H_2O molecule).

Kim's Way Out

In a well-known series of essays, now published as Kim (1993), Jaegwon Kim has offered a view that attempts to reconcile the closure of the physical (the view that every physical event has a set of causes that are both sufficient and physical) with our intuitions about the efficacy of the mental. It is not appropriate here to attempt a complete discussion of Kim's carefully drawn distinctions and thorough argumentation. We may, however, describe a crude example (not Kim's own) and express some of Kim's cardinal points in its terms.

Consider the world of pumps. These come in different kinds -- piston, diaphragm, rotary, etc. The status of an object as a pump arguably depends on its being part of a larger system. For example, we can use a fan to pump some noxious gas from one room to another, or to the outside of a building; but if we set a fan on the lawn, it may beat the air, but is not (functioning as) a pump. Pumps have causal properties --

e.g., they can move fluids. They have these causal properties in virtue of the properties of their parts; e.g., the imperviousness (to gas molecules) and angle of the fan blade cause gas molecules that hit it to move in certain directions.

The properties just listed match properties of mental events, on many views of the latter. Thus, the same belief (i.e., belief in the truth of the same proposition) may be realized as different neural states in different people (or the same person at different times); an occurrence of a belief in a (part of a) brain would not have the causal properties it is often thought of as having unless it were surrounded by a functioning whole brain and body; and a belief would not have the causal properties it is often thought of as having unless it had a neural constitution that could affect other neural (and eventually muscular) events.

Now, let us imagine a "pump epiphenomenalist" (P-Eist), who argues as follows. "Pumps have no efficacy. Of course, objects are caused to be pumps, i.e., their being pumps is a consequence of their parts having the properties they have and standing in the relations in which they stand. But all the causal work of pumps is explained by the properties and relations of the parts. It would be ludicrous to say that the movement of fluids has *two* causes, i.e., the push of the blades *and* the pumping. Explanation of the movements of fluids in terms of such things as rigidity and motion of the pump's parts *has* to be mentioned in any full account of those movements, and this kind of explanation explains everything that needs to be explained; thus the explanation in terms of rigidity and motion of the pump's parts *excludes* the property of being a pump from any explanatory role. Our intuition that pumps move fluids is just an illusion. Being a pump is a mere epiphenomenal property."

What the P-Eist says about pumps is in every way correct -- except for the conclusion. Thus, if mental event properties are analogous to the property of being a pump, then all the reasons that lead to epiphenomenalism can be granted, without granting the epiphenomenalist conclusion. Or, to put the matter somewhat differently, if epiphenomenalism leaves mental event properties no worse off than the property of being a pump, then we can accept epiphenomenalism of mental event properties with equanimity. Despite the points made in the preceding paragraph, we will not feel that we are making a mistake when we say that pumps move fluids; and no more should we feel that we are making a mistake if we say that S reaches for the glass because S believes it contains water.

Remark on Kim's Way Out

As Kim is well aware, this attractive solution depends on a key point in the analogy between mental properties and the property of being a pump. We feel it is idle to deny that pumps move fluids because pumps are *realized* in (or by) the system of parts whose properties and relations figure in the fully detailed causal explanation of the movement of fluids. Because there can be many kinds of pumps, we cannot *identify* being a pump with any of the configurations that realize pumps. But we have insight into why it is that a given configuration of parts, surrounded by a certain kind of system, is *necessarily* a pump. If we can say that what causes the movement of certain fluids *had* to be a pump, the claim that being a pump has nothing to do with the movement would seem to be merely a tendentious and

misleading description of the circumstances.

Now, there are some mental properties for which it is plausible that the key analogy just described holds. These are the properties for which functionalism is most plausible, e.g., beliefs and desires. That is, it does not seem out of the question that a view can be sustained according to which S's neural condition at time t, in S's surrounding's at t (and, perhaps, given S's history) *necessarily constitutes* S as a believer of p, or a desirer of q. If such an account can be made out, then there is a robust sense to the idea that beliefs and desires are *realized by* the physical conditions just listed; that they are thus no worse off than pumps in the realm of efficacy; and that there will be no violation of our intuitions that S's actions occur because of what S believes and desires.

But there are other mental properties that resist the kind of story that seems plausible for beliefs and desires. As Kim again explicitly recognizes (1993, p. 366), these are the properties that have long caused difficulties for functionalism, namely, the qualities of phenomenal experiences such as pains, itches, tastes, smells, afterimages, and so on. The "explanatory gap" (Levine, 1983) or "unintelligibility" (Robinson, 1982a) in the connection between neural events and phenomenal qualities can be otherwise expressed as our inability to see any necessity in that connection. We are unable to understand why it should be that a series of neural activations occurring in various degrees of intensity and temporal relations should always be accompanied by pain, or itch, or, indeed, by any phenomenal quality whatever. Inability to see any such necessity is, of course, not a proof that such a necessity does not obtain. Nonetheless, absent insight into the necessity of the connection between neural properties and qualitative properties, we are arguably in an *explanatory* position similar to traditional epiphenomenalism. That is, we will have a sufficient explanation of behavioral reactions to stimuli that invokes exclusively neural properties. In addition, we may hold the view that these neural properties are necessarily connected to qualitative properties; but, lacking explanation of this necessity, this connection will contribute no understanding of how qualitative properties could make a difference to behavior. Because this difficulty has not been removed in the case of qualia, the success or failure of the previously discussed Traditional Arguments remains relevant to contemporary thinking about epiphenomenalism.

Epiphenomenalism and Intrinsic Properties

Frank Jackson (1982) has given an epiphenomenalistic argument that has spawned lively responses from many quarters. This argument turns on the concept of physical information, where "physical information" is information about ourselves and our world of the kind that is obtainable in the physical, chemical, and biological sciences. In Jackson's argument, a brilliant scientist, Mary, has learned all the physical information there is about color vision. Having been confined to a black and white room, however, Mary has never had a color experience. Jackson asks whether Mary will learn anything when she is released from her confinement and thus comes for the first time to have color experiences. It seems compelling that she would learn something; but as she already has all the physical information there is, what she learns must be some other kind of information, which we may call "phenomenal information". This "knowledge argument" has been regarded as a strong reason to accept a dualistic view of our experiences. When combined with the traditional arguments (Pro) given above, it becomes a potent source of support

for epiphenomenalism.

David Lewis (1988) undertakes a thorough response to the knowledge argument. Among Lewis's many considerations, there is one that seeks to enforce a connection between phenomenal information *per se* and epiphenomenalism. According to Lewis's argument, even if one says that phenomenal *events* are identical with physical events, and even if one says that phenomenal events produce physical effects in violation of physical laws, one will still be led to a form of epiphenomenalism if one says there is phenomenal information that is irreducibly different from physical information. To put the argument in ruthlessly summary form, let V1 and V2 be two possibilities for the phenomenal information that one acquires by, and only by, tasting Vegemite. Suppose that P1 is a physical state produced by the taste of Vegemite. That the taste of Vegemite has this physical effect is a piece of physical information. But this same physical information is compatible with two possibilities, (a) V1 is related by a law, L1, to P1; and (b) V2 is related by a different law, L2, to P1. Now, either of these possibilities is compatible with all the physical information we have; i.e., their difference makes no physical difference. Thus, that the phenomenal information in the taste of Vegemite is, say, V1 rather than V2 can make no difference to anything physical, i.e., V1 is epiphenomenal. Lewis's point here is not to argue for or against epiphenomenalism; rather, he assumes epiphenomenalism is false, and uses the fact that the hypothesis of phenomenal information leads to it as an argument against that hypothesis.

Denis Robinson (1993) raises the possibility that Lewis's argument can be extended to produce a far-reaching and puzzling result. Suppose that I1 and I2 are two possibilities for an *intrinsic* property of a basic physical entity, e.g., a quark. Everything relevant to physics can be expressed by the lawlike relations in which quarks stand to fundamental physical objects and properties. Let this set of relations be [S]. It appears that there are two possibilities, (a) I1 is related by a set of laws, [L1], to [S]; or (b) I2 is related by a different set of laws, [L2], to [S]. Either of these possibilities is compatible with all the physics we have, i.e., their difference makes no physical difference. Thus, that the intrinsic property of quarks is, say, I1 rather than I2 can make no difference to physics, i.e., I1 is epiphenomenal. The generalization of this point is that the intrinsic properties of the fundamental objects of physics must be epiphenomenal.

It thus appears that we must either (1) deny that fundamental objects of physics have any intrinsic properties, or (2) deny that Lewis's argument for the connection of phenomenal information with epiphenomenalism is sound, or (3) deny that Lewis's argument can be paralleled in the suggested way for the case of intrinsic physical properties, or (4) admit an epiphenomenalism of intrinsic properties into our view of the basic structure of physical reality.

Bertrand Russell (1927, p. 382) held the view that physical theory can reveal only causal structure, or "formal properties" of matter, and that "by examining our percepts we obtain knowledge which is not purely formal as to the matter of our brains." This idea is taken up sympathetically (with substantial reworking in a quantum mechanical context) by Lockwood (1993). Chalmers (1996) offers a useful discussion of the view, and expresses some sympathy for it. Denis Robinson (1993), however, regards intrinsic similarity of fundamental physical entities as different from similarity of phenomenal properties.

If phenomenal properties are intrinsic properties of fundamental physical objects, and the latter stand in lawlike relations, then lawlike relations will hold between phenomenal properties and some physical occurrences. This conclusion appears to give a causal role to phenomenal properties and thus to suggest a way out of epiphenomenalism. But if intrinsicality carries epiphenomenality, as D. Robinson's extension of Lewis's argument suggests, then this way out of epiphenomenalism would be blocked.

Libet's Unconscious Cerebral Initiatives

Benjamin Libet (1985) argues that experiments done by himself and others shows that certain voluntary actions are preceded by neural events that occur prior to awareness of the intention to act. Interpretation of Libet's results is extremely controversial (see the Open Peer Commentaries following Libet's (1985) target article), and Libet himself does not draw an epiphenomenalistic conclusion. Some philosophers have, however, noted that if Libet's own claims are accepted, then some of our actions are initiated prior to our conscious intention to perform them. If this is correct, then at least in some cases our intuitive judgments that our conscious intentions are causing our movements must be illusory.

Historical Note on Automatism and the Term "Epiphenomenalism"

James Ward's *Encyclopedia Britannica* (Tenth edition, 1902) article, "Psychology", contains the following summary of T. H. Huxley's view: "physical changes are held to be independent of psychical, whereas psychical changes are declared to be their "collateral products." They are called *collateral* products, or "epiphenomena" to obviate the charge of materialism . . . ." McDougall (1911) roundly declares, referring to Huxley, that "to him [the doctrine] owes the name by which it is generally known; for he it was who suggested that the stream of consciousness should be called epiphenomenal, or the epiphenomenon of the brain-process." In Carington (1949), H. H. Price expresses his belief that the term "epiphenomenalism" was introduced by T. H. Huxley.

It is interesting, therefore, that the term "epiphenomenalism" does not occur in Huxley's (1874) essay on our topic; nor have I been able to find it elsewhere in his published work. (Neither does Huxley use the terms "stream of consciousness" or "brain-process".) Of course, it is possible that Huxley made oral use of "epiphenomenalism" in lecturing. This seems unlikely, however, as he had at his disposal another brief term for the view he was concerned to promote, the meaning of which would have been more immediately accessible to most audiences, namely, "automatism". This is the term that occurs in his 1874 essay, which bears the title "On the Hypothesis that Animals are Automata, and its History". Besides containing the analogy of the steam-whistle that contributes nothing to the locomotive's work, this essay compares consciousness to the sound of the bell of a clock that has no role in keeping the time, and treats volition as a symbol in consciousness of the brain-state cause of an action. As Ward correctly noted, nonefficacious mental events are referred to in this essay as "collateral products" of their physical causes. The essay is not solely concerned with animals: to the best of Huxley's judgment, "the argumentation which applies to brutes holds equally good of men".

Huxley and his contemporaries seem to have been impressed by preparations in which frogs had had various portions of their brains removed. Reasoning by analogy with humans lesioned by disease or battle, Huxley finds it plausible that the frogs are not conscious, or not exercising volition; yet when thrown into water, for example, they swim just as well as undamaged frogs. Huxley also discusses at some length the case of a Seargent F., who had sustained a shot that fractured his left parietal bone. Once or twice a month, this soldier would have a day-long bout in which he exhibited complex behavior (e.g., singing, writing a letter, "reloading", "aiming", and "firing" his cane with motions exactly appropriate to a rifle in a skirmish) while being plausibly unconscious, as evidenced by insensitivity to pins and shocks, sound, smell and taste, and to a great extent, vision. Huxley allows that there can be no direct evidence showing that the soldier is conscious or not conscious; but he concludes that he may be devoid of consciousness, while performing his complex and apparently purposeful movements.

Huxley was not alone among 19th century figures who gave vigorous and clear expositions of an epiphenomenalistic view. S. Hodgson (1870), W. K. Clifford (1874) and H. Maudsley (1886) were exponents of the view. Romanes' posthumous (1896) contains an excellent statement of the view, which was first published in the early 1880s; and William James (1879) likewise offers an early clear statement of it. Both Romanes and James follow their statements of the view with arguments against its acceptance.

None of the works just mentioned include the term "epiphenomenalism". I have located three articles in *Mind* in the 1890s that do use the term (the earliest, in 1893, hyphenates it as "epi-phenomenalism"). The earliest occurrence of the term for referring to automatism that I have been able to locate is in William James's *The Principles of Psychology*, first published in 1890. It occurs in his chapter "The Automaton-Theory" once, in scare quotes; the rest of the time, the view is referred to as the "automaton-theory" or the "conscious automaton-theory". James attributes the origination of the view to Shadworth Hodgson, in *The Theory of Practice* (1870). A section of this work titled "Dependence of consciousness on nerve movement" does indeed contain a forthright statement of the view (without "epiphenomenalism", "automatism" or any other "-ism" tag).

Early in his discussion of automatism, James (1890) includes some remarks about his intellectual development, and refers to his early study of medicine. "Epiphenomenon" has a use in this field, meaning a symptom concurrent with, but not causally contributory to, the course of a disease. Some early twentieth century dictionaries list only this meaning of the term; by mid-twentieth century, the focal philosophical meaning is standardly given. My present surmise is that the term "epiphenomenalism" came into philosophy from medicine in the late nineteenth century, possibly, though less certainly, through William James's use of the term in his influential *Principles of Psychology* (1890).

Bibliography

An extensive bibliography is available under the entry for epiphenomenalism in the CD-Rom edition of *The Philosopher's Index*. The list below contains all items referred to in the foregoing article, and a few other sources that offer particularly helpful discussions.

- Benecke, E. C. (1901) "On the Aspect Theory of the Relation of Mind to Body", *Aristotelian Society Proceedings, 1900-1901* n.s. 1:18-44.
- Bieri, P. (1992) "Trying Out Epiphenomenalism", *Erkenntnis* 36:283-309.
- Carington, W. (1949) *Matter, Mind, and Meaning* (New Haven: Yale U. P.). H. H. Price edited this work, and wrote an introduction and notes for it.
- Chalmers, D. J. (1996) *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford U. P.).
- Clifford, W. K. (1874) "Body and Mind", lecture originally given to the Sunday Lecture Society, Nov. 1, 1874. Published in *The Fortnightly Review*, December, 1874, n.s.16:714-736. Reprinted in L. Stephen & F. Pollock, eds., *Lectures and Essays of the late W. K. Clifford*, (London: Macmillan, 1879).
- Davidson, D. (1970) "Mental Events", in Lawrence Foster and J. W. Swanson, eds., *Experience and Theory* (London: Duckworth). Reprinted, with other relevant papers, in D. Davidson, *Actions and Events* (Oxford: Clarendon, 1980).
- Dennett, D. C. (1991a) "Real Patterns", *The Journal of Philosophy* 88:27-51.
- Dennett, D. C. (1991b) *Consciousness Explained* (Boston: Little, Brown).
- Descartes, R. (1649) *The Passions of the Soul*, Part I, art. xxxiv.
- Fodor, J. A. (1989) "Making Mind Matter More", *Philosophical Topics* 17:59-79.
- Hodgson, S. (1870) *The Theory of Practice* (London: Longmans, Green, Reader, & Dyer).
- Honderich, T. (1982) "The Argument for Anomalous Monism", *Analysis* 42:59-64.
- Honderich, T. (1983) "Anomalous Monism: Reply to Smith", *Analysis* 43:147-149.
- Honderich, T. (1984) "Smith and the Champion of Mauve", *Analysis* 44:86-89.
- Huxley, T. H. (1874) "On the Hypothesis that Animals are Automata, and its History", *The Fortnightly Review*, n.s.16:555-580. Reprinted in *Method and Results: Essays by Thomas H. Huxley* (New York: D. Appleton and Company, 1898).
- Jackson, F. (1982) "Epiphenomenal Qualia", *The Philosophical Quarterly* 32:127-136.
- James, W. (1879) "Are We Automata?" *Mind* 4:1-22.
- James, W. (1890) *The Principles of Psychology* (H. Holt)
- Kim, J. (1993) *Supervenience and Mind: Selected Philosophical Essays* (Cambridge: Cambridge U. P.).
- Lalor, B. J. (1997) "It Is What You Think: Intentional Potency and Anti-individualism", *Philosophical Psychology* 10:165-178.
- LePore, E. & Loewer, B. (1987) "Mind Matters", *The Journal of Philosophy* 84:630-642.
- Lewis, D. (1988) "What Experience Teaches", *Proceedings of the Russellian Society*, J. Copley-Coltheart, ed. University of Sydney. Reprinted in Lycan, W. G., ed., *Mind and Cognition* (Cambridge, MA: MIT Press, 1990).
- Libet, B. (1985) "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action", *Behavioral and Brain Sciences* 8:529-566.
- Lockwood, M. (1993) "The Grain Problem", in H. Robinson, ed., *Objections to Physicalism* (Oxford: Oxford University Press).
- Maudsley, H. (1886) *Body and Mind: An Inquiry into Their Connection and Mutual Influence, Specially in Reference to Mental Disorders* (New York: D. Appleton and Co.).

- McDougall, W. (1911) *Body and Mind: A History and Defense of Animism* (London: Methuen).
- Popper, K. & Eccles, J. (1977) *The Self and Its Brain* (New York: Springer-Verlag).
- Putnam, H. (1975) "The Meaning of 'Meaning'", in K. Gunderson, ed., *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science VII (Minneapolis: U. of Minnesota Press). Reprinted in H. Putnam, *Mind, Language and Reality: Philosophical Papers, vol. 2* (Cambridge: Cambridge U. P.).
- Robinson, D. (1993) "Epiphenomenalism, Laws and Properties", *Philosophical Studies* 69:1-34.
- Robinson, W. S. (1982a) "Why I Am a Dualist", in E. D. Klemke, A. D. Kline & R. Hollinger, eds., *Philosophy: The Basic Issues* (New York: St. Martin's Press).
- Robinson, W. S. (1982b) "Causation, Sensations and Knowledge", *Mind* 91:524-540.
- Robinson, W. S. (1990) "States and Beliefs", *Mind* 99:33-51.
- Robinson, W. S. (1995) "Mild Realism, Causation, and Folk Psychology", *Philosophical Psychology* 8:167-187.
- Romanes, G. J. (1896) *Mind and Motion, and Monism* (London: Longmans, Green, and Co.). This book is an edition of material that first appeared in 1882 through 1886.
- Russell, B. (1927) *The Analysis of Matter* (New York: Harcourt, Brace).
- Smith, P. (1982) "Bad News for Anomalous Monism?", *Analysis* 42:220-224.
- Smith, P. (1984) "Anomalous Monism and Epiphenomenalism: A Reply to Honderich", *Analysis* 44:83-86.
- Stich, S. (1983) *From Folk Psychology to Cognitive Science* (Cambridge, MA: MIT Press).
- Taylor, R. (1963) *Metaphysics* (Englewood Cliffs, NJ: Prentice Hall).
- Van Rooijen, J. (1987) "Interactionism and Evolution: A Critique of Popper", *British Journal for the Philosophy of Science* 38:87-92.
- Ward, J. (1902) *Encyclopedia Britannica*, 10th edition, vol. 32, article "Psychology". Material quoted above appears in a section titled "Relation of Body and Mind: Psychophysical Parallelism" which did not appear in the 9th (1883) edition.
- Wundt, W. (1912) *An Introduction to Psychology*, translated from the second German edition by R. Pintner. (London: George Allen).

Other Internet Resources

- David Chalmers' bibliography

Related Entries

consciousness | Descartes, René | dualism | emergent properties | functionalism | mind: philosophy of | multiple realizability | qualia | supervenience

William S. Robinson
Iowa State University

wsrob@iastate.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 18, 1999
Content last modified: January 18, 1999

Stanford Encyclopedia of Philosophy
Notes to Epiphenomenalism

Notes

[1.] Strictly speaking, one can distinguish two versions of epiphenomenalism, (a) mental events have no effects; and (b) mental events may have effects on other mental events, so long as the latter do not have any physical effects, nor any effects on further mental events that eventually have any physical effects. Because observing this distinction would contribute nothing but a tedious duplication of arguments that would differ only trivially, it will be ignored in the remainder of the article.

William S. Robinson
wsrob@iastate.edu

First published: January 18, 1999
Content last modified: January 18, 1999

Multiple Realizability

In the philosophy of mind, multiple realizability is the contention that a given mental kind (property, state, event) is realized by distinct physical kinds. The classic example (presented below) is pain: a wide variety of physical states and events, sharing nothing in common at that level of description, can presumably realize the same pain state. This contention served as a premise in the most influential argument against early psychoneural identity theories. It also served indirectly in early arguments for functionalism. More recently, it has been adopted by nonreductive physicalists to challenge all varieties of psychophysical reduction. It has even been employed to challenge the functionalism it initially motivated. A variety of recent reductionist programs have either attacked the argument from multiple realizability to irreducibility or revised aspects of classical reductionism that made the latter susceptible to this challenge. Even the truth of the multiple realizability contention has been questioned.

What is meant by "realization" in this context? Typically a standard characterization is given. An event e's being F realizes e's being G just in case (i) e is F, (ii) e is G, (iii) for all e it is (physically) necessary that if e is F then e is G, and (iv) e's being F explains e's being G. The last condition requires extensive clarification, but the key point is that realization is stronger than "mere" (physical) necessity. This is because some (physically) necessary property connections are not explanatory. It should be noted that this standard characterization does not capture the relation championed by some functionalists, who hold that a physical state can realize a functional state contingently. The appropriate characterization of realization, like that of supervenience in related contexts, remains a contentious matter.

Putnam's Early Arguments

In a series of papers published throughout the 1960s (and reprinted in his 1975 anthology, Hilary Putnam introduced multiple realizability into the philosophy of mind. Against the "brain state theorist," who holds that every mental kind (like pain) is identical to some as-yet-undiscovered neural kind ("C-fiber firing" was a favorite expression), Putnam considers the wide variety of terrestrial creatures seemingly capable of the former. Humans, other primates, other mammals, birds, reptiles, amphibians, and even mollusks (e.g., octopi) seem reasonable candidates for pain-bearing. But then for the brain state theory to be true, there must be some physical-chemical kind common to this wide variety of neural structures, and correlated exactly with each occurrence of the mental kind. This is a necessary condition of the hypothesized type-identity. Comparative neuroanatomy and physiology, facts about convergent evolution, and the corticalization of function (especially sensory function) as cortical mass increases across species all speak against this consequence. Furthermore, early identity theories insisted that these identities, while contingent, hold by virtue of natural (scientific) law. But then any physically possible cognizer (e.g., pain-bearer) must also be capable of that physical-chemical kind. Here the well-known philosophers' fantasies enter the discussion: silicon-based androids, artificially intelligent electronic robots, Martians with green slime pulsating through their heads. Even further, these identity theories were supposed to be completely general: every mental kind was held to be identical to some neural kind. So the critic needed to find only one mental kind realized differently at the physical-chemical level. As Putnam admitted, the early identity theories were empirical hypotheses. But one consequence was "certainly ambitious," and (highly) probably false.

The argument from multiple realizability for functionalism is more indirect. Ned Block and Jerry Fodor (reprinted as Chapter 4 in Fodor 1981)once noted that this feature of the mental shows that any physical type-identity hypothesis will be insufficiently abstract. Functionalism, which seeks to identify mental kinds with functional kinds characterized exhaustively in terms of their causes and effects, seems to be the next level of abstraction up in the property hierarchy from physical mechanisms. Multiple realizability at the level of physical description is a common characteristic of functional kinds like mousetrap and valve lifter. Characterizing mental kinds as functional kinds thus appears to be a reasonable empirical hypothesis in light of multiple physical realizability.

Fodor's Elaboration and the Legacy of Multiple Realizability

One reply a physical type-identity theorist might make identifies mental kinds with the disjunction of the physical kinds realizing it. (Jaegwon Kim once defended this strategy; the article is reprinted as Chapter 8 in Kim 1993.) Putnam dismissed this strategy as unworthy of serious consideration, but Fodor saw the deep difficulty that the required disjunctions create. In Chapter 1 of his (1975) he argued that reductionism imposes too strong a constraint on acceptable theories in special sciences like psychology. Fodor characterized reductionism as the conjunction of "token physicalism" with the claim that there are natural kind predicates in an ideally completed physics corresponding to each natural kind term in any ideally completed special science. (He characterized "token physicalism" as the claim that all events that science talks about are physical events.) Fodor gave reductionists the best-developed theory of reduction

at the time: in essence, Ernest Nagel's (1961) "derivability" account, which "connected" disparate elements of the reduced and reducing vocabularies via "bridge laws" and claimed a reduction when the laws of the reduced theory were derived from the laws of the reducing and the bridge laws. According to Fodor (1975), if reductionism is to establish physicalism, these cross-theoretic bridge laws must establish contingent identities of reduced and reducing kinds. But given multiple realizability, the only way this can obtain is if the physical science constituent of a psychophysical bridge law is a disjunction of all the terms denoting possible physical realizations of the mental kind. Given the extent and variety of the latter, it is overwhelmingly likely that the disjunctive component will not be a kind-predicate, nor the entire statement a law, of any physical science. Multiple realizability demonstrates that the additional component of reductionism (beyond token physicalism) is empirically untenable.

In a pair of famous examples illustrating multiple realizability in special sciences (economics and psychology), Fodor (1981, Chapter 5) implicitly distinguishes two types of the relation. Call the type Putnam emphasized multiple realizability "across physical structure-types." A more radical type would obtain when a token nervous system realizes a given mental kind via distinct neural events at different times. Call this sense multiple realizability "in a token system across times." (These terms are from John Bickle 1998, Chapter 4.) This second sense will typically increase the disjunctive components of psychophysical bridge laws, because there will be a disjunction of physical predicates realizing each mental kind for every token cognizer.

Following Fodor, psychologist Zenon Pylyshyn (1984) used multiple realizability to ground a methodological criticism of reductionism. He describes a pedestrian, having just witnessed an automobile accident, rushing into a nearby phone booth and dialing a 9 and a 1. What will this person do next? Dial another 1, with overwhelming likelihood. Why? Because of a systematic generalization holding between what he recognized, his background knowledge, his resulting intentions, and that action (intentionally described). But we won't discover that generalization if we focus on his neurophysiology and the resulting muscular contractions. That level of explanation is too weak, for it cannot tell us that this sequence of neural events and muscular contractions corresponds to the action of dialing a 1. A given physiological explanation only links one way of learning the emergency phone number to one way of coming to know that an emergency occurred to one sequence of neural events and resulting muscular contractions producing the behavior (nonintentionally described). However, the number of physical events constituting each of these cognitive classes--the learning, the coming to know, and the action of dialing--is potentially unlimited, with the constituents of each class often unrelated to each other at the physiological level of description. This is a consequence of multiple realizability. So if there is a generalization at the higher level of description available for capturing--and in the pedestrian example there surely is--an exclusively reductionist approach to psychological explanation will miss it. Thus because of multiple realizability, reductionism violates a tenet of scientific methodology: seek to capture all capturable generalizations. (Fodor 1975, Chapter 5, and Terence Horgan 1993 have also raised related methodological caveats about reductionism resting ultimately on multiple realizability. Bickle 1998, Chapter 4, has recently responded to these.)

Contemporary nonreductive physicalism is currently the dominant position in Anglo-American philosophy of mind. It accepts without alteration or amendment the multiple realizability challenge to all

versions of reductive materialism. Ernest LePore and Barry Loewer (1989) have recently called the nonidentity of mental content and physical properties "practically received wisdom" among philosophers of mind. This generalization of the multiple realizability argument also traces back to Fodor (1975 and 1981, Chapter 5). While he targeted explicitly reductionism built on the classical logical empiricist account, he also suggested that his argument applied to more "liberal" versions of reduction then under development. Nonreductive materialists part company with functionalists over the latter's attempt to identify mental kinds with functional kinds. Most of the arguments hinge on issues about individualism in psychology. But a multiple realizability argument has surfaced here. In specifying the nature of mental kinds, many functionalists followed Putnam and Fodor by adopting "Turing machine functionalism": mental kinds are identical to the computational kinds of a suitably programmed universal Turing machine. Putnam (1988) has recently argued that mental kinds are both "compositionally" and "computationally" plastic. The first point is his familiar multiple realizability contention of the mental on the physical. The second contends that the same mental kind can be a property of systems which are not of the same (Turing) computational state. Multiple realizability strikes back at the very theory of mind it initially motivated!

Early Reductionist Replies

David Lewis (reprinted as Chapter 18 in Block 1980) offered the earliest influential reply to Putnam's multiple realizability argument. The inconsistency between the reductionist's thesis and multiple realizability evaporates when we note a tacit relativity to context. A common sense example illustrates Lewis's point. The following three claims appear inconsistent: (1) There is only one winning lottery number. (2) The winning lottery number is 03. (3) The winning lottery number is 61. These three claims seem similarly inconsistent: (1′) (the reductionist thesis) There is only one physical-chemical realization of pain. (2′) The physical-chemical realization of pain is C-fiber firing. (3′) The physical-chemical realization of pain is . . . (something else entirely). ((2′) and (3′) reflect the multiple realizability contention.) But there is no mystery in reconciling (1) -- (3). Append "per week" to (1), "this week" to (2), and "last week" to (3). Similarly, append "per structure-type" to (1′), "in humans" to (2′), and "in mollusks" to (3′). Inconsistency evaporates. Lewis's point is that reductive identities are always specific to a domain.

Many philosophers have elaborated on Lewis's point. Patricia Churchland (1986, Chapter 7), Clifford Hooker (1981), Berent Enç (1983), and other philosophers of science have described historical intertheoretic reductions where a given reduced concept is multiply realized at the reducing level. A common example is the concept of temperature from classical equilibrium thermodynamics. Temperature in a gas is identical to mean molecular kinetic energy. Temperature in a solid, however, is identical to mean maximal molecular kinetic energy, since the molecules of a solid are bound in lattice structures and hence restricted to a range of vibratory motions. Temperature in a plasma is something else entirely, since the molecular constituents of a plasma have been ripped apart. Even a vacuum can have a (blackbody) temperature, though it contains no molecular constituents. Temperature of classical thermodynamics is multiply realized microphysically in a variety of distinct physical states. Yet this is a "textbook" intertheoretic reduction and cross-theory identification. The reduction and identification are

specific to the domain of physical state.

Lewis's insight also underlies Jaegwon Kim's recent appeal to structure-specific "local reductions" (reprinted as Chapters 14 and 16 in Kim 1993). Kim agrees that multiple realizability rules out a general reduction of (structure-independent) psychology to the physical sciences. But it permits (even sanctions) a local reduction to a theory of the physical mechanisms of a given structure-type. (Kim admits that the relevant structure-types here will probably be narrower than biological species.) Local reductions involve "structure-specific bridge laws" where the mental-physical biconditional occurs as the consequent of a conditional whose antecedent denotes a specific structure-type. Conditionals whose antecedents denote different structure types will typically have biconditionals as consequents whose mental term-constituents are coreferential but whose physical term- constituents denote different physical events. Multiple realizability forces this much revision to the bridge laws of classical reductionism. But according to Kim, local reductions are the rule rather than the exception in science generally, and are sufficient for any reasonable scientific or philosophical purpose. Kim's is yet another way to express the tacit specificity to a domain in scientific reductions.

Multiple Realizability Within Structure-Types and Token Systems

The scope of Lewis's strategy and its recent variations is limited, however. They only adequately address the "across structure-types" version of the multiple realizability argument. Recent anti-reductionists have stressed the more radical type. To capture that sense, the context or domain to which a reduction must be relativized is a token system of a structure-type at a time. This much "domain" or "context" specificity seems inconsistent with even a minimally acceptable degree of generality in scientific theorizing. This reply goes back to Ned Block's work in the late 1970s (reprinted as Chapter 22 in Block 1980). Block insisted that the necessary narrowing of psychological kinds renders psychology incapable of capturing whatever generalizations hold across species. Ronald Endicott (1993) gives Block's reply an interesting empirical twist by noting facts about plasticity in the human brain. The capacity for distinct neural structures and processes to subserve a given psychological function owing to trauma, damage, changing task demands, development, and other factors is extensive. A psychology narrowed enough to handle the more radical type of multiple realizability might not be sufficiently general to capture generalizations even within a species whose brains display human-like plasticity.

Against this reply, Kim (1993, Chapter 16) and Bickle (1998, Chapter 4) independently remind us that a guiding methodological principle in contemporary neuroscience assumes some continuity of underlying neural mechanisms. This assumption informs most experimental techniques and paradigms, and theoretical conclusions drawn from experimental data. Continuity is assumed both within and across species. If the radical sense of multiple realizability really obtained, to the extent necessary to circumvent the Lewis-inspired replies to the initial multiple realizability argument, contemporary neuroscientific experimental techniques should bear little fruit. (Why study the macaque visual system to investigate human visual processing, for example, if we can't safely assume some continuity across species? Why

should PET scans reveal common areas of high metabolic activity across and within individual humans, down to less than a centimeter of resolution? Standard neuroscientific experimental procedures and even clinical diagnostic tools would be hopelessly naïve.) These procedures do work, however (and are not hopelessly naïve), and this is evidence that psychological functions might not be so radically multiply realized as recent anti-reductionists pretend. Even neural plasticity is systematic. It has a regular progression following damage to a principal structure; there are underlying neural mechanisms that subserve it. Furthermore, function following damage is often seriously degraded. Persons can still talk, manipulate spatial representations, or move their extremities, but their performance is often qualitatively and quantitatively less than normal. This fact gives rise to tricky questions about individuation of psychological function. Are these alternative neural structures realizing the same psychological function (the same mental kind) as before?

Still, one would like a more direct reply to this more radical type of multiple realizability. Following suggestions by Hooker (1981) and Enc (1983), Bickle (1998, Chapter 4) argues that this feature is common to scientific reductions generally. For example, it obtains in the "textbook" reduction of classical equilibrium thermodynamics to microphysics via statistical mechanics. For any token aggregate of gas molecules, there is an indefinite number of realizations of a given temperature, a given mean molecular kinetic energy. Microphysically, the most fine-grained theoretical specification of a gas is its microcanonical ensemble, in which the momentum and the location (and thus the kinetic energy) of each molecule is specified individually. Indefinitely many distinct microcanonical ensembles of a token volume of gas molecules can yield the same mean molecular kinetic energy. Thus at the lowest level of microphysical description, temperature is vastly multiply realizable in the same token system over time. So even multiple realizability in this more radical sense is not by itself a barrier to reducibility. (Bickle also argues that exactly this relation obtains within "connectionist" theories of representational content.)

Other Replies to Multiple Realizability Arguments

In searching for reductive unity underlying the variety of cognitive systems, Paul Churchland (1982) once suggested descending "below" neurobiology and even biochemistry to the level of nonequilibrium thermodynamics. He insisted that finding reductive unity there was more than a bare logical possibility, because of some parallels between biological processes, whose multiply realized kinds find reductive unity there, and cognitive activity (especially learning). Robert Richardson (1979) once suggested that the Putnam-Fodor challenge resulted from an incomplete understanding of Ernest Nagel's classic account of reduction. Although Nagel's examples involve biconditional cross-theory connections, one-way conditional connections expressing sufficient conditions at the reducing level were all his "principle of derivability" required. (Richardson cites the relevant passages from Nagel's classic exposition.) Multiple realizability only challenges the necessity of (nondisjunctive) reducing conditions, and so is not a challenge to even a projected Nagelian reduction of psychology to the physical sciences.

Besides his appeal to species-specific bridge laws and local reducibility, Kim (1993, Chapter 14) has recently offered two additional replies to the multiple realizability argument. His "projectability" reply starts from the familiar fact that the kind "jade" fragments into jadeite and nephrite. This renders jade

incapable of passing the projectability test for nomicness because of its genuinely disjunctive nature. Multiple realizability of psychological kinds yields the same consequence. Instead of rendering psychology an autonomous special science, multiple realizability implies that there is no structure-independent scientific psychology. There are only "local" scientific psychologies reducible to the theory of the underlying physical mechanisms of the structure-type in question. Closely related is his "causal powers" reply. Scientific kinds are individuated by their causal powers, and the causal powers of each instance of some realized kind are identical to those of its realizer. From these principles it follows that instances of a mental kind with different physical realizations are distinct kinds. Thus (structure-independent) mental kinds are not causal kinds, and hence are disqualified as proper scientific kinds. Multiple realizability yields the failure of structure-independent mental kinds to meet the standards of scientific kinds. (Terence Horgan 1996 has mounted some interesting rejoinders to these related arguments.)

With regard to Pylyshyn's (1984) attack on reductionist methodology, Patricia Churchland (1986, Chapter 9) suggests that lower level sciences themselves can construct functional theories. This inserts a new level of theory between that of the structure of the lower level kinds and that of purely functional kinds: between, for example, the physiology of individual neurons and cognitive psychology. We might find a common neurofunctional property for a given type of psychological state. And if the scope of the macro-theory doesn't extend beyond that of its microfunctional counterpart, then reduction will be achieved despite vast multiple realizability at the microstructural level. Neurocomputational approaches that have blossomed recently give empirical credence to Churchland's suggestion. Bickle (1998, Chapters 3 and 4) has tried to explain how such a result is sufficient for reduction by building some suggestions from Hooker (1981, Part III) about "cross categorial" and "token-to-token" reductions into a general model of the intertheoretic reduction relation. The cross-classifications implied by multiple realizability might require that a special kind of reduction relation obtain between psychology and the physical sciences. But this will be reduction enough, with plenty of scientific precedent.

Finally, notice that all reductionist responses discussed so far refrain from attacking the multiple realizability contention itself. Even this has been attacked. Nick Zangwill (1992) insists that multiple realizability across biological species has not been proven. The multiple realizability contention assumes a type-identity of mental kinds across species. This assumption is problematic, given that the obvious sensory and motor differences alone will yield different cause-and-effect patterns at all but the grossest level of description. We are back to tricky issues about how psychology individuates types. It is interesting to note in closing that nobody has developed the argument that the multiple realizability contention rests precariously on "folk" intuitions about mental type-individuation.

Bibliography

- **Bickle, John** (1998). *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.
- **Block, Ned** (ed.) (1980) *Readings in the Philosophy of Psychology*, vol. 1. Cambridge, MA: Harvard University Press.
- **Churchland, Patricia** (1986). *Neurophilosophy*. Cambridge, MA: MIT Press.

- **Churchland, Paul** (1982). "Is Thinker a Natural Kind?" *Dialogue* 21, 223-238
- **Enç, Berent** (1983). "In Defense of the Identity Theory." *Journal of Philosophy* 80, 279-298.
- **Endicott, Ronald** (1993). "Species-Specific Properties and More Narrow Reductive Strategies." *Erkenntnis* 38, 303-321.
- **Fodor, Jerry** (1975). *The Language of Thought*. New York: Thomas Crowell.
- **Fodor, Jerry** (1981). *Representations*. Cambridge, MA: MIT Press
- **Hooker, Clifford** (1981). "Towards a General Theory of Reduction. Part III: Cross-Categorial Reductions." *Dialogue* 20, 496-529.
- **Horgan, Terence** (1993). "Nonreductive Materialism and the Explanatory Autonomy of Psychology.: In S. Wagner and R. Warner (eds.), *Naturalism: A Critical Appraisal*. Notre Dame, IN: University of Notre Dame Press.
- **Horgan, Terence** (1996). "Kim on the Mind-Body Problem." *British Journal for the Philosophy of Science* 47, 579-607.
- **Kim, Jaegwon** (1993). *Supervenience and Mind*. Cambridge: Cambridge University Press.
- **LePore, Ernest** and **Loewer, Barry** (1989). "More on Making Mind Matter." *Philosophical Topics* 17: 175-191.
- **Nagel, Ernest** (1961). *The Structure of Science*. New York: Harcourt, Brace, and World.
- **Putnam, Hilary** (1975). *Mind, Language, and Reality: Philosophical Papers*, vol. 2. Cambridge: Cambridge University Press.
- **Putnam, Hilary** (1988). *Representation and Reality*. Cambridge, MA: MIT Press.
- **Pylyshyn, Zenon** (1984). *Computation and Cognition*. Cambridge, MA: MIT Press.
- **Richardson, Robert** (1979). "Functionalism and Reductionism." *Philosophy of Science* 46, 533-558.
- **Zangwill, Nick** (1992). "Variable Reduction Not Proven." *Philosophical Quarterly* 42: 214-218

Other Internet Resources

David Chalmers' Bibliography on Reduction and Multiple Realizability

Related Entries

functionalism | physicalism | reduction and reductionism

John Bickle
bicklejw@email.uc.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 23, 1998
Content last modified: November 23, 1998

Physicalism

Physicalism is the thesis that everything is physical, or as contemporary philosophers sometimes put it, that everything supervenes on the physical. The thesis is usually intended as a metaphysical thesis, parallel to the ancient Greek philosopher Thales's thesis that everything is water, or the idealism of the 18th Century philosopher Berkeley, that everything is mental. The general idea is that the nature of the actual world (i.e. the universe and everything in it) conforms to a certain condition, the condition of being physical. Of course, physicalists don't deny that the world might contain many items that at first glance don't *seem* physical -- items of a biological, or psychological, or moral, or social nature. But they insist nevertheless that at the end of the day such items are wholly physical.

Physicalism is sometimes known as materialism. Historically, materialists held that everything was matter -- where matter was conceived as "an inert, senseless substance, in which extension, figure, and motion do actually subsist" (Berkeley, *Principles of Human Knowledge*, par. 9). The reason for speaking of physicalism rather than materialism is to abstract away from this historical notion, which is usually thought of as too restrictive -- for example, forces such as gravity are physical but it is not clear that they are material in the traditional sense (Dijksterhuis 1961, Yolton 1983). It is also to emphasize a connection to physics and the physical sciences. Indeed, physicalism is unusual among metaphysical doctrines in being associated historically with a commitment both to the sciences and to a particular branch of science, namely physics.

---

1. A Framework for Discussion

In approaching the topic of physicalism, one may distinguish what I will call *the interpretation question* from *the truth question*. The interpretation question asks:

- What does it *mean* to say that everything is physical?

The truth question asks:

- Is it *true* to say that everything is physical?

There is obviously a sense in which the second question here presupposes an answer to the first -- you need to know what a statement means before you can ask whether it's true -- and I will begin with the interpretation question. Nevertheless, the issues here turn out to be somewhat technical, and those new to the topic might like to read only the first section of my discussion of the interpretation question, which is: Supervenience Physicalism: Introductory, and then turn directly to the truth question which begins at The Case Against Physicalism I: Qualia and Consciousness.

The interpretation question itself divides into two sub-questions, which I will call *the completeness question* and *the condition question*. The completeness question asks:

- What does it mean to say that *everything* is physical?

In other words, the completeness question holds fixed the issue of what it means for something to satisfy the condition of being physical, and asks instead what it means for *everything* to satisfy that condition. Notice that a parallel question could be asked of Thales: assuming we know what condition you have to satisfy to be water, what does it mean to say that everything satisfies that condition?

The condition question asks:

- What does it mean to say that everything is *physical*?

In other words, the condition question holds fixed the issue of what it means for everything to satisfy some condition or other, and asks instead what *is* the condition, being physical, that everything satisfies. Notice again that a parallel question could be asked of Thales: assuming we know what it is for everything to satisfy some condition or other, what is the condition, being water, that according to Thales, everything satisfies? In discussing the interpretation question, I will turn first to the completeness question, and then consider the condition question.

2. Supervenience Physicalism: Introductory

In attempting to answer the completeness question, it has become customary since Davidson 1970 to look to the notion of supervenience. (The notion of supervenience is historically associated with meta-ethics, but it has received extensive discussion in the general metaphysics and logic literature. For a survey, see Kim 1993.)

The idea of supervenience might be introduced via an example due to David Lewis of a dot-matrix picture:

> A dot-matrix picture has global properties -- it is symmetrical, it is cluttered, and whatnot -- and yet all there is to the picture is dots and non-dots at each point of the matrix. The global properties are nothing but patterns in the dots. They supervene: no two pictures could differ in their global properties without differing, somewhere, in whether there is or there isn't a dot (1986, p. 14).

Lewis's example gives us one way to introduce the basic of idea of physicalism. The basic idea is that the physical features of the world are like the dots in the picture, and the psychological or biological or social features of the world are like the global properties of the picture. Just as the global features of the picture are nothing but a pattern in the dots, so too the psychological, the biological and the social features of the world are nothing but a pattern in the physical features of the world. To use the language of supervenience, just as the global features of the picture supervene on the dots, so too everything supervenes on the physical, if physicalism is true.

It is desirable to have a more explicit statement of physicalism, and here too Lewis's example gives us direction. Lewis says that, in the case of the picture, supervenience means that "no two pictures can be identical in the arrangement of dots but different in their global properties". Similarly, one might say that, in the case of physicalism, no two possible worlds can be identical in their physical properties but differ, somewhere, in their mental, social or biological properties. To put this slightly differently, we might say that if physicalism is true at our world, then no *other* world can be physically identical to it without being identical to it in all respects. This suggests the following account of what physicalism is:

> (1) Physicalism is true at a possible world *w* iff any world which is a physical duplicate of *w* is a duplicate of *w simpliciter*.

If physicalism is construed along the lines suggested in (1), then we have an answer to the completeness question. The completeness question asks: what does it mean to say that *everything* is physical. According to (1), what this means is that if physicalism is true, there is no possible world which is identical to the actual world in every physical respect but which is not identical to it in a biological or social or psychological respect. It will be useful to have a name for physicalism so defined, so let us call it *supervenience physicalism*.

3. Supervenience Physicalism: Further Issues

Supervenience physicalism is relatively simple and clear, but when construed as a formulation of physicalism, it faces three initial problems: (a) the epiphenomenal ectoplasm problem; (b) the lone ammonium molecule problem; (c) the modal status problem; and (d) the necessary beings problem.

The epiphenomenal ectoplasm problem

(Cf. Horgan 1983, Lewis 1983.) Imagine a possible world *W* that is exactly like our world in respect of the distribution of physical and mental properties, but for one difference: it contains some pure experience which does not interact causally with anything else in the world -- *epiphenomenal ectoplasm*, to give it a name. The problem this possibility presents for (1) is that, if (1) provides the correct definition of physicalism, and if physicalism is true at the actual world, then there is *no* possible world of the kind we just described, i.e., *W* does not exist. The reason is that *W* is by assumption a physical duplicate of our world; but then, if physicalism is true at our world, *W* should be a duplicate simpliciter of our world. But *W* is patently not a duplicate of our world: it contains some epiphenomenal ectoplasm that our world lacks. On the other hand, it seems quite wrong to say that *W* is an impossibility -- at any rate, physicalism should not *entail* that it is impossible.

In order to solve the epiphenomenal ectoplasm problem, we need to adjust (1) so that it does not have the truth of physicalism ruling out *W* as a possible world. While there are a number of different proposals about how to do this, the simplest is due to Frank Jackson (cf. Jackson 1993. For earlier proposals and further discussion, see Horgan 1983 and Lewis 1983.) He proposes replacing (1) with:

> (2) Physicalism is true at a possible world *w* iff any world which is a minimal physical duplicate of *w* is a duplicate of *w simpliciter*

By ‘minimal physical duplicate’, Jackson means a possible world that is identical in all physical respects to the actual world, but which does not contain anything else; in particular, it does not contain any epiphenomenal ectoplasm. Unlike (1), (2) does not have physicalism ruling out *W*, and so (2) is preferable to (1), as a statement of physicalism, and it is (2) with which we shall work in this entry.

The lone ammonium molecule problem

(Cf. Kim 1993.) Imagine a possible world *W** that is physically exactly like our world except in one trivial respect: it has one extra ammonium molecule located, say, on Saturn's rings. It is natural to suppose that at *W**, the distribution of mental properties is exactly as it is in the actual world -- the presence of an extra molecule does not make that much of a difference. On the other hand, for all (2) says, such a world might be *radically different* in terms of the distribution of mental properties. Since (2) only refers to worlds that are exact minimal physical duplicates of our world, it is silent on worlds that are different *even in minute details*, and hence is silent on *W**. It thus leaves open the possibility that, in *W**, everything has mental properties, or nothing has or the only things that have mental properties are Saturn's rings! But that seems absurd: while *W** is clearly not a physical duplicate of our world, for it contains an extra molecule, the distribution of mental properties in *W** would nevertheless match that of the actual world.

Once again there are a number of different responses to this problem in the literature (cf. Kim 1993). But perhaps the simplest response is that the problem conflates two issues that are better kept apart: the question of what physicalism itself tells us about *W**, and the question of what our general knowledge tells us about *W**. It is true that physicalism itself does not tell us anything about the distribution of mental properties at *W**. Nevertheless, we know independently what the distribution is -- we know independently that the presence or absence of molecules on Saturn doesn't affect things like who has mental properties here on Earth. But why should one assume that this last piece of knowledge should be a consequence of physicalism? To put the point slightly differently, imagine that we discover that who has mental properties on Earth *is* in part a function of the behavior of molecules on Saturn. That would of course tell us that we are deeply wrong in our assumptions about how the world works. But it would not tell us that we are deeply wrong about physicalism. (For further discussion of this point, see Paull and Sider 1992, and Stalnaker 1996.)

The modal status problem

Some philosophers (e.g. Davidson 1970) have thought of physicalism as a conceptual or necessary truth, if it is true at all. But most have thought of it as contingent, a truth about our world which might have been otherwise. The statement of physicalism encoded in (2) allows a way in which this might be so. (2) tells us that physicalism is true at a world just in case the world in question conforms to certain conditions. But it leaves it open whether or not the actual world conforms to those conditions *as a matter of fact*. Perhaps it is *not* true of our world that a physical duplicate of it would be a psychological duplicate. If so, physicalism would not be true at our world.

But for some it is puzzling that physicalism is stated using modal notions (i.e. notions such as possible worlds) and nonetheless is contingent. To see the problem, notice first that, supervenience physicalism tells us that the minimal physical truths of the world entail *all* the truths; hence

> (3) The minimal physical truths entail all the truths.

Now suppose that *S* is a statement which specifies the minimal physical nature of the actual world and *S**

is a statement which specifies the total nature of the world. (It might be that neither *S* nor *S** are expressible in languages we can understand, but let us set this aside.) If supervenience physicalism is true, it will then be true that:

> (4) *S* entails *S**

On the other hand, (4) is clearly a necessary truth. However, if (4) is a necessary truth, how can *physicalism* be contingent? After all, (4) seems equivalent to physicalism. But if the two are equivalent, how can one be necessary and the other contingent?

But the response to this problem is straightforward. (4) is necessary, but it is *not* equivalent to physicalism. Rather, (4) *follows* from physicalism given various contingent assumptions, in particular the assumptions that *S* and *S are* the statements we say they are -- it is contingent fact, for example that *S** summarizes the total nature of the world. On other hand, (3) *is* equivalent to physicalism but it is *not* necessary. (It is important to bear in mind here that not all entailment claims are necessary. Consider 'my aunt's favorite statement entails my uncle's favorite' -- that statement is contingent even though it is most naturally thought of as an entailment claim.)

The necessary beings problem

(Cf. Jackson 1998) Imagine a necessary being -- that is, a being which exists in all possible worlds -- which is essentially nonphysical. (Some theists believe that God provides an example of such a being.) If such a non-physical being exists, it is an natural to suppose that physicalism is false. But if physicalism is defined according to (2), the existence of such a being is compatible with physicalism. For consider: if the actual world is wholly physical, apart from the necessary non-physical being, any minimal physical duplicate of the actual world is a duplicate simpliciter. Since the non-physical being exists in all possible worlds, it exists at all worlds which are minimal physical duplicates of the actual world. So we seem to face a problem: the existence of the non-physical necessary being entails that physicalism is false, but the definition of physicalism permits it to be true in this case.

This problem is not so easily answered as the previous three. Lying behind the problem is a deeper issue about the correct interpretation of necessity and possibility -- the modal notions one uses to formulate supervenience. On one way of interpreting these notions, the existence of a necessary being of this sort is incoherent. A reason is that it would violate David Hume's famous dictum that there are no necessary connections between distinct existences -- the being is distinct from the physical world and yet is necessitated by it. On another way of interpreting these notions, however, there is nothing incoherent in the idea of such a being. The correct way to think about modal notions, however, is a topic that is well beyond the scope of our discussion here. The problem seems to be that the supervenience definition of physicalism in effect presupposes something like Hume's dictum, in that it uses failure of necessitation as a test for distinctness. But this means that someone who denies the dictum will have to find an alternative way of formulating physicalism.

4. Minimal Physicalism and Philosophy of Mind

We saw earlier that physicalism is intended as a very general claim about the nature of the world. Nevertheless, by far the most discussion of physicalism in the literature has been in the philosophy of mind The reason for this is that it is in philosophy of mind that we find the most plausible and compelling arguments that physicalism is false. Indeed, as we will see later on, arguments about qualia and consciousness are usually formulated as arguments for the conclusion that physicalism is false. Thus, much of philosophy of mind is devoted to a discussion of physicalism. Most contemporary philosophers are physicalists. But some, dualists, deny physicalism and thus deny that consciousness supervenes on the physical.

While the issue of physicalism is central to philosophy of mind, however, it is important also to be aware that supervenience physicalism is neutral on a good many of the questions that are pursued in philosophy of mind, and pursued elsewhere for that matter. If you read over the philosophy of mind literature, you will often find people debating a number of different issues: whether there *are* mental states at all; what *sort* of thing mental states are; to what extent mental states are environmentally determined. Given the multifariousness of mental states, it is quite likely that the correct position will be some kind of combination of these positions. But this is a question of further inquiry that is irrelevant to physicalism itself. So physicalism itself leaves many debates in the philosophy of mind unanswered.

This point is sometimes expressed by saying that *supervenience physicalism is minimal physicalism* (Lewis 1983): it is intended to capture the minimal or core commitment of physicalism. Physicalists may differ from one another in many ways, but all of them must at least hold supervenience physicalism. (Notice that the idea that (2) captures the minimal commitment of physicalism is a distinct idea from that of a minimal physical duplicate which it uses to capture minimal physicalism.)

Two issues here require further comment. First, in some discussions in philosophy of mind, the term ‘physicalism’ is used to refer to the identity theory, the idea that mental states or properties are neurological states or properties (Block 1980). In this use of the term, one can reject physicalism by rejecting the identity theory -- so by that standard a behaviorist or functionalist in philosophy of mind would not count as a physicalist. Obviously, this is a much more restricted use of the term than is being employed here.

Second, one might think that supervenience physicalism is inconsistent with eliminativism, the claim that psychological states do not exist, for the following reason. Suppose psychological states supervene on physical states. Doesn't that mean, contrary to eliminativism, that there must *be* some psychological states? The answer to this question is ‘no.’ For consider: the telephone on my desk has no psychological states whatsoever. Nevertheless it is still true (though, admittedly, a little odd) to say that a telephone which is identical to my telephone in all physical respects will be identical to it in all psychological respects. In the sense intended, therefore, one thing can psychologically identical to another even when neither *has* any psychological states.

5. Token and Type Physicalism

To what extent does supervenience physicalism capture minimal physicalism, the core commitment of all physicalists? In order to answer this question it is worth comparing and contrasting supervenience physicalism with two alternative statements of physicalism that one finds in the literature: token and type physicalism.

Token physicalism is the view that every particular thing in the world is a physical particular. Here is one formulation of this idea:

> **Token physicalism**:
> For every actual particular (object, event or process) *x*, there is some physical particular *y* such that *x* = *y*.

Supervenience physicalism neither implies nor is implied by token physicalism. To see that token physicalism does not imply supervenience physicalism, one need only note that the former is consistent with a version of dualism, namely property dualism. The mere fact that every particular has a physical property does not rule out the possibility that some particulars also have non-supervenient mental properties, i.e. mental properties that are only contingently related to the physical. But supervenience physicalism *does* rule out this possibility. Since token physicalism does not rule out property dualism but supervenience physicalism does, the first does not imply the second.

To see that supervenience physicalism does not imply token physicalism is more difficult. The crucial point is that token physicalism requires that for every psychological or social particular, there is some physical particular with which it is identical. But this is by no means obviously true. Consider the United States Court of Appeals for the Seventh Circuit. This might thought of as a social or legal object. But then, according to token physicalism, there must be some physical object for it to be identical with. But there might be no physical object (in any natural sense of the term) which is identical to the Court of Appeals for the Seventh Circuit. On the other hand, supervenience physicalism imposes no such requirement, and so supervenience physicalism does not imply token physicalism (For the classic presentation of this point, see Haugeland 1983).

The point that supervenience physicalism is logically distinct from token physicalism is an important one. One thing it shows is that token physicalism (since it is consistent with property dualism) does not capture minimal physicalism, and so the distinction between token physicalism and supervenience physicalism is no objection to the latter. But the difference between the two theses also raises a different question. Given that token physicalism does not capture the minimal commitment of physicalism, why has token physicalism been the subject of such discussion? One reason is that token physicalism provides one version of the idea that upper level scientific claims requires physical mechanisms. Supervenience physicalism does not on its own entail this. But token physicalism is often seen as a way to ensure this requirement. (For the classic presentation of this point, see Fodor 1974; see also Papineau 1996)

Having considered token physicalism, we can now turn to type physicalism. Type physicalism is a generalization and extension of the identity theory, which we considered above. It holds that that every property (or at least every property that is or could be instantiated in the actual world) is identical with some physical property. Here is a statement of this sort of idea:

> **Type physicalism**:
> For every actually instantiated mental property *F*, there is some physical property *G* such that *F*=*G*.

Unlike token physicalism, type physicalism certainly *does* entail supervenience physicalism: if every property instantiated in the actual world is identical with some physical property, then a world identical to our world in physical respects will of course be identical to it in all respects.

Nevertheless the reverse entailment does not hold. Supervenience physicalism, as we have been understanding it, is a contingent thesis that is consistent with the possibility (if not the actuality) of disembodiment. But type physicalism as defined here is inconsistent with this possibility. To that extent, supervenience physicalism does not entail type physicalism.

Earlier we noted that philosophers such as Davidson have thought that physicalism is a necessary truth. Even on that assumption, however, it is still not completely obvious that supervenience physicalism entails type physicalism. The reason for this has to do with questions concerning the logical (or Boolean) closure of the set of physical properties -- if *P*, *Q* and *R* are physical properties, which of the various logical permutations of *P*, *Q* and *R* are likewise physical properties? On some assumptions concerning closure and supervenience, supervenience physicalism (construed as a necessary truth) entails type physicalism; on other assumptions, it doesn't. But the problem is that the assumptions themselves are difficult to interpret and evaluate, and so the issue remains a difficult one. It is not necessary for our purposes to settle the question concerning closure here. (For further discussion of these issues see Kim 1993, Bacon 1990, Van Cleve 1990, Stalnaker 1996.)

6. Reductive and Non-Reductive Physicalism

Before the development and study of the notion of supervenience, physicalism was often stated as a reductionist thesis. It will therefore be useful to contrast the supervenience formulation of physicalism with various reductionist proposals, and also turn to a question that has received a lot of attention in the literature, viz., whether a physicalist must be a reductionist.

The main problem in assessing whether a physicalist must be a reductionist is that there are various non-equivalent versions of reductionism.

One idea is tied to the notion of conceptual or reductive analysis. When philosophers attempt to provide an analysis of some concept or notion, they usually try to provide a reductive analysis of the notion in question, i.e. to analyze it in other terms. Applied to the philosophy of mind, this notion might be thought

of entailing the idea that every mental concept or notion is analyzed in terms of a physical concept or notion. A formulation of this idea is (5):

> (5) Reductionism is true iff for each mental predicate *F*, there is a physical predicate *G* such that a sentence of the form ‘ *x* is *F* iff *x* is *G*’ is analytically true.

While one occasionally finds in the literature the suggestion that physicalists are committed to (5) in fact, no physicalist since before Smart (1959) has (unqualifiedly) held anything like (5). Adapting Ryle (1949), Smart supposed that in addition to physical expressions there is a class of expressions which are topic-neutral, i.e. expressions which were neither mental nor physical but when conjoined with any theory would greatly increase the expressive power of the theory. Smart suggested that one might analyze mental expressions in topic-neutral (but not physical) terms, which in effect means that a physicalist could reject (5). It is fair to say that this move is one of the central innovations of philosophy of mind, a move to a large extent endorsed and developed by functionalists and cognitive scientists.

A different notion of reduction derives from the attempts of philosophers of science to explain intertheoretic reduction. The classic formulation of this notion was given by Ernest Nagel (1961). Nagel said that one theory was reduced to another if you could logically derive the first from the second together with what he called bridge laws, i.e., laws connecting the predicates of the reduced theory (the theory to be reduced) with the predicates of the reducing theory (the theory to which one is reducing). Here is a formulation of this idea, where the theories in question are psychology and neuroscience:

> (6) Reductionism is true iff for each mental predicate *F* there is a neurological predicate *G* such that a sentence of the form ‘*x* is *F* iff *x* is *G*’ expresses a bridge law.

Once again, however, there is no reason at all why physicalists need to accept that reductionism is true in the sense of (6). Indeed, many philosophers have argued that there are very strong empirical reasons to deny that anything like (6) is going to be the case. The reason is this. Many different neurological processes (whether in our own species or a different one) could underlie the same psychological process -- indeed, given science fiction, even non-neurological processes might underlie the same psychological process. But if multiple realizability -- as this sort of idea is called -- is true, then (6) seems to be false. (Fodor 1974, but for recent alternative views, see Kim 1993).

A third notion of reductionism is more metaphysical in focus than either the conceptual or theoretical ideas reviewed so far. According to this notion, reductionism means that the properties expressed by the predicates of (say) a psychological theory are identical to the properties expressed by the predicates of (say) a neurological theory -- in other words, this version of reductionism is in essence a version of type physicalism or the identity theory. However, as we have seen, if physicalists are committed only to supervenience physicalism, they are not committed to type physicalism. Hence a physicalist need not be a reductionist in this metaphysical sense.

A final notion of reductionism that needs to be distinguished from the previous three concerns whether mental statements follow *a priori* from non-mental statements. Here is a statement of this sort of idea,

> (7) Reductionism is true iff for each mental predicate *F* there is non-mental predicate *G* such that a sentence of the form ‘ if *x* is *F* then *x* is *G*’ is *a priori*.

What (7) says is that if reductionism is true, *a priori* knowledge alone, plus knowledge of the physical truths will allow one to know the mental truths. This question is in fact a highly vexed one in contemporary philosophy. However, this question is usually debated in the context of another, viz., the question of *a posteriori* and *a priori* physicalism. It is to that question, therefore, to which we will now turn.

7. *A Priori* and *A Posteriori* Physicalism

We saw earlier that if physicalism is true then (4) is true, where ‘*S*’ is a sentence that reports the entire physical nature of the world and ‘*S**’ is a sentence that reports the entire nature of the world:

> (4) *S* entails *S**

Another way to say this is to say that if physicalism is true, then the following conditional is necessarily true:

> (8) If *S* then *S**

Indeed, this is a general feature of physicalism: if it is true then there will always be a necessary truth of the form of (8).

Now, if (8) is necessary the question arises whether it is *a priori*, i.e. knowable independent of empirical experience, or whether it is *a posteriori*, i.e. knowable but not independently of empirical experience. Traditionally, every statement that was necessary was assumed to be *a priori*. However, since Kripke's classic work *Naming and Necessity* (1980), philosophers have become used to the idea that there are truths which are both necessary and *a posteriori*. Accordingly many recent philosophers have defended a *posteriori physicalism*: the claim that statements such as (8) are necessary and *a posteriori* (cf. Loar 1997). Moreover, they have used this point to try to disarm many objections to physicalism, including those concerning qualia and intentionality that we will consider in a moment. Indeed, as we have just noted, some philosophers have suggested that the necessary *a posteriori* provides the proper interpretation of non-reductive physicalism.

The appeal to the necessary *a posteriori* is on the surface an attractive one, but it is also controversial. One problem arises from the fact that Kripke's idea that there are necessary and *a posteriori* truths can be interpreted in two rather different ways. On the first interpretation -- I will call it the derivation view -- while there are necessary *a posteriori* truths, these truths can be derived *a priori* from truths which are *a posteriori* and contingent. On the second interpretation -- I will call it the non-derivation view -- there are

non-derived necessary *a posteriori* truths, i.e. necessary truths which are not derived from any contingent truths (or any *a priori* truths for that matter). The problem is that when one combines the derivation view with the claim that (8) is necessary and *a posteriori*, one encounters a contradiction. If the derivation view is correct, then there is some contingent and *a posteriori* statement *S#* that logically entails (8). However, if *S#* logically entails (8) then (since ‘If *C*, then if *A* then *B*’ is equivalent to ‘If *C* & *A*, then *B*’) we can infer that the following is both necessary and *a priori*:

> (9) If *S* & *S#* then *S**.

One the other hand, if physicalism is true, and *S* summarizes the total nature of the world it seems reasonable to suppose that *S# was already implicitly included in S*. In other words it seems reasonable to suppose that (9) is simply an expansion of (8). But if (9) is just an expansion of (8), then if (9) is *a priori*, (8) must also be *a priori*. But that means our initial assumption is false: (8) is not a necessary *a posteriori* truth after all (Jackson 1998).

How might an *a posteriori* physicalist respond to this objection? The obvious response is to reject the derivation view of the necessary *a posteriori* in favor of the non-derivation view. But this is just to say that if one wants to defend *a posteriori* physicalism, one will have to defend the non-derivation view of the necessary *a posteriori*. However, the problem here is that the non-derivation view is very controversial. Indeed, the question of which interpretation of Kripke's work is the right one, is one of the most vexed in contemporary analytic philosophy. So it is not something that we can hope to solve here. (For discussion, see Byrne 1999, Chalmers 1996 1999, Jackson 1998, Loar 1997, 1999, Lewis 1994, Yablo 1999.)

8. Physicalism and Emergentism

A further issue needs to be mentioned in connection with the distinctions between non-reductive and reductive physicalism, and with *a posteriori* and *a priori* physicalism. Kim and others have suggested that non-reductive physicalism is a form of *emergentism*, the view that supervenience provides a way to interpret the relation between the psychological and the physical in such a way that the psychological is *genuinely novel.* (Emergentism was influential in the first forty years of the twentieth century, but it is not unfair to say that similar positions are defended by many contemporary philosophers. For the historical background, see MacLaughlin 1992.)

Now it is difficult to evaluate emergentism because it is unclear what genuine novelty is supposed to be. On one interpretation, what the emergentists meant by ‘genuine novelty’ was non-predictability in principle, i.e. the idea that no matter how much physical information you had about a creature you could not predict on that basis alone what experiences, if any, they might have. On this interpretation, emergentism seems very similar to *a posteriori* physicalism (Byrne 1993). On a different interpretation, what the emergentists meant by ‘genuine novelty’ was the idea that there was only a contingent connection between psychological states and physical state, a connection perhaps mediated by contingent psycho-physical laws. On this interpretation, however, emergentism seems simply to be a denial of

physicalism as we have defined it here.

There is, however, a third interpretation of ‘genuine novelty’ which requires separate treatment. On this interpretation, the idea of genuine novelty is the idea of a layered world: a world that has genuine levels (i.e. levels distinct from one another), and that each of these levels are necessarily connected to others. So interpreted, emergentism is the view that our world is such a layered world.

Now, this version of emergentism does present a problem for our account. To see this, imagine that you begin with a classical form of dualism, and then discover that the laws which related the mental level to the physical level are metaphysically necessary, rather than contingent. In that situation, it is not clear that you have discovered that dualism is false. So emergentism seems to be consistent with dualism. On the other hand, emergentism seems also to be consistent with supervenience physicalism, since according to emergentism, any world physically identical to the actual world will be identical to it in all respects. However -- and here is the final point -- we have been assuming all along that supervenience physicalism is inconsistent with dualism. In short, the problem is this: (a) supervenience physicalism is consistent with emergentism; (b) emergentism is consistent with dualism; but (c) supervenience physicalism is *in*consistent with dualism.

How are we to respond to this problem? I think the best thing to say is that emergentism and physicalism are inconsistent, and hence that (a) is false. However, the inconsistency does not have its source in the formal notion of supervenience. Instead it has its source in the interpretation that both views assign to that notion. The emergentist is obviously being guided by the metaphor of layers, and interprets supervenience in that light. However, while one sometimes uses the metaphor of layers to describe the world as portrayed by supervenience physicalism, it would be more apt to say -- as Lewis says in the example of the dot-matrix picture that we considered above -- that that doctrine presents the psychological, the biological and so on as *patterns in* the physical, rather than *layers on top of* the physical. So the picture implicit in emergentism is that of *a layered world*, whereas the picture implicit in supervenience physicalism is that of *a patterned world*. Since these pictures are inconsistent, (a) is false.

Even if emergentism is distinct from supervenience physicalism, however, it remains a controversial issue whether the emergentist picture can be made fully coherent (Stalnaker 1996). One sort of argument against it is that it seems to violate Hume's dictum that there are no necessary connections between distinct existences: according to emergentism, the levels of the world are wholly distinct from each other, and yet are necessarily connected (Jackson 1993) However, as we saw in our discussion above of the necessary beings problem, the proper interpretation of Hume's dictum is itself a matter of controversy, so emergentism remains controversial.

9. Understanding ‘Physical’: Introductory

Earlier we distinguished two interpretative questions with respect to physicalism, the completeness question and the condition question. So far we have been concerned with the completeness question. I turn now to the condition question, the question of what it is for something (an object, an event, a process,

a property) to be a physical.

The condition question that has received less attention in the literature than the questions we have been studying so far. But it is just as important. Without any understanding of what the physical is, we can have no serious understanding of what physicalism is. After all, if we say that, no two possible worlds can be minimal physical duplicates without being duplicates simpliciter, we don't know what we've said unless we understand what it would take to be a minimal physical duplicate, as opposed (say) to a chemical duplicate or a financial duplicate. (The point here is a quite general one: if Thales says that everything is water, or Up-to-Date-Thales says everything supervenes on water, we don't understand what he says unless he says something about what water is. The physicalist is in the same position.)

So what is the answer to the condition question? If we concentrate for simplicity on the notion of a physical property, we can discern two kinds of answers to this question in the literature. The first ties the notion of a physical property to a notion of a physical theory, for this reason we can call it the theory based conception of a physical property:

> **The theory-based conception**:
> A property is physical iff it either is the sort of property that physical theory tells us about or else is a property which metaphysically (or logically) supervenes on the sort of property that physical theory tells us about.

According to the theory-based conception, for example, if physical theory tells us about the property of having mass, then having mass is a physical property. Similarly, if physical theory tells us about the property of being a rock -- or, what is perhaps more likely, if the property of being a rock supervenes on properties which physical theory tell us about -- then it too is a physical property. (The theory-based conception bears some relation to the notion of physical1 discussed in Feigl 1965; more explicit defense is found in Smart 1978, Lewis 1994, Braddon-Mitchell and Jackson 1996, and Chalmers 1996.)

The second kind of answer ties the notion of a physical property to the notion of a physical object, for this reason we can call it the object-based conception of a physical property:

> **The object-based conception**:
> A property is physical iff: it either is the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents or else is a property which metaphysically (or logically) supervenes on the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents.

According to the object-based conception, for example if rocks, trees, planets and so on are paradigmatic physical objects, then the property of being a rock, tree or planet is a physical property. Similarly, if the property of having mass is required in a complete account of the intrinsic nature of physical objects and their constituents, then having mass is a physical property. (The best examples of philosophers who operate with the object-conception of the physical are Meehl and Sellars 1956 and Feigl 1965; more

recent defense is to be found in Jackson 1998.)

It is important to note that both conceptions of the physical remain silent on the question of whether topic-neutral or functional properties should be treated as physical or not. To borrow a phrase from Jackson (1998), however, it seems best to treat these properties as onlooker properties: given any set of physical properties, one might add onlooker properties without compromising the integrity of the set. But onlooker properties should not be treated as being physical by definition.

10. Understanding ‘Physical’: Further Issues

Along with the concepts of space, time, causality, value, meaning, truth and existence, the concept of the physical is one of the central concepts of human thought. So it should not be surprising that any attempt to come to grips with what a physical property will be controversial. The theory and object conceptions are no different: each has provoked a number of different criticisms. In this section, I will review some main ones.

Circularity

One might object that both conceptions are inadequate because they are circular, i.e., both appeal to the notion of something physical (a theory or an object) to characterize a physical property. But how can you legitimately explain the notion of one sort of physical thing by appealing to another?

However, the response to this is that circularity is only a problem if the conceptions are interpreted as providing a reductive analysis of the notion of the physical. But there is no reason why they should be interpreted as attempting to provide a reductive analysis. After all, we have many concepts that we understand without knowing how to analyze (cf. Lewis 1970). So there seems no reason to suppose that either the theory or object conception is providing anything else a way of understanding the notion of the physical.

The point here is an important one in the context of the condition question. Earlier we said that the condition question was perfectly legitimate because it is legitimate to ask what the condition of being physical is that, according to physicalism, everything has. But this legitimate question should not be interpreted as the demand for a reductive analysis of the notion of the physical. Consider Thales again: it is right to ask Thales what he means by ‘water’ -- and in so doing demand an understanding of the notion of water -- but it is wrong to demand of him a conceptual analysis of water.

Hempel's dilemma

One might object that any formulation of physicalism which utilizes the theory-based conception will be either trivial or false. Carl Hempel (cf. Hempel 1970, see also Crane and Mellor 1990) provided a classic formulation of this problem: if physicalism is defined via reference to contemporary physics, then it is

false -- after all, who thinks that contemporary physics is complete? -- but if physicalism is defined via reference to a future or ideal physics, then it is trivial -- after all, who can predict what a future physics contains? Perhaps, for example, it contains even mental items. The conclusion of the dilemma is that one has no clear concept of a physical property, no concept that is clear enough to do the job that philosophers of mind want the physical to play.

One response to this objection is to take its first horn, and insist that, at least in certain respects contemporary physics really is complete or else that it is rational to believe that it is (cf. Smart 1978, Lewis 1994 and Melnyk 1997). But while there is something right about this, there is also something wrong about it. What is right about it is that there is a sense in which it is rational to believe that physics is complete. After all, isn't it rational to believe that the most current science is true? But even so -- and here is what is wrong about the suggestion -- it is still mistaken to *define* physicalism with respect to the physics that happens to be true in this world. The reason is that whether a physical theory is true or not is a function of the contingent facts; but whether a property is physical or not is a not function of the contingent facts. For example, consider medieval impetus physics. Medieval impetus physics is false (though of course it might not have been) and thus it is irrational to suppose it true. Nevertheless, the property of having impetus -- the central property that objects have according to impetus physics -- is a physical property, and a counterfactual world completely described by impetus physics would be a world in which physicalism is true. But it is hard to see how any of this could be right if physicalism were defined by reference to the physics that we have now or by the physics that happens to be true in our world.

A different response to Hempel's dilemma is that what it shows, if it shows anything, is that a particular proposal about how to define a physical property -- namely, via reference to physics at a particular stage of its development -- is mistaken. But from this one can hardly conclude that we have no clear understanding of the concept at all. As we have seen, we have many concepts that we don't know how to analyze. So the mere fact -- if indeed it is a fact -- that a certain style of analysis of the notion of the physical fails does not mean that there is no notion of the physical at all, still less that we don't understand the notion.

One might object that, while these remarks are perfectly true, they nevertheless don't speak to something that is right about Hempel's dilemma, namely that for the theory-conception to be complete one needs to say a little more about what physical theory is. Here, however, we can appeal to the fact that we have a number of paradigms of what a physical theory is: common sense physical theory, medieval impetus physics, Cartesian contact mechanics, Newtonian physics, and modern quantum physics. While it seems unlikely that there is any *one* factor that unifies this class of theories, it does not seem unreasonable that there is a cluster of factors -- a common or overlapping set of theoretical constructs, for example, or a shared methodology. In short, we might say that the notion of a physical theory is a Wittgensteinian family resemblance concept, and this should be enough to answer the question of how to understand physical theory.

The panpsychism problem

Hempel's dilemma against the theory-conception is similar to an objection that one often hears propounded against the object-conception (cf. Jackson 1998). Imagine the possibility of panpsychism, i.e. the possibility that all the physical objects of our acquaintance are conscious beings just as we are. Would physicalism be true in that situation? It seems intuitively not; however, if physicalism is defined via reference to the object-conception of a physical property then it is hard to see why not. After all, according to that conception, something is a physical property just in case it is required by a complete account of paradigmatic physical objects. But this makes no reference to the *nature* of paradigmatic physical objects, and so allows the possibility that physicalism is true in the imagined situation.

The first thing to say in response is that the mere possibility of panpsychism cannot really be what is at issue in this objection. For no matter how implausible and outlandish it sounds, panpsychism per se is not inconsistent with physicalism (cf. Lewis 1983). After all, the fact that there are *some* conscious beings is not contrary to physicalism -- why then should the possibility that *everything* is a conscious being be contrary to physicalism? So what is at issue in the objection is not panpsychism so much as the possibility that the paradigms or exemplars in terms of which one characterizes the notion of the physical might turn out to be radically different from what we normally assume -- for example, they might turn out to be in some essential or ultimate respect mental. If that were so, it certainly does seem strange to say that physicalism would or could be true.

Once the problem is put like that, however, it is clear that that the problem has a rather similar structure to other problems that arise when one tries to understand a concept in terms of paradigmatic objects which fall under the concept. Suppose one tried to define the concept red in terms of similarity to paradigmatic red things, such as blood. Pursuing this strategy commits one to the idea that the belief that blood is red is a piece of common knowledge shared among all those who are competent with the term. But that seems wrong -- someone who thought that blood was green would be mistaken about blood but not about red. Now this problem is a difficult problem, however -- and this is the crucial point for our purposes -- the problem is also a quite general problem, and not particularly tied to the notion of the physical. So to that extent, the concept of the physical does not seem to be any worse off than the concept of red. (For discussion of the general strategy see Lewis 1997)

The relation between the two conceptions

Perhaps the most interesting issue concerning the theory and object conceptions of a physical property concerns the question of whether they characterize the same class of properties. There are a number of different possibilities here, not of all of which we can discuss. But one that has received some attention in the literature is that physical theory only tells us about the dispositional properties of physical objects, and so does not tell us about the categorical properties, if any, that they have -- a thesis of this sort has been defended by a number of philosophers, among them Russell (1927), Armstrong (1968), Blackburn (1992) and Chalmers (1996). However, if this is correct, it would seem that the physical properties described by the theory conception are only a sub-class of the physical properties described by the object conception. For if physical objects do have categorical properties, those properties will not count as physical by the standards of the theory conception. On the other hand, there seems no reason not to count them as

physical in some sense or other. If that is right, however, then the possibility emerges that the theory- and the object-conceptions characterize distinct classes of properties.

11. Physicalism and the Physicalist World Picture

Perhaps because of its connection to the physical sciences, physicalism is sometimes construed as an entire package of views, which contains the metaphysical thesis I have isolated for discussion as only one part. If we want a name for the entire package of views including the metaphysical claim we might call it the *Physicalist World Picture*. I will close our discussion of the interpretation question by considering the relation between physicalism (the metaphysical claim) and various other items that at least sometimes have been thought to be a part of the Physicalist World Picture.

> **(a) Methodological Naturalism**: the idea that the mode of inquiry typical of the physical sciences will provide theoretical understanding of world, to the extent that this sort of understanding can be achieved. Physicalism is not methodological naturalism because physicalism is a metaphysical thesis not a methodological thesis.
>
> **(b) Epistemic Optimism:** the idea that the mode of understanding typical of the sciences can be used by us, i.e. by human beings, to explain the world in total, to provide a final theory of the world.. Physicalism is not epistemic optimism because, since commitment to physicalism does not commit you to methodological naturalism, it clearly does not commit you to any optimism about the success of that method in the long run.
>
> (c) **Final Theory**: the idea that there *is* a final and complete theory of the world, regardless of whether we can formulate it. One might think it obvious that if physicalism is true, there is a final theory of the world. However, because of some unclarity in the notion of a theory, the issues here are not cut and dried. According to some views, something is a theory only if it is finitely stateable in a language we can understand. If that is so, clearly physicalism does not entail the idea of a final theory. On a looser conception of a theory, however, it is reasonable to say that physicalism entails that there is a final theory.
>
> (d) **Objectivity:** the idea that the final and complete theory of world, if it exists, will not involve any essential reference to particular points of view or experiences. It is reasonable to say that physicalism entails objectivity. However, given the possibilities of non-reductive or *a posteriori* physicalism even here the issues are not settled. On those approaches, it seems possible to have irreducible points of view or experiences supervening on something physical, which compromises objectivity.
>
> (e) **Unity of Science**: the idea that all the branches of sciences developed by us will or should be unified into a single science, usually (but not always) thought of as physics. This thesis is clearly a methodological thesis about how science ought to proceed. As we have seen, however, physicalism is a metaphysical thesis rather than a methodological thesis

about how science ought to proceed. Hence it is not equivalent to the unity of science thesis.

(f) **Explanatory Reductionism**: the idea that all genuine explanations must be couched in the terms of physics, and that other explanations, while pragmatically useful, can or should be discarded as knowledge develops. Physicalism is not explanatory reductionism because, as we saw in our discussion of non-reductive physicalism, physicalism is consistent with the idea that special sciences are quite distinct from physics. One might say that the special sciences are concerned with patterns in the physical that physicists themselves are not concerned with. For that reason the subject matter of the special sciences is distinct from the subject matter of physics.

(g) **Generality of Physics:** the idea that every particular event or process which falls under a law of the special sciences (i.e. sciences other than physics) also falls under a law of physics. In general, this view presupposes a view about laws and explanation -- for example, it implies or seems to imply that special sciences have laws. But physicalism does not entail any such thesis.

(h) **Causal Closure of the Physical**: the idea that every event has a physical cause, assuming it has a cause at all. Strictly speaking, physicalists are not committed to realism about causation, so they are not committed to causal closure. (Of course, many physicalists do think that causal closure is true, as we will see below, but their position does not entail causal closure.)

(i) **Empiricism**: the idea that all knowledge (with the possible exception of conceptual knowledge) is ultimately founded on sensory or perceptual experience. Empiricism can be given a descriptive or a normative reading. On its descriptive reading, it is most likely false. Most of the information that normal humans come to deploy seems to be caused by of both experience *and* inborn structure and maturation. On the normative reading, the claim is that justification is, at the end of the day, based on experience. But this epistemological thesis has nothing to do with physicalism.

(j) **Nominalism**: the idea that there are no abstract objects, i.e., entities not located in space and time, such as numbers, qualities or propositions. If we assume that abstract objects, if they exist, exist necessarily, i.e., exist in all possible worlds, then supervenience physicalism is completely silent on the question of whether abstract objects exist. All supervenience says is that if a world is a minimal physical duplicate of the actual world, it is a duplicate simpliciter. But if abstract objects exist then they clearly exist in both the actual world and any duplicate of the actual world. What this suggests is that nominalism is a distinct issue from physicalism (Schiffer 1987, Stoljar 1996).

(k) **Atheism:** the idea that there is no God as traditionally conceived. In the 17th and 18th century, physicalism (or materialism, as it was then known) was widely viewed as

inconsistent with belief in God (Yolton 1983). Nowadays, this issue is somewhat less discussed. Nevertheless, as we noted previously, *if* God is thought of as essentially non-physical, then Atheism *does* seem to be a consequence of physicalism, at least on some interpretations of the background modal notions.

12. The Case Against Physicalism I: Qualia and Consciousness

Having provided an answer to the interpretation question, I now turn to the truth question: is physicalism (as we have interpreted it so far) true? I will first discuss three reasons for supposing that physicalism is not true. Then I will consider the case for physicalism.

The main argument against physicalism is usually thought to concern the notion of qualia, the felt qualities of experience. The notion of qualia raises puzzles of its own, puzzles having to do with its connection to other notions such as consciousness, introspection, epistemic access, acquaintance, the first-person perspective and so on. However the idea that we will discuss here is the apparent contradiction between the existence of qualia and physicalism.

Perhaps the clearest version of this argument is Jackson's knowledge argument. (There are also a number of other arguments in this area -- for a very good recent discussion, see Chalmers 1996). This argument asks us to imagine Mary, a famous neuroscientist confined to a black and white room. Mary is forced to learn about the world via black and white television and computers. However, despite these hardships Mary learns (and therefore knows) all that physical theory can teach her. Now, if physicalism were true, it is plausible to suppose that Mary knows everything about the world. And yet -- and here is Jackson's point -- it seems she does not know everything. For, upon being released into the world of color, it will become obvious that, inside her room, she did not know what it is like for both herself and others to see colors -- that is, she did not know about the qualia instantiated by particular experiences of seeing colors. Following Jackson (1986), we may summarize the argument as follows:

> P1. Mary (before her release) knows everything physical there is to know about other people.
>
> P2. Mary (before her release) does not know everything there is to know about other people (because she learns something about them on being released).
>
> Conclusion. There are truths about other people (and herself) that escape the physicalist story.

Clearly this conclusion entails that physicalism is false: for if there are truths which escape the physicalist story how can everything supervene on the physical. So a physicalist must either reject a premise or show that the premises don't entail the conclusion.

There are many possible responses to this argument, but here I will briefly mention only three. The first is *the ability hypothesis* due to Lawrence Nemerow (1988) and developed and defended by David Lewis (1994). The ability hypothesis follows Ryle (1949) in drawing a sharp distinction between propositional knowledge or knowledge-that (such as ‘Mary knows that snow is white’) and knowledge-how (such as ‘Mary knows how to ride a bike’), and then suggests that all Mary gains is the latter. On the other hand, P2 would only be true if Mary gained propositional knowledge.

A second response appeals to the distinction between *a priori* and *a posteriori* physicalism. As we saw above, the crucial claim of *a posteriori* physicalism is that (4) -- i.e. the claim that *S* entails *S** -- is *a posteriori*. Since (4) is *a posteriori*, you would need certain experience to know it. But, it is argued, Mary has not had (and cannot have) the relevant experience. Hence she does not know (4). On the other hand, the mere fact that Mary has not had (and cannot have) the experience to know (4) does not remove the possibility that (4) is true. Hence *a posteriori* physicalism can avoid the knowledge argument. (It is an interesting question which premise of the knowledge argument is being attacked by this response. The answer depends on whether (4) is physical or not: if (4) is physical, then the response attacks P1. But if (4) is not physical, the response is that the argument is invalid.).

A third response is to distinguish between various conceptions of the physical. We saw above that potentially the class of properties defined by the theory-conception of the physical was distinct from the class of properties defined by the object-conception. But that suggests that the first premise of the argument is open to interpretation in either of two ways. On the other hand, Jackson's thought experiment only seems to support the premise if it is interpreted in the one way, since Mary learns by learning all that physical *theory* can teach her. But leaves open the possibility that one might appeal to the object-conception of the physical to define a version of physicalism which evades the knowledge argument.

One of the most lively areas of philosophy of mind concerns the issue of which if any of these responses to the knowledge argument will be successful. The ability response raises questions about whether know-how is genuinely non-propositional (cf. Lycan (1996), Loar (1997) and Stanley and Williamson (forthcoming)), and about whether it gets the facts right to begin with (Braddon Mitchell and Jackson 1996). As against *a posteriori* physicalism, it has been argued both that it rests on a mistaken approach to the necessary *a posteriori* (Chalmers 1996, 1999, Jackson 1998), and that the promise of the idea is chimerical anyway (cf. Stoljar 2000). The third response raises questions about the distinction between the object and the theory conception of the physical and associated issues about dispositional and categorical properties (Cf. Chalmers 1996, Lockwood (1992), and Stoljar 2000, 2001.)

13. The Case Against Physicalism II: Meaning and Intentionality

Philosophers of mind often divide the problems of physicalism into two: first, there are the problems of qualia, typified by the knowledge argument; second, there are problems of intentionality. The intentionality of mental states is their aboutness, their capacity to represent the world as being a certain

way. One does not simply think, one thinks *of* (or *about*) Vienna; similarly, one does not simply believe, one believes *that* snow is white. Just as in the case of qualia, some of the puzzles of intentionality derive from facts internal to the notion, and from the relation of this notion to the others such as rationality, inference and language. But others derive from the fact that it seems difficult to square the fact that mental states have intentionality with physicalism. There are a number of ways of developing this criticism but much recent work as concentrated on a certain line of argument that Saul Kripke has found in the work of Wittgenstein (1982).

Kripke's argument is best approached by first considering what is often called a dispositional theory of linguistic meaning. According to the dispositional theory, a word means what it does -- for example, the word ‘red’ means red -- because speakers of the word are disposed to apply to word to red things. Now, for a number of reasons, this sort of theory has been very popular among physicalists. First, the concept of a disposition at issue here is clearly a concept that is compatible with physicalism. After all, the mere fact that vases are fragile and sugar cubes are soluble (both are classic examples of dispositional properties) does not cause a problem for physicalism, so why should the idea that human beings have similar dispositional properties? Second, it seems possible to develop the dispositional theory of linguistic meaning so that it might apply also to intentionality. According to a dispositional theory of intentionality, a mental concept would mean what it does because thinkers are disposed to employ the concept in thought in a certain way. So a dispositional theory seems to hold out the best promise of a theory of intentionality that is compatible with physicalism.

Kripke's argument is designed to destroy that promise. (In fact, Kripke's argument is designed to destroy considerably more than this: the conclusion of his argument is a paradoxical one to the effect that there can be no such a thing as a word's having a meaning. However, we will concentrate on the aspects of the argument that bear on physicalism.) In essence his argument is this. Imagine a situation in which (a) the dispositional theory is true; (b) the word ‘red’ means red for a speaker S; and yet (c) the speaker misapplies the word -- for example, S is looking at a white thing through rose-tinted spectacles and calls it red. Now, in that situation, it would seem that S is disposed to apply ‘red’ to things which are (not merely red but) either-red-or-white-but-seen-through-rose-tinted-spectacles. But then, by the theory, the word ‘red’ means (not red but) either-red-or-white-as-seen-through-rose-tinted-spectacles. But that contradicts our initial claim (b), that ‘red’ means red. In other words, the dispositional theory, when combined with a true claim about the meaning of word, plus a truism about meaning -- that people can misapply meaningful words -- leads to a contradiction and is therefore false.

How might a physicalist respond to Kripke's argument? As with the knowledge argument, there are many responses but here I will mention only two. The first response is to insist that Kripke's argument neglects the distinction between *a priori* and *a posteriori* physicalism. Kripke often does say that according to the dispositionalist, one should be able to ‘read off’ truths about meaning from truths a physicalist can reject. (For a proposal like this, see Horwich 2000.) However, the problem with this proposal is, as we have seen, that its background account of the necessary *a posteriori* is very controversial. As we saw, *a posteriori* physicalists are committed to what we called the non-derived view about necessary *a posteriori* truths. But the non-derived view has come under strident attack in recent times.

The second response is to defend the dispositional theory against Kripke's argument. One way to do this is to argue that Kripke's argument only works against a very simple dispositionalism, and that a more complicated version of such a theory would avoid these problems. (For a proposal along these lines, see Fodor 1992 and the discussion in Braddon-Mitchell and Jackson 1996). A different proposal is to argue that Kripke's argument underestimates the complexity in the notion of a disposition. The mere fact that in certain circumstances someone would apply ‘red’ to white things does not mean that they are disposed to apply red to white things -- after all, the mere fact that in certain circumstances something would burn does not mean that it flammable in the ordinary sense. (For a proposal along these lines see Hohwy 1998, and Heil and Martin 1998)

As with the knowledge argument, the issues surrounding Kripke's argument are very much wide open. But it is important to note that most philosophers don't consider the issues of intentionality as seriously as the issue of qualia when it comes to physicalism. In different vocabularies, for example, both Block (1995) and Chalmers (1996) distinguish between the intentional aspects of the mind or consciousness, and the phenomenal aspects or qualia, and suggest that it is really the latter that is the central issue. As Chalmers notes (1996; p. 24), echoing Chomsky's famous distinction, the intentionality issue is a *problem*, but the qualia issue is a *mystery*.

14. The Case Against Physicalism III: Methodological Issues

The final argument I will consider against physicalism is of a more methodological nature. It is sometimes suggested, not that physicalism is false, but that the entire ‘project of physicalism’ -- the project in philosophy of mind of debating whether physicalism is true, and trying to establish or disprove its truth by philosophical argument -- is misguided. This sort of argument has been mounted by a number of writers, but perhaps its most vocal advocate has been Noam Chomsky (2000; see also Searle 1992, 1999).

It is easiest to state Chomsky's criticism by beginning with two points about methodological naturalism. In general it seems rational to agree with the methodological naturalists that the best hope for a theoretical understanding of the world is by pursuing the methods which are typical of the sciences. It would then seem rational as a special case that our best hope for a theoretical understanding of consciousness or experience is by pursuing the methods of the sciences -- by pursuing, as we might put it, the naturalistic project with respect to consciousness. So Chomsky's first point is that it is rational to pursue the naturalistic project with respect to consciousness.

Chomsky's second point is that the physicalist project in philosophy of mind is on the face of it rather different from the naturalistic project. In the first place, the physicalist project is, as we have noted, usually thought of a piece of metaphysics. But there is nothing metaphysical about the naturalistic project, it simply raises questions about what we can hope to explain. In the second place, the physicalist project is normally thought of as being amenable to philosophical argument, whereas it is completely unclear where philosophical argument would enter the naturalistic project. In short, there doesn't seem anything

particularly ‘philosophical’ about the naturalistic project -- it simply applies the methods of science to consciousness. But the physicalist project is central to analytic philosophy.

It is precisely at the place where the physicalist project departs from the naturalistic project that Chomsky's criticism begins to take shape. For insofar as it is different from the naturalistic project, there are a number of ways in which the physicalist project is questionable. First, it is hard to see what the project might be -- it is true that throughout the history of philosophy and science one encounters suggestions that one might find out about the world in ways that are distinct from the ones used in the sciences, but these suggestions have always been rather obscure. Second, it is hard see how this sort of project could recommend itself to physicalists *themselves* -- such a project seems to be a departure from methodological naturalism but most physicalists endorse methodological naturalism as a matter of fact. On the other hand, if the physicalist project does not depart from the naturalistic project, then the usual ways of talking and thinking about that project are highly misleading. For example, it is misleading to speak of it a piece of metaphysics and philosophy as opposed to a piece of ordinary science.

In sum, Chomsky's criticism is best understood as a kind of dilemma. The physicalist project is either identical to the naturalistic project or it is not. If it is identical, then the language and concepts that shape the project are potentially extremely misleading; but if it is not identical, then there are a number of ways in which it is illegitimate.

How is one to respond to this criticism? In my view, the strongest answer to Chomsky accepts the first horn of his dilemma and suggests that what philosophers of mind are really concerned with is the naturalistic project. Now, of course, what concerns them is not the details of the project -- that would not distinguish them from working scientists. Rather they are concerned with what the potential limits of the project are.

This is a theme which has reached its best expression in the work of Thomas Nagel (1980, 1984, 1999) and allied work by Bernard Williams (1984). According to them, any form of scientific inquiry will at least be objective, or will result in an objective picture of the world. On the other hand, we have a number of arguments -- the most prominent being the knowledge argument -- which plausibly show that there is no place for experience or qualia in a world that is described in purely objective terms. If Nagel and Williams are right that any form of scientific inquiry will yield a description of the world in objective terms, the knowledge argument is nothing less than a negative argument to the effect that the naturalistic project with respect to consciousness will not succeed.

If what is at issue is the limits of the naturalist project, why is the debate so often construed as a metaphysical debate rather than a debate about the limits of inquiry? In answer to this question, we need to sharply divorce the background metaphysical framework within which the problems of philosophy of mind find their expression, and the problems themselves. Physicalism is the background metaphysical assumption against which the problems of philosophy of mind are posed and discussed. Given that assumption, the question of the limits of the naturalistic project *just is* the question of whether there can be experience in a world that is totally physical. Nevertheless, when properly understood, the problems that philosophers of mind are interested in are not with the framework themselves, and to that extent are

not metaphysical. Thus, the common phrase 'metaphysics of mind' is misleading.

15. The Case for Physicalism

Having considered one side of the truth question, I will now turn to the other: what reason is there for believing that physicalism is true?

The first thing to say when considering the truth of physicalism is that we live in an overwhelmingly physicalist or materialist intellectual culture. The result is that, as things currently stand, the standards of argumentation required to persuade someone of the truth of physicalism are much lower than the standards required to persuade someone of its negation. (The point here is a perfectly general one: if you already believe or want something to be true, you are likely to accept fairly low standards of argumentation for its truth.)

However, while it might be difficult to assess dispassionately the arguments for or against physicalism, this is still something we should endeavor to do. Here I will review two arguments that are commonly thought to establish the truth of physicalism. What unites the arguments is that each takes something from the physicalist world-picture which we considered previously and tries to establish the metaphysical claim that everything supervenes on the physical.

The first argument is (what I will call) *The Argument from Causal Closure*. The first premise of this argument is the thesis of the Causal Closure of the Physical -- that is, the thesis that every event which has a cause has a physical cause. The second premise is that mental events cause physical events -- for example we normally think that events such wanting to raise your arm (a mental event) cause events such as the raising of your arm (a physical event). The third premise of the argument is a principle of causation that is often called the exclusion principle (Kim 1993, Yablo 1992). The correct formulation of the exclusion principle is a matter of some controversy but a formulation that is both simple and plausible is the following:

> **Exclusion Principle**
> If an event *e* causes event *e**, then there is no event *e#* such that *e#* is non-supervenient on *e* and *e#* causes *e**.

The conclusion of the argument is the mental events are supervenient on physical events, or more briefly that physicalism is true. For of course, if the thesis of Causal Closure is true then behavioral events have physical causes, and if mental events also cause behavioral events, then they must supervene on the physical if the exclusion principle is true.

The Argument from Causal Closure is perhaps the dominant argument for physicalism in the literature today. But it is somewhat unclear whether it is successful. The most promising response for the anti-physicalist is to reject the second premise and to adopt a version of what is called epiphenomenalism, the view that mental events are caused by, and yet do not cause, physical events. The argument against this

position is usually epistemological: if pains don't cause pain behavior how can it be that your telling me that you are in pain gives me any reason for supposing you are? It might seem that epiphenomenalists are in trouble here, but as a number of recent philosophers have argued, the issues here are very far from being settled (Chalmers 1996, Hyslop 1999). The crucial point is that the causal theory of evidence is open to serious counterexamples so it is unclear that it can be used against epiphenomenalism effectively.

A different sort of response is to reject the causal principles on which the argument is based. As against the exclusion principle, for example, it is often pointed out that certain events are overdetermined. The classic example is the firing squad: both the firing by soldier A and by soldier B caused the prisoner's death but since these are distinct firings, the exclusion principle is false. However, while this line of response is suggestive, it is in fact rather limited. It is true that the case of the firing squad represents an exception to the exclusion principle -- an exception that the principle must be emended to accommodate. But is difficult to believe that it represents an exception that can be widespread. A more searching response is to reject the very idea of causal closure on the grounds, perhaps, that (as Bertrand Russell (1917) famously argued) causation plays no role in a mature portrayal of the world. Once again, however, the promise of this response is more imagined than real. While it is true that many sciences do not explicitly use the notion of causation, it is extremely unlikely that they do not imply that various causal claims are true.

The second argument for physicalism is (what I will call) *The Argument from Methodological Naturalism*. The first premise of this argument is that it is rational to be guided in one's metaphysical commitments by the methods of natural science. Lying behind this premise are the arguments of Quine and others that metaphysics should not be approached in a way that is distinct from the sciences but should rather be thought of as continuous with it. The second premise of the argument is that, as a matter of fact, the metaphysical picture of the world that one is led to by the methods of natural science is physicalism. The conclusion is that it is rational to believe physicalism, or, more briefly that physicalism is true.

The Argument from Methodological Naturalism has received somewhat less attention in the literature than the Argument from Causal Closure. But it seems just as persuasive -- in fact, rather more so. For how might one respond? One possibility is to reject its first premise. But this is not something that most people are attracted to (or at least are attracted to explicitly.)

The other possibility is to reject its second premise. However, once it is appreciated what physicalism is -- and, more important, what it is not -- it is not terribly clear what this would amount to or what the motivation for it would be. In the first place, our earlier discussion shows that physicalism is not inconsistent with explanatory autonomy of the various sciences, so that one should not reject physicalism merely because one can't see how to reduce those sciences to others. In the second place, while it is perfectly true that there are examples of non-physicalist approaches to the world -- vitalism in biology is perhaps the best example -- this is beside the point. The second premise of the Argument from Methodological Naturalism does not deny that other views are possible, it simply says that physicalism is the most likely view at the moment. Finally, one might be inclined to appeal to arguments such as the knowledge argument to show that physicalism is false, and hence that methodological naturalism could not show that physicalism is false. However, this suggestion represents a sort of confusion about the

knowledge argument. As we saw above, if successful the knowledge argument suggests, not simply that physicalism is false but that any approach to the world that is compatible with methodological naturalism is false. But if that is so, it is mistaken to suppose that the knowledge argument gives one any reason to endorse anti-physicalism if that is supposed to be a position compatible with methodological naturalism.

16. Concluding Remarks and Further Questions

That completes our discussion of physicalism. As you can see, it is sketchy in many places, and this suggests that there is much further work to be done before we arrive at a final assessment of the doctrine of physicalism and the role that it plays in contemporary thought. It may be helpful to end, however, by identifying three areas that seem to me especially deserving of attention in the future.

First, *a priori* and *a posteriori* physicalism. As we have seen, the distinction between *a posteriori* and *a priori* is a crucial one in a number of respects. If *a posteriori* physicalism can be made out, then we have potential answers to both the qualia problem and the intentionality problem. In addition, we have an interpretation of emergentism. On the other hand, it is unclear if the notion is at the end of the day coherent.

Second, the relation between the theory-conception and the object-conception of the physical. As we have also seen, this distinction is an important one in that it allows us to answer the knowledge argument without appealing to the non-derivation view of the necessary *a posteriori*. On the other hand, the distinction is also controversial in a number of respects.

Finally, the relation between objectivity and physicalism. As we have seen Nagel and Williams both think that objectivity is a presupposition of the methodological naturalism that so many contemporary philosophers find attractive. But if that is right there is no point developing a version of physicalism -- or any approach to the world -- that rejects it.

Bibliography

- Armstrong, D., 1968, *A Materialist Theory of the Mind*, Routledge.
- Berkeley, G., 1710, *Principles of Human Knowledge*.
- Bacon, J., 1990, ‘Van Cleve Versus Closure’, *Philosophical Studies*, 58.
- Blackburn, S., 1992, ‘Filling in Space’, *Analysis*, pp. 60-65.
- Block, N., 1980, ‘Troubles with Functionalism’, in N. Block (ed), *Readings in the Philosophy of Psychology*, Vol. I, Cambridge, MA: Harvard University Press, 1980.
- Block, N., 1995, ‘On a Confusion About a Function of Consciousness’, *Behavioral and Brain Sciences*, 18/(2): 227-287.
- Block, N., and Stalnaker, R., 1999, ‘Conceptual Analysis, Dualism and the Explanatory Gap’, *Philosophical Review*, forthcoming.
- Braddon-Mitchell, D., and Jackson, F., 1996, *Philosophy of Mind and Cognition*, Blackwell.

- Byrne, A., 1993, *The Emergent Mind*, Ph.D. Dissertation. Princeton.
- Byrne, A., 1999, ‘Cosmic Hermeneutics’, in J. Tomberlin (ed), *Philosophical Perspectives*
- Chalmers, D., 1996, *The Conscious Mind*, New York: Oxford University Press
- Chalmers, D., 1999, ‘Materialism and the Metaphysics of Modality’, *Philosophy and Phenomenological Research*, forthcoming.
- Chomsky, N., 1994, ‘Language and Nature’, *Mind* 104/413, 1995.
- Chomsky, N., 1994b, ‘Noam Chomsky’, in S. Guttenplan (ed), *A Companion to the Philosophy of Mind*, Blackwell.
- Chomsky, N., 2000, *New Horizons in the Study of Language and Mind*, Cambridge.
- Crane, T. and Mellor, D.H., 1990, ‘There is no Question of Physicalism’, *Mind*, 99: 185.
- Davidson, D. 1970 . ‘Mental Events’. In Davidson, D (1980) *Essays on Actions and Events*. Oxford.
- Descartes, R., 1641, *Meditations on First Philosophy*, in Cottingham, et al. (eds), *The Philosophical Writings of Rene Descartes*, Cambridge, 1985.
- Dijksterhuis, E.J., 1961, *The Mechanization of the World-Picture*, Oxford: Clarendon.
- Field, H., 1972, ‘Tarski's Theory of Truth’, *Journal of Philosophy*, 69, 347-75.
- Field, H., 1992, ‘Physicalism’, in J. Earman (ed), *Inference, Explanation and Other Frustrations*, Berkeley: University of California Press
- Field, H., 1994, ‘Deflationist Views of Meaning and Content’, *Mind*, 103/411.
- Feigl, H., 1967, ‘The "Mental" and the "Physical"’, Minneapolis: University of Minnesota Press (originally published in 1958).
- Feinberg, G., 1966, ‘Physics and the Thales Problem’, *Journal of Philosophy*, 63.
- Fodor, J.A., 1974, ‘Special Sciences: Or, The Disunity of Science as a Working Hypothesis’, reprinted in J. Fodor, *Representations*, MIT Press, 1981.
- Fodor, J.A., 1992, *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press
- Foster, J., 1982, *The Case for Idealism*, London: Routledge
- Foster, J., 1991, *The Immaterial Self: A Defence of the Cartesian Dualist Conception of Mind*, London: Routledge.
- Gold, I., and Stoljar, D., 1999, ‘A neuron doctrine in the philosophy of neuroscience’, *Behavioral and Brain Sciences*, 22/5
- Haugeland, J., 1983, ‘Weak Supervenience’, *American Philosophical Quarterly*
- Hempel, C., 1970, ‘Reduction: Ontological and Linguistic Facets’, in S. Morgenbesser, et al. (eds), *Essays in Honor of Ernest Nagel*, New York: St Martin's Press.
- Horgan, T., 1983, ‘Supervenience and Microphysics’, *Pacific Philosophical Quarterly*, 63, pp. 29-43.
- Horgan, T., 1993, ‘From Supervenience to Superdupervenience: Meeting the Demands of a Material World’, *Mind*, 102/408.
- Horwich, P., 1995, ‘Meaning, Use, and Truth’, *Mind*, 204/414.
- Horwich, P., 2000, *Meaning*, Oxford.
- Hohwy, J., 1998, *Meaning as Use*, Ph.D Dissertation, Australian National University.
- Hyslop, A., 1999, ‘Methodological Epiphenomenalism’, *Australasian Journal of Philosophy*.
- Jackson, F., 1982, ‘Epiphenomenal Qualia’, *Philosophical Quarterly*, 32: 127-36
- Jackson, F., 1986, ‘What Mary Didn't Know’, *Journal of Philosophy*, 83, pp. 291-5

- Jackson, F., 1998, *From Metaphysics to Ethics: A Defense of Conceptual Analysis*, Oxford: Clarendon.
- Jackson, F., 1993, 'Armchair Metaphysics', in J. Hawthorne and M. Michael (eds), *Philosophy in Mind*, Amsterdam: Kluwer.
- Jackson, F., and Pettit, P., 1992, 'In Defense of Explanatory Ecumenism', *Economics and Philosophy*, 8, pp. 1-21.
- Kim, J., 1993, *Mind and Supervenience*, Cambridge: Cambridge University Press.
- Kim, J., 1998, *Mind in a Physical World*, Cambridge: Cambridge University Press.
- Kripke, S., 1980, *Naming and Necessity*, Cambridge, MA: Harvard University Press.
- Kripke, S., 1982, *Wittgenstein on Rules and Private Language: An Elementary Exposition*, Oxford: Basil Blackwell.
- Lewis, D., 1970, 'How to Define Theoretical Terms', *Journal of Philosophy* 67: 427-46.
- Lewis, D., 1983, 'New Work for a Theory of Universals', *Australasian Journal of Philosophy*, 61/4.
- Lewis, D., 1986, *On the Plurality of Worlds*, Oxford: Blackwell.
- Lewis, D., 1994, 'Reduction of Mind', in S. Guttenplan (ed), *A Companion to the Philosophy of Mind*, Oxford: Blackwell.
- Lewis, D., 1997, 'Naming the Colours', *Australasian Journal of Philosophy*, 75, pp. 325-342.
- Loar, B., 1997, 'Phenomenal States', in N. Block, et al. (eds), *The Nature of Consciousness: Philosophical Debates*, Cambridge, MA: MIT Press.
- Loar, B., 1999, "David Chalmers' *The Conscious Mind&*", *Philosophy and Phenomenological Research*, forthcoming.
- Lockwood, M., 1989, *Mind, Brain and Quantum*, Oxford: Basil Blackwell.
- Lycan, W., 1996, *Consciousness and Experience*, Cambridge, MA: MIT Press.
- Maudlin, T., 1996, 'On the unification of physics', *Journal of Philosophy*, 93: 129-144.
- Melnick A., 1997, "How To Keep The 'Physical' in Physicalism", *Journal of Philosophy*, 94: 622-637.
- MacLaughlin, B., 1992, 'The Rise and Fall of British Emergentism', in Beckerman, et al. (eds), *Emergence or Reduction?*, Berlin: De Gruyter.
- Nagel, E., 1961, *The Structure of Science*, New York: Harcourt, Brace and World.
- Nagel, T., 1974, 'What is it like to be a bat', *Philosophical Review*, 4: 435-50.
- Nagel, T., 1983, *The View from Nowhere*, New York: Oxford.
- Nemerow, L., 1988, 'Physicalism and the Cognitive Role of Acquaintance', in W. Lycan (ed), *Mind and Cognition*, Oxford: Blackwell.
- Papineau, D., 1996, *Philosophical Naturalism*, Oxford: Blackwell.
- Paull, C., and Sider, T., 1992, 'In Defense of Global Supervenience', *Philosophical and Phenomenological Research*, 52.
- Poland, J., 1994, *Physicalism: The Philosophical Foundations*, Oxford: Clarendon.
- Putnam, H., 1975, 'Philosophy and our mental life', in H. Putnam, *Mind, Language and Reality: Philosophical Papers*, Vol. 2, Cambridge: Cambridge University Press.
- Russell, B., 1917, 'On the Notion of Cause', in B. Russell, *Mysticism and Logic*, Penguin, 1963.
- Russell, B., 1927, *The Analysis of Matter*, London: Kegan Paul.
- Ryle, G., 1949, *The Concept of Mind*, London: Routledge.

- Schiffer, S., 1987, *Remnants of Meaning*, Cambridge, MA: MIT Press.
- Shoemaker, S., 1994, 'Phenomenal Character', *Nous*, 28, pp.21-38
- Smart, J.J.C., 1959, 'Sensations and Brain Processes', reprinted in D. Rosenthal (ed), *Materialism and the Mind-Body Problem*, Hackett, 1987.
- Smart, J.J.C., 1978, 'The Content of Physicalism', *Philosophical Quarterly*, 28, pp. 239-41.
- Smith, M., and Stoljar, D., 1998, 'Global Response-Dependence and Noumenal Realism', *The Monist*, 81/1.
- Steward, H., 1996, *The Ontology of Mind*, Oxford: Clarendon.
- Stanley, J., and Williamson, T., forthcoming, 'Knowing How'.
- Stalnaker, R., 1996, 'Varieties of Supervenience', *Philosophical Perspectives*, 10: 221-241.
- Stoljar, D., 1996, 'Nominalism and Intentionality', *Noûs*, 30/2, pp. 261-281.
- Stoljar, D., 2000, 'Physicalism and the Necessary *A Posteriori*', *Journal of Philosophy*, 97/1 (January), pp. 33-54.
- Stoljar, D., 2001, 'Two Conceptions of the Physical', *Philosophy and Phenomenological Research*, forthcoming.
- Stoljar, D., 2001a, 'The Conceivability Argument and Two Conceptions of the Physical', *Philosophical Perspectives*, forthcoming.
- Stoljar, D., 2001b, 'Causation: Physical, Mental and Social', *International Encyclopedia of the Social and Behavioral Sciences*, forthcoming.
- Stroud, B., 1986, 'The Physical World', *Proceedings of the Aristotelian Society*.
- Van Cleve, J., 1990, 'Supervenience and Closure', *Philosophical Studies*, 58.
- Williams, D., 1985, *Ethics and the Limits of Philosophy*, Fontana.
- Yablo, S., 1992, 'Mental Causation', *The Philosophical Review*, (April), pp. 245-280.
- Yablo, S., 1999, 'Concepts and consciousness', *Philosophy and Phenomenological Research*, forthcoming.
- Yolton, R., 1983, *Thinking Matter*, Minneapolis: University of Minnesota Press.

Other Internet Resources

- Contemporary Philosophy of Mind: An Annotated Bibliography
  , by David Chalmers (U. Arizona)

Related Entries

behaviorism | color | Davidson, Donald | epiphenomenalism | multiple realizability | qualia

Acknowledgements

The author would like to thank Rich Cameron, Robert Pasnau, Stewart Saunders, and particularly David Chalmers for their help in constructing this entry. The principal editor would like to thank one of the

Encyclopedia readers, Joshua R. Stern, for his help proofreading and discovering numerous typographical errors. His volunteer efforts were entirely unsolicited and very much appreciated.

Daniel Stoljar
daniel.stoljar@colorado.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 13, 2001
Content last modified: February 13, 2001

Stanford Encyclopedia of Philosophy

Color

Colors are of philosophical interest for two kinds of reason. One is that colors comprise such a large and important portion of our social, personal and epistemological lives and so a philosophical account of our concepts of color is highly desirable. The second reason is that trying to fit colors into accounts of metaphysics, epistemology and science leads to philosophical problems that are intriguing and hard to resolve. Not surprisingly, these two kinds of reasons are related. The fact that colors are so significant in their own right, makes more pressing the philosophical problems of fitting them into more general metaphysical and epistemological frameworks.

The Philosophy of Color

The visual world, the world as we see it, is a world populated by colored objects. Typically, we see the world as having a rich tapestry of colors or colored forms---fields, mountains, oceans, hairstyles, clothing, fruit, plants, animals, buildings, and so on. Colors are important in both identifying objects, i.e., in locating them in space, and in re-identifying them. So much of our perception of physical things involves our identifying objects by their appearance, and colors are typically essential to an object's appearance, that any account of visual perception must contain some account of colors. Since visual perception is one of the most important species of perception and hence of our acquisition of knowledge of the physical world, and of our environment, including our own bodies, a theory of color is doubly important.

Despite much thought, over thousands of years, by philosophers and scientists, however, we seem little closer now to an agreed account of color than we ever were. The disagreement is reflected in the fact that some theorists believe colors to be perceiver-relative, e.g., dispositions or powers to induce experiences of a certain kind, or to appear in certain ways to observers of a certain kind. Others take them to be objective, physical properties of objects. Among the latter group, some take these properties to be physical microstructures, while others regard colors as sui generis irreducible properties of physical bodies, and yet others take them to be dispositional properties to affect light. Finally, there are even some who deny that there are colors in the world at all: there are none of the colors, it is claimed, that we naturally and normally and unreflectingly attribute to objects.

The major problem with color has to do with fitting what we seem to know about colors into what science, particularly physics, tells us about physical bodies and their qualities. More specifically, we experience color as an intrinsic feature of the surfaces of physical bodies, or as a property spread throughout a volume, e.g., of wine. But, or so it seems, the physical account of these physical objects finds no place for such qualities. It is this problem that historically has led the major physicists who have thought about color, to hold a common view: that the colors we ordinarily and naturally take objects to possess, are such that physical objects do not actually have them. Oceans and skies are not blue in the way that we naively think, nor are apples red, (nor green). Colors of that kind, it is believed, have no place in the physical account of the world that has developed from the 16th Century to this century. Physicists who have subscribed to this doctrine include the luminaries: Galileo, Boyle, Descartes, Newton, Young, Maxwell and Helmholtz. In this doctrine, they were joined by a number of fellow travelers, including most famously, John Locke.

Such a view is clearly paradoxical, given what was said above, about the ubiquity of colors in the perceived world, and about the importance of colors in the identification and re-identification of physical objects. It is possible to mitigate the paradoxical character of the doctrine by drawing a distinction between two concepts of color: (i) color as a sensory quality, intrinsic to our sensory experiences; (ii) color as a power, to induce sensory experiences with color, understood as a sensory quality. On this account, color terms have a systematic ambiguity. Provided we take account of the ambiguity, no harm is done, and much benefit derived. According to this view, then, in one sense of ‘color’, physical objects have colors, for they have the power to induce experiences of color, but in the other sense, they do not. That is to say, there is no problem if the second concept, color-as-it-is-in-experience, is restricted in its use so as to apply to the experiential quality. Problems arise, however, when it is used, naively and unreflectingly, to apply to physical bodies: we naively suppose the experiential quality to be an intrinsic

quality of the physical object. When we enjoy visual experiences, then in some sense we project the sensory quality in our experience on to physical objects. One who exploited this idea to great effect was David Hume, who used our experience of color as a model for thinking about the way we attribute causal connections, necessity and moral predicates to objects and situations in the world (remembering that Hume adapted the model to his own terminology of ‘impressions’ and ‘ideas’).

With this in mind, the Descartes-Locke position is best expressed as implying that it is possible for perceivers when applying color concepts to physical bodies, to use different concepts, reflecting different attitudes which one may adopt. One attitude is what might be called a 'natural' attitude: a naive, pre-reflective, natural attitude; the other involves a more sophisticated attitude. One concept is a pre-reflective, pre-theoretical concept, the other is more sophisticated. While the Descartes-Locke view has had, and continues to have, a strong influence among many of the scientists who work on color, there has always been strong opposition to it among the philosophical community.

Various reasons have been given for dissatisfaction with the physicists' position. It has been variously argued that: (i) the notion of ‘color as it is in experience’ is incoherent; (ii) the physicists' doctrine encapsulates a confusion about what the ordinary, natural concept of color is; (iii) those who defend the doctrine forget that there are other sciences besides physics. For example, there are biological sciences such as zoology, botany, ecology, and so on, in which colors do have a role to play that they do not have in physics and chemistry; (iv) those who defend the doctrine forget that color has a social role to play; that colors are important in social life, and the criteria for application of color predicates are based on that social life; (v) the ordinary, natural concept of color, the ‘folk’ concept is not as the physicists and their friends believe.This last criticism takes several different forms:

1. according to the natural notion, colors are simple, objective, intrinsic qualities of objects, qualities whose natures are manifest in perception: John Campbell, P.M.S.Hacker;
2. according to the natural notion, colors are unknown qualities of physical bodies, qualities which appear to us in distinctive ways: Thomas Reid, Frank Jackson, David Lewis, Brian McLaughlin.
3. according to the natural notion, colors are dispositional properties, powers to appear in appropriate ways: Michael Dummett, Gareth Evans.

The criticism of the physicists' position that their notion of ‘color as it is in experience’ is confused is widely held. [See Westphal (1987), Hacker (1987) and Evans (1980)]. This criticism has in turn been attacked on the ground that it is based on uncharitable interpretations of the texts, ones made without serious attempt to make sense of the physicists' position. [See Maund (1991) and (1995) pp. 6-24; 30-33]

The first of these formulations of the objection, we should note, opens the way for a modification of the physicists' position on color, as expressed by Descartes and Locke. It construes the natural/folk concept of color in such a way that colors are taken to be perceiver-independent, intrinsic, qualitative features of physical surfaces, volumes and other physical entities such as skies, rainbows and flames. This is the kind of color that our visual experiences represent objects as having. The Descartes-Locke position can be reframed so as to adopt this formulation of the folk concept, and to argue that no instances of this concept are physically actualised. According to this way of thinking, Descartes and Locke were right that given

the natural (naive, pre-reflective) concept of color, we can conclude that, in this sense, objects do not have colors. Their way of characterising the concept, however, is at fault. It should not be described as 'color as it is in experience'. There may very well be a coherent notion of 'color-as-it-is-in-experience' or 'color as a sensory quality', but that notion is not the natural concept of color. The natural concept is more plausibly construed as a concept of a certain kind of property: it is a perceiver-independent, intrinsic, qualitative feature of physical surfaces (i.e. it is not a dispositional property either to affect light or to appear to observers). This re-formulation of the Descartes-Locke view may be described as the Illusion theory of Colors.

These theorists can be regarded as being right that there is a natural concept of color which is such that objects do not have colors. It is their way of characterising the concept, however, that is at fault. This concept should not be expressed as "color as it is in experience". There may very well be a coherent notion of "color-as-it-is-in-experience" or "color as a sensory quality", but that notion is not the natural concept of color. The natural concept is more plausibly construed as a concept of a certain kind of property: an intrinsic perceiver-independent feature of physical surfaces (i.e. it is not a dispositional property either to affect light or to appear to observers). It is possible, moreover, to specify the natural concept in more detail: to be red is to have a certain feature, one that satisfies a range of conditions. One such condition is that colors together form a system of qualitative features, that resemble and differ from each other in systematic ways. A second condition is that the color has a causal role to play in the visual identification or recognition of the color.

Given that colors can be specified as properties of this kind, it is then possible to argue that there are no actual properties that satisfy all the conditions set down. Accordingly, those in the Locke-Descartes-Helmholtz tradition who emphasize the need for a distinction between color-as-in-physical-objects and color-as-in-experience are best interpreted as thinking that there is no physical feature in physical objects that satisfies all the requirements (that serves all the required roles) of color, as it is naturally conceived. This reconstruction of the Descartes-Locke preserves the other element of that position, namely that the right way to think of colors is as mind-dependent dispositional properties. This is the best way, it is claimed, to make sense of colors, taken to be properties of physical bodies.

Clearly those in the Descartes-Locke tradition make two substantial claims. One concerns the character of the ordinary, naive, pre-reflective concept of color; the other is a proposal, of what form our color concepts should take, for scientific and metaphysical purposes, i.e., if we want to think clearly and scientifically about colors. the proposal is that color, thought of as a property that physical objects possess, should be thought of as dispositional property: a power to induce experiences of color, in normal perceivers, in the right kind of circumstances. [This concept of color, it should be noted, requires n addition, another concept of color, color-as-it-is-in-experience.]

There is a substantial body of opposition to this element in the Descartes-Locke tradition as well. For some thinkers, the ones referred to above, the opposition is coupled with, and allows for, the ordinary concept. For others, the question of what colors are essentially is decided on grounds separate from the question of how colors are conceptualized, by ordinary perceivers.

We can set out, in summary form, the set of leading rival theories:

1. Colors are objective, intrinsic properties: they are sui generis, irreducible properties, supervenient on microstructural properties of the type cited in 2, below. [P.M.S. Hacker, J. Campbell]
2. Colors are objective, intrinsic properties of physical bodies. They are to be identified with, or are reducible to, microstructural properties of the bodies that possess them. [F. Jackson, T.Reid]
3. Colors are objective properties of bodies but they are dispositional, light-related properties; dispositions to modify light (to emit, reflect, absorb, differentially reflect and absorb, transmit or scatter light), in the right way, and in the right proportion. [D.M. Armstrong, J. Westphal, D.R. Hilbert]
4. Colors are dispositional properties: powers to induce in observers of a specifiable kind, physical responses of a kind that are peculiar to those observers.
5. Colors are perceiver-dependent dispositional properties: powers to appear in distinctive ways, to normal perceivers, in contextualised standard conditions. [M. Dummett, G. Evans, J. McDowell]
6. Colors are perceiver-dependent but hybrid properties: to have a specific color is to have some intrinsic feature by virtue of which the object has the power to appear in a distinctive way (e.g., as in 4). [R. Descartes, J. Locke]
7. Colors are relational properties: they are physical properties, e.g., spectral reflectances, which are perceived in a distinctive way. [E. Thompson, M. Tye]
8. Colors are objective, perceiver-independent properties: they are the occupants of certain perceiver-dependent functional roles, e.g. roles defined in terms of the way objects look. [B. McLaughlin, J.Cohen]
9. Colors are socially or culturally constructed properties. To be red is for an object to satisfy such criteria that make it worthy of having the predicate "red" applied. [J. Van Brakel]

The Aim of Philosophical Theories of Color

To assess the rival claims about the status of color, we need to clarify what the aim of a philosophical theory of color should be. Philosophical discussions of color are sometimes framed in terms of answering such questions as "what is the nature of color?" "what is color essentially?" "what is the essence of color?". Sometimes, they are framed in terms of answering the question of what kind of understanding a person must have in order to understand color concepts or to be able to use color terms with understanding. Here we are asking a question about color concepts and it will be important to clarify first, whether the concepts in question are concepts are concepts of natural language, or technical concepts introduced for scientific or industrial purposes, and then with respect to both, whether there are different kinds of concepts.

On the face of it, there are two different exercises here: identifying the nature of colors, i.e., what colors are essentially; and specifying what the concept of color is. It seems that one exercise requires looking at the world and the other looking at the thinkers. However both exercises would seem to be an integral part of any philosophical theory of color. For there appear to be two prominent facts about colors that any theory would need to respect: (1) that colors are properties in the world (i.e., properties of physical

objects), to which one's color vision is sensitive; (2) that colors are qualities that perceptual experience represent (or presents) objects as having. At least, if any theory denies that these are facts about colors, then an extremely good explanation is called for. One theory that comes close to denying that the first is a fact is the Descartes-Locke tradition. This theory is more subtle, however, than this would suggest. It draws a distinction between two concepts or senses of color. In one sense objects do have colors but this is not the sense in which objects are represented as having colors; while in the other sense, objects are represented as having colors, but these are not properties which objects actually have.

There is a stronger reason for thinking that the two exercises might be related. This reason depends on the fact that there are different available models for thinking about concepts. For certain kinds of concepts, the understanding required in order to possess the concept of X provides the only answer to the question of what X's are essentially. (The nominal essence is the same as the real essence.) On the other hand, if externalists are right then the content of some mental states are broad. Hence, if concepts are constituents of the content then individuating these concepts will require identifying some object, property or natural kind. Accordingly, depending on which model for color concepts we adopt, the two exercises (looking at the world and looking at the thinkers) may be related.

In any case, if those in the Descartes-Locke tradition are right in drawing distinctions between different concepts of color, then there will be different answers to the question of what the essence of color is, depending on which concept we have in mind. There are, within this tradition, two important distinctions. One is between color as a property of physical bodies and color as a quality in experience, i.e a phenomenal or subjective quality: color-as-it-is-in-experience. In the second place, if we concentrate on color conceived as a property of physical bodies, we can distinguish between a naive, pre-reflective concept and a sophisticated, critical concept. For Descartes and Locke, the former concept is confused or at least has a faulty assumption built into it. Accordingly, the answer to the question of what colors are essentially, will be "none" for the pre-reflective concept and "a mind-dependent dispositional property" for the critical concept.

Much of the opposition to the Descartes-Locke account consists in challenging their characterization of the natural or folk concept. Some philosophers, Dummett, Evans, McDowell, argue that the natural concept is dispositional and does not need to be reconstructed in the way that Descartes and Locke suggest. A different form of opposition is found among the many objectivists who claim that the natural concept is objectivist. Opinion divides between those who hold that colors have hidden essences, e.g., microstructural features such as spectral reflectances, and those who hold that their essences are manifest. Most modern objectivists believe not only that colors have essences but that these essences have either been discovered by scientists, or are close to being discovered.

It is possible, however, to take a different view: that the natural concept of color is objectivist, that is, that colors are conceptualized as objective, intrinsic features of physical bodies but, so it happens, there are no such features in nature. The essences are virtual, not actual. On this account, colors are virtual properties (have virtual essences) just as phlogiston and caloric are virtual natural kinds. Given that color perceivers have experiences of color, i.e., experiences which represent objects as having colors, this account implies an ‘illusion theory’ or ‘error theory’ of color experience and perception. Objects are perceived as having

colors which they do not in fact have: there are no such colors.

However the dispute about the natural pre-reflective concept is resolved, we still need to answer the question of how we ought to think of color. If Descartes and Locke are right, then the natural concept needs to be replaced by the critical concept. However, even if they are wrong and dispositionalists such as Dummett and Evans were right about the natural concept, we might want to argue that the concept should be either replaced by, or supplemented by, a reconstructed, revised concept, in the way for example, that concepts of heat, sound, force, and solidity have been refined or revised or supplemented by scientific concepts.

Caution is necessary, however. There appear to be two reasons why there might be a need for a reconstructed concept of color. One could be that the natural concept needs to be eliminated. The other might be that it needs to be supplemented by a new, technical concept. Each reason is related to the fact that the natural concept serves a range of purposes. If it turns out that all, or nearly all, the major purposes are best served by a new revised concept then that would be a case for replacing the old concept. If only certain of the purposes are better served, then that constitutes a case for supplementing the old concept. Given that there is a natural concept of color which is employed both in color vision and in the many social practices concerned with color, then it will be important to specify what the natural concept is, in order to appreciate how if at all it needs to be revised.

It is crucial to remember, moreover, that in the case of color, almost all the purposes to do with colors and hence reasons for having concepts of color have to do with the perception of color. Colors function predominantly as natural and conventional signs, i.e., for various practical epistemological and social purposes. To the extent, therefore, that the natural concept needs to be revised or replaced by any reconstructed concept, the new concept(s) would need to be capable of serving those various purposes. Accordingly an examination of what the natural concept is will be vital for the justification of any theory about how colors should be conceived.

The Natural Concept of Color

The physicists' account of color, both in its original and in its reconstructed version, depends on a certain view of the natural concept of color. So too, in their differing ways, do the accounts of their most important critics. Getting clear about the natural concept of color will thus be essential for resolving that philosophical debate. Quite apart from that, the natural concept is important to describe for its own sake. Getting clear about the natural concept of color is necessary for a philosophical understanding of color.

Discussions of what kind of property color can often be driven by epistemological and metaphysical concerns, which can cause us to lose sight of what seem to be the plainest and most obvious truths about colors. Two of the most obvious truths have already been cited: (1) that colors are properties in the world, (i.e., properties of physical objects) to which one's color vision is sensitive; (2) that colors are qualities that perceptual experience represent (or presents) objects as having. There might be a theory according to which these are not truths at all but only ‘truths’. Such a theory cannot be ruled out, but to be acceptable it

would need to explain why the ‘truths’ have the force that they do as apparent truths.

There is another truth that is more often ignored: that colors play significant roles in the epistemological, personal and social lives of human beings. Colors are important epistemologically, as natural signs or indicators for the identification and re-identification of physical objects. From a social point of view, colors serve a variety of purposes. While retaining their function as natural signs, they also serve as conventional signs, e.g., as badges, uniforms, in ceremony, ritual etc. They may also be said to have a ‘life of their own’. As well as having emotional and aesthetic effects, colors are used in social life to amuse, to entertain, to delight, to shock, to impress, to astound, to warn, to attract, to be enjoyed, and so on, in contexts having to do with pageantry, ceremonial, courtship, painting, lighting, plays, clothing, dining, drinking, and so on.

Recognizing these obvious truths highlights the importance of the natural concept of color, for it is the way in which colors play their various roles that the natural concept is significant. In the first place, colors play their roles through the exercise of color vision, which in turn involves the exercise of color concepts. The perceiver must have perceptual experiences, or acquire perceptual states, which have a certain content. It is through such experiences that colors serve as natural signs, i.e., for the identification and re-identification of objects. Some account needs to be given of what constitutes the content of these perceptual experiences or states, i.e., of what kind of property color is presented or represented as being.

In the second place, there is a wealth of uses for color in our cultural and social life, giving rise to a flourishing color vocabulary. Examples of the sorts of practices are the use of color language, i.e., in the use of predicates such as ‘blue’, ‘red’, ‘white’ etc.; the teaching and learning of color predicates, by the use of paradigm examples; the sorting and classifying of objects; the placing of color samples in ordered structured arrays, the use of colors to impress, delight, astound, court, entertain, and so on. Central among such practices are those involving the use of colors as both natural signs or indicators, and as conventional signs.

If we concentrate on the use of color predicates such as ‘red’, ‘blue’, ‘olive’ etc., in natural language, it is possible to specify what we might call the ‘folk concept’ of color, one expressed by such terms. There is some advantage, however, in using the term ‘natural concept’ to emphasize that the folk concept is built upon the use of a biological endowment, one that is exhibited in the use of colors as natural signs, for the identification and re-identification of physical objects. Whatever it is called, it is clear that, if we wish to give an account of the epistemological, personal and social roles served by colors, then we need to give an account of the natural or folk concept of color, the concept which is embedded in the activities and practices that form the basis of such roles.

To specify the natural or folk concept of color, therefore, requires studying the variety of activities and practices, linguistic and non-linguistic, in which colors play a role. To specify this concept is a central task for any theory of colors to perform. Color is not just a topic for scientific experts. The ordinary folk are experts too. They have expertise in recognizing colors, in sorting and classifying them, in using colors and in responding to them. Color experts are not just those who study color in a scientific way, nor those who paint in colors, nor those who are industrial chemists. There are, in other words, different ranges and

levels of expertise. Those of us who are competent with colors know a lot: we know what color blue is, how it differs from red and from yellow and green; we know how dark blue differs from light blue; we use terms such as rich, pale, faded, intense, brilliant, bright, pure, mixed, and so on to convey and exploit what we know.

Recognizing the expertise of the ‘folk’ should also make us alert to the dangers of using the term "the folk concept" of color. The term ‘natural concept’ is much safer to use. For one thing, ‘folk concept’ gets easily conflated with ‘folk theory’, i.e., some doctrine that we in our naive moments would articulate if asked, and if we had the time to reflect on. There are two things wrong with this slide. In the first place, the natural concept is a concept (or set of related concepts) not a theory (though its possession may presuppose certain beliefs). In the second place, by calling it a theory, it is easy is to over-intellectualize the concept. The concept is one embedded in a vast set of conceptual practices, engaged in by color experts, those who are competent in the perception, recognition and use of colors. The knowledge is implicit as well as explicit, and it involves know-how besides.

In providing an account of the natural/folk concept of color, there are two sorts of description that we can provide: (i) a description of the way color is conceptualized, i.e., the kind of property color is conceptualized as being; (ii) a description of the kind of concept the natural/folk concept is, i.e., a description of how the concept is acquired, how it is exercised, the purposes it serves, and so on. In this respect it is possible to describe the folk concept of color as follows:

1. color concepts are perceptual concepts;
2. color terms such as ‘red’, ‘green’, ‘blue’, etc are taught by the use of paradigm examples;
3. there is a distinctive color vocabulary the terms of which are taught by paradigm examples.
4. it is through the exercise of color concepts (through the way colored objects appear), that colors fulfil their (practical) epistemological role: to serve as the signs for identification and re-identification of physical objects.
5. it is through the exercise of color concepts (through the way colored objects appear), that colors fulfil their social roles: to serve as conventional signs
6. it is through the exercise of color concepts (through the way colored objects appear), that colors fulfil their aesthetic and emotional roles.

It is through the study of the activities and practices which involve the acquisition and exercise of such concepts, that we can state certain principles about the kind of properties colors are conceptualized as being. These principles are implicit or explicit in the activities and practices. One such principle is that colors are perceptually salient, i.e are the sorts of properties that color vision is sensitive to and which are presented or represented in perceptual experience. Other important principles are those having to do with fact that colors as a group form structured arrays, with characteristic internal structures.

This last feature is perhaps the most significant feature about colors, the colors objects are represented as having. Colors are properties that as a group, form an internally-related 4+2 structure, built on the four unique, primary hues: green, red, blue and yellow, and related to the black/white pair. Colors can be placed in systematically ordered arrays, along three dimensions, e.g., hue, saturation, and brightness.

There are different arrays according to whether the colors are surface colors, film(aperture) colors, volume colors, light colors etc. Each array has a distinctive, complex character: a fourfold structure, based upon four distinctive, unique hues: green, red, blue and yellow. All colors can be mapped according to how near to or far from they are to any two of these unique, primary colors. Moreover, more than one such color system can be constructed even for the one mode of appearance: the Oswald, Munsell and Swedish Natural Color systems are examples. The last-named for example, uses dimensions of hue, chromaticness and whiteness/blackness, whereas the Munsell system uses hue, chroma and lightness. These dimensions are most suitable for surface colors, whereas hue, saturation and brightness are more appropriate for aperture colors.

In short, we can state that, given the natural/folk concept of color, colors are the following kind of properties:

1. Colors are perceptually salient:
    - they are the sorts of properties that one discerns directly by looking, that one recognizes the objects as having.
    - they are qualities that are presented (or represented) in visual experiences. Visual experiences characteristically are experiences of colors, and of shapes and other qualities of objects in a 3-dimensional space.
    - they are properties causally connected with perception and identification of specific objects as red, yellow, blue, and so on.
2. Colors are the sorts of properties that can be arranged systematically in ordered arrays. That is, a set of internal relationships hold between the range of colors.
3. Colors comprise the kind of property which make true a wide range of first order color principles or statements. Some examples are the following:
    - tomatoes are red, so are sunsets and blood, and so on.
    - red merges gradually into yellow in one direction, and into blue in another direction, but not into green except through either yellow or blue; and so on.
    - ripening bananas goes from green to yellow, certain ripening apples go from green to red, and so on.

The statements cited in 3 are meant to be illustrative examples. There is a considerable variety in the types of statement involved. Some are causal truths e.g., ripening pears and wheat go from green to yellow, acids turn blue litmus paper red, spiders with red stripes on the back are venomous, black hair tends to grow grey with age, and so on. In the second place, there are certain aesthetic and emotional facts. For many people, green is soothing. Soft pastels are suitable in certain contexts; mauves and browns are not. Light colors and dark colors have different effects on mood, and so on. Certain colors are harmonious, and others jarring. In addition, there are truths comprising the way perceived colors vary with illumination, distance, orientation, and so on. There are other truths concerning how the color of surfaces is affected by the background against which an object is seen.

What has been just given is a partial characterisation of colors, given the natural or folk concept of color. It is plausible to hold that a fuller characterisation than this minimalist version can be provided.

Unfortunately any such account is likely to be controversial. Frank Jackson (1996) describes what he calls ‘a prime intuition’ about colors that there is something peculiarly visually conspicuous about colors: "Redness is visually presented in a way that having inertial mass and being fragile, for instance, are not, When we teach the meanings of the color words, we aim to get our hearers to grasp the fact that they are words for the properties putatively presented in visual experience when things look colored." [p. 199.] This prime intuition, Jackson states, is simply that red is the property objects look to have when they look red. This deceptively simple prime intuition is said to tell us something important about the metaphysics of color when we combine it with plausible views about what is required for an experience to be the presentation of a property: a necessary condition for experience E to be the presentation of property P is that there be a causal connection in normal cases. [p. 200.]

The idea behind talk of ‘prime intuitions’ is that they are the intuitions formed by those competent in the use of color language, especially in the identification and recognition of colors. Those competent perceivers and users of color language are persons capable of reflecting on the way in which colors are presented (or represented) in color experience. Accordingly, it is plausible to hold that colors are presented as follows: as objective, perceiver-independent, intrinsic features of physical bodies, i.e., physical surfaces, volumes, light sources, illuminations, films, media, and so on.

The expression of this intuition adds little to the previous minimalist characterisation of the folk/natural concept of color. It is plausible to go further and hold that colors are not only intrinsic features of physical bodies, but are presented as manifest, sensuous properties. The way they are manifest is that their nature is open and manifest, not hidden. Some philosophers are prompted to respond to this claim by saying that it begs the question against those theorists who hold that colors have hidden essences, e.g., physical microstructures. The counter-reply to this response is that the view that colors have hidden essences is not the right theory about the natural concept or folk concept of color. It may be plausible as a theory of what a reformed concept of color is or should be, but not as a theory of what the folk concept is.

Admittedly it is not easy to convince people that there are manifest properties. One of the problems is that many concepts begin their life as concepts of manifest properties, but then evolve into more complex concepts. Children's concepts of say ‘horse’, ‘dog’, ‘man’ and so on, are cases in point. For them, the concept is almost exclusively defined by the corresponding appearances. There are, however, more sophisticated examples. Take concepts such as brilliant, dazzling, sprightly, po-faced, cheery, glum, picturesque, grim-faced, pale, etc. These are terms that characteristically apply to appearances. All of them, it is plausible to suggest, are manifest properties with essences they ‘wear on their faces’, and are not hidden.

Many properties are not manifest: being poisonous, being made by a robot, containing water as a constituent, coming from Virginia, and so on, but some clearly are. Included among these are colors. Someone who is taught color terms and who understands how they are used, knows what it is for something to be red, to be blue, or whatever: it is to have that feature which the perceiver is capable of recognizing. Reflecting on the way colors are represented, the thoughtful observer can say that the sort of property colors are represented as being are as color ‘stuffs’ spread on the surface of physical bodies (or through volumes, etc.). They are intrinsic features of physical surfaces (volumes, . . .), spread over the

surface. It seems only too clear that we experience the redness of a ripe apple as an objective quality of the apple, the redness being in an objective space just as much as are the shape, the contour, the texture of the apple. This point can be neatly illustrated by quotes from two eminent workers on the physiology/psychology of color, Hering and Boynton. Hering for example, writes:

> When we open our eyes in an illuminated room we see a manifold of spatially extended forms that are differentiated or separated from one another through differences in their colors . . . Colors are what fill in the outlines of these forms, they are the stuff out of which visual phenomena are built up; our visual world consists solely of differently formed colors; and objects, from the point of view of seeing them, that is, seen objects, are nothing other than colors of different kinds and forms. [Hering (1964), p. 1]

In similar vein, the physiological psychologist Robert Boynton writes in ‘Color in Contour and Object Perception’: "From early childhood we are easily able to recognize a property of objects, usually associated with their surfaces, that we call color. No child, and relatively few adults, will doubt that color is on (or sometimes in) objects." [Boynton (1978), p. 175] In addition, one is aware of the different character of the way colors appear in different modes, i.e., for object surfaces such as apples, patches of light on screens, volumes such as wine, scattering media such as skies, light sources such as globes, and so on.

There is one more prime intuition which is one of the most important. It is part of the folk concept, another ‘prime intuition’, that colors are represented as qualitative, sensuous features. This point will no doubt be controversial, but it ought not be. Reference to the sensuous nature of colors is crucial. These qualitative features that colors are represented as being are ‘sensuous’ in the widest sense. This is not an issue of deep metaphysics. The term "sensuous" is often used on such a way as to apply to phenomenal, i.e., to ontologically subjective qualities. However, there is a wider sense which does not have this commitment. There is a neutral use for the term. An illustration is an example which H.H.Price borrows from Husserl: ‘When I see a tomato hanging on a vine then a ripe tomato hanging on a vine is "leibhaft gegeben": it is given to me with its sensuous qualities.’ [Price (1932), p. 231.] This sound much better, of course, in German, but in English the point is that the tomato (better still, grapes) and the vine are given in perception with the sensuous features. English speakers understand that in perceiving tomatoes, grapes, etc one is acquainted with sensuous features. Price was acknowledging that whatever one's theory of perception, and especially whether one thought that the perceiver is directly aware of physical objects or sensory presentations, one was acquainted with sensuous features.

A similar point is made by Evan Thompson (1995) though with respect to the term ‘phenomenal’, rather than ‘sensuous’. Research in psychophysics and visual physiology, he writes, is constrained by the ‘phenomenal structure of color’. By this term he means to refer primarily to the three dimensions of color, known as hue, saturation and lightness, as well as to the relations that colors exhibit among themselves (p. 39). As he points out, textbooks often classify these properties of color as ‘subjective color phenomena’ or as features of ‘color experience’. Thompson prefers to use the term ‘phenomenal’ to describe them because they are first and foremost features of how colors appear: "I thus intend to use the term ‘phenomenal’ in its older sense of pertaining to appearances, not in the current sense of subjective."

The neutral notions of ‘sensuous’ and ‘phenomenal’ are ones that can be shared by writers with very different philosophical commitments. It is such a notion that Michael Tye employs, when he states that when philosophers appeal to the phenomenology of perceptual consciousness, in making claims about the phenomenal character of experience, they are mistaking intrinsic features of the content of experience for intrinsic features of experience itself.

Accordingly, we can represent writers as diverse as Price, Thompson and Tye, despite their philosophical differences, as in agreement. There is a neutral sense of ‘sensuous’, or ‘phenomenal’, according to which it is possible for physical objects to have sensuous or phenomenal properties. Most importantly, the color properties that the natural concept of color attributes to physical objects are sensuous properties. It is of course a separate question of whether physical objects do have the sensuous features that they are represented as having. Price thinks that they do not, but he also thinks that a further argument is required to show that they are not.

To conclude: given the characterisation of the natural concept of color, color is a certain kind of property. Which kind it is can be specified, in part, by saying that it is an objective, perceiver-independent, manifest and sensuous kind. In addition the property is one with certain kinds of causal powers vis a vis the presentation of color in the perception, recognition and identification of colors. Finally, colors are the kinds of properties that fit together in characteristic ways to form structured color arrays, with a distinctive 3-dimensional character. They are properties that as a group, form an internally related 4+2 structure, built on the four unique, primary hues: green, red, blue and yellow, and related to the black/white pair. Some parts of this characterisation of the natural concept are contentious, e.g., the claims that colors are manifest and sensuous. Some of the most significant parts of the characterisation which have the most far-reaching implications are not controversial: that colors have causal powers as described above, and that collectively form astructured system.

Science of Color or Color Science?

There is a flourishing field of color science, one that goes back to Newton and includes such famous figures as Young, Maxwell, Helmholtz, and Hering. Almost all of this field has to do with the perception of color, that is, to studying the conditions that cause or contribute to colors being seen, or to ways in which the colors as they appear may be studied. An example of the second kind of research is the many studies in colorimetry, involving the specification of colors. Typically these involve matching techniques in which the subject is asked to match some stimulus with one or another of a range of standardized cases.

One of the most vigorous areas of research, especially more recently, is the study of color vision, i.e., of the mechanisms involved in the perception of color. Helmholtz and Hering were pioneers in the physiology of this area, but much has been done recently in research on the neural processes involved in color perception. A crucial development has been the growth in opponent-process theory. [See Kaiser and Boynton (1996) for technical discussion. See Hardin (1988) and Thompson (1995) for philosophically-informed discussions.] Also of significance are the experiments by Land and his colleagues and the

development of his retinex theory of color vision. [See Land (1983) and also Hardin (1988) and Thompson (1995) for critical discussion.] Uncovering the mechanisms that underlie color vision is an exciting current field of research. The major philosophical relevance of such research is that it promises to help explain why some of the appearances of color have the character that they do, e.g., why there are no reddish-greens nor bluish-yellows. If it becomes clear that appearances have a certain character which no set of objective physical features have, and that character can be found to based on the physiological/neural processes, then the research may be crucial in establishing that color is best thought of not as some objective feature of the world that color vision detects, but rather as something constructed by one's color vision.

Another area of color science has to deal with the construction of color systems, i.e., of ways of ordering the range of colors in a systematic fashion. Usually this is done by constructing three dimensional color solids. It is interesting, however that there are different systems that have been constructed. For one thing different dimensions are used depending on the way in which color appears. Colors as properties of surfaces, in general, have a different mode of appearance from colors as properties of volumes such as wine, and yet again from that for film color or aperture color. These different modes of appearance suit different dimensions of color. For surfaces the dimensions (at least in some systems) are hue, chromaticness and whiteness/blackness; for aperture or film colors the dimensions are hue, saturation and brightness.

Yet another field in which the way colors appear is crucial is the field of color psychology: the field in which color-constancy, simultaneous contrast, the effects of various backgrounds on color perceptions, and so on, are examined, and competing explanations debated.

Almost all of this research in color science is devoted to the way color appears, i.e., to the conditions under which one perceives color or experiences it or to the character of the way color appears. Almost none of it is concerned with the other color ‘truths’, that is, to what we might call ‘causal truths and principles’. This is not to say that those ‘truths’ or principles are false or are invalid. Biology and chemistry and indeed physics all use color concepts and claims in their theories (some of them) and explanations, but there has not developed what might be called a "science of color", except for the study of the way color appears, or one might say, of the way colors are represented, and of the causes and conditions conducive to the way they appear and are recognized. That there is a flourishing field of color science but not a science of color reflects the special place color has in science.

There is an important difference between color science, on the one hand, and the science of shapes, geometry, the science of heat and temperature, and the science of sound, on the other hand. In the case of shapes and heat/temperature, and sounds and weights, we have properties of physical objects which we can detect, naturally and unreflectingly, by the use of our senses. These properties, however, are different from colors. In the case of shape and heat and weight and sound, there has developed a science in which the principles of sound, weight, heat, and shape are studied. There is however no parallel science of color. There are few color principles to serve as the basis for a science of color. Color science is a large field, but it is built around the way that colors appear and to the conditions under which colors can be perceived, and the causes which lead to the perception of colors. If colors ceased to appear in the

distinctive ways then color science would disappear.

The field of color science has developed through building up theories and color facts which contribute to our understanding of the perception of color, as well as to provide an objective specification of color. With both aims, the scientific account has to take account of a range of color facts that hold of the practices and behaviour of color-perceivers. That is, before we discovered any detailed scientific knowledge about color, we had - and still have - a considerable body of knowledge about color.

This body of color knowledge is contained within the conceptual practices specific to color. By studying these practices, we can draw up a set of general color principles and 'truths'. The range and extent of the general principles, have been emphasized by Justin Broackes (1992). [For further discussion, see Maund (1995).] These general facts or principles include causal truths, although the nature of the causal powers may be difficult to discern. It is easy to see that color science has both filled out the details of some of the color principles described previously, e.g., in respect to the internal relations and to the conditions under which color is perceived, and in some case modified them. Furthermore, the discovery of color-mixing laws and of the mechanisms underlying color vision has added to our knowledge of color.

A Unifying Framework for Colors

There is a set of color truths and color principles that can be said to comprise our color-knowledge. It is to this body of knowledge that color science contributes. It is possible to find some unifying order that brings together in a unitary scheme these various color truths and principles. The set of truths and principles can be divided into separate categories, which bring out the different roles played by colors and color-concepts in (i) the acquisition of color terms in language; (ii) the appearances of colors (iii) the development of first-order truths or 'truths' about colors and (iv) servicing social and epistemological purposes.

L: Principles about the Use of Color Terms:

1. Color terms are taught and learned by the use of paradigms. This is to say that the paradigms are identified and recognized by the way they look (appear).
2. Colors are properties of bodies that play a causal role in the learning of color terms and the communication about colors.
3. Cross-cultural comparisons indicate that there are certain basic color terms which are systematically related. [See Berlin and Kay (1969), and Boynton and Olson (1990), and for a contrary view, Van Brakel (1993)]

A: Principles about Appearances and the Perception of Color:

1. Specific colors have distinctive appearances, characteristic of each color.
2. The way colors are identified and recognized is by the way they appear to perceivers. There are no color thermometers or other measuring devices.

3. Colors take a different mode of appearance, i.e., have a different characteristic appearance, when they are features of physical surfaces, films, volumes, light sources, etc.
4. There are principles governing the conditions under which colors are perceived. Certain conditions are better than others for identifying colors; certain people are better than others at identifying colors. Colored bodies can appear differently when viewed at different distances, in different illuminations, and against different backgrounds.
5. Among the principles in A4 are principles governing constancy effects: tendencies for objects to look the same under different conditions.
6. There is a certain distinctive form to the way colors appear. Visual experiences represent colors in a certain way, as qualitative features which are "sensuous" in the widest sense.

T: Color Truths of the First Order:

1. There is a vast range of specific color truths: ripe bananas are yellow; certain sunsets are golden; claret wine is claret red and so on.
2. Colors can be combined together in structured, systematically ordered arrays, with a distinctive character. They are qualitative features which are "sensuous" in the widest sense. These arrays are different depending on whether the colors are colors of surfaces, volumes, films, scattering media, lights and so on.
3. There are general causal truths e.g., ripening pears, bananas and wheat go from green to yellow, acids turn blue litmus paper red, spiders with red stripes on the back are venomous, black hair tends to grow grey with age, and so on.
4. Different colors have different specific aesthetic effects, including principles of harmony, balance, contrast, etc.
5. Different colors have different emotional effects.

R: Principles about Colors and Their Roles:

1. Colors are natural signs, i.e., are easily identifiable features of objects that enable perceivers to identify and re-identify kinds of objects.
2. Colors serve as conventional signs, for similar purposes as those in R1.
3. Colors, often because of effects in T4 and T5, serve certain social and psychological roles.

It needs to be emphasized that these categories are not meant to be exclusive. For example, principle R3, concerning the social and psychological role of colors is related to principles T4 and T5, concerning the specific aesthetic and emotional roles of colors. Likewise the principles in A6 and T2 both refer to the sensuous nature of colors.

The Illusion Theory of Colors

The natural or folk concept of color conceptualizes color as a certain kind of property: an objective, perceiver-independent, intrinsic feature of physical bodies. The property is, moreover, one with a certain character and with certain causal properties. The most straight-forward view of colors is that they are what they seem. Colors are properties whose essences are manifest. Since colors here are taken to be simple, primary quality of objects, we might call this view "The Simple Objectivist View of Color". It stands in clear contrast to those forms of Objectivism according to which the essences of colors are hidden, not manifest. [See OBJ].

Opposed to both forms of Objectivism is another view: ‘The Illusion Theory of Colors’. In this account, it is held that once we spell out the character of the features specified by the folk or natural concept, we discover that there is in nature no such features: colors as they are conceptualized are properties not found in nature. The colors objects are represented as having, in visual experience, are ones that no object actually has.

The natural concept of color conceptualizes color as a certain kind of property. Which kind it is can be specified, in part, by saying that it is an objective, perceiver-independent, manifest and sensuous kind. In addition the property is one with certain kinds of causal powers vis a vis the presentation of color in the perception, recognition and identification of colors. Finally, colors are the kinds of properties that fit together in characteristic ways to form structured color arrays, with a distinctive 3-dimensional character. They are properties that as a group, form an internally related 4+2 structure, built on the four unique, primary hues: green, red, blue and yellow, and related to the black/white pair. Some parts of this characterisation of the natural concept are contentious, e.g., the claims that colors are manifest and sensuous. Some of the most significant parts of the characterisation which have the most far-reaching implications are not controversial: that colors have causal powers as described above, and that collectively form a structured system.

Defenders of the Illusion Theory of Color exploit the presence of these features in the natural concept to argue that, given this concept of color, there are in fact no colors in nature, that objects are presented in experience as having colors which neither they nor any object have. Crucially, there are no properties that both have the causal powers in question and which collectively have the right character. In short, there are no colors that are intrinsic, non-relational, perceiver-independent properties and which satisfy the requirements of the three-dimensional color solid. None that is, that allow us to make sense of the way in which we perceive and identify and recognize colors.

No properties of physical objects stand, it has been said, in the right kinds or relations that are characteristic of the structured color arrays. It is true that we can arrange physical samples in ordered arrays but the ordering principles depend on the way they appear. What is crucial to the principle of ordering is the way the colors are represented as being, or rather, the character of the way colors are represented. It is because there is a distinctive appearance associated with each color that the colors are capable of being systematically ordered in the way that they are.

It would seem that, as far as our conceptual practices governing color are concerned, physical objects do not have the kinds of color they are represented as having. The colors that objects are represented as having are illusory: no physical object actually has those colors. The colors might be said to be "virtual properties": they are properties objects do not have, but might have had: in some other possible world but not in this one. If we speak of colors as having essences, then they have virtual essences. Colors are virtual properties, just as phlogiston and caloric are virtual natural kinds.

The illusion theory, or virtual essence theory, of colors leaves us with a problem. If there are no properties that satisfy the requirements for being colors: how did the natural concept develop? The solution to this problem is found in the fact that the way that the concepts of color operate, to serve their various functions and roles, is through the way colors appear. For these purposes and roles, objects do not need the actual colors. It will be sufficient if they appear to have colors. For these purposes, it is sufficient that "it is as if they have the colors".

There are two major functions for color concepts. One reflects an epistemological purpose: colors are signs used to indicate the presence of objects of interest. The signs are either natural or conventional, the latter being ones designed for various social purposes. The purposes are equally well served even if objects do not have colors, but have the right appearances. All that is needed is that they are represented as having colors. The second major purpose of color concepts is aesthetic, understood in the widest sense. Color is significant in painting, decorating, clothing, theatre, make-up, advertising, showing off, sexual appeal and so on. Again, it matters not in the least that objects do not have these properties. All that is required is that they be represented as having them.

The significance of appearances is widespread. As we have seen, they provide the basis for the ordering principles governing color systems. Likewise, the causal truths and principles that employ color terms are ones connected with the way colors appear. For example, we can explain how there are such ‘truths’ as "ripening pears (bananas, . . .) go from green to yellow". This is a truth. For a pear to be represented as green under the right conditions is a sign that the pear is not yet ripe; for it to be represented as yellow is a sign that it is ripe. In other words, whatever causal truths we have concerning color, can be explained by interpreting colors as signs or as indicators for other physical features, where those physical features serve the causal roles.

In a previous section a distinction was made between color science and the science of color. While the former field is flourishing there is little science of color. One way of understanding how this situation has arisen is that there are no actual colors in physical reality. What there are are experiences which represent objects as having colors, colors which in fact they do not have. That is, colors are virtual properties. Our visual experiences present us with systematic illusions. If this were the case, we would still have the same color science, exactly as we have now, for we would still need to know how colors are represented, and what causes them to be represented in the way that they are, and how the various conditions under which we have color experiences systematically differ can be explained. Since one of the central roles colors serve is to act as signs or indicators for physical objects, and any theory of color has to acknowledge this role anyway, it would seem that any fledgling science of color is best dispensed with in terms of other

sciences, and color science left to the science of color perception. This does not stop it from being the case that there is an important theory of color vision and perception, and of the role colors play as signs or indicators.

Clearly, the concept of color can be used to serve many of its normal purposes even if the representations of color are illusory, provided that the illusions are systematic, which on a proper theory, of course, they will be.

Colors as Simple Intrinsic Objective Qualities

The Simple Objectivist View of Color is that there are in nature colors of the kind specified by the natural concept. Colors are simple intrinsic, non-relational, non-reducible properties, supervenient on micro-physical features. Such a view has been presented by P.M.S. Hacker (1987) and by J. Campbell (1994).

The main problem the illusion theorist finds with the Simple Objectivist View is with reconciling the putative character of the intrinsic color features that fit together to comprise color solids with the distinctive structure, with the causal role of such features in the recognition and identification of colors. The problem is addressed by Hacker in his defense of the claim that colors are intrinsic features of physical bodies. He forthrightly rejects not only the physicists' view, and Reid's view on colors, but also the dispositionalist account offered by McGinn, McDowell and Dummett. He insists that colors are properties which are used to provide causal explanations. There is no more reason to deny this, he says, than there is to deny the parallel claim for solidity and liquidity. In particular, he claims that we can provide causal explanations for why colors affect color perceivers. The explanation is not vitiated by the discovery that microstructural processes are involved, any more than explanations concerning solidity and liquidity are rendered otiose by the discovery of the microstructural base for these properties.

It is doubtful that this maneuvre works, for a number of reasons. One is that we would need to specify the criteria that make it the case that an object is intrinsically red. Not all perceivers agree in their judgements. It is not that there are color blind people who can, after all, be said to be color-deficient. There is a small but still significant number of color-anomalous people, who can make all the same color discriminations as regular people, but who disagree about which samples are pure red, green, etc. That is, it seems that their color solid is skewed from the normal. It seems arbitrary that we decide that the real color is the one that the majority pick. Secondly, if there were an evolutionary shift, or an eugenics program, the minority could become the majority.

There is a more important reason against Hacker's proposal, however, which depends on the fact that for colors, microstructural explanations cannot be provided for all the relevant, important features. Specifically the complex internal relationships between the colors cannot be explained by the microstructural properties of physical bodies, except through their affecting the perceivers. That is, to explain why the colors have the relationships they do requires giving an account of the structure of the perceiver's perceptual apparatus. At a minimum, this requires an account of the response curves of cells in the retinae, but also required would be an account of the appropriate neural processes. In short, the

explanation will have to work via an explanation of how things appear, that is, of how one's perceptual experiences have the content that they do.

There is a difference between solidity/liquidity and the colors. In the case of solidity and its sister concepts, there is a range of features that are associated with them, including causal relationships. If we have been given adequate scientific explanations at the microstructural level for solidity, then adequate microstructural explanations will need to be given for these other features. The reason why it is important to preserve the concepts of solidity and liquidity is that such concepts unify sets of properties that are useful to have unified, and this unification is lost if we retreat to the microstructural level.

The important difference for colors is that there are crucial features of colors that are not reproduced at the microstructural level of the physical objects, nor are they explained at that level. The features are those that colors have, by virtue of which they are capable of forming systems of properties with internal relationships. This structural property is not explained at the microstructural level of physical samples of colors. To try to explain the structure physically, the best we could hope to do is to try to explain it in terms of dispositions, e.g., to induce a certain ratio of light sensitive retinal cells. Even if that were to work, it is the wrong kind of explanation to help Hacker. He wants to hold that colors are intrinsic qualities of physical objects, not relational, dispositional ones.

Objectivism

The Simple Objectivist view is that the objective essences of colors are manifest. The more common form of objectivism, however, is that colors have hidden objective essences. Advocates of this view have differed, though, on whether the hidden essence of color is a microstructural feature, intrinsic to the colored body or whether it is a light-related dispositional property, e.g., one connected with reflectance profiles. Thomas Reid was probably the earliest example of the first type and Frank Jackson (1996) the latest. Recent examples of the second type are D.M. Armstrong (1969), J. Westphal (1987) and D. Hilbert (1987).

The attempt to locate the essence of color among the microstructural features of colored bodies seems unpromising. One of the major problems is ‘the problem of multiple realizations’. Given the range of bodies that have colors _ surfaces, volumes, light-sources, illuminations, luminescent bodies, films, expanses _ the intrinsic physical features that provide the causes for the way colors appear show a bewildering variety. Even if we concentrate on the first type of color, surface color, : we find that there is a wide variety of underlying physical microstructures, responsible for objects' appearing blue, yellow, etc. The causes of the colors objects appear to have are many and varied. The same type of micro-structure consistently appears the same color (within limits) under different conditions, but different microstructures may appear the same.

It would seem, therefore, that the most plausible candidates for objective essences are light-related dispositional properties, e.g., capacities to emit, reflect, absorb, transmit or scatter light to varying degrees. However, the problem of multiple realizations has merely been postponed for, depending upon

the type of object in question, a different candidate for the objective essence has to be found. For physical surfaces, it would have to be related to the object's reflectance curve, e.g., the capacity to differentially reflect wavelengths from different regions of the incident illumination; in the case of volumes, it would be related to the object's transmittance; in the case of such objects as the sky, to the scatterance; in the case of aperture color or film color, it would be related to the pattern of light received at a particular place or at the source of the light (the reflecting source in the case of physical surfaces) and so on. Even in the case of ordinary objects, the colors may be caused in a variety of ways. The blue of a bird's coat may result from scatterance, the red in a different way.

Nevertheless, progress can be made if we concentrate on one type of color, surface color. The most plausible attempt is to try to identify surface color with a light-modifying disposition, e.g., with a disposition to reflect (or absorb) certain proportions of standardized illumination, or, if one prefers, certain proportions of light of the wavelengths from the visible spectrum. Objects with neutral or achromatic colors are ones which reflect all wavelengths to (roughly) the same degree, with whites reflecting a higher percentage than greys and blacks. Objects with chromatic colors are those which differentially reflect or absorb light at different wavelengths. Accordingly surface color would be identified with some feature of an object's spectral reflectance curve.

A special case of the problem of multiple realizations is posed by the occurrence of metamers. In the case of physical surfaces there are metamers, i.e objects with very different reflectance curves that have identical appearances of color. The situation is far more pronounced in the case of film colors or aperture colors. Here there are innumerable different combinations of light that will give the same hue. It would seem that the property shared by physical objects with the same film color is a disposition to incite the three light-sensitive cones in the retinae according to the same ratio: x: y: z.

The major problem, however, with any of the objectivist accounts of surface color is that, given the way the natural or folk concept has been characterized, it is hard to see how there could be an objective essence with the right characteristics. For, given the natural or folk concept of color, it seems that color is a certain kind of property: a perceiver-independent, intrinsic property of objects, one that satisfies certain constraints. But it is hard to see how there could be any intrinsic, physical features that satisfy all the constraints. The most crucial requirement is that colors, as a group, have to stand together in the right kind of relationships. There are no manifest, intrinsic features that satisfy this requirement. Nor does any set of microstructures stand the remotest chance of satisfying the appropriate constraints. In particular, there are no physical features, either of microstructure, or of the object's contribution to light, such that the right kind of internal relationships hold. [See Hardin (1988) and Maund (1995).] Neither can color be a dispositional property, say spectral reflectance, or a disposition to produce physiological responses of a certain kind, since spectral reflectances don't fit together in the right kind of ways.

There are two kinds of response an objectivist might make. One would be to deny the account of the natural concept of color as expounded here. Instead, it is held, it is part of the way color terms operate in the language, that it is understood that colors may well turn out to be hidden essences. The other response involves not challenging the account of the natural concept but insisting instead that it needs to be revised or reconstructed for, say, scientific purposes.

One proposal that illustrates the first type of response is the ‘functionalist’ proposal: red is the property that disposes its bearers to look red [Cohen (2001), McLaughlin (2001)] This proposal conforms to a more general approach: to hold that that the disposition to appear is not part of what colors are essentially, but it is part of the (folk) conceptualization of a certain perceiver-independent property, which is color. The folk, it is held, use the way colors appear in order to characterize a certain property, but the way objects appear is essential to the characterisation, not the property. Here we are depending on a distinction between a property and the mode of presentation (or Fregean sense) through which the property is presented (or is thought about). Accordingly, appearances would be tied to the concept of color without being part of the property of color. It would then be open to us to identify the property through some scientific means.

The second type of response is a revisionary proposal. With this response the objectivist can concede that there are no colors in the way ordinarily conceived but hold that, nevertheless, there are colors as conceived in another manner, i.e., in an objectivist manner. After all, there are not atoms as Dalton conceived them, nor oxygen as Lavoisier conceived it, nor planets as pre-Aristotelians conceived them, but still there are atoms, oxygen and planets, all the same. In other words, the objectivist can propose a revisionary concept of color as an objective property, e.g some microstructural property or a spectral reflectance or some other light-modifying feature.

The assessment of this proposal will depend on the nature of the revision recommended. It is one thing to propose the introduction of a new concept; it is another thing to propose it as a replacement for the existing concept in the spirit of eliminating it and other competing revisions. There is no reason in principle why we should not introduce a new, different concept of color, ‘physical color’ which, we may assume, takes over the causal role specified in our characterisation of color. Such a move is legitimate, but it leaves open the possibility that there is still a need for another concept, for the causal requirement was only one of the requirements for the original concept. If the new physical concept cannot service other legitimate requirements, then we need another concept to serve these purposes. One possibility is that two new concepts should emerge. As Ian Hacking has pointed out, it is plausible that the original concept of "acid" later split into two new concepts, each perfectly legitimate. Another example is the replacement of Newtonian mass by the two new concepts of mass in relativity theory. Where previously it had been assumed that there was a single essence, it is now the case that there are two essences.

Assessing the merits of any revisionary proposal will depend on examining the reasons for modifying the original concept, and on whether there is any available rival. Before considering the objectivist's revisionary proposal, let us consider the other response the objectivist can make: to reject the account offered above of the folk or natural concept. There are two forms this response can take, a simple form and a more sophisticated one. In the simple version, the way colors appear is used as a criterion for detecting the presence of the hidden essence, but it is not essential to what it is for something to be colored. After all, gold is acknowledged as having both a hidden essence and an appearance: the real essence of gold, its atomic number, plays a causal role in producing a certain golden appearance, one that is used by language-users to identify, loosely, pieces of gold. In like manner, it is argued, colors have hidden essences which play a causal role in their having the appearances they do.

As an account of how color terms operate, this view is implausible. It has the consequence that the appearance is not essential to color; that if objects were to cease to have their distinctive appearances, while retaining their reflectance-profiles, then our color vocabulary would largely remain untouched. Our summary of the color-principles, however, revealed that all of these principles either directly involve color appearances or it was the case that the way they worked was through color appearances. Without appearances, colors would not be of any interest whatever. Just as wines would cease to have interest, even to dedicated wine-growers, were they to lose their distinctive tastes, so too would colors were they to lose their appearances. There are two possibilities. One is that through genetic change, humans became incapable of seeing objects except in terms of shades of grey. So no object has the distinctive color appearances. The second situation is the same as the first, except that 40 years later, technologists have devised spectacles (or implants) that allow people once again to see objects as colored. However they cannot match appearances with reflectances. Tomatoes have become blue, skies appear red except at sunsets when they are appear blue (or sometimes greenish-blue). It seems implausible that in such circumstances color vocabulary would go with the light-dependent properties rather than with the appearance dispositional properties.

It needs to be remembered that the situation as far as colors are concerned is very different from that for gold or aluminum. As Putnam points out, though most of us identify gold through its appearance, the appearance does not constitute (part of) the essence for gold. The reason we do so is that the appearance is a trivial criterion. It is easy to imagine circumstances in which gold would lose its lustre and in any case we distinguish fools' gold from real gold. The situation is very different for colors. Unlike gold the appearance associated with colors is crucial. The important point, as far as colors are concerned, is that colored objects have characteristic appearances and that those appearances are of great interest to us. It is because we have that interest that we need a concept of dispositional color - the power to appear in characteristic ways. It is because of the way colors appear that they are important to us both biologically and socially. It is because colored bodies appear that way (i.e., the way they do) that colors perform their various functions. To tie fool's yellow with appearance and real yellow wish some microstructural property seems absurd. Of course, if we distinguish between say physical color and psychological color, then we could imagine circumstances in which two objects which had the same psychological yellow color were ones in which one object had real physical yellow, and the other fool's physical yellow.

The Complexity of Scientific Identifications

Objectivists who hold that colors have hidden objective essences customarily draw upon standard examples within science, where microstructural properties are called upon to explain why objects have certain ‘surface’ properties, e.g., solidity, solubility, temperature, elasticity, fitness, refractive power, and so on. Such examples are meant to provide models for thinking about colors, in that they offer cases in which the surface property is reduced to, or identified with, a certain microstructural property, e.g., where light is taken to be a form of electromagnetic radiation, temperature of a gas mean kinetic energy of the constituent molecules, and so on.

The logic of scientific identifications or reductions, however, is not as clear as it might be. The concept of temperature, for example, had a distinguished scientific status long before the atomic/molecular theory emerged. If temperature originally was conceptualized as ‘that property which occupied such and such a causal role’ then the way was left open to identify temperature with mean kinetic energy. It is not at all clear, however, that this was the original scientific concept of temperature. A different way of viewing the situation is to take it that the scientific reduction (or identification) of temperature to (with) kinetic energy worked as a two stage process. It first involves a replacement of the original concept by a reconstructed one, followed by a second stage in which the reconstructed concept allows for the identification of temperature with the appropriate property described within statistical mechanics.

However not every case in science in which we can explain surface properties in terms of microstructural processes and properties are ones that involve identification or reduction. Take the example of solidity. It would seem that we have discovered what the microstructural property is that provides the causal basis for solidity. It is not at all clear, however, that solidity has been reduced to or identified with, that microstructure. It all depends on what exactly the concept of solidity is and there are different models for thinking about this concept. Not all of them lend themselves to identification and reduction, and those that do not have strong claims for legitimacy.

To appreciate such models, let us consider what reduction or identification in the case of solidity requires. To explain how solidity has been reduced, we need to specify what the original concept of solidity was. There is, however, more than one candidate. Each candidate is associated with certain causal powers: relative impenetrability, stability of a certain kind, capacities to resist, and so on. There are however, different ways in which solidity might be related to these causal capacities:

1. solidity = the causal capacities and powers to . . .
2. solidity = that microstructural basis whatever it is, that is the causal basis for the causal capacities and powers, as in 1.
3. solidity = the property of having some intrinsic structure whereby the object has the capacities as in 1.

We need to distinguish account 2 from a cousin:

> 2*. solidity = that microstructural basis whatever it is, which, as it happens, is the causal basis for the causal capacities and powers, as in 1.

The difference between Models 2 and 2* rests on how the relevant causal powers are related to the microstructure. The term ‘solidity’ is understood on either model as functioning as a name for the property but they are different types of name. In model 2 the characterisation in terms of the causal powers is essential to the understanding of the name, whereas in Model 2* it is not. In the latter case, the causal capacities are used to refer to the microstructure, but any of a number of other characterizations might have served. The causal capacities might not even hold of that particular microstructure, and yet it could refer all the same. Model 2* seems to fit names such as ‘gold’ and ‘water’ -- at least as far as capacities such as appearances and tastes are concerned. Model 2 would seem to fit terms such as

'electron', 'proton', 'gravity wave', 'force', etc., and it seems more appropriate than the other for 'solidity' and 'liquidity'.

On model 2, solidity can be identified with some microstructural property. Reference to the relevant causal capacities is essential to the conceptualization of solidity but the causal capacities are not essential to solidity itself. On the other hand, with respect to Models 1 and 3, the causal powers are essential to solidity. Given that solidity is conceptualized in these ways the causal powers are essential to the property of solidity and not just to the conceptualization.

Whichever model we adopt, we can agree that solidity has been explained by reference to certain microstructures. Only on Model 2 (or 2*) has the property been reduced to, or identified with, any microstructural property. On the other models it has not. Moreover there seems no compelling reason to favour Model 2. Scientific practice does not point in favour of it, and even if it did, it would do so only after philosophical work has been done. The fact that some modern scientists, or even most of them, say things such as "we have learned that solidity is XYZ" is not decisive. Such sayings can be taken as elliptical for statements such as "we have learned that the explanatory basis (or causal grounds) for solidity is XYZ". If it turns out that there is no single microstructural basis that is the causal basis for the causal powers, then this eventuality is well handled on model 3. This model requires that there is some basis for the causal powers, not that there is a unique basis. Model 2 on the other hand, requires a unique property. The only possible way to handle the eventuality of multiple realizations, given this account, is to say that there are different kinds of solidity each with its own unique basis. Presumably, each kind of solidity is possessed by one of a limited range of objects. In the case of temperature, it would seem that the physical basis must be different for gases, as against liquids, solids, and sub-atomic processes.

Multiple realizability of states such as temperature and solidity and potentially color, might be handled by a modification of model 2. It could be that in our reconstruction, we relativize the concept so that temperature is relativized to a range of objects, being that state which, for that range of objects, plays the distinctive causal role for that property. It is plausible to say, for example, that any eye is an eye for an organism of a certain kind. In each type of organism the eye plays a similar causal role, but it is realized by different structures in each type of organism. On this way of thinking, an eye for a spider is one thing, an eye for a human another, but what makes them both eyes is the kind of causal role they play (specified in a formal, abstract way).

The lesson the example of solidity teaches us this. First, we only have reduction or identification if the original concept is of a certain type. If it is not, if it is, for example, a pure dispositional property, then we do not have reduction or identification. In such a case, we might treat the original concept as replaceable by a revised concept. The justification would be, for example, that not only would nothing important be lost by the change, but scientific practice, say, would be enhanced. It would then be possible to have replacement plus identification (or reduction). However, that would require taking the new concept to be of a certain type, e.g., as in Model 2 or 2*, rather than of a mixed type, as in Model 3. We should not assume that in general the first type is superior to the mixed type, and should expect that it sometimes is not.

Second, if it turns out that there are multiple realizations for solubility or temperature or whatever, the only kind of identification is a relativized one. Such a relativized identification is admissible, but it requires some uniting principle. In the case of solidity, the uniting principle seems to be that of having certain causal capacities and powers. What makes the relevant microstructural property count, in the proper context, as solidity, is that it occupies a certain causal role. In the case of colors such a relativized account, one relativized to observers and the way they appear, would seem to be the most appropriate account.

Objectivist And Subjectivist Accounts

Major opposition to the objectivists come from those philosophers who hold that colors are essentially mind-dependent dispositional properties: powers to appear in distinctive kinds of ways to perceivers of the right kind. Such accounts are often called ‘subjectivist’. In comparing the merits of objectivist and subjectivist accounts, it is helpful to study the examples of solidity and liquidity, for these examples provide a set of parallels for thinking about color.

In the case of solidity and liquidity, there is a range of causal capacities that historically have been thought to be constitutive of these properties. The growth of science has seen the discovery of the microstructural properties that form the causal ground of these capacities and powers (at least in broad terms). This discovery does not mean that such microstructures constitute the essences of solidity and liquidity. There are at least two ways of thinking about what solidity and liquidity are essentially. On one model, solidity is essentially the microstructural property and the description of the causal powers forms an essential part of the way the property is characterized, rather than an essential part of the property itself. There is a second way of thinking about solidity, according to which solidity is not identified withe microstructural basis even if the latter is unique. Rather the causal powers are essential to solidity, either because solidity is identified with them, or because solidity is taken to be a second order property: for something to be solid is for it to have some property which is the basis for the relevant causal powers.

As far as color is concerned, it would seem that the objectivist would need to depend on either of the last two models. For most objectivists take colors to be essentially dispositional properties, ones characterized in terms of reflectance profiles. An object's reflectance curve represents a dispositional property: a power to differentially absorb or reflect light from the range of wavelengths constituting daylight (or a standardized equivalent).

It would seem, therefore, that as far as revisionary accounts of color are concerned, the choice is between different kinds of dispositionalist accounts: objectivist and perceiver-dependent (subjectivist). On both analyses, what colors are essentially is given by a description of appropriate causal powers. In one case these causal powers are objective ones; in the other, they are special perceiver-dependent causal powers. Let us call the respective kinds of color ‘objective color’ and ‘psychological color’. In assessing these accounts as providing revisionary proposals, the important question to ask is what can be achieved by adopting the respective proposals. There is reason to think that psychological color is superior, or at least as good.

The psychological, i.e., mind-dependent property is usually presented as being a pure disposition: to call something ‘red’ is to say that it has the power to appear red to observers of the appropriate kind. A far better proposal, which can be found in the writings of Descartes and Locke, is one that presents color in terms of a mixed disposition:

> *x* is red = *x* has some feature by virtue of which x appears red, . . .

This concept has all the advantages of the objectivist concept, and added virtues of its own. It allows for multiple realizations of the disposition, and hence of the color. It does not require that for each color there is a unique physical basis. Second, by placing emphasis on appearances, it provides the means to unite the various kinds of color: surface color, volume-color, aperture color, illumination-color, etc. And finally, it can perform the one function that the physical concept does very well: it shows how colors can have a causal role in relation to the perception of color, and the social roles played by colors.

That the mixed dispositionalist account readily solves the problem of multiple realizations is obvious. It is one of the central advantages of the psychological account, however, that it both provides a connecting link between all the various kinds of color: surface color, volume-color, aperture color, illumination-color, etc., in that it unites them all as colors while at the same time it makes intelligible their differences. For each of these colors, there is a distinct mode of appearance. For each mode of appearance, colors can be organized into systematic 3-dimension color arrays. And for each array, hue is one of the dimensions: colors can be ordered with respect to how close they are to red, green, yellow and blue. There are, however, important differences. For aperture colors, the other dimensions are saturation and lightness. for surface colors, they are chroma and value, or chromaticness and blackness/whiteness. As well, the greens, blues, yellows, etc of surfaces are different types of green, blue, yellow, etc from those of films _ but they are greens, blues, yellows, etc, for all that. It is through characterising surfaces, films, illumination-sources, and so on, as providing appearances that sense can readily be made of the range of similarities and differences between the various kinds of color. The point of having a dispositional concept framed in terms of the way things appear is that it helps provide principles of unity and diversity for the available range of color systems.

Finally the mixed dispositional concept can perform the same function that the objectivist concept can serve: it can be used to show how colors are causally relevant to the perception of color. The mixed dispositional concept retains the emphasis on red objects having the right kind of power, but it allows that the object, in having that power, has some physical feature (which may be different in different objects) which is the basis for that power. Clearly, on this analysis, the underlying physical feature has all the causal powers one could wish. The mixed dispositional analysis combines this with the advantage of keeping colors tied to the way they appear.

In addition, there is reason to think that colored objects appear is an essential part of the range of conceptual practices. The point here is that while reflectances are causally relevant, as for that matter are microstructures, so too are appearances. In the case of color there is a deeply entrenched set of activities and practices central to which is the operation of causal powers to appear, i.e., powers to cause perception

of objects as red, blue, etc. These causal powers are also central to the field of color science. Given this twofold fact, then if we are considering a revisionary concept of color, then, by analogy with solidity, there is good reason to propose a dispositionalist concept of color, for which the power to appear in a way distinctive for individual colors is essential.

The important point, as far as colors are concerned, is that colored objects have characteristic appearances and that those appearances are of great interest to us. It is because we have that interest that there is point to having a concept of dispositional color - the power to appear in characteristic ways. It is because of the way colors appear that they are important to us both biologically and socially. It is because colors have a characteristic appearance that: the colors can be ordered systematically in color arrays; they have emotional effects; principles of harmony and contrast apply; there are principles governing phenomena of color contrast. It is true that physical features both of physical objects and of retinal cells contribute causally to these phenomena, but central to all of these color principles is the way color appears.

At this stage, an objectivist might argue that there is a more fundamental causal power, one associated with reflectance curves, and for this reason it would be preferable to adopt, as a revisionary proposal, a concept of color whereby this more fundamental causal power is essential. If these two proposals are seen as competing theories, it is not clear that one is preferable to the other. Moreover it is not clear why we could not adopt both proposals and have two concepts of color, just as in the case of the geometrical property, size, we have two concepts: absolute (intrinsic) size, and angular size. That this ecumenical solution represents a viable option receives support from the consideration that if we are thinking of the proposal for the objectivist concept we have at least two proposals: one framed in terms of intrinsic microstructural properties, and one framed in terms of dispositional light-related properties. Again rather that having to choose between them, we could adopt both proposals and admit that there are different kinds of color.

In conclusion, therefore, it would appear that in so far as the objectivist is offering us a reconstruction of our original concept of color, there is reason to think that a dispositional analysis would provide a construction that is at least as good. There is strong reason, however, to think that an ecumenical solution to the problems of color can be found: that there is a place for different concepts of color, and with them different essences.

Objectivism: Problems and Solutions

It has been argued that there is room for both objectivist and psychological concepts of color, even though those providing analyses built on these concepts commonly see them as competitors. Properly interpreted, these accounts do not have to be seen as rivals. This claim can receives further support from consideration of two very different objectivist accounts, one by Jackson and Pargetter (1987), the other by David Hilbert (1987). Their respective solutions to some of the problems facing objectivism illustrate how each provides an objectivist concept of color that fits comfortably with a psychological concept, as well as with each other.

Jackson and Pargetter, who claim that each color can be identified with a physical property, have the explicit aim of overcoming the problem of multiple realizations. They concede that there is no single physical feature that is the basis for each color say, blue, but maintain that this does not matter. Blue is identified with a different physical property on different occasions, depending on what kind of physical object has it. This means relativizing the concept of color, to kinds of objects and circumstances. In principle there is no reason why there should not be a concept of physical color, in the way described by Jackson and Pargetter. The issue though is that we need some way to unify the various properties so as to bring them under the umbrella of color, and on the face of it, the psychological concept seems necessary.

The point is a general one. The objectivist who attempts to identify the objective essence for color must relate that essence to the way colors appear. Given that the criteria used by competent color perceivers to identify colors, depend on the appearances, it is necessary for the objectivist to spell out the nature of this relation. For the subjectivist the appearances constitute (part of) what the essence is. for the objectivist the appearance picks out the essence which is independent of the appearance.

In determining the right objectivist candidate, our aim is not simply to show how color vision enables, say, the observer to distinguish objects with different spectral reflectance characteristics. Rather we need to explain why one reflectance profile deserves to be classified as blue, and likewise why other profiles are related to similar and differing colors. Consequently to identify one reflectance profile as that corresponding to unique green, we need to be able to specify standard conditions and normal observers. As Hardin has persuasively pointed out, this cannot be done except in a highly arbitrary way. Not only is there a minority of color perceivers who are anomalous (only slightly, but appreciably so) with respect to normal observers, but there is a considerable statistical spread even within the group of normal observers. The reflectance profile for unique green will differ for different members of the "normal group". One can decide, of course, on a standard and fix one reflectance profile as green, but the procedure is highly arbitrary. As we have seen, there are few interesting causal powers associated with colors apart from the way objects affect perceivers. There is an alternative, however, and that is to tie color to appearance and, in consequence, relativize color to observers, with as much freedom or restriction, according to context, as is required. The most natural way to relativize color is through a dispositional concept that ties color to the way it appears to observers of the right kind. Jackson's concept of physical color would need to be supplemented by the dispositional concept. The admissibility of such a concept would not mean that the dispositional concept ought to be eliminated.

A novel twist to the objectivist program has been provided by David Hilbert (1987), with his account known as ‘Anthropocentric Realism’. It provides a solution to the multiple realizations problem but one that still seems to supplement the psychological concept, and not to dispense with it. On this view, colors are identified with spectral reflectances, at least surface colors are. A distinction is drawn between this kind of color and anthropocentric color. Individual anthropocentric colors are associated with groups of spectral reflectances. Color perception and color language ‘give us anthropocentrically defined kinds of colors and not colors themselves’. [Hilbert (1987) p. 27.] Terms such as ‘red’, ‘blue’, ‘yellow’, etc are associated with anthropocentric colors. To be red, for example, is to have a reflectance that falls within a particular class of reflectances. These classes, in general, are highly anthropocentric, sharing few interesting causal powers, and being of little consequence, apart from how they connect with the

peculiarities that underlie human color vision. The principle of grouping is that a given perceived color is associated with ‘a triple of integrated reflectances’. This association is based on the fact that human color vision depends on the use of ‘three types of broad band sensors’, i.e., the three types of light-sensitive receptors [Hilbert (1987) p. 111]

Colors, on this view, are both objective and anthropocentric. This would help explain why there is a color science and no science of color. It can also be readily modified so as to handle that problem whose resolution seems to require the relativization of color to kinds of observers. Once the concept of anthropocentric color is in place, it can be relativized, if necessary, to groups of observers.

It is Hilbert's claim that with this analysis, many of the ‘common-sense claims’ about color can be preserved, e.g., that orange is more similar to red than it is to blue. The point is that the triples of integrated reflectances can be taken as co-ordinates in a three-dimensional space, thus defining a color space. Similar colors will be located at adjacent points in this space. It is claimed that the right interpretation of statements of color similarity and dissimilarity is in terms of statements about relative location in color space.

It is with this claim that scepticism will most naturally arise. On the face of it, there is a certain qualitative character to ostensively defined color space, e.g as expressed in the Munsell or Swedish Natural color systems, that is not captured by the triple-reflectance color space. One measure of this fact is that changes along the dimensions of brightness and saturation have a different character from changes of hue from unique green to unique yellow to red to blue. Changes in Hilbert's color space don't seem to be of the right kind _ which of course is not to deny that they may not contribute to a causal explanation for why the psychological color spaces have the character that they do.

To conclude: the existence of the various problems facing the objectivist proposals do not demonstrate that the objectivist concepts are not viable. The solutions offered, however, cast doubt on the claim that objectivist concepts stand in no need of supplementary psychological concepts of color.

An Ecological View of Color

There is another theory of colors which has something in common with the illusion theory, in that it rejects objectivist accounts, but which is crucially different. It is the theory defended by Evan Thompson, the Ecological View of Colors, and is designed to be consonant with J.J.Gibson's views on perception. On this account, colors are taken to be dependent, in part, on the perceiver and so are not intrinsic properties of a perceiver-independent world. This account is not the same as an illusion theory. Being colored, instead, is construed as a relational property of the environment, connecting the environment with the perceiving animal. In the case of the color of physical surfaces, "being colored corresponds to the surface spectral reflectance as visually perceived by the animal". [Thompson (1995) Ch. 5, pp. 242-50.]

In more detail this account is spelled out in the following way: "being colored a particular determinate color or shade is equivalent to having a particular spectral reflectance, illuminance, or emittance that

looks that color to a particular perceiver in specific viewing conditions" [p.245]. Thompson insists that this account is to be distinguished from both a Lockean dispositionalist account and an illusion theory of colors. It is difficult to see, however, how he can maintain this stand. For one thing, he concedes that we see colors as perceiver-independent properties of things while maintaining that colors are perceiver-dependent properties. His answer to this difficulty, i.e., to why this is not a form of the illusion theory, is that on the ecological view it is not possible to perceive color as relational. That is, the relational nature of color does not allow the perceiver to perceive colors as relational. But this answer is not an answer to the question posed. What it explains is why one should not be surprised to find that, on the ecological view, that colors are experienced as perceiver-independent properties. But this is to admit that the way colors are represented in experience is not the way they are. The illusion theory denies that objects have the property (the color) they are represented as having. It need not deny that it is possible to formulate another concept of color that objects do satisfy. What it insists upon is that there is a need for the concept of color in the illusory sense.

A Pluralist Framework

If analysis of the natural concept of color leads to an illusion theory, or to the theory of colors as virtual properties, we still need to develop an account that prescribes how we should, in the future, think about color, at least in general terms. For practical purposes, it does not matter at all that colors are virtual properties. For these purposes, it is sufficient if "it is as if there are colors"; i.e, these purposes are served equally as well if objects appear to be colored. They do not need to be really colored.

There are other, more theoretical purposes for which we need to develop a more comprehensive account of color, one that specifies other concepts of color. The best such account is one that sets out a pluralist framework, one that allows for a variety of different concepts of color, including objectivist and psychological concepts, and arguably, ecological and phenomenal concepts. Moreover, such a framework does not require us to reject the natural or folk concept.

That there is scope for more than one concept of color should not be surprising. The natural concept of color is intended to serve a range of purposes. We find, though, that nothing exists that satisfies all the requirements. However, all is not lost. It is possible to develop a new set of color concepts that as a whole serves all or most of the previous purposes. None of them taken singly serves all, but each serves some. It is built into the natural or folk concept that colors have, broadly speaking, two major roles: (i) colors have a causal role to play in color perception; (ii) colors serve a variety of epistemological, aesthetic and emotional purposes. Colors serve the latter set of purposes through the way colors appear. Once it is recognized that colors, as specified by the traditional concept, are virtual properties and that there is no property that serves all the functions relevant to that concept, the way is open to recognize two new concepts of color: dispositional, psychological color, to take over and consolidate the role served by the appearance, and physical color, to take over the causal role. Moreover, once it is revealed that the cause of color perception are complex, it is open for us to see the point of having several physical or objectivist concepts of color, one framed in terms of microstructural properties, the other in terms of light-related properties.

To argue in this way for the place of a number of concepts of color, and for the possibility of an objectivist concept, to supplement other concepts of color, is to argue for a pluralist framework for colors. This framework has the advantage of allowing a place for an objective concept of color, while not making it mandatory. Whether or not there is any point in having an objective concept, there is, as we have seen, a need for a dispositional concept, one tied to the appearance of color. The dispositional concept is a crucial part of the pluralist framework.

But once we become enlightened by accepting the theory of virtual colors, how should we then think of the dispositional concept? What exactly does the exercise of the disposition consist in? What exactly is the content of the dispositional concept? The right answer is that there are two parts to the dispositionalist concept. One part refers to the way objects appear, and the other to the feature, whatever it is, which is the causal basis for the appearance. That is, the disposition is not pure but ‘mixed’. BlueD objects are objects that have some feature by virtue of which they look as if they are blue, i.e., blue in the intrinsic sense, i.e., blue in the virtual-color sense. To say that this sense of color is the virtual-color sense is not to say that colors are ordinarily conceived of as virtual. It is to say that the properties colors are conceived of as being are virtual. The content of the dispositional concept thus presupposes the virtual-color concept. This means that there is point in retaining this concept, even when we come to know that no objects have the property. The fact that I do not believe that this property of intrinsic blueness is ever instantiated does not mean that I should give up the concept, any more than disbelievers in Satan should give up the concept of satanic.

In this state of theoretical sophistication, my use of the natural concept to describe things requires me to adopt the naive attitude to color or, preferably, the engaged attitude typical of the playgoer who, at the theatre, suspends his belief that ‘it is all a pack of lies’. Of course as philosophers, we need to understand why we have this virtual-color concept and what role it plays, and how it works. But none of that stops me from continuing to employ the virtual-color concept, whether as scientist, artist, consumer, town-planner, interior decorator or philosopher. As for serving functions such as being signs or as being aesthetically or emotionally significant, virtual colors are as good as real colors.

There is no need therefore to jettison the natural concept. Realizing however that the color properties are virtual properties means that, for our understanding of how such a concept should apply and why it is so beneficial. Part of this understanding is provided by the explanation for why we have the natural concept that we do. The explanation for why the natural concept is beneficial is that the purposes served by the concepts are equally well served if objects merely appear to be colored and are not actually colored.

Colors as Phenomenal Qualities

Given the virtual colors theory, there is a further problem that needs to be resolved: to explain how it is that the natural concept takes the form that it does, and in particular how it contains the basis for the structure that underpins the color systems for ordering colors. One of the characteristics of the color properties captured in the natural concept is that colors are the sorts of properties among which a set of

internal relationships hold. That is, colors as a block, have a qualitative, sensuous character that enables objects having them to be systematically ordered and arranged. There are no physical features that have this character. One explanation for why this character is part of the virtual color property is that our sensory representations (or the elements/aspects in our visual experiences) have the qualitative character. It is because the sensory representations have the qualitative character that they do, that they represent physical objects as having the qualitative character in question.

The way that the phenomenal concept, i.e., the concept of color as a phenomenal property is introduced is that it serves to explain why the natural concept of color has the character that it does. When we have color experiences, typically we form sensory representations of the world. These representations represent objects in the physical world as having (virtual) colors, and they do so because the representations have the character implicit in three-dimensional color arrays. The representations do not have virtual colors (they have the right kind of structure, but they do not have the right causal powers), but they represent physical objects as having those colors. Sensory representations, in other words, have the phenomenological character that physical objects might have had but do not.

The qualitative character that the sensory representations have is sensory and phenomenal in the strongest sense. The character is ontologically subjective. In visual experience we experience the sensory color qualities as being in a public three dimensional space. That is, our experiences, and our sensory representations, represent the color qualities as being on the surfaces of physical objects, or as otherwise located in physical space. Contrary to what some philosophers believe, there is no more problem in experiencing phenomenal qualities in such a way than there is in feeling a pain in a foot or an elbow. In the case of pains, the phenomenal quality is felt on a bodily location, e.g., behind the eye, or in an elbow, etc. It needs to be said that although our sensory representations have phenomenal color qualities, which we are aware of, we are not aware of them as phenomenal qualities, that is, as phenomenal qualities of physical objects. We use the sensory representations as signs for physical objects, but we are not aware of the sign as a sign. It needs to be stressed that this account does not require colors, either phenomenal or virtual, to be projected into space. Just as they represent objects as having virtual colors, so they represent objects as having spatial properties (and relations) through themselves having phenomenal spatial properties.

Conclusion

It has been argued an adequate account of color must, in the first place, provide an account of the folk concept or natural concept of color. Such an account, there is reason to believe, is an illusion theory of color. Supposing that colors, as we normally think of them, are virtual properties, we are faced with the question of how, if at all, should we adjust our ideas in thinking of colors. If normally our perception of color involves ‘false consciousness’, what is the right way to think of colors? The answer to that is that for many purposes we should continue to think of them in the same way as we always did. In the case of color, unlike other cases, false consciousness should be a cause for celebration.

Although for most practical purposes, it does not matter that colors are virtual properties, there are more

theoretical purposes, however, for which we need to develop a more comprehensive account, a pluralist account of color. The different elements of the natural concept of color reflect different functions that colors are meant to play. Given these different functions and the fact that there is no property that satisfies all of them, it is open to us to develop a pluralist framework in which different concepts of color take over different functions. This pluralist framework makes room for the introduction of objectivist concepts of color, but such concepts need to stand beside a dispositionalist concept which makes reference to the way colors appear to perceivers, and arguably, with the latter, a phenomenal concept. It is such a framework that is necessary to give an adequate account of the rich epistemological and socially important roles that colors play.

Bibliography

- Agoston, G. (1987), *Color Theory and Application in Art and Design*, Berlin: Springer.
- Armstrong, D. M. (1969), 'Color-Realism and the Argument from Microscopes', Brown and Rollins (1969), *Contemporary Philosophy in Australia*, London: Allen and Unwin, pp. 119-31.
- Berlin, B., and Kay, P. (1969), *Basic Color Terms*, Berkeley and Los Angeles: University of California Press.
- Boynton, R.M. (1978), 'Color in Contour and Object Perception', in Carterette and Friedman, (eds.) (1978), Handbook of Perception, vol. 8, New York: Academic Press, pp. 173-98.
- Boynton R.M. and Olson C.X. (1990), 'Salience of chromatic basic color terms confirmed by three measures', *Vision Research*, **30**, 1311-17.
- Broackes, Justin (1992), 'The Autonomy of Color', in Charles, David, and Lennon, Kathleen (eds.) (1992), *Reduction, Explanation and Realism*, Oxford: Clarendon Press, pp. 421-66.
- Byrne, Alex and Hilbert, David R., (1997), *Readings on Color, Vol. I: The Philosophy of Color*, Camb.Mass. : M.I.T Press.
- Byrne, Alex and Hilbert, David R., (1997), *Readings on Color, Vol. II: The Science of Color*, Camb.Mass. : M.I.T Press.
- Campbell, J. (1994), 'A Simple View of Color', in Haldane, John, and Wright, Crispin (eds.) (1994), *Reality, Representation and Projection*, Oxford: Clarendon Press, pp. 257-69.
- Campbell, Keith (1969), 'Colors', in Brown and Rollins (1969), *Contemporary Philosophy in Australia*, London: Allen and Unwin, pp. 132-57.
- Cohen, Jonathan (2001), 'Subjectivism, Physicalism, or None of the Above . . .', *Consciousness and Cognition*, 10, pp. 94-104.
- Evans, Gareth, (1980), 'Things Without the Mind', in Z. v Straaten, *Philosophical Subjects*, **10**, pp. 76-116.
- Hacker, P. M. S. (1987), *Appearance and Reality*, Oxford: Blackwell Publisher.
- Hard, Anders, and Sivik, Lars (1981), 'NCS-Natural Color System: A Swedish Standard for Color Notation', *Color Research and Application*, **6**, pp. 129-38.
- Hardin, C. L. (1988/1993), *Color for Philosophers*, Indianapolis, Ind.: Hackett.
- Hardin C.L. & Maffi L. (1997), *Color categories in thought and language*, Cambridge: C.U.P.
- Hering, E. (1964), *Outlines of a Theory of the Light Sense*, trans. L. Hurvich and D. Jameson, Cambridge, Mass.: Harvard University Press.

- Hilbert, D. R. (1987), *Color and Color Perception*, Stanford, Calif.: C.S.L.I.
- Jackson Frank, (1996), ‘The Primary Quality View of Color’, *Philosophical Perspectives*, **10**, pp. 199-219.
- Jackson F. & Pargetter R., (1987), ‘An objectivist's guide to subjectivism about color’, *Revue Internationale de Philosphie*, **160**, pp. 129-41.
- Kaiser P.K. and Boynton R.M. (1996), *Human Color Vision*, (2nd edition) Washington: Optical Society of America.
- Kuehni, R. (1997) *Color* New York: J. Wiley and Sons.
- Land, E. H. (1983), ‘Recent Advances in Retinex Theory . . .’, *Proceedings of the National Academy of Sciences*, **80**, pp.5163-9.
- Landesman, C. (1989), *Color and Consciousness*, Philadelphia: Temple University Press.
- Lewis, David, (1997), ‘Naming the Colors’, *Australasian Journal of Philosophy*, **75**, pp. 325-42.
- Maund, J. B. (Barry) (1991), ‘The Nature of Color’, *History of Philosophy Quarterly*, **8**, pp. 253-63.
- Maund, Barry (1995), *Colors: Their Nature and Representation*, Cambridge: Camb.University Press.
- Matthen, M., (2000) ‘The Disunity of Color’, *Philosophical Review*, 108(1), pp. 47-84.
- McLaughlin, B. (2001), ‘The Place of Color in Nature’, in R.Mausfield and D.Heyer (Eds.), *Color Percepton: From Light to Object*, New York: Oxford University Press.
- Nassau, K. (1983), *The Physics and Chemistry of Color*, New York: Wiley.
- Price H. H. (1932). *Perception*, London: Methuen.
- Ross, P. (2001), ‘The location problem for color subjectivism’, *Consciousness and Cognition*, 10, pp. 42-58.
- Stroud,B. (2000), *The Quest for Reality: Subjectivism and the Metaphysics of Color*, New York: Oxford University Press.
- Thompson, Evan (1995), *Color Vision*, London: Routledge.
- Tye, M. (2000), *Consciousness, Color, and Content*, Cambridge, Mass.: MIT/Bradford.
- Van Brakel, J. (1993), ‘The Plasticity of Categories: The Case of Color’, *British Journal for the Philosophy of Science*,**XL** 44, pp. 103-35.
- Westphal, Jonathan (1987), *Color: A Philosophical Introduction*, 1st ed., Oxford: Blackwell Publisher.

Other Internet Resources

- The Munsell Color System (Adobe Technical Guide)

[Please contact the author with other suggestions.]

Related Entries

concepts | Descartes, René | Locke, John | qualia | realism | reduction and reductionism

Barry Maund
jbmaund@cyllene.uwa.au.edu

First published: December 1, 1997
Content last modified: July 9, 2002

Realism

The nature and plausibility of realism is one of the most hotly debated issues in contemporary metaphysics, perhaps even the most hotly debated issue in contemporary philosophy. The question of the nature and plausibility of realism arises with respect to a large number of subject matters, including ethics, aesthetics, causation, modality, science, mathematics, semantics, and the everyday world of macroscopic material objects and their properties. Although it would be possible to accept (or reject) realism across the board, it is more common for philosophers to be selectively realist or non-realist about various topics: thus it would be perfectly possible to be a realist about the everyday world of macroscopic objects and their properties, but a non-realist about aesthetic and moral value. In addition, it is misleading to think that there is a straightforward and clear-cut choice between being a realist and a non-realist about a particular subject matter. It is rather the case that one can be more-or-less realist about a particular subject matter. Also, there are many different forms that realism and non-realism can take.

The question of the nature and plausibility of realism is so controversial that no brief account of it will satisfy all those with a stake in the debates between realists and non-realists. This article offers a broad brush characterisation of realism, and then fills out some of the detail by looking at a few canonical examples of opposition to realism. The discussion of forms of opposition to realism is far from exhaustive and is designed only to illustrate a few paradigm examples of the form such opposition can take.

There are two general aspects to realism, illustrated by looking at realism about the everyday world of macroscopic objects and their properties. First, there is a claim about *existence*. Tables, rocks, the moon, and so on, all exist, as do the following facts: the table's being square, the rock's being made of granite, and the moon's being spherical and yellow. The second aspect of realism about the everyday world of macroscopic objects and their properties concerns *independence*. The fact that the moon exists and is spherical is independent of anything anyone happens to say or think about the matter. Likewise, although there is a clear sense in which the table's being square is dependent on us (it was designed and constructed by human beings after all), this is not the type of dependence that the realist wishes to deny. The realist wishes to claim that apart from the mundane sort of empirical dependence of objects and their properties familiar to us from everyday life, there is no *further* sense in which everyday objects and their properties can be said to be dependent on anyone's linguistic practices, conceptual schemes, or whatever.

In general, where the distinctive objects of a subject-matter are *a*, *b*, *c*, … , and the distinctive properties are …*is F*, …*is G*, …*is H* and so on, realism about that subject matter will typically take the form of a claim like the following:

> **Generic Realism**:
> *a*, *b*, and *c* and so on exist, and the fact that they exist and have properties such as *F-ness*, *G-ness*, and *H-ness* is (apart from mundane empirical dependencies of the sort sometimes encountered in everyday life) independent of anyone's beliefs, linguistic practices, conceptual schemes, and so on.

Non-realism can take many forms, depending on whether or not it is the existence or independence dimension of realism that is questioned or rejected. The forms of non-realism can vary dramatically from subject-matter to subject-matter, but error-theories, non-cognitivism, instrumentalism, nominalism, certain styles of reductionism, and eliminativism typically reject realism by rejecting the existence dimension, while idealism, subjectivism, and anti-realism typically concede the existence dimension but reject the independence dimension. Philosophers who subscribe to quietism deny that there can be such a thing as substantial metaphysical debate between realists and their non-realist opponents.

---

1. Preliminaries

Three preliminary comments are needed. Firstly, there has been a great deal of debate in recent philosophy about the relationship between realism, construed as a metaphysical doctrine, and doctrines in the theory of meaning and philosophy of language concerning the nature of truth and its role in accounts of linguistic understanding (see Dummett 1978 and Devitt 1991a for radically different views on the issue). Independent of the issue about the relationship between metaphysics and the theory of meaning, the well-known disquotational properties of the truth-predicate allow claims about objects, properties, and facts to be framed as claims about the truth of sentences. Since:

> (1) ‘The moon is spherical’ is true if and only if the moon is spherical.

The claim that the moon exists and is spherical can be framed independently of anyone's beliefs, linguistic practices and conceptual schemes, as the claim that the sentences ‘The moon exists’ and ‘The moon is spherical’ are true independently of anyone's beliefs, linguistic practices, conceptual schemes and so on. As Devitt points out (1991b: 46) availing oneself of this way of talking does not entail that one sees the metaphysical issue of realism as ‘really’ a semantic issue about the nature of truth (if it did, any question about any subject matter would turn out to be ‘really’ a semantic issue).

Secondly, although in introducing the notion of realism above mention is made of objects, properties, and facts, no theoretical weight is attached to the notion of a ‘fact’, or the notions of ‘object’ and ‘property’. To say that it is a fact that the moon is spherical is just to say that the object, the moon, instantiates the property of being spherical, which is just to say that the moon is spherical. There are substantial metaphysical issues about the nature of facts, objects, and properties, and the relationships between them (see Mellor and Oliver 1997 and Lowe 2002, part IV), but these are not of concern here.

Thirdly, as stated above, Generic Realism about the mental or the intentional would strictly speaking appear to be ruled out *ab initio*, since clearly Jones' believing that Cardiff is in Wales is not independent of facts about belief: trivially, it is dependent on the fact that Jones believes that Cardiff is in Wales. However, such trivial dependencies are not what are at issue in debates between realists and non-realists about the mental and the intentional. A non-realist who objected to the independence dimension of realism about the mental would claim that Jones' believing that Cardiff is in Wales depends in some *non-trivial* sense on facts about beliefs, etc.

2. Against the Existence Dimension (I): Error-Theory and Arithmetic

There are at least two distinct ways in which a non-realist can reject the existence dimension of realism about a particular subject matter. The first of these rejects the existence dimension by rejecting the claim that the distinctive objects of that subject-matter exist, while the second admits that those objects exist but denies that they instantiate any of the properties distinctive of that subject-matter. Non-realism of the first kind can be illustrated via Hartry Field's error-theoretic account of arithmetic, and non-realism of the second kind via J.L. Mackie's error-theoretic account of morals. This will show how realism about a subject-matter can be questioned on both epistemological and metaphysical grounds.

According to a *platonist* about arithmetic, the truth of the sentence ‘7 is prime’ entails the existence of an *abstract object*, the number 7. This object is abstract because it has no spatial or temporal location, and is causally inert. A *platonic realist* about arithmetic will say that the number 7 exists and instantiates the property of being prime independently of anyone's beliefs, linguistic practices, conceptual schemes, and so on. A certain kind of nominalist rejects the existence claim which the platonic realist makes: there are no abstract objects, so sentences such as ‘7 is prime’ are *false* (hence the name ‘error-theory’). Platonists divide on their account of the epistemology of arithmetic: some claim that our knowledge of arithmetical fact proceeds by way of some quasi-perceptual encounter with the abstract realm (Gödel 1983), while

others have attempted to resuscitate a qualified form of Frege's *logicist* project of grounding knowledge of arithmetical fact in knowledge of logic (Wright 1983, Hale 1987, Hale and Wright 2001).

The main arguments against platonic realism turn on the idea that the platonist position precludes a satisfactory epistemology of arithmetic. For the classic exposition of the doubt that platonism can square its claims to accommodate knowledge of arithmetical truth with its conception of the subject matter of arithmetic as causally inert, see Benacerraf (1973). Benacerraf argued that platonism faces difficulties in squaring its conception of the subject-matter of arithmetic with a general causal constraint on knowledge (roughly, that a subject can be said to know that *P* only if she stands in some causal relation to the subject matter of *P*). In response, platonists have attacked the idea that a plausible causal constraint on ascriptions of knowledge can be formulated (Wright 1983 Ch.2, Hale 1987 Ch.4). In response, Hartry Field, on the side of the anti-platonists, has developed a new variant of Benacerraf's epistemological challenge which does not depend for its force on maintaining a generalised causal constraint on ascriptions of knowledge. Rather, Field's new epistemological challenge to platonism arises from his reasonable observation that ‘we should view with suspicion any claim to know facts about a certain domain if we believe it impossible to explain the reliability of our beliefs about that domain’ (Field 1989: 232-3). Field's challenge to the platonist is to offer an account of what such a platonist should regard as a datum—i.e. that when ‘*p*’ is replaced by a mathematical sentence, the schema (2) holds in most instances :

> (2) If mathematicians accept ‘*p*’ then *p*. (1989: 230)

Field's point is not simply, echoing Benacerraf, that no causal account of reliability will be available to the platonist, and therefore to the platonic realist. Rather, Field conceives what is potentially a far more powerful challenge to platonic realism when he suggests that not only has the platonic realist no recourse to any explanation of reliability that is causal in character, but that she has no recourse to any explanation that is non-causal in character either. He writes:

> (T)here seems prima facie to be a difficulty in principle in explaining the regularity. The problem arises in part from the fact that mathematical entities as the [platonic realist] conceives them, do not causally interact with mathematicians, or indeed with anything else. This means we cannot explain the mathematicians beliefs and utterances on the basis of the mathematical facts being causally involved in the production of those beliefs and utterances; or on the basis of the beliefs or utterances causally producing the mathematical facts; or on the basis of some common cause producing both. Perhaps then some sort of non-causal explanation of the correlation is possible? Perhaps; but it is very hard to see what this supposed non-causal explanation could be. Recall that on the usual platonist picture [i.e. platonic realism], mathematical objects are supposed to be mind- and language-independent; they are supposed to bear no spatiotemporal relations to anything, etc. The problem is that the claims that the [platonic realist] makes about mathematical objects appears to rule out any reasonable strategy for explaining the systematic correlation in question. (1989: 230-1)

This suggests the following dilemma for the platonic realist:

i. Platonic realism is committed to the existence of acausal objects and to the claim that these objects, and facts about them, are independent of anyone's beliefs, linguistic practices, conceptual schemes, and so on (in short to the claim that these objects, and facts about them, are language- and mind-independent).
ii. Any causal explanation of reliability is incompatible with the acausality of mathematical objects.
iii. Any non-causal explanation of reliability is incompatible with the language- and mind-independence of mathematical objects.
iv. Any explanation of reliability must be causal or non-causal.
v. There is no explanation of reliability that is compatible with both the acausality and language- and mind-independence of mathematical objects.

Therefore,

vi. There is no explanation of reliability that is compatible with platonic realism.

Whether there is a version of platonic realism with the resources to see off Field's epistemological challenge is very much a live issue (see Hale 1994, Divers and Miller 1999).

What does Field propose as an alternative to platonic realism in arithmetic? Field's answer (1980, 1989) is that although mathematical sentences such as ‘7 is prime’ are false, the utility of mathematical theories can be explained otherwise than in terms of their truth. For Field, the utility of mathematical theories resides not in their truth but in their *conservativeness*, where a mathematical theory *S* is conservative if and only if for any nominalistically respectable statement *A* (i.e. a statement whose truth does not imply the existence of abstract objects) and any body of such statements *N*, *A* is not a consequence of the conjunction of *N* and *S* unless *A* is a consequence of *N* alone (Field 1989: 125). In short, mathematics is useful, not because it allows you to derive conclusions that you couldn't have derived from nominalistically respectable premises alone, but rather because it makes the derivation of those (nominalistically respectable) conclusions easier than it might otherwise have been. Whether or not Field's particular brand of error-theory about arithmetic is plausible is a topic of some debate, which unfortunately I cannot pursue further here (see Hale and Wright 2001).

3. Against the Existence Dimension (II): Error-Theory and Morality

According to Field's error-theory of arithmetic, the objects distinctive of arithmetic do not exist, and it is this which leads to the rejection of the existence dimension of arithmetical realism, at least as platonistically conceived (for a non-platonistic view of arithmetic which is at least potentially realist, see Benacerraf 1965; for incisive discussion, see Wright 1983, Ch.3). J. L. Mackie , on the other hand, proposes an error-theoretic account of morals, not because there are no objects or entities that could form the subject matter of ethics (it is no part of Mackie's brief to deny the existence of persons and their

actions and so on), but because it is implausible to suppose that the sorts of properties that moral properties would have to be are ever instantiated in the world (Mackie 1977, Ch.1). Like Field on arithmetic, then, Mackie's central claim about the atomic, declarative sentences of ethics (such as ‘Napoleon was evil’) is that they are systematically and uniformly false. How might one argue for such a radical-sounding thesis? The clearest way to view Mackie's argument for the error-theory is as a conjunction of a conceptual claim with an ontological claim (following Smith 1994, pp.63-66). The conceptual claim is that our concept of a moral fact is a concept of an objectively prescriptive fact, or, equivalently, that our concept of a moral property is a concept of an objectively prescriptive quality (what Mackie means by this is explained below). The ontological claim is simply that there are no objectively prescriptive facts, that objectively prescriptive properties are nowhere instantiated. The conclusion is that there is nothing in the world answering to our moral concepts, no facts or properties which render the judgements formed via those moral concepts true. Our moral judgements are all of them false. We can thus construe the error-theory as follows:

> **Conceptual Claim:** our concept of a moral fact is a concept of an objectively prescriptive fact, so that the truth of an atomic, declarative moral sentence would require the existence of objectively and categorically prescriptive facts.
>
> **Ontological Claim:** There are no objectively and categorically prescriptive facts.
>
> So,
>
> **Conclusion:** there are no moral facts; atomic, declarative moral sentences are systematically and uniformly false.

This argument is clearly valid, so the question facing those who wish to defend at least the existence dimension of realism in the case of morals is whether the premises are true.

Mackie's conceptual claim is that our concept of a moral requirement is the concept of an objectively, categorically prescriptive requirement. What does this mean? To say that moral requirements are prescriptive is to say that they tell us how we ought to act, to say that they give us reasons for acting. Thus, to say that something is morally good is to say that we ought to pursue it, that we have reason to pursue it. To say that something is morally bad is to say that we ought not to pursue it, that we have reason not to pursue it. To say that moral requirements are categorically prescriptive is to say that these reasons are categorical in the sense of Kant's categorical imperatives. The reasons for action that moral requirements furnish are not contingent upon the possession of any desires or wants on the part of the agent to whom they are addressed: I cannot release myself from the requirement imposed by the claim that torturing the innocent is wrong by citing some desire or inclination that I have. This contrasts, for example, with the requirement imposed by the claim that perpetual lateness at work is likely to result in one losing one's job: I can release myself from the requirement imposed by this claim by citing my desire to lose my job (perhaps because I find it unfulfilling, or whatever). Reasons for action which are contingent in this way on desires and inclinations are furnished by what Kant called hypothetical imperatives.

So our concept of a moral requirement is a concept of a categorically prescriptive requirement. But Mackie claims further that our concept of a moral requirement is a concept of an objectively categorically prescriptive requirement. What does it mean to say that a requirement is objective? Mackie says a lot of different-sounding things about this, and the following is by no means a comprehensive list (references are to Ch. 1 of Mackie 1977). To call a requirement objective is to say that it can be an object of knowledge (24, 31, 33), that it can be true or false (26, 33), that it can be perceived (31, 33), that it can be recognised (42), that it is prior to and independent of our preferences and choices (30, 43), that it is a source of authority external to our preferences and choices (32, 34, 43), that it is part of the fabric of the world (12), that it backs up and validates some of our preferences and choices (22), that it is capable of being simply true (30) or valid as a matter of general logic (30), that it is not constituted by our choosing or deciding to think in a certain way (30), that it is extra-mental (23), that it is something of which we can be aware (38), that it is something that can be introspected (39), that it is something that can figure as a premise in an explanatory hypothesis or inference (39), and so on. Mackie plainly does not take these to be individually necessary: facts about subatomic particles, for example, may qualify as objective in virtue of figuring in explanatory hypotheses even though they cannot be objects of perceptual acquaintance. But his intention is plain enough: these are the sorts of conditions whose satisfaction by a fact renders it objective as opposed to subjective. Mackie's conceptual claim about morality is thus that our concept of a moral requirement is a concept of a fact which is objective in at least some of the senses just listed, while his ontological claim will be that the world does not contain any facts which are both candidates for being moral facts and yet which play even some of the roles distinctive of objective facts.

How plausible is Mackie's conceptual claim? This issue cannot be discussed in detail here, except to note that while it seems plausible to claim that *if* our concept of a moral fact is a concept of a reason for action *then* that concept must be a concept of a categorical reason for action, it is not so clear why we have to say that our concept of a moral fact is a concept of a reason for action at all. If we deny this, we can concede the conditional claim whilst resisting Mackie's conceptual claim. One way to do this would be to question the assumption, implicit in the exposition of Mackie's argument for the conceptual claim above, that an ‘ought’-statement that binds an agent A provides that agent with a reason for action. For an example of a version of moral realism that attempts to block Mackie's conceptual claim in this way, see Railton (1986). For defence of Mackie's conceptual claim, see Smith (1994), Ch.3. For exposition and critical discussion, see Miller (2003a), Ch.9.

What is Mackie's argument for his ontological claim? This is set out in his ‘argument from queerness’ (Mackie has another argument, the ‘argument from relativity’(1977: 36-38), but this argument cannot be discussed here).The argument from queerness has both metaphysical and epistemological components. The metaphysical problem with objective values concerns ‘the metaphysical peculiarity of the supposed objective values, in that they would have to be intrinsically action-guiding and motivating’(49). The epistemological problem concerns ‘the difficulty of accounting for our knowledge of value entities or features and of their links with the features on which they would be consequential’(49). Let's look at each type of worry more closely in turn.

Expounding the metaphysical part of the argument from queerness, Mackie writes: “If there were

objective values, then they would be entities or relations of a very strange sort, utterly different from anything else in the universe."(38) What is so strange about them? Mackie says that Plato's Forms (and for that matter, Moore's non-natural qualities) give us a 'dramatic picture' of what objective values would be, if there were any:

> The Form of the Good is such that knowledge of it provides the knower with both a direction and an overriding motive; something's being good both tells the person who knows this to pursue it and makes him pursue it. An objective good would be sought by anyone who was acquainted with it, not because of any contingent fact that this person, or every person, is so constituted that he desires this end, but just because the end has to-be-pursuedness somehow built into it. Similarly, if there were objective principles of right and wrong, any wrong (possible) course of action would have not-to-be-doneness somehow built into it. Or we should have something like Clarke's necessary relations of fitness between situations and actions, so that a situation would have a demand for such-and-such an action somehow built into it (40).

The obtaining of a moral states of affairs would be the obtaining of a situation 'with a demand for such and such an action somehow built into it'; the states of affairs which we find in the world do not have such demands built into them, they are 'normatively inert', as it were. Thus, the world contains no moral states of affairs, situations which consist in the instantiation of a moral quality.

Mackie now backs up this metaphysical argument with an epistemological argument:

> If we were aware [of objective values], it would have to be by some special faculty of moral perception or intuition, utterly different from our ways of knowing everything else. These points were recognised by Moore when he spoke of non-natural qualities, and by the intuitionists in their talk about a faculty of moral intuition. Intuitionism has long been out of favour, and it is indeed easy to point out its implausibilties. What is not so often stressed, but is more important, is that the central thesis of intuitionism is one to which any objectivist view of values is in the end committed: intuitionism merely makes unpalatably plain what other forms of objectivism wrap up (38).

In short, our ordinary conceptions of how we might come into cognitive contact with states of affairs, and thereby acquire knowledge of them, cannot cope with the idea that the states of affairs are objective values. So we are forced to expand that ordinary conception to include forms of moral perception and intuition. But these are completely unexplanatory: they are really just placeholders for our capacity to form correct moral judgements (the reader should here hear an echo of the complaints Benecerraf and Field raise against arithmetical platonism).

Evaluating the argument from queerness is well outwith the scope of the present entry. While Railton's version of moral realism attempts to block Mackie's overall argument by conceding his ontological claim whilst rejecting his conceptual claim, other versions of moral realism agree with Mackie's conceptual claim but reject his ontological claim. Examples of the latter version, and attempts to provide the owed

response to the argument from queerness, can be found in Smith (1994), Ch.6, and McDowell (1998a), Chs 4-10 .

There are two main ways in which one might respond to Mackie's argument for the error-theory: directly, via contesting one of its premises or inferences, or indirectly, pointing to some internal tension within the error-theory itself. Some possible direct responses have already mentioned, responses which reject either the conceptual or ontological claims that feature as premises in Mackie's argument for the error-theory. An indirect argument against the error-theory has been developed in recent writings by Crispin Wright (this argument is intended to apply also to Field's error-theory of arithmetic).

Mackie claims that the error-theory of moral judgement is a second-order theory, which does not necessarily have implications for the first order practice of making moral judgements (1977: 16). Wright's argument against the error-theory takes off with the forceful presentation of the opposing suspicion:

> The great discomfort with [Mackie's] view is that, unless more is said, it simply relegates moral discourse to bad faith. Whatever we may once have thought, as soon as philosophy has taught us that the world is unsuited to confer truth on any of our claims about what is right, or wrong, or obligatory, etc., the reasonable response ought surely to be to forgo the right to making any such claims …. If it is of the essence of moral judgement to aim at the truth, and if philosophy teaches us that there is no moral truth to hit, how are we supposed to take ourselves seriously in thinking the way we do about any issue which we regard as of major moral importance? (1996: 2; see also 1992: 9).

Wright realises that the error-theorist is likely to have a story to tell about the point of moral discourse, about "some norm of appraisal besides truth, at which its statements can be seen as aimed, and which they can satisfy."(1996: 2) And Mackie has such a story: the point of moral discourse is—to simplify—to secure the benefits of social co-operation (1973: chapter 5 passim; note that this is the analogue in Mackie's theory of Field's notion of the conservativeness of mathematical theories). Suppose we can extract from this story some subsidiary norm distinct from truth, which governs the practice of forming moral judgements. Then, for example, 'Honesty is good' and 'Dishonesty is good', although both false, will not be on a par in point of their contribution to the satisfaction of the subsidiary norm: if accepted widely enough, the former will presumably facilitate the satisfaction of the subsidiary norm, while the latter, if accepted widely enough, will frustrate it. Wright questions whether Mackie's moral sceptic can plausibly combine such a story about the benefits of the practice of moral judgement with the central negative claim of the error-theory:

> [I]f, among the welter of falsehoods which we enunciate in moral discourse, there is a good distinction to be drawn between those which are acceptable in the light of some such subsidiary norm and those which are not—a distinction which actually informs ordinary discussion and criticism of moral claims—then why insist on construing *truth* for moral discourse in terms which motivate a charge of global error, rather than explicate it in terms of the satisfaction of the putative subsidiary norm, whatever it is? The question may have a good answer. The error-theorist may be able to argue that the superstition that he finds in

> ordinary moral thought goes too deep to permit of any construction of *moral* truth which avoids it to be acceptable as an account of moral truth. But I do not know of promising argument in that direction (1996: 3; see also 1992: 10).

Wright thus argues that even if we concede to the error-theorist that his original scepticism about moral truth is well-founded, the error-theorist's own positive proposal will be inherently unstable. For an attempt to respond to Wright's argument, on behalf of the error-theorist, see Miller 2002.

4. Reductionism and Non-Reductionism

Although some commentators (e.g. Pettit 1991) require that a realistic view of a subject matter be non-reductionist about the distinctive objects, properties, and facts of that subject matter, the reductionist/non-reductionist issue is really orthogonal to the various debates about realism. There are a number of reasons for this, with the reasons varying depending on the type of reduction proposed.

Suppose, first of all, that one wished to deny the existence claim which is a component of platonic realism about arithmetic. One way to do this would be to propose an *analytic reduction* of talk seemingly involving abstract entities to talk concerning only concrete entities. This can be illustrated by considering a language the truth of whose sentences seemingly entails the existence of a type of abstract object, directions. Suppose there is a first order language L, containing a range of proper names ‘*a*’, ‘*b*’, ‘*c*’, and so on, where these denote straight lines conceived as concrete inscriptions. There are also predicates and relations defined on straight lines, including ‘ … is parallel to …’. ‘*D*( )’ is a singular term forming operator on lines, so that inserting the name of a concrete line, as in ‘*D*(*a*)’, produces a singular term standing for an abstract object, the direction of *a*. A number of contextual definitions are now introduced:

> (A) ‘$D(a) = D(b)$’ is true if and only if *a* is parallel to *b*.
>
> (B) ‘$\Pi D(x)$’ is true if and only if ‘Fx’ is true, where ‘… is parallel to …’ is a congruence for ‘*F*( )’.
>
> (To say that ‘… is parallel to …’ is a congruence for ‘*F*( )’ is to say that if *a* is parallel to *b* and *Fa*, then it follows that *Fb*).
>
> (C) ‘$(\exists x)\Pi x$’ is true if and only if ‘$(\exists x)Fx$’ is true, where ‘Π’ and ‘*F*’ are as in (B).

According to a platonic realist, directions exist and have a nature which is independent of anyone's beliefs, linguistic practices, conceptual schemes, and so on. But doesn't the availability of (A), (B), and (C) undermine the existence claim at the heart of platonic realism? After all, (A), (B), and (C) allow us to paraphrase any sentence whose truth appears to entail the existence of abstract objects into a sentence whose truth involves only the existence of concrete inscriptions. Doesn't this show that an analytic reduction can aid someone wishing to question the existence claim involved in a particular form of

realism? There is a powerful argument, first developed by William Alston (1958), and recently resuscitated to great effect by Crispin Wright (1983, Ch.1), that suggests not. The analytic reductionist who wishes to wield the contextual definitions against the existence claim at the heart of platonic realism takes them to show that the apparent reference to abstract objects on the left-hand sides of the definitions is *merely* apparent: in fact, the truth of the relevant sentences entails only the existence of a range of concrete inscriptions. But the platonic realist can retort: what the contextual definitions show is that the apparent *lack* of reference to abstract objects on the *right-hand* sides is merely apparent. In fact, the platonic realist can say, the truth of the sentences figuring on the right-hand sides implicitly involves reference to abstract objects. If there is no way to break this deadlock the existence of the analytic reductive paraphrases will leave the existence claim at the heart of the relevant form of realism untouched. So the issue of this style of reductionism appears to be orthogonal to debates between realists and non-realists.

Can the same be said about non-analytic styles of reductionism? Again, there is no straightforward connection between the issue of reductionism and the issue of realism. The problem is that, to borrow some terminology and examples from Railton 1989, some reductions will be *vindicative* whilst others will be *eliminativist*. For example, the reduction of water to H_20 is vindicative: it vindicates our belief that there is such a thing as water, rather than overturning it. On the other hand:

> … the reduction of ‘polywater’—a peculiar form of water thought to have been observed in laboratories in the 1960's—to ordinary-water-containing-some impurities-from-improperly-washed-glassware contributed to the conclusion that there really is no such substance as polywater (1989: 161).

Thus, a non-analytic reduction may or may not have implications for the existence dimension of a realistic view of a particular subject matter. And even if the existence dimension is vindicated, there is still the further question whether the objects and properties vindicated are independent of anyone's beliefs, linguistic practices, and so on. Again, there is no straightforward relationship between the issue of reductionism and the issue of realism.

5. Against the Existence Dimension (III): Expressivism about Morals

We saw above that for the subject-matter in question the error-theorist *agrees* with the realist that the truth of the atomic, declarative sentences of that area requires the existence of the relevant type of objects, or the instantiation of the relevant sorts of properties. Although the realist and the error-theorist agree on this much, they of course disagree on the question of whether the relevant type of objects exist, or on whether the relevant sorts of properties are instantiated: the error-theorist claims that they don't, so that the atomic, declarative sentences of the area are systematically and uniformly false, the realist claims that at least in some instances the relevant objects exist or the relevant properties are instantiated, so that the atomic, declarative sentences of the area are at least in some instances true. We also saw that an error-

theory about a particular area could be motivated by epistemological worries (Field) or by a combination of epistemological and metaphysical worries (Mackie).

Another way in which the existence dimension of realism can be resisted is via expressivism about morals. Whereas the realist and the error-theorist agree that the sentences of the relevant area are *truth-apt*, apt to be assessed in terms of truth and falsity, the realist and the expressivist (alternatively non-cognitivist, projectivist) disagree about the truth-aptness of those sentences. It is a fact about English that sentences in the declarative mood ('The beer is in the fridge') are conventionally used for making assertions, and assertions are true or false depending on whether or not the fact that is asserted to obtain actually obtains. But there are other grammatical moods that are conventionally associated with different types of speech-act. For example, sentences in the imperatival mood ('Put the beer in the fridge') are conventionally used for giving orders, and sentences in the interrogative mood ('Is the beer in the fridge?') are conventionally used for asking questions. Note that we would not ordinarily think of orders or questions as even apt for assessment in terms of truth and falsity: they are not truth-apt. Now the conventions mentioned here are not exceptionless: for example, one can use sentences in the declarative mood ('My favourite drink is Belhaven 60 shilling') to give an order (for some Belhaven 60 shilling), one can use sentences in the interrogative mood ('Is the Pope a Catholic?') to make an assertion (of whatever fact was the subject of the discussion), and so on. The expressivist about a particular area will claim that the realist is misled by the syntax of the sentences of that area into thinking that they are truth-apt: she will say that this is a case where the conventional association of the declarative mood with assertoric force breaks down. 'Stealing is wrong' is no more truth-apt than 'Put the beer in the fridge': it is just that the truth-aptness of the latter is worn on its sleeve, while the lack of truth-aptness of the former is veiled by its surface syntax.

So, if moral sentences are not conventionally used for the making of assertions, what are they conventionally used for? According to one classical form of expressivism, *emotivism*, they are conventionally used for the expression of emotion, feeling, or sentiment. Thus, A.J. Ayer writes:

> If I say to someone, 'You acted wrongly in stealing that money', I am not stating anything more than if I had simply said, 'You stole that money'. In adding that this action is wrong, *I am not making any further statement about it. I am simply evincing my moral disapproval about it.* It is as if I had said, 'You stole that money', in a peculiar tone of horror, or written with the addition of some special exclamation marks. The tone, or the exclamation marks, adds nothing to the literal meaning of the sentence. *It merely serves to show that the expression of it is attended by certain feelings in the speaker* (Ayer 1946: 107, emphases added).

It follows from this that:

> If I now generalise my previous statement and say, 'Stealing money is wrong,' I produce a sentence which has no factual meaning—that is, expresses no proposition that can be either true or false (1946: 107).

Emotivism faces many problems, discussion of which is not possible here (for a survey, see Miller 2003a Ch.3). One problem that has been the bugbear of all expressivist versions of non-realism, the 'Frege-Geach Problem', is so-called because the classic modern formulation is by Peter Geach (1960), who attributes the original point to Frege.

According to emotivism, when I sincerely utter the sentence 'Murder is wrong' I am not expressing a belief or making an assertion, but rather expressing some non-cognitive sentiment or feeling, incapable of being true or false. Thus, the emotivist claims that in contexts where 'Murder is wrong' is apparently being used to assert that murder is wrong it is in fact being used to express a sentiment or feeling of disapproval towards murder. But what about contexts in which it is not even apparently the case that 'Murder is wrong' is being used to make an assertion? An example of such a sentence would be 'If murder is wrong, then getting little brother to murder people is wrong'. In the antecedent of this 'Murder is wrong' is clearly not even apparently being used to make an assertion. So what account can the emotivist give of the use of 'Murder is wrong' within 'unasserted contexts', such as the antecedent of the conditional above? Since it is not there used to express disapproval of murder, the account of its semantic function must be different from that given for the apparently straightforward assertion expressed by 'Murder is wrong'. But now there is a problem in accounting for the following apparently valid inference:

> (1) Murder is wrong.
>
> (2) If Murder is wrong, then getting your little brother to murder people is wrong.
>
> Therefore:
>
> (3) Getting your little brother to murder people is wrong.

If the semantic function of 'Murder is wrong' as it occurs within an asserted context in (1) is different from its semantic function as it occurs within an unasserted context in (2), isn't someone arguing in this way simply guilty of equivocation? In order for the argument to be valid, the occurrence of 'Murder is wrong' in (1) has to *mean the same thing* as the occurrence of 'Murder is wrong' in (2). But if 'Murder is wrong' has a different semantic function in (1) and (2), then it certainly doesn't mean the same thing in (1) and (2). So the above argument is apparently no more valid than:

> (4) My beer has a head on it.
>
> (5) If something has a head on it, then it must have eyes and ears.
>
> Therefore:
>
> (6) My beer must have eyes and ears.

This argument is obviously invalid, because it relies on an equivocation on two senses of 'head', in (4)

and (5) respectively.

It is perhaps worth stressing why the Frege-Geach problem doesn't afflict ethical theories which see ‘Murder is wrong’ as truth-apt, and sincere utterances of ‘Murder is wrong’ as capable of expressing straightforwardly truth-asessable beliefs. According to theories like these, moral *modus ponens* arguments such as the argument above from (1) and (2) to (3) are just like non-moral cases of *modus ponens* such as

> (7) It is raining;
>
> (8) If it is raining then the streets are wet;
>
> Therefore,
>
> (9) the streets are wet.

Why is this non-moral case of *modus ponens* not similarly invalid in virtue of the fact that ‘It is raining’ is asserted in (7), but not in (8)? The answer is of course that the state of affairs asserted to obtain by ‘It is raining’ in (7) is the same as that merely hypothesised to obtain in (8). In (7) ‘It is raining’ is used to assert that a state of affairs obtains (it's raining), and in (8) it is asserted that if that state of affairs obtains, so does another (the streets being wet). Throughout, the semantic function of the sentences concerned is given in terms of the states of affairs asserted to obtain in simple assertoric contexts. And it is difficult to see how an emotivist can say anything analogous to this with respect to the argument from (1) and (2) to (3): it is difficult to see how the semantic function of ‘Murder is wrong’ in the antecedent of (2) could be given in terms of the sentiment it allegedly expresses in (1).

The Frege-Geach challenge to the emotivist is thus to answer the following question: how can you give an emotivist account of the occurrence of moral sentences in ‘unasserted contexts’—such as the antecedents of conditionals—without jeopardising the intuitively valid patterns of inference in which those sentences figure? Philosophers wishing to develop an expressivistic alternative to moral realism have expended a great deal of energy and ingenuity in devising responses to this challenge. See in particular Blackburn's development of ‘quasi-realism’, in his (1984) Chs 5 and 6, (1993) Ch.10, (1998) Ch.3 and Gibbard's ‘norm-expressivism’, in his (1990) Ch.5. For criticism see Hale (1993) and (2002). For an overview, see Miller (2003), Chs 4 and 5.

6. Against the Independence Dimension (I): Semantic Realism

Challenges to the existence dimension of realism have been outlined in previous sections. In this section some forms of non-realism that are neither error-theoretic nor expressivist will be briefly introduced. The forms of non-realism view the sentences of the relevant area as (against the expressivist) truth-apt, and (against the error-theorist) at least sometimes true. The existence dimension of realism is thus left intact.

What is challenged is the independence dimension of realism, the claim that the objects distinctive of the area exist, or that the properties distinctive of the area are instantiated, independently of anyone's beliefs, linguistic practices, conceptual schemes, and so on.

Classically, opposition to the independence dimension of realism about the everyday world of macroscopic objects took the form of *idealism*, the view that the objects of the everyday world of macroscopic objects are in some sense *mental*. As Berkeley famously claimed, tables, chairs, cats, the moons of Jupiter and so on, are nothing but ideas in the minds of spirits:

> All the choir of heaven and furniture of the earth, in a word all those bodies which compose the mighty frame of the world, have not any subsistence without a mind (Berkeley 1710: §6).

Idealism has long been out of favour in contemporary philosophy, but those who doubt the independence dimension of realism have sought more sophisticated ways of opposing it. One such philosopher, Michael Dummett, has suggested that in some cases it may be appropriate to reject the independence dimension of realism via the rejection of semantic realism about the area in question (see Dummett 1978 and 1993). This section contains a brief explanation of semantic realism, as characterised by Dummett, Dummett's views on the relationship between semantic realism and realism construed as a metaphysical thesis, and an outline of some of the arguments in the philosophy of language that Dummett has suggested might be wielded against semantic realism.

It is easiest to characterise semantic realism for a mathematical domain. It is a feature of arithmetic that there are some arithmetical sentences for which the following holds true: we know of no method that will guarantee us a proof of the sentence, and we know of no method that will guarantee us a disproof or a counterexample either. One such is Goldbach's Conjecture:

> (G) Every even number is the sum of two primes.

It is possible that we may come across a proof, or a counterexample, but the key point is that we do not know a method, or methods, the application of which is guaranteed to yield one or the other. A semantic realist, in Dummett's sense, is one who holds that our understanding of a sentence like (G) consists in knowledge of its truth-condition, where the notion of truth involved is *potentially recognition-transcendent* or *bivalent*. To say that the notion of truth involved is potentially recognition-transcendent is to say that (G) may be true (or false) even though there is no guarantee that we will be able, in principle, to recognise that that is so. To say that the notion of truth involved is bivalent is to accept the unrestricted applicability of the law of bivalence, that every meaningful sentence is determinately either true or false. Thus the semantic realist is prepared to assert that (G) is determinately either true or false, regardless of the fact that we have no guaranteed method of ascertaining which. (Note that the precise relationship between the characterisation in terms of bivalence and that in terms of potentially recognition-transcendent truth is a delicate matter that will not concern us here. See the Introduction to Wright 1993 for some excellent discussion. It is also important to note that in introducing the idea that a speaker's understanding of a sentence consists in her knowledge of its truth-condition, Dummett is packing more

into the notion of truth than the disquotational properties made use of in §1 above. See Dummett's essay 'Truth', in his 1978).

Dummett makes two main claims about semantic realism. First, there is what Devitt (1991a) has termed the *metaphor thesis*: This denies that we can even *have* a literal, austerely metaphysical characterisation of realism of the sort attempted above with Generic Realism. Dummett writes, of the attempt to give an austere metaphysical characterisation of realism about mathematics (platonic realism) and what stands opposed to it (intuitonism):

> How [are] we to decide this dispute over the ontological status of mathematical objects[?] As I have remarked, we have here two metaphors: the platonist compares the mathematician with the astronomer, the geographer or the explorer, the intuitionist compares him with the sculptor or the imaginative writer; and neither comparison seems very apt. The disagreement evidently relates to the amount of freedom that the mathematician has. Put this way, however, both seem partly right and partly wrong: the mathematician has great freedom in devising the concepts he introduces and in delineating the structure he chooses to study, but he cannot prove just whatever he decides it would be attractive to prove. How are we to make the disagreement into a definite one, and how can we then resolve it? (1978: xxv).

According to the *constitution thesis*, the literal content of realism *consists in* the content of semantic realism. Thus, the literal content of realism about the external world is constituted by the claim that our understanding of at least some sentences concerning the external world consists in our grasp of their potentially recognition-transcendent truth-conditions. The spurious 'debate' in metaphysics between realism and non-realism can thus become a genuine *debate* within the theory of meaning: should we characterise speakers' understanding in terms of grasp of potentially recognition-transcendent truth-conditions? As Dummett puts it:

> The dispute [between realism and its opponents] concerns the notion of truth appropriate for statements of the disputed class; and this means that it is a dispute concerning the kind of *meaning* which these statements have (1978: 146).

Few have been convinced by either the metaphor thesis or the constitution thesis. Consider Generic Realism in the case of the world of everyday macroscopic objects and properties:

> (GR1) Tables, rocks, mountains, seas, and so on exist, and the fact that they exist and have properties such as mass, size, shape,colour, and so on, is (apart from mundane empirical dependencies of the sort sometimes encountered in everyday life) independent of anyone's beliefs, linguistic practices, conceptual schemes, and so on.

Dummett may well call for some non-metaphorical characterisation of the independence claim which this involves, but it is relatively easy to provide one such characterisation by utilising Dummett's own notion

of recognition-transcendence:

> (GR2) Tables, rocks, mountains, seas, and so on exist, and the fact that they exist and have properties such as mass, size, shape,colour, and so on, is (apart from mundane empirical dependencies of the sort sometimes encountered in everyday life) independent of anyone's beliefs, linguistic practices, conceptual schemes, and so on. Tables, rocks, mountains, seas, and so on exist, and in general there is no guarantee that we will be able, even in principle, to recognise the fact that they exist and have properties such as mass, size, shape, colour, and so on.

On the face of it, there is nothing metaphorical in (GR2) or, at least if there is, some argument from Dummett to that effect is required. This throws some doubt on the metaphor thesis. And there is nothing distinctively semantic about (GR2), and this throws some doubt on the constitution thesis. Whereas for Dummett, the essential realist thesis is the meaning-theoretic claim that our understanding of a sentence like (G) consists in knowledge of its potentially recognition-transcendent truth-condition, for Devitt:

What has truth to do with Realism? On the face of it, nothing at all. Indeed, "Realism says nothing semantic at all beyond … making the negative point that our semantic capacities do *not* constitute the world." (1991a: 39)

Devitt's main criticism of the constitution thesis is this: the literal content of realism about the external world is not given by semantic realism, since semantic realism is consistent with an *idealist* metaphysics of the external world. He writes:

> Does [semantic realism] entail Realism? It does not. Realism … requires the objective independent existence of common-sense physical entities. Semantic Realism concerns physical *statements* and has no such requirement: *it says nothing about the nature of the reality that makes those statements true or false*, except that it is [at least in part potentially beyond the reach of our best investigative efforts]. An idealist who believed in the … existence of a purely mental realm of sense-data could subscribe to [semantic realism]. He could believe that physical statements are true or false according as they do or do not correspond to the realm of sense-data, whatever anyone's opinion on the matter: we have no 'incorrigible knowledge' of sense-data. … In sum, mere talk of truth will not yield any particular ontology. (1983: 77)

Suppose that Dummett's metaphor and constitution theses are both implausible. Would it follow that the arguments Dummett develops against semantic realism have *no* relevance to debates about the plausibility of realism about everyday macroscopic objects (say), construed as a purely metaphysical thesis as in (GR2)? Dummett's arguments can retain their relevance to metaphysical debate even if the metaphor and constitution theses are false, and, indeed, even if Dummett's view (1973: 669) that the theory of meaning is the foundation of all philosophy is rejected. For a full development of this line of argument, see Miller 2003b.

Suppose that we are considering some region of discourse *D*, the sentences of which we intuitively understand. Suppose, for *reductio*, that the sentences of *D* have potentially recognition-transcendent truth-conditions. Thus,

> (1) We understand the sentences of *D*.

> (2) The sentences of *D* have recognition-transcendent truth-conditions.

Now, from (1) together with the Fregean thesis that to understand a sentence is to know its truth-conditions (see Miller 1998, Chs 1 and 2), we have:

> (3) We know the truth-conditions of the sentences of *D*.

We now add the apparently reasonable constraint on ascriptions of knowledge:

> (4) If a piece of knowledge is ascribed to a speaker, then it must be at least in principle possible for that speaker to have *acquired* that knowledge.

So,

> (5) It must be at least in principle possible for us to have acquired knowledge of the recognition-transcendent truth-conditions of *D*.

But,

> (6) There is no plausible story to be told about how we could have acquired knowledge of recognition-transcendent truth-conditions.

So, by *reductio*, we reject (2) to get:

> (7) The sentences of *D* do not have recognition-transcendent truth-conditions, so semantic realism about the subject matter of *D* must be rejected.

The crucial premise here is obviously (6). Wright puts the point as follows:

How are we supposed to be able to *form* any understanding of what it is for a particular statement to be true if the kind of state of affairs which it would take to make it true is conceived, *ex hypothesi*, as something beyond our experience, something which we cannot confirm and which is insulated from any distinctive impact on our consciousness?(1993: 13).

However, Wright then more or less concedes that the acquisition argument can be neutralised by invoking

the *compositionality* of meaning and understanding:

> [T]he realist seems to have a very simple answer. Given that the understanding of statements in general is to be viewed as consisting in possession of a concept of their truth-conditions, acquiring a concept of an evidence-transcendent state of affairs is simply a matter of acquiring an understanding of a statement for which that state of affairs would constitute the truth-condition. And such an understanding is acquired, like the understanding of any previously unheard sentence in the language, by understanding the constituent words and the significance of their mode of combination. (1993: 16)

Dummett's challenge to semantic realism, then, turns on his second argument, the *manifestation argument*. Suppose that we are considering region of discourse *D* as before. Then:

> (1) We understand the sentences of *D*.

Suppose, for *reductio*, that

> (2) The sentences of *D* have recognition-transcendent truth-conditions.

From (1) and and the Fregean thesis that to understand a sentence is to know its truth-conditions, we have:

> (3) We know the truth-conditions of the sentences of *D*.

We then add the following premise, which stems from the Wittgensteinian insight that understanding does not consist in the possession of an inner state, but rather in the possession of some practical ability (see Wittgenstein 1958):

> (4) If speakers possess a piece of knowledge which is constitutive of linguistic understanding, then that knowledge should be *manifested* in speakers' use of the language i.e. in their exercise of the practical abilities which constitute linguistic understanding.

It now follows from (1), (2) and (3) that:

> (5) Our knowledge of the recognition-transcendent truth-conditions of the sentences of *D* should be manifested in our use of those sentences, i.e. in our exercise of the practical abilities which constitute our understanding of *D*.
>
> Since
>
> (6) Such knowledge is never manifested in the exercise of the practical abilities which constitute our understanding of *D*,

It follows that

> (7) We do not possess knowledge of the truth-conditions of *D*.

(7) and (3) together give us a contradiction, whence, by *reductio*, we reject (2) to obtain:

> (8) The sentences of *D* do not have recognition-transcendent truth-conditions, so semantic realism about the subject matter of *D* must be rejected.

The basic point is that, so far as an account of speakers' understanding goes, the ascription of knowledge of recognition-transcendent truth-conditions is simply *redundant*: there is no good reason for ascribing it. Consider one of the sentences introduced earlier as a candidate for possessing recognition-transcendent truth-conditions ‘Every even number greater than two is the sum of two primes’. The semantic realist views our understanding of sentences like this as consisting in our knowledge of a potentially recognition-transcendent truth-condition. But:

How can that account be viewed as a description of any *practical* ability of use? No doubt someone who understands such a statement can be expected to have many relevant practical abilities. He will be able to appraise evidence for or against it, should any be available, or to recognize that no information in his possession bears on it. He will be able to recognize at least some of its logical consequences, and to identify beliefs from which commitment to it would follow. And he will, presumably, show himself sensitive to conditions under which it is appropriate to ascribe propositional attitudes embedding the statement to himself and to others, and sensitive to the explanatory significance of such ascriptions. In short: in these and perhaps other important respects, he will show himself competent to use the sentence. But the headings under which his practical abilities fall so far involve no mention of evidence-transcendent truth-conditions (1993: 17).

This establishes (6), and the conclusion follows swiftly.

A detailed assessment of the plausibility of Dummett's arguments is impossible here. For a full response to the manifestation argument, see Miller 2003c. For the acquisition argument, see Miller 2003d. Wright develops a couple of additional arguments against semantic realism. For these—the argument from rule-following and the argument from normativity—see the Introduction to Wright 1993. For an excellent survey of the literature on Dummett's arguments against semantic realism, see Hale 1997.

7. Against the Independence Dimension (II): More Forms of Anti-Realism

Suppose that one wished to develop a non-realist alternative to, say, moral realism. Suppose also that one is persuaded of the unattractiveness of both error-theoretic and expressivist forms of non-realism. That is

to say, one accepts that moral sentences are truth-apt, and, at least in some cases, true. Then the only option available would be to deny the independence dimension of moral realism. But so far we have only seen one way of doing this: by admitting that the relevant sentences are truth-apt, sometimes true, and possessed of truth-conditions which are not potentially recognition-transcendent. But this seems weak: it seems implausible to suggest that a moral realist must be committed to the potential recognition-transcendence of moral truth. It therefore seems implausible to suggest that a non-expressivistic and non-error-theoretic form of opposition to realism must be committed to simply denying the potential recognition-transcendence of moral truth, since many who style themselves moral realists will deny this too. As Wright puts it:

> There are, no doubt, kinds of moral realism which do have the consequence that moral reality may transcend all possibility of detection. But it is surely not essential to any view worth regarding as realist about morals that it incorporate a commitment to that idea. (1992: 9)

So, if the debate between a realist and a non-realist about the independence dimension doesn't concern the plausibility of semantic realism as characterised by Dummett, what does it concern? (Henceforth a non-error-theoretic, non-expressivist style of non-realist is referred to as an anti-realist). Wright attempts to develop some points of contention, (or ‘realism-relevant cruces’ as he calls them) over which a realist and anti-realist could disagree. Wright's development of this idea is subtle and sophisticated and only a crude exposition of a couple of his realism-relevant Cruces can be given here.

The first of Wright's realism-relevant Cruces to be considered here concerns the capacity of states of affairs to figure ineliminably in the explanation of features of our experience. The idea that the explanatory efficacy of the states of affairs in some area has something to do with the plausibility of a realist view of that area is familiar from the debates in meta-ethics between philosophers such as Nicholas Sturgeon (1988), who believe that irreducibly moral states of affairs do figure ineliminably in the best explanation of certain aspects of experience, and opponents such as Gilbert Harman (1977), who believe that moral states of affairs have no such explanatory role. This suggests a ‘best explanation test’ which, crudely put, states that realism about a subject matter can be secured if its distinctive states of affairs figure ineliminably in the best explanation of aspects of experience. One could then be a non-expressivist, non-error-theoretic, anti-realist about a particular subject matter by denying that the distinctive states of that subject do have a genuine role in best explanations of aspects of our experience. And the debate between this style of anti-realist and his realist opponent could proceed independently of any questions concerning the capacity of sentences in the relevant area to have potentially recognition-transcendent truth values.

For reasons that needn't detain us here, Wright suggests that this ‘best explanation test’ should be superseded by questions concerning what he calls *width of cosmological role* (1992, Ch.5). The states of affairs in a given area have narrow cosmological role if they cannot contribute to the explanation of things *other* than our beliefs about that subject-matter (or other than *via* explaining our beliefs about that subject matter). This will be an anti-realist position. One style of realist about that subject matter will say that its states of affairs have wide cosmological role: they do contribute to the explanation of things other than

our beliefs about the subject matter in question (or other than via explaining our beliefs about that subject matter). It is relatively easy to see why width of cosmological role could be a bone of contention between realist and anti-realist views of a given subject matter: it is precisely the width of cosmological role of a class of states of affairs—their capacity to explain things other than, or other than via, our beliefs, in which their independence from our beliefs, linguistic practices, and so on, consists. Again, the debate between someone attributing a narrow cosmological role to a class of states of affairs and someone attributing a wide cosmological role could proceed independently of any questions concerning the capacity of sentences in the relevant area to have potentially recognition-transcendent truth values.

Wright thinks that it is arguable that moral discourse does not satisfy width-of-cosmological role. Whereas a physical fact—such as a pond's being frozen over—can contribute to the explanation of *cognitive effects* (someone's believing that the pond is frozen over), *effects on sentient, but non-conceptual creatures* (the tendency of goldfish to cluster towards the bottom of the pond), *effects on us as physically interactive agents* (someone's slipping on the ice), and *effects on inanimate matter* (the tendency of a thermometer to read zero when placed on the surface), moral facts seem only to contribute to the explanation of the first sort of effect:

> [I]t is hard to think of anything which is true of sentient but non-conceptual creatures, or of mobile organisms, or of inanimate matter, which is true because a … moral fact obtains and in whose explanation it is unnecessary to advert to anyone's appreciation of that moral fact (1996: 16).

Thus, we have a version of anti-realism about morals that is non-expressivist and non-error-theoretic and can be framed independently of considerations about the potential of moral sentences to have recognition-transcendent truth-values: moral sentences are truth-apt, sometimes true, and moral states of affairs have narrow cosmological role.

The second of Wright's realism-relevant Cruces to be considered involves the notion of judgement-dependence. Suppose that we are considering a particular region of discourse *D* in which ‘*P*’ is a representative central predicate. Consider the opinions formed by the practitioners of that discourse, formed under conditions which are, for that discourse, cognitively ideal: call such opinions *best opinions*, and the cognitively ideal conditions the *C-conditions.* Suppose we find that the best opinions formed by the practitioners covary with the facts about the instantiation of ‘*P*’. Then, Wright suggests, there are two ways in which we might seek to explain this covariance. On the one hand, we might take best opinions to be playing at most a *tracking* role: best opinions are just extremely good at tracking independently constituted truth-conferring states of affairs. In such a case, best opinion plays merely an *extension-reflecting* role, serving merely to reflect the independently determined extensions of the central predicates of *D* (or equivalently, the independently determined extension of the truth-predicate applicable in *D*). On the other hand, we might try to explain the covariance of best opinion and fact by assigning to best opinion an altogether different sort of role. Rather than viewing best opinion as merely tracking the facts about the extensions of the central predicates of *D*, we can view them as themselves *determining* those very extensions. Best opinion, on this sort of view, does not serve merely to track independently constituted states of affairs which determine the extensions of the central predicates of *D*: rather, best

opinion serves to determine those extensions and so to play an *extension-determining* role. When the covariance of best opinion and the facts about the instantiation of the central predicates of a region of discourse admits of this latter sort of explanation, the predicates of that region are said to be *judgement-dependent*; when it admits only of the former sort of explanation, the predicates are said to be *judgement-independent*.

How do we determine whether the central predicates of a region of discourse are judgement-dependent? Wright's discussion proceeds by reference to what he terms *provisional equations*. These have the following form:

> (PE) $\forall x[C \rightarrow (\text{A suitable subject } s \text{ judges that } Px \leftrightarrow Px)]$

where '*C*' denotes the conditions (the *C*-conditions) which are cognitively ideal for forming the judgement that *x* is *P*. The predicate 'P' is then said to be judgement-dependent if and only if the provisional equation meets the following four conditions:

> *The A Prioricity Condition:* The provisional equation must be *a priori* true: there must be *a priori* covariance of best opinions and truth. (Justification: 'the truth, if it is true, that the extensions of [a class of concept] are constrained by idealised human response—best opinion—ought to be available purely by analytic reflection on those concepts, and hence available as knowledge *a priori*'(Wright 1992: 117)). This is because the thesis of judgement-dependence is the claim that, for the region of discourse concerned, best opinion is the *conceptual ground* of truth).
>
> *The Substantiality Condition* The *C*-conditions must be specifiable *non-trivially*: they cannot simply be described as conditions under which the subject has 'whatever it takes' to form the right opinion concerning the subject matter at hand.(Justification: without this condition, *any* predicate will turn out to be judgement-dependent, since for any predicate *Q* it is going to be an *a priori* truth that our judgements about whether *x* is *Q*, formed under conditions which have 'whatever it takes' to ensure their correctness, will covary with the facts about the instantiation of *Q*-ness. We thus require this condition on pain of losing the distinction between judgement-dependent and judgement-independent predicates altogether).
>
> *The Independence Condition*: The question as to whether the *C*-conditions obtain in a given instance must be logically independent of the class of truths for which we are attempting to give an extension-determining account: for specifying what makes an opinion best must not presuppose some logically prior determination of the extensions putatively determined by best opinions. (Justification: if we have to assume, say, certain facts about the extension of *P* in the specification of the conditions under which opinions about *P* count as best, then we cannot view best opinions as somehow constituting those facts, since specifying whether a given opinion is best would then presuppose some logically *prior* determination of the very facts allegedly constituted by best opinions).

> *The Extremal Condition*: There must be no better way of accounting for the *a priori* covariance: no better account, other than according best opinion an extension-determining role, of which the satisfaction of the foregoing three conditions is a consequence. (Justification: without this condition, the satisfaction of the foregoing conditions would be consistent with the thought that certain states of affairs are judgement-independent even though infallibly detectable,'states of affairs in whose determination facts about the deliverances of best opinions are in no way implicated although there is, *a priori,* no possibility of their misrepresentation' (Wright 1992: 123)).

When all of the above conditions can be shown to be satisfied, we can accord best opinion an extension-determining role, and describe the subject matter as judgement-dependent. If these conditions cannot collectively be satisfied, best opinion can be assigned, at best, a merely extension-reflecting role.

Two points are worth making. First, it is again relatively easy to see why the question of judgement-dependence can mark a bone of contention between realism and anti-realism. If a subject matter is judgement-dependent we have a concrete sense in which the independence dimension of realism fails for that subject matter: there is a sense in which that subject matter is not entirely independent of our beliefs, linguistic practices, and so on. Second, the debate about the judgement-dependence of a subject matter is, on the face of it at least, independent of the debate about the possibility of recognition-transcendent truth in that area.

Wright argues (1989) that facts about colours and intentions are judgement-dependent, so that we can formulate a version of anti-realism about colours (intentions) that views ascriptions of colours (intentions) as truth-apt and sometimes true, and truth in those areas as judgement-dependent. In contrast to this, Wright argues (1988) that morals cannot plausibly be viewed as judgement-dependent, so that a thesis of judgement-dependence is not a suitable vehicle for the expression of a non-expressivistic, non-error-theoretic, version of anti-realism about morality.

For discussion of further allegedly realism-relevant Cruces, such as *cognitive command*, see Wright 1992. For critical discussion of Wright on cognitive command, see Shapiro and Taschek 1996.

8. Undermining the Debate: Quietism

Some of the ways in which non-realist theses about a particular subject matter can be formulated and motivated have been described above. Quietism is the view that significant metaphysical debate between realism and non-realism is impossible. Gideon Rosen nicely articulates the basic quietist thought:

> We *sense* that there is a heady metaphysical thesis at stake in these debates over realism—a question on a par with the issues Kant first raised about the status of nature. But after a point, when every attempt to say just what the issue is has come up empty, we have no real

> choice but to conclude that despite all the wonderful, suggestive imagery, there is ultimately nothing in the neighborhood to discuss (1994: 279).

Quietism about the ‘debate’ between realists and their opponents can take a number of forms. One form might claim that the idea of a significant debate is generated by unsupported or unsupportable philosophical theses about the relationship of the experiencing and minded subject to their world, and that once these theses are exorcised the ‘debate’ will gradually wither away. This form of quietism is often associated with the work of the later Wittgenstein, and receives perhaps its most forceful development in the work of John McDowell (see in particular McDowell 1994). Other forms of quietism may proceed in a more piecemeal fashion, taking constraints such as Wright's realism-relevant Cruces and arguing on a case-by-case basis that their satisfaction or non-satisfaction is of no metaphysical consequence. This is in fact the strategy pursued in Rosen 1994. He makes the following remarks regarding the two realism-relevant Cruces considered in the previous section.

Suppose that:

> (F) It is *a priori* that: *x* is funny if and only if we would judge *x* funny under conditions of full information about *x*s relevant extra-comedic features

and suppose that (F) satisfies (in addition to a prioricity) the various other constraints that Wright imposes on his provisional equations ((F) is actually not of the form of a provisional equation, but this is not relevant to our purposes here). Rosen questions whether this would be enough to establish that the facts about the funny are in some metaphysically interesting sense ‘less real’ or ‘less objective’ than facts (such as, arguably, facts about shape) for which a suitable equation cannot be constructed.

In a nutshell, Rosen's argument proceeds by inviting us to assume the perspective of an anthropologist who is studying us and who ‘has gotten to the point where he can reliably determine which jokes we will judge funny under conditions of full relevant information’(1994: 302). Rosen writes:

> [T]he important point is that from [the anthropologist's] point of view, the facts about the distribution of [the property denoted by our use of ‘funny’] are ‘mind-dependent’ only in the sense that they supervene directly on facts about our minds. But again, this has no tendency to undermine their objectivity … [since] we have been given no reason to think that the facts about what a certain group of people would think after a certain sort of investigation are anything but robustly objective (1994: composed from 300 and 302).

How plausible is this attempt to deflate the significance of the discovery that the subject matter of a particular area is, in Wright's sense, judgement-dependent? Argument—as opposed to the trading of intuitions—at this level is difficult, but Rosen's claim here is very implausible. Suppose we found out that facts about the distribution of gases on the moons of Jupiter supervened directly on facts about our minds. Would the threat we then felt to the objectivity of facts about the distribution of gases on the moons of Jupiter be at all assuaged by the reflection that facts about the mental might themselves be susceptible to

realistic treatment? It seems doubtful. Fodor's *Psychosemantics* would not offer much solace to realists in the world described in Berkeley's *Principles*. Rosen's claim derives some of its plausibility from the fact that he uses examples, such as the funny and the constitutional, where our pre-theoretical attachment to a realist view is very weak: it may be that the judgement-dependence of the funny doesn't undermine our sense of the objectivity of humour simply because the level of objectivity we pretheoretically expect of comedy is quite low. So although there is no knock-down argument to Rosen's claim, it is much more counterintuitive than he would be willing to admit.

Rosen also questions whether there is any intuitive connection between considerations of width of cosmological role and issues of realism and non-realism. Rosen doubts in particular that there is any tight connection between facts of a certain class having only narrow cosmological role and mind-dependence in any sense relevant to the plausibility of realism. He writes:

> It is possible to imagine a subtle physical property *Q* which, though intuitively thoroughly objective, is nonetheless nomically connected in the first instance only with brain state *B*—where this happens to be the belief that things are *Q*. This peculiar discovery would not undermine our confidence that *Q* was an objective feature of things, as it should if [a feature of objects is less than fully objective if it has narrow cosmological role] (1994: 312).

It seems to me that, at least in the first instance, Wright has a relatively quick response to this point at his disposal. Waiving the point that in any case the width of cosmological role constraint applies to *classes* of properties and facts, he can point out that in the example constructed by Rosen the narrowness of Q's cosmological role is an *a posteriori* matter. Whereas what we want is that the narrowness of cosmological role is an *a priori* matter: one does not need to conduct an empirical investigation to convince oneself that facts about the funny fail to have wide cosmological role.

Wright thus has the beginnings of answers to Rosen's quietist attack on his use of the notions of judgement-dependence and width of cosmological role. It is not possible to deal fully with these arguments here, let alone with the other quietist arguments in Rosen's paper, or the arguments of other quietists such as McDowell, beyond giving a flavour of how quietism might be motivated and how those active in the debates between realists and their opponents might start to respond.

9. Concluding Remarks and Apologies

This discussion of realism and of the forms that non-realist opposition may take is far from exhaustive, and aims only to give the reader a sense of what to expect if they delve deeper into the issues. In particular, nothing has been mentioned about the work of Hilary Putnam, his characterisation of 'metaphysical realism', and his so-called 'model-theoretic' argument against it. Putnam's writings are extensive, but one could begin with Putnam 1981 and 1983. For critical discussion, see Hale and Wright 1997. Nor have issues about the metaphysics of modality and possible worlds been discussed. The *locus classicus* in this area is Lewis 1986. For commentary, see Divers 2002. And the very important topic of

scientific realism has not been touched upon. For an introductory treatment and suggestions for further reading, see Bird 1998 Ch. 4. Finally, it has not been possible to include any discussion of realism about intentionality and meaning. The *locus classicus* in recent philosophy is Kripke 1982. For a robustly realistic view of the intentional, see Fodor 1987. For a collection of some of the central secondary literature, see Miller and Wright 2002.

Bibliography

- Alston, W., (1958) “Ontological Commitment,” *Philosophical Studies* 9, pp.8-17.
- Ayer, A.J., (1946) *Language, Truth, and Logic* (New York: Dover Publications, 2nd Edition).
- Benacerraf, P., (1965) “What Numbers Could Not Be,” *Philosophical Review* 74, pp.47-73.
- -----, (1973) “Mathematical Truth,” *Journal of Philosophy* 70, pp.661-679.
- Berkeley, G., (1710) *The Principles of Human Knowledge* (many editions).
- Bird, A., (1998) *The Philosophy of Science* (London: UCL Press).
- Blackburn, S., (1984) *Spreading The Word* (Oxford: Oxford University Press).
- -----, (1993) *Essays in Quasi-Realism* (Oxford: Oxford University Press).
- -----, (1998) *Ruling Passions* (Oxford: Oxford University Press).
- Devitt, M., (1983) “Dummett's Anti-Realism,” *Journal of Philosophy* 80, pp.73-99.
- -----, (1991a) *Realism and Truth* (Princeton: Princeton University Press, 2nd Edition).
- -----, (1991b) “Aberrations of the Realism Debate,” *Philosophical Studies* 61, pp.43-63.
- Divers, J., (2002) *Possible Worlds* (London: Routledge forthcoming).
- Divers, J. and Miller, A., (1999) “Arithmetical Platonism: Reliability and Judgement-Dependence,” *Philosophical Studies* 95, pp.277-310.
- Dummett, M., (1973) *Frege: Philosophy of Language* (London: Duckworth).
- -----, (1978) *Truth and Other Enigmas* (London: Duckworth).
- -----, (1993) *The Seas of Language* (Oxford: Oxford University Press).
- Field, H., (1980) *Science Without Numbers* (Oxford: Blackwell).
- -----, (1989) *Realism, Mathematics, and Modality* (Oxford: Blackwell).
- Fodor, J., (1987) *Psychosemantics* (Cambridge, MA: MIT Press).
- Geach, P., (1960) “Ascriptivism,” *Philosophical Review* 69, pp.221-225.
- Gibbard, A., (1990) *Wise Choices, Apt Feelings* (Oxford: Clarendon Press).
- Godel, K., (1983) “What is Cantor's Continuum Problem?” in P. Benacerraf and H. Putnam (eds) *Philosophy of Mathematics: Selected Readings* (Cambridge: Cambridge University Press), pp.470-485.
- Hale, B., (1987) *Abstract Objects* (Oxford: Blackwell).
- -----, (1993) “Can There Be a Logic of Attitudes?” in J. Haldane and C. Wright (eds) *Reality, Representation, and Projection* (Oxford: Oxford University Press), pp.337-363.
- -----, (1994) “Is Platonism Epistemologically Bankrupt?” *Philosophical Review* 103, pp.299-325. Reprinted in Hale and Wright 2001.
- -----, (1997) “Realism and its Oppositions,” in B. Hale and C. Wright (eds) *A Companion to the Philosophy of Language* (Oxford: Blackwell), pp.271-308.
- -----, (2002) “Can Arboreal Knotwork Help Blackburn Out of Frege's Abyss?’, *Philosophy and*

Phenomenological Research, forthcoming.
- Hale, B. and Wright, C., (1997) “Putnam's Model-Theoretic Argument Against Metaphysical Realism,” in B. Hale and C. Wright (eds) *A Companion to the Philosophy of Language* (Oxford: Blackwell), pp.427-457.
- Hale, B. and Wright, C., (2001) *The Reason's Proper Study* (Oxford: Oxford University Press).
- Harman, G., (1977) *The Nature of Morality* (Oxford: Oxford University Press).
- Kripke, S., (1982) *Wittgenstein on Rules and Private Language* (Oxford: Blackwell).
- Lowe, E.J., (2002) *A Survey of Metaphysics* (Oxford: Oxford University Press).
- McDowell, J., (1994) *Mind and World* (Cambridge, MA: Harvard University Press).
- -----, (1998a) *Mind, Value, and Reality* (Cambridge, MA: Harvard University Press).
- Mackie, J. L., (1977) *Ethics: Inventing Right and Wrong* (Harmondsworth: Penguin).
- Mellor, D.H and Oliver, A., (1997) *Properties* (Oxford: Oxford University Press).
- Miller, A., (1998) *Philosophy of Language* (Montreal: McGill-Queens University Press)
- -----, (2002) “Wright's Argument Against Error-Theories,” *Analysis* 62, pp.98-103.
- -----, (2003a) *An Introduction To Contemporary Metaethics* (Cambridge: Polity Press).
- -----, (2003b) “The Significance of Semantic Realism,” *Synthese* (forthcoming).
- -----, (2003c) “What is the Manifestation Argument?” *Pacific Philosophical Quarterly* (forthcoming).
- -----, (2003d) “What is the Acquisition Argument?” in A. Barber (ed) *Epistemology of Language* (Oxford: Oxford University Press, forthcoming).
- Miller, A. and Wright, C. eds (2002) *Rule-Following and Meaning* (London: Acumen).
- Pettit, P., (1991) ‘Realism and Response-Dependence’, *Mind* 100, pp.587-626.
- Putnam, H., (1981) *Reason, Truth and History* (Cambridge: Cambridge University Press).
- -----, (1983) *Realism and Reason* (Cambridge: Cambridge University Press).
- Railton, P., (1986) “Moral Realism,” *Philosophical Review* 95, pp.163-207.
- -----, (1989) “Naturalism and Prescriptivity,” *Social Philosophy and Policy* 7, pp.151-174.
- Rosen, G., (1994) “Objectivity and Modern Idealism: What is the Question?”, in M. Michael and J. O'Leary-Hawthorne (eds) *Philosophy in Mind: The Place of Philosophy in the Study of Mind* (Dordrecht: Kluwer Academic Publishers), pp.277-319.
- Shapiro, S., and Taschek, W., (1996) “Intuitionism, Pluralism, and Cognitive Command,” *Journal of Philosophy* 93, pp.74-88.
- Smith, M., (1994) *The Moral Problem* (Oxford: Blackwell).
- Sturgeon, N., (1988) “Moral Explanations,” in G. Sayre-McCord (ed) *Essays on Moral Realism* (Ithaca: Cornell University Press), pp.229-255.
- Wittgenstein, L., (1958) *Philosophical Investigations* (Oxford: Blackwell).
- Wright, C., (1983) *Frege's Conception of Numbers as Objects* (Aberdeen: Aberdeen University Press).
- -----, (1988) “Moral Values, Projection, and Secondary Qualities,” *Proceedings of the Aristotelian Society* Supplementary Volume, pp.1-26.
- -----, (1989) “Meaning and Intention as Judgement-Dependent,” reprinted in Miller and Wright (op. cit)., pp.129-140.
- -----, (1992) *Truth and Objectivity* (Cambridge, MA: Harvard University Press).
- -----, (1993) *Realism, Meaning, and Truth* (Oxford: Blackwell, 2nd Edition).

- -----, (1996) “Truth in Ethics,” in B. Hooker (ed) *Truth in Ethics* (Oxford: Blackwell), pp.1-18.

Other Internet Resources

- The Metaphysics and Epistemology of Modality page maintained at Arché Centre for the philosophy of logic, language, mathematics and mind (University of St. Andrews)

Related Entries

cognitivism vs. non-cognitivism, moral | dependence, ontological | fictionalism | fictionalism: modal | metaethics | moral realism | possible worlds | realism: semantic challenges to | relativism

Alexander Miller
millera1@cardiff.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 8, 2002
Content last modified: July 8, 2002

Modal Fictionalism

Questions about necessity (or what has to be, or what cannot be otherwise) and possibility (or what can be, or what could be otherwise) are questions about *modality*. *Fictionalism* is an approach to theoretical matters in a given area which treats the claims in that area as being in some sense analogous to fictional claims: claims we do not literally accept at face value, but which we nevertheless think serve some useful function. However, despite its name, "Modal Fictionalism" in its usual manifestations is not primarily fictionalism about claims of necessity and possibility, but rather a fictionalist approach to claims about possible worlds. (For instance, modal fictionalism is not normally fictionalist about the claim that "it is possible that there be a species of tail-less kangaroo", but rather about the claim that "there is a possible world in which there is a species of tail-less kangaroo".) The practice of taking possible worlds to be merely convenient fictions, or of treating talk about possible worlds as being useful without being literally correct, is quite common in philosophical circles. It is only recently, however, that philosophers have seriously examined the implications of taking possible worlds to be merely fictional objects, like Sherlock Holmes or a frictionless surface.

Theories employing possible worlds terminology have been found to be very useful in philosophy, e.g. when engaging in thought experiments; distinguishing various claims in metaphysics, or in the philosophy of language, mind, knowledge or ethics; and in areas other than philosophy, like linguistics, modal logic, and probability theory. Many have found the status of these worlds and their contents to be puzzling, to say the least. What are they? Where, if anywhere, are they supposed to be? How are we supposed to discover facts about them? Isn’t it extravagant to believe that just because a situation is possible, it must in some sense exist? Modal fictionalists take theories committed to the existence of possible worlds, merely hypothetical situations, non-actual but possible objects etc. to be strictly and literally false, and so they avoid the problems of believing in possible worlds. Nevertheless, they claim, they can enjoy the benefits of using these seemingly problematic theories.

Modal fictionalism should be of interest to those concerned with the metaphysics of modality, since theories committed to the literal existence of possible worlds (and, even more worryingly, the literal existence of merely possible objects ‘contained’ in these worlds) come at a cost, both to economy and to many people’s intuitions. But it is, or should be, of wider interest as well, since it is one of the most discussed applications of a fictionalist treatment of abstract objects, along with mathematical fictionalism. Lessons learned in the case of modal fictionalism can hopefully be applied to other areas in which we may wish to evade literal theoretical commitments.

I shall begin by discussing the motivation for modal fictionalism, and distinguishing some of its varieties.

Next, I shall seek to put fictionalism in a slightly broader theoretical context, by discussing its connections with instrumentalism and eliminativism, and by discussing what connection there might be between "fictionalism" and treatments of paradigmatic fictions. I shall then discuss the debate about the "Brock/Rosen objection" and a problem raised by Bob Hale, both of which turn on technical problems concerning modal claims about the status of the modal fiction. Finally, in section 4, other concerns about modal fictionalism will be discussed.

---

1. Types of Modal Fictionalism

Modal fictionalists often focus on the claim that possible worlds are merely fictional entities, and apparent commitment to possible worlds is to be explained in the same sort of way that apparent commitment to ideal gases or frictionless surfaces is to be explained. Rosen 1990 and others have formulated modal fictionalism as a theory that takes talk of possible worlds to be on a par with talk about paradigmatically fictional objects, e.g. Sherlock Holmes ("There is a (non-actual) possible world at which there are blue swans" is to be understood on the model of "There is a brilliant detective at 221b Baker Street", in Rosen's example). This goes with an at least partial account of how we are to treat paradigmatically fictional claims: that they are, literally and strictly speaking, false. The literal truth, according to modal fictionalists, is that there are no merely possible worlds (or merely possible situations, or outcomes), and there are no merely possible objects. Strictly and literally speaking, there is no sculpture that I spent this morning making, though there could have been. When the flipped coin comes down heads, there is strictly speaking no outcome of that very throw in which it comes down tails.

What is literally true, however, is that *according to the modal fiction*, or *according to the fiction of possible worlds* there is a (merely possible) sculpture I could have spent this morning making, and there is an (unactualised) outcome of the toss in which the coin came down tails. What is said in talk about merely possible worlds and merely possible objects is generally literally false, but the slightly more longwinded talk about what is true according to the fiction of possible worlds is literally true. One might think (as Hinckfuss 1993 does) that talk about possible worlds is (or should be) governed by implicit presuppositions known to be false so that what is *said* in the language of possible worlds does not commit one to the existence of possible worlds, but only to some more economical proposition: something of the kind "if there were possible worlds of such-and-such a sort, then ...", or "given the

presupposition that there are possible worlds, ...". Or you might have some other account of the functioning of talk about possible worlds: Nolt 1986 suggests we should take typical "possibilistic discourse" to be a game of make-believe (Nolt 1986, p. 440), and while Nolt does not tell us specifically what theory of make-believe he has in mind, there are many theories of make-believe (most famously Walton 1990) that might be employed by a modal fictionalist to explain the behaviour of our typical utterances about possible worlds.

The main benefit which a fictional approach to possible worlds offers is, of course, the advantage of using the language of possible worlds, without the stiff ontological cost of literal commitment to such worlds. It is an especially tempting account of merely possible objects (like blue swans, or dragons, or the Holy New Zealand Empire): even those who accept some abstractionist account of possible worlds (see van Inwagen 1986) might well be reluctant to accept the literal existence of their contents. After all, it is often thought that the distinguishing mark of *merely* possible objects is that they do not actually exist.

Central to fictionalist treatments of possible worlds are biconditionals connecting truths about necessity and possibility, on the one hand, and the contents of the modal fiction, on the other. Central biconditionals will be all the instances of the following schemas (where P expresses a proposition):

> Possibly P iff according to the fiction of possible worlds, P is true at some possible world.

> Necessarily P iff according to the fiction of possible worlds, P is true at all possible worlds.

Either schema will be adequate to yield the other, given the standard inter-definition of possibility and necessity, provided enough logical machinery is available for reasoning within the scope of the "according to the fiction" operator. As a matter of fact, the above is a simplification, since, according to Rosen (1990, p 335), what is true according to the fiction of possible worlds may be only a proposition connected with P: the paraphrase of P into the language of possible worlds. In general, for Rosen, this will be the analysis of P in Lewis's theory of possible worlds. Rosen states the form of the fictionalist biconditionals at its most general as:

> P iff according to PW, P*.

where "PW" is the fiction of possible worlds, P is any proposition, and P* is its possible-worlds "paraphrase" (Rosen 1990, p. 335). In simple cases the above biconditionals will do as they are: for example,

> Possibly, pigs fly iff according to the fiction of possible worlds (or according to PW), at some possible world, pigs fly.

In less straightforward cases, however, the proposition expressed by P* may have to differ from that expressed by P.[1] Fictionalists may also employ other biconditionals when constructing their fiction -

one example is a biconditional to ensure that every proposition that is (really) true is according to the fiction true in the actual world.[2] A modal fiction may require more contents than those yielded by such biconditionals: what other contents the fiction of possible worlds might contain is an important question, and one unsurprisingly which different modal fictionalists will answer differently.

1.1 The Contents of The Fiction of Possible Worlds

Many fictionalists are far from explicit about exactly what the content of the fiction of possible worlds is to be. (Often they are also silent on the pressing issue of how one would justify one fiction rather than another: but more on this in section 4). Two explicit proposals were made in 1989 and 1990 about which fiction to use. Gideon Rosen's 1990 proposed using a slightly modified version of David Lewis's 1986 theory of possible worlds as the modal fiction. The theory had been proposed by Lewis as the literal truth, but by treating it as a mere fiction Rosen provided a ready-made story about possible worlds, their extent and nature. The other proposal was that of D.M. Armstrong's 1989. Armstrong proposed a "two-step" fiction, according to which there was a "great fiction", which asserted the existence of a lot of "little fictions", each of which completely described a possible world. Armstrong resisted identifying these complete descriptions of ways things could be with the possible worlds, as an abstractionist might, since Armstrong believed that worlds are supposed to be objects like our cosmos, rather than descriptions, properties, or other such abstract objects. Armstrong employed the two stage fiction, on the other hand, because he held that it is true at each world that it is the only world, and if the fiction were of a Lewisian pluriverse each world would (according to the fiction) be such that it was one of many worlds. In this respect, though Armstrong spells out the concern in quite a different manner, it can be seen as an anticipation of the Brock/Rosen objection (see section 3, below). In both the theories of Armstrong and Rosen the "worlds" in the story are thought to be 'concrete' cosmoi, like our own.[3]

For any set of principles governing the fiction of possible worlds, the question obviously arises as to why that set should be chosen. Alternatively, there are many stories we could tell about other worlds (whether they are cosmoi or fictional abstract objects), and the question arises as to why one rather than any other be chosen to be *the* modal fiction. (One could of course be more pluralist than this, and allow that different fictions are suitable for different purposes. The relativised question may still be asked, however: why *this* fiction, rather than any other, for *this* particular purpose?). This task of selecting and justifying one or more particular stories about worlds will be one of the challenges discussed in section 4. However, the next distinction to be discussed makes a big difference to what sort of answer to this question will be acceptable.

1.2 Strong vs. Timid

This next distinction concerns the role the modal fiction is to play in the theory. Is the fiction of possible worlds intended to provide an explanation of the applicability modal vocabulary, or not? The view that the truth of modal claims depends on, or is to be explained by, the contents of the fiction of possible worlds is often called "strong modal fictionalism", following Rosen's description of such a view (Rosen 1990, p. 354, Nolan 1997a). The view that modal truth does not depend on the contents of the modal

fiction (and usually, that what the contents of the modal fiction are depends on the modal truth), on the other hand, is known as "timid modal fictionalism", again following Rosen (Rosen 1990, p. 354).

Both the view put forward by Armstrong 1989 and the primary view discussed (but not endorsed) by Rosen 1990 seem to have been fictionalisms of the ‘strong’ kind (see Nolan 1997a, p. 263). ‘Timid’ fictionalism about possible worlds is also mentioned by Rosen 1990, and has been endorsed in passing by Field (1989, p. 41, 86). A modal fictionalist theory which sees the fiction of possible worlds as providing the resources for an analysis of modal statements would of course do important theoretical work: since the analysis of claims involving modal operators is among the most controversial and difficult issues in metaphysics. The downside is that strong modal fictionalism seems to face serious objections: see section 4.

Apart from any objections that might be leveled at strong modal fictionalism, the advocate of this variety of modal fictionalism faces the challenge of stating the contents of the modal fiction without relying on modal notions in a way that would make the explanation of those modal notions circular. Since the fiction explains the truth of modal claims, explaining the content of the modal fiction by reference to modal claims would be circular (e.g. by stipulating that all of the propositions which are necessary hold in all worlds, or that objects have in all worlds in which they exist the properties which they in fact have essentially). Stating, in non-modal terms, general principles about possible worlds that would enable one to determine, when presented with a proposition, whether it holds at all worlds, at some but not others, or none, is a difficult and controversial matter. Plausibly, a strong modal fictionalist will owe us such a non-modal specification of the contents of the modal fiction, if our understanding of the contents of the modal fiction is to allow us a useful method of assessing the truth of modal claims. Furthermore the strong modal fictionalist seems committed to *there being* a fact of the matter about the content of the modal fiction which is not itself to be explained or analysed in terms of further modal facts of the matter.

One particularly pressing instance of this difficulty (noted e.g. by Rosen 1990, p. 344) is the problem of how to understand the "according to the fiction of possible worlds..." operator. Natural initial glosses, as Rosen points out, include:

> If PW were true, then P would be true; If we suppose PW, P follows; It would be impossible for PW to be true without P being true as well. (Rosen 1990, p. 344)

The problem is that these seem modal.
Rosen offers several possible responses the fictionalist could make to this problem. He could admit that his theory does indeed contain a modal primitive, but claim nevertheless that it is some sort of advance in analysis, reducing all modal primitives to one. (Though as Rosen says, "According to the fiction of possible worlds" seems like a very odd primitive.) Or he could instead attempt to spell out the prefix in a non-modal way. This has not to date been attempted by many avowed modal fictionalists, but is one of the many tasks facing a modal fictionalist who sees the invocation of the fiction as the method for analysing or explaining modality.

If the modal fictionalist were only timid, on the other hand, then the fictionalist biconditional could be used to generate a great deal of the content of the modal fiction: whether or not the fiction claims that a given proposition is true in all possible worlds (or all accessible worlds) depends on whether that proposition is indeed necessarily true; whether or not a proposition is true at some world (or some accessible world) depends on whether it is possibly true; and so on. Furthermore, an analysis of truth in fiction in terms of some modal notion (whether a counterfactual conditional, strict implication, or whatever) would not make the account circular, since analysing modal discourse would be no part of the purpose of the fictional machinery. One recent worked-out modal fictionalist approach which explicitly analyses the relevant "according to the modal fiction" operator in terms of modal operators is Divers 1999b. Among other benefits, Divers argues that this sort of definition enables one to prove a "modal safety result": that when we have two modal claims A and B, and we have the possible-worlds-analogues of the two modal claims (call them A* and B* respectively), "Necessarily, if B* is a consequence of A* then B is a consequence of A" (Divers 1999b, p. 330). Such a safety result would be welcome, for it would provide the guarantee we needed that "detouring" in our reasoning through claims like A* and B* would not lead us astray when trying to determine whether B followed from A.

A drawback of timid modal fictionalism is that it leaves the important issue of a theory of modality to one side. Without some other positive story, even if it were primitivism about modality, it might seem that it would be difficult to motivate timid fictionalism over agnosticism about the status of possible worlds. Many of the leading candidates for analysis of modal claims are those that explain the truth of modal claims in terms of the literal existence and nature of possible worlds, so without a positive alternative story, it might seem premature to reject this style of explanation of modality. However, the project of endorsing fictionalism about possible worlds and possible objects without endorsing the further claim that this fiction provides the material to explain or analyse our modal notions, while in some ways less philosophically interesting than its strong cousin, is proportionally less open to serious objections.

1.3 Normative vs. Descriptive

Modal fictionalism can be interpreted as a descriptive theory of what our talk in fact amounts to, or as a normative proposal for how we should use talk of possible worlds. (In the terminology of Burgess and Rosen 1997, this is the distinction between a "hermeneutic" construal of modal fictionalism and a "revolutionary" construal of modal fictionalism). If the theory is that in fact we take possible worlds to be no more than convenient fictions, and, in the case of strong modal fictionalism, that facts about the content and nature of the fiction of possible worlds explain and/or provide the basis of the analysis of our modal locutions, then the theory is descriptive. This sociological claim that most philosophers who talk about possible worlds take this talk to be analogous to talk of e.g. ideal gases is a dubious one: my own impression is that modal fictionalism is a minority view amongst those philosophers who work extensively on the philosophy of possible worlds and its applications, though perhaps fictionalism might be the majority view when all who have opinions on possible worlds are taken into account.

In any case, fictionalism about possible worlds might be important even if the descriptive version is

incorrect. A normative claim to the effect that talk of possible worlds *ought* to be interpreted as merely fictional discourse, or the corresponding strong fictionalist claim that modal statements *ought* to be reinterpreted so as to be explained or analysed as statements about the content of the fiction of possible worlds, might be found attractive even if it were conceded that most people who employed the discourse were not talking fictively, or that the actual commitments of the folk and most philosophers who employ everyday modal idiom are to be cashed out in some other way (e.g. as implicit commitment to some false theory, perhaps one which took modal statements to have truth conditions in terms of an objective modal reality).

The two questions, of whether the fictionalist theory is supposed to describe our current practice, and whether it is supposed to describe a practice we should adopt instead of realism, are independent. Two people could agree that normal usage of possible-worlds talk is not intended literally, but disagree about whether we should accept the literal existence of possible worlds (that is, they could be descriptive fictionalists while disagreeing about the normative question), and likewise two people could be normative fictionalists, and claim that we should take ordinary possible worlds talk as being only fictionally true or appropriate, while disagreeing about whether in fact current usage reflects this fictionalism or instead reveals non-fictionalist commitments among the users of the vocabulary. (This independence may be overlooked if we employ Burgess and Rosen’s vocabulary, which suggests that fictionalists face a binary choice of a "hermeneutic" fictionalist approach or a "revisionary" approach.)

Unfortunately, modal fictionalists have often not been explicit about whether their theory is to be an analysis of the possible-worlds talk and perhaps modal discourse which is actually employed, or a normative suggestion about how we might move to a superior theory. It is to be hoped that those seeking to provide a fully worked out modal fictionalist position would make more explicit the status of their proposal.

1.4 Fictionalism about Modality and Modal Fictionalism

So far, it has been taken for granted that the modal fictionalist treats ordinary modal claims as sometimes literally true, and it is only claims about possible worlds and their contents which the modal fictionalist will wish to claim are literally false, but true according to a story. A more radical modal fictionalism is possible: one in which modal claims themselves (such as the claim that there could have been blue swans, or that necessarily, everything is self-identical) are not literally true, but only true according to a fiction. (In Nolan 1997a I called a version of modal fictionalism which took both claims about possible worlds and modal claims to be true only according to fictions *broad* modal fictionalism.)

Such extended fictionalism about modality could, I think, come in two main varieties. One would maintain that modal operators (and associated pieces of language like essential predication, counterfactual conditionals, and perhaps other things like assignments of probabilities) lacked literal application: that all statements prefixed with a modal operator were either uniformly literally truthvalueless or uniformly literally false. This faces several immediate formal difficulties, given the usual characterisation of the modal operators.[4]

A less blanket rejection of modality might only insist on the falsehood of some modal claims: think of the analogy of a moral nihilist who rejects only "positive" moral claims, such as that taking property without permission is wrong, or that it is good to help the ill, but accepts that it is not wrong to take property without permission (since nothing is wrong) and that it is not good to help the ill (since nothing is good). This fictionalist about modality might reject the literal truth of all claims about necessity, for example (and, in virtue of the interdefinition, accept the literal truth of all possibility claims). Or she might assert as little as is needed to preserve the basic modal inferences from actuality: for example, every true P is possibly true, and actually true, but this is the limit of literal modal truth. This would be more analogous to the fictionalist about possible worlds who nevertheless thinks that there is one possible world -- the actual, and one collection of possible objects -- the actual objects.

Fictionalism of this sort extending to modal discourse as well should be a position worthy of investigation by those attracted to fictionalist strategies and mistrustful of modality. Since it seems to lack contemporary advocates, however, it shall not be dwelled upon here.[5]

1.5 Further Extensions

Modal fictionalism has traditionally been conceived as fictionalism about possible worlds, and implicitly also about their contents. One natural way to extend modal fictionalism so understood is to admit impossible worlds as well. Use of a fiction of impossible worlds is unlikely to seem too extravagant for many modal fictionalists, but it might also provide a mechanism for non-fictionalists about possible worlds to talk of impossible worlds without the extra theoretical costs they fear. So there is a partially fictionalist option of being a realist of some stripe about possible worlds but a fictionalist about impossible worlds (and perhaps other situations, e.g. incomplete or underdetermined ones). This partial fictionalism will face some of the same sorts of questions, and employ the same sorts of strategies, as modal fictionalism about *possible* worlds.

Another mixed strategy is rarely explicitly endorsed, though it may be implicit in many theorists' talk. By far the majority of realists about possible worlds take them to be abstract objects of some sort: sets of sentences or propositions, uninstantiated world-properties, or unactualised maximal facts, or perhaps even as sui generis simples. However, their approach to merely possible objects is often less explicitly spelled out. As well as possible worlds, merely possible objects such as blue swans, my counterparts, Newtonian masses, XYZ, and many others are discussed and quantified over. Some abstractionists will wish to identify such merely possible objects with actual abstract objects - perhaps uninstantiated properties that are less than world-properties, descriptions of parts of worlds, set-theoretic representatives, or whatever. But many more will not. This is for at least three reasons. The first is the intuition, never entirely quelled, that a merely possible blue swan should be *blue* and a *swan*. Abstract objects are rarely, if ever, either. The second is that it is often thought that the whole point about merely possible objects is that they do not exist, but might have. Admitting their literal existence by identifying them with actual existing abstract objects will go against the grain for many. Finally, many theories of possible worlds lack obvious candidates to play the roles of merely possible objects. This is most obvious for those theories that take possible worlds to be simples, but it is also not straightforward to divide a set

of propositions into "object shaped" components, or to distinguish one merely possible object from its duplicates if one identifies worlds with world-properties, and furthermore seeks to divide these up to yield merely possible less-than-world-sized objects. Fictionalism about merely possible objects is an underappreciated theory, and should perhaps recommend itself to more abstractionists about possible worlds than it in fact does.

2. Fictionalism in Context

2.1 Modal Fictionalism and Fiction

A theory that holds that possible worlds are to be treated as having the status of fictional objects immediately gives rise to questions about the status of fictions and fictional objects more generally. The issue of how to understand the "according to the fiction..." operator has already been mentioned. Other issues in the treatment of fiction are the question of whether "theoretical" fictions are to be seen in the same light as literary fictions, and the question of what ontological status fictions possess.

There is room in principle to distinguish claims about paradigmatically fictional characters (like Sherlock Holmes or hobbits) from the treatment one wishes to apply to claims which serve useful theoretical purposes but which are not to be taken literally, or at least not at face value: to distinguish "fabulous" and "fictional" entities, to use Bentham’s not entirely perspicuous terminology (Bentham 1959). One might wish to do this, for example, if one thought that paradigmatic story-telling consisted of utterances that were not truth-apt (not anyway through the same mechanism as ordinary utterances), while claims about, say, ideal gases were better construed as standard sorts of assertions that were literally false. In practice, however, recent modal fictionalism discussions have proceeded as if the theory holds that possible worlds are fictional in the fullest sense of the word.

The ontological status of fictions is an important issue to settle if we are to determine whether modal fictionalism is any advance metaphysically on rival realist theories of possible worlds. On pain of circularity, the ontology of fiction, as conceived by the fictionalist, had best not include possible worlds (as the ontology of fiction offered in Lewis 1978 does), but there are other sources of concern. Rosen points out that fictionalists must believe in fictions, stories, theories, or somesuch, and that if these are construed as abstract objects, they will not be philosophically uncontroversial (Rosen 1990, p. 338). Rosen argues that belief in abstract objects like stories and theories is less revisionary than belief in Lewisian possible worlds, at least. However, it is hardly clear that an ontology of abstract representational entities is any more or less objectionable than the ontology of abstractionist theories of possible worlds. (Lycan 1993, p. 16, Nolan 1997a, pp. 271-273). Alternatively, fictionalists could commit themselves to more concrete, "nominalistically respectable" fictions only: the relevant fictions or theories of possible worlds might be conceived of only as ink-marks on pieces of paper, or information states inside brains, or perhaps as some amalgam of these and other actual, concrete ontology. This approach is not problem-free either (see section 4): it is far from clear that it will serve as a basis for paradigm fictions, let alone be enough to explain possible worlds.

2.2 Fictionalism and Instrumentalism

Fictionalism about possible words, in the sense of "fictionalism" that has been used here and that is standard in the literature, can be distinguished from instrumentalism, in a way that is also standard for those who bother to distinguish them. The main difference, as usually conceived, is that a fictionalist holds that claims in the target discourse, at least made without presuppositions, are literally significant and truth-apt, but in fact false (though true according to a fiction, theory, or whatever). A traditional instrumentalist, on the other hand, is often characterised as taking the claims to lack truth-value. Some instrumentalists will go further, and say that such apparent claims are not capable of being genuinely asserted at all - their linguistic function is a different one. Be that as it may, the difference is not terribly great. In both cases, there is a realm of literal truths (e.g. truths of modality), and a realm of ‘claims’ that are not literally true, but which it is useful to advance (or manipulate in a similar way) "as if" they were true, in some respects at least. In both cases some set of biconditionals will be needed to move from literal truths to what is true according to the fiction (or acceptable in the instrumentalist discourse), and back, and both views will join in repudiating realist theories of possible worlds, while wishing to retain the legitimacy of talk *prima facie* about them.

Other versions of instrumentalism will be even more similar to fictionalism. So-called "epistemic instrumentalism", for example, does not deny that the claims of the relevant theory are truth-apt, but says merely that we should withhold belief from them (a *locus classicus* of this sort of instrumentalism about unobservables is van Fraassen 1980). Such a view is virtually the same as fictionalism, except for the cosmetic difference that it is more likely to be cast in terms of what is true "according to the theory..." (or "according to the models of the theory...") instead of "according to the fiction...", and the slightly more noteworthy difference that such instrumentalism is neutral on the question of whether the theory in question is true or false, whereas fictionalism is committed to its being false. In the case of possible worlds, one interpretation of the theory of Merrill 1978 is that it should be seen as a variety of van Fraassen-style "instrumentalism", employing semantic models of possible worlds without belief in the possible worlds themselves (indeed, while being convinced the worlds themselves do not exist). Merrill himself describes his position as "instrumentalism" about possible worlds.

Other positions in the literature are sometimes labelled "instrumentalism": for instance Forbes 1983 names his position "instrumentalism", even though he believes in the literal truth of many positive possible-worlds statements, since he claims these statements are to be understood not as straightforward quantifications over possible worlds and possible objects, but rather as equivalent to statements not quantifying over such objects. Whatever it is called, Forbes’s position on possible worlds is far removed from the other positions labelled instrumentalism here.

2.3 Fictionalism and Eliminativism

Fictionalists have in many respects only small differences from instrumentalists -- though of course the significance of differences can vary from inquiry to inquiry, and the philosophy of language issues that

arise in making the distinction are to some both interesting and important. The disagreement between fictionalists and eliminativists about possible worlds is more significant, on at least one understanding of what it is to be an eliminativist. In one sense, of course, fictionalism is a species of eliminativism, if eliminativism about possible worlds is just the doctrine that there are no such things. If, however, eliminativism about a domain is the doctrine that not only do the putative objects committed to not exist, but that discourse about (or apparently about) such objects is to be done away with, then clearly this doctrine will be in conflict with the fictionalism which seeks to give a fictionalising interpretation to possible worlds discourse just in order to salvage a useful theoretical and heuristic device without paying the ontological (and other theoretical) costs. An analogy might be drawn in the realm of theoretical entities: while fictionalism is the standard approach to apparent commitment to ideal gases, eliminativism is the contemporary approach to luminiferous ether.

Again, distinguishing fictionalism about possible worlds (and a fictionalist treatment of apparent commitment to possible worlds) from eliminativism about possible worlds (and an approach of rejection and cessation of apparent commitment to possible worlds) is to some extent an exercise in stipulation. For one could, with equal justice, count as fictionalist a position which held that possible worlds had the status of other fictional objects (i.e. non-existence), that talk of them was at best to be treated fictionalistically, but that furthermore such talk was in the end misleading or worthless or defective to the point that it should be abandoned. If one were to go this way instead, fictionalism and eliminativism would overlap. Including in fictionalism a commitment to the worth of possible-worlds discourse would rule out the view just outlined as a *fictionalist*, as opposed to eliminativist, position.

Whatever terms are used to describe the distinction (and it is unlikely that any short expressions will conveniently mark the difference without some stipulation), the distinction between those who think that our practice of apparently quantifying over possible worlds and making claims about their contents is a valuable one worth preserving, even if there are no (merely) possible worlds, or literal truths about them, and those who think that this practice should be dispensed with, perhaps *because* there are no (merely) possible worlds, or truths about them. This distinction marks an important divide among those who reject the existence of possible worlds and their contents: and however exactly the divide is drawn, typical fictionalists disagree with 'hard-core' eliminativists on this issue.

While a fictionalist, so distinguished, and an eliminativist, so distinguished, may agree to a great extent about the fundamental metaphysical issues, they may have very significant differences when it comes to philosophical practice. For an eliminativist will eschew explanations couched in terms of possible worlds, whether that be in the semantics of conditional statements, analyses of supervenience claims in ethics or philosophy of mind, possible-worlds accounts of content, explications of the operation of modal operators or probability assignments in terms of possible worlds, and so on. Fictionalists, on the other hand, will typically hope to salvage some or all of these explanations, even if an appeal to possible worlds is not quite the end of the explanatory story. It is this difference in the spirit of the eliminativist and fictionalist approaches (as here distinguished) that is the interesting difference between the views.

3. Technical Problems

There are a variety of technical objections to modal fictionalism, and if these objections succeeded they would derail specific proposals like that of Rosen 1990 and perhaps cause difficulties for modal fictionalism in general. The most discussed of these objections is the Brock/Rosen objection, an objection to the theory of Rosen 1990, published independently in Brock 1993 and Rosen 1993. (In the case of Rosen 1993, a variety was also leveled at the fictionalism of Armstrong 1989).

3.1 The Brock/Rosen Objection

The reader is advised to consult Brock 1993 and Rosen 1993 for an exact statement of their objections, but in essence the problem arises when we consider the modal status of certain claims about possible worlds themselves. One of the principal fictionalist biconditionals, as we have seen, is the biconditional connecting necessary truths and what the fiction asserts about all possible worlds:

> Necessarily P iff, according to the modal fiction, at all worlds, P*,

where P* is the possible-worlds paraphrase of P. According to the fiction employed by Rosen 1990 (i.e. the theory which David Lewis offers as fact), there are many co-existing concrete possible worlds - each one its own cosmos. Furthermore, this claim is true at any of these possible worlds. So, according to the modal fiction, at all worlds it is true that there are many other possible worlds. However, it follows from the biconditional that since, according to the modal fiction, at all worlds, there are many possible worlds, it follows that necessarily, there are many possible worlds. Since necessarily P implies P, it follows that there are (literally!) many possible worlds. But the whole point of modal fictionalism was to deny (or at least avoid asserting) that there were many possible worlds. So modal fictionalism, at least of the variety described, is self-refuting.

So the objection goes. (The version presented above is Brock's from Brock 1993 -- for some of the detail of Rosen's argument, see below or Rosen 1993.) Two direct responses have been offered to this problem in the literature. The first, by Peter Menzies and Philip Pettit (Menzies and Pettit 1994), conceded that the objection "is decisive against the letter of the Rosen proposal" (p. 29), and sought to provide modified fictionalist biconditionals to produce a modal fictionalist theory which would not be susceptible to the Brock/Rosen objection. Nolan and O'Leary-Hawthorne 1996 produced a version of the Brock/Rosen objection which they believed circumvented the Menzies and Pettit solution offered in section 3 of Menzies and Pettit's paper. Section 5 of Menzies and Pettit gave another translation scheme meant to avoid the Brock/Rosen objection; it is argued in the following supplementary document that this scheme too is unsatisfactory.

> [Supplementary Document: A Persisting Problem For Fictionalism About Possible Worlds]

I shall not dwell on the details of the Menzies/Pettit position, since the other response to the Brock/Rosen objection has been more influential: the response of Noonan 1994. Noonan claims that the Brock/Rosen

objection is not even successful against the letter of the original proposal in Rosen 1990 (Noonan 1994, p. 133). He argues that careful attention to the procedures actually given by Lewis (Lewis 1968) for paraphrasing modal claims into claims in the language of possible worlds will show that "according to the fiction of possible worlds, at any world, there are many worlds" will not be able to be derived, and so the move to "necessarily, there are many worlds" cannot be carried through. Thus if the fictionalist sticks closely to the possible-worlds "translations" given in Lewis 1968, s/he will be able to avoid the threatened collapse.

The way the Brock/Rosen objection is raised by Rosen begins with an apparently harmless statement of contingency: (numbering of claims are Rosen's, and while Brock and Rosen's presentations are reasonably informal, the formal translations are given by Noonan, and his numberings for those are provided)

> (2) Necessarily, it is contingent whether kangaroos exist

or to put it in formal modal logic ($Kx = x$ is a kangaroo):

> $\Box (\Diamond \exists x Kx \ \& \ \Diamond {\sim} \exists x Kx)$

From this, Rosen claims, the "standard analysis" delivers:

> (3) At all worlds, there are worlds where kangaroos exist and worlds where they don't.

Ignoring for our purposes the complications which need to be introduced if we are to add accessibility relations between worlds, the Lewis 1968 equivalent of (2) is

> (3L) $\forall w(Ww \rightarrow \exists w'(Ww' \ \& \ \exists x(Ixw' \ \& \ Kx)) \ \& \ \exists w''(Ww'' \ \& \ \neg\exists x(Ixw'' \ \& \ Kx))))$

Noonan points out that this formula does not imply

> (4L) $\forall w(Ww \rightarrow \exists w'(Ww' \ \& \ Iw'w \ \& \ \exists w''(Ww'' \ \& \ Iw''w \ \& \ \neg w'=w'')))$

which is the formula required in Lewis 1968 to be able to move back to "necessarily, there are two worlds", and which is the formula Brock and Rosen would need to derive if they were to show that Rosen's (2) (or Brock's equivalent) led the modal fictionalist to have to say that there are literally several worlds. The reason why (3L) does not imply (4L) is that it is not enough for there to be two worlds at a given world (V, let us call it) that there be existential quantification over two worlds within the scope of an existential quantifier committing us to V: the two worlds must also be "in" V, in the sense that the two place predicate "I" must hold between V and each of the worlds. This is what happens in (4L), but it does not happen in (3L), where there are two existential quantifiers over worlds in the scope of the outside universal quantifier, but where the worlds existentially quantified over are not asserted to be "in" any of the worlds *w*.

Rosen 1995 has accepted Noonan's resolution of the problem apparently posed by the Brock/Rosen objection. Rosen has thus changed his preferred proposal, so that instead of a general endorsement of the position outlined by Lewis 1986, Rosen's recommendation for modal fictionalists now relies more heavily on Lewis 1968. Instead of the simpler biconditionals discussed near the beginning of this entry and in Rosen 1990, the revised proposal is to take equivalences asserted by Lewis 1968 between modal claims and claims couched in terms of quantification over possible worlds, and treat those equivalences instead as specifying connections between modal statements and claims about what is true according to the fiction. If a modal claim is literally true, its associated world-claim is true according to the fiction, and vice versa.

So the state of play seems to be this: while the Brock/Rosen style objection remains something for a modal fictionalist to be wary of when constructing the fiction of possible worlds, it is possible to avoid the objection by being suitably careful about what fictionalist biconditionals to employ: and if Noonan is right, strict adherence to the fictionalist modification of the equivalences offered by Lewis 1968 provides one suitable way of being careful. It is not uncontroversial, of course, that Noonan's strategy is the one a modal fictionalist should employ: Divers 1999a and 1999b argue that it is not.

3.2 Hale's Dilemma

Bob Hale (in Hale 1995b) posed a dilemma for modal fictionalism (more specifically, Rosen's version of modal fictionalism, though other varieties face a similar dilemma). A modal fictionalist who maintains the version outlined in Rosen 1990 believes that the fiction of possible worlds (PW) is not literally true. A question arises about the modal status of the fiction: is it necessarily false, or contingently false? In either case, Hale argues, the modal fictionalist is in trouble.

Should modal fictionalists claim that the story of possible worlds is necessarily false, then Hale argues that they cannot gloss their "according to the fiction of possible worlds ... ." prefix as "were the fiction of possible worlds true, then ... would be true". This is because, according to Hale, conditional claims with antecedents which are necessarily false are automatically true, so if the fiction of possible worlds is taken to be necessarily false, then all conditionals of the form "were the fiction of possible worlds true then ..." are true, and not merely the ones that the modal fictionalist wishes to endorse. If the modal fiction is to be useful, not everything should be true according to it: examples of claims that had better not be true according to it include the claim that 2+2=7, or the claim that there are no possible worlds.

On the other hand, if the fiction of possible worlds (PW) is only contingently false, Hale claims this also lands the Rosen's fictionalism in trouble, since if its falsehood is only contingent, then the fiction might have been literally true (or it is possible that the fiction be true). But according to Hale the "official fictionalist paraphrase" of what this possibility would amount to "cannot adequately capture the content of the claim that possibly PW is true". (p 65) Hale claims this is so because the claim "According to PW, there is a possible world at which PW is true" is equivalent for Rosen's fictionalist to "If PW were true, there would be a world at which PW is true": and this conditional is one which would be true whether or

not PW was true.

A modal fictionalist might try to resist either horn of the dilemma. On the first horn, modal fictionalists might employ another gloss on what it is to be true according to PW, or they might endorse one of the various theories of conditionals on which conditionals with necessarily false antecedents are not automatically true. On the second horn, even fictionalists who accepted that they were committed to analysing their claim that PW could have been true as "If PW were true, there would be a world at which PW is true" could dispute Hale's claim that this is inadequate (see for example Divers 1999b p 325-326).

A third option, explored by Rosen 1995, is not to take PW to be false, but rather altogether lacking a truth-value -- e.g., in virtue of employing terms with no literal application, such as "... is a world-mate of...". Hale's dilemma is directed primarily against fictionalists who take the literal content of their fiction to be false, and those fictionalists prepared to ascribe some other status to their fictional claims avoid the dilemma as initially stated (though this route may encounter difficulties of its own, especially if it retains some sort of conditional analysis of the "according to PW" prefix).

Rosen 1995 and Divers 1999b are among the responses to Hale, and Hale has in turn responded to Rosen 1995 in Hale 1995a, where Hale argues that several of the responses suggested by Rosen in turn face serious problems.

4. Other Concerns

In addition to the technical challenges facing modal fictionalism outlined in the previous section, modal fictionalism has been challenged on a number of less technical grounds. Not all of these challenges are equally cogent against every variety of modal fictionalism, and some explicitly have as their targets only some versions of the doctrine.

4.1 Artificiality

Fictions are human products: they have authors, and those authors have a good deal of control over what is true according to them (though some fashionable postmodernists might disagree). Alternatively, if it is thought that fictions are timeless Platonic abstract entities (sets of propositions, perhaps), we should say that which fictions we consider and express is a matter of human activity: and the ‘authors’ of those eternal Platonic fictions actually expressed have a good deal of control over which of a variety of such fictions they express. However, "theoretical fictions" introduced as an alternative to realist theories, and which are supposed to play a serious role in inquiry, do not seem to be able to be as arbitrary. Not any old story told about possible worlds will serve as the modal fiction, at least if it is to provide the heuristic and other advantages of talk about possible worlds. The suspicion is that in some respects talk of possible worlds cannot be like paradigm fiction, since the ‘choice’ of which story about possible worlds should count as the modal fiction is not as up to us as what to say about Sherlock Holmes was up to Conan Doyle.[6]

Modal fictionalists can certainly respond to the more general worry that fictionalism makes it too arbitrary which particular story about possible worlds ought to be employed. After all, the purposes of the fiction constrain what sorts of fiction are suitable, just as not any old story about a gas will serve to provide an "ideal gas" which has a behaviour approximated by real gases. Some story must be told about what sorts of constraints are appropriate, and why: and this can be difficult in its own right (see the "which fiction should be employed" section, below). Even if there are substantial constraints on what story will be adequate, and these constraints are not due merely to facts about us or our choices, there may still be some scope for the story to be artificial to a small degree, for some points of detail may be left underdetermined by the constraints. This may also be considered a problem, but it is unlikely to be fatal.

There is a more specific worry that remains even if a suitable account of what constraints there are on the modal fiction can be given. This is that it is too contingent a matter whether there is a modal fiction at all. After all, if there had never been any sentient creatures, no stories would ever have been told, and even if the modal fiction is construed as a Platonic entity (a collection of propositions, perhaps), it may never have been a fiction if it was never expressed by story-tellers. (Of course if nothing hangs on whether or not it is a fiction, then this will not worry the Platonist modal fictionalist). This worry, again, is particularly pressing if modal truth is to depend on the contents of the fiction, since it does not seem that whether or not blue swans are possible, for example, depends on whether or not anyone ever told stories. There are responses to this worry, and responses to these responses: it is suggested that the interested reader consult (Nolan 1997a).

4.2 Incompleteness

Fictions are often incomplete: they are silent about some issues. The Sherlock Holmes stories make no representation one way or the other about the exact population of India, or whether the number of hairs on Dr Watson's head is odd or even. Arguably, the modal fiction will be incomplete too: there will be some propositions such that neither they nor their negations will be true according to the fiction. This prospect raises several worries.

First, there is the "incompleteness problem" discussed by Rosen in Rosen 1990, pp. 341-345. There are some modal issues (and corresponding issues about the nature of possible worlds) that a realist may well be silent on: not because they believe there is no answer, but rather because they believe themselves ignorant of the answer. A fictionalist who treats the realist's theory as a fiction, on the other hand, will be silent upon the same issues - but this can lead to a more serious problem. If the fiction is silent on an issue (Rosen's example concerns the size of worlds), it is not that the issue is unknown -- it is that the fiction does not represent a fact of the matter one way or the other. So it might appear for the fictionalist there is not an unknown modal fact - either the claim is false because the corresponding worlds-claim is not true according to the fiction, or something involving a truth-value gap is going on. Rosen also discusses what effect this might have on modal claims corresponding to such silences. A detailed discussion and critique of Rosen on this issue can be found in the following supplementary document:

[Supplementary Document: Rosen's Incompleteness Worry]

What is uncontentious is that modal fictionalists operating with fictions which are incomplete in the way, e.g., that the fiction of Rosen 1990 is, will face difficulties, or at least departures from orthodoxy which will be found unattractive by some.

Another "incompleteness" worry in the literature is that expressed in Nolan 1997a. This is also a worry that the modal fiction will not represent as much as is desirable, though the concern is not confined to those areas in which realists might confess ignorance. (The concern resembles an objection Lewis brings against "sparse linguistic ersatzism" in Lewis 1986, pp. 142-165.) A modal fiction, to be adequate, must represent a very great deal about possible worlds, since there are infinitely many claims about possible worlds that must be part of the content of the fiction if there are to be enough possible-worlds claims to correspond to all the modal claims we would accept. Only a tiny proportion of the propositions about possible worlds needed will be able to be stated explicitly by the modal fictionalist: constraints of time and space and publication costs will mean that the fictionalist will need to describe the fictional worlds in only a few volumes, while an exhaustive explicit description of even a single possible world as complex as our actual world is beyond our finite resources.

What modal fictions will presumably have, however, are generalisations about possible worlds: for instance, principles of recombination and plenitude, principles about what truths are respected by all worlds, and so on. The modal fictionalist might reasonably hope that these generalisations *imply* all of the specific claims needed by the fiction. Implication is, however, a modal notion: not that this is automatically a problem, but it will be a problem for the "strong" modal fictionalist, who seeks to reductively analyse modality in terms of what is true according to the fiction. The strong modal fictionalist's analysis will be circular if he has to appeal to something like implication (or related modal notions) to spell out what claims are represented as true by the fiction, as it seems he must if the bulk of the claims are to be represented only implicitly. It seems that a strong modal fictionalist will be stuck with a radically incomplete fiction, if he relies only on what his modal fiction explicitly says, or he faces the task of specifying the implicit content of the modal fiction without recourse to modal notions like implication.

A strong modal fictionalist could attempt to capture non-explicit content in ways that did not rely on modal resources: one way this could be attempted would be to offer a syntactic account (or some other non-modal account) of some sort of consequence relation, and to stipulate that the fiction is to be considered closed under the relation thus specified. However, this is not easy to do in such a way that all of the necessary semantic consequences are indeed "consequences" of the explicit generalisations given. Furthermore, success at this project threatens to undermine the strong modal fictionalist's project in another way: for if it was possible to give a specification of a relation of "consequence" without relying on primitive modal notions, and that did the work of semantic consequence, then this would offer an analysis of "broadly logical" consequence (and presumably related notions, such as logical necessity and possibility) directly, rather than in terms of what was true according to a modal fiction, thus making the strong modal fictionalist's analysis of modality in terms of the fiction redundant. So the strong modal

fictionalist faces a serious challenge in providing a fiction capable of representing what is needed for his theory to be adequate.

4.3 Which Fiction Should Be Employed?

An essential part of an adequate modal fictionalist theory is a specification of the fiction of possible worlds which is to be employed. As well as selecting one of the many potential candidate stories about worlds, it is also essential to provide an explanation and justification of the choice. This is very rarely done by modal fictionalists (Armstrong 1989 provides an exception). This is not to say that it cannot be done, or cannot be done plausibly: but justifying the choice of fiction is not something that can be neglected if a modal fictionalist theory is to be convincing.

As with so many other challenges, timid modal fictionalism can immediately provide the outlines of an answer to this question. (Though it is to be remembered that timid modal fictionalism is able to avoid so many theoretical difficulties only because the fiction is not asked to do much theoretical work). If the truths of possibility and necessity (and conditionality, and other modal truths) obtain without dependence on the content of the modal fiction, it is surely reasonable to suppose that whichever fiction it is correct to employ, it must respect those independently obtaining modal truths. Strong modal fictionalists must also ensure that the contents of the fiction are associated with the modal claims they wish to make in the appropriate way, of course, but this will be of less help to them in establishing the content of the modal fiction. For if the content of the modal fiction is to explain the truth of the modal claims, it must be able to be fixed independently, on pain of circularity. This is especially so if the strong modal fictionalist holds, as one well might, that it is our understanding of the modal fiction that provides (perhaps implicitly) our epistemic access to which modal claims are true and which false. Giving a non-circular specification of the content of the modal fiction is one of the very difficult challenges facing the strong modal fictionalist.

While strong modal fictionalists cannot appeal to an independently constituted body of modal truths, one thing they can do is insist that the modal fiction respect our ordinary modal *judgments*: that is, that by and large if we *accept* a modal claim as true, the associated claim involving possible worlds will be true according to the modal fiction. (Rosen 1990, p. 337, speaks of the desideratum that modal fictionalism "*ratify* a substantial body of prior modal opinion".) This is presumably not to forbid any departures from our pre-theoretic modal judgements, should they be required, but it would provide a way even for the strong fictionalist to rule out gratuitous departures from our modal opinions.

The next obvious source of content for the modal fiction is the literal truth about our actual world (Rosen 1990, p. 335). The addition of all literally true non-modal propositions (in a suitable sense of "non-modal") to the fiction as part of its description of the actual world is useful, since it provides a rich source of content that can be extended by, for instance, a principle of recombination, to yield claims about non-actual worlds. It would also seem to be required, for if the fiction fails to be committed to the actual world verifying a certain non-modal truth *q*, the inference from *q* to Actually-*q* and back will be jeopardised. Some particular non-modal truths may prove especially useful: Armstrong 1989 pp 138-139

mentions analytic truths, truths in virtue of the meanings of terms, in this connection. One can either add the actual-world non-modal content by including an "encyclopedia" in the fiction, as Rosen does, or one could allow it in by, for example, stipulating as extra bridge-laws biconditionals of the form:

P iff According to PW, at the actual world, P

for all non-modal propositions P.

As well as conformity with our pre-theoretic modal judgements and inclusion of an encyclopedia of actual non-modal truths, Rosen 1990 mentions another source of information to apply in specifying the modal fiction. We have practices of forming modal beliefs involving imagining situations in accord with principles of recombination, non-arbitrariness, and so on (p. 339-40). Rosen points out that while a realist has the challenge of explaining why this practice of imagining should be a guide to modal truth, the fictionalist need not face this challenge if those practices are part of the process of constructing the fiction of possible worlds. If the constraints, or limits, on our imaginative practices when considering hypothetical situations are vital in our practice of making many of our modal judgements, it would make sense to similarly constrain the modal fiction.

There are no doubt many other sorts of constraints which a modal fictionalist may appeal to in order to narrow down the class of fictions about possible worlds which are acceptable for her purposes. Even after all of these constraints are in place, however, there may still be the theoretical possibility that more than one fiction about possible worlds (complete or incomplete) satisfies them equally well. A fictionalist facing a choice between equally deserving fictions would need to address the issue of what attitude to take to other modal fictionalists who choose differently. (Should they be judged incorrect? Correct, because judgements about the content of the modal fiction are relative to which (acceptable) fiction is adopted? Or should they be judged to be talking about something else?). Or perhaps the fictionalist could find some way to avoid making the choice of one single fiction.

If the fictions disagreed sufficiently, there may even be fictions which meet the constraints but which differ on matters which are linked through the fictionalist’s biconditionals with literal modal claims. (This is only possible if the modal truths themselves are not being appealed to as constraints on acceptable fictions, so is not a problem which faces the timid modal fictionalist). If the constraints are not enough to uniquely determine the truth-value of every modal claim, then not only the determinacy of the content of the fiction but the determinacy of the truth-value of some modal claims is at stake. Are those modal claims true or false, or neither? Might they be fiction-relative, so there are no-fault disagreements about them?

This is not the place to attempt to settle the matter of whether constraints are likely to uniquely determine a modal fiction, nor whether it would be genuinely objectionable if they did not do so. Rather, the issues are mentioned as ones to be kept in mind when formulating or defending a modal fictionalist theory.

4.4 The Theoretical Primitives of Modal Fictionalism

Metaphysical theories often rely on resources which are taken as "primitive": roughly, theoretical resources which are not to be further explained or analysed. Different theories of the same subject matter will often take different resources to be primitive, and while it is a difficult question to decide whether one set of primitives is better or worse than another, evaluation of the relative simplicity, naturalness, or other merits of theoretical primitives is part of the evaluation of rival theories. This sort of comparison can can be especially relevant in areas where disputes between rival theories are not to be settled easily by experiment or observation. Such disputes make up one of the battlegrounds between fictionalists and their rivals, with anti-fictionalists claiming that the unanalysed theoretical resources which fictionalists rely on render fictionalist theories unattractive, or at least relatively unattractive compared to some rival or other.

The central piece of theoretical machinery the modal fictionalist employs is the "According to PW ..." operator. When it is glossed in tempting ways, as "if PW were true, then ..." or "it follows from PW that ...", it seems to be a modal notion: and if this is not to be further explained, the modal fictionalist cannot use the fiction of possible worlds and its contents as the basis of a reductive analysis of modality. This will only be of concern to some modal fictionalists, of course -- timid fictionalists will not have been looking for a reductive analysis of modality based on their fiction in the first place - and some timid fictionalists such as Divers 1999b explicitly endorse modal explanations of the fictionalist operator (Divers 1999b p 335). Such fictionalists may be happy to take advantage of possible analyses of "according to the fiction" operators in modal terms, and in so doing provide an answer to the question of how to understand such expressions: but on the other hand, their position may not be attractive to someone primarily concerned to analyse modal operators. (Even timid fictionalism is compatible with a reductionist account of modality, of course, since the timid fictionalist may seek to explain modality in some other terms. It is just that it is not hospitable to reductionist accounts of modality in terms of possible worlds).

A fictionalist who wishes to provide an analysis of modality, on the other hand, had better not take their "according to PW ..." operator to be analysed in terms of standard modal devices, or alternatively in terms of possible worlds (see Rosen 1990, pp. 344-345). The canonical version of the theory that Rosen presents takes the "According to PW ..." operator to be a primitive one: that is, one which is not to be further analysed, in modal or non-modal terms (Rosen 1995, p. 70). Rosen points out that one might think that his favoured prefix is a modal locution, and if so even his position cannot be said to entirely reduce the modal in favour of the non-modal (Rosen 1990, pp. 344-345). Nevertheless, as he points out, it may still be thought to be some theoretical advance to be able to explain all of the other modal notions using only this one. It is hard to know how the issue of whether or not "according to PW.." should count as a modal operator is to be decided: in any case, it will not be further pursued here.

Regardless of its status as a modal locution, Rosen recognizes that it is a very unsatisfying primitive: the notion of a proposition being true according to PW is an unlikely one to be considered basic and unanalysable. Whether or not this is a fatal flaw of Rosen’s proposal is, he acknowledges, "a matter of somewhat delicate judgement" (Rosen 1990, p. 349). What he does have to say about it, however, is that arguably many realists about possible worlds have also not provided a satisfactory analysis of the

"according to the fiction ..." operator, and so face the same challenge.[7]

The issue of whether "According to PW..." is a satisfactory theoretical primitive is presumably partly to be settled by seeing what rival theories are possible, and what primitives they need to rely on to account for modality and for fiction. Beyond that, how to settle disputes about the relative attractiveness of primitives is a difficult issue in philosophical methodology. Taking such an apparently complex operator to be unanalysable looks unattractive (Nolan 1997a, pp. 273-274), but the position is perhaps not untenable. A better option for the modal fictionalist interested in analysing modality in terms of the modal fiction might be to attempt a non-modal explanation of what is true according to fiction. In any case, this problem, like many problems for modal fictionalism, does not arise for the timid modal fictionalist. For those fictionalists for which it is a problem, however, the unattractiveness and unintuitiveness of taking "According to PW..." or a similar device to be primitive remain a largely unaddressed challenge.

4.5 The Threat from Abstractionism

For modal fictionalism to become the preferred treatment of possible worlds, it must not only be able to perform adequately the tasks assigned to talk of possible worlds, it must also do better than its rivals, or at least possess virtues that those rivals lack. However, modal fictionalism faces a close rival that apparently shares its benefits and avoids some of its vices. Instead of a fiction of possible worlds, some rival views identify possible worlds with certain maximal representations. This sort of "abstractionist" view (following the terminology of Van Inwagen 1986), or "ersatz" view (in the terminology of Lewis 1986) is prima facie committed only to representations, as a modal fictionalist must it seems also be, but the abstractionist is a realist about possible worlds, and thus has prima facie a more straightforward approach than the fictionalist's.

This is particularly true in the case of the platonist modal fictionalist. If the modal fictionalist accepts that the modal fiction is a collection of platonistic propositions, then that very collection of propositions will also do as an abstractionist's "world book": and if the fiction provides a separate description of each possible world (or such a description can be constructed from the resources given), then these complete representations will just be those things which some abstractionists take to be possible worlds. A modal fictionalist may be driven to accept that the fiction is a collection of propositions in response to any of several objections: the worry about artificiality, the worry about incompleteness, or alternatively on more general grounds (given that taking fictions to be collections of propositions is attractive quite apart from considerations about modal fictionalism). (This worry is mentioned in Nolan 1997a p 272.)

Suppose that a modal fictionalist does accept an ontology of propositions rich enough to provide for maximal consistent collections of such propositions. Why then would fictionalism be preferred to abstractionism, or vice versa? Abstractionism would have the advantage of being a more straightforward treatment of normal quantification over possible worlds, since there would be no need to suppose that there is (or should be) a silent "according to the fiction of possible worlds" or "according to the presupposition that there are worlds" governing such possible-worlds talk. Nor would abstractionism face

the technical challenges that fictionalism faces, since all the merely possible worlds would in fact exist; there would be no problem of switching back and forth between fictional and literal discourse. Furthermore, an abstractionist would not need to face the worries of accounting for the "according to the fiction..." operator: the abstractionist's overall theory would need to make room for this operator somewhere, but she could hold out the promise of being able to use modal locutions and talk of possible worlds in its explication without risk of circularity.

One reason that might be offered for preferring modal fictionalism to some form of abstractionism is that abstractionism faces a battery of well-known objections, levelled against it by Lewis 1986, chapter 3. (Rosen 1990 p 328-9 mentions this as a motivation for fictionalism against abstractionism.) It is far from clear that fictionalism avoids these objections, however: and it seems that fictionalism committed to Platonic propositions prima facie faces the same worries about representation, primitive modality, and mysterious ontology. Whether the abstractionist can answer Lewis's objections, and whether the fictionalist can answer or avoid them as well or better, is an issue beyond the scope of this entry. I merely note that the Platonist fictionalist in particular should be cautious in drawing too much comfort from these arguments.

Rosen 1990 suggests another reason. It might be thought that there are good arguments to show that possible worlds, if there are any such things, must be concrete cosmoi like the one in which we inhabit, and cannot be abstract objects, especially abstract objects like collections of Platonic propositions. (He assumes that this has been established for the purposes of his paper on p. 329.) Indeed, if there were arguments to show that collections of propositions were non-starters as candidates to be possible worlds, this would damn the project of abstractionist theories of this form. It is hard to find in Lewis, or elsewhere, arguments that our conception of possible worlds is so tied to their being concrete that we should prefer to believe that there were no merely possible worlds than to believe that they turn out to be abstract objects, however, though Armstrong 1989 (p 46, 49) offers this as a reason for being a fictionalist rather than an ersatzer (abstractionist).

Finally, a modal fictionalist might reject abstractionism because he rejects the associated ontology. This move does not seem open to a Platonist modal fictionalist, since the ontology is one of collections of propositions in both cases. A non-platonist fictionalist, who is happy to rely on fictions construed as collections of marks on paper, or noises in air, or perhaps a combination of these and mental states of speakers and listeners (or writers and readers), can then reject the abstractionist accounts of possible worlds at issue precisely because they are committed to abstract representing entities. Such a fictionalist needs to deal with the worries about artificiality and incompleteness (see above), which arise in more acute forms than face the platonist. He also has a further difficulty in that many accounts of fiction themselves refer to propositions, and are committed to them. The fictionalist who eschews propositions will need to provide an account of fiction and of sentences being true according to fictions compatible with repudiating commitment to propositions. This is not an easy task. However, if it could be carried out, it would be clear that in one respect at least - the respect of ontology - the fictionalist would have a theory with definite advantages over abstractionism. Since ontological concerns are among the primary motivations for modal fictionalism, this is no doubt a path some modal fictionalists will attempt.

4.6 Does Modal Fictionalism Deliver Possible Worlds Semantics?

John Divers in Divers 1995 argues that modal fictionalism cannot deliver the benefits of the standard possible worlds semantics for modal discourse. There is a discussion of Divers's argument in the following supplementary document:

[Supplementary Document: Modal Fictionalism and Possible Worlds Semantics]

4.7 Concern about Concern

Another worry about modal fictionalism is discussed by Rosen 1990 (pp. 349-354): the "argument from concern". An "argument for concern" was originally developed as an objection to the (realist) theory of possible worlds proposed by David Lewis. Lewis claimed that the truth of counterfactual conditional claims could be analysed as the truth of claims about the goings-on in other possible worlds: to take the classic example, the claim "Hubert Humphrey might have won" is true because there is a possible world very similar to ours in which someone much like Hubert Humphrey *did* win. Saul Kripke in Kripke 1980 suggested in a footnote (p. 45) that there was a problem for this view: while Humphrey cares a great deal about the fact that he might have won, he presumably does not care about whether someone a lot like him but who is not him wins in another cosmos. ("Probably, however, Humphrey could not care less whether someone *else*, no matter how much resembling him, would have been victorious in another possible world"). In any case, it is hard to believe that his concern about the first fact is a concern about the second. Examples can of course be multiplied: we often care about modal features of our lives (what could have been, and what would have been), but non-philosophers perhaps seldom even think about whether people much like them have different experiences in different cosmoi, let alone care deeply about such things. So expanded, the "argument from concern" is that the analysis of the truth-conditions of modal statements about objects in our world turns matters we care about into matters we do not care about, and so fails to be a plausible analysis. (Note that Kripke himself does not expand his passing comment in this way).

The analogous "argument for concern" can be run for modal fictionalism (and indeed for almost any account of the truth-conditions of modal claims: see Lewis 1986, pp. 195-197, who argues in part that ersatzers are in the same boat as he is). Humphrey cares about whether he could have won the election, or whether he would have won if some things had been done differently, but it is hard to believe that he cares particularly whether according to a certain story people like him win in other worlds, or even whether according to a certain unusual story he himself wins in other worlds. So it seems implausible to suppose the question of whether or not he could have won is the same question as the question of whether according to the story he does win in certain other worlds.

As Rosen points out in the case of Kripke's objection to Lewis, "this by itself is not a logical objection to the claim that the facts are identical" (p. 349). One may care deeply about something, not realising that it is identical to something else one claims not to care about (just as I might greatly admire the speeches of Cicero, and honestly claim to have no time for the speeches of Tully). Even if the issue of whether

Humphrey could have won is just the issue of whether the modal fiction says that counterparts of Humphrey win at other worlds, this would not mean that this philosophical analysis will be obvious to workaday politicians like Humphrey, nor need it be reflected in his views. Rosen suggests that the objection might have a more ‘pragmatic’ force (p. 350). The objection might be something like this (the way of putting it is not Rosen’s, but I take it the sentiments are): we care about what might have been, and if the modal fictionalist is right, what might have been is a matter of what the fiction says about what goes on in other worlds, and what some complicated story says about other cosmoi is not something we currently have more than an academic interest in, it seems. So if we accept this theory, we should revise what we care about (since we shall think the two come to the same thing): either by becoming as indifferent to modal matters as we are to what story the modal fiction tells - or alternatively by becoming as concerned about the contents of the modal fiction as we currently are about what could have been and would have been, had we acted differently. Either option requires large revisions of our concerns, and it is a cost of a theory to require such revision.

Rosen’s response to this, on behalf of the modal fictionalist, is that this price is worth paying, particularly if we extend our concern to the contents of the fiction, rather than the much harder task of ceasing to care about modal matters of fact. He then goes on to point out that this raises another worry, that of arbitrariness: why care so much about the contents of the modal fiction as opposed, say, to any other story about many worlds and what happens in each? Rosen discusses several possible replies to this question (pp. 352-353), though leaves the final answer open. It seems to this writer, however, that this might not be a particular problem for modal fictionalists: for the question of why we should care about modal facts as opposed to truths expressed with any other conceivable intensional operators looks equally pressing, and any realist who, for instance, analyses modality in terms of the nature of possible worlds will face the question of why we should care about what is true according to the various worlds, as opposed to what is "shtrue" at these worlds, where the "shtrue at" relation is some other relation between worlds and propositions. At the very least, one would want a theory of how our concern for modal truths might be justifiable before seriously worrying about whether the same sort of thing could be said about concern about the content of the modal fiction.

Finally, the whole argument from concern presented above only really gets a hold on the modal fictionalist who thinks that what is the case modally is just a matter of the contents of the modal fiction. There is no need to think this, of course, and timid modal fictionalists will reject it. (Rosen also points out that the modal fictionalist can sidestep the argument if he does not take the fiction to provide the materials for an analysis of modality.) Even strong modal fictionalists can in principle think that there is some sort of analysis or reduction of the modal to what the contents of the fiction are without taking the further step of thinking that the modal facts and facts about the content of the fiction are one and the same. (This would be one way of linking the two, but they might think that the modal facts are constituted by facts about the fiction without thereby being identical with them, for example). So the argument from concern creates problems only for some modal fictionalists.

Bibliography

- Armstrong, D. M. 1989. *A Combinatorial Theory of Possibility*. Cambridge University Press, Cambridge.
- Armstrong, D. M. 1993. "Reply to Lycan", in Bacon, Campbell and Reinhardt 1993: 18-22
- Bacon, J., Campbell, K., and Reinhardt, L. (eds). 1993. *Ontology, Causality and Mind*. Cambridge University Press, Cambridge.
- Baldwin, Thomas. 1998. "Modal Fictionalism and the Imagination". *Analysis*. 58/2: 72-75.
- Bentham, Jeremy. [1959] (1814-1832) "The Theory of Fictions", in Ogden, C. K. 1959. *Bentham's Theory of Fictions*. Littlefield, Adams and Co., Paterson, New Jersey.
- Brock, Stuart. 1993. "Modal Fictionalism: A Response to Rosen". *Mind*. 102/405: 147-150
- Burgess, John P. and Rosen, Gideon. 1997. *A Subject With No Object*. Oxford University Press, Oxford.
- Chihara, Charles. 1998. *The Worlds of Possibility*. Clarendon, Oxford.
- Divers, John. 1999a. "A Genuine Realist Theory of Advanced Modalising". *Mind*. 108: 217-239
- Divers, John. 1999b. "A Modal Fictionalist Result". *Nous*. 33/3: 317-346
- Divers, John. 1995. "Modal fictionalism cannot deliver possible worlds semantics". *Analysis*. 55/2: 81-88
- Field, Hartry. 1989. *Realism, Mathematics and Modality*. Basil Blackwell, Oxford.
- Forbes, G. 1983. "Physicalism, Instrumentalism and the Semantics of Modal Logic". *Journal of Philosohpical Logic* 12: 271-298
- Hale, Bob. 1995a. "A Desperate Fix". *Analysis*. 55/2: 74-81
- Hale, Bob. 1995b. "Modal Fictionalism: A Simple Dilemma". *Analysis*. 55/2: 63-67
- Hinckfuss, Ian. 1993. "Suppositions, Presuppositions, and Ontology". *Canadian Journal of Philosophy*, 23/4: 595-618
- Kripke, Saul. 1980. *Naming and Necessity*. Blackwell, Oxford.
- Lewis, David. 1992. "Critical Notice of Armstrong, D.M. *A Combinatorial Theory of Possibility*". *Australasian Journal of Philosophy*. 70/2: 211-224
- Lewis, David. 1968. "Counterpart Theory and Quantified Modal Logic". *The Journal of Philosophy*. 65/5: 113-126
- Lewis, David. 1986. *On The Plurality of Worlds*. Blackwell, Oxford.
- Lewis, David. 1978. "Truth in Fiction". *American Philosophical Quarterly*. 15/1: 37-46
- Lycan, William G. 1993. "Armstrong's New Combinatorialist Theory of Modality", in Bacon, Campbell and Reinhardt 1993: 3-17
- Lycan, William G. 1994. *Modality and Meaning*. Kluwer, Dordrecht.
- Menzies, Peter and Pettit, Philip. 1994. "In Defence of Fictionalism about Possible Worlds". *Analysis*. 54/1: 27-36
- Merrill, G. H. 1978. "Formalization, Possible Worlds and the Foundations of Modal Logic". *Erkenntnis*. 12: 305-327
- Nolan, Daniel. 1997a. "Three Problems for 'Strong' Modal Fictionalism". *Philosophical Studies*. 87/3: 259-275
- Nolan, Daniel. 1997b. "Impossible Worlds: A Modest Approach". *Notre Dame Journal of Formal Logic*. 38/4: 535-572
- Nolan, D., and O'Leary-Hawthorne, J. 1996. "Reflexive fictionalisms". *Analysis*. 56/1: 26-32
- Nolt, J. 1986. "What Are Possible Worlds?". *Mind*. 95: 432-445

- Noonan, Harold. 1994. "In Defence of the Letter of Fictionalism". *Analysis*. 54/3: 133-139
- Rescher, Nicholas. 1975. *A Theory of Possibility*. Blackwell, Oxford.
- Rosen, Gideon. 1993. "A Problem for Fictionalism About Possible Worlds". *Analysis*. 53/2: 71-81
- Rosen, Gideon. 1990. "Modal Fictionalism". *Mind*. 99/395: 327-354
- Rosen, Gideon. 1995. "Modal Fictionalism Fixed". *Analysis*. 55/2: 67-73
- Routley, Richard. 1980. *Exploring Meingong's Jungle and Beyond*. ANU Departmental Monograph #3, Canberra.
- Stalnaker, Robert. 1976. "Possible Worlds". *Nous*. 10: 65-75
- van Fraassen, Bas C. 1980. *The Scientific Image*. Clarendon, Oxford.
- van Inwagen, Peter. 1986. "Two Concepts of Possible Worlds", in *Studies in Essentialism, Midwest Studies in Philosophy*. XI: 185-213
- Vision, Gerald. 1994. "Fiction and Fictionalist Reductions". *Pacific Philosophical Quarterly*. 74: 150-174
- Walton, Kendall L. 1990. *Mimesis as Make-Believe*. Harvard University Press, Cambridge MA
- Zalta, Edward N. 1997. "A Classically-Based Theory of Impossible Worlds". *Notre Dame Journal of Formal Logic*. 38/4: 640-660

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

abstract objects | actualism | fictionalism | modality, metaphysics of | possible worlds

Daniel Nolan
dpnolan@syr.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 14, 2002
Content last modified: May 14, 2002

Stanford Encyclopedia of Philosophy
Notes to Modal Fictionalism

Notes

1. Examples include cases where P itself contains nested modal locutions, or where it contains names or other rigid designators subject to counterpart-theoretic paraphrase.

2. Other biconditionals which might be needed besides the ones associated with the sentential operators of necessity and possibility are ones connecting counterfactual statements and "closeness" relations between possible worlds, and biconditionals connecting probability values and measures on sets of possible worlds. De re modality may call for even more biconditionals, linking de re modal claims to claims about which counterpart relations, or trans-world identities, hold according to the fiction.

3. A modal fictionalist need not take possible worlds to be concrete, spatio-temporal (albeit fictional) objects: one could easily be a fictionalist whose fiction of possible worlds describes them as being abstract objects of some sort. This "abstract object" approach has not received much support in print, however.

4. This blanket approach would require rejection of some widely held basic principles of modal logic: that "possibly P" is true whenever "P" is, for example, or that "actually P" is true for all true P (if "actually" is to be treated as a modal operator, as it often is taken to be). This view would also run into difficulties if we retain the standard interdefinition of "necessarily P" with "not-possibly not-P" and "possibly P" with "not-necessarily not-P". If both "necessarily P" and "possibly not-P" were without truth value, excluded middle would fail, since "possibly not-P or not-possibly not-P" would fail to be true. If statements prefixed with a modal operator were all treated as false, on the other hand, then if the standard interdefinitions applied, violations of the law of non-contradiction could be generated. One counterexample to non-contradiction could be derived from the falsity of "possibly not-P or not-possibly-not P". Since this is false, its negation is true (plausibly), and hence, by application of a De Morgan law, "not-possibly not-P and not-not-possibly not-P" follows, which is a contradiction. It is, of course, open to a fictionalist to reject both the basic modal inferences mentioned and the standard interdefinition of possibility and necessity. Such a theorist could still allow that these moves and the interdefinition preserve truth *according to the fiction of modality*, which may salvage enough of the basic principles and the interdefinition for a fictionalist’s purposes. ss

5. Jeremy Bentham, (Works VII 76-9, 1959, pp. 53-58) advocates fictionalism about the "qualities" of impossibility and possibility, as well as probability and improbability. It is clear that he understood these "qualities" as epistemic rather than alethic ones, but his general approach suggests he might have said the same about the ascription of alethic modalities. He would then count as an early fictionalist about

modality itself.

6. The concern is exacerbated if the fiction of possible worlds is intended to provide an analysis or reductive explanation of modal truth (i.e. if "strong modal fictionalism" is intended). For the modal truths are not up to us, it seems. Whether or not it is possible for it to be simultaneously raining and not raining (in the same location, at the same time, in the same respect, etc. etc.) does not seem to be anything we have any control over. Furthermore, some propositions being necessary and others possible, some de re modal claims being true and others false, does not seem to depend on whether or not humans bothered to tell stories about possible worlds. Finally, if things like natural laws and causation are modal matters (e.g. if causation is a matter of the obtaining of certain counterfactuals), the artificiality of modality would seem to imply that even natural laws and causation are, to some extent, artificial. This anti-realism will be found objectionable by many (though not by all: Rescher reasons from conceptualism about modality to idealism virtually across the board through these sorts of connections (Rescher 1975)). Taking modal facts to depend on accidents of our storytelling seems to make the modal truths too artificial.

7. Though Rosen does not say so in quite so many words, he seems to have particularly in mind David Lewis and the account of fiction he offered in Lewis 1979. Lewis's account of fiction is one in which what is true in a fiction is what is true in certain possible worlds determined by the text of that fiction. Rosen points out that such accounts would have trouble dealing with fictions about the space of possible worlds as a whole (Rosen 1990, pp. 345-346). If the realist's analysis of truth in fiction does not satisfactorily deal with what is true according to fictions like PW, the realist cannot claim to have an account of "According to PW..." that the fictionalist lacks (p. 346). So if Rosen is right, the realist cannot analyse away the locution that the fictionalist takes as primitive, so the realist cannot claim to be better off in choice of primitives. Rosen recognizes that the realist might eventually be able to account for even this class of fictions using possible worlds, and should that day come the realist will be able to reclaim the advantage of avoiding the "according to PW" primitive.

The sort of treatment offered by Lewis 1979 notoriously faces difficulties in dealing with impossible fictions more generally, especially ones in which it is not straightforward to divide the fictions into consistent subfictions. But there are world-based treatments of fiction that are arguably more successful at dealing with these problems: Routley 1990 (Ch. 7) is one example. Another way of developing a more adequate account of impossible fictions is to stay classical but supplement the theory of Lewis 1979 with impossible worlds (classical accounts of impossible worlds include Nolan 1997b and Zalta 1997, among others). So Rosen's *tu quoque* is shaky, at best. In any case, an analysis of fiction which relied in part on some modal resource or other need not look very much like the traditional approaches employing possible worlds, and it does not seem that Rosen has given us reason to suppose that such analyses of fiction could not be made to work.

Daniel Nolan
dpnolan@syr.edu

First published: May 14, 2002
Content last modified: May 14, 2002

Stanford Encyclopedia of Philosophy;
Supplement to Modal Fictionalism

A Persisting Problem For Fictionalism About Possible Worlds[*]

At least since Gideon Rosen's 1990, the theory that talk about possible worlds should be interpreted as a useful fiction has been taken seriously by many philosophers as a way we might be able to have, in Lewis's phrase, "paradise on the cheap" (Lewis 1986, p. 136).

The point of fictionalism about possible worlds is that if it is successful, it enables one to use the conceptual resources of possible-world semantics without needing the ontological baggage that realism about such worlds carries. The strategy is to employ an account of possible worlds (e.g. David Lewis's)[1], but insist that the account is only a useful fiction. A modal claim will then be true iff its translation into possible-world semantics is true in the fiction (let us call the account treated as fiction by Rosen the "Lewis story"[2]). So, for example, P is possible iff, in the Lewis story, there exists a possible world where P is true. Of course, this does not imply that in reality there is a possible world such that P is true (at least as far as the fictionalist is concerned), any more than its being true in the Arthur legend that King Arthur defeated the Emperor Augustus implies that Arthur (if he existed) really defeated Augustus in battle.

A problem was raised for modal fictionalism independently by Rosen (1993) and Stuart Brock (1993). This problem, to put it in a perhaps oversimplified form, is that since in the Lewis story it is true at each possible world that there are many worlds, it is, *according to the fiction* (or, according to the story), true at every possible world that there exist many worlds. But the truth conditions offered by a Rosen-style fictionalist state that if according to the fiction a proposition P is true at every world, then P is (in fact) necessarily true. Therefore, it follows in this case that necessarily there are several possible worlds, and so since (necessarily P) implies P, there in fact exist several possible worlds, which is not something the fictionalist should countenance.

This problem is discussed in a paper by Peter Menzies and Philip Pettit (1994) who claim to show that the objection is not fatal, and that there is a coherent and plausible fictionalist way of avoiding it. Menzies and Pettit propose that the fictionalist operator needs to be applied to every quantifier in a modal claim when translating into possible-world talk, either by employing a "prefixing" strategy or an "indexing" strategy (for details see their paper). So while it is not true (for the Menzies/Pettit fictionalist) that "necessarily there exist several worlds", what is in fact true in virtue of there being several possible worlds that exist at each fictional world is that "necessarily, according to the fiction there exist several worlds", which is not a claim incompatible with fictionalism at all. However, they recognize a *further* problem for fictionalism: a problem involving a so-called "modal dangler", or a modal claim that

concerns the truth (or in this case, possible truth) of the theory itself.[3] This supplement will consist of a discussion of this problem, an examination of the solution proposed for it by Menzies and Pettit, and an explanation of why the proposed solution is so far unsatisfactory.

This further objection is more or less as follows: The Lewis story (or the "hypothesis of the plurality of worlds" (PW) as it is referred to by Brock and Menzies/Pettit) will, according to the Lewis story, be something that is true at every world in the Lewis story. This causes the modal fictionalist a real problem, because given the standard paraphrase into possible world semantics, something is necessary iff it is the case at every possible world. Since, in the Lewis story, the Lewis story is correct at every possible world, then the fictionalist who is prepared to translate modal claims into claims about possible worlds (albeit fictional ones) is forced to admit that the truth of the Lewis story is necessary, and thus the Lewis "story" is no mere fiction, but actually correct. (Necessarily-P implies P). However, it is precisely PW, or the Lewis story, that the fictionalist wants to argue is *false* in the actual world. It appears that the fictionalist (at least one who follows the strategy described) is committed to the truth, and even necessary truth, of the very thesis fictionalism was developed to deny.

Menzies and Pettit suggest an alternative way of constructing one's possible-worlds fictionalism to avoid any commitment to the necessary truth of the Lewis story. They start by pointing to the special status that the actual world has among worlds. Other worlds are (according to our fictionalist) purely fictional entities, so what is true at[4] them is completely determined by the fiction. What is true at the actual world, however, is not determined by what (modal-) fictional stories we choose to tell about it. The truth conditions for the "modal danglers" (statements concerning the modal status of the modal fictionalist theory itself) are to reflect this, on the Menzies-Pettit proposal. According to the proposed paraphrase for the part of modal discourse concerned with modal danglers,

> Possibly P if and only if ‘P’ holds at the actual world or at one of the other worlds in PW.
>
> Necessarily P if and only if ‘P’ holds at the actual world and at all of the other worlds posited in PW" (1994, p. 36)

Note that these truth conditions do not say "if ‘P’ holds in the actual world *on the PW account*". These truth conditions involve what the actual world is really like, and not merely what it is like in a given fictional account. Thus this view allows us to conclude that PW, or the Lewis story, is not necessary, and so we are not forced to admit that it is true.

This modification works to defuse Brock's specific objection about the expressibility of the falsehood of PW, but it does not defuse objections based on counterfactual situations where we would want to say that the Lewis story is not correct. For instance, consider the quite plausible counterfactual "If there were a king of France, still PW would be false". A fictionalist should want to accept this: after all, changes in the political situation in France will not produce changes in the fundamental truths of the metaphysics of modality (France is not *that* important). However, consider this modal claim: "It is possible that there be a present king of France and also that PW be false". Now, it is not true of the actual world that there is a

king of France, and it is not true of any of the (supposedly fictional) possible worlds that PW is false there (after all, in the Lewis story, the Lewis story (or PW) holds at every world). The above modal claim about the king and PW is thus false. For a counterfactual with a contingent antecedent to be true it must be possible that the antecedent and consequent be true together. Since it is impossible that there be a king of France and PW be false, it seems that the initial plausible counterfactual cannot be held to be true by the fictionalist.

This sort of objection, of course, is a general one: nothing hinges on the mentioning of the present king of France, as almost any counterfactual will do. The general point is that a fictionalist should be committed, on the grounds of the irrelevance of most changes in the world to the truth of fictionalism, to saying that even if the world had turned out slightly differently, still the Lewis story would have been merely fiction: and this is something that cannot be coherently maintained by a fictionalist who accepts the Menzies/Pettit truth conditions for modal statements.

One might be tempted to think that the entire problem of the modal status of modal danglers (i.e. those statements that concern the truth or modal status of the theory itself) is one that can be ignored, or at least easily avoided, by the fictionalist. After all, the modal danglers are a pretty special class of truths, unlikely to make an impact on science or semantics or other fields in which possible worlds are useful. One way of avoiding the problem of "modal danglers" is to restrict fictionalism so that it does not consider the modal status of the theory itself, but that the theory functions as some sort of "metalanguage" immune to the modal interpretation available for all other areas of discourse. There are at least two good reasons for resisting this strategy, however.[5]

Firstly, it seems *ad hoc*. Possibility seems to be the same thing in both the sentences "Possibly PW is false" and "Possibly Phlogiston Theory is false": merely using the word "metalanguage" does not provide explanatory power. Secondly, it leaves the "possibly" operator in modal danglers as a primitive. If we are going to have primitive unanalyzed modalities besides that of "truth in fiction" in our theory anyway, why bother with an analysis of modality in terms of truth-in-fiction? Why not say that all of the modal operators reflect primitive features of the world and abandon the attempt to analyse modality in terms of a modal fiction?

It is too early to say that a Rosen-style modal fictionalism must be abandoned -- after all, merely showing that it has not yet succeeded in analysing the full range of modal claims (e.g. modal danglers) does not imply that it cannot do so, but only that it has not. However, even given the Menzies/Pettit modifications to Rosen's original proposal, modal danglers remain a difficulty for modal fictionalism. Those eager to promote fictionalist analyses of possible worlds talk owe us an answer to this problem.[6]

Notes

[*] Another version of the central objection in this piece to the Menzies/Pettit proposal has been developed independently by David Lewis.

[1.] This is the account of possible worlds that is treated as the fiction about possible worlds by Gideon Rosen, (1990) and by Menzies and Pettit (1994), and by Brock (1993).

[2.] I prefer to use the term "Lewis story" it when discussing the specific proposal that takes Lewis' theory as the fiction rather than using the term "PW" or "the fiction of a plurality of worlds", as obviously one could employ a story about the existence of many possible worlds without it agreeing with the story told by David Lewis. (One could have trans-world identity rather than counterparts, for example).

[3.] See Menzies and Pettit (1994), p. 35, for this "Last Difficulty". They attribute this further objection to Stuart Brock.

[4.] At_1, for those who wish to be precise in the manner described in Menzies and Pettit (1994), p. 33

[5.] These reasons for rejecting this move are the same sorts of reasons that Brock (1993, pp. 149-150) outlines for not arbitrarily ruling "Necessarily there are many worlds" out of consideration by a fictionalist theory.

[6.] I am especially indebted to Peter Menzies and Philip Pettit for encouragement as well as helpful comments and discussion.

Daniel Nolan
dpnolan@syr.edu

Return to Modal Fictionalism

First published: May 14, 2002
Content last modified: May 14, 2002

Stanford Encyclopedia of Philosophy
Supplement to Modal Fictionalism

Rosen's Incompleteness Worry

The "incompleteness problem" discussed by Rosen (1990, pp. 341-345) can be expressed as follows: there are some modal issues (and corresponding issues about the nature of possible worlds) that a realist may well be silent on -- not because they believe there is no answer, but rather because they believe themselves ignorant of the answer. A fictionalist who treats the realist's theory as a fiction, on the other hand, will be silent upon the same issues -- but this can lead to a more serious problem. Rosen, who uses as his fiction the theory that David Lewis proposes as fact, takes as an example an issue Lewis 1986 is silent on: the maximum "size" of possible worlds (or in particular, the maximum number of non-overlapping physical objects in a single world). Rosen's worry is this: according to the view he develops, (the numbering is Rosen's, and the details are from his example on p. 342 of 1990):

> (10) There might have been κ non-overlapping physical objects.
>
> if and only if
>
> (10f) According to PW, there is a [world] containing κ non-overlapping physical objects.

(10f) is not true, since PW is silent on the issue. So (10) is not true. But the same thing happens for the negation of (10):

> (11) It is not the case that there might have been κ non-overlapping physical objects.
>
> is plausibly true if and only if
>
> (11f) According to PW, no [world] contains κ [non-overlapping] physical objects.

If (11) is really the negation of (10), then classically one of them must be true. If one of them is literally false, then its negation should be true. Rosen is inclined to accept (on the fictionalist's behalf) that they are genuine contradictories, and both lack a truth-value. He points out, however, two difficulties with this: firstly, his proposal makes (10f) and (11f) truth-valueless too, when they seem clearly false (at least if "according to PW.." works like "according to the fiction ..." operators standardly do); and secondly, the disjunction of (10) and (11) is true and a logical truth, so we have the truth of a disjunction without the truth of either disjunct. (Not that this is unknown in the treatment of truth-value gaps, as Rosen points out.) Such logical revision might be thought a high price to pay.

In addition, Rosen rightly points out (p. 342), the fictionalist has settled the question of whether it is true that there might have been κ many non-overlapping physical objects: it is not true. Those who were inclined to think that we are in ignorance of this piece of modal information (or any other piece of modal information about which the story is "silent" in a similar way) will not like this result. The fictionalist does not make room for modal ignorance to the same extent the realist does.

Rosen's discussion of this worry is unhappy in several places. The first is that (11f) is not his only official translation of (11), assuming "might have been" is to be treated as "possibly" in this context: controposing his biconditional for possiblity (with the appropriate substitutions for P) yields:

> It is not the case that there might have been κ non-overlapping physical objects
>
> if and only if
>
> (11f*) It is not the case that, according to PW, there is a world containing κ non-overlapping physical objects.

(11f*) yields the result that (11) is simply true, if we interpret "according to PW" in the usual way (and not the way which makes "According to PW, Q" truthvalueless if Q is "not determinately settled as true or false by the theory PW", whatever that might mean in this context). Since (11) *is* the contradictory of (10), and (10) is false by the lights of Rosen's original proposal, this should not be surprising. This result has unpleasant consequences of its own, of course, since accepting this theory would commit one to denying the interdefinability of possibility and necessity. Normally, "possibly P" is taken to be equivalent to "not-necessarily not-P", and "necessarily P" is taken to be equivalent to "not-possibly not-P". But the second will fail in this case (and the first will fail in other similar cases). For (11f) is the appropriate correlate of the claim that "necessarily, there are not κ non-overlapping physical objects", and this claim is standardly taken to be equivalent to (11). For the reasons Rosen gives, however, the correlate of (11f) is not true, whereas (11) is, due to the truth of (11f*).

Denying the interdefinability of necessity and possibility in the standard way is perhaps not as big a modification to our logic and semantics as introducing truth-value gaps. It will still be seen as very unattractive by many. There is another modification to the basic theory which might be made to deal with these cases, which is a different way of implementing the spirit of Rosen's proposed strategy here. Rosen says the fictionalist can modify his theory by "declaring that in general when the paraphrase P* of a modal claim is not determinately settled as true or false by the theory PW, the modal claim P is to lack a truth-value" (p. 343). What is puzzling about Rosen's implementation of this proposal is that he takes it that this implies that the "According to PW, P*" claim has to be truthvalueless when PW neither says P* or its negation. This has the further unwelcome consequence that "According to PW" will not function like "according to the fiction..." operators are standardly thought to. A more natural way of going, surely, is to treat "According to PW ..." in the standard way (i.e. it is just false that "According to PW, P*", when PW is neither committed to P* or to its negation), but to provide for truth-value gaps for the modal claims by restricting the fictionalist biconditional. Instead of the general scheme:

> P iff According to PW, P*,

the fictionalist could instead accept the clumsier:

> if P, then According to PW, P*, and if not-P, then According to PW, not-P*, and if According to PW, P*, then P, and if According to PW, not-P*, then not-P

and further stipulate that P is truthvalueless iff neither according to PW, P*, nor according to PW, not-P*.

When the fiction says something about whether or not P*, which will be the usual case, this longer set of conditions will permit the fictionalist to move back and forth from modal language to talk of possible worlds in the usual way: it is only when PW is silent about the relevant issue that the corresponding modal claim suffers a lack of truth-value. In modifying the central biconditional, it is true that the fictionalist is sacrificing some of the elegance of the original theory, though the unpleasant looking list of conditions can be rewritten so that they have more of the appearance of a minor alteration of the original:

> P iff According to PW, P*, (unless PW is silent about P*, in which case P is truth-valueless).

This still has the feature that a disjunction may be true without either of its disjuncts being true, and other features common to many treatments of truth-value gaps, and it still has the feature that the theory delivers definite answers of a sort for certain modal matters (i.e. that P has no truth-value) when it might have seemed that we imagined we were merely ignorant: it is still an approach with many of the features of Rosen’s amendment. It however lacks the most objectionable feature possessed by Rosen’s proposal of keeping the biconditional intact and declaring that "According to PW ..." is gappy.

In any case, Rosen’s incompleteness worry arises only for those modal fictionalist theories which do indeed admit that the modal fiction is silent on some relevant issues. Some are indeed likely to (and Rosen’s is one), but it should be remembered that this does not seem to be an unavoidable feature of such theories. One could attempt to specify enough about the content of the modal fiction so that it settled, at least in principle, all of the relevant issues. (It may be difficult to tell how they are settled, of course, if the content of the fiction depends, e.g., on facts about the arrangement of the actual world, but this is not a problem of completeness, but only of modal epistemology). In particular, timid fictionalists can stipulate the content of the modal fiction in terms of the modal truths (so, e.g. the fiction is to represent that at some world, Q*, just in case possibly Q). This sort of specification would only be circular if the status of the modal truths was to be analysed in terms of the content of the fiction, but the timid fictionalist has no such pretensions. While the timid fictionalist’s fiction may leave some questions unsettled (how many indiscernible worlds there are, for example), the timid fictionalist should be able to make the fiction determinate to the extent that all of the correlates of modal claims are represented, by brute stipulation if necessary. It may be harder for a strong modal fictionalist, but it has not been shown to be impossible.

In summary, then, while Rosen's incompleteness concerns may well not give rise to the problems he alleges, there are difficulties (or at least departures from orthodoxy) which will be faced by fictions which are incomplete in the sorts of ways that, for instance, Rosen's candidate is.

Daniel Nolan
dpnolan@syr.edu

Return to Modal Fictionalism

First published: May 14, 2002
Content last modified: May 14, 2002

Stanford Encyclopedia of Philosophy
Supplement to Modal Fictionalism

Modal Fictionalism and Possible Worlds Semantics

John Divers in (Divers 1995) argues that "Modal fictionalism cannot deliver possible worlds semantics" for modal logics. Since Kripke, it has become standard to do the formal semantics for languages with modal operators (like necessity and possibility operators) by providing models of worlds and of objects in those worlds, and counting modal statements as true according to the relevant models if the worlds in those models meet certain conditions (e.g. where *p* is an atomic claim, "necessarily *p*" is true at a world *w* in the model just in case all of the possible worlds "accessible" from *w* are worlds at which *p* holds). On one construal of what benefits this sort of semantics offers, modal fictionalism arguably cannot do as well as its realist rivals.

The first thing that must be done is to be clear on what is meant by "possible worlds semantics". There is a common useage of that expression in which providing such a "semantics" for e.g. a language containing modal operators is a matter of providing certain sorts of semantic models for it. Formal specifications of these models can provide the basis for exploring the formal features of the system (such as whether it is decidable, or whether the consequence relation for the system is compact). Another benefit is that by specifying models with respect to which the axioms and rules of the system are demonstrably complete, we provide a mechanism for producing counter-examples (counter-models), as well as another method, besides e.g. an axiomatic proof procedure, for seeing what conclusions we might be able to draw from a given premise set. The capacity for constructing counter-examples is very important, since we are often as interested in which arguments turn out to be invalid as in which arguments are valid. Historically, providing "world" models for modal logic was also useful for bringing out dimensions of variation among modal logics: by putting various constraints on the accessibility relation, the behaviour of the necessity and possibility operators could be varied in ways it was hard to imagine before possessing the device of models.

However, none of these benefits require possible worlds *per se*, even for so-called "possible worlds semantics" in this sense (see Lewis 1986, pp. 17-20). The formal semantics officially consists of constructing models using sets of various kinds and functions from those sets to propositions or sentences (or other objects representing propositions or sentences: other sets do nicely). Certain objects in these models are often called "worlds", but this is just a convenient label: the blander word "indices" is sometimes recommended by logicians wanting to avoid the appearance of trucking with metaphysics. Interpreting a formal language with modal operators as having its truth conditions given by which indices stand in which relations need not bring possible worlds in the full-blown sense into the picture at all, any more than the models of modal logics that define truth in terms of geometrical features of cylinders commits users of these logics to the belief that there are special modal cylinders out there, about which we constantly talk when we talk of what has to be or what might have been.

Divers intends much more by a "possible worlds semantics" than merely providing a class of models for a set of axioms and rules of inference in the way mentioned above: he agrees that one can "interpret a modal language extensionally by invoking a model-theoretic approach in which the indices ('worlds') have (intuitively) no modal reference" (Divers 1995, p. 81), and so would agree that he has more in mind than this limited formal task. However, the sense in which providing a "possible worlds semantics" comes to no more than providing traditional Kripke-syle models using indices which are adequate for the job described above does seem to be what most logicians mean by the expression, and I suspect it is what many philosophers of modality mean as well. If so, then Divers's charge is misleadingly stated, since he expects much more from a possible worlds semantics. With this important caveat in mind, let us look at what Divers thinks modal fictionalism cannot deliver and modal realism can.

For Divers, "A possible worlds semantics is a semantic theory which has two components: (i) an interpretation which translates some range of ordinary modal sentences into a medium involving quantification over worlds, and (ii) a theory of validity (typically model-theoretic) that determines formally whether or not arguments formulated in the interpreting medium are valid." (Divers 1995 p. 80) He thinks this "translation" is literally translation, and not just an interpretation for model-theoretic purposes: for the connection between modal sentences and sentences quantifying over possible worlds is meant to be an *analytic* one (p. 80). The other important feature Divers thinks that "possible worlds semantics" has is that the claims about possible worlds are *extensional* claims, which allows the validity of inferences involving them to be determined by the usual methods of "the fully extensional predicate calculus" (p. 81).

Divers is of course allowed to stipulate how he will use his terms as he likes, but insofar as this is intended as an explication of what a possible worlds semantics (in some sense richer than the minimal logican's sense) should be, a point of controversy may be noted. Realists who think that quantification over possible worlds gives insight into the nature of modal truth need not think that such quantificational claims are *analytically*connected with modal claims. They may instead think that possible worlds provide the best theoretical explanation of the truth of modal claims, or are what make those claims true in a way that is to be discovered by synthetic methods. Indeed, in light of the fact that some *prima facie* competent deployers of modal vocabulary (including thoughtful philosophers) do not accept the truth of the possible-worlds "translations" of modal claims that they do accept, we should be reluctant to think that the former are obvious analytic consequences of, let alone synonymous with, the modal claims. Divers cites Graeme Forbes's arguments in Forbes 1985 that we should assume that modal claims and the associated possible worlds claims are synonymous if we are to explain the behaviour of our modal claims by appeal to the possible worlds claims: this is however one side of a controversy between groups that both take themselves to be employing "possible worlds semantics", and employing *worlds* in these semantics, rather than merely employing indicies in formal models of modal languages.

As for the second part of Divers's definition, it is not entirely unambiguous what it is for a claim to be "extensional": there are many ways of drawing the extensional/intensional divide. Divers does not say exactly what he means, but he seems to be using the word in the common sense in which a sentence is extensional if co-referring (or more generally, co-designating) expressions can be substituted *salva*

veritate. Sentences consisting of an "According to PW ..." operator followed by a sub-sentence are thus not extensional, since substituting another sub-sentence (materially) equivalent in truth-value does not always preserve truth. In Divers's example, "There is a world in which there are red dragons" and "1 = 0" will be judged by the fictionalist to have the same truth value, but the expressions "According to PW, there is a world in which there are red dragons" and "According to PW, 1 = 0" will not have the same truth-values (Divers 1985, p. 85). There is another sense in which a claim might be said to be extensional: it is extensional if it does not depend for its truth-value at a world on the truth-value of claims in any other possible world. In this sense, of course, the fictionalist's "According to PW ..." sentences are still not extensional, but neither are the realist's possible worlds "translations" of everyday modal locutions, and possible-worlds "translations" of modal locutions should not in general be thought to be extensional. (One might be tempted more by this view if, for example, one thought that extensionality or intensionality were preserved under meaning equivalence, and accepted that the relevant equivalences were equivalences of meaning, as Divers takes the realist to). There is nothing wrong with Divers's useage, but the ambiguity is possibly confusing. This will become relevant when the question of what advantage extensionality would bring is discussed below.

But let us agree that a "possible worlds semantics" for current purposes has the features Divers ascribes to it. At this point many, realists and fictionalists alike, will already reject the project of attaining a "possible worlds semantics" in Divers's sense (though many would be concerned if they could not have a possible worlds semantics for their favourite formal modal systems, in the more minimal sense). However, let us also grant for the sake of discussion that a possible worlds semantics is desirable, and it would count against a theory if it did not furnish us with one. The next thing that is worth noting is that some possible modal fictionalist theories will straightforwardly provide us with such a semantics. *Broad* modal fictionalists (see main text) take both the talk committed to possible worlds and many modal claims (claims of necessity, and claims about the possible truth of actual falsehoods, etc.) to be true only according to a certain fiction. It is open to such modal fictionalists to accept straightforwardly that the modal sentences were analytically equivalent to their paraphrases in terms of possible worlds. The predicate calculus could then be used to determine validity for modal inferences, and they could have a "possible worlds semantics" for modal claims in as rich a sense as Divers could wish. Divers does not consider this option for modal fictionalists, though his paper is presented as a discussion of Rosen 1990, and Rosen does not mention broad modal fictionalists. Rather, Rosen only considers modal fictionalists who take the modal claims to be equivalent in truth-value to claims about possible worlds prefixed with "according to PW ..." operators. The broad modal fictionalists I have been considering take the equivalence to be between modal claims and *unprefixed* claims about possible worlds, and so arguably fall outside the scope of Divers's discussion. They are worth mentioning even so, since Divers quickly moves from talking about Rosen's particular presentation to a discussion of "modal fictionalism" in general.

Let us then concentrate on the remaining views which come under Divers's attack: those views which take modal statements to be equivalent in truth-value to prefixed claims about possible worlds i.e., claims of the form "According to PW ...". Do these views have to eschew "possible worlds semantics", in Divers's sense? Divers claims that they do, since the analysis in terms of talk about possible worlds provided for modal statements is an analysis in terms of an intensional context, the "according to PW ..."

operator, into which propositions with the same truth-value cannot be substituted *salva veritate*. For those modal fictionalists who do offer this as an analysis, Divers is right that their analysis is not an extensional one. What is not clear is that such fictionalists cannot nevertheless have the advantages associated with extensional semantic treatments of modal language.

Divers says "the distinctive attraction of possible worlds semantic theories is that they combine the virtues of analyticity and extensionality in the treatment of modal languages". Analyticity is presumably held to be a virtue because it is analyticity that provides possible worlds semantic theories "their claim to be the right and genuinely explanatory theories of the semantic features of modal languages", and this is backed in part by an appeal to the arguments of Forbes 1985. Extensionality, the stumbling block for fictionalists according to Divers, is apparently advantageous because (or mainly because) it allows us to determine the validity of "ordinary modal arguments" by "assessing their validity by application of the methods of the fully extensional predicate calculus" (p. 85). Divers also indicates that this advantage is a matter of it being correct according to the theory that "the validity of intuitively valid modal arguments can be demonstrated by first-order methods" (p. 86); he points out, rightly, that merely having this true according to the fiction would not be the same as being able to literally claim it. Divers stresses that it is the combination of these two advantages that is the distinctive benefit fictionalists must forgo (p. 85).

While the modal fictionalists we are now considering do not take ordinary language modal claims to have extensional translations in the language of possible worlds, they may still have a claim to the combined virtue with which Divers is concerned in another, albeit slightly more indirect, way. If they do, then Divers has failed to make his case, since he maintains not only that there is a feature that realist theories have (or perhaps better, can have) that modal fictionalist positions do not (or cannot), but also that the relevant feature is, *prima facie* at least, an advantage. So how may modal fictionalists who take the equivalence in meaning to be one between modal claims and "According to PW ..." claims also take advantage of first-order methods to determine validity for arguments couched in the form of admittedly intensional statements beginning "According to PW ..."?

The most obvious way for a modal fictionalist to attempt to secure this advantage is to insist both that the relevant fictionalist biconditionals are analytic, and that PW is closed under first-order consequence: so that when s follows from r, if "According to PW, r" is true then "According to PW, s" is true, and for any propositions p_1, p_2 ... p_n, when "According to PW, p_1", "According to PW, p_2" ... "According to PW, p_n" are all true, then "According to PW, p_1, p_2, p_3, ... , and p_n" will be true too. Then when one has a collection of premises all suitably prefixed, one will be able to determine whether a conclusion prefixed with "According to PW ..." follows from the premises plus general truths about the modal fiction (such as its closure under first-order consequence) by first-order methods, by seeing whether the proposition embedded in the conclusion follows from the propositions embedded in the premises. This might not quite capture all of the conclusions derivable (conclusions which follow simply from general claims about the fiction will follow in such cases regardless of the specific prefixed premises, for example), but in the ordinary cases when a realist would appeal to the putative analytic equivalences to test the modal arguments for validity, the fictionalist will be able to have recourse to this procedure, which reduces the problem to that of checking statements to see if one is a first-order consequence of the others. The fictionalist’s justification for this will *ex hypothesi* be an appeal to the analytic interderivability of claims

about what is true according to the fiction and modal claims, and so this first-order procedure is available and justifiable in virtue of the analytic connection between the modal claims and the claims involving quantification over possible worlds (albeit that those claims quantify over possible worlds in the scope of "According to PW" operators). True, this advantage will not be because of any extensional features of the "According to PW ..." claims, and so Divers is right in one sense to say that this method is not "extensional". Since the fictionalist who combines analyticity of the relevant equivalences with the above method for securing the use of a first-order decision procedure for validity can have exactly the same advantage that Divers claims for the combination of "the virtues of analyticity and extensionality", Divers's case that there is some *advantage* that realists can claim over fictionalists does not seem to have been made out.

In a later paper (Divers 1999b) Divers sets out to prove a "safety result" which would guarantee the use of extensional reasoning in the scope of an "according to PW" operator for a modal fictionalist. Divers 1999b can thus be seen as a response to Divers's own 1995 challenge, showing how the fictionalist can secure the advantage which Divers 1995 suggests is not available to the modal fictionalist. Not all modal fictionalists would be happy with the resources Divers employs to achieve his result. In particular, Divers employs a primitive necessity operator (p 334), and so his proof as it stands would not be available to an advocate of strong modal fictionalism. (Divers himself notes that the strategy is unavailable to those modal fictionalists who seek a reduction of the modal operators). Fictionalists not concerned to analyse modality, however, may find in Divers 1999b a sufficient answer to Divers 1995.

So to summarise: Divers's claim that "modal fictionalism cannot deliver possible worlds semantics" need not worry modal fictionalists. In the first place, the sense of "possible worlds semantics" Divers employs is not the sense in which many practitioners working with modal logic or the language of possible worlds are most concerned with. Secondly, even given Divers's richer concept of "possible worlds semantics", a conception rich enough that many realists about possible worlds would also not wish to suppose that they should have such a semantics, some modal fictionalists can have such a semantics in exactly Divers's sense (the broad modal fictionalists). Finally, for the remaining modal fictionalists prepared to postulate the appropriate analytic connections, it is possible, at least for all Divers has said, that one might have a semantic treatment of modal claims which treated them as analytically equivalent to claims in the language of possible worlds (albeit prefixed ones) *and* has the virtue of being able to employ first-order methods to determine the validity of many arguments employing premises and conclusions prefixed with "According to PW ..." operators, and so achieve the same advantage Divers claims for analytic, extensional translations. If this is so, then even if there is a sense in which these fictionalists do not have a "possible worlds semantics", they have something just as good, as far as the advantages Divers is concerned with go. Divers 1995 does however provide a challenge for modal fictionalists: if they do want "possible worlds semantics" in Divers's rich sense, they need to provide the details of how it is that they are entitled to the corresponding advantages: perhaps through some analytic equivalence of the sort explored in this note, or perhaps through a system like that of Divers 1999b, or perhaps by some other strategy. (Note that strong modal fictionalists who take their theory to not be descriptive of current usage are unlikely to be tempted by the claim of analytic equivalence and will have to demur from Divers's timid fictionalist proposal, so they at least face an open question if they want the equivalent of a "possible worlds semantics" in Divers's rich sense).

Daniel Nolan
dpnolan@syr.edu

Return to Modal Fictionalism

First published: May 14, 2002
Content last modified: May 14, 2002

Abstract Objects

It is widely supposed that every object falls into one of two categories: Some things are concrete; the rest abstract. The distinction is supposed to be of fundamental significance for metaphysics and epistemology. The present article surveys a number of recent attempts to say how it should be drawn.

Introduction

The abstract/concrete distinction has a curious status in contemporary philosophy. It is widely agreed that the distinction is of fundamental importance. But there is no standard account of how the distinction is to be explained. There is a great deal of agreement about how to classify certain paradigm cases. Thus it is universally acknowledged that numbers and the other objects of pure mathematics are abstract, whereas rocks and trees and human beings are concrete. Indeed the list of paradigms may be extended indefinitely:

ABSTRACTA	CONCRETA
Classes	Stars
Propositions	Protons

Concepts	The electromagnetic field
The letter A	Stanford University
Dante's *Inferno*	James Joyce's copy of Dante's *Inferno*
...	...

The challenge remains, however, to say what underlies this alleged dichotomy. In the absence of such an account, the philosophical significance of the contrast remains uncertain. We may know how to classify things as abstract or concrete by appeal to "intuition". But unless we know what makes for abstractness and concreteness, we cannot know what (if anything) hangs on the classification.

Historical Remarks

The contemporary distinction between abstract and concrete is not an ancient distinction. Indeed, there is a strong case for the view that despite occasional anticipations, it plays no significant role in philosophy before the 20th century. The modern distinction bears some resemblance to Plato's distinction between Forms and Sensibles. But Plato's Forms were supposed to be causes *par excellence*, whereas abstract objects are normally supposed to be causally inert in every sense. The original "abstract"/"concrete" distinction was a distinction among words or terms. Traditional grammar distinguishes the abstract noun "whiteness" from the concrete noun "white" without implying that this linguistic contrast corresponds to a metaphysical distinction in what they stand for. In the 17th century this grammatical distinction was transposed to the domain of ideas. Locke speaks of the general idea of a triangle which is "neither Oblique nor Rectangle, neither Equilateral, Equicrural nor Scalenon; but all and none of these at once," remarking that even this idea is not among the most "abstract, comprehensive and difficult" (Essay IV.vii.9). Locke's conception of an abstract idea as one that is formed from concrete ideas by the omission of distinguishing detail was immediately rejected by Berkeley and then by Hume. But even for Locke there was no suggestion that the distinction between abstract ideas and concrete or particular ideas corresponds to a distinction among objects. "It is plain, …" Locke writes, "that General and Universal, belong not to the real existence of things; but are Inventions and Creatures of the Understanding, made by it for its own use, and concern only signs, whether Words or Ideas" (III.iii.11).

The abstract/concrete distinction in its modern form is meant to mark a line in the domain of objects. So conceived, the distinction becomes a central focus for philosophical discussion only in the twentieth century. The origins of this development are obscure. But one crucial factor appears to have been the breakdown of the allegedly exhaustive distinction between the mental and the material that had formed the main division for ontologically minded philosophers since Descartes. One signal event in this development is Frege's insistence that the objectivity and a priori of the truths of mathematics entail that numbers are neither material beings nor ideas in the mind. If numbers were material things (or properties of material things), the laws of arithmetic would have the status of empirical generalizations. If numbers were ideas in the mind, then the same difficulty would arise, as would countless others. (Whose mind contains the number 17? Is there one 17 in your mind and another in mine? In that case, the appearance of a common mathematical subject matter is an illusion.) In *The Foundations of Arithmetic* (1884), Frege

concludes that numbers are neither external 'concrete' things nor mental entities of any sort. Later, in his essay "The Thought" (Frege 1918), he claims the same status for the items he calls thoughts -- the senses of declarative sentences -- and also, by implication, for their constituents, the senses of subsentential expressions. Frege does not say that senses are "abstract". He says that they belong to a "third realm" distinct both from the sensible external world and from the internal world of consciousness. Similar claims had been made by Bolzano (1837), and later by Brentano (1874) and his pupils, including Meinong and Husserl. The common theme in these developments is the felt need in semantics and psychology as well as in mathematics for a class of objective (i.e., non-mental) supersensible entities. As this new "realism" was absorbed into English speaking philosophy, the traditional term "abstract" was enlisted to apply to the denizens of this "third realm".

The Way of Negation

Frege's way of drawing the distinction is an instance of what Lewis (1986) calls the **Way of Negation**. Abstract objects are defined as those that lack certain features possessed by paradigmatic concrete things. Nearly every explicit characterization in the literature has this feature. There are, however, several significant difficulties with this approach, at least in its most familiar implementations.

According to Frege's explicit account, the items in the "third realm" are non-mental and non-sensible. But it is unclear what it means to call an object mental or mind-dependent; and to the extent that the notion is intelligible, it is quite unclear whether abstract objects in general satisfy the condition. It is commonly supposed, for example, that the game of chess is an abstract entity (Dummett 1973). But there is certainly a sense in which the game would not have existed were it not for the mental activity of human beings. So at least one sort of mind-dependence would appear to be compatible with abstractness. Moreover, it has sometimes been maintained that the paradigmatic abstract entities -- mathematical objects, universals -- exist only as ideas in the mind of God. The view may be outlandish; but is it a view according to which abstract entities do not exist? Or is it rather a view according to which certain abstract entities are also mind-dependent? Insofar as the latter interpretation is not straightforwardly contradictory, the definition of "abstract" should not require mind-independence.

Perhaps more importantly, Frege's identification of the abstract with the realm of non-sensible non-mental things entails that unobservable physical objects such as quarks and electrons should be classified as abstract entities. But this is at odds with standard usage, and almost certainly with Frege's intention.

The Non-Spatiality Criterion

Contemporary purveyors of the Way of Negation standardly amend Frege's criterion by requiring that abstract objects be **non-spatial** or **causally inefficacious** or both. Indeed, if any characterization of the abstract deserves to be regarded as the standard one, it is this: An abstract entity is a non-spatial (or non-spatiotemporal) causally inert thing. But this standard characterization presents a number of perplexities.

Consider the requirement that abstract objects be non-spatial or non-spatiotemporal. Some of the paradigms of abstractness are non-spatiotemporal in a straightforward sense. It makes no sense to ask where the cosine function is. Or if it does make sense to ask, the only sensible answer is that it is nowhere. Similarly, it makes little sense to ask when the Pythagorean theorem came to exist. And if it does make sense to ask, the only sensible answer is that it has always existed, or perhaps, that it does not exist 'in time' at all. These paradigmatic abstracta have no non-trivial spatial or temporal properties. They have no spatial location, and they exist nowhere in particular in time. But consider the game of chess. Some philosophers take the view that chess is like a mathematical object in these respects. But that is certainly not the most natural view. The natural view is that chess was invented at a certain place and time (though it may be hard to say exactly where or when); that before it was invented it did not exist at all; that it was imported from India into Persia in the 7th century; that it has changed in various respects over the years, and so on. The only reason to resist this natural description would appear to be the thought that since chess is clearly an abstract object (it's not a physical object, after all!), and since abstract objects do not exist in spacetime (by definition!), chess must resemble the cosine function in its relation to space and time. However, one might with equal justice regard the case of chess and other "artificial" abstract entities as a counterexample to the view that abstract objects in general possess only trivial spatial and temporal properties.

This is not necessarily ground for abandoning the non-spatiotemporality criterion. Even if there is a sense in which some abstract entities possess non-trivial spatiotemporal properties, it might still be said thought that concrete entities 'exist in spacetime' *in a distinctive way*, and that abstract entities may be characterized as items that fail to exist in space and time in the manner characteristic of concrete objects.

The paradigmatic concrete objects generally occupy a relatively determinate spatial volume at each time at which they exist, or a determinate volume of spacetime over the course of their existence. It makes sense to ask of any such object, "Where is it now and how much space does it occupy?", even if the answer must in some cases be somewhat vague. By contrast, even if the game of chess is somehow "implicated" in space and time, it makes no sense to ask how much space it now occupies -- or if it does make sense to ask, the only sensible answer is that it occupies no space at all (which is not to say that it occupies a spatial point.) And so it might be said: An object is abstract if it fails to occupy anything like a determinate region of space (or spacetime).

This promising suggestion faces two sorts of difficulty. First, according to some interpretations of quantum mechanics, microscopic physical objects fail to occupy anything like a determinate region of space. If we consider an isolated proton whose position has not been measured for some time, the question "Where is it now and how much space does it occupy?" will have no straightforward answer. And yet no one would suggest that an unobserved proton is an abstract entity. Second, it is not out of the question that certain items that are standardly regarded as abstract may nonetheless occupy determinate volumes of space and time. It is generally agreed that sets and functions are abstract entities. So consider the various sets composed from Peter and Paul: {Peter, Paul}, {{Peter}, {Peter, Paul}}, etc. The question, "Where are these things and how much space do they occupy?" does not arise in the normal course of inquiry. Moreover, many philosophers will be inclined to say that either the question makes no sense, or the answer is a simple "Nowhere. None." But this would appear to be another unreflective

application of the unpersuasive inference noted above. In this case: Sets are abstract; abstract objects do not exist in space. So sets must not exist in space. But as before, there is reason to doubt the cogency of such an inference. Let it be granted that pure sets are like the cosine function: located nowhere in space a nd nowhere in particular in time. Is there a principled objection to the view that impure sets exist where and when their members do? It is not unnatural to say that a set of books is located on a certain shelf in the library. So why not say that the sets containing Peter and Paul exist wherever and whenever Peter and Paul themselves exist, and that in general an impure set exists where and when its spatiotemporally located ur-elements are located? To be sure, nothing in set theory forces us to say this. But the applications of set theory to the concrete domain are not inconsistent with this manner of speaking. So, while it may be clear that the impure sets are abstract and not concrete, it is quite unclear whether they fail to exist in space in much the same sense in which paradigmatic concreta exist in space. This suggests that it may have been a mistake from the start to suppose that the distinction between concrete and abstract is at bottom a matter of spatiotemporal locatedness.

The Causal Inefficacy Criterion

The most widely accepted version of the Way of Negation has it that abstract objects are distinguished by their causal inefficacy. Concrete objects (whether mental or physical) have causal powers; numbers a nd functions and the rest make nothing happen. There is no such thing as causal commerce with the game of chess. And even if impure sets do in some sense exist in space, it is easy enough to believe that they make no distinctive causal contribution to what transpires. Peter and Paul may have effects individually; and they may have effects together which neither has on his own. But these joint effects are naturally construed as effects of two concrete objects acting jointly, or perhaps as effects of their mereological aggregate (itself a paradigm concretum), rather than as effects of some set-theoretic construction. (Suppose Peter and Paul together tip a balance. If we entertain the possibility that this event is caused by a set, we shall have to ask which set caused it: the set containing just Peter and Paul? Some more elaborate construction based on them? Or perhaps the set containing the molecules that compose Peter and Paul? This proliferation of possible answers suggests that it was a mistake to credit causal powers to sets in the first place.)

There are no decisive intuitive counterexamples to this account of the abstract/concrete distinction. The chief difficulty is rather conceptual. The causal relation, strictly speaking, is a relation among events. If we say that the rock caused the window to break, what we mean is that some event involving the rock caused the breaking. If the rock itself is a cause, it is a cause in some derivative sense. But this derivative sense has proved elusive. The rock's hitting the window is an event in which the rock "participates" in a certain way, and it is because the rock participates in events in this way that we credit the rock itself with causal efficacy. But what is it for an object to participate in an event? Suppose John is thinking about the Pythagorean Theorem and you ask him to say what's on his mind. His response is an event: the utterance of a sentence; and one of its causes is the event of John's thinking about the theorem. Does the Pythagorean Theorem "participate" in this event? There is surely *some* sense in which it does. The event consists in John's coming to stand in a certain relation to the theorem, just as the rock's hitting the window consists in the rock's coming to stand in a certain relation to the window. But we do not credit

the Pythagorean Theorem with causal efficacy simply because it participates in this sense in an event which is a cause. The challenge is therefore to characterize the distinctive manner of "participation in the causal order" which distinguishes the concrete entities. This problem has received relatively little attention. There is no reason to believe that it cannot be solved. But in the absence of a solution, this standard version of the Way of Negation must be reckoned unsatisfactory.

The Way of Example

In addition to the Way of Negation, Lewis identifies three main strategies for explaining the abstract/concrete distinction. According to the **Way of Example**, it suffices to list paradigm cases of abstract and concrete entities in the hope that the sense of the distinction will somehow emerge. If the distinction were primitive and unanalyzable, this might be the only way to explain it. But as we have remarked, this approach is bound to call the interest of the distinction into question. The abstract/concrete distinction matters because abstract objects as a class appear to present certain general problems in epistemology and the philosophy of language. It is supposed to be unclear how we come by our knowledge of abstract objects in a sense in which it is not unclear how we come by our knowledge of concrete objects (Benacerraf 1973). It is supposed to by unclear how we manage to refer determinately t o abstract entities in a sense in which it is not unclear how we manage to refer determinately to other things (Benacerraf 1973, Hodes 1984). But if these are genuine problems, there must be some account of why abstract objects as such should be especially problematic in these ways. It is hard to believe that it is simply their primitive abstractness that makes the difference. It is much easier to believe that it is their non-spatiality or their causal inefficacy or something of the sort. It is not out of the question that the abstract/concrete distinction is fundamental, and that the Way of Example is the best we can do by way of elucidation. But if so, it is quite unclear why the distinction should make a difference.

The Way of Conflation

According to the **Way of Conflation**, the abstract/concrete distinction is to be identified with one or another metaphysical distinction already familiar under another name: as it might be, the distinction between sets and individuals, or the distinction between universals and particulars. There is no doubt that some authors have used the terms in this way. But this sort of conflation is relatively rare nowadays. As most philosophers use the term, a claim to the effect that sets (or universals) are the only abstract objects would amount to a substantive metaphysical thesis in need of substantive defense.

The Way of Abstraction

The most important alternative to the Way of Negation is what Lewis calls the **Way of Abstraction**. According to a longstanding tradition in philosophical psychology, abstraction is a distinctive mental process in which new ideas or conceptions are formed by considering several objects or ideas and omitting the features that distinguish them. One is given a range of white things of varying shapes and

sizes; one ignores or "abstracts from" the respects in which they differ, and thereby attains the abstract idea of whiteness. Nothing in this tradition requires that ideas formed in this way represent or correspond to a distinctive class of objects. But it might be maintained that the distinction between abstract and concrete objects should be explained by reference to the psychological process of abstract ion or something like it. The simplest version of this strategy would be to say that an object is abstract if it is (or might be) the referent of an abstract idea, i.e., an idea formed by abstraction.

So conceived, the Way of Abstraction is wedded to an outmoded philosophy of mind. But a related approach has gained considerable currency in recent years. Crispin Wright (1983) and Bob Hale (1987) have developed an account of abstract objects that takes leave from certain suggestive remarks in Frege (1884). Frege notes (in effect) that many of the singular terms that refer to abstract entities are formed by means of functional expressions. We speak of *the shape of* an object, *the direction of* a line, *the number of* books. Of course many singular terms formed by means of functional expressions denote ordinary concrete objects: "the father of Plato", "the capital of France". But the functional terms that pick out abstract entities are distinctive in the following respect: W here '$f(a)$' is such an expression, there is typically an equation of the form

$$f(a) = f(b) \text{ if and only if } a\ R\ b,$$

where R is an equivalence relation. (An equivalence relation is a relation that is reflexive, symmetric and transitive.) For example,

> The direction of *a* = the direction of *b* iff *a* is parallel to *b*.
>
> The number of *F*s = the number of *G*s iff there are just as many *F*s as *G*s.

Moreover, these equations (or abstraction principles, as they are sometimes called) appear to have a special semantic status. While they are not strictly speaking definitions of the functional expression that occurs on the left, they would appear to hold in virtue of the meaning of that expression. To understand the term "direction" is (in part) to know that "the direction of a" and "the direction of b" refer to the same entity if and only if the lines a and b are parallel. Moreover, the equivalence relation that appears on the right hand side of the equation would appear to be semantically and perhaps epistemologically prior to the functional expression on the left (Noonan 1978). Mastery of the concept of a direction presupposes mastery of the concept of parallelism, but not vice versa.

The availability of abstraction principles meeting these conditions may be exploited in several ways to yield an account of the distinction between abstract and concrete objects. When 'f' is a functional expression governed by an abstraction principle, there will be a corresponding concept $\boldsymbol{K}f$ such that X is $\boldsymbol{K}f$ iff for some y, $x = f(y)$. The simplest version of this approach to the Way of Abstraction is then to say that X is an abstract object if (and only if?) X is an instance of some kind $\boldsymbol{K}f$ whose associated functional expression 'f' is governed by a suitable abstraction principle. m

This simple account is liable to a number of objections.

- As we have noted, pure sets are paradigmatic abstract objects. But it is not clear that they satisfy the proposed criterion. According to naïve set theory, the functional expression ‘set of’ is indeed characterized by a putative abstraction principle.

  The set of *F*s = the set of *G*s iff for all *x*, *x* is *F* iff *x* is *G*.

  But this principle is inconsistent, and so fails to characterize an interesting concept. In contemporary mathematics, the concept of a set is not normally introduced by abstraction. It remains an open question whether something like the mathematical concept of a set can be characterized by a suitably restricted version of the naïve abstraction principle. But even if such a principle is available, it is unlikely that the epistemological priority condition will be satisfied. (That is, it is unlikely that mastery of the concept of set will presuppose mastery of the equivalence relation that figures on the right hand side.) It is therefore uncertain whether the Way of Abstraction so understood will classify the objects of pure mathematics as abstract entities (as it presumably must).
- As Dummett (1973) has noted, in many cases the standard names for paradigmatically abstract objects do not assume the functional form to which the definition adverts. Chess is an abstract entity. But we do not understand the word "chess" as a synonymous with an expression of the form "*f*(*x*)" where "*f*" is governed by an abstraction principle. Similar remarks would seem to apply to such things as the English language, social justice, architecture, Audrey Hepburn's smile. (In this last case, we are to imagine that Hepburn's smile is essentially connected to its bearer. Someone else might smile just like Hepburn, but her smile would not be *Hepburn's* smile.) If so, the Fregean approach undergenerates: At best it may be said to characterize a special case of the general concept of an abstract entity.
- As formulated, the account would appear to admit of counterexamples. A mereological fusion of concrete objects is itself a concrete object. But the concept of a mereological fusion is governed by a principle with all the marks of an abstraction principle:

  The fusion of the *F*s = the fusion of the *G*s iff the *F*s and *G*s cover one another.
  (The *F*s cover the *G*s iff every part of every *G* has a part in common with an *F*.)

  Or consider: A train is a maximal string of railroad carriages, all of which are connected to one another. We may define a functional expression, ‘the train of *x*’, by means of an "abstraction" principle:

  The train of *x* = the train of *y* iff *x* and *y* are carriages and *x* and *y* are connected.

  We may then say that *x*is a train iff for some carriage *y*, *x*is the train of *y*. The simple account thus yields the consequence that trains are to be reckoned abstract entities.

It is unclear whether these objections apply to the more sophisticated abstractionist proposals of Wright and Hale. This Fregean approach to the abstract/concrete distinction is clearly promising. But like most other approaches to explaining the distinction, it has not yet assumed its final form. Definitive assessment would therefore be premature.

Further Reading

Zalta (1983) is an axiomatic theory of abstract objects. Putnam (1975) makes the case for abstract objects on scientific grounds. Field (1980) and (1989) make the case against abstract objects. Bealer (1993) and Tennant (1997) present a priori arguments for the necessary existence of abstract entities. The dispute over the existence of abstracta is reviewed in Burgess and Rosen (1997).

Bibliography

- Bealer, George. (1993), "Universals", *Journal of Philosophy* 90 (1).
- Benacerraf, Paul. (1973), "Mathematical Truth", *Journal of Philosophy* 70.
- Bolzano, Bernard. (1837) *Wissenschaftslehre*, translated as *Theory of Science,* edited with an introd. by Jan Berg, trans., Burnham Terrell, D. Reidel, 1973.
- Brentano, Franz (1874). *Psychologie vom empirischen Standpunkt*. Translated as *Psychology from an Empirical Standpoint*, edited by Oskar Kraus ; English edition edited by Linda L. McAlister, translated by Antos C. Rancurello, D.B. Terrell, and Linda L. McAlister, Routledge, 1995.
- Burgess, John and Gideon Rosen (1997). *A Subject with No Object*, Oxford.
- Dummett, Michael. (1973) *Frege: Philosophy of Languag*e, Duckworth.
- Field, Hartry. (1980) *Science without Numbers* , Princeton.
- Field, Hartry. (1989) *Realism, Mathematics and Modality*, Basil Blackwell.
- Frege, Gottlob. (1884) *Die Grundlagen der Arithmetik*, translated by J. L. Austin as *The Foundations of Arithmetic*, Oxford, 1959.
- Frege, Gottlob. (1918) "Der Gedanke: Eine Logische Untersuchung", translated by A. Quinton and M. Quinton as "The Thought: A Logical Enquiry" in Klemke, ed., *Essays on Frege*, Chicago, 1968.
- Hale, Bob. (1987) *Abstract Objects*, Basil Blackwell.
- Hodes, Harold (1984). "Logicism and the Ontological Commitments of Arithmetic", *Journal of Philosophy* 81.
- Lewis, David. (1986), *On the Plurality of Worlds*, Basil Blackwell.
- Noonan, Harold. (1978), "Count Nouns and Mass Nouns", *Analysis* 38.4.
- Putnam, Hilary. (1975) "Philosophy of Logic", in his *Mathematics, Matter and Method*, Cambridge.
- Tennant, Neil. (1997) "On the Necessary Existence of Numbers," *Nous*, 31.
- Wright, Crispin. (1983), *Frege's Conception of Numbers as Objects*, Aberdeen University Press.
- Zalta, Edward. (1983) *Abstract Objects: An Introduction to Axiomatic Metaphysics*, D. Reidel.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Frege, Gottlob | Frege, Gottlob: logic, theorem, and foundations for arithmetic | individual | object | properties

Gideon Rosen
grosen@princeton.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 19, 2001
Content last modified: July 19, 2001

Frege's Logic, Theorem, and Foundations for Arithmetic

Frege formulated two distinguished formal systems and used these systems in his attempt both to express certain basic concepts of mathematics precisely and to derive certain mathematical laws from the laws of logic. In his *Begriffsschrift* of 1879, he developed a second-order predicate calculus and used it both to define interesting mathematical concepts and to state and prove mathematically interesting propositions. However, in his *Grundgesetze der Arithmetik* of 1893/1903, Frege added (as an axiom) what he thought was a distinguished logical proposition (Basic Law V) and tried to derive the fundamental theorems of various mathematical (number) systems from this proposition. Unfortunately, not only did Basic Law V fail to be a logical proposition, but the resulting system proved to be inconsistent, for it was subject to Russell's Paradox.

Although the inconsistency in Frege's *Grundgesetze* is widely known, it is not very well known that a deep theoretical accomplishment can be extracted from his work. The *Grundgesetze* contains all the essential steps of a valid proof (in second-order logic) of the fundamental propositions of arithmetic from a single consistent principle. This consistent principle, known in the literature as "Hume's Principle", asserts that for any concepts *F* and *G*, the number of *F*-things is equal to the number *G*-things if and only if there is a one-to-one correspondence between the *F*-things and the *G*-things. In the *Grundgesetze*, Frege used Basic Law V to derive Hume's Principle, but the derivations of the fundamental propositions of arithmetic from Hume's Principle do not essentially require Basic Law V. So by setting aside the derivation of Hume's Principle from the inconsistent Basic Law V and focusing on Frege's proofs of the basic propositions of arithmetic, his theoretical accomplishment emerges much more clearly, for his work shows us how to prove the Dedekind/Peano axioms for number theory from Hume's Principle in second-order logic. This achievement, which involves some remarkably subtle chains of definitions and logical reasoning, has become known as Frege's Theorem. [See Boolos (1990), p. 268.]

The principal goals of this essay are: (1) to review in some detail the essential features of Frege's logical systems, (2) to work through the derivations involved in Frege's Theorem, and (3) to frame the most important philosophical questions that arise in connection with this theorem. In addition, we hope to prepare students of Frege to read his original work (in translation) and to prepare the reader to understand a number of excellent articles in the secondary literature on Frege's work.

To accomplish these goals, we presuppose only a familiarity with the first-order predicate calculus. We show how to extend this language and logic to include the most salient features of Frege's second-order

predicate calculus, his theory of concepts, and his theory of extensions. Our discussion will be largely based upon material drawn from Frege's three principal published works:

- *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* (‘Concept Notation: A formula language of pure thought, modelled upon that of arithmetic’), 1879
- *Die Grundlagen der Arithmetik* (‘The Foundations of Arithmetic’), 1884
- *Grundgesetze der Arithmetik* (‘Basic Laws of Arithmetic’), 1893/1903

We will refer to these works with boldfaced abbreviations of their German titles: **Begr**, **Gl** and **Gg**, respectively. Those readers already familiar with parts of Frege's texts may wish to skip the discussion of that material.

[Link to Complete Table of Contents (With Listing of Subsections)]

---

§1: Frege's Predicate Calculus and Theory of Concepts

In this section, we describe the language and logic of Frege's predicate calculus. We explain his function-argument analysis of atomic sentences and his definition of concepts in terms of functions, give examples of his ‘concept script’, and discuss the Rule of Substitution in his logic. We also show how Frege's Rule of Substitution corresponds to a comprehension principle for concepts in second-order logic, and we introduce and explain λ-notation to help us distinguish open formulas and complex names of concepts. Readers who are already familiar with these ideas may wish to skip ahead to Section 2.

The Language

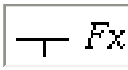
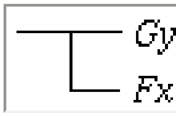
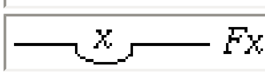
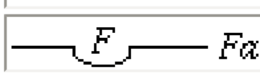
In **Begr**, Frege invented the predicate calculus. It will soon become clear that the language and logic of his predicate calculus are ‘second-order’. The language included not only the variables *x*,*y*,*z*, ... , which

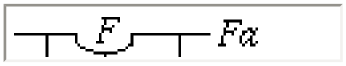
range over *objects*, but also included the variables *f*,*g*,*h*, ... , which range over *functions*. Frege rigidly distinguished objects from functions and so we may think of these variables as ranging over separate, mutually exclusive domains. Frege took functional application ‘*f*(*x*)’ as the principal operation for forming complex names of objects in his language. The expression ‘*f*(*x*)’ denotes the object to which the function *f* maps the object *x*. Frege called the object *x* the ‘argument’ of the function *f* and called *f*(*x*) the ‘value’ of the function. Since Frege also recognized two special objects he called *truth-values* (The True and The False), he defined a *concept* to be any function that always maps its arguments to truth-values. For example, whereas ‘x^2 +3’ and ‘father-of(*x*)’ denote ordinary functions, the expressions ‘Happy(*x*)’ and ‘*x* > 5’ denote concepts. The former denotes a concept which maps any object that is happy to The True and all other objects to The False; the latter denotes a concept that maps any object that is greater than 5 to The True and all other objects to The False. Given that concepts like *being happy* and *being greater than* 5 map their arguments to truth values, the atomic sentences of Frege's language, such as ‘Happy(*b*)’ and ‘*4* > 5’, become *names* of truth-values.

In what follows, we use the symbols *F*,*G*, ... as variables ranging over concepts and we often write ‘*Fx*’ (instead of ‘*F*(*x*)’) to express the claim that concept *F* maps *x* to The True. When this claim is true, Frege would say that *x falls under* the concept *F*.

When *f* is a function of two arguments *x* and *y* and *f* always maps its pair of arguments to a truth value, Frege would say that *f* is a relation. We shall use the expression ‘*Rxy*’ (or sometimes ‘*R*(*x*,*y*)’) to assert that the relation *R* maps *x* and *y* (in that order) to The True. In what follows, we shall sometimes write the symbol that denotes a mathematical relation in the usual ‘infix’ notation; for example, ‘>’ denotes the greater-than relation in the expression ‘*x* > *y*’.

Now that we have explained Frege's analysis of the atomic statements ‘*Fx*’ and ‘*Rxy*’ familiar to modern students of logic, we turn next to the more complex statements of his language. Frege developed his own graphical notation for asserting complex statements involving negations, conditionals, and universal quantification. If we ignore the fact that Frege used Gothic letters as variables of quantification, certain letters as bound variables in names of courses-of-values, and certain other letters as placeholders in the names of functions, then Frege's notation for the logical notions ‘not’, ‘if-then’, ‘every’ and ‘some’ can be described in the following table:

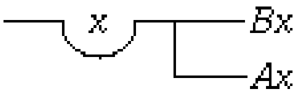
Logical Notion	**Modern Notation**	**Frege-Style Notation**
It is not the case that *Fx*	$\neg Fx$	
If *Fx* then *Gy*	$Fx \rightarrow Gy$	
Every *x* is such that *Fx*	$\forall xFx$	
Some *x* is such that *Fx*	$\neg\forall x\neg Fx$, i.e., $\exists xFx$	
Every *F* is such that *Fa*	$\forall F\, Fa$	

Some F is such that Fa	$\neg\forall F \neg Fa$, i.e., $\exists F\ Fa$	

So, for example, whereas a modern logician would symbolize the claim ‘All *A*s are *B*s’ as:

$\forall x(Ax \rightarrow Bx)$

Frege would symbolize this claim as follows:



However, since Frege's notation was never adopted as a standard, we shall instead use the more familiar modern notation in the remainder of this essay. [See Beaney (1997, Appendix 2) and Furth (1967) for more detailed introduction to Frege's notation.] We shall assume that the reader is familiar with the fact that negations (‘$\neg\varphi$’) and conditionals (‘$\varphi \rightarrow \psi$’) can be used to define the other molecular formulas such as conjunctions (‘φ & ψ’), disjunctions (‘φ v ψ’), and biconditionals (‘$\varphi \equiv \psi$’). Moreover, it is important to mention that Frege took identity statements of the form ‘$x = y$’ as primitive in his language. Whereas ‘$2^2 = 4$’ names The True, ‘$2^2 = 3$’ names The False. The statement form ‘$f(x) = y$’ plays an important role in Frege's axioms and definitions. Note finally that since Frege allowed quantification over both objects and functions, the language of his predicate calculus becomes ‘second-order’.

The Logic

Frege's logic consisted of basic axioms and rules of inference that governed the permissible inferences within his system. His axioms included familiar axioms of propositional logic, second-order predicate logic, and the logic of identity. For example, where φ and ψ are any formulas and ‘*a*’ is any object term and ‘*P*’ is any concept term, then the following were among the basic laws of Frege's system:

- $\varphi \rightarrow (\psi \rightarrow \varphi)$
- $(\forall x Px) \rightarrow Pa$
- $(\forall F\ Fa) \rightarrow Pa$
- $a = b \rightarrow \forall F(Fa \equiv Fb)$

[Frege's most well-known codification of these laws occurs in **Gg I**, §47; however, the above laws are first introduced in **Gg I**, §§18, 20, 25, and 20, respectively.] We shall simplify our discussion in what follows by assuming that the usual axioms of the modern second-order predicate calculus apply to Frege's system. These are essentially the same as the axioms for the first-order predicate calculus, except for the addition of laws for the second-order quantifiers $\forall F$ and $\exists F$ which correspond to the laws governing the first-order quantifiers $\forall x$ and $\exists x$.

Although these axioms of Frege's logic are familiar to us, the rules of inference in Frege's system are not as familiar. The reason is that the rules govern not only his graphical notation for molecular and quantified formulas, but also his special purpose symbols, such as certain lowercase letters used as placeholders, certain Gothic and letters used as bound variables, and various other signs of his system we have not yet mentioned. Since these will play no role in the discussion that follows, we shall again simplify our discussion by assuming that the usual rules of the modern second-order predicate calculus apply to Frege's system. Again, these are essentially the same as the rules for the first-order predicate calculus, except for the addition of new rules for the second-order quantifiers that correspond to the generalization and instantiation rules (i.e., introduction and elimination rules) for the first-order quantifiers.

The Rule of Substitution

There is, however, one distinguished rule of Frege's system that will play an important role in what follows, namely, his Rule of Substitution. For the purposes of this discussion, we may formulate the rule in the following somewhat simplified manner:

> **Rule of Substitution** (Simplified Version):
> In any statement of the form ...*Fx*... (in which the variable *F* is free) which is derivable as a theorem of logic, we may substitute any open formula $\varphi(x)$ (with the free variable *x*) for all the occurrences of the atomic formula *Fx* in ...*Fx*... .

To see this rule in action, first consider the following theorem of (Frege's) second-order predicate logic:

(A) $\forall x(Fx \equiv Fx)$.

Now Frege's Rule of Substitution not only allows us to substitute the atomic formula '*Ox*' (which might represent the claim '*x is odd*') for the formula *Fx* to derive the true statement $\forall x(Ox \equiv Ox)$, but also allows us to substitute complex formulas with a free variable *x* for '*Fx*'. So, for example, we are allowed substitute the formula '*Ox* & *x* > 5' ('*x* is odd and *x* is greater than 5') for '*Fx*' in (A) to derive the following from (A):

(B) $\forall x(Ox \,\&\, x > 5 \equiv Ox \,\&\, x > 5)$

Inferences such as this will be valid no matter what complex formula with *x* free we substitute for *Fx* in our universal claim (A). This is what justifies Frege's Rule of Substitution.

In what follows, we will assume that the Rule of Substitution can be generalized to relations, so that we can uniformly replace the formula *Rxy* (in a theorem of logic with *R* free) by a complex formula $\varphi(x,y)$ (in which both *x* and *y* are free).

The Theory of Concepts

The Rule of Substitution has rather powerful consequences. It implies that there exists a concept corresponding to every open formula with a free variable *x*. To see that this is a consequence of the rule, note that it follows from (A) by existential generalization that:

$$\exists G \forall x(Gx \equiv Fx)$$

Frege's Rule of Substitution now allows us to substitute any formula with free variable *x* for *Fx*. In other words, every instance of the following Comprehension Principle for Concepts is derivable in Frege's system:

> **Comprehension Principle for Concepts**:
> $\exists G \forall x(Gx \equiv \varphi(x))$,
> where $\varphi(x)$ is any formula which has *x* free and which has no free *G*s.

Similarly, from the theorem of logic:

$$\forall x \forall y(Rxy \equiv Rxy)$$

one can generalize and then use the Rule of Substitution to derive the following Comprehension Principle for Relations:

> **Comprehension Principle for Relations**:
> $\exists R \forall x \forall y(Rxy \equiv \varphi(x,y))$,
> where $\varphi(x,y)$ is any formula with *x* and *y* free and which has no free *R*s.

Although Frege didn't explicitly formulate these Comprehension Principles, they constitute a very important generalization about his system that reveals its underlying theory of concepts and relations. We can see these principles at work if we return to the example used above. The following is an instance of the Comprehension Principle for Concepts and so constitutes a theorem of Frege's system:

$$\exists G \forall x(Gx \equiv Ox \ \& \ x > 5)$$

This asserts: there exists a concept *G* such that for every object *x*, *x* falls under *G* if and only if *x* is odd and greater than 5. We can see, therefore, that Frege's Rule of Substitution essentially treats an open formula like ‘*Ox* & *x* > 5’ as if it were a name of a complex concept. Similarly, the following is an instance of the Comprehension Principle for Relations:

$$\exists R \forall x \forall y(Rxy \equiv Ox \ \& \ x > y)$$

This asserts the existence of a relation that objects x and y bear to one another just in case the complex condition Ox & $x > y$ holds.

Logicians nowadays typically distinguish the open formula $\varphi(x)$ from the corresponding name of a concept. They use the notation $[\lambda x\ Ox\ \&\ x > 5]$ as the name of the concept *being an object x such that x is odd and x is greater than* 5 (or, more naturally, ‘being odd and greater than 5’). The term-forming operator ‘λx’ (‘being an x such that’) combines with a formula $\varphi(x)$ in which x is free to produce $[\lambda x\ \varphi(x)]$. The λ-expression is a name of the concept expressed by the formula. This notation can be extended for relational concepts. The expression:

$$[\lambda xy\ Ox\ \&\ x > y]$$

names the 2-place relation *being an x and y such that x is odd and greater than y*. So we will use expressions of the more general form $[\lambda xy\ \varphi(x,y)]$ in what follows. The reader should note, however, that insofar as λ-expressions are taken to be predicates, Frege would not have accepted the idea that these are complete expressions that name concepts. But we shall not discuss Frege's reasons for this in the present essay.

This λ-notation is governed by the following simple logical principle known as λ-Conversion. Let $\varphi(x)$ be any formula in which the variable x is free, and let $\varphi(y/x)$ be the result of substituting the variable y for x everywhere in $\varphi(x)$. Then the principle of λ-Conversion is:

λ-Conversion:
$$\forall y([\lambda x\ \varphi(x)]y \equiv \varphi(y/x))$$

This asserts that an object y falls under the concept $[\lambda x\ \varphi(x)]$ if and only if $\varphi(y/x)$ holds. So, using our example, the following is an instance of λ-conversion:

$$\forall y([\lambda x\ Ox\ \&\ x > 5]y \equiv Oy\ \&\ y > 5)$$

This asserts that an object y falls under the concept *being odd and greater than* 5 if and only if y is odd and greater than 5. Note that when the variable y is instantiated to some object term, the resulting instance of λ-Conversion is a biconditional. Some logicians call the rule of inference derived from the right-to-left direction of such biconditionals ‘λ-Abstraction’. For example, the inference from

$$O6\ \ \&\ \ 6 > 5$$

to

$$[\lambda x\ Ox\ \&\ x > 5]6$$

is justified by λ-Abstraction.

The principle of λ-Conversion can be generalized, so that it covers relations as well:

$$\forall z \forall w([\lambda xy\ \varphi(x,y)]zw \equiv \varphi(z/x, w/y))$$

The reader should construct an instance of this principle using our example [λ*xy Ox* & *x* > y].

To reiterate, then, Frege's Rule of Substitution allows us to instantiate $\varphi(x)$ for the free variable *F* in theorems of logic as if $\varphi(x)$ were a λ-expression and constituted a name of a concept. In what follows, we shall make use of this λ-notation. Indeed, λ-notation is required if we are to give a more precise formulation of the Rule of Substitution; the precise formulation of the rule for concepts is:

> **Rule of Substitution**:
> The λ-expression $[\lambda x\ \varphi(x)]$ may be uniformly substituted for the occurrences of the variable *F* in any theorem of logic containing *F* free.

(The formulation for relations is similar.) Moreover, the principle of λ-Conversion simplifies the strict proof of the equivalence of Frege's Rule of Substitution and the Comprehension Principle for Concepts. As it turns out, not only does Frege's Rule of Substitution imply the Comprehension Principle for Concepts, but the converse also holds: the Comprehension Principle for Concepts implies the Rule of Substitution. [For a proof sketch, see Boolos (1985) p. 161-162. Note that instead of $[\lambda x\ \varphi(x)]$, Boolos uses the notation {*a*: *Aa*}; elsewhere, in (1987) for example, Boolos uses the notation [*x* : *A*(*x*)] to denote concepts.]

It is important to appreciate that the system we have just described, i.e., Frege's system of second order logic and the theory of (relational) concepts that he developed in **Begr**, is consistent. (It is only later in **Gg**, when Frege added Basic Law V to this consistent basis, that the resulting system became inconsistent.) Its underlying comprehension principle for concepts ensures that the domain of concepts is very rich. Each concept has a negation, every pair of concepts has a conjunction, every pair of concepts has a disjunction, etc. The reader should be able to write down instances of the comprehension principle which demonstrate these claims. In Part III of **Begr**, Frege applied his system to the ‘theory of sequences’ (we call these ‘*R*-series’ below). It is here that Frege presents his celebrated definition of the ‘ancestral’ of a relation and first proves the generalized analogues of the principle of mathematical induction, as well as various structural properties of the ancestral. We shall postpone further discussion of this work until §§4 and 5, where we reproduce Frege's definition of the ancestral of a relation and show how Frege incorporated this definition into the proof of mathematical induction, respectively.

§2: Frege's Theory of Extensions: Basic Law V

[Note: This section is included to give an historical understanding of Frege's system. It is not required for

understanding the proof of Frege's Theorem.]

The principle that undermined Frege's system (Basic Law V) was one that attempted to systematize the notions ‘course-of-values of a function’ and ‘extension of a concept’. The course-of-values of a function f is something like a set of ordered pairs that records the value $f(x)$ for every argument x. For example, the course-of-values of the function *father of x* records, among other things, that Bill Clinton is the value of the function when Chelsea Clinton is the argument. The course-of-values for the function x^2 records, among other things, that the number 4 is the value when the number 2 is the argument, that 9 is the value when 3 is the argument, etc. The extension of a concept is something like the set of all objects that fall under the concept. For example, the extension of the concept *x is a positive even integer less than 8* is something like the set consisting of the numbers 2, 4,and 6 (strictly speaking, the extension of this concept records The True as the value when 2, 4 and 6 are supplied as argument, and records that The False is the value when anything else is supplied as argument). Since concepts are just functions from objects to truth values, the extension of a concept is simply the course-of-values which records which objects that concept maps to The True.

Notation for Courses-of-Values

Frege introduces notation for courses-of-values in **Gg I**, §9. He switched to the lower case letters ε and α when writing the names of courses-of-values and extensions. He used something like the notation $\acute{\varepsilon} f(\varepsilon)$ to designate the course-of-values of the function f. In the special case where f is the concept F, he used the notation $\acute{\varepsilon} F\varepsilon$ to designate the extension of the concept F.

Here are two pairs of examples of Frege's notation---the first are examples of courses-of-values and the second are examples of extensions. The first pair of examples comes from **Gg I**, §9. Frege uses the notation:

$$\acute{\varepsilon}(\varepsilon^2 - \varepsilon)$$

to denote the course-of-values of the function:

$$x^2 - x$$

He also uses:

$$\acute{\alpha}(\alpha \cdot (\alpha - 1))$$

to denote the course-of-values of the function:

$$x \cdot (x - 1)$$

Frege then notes that:

$$\forall x[x^2 - x = x \cdot (x - 1)]$$

always has the same truth value as the following:

$$\grave{\varepsilon}(\varepsilon^2 - \varepsilon) = \grave{\alpha}(\alpha \cdot (\alpha - 1))$$

This equivalence will become embodied in Basic Law V.

Here is a second example, this time involving a pair of concepts. Consider the concept *that which when added to* 4 *equals* 5, or using λ-notation, the following concept:

$$[\lambda x\, x + 4 = 5]$$

Frege would use the following notation to denote the extension of this concept:

$$\grave{\varepsilon}(\varepsilon + 4 = 5)$$

Now consider the concept *that which when added to* 2^2 *equals* 5 (i.e., $[\lambda x\, x + 2^2 = 5]$). Frege would use the following notation to denote the extension of this concept:

$$\grave{\alpha}(\alpha + 2^2 = 5)$$

Note that it seems natural to identify these two extensions whenever all and only the objects that fall under the first concept fall under the second.

From these examples, it should be clear that when $\varphi(x)$ is any formula in which the variable x is free, we may write $\grave{\varepsilon}(\varphi(\varepsilon/x))$ to designate the extension of the concept $[\lambda x\, \varphi(x)]$. Those readers already familiar with the 'λ-calculus' should remember that $\grave{\varepsilon}(\varphi(\varepsilon))$ denotes an object, that $[\lambda x\, \varphi(x)]$ denotes a concept, and that Frege rigorously distinguished objects and concepts and supposed them to constitute mutually exclusive domains. One subtlety that we have ignored in the examples of the previous paragraph is the difference between Frege's notation

$$\grave{\varepsilon}(\varepsilon + 4 = 5)$$

and our notation:

$$\grave{\varepsilon}([\lambda x\, x + 4 = 5]\varepsilon)$$

However, Frege's notation is equivalent to this latter notation, given that λ-Conversion allows us to convert

$$[\lambda x\; x + 4 = 5]\varepsilon$$

to:

$$\varepsilon + 4 = 5$$

In general, whenever we have a complex formula $\varphi(x)$ in which the variable x is free, and $\varphi(\varepsilon/x)$ is the result of replacing ε for x everywhere in $\varphi(x)$, we may always rewrite:

$$\grave{\varepsilon}([\lambda x\; \varphi(x)]\varepsilon)$$

as:

$$\grave{\varepsilon}(\varphi(\varepsilon/x))$$

Frege doesn't require this ‘Rewrite Rule’, since he doesn't use λ-notation. But since it is logically more perspicuous to distinguish open formulas from the corresponding names of concepts, it is important to make this rule explicit.

Membership in an Extension

If we remember that the extension of a concept is something like the set of objects that fall under the concept, then we could replace Frege's talk of ‘extensions’ by talk of ‘sets’ and use the following ‘set notation’ to refer to the set of objects that when added to 4 yield 5 and the set of objects that when added to 2^2 yield 5, respectively:

$$\{x \mid x + 4 = 5\}$$

$$\{x \mid x + 2^2 = 5\}$$

In what follows, we sometimes render Frege's notation in this more modern notation.

Frege took advantage of his second-order language to *define* what it is for an object to be a member of an extension. Although Frege used the notation $x \frown y$ to designate the membership relation, we shall follow the more usual practice of using $x \in y$. Thus, the following captures the main features of Frege's definition of membership in **Gg I**, §34:

$x \in y =_{df} \exists G(y = \acute{\varepsilon} G\varepsilon \ \& \ Gx)$

In other words, *x* is an element of *y* just in case *x* falls under a concept of which *y* is the extension. For example, given this definition, one can prove that John is a member of the extension of the concept *being happy* (formally: $j \in \acute{\varepsilon} H\varepsilon$) from the premise that John falls under the concept *being happy* ('*Hj*'). Here is a simple proof:

1. Hj	Premise
2. $\acute{\varepsilon} H\varepsilon = \acute{\varepsilon} H\varepsilon$	= Introduction
3. $\acute{\varepsilon} H\varepsilon = \acute{\varepsilon} H\varepsilon \ \& \ Hj$	from 1,2, by & Introduction
4. $\exists G(\acute{\varepsilon} H\varepsilon = \acute{\varepsilon} G\varepsilon \ \& \ Gj)$	from 3, by Existential Introduction
5. $j \in \acute{\varepsilon} H\varepsilon$	from 4, by definition of $\in$

Some readers may wish to examine a somewhat more complex example, in which the above definition of membership is used to prove that $1 \in \acute{\alpha}(\alpha + 2^2 = 5)$ given the premise that $1 + 2^2 = 5$. (A More Complex Example)

Basic Law V

Frege attempted to axiomatize courses-of-values and extensions by formulating Basic Law V. The reader should now be in a position to see how the following formulation of Basic Law V corresponds to Frege's formulation in **Gg I**, §20:

Basic Law V:
$\acute{\varepsilon} f(\varepsilon) = \acute{\alpha} g(\alpha) \equiv \forall x[f(x) = g(x)]$

This principle asserts: the course-of-values of the function *f* is identical to the course-of-values of the function *g* if and only if *f* and *g* map every object to the same value. [Actually, Frege uses an identity sign instead of the biconditional sign as the main connective of the principle. The reason he could do this is that, in his system, when two sentences are materially equivalent, they *name* the same truth value.]

Basic Law V has the following special case, when the functions *f* and *g* are the concepts *F* and *G*:

Basic Law V (Special Case):
$\acute{\varepsilon} F\varepsilon = \acute{\alpha} G\alpha \equiv \forall x(Fx \equiv Gx)$

[Here, again, Frege used an identity sign in place of the biconditional signs.] In this special case, Basic

Law V asserts: the extension of the concept *F* is identical to the extension of the concept *G* if and only if all and only the objects that fall under *F* fall under *G* (i.e., if and only if the concepts *F* and *G* are materially equivalent). In more modern guise, Frege's Basic Law V asserts that the set of *F*s is identical to the set of *G*s if and only if *F* and *G* are materially equivalent:

$$\{x|Fx\} = \{y|Gy\} \equiv \forall z(Fz \equiv Gz)$$

The second example discussed in the subsection Notation for Courses of Values can now be seen as an instance of Basic Law V:

$$\grave{\varepsilon}(\varepsilon + 4 = 5) = \grave{\alpha}(\alpha + 2^2 = 5) \equiv \forall x(x + 4 = 5 \equiv x + 2^2 = 5)$$

This simply asserts that the extension of the concept *that which added to* 4 *yields* 5 is identical to the extension of the concept *that which added to* 2^2 *yields* 5 if and only if all and only the objects that when added to 4 yield 5 are objects that when added to 2^2 yield 5.

Basic Law V looks like it asserts a very general truth. In fact, it does correctly imply the Law of Extensions and the Principle of Extensionality. The Law of Extensions (cf. **Gg I**, §55, Theorem 1) asserts that an object is a member of the extension of a concept if and only if it falls under that concept:

Law of Extensions:
$\forall F \forall x(x \in \grave{\varepsilon}F\varepsilon \equiv Fx)$

(Proof of the Law of Extensions)

Basic Law V also correctly implies the Principle of Extensionality. This principle asserts that if two extensions have the same members, they are identical. If we let the expression '*Extension*(*x*)' abbreviate the formula '$\exists F(x = \grave{\varepsilon}F\varepsilon)$' then we may formally represent and derive the principle of extensionality as follows:

Principle of Extensionality:
$Extension(x)\ \&\ Extension(y) \rightarrow [\forall z(z \in x \equiv z \in y) \rightarrow x = y]$

(Proof of the Principle of Extensionality)

Despite these deceptive successes of Basic Law V, the fact is that it can't be consistently added to Frege's system. In the following subsections, we shall show how Basic Law V proves to be inconsistent with the rest of Frege's second order logic and theory of concepts. The proofs depend essentially on the second order character of Frege's system and on the second-order definition of the membership relation. Frege was made aware of the inconsistency by Bertrand Russell, who sent him a letter formulating 'Russell's Paradox' just as the second volume of **Gg** was going to press. Frege quickly added an Appendix to the

second volume, describing two distinct ways of deriving a contradiction from Basic Law V. The first establishes the contradiction directly, without any special definitions. The second deploys the membership relation and more closely follows Russell's Paradox. We will examine both derivations of the contradiction in what follows.

Both derivations of the contradiction turn on an important corollary to Basic Law V, namely, that every concept has an extension:

> **Corollary to Basic Law V**:
> $\forall F \exists x(x = \acute{\alpha} F\alpha)$

To see that this is a consequence of Basic Law V, note that when we instantiate the variable *G* to *F* in Basic Law V, we can establish:

$$\acute{\varepsilon} F\varepsilon = \acute{\alpha} F\alpha \equiv \forall x(Fx \equiv Fx)$$

Since the right side of this instance of Law V can be derived by logic alone, it follows that $\acute{\varepsilon} F\varepsilon = \acute{\alpha} F\alpha$. But, then, by existential generalization, it follows that:

$$\exists x(x = \acute{\alpha} F\alpha)$$

But now our *Corollary* follows by universal generalization on the concept variable *F*. However, the combination of Frege's Rule of Substitution (which ensures that there is a concept corresponding to every formula with free variable *x*) and Basic Law V (which ensures that each concept has an extension that behaves in a certain way), turns out to be a volatile mix.

First Derivation of the Contradiction

In the Appendix to **Gg II**, Frege shows that a contradiction can be derived once we formulate the concept *being the extension of a concept which you don't fall under*. The following open formula expresses this concept:

$$\exists F(x = \acute{\alpha} F\alpha \ \& \ \neg Fx)$$

From the Comprehension Principle for Concepts (or Frege's Rule of Substitution), we know that there exists a concept corresponding to this formula and we may use the following λ-expression to name it:

$$[\lambda x \ \exists F(x = \acute{\alpha} F\alpha \ \& \ \neg Fx)]$$

Now by the *Corollary* to Basic Law V and our Rewrite Rule, the extension of this concept exists and can

be designated as follows:

$$\acute{\varepsilon}[\exists F(\varepsilon = \acute{\alpha}F\alpha \ \& \ \neg F\varepsilon)]$$

It can now be proved that this extension falls under the concept $[\lambda x\, \exists F(x = \acute{\alpha}F\alpha \ \& \ \neg Fx)]$ if and only if it does not. (First Derivation of the Contradiction.)

Second Derivation of the Contradiction

Frege next (in the Appendix to **Gg II**) explained how Basic Law V implies the Naive Comprehension Axiom for extensions or sets, which Russell's Paradox shows to be inconsistent. From the Law of Extensions (which was derived from Basic Law V above), one can establish the Naive Comprehension Axiom for extensions in three simple steps. First we instantiate the Law of Extensions to the free variable *F*, to yield:

$$\forall x(x \in \acute{\varepsilon}F\varepsilon \equiv Fx)$$

Then by generalizing on the extension $\acute{\varepsilon}F\varepsilon$, it follows that:

$$\exists y \forall x(x \in y \equiv Fx)$$

Now at this point, we may universally generalize on the variable *F* to get the following second-order Naive Comprehension Axiom for extensions, which asserts that for every concept *F*, there is an extension which has as members all and only the objects that fall under *F*:

Naive Comprehension Axiom for Extensions:
$\forall F \exists y \forall x(x \in y \equiv Fx)$

Alternatively, instead of generalizing, we could have appealed to Frege's Rule of Substitution to show that all of the instances of the following Naive Comprehension Schema for extensions are derivable in Frege's system:

Naive Comprehension Schema for Extensions:
$\exists y \forall x(x \in y \equiv \varphi(x))$, where $\varphi(x)$ is any formula in which x is free and which contains no free occurrences of y

This asserts that for any formula $\varphi(x)$ defining a condition on objects, there is an extension which has as members all and only the objects that meet the condition.

Both the Naive Comprehension Axiom and the Naive Comprehension Schema immediately give rise to

Russell's Paradox in the context of Frege's logic. In the case of the axiom, the contradiction follows by instantiating the quantified variable F to the concept $[\lambda z\, \neg(z \in z)]$. In the case of the schema, the contradiction follows by taking $\varphi(x)$ to be $\neg(x \in x)$, as follows:

$$\exists y \forall x(x \in y \equiv \neg(x \in x))$$

In either case, the proof of the contradiction goes through. The derivation of the contradiction from the above instance of the schema is particularly easy. For suppose the object b is such a y. Then:

$$\forall x(x \in b \equiv \neg(x \in x))$$

But we can now instantiate the universal claim to the object b to yield the following contradiction:

$$b \in b \equiv \neg(b \in b)$$

(See the entry on Russell's Paradox.)

How the Paradox is Engendered

Philosophers have diagnosed the inconsistency in Frege's system in various ways, and it is safe to say that the matter is still somewhat controversial. In this subsection, we discuss only the basic elements of the problem. Most philosophers and logicians agree that the reason Frege's second-order logic and theory of extensions is inconsistent is that they jointly require the impossible situation in which the domain of concepts has to be strictly larger than the domain of extensions while at the same time the domain of extensions has to be as large as the domain of concepts. This impossible situation is strikingly analogous to the impossible situation set up in the proof by *reductio* of Cantor's Theorem (Cantor's Theorem asserts that if A is any set, and B is the power set of A (i.e., B is the set of all subsets of A), then B has more members than A; the proof by reductio shows that it is impossible for there to be a function from A *onto* B).

To analyze the inconsistency in Frege's system in more detail, it is important to discuss the conditions under which concepts are to be identified. Although Frege did not believe that statements of the form '$F = G$' were meaningful, it is evident from the study of **Gg** that the material equivalence of concepts F and G serves as the proxy identity conditions of F and G. So, whenever it is *not* the case that all and only the objects that fall under F fall under G, F and G are distinct concepts.

With this in mind, we can see how the paradox is engendered. Recall first that the *Corollary* to Basic Law V reveals that Basic Law V correlates each concept with an extension. Each direction of Basic Law V requires that this correlation have certain properties. We shall see, for example, that the right-to-left direction of Basic Law V (i.e., Va) requires that no concept gets correlated with two distinct extensions. [Frege uses the label 'Va' to designate the right-to-left direction of Basic Law V. See, for example, **Gg I**, §52. However, many commentators use 'Va' to designate the left-to-right direction. We shall follow

Frege's use, since that will make sense of his Appendix to **Gg II**, in which he discusses the paradoxes.] Va asserts:

> **Basic Law Va**:
> $\forall x(Fx \equiv Gx) \rightarrow \acute{\varepsilon}F\varepsilon = \acute{\alpha}G\alpha$

If we think in terms of its contraposition and remember the identity conditions for concepts, Va in effect asserts that whenever extensions differ, the concepts with which they are correlated differ. This means that the correlation between concepts and extensions that Basic Law V sets up must be a function---no concept gets correlated with two distinct extensions (though for all Va tells us, distinct concepts might get correlated with the same extension). Frege noted (in the Appendix to **Gg II**) that this direction of Basic Law V doesn't seem problematic.

However, the left-to-right direction of Basic Law V (i.e., Vb) is more serious. Vb asserts:

> **Basic Law Vb**:
> $\acute{\varepsilon}F\varepsilon = \acute{\alpha}G\alpha \rightarrow \forall x(Fx \equiv Gx)$

If we consider the contrapositive of this and remember the identity conditions for concepts, then Vb, in effect, asserts that whenever the concepts *F* and *G* differ, the extensions of *F* and *G* differ. So, the correlation that Basic Law V sets up between concepts and extensions will have to be one-to-one; i.e., it correlates distinct concepts with distinct extensions. Since every concept is correlated with some extension, there have to be at least as many extensions as there are concepts.

But the problem is that Frege's system *as a whole* requires that there be *more* concepts than extensions. The requirement that there be more concepts than extensions is imposed jointly by the Comprehension Principle for Concepts *and* the new significance this principle takes on in the presence of Basic Law V. The Comprehension Principle for Concepts asserts the existence of a concept for every condition on objects expressible in the language. Now though it may seem that this forces the domain of concepts to be larger than the domain of objects, it is a model-theoretic fact that there are models of second-order logic with the Comprehension Principle for Concepts (but without Basic Law V) in which the domain of concepts need not be larger than the domain of objects.[1] However, the Comprehension Principle for Concepts takes on new significance when Basic Law V is added to Frege's system. The synergism of the Comprehension Principle for Concepts and Basic Law V force the domain of concepts to be larger than the domain of objects (and so larger than the domain of extensions). However, as we saw in the last paragraph, Vb requires that there be at least as many extensions as there are concepts.

Thus, Frege's second-order logic and theory of extensions together required the impossible situation in which the domain of concepts has to be strictly larger than the domain of extensions while at the same time the domain of extensions has to be as large as the domain of concepts.

Recently, there has been a lot of interest in discovering ways of repairing Frege's system. The traditional

view is that one must either restrict Basic Law V or restrict the Comprehension Principle for Concepts. Recently, Boolos (1986, 1993) developed one of the more interesting suggestions for revising Basic Law V without abandoning second-order logic and its comprehension principle for concepts. On the other hand, there have been many suggestions for restricting the Comprehension Principle for Concepts. The most severe of these is to abandon second-order logic (and the Comprehension Principle for Concepts) altogether. Schroeder-Heister (1987) conjectured that the first-order portion of Frege's system (i.e., the system which results by adding Basic Law V to the first-order predicate calculus) was consistent and this was proved by Parsons (1987) and Burgess (1998).[2] Heck (1996) and Wehmeier (1999) consider less drastic moves. They investigate systems of second-order logic which have been extended by Basic Law V but in which the Comprehension Principle for Concepts is restricted in some way. We will not discuss their investigations further in the present entry, for it is not clear which of their alternatives, or others, would have been acceptable to Frege. Instead, we focus on the theoretical accomplishment revealed by Frege's work in **Gg**.

Despite the failure of Basic Law V, Frege validly proved a rather deep fact about the natural numbers, namely, that the Dedekind/Peano axioms for number theory could be derived in second-order logic with the help of a single additional principle. The principle in question is known as Hume's Principle (discussed below). Although both Parsons (1965) and Wright (1983) had recently noted that Hume's Principle was powerful enough for the derivation of the Dedekind/Peano axioms, Heck (1993) showed that although Frege did use Basic Law V to derive Hume's principle, his (Frege's) subsequent derivations of the Dedekind/Peano axioms of number theory from Hume's Principle never made an *essential* appeal to Basic Law V. Since Hume's Principle just by itself is consistent with second-order logic, this means that Frege validly derived the basic laws of number theory. It will be the task of the next few sections to explain Frege's accomplishments in this regard. We will do this in two stages. In §3 we study Frege's attempt to derive Hume's Principle from Basic Law V by analyzing cardinal numbers as extensions. Then, we put this aside in §§4 and 5 to examine how Frege was able to derive the Dedekind/Peano axioms of number theory from Hume's Principle alone.

§3: Frege's Analysis of Cardinal Numbers

Cardinal numbers are the numbers that can be used to answer the question ‘How many?’, and Frege discovered that such numbers bear an interesting relationship to the natural numbers. Frege's insights concerning this relationship trace back to his work in **Gl**, in which the notion of an extension played very little role. The seminal idea of **Gl**, §46, was the observation that a statement of number (e.g., "There are nine planets") is an assertion about a concept. To explain this idea, Frege noted that one and the same external phenomenon can be counted in different ways; for example, a certain external phenomenon could be counted as one army, 5 divisions, 25 regiments, 120 companies, 400 platoons, or 4000 people. Each different way of counting this phenomenon corresponds to the manner of its conception. The question "How many are there?" is only properly formulated as the question "How many *F*s are there?" where a concept *F* is supplied. On Frege's view, the statements of number which answer such questions (e.g., "There are *n* *F*s") tell us something about the concept involved. For example, the statement "There are nine planets in the solar system" tells us that the ordinary, *first-level* concept *planet in the solar system*

falls under the *second-level* numerical concept *concept under which nine objects fall*.

Frege then moves from this realization, in which statements of numbers are analyzed as predicating second-level numerical concepts of first-level concepts, to develop an account of the cardinal and natural numbers as ‘self-subsistent’ objects. He introduces a ‘cardinality operator’ on concepts, namely, ‘the number belonging to the concept *F*’, which will designate the cardinal number which numbers the objects falling under *F*. In what follows, we say this more simply as ‘the number of *F*s’ and use the simple notation ‘#F’. Frege offers both an implicit and an explicit definition of this operator in **Gl**. Both of these definitions require a preliminary definition of when two concepts *F* and *G* are in one-to-one correspondence or ‘equinumerous’. After developing the definition of equinumerosity, we then discuss Frege's implicit and explicit definition of the number of *F*s.

Equinumerosity

In order to state the definition of equinumerosity, we shall employ the well-known logical notion ‘there exists a unique x such that $\varphi(x)$’. To say that there exists a unique x such that $\varphi(x)$ is to say: there is some x such that $\varphi(x)$ and anything y which is such that $\varphi(y/x)$ is identical to x. In what follows, we use the notation ‘$\exists!x\varphi(x)$’ to abbreviate this notion of unique existence, and we define it formally as follows:

$$\exists!x\varphi(x) \quad =_{df} \quad \exists x[\varphi(x) \;\&\; \forall y(\varphi(y/x) \rightarrow y{=}x)]$$

Now, in terms of this logical notion of unique existence, we can state Frege's definition of equinumerosity (**Gl**, §71, 72) as follows:

> *F* and *G* are *equinumerous* (or, *F* and *G* are in *one-to-one correspondence*) just in case there is a relation *R* such that: (1) every object falling under *F* is *R*-related to a unique object falling under *G*, and (2) every object falling under *G* is such that there is a unique object falling under *F* which is *R*-related to it.

If we let ‘$F \approx G$’ stand for equinumerosity, then the definition of this notion can be rendered formally as follows:

$$F \approx G \quad =_{df}$$
$$\exists R[\forall x(Fx \rightarrow \exists!y(Gy \;\&\; Rxy)) \quad \& \quad \forall x(Gx \rightarrow \exists!y(Fy \;\&\; Ryx))]$$

To see that Frege's definition of equinumerosity works correctly, consider the following two examples. In the first example, we have two concepts that are equinumerous:

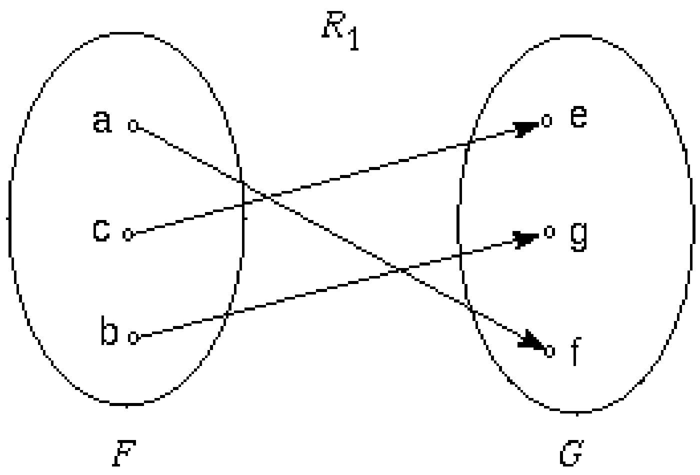


Figure 1

Although there are several different relations *R* which would demonstrate the equinumerosity of *F* and *G* the particular relation used in Figure 1 is:

$$R_1 = [\lambda xy\ (x{=}a\ \&\ y{=}f) \vee (x{=}b\ \&\ y{=}g) \vee (x{=}c\ \&\ y{=}e)]$$

It is a simple exercise to show that R_1, as defined, is a ‘witness’ to the equinumerosity of *F* and *G* (according to the definition).

In the second example, we have two concepts that are not equinumerous:

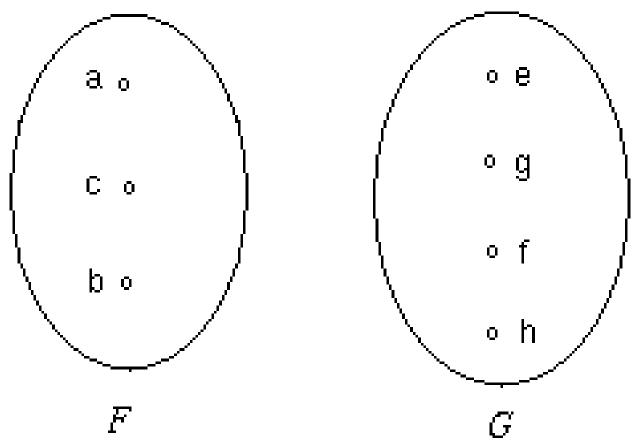


Figure 2

In this example, no relation *R* can satisfy the definition of equinumerosity.

Clearly, then, the concepts *F* and *G* will be equinumerous whenever the number of objects falling under *F* is identical to the number of objects falling under *G*. This fact will be codified by Hume's Principle. Before moving ahead to the discussion of this principle, the reader should convince him- or herself of the following four facts: (1) that the material equivalence of two concepts implies their equinumerosity, (2) that equinumerosity is reflexive, (3) that equinumerosity is symmetric, and (4) that equinumerosity is

transitive. In formal terms, the following facts are provable:

> **Facts About Equinumerosity**:
> 1. $\forall x(Fx \equiv Gx) \rightarrow F \approx G$
> 2. $F \approx F$
> 3. $F \approx G \rightarrow G \approx F$
> 4. $F \approx G \;\&\; G \approx H \rightarrow F \approx H$

The proofs of these facts, in each case, require the identification of a relation that is a witness to the relevant equinumerosity claim. In some cases, it is easy to identify the relation in question. In other cases, the reader should be able to ‘construct’ such relations (using λ-notation) by considering the examples described above. Facts (2) - (4) establish that equinumerosity is an ‘equivalence relation’ which divides up the domain of concepts into ‘equivalence classes’ of equinumerous concepts. Material equivalence is also an equivalence relation which divides up the domain of concepts into equivalence classes of materially equivalent concepts.

Contextual Definition of ‘The Number of *F*s’: Hume's Principle

Frege contextually defined ‘the number of *F*s’ in terms of the principle now known as Hume's Principle:[3]

> **Hume's Principle**:
> The number of *F*s is identical to the number of *G*s if and only if *F* and *G* are equinumerous.

Using our notation ‘#*F*’ to abbreviate ‘the number of *F*s’, we may formalize Hume's Principle as follows:

> **Hume's Principle**:
> $\#F = \#G \;\equiv\; F \approx G$

This contextual definition governing cardinal numbers is the basic principle upon which Frege forged his development of the theory of natural numbers.[4] In **Gl**, Frege sketched the derivations of the basic laws of number theory from Hume's Principle; these sketches were developed into more rigorous proofs in **Gg I**. We will examine these derivations in the following sections.

Once Frege had a contextual definition of #*F*, he then defined a cardinal number as any object which is the number of some concept:

> *x is a cardinal number* $=_{df}$ $\exists F(x = \#F)$

This definition appears in **Gl**, §72.

Notice that Hume's Principle bears an obvious formal resemblance to Basic Law V. Both are biconditionals asserting the equivalence of an identity among singular terms (the left-side condition) with an equivalence relation on concepts (the right-side condition). Indeed, both correlate concepts with certain objects. In the case of Hume's Principle, each concept F is correlated with $\#F$. However, whereas Basic Law V problematically asserts that the correlation between concepts and extensions is one-to-one, Hume's Principle only asserts that the correlation between concepts and numbers is many-to-one. Hume's Principle often correlates distinct concepts with the same number. For example, the distinct concepts *author of Principia Mathematica* (‘$[\lambda x\, Axp]$’) and *number between* 1 *and* 4 (‘$[\lambda x\, 1 < x < 4]$’) are equinumerous (both both have two objects falling under them). So $\#[\lambda x\, Axp] = \#[\lambda x\, 1 < x < 4]$. Thus, Hume's Principle, unlike Basic Law V, does not require that the domain of numbers be as large as the domain of concepts. Indeed, Hume's Principle has recently been proved consistent with second-order logic. This was shown independently by Burgess (1984) and Hazen (1985).

Explicit Definition of ‘The Number of *F*s’

[Note: The remaining two subsections are not strictly necessary for understanding the proof of Frege's Theorem. They are included here for those who wish to have a more complete understanding of what Frege in fact attempted to do. They presuppose the material in §2. Readers interested in just the positive aspects of Frege's accomplishments should skip directly to §4.]

Before we examine the powerful consequences that Frege derived from Hume's Principle, it is worth digressing to describe his failed attempt to explicitly define ‘$\#F$’ and to derive Hume's Principle from Basic Law V. The idea behind this attempt was the realization that if given any concept F, the notion of equinumerosity can be used to define the second-level concept *being a concept G that is equinumerous to F* (‘$G \approx F$’). Frege found a way to collect all of the concepts equinumerous to a given concept F into a single extension. In **Gl**, he informally took this to be an extension consisting of first-order concepts. In that work, he defined informally (§68):

> the number of Fs $=_{df}$
> the extension of the second-level concept: *being a first-level concept equinumerous to F*

In terms of the example used at the end of the previous subsection, this definition identifies the number of the concept *author of Principia Mathematica* as the extension consisting of all and only those first-level concepts that are equinumerous to this concept; this extension has both $[\lambda x\, Axp]$ and $[\lambda x\, 1 < x < 4]$ as members. Frege in fact identifies the cardinal number 2 with this extension, for it contains all and only those concepts under which two objects fall. Similarly, Frege identifies the cardinal number 0 with the extension consisting of all those first-level concepts under which no object falls; this extension would include such concepts as *unicorn*, *centaur*, *prime number between* 3 *and* 5, etc. Frege's insight here inspired Russell to develop a somewhat similar definition in his work, and it is now common to see references to the so-called "Frege-Russell definition of the cardinal numbers" as classes of equinumerous concepts or sets.[5] Of course, this explicit definition of ‘the number of *F*s’ stands or falls with a coherent

conception of ‘extension’. We know that Basic Law V does not offer such a coherent conception.

Derivation of Hume's Principle

Frege's derivation of Hume's Principle was invalidated by the fact that it appeals to the inconsistent Basic Law V. Nevertheless, it is instructive to consider why Frege thought the derivation was valid. In **Gl**, §73, Frege sketches an informal proof of the right-to-left direction of Hume's Principle using the above explicit definition of the number of *F*s. The derivation appeals to the fact that a concept *G* is a member of the extension of the second-level concept *concept equinumerous to F* if and only if *G* is equinumerous to *F*. In other words, the proof relies on a kind of higher-order version of the Law of Extensions (described above), the ordinary version of which we know to be a consequence of Basic Law V.[6] Here is a reconstruction of Frege's proof in **Gl**, §73, extended so as to cover both directions of Hume's Principle.

> Frege's Derivation of Hume's Principle in **Gl**

However, in the development of **Gg**, Fregean extensions do *not* contain concepts as members but rather objects. So Frege had to find another way to express the explicit definition described in the previous subsection. His technique was to let extensions go proxy for their corresponding concepts. Since a full explanation of this technique and the proof of Hume's Principle in **Gg** would constitute a digression for the present exposition, we shall describe the details for interested readers on a separate page:

> Frege's `Derivation' of Hume's Principle in **Gg**

As noted on several occasions, the inconsistency in Basic Law V invalidated Frege's derivation of Hume's Principle. But Hume's Principle, in and of itself, is a powerful and consistent principle.

§4: Frege's Analysis of Predecessor, Ancestrals, and the Natural Numbers

In what follows, we shall suppose that Hume's Principle has replaced Basic Law V in Frege's second-order system. This requires that we replace the operator "the course of values of the function *f*" (and "the extension of concept *F*") with the primitive operator "the number of Fs". As we have mentioned, Frege made the insightful discovery that the basic laws of number theory could be derived from Hume's Principle alone. This is Frege's Theorem. In this section, we introduce the definitions required for the proof of Frege's Theorem. In the next section, we go through the proof. In the final section, we conclude with a discussion of the philosophical questions that arise when we consider Hume's Principle as a replacement for Basic Law V.

The insight behind Frege's analysis of the natural numbers was the realization that one can define the finite cardinal numbers in terms of the following concepts:

$C_0 = [\lambda x \;\; x \neq x]$
$C_1 = [\lambda x \;\; x = \#C_0]$
$C_2 = [\lambda x \;\; x = \#C_0 \;\; \vee \;\; x = \#C_1]$
$C_3 = [\lambda x \;\; x = \#C_0 \;\; \vee \;\; x = \#C_1 \;\; \vee \;\; x = \#C_2]$
etc.

Note that starting with C_1, each concept C_k has the following property: all and only the numbers of concepts preceding C_k in the sequence fall under C_k. So, for example, the concepts preceding C_3 are C_0, C_1, and C_2. Accordingly, all and only the following numbers fall under C_3: $\#C_0$, $\#C_1$, and $\#C_2$.

Frege noticed that these concepts can be used, respectively, to define the the finite cardinal numbers, as follows:

$0 = \#C_0$
$1 = \#C_1$
$2 = \#C_2$
etc.

This insight, however, was only the first step in Frege's plan. He realized that though this seems to define a sequence of numbers with which we can identify the natural numbers, we have not as yet defined the concept ‘natural number’ so that it applies to all and only the cardinal numbers defined in the second sequence. Such a concept is required if we are to prove *as theorems* the following axioms of Dedekind/Peano number theory:

Dedekind/Peano Axioms for Number Theory:

- 0 is a natural number.
- 0 is not the successor of any natural number.
- No two natural numbers have the same successor.
- If both (a) 0 falls under *F*, and (b) for any two natural numbers *n* and *m*, the fact that *n* falls under *F* implies that *m* falls under *F*, then every natural number falls under *F*. (Principle of Induction)
- Every natural number has a successor.

Moreover, Frege recognized the need to employ the Principle of Induction in the proof that every number has a successor. One cannot prove the claim that *every number has a successor* simply by producing the sequence of expressions for cardinal numbers (e.g., the second of the two sequences described above). All such a sequence demonstrates is that for every expression listed in the sequence, one can define an expression of the appropriate form to follow it in the sequence. This is *not* the same as proving that *every natural number* has a successor.

Predecessor

To accomplish these further goals, Frege proceeded by defining the concept *x* (*immediately*) *precedes y* as follows (**Gl**, §76, and **Gg I**, §43):

> *x* (*immediately*) *precedes y* if and only if there is a concept *F* and an object *w* such that: (a) *w* falls under *F*, (b) *y* is the number of *F*s, and (c) *x* is the number of the concept *object falling under F other than w*

In formal terms, the definition takes the following form:

$$Precedes(x,y) \ =_{df}$$
$$\exists F \exists w(Fw \ \ \& \ \ y = \#F \ \ \& \ \ x = \#[\lambda z\, Fz \,\&\, z \neq w])$$

Even though we can't as yet assume that we have defined the natural numbers 1 and 2, we can use them intuitively to show that the definition properly predicts that *Precedes*(1,2) if given certain facts about the numbers of certain concepts. Let the expression '[λ*z Azp*]' denote the concept *author of Principia Mathematica*. Only Bertrand Russell ('*r*') and Alfred Whitehead fall under this concept. Let the expression '[λ*z Azp* & *z*≠*r*]' denote the concept *author of Principia Mathematica other than Russell*.[7] Then the following may, for the purposes of this example, be taken as facts:

- Russell falls under the concept *author of Principia Mathematica*, i.e.,
  $[\lambda z\, Azp]r$
- 2 is the number of the concept *author of Principia Mathematica*, i.e.,
  $2 = \#[\lambda z\, Azp]$
- 1 is the number of the concept *author of Principia Mathematica other than Russell*, i.e.,
  $1 = \#[\lambda z\, Azp \,\&\, z \neq r]$

If we assemble these truths into a conjunction and apply existential generalization in the appropriate places, the result is the definiens of the definition of predecessor instantiated to the numbers 1 and 2. Thus, if given certain facts about the number of objects falling under the certain concepts, the definition of predecessor correctly predicts that *Precedes*(1,2).

The Ancestral of Relation *R*

Frege next defines the relational concept *x is an ancestor of y in the R-series*. This new relation is called 'the ancestral of the relation *R*' and we henceforth designate this relation as *R**. Frege first defined the ancestral of relation *R* in **Begr** (Part III, Proposition 76), though the word 'ancestral' comes to us from Russell and Whitehead. Frege's term for the ancestral is "*x* comes before *y* in the *R*-series"; alternatively, "*y* follows *x* in the *R*-series". (See also **Gl**, §79 and **Gg I**, §45.) The intuitive idea is easily grasped if we

consider the relation *x* is the father of *y*. Suppose that *a* is the father of *b*, that *b* is the father of *c*, and that *c* is the father of *d*. Then '*x* is an ancestor of *y* in the fatherhood-series' is defined so that *a* is an ancestor of *b*, *c*, and *d*, that *b* is an ancestor of *c* and *d*, and that *c* is an ancestor of *d*.

Frege's definition of the ancestral of *R* requires a preliminary definition:

> *the concept F is hereditary in the R-series* if and only if any pair of *R*-related objects *x* and *y* are such that *y* falls under *F* whenever *x* falls under *F*

In formal terms:

$$Her(F,R) =_{abbr} \forall x \forall y(Rxy \mathbin{\&} \mathrm{F}x \rightarrow \mathrm{F}y)$$

Intuitively, the idea is that *F* is hereditary in the *R*-series if *F* is always 'passed' from *x* to *y* whenever *x* and *y* are a pair of *R*-related objects. (We warn the reader here that the notation '*Her*(*F*,*R*)' is merely an abbreviation of a much longer statement. It is *not* a formula of our language having the form '*R*(*x*,*y*)'. In what follows, we sometimes introduce other such abbreviations.)

Frege's definition of the ancestral of *R* can now be stated as follows:

> *x comes before y in the R-series* $=_{df}$ *y* falls under all those *R*-hereditary concepts *F* under which falls every object to which *x* is *R*-related

In other words, *y* follows *x* in the *R*-series whenever *y* falls under every hereditary concept *F* which *x* 'passes on' to all of its immediate descendants. In formal terms:

$$R^*(x,y) =_{df} \forall F[\forall z(Rxz \rightarrow Fz) \mathbin{\&} Her(F,R) \rightarrow Fy]$$

For example, Clinton's father stands in relation *father* of* (i.e., *forefather*) to Chelsea because she falls under every hereditary concept that Clinton and his brother inherited from Clinton's father. However, Clinton's brother is not one of Chelsea's forefathers, since he fails to be her father, her grandfather, or any of the other links in the chain of fathers from which Chelsea descended.

It is important to grasp the differences between a relation *R* and its ancestral *R**. *Rxy* implies *R**(*x*,*y*) (e.g., if Clinton is a father of Chelsea, then Clinton is a forefather of Chelsea), but the converse doesn't hold (Clinton's father is a father* of Chelsea, but he is not a father of Chelsea). Indeed, a grasp of the definition of *R** should leave one able to prove the following easy consequences, many of which correspond to theorems in **Begr** and **Gg**:[8]

Facts About *R**:

1. $Rxy \rightarrow R^*(x,y)$
2. $\neg(R^*(x,y) \rightarrow Rxy)$
3. $\exists x R^*(x,y) \rightarrow \exists x\, Rxy$
4. $Rxy \;\&\; R^*(y,z) \rightarrow R^*(x,z)$
5. $[R^*(x,y) \;\&\; \forall z(Rxz \rightarrow Fz) \;\&\; Her(F,R)] \rightarrow Fy$
6. $[Fx \;\&\; R^*(x,y) \;\&\; Her(F,R)] \rightarrow Fy$
7. $R^*(x,y) \;\&\; R^*(y,z) \rightarrow R^*(x,z)$

The reader should consider what happens when *R* is taken to be the relation *precedes*. Appealing to our intuitive grasp of the numbers, we can say that it is an instance of Fact (1) that if 10 precedes 12, then 10 precedes* 12; and that it is an instance of Fact (2) that 10's preceding* 12 does not imply that 10 precedes 12. An instance of Fact (7) is that precedes* is transitive. When we restrict ourselves to the natural numbers, it is intuitive to think of the difference between precedes and precedes* as the difference between *immediately precedes* and *less-than*.

The Weak Ancestral of *R*

Given the notion of the ancestral of relation *R*, Frege then defines its weak ancestral, which he termed "*y is a member of the R-series beginning with x*" (cf. **Begr**, Part III, Proposition 99; **Gl**, §81, and **Gg I**, §46):

> *y is a member of the R-series beginning with x* if and only if either *x* bears the ancestral of *R* to *y* or *x* = *y*

In formal terms:

$$R^+(x,y) \;=_{df}\; R^*(x,y) \vee x=y$$

We note here that Frege would also read $R^+(x,y)$ as: *x* is a member of the *R*-series ending with *y*! Logicians call R^+ the ‘weak-ancestral’ of *R* because it is a weakened version of R^*. When *R* is *precedes*, we can intuitively regard its weak ancestral, *precedes*$^+$, as the relation *less-than-or-equal-to* on the natural numbers.

The general definition of the weak ancestral of *R* yields the following facts, many of which correspond to theorems in **Gg**:[9]

Facts About R^+:

1. $Rxy \rightarrow R^+(x,y)$
2. $R^+(x,y) \;\&\; Ryz \rightarrow R^*(x,z)$
3. $R(x,y) \;\&\; R^+(y,z) \rightarrow R^*(x,z)$
4. $R^*(x,y) \;\&\; Ryz \rightarrow R^+(x,z)$

5. $R^*(x,y) \rightarrow \exists z[Rzy \,\&\, R^+(x,z)]$ (Proof of Fact 5 Concerning the Weak Ancestral)
6. $[Fx \,\&\, R^+(x,y) \,\&\, Her(F,R)] \rightarrow Fy$
7. $R^*(x,y) \,\&\, Rzy \,\&\, R$ is 1-1 $\rightarrow R^+(x,z)$[10]
8. $R^+(x,x)$ (Reflexivity)

The proofs of these facts are left as exercises.

The Concept *Natural Number*

Frege's definition of *natural number* requires one more preliminary definition. It may be recalled that that Frege identified the number 0 as the (cardinal) number of the concept *being non-self-identical*. That is:

$$0 =_{df} \#[\lambda x\, x \neq x]$$

Since the logic of identity guarantees that no object fails to be self-identical, nothing falls under the concept *being non-self-identical*. Had one of Frege's explicit definitions of the cardinal numbers worked as he had intended, the number 0 would, in effect, be identified with the extension of all (extensions of) concepts under which nothing falls. However, for the present purposes, we may note that 0 is defined in terms of the primitive notion ‘the number of *F*s’ and a particular complex concept the existence of which is guaranteed in Frege's theory of concepts and second-order logic with identity. It is straightforward to prove the following Lemma Concerning Zero from this definition of 0:

> **Lemma Concerning Zero**:
> $\#F = 0 \equiv \neg\exists x Fx$
>
> (Proof of Lemma Concerning Zero)

Note that the proof appeals to Hume's Principle and facts about equinumerosity.

Frege's definition of the concept *natural number* can now be stated in terms of the weak-ancestral of Predecessor:

> *x is a natural number* if and only if *x* is a member of the predecessor-series beginning with 0

This definition appears in **Gl**, §83, and **Gg I**, §46 as the definition of ‘finite number’. Indeed, the natural numbers are precisely the finite cardinals. In formal terms, Frege's definition becomes:

$$\mathbb{N}x =_{df} Precedes^+(0,x)$$

In what follows, we shall sometimes use the variables *m*, *n*, and *o* to range over the natural numbers.

§5: Frege's Theorem

Frege's Theorem is that the five Dedekind/Peano axioms for number theory can be derived from Hume's Principle in second-order logic. In this section, we reconstruct the proof of this theorem which can be extracted from Frege's work using the definitions and theorems assembled so far. Some of the steps in this proof can be found in **Gl**. (See the Appendix to Boolos (1990) for a reconstruction.) Our reconstruction follows Frege's **Gg** in spirit and in most details, but we have tried to simplify the presentation in several places. For a more strict description of Frege's **Gg** proof, the reader is referred to Heck (1993). The following should help prepare the reader for Heck's excellent essay.

Zero is a Number

The following is an immediate consequence of the definition of *natural number*:

> **Theorem 1**:
> $\mathbb{N}0$
>
> *Proof*: It is a simple consequence of the definition of 'weak ancestral' that R^+ is reflexive (see Fact 8 about R^+ in our subsection on the Weak Ancestral in §4). So $Precedes^+(0,0)$. Hence, by the definition of number, 0 is a number.

It seems that Frege never actually identified this fact explicitly in **Gl** or labeled this fact as a numbered Theorem in **Gg I**. It is possible that he thought it was too obvious to mention.

Zero Isn't the Successor of Any Number

It is also a simple consequence of the foregoing that 0 doesn't succeed any number. This can be represented formally as follows:

> **Theorem 2**:
> $\neg\exists x(\mathbb{N}x \;\&\; Precedes(x,0))$
>
> *Proof*: Assume, for *reductio*, that some object, say *b*, does precede 0. Then, by the definition of predecessor, it follows that there is a concept, say *Q* and an object falling under *Q*, say *c*, such that 0 is #*Q*. But this contradicts the Lemma Concerning Zero (above). Since nothing precedes 0, no natural number precedes 0.

See **Gl**, §78, Item (6); and **Gg I**, §109, Theorem 126.

No Two Numbers Have the Same Successor

The fact that no two numbers have the same successor is somewhat more difficult to prove (cf. **Gl**, §78, Item (5); **Gg I**, §95, Theorem 89). We may formulate this theorem as follows, with m, n, and o as restricted variables ranging over the natural numbers:

> **Theorem 3**:
> $\forall m \forall n \forall o[Precedes(m,o)\ \&\ Precedes(n,o) \rightarrow m = n]$

In other words, this theorem asserts that predecessor is a one-to-one relation on the natural numbers. To prove this theorem, it suffices to prove that predecessor is a one-to-one relation full stop. One can prove that predecessor is one-to-one from Hume's Principle, with the help of the following Equinumerosity Lemma, the proof of which is rather long and involved. The Equinumerosity Lemma asserts that when *F* and *G* are equinumerous, *x* falls under *F*, and *y* falls under *G*, then the concept *object falling under F other than x* is equinumerous to the concept *object falling under G other than y*. The picture is something like this:

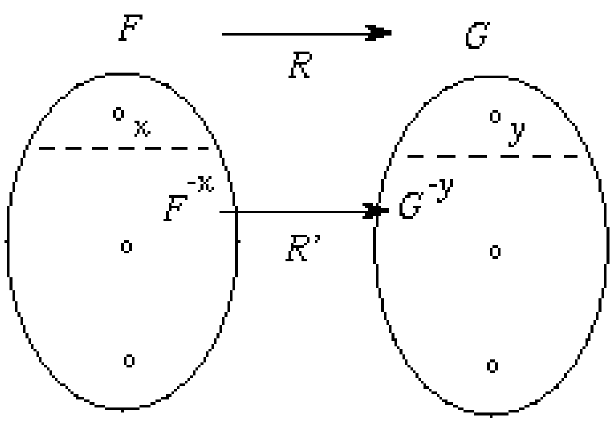


Figure 3

In terms of Figure 3, the Equinumerosity Lemma tells us that if there is a relation *R* which is a witness to the equinumerosity of *F* and *G*, then there is a relation R' which is a witness to the equinumerosity of the concepts that result when you restrict *F* and *G* to the objects other than *x* and *y*, respectively.

To help us formalize the Equinumerosity Lemma, let F^{-x} abbreviate the concept $[\lambda z\ Fz\ \&\ z \neq x]$ and let G^{-y} abbreviate the concept $[\lambda z\ Gz\ \&\ z \neq y]$. Then we have:

> **Equinumerosity Lemma**:
> $F \approx G\ \&\ Fx\ \&\ Gy \rightarrow F^{-x} \approx G^{-y}$

(Proof of Equinumerosity Lemma)

Now we can prove that Predecessor is a one-to-one relation from this Lemma and Hume's Principle (cf. **Gg I**, §108):

> **Predecessor is One-to-One**:
> $\forall x \forall y \forall z[Precedes(x,z) \,\&\, Precedes(y,z) \rightarrow x = y]$
>
> *Proof*: Assume that both *a* and *b* are precedessors of *c*. By the definition of predecessor, we know that there are concepts and objects *P*, *Q*, *d*, and *e*, such that:
>
> - $Pd \;\&\; c = \#P \;\&\; a = \#P^{-d}$
> - $Qe \;\&\; c = \#Q \;\&\; b = \#Q^{-e}$
>
> But if both $c = \#P$ and $c = \#Q$, then $\#P = \#Q$. So, by Hume's Principle, $P \approx Q$. So, by the Equinumerosity Lemma, it follows that $P^{-d} \approx Q^{-e}$. If so, then by Hume's Principle, $\#P^{-d} = \#Q^{-e}$. But then, $a = b$.

So, if Predecessor is a one-to-one relation, it is a one-to-one relation on the natural numbers. Therefore, no two numbers have the same successor.

It is important to mention here that not only is Predecessor a one-to-one relation, it is also a function:

> **Predecessor is a Function**:
> $\forall x \forall y \forall z[Precedes(x,y) \,\&\, Precedes(x,z) \rightarrow y = z]$

This fact can be proved with the help of a kind of converse to the Equinumerosity Lemma:

> **Equinumerosity Lemma 'Converse'**:
> $F^{-x} \approx G^{-y} \;\&\; Fx \;\&\; Gy \;\rightarrow\; F \approx G$

We leave the proof of the Equinumerosity Lemma 'Converse' and the proof that Predecessor is a function as exercises for the reader. The fact that Predecessor is a function will play a part in the proof that every number has a successor.

The Principle of Mathematical Induction

Let us say that a concept *F* is *hereditary on the natural numbers* just in case every 'adjacent' pair of numbers *n* and *m* (*n* preceding *m*) is such that *m* falls under *F* whenever *n* falls under *F*, i.e.,

> $HerOn(F,\mathbb{N}) \;=_{abbr}\; \forall n \forall m[Precedes(n,m) \rightarrow (Fn \rightarrow Fm)]$

Then we may state the Principle of Mathematical Induction as follows: if (a) 0 falls under *F* and (b) *F* is

hereditary on the natural numbers, then every natural number falls under *F*. In formal terms:

> **Principle of Mathematical Induction**:
> $F0 \;\&\; HerOn(F,\mathbb{N}) \rightarrow \forall n\, Fn$

Frege actually proves the Principle of Mathematical Induction from a more general principle that governs any *R*-series whatsoever. We will call the latter the General Principle of Induction. It asserts that whenever *a* falls under *F*, and *F* is hereditary on the *R*-series beginning with *a*, then every member of that *R*-series falls under *F*. We can formalize the General Principle of Induction with the help of a more strict understanding of 'hereditary on the *R*-series beginning with *a*'. Here is a definition:

$$HerOn(F, {}^{a}R^{+}) =_{abbr} \forall x \forall y[R^{+}(a,x) \;\&\; R^{+}(a,y) \;\&\; Rxy \;\rightarrow\; (Fx \rightarrow Fy)]$$

In other words, *F* is hereditary on the members of the *R*-series beginning with *a* just in case every adjacent pair *x* and *y* in this series (with *x* bearing *R* to *y*) is such that *y* falls under *F* whenever *x* falls under *F*. Now given this definition, we can reformulate the General Principle of Induction more strictly as:

> **General Principle of Induction**:
> $[Fa \;\&\; HerOn(F, {}^{a}R^{+})] \rightarrow \forall x[R^{+}(a,x) \rightarrow Fx]$

This is a version of Frege's Theorem 152 in **Gg I**, §117.

Frege's proves this claim by making an insightful appeal to his Rule of Substitution. We may sketch the proof strategy as follows. Assume that the antecedent of the General Principle of Induction holds for an arbitrarily chosen concept, say *P*. That is, assume:

$$Pa \;\;\&\;\; HerOn(P, {}^{a}R^{+})$$

Now to show $\forall x(R^{+}(a,x) \rightarrow Fx)$, pick an arbitrary object, say *b*, and further assume $R^{+}(a,b)$. We then simply have to show *Pb*. Frege does this by using the Rule of Substitution on Fact (6) about R^{+} (in our subsection on the Weak Ancestral in §4). Recall that Fact (6) is:

$$[Fx \;\&\; R^{+}(x,y) \;\&\; Her(F,R)] \rightarrow Fy$$

This is a theorem of logic containing the free variables *x*, *y*, and *F*. Frege instantiates *x* and *y* to *a* and *b*, respectively. He then, as we might put it, instantiates *F* to the concept $[\lambda z\, R^{+}(a,z) \;\&\; Pz]$ and applies λ-Conversion. (This is where Frege used his Rule of Substitution.) The concept being instantiated for *F* is the concept *member of the R-series beginning with a and which falls under P*. The result of instantiating all the free variables in Fact (6) and then applying λ-Conversion yields a rather long conditional, with numerous conjuncts in the antecedent and the claim that *Pb* in the consequent. Thus, if the antecedent can

be established, the proof is done. However, for those following along with pencil and paper, all of the conjuncts to this conditional are things we already know, with the exception of the claim that [λz R^+(*a*,*z*) & *Pz*] is hereditary on *R*. However, this claim can be established straightforwardly from things we know to be true (and, in particular, from facts contained in the antecedent of the Principle we are trying to prove, which we assumed as part of our conditional proof). The reader is encouraged to complete the proof as an exercise. For those who would like to check their work, we give the complete Proof of the General Principle of Induction here. (Proof of the General Principle of Induction)

Now to derive Principle of Mathematical Induction from the General Principle of Induction, we formulate the instance of the latter in which *a* is 0 and *R* is *Precedes*:

$$[F0 \ \& \ HerOn(F, {}^{0}Precedes^{+})] \ \rightarrow \ \forall x[Precedes^{+}(0,x) \rightarrow Fx]$$

When we expand the defined notation for *HerOn*, substitute the notation ℕ*x* and ℕ*y* for $Precedes^+(0,x)$ and $Precedes^+(0,y)$, respectively, and then employ our restricted quantifiers ∀*n*(...*n*...) and ∀*m*(...*m*...) for the claims of the form ∀*y*(ℕ*y* → ...*y*...) and ∀*x*(ℕ*x* → ...*x*...), respectively, the result is the Principle of Mathematical Induction (in which the notation *HerOn*(*F*,ℕ) has been eliminated in terms of its definiens).

Every Number Has a Successor

Frege uses the Principle of Mathematical Induction to prove that every number has a successor in the natural numbers. We may formulate the theorem as follows:

> **Theorem 5**:
> $\forall x[\mathbb{N}x \rightarrow \exists y(\mathbb{N}y \ \& \ Precedes(x,y))]$

To understand Frege's strategy for proving this theorem, recall that the weak ancestral of the Predecessor relation, i.e., $Precedes^+(x,y)$, can be read as: *x* is a member of the predecessor-series ending with *y*. Frege then considers the concept *member of the predecessor-series ending with n*, i.e., [λz $Precedes^+(z,n)$], where *n* is a natural number. Frege then shows, by induction, that every natural number *n* precedes the number of the concept *member of the predecessor-series ending with n*. That is, Frege proves that every number has a successor by proving the following Lemma on Successors by induction:

> **Lemma on Successors**:
> $\forall n \ Precedes(n, \#[\lambda z \ Precedes^{+}(z,n)])$

This asserts that every number *n* precedes the number of numbers in the predecessor series ending with *n*. Frege can establish Theorem 5 by proving the Lemma on Successors and by showing that the successor of a natural number is itself a natural number.

To see an intuitive picture of why the Lemma on Successors gives us what we want, we may temporarily

regard Precedes$^+$ as the relation $\leq$. (One can prove that Precedes$^+$ has the properties that $\leq$ has on the natural numbers.) Although we haven't yet defined the natural numbers following 0, the following intuitive sequence is driving Frege's strategy:

> 0 precedes $\#[\lambda z\, z \leq 0]$
> 1 precedes $\#[\lambda z\, z \leq 1]$
> 2 precedes $\#[\lambda z\, z \leq 2]$
> etc.

For example, the third member of this sequence is true because there are 3 natural numbers (0, 1, and 2) that are less than or equal to 2; so the number 2 precedes the number of numbers less than or equal to 2. Frege's strategy is to show that the general claim, that *n* precedes the number of numbers less than or equal to *n*, holds for every natural number. So, given this intuitive understanding of the Lemma on Successors, Frege has a good strategy for proving that every number has a successor. (For the remainder of this subsection, the reader may wish to continue to think of Precedes$^+$ in terms of $\leq$.)

Now to prove the Lemma on Successors by induction, we need to reconfigure this Lemma to a form which can be used as the consequent of the Principle of Induction; i.e., we need something of the form $\forall n\, Fn$. We can get the Lemma on Successors into this form by 'abstracting out' a concept from the Lemma using the right-to-left direction of λ-Conversion to produce the following equivalent statement of the Lemma:

$$\forall n\, [\lambda y\, Precedes(y, \#[\lambda z\, Precedes^+(z,y)])]n$$

The concept 'abstracted out' is the following:

$$[\lambda y\, Precedes(y, \#[\lambda z\, Precedes^+(z,y)])]$$

This is the concept: *being an object y which precedes the number of the concept: member of the predecessor series ending in y*. Let us abbreviate the λ-expression that denotes this concept as '*Q*'. Then Frege's strategy is to instantiate the variable *F* in the Principle of Induction (using his Rule of Substitution) to *Q*. The result is therefore something that we may take as having been proved:

$$Q0\ \&\ HerOn(Q,\mathbb{N}) \rightarrow \forall n\, Qn$$

Since the consequent is the Lemma on Successors, Frege can prove this Lemma by proving both that 0 falls under *Q* (cf. **Gg I**, Theorem 154) and that *Q* is hereditary on the natural numbers (cf. **Gg I**, Theorem 150):

> Proof that 0 falls under *Q*

Proof that *Q* is hereditary on the natural numbers

Given this proof of the Lemma on Successors, Theorem 5 is not far away. The Lemma on Successors shows that every number precedes some cardinal number of the form *#F*. We still have to show that such successor cardinals are natural numbers. That is, it still remains to be shown that if a number *n* precedes something *y*, then *y* is a natural number:

Successors of Natural Numbers are Natural Numbers:
$\forall n \forall y(Precedes(n,y) \rightarrow \mathbb{N}y)$

Proof: Suppose that *Precedes*(*n*,*a*). Then, by definition, since *n* is a natural number, $Precedes^{+}(0,n)$. So by Fact (2) about R^{+} (in the subsection on the Weak Ancestral in §4), it follows that $Precedes^{*}(0,a)$, and so by the definition of $Precedes^{+}$, it follows that $Precedes^{+}(0,a)$; i.e., *a* is a natural number.

Theorem 5 now follows from the Lemma on Successors and the fact that successors of natural numbers are natural numbers. With the proof of Theorem 5, we have completed the proof of Frege's Theorem. Before we turn to the last section of this entry, it is worth mentioning the mathematical significance of this theorem.

Arithmetic

From Frege's Theorem, one can derive arithmetic. It is an immediate consequence of the functionality of Predecessor that every number has a unique successor. That means we can define the successor function:

$n' =_{df}$ the *x* such that *Precedes*(*n*,*x*)

We may then define the sequence of natural numbers succeeding 0 as follows:

$1 = 0'$
$2 = 1'$
$3 = 2'$
etc.

Moreover, the recursive definition of addition can now be given:

$n + 0 = n$
$n + m' = (n + m)'$

We may also officially define:

$n < m =_{\mathrm{df}} \mathit{Precedes}^*(n,m)$
$n \leq m =_{\mathrm{df}} \mathit{Precedes}^+(n,m)$

These definitions constitute the foundations of arithmetic. Frege has insightfully isolated a group of basic laws in which they may be grounded.

§6: Philosophical Questions Surrounding Frege's Theorem

Frege's Theorem is an elegant derivation of the basic laws of arithmetic which can be carried out independently of the portion of Frege's system which led to inconsistency. Frege himself never identified "Frege's Theorem" as a "result". In **Gg**, he attempted to derive Hume's Principle and the Dedekind-Peano axioms from Basic Law V, but once the contradiction became known to him, he never officially retreated to the ‘fall-back’ position of claiming that the proof of the Dedekind-Peano axioms from Hume's Principle alone constituted an important result. One of several reasons why he didn't adopt this fall-back position is that he didn't regard Hume's Principle as a sufficiently general principle---he didn't believe it was strong enough, from an epistemological point of view, to help us answer the question, "How are numbers given to us?". We discuss the reasons for his attitude, among other things, in what follows.

A discussion of the philosophical questions surrounding Frege's Theorem should begin with some statement of how Frege conceived of his own project when writing **Begr**, **Gl**, and **Gg**. It seems clear that epistemological considerations in part motivated Frege's work on the foundations of mathematics. It is well documented that Frege had the following goal, namely, to explain our knowledge of the basic laws of arithmetic by giving an answer to the question "How are numbers ‘given’ to us?" which makes no appeal to the faculty of intuition. If Frege could show that the basic laws of number theory are derivable from analytic truths of logic, then he could argue that we need only appeal to the faculty of understanding (as opposed to some faculty of intuition) to explain our knowledge of the truths of arithmetic. Frege's goal then stands in contrast to the Kantian view of the exact mathematical sciences, according to which general principles of reasoning must be supplemented by a faculty of intuition if we are to achieve mathematical knowledge. The Kantian model here is that of geometry; Kant thought that our intuitions of figures and constructions played an essential role in the demonstrations of geometrical theorems. (In Frege's own time, the achievements of Frege's contemporaries Pasch, Pieri and Hilbert showed that such intuitions were not essential.)

Frege's Goals and Strategy in His Own Words

Frege's strategy then was to show that no appeal to intuition is required for the derivation of the theorems of number theory. This in turn required that he show that the latter are derivable using only rules of inference, axioms, and definitions that are purely analytic principles of logic. This view has become known as ‘Logicism’. Here is what Frege says:

[**Begr**, Preface, p. 5:]
To prevent anything intuitive from penetrating here unnoticed, I had to bend every effort to keep the chain of inferences free of gaps.
[from the Bauer-Mengelberg translation in van Heijenoort (1967)]

[**Begr**, Part III, §23:]
Through the present example, moreover, we see how pure thought, irrespective of any content given by the senses or even by an intuition a priori, can, solely from the content that results from its own constitution, bring forth judgements that at first sight appear to be possible only on the basis of some intuition. ... The propositions about sequences [*R*-series] in what follows far surpass in generality all those that can be derived from any intuition of sequences.
[from the Bauer-Mengelberg translation in van Heijenoort (1967)]

[**Gl**, §62:]
How, then, are numbers to be given to us, if we cannot have any ideas or intuitions of them? Since it is only in the context of a proposition that words have any meaning, our problem becomes this: To define the sense of a proposition in which a number word occurs.
[from the Austin translation in Frege (1974)]

[**Gl**, §87:]
I hope I may claim in the present work to have made it probable that the laws of arithmetic are analytic judgements and consequently a priori. Arithmetic thus becomes simply a development of logic, and every proposition of arithmetic a law of logic, albeit a derivative one.
[from the Austin translation in Frege (1974)]

[**Gg I**, §0:]
In my *Grundlagen der Arithmetik*, I sought to make it plausible that arithmetic is a branch of logic and need not borrow any ground of proof whatever from either experience or intuition. In the present book, this shall be confirmed, by the derivation of the simplest laws of Numbers by logical means alone.
[from the Furth translation in Frege (1967)]

[**Gg II**, Appendix:]
The prime problem of arithmetic is the question, In what way are we to conceive logical objects, in particular, numbers? By what means are we justified in recognizing numbers as objects? Even if this problem is not solved to the degree I thought it was when I wrote this volume, still I do not doubt that the way to the solution has been found.
[from the Furth translation in Frege (1967)]

The Basic Problem for Frege's Strategy

The basic problem for Frege's strategy, however, is that for his logicist project to succeed, his system must at some point include (either as an axiom or theorem) statements that explicitly assert the existence of certain kinds of abstract entities and it is not obvious how to justify the claim that we know such explicit existential statements. Given our description of his system, it should be clear that Frege's logical system includes existence claims for the following entities:

- concepts
- extensions
- truth-values
- numbers

Although Frege attempted to reduce the latter two kinds of entities (truth-values and numbers) to extensions, the fact is that the existence of concepts and extensions are implied by his Rule of Substitution and Basic Law V, respectively. Logic, it is often argued, should be free of such existence assumptions. A Kantian might well complain both that explicit existence claims seem to be synthetic rather than analytic (i.e., such claims don't seem to be true in virtue of the meanings of the words involved) and that since the Rule of Substitution and Basic Law V imply existence claims, Frege cannot claim that such principles are purely analytic principles of logic. If so, then some other faculty (such as intuition) might still be needed to account for our knowledge of (the existence claims of) arithmetic.

The Existence of Concepts

Boolos (1985) was the first to note that the Rule of Substitution causes a problem of this kind for Frege's program, since it is equivalent to a quite liberal existential claim, namely, the Comprehension Principle for Concepts. Boolos suggests a defense for Frege with respect to this particular aspect of his logic, namely, to reinterpret (by paraphrasing) the second-order quantifiers so as to avoid commitment to concepts. (See Boolos (1985) for the details.) Boolos's suggestion, however, is one which would require Frege to abandon his realist theory of concepts. Moreover, although Boolos' suggestion might lead us to an epistemological justification of the Comprehension Principle for Concepts, it doesn't do the same for the Comprehension Principle for Relations, for his reinterpretation of the quantifiers works only for the ‘monadic’ quantifiers (i.e., those ranging over concepts having one argument) and thus doesn't offer a paraphrase for quantification over relational concepts.

Another problem for a strategy of the type suggested by Boolos is that if the second order quantifiers are interpreted so that they do not range over a separate domain of entities, then there is nothing appropriate to serve as the denotations of λ-expressions. Although Frege wouldn't quite put it this way, we have seen that his system treats open formulas with free object variables as if they denoted concepts. Although Frege doesn't use λ-notation, the use of such notation seems to be the most logically perspicuous way of reconstructing his work. The use of such notation faces the same epistemological puzzles that Frege's Rule of Substitution faces.

To see why, note that the Principle of λ-Conversion:

$$\forall y([\lambda x\, \varphi(x)]y \equiv \varphi(y/x))$$

seems to be an analytic truth of logic. It says this:

> An object *y* exemplifies the complex property *being an x such that* $\varphi(x)$ if and only if *y* satisfies (in Tarski's sense) φ

One might argue that this is true in virtue of the very meaning of the λ-expression, the meaning of $\equiv$, and the meaning of the statement form *Fx*. However, λ-Conversion also implies the Comprehension Principle for Concepts, for the latter follows from the former by existential generalization:

$$\exists F \forall y(Fy \equiv \varphi(y/x))$$

The point here is that the fact that an existential claim is derivable casts at least some doubt on the purely analytic status of λ-Conversion. The question of how we obtain knowledge of such principles is still an open question in philosophy. It is an important question to address, since Frege's most insightful definitions are cast using quantifiers ranging over concepts and relations (e.g., the ancestrals of a relation) and it would be useful to have a philosophical explanation of how such entities and the principles which govern them become known to us. In contemporary philosophy, this question is still poignant, since many philosophers do accept that *properties* and *relations* of various sorts exist. These entities are the contemporary analogues of Frege's concepts.

The Existence of Extensions

We have also seen (§2) that the Corollary to Basic Law V implies the existence of extensions. The question for Frege's project, then, is why should we accept as a law of logic a statement that implies the existence of individuals? Frege did conceive of Basic Law V as a law of logic:

> [**Gg I**, Preface, p. 3:]
> A dispute can arise, so far as I can see, only with regard to my Basic Law concerning courses-of-values (V)... I hold that it is a law of pure logic.
> [from the Furth translation in Frege (1967)]

Moreover, he thought that an appeal to extensions would answer one of the questions that motivated his work:

> [**Letter to Russell**, July 28, 1902:]
> I myself was long reluctant to recognize ranges of values and hence classes [sets]; but I saw no other possibility of placing arithmetic on a logical foundation. But the question is, How

> do we apprehend logical objects? And I have found no other answer to it than this, We apprehend them as extensions of concepts, or more generally, as ranges of values of functions.
>
> [from the Kaal translation in Frege (1980)]

Now it is unclear why Frege thought that he could answer the question posed here with the reply "We apprehend numbers as extensions of concepts". He seems to think we can answer the obvious next question "How do we apprehend extensions?" by saying "by way of Basic Law V". His idea here seems to be that since Basic Law V is supposed to be purely analytic or true in virtue of the meanings of its terms, we apprehend a pair of extensions whenever we truly judge that concepts *F* and *G* are materially equivalent. Some philosophers argue that Frege would have been correct to argue in just this way (had Basic Law V been consistent). They argue that Basic Law V (or consistent principles having the same logical form) justifies *reference* to the entities described in the left-side condition by grounding such reference in the *truth* of the right-side condition.[11]

But this, of course, raises an obvious problem. To justify reference to extensions, we must first justify the claim that those extensions exist. It is not clear that the claim that concepts are materially equivalent can justify such an existence claim. But given Frege's view that Basic Law V is analytic, it seems that he must hold that the right-side condition implies the corresponding left-side condition as a matter of meaning.[12]

This view, however, runs up against the following argument. Suppose the right hand condition implies the left-side condition as a matter of meaning. That is, suppose that (R) implies (L) as a matter of meaning:

(R) $\forall x(Fx \equiv Gx)$

(L) $\grave{\varepsilon}F\varepsilon = \grave{\alpha}G\alpha$

Now note that (L) itself can be analyzed, from a logical point of view. No matter whether we construe $\grave{\varepsilon}$ as an operator or a kind of definite description, (L) (i.e., "the extension of *F* = the extension of *G*") can be analyzed as the claim:

> There is an object *x* and an object *y* such that:
> (1) *x* is a unique extension of *F*,
> (2) *y* is a unique extension of *G*, and
> (3) *x* = *y*.

That is, for some defined or primitive notion *Extension*(*x*,*F*) (‘*x* is an extension of *F*’), (L) implies the analysis (D) as a matter of meaning:

(D) $\exists x\exists y[Extension(x,F)\ \&\ \forall z(Extension(z,F) \rightarrow z{=}x)\ \&$
$Extension(y,G)\ \&\ \forall z(Extension(z,G) \rightarrow z{=}y)\ \&$

$x = y$]

But if (R) implies (L) as a matter of meaning, and (L) implies (D) as a matter of meaning, then (R) implies (D) as a matter of meaning. This seems doubtful. The material equivalence of *F* and *G* does not imply the existence claim (D) as a matter of meaning, whatever notion of meaning is involved. [This argument attempts to show why Va (i.e., the right-to-left direction of Basic Law V) is not analytic. Below, it will be adapted to show that the right-to-left direction of Hume's Principle is not analytic. See Boolos (1997, 307 - 309), for reasons why Vb and the left-to-right direction of Hume's Principle are not analytic.]

The moral to be drawn here is that the modern Fregean must attempt to explain our knowledge of existence claims for abstract objects such as extensions head on, and not try to justify them indirectly, by attempting to justify claims that imply such existence claims. Even if we follow Frege in conceiving of extensions as ‘logical objects’, the question remains as to how the very claims that such objects exist can be true on logical or analytic grounds alone. We might agree that there must be logical objects of some sort if logic is to have a subject matter, but if Frege is to achieve his goal of showing that our knowledge of arithmetic is free of intuition, then the logical knowledge with which he identifies arithmetical knowledge must be either be *purely* analytic or shown otherwise to be free of intuition. We'll return to this theme in the final subsection.

The Existence of Numbers and Truth-Values: The Julius Caesar Problem

Given that the proof of Frege's Theorem makes no appeal to Basic Law V, some philosophers have argued Frege's best strategy for achieving his goal is to replace Basic Law V with Hume's Principle and argue that Hume's Principle is an analytic principle of logic.[13] However, we have just seen one reason why such a strategy does not suffice. The claim that Hume's Principle is an analytic principle of logic is subject to the same problem just posed for Basic Law V. The equinumerosity of *F* and *G* does not, as a matter of meaning, imply (identity claims that entail) the existence of numbers. When we analyze "$\#F = \#G$" in the same way that we analyzed "$\acute{\varepsilon}F\varepsilon = \acute{\alpha}G\alpha$" (i.e., by analyzing away the operator # or definite description "the number of *F*s" in terms of existence and uniqueness claims), it becomes clear that the equinumerosity of *F* and *G* does not, as a matter of meaning, imply the result of the analysis.

Moreover, Frege had his own reasons for not replacing Basic Law V with Hume's Principle. One reason was that he thought Hume's Principle offered no answer to the epistemological question, ‘How do we grasp or apprehend logical objects, such as the numbers?’. But Frege had another reason for not substituting Hume's Principle for Basic Law V, namely, that Hume's Principle would be subject to ‘the Julius Caesar problem’. Frege first raises this problem in connection with an inductive definition of ‘$n = \#F$’ that he tries out in **Gl**, §55. Concerning this definition, Frege says:

> [**Gl**, §55:]
> ... but we can never -- to take a crude example -- decide by means of our definitions whether any concept has the number Julius Caesar belonging to it, or whether that

conqueror of Gaul is a number or is not.

[from the Austin translation in Frege (1974)]

Frege raises this same concern again for a contextual definition that gives a ‘criterion of identity’ for the objects being defined. In **Gl** §66, Frege considers the following contextual definition of ‘the direction of line *x*’:

The direction of line *a* = the direction of line *b* if and only if *a* is parallel to *b*.

With regard to this definition, Frege says:

[**Gl**, §66:]

It will not, for instance, decide for us whether England is the same as the direction of the Earth's axis---if I may be forgiven an example which looks nonsensical. Naturally no one is going to confuse England with the direction of the Earth's axis; but that is no thanks to our definition of direction.

[from the Austin translation in Frege (1974)]

Now trouble for Hume's Principle begins to arise when we recognize that it is a contextual definition that has the same logical form as this definition for directions. It is central to Frege's view that the numbers are *objects*, and so he believes that it is incumbent upon him to say *which* objects they are. But the ‘Julius Caesar problem’ is that Hume's Principle, if considered as the sole principle offering identity conditions for numbers, doesn't describe the conditions under which an arbitrary object, say Julius Caesar, is or is not to be identified with the number of planets. That is, Hume's Principle doesn't define the condition ‘$\#F=x$’, for arbitrary *x*. It only offers identity conditions when *x* is an object we know to be a cardinal number (for then $x=\#G$, for some *G*, and Hume's Principle tells us when $\#F=\#G$).

In **Gl**, Frege solves the problem by giving his explicit definition of numbers in terms of extensions. (We described this in §4 above.) Unfortunately, this is only a stopgap measure, for when Frege later systematizes extensions in **Gg**, Basic Law V has the same logical form as Hume's Principle and the above contextual definition of directions. Frege is aware that the Julius Caesar problem affects Basic Law V, though. In **Gg I**, §10, Frege appears to raise the Julius Caesar problem for extensions of concepts. With respect to Basic Law V, he says:

[**Gg I**, §10:]

...this by no means fixes completely the denotation of a name like ‘$\grave{\varepsilon}\Phi(\varepsilon)$’. We have only a means of always recognizing a course-of-values if it is designated by a name like ‘$\grave{\varepsilon}\Phi(\varepsilon)$’, by which it is already recognizable as a course-of-values. But, we can neither decide, so far, whether an object is a course-of-values that is not given us as such ...

[from the Furth translation in Frege (1967)]

In other words, Basic Law V does not tell us the conditions under which an arbitrarily chosen object *x*

may be identified with some given extension, such as $\grave{\varepsilon} F\varepsilon$.

Until recently, it was thought that Frege solved this problem in §10 by restricting the universal quantifier $\forall x$ of his **Gg** system so that it ranges only over extensions. If Frege could have successfully restricted this quantifier to extensions, then when the question arises, is (arbitrarily chosen) object *x* is identical with $\grave{\varepsilon} F\varepsilon$, one could answer that *x* has to be the extension of some concept, say *G* and that Basic Law V would then tell you the conditions under which *x* is identical to $\grave{\varepsilon} F\varepsilon$. On this interpretation of §10, Frege is alleged to have restricted the quantifiers when he identified the two truth values (The True and The False) with the two extensions that contain just these objects as members, respectively. By doing this, it was thought that all of the objects in the range of his quantifier $\forall x$ in **Gg** become extensions which have been identified as such, for the truth values were the only two objects of his system that had not been introduced as extensions or courses of value.

However, recent work by Wehmeier (1999) suggests that, in §10, Frege was not attempting to restrict the quantifiers of his system to extensions (nor, more generally, to courses-of-values). The extensive footnote to §10 indicates that Frege considered, but did not hold much hope of, identifying every object in the domain with the extension consisting of just that object.[14] But, more importantly, Frege later considers cases (in **Gg**, Sections 34 and 35) which seem to presuppose that the domain contains objects which aren't extensions. (In these sections, Frege considers what happens to the definition of '*x* is a member of *y*' when *y* is not an extension.)[15]

Even if Frege somehow could have successfully restricted the quantifiers of **Gg** to avoid the Julius Caesar problem, he would no longer have been able to extend his system to include names of ordinary non-logical objects. For if he were to attempt to do so, the question, "Under what conditions is $\grave{\varepsilon} F\varepsilon$ identical with Julius Caesar?", would then be legitimate but have no answer. That means his logical system could not be used for the analysis of ordinary language. But it was just the analysis of ordinary language that led Frege to his insight that a statement of number is an assertion about a concept.

Final Observations

Even when we replace the inconsistent Basic Law V with the powerful Hume's Principle, Frege's work still leaves two questions unanswered: (1) How do we know that numbers exist?, and (2) How do we precisely specify which objects they are? The first question arises because Hume's Principle doesn't seem to be a purely analytic truth of logic; by what faculty do we come to know (the truth of) the existential claim that numbers exist if neither Hume's Principle nor this existential claim is analytically true? The second question arises because Frege's work offers no general condition under which we can identify an arbitrarily chosen object *x* with a given number such as the number of planets; how then can Frege claim to have precisely specified which objects the numbers are within the domain of all logical and non-logical objects? So questions about the very existence and identity of numbers still plague Frege's work.

These two questions arise because of a limitation in the logical form of these Fregean biconditional

principles such as Hume's Principle and Basic Law V. These contextual definitions attempt to do two jobs which modern logicians now typically accomplish with separate principles. A properly reformulated theory of ‘logical’ objects should have: (1) a separate *non-logical* comprehension principle which explicitly asserts the existence of *logical objects*, and (2) a separate identity principle which asserts the conditions under which *logical objects* are identical. The latter should specify identity conditions for logical objects in terms of their most salient characteristic, one which distinguishes them from other objects. Such an identity principle would then be more specific than the global identity principle for all objects (Leibniz's Law) which asserts that if objects x and y fall under the same concepts, they are identical.

By way of example, consider modern set theory. Zermelo set theory (Z) has a distinctive *non-logical* comprehension principle for sets:

> **Subset Axiom of Z**:
> $\forall x[Set(x) \rightarrow \exists y[Set(y) \,\&\, \forall z(z \in y \equiv z \in x \,\&\, \varphi(z))]]$,
> where $\varphi(z)$ is any formula in which the variable z is free and which has no free variables y

Z has a separate identity principle:

> **Identity Principle for Sets**:
> $Set(x) \,\&\, Set(y) \rightarrow [\forall z(z \in x \equiv z \in y) \rightarrow x = y]$

Note that the second principle offers identity conditions in terms of the most salient features of sets, namely, the fact that they, unlike other objects, have members. The identity conditions for objects which *aren't* sets, then, can be the standard principle that identifies objects whenever they fall under the same concepts. This leads us naturally to a very general principle of identity for any objects whatever:

> **General Principle of Identity**:
> $x = y \;\; =_{df} \;\; [Set(x) \,\&\, Set(y) \,\&\, \forall z(z \in x \equiv z \in y)] \;\; \vee$
> $\qquad\qquad [\neg Set(x) \,\&\, \neg Set(y) \,\&\, \forall F(Fx \equiv Fy)]$

Now, if something is given to us *as a set* and we ask whether it is identical with an arbitrarily chosen object x, this specifies a clear condition that settles the matter. The only questions that remain for the theory Z concern its existence principle: Do we know that the comprehension principle is true, and if so, how? The question of existence is thus laid bare. We do not approach it by attempting to justify a principle that implies the existence of sets via definite descriptions which we don't yet know to be well-defined.

In his classic essays (1987) and (1986), Boolos appears to recommend this very procedure of using separate existence and identity principles. In those essays, he eschews the primitive mathematical relation of set membership and suggests that Frege formulate his theory of numbers (‘Frege Arithmetic’) by using a single *nonlogical* comprehension axiom which employs a special instantiation relation that holds

between a concept G and an object x whenever, intuitively, x is an extension consisting solely of concepts and G is a concept ‘in’ x. He calls this nonlogical axiom ‘Numbers’ and uses the notation ‘$G\eta x$’ to signify that G is in x:

> **Numbers**:
> $\forall F \exists ! x \forall G(G \eta x \equiv G \approx F)$

[See Boolos (1987), p. 5; and (1986), p. 140.] This principle asserts that for any concept F, there is a unique object which contains in it all and only those concepts G which are equinumerous to F. Boolos then makes two observations: (1) that Frege can then define #F as "the unique object x such that for all concepts G, G is in x iff G is equinumerous to F", and (2) that Hume's Principle is derivable from Numbers. [See Boolos (1986), p. 140.] Given these observations, we know from our work in §§4 and 5 above that Numbers suffices for the derivation of the basic laws of arithmetic.

Since Boolos calls this principle ‘Numbers’, it is no stretch to suppose that he would accept the following explicit reformulation (in which ‘*Number*(x)’ is an undefined, primitive notion):

> **Numbers**:
> $\forall F \exists ! x[\mathit{Number}(x) \;\&\; \forall G(G \eta x \equiv G \approx F)]$

Though Boolos doesn't explicitly formulate an identity principle to complement Numbers, it seems clear that the following principle would offer identity conditions in terms of the most distinctive feature of numbers:

> **Identity Principle for Numbers**:
> $\mathit{Number}(x) \;\&\; \mathit{Number}(y) \rightarrow [\forall G(G \eta x \equiv G \eta y) \rightarrow x = y]$

It is then straightforward to formulate a general principle of identity, as we did in the case of the set theory Z:

> **General Principle of Identity**:
> $x = y \;=_{df}\; [\mathit{Number}(x) \;\&\; \mathit{Number}(y) \;\&\; \forall F(F \eta x \equiv F \eta y)] \;\lor$
> $[\neg \mathit{Number}(x) \;\&\; \neg \mathit{Number}(y) \;\&\; \forall F(Fx \equiv Fy)]$

This formulation of Frege Arithmetic, in terms of Numbers and the General Principle of Identity, puts the Julius Caesar problem (described above) into better perspective; the condition ‘#F=x’ is defined for arbitrary concepts F and objects x. It openly faces the epistemological questions head-on: Do we know that Numbers is true, and if so, how? This is where philosophers need to concentrate their energies. [For a reconstruction of Frege Arithmetic with a more general version of the special instantiation relation η, see Zalta (1999).]

By replacing Fregean biconditionals such as Hume's Principle with explicit existence and identity

principles, we reduce two problems to one and and isolate the real problem for Fregean foundations of arithmetic, namely, the problem of giving an epistemological justification of existence claims (e.g., Numbers) for abstract objects of a certain kind. For anything like Frege's program to succeed, it must at some point explicitly assert (as an axiom or theorem) the existence of (logical) objects of some kind. Those explicit existence claims should be the focus of attention, for they are the point at which logic and metaphysics dovetail. The theory of logical objects, if carried out without any mathematical primitives, should simply be acknowledged as a nonlogical metaphysical theory, not a piece of logic. A proper epistemology for such a theory should offer some epistemological justification of the explicit existence claims that serve as the basic axioms of the theory. That is the moral to be drawn from Frege's work.

Bibliography

- Beaney, M., 1997, *The Frege Reader*, Oxford: Blackwell
- Bell, J. L., 1995, ‘Type-Reducing Correspondences and Well-Orderings: Frege's and Zermelo's Construction Re-examined’, *Journal of Symbolic Logic*, **60**: 209-221.
- Bell, J. L., 1999, ‘Frege's Theorem in a Constructive Setting’, *Journal of Symbolic Logic*, **64**/2: 486-488
- Bell, J.L., 1994, ‘Fregean Extensions of First-Order Theories’, *Mathematical Logic Quarterly*, **40**: 27-30; reprinted in Demopoulos 1995, 432-437.
- Boolos, G., 1985, ‘Reading the *Begriffsschrift*’, *Mind*, **94**: 331 - 344; reprinted in Boolos (1998): 155-170. [Page references are to the reprint.]
- Boolos, G., 1986, ‘Saving Frege From Contradiction’, in *Proceedings of the Aristotelian Society*, **87** (1986/1987): 137 - 151; reprinted in Boolos (1998): 171-182. [Page references are to the original.]
- Boolos, G., 1987, ‘The Consistency of Frege's *Foundations of Arithmetic*’, in *On Being and Saying*, J. J. Thomson (ed.), Cambridge, MA: MIT Press, pp. 3-20; reprinted in Boolos (1998): 183-201. [Page references are to the original.]
- Boolos, G., 1990, ‘The Standard of Equality of Numbers’, in *Meaning and Method: Essays in Honor of Hilary Putnam*, G. Boolos (ed.), Cambridge: Cambridge University Press, pp. 261-277; reprinted in Boolos (1998): 202-219. [Page references are to the original.]
- Boolos, G., 1993, ‘Whence the Contradiction?’, in *Aristotelian Society Supplementary Volume*, **67**: 213-233; reprinted in Boolos (1998): 220-236.
- Boolos, G., 1994, ‘The Advantages of Honest Toil Over Theft’, in *Mathematics and Mind*, Alexander George (ed.), Oxford: Oxford University Press, 27-44; reprinted in Boolos (1998): 255 - 274.
- Boolos, G., 1997, ‘Is Hume's Principle Analytic?’, in Heck (1997a), 245-262; reprinted in Boolos (1998): 301-314. [Page references are to the reprint.]
- Boolos, G., 1998, *Logic, Logic, and Logic*, J. Burgess and R. Jeffrey (eds.), Cambridge, MA: Harvard University Press.
- Burgess, J., 1984, ‘Review of Wright (1983)’, *The Philosophical Review*, **93**: 638-40.
- Burgess, J., 1998, ‘On a Consistent Subsystem of Frege's *Grundgesetze*’, *Notre Dame Journal of Formal Logic*, **39**: 274-278

- Demopoulos, W., (ed.), 1995, *Frege's Philosophy of Mathematics*, Cambridge: Harvard University Press.
- Demopoulos, W., forthcoming-a, ‘Gottlob Frege’, *The Garland Encyclopedia of Philosophy of Science*, M. Hallett (ed.), Garland Press
- Demopoulos, W., 1998, ‘The Philosophical Basis of Our Knowledge of Number’, *Nous*, **32**: 481-503
- Dummett, M., 1991, *Frege: Philosophy of Mathematics*, Cambridge: Harvard University Press.
- Dummett, M., 1997, ‘Neo-Fregeans: In Bad Company?’, in Schirn (1997)
- Field, H., 1984, ‘Critical Notice of Crispin Wright: *Frege's Conception of Numbers as Objects*’, *Canadian Journal of Philosophy*, **14**: 637-632; reprinted under the title ‘Platonism for Cheap? Crispin Wright on Frege's Context Principle’ in H. Field, *Realism, Mathematics, and Modality*, Oxford: Blackwell, 1989, pp. 147-170
- Frege, G., 1980, *Philosophical and Mathematical Correspondence*, G. Gabriel, H. Hermes, F. Kambartel, C. Thiel, and A. Veraart (eds. of the German edition), abridged from the German edition by Brian McGuinness, translated by Hans Kaal, Chicago: University of Chicago Press
- Frege, G., 1974, *The Foundations of Arithmetic*, J. L. Austin (trans.), Oxford: Basil Blackwell
- Frege, G., 1967, *The Basic Laws of Arithmetic*, M. Furth (trans.), Berkeley: University of California
- Furth, M., 1967, ‘Editor's Introduction’, in G. Frege, *The Basic Laws of Arithmetic*, M. Furth (translator and editor), Berkeley: University of California Press, pp. v-lvii
- Goldfarb, W., 2001, ‘First-Order Frege Theory is Undecidable’, *Journal of Philosophical Logic*, **30**: 613-616.
- Hale, B., 1994, ‘Dummett's Critique of Wright's Attempt to Resuscitate Frege’, *Philosophia Mathematica*, (Series III), **2**: 122-147.
- Hazen, A., 1985, ‘Review of Crispin Wright's *Frege's Conception of Numbers as Objects*’, *Australasian Journal of Philosophy*, **63**/2 (June): 251-254
- Heck, R., 1996, ‘The Consistency of Predicative Fragments of Frege's *Grundgesetze der Arithmetik, History and Philosophy of Logic*, **17**: 209-220
- Heck, R., (ed.), 1997a, *Language, Thought, and Logic: Essays in Honour of Michael Dummett*, Oxford: Oxford University Press.
- Heck, R., 1997b, ‘The Julius Caesar Objection’, in Heck (1997a), 273-308
- Heck, R., 1993, ‘The Development of Arithmetic in Frege's *Grundgesetze der Arithmetik*’, *Journal of Symbolic Logic*, **58**/2 (June): 579-600; reprinted in Demopoulos (1995).
- Heck, R., 1996, ‘The Consistency of Predicative Fragments of Frege's *Grundgesetze der Arithmetik*’, *History and Philosophy of Logic*, **17**: 209-220
- Parsons, C., 1965, ‘Frege's Theory of Number’, *Philosophy in America*, M. Black (ed.), Ithaca: Cornell University Press, pp. 180-203; reprinted with Postscript in Demopoulos (1995), pp. 182-210
- Parsons, T., 1987, ‘The Consistency of the First-Order Portion of Frege's Logical System’, *Notre Dame Journal of Formal Logic*, **28**: 161-68.
- Pelletier, F.J., 2001, "Did Frege Believe Frege's Principle", *Journal of Logic, Language, and Information*, **10**/1: 87-114
- Resnik, M., 1980, *Frege and the Philosophy of Mathematics*, Ithaca: Cornell University Press

- Rosen, G., 1993, ‘The Refutation of Nominalism(?)’, *Philosophical Topics*, **21**/2: 149-186
- Schirn, M., (ed.), 1997, *Philosophy of Mathematics Today*, Oxford: Oxford University Press.
- Schroeder-Heister, P., 1987, ‘A model-theoretic reconstruction of Frege's Permutation Argument’, *Notre Dame Journal of Formal Logic*, **28**/1: 69-79
- Sullivan, P. and Potter, M., 1997, ‘Hale on Caesar’, *Philosophia Mathematica*, (Series III), **5**: 135-152.
- Tabata, H., 2000, ‘Frege's Theorem and His Logicism’, *History and Philosophy of Logic*, **21**/4: 265-295
- van Heijenoort, J., 1967, ed., *From Frege to Gödel: A Sourcebook in Mathematical Logic*, Cambridge: Harvard University Press
- Wehmeier, K., 1999, ‘Consistent Fragments of *Grundgesetze* and the Existence of Non-Logical Objects’, *Synthese*, **121**: 309-328
- Whitehead, A. N. and Russell, B., 1912, *Principia Mathematica Vol. II*, Cambridge: Cambridge University Press.
- Wright, C., 1983, *Frege's Conception of Numbers as Objects*, Aberdeen: Aberdeen University Press.
- Wright, C., 1997a, ‘Response to Dummett’, in Schirn (1997)
- Wright, C., 1997b, ‘On the Philosophical Significance of Frege's Theorem’, in Heck (1997), 201-244
- Zalta, E., 1999, ‘Natural Numbers and Natural Cardinals as Abstract Objects: A Partial Reconstruction of Frege's *Grundgesetze* in Object Theory’, *Journal of Philosophical Logic*, **28**/6 (1999): 619-660

Other Internet Resources

- Die Grundlagen der Arithmetik, original German text (maintained by Alain Blachair, Académie de Nancy-Metz)

Related Entries

Frege, Gottlob | *Principia Mathematica* | Russell, Bertrand | Russell’s paradox

Acknowledgements

I am indebted to William Demopoulos, whose short essay "Gottlob Frege" for the forthcoming Garland Encyclopedia motivated me to write the present entry. Demopoulos has kindly allowed me to quote certain passages from that essay in the footnotes to the present entry. I am also indebted to Roberto Torretti, who carefully read this piece and identified numerous infelicities, and to Xu Mingming, who noticed that Fact 7 about the Weak Ancestral (Section 4, subsection "The Weak Ancestral of *R*") was missing an important condition (namely, that *R* must be 1-1). Finally, I am indebted to Kai Wehmeier,

who reminded me that, strictly speaking, the result of replacing Basic Law V by Hume's Principle in Frege's system does not result in a subsystem of the original, since we have to replace the primitive notion "the course of values of the function *f*" with the primitive notion "the number of *F*s".

Edward N. Zalta
zalta@stanford.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 10, 1998
Content last modified: May 9, 2002

Frege's Logic and Foundations for Arithmetic

Complete Table of Contents
With Listing of Subsections

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

A More Complex Example

Example of an Inference Using the Definition of Membership

If given the premise that

$$1 + 2^2 = 5$$

one can prove that

$$1 \in \acute{\epsilon}(\epsilon + 2^2 = 5)$$

For it follows from our premise (by λ-Abstraction) that:

$$[\lambda z\, z + 2^2 = 5]1$$

Independently, by the logic of identity, we know:

$$\acute{\epsilon}([\lambda z\, z + 2^2 = 5]\epsilon) \;=\; \acute{\epsilon}([\lambda z\, z + 2^2 = 5]\epsilon)$$

So we may conjoin this fact and the result of λ-Abstraction to produce:

$$\acute{\epsilon}([\lambda z\, z + 2^2 = 5]\epsilon) = \acute{\epsilon}([\lambda z\, z + 2^2 = 5]\epsilon) \;\;\&\;\; [\lambda z\, z + 2^2 = 5]1$$

Then, by existential generalization on the concept $[\lambda z\, z + 2^2 = 5]$, it follows that:

$$\exists G[\acute{\epsilon}([\lambda z\, z + 2^2 = 5]\epsilon) = \acute{\epsilon}G\epsilon \;\;\&\;\; G1]$$

By the definition of membership, we obtain:

$$1 \in \acute{\epsilon}([\lambda z\, z + 2^2 = 5]\epsilon)$$

And, finally, by our Rewrite Rule, we establish what we set out to prove:

$1 \in \grave{\varepsilon}(\varepsilon + 2^2 = 5)$

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: October 12, 1998
Content last modified: October 12, 1998

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Proof of the Law of Extensions

We want to show, for an arbitrarily chosen concept P and an arbitrarily chosen object c, that $c \in \grave{\varepsilon}P\varepsilon \equiv Pc$.

($\rightarrow$) Assume $c \in \grave{\varepsilon}P\varepsilon$ (to show Pc). Then, by the definition of $\in$, it follows that:

$$\exists H(\grave{\varepsilon}P\varepsilon = \grave{\alpha}H\alpha \;\&\; Hc)$$

Suppose that Q is such a property. Then, we know:

$$\grave{\varepsilon}P\varepsilon = \grave{\alpha}Q\alpha \;\&\; Qc$$

But, by Basic Law V, the first conjunct implies $\forall x(Px \equiv Qx)$. So from the fact that Qc, it follows that Pc.

($\leftarrow$) Assume Pc (to show $c \in \grave{\varepsilon}P\varepsilon$). Then, by the Corollary to Basic Law V, P has an extension, namely, $\grave{\varepsilon}P\varepsilon$. So by the laws of identity, we know $\grave{\varepsilon}P\varepsilon = \grave{\varepsilon}P\varepsilon$. We may conjoin this with our assumption to conclude:

$$\grave{\varepsilon}P\varepsilon = \grave{\varepsilon}P\varepsilon \;\&\; Pc$$

Now by existential generalizing on the concept P, it follows that:

$$\exists H(\grave{\varepsilon}P\varepsilon = \grave{\varepsilon}H\varepsilon \;\&\; Hc)$$

Thus, by the definition of $\in$, it follows that $c \in \grave{\varepsilon}P\varepsilon$.

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Proof of the Principle of Extensionality from Basic Law V

Assume $Extension(x)$ and $Extension(y)$. Then $\exists F(x = \grave{\varepsilon} F\varepsilon)$ and $\exists F(y = \grave{\alpha} F\alpha)$. Let, P, Q be arbitrary such concepts; i.e., suppose $x = \grave{\varepsilon} P\varepsilon$ and $y = \grave{\alpha} Q\alpha$.

Now to complete the proof, assume $\forall z(z \in x \equiv z \in y)$. It then follows that $\forall z(z \in \grave{\varepsilon} P\varepsilon \equiv z \in \grave{\alpha} Q\alpha)$. So, by the Law of Extensions and the principles of predicate logic, we may convert both conditions in the universalized biconditional to establish that $\forall z(Pz \equiv Qz)$. So, by Basic Law V, $\grave{\varepsilon} P\varepsilon = \grave{\alpha} Q\alpha$. So $x = y$.

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

First Derivation of the Contradiction

The λ-expression which denotes the concept *being the extension of a concept which you don't fall under* is:

$$[\lambda x\, \exists F(x = \grave{\alpha} F\alpha \ \&\ \neg Fx)]$$

As we saw in the text, we know that such a concept as this exists, by the Comprehension Principle for Concepts. Let 'G' abbreviate this name of the concept. So $\grave{\varepsilon}G\varepsilon$ exists, by the Corollary to Basic Law V. Now suppose $\grave{\varepsilon}G\varepsilon$ falls under the concept G; i.e., suppose:

$$[\lambda x\, \exists F(x = \grave{\alpha} F\alpha \ \&\ \neg Fx)]\, \grave{\varepsilon}G\varepsilon$$

Then, by the principle of λ-conversion, it follows that:

$$\exists F(\grave{\varepsilon}G\varepsilon = \grave{\alpha}F\alpha \ \&\ \neg F\grave{\varepsilon}G\varepsilon)$$

Let H be an arbitrary such concept. So we know the following about H:

$$\grave{\varepsilon}G\varepsilon = \grave{\alpha}H\alpha \ \&\ \neg H\grave{\varepsilon}G\varepsilon$$

Now given Law V, it follows from the first conjunct that $\forall x(Gx \equiv Hx)$. So since $\grave{\varepsilon}G\varepsilon$ fails to fall under H, it fails to fall under G, contrary to hypothesis.

So then suppose $\grave{\varepsilon}G\varepsilon$ fails to fall under G. Then, again by λ-conversion, we know:

$$\neg\exists F(\grave{\varepsilon}G\varepsilon = \grave{\alpha}F\alpha \ \&\ \neg F\grave{\varepsilon}G\varepsilon),$$

i.e.,

$$\forall F(\grave{\varepsilon}G\varepsilon = \grave{\alpha}F\alpha \rightarrow F\grave{\varepsilon}G\varepsilon)$$

But by instantiating this universal claim to G, it follows from the self-identity of $\grave{\varepsilon}G\varepsilon$ that $\grave{\varepsilon}G\varepsilon$ does fall under the concept G, contrary to hypothesis. Contradiction.

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Stanford Encyclopedia of Philosophy
Notes to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Notes

1. These models of second-order logic with a Comprehension Principle for Concepts are called ‘general models’ (as opposed to ‘standard’ models in which the domain of concepts is taken to be the power set of domain of objects). These general models exploit the fact that there are only a denumerably infinite number of conditions on objects expressible in the language and hence, only a denumerably infinite number of instances of comprehension. These general models include in the domain of concepts only enough concepts to make these instances of comprehension true. Thus, only a denumerably infinite number of concepts are required, even if the domain of objects is denumerably infinite. So we emphasize that it is the interaction of the Comprehension Principle for Concepts with Vb that engenders the paradox.

2. It is important to note here that Frege's definitions of the membership relation and the notion of equinumerosity require a second-order language, since both definitions involve quantification over concepts.

3. Frege doesn't call this principle ‘Hume's Principle’ in his own writings. The label was instead introduced in Boolos (1987). Frege did cite Hume when he introduced this principle in **Gl**. In **Gl**, §63, he quotes Hume's *Treatise* (I, iii, 1):

> When two numbers are so combined as that one has always an unite answering to every unite of the other, we pronounce them equal.

The idea in Hume does bear some resemblance to the principle Frege constructs, and so we shall continue to use Boolos' label for this principle.

4. We call this an implicit or contextual definition rather than an explicit definition because the notation $\#F$ can only be eliminated when it appears in a context of the form ‘$\#F = \#G$’. By contrast, an explicit definition would take the form:

$$\#F \;\; =_{df} \;\; \text{the object } x \text{ such that } \wp(x,F),$$

where $\wp(x,F)$ is some condition on x and F. This would allow us to eliminate the $\#F$ no matter in which context it appears. We shall examine Frege's attempt to give such a definition momentarily.

The reader might also find the following observation by Demopoulos useful:

> Moreover, Frege's contextual definition (i.e., Hume's Principle) is not ‘conservative’ over the language L = {0, S, N} of second order arithmetic. (It is not conservative because it allows one to prove statements that are otherwise unprovable using this language and second-order logic alone. A proper, explicit definition only introduces simplifying notation --- the new theorems formulable with the new notation introduced by an explicit definition would still have been provable had the new notation been eliminated in terms of primitive notation. As such, explicit definitions are conservative.) Indeed, the contextual definition allows for the proof *both* of the infinity of the sequence of natural numbers *and* of the existence of an infinite cardinal (which Frege called ‘endlos’ in **Gl**).

[from Demopoulos (forthcoming-a)]

5. The reader might find the following observation by Demopoulos useful.

> The characterization "Frege-Russell", nearly universal and certainly well-established, actually slurs over the fact that, from the point of view of Russell's Simple Theory of Types, the number associated with a set (the analogue, in this setting, of a concept of first level) is an entity of higher type than the set itself. Beginning with individuals -- entities of lowest type -- Russell proceeds first to sets of individuals (corresponding to Frege's first-level concepts) and thence to classes of such sets (corresponding to Frege's concepts of second level). For Russell, since numbers are classes of equinumerous sets, they are of higher type than sets. But for Frege, extensions, and therefore numbers, belong to the totality of objects *whatever the level of concept with which they are associated*. Thus, while Russell and Frege both subscribe to *some* version of Hume's Principle, their conceptions of the logical form of the cardinality operator, and therefore, that of the principle itself, are quite different: the operator is "type-raising" for Russell, since it takes us from a set to a class; while for Frege it is "type-lowering", since it takes a concept (set) to an object (individual). This difference is fundamental, since it enables Frege to establish -- on the basis of Hume's principle -- those of the Peano-Dedekind axioms of arithmetic which assert that the system of natural numbers is, in Dedekind's phrase, "simply infinite" (Dedekind infinite). By contrast, when the cardinality operator is type raising, Hume's principle is rather weak, allowing for models of every finite power. (See Bell (1996) and Boolos (1994) for further discussion of these matters.)

[from Demopoulos (forthcoming-a)]

6. The higher-order version of the Law of Extensions asserts that a concept *G* is a member of the extension of the second-order concept *concept equinumerous to F* iff *G* is equinumerous to *F*. If we temporarily suppose that the variable ε can be used a variable ranging over concepts, then we could represent the extension of the second-order concept just described as:

$\grave{\varepsilon}(\varepsilon \approx F)$

Then, the higher-order law of extensions would be formalizable as follows:

$$G \in \grave{\varepsilon}(\varepsilon \approx F) \;\equiv\; G \approx F$$

This principle is used implicitly on several occasions in the derivation of Hume's Principle in **Gl**. Those readers who read the material on the derivation of Hume's Principle in **Gg** will see that this principle gets reformulated as the Lemma to the Proof of Hume's Principle.

7. Strictly speaking, we should represent this concept as follows:

$$[\lambda z\, [\lambda y\, Ayp]z \;\&\; z \neq r]$$

But we have applied the following instance of λ-Conversion to the first conjunct within the matrix of the λ-expression:

$$[\lambda y\, Ayp]z \equiv Azp$$

We thereby simplify the entire expression to:

$$[\lambda z\, Azp \;\&\; z \neq r]$$

8. The Facts numbered 3, 4, 5, and 6 correspond to Theorems 124, 129, 123, and 128, respectively, in **Gg I**. Facts 1, 6, and 7 correspond to Propositions 91, 84, and 98, respectively, in Part III of **Begr**.

9. Facts 2, 3, 5, and 6 correspond to Theorems 134, 132, 141, and 144, respectively, in **Gg I**.

10. A relation *R* is one-to-one ("*R* is 1-1") just in case it satisfies the following condition:

$$Rxz \;\&\; Ryz \rightarrow x=y$$

So Fact 7 in the text is a fact about the weak ancestral whenever the relation *R* in question is 1-1. We shall prove that the Predecessor relation is 1-1 in the third subsection of Section 5. Then Fact 7 and the fact that Predecessor is 1-1 will both play a crucial role in the proof that every number has a successor.

To prove Fact 7, assume that $R^*(a,b)$, *Rcb* and that *R* is 1-1. We want to show $R^+(a,c)$. Now by Fact 5 concerning the weak ancestral, we know that it follows from $R^*(a,b)$ that $\exists z[Rzb \;\&\; R^+(a,z)]$. So call an arbitrary such object "*d*". So we know $Rdb \;\&\; R^+(a,d)$. Now since R is 1-1, it follows from *Rdb* and *Rcb*,

that $c=d$. So, $R^+(a,c)$, which is what we had to show.

11. See the work by Wright cited in the Bibliography for a defense of something like this position. Wright justifies this position on Fregean grounds by appealing to Frege's Context Principle, which asserts that a word has no meaning (reference) except in the context of a proposition (truth).

12. See the paper by Rosen in the Bibliography for a full discussion of how someone might claim that the right-hand condition of an instance might imply its corresponding left-hand condition.

13. Again, see the work by Wright cited in the Bibliography.

14. In the long footnote to §10, Frege seems to suggest that the idea of replacing the truth values with their unit classes cannot be extended to the case of every object in the domain without conflicting with his earlier stipulations (in **Gg I**, §§3, 9 and 20), and in particular, with Basic Law V.

15. Wehmeier (1999) also shows that Frege could not have had much luck restricting the quantifiers of **Gg** to extensions. Wehmeier considers two consistent subsytems that Frege might have adopted to avoid the contradiction, namely, the system H described in Heck (1996) and the system Wehmeier himself describes and labels T△. Both of these systems retain Basic Law V but place restrictions on the Comprehension Principle for Concepts. However, Wehmeier shows that both systems imply the existence of objects which are not extensions (or courses-of-values), and indeed, they imply an infinite number of such objects.

Edward N. Zalta
zalta@stanford.edu

First published: May 7, 2001
Content last modified: May 7, 2001

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Frege's Derivation of Hume's Principle in Gl

The Derivation of Hume's Principle in **Gl**, #73, appeals to the following principle, which is a higher-order version of the Law of Extensions:

> *Principle*: G is an member of the extension of the second-order concept *equinumerous to F* iff G is equinumerous to F

Now Hume's Principle is that the number of Fs is identical to the number of Gs iff F and G are equinumerous. We prove the biconditional in stages.

($\rightarrow$) Assume that the number of Ps is identical to the number of Qs. Then, by the definitions of 'the number of Ps' and 'the number of Qs', we know that the extension of the concept *equinumerous to P* is identical with the extension of the concept *equinumerous to Q*. But it is a fact about equinumerosity that P is equinumerous to P. So by the above Principle, the extension of the concept *equinumerous to P* has P as a member. So, by substitution of identicals, the extension of the concept *equinumerous to Q* has P as an member. So P is equinumerous to Q, by the above Principle.

($\leftarrow$) Assume P is equinumerous with Q. We want to show that the number of Ps is identical to the number of Qs. So, by definition, we have to show that the extension of the concept *equinumerous to P* is identical to the extension of the concept *equinumerous to Q*. By the Principle of Extensionality, then, we have to show that these two extensions have the same members! So we pick an arbitrary concept S and show that S is a member of the extension of the concept *equinumerous to P* iff S is a member of the extension of the concept *equinumerous to Q*.

> ($\rightarrow$) Assume S is a member of the extension of the concept *equinumerous to P* (to show: S is a member of the extension of the concept *equinumerous to Q*). Then, by the above Principle, S is equinumerous to P. So by the transitivity of equinumerosity (this is Fact 4 in the subsection on Equinumerosity in the main portion of the entry), S is equinumerous to Q. So, by the above Principle, S is in the extension of the concept *equinumerous to Q*
>
> ($\leftarrow$) Assume S is a member of the extension of the concept *equinumerous to Q* (to show: S is a member of the extension of the concept *equinumerous to P*). Then S is equinumerous to Q, by the above Principle. By the symmetry of equinumerosity (Fact 3 in the subsection on Equinumerosity), it follows that Q is equinumerous to S. So, given our hypothesis that P is

symmetry of equinumerosity (Fact 3 in the subsection on Equinumerosity), it follows that Q is equinumerous to S. So, given our hypothesis that P is equinumerous to Q, it follows by the transitivity of equinumerosity, that P is equinumerous to S. So, again by symmetry, we have: S is equinumerous to P. And, by the above Principle, it follows S is in the extension of the concept *equinumerous to* P.

⋈

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Frege's ‘Derivation’ of Hume's Principle in Gg

In **Gg**, Fregean extensions do *not* contain concepts as members but rather objects. So Frege had to find another way to express the explicit definition of #*F*. His technique was to let extensions go proxy for their corresponding concepts. We may describe Frege's technique as follows. (What follows is an adaptation and simplification of the strategy Frege outlines in **Gg** I, §34ff.) Instead of defining the number of *F*s as the extension consisting of all those first-order concepts that are equinumerous to *F*, he defined it as the extension consisting of all the extensions of concepts equinumerous to *F*. Here is a formula which says: *x is an extension of a concept that is equinumerous to F*:

$$\exists H(x = \acute{\alpha} H\alpha \quad \& \quad H \approx F)$$

We can name this concept using our λ-notation as follows:

$$[\lambda x\, \exists H(x = \acute{\alpha} H\alpha \quad \& \quad H \approx F)]$$

Instead of writing out this lengthy expression *being an x which is an extension of a concept equinumerous to F*, let us abbreviate our λ-notation for this concept as ‘$F^{\#}$’. Note that the extension of this concept, $\acute{\varepsilon} F^{\#}(\varepsilon)$, contains only extensions as members.

Now Frege's explicit definition of ‘the number of *F*s’ can be given as follows:

$$\#F \quad =_{df} \quad \acute{\varepsilon} F^{\#}(\varepsilon)$$

This definition identifies the number of *F*s as the extension that contains all and only those extensions of concepts that are equinumerous to *F*.

We can complete our preliminary work for the proof of Hume's Principle by formulating and proving the following Lemma (derived from Basic Law V), which simplifies the proof of Hume's Principle:

Lemma for Hume's Principle:
$\acute{\varepsilon} G\varepsilon \in \#F \quad \equiv \quad G \approx F$

(Proof of the Lemma for Hume's Principle)

This Lemma tells us that an extension such as ἐGε will be a member of #*F* just in case *G* is equinumerous to *F*. Clearly, since *F* is equinumerous to itself, it follows that #*F* contains ἐFα as a member. From these facts, one can get a sense of how Frege derived Hume's Principle Basic Law V in **Gg**. Here is a reconstruction of the argument.

Proof of Hume's Principle from Basic Law V

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Proof of the Lemma for Hume's Principle

Let P, Q be arbitrarily chosen concepts. We want to show:

$$\grave{\varepsilon}Q\varepsilon \in \#P \equiv Q \approx P$$

So, by definition of $\#F$, we have to show:

$$\grave{\varepsilon}Q\varepsilon \in \grave{\alpha}P^{\#}\alpha \equiv Q \approx P$$

We prove this by appealing to the Law of Extensions, which yields the following Fact:

$$Fact\text{: } \grave{\varepsilon}Q\varepsilon \in \grave{\alpha}P^{\#}\alpha \equiv P^{\#}(\grave{\varepsilon}Q\varepsilon)$$

($\rightarrow$) Assume $\grave{\varepsilon}Q\varepsilon \in \grave{\alpha}P^{\#}\alpha$ (to show: $Q \approx P$). Then, by the above Fact, we know $P^{\#}(\grave{\varepsilon}Q\varepsilon)$, i.e.,

$$[\lambda x\, \exists H(x = \grave{\alpha}H\alpha \;\&\; H \approx P)](\grave{\varepsilon}Q\varepsilon)$$

By λ-conversion, this implies:

$$\exists H(\grave{\varepsilon}Q\varepsilon = \grave{\alpha}H\alpha \;\&\; H \approx P)$$

Let R be such a concept:

$$\grave{\varepsilon}Q\varepsilon = \grave{\alpha}R\alpha \;\&\; R \approx P \tag{1}$$

But, by Basic Law V, the first conjunct implies $\forall x(Qx \equiv Rx)$. Since the material equivalence of two concepts implies their equinumerosity (this was noted as Fact 1 in the subsection on Equinumerosity in the main part of the entry), it follows that $Q \approx R$. So from this result and the second conjunct of (1), it follows that $Q \approx P$, by the transitivity of equinumerosity (Fact 4 in the subsection on Equinumerosity).

($\rightarrow$) Assume $Q \approx P$ (to show: $\grave{\varepsilon}Q\varepsilon \in \grave{\alpha}P^{\#}\alpha$). Then, by identity introduction, we know: $\grave{\varepsilon}Q\varepsilon = \grave{\varepsilon}Q\varepsilon \;\&\; Q \approx P$. So, by existential generalization:

$$\exists H(\grave{\varepsilon}Q\varepsilon = \grave{\alpha}H\alpha \;\&\; H \approx P)$$

$$\exists H(\grave{\varepsilon}Q\varepsilon = \grave{\alpha}H\alpha \;\&\; H \approx P)$$

And by λ-Conversion:

$$[\lambda x\, \exists H(x = \grave{\alpha}H\alpha \;\&\; H \approx P)](\grave{\varepsilon}Q\varepsilon)$$

So, by the Law of Extensions, $\grave{\varepsilon}Q\varepsilon \in \grave{\alpha}P\#\alpha$.

Return to Frege's ‘Derivation’ of Hume's Principle in **Gg**

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Proof of Hume's Principle from Basic Law V
Grundgesetze-Style

Let P, Q be arbitrarily chosen concepts. We want to show:

$$\#P = \#Q \equiv P \approx Q$$

($\rightarrow$) Assume $\#P = \#Q$ (to show: $P \approx Q$). Note that since $P \approx P$ (this is Fact 2 in the subsection on Equinumerosity), we know by the previous Lemma that $\grave{\varepsilon}P\varepsilon \in \#P$. But then, by identity substitution, $\grave{\varepsilon}P\varepsilon \in \#Q$. So, by our previous Lemma, $P \approx Q$.

($\leftarrow$) Assume $P \approx Q$ (to show: $\#P = \#Q$). By definition of $\#$, we have to show $\grave{\varepsilon}P^{\#}\varepsilon = \grave{\alpha}Q^{\#}\alpha$. So, by Basic Law V, we have to show $\forall x(P^{\#}x \equiv Q^{\#}x)$. Since pick an arbitrary object b (to show: $P^{\#}b \equiv Q^{\#}b$).

> ($\rightarrow$) Assume $P^{\#}b$. Then, by definition of $P^{\#}$ and λ-Conversion, $\exists H(b = \grave{\alpha}H\alpha \ \& \ H \approx P)$. Let R be an arbitrary such concept; so $b = \grave{\alpha}R\alpha \ \& \ R \approx P$. From the second conjunct and our initial hypothesis, it follows (by the transitivity of equinumerosity) that $R \approx Q$. So, reassembling what we know, it follows that $b = \grave{\alpha}R\alpha \ \& \ R \approx Q$. By existential generalization, it follows that $\exists H(b = \grave{\alpha}H\alpha \ \& \ H \approx Q)$. So by λ-Conversion,
>
> $$[\lambda x \ \exists H(b = \grave{\alpha}H\alpha \ \& \ H \approx Q)]b$$
>
> It follows from this, by definition, that $Q^{\#}b$.
>
> ($\leftarrow$) (Exercise)

Return to Frege's 'Derivation' of Hume's Principle in **Gg**

zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Proof of Fact 5 Concerning the Weak Ancestral

Fact 5 concerning the weak ancestral R^+ of Rasserts:

> **Fact 5 (R^+)**:
> $R^*(x,y) \rightarrow \exists z[Rzy \mathbin{\&} R^+(x,z)]$

To prove this, we shall appeal to Fact 5 concerning the ancestral R^* of R:

> **Fact 5 (R^*)**:
> $[R^*(x,y) \mathbin{\&} \forall u(Rxu \rightarrow Fu) \mathbin{\&} \text{Her}(F,R)] \rightarrow Fy$,

for any concept F and objects x and y:

Now to prove Fact 5 (R^+), assume $R^*(a,b)$. We want to show:

> $\exists z[Rzb \mathbin{\&} R^+(a,z)]$

Notice that by λ-Conversion, it suffices to show:

> $[\lambda w\ \ \exists z[Rzw \mathbin{\&} R^+(a,z)]b$

Let us use 'P' to denote this concept under which (we have to show) b falls. Notice that we could prove Pb by instantiating Fact 5 (R^*) to P, a, and b and establishing the antecedent of the result. In other words, by Fact (R^*), we know:

> $[R^*(a,b) \mathbin{\&} \forall u(Rau \rightarrow Pu) \mathbin{\&} \text{Her}(P,R)] \rightarrow Pb$

So if we can show the conjuncts of the antecedent, we are done. The first conjunct is already established, by hypothesis. So we have to show:

> (1) $\forall u(Rau \rightarrow Pu)$
> (2) $\text{Her}(P,R)$

To see what we have to show for (1), we expand our defined notation and simplify by using λ-Conversion. Thus, we have to show:

(1) $\forall u[Rau \rightarrow \exists z(Rzu \,\&\, R^{+}(a,z))]$

So assume *Rau*, to show the consequent of (1). But it is an immediate consequence of the definition of the weak ancestral R^{+} that R^{+} is reflexive. (This is Fact 8 concerning the weak ancestral, in Section 4, "The Weak Ancestral of *R*".) So we may conjoin and conclude $Rau \,\&\, R^{+}(a,a)$. From this, we may infer consequent of (1), by existential generalization, which is what we had to show.

To show (2), we have to show that *P* is hereditary on *R*. If we expand our defined notation and simplify by using λ-Conversion), then we have to show:

(2) $Rxy \rightarrow [\exists z(Rzx \,\&\, R^{+}(a,z)) \rightarrow \exists z(Rzy \,\&\, R^{+}(a,z))]$

So assume

(A) $Rxy \,\&\, \exists z(Rzx \,\&\, R^{+}(a,z))$

to show: $\exists z(Rzy \,\&\, R^{+}(a,z))$. From the second conjunct of (A), we know that there is some object, say *d*, such that:

$Rdx \,\&\, R^{+}(a,d)$; i.e.,
$R^{+}(a,d) \,\&\, Rdx$

So, by Fact 2 about the weak ancestral (Section 4, "The Weak Ancestral of *R*"), it follows that $R^{*}(a,x)$, from which it immediately follows that $R^{+}(a,x)$, by definition of R^{+}. So, by appealing to the first conjunct of (A), we have:

$Rxy \,\&\, R^{+}(a,x)$,

from which it follows that:

$\exists z(Rzy \,\&\, R^{+}(a,z))$,

which is what we had to show.

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

[Edward N. Zalta](http://mally.stanford.edu/zalta.html)
zalta@stanford.edu

First published: November 22, 1999
Content last modified: November 22, 1999

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Proof of Lemma Concerning Zero

Let P be an arbitrarily chosen concept. We want to show $\#P = 0 \equiv \neg\exists x Px$.

($\rightarrow$) Assume $\#P = 0$. Then, by definition of 0, $\#P = \#[\lambda z\, z \neq z]$. So by Hume's Principle, P is equinumerous to $[\lambda z\, z \neq z]$. So, by the definition of equinumerosity, there is an R that maps every object falling under P to a unique object falling under $[\lambda z\, z \neq z]$ and vice versa. Suppose, for reductio, that $\exists x Px$, say Pa. Then there is an object, say b, such that Rab and $[\lambda z\, z \neq z]b$. But, then, by λ-Conversion, b is not self-identical, which contradicts the laws of identity.

($\leftarrow$) Suppose $\neg\exists x Px$. Now as we have seen, the laws of identity guarantee that no object falls under the concept $[\lambda z\, z \neq z]$. But then any relation you please bears witness to the fact that P is equinumerous with $[\lambda z\, z \neq z]$. For let R be some arbitrary relation. Then (a) every object falling under P bears R to a unique object falling under $[\lambda z z \neq z]$ (since there are no objects falling under P), and (b) every object falling under $[\lambda z\, z \neq z]$ is such that there is a unique object falling under P that bears R to it (since there are no objects exemplifying $[\lambda z\, z \neq z]$). Since P is therefore equinumerous with $[\lambda z\, z \neq z]$, it follows by Hume's Principle, that $\#[\lambda z z \neq z] = \#P$. But, then, by definition, $0 = \#P$.

$\bowtie$

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Proof of Equinumerosity Lemma

In this proof of the Equinumerosity Lemma, we utilize the following abbreviation:

$$x = \imath y\varphi(y) =_{abbr} \exists y[\varphi(y) \ \& \ \forall z(\varphi(z/y) \rightarrow z = y) \ \& \ x = y]$$

We may read this as follows:

> x is identical to *the* object y which satisfies the condition φ $=_{df}$ there exists a y such that: (a) y satisfies the condition φ, (b) everything which satisfies the condition φ is identical to y, and (c) x is identical to y

It will be seen how this abbreviation is employed to simplify the definition of new relations. Given this new notation, it is straightforward to show:

> Principle of Descriptions: $x = \imath y\varphi(y) \rightarrow \varphi(x/y)$

In other words, if x is *the* object y that satisfies the condition $\varphi(y)$, then x satisfies the condition φ. (The proof is a simple exercise.) The appeal to this principle will be obvious in what follows.

Proof of Equinumerosity Lemma. Assume that $P \approx Q$, Pa, and Qb. So there is a relation, say R, such that (a) R maps every object falling under P to a unique object falling under Q and (b) for every object falling under Q there is a unique object falling under P which is R-related to it. Now we use P^{-a} to designate $[\lambda z\, Pz \,\&\, z \neq a]$, and we use Q^{-b} to designate $[\lambda z\, Qz \,\&\, z \neq b]$. We want to show that $P^{-a} \approx Q^{-b}$. By the definition of equinumerosity, we have to show that there is a relation R' which is a one-to-one function from the objects falling under P^{-a} onto the objects falling under Q^{-b}. We prove this by cases.

Case 1: Suppose Rab. Then we choose R' to be R itself. Clearly, R is then a one-to-one function from the objects of P^{-a} to the objects of Q^{-b}. But the proof can be given as follows. We show: (A) that R is a function from the objects of P^{-a} to the objects of Q^{-b}, and then (B) that R is a one-to-one function from the objects of P^{-a} onto the objects of Q^{-b}.

(A) Pick an arbitrary object, say c, such that $P^{-a}c$. We want to show that there is a unique object which falls under Q^{-b} and to which c bears R. Since $P^{-a}c$, we know that $Pc \ \& \ c \neq a$, by the definition of P^{-a}. But if Pc, then by our hypothesis that R is a witness to the equinumerosity of P and

then by our hypothesis that R is a witness to the equinumerosity of P and

Proof (cont'd)

Return to Main Text

Q, it follows that there is a unique object, say d, such that Qd and Rcd. But we are considering the case in which Rab and so from the established facts that Rcd and $c \neq a$, it follows by the one-to-one character of R that $b \neq d$. So we have that Qd and $d \neq b$, which establishes that $Q^{-b}d$. And we have also established that Rcd. So it remains to show that every other object that falls under Q^{-b} to which c bears R just is identical to d. So suppose $Q^{-b}e$ and Rce. Then by definition of Q^{-b}, it follows that Qe. But now $e = d$, for d is the unique object falling under Q to which c bears R. So there is a unique object which falls under Q^{-b} and to which c bears R.

(B) Pick an arbitrary object, say d, such that $Q^{-b}d$. We want to show that there is a unique object falling under P^{-a} that bears R to d. Since $Q^{-b}d$, we know Qd and $d \neq b$. From Qd and the fact that R witnesses the equinumerosity of P and Q, we know that there is a unique object, say c, that falls under P and which bears R to d. Since we are considering the case in which Rab, and we've established Rcd and $d \neq b$, it follows that $a \neq c$, by the functionality of R. Since we now have Pc and $c \neq a$, we have established that c falls under P^{-a}, and moreover, that Rcd. So it remains to prove that any other object that falls under P^{-a} and which bears R to d just is (identical to) c. But if f, say, falls under P^{-a} and bears R to d, then Pf, by definition of P^{-a}. But recall that c is the unique object falling under P which bears R to d. So $f = c$.

Case 2: Suppose $\neg Rab$. Then we choose R' to be the relation:

$$[\lambda xy\,(x \neq a \;\&\; y \neq b \;\&\; Rxy) \vee (x = \imath u(Pu \;\&\; Rub) \;\&\; y = \imath u(Qu \;\&\; Rau))]$$

To see that there is such a relation, note that once we replace the abbreviations $x = \imath u(Pu \;\&\; Rub)$ and $y = \imath u(Qu \;\&\; Rau)$ by primitive notation, the matrix of the λ-expression is a formula of the form $\varphi(x,y)$ which can be used in an instance of the Comprehension Principle for Relations.

Now we want to show that R' is a one-to-one function from the objects of P^{-a} onto the objects of Q^{-b}. We show (A) that R' is a function from the objects of P^{-a} to the objects of Q^{-b}, and then (B) that R' is a one-to-one function from the objects of P^{-a} onto the objects of Q^{-b}.

(A) To show that R' is a function from the objects of P^{-a} to the objects of Q^{-b}, pick an arbitrary object, say c, such that $P^{-a}c$. Then by definition of P^{-a}, we know that Pc and $c \neq a$. We need to find an object, say d for which the following three things hold: (i) $Q^{-b}d$, (ii) $R'cd$, and (iii) $\forall w(Q^{-b}w \;\&\; R'cw \rightarrow w = d)$. We find such a d in each of the following, mutually exclusive cases:

cases:

Proof (cont'd)

Return to Main Text

Case 1: Rcb. So, since we know that each object falling under Q is such that there is a unique object falling under P that is R-related to it, we know that $c = \imath u(Pu \,\&\, Rub)$. Then, since we know R maps a to a unique object falling under Q, we let d be that object. That is, d satisfies the defined condition $d = \imath u(Qu \,\&\, Rau)$. So Qd, Rad, and $\forall w(Qw \,\&\, Raw \rightarrow w = d)$. We now show that (i), (ii) and (iii) hold for d:

i) Since we know Qd, all we have to do to establish $Q^{-b}d$ is to show $d \neq b$. But we know Rad and we are considering the case where $\neg Rab$. So, by the laws of identity, $d \neq b$.

ii) To show $R'cd$, we need to establish:

$$(c \neq a \,\&\, d \neq b \,\&\, Rcd) \vee (c = \imath u(Pu \,\&\, Rub) \,\&\, d = \imath u(Qu \,\&\, Rau))$$

But the conjunctions of the right disjunct are true (by assumption and by choice, respectively). So $R'cd$.

iii) Suppose $Q^{-b}e$ (i.e., Qe and $e \neq b$) and $R'ce$. We want to show: $e = d$. Since $R'ce$, then:

$$(c \neq a \,\&\, e \neq b \,\&\, Rce) \vee (c = \imath u(Pu \,\&\, Rub) \,\&\, e = \imath u(Qu \,\&\, Rau))$$

But the left disjunct is impossible (we're considering the case where Rcb, yet the left disjunct asserts Rce and $e \neq b$, which together contradict the functionality of R). So the right disjunct must be true, in which case it follows from the fact that $e = \imath u(Qu \,\&\, Rau)$ that $e = d$, by the definition of d.

Case 2: $\neg Rcb$. We are under the assumption $P^{-a}c$ (i.e., Pc and $c \neq a$), and so we know by the definition of R and the fact that Pc that there is a unique object which falls under Q and to which c bears R. Choose d to be this object. So Qd, Rcd, and $\forall w(Qw \,\&\, Rcw \rightarrow w = d)$. We can now show that (i), (ii) and (iii) hold for d:

i) Since we know Qd, all we have to do to establish that $Q^{-b}d$ is to show $d \neq b$. We know that Rcd and we are considering the case where $\neg Rcb$. So it follows that $d \neq b$, by the laws of identity. So $Q^{-b}d$.

ii) To show $R'cd$, we need to establish:

$$(c \neq a \,\&\, d \neq b \,\&\, Rcd) \vee (c = \imath u(Pu \,\&\, Rub) \,\&\, d = \imath u(Qu \,\&\, Rau))$$

$$(c \neq a \ \& \ d \neq b \ \& \ Rcd) \vee (c = \iota u(Pu \ \& \ Rub) \ \& \ d = \iota u(Qu \ \& \ Rau))$$

[Proof (cont'd)]

[Return to Main Text]

But the conjuncts of the left disjunct are true, for $c \neq a$ (by assumption), $d \neq b$ (we just proved this), and Rcd (by the definition of d). So $R'cd$.

iii) Suppose $Q^{-b}e$ (i.e., Qe and $e \neq b$) and $R'ce$. We want to show: $e = d$. Since $R'ce$, then:

$$(c \neq a \;\&\; e \neq b \;\&\; Rce) \vee (c = \imath u(Pu \;\&\; Rub) \;\&\; e = \imath u(Qu \;\&\; Rau))$$

But the right disjunct is impossible (we're considering the case where $\neg Rcb$, yet the right disjunct asserts $c = \imath u(Pu \& Rub)$, which implies Rcb, a contradiction). So $c \neq a \;\& e \neq b \;\& Rce$. Since we now know that Qe and Rce, we know that $e = d$, since d is, by definition, the unique object falling under Q to which c bears R.

(B) To show that R' is a one-to-one function from the objects of P^{-a} onto the objects of Q^{-b}, pick an arbitrary object, say d, such that $Q^{-b}d$. Then by definition of Q^{-b}, we know that Qd and $d \neq b$. We need to find an object, say c, for which the following three things hold: (i) $P^{-a}c$, (ii) $R'cd$, and (iii) $\forall w(P^{-a}w \;\& R'wd \rightarrow w = c)$. We find such a c in each of the following, mutually exclusive cases:

Case 1: Rad. So $d = \imath u(Qu \;\&\; Rau)$. Then choose $c = \imath u(Pu \;\&\; Rub)$ (we know there is such an object). So Pc, Rcb, and $\forall w(Pw \& Rwb \rightarrow w = c)$. We now show that (i), (ii) and (iii) hold for c:

i) Since we know Pc, all we have to do to establish $P^{-a}c$ is to show $c \neq a$. But we know Rcb, and we are considering the case where $\neg Rab$. So, by the laws of identity, it follows that $c \neq a$.

ii) To show $R'cd$, we need to establish:

$$(c \neq a \;\&\; d \neq b \;\&\; Rcd) \vee (c = \imath u(Pu \;\&\; Rub) \;\&\; d = \imath u(Qu \;\&\; Rau))$$

But the conjuncts of the right disjunct are true (by choice and by assumption, respectively). So $R'cd$.

iii) Suppose $P^{-a}f$ (i.e., Pf and $f \neq a$) and $R'fd$. We want to show: $f = c$. Since $R'fd$, then:

$$(f \neq a \;\&\; d \neq b \;\&\; Rfd) \vee (f = \imath u(Pu \;\&\; Rub) \;\&\; d = \imath u(Qu \;\&\; Rau))$$

But the left disjunct is impossible (we're considering the case where Rad, yet the left disjunct asserts Rfd and $f \neq a$, which

where Rad, yet the left disjunct asserts Rfd and $f \neq a$, which

Proof (cont'd)

Return to Main Text

together contradict the fact that R is one-to-one). So the right disjunct must be true, in which case it follows from the fact that $f = \iota u(Pu \,\&\, Rub)$ that $f = c$, by the definition of c.

Case 2: $\neg Rad$. We are under the assumption $Q^{-b}d$ (i.e., Qd and $d \neq b$), and so we know by the definition of R and the fact that Qd that there is a unique object which falls under P and which bears R to d. Choose c to be this object. So Pc, Rcd, and $\forall w(Pw \,\&\, Rwd \rightarrow w = c)$. We can now show that (i), (ii), and (iii) hold for c:

i) Since we know Pc, all we have to do to establish that $P^{-a}c$ is to show $c \neq a$. But we know that Rcd, and we are considering the case in which $\neg Rad$. So it follows that $c \neq a$, by the laws of identity. So $P^{-a}c$.

ii) To show $R'cd$, we need to establish:

$$(c \neq a \,\&\, d \neq b \,\&\, Rcd) \vee (c = \iota u(Pu \,\&\, Rub) \,\&\, d = \iota u(Qu \,\&\, Rau))$$

But the conjuncts of the left disjunct are true, for $c \neq a$ (we just proved this), $d \neq b$ (by assumption), and Rcd (by the definition of c). So $R'cd$.

iii) Suppose $P^{-a}f$ (i.e., Pf and $f \neq a$) and $R'fd$. We want to show: $f = c$. Since $R'fd$, then:

$$(f \neq a \,\&\, d \neq b \,\&\, Rfd) \vee (f = \iota u(Pu \,\&\, Rub) \,\&\, d = \iota u(Qu \,\&\, Rau))$$

But the right disjunct is impossible (we're considering the case where $\neg Rad$, yet the right disjunct asserts $d = \iota u(Qu \,\&\, Rau)$, which implies Rad, a contradiction). So $f \neq a \,\&\, d \neq b \,\&\, Rfd$. Since we now know that Pf and Rfd, we know that $f = c$, since c is, by definition, the unique object falling under P which bears R to d. $\bowtie$

Return to Main Text

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Proof of the General Principle of Induction

Assume the antecedent of the principle, eliminating the defined notation for $HerOn(F, {}^{a}R^{+})$:

$$Pa \;\&\; \forall x,y(R^{+}(a,x) \;\&\; R^{+}(a,y) \;\&\; Rxy \rightarrow (Px \rightarrow Py)).$$

We want to show, for an arbitrary object b, that if $R^{+}(a,b)$ then Pb. So assume $R^{+}(a,b)$. To show Pb, we appeal to Fact (6) about R^{+} (in our subsection on the Weak Ancestral in §4):

$$Fx \;\&\; R^{+}(x,y) \;\&\; Hereditary(F,R) \rightarrow Fy$$

Instantiate the variable F in this Fact to the property $[\lambda z\, R^{+}(a,z) \;\&\; Pz]$ (that there is such a property is guaranteed by the Comprehension Principle for Relations), and instantiate the variables x and y to the objects a and b, respectively. The result (after applying λ-Conversion) is therefore something that we have established as true :

$$R^{+}(a,a) \;\&\; Pa \;\&\; R^{+}(a,b) \;\&\; Hereditary([\lambda z\, R^{+}(a,z) \;\&\; Pz], R) \rightarrow R^{+}(a,b) \;\&\; Pb$$

So if we can establish the antecedent of this fact, we establish Pb. But we know that the first conjunct is true, by the definition of R^{+}. We know the second conjunct is true, by assumption. We know that the third conjunct is true, by further assumption. So if we can establish:

$$Hereditary([\lambda z\, R^{+}(a,z) \;\&\; Pz], R),$$

we are done. But, by the definition of heredity, this just means:

$$\forall x,y[Rxy \rightarrow ((R^{+}(a,x) \;\&\; Px) \rightarrow (R^{+}(a,y) \;\&\; Py)].$$

To prove this claim, we assume Rxy, $R^{+}(a,x)$, and Px (to show: $R^{+}(a,y)$ & Py). But from the facts that $R^{+}(a,x)$ and Rxy, it follows from Fact (2) about R^{+} (in our subsection on the Weak Ancestral) that $R^{*}(a,y)$. This implies $R^{+}(a,y)$, by the definition of R^{+}. But since we now have $R^{+}(a,x)$, $R^{+}(a,y)$, Rxy, and Px, it follows from the first assumption in the proof that

implies $R^+(a, y)$, by the definition of R^+. But since we now have $R^+(a, x)$, $R^+(a, y)$, Rxy, and Px, it follows from the first assumption in the proof that Py. ⋈

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Proof that 0 Falls Under *Q*

The proof that 0 falls under *Q* is relatively straightforward. We want to show:

$[\lambda y\, Precedes(y,\#[\lambda z\, Precedes^{+}(z,y)])]0$

By λ-Conversion, it suffices to show:

$Precedes(0, \#[\lambda z\, Precedes^{+}(z,0)])$

So, by the definition of Predecessor, we have to show that there is a concept *F* and object *x* such that:

(1) Fx
(2) $\#[\lambda z\, Precedes^{+}(z,0)] = \#F$
(3) $0 = \#[\lambda u\, Fu \,\&\, u \neq x]$

We can demonstrate that there is an *F* and *x* for which (1), (2) and (3) hold if we pick *F* to be $[\lambda z\, Precedes^{+}(z,0)]$ and pick *x* to be 0. We now establish (1), (2), and (3) for these choices.

To show that (1) holds, we have to show:

$[\lambda z\, Precedes^{+}(z,0)]0$

But we know, from the definition of $Precedes^{+}$, that $Precedes^{+}(0,0)$, So by abstraction using λ-Conversion, we are done.

To show that (2) holds, we need do no work, since our choice of *F* requires us to show:

$\#[\lambda z\, Precedes^{+}(z,0)] = \#[\lambda z\, Precedes^{+}(z,0)]$,

which we know by the logic of identity.

To show (3) holds, we need to show:

(A) $0 = \#[\lambda u\ Precedes^{+}(u,0)\ \&\ u \neq 0]$

[Note that the λ-expression in (A) has been simplified by applying λ-Conversion to the following (which, strictly speaking, is what results when you substitute our choice for *F* in (3)):

$$[\lambda u\ [\lambda z\ Precedes^{+}(z,0)]u\ \&\ u \neq 0]$$

In what follows, we use the simplified version of this λ-expression.]

To show (A), it suffices to show the following, in virtue of the Lemma Concerning Zero (in our subsection on The Concept *Natural Number* in §4):

$$\neg\exists x\ ([\lambda u\ Precedes^{+}(u,0)\ \&\ u \neq 0]x)$$

And by λ-Conversion, it suffices to show:

(B) $\neg\exists x\ (Precedes^{+}(x,0)\ \&\ x \neq 0)$

We establish (B) as follows.

When we established Theorem 2 (i.e., the fact that 0 is not the successor of any number), we proved that nothing precedes 0:

$$\neg\exists x\ Precedes(x,0)$$

From this, and Fact (3) about R^* (in the subsection on the Ancestral of *R*, in §4), it follows that nothing ancestrally precedes 0:

$$\neg\exists x\ Precedes^{*}(x,0)$$

Now suppose (for *reductio*) the negation of (B); i.e, that there is some object, say *a*, such that $Precedes^{+}(a,0)$ and $a \neq 0$. Then, by definition of $Precedes^{+}$, it follows that either $Precedes^{*}(a,0)$ or $a = 0$. But since our reductio hypthesis includes that $a \neq 0$, it must be that $Precedes^{*}(a,0)$, which contradicts the fact displayed immediately above.

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: June 10, 1998

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Proof that Q is Hereditary on the Natural Numbers

We want to prove the following claim:

> $HerOn(Q,\mathbb{N})$

The proof of this claim appeals to the following Lemma (cf. **Gg** I, Theorem 149):

> **Lemma on *Predecessor*$^+$:**
> $\mathbb{N}x \;\&\; Precedes(y,x) \rightarrow$
> $\#[\lambda z\, Precedes^+(z,y)] = \#[\lambda z\, Precedes^+(z,x) \;\&\; z \neq x]$
>
> (Proof of the Lemma on *Predecessor*$^+$)

Intuitively, the Lemma on Predecessor$^+$ tells us that if y precedes a number x, then $\#[\lambda z\, z \leq y]$ is identical to $\#[\lambda z\, z \leq x \;\&\; z \neq x]$.

Now to show $HerOn(Q,\mathbb{N})$, we have to show:

> $\forall n \forall m[Precedes(n,m) \rightarrow (Qn \rightarrow Qm)]$

If we replace 'Q' with its definition and simplify the result by λ-Conversion, then what we have to show is:

> $\forall n \forall m(Precedes(n,m) \rightarrow$
> $Precedes(n,\#[\lambda z\, Precedes^+(z,n)]) \rightarrow Precedes(m,\#[\lambda z\, Precedes^+(z,m)])$

(Intuitively, we have to show that if n precedes m, then if n precedes the number of numbers less than or equal to n, then m precedes the number of numbers less than or equal to m.) So, letting n and m be arbitrary, we assume both:

> (A) $Precedes(n,m)$
> (B) $Precedes(n,\#[\lambda z\, Precedes^+(z,n)])$

to show:

$Precedes(m,\#[\lambda z\ Precedes^{+}(z,m)])$

By the definition of Predecessor, we have to show that there is a concept *F* and object *x* such that:

(1) Fx
(2) $\#[\lambda z\ Precedes^{+}(z,m)] = \#F$
(3) $m = \#[\lambda u\ Fu \ \&\ u \neq x]$

We can demonstrate that there is an *F* and *x* for which (1), (2) and (3) hold if we pick *F* to be $[\lambda z\ Precedes^{+}(z,m)]$ and pick *x* to be *m*. We now establish (1), (2), and (3) for these choices.

To show that (1) holds, we have to show:

$[\lambda z\ Precedes^{+}(z,m)]m$

But we know, from the definition of $Precedes^{+}$, that $Precedes^{+}(m,m)$. So by abstraction using λ-Conversion, we are done.

To show that (2) holds, we need do no work, since our choice of *F* requires us to show:

$\#[\lambda z\ Precedes^{+}(z,m)] = \#[\lambda z\ Precedes^{+}(z,m)]$,

which we know by the logic of identity.

To show (3) holds, we need to show:

(C) $m = \#[\lambda u\ Precedes^{+}(u,m) \ \&\ u \neq m]$

[Note that the λ-expression in the above has been simplified by applying λ-Conversion to the following (which, strictly speaking, is what results when you substitute our choice for *F* in (3)):

$[\lambda u\ [\lambda z\ Precedes^{+}(z,m)]u \ \&\ u \neq m]$

In what follows, we use the simplified version of this λ-expression.]

Now in virtue of (A), (B) and the functionality of Predecessor (the proof of which was left as an exercise in the subsection No Two Numbers Have the Same Successor in §5), we know $m = \#[\lambda z\ Precedes^{+}(z,n)]$. So, substituting for *m* in (C), we have to show:

$\#[\lambda z\, Precedes^{+}(z,n)] = \#[\lambda u\, Precedes^{+}(u,m) \,\&\, u \neq m]$

But we can demonstrate this by appealing to the Lemma on Predecessor^{+} mentioned at the outset. We may instantiate the variables *x* and *y* in this Lemma to *m* and *n*, respectively, yielding:

$$\mathbb{N}m \,\&\, Precedes(n,m) \rightarrow \\ \#[\lambda z\, Precedes^{+}(z,n)] = \#[\lambda z\, Precedes^{+}(z,m) \,\&\, z \neq m]$$

Since the consequence is what we had to show, we are done, for the conjuncts of the antecedent are things we assumed to be true at the beginning of our conditional proof.

Return to Frege's Logic, Theorem, and Foundations for Arithmetic

Edward N. Zalta
zalta@stanford.edu

First published: June 10, 1998
Content last modified: November 22, 1999

Stanford Encyclopedia of Philosophy
Supplement to Frege's Logic, Theorem, and Foundations for Arithmetic

---

Proof of the Lemma on Predecessor$^+$

We wish to prove the Lemma on Predecessor$^+$, which asserts:

> **Lemma on *Predecessor*$^+$:**
> $\mathbb{N}x \,\&\, Precedes(y,x) \rightarrow$
> $\#[\lambda z\, Precedes^+(z,y)] = \#[\lambda z\, Precedes^+(z,x) \,\&\, z \neq x]$

Now this Lemma can be proved with the help of the following Lemma:

> **Lemma: No Number Ancestrally Precedes Itself**
> $\forall x[\mathbb{N}x \rightarrow \neg Precedes^*(x,x)]$
>
> (Proof) [Exercise for the Reader]

Now to prove the Lemma on Predecessor$^+$, assume that $\mathbb{N}n$ and $Precedes(m,n)$. We want to show:

$$\#[\lambda z\, Precedes^+(z,m)] = \#[\lambda z\, Precedes^+(z,n) \,\&\, z \neq n]$$

By Hume's Principle, it suffices to show:

$$[\lambda z\, Precedes^+(z,m)] \approx [\lambda z\, Precedes^+(z,n) \,\&\, z \neq n]$$

Now it is a fact about equinumerosity (see the Facts About Equinumerosity, in §3, Equinumerosity, Fact 1) that if two concepts are materially equivalent, then they are equinumerous. It therefore suffices to show that

$$\forall x([\lambda z\, Precedes^+(z,m)]x \equiv [\lambda z\, Precedes^+(z,n) \,\&\, z \neq n]x)$$

And by λ-Conversion, it suffices to show:

$$\forall x[Precedes^+(x,m) \equiv Precedes^+(x,n) \,\&\, x \neq n]$$

So let us pick an arbitrary object, say *a*, and show:

$Precedes^{+}(a,m) \equiv Precedes^{+}(a,n)\ \&\ a \neq n$

($\rightarrow$) Assume $Precedes^{+}(a,m)$. Then from our assumption that $Precedes(m,n)$ and a Fact about R^{+} (see §4, Weak Ancestral, Facts About R^{+}, Fact 2), it follows that $Precedes^{*}(a,n)$. *A fortiori*, then, $Precedes^{+}(a,n)$. Now by the Lemma mentioned at the outset of this proof, namely, No Natural Number Ancestrally Precedes Itself, we know that $\neg Precedes^{*}(n,n)$. Since *a* ancestrally precedes *n* and *n* does not, it follows that $a \neq n$. We have therefore proved what we were after, namely:

$Precedes^{+}(a,n)\ \&\ a \neq n$

($\leftarrow$) Assume $Precedes^{+}(a,n)$ and $a \neq n$. By definition, the first conjunct tells us that either $Precedes^{*}(a,n)$ or $a = n$. Since we know by assumption that the latter disjunct is not true, we know $Precedes^{*}(a,n)$. But from this fact, the fact that Predecessor is 1-1, and our assumption that $Precedes(m,n)$, it follows from a fact about R^{+} (see §4, Weak Ancestral, Facts About R^{+}, Fact 7), that $Precedes^{+}(a,m)$, which is what we had to show.

Return to Proof that Q is Hereditary on the Natural Numbers

Edward N. Zalta
zalta@stanford.edu

First published: June 28, 1998
Content last modified: November 22, 1999

Semantic Challenges to Realism

According to realism, the world is as it is independently of how humans take it to be. The objects the world contains, together with their properties and the relations they enter into, fix the world's nature and these objects exist independently of our ability to discover they do. Unless this is so, realists argue, none of our beliefs about our world could be objectively true since true beliefs tell us how things are and beliefs are objective when true or false independently of what anyone might think. The issue of objectivity affects all of us deeply -- when we think the State has an obligation to provide adequate health care to all its citizens we mean to be describing what the State's obligations really are, independently of what anyone might think about the matter. If someone disagrees with us over this matter, we think they've got it wrong -- are mistaken about how things are as regards the State and its obligations. If there can be no objectivity without a mind-independent world, as realists claim, then there had better be a mind-independent world.

Many philosophers believe realism is just plain common sense. Others believe it to be a direct implication of modern science which paints humans as fallible creatures adrift in an inhospitable world not of their making. Nonetheless, realism is controversial. There are epistemological problems connected with it -- how can we obtain knowledge of a mind-independent world? There are also prior semantic problems -- how are the links between our beliefs and the mind-independent states of affairs they allegedly represent set up? This is the Representation Problem.

Anti-realists deny the world is mind-independent. Believing the Representation Problem to be insoluble for realists, they conclude realism must be false. In this article I review a number of anti-realist semantic challenges to realism all based on the Representation Problem: (i) The Manifestation Challenge claims the cognitive and linguistic behaviour of an agent provides no evidence that realist mind/world links exist. (ii) The Language Acquisition Challenge claims that if such links were to exist language learning would be impossible. (iii) The Brain in a Vat Challenge holds that realism entails both that we could be massively deluded ("Brains in a Vat") and that if we were we could not even form the belief that we were. (iv) The Conceptual Relativity Challenge alleges it is senseless to ask what the world contains independently of how we conceive of it. The objects that exist depend on the conceptual scheme used to classify them. (v) The Model-Theoretic Challenge contends realist must either hold that an ideal theory passing every conceivable test could be false or that perfectly determinate terms like ‘cat’ are massively indeterminate. Both alternatives are absurd according to anti-realists.

I proceed by first defining realism, illustrating its distinctive mind-independence claim with some examples and distinguishing it from a doctrine with which it is often confused, factualism. I then outline the Representation Problem, saying why it is a problem for realism before presenting the anti-realist's

semantic challenges to realism which are all based on it. I discuss realist responses to these challenges, indicating how the debates have proceeded, suggesting various alternatives on the way and countenancing anti-realist replies. I finish with my own evaluation of the problems facing realists and anti-realists and prospects for resolution.

---

1. What is Realism

Realism is the thesis that the objects, properties and relations the world contains exist independently of our thoughts about them or our perceptions of them. Anti-realists either doubt or deny the existence of the entities the realist believes in or else doubt or deny their independence from our conceptions of them.

Some of the metaphysical issues over which realism and anti-realism have locked horns are the following: Are there moral values? Are there abstract objects like numbers or points or spaces or symphonies or jokes? What about selves or souls or minds? Are there colours? What of after-images, pains and other sensations? Is the past real? What of the future? Are fictional characters real? Do muons and quarks and other theoretical entities exist?

These represent only a small sample of current realist/anti-realist debates about the existence of certain sorts of entities or properties. Similar questions could be and have been raised about the mind-independence of the disputed entities or properties.

For example, many think that colours are mind-dependent. That is, whilst they do not dispute the existence of things that are red, green, blue etc., they think that if no sentient creatures had ever evolved nothing would have been coloured. The tomato's redness is, if not in the eye of the perceiver, non-existent without it. Others deem moral values mind-dependent. Once more, they do not doubt the existence of moral values but instead question their independence from the mind, believing them instead to be psychological or social constructs of some sort or other.

An obvious problem with this characterization of realism is that it seems to make anti-realism about

minds or experience obligatory! For the existence of minds is surely ‘mind-dependent’ and indeed it is in the sense that if there were no minds there'd be no minds and there'd be no experiences either.

This is not what is intended by the ‘mind-independence’ formulation, however. A realist about mental states or conscious experiences is one who holds that our world is a world containing creatures who are sometimes in states of believing, desiring, remembering, perceiving, etc. and that the world's containing (creatures who enjoy) such states is in no way dependent upon the ability of those creatures themselves to determine, either conceptually or perceptually, that it does. The world is as it is independently of what we think about it.

This characterization of realism is not universally accepted. Some anti-realists maintain that realism is committed to a distinctive (and tendentious) conception of truth [Compare Putnam 1981, 1985, 1992] or, more radically, that realism just is a thesis about the nature of truth - that truth can transcend the possibility of verification [Compare Dummett 1978, 1991, 1993].

These semantic formulations of realism are unacceptable to realists who are deflationists about truth [see the entry truth: deflationary theory of] however with the unfortunate consequence that many such realists tend to simply ignore the anti-realist's legitimate semantic challenges to their position. For deflationary theories of truth deny that truth is a substantive notion which can be used to characterise alternative metaphysical views.

Some examples may help to illustrate the mind-independence characterization. Most of us are realists about elephants. We believe the world contains elephants and that its containing them is in no way dependent upon our having perceived them or thought about them.

What about Yeti? Yeti realism is nowhere near as persuasive as pachyderm realism. Yes, there are lots of stories about large woolly ape-men hiding in the recesses of the Himalayas and occasionally interacting with humans (scaring them, mainly), but we're not sure what status to accord those stories. May be they are veridical observations of some enormous rare ape mistaken for an ‘ape- man’ or may be just erroneous observations of sociopaths in Yak coats? Alternatively, these supposed ‘sightings’ might simply be stories Tibetans tell their children.

If Yeti reports were to prove credible as reports and definite enough we might decide that we knew what it would take for a creature to be a Yeti and, after an exhaustive search, that the world contained no such creatures or admit, in the absence of any such search, that the world might well contain Yeti, and that whether it does or doesn't in no way depends upon our speculations on the matter.

We would be Eliminativists about Yeti if we decided that the world contained none, adopting an Error Theory towards the reports of Yeti, alleging some perceptual or other type of error behind them. We would be agnostic about Yeti if we decided that though the reports may be veridical and there may well be such creatures, there was as of now no good reason to believe there are any.

Suppose, on the other hand, that all we ever will have are the vague Yeti legends we now have. In that circumstance we might wonder whether the ‘sightings’ were even supposed to be reports. Perhaps the Yeti stories are like Santa Claus stories -- not intended to be taken literally? Their function might be to chasten obstreperous Tibetan children just as Santa Claus stories function to reward good Western children -- even though, naturally, both stories can only discharge their roles if the children mistakenly believe they are factual. If this was our assessment, we'd be taking a Non-Factualist attitude to Yeti stories. The function of such stories is not to describe reality.

Most of us are Eliminativists about Yeti. We think there aren't any. This is because we either think that the reports are erroneous or that the ‘sightings’ are not meant to be taken literally. Those who are agnostic about Yeti take a more charitable attitude to the reports -- they think that the world may well contain the sort of creature described therein. This sounds pretty unlikely to most ears. That's why we're Yeti Eliminativists for the most part and not Yeti Agnostics.

One thing we don't think is that it would never be rational to believe any report of a Yeti. To the contrary. The discovery of some huge new hairy ape by a team of reputable scientists in roughly the right vicinity should certainly lead us to reassess our Non-Factualist attitudes to Yeti reports. Given the indeterminacy of our ‘Yeti specifications’, such a discovery would still leave it open whether the creature was a Yeti. A more temperate conclusion would be that it had spawned the (rather fanciful)Yeti stories.

What of a new hairy hominid? This could very plausibly be a Yeti since the stories all agree that the Yeti is an ape-man. Still we would need further information before rushing to claim it was. Perhaps this hominid had never been seen before its discovery by the scientists and the legends were all based on perceptual errors of the sort canvassed above? It would just be a fluke that this creature fitted the descriptions. Homo Himalayus might then be called ‘Yeti’ by the scientists as a way of paying deference to the legends, rather than vindicating their authenticity.

This provides a contrast with the anti-realist views to be discussed below. They do think that it would never be rational to believe in the existence of the mind-independent entities the realist believes in. No evidence could convince us that some entity existed mind-independently because the very idea of mind-independent existence is incoherent.

We should note, though, that anti-realism is not intrinsically a global view. One may be a realist about certain domains and an anti-realist about others. Indeed it is not uncommon for scientific realists to be ethical anit-realists since many such thinkers hold that whilst black holes and baby universes may be mind-independent, moral values and norms are distinctively human constructions.

2. Realism and Factualism

It is often claimed that realism is a thesis about discourses or theories -- that the sentences in some discourse or theory are to be construed literally as fact- stating ones. This is factualism. It is often added that what these sentences state is largely true, although this is not a requirement of factualism per se.

It is a mistake to identify realism with factualism, however. The anti-realist views to be discussed below are factualist about the discourses describing certain contentious domains. What they contest is that the entities in these domains exist mind-independently as they find the notion of mind-independent existence incoherent.

Amongst views that generally accept the notion of mind-independent existence, a selective realism about Fs vies with selective eliminativism and agnosticism as answers to the question “Does the world contain Fs?” The realist says “Yes”, the eliminativist says “No”, the agnostic says “We can't say”. All three views agree that the existence of Fs in no way depends upon our ability to determine that they exist.

Some anti-realists agree that there are Fs but deny that anything could exist mind-independently. Unlike eliminativists and agnostics who accept it, they reject the notion of mind-independent existence. They are factualists about F discourse and anti-realist about Fs. However, such anti-realists differ from realists in their understanding of ‘Fs exist’ and, as a result, their interpretation of F discourse in general varies from that of the realist's.

Clearly, adopting a non-factualist or error-theoretic interpretation of discourse about Fs commits one to anti-realism about Fs. If we think Yeti reports are systematically mistaken or that such reports are not meant to be taken literally in the first place, we will deny that the world contains any Yetis. This means factualism is a necessary condition for realism.

It is not a sufficient condition though. Verificationists who reject the idea that something might exist even though we might never be able to confirm that it did, can be factualists about F discourse. Still, they are anti-realist about Fs since they deny that Fs could exist mind-independently.

Realism about a given class of entities may entail that discourse about them is factual but it certainly does not entail that most of our assertions concerning them are true. To the contrary, all our claims about the relevant domain could be mistaken. Consider mediaeval discourse about the cosmos. The mediaevals surely did believe that the cosmos was as it was, independently of what anyone thought about it. They were realists about the cosmos. Yet most of their beliefs were false, not true.

3. The Problem of Representation

Realism is a metaphysical thesis about what the world is like, what it contains. It is not a semantic thesis concerned with how humans represent the world in thought or language. How can it then be vulnerable to semantic challenge?

Quite simply. If the world is as resolutely mind-independent as the realist makes out, there is a problem about how we get to know about it in the first place. On what grounds can we trust our theories if they could all be radically mistaken? Wouldn't a truly mind-independent world make any representation of it in thought or language unreliable or even impossible? These are precisely the questions anti-realist urge in

their various semantic challenges to realism.

Realists do think, in the main, that we are able to represent the world reliably. Scientific realists think science is the best representation that we have of what the world is like and that its representations correspond pretty closely to the way things actually are. Yet it is crucial to their position that even our best scientific theories -- General Relativity, Quantum Theory, Theory of Evolution etc. -- could be radically mistaken. Be that as it may, when scientists talk about muons and quarks, gravitational constants, entropy, quantum fluctuations, the curvature of space-time etc. they are not just exploiting some useful linguistic devices for organizing their observational data, they are telling us what the world contains, independently of what anyone might think it contains, according to scientific realists.

For the scientist's representations of the world to be reliable, there must be a correlation between these representations and the states of affairs which they portray. So the cosmologist who utters the statement ‘The entropy of the Big Bang was remarkably low’ has uttered a truth if and only if the entropy of the Big Bang was remarkably low.

A natural question to ask is how the correlation between the statement and the mind-independent state of affairs which makes it true is supposed to be set up. How does it come about that the word ‘entropy’ refers to the amount of disorder in a system, that the descriptive name ‘The Big Bang’ refers to the event with which the Universe began? Is it that the shock waves of that cataclysmic event continue to reverberate some sixteen billion years later in human minds, dislodging a mental symbol as if it were a loose tooth, and that this mental symbol refers to whatever it was that shook it free? Clearly not. How then does that mental symbol get to refer to the Big Bang?

The only plausible answer has to do with us as cognitive beings. It is something about the way we use our words or deploy our mental symbols in thought and action which effects this correlation between mental symbol and worldly referent.

Suppose this is not so. Assume instead that God or Nature has solved this problem for us. God or Nature has set up just the right connections between our mental symbols and the bits of the world which we take ourselves to be referring to in thought.

Still we face the problem of finding evidence that this has occurred. Yet it seems the relevant evidence will be just what it was if God or Nature had not been so obliging -- linguistically, it will be the use speakers make of their words, the statements they endorse and the statements they dissent from, the rationalizations they provide for their actions, their defence and explanations of their views and criticisms of opposing views etc.; cognitively, it will be the functional role of mental symbols in thought, perception, language learning etc.

4. The Semantic Challenges to Realism

Manifestation

The first anti-realist challenge to consider focuses on the use we make of our words and sentences. The challenge is simply this: what aspect of our linguistic use could provide the necessary evidence for the realist's correlation between sentences and mind-independent states of affairs? Which aspects of our semantic behaviour manifest our grasp of these correlations, assuming they do hold?

When we look at how speakers actually do use their sentences, anti-realists claim, we see them responding not to states of affairs that they cannot in general detect but rather to agreed upon conditions for asserting these sentences. Scientists assert ‘The entropy of the Big Bang was remarkably low’ because they all concur that the conditions justifying this assertion have been met.

What prompts us to use our sentences in the way that we do are the public justification conditions associated with those sentences, justification conditions forged in linguistic practices which imbue these sentences with meaning.

The realist believes we are able to mentally represent mind-independent states of affairs. But what of cases where everything that we know about the world leaves it unsettled whether the relevant state of affairs obtains? Did Socrates sneeze in his sleep the night before he took the hemlock or did he not? How could we possibly find out? Yet realists hold that the sentence ‘Socrates sneezed in his sleep the night before he took the hemlock’ will be true if Socrates did sneeze then and false if he did not and that this is a significant semantic fact.

The Manifestation challenge to realism is to isolate some feature of the use agents make of their words or their mental symbols which effects the link between mind-independent states of affairs and the thoughts and sentences that represent them. Nothing in the thinker's linguistic behaviour, according to the anti-realist, provides evidence that this link has been forged since linguistic use is keyed to public assertibility conditions, not undetectable truth- conditions. In those cases, such as the Socrates one, where we cannot find out whether the truth-condition is satisfied or not, it is simply gratuitous to believe that there is anything we can think or say or do which could provide evidence that the link has been set up in the first place. So the anti-realist claims. [Compare Dummett 1978, 1991, 1993 Tennant 1997; Wright 1993.]

Why should we expect the evidence to be behavioural rather than, say, neurophysiological? The reason anti-realists give is that the meanings of our word and (derivatively for them) the contents of our thoughts are essentially communicable and thus must be open for all speakers and thinkers to see [Dummett 1978, 1993]

An interesting question arises as to whether our linguistic dispositions suffice to determine what we mean by our words. Saul Kripke has argued, on behalf of Wittgenstein, that the answer to this is ‘No’ -- that there are simply no facts that correspond to one's meaning Yeti by the word ‘Yeti’ irrespective of whether these facts are restricted to the behavioural. The resultant meaning scepticism has been argued by some to lead to a very radical global anti-realism which is dubiously coherent [Boghossian 1989, Wright 1984].

Language Acquisition

The second challenge to be considered concerns our acquisition of language. Suppose God had linked our mental representations to just the right states of affairs in the way required by the realist. If so, this is a semantically significant fact. Anyone learning their native language would have to grasp these correspondences between sentences and states of affairs. How can they do this if even the competent speakers whom they seek to emulate cannot detect when these correspondences hold? In short, competence in one's language would be impossible to acquire if realism were true. [Compare Dummett 1978, 1993; Wright 1993.]

Brains in a Vat?

States of affairs that are truly mind-independent go hand in glove with radical scepticism. The sceptic contends that for all we could tell we could be Brains in a Vat -- brains kept alive in a bath of nutrients by mad alien scientists. All our thoughts, all our experience, all that passed for science would be systematically mistaken if we were. We'd have no bodies although we thought we did, the world would contain no physical objects, yet it would seem to us that it did, there'd be no Earth, no Sun, no vast universe, only the brain's deluded representations of such. At least this could be the case if our representations derived even part of their content from links with mind-independent objects and states of affairs. Since realism implies such an absurd possibility could hold without our being able to detect it, it has to be rejected according to anti- realists.

A much stronger anti-realist argument uses the Brain in a Vat hypothesis to show that realism is internally incoherent rather than, as before, simply false. A crucial assumption of the argument is Semantic Externalism -- the thesis that the reference of our words and mental symbols is partially determined by contingent relations between thinkers and the world. This is a semantic assumption many realists independently endorse.

Given Semantic Externalism, the argument proceeds by claiming that if we were brains in a vat we could not possibly have the thought that we were. For, if we were so envatted, we could not possibly mean by ‘brain’ and ‘vat’ what unenvatted folk mean by these words since our words would be connected only to neural impulses or images in our brains where the unenvatteds' words are connected to real-life brains and real-life vats. Similarly the thought we pondered whenever we posed the question ‘Am I a Brain in a Vat?’ could not possibly be the thought unenvatted folk pose when they ask themselves the same-sounding question in English. But realism entails that we could indeed be Brains in a Vat. As we have just shown that were we to be so, we could not even entertain this as a possibility, realism is incoherent. [Compare Putnam 1981.]

Conceptual Relativity

If the notion of mind-independent existence is incoherent, as anti-realists contend, what should we put in

its stead? Berkeley famously answered “Mind- dependent existence!” where the Mind in question, for the good Bishop, was, of course, the Mind of God. Modern anti-realists tend not to be theists and tend not to relativize existence to any single mind. Instead of God they posit conceptual schemes as that on which the notion of existence depends. To that extent they follow Kant rather than Berkeley, though unlike Kant they tend to be pluralists -- it is conceptual schemes which they endorse rather than The One Conceptual Scheme which Kant held to be obligatory for all rational creatures.

According to this view, there can no more be an answer to the question “What objects and properties does the world contain?” outside of some scheme for classifying entities than there can be an answer to the question of whether two events A and B are simultaneous outside of some inertial frame for dating those events. The objects which exist are the objects some conceptual scheme says exists. ‘mesons exist’ really means ‘mesons exist relative to the conceptual scheme of current physics’.

Realists think there is a unitary sense of ‘object’, ‘property’ etc. for which the question ‘What objects and properties does the world contain?’ makes sense. Any answer which succeeded in listing all the objects, properties, events etc. which the world contains would comprise a privileged description of that totality. Anti-realists reject this. For them ‘object’, ‘property’ etc. shift their senses as we move from one conceptual scheme to another. Some anti-realists argue that there cannot be a totality of all the objects the world contains since the notion of ‘object’ is indefinitely extensible and so, trivially, there cannot be a privileged description of any such totality.

How does the anti-realist defend conceptual relativity? By arguing that there can be two complete theories of the world which are descriptively equivalent yet logically incompatible from the realist's point of view. For example theories of space-time can be formulated in one of two mathematically equivalent ways: either with an ontology of points, spatiotemporal regions being defined as sets of points or with an ontology of regions, points being defined as convergent sets of regions. Such theories are descriptively equivalent since mathematically equivalent and yet are logically incompatible from the realist's point of view, anti-realists contend. [Compare Putnam 1985, 1990.]

The Model-Theoretic Argument

This is the most technical of the arguments we have so far considered although we shall not reproduce the technicalities here - the central ideas can be conveyed informally, although some technical concepts will be mentioned where necessary. The argument purports to show that the Problem of Representation introduced before is insoluble for realists. That problem, to recall, was to explain how our mental symbols and words get hooked up to mind-independent objects and how our sentences and thoughts target mind-independent states of affairs.

According to the Model-Theoretic Argument, there are simply too many ways in which our mental symbols can be mapped onto items in the world. A consequence of this is that realists must either accept massive indeterminacy in what our symbols refer to or insist dogmatically that even an ideal theory, whose terms and predicates can demonstrably be mapped onto objects and properties of some sort in the

world so as to make its theses come out true might still be false, i.e. that such a mapping might not be the right one, the one ‘intended’.

Neither alternative can be defended, according to anti-realists. Massive indeterminacy in perfectly determinate terms is absurd, whilst for realists to contend that even an ideal theory could be false is to resort to dogmatism, since on their own admission we cannot tell which mapping the world has set up for us. Such dogmatism leaves the realist with no answer to a scepticism which undermines any capacity to reliably represent the world, anti-realists maintain.

Why can't an ideal theory be false? To admit that this is possible is to admit that there is a gap between what is true and what is ideally warranted by our best theory, something no anti-realist can afford to do. But an argument is needed to show this is not possible.

Anti-realists have one in the Model-Theoretic Argument. It proceeds thus: We imagine that we have an ideal theory T which passes every observational and theoretical test we can conceive of. Assume we can formalize T in first order logic. Assume also that the world is infinite in size and that our formal theory T is consistent. Then by the Completeness Theorem for first order logic, T will have a model M of the same size as the world (since by that theorem T will have models of every infinite size). Match up the objects in the model M one to one with the objects the world contains and use this mapping to define the relations of M directly in the world. We now have a correspondence between the expressions of the language L in which T is expressed and (sets of) objects in the world. T will then be true if ‘true’ just means ‘true-in-M’.

If T is not guaranteed true by this procedure it can only be because M is not the intended model. Yet all our observation sentences come out true according to M and the theoretical constraints must be satisfied because T's theses all come out true in M also. So the realist owes us an explanation of what constraints a model has to satisfy for it to be ‘intended’ over and above its satisfying every observational and theoretical constraint we can conceive of. Suppose on the other hand that the realist is able to somehow specify the intended model. Call this model M*. Then nothing the realist can do can possibly distinguish M* from a permuted variant P which can be specified following Putnam 1994b, 356-357:

We define properties of being a cat* and being a mat* such that:
(i) In the actual world cherries are cats* and trees are mats*
(ii) In every possible world the two sentences "A cat is on a mat" and "A cat* is on a mat*" have precisely the same truth value.
Instead of considering two sentences "A cat is on a mat" and "A cat* is on a mat*" now consider only the one "A cat is on a mat", allowing its interpretation to change by first adopting the standard interpretation for it and then adopting the non-standard interpretation in which the set of cats* are assigned to ’cat’ in every possible world and the set of mats* are assigned to ’mat’ in every possible world. The result will be the truth-value of "A cat is on a mat" will not change and will be exactly the same as before in every possible world. Similar non-standard reference assignments could be constructed for all the predicates of a language. [Compare Putnam 1985, 1994b.]

5. Realist Responses

We now turn to some realist responses to these challenges. The Manifestation and Language Acquisition challenges allege there is nothing in an agent's cognitive or linguistic behaviour that could provide evidence that they had grasped what it was for a sentence to be true in the realist's sense of ‘true’. How can you manifest a grasp of a notion which can apply or fail to apply without you being able to tell which? How could you ever learn to use such a concept?

One possible realist response is that the concept of truth is actually a very simple one and the demand that one always be able to determine whether a concept applies or fails to apply spurious. As to the first part, it is often argued that all there is to the notion of truth is what is given by the formula “’p’ is true if and only if p”. The function of the truth-predicate is to disquote sentences in the sense of undoing the effects of quotation -- thus all that one is saying in calling the sentence ‘Yeti are vicious’ true is that Yeti are vicious.

It is not clear that this response really addresses the anti-realist's worry, however. It may well be that there is a simple algorithm for learning the meaning of ‘true’ and that, consequently, there is no special difficulty in learning to apply the concept. But that by itself does not tell us whether the predicate ‘true’ applies to cases where we cannot ascertain that it does. All the algorithm tells us, in effect, is that if it is legitimate to assert p it is legitimate to assert that ‘p ‘ is true. So are we entitled to assert ‘Either Socrates did or did not sneeze in his sleep the night before he took the hemlock’ or are we not? Presumably that will depend on what we mean by the sentence, whether we mean to be adverting to two states of affairs neither of which we have any prospect of ever confirming.

Anti-realists follow verificationists in rejecting the intelligibility of such states of affairs and tend to model their rules for assertion on intuitionistic logic which rejects the universal applicability of the Law of Bivalence, the principle that every statement is either true or false. This law is a foundational semantic principle for classical logic.

As to the analogy with vagueness, this is a little more involved. If one accepts the Epistemic conception of vagueness then one will hold that a ‘penumbral’ case of red could indeed be red even though we could not in principle determine that it was. Since this is precisely how the realist thinks of truth, as applying or not independently of our capacity to determine this, the analogy would be apt. But the Epistemic theory of vagueness is highly controversial and other theories of vagueness deny that borderline red surfaces must either be red or not. Perhaps the realist could then link the two theories, claiming that since there is no incoherence in the Epistemic interpretation of vagueness, there is no incoherence in the realist notion of truth? Predictably, though, the anti- realist will reply that if these two theories really must stand or fall together, then they fall together.

A more direct realist response to the Manifestation challenge points to the prevalence in our linguistic practices of realist-inspired beliefs which we give expression to in what we say and do. The fact is that we do assert things like ‘Elephants exist independently of what anyone believes’ and all our actions and other

assertions confirm that we really do believe this. For example, we all agree that even if humans had never evolved on this planet the world could still have contained elephants. Furthermore, the overwhelming acceptance of classical logic by mathematicians and scientists and their rejection of intuitionistic logic for the purposes of mainstream science provides very good evidence for the coherence and usefulness of a distinctively realist understanding of truth.

Anti-realists reject this reply. They argue that all we make manifest by asserting things like ‘Either Socrates sneezed in his sleep the night before he took the hemlock or else he didn't’ is our pervasive misunderstanding of the notion of truth. They apply the same diagnosis to our belief in the mind-independence of elephants and to the counterfactual above which expresses this belief. We overgeneralize the notion of truth, believing that it applies in cases where it does not. A consequence of their view is that reality is indeterminate in surprising ways -- we have no grounds for asserting that Socrates did sneeze in his sleep the night before he took the hemlock and no grounds for asserting that he did not and no prospect of ever finding out which. Does this mean that for anti-realist the world contains no such fact as the fact that Socrates did one or the other of these two things? Not necessarily. For anti-realists who subscribe to intuitionistic principles of reasoning, the most that can be said is that there is no present warrant to assert that Socrates either did or did not sneeze in his sleep the night before he took the hemlock.

Perhaps anti-realists are right about all this. But if so, they need to explain how a practice based on a pervasive illusion can be as successful as modern science is. The fact is that anti-realists perturbed by the manifestability of realist truth are revisionists about parts of our linguistic practice and the consequence of this revisionist stance is that mathematics and science require extensive and non-trivial revision.

Much could be and has been said by anti-realists in response to this point. Standing back from the debate between the two sides is not always easy but at least this point should be made. Nothing said so far solves the Representation Problem, the problem of how our mental symbols get to target mind-independent entities in the first place, let alone the right ones. Some natural mechanism for effecting the right links must be at work for it cannot just be a primitive inexplicable fact that ‘The Big Bang’ refers to the Big Bang. If this problem could be solved, the Manifestation and Acquisition challenges would, presumably, be answered. It is, of course, the burden of the other anti-realist challenges to show that the realist cannot solve the Representation Problem.

Brains in a Vat

Realists who are naturalistic in their thinking are perhaps better placed than others to respond to this particular challenge. Recall that the Brain in a Vat argument purports to show that realism is incoherent on the grounds that it is both committed to the genuine possibility of our being Brains in a Vat and yet entails something inconsistent with this: namely, that were we to be so envatted we could not possibly have the thought that we were!

Realists have two obvious responses. They may either forswear commitment to Brains in a Vat or else

deny the Semantic Externalism which allegedly implies we could not think that we were Brains in a Vat were we to be so.

Naturalistic realists do question the coherence of the very idea of our being Brains in a Vat. For them there is no external vantage point from which one can assess our best overall theory and yet the sceptic's hypothesis feigns to occupy just such a vantage point. How so? By using terms which derive their meaning from successful theory to pose a problem which, if intelligible, would rob those very terms of meaning. In a similar vein some naturalistic realists have claimed that the mad scientists face an insoluble problem of combinatorial explosion the moment they give you any significant exploratory and volitional powers in the virtual world in which you are imprisoned.

As to the latter, it may be that the clever alien scientists have generated a convincing illusion of significant exploratory and volitional powers in the mind of the poor envatted Brain. Whether the sceptic's prospect is intelligible only at the cost of robbing the very terms in which it is framed of meaning is much more difficult to assess, however.

What of denying Semantic Externalism? Is this really a live option for realists? The answer is ‘Yes’. There are many realists who think the Representation Problem is just a pseudo-problem. When we say things like “’cat’ refers to cats” or “’quark’ refers to quarks” we are simply registering our dispositions to call everything we consider sufficiently cat-like/quark-like, ‘cat’/’quark’. According to these Semantic Deflationists, it is just a confusion to ask how the link was set up between our use of ‘The Big Bang’ and the event of that name which occurred some sixteen billion years ago. Some naturalistic story can, presumably, be told about how creatures like us developed the linguistic dispositions we did, in the telling of which it will emerge how we come to assert things like ‘The entropy of the Big Bang was very low’.

It is a moot question whether Semantic Deflationism really dissolves the Representation Problem or merely refuses to face up to it, though. However the story about the origins of our linguistic dispositions is told, it had better be that our utterances of ‘The entropy of the Big Bang was very low’ somehow end up evincing just the right sort of differential sensitivity to the Big Bang's having low entropy. For if all there is to the story are our linguistic dispositions and the conditions to which they are presently attuned, the case has effectively been ceded to the anti-realist who denies it is possible to set up a correlation between our utterances or thoughts and the mind-independent states of affairs which uniquely make them true.

A different response questions the implementation of Externalist constraints in the argument. It may well be that if we were Brains in a Vat we could not express the thought the unenvatted express when they say ‘we might be Brains in a Vat’ but this does not prove this thought is inexpressible tout court for such a Brain. Perhaps the Brain can contemplate the possibility of its own incarceration using some sophisticated indirect theoretical reasoning?

Realists in general see it as a fatal weakness of anti-realism that it does not permit fallible, finite creatures to be radically mistaken in the beliefs they form about the world. Many realists favourably disposed to Semanic Externalism do wish to hold both that we could indeed be Brains in a Vat and that even so we

could form the conjecture that we were. Of course the burden of proof is then placed on the realist to show how, compatibly with Externalism, the Brains can become aware of the possibility that they are envatted.

Realists influenced by Saul Kripke's views on metaphysics and epistemology might wish to argue that we do in fact know a priori that we are not brains in a vat. However, this is not because it is incoherent to suppose that we are. To the contrary, since we could have been brains in a vat the speculation that we in fact are is perfectly coherent. It is simply false. That we are not brains in a vat is thus a contingent a priori truth for such realists who see the brain-in-the-vat argument as conflating epistemological questions of what can be known a priori with metaphysical questions as to what is and is not genuinely possible. This response is all well and good privided we really can know a priori that we are not brains in a vat. Yet it is difficult to shake the doubt that we can know any such thing a priori - isn't it merely a better explanation of the actual course of our experience that we are not envatted?

At least this should be said. The anti-realists who reject the sceptic's thesis as unintelligible are not alone in doing this. Naturalistic realists often do as well. However, a demonstration that anti-realism alone can justifiably reject scepticism would be a very powerful point in its favour.

Yet it has to be said that the reasons anti-realists have so far offered for thinking they alone can confute the sceptic are not fully convincing. Either they give the hypothesis that I am a Brain in a Vat the same short shrift some naturalistic realists do, though for different reasons (having to do with the lack of assertibility conditions for the sentence ‘I am a Brain in a Vat’) or else they attempt to show that if I were a Brain in a Vat I'd be able to deduce that I am not since my utterances of ‘I am a Brain in a Vat’ would come out uniformly untrue. In the latter case, even if the conclusion is sustained by the reasoning (which is highly debatable) it is open to the realist to endorse it. So there is no ground at present for thinking that anti-realism alone can stave off radical scepticism as unintelligible.

Of course, anti-realists do not need to reject radical scepticism as unintelligible. They might join Kripkean realists in claiming that we know a priori that the brain-in-the-vat hypothesis is false.

Conceptual Relativity

To the extent that it seems to make the existence of all things relative to the classificatory skills of minds, the thesis of conceptual relativity looks highly counter-intuitive. Whilst it may be quite plausible to think that colours or moral values might disappear with the extinction of sentient life on Earth, it is not at all plausible to think that pachyderms and Yetis (if there are any) let alone trees, rocks and microbes would follow in their train! If our intuitions are anything to go by, then, the idea of conceptual relativity looks highly suspect.

Kant had a story to tell about why these intuitions were unreliable. This had to do with his distinction between empirical realism and transcendental idealism. According to the latter, our knowledge of what exists is nothing other than knowledge of how various objects appear to us. Of necessity, the knowing

mind cannot reach behind those appearances to how things are in themselves.

Dividing Kant's One True Conceptual Scheme into The Many suggested by modern anti-realists need not alter the basic distinction between how things are and how knowing minds represent them -- unless, of course, that distinction is itself questionable. In fact, many anti-realists do reject any such division, finding the whole idea of our being able to factor our knowledge of the world into separable contributions made by representational scheme and represented reality, quite objectionable. To them, Kant's problem with ‘noumena’ stems from a lingering, unrecognized attachment to realist metaphysics.

To the realist who complains that elephants would not cease to exist if humans vanished from the planet, the anti-realist should reply “Of course not!” To the contrary we accept a theory which licenses us to assert ‘Elephants exist’ and also licenses us to assert ‘If humans were to disappear from this planet, elephants need not follow in their train’ since the theory assures us that the existence of elephants in no way causally depends on the existence of humans. For the anti-realist the true picture is that our well-founded practices of assertion ground at one and the same time our conception of the world and our conception of humanity's place within it.

Realists are unlikely to be satisfied with this response, however. The worry is not so much that elephants might disappear, along with the rest of the mind- dependent world, with some plague that wiped out humanity but rather that whether there are to be any pachyderms in the first place apparently depends upon the conceptual schemes humans happen to chance upon! The relativity of existence to conceptual scheme is, in this respect, quite unlike the relativity of simultaneity to frame of reference.

Still, we have actual instances of conceptual schemes which explain the same phenomena equally well yet which are logically incompatible from the realist's point of view, anti-realists maintain. The earlier example of competing theories of space-time was a case in point. Recall that according to the first theory space-time consists of unextended spatiotemporal points and regions of space- time were to be explained as sets of these points whilst according to the second space-time consists of extended spatiotemporal regions and points were merely logical constructions, identifiable with convergent sets of regions.

In order to assess such examples we need a criterion of descriptive equivalence. Anti-realists have suggested that two theories are descriptively equivalent if each theory can be interpreted in the other and both theories explain the same phenomena.

Realists reject the anti-realist claim that there are two descriptively equivalent logically incompatible theories in cases such as the space-time one. Within the context of the relative interpretation of the one theory within the other, all the two alternative constructions of points in terms of regions and regions in terms of points actually show, the realist will say, is that there is a systematic way of assigning a point space to a region space and vice versa.

Anti-realist respond that the two theories really are incompatible since the region theory denies that points are physical entities. It is very hard to see how the region theory and the point theory can be both descriptively equivalent and logically incompatible, however. For if we restrict ourselves to the topology

of space-time the punctate and the region theory are descriptively equivanlent in the sense that each can be translated into the other: points as convergent sets of regions, regions as sets of points. So it is hard to see how the two theories can be logically incompatible topologically.

Differences do emerge over the contents of space-time: the properties, relations and functions definable on space or time. For punctate theories may contain details that are not duplicated in the region theories: at the stroke of midnight Cinderella's carriage changes into a pumpkin - it is a carriage up to midnight, a pumpkin thereafter. According to the region-based theory which takes temporal intervals as its primitives, that's all there is to it. But if there are temporal points, instants, there is a further fact left undecided by the story so far - viz, at the moment of midnight is the carriage still a carriage or is it a pumpkin?

So does the region-based theory fail to recognize certain facts or are these putative facts merely artefacts of the punctate theory's descriptive resources, reflecting nothing in reality? We cannot declare the two theories descriptively equivalent until we resolve this question at least.

In general, then, realists either dismiss cases of apparent logical incompatibility between two descriptively equivalent rival theories as merely apparent or question the descriptive equivalence of the two theories.

The Model Theoretic Argument

If realism is to be tenable at all, it must be possible for even our best theories to be mistaken. So realists must reject the Model-Theoretic Argument which purports to show that this is not possible. Realists have responded to the argument by rejecting the claim that a model M of the hypothetical ideal theory T passes every theoretical constraint simply because all of the theory's theses come out true in it. For there is no guarantee, they claim, that terms stand in the right relation of reference to the objects to which M links them. To be sure, if we impose a ‘right reference constraint’ as another theoretical constraint, M (or some model based on it) can interpret this constraint in such a way as to make it come out true. But there is a difference between a model's making some description of a constraint come out true and its actually conforming to that constraint, realists insist.

For their part, anti-realists have taken realist's insistence on a right reference constraint to be ‘just more theory’. This is understandable since from their point of view what it is for a model to conform to a constraint C is for us to be justified in asserting that it does. Unfortunately, this has led to something of a stand-off. Realists think that anti-realists are refusing to acknowledge a clear and important distinction. Anti-realists think that realists are simply falling back on dogmatism at a crucial point in the argument.

Many have concluded from the apparent stalemate that this particular debate is dialectically intractable -- one either sides with the realists or the anti- realists, depending on whether one thinks of truth as the realist does or as the anti-realist does. This conclusion, which would put the argument beyond the pale of rational appraisal, is a little premature, though. There are ways of explaining and illustrating the crucial issues concerning constraint satisfaction that are intelligible to anti-realists, ways which do not appeal to any verification-transcendent notions.

The Permutation Argument presents a genuine challenge to any realist who believes in determinate reference. But it does not refute realism unless realism is committed to determinate reference in the first place and it is not at all obvious that a belief in the mind-independence of reality does commit the realist to determinate reference. Realist responses to this argument vary widely. At one extreme are the 'determinatists', those who believe that Nature has set up significant, determinate referential connections between our mental symbols and items in the world. They contend that all the argument shows is that the distribution of truth-values across possible worlds is not sufficient to determine reference.

At another extreme are 'indeterminatists', realists who concede the conclusion, agreeing that it demonstrates that word-world reference is massively indeterminate or 'inscrutable'. Some infer from this that reference could not possibly consist in correspondences between mental symbols and objects in the world. For them all that makes 'elephant' refer to elephants is that our language contains the word 'elephant'. This is Deflationism about reference.

In between these two extremes are those prepared to concede the argument establishes the real possibility of significant and surprising indeterminacy in the reference of our mental symbols but who take it to be an open question whether other constraints can be found which pare down the range of reference assignments to just the intuitively acceptable ones. On this view 'elephant' may partially refer to elephants according to one acceptable reference assignment and may partially refer to elephant-stages or undetached elephant parts according to other such assignments, but not refer, even partially, to quolls or quarks.

6. Summary

We have considered a number of semantic challenges to realism, the thesis that the objects and properties that the world contains exist independently of our conception or perception of them. On all fronts, debate between realists and their anti-realist opponents is still very much open. If realists could provide a plausible theory about how correspondences between mental symbols and the items in the world to which they refer might be set up, many of these challenges could be met. Alternatively, if they could explain how, consistently with our knowledge of a mind-independent world, no such correspondences are required to begin with, many of the anti-realist objections would fall away as irrelevant. In the absence of such explanations it is still entirely reasonable for realists to believe that the correspondences are in place, however, and there can, indeed, be very good evidence for believing this. Ignorance of Nature's reference-fixing mechanism is no reason for denying it exists.

For their part, anti-realists themselves need to say more than they have so far said about how mental and semantic content is grounded in linguistic and cognitive practice. It is not obvious that they have any satisfactory answer to their own Representation Problem -- how are correlations between mental symbols and mind-dependent objects set up? Merely gesturing in the direction of accepted practices for asserting sentences is no satisfactory answer to this question if it is simply assumed that the asserted sentences have determinate meanings. How does human intervention succeed where Nature fails?

Bibliography

- Boghossian, Paul 1989 "The Rule Following Considerations" Mind 93, pp 507-49
- Devitt, Michael 1991 Realism and Truth, Second Edition, Princeton University Press.
- Dummett, Michael 1978 Truth and Other Enigmas, Duckworth.
- Dummett, Michael 1993 The Seas of Language, Oxford University Press.
- Dummett, Michael 1991 The Logical Basis of Metaphysics, Cambridge, Mass.: Harvard University Press.
- Dummett, Michael 2000 "Is Time a Continuum of Instants?", Philosophy 75, pp 497-515
- Field, Hartry 1978 “Mental Representation”, Erkenntnis 13, 9-61.
- Field, Hartry 1998 “Some Thoughts on Radical Indeterminacy”, The Monist, vol. 81, no. 2, 253-273.
- Horwich, Paul 1990 Truth, Oxford, Blackwell.
- Lewis, David 1983 New Work for a Theory of Universals, Australasian Journal of Philosophy 61, pp 343-77
- Lewis, David 1984 Putnam's Paradox, Australasian Journal of Philosophy 62, pp 221-36
- McDowell, John 1976 Truth-conditions, Bivalance and Verificationism: in Truth and Meaning: Essays in Semantics Evans, G & McDowell, J (eds) Oxford, Clarendon Press
- McGinn, Colin 1976 "Truth and Use", in Reference, Truth and Reality, Platts, M(ed) Routledge & Kegan Paul
- McGinn, Colin 1979 "An Apriori Argument for Realism", Journal of Philosophy 76, pp 113-33
- McGinn, Colin 1982 "Realist Semantics and Content Ascription", Synthese 52, pp 113-34
- Nagel, Tom 1986 The View from Nowhere, Oxford
- Nagel, Tom 1997 The Last Word, Oxford
- Papineau, David (ed) 1996 The Philosophy of Science, Oxford Readings in Philosophy, Oxford
- Putnam, Hilary 1981 Reason,Truth and History, Cambridge University Press.
- Putnam, Hilary 1985 Realism and Reason: Philosophical Papers, Volume 3, Cambridge University Press.
- Putnam, Hilary 1990 Realism with a Human Face, Harvard University Press.
- Putnam, Hilary 1992 Renewing Philosophy, Harvard University Press.
- Putnam, Hilary 1994 ‘The Dewey Lectures’, Journal of Philosophy, 91, 445- 517.
- Putnam, Hilary 1994 Words and Life, Harvard University Press.
- Tennant, Neil 1997 The Taming of the True Clarendon Press, Oxford.
- Van Fraassen, B.C. 1980 The Scientific Image Oxford
- Wright, Crispin 1984 Kripke's Account of the Argument against Private Language, Journal of Philosophy 81, pp 759-78
- Wright, Crispin 1993 Realism, Meaning and Truth, Blackwell.
- Wright, Crispin 1992 Truth and Objectivity, Harvard University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

mental representation | truth | truth: coherence theory of | truth: deflationary theory of

Acknowledgements

Thanks to Peter Forrest, Peter Roeper and a subject editor for the Stanford Encyclopedia of Philosophy.

Drew Khlentzos
dkhlentz@metz.une.edu.au

First published: January 11, 2001
Content last modified: January 11, 2001

The Deflationary Theory of Truth

According to the deflationary theory of truth, to assert that a statement is true is just to assert the statement itself. For example, to say that ‘snow is white’ is true, or that it is true that snow is white, is equivalent to saying simply that snow is white, and this, according to the deflationary theory, is all that can be said significantly about the truth of ‘snow is white’.

There are many implications of a theory of this sort for philosophical debate about the nature of truth. Philosophers often make suggestions like the following: truth consists in correspondence to the facts; truth consists in coherence with a set of beliefs or propositions; truth is the ideal outcome of rational inquiry. According to the deflationist, however, such suggestions are mistaken, and, moreover, they all share a common mistake. The common mistake is to assume that truth *has* a nature of the kind that philosophers might find out about and develop theories of. For the deflationist, truth has no nature beyond what is captured in ordinary claims such as that ‘snow is white’ is true just in case snow is white. Philosophers looking for the nature of truth are bound to be frustrated, the deflationist says, because they are looking for something that isn't there.

The deflationary theory has gone by many different names, including at least the following: the redundancy theory, the disappearance theory, the no-truth theory, the disquotational theory, and the minimalist theory. There is no terminological consensus about how to use these labels: sometimes they are used interchangeably; sometimes they are used to mark distinctions between different versions of the same general view. Here I will use ‘deflationism’, and ‘the deflationary theory of truth’ to denote the general view I want to discuss, and reserve other names for specific versions of that view.

History of Deflationism

The deflationary theory has been one of the most popular approaches to truth in the twentieth century, having received explicit defense by Frege, Ramsey, Ayer, and Quine, as well as sympathetic treatment from many others. (According to Dummett 1959, the view originates with Frege.) The following passages all contain recognizable versions of the doctrine, though they differ on points of detail.

> It is worthy of notice that the sentence 'I smell the scent of violets' has the same content as the sentence 'it is true that I smell the scent of violets'. So it seems, then, that nothing is added to the thought by my ascribing to it the property of truth. (Frege 1918)

> Truth and falsity are ascribed primarily to propositions. The proposition to which they are ascribed may be either explicitly given or described. Suppose first that it is explicitly given; then it is evident that 'It is true that Caesar was murdered' means no more than that Caesar was murdered, and 'It is false that Caesar was murdered' means no more than Caesar was not murdered. They are phrases which we sometimes use for emphasis or stylistic reasons, or to indicate the position occupied by the statement in our argument....In the second case in which the proposition is described and not given explicitly we have perhaps more of a problem, for we get statements from which we cannot in ordinary language eliminate the words 'true' or 'false'. Thus if I say 'He is always right', I mean that the propositions he asserts are always true, and there does not seem to be any way of expressing this without using the word 'true'. But suppose we put it thus 'For all p, if he asserts p, p is true', then we see that the propositional function p is true is simply the same as p, as e.g. its value 'Caesar was murdered is true' is the same as' Caesar was murdered'. (Ramsey 1927)

> ...it is evident that a sentence of the form "p is true" or "it is true that p" the reference to truth never adds anything to the sense. If I say that it is true that Shakespeare wrote *Hamlet*, or that the proposition "Shakespeare wrote *Hamlet*" is true, I am saying no more than that Shakespeare wrote *Hamlet*. Similarly, if I say that it is false that Shakespeare wrote the *Illiad*, I am saying no more than that Shakespeare did not write the *Illiad*. And this shows that the words 'true' and 'false' are not used to stand for anything, but function in the sentence merely as assertion and negation signs. That is to say, truth and falsehood are not genuine concepts. Consequently there can be no logical problem concerning the nature of truth. (Ayer 1935).

> The truth predicate is a reminder that, despite a technical ascent to talk of sentences, our eye is on the world. This cancellatory force of the truth predicate is explicit in Tarski's paradigm:
>
> > 'Snow is white' is true if and only if snow is white.

> Quotation marks make all the difference between talking about words and talking about snow. The quotation is a name of a sentence that contains a name, namely ‘snow’, of snow. By calling the sentence true, we call snow white. The truth predicate is a device for disquotation. (Quine 1970).

In addition to being popular historically, the deflationary theory has been the focus of much recent work. Perhaps its most vociferous contemporary defenders are Hartry Field and Paul Horwich.

One reason for the popularity of deflationism is its anti-metaphysical stance. Deflationism seems to deflate a grand metaphysical puzzle, a puzzle about the nature of truth, and much of modern philosophy is marked by a profound scepticism of metaphysics. Another reason for the popularity of deflationism concerns the fact that truth is a semantic notion, and therefore takes its place along with other semantic notions, such as reference, meaning, and content. Many philosophers are concerned with trying to understand these semantic notions. The deflationary theory is attractive since it suggests that, at least in the case of truth, there is less to be puzzled about here than one might expect.

The Equivalence Schema

Perhaps because of the widespread interest in the deflationism, the theory has received many different formulations. The result is that there is not so much *a* deflationary theory of truth as many. In recent times, however, the deflationary theory has most often been presented with the help of a schema, which is sometimes called *the equivalence schema* (in this schema ‘n’ is a name for ‘p’):

> (ES) n is true if and only p.

With the help of (ES), we can formulate deflationism as the view, first, that someone has the concept of truth just in case he or she accepts all (nonparadoxical) instances of this schema, and, second, that saying this captures everything significant that can be said about truth. In most cases, theories which depart from deflationism don't deny the first part of this claim; what they deny is that the equivalence schema tells us the whole truth about truth. Since such theories *add* to the equivalence schema, they are often called *inflationary* theories of truth. (The equivalence schema is associated with Alfred Tarski (1944, 1958), but it is far from obvious that Tarski was any sort of deflationist. We will set Tarski aside here.)

Two important points need to be noted about this formulation of deflationism. The first point concerns the parenthetical word ‘nonparadoxical’. Why do we not say more simply that, according to deflationism, someone has the concept of truth just in case he or she accepts *all* instances of (ES). The reason is that some instances of the schema -- most notoriously, those associated with the liar paradox -- are either false or have no truth value, and are on that ground presumably would not be accepted by someone who has the concept of truth. We can mark these exceptions by restricting the formulation to nonparadoxical instances of the schema. (For further discussion of this issue, see the discussion of objection from truth value gaps below.)

The second point is that it is important to see that deflationists are not suggesting that the equivalence schema is an *explicit* definition of truth. First, (ES) is not a definition of anything. Second, the deflationary theory does not, strictly speaking, provide an explicit definition of truth at all. What the deflationary theory provides instead is an explicit definition of *having* the concept of truth. To be more precise, the suggestion of the deflationary theory is that someone has the concept of truth just in case he or she is disposed to accept all (noncontroversial) instances of the equivalence schema, i.e., every sentence of the form ‘n is true iff p’ that is not paradoxical or in some other way deviant. Of course, deflationists *do* think that, in saying something about what it is to have the concept of truth, they have told us what the concept of truth is. But the latter is a by-product of the former; for this reason, we can say that deflationists are proposing an *implicit* definition of the concept of truth. This marks an important divide between deflationist theories and other kinds of theories of truth which attempt to provide an explicit definition of truth in other terms.

Are there versions of deflationism, or positions allied to deflationism, which do not employ the equivalence schema or some similar device? Yes, but we shall mention them here only to set them aside. One such view-which may be called *expressivism* -- is the analogue of emotivism in ethics. (This view of truth is often associated with Strawson 1950, though the attribution is a difficult one.) According to emotivism, at least in one of its most traditional forms, utterances of the form ‘torture is wrong’ do not, despite appearances, predicate ‘is wrong’ of torture; rather utterances of ‘torture is wrong’ merely indicate a negative attitude on the part of the speaker toward torture. Expressivism is the parallel position about truth. According to expressivism, utterances of the form ‘S is true’ do not, despite appearances, predicate ‘is true’ of S; rather ‘S is true’ merely indicates preparedness on the part of the speaker to assert S.

Another such view is the *prosentential theory of truth* advanced by Dorothy Grover (see Grover, Camp and Belnap 1973, and Grover 1992) . According to this theory, sentences formed with the predicate ‘is true’ are *prosentences*, where a prosentence is a device for achieving anaphoric cross-reference to sentences uttered previously in a conversation, just as pronouns are devices for achieving anaphoric cross-reference to names uttered previously in a conversation. According to the prosentential theory, for example, just as in

> (1) Mary wanted to buy a car, but she could only afford a motorbike.

we interpret ‘she’ as a pronoun anaphorically dependent on ‘Mary’, so too in

> (2) Snow is white. That is true, but it rarely looks white in Pittsburgh

we interpret ‘That is true’ as a prosentence anaphorically dependent on ‘Snow is white’.

Expressivism and the prosentential theory are close cousins of deflationism, and, in some uses of the term, might reasonably be called deflationary. However, they are also sufficiently different from those versions of deflationism that utilize the equivalence schema to be set aside here. The important difference between expressivism and the prosentential theory on the one hand, and deflationism as we are

understanding it on the other, concerns the logical structure of sentences such as ‘S is true’. For the deflationist, the structure of such sentences is very straightforward: ‘’S’ is true’ predicates the property expressed by ‘is true’ of the thing denoted by ‘S’. We might express this by saying that, according to deflationism, ‘S is true’ says *of* S that it is true, just as ‘apples are red’ says, *of* apples, that they are red or ‘John sleeps’ says, *of* John, that he sleeps. Both expressivism and the prosentential theory deny this, though for different reasons. According to expressivism, ‘S is true’ is properly interpreted not even of subject-predicate form; rather it has the structure ‘Hooray to S’. Obviously, therefore, it does not say, *of* S, that it is true. According to prosententialism, by contrast, while ‘S is true’ has a subject-predicate structure, it would still be mistaken to interpret it as being about S. For consider: according to the prosentential theory, ‘S is true’ is a prosentence which stands in for the sentence denoted by S just as ‘she’ in (1) is a pronoun which stands in for the name ‘Mary’. But we do not say that ‘she’ in (1) is about the *name* ‘Mary’; similarly, according to the prosentential theory, we should not say that ‘S is true’ is about S. To suppose otherwise would be to misconstrue the nature of anaphora.

Varieties of Deflationism

Different interpretations of the equivalence schema yield different versions of deflationism.

One important question concerns the issue of what instances of the equivalence schema are assumed to be about (equivalently: to what the names in instances of the equivalence schema are assumed to refer). According to one view, instances of the equivalence schema are about sentences, where a name for a sentence can be formulated simply by quoting the sentence -- thus "’Brutus killed Caesar’" is a name for ‘Brutus killed Caesar’. In other words, for those who hold what might be called a *sententialist* version of deflationism the equivalence schema has instances like (3):

> (3) ‘Brutus killed Caesar’ is true if and only if Brutus killed Caesar.

To make this explicit, we might say that, according to sententialism, the equivalence schema is (ES-sent):

> (ES-sent) The sentence ‘s’ is true if and only if s

According to those who hold what might be called a *propositionalist* version of deflationism, by contrast, instances of the equivalence schema are about propositions, where names of propositions are, or can be taken to be, expressions of the form ‘the proposition that p’ -- thus, ‘the proposition that Brutus killed Caesar’ is a name for the proposition that Brutus killed Caesar. For the propositionalist, in other words, instances of the equivalence schema are properly interpreted not as being about sentences but about propositions, i.e., more like (4) than (3):

> (4) The proposition that Brutus killed Caesar is true if and only if Brutus killed Caesar.

To make this explicit, we might say that, according to propositionalism, the equivalence schema is (ES-

prop):

(ES-prop) The proposition that p is true if and only if p.

To interpret the equivalence schema as (ES-sent) rather than (ES-prop), or vice versa, is to yield a different deflationary theory of truth. Hence sententialism and propositionalism are different versions of deflationism. (There are also some further ways to interpret the equivalence schema, but we shall set them aside here.)

The other dimension along which deflationary theories vary concerns the nature of the equivalence that the theories interpret instances of the equivalence schema as asserting. On one view, the right hand side and the left hand side of such instances are analytically equivalent. Thus, for sententialists, (3) asserts that, "'Brutus killed Caesar' is true" means the same as 'Brutus killed Caesar'; while for propositionalists (4) asserts that 'the proposition that Brutus killed Caesar is true' means the same as 'Brutus killed Caesar'. A second view is that the right hand side and the left hand side of claims such as (3) and (4) are only materially equivalent; this view interprets the 'if and only if' in both (3) and (4) as the biconditional of classical logic. And a third view is that claims such as (3) and (4) assert a necessary equivalence between their right hand sides and their left hand sides; that is, both (3) and (4) are to be interpreted as material biconditionals that hold of necessity.

This tripartite distinction between analytic, necessary, and material equivalence, when combined with the distinction between sententialism and propositionalism, yields six different versions of deflationism:

	Sentential	Propositional
Analytic	A	B
Material	C	D
Necessary	E	F

It is this variegated nature of deflationism that to a large extent dictates the many names that have been used for the theory. The labels 'redundancy theory', 'disappearance theory' and 'no-truth theory' have been used mainly to apply to analytic versions of deflationism: positions A or B. The label 'disquotational theory' tends to apply to sententialist versions, and in fact to material sentential deflationism: position C. The label 'minimalist theory' is a label used recently by Paul Horwich (1990) to apply to necessary versions, and in fact to necessary propositional deflationism: position F. It will not be important for us to examine all of these versions of deflationism in detail; to a large extent philosophers prefer one or other versions of these views on the basis of views from other parts of philosophy, views about the philosophy of language and metaphysics. However, it will be convenient here to settle on one version of the view. I will therefore follow Horwich in concentrating mainly on position F. Horwich calls this view 'minimalism', but I will continue simply with 'deflationism'.

The Utility of Deflationary Truth

The deflationist idea that the equivalence schema (ES-prop) provides an implicit definition of the concept of truth suggests that truth is, as the label ‘redundancy theory’ suggests, a redundant concept, a concept that we could do without. On the contrary, however, advocates of the deflationary theory (particularly those influenced by Ramsey) are at pains to point out that anyone who has the concept of truth in this sense is in possession of a very useful concept indeed; in particular, anyone who has this concept is in a position to form generalizations that would otherwise require logical devices of infinite conjunction.

Suppose, for example, that Jones for whatever reason decides that Smith is an infallible guide to the nature of reality. We might then say that Jones believes everything Smith says. To say this much, however, is not to capture the content of Jones's belief. In order to do that we need some way of expressing an infinite conjunction of something like the following form:

> If Smith says that snow is white, then snow is white, and if he says snow is pink, then snow is pink, and if he says that snow is chartreuse, then snow is chartreuse,...and so on.

The equivalence schema (ES-prop) allows us to capture this infinite conjunction. For, on the basis of the schema, we can reformulate the infinite conjunction as:

> If Smith says that snow is white, then the proposition that snow is white is true, and if he says snow is pink, then the proposition that snow is pink is true, and if he says that snow is chartreuse, then the proposition that snow is chartreuse is true,...and so on.

In turn, this reformulated infinite conjunction can be expressed as a statement whose universal quantifier ranges over propositions:

> For every proposition x, if what Smith said = x, then x is true.

Or, to put the same thing more colloquially:

> Everything Smith says is true.

This statement give us the content of Jones's belief. And the important point for deflationists is that we could not have stated the content of this belief unless we had the concept of truth as described by the deflationary theory. In fact, for most deflationists, it is this feature of the concept of truth -- its role in the formation of generalizations -- that explains why we have a concept of truth at all. This is, as it is often put, the raison d'etre of the concept of truth.

Is Truth A Property?

It is commonly said that, according to the deflationary theory, truth is not a property and therefore that,

according to the theory, if a proposition is true, it is mistaken to say that the proposition has a property, the property of being true. There is something right and something wrong about this view, and to see what is wrong and right about it will help us to understand the deflationary theory.

Consider the two true propositions (5) and (6):

(5) Caracas is the capital of Venezuela

(6) The earth revolves around the sun.

Do these propositions share a property of being true? Well, in one sense of course they do: since they are both true, we can say that there both have the property of being true. In this sense, the deflationary theory is not denying that truth is a property: truth is the property that all true propositions have.

On the other hand, when we say that two things share a property F, we often mean more than simply that they are both F; we mean in addition that there is intuitively a common explanation as to why they are both F. It is in this second sense in which deflationists are denying that truth is a property. Thus, in the case of our example, what explains the truth of (5) is that Caracas is the capital of Venezuela; and what explains this is the political history of Venezuela. On the other hand, what explains the truth of (6) is that the earth revolves around the sun; and what explains this is the nature of the solar system. The nature of the solar system, however, has nothing to do with the political history of Venezuela (or if is does the connections are completely accidental!) and to that extent there is no shared explanation as to why (5) and (6) are both true. Therefore, in this stronger sense, they have no property in common.

It will help to bring out the contrast being invoked here if we consider two properties that have nothing to do with truth, the property of being, i.e. the property of having existence, and the property of being a mammal. Consider Hillary Rodham Clinton and the Great Wall of China. Do these objects have the property of existence? Well, in one sense, they do: they both exist so they both have the property of existence. On the other hand, however, there is no common explanation as to why they both exist. What explains the existence of the Great Wall is the architectural and defense policies of classical China; what explains the existence of Hillary Rodham Clinton is Mr and Mrs Rodham. We might then say that existence is not a property and mean by this that it does not follow from the fact that two things exist that there is a common explanation as to why they exist. But now compare the property of existence with the property of being a mammal. If two things are mammals, they have the property of being a mammal, but in addition there is some common explanation as to why they are both mammals-both are descended from the same family of creatures, say. According to deflationism, the property of being true is more like the property of existence than it is like the property of being a mammal, and this is what deflationists mean when they say that truth is not a property.

The Deflationary Theory of Falsity

Truth and falsity are a package deal. It would be hard to imagine someone having the concept of truth

without also having the concept of falsity. One obvious question to ask the proponent of the deflationary theory of truth, then, is how the theory is to be extended to falsity.

A natural account of the concept of falsity defines it in terms of the concept of truth. Thus, someone has the concept of falsehood just in case they accept instances of the schema:

> (F-prop) The proposition that P is false if and only if the proposition that P is not true

A second, and initially slightly different, account of falsity defines it directly in terms of negation. According to this view, someone has the concept of falsity just in case they accept instances of the schema:

> (F-prop*) The proposition that P is false if and only if it is not the case that P

Many deflationists suppose that that (F-prop) and (F-prop*) in fact implicitly define the same concept of falsity (cf Horwich 1994). The key idea here is that there seems no reason to distinguish *being true* from *being the case*. If there is no distinction between being true and being the case, presumably there is also no distinction between 'It is not the case that p' and 'It is not true that p'. In addition, however, 'It is not true that p' is plausibly synonymous with 'the proposition that p is not true'; and this means that (F-prop) and (F-prop*) are equivalent. As we will shortly see, this account of falsity, though certainly a natural one, leaves the deflationary theory open to an important objection concerning truth-value gaps.

Objections to Deflationism

Our concern to this point has been only with what the deflationary theory is. In the remainder of this article, I consider five objections. These are by no means the only objections that have been advanced against deflationism---Horwich (1990) considers thirty-nine different objections!---but they do seem particularly obvious and important.

Objection #1: Propositions Versus Sentences.

We noted earlier that deflationism can be presented in either a sententialist version or a propositionalist version. Some philosophers have suggested, however, that the choice between these two versions constitutes a dilemma for deflationism (Jackson, Oppy and Smith 1994). The objection is that if deflationism is construed in accordance with propositionalism, then it is trivial, but if it is construed in accordance with sententialism it is false. To illustrate the dilemma, consider the following claim:

> (7) *Snow is white* is true if and only if snow is white

Now, does *snow is white* refer to a sentence or a proposition? If, on the one hand, we take (7) to be about a sentence, then, assuming (7) can be interpreted as making a necessary claim, (7) is false. On the face of

it, after all, it takes a lot more than snow's being white for it to be the case that ‘snow is white’ is true. In order that ‘snow is white’ be true, it must be the case not only that snow is white, it must in addition be the case that ‘snow is white’ *means that* snow is white. But this is a fact about language that (7) ignores. On the other hand, suppose we take *snow is white* to denote a proposition; in particular, suppose we take it to denote the proposition that snow is white. Then the theory looks to be trivial, since the proposition that snow is white is defined as being true just in case snow is white. In short, the deflationist is faced with a dilemma: take deflationism to be a theory of sentences and it is false; take it to be a theory of propositions, on the other hand, and it is trivial.

Of the two horns of this dilemma, it might seem that the best strategy for deflationists is to remain with the propositionalist version of their doctrine and accept its triviality. A trivial doctrine, after all, at least has the advantage of being true. Moreover, the charge of triviality is something that deflationists might well be expected to wear as a badge of honor: since deflationists are advocating their theory as following from mundane facts about which everyone can agree, it is no wonder that the theory they advocate is trivial.

However, there are a number of reasons why deflationists have typically not endorsed this option. First, the triviality at issue here does not have its source in the concept of truth, but rather in the concept of a proposition. Second, a trivial version of deflationism says nothing about the theory of meaning, where by ‘theory of meaning’, I mean an account of the connections between sentences of natural language and the propositions they express. After all, if deflationists are attending only to propositions, they are evidently *not* attending to the relation between sentences and propositions. Of course, one might point out that other theories of truth are also silent on the theory of meaning-why then can deflationism not be? However, the fact is that many deflationists present their doctrine as a central part of a much bigger philosophical project, viz., to provide a deflationary account of all the semantic notions, that is, notions such as truth, reference, and meaning. The problem for deflationists who grasp the second horn of the dilemma is that they must admit that there is no way to complete this project: the deflationary theory of truth can only be maintained by remaining silent about the theory of meaning. And this means that deflationism should be understood as a much more modest project than it is often taken to be.

The other possible response to this dilemma is to accept that deflationism applies inter alia to sentences, but to argue that the sentences to which it applies must be *interpreted* sentences, i.e., sentences which have meaning. Of course, if the sentences to which deflationism applies are interpreted sentences, then there will be no force to the objection that deflationism is ignoring the fact that sentences have meaning. Deflationism, on this interpretation, is not so much ignoring this fact as *assuming* it.

However, if it is to be conceded that the deflationary theory of truth applies only to sentences which have meaning, the deflationist takes on a dual task: first, to provide some *other* account of what it is for a sentence to mean what it does; second, to provide an account that does not employ the concept of truth. For after all, if deflationists *did* appeal to the concept of truth in building their theory of meaning, the doctrine would be obviously circular. Since most theories of meaning do in fact appeal to the concept of truth, this task is by no means easy. However, recent defenders of deflationism have begun it: both Paul Horwich and Hartry Field have in different ways defended a version of a use theory of meaning (cf. Field

1994, Horwich 1995). There is, however, a lot of work to be done before a use theory can be regarded as a successful theory of meaning.

Objection #2: Correspondence

It is often said that what is most obvious about truth is that truth consists in correspondence to the facts-for example, that the truth of the proposition that the earth revolves around the sun consists in its correspondence to the fact that the earth revolves around the sun. The so-called correspondence theory of truth is built around this intuition, and tries to explain the notion of truth by appeal to the notions of correspondence and fact. Even if one does not *build* one's theory of truth around this intuition however, many philosophers regard it as a condition of adequacy on any theory of truth that the theory accommodates the correspondence intuition.

It is often objected to deflationism, however, that the doctrine has particular trouble meeting this adequacy condition. One way to bring out the problem here is by focusing on a particular articulation of the correspondence intuition, an articulation favoured by deflationists themselves (Horwich 1990). According to this way of spelling it out, the intuition that a certain sentence or proposition ‘corresponds to the facts’ is the intuition that the sentence or proposition is true *because* of a certain way the world is; that is, the truth of the proposition is *explained* by some contingent fact which is usually external to the proposition itself. We might express this by saying that someone who endorses the correspondence intuition so understood would endorse:

> (8) The proposition that snow is white is true *because* snow is white

Now, the problem with (8) is that, when we combine it with the deflationary theory-or at least with a necessary version of that theory-we can derive something that is plainly false. Someone who holds a necessary version of deflationism would clearly be committed to the necessary truth of:

> (9) The proposition that snow is white is true iff snow is white.

And, since (9) is a necessary truth, it is very plausible to suppose that (8) and (9) together entail:

> (10) Snow is white because snow is white.

Unfortunately, however, (10) is false. The reason is that the relation reported by ‘because’ in (8) and (10) is a causal or explanatory relation, and such relations must obtain between distinct relata. But the relata in (10) are (obviously) not distinct. Hence (10) is false. But this means that the conjunction of (8) and (9) must be false, and that deflationism is inconsistent with the correspondence intuition. To borrow a phrase of Mark Johnston's -- who mounts a similar argument in a different context -- we might put the point differently by saying that, if deflationism is true, then what seems to be a perfectly good explanation in (8) *goes missing*; if deflationism is true, after all, then (8) is equivalent to (10), and (10) is not an explanation of anything.

How might a deflationist respond to this objection? One response is to provide a different articulation of the correspondence intuition. For example, one might point out that the connection between the proposition that snow is white and snow's being white is not a contingent connection, and suggest that this rules out (8) as a successful articulation of the correspondence intuition. That intuition (one might continue) is more plausibly given voice by (8*):

> (8*) ‘Snow is white’ is true because snow is white.

However, when (8*) is conjoined with (9), one cannot derive the problematic (10), and thus, one might think, the objection from correspondence might be avoided. Now certainly this is a possible suggestion; the problem with it, however, is that a deflationist who thinks that (8*) is true is most plausibly construed as holding a sententialist, rather than a propositionalist, version of deflationism. A sententialist version of deflationism, on the other hand, will in turn supply a version of (9), viz.:

> (9*) ‘Snow is white’ is true iff snow is white

which, at least it is interpreted as a necessary truth, will conspire with (8*) to yield (10). And we are back where we started.

Another response would be to object that ‘because’ creates an opaque context -- that is, the kind of context within which one cannot substitute co-referring expressions and preserve truth. If ‘because’ creates an opaque context, then it would be illegitimate to suppose that (8) and (9) entail (10). This too is a possibility; however, it is not clear that ‘because’ creates opaque context of the right kind. In general we can distinguish two kinds of opaque context: intensional contexts, which allow the substitution of necessarily co-referring expressions but not contingently co-referring expressions; and hyper-intensional contexts, which do not even allow the substitution of necessarily co-referring expressions. If the inference from (8) and (9) to (10) is to be successfully blocked, it is necessary that ‘because’ creates a hyper-intensional context. However, it is open to a friend of the correspondence objection to argue that, while ‘because’ creates an intensional context, it does not create a hyper-intensional context.

A final, and most radical, response would be to reject the correspondence intuition outright. This response is not in fact as drastic as it sounds. In particular, the deflationist does not have say that someone who says ‘the proposition that snow is white corresponds to the facts’ is speaking falsely. Deflationists would do better to say that such a person is simply using a picturesque or ornate way of saying that the proposition is true, where truth is understood in accordance with the deflationary theory. Indeed, the deflationist can even agree that for certain rhetorical or conversational purposes, it might be more effective to use the ‘correspondence to the facts’ talk. Nevertheless, it is important to see that this response does involve a burden, since it involves rejecting a condition of adequacy that many regard as binding on a theory of truth

Objection #3: Truth-value Gaps.

Philosophy of language has isolated a class of propositions that are supposed to fail of truth-value. According to some moral philosophers, for example, moral propositions -- such as the injunction that one ought to return people's phonecalls -- are neither true nor false. The same thing is true, according to some philosophers of language, about propositions which presuppose the existence of something which does not in fact exist -- such as the claim that the present King of France is bald; about propositions that are vague -- such as the proposition that wall hangings are furniture; and about propositions that are paradoxical, such as those that arise in connection with the liar paradox. Let us call this thesis *the gap*, since it finds a gap in the class of propositions between those that have truth-values and those that don't.

The deflationary theory of truth is inconsistent with there being a gap in the class of propositions, and this has been thought by many to be an objection to the theory. The reason for the inconsistency is very simple, and flows directly from the deflationist theory of falsity that we considered earlier. Suppose, for reductio, that the gap is correct and thus that there is a proposition Q which lacks a truth-value. Obviously, since Q lacks a truth-value, it is not the case that it is true or false. But now consider the equivalence schema (F-prop):

> (F-prop) The proposition that P is false if and only if the proposition that P is not true.

It is clear from (F-prop) that if it is not the case that Q is true or false, then it is not the case that Q is true or not true. But that is a contradiction: it must be the case that Q is true or not true. It follows that Q must have a truth value, and, of course, that there is no gap in the class of propositions.

Clearly, then, one must give up, or modify, either deflationism or the gap. Which? One strategy modifies deflationism by jettisoning the account of falsity that the deflationist offers, while hanging on to the account of truth. This strategy is a fairly desperate one, however. To begin with, if we give up the account of falsehood, it is not clear that we have an account of truth. Truth and falsehood are, as I have said, a package deal. Moreover, the deflationary theories of falsity that we considered are motivated in large part by classical logic. Presumably, it would be desirable to maintain classical logic if at all possible, and this means that we should maintain the deflationist account of falsity. Finally, one can generate a problem for the gap even if we operate without falsity, and only with truth (Rescher 1969). Suppose, again for reductio, that there is a proposition Q that is neither true nor false. Obviously, if Q is neither true nor false, then the proposition that Q is true will be false. But this means that for at least one instance of the equivalence schema, one side of the biconditional will be false, and the other side will be neither true nor false. On all logics that involve truth-value gaps, however, such a biconditional will be counted either as false or else as neither true nor false. Either way, the result is that the equivalence schema is not true in all instances. And this contradicts deflationism.

A second strategy argues that the gap, as I have presented it, is malformed. According to this strategy, one should not respond to the phenomena that prompt the gap by suggesting that certain propositions lack truth values; one should rather suggest that certain declarative sentences lack truth values, i.e., because they fail to express propositions at all. Thus, if we take presupposition failure as our example, the suggestion is that instead of supposing that the *proposition* that the present King of France is bald does

not have a truth value if the King of France does not exist, one should rather suppose that the sentence ‘the present King of France is bald’ does not express a proposition, and therefore fails to have a truth value. This kind of approach removes any conflict between the gap and deflationism. The gap says, or implies, that certain sentences fail to express propositions; deflationism says, or implies, that if those sentences *did* express propositions, they would have truth values. But there is clearly no contradiction in supposing, on the one hand, that a certain sentence fails to express a proposition and, on the other, that if it did, it would have a truth value.

A final strategy is to reject the gap entirely, and to simply agree that there is no gap which divides either propositions or sentences. This may initially seem to be an overreaction to the inconsistency of deflationism and the gap; however, what lies behind this strategy is the thought that it is not clear that the various phenomena that motivate the gap ought to be regarded as phenomena which involve failure of truth value, whether of sentences or propositions. In the case of presupposition failure, for example, it is not clear that the problem is best explained by a failure of certain sentences to have truth values, or by the presence of conventional or conversational implicatures that govern utterances of those sentences. The possibility of a broadly pragmatic account of the phenomena suggests that one might accommodate the intuitions behind the gap without supposing that there is a gap in the class of propositions (for an example, see Stalnaker 1975).

Of course, at the present stage of investigation, both of these strategies are only schematic: it is not clear that both or either of them will be sufficient to account for the various linguistic phenomena that prompt the gap. Clearly, work is required on the part of deflationists to show that these strategies are in fact up to the task.

Objection #4: Normativity.

It is commonly said that our beliefs and assertions aim at truth. The idea here, of course, is not that our beliefs and assertions are always true in a statistical sense, or even that they are mostly true. The idea is rather that truth is a *norm* of assertion. This fact about assertion and truth has often been seen to suggest that deflationism must be false. However, the felt contradiction between normativity and deflationism is difficult to make precise.

The first thing to say is that there is certainly a sense in which deflationism is not inconsistent with the idea that truth is a norm of assertion. To illustrate this, notice that we can obtain an intuitive understanding of the content of this idea without mentioning truth at all, so long as we focus on a particular case. Suppose for whatever reason that Mary sincerely believes that snow is green, has good evidence for this belief, and on the basis of this belief and evidence asserts that snow is green. We might say that there is a norm of assertion which implies that Mary is in this case open to criticism. After all, since snow is evidently not green, there must be something *incorrect* or *defective* about Mary's assertion that it is. It is this incorrection or defectiveness that the idea that truth is a norm of assertion is trying to capture.

But now let us see if we can give a general statement of the norm that lies behind this particular case. The problem of providing a general statement seems to be difficult, and for reasons that by now should be familiar. To state the norm in general we would need to be able to do something we cannot really do, namely, to complete an infinite conjunction of something like the following form:

> If someone asserts that snow is green, and snow is not green then he or she is open to criticism, and if someone asserts that grass is purple, and grass is not purple then he or she is open to criticism,...and so on.

Given the equivalence schema (F-prop*) provided by the deflationary theory of falsity, however, this infinite conjunction can be reformulated as:

> If someone asserts that snow is green and the proposition that snow is green is false, then he or she is open to criticism, and if someone asserts that grass is purple and the proposition that grass is purple is false, then he or she is open to criticism, and so on.

In turn, this reformulated infinite conjunction can be reformulated as a statement whose universal quantifier ranges over propositions:

> For all propositions p, if someone asserts that p, and p is false, then he or she is open to criticism

Or, to put it as some philosophers might:

> Truth is a norm of assertion.

For after all, if truth is a norm of assertion, then, if you assert something false, you are open to criticism. In short, then, deflationists are certainly not denying that truth is a norm of assertion; on the contrary, the concept of truth is required to state that very generalization.

If the problem of normativity is not the straightforward one that deflationists cannot account for the idea that truth is a norm of assertion, what is the problem? Crispin Wright argues that the problem is not so much that deflationists cannot account for normativity; rather, he suggests that the problem is twofold: first, that any theory of truth that *does* account for normativity is ipso facto not a *deflationary* theory properly so-called, and second, that any theory of truth which employs the equivalence schemas can account for normativity (Wright 1992; and see Price 1996 for discussion). The result is that, since most contemporary varieties of deflationism evidently employ the equivalence schemas, most contemporary varieties of deflationism are not varieties of deflationism properly so-called.

Wright's objection from normativity is a difficult one to assess. For one thing, it is difficult to find Wright's reason for supposing that the equivalence schemas play such a central role in the explanation of normativity. As we have seen, the equivalence schemas are crucial in providing a general statement of the

idea that truth is a norm of assertion, but there seems for all that no internal connection between truth and the norm in question, and thus no *internal* connection between the equivalence schemas and that norm (cf. Price 1996). Nor is it clear what role normativity plays in the distinction between an inflationary and a deflationary theory of truth. Certainly it is not good enough to simply *define* deflationism so that any deflationary theory cannot account for normativity. Of course, it is a consequence of a definition of this sort that a theory of truth is either inflationary or false; but then again, no deflationist will accept the definition.

Whatever one thinks of the details of Wright's objection, however, it does have far-reaching consequences for deflationism about truth. What the objection forces us to consider is the possibility that there is no very clear distinction between an inflationary and a deflationary theory of truth. Indeed, this possibility-that there is no clear inflationary/deflationary distinction-is the topic of the final objection to deflationism that I will discuss.

Objection #5: Inflationist Deflationism?

The final objection begins by drawing attention to a little known doctrine about truth that G.E.Moore held at the beginning of the century. Richard Cartwright describes the view as follows: "a true proposition is one that has a certain simple unanalyzable property, and a false proposition is one that lacks the property" (1987, p.73). This doctrine about truth is, of course, to be understood as the analogue for truth of the doctrine that Moore held about good, namely that good is a simple, unanalyzable quality.

The problem that this Moorean view about truth presents for the deflationary theory might best be expressed in the form of a question: what is the difference between the Moorean view and deflationism? Of course, there is a sense in which the flavour of the Moorean view is very different from the flavour of the deflationist theory about truth. After all, what could be more inflationary than thinking that truth is a property of a proposition that is unanalyzable? Certainly Moore's view about good has been viewed in this light. However, the fact that one view has a different flavour from another does not mean that, at bottom, they are not the same view. One might perhaps suggest that, according to the deflationary theory, the concept of truth has an important logical role, i.e., to capture generalizations. However, this doesn't really answer our question. For one thing, it isn't clear that Moore's notion might not also capture generalizations. For another, the idea that the deflationary concept of truth plays an important logical role doesn't distinguish the metaphysics of deflationism from the metaphysics of the Moorean view; and it is the metaphysics of the matter that the present objection really brings into focus. Alternatively, one might suggest that the distinction between truth according to deflationism and truth according to Moore's view is the distinction between having a simple unanalyzable nature, and not having a nature at all. However, what is that distinction? It is certainly not obvious that there is any distinction between having a nature about which nothing can be said and having no nature at all.

The problem is particularly acute in light of the fact that deflationism has often been discussed in the context of various claims about reductionism. In many discussions of deflationism, for example, the opponent is assumed to be a particular version of a correspondence theory that attempts to reduce the

correspondence relation to certain relations of causation (Field 1986 is a good example). However, it should be noted that this kind of view is also opposed to the kind of position that takes semantic facts-such as a proposition's being true-as primitive (Field 1972 is a good example). And the problem that we are considering for deflationism is that these two views are not simply identical in being opposed to the kind of view that explains correspondence in terms of causation: it is that they are identical *simpliciter*. The suggestion, in short, is that deflationism is identical to what initally seems to be its complete opposite, Moorean inflationism.

The decision to be an inflationist or a deflationist about truth has been called "the biggest decision a theorist of truth must make" (Boghossian 1990). Certainly this is true at an intuitive level. But it is sobering also to realize that it is not exactly clear what this decision amounts to when subjected to philosophical scrutiny. And this suggests that there is still a lot of work to be done before we can arrive at a final evaluation of the deflationary theory of truth.

Bibliography

- Ayer, A.J. 1935: ‘The Criterion of Truth’, Analysis, 3.
- Boghossian, P.A. 1990: ‘The Status of Content’, *The Philosophical Review*, Vol. XCIX, No. 2.
- Cartwright, R. 1987: ‘A Neglected Theory of Truth’. In Cartwright, R 1987: *Philosophical Essays*, Cambridge, Mass., MIT Press.
- David, M 1996: *Correspondence and Disquotation: An Essay on the Nature of Trut*h. New York, Oxford University Press.
- Devitt, M 1990: *Realism and Truth*. 2nd Edition, Princeton, Princeton University Press.
- Dummett, M 1959: ‘Truth’, *Proceedings of the Aristotelian Society*, n.s. 59. Reprinted in Dummett, M 1978 *Truth and Other Enigmas*, Oxford, Clarendon Press.
- Field, H. 1972: ‘Tarski's Theory of Truth’, *Journal of Philosophy*, 69, 347-75.
- Field, H 1986: ‘The Deflationary Conception of Truth. MacDonald, G and Wright, C. (eds.) *Fact, Science and Morality*, Oxford, Blackwell.
- Field, H. 1994: ‘Deflationist Views of Meaning and Content’, *Mind*, Vol. 103, No.411.
- Grover, D 1992: *A Prosentential Theory of Truth*, Princeton, N.J. : Princeton University Press
- Grover, D., Camp, J., and Belnap, N. 1975: ‘A Prosentential Theory of Truth’ *Philosophical Studies*, 27.
- Holton, R. 1996: ‘Minimalism and Truth-Value Gaps’. Forthcoming
- Horwich, P. 1990: *Truth*, Oxford, Blackwell.
- Horwich, P. 1994: (ed.) *Theories of Truth*. New York, Dartmouth.
- Horwich, P. 1995: ‘Meaning, Use, and Truth’, *Mind*, Vol 204, No 414.
- Jackson, F. , Oppy, G. and Smith, M. ‘Minimalism and Truth Aptness’, *Mind*, Vol 103, No 411.
- Kirkham, R.L. 1992: *Theories of Truth*, Cambridge, Mass.: MIT Press.
- Leeds, S. ‘Theories of Truth and Reference’, *Erkenntnis*, 13.
- O'Leary Hawthorne, J. and Oppy, G. and. Forthcoming. ‘Minimalism and Truth’. *Nous*.
- Price, H. 1996: ‘Three Norms of Assertability’. Forthcoming
- Quine, W.V 1970: *Philosophy of Logic*. Englewood Cliffs, Prentice Hall.

- Ramsey, F.P. 1927 ‘Facts and Propositions’, *Proceedings of the Aristotelian Society*, Vol.7.
- Rescher, N. 1969 *Many-Valued Logic*. New York, McGraw-Hill
- Strawson, P. 1950: ‘Truth’, *Proceedings of the Aristotelian Society*, Vol.24.
- Stalnaker, R. 1975: ‘Pragmatic Presupposition’. Reprinted in Davis, S. (ed.) 1991. *Pragmatics: A Reader*. New York : Oxford University Press.
- Tarksi, A. 1944: ‘The Semantic Conception of Truth’, *Philosophy and Phenomenological Research*, IV, pp.341-75
- Tarski, A. 1958: ‘The Concept of Truth in Formalized Languages’. In Tarski, A. 1958: *Logic, Semantics, Metamathematics: Papers form 1923 to 1938*, Oxford, Oxford University
- Wright, P. 1992. *Truth and Objectivity*, Cambridge, Mass., Harvard University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

propositions | truth | truth: correspondence theory of

Acknowledgements:

I would like to express my thanks to James Chase, Jacob Hohwy, Graham Oppy, and Huw Price for help in constructing this entry.

Daniel Stoljar
Australian National University and University of Colorado
daniel.stoljar@colorado.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 28, 1997
Content last modified: August 28, 1997

The Correspondence Theory of Truth

Narrowly speaking, the correspondence theory of truth is the view that truth is correspondence to a fact -- a view that was advocated by Russell and Moore early in the 20th century. But the label is usually applied much more broadly to any view explicitly embracing the idea that truth consists in a relation to reality, i.e., that truth is a relational property involving a characteristic relation (to be specified) to some portion of reality (to be specified). During the last 2300 years this basic idea has been expressed in many ways, resulting in a rather extended family of views, theories, and theory sketches. The members of the family employ various concepts for the relevant relation (correspondence, conformity, congruence, agreement, accordance, copying, picturing, signification, representation, reference, satisfaction) and/or various concepts for the relevant portion of reality (facts, states of affairs, situations, events, objects, sequences of objects, sets, properties, tropes). The resulting multiplicity of versions and reformulations of the theory is due to a blend of substantive and terminological differences.

The correspondence theory of truth is often associated with metaphysical realism. Its traditional competitors, coherentist, pragmatist, and verificationist theories of truth, are often associated with idealism, anti-realism, or relativism. In recent years, the traditional competitors have been virtually replaced (at least from publication-space) by deflationary theories of truth -- and to a lesser extent by the identity theory -- which now lead the attack against correspondence theories.

---

1. History of the Correspondence Theory

The correspondence theory is often traced back to Aristotle's well-known definition of truth (*Metaphysics* 1011b25): “To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, or of what is not that it is not, is true” -- but virtually identical formulations can be found in Plato (*Cratylus* 385b2, *Sophist* 263b). It is noteworthy that this definition does not highlight the basic correspondence intuition. Although it does invoke a relation (saying something *of* something) to reality (what is), the relation is not made very explicit, and there is no specification of what on the part of reality is responsible for the truth of a saying. The definition offers a muted, relatively minimal version of a correspondence theory. (For this reason it has also been interpreted as an impure precursor of deflationary theories of truth.) Aristotle sounds much more like a genuine correspondence theorist in the *Categories* (12b5, 14b15), where he talks of “underlying things” that make statements true and implies that these “things” (*pragmata*) are logically structured situations or facts (viz., his sitting, his not sitting). Most influential is his claim in *De Interpretatione* (16a3) that thoughts are “likenessess” (*homoiosis*) of things. Although he nowhere defines truth in terms of a thought's likeness to a thing or fact, it is clear that such a definition would fit well into his overall philosophy of mind.

In medieval authors we find a division between “metaphysical” and “semantic” versions of the correspondence theory. The former are indebted to the truth-as-likeness theme suggested by Aristotle's overall views, the latter are modeled on Aristotle's definition.

The best known is the metaphysical version presented by Thomas Aquinas: “*Veritas est adaequatio rei et intellectus*” -- truth is the equation of thing and intellect -- which he restates as: “A judgment is said to be true when it conforms to the external reality” (*De Veritate* Q.1, A.1&3; cf. *Summa Theologiae* Q.16). Aquinas credits the Neoplatonist Isaac Israeli with this definition. But there is no such definition in Isaac. It originated with the Arabic philosophers Avicenna and Averroes (*Tahafut*, 103, 302), and was introduced to the scholastics by William of Auxerre (cf. Boehner 1958; Wolenski 1994). The formula “equation of thing and intellect” is intended to leave room for the idea that “true” can be applied not only to thoughts and judgments but also to things (e.g. a true friend). Aquinas explains that a thought is said to be true because it conforms to reality, whereas a thing is said to be true because it conforms to a thought (a friend is true insofar as, and because, he conforms to our, or God's, conception of what a friend ought to be). This notion of thing-truth, which played an important role in ancient and medieval thinking, is disregarded by modern and analytic philosophers but survives to some extent in existentialist and continental philosophy.

Medieval authors who prefer a semantic version of the correspondence theory often use a peculiarly

truncated formula to render Aristotle's definition: A (mental) sentence is true iff, as it signifies, so it is (*sicut significat*, *ita est*). This emphasizes the semantic relation of signification while remaining maximally elusive about what it is that is signified by a true sentence. Foreshadowing a favorite approach of the 20th century, medieval semanticists like Ockham (*Summa Logicae II*) and Buridan (*Sophismata*) give exhaustive lists of different truth-conditional clauses for sentences of different grammatical categories; they systematically refrain from associating true sentences in general with a single ontological category.

Authors of the modern period generally convey the impression that the correspondence theory of truth is far too obvious to merit much, or any, discussion. Brief statements of some version or other can be found in almost all major writers; see e.g. Descartes 1639, ATII 597; Spinoza, *Ethics*, axiom vi; Locke, *Essay*, IV.v.i; Leibniz, *New Essays*, IV.v.ii; Hume, *Treatise*, 3.1.1; and Kant 1787, B82 -- Berkeley, who does not seem to offer any account of truth, is a potentially significant exception. Due to the influence of Thomism, metaphysical versions of the theory are more popular with the moderns than semantic versions. But since the moderns generally subscribe to a representational theory of the mind (the theory of ideas), they would seem to be committed to spelling out concepts like correspondence or conformity in terms of a psycho-semantic representation relation holding between ideas, or sequences of ideas, and appropriate portions of reality.

The now classical formulation of the theory appears in Moore (1910-11, chap. 15) and Russell: “Thus a belief is true when there is a corresponding fact, and is false when there is no corresponding fact” (1912, 129; cf. also his 1906). The self-conscious emphasis on facts as the corresponding portions of reality -- and a more serious concern with falsehood -- distinguishes this version from its precursors. Russell and Moore's forceful advocacy of truth as correspondence to fact was, at that time, an integral part of their defense of metaphysical realism. Their formulation is indebted to one of their idealist opponents, H. H. Joachim (1906), an early advocate of the coherence theory, who had set up a slightly more involved correspondence-to-fact account of truth as the main target of his attack on realism. Later, Wittgenstein (1921) and Russell (1918) developed “logical atomism”, which introduces an important modification of the correspondence approach. Further modification, and a return to more overtly semantic versions of the approach, was influenced by Tarski's (1935) technical work on truth (cf. Field 1972, Popper 1972).

2. Truth and Truthbearers

Correspondence theories of truth have been given for beliefs, thoughts, ideas, judgments, statements, assertions, utterances, sentences, and propositions. It has become customary to talk of “truthbearers” whenever one wants to stay neutral between these choices. Two points should be kept in mind: (i) The term is somewhat misleading; it is intended to refer to bearers of truth or falsehood (or alternatively, to things of which it makes sense to ask whether they are true or false, thus allowing for the possibility that some of them might be neither); (ii) One distinguishes between secondary and primary truthbearers. Secondary truthbearers are those whose truth-values (truth or falsehood) are derived from the truth-values of primary truthbearers, whose truth-values are not derived from any other truthbearers. It is often unproblematic to advocate one theory of truth for bearers of one type and another theory for bearers of a

different type (e.g., an identity theory of truth for propositions could be a component of a correspondence theory of truth for sentences): different theories applied to bearers of different types do not automatically compete. The standard segregation of truth theories into competing camps proceeds under the assumption, or pretense, that they are intended for primary truthbearers. Confusingly, there is little agreement as to which entities are properly taken to be primary truthbearers. Nowadays, the main contenders are public language sentences, sentences of the language of thought (mental representations), and propositions.

3. Simple Versions of the Correspondence Theory

Two simple forms of correspondence definitions of truth should be distinguished ("*x*" refers to whatever truthbearers might be taken as primary; the notion of correspondence might be replaced by various related notions):

> (1) *x* is true iff *x* corresponds to some fact;
> *x* is false iff *x* does not correspond to any fact;
>
> (2) *x* is true iff *x* corresponds to some state of affairs that obtains;
> *x* is false iff *x* corresponds to some state of affairs that does not obtain.

Both forms invoke portions of reality -- facts/states of affairs -- that are denoted by that-clauses or by sentential gerundives, viz., the fact/state of affairs *that snow is white*, or the fact/state of affairs of *snow's being white*. (2) is committed to the existence of non-obtaining entities of this sort; (1) is not, for to say that a fact does not obtain means, at best, that there is no such fact. It should be noted that this terminology is not standardized: some authors use "state of affairs" much like "fact" is used here (e.g. Armstrong 1997). The question whether non-obtaining entities of the relevant sort are to be accepted is the substantive issue behind such terminological variations. The difference between (2) and (1) is akin to the difference between Platonism about properties (embraces uninstantiated properties) and Aristotelianism (rejects uninstantiated properties).

Advocates of (2) typically hold that facts are states of affairs that obtain, i.e., they hold that their account of truth is in effect an analysis of (1)'s account of truth. So disagreement turns largely on the treatment of falsehood, which (1) simply identifies with the absence of truth.

The following points can be made for preferring (2) over (1): (a) Form (2) does not imply that things outside the category of truthbearers are false just because they don't correspond to facts. (b) Form (2) allows for items within the category of truthbearers that are neither true nor false, i.e., it allows for the failure of bivalence. (c) If the primary truthbearers are sentences or mental states, then states of affairs can be their meanings or contents, and the correspondence relation in (2) can be understood, accordingly, as the relation of representation, signification, meaning, or having-as-content. Facts, on the other hand, cannot be identified with the meanings or contents of sentences or mental states, on pain of the absurd consequence that false sentences and beliefs have no meaning or content. (d) Take a truth of the form '*p*

or *q*', where '*p*' is true and '*q*' false. What are the constituents of the corresponding fact? Since '*q*' is false, they cannot both be facts. (2) allows that the corresponding fact is a disjunctive state of affairs composed of a state of affairs that obtains and a state of affairs that does not obtain.

The main point in favor of (1) over (2) is that (1) is not committed to counting non-obtaining states of affairs, like the state of affairs that snow is green, as constituents of reality.

Both forms (1) and (2) should be distinguished from

> (3) *x* is true iff *x* corresponds to some fact that exists;
> *x* is false iff *x* corresponds to some fact that does not exist,

which is a confused version of (1), or a confused version of (2), or, if unconfused, implies commitment to Meinongianism, i.e., the thesis that there are things that do not exist. The lure of (3) stems from the desire to give a good account of falsehood while avoiding commitment to non-obtaining states of affairs. Moore sometimes succumbs to (3)'s temptations (1910-11, 269). It can also be found in Wittgenstein (1921, 4.25), who uses "*Sachverhalt*" to refer to (atomic) facts. The problem of falsehood -- How can we think what is not the case? -- which is related to the problem of nonexistence -- How can we think of what does not exist? -- was brought out nicely in Plato's *Theaetetus* (188c-89).

4. Arguments for the Correspondence Theory

The main positive argument given by advocates of the correspondence theory of truth is its obviousness. Descartes: "I have never had any doubts about truth, because it seems a notion so transcendentally clear that nobody can be ignorant of it…the word 'truth', in the strict sense, denotes the conformity of thought with its object" (1639, AT II 597); Kant: "The nominal definition of truth, that it is the agreement of [a cognition] with its object, is assumed as granted" (1787, B82); *Oxford English Dictionary*: "Truth, n. Conformity with fact; agreement with reality". Since the (relatively recent) arrival of apparently competing theories, correspondence theorists have added negative arguments to their arsenal, defending their view against objections and attacking (sometimes ridiculing) competing views.

5. Objections to the Correspondence Theory

***Objection 1*:** Definitions like (1) or (2) are too broad; although they apply to truths from some domains of discourse, e.g., the domain of science, they fail for others, e.g., the domain of morality: there are no moral facts.

The objection recognizes moral truths, but rejects the idea that reality contains moral facts. Logic provides another example of a domain that has been "flagged" in this way: the logical positivists recognized logical truths but rejected logical facts.

There are four possible responses to objections of this sort: (a) Error theory, which says that all claims from the flagged domain are false; (b) Noncognitivism, which says that, despite appearances to the contrary, claims from the flagged domain are not truth-evaluable to begin with, they are commands or emotions disguised as truthbearers; (c) Reductionism, which says that truths from the flagged domain correspond to facts of a different domain regarded as unproblematic, e.g., moral truths correspond to social-behavioral facts, logical truths correspond to facts about linguistic conventions; (d) Standing firm, i.e., embracing facts of the flagged domain. The last option can be supported by the following observation. The objection in effect maintains that there are different brands of *truth* (not just different brands of truths) for the different domains. This makes it hard to explain why there are many obviously valid arguments combining premises from flagged and unflagged domains.

***Objection 2*:** Correspondence theories are too obvious; they are trivial, vacuous, engaging in mere platitudes. Locutions from the “corresponds to the facts”-family are used regularly in everyday language as idiomatic substitutes for “true”. Such common turns of phrase should not be taken to indicate commitment to a correspondence *theory* in any serious sense. Definitions like (1) or (2) merely condense some trivial idioms into handy formulas; they don't deserve the grand label “theory”: there is no theoretical weight behind them (cf. Woozley 1949, chap. 6; Blackburn 1984, chap. 7.1).

In response, one could point out: (a) Definitions like (1) or (2) are “mini-theories” -- mini-theories are quite common in philosophy, and it is not obvious that they are vacuous merely because they are modeled on common usage; (b) There are correspondence theories that go beyond these definitions; (c) The complaint implies that definitions like (1) or (2) are generally accepted and are, moreover, so shallow that they are compatible with any deeper theory of truth. This makes it difficult to explain why a considerable number of thinkers emphatically reject all correspondence formulations. Moreover, the suggestion that the correspondence of *S*s belief to a fact could be said to consist in, e.g., its coherence with *S*s belief system seems quite implausible, even on the most shallow understanding of “correspondence” and “fact”.

***Objection 3*:** Correspondence theories are too obscure.

Objections of this sort, which are the most common, protest that the central notions of a correspondence theory carry unacceptable commitments and/or cannot be accounted for in any respectable manner. They might be divided into objections primarily aimed at the correspondence relation, or its relatives, and objections primarily aimed at the notions of fact or state of affairs:

3.C1: The correspondence relation must be some sort of resemblance relation. But truthbearers don't resemble anything in the world except other truthbearers -- echoing Berkeley's “an idea can be like nothing but an idea”.

3.C2: The correspondence relation is very mysterious: it seems to reach into the most distant regions of space (faster than light?) and time (past and future). How could such a relation possibly be accounted for within a naturalistic framework? What physical relation could it possibly be?

3.F1: Given the great variety of complex truthbearers, a correspondence theory will be committed to all sorts of complex “funny facts” that are ontologically disreputable: negative, disjunctive, conditional, universal, probabilistic, subjunctive, and counterfactual facts have all given cause for complaint on this score.

3.F2:All facts, even the most simple ones, are disreputable. Fact-talk, being wedded to that-clauses, is entirely parasitic on truth-talk. Facts are too much like truthbearers. Facts are fictions, spurious sentence-like slices of reality, “projected from true sentences for the sake of correspondence” (Quine 1987, 213; cf. Strawson 1950).

6. Correspondence as Isomorphism

A correspondence theory is usually expected to go beyond a mere definition like (1) or (2) and discharge a triple task: it should tell us about the workings of the correspondence relation, about the nature of facts, and about the conditions that determine which truthbearers correspond to which facts. It is natural to tackle this by construing correspondence as an *isomorphism* between truthbearers and facts (cf. Kirkham 1992, chap. 4). The basic idea is that truthbearers and facts are both complex structured entities: truthbearers are composed of words, or concepts, and other truthbearers; facts are composed of things, properties, relations, and other facts or states of affairs. The aim is to show how the correspondence relation is generated from underlying relations between the ultimate constituents of truthbearers and the ultimate constituents of their corresponding facts. One part of the project will be concerned with these correspondence-generating relations: it will lead into a theory that addresses the question how simple words, or concepts, can be *about* things, properties, and relations; i.e., it will merge with semantics or psycho-semantics (depending on what the truthbearers are taken to be). The other part of the project, the specifically ontological part, will have to provide identity criteria for facts and explain how their simple constituents combine into complex wholes. Putting all this together should yield an account of the conditions determining which truthbearers correspond to which facts.

The isomorphism approach offers an answer to objection 3.C1. Although the truth that the cat is on the mat does not resemble the cat or the mat (the truth doesn't smell, etc.), it does resemble the fact that the cat is on the mat. This is not a qualitative resemblance; it is a more abstract, structural resemblance.

The approach also puts objection 3.C2 in some perspective. The correspondence relation is supposed to reduce to underlying relations between words, or concepts, and reality. Consequently, a correspondence theory is little more than a spin-off from semantics and/or psycho-semantics (the theory of intentionality). This reminds us that, as a relation, correspondence is no more -- but also no less -- mysterious than semantic relations in general. Such relations have some curious features, and they raise a host of puzzles and difficult questions -- most notoriously: Can they be explained in terms of natural (causal) relations, or do they have to be regarded as irreducibly non-natural aspects of reality? Some philosophers have claimed that semantic relations are too mysterious to be taken seriously, usually on the grounds that they are not explainable in naturalistic terms. But one should bear in mind that this is a very

general and extremely radical attack on semantics as a whole, on the very idea that words and concepts can be about things. The common practice to aim it specifically at the correspondence theory seems misleading. As far as the intelligibility of the correspondence relation is concerned, the correspondence theory will stand, or fall, with the general theory of reference and intentionality.

On a straightforward implementation of the isomorphism approach, correspondence will be a one-one relation between truths and corresponding facts, which leaves the approach vulnerable to objections against funny facts (3.F1): each true truthbearer, no matter how complex, will be assigned a matching fact. Moreover, since the approach assigns corresponding entities to all (relevant) constituents of truthbearers, complex facts will contain objects corresponding to the logical constants ("not", "or", "if-then", etc.), and these "logical objects" will have to be regarded as constituents of the world. Many philosophers have found it hard to believe in the existence of all these funny facts and objects.

The isomorphism approach has never been advocated in a fully naïve form, assigning corresponding objects to each and every wrinkle of our verbal or mental utterings. Instead, proponents try to isolate the "relevant" constituents of truthbearers through meaning analysis, aiming to uncover the logical form, or deep structure, behind ordinary language and thought. This deep structure might then be expressed in an "ideal-language" (e.g., the language of predicate logic) whose syntactic structure is designed to mirror perfectly the ontological structure of reality. The resulting view -- correspondence as isomorphism between properly analyzed truthbearers and facts -- avoids assigning strange objects to such phrases as "the average husband", "the present king of France", or "the sake of", the view remains committed to logically complex facts and to logical objects corresponding to the logical constants.

Austin (1950) rejects the isomorphism approach on the grounds that it reads the structure of our language into the world. On his version of the correspondence theory (a more elaborated variant of (2) applied to statements), a statement as a whole is correlated to a state of affairs by arbitrary linguistic conventions without mirroring the inner structure of its correlate. This approach appears vulnerable to the objection that it avoids funny facts at the price of neglecting systematicity. Language does not provide separate linguistic conventions for each statement: that would require too vast a number of conventions. Rather, it seems that the truth-values of statements are systematically determined, via a relatively small set of conventions, by the semantic values (relations to reality) of their simpler constituents. Recognition of this systematicity is built right into the isomorphism approach.

7. Modified Versions of the Correspondence Theory

7.1 *Logical Atomism*

Wittgenstein (1921) and Russell (1918) propose a modified correspondence account of truth as part of their program of "logical atomism". Such an account proceeds in two stages: (A) the basic truth-definition, say (1), is restricted to a special subclass of truthbearers, the so-called elementary, or atomic, truthbearers; (B) the truth-values of non-elementary, or molecular, truthbearers are explained *recursively*

in terms of their logical structure and the truth-values of their simpler constituents: for example, a sentence of the form ‘not-*p*’ is true iff ‘*p*’ is false; a sentence of the form ‘*p* and *q*’ is true iff ‘*p*’ is true and ‘*q*’ is true; a sentence of the form ‘*p* or *q*’ is true iff ‘*p*’ is true or ‘*q*’ is true, etc. These recursive clauses (called “truth conditions”) can be reapplied until the truth of a molecular sentence of arbitrary complexity is reduced to the truth or falsehood of its atomic constituents. Definition (1), restricted to atomic truthbearers, serves as the base-clause for the truth-conditional recursions.

Such an account of truth is designed to go with the ontological view that the world is the totality of atomic facts (cf. Wittgenstein 1921, 2.04); i.e., atomic facts are all the facts there are -- although logical atomists tend to allow conjunctive facts, regarding them as mere aggregates of atomic facts. An atomic truth is true because it corresponds to an atomic fact: correspondence is still isomorphism, but it holds exclusively between atomic truths and atomic facts. Molecular truths are not assigned any matching facts: strictly speaking, they do not correspond to facts at all; but their truth-values are explained in terms of logical structure and correspondence of atomic constituents. One-one correspondence is restricted to the atomic level; above that level, there is no one-one correlation between truths and facts (e.g., ‘*p*’, ‘*p* or *q*’, and ‘*p* or *r*’ might all be true merely because ‘*p*’ corresponds to a fact). The trick for avoiding logically complex facts lies in not assigning any entities to the logical constants. Logical complexity, so the idea goes, belongs to the structure of language and/or thought; it is not a feature of the world.

While Wittgenstein and Russell seem to have held that the constituents of atomic facts are to be determined on the basis of *a priori* considerations, Armstrong (1997) advocates an *a posteriori* form of logical atomism. On his view, atomic facts are composed of particulars and simple universals (properties and relations). The latter are objective features of the world that ground the objective resemblances between particulars and explain their causal powers. Accordingly, what particulars and universals there are will have to be determined on the basis of total science.

Logical atomism is not easy to sustain and has rarely been held in a pure form. Among its difficulties are the following: (a) There are molecular truthbearers, like subjunctives and counterfactuals, that tend to provoke the funny-fact objection but cannot be handled by simple truth-conditional clauses because their truth-values do not seem to be determined by the truth-values of their atomic constituents. (b) Are there universal facts corresponding to true universal generalizations? Wittgenstein (1921) disapproves of universal facts; apparently, he wants to reanalyze universal generalizations as infinite conjunctions of their instances. Russell (1918) and Armstrong (1997) reject this analysis; they admit universal facts. (c) Negative truths are the most notorious problem case, because they clash with an appealing principle, the “truthmaker principle” (Armstrong 1997, chap. 8), which says that for every truth there must be something in the world that makes it true, i.e., every true truthbearer must have a truthmaker. On the account given above, ‘not-*p*’ is true iff ‘*p*’ is false iff ‘*p*’ does not correspond to any fact; hence, ‘not-*p*’ is not made true by any fact: it does not seem to have a truthmaker. Russell finds himself driven to admit negative facts, regarded by many as paradigmatically disreputable portions of reality. Wittgenstein (cf. 1921, 2.06) sometimes talks of atomic facts that do not exist and calls their very nonexistence a negative fact -- but this is hardly an atomic fact itself. Armstrong (1997, chap. 8.7) holds that negative truths are made true by a second-order “totality fact” which says of all the (positive) first-order facts that they are all the first-order facts.

Logical atomism is designed to address objections to funny facts (3.F1). It is not designed to address objections to facts in general (3.F2). Here logical atomists will respond by defending (atomic) facts. According to one defense, facts are needed because mere objects are not sufficiently *articulated* to serve as truthmakers. If *a* were the sole truthmaker of ‘*a* is *F*’, then the latter should imply ‘*a* is *G*’, for any ‘*G*’. So the truthmaker for ‘*a* is *F*’ needs at least to involve *a* and *F*ness. But since *F*ness is a universal, it could be instantiated in another object, *b*, hence the mere existence of *a* and *F*ness is not sufficient for making true the claim ‘*a* is *F*’: *a* and *F*ness need to be tied together in the fact of *a*'s being *F*. Armstrong (1997) and Olson (1987) also maintain that facts are needed to make sense of the tie that binds particular objects to universals. In this context it is usually emphasized that facts do not supervene on, hence, are not reducible to, their constituents. Facts are entities *over and above* the particulars and universals of which they are composed: *a*'s loving *b* and *b*'s loving *a* are not the same fact even though they have the same constituents. Another defense, surprisingly rare, would point out that many facts are observable: one can see that the cat is on the mat; and this is different from seeing the cat, or the mat, or both -- the objection that many facts are not observable invites the rejoinder that many objects are not observable either.

Some logical atomists propose a version of (1) without facts because they regard facts as too sentence-like. Instead, they propose events and/or objects-plus-tropes as the corresponding portions of reality. It is claimed that these items are more “thingy” than facts but still sufficiently articulated -- and sufficiently abundant -- to serve as adequate truthmakers (cf., e.g., Mulligan, Simons, and Smith 1984).

7.2 *Logical “Subatomism”*

Logical atomism aims at getting by without complex truthmakers by restricting definitions like (1) or (2) to atomic truthbearers and accounting for the truth-values of molecular truthbearers recursively in terms of their logical structure and atomic truthmakers (atomic facts, events, objects-plus-tropes). More radical modifications of the correspondence theory push the recursive strategy even further, entirely discarding definitions like (1) or (2), and hence the need for atomic truthmakers, by going, as it were, “*subatomic*”.

Such accounts analyze truthbearers, e.g., sentences, into their subsentential constituents and dissolve the relation of correspondence into appropriate semantic subrelations: names *refer* to, or denote, objects; predicates (open sentences) apply to objects, or are *satisfied* by objects. Satisfaction of complex predicates can be handled recursively in terms of logical structure and satisfaction of simpler constituent predicates: an object *o* satisfies ‘*x* is not *F*’ iff *o* does not satisfy ‘*x* is *F*’; *o* satisfies ‘*x* is *F* or *x* is *G*’ iff *o* satisfies ‘*x* is *F*’ or *o* satisfies ‘*x* is *G*’; and so on. These recursions are anchored in a base-clause addressing the satisfaction of *primitive* predicates: e.g., *o* satisfies ‘*x* is *F*’ iff *o* instantiates the property expressed by ‘*F*’ -- some would prefer a more nominalistic base-clause for satisfaction, hoping to get by without seriously invoking properties. Truth for singular sentences, consisting of a name and an arbitrarily complex predicate, is defined thus: A singular sentence is true iff the object denoted by the name satisfies the predicate. Logical machinery provided by Tarski (1935) can be used to turn this simplified sketch into a more general definition of truth -- a definition that handles sentences containing

relational predicates and quantifiers and covers molecular sentences as well. (Whether Tarski's own definition of truth can be regarded as a correspondence definition even in this modified sense is under debate; cf. Popper 1972; Field 1972, 1986; Kirkham 1992, chaps. 5-6; Soames 1999.)

Since it promises to avoid facts and all similarly articulated, sentence-like slices of reality, correspondence theorists who take seriously objection 3.F2 favor this subatomistic approach: not even atomic truthbearers are assigned any matching truthmakers. The correspondence relation itself has given way to two semantic relations between constituents of truthbearers and objects: denotation and satisfaction -- relations central to any semantic theory. Some advocates envision causal accounts of denotation and satisfaction (cf., Field 1972; Devitt 1982, 1984; Schmitt 1995; Kirkham 1992, chaps. 5-6). It turns out that relational predicates require talk of satisfaction by ordered *sequences* of objects. Davidson (1969, 1977) maintains that satisfaction by sequences is all that remains of the traditional idea of correspondence to facts; he regards denotation and satisfaction as "theoretical constructs" not in need of causal, or any, explanation.

Problems: (a) The subatomistic approach accounts for the truth-values of molecular truthbearers in the same way as the atomistic approach; consequently, molecular truthbearers that are not truth-functional pose similar problems. (b) Belief attributions and modal claims pose special problems; e.g., it seems that "believes" is a relational predicate, so that "John believes that snow is white" is true iff "believes" is satisfied by John and the object denoted by "that snow is white"; but the latter appears to be a proposition or state of affairs, which threatens to let in through the back-door the very sentence-like slices of reality the subatomic approach was supposed to avoid; (c) The phenomenon of referential indeterminacy threatens to undermine the idea that the truth-values of atomic truthbearers are always determined by the denotation and/or satisfaction of their constituents; e.g., pre-relativistic uses of the term "mass" are plausibly taken to lack determinate reference (referring determinately neither to relativistic mass nor to rest mass); yet a claim like "The mass of the earth is greater than the mass of the moon" seems to be determinately true even when made by Newton (cf. Field 1973).

Problems for both versions of modified correspondence theories: (a) It is not known whether an entirely general recursive definition of truth, one that covers all truthbearers, can be made available. This depends on unresolved issues concerning the extent to which truthbearers are amenable to the kind of structural analyses that are presupposed by the recursive clauses. The more an account of truth wants to exploit the internal structure of truthbearers, the more it will be hostage to the (limited) availability of appropriate structural analyses of the relevant truthbearers. (b) Any account of truth that employs recursions may be virtually committed to taking sentences (maybe sentences of the language of thought) as primary truthbearers. After all, the recursive clauses rely heavily on what appears to be the logico-syntactic structure of truthbearers, and it is unclear whether anything but sentences can plausibly be said to possess this kind of structure. The thesis that sentences of any sort can be regarded as primary truthbearers is contentious. (c) If clauses like "'*p* or *q*' is true iff '*p*' is true or '*q*' is true" are to be used in a recursive account of *our* notion of *truth*, as opposed to some other notion, it has to be presupposed that 'or' expresses *disjunction*: one cannot define 'or' and 'true' at the same time. To avoid circularity, a modified correspondence theory (be it atomic or subatomic) must hold that the logical connectives can be understood without reference to correspondence truth.

8. The Correspondence Theory and Its Competitors

A. Against the traditional competitors -- coherentist, pragmatist, and verificationist and other epistemic theories of truth -- correspondence theorists raise two main sorts of objections. First, these competing accounts tend to lead into relativism. Take, e.g., a coherentist account of truth. Since it is possible that '*p*' coheres with the belief system of *S* while 'not-*p*' coheres with the belief system of *S**, the coherentist account seems to imply, absurdly, that contradictories, '*p*' and 'not-*p*', could both be true. To avoid embracing contradictions, coherentists often commit themselves (if only covertly) to the objectionable relativistic view that '*p*' is true-for-*S* and 'not-*p*' is true-for-*S**. Second, the competing accounts tend to lead into some form of idealism or anti-realism. E.g., it is possible for the belief that *p* to cohere with someone's belief system even though it is not a fact that *p*; also, it is possible for it to be a fact that *p* even though no one believes that *p*, or the belief that *p* does not cohere with anyone's belief system. Cases of this form are frequently cited as counterexamples to coherentist accounts of truth. Coherentists tend to reject such counterexamples by insisting that they are not possible after all -- a reaction that commits them to the anti-realist view that the facts are (largely) determined by what we believe.

B. According to the identity theory of truth, true propositions do not correspond to facts, they are (identical with) facts: the true proposition that snow is white = the fact that snow is white. This non-traditional competitor of the correspondence theory threatens to collapse the correspondence relation into identity. In response, a correspondence theorist might point out: First, the identity theory is defensible only for propositions as truthbearers, and only if propositions are construed in a certain way, namely as having objects and properties as constituents rather than ideas or concepts of objects and properties. Hence, even if the identity theory of truth were accepted for propositions (so construed), there would still be ample room (and need) for correspondence accounts of truth with respect to other types of truthbearers. Second, the identity theory rests on the assumption that that-clauses always denote propositions, so that the that-clause in "the fact that snow is white" denotes the proposition that snow is white. The assumption can be questioned. That-clauses can be understood as ambiguous names, sometimes denoting propositions and sometimes denoting facts. The descriptive phrases "the proposition…" and "the fact…" can be regarded as serving to disambiguate the succeeding ambiguous that-clauses -- much like the descriptive phrases in "the philosopher Socrates" and "the soccer-player Socrates" serve to disambiguate the ambiguous name "Socrates" (cf. David 2002).

C. At present the most noticeable competitors to correspondence theories are deflationary accounts of truth. Deflationists maintain that correspondence theories need to be deflated. Their central notions, correspondence and fact (and their relatives), are said to play no legitimate role in an adequate account of truth: they can be excised without loss. A correspondence-type formulation like

> (4) "Snow is white" is true iff it corresponds to the fact that snow is white,

is to be deflated to

(5) “Snow is white” is true iff snow is white,

which, according to deflationists, says all there is to be said about the truth of “Snow is white” without superfluous embellishments (cf. Quine 1987, 213).

Correspondence theorists protest that (5) cannot lead to anything deserving to be regarded as an account or theory of truth because it resists generalization. (5) is a substitution instance of the schema

(6) “*p*” is true iff *p*,

which does not say anything itself and cannot be turned into a genuine generalization about truth; moreover, no genuine generalizations about truth can be accounted for on the basis of (6). Correspondence definitions, on the other hand, i.e., definitions like (1) or (2), do yield genuine generalizations about truth. (It should be noted that (4), which lends itself to deflating excisions, misrepresents the correspondence theory. According to (4), corresponding to the fact that snow is white is sufficient and necessary for “Snow is white” to be true. Yet, according to (1) and (2), it is sufficient but not necessary: “Snow is white” will be true as long as it corresponds to some fact or other. The genuine article, (1) or (2), is not as easily deflated as (4).)

This debate turns crucially on the question whether anything deserving to be called an “account” or “theory” of truth ought to take the form of a genuine generalization (and ought to be able to account for genuine generalizations involving truth). Correspondence theorists tend to regard this as a (minimal) requirement. Deflationists argue that truth is a shallow (sometimes “logical”) notion -- a notion that has no serious explanatory role to play: as such it does not require a full-fledged account, a real theory, that would have to take the form of a genuine generalization (cf. Devitt 1984; Field 1986; Kirkham 1992; Gupta 1993; David 1994; Schmitt 1995; and the essays in Blackburn and Simmons 1999, and in Schantz 2002).

9. More Objections to the Correspondence Theory

Two final objections to the correspondence theory should be mentioned.

9.1 *The Big Fact*

Inspired by a similar argument of Frege's, Davidson (1969) argues that the correspondence theory is bankrupt because it cannot avoid the consequence that all true sentences correspond to the same fact: the Big Fact.

Assume that a given sentence, *s*, corresponds to the fact that *p*; assume furthermore that ‘*p*’ and ‘*q*’ have the same truth-value. Now, since

p

implies

$\{x: x = x \;\&\; p\} = \{x: x = x\}$,

which implies

$\{x: x = x \;\&\; q\} = \{x: x = x\}$,

which in turn implies

q,

it follows that

s corresponds to the fact that *p*

implies

s corresponds to the fact that *q*.

Since the only restriction on ‘*q*’ was that it has the same truth-value as ‘*p*’, it would follow that any sentence that corresponds to any fact corresponds to every fact, so that all true sentences correspond to the same facts, thereby proving the emptiness of the correspondence theory -- the conclusion of the argument is taken as tantamount to the conclusion that every true sentence corresponds to the totality of all the facts, i.e, the Big Fact, i.e., the world as a whole.

This argument, which is a variation on the so-called “slingshot argument”, has been criticized repeatedly. Critics point out that it relies on two assumptions: (i) logically equivalent sentences can be substituted in the context “corresponds to the fact that…”; and (ii) if definite descriptions denoting the same thing can be substituted for each other in a given sentence, then they can still be so substituted if that sentence is embedded within the context “corresponds to the fact that…” (in the version above the relevant descriptions are “$\{x: x = x \;\&\; p\}$” and “$\{x: x = x \;\&\; q\}$”). It is far from obvious why correspondence theorists should be tempted by either one of these assumptions (cf. Olson 1987).

9.2 *No Independent Access*

The objection that may well have been the most effective in causing discontent with the correspondence theory is based on an epistemological concern. In a nutshell, the objection is that a correspondence theory of truth must inevitably lead into skepticism about the external world because the required correspondence between our thoughts and reality is not ascertainable. Ever since Berkeley's attack on the representational theory of the mind, objections of this sort have enjoyed considerable popularity. It is typically pointed out that we cannot step outside our own minds to compare our thoughts with mind-independent reality. Yet -- so the objection continues -- on the correspondence theory of truth, this is precisely what we would have to do to gain knowledge. We would have to access reality as it is in itself, independently of our cognition of it, and determine whether our thoughts correspond to it. Since this is impossible, since all our access to the world is mediated by our cognition, the correspondence theory makes knowledge impossible. Assuming that the resulting skepticism is unacceptable, the correspondence theory has to be rejected. This type of objection brings up a host of issues in epistemology, the philosophy of mind, and general metaphysics. All that can be done here is to hint at a few pertinent points.

The objection makes use of the following line of reasoning: “If truth is correspondence, then, since knowledge requires truth, we have to know that our beliefs correspond to reality, if we are to know anything about reality”. There are two assumptions implicit in this line of reasoning; both of them debatable. (i) It is assumed that *S* knows *x* only if *S* knows that *x* is true -- a requirement not underwritten by standard definitions of knowledge, which tell us that *S* knows *x* only if *x* is true and *S* is justified in believing *x*. The assumption may rest on confusing requirements for knowing *x* with requirements for knowing that one knows *x*. (ii) It is assumed that, if truth = *F*, then *S* knows that *x* is true only if *S* knows that *x* has *F*. This seems highly implausible. By the same standard it would follow that no one who does not know that water is H_2O can know that the Nile contains water -- which would mean, of course, that until fairly recently nobody knew that the Nile contained water (that there were stars in the sky, whales in the sea, or that the sun gives light). Moreover, even if one does know that Water is H_2O, one's strategy for finding out whether the liquid in one's glass is water does not have to involve chemical analysis. Similarly, it seems the correspondence theory does not entail that we have to know that a belief corresponds to a fact in order to know that it is true, or that our method of finding out whether a belief is true has to involve a strategy of actually comparing a belief with a fact -- although the theory does of course entail that obtaining knowledge amounts to obtaining a belief that corresponds to a fact.

More generally, one might question whether the objection still has much bite once the metaphors of “accessing” and “comparing” are spelled out with more attention to the psychological details involved in belief formation and the epistemological issues involved in the question under what conditions a belief is justified or warranted. Finally, one might wonder whether competing accounts of truth actually enjoy any significant advantage over the correspondence theory when held to the standards set up by this sort of objection.

Bibliography

- Alston, W. P., 1996, *A Realist Conception of Truth*, Ithaca and London: Cornell University Press.
- Armstrong, D. M., 1997, *A World of States of Affairs*, Cambridge: Cambridge University Press.
- Austin, J. L., 1950, ‘Truth’, reprinted in *Philosophical Papers*, 3d ed., Oxford: Oxford University Press 1979, 117-33.
- Averroes, *Tahafut Al-Tahafut*, trans. by S. Van Den Berg, *The Incoherence of the Incoherence*, London: Luzac & Co. 1954.
- Blackburn, S., 1984, *Spreading the Word: Groundings in the Philosophy of Language*, Oxford: Clarendon Press.
- Blackburn, S., and Simmons, K., (eds.), 1999, *Truth*, Oxford: Oxford University Press.
- Boehner, P., 1945, ‘Ockham's Theory of Truth’, in *Collected Articles on Ockham*, St. Bonaventure, N.Y.: Franciscan Institute 1958.
- David, M., 1994, *Correspondence and Disquotation: An Essay on the Nature of Truth*, Oxford: Oxford University Press.
- -----, 2002, ‘Truth and Identity’, in J. K. Campbell, M. O'Rourke, and D. Shier, (eds.), *Meaning and Truth: Investigations in Philosophical Semantics*, New York-London: Seven Bridges Press, 124-41.

- Davidson, D., 1969, ‘True to the Facts’, *The Journal of Philosophy* 66: 748-64.
- -----, 1977, ‘Reality Without Reference’, *Dialectica* 31: 247-53.
- Descartes, R., 1639, ‘Letter to Mersenne: 16 October 1639’, in *The Philosophical Writings of Descartes*, vol. 3, Cambridge: Cambridge University Press 1991, 138-40.
- Devitt, M., 1982, *Designation*, New York: Columbia University Press.
- -----, 1984, *Realism and Truth*, 2d ed., Oxford: Blackwell 1991.
- Gupta, A., 1993, ‘A Critique of Deflationism’, *Philosophical Topics* 21: 57-81.
- Field, H., 1972, ‘Tarski's Theory of Truth’, *The Journal of Philosophy* 69: 347-75.
- -----, 1973, ‘Theory Change and the Indeterminacy of Reference’, *The Journal of Philosophy* 70: 462-81.
- -----, 1986, ‘The Deflationary Concept of Truth’, in G. Macdonald and C. Wright, (eds.), *Fact, Science and Morality: Essays on A. J. Ayer’s ‘Language, Truth & Logic’*, Oxford: Basil Blackwell, 55-117.
- Forbes, G., 1986, ‘Truth, Correspondence and Redundancy’, in G. Macdonald and C. Wright, (eds.), *Fact, Science and Morality: Essays on A. J. Ayer's Language, Truth & Logic*, Oxford: Basil Blackwell, 27-54.
- Joachim, H. H., 1906, *The Nature of Truth*, 2d ed., Oxford: Oxford University Press, 1936.
- Kant, I., 1787, *Critique of Pure Reason*, New York: St. Martin's Press 1929.
- Kirkham, R. L., 1992, *Theories of Truth: A Critical Introduction*, Cambridge, Mass.: MIT Press.
- Moore, G. E., 1910-11, *Some Main Problems of Philosophy*, London: George Allen & Unwin 1953.
- Mulligan, K., Simons, P., and Smith, B., 1984, ‘Truth makers’, *Philosophy and Phenomenological Research* 44: 287-321. [Preprint available online]
- O'Connor, D. J., 1975, *The Correspondence Theory of Truth*, London: Hutchinson.
- Olson, K. R., 1987, *An Essay on Facts*, Stanford: CSLI lecture notes.
- Popper, K., 1972, ‘Philosophical Comments on Tarski's Theory of Truth’, in *Objective Knowledge: An Evolutionary Approach*, Oxford: Clarendon Press, 319-40.
- Quine, W. V. O., 1987, *Quiddities: An Intermittently Philosophical Dictionary*, Cambridge, Mass.: Harvard University Press.
- Russell, B., 1906-07, ‘On the Nature of Truth’, *Proceedings of the Aristotelian Society* 7: 28-49.
- -----, 1912, *Problems of Philosophy*; reprinted at Oxford: Oxford University Press 1971.
- -----, 1918, ‘The Philosophy of Logical Atomism’, in *Logic and Knowledge: Essays 1901-1950*, London: George Allen and Unwin 1956, 177-281.
- Schantz, R., ed., 2002, *What is Truth?*, Berlin-New York: De Gruyter.
- Schmitt, F. F., 1995, *Truth: A Primer*, Boulder: Westview Press.
- Soames, S., 1999, *Understanding Truth*, Oxford-New York: Oxford University Press.
- Strawson, P. F., 1950, ‘Truth’; reprinted in G. Pitcher, (ed.), *Truth*, Englewood Cliffs: Prentice-Hall 1964, 32-53.
- Tarski, A., 1935, ‘The Concept of Truth in Formalized Languages’, in *Logic, Semantics, Metamathematics*, 2d ed., Indianapolis: Hackett 1983, 152-278.
- Taylor, B., 1976, ‘States of Affairs’, in G. Evans and J. McDowell, (eds.), *Truth and Meaning: Essays in Semantics*, Oxford: Clarendon, 263-84.
- Vision, G., 1988, *Modern Anti-Realism and Manufactured Truth*, London & New York:

Routledge.
- Wittgenstein, L., 1921, *Tractatus Logico-Philosophicus*, London: Routledge 1961.
- Wolenski, J., 1994, 'Contributions to the History of the Classical Truth-Definition', in *Logic, Methodology and Philosophy of Science* 9: 481-95.
- Woozley, A. D., 1949, *Theory of Knowledge*, London: Hutchinson.

Other Internet Resources

- Truth and Correspondence, by Laurence BonJour (Chapter 3, Ph.D. dissertation, Princeton, 1969)

Related Entries

Aristotle | language of thought hypothesis | Meinong, Alexius | Moore, George Edward | propositions | realism | Russell, Bertrand | slingshot argument | Tarski, Alfred | tropes | truth: coherence theory of | truth: deflationary theory of | truth: identity theory of | Wittgenstein, Ludwig

Marian David
University of Notre Dame
david.1@nd.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 10, 2002
Content last modified: May 10, 2002

The Coherence Theory of Truth

A coherence theory of truth states that the truth of any (true) proposition consists in its coherence with some specified set of propositions. The coherence theory differs from its principal competitor, the correspondence theory of truth, in two essential respects. The competing theories give conflicting accounts of the relation between propositions and their truth conditions. (In this article, 'proposition' is not used in any technical sense. It simply refers to the bearers of truth values, whatever they may be.) According to one, the relation is coherence, according to the other, it is correspondence. The two theories also give conflicting accounts of truth conditions. According to the coherence theory, the truth conditions of propositions consist in other propositions. The correspondence theory, in contrast, states that the truth conditions of propositions are not (in general) propositions, but rather objective features of the world. (Even the correspondence theorist holds that propositions about propositions have propositions as their truth conditions.)

Versions of the Coherence Theory of Truth

The coherence theory of truth has several versions. These versions differ on two major issues. Different versions of the theory give different accounts of the coherence relation. Different varieties of the theory also give various accounts of the set (or sets) of propositions with which true propositions cohere. (I will refer to such a set as a *specified set*.)

According to some early versions of the coherence theory, the coherence relation is simply consistency. On this view, to say that a proposition coheres with a specified set of propositions is to say that the proposition is consistent with the set. This account of coherence is unsatisfactory for the following reason. Consider two propositions which do not belong to a specified set. These propositions could both

be consistent with a specified set and yet be inconsistent with each other. If coherence is consistency, the coherence theorist would have to claim that both propositions are true, but this is impossible.

A more plausible version of the coherence theory states that the coherence relation is some form of entailment. Entailment can be understood here as strict logical entailment, or entailment in some looser sense. According to this version, a proposition coheres with a set of propositions if and only if it is entailed by members of the set.

The second point on which coherence theorists (coherentists, for short) differ is the constitution of the specified set of propositions. Coherentists generally agree that the specified set consists of propositions believed or held to be true. They differ on the questions of who believes the propositions and when. At one extreme, coherence theorists can hold that the specified set of propositions is the largest consistent set of propositions currently believed by actual people. For such a version of the theory, see Young (1995). According to a moderate position, the specified set consists of those propositions which will be believed when people like us (with finite cognitive capacities) have reached some limit of inquiry. For such a coherence theory, see Putnam (1981). At the other extreme, coherence theorists can maintain that the specified set contains the propositions which would be believed by an omniscient being. Some idealists seem to accept this account of the specified set.

If the specified set is a set actually believed, or even a set which would be believed by people like us at some limit of inquiry, coherentism involves the rejection of realism about truth. Realism about truth involves acceptance of the principle of bivalence (according to which every proposition is either true or false) and the principle of transcendence (which says that a proposition may be true even though it cannot be known to be true). Coherentists who do not believe that the specified set is the set of propositions believed by an omniscient being are committed to rejection of the principle of bivalence since it is not the case that for every proposition either it or a contrary proposition coheres with the specified set. They reject the principle of transcendence since, if a proposition coheres with a set of beliefs, it can be known to cohere with the set.

Arguments for Coherence Theories of Truth

Two principal lines of argument have led philosophers to adopt a coherence theory of truth. Early advocates of coherence theories were persuaded by reflection on metaphysical questions. More recently, epistemological and semantic considerations have been the basis for coherence theories.

The Metaphysical Route to Coherentism

Early versions of the coherence theory were associated with idealism. Walker (1989) attributes coherentism to Spinoza, Kant, Fichte and Hegel. Certainly a coherence theory was adopted by a number of British Idealists in the last years of the nineteenth century and the first decades of the twentieth. See, for example, H.H. Joachim (1906).

Idealists are led to a coherence theory of truth by their metaphysical position. Advocates of the correspondence theory believe that a belief is (at least most of the time) ontologically distinct from the objective conditions which make the belief true. Idealists do not believe that there is an ontological distinction between beliefs and what makes beliefs true. From the idealists' perspective, reality is something like a collection of beliefs. Consequently, a belief cannot be true because it corresponds to something which is not a belief. Instead, the truth of a belief can only consist in its coherence with other beliefs. A coherence theory of truth which results from idealism usually leads to the view that truth comes in degrees. A belief is true to the degree that it coheres with other beliefs.

In recent years metaphysical arguments for coherentism have found few advocates. This is due to the fact that idealism is not widely held.

Epistemological Routes to Coherentism

Blanshard (1939, ch. XXVI) argues that a coherence theory of justification leads to a coherence theory of truth. His argument runs as follows. Someone might hold that coherence with a set of beliefs is the test of truth but that truth consists in correspondence to objective facts. If, however, truth consists in correspondence to objective facts, coherence with a set of beliefs will not be a test of truth. This is the case since there is no guarantee that a perfectly coherent set of beliefs matches objective reality. Since coherence with a set of beliefs is a test of truth, truth cannot consist in correspondence.

Blanshard's argument has been criticised by, for example, Rescher (1973). Blanshard's argument depends on the claim that coherence with a set of beliefs is the test of truth. Understood in one sense, this claim is plausible enough. Blanshard, however, has to understand this claim in a very strong sense: coherence with a set of beliefs is an infallible test of truth. If coherence with a set of beliefs is simply a good but fallible test of truth, as Rescher suggests, the argument fails. The "falling apart" of truth and justification to which Blanshard refers is to be expected if truth is only a fallible test of truth.

Another epistemological argument for coherentism is based on the view that we cannot "get outside" our set of beliefs and compare propositions to objective facts. A version of this argument was advanced by some logical positivists including Hempel (1935) and Neurath (1983). This argument, like Blanshard's, depends on a coherence theory of justification. The argument infers from such a theory that we can only know that a proposition coheres with a set of beliefs. We can never know that a proposition corresponds to reality.

This argument is subject to at least two criticisms. For a start, it depends on a coherence theory of justification, and is vulnerable to any objections to this theory. More importantly, a coherence theory of truth does not follow from the premisses. We cannot infer from the fact that a proposition cannot be known to correspond to reality that it does not correspond to reality. Even if correspondence theorists admit that we can only know which propositions cohere with our beliefs, they can still hold that truth consists in correspondence. If correspondence theorists adopt this position, they accept that there may be truths which cannot be known. Alternatively, they can argue, as does Davidson (1986), that the

coherence of a proposition with a set of beliefs is a good indication that the proposition corresponds to objective facts and that we can know that propositions correspond.

Coherence theorists need to argue that propositions cannot correspond to objective facts, not merely that they cannot be known to correspond. In order to do this, the foregoing argument for coherentism must be supplemented. One way to supplement the argument would be to argue as follows. As noted above, the correspondence and coherence theories have differing views about the nature of truth conditions. One way to decide which account of truth conditions is correct is to pay attention to the process by which propositions are assigned truth conditions. Coherence theorists can argue that the truth conditions of a proposition are the conditions under which speakers make a practice of asserting it. Coherentists can then maintain that speakers can only make a practice of asserting a proposition under conditions the speakers are able to recognise as justifying the proposition. Now the (supposed) inability of speakers to "get outside" of their beliefs is significant. Coherentists can argue that the only conditions speakers can recognise as justifying a proposition are the conditions under which it coheres with their beliefs. When the speakers make a practice of asserting the proposition under these conditions, they become the proposition's truth conditions. For an argument of this sort see Young (1995).

Criticisms of Coherence Theories of Truth

Any coherence theory of truth faces two principal challenges. The first may be called the specification objection. The second is the transcendence objection.

The Specification Objection

According to the specification objection, coherence theorists have no way to identify the specified set of propositions without contradicting their position. This objection originates in Russell (1907). Opponents of the coherence theory can argue as follows. The proposition (1) ‘Jane Austen was hanged for murder’ coheres with some set of propositions. (2) ‘Jane Austen died in her bed’ coheres with another set of propositions. No one supposes that the first of these propositions is true, in spite of the fact that it coheres with a set of propositions. The specification objection charges that coherence theorists have no grounds for saying that (1) is false and (2) true.

Some responses to the specification problem are unsuccessful. One could say that we have grounds for saying that (1) is false and (2) is true because the latter coheres with propositions which correspond to the facts. Coherentists cannot, however, adopt this response without contradicting their position. Sometimes coherence theorists maintain that the specified system is the most comprehensive system, but this is not the basis of a successful response to the specification problem. Coherentists can only, unless they are to compromise their position, define comprehensiveness in terms of the size of a system. Coherentists cannot, for example, talk about the most comprehensive system composed of propositions which correspond to reality. There is no reason, however, why there cannot be two or more equally large systems. Other criteria of the specified system, to which coherentists frequently appeal, are similarly unable to solve the specification problem. These criteria include simplicity, empirical adequacy and

others. Again, there seems to be no reason why two or more systems cannot equally meet these criteria.

Although some responses to the Russell's version of the specification objection are unsuccessful, it is unable to refute the coherence theory. Coherentists do not believe that the truth of a proposition consists in coherence with any arbitrarily chosen set of propositions. Rather, they hold that truth consists in coherence with a set of beliefs, or with a set of propositions held to be true. No one actually believes the set of propositions with which (1) coheres. Coherence theorists conclude that they can hold that (1) is false without contradicting themselves.

A more sophisticated version of the specification objection has recently been advanced by Walker (1989); for a discussion, see Wright (1995). Walker argues as follows. In responding to Russell's version of the specification objection, coherentists claim that some set of propositions, call it *S*, is believed. They are committed to the truth of (3) ‘*S* is believed.’ The question of what it is for (3) to be true then arises. Coherence theorists might answer this question by saying that "‘*S* is believed’ is believed" is true. If they give this answer, they are apparently off on an infinite regress, and they will never say what it is for a proposition to be true. Their plight is worsened by the fact that arbitrarily chosen sets of propositions can include propositions about what is believed. So, for example, there will be a set which contains ‘Jane Austen was hanged for murder,’ "‘Jane Austen was hanged for murder’ is believed," and so on. The only way to stop the regress seems to be to say that the truth conditions of (3) consist in the fact *S* is believed. If, however, coherence theorists adopt this position, they seem to contradict their own position by accepting that the truth conditions of some proposition consist in facts, not in propositions in a set of beliefs.

There is some doubt about whether Walker's version of the specification objection succeeds. Coherence theorists can reply to Walker by saying that nothing in their position is inconsistent with the view that there is a fact about which set of propositions is believed. Even though this fact obtains, however, the truth conditions of propositions, including propositions about which sets of propositions are believed, are the conditions under which they cohere with a set of propositions. For a defence of the coherence theory again Walker's version of the specification objection, see Young (2001).

A coherence theory of truth gives rise to a regress, but it is not a vicious regress and the correspondence theory faces a similar regress. If we say that *p* is true if and only if it coheres with a specified set of propositions, we may be asked about the truth conditions of ‘*p* coheres with a specified set.’ Plainly, this is the start of a regress, but not one to worry about. It is just what one would expect, given that the coherence theory states that it gives an account of the truth conditions of all propositions. The correspondence theory faces a similar benign regress. The correspondence theory states that a proposition is true if and only if it corresponds to certain objective conditions. The proposition ‘*p* corresponds to certain objective conditions’ is also true if and only if it corresponds to certain objective conditions, and so on.

The Transcendence Objection

The transcendence objection charges that a coherence theory of truth is unable to account for the fact that some propositions are true which cohere with no set of beliefs. According to this objection, truth transcends any set of beliefs. Someone might argue, for example, that the proposition ‘Jane Austen wrote ten sentences on November 17th, 1807’ is either true or false. If it is false, some other proposition about how many sentences Austen wrote that day is true. No proposition, however, about precisely how many sentences Austen wrote coheres with any set of beliefs and we may safely assume that none will ever cohere with a set of beliefs. Opponents of the coherence theory will conclude that there is at least one true proposition which does not cohere with any set of beliefs.

Some versions of the coherence theory are immune to the transcendence objection. A version which holds that truth is coherence with the beliefs of an omniscient being is proof against the objection. Every truth coheres with the set of beliefs of an omniscient being. All other versions of the theory, however, have to cope with the objection, including the view that truth is coherence with a set of propositions believed at the limit of inquiry. Even at the limit of inquiry, finite creatures will not be able to decide every question, and truth may transcend what coheres with their beliefs.

Coherence theorists can defend their position against the transcendence objection by maintaining that the objection begs the question. Those who present the objection assume, generally without argument, that it is possible that some proposition be true even though it does not cohere with any set of beliefs. This is precisely what coherence theorists deny. Coherence theorists have arguments for believing that truth cannot transcend what coheres with some set of beliefs. Their opponents need to take issue with these arguments rather than simply assert that truth can transcend what coheres with a specified system.

Bibliography

- Blanshard, B., 1939, *The Nature of Thought*, George Allen and Unwin, London.
- Davidson, D., 1986, "A Coherence Theory of Truth and Knowledge," *Truth And Interpretation, Perspectives on the Philosophy of Donald Davidson*, ed. Ernest LePore, Basil Blackwell, Oxford, 307-19.
- Hempel, C., 1935, "On the Logical Positivists' Theory of Truth," *Analysis* **2**,49-59.
- Joachim, H.H., 1906, *The Nature of Truth*, Oxford University Press, Oxford.
- Neurath, O., 1983, *Philosophical Papers 1913-46*, eds. Robert S. Cohen and Marie Neurath, D. Reidel, Dordrecht and Boston.
- Putnam, H., 1981, *Reason, Truth and History*, Cambridge University Press, Cambridge.
- Rescher, N., 1973, *The Coherence Theory of Truth*, Oxford University Press, Oxford.
- Russell, B., 1907, "On the Nature of Truth," *Proceedings of the Aristotelian Society* **7**, 228-49.
- Walker, R.C.S., 1989, *The Coherence Theory of Truth: Realism, anti-realism, idealism*, Routledge, London and New York.
- Wright, C., 1995, "Critical Study: Ralph C.S. Walker, *The Coherence Theory of Truth: Realism, anti-realism, idealism*," *Synthese* **103**, 279- 302.
- Young, J.O., 1995, *Global Anti-realism*, Avebury, Aldershot.

- Young, J.O., 2001, "A Defence of the Coherence Theory of Truth," *The Journal of Philosophical Research*, **26**, 89-101.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

idealism: British | knowledge: analysis of | logical positivism | realism: semantic challenges to | Russell, Bertrand | truth: correspondence theory of | truth: deflationary theory of | truth: identity theory of | truth: revision theory of

James O. Young
joyoung@uvvm.uvic.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 3, 1996
Content last modified: May 30, 2001

The Identity Theory of Truth

The simplest and most general statement of the identity theory of truth is that when a truth-bearer (e.g. a proposition) *is* true, there is a truth-maker (e.g. a fact) with which it is identical and the truth of the former *consists in* its identity with the latter. The theory is best understood by contrast with a rival such as the correspondence theory, according to which the relation of truth-bearer to truth-maker is correspondence rather than identity.

Section Links:

Sources

The theory is a response to certain intellectual pressures. One such pressure is the wish that there should be no gap between mind and world: that when we think truly, we think *what is the case*. Another is dissatisfaction with the correspondence theory of truth, of the sort expressed by Frege [Frege (1918), p. 3]:

> A correspondence, moreover, can only be perfect if the corresponding things coincide and so are just not different things. ... It would only be possible to compare an idea with a thing if the thing were an idea too. And then, if the first did correspond perfectly with the second, they would coincide. But this is not at all what people intend when they define truth as the correspondence of an idea with something real. For in this case it is essential precisely that the reality shall be distinct from the idea. But then there can be no complete correspondence, no complete truth. So nothing at all would be true; for what is only half true is untrue. Truth does not admit of more and less.

Frege then goes on to deploy a charge of circularity against the likely reply that all the correspondence theory requires is correspondence in a certain respect. He himself concluded that truth was indefinable; but some have thought it possible to formulate an identity theory of a recognizably Fregean sort.

[Return to Section Links]

A Genuine Theory?

This theory is notably absent from textbook discussions of truth; and there is controversy over whether it is a theory of truth at all. Those who think that it is not are likely to make one or both of the following objections: it is obviously absurd; no one has ever held it. The remainder of this section is devoted to considering these two objections.

Can the absurdity charge be met?

The identity theory is clearly absurd from the point of view of those who, for instance, believe that truth-bearers are sentences and truth-makers non-linguistic states of affairs. But it may be available to those who hold the kinds of metaphysical views which make truth-bearers and truth-makers more alike. (For ease of expression, I shall from now on use the vocabulary of ‘judgments’ and ‘facts’ for truth-bearers and truth-makers respectively, recognizing that these terms can be tendentious -- especially in expressing the views of philosophers who abjured them.)

Making judgments more like facts

Some philosophers have tried to make judgments more like facts. Russell, reacting against idealism, at one stage adopted a view of judgment which did not regard it as an intermediary between the mind and the world: instead, the constituents of judgments are the very things the judgments are about. This involves a kind of realism about judgments, and looks as though it offers the possibility of an identity theory of truth. But since both true and false judgments are equally composed of real constituents, truth would not be distinguished from falsehood by being identical with reality; an identity theory of truth is thus unavailable on this view of judgment because it would be rendered vacuous by being inevitably accompanied by an identity theory of falsehood. Those who have held this sort of view of judgments, such as Moore and Russell, have accordingly been forced to hold that truth is an unanalyzable property of some judgments. If one looks for an identity theory here, one finds what might be called an identity theory of judgment rather than of truth. [Less brutally condensed accounts of these matters can be found in Baldwin (1991), Candlish (1989) and Candlish (1996).]

Making facts more like judgments

Other philosophers, notably those who have held the idealist view that reality is experience, have implied that facts are more like judgments. One such is F.H. Bradley, who explicitly embraced an identity theory

of truth, regarding it as the only account capable of resolving the difficulties he finds with the correspondence theory. [See Bradley (1907).] The way he reaches it is worth describing in a little detail, for it shows how he could avoid allowing the theory to be rendered vacuous by an accompanying identity theory of falsehood.

Bradley argues that the correspondence theory's view of facts as real and mutually independent entities is unsustainable: the impression of their independent existence is the outcome of the illegitimate projection on to the world of the divisions with which thought must work, a projection which creates the illusion that a judgment can be true by corresponding to part of a situation: as, e.g., the remark ‘The pie is in the oven’ might appear to be true despite its (by omission) detaching the pie from its dish and the oven from the kitchen. His hostility to such abstraction ensures that, according to Bradley's philosophical logic, at most one judgment can be true -- that which encapsulates reality in its entirety. This allows his identity theory of truth to be accompanied by a non-identity theory of falsehood, since he can account for falsehood as a falling short of this vast judgment and hence as an abstraction of part of reality from the whole. The result is his adoption of the idea that there are degrees of truth: that judgment is the least true which is the most distant from the whole of reality. Although the consequence is that all ordinary judgments will turn out to be more or less infected by falsehood, Bradley allows some sort of place for false judgment and the possibility of distinguishing worse from better. One might argue that the reason the identity theory of truth remains only latent in Russell and Moore is the surrounding combination of their atomistic metaphysics and their assumption that truth is not a matter of degree.

For Bradley, then, at most one judgment can be fully true. But even this one judgment has so far been conceived as *describing* reality, and its truth as consisting in correspondence with a reality not distorted by being mentally cut up into illusory fragments. Accordingly, even this one, for the very reason that it remains a description, will be infected by falsehood unless it ceases altogether to be a judgment and *becomes* the reality it is meant to be *about*. This apparently bizarre claim becomes intelligible if seen as both the most extreme expression of his hostility to abstraction and a reaction to the most fundamental of his objections to the correspondence theory, which is the same as Frege's: that for there to be correspondence rather than identity between judgment and reality, the judgment must differ from reality and in so far as it does differ, to that extent must distort and so falsify it.

Thus Bradley's version of the identity theory turns out to be misleadingly so-called. For it is in fact an eliminativist theory: when truth is attained, judgments disappear and only reality is left. It is not surprising that Bradley, despite expressing his theory in the language of identity, talked of the attainment of complete truth in terms of thought's suicide. In the end, then, even the attribution of the identity theory of truth to one who explicitly endorsed it turns out to be dubious. [For a more detailed version of this section, see Candlish (1995). For other doubts about whether Bradley was an identity theorist, see Walker (1998).]

A metaphysically neutral identity theory

More recently there have been attempts, consciously taking inspiration from Frege, to defend a

metaphysically neutral version of the theory: holding that truth-bearers are the contents of thoughts, and that facts are simply true thoughts rather than the metaphysically weighty sorts of things envisaged in correspondence theories. That is, the identity is not conceived as a (potentially troublesome) relation between an apparently mind-dependent judgment and an apparently mind-independent fact. A claimed benefit of this version is that it is not immediately disabled by the inevitable accompaniment of an identity theory of falsehood. The difficulty for these attempts is to make out the claim that they involve a *theory* of truth at all, since they lack independent accounts of truth-bearer and truth-maker to give the theory substance. [See Candlish (1995), Dodd and Hornsby (1992), Dodd (1996), Hornsby (1997).]

The most thorough account of this type is found in Dodd (2000). But although this book in its very title proclaims its author's adherence to an identity theory, it actually defends a variety of deflationism: ‘truth is *nothing more than* that whose expression in a language gives that language a device for the formulation of indirect and generalized assertions’ (p. 133, emphasis Dodd's). What became of the identity theory? The answer lies in the fact that Dodd conceives his identity theory as consisting entirely in the denial of correspondence and the identification of facts with true thoughts. It actually has nothing to say about ‘the nature of truth’, as traditionally conceived, offering no definition of ‘is true’, no explanation of what truth consists in or of the difference between truth and falsehood. This theory is ‘modest’, to use Dodd's expression, as opposed to ‘robust’ identity theories which begin from the bipolar recognition of independent conceptions of fact (conceived as truth-maker) and proposition (conceived as truth-bearer) employed in correspondence theories, and then attempt in one way or another to eliminate the apparent gap between them. Dodd's view is that his ‘modest’ theory gets some bite from its opposition to correspondence theories; and he urges (as does Hornsby) that we should anyway scale down our expectations of what a theory of truth can provide. However, the history of identity theories of truth reveals them as tending to mutate into other theories when put under pressure, as one can see from the discussion in the present article. Dodd holds that this is a problem only for robust theories. Yet his theory also exemplifies a variety of this tendency: in the end, it evolves into deflationism.

Can the no-holder charge be met?

Although it is difficult to find a completely uncontroversial attribution of the identity theory, there is evidence of its presence in the thought of a few major philosophers. As one might expect, mystical philosophers attracted by the idea that the world is a unity express views which at least resemble the theory (for example, Plotinus, *The Enneads*: 5th Ennead, 3rd Tractate, §5; 5th Ennead, 5th Tractate, §2). Bradley may also fall into this category; in any case, he and Frege have already been mentioned. Bolzano and Meinong are other possibilities: Findlay, for example, believes Meinong to have held an identity theory, reminding us that on his view, there are no entities between our minds and the facts; facts themselves are true in so far as they are the objects of judgments. [See Findlay (1933), Ch. III sec. ix.] C.A. Baylis defended a similar account of truth in 1948, and Roderick Chisholm endorsed a recognizably Meinongian account in his *Theory of Knowledge*. A sketchy version of the theory is embraced in Woozley's *Theory of Knowledge*. There are also the attempts, once again already mentioned, to establish a metaphysically neutral version: these show that there can be no doubt that some philosophers have tried to defend something that they wished to call an identity theory of truth.

Thomas Baldwin argues that the identity theory of truth, though itself indefensible, has played an influential but subterranean role within philosophy from the nineteenth century onwards, citing as examples philosophers of widely different convictions. [See Baldwin (1991). One of his attributions is queried in Stern (1993), others in Candlish (1995).] Whether or not Baldwin is right -- and it is possible that the theory is no more than an historical curiosity -- the identity theory of truth in its full-blooded form may turn out to be best thought of as comparable to solipsism: rarely, if ever, consciously held, but the inevitable result of thinking out the most extreme consequences of assumptions which philosophers often just take for granted.

[Return to Section Links]

Bibliography

In each case, the date shown immediately after the author's name is the date of original publication. A separate date is shown for the edition cited only where this differs from the original.

- Baldwin, T. (1991), ‘The Identity Theory of Truth’, *Mind* 100, pp. 35-52.
- Baylis, C.A. (1948), ‘Facts, Propositions, Exemplification and Truth’, *Mind*, LVII, pp. 459-79.
- Bolzano, B. (1837), *Wissenschaftslehre* (Leipzig: Felix Meiner 1929), Vol. I, sections 19-33.
- Bradley, F.H. (1907), ‘On Truth and Copying’, *Essays on Truth and Reality* (Oxford: Clarendon Press, 1914), pp. 107-26.
- Candlish, S. (1989), ‘The Truth about F.H. Bradley’, *Mind* 98, pp. 331-48.
- -----. (1995), ‘Resurrecting the Identity Theory of Truth’, *Bradley Studies* 1, pp. 116-24.
- -----. (1996), ‘The Unity of the Proposition and Russell's Theories of Judgment’, in *Bertrand Russell and the Origins of Analytical Philosophy*, ed. Ray Monk and Anthony Palmer (Bristol: Thoemmes).
- -----. (1999), ‘Identifying the Identity Theory of Truth’, *Proceedings of the Aristotelian Society*, XCIC, pp. 233-40.
- -----. (1999), ‘A Prolegomenon to an Identity Theory of Truth’, *Philosophy*, 74, pp. 199-221.
- Cartwright, R. (1987), ‘A Neglected Theory of Truth’, in his *Philosophical Essays* (Cambridge, MA and London: The MIT Press).
- Chisholm, R.M. (1966), *Theory of Knowledge* (Englewood Cliffs, N.J.: Prentice-Hall), Ch. 7.
- Dodd, J. (1995), ‘McDowell and Identity Theories of Truth’, *Analysis* 55, pp.160-5.
- -----. (1996), ‘Resurrecting the Identity Theory of Truth: A Reply to Candlish’, *Bradley Studies* 2, pp. 42-50.
- -----. (1999), ‘Hornsby on the Identity Theory of Truth’, *Proceedings of the Aristotelian Society*, XCIC, pp. 225-32.
- -----. (2000), *An Identity Theory of Truth* (London: Macmillan).
- Dodd, J. and Hornsby, J. (1992), ‘The Identity Theory of Truth: Reply to Baldwin’, *Mind* 101, pp. 319-22.
- Findlay, J.N. (1933), *Meinong's Theory of Objects* (Oxford: Oxford University Press).

- Frege, G. (1918), ‘Thoughts’, in his *Logical Investigations* (Oxford: Blackwell, 1977).
- Hornsby, J. (1997), ‘Truth: The Identity Theory’, *Proceedings of the Aristotelian Society* XCVII, pp. 1-24.
- -----. (1999), ‘The Facts in Question: a Response to Dodd and to Candlish’, *Proceedings of the Aristotelian Society*, XCIC, pp. 241-45.
- Plotinus (301), *The Enneads*, transl. Stephen MacKenna (London: Faber and Faber 1917; 3rd edn revised by B.S. Page, 1962).
- Stern, R. (1993), ‘Did Hegel Hold an Identity Theory of Truth?’, *Mind* 102, pp. 645-47.
- Walker, R.C.S. (1998), ‘Bradley's Theory of Truth’, in *Appearance versus Reality*, ed. Guy Stock (Oxford: Clarendon Press), pp. 93-109.
- Woozley, A.D. (1949), *Theory of Knowledge* (London: Hutchinson), Ch. 7.

[Return to Section Links]

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Bradley, Francis Herbert | facts | Frege, Gottlob | Meinong, Alexius | Moore, George Edward | propositions | Russell, Bertrand | truth: coherence theory of | truth: correspondence theory of | truth: deflationary theory of | truth: revision theory of

Stewart Candlish
candlish@arts.uwa.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 28, 1996
Content last modified: August 6, 2002

(*Reproduced by kind permission of Dr T.J. Winnifrith*)

Francis Herbert Bradley

F.H. Bradley (1846-1924) was the most famous, original and philosophically influential of the British Idealists. These philosophers came to prominence in the closing decades of the nineteenth century, but their effect on British philosophy and society at large -- and, through the positions of power attained by some of their pupils in the institutions of the British Empire, on much of the world -- persisted well into the first half of the twentieth. They stood out amongst their peers in consciously rejecting the tradition of their earlier compatriots, such as Hume and Mill, and responding rather to the work of Kant and Hegel.

It is for his metaphysics that Bradley has become best known. He argued that our everyday conceptions of the world (as well as those more refined ones common among his philosophical predecessors) contain hidden contradictions which appear, fatally, when we try to think out their consequences. In particular, Bradley rejected on these grounds the view that reality can be understood as consisting of many objects existing independently of each other (pluralism) and of our experience of them (realism). Consistently, his own view combined monism -- the claim that reality is one, that there are no real separate things -- with absolute idealism -- the claim that reality consists solely of idea or experience. This vision of the world had a profound effect on the verse of T.S. Eliot, who studied philosophy at Harvard and wrote a Ph.D. thesis on Bradley.

On philosophers, however, Bradley's contributions to moral philosophy and the philosophy of logic were far more influential than his metaphysics. His critical examination of hedonism -- the view that the goal

of morality is the maximization of general pleasure -- was seminal and stands as a permanent contribution to the subject which can still be read with profit today. Some of the doctrines of his logic have become standard and unnoticed assumptions through their acceptance by Bertrand Russell, an acceptance which survived Russell's subsequent repudiation of idealist logic and metaphysics.

Other notable figures among the British Idealists were Bernard Bosanquet, Edward Caird, T.H. Green, Harold Joachim and J.M.E. McTaggart.

Section Links:

Life

Bradley was born on 30th January, 1846 in Clapham (then in the county of Surrey, since absorbed into a much expanded London). He was the fourth child and eldest surviving son of Charles Bradley, a prominent Evangelical preacher, and his second wife, Emma Linton. The family was talented and well connected: George Granville Bradley, a son from the first marriage, was successively Head Master of Marlborough College, Master of University College, Oxford, and Dean of Westminster Abbey; A.C. Bradley, a younger son from the second marriage, taught philosophy at Oxford until 1881, and, after moving to literary studies, held chairs at Liverpool and Glasgow, refused one at Cambridge, and became the most distinguished Shakespearean critic of his day. Charles Bradley's ‘Clapham Sect’ (as this actively evangelical humanitarian group was known at the time) had strong imperial connections, including among its members a Governor-General of Bengal, a Governor of Sierra Leone, several members of Parliament and a permanent head of the Colonial Office.

In 1856 F.H. Bradley's schooling began at Cheltenham College; in 1861 he was transferred to Marlborough College, then under his half-brother's Headship. While at Cheltenham he began learning German; he read at least some of Kant's *Critique of Pure Reason* while still at school, though it is not clear that this was in the original language. In the winter of 1862-3 he contracted typhoid fever (at one

stage expected to kill him), followed shortly by pneumonia. Surviving both, he was protected from further exposure to the rigours of English public school life by leaving Marlborough in 1863.

In 1865 Bradley entered University College, Oxford, as a Scholar, getting a first in classical moderations (Mods) in 1867 but only an unexpected second in *literae humaniores* (Greats) in 1869. A.E. Taylor, a later admirer of Bradley and sympathetic to his idealism, attributed his reverse in Greats to ‘the complete incapacity of examiners whose philosophical scriptures were the writings of John Stuart Mill to comprehend what philosophy meant to the brilliant younger men who were shortly to revolutionize philosophical studies in Great Britain.’ Whether or not this is true, there is certainly an undisguised contempt for Mill and his followers exhibited in Bradley's *Principles of Logic*. After more than one failure to obtain a college fellowship, he was in December 1870 elected to one at Merton College Oxford, tenable for life, with no teaching duties, and terminable only on marriage. He never married, and remained in his fellowship until his death.

In June 1871 Bradley suffered a severe inflammation of the kidneys which appears to have had permanent effects. It has been suggested, possibly with malice, that the Bradleys in general were disposed to hypochondria; be that as it may, he was prone thereafter to be incapacitated by cold, physical exhaustion or anxiety, and in consequence lived a retired life. He took an active part in the running of his college, but avoided public occasions, to the extent, for example, of declining an invitation to become a founding member of the British Academy. Collingwood records of Bradley in his *Autobiography*, ‘[A]lthough I lived within a few hundred yards of him for sixteen years, I never to my knowledge set eyes on him.’ This relative seclusion added an element of mystery to his philosophical reputation, a mystery enhanced by the dedication of some of his books to a person identified only by the initials ‘E.R.’

But although Bradley devoted himself to philosophy, so that the history of his public life is largely that of his books and articles, it is clear that his was not a narrowly bookish existence. To protect his health, he frequently escaped the damp chill of Oxford winters for the kinder weather of southern English and Mediterranean seaside resorts. His metaphysics, a striking combination of the rational and the mystical, makes more than grudging room for the life of the senses and emotions, and his writings, especially his posthumously published *Aphorisms*, could not be the work of a man whose experience had been confined to the study. He liked guns and disliked cats, indulging his preferences economically by using the former to shoot the latter in the college grounds at night.

Bradley's political views are said to have been conservative, though not of a narrowly doctrinaire kind. Although his writings reveal a religious temperament, he seems (judging by a letter of 1922) to have found the evangelical religiosity of his father's household oppressive, and, perhaps in consequence, the attitude to Christianity displayed later in his writings exhibits a certain ambivalence; on the whole, he appears to have been a freethinker. (To imagine growing up amongst the members of the Clapham Sect, we might use John Sutherland's suggestion that the characters of Edmund and Fanny in Jane Austen's *Mansfield Park* give us some idea of what they would have been like.)

Bradley's public recognition included the conferring of the honorary degree LL.D. by the University of

Glasgow (1883), election to membership of the Royal Danish Academy (1921), of the *Accademia dei Lincei* and the *Reale Istituto Lombardo* of Milan (1922), and election to an Honorary Fellowship of the British Academy (1923). In 1924, King George V bestowed on him, the first philosopher to be singled out for this very rare honour, the Order of Merit. Three months later, after a few days' illness, he died from blood poisoning on 18th September, 1924. He is buried in Holywell Cemetery, Oxford.

[Return to Section Links]

Reputation

As the above (by no means complete) account of his public recognition reveals, in his own day Bradley's intellectual reputation stood remarkably high: he was widely held to be the greatest English philosopher of his generation, and although the idealists were never a dominant majority, amongst some philosophers the attitude towards him seems to have been one almost of veneration.

This reputation began to collapse fairly quickly after his death. The reasons for this are complex, and include matters seemingly extraneous to philosophy itself, such as the reaction against British imperialism (whose moral and spiritual mission had been justified by some idealist philosophers and undertaken by their pupils) following the Great War. One more locally significant factor was the tendentious but still damaging accounts of his views which appeared in the writings of Moore and Russell following their defection from the idealist camp. Another was logical positivism: for example, in the first chapter of A.J. Ayer's anti-metaphysical tract *Language, Truth and Logic*, Bradley is presented solely as a metaphysician and, on the basis of a single out-of-context sentence, selected for ridicule. Consequent upon such influences was a change, inimical to idealism, in the whole style of doing philosophy, a change characterized by the development of formal logic and the new respect paid to the deliverances of common sense and of ordinary language. Bradley's highly wrought prose and his confidence in the metaphysician's right to adjudicate on the ultimate truth began to seem alien to a later generation of philosophers reared on a mixture of plain talk and formalization and encouraged to defer to mathematics and empirical science.

Such influences ensured that a misleading and dismissive stereotype of Bradley became current among analytic philosophers and established in their textbooks, so that serious discussion of his work largely disappeared. One result has been that, despite his seminal influence on Russell and their extended controversy over fundamental matters, books and articles on Russell can contain few or even no references to Bradley. Another is that the incidental textbook references to some of Bradley's most characteristic, original and significant views, e.g. on relations and on truth, are often based on hostile caricatures. With a few exceptions (for instance, McTaggart's argument for the unreality of time), discussion of the work of the idealists has been sparse since the nineteen thirties. Discussion of Bradley began to revive, as did his reputation, in the nineteen seventies. At the time of writing it is clear that he is still widely underrated; it is, however, far from clear that his reputation will ever again stand as high as it did in his own lifetime.

[Return to Section Links]

Philosophy of History

Bradley's first substantial contribution to philosophy was the publication in 1874 of his pamphlet ‘The Presuppositions of Critical History’. Although it was not widely noticed at the time, it did have an impact on the thinking of R.G. Collingwood, whose epistemology of history, like Bradley's, evinces a certain scepticism concerning historical facts and the authority of testimony, and it has had a considerable subsequent influence. Bradley's views were inspired by his reading of German biblical critics, and such views have been prominent since in religious studies, where a reluctance to take at face value testimony of the occurrence of miracles which violate the laws of nature is appropriate. But Bradley's attempt to extend this reluctance to historical reports in general underestimates the contrast between the uniformity of nature and the variety of human history.

Although its overall argument cannot be regarded as satisfactory, the pamphlet is nevertheless worth reading both for its historical significance and for its value as a fairly brief introduction to Bradley's thought. Some characteristic later themes, such as the fallibility of individual judgments and the rejection of correspondence accounts of truth, here make an early appearance; and Bradley's philosophical style -- often obscure, typically disdainful of illustrative example, and by late twentieth-century standards uncomfortably literary -- can be seen in high relief.

[Return to Section Links]

Ethics

Bradley's views on ethics were expressed at length in his first widely acknowledged publication, *Ethical Studies* (1876). One reason it was noticed is that the book is highly polemical. (Sidgwick called it ‘vehemently propagandist’ in his *Mind* review.) He did not change these views significantly in later years: in 1893 he described it as ‘a book which, in the main, still expresses my opinions’ (*Appearance and Reality*, p. 356n) and at the time of his death was working on a second edition which, characteristically, was to retain the original text intact but incorporate additional matter.

Bradley says in his Preface that his object is ‘mainly critical’ and that the ethical theory of his time rests on ‘preconceptions metaphysical and psychological’, which are ‘confused or even false’. In this the most Hegelian of his books, his approach is, in a series of connected essays, to work dialectically through these erroneous theories towards a proper understanding of ethics. Accordingly he tells us that the essays ‘must be read in the order in which they stand’, and a corollary of this is that the common practice of extracting one or two of them (usually the brilliantly written ‘Pleasure for Pleasure's Sake’ and ‘My Station and Its Duties’) from the whole, on the basis of their individual merits, can result in a misleading impression of their significance within Bradley's moral thinking: neither represents some finished

position.

The development of this proper understanding begins by examining the ‘vulgar’ notion of moral responsibility and the apparent threats to it posed by the philosophical doctrines of determinism and indeterminism, threats which he argues evaporate once we examine the reality of human action. (A prominent theme in the book is that everyday moral thought is not to be overturned by moral philosophy.) It proceeds by turning to the question ‘Why should I be moral?’, which he answers by suggesting that the moral end for each of us is self-realization. What this is, is then gradually unfolded through examination of representative philosophical theories each of which is rejected as unsatisfactory because of its one-sided concentration upon particular features of the moral life. Nevertheless, he thinks, each theory captures something important which must not be forgotten in the proper understanding he aims at. For example, in the third essay, ‘Pleasure for Pleasure's Sake’, a still-classic critique of hedonistic utilitarianism, Bradley argues that its individualism is insupportable, as is its hedonistic conception of happiness as a pleasurable state identifiable independently of the means by which it is attained (so that it could in principle be achieved more conveniently than through moral behaviour). But purged of these errors, the essential utilitarian insight of the importance of happiness as the point of morality can be retained. Likewise, in the next essay's examination of a Kantian (if not quite Kant's) ethics of duty, he argues that from this conception of morality we should abandon, as the result of a false abstraction, its idea that duty should be done just for duty's sake. We can, however, retain the insight that morality requires the performance of individual *duties*, provided we are clear that their obligatoriness arises from the nature of each duty rather than from some formal principle.

These theories are inadequate because they have a deficient conception of the self, a deficiency he begins to remedy in the fifth essay, the famous ‘My Station and Its Duties’, where he outlines a social conception of the self and of morality with such vigour that it is understandable that the mistaken idea that it expresses his own position has gained some currency. This Hegelian account of the moral life, in which the self is fully realized by fulfilling its role in the social organism which grounds its duties, is clearly one which greatly attracted Bradley, and he seems never to have noticed the implicit tension between the metaphysical account of the self as necessarily social and the moral injunction to realize the self in society. But he finally acknowledges its inadequacy, pointing out, for instance, that any actual society may exhibit moral imperfections requiring reform from the standpoint of an ideal which cannot be exemplified in the roles available within that society. This leads him naturally into the next essay's consideration of ideal morality, where he discusses the scope of morality's demands on the individual, and, by a further natural extension, into the seventh essay's discussion of the distinction between the good and the bad self, a discussion which involves an attempted demonstration that the bad self is a kind of unrealizable parasite on the good. This is necessary to his enterprise: without it, he could not hope to make plausible his suggestion that the aim of morality is self-realization. But in one way the enterprise still founders: the final essay argues that morality is ultimately self-contradictory, depending for its existence on the evil it seeks to overcome. Realization of the ideal self is thus unattainable through morality, but the book closes by suggesting that it is still possible in religion.

Some of Bradley's metaphysical ideas are displayed in his defence of his moral philosophy. An example is his claim that the self is a concrete universal and that the ethical doctrines he criticizes are damaged by

their reliance upon abstract notions of the self. The self is *universal* in that it retains its identity over time and through many different actions, thus collecting together the series of abstract particulars which make up its history in a way analogous to that in which the abstract universal *red* collects together its scattered individual instances (now often called ‘tropes’); it is *concrete* in that, unlike *red* it is a real non-abstract individual. For such claims to be fully convincing, a developed system in which the underlying metaphysical ideas are fully worked out is needed, as he himself admitted. But in this later working out, most of it in *Appearance and Reality*, the expression ‘concrete universal’ almost disappears from Bradley's vocabulary, mainly because he eventually concludes that there can be only one such thing; nevertheless, the idea involved remains, reappearing in the form of the recurring theme that abstraction is falsification, and in this form is central to his logic and his metaphysics.

[Return to Section Links]

Logic

Bradley's most sustained treatment of logic comes in *The Principles of Logic*, published contemporaneously with Frege's *Grundlagen*. The benefit of hindsight provides a striking contrast between these works, the former apparently looking back to the nineteenth century, the latter anticipating the twentieth. While both books eschew formal methods, in Frege's case this results merely from an attempt to give a readable account of some applications of mathematical logic. But the absence of formulae (theorems, axioms, rules of inference) from Bradley's book is intrinsic to it, expressing an opposition (shared by Mill) to the formalization of reasoning *in principle*, as detaching inference from the practical acquisition of scientific knowledge. This, together with the fact that familiar terms (e.g. ‘contradiction’) are used in unfamiliar ways, gives the book an archaic feel. Nevertheless, and despite the fact that *Principles* would no longer ordinarily be consulted by a modern logician unless for historical purposes, it focuses on issues central to logic, and the impression of its being backward-looking is to some extent misleading: for example, it uses the older vocabulary of ‘ideas’ and ‘judgments’ to express views which, often through their (selective) impact upon Russell, gave rise to doctrines subsequently expressed in terms of sentences and propositions; and it effectively exposed the notion of meaning to a sceptical scrutiny which has continued long since.

Although the treatment is less rigidly dialectical than that of *Ethical Studies,* Bradley develops his views through criticism of others, and alters them as he goes along. One result is that the book is far from easy to *consult*, and a reader determined to find out what Bradley thinks must be prepared to follow its argument through many twists and turns.

Traditionally, logic books came divided into three parts, dealing respectively with Conception (usually via *ideas*, the traditional components of judgments), Judgment and Inference. Bradley both inherits and transforms this tradition, keeping the three-part format but devoting the first to Judgment and both second and third parts to Inference, thus dropping the separate treatment of Conception. This is significant in that it reflects his rejection of the standard view that judgments are formed by somehow conjoining ideas: for example, the Port-Royal *Logic*'s Aristotelian claim that they are ‘necessarily

composed of three elements -- the subject-idea, the attribute, and the joining of these two ideas'. Bradley attacks such doctrines on more than one front.

He argues, for instance, that those who, like Hume, think judgments to consist of separable ideas, fail to identify the sense of 'idea' in which ideas are important to logic: ideas in this sense are not separate and datable psychological events (such as my now visualizing a rainbow) but abstract universals. Once ideas are properly understood, he suggests, they can no longer even plausibly be thought of as individual and mutually independent entities which can be put together to create a judgment (as Locke maintains in Chapter XIV of Book IV of *An Essay Concerning Human Understanding*): the order of dependence is the opposite, ideas being abstractions from complete judgments. Here, albeit in his archaic vocabulary, Bradley identifies in advance the difficulties which Russell was later to face in trying to reconcile the unity of the proposition with what he thought to be the mutual independence of its constituents, difficulties which appeared in another guise for Frege in his attempt to maintain a strict division between concepts and objects.

Further, given that ideas are universals, accounts like that of Port-Royal make it impossible to see how judgment can be about reality, since its ideas represent kinds of things, while those real things themselves are particular; so long as judgment is confined to ideas, there can be no unique identification of any item about which we judge. Bradley applies the point to language, arguing that even grammatically proper names and demonstratives are disguised general terms. He thus anticipates that application of Russell's Theory of Descriptions in which it is used to eliminate grammatical names in favour of quantified general sentences. Whether or not this is actually the origin of that theory, there is no doubt in another case: Russell, who claimed in correspondence to have read *Principles* closely, acknowledged openly that he was convinced by Bradley's argument that the logical form of universal sentences is hypothetical (so that, e.g., 'All cows eat grass' is to be understood as saying 'If anything is a cow then it eats grass'). In this way, Bradley had a significant, if indirect, impact on predicate calculus.

Bradley's own account of judgment is that it is 'the act which refers an ideal content ... to a reality beyond the act', so that the logical form of every judgment is 'Reality is such that, if anything is S then it is P'. This formulation makes intelligible what is superficially paradoxical in Bradley, when he says: 'All judgments are categorical, for they all do affirm about the reality, and assert their content of that. Again, all are hypothetical, for not one of them can ascribe to reality its content unconditionally' (*Principles*, Bk I, Ch. II, sec. 79, modified according to Bradley's notes to the Second Edition). It is not hard to see in this an informal anticipation of the representation of sentences in terms of a combination of universal quantifier and object- and predicate-variables. (Here as elsewhere the book looks forward as well as back.) But it is an exaggeration to claim, as some have done, that Bradley's strictures on the account of judgment as a combination of ideas mean that he is straightforwardly opposed to psychologism in logic, for it is clear that he thinks logic's subject matter to be mental acts, not sentences or statements.

Bradley continues to criticize traditional logic when he turns from judgment to inference. Just as he rejected the Aristotelian account of judgments as combinations of subject and predicate, he rejects Aristotelian syllogistic (for the same reason as he later rejects Mill's canons of induction): it misses the fact that reasoning can take place only through the generality involved in universals. Universals are thus

essential to inference, and for this reason Hume's account of inference in terms of the association of ideas collapses: Humean ideas are particulars, fleeting episodes which cannot be revived by association. This does not mean that association of ideas is impossible, but genuine association (which Bradley calls ‘redintegration’) can involve only universals.

Surprisingly for those who subscribe to the common view, first broadcast by Russell in 1900 in *A Critical Exposition of the Philosophy of Leibniz* and much repeated thereafter, that Bradley thought all judgments to be of subject/predicate form and accordingly failed to recognize relational judgments as a distinct kind, Bradley's treatment of inference includes the complaint that the mathematical logics of his time cannot represent valid relational inferences. His own initial account of inference is that it is ‘ideal experiment’: ‘ideal’ in that these are thought-experiments which remain in the realm of idea, but nevertheless experiments in that their results are not guaranteed in advance by a complete set of logical laws which infallibly determine their own application (a view reminiscent of Wittgenstein). But later, after a long and tangled consideration of the question of how it is possible for a deductive inference to be reflected in reality, he comes up with a revised account: ‘Every inference is the ideal self-development of an object taken as real’ (*Principles*, Terminal Essay I, p. 598). Bradley seems here to be following the Humean idea that there are no logical relations between distinct existences: the reason that valid inference can be reflected in reality is that it can never take one beyond the original subject matter.

Much of *The Principles of Logic* is polemical, and it affords occasional examples of Bradley at his funniest and most acerbic, such as this note to a short chapter criticizing Herbert Spencer's view of the nature of inference (Bk II, Pt II, Ch. II, sec. 14, n. 3),

> With regard to Mr. Spencer's view I would suggest, as a possibility, that it never was taken from the facts, but was a development of or from something about Comparison which he found in Hamilton. Reading so few books, Mr. Spencer was naturally more at the mercy of those he did read.

and this passing swipe at Hamilton himself (Bk II, Pt II, Ch. I, sec. 9),

> This may be called the law of Redintegration. For we may take this name from Sir W. Hamilton (*Reid*, p. 897), having found nothing else that we could well take.

It is clear that much of Bradley's criticism of his predecessors and contemporaries expresses his hostility to the sort of psychological atomism evident in extreme form in Hume but equally to be found presupposed in accounts of judgment like those mentioned above. What Bradley particularly objected to about such views is that the particulars (ideas) which they treated as realities in their own right, and out of which judgments are said to be composed, are anything but: far from being themselves genuine individuals, they are abstractions from the continuous whole of psychological life and incapable of independent existence. This is an early version of a holism which has since had many adherents. But he then goes on to point out that judgments too involve abstractions, since the subject matter of any judgment is necessarily detached from its background (as, for example, ‘Julius Caesar crossed the

Rubicon' detaches the river from its location and the general from his army) and this process inevitably misrepresents the way things really are. Thus the objections which Bradley deployed against misleading accounts of logic now begin to pose a threat against logic itself by eroding the integrity of the judgments which go into its inferences, and he ends *Principles* appropriately by suggesting that no judgment is ever really true nor any inference fully valid.

At this point Bradley's attempt to write a book on logic without getting entangled in metaphysics begins to succumb to his doubts about the notion of truth. He holds that logic presupposes a correspondence theory of truth (he calls it the 'copy' theory), but it is apparent that he thinks this theory metaphysically inadequate: indeed, he marshals against it counter-examples drawing on, e.g., disjunctions, counter-examples which had to await the theory of truth-functions before they could be accommodated. In *Essays on Truth and Reality* he takes these ideas further, arguing for 'the identity of truth knowledge and reality'. It could hardly be more clear that Bradley holds an identity theory of truth, and although he is commonly believed to have been a supporter of a coherence theory of truth (and is standardly identified as such in the textbooks), this common belief is at the very least greatly misleading. However, the combination of the identity theory and his metaphysical doctrine that reality is a unified whole enables coherence to be deduced from his views as a consequence, and he himself thought the *test* of truth to be 'system', a notion under which he included what is commonly meant by coherence; this explains why he has so often been thought to be a coherence theorist. It might be thought that his famous attack on the Hegelian idea that the rational is the real (*Principles* Bk III, Pt II, Ch. IV, sec. 16) is inconsistent with his holding an identity theory of truth: but the two are reconciled through his doctrine of degrees of truth, a doctrine which has to be understood within the context of his metaphysics.

[Return to Section Links]

Metaphysics

After the completion of *The Principles of Logic*, Bradley turned to the task of giving a full account of his metaphysics. The result was *Appearance and Reality* (1893). But Bradley was philosophically active for a further thirty years thereafter, continuing to elucidate, defend and refine his views, and engaging with critics and rivals (notably, and revealingly for both sides, with Russell). Concentration upon *Appearance and Reality* alone, therefore, risks placing undue weight upon what turn out to be temporary features of thought or expression, and this has in fact contributed to the distorted impressions of his thinking so often to be found in the textbooks of analytic philosophy.

Appearance and Reality is divided into two books. The first, 'Appearance', is brief, and its aim destructive, arguing that 'the ideas by which we try to understand the universe' all bring us ultimately to contradictions when we try to think out their implications. Some of these ideas belong especially to philosophy, such as the view that only the primary qualities are real; others, for instance the notions of cause, motion, self, space, thing and time, are deployed in everyday life. The second book, 'Reality', is long; its aim is to provide a positive account of the Absolute -- the ultimate, unconditioned reality as it is in itself, not distorted by projection through the conceptual mechanisms of thought. A large proportion of

his discussion is devoted to consideration of natural objections to this positive account.

Much of Book I involves presentation of familiar suggestions which make only part of Bradley's case: he alleges, for example, that motion involves paradoxes, and that primary qualities alone cannot give us reality, for they are inconceivable without secondary qualities. But Chapter III, entitled ‘Relation and Quality’, is uniquely Bradleian, alarming in the breadth of its implications, and has caused intermittent controversy ever since. In generalized form, its contention is that relations (such as *greater than*) are unintelligible either with or without terms, and, likewise, terms unintelligible either with or without relations. Bradley himself says of the arguments he wields in support of this contention,

> The reader who has followed and has grasped the principle of this chapter, will have little need to spend his time on those which succeed it. He will have seen that our experience, where relational, is not true; and he will have condemned, almost without a hearing, the great mass of phenomena.

And it is clear that his views on relations are both highly controversial and central to his thought.

In view of this, it was a serious tactical error on Bradley's part to present his arguments so sketchily and unconvincingly that even sympathetic commentators have not found it easy to defend him, while C.D. Broad was able to say later, ‘Charity bids us avert our eyes from the pitiable spectacle of a great philosopher using an argument which would disgrace a child or a savage.’

The impression that Bradley's crucial metaphysical arguments are negligible arises in part from reading them as designed to prove the doctrine of the internality of all relations (i.e., their reducibility to qualities, or their holding necessarily, depending on the sense of ‘internal’, Russell having interpreted the doctrine in the former way, Moore in the latter). Whichever sense we take, this is a misreading -- and an impossible one, if we take ‘internal’ in Russell's sense, because of Bradley's rejection of the subject/predicate account of judgment. If, however, we use Moore's sense of ‘internal’, the reading is understandable: in Chapter III Bradley confusingly applies this word to relations in a metaphysically innocent way which has no connection with the doctrine of internality, without drawing attention to this fact; while in other parts of *Appearance and Reality* he openly flirts with the doctrine of internality, repudiating it clearly only in later works less often read, such as the important essay ‘Relations’ left incomplete at his death. Further, Bradley does uniformly reject the reality of external relations, and it is natural, though not logically inevitable, to interpret this as a commitment to the doctrine of internality.

His considered view, though, is that neither external nor internal relations, nor yet their terms, are real; and that is the proper conclusion of his arguments in the chapter in question, arguments which he deploys as a team, systematically excluding the possible positions available to those who would disagree. The member of this team which has attracted the greatest attention is the one which alleges that if a relation were a further kind of real thing along with its terms (as, e.g., Russell later assumed in his multiple relation theory of judgment), then a further relation would be required to relate it to its terms, and so on *ad infinitum*. It is clear from this argument (which is an obvious descendant of *The Principles of Logic*'s

attack on the traditional analysis of judgment), as well as from his own explanation, that for him ‘real’ is a technical term: to be real is to be an individual substance (in the sense commonly found in Descartes, Leibniz and Spinoza), so that to deny the reality of relations is to deny that they are independent existents. It is this which explains reactions like Broad's: in common with others, he took Bradley to be assuming that relations are a kind of object, when what Bradley was doing was arguing by a kind of *reductio* against that very assumption.

Some, however, have thought that the denial of the reality of relations amounts to the assertion that all relational judgments are false, so that it is, for example, not true that 7 is greater than 3 or that hydrogen is lighter than oxygen. Such an interpretation is made credible by Bradley's account of truth, for on that account no ordinary judgment is ever perfectly true; in consequence, to one who reads him under the influence of the later but anachronistic assumption that truth is two-valued, his claim appears to be that relational judgments are all false. On Bradley's account of truth, however, while for ordinary purposes it is true that 7 is greater than 3 and false that oxygen is lighter than hydrogen, once we try to meet the more exacting demands of metaphysics we are forced to recognize that truth admits degrees and that, while the former is undoubtedly more true than the latter, it is not fully true. The imperfection of even the more true of these judgments, though, is nothing to do with the its being relational rather than predicative. For, as was observed above in the section on Logic, Bradley thought all judgments to be defective in that representation can proceed only on the basis of separating in thought what is not separate in reality: when, for example, we say ‘These apples are hard and sour’, we not only implicitly abstract the apples from their container but detach the hardness and sourness from each other and abstract them from the apples themselves. A perfect truth, one completely faithful to reality, would thus have to be one which did not abstract from reality at all; and this means that it would have to be identical with the whole of reality and accordingly no longer even a judgment. The final truth about reality is, on Bradley's view, quite literally and in principle inexpressible.

It is, however, possible to give an outline. The impression of reality's consisting of a multiplicity of related objects is a result of the separations imposed by thought; in fact ‘the Absolute is not many; there are no independent reals.’ (All quotations from here on are from *Appearance and Reality*, Ch. XIV.) Reality is one -- but one what? Experience, he says, in a wide sense of the term: ‘Feeling, thought and volition (any groups under which we class psychological phenomena) are all the material of existence, and there is no other material, actual or even possible.’ The immediate argument he gives for this unintuitive doctrine is brief to the point of offhandedness, merely challenging the reader to think otherwise without self-contradiction; his greater concern is to make it quite clear that this experience does not belong to any individual mind, and his doctrine not a form of solipsism. But he is not quite as offhand as he appears, for he soon makes clear that he thinks the whole book to be a best-explanation argument for this objective (or absolute) idealism: ‘This conclusion will, I trust, at the end of my work bring more conviction to the reader; for we shall find that it is the one view which will harmonize all facts.’

So ‘the Absolute is one system, and ... its contents are nothing but sentient experience. It will hence be a single and all-inclusive experience, which embraces every partial diversity in concord. For it cannot be less than appearance, and hence no feeling or thought, of any kind, can fall outside its limits.’ But how

can we understand this diversity to be possible, when it cannot be accounted for through terms and relations? Bradley's answer is that we cannot understand this in detail, but can get some grasp on what he means by considering a pre-conceptual state of immediate experience in which there are differences but no separations, a state from which our familiar, cognitive, adult human consciousness arises by imposing conceptual distinctions upon the differences. Reality is like this primitive state, but not exactly like, for it transcends thought rather than falls short of it, and everything, even conceptual thought itself, is included in one comprehensive and harmonious whole. Appearances thus contribute to Reality in a fashion analogous to the ways in which segments of a painting contribute to the whole work of art: detached from their background, they would lose their significance and might in isolation even be ugly; in context, they can themselves be beautiful and make an essential contribution to the beauty and integrity of the whole. Such limited comparisons are all the help we can get in understanding the Absolute and its relation to its appearances: Bradley rejects as impossible the demand for detailed explanations of how phenomena like error and evil belong to the Absolute, instead trying to shift the burden of proof to critics who express confidence in their incompatibility. His general answer is that anything that exists, even the worst of evils, is somehow real: the Absolute must comprehend both evil and good. But, just as truth admits of degrees, a judgment being less true the further it is from comprehending the whole of reality, so (consistent with ‘the identity of truth knowledge and reality’) reality itself admits of degrees, a phenomenon being the less real the more it is just a fragmentary aspect of the whole. The Absolute is in such a way further from evil than from good, but is itself neither, transcending them both as it transcends even religion -- it is in a sense a Supreme Being, but not a personal God.

In Bradley's often rhapsodic descriptions of the Absolute, a conception of the world based both on his sceptical scrutiny of the inadequacies of philosophers' accounts of judgment and, it is clear, on a kind of personal experience of a higher unity which in another context might have made him one of the world's revered religious mystics, we can see why, at the start of this article, his metaphysics was described as ‘a striking combination of the rational and the mystical’. The very idiosyncrasy of this combination has meant that few philosophers have been convinced by it. Nevertheless, in its bold and direct confrontation of what he called ‘the great problem of the relation between Thought and Reality’, it stands in Western philosophy as a permanent and unsettling challenge to the capacity of discursive thought to display the world without distortion; unsettling because it arises, not from the imposition of an external standard which could be rejected as arbitrary or inappropriate, but from the demand that our mechanisms of representation meet the standards they themselves implicitly set.

[Return to Section Links]

Bibliography

Works by Bradley

- *Ethical Studies* (London: Oxford University Press, 1876; second edition, with notes: London: Oxford University Press, 1927).

- *The Principles of Logic* (London: Oxford University Press, 1883; second edition, revised, with commentary and terminal essays, London: Oxford University Press, 1922; corrected impression, 1928).
- *Appearance and Reality* (London: Swan Sonnenschein, 1893; second edition, with an appendix, London: Swan Sonnenschein, 1897; ninth impression, corrected, Oxford: Clarendon Press, 1930).
- *Essays on Truth and Reality* (Oxford: Clarendon Press, 1914).
- *Aphorisms* (Oxford: privately printed at the Clarendon Press, 1930).
- *Collected Essays* (Oxford: Clarendon Press, 1935).
- *Writings on Logic and Metaphysics* edited and with introductions by James W. Allard and Guy Stock (Oxford: Clarendon Press, 1994).
- *The Collected Works of F.H. Bradley*, 12 volumes, edited and introduced by W.J. Mander and Carol A. Keene (Bristol: Thoemmes, 1999).

The more recent of the editions produced in Bradley's lifetime are the ones now usually cited and the most useful: while the earlier text is left intact, Bradley's later thoughts are added in the form of notes, appendices and essays, enabling the reader to trace the changes in his ideas. (Such additional material is particularly extensive in the *Logic*, where Bradley frequently defers to Bosanquet's criticisms of the first edition.) *Collected Essays* contains the two pamphlets ‘The Presuppositions of Critical History’ (1874) and ‘Mr Sidgwick's Hedonism’ (1877) as well as the valuable unfinished essay on relations (1923-4) and a good bibliography. Between them, this book and the important *Essays on Truth and Reality* contain all his articles of any substance; these are the versions normally cited. *Aphorisms*, after many years out of print, appeared in 1993 (bound together with ‘Presuppositions of Critical History’ and an introduction by Guy Stock) in a facsimile edition (Bristol: Thoemmes Press). Bradley's unpublished papers, notebooks and letters received are in the library of Merton College, Oxford. Correspondence between Bradley and Russell is in the Russell Archives at McMaster University; interesting extracts appear on pp. 349-353 of Volume 6 of *The Collected Papers of Bertrand Russell* (London: Routledge 1992). The John Rylands Library of the University of Manchester has letters from Bradley to Samuel Alexander. Much previously unpublished material has been made available in the 1999 *Collected Works*.

Other Authors

- Basile, P. (1999) *Experience and Relations: An Examination of F.H. Bradley's Conception of Reality* (Berne: Paul Haupt).
- Bradley, J., ed., (1996) *Philosophy after F.H. Bradley* (Bristol: Thoemmes).
- Campbell, C.A. (1931) *Scepticism and Construction: Bradley's Sceptical Principle as the Basis of Constructive Philosophy* (London: George Allen and Unwin Ltd).
- Candlish, S. (1978) ‘Bradley on My Station and Its Duties’, *Australasian Journal of Philosophy*, 56 (2): pp. 155-70.
- Candlish, S. (1989) ‘The Truth about F.H. Bradley’, *Mind*, 98 (391): pp. 331-48.
- Candlish, S. (1996) ‘The Unity of the Proposition and Russell's Theories of Judgement’, in Monk, R. and Palmer, A., eds, *Bertrand Russell and the Origins of Analytic Philosophy* (Bristol: Thoemmes).

- Coady, C.A.J. (1992) *Testimony* (Oxford: Clarendon Press).
- Eliot, T.S. (1916) *Knowledge and Experience in the Philosophy of F.H. Bradley* (London: Faber, 1964).
- Gaskin, R. (1995) ‘Bradley's Regress, the Copula and the Unity of the Proposition’, *The Philosophical Quarterly*, 45 (179): pp. 161-80.
- Horstmann, R.-P. (1984) *Ontologie und Relationen* (Koenigstein: Athenaeum).
- Hylton, P. (1990) *Russell, Idealism, and the Emergence of Analytic Philosophy* (Oxford: Clarendon Press).
- Ingardia, R., ed., (1991) *Bradley: A Research Bibliography* (Bowling Green, Ohio: Philosophy Documentation Center). [Warning: This volume contains many errors, mostly trivial; e.g., many of the articles attributed to Cresswell are by Crossley.]
- MacEwen, P., ed., (1996) *Ethics, Metaphysics and Religion in the Thought of F.H. Bradley* (Edwin Mellen).
- Mander, W. (1994) *An Introduction to Bradley's Metaphysics* (Oxford: Clarendon Press).
- Mander, W. (1995) ‘Bradley's Philosophy of Religion’, *Religious Studies*, 31 (3): pp. 285-301.
- Mander, W., ed., (1996) *Perspectives on the Logic and Metaphysics of F.H. Bradley* (Bristol: Thoemmes).
- Manser, A. (1983) *Bradley's Logic* (Oxford: Blackwell).
- Manser, A. and Stock, G., eds, (1984) *The Philosophy of F.H. Bradley* (Oxford: Clarendon Press, reprinted in paperback 1986).
- Nicholson, P. (1990) *The Political Philosophy of the British Idealists: Selected Studies* (Cambridge: Cambridge University Press), ‘Study I’.
- Passmore, J. (1969) ‘Russell and Bradley’, in Brown, R. and Rollins, C.D., eds, *Contemporary Philosophy in Australia* (London: George Allen and Unwin).
- Sprigge, Timothy (1993) *James and Bradley: American Truth and British Reality* (Chicago & La Salle, Illinois: Open Court).
- Stock, G., ed., (1998) *Appearance versus Reality* (Oxford: Clarendon Press).
- Taylor, A.E. (1924-5) ‘Francis Herbert Bradley, 1846-1924’, *Proceedings of the British Academy*, xi (2): pp. 458-468.
- Wollheim, R. (1956) ‘F.H. Bradley’, in Ayer, A.J. et al., *The Revolution in Philosophy* (London: Macmillan), pp. 12-25.
- Wollheim, R. (1969) *F.H. Bradley* (Harmondsworth: Penguin), second edition.

There is also a journal, *Bradley Studies*, which (in its own words) "aims to publish critical and scholarly articles on philosophical issues arising from Bradley's writings and from those of related authors [and] to include each year an ongoing list of what has been published on Bradley and related themes." The journal is distributed to all members of the Bradley Society as a part of their annual membership, but may also be bought separately by individuals and institutions. Enquiries about the journal should be directed to its Editor William Mander.

[Return to Section Links]

Other Internet Resources

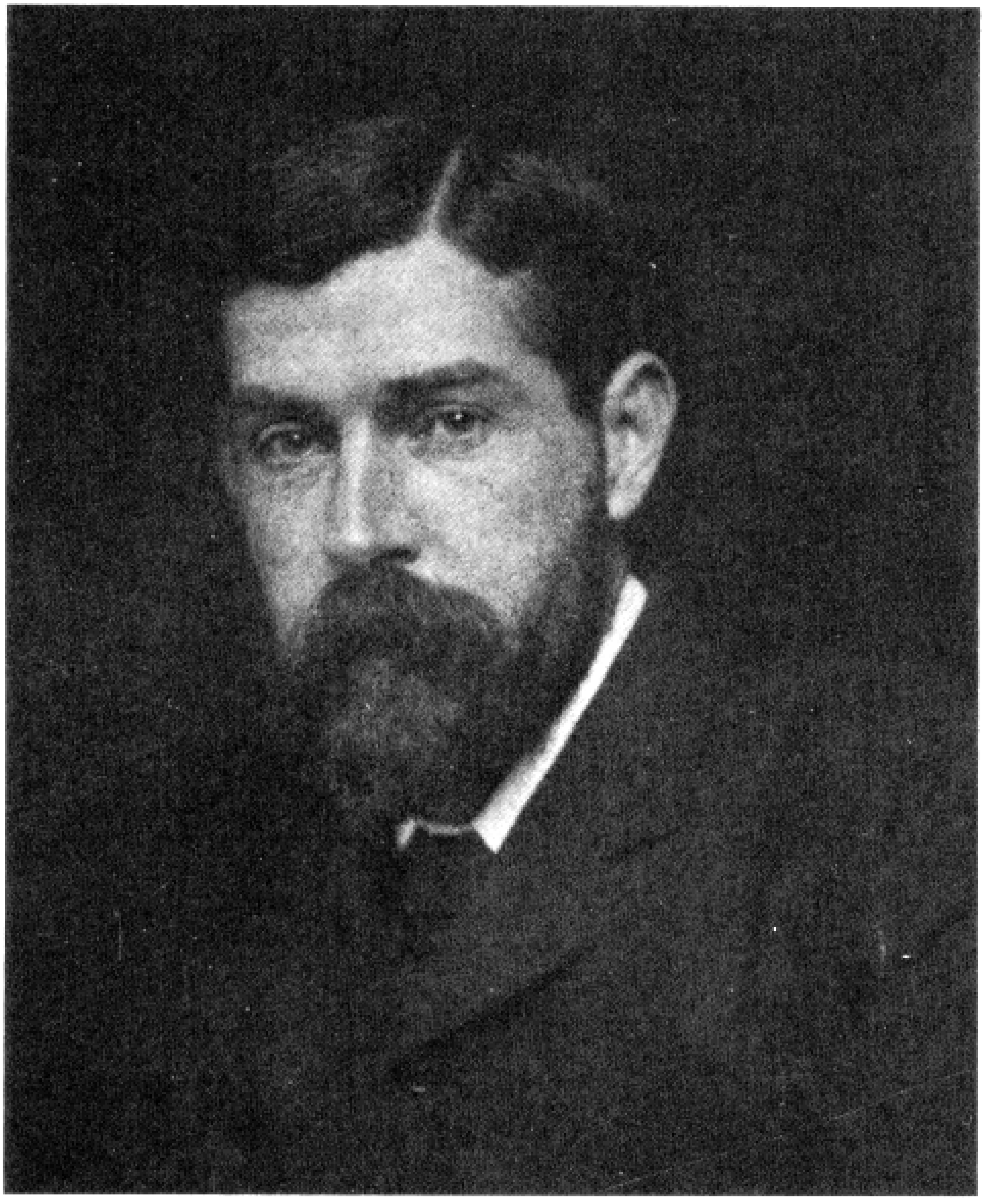
- The publisher's description of the Thoemmes edition of the *Collected Works* (with links to the editors' useful introductions to the volumes.

Related Entries

Bosanquet, Bernard | Frege, Gottlob | Moore, George Edward | Russell, Bertrand | truth: identity theory of

Stewart Candlish
candlish@arts.uwa.edu.au

First published: May 9, 1996
Content last modified: August 2, 2002



Bernard Bosanquet

Bernard Bosanquet (1848-1923), British philosopher, political theorist and social reformer, was one of the principal exponents (with F.H. Bradley) of late nineteenth and early twentieth century 'Absolute Idealism.'

Life

Bernard Bosanquet was born on July 14, 1848 in Rock Hall (near Alnwick), Northumberland, England. He was the youngest of five sons of the Reverend Robert William Bosanquet by the latter's second wife, Caroline (MacDowall). Bernard's eldest brother, Charles, was one of the founders of the Charity Organization Society and its first Secretary. Another brother, Day, was an Admiral in the Royal Navy and served as Governor of South Australia. Yet another, Holford, was elected to the Royal Society and was a fellow of St John's College, Oxford.

Bosanquet studied at Harrow (1862-1867) and at Balliol College, Oxford (1867-1870), where he fell

under the influence of idealist ‘German’ philosophy, principally through the work of Edward Caird and T.H. Green. (Green described him as "the most gifted man of his generation.") Bosanquet received first class honors in classical moderations (1868) and *literae humaniores* (1870) and, upon graduation, was elected to a Fellowship of University College, Oxford, over F.H. Bradley. While at University College, Bosanquet taught the history of logic and the history of moral philosophy; his only published work during this time was a translation of G.F. Schoemann's *Athenian Constitutional History*.

Upon receipt of a small inheritance in 1881, Bosanquet left Oxford for London, where he became active in adult education and social work through such organizations as the London Ethical Society (founded 1886), the Charity Organisation Society, and the short-lived London School of Ethics and Social Philosophy (1897-1900). During this time he met and married (in 1895) Helen Dendy, an activist in social work and social reform, who was to be a leading figure in the Royal Commission on the Poor Laws (1905-1909).

While in London, Bosanquet was also able to engage in philosophical work, and many of his major publications date from this time. Some of them--such as *The Philosophical Theory of the State* and *Psychology of the Moral Self*--were developed from lectures that he gave to adult education groups. He was an early member of the Aristotelian Society, and served as its Vice President (1888) and President (1894-1898).

At the age of 55, Bosanquet briefly returned to professorial life, as Professor of Moral Philosophy at the University of St Andrews in Scotland (1903-1908), but his health was not good and he wished to devote more time to original philosophical writing. He retired to Oxshott, Surrey, where he nevertheless remained active in social work and philosophical circles. In 1911 and 1912, Bosanquet was elected Gifford Lecturer in the University of Edinburgh. The text of these lectures--*The Principle of Individuality and Value* and *The Value and Destiny of the Individual*--serve as the most developed statement of his metaphysical views. It is important for a proper understanding of Bosanquet's philosophy that one recognize that the elaboration of his metaphysics came some time *after* his work in ethics, social work and political philosophy.

The publication of the Gifford lectures incited a good deal of critical reaction to Bosanquet's views, particularly in metaphysics (e.g., on the ‘idealism/materialism’ controversy and on the nature of finite individuality), logic (e.g., concerning the status of propositions and the nature of inference), and ethics. Despite his vigorous participation in such exchanges, present throughout Bosanquet's work is his desire to find common ground among philosophers of various traditions and to show relationships among different schools of thought, rather than to dwell on what separates them.

In spite of the challenges to idealism from both within and outside of the academic world, discussion of Bosanquet's work continued through the early decades of the 20th century. He died in his 75th year in London on February 8, 1923.

At the time of his death, Bosanquet was arguably "the most popular and the most influential of the

English idealists" (J.H. Randall). He was the author or editor of more than 20 books and some 150 articles. The breadth of his philosophical interests is obvious from the range of topics treated in his books and essays--logic, aesthetics, epistemology, social and public policy, psychology, metaphysics, ethics and political philosophy. For his contributions to philosophy and to social work, he had been made a Fellow of the British Academy in 1907, and had received honorary degrees from Glasgow, Birmingham, Durham, and St Andrews.

Bosanquet was one of the earliest philosophers in the Anglo-American world to appreciate the work of Edmund Husserl, Benedetto Croce, Giovanni Gentile and Emile Durkheim, and the relation of his thought to that of Ludwig Wittgenstein, G.E. Moore and Bertrand Russell is significant, though still largely unexplored. Although F.H. Bradley is today far better known in philosophical circles, in his obituary in the *Times*, Bosanquet was said to have been "the central figure of British philosophy for an entire generation."

General Background

Bosanquet's philosophical views were in many ways a reaction to 19th century Anglo-American empiricism and materialism (e.g., that of Jeremy Bentham, John Stuart Mill and Alexander Bain), but also to that of contemporary personalistic idealism (e.g. that of Andrew Seth Pringle-Pattison, James Ward, Hastings Rashdall, W.R. Sorley, and J.M.E. McTaggart) and organicism (e.g. Herbert Spencer). Bosanquet held that the inspiration of many of his ideas could be found in Hegel, Kant, and Rousseau and, ultimately, in classical Greek thought. Indeed, while at the beginning of his philosophical career Bosanquet described Kant and Hegel as "the great masters who ‘sketched the plan’," he said that the most important influence on him was that of Plato. The result was a brand of idealist philosophical thought that combined the Anglo-Saxon penchant for empirical study with a vocabulary and conceptual apparatus borrowed from the continent. Bosanquet is generally considered to be one of the most ‘Hegelian’ of the British Idealists, though the extent to which the term ‘Hegelian’ is appropriate or illuminating in describing his work has been a matter of some recent debate.

More directly, Bosanquet's thought shows a number of similarities to that of T.H. Green, his teacher, and to Bradley, his contemporary. Bosanquet himself acknowledges that these similarities are far from coincidental. He frequently admits his debt to Green's works and, as late as 1920, he wrote that "since the appearance of Ethical Studies... I have recognized [Bradley] as my master; and there is never, I think, any more than a verbal difference or difference of emphasis, between us".

There is, however, at least some hyperbole in such comments. Bosanquet did not follow either Green or Bradley blindly, and there are important differences in his work. While he defended Green's ethical theory and many of Green's conclusions, he addressed a number of issues never dealt with in Green's corpus. Moreover, while it is clear that Bosanquet considered Bradley's work in metaphysics and ethics to have been momentous, this admiration was no doubt influenced by the fact that Bradley's philosophy and method reflected interests and an approach that Bosanquet had arrived at quite independently.

Principal Contributions

Religion

Bosanquet's philosophical views on religion were in large part influenced by early nineteenth century biblical studies--initially, mediated through the writings of his Oxford tutors, Edward Caird and Benjamin Jowett.

The work of David Strauss, Ferdinand Baur, and others, at the beginning of the nineteenth century, marked a turn in the scholarly approach to religion and scripture, towards what is now called ‘the scientific study of religion.’ Religious experience, sacred texts, and religious practice were now to be seen as phenomena open to critical investigation and which could--and should--be examined independently of one's religious commitment, and according to the principles of literary and historical analysis. Strauss and his followers challenged the conflation of religious dogmas and creeds with original religious experience, and they were particularly doubtful whether one could recover much knowledge of such experience from ‘events’ recorded in scripture.

By the mid-nineteenth century, this approach to the study of religion had established itself in Britain, particularly in Oxford. Figures such as Jowett and Caird, and others in the Church of England ‘Broad Church movement’ (such as Frederick Temple, Bishop J.W. Colenso, and Thomas Arnold) argued for a more analytical and ‘rational’ approach to understanding religious belief--though they were frequently criticised for this by Church authorities.

The distinction of practice from dogma and experience from creeds was, however, also a feature of the evangelical movement within the Church of England. Bosanquet, like many of his fellow idealists, was raised in an Evangelical household; his later philosophical views, then, can be seen as a continuation, rather than an interruption or contradiction of, his early religious convictions.

Despite his conventional religious upbringing, Bosanquet was not an orthodox Christian. While he did claim that religion was not only central to one's life, but was that which made life worth living, he held that, taken literally or at face value, many particular religious beliefs are either incoherent or false. Bosanquet notes that, in religion, "rationalism, curiosity, metaphor, and deduction from metaphor, operate by way of distortion" (*What Religion Is*, p. 68), and that, to help one read biblical texts, one must engage in a hermeneutical enterprise, and ‘learn to interpret’ them--though, even here, he doubted whether ‘the sacred books of a Church can ever be understood in their actual meaning.’ Moreover, some religious beliefs do not mean what many take them to mean. Bosanquet argues, for example, that, if we examine the idea of God--who is often described as an ‘infinite individual--we will find that to attribute ‘infinity’ to a being would be inconsistent with "every predicate which we attach to personality." Finally, Bosanquet held that religious belief in general is not about some supernatural being or transcendent realm, entering into our daily lives. It focuses, rather, on what takes place in the world. His analysis of religion and religious belief is, then, ‘immanentist.’

Bosanquet distinguished religious beliefs about particular persons or events from ‘religion’ (or, what was the same thing for him, ‘religious belief as a whole’ or ‘religious consciousness’). Still, he did not see himself as either an agnostic or atheist, or as reducing ‘religion’ to the ‘ethical’. While he states that there is much in Christianity that is no longer intelligible, he insists that religion--in the sense of religious consciousness--is needed for morality, and that an ethics cut off from religion is "without sap or life." Similarly, Bosanquet's opposition to seeing religion or religious belief as a faith in something supernatural does not mean that he denied the existence of the spiritual or held a ‘reductionist’ view of reality. When it comes to human consciousness, he argued, the spiritual--the awareness of the infinite in our world--is at least as much a part of what exists as the material. This ‘infinite’ here is what Bosanquet called the ‘Absolute’.

Human beings are, Bosanquet noted, aware of something infinite that bears directly on their lives, and in his entry on ‘Philosophy of Religion,’ for J.M. Baldwin's *Dictionary of Philosophy and Psychology* (1902), he writes that it is this awareness, and one's commitment to "that set of objects, habits, and convictions, whatever it might prove to be, which [one] would rather die for than abandon, or at least would feel himself excommunicated from humanity if he did abandon" that constitutes what religion is. (While some idealists, such as Pringle-Pattison, seem to have held that such an Absolute is God, Bosanquet did not--though neither does he explicitly reject the existence of God.) Still, religious belief is neither the same as, nor tied essentially to, rituals and practices. Neither does it require adherence or assent to a set of propositions or dogmas--and certainly not to a set of propositions focusing on beings or events in the history of a community of believers. Religious belief is, in short, quite distinct from ‘theism.’

While one finds religious belief and religious consciousness throughout history and throughout the world, Bosanquet rejects the view that all religions are on a par. Religious consciousness has evolved and higher forms of religion--i.e., those which show a *unity* of the Divine and human nature--are the more ‘true.’ What Bosanquet is ultimately interested in, then, is religion in its highest or most developed form--what Caird called ‘Absolute Religion.’ Though Bosanquet does not develop what, specifically, this means, his Gifford lectures give some hint as to the direction of his thought.

Despite his criticisms of, and challenges to, Christianity, Bosanquet believed that the world had benefitted from Christian civilisation and culture, and that Christianity was a progress over ‘earlier’ stages of religion. Moreover, he not only frequently employs allusions to Christian religious belief and practice to illustrate his general views, but retains elements from Christian doctrine, such as the ideas of the atonement and of justification by faith--though in a highly modified form. The doctrines of the atonement (to which Bosanquet often referred, using the words of Goethe, as ‘dying in order to live’) and of ‘justification by faith’ (which emphasised the presence of religious consciousness in ‘works’) have a practical rather than a theological significance. The former reflected the notion of ‘self-sacrifice,’ involved in the achievement of self-realisation--where one had to ‘die’ to the desires of one's ‘private will’ in order to ‘live’ as a more complete moral agent. And the latter doctrine was a reminder that one's actions could have a moral and spiritual character only so far as they were carried out, out of a set of dominant ideas to which one was committed.

Bosanquet holds that religion is reasonable, and that any rational person would be religious. He insists that religious belief as a whole is not superstition, and that it is true so far as it is an expression of a ‘nisus to totality’ or a ‘move to wholeness.’ Again, since particular religious beliefs purport to be cognitive, they must, at least in principle, be able to be known by believers and non-believers alike. (He is, however, sceptical about the relevance of traditional apologetics.) In both cases, the standard that Bosanquet employs in order to assess truth in religion is the same as that which he uses to assess the truth in general--namely, coherence.

Though Bosanquet's analysis of religious belief reflects an understanding that, broadly speaking, was shared by a number of his fellow idealists, it is significantly different from other late 19th century and early 20th century perspectives, such as those of William Clifford, John Henry Newman, and William James, and can be seen as an alternative to them. Given its immanentist character and insistence on separating religion from dogma and theology, it is close to the view of religion that one finds in recent work by R.B. Braithwaite, R.M. Hare, W. Cantwell Smith, D.Z. Phillips, and Hendrik Hart, and there is some similarity to the contemporary ‘Sea of Faith’ movement, advanced by the Anglican theologian, Don Cupitt. Bosanquet's views, like those of these authors, have been challenged (for example, by C.C.J. Webb, François Houang, and Alan P.F. Sell) for not only being inconsistent with any orthodox theism, but as presenting in its stead a ‘generic religion’ (which, some critics hold, is not religion at all). It is, however, important to recognise that Bosanquet is not advancing a non-cognitivist or fideist view of religion, and that he maintains that both religious beliefs as a whole and particular religious beliefs must meet appropriate ‘rational’ standards.

Social and Political Philosophy

Bosanquet's social and political philosophy is called ‘idealist’ because of his view that social relations and institutions were not ultimately material phenomena, but best understood as existing at the level of human consciousness. Writing largely in reaction to the utilitarianism of Bentham and Mill and to the natural-rights based theory of Herbert Spencer, Bosanquet's views show both a strong influence of Hegel and an important debt to Kant and to the classical Greek thought of Plato and Aristotle. Indeed, Bosanquet often spoke of his political theory as reflecting principles found in ‘classical philosophy,’ and one of his early works was a commentary on Plato's Republic. Nevertheless, his political thought lies clearly within the tradition of liberalism.

The main source for Bosanquet's social and political philosophy is *The Philosophical Theory of the State* (1899; 4th ed., 1923), though many of his ideas are developed in dozens of articles and essays which he wrote for professional academic journals, for publications of the Charity Organisation Society and for the popular press. Like many of his fellow idealists (notably T.H. Green, D.G. Ritchie, William Wallace, John Watson and, to a lesser degree, F.H. Bradley). Bosanquet's principal concern was to explain the basis of political authority and the state, the place of the citizen in society, and the nature, source and limits of human rights. The political theory that he develops is importantly related to his metaphysics and logic--particularly to such notions as the individual, the general will, ‘the best life’, society, and the state. In order to provide a coherent account of such issues, Bosanquet argued, one must abandon some of the

assumptions of the liberal tradition--particularly those that reveal a commitment to ‘individualism’.

Bosanquet saw authority and the state neither as based on individual consent or a social contract, nor as simply institutions where there is a general recognition of a sovereign, but as products of the natural development of human life, and as expressions of what he called the ‘real’ or general will. On Bosanquet's view, the will of the individual is "a mental system" whose parts--"ideas or groups of ideas"--are "connected in various degrees, and more or less subordinated to some dominant ideas which, as a rule, dictate the place and importance of the others" (i.e., of the other ideas that one has). Thus, Bosanquet writes that, "[i]n order to obtain a full statement of what we will, what we want at any moment must at least be corrected and amended by what we want at all other moments." But the process does not stop there. He continues: "this cannot be done without also correcting and amending it so as to harmonise it with what others want, which involves an application of the same process to them." In other words, if we wish to arrive at an accurate statement of what our will is, we must be concerned not only with what we wish at some particular moment, but also with all of the other wants, purposes, associations and feelings that we and others have (or might have) given all of the knowledge available. The result is one's ‘real’ or the ‘general will’.

Bosanquet sees a relation between the ‘real’ or ‘general will’ and the ‘common good.’ He writes that "The General Will seems to be, in the last resort, the ineradicable impulse of an intelligent being to a good extending beyond itself." This ‘good’ is nothing other than "the existence and the perfection of human personality" which he identifies with "the excellence of souls" and the complete realisation of the individual. It is so far as the state reflects the general will and this common good that its authority is legitimate and its action morally justifiable. Bosanquet describes the function of the state, then, as ‘the hindrance of hindrances’ to human development.

The influence of Rousseau and Hegel is clearly evident here. Indeed, Bosanquet saw in Hegel's *Philosophy of Right* a plausible account of the modern state as an ‘organism’ or whole united around a shared understanding of the good. Moreover, like Hegel, he argued that the state, like all other social ‘institutions,’ was best understood as an ethical idea and as existing at the level of consciousness rather than just material reality. Within nation states, Bosanquet held that the authority of the state is absolute, because social life requires a consistent co-ordination of the activities of individuals and institutions.

Still, although Bosanquet believed that the state was absolute, he did not exclude the possibility of an organized system of international law. The conditions for an effective recognition and enforcement of such a system were, he thought, absent at that moment--though he held out hope that the League of Nations reflected the beginnings of the consciousness of a genuine human community and that it might provide a mechanism by which multinational action could be accomplished.

Because the state can be said to reflect the general will that is also each individual's real will, Bosanquet held (following Rousseau) that sometimes individuals can be required to engage in certain activities for their own good--that is they can be ‘forced to be free.’ Moreover, he maintained that it is in terms of the ‘common good’ that one's ‘station’ or ‘function’ in society is defined, and it is the conscientious carrying out of the duties that are attached to one's ‘station’ that constitutes ethical behaviour. In fact, on

Bosanquet's account, it is primarily in light of one's service in the state that a person has the basis for speaking of his or her particular identity. Not surprisingly, then, Bosanquet was frequently challenged by those who claimed that he was anti-democratic and that his philosophical views led to a devaluation of the individual. Such attacks ignore, however, Bosanquet's insistence on liberty as the essence and quality of the human person and his emphasis on the moral development of the human individual and on limiting the state from directly promoting morality (which reflects both his own reading of Kant and the influence of Green's Kantianism.) Moreover, while Bosanquet did not hold that there were any *a priori* restrictions on state action, he held that there were a number of practical conditions that did limit it. For example, while law was seen as necessary to the promotion of the common good, it could not make a person good, and social progress could often be better achieved by volunteer action. (It is just this emphasis that Bosanquet found and defended in the approach to social work of the Charity Organisation Society.)

Although the state and law employ compulsion and restraint, they were considered to be ‘positive’ in that they provided the material conditions for liberty, the functioning of social institutions, and the development of individual moral character. For Bosanquet, then, there was no incompatibility between liberty and the law. Moreover, since individuals are necessarily social beings, their rights were neither absolute and inalienable, but reflected the ‘function’ or ‘positions’ they held in the community. For such rights to have not only moral but legal weight, Bosanquet insisted that they had to be ‘recognized’ by the state in law. Strictly speaking, then, there could be no rights against the state. Nevertheless, Bosanquet acknowledged that, where social institutions were fundamentally corrupt, even though there was no right to rebellion, there could be a duty to resist.

Although Bosanquet is sometimes regarded as a conservative, recent studies have pointed out that he was an active Liberal and, in the 1910s, supported the Labour Party. He insisted on the positive role that the state can have in the promotion of social well being and he was in favour of worker ownership. It is also worth noting that Bosanquet's audience was as much the professional in social work or the politician, as the philosopher. He was well-informed of the political situation in Britain, on the continent, and in the United States. His interests extended to economics and social welfare, and his work in adult education and social work provides a strong empirical dimension to his work. This background provided him with a broad base from which to reply to challenges from many of his critics-- e.g., from philosophers, like Mill and Spencer, and from social reformers, such as Sidney and Beatrice Webb and, the founder of the Salvation Army, General William Booth. Despite charges that Bosanquet's political philosophy is simplistic, inconsistent, or naive, Adam Ulam notes that *The Philosophical Theory of the State* "has a comprehensiveness and an awareness of conflicting political and philosophical opinions which give it a supreme importance in modern political thought. Bosanquet is both a political theorist and a political analyst."

It has sometimes been suggested that the influences of Kant and Hegel lead to a tension in Bosanquet's political thought. Bosanquet's emphasis on the moral development of the human individual and on limiting the state from directly promoting morality clearly reflects both his own reading of Kant and the Kantian influences on Green. Moreover, Bosanquet believed that the ‘best life’ that he describes as the ‘end’ of the individual and of the state alike, approximates what Kant referred to as ‘the kingdom of ends’. Even Bosanquet's justification of the authority of the state can be seen as a reflection of a Kantian

imperative that one wills the state as a necessary means to the moral end.

General Assessment

Interest in Bosanquet's work--as with idealism as a whole--waned during the middle decades of the 20th century. Of the idealists, the writings of Bradley and, in political theory, Green, are now much better known. There is no simple explanation of this; many factors seem relevant.

First, some of the work that made Bosanquet's reputation in his time--his popular essays, the books and articles that came out of his university extension courses, and his involvement in social policy--now seems largely dated. For example, several of his essays lack the logical rigor that one finds in material destined for the more specialized audience of academic philosophers. While insightful and wide ranging--and while accessible to a much wider audience than the work of other idealists, such as Bradley and J.M.E. McTaggart--Bosanquet's writings lack the sharpness, the density, and, at times, the outrageousness of those of some of his contemporaries.

It has been suggested, as well, that some of the concepts central to Bosanquet's work are not clearly defined, and Bosanquet himself was an indifferent literary stylist. His work often betrays a looseness that one tends to find in texts based on lectures prepared for general audiences or for classes, and even his early work on logic was remarked upon for its "stiffness." But these primarily stylistic concerns may also be a product of refusing to sever the analysis of concepts from the experience which Bosanquet was trying to describe.

There are other reasons that no doubt contributed to the decline of interest in Bosanquet's work. Aside from the general collapse of idealism as a philosophical movement--by the early part of the 20th century, it was seen by many as a philosophical dead- end--and the suspicion of what was regarded by later generations as its obscure vocabulary, Bosanquet's association with the majority report of the Poor Law Reform Commission and his alleged championing of the nation state, led many to see him as a conservative if not reactionary thinker whose contributions to philosophy and politics were outdated almost as soon as they had been published.

In recent years, however, there has been a renewed interest in Bosanquet's work--particularly concerning his philosophical and social thought, which is experiencing a revival in the work of some contemporary liberal theorists. Given the number of studies published during the past twenty years on Hegel, Green and, more recently, Bradley, and given the reevaluation of the significance of the work of British idealism and its place in the history of philosophy, it seems likely that there will be a reconsideration of the contribution of Bosanquet's philosophy as well.

Bosanquet's Works

A Comprehensive Listing:

The most comprehensive list to date of Bosanquet's work is found in Peter P. Nicholson, "A Bibliography of the Writings of Bernard Bosanquet (1848-1923)," *Idealistic Studies*, 8 (1978): 261-280.

Principal Works

The publication of a 20 volume set of *The Works of Bernard Bosanquet* (edited by William Sweet) is planned by Thoemmes Press (Bristol, U.K.) for 1999. This will include the following principal works:

- *Knowledge and Reality, A Criticism of Mr. F. H. Bradley's 'Principles of Logic'*. London: Kegan Paul, Trench, 1885.
- *Logic, or the Morphology of Knowledge*. Oxford: Clarendon Press, 1888. 2d ed., 1911.
- *Essays and Addresses*. London, Swan Sonnenschein, 1889.
- *A History of Aesthetic*, London: Swan Sonnenschein, 1892. 2d ed., 1904.
- *The Civilization of Christendom and Other Studies.* London: Swan Sonnenschein, 1893.
- *The Essentials of Logic: Being Ten Lectures on Judgement and Inference*. London and New York: Macmillan, 1895.
- *Aspects of the Social Problem*, London, 1895.
- *A Companion to Plato's Republic for English Readers: Being a Commentary adapted to Davies and Vaughan's Translation*. New York/London, 1895.
- *The Philosophical Theory of the State*, London, 1899; 4th ed., 1923.
- *Psychology of the Moral Self*, London and New York: Macmillan, 1897.
- *The Principle of Individuality and Value. The Gifford Lectures for 1911 delivered in Edinburgh University*. London: Macmillan, 1912.
- *The Value and Destiny of the Individual. The Gifford Lectures for 1912 delivered in Edinburgh University*. London: Macmillan, 1913.
- *The Distinction Between Mind and its Objects. The Adamson Lecture for 1913 with an Appendix.* Manchester: University Press, 1913
- *Three Lectures on Aesthetic*, London: Macmillan, 1915.
- *Social and International Ideals: Being Studies in Patriotism*, London: Macmillan, 1917.
- *Some Suggestions in Ethics*, London: Macmillan, 1918; 2nd ed. 1919.
- "Do Finite Individuals possess a substantive or an adjectival mode of being?", *Life and Finite Individuality*, (ed. H. Wildon Carr), *Proceedings of the Aristotelian Society*, supp. vol. 1, (1918): 75-102; 179-194 (Reprinted from *Proceedings of the Aristotelian Society*, n.s. XVIII (1917-1918): 479-506.)
- *Implication and Linear Inference*, London: Macmillan, 1920.
- *What Religion Is*, London: Macmillan, 1920.
- *The Meeting of Extremes in Contemporary Philosophy*. London: Macmillan, 1921.
- *Three Chapters on the Nature of Mind*, London: Macmillan, 1923.
- "Life and Philosophy," *Contemporary British Philosophy*, (ed. J.H. Muirhead), London, 1924, pp. 51-74.
- *Science and Philosophy and Other Essays by the Late Bernard Bosanquet*, (ed. J.H. Muirhead and

R.C. Bosanquet), London, Allen and Unwin, 1927.

Select Bibliography

- Acton, H.B. "Bernard Bosanquet," *The Encyclopedia of Philosophy*. Ed. Paul Edwards, New York, 1967, Vol. 1, pp. 347-350.
- Acton, H.B. "The Theory of Concrete Universals," *Mind*, n.s. XLV (1936): 417-31; n.s. XLVI (1937): 1-13.
- Armour, Leslie. "The Dialectics of Rationality: Bosanquet, Newman and the Concept of Assent," in *Rationality Today,* Ottawa, ON: University of Ottawa Press, 1979, pp. 491-497.
- Bedau, Hugo Adam. "Retribution and the Theory of Punishment," *Journal of Philosophy* 75 (1978): 601-620.
- Bosanquet, Helen. *Bernard Bosanquet: A Short Account of his Life* . London, 1924.
- Bradley, James. "Hegel in Britain: A Brief History of British Commentary and Attitudes," *The Heythrop Journal,* Vol. 20, (1979): 1-24; 163-182.
- Broad, C.D. "The Notion of a General Will," *Mind*, n.s. XXVIII, (1919): 502-504.
- Bussey, Gertrude Carman. "Dr. Bosanquet's Doctrine of Freedom," *Philosophical Review*, XXV (1916): 711-719 and 728-730.
- Carritt, E.F. *Morals and Politics: Theories of their Relation from Hobbes and Spinoza to Marx and Bosanquet*. Oxford, 1935.
- Cole, G.D.H. "Loyalties," *Proceedings of the Aristotelian Society*, n.s. XXVI (1925-1926): 151-170.
- Collini, S. "Hobhouse, Bosanquet and the State: Philosophical Idealism and Political Argument in England: 1880-1918," *Past and Present,* 72 (1976): 86-111.
- Collini, S. "Sociology and Idealism in Britain: 1880-1920," *Archives europeennes de sociologie,* 19 (1978): 3-50.
- Crane, Marion Delia. "Dr. Bosanquet's Doctrine of Freedom," *Philosophical Review*, XXV (1916): 719-728.
- Crane, Marion Delia. "The Method in the Metaphysics of Bernard Bosanquet," *Philosophical Review*, XXIX (1920): 437-452.
- Crane (Carroll), Marion. *The Principles of Absolutism in the Metaphysics of Bernard Bosanquet*. New York. Ph.D. thesis in philosophy, Cornell University, 1921. (Reprinted in "The Principle of Individuality in the Metaphysics of Bernard Bosanquet," *Philosophical Review*, XXX (1921): 1-23 and "The Nature of the Absolute in the Metaphysics of Bernard Bosanquet, *Philosophical Review*, XXX (1921): 178- 191.)
- Cunningham, G. Watts. "Bosanquet on Philosophic Method," *Philosophical Review*, XXXV (1926): 315-327.
- Cunningham, G. Watts. "Bosanquet on Teleology as a Metaphysical Category," *Philosophical Review*, XXXII (1923): 612-624.
- Cunningham, G. Watts. *The Idealist Argument in Recent British and American Philosophy*. New York, 1933.
- den Otter, Sandra. *British Idealism and Social Explanation: A Study in Late Victorian Thought,*

Oxford: Clarendon Press, 1996.

- Dockhorn, Klaus. *Die Staatsphilosophie des Englischen Idealismus*. Köln/Bochum-Langendreer: Heinrich Poppinghaus o. H.-G., 1937. (Bosanquet is discussed on pp. 61-116.)
- Emmet, Dorothy. "Bosanquet's Social Theory of the State," *The Sociological Review*, 37 (1989): 104-127.
- Feinberg, Walter. *A Comparative Study of the Social Philosophies of John Dewey and Bernard Bosanquet*. Ph.D. thesis in philosophy, Boston University, 1966.
- Fisher, John. "The Ease and Difficulty of Theory," *Dialectics and Humanism*, 3 (1976)): 117- 124.
- Gaus, Gerald. "Green, Bosanquet and the philosophy of coherence" in *Routledge History of Philosophy, Volume 7 - The Nineteenth Century*, Ed. C.L. Ten, London, 1994.
- Gaus, Gerald. *The Modern Liberal Theory of Man*. Canberra: Croom Helm, 1983.
- Gibbins, John R. "Liberalism, Nationalism and the English Idealists," in *History of European Ideas*, 15 (1992): 491-497.
- Gilbert, K. "The Principle of Reason in the Light of Bosanquet's Philosophy," *Philosophical Review*, XXXII (1923): 599-611.
- Ginsberg, Morris. "Is there a general will?," *Proceedings of the Aristotelian Society*, XX (1919-1920): 89-112.
- Harris, Frederick Philip. *The Neo-Idealist Political Theory: Its Continuity with the British Tradition*. New York. King's Crown Press, 1944 (Ph.D. thesis, Columbia University).
- Haldar, Hira-lal. *Neo-hegelianism*. London, 1927.
- Hobhouse, Leonard T. *The Metaphysical Theory of the State*. London, 1918.
- Hoernlé, R.F.A. "Bernard Bosanquet's Philosophy of the State," *Political Science Quarterly*, 34 (1919): 609-631.
- Hogdson, S.H. "Bernard Bosanquet's Recent Criticism of Green's Ethics," *Proceedings of the Aristotelian Society*, II (1901-1902): 66-71.
- Houang, François. *De l'humanisme á l'absolutisme: l'évolution de la pensée religieuse du néo-hegelien anglais Bernard Bosanquet*. Paris, Vrin, 1954.
- Houang, François. *Le neo-hegelianisme en Angleterre: la philosophie de Bernard Bosanquet (1848-1923)*. Paris: Vrin, 1954.
- Jacobs, Ellen. *Bernard Bosanquet: Social and Political Thought*. Ph.D. thesis, City University of New York, 1986.
- Jacquette, Dale, "Bosanquet's Concept of Difficult Beauty," *Journal of Aesthetics and Art Criticism*, 43 (1984): 79-88.
- Lang, Berel. "Bosanquet's Aesthetic: A History and Philosophy of the Symbol," *Journal of Aesthetics and Art Criticism*, 26 (1968): 377-387.
- Laski, H. "Bosanquet's Theory of the General Will," *Proceedings of the Aristotelian Society*, n.s. supp. vol. VIII (1928): 45-61.
- LeChevalier, Charles. *La pensée morale de Bernard Bosanquet (1848- 1923): Étude sur l'univers moral de l'idéalisme anglais au 19e siecle*. (Thèse complementaire pour le doctorat ès lettres) Paris: Vrin, 1963. (Republished under the title *Éthique et idéalisme: le courant néo-hegelien en Angleterre, Bernard Bosanquet et ses amis*. Paris: Vrin, 1963.)
- Lindsay, A.D. "Bosanquet's Theory of the General Will," *Proceedings of the Aristotelian Society*, n.s., supp. vol. VIII (1928): 31-44.

- Lindsay, A.D. "Sovereignty," *Proceedings of the Aristotelian Society*, XXIV (1923-1924): 235-254.
- MacAdam, James I. "What Rousseau Meant by the General Will", in *Rousseau's Response to Hobbes*, Eds. Howard R. Cell and James I. MacAdam, New York: Peter Lang, 1988, pp. 152-153. (This chapter originally appeared as an article in *Dialogue*, V, (1966-1967): 498-515.)
- MacIver, R.M. *Community: A Sociological Study*. New York, 1917. (Esp. Appendix A on the individual, the association, and the community, pp. 421-425, and Appendix B, "A Criticism of the Neo- Hegelian Identification of Society and the State," pp. 425-433.)
- MacIver, R.M. *Politics and Society*, Ed. David Spitz, New York: Atherton Press, 1969. (Contains letters between Bosanquet and MacIver on the distinction between society and the state.)
- Marcuse, Herbert. *Reason and Revelation: Hegel and the rise of Social Theory*. Boston: Beacon Press, 1960.
- Mathew, M.C. "Bosanquet's Logical Theory," *Philosophical Quarterly of India* 17 :314-324.
- McBriar, A.M. *An Edwardian Mixed Doubles: The Bosanquets versus the Webbs; A Study in British Social Policy*. Oxford, 1987.
- McTaggart, J.M.E. *Studies in Hegelian Cosmology*. Cambridge: Cambridge University Press, 1901.
- Meadowcroft, James. *Conceptualizing the State: Innovation and Dispute in British Political Thought 1880-1914*. Oxford: Clarendon Press, 1995.
- Metz, Rudolf. *Die philosophischen Stromungen der Gegenwart in Groβbritannien*. Leipzig: Felix Meiner Verlag, 1935; (*A Hundred Years of British Philosophy*. Trans. J.W. Harvey, T.E. Jessop and Henry Sturt; Ed. J.H. Muirhead. London, 1938).
- Milne, A.J.M. *The Social Philosophy of English Idealism*. London, 1962.
- Morrow, John. "Ancestors, Legacies and Traditions: British Idealism in the History of Political Thought," *History of Political Thought*, 6 (1985): 491-515.
- Morrow, John. "Liberalism and British Idealist Political Philosophy: A Reassessment," *History of Political Thought*, 5 (1984): 91-108.
- Morris-Jones, Huw. "Bernard Bosanquet," *International Encyclopedia of the Social Sciences*. Ed. David L. Sills, New York: The Free Press, 1968, Vol. 2, pp. 131-134.
- Moser, Claudia. *Die Erkenntnis- und Realitaets-problematik bei F.H. Bradley und B. Bosanquet*. Würzburg, 1989.
- Mowat, Charles L. *The Charity Organization Society*. London, 1961.
- Muirhead, J.H. (ed.) *Bernard Bosanquet and his Friends*. London, 1935.
- Nicholson, Peter P. "Philosophical Idealism and International Politics: A Reply to Dr. Savigear," *British Journal of International Studies*, 2 (1976): 76-83.
- Nicholson, Peter P. *The Political Philosophy of the British Idealists: Selected Studies*. Cambridge, 1990.
- O'Sullivan, Noel. *The Problem of Political Obligation*. London, 1986.
- Oakeshott, Michael. "Review of Bertil Pfannenstill, *Bernard Bosanquet's Philosophy of the State*," *Philosophy*, 11 (1936): 482-483.
- Pant, Nalini. *Theory of Rights: Green, Bosanquet, Spencer, and Laski*. Varanasi, Vishwavidyalaya Prakashan, 1977.
- Parker, Christopher. "Bernard Bosanquet, Historical Knowledge, and the History of Ideas,"

Philosophy of Social Science, 18 (1988): 213-230.

- Pearson, Robert and Geraint Williams, *Political Thought and Public Policy in the Nineteenth Century: An introduction*. London, 1984.
- Pfannenstill, Bertil. *Bernard Bosanquet's Philosophy of the State*. Lund, 1936.
- Primoratz, Igor. "The Word 'Liberty' on the Chains of Galley-Slaves: Bosanquet's Theory of the General Will," *History of Political Thought*, 15 (1994): 249-267.
- Pucelle, Jean. *L'idéalisme en angleterre de Coleridge á Bradley*. Neuchatel, 1955.
- Quinton, Anthony. "Absolute Idealism," *Proceedings of the British Academy*, LVII (1971): 303-329.
- Randall, J.H., Jr. "Idealistic Social Philosophy and Bernard Bosanquet," *Philosophy and Phenomenological Research*, XXVI (1966): 473-502. (Reprinted in *The Career of Philosophy*. 3 vols., Vol. 3, New York, Columbia University Press, 1977, pp. 97- 130.)
- Robbins, Peter. *The British Hegelians: 1875-1925*. New York, 1982.
- Robinson, Jonathan. "Bradley and Bosanquet," *Idealistic Studies*, 10 (1980): 1-23.
- Russell, Bertrand, C. Delisle Burns, and G.D.H. Cole. "The Nature of the State in its External Relations," *Proceedings of the Aristotelian Society*, n.s. vol. XVI (1915- 1916): 290-310. (A round table, with a discussion of Bosanquet's theory of international politics.)
- Sabine, George. "Bosanquet's Theory of the Real Will," *Philosophical Review*, XXXII (1923): 633-651.
- Sabine, George. *A History of Political Theory*. 4th ed., Hinsdale, IL: The Dryden Press, 1973. (A discussion and critique of Bosanquet and T.H. Green, pp. 725- 753.)
- Salomaa, J.A. *Idealismus und Realismus in der englischen Philosophie der Gegenwart*. Helsinki, 1929.
- Sell, Alan P.F. *Philosophical Idealism and Christian Belief*. New York: St. Martin's Press, 1995.
- Seth Pringle-Pattison, Andrew. "Do Finite Individuals possess a substantive or an adjectival mode of being?", in *Life and Finite Individuality*, Ed. H. Wildon Carr, *Proceedings of the Aristotelian Society*, supp. vol. I, (1918): 103-126.
- Seth Pringle-Pattison, Andrew. *The Idea of God in the Light of Recent Philosophy*. Oxford, 1917.
- Spiller, Gustav. *The Ethical Movement in Britain: A Documentary History*. London, 1934.
- Stedman, R.E. "Nature in the Philosophy of Bosanquet," *Mind*, n.s. XLIII (1934): 321-334.
- Stedman, Ralph. "Bosanquet's Doctrine of Self-Transcendence," *Mind*, n.s. XL (1931):
- Sturt, Henry. *Idola Theatri: A Criticism of Oxford Thought and Thinkers from the Standpoint of Personal Idealism*. London, 1906.
- Sweet, William. "Bernard Bosanquet and the Development of Rousseau's Idea of the General Will," in *Man and Nature - L'homme et la nature*, X (1991): 179-197.
- Sweet, William. "Bosanquet and British Political Thought," in *Laval theologique et philosophique*, forthcoming
- Sweet, William. "Bosanquet et les droits de la personne," in *Cahiers de l'équipe de recherche en éthique sociale*, no. 9701, Montréal: Équipe de recherche en éthique sociale, 1997.
- Sweet, William. "British Idealism," in *The Philosophy of Law: An Encyclopedia*, (ed. Christopher B. Gray), New York: Garland Publishing, forthcoming 1999.
- Sweet, William. "Critical Review of Peter P. Nicholson, *The Political Philosophy of the British Idealists: Selected Studies*," in *Laval theologique et philosophique*, Vol. 48 (1992): 477-480.

- Sweet, William. "F.H. Bradley and Bernard Bosanquet," in *Philosophy after F.H. Bradley*, Ed. James Bradley, Bristol, UK: Thoemmes Press, 1996.
- Sweet, William. *Idealism and Rights*, Lanham, MD: University Press of America, 1997.
- Sweet, William. "Individual Rights, Communitarianism, and British Idealism," in *The Bill of Rights: Bicentennial Reflections*, (ed. Yeager Hudson and Creighton Peden), Lewiston, NY: Edwin Mellen Press, 1993, pp. 261-277.
- Sweet, William. "Is Later British Idealist Political Theory Fundamentally Conservative?," in *The European Legacy: Toward New Paradigms*, Vol. 1, No. 1 (Cambridge, MA: MIT Press, 1996): 403-408.
- Sweet, William. "Law and Liberty in J.S. Mill and Bernard Bosanquet," in *The Social Power of Ideas*, (ed. Yeager Hudson and W. Creighton Peden), Lewiston, NY: Edwin Mellen Press, 1995, pp. 361-385.
- Sweet, William. "The Legitimacy of Law: From Contract to Community," in *Indian Socio-Legal Journal*, Vol. XIX, No. 2 (1993): 69-84.
- Sweet, William. "Liberalism, Bosanquet and the Theory of the State," in *Liberalism, Oppression, and Empowerment*, (ed. Creighton Peden and Yeager Hudson), Lewiston, NY: Edwin Mellen Press, 1995, pp. 3-34.
- Sweet, William. "L'individu et les droits de la personne selon Maritain et Bosanquet," *Études Maritainiennes / Maritain Studies*, VI (1990): 141- 166.
- Sweet, William. "Was Bosanquet a Hegelian?," in *Bulletin of the Hegel Society of Great Britain*, No. 31 (1995): 39-60.
- Tallon, Hugh Joseph. *The Concept of Self in British and American Idealism*. Washington, D.C.: Catholic University of America Press, 1939.
- Thakurdas, Frank. *The English Utilitarians and the Idealists*. Delhi: Vishal Publication, 1978.
- Tsanoff, Radoslav A. "The Destiny of the Self in Professor Bosanquet's Theory," *Philosophical Review*, XXIX (1920): 59-79.
- Turner, Frank M. *The Greek Heritage in Victorian Britain*. New Haven: Yale University Press, 1981.
- Ulam, Adam. *The Philosophical Foundations of English Socialism*. Cambridge, MA, 1951.
- Vincent, Andrew. "Citizenship, poverty and the real will," *The Sociological Review*, 40 (1992): 702-725.
- Vincent, Andrew and Raymond Plant. *Philosophy, Politics and Citizenship: the Life and Thought of the British Idealists*. Oxford, 1984.
- von Trott, A. "Bernard Bosanquet und der Einfluβ Hegels auf die englische Staatsphilosophie,"*Zeitschrift für Deutsche Kulturphilosophie*, Band 4, Heft 2 (1938): 193-199.
- Wahl, Jean. *Les philosophes pluralistes d'angleterre et d'amerique*. Paris, 1920.
- Watson, John. "Bosanquet on Mind and the Absolute," *Philosophical Review*, XXXIV (1925): 427-442.
- Weldon, T.D. *States and Morals*. London, 1946.
- White, David A. "Revealment: A Meeting of Extremes in Aesthetics" *Journal of Aesthetics and Art Criticism*, 515-520.
- Willis, Kirk. "The Introduction and Critical Reception of Hegelian Thought in Britain 1830-1900," *Victorian Studies*, 32 (1988): 85-111.

Other Internet Resources

- Bernard Bosanquet

Related Entries

Bradley, Francis Herbert | idealism: British | liberalism

William Sweet
wsweet@stfx.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 15, 1997
Content last modified: September 20, 1998

The Revision Theory of Truth

The revision theory of truth was developed independently by Gupta (1982) and Herzberger (1982) in an attempt to analyze paradoxes such as the liar paradox that appear to show that common-sense beliefs about truth are inconsistent.

Liar Paradox

Consider the following "liar sentence."

(L) Sentence (L) is not true.

Sentence (L) says of itself that it is not true. A contradiction can be derived from (L) from apparently trivial principles. One basic intuition about truth needed is that a sentence is true if and only if what it asserts is the case. For example, the sentence "there is gold on the moon" is true if and only if there is gold on the moon. Applying this principle in the particular case of sentence (L) results in:

(1) Sentence (L) is true if and only if sentence (L) is not true.

Notice that sentence (L) asserts precisely that it is not true. So by the principle just mentioned it is true if and only if it is not true.

Another principle needed to obtain an explicit contradiction is the following:

(2) Sentence (L) is either true or not.

Sentence (1) may already appear to be obviously contradictory. If not, a contradiction can be quickly derived using (2) by breaking into the two cases it allows: that (L) is true, or that (L) is not.

Case 1. Sentence (L) is true. Then by (1) sentence (L) is not true. So it is both true and not true, which is impossible.

Case 2. Sentence (L) is not true. Then by (1) sentence (L) is true. Again it is both true and not true, which is impossible.

So either of the two cases is seen to be impossible by applying (1). This seems to imply that at least one of these basic intuitions expressed in (1) and (2) is wrong.

Description of the Theory

The revision theory of truth seeks to explain the meaning of truth. According to the theory the meaning of truth for a collection of sentences (a "language") is given by the "Tarski biconditionals" as they are called in Gupta (1982). An example of a Tarski biconditional is (1) above. In general, where P is a sentence of the language and S is a name of the sentence, the Tarski biconditional for the sentence is:

> S is true if and only if P

One way of forming names of sentences is by quotation, so that if the sentence in question is "there is gold on the moon," the Tarski biconditional for the sentence is:

> "there is gold on the moon" is true if and only if there is gold on the moon.

Another method of giving names to sentences is by Goedel numbering, assigning numbers to sentences in some typically systematic fashion.

While the Tarski biconditionals seem to be quite trivial, they are seen to lead to blatant contradictions when applied to sentences such as the liar sentence.

According to the revision theory, while the Tarski biconditionals give the meaning of truth, special semantical tools are needed to show how they generate the concept of truth. In particular, the theory accepts that truth is a circular concept, and provides special tools for understanding circular concepts such as truth. Thus, the reasoning above that resulted in a contradiction from would be seen as misapplying the information expressed in (1), the Tarski biconditional for the liar sentence.

On the revision theory, Tarski biconditionals such as (1) should be understood as having a hypothetical character. While they entirely define the concept of truth, they do so only in virtue of the special role

given to them by the revision theory. In particular, the biconditionals are seen as providing a method for obtaining better and better approximations of the extension of the truth predicate. Thus, they do not simply provide the extension of the truth predicate, but provide an improvement on any temporary extension that might be suggested.

Thus, let M be an ordinary first-order model that also gives an arbitrary extension to the truth predicate. The Tarski biconditionals provide a method for obtaining an improved model M*. Namely, for any sentence P having name S, S is assigned to the extension of the truth predicate in M* if P evaluates as true in M, and not assigned to the extension of truth otherwise. Thus, given any model M with any initial extension to the truth predicate, the biconditionals generate a series of models M*, M**, M***, etc., that are constructed using the biconditionals by evaluating sentences in the previous member of the series.

The series is also extended into the transfinite by collecting together the results from earlier stages at limit ordinals. One method of doing this is, at a limit stage, letting the extension of truth at the limit stage consist of all (names of) sentences that have stabilized as the sequence approached the limit. That is, if at some stage in the sequence a sentence is declared true at every subsequent stage below the limit stage, then put it in the extension of truth at the limit. Many other reasonable limit rules are possible here.

Consequences and Properties

A "revision sequence" is any sequence of models beginning with an arbitrary model M that is generated by the Tarski bicondionals according to the revision theory of truth.

Some sentences will stabilize eventually in every revision sequence. For example, where "T" is the truth predicate, let S be the name of the sentence "T(T(F(b)))" where "F" is an arbitrary one-place predicate and "b" is an arbitrary name.

Let M(T) represent the extension assigned to T by M. Then S is in M**(T) if and only if "T(F(b))" is in M*(T). But "T(F(b))" is in M*(T) if and only if "F(b)" is in M(T). Thus, S is in M**(T) if and only if "F(b)" is in M(T), that is, if and only if b is in M(F). Likewise, from the second revision onwards, S is assigned to the extension of truth if and only if b is in the extension of the predicate F. So beginning with any model M the sentence S stabilizes as either true or false in every revision sequence depending on whether "F(b)" is evaluated as either true or false in the original model M.

Intuitively "normal" sentences stabilize in every sequence. Sentences such as the liar sentence exhibit unusual behavior in the framework of the revision theory. For example, the liar sentence alternates between true and false at successive revisions in every revision sequence. Hence it exhibits extremely unstable behavior.

Many other classifications of sentences with respect to their behavior over various revision sequences are possible. Some will stabilize as false in all sequences. Some will stabilize as true in some but not all sequences. Some will stabilize as true in some sequences and false in the rest. This apparatus provides

tools for fine-grained classifications of various types of sentences into different semantical categories.

Because the theory deals with sequences of classical models, every logical truth will stabilize as true in every sequence and every logical falsehood as false in every sequence. Hence one of the advantages often claimed for the revision theory is that it supports classical reasoning in contrast to various other approaches to truth.

Along with other recent theories of truth, the revision theory has the feature that given any revision, eventually a stage will be reached at which every sentence that will ever stabilize as true or as false in the sequence has already stabilized.

Bibliography

- Belnap, Nuel (1982) "Gupta's Rule of Revision Theory of Truth." *Journal of Philosophical Logic* 11: 103 - 16.
- Gupta, Anil (1981) "Truth and Paradox" (abstract). *Journal of Philosophy* 78: 735 - 6.
- Gupta, Anil (1982) "Truth and Paradox." *Journal of Philosophical Logic* 11: 1 - 60. Reprinted in Martin (1984).
- Gupta, Anil and Nuel Belnap (1993) *The Revision Theory of Truth* Cambridge, MA: The MIT Press.
- Herzberger, Hans (1982) "Notes on Naive Semantics." *Journal of Philosophical Logic* 11: 61 - 102. Reprinted in Martin (1984).
- Herzberger, Hans (1982) "Naive Semantics and the Liar Paradox." *Journal of Philosophy* 79: 479 - 97.
- Martin, Robert, ed. (1984) *Recent Essays on Truth and the Liar Paradox* Oxford: Oxford University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

truth: coherence theory of | truth: correspondence theory of | truth: deflationary theory of | truth: identity theory of

Eric M. Hammer
v-erhamm@microsoft.com

First published: December 15, 1995
Content last modified: January 2, 1996

Folk Psychology as a Theory

Many philosophers and cognitive scientists claim that our everyday or "folk" understanding of mental states constitutes a theory of mind. That theory is widely called "folk psychology" (sometimes "commonsense" psychology). The terms in which folk psychology is couched are the familiar ones of "belief" and "desire", "hunger", "pain" and so forth. According to many theorists, folk psychology plays a central role in our capacity to predict and explain the behavior of ourselves and others. However, the nature and status of folk psychology remains controversial.

Historical Background

One important source of the idea that our everyday understanding of mental states constitutes a folk theory of mind is Wilfred Sellars's attack on what he called "the myth of the given" (Sellars 1956). Sellars denied that the content of our mental life is simply presented to us; that is, he denied that our beliefs about our own mental states enjoy a privileged epistemic status. To weaken the grip of the myth of the given he presented an alternative myth in which our ancestors, initially limited to a purely behavioristic understanding of action, learned a new theory of action that posits inner episodes as the causes of overt behavior. At first our ancestors only applied the new theory to others, but then they learned to "read" their own mental states off their behavior. In the final stages of the myth they became adept at mental state self-attribution without theorizing from their behavior; nevertheless the self-attributed states remain the posits of an introduced theory of mind.

It need hardly be said that Sellars did not take himself to have accurately accounted for the historical

origins of our capacity for mental state attribution. Rather, his aim was to open a new space in the debate about the status of mental state attributions by pointing out that mental states could be the posits of a theory of mind. As we shall see, subsequent generations took Sellars's suggestion entirely seriously, and were right to do so.

Another important historical source of the idea that our everyday understanding of mental states constitutes a folk theory of mind is the rise of cognitivism in the 1960s. Following the widespread perception that behaviorism had failed, cognitive scientists began to posit internal episodes as causes of overt behavior. These internal episodes were typically taken to be *representations,* and the term "theory" was applied to posited representational structures of sufficient complexity. For example, cognitive scientists seeking to explain our capacity to manipulate middle-sized physical objects posited an internally represented theory of dynamics, often referred to as "folk physics". (See for example McCloskey 1983.) It was only natural, therefore, that cognitive scientists adopted the explanatory strategy of positing internally represented theories when they attempted to explain our folk capacity to predict and explain behavior. The label "folk psychology" was widely adopted for that posit.

It is worth mentioning a third historical strand. Since at least the 1940s social psychologists have been interested in our capacity to attribute mental states to others. In an important early study, Heider and Simmel showed subjects a short movie which consisted of geometric shapes moving on a screen (Heider & Simmel 1944). When asked to report what they saw, almost every subject attributed propositional attitudes to the shapes, suggesting the existence of a universal and largely automatic capacity for propositional attitude attribution. In subsequent decades, social psychologists explored the accuracy and limitations of this capacity. For example, Nisbett and Wilson (1977) investigated a range of circumstances under which subjects *mispredict* behavior, and discovered that the circumstances under which mispredictions occur are just those circumstances under which subjects offer contrived or confabulated explanations of behavior.

Two Senses of "Folk Psychology"

I remarked above that many philosophers and cognitive scientists claim that our folk understanding of mental states constitutes a theory of mind. But we can unpack the idea of an *understanding* of mental states in two ways, yielding two different senses of "folk psychology": an *externalist* sense and an *internalist* sense (Stich & Ravenscroft 1994).

On the externalist account of folk psychology, folk psychology is a theory of mind implicit in our everyday talk about mental states. In the everyday traffic of our lives we make remarks linking sensory experiences to mental states; mental states to other mental states; and mental states to behavior. Thus we remark that the smell of freshly baked bread made Sally feel hungry; that Sally wanted to go on a diet because she thought that she was overweight; and that Sally went to the fridge because she desired a piece of chocolate cake. According to some philosophers, remarks such as these (or suitable generalisations of remarks such as these) function as a term-introducing theory which implicitly defines terms such as "believe", "want" and "desire". (See for example Lewis 1972.)

On the internalist account of folk psychology, folk psychology is a theory of human psychology which is represented in the mind-brain and which underpins our everyday capacity to predict and explain the behavior of ourselves and others. On this view, folk psychology is a *data structure* or *knowledge representation* which mediates between our observations of behavior-in-circumstances and our predictions and explanations of that behavior.

In the following two sections I give further details about the externalist and internalist accounts of folk psychology, respectively. In the final section I raise some issues about the relationship between folk psychology (internal) and folk psychology (external). However, before moving on it will be helpful to introduce a further piece of terminology which is often used in the literature.

The claim that our everyday understanding of mental states constitutes a folk theory of mind is often called the "theory theory". We have seen that two senses of "folk psychology" can be distinguished; similarly, two senses of "theory theory" can be distinguished. On the *externalist* reading of "theory theory", our everyday talk about mental states implicitly constitutes a theory of mind: folk psychology (external). On the *internalist* reading of "theory theory", our everyday capacity to predict and explain behavior is underpinned by an internally represented theory of mind: folk psychology (internal). Unfortunately, theory theorists are not always as clear as one might hope about which sense of "theory theory" they are endorsing.

The Nature and Status of Folk Psychology (External)

The principal architect of theory theory (external) is David Lewis (see especially Lewis 1972). Lewis instructs us to:

> Collect all the platitudes ... regarding the causal relations of mental states, sensory stimuli, and motor responses. ... Add also all the platitudes to the effect that one mental state falls under another ... Perhaps there are platitudes of other forms as well. Include only the platitudes which are common knowledge amongst us: everyone knows them, everyone knows that everyone else knows them, and so on. (Lewis 1972: 256.)

Having assembled the platitudes we can then form their conjunction. Let $m_1, \ldots, m_n$ be the mental state terms used in these platitudes. We can then express the conjunction of platitudes as:

$$S_1(m_1, \ldots ,m_n) \ \& \ S_2(m_1, \ldots ,m_n) \ \& \ \ldots \ \& \ S_j(m_1, \ldots ,m_n)$$

where each $S_i(m_1, \ldots ,m_n)$ is a sentence in which some or all of the mental state terms m_i occur. This conjunction will also contain a variety of terms which name non-mental states. For example, it will

contain terms referring to types of sensory input (sharp blows; bright lights; gentle strokings) and to types of behavioral output (saying "ouch"; shielding the eyes; smiling). Following Lewis we can call these terms the *O-terms* ($O_1, \dots, O_n$). In the interests of clarity, these terms have been suppressed.

We can now replace (each occurrence of) mental state term m_i by a corresponding free variable x_i:

$$S_1(x_1, \dots ,x_n) \ \& \ S_2(x_1, \dots ,x_n) \ \& \ \dots \ \& \ S_j(x_1, \dots ,x_n)$$

Prefixing an existential quantifier we obtain the *Ramsey sentence* for folk psychology:

$$\exists x_1 \dots x_n[S_1(x_1, \dots ,x_n) \ \& \ S_2(x_1, \dots ,x_n) \ \& \ \dots \ \& \ S_j(x_1, \dots ,x_n)]$$

The Ramsey sentence for folk psychology says that there exists a set of entities $x_1, \dots ,x_n$ which exhibit just those relations which the states named by the term $m_1, \dots , m_n$ exhibit. It is possible to obtain from the Ramsey sentence an explicit definition of any mental state term m_i. (See Lewis 1972 for the formal details.) Lewis has thus shown how to obtain an explicit definition of any mental state term m_i from the collected platitudes; in other words, he has shown how we can treat our everyday talk about mental states as a term-introducing theory of mind.

To clarify Lewis's Ramsey sentence approach, assume that our everyday talk about mental states consists of just three platitudes:

P1 Bodily damage causes pain.

P2 People who are in pain experience acute distress.

P3 People who are in pain nurse the afflicted body part.

These platitudes express the causal relationships between bodily damage and pain; between pain and states of acute distress; and between pain and a certain sort of behavior (nursing the afflicted body part).

Using "m_1" for "pain" and "m_2" for "acute distress", we can write the conjunction of P1 to P3 as:

$$S_1(m_1) \ \& \ S_2(m_1, m_2) \ \& \ S_3(m_1)$$

Once again the O-terms have been suppressed in the interests of clarity, but it is worth bearing in mind that the O-terms include a name referring to bodily damage and a name referring to a certain sort of behavior, *viz*, nursing the afflicted body part.

Replacing m_1 and m_2 with free variables x_1 and x_2, respectively, we obtain:

$$S_1(x_1) \& S_2(x_1, x_2) \& S_3(x_1).$$

Prefixing an existential quantifier we obtain the Ramsey sentence for our toy theory:

$$\exists x_1, x_2[S_1(x_1) \& S_2(x_1, x_2) \& S_3(x_1)].$$

The Ramsey sentence says that there exists states x_1 and x_2 that (respectively) play the roles accorded to pain and acute distress by the platitudes P1 to P3.

From the Ramsey sentence we can obtain an explicit definition of, say, m_1. (Again readers are encouraged to consult Lewis 1972 for the formal details.)

> m_1 (ie pain) = the unique x_1 such that x_1 is caused by bodily damage, causes acute distress, and causes the nursing of the afflicted body part.

Notice that the definition of m_1 was obtained from the platitudes: nothing was added during the process of defining m_1. It is clear, therefore, that the definition was implicit in the platitudes all along.

This example makes it clear that Lewis is primarily concerned with those platitudes which detail "the causal relations of mental states, sensory stimuli, and motor responses" (Lewis 1972: 256). Lewis is therefore interpreting folk psychology as a *functionalist* theory; that is, as a theory which identifies mental states in terms of their causal-functional relations. Indeed, some authors use the terms "theory theory" and "functionalism" interchangeably. Attractive though Lewis's position is, it is at least partly hostage to fortune. For it is an open question whether the theory implicit in our everyday platitudes about mental states really is a strictly functionalist one. Many authors have doubted that, for example, our talk about qualia can be adequately cashed out in functionalist terms. (See for example Chalmers 1996.) Indeed, it is an open question whether our everyday talk about mental states is sufficiently systematic to support Lewis's Ramsey sentence approach.

There is, moreover, a largely empirical question to be raised about folk psychology (external). For even if we accept that our everyday talk about mental states implicitly constitutes a theory of mind, it remains to be determined if that theory is *true*. Maybe future research in psychology or neuroscience will establish that folk psychology (external) is false. And if folk psychology (external) is false, it would seem to follow that there are no such thing as beliefs and desires, pains, hungers and tickles. This surprising doctrine is called *eliminativism*, and has been a major focus of discussion amongst philosophers of mind over the last 15 years. (See Churchland 1981; Horgan & Woodward 1985.)

The Nature and Status of Folk Psychology (Internal)

In our everyday social interactions we both predict and explain behavior, and our explanations are couched in a mentalistic vocabulary which includes terms like "belief" and "desire". For brevity's sake we can refer to such activities as *mentalizing*, and we can ask about the cognitive mechanisms which underpin our capacity to mentalize. According to the theory theory (internal), our capacity for mentalization crucially involves an internally represented theory of mind: folk psychology (internal). On this view, predicting and explaining human behavior is akin to predicting and explaining the movements of the heavenly bodies using Newton's mechanics. As we shall see, this analogy is very rough, but it emphasizes the central place given to theorizing on the theory theory model.

It is important to stress that internalist theory theorists need not be committed to any particular theory of mental representation. In particular, they need not be committed to the language of thought hypothesis. Folk psychology (internal) may be represented in the language of thought, or by a distributed connectionist network, or by some other means. (This point is well made by Stich & Nichols 1992.) Of course, since theories are essentially *representational* structures, theory theory (internal) is incompatible with radically *anti-representationalist* theories of mind. I will not, though, consider anti-representationalism further.

It is also important to stress that internalist theory theorists need not be committed to the claim that folk psychology (internal) is learned in the way that we learn, say, physics or chemistry. (It is here that the analogy with Newtonian mechanics breaks down.) Some internalist theory theorists have argued that folk psychology is indeed largely learned (see Gopnik & Wellman 1992). That position has been challenged by a version of Chomsky's "poverty of stimulus" argument. Empirical studies suggest that young children become fairly competent folk psychologists by four or five years (Wimmer & Perner 1983). Is it really plausible that four year olds have been exposed to sufficient examples of behavior-in-circumstance to construct, via the principles of induction, full-blown folk psychology?

Other internalist theory theorist argue that folk psychology (internal) is largely innate; or at least that we are born with a mechanism dedicated to the acquisition of folk psychology (see for example Fodor 1992 and Carruthers 1996: especially Section 1.7). There are strong parallels here with debates about nativism in psycholinguistics. Some psycholinguists argue that our capacity to use language is a product of natural selection (see for example Pinker 1994). Similarly, some internalist theory theorists argue that our capacity to predict and explain behavior is a product of natural selection (see for example Baron-Cohen 1992).

It should be clear that the theory theory (internal) is a very attractive doctrine. Nevertheless, it has not remained unchallenged. Simulation theorists have argued that our mentalizing capacity is not primarily a capacity to theorize but is rather a capacity to *simulate* the mental processes of others. (See folk psychology, as simulation.) The debate between simulationists and internalist theory theorists has both philosophical and empirical dimensions, and it is fair to say that, at present, the issue is very much open. (For an excellent introduction to this debate see Davies & Stone 1995.)

Concluding Remarks

I have identified two distinct senses of "folk psychology". What is the relationship between the theories to which those terms refer? It is implausible that the theory of mind implicit in our everyday talk about mental states is simply identical to the internally represented theory of mind which underpins our capacity to mentalize. Rather, it is likely that folk psychology (internal) is partly inaccessible to consciousness, and that folk psychology (external) is an articulation of that fragment of folk psychology (internal) which is available to conscious reflection. It follows that our everyday talk about the mind is only a rough guide to folk psychology (internal).

The previous paragraph assumed that the theory theory is true on both its internalist and its externalist readings. But if simulation theory is true, our capacity to mentalize is not underpinned by folk psychology (internal) and so the internalist version of the theory theory is false. Note, though, that the externalist version of the theory theory could remain true *even if* the internalist version were false: simulation theory is compatible with the idea that our everyday talk about mental states implicitly constitutes a theory of mind. These options, and the consequences for eliminativism, are considered in more detail in Stich & Ravenscroft 1994.

Bibliography

- Baron-Cohen, S. (1992): *Mindblindness* (Cambridge MA: MIT Press).
- Carruthers, P. (1996): *Language, Thought and Consciousness* (Cambridge: Cambridge University Press).
- Chalmers, D. (1996): *The Conscious Mind* (New York: Oxford University Press).
- Churchland, P. (1981): "Eliminative Materialism and the Propositional Attitudes". *Journal of Philosophy* 78: 67-90.
- Davies, M. & Stone, T. (1995): *Folk Psychology* (Oxford: Blackwell).
- Fodor, J. (1992): "A Theory of the Child's Theory of Mind". *Cognition* 44: 283-96.
- Gopnik, A. & Wellman, H. (1992): "Why the Child's Theory of Mind Really is a Theory. *Mind and Language* 7: 145-71.
- Heider, F. & Simmel, M. (1944): "An Experimental Study of Apparent Behavior". *American Journal of Psychology* 57: 243-59.
- Horgan, T. & Woodward, J. (1985): "Folk Psychology is Here to Stay". *Philosophical Review* 94: 197-226.
- Lewis, D. (1972): "Psychophysical and Theoretical Identifications". *Australasian Journal of Philosophy* 50: 249-58.
- McCloskey, M. (1983): "Naive Theories of Motion". In D. Gentner & A. Stevens (eds), *Mental Models* (Hillsdale: Erlbaum).
- Nisbett, R. & Wilson, T. (1977): "Telling More than We Know: Verbal Reports on Mental Processes". *Psychological Review* 84: 231-59.
- Pinker, S. (1994): *The Language Instinct* (New York: William Morrow).

- Sellars, W. (1956): "Empiricism and the Philosophy of Mind". In H. Feigl and M. Scriven (eds), *Minnesota Studies in the Philosophy of Science* (vol. 1) (Minneapolis: University of Minesota Press).
- Stich, S. & Nichols, S. (1992): "Folk Psychology: Simulation or Tacit Theory?". *Mind and Language* 7: 35-71.
- Stich, S. & Ravenscroft, I. (1994): "What is Folk Psychology?". *Cognition* 50: 447-68.
- Wimmer, H. & Perner, J. (1983): "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception". *Cognition* 13: 103-28.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

folk psychology: as mental simulation | functionalism | materialism: eliminative | Sellars, Wilfrid

Ian Ravenscroft
The Flinders University of South Australia
Ian.Ravenscroft@flinders.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 22, 1997
Content last modified: September 30, 1997

Folk Psychology as Mental Simulation

The simulation (or, "mental simulation") theory maintains that human beings are able to use the resources of their own minds to simulate the psychological causes of the behavior of others, typically by making decisions within a "pretend" context. The theory is usually, though not always, taken to present a serious challenge to the assumption that a *theory* underlies everyday human competence in predicting and explaining behavior, including the capacity to ascribe mental states to others. Unlike earlier controversies concerning the role of empathetic understanding and historical reenactment in the human sciences, the current debate between the simulation theory and the "theory" theory appeals to empirical findings, particularly experimental results concerning children's development of psychological competence. These are detailed in what follows.

What is Meant by ‘Simulation’?

Like the term ‘theory,’ ‘simulation’ has come to be used broadly and in a variety of ways. Simulation is sometimes equated with role-taking, or "putting oneself in the other's place." However, it is often taken to include mere "projection," or reliance on a shared world of facts and emotive and motivational charges, without adjustments in imagination; e.g., where there is no need to *put* oneself in the other's place, as one is, in all relevant respects, already there. (Gordon calls this the default mode of simulation.) Sometimes it is taken to include as well automatic responses such as the subliminal mimicry of facial expressions and bodily movements. Stephen Stich and Shaun Nichols, whose critical papers have clarified the issues and helped refine the theory, urge that the term be dropped in favor of a finer-grained terminology.

Simulation is often conceived in cognitive-scientific terms: one's own behavior control system is

employed as a manipulable model of other such systems. The system is first taken off-line, so that the output is not actual behavior but only predictions or anticipations of behavior, and inputs and system parameters are accordingly not limited to those that would regulate one's own behavior. Many proponents hold that, because one human behavior control system is being used to model others, general information about such systems is unnecessary. The simulation is thus said to be process-driven rather than theory-driven (Goldman).

The simulation theory is often thought to require that, to anticipate or to explain another's behavior, one has to make decisions in the role of the other--something we are not frequently aware of doing. However, decision-making, insofar as it results in a decision to perform a definite action, would always yield a definite prediction. Something short of decision-making would better account for our actual capacity to anticipate behavior, limited as it is. For people commonly allow a range of indeterminacy in their expectations of what others will do: some actions are seen as unsurprising given the person and the situation, and others as very surprising. Even if one does not make a decision in the role of the other, one can, by making adjustments in imagination, make some possible actions appear attractive (and thus unsurprising) and others unattractive (and thus surprising).

Varieties of Simulation Theory

Alvin Goldman and the psychologist Paul Harris conceive simulation differently from Robert Gordon and Jane Heal, the philosophers who, working independently, introduced the theory in 1986. According to Goldman and (less clearly) Harris, to ascribe mental states to others by simulation, one must already be able to ascribe mental states to oneself by introspection, and thus must already possess the relevant mental state concepts. Gordon holds a contrary view suggested by both Kant and Quine: Only those who can simulate can understand an ascription of, e.g., belief--that *to S* it is the case that p. While no simulation theorist claims that all our everyday explanations and predictions of the actions of other people are based on role-taking, Heal in particular has been a moderating influence, arguing for a hybrid simulation-and-theory account that reserves simulation primarily for items with rationally linked content, such as beliefs, desires, and actions.

The introspectionist account of simulation may suggest that simulation is just an application of the argument from analogy. According to one version of this argument,

> I am conscious in myself of a series of facts connected by an uniform sequence, of which the beginning is modifications of my body, the middle is feelings, the end is outward demeanour. In the case of other human beings I have the evidence of my senses for the first and last links of the series, but not for the intermediate link....by supposing the link to be of the same nature as in the case of which I have experience,...I bring other human beings, as phenomena, under the same generalizations which I know by experience to be the true theory of my own existence. -- J.S. Mill, *An Examination of Sir William Hamilton's Philosophy*. 6th edition. London, 1869.

Likewise, the "one system modeling another"account may suggest that simulation is a device for discerning what goes on "inside" another, based on an assumption of the internal similarity of the simulating system and the target system. However, where one explains or predicts another's behavior in terms of a shared, jointly known world, there is no question of internal resemblance between simulator and target, only one of *what it is about the world* that moves the other to action. What is presumed is not similarity but access, not that the other believes as one does but that the other has access to (what one presumes to be) the world.

Areas of Empirical Investigation

Three main areas of empirical investigation have been thought especially relevant to the debate:

- ***False belief.*** Taking into account another's ignorance or false belief when predicting or explaining their behavior requires imaginative modifications of one's own beliefs, according to the simulation theory. Thus the theory offers an explanation of the results of numerous experiments showing that younger children fail to take such factors into account. It would also explain the correlation, in autism, of failure to take into account ignorance or false belief and failure to engage in spontaneous pretend-play, particularly role play. Although these results can also be explained by certain versions of theory theory (and were so interpreted by the experimenters themselves), the simulation theory offers a new interpretation.
- ***Priority of self- or other-ascription.*** A second area of developmental research asks whether children ascribe mental states to themselves before they ascribe them to others. Versions of the simulation theory committed to the view that we recognize our own mental states as such and make analogical inferences to others' mental states seem to require an affirmative answer to this question; other versions of the theory seem to require a negative answer. Some experiments suggest a negative answer, but debate continues on this question.
- ***Cognitive impenetrability.*** Stich and Nichols suppose simulation to be "cognitively impenetrable" in that it operates independently of any general knowledge the simulator may have about human psychology. Yet they point to results suggesting that when subjects lack certain psychological information, they sometimes make incorrect predictions, and therefore must not be simulating. Because of problems of methodology and interpretation, as noted by a number of philosophers and psychologists, the cogency of this line of criticism is unclear.

The numerous other empirical questions of possible relevance to the debate include the following:

> Does brain imaging reveal that systems and processes employed in decision-making are reemployed in the explanation and prediction of others' behavior?
>
> Does narrative (including film narrative) create emotional and motivational effects by the same processes that create them in real-life situations?

Some philosophers think the simulation theory may shed light on issues in traditional philosophy of mind and language concerning intentionality, referential opacity, broad and narrow content, the nature of mental causation, Twin Earth problems, the problem of other minds, and the peculiarities of self-knowledge. Several philosophers have applied the theory to aesthetics, ethics, and philosophy of the social sciences. Success or failure of these efforts to answer philosophical problems may be considered empirical tests of the theory, in a suitably broad sense of "empirical."

Bibliography

Principal Sources:

- Goldman, A., 1989, "Interpretation Psychologized." *Mind and Language* 4, 161-185; reprinted in Davies, M. and Stone T., eds., 1995, Folk Psychology: The Theory of Mind Debate. Oxford: Blackwell Publishers.
- Gordon, Robert M., 1986, "Folk Psychology as Simulation", *Mind and Language* 1, 158-171; reprinted in Davies, M. and Stone T., eds., 1995, Folk Psychology: The Theory of Mind Debate. Oxford: Blackwell Publishers.
- Harris, P., 1989, *Children and Emotion*, Oxford: Blackwell Publishers.
- Heal, J., 1986, "Replication and Functionalism", in *Language, Mind, and Logic*, J. Butterfield (ed.), Cambridge: Cambridge University Press; reprinted in Davies, M. and Stone T., eds., 1995, Folk Psychology: The Theory of Mind Debate. Oxford: Blackwell Publishers.

Collections:

- Carruthers, P. & Smith, P., eds., 1996, *Theories of Theories of Mind*. Cambridge: Cambridge University Press.
- Davies, M. and Stone T., eds., 1995, *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell Publishers. (The introductory chapter offers an excellent overview and analysis of the initial debate.)
- Davies, M. and Stone T., eds., 1995, *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell Publishers.

Further Readings

- Goldman, A., 1993, ‘The Psychology of Folk Psychology,’ *The Behavioral and Brain Sciences*, 16: 15-28.
- Gordon, R. M., and J. Barker, 1994, ‘Autism and the "theory of mind" debate.’ In *Philosophical Psychopathology: A Book of Readings*, G. Graham and L. Stephens, eds. MIT Press, pp. 163-181.
- Gordon, R.M., 1995, ‘Sympathy, Simulation, and the Impartial Spectator,’ *Ethics* 105:727-742. Reprinted in *Mind and Morals: Essays on Ethics and Cognitive Science*, L. May, M. Friedman, & A. Clark, eds. MIT Press, 1996.

- Harris, P., 1989, *Children and Emotion*. Oxford: Blackwell Publishers.
- Peacocke, C., ed., 1994, *Objectivity, Simulation, and the Unity of Consciousness*. Oxford: Oxford University Press.
- Perner, J., 1991, *Understanding the Representational Mind*. Cambridge, MA: MIT Press.
- Wellman, H. M., 1990, *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Other Internet Resources

- NEH Summer Seminar on Simulation Theory page, maintained by Nigel Thomas

Related Entries

folk psychology: as a theory | materialism: eliminative

Acknowledgment

A large portion of this entry is excerpted, with permission, from "Simulation vs Theory Theory", *MIT Encyclopedia of Cognitive Science* (MIT Press, 1999

Robert M. Gordon
gordon@umsl.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 8, 1997
Content last modified: March 8, 2001

Wilfrid Sellars

Wilfrid Stalker Sellars (b. 1912, d. 1989) was a profoundly creative and synthetic thinker whose work both as a systematic philosopher and as an influential editor helped set and shape the Anglo-American philosophical agenda for over four decades. Sellars is perhaps best known for his classic 1956 essay "Empiricism and the Philosophy of Mind", a comprehensive and sophisticated critique of "the myth of the given" which played a major role in the postwar deconstruction of Cartesianism, but his published corpus of three books and more than one hundred essays includes numerous original contributions to ontology, epistemology, and the philosophies of science, language, and mind, as well as sensitive historical and exegetical studies.

Sellars' Life and Career

- 1912 born May 20 in Ann Arbor, MI
- 1933 receives AB at the University of Michigan
- 1934 receives AM at the University of Buffalo, NY, enters Oriel College, Oxford, as a Rhodes Scholar
- 1936 receives BA with First Class Honours in Philosophy, Politics, and Economics (MA 1940)
- 1938 becomes Assistant Professor of Philosophy, University of Iowa
- 1943 enters U.S. Naval Reserve, assigned to Air Combat Intelligence
- 1946 becomes Assistant Professor of Philosophy, University of Minnesota
- 1950 founds *Philosophical Studies* with Herbert Feigl, the first scholarly forum explicitly created

for the new hybrid "analytic philosophy"
- 1951 becomes Professor of Philosophy, University of Minnesota
- 1956 serves as Special Lecturer in Philosophy at the University of London, published as "Empiricism and the Philosophy of Mind"
- 1958 moves to Yale University, CN, first as a visitor, subsequently as Professor of Philosophy
- 1963 assumes the position of University Professor of Philosophy and Research Professor of Philosophy at the University of Pittsburgh, PA, publishes *Science, Perception and Reality*
- 1965 delivers John Locke Lectures for 1965-66 at Oxford University, subsequently published as *Science and Metaphysics*
- 1970 serves as president of the Eastern Division of the American Philosophical Association
- 1971 delivers Matchette Foundation Lectures, University of Texas, subsequently published as "The Structure of Knowledge"
- 1973 delivers John Dewey Lectures for 1973-74, University of Chicago, IL, subsequently published as *Naturalism and Ontology*
- 1977 delivers Paul Carus Lectures for 1977-78 at the Eastern Division meetings of the American Philosophical Association, later published as "Foundations for a Metaphysics of Pure Process"
- 1987 Colloquium in Sellarsian Philosophy held at University of Pittsburgh in honor of Sellars' 75th birthday
- 1989 dies at home in Pittsburgh, PA, on July 2

Sellars' Metaphilosophy

Although Wilfrid Sellars is best known for his his ground-breaking essay "Empiricism and the Philosophy of Mind" [EPM] and his critique of what he there called "the myth of the given", he was in fact a systematic philosopher *par excellence*. "The aim of philosophy," he wrote, "is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term." [PSIM, 37] This image of the philosopher as a *reflective generalist* recurs frequently in Sellars' metaphilosophical reflections. His most explicit account of the central task confronting contemporary philosophy aligns it firmly with the modernist project of achieving a *rapprochement* between our humanistic understanding of ourselves as free and rational agents, at home among meanings and values, and the thoroughly "disenchanted" picture of the world being painted by an increasingly comprehensive natural science. Sellars thematized this contrast as a confrontation of two "images": the "*manifest image*" whose primary objects are *persons*, beings who can and do conceive of themselves as sentient perceivers, cognitive knowers, and deliberative agents, and the "*scientific image*", whose primary entities are some sophisticated version of "atoms in the void". "The scientific image," Sellars wrote, "presents itself as a *rival* image. From its point of view the manifest image on which it [methodologically] rests is an ‘inadequate’ but pragmatically useful likeness of a reality which first finds its adequate (in principle) likeness in the scientific image." [PSIM, 57] As Sellars saw it, the goal of philosophy was to transform this tension between our lived self-conception and our hard won explanatory understanding of the world into a single "stereoscopic" image, a *synoptic* vision of persons-in-the-world. Much of his philosophical work is addressed to three central moments of this complex undertaking: accommodating the *intentional contents* of thought and language, the *sensuous contents* of perception and imagination, and the

normative dimensions of knowledge and conduct within such a stereoscopic image - all the while resolutely maintaining a robust scientific realism, for "in the dimension of describing and explaining the world, science is the measure of all things, of what is that it is, and of what is not that it is not." [EPM, 173]

Sellars' Philosophy of Science and Epistemology

Sellars' interpretation of the epistemology of natural science departed decisively from the received view according to which explanation was identified with derivation - singular matters of empirical fact being explained by deriving descriptions of them from ("inductively" established) empirical generalizations (along with appropriate statements of initial conditions), and these "empirical laws" in turn being explained by deriving them from theoretical postulates and correspondence rules. On this received Positivist view, theories (e.g., microtheories) explain empirical matters of fact only indirectly, by implying generalizations framed in an observation-language that explain them directly. In consequence, as Hempel pointed out in "The Theoretician's Dilemma", such theories, although perhaps convenient aids to calculation and compact representation, are in principle utterly dispensable.

Sellars regarded this "layer-cake model" or "levels picture" of theories as fundamentally misguided. He argued that there is no autonomous stratum of empirical counterparts to theoretical laws. The empirical generalizations corresponding to theoretical laws become salient only from the theoretical perspective. Generalizations arrived at autonomously at the observational level, however reliable, are not laws of nature, and theories consequently cannot be in the business of explaining such lower-level generalizations by entailing them. Rather, "theories explain laws by explaining why the objects of the domain in question obey the laws that they do to the extent that they do." (LT,123)

> [That is,] they explain why individual objects of various kinds and in various circumstances in the observation framework behave in those ways in which it has been inductively established that they do behave. Roughly, it is because a gas is ... a cloud of molecules which are behaving in certain theoretically defined ways, that it obeys the empirical Boyle-Charles Law. (LT,121)

On Sellars' view stories that postulate "theoretical entities" are not merely manageable second-class surrogates for more complicated and unwieldy stories about entities that we have good, i.e., observational, reasons to believe actually exist. Theoretical entities, rather, are those entities we warrantedly believe to exist for good and sufficient *theoretical* reasons. On this understanding, scientific theories explanatorily "save the appearances" precisely by characterizing the reality of which the appearances are appearances.

Like Quine, Sellars was deeply influenced by the work of Rudolf Carnap. Sellars' sophisticated account of the nature and import of theoretical reasoning in natural science, however, enabled him to develop a systematic naturalistic *alternative* to Quine's influential critique of Carnapian *logical empiricism*. In particular, the epistemological contrast between two sorts of empirical generalizations - those adopted on

narrowly inductive grounds and those expressing constitutive principles of postulational theories adopted on broadly empirical, i.e., explanatory grounds - enabled Sellars to distinguish among three different grades of "observational involvement": observations and general claims individually validated "inductively" by way of *direct* appeals to observational backing, the constitutive posits of postulational theories holistically validated by way of *indirect, explanatory* appeals to observational backing, and purely formal claims expressing necessary conditions for the formulation of scientific hypotheses in general. Consequently, where Quine rejected the classical Kantian analytic-synthetic dichotomy out of hand, Sellars argued that there were two quite different distinctions tangled up in the single dichotomy that Carnap had inherited from the Kantian tradition: the distinction between logical and empirical (matter-of-factual) claims (analytic$_2$-synthetic$_2$), and the distinction between claims whose revision requires abandonment or modification of the system of (theoretical) concepts in terms of which they are framed and claims revisable on the basis of observations formulated in terms of a system of (theoretical) concepts which remained fixed throughout (analytic$_1$-synthetic$_1$). Like Quine, then, Sellars moved decisively away from classical Kantian rationalism, but in the direction of a *Kantian* empiricism which preserved logical space for a theory of semantic meaning and the correlative distinctions between individual matter-of-factual truths and truths which, although belonging to theoretical systems themselves adopted on broadly empirical (synthetic$_2$) grounds, were, relative to such a system, true *ex vi terminorum* (analytic$_1$):

Kant's Rationalism

Grounded in experience ("a posteriori", simple induction)	*Not so grounded* ("a priori")	
Synthetic		**Analytic**
Empirical Laws (regularities)	Arithmetic, Geometry, Mechanics ("synthetic a priori")	Logic
	"Our conceptual framework" (innate principles)	

Kantian Empiricism

Grounded in experience (Empirical)		*Not so grounded*
Synthetic$_2$		**Analytic$_2$ (L-true)**
Synthetic$_1$	**Analytic$_1$**	
Observation, Simple Induction (Operational geometry, mechanics)	Postulation (Physical geometry, idealizing scientific theories, mechanics, micro-physics)	Logic, arithmetic, mathematical analysis (Pure geometry *qua* calculus)

Sellars' Philosophy of Language and Mind

Essential to Sellars' thoroughgoing naturalism is an account of semantic meaning that requires no recourse to irreducibly platonistic or mentalistic idioms. Sellars consequently resolutely locates the normative conceptual order within the causal order and advances a naturalistic interpretation of the modes of causality exercised by linguistic rules centered on the notion of *pattern-governed* behavior, i.e.:

> behavior which exhibits a pattern, not because it is brought about by the intention that it exhibit this pattern, but because the propensity to emit behavior of the pattern has been selectively reinforced, and the propensity to emit behavior which does not conform to this pattern selectively extinguished. (MFC,423)

Pattern-governed behavior characteristic of a species - e.g., the dance of the bees - can arise from processes of natural selection on an evolutionary time scale, but, crucially, pattern-governed behavior can also be developed in individual "trainees" by deliberate selective reinforcement on the part of other individuals, the trainers, acting under the guidance of linguistic *rules of criticism*. In contrast to linguistic *rules of action*, e.g., "Ceteris paribus, one ought to (or: may) say such and such if in circumstances C", which can be efficacious in guiding linguistic activity only to the extent that their subjects already possess the concepts of "saying such-and-such", "being in circumstances C", and, indeed, obeying a rule (i.e., doing something because it is enjoined or permitted by a rule), rules of criticism are ought-to-be's - e.g., "Westminster clock chimes ought to strike on the quarter hour" (LTC,95) - whose subjects, although their performances may be assessed according to such rules, need not themselves have the concept of a rule nor, indeed, any concepts at all. Thus a trainer can be construed as reasoning

> Patterned-behavior of such and such a kind ought to be exhibited by trainees, hence we, the trainers, ought to do this and that, as likely to bring it about that it is exhibited. (MFC,423)

And, in consequence of the conducts of trainers under the guidance of such rules of action, the behavior of a language-learner can come to *conform* to the relevant rules of criticism without his "grasping" them himself in any other sense. "Trainees conform to ought-to-be's because trainers obey corresponding ought-to-do's." (MFC,423)

Against this background, then, Sellars advanced an account of meaning as *functional classification* according to which semantic idioms in the first instance mark contexts within which structurally distinct "natural-linguistic objects" (e.g., utterings or inscribings) are classified in terms of their roles or functions in *language entry transitions* (linguistic responses to perceptual stimuli), *language exit transitions* (causal-linguistic antecedents of non-linguistic conduct), and *intra-linguistic moves* (inferential

transitions from one linguistic representing to another). In particular, ‘means’ is interpreted as a specialized form of the copula, tailored to metalinguistic contexts, according to which the right side of the superficially relational form "___ means …" is properly understood as mentioning or exhibiting a linguistic item.

On Sellars' view, such special copulae and metalinguistic indicators initially arise in response to the need to abstract from our domestic sign designs in order to classify items of different languages on the basis of such functional criteria. In this project, ordinary quotation suffers from a systematic ambiguity regarding the criteria - structural (e.g., geometric, acoustic) or functional - according to which linguistic tokens are classifiable as belonging to this or that linguistic type. Accordingly, Sellars introduced a more straightforward device of two separate styles of quotation marks, star-quotes and dot-quotes, tied respectively to the structural and functional modes of sorting and individuating lexical items. Both star- and dot-quotes are illustrating, and thus indexical, devices, but dot-quotes are, in a sense, doubly so. For, whereas star-quotes form a common noun that is true of inscriptions (empirical structures) appropriately design-isomorphic to the token exhibited between them, dot-quotes form a common noun true of items in any language that play the role or do the job performed in our language by the tokens exhibited between them. In terms of this notational apparatus, then, such semantic claims as, for example,

(1s) (In German) ‘rot’ means red.
(2s) (In German) ‘Schnee ist weiss’ means snow is white.

can be more perspicuously expressed by

(1*) (In the German linguistic community) *rot*s are .red.s.
(2*) (In the German linguistic community) *Schnee ist weiss*s are .snow is white.s.

Once such a distinction between functional and structural classification of linguistic representing items is in hand, it is a straightforward matter to extend it to an account of *mental* representations, i.e., *thoughts*, as well. Unlike Quine, Sellars never abandoned the classical notion of thoughts as intentional *inner episodes* that play a causal-explanatory role vis-à-vis overt, paradigmatically linguistic, behavior. Consistent with his thoroughgoing naturalism, however, correlative to his ontological "linguistic nominalism", Sellars embraced a form of "psychological nominalism", whose leitmotif was

> . . . the denial of the claim, characteristic of the realist tradition, that a "perception" or "awareness" of abstract entities is the root mental ingredient of mental acts and dispositions. (EAE,445)

Instead, Sellars argued, the proper account of the distinctive intentionality of thought is also to be drawn in terms of the forms and functions of natural linguistic items. The positive thesis correlative to psychological nominalism, consequently, is modeled by what Sellars came to call "verbal behaviorism".

> According to VB [verbal behaviorism], thinking ‘that-p,’ where this means ‘having the

> thought occur to one that-p,' has as its primary sense [an event of] saying 'p'; and a secondary sense in which it stands for a short term proximate propensity [dispositional] to say 'p'. (MFC,419)

The origins of Sellars' mature forms of verbal behaviorism lie in the revolutionary theses of his classic essay "Empiricism and the Philosophy of Mind", and, in particular, in his mythical story of our Rylean ancestors and the genius Jones. The story begins *in medias res* with people who have mastered a "Rylean language", a sophisticated expressive system, including logical operators and subjunctive conditionals, whose fundamental descriptive vocabulary pertains to public spatio-temporal objects. Consonant with the Sellarsian account of linguistic meaning as functional classification, this hypothetical Rylean language, although lacking any resources for speaking of inner episodes, thoughts or experiences has been enriched by the fundamental resources of semantical discourse - enabling our ancestors to say of the their peers' utterances that they mean this or that, that they stand in various logical relations to one another, that they are true or false, and so on. In this milieu now appears the genius Jones.

> [In] the attempt to account for the fact that his fellow men behave intelligently not only when their conduct is threaded on a string of overt verbal episodes . . . but also when no detectable verbal output is present, Jones develops a theory according to which overt utterances are but the culmination of a process which begins with certain inner episodes. . . . [His] model for these episodes which initiate the events which culminate in overt verbal behavior is that of overt verbal behavior itself. (EPM,186)

Although the primary use of semantical terms remains the semantical characterization of overt verbal episodes, this Jonesean theory thus carries over the applicability of those semantical categories to its postulated inner episodes. i.e., to (occurrent) thoughts. The point of the Jonesean myth is to suggest that the *epistemological* status of thoughts (qua inner episodes) vis-à-vis candid public verbal performances is most usefully understood as *analogous* to the epistemological status of, e.g., molecules vis-à-vis the public observable behavior of gases.

> [Thought] episodes are 'in' language-using animals as molecular impacts are 'in' gases, not as 'ghosts' are in 'machines'. (EPM,187)

Unlike molecules, however, which are introduced into kinetic gas theory as having a specific empirical character (represented by the posited essentially Newtonian lawfulness of their dynamic interactions), the thought episodes postulated by that theory as covert states of persons are introduced by a *purely functional* analogy. The concept of an occurrent thought is that of a causally-mediating logico-semantic role player, whose determinate empirical/ontological character, and thereby logical space for some form of "identity theory" is so far left open.

> [The] fact that [thoughts] are not introduced as physiological entities does not preclude the possibility that at a later methodological stage they may, so to speak, 'turn out' to be such. Thus, there are many who would say that it is already reasonable to suppose that these

> thoughts are to be ‘identified’ with complex events in the cerebral cortex . . . (EPM,187-8)

Since, on Sellars' account, the concept of a thought is fundamentally the concept of a functional kind, no ontological tensions would be generated by the identification within the scientific image of items belonging to that functional kind with, for instance, states and episodes of an organism's central nervous system. The manifest image's conception of person as thinkers, Sellars concludes, can fuse smoothly with the scientific image's conception of persons as complex material organisms having a determinate physiological and neurological structure.

The idea that the intentionality of the mental is to be understood in terms of epistemologically theoretical transpositions of the semantic categories of public language, themselves interpreted as modes of *functional* classification earn Sellars a definitive place in contemporary analytic philosophy of mind. As Dennett puts it,

> Thus was contemporary functionalism in the philosophy of mind born, and the varieties of functionalism we have subsequently seen are in one way or another enabled, and directly or indirectly inspired, by what was left open in Sellars' initial proposal ... (MTE,341)

Sellars' proposal that we can illuminate the epistemic status of mental concepts by an appeal to the contrast between theoretical and non-theoretical discourse makes sense only against the background of another central element of his philosophical thought, his comprehensive critique of the "myth of the given". The philosophical framework of givenness historically takes on many guises, including not only the idea that empirical knowledge rests on a foundation, but also, crucially, the assumption that the "privacy" of the mental and one's "privileged access" to one's own mental states are fundamental features of experience, both logically and epistemologically prior to all intersubjective concepts pertaining to inner episodes.

Sellars argues, on the contrary, that what begins in the case of inner episodes as a language with a purely theoretical use can acquire a first-person reporting role. It can turn out to be possible to train people, in essence by a process of operant conditioning, to have "privileged access" to some of their inner episodes, that is, to respond directly and non-inferentially to the occurrence of one thought with another (meta-) thought to the effect that one is thinking it. It is a special virtue of this aspect of Sellars' Jonesean story that it shows how the essential intersubjectivity of language can be reconciled with the "privacy" of inner episodes, i.e.,

> . . . that it helps us understand that concepts pertaining to such inner episodes as thoughts are primarily and essentially inter-subjective, as inter-subjective as the concept of a positron, and that the [first-person] reporting role of these concepts . . . constitutes a dimension of [their] use . . . which is built on and presupposes this inter-subjective status. (EPM,189)

At the heart of Sellars' general case against the Myth of the Given is his articulate recognition of the

irreducibly normative character of epistemic discourse.

> The essential point is that in characterizing an episode or a state as that of knowing, we are not giving an empirical description of that episode or state, we are placing it in the logical space of reasons, of justifying and being able to justify what one says. (EPM,169)

Once it is acknowledged that the senses per se grasp no facts, that all knowledge that something is such-and-so (all "subsumption of particulars under universals") presupposes learning, concept formation, and even symbolic representation, it follows that ". . . instead of coming to have a concept of something because we have noticed that sort of thing, to have the ability to notice a sort of thing is already to have the concept of that sort of thing, and cannot account for it." (EPM,176)

Sellars follows Kant in rejecting the Cartesian picture of a sensory-cognitive continuum. The "of-ness" of *sensations* - e.g., a sensation's being of a red triangle or of a sharp shooting pain - he insists, is not the intentional "of-ness" ("aboutness") of thoughts. The "rawness" of "raw feels" is rather their *non-conceptual* character. (cf. IAMBP,376) Consequently, while his *epistemological* views regarding sensory episodes parallel his treatment of the epistemology of occurrent thoughts, Sellars' account of the *ontology* of sensations diverges dramatically from his functionalist account of thoughts.

In a final episode of the Jonesean myth, sensations are introduced as elements of an explanatory account of the occurrence in various circumstances of perceptual cognitions, having determinate semantic contents:

> . . . the hero . . . postulates a class of inner - theoretical - episodes which he calls, say, impressions, and which are the end results of the impingement of physical objects and processes on various parts of the body. . . (EPM,191)

This time, however, the model for Jones' theory is not that of functionally-individuated families of sentences, but rather "a domain of ‘inner replicas’ which, when brought about in standard conditions share the perceptible characteristics of their physical sources". (EPM,191) The leading idea of this model is the occurrence, ‘in’ perceivers of "replicas" *per se*, not of perceivings of "replicas" (which would mistakenly inject into the account of impressions the intentionality of thought), and, although the entities of this *model* are particulars, the entities introduced by the theory are not particulars but rather states of a perceiving subject. Thus, although talk of the "of-ness" of sensations, like that of the "of-ness" of thoughts is, on Sellars' view, fundamentally classificatory, the classification at issue is based not on a functional (logical, semantic) analogy but rather on analogies that, although in the first instance extrinsic and causal, ultimately attribute to sensations a determinate intrinsic content. The specific point of the model is to insist that states of, e.g., sensing [red triangle]ly (to highlight the status of ‘sensation’ as a "verbal noun"), characteristically brought about in normal perceivers in standard conditions by the action of red triangular objects on the eyes, can discharge their explanatory jobs in relation to cognitive perceptual takings (especially non-veridical perceptual judgments) only if they are conceived as resembling and differing from other sensory states - e.g., sensing [green triangular]ly, sensing [red

square]ly, etc. - in a manner formally analogous to the way in which objects of the "replica" model - e.g., red and triangular, green and triangular, and red and square "wafers" - are conceived to resemble and differ from one another.

If that were the end of Sellars' ontological story regarding sensations, matters would be complicated enough. But Sellars proceeds to develop this core account in a variety of different directions, in consequence of which his full theory of sensations has emerged as being one of the most difficult and controversial aspects of his philosophy.

The first complication of Sellars' theory of sensation results from his conviction that, in the case of sensations, Jones' theory is *interpretive*. It does not introduce new domains of entities, but rather reinterprets the categorial/ontological status of sensory contents as states of perceivers. The crux of the original Jonesean theory that the very color quanta of which we are perceptually aware *as* existing in space are instead actually states of persons-qua-perceivers. Already within the manifest image, then, the ontological status ultimately accorded to sensory "content qualia" is incompatible with their being instantiated in physical space.

The second complication of Sellars' theory of sensations arises from the further conclusion that it is this manifest image conception of sensory contents as states of perceivers which must ultimately be synoptically "fused" with the scientific image, and that the latter's commitment to the idea that those perceivers themselves are complex systems of micro-physical particles constitutes a barrier to doing so in any straightforward way. Sellars notoriously concludes that sensory contents can be synoptically integrated into the scientific image only after both they and the currently-fundamental micro-physical particulars of that image as well undergo yet another categorial transposition into a categorially monistic ontology whose fundamental entities are all "absolute processes". Sensings qua absolute processes would then be physical, he writes,

> . . . not only in the weak sense of not being mental (i.e., conceptual), for they lack intentionality, but in the richer sense of playing a genuine causal role in the behavior of sentient organisms. They would, as I have used the terms, be physical-l but not physical-2. Not being epiphenomenal, they would conform to a basic metaphysical intuition: to be is to make a difference. (CL,III,126)

A Final Remark

Lengthy as this discussion has been, it only begins to capture the scope, depth, and systematic character of Sellars' philosophical accomplishments. Many themes from his work have simply gone unmentioned - his anticipation of epistemological externalism and defense of a strong internalist alternative, his insightful analysis of predication and correlative nominalistic alternative to classical Platonistic categorial ontology, his sophisticated account of induction as a form of vindicatory practical reasoning, his significant contributions to ethical theory and the theory of action, and his masterful interpretations of the work of many of the discipline's great historical figures, not as scholarly museum exhibits, but always

as active participants in a continuing philosophical conversation. The bibliographies and Internet resources listed below will point the way to both more comprehensive and more detailed accounts of the work of this towering philosophical figure of the postwar era.

Principal Works by Wilfrid Sellars

Books

- *Pure Pragmatics and Possible Worlds-The Early Essays of Wilfrid Sellars*, [PPPW], ed. by Jeffrey F. Sicha, (Ridgeview Publishing Co; Atascadero, CA; 1980). [Contains a long introductory essay by Sicha and an extensive bibliography of Sellars' work through 1979.]
- *Science, Perception and Reality*, [SPR], (Routledge & Kegan Paul Ltd; London, and The Humanities Press: New York; 1963) [Reissued in 1991 by Ridgeview Publishing Co., Atascadero, CA. This edition contains a complete bibliography of Sellars' published work through 1989.]
- *Philosophical Perspectives*, [PP], (Charles C. Thomas: Springfield, IL; 1967). Reprinted in two volumes, *Philosophical Perspectives: History of Philosophy* and *Philosophical Perspective: Metaphysics and Epistemology*, (Ridgeview Publishing Co.; Atascadero, CA; 1977).
- *Science and Metaphysics: Variations on Kantian Themes*. [S&M], (Routledge & Kegan Paul Ltd; London, and The Humanities Press; New York; 1968). The 1966 John Locke Lectures. [Reissued in 1992 by Ridgeview Publishing Co., Atascadero, CA. This edition contains a complete bibliography of Sellars' published work through 1989, a register of Sellars' philosophical correspondence, and a listing of circulated but unpublished papers and lectures.]
- *Essays in Philosophy and Its History*, [EPH], (D. Reidel Publishing Co.; Dordrecht, Holland; 1975).
- *Naturalism and Ontology*, [N&O], (Ridgeview Publishing Co.; Atascadero, CA: 1979). [An expanded version of the 1974 John Dewey Lectures]
- *The Metaphysics of Epistemology, Lectures by Wilfrid Sellars*, edited by Pedro Amaral, (Ridgeview Publishing Co.; Atascadero, CA; 1989). [Contains a complete bibliography of Sellars' published work through 1989.]
- *Empiricism and the Philosophy of Mind* [EPM*], edited by Robert Brandom, (Harvard University Press.; Cambridge, MA; 1997). [The original, 1956, version of [EPM] (see below), lacking footnotes added in [SPR], with an Introduction by Richard Rorty and Study Guide by Brandom.]

Selected Essays

- [AAE], "Actions and Events", *Noûs* 7, 1973, pp. 179-202.
- [AE], "Abstract Entities", *Review of Metaphysics* 16, 1983; reprinted in [PP], pp. 229-69.
- [CDCM], "Counterfactuals, Dispositions, and the Causal Modalities", in *Minnesota Studies in the Philosophy of Science*, Vol. II, ed. by H. Feigl, M. Scriven, and G. Maxwell, (University of Minnesota Press; Minneapolis, MN: 1957), pp. 225-308.
- [CL], "Foundations for a Metaphysics of Pure Process", The Carus Lectures for 1977-78,

published in *The Monist* 64, No. 1, 1981.

- [EAE], "Empiricism and Abstract Entities", in *The Philosophy of Rudolph Carnap*, ed. by P.A. Schilpp (Open Court; LaSalle, IL; 1963); reprinted in [EPH], pp. 245-86.
- [EPM], "Empiricism and the Philosophy of Mind", in *The Foundations of Science and the Concepts of Psychoanalysis, Minnesota Studies in the Philosophy of Science, Vol. I*, ed. by H. Feigl and M. Scriven (University of Minnesota Press; Minneapolis, MN; 1956); reprinted in [SPR], pp. 127-96).
- [FD], "Fatalism and Determinism", in Keith Lehrer, ed., *Freedom and Determinism*, (Random House; New York, NY: 1966), pp. 141-74.
- [GEC], "Givenness and Explanatory Coherence", *Journal of Philosophy* 70, 1973, pp. 612-24.
- [I], "…this I or he or it (the thing) which thinks", the 1970 Presidential Address, American Philosophical Association (Eastern Division), reprinted in [EPH].
- [IAMBP], "The Identity Approach to the Mind-Body Problem", *Review of Metaphysics* 18, 1965; reprinted in [PP], pp. 370-88.
- [IKTE], "The Role of Imagination in Kant's Theory of Experience", The 1977 Dotterer Lecture, in H.W. Johnstone, Jr., ed., *Categories: A Colloquium*, (Pennsylvania State University Press: 1977), pp. 231-45.
- [IV], "Induction as Vindication", *Philosophy of Science* 31, 1964; reprinted in [EPH], pp. 367-416.
- [ISRT], "Is Scientific Realism Tenable", *Proceedings of the PSA*, Volume 2, 1976, pp. 307-34.
- [KTE], "Some Remarks on Kant's Theory of Experience", *Journal of Philosophy* 64, 1967, pp. 633-47.
- [LT], "The Language of Theories", in *Current Issues in the Philosophy Science*, ed. by H. Feigl and G. Maxwell (Henry Holt, Rhinehart and Winston; New York, NY; 1961): reprinted in [SPR], pp. 106-26.
- [LTC],"Language as Thought and Communication", *Philosophy and Phenomenological Research* 29. 1969; reprinted in [EPH], pp. 93-117.
- [MFC], "Meaning as Functional Classification", *Synthese* 27, 1974; pp. 417-37. (Issue also contains comments by Daniel Dennett and Hilary Putnam and Sellars' replies.)
- [MEV], "Mental Events", *Philosophical Studies* 81, 1981; pp. 325-45.
- [MGEC], "More on Givenness and Explanatory Coherence", in George S. Pappas, ed., *Justification and Knowledge*, (D. Reidel Publishing Co.; Dordrecht, Holland: 1979), pp. 169-82.
- [NDL], "Are There Non-Deductive Logics?", in N. Rescher *et al*, eds., *Essays in Honor of Carl G. Hempel*, Synthese Library, (D. Reidel Publishing Co.; Dordrecht, Holland: 1970), pp. 83-103.
- [OAFP], "On Accepting First Principles", in J. Tomberlin, ed., *Philosophical Perspectives 2: Epistemology, 1988*, (Ridgeview Publishing Co.; Atascadero, CA: 1988), pp. 301-14.
- [P], "Phenomenalism", in [SPR], pp. 60-105.
- [PSIM], "Philosophy and the Scientific Image of Man", in *Frontiers of Science and Philosophy*, ed. by Robert Colodny (University of Pittsburgh Press; Pittsburgh, PA; 1962); reprinted in [SPR], pp. 1-40.
- [SK], "The Structure of Knowledge", The Matchette Foundation Lectures for 1971, published in Castañeda, ed., *Action, Knowledge, and Reality* (see below).
- [SSMB], "A Semantical Solution of the Mind-Body Problem", *Methodos* 5, 1953, pp. 45-82.

Reprinted in [PPPW].

- [TA], "Thought and Action", in Keith Lehrer, ed., *Freedom and Determinism*, (Random House; New York, NY: 1966), pp. 105-39.
- [TWO], "Time and the World Order", in *Minnesota Studies in the Philosophy of Science*, Vol. III, ed. by H. Feigl and G. Maxwell, (University of Minnesota Press; Minneapolis, MN: 1962), pp. 527-616.

Bibliography

Major Critical Studies

- Castañeda, H-N., ed. *Action, Knowledge, and Reality* [AK&R] (Bobbs-Merrill; Indianapolis, IN; 1975). [Also contains an extensive bibliography of Sellars' work through 1974, Sellars' intellectual autobiography, and 'The Structure of Knowledge' (see above).]
- deVries, Willem A., and Timm Triplett, *Knowledge, Mind, and the Given: Reading Wilfrid Sellars' "Empiricism and the Philosophy of Mind"*, (Hackett Publishing Co.; Indianapolis, IN & Cambridge, MA; 2000). [A detailed commentary on [EPM] (see above), including the complete text as published with additional footnotes in [SPR], 1963. The best general introduction to Sellars' classic essay.]
- Delaney, C.F., Michael J. Loux, Gary Gutting, and W. David Solomon, *The Synoptic Vision: Essays on the Philosophy of Wilfrid Sellars* (University of Notre Dame Press; Notre Dame. IN; 1977). [Also contains an extensive bibliography.]
- Pitt, Joseph C., ed., *The Philosophy of Wilfrid Sellars: Queries and Extensions* [PSQE] (D. Reidel Publishing Co; Dordrecht, Holland; 1978). [Revised proceedings of a workshop on the Philosophy of Wilfrid Sellars held at Virginia Polytechnic Institute and State University in Blacksburg, VA, in November 1976.]
- Pitt, Joseph C., *Pictures, Images, and Conceptual Change: An Analysis of Wilfrid Sellars' Philosophy of Science* (D. Reidel Publishing Co.; Dordrecht, Holland; 1981).
- Seibt, Johanna, *Properties as Processes*, A Synoptic Study of Wilfrid Sellars' Nominalism", (Ridgeview Publishing Co.; Atascadero, CA; 1990.

- *Noûs*, Vol. 7, No. 2, 1973. [Special issue devoted to the philosophy of Wilfrid Sellars.]
- *The Monist*, Vol. 65, No. 3, 1982. [Issue devoted to the philosophy of Wilfrid Sellars.]
- *Philosophical Studies*, Vol. 54, No. 2, 1988. [Revised proceedings of the colloquium on Sellars' philosophy held in October 1987 at the University of Pittburgh's Center for Philosophy of Science.]
- *Philosophical Studies*, Vol. 101, Nos. 2-3, 2000. [Special issue devoted to the philosophy of Wilfrid Sellars.]

Supplementary Bibliography

- **Alanen, L.**, "Thought-Talk: Descartes and Sellars on Intentionality", *American Philosophical Quarterly*, 29, 1992, pp. 19-34.
- **Aune, Bruce**, "Sellars' Two Images of the World", *Journal of Philosophy*, 87, 1990, pp. 537-45.
- **Bernstein, Richard J.**, "Sellars' Vision of Man-in-the-Universe", *Review of Metaphysics*, 20, 1965-66, pp. 290-316.
- **Brandom, Robert**, *Making It Explicit*, (Harvard University Press; Cambridge, MA; 1995).
- ____________, “Study Guide”, in EPM* (see above).
- ____________, *Articulating Reasons: An Introduction to Inferentialism*, (Harvard University Press; Cambridge, MA;, 2000).
- **Clark, Romane**, "Sensibility and Understanding: The Given of Wilfrid Sellars", *The Monist*, 65, 1982, 350-64.
- **Cornman, James**, "Sellars, Scientific Realism, and Sensa", *Review of Metaphysics*, 23, 1969-70, pp. 417-51.
- ____________, "Sellars on Scientific Realism and Perceiving", in *Proceedings of the PSA*, Volume 2, ed. by F. Suppe and P.D. Asquith, 1976, pp. 344-58.
- **Dennett, Daniel C.**,[MTE], "Mid-Term Examination: Compare and Contrast", in *The Intentional Stance* (Bradford Books, The MIT Press; Cambridge, MA; 1987), pp. 339-50.
- **Echelbarger, Charles**, "Sellars on Thinking and the Myth of the Given", *Philosophical Studies* 25, 1974, pp. 231-46.
- ____________, “An Alleged Legend”, *Philosophical Studies*, 39, 1981, pp. 227-46.
- **Garfield, Jay**, “The Myth of Jones and the Mirror of Nature: Reflections on Introspection”, *Philosophy and Phenomenological Research*, 50, 1989, pp. 1-23.
- **Geiger, L.**, *Die Logik der seelischen Ereignisse. Zu Theorien von L. Wittgenstein und W. Sellars*, (Suhrkamp Verlag; Frankfurt/M: 1969).
- **Habermas, Juergen**, "Sprachspiel, Intention und Bedeutung. Zu Motiven bei Sellars und Wittgenstein", in Wiggerhaus, R., (ed.), *Sprachanalyse und Soziologie. Die sozialwissenschalfliche Relevanz von Wittgensteins Sprachphilosophie*, (Suhrkamp Verlag; Frankfurt/M: 1975), pp. 319-40.
- **Harman, Gilbert H.**, "Sellars' Semantics", *The Philosophical Review* 79, 1970, pp. 404-19.
- **Hooker, C.A.**, "Sellars' Argument for the Inevitability of the Secondary Qualities", *Philosophical Studies* 32, 1977, pp. 335-48.
- **Koch, Anton F.**, *Vernunft und Sinnlichkeit im praktischen Denken. Eine sprachbehavioristische Rekonstruktion Kantisher Theoreme gegen Sellars*, (Verlag Königshausen + Neumann; Würzburg: 1980).
- **Kurthen, M.**, "Qualia, Sensa und Absolute Prozesse. Zu W. Sellars' Kritik des psychocerebalen Reduktionismus", *Journal for General Philosophy of Science (Zeitschrift für Allgemeine Wissenschaftstheorie)*, 21, 1990, 25-41.
- **Marras, Antonio**, "Sellars on Thought and Language", *Noûs* 7, 1973, pp. 152-63.
- ____________, "On Sellars' Linguistic Theory of Conceptual Activity", *Canadian Journal of Philosophy*, 2, 1973, pp. 471-83.
- ____________, "Reply to Sellars", *Canadian Journal of Philosophy*, 2, 1973, pp. 495-501.
- ____________, "Sellars' Behaviourism: A Reply to Fred Wilson", *Philosophical Studies*, 30, 1976, pp. 413-18.

- **McDowell, John**, *Mind and World*, (Harvard University Press; Cambridge, MA; 1994).
- ____________, "Having the World in View: Sellars, Kant, and Intentionality", *Journal of Philosophy*, 95, 1998, pp. 431-91.
- **McGilvray, J.A.**, "Pure Process(es)?", *Philosophical Studies* 43, 1983, pp. 243-51.
- **Meyers, R.G.**, "Sellars' Rejection of Foundations", *Philosophical Studies*, 39, 1981, pp. 61-78.
- **Pohlenz, G.**, "Phänomenale Realität und naturalistische Philosophie. Eine systematische Widerlegung der Feigl'schen und Sellars'schen Theorien phänomenaler Qualitäten und Skizze einer alternativen Theorie", *Zeitschrift für philosophische Forschung*, 44, 1990, 106-42.
- **Richardson, R.C. and Muilenburg, G.**, "Sellars and Sense Impressions", *Erkenntnis*, 17, 1982, pp. 171-211.
- **Rosenberg, Jay F.**, "The Elusiveness of Categories, the Archimedean Dilemma, and the Nature of Man", in Castañeda, ed., [AK&R](see above), pp. 147-84.
- ____________, "Linguistic Roles and Proper Names", in Pitt, [PSQE] (see above), pp. 189-216.
- ____________, "The Place of Color in the Scheme of Things: A Roadmap to Sellars' Carus Lectures", *The Monist*, 65, 3, 1982, pp. 315-35.
- ____________, "Wilfrid Sellars' Philosophy of Mind" in *Contemporary Philosophy, 4: Philosophy of Mind*, ed. by Guttorm Floistad, (Martinus Nijhoff Publishers; 1983), pp. 417-39.
- ____________, [FI] "Fusing the Images: Nachruf for Wilfrid Sellars", *Journal for General Philosophy of Science* (*Zeitschrift für allgemeine Wissenschaftstheorie*), Vol. XXI, No. 1, 1990, pp. 3-25.
- ____________, "Response to Aune, "Sellars' Two Images of the World"", (Abstract), *The Journal of Philosophy*, Vol. 87, No. 10, October, 1990, pp. 546-7.
- ____________, "Wilfrid Sellars und die Theorie-Theorie", *Deutsche Zeitschrift für Philosophie*, 48, 2000, pp. 639-655.
- ____________, "Sellars, Wilfrid Stalker", entry in *A Companion to Analytic Philosophy*, A. Martinich & D. Sosa, eds., (Blackwell Publishing Ltd; Oxford: 2001), pp. 239-53.
- **Rottschaefer, W.A.**, "Verbal Behaviorism and Theoretical Mentalism: An Assessment of the Marras-Sellars Dialogue", *Philosophical Research Archives*, 9, 1983, pp. 511-33.
- **Seibt, Johanna**, "Analysis without synopsis must be blind. Obituary for W. Sellars", *Erkenntnis*, 33, 1990, pp. 5-8.
- ____________, "Wilfrid Sellars' systematischer Nominalismus", *Information Philosophie*, 3, 1995, pp. 22-6.
- **Sicha, Jeffrey**, *The Metaphysics of Elementary Mathematics*, (University of Massachusetts Press; Amherst, MA: 1974).
- **Smart, J.J.C.**, "Sellars on Process", *The Monist* 65, 1982, pp. 302-14.
- **Sosa, Ernest**, "Mythology of the Given", *History of Philosophy Quarterly*, 14, 1997, pp. 275-87.
- **Tye, Michael**, "The Adverbial Theory: A Defense of Sellars against Jackson", *Metaphilosophy*, 6, 1975, pp. 136-43.
- **van Fraassen, Bas C.**, "Wilfrid Sellars on Scientific Realism", *Dialogue* 14, 1975, pp. 606-16.
- ____________, "On the Radical Incompleteness of the Manifest Image", *Proceedings of the PSA*, Volume 2, ed. by F. Suppe and P.D. Asquith, 1976, pp. 335-43.
- **Vinci, T.**, "Sellars and the Adverbial Theory of Sensation", *Canadian Journal of Philosophy*, 11, 1981, pp. 199-217.

- **Wilson, Fred**, "Marras on Sellars on Thought and Language", *Philosophical Studies*, 28, 1975, pp. 91-102.
- **Woods, M.**, "Sellars on Kantian Intuitions", *Philosophy and Phenomenological Research*, 44, 1984, pp. 413-18.
- **Wright, E.L.**, "A Defense of Sellars", *Philosophy and Phenomenological Research*, 46, 1985, pp. 73-90.

Other Internet Resources

- Wilfrid Sellars Web Site

Related Entries

Carnap, Rudolf | functionalism | intentionality | Kant, Immanuel | language: philosophy of | meaning | mind: philosophy of | Quine, Willard van Orman | science, philosophy of

Jay F. Rosenberg
jfr@email.unc.edu

First published: February 22, 1997
Content last modified: April 24, 2002

Representational Theories of Consciousness

Consciousness is said in many ways. A theory "of consciousness" can be about practically anything mental; just look through the various current book series and anthologies and journals that feature the word "consciousness" in their titles. But the notions of consciousness addressed here will be only the following: (1) Conscious awareness of one's own mental states, and "conscious states" in the particular sense of: states whose subjects are aware of being in them. (2) "Phenomenal consciousness," as Block (1995) calls it, viz., being in a sensory state that has a distinctive qualitative or phenomenal character. (3) The matter of "what it is like" for the subject to be in a particular mental state, especially for that subject to experience a particular phenomenal property.

The idea of representation has been central in discussions of intentionality for many years, at least since Sellars (1956). But only more recently has it begun playing a major role in theories of consciousness. For each of the three foregoing notions of consciousness, some philosophers have claimed that that type of consciousness is entirely or largely explicable as a kind of representing.

§1: Awareness of One's Own Mental States

Occurrent mental or psychological states fall roughly into three categories: those whose subjects are aware of being in them; those whose subjects are not aware of being in them, but could have been had they taken notice; and those, such as language-processing states, which are entirely subterranean and inaccessible to introspection. A theory is needed to explain these differences.

Two Representational Theories of Awareness

There are two main representational theories of conscious awareness and conscious states. (Important terminological note: Again, I am using the term "conscious state" to mean a mental state that one is aware of being in, and "awareness" as a term for that phenomenon. The phrase "conscious state" has been used in at least one entirely different sense, as by Dretske (1993). Failure to keep these senses straight can lead to much confusion.) The two theories are the Lockean "inner sense" or "higher-order perception" (HOP) theory offered by Armstrong (1968, 1981) (see also Lycan (1987, 1996)), and the "higher-order thought" (HOT) theory defended by Rosenthal (1993) (see also Gennaro (1995) and Carruthers (2000)). According to each of the theories, a subject's awareness of her/his own mental state consists in a representing of that state itself. HOP has it that the representing is done quasi-perceptually, by a set of functionally specified internal attention mechanisms of some kind that scan or monitor first-order mental/brain states. HOT theorists say that merely having a thought about the first-order state will suffice, provided that the thought was directly caused by the state itself.

Each of HOP and HOT easily explains the differences between our three categories of first-order state; a state is, or is not, or could not be a conscious state accordingly as it itself is, or is not, or psychofunctionally could not be the object of a higher-order quasi-perception or thought. Further motivation for a representational theory of awareness is obvious enough as well: In general, to be aware of any thing or state of affairs is to represent that item in some way, and to be unaware or unconscious of it is to fail to represent it. And when we deliberately introspect and thereby become aware of a first-order mental state that we had not realized we were in, the awareness is quasi-perceptual or at least takes the form of a mental state of some kind itself directed upon the first-order state; it feels as though we are "looking at" a particular sector of our cognitive or phenomenal field.

Some Criticisms of Representational Theories of Awareness

Two misguided objections to representational theories need to be warded off. First, it is sometimes complained that such theories do not in any way explain the qualitative or phenomenal character of experience; a mere higher-order representation could hardly bring a phenomenal property into being. So how can they claim to be theories of *consciousness*? But it is no claim of either HOP or HOT (at least as I understand them here) to have explained anything about phenomenal character; they are theories only of the distinction between mental states one is aware of being in and mental states one is not aware of being in. Some other theory must be given of phenomenal character.

Second, some philosophers have feared a regress. If the second-order representation is to confer consciousness on the first-order state, it must itself be a conscious state; so there must be a third-order

representation of it, and so on forever. But HOP and HOT theorists reject the opening conditional premise. The second-order representation need not itself be a conscious state. (Of course, it *may* be a conscious state, if there does happen to be a higher-order representation of it in turn.)

But some further criticisms are not so easily rebutted. (I here ignore objections that apply selectively, to HOP alone or to HOT alone.)

Fallibility. Some philosophers have complained that representational theories leave introspective beliefs too fallible and underrate the privileged access we have to our own mental states. An internal monitor, or whatever device produces a higher-order thought, is a *mechanism*, and every mechanism is fallible and works only contingently. But the objectors contend that our awareness of our own mental states is either infallible or, if not flatly infallible, noncontingently constrained against unreliability. Shoemaker (1994b), for example, grants that pain can "occasionally" escape awareness, but he insists that that could not happen "as a matter of course; it may be true in Lake Wobegon that all of the children are above average, but it can't be true everywhere" (273-274).

There is a special problem about first-order states that have wide intentional content, i.e., content that is not determined by what is in the subject's head but consists in part of relations the subject bears to external objects (Davidson (1987)). Belief contents are normally wide; two people could be molecularly indistinguishable but have different beliefs, if one were on earth and believed something about about water (H_2O), while the other was on Putnam's Twin Earth and had the corresponding belief about the qualitatively similar stuff, XYZ. But (Dretske (1995)) an internal monitor, or whatever device produces a higher-order thought, would be located inside the head and would not be able to look outside the head; so it would not be able to tell, for example, whether its owner's belief was a belief about water or a belief about XYZ. Yet we do know introspectively what it is we believe; so much the worse for the higher-order representation views. (Burge (1988) and Heil (1988) have offered what is now a standard reply to this argument; there has been considerable controversy over that reply.)

Finally, Bar-On and Long (in press) argue more radically that privileged access is not an epistemic matter at all; that is, they say, first-person privilege is not a matter of evidence or especially reliable indication or justification of any other sort. Knowing what one believes or what one phenomenally feels is not in that way like proprioceptively knowing the position of my own limbs without looking. (Bar-On and Long go on to suggest a sophisticated expressive theory of first-person avowals.)

Objections of this sort are really objections to higher-order representation theories considered as theories of self-knowledge and first-person privilege. They do not obviously carry over to the theories' original explanandum, which was the distinction between conscious states and un-, sub-, pre-, or otherwise nonconscious states. It may be that the hypothesis of higher-order representation is needed to explain the latter distinction even if it does not satisfactorily explain privileged access.

False Positives. Shoemaker's complaint was about unfelt pains and other sensations of which, on a representationalist view, one might be systematically unaware. But the fallibility problem for

representational theories gets worse: There is also a danger of false positives. Might not an internal scanner, or whatever mechanism produces a higher-order thought, be defective in that it fires mistakenly, introspecting a visual sensation or a pain that never occurred? A subject might seem to her-/himself to suffer terrible pain for weeks, while actually having no pain at all. (This is not a case of psychosomatic pain; psychosomatic pain is real pain, though of no known physical cause.) Does that consequence of representationalism even make clear sense? Neander (1998) prosecutes this objection; Lycan (1998) rejoins (none too effectively).

Ubiquity. Rey (1983) objects that if all it takes to make a first-order state a conscious state is that the state be the object of a higher-order representation, then consciousness is a lot more prevalent than we think. Any laptop computer, for example, has monitoring devices that keep track of its "psychological" states. Perhaps no existing computer has genuinely psychological states, but Rey argues that once we had done whatever needs to be done in order to fashion a being that did have nonconscious first-order intentional and sensory states, the addition of an internal monitor or two would be a trifling afterthought, hardly the sort of thing that could turn a simply nonconscious being into a conscious being.

This objection clearly calls for a thoughtful response. White (1987) and Lycan (1996) have tried to rebut it.

Computational/Cognitive Overload. Carruthers (2000) points out that, given the richness of a person's conscious experience at a time, the alleged higher-order representing devices would be kept very busy. The higher-order representations would have to keep pace with every nuance of the total experience. The complexity of the experience would have to be matched in every detail by a higher-order perception or thought. It is hard to imagine that a human being would have so great a capacity for complex higher-order representation, much less that a small child or a nonhuman animal would have it. Carruthers concludes that if HOP or HOT is true, then to say the least, few if any creatures besides human adults have conscious experiences. Perhaps representational theorists can live with that apparent consequence; perhaps not.

Thus, representational theories of consciousness in the sense of awareness of one's own mental states have serious difficulties to overcome.

§2: Qualitative/Phenomenal Character

Qualitative features of mental states are often called "qualia" (singular, "quale"). In recent philosophy of mind that term has been used in a number of confusingly different ways, but here I shall use it in the specific, strict sense due to C.I. Lewis (1929): A quale in this sense is an introspectible monadic qualitative or phenomenal property inhering in a mental state, and I shall be talking only of sensory states in particular. A quale can be thought of as the distinctive phenomenal property of an apparent phenomenal individual. (An "apparent phenomenal individual" is anything of the sort that Russell would have taken to be a "sense-datum," such as a colored region of ones visual field, or a heard sound or an experienced smell.) A paradigm of a quale is the color of an after-image. For example, Bertie is

experiencing a green after-image as a result of seeing a red flash bulb go off; the greenness of the after-image is the quale.

Qualia in this sense pose a problem for materialist theories of the mind. For where, ontologically speaking, are they located? Actual Russellian sense-data are immaterial individuals; so the materialist cannot admit that the greenness of the after-image is a property of an actual sense-datum. Nor is it plausible to suggest that the greenness is exemplified by anything physical in the brain (if there is some green physical thing in your brain, you are probably in big trouble). To sharpen the problem, suppose there is no green physical object in Bertie's visible environment either: So there is no green physical thing either inside his head or visibly outside it. But since there is a green thing that he is experiencing, it must after all be a nonphysical, immaterial thing.

There is a representational theory of qualia, due in modern times to Anscombe (1965) and Hintikka (1969); adherents include Kraut (1982), Lewis (1983), Lycan (1987, 1996), Harman (1990), Shoemaker (1994a), Tye (1994, 1995), and Dretske (1995). The representational theory is an attempt to resolve the foregoing dilemma compatibly with materialism. According to it, qualia are actually intentional contents, represented properties of represented objects. Suppose Ludwig is seeing a real tomato in good light, and naturally it looks red to him; there is a corresponding red patch in his visual field. He is visually representing the actual redness of the tomato, and the redness of the "patch" is just the redness of the tomato itself. But suppose George Edward is hallucinating a similar tomato, and there is a tomato-shaped red patch in his visual field just as there is in Ludwig's. George Edward too is representing the redness of an external, physical tomato. It is just that in his case the tomato is not real; it and its redness are intentional inexistents. But the redness is still the redness of the illusory tomato. (Note that the representation going on here is good old first-order representation of environmental features, not higher-order representation as in the HOP and HOT theories of awareness.)

What about Bertie's green after-image? On the representationalist analysis, for Bertie to experience the green after-image is for Bertie to be visually-representing a green blob located at such-and-such a spot in the room. Since in reality there is no green blob in the room with Bertie, his visual experience is unveridical; after-images are illusions. The quale, the greenness of the blob, is (like the blob itself) an intentional inexistent.

And that is how the representationalist resolves our dilemma. There is a green thing that Bertie is experiencing, but it is not an actual thing. That "there is" is the same lenient non-actualist "there is" that occurs in "There is something that Bertie believes in but that doesn't exist" and in "There is a mythical god that the Greeks worshipped but no one worships any more." (In defending his sense-data, Russell mistook a nonactual material thing for an actual immaterial thing.)

Most representationalists agree that the representation of color and other sensible properties is "nonconceptual" in some sense--at least in that the qualitative representations need not be easily translatable into the subject's natural language. Of course, some psychosemantics would be needed to explain what it is in virtue of which a brain item represents greenness in particular. Dretske (1995) offers one, as does Tye (1995); both accounts are teleologized versions of "indicator" semantics.

On pain of circularity, the representational theory requires color realism. In this discussion, "green" has meant the objective, public property that inheres in some physical objects. (One could not, without circularity, explicate phenomenal greenness in terms of represented real-world color and then turn around and construe real physical greenness as a disposition to produce sensations of phenomenal greenness.) What sort of real-world property is an "objective," physical color? There is a variety of realist answers, though none of them is uncontroversial. (Dretske (1995), Tye (1995), Lycan (1996) and Lewis (1997) each gesture toward one.)

Of course, the mere representation of redness does not suffice for phenomenal red, for something's looking red to a subject. (One could say the word "red" aloud, or semaphore it from a cliff, or send it in Morse code, or write the French word "rouge" on a blackboard, or point to a color chip.) The representation must be specifically a visual representation, produced by either a normal human visual system or by something functionally like one. Thus, the representational theory of qualia cannot be purely representational, but must appeal to some further factor. Dretske (1995) cites only the fact that visual representation is sensory and what he calls "systemic." Tye (1995) requires that the representation be nonconceptual and "poised," and also argues that visual representations of color would differ from other sorts of representations in being accompanied by further representational differences. Lycan (1996) appeals to functional considerations. (The latter mixed view is what Block (1996) calls "quasi-representationism.")

Thus we may distinguish different grades of representationalism about qualia. *Purest* representationalism would be the view that representation is all it takes to make a quale; but no one holds that view. *Strong* representationalism is what is defended by Dretske, Tye and Lycan: that representation of a certain kind suffices for a quale, where the kind can be specified in functionalist or other familiar materialist terms, without recourse to properties of any ontologically new sort. *Weak* representationalism says only that qualitative states have representational content, which admission is compatible with qualia also necessarily involving features that are ontologically "new" (Block (1990, 1996), Chalmers (1996)). (There is a further question, of whether qualia themselves, in our very specific sense, exhaust all of what has usually been thought of as a sensory state's overall phenomenal character. Lycan (1998) argues that they do not.) Throughout the rest of this article, unless otherwise noted, "representationalism" shall mean the strong representationalist view.

§2.1: Arguments in Favor of the Representational Theory of Qualia

There are at least four direct arguments in favor of the representational theory.

First, the theory is the only very promising way to preserve materialism while accommodating qualia. For the only viable alternative resolution of our Bertie dilemma seems to be belief in actual Russellian sense-data or at least in immaterial properties. (The anti-materialist may not mind sense-data ontologically, but s/he will also inherit the nasty epistemological problems that Russell never succeeded in overcoming; the external world is a nice thing to have. More likely, an opponent will hold the line at

property dualism, as do Jackson (1982) and Chalmers (1996).) A materialist might suggest a type-identity of Bertie's phenomenal greenness with something neurophysiological, but that would be to do away with the important claim that greenness itself, rather than some surrogate property, figures in Bertie's experience; the suggestion would be an error theory, and would have to explain away the intuition that, whatever the ultimate ontology, Bertie really is experiencing an instance of greenness.

Second (Dretske (1996)), there is nothing intrinsic to the brain that constitutes the difference between a red quale and a green one; unless there are Russellian sense-data or at least immaterial properties, what distinguishes the two qualia must be relational, and the only obvious candidate is, *representing* red or green. (A surprising but harmless consequence of this view is that qualia in our strict (C.I.-)Lewisian sense are not themselves properties of the experiences that present them: Qualia are represented properties of represented objects, and so they are only intentionally present in experiences. As before, the relevant properties *of* the experiences are, representing this quality or that. Of course, one could neologize slightly and speak of "qualia" as properties of experiences, identifying them with representational features.) But to this second argument the neurophysiological type-identity theorist would have a stronger rebuttal.

Third, we distinguish between veridical and unveridical visual experiences. How so? I would take it to be fairly uncontentious that Bertie's experience is as of a green blob and has greenness as an intentional object, and that what the experience reports is false. (The only type of theorist I know who disputes this is an unreconstructed Russellian sense-datum merchant: If one thinks of the after-image itself as an actually and independently existing individual--indeed one of the world's basic building blocks--one will not also think of it as representational.) Moreover, the experience's veridicality condition, i.e., there being a green blob where there seems to Bertie to be one, seems to exhaust not only its representational content but its qualitative content. Once the *greenness* has already been accounted for, what qualitative content is left? (But we shall see in §3 below that that question has a serious nonrhetorical answer.)

Fourth, the transparency argument (Harman (1990)): We normally "see right through" perceptual states to external objects and do not even notice that we are *in* perceptual states; the properties we are aware of in perception are attributed to the objects perceived. "Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree, including relational features of the tree ‘from here’" (p. 39). Tye (1995) extends this argument to bodily sensations such as pain.

The transparency argument can be extended also to the purely hallucinatory case. Suppose you are looking at a real, richly red tomato in good light. Suppose also that you then hallucinate a second, identical tomato to the right of the real one. (You may be aware that the second tomato is not real.) Phenomenally, the relevant two sectors of your visual field are just the same; the appearances are just the same in structure. The redness involved in the second-tomato appearance is exactly the same property as is involved in the first. But if we agree that the redness perceived in the real tomato is just the redness of the tomato itself, then the redness perceived in the hallucinated tomato--the red quale involved in the second-tomato appearance--is just the redness of the hallucinated tomato itself.

§2.2: Objections to the Representational Theory of Qualia

Objections to the representational theory come mainly in the form of counterexamples: cases in which either two experiences share their intentional content and differ in their qualia or they differ entirely in their intentional content but share qualia. Peacocke (1983) gave three examples of the former kind, Block (1990) one of the latter. (Block (1995, 1996) also offers some of the former kind; for discussion, see Lycan (1996).)

In Peacocke's first example, your experience represents two (actual) trees, at different distances from you but as being of the same physical height and other dimensions; "[y]et there is also some sense in which the nearest tree occupies more of your visual field than the more distant tree" (p. 12). That sense is a qualitative sense, and Peacocke maintains that the qualitative difference is unmatched by any representational difference. The second and third examples concern, respectively, binocular vision and the reversible-cube illusion. In each case, Tye (1995) and Lycan (1996) have rejoined that there are after all identifiable representational differences constituting the qualitative differences.

Block appeals to an "Inverted Earth," a planet exactly like Earth except that its real physical colors are (somehow) inverted with respect to ours. The Inverted Earthlings' speech sounds just like English, but their intentional contents in regard to color are inverted relative to ours: When they say "red," they mean green (if it is green Inverted objects that correspond to red Earthly objects under the inversion in question), and green things *look* green to them even though they call those things "red." Now, an Earthling victim is chosen by the customary mad scientists, knocked out, fitted with color inverting lenses, transported to Inverted Earth, and repainted to match that planet's human skin and hair coloring. Block contends that after some length of time--a few days or a few millennia--the victim's word meanings and propositional-attitude contents and all other intentional contents will shift to match the Inverted Earthlings' contents, but, intuitively, the victim's qualia will remain the same. Thus, qualia are not intentional contents.

The obvious representationalist reply is to insist that if the intentional contents would change, so too would the qualitative contents; what is Block's argument for denying this? Block's nearly explicit argument is that qualia are "narrow," in that they supervene on head contents (on this view, two molecularly indistinguishable people could not experience different qualia), while the intentional contents shift under environmental pressure precisely because they are "wide." If qualia are indeed narrow, and all the intentional contents are wide and would shift, then Block's argument succeeds. (Stalnaker (1996) gives a version of Block's argument that does not depend on the assumption that qualia are narrow; Lycan (1996) rebuts it.)

Three replies are available, then: (i) To insist that not all the intentional contents would shift. Word meanings would shift, but it does not follow that visual contents ever would. (ii) To hold that although all the ordinary intentional contents would shift, there is a special class of narrow though still representational contents underlying the wide contents; qualia can be identified with the special narrow contents. (iii) To deny that qualitative content is narrow and argue that it is wide, i.e., that two molecularly indistinguishable people could indeed experience different qualia. This last is the position

that Dretske (1996) has labelled "phenomenal externalism."

Reply (i) has not been much pursued. (ii) has, a bit, by Tye (1994) and especially Rey (1998). Rey argues vigorously that qualia are narrow, and then offers a narrow representational theory. (But it turns out that Rey's theory is not a theory of qualia in the strict Lewisian sense of quale used here; see below.)

Reply (iii), phenomenal externalism, has been defended by Dretske (1995, 1996), Tye (1995) and Lycan (1996). A number of people (even Tye himself (1998)) have since called the original contrary assumption that qualia are narrow a "deep / powerful / compelling" intuition, but it proves to be highly disputable. Here are two arguments, though not very strong arguments, for the claim that qualia are wide.

First, if the representational theory is correct, then qualia are determined by whatever determines a psychological state's intentional content; in particular, the color properties represented are taken to be physical properties instanced in the subject's environment. What determines a psychological state's intentional content is given by a *psychosemantics*, in Fodor's (1987) sense. But every known plausible psychosemantics makes intentional contents wide. Of course, the representational theory is just what is in question; but:

Second, suppose qualia are narrow. Then Block's Inverted Earth argument is plausible, and it would show that either qualia are narrow functional properties or they are properties of a very weird kind whose existence is suggested by nothing else we know (see Ch. 6 of Lycan (1996)). But qualia are not functional properties, at least not narrow ones: Recall the Bertie dilemma. Also, qualia are monadic properties, while functional properties are all relational; and see further Block's anti-functionalist arguments in Block (1978). So, either qualia are wide or weirdness is multiplied beyond necessity. Of course, that dichotomy will be resisted by anyone who offers a narrow representationalist theory as in (ii) above.

§2.4: Arguments For Narrow Qualia

Although until a few years ago the assumption that qualia are narrow had been tacit and entirely undefended, opponents of representationalism have since begun defending it with vigor. Here are (only) some of their arguments, with sample replies.

Introspection. An Earthling suddenly transported to Inverted Earth or some other relevant sort of Twin Earth would notice nothing introspectively, despite a change in representational content; so the quale must remain unchanged and so is narrow.

Reply: The same goes for propositional attitudes, i.e., the transported Earthling would notice nothing introspectively. Yet the attitude contents are still wide. Wideness does not entail introspective change under transportation.

Narrow content. In the propositional-attitude literature, the corresponding transportation argument has

been taken as the basis of an argument for "narrow content," viz., for something that is intentional content within the meaning of the act but is narrow rather than, as usual, wide. The self-knowledge problem aforementioned, and the problem of "wide causation" (Fodor (1987), Kim (1995)), have also been used to motivate narrow content. And, come to think of it, any general argument for narrow content will presumably apply to sensory representation as well as to propositional attitudes. If there is narrow content at all, and sensory content is representational, then probably sensory states have narrow content too. Thus, qualia can and should be taken to be the narrow contents of such states.

Replies: First, this begs the question against the claim that qualia are wide. Even if there are indeed narrow contents impacted within sensory states, independent argument is needed for the identification of qualia with those contents rather than with wide contents. Second and more strongly, narrow sensory contents still would not correspond to qualia in the Lewisian sense. The redness of a patch in my visual field is (so far as has been shown) still a wide property, even if some other, narrow property underlies it in the same way that mysterious, ineffable narrow contents are supposed to underlie beliefs and desires.

Modes of Presentation. (Rey (1998)) There is no such thing as representation without a mode of presentation. If a quale is a representatum, then it is represented under a mode of presentation, and modes of presentation may be narrow even when the representational content itself is wide. Indeed, many philosophers of mind take modes of presentation to be internal causal or functional roles played by the representations in question. Surely they are strong candidates for qualitative content. Are they not narrow qualia?

Reply: Remember, the qualia themselves are properties like phenomenal greenness and redness, which according to the representational theory are representata. The modes or guises *under which* greenness and redness are represented in vision are something else again.

It can plausibly be argued that such modes and guises are qualitative or phenomenal properties of some sort, perhaps higher-order properties. See the next section.

Qualitative but Nonrepresenting States. (Rey (1998)) Even if perceptual states are representational, bodily sensations and moods do not have that same feature. Yet bodily sensations and moods do have qualitative character. "Many have noted that states like that of elation, depression, anxiety, pleasure, orgasm seem to be just overall states of *oneself*, and not features of presented objects" (p. 441, italics original).

Here the representational theorist must tough it out, and insist both that bodily sensations and moods do have intentional content and that their intentional content exhausts their qualitative character. Of course, each of those claims, especially the first, requires hefty independent defense. It is easy enough to argue that pains and tickles and even orgasms have some representational features (see Tye (1995) and Lycan (1996)). For example, a pain is felt as being in a certain part of one's body, *as if* that part is disordered in a certain way; that is why pains are described as "burning," "stabbing," "throbbing" and the like. It is harder to show that the sensations' qualitative contents are exhausted by their representational features. It is perhaps still harder to maintain that a mood has intentional content, though it could be argued that a

state of elation, for example, represents *the world* as a beautiful and exciting place.

Memory. (Block (1996)) "[Y]ou remember the color of the sky on your birthday last year, the year before that, ten years before that, and so on, and your long-term memory gives you good reason to think that the phenomenal character of the experience has not changed.... Of course, memory *can* go wrong, but why should we suppose that it **must** go wrong here?" (pp. 43-44, italics and boldface original). The idea is that memory acts as a check on the qualia, and can be used to support the claim that the qualia have remained unchanged despite the wholesale shift in representational contents.

Reply: Memory contents are wide, and so by Block's own reasoning they will themselves undergo the representational shift to the Inverted-Earth complementary color. Thus, your post-shift memories of good old Earth are false. When you say or think to yourself, "Yes, the sky looks as blue as it did thirty years ago," you are not expressing the same memory content as you would have when you had just arrived on Inverted Earth. You are now remembering or remembering that the sky looked yellow, since for you "blue" now means yellow. And that memory is false, since on the long-ago occasion the sky looked blue to you, not yellow; memory is not after all a reliable check on the qualia. (Lycan (1996) takes this line; Tye (1998) expands it in more detail.)

Hardly anyone will accept all of the foregoing replies. But no one should now find it obvious either that qualia are narrow or that they are wide. The matter is likely to remain controversial for some time to come.

§3: What It's Like

Some philosophers (e.g., Dretske (1995), Tye (1995)) use this troublesome expression simply to mean a quale, and this is one of the two meanings it has had in recent philosophy of mind. But in the opening paragraph of this article, the phrase was introduced in the context, "‘what it is like’ for the subject to be in a particular mental state, especially for that subject to experience a particular phenomenal property [or quale]," which suggests that there is another sense in which (when the mental state does involve a quale) the "what it's like" is something over and above the quale itself. In fact, since this second "what it's like" is itself a property of the quale, it cannot very well be identical with the quale. (Block (1995), like many other writers, fails to distinguish this from "phenomenal consciousness" in our original sense. But Carruthers (2000) elaborates nicely on the distinction.)

Here are two further reasons for maintaining such a distinct sense of the phrase. First, Armstrong (1968), Nelkin (1989), Rosenthal (1991), and Lycan (1996) have argued that qualia can fail to be conscious in the earlier sense of awareness; a quale can occur without its being noticed by its subject. But in such a case, there is a good sense in which it would not be like anything for the subject to experience that quale. (Of course, in the first, Dretske-Tye sense there would be something it was like, since the quale itself is that. But in another sense, if the subject is *entirely unaware* of the quale, it is odd even to speak of the subject as "experiencing" it, much less of there being something it is like for the subject to experience it.) So even in the case in which one is aware of one's quale, the second type of "what it's like" requires

awareness and so is something distinct from the quale itself.

Second, a quale can be described in one's public natural language, while what it is like to experience the quale seems to be ineffable. Suppose Ludwig asks Bertie, "How, exactly, does the after-image look to you as regards color?" Bertie replies, "I told you, it looks green." "Yes," says Ludwig, "but can you tell me what it's like to experience that ‘green’ look?" "Well, the image looks the same color as that," says Bertie, pointing to George Edward's cloth coat. "No, I mean, can you tell me what it's like intrinsically, not comparatively?" "Um,...." --In one way, Bertie can describe the phenomenal color, paradigmatically as "green." But when asked what it is like to experience that green, he goes mute. So there is a difference between (a) "what it's like" in the bare sense of the quale, the phenomenal color that can be described using ordinary color words, and (b) "what it's like" to experience that phenomenal color, which cannot easily be described in public natural language at all.

It is the second sense of "what it's like" that figures in anti-materialist arguments from subjects' "knowing what it's like," primarily Nagel's (1974) "Bat" argument and Jackson's (1982) "Knowledge" argument. Jackson's character Mary, a brilliant color scientist trapped in an entirely black-and-white laboratory, nonetheless becomes omniscient as regards the physics and chemistry of color, the neurophysiology of color vision, and every other public, objective fact conceivably relevant to human color experience. Yet when she is finally released from her captivity and ventures into the outside world, she sees colors for the first time, and learns something: namely, she learns what it is like to see red and the other colors. Thus she seems to have learned a new fact, one that by hypothesis is not a public, objective fact. It is an intrinsically perspectival fact. This is what threatens materialism, since according to that doctrine, every fact about every human mind is ultimately a public, objective fact.

Upon her release, Mary has done two things: She has at last hosted a red quale, and she has learned what it is like to experience a red quale. In experiencing it she has experienced a "what it's like" in the first of our two senses. But the fact she has learned has the ineffability characteristic of our second sense of "what it's like"; were Mary to try to pass on her new knowledge to a still color-deprived colleague, she would not be able to express it in English.

We have already surveyed the representational theory of qualia. But there are also representational theories of "what it's like" in the second sense. A common reply to the arguments of Nagel and Jackson (Horgan (1984), Van Gulick (1985), Churchland (1985), Tye (1986), Lycan (1987, 1990, 1996), Loar (1990), Rey (1991), Leeds (1993)) is to note that a knowledge difference does not entail a difference in fact known, for one can know a fact under one representation or mode of presentation but fail to know one and the same fact under a different mode of presentation. Someone might know that water is splashing but not know that H_2O molecules are moving, and vice versa; someone might know that person X is underpaid without knowing that she herself is underpaid, even if she herself is in fact person X. Thus, from Mary's before-and-after knowledge difference, Jackson is not entitled to infer the existence of a new, weird fact, but at most that of a new way of representing. Mary has not learned a new fact, but has only acquired a new, introspective or first-person way of representing one that she already knew in its neurophysiological guise.

(As noted above, the posited introspective modes of presentation for qualia in the Lewisian sense are strong candidates for the title of qualia in a distinct, higher-order sense of the term, and they may well be narrow rather than wide. This is what Rey (1998) seems to be talking about.)

This attractive response to Nagel and Jackson--call it the "perspectivalist" response--requires that the first-order qualitative state itself be represented (else how could it be newly known under Mary's new mode of presentation?). And that hypothesis in turn encourages a representational theory of higher-order conscious awareness and introspection. Since we have seen that representational theories of awareness face powerful objections, the perspectivalist must either buy into such a theory despite its liabilities, or find some other way of explicating the idea of an introspective or first-person perspective without appealing to higher-order representation. The latter option does not seem promising.

Bibliography

- Anscombe, G.E.M. (1965). ‘The Intentionality of Sensation: A Grammatical Feature’, in R.J. Butler (ed.), *Analytical Philosophy: Second Series*. (Oxford: Basil Blackwell.)
- Armstrong, D.M. (1968). *A Materialist Theory of the Mind*. (London: Routledge and Kegan Paul.)
- Armstrong, D.M. (1981). ‘What is Consciousness?’, in *The Nature of Mind*. (Ithaca, NY: Cornell University Press.)
- Bar-On, D., and D. Long (in press). ‘Avowals and First-Person Privilege’, *Philosophy and Phenomenological Research*.
- Block, N.J. (1978). ‘Troubles with Functionalism’, in W. Savage (ed.), *Perception and Cognition: Minnesota Studies in the Philosophy of Science, Vol. IX*. (Minneapolis: University of Minnesota Press).
- Block, N.J. (1990). ‘Inverted Earth’, in Tomberlin (1990). Reprinted in Lycan (1999).
- Block, N.J. (1995). ‘On a Confusion about a Function of Consciousness’, *Behavioral and Brain Sciences* 18: 227-47.
- Block, N.J. (1996). ‘Mental Paint and Mental Latex’, in Villanueva (1996).
- Burge, T. (1988). ‘Individualism and Self-Knowledge’, *Journal of Philosophy* 85: 649-53.
- Carruthers, P. (2000). *Phenomenal Consciousness*. (Cambridge: Cambridge University Press.)
- Chalmers, D. (1996). *The Conscious Mind*. (Oxford: Oxford University Press.)
- Churchland, P.M. (1985). ‘Reduction, Qualia, and the Direct Introspection of Brain States’, *Journal of Philosophy* 82: 8-28.
- Davidson, D. (1987). ‘Knowing Ones Own Mind’, *Proceedings and Addresses of the American Philosophical Association*, Vol. 60, No. 3.
- Davies, M., and G. Humphreys (eds.) (1993). *Consciousness*. (Oxford: Basil Blackwell.)
- Dretske, F. (1993). ‘Conscious Experience’, *Mind* 102: 263-83.
- Dretske, F. (1995). *Naturalizing the Mind*. (Cambridge, MA: Bradford Books / MIT Press.)
- Dretske, F. (1996). ‘Phenomenal Externalism’, in Villanueva (1996).
- Fodor, J.A. (1987). *Psychosemantics*. (Cambridge, MA: Bradford Books/MIT Press.)
- Gennaro, R. (1995). *Consciousness and Self-Consciousness*. (Philadelphia: John Benjamins.)
- Harman, G. (1990).‘ The Intrinsic Quality of Experience’, in Tomberlin (1990). Reprinted in

Lycan (1999).

- Heil, J. (1988). ‘Privileged Access’, *Mind* 47: 238-51. Reprinted in Lycan (1999).
- Hintikka, K.J.J. (1969). ‘On the Logic of Perception’, in N.S. Care and R.H. Grimm (eds.), *Perception and Personal Identity*. (Cleveland, OH: Case Western Reserve University Press.)
- Horgan, T. (1984). ‘Jackson on Physical Information and Qualia’, *Philosophical Quarterly* 34: 147-52.
- Jackson, F. (1982). ‘Epiphenomenal Qualia’, *Philosophical Quarterly* 32: 127-36. Reprinted in Lycan (1999).
- Kim, J. (1995). ‘Mental Causation: What, Me Worry?’, in E. Villanueva (ed.), *Philosophical Issues, 6: Content*. (Atascadero, CA: Ridgeview Publishing.)
- Kraut, R. (1982). ‘Sensory States and Sensory Objects’, *Noûs* 16: 277-95.
- Leeds, S. (1993). ‘Qualia, Awareness, Sellars’, *Noûs* 27: 303-30.
- Lewis, C.I. (1929). *Mind and the World Order*. (New York: C. Scribners Sons.)
- Lewis, D. (1983). ‘Individuation by Acquaintance and by Stipulation’, *Philosophical Review* 92: 3-32.
- Lewis, D. (1997). ‘Naming the Colours’, *Australasian Journal of Philosophy* 75: 325-42.
- Loar, B. (1990). ‘Phenomenal States’, in Tomberlin (1990).
- Lycan, W.G. (1987). *Consciousness*. (Cambridge, MA: Bradford Books / MIT Press.)
- Lycan, W.G. (1990). ‘What is the Subjectivity of the Mental?’, in Tomberlin (1990).
- Lycan, W.G. (1996). *Consciousness and Experience*. (Cambridge, MA: Bradford Books / MIT Press.)
- Lycan, W.G. (1998). ‘In Defense of the Representational Theory of Qualia’ (Replies to Neander, Rey and Tye), in Tomberlin (1998).
- Lycan, W.G. (ed.) (1999). *Mind and Cognition*, Second Edition. (Oxford: Basil Blackwell.)
- Nagel, T. (1974). ‘What Is It Like to Be a Bat?’, *Philosophical Review* 82: 435-56.
- Neander, K. (1998). ‘The Division of Phenomenal Labor: A Problem for Representational Theories of Consciousness’, in Tomberlin (1998).
- Nelkin, N. (1989). ‘Unconscious Sensations’, *Philosophical Psychology* 2: 129-41.
- Peacocke, C. (1983). *Sense and Content*. (Oxford: Oxford University Press.)
- Rey, G. (1983). ‘A Reason for Doubting the Existence of Consciousness’, in Davidson, R., G.E. Schwartz and D. Shapiro (eds.), *Consciousness and Self-Regulation*, Vol. 3. (New York: Plenum Press.)
- Rey, G. (1991). ‘Sensations in a Language of Thought’, in Villanueva (1991).
- Rey, G. (1998). ‘A Narrow Representationalist Account of Qualitative Experience’, in Tomberlin (1998).
- Rosenthal, D. (1991). ‘The Independence of Consciousness and Sensory Quality’, in Villanueva (1991).
- Rosenthal, D. (1993). ‘Thinking that One Thinks’, in Davies and Humphreys (1993).
- Sellars, W. (1956). ‘Empiricism and the Philosophy of Mind’, in H. Feigl and M. Scriven (eds.), *Minnesota Studies in the Philosophy of Science, Vol. I*. (Minneapolis: University of Minnesota Press.)
- Shoemaker, S. (1994a). ‘Phenomenal Character’, *Noûs* 28: 21-38.
- Shoemaker, S. (1994b). ‘Self-Knowledge and Inner Sense. Lecture II: The Broad Perceptual

Model’, *Philosophy and Phenomenological Research* 54: 271-90.

- Stalnaker, R. (1996). ‘On a Defense of the Hegemony of Representation’, in Villanueva (1996).
- Tomberlin, J.E. (ed.) (1990). *Action Theory and Philosophy of Mind (Philosophical Perspectives, Vol. 4)*. (Atascadero, CA: Ridgeview Publishing.)
- Tomberlin, J.E. (ed.) (1998). *Language, Mind, and Ontology (Philosophical Perspectives, Vol. 12)*. (Atascadero, CA: Ridgeview Publishing.)
- Tye, M. (1986). ‘The Subjectivity of Experience’, *Mind* 95: 1-17.
- Tye, M. (1992). ‘Visual Qualia and Visual Content’, in T. Crane (ed.), *The Contents of Experience*. (Cambridge: Cambridge University Press.)
- Tye, M. (1994). ‘Qualia, Content, and the Inverted Spectrum’, *Noûs* 28: 159-83.
- Tye, M. (1995). *Ten Problems of Consciousness*. (Cambridge, MA: Bradford Books/MIT Press.)
- Tye, M. (1998). ‘Inverted Earth, Swampman, and Representationism’, in Tomberlin (1998).
- Van Gulick, R. (1985). ‘Physicalism and the Subjectivity of the Mental’, *Philosophical Topics* 13: 51-70.
- Villanueva, E. (ed.) (1991). *Philosophical Issues, 1: Consciousness*. (Atascadero, CA: Ridgeview Publishing.)
- Villanueva, E. (ed.) (1996). *Philosophical Issues, 7: Perception*. (Atascadero, CA: Ridgeview Publishing.)
- White, S. (1987). ‘What Is It Like to Be a Homunculus?’, *Pacific Philosophical Quarterly* 68: 148-74.

Other Internet Resources

- Consciousness and Qualia, Part 1 of *Contemporary Philosophy of Mind: An Annotated Bibliography*, by David Chalmers
- Directory of online papers on consciousness (compiled by David Chalmers)

Related Entries

consciousness | consciousness: higher-order theories | consciousness: unity of

William Lycan
ujanel@isis.unc.edu

First published: May 22, 2000
Content last modified: June 13, 2000

The Unity of Consciousness

Our consciousness evidences a striking unity. As we will see, unified consciousness can take more than one form but the following form is central. When we are conscious, we are conscious of the contents of a number of conscious states at the same time and as related to one another in various ways. I am aware not of A and, separately, of B and, separately, of C, but of A-and-B-and-C, all at the same time -- or better, as all parts of the contents of a single conscious state. Since at least the time of Kant (1781/7), this phenomenon has been called the *unity of consciousness*.

Historically, the notion of the unity of consciousness has played a very large role in thought about the mind. Indeed, as we will see, it figured centrally in some of the most influential arguments about the mind from the time of Descartes to the 20th century. In the early part of the 20th century, the notion largely disappeared for a time. Analytic philosophers began to pay attention to it again only in the 1960s. We will trace this history up to about 1900. At that point, we will have to delineate the unity of consciousness more carefully and examine some evidence from neuropsychology because both are necessary to understand the recent work on the issue. We will then examine that work. We will conclude with the question of whether consciousness' being unified has implications for other issues concerning consciousness and the mind.

1. History of the Notion

The unity of consciousness has had an interesting history in philosophy and psychology. Taking Descartes to be the first major philosopher of the modern period, the unity of consciousness was central to the study

of the mind for the whole of the modern period until the 20th century. The notion figured centrally in the work of Descartes, Leibniz, Hume, Reid, Kant, Brentano, James -- indeed, in most of the major precursors of contemporary philosophy of mind and cognitive psychology. It played a particularly important role in Kant's work (Brook, 1994).

A couple of examples will illustrate the role that the notion of the unity of consciousness played in this long literature. Consider a classical argument for dualism (the view that the mind is not the body, indeed is not made out of matter at all). It starts like this:

> When I consider the mind, that is say, myself inasmuch as I am only a thinking thing, I cannot distinguish in myself any parts, but apprehend myself to be clearly one and entire. [Descartes, 1641, p. 196]

Descartes then asserts that if the mind is not made up of parts, it cannot be made of matter, presumably because, as he saw it, anything material has parts. He then goes on to say that this would be enough to prove dualism by itself, had he not already proved it elsewhere. Notice where it is that I cannot distinguish any parts. It is in the unified consciousness that I have of myself.

Here is another, somewhat more elaborate argument based on unified consciousness. The conclusion will be that unified consciousness could never be achieved by any system of components acting in concert. William James' well-known version of the argument starts as follows:

> Take a sentence of a dozen words, take twelve men, and to each one word. Then stand the men in a row or jam them in a bunch, and let each think of his word as intently as he will; nowhere will there be a consciousness of the whole sentence. [James, 1890, Vol. 1, p.160]

James generalizes this observation to all conscious states. To get dualism out of this, we need to add a premise: that if the mind were made out of matter, conscious states would have to be distributed over some group of components in some relevant way. But, this thought experiment is meant to show, conscious states cannot be so distributed. Therefore, the conscious mind is not made out of matter. Q.E.D. Call the argument that James is using here the Unity Argument. Clearly, the idea that our consciousness of, in this case, the parts of a sentence is unified is at the center of the Unity Argument. Like the first, this argument goes all the way back to Descartes. Versions of it can be found in thinkers otherwise as different from one another as Leibniz, Reid, and James. The Unity Argument continued to be influential into the 20th century. That the argument was considered to be a powerful reason for concluding that the mind is not the body is illustrated in a somewhat backhanded way by Kant's treatment of it (as he found it in Descartes and Leibniz, not James, of course).

Kant did not think that we could *demonstrate* anything about the nature of the mind, including whether or not it is made out of matter. To make the case for this view, he had to show that all existing arguments that the mind is not material do not work and he set out to do just this in the chapter in the *Critique of Pure Reason* on the Paralogisms of Pure Reason (1781) (paralogisms are faulty inferences about the nature of the mind). The Unity Argument is the target of a major part of that chapter; if one is going to

show that we cannot know what the mind is like, we must dispose of the Unity Argument, which purports to show that the mind is not made out of matter. Kant's argument that the Unity Argument does not support dualism is simple. He urges that the idea of unified consciousness being achieved by something that has no parts or components is no less mysterious than its being achieved by a system of components acting together (1781, A352). Remarkably enough, even though no philosopher has ever met this challenge of Kant's and no account exists of what an immaterial mind *not* made out of parts might be like, philosophers continued to rely on the Unity Argument until well into the 20th century. It may be a bit difficult for us to capture this now but the idea that unified consciousness could not be realized by any system of components, and *a fortiori* any system of material components, had a strong intuitive appeal for a long time.

The notion that consciousness is unified was also central to one of Kant's own famous arguments, his ‘transcendental deduction of the categories’. In this argument, boiled down to its essentials, Kant claims that in order to tie various objects of experience together into a single unified conscious representation of the world, something that he simply assumed that we could do, we must be able to apply certain concepts to the items in question. In particular we have to apply concepts from each of four fundamental categories of concept: quantitative, qualitative, relational, and what he called ‘modal’ concepts. Modal concepts concern whether an item might exist, does exist, or must exist. Thus, the four kinds of concept are concepts for how many units, what features, what relations to other objects, and what existence status are represented in an experience.

It was relational concepts that most interested Kant and of relational concepts, he thought the concept of cause-and-effect to be by far the most important. Kant wanted to show that natural science (which for him meant primarily physics) was genuine knowledge (he thought that Hume's skeptical treatment of cause and effect relations challenged this status). He believed that if he could prove that we must tie items in our experience together causally if we are to have a unified awareness of them, he would have put physics back on "the secure path of a science". The details of his argument have exercised philosophers for over two hundred years. We will not go into them here, but the argument illustrates how central the notion of the unity of consciousness was in Kant's thinking about the mind and its relation to the world.

Even though the unity of consciousness had been at the center of pre-20th century research on the mind, early in the 20th century the notion almost disappeared. Logical atomism in philosophy and behaviorism in psychology were both unsympathetic to the notion. Logical atomism focused on the atomic elements of cognition (sense data, simple propositional judgments, etc.), rather than on how these elements are tied together to form a mind. Behaviorism urged that we focus on behavior, the mind being either a myth or at least something that we cannot and do not need to study in a science of the human person. This attitude extended to consciousness, of course. The philosopher Daniel Dennett summarizes the attitude prevalent at the time this way:

> Consciousness appear[ed] to be the last bastion of occult properties, epiphenomena, immeasurable subjective states -- in short, the one area of mind best left to the philosophers. Let them make fools of themselves trying to corral the quicksilver of ‘phenomenology’ into a respectable theory. [1978, p. 149]

The unity of consciousness next became an object of serious attention in analytic philosophy only as late as the 1960s. In the years since, new work has appeared regularly. The accumulated literature is still not massive but the unity of consciousness has once again become an object of serious study. Before we examine the more recent work, we need to explicate the notion in more detail than we have done so far and introduce some empirical findings. Both are required to understand recent work on the issue.

2. What the Unity of Consciousness Is

To expand on our earlier notion of the unity of consciousness, we need to introduce a pair of distinctions. Current work on consciousness labors under a huge, confusing terminology. Different theorists talk about access consciousness, phenomenal consciousness, self-consciousness, simple consciousness, creature consciousness, state consciousness, monitoring consciousness, awareness as equated with consciousness, awareness distinguished from consciousness, higher order thought, higher order experience, qualia, the felt qualities of representations, consciousness as displaced perception, ... and on and on and on. We can ignore most of this profusion but we do need two distinctions: between consciousness of objects and consciousness of our representations of objects, and between consciousness of representations and consciousness of self.

Consciousness of objects is closely related to sentience and to being awake. It is (at least) being in a certain informationally and behaviorally responsive state to one's immediate environment. It is the ability, for example, to process and act responsively to information about food, friends, foes, and other items of relevance. One finds consciousness of objects in creatures much less complex than human beings. It is what we (at any rate first and primarily) have in mind when we say of some person or animal as it is coming out of a general anaesthesia, ‘It is regaining consciousness’. Consciousness of objects is not just any form of informational access to the world. It is *knowing about, being conscious of,* things in the world. We will return to this point in a moment.

We are conscious of our representations when we are conscious, not (just) of some object, but of our representations: ‘I am seeing [as opposed to touching, smelling, tasting] and seeing clearly [as opposed to dimly].’ Consciousness of our own representations it is the ability to process and act responsively to information about oneself, but it is not just any form of such informational access. It is *knowing about, being conscious of,* one's own psychological states. In Nagel's famous phrase (1974), when we are conscious of our representations, it is ‘like something’ to have them. If, as seems likely, there are forms of consciousness that do not involve consciousness of objects, they might well consist in consciousness of representations, though some theorists would insist that this kind of consciousness is not *of* representations either (*via* representations, perhaps, but not *of* them).

The distinction just drawn between consciousness of objects and consciousness of our representations of objects may seem to be similar to Block's (1995) well-known distinction between P- [phenomenal] and A- [access] consciousness. Here is his definition of ‘A-consciousness’: "A state is A-conscious if it is poised for direct control of thought and action" (1995, p. 233). He tells us that he cannot define ‘P-

consciousness' in any "remotely noncircular way" but will use it to refer to what he calls "experiential properties", what it is like to have certain states (1995, p. 231). Our consciousness of objects may appear to be rather like Block's A-consciousness. It is not. It is a form of P-consciousness. Consciousness of an object is -- how else can we put it? -- *consciousness* of the object. Even if consciousness *just is* informational access of a certain kind (something that Block would deny), it is not all forms of informational access and we are talking about *conscious* access here. Recall the idea that it is like something to have a conscious state. Other closely related ideas are that in a conscious state, something appears to one, that conscious states have a 'felt quality' (see the entry on qualia). A term for all this is *phenomenology:* conscious states have a phenomenology. (Thus some philosophers speak of *phenomenal consciousness* here.) We could now state the point we are trying to make this way. If I am conscious of an object, then it is like something to have that object as the content of a representation.

(Some theorists would insist that this last statement be qualified. While such a representation of an object may provide everything that a *representation* has to have for its contents to be like something to me, they would urge, something more is needed. Different theorists would add different elements. For some, I would have to be aware, not just of the object, but of my representation of it. For others, I would have to direct attention a certain way or something. We cannot go into this controversy here. Here we are merely making the point that consciousness of objects is more than Block's A-consciousness.)

Consciousness of self involves, not just consciousness of states that it is like something to have, but consciousness of the thing that has them, i.e., of oneself. It is the ability to process and act responsively to information about oneself, but again it is more than that. It is *knowing about, being conscious of,* oneself, indeed *of* oneself *as* oneself. Consciousness of oneself in this way is often called consciousness of oneself as *the subject of experience.* Consciousness of oneself as oneself seems to require an indexical ability and a rather special indexical ability at that, not just an ability to pick out something out but to pick something out *as oneself.* Human beings have such self-referential indexical ability. Whether any other creatures have it is controversial. The leading nonhuman candidate would be chimpanzees and other primates who have been taught enough language to use first-person pronouns.

The literature on consciousness sometimes fails to distinguish consciousness of objects, consciousness of one's own representations, and consciousness of self, or treats one of the three, usually consciousness of one's own representations, as the totality of consciousness. (There may also be conscious states that do not have objects, yet are not consciousness of a representation either. We cannot pursue that complication here.) The term 'conscious' and cognates are ambiguous in everyday English. We speak of someone regaining consciousness -- where we mean simple consciousness of the world. Yet we also say things like, 'She wasn't conscious of what motivated her to say that' -- where we do not mean that she lacked either consciousness of the world or consciousness of self but rather that she was not conscious of certain things *about herself,* specifically, certain of her own representational states. To understand the unity of consciousness, it is important to make these distinctions. The reason is this: the unity of consciousness takes a somewhat different form in consciousness of self than it takes in either consciousness of one's own representations or consciousness of objects.

So what is unified consciousness? As we said, the predominant form of the unity of consciousness is

being aware of a number of things at the same time. Intuitively, this is the notion of a number of representations being aspects of a single encompassing conscious state. A more informative idea can be gleaned from the way philosophers have written about unified consciousness. As emerges from what they have said, the central feature of unified consciousness is taken to be something like this:

> Unity of consciousness: A group of representations being related to one another such that to be conscious of any of them is to be conscious of others of them and of the group of them as a single group.

Call this notion UC(1). Now, unified consciousness *of some sort* can be found in all three of the kinds of consciousness we delineated. (It can be found in a fourth, too, as we will see in a moment.) We can have unified consciousness of: objects represented to us; the representations themselves; and oneself, the thing having the representations. In the first case, the represented objects would appear as aspects of a single encompassing conscious states. In the second case, the representations themselves would thus appear. In the third case, one is aware of oneself as a single, unified subject. Does UC(1) fit all three (or all four, including the fourth yet to be introduced)? It does not. At most, it fits the first two. Let us see how this unfolds.

2.1 Unity of consciousness of objects

Unified consciousness of objects is the consciousness that one has of the world around one (including one's own body) as aspects of a single world, of the various items in it as linked to other items in it. What makes it *unified* can be illustrated by an example. Suppose that I am aware of the computer screen in front of me and also of the car sitting in my driveway. If awareness of these two items is not unified, I will lack the ability to compare the two. If I cannot bring the car as I am aware of it to the state in which I am aware of the computer screen, I could not answer questions such as, Is the car the same color as the WordPerfect icon?, or even, As I am experiencing them, is the car to the left or to the right of the computer screen? We can compare represented items in these ways only if we are aware of both items together, as parts of the same field or state or act of conscious. That is what unified consciousness does for us. UC(1) fits this kind of unified consciousness well. There are a couple of disorders of consciousness in which this unity seems to break down or be missing. We will examine them shortly.

2.2 Unified consciousness of one's own representations

Unified consciousness of one's own representations is the consciousness that we have of our representations, consciousness of our own psychological states. The representations by which we are conscious of the world are particularly important but, if those theorists who maintain that there are forms of consciousness that do not have objects are right, they are not the only ones. What makes consciousness of our representations unified? We are aware of a number of representations together, in such a way that they appear as aspects of a single state of consciousness. As with unified consciousness of the world, here too we can compare items of which we have unified consciousness. For example, we can compare what it is like to *see* an object to what it is like to *touch* the same object. Thus, UC(1) fits this kind of unified

consciousness well, too.

2.3 Unified consciousness of self

When one has unified consciousness of self, one is aware of oneself not just as the subject but, in Kant's words (A350), the ‘single common subject’ of a number of representations and the single common agent of various acts of deliberation and action.

This is one of the two forms of unified consciousness that UC(1) does not fit. When one is aware of oneself as the common subject of experiences, the common agent of actions, one is not aware of *a number* of objects. Some think that when one is aware of oneself as subject, one is not aware of oneself as an *object* at all. Kant held this view (A382, A402, B429). Whatever the merits of this view, it is clear that when one is aware of oneself as the single common subject of many representations, one is not aware of a number of things. Rather, one is aware of, and knows that one is aware of, one and the same thing -- via a number of representations. Call this kind of unified consciousness UC(2). Even though UC(2) is different from UC(1), we still have the core idea: unified consciousness consists in tying what is contained in a number of representations, in this case a number of representations of oneself, together so that they are all part of a single field or state or act of consciousness.

Unified consciousness of self has been argued to have some very special properties. In particular, there is a small but important literature on the idea that the reference to oneself as oneself by which one achieves awareness of oneself as subject involves no ‘identification’ (Castañeda, 1966; Shoemaker, 1968; Perry, 1979). Generalizing the notion a bit, some claim that reference to self does not proceed by way of attribution of properties or features to oneself at all. One argument for this view is that one is or could be aware of oneself as the subject of each and every one of one's conscious experiences. If so, awareness of self is not what Bennett calls ‘experience-dividing’ -- statements expressing it have "no direct implications of the form ‘I shall experience C rather than D’" (Bennett, 1974, p. 80). If this is so, the linguistic activities using first person pronouns by which we refer to ourselves as subject and the representational states that result have to have some unusual properties.

2.4 Unity of focus

Finally, as we said, we need to distinguish a fourth site of unified consciousness. Let us call it unity of focus. Unity of focus is our ability to pay unified *attention* to objects, one's representations, and one's own self. It is different from the other sorts of unified consciousness. In the other three situations, consciousness ranges over either many objects or many instances of consciousness of an object (in unified consciousness of oneself). Unity of focus picks out one such item (or a small number of them). Wundt captures what I have in mind well in his distinction between the field of consciousness *(Blickfeld)* and the focus of consciousness *(Blickpunkt).* The consciousness of a single item on which one is focusing is unified because one is aware of many aspects of the item in one state or act of consciousness (especially relational aspects, e.g., any dangers it poses, how it relates to one's goals, etc.) and one is aware of many different considerations with respect to it in one state or act of consciousness (goals, how well one is

achieving them with respect to this object, etc.). UC(1) does not fit this kind of unified consciousness any better than it fit unified consciousness of self. Here too we are not, or need not be, aware of a *number* of items. Rather, one is integrating a number of *properties* of an item, especially properties that involve *relationships* to oneself, and integrating a number of one's *abilities* and applying them to the item, and so on. Call this form of unified consciousness UC(3). One way to think of the relationship of UC(3) (unified focus) to UC(1) and UC(2) is this. UC(3) occurs *within* UC(1) and UC(2) -- within unified consciousness of world and self.

Though this has often been overlooked, all forms of unified consciousness come in both simultaneous and across-time versions. That is to say, the unity can consist in links of certain kinds among phenomena occurring at the same time (synchronically) and it can consist in links of certain kinds among phenomena occurring at different times (diachronically). In its synchronic form, it consists in such things as our ability to compare items to one another, for example, to see if an item fits into another item. Diachronically, it consists in a certain crucial form of memory, namely, our ability to retain a representation of an earlier object in the right way and for long enough to bring it as recalled into current consciousness of currently represented objects in the same as we do with simultaneously represented objects. Though this process across time has always been called the unity of consciousness, sometimes even to the exclusion of the synchronic unity just delineated, another good name for it would be *continuity of consciousness.* Note that this process of relating earlier to current items in consciousness is more than, and perhaps different from, the learning of new skills and associations. Even severe amnesiacs can do the latter.

That consciousness can be unified across time as well as at a given time points up just how central unity of consciousness is to cognition. Without the ability to retain representations of earlier objects and unite them with current represented objects, most complex cognition would simply be impossible. The only bits of language that one would be able to understand, for example, would be single words; the simplest of sentences is an entity spread over time. Now, unification *in consciousness* might not be the only way to unite earlier cognitive states (earlier thoughts, earlier experiences) with current ones but it is certainly a central way and the one best known to us. The unity of consciousness is central to cognition.

2.5 Other Forms of Unity

We will close this section by noting that UC(1), UC(2) and UC(3) are not the only kinds of mental unity. Our remarks about UC(3), specifically about what can be integrated in focal attention, might already have suggested as much. There is unity in the exercise of our cognitive capacities, unity that consists of integration of motivating factors, perceptions, beliefs, etc., and there is unity in the outputs, unity that consists of integration of behavior.

Human beings bring a strikingly wide range of factors to bear on a cognitive task such as seeking to characterize something or trying to reach a decision about what to do about something. For example, we can bring to bear: what we want; what we believe; our attitudes to self, situation, and context; input from each of our various senses; information about the situation, other people, others' beliefs, desires, attitudes, etc.; the resources of however many languages we have available to us; various kinds of memory; bodily

sensations; our various and very diverse problem-solving skills; and so on. Not only can we bring all these elements to bear, we can integrate them in a way that is highly structured and ingeniously appropriate to our goals and the situation(s) before us. This form of mental unity could appropriately be called *unity of cognition.* Unity of consciousness often goes with unity of cognition because one of our means of unifying cognition with respect to some object or situation is to focus on it consciously. However, there is at least some measure of unified cognition in many situations of which we are not conscious, as is testified by our ability to balance, control our posture, manoeuver around obstacles while our consciousness is entirely absorbed with something else, and so on.

At the other end of the cognitive process, we find an equally interesting form of unity, what we might call *unity of behavior,* our ability to coordinate our limbs, eyes, bodily attitude, etc. The precision and complexity of the behavioral coordination we can achieve would be difficult to exaggerate. Think of a concert pianist performing a complicated work.

3. Empirical Phenomena Related to Unified Consciousness

One of the most interesting ways to study psychological phenomena is to see what happens when they or related phenomena break down. Phenomena that look simple and seamless when functioning smoothly often turn out to have all sorts of structure when they begin to malfunction. Like other psychological phenomena, we would expect unified consciousness to be open to being damaged, distorted, etc., too. And if the unity of consciousness is as important to cognitive functioning as we have been suggesting, such damage or distortion should create serious problems for the people to whom it happens. The unity of consciousness is damaged and/or distorted in both naturally-occurring and experimental situations, and some of these situations are indeed very serious for those undergoing them.

In fact, unified consciousness can break down in what look to be two distinct ways. There are situations in which it is natural to say that one unified conscious being has split into two unified conscious beings without the unity itself being destroyed or even significantly damaged, and situations in which at all times we have one being with one instance of consciousness but the unity itself appears to be damaged or even destroyed. In the former cases, there is reason to think that a single instance of unified consciousness has become two (or something like two). In the latter cases, unity of consciousness has been compromised in some way but nothing suggests that anything has split.

3.1 Split Consciousness

First, situations in which we are inclined to say that something has split. Some such description seems natural in at least three different kinds of situation.

(1) Brain Bisection Operations

One is ‘brain bisection’ operations (commissurotomies), specifically certain results of them. In these operations, the corpus callosum is cut to stop the spread of epileptic seizures from one hemisphere to the other. The corpus callosum is a large strand of about 200,000,000 neurons running from one hemisphere to the other. When present, it is the chief channel of communication between the hemispheres. These operations, done mainly in the 1960s, were a last-ditch effort to control certain kinds of severe epilepsy by stopping seizures in one lobe of the cerebral cortex from spreading to the other lobe. Under certain laboratory conditions, two ‘centers of consciousness’ seem to appear in patients who have had this operation (Nagel, 1971; Marks, 1981). Here is a couple of examples of the kinds of behavior that prompt such an assessment.

The human retina is split vertically in such a way that the left half of the retina is primarily hooked up to the left hemisphere of the brain and the right half of the retina is hooded up to the right hemisphere of the brain. Now suppose that we flash the word TAXABLE on a screen in front of a brain bisected patient in such a way that the letters TAX hit one side of the retina, the letters ABLE the other and we put measures in place to ensure that the information hitting each retina stays in one lobe and is not fed to the other. If such a patient is asked what word is being shown, the mouth will say TAX while the hand controlled by the hemisphere that does not control the mouth (usually the left hand) will write ABLE. Or, if the hemisphere that controls a hand is asked to do arithmetic in a way that does not penetrate to the hemisphere that controls the mouth and the hands are shielded from the eyes, the mouth will insist that it is not doing arithmetic, has not even thought of arithmetic today, etc., -- while the appropriate hand is busily doing arithmetic! Notice that since the two ‘centers’ coexist and are active at the same time, whatever breach of unified consciousness there is in these cases is a breach of synchronically unified consciousness. These operations have received a huge amount of attention from philosophers in the past few decades and we will return to them.

(2) Hemi-neglect

Another phenomenon where we may find something like a split without diminished or destroyed unity is hemi-neglect, the strange phenomenon of losing all sense of one side of one's body or sometimes a part of one side of the body. Whatever it is exactly that is going on in hemi-neglect, unified consciousness remains. It is just that its ‘range’ has been bizarrely circumscribed. It ranges over only half the body (in the most common situation), not seamlessly over the whole body. Where we expect proprioception and perception of the whole body, in these patients they are of (usually) only one-half of the body.

(3) Dissociative Identity Disorder

A third candidate phenomenon is what used to be called Multiple Personality Disorder, now, more neutrally, Dissociative Identity Disorder (DID). Everything about this phenomenon is controversial, including whether there is any real multiplicity *of consciousness* at all, but one common way of describing what is going on in at least some central cases is to say that the units (whether we call them persons, personalities, sides of a single personality, or whatever) ‘take turns’, usually with pronounced changes in personality. When one is active, the other(s) usually is(are) not. If this is an accurate description, then here

too we have a breach in unity of some kind in which unity is nevertheless not destroyed. Notice that whereas in brain bisection cases the breach, whatever it is like, is synchronic (at a time), here it is diachronic (across time), different unified ‘packages’ of consciousness taking turns, and the breach consists primarily in some pattern of reciprocal (or sometimes one way) amnesia -- some pattern of each ‘package’ not remembering having the experiences or doing the things had or done when another ‘package’ was in charge.

3.2 Shattered Consciousness

By contrast to brain bisection and DID cases, there are phenomena in which unified consciousness does not seem to split and does seem to be damaged or even destroyed altogether. In brain bisection and dissociative identity cases, the most that is happening is that unified consciousness is splitting into two or more relatively intact units -- two or more at a time or two or more across time. It is a matter of controversy whether even that is happening, especially in DID cases, but we clearly do not have more than that. In particular, the unity itself does not disappear; while it may split, we could say, it does not shatter. There are at least three kinds of case in which unity does appear to shatter.

(1) Schizophrenia

One is some particularly severe forms of schizophrenia. Here the victim seems to lose the ability to form an integrated, interrelated representation of his or her world and his or her self altogether. The person speaks in ‘word salads’ that never get anywhere, indeed sometimes never become complete sentences. The person is unable to put together integrated plans of actions even at the level necessary to obtain sustenance, tend to bodily needs, or escape painful irritants. And so on. Here, it seems more correct to say that unity of consciousness has shattered than split. The behavior of these people seems to express no more than what we might call experience-fragments, each lasting a tiny length of time and unconnected to any others. In particular, except for the (usually semantically irrelevant) associations that lead these people from each entry to the next in the word salads they create, to be aware of one of these states is not to be aware of any others -- or so the evidence suggests.

(2) Dysexecutive Syndrome

In schizophrenia of this sort, the shattering of unified consciousness is part of a general breakdown or deformation of mental functioning: affect, desire, belief, even memory all suffer massive distortions. In another kind of case, the normal unity of consciousness seems to be just as absent but there does not seem to be general disturbance of the mind. This is what some researchers call dysexecutive syndrome (Dawson, 1998, p. 215). What characterizes the breakdown in the unity of consciousness here is that subjects are unable to consider two things together, even things that are directly related to one another. For example, such people cannot figure out whether a piece of a puzzle fits into a certain place even when the piece and the puzzle are both clearly visible and the piece obviously fits. They cannot crack an egg into a pan. And so on.

(3) Simultagnosia

A disorder presenting similar symptoms is simultagnosia or Balint's syndrome (Balint was an earlier 20th century German neurologist). In this disorder, which is fortunately rare, patients see only one object located at one ‘place’ in the visual field at a time. Outside of a few ‘degrees of arc’ in the visual field, these patients say they see nothing and seem to be receiving no information (Hardcastle, in progress). In both dysexecutive disorder and simultagnosia (if we have two different phenomena here), subjects seem not to be aware of even two items in a single conscious state.

We can pin down what is missing in each case a bit more precisely. Recall the distinction between being conscious of individual objects and having unified consciousness of a number of objects at the same time introduced at the beginning of this article. Broadly speaking, we can think of the two phenomena isolated by this distinction as two stages. First, the mind ties together various sensory inputs into representations of objects. In contemporary cognitive research, this activity has come to be called binding (Hardcastle 1998 is a good review). Then, the mind ties these represented objects together to achieve unified consciousness of a number of them at the same time. (The first theorist to separate these two stages was Kant, in his doctrine of synthesis.) The first stage continues to be available to dysexecutive and simultagnosia patients: they continue to be aware of individual objects, events, etc. The damage seems to be to the second stage: it is the tying of objects together in consciousness that is impaired or missing altogether. The distinction can be made this way: these people can achieve some UC(S), unity of focus with respect to individual objects, but little or no unified consciousness of any of the three kinds over a number of objects.

The same distinction can also help make clear what is going on in the severe forms of schizophrenia just discussed. Like dysexecutive syndrome and simultagnosia patients, severe schizophrenics lack the ability to tie represented objects together, but they also seem to lack the ability to form unified representations of individual objects. In a different jargon, these people seem to lack even the capacity for object constancy. Thus their cognitive impairment is much more severe than that experienced by dysexecutive syndrome and simultagnosia patients.

With the exception of brain bisection patients, who do not evidence distortion of consciousness outside of specially contrived laboratory situations, the split or breach occurs naturally in all the patients just discussed. Indeed, they are a central class of the so-called ‘experiments of nature’ that are the subject-matter of contemporary neuropsychology. Since all the patients in whom these problems occur naturally are severely disadvantaged by their situation, this is further evidence that the ability to unify the contents of consciousness is central to proper cognitive functioning.

3.3 What Happens When Unified Consciousness Breaks Down?

Is there anything common to the six situations of breakdown in unified consciousness just sketched? And how do they relate to UC(1), UC(2) or UC(3)?

In brain bisection cases, the key evidence for a duality of some kind is that there are situations in which whatever is aware of some items being represented in the body in question is not aware of other items being represented in that same body at the same time. We looked at two examples of the phenomenon in Section 3 in connection with the word TAXABLE and the doing of arithmetic. With respect to these represented items, there is a significant and systematically extendable situation in which to be aware of some of these items is *not* to be aware of others of them. This seems to be what motivates the judgment in us that these patients evidence a split in unified consciousness. If so, brain bisection cases are a straightforward case of a failure to meet the conditions for UC(1). However, they are more than that. Because the ‘centers of consciousness’ created in the lab do not communicate with one another except in the way that any mind can communicate with any other mind, there is also a breakdown in UC(2). One subject of experience aware of itself as the single common subject of its experience seems to become two (in some measure at least).

In DID cases, a central feature of the case is some pattern of amnesia. Again, this is a situation in which being conscious of some represented objects goes with not being conscious of others in a systematic way. The main difference is that the breach is at a time in brain bisection cases, across time in DID cases. So again the breakdown in unity consists in a failure to meet the conditions for UC(1). However, DID being diachronic, there is also a breakdown in UC(2) across time -- though there is continuity across time within each personality, there seems to be little or no continuity, conscious continuity at any rate, from one to another.

The same pattern is evident in the cases of severe schizophrenia, dysexecutive disorder and simultagnosia that we considered. In all three cases, consciousness of some items goes with lack of consciousness of others. In these cases, to be aware of a given item is precisely *not* to be aware of other relevant items. However, in the severe schizophrenia cases we considered, there is also a failure to meet the conditions of UC(3).

Hemi-neglect is a bit different. Here we do not have two or more ‘packages’ of consciousness and we do not have individual conscious states that are not unified with other conscious states. (Not so far as we know -- for there to be conscious states not unified with the states on which the patient can report, there would have to be consciousness of what is going on in the side neglected by the subject with whom we can communicate and there is no evidence for this.) Here none of the conditions for UC(1), UC(2) or UC(3) fail to be met -- but that may be because hemi-neglect is not a *split or a breakdown* in unified consciousness in the first place. It may be simply a shrinking of the range of phenomena over which otherwise intact unified consciousness extends.

It is interesting that none of the breakdown cases we have considered evidence damage to or destruction of the unity in UC(2). We have seen cases in which unified consciousness it might split at a time (brain bisection cases) or over time (DID cases) but not cases in which the unity itself is significantly damaged or destroyed. Nor is our sample unrepresentative; the cases we have considered are the most widely discussed cases in the literature. There do not seem to be many cases in which it is plausible to say that UC(2), awareness of oneself as a single common subject, has been damaged or destroyed.

4. Recent Philosophical Work

4.1 The Kantian Approach

After a long hiatus, serious work on the unity of consciousness began in recent philosophy with two books on Kant, P. F. Strawson (1966) and Jonathan Bennett (1966). Both of them had an influence far beyond the bounds of Kant scholarship. (A second book by Bennett (1974) should also be mentioned). Central to these works is an exploration of the relationship between unified consciousness, especially unified consciousness of self, and our ability to form an integrated, coherent representation of the world, a linkage that the authors took to be central to Kant's transcendental deduction of the categories. Whatever the merits of the claim (see Brook 1994 for a skeptical judgment), their work set off a long line of writings on the supposed link. Sydney Shoemaker (1984 and 1996), Karl Ameriks (1983), Paul Guyer (1987), Patricia Kitcher (1990), and Quassim Cassam (1997) are some of the philosophers who have worked on unified consciousness from a broadly Kantian point of view. Quite recently the approach prompted a debate about unity and objectivity among Michael Lockwood, Susan Hurley and Anthony Marcel in Peacocke (1994) (note that ‘unity of consciousness’ even occurs in the book's title).

4.2 Philosophical Reflections on Brain Bisection Operations

Another issue that led philosophers back to the unity of consciousness, perhaps the next historically, was the neuropsychological results of brain bisection operations that we explored earlier. Starting with Thomas Nagel (1971) and continuing in the work of Charles Marks (1981), Derek Parfit (1971 and 1984), Lockwood (1989), Hurley (1998) and many others, these operations have been a major theme in work on the unity of consciousness since the 1970s. Much ink has been spilled on the question of what exactly is going on in the phenomenology of brain bisection patients. Nagel goes so far as to claim that there is no whole number of ‘centers of consciousness’ in these patients: there is too much unity to say "two", yet too much splitting to say "one".

Some recent work by Jocelyne Sergent (1990) might seem to support this conclusion. She found, for example, that when a sign ‘6’ was sent to one hemisphere of the brain in these subjects and a sign ‘7’ was sent to the other in such a way that crossover of information from one hemisphere to the other was extremely unlikely, they could say that the 6 is a smaller number than the 7 but could not say whether the signs were the same or different. It is not certain that Sergent's work does support Nagel's conclusions. First, Sergent's claims are controversial -- not all researchers have been able to replicate them. Second, even if the data are good, the interpretation of them is far from straightforward. In particular, they seem to be consistent with there being a clear answer to any *precise* ‘one or two?’ question that we could ask. (’Unified consciousness of the two signs with respect to numerical size?’ Yes. ‘Unified consciousness of the visible structure of the signs?’ No). If so, it is not obvious that the fact that there is mixed evidence, some pointing to the conclusion ‘one’, some pointing to the conclusion ‘two’, supports the view expressed by Nagel that there may be no whole number of subjects that these patients are.

Much of the work since Nagel has focused on the same issue of the kind of split that the laboratory manipulations of brain bisection patients induces. Some attention has also been paid to the implications of these splits. For example, could one hemisphere commit a crime in such a way that the other could not justifiably be held responsible for it? Or, if such splitting occurred regularly and was regularly followed by merging with ‘halves’ from other splits, what would the implications be for our traditional notion of what philosophers call ‘personal identity’, namely, being or remaining one and the same thing. (Here we are talking about identity in the philosopher's sense of being or remaining one thing, not in the sense of the term that psychologists use when they talk of such things as ‘identity crises’.)

Parfit has made perhaps the largest contributions to the issue of the implications of brain bisection cases for personal identity. Phenomena relevant to identity in things others than persons can be a matter of degree. This is well illustrated by the famous ship of Theseus example. Suppose that over the years, a certain ship in Theseus was rebuilt, board by board, until every single board in it has been replaced. Is the ship at the end of the process the ship that started the process or not? Now suppose that we take all those rotten, replaced boards and reassemble them into a ship? Is *this* ship the original ship of Theseus or not? Many philosophers have been certain that such questions cannot arise for persons ; identity in persons is completely clear and unambiguous, not something that could be a matter of degree as related phenomena obviously can be with other objects (Bishop Joseph Butler [1736] is a well-known example). As Parfit argues, the possibility of persons (or at any rate minds) splitting and fusing puts real pressure on such intuitions about our specialness; perhaps the continuity of persons can be as partial and tangled as the continuity of other middle-sized objects.

Lockwood's exploration of brain bisections cases goes off in a different direction, two different directions in fact (we will examine the second below). Like Nagel, Marks, and Parfit, Lockwood has written on the extent to which what he calls ‘co-consciousness’ can split. (’Co-consciousness’ is the term that many philosophers now use for the unity of consciousness; roughly, two conscious states are said to be co-conscious when they are related to one another as conscious states are related to one another in unified consciousness.) He also explores the possibility of psychological states that are not determinately in any of the available ‘centers of consciousness’ and the implications of this possibility for the idea of the specious present, the idea that we are directly and immediately aware of a certain tiny spread of time, not just the current infinitesimal moment of time. He concludes that the determinateness of psychological states being in an available ‘center of consciousness’ and the notion that psychological states spread over at least a small amount of time in the specious might well present stand or fall together.

4.3 Hurley on the Unity of Consciousness

Some philosophers interested in pathologies of unified consciousness examine more than brain bisection cases. In what is perhaps the most complex work on the unity of consciousness to date, Hurley examines most of the kinds of breakdown phenomena that we introduced earlier. She starts with an intuitive notion of co-consciousness that she does not formally define. She then explores the implications of a wide range of ‘experiments of nature’ and laboratory experiments for the presence or absence of co-consciousness across the psychological states of a person. For example, she considers acallosal patients (people born without a corpus callosum). When present, the corpus callosum is the chief channel of communication

between the hemispheres. When it is cut, it is possible to generate what looks like two centers of consciousness, two internally co-conscious systems that are not co-consciousness with one another. (We examined the kind of evidence that leads to this appearance in Section 3.) Hurley argues that in patients in whom it never existed, things are not so clear. Even though the channels of communication in these patients are often in part external (behavioral cuing activity, etc.), the result may still be a single co-conscious system. That is to say, the neurological and behavioral basis of unified consciousness may be very different in different people.

Hurley also considers research by Trewarthen in which a patient is conscious of some object seen by, say, the right hemisphere until her left hand, which is controlled by the right hemisphere, reaches for it. Somehow the act of reaching for it seems to obliterate the consciousness of it. Very strange -- how can something pop into and disappear from unified consciousness in this way? This leads her to consider the notion of *partial unity*. Could two centers of consciousness, A and B, though not co-conscious with one another, nonetheless both be co-conscious with some third thing, e.g., the volitional system B (the system of intentions, desires, etc.). If so, ‘co-conscious’ is not a transitive relationship -- A could be co-conscious with B and C could be co-conscious with B without A being co-conscious with C. This is puzzling enough. Even more puzzling would be the question of how activation of the system B with which both A and C are co-conscious could result in either A or C ceasing to be conscious of an object aimed at by B.

Hurley's response to Trewarthen's cases (and also Sergent's cases that we examined in the previous section) is to accept that intention can obliterate consciousness and then distinguish time periods (1998, p. 216). At any given time in Trewarthen's cases, the situation with respect to unity is clear. That the picture does not conform to our usual expectations for diachronic singularity or transitivity then becomes simply an artefact of the cases, not a problem. It is not made clear how this reconciles Sergent's evidence with unity. One strategy would be the one we considered earlier: make the questions more precise. For precise questions, there seems to be a coherent answer about unity for every phenomenon Sergent describes.

Hurley also considers what she calls Marcel's case. Here subjects are asked to report the appearance of some item in consciousness in three ways at the same time -- say, by blinking, pushing a button, and saying, ‘I see it’. Remarkably, any of these acts can be done without the other two. And the question is, What does this imply for unified consciousness? In a case in which the subject pushes the button but neither blinks nor says anything, for example, is the hand-controller aware of the object while the blink-controller and the speech-controller are not? How could the conscious system become fragmented in such a way?

Hurley's suggestion is: they can't. What induces the appearance of incoherence about unity is the short time scale. Suppose that it takes some time to achieve unified consciousness, perhaps because some complex feedback processes are involved. If that were the case, then we do not have a stable unity situation in Marcel's case. The subjects are not given enough time to achieve unified consciousness of any kind (1998, p. 216).

There is a great deal more to Hurley's work. She urges, for example, that there a normative dimension to unified consciousness -- conscious states have to cohere for unified consciousness to result. And systems

in the brain have to achieve she calls ‘dynamic singularity’ -- being a single system -- for unified consciousness to result.

4.4 The Binding Problem

A third issue that got philosophers working on the unity of consciousness again is binding (again, see Hardcastle's 1998 review). Here the connection is more distant because binding as usually understood is not unified consciousness as we have been discussing it. Recall the two stages of cognition laid out earlier. First, the mind ties together various sensory inputs into representations of objects. Then the mind ties these represented objects to one other to achieve unified consciousness of a number of them at the same time. It is the first stage that is usually called binding. The representations that result at this stage need not be conscious in any of the ways delineating earlier -- many perfectly good representations affect behavior and even enter memory without ever becoming conscious. Representations resulting from the second stage need not be conscious, either, but when they are, we have at least some of the kinds of unified consciousness delineated in Section 2.

4.5 Neurophysiology and the Unity of Consciousness

In the past few decades, philosophers have also worked on how unified consciousness relates to the brain. Lockwood, for example, thinks that relating consciousness to matter will involve more issues on the side of matter than most philosophers think. (We mentioned that his work goes off in two new directions. This is the second one.) Quantum mechanics teaches us that the way in which observation links to physical reality is a subtle and complex matter. Lockwood urges that our conceptions will have to be adjusted on the side of matter as much as on the side of mind if we are to understand consciousness as a physical phenomenon and physical phenomena as open to conscious observation. If it is the case not only that our understanding of consciousness is affected by how we think it might be implemented in matter but also that processes of matter are or can be affected by our (conscious) observation of them, then our picture of consciousness stands as ready to affect our picture of matter as vice-versa.

The Churchlands, Paul M. and Patricia S. (see for example P. M. Churchland 1995, p. 214), and Daniel Dennett (1991) have fairly radical views of the underlying architecture of unified consciousness. The Churchlands see unity itself much as other philosophers do. They do argue that the term ‘consciousness’ covers a range of different phenomena that need to be distinguished from one another but the important point here is that they urge that the architecture of the underlying processes probably consist not of transformations of symbolically encoded objects of representations, as most philosophers have believed, but of vector transformations in what are called phase spaces. Dennett articulates an even more radical view, encompassing both unity and underlying architecture. For him, unified consciousness is simply a temporary ‘virtual captain’, a small group of related information-parcels that happens to gain temporary dominance in a struggle for control of such cognitive activities as self-monitoring and self-reporting in the vast array of microcircuits of the brain. We take these transient phenomena to be more than they are because each of them is the ‘me’ of the moment; the temporary coalition of conscious states winning at the moment is what I am, is the self. Radical implementation, narrowed range and transitoriness

notwithstanding, when unified consciousness is achieved, these philosophers tend to see it in the way we have presented it.

Dennett's and the Churchlands' views fit naturally with a dynamic systems view of the underlying neural implementation. The dynamic systems view is the view that unified consciousness is a result of certain self-organizing activities in the brain. Dennett thinks that given the nature of the brain, a vast assembly of neurons receiving electrochemical signals from other neurons and passing such signals to yet other neurons, cognition could not take any form other than something like a pandemonium of competing bits of content, the ones that win the competition being the ones that are conscious. The Churchlands don't tend to agree with Dennett about this. They see consciousness as a state of the brain, the ‘wetware’, not a result of information processing, of ‘software’. They also advocate a different picture of the underlying neurological process. As we said, they think that transformations of complex vectors in a multi-dimensional phase space are the crucial processes, not competition among bits of content. However, they agree that it is very unlikely that the processes that subserve unified consciousness are sentence-like or language-like at all. It is too early to say whether these radically novel pictures of what the system that implements unified consciousness is like will hold any important implications for what unified consciousness is or when it is present.

Hurley is also interested in the relationship of unified consciousness to brain physiology. It would be truer to say of her that she resists certain standard ways of linking them, however, than to say that she herself links them. In particular, while she clearly thinks that physiological phenomena have all sorts of implications and give rise to all sorts of questions about the unity of consciousness, she strongly resists any simplistic patterns of connection. Many researchers have been attracted by some variant of what she calls the *isomorphism hypothesis*. This is the idea that changes in consciousness will parallel changes in brain structure or function. She wants to insist, to the contrary, that often two instances of exactly the same change in consciousness will go with very different changes in the brain. We saw an example in the last section. In most of us, unified consciousness is closely linked to an intact, functioning corpus callosum. However, in acallosal people, there may be the same unity but achieved by mechanisms such as cuing activity external to the body that are utterly different from communication though a corpus callosum. Going the opposite way, different changes in consciousness can go with the same changes to structure and function in the brain.

4.6 Other Work

Two philosophers have gone off in directions different from any of the above, Stephen White (1991) and Christopher Hill (1991). White's main interest is not the unity of consciousness as such but what one might call the unified locus of responsibility -- what it is that ties something together to make it a single agent of actions, i.e., something to which attributions of responsibility can appropriately be made. He argues that unity of consciousness is one of the things that go into becoming unified as such an agent but not the only thing. Focused coherent plans, a continuing single conception of the good, reasonably good autobiographical memory, certain future states of persons mattering to us in a special way (mattering to us because we take them to be future states of ourselves, one would say if it were not blatantly circular), a certain continuing kind and degree of rationality, certain social norms and practices, and so on and so

forth. In his picture of moral responsibility, unbroken unity of consciousness at and over time is only a small part of the story.

Hill's fundamental claim is that a number of different relationships between psychological states have a claim to be considered unity relationships, including: being owned by the same subject, being [phenomenally] next to (and other relationships that states in the field of consciousness appear to have to one another), both being the object of a single other conscious state, and jointly having the appropriate sorts of effects (functions). An interesting question, one that Hill does not consider, is whether all these relations are what interests us when we talk about the unity of consciousness or only some of them (and if only some of them, which ones). Hill also examines scepticism about the idea that clearly bounded individual conscious states exist. Since we have been assuming throughout that such states do exist, it is perhaps fortunate that Hill argues that we could safely do so.

In some circles, the idea that consciousness has a special kind of unity has fallen into disfavor. Nagel (1971), Donald Davidson (1982), and Dennett (1991) have all urged that the mind's unity has been greatly overstated in the history of philosophy. The mind, they say, works mostly out of the sight *and the control* of consciousness. Moreover, even states and acts of ours that are conscious can fail to cohere. We act against what we know perfectly well to be our own most desired course of action, for example, or do things while telling ourselves that we must avoid doing them. There is an approach to the small incoherencies of everyday life that does not requires us to question whether consciousness is unified in this way, the Freudian approach (e.g., Freud 1916/17). This approach accepts that the unity of consciousness exists much as it presents itself but argues that the range of material over which it extends is much smaller than philosophers once thought. This latter approach has some appeal. If something is out of sight and/or control, it is out of the sight or control of what? The answer would seem to be, the unified conscious mind. If so, the only necessary difference between the pre-twentieth century vision of unified consciousness as ranging over everything in the mind and our current vision of unified consciousness is that the range of psychological phenomena over which unified consciousness ranges has shrunk.

A final historical note. At the beginning of the 21st century, work on the unity of consciousness continues apace. For example, a major conference was recently devoted to the unity of consciousness, the *Association for the Scientific Study of Consciousness Conference* held in Brussels in 2000 *(ASSC5)* Encyclopedias of philosophy (such as this one) and of cognitive science are commissioning articles on the topic. Psychologists are taking up the issue. Bernard Baars (1988, 1997) notion of the global workspace is an example. Another example is work on the role of unified consciousness in precise control of attention. However, the topic is not yet at the center of consciousness studies. One illustration of this is that it can still be missing entirely in anthologies of current work on consciousness.

5. A Background Question

We will close with a different issue. As we saw, philosophers used to think that the unity of consciousness has huge implications for the nature of the mind, indeed entails that the mind could not be made out of matter. We also saw that the prospects for this inference are not good. What about the *nature* of

consciousness? Does the unity of consciousness have any implications for this issue?

There are currently at least three major camps on the nature of consciousness. One camp sees the ‘felt quality’ of representations as something quite unique, in particular as quite different from the power of representations to change other representations and shape belief and action. On this picture, representations could function much as they do without it being like anything to have them. They would merely not be conscious. If so, consciousness may not play any important cognitive role at all, its unity included (Jackson 1986; Chalmers 1996). A second camp holds, to the contrary, that consciousness is simply a special kind of representation (Rosenthal 1991; Dretske 1995; Tye 1995). And a third holds that what we label ‘consciousness’ is really something else. On this view, consciousness will in the end be ‘analyzed away’ -- the term is too coarse-grained and presents things in too unquantifiable a way to have any use in a mature science of the mind (P. S. Churchland 1983).

It is not obvious that the unity of consciousness has strong implications for the truth or falsity of any of these views. If it is as central and undeniable as many have suggested (we saw some of the arguments earlier), its existence may cut against the third, eliminativist position a bit. With respect to the other two positions, the unity of consciousness seems neutral.

Whatever its implications for other issues, the unity of consciousness seems to be a real feature of the human mind, indeed central to it. If so, any complete picture of the mind will have to provide an account of it. Even those who hold that the extent to which consciousness is unified has been overrated owe us an account of *what* has been overrated.

Bibliography

- Ameriks, K. 1983. *Kant's Theory of Mind*. Oxford: Oxford University Press.
- Baars, B. 1988. *A Cognitive Theory of Consciousness.* Cambridge: Cambridge University Press.
- Baars, B. 1997. *In the Theatre of Consciousness: The Workspace of the Mind*. Oxford: Oxford University Press
- Bennett, J. 1966. *Kant's Analytic*. Cambridge: Cambridge University Press.
- Bennett, J. 1974. *Kant's Dialectic*. Cambridge: Cambridge University Press.
- Block, N. 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18: 227-47.
- Brook, A. 1994. *Kant and the Mind*. New York: Cambridge University Press.
- Butler, J. 1736. Dissertation I: Of Personal Identity. In his *Analogy of Religion*. London: Dent, 1906.
- Castañeda, Hector-N. 1966. ‘He’: A study in the logic of self-consciousness. *Ratio* 8, pp.130-57.
- Chalmers, D. 1996. *The Conscious Mind*. Oxford: Oxford University Press.
- Churchland, P. M. 1995. *The Engine of Reason, the Seat of the Soul*. MIT Press/A Bradford Book
- Churchland, P. S. 1983. Consciousness: the transmutation of a concept. *Pacific Philosophical Quarterly* 64: 80-95.
- Davidson, D. 1980. Paradoxes of irrationality. In R. Wollheim and J. Hopkins, eds. *Philosophical*

Essays on Freud. Cambridge: Cambridge University Press.

- Dawson, M. 1998. *Understanding Cognitive Science.* Oxford: Blackwell Publishers.
- Dennett, D. 1978. Toward a cognitive theory of consciousness. In *Brainstorms.* Montgomery, VT: Bradford Books, pp. 149-73.
- Dennett, D. 1991. *Consciousness Explained.* New York: Little, Brown.
- Descartes, René. 1641. *Meditations on First Philosophy.* E. S. Haldane and G. R. T. Ross, trans. In *The Philosophical Works of Descartes,* Vol. 1. Cambridge: Cambridge University Press, 1970.
- Dretske, F. 1995. *Naturalizing the Mind*. MIT Press/A Bradford Book.
- Freud, S. 1916/17. *Lectures on Psychoanalysis. The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vols.. XV and XVI. James Strachey, trans. and ed. London: Institute of Psychoanalysis and the Hogarth Press.
- Guyer, P. 1987. *Kant and the Claims of Knowledge.* Cambridge: Cambridge University Press
- Hardcastle, V. 1998. The binding problem. In *A Companion to Cognitive Science*, Wm. Bechtel and G. Graham, eds. Oxford: Blackwell Publishers.
- Hardcastle, V. In progress. Attention versus consciousness: a distinction with a difference. http://www.phil.vt.edu/Valerie/papers/attencons.html.
- Hill, C. 1991. *Sensations: A Defense of Type Materialism.* New York: Cambridge University Press.
- Hurley, S. 1994. Unity and Objectivity. In Peacocke (1994).
- Hurley, S. 1998. *Consciousness in Action.* Cambridge, MA: Harvard University Press.
- Jackson, F. 1986. What Mary didn't know. *Journal of Philosophy* 83(5): 291-5.
- James, W. 1890. *Principles of Psychology,* two volumes. London: Macmillan.
- Kant, I. 1781/87. *Critique of Pure Reason*. P. Guyer and A. Wood, trans. and eds. Cambridge: Cambridge University Press (cited as Axxx for the first edition of 1781and Bxxx for the second of 1787).
- Kitcher, P. 1990. *Kant's Transcendental Philosophy.* New York: Oxford University Press.
- Lockwood, M. 1989. *Mind, Brain and the Quantum.* Oxford: Blackwell Publishers.
- Lockwood, M. 1994. Issues of Unity and Objectivity. In Peacocke (1994).
- Marcel, A. 1994. What is Relevant to the Unity of Consciousness? In Peacocke (1994).
- Marks, C. 1981. *Commissurotomy, Consciousness and Unity of Mind.* Cambridge, MA: MIT Press.
- Nagel, T. 1965. Physicalism. *Philosophical Review* 74: 339-56.
- Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese* 22: 396-413.
- Nagel, T. 1974. What it is like to be a bat. *Philosophical Review* 83: 435-50.
- Parfit, D. 1971. Personal Identity. *Philosophical Review* 80: 3-27.
- Parfit, D. 1984. *Reasons and Persons.* Oxford: Oxford University Press.
- Peacocke, C. 1994. *Objectivity, Simulation, and the Unity of Consciousness* Oxford: Oxford University Press (for the British Academy).
- Perry, J. 1979. The problem of the essential indexical. *Noûs* 13: 3-21.
- Sergent, J. 1990. Furtive incursions into bicameral minds. *Brain* 113: 537-68.
- Rosenthal, D. 1991. *The Nature of Mind.* Oxford: Oxford University Press.
- Shoemaker, S. 1968. Self-reference and self-awareness. *Journal of Philosophy* 65: 555-67.
- Shoemaker, S. 1984. Commentary: Self-consciousness and Synthesis. In A. Wood, ed. *Self and*

Nature in Kant's Philosophy. Ithaca, NY: Cornell University Press, pp. 148-155.

- Shoemaker, S. 1996. Unity of consciousness and consciousness of unity. In *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.
- Strawson, P. F. 1966. *The Bounds of Sense*. London: Methuen Ltd.
- Tye, M. 1995. *Ten Problems of Consciousness*. Cambridge, MA: MIT Press/A Bradford Book.
- White, S. 1990. *The Unity of the Self*. Cambridge, MA: MIT Press/A Bradford Book.

Other Internet Resources

- Bibliography on the Unity of Consciousness, by David Chalmers (U. Arizona)
- Attention versus consciousness: a distinction with a difference, a paper in progress, by V. Hardcastle (Virginia Polytechnic Institute and State University)

Related Entries

consciousness | qualia

Andrew Brook
abrook@ccs.carleton.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 27, 2001
Content last modified: March 27, 2001

Stanford Encyclopedia of Philosophy

Mental Imagery

Mental imagery (sometimes colloquially called visualization, or "seeing in the mind's eye") is experience that resembles perceptual experience, but which occurs in the absence of the appropriate stimuli for the relevant perception (cf. Finke, 1989; McKellar, 1957). Very often these experiences are understood by their subjects as echoes or reconstructions of actual perceptual experiences from their past; at other times they may seem to anticipate possible, often desired or feared, future experiences. Thus imagery has often been believed to play a very large, even pivotal, role in both memory (Yates, 1966; Paivio, 1986) and motivation (McMahon, 1973). It is also commonly believed to be centrally involved in visuo-spatial reasoning and inventive or creative thought. Indeed, it has usually been regarded as crucial for *all* thought processes, although, during the 20th century in particular, this has been called into question.

1. Terminological and Definitional Problems

We have defined mental imagery as a form of experience, but, of course, evidence for the occurrence of any experience is necessarily subjective. Because of this, some authors, most notably the arch-behaviorist J.B. Watson (1913a), have cast doubt on the scientific status and even the existence of imagery. However, if imagery serves certain functions in our mental life (as suggested above) then perhaps some

objective validation and study of it might be possible through the study of the performance of these functions. In the light of this, some authors (notably the psychologist Stephen Kosslyn, who is probably the most influential contemporary imagery theorist) prefer an alternative definition of "imagery" to that given above. Instead of understanding it primarily as a sort of experience, they prefer to view the term as referring to the particular type of cognitive process or "underlying representation" (Kosslyn, 1983) that is involved in these functions. These representations or processes are generally understood to be such that their presence or activity can (but need not always) be consciously experienced as imagery in our original sense.

However, characterizing imagery in this way (as explanans rather than explanandum) begs important questions about the nature of the mind and about the causes of imagery experiences (conceivably they are *not* experiences of cognitive processes or underlying representations). On the other hand, it should be admitted that defining imagery as a form of experience, is also problematic, and might deflect attention away from the possibility that importantly similar underlying representations or mechanisms may be operative both when we experience imagery *and* during certain unconscious mental processes (some evidence suggests that this is so). To avoid such problems we might replace "imagery" with some special jargon: we could speak of "quasi-perceptual experiences" on the one hand and "image representations (or processes)" on the other. However, this is not an established convention, and using these terms exclusively throughout this entry would seriously complicate discussion of the views of those thinkers (probably the vast majority) who fail to disentangle these notions. Thus, the (more or less) ordinary language term "imagery" will continue to be used where appropriate.

But our initial definition of "imagery" may well be thought unsatisfactory even in its own terms. Not only does it duck the difficult task of specifying what dimensions and degrees of similarity to perception are necessary for an experience to count as imagery; it also elides the controversial question of whether imagery is a *sui generis* phenomenon, conceptually quite distinct from true perceptual experience despite the surface resemblance, or whether it is more appropriately regarded as lying at one end of a continuum stretching from ordinary veridical perception at one end, to ‘pure’ imagery, where the character of the experience seems to be quite independent of any current stimulus input, at the other. In between would come cases, often held to be due to the effects of imagination, where the character of the experience seems to be only *partially* determined by the character of the current stimulus: both mistaken or illusive perception and non-deceptive *seeing as* (such as seeing the notorious duck-rabbit figure *as* a duck [or rabbit], or, for example, "seeing" the shapes of animals, or whatever, in the clouds or constellations). Many philosophers and cognitive theorists implicitly take this line, treating *percepts* as, essentially, special cases of imagery, differing only in causal history and, perhaps, "vivacity". For example: for Descartes (in the *Treatise on Man*) both images and percepts are ultimately embodied as pictures picked out on the surface of the pineal gland by the flow of animal spirits; for Kosslyn (1994) both are depictive representations in the brain's "visual buffer"; for Hinton (1979) both are "structural descriptions" in working memory. However, other theorists (e.g. Sartre, 1936) try to draw a sharp conceptual and phenomenological distinction between perceptual and imaginal experience.

But in the absence of consensus about such issues, or about the underlying mechanisms and the psychological functions of imagery, our initial rough characterization is probably about the best we can

do without begging important questions. Perhaps it is sufficient. Imagery is a common, everyday phenomenon that is indicated by a whole range of colloquial expressions: "having a picture in the head", "picturing", "visualizing", "having/seeing a mental image/picture", "seeing in the mind's eye", and, in some contexts, simply "imagining". Although a small percentage of people seem inclined to deny ever experiencing it, for the vast majority of us, our imagery, like our consciousness itself, is something with which we seem to be thoroughly familiar and intimate.

However, the term "mental imagery", and all the colloquial equivalents mentioned above, may be potentially misleading in itself. For one thing, all these expressions suggest, more or less strongly, a purely visual phenomenon. In fact, most discussions of imagery, in the past and today, have indeed focused upon the visual mode. Nevertheless, there is every reason to believe that other modes of quasi-perceptual experience are just as common and important (Newton, 1982), and "imagery" has come to be the accepted scientific term for referring to them too: interesting studies of "auditory imagery", "kinaesthetic imagery", "haptic (touch) imagery", and so forth, can be found in the contemporary psychological literature.

A related, and perhaps a more serious problem with the term "imagery" and with most of the colloquial alternatives is that they strongly suggest that the phenomenon involves some sort of *picture* (*the image*) entering into or being created in the mind. Indeed, this theoretical story seems to have gone virtually unquestioned during past ages (which may explain how the terminology in question became entrenched), and probably remains the majority, lay and expert, view today. Nevertheless, during this century it has come under strong challenge, and can no longer be regarded as uncontroversial. The confusions arising from this (as well as the other ambiguities of the term "imagery" that we have mentioned) continue to bedevil discussions of the topic. In particular, people who deny the existence of *mental pictures* seem frequently to be misunderstood as (implausibly) denying the occurrence of quasi-perceptual experiences, and in some cases they may themselves come to believe that the first denial commits them to the second (Thomas, 1989). Indeed, there is some reason to think (although it is certainly not established) that that minority of people (about 10% of the population by some estimates) who deny ever experiencing imagery, or who deny that it plays any significant role in their mental lives, may simply be understanding the terminology in a somewhat idiosyncratic fashion: what they intend to deny may not be so much that they *have* quasi-perceptual experience, but, rather, that what they do have is predominantly visual, or that it involves inner pictures, or that it resembles perceptual experience to the extent that they (perhaps wrongly) understand other people to be claiming for *their* imagery (or some combination of these claims). This is a theoretically important issue because if it is true that some people really do not experience any imagery then imagery (understood as experience rather than representation) *cannot* play the vital role in mental life that has very often been attributed to it.

On a more consensual note, with only rare exceptions (e.g. Wright, 1983) nearly all serious discussions of imagery take it for granted (explicitly or implicitly) that it exhibits *intentionality* (i.e. imagery is normally *of* something or other, in the same sense that perception is perception *of* something), and that it is, for the most part, subject to conscious control. Although images often come into the mind unbidden, and sometimes it is hard to shake off unwanted imagery (say, of the horrible accident that one cannot get out of one's mind) in general one can conjure-up, manipulate, and dismiss images at will. In this regard,

imagery appears as an unequivocally *mental* phenomenon, quite distinct from other quasi-perceptual experiences, such as after-images and phosphenes (Oster, 1970), that are *not* subject to direct conscious control, and which are probably best explained in straightforwardly physiological terms. It is also distinguished from cognitive and representational, but nevertheless unconscious and automatic functions such as the postulated high capacity but very short term visual memory store known as "iconic memory" (Neisser, 1967). On the other hand, so called *eidetic imagery*, if, indeed, it exists at all as a distinct phenomenon (see Haber, 1979, and the appended commentaries), is probably best understood as a species of mental imagery proper, despite the fact that it is characterized by a vividness, detailed articulation, and stability that far exceed what most normal subjects seem to want to claim for their imagery experiences.

It may also be worth pointing out that mental imagery should be distinguished from "imagery" as the term has come to be used in a literary context, where it seems to refer to a linguistic trope that employs highly concrete, perceptually specific language in order to evoke certain emotions or otherwise convey some more abstract and elusive underlying sense. Very likely, literary imagery originally got its name from a supposed power of the words in question to induce mental imagery in a reader, and some contemporary literary critics defend such an interpretation (Esrock, 1994; Scarry, 1999), but it is surely not the case that the expression is now universally, or even generally, understood this way.

3. The Eclipse of Imagery in Scientific Psychology

For the late 19th century researchers who established psychology as an empirical scientific discipline, mental images (usually, in English, referred to as *ideas*) held just the same central place in the explanation of cognition that they had held for philosophical psychologists of earlier times. However, developments within psychology at the beginning of the 20th century began to cast doubt on this long established consensus. A group of psychologists working in Würzburg, Germany claimed to have found empirical evidence for conscious thought contents that were not imaginal or perceptual in character. Their results were challenged on several grounds, and were certainly never definitively established. Nevertheless, the bitter dispute that ensued, the so called *imageless thought* controversy, had a profound effect on the development of psychology as a science (and, I would argue, on philosophy also). Most psychologists became, in effect, profoundly disillusioned with the notion of mental imagery, and either avoided seriously considering the topic, treated it dismissively, or, in some extreme cases, denied the existence of the phenomenon outright. These attitudes noticeably influenced other disciplines, including philosophy. Although the psychological study of imagery revived with the rise of cognitivism in the 1960s and 70s, when new experimental techniques were developed that enabled a truly experimental study of the phenomenon, current views about, and attitudes towards, mental imagery cannot be properly understood without an awareness of this history, versions of which, of varying degrees of accuracy, have passed into the folklore of psychology.

3.1 Founders of Experimental Psychology: Wilhelm Wundt and William James

When psychology first began to emerge as an experimental science, in the philosophy departments of the German universities in the late 19th century, the central role of imagery in mental life was not in question. Wilhelm Wundt, acclaimed "the father of experimental psychology", established the first psychological research and teaching laboratory in the Leipzig Philosophy Department in around 1876 (Fancher, 1996). He regarded his psychology as a branch of philosophy, an attempt to apply the experimental method of natural science (particularly, the physiology of Helmholtz) to essentially philosophical problems concerning the nature of mind and its metaphysical status. This view of the subject persisted, in Germany, at least until the Nazi era. Wundt's research program aimed to investigate the "elements of consciousness," and the laws governing the combination of these elements (Wundt, 1912). Although his theoretical system made a place for emotional *feelings* as one class of element, in practice the main focus of Wundt's experimentally based research program was on the elements of *sensation* and their compounding into *ideas*. As has been the case in the empiricist philosophical tradition, these *ideas* were conceived of as, to all intents and purposes, mental images. Indeed, Wundt insists, much in the spirit of Hume, that there is no fundamental difference in kind between the ideas arising directly from perception and "memory images" (Wundt, 1912). Thus, Wundtian experimental psychology was largely a study of cognitive processes, and, for him (and most of his numerous students and imitators), the mental image (under the rubric *idea*) played essentially the same crucial, representational role in cognition that it had played for most of his philosophical predecessors.

Wundt's American counterpart, and contemporary, William James, took a not dissimilar view, although he was careful to acknowledge that in some people the "thought stuff," as he called it, might consist not so much of *visual* imagery as of imagery of other modes, especially the "verbal images" of inner speech (James, 1890 ch. 18). In his textbook *The Principles of Psychology* (1890) James has much that is insightful to say about psychological processes in general, and about the role of imagery in them in particular, but, although he carried out experimental demonstrations in his psychology teaching at Harvard, James had little interest in the actual pursuit of experimental research, and established no graduate teaching program in experimental psychology (Fancher, 1996). Thus, despite the lucidity of his justly famous text, and the wide readership it has continued to find, his direct influence on the disciplinary development of scientific psychology, even in his native America, probably never equaled that of Wundt (or even lesser German pioneers, such as G.E. Müller), who trained many Americans (as well as many Germans, and students of other nationalities) in the techniques of experimental research. Just around this time, when psychology was the latest intellectual fashion, the American Universities were undergoing a tremendous expansion. Thus many of these students were able to return from Germany to the United States to found experimental psychology teaching programs of their own. It was because of this, much more than the intellectual influence of James, that, well before it grew into a dominant world power and achieved its current leadership in the sciences generally, the U.S.A. quickly grew to rival, and eventually surpass, Germany's initial preeminence in scientific psychology.

Although psychologists of this era have often been portrayed (notably by Boring (1950)) as using an introspective methodology, in fact Wundt, in particular, was very sensitive to standard criticisms of introspection, such as the contention that the very attempt to observe our own mental activities will itself alter them. He thus limited its use to situations where he was satisfied that the causes of the relevant

mental events, the experimental stimuli, could be strictly controlled and the results shown to be replicable, with any introspective reports being made unreflectively, as soon as the relevant content appeared in the mind (Mischel, 1970; Danziger, 1980). Wundt's research did *not* rely upon discursive descriptions of mental contents. An "introspective" report in his laboratory might typically have involved no more than indicating the moment when a certain sensation entered consciousness, or saying whether a musical tone seemed higher or lower than the one presented just before. Such "introspective reports" differ little from the sorts of responses that might be called for in a modern cognitive psychology experiment. Wundt's methodological discipline meant that the data collected in his laboratory were primarily such things as reaction times or discrimination thresholds, rather than discursive introspective reports; it also meant, in practice, that his experiments were restricted almost entirely to the study of "lower" psychological processes, principally sensation and perception. Thus, although Wundt did hold that "higher" mental process, such as thought and memory, depended largely upon mental images (including verbal imagery, silent speech), in practice his experimental work did little directly to illuminate these. "Higher" mental processes, for Wundt, were best investigated non-experimentally, via a methodology that he called *völkerpsychologie*, a hermeneutic study of cultural products to which he devoted much of his later career, but which never achieved anything like the influence of his experimentally based work.

3.2 Edward B. Titchener: The Complete Iconophile

An Englishman, Edward B. Titchener, became one of Wundt's most influential students. After graduate studies with Wundt, Titchener moved to the United States and became professor of Psychology at Cornell, where, as well as being responsible for translating many of the more experimentally oriented works of Wundt into English, he established a successful graduate school and a vigorous research program (Tweney, 1987). Despite the fact that Wundt's and Titchener's philosophical and theoretical views, and their scientific methodologies, differed in important ways (Leahey, 1981), Titchener, much more than most of his American born colleagues, shared Wundt's vision of psychology as a pure science, with essentially philosophical rather than pragmatic ends, and he gained the reputation of being Wundt's leading disciple and representative in the English speaking world. However, he had no interest in his master's *völkerpsychologie*. Titchener had been deeply influenced by positivist optimism as to the scope of science, and he hoped to study even the "higher" thought processes experimentally (Danziger, 1979, 1980). Thus he attempted to push the method of controlled laboratory introspection far beyond the bounds that Wundt had so carefully set for it.

Titchener appears to have been both a particularly vivid imager, and a firm believer in imagery's cognitive importance. He had studied British Empiricist philosophy whilst an undergraduate at Oxford, and was well aware of Berkeley's argument that "general ideas" (i.e. mental images that, in-and-of-themselves, represent a kinds or categories of things, rather than particulars) are inconceivable. Berkeley argues that, for instance, the general idea of a triangle, which would need to be:

> neither oblique nor rectangle, neither equilateral, equicrural, nor scalenon, but *all and none* of these at once. In effect it is something imperfect that cannot exist, an idea wherein some

> parts of several different and *inconsistent* ideas are put together. (Berkeley, 1734).

Many philosophers take Berkeley's argument to amount to a knock-down refutation of the traditional theory -- first articulated by Aristotle (*De Interpretatione* 16a; *De Anima* 420b), and reiterated by Locke (1700) -- that images (ideas) are the primary vehicles of thought and that they ground linguistic meaning. If mental images can only, intrinsically represent particulars (as Berkeley, relying on the empiricist view of the nature of imagery as consisting of copies or fading echoes of sensory impressions, argued) then they are surely inadequate for grounding the meanings of the general, categorical terms that are fundamental to thought. However, Titchener, on introspective grounds, flatly rejected Berkeley's claim:

> But I can quite well get . . . the triangle that is no triangle at all and all triangles at one and the same time. It is a flashy thing, come and gone from moment to moment: it hints two or three red angles, with the red lines deepening into black, seen on a dark green ground. It is not there long enough to say whether the angles join to form the complete figure, or even whether all three of the necessary angles are given. Nevertheless, it means triangle; it is Locke's general idea of a triangle; (Titchener, 1909).

Of course, Titchener was well aware that the image described here was thoroughly idiosyncratic. However, he did want to claim that such images (in virtue not so much of their individual, intrinsic characteristics, but of their place in a whole associative network of imagery) do carry meaning, and are thus fitted to be the vehicles of thought. He also described examples of his own visualizations of abstract concepts (such as the concept of *meaning* itself: "the blue-grey tip of a kind of scoop … digging into a dark mass of what appears to be plastic material") and even claimed to experience imaginal meanings of connectives such as *but* (Titchener, 1909). Titchener plainly held that (together with actual sensation) mental content *is* mental imagery.

3.3 The Perky Experiment

Titchener's theories, and, to a very large extent, the introspection based experimental methods he used to test and refine them, have long since fallen into disrepute. (By contrast, Wundt's reputation has seen a considerable revival in recent decades (e.g. Blumenthal, 1975; Bringman & Tweney, 1980; Fancher, 1996).) However, one series of experiments carried out in Titchener's laboratory, by his student C.W. Perky (1910), has achieved something of a classic status in the literature on imagery. Perky asked her subjects to fixate a point on a screen in front of them and to visualize various objects there, such as a tomato, a book, a leaf, a banana, an orange, or a lemon. As the subjects did this, and unbeknownst to them, a faint patch of color, of an appropriate size and shape, and just above the normal threshold of visibility, was back projected (in soft focus) onto the screen. Apart from on a couple of occasions when the projection apparatus was mishandled, none of Perky's subjects (ranging from a ten year old child to her colleagues, the trained and experienced introspectors of Titchener's laboratory) ever realized that they were experiencing real percepts; they took what they "saw" on the screen to be entirely the products of their imagination. In fact, however, the projections did influence their experiences: some subjects expressed surprise at finding themselves imagining a banana "upright" rather than the horizontally

oriented one they had been trying for; one was surprised to wind up imagining an elm leaf after trying for a maple. On the other hand, purely imaginary details were also reported: One subject could "see" the veins of the leaf; another claimed that the title on the imagined book was readable.

Perky's results have been read as evidence that imagery may be systematically confused with genuine visual experience, that images and percepts, as Hume believed, differ subjectively in, at most, only their degree of "vivacity" or vividness. However we should note that the projected color patches were clearly visible as such to people who were not under instructions to form an image (Perky, 1910). Furthermore, Segal (1971b) reports that when she initially tried to replicate the "Perky effect" with "the suspicious, pragmatic students who populated our campuses in the late 1950s and early 1960s," they quickly saw through the deception. Eventually, she achieved better replications by taking steps to induce a state of relaxation in her subjects (Segal & Nathan, 1964). Several subjects, for example, asked to imagine a New York skyline whilst a faint image of a tomato was projected on the screen, reported imagining New York at sunset (Segal, 1972). Nevertheless, Segal concludes, from her extensive experimental studies over many years, that the *Perky effect* arises not so much from the indistinguishability of mental images and (faint) percepts, as from the fact that the effort to form an image, under certain circumstances, interferes with the normal course of perception and raises perceptual detection thresholds (Segal, 1971b; Segal & Fusella, 1971).

3.4 The *Imageless Thought* Controversy

Perhaps Wundt's most important German student was Oswald Külpe, who had for several years served as Wundt's assistant professor, but eventually left to set up his own laboratory in the philosophy department of Würzburg University. He and his students there developed a direct challenge to the prevalent imagery theory of thought. Under the influence of both Machian positivism and, later, the act psychology of Brentano and the phenomenology of Husserl, Külpe, like Titchener (whom he had helped train), rejected what he saw as Wundt's unnecessarily strict methodological restrictions on the scope of empirical science, and encouraged his students to extend the scope of the introspective "experimental" method to the study of the "higher" processes of thought and reasoning (Danziger, 1979, 1980; Ash, 1998). In 1901, two of these students, Mayer and Orth, performed a word association experiment in which subjects were asked to report everything that had passed through their mind between hearing the stimulus word and giving the response. Note that it was normal practice, in this era of psychology, for experimental subjects, or "observers" as they were often called, to be drawn from among fellow researchers within the same laboratory, often including the supervising professor. Present day psychologists would, with good reason, suspect such subjects of being liable to produce results strongly biased by theoretical preconceptions (Orne, 1962; Intons-Peterson, 1983). Great pains are usually taken, today, to ensure that subjects in psychological experiments have no idea what hypothesis the experiment is supposed to be testing. In 1901 however, it was thought that experienced and knowledgeable "observers" were more likely to produce consistent and meaningful results than the psychologically untrained. In the case of the Meyer and Orth experiment, two amongst the four subjects were Meyer and Orth themselves. Nevertheless, they professed to be surprised by some of their findings. In particular:

> The subjects frequently reported that they experienced certain events of consciousness which they could quite clearly designate neither as definite images nor yet as volitions. For example, the subject Meyer made the observation that, in reference to the stimulus word "metre" a peculiar event of consciousness intervened which could not be characterized more exactly, and which was succeeded by the spoken response "trochee".
> (Meyer & Orth, as quoted and translated by Humphrey, 1951)

The jargon term *bewusstseinslagen* ("states of consciousness" -- Humphrey, 1951) was coined to designate these indescribable non-sensorial states, and they soon began to turn up in more and more profusion in the introspective reports generated in the Würzburg laboratory, taking on an increasing theoretical significance as time went by. In 1905 another Würzburg researcher, Ach, also introduced the largely overlapping, but more explicitly intentionalistic concept of *bewusstheit* or "awareness", an unanalysable "impalpably given ’knowing’" (Ach, quoted and translated by Humphrey, 1951), and by 1907, Karl Bühler, perhaps the most radical of Külpe's students, was simply referring to *gedanken* ("thoughts"). Bühler's experiments might, for example, involve giving a subject (often professor Külpe himself) a somewhat gnomic sentence to interpret (e.g. "Thinking is so extraordinarily difficult that many prefer to judge.") and then collecting introspective reports of the conscious, but allegedly non-imaginal, *gedanken* that had occurred between the hearing of the sentence and the giving of the interpretation. Although the Würzburg school never denied that imagery does occur, by this time the greater part of the conscious contents of minds examined in Würzburg seemed to be non-imaginal.

Unsurprisingly, Wundt, and others, refused to accept these new methods and conclusions, and a heated debate, the so called *imageless thought* controversy, ensued. Though Wundt was surely skeptical of the existence of imageless thoughts, his primary criticisms were methodological. He was very much concerned with the fact that the experiments were necessarily constructed so that the introspective reports were given *after* the completion of the experimental task (word association, sentence interpretation, or whatever). The Würzburg research thus involved discursive recollection (or was it reconstruction?) of conscious contents that were no longer present to the mind. Such experiments, Wundt argued, were open invitations to suggestion, and, indeed, were

> not experiments at all in the sense of scientific methodology: they are counterfeit experiments that seem methodical simply because they are ordinarily performed in a psychological laboratory and involve the coöperation of two persons, who purport to be experimenter and observer. In reality, they are as unmethodical as possible; they possess none of the special features by which we distinguish the introspections of experimental psychology from the casual introspections of everyday life. (Wundt, quoted and translated by Titchener, 1909. Original German, 1907.)

Titchener also strongly objected to the *imageless thought* demonstrations, but for different reasons. He did not object to the aims or the introspective methodology of the Würzburg school, but to their purported results, and, for him, the experiments were not so much misconceived as incompetently executed: In particular, he felt, the "observers" (experimental subjects) in Würzburg had been inadequately trained in the art of introspection. According to Titchener, the main pitfall of introspection

was what he called the "stimulus error," the strong tendency to confound the conscious experience itself with whatever it might represent: Thus, to report, when looking at a rectangular table top, that one experiences a rectangle, would be to commit the stimulus error: The "real" conscious content would (on Titchener's view) have the trapezoidal shape that the table top projects upon the retina. For Titchener, the intentionality generally ascribed to *imageless thoughts* only showed that the Würzburg introspectors were systematically committing the stimulus error: They were not reporting the intrinsic nature of their conscious contents, but what those contents signified. Titchener suggested that the purported *bewusstseinslagen* etc. were, in fact, faint and fleeting kinaesthetic sensations, feelings of muscular tension and the like (Tweney, 1987). In his laboratory, experiments quite similar to those done in Würzburg, but carried out using introspective "observers" well trained in avoiding the stimulus error (Titchener himself, or his own graduate students), produced no reports of imageless thoughts. Instead, they found the fleeting imagery or the subtle bodily sensations that Professor Titchener's theory predicted (Titchener, 1909; Humphrey, 1951).

This work of Titchener's (like other responses to the *imageless thought* controversy from America, Britain, and elsewhere) had relatively little impact in Germany, which, with some justification at that time, still regarded itself as very much preeminent in psychological science. Nevertheless, on both sides of the Atlantic the controversy was recognized as touching on deep foundational issues in the science of mind. Although largely forgotten today, it seems to have had a lasting impact on the development not only of psychology, but (especially in the German speaking world, where the fields were more closely intellectually and institutionally entwined) philosophy as well. The Würzburg school's claims, despite their shaky basis, undoubtedly contributed to a sense that imagery could not be so psychologically important as had traditionally been assumed, and that an alternative way of thinking about cognitive content was needed. Many psychologists and philosophers of this era came, partly for this reason, to feel that thought should be understood in terms of language *per se*, and that it was a serious mistake ever to have thought that the representational power of language derives from that of some more fundamental form of representation, such as mental imagery. Bloor (1983) goes so far as to suggest (though without citing any evidence) that the work of the later Wittgenstein largely grew out of the reaction to the *imageless thought* affair. Bloor may be overstating the case, but certainly a leading Würzburg alumnus, Karl Bühler, was established in Vienna during the inter-war years, and Wittgenstein is known to have met him there, and seems to have reacted strongly to his views (Toulmin, 1969; Bartley, 1973). Bühler also taught, and deeply influenced, the young Karl Popper (Popper, 1976), and undoubtedly his views would also have been quite familiar to the Vienna Circle positivists.

But the *imageless thought* controversy was never satisfactorily resolved, at least in the terms in which it was originally posed. Although the Würzburg school has been praised for drawing psychological attention to the intentionality of mental contents, and for the introduction of once important concepts such as "mental set" into psychology, it would certainly be grossly misleading to suggest that their work provides evidence for the existence of non-sensorial conscious mental contents (i.e. imageless thoughts) that comes anywhere close to meeting contemporary scientific standards. Indeed, the fact that Külpe's and Titchener's laboratories each produced results that fitted their directors' contrasting preconceptions did not go unnoticed by their contemporaries. The unresolvable debate contributed significantly to a growing sense of intellectual crisis within psychology, leading to a deep loss of confidence (persisting to

the present) in the scientific value of introspection. It also led to a precipitous decline in scientific interest in imagery. On the one hand its importance in the cognitive economy was now subject to doubt; on the other hand it had come to seem that it was very difficult, if not impossible, to study it experimentally and objectively.

3.5 European Responses: Jaensch, Freud, and Gestalt Psychology

In Germany, some psychologists responded to this crisis by turning away from the experimental study of "cognitive" questions about the workings of the mind in general, and moved instead toward an understanding of their subject as concerned with interpretive studies of persons, or the differences between them. They, generally, became more interested in their subjects' dispositions, values, motives, etc. than in either their imagery (unless, perhaps, its contents were interestingly idiosyncratic) or their *bewusstseinslagen* (if any such existed) (Danziger, 1980).

An exception is the work of Jaensch (1930) on *eidetic* imagery (i.e. visual imagery that is experienced as before the eyes rather than "in the head," and that is unusually vivid and stable -- most evidence for the existence of eidetic imagery comes from studies of children, and it seems to be rare in adults (Haber, 1979)). However, although this work has not been without influence, and is not necessarily entirely devoid of scientific value, it is deeply tainted by Jaensch's enthusiasm for the Nazi racist ideology that was then taking hold in Germany. Eidetic imagery, he claims (on meager evidence), is characteristic of the less developed minds of not only children, but also members of "southern," "sun adapted" (i.e. darker skinned) races. (Jaensch later won notoriety for performing an experiment designed to show that "northern" chickens are racially superior -- as evidenced by more careful and intelligent pecking -- to "southern" ones (Ash, 1998).)

However, the idea that thought processes that rely upon visual imagery (as opposed to verbal thought) are characteristic of minds that are somehow defective or inferior is not confined to Nazi thinkers such as Jaensch in this era. Sigmund Freud (a Jew, who had to flee his native Austria to escape the Nazis) seems implicitly to have regarded visual images reported by his patients as part and parcel of their neuroses, as something to be exorcized and replaced by verbally mediated, "rational" insights (Esrock, 1994 ch. 3). This may well be related to the sensibility that Jay (1993) finds to be pervasive in 20th century French intellectual life, wherein visually based thought and experience is actively disvalued in comparison to other modes of sense experience, and verbally mediated thinking. Arguably, signs of a similar attitude are evident some decades earlier in England, in the responses Francis Galton received to his pioneering questionnaire about mental imagery vividness. Unlike the regular folk he questioned, many of the scientists and other intellectuals amongst Galton's respondents were distinctly unwilling to admit to ever experiencing mental imagery (Galton, 1880, 1883), a finding that more recent research has failed to reproduce (Roe, 1951; and see Ferguson, 1977, 1992; Shepard, 1978a,b; Deutsch, 1981; Miller, 1984). It is hard to say how widespread such attitudes were, or how they originated (or why they now seem to have faded), but they may well have contributed to the sharp decline in intellectual interest in imagery, apparent not only in psychology but also philosophy and literary studies, that is very apparent in the early decades of the 20th century (Esrock, 1994), and which, among philosophers and literary critics at any

rate, has only quite recently shown signs of reversal (e.g. Rollins, 1989; Ellis, 1995; Scarry, 1999).

Many other German psychologists, in the wake of the *imageless thought* controversy, continued to adhere to the Wundtian ambition of developing an experimental science of the mind, and returned to something like the sort of methodological caution in the use of introspective reports that Wundt himself had advocated, often insisting on the direct corroboration of introspective evidence by observable effects on behavior (Danziger, 1980). This usually meant that, as with Wundt himself, although their experimentally based psychology did not explicitly repudiate the essential role traditionally assigned to imagery in thought and memory, in practice it had rather little to say about it. (Plausible behavioral correlates of imagery processes were not established until the rise of the cognitive psychology movement.)

Perhaps the most influential movement arising from this strand of German psychological thought was Gestalt Psychology. It was also perhaps the last German bred movement to make a major impact in the United States, where it became a sort of "official opposition" to the indigenous and dominant Behaviorism. This was facilitated by the fact that, under the pressure of the rising tide of German Naziism, a significant number of Gestalt Psychology's adherents -- including the acknowledged leaders, Max Wertheimer, Wolfgang Köhler, and Kurt Koffka -- emigrated to America during the 1920s and 30s (Ash, 1998). Gestalt Theory attempted to explain "higher" thought processes in terms of a sort of hypothetical neuroscience (*field theory*) rather than in terms of the vicissitudes of introspected thought contents (Thomas, 1987; Ash, 1998). Although the Gestalt psychologists were much concerned with the experimental investigation of subjective experience (from whence they sought most of the evidential support for their views), in practice this research focused almost entirely on *perceptual* experience. The typical Gestaltist experiment sought immediate, unreflective descriptions of the appearance of a carefully constructed stimulus (frequently complex and illusional), and preferentially used subjects naïve to the theoretical views and concerns of the experimenter. This was something very unlike the deliberate "looking within" practiced by the psychologically sophisticated, trained introspectors of Titchener's or Külpe's laboratories. In certain respects Gestalt psychology foreshadowed, and, indeed, importantly influenced, the cognitivist movement of recent decades (Gardner, 1987). Nevertheless, it had little directly to say about the nature or function of imagery.

3.6 The American Response: Behaviorist Iconophobia (and Motor Theories of Imagery)

Where the Gestalt Psychologists, for the most part, ignored the concept of imagery, the Behaviorist movement, which came to dominate American (and, eventually, international) scientific psychology for almost half a century, actively attacked it. To borrow a coinage from Dennett (1978), Behaviorist psychology was thoroughly *iconophobic*. Although the rapid rise of Behaviorism in the United States in the early years of the 20th century certainly had multiple causes, social and institutional as well as intellectual (O'Donnell, 1985), the *imageless thought* controversy, and the questions it raised about introspection as a viable scientific methodology, was certainly prominent amongst the intellectual causes. In the famous "manifesto" by which John B. Watson publicly launched Behaviorism as a self-conscious

movement, the controversy over imageless thoughts is cited as the prime example (indeed, the only really explicit example) of the malaise of psychological methodology, for which Behaviorism would be the cure (Watson, 1913a). In a lengthy footnote to this paper, and in a follow-up article, Watson (1913b) cast doubt on the very existence of mental imagery, a position he was to state more forcefully in later work, where he stigmatized the concept (together with all other remotely mentalistic concepts) as a thoroughly unscientific, "medieval" notion, inextricably bound up with religious belief in an immortal soul, and, as such, barely one step away from "old wives tales" and the superstitions of "savagery" (Watson, 1930). He described personal reports of such things as memory images of one's childhood home as "sheer bunk," nothing more than the sentimental "dramatizing" of verbally mediated memories (i.e. conditioned tendencies to *say* certain things, either out loud or sub-vocally) (Watson, 1928).

Not all American psychologists, even overt Behaviorists, were quite as vehement as Watson in their denunciation of mentality in general, or imagery in particular, but his views certainly resonated with many. The publication of Watson's manifesto (1913a) had, in fact, been preceded by several less radical critiques of introspective methodology from other American psychologists (Danziger, 1980). Particularly relevant here is Knight Dunlap's "The Case Against Introspection" (1912), because Dunlap, who was a junior colleague of Watson in the Johns Hopkins University Psychology Department, seems to have played a crucial if inadvertent role in the formation of Watson's attitude towards imagery, and, thereby, in the crystallization of Behaviorism itself (Cohen, 1979; Thomas, 1989).

During his early days at Johns Hopkins (where he arrived in 1908) Watson, by his own account, believed that "centrally aroused visual sensations [i.e., images] were as clear as those peripherally aroused" (Watson, 1913a), and when Dunlap told Watson of his skepticism concerning what he (Dunlap) called "the old doctrine of images" Watson initially demurred, insisting that he himself made important use of visual imagery, for example in the process of designing experimental apparatus (Dunlap, 1932; cf. Watson, 1936).

However, by this time Watson already seems to have been ambitious to approach human psychology using the methodology that he had already successfully developed for the study of animal behavior (Watson, 1924, 1936). By 1910, and perhaps before, the only real factor preventing Watson from conceiving of the study of behavior as embracing the whole of psychology seems to have been "the problem of the higher thought processes" (Burnham, 1968): Thought was supposed to be carried on primarily in imagery, and imagery was not behavior (see Watson, 1913b). Dunlap's objections to the "old doctrine" that held visual imagery to consist in "centrally aroused visual sensations" seems to have played a crucial role in emboldening Watson to deny the existence of imagery altogether, thus enabling him to present the study of behavior as a fully sufficient methodology for psychology (Watson, 1924; Thomas, 1989).

However, Dunlap never became a Behaviorist himself (Dunlap, 1932), and when his actual views about imagery are examined (Dunlap, 1914) it becomes apparent that he did not intend to deny that people have experiences that are, in a significant sense, quasi-perceptual. Although he described himself as an "iconoclast" (1932), and held that "the image, as a copy or reproduction of sensation . . . does not exist," (Dunlap,1914), Dunlap also asserted that Watson went much too far in rejecting "imagination" as well as

"images" (Dunlap, 1932), and he continued to hold that we are in need of an account of the nature of "ideas". Something, something mental and, indeed, quasi-perceptual, is needed to fill the functional role that images played in the traditional psychology of thinking. It is clear that he (unlike Watson) did not deny the existence of imagery in the sense in which it was defined at the beginning of this article (i.e. quasi-perceptual experience). Dunlap's theory would seem to be best understood as a pioneering (though perhaps, ultimately, unconvincing) attempt to explain both the experience of imagery, and the functional role that it plays in thinking, in a way that avoids postulating the presence of pictures in the head, or inner copies of former sense impressions.

According to Dunlap, *ideas* are actually complexes of muscular sensations, caused by outwardly imperceptible movements, or, at least, tensings, of the muscles, particularly (though not exclusively) the muscles associated with the sense organs themselves, such as those that move the eyes. Particular patterns of muscular response, Dunlap holds, occur during the perception of particular types of objects or events, and may be aroused not only in the course of the actual perception of a relevant object, but also through associative links with the sensations produced by other muscular response patterns appropriate to other sorts of objects or events. These latter patterns may have arisen in actual perception, or may themselves have been aroused associatively in a similar way. Thus, associative trains of thought can be sustained. When the muscular response is aroused associatively, rather than by the actual perceptual presence of the relevant object, we experience the idea, or image, of the object. Visual imagery consists not of copies or echoes of visual sensations, but rather of actual current sensations in the muscles involved in the process of seeing something.

> There is indeed a present content essentially connected with imagination or thought; but this present content is in each case a muscle sensation, or a complex of muscle sensations. We are therefore, in investigating images, dealing not with copies, or pale ghosts, of former sensations but with actual present sensations. (Dunlap, 1914)

These muscle sensations are, explicitly, not to be confused with the impalpable *imageless thoughts* of Würzburg, rather, "This sensation is the true *image*" (Dunlap, 1914, emphasis in original). (For a more extensive account of Dunlap's theory of imagery, and its influence on Watson, see Thomas (1989).)

Dunlap's theory of imagery/ideas was publicly presented only in one brief and rather obscurely published article (Dunlap, 1914) and (apart from its unintended and covert influence on Watson), it seems to have attracted very little interest from his contemporaries. The theory probably owes much to the influence of Hugo Münsterberg, whose "motor theory" of the mind had a considerable vogue amongst American psychologists at the time, but which was soon to be eclipsed by the rise of Behaviorism (Scheerer, 1984). Münsterberg was a German, a former student of Wundt, who had been hired to teach psychology at Harvard when William James moved on, and Dunlap had studied under him before coming to Johns Hopkins (Dunlap, 1932). An earlier "motor" theory of imagery can also be found in the work of the French psychologist Theodule Ribot (1890, 1900) (predating Münsterberg's influence), but the most developed version was surely that of Margaret Floy Washburn, a former student of Titchener. Washburn (unlike Dunlap) is quite open in acknowledging her intellectual debt to Münsterberg (Washburn, 1932), and her book *Movement and Mental Imagery* (Washburn, 1916) goes into considerable, if speculative,

physiological as well as psychological detail. However, by the time this was published Behaviorist iconophobia was already taking firm hold, and Washburn's version of the motor theory of imagery seems to have failed to attract any more adherents than Dunlap's had.

Infamously, during the period of Behaviorist dominance, up until about 1960, mental imagery received minimal attention from scientific psychologists. According to Paivio (1971), the 1920s and 1930s were "the most arid period" for imagery research, but Kessel (1972) reports that even through the 1940s and 1950s a scant five references to imagery are to be found in *Psychological Abstracts*. Admittedly some interest in the psychology of imagery continued outside the United States in this era. In Britain, for example, psychologists such as Pear (1925, 1927), Bartlett (1927, 1932), and, latterly, McKellar (1957) kept an interest in the topic alive. However, this work did not have much contemporaneous impact in the United States, which by the 1930s had already achieved its dominant superpower status in psychology, if not yet in other domains. A general revival of interest in imagery did not get under way in America before the1960s. By that time the Behaviorist consensus was beginning to break down (as can be seen in the work of Mowrer (1960), who tried to patch-up Behaviorist learning theory by introducing the alien concept of imagery into it), and new and striking empirical findings about imagery emerged to play a significant role in the cognitive revolution.

Bibliography

[Particularly seminal, influential, or useful contributions to the imagery literature are marked with a •. Items that are cited in this article but that say little or nothing directly about mental imagery are marked with a •. Some items are annotated.]

- Ahsen A. (1984). ISM: The Triple Code Model for Imagery and Psychophysiology. *Journal of Mental Imagery* (8) 15-42.
- Anderson, J.R. (1978). Arguments Concerning Representations for Mental Imagery. *Psychological Review* (85) 249-77.
  [Argues that the "analog vs.propositional" (picture vs. description) question is ill posed.]
- Anderson, J.R. (1979). Further Arguments Concerning Representations for Mental Imagery: A Response to Hayes-Roth and Pylyshyn. *Psychological Review* (86) 395-406.
- Ash, M.G. (1998). *Gestalt Psychology in German Culture, 1890-1967*. Cambridge: Cambridge University Press.
- Audi, R. (1978). The Ontological Status of Mental Images. *Inquiry* (21) 348-361.
- Aveling E. (1927). The Relevance of Visual Imagery to the Process of Thinking 2. *British Journal of Psychology* (18) 15-22.
  [A companion piece to Pear (1927) and Bartlett (1927).]
- Baars, B.J. (Ed.) (1996). Special issue on mental imagery of *Consciousness and Cognition* (5-iii).

- Barsalou, L.W. (1999). Perceptual Symbol Systems (with commentaries and author's reply). *Behavioral and Brain Sciences* (22) 577-660. [Available online]

  [Purportedly not directly about imagery, but deals with the closely adjacent topic of mental representations that are inherently perceptual in character, and argues that they are adequate to account for cognition, and explanatorily superior to "amodal" conceptions of representation (such as *mentalese*).]

- Bartlett, F.C. (1927). The Relevance of Visual Imagery to the Process of Thinking. *British Journal of Psychology* (18) 23-29.
  [A companion piece to Pear (1927) and Aveling (1927).]

- Bartlett, F.C. (1932). *Remembering.* Cambridge: Cambridge University Press.

- Bartley, W.W. (1973). *Wittgenstein*. New York: Lippincott.

- Bartolomeo, P., Bachoud-Lévi, A-C., De Gelder, B. Denes, G., G., Dalla Barba, G., Brugieres, P. & Degos, J.-P. (1998). Multiple-Domain Dissociation between Impaired Visual Perception and Preserved Mental Imagery in a Patient with Bilateral Extrastriate Lesions. *Neuropsychologia* (36) 239-249.
  [Suggests that imagery does *not* depend on activity in the early visual areas of the brain. For an opposing view see Kosslyn, Alpert *et al*. (1993), Kosslyn, Thompson *et al*. (1995), Kosslyn, Pascual-Leone *et al*. (1999).]

- Bartolomeo, P., Bachoud-Lévi, A-C., & Denes, G. (1997). Preserved Imagery for Colours in a Patient with Cerebral Achromatopsia. *Cortex* (33) 369-378.
  [See note on previous item.]

- Bartolomeo, P., D'Erme, P., & Gainotti, G. (1994). The Relation between Visuospatial Neglect and Representational Neglect. *Neurology* (44) 1710-1714.
  [See Bisiach & Luzzatti (1978).]

- Basso, A., Bisiach, E., & Luzzatti, C. (1980). Loss of Mental Imagery: A Case Study. *Neuropsychologia* (18) 435-442.

- Baylor, G.W. (1972). *A Treatise on the Mind's Eye: An Empirical Investigation of Visual Mental Imagery*. Unpublished Ph.D. thesis, Carnegie-Mellon University, Pittsburgh, PA. (University Microfilms 72-12, 699.)
  [The first serious attempt to simulate imagery computationally. The major inspiration for the description theory of Pylyshyn (1973).]

- Baylor, G.W. (1973). Modelling the Mind's Eye. In A. Elithorn & D. Jones (Eds.), *Artificial and Human Thinking*. Amsterdam: Elsevier.
  [A brief sketch of the model detailed in Baylor (1972).]

- Berkeley, G. (1734). *A Treatise Concerning the Principles of Human Knowledge.* (2nd edn.) In M.R. Ayers (Ed.). *George Berkeley: Philosophical Works Including the Works on Vision.* London: Dent, 1975.
  [The *ideas* of Berkeley's philosophy are, to all intents and purposes, mental images.]

- Bexton, W.H., Heron, W., & Scott, T.H. (1954). Effects of Decreased Variation in the Sensory Environment. *Canadian Journal of Psychology* (8) 70-76.
  [Sensory deprivation discovered to give rise to spontaneous and bizarre imagery.]
- Bisiach, E. & Berti, A. (1990). Waking Images and Neural Activity. In R.G. Kunzendorf & A.A. Sheikh (Eds.) *The Psychophysiology of Mental Imagery: Theory, Research and Application*. Amityville, NY: Baywood.
- Bisiach, E. & Luzzatti, C. (1978). Unilateral Neglect of Representational Space. *Cortex* (14) 129-133.
  [Brain damaged patients who ignore things to their left also ignore the left side in their imagery. Also see the next item, and: Bartolomeo, D'Erme, & Gainotti, (1994), Coslett (1997).]
- Bisiach, E., Luzzatti, C., & Perani, D. (1979). Unilateral Neglect, Representational Schema and Consciousness. *Brain* (102) 609-618.
- Blachowicz, J. (1997). Analog Representation Beyond Mental Imagery. *Journal of Philosophy* (94) 55-84.
- Block, N. (Ed.) (1981a). *Imagery*. Cambridge, MA: MIT Press.
  [Widely read collection of pieces concerned with the *analog/propositional* debate..]
- Block, N. (Ed.) (1981b). *Readings in Philosophy of Psychology, Vol. 2*. London: Methuen.
  [Section on imagery adds to and complements the above.]
- Block, N. (1983a). Mental Pictures and Cognitive Science. *Philosophical Review* (92) 499-539.
- Block, N. (1983b). The Photographic Fallacy and the Debate about Mental Imagery. *Noûs* (17) 651-661.
- Bloor, D. (1983). *Wittgenstein: A Social Theory of Knowledge*. London: Macmillan.
- Blumenthal, A.C. (1975). A Reappraisal of Wilhelm Wundt. *American Psychologist* (30) 1081-1088.
- Blumenthal, H.J. (1977-8). Neoplatonic Interpretations of Aristotle on *Phantasia*. *Review of Metaphysics* (31) 242-257.
- Boring, E.G. (1950). *A History of Experimental Psychology* (2nd edn.). New York: Appleton.
- Bower, K.J. (1984). Imagery: From Hume to Cognitive Science. *Canadian Journal of Philosophy* (14) 217-234.
- Brandt, S.A. & Stark, L.W. (1997). Spontaneous Eye Movements During Visual Imagery Reflect the Content of the Visual Scene. *Journal of Cognitive Neuroscience* (9) 27-38.
  [Some direct experimental support of a *Perceptual Activity* theory of imagery. Cf. Hong *et al*. (1997).]

- Brann, E.T.H. (1991). *The World of the Imagination: Sum and Substance*. Savage, MD: Rowman & Littlefield.
  [An ambitious philosophical history of conceptions of imagination and imagery, from ancient to contemporary times.]
- Bringman, W.G. & Tweney, R.D. (Eds.) (1980). *Wundt Studies.* Toronto: Hogrefe.
- Brodie, A. (1986-7). Medieval Notions and the Theory of Ideas. *Proceedings of the Aristotelian Society* (86) 153-167.
- Brooks, L.R. (1967). The Suppression of Visualization by Reading. *Quarterly Journal of Experimental Psychology* (19) 287-299.
- Brooks, L. R. (1968). Spatial and Verbal Components of the Act of Recall. *Canadian Journal of Psychology* (22) 349-368.
  [Together with the above, this demonstrates selective interference between spatial perception and spatial (including visual) imagery. See Hampson & Duffy (1984) for a replication in congenitally blind subjects.]
- Bugelski, B.R. (1970). Words and Things and Images. *American Psychologist* (25) 1002-10012.
  [On imagery effects in verbal learning experiments.]
- Bugelski, B.R. (1977). *Mnemonics*. In *International Encyclopedia of Psychiatry, Psychology, Psychoanalysis, and Neurology, Vol. 7*. New York: Van Nostrand Reinhold.
- Bugelski, B.R. (1984). *Imagery*. In R.J. Corsini (Ed.). *Encyclopedia of Psychology, Vol. 2*. New York: Wiley.
- Burnham, J.C. (1968). On the Origins of Behaviorism. *Journal of the History of the Behavioral Sciences* (4) 143-151.
- Candlish, S. (1975). Mental Images and Pictorial Properties. *Mind* (84) 260-262.
  [A critique of Hannay's (1971) defense of pictorialism.]
- Candlish, S. (1976). The Incompatibility of Perception and Imagery: A Contemporary Orthodoxy. *American Philosophical Quarterly* (13) 63-68.
  [Stewart Candlish informs me that the title of this article was misprinted in the published version. The title given above is the one he intended.]
- Candlish, S. (2001). Mental Imagery. In S. Schroeder (Ed.). *Wittgenstein and Contemporary Philosophy of Mind*. London: Palgrave.
  [Discusses Wittgenstein's views on imagery, and their influence.]
- Carpenter, P.A. & Eisenberg, P. (1978). Mental Rotation and the Frame of Reference in Blind and Sighted Individuals. *Perception and Psychophysics* (23) 117-124.
  [Mental rotation effect demonstrated in congenitally blind subjects.]

- Casey, E.S. (1971). Imagination: Imagining and the Image. *Philosophy and Phenomenological Research* (31) 475-90.
- Casey, E.S. (1976). *Imagining: A Phenomenological Study*. Bloomington, IN: Indiana University Press.
- Casey, E.S. (1977-8). Imagining and Remembering. *Review of Metaphysics* (31) 187-209.
- Chambers, D. (1993). Images are Both Depictive and Descriptive. In B. Roskos-Ewoldsen, M.J. Intons-Peterson & R.E. Anderson (Eds.). *Imagery, Creativity and Discovery: A Cognitive Perspective.* Amsterdam: Elsevier. (pp. 77-97)
  [If neither theory fits the facts, why not choose both.]
- Chambers, D. & Reisberg, D. (1985). Can Mental Images be Ambiguous? *Journal of Experimental Psychology: Human Perception and Performance* (11) 317-328.
  [A very striking experiment; but see Peterson *et al.* (1992), Rollins (1994), Cornoldi *et al*, (1996), Slezak (1991, 1995), and other listed works by Chambers and/or Reisberg for related (and often conflicting) experimental results, and competing interpretations.]
- Chambers, D. & Reisberg, D. (1992). What an Image Depicts Depends on What an Image Means: An Image of a Duck Does Not Include a Rabbit's Nose. *Cognitive Psychology* (24) 145-174.
- Cohen, D. (1979). *J. B. Watson -- the Founder of Behaviorism: A Biography*. London: Routledge & Kegan Paul.
- Cohen, J. (1996). The Imagery Debate: A Critical Assessment. *Journal of Philosophical Research* (21) 149-182.
- Cornoldi, C., Logie, R.H., Brandimonte, M.A., Kaufmann, G., & Reisberg, D. (1996). *Stretching the Imagination: Representation and Transformation in Mental Imagery*. Oxford: Oxford University Press.
  [See note at Chambers & Reisberg (1985).]
- Coslett, H.B. (1997). Neglect in Vision and Visual Imagery: A Double Dissociation. *Brain* (120) 1163-1171.
  [See note at Bisiach & Luzzatti (1978).]
- Crammond, D.J. (1997). Motor Imagery: Never in Your Wildest Dreams. *Trends in Neuroscience* (20-2) 54-57.
- Currie, G. (1995). Visual Imagery as the Simulation of Vision. *Mind and Language* (10) 25-44.
- Currie, G. & Ravenscroft, I. (1997). Mental Simulation and Motor Imagery. *Philosophy of Science* (64) 161-180.
- Danto, A.C. (1958). Concerning Mental Pictures. *Journal of Philosophy* (55) 12-20.

- Danziger, K. (1979). The Positivist Repudiation of Wundt. *Journal of the History of the Behavioral Sciences* (15) 205-230.
- Danziger, K. (1980). The History of Introspection Reconsidered. *Journal of the History of the Behavioral Sciences* (16) 241-262.
- Daston, L. (1998). Fear and loathing of the imagination in science. *Dædalus* (127-1) 73-95.
- Denis, M., Engelkamp, J., & Richardson, J.T.E. (Eds.) (1988). *Cognitive and Neuropsychological Approaches to Mental Imagery*. Dordrecht, Netherlands: Martinus Nijhoff.
- Denis, M. & Carfantan, M. (1985). People's Knowledge About Images. *Cognition* (20) 49-60.
  [An empirical study of the folk psychology of imagery.]
- Dennett, D.C. (1969). *Content and Consciousness*. London: Routledge & Kegan Paul.
  [Argues that the inherent vagueness of images suggests that they are more like descriptions than pictures. (See, e.g. Hannay, 1971, Block, 1983b, and Tye, 1991, for counter arguments.)]
- Dennett, D.C. (1978). Two Approaches to Mental Imagery. In his *Brainstorms*. Montgomery, VT: Bradford Books.
- Dennett, D.C. (1991). *Consciousness Explained*. Boston, MA: Little, Brown.
  [Chapter 10 attempts to integrate Kosslyn's quasi-pictorial theory of imagery into Dennett's philosophical framework.]
- Deutsch, M. (1981). Imagery and Inference in Physical Research. In Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (Eds.), *On Scientific Thinking* (pp. 354-360). New York: Columbia University Press. (Extract from original work of 1959.)
- Dror, I.E., Ivey, C., & Rogus, C. (1997). Visual Mental Rotation of Possible and Impossible Objects. *Psychonomic Bulletin and Review* (4) 242-247.
- Dunlap, K. (1912). The Case Against Introspection. *Psychological Review* (19) 404-413.
- Dunlap, K. (1914). Images and Ideas. *Johns Hopkins University Circular* (3 -- March 1914) 25-41.
  [A *motor* theory of imagery. See Washburn (1916) for a related view, and Thomas (1989) for discussion.]
- Dunlap, K. (1932). Knight Dunlap. In C. Murchison (Ed.), *A History of Psychology in Autobiography* (Vol. 2, pp. 35-61). Worcester, MA: Clark University Press.
- Ellis, R.D. (1995). *Questioning Consciousness: The Interplay of Imagery, Cognition, and Emotion in the Human Brain*. Amsterdam: John Benjamins.
  [Gives an imagery based theory of thought and semantics. See Thomas (1997b) for discussion.]

- Esrock, E.J. (1994). *The Reader's Eye: Visual Imaging as Reader Response*. Baltimore, MD. Johns Hopkins University Press.
  [A historical treatment of the role of the concept of mental (as opposed to verbal) imagery in 20th century literary criticism, and a proposal, drawing on cognitive psychology research, for a mental imagery based theory of response to literature. Cf. Scarry (1999).]
- Fancher, R.E. (1996). *Pioneers of Psychology* (3rd edn.). New York: W.W. Norton.
- Farah, M.J. (1984). The Neurological Basis of Mental Imagery: A Componential Analysis. *Cognition* (18) 245-72.
- Farah, M.J. (1988). Is Visual Imagery Really Visual? Overlooked Evidence from Neuropsychology. *Psychological Review* (95) 307-317.
- Farah, M. J., Hammond, K. M., Levine, D. N., & Calvanio, R. (1988). Visual and Spatial Mental Imagery: Dissociable Systems of Representation. *Cognitive Psychology* (20) 439-462.
- Farah, M. J., Soso, M. J., & Dasheif, R. M. (1992). Visual Angle of the Mind's Eye Before and After Unilateral Occipital Lobectomy. *Journal of Experimental Psychology: Human Perception and Performance* (18) 241-246.
- Farley, A.M. (1974). *VIPS: A Visual Imagery Perception System; the Result of Protocol Analysis*. Unpublished Ph.D. thesis, Carnegie-Mellon University, Pittsburgh, PA.
  [Computer model of imagery based on the *perceptual activity* theory of Hochberg (1968).]
- Farley, A.M. (1976). A Computer Implementation of Constructive Visual Imagery and Perception. In R.A. Monty J.W. Senders (Eds.) *Eye Movements and Psychological Processes*. Hillsdale, NJ: Erlbaum.
  [A concise account of the model developed by Farley (1974).]
- Ferguson, E.S. (1977). The Mind's Eye: Nonverbal Thought in Technology. *Science* (197) 827-836.
- Ferguson, E.S. (1992). *Engineering and the Mind's Eye*. Cambridge, MA: MIT Press.
- Finke, R.A. (1980). Levels of Equivalence in Imagery and Perception. *Psychological Review* (87) 113-132.
- Finke, R.A. (1985). Theories Relating Imagery to Perception. *Psychological Bulletin* (98) 236-259.
- Finke, R.A. (1986). Mental Imagery and the Visual System. *Scientific American* (245 #iii, March) 76-83.
- Finke, R.A. (1989). *Principles of Mental Imagery*. Cambridge, MA: MIT Press.
  [Useful textbook of the experimental cognitive psychology of imagery.]

- Finke, R.A., Pinker, S., & Farah, M.J. (1989). Reinterpreting Visual Patterns in Mental Imagery. *Cognitive Science* (13) 51-78.

- Finke, R.A. & Shepard, R.N. (1986). Visual Functions of Mental Imagery. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.). *Handbook of Perception and Human Performance, Vol. 2*. New York: Wiley-Interscience.

- Finke, R.A., Ward, T.B., & Smith, S.M. (1992). *Creative Cognition: Theory, Research, and Applications*. Cambridge, MA: MIT Press.
[Gives imagery a large role in inventive thinking.]

- Flew, A. (1953). Images, Supposing and Imagining. *Philosophy* (28) 246-254.

- Fodor, J.A. (1975). *The Language of Thought*. New York: Thomas Crowell. (Paperback edition: Harvard University Press, 1980)
[Argues that imagery representations must be semantically dependent on representations that are linguistic in form.]

- Freyd, J.J. (1987). Dynamic Mental Representations. *Psychological Review* (94) 427-38.

- Furlong, E.J. (1953). Abstract Ideas and Images. *Proceedings of the Aristotelian Society* (Supplementary volume 27) 121-136.

- Furlong, E.J. (1961). *Imagination*. London: Allen & Unwin.

- Galton, F. (1880). Statistics of Mental Imagery. *Mind* (5) 301-318.
[Pioneering individual differences survey.]

- Galton, F. (1883). *Inquiries into Human Faculty and its Development*. London: Macmillan.
[Summarizes and discusses results of the above.]

- Gardner, H. (1987). *The Mind's New Science: A History of the Cognitive Revolution* (2nd edition). New York: Basic Books.
[Includes a fairly good account of the "analog-propositional" debate.]

- Georgopoulos, A.P., Lurito, J.T., Petrides, M., & Schwartz, A.B. (1989). Mental Rotation of the Neuronal Population Vector. *Science* (243) 234-236.
[A neuroscientific study of the mental rotation effect (in monkeys) which links it to motor control.]

- Giaquinto, M. (1992). Visualizing as a Means of Geometrical Discovery. *Mind and Language* (7) 382-401.

- Giaquinto, M. (1993). Visualizing in Arithmetic. *Philosophy and Phenomenological Research* (53) 385-396.

- Gibson, J.J. (1970). On the Relation Between Hallucination and Perception. *Leonardo* (3) 425-7.

- Gibson, J.J. (1974). Visualizing Conceived as Visual Apprehending Without Any Particular Point of Observation. *Leonardo* (7) 41-42.
- Glasgow, J.I. (1993). The Imagery Debate Revisited: A Computational Perspective. *Computational Intelligence* (9) 310-333.
  [Printed with numerous peer commentaries and author's reply.]
- Glasgow, J. & Papadias, D. (1992). Computational Imagery. *Cognitive Science* (16) 355-394.
- Goldenberg, G. (1989). The Ability of Patients with Brain Damage to Generate Mental Visual Images. *Brain* (112) 305-325.
- Gray, C.R. & Gummerman, K. (1975). The Enigmatic Eidetic Image: A Critical Examination of Methods, Data, and Theories. *Psychological Bulletin* (82) 383-407.
- Haber, R.N. (1979). Twenty Years of Haunting Eidetic Imagery: Where's the Ghost? *Behavioral and Brain Sciences* (2) 583-629.
  [With appended commentaries.]
- Hampson, P.J. (1979). *The Role of Imagery in Cognition*. Unpublished Ph.D. thesis, University of Lancaster, Lancaster, U.K.
- Hampson, P.J. & Duffy, C. (1984). Verbal and Spatial Interference Effects in Congenitally Blind and Sighted Subjects. *Canadian Journal of Psychology* (38) 411-20.
  [Selective interference effects (see Brooks (1967, 1968)) demonstrated between spatial perception and spatial imagery in the congenitally blind.]
- Hampson, P.J., Marks, D.F., & Richardson, J.T.E. (Eds.) (1990). *Imagery: Current Developments*. London: Routledge.
- Hampson, P.J. & Morris, P.E. (1978). Unfulfilled Expectations: A Critique of Neisser's Theory of Imagery. *Cognition* (6) 79-85.
  [A critique of Neisser's (1976) *perceptual activity* theory of imagery. See Neisser (1978) for reply.]
- Hampson, P.J. & Morris, P.E. (1979). Cyclical Processing: A Framework for Imagery Research. *Journal of Mental Imagery* (3) 11-22.
  [An attempt to synthesize the *quasi-pictorial* and *perceptual activity* theories.]
- Hannay, A. (1971). *Mental Images -- A Defence*. London: Allen & Unwin.
  [Argues for the reality of inner pictures.]
- Hannay, A. (1973). To See a Mental Image. *Mind* (82) 161-262.
- Harrison, B. (1962-3). Meaning and Mental Images. *Proceedings of the Aristotelian Society* (63) 237-250.

- Harvey, E.R. (1975). *The Inward Wits: Psychological Theory in the Middle Ages and the Renaissance*. London: Warburg Institute, University of London.
- Hayes, J.R. (1973). On the Function of Visual Imagery in Elementary Mathematics. In W.G. Chase (Ed.) *Visual Information Processing*. New York: Academic Press.
- Hayes-Roth, F. (1979). Distinguishing Theories of Mental Representation: A Critique of Anderson's ‘Arguments Concerning Mental Imagery’. *Psychological Review* (86) 376-382.
- Hebb, D.O. (1968). Concerning Imagery. *Psychological Review* (75) 466-477.
  [Outlines a version of *motor* or *perceptual activity* theory.]
- Hebb, D.O. (1969). The Mind's Eye. *Psychology Today* (2) 54-57 & 67-68.
- Hegarty, M. (1992). Mental Animation: Inferring Motion from Static Displays of Mechanical Systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (18) 1084-1102.
  [Animated mental images.]
- Heil, J. (1982). What Does the Mind's Eye Look At? *Journal of Mind and Behavior* (3) 143-149.
  [An *adverbial* account of imagery, which may be considered the philosophical counterpart (at the level of language analysis) to the *perceptual activity* theory in cognitive science. Imagery is regarded not as the having of a mental object (an image) in the mind, rather it is a type of activity, a way of thinking about some actual or possible real-world object. See Rabb (1975), Tye (1984) for other versions of *adverbial* theory.]
- Heuer, F., Fischman, D., & Reisberg, D. (1986). Why Does Vivid Imagery Hurt Colour Memory? *Canadian Journal of Psychology* (40) 161-175.
  [Individual differences study using the VVIQ questionnaire of Marks (1973). A companion piece to Reisberg, Culver, Heuer, & Fischman (1986).]
- Hilgard, E.R. (1981). Imagery and Imagination in American Psychology, *Journal of Mental Imagery* (5) 5-66.
  [Historical reflections, with appended commentaries.].
- Hinton, G. (1979). Some Demonstrations of the Effects of Structural Descriptions in Mental Imagery. *Cognitive Science* (3) 231-250.
  [Argues for the view that images are "structural descriptions". A version of the "propositional" theory defended by Pylyshyn.]
- Hochberg, J. (1968). In the Mind's Eye. In R.N. Haber (Ed.). *Contemporary Theory and Research in Visual Perception*. Holt Rinehart & Winston. New York. pp. 309-331.
  [Argues for a *perceptual activity* approach.]
- Holt, R.R. (1964). Imagery: The Return of the Ostracised. *American Psychologist* (19) 254-266.
  [Influential account of the history of imagery in scientific psychology.]

- Hong, C.C.-H., Potkin, S.G., Antrobus, J.S., Dow, B.M., Callaghan, G.M., & Gillin, J.C. (1997) REM Sleep Eye Movement Counts Correlate with Visual Imagery in Dreaming: A Pilot Study. *Psychophysiology* (34) 377-381.
  [Cf. Brandt & Stark (1997).]
- Horne, P.V. (1993). The Nature of Imagery. *Consciousness and Cognition* (2) 58-82.
  [With commentary.]
- Humphrey, G. (1951). *Thinking*. London: Methuen.
  [Contains what is probably still the best account in English of the views of the influential *imageless thought* school of German introspective psychology, including translations from primary sources.]
- Intons-Peterson, M.J. (1983). Imagery Paradigms: How Vulnerable are They to Experimenter's Expectations? *Journal of Experimental Psychology: Human Perception and Performance* (9) 394-412.
  [Shows that results of some imagery experiments may be seriously distorted by experimenters' expectations.]
- Intons-Peterson, M.J. & Roskos-Ewoldsen, B.B. (1989). Sensory Perceptual Qualities of Images. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (15) 188-199.
- Ishiguro, H. (1967). Imagination. *Proceedings of the Aristotelian Society, Supplementary Volume* (41) 37-56.
  [Images as intentional objects. (Strongly influenced by Wittgenstein and Ryle.)]
- Jaensch, E.R. (1930). *Eidetic Imagery and Typological Methods of Investigation.* (Translated from the German by O.A. Oeser.) London: Routledge & Kegan Paul.
  [A seminal study of eidetic imagery, but seriously tainted by the racist assumptions of its Nazi milieu.]
- James, W. (1890). *The Principles of Psychology*. New York: Holt. [Harvard University Press edition of 1983].
- Janssen, W. (1976). *On the Nature of Mental Imagery*. Soesterburg, Netherlands: Institute for Perception TNO.
- Jay, M. (1993). *Downcast Eyes: The Denigration of Vision in Twentieth-Century French Thought.* Berkeley, CA: University of California Press.
- Jonides, J., Kahn, R., & Rozin, P. (1975). Imagery Instructions Improve Memory in Blind Subjects. *Bulletin of the Psychonomic Society* (5) 424-6.
- Kaufmann, G. (1980). *Imagery, Language and Cognition*. Oslo, Norway: Universitetsforlaget.
- Keilkopf, C.F. (1968). The Pictures in the Head of a Man Born Blind. *Philosophy and Phenomenological Research* (28) 501-513.

- Kerr, N.H. (1983). The Role of Vision in 'Visual Imagery' Experiments: Evidence from the Congenitally Blind. *Journal of Experimental Psychology: General* (112) 265-77.
  [Many "classic" experimental effects attributed to imagery can be reproduced in blind subjects.]
- Kessel, F.S. (1972). Imagery: A Dimension of Mind Rediscovered. *British Journal of Psychology* (63) 149-62.
- Kolers, P.A. (1987). Imaging. In R.L. Gregory & O.L. Zangwill (Eds.). *The Oxford Companion to the Mind*. Oxford: Oxford University Press.
- Kolers, P.A. & Smythe, W.E. (1979). Images, Symbols, and Skills. *Canadian Journal of Psychology* (33) 158-184.
- Kosslyn, S.M. (1980). *Image and Mind*. Cambridge, MA: Harvard University Press.
  [Seminal statement and defence of the computational *Quasi-Pictorial* theory of imagery, which has become the dominant view in cognitive science.]
- Kosslyn, S.M. (1981). The Medium and the Message in Mental Imagery: A Theory. *Psychological Review* (88) 46-66.
- Kosslyn, S.M. (1987). Seeing and Imagining in the Cerebral Hemispheres: A Computational Approach. *Psychological Review* (94) 148-75.
- Kosslyn, S.M. (1983). *Ghosts in the Mind's Machine: Creating and Using Images in the Brain*. New York: Norton.
  [A popularization of the *Quasi-Pictorial* theory.]
- Kosslyn, S.M. (1994). *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: MIT Press.
  [Updates the *Quasi-Pictorial* theory with an account of how imagery might be neurologically embodied.]
- Kosslyn, S. M., Alpert, N. M., Thompson, W. L., Maljkovic, V., Weise, S. B., Chabris, C. F., Hamilton, S. E., Rauch, S. L., & Buonanno, F. S. (1993). Visual mental imagery activates topographically organized visual cortex: PET investigations. *Journal of Cognitive Neuroscience* (5) 263-287.
  [Suggests that imagery depends on activity in the early, retinotopically mapped visual areas of the brain. For contrary evidence see: Roland & Gulyàs (1994), Mellet *et al*. (1996), Bartolomeo *et al*. (1997), Bartolomeo *et al*. (1998).]
- Kosslyn, S.M. & Hatfield, G. (1984). Representation Without Symbol Systems. *Social Research* (51) 1019-1045.

- Kosslyn, S.M., Pascual-Leone, A., Felician, O., Camposana, S., Keenan, J.P., Thompson, W.L., Ganis, G., Sukel, K.E.. & Alpert, N.M. (1999). The Role of Area 17 in Visual Imagery: Convergent Evidence from PET and rTMS. *Science* (284) 167-170.
  [Suggests that imagery depends on activity in the early, retinotopically mapped visual areas of the brain. For contrary evidence see: Roland & Gulyàs (1994), Mellet *et al*. (1996), Bartolomeo *et al*. (1997), Bartolomeo *et al*. (1998).]

- Kosslyn, S.M., Pinker, S., Smith, G.E., & Shwartz, S.P. (1979). On the Demystification of Mental Imagery. *Behavioral & Brain Sciences* (2) 535-581.
  [With appended commentaries.]]

- Kosslyn, S.M. & Pomerantz, J.R. (1977). Imagery, Propositions and the Form of Internal Representations. *Cognitive Psychology* (9) 52-76.
  [A defence of *Quasi-Pictorial* theory against "propositional"/descriptional alternatives.]

- Kosslyn, S.M. & Shwartz, S.P. (1977). A Simulation of Visual Imagery. *Cognitive Science* (1) 265-295.
  [Computer model of *Quasi-Pictorial* theory.]

- Kosslyn, S.M., Sukel, K.E., & Bly, B.M. (1999). Squinting with the Mind's Eye: Effects of Stimulus Resolution on Imaginal and Perceptual Comparisons. *Memory and Cognition* (19) 276-282.

- Kosslyn, S.M., Thompson, W.L., Kim, I.J., & Alpert, N.M. (1995). Topographical Representation of Mental Images in Primary Visual Cortex. *Nature* (378) 496-498.
  [Suggests that imagery depends on activity in the early, retinotopically mapped visual areas of the brain. For contrary evidence see: Roland & Gulyàs (1994), Mellet *et al*. (1996), Bartolomeo *et al*. (1997), Bartolomeo *et al*. (1998).]

- Kreiman, G., Koch C., & Freid, G. (2000). Imagery Neurons in the Human Brain. *Nature* (408) 357-361.

- Kunzendorf, R.G., Justice, M., & Capone, D. (1997). Conscious Images as "Centrally Excited Sensations": A Developmental Study of Imaginal Influences on the ERG. *Journal of Mental Imagery* (21) 155-166.

- Kunzendorf, R.G. & Sheikh, A.A. (Eds.) (1990). *The Psychophysiology of Mental Imagery: Theory, Research and Application*. Amityville, NY: Baywood.

- Lawrie, R. (1970). The Existence of Mental Images. *Philosophical Quarterly* (20) 253-7.

- Leahey, T.H. (1981). The Mistaken Mirror: On Wundt's and Titchener's Psychologies. *Journal of the History of the Behavioral Sciences* (17) 273-282.

- Levine, D. N., Warach, J., & Farah, M. (1985). Two Visual Systems in Mental Imagery: Dissociation of "What" and "Where" in Imagery Disorder Due to Bilateral Posterior Cerebral Lesions. *Neurology* (35) 1010-1018.
- Locke, J. (1700). *An Essay Concerning Human Understanding.* [Edition of S. Pringle-Pattison (1924). Oxford: Oxford University Press.]
  [It is not entirely clear whether Locke's term *idea* can always properly be equated with *mental image*. However, he certainly sometimes seems to have imagery in mind, and *idea* was generally treated as equivalent to *mental image* by many of his most important successors, including Berkeley and Hume.]
- Logie, R.H. & Denis, M. (Eds.) (1991). *Mental Images in Human Cognition.* Amsterdam: Elsevier Science Publishers B.V.
- Long, A.A. (1986). *Hellenistic Philosophy: Stoics, Epicureans, Sceptics*. Berkeley, CA: University of California Press.
  [Recounts the central role of imagery (*phantasia*) in Stoic and Epicurean epistemology.]
- Loverock, D.S. & Modigliani, V. (1995). Visual Imagery and the Brain: A Review. *Journal of Mental Imagery* (19) 91-132.
- Lowe, E.J. (1996). *Subjects of Experience*. Cambridge: Cambridge University Press.
  [Contains a sophisticated philosophical defense of the Lockean view that the meanings of linguistic utterances are rooted in imagery. Cf. Ellis (1995), Thomas (1997b).]
- Luria, A.R. (1968). *The Mind of a Mnemonist*. (Trans. L. Solotaroff.) New York: Basic Books.
  [Seminal case study of a "hyper-imager".]
- Marks, D.F. (1973). Visual Imagery Differences in the Recall of Pictures. *British Journal of Psychology* (64) 17-24.
  [Introduces the VVIQ questionnaire, used for measuring individual differences in imagery vividness.]
- Marks, D.F. (1983). Mental Imagery and Consciousness: A Theoretical Review. In A.A. Sheikh (Ed.) *Imagery: Current Theory, Research, and Application*. New York: Wiley.
- Marks, D.F. (Ed.) (1986). *Theories of Image Formation*. New York: Brandon House.
- Marks, D.F. (1999). Consciousness, Mental Imagery and Action. *British Journal of Psychology* (90) 567-585.
  [Reviews work on individual differences in imagery vividness, and relates it to the psychology of action.]
- Matthews, G.B. (1969). Mental Copies. *Philosophical Review* (78) 53-73.
- McKellar, P. (1957). *Imagination and Thinking*. London: Cohen & West.
- McMahon, C.E. (1973). Images as Motives and Motivators: A Historical Perspective. *American Journal of Psychology* (86) 465-90.

- Mel, B.W. (1986). A Connectionist Learning Model for 3- Dimensional Mental Rotation, Zoom, and Pan. In *Program of the Eighth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Mel, B.W. (1990). *Connectionist Robot Motor Planning*. San Diego, CA: Academic Press.
  [A connectionist account of imagery, that links it to action plans.]
- Mellet, E., Tzourio, N., Crivello, F., Joliot, M., Denis, M., & Mazoyer, B. (1996). Functional anatomy of spatial mental imagery generated from verbal instructions. *Journal of Neuroscience* (16) 6504-6512.
  [Suggests that imagery does *not* depend on activity in the early, retinotopically mapped visual areas of the brain. For an opposing view see Kosslyn, Alpert et *al*. (1993), Kosslyn, Thompson *et al*. (1995), Kosslyn, Pascual-Leone *et al*. (1999).]
- Miller, A.I. (1984). *Imagery in Scientific Thought: Creating 20th Century Physics*. Boston MA: Birkhäuser.
  [Argues for an essential role for imagery in modern physical thought (and scientific thought in general).]
- Mischel, T. (1970). Wundt and the Conceptual Foundations of Psychology. *Philosophy and Phenomenological Research* (31) 1-26.
- Modrak, D.K.W. (1987). *Aristotle: The Power of Perception*. Chicago: University of Chicago Press.
- Moran, T.P. (1973). *The Symbolic Imagery Hypothesis: A Production System Model*. Unpublished Ph.D. thesis. Carnegie-Mellon University, Pittsburgh, PA. (University Microfilms 74-14,657.).
- Morris, P.E. & Hampson, P.J. (1983). *Imagery and Consciousness*. Academic Press. London.
  [Usefully summarizes much experimental evidence. Covers *quasi-pictorial*, *description*, **and** *perceptual activity* theories, and attempts a theoretical synthesis.]
- Mowrer, O.H. (1960). *Learning Theory and the Symbolic Processes*. New York: Wiley.
  [An attempt to introduce imagery into Behaviorist theory.]
- Mowrer, O.H. (1977). Mental Imagery: An Indispensible Psychological Concept. *Journal of Mental Imagery* (2) 303-321.
- Nadaner, D. (1988). Visual Imagery, Imagination, and Education. In K. Egan & D. Nadaner (Eds.). *Imagination and Education*. Milton Keynes, U.K.: Open University Press.
- Narayanan, N.H. (1993). Imagery: Computational and Cognitive Perspectives. *Computational Intelligence* (9) 303-308.
- Neisser, U. (1967). *Cognitive Psychology*. Englewood Cliffs, NJ: Prentice-Hall.
- Neisser, U. (1970). Visual Imagery as Process and as Experience. In J.S. Antrobus (Ed.). *Cognition and Affect*. Boston, MA: Little, Brown & Co.

- Neisser, U. (1972a). Changing Conceptions of Imagery. In P.W. Sheehan (Ed.). *The Function and Nature of Imagery*. London: Academic Press.
- Neisser, U. (1972b). A Paradigm Shift in Psychology. *Science* (176) 628-30.
  [A major player in the cognitive revolution places the revival of imagery research at its heart.]
- Neisser, U. (1976). *Cognition and Reality*. San Francisco, CA: W.H. Freeman.
  [Proposes a *perceptual activity* theory of imagery, an alternative to both pictorial and propositional/descriptional accounts.]
- Neisser, U. (1978a). Anticipations, Images and Introspection. *Cognition* (6) 167-174.
  [Defends the theory of Neisser (1976) from the critique of Hampson & Morris (1978).]
- Neisser, U. (1978b). Perceiving, Anticipating and Imagining. *Minnesota Studies in the Philosophy of Science* (9) 89-106.
  [Summary version of the theory of Neisser (1976).]
- Newton, N. (1982). Experience and Imagery. *The Southern Journal of Philosophy* (21) 475-487.
  [Argues the importance of non-visual modes of imagery in human experience.]
- Newton, N. (1989). Visualizing is Imagining Seeing: a reply to White. *Analysis* (49) 77-81.
- Nicholas, J.M. (Ed.) (1977). *Images, Perception and Knowledge*, (Western Ontario Studies in the Philosophy of Science, #8). Dordrecht/Boston: Reidel.
- Nussbaum, M.C. (1978). The Role of *Phantasia* in Aristotle's Explanation of Action. In her *Aristotle's De Motu Animalium: Text with Translation, Commentary, and Interpretative Essays*. Princeton, NJ: Princeton University Press.
- O'Donnell, J.M. (1985). *The Origins of Behaviorism: American Psychology, 1870-1920*. New York: New York University Press.
- Orne, M.T. (1962). On The Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and their Implications. *American Psychologist* (17) 776-783.
- Oster, G. (1970, February). Phosphenes. *Scientific American* (222-ii) 82-87.
  [Phosphenes should not be confused with mental images.]
- Paivio, A. (1971). *Imagery and Verbal Processes*. New York: Holt, Rinehart and Winston. [Reprinted in 1979 -- Hillsdale, NJ: Erlbaum]
  [Classic statement of the *Dual Coding* (imaginal and linguistic) theory of memory and mental representation, with much empirical evidence on the mnemonic effects of imagery. Paivio's work (together with Shepard's "mental rotation" experiments) probably played the key role in making imagery a scientifically respectable topic of investigation in cognitive science.]
- Paivio, A. (1977). Images, Propositions and Knowledge. In J.M. Nicholas (ed.). *Images, Perception and Knowledge*. Dordrecht/Boston, MA: Reidel.

- Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. New York: Oxford University Press.
  [A major restatement and defense of "dual-coding" theory.]
- Paivio, A. (1991). Dual Coding Theory: Retrospect and Current Status. *Canadian Journal of Psychology* (45) 255-287.
- Paivio, A. (1995). Imagery and Memory. In M.S. Gazzaniga (Ed.) *The Cognitive Neurosciences.* Cambridge, MA: MIT Press. (pp. 977-986.)
  [For an even more recent statement and defense of *Dual Coding Theory*, see Sadoski & Paivio, 2001.]
- Palmer, S.E. (1978). Fundamental Aspects of Cognitive Representation. In E. Rosch & B.B. Lloyd (Eds.), *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.
  [Argues that the "analog/propositional" debate over imagery misses the point about the nature of representation in computational theories of mind.]
- Pear, T.H. (1924). Imagery and Mentality. *British Journal of Psychology* (14) 291-299.
- Pear, T.H. (1927). The Relevance of Visual Imagery to the Process of Thinking 1. *British Journal of Psychology* (18) 1-14.
  [A companion piece to Bartlett (1927) and Aveling (1927).]
- Perky, C.W. (1910) An Experimental Study of Imagination. *American Journal of Psychology* (21) 422-52.
  [A famous study showing that mental images could be confused with (faint) percepts under certain conditions. See Segal (1971, 1972) for a modern attempt at replication.]
- Peterson, M.A., Kihlstrom, J.F., Rose, P.M., & Glisky, M.L. (1992). Mental Images Can be Ambiguous: Reconstruals and Reference Frame Reversals. *Memory and Cognition* (20), 107-123.
  [See the comment on Chambers & Reisberg (1985).]
- Petre, M. & Blackwell, A.F. (1999). Mental Imagery in Program Design and Visual Programming. *International Journal of Human-Computer Studies* (51) 7-30.
  [A study of the (apparently quite significant) role played by imagery in the thought processes of computer programming.]
- Piaget, J. & Inhelder, B. (1971). *Mental Imagery in the Child*. London: Routledge & Kegan Paul. [Originally published in French as *L'Image Mentale chez L'Enfant*. Presses Universitaires de France, 1966.]
- Pinker, S. (1980). Mental Imagery and the Third Dimension. *Journal of Experimental Psychology: General* (109) 354-71.

- Pinker, S. (1988). A Computational Theory of the Mental Imagery Medium. In M. Denis, J. Engelkamp, & J.T.E. Richardson (eds.). *Cognitive and Neuropsychological Approaches to Mental Imagery*. Dordrecht, Netherlands: Martinus Nijhoff.
  [A three-dimensional version of the "picture" (or array) theory.]

- Popper, K.R. (1976). *Unended Quest: An Intellectual Autobiography*. London: Fontana/Collins.

- Price, H.H. (1953). *Thinking and Experience*. London: Hutchinson.
  [Contains a defense of an imagery based account of thinking and meaning.]

- Pylyshyn, Z.W. (1973). What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery. *Psychological Bulletin* (80) 1-25.
  [A seminal attack on pictorial accounts of imagery. This was the opening salvo of the infamous *analog/propositional* dispute.]

- Pylyshyn, Z.W. (1978). Imagery and Artificial Intelligence. *Minnesota Studies in the Philosophy of Science* (9) 19-55.
  [Pylyshyn argues that images are best conceived of as propositional descriptions within a general computational account of mental representation.]

- Pylyshyn, Z.W. (1981). The Imagery Debate: Analogue Media Versus Tacit Knowledge. *Psychological Review* (88) 16-45.
  [A restatement of the *propositional/descriptional* account of imagery that squarely confronts the empirical arguments brought by pictorialists.]

- Rabb, J.D. (1975). Imaging: An Adverbial Analysis. *Dialogue* (14) 312-318.
  [An *adverbial* theory of imagery. Cf. Heil (1982), Tye (1984).]

- Rees, D.A. (1971). Aristotle's Treatment of *Phantasia*. In J.P. Anton & G.L. Kustas (Eds.) *Essays in Ancient Greek Philosophy*. Albany, NY: State University of New York Press.

- Reisberg, D. (Ed.) (1992). *Auditory Imagery*. Hillsdale, NJ: Erlbaum.

- Reisberg, D. (1994). Equipotential Recipes for Unambiguous Images: A Reply to Rollins. *Philosophical Psychology* (7) 359-366.
  [See Rollins (1994) and the annotation to Chambers & Reisberg (1985).]

- Reisberg, D. & Chambers, D. (1991). Neither Pictures Nor Propositions: What Can We Learn From a Mental Image? *Canadian Journal of Psychology* (45) 336-352.
  [See annotation to Chambers & Reisberg (1985).]

- Reisberg, D., Culver, L.C., Heuer, F., & Fischman, D. (1986). Visual Memory: When Imagery Vividness Makes a Difference. Journal of Mental Imagery (10) 51-74.
  [Individual differences study using the VVIQ questionnaire of Marks (1973). Vivid imagers show worse color memory than less vivid imagers. A companion piece to Heuer, Fischman, & Reisberg (1986).]

- Reisberg, D., Smith, J.D., Baxter, D.A., & Sonenshine, M. (1989). "Enacted" Auditory Images are Ambiguous; "Pure" Auditory Images are Not. *Quarterly Journal of Experimental Psychology* (41A) 619-641.
  [An auditory analogue of the effect discovered by Chambers & Reisberg (1985).]
- Reisberg, D., Wilson, M., & Smith, J.D. (1991). Auditory Imagery and Inner Speech. In R.H. Logie & M. Denis (Eds.). *Mental Images in Human Cognition.* Amsterdam: Elsevier Science Publishers B.V. (pp. 59-81).
- Rey, G. (1981). Introduction: What are Mental Images? In N. Block (Ed.) *Readings in the Philosophy of Psychology, Vol. 2*. London: Methuen.
- Rhem, L.P. (1973). Relationships Among Measures of Visual Imagery. *Behavior Research and Therapy* (11) 265-270.
- Rhodes, G. & O'Leary, A. (1985) Imagery Effects on Early Visual Processing. *Perception and Psychophysics* (37) 382-388.
- Ribot, T. (1890). *Psychologie de L'Attention.* Paris: Alcan. [Translated as: *The Psychology of Attention*. Chicago: Open Court, 1903]
  [Sketches a "motor" theory of imagery.]
- Ribot, T. (1900). *Essai sur L'Imagination Créatrice*. Paris: Alcan. [Translated as: *Essay on the Creative Imagination*. Chicago: Open Court, 1906.]
  [Includes a "motor" theory of imagery, related to those of Dunlap (1914) and Washburn (1916).]
- Richards, N. (1977). Depicting and Visualising. *Mind* (82) 218-229.
- Richardson, A. (1969). *Mental Imagery*. London: Routledge & Kegan Paul.
- Richardson, J.T.E. (1980). *Mental Imagery and Human Memory*. London: Macmillan.
  [Although the book is mainly concerned with empirical issues, chapter two is a Wittgenstein influenced philosophical discussion of the concept of imagery.]
- Richardson, J.T.E. (1999). *Mental Imagery*. Psychology Press: Hove, U.K.
  [Useful textbook surveying the cognitive psychology of imagery, including individual differences research.]
- Robson, J. (1986). Coleridge's Images of Fantasy and Imagination. In D.G. Russell, D.F. Marks, & J.T.E. Richardson (Eds.) *Imagery 2*. Dunedin, New Zealand: Human Performance Associates.
  [Imagery in Romantic psychological theory.]
- Roe, A. (1951). A Study of Imagery in Research Scientists. *Journal of Personality* (19) 459-70.

- Roland, P.E. & Gulyàs B. (1994). Visual Imagery and Visual Representation. *Trends in Neuroscience* (17) 281-286.
  [Suggests that imagery does *not* depend on activity in the early, retinotopically mapped visual areas of the brain. For an opposing view see Kosslyn, Alpert et *al*. (1993), Kosslyn, Thompson *et al*. (1995), Kosslyn, Pascual-Leone *et al*. (1999).]
- Rollins, M. (1989). *Mental Imagery: On the Limits of Cognitive Science*. New Haven, CT: Yale University Press.
- Rollins, M. (1994). Re: Reinterpreting Images. *Philosophical Psychology* (7) 345-358.
  [See Reisberg (1994) and the annotation to Chambers & Reisberg (1985).]
- Roskos-Ewoldsen, B., Intons-Peterson, M.J., & Anderson, R.E. (Eds.) (1993). *Imagery, Creativity and Discovery: a Cognitive Perspective*. Amsterdam: Elsevier.
- Roth, R.J. (1963). The Aristotelian Use of *Phantasia* and *Phantasma*. *The New Scholasticism* (37) 312-326.
- Russell, D.G., Marks, D.F., & Richardson ,J.T.E. (Eds.) *Imagery 2*. Dunedin, New Zealand: Human Performance Associates.
  [Proceedings of the Second International Imagery Conference (Swansea, Wales, 1985).]
- Russow, L.-M. (1978). Some Recent Work on Imagination. *American Philosophical Quarterly* (15) 57-66.
- Russow, L.-M. (1980). Towards a Theory of Imagination. *Southern Journal of Philosophy* (28) 353-369.
- Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.
  [Chapter 8 contains a seminal critique of pictorial accounts of imagery and questions the traditional concept of imagination as the image producing faculty. It is suggested that both imagination and imagery are conceptually related to *pretending*.]
- Ryle, G. (1971). Phenomenology versus The Concept of Mind. In his *Collected Papers, Volume 1: Critical Essays*. London: Hutchinson.
  [Some qualifications of the view expressed in Ryle (1949).]
- Sadoski, M. & Paivio, A. (2001). *Imagery and Text: A Dual Coding Theory of Reading and Writing*. Mahwah, NJ: Erlbaum.
- Samuels, M. & Samuels, N. (1975). *Seeing with the Mind's Eye: The History, Techniques and Uses of Visualization*. New York/Berkeley, CA: Random House/The Bookworks.
  [Not a scholarly work.]
- Sandbach, F.H. (1971). *Phantasia Kataleptike*. In A.A. Long (ed.). *Problems in Stoicism*. London: Athlone Press.
  [Imagery in Stoic epistemology.]

- Sarbin, T.R. (1972). Imagination as Muted Role Taking. In P.W. Sheehan (ed.). *The Function and Nature of Imagery.* Academic Press. New York. pp. 333-354.
  [A version of *perceptual activity* imagery theory, strongly influenced by Ryle (1949).]
- Sartre, J.-P. (1936/1962). *Imagination: A Psychological Critique*. Ann Arbor, MI: University of Michigan Press. (Translated by F. Williams from the original French of 1936.)
  [Useful historical material, as well as the philosophical discussion.]
- Sartre, J.-P. (1940/1948). *The Psychology of Imagination*. New York: Philosophical Library. (Translated by B. Frechtman from the original French of 1940.)
  [Argues the intentionality of imagery. Images are not inner objects.]
- Scarry, E. (1999). Dreaming by the Book. Princeton NJ; Princeton University Press.
  [A literary critic on the power of language to evoke mental imagery, and the importance of such imagery in the proper appreciation of literature. Cf. Esrock (1994).]
- Scheerer, E. (1984). Motor Theories of Cognitive Structure: A Historical Review. In W.Prinz & A.F. Sanders (Eds.), *Cognition and Motor Processes*. Berlin/Heidelberg: Springer-Verlag. (pp. 77-98).
  [Includes a brief description of Washburn's (1916) *motor* theory of imagery.]
- Schofield, M. (1978). Aristotle on the Imagination. In G.E.R. Lloyd & G.E.L. Owen (Eds.) *Aristotle on the Mind and the Senses*. Cambridge: Cambridge University Press.
- Segal, S.J. (Ed.) (1971a). *Imagery: Current Cognitive Approaches*. New York: Academic Press.
- Segal, S.J. (1971b). Processing of the Stimulus in Imagery and Perception. In S.J. Segal (1971a) pp. 73-100.
  [On attempting to replicate the Perky (1910) experiment.]
- Segal, S.J. (1972). Assimilation of a Stimulus in the Construction of an Image: The Perky Effect Revisited. In P.W. Sheehan (Ed.), *The Function and Nature of Imagery.* (pp. 203-230). New York & London: Academic Press.
- Segal, S.J. & Fusella, V. (1971). Effects of Images in Six Sense Modalities on Detection (d') of Visual Signal from Noise. *Psychonomic Science* (24) 55-56.
- Segal, S.J. & Nathan, S. (1964). The Perky Effect: Incorporation of an External Stimulus into Imagery Experience under Placebo and Control Conditions. *Perceptual and Motor Skills* (18) 385-395.
- Sergent, J. (1990). The Neuropsychology of Visual Image Generation: Data, Method, and Theory. *Brain and Cognition* (13) 98-129.

- Sheehan, P.W. (Ed.) (1972). *The Function and Nature of Imagery*. Academic Press. New York & London.
  [Valuable anthology of the "state of the art" at the time.]
- Sheehan, P.W. (1978). Mental Imagery. In B.M. Foss (Ed.) *Psychology Survey. No.1*. London: Allen & Unwin.
  [Helpful review article, but now dated.]
- Sheikh, A.A. (Ed.) (1983). *Imagery: Current Theory, Research, and Application*. New York: Wiley.
  [Useful, wide ranging collection.]
- Shepard, R.N. (1975). Form, Formation, and Transformation of Internal Representations. In R.L. Solso (Ed.) *Information Processing and Cognition: the Loyola Symposium*. Hillsdale, NJ: Erlbaum.
  [Defends "analog" account of imagery. Introduces concept of "second order isomorphism".]
- Shepard, R.N. (1978a). Externalization of Mental Images and the Act of Creation. In B.S. Randhawa & B.F. Coffman (Eds.). *Visual Learning, Thinking and Communication*. London: Academic Press.
- Shepard, R.N. (1978b). The Mental Image. *American Psychologist* (33) 125-137.
- Shepard, R.N. (1981). Psychophysical Complementarity. In M. Kubovy & J.R. Pomerantz (Eds.) *Perceptual Organization.* Hillsdale, NJ: Erlbaum.
- Shepard, R.N. (1984). Ecological Restraints on Internal Representation. *Psychological Review* (91) 417-447.
- Shepard, R.N., Cooper, L.A., *et al*. (1982). *Mental Images and Their Transformations*. Cambridge, MA: MIT Press.
  [A useful compendium of the seminal work by Shepard and his students on the "mental rotation" of images (and on related phenomena).]
- Shepard, R.N. & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science* (171) 701-703.
  [A classic psychological experiment. The first, most striking, and best known of the mental rotation studies. Together with the work on the mnemonic effects of imagery (see Paivio, 1971) this played a major role in re-establishing the scientific respectability of imagery research.]
- Shepard, R.N. & Podgorny, P. (1978). Cognitive Processes That Resemble Perceptual Processes. In W.K. Estes (Ed.) *Handbook of Learning and Cognitive Processes*. Hillsdale, NJ: Erlbaum.
- Shorter, J.M. (1952). Imagination. *Mind* (61) 528-542.
  [Perhaps the earliest suggestion that imagining is more like describing than like seeing a picture (C.f. Dennett, 1969).]

- Simon, H.A. (1972). What is Visual Imagery? An Information Processing Interpretation. In L.W. Gregg (Ed.). *Cognition in Learning and Memory*. New York: Wiley.
  [Early sketch of a computational model of imagery.]

- Slezak, P. (1991). Can Images be Rotated and Inspected? A Test of the Pictorial Medium Theory. In *Proceedings, Thirteenth Annual Conference of the Cognitive Science Society* (pp. 55-60). Hillsdale, NJ: Erlbaum.
  [See note at Chambers & Reisberg (1985).]

- Slezak, P. (1995). The "Philosophical" Case Against Visual Imagery. In P. Slezak, T. Caelli, & R. Clark (Eds.) *Perspectives on Cognitive Science: Theories, Experiments and Foundations*. Norwood, NJ: Ablex.
  [A recent attempt to press the cognitivist case against pictorialism by a psychologically sophisticated philosopher.]

- Sober, E. (1976). Mental Representations. *Synthése* (33) 101-148.

- Squires, J.E.R. (1968). Visualising. *Mind* (77) 58-67.

- Sterelny, K. (1986). The Imagery Debate. *Philosophy of Science* (53) 560-583.
  [A philosopher's take on the "analog/propositional" debate.]

- Taylor, J.G. (1973). A Behavioural Theory of Images. *South African Journal of Psychology* (3) 1-10.
  [A rare attempt to assimilate imagery into Behaviorist theory.]

- Taylor, P. (1981). Imagination and Information. *Philosophy and Phenomenological Research* (42) 205-223.
  [Despite arguments to the contrary from Sartre and Wittgenstein, we *can* gain new information from our mental imagery.]

- Thomas, N.J.T. (1987). *The Psychology of Perception, Imagination and Mental Representation, and Twentieth Century Philosophies of Science*. Unpublished Ph.D. thesis, Leeds University, Leeds, U.K. (A.S.L.I.B. Index to Theses 37-iii No. 37-4561).

- Thomas, N.J.T. (1989). Experience and Theory as Determinants of Attitudes toward Mental Representation: The Case of Knight Dunlap and the Vanishing Images of J.B. Watson. *American Journal of Psychology* (102) 395-412. [Available online]
  [Discusses the historical circumstances surrounding the "banishment" of imagery from psychological theory in the Behaviorist tradition, and considers certain conceptual confusions that may induce some people to discount the psychological significance of imagery. Dunlap's (1914) theory is outlined.]

- Thomas, N.J.T. (1997a). Imagery and the Coherence of Imagination: a Critique of White. *Journal of Philosophical Research*, (22) 95-127. [Available online]
  [Defends the traditional (Aristotelian) view of the concept of imagination as derivative from the concept of imagery, and argues that the root concept of both is *perceiving as*. Traces resistance to the Aristotelian view to unsupported pictorialist assumptions.]

- Thomas, N.J.T. (1997b). A Stimulus to the Imagination. *Psyche* (3) (On-line serial). [Available online]

  [An essay review of Ellis (1995), which sets his work in a historical context and reviews some standard objections to the sort of imagery based semantics he proposes.]

- Thomas, N.J.T. (1999a). Imagination. In *Dictionary of Philosophy of Mind* (online dictionary), Chris Eliasmith (ed.). [Available online]

  [Provides a brief sketch of the history of the concept, from Aristotle to the present.]

- Thomas, N.J.T. (1999b). Are Theories of Imagery Theories of Imagination? An Active Perception Approach to Conscious Mental Content. *Cognitive Science* (23) 207-245. [Available online]

  [Assesses cognitive theories of imagery both in empirical terms and in the light of their relationship to traditional views of imagination. Proposes and defends *Perceptual Activity Theory* as an alternative that is empirically and conceptually superior to both *quasi-pictorial* and *propositional* theories.]

- Titchener, E.B. (1909). *Lectures on the Experimental Psychology of the Thought-Processes*. New York: Macmillan.
  [A radical defense of an image centered introspective psychology against the claims of the Wurzburg *imageless thought* school of introspectors.]

- Trehub, A. (1977). Neuronal Models for Cognitive Processes: Networks for Learning, Perception and Imagination. *Journal of Theoretical Biology* (65) 141-169.

- Trehub, A. (1991). *The Cognitive Brain*. Cambridge, MA: MIT Press.
  [Ambitious neuroscientific theory, treating imagery as activity in the retinotopic maps of the visual system.]

- Toulmin, S. (1969). Ludwig Wittgenstein. *Encounter* (32) 58-71.

- Tweedale, M.M. (1990). Mental Representations in Later Medieval Scholasticism. In J.-C. Smith (Ed.). *Historical Foundations of Cognitive Science*. Dordrecht, Netherlands: Kluwer.

- Tweney, R.D. (1987). Programmatic Research in Experimental Psychology: E.B. Titchener's Laboratory Investigations, 1891-1927. In M.G. Ash & W.R. Woodward (Eds.), *Psychology in Twentieth Century Thought and Society* (pp.34-57). Cambridge: Cambridge University Press.

- Tweney, R.D., Doherty, M.E., & Mynatt, C.R. (Eds.) (1981). *On Scientific Thinking*. New York: Columbia University Press.
  [Contains anecdotal but very suggestive extracts concerning the key role that imagery can play in the thought processes of scientists.]

- Tye, M. (1984). The Debate About Mental Imagery. *Journal of Philosophy* (81) 678-691.
  [An *adverbial* account of imagery that is abandoned in Tye's later writings on the subject. Cf. Rabb (1975) and Heil (1982) for other defenses of the *adverbial* theory.]

- Tye, M. (1988). The Picture Theory of Mental Images. *Philosophical Review* (97) 497-520.
  [A persuasive defense of "quasi-pictorial" theory against "descriptionist" criticisms.]

- Tye, M. (1991). *The Imagery Debate*. Cambridge, MA: MIT Press.
  [This fills out the argument of Tye (1988) and gives an excellent philosophical account of the "analog/propositional" debate and the conceptual basis of (quasi-)pictorialism. However, it fails to look seriously beyond this context, and is rather unreliable on historical and empirical issues.]
- von, Eckardt B. (1988). Mental Images and Their Explanations. *Philosophical Studies* (53) 441-460.
  [A critique of Tye's (1984) *adverbial* theory.]
- Warnock, M. (1976). *Imagination*. London: Faber & Faber.
  [Imagery and imagination in Hume and Kant, in Romantic theory, and in Sartre and Wittgenstein.]
- Washburn, M.F. (1916). *Movement and Mental Imagery*. Boston, MA: Houghton Mifflin.
  [A *motor* theory of imagery. See Dunlap (1914) for another version.]
- Washburn, M.F. (1932). Some Recollections. In C. Murchison (Ed.), *A History of Psychology in Autobiography, Vol.2*. Worcester, MA: Clark University Press. (pp. 333-358). [Available online]
- Watson, G. (1982). *Phantasia* in Aristotle, *De Anima 3.3*. *Classical Quarterly* (32) 100-113.
- Watson G. (1988). *Phantasia in Classical Thought*. Galway, Republic of Ireland: Galway University Press.
- Watson, J.B. (1913a). Psychology as the Behaviorist Views It. *Psychological Review* (20) 158-177. [Available online]
  [The classic "Behaviorist manifesto". Questions the very existence of imagery.]
- Watson, J.B. (1913b). Image and Affection in Behavior. *Journal of Philosophy, Psychology and Scientific Methods* (10) 421-8.
  [A more careful and detailed version of the anti-imagery position put forward in Watson (1913a).]
- Watson, J.B. (1924). *Psychology from the Standpoint of a Behaviorist* (2nd ed.). Philadelphia, PA: Lippincott.
- Watson, J.B. (1928). *The Ways of Behaviorism.* New York: Harper.
  [Reports of memory images are "sheer bunk".]
- Watson, J.B. (1930). *Behaviorism* (2nd ed.). Chicago: University of Chicago Press.
- Watson, J.B. (1936). John Broadus Watson. In C. Murchison (Ed.), *A History of Psychology in Autobiography* (Vol. 3, pp. 271-281). Worcester, MA: Clark University Press.
- Wekker, L.M. (1966). On the Basic Properties of the Mental Image and a General Approach to their Analogue Simulation. In *Psychological Research in the U.S.S.R.* Moscow: Progress Publishers.
  [Imagery theory in the Soviet psychological tradition. Somewhat similar to the *motor* theories of Dunlap (1914) and Washburn (1916).]

- Wexler, M., Kosslyn, S.M., & Berthoz, A. (1998). Motor Processes in Mental Rotation. *Cognition* (68) 77-94.
- Wheeler, M.E., Petersen, S.E., & Buckner, R.L. (2000). Memory's Echo: Vivid Remembering Reactivates Sensory Specific Cortex. *Proceedings of the National Academy of Sciences of the U.S.A.* (97) 11125-11129.
- White, A.R. (1990). *The Language of Imagination*. Oxford: Blackwell.
  [Part one is an excellent, if selective, concise history of the concept of imagination. Part 2 argues (in the teeth of the strong historical consensus detailed in part 1) that there is *no* conceptual connection between imagination and imagery. See Thomas (1997a) for a critique of this view.]
- Wright, E. (1983). Inspecting Images. *Philosophy* (58) 57-72.
- Wundt, W. (1912). *An Introduction to Psychology* (2nd edn.). New York: Macmillan. [Translated from the German.]
- Yates, F.A. (1966). *The Art of Memory*. London: Routledge and Kegan Paul.
  [A celebrated and seminal history of mnemonic uses of imagery, from ancient to early modern times. Such techniques are shown to have had a previously unrecognized importance in the history of western thought.]
- Yuille, J.C. (Ed.) (1983). *Imagery, Memory and Cognition: Essays in Honour of Allan Paivio*. Hillsdale, NJ: Erlbaum.
- Zemach, E.M. (1969). Seeing, "Seeing", and Feeling. *Review of Metaphysics* (23) 3-24.
- Zimler, J. & Keenan, J.M. (1983). Imagery in the Congenitally Blind: How Visual are Visual Images? *Journal of Experimental Psychology: Learning, Memory, and Cognition* (9) 269-282.

Other Internet Resources

- Imagination, Mental Imagery, Consciousness, and Cognition: Scientific, Philosophical, and Historical Approaches. (Nigel Thomas, California State University/L.A.)
- Stanford Encyclopedia of Philosophy: "Aristotle on imagination", supplement to Aristotle's Psychology, by Christopher Shields (U. Colorado)

Related Entries

behaviorism | cognitive science | consciousness | Descartes, René | intentionality | mental representation | perception

Nigel J. T. Thomas
n.j.thomas70@members.leeds.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 18, 1997
Content last modified: November 10, 2001

Stanford Encyclopedia of Philosophy
Supplement to Aristotle's Psychology

The Subordinate Psychic Faculties of Imagination and Desire

Imagination

Aristotle sometimes recognizes as a distinct capacity, on par with perception and mind, imagination (*phantasia*) (*De Anima* iii 3, 414b33-415a3). Although he does not discuss it at length, or even characterize it intrinsically in any detailed way, Aristotle does take pains to distinguish it from both perception and mind. In a brief discussion dedicated to imagination (*De Anima* iii 3), Aristotle identifies it as "that in virtue of which an image occurs in us" (*De Anima* iii 3, 428aa1-2), where this is evidently given a broad range of application, including the activities involved in thoughts, dreams, and memories. Aristotle is, however, mainly concerned to distinguish imagination from perception and mind. He distinguishes it from perception on a host of grounds, including: (i) imagination produces images when there is no perception, as in dreams; (ii) imagination is lacking in some lower animals, even though they have perception, which shows that imagination and perception are not even co-extensive; and (iii) perception is, Aristotle claims, always true, whereas imagination can be false, false even in fantastic ways (*De Anima* iii 3, 428a5-16). He also denies that imagination can be identified with mind or belief, or any combination of belief and perception (*De Anima* iii 3, 428a16-b10), even though it comes about through sense perception (*De Anima* iii 3, 429a1-2; *De Insomniis* 1, 459a17). The suggestion, then, is that imagination is a faculty in humans and most other animals which produces, stores, and recalls the images used in a variety of cognitive activities, including those which motivate and guide action (*De Anima* iii 3, 429a4-7, *De Memoria* 1, 450a22-25).

Because he tends to treat imagination pictographically (*De Anima* iii 3, 429a2-4; cf. *De Sensu* 1, 437a3-17; 3, 439b6), Aristotle seems to regard the images used in cognitive processes as representations best thought of on the model of copies or likenesses of external objects. This much he holds in common with many other empirically oriented cognitive psychologists. Typically he will suggest, in this vein, that thought requires images, both genetically and concurrently, so that "whenever one contemplates, one necessarily at the same time contemplates in images" (*De Anima* iii 8, 432a8-9, 431a16-17; *De Memoria* 1, 449b31-450a1). His suggestions in this direction may seem unfortunate, since for a broad range of thoughts, images, construed naturally and narrowly as pictorial representations, seem unnecessary or even plainly irrelevant. (It is hard to fathom, e.g., what image corresponds to the thought that gerunds make for ungainly syntax--still less is it clear what grounds could compel one to agree that some image or other must accompany it). Perhaps, though, his remarks should be tempered by the recognition that Aristotle accepts the existence of a thinking god whose activity is exhausted by thinking, but whose thinking is not plausibly regarded as imagistic (*Metaphysics* xii 7,1072b26-30). If that is so, Aristotle

could not accept the thesis that for any episode of thought *t*, necessarily *t* is or is directed upon a pictorial image. Still, Aristotle clearly expects images, so construed, to play a central or even indispensable role in human cognition.

Desire

[Under Construction]

Christopher Shields
shields@colorado.edu

Return to Aristotle's Psychology

First posted: January 11, 2000
Last modified: January 11, 2000

Aristotle's Psychology

Aristotle (384-322 BC) was born in what was to become Macedon in northern Greece, but spent most of his adult life in Athens. His life in Athens divides into two periods, first as a member of Plato's Academy (367-347) and later as director of his own school, the Lyceum (334-323). The intervening years were spent mainly in Assos and Lesbos, and briefly back in Macedon. His years away from Athens were predominantly taken up with biological research and writing. Judging on the basis of their content, Aristotle's most important psychological writings probably belong to his second residence in Athens, and so to his most mature period. His principal work in psychology, *De Anima*, reflects in different ways his pervasive interest in biological taxonomy and his most sophisticated physical and metaphysical theory.

Because of the long tradition of exposition which has developed around Aristotle's *De Anima*, the interpretation of even its most central theses is sometimes disputed. Moreover, because of its evident affinities with some prominent approaches in contemporary philosophy of mind, Aristotle's psychology has received renewed interest and has incited intense interpretative dispute in recent decades. Consequently, this entry proceeds on two levels. The main article recounts the principal and distinctive claims of Aristotle's psychology, avoiding so far as possible exegetical controversy and critical commentary. At the end of appropriate sections of the main article, readers are invited to explore problematic or advanced features of Aristotle's theories by moving to a lower level.

Aristotle's Psychological Writings

Aristotle investigates psychological phenomena primarily in *De Anima* and a loosely related collection of short works called the *Parva Naturalia*, whose most noteworthy pieces are *De Sensu* and *De Memoria*. He also touches upon psychological topics, often only incidentally, in his ethical, political, and metaphysical treatises, as well as in his scientific writings, especially *De Motu Animalium*. The works in the *Parva Naturalia* are, in comparison with *De Anima*, empirically oriented, investigating, as Aristotle says, "the phenomena common to soul and body" (*De Sensu* 1, 436a6-8). This contrasts with *De Anima*, which introduces as a question for consideration "whether all affections are common to what has the soul or whether there is some affection peculiar to the soul itself" (*De Anima* i 1, 402a3-5). That is, in *De Anima* Aristotle wants to know whether all psychological states are also material states of the body. "This," he remarks, "it is necessary to grasp, but not easy" (*De Anima* i 1, 402a5). In this way, *De Anima* proceeds at a higher level of abstraction than the *Parva Naturalia*. It is generally more theoretical, more self-conscious about method, and more alert to general philosophical questions about perception, thinking, and soul-body relations.

In both *De Anima* and the *Parva Naturalia*, Aristotle assumes something which may strike some of his modern readers as odd. He takes psychology to be the branch of science which investigates the soul and its properties, but he thinks of the soul as a general principle of life, with the result that psychology studies all living beings, and not merely those he regards as having minds, human beings. So, in *De Anima*, he takes it as his task to provide an account of the life activities of plants and animals, along side those of humans (*De Anima* ii 11, 423a20-6, cf. ii 1, 412a13; cf. *De Generatione Animalium* ii 3, 736b13; *De Partibus Animalium* iv 5, 681a12). In comparison with the modern discipline of Psychology, then, Aristotle's psychology is broad in scope. He even devotes attention to the question of the nature of life itself, a subject which falls outside the purview of psychology in most contemporary contexts. On Aristotle's approach, psychology studies the soul (*psuchê* in Greek, or *anima* in Latin); so it naturally investigates all ensouled or animate beings.

Hylomorphism in General

In *De Anima*, Aristotle makes extensive use of technical terminology introduced and explained elsewhere in his writings. He claims, for example, using vocabulary derived from his physical and metaphysical theories, that the soul is a "first actuality of a natural organic body" (*De Anima* ii 1, 412b5-6), that it is a "substance as form of a natural body which has life in potentiality" (*De Anima* ii 1, 412a20-1) and, similarly, that it "is a first actuality of a natural body which has life in potentiality" (*De Anima* ii 1, 412a27-8), all claims which apply to plants, animals and humans alike.

In characterizing the soul and body in these ways, Aristotle applies concepts drawn from his broader *hylomorphism*, a conceptual framework which underlies virtually all of his mature theorizing. It is accordingly necessary to begin with a brief overview of that framework. Thereafter it will be possible to recount Aristotle's general approach to soul-body relations, and then, finally, to consider his analyses of

the individual faculties of soul.

‘Hylomorphism’ is simply a compound word composed of the Greek terms for matter (*hulê*) and form or shape (*morphê*); thus one could equally describe Aristotle's view of soul and body as an instance of his "matter-formism". That is, when he introduces the soul as the *form* of the body, which in turn is said to be the *matter* of the soul, Aristotle treats soul-body relations as a special case of a more general relationship which obtains between the components of all generated compounds, natural or artifactual.

The notions of form and matter are themselves, however, developed within the context of a general theory of causation and explanation which appears in one guise or another in all of Aristotle's mature works. According to this theory, when we wish to explain what there is to know, for example, about a bronze statue, a complete account necessarily alludes to at least the following four factors: the statue's matter, its form or structure, the agent responsible for that matter's manifesting its form or structure, and the purpose for which the matter was made to realize that form or structure. These four factors he terms the four causes (*aitiai*):

The *material cause*: that from which something is generated and out of which it is made, e.g. the bronze of a statue.
The *formal cause*: the structure which the matter realizes and in terms of which the matter comes to be something determinate, e.g. the Hermes shape in virtue of which this quantity of bronze is said to be a statue of Hermes.
The *efficient cause*: the agent responsible for a quantity of matter's coming to be informed, e.g. the sculptor who shaped the quantity of bronze into its current Hermes shape.
The *final cause*: the purpose or goal of the compound of form and matter, e.g. the statue was created for the purpose of honoring Hermes.

For a broad range of cases, Aristotle implicitly makes twin claims about these four causes: (i) a complete explanation requires reference to all four; and (ii) once such reference is made, no further explanation is required. Thus, when appropriate, appeal to the four causes is both necessary and sufficient for complete and adequate explanation. Although not all things which admit of explanation have all four causes, e.g. geometrical figures are not efficiently caused, even a brief overview of his psychological writings reveals that Aristotle regards all four causes as in play in the explanation of living beings. A monkey, for example, has matter, its body; form, its soul; an efficient cause, its parent; and a final cause, its function. Moreover, he holds that the form is the *actuality* of the body which is its matter: an indeterminate lump of bronze becomes a statue only when it realizes some particular statue-shape. So, Aristotle suggests, that matter is *potentially* some F until it acquires an actualizing form, when it becomes actually F. Given his overarching explanatory schema, it is hardly surprising that Aristotle should advance a hylomorphic account of soul and body; this is, for him, standard explanatory procedure.

Still, it is noteworthy that this four-causal framework of explanation is developed initially in response to

some puzzles about change and generation. Aristotle argues with some justification that all change and generation require the existence of something complex: when a statue comes to be from a lump of bronze, there is some continuing subject, the bronze, and something it comes to acquire, its new form. Thus the statue is, and must be, a certain kind of compound, one of form and matter. Without this type of complexity, generation would be impossible; since generation in fact occurs, form and matter must be genuine features of generated compounds. Similarly, but less obviously, qualitative change requires much the same apparatus: when a statue is painted, there is some continuing subject, the statue, and a new feature acquired, its new color. Here too there is complexity, and complexity which is readily articulated in terms of form and matter, but now of form which is evidently inessential to the continued existence of the entity whose form it is. The statue continues to exist, but receives a form which is accidental to it; it might lose that form without going out of existence. By contrast, should the statue lose its essential form, as would happen for example if the bronze which constitutes it were melted, divided, and recast as twelve dozen letter openers, it would cease to exist altogether.

For the purposes of understanding Aristotle's psychology, the origin of Aristotle's hylomorphism is significant for two reasons. First, from its inception, Aristotle's hylomorphism exploits two distinct but related notions of form, one of which is essential to the compound whose form it is, and the other of which is accidental to its subject. In advancing his view of the soul and its capacities, Aristotle employs both of these notions: the soul is an essential form, whereas perception involves the acquisition of accidental forms. Second, because Aristotle's hylomorphism was initially developed to handle puzzles of change and generation, its deployment in philosophical psychology is sometimes strained, insofar as Aristotle is not immediately willing to treat every instance of perception and thought as a straightforward instance of change in some continuing subject.

Hylomorphic Soul-Body Relations

In applying his general hylomorphism to soul-body relations, Aristotle contends that the following general analogy obtains:

soul : body : : form : matter : : Hermes-shape : bronze

If the soul bears the same relation to the body which the shape of a statue bears to its material basis, then we should expect some general features to be common to both; and we should be able to draw some immediate consequences regarding the relationship between soul and body. To begin, some questions about the unity of soul and body, an issue of concern to substance dualists and materialists alike, receive a ready response. Materialists hold that all mental states are also physical states; substance dualists deny this, because they hold that the soul is a subject of mental states which can exist alone, when separated from the body. In a certain way, the questions which give rise to this dispute simply fall by the wayside. If we do not think there is an interesting or important question concerning whether the Hermes-shape and its material basis are one, we should not suppose there is a special or pressing question about whether the soul and body are one. So Aristotle contends: "It is not necessary to ask whether soul and body are one,

just as it is not necessary to ask whether the wax and its shape are one, nor generally whether the matter of each thing and that of which it is the matter are one. For even if one and being are spoken of in several ways, what is properly so spoken of is the actuality" (*De Anima* ii 1, 412b6-9). Aristotle does not here eschew questions concerning the unity of soul and body as meaningless; rather, he seems, in a deflationary vein, to suggest that they are readily answered or somehow unimportant. If we do not spend time worrying about whether the wax of a candle and its shape are one, then we should not exercise ourselves over the question of whether the soul and body are one. The effect, then, is to fit soul-body relations into a larger pattern of explanation, hylomorphism, in terms of which questions of unity do not normally arise.

It should be emphasized, however, that Aristotle does not here decide the question by insisting that the soul and body are identical, or even that they are one in some weaker sense; indeed, this is something he evidently denies (*De Anima* ii 1, 412a17; ii 2, 414a1-20). Instead, just as one might well insist that the wax of a candle and its shape are distinct, on the grounds that the wax could easily exist when the particular shape is no more, or, less obviously, that the particular shape could survive the replenishment of its material basis, so one might equally deny that the soul and body are identical. In a fairly direct way, though, the question of whether soul and body are one loses its force when it is allowed that it contains no implications beyond those we establish for any other hylomorphic compound, including houses and other ordinary artifacts.

One way of appreciating this is to consider a second general moral Aristotle derives from hylomorphism. This concerns the question of the separability of the soul from the body, a possibility embraced by substance dualists from the time of Plato onward. Aristotle's hylomorphism commends the following attitude: if we do not think that the Hermes-shape persists after the bronze is melted and recast, we should not think that the soul survives the demise of the body. So, Aristotle claims, "It is not unclear that the soul-or certain parts of it, if it naturally has parts-is not separable from the body" (*De Anima* ii 1, 413a3-5). So, unless we are prepared to treat forms in general as capable of existing without their material bases, we should not be inclined to treat souls as exceptional cases. Hylomorphism, by itself, gives us no reason to treat souls as separable from bodies, even if we think of them as distinct from their material bases. At the same time, Aristotle does not appear to think that his hylomorphism somehow refutes all possible forms of dualism. For he appends to his denial of the soul's separability the observation that some parts of the soul may in the end be separable after all, since they are not the actualities of any part of the body (*De Anima* ii 1, 413a6-7). Aristotle here prefigures his complex attitude toward mind (*nous*), a faculty he repeatedly describes as exceptional among capacities of the soul.

Still, in general, the soul is the form of the body in much the same way the form of a house structures the bricks and mortar from which it is built. When the bricks and mortar realize a certain shape, they manifest the function definitive of houses, namely that of providing shelter. Thus, the presence of the form makes those bricks and that mortar a house, as opposed, e.g., to a wall or an oven. As we have seen, Aristotle will say that the bricks and mortar, as matter, are potentially a house, until they realize the form appropriate to houses, in which case the form and matter together make an actual house. So, in Aristotle's terms, the form is the actuality of the house, since its presence explains why this particular quantity of

matter comes to be a house as opposed to some other kind of artifact.

In the same way, then, the presence of the soul explains why this matter is the matter of a human being, as opposed to some other kind of thing. Now, this way of looking at soul-body relations as a special case of form-matter relations treats reference to the soul as an integral part of any complete explanation of a living being, of any kind. To this degree, Aristotle thinks that Plato and other dualists are right to stress the importance of the soul in explanations of living beings. At the same time, he sees their commitment to the separability of the soul from the body as unmotivated by a mere appeal to formal causation: he will allow that the soul is distinct from the body, and is indeed the actuality of the body, but he sees that these concessions by themselves provide no grounds for supposing that the soul can exist without the body. His hylomorphism, then, embraces neither reductive materialism nor Platonic dualism. Instead, it seeks to steer a middle course between these alternatives by pointing out, implicitly, and rightly, that these are not exhaustive options.

Psychic Faculties

Although willing to provide a common account of the soul in these general terms, Aristotle devotes most of his energy in *De Anima* to detailed investigations of the soul's individual capacities, which include nutrition, perception, and mind, with perception receiving the lion's share of attention. The broadest is nutrition, which is shared by all natural living organisms; animals have perception in addition; and among natural organisms humans alone have mind. Aristotle maintains that various kinds of souls, nutritive, perceptual, and intellectual, form a kind of hierarchy. Any creature with reason will also have perception; any creature with perception will also have the ability to take on nutrition and to reproduce; but the converse does not hold. Thus, plants show up with only the nutritive soul, animals have both perceptual and nutritive faculties, and humans have all three. The reasons why this should be so are broadly teleological. In brief, every living creature as such grows, reaches maturity, and declines. Without a nutritive capacity, these activities would be impossible (*De Anima* iii 12, 434a22-434b18; cf. *De Partibus Animalium* iv 10, 687a24-690a10; *Metaphysics*xii 10, 1075a16-25). So, Aristotle concludes, psychology must investigate not only perceiving and thinking, but also nutrition.

There is some dispute about which of the psychic abilities mentioned by Aristotle in *De Anima* qualify as full-fledged or autonomous faculties. He evidently accepts the three already mentioned as centrally important. Indeed, he is willing to demarcate a hierarchy of life in terms of them. Even so, he also discusses two other capacities, imagination (*De Anima* iii 3) and desire (*De Anima* iii 9 and 10), and appeals to them both in his account of thinking and his philosophy of action. He does little, however, to characterize them in any intrinsic way; he evidently regards them as subordinate faculties, integrated in various ways with the faculties of nutrition, perception, and thought. Each of them raises interesting questions about how Aristotle views the various capacities of soul as integrating into unified forms. His main deployment of a hylomorphic analysis, however, extends only to the individual capacities of nutrition, perception, and mind.

Nutrition

When turning to these individual faculties of the soul, Aristotle considers nutrition first, for two related reasons. The first is straightforward: psychology considers all animate entities, and the nutritive soul belongs to all naturally living things, since it is "the first and most common capacity of soul, in virtue of which life belongs to all living things" (*De Anima* ii 4, 415a24-25). The second is slightly more complex, being at root teleological. Given that the higher forms of soul presuppose nutrition, its explication is prior to them in the order of Aristotle's exposition.

Aristotle approaches his account of the nutritive soul by relying on a methodological precept which informs much of his psychological theorizing, namely that a capacity is individuated by its objects, so that, e.g., perception is distinguished from mind by being arrayed toward sensible qualities rather than intelligible forms (*De Anima* ii 4, 415a20-21). This induces him to offer what may sound initially like a pedestrian observation, that in nutrition there are three components, "that which is nourished, that by which it is nourished, and what nourishes (i.e. that which engages in nutrition)." This, however, Aristotle unpacks by maintaining that "what nourishes is the primary soul; what is nourished is the body which has this soul; and that by which it is nourished is nourishment (i.e. food)" (*De Anima* ii 4, 416b20-23). The interest of this suggestion lies in the implication that all and only living systems can be nourished, a consequence Aristotle makes more explicit by claiming that "nothing is nourished which does not have a share in life" (*De Anima* ii 4. 415b27-28) and that "since nothing is nourished which does not partake of life, what is nourished will be the ensouled body insofar as it is ensouled, with the result that nourishment (i.e. food) is related to the ensouled, and not coincidentally" (*De Anima* ii 4, 416b9-11). Here Aristotle means that food, as food, is definitionally related to life. Whatever is food is already such as to be necessarily related to living beings.

The significance of this observation resides in the thought that any adequate account of nutrition will make ineliminable reference to life as such. This in turn entails that it will not be possible to *define* life as the capacity for taking on nutrition. For then we would have a vicious circularity: a living system is the sort of thing which can take on nutrition, while nutrition is whatever stuff is such as to sustain a living system. So, if living systems cannot be reductively defined in some other way, it will follow that no reductive account of life will be forthcoming. Consequently, Aristotle's discussion of nutrition provides some reason for thinking that he will resist any attempt to define life in terms which do not themselves implicitly appeal to life itself. That is, he will resist any reductive account of life.

This also seems to be the purport of Aristotle's rejection of the simple mechanistic accounts of growth which he considers when discussing the nutritive soul (*De Anima* ii 4, 415b27-416a20; cf. *De Generatione et Corruptione* i 5). Aristotle objects to those who want to account for growth merely in terms of the natural tendencies of material elements. For growth is a *constrained* pattern of development, the source of which Aristotle ascribes to the soul. He takes it as manifest that growth in organisms proceeds along structured paths, in end-directed ways. These structures in turn manifest capacities whose explication cannot be given in crude materialistic terms; for materialistic terms, as Aristotle understands them, fail to account for the fact that mature members of species cease growing, having realized the

structures characteristic of their kind. Fire, for example, by contrast "grows" haphazardly, without directionality, moving toward the combustible without end, until hindered by external impediments or lack of fuel.

Now, the forms of materialist explanations Aristotle considers are primitive. One critical question about his treatment of these explanations concerns whether he is right to suggest that facts about constrained patterns of development are incompatible with more explanatorily advanced forms of materialism, and, if so, whether those forms of materialism will be reductive in the sense that they will avoid all implicit or explicit reference to life. So far, there is little reason to think that Aristotle has been proven wrong; that is, there is at present no reductive account of life which enjoys universal or even broad support.

In any case, Aristotle's discussion of nutrition is characteristic of his general approach to the soul's faculties. His discussions often proceed on two levels. On the one hand, he simply seeks to provide an account of the relevant phenomena. At the same time, his interests in definition are conditioned by a host of broader methodological and metaphysical concerns. Consequently, he attempts to capture the nature of the individual faculties while at the same time investigating whether reductive accounts of them are plausible. In this way, at least, Aristotle's investigations reflect sensitivity to a host of questions in definitional methodology, including most notably questions about the plausibility of reductive approaches to life's most characteristic features. These same interests are apparent in his discussions of perception and mind.

Perception

Aristotle devotes a great deal of attention to perception, discussing both the general faculty and the individual senses. In both cases, his discussions are cast in hylomorphic terms. Perception is the capacity of the soul which distinguishes animals from plants; indeed, having a perceptive faculty is definitive of being an animal (*De Sensu* 1, 436b10-12). In broad terms at least, animals must have perception if they are to live. So, Aristotle supposes, there are defensible teleological grounds for treating animals as essentially capable of perceiving (*De Anima* ii 3, 414b6-9, 434a30-b4; *De Sensu* 1, 436b16-17). If an animal is to grow to maturity and propagate, it must be able to take in nourishment and to navigate its way through the world. Perception serves these ends.

This much, however, does not explain how perception occurs. Aristotle claims that perception is best understood on the model of hylomorphic change generally: just as a house changes from blue to white when acted upon by the agency of a painter applying paint, so "perception comes about with <an organ's> being changed and affected. . .for it seems to be a kind of alteration" (*De Anima* ii 5, 416b33-34). So, in line with his general account of alteration, Aristotle treats perception as a case of interaction between two suitable agents: objects capable of acting and capacities capable of being affected. That the agents and patients must both be *suitable* is important, since we need to distinguish between two ways, e.g., an odor might affect something. By being placed in its vicinity, a clove of garlic might affect a block of tofu. The tofu might come to take on the odor of the garlic. But we would not want to say that the tofu perceives the garlic. By contrast, when an animal is affected by the same clove, it perceives the odor.

Since the garlic is the same in both cases, the difference in these cases must reside in the character of the object affected: when animals receive forms, perception results.

In both kinds of alterations, Aristotle is happy to speak of an affected thing as receiving the form of the agent which affects it and of the change consisting in the affected thing's "becoming like" the agent (*De Anima* ii 5, 418a3-6; ii 12, 424a17-21). So, there is in both cases, a hylomorphic model of alteration involving "enforming", that is, a model according to which change is explained by the acquisition of a form by something capable of receiving it. Consequently, whatever is changed in a given way is necessarily such that it is capable of being changed in that way. This is not the mere triviality that whatever becomes actually F must already be possibly F. Instead, it is the recognition that specific forms of change require suitable capacities in the changing subjects, and that, consequently, analyses of specific forms of change will necessarily involve consideration of those capacities. No marshmallow can receive the form of an actual automobile; and only entities capable of perceiving can receive the perceptible forms of objects. This is Aristotle's meaning when he claims: "the perceptive faculty is in potentiality such as the object of perception already is in actuality" and that when something is affected by an object of perception, "it is made like it and is such as that thing is" (*De Anima* ii 5, 418a3-6).

This hylomorphic restriction on the suitability of subjects of change has the effect of limiting cases of actual perception to those instances of form-reception which involve living beings endowed with the appropriate faculties. It does not, however, explain just what those faculties are, nor even how they are "made like" their objects of perception. Minimally, though, Aristotle claims that for some subject S and some sense object O:

S perceives O if and only if: (i) S has the capacity requisite for receiving O's sensible form; (ii) O acts upon that capacity by enforming it; and, as a result, (iii) S's relevant capacity becomes isomorphic with that form.

Each of these clauses requires unpacking. The plausibility of Aristotle's theory turns on their eventual explications. The first clause (i) is intended to distinguish the active capacities of animals from the merely passive capacities of lifeless material bodies, including the media through which sensible forms travel. (Just as we do not want to say that the tofu in the refrigerator perceives the garlic next to it, we do not want to say that air perceives the color blue when affected by the color of a car.) But it does not yet specify what is required for having the requisite active capacities. Also difficult is the notion of isomorphism appealed to in (iii). As stated, (iii) invites, and has received, scrutiny. Interpretations range from treating the form of isomorphism as direct and literal, so that, e,g., the eyes become speckled when viewing a robin's egg, to attenuated, where the isomorphism is more akin to that enjoyed between a house and its blue-print. Here especially the plausibility of Aristotle's hylomorphic analysis of perception hangs in the balance.

Mind

Aristotle describes mind (*nous*) as "the part of the soul by which it knows and understands" (*De Anima* iii 4, 429a9-10; cf. iii 3, 428a5; iii 9, 432b26; iii 12, 434b3), thus characterizing it in broadly functional terms. It is plain that humans can know and understand things; indeed, Aristotle supposes that it is our very nature to desire knowledge and understanding (*Metaphysics* i 1, 980a21; *De Anima* ii 3, 414b18; iii 3, 429a6-8). In this way, just as the having of sensory faculties is essential to being an animal, so the having of a mind is essential to being a human. Investigating this capacity of soul thus has a special significance for Aristotle: in investigating mind, he is investigating what makes humans human.

His primary investigation of mind occurs in two chapters of *De Anima*, both of which are richly suggestive, but neither of which admits of easy or uncontroversial exposition. In them, *De Anima* iii 4 and 5, Aristotle approaches the nature of thinking by once again deploying a hylomorphic analysis, given in terms of form reception. Just as perception involves the reception of a sensible form by a suitably qualified sensory faculty, so thinking involves the reception of an intelligible form by a suitably qualified intellectual faculty (*De Anima* iii 4, 429a13-18). According to this model, thinking consists in a mind's becoming enformed by some object of thought, so that actual thinking occurs whenever some suitably prepared mind is "made like" its object by being affected by it.

This hylomorphic analysis of thinking is evidently a simple extension of the general model of hylomorphic change exploited by Aristotle in a host of similar contexts. Accordingly, Aristotle's initial account of thinking will directly parallel his analysis of perception (*De Anima* iii 4, 429a13-18). That is, at least in schematic outline, Aristotle will offer the following approach, for any given thinker S and an arbitrary object of thought O:

S thinks O if and only if: (i) S has the capacity requisite for receiving O's intelligible form; (ii) O acts upon that capacity by enforming it; and, as a result, (iii) S's relevant capacity becomes isomorphic with that form.

Unsurprisingly, the same questions which arose in the case of perception also arise here. Most immediately, to understand Aristotle's approach to thinking, it is necessary to determine what it means to say that a thinker's mind and its object become isomorphic.

Here, at least, Aristotle points out what is obvious, that when a thinker's soul is made like its cognitive object, it does not become one with some hylomorphic compound, but with its form: "for it is not the stone which is in the soul, but its form" (*De Anima* iii 8, 431b29-432a1; cf. iii 4, 429a27). The suggestion is, then, that when S comes to think of a stone, as opposed to merely perceiving some particular stone, S has a faculty which is such that it can become one in form with that stone. Aristotle sometimes infers from this sort of consideration that thought is of universals, whereas perception is of particulars (*De Anima* ii 5, 417b23, *Posterior Analytics* i 31, 87b37-88a7), though he elsewhere will allow that we also have knowledge of individuals (*De Anima* ii 5, 417a29; *Metaphysics* xiii 10, 1087a20). These passages are not contradictory, since Aristotle may simply be emphasizing that thought tends to proceed at a higher level of generality than perception, because of its trading in comparatively abstract structural features of its objects. A person can think of what it is to be a stone, but cannot, in any direct and literal

sense of the term, perceive this.

However that may be, Aristotle's conception of thinking implicates him in supposing that thought involves grasping the structural features of the objects of thought. To take an initially favorable case, when thinking *that tree frogs are oviparous*, S will be in a psychic state whose internal structural states are, among other things, one in form with tree frogs. Since S's soul does not become a tree frog when thinking of tree frogs, this form of isomorphism cannot be mere instantiation of the form *being a tree frog*. Rather, S's mind will evidently be one in form with the tree frog, to revert to our earlier analogy, in something like the way a blueprint and the house of which it is the blueprint are one in form. There must be a determinate and expressible structural isomorphism, even though one could not say that the blueprint realizes the form of the house: houses are, after all, necessarily three-dimensional.

For Aristotle, it is not a contingent state of affairs that S's mind does not realize the form being a tree frog in the way that tree frogs themselves do. On the contrary, the mind *cannot* realize a broad range of forms: the mind is, according to Aristotle, not "mixed with the body", insofar as it, unlike the perceptual faculty, lacks a bodily organ (*De Anima* iii 4, 429a24-7). As such, it would not be possible for the mind to realize the form of a house in the way bricks and mortar instantiate such a form: houses provide shelter, something a mind, so understood, cannot do. Consequently, when claiming that minds become isomorphic with their objects, Aristotle must understand the way in which minds become enformed as somehow attenuated or non-literal. Perhaps, though, this should be plain enough. If a mind thinks something by being made like it, then the way it is likened to what it thinks must be somehow representational. Consequently, Aristotle is reasonably understood as holding that S thinks some object of thought O whenever S's mind is made like that object by representing salient structural features of O by being directly isomorphic with them, without, that is, by simply realizing the form of O in the way O does.

This approach to the nature of thinking has some promising features. Both in its own terms, and in virtue of its fitting into a broader pattern of explanation, Aristotle's hylomorphic analysis merits serious consideration. At the same time, one of its virtues may appear also as a vice. We noted in discussing Aristotle's hylomorphic analysis of change generally that his account requires the existence of suitably disposed subjects of change. Only surfaces can be affected so as to be changed in color. An action, such as Socrates' becoming unnerved by a glance of Alcibiades, cannot be made white; it is simply not the appropriate sort of subject. So, hylomorphic change requires at least the following two components: (i) something pre-existing to be the patient of the change, and (ii) that thing's being categorially suited to be changed in the way specified.

Already at the first stage, however, Aristotle's application of this hylomorphic analysis of change to thinking may seem an over-extension. For he maintains directly that mind is "none of the things existing in actuality before thinking" (*De Anima* iii 4, 429a24). His reasons for maintaining this thesis are complex, but derive ultimately from the forms of plasticity Aristotle believes the mind must manifest if it is to be capable of thinking all things (*De Anima* iii 4, 429a18). Now, if the mind is indeed nothing in actuality before thinking, it is hard to understand how the hylomorphic analysis of change and affection could be brought to bear in this arena. If some dough is made cookie-shaped, it is actually dough before

being so enformed; even the sense organs, when made like their objects, are actually existing organs before being affected by the objects of perception. So, given a conception of mind as not existing in actuality before thinking, it is hard to appreciate how thinking lends itself to an analysis in terms of any recognizable hylomorphic approach to change.

How great a problem this will be depends in part upon how entrenched Aristotle's commitment to mind's being nothing in actuality before thinking turns out to be. It equally turns on how adaptable Aristotle's hylomorphic account of change proves. On this latter point, Aristotle notes that according to that account, there are various different types of change and alteration, illustrated by the difference between a brown fence's being painted white and a builder's taking up his tools and beginning to build. In the first case, there is a destruction and a loss, of the fence's original color; in the second case, nothing is destroyed, but rather that which is already dispositionally F becomes occurently F by engaging in some F-ish activity. A builder is as such already able to build. When he begins building he becomes fully and actually a builder for the duration of his working. In this way, he loses nothing, but instead realizes an already established potential.

This second type of change, which Aristotle maintains is the appropriate model for many psychic activities, is either "not an instance of alteration. . .or is a different kind of alteration," where one "should not speak of being affected, unless <one allows that> there are two kinds of alteration" (*De Anima* ii 5, 417b6-16). Perhaps Aristotle's position will then be that the mind, at least insofar as its cognitive capacities for thought are concerned, is simply such as to be enformed by any of an infinite range of objects of thought. This would involve its being nothing determinate in itself; and far from being anomalous for Aristotle, the mind would be in the cognitive realm precisely what the most basic stuff, if there is a most basic stuff, would be in the material realm. Both would manifest unconstrained plasticity; and so each would be characterized essentially in terms of their range of potentialities.

That said, it should be noticed that when it is detached from the idiosyncratic thesis that the mind is nothing in actuality before thinking, Aristotle's hylomorphic analysis of thought retains whatever plausibility it may have independently. For the suggestion that thinking is to be understood at least partially in terms of isomorphisms between our representational capacities and the objects of our cognition has had, for good reason, a durable appeal. To the degree that hylomorphism is generally defensible, then, its application in this domain provides a theoretically rich framework for investigating the nature of thought.

Supplementary Discussion of Mind:

> The Subordinate Psychic Faculties of Imagination and Desire

Bibliography

Translations and Commentaries

- Apostle, Hippocrates (1981), *Aristotle's On the Soul* (Grinell, Iowa: Peripatetic Press)
- Beare, J. I. and Ross, G.R.T. (1908), *The Parva Naturalia* (Oxford: Clarendon Press)
- Hamlyn, D. W. (1993) *Aristotle: De Anima Books II and III* (Oxford: Clarendon Press)
- Hicks, R. D. (1907) *Aristotle: De Anima* (Cambridge: Cambridge University Press)
- Lawson-Tancred, H. (1986), *Aristotle: De Anima* (Harmondsworth: Penguin)
- Rodier, G.1900), Aristote: *Traité de l'âme*(Paris: Leux)
- Ross, W. D. (1955), *Aristotle: Parva Naturalia* (Oxford: Clarendon Press)
- Ross, W.D. (1956), *Aristotle's de Anima* (Oxford: Clarendn Press)
- Sorabji, R. (1972), *Aristotle on Memory* (London: Duckworth)
- Theiler, W. (1979), *Aristoteles: Über die Seele* (Berlin: Akademie Verlag)

Secondary Sources

- Ackrill, J.L., "Aristotle's Definitions of Psuchê," in *Essays on Plato and Aristotle* (Oxford: Clarendon Press), 163-178
- Barnes, J. (1971), "Aristotle's Concept of Mind," *Proceedings of the Aristotelian Society* 75, 101-114
- Brentano, F. (1977), *The Psychology of Aristotle*, trans. R. George (Berkeley: University of California Press)
- Burnyeat, M., "Is an Aristotelian Philosophy of Mind Still Credible?" (1992) in M. Nussbaum and A. Rorty, edd., *Essays on Aristotle's de Anima* (Oxford: Oxford University Press), 15-26
- Everson, Stephen (1995), "*Psychology*," in J. Barnes, ed., *The Cambridge Companion to Aristotle* (Cambridge: Cambridge University Press, 168-194
- Irwin, T.H. (1991), "Aristotle's Philosophy of Mind," in S. Everson, ed., *Companions to Ancient Thought 2: Psychology* (Cambridge: Cambridge University Press)
- Matthews, G.B. (1992), "De Anima 2.2-4 and the Meaning of Life," in M. Nussbaum and A. Rorty, edd., *Essays on Aristotle's de Anima* (Oxford: Oxford University Press), 185-194
- Shields, C. (1988), "Soul and Body in Aristotle," *Oxford Studies in Anceint Philosophy* 6, 103-138
- Sorabji, R. (1971), "Aristotle on Demarcating the Five Senses," *Philosohical Review* 80, 55-79
- Wedin, M. (1988), *Mind and Imagination in Aristotle* (New Haven: Yale University Press)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle

Christopher Shields
shields@colorado.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 11, 2000
Content last modified: January 14, 2000

The Philosophy of Neuroscience

Over the past three decades, philosophy of science has grown increasingly "local." Concerns have switched from general features of scientific practice to concepts, issues, and puzzles specific to particular disciplines. Philosophy of neuroscience is a natural result. This emerging area was also spurred by remarkable recent growth in the neurosciences. Cognitive and computational neuroscience continues to encroach upon issues traditionally addressed within the humanities, including the nature of consciousness, action, knowledge, and normativity. Empirical discoveries about brain structure and function suggest ways that "naturalistic" programs might develop in detail, beyond the abstract philosophical considerations in their favor.

The literature distinguishes "philosophy of neuroscience" and "neurophilosophy." The former concerns foundational issues within the neurosciences. The latter concerns application of neuroscientific concepts to traditional philosophical questions. Exploring various concepts of representation employed in neuroscientific theories is an example of the former. Examining implications of neurological syndromes for the concept of a unified self is an example of the latter. In this entry, we will assume this distinction and discuss examples of both.

Before and After *Neurophilosophy*

Contrary to some opinion, actual neuroscientific discoveries have exerted little influence on the details of materialist philosophies of mind. The "neuroscientific milieu" of the past four decades has made it harder

for philosophers to adopt dualism. But even the "type-type" or "central state" identity theories that rose to brief prominence in the late 1950s (Place, 1956; Smart, 1959) drew upon few actual details of the emerging neurosciences. Recall the favorite early example of a psychoneural identity claim: pain is identical to C-fiber firing. The "C fibers" turned out to be related to only a single aspect of pain transmission (Hardcastle, 1997). Early identity theorists did not emphasize psychoneural identity hypotheses, admitting that their "neuro" terms were placeholders for concepts from future neuroscience. Their arguments and motivations were philosophical, even if the ultimate justification of the program was held to be empirical.

The apology for this lacuna by early identity theorists was that neuroscience at that time was too nascent to provide any plausible identities. But potential identities were afoot. David Hubel and Torsten Wiesel's (1962) electrophysiological demonstrations of the receptive field properties of visual neurons had been reported with great fanfare. Using their techniques, neurophysiologists began discovering neurons throughout visual cortex responsive to increasingly abstract features of visual stimuli: from edges to motion direction to colors to properties of faces and hands. More notably, Donald Hebb had published *The Organization of Behavior* (1949) a decade earlier. Therein he offered detailed explanations of psychological phenomena in terms of known neural mechanisms and anatomical circuits. His psychological explananda included features of perception, learning, memory, and even emotional disorders. He offered these explanations as potential identities. (See the Introduction to his 1949). One philosopher who did take note of some available neuroscientific detail was Barbara Von Eckardt-Klein (1975). She discussed the identity theory with respect to sensations of touch and pressure, and incorporated then-current hypotheses about neural coding of sensation modality, intensity, duration, and location as theorized by Mountcastle, Libet, and Jasper. Yet she was a glaring exception. By and large, available neuroscience at the time was ignored by both philosophical friends and foes of early identity theories.

Philosophical indifference to neuroscientific detail became "principled" with the rise and prominence of functionalism in the 1970s. The functionalists' favorite argument was based on multiple realizability: a given mental state or event can be realized in a wide variety of physical types (Putnam, 1967; Fodor, 1974). So a detailed understanding of one type of realizing physical system (e.g., brains) will not shed light on the fundamental nature of mind. A psychological state-type is autonomous from any single type of its possible realizing physical mechanisms. (See the entry on "Multiple Realizability" in this Encyclopedia, linked below.) Instead of neuroscience, scientifically-minded philosophers influenced by functionalism sought evidence and inspiration from cognitive psychology and "program-writing" artificial intelligence. These disciplines abstract away from underlying physical mechanisms and emphasize the "information-bearing" properties and capacities of representations (Haugeland, 1985). At this same time neuroscience was delving directly into cognition, especially learning and memory. For example, Eric Kandel (1976) proposed presynaptic mechanisms governing transmitter release rate as a cell-biological explanation of simple forms of associative learning. With Robert Hawkins (1984) he demonstrated how cognitivist aspects of associative learning (e.g., blocking, second-order conditioning, overshadowing) could be explained cell-biologically by sequences and combinations of these basic forms implemented in higher neural anatomies. Working on the post-synaptic side, neuroscientists began unraveling the cellular mechanisms of long term potentiation (LTP) (Bliss and Lomo, 1973). Physiological psychologists quickly

noted its explanatory potential for various forms of learning and memory.[1] Yet few "materialist" philosophers paid any attention. Why should they? Most were convinced functionalists. They believed that the "engineering level" details might be important to the clinician, but were irrelevant to the theorist of mind.

A major turning point in philosophers' interest in neuroscience came with the publication of Patricia Churchland's *Neurophilosophy* (1986). The Churchlands (Pat and husband Paul) were already notorious for advocating eliminative materialism (see the next section). In her (1986) book, Churchland distilled eliminativist arguments of the past decade, unified the pieces of the philosophy of science underlying them, and sandwiched the philosophy between a five-chapter introduction to neuroscience and a 70-page chapter on three then-current theories of brain function. She was unapologetic about her intent. She was introducing philosophy of science to neuroscientists and neuroscience to philosophers. Nothing could be more obvious, she insisted, than the relevance of empirical facts about how the brain works to concerns in the philosophy of mind. Her term for this interdisciplinary method was "co-evolution" (borrowed from biology). This method seeks resources and ideas from anywhere on the theory hierarchy above or below the question at issue. Standing on the shoulders of philosophers like Quine and Sellars, Churchland insisted that specifying some point where neuroscience ends and philosophy of science begins is hopeless because the boundaries are poorly defined. Neurophilosophers would pick and choose resources from both disciplines as they saw fit.

Three themes predominate Churchland's philosophical discussion: developing an alternative to the logical empiricist theory of intertheoretic reduction; responding to property-dualistic arguments based on subjectivity and sensory qualia; and responding to anti-reductionist multiple realizability arguments. These projects have remained central to neurophilosophy over the past decade. John Bickle (1998) extends the principal insight of Clifford Hooker's (1981) post-empiricist theory of intertheoretic reduction. He quantifies key notions using a model-theoretic account of theory structure adapted from the structuralist program in philosophy of science (Balzer, Moulines, and Sneed, 1987). He also makes explicit the form of argument scientists employ to draw ontological conclusions (cross-theoretic identities, revisions, or eliminations) based on the nature of the intertheoretic reduction relations obtaining in soecific cases. For example, physicists concluded that visible light, a theoretical posit of optics, is electromagnetic radiation within specified wavelengths, a theoretical posit of electromagnetism: a cross-theoretic ontological identity. In another case, however, chemists concluded that phlogiston did not exist: an elimination of a kind from our scientific ontology. Bickle explicates the nature of the reduction relation in a specific case using a semi-formal account of ‘intertheoretic approximation’ inspired by structuralist results. Paul Churchland (1996) has carried on the attack on property-dualistic arguments for the irreducibility of conscious experience and sensory qualia. He argues that acquiring some knowledge of existing sensory neuroscience increases one's ability to ‘imagine’ or ‘conceive of’ a comprehensive neurobiological explanation of consciousness. He defends this conclusion using a thought-experiment based on the history of optics and electromagnetism. Finally, the literature critical of the multiple realizability argument has begun to flourish. Although the multiple realizability argument remains influential among nonreductive physicalists, it no longer commands the universal acceptance it once did. Replies to the multiple realizability argument based on neuroscientific details have appeared. For example, William Bechtel and Jennifer Mundale (1997, in press) argue that neuroscientists use

psychological criteria in brain mapping studies. This fact undercuts the likelihood that psychological kinds are multiply realized. (For a review of recent developments see the final sections of the entry on 'Multiple Realizability' in this Encyclopedia, linked below.)

Eliminative Materialism and Philosophy Neuralized

Eliminative materialism (EM) is the conjunction of two claims. First, our common sense 'belief-desire' conception of mental events and processes, our 'folk psychology,' is a false and misleading account of the causes of human behavior. Second, like other false conceptual frameworks from both folk theory and the history of science, it will be replaced by, rather than smoothly reduced or incorporated into, a future neuroscience. Folk psychology is the collection of common homilies about the causes of human behavior. You ask me why Marica is not accompanying me this evening. I reply that her grant deadline is looming. You nod sympathetically. You understand my explanation because you share with me a generalization that relates beliefs about looming deadlines, desires about meeting professionally and financially significant ones, and ensuing free-time behavior. It is the collection of these kinds of homilies that EM claims to be flawed beyond significant revision. Although this example involves only beliefs and desires, folk psychology contains an extensive repertoire of propositional attitudes in its explanatory nexus: hopes, intentions, fears, imaginings, and more. To the extent that scientific psychology (and neuroscience!) retains folk concepts, EM applies to it as well.

EM is physicalist in the classical sense, postulating some future brain science as the ultimately correct account of (human) behavior. It is eliminative in predicting the future removal of folk psychological kinds from our post-neuroscientific ontology. EM proponents often employ scientific analogies (Feyerabend 1963; Churchland, 1981). Oxidative reactions as characterized within elemental chemistry bear no resemblance to phlogiston release. Even the "direction" of the two processes differ. Oxygen is gained when an object burns (or rusts), phlogiston was said to be lost. The result of this theoretical change was the elimination of phlogiston from our scientific ontology. There is no such thing. For the same reasons, according to EM, continuing development in neuroscience will reveal that there are no such things as beliefs and desires as characterized by common sense.

Here we focus only on the way that neuroscientific results have shaped the arguments for EM. Surprisingly, only one argument has been strongly influenced. (Most arguments for EM stress the failures of folk psychology as an explanatory theory of behavior.) This argument is based on a development in cognitive and computational neuroscience that might provide a genuine alternative to the representations and computations implicit in folk psychological generalizations. Many eliminative materialists assume that folk psychology is committed to propositional representations and computations over their contents that mimic logical inferences (Paul Churchland, 1981; Stich, 1983; Patricia Churchland, 1986).[2] Even though discovering such an alternative has been an eliminativist goal for some time, neuroscience only began delivering on this goal over the past fifteen years. Points in and trajectories through vector spaces, as an interpretation of synaptic events and neural activity patterns in biological neural networks are key features of this new development. This argument for EM hinges on the differences between these notions of cognitive representation and the propositional attitudes of folk psychology (Churchland, 1987).

However, this argument will be opaque to those with no background in contemporary cognitive and computational neuroscience, so we need to present a few scientific details. With these details in place, we will return to this argument for EM (five paragraphs below).

At one level of analysis the basic computational element of a neural network (biological or artificial) is the neuron. This analysis treats neurons as simple computational devices, transforming inputs into output. Both neuronal inputs and outputs reflect biological variables. For the remainder of this discussion, we will assume that neuronal inputs are frequencies of action potentials (neuronal "spikes") in the axons whose terminal branches synapse onto the neuron in question. Neuronal output is the frequency of action potentials in the axon of the neuron in question. A neuron computes its total input (usually treated mathematically as the sum of the products of the signal strength along each input line times the synaptic weight on that line). It then computes a new activation state based on its total input and current activation state, and a new output state based on its new activation value. The neuron's output state is transmitted as a signal strength to whatever neurons its axon synapses on. The output state reflects systematically the neuron's new activation state.[3]

Analyzed at this level, both biological and artificial neural networks are interpreted naturally as *vector-to-vector transformers*. The input vector consists of values reflecting activity patterns in axons synapsing on the network's neurons from outside (e.g., from sensory transducers or other neural networks). The output vector consists of values reflecting the activity patterns generated in the network's neurons that project beyond the net (e.g., to motor effectors or other neural networks). Given that neurons' activity depends partly upon their total input, and total input depends partly on synaptic weights (e.g., presynaptic neurotransmitter release rate, number and efficacy of postsynaptic receptors, availability of enzymes in synaptic cleft), the capacity of biological networks to change their synaptic weights make them *plastic* vector-to-vector transformers. In principle, a biological network with plastic synapses can come to implement any vector-to-vector transformation that its composition permits (number of input units, output units, processing layers, recurrency, cross-connections, etc.) (Churchland, 1987).

The anatomical organization of the cerebellum provides a clear example of a network amendable to this computational interpretation. Consider Figure 1. The cerebellum is the bulbous convoluted structure dorsal to the brainstem. A variety of studies (behavioral, neuropsychological, single-cell electrophysiological) implicate this structure in motor integration and fine motor coordination. Mossy fibers (axons) from neurons outside the cerebellum synapse on cerebellular granule cells, which in turn project to parallel fibers. Activity patterns across the collection of mossy fibers (frequency of action potentials per time unit in each fiber projecting into the cerebellum) provide values for the input vector. Parallel fibers make multiple synapses on the dendritic trees and cell bodies of cerebellular Purkinje neurons. Each Purkinje neuron "sums" its post-synaptic potentials (PSPs) and emits a train of action potentials down its axon based (partly) on its total input and previous activation state. Purkinje axons project outside the cerebellum. The network's output vector is thus the ordered values representing the pattern of activity generated in each Purkinje axon. Changes to the efficacy of individual synapses on the parallel fibers and the Purkinje neurons alter the resulting PSPs in Purkinje axons, generating different axonal spiking frequencies. Computationally, this amounts to a different output vector to the same input activity pattern (plasticity).[4]

This interpretation puts the useful mathematical resources of *dynamical systems* into the hands of computational neuroscientists. *Vector spaces* are an example. For example, learning can be characterized fruitfully in terms of changes in synaptic weights in the network and subsequent reduction of error in network output. (This approach goes back to Hebb, 1949, although within the vector-space interpretation that follows.) A useful representation of this account is on a *synaptic weight-error space*, where one dimension represents the global error in the network's output to a given task, and all other dimensions represent the weight values of individual synapses in the network. Consider Figure 2. Points in this multi-dimensional state space represent the global performance error correlated with each possible collection of synaptic weights in the network. As the weights change with each performance (in accordance with a biologically-implemented learning algorithm), the global error of network performance continually decreases. Learning is represented as synaptic weight changes correlated with a descent along the error dimension in the space (Churchland and Sejnowski, 1992). Representations (concepts) can be portrayed as *partitions* in multi-dimensional vector spaces. An example is a *neuron activation* vector space. See Figure 3. A graph of such a space contains one dimension for the activation value of each neuron in the network (or some subset). A point in this space represents one possible pattern of activity in all neurons in the network. Activity patterns generated by input vectors that the network has learned to group together will cluster around a (hyper-) point or subvolume in the activity vector space. Any input pattern sufficiently similar to this group will produce an activity pattern lying in geometrical proximity to this point or subvolume. Paul Churchland (1989) has argued that this interpretation of network activity provides a quantitative, neurally-inspired basis for prototype theories of concepts developed recently in cognitive psychology.

Using this theoretical development, Paul Churchland (1987, 1989) has offered a novel argument for EM. According to this approach, activity vectors are the central kind of representation and vector-to-vector transformations are the central kind of computation in the brain. This contrasts sharply with the *propositional* representations and *logical/semantic* computations postulated by folk psychology. Vectorial content is unfamiliar and alien to common sense. This cross-theoretic difference is at least as great as that between oxidative and phlogiston concepts, or kinetic-corpuscular and caloric fluid heat concepts. Phlogiston and caloric fluid are two "parade" examples of kinds eliminated from our scientific ontology due to the nature of the intertheoretic relation obtaining between the theories with which they are affiliated and the theories that replaced these. The structural and dynamic differences between the folk psychological and emerging cognitive neuroscientific kinds suggest that the theories affiliated with the latter will also correct significantly the theory affiliated with the former. This is the key premise of an eliminativist argument based on predicted intertheoretic relations. And these intertheoretic contrasts are no longer just an eliminativist's goal. Computational and cognitive neuroscience has begun to deliver an alternative kinematics for cognition, one that provides no structural analogue for the propositional attitudes.

Certainly the replacement of propositional contents by vectorial alternatives implies significant correction to folk psychology. But does it justify EM? Even though this central feature of folk-psychological posits finds no analogues in one hot theoretical development in recent cognitive and computational neuroscience, there might be other aspects of cognition that folk psychology gets right. Within

neurophilosophy, concluding that a cross-theoretic identity claim is true (e.g., folk psychological state F is identical to neural state N) or that an eliminativist claim is true (there is no such thing as folk psychological state F) depends on the nature of the intertheoretic reduction obtaining between the theories affiliated with the posits in question (Hooker, 1981; Churchland, 1986; Bickle, 1998). But the underlying account of intertheoretic reduction recognizes a spectrum of possible reductions, ranging from relatively "smooth" through "significantly revisionary" to "extremely bumpy".[5] Might the reduction of folk psychology and a "vectorial" neurobiology occupy the middle ground between "smooth" and "bumpy" intertheoretic reductions, and hence suggest a "revisionary" conclusion? The reduction of classical equilibrium thermodynamics to statistical mechanics to microphysics provides a potential analogy. John Bickle (1992, 1998, chapter 6) argues on empirical grounds that such a outcome is likely. He specifies conditions on "revisionary" reductions from historical examples and suggests that these conditions are obtaining between folk psychology and cognitive neuroscience as the latter develops. In particular, folk psychology appears to have gotten right the grossly-specified functional profile of many cognitive states, especially those closely related to sensory input and behavioral output. It also appears to get right the "intentionality" of many cognitive states--the object that the state is of or about--even though cognitive neuroscience eschews its implicit linguistic explanation of this feature. Revisionary physicalism predicts significant *conceptual change* to folk psychological concepts, but denies total elimination of the caloric fluid-phlogiston variety.

The philosophy of science is another area where vector space interpretations of neural network activity patterns has impacted philosophy. In the Introduction to his (1989) book, Paul Churchland asserts that it will soon be impossible to do serious work in the philosophy of science without drawing on empirical work in the brain and behavioral sciences. To justify this claim, in Part II of the book he suggests neurocomputational reformulations of key concepts from this area. At the heart is a neurocomputational account of the structure of scientific theories (1989, chapter 9). Problems with the orthodox "sets-of-sentences" view have been well-known for over three decades. Churchland advocates replacing the orthodox view with one inspired by the "vectorial" interpretation of neural network activity. Representations implemented in neural networks (as discussed above) compose a system that corresponds to important distinctions in the external environment, are not explicitly represented as such within the input corpus, and allow the trained network to respond to inputs in a fashion that continually reduces error. These are exactly the functions of theories. Churchland is bold in his assertion: an individual's theory-of-the-world is a specific point in that individual's error-synaptic weight vector space. It is a configuration of synaptic weights that partitions the individual's activation vector space into subdivisions that reduce future error messages to both familiar and novel inputs. (Consider again Figure 2 and Figure 3.) This reformulation invites an objection, however. Churchland boasts that his theory of theories is preferable to existing alternatives to the orthodox "sets-of-sentences" account--for example, the *semantic* view (Suppe, 1974; van Fraassen, 1980)--because his is closer to the "buzzing brains" that use theories. But as Bickle (1993) notes, neurocomputational models based on the mathematical resources described above are a long way into the realm of abstractia. Even now, they remain little more than novel (and suggestive) applications of the mathematics of quasi-linear dynamical systems to simplified schemata of brain circuitries. Neurophilosophers owe some account of identifications across ontological categories before the philosophy of science community will accept the claim that theories are points in high-dimensional state spaces implemented in biological neural networks. (There is an important

methodological assumption lurking in this objection, however, which we will discuss toward the end of the next paragraph.)

Churchland's neurocomputational reformulations of scientific and epistemological concepts build on this account of theories. He sketches "neuralized" accounts of the theory-ladenness of perception, the nature of concept unification, the virtues of theoretical simplicity, the nature of Kuhnian paradigms, the kinematics of conceptual change, the character of abduction, the nature of explanation, and even moral knowledge and epistemological normativity. Conceptual redeployment, for example, is the activation of an already-existing prototype representation--the centerpoint or region of a partition of a high-dimensional vector space in a trained neural network--to a novel type of input pattern. Obviously, we can't here do justice to Churchland's many and varied attempts at reformulation. We urge the intrigued reader to examine his suggestions in their original form. But a word about philosophical methodology is in order. Churchland is *not* attempting "conceptual analysis" in anything resembling its traditional philosophical sense and neither, typically, are neurophilosophers.(This is why a discussion of neurophilosophical reformulations fits with a discussion of EM.) There are philosophers who take the discipline's ideal to be a relatively simple set of necessary and sufficient conditions, expressed in non-technical natural language, governing the application of important concepts (like justice, knowledge, theory, or explanation). These analyses should square, to the extent possible, with pretheoretical usage. Ideally, they should preserve synonymy. Other philosophers view this ideal as sterile, misguided, and perhaps deeply mistaken about the underlying structure of human knowledge (Ramsey, 1992). Neurophilosophers tend to reside in the latter camp. Those who dislike philosophical speculation about the promise and potential of nascent science in an effort to reformulate ("*reform*-ulate") traditional philosophical concepts have probably already discovered that neurophilosophy is not for them. But the charge that neurocomputational reformulations of the sort Churchland attempts are "philosophically uninteresting" or "irrelevant" because they fail to provide "adequate analyses" of theory, explanation, and the like will fall on deaf ears among many contemporary philosophers, as well as their cognitive-scientific and neuroscientific friends.

Before we leave the neurophilosophical applications of this theoretical development from recent cognitive/computational neuroscience, one more point of scientific detail is in order. The popularity of treating the neuron as the basic computational unit among *neural* modelers, as opposed to cognitive modelers, is declining rapidly. *Compartmental modeling* enables computational neuroscientists to mimic activity in and interactions between patches of neuronal membrane (Bower and Beeman, 1995). This permits modelers to control and manipulate a variety of subcellular factors that determine action potentials per time unit (including the topology of membrane structure in individual neurons, variations in ion channels across membrane patches, field properties of post-synaptic potentials depending on the location of the synapse on the dendrite or soma). Modelers can "custom build" the neurons in their target circuitry without sacrificing the ability to study circuit properties of networks. For these reasons, few serious computational *neuro*scientists continue to work at a level that treats neurons as unstructured computational devices. But the above interpretative points still stand. With compartmental modeling, not only are simulated neural networks interpretable as vector-to-vector transformers. The neurons composing them are, too.

Philosophy of science and scientific epistemology are not the only areas where philosophers have lately urged the relevance of neuroscientific discoveries. Kathleen Akins (1996) argues that a "traditional" view of the senses underlies the variety of sophisticated "naturalistic" programs about intentionality. (She cites the Churchlands, Daniel Dennett, Fred Dretske, Jerry Fodor, David Papineau, Dennis Stampe, and Kim Sterelny as examples, with extensive references.) Current neuroscientific understanding of the mechanisms and coding strategies implemented by sensory receptors shows that this traditional view is mistaken. The traditional view holds that sensory systems are "veridical" in at least three ways. (1) Each signal in the system correlates with a small range of properties in the external (to the body) environment. (2) The structure in the relevant relations between the external properties the receptors are senstive to is preserved in the structure of the relations between the resulting sensory states. And (3) the sensory system reconstructs faithfully, without fictive additions or embellishments, the external events. Using recent neurobiological discoveries about response properties of thermal receptors in the skin as an illustration, Akins shows that sensory systems are "narcissistic" rather than "veridical." All three traditional assumptions are violated. These neurobiological details and their philosophical implications open novel questions for the philosophy of perception and for the appropriate foundations for naturalistic projects about intentionality. Armed with the known neurophysiology of sensory receptors, for example, our "philosophy of perception" or of "perceptual intentionality" will no longer focus on the search for correlations between states of sensory systems and "veridically detected" external properties. This traditional philosophical (and scientific!) project rests upon a mistaken "veridical" view of the senses. Neuroscientific knowledge of sensory receptor activity also shows that sensory experience does not serve the naturalist well as a "simple paradigm case" of an intentional relation between representation and world. Once again, available scientific detail shows the naivity of some traditional philosophical projects.

Focusing on the anatomy and physiology of the pain transmission system, Valerie Hardcastle (1997) urges a similar negative implication for a popular methodological assumption. Pain experiences have long been philosophers' favorite cases for analysis and theorizing about conscious experience generally. Nevertheless, every position about pain experiences has been defended recently: eliminativism, a variety of objectivist views, relational views, and subjectivist views. Why so little agreement, despite agreement that pain experiences are the place to start an analysis or theory of consciousness? Hardcastle urges two answers. First, philosophers tend to be uninformed about the neuronal complexity of our pain transmission systems, and build their analyses or theories on the outcome of a single component of a multi-component system. Second, even those who understand some of the underlying neurobiology of pain tend to advocate gate-control theories.[6] But the best existing gate-control theories are vague about the neural mechanisms of the gates. Hardcastle instead proposes a dissociable dual system of pain transmission, consisting of a pain sensory system closely analogous in its neurobiological implementation to other sensory systems, and a descending pain inhibitory system. She argues that this dual system is consistent with recent neuroscientific discoveries and accounts for all the pain phenomena that have tempted philosophers toward particular (but limited) theories of pain experience. The neurobiological uniqueness of the pain inhibitory system, contrasted with the mechanisms of other sensory modalities, renders pain processing atypical. In particular, the pain inhibitory system dissociates pain sensation from stimulation of nociceptors (pain receptors). Hardcastle concludes from the neurobiological uniqueness of pain transmission that pain experiences are atypical conscious events, and hence not a good place to start theorizing about or analyzing the general type.

Neuroscience and Psychosemantics

Developing and defending theories of content is a central topic in current philosophy of mind. A common desideratum in this debate is a theory of cognitive representation consistent with a physical or naturalistic ontology. We'll here describe a few contributions neurophilosophers have made to this literature.

When one perceives or remembers that he is out of coffee, his brain state possesses intentionality or "aboutness." The percept or memory is about one's being out of coffee; it represents one as being out of coffee. The representational state has content. A psychosemantics seeks to explain what it is for a representational state to be about something: to provide an account of how states and events can have specific representational content. A physicalist psychosemantics seeks to do this using resources of the physical sciences exclusively. Neurophilosophers have contributed to two types of physicalist psychosemantics: the Functional Role approach and the Informational approach.

The core claim of a functional role semantics holds that a representation has its content in virtue of relations it bears to other representations. Its paradigm application is to concepts of truth-functional logic, like the conjunctive ‘and’ or disjunctive ‘or.’ A physical event instantiates the ‘and’ function just in case it maps two true inputs onto a single true output. Thus it is the relations an expression bears to others that give it the semantic content of ‘and.’ Proponents of functional role semantics propose similar analyses for the content of all representations (Block 1986). A physical event represents birds, for example, if it bears the right relations to events representing feathers and others representing beaks. By contrast, informational semantics ascribe content to a state depending upon the causal relations obtaining between the state and the object it represents. A physical state represents birds, for example, just in case an appropriate causal relation obtains between it and birds. At the heart of informational semantics is a causal account of information (Dretske, 1981, 1988). Red spots on a face carry the information that one has measles because the red spots are caused by the measles virus. A common criticism of informational semantics holds that mere causal covariation is insufficient for representation, since information (in the causal sense) is by definition always veridical while representations can misrepresent. A popular solution to this challenge invokes a teleological analysis of ‘function.’ A brain state represents X by virtue of having the function of carrying information about being caused by X (Dretske 1988). These two approaches do not exhaust the popular options for a psychosemantics, but are the ones to which neurophilosophers have contributed.

Paul Churchland's allegiance to functional role semantics goes back to his earliest views about the semantics of terms in a language. In his (1979) book, he insists that the semantic identity (content) of a term derives from its place in the network of sentences of the entire language. The functional economies envisioned by early functional role semanticists were networks with nodes corresponding to the objects and properties denoted by expressions in a language. Thus one node, appropriately connected, might represent birds, another feathers, and another beaks. Activation of one of these would tend to spread to the others. As ‘connectionist’ network modeling developed, alternatives arose to this one-representation-per-node ‘localist’ approach. By the time Churchland (1989) provided a neuroscientific elaboration of functional role semantics for cognitive representations generally, he too had abandoned the ‘localist’

interpretation. Instead, he offered a 'state-space semantics'.

We saw in the section just above how (vector) state spaces provide a natural interpretation for activity patterns in neural networks (biological and artificial). A state-space semantics for cognitive representations is a species of a functional role semantics because the individuation of a particular state depends upon the relations obtaining between it and other states. A representation is a point in an appropriate state space, and points (or subvolumes) in a space are individuated by their relations to other points (locations, geometrical proximity). Churchland (1989, 1995) illustrates a state-space semantics for neural states by appealing to sensory systems. One popular theory in sensory neuroscience of how the brain codes for sensory qualities (like color) is the *opponent process account* (Hardin 1988). Churchland (1995) describes a three-dimensional activation vector state-space in which every color perceivable by humans is represented as a point (or subvolume). Each dimension corresponds to activity rates in one of three classes of photoreceptors present in the human retina and their efferent paths: the red-green opponent pathway, yellow-blue opponent pathway, and black-white (contrast) opponent pathway. Photons striking the retina are transduced by the receptors, producing an activity rate in each of the segregated pathways. A represented color is hence a triplet of activation frequency rates. As an illustration, consider again Figure 3. Each dimension in that three-dimensional space will represent average frequency of action potentials in the axons of one class of ganglion cells projecting out of the retina. Each color perceivable by humans will be a region of that space. For example, an orange stimulus produces a relatively low level of activity in both the red-green and yellow-blue opponent pathways (x-axis and y-axis, respectively), and middle-range activity in the black-white (contrast) opponent pathway (z-axis). Pink stimuli, on the other hand, produce low activity in the red-green opponent pathway, middle-range activity in the yellow-blue opponent pathway, and high activity in the black-white (contrast) opponent pathway.[7] The location of each color in the space generates a 'color solid.' Location on the solid and geometrical proximity between regions reflect structural similarities between the perceived colors. Human gustatory representations are points in a four-dimensional state space, with each dimension coding for activity rates generated by gustatory stimuli in each type of taste receptor (sweet, salty, sour, bitter) and their segregated efferent pathways. When implemented in a neural network with structural and hence computational resources as vast as the human brain, the state space approach to psychosemantics generates a theory of content for a huge number of cognitive states.[8]

Jerry Fodor and Ernest LePore (1992) raise an important challenge to Churchland's psychosemantics. Location in a state space alone seems insufficient to fix a state's representational content. Churchland never explains why a point in a three-dimensional state space represents a *color*, as opposed to any other quality, object, or event that varies along three dimensions.[9]. Churchland's account achieves its explanatory power by the interpretation imposed on the dimensions. Fodor and LePore allege that Churchland never specifies how a dimension comes to represent, e.g., degree of saltiness, as opposed to yellow-blue wavelength opposition. One obvious answer appeals to the stimuli that form the 'external' inputs to the neural network in question. Then, for example, the individuating conditions on neural representations of colors are that opponent processing neurons receive input from a specific class of photoreceptors. The latter in turn have electromagnetic radiation (of a specific portion of the visible spectrum) as their activating stimuli. However, this appeal to 'external' stimuli as the ultimate individuating conditions for representational content makes the resulting approach a version of

informational semantics. Is this approach consonant with other neurobiological details?

The neurobiological paradigm for informational semantics is the *feature detector*: one or more neurons that are (i) maximally responsive to a particular type of stimulus, and (ii) have the function of indicating the presence of that stimulus type. Examples of such stimulus-types for visual feature detectors include high-contrast edges, motion direction, and colors. A favorite feature detector among philosophers is the alleged fly detector in the frog. Lettvin *et al*. (1959) identified cells in the frog retina that responded maximally to small shapes moving across the visual field. The idea that these cells' activity functioned to detect flies rested upon knowledge of the frogs' diet. (Bechtel 1998 provides a useful discussion.) Using experimental techniques ranging from single-cell recording to sophisticated functional imaging, neuroscientists have recently discovered a host of neurons that are maximally responsive to a variety of stimuli. However, establishing condition (ii) on a feature detector is much more difficult. Even some paradigm examples have been called into question. David Hubel and Torsten Wiesel's (1962) Nobel Prize-winning work establishing the receptive fields of neurons in striate cortex is often interpreted as revealing cells whose function is edge detection. However, Lehky and Sejnowski (1988) have challenged this interpretation. They trained an artificial neural network to distinguish the three-dimensional shape and orientation of an object from its two-dimensional shading pattern. Their network incorporates many features of visual neurophysiology. Nodes in the trained network turned out to be maximally responsive to edge contrasts, but did not appear to have the function of edge detection. (See Churchland and Sejnowski 1992 for a review.)

Kathleen Akins (1996) offers a different neurophilosophical challenge to informational semantics and its affiliated feature-detection view of sensory representation. We saw in the previous section how Akins argues that the physiology of thermoreception violates three necessary conditions on ‘veridical’ representation. From this fact she draws doubts about looking for feature detecting neurons to ground a psychosemantics generally, including thought contents. Human thoughts about flies, for example, are sensitive to numerical distinctions between particular flies and the particular locations they can occupy. But the ends of frog nutrition are well served without a representational system sensitive to such ontological refinements. Whether a fly seen now is numerically identical to one seen a moment ago need not, and perhaps cannot, figure into the frog's feature detection repertoire. Akins' critique casts doubt on whether details of sensory transduction will scale up to provide an adequate unified psychosemantics. It also raises new questions for human intentionality. How do we get from activity patterns in "narcissistic" sensory receptors, keyed not to "objective" environmental features but rather only to effects of the stimuli on the patch of tissue innervated, to the human ontology replete with enduring objects with stable configurations of properties and relations, types and their tokens (as the "fly-thought" example presented above reveals), and the rest? And how did the development of a stable, rich ontology confer survival advantages to human ancestors?

Consciousness Explained?

Consciousness has re-emerged as a topic in philosophy of mind and the cognitive and brain sciences over the past three decades. Instead of ignoring it, many physicalists now seek to explain it (Dennett, 1991).

Here we focus exclusively on ways that neuroscientific discoveries have impacted philosophical debates about the nature of consciousness and its relation to physical mechanisms. (See links to other entries in this encyclopedia below for broader discussions about consciousness and physicalism.)

Thomas Nagel (1974) argues that conscious experience is subjective, and thus permanently recalcitrant to objective scientific understanding. He invites us to ponder 'what it is like to be a bat' and urges the intuition that no amount of physical-scientific knowledge (including neuroscientific) supplies a complete answer. Nagel's intuition pump has generated extensive philosophical discussion. At least two well-known replies make direct appeal to neurophysiology. John Biro (1991) suggests that part of the intuition pumped by Nagel, that bat experience is substantially different from human experience, presupposes systematic relations between physiology and phenomenology. Kathleen Akins (1993a) delves deeper into existing knowledge of bat physiology and reports much that is pertinent to Nagel's question. She argues that many of the questions about bat subjectivity that we still consider open hinge on questions that remain unanswered about neuroscientific details. One example of the latter is the function of various cortical activity profiles in the active bat.

More recently philosopher David Chalmers (1996) has argued that any possible brain-process account of consciousness will leave open an 'explanatory gap' between the brain process and properties of the conscious experience.[10] This is because no brain-process theory can answer the "hard" question: Why should that particular brain process give rise to conscious experience? We can always imagine ("conceive of") a universe populated by creatures having those brain processes but completely lacking conscious experience. A theory of consciousness requires an explanation of how and why some brain process causes consciousness replete with all the features we commonly experience. The fact that the hard question remains unanswered shows that we will probably never get a complete explanation of consciousness at the level of neural mechanism. Paul and Patricia Churchland (1997) have recently offered the following diagnosis and reply. Chalmers offers a *conceptual argument*, based on our ability to imagine creatures possessing brains like ours but wholly lacking in conscious experience. But the more one learns about how the brain produces conscious experience--and a literature is beginning to emerge (e.g., Gazzaniga, 1995)--the harder it becomes to imagine a universe consisting of creatures with brain processes like ours but lacking consciousness. This is not just bare assertion. The Churchlands appeal to some neurobiological detail. For example, Paul Churchland (1995) develops a neuroscientific account of consciousness based on recurrent connections between thalamic nuclei (particularly "diffusely projecting" nuclei like the intralaminar nuclei) and cortex.[11] Churchland argues that the thalamocortical recurrency accounts for the selective features of consciousness, for the effects of short-term memory on conscious experience, for vivid dreaming during REM (rapid-eye movement) sleep, and other "core" features of conscious experience. In other words, the Churchlands are claiming that when one learns about activity patterns in these recurrent circuits, one can't "imagine" or "conceive of" this activity occurring without these core features of conscious experience. (Other than just mouthing the words, "I am now imagining activity in these circuits without selective attention/the effects of short-term memory/vivid dreaming/...").

A second focus of skeptical arguments about a complete neuroscientific explanation of consciousness is sensory *qualia*: the introspectable qualitative aspects of sensory experience, the features by which subjects discern similarities and differences among their experiences. The colors of visual sensations are a

philosopher's favorite example. One famous puzzle about color qualia is the alleged conceivability of spectral inversions. Many philosophers claim that it is conceptually possible (if perhaps physically impossible) for two humans not to differ neurophysiologically, while the color that fire engines and tomatoes appear to have to one subject is the color that grass and frogs appear to have to the other (and vice versa). A large amount of neuroscientifically-informed philosophy has addressed this question. (C.L. Hardin 1988 and Austen Clark 1993 are noteworthy examples.) A related area where neurophilosophical considerations have emerged concerns the metaphysics of colors themselves (rather than color experiences). A longstanding philosophical dispute is whether colors are objective properties existing external to perceivers or rather identifiable as or dependent upon minds or nervous systems. Some recent work on this problem begins with characteristics of color experiences: for example, that color similarity judgments produce color orderings that align on a circle (Clark 1993). With this resource, one can seek mappings of phenomenology onto environmental or physiological regularities. Identifying colors with particular frequencies of electromagnetic radiation does not preserve the structure of the hue circle, whereas identifying colors with activity in opponent processing neurons does. Such a tidbit is not decisive for the color objectivist-subjectivist debate, but it does convey the type of neurophilosophical work being done on traditional metaphysical issues beyond the philosophy of mind. (For more details on these issues, see the entry on Color in this Encyclopedia, linked below.)

We saw in the discussion of Hardcastle (1997) two sections above that neurophilosophers have entered disputes about the nature and methodological import of pain experiences. Two decades earlier, Dan Dennett (1978) took up the question of whether it is possible to build a computer that feels pain. He compares and notes tension between neurophysiological discoveries and common sense intuitions about pain experience. He suspects that the incommensurability between scientific and common sense views is due to incoherence in the latter. His attitude is wait-and-see. But foreshadowing Churchland's reply to Chalmers, Dennett favors scientific investigations over conceivability-based philosophical arguments.

Neurological deficits have attracted philosophical interest. For thirty years philosophers have found implications for the unity of the self in experiments with commissurotomy patients (Nagel 1971).[12] In carefully controlled experiments, commissurotomy patients display two dissociable seats of consciousness. In chapter 5 of her (1986) book, Patricia Churchland scouts philosophical implications of a variety of neurological deficits. One deficit is blindsight. Some patients with lesions to primary visual cortex report being unable to see items in regions of their visual fields, yet perform far better than chance in forced guess trials about stimuli in those regions. A variety of scientific and philosophical interpretations have been offered. Ned Block (1988) worries that many of these conflate distinct notions of consciousness. He labels these notions ‘phenomenal consciousness’ (‘P-consciousness’) and ‘access consciousness’ (‘A-consciousness’). The former is the ‘what it is like’-ness of experience. The latter is the availability of representational content to self-initiated action and speech. Block argues that P-consciousness is not always representational whereas A-consciousness is. Dennett (1991, 1995) and Michael Tye (1993) are skeptical of non-representational analyses of consciousness in general. They provide accounts of blindsight that do not depend on Block's distinction.

We break off our brief overview of neurophilosophical work on consciousness here. Many other topics are worth neurophilosophical pursuit. We mentioned commissurotomy and the unity of consciousness and

the self, which continues to generate discussion. Qualia beyond those of color and pain have begun to attract neurophilosophical attention (Akins 1993a, 1993b, 1996; Clark 1993), as has self-consciousness (Bermudez 1998).

Location of Cognitive Function: From Lesion Studies to Recent Neuroimaging

One of the first issues to arise in the 'philosophy of neuroscience' (before there was a recognized area) was the localization of cognitive functions to specific neural regions. Although the 'localization' approach had dubious origins in the phrenology of Gall and Spurzheim, and was challenged severely by Flourens throughout the early nineteenth century, it re-emerged in the study of aphasia by Bouillaud, Auburtin, Broca, and Wernicke. These neurologists made careful studies (where possible) of linguistic deficits in their aphasic patients followed by brain autopsies post mortem.[13] Broca's initial study of twenty-two patients in the mid-nineteenth century confirmed that damage to the *left cortical hemisphere* was predominant, and that damage to the second and third frontal convolutions was necessary to produce speech production deficits. Although the anatomical coordinates Broca postulated for the 'speech production center'' do not correlate exactly with damage producing production deficits, both this area of frontal cortex and speech production deficits still bear his name ('Broca's area' and 'Broca's aphasia'). Less than two decades later Carl Wernicke published evidence for a second language center. This area is anatomically distinct from Broca's area, and damage to it produced a very different set of aphasic symptoms. The cortical area that still bears his name ('Wernicke's area') is located around the first and second convolutions in temporal cortex, and the aphasia that bears his name ('Wernicke's aphasia') involves deficits in language comprehension. Wernicke's method, like Broca's, was based on lesion studies: a careful evaluation of the behavioral deficits followed by post mortem examination to find the sites of tissue damage and atrophy. Lesion studies suggesting more precise localization of specific linguistic functions remain a cornerstone to this day in aphasic research.

Lesion studies have also produced evidence for the localization of other cognitive functions: for example, sensory processing and certain types of learning and memory. However, localization arguments for these other functions invariably include studies using animal models. With an animal model, one can perform careful behavioral measures in highly controlled settings, then ablate specific areas of neural tissue (or use a variety of other techniques to block or enhance activity in these areas) and remeasure performance on the same behavioral tests. But since we lack an animal model for (human) language production and comprehension, this additional evidence isn't available to the neurologist or neurolinguist. This fact makes the study of language a paradigm case for evaluating the logic of the lesion/deficit method of inferring functional localization. Philosopher Barbara Von Eckardt (1978) attempts to make explicit the steps of reasoning involved in this common and historically important method. Her analysis begins with Robert Cummins' early analysis of functional explanation, but she extends it into a notion of *structurally adequate* functional analysis. These analyses break down a complex capacity C into its constituent capacities c_1, c_2, . . ., c_n, where the constituent capacities are consistent with the underlying structural details of the system. For example, human speech production (complex capacity C) results from

formulating a speech intention, then selecting appropriate linguistic representations to capture the content of the speech intention, then formulating the motor commands to produce the appropriate sounds, then communicating these motor commands to the appropriate motor pathways (constituent capacities c_1, c_2, . . ., c_n). A functional-localization hypothesis has the form: brain structure S in organism (type) O has constituent capacity c_i, where c_i is a function of some part of O. An example might be: Broca's area (S) in humans (O) formulates motor commands to produce the appropriate sounds (one of the constituent capacities c_i). Such hypotheses specify aspects of the structural realization of a functional-component model. They are part of the theory of the neural realization of the functional model.

Armed with these characterizations, Von Eckardt argues that inference to a functional-localization hypothesis proceeds in two steps. First, a functional deficit in a patient is hypothesized based on the abnormal behavior the patient exhibits. Second, localization of function in normal brains is inferred on the basis of the functional deficit hypothesis plus the evidence about the site of brain damage. The structurally-adequate functional analysis of the capacity connects the pathological behavior to the hypothesized functional deficit. This connection suggests four adequacy conditions on a functional deficit hypothesis. First, the pathological behavior P (e.g., the speech deficits characteristic of Broca's aphasia) must result from failing to exercise some complex capacity C (human speech production). Second, there must be a structurally-adequate functional analysis of how people exercise capacity C that involves some constituent capacity c_i (formulating motor commands to produce the appropriate sounds). Third, the operation of the steps described by the structurally-adequate functional analysis minus the operation of the component performing c_i (Broca's area) must result in pathological behavior P. Fourth, there must not be a better available explanation for why the patient does P. Arguments to a functional deficit hypothesis on the basis of pathological behavior is thus an instance of argument to the best available explanation. When postulating a deficit in a normal functional component provides the best available explanation of the pathological data, we are justified in drawing the inference.

Von Eckardt applies this analysis to a neurological case study involving a controversial reinterpretation of agnosia.[14] Her philosophical explication of this important neurological method reveals that most challenges to localization arguments either argue only against the localization of a particular type of functional capacity or against generalizing from localization of function in one individual to all normal individuals. (She presents examples of each from the neurological literature.) Such challenges do not impugn the validity of standard arguments for functional localization from deficits. It does not follow that such arguments are unproblematic. But they face difficult factual and methodological problems, not logical ones. Furthermore, the analysis of these arguments as involving a type of functional analysis and inference to the best available explanation carries an important implication for the biological study of cognitive function. Functional analyses require functional theories, and structurally adequate functional analyses require checks imposed by the lower level sciences investigating the underlying physical mechanisms. Arguments to best available explanation are often hampered by a lack of theoretical imagination: the available explanations are often severely limited. We must seek theoretical inspiration from any level of theory and explanation. Hence making explicit the ‘logic’ of this common and historically important form of neurological explanation reveals the necessity of joint participation from all scientific levels, from cognitive psychology down to molecular neuroscience. Von Eckardt (1978)

anticipated what came to be heralded as the 'co-evolutionary research methodology,' which remains a centerpiece of neurophilosophy to the present day.

Over the last two decades, evidence for localization of cognitive function has come increasingly from a new source: the development and refinement of neuroimaging techniques. The form of localization-of-function argument appears not to have changed from that employing lesion studies (as analyzed by Von Eckardt). Instead, these imaging technologies resolve some of the methodological problems that plauge lesion studies. For example, researchers do not need to wait until the patient dies, and in the meantime probably acquires additional brain damage, to find the lesion sites. Two functional imaging techniques are prominent: positron emission tomography, or PET, and functional magnetic resonance imaging, or fMRI. Although these measure different biological markers of functional activity, both now have a resolution down to around 1mm.[15] As these techniques increase spatial and temporal resolution of functional markers and continue to be used with sophisticated behavioral methodologies, the possibility of localizing specific psychological functions to increasingly specific neural regions continues to grow.[16]

A Result of the Co-evolutionary Research Ideology: Cognitive and Computational Neuroscience

What we now know about the cellular and molecular mechanisms of neural conductance and transmission is spectacular. (For those in doubt, simply peruse for five minutes a recent volume of *Society for Neuroscience Abstracts*.) The same evaluation holds for all levels of explanation and theory about the mind/brain: maps, networks, systems, and behavior. This is a natural outcome of increasing scientific specialization. We develop the technology, the experimental techniques, and the theoretical frameworks within specific disciplines to push forward our understanding. Still, a crucial aspect of the total picture gets neglected: the relationship between the levels, the 'glue' that binds knowledge of neuron activity to subcellular and molecular mechanisms, network activity patterns to the activity of and connectivity between single neurons, and behavior to network activity. This problem is especially glaring when we focus on the relationship between 'cognitivist' psychological theories, postulating information-bearing representations and processes operating over their contents, and the activity patterns in networks of neurons. Co-evolution between explanatory levels still seems more like a distant dream rather than an operative methodology.

It is here that some neuroscientists appeal to 'computational' methods (Churchland and Sejnowski 1992). If we examine the way that computational models function in more developed sciences (like physics), we find the resources of *dynamical systems* constantly employed. Global effects (such as large-scale meteorological patterns) are explained in terms of the interaction of 'local' lower-level physical phenomena, but only by dynamical, nonlinear, and often chaotic sequences and combinations. Addressing the interlocking levels of theory and explanation in the mind/brain using computational resources that have worked to bridge levels in more mature sciences might yield comparable results. This methodology is necessarily interdisciplinary, drawing on resources and researchers from a variety of levels, including higher levels like experimental psychology, 'program-writing' and 'connectionist' artificial intelligence,

and philosophy of science.

However, the use of computational methods in neuroscience is not new. Hodgkin, Huxley, and Katz (1952) incorporated values of voltage-dependent potassium conductance they had measured experimentally in the squid giant axon into an equation from physics describing the time evolution of a first-order kinetic process. This equation enabled them to calculate best-fit curves for modeled conductance versus time data that reproduced the S-shaped (sigmoidal) function suggested by their experimental data. Using equations borrowed from physics, Rall (1959) developed the cable model of dendrites. This theory provided an account of how the various inputs from across the dendritic tree interact temporally and spatially to determine the input-output properties of single neurons. It remains influential today, and has been incorporated into the GENESIS software for programming neurally realistic networks (Bower and Beeman 1995). More recently, David Sparks and his colleagues have shown that a vector-averaging model of activity in neurons of superior colliculi correctly predicts experimental results about the amplitude and direction of saccadic eye movements (Lee, Rohrer, and Sparks 1988). Working with a more sophisticated mathematical model, Apostolos Georgopoulos and his colleagues have predicted direction and amplitude of hand and arm movements based on averaged activity of 224 cells in motor cortex. Their predictions have borne out under a variety of experimental tests (Geogopoulos *et al*. 1986). We mention these particular studies only because we are familiar with them. We could multiply examples of the fruitful interaction of computational and experimental methods in neuroscience easily by one-hundred-fold. Many of these extend back before ‘computational neuroscience’ was a recognized research endeavor.

We've already seen one example, the vector transformation account, of neural representation and computation, under active development in cognitive neuroscience. Other approaches using ‘cognitivist’ resources are also being pursued.[17] Many of these projects draw upon ‘cognitivist’ characterizations of the phenomena to be explained. Many exploit ‘cognitivist’ experimental techniques and methodologies. Some even attempt to derive ‘cognitivist’ explanations from cell-biological processes (e.g., Hawkins and Kandel 1984). As Stephen Kosslyn (1997) puts it, cognitive neuroscientists employ the ‘information processing’ view of the mind characteristic of cognitivism without trying to separate it from theories of brain mechanisms. Such an endeavor calls for an interdisciplinary community willing to communicate the relevant portions of the mountain of detail gathered in individual disciplines with interested nonspecialists: not just people willing to confer with those working at related levels, but researchers trained in the methods and factual details of a variety of levels. This is a daunting requirement, but it does offer some hope for philosophers wishing to contribute to future neuroscience. Thinkers trained in both the ‘synoptic vision’ afforded by philosophy and the factual and experimental basis of genuine graduate-level science would be ideally equipped for this task. Recognition of this potential niche has been slow among graduate programs in philosophy, but there is some hope that a few programs are taking steps to fill it. (See, e.g., "Other Internet Resources," linked below.)

Bibliography

- **Akins, Kathleen** (1993a) ‘What Is It Like to be Boring and Myopic?’ In B. Dahlboom (ed.)

Dennett and His Critics. New York: Basil Blackwell.

- ----- (1993b) ‘A Bat Without Qualities.’ In M. Davies and G. Humphreys (eds.) *Consciousness: Psychological and Philosophical Essays*. New York: Basil Blackwell.
- ----- (1996) ‘Of Sensory Systems and the ‘Aboutness’ of Mental States.’ *Journal of Philosophy*. 93, 337-372.
- **Aston-Jones, G., Desimone, R., Driver, J., Luck, S., and Posner, M.** (1999) ‘Attention.’ In Zigmond *et al*.
- **Balzer, Wolfgang, Moulines, C. Ulises, and Sneed, Joseph** (1987) *An Architectonic for Science*. Dordrecht: Reidel.
- **Bechtel, William** (1998) ‘Representations and Cognitive Explanations: Assessing the Dynamicist Challenge in Cognitive Science.’ *Cognitive Science*. 22, 295-318.
- **Bechtel, William, and Mundale, Jennifer** (1997) "Multiple Realizability Revisited." *Proceedings of the Australian Cognitive Science Society.*
- **Bechtel, William, and Mundale, Jennifer** (1999) "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science.* 66, 175-207.
- **Bechtel, William, and Richardson, Robert** (1993) *Discovering Complexity*. Princeton: Princeton University Press.
- **Bechtel, W., Mandik, P., Mundale, J., and Stufflebeam, R.S.** (forthcoming) *Philosophy and the Neurosciences: A Reader*. Oxford: Blackwell. [Table of Contents available online]
- **Bermudez, J.L** (1998) *The Paradox of Self-Consciousness.* Cambridge, MA: MIT Press.
- **Bickle, John** (1992) "Revisionary Physicalism." *Biology and Philosophy*. 7, 411-430.
- ----- (1995) "Psychoneural Reduction of the Genuinely Cognitive: Some Accomplished Facts." *Philosophical Psychology*. 8, 265-285.
- ----- (1998) *Psychoneural Reduction: The New Wave.* Cambridge, MA: MIT Press.
- **Biro, John** (1991) ‘Consciousness and Subjectivity’, In E. Villaneuva (ed.) *Philosophical Issues*. Atascadero, CA: Ridgeview.
- **Bliss,T.V.P. and Lomo, T.** (1973) "Long-Lasting Potentiation of Synaptic Transmission in the Dentate Area of the Anaesthetized Rabbit Following Stimulation of the Perforant Path." *Journal of Physiology (London)* 232, 331-356.
- **Block, Ned** (1986) ‘Advertisement for a Semantics for Psychology.’ In French, Uehling, and Wettstein (eds.) *Midwest Studies in Philosophy.* 10, 617-678.
- ----- (1988) ‘On a Confusion About a Function of Consciousness.’ *Behavioral and Brain Sciences.* 18, 227-247.
- **Bower, James and Beeman, David** (1995) *The Book of GENESIS*. New York: Springer-Verlag.
- **Caplan, D., Carr, T., Gould, J., and Martin, R.** (1999) ‘Language and Communication.’ In Zigmond et al.
- **Chalmers, David** (1996) *The Conscious Mind.* Oxford: Oxford University Press.
- **Churchland, Patricia** (1986) *Neurophilosophy*. Cambridge, MA: MIT Press.
- **Churchland, Patricia and Sejnowski, Terence** (1992) *The Computational Brain.* Cambridge, MA: MIT Press.
- **Churchland, Paul** (1979) *Scientific Realism and the Plasticity of Mind.* Cambridge: Cambridge University Press.
- **Churchland, Paul** (1981) "Eliminative Materialism and the Propositional Attitudes." *Journal of*

Philosophy. 78, 67-90.
- ----- (1987) *Matter and Consciousness*, revised edition. Cambridge, MA: MIT Press.
- ----- (1989) *A Neurocomputational Perspective*. Cambridge, MA: MIT Press.
- ----- (1995) *The Engine of Reason, The Seat of the Soul*. Cambridge, MA: MIT Press.
- ----- (1996) ‘The Rediscovery of Light.’ *Journal of Philosophy* 93, 211-228.
- **Churchland, Paul, and Churchland, Patricia** (1997) ‘Recent Work on Consciousness: Philosophical, Empirical and Theoretical.’ *Seminars in Neurology*. 17, 101-108.
- **Clark, Austen** (1993) *Sensory Qualities*. Cambridge: Cambridge University Press.
- **Dennett, Daniel** (1978) ‘Why You Can't Make a Computer That Feels Pain.’ *Synthese*. 38, 415-456.
- ----- (1991) *Consciousness Explained*. New York: Little Brown.
- ----- (1995) ‘The Path Not Taken.’ *Behavioral and Brain Sciences*. 18, 252-253.
- **Dretske, Fred** (1981) *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.
- ----- (1988) *Explaining Behavior*. Cambridge, MA: MIT Press.
- **Feyerabend, Paul** (1963) "Mental Events and the Brain." *Journal of Philosophy*. 60, 295-296.
- **Fodor, Jerry** (1974) "Special Sciences." *Synthese*. 28, 77-115.
- ----- (1981) *RePresentations*. Cambridge, MA: MIT Press.
- ----- (1987) *Psychosemantics*. Cambridge, MA: MIT Press.
- **Fodor, Jerry and LePore, Ernest** (1992) *Holism: A Shopper's Guide*. Cambridge, MA: MIT Press.
- **Gazzaniga, Michael** (ed.) (1995) *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.
- **Georgopoulos, A., Schwartz, A., and Kettner, R.** (1986) ‘Neuronal Population Coding of Movement Direction.’ *Science*. 233, 1416-1419.
- **Hardcastle, Valerie Gray** (1997) ‘When a Pain Is Not.’ *Journal of Philosophy*. 94, 381-409.
- **Hardin, C.L.** (1988) *Color for Philosophers*. Hackett.
- **Haugeland, John** (1985) *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.
- **Hawkins, Richard and Kandel, Eric** (1984) ‘Is There a Cell-Biological Alphabet for Learning?’ *Psychological Review*. 91, 375-391.
- **Hebb, Donald** (1949) *The Organization of Behavior*. New York: Wiley.
- **Hooker, Clifford** (1981) "Towards a General Theory of Reduction. Part I: Historical and Scientific Setting. Part II: Identity in Reduction. Part III: Cross-Categorial Reduction." *Dialogue*. 20, 38-59, 201-236, 496-529.
- **Horgan, Terence and Graham, George** (1991) "In Defense of Southern Fundamentalism." *Philosophical Studies*. 62, 107-134.
- **Hubel, David and Wiesel, Torsten** (1962) ‘Receptive Fields, Binocular Interaction and Functional Architecture In the Cat's Visual Cortex.’ *Journal of Physiology* (London). 160, 106-154.
- **Jackson, Frank and Pettit, Philip** (1990)."In Defense of Folk Psychology." *Philosophical Studies*. 59, 31-54.
- **Kandel, Eric** (1976) *Cellular Basis of Behavior*. San Francisco: W.H. Freeman.
- **Kolb, Bryan and Whishaw, Ian** (1996). *Fundamentals of Human Neuropsychology*, 4th edition. W.H. Freeman.
- **Kosslyn, Stephen** (1997) ‘Mental Imagery.’ In S. Gazzaniga (ed.) *Conversations in the Cognitive*

Neurosciences. Cambridge, MA: MIT Press.

- **Lee, C.W., Rohrer, R., and Sparks, D.** (1988) ‘Population Coding of Saccadic Eye Movements by Neurons in the Superior Colliculus.’ *Nature*. 332, 357-360.
- **Lehky, S.R. and Sejnowski, T.** (1988) ‘Network Model of Shape-from-Shading: Neural Function Arises from Both Receptive and Projective Fields.’ *Nature*. 333, 452-454.
- **Lettvin, J.Y., Maturana, H.R., McCulloch, W.S., and Pitts, W.H.** (1959) ‘What the Frog's Eye Tells the Frog's Brain.’ *Proceedings of the IRF*. 47, 1940-1951.
- **Levine, Joseph** (1983) ‘Materialism and Qualia: The Explanatory Gap.’ *Pacific Philosophical Quarterly*. 64, 354-361.
- **Llinás, Rodolfo** (1975) ‘The Cortex of the Cerebellum." *Scientific American* 232, 56-71.
- **Llinás, Rodolfo, and Churchland, Patricia** (eds.) (1996) *The Mind-Brain Continuum*. Cambridge, MA: MIT Press.
- **Magistretti, P.** (1999) ‘Brain Energy Metabolism.’ In Zigmond et al.
- **Nagel, Thomas** (1971) ‘Brain Bisection and the Unity of Consciousness’ *Synthese*. 25, 396-413.
- **-----** (1974) "What Is It Like to Be A Bat?" *Philosophical Review*. 83, 435-450.
- **Place, U.T.** (1956) ‘Is Consciousness a Brain Process?’ *The British Journal of Psychology*. 47, 44-50.
- **Putnam, Hilary** (1967) ‘Psychological Predicates.’ In Capitan and Merrill (eds.), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press.
- **Rall, W.** (1959) ‘Branching Dendritic Trees and Motoneuron Membrane Resistivity.’ *Experimental Neurology*. 1, 491-527.
- **Ramsey, William** (1992) ‘Prototypes and Conceptual Analysis." *Topoi* 11, 59-70.
- **Rumelhart, David, Hinton, Geoffrey, and McClelland, James** (1986) "A Framework for Parallel Distributed Processing." In Rumelhart and McClelland (eds.), *Parallel Distributed Processing*. Cambridge, MA: MIT Press.
- **Sacks, Oliver** (1985). *The Man Who Mistook his Wife for a Hat*. New York: Summit Books
- **Schaffner, Kenneth** (1992) ‘Philosophy of Medicine.’ In M. Salmon, J. Earman, C. Glymour, J. Lennox, P. Machamer, J. McGuire. J. Norton, W. Salmon, and K. Schaffner (eds.), *Introduction to the Philosophy of Science*. Englewood Cliffs, NJ: Prentice-Hall.
- **Smart, J.J.C.** (1959) ‘Sensations and Brain Processes.’ *Philosophical Review*. 68, 141-156.
- **Stich, Stephen** (1983) *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.
- **Stufflebam, Robert and Bechtel, William** (1997) ‘PET: Exploring the Myth and the Method.’ *Philosophy of Science* (Supplement), S95-S106.
- **Suppe, Frederick** (1974) *The Structure of Scientific Theories*. Urbana: University of Illinois Press.
- **Tye, Michael** (1993) ‘Blindsight, the Absent Qualia Hypothesis, and the Mystery of Consciousness.’ In C. Hookway (ed.), *Philosophy and the Cognitive Sciences*. Cambridge: Cambridge University Press.
- **Van Fraassen, Bas C.** (1980). *The Scientific Image* Oxford University Press.
- **Von Eckardt Klein, Barbara** (1975). ‘Some Consequences of Knowing Everything (Essential) There is to Know About one's Mental States.’ *Review of Metaphysics*.
- **Von Eckardt Klein, Barbara** (1978) ‘Inferring Functional Localization from Neurological Evidence.’ In E. Walker (ed.), *Explorations in the Biology of Language*. Cambridge, MA: MIT

Press.

- **Zigmond, M., Bloom, F., Landis, S., Roberts, J., and Squire, L.** (eds.) (1999) *Fundamental Neuroscience*. San Diego: Academic Press.

Other Internet Resources

- Pete Mandik's ‘Philosophy and the Neurosciences’ site
- Valerie Hardcastle's ‘Mind/Brain Resources’ site
- Philosophy-Neuroscience-Psychology Program at Washington University in St. Louis

Related Entries

cognitive science | color | connectionism | Feyerabend, Paul | functionalism | identity theory of mind | multiple realizability | physicalism | qualia | reduction and reductionism

John Bickle
bicklejw@email.uc.edu
and
Pete Mandik
mandikp@wpunj.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 7, 1999
Content last modified: May 16, 2001

Stanford Encyclopedia of Philosophy
Notes to Philosophy of Neuroscience

Notes

1. See Bickle 1995 and 1998, Chapter 5, and Schaffner 1992 for non-technical overviews emphasizing consequence for the intertheoretic reducibility of cognitive psychology to neuroscience.

2. This assumption is not gratuitous. We saw in the first paragraph of this section that folk psychological generalizations advert to a variety of propositional attitudes. Eliminativists are also not the only theorists who make this assumption. Incarnations of Jerry Fodor's (1981, 1987) influential (and noneliminative) Representational Theory of Mind assume explicitly this account of folk psychology. There have been dissenters, however: Horgan and Graham, 1991 and Jackson and Pettit, 1990.

3. See Rumelhart *et al*., 1986 for an overview of this mathematical framework for parallel distributed processing in artificial systems structured like brains.

4. For a good overview of this circuitry with helpful illustrations, see Llinás, 1975.

5. Hooker, 1981 suggests this terminology. See Bickle 1998, chapter 3 for an attempt to quantify these notions.

6. Gate-control theories hypothesis that the extensive feedback projections in the mammalian pain transmission system serve to inhibit, enhance, or distort feedforward pain information (from nociceptors (pain receptors) to central nervous system). See Hardcastle, 1997 for a critical discussion of these theories, with extensive references.

7. See Churchland, 1995, for an illustration of the location of a number of color stimuli in this space.

8. See Churchland 1989 for an eye-opening approximation of this number.

9. Churchland (1995) himself provides a non-color example of a three-dimensional face space, where the dimensions represent nose width, eye separation, and mouth fullness.

10. For an early articulation of the ‘explanatory gap,’ see Levine 1983.

11. The thalamus is a bilateral structure at the rostral tip of the brainstem. Its various nuclei are densely

connected with cortex.

12. Commissurotomy is a surgical procedure to treat severe epilepsy resistant to other regimes. It involves ablating portions of the corpus callosum, a large band of axon fibers connecting the two cerebral hemispheres.

13. The aphasias are specifically linguistic deficits. They can include, but are not limited to, speech production and speech comprehension deficits. See Kolb and Whishaw, 1996.

14. Agnosia is a recognitional deficit that does not involve specifical sensory deficits. A visual agnosic, for example, can describe features of objects presented visually and even accurately draw the object, but will be unable to identify it (e.g., "rose"); though typically he will be able to identify the object via other sense modalities (olfaction). A prosopagnosic will be able to identify features of faces, but will be unable to identify whose face it is, even if the face has been presented many times. See Kolb and Whishaw (1996) for a good overview. See Sacks (1985) for some clinical descriptions that depict the human side of the deficit.

15. Aside from the potential neurophilosophical impact of being able to image specific neural activity during specific behavioral and cognitive tasks, the underlying science of these techniques is both fascinating and not yet entirely understood. For PET, water or sugar molecules are labeled with unstable radionuclides possessing excessive protons (Magistretti 1999). This extra proton is converted into a neutron by the normal process of radioactive decay. This process emits a positron (a positively-charged electron) which collides with an electron and releases two photons with opposite trajectories. Special detectors located around the head respond to these photons. When two photons simultaneously reach detectors oriented 180 degrees to each other, the positron-electron collision can be localized to a resolution of a few millimeters. Water molecules labeled with oxygen-15 are used to measure amount of blood flow, and deoxyglucose molecules labeled with fluorine-18 are used to measure glucose utilization. Both blood flow and glucose utilization are correlated directly with level of neural and glial cell activity, so a PET scan provides an extremely accurate measure of location of neural activity in baseline and test situations. Sophisticated algorithms and computer graphics produce an image of different activity levels in different neural regions. Activity profiles can be analyzed at a variety of imaged ‘slices’ or ‘cuts’ through the tissue (hence the term, ‘tomography’). Researchers have used PET studies to obtain evidence for, e.g., activity in anterior cingulate cortex as crucial to the executive control of attention (reviewed in Aston-Jones *et al*. 1999). PET technology has also been used extensively to compare and contrast the various neural areas active during distinct linguistic tasks like reading and writing (reviewed in Caplan *et al*. 1999).

fMRI technology provides a way to measure directly the amount of oxygenation or phosphorylation in specific regions of neural tissue (Magistretti 1999). These markers are directly tied to neural activity level since they indicate cell respiration and ATP utilization. Depending on its degree of saturation by oxygen, hemoglobin in the blood alters the signal of biological tissue subjected to a strong magnetic field and then released. Biophysicists still debate the exact mechanisms of these patterns of detectable energy

release. But sensors located around the tissue can convert this energy into coordinates to compute resolution of activity to a few millimeters (roughly equivalent to PET). Although the clinical and experimental applications of this technology are still in their infancy, fMRI has been used to study the localization of specific linguistic functions, memorial functions, and executive and planning functions of frontal cortex.

16. For a useful evaluation and philosophical discussion of PET, seesee Stufflebeam and Bechtel 1997. For further philosophical discussion of localization techniques and arguments see Bechtel and Richardson 1993 and Bechtel and Mundale 1997 and in press.

17. See Llinás and Churchland 1996, especially the essay by Merzenich and deCharms, for the variety of ‘representation’ concepts employed in modern neurobiological theorizing.

John Bickle
bicklejw@email.uc.edu
and
Peter Mandik
pjmandik@artsci.wustl.edu

First published: June 7, 1999
Content last modified: June 7, 1999

FIGURE 1 ARCHITECTURE OF A CEREBELLAR NETWORK

See text for full description (adapted with revisions from Paul Churchland, 1987.)

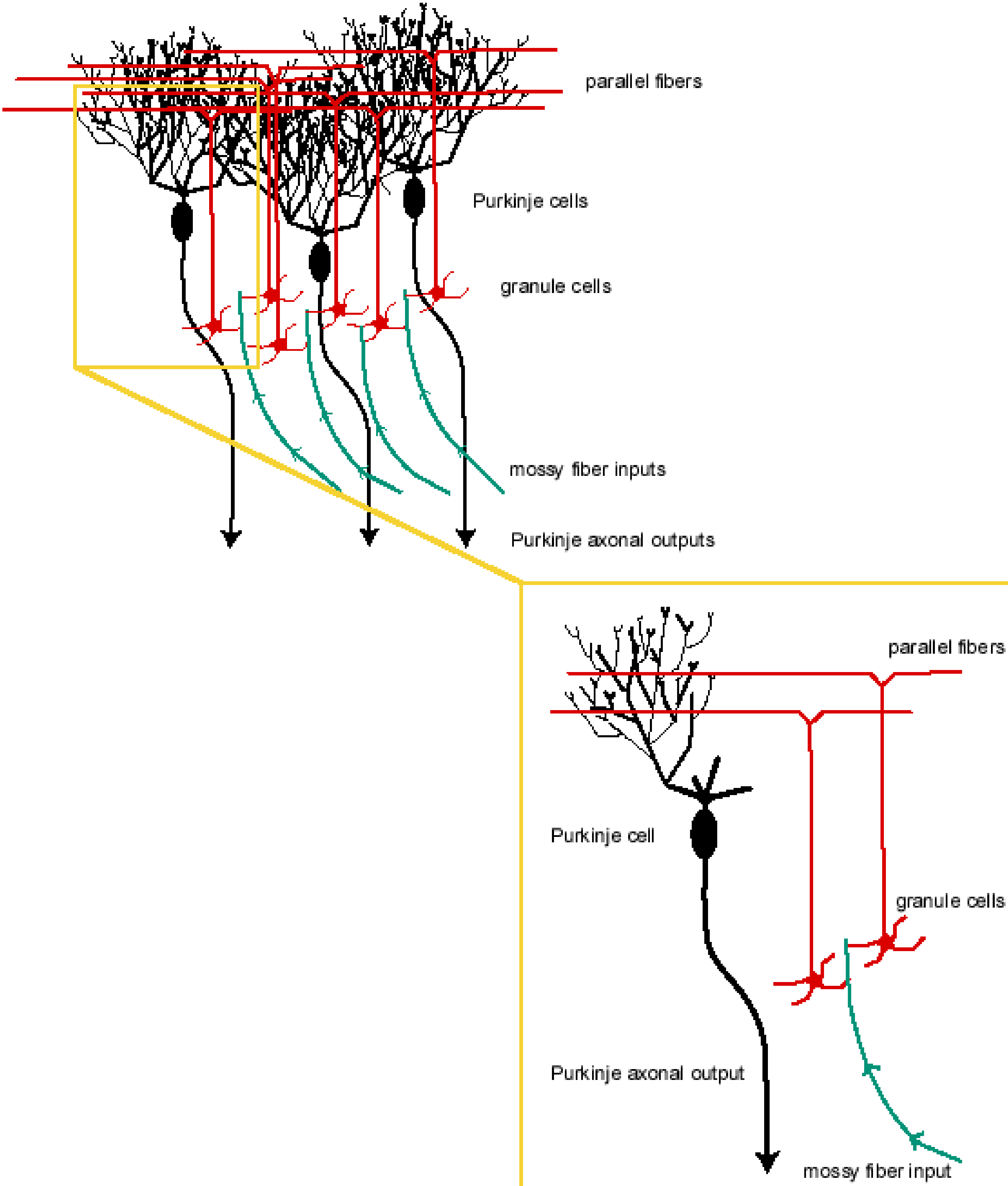


FIGURE 2 LEARNING CHARACTERIZED AS GRADIENT DESCENT IN ERROR-SYNAPTIC WEIGHT SPACE

One axis (y) reflects the global error measure of the network's output to a given input.
The other axes refect weight values of two synapses in the network.
The complete error weight space will have n+1 dimensions, where n is the number of synapses in the network.

See text for full description. (Adapted from Paul Churchland, 1987.)

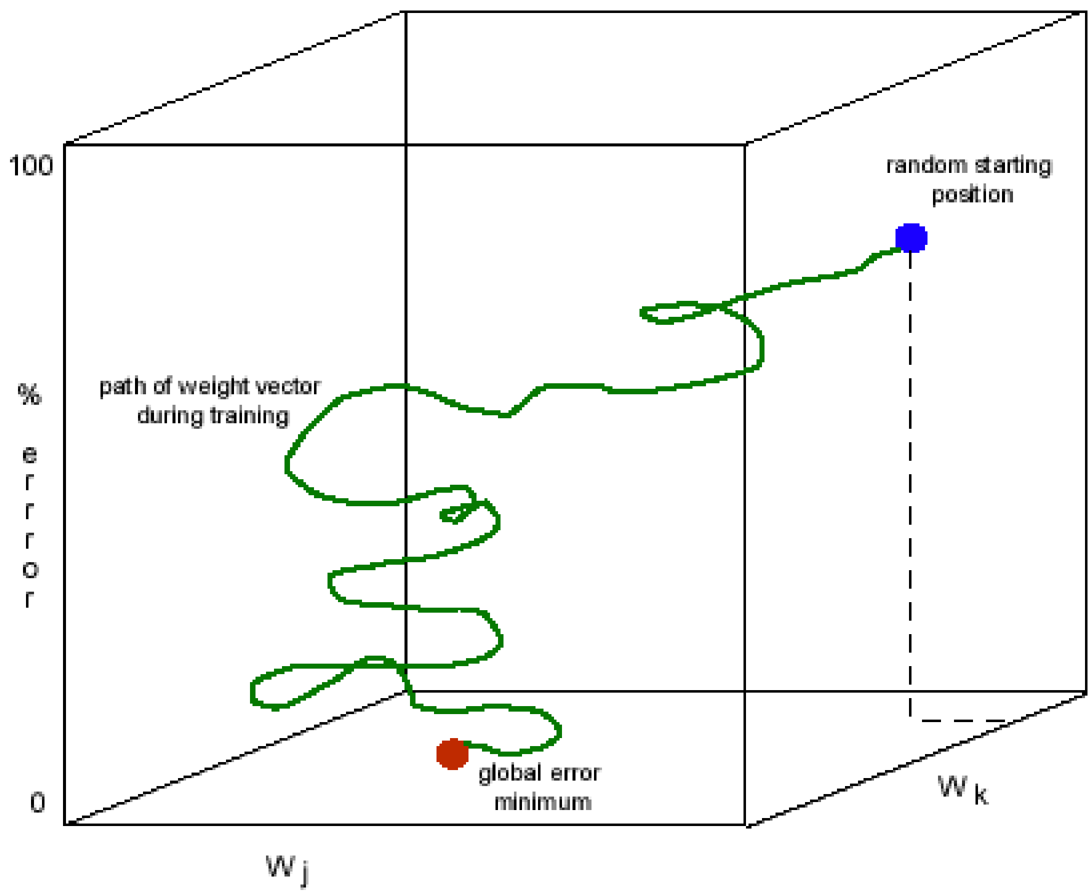
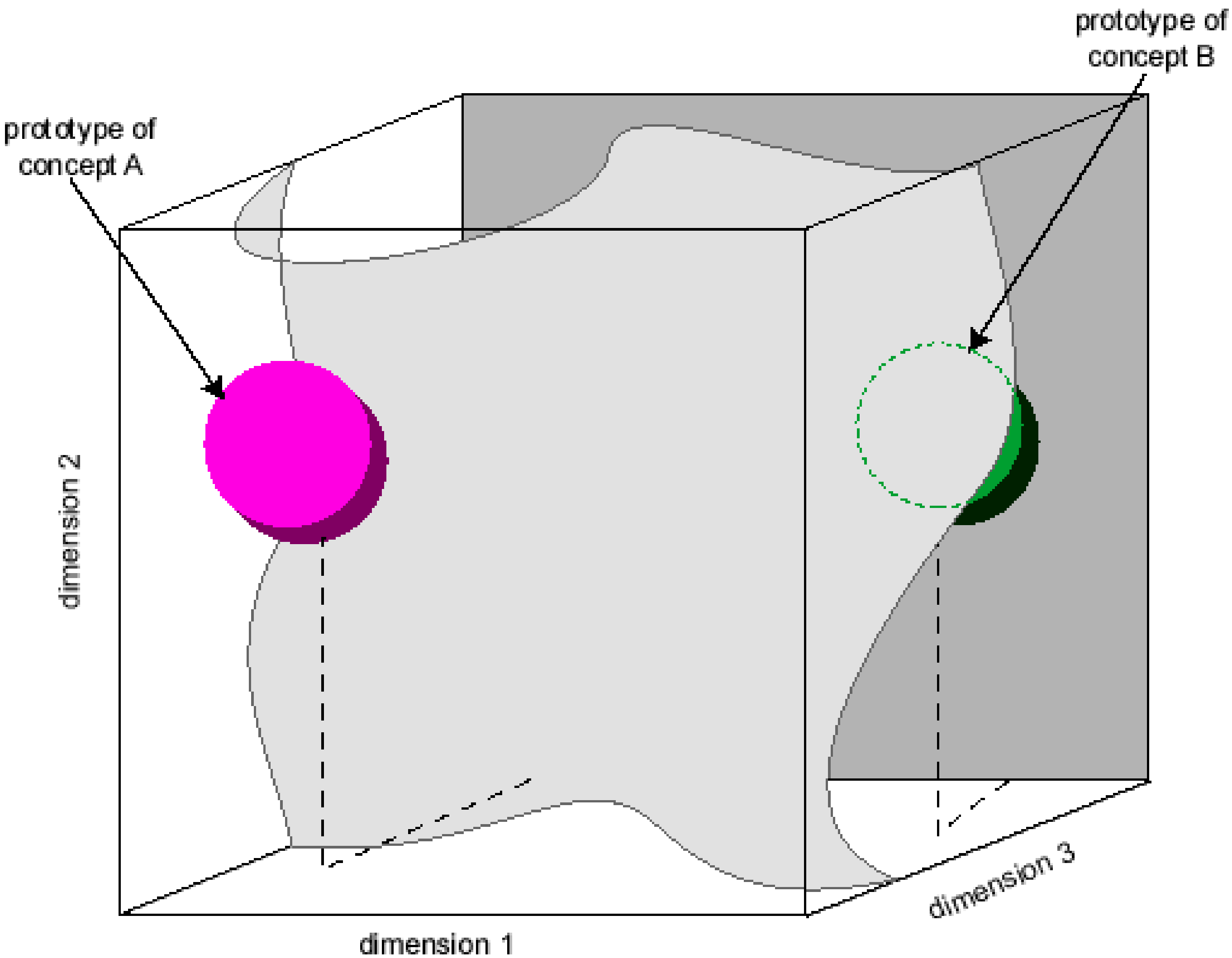


FIGURE 3 REPRESENTATIONAL CONTENT (CONCEPTS) CHARACTERIZED AS PARTITIONS IN HIGH-DIMENSIONAL VECTOR STATE-SPECES

A prototype of the concept is the center point or subvolume in the partition.

See text for full details. (Adapted from paul Churchland, 1989.)



Paul Feyerabend

Paul Feyerabend (b.1924, d.1994), having studied science at the University of Vienna, moved into philosophy for his doctoral thesis, made a name for himself both as an expositor and (later) as a critic of Karl Popper's ‘critical rationalism’, and went on to become one of this century's most famous philosophers of science. An imaginative maverick, he became a critic of philosophy of science itself, particularly of ‘rationalist’ attempts to lay down or discover rules of scientific method.

1. A Brief Chronology of Feyerabend's Life and Work.

1924 Born in Vienna. Son of a civil servant and a seamstress.

1940 Was inducted into the *Arbeitsdienst* (the work service introduced by the Nazis).

1942 Drafted into the Pioneer Corps of the German army. After basic training, volunteered for Officers' School.

1943 Learned of his mother's suicide.

1944 Decorated, Iron Cross. Advanced to Lieutenant. Lectured to Officers' School.

1945 Shot in the spine during the retreat from the Russian Army.

1946 Received a fellowship to study singing and stage-management in Weimar. Joined the ‘Cultural Association for the Democratic Reform of Germany’.

1947 Returned to Vienna to study history and sociology at the University. Soon transferred to physics. First article, on the concept of illustration in modern physics, published. Feyerabend ‘a raving positivist’ at the time.

1948 First visit to the Alpbach seminar of the Austrian College Society. Became secretary of the seminars. Met Karl Popper and Walter Hollitscher. Married first wife, Edeltrud.

1949 Became student leader of the ‘Kraft Circle’, a student philosophy club centred around Viktor Kraft, Feyerabend's dissertation supervisor and a former member of the Vienna Circle. Ludwig Wittgenstein visited the Kraft Circle to give a talk. Feyerabend also met Bertholt Brecht.

1951 Received doctorate in philosophy for his thesis on ‘basic statements’. Applied for a British Council scholarship to study under Wittgenstein at Cambridge. But Wittgenstein died before Feyerabend arrived in England, so Feyerabend chose Popper as his supervisor instead.

1952 Came to England, to study under Popper at the London School of Economics. Concentrated on the quantum theory and Wittgenstein. Studied the typescript of Wittgenstein's *Philosophical Investigations*, and prepared a summary of the book. Befriended another of Popper's students, Joseph Agassi.

1953 Feyerabend returned to Vienna. Popper applied for an extension to his scholarship, but Feyerabend decided to remain in Vienna instead. Translated Popper's *The Open Society and its Enemies* into German. Declined the offer to become Popper's research assistant. Agassi took the post. Feyerabend became research assistant to Arthur Pap in Vienna.

1954 First articles on quantum mechanics and on Wittgenstein published. Pap introduced Feyerabend to Herbert Feigl.

1955 Took up his first full-time academic appointment as lecturer in philosophy at the University of Bristol, England. His summary of Wittgenstein's *Philosophical Investigations* appeared as a review of the book in *The Philosophical Review*.

1956 Married second wife, Mary O'Neill. Published an article on the ‘paradox of analysis’. Feyerabend got to know the quantum physicist David Bohm, whose ideas were to influence him substantially.

1957 Gave a paper on the quantum theory of measurement to the Colston Research Symposium at the University of Bristol.

1958 Took up visiting lectureship at the University of California, Berkeley. Two of his most important early papers, ‘An Attempt at a Realistic Interpretation of Experience’, and ‘Complementarity’ appeared in the proceedings of the Aristotelian Society. In them, Feyerabend argued against positivism and in favour of a scientific realist account of the relation between theory and experience, largely on grounds familiar from Karl Popper's falsificationist views.

1959 Accepted a permanent position at Berkeley, emigrated to the US, becoming a naturalized US citizen.

1960 As a result of earlier discussions with Herbert Feigl, Feyerabend published ‘Das Problem der Existenz theoretischer Entitäten’, in which he argued that there is no special ‘problem’ of theoretical entities, and that *all* entities are hypothetical. Gave two lectures to Oberlin College, Ohio, in which he embroidered on Popper's views about the pre-Socratic thinkers.

1962 ‘Explanation, Reduction, and Empiricism’ appeared. Criticised existing empiricist accounts of explanation and theoretical reduction (Hempel, Nagel), and introduced the concept of incommensurability, based on the ‘contextual theory of meaning’ which Feyerabend claimed to find in Wittgenstein's *Investigations*.

1963 ‘How to be a Good Empiricist’, a position paper summing up his point of view, was published, along with his two main articles on the Mind/Body Problem in which he introduced the position now known as ‘eliminative materialism’.

1965 Publication of the first part of the essay ‘Problems of Empiricism’, and his ‘Reply to Criticism’, in which Feyerabend made his last serious attempt to construct a ‘tolerant’, ‘disinfected’ empiricism. Although beginning to put some distance between himself and Popper, Feyerabend was still able to write a glowing review of Popper's *Conjectures and Refutations*.

1967-8 Focus of his published papers had by now moved to ‘theoretical pluralism’, the view that in order to maximise the chances of falsifying existing theories, scientists should construct and defend as many alternative theories as possible. Feyerabend's articles ‘On a Recent Critique of Complementarity’ defended Niels Bohr's views against Popper's critique. Popper not amused.

1969 In a tiny article, ‘Science Without Experience’, Feyerabend finally gave up the attempt to be an empiricist, arguing that in principle experience is necessary at *no* point in the construction, comprehension or testing of empirical scientific theories.

1970 Publication of ‘Consolations for the Specialist’, in which Feyerabend attacked Popper from a Kuhnian point of view, and the essay version of ‘Against Method: Outline of an Anarchistic Theory of Knowledge’, in which ‘epistemological anarchism’ was revealed for the first time. Feyerabend claimed to be applying the liberalism of John Stuart Mill's *On Liberty* to scientific methodology. Published little during the next few years.

1974 Death of Feyerabend's friend Imre Lakatos, putting paid to their plans to produce a dialogue volume, *For and Against Method*. Feyerabend, lecturing at the University of Sussex, was ill too. Published a scathing review of Popper's *Objective Knowledge*.

1975 Appearance of Feyerabend's first book, *Against Method*, setting out ‘epistemological anarchism’, whose main thesis was that there is no such thing as *the* scientific method. Great scientists are methodological opportunists who use any moves that comes to hand, even if they thereby violate canons of empiricist methodology.

1976-7 Feyerabend replies to most of the major reviewers of *Against Method*. Got depressed. Published his first major article on relativism: the first time he explicitly endorsed the view.

1978 *Science in a Free Society* appears, including replies to reviewers of *Against Method*. Some clarification of epistemological anarchism, and very little retreat from the position set out in *AM*. Explored further the political implications of epistemological anarchism. The book also included one of Feyerabend's major endorsements of relativism, one of the views for which he was becoming known. First volume of the German edition of Feyerabend's philosophical papers appears. (Feyerabend published increasingly in German from this point onwards).

1981 English publication of the first two volumes of Feyerabend's *Philosophical Papers*, with new material in introductory chapters.

1983 Met Grazia Borrini at his Berkeley lectures.

1984 Publishes 'Science as an Art', in which he defends an explicitly relativistic account of the history of science according to which there is change, but no 'progress'. Also continues his campaign to rehabilitate Ernst Mach.

1987 Publication of *Farewell to Reason*, a volume collecting some of the papers Feyerabend had published between 1981 and 1987. Relativism again at the forefront, especially in its 'Protagorean' version.

1988 Second, revised edition of *Against Method*, omitting the long chapter on the history of the visual arts, but now incorporating parts of *Science in a Free Society*, appeared.

1989 Paul and Grazia married in January. Left for Italy and Switzerland in the fall, at least partly because of the effects of the October earthquake in California.

1990 Officially resigned from Berkeley in March.

1991 Retired from Zurich. *Three Dialogues on Knowledge* and *Beyond Reason*, a festschrift edited by a former pupil, Gonzalo Munévar, published. Also lots of small publications, many of them in *Common Knowledge*. Signs of an increasing unhappiness with relativism in Feyerabend's publications around this time. But still vigorously opposed to 'objectivism'.

1993 Third edition of *Against Method* published. Feyerabend developed an inoperable brain tumour, and was hospitalized.

1994 Feyerabend died at home in Zurich, February 11th. Several major memorial symposia and colloquia took place over the next two years.

1995 *Killing Time: The Autobiography of Paul Feyerabend* published.

2. Feyerabend's Life and Work: A Critical Appraisal.

(Unless otherwise stated, page references are to *Killing Time: The Autobiography of Paul Feyerabend*, (Chicago: University of Chicago Press, 1995), henceforth referred to as *'KT'*).

Early Life (1924-1938).

Paul Karl Feyerabend was born into a middle-class Viennese family in 1924. Times were hard in Vienna in the nineteen-twenties: in the aftermath of the First World War there were famines, hunger riots, and runaway inflation. Feyerabend's family had a three-room apartment on the *Wolfganggasse*, 'a quiet street lined with oak trees' (p.11). The first chapters of his autobiography give the impression of his being a strange child, whose activities were entirely centred around his own family, and who was cut off from neighbors, other children and the outside world because "[t]he world is a dangerous place" (p.15). Between the ages of three and six, Feyerabend recalls, he spent most of his time in the apartment's kitchen and bedroom. Occasional visits to the cinema and numerous stories, especially stories with a magical aura, seem to have taken the place usually filled by childhood friends. He was a sickly child, but ran away from home once, when he was five years old (p.7). When he started school at the age of six, he

"had no idea how other people lived or what to do with them" (p.16). The world seemed to be filled with strange and inexplicable happenings. It took him some while to get used to school, which initially made him sick. But when he did so, his health problems had disappeared. When he learned to read, he found the new and magical world of books waiting for him, and indulged himself to the full (p.25). But his sense of the world's inexplicability took some time to dissipate - he recalls feeling that way about events during the nineteen-thirties and throughout the second world war.

Feyerabend attended a *Realgymnasium* (High School) at which he was taught Latin, English, and science. He was a *Vorzugsschüler*, that is, ‘a student whose grades exceeded a certain average’ (p.22), and by the time he was sixteen he had the reputation of knowing more about physics and math than his teachers. But he also got thrown out of school on one occasion.

Feyerabend ‘stumbled into drama’ (p.26) by accident, becoming something of a ham actor in the process. This accident then led to another, when he found himself forced to accept philosophy texts among the bundles of books he had bought for the plays and novels they contained. It was, he later claimed, "the dramatic possibilities of reasoning and... the power that arguments seem to exert over people" (p.27) with which philosophy fascinated him. Although his reputation was as a philosopher, he preferred to be thought of as an entertainer. His interests, he said, were always somewhat unfocussed (p.27).

However, Feyerabend's school physics teacher Oswald Thomas inspired in him an interest in physics and astronomy. The first lecture he gave (at school) seems to have been on these subjects (p.28). Together with his father, he built a telescope and ‘became a regular observer for the Swiss Institute of Solar Research’ (p.29). He describes his scientific interests as follows:

> I was interested in both the technical and the more general aspects of physics and astronomy, but I drew no distinction between them. For me, Eddington, Mach (his *Mechanics* and *Theory of Heat*), and Hugo Dingler (*Foundations of Geometry*) were scientists who moved freely from one end of their subject to the other. I read Mach very carefully and made many notes. (p.30).

Feyerabend does not tell us how he became acquainted with another one of his main preoccupations - singing. He was proud of his voice, becoming a member of a choir, and took singing lessons for years, later claiming to have remained in California in order not to have to give up his singing teacher. In his autobiography he talks of the pleasure, greater than any intellectual pleasure, derived from having and using a well-trained singing voice (p.83). During his time in Vienna in the second world war, his interest led him to attend the opera (first the *Volksoper*, and then the *Staatsoper*) together with his mother. A former opera singer, Johann Langer, gave him singing lessons and encouraged him to go to an academy. After passing the entrance examination, Feyerabend did so, becoming a pupil of Adolf Vogel. At this point in his life, he later recalled:

> The course of my life was... clear: theoretical astronomy during the day, preferably in the domain of perturbation theory; then rehearsals, coaching, vocal exercises, opera in the

> evening...; and astronomical observation at night... The only remaining obstacle was the war. (p.35).

The *Anschluss* (1938).

Feyerabend tells how, without falling for Adolf Hitler's charisma, he appreciated Hitler's oratorial style. Austria was re-unified with Germany in 1938. Jewish schoolmates were treated differently, and Jewish neighbours and acquaintances started disappearing. But, as usual, Feyerabend had no clear view of the situation:

> Much of what happened I learned only after the war, from articles, books, and television, and the events I did notice either made no impression at all or affected me in a random way. I remember them and I can describe them, but there was no context to give them meaning and no aim to judge them by. (pp.37-8).

> For me the German occupation and the war that followed were an inconvenience, not a moral problem, and my reactions came from accidental moods and circumstances, not from a well-defined outlook. (p.38).

The general impression given by his autobiography is of an imaginative but fairly solitary person with no stable or well-defined personality. Rather, his decisions and courses of action seem to have been the result of a struggle between his tendency to conform and his contrariness. Just as when he was a child, events happening around him seemed strange, distant, and out of context. It is very difficult to see him identifying with any group, and he must have made an unlikely soldier.

The War (1939-1945).

As far as his army record goes, Feyerabend claims in his autobiography that his mind is a blank. But in fact this is one of the periods he tells us most about. Having passed his final high school exams in March 1942, he was drafted into the *Arbeitsdienst* (the work service introduced by the Nazis), and sent for basic training in Pirmasens, Germany. Feyerabend opted to stay in Germany to keep out of the way of the fighting, but subsequently asked to be sent to where the fighting was, having become bored with cleaning the barracks! He even considered joining the SS, for aesthetic reasons. His unit was then posted Quelerne en Bas, near Brest, in Brittany. Still, the events of the war did not register. In November 1942, he returned home to Vienna, but left before Christmas to join the Wehrmacht's Pioneer Corps.

Their training took place in Krems, near Vienna. Feyerabend soon volunteered for officers' school, not because of an urge for leadership, but out of a wish to survive, his intention being to use officers' school as a way to avoid front-line fighting. The trainees were sent to Yugoslavia. In Vukovar, during July 1943, he learnt of his mother's suicide, but was absolutely unmoved, and obviously shocked his fellow officers by displaying no feeling. In December that same year, Feyerabend's unit was sent into battle on the northern part of the Russian front, but although they blew up buildings, they never encountered any

Russian soldiers.

Despite the fact that Feyerabend reports of himself that he was foolhardy during battle, treating it as a theatrical event, he received the Iron Cross (second class) early in March 1944, for leading his men into a village under enemy fire, and occupying it. He was advanced from private soldier to lance corporal, to sergeant, and then, at the end of 1944, to lieutenant. At the end of November that year, he gave a series of lectures to the officers' school at Dessau Rosslau, near Leipzig. Their theme was the (‘historicist’) one that "historical periods such as the Baroque, the Rococo, the Gothic Age are unified by a concealed essence that only a lonely outsider can understand" (p.49). His description of these lectures, and of his notebook entries at the time, reveals the influence of Friedrich Nietzsche in their fascination with this ‘lonely outsider’, ‘the solitary thinker’ (p.48).

Having returned home for Christmas 1944, Feyerabend again boarded the train for the front, this time for Poland, in January 1945. There he was put in charge of a bicycle company. Although he claims to have relished the role of army officer no more than he later did that of university professor, he must have been at least a competent soldier, since in the field he came to take the place of a sequence of injured officers: first a lieutenant, then a captain, and then a major, before he was shot during another heroic act of carelessness performed in the 1945 retreat westwards from the Russian army. The bullet lodged in his spine left him temporarily paralysed from the waist down, meaning that he spent time in a wheelchair, then on crutches, and thereafter walked with the aid of a stick. The war ended as he was recovering from his injury, in a hospital in Apolda, a little town near Weimar, while fervently hoping not to recover before the war was over. Germany's surrender came as a relief, but also as a disappointment relative to past hopes and aspirations. He later said of his stint in the army that it was "an interruption, a nuisance; I forgot about it the moment it was over" (p.111).

Post-War Activities (1945-1947).

However, the war took its toll even on Feyerabend. The bullet in his spine left him impotent for the rest of his life. (His descriptions of subsequent sexual encounters are one of the more amusing parts of his autobiography). Although he started off completely ignorant of women, he married four times, and had, by his own account, plenty of affairs. But he seems to have been distant not just in his relationship with his parents, but in some of his marriages too. He hated the slavery love seemed to imply, but hated equally the freedom achieved by taking evasive action. He got bogged down in cycles of dependence, isolation, and renewed dependence, which only dissolved into a more balanced pattern after many years.

At the end of the war, Feyerabend went to the mayor of Apolda and asked for a job. He was assigned to the education section, given an office and a secretary and, fittingly, put in charge of entertainment.

In 1946, having recovered from paralysis, he received a state fellowship to return to study singing and stage-management for a year at the *Musikhochschule* in Weimar. He moved from Apolda to Weimar after about three months. At the Weimar *Institut zur Methodologischen Erneuerung des Deutschen Theaters* he studied theatre, and at the Weimar academy he took classes in Italian, harmony, piano,

singing and enunciation. Singing remained one of his life's major interests. He attended performances (drama, opera, ballet, concerts) at Weimar's *Nationaltheater*, and later reminisced about opera stars of the time, recalling debates and arguments about theatre (e.g. the stereotyping of roles and plays) with Maxim Vallentin, Hans Eisler, etc. He also played a small part in one of the films of G.W.Pabst, a notable German film-director. Although, by his own account, he led a full life, he became restless and decided to move.

Return to Vienna: University Life, Alpbach, and Popper (1947-1948).

Feyerabend therefore returned, still on crutches, to his parents' apartment house in Vienna's 15th district. Although he planned to study physics, maths and astronomy, he chose instead to read history and sociology at the University of Vienna's *Institut für Osterreichische Geschichtsforschung*, thinking that history, unlike physics, is concerned with real life. But he became dissatisfied with history, and returned to theoretical physics. Together with a group of science students, who all regarded themselves as far superior to students of other subjects, Feyerabend invaded philosophy lectures and seminars. Although this was not his first contact with philosophy, it seems to have been the period which cemented his interest. He recalls that in all interventions he took the radical positivist line that science is the basis of knowledge; that it is empirical; and that nonempirical enterprises are either logic or nonsense (p.68). These views would have been familiar from the climate of Logical Positivism which found its main root in the Vienna Circle, a group of scientifically-minded philosophers who, in the nineteen-twenties and 'thirties sought to deploy the newly-revitalised formal logic of Gottlob Frege and Russell and Whitehead's *Principia Mathematica* to represent the structure of human knowledge. As we shall see, Feyerabend's youthful positivist scientism makes quite a contrast with his later conclusions.

In August 1948, at the first meeting of the international summer seminar of the Austrian College Society in Alpbach which he attended, Feyerabend met the philosopher of science Karl Popper, who had already made a name for himself as the Vienna Circle's 'official opposition'. (The Austrian College Society had been founded in 1945 by Austrian resistance fighters, 'to provide a forum for the exchange of scholars and ideas and so to prepare the political unification of Europe' (*Science in a Free Society*, p.109)). In his 1934 book *Logik der Forschung* Popper had elaborated the straightforward and appealing falsificationist view that great science could be characterised as a process in which thinkers put forward bold conjectures and then do their best to improve them by trying to refute them. Instead of trying to develop an inductive logic, Popper argued for the (deductivist) view that scientific method could be characterised in terms of logically valid deductive inferences.

Popper's own autobiography, unfortunately, tells us nothing about their meeting or their relationship, despite the fact that he was to be the largest single influence (first positive, then negative) on Feyerabend's work. For those hoping that Feyerabend might use the occasion of his autobiography to settle accounts with his erstwhile philosophical conscience, it is disappointing that the book tells us so little about his acquaintance with Popper. Elsewhere Feyerabend tells us that he

> admired [Popper's] freedom of manners, his cheek, his disrespectful attitude towards the

> German philosophers who gave the proceedings weight in more senses than one, his sense of humour... [and] his ability to restate ponderous problems in simple and journalistic language. Here was a free mind, joyfully putting forth his ideas, unconcerned about the reaction of the 'professionals'. (*SFS*, p.115).

But Popper's ideas themselves, Feyerabend alleges, were not new to him, deductivism having been defended as early as 1925 by Viktor Kraft, and falsificationism being 'taken for granted' at Alpbach. Popper's ideas, he remarks, were also similar to those of another Viennese philosopher, Ludwig Wittgenstein (!), although 'more abstract and anaemic' (*SFS*, p.116). Over the following years, Feyerabend attended the Alpbach symposium about fifteen times, first as a student, then as a lecturer and seminar chair. He was offered, and accepted, the post of 'scientific secretary' to the society, and this he calls 'the most decisive step of my life' (p.70). In fact, it is this decision which answers his self-addressed questions about the origin of his career, his reputation, and his situation at the time of writing his autobiography, since he traces his situation back to it.

At Alpbach he was also approached by communists, including the Marxist intellectual Walter Hollitscher, who became his teacher and friend. Feyerabend resisted Hollitscher's political arguments on the basis of his own 'youthful elitism' and 'an almost instinctive aversion to group thinking' (p.73). But although Feyerabend later described himself as having been 'a raving positivist' at the time, it was Hollitscher, he says, who persuaded him of the cogency of realism about the 'external world' (Popper's important arguments for realism came somewhat later). The considerations Hollitscher deployed were, first, that scientific research was conducted on the assumption of realism, and could not be otherwise conducted, and, second, that realism is fruitful and productive of scientific progress, whereas positivism was simply a commentary on scientific results, barren in itself.

> Hollitscher never presented an argument that would lead, step by step, from positivism to realism and he would have regarded the attempt to produce such an argument as philosophical folly. He rather developed the realist position itself, illustrated it by examples from science and commonsense, showed how closely it was connected with scientific research and everyday action and so revealed its strength. (*SFS*, p.113).

Feyerabend eventually developed these thoughts in a fascinating series of papers beginning in 1957, arguing that science needs realism in order to progress, and that positivism would stultify such progress. The argument was entirely in line with Popper's approach, as well as with his conclusions.

Early Contact with Wittgenstein (1948-1952).

Feyerabend's principal intellectual engagement in the late 1940s and early 1950s was in his capacity as student leader of the 'Kraft Circle'. Viktor Kraft was a former member of the Vienna Circle, and became Feyerabend's dissertation supervisor. The Kraft Circle was a philosophy club centred around Kraft, which constituted another part of the Austrian College Society. Bela Juhos, Walter Hollitscher, Georg Henrik von Wright, Elizabeth Anscombe and Wittgenstein were all visiting speakers. Feyerabend reports

that the Circle held meetings from 1949 to 1952 or ‘53 (*SFS*, p.109), that they set themselves the task of ‘considering philosophical problems in a nonmetaphysical manner and with special reference to the findings of the sciences’ (‘Herbert Feigl: A Biographical Sketch’, in P.K.Feyerabend & G.Maxwell (eds.), *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*, (Minneapolis: University of Minnesota Press, 1966), pp.1-2) and that their main topics of discussion were the questions of the reality of theoretical entities and of the ‘external world’. About Wittgenstein's lecture, Feyerabend recalls the following:

> Not even a brief and quite interesting visit by Wittgenstein himself (in 1952) could advance our discussion. Wittgenstein was very impressive in his way of presenting concrete cases, such as amoebas under a microscope... but when he left we still did not know whether or not there was an external world, or, if there was one, what the arguments were in favour of it. (Feyerabend & Maxwell ibid., p.4. Note that Feyerabend must have got the date wrong, since Wittgenstein died in April 1951).
>
> Wittgenstein, who took a long time to make up his mind and then appeared over an hour late gave a spirited performance and seemed to prefer our disrespectful attitude to the fawning admiration he encountered elsewhere. (*SFS*, p.109).

In 1949, Feyerabend was introduced to Bertholt Brecht, and Hollitscher offered him the opportunity to become one of Brecht's production assistants, but he turned it down, later describing this as one of the biggest mistakes of his life (*SFS*, p.114). In the autobiography, however, he retracts this statement, saying that he would not have enjoyed being part of the closely knit group that surrounded Brecht. (The reasons for his later defection from the Popperian camp seem to have been similar).

The University of Vienna's physicists were Hans Thirring, Karl Przibram, and Felix Ehrenhaft. Feyerabend admired Thirring and Ehrenhaft, and was influenced by Ehrenhaft, who had lectured on physics there from 1947. Ehrenhaft was known as a fierce and independent critic of all kinds of orthodoxy in physics, but was sometimes thought of as a charlatan. Feyerabend reports that he and his fellow science students looked forward to exposing him as a fraud, but in fact were treated, at the 1949 Alpbach seminar, to a battle between Ehrenhaft and the orthodox in which the former presented his experiments but the latter defended their position by using strategies which Galileo's opponents would have been proud of, ridiculing Ehrenhaft's phenomena as mere *Dreckeffects*. Feyerabend commented that "Only much later did Ehrenhaft's lesson sink in and our attitude at the time as well as the attitude of the entire profession provided me then with an excellent illustration of the nature of scientific rationality" (*SFS*, p.111). Ehrenhaft did not convince the theoreticians, who protected themselves with an iron curtain of dogmatic belief of exactly the same kind as that deployed by Galileo's opponents. His audience remained staunch empiricists, never doubting that science had to be adapted to facts. Feyerabend commented that the day-to-day business of science, what Thomas Kuhn called ‘normal science’, cannot exist without this kind of ‘split consciousness’.

At the University of Vienna, although he had originally planned to submit a thesis on physics,

Feyerabend swapped to philosophy when he got nowhere with the electrodynamics problem he was calculating (the philosopher of science as failed scientist?). He completed his doctoral thesis, '*Zur Theorie der Basissätze*' in 1951 under Kraft's supervision. The subject of the thesis was 'basic sentences', or 'protocol sentences', i.e. the kind of sentences that, the Logical Positivists had theorised, comprise the foundations of scientific knowledge. He later reported that in his philosophical work he had "started from and returned to the discussion of protocol statements in the Vienna Circle" ('Concluding Unphilosophical Conversation', in Munévar (1991), p.526). This is unsurprising, given that Kraft was then the Vienna Circle's only survivor in Vienna. However, Kraft's influence on Feyerabend has only recently been emphasised. Much of the material from Feyerabend's thesis was presented at (or gleaned from) meetings of the Kraft Circle, and also appears in his early articles, such as 'An Attempt at a Realistic Interpretation of Experience' (1958). The thesis itself was 'a condensed version of the discussions in the Kraft Circle' (p.115).

In the early 1950's, Feyerabend published several German papers on Wittgenstein, written as a result of having read the proofs of the *Philosophical Investigations*, lent to him by Elizabeth Anscombe. Feyerabend first met Anscombe when lecturing on Descartes to the Austrian College Society. Anscombe had come to Vienna to perfect her German in order to translate Wittgenstein's works.

> She gave me manuscripts of Wittgenstein's later writings and discussed them with me. The discussions extended over months and occasionally proceeded from morning over lunch until late into the evening. They had a profound influence upon me though it is not at all easy to specify particulars. (*SFS*, p.114).

Feyerabend planned to study with Wittgenstein in Cambridge, and Wittgenstein was prepared to take him on as a student, but he died before Feyerabend arrived in England. Karl Popper became his supervisor instead.

Life at the London School of Economics (1952-1953).

In Feyerabend's autobiography, we are told a little about Popper's lectures and his famous LSE seminar. The lectures began with the claim that there is no method in science, but that there *are* some simple and helpful rules of thumb. Popper tried to show "how simple ideas that were derived from equally simple requirements brought order into the complex world of research" (pp.88-9). Having being convinced by Popper's and Pierre Duhem's critiques of inductivism (the view that science proceeds through generalisation from facts recorded in experience), Feyerabend considered falsificationism a real option, and, he says, "fell for it" (p.89), applying falsificationism in his papers and lectures. This is not his first admission that he was a falsificationist, but it is notable that he did not see it as entailing his having been a Popperian. Feyerabend was (usually) a fairly *liberal* falsificationist, always emphasising the tenacity with which scientists should defend their theories, and allowing that scientific theories can start by being untestable. Faithful Popperians like John Watkins and Joseph Agassi, he emphasises, continually ticked him off for being unorthodox (he was later accused, by Agassi, of plagiarising from Popper). Instead he later saw this interlude as an example of the dangers of abstract reasoning. Rationalism is already

dangerous, since it ‘paralyses our judgment’ (p.89) and is invested with ‘an almost superhuman authority’ (p.90). But Popper added a further dangerous element: *simplicity*. Such a philosophy, complains Feyerabend, "may be out of touch with reality... [that is], with scientific practice" (p.90).

Feyerabend is here referring to Popper's approach to the epistemology of science, which he himself followed and furthered for quite a while. In chapter II of *The Logic of Scientific Discovery* (1934), Popper had distinguished between scientific practice and scientific standards, principles, or methodology. Arguing against a ‘naturalistic’ theory of method which makes standards depend on practice, Popper opted instead for a strongly normative epistemology, a discipline which lays down optimum rules of method for scientists to follow. This is one of the most important aspects of the Popperian perspective which Feyerabend originally took on board.

Such an epistemology, Feyerabend now complains, makes the false assumption that ‘rational’ standards can lead to a practice that is as mobile, rich and effective as the science we already have. Falsificationism would destroy science as we know it. Science did not develop in accordance with Popper's model. It is not ‘irrational’, but it contains no overarching pattern. Popper's rules could produce a science, but not the science we now have. (Feyerabend remarks that the Logical Positivist Otto Neurath had already put this criticism of Popper some time before (p.91)).

In 1952, Feyerabend presented his ideas on scientific change to Popper's LSE seminar and to a gathering of illustrious Wittgensteinians (Elizabeth Anscombe, Peter Geach, H.L.A.Hart and Georg Henrik von Wright) in Anscombe's Oxford flat. This meeting seems to have been the first airing of the important concept of *incommensurability* (although not the term itself, which crept into publications only a decade later):

> On one occasion which I remember vividly Anscombe, by a series of skilful questions, made me see how our conception (and even our perceptions) of well- defined and apparently self-contained facts may depend on circumstances not apparent in them. There are entities such as physical objects which obey a ‘conservation principle’ in the sense that they retain their identity through a variety of manifestations and even when they are not present at all while other entities such as pains and after-images are ‘annihilated’ with their disappearance. The conservation principles may change from one developmental stage of the human organism to another and they may be different for different languages (cf. Whorf's ‘covert classifications’...). I conjectured that such principles would play an important role in science, that they might change during revolutions and that deductive relations between pre-revolutionary and post-revolutionary theories might be broken off as a result. (*SFS*, p.115).
>
> Major discoveries, I said, are not like the discovery of America, where the general nature of the discovered object is already known. Rather, they are like recognizing that one has been dreaming. (*KT*, p.92).

These thoughts received an unenthusiastic reception from Hart, von Wright and Popper.

Feyerabend's articles on Wittgenstein culminated in his review of the *Philosophical Investigations*, the text of which he studied in detail while he was in London. (‘Being of a pedantic turn of mind’, he says, ‘I rewrote the book so that it looked more like a treatise with a continuous argument’. (*SFS*, p.116)). Anscombe translated Feyerabend's summary into English and sent it to *The Philosophical Review*. It was accepted by the editor, Norman Malcolm (having been turned down by Gilbert Ryle, editor of *Mind* - see *KT*, p.93). This review was Feyerabend's first English-language publication; he called it his ‘Wittgensteinian monster’ (p.115). He later commented:

> I knew that Wittgenstein did not want to present a theory (of knowledge, or language), and I did not expressly formulate a theory myself. But my arrangements made the text speak like a theory and falsified Wittgenstein's intentions. (*KT*, p.93).
>
> Wittgenstein's emphasis on the need for concrete research and his objections to abstract reasoning (‘Look, don't think!) somewhat clashed with my own inclinations and the papers in which his influence is noticeable are therefore mixtures of concrete examples and sweeping principles. (*SFS*, p.115).

In his review of the *Philosophical Investigations*, he summarised the book in a very effective way, drawing particular attention to Wittgenstein's critique of a family of ‘realist’ or ‘essentialist’ theories of meaning according to which the meaning of a word is the object designated or referred to by that word. Feyerabend argued that Wittgenstein was attempting a *reductio ad absurdum* of realist theories, showing that they had the untenable implication that we could not be said to know the meaning of words which we nevertheless constantly use in totally unproblematic ways.

Unfortunately, as I have argued at length elsewhere, (Preston 1997, ch.2), Feyerabend completely failed to follow up this insight by endorsing Wittgenstein's non-representationalist conception of meaning, according to which the meaning of a term is determined by its *use*. Instead, wrongly associating the idea that meaning is use with positivism, Feyerabend proferred what he called a ‘contextual’ theory of meaning, which identified the meaning of a term or statement with whatever role it plays in *theoretical* contexts. But he over-extended the idea of the theoretical to cover *any* context whatever, thus completely depriving it of content. For Feyerabend, the theoretical contrasts with nothing at all.

The book review was also critical of Wittgenstein, though. Notably, it railed against Wittgenstein's conception of philosophy (as ‘philosophical analysis’). In a short article published the next year (1956), Feyerabend expanded on his critique, arguing that consideration of G.E.Moore's famous ‘paradox of analysis’ showed that "*philosophy cannot be analytic and scientific*, i.e., interesting, progressive, about a certain subject matter, informative *at the same time*" (‘A Note on the Paradox of Analysis’, p.95). Feyerabend thenceforth plumped for (what he conceived of as) *scientific* philosophy. Like Popper, he had very little time for the kind of ‘analytic’ philosophy or ‘linguistic’ philosophy which followed in Wittgenstein's wake, and with which Oxford University dominated the philosophical scene in the 1950s

and early 1960s.

One of the things that comes across most clearly from his autobiography is the consistently malleable nature of Feyerabend's views. He records that his friend Agassi caused him completely to change his mind about a book he considered translating. When Agassi urged Feyerabend to become a faithful Popperian, Feyerabend's resistance seems to have been based mainly on his aversion to groups.

Return to Vienna (1953-1955).

By the summer of 1953, when Popper had to apply for extra funds to allow Feyerabend to work as his assistant, Feyerabend had decided to leave the Popperian church and return to Vienna. Although the assistantship was soon approved, Feyerabend ‘felt quite uncomfortable. I couldn't put my finger on it; I only knew that I wanted to remain in Vienna’ (p.99).

During this period Feyerabend, having nothing to do and needing the money, translated Popper's ‘war effort’, *The Open Society and its Enemies* into German, wrote articles on ‘Methodology’ and ‘Philosophy of Nature’ for a French encyclopaedia, produced a report on post-war developments in the Humanities in Austria for the U.S. Library of Congress, and made a mess of his first professional opportunity as a singer (p.98). But he also felt that he did not know what to do in the long run, so he applied for jobs in various universities.

He then met Arthur Pap, ‘who had come to Vienna to lecture on analytic philosophy and who hoped, perhaps somewhat unrealistically, that he would be able to revive what was left from the great years of the Vienna Circle and the analytic tradition there’ (‘Herbert Feigl: A Biographical Sketch’, p.3). Feyerabend became Pap's assistant. Pap arranged for him to meet Herbert Feigl in Vienna in 1954, and together they studied Feigl's papers. Feigl had been a member of the Vienna Circle until his emigration to the USA in 1930, but he had never given up the ‘realist’ view that there is a knowable external world. He convinced Feyerabend that the positivism of Kraft and Pap had not solved the traditional problems of philosophy. His paper ‘Existential Hypotheses’ (1950), together with Kraft's contributions and certain ideas Popper had put forward at Alpbach in 1948 and 1949, greatly diminished Feyerabend's doubts about realism (ibid., p.4). Here is how Feyerabend recounts Feigl's influence:

> It was ... quite a shock to hear Feigl expound fundamental difficulties and to hear him explain in perfectly simple language without any recourse to formalism why the problem of application [of the probability-calculus] was still without a solution. Formalization, then, was not the last word in philosophical matters. There was still room for fundamental discussion-for speculation (dreaded word!); there was still a possibility of overthrowing highly formalized systems with the help of a little common sense! (ibid., p.5).

1954 saw the publication of the first of Feyerabend's many articles on the philosophy of quantum mechanics, the first fruits of the time he spent studying with Popper. In these publications, he generally took the line that the dominance achieved by the ‘Copenhagen Interpretation’ of the quantum theory was

undeserved. Feyerabend was particularly keen to argue that it had not and could not be shown that this interpretation of the theory was a general panacea for the problems of microphysics, or that its defenders could justifiably believe it to be unassailable. He came to defend the right of ‘hidden-variables’ theorists such as Louis de Broglie, David Bohm, and Jean-Pierre Vigier to hypothesise the existence of an unobserved deterministic substructure underpinning the apparently indeterministic cavortings of objects on the quantum-mechanical level.

However, Feyerabend also came to think that Popper's earlier critique of the Copenhagen orthodoxy had been somewhat limited and superficial. According to Popper, the Copenhagen Interpretation was simply the result of some bad positivistic philosophising. Niels Bohr and Werner Heisenberg, on this view, had been seduced by positivist philosophers (like Ernst Mach and his ostensible followers, the Vienna Circle) into thinking that their theory was not conjectural but was merely a compendious, economical but non-hypothetical *description* of experience. Feyerabend argued that, on the contrary, the Copenhagen theorists had some perfectly good ‘physical’, ‘scientific’, or ‘factual’ arguments for thinking that their view alone was currently compatible with the observed results of experiments. He therefore put forward a *defence* of their instrumentalist interpretation of the quantum theory. But the defence was only tactical, since he ultimately argued that the observed results of experiments themselves needed to be *challenged* by a point of view which would reveal their truth or falsity. So Feyerabend used the quantum case to push for a reconsideration of the methodological rules to which scientists subscribe. This is the genesis of his idea of a ‘pluralistic’ test model, in which theories are compared with one another, as well as against ‘experience’. (Note, however, that this idea can already be found in Popper, and that Feyerabend did initially acknowledge this fact). According to Feyerabend, only by endorsing scientific realism can the scientist cleave to a methodology which would consistently bring out the (conceptually) *revolutionary* potential of scientific theories, rekindling the kind of fire Galileo had lit under the Aristotelian world-view. Such a realism interprets theories *not* as summaries of experience, but as genuine conjectures about a mind-independent reality. It also puts the observation-language of science in the same epistemological boat as its theoretical terms: observations, he urged, are just as ‘theoretical’ (that is, hypothetical) as theories: "Logically speaking, all terms are ‘theoretical’" (*Philosophical Papers, Volume 1*, p.32 note).

First Academic Appointment: the University of Bristol (1955-1958).

In 1955, with the help of references from Popper and Erwin Schrödinger, as well as his own big mouth (*SFS*, p.116, *KT*, p.102), Feyerabend secured his first academic post lecturing in philosophy of science at the University of Bristol, England. In his autobiography (pp.103-4) he describes how Agassi had to help him prepare for these lectures, since they covered a subject Feyerabend had never studied (see also *SFS*, p.116). He also describes how for some time he felt directionless and unsettled: he was ‘killing time’.

In the summer of that year, he again visited Alpbach, where he met the philosopher of science Philipp Frank (another former Logical Positivist), who exerted on him a (somewhat delayed) influence:

> Frank argued that the Aristotelian objections against Copernicus agreed with empiricism, while Galileo's law of inertia did not. As in other cases, this remark lay dormant in my

> mind for years; then it started festering. The Galileo chapters of *Against Method* are a late result. (*KT*, p.103. See also *SFS*, p.112).

Around the same time, Feyerabend met David Bohm, who was lecturing in physics at the University of Bristol. Bohm had been the favoured protegé of Niels Bohr, and his first book (*Quantum Theory*, (Englewood Cliffs, NJ: Prentice-Hall, 1951))was a lengthy defence of the Copenhagen Interpretation of the quantum theory. But in the early 1950s Bohm rejected his former view, and became one of the leading defenders of the then-unpopular ‘hidden-variables’ theory. He was to be a significant influence on Feyerabend, weaning him away from Popper with his somewhat Hegelian account of the structure of reality. In their later work, Bohm and Feyerabend moved in parallel directions, towards an interest in ‘fringe’ science. Feyerabend produced a critical study of Bohm's 1957 book *Causality and Chance in Modern Physics* in 1960, when he was still very much under Popper's influence. But, as always with Feyerabend, Bohm's ideas sunk in gradually, and had visible effect only in his productions of the early 1970s.

In 1956, Feyerabend got married for the second time, this time to one of his former students, Mary O'Neill. But this relationship seems to have been very short-lived, for he reports that his wife spent Christmas 1957 away from him with her parents, that she subsequently had an affair, and that the last time he saw her was 1958.

Feyerabend remembers his Bristol lecture course on quantum mechanics as being a disaster. However, in the summer of 1956, along with Alfred Landé, he chaired a successful seminar on philosophical issues in quantum mechanics at Alpbach. A related success was his contribution to the 1957 Colston Research Symposium, where he gave a paper ‘On the Quantum Theory of Measurement’. Here Feyerabend introduced what was to become a long-running theme in his work: that there is no separate and neutral ‘observation-language’ or ‘everyday language’ against which the theoretical statements of science are tested, but that ‘the everyday level is *part of* the theoretical rather then something self-contained and independent’ (*Philosophical Papers, Volume I*, p.217, emphasis added). This was his principal contribution to his central subject, the relation between theory and experience. It constituted not only a decisive break with the positivist conception of theories, but also something of a step beyond Popper's conception.

The University of California at Berkeley: Early Years (1958-1964).

In the summer of 1957, Feyerabend accepted an invitation from Michael Scriven to visit the Minnesota Center for the Philosophy of Science in Minneapolis. The Center was, as Feyerabend later said, ‘one of the foremost institutions in the field’ (p.115). There he met Feigl, Carl Hempel, Ernest Nagel, Hilary Putnam, Adolf Grünbaum, Grover Maxwell, E.L.Hill, Paul Meehl, and others. He returned to the Center in 1958, having accepted another invitation to work there, backed by an NSF grant. He often went back there in subsequent years.

Around this time, many of Feyerabend's most important early papers were published. In them, under the

influence of both Popper and Wittgenstein, Feyerabend initiated a vigourous critique of the then-orthodox philosophies of science provided by descendants of the Vienna Circle, ‘Logical Empiricist’ thinkers such as Rudolph Carnap, Feigl, Nagel, and Hempel. This critique was conducted through a study of the relationship between observation and theory.

In perhaps the most important of these early publications, ‘An Attempt at a Realistic Interpretation of Experience’ (1958), Feyerabend argued against positivism and in favour of a scientific realist account of the relation between theory and experience, largely on grounds familiar from Karl Popper's falsificationist views. Positivist theories of meaning, he complained, have consequences which are ‘at variance with scientific method and reasonable philosophy’ (*Philosophical Papers, Volume 1*, p.17). In particular, they imply what Feyerabend dubbed the ‘stability thesis’, that even major changes in theory will not affect the meanings of terms in the scientific observation-language. Against this supposition, Feyerabend defended what he called ‘Thesis I’, the idea that

> *the interpretation of an observation-language is determined by the theories which we use to explain what we observe, and it changes as soon as those theories change*. (ibid., p.31).

Thesis I reversed the direction of interpretation which the positivists had presupposed. Instead of meaning seeping upwards from the level of experience (or the observation-language), Feyerabend had it trickling down *from* theory *to* experience. For him, theory is meaningful independently of experience, rather than vice-versa. The roots of this view clearly lie in his contextual theory of meaning, according to which meaning is conferred on terms by virtue of their participation in *theoretical* contexts. It seems to imply that there is no principled semantic distinction between theoretical terms and observation terms. And Feyerabend soon followed up this implication with his ‘Pragmatic Theory of Observation’, according to which what is important about observation-sentences is not their having a special core of empirical meaning, but their causal role in the production and refutation of theories.

In 1958, Feyerabend had been invited to spend one year at the University of California at Berkeley, and accepted. When this visiting appointment ended, the University administration decided to hire him on the basis of his publications and, of course, his big mouth (p.115). But because of his grant to work at Minneapolis, he only started lecturing full-time at Berkeley in 1960. There he encountered Thomas Kuhn, and read Kuhn's forthcoming book *The Structure of Scientific Revolutions* in draft form. He then wrote to Kuhn about the book (these letters have recently been published in *Studies in the History and Philosophy of Science*, **26**, 1995). But he was not quite ready to take on Kuhn's descriptive-historical approach to the philosophy of science. Although more and more historical examples peppered his published work, he was still using them to support fairly orthodox falsificationist conclusions.

In his meta-methodology, Feyerabend applied to the dispute over the interpretation of scientific theories a strong measure of Popperian methodological conventionalism, arguing that the dispute between realists and instrumentalists is not a factual issue but a matter of choice. We can choose to see theories either as descriptions of reality (scientific realism) or as instruments of prediction (instrumentalism), depending on what ideals of scientific knowledge we aspire to. Adherence to these competing ideals (roughly: high

informative content on the one hand, and sense-certainty on the other) is to be judged by their respective consequences. Stressing that philosophical theories have not merely reflected science but have *changed* it, Feyerabend argued further that the *form* of our knowledge can be altered to fit our ideals. So we can *have* certainty, and theories that merely summarise experience, if we wish. But, mobilising the usual equation between empirical content and testability (common to Carnap, Popper and Feyerabend), he urged that we should decisively reject the ideal of certainty and opt instead for theories which go beyond experience and say something informative about reality itself. In this respect, he clearly followed Popper's lead, reconstruing empiricism as a doctrine about the most desirable form for our theories, rather than as a view about the sources of knowledge.

Feyerabend argued that the idea, common to positivists, that the interpretation of observation terms doesn't depend upon the status of our theoretical knowledge, has consequences undesirable to positivists. One of these is that "every positivistic observation language is based upon a metaphysical ontology" (*Philosophical Papers, Volume 1*, p.21). Another follows from the thesis, which he relishes, that the theories we hold influence our language, and maybe even our perceptions. This implies that as long as we use only one empirically adequate theory, we will be unable to imagine alternative accounts of reality. If we also accept the positivist view that our theories are summaries of experience, those theories will be void of empirical content and untestable, and hence there will be a diminution in the critical, argumentative function of our language. Just as purely transcendent metaphysical theories are unfalsifiable, so too what began as an all-embracing scientific theory offering certainty will, under these circumstances, have become an irrefutable dogma, a *myth*. Elsewhere (Preston 1997, chapter 5) I have argued that his antipathy toward this ‘myth predicament’ was one the main driving-forces behind Feyerabend's views at the time.

Feyerabend defended a realism according to which "the interpretation of a scientific theory depends upon nothing but the state of affairs it describes" (*Philosophical Papers, Volume 1*, p.42). At the same time he claimed to find in Wittgenstein's *Philosophical Investigations* a contextual theory of meaning according to which the meaning of terms is determined not by their use, nor by their connection with experience, but by the role they play in the wider context of a theory or explanation. Thesis I, the key proposition of Feyerabend's early work, is supposed to encapsulate *both* the contextual theory of meaning and scientific realism. Only realism, by insisting on interpreting theories in their most vulnerable form as universally-quantified statements which strive for truth, leads to scientific progress instead of stagnation, he argued. Only realism allows us to live up to the highest intellectual ideals of critical attitude, honesty, and testability.

Unlike positivism, which conflicts with science by taking experiences as unanalysable building-blocks, realism treats experiences as analysable, explaining them as the result of processes not immediately accessible to observation. Experiences and observation-statements are thus revealed as more complex and structured than positivism had realised. Feyerabend over-extended the contextual theory of meaning to apply not only to theoretical terms but to observation terms too, arguing that there is no special ‘problem’ of theoretical entities, and that the distinction between observation terms and theoretical terms is a purely pragmatic one. If, as the contextual theory also implies, observation-statements depend on theoretical principles, any inadequacy in these principles will be transmitted to the observation-

statements they subtend, whence our beliefs about what is observed may be in error, and even our experiences themselves can be criticised for giving only an approximate account of what is going on in reality. *All* our statements, beliefs and experiences are ‘hypothetical’. Observations and experiments always need interpretation, and different interpretations are supplied by different theories. If existing meanings embody theoretical principles, then instead of passively accepting observation- statements, we should attempt to find and test the theoretical principles implicit in them, which may require us to change those meanings.

Feyerabend therefore idolised semantic instability, arguing that the semantic stability presupposed by positivist accounts of reduction, explanation and confirmation, has been and should be violated if we want progress in science. If meaning is determined by theory, terms in very different theories simply cannot share the same meaning: they will be ‘incommensurable’. Any attempt to derive the principles of an old theory from those of a new one must either be unsuccessful or must effect a change in the meaning of the old theory's terms. The ‘theoretical reduction’ beloved of Logical Empiricists is therefore actually more like *replacement* of one theory and its ontology by another. At the end of his well-known 1962 paper ‘Explanation, Reduction, and Empiricism’, in which he introduced the concept of incommensurability, Feyerabend concluded that this concept precluded any formal account of explanation, reduction or confirmation. (Kuhn's book *The Structure of Scientific Revolutions*, in which the same term was used to characterize a related concept, was published in the same year).

In his first major published excursion from the philosophy of science, Feyerabend applied these ideas to the mind/body problem. In two papers published in 1963, he sought to defend materialism (roughly, the view that everything which exists is physical) against the supposition that the mind cannot be a physical thing. Although these papers exhibit a rather unclear mixture of views, they are now remembered primarily for having ushered in the position known as *‘eliminative materialism’*, according to which our way of conceiving the mind and mental phenomena amounts to a seriously inadequate theory which is in conflict with a (materialistic) scientific account of those same things. Feyerabend suggested that the two theories in question were incommensurable, but that nevertheless we ought to prefer the materialistic one on general methodological grounds. This radical view of the mind/body problem has been one of Feyerabend's most important legacies. Even though Feyerabend himself seems to have given it up in the late 1970s, it was taken up by Richard Rorty and, more recently, by Paul and Patricia Churchland.

In Feyerabend's version of the incommensurability thesis, the semantic principles of construction underpinning a theory (in its *realist* interpretation) can be violated or ‘suspended’ by another theory. As a result, theories cannot always be compared with respect to their content, as ‘rationalists’ would like. It took Feyerabend a while to see it, for he did *not* officially subscibe to this view until the late 1960s, but this opens the door to relativism, the view that there is no objective way of choosing between theories or traditions. This is perhaps the most notorious and widely-reviled consequence of the contextual theory of meaning.

In the ground-breaking central papers from this period of Feyerabend's oeuvre such as ‘How to be a Good Empiricist’ (1963), ‘Realism and Instrumentalism’ (1964), ‘Problems of Empiricism’ and ‘Reply to Criticism’ (1965), his most important argument for scientific realism was methodological: realism is

desirable because it demands the proliferation of new and incompatible theories. This leads to scientific progress because it results in each theory having more empirical content that it otherwise would, since a theory's testability is proportional to the number of potential falsifiers it has, and the production of alternative theories is the only reliable way to ensure the existence of potential falsifiers. So scientific progress comes through *‘theoretical pluralism’*, allowing a plurality of incompatible theories, each of which will contribute by competition to maintaining and enhancing the testability, and thus the empirical content, of the others. According to Feyerabend's *pluralistic test model*, theories are tested against one another. He thus idealised what Kuhn called ‘pre-paradigm’ periods and ‘scientific revolutions’, occasions when there are many incompatible theories, all forced to develop through their competition with each other. But he downplayed the idea that theories are still compared with one another primarily for their ability to account for the results of observation and experiment. For Feyerabend, this idea was an empiricist myth which disguised the role of aesthetic and social factors in theory-choice.

Thus far, the argument for theoretical pluralism largely follows that of John Stuart Mill's *On Liberty* (1859), to which Feyerabend often paid homage. But Feyerabend went on to try to demonstrate a mechanism whereby theories can augment their empirical content. According to this part of the argument, theories may face difficulties which can only be discovered with the help of alternative theories. A theory can be incorrect without our being able to discover this in a direct way: sometimes the construction of new experimental methods and instruments which would reveal the incorrectness is excluded by laws of nature, sometimes the discrepancy (were it to be discovered) might be regarded as an oddity, and might never be given its correct interpretation. Circumstances can thus conspire to hide from us the infirmities of our theory. The methodological ‘principle of testability’ demands that we develop alternative theories incompatible with the existing theory, and develop them in their strongest form, as descriptions of reality, not mere instruments of prediction. Instead of waiting until the current theory gets into difficulties, and only then starting to look for alternatives, we ought vigorously to proliferate theories and tenaciously defend them in the hope that they may afford us an *indirect refutation* of our existing theory. Only theories which are empirically adequate will thus contribute to raising the empirical content of their fellows. But Feyerabend insists that *any* theory, no matter how weak, may *become* empirically adequate, and so may contribute to this process. To be a realist, he therefore suggests, involves demanding support for any theory, including implausible conjectures having no independent empirical support, conjectures which are inconsistent with data and well-confirmed laws. We should retain theories that are in trouble, and invent and develop theories that contradict the observed phenomena, just because in doing so we will be respecting the intellectual ideal of testability.

In thus appealing to the ‘principle of testability’ as the supreme methodological maxim, Feyerabend forgets that testability must be traded-off against other theoretical virtues. Only his pathological fear of theories losing their empirical content and becoming myths leads him to want to maximise testability and embrace an absolutely unrestricted principle of proliferation. He also disregards historical evidence that anti-realist approaches can be just as pluralistic as realism.

At Alpbach in 1964, Feyerabend and Feigl jointly directed a seminar on the recent development of analytic philosophy. There Feyerabend re-encountered the leading light of the Logical Positivist movement, Rudolph Carnap (whom he had already met at UCLA). Carnap tried to convince Feyerabend

of the virtues of clarity, but failed. Feyerabend was still attached to ‘scientific’ philosophy, and considered philosophy worthless unless it made a positive and quantifiable contribution to the growth of knowledge (which, of course, meant science).

But a seminar in Hamburg in 1965, at which Feyerabend discussed the foundations of quantum theory with the physicist C.F. von Weizsäcker, did have a lasting, if somewhat delayed, impact:

> Von Weizsäcker showed how quantum mechanics arose from concrete research while I complained, on general methodological grounds, that important alternatives had been omitted. The arguments supporting my complaint were quite good... but it was suddenly clear to me that imposed without regard to circumstances they were a hindrance rather than a help: a person trying to solve a problem whether in science or elsewhere *must be given complete freedom* and cannot be restricted by any demands, norms, however plausible they may seem to the logician or the philosopher who has thought them out in the privacy of his study. Norms and demands must be checked by research, not by appeal to theories of rationality. In a lengthy article I explained how Bohr had used this philosophy and how it differs from more abstract procedures. Thus Professor von Weizsäcker has prime responsibility for my change to ‘anarchism’ - though he was not at all pleased when I told him so in 1977. (*SFS*, p.117).

The Impact of the ‘Student Revolution’.

The mid-to-late 1960s was a time of ferment in Western culture, and Feyerabend was in the thick of it. In Berkeley, naturally, he ran into the Free Speech Movement, and he encountered the ‘student revolution’ there too, as well as in London and Berlin. This obviously fired his interest in political philosophy, more especially in political questions about science. Of his post at Berkeley, he later said:

> My function was to carry out the educational policies of the State of California which means I had to teach people what a small group of white intellectuals had decided was knowledge. (*SFS*, p.118).

However, Feyerabend's experience under these educational policies was undoubtedly one of the defining periods of his intellectual life, a time in which he became deeply suspicious of these intellectuals and ‘Western rationalism’ as a whole:

> In the years 1964ff. Mexicans, Blacks, Indians entered the university as a result of new educational policies. There they sat, partly curious, partly disdainful, partly simply confused hoping to get an ‘education’. What an opportunity for a prophet in search of a following! What an opportunity, my rationalist friends told me, to contribute to the spreading of reason and the improvement of mankind! I felt very differently. For it dawned on me that the intricate arguments and the wonderful stories I had so far told to my more or less sophisticated audience might just be dreams, reflections of the conceit of a small

> group who had succeeded in enslaving everyone else with their ideas. Who was I to tell these people what and how to think? (ibid. See also *KT*, p.123).

At this time, Feyerabend gave two lectures, one on general philosophy, and one on philosophy of science. He seems to have got into some trouble at Berkeley by running his seminar on unacceptably loose lines, regularly cancelling lectures, and failing to prepare for the lectures he did give:

> I often told the students to go home--the official notes would contain everything they needed. As a result an audience of 300, 500, even 1,200 shrank to 50 or 30. I wasn't happy about that; I would have preferred a larger audience, and yet I repeated my advice until the administration intervened. Why did I do it? Was it because I disliked the examination system, which blurred the line between thought and routine? Was it because I despised the idea that knowledge was a skill that had to be acquired and stabilized by rigorous training? Or was it because I didn't think much of my own performance? All these factors may have played a role. (p.122).

But although he sympathised with the original aims of the student movement, Feyerabend was unimpressed by their leaders, feeling that their ideas were as authoritarian as those they were trying to replace. He reports having cut fewer lectures during the student strike than either before or afterwards! Nevertheless, by holding his lectures off-campus during this campus war, Feyerabend antagonised the administration that had hired him. Tales of him giving ‘A’ grades to every student in his class, regardless of their production (or lack of it), abound. He had the impression that some of his colleagues, especially John Searle, wanted to have him fired, and that they only gave up when they realised how much paperwork would be involved (p.126).

The Late Sixties.

During the summer of 1966, Feyerabend lectured on church dogma at Berkeley. (‘Why church dogma? Because the development of church dogma shares many features with the development of scientific thought’ (pp.137-8)). He eventually turned these thoughts into a paper on ‘Classical Empiricism’, published in 1970, in which he argued that empiricism shared certain problematic features with protestantism. He had already come some way from his 1965 defence of a ‘disinfected’, ‘tolerant’ form of Empiricism. The publication, in 1969, of the four- page article, ‘Science Without Experience’, which argued that in principle experience is necessary at *no* point in the construction, comprehension or testing of empirical scientific theories finally gave notice that Feyerabend was no longer concerned to present himself as any kind of empiricist.

Despite taking his academic duties and responsibilities decreasingly seriously, and coming into conflict with his own university's administration as a result, Feyerabend had not yet fouled his substantial reputation as a serious philosopher of science. He reports that he received job offers from London, Berlin, Yale, and Auckland, that he was invited to become a fellow of All Souls College, Oxford, and that he corresponded with Friedrich von Hayek (whom he already knew from the Alpbach seminars)

about a job in Freiburg (p.127). He accepted the posts in London, Berlin, and Yale. In 1968, he resigned from UC Berkeley and left for Minneapolis, but grew homesick, got re-appointed, and returned to Berkeley almost immediately.

In London, lecturing to University College and the LSE, he met Imre Lakatos. The two became great friends, corresponding with one another regularly and voluminously until Lakatos' death. Feyerabend recalls that Lakatos, whose office was across the corridor from the LSE lecture hall, used to intervene in his lectures when Feyerabend made a point he disagreed with (*SFS*, p.13, *KT*, p.128).

Against Method (1970-75).

After stints in London, Berlin, and Yale (all of them running alongside his post at UC Berkeley), Feyerabend took up a chair at the University of Auckland, New Zealand, and lectured there in 1972 and 1974 (pp.134-5). He even considered settling down in New Zealand around that time (p.153), although this hardly seems compatible with his jet-setting lifestyle.

By the early 1970s Feyerabend had flown the falsificationist coop and was ready expound his own perspective on scientific method. In 1970, he published a long article entitled ‘Against Method’ in which he attacked several prominent accounts of scientific methodology. In their correspondence, he and Lakatos subsequently planned the construction of a debate volume, to be entitled *For and Against Method*, in which Lakatos would put forward the ‘rationalist’ case that there was an identifiable set of rules of scientific method which make all good science science, and Feyerabend would attack it. Lakatos' unexpected death in February 1974, which seems to have shocked Feyerabend deeply, meant that the rationalist part of the joint work was never completed.

Later that year, Feyerabend found himself lecturing at the University of Sussex:

> I have no idea why and how I went to the University of Sussex at Brighton... what I do remember is that I taught two terms (1974/1975) and then resigned; twelve hours a week (one lecture course, the rest tutorials) was too much. (p.153).

A member of Feyerabend's audience recalls things in rather more detail:

> Sussex University: the start of the Autumn Term, 1974. There was not a seat to be had in the biggest Arts lecture theatre on campus. Taut with anticipation, we waited expectantly and impatiently for the advertized event to begin. He was not on time - as usual. In fact rumour had it that he would not be appearing at all that illness (or was it just ennui? or perhaps a mistress?) had confined him to bed. But just as we began sadly to reconcile ourselves to the idea that there would be no performance that day at all, Paul Feyerabend burst through the door at the front of the packed hall. Rather pale, and supporting himself on a short metal crutch, he walked with a limp across to the blackboard. Removing his sweater he picked up the chalk and wrote down three questions one beneath the other:

> What's so great about knowledge? What's so great about science? What's so great about truth? We were not going to be disappointed after all!
>
> During the following weeks of that term, and for the rest of his year as a visiting lecturer, Feyerabend demolished virtually every traditional academic boundary. He held no idea and no person sacred. With unprecedented energy and enthusiasm he discussed anything from Aristotle to the Azande. How does science differ from witchcraft? Does it provide the only rational way of cognitively organizing our experience? What should we do if the pursuit of truth cripples our intellects and stunts our individuality? Suddenly epistemology became an exhilarating area of investigation.
>
> Feyerabend created spaces in which people could breathe again. He demanded of philosophers that they be receptive to ideas from the most disparate and apparently far-flung domains, and insisted that only in this way could they understand the processes whereby knowledge grows. His listeners were enthralled, and he held his huge audiences until, too ill and too exhausted to continue, he simply began repeating himself. But not before he had brought the house down by writing 'Aristotle' in three-foot high letters on the blackboard and then writing 'Popper' in tiny, virtually illegible letters beneath it! (John Krige, *Science, Revolution and Discontinuity*, (Sussex: Harvester Press, 1980), pp.106-7).

Because his health was poor, Feyerabend started seeing a healer who had been recommended to him. The treatment was successful, and thenceforth Feyerabend used to refer to his own case as an example of both the failures of orthodox medicine and the largely unexplored possibilities of 'alternative' or traditional remedies.

Instead of the volume written jointly with Lakatos, Feyerabend put together his tour de force, the book version of *Against Method* (London: New Left Books, 1975), which he sometimes conceived of as a letter to Lakatos (to whom the book is dedicated). A more accurate description, however, is the one given in his autobiography:

> *AM* is not a book, it is a collage. It contains descriptions, analyses, arguments that I had published, in almost the same words, ten, fifteen, even twenty years earlier... I arranged them in a suitable order, added transitions, replaced moderate passages with more outrageous ones, and called the result 'anarchism'. I loved to shock people... (pp.139, 142).

The book contained many of the themes mentioned so far in this essay, sprinkled into a case study of the transition from geocentric to heliocentric astronomy. But whereas he had previously been arguing in favour of methodology (a 'pluralistic' methodology, that is), he had now become dissatisfied with *any* methodology. He emphasised that older scientific theories, like Aristotle's theory of motion, had powerful empirical and argumentative support, and stressed, correlatively, that the heroes of the scientific revolution, such as Galileo, were not as scrupulous as they were sometimes represented to be. He portrayed Galileo as making full use of rhetoric, propaganda, and various epistemological tricks in order

to support the heliocentric position. The Galileo case is crucial for Feyerabend, since the 'scientific revolution' is his paradigm of scientific progress and of radical conceptual change, and Galileo is his hero of the scientific revolution. He also sought further to downgrade the importance of empirical arguments by suggesting that aesthetic criteria, personal whims and social factors have a far more decisive role in the history of science than rationalist or empiricist historiography would indicate.

Against Method explicitly drew the 'epistemological anarchist' conclusion that there are no useful and exceptionless methodological rules governing the progress of science or the growth of knowledge. The history of science is so complex that if we insist on a general methodology which will not inhibit progress the only 'rule' it will contain will be the useless suggestion: 'anything goes'. In particular, logical empiricist methodologies and Popper's Critical Rationalism would inhibit scientific progress by enforcing restrictive conditions on new theories. The more sophisticated 'methodology of scientific research programmes' developed by Lakatos either contains ungrounded value-judgements about what constitutes good science, or is reasonable only because it is epistemological anarchism in disguise. The phenomenon of incommensurability renders the standards which these 'rationalists' use for comparing theories inapplicable. The book thus (understandably) had Feyerabend branded an 'irrationalist'. At a time when Kuhn was downplaying the 'irrationalist' implications of his own book, Feyerabend was perceived to be casting himself in the role others already saw as his for the taking. (He did not, however, commit himself to *political* anarchism. His political philosophy was a mixture of liberalism and social democracy).

He later said:

> One of my motives for writing *Against Method* was to free people from the tyranny of philosophical obfuscators and abstract concepts such as 'truth', 'reality', or 'objectivity', which narrow people's vision and ways of being in the world. Formulating what I thought were my own attitude and convictions, I unfortunately ended up by introducing concepts of similar rigidity, such as 'democracy', 'tradition', or 'relative truth'. Now that I am aware of it, I wonder how it happened. The urge to explain one's own ideas, not simply, not in a story, but by means of a 'systematic account', is powerful indeed. (pp.179-80).

The Political Consequences of Epistemological Anarchism: *Science in a Free Society* (1978).

The critical reaction to *Against Method* seems to have taken Feyerabend by surprise. He was shocked to be accused of being aggressive and nasty, so he replied by accusing his accusers of the very same thing. He felt it necessary to respond to most of the book's major reviews in print, and later assembled these replies into a section of his next book, *Science in a Free Society*, entitled 'Conversations with Illiterates'. Here he berated the unfortunate reviewers for having misread *Against Method*, as well as for being constitutionally incapable of distinguishing between irony, playfulness, argument by reductio ad absurdum, and the (apparently rather few) things he had really committed himself to in *AM*. The spectacle of Feyerabend levelling these accusations at others is not itself without irony. (His widow

reports that in his later years, *SFS* was the book he would most like to have distanced himself from). In the commotion surrounding *AM*, Feyerabend succumbed to depression:

> ... now I was alone, sick with some unknown affliction; my private life was in a mess, and I was without a defense. I often wished I had never written that fucking book. (*KT*, p.147).

Feyerabend saw himself as having undermined the arguments for science's privileged position within culture, and much of his later work was a critique of the position of science within Western societies. Because there is no scientific method, we can't justify science as the best way of acquiring knowledge. And the *results* of science don't prove its excellence, since these results have often depended on the presence of non-scientific elements, science prevails only because ‘the show has been rigged in its favour’ (*SFS*, p.102), and other traditions, despite their achievements, have never been given a chance. The truth, he suggests, is that

> science is much closer to myth than a scientific philosophy is prepared to admit. It is one of the many forms of thought that have been developed by man, and not necessarily the best. It is conspicuous, noisy, and impudent, but it is inherently superior only for those who have already decided in favour of a certain ideology, or who have accepted it without ever having examined its advantages and its limits (*AM*, p.295).

The separation of church and state should therefore be supplemented by the separation of science and state, in order for us to achieve the humanity we are capable of. Setting up the ideal of a free society as ‘a society in which all traditions have equal rights and equal access to the centres of power’ (*SFS*, p.9), Feyerabend argues that science is a threat to democracy. To defend society against science we should place science under democratic control and be intensely sceptical about scientific ‘experts’, consulting them only if they are controlled democratically by juries of laypeople.

Ten Wonderful Years: The Eighties in Berkeley and Zurich.

Out of all Feyerabend's many academic positions, perhaps the one he enjoyed most was his tenure throughout the 1980s at the *Eidgenössische Technische Hochschule*, Zurich. Feyerabend applied for the post after his friend Eric Jantsch had told him that the Polytechnic was looking for a philosopher of science. The selection process was, by Feyerabend's account, very long and somewhat involved (pp.154ff.). Having recently left another post in Kassel, he apparently gave up hopes of being hired by the Swiss, and ‘decided to remain in Berkeley and stop moving about’ (p.158). But, after several stages in the decision-making procedure, he was finally given the job, and ‘ten wonderful years of half-Berkeley, half-Switzerland’ (p.158) turned out to be exactly what he had been looking for. At Zurich he lectured on Plato's *Theaetetus* and *Timaeus,* and then on Aristotle's *Physics*. The two-hour seminars, many of which were organised by Christian Thomas (with whom Feyerabend was to edit anthologies) were run on the same lines as Berkeley: no set topic, but presentations by the participants (p.160). Feyerabend later considered this to be the period in which he ‘got his intellectual act together’ (p.162), meaning that he recovered from the critical reactions to *Against Method* and was finally freed from the

necessity of defending it against all criticism. However, this didn't seem to have affected his attitude towards work: in Zurich he refused offers of an office, because no office meant no office hours, and therefore no waste of time (pp.131, 158)!

Many of the more important papers Feyerabend published during the mid- 1980s were collected together in *Farewell to Reason* (London: Verso, 1987). The major message of this book is that relativism is the solution to the problems of conflicting beliefs and of conflicting ways of life. Feyerabend starts by suggesting that the contemporary intellectual scene in Western culture is by no means as fragmented and cacophonous as many intellectuals would have us believe. The surface diversity belies a deeper uniformity, a monotony generated and sustained by the cultural and ideological imperialism which the West uses to beat its opponents into submission. Such uniformity, however, can be shown to be harmful even when judged by the standards of those who impose it. Cultural diversity, which already exists in some societies, is a good thing not least because it affords the best defence against totalitarian domination.

Feyerabend proposes to support the idea of cultural diversity both positively, by producing considerations in its favour, and negatively ‘by criticising philosophies that oppose it’ (*FTR*, p.5). Contemporary philosophies of the latter type are said to rest on the notions of *Objectivity* and *Reason*. He seeks to undermine the former notion by pointing out that confrontations between cultures with strongly held opinions which are each believed by members of the cultures in question to be objectively true can turn out in different ways. The result of such confrontation may be the persistence of the old views, fruitful and mutual interaction, relativism, or argumentative evaluation. ‘Relativism’ here means the decision to treat the other people's form of life and the beliefs it embodies as ‘true-for-them’, while treating our own views as ‘true-for-us’. Feyerabend feels that this is an appropriate way to *resolve* such confrontation.

Admittedly, these outcomes are indeed possible. But this does not establish any form of relativism. Indeed, we might as well turn the argument around, and say that the possibility of the dispute being resolved by one participant freely coming around to the other's point of view shows the *un*tenability of relativism.

Feyerabend complains that the ideas of reason and rationality are ‘ambiguous and never clearly explained’ (*FTR*, p.10); they are deified hangovers from autocratic times which no longer have any content but whose ‘halo of excellence’ (ibid.) clings to them and lends them spurious respectability:

> [R]ationalism has no identifiable content and reason no recognisable agenda over and above the principles of the party that happens to have appropriated its name. All it does now is to lend class to the general drive towards monotony. It is time to disengage Reason from this drive and, as it has been thoroughly compromised by the association, to bid it farewell. (*FTR*, p.13).

Relativism is the tool with which Feyerabend hopes to ‘undermine the very basis of Reason’ (ibid.). But

is it Reason with a capital ‘R’, the philosophers' abstraction alone, that is to be renounced, or reason itself too? Feyerabend is on weak ground when he claims that ‘Reason’ is a philosophers' notion which has no content, for it is precisely the philosopher who *is* willing to attach a specific content to the formal notion of rationality (unlike the layperson, whose notion of reason is closer to what Feyerabend calls the ‘material’ conception, where to be rational is ‘to avoid certain views and to accept others’ (ibid., p.10)).

Relativism is a result of cultural confrontation, an ‘attempt to make sense of the phenomenon of cultural variety’ (*FTR*, p.19). Feyerabend is well aware that the term ‘relativism’ itself is understood in many different ways. But his attempt to occupy a substantial yet defensible relativist position is a failure. At some points he merely endorses views which no-one would deny, but which do not deserve to be called relativist (such as the idea that people may profit from studying other points of view, no matter how strongly they hold their own view (*FTR*, p.20)). At others he does manage to subscribe to a genuinely relativist view, but fails to show why it must be accepted.

It was only in 1988, on the 50th anniversary of Austria's unification with Germany, that Feyerabend became interested in his past (p.1). The Feyerabends left California for life in Switzerland and Italy in the fall of 1989 (p.2). It was during this move that Feyerabend re-discovered his mother's suicide note (p.9), which may have been one of the factors that spurred him to write his autobiography. Feyerabend looked forward to his retirement, and he and Grazia decided to try to have children. He claimed to have forgotten the thirty-five years of his academic career almost as quickly as he had earlier forgotten his military service (p.168).

Feyerabend in the Nineties.

Feyerabend published a surprisingly large number of papers in the 1990s (although many of them were short ones with overlapping content). Several appeared in a new journal, *Common Knowledge*, in whose inauguration he lent a hand, and which set out to integrate insights from all parts of the intellectual landscape.

Although these papers were on scattered subjects, there are some strong themes running through them, several of which (as I have argued elsewhere (Preston [forthcoming])) bear comparison with what gets called ‘post-modernism’. Here I shall sketch only the main ones.

One of the projects which Feyerabend worked on for a long time, but never really brought to completion, went under the name ‘The Rise of Western Rationalism’. Under this umbrella he hoped to show that Reason (with a capital ‘R’) and Science had displaced the binding principles of previous world-views *not* as the result of having won an argument, but as the result of power-play. While the first philosophers (the pre-Socratic thinkers) had interesting views, their attempt to replace, streamline or rationalise the folk-wisdom which surrounded them was eminently resistible. Their introduction of the appearance/reality dichotomy made nonsense of many of the things people had previously known. Even nowadays, indigenous cultures and counter-cultural practices provide alternatives to Reason and that nasty Western science.

However, Feyerabend recognised that this is to present science as too much of a monolith. In most of his work after *Against Method*, emphasises what has come to be known as the ‘disunity of science’. Science, he insists, is a collage, not a system or a unified project. Not only does it include plenty of components derived from distinctly ‘non-scientific’ disciplines, but these components are often vital parts of the ‘progress’ science has made (using whatever criterion of progress you prefer). Science is a collection of theories, practices, research traditions and world-views whose range of application is not well-determined and whose merits vary to a great extent. All this can be summed up in his slogan: "Science is not one thing, it is many".

Likewise, the supposed ontological correlate of science, ‘the world’, consists not only of one kind of thing but of countless kinds of things, things which cannot be ‘reduced’ to one another. In fact, there is no good reason to suppose that the world has a single, determinate nature. Rather we inquirers construct the world in the course of our inquiries, and the plurality of our inquiries ensures that the world itself has a deeply plural quality: the Homeric gods and the microphysicist's subatomic particles are simply different ways in which ‘Being’ responds to (different kinds of) inquiry. How the world is ‘in-itself’ is for ever unknowable. In this respect, Feyerabend's last work can be thought of as aligned with ‘social constructivism’.

Conclusion: Last Things.

Feyerabend's autobiography occupied him right up until his death on February 11th, 1994, at his home in Meilen on Lake Zurich. At the end of the book, he expressed the wish that what should remain of him would be ‘*not* papers, *not* final declarations, but love’ (p.181).

His autobiography was published in 1995, his last book *The Conquest of Abundance*, is now being prepared, and a third volume of his *Philosophical Papers* will appear in 1998.

Although the focus of philosophy of science has moved away from interest in scientific methodology in recent years, this is not due in any great measure to acceptance of Feyerabend's anti-methodological argument. His critique of science (which gave him the reputation for being an ‘anti-science philosopher’, ‘the worst enemy of science’, etc.) is patchy. Its flaws stem directly from his scientific realism. It sets up a straight confrontation between science and other belief-systems as if they are all aiming to do the *same* thing (give us ‘knowledge of the world’) and must be compared for how well they deliver the goods. A better approach would be, in Gilbert Ryle's words, ‘to draw uncompromising contrasts’ between the businesses of science and those of other belief-systems. Such an approach fits far better with the theme Feyerabend approached later in his life: that of the disunity of science.

Feyerabend came to be seen as a leading cultural relativist, not just because he stressed that some theories are incommensurable, but also because he defended relativism in politics as well as in epistemology. His denunciations of aggressive Western imperialism, his critique of science itself, his conclusion that ‘objectively’ there may be nothing to choose between the claims of science and those of

astrology, voodoo, and alternative medicine, as well as his concern for environmental issues ensured that he was a hero of the anti-technological counter-culture.

Different components and phases of Feyerabend's work have influenced very different groups of thinkers. His early scientific realism, contextual theory of meaning, and the way he proposed to defend materialism were taken up by Paul and Patricia Churchland. Richard Rorty, for a time, also endorsed eliminative materialism. Feyerabend's critique of reductionism has influenced Cliff Hooker and John Dupré, and his general point of view influenced books such as Alan Chalmers' well-known introduction to philosophy of science *What is this thing called science?* (Milton Keynes: Open University Press, 1978).

Feyerabend has also had considerable influence within the social studies. He directly inspired books like D.L.Phillips' *Abandoning Method* (San Francisco, 1973), in which the attempt was made to transcend methodology. Less directly, he has exerted enormous influence on a generation of sociologists of science through his relativism, social constructivism, and apparent irrationalism. It is still far too early to say whether, and in what way, his philosophy will be remembered.

3. Feyerabend's Major Writings

- 'Problems of Empiricism', *Beyond the Edge of Certainty: Essays in Contemporary Science and Philosophy*, ed. R.G.Colodny (New Jersey: Prentice-Hall, 1965), pp.145-260.
- *Against Method* (London: Verso, 1975).
- *Science in a Free Society* (London: New Left Books, 1978).
- *Realism, Rationalism, and Scientific Method: Philosophical Papers, Volume 1* (Cambridge: Cambridge University Press,1981).
- *Problems of Empiricism: Philosophical Papers, Volume 2* (Cambridge: Cambridge University Press, 1981).
- *Farewell to Reason* (London: Verso/New Left Books, 1987).
- *Against Method* (London: 1975); Revised edn (London: Verso, 1988).
- *Three Dialogues on Knowledge* (Oxford: Basil Blackwell, 1991).
- *Killing Time: The Autobiography of Paul Feyerabend*, (Chicago: University of Chicago Press, 1995).
- *Conquest of Abundance: A Tale of Abstraction Versus the Richness of Being*, ed. B.Terpstra (Chicago: University of Chicago Press, 1999).
- *Knowledge, Science and Relativism: Philosophical Papers, Volume 3*, ed. J.Preston, (Cambridge: Cambridge University Press, 1999).

Bibliography

- Achinstein, P. [1964]: 'On the Meaning of Scientific Terms', *Journal of Philosophy*, **61**.
- Achinstein, P. [1968]: *Concepts of Science*. Baltimore: Johns Hopkins University Press.

- Agassi, J. [1976]: Review of *Against Method*, *Philosophia*, **6**.
- Alford, C.F. [1985]: ‘Yates on Feyerabend's Democratic Relativism’, *Inquiry*, **28**.
- Andersson, G. [1994]: *Criticism and the History of Science: Kuhn's, Lakatos's and Feyerabend's Criticisms of Critical Rationalism*. (Leiden: Brill).
- Baertschi, B. [1986]: ‘Le Réalisme Scientifique de Feyerabend’, *Dialogue*, **25**.
- Bearn, G.C.F. [1986]: ‘Nietzsche, Feyerabend, and the Voices of Relativism’, *Metaphilosophy*, **17**.
- Bernstein, R.J. [1983]: *Beyond Objectivism and Relativism*. Oxford: Basil Blackwell.
- Bhaskar, R. [1975]: ‘Feyerabend and Bachelard: Two Philosophies of Science’, *New Left Review*, **94**.
- Brown, H.I. [1976]: ‘Reduction and Scientific Revolutions’, *Erkenntnis*, **10**.
- Brown, H.I. [1983]: ‘Incommensurability’, *Inquiry*, **26**.
- Burian, R.M.: ‘Scientific Realism and Incommensurability: Some Criticisms of Kuhn and Feyerabend’, in *Methodology, Metaphysics and the History of Science*, eds. R.S.Cohen and M.W.Wartofsky (Dordrecht: D.Reidel, 1984).
- Butts, R.E. [1966]: ‘Feyerabend and the Pragmatic Theory of Observation’, *Philosophy of Science*, **33**.
- Chalmers, A. [1986]: ‘The Galileo that Feyerabend Missed: An Improved Case Against Method’, in J.A.Schuster & R.R.Yeo (eds.), *The Politics and Rhetoric of Scientific Method*. (Dordrecht: D.Reidel).
- Churchland, P.M. [1979] *Scientific Realism and the Plasticity of Mind* (Cambridge: Cambridge University Press).
- Churchland, P.M. [1981]: ‘Eliminative Materialism and the Propositional Attitudes’, *Journal of Philosophy*, **78**.
- Churchland, P.S. [1986]: *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge Mass.: MIT Press.
- Coffa, J.A. [1967]: ‘Feyerabend on Explanation and Reduction’, *Journal of Philosophy*, **64**.
- Collier, J. [1984]: ‘Pragmatic Incommensurability’, in P.D.Asquith & P.Kitcher (eds.), *PSA 1984, Volume 1*. East Lansing, Mi: Philosophy of Science Association.
- Couvalis, S.G. [1986]: ‘Should Philosophers Become Playwrights?’, *Inquiry*, **29**.
- Couvalis, S.G. [1987]: ‘Feyerabend's Epistemology and Brecht's Theory of the Drama’, *Philosophy and Literature*, **11**.
- Couvalis, S.G. [1988a]: ‘Feyerabend, Ionesco, and the Philosophy of the Drama’, *Critical Philosophy*, **4**.
- Couvalis, S.G. [1988b]: ‘Feyerabend and Laymon on Brownian Motion’, *Philosophy of Science*, **55**.
- Couvalis, S.G. [1989] *Feyerabend's Critique of Foundationalism* (Aldershot: Avebury Press).
- Couvalis, S.G. [2001] ‘Recent Feyerabendiana’, *Metascience*, **10**.
- Davidson, D. [1973]: ‘On the Very Idea of a Conceptual Scheme’, *Proceedings of the American Philosophical Association*, **47**. (Reprinted in Krausz and Meiland [1982]).
- Devitt, M. [1979]: ‘Against Incommensurability’, *Australasian Journal of Philosophy*, **57**.
- Dürr, H-P., ed. [1980]: *Versuchungen: Aufsätze zur Philosophie Paul Feyerabend's*. Erster Band. Frankfurt: Suhrkamp.

- Dürr, H-P., ed. [1981]: *Versuchungen: Aufsätze zur Philosophie Paul Feyerabend's*. Zweiter Band. Frankfurt: Suhrkamp.
- Dusek, V. [1998] ‘Brecht and Lukács as Teachers of Feyerabend and Lakatos: The Feyerabend-Lakatos Debate as Scientific Recapitulation of the Brecht-Lukács Debate’, *History of the Human Sciences*, **11**.
- Everitt, N. [1981]: ‘A Problem for the Eliminative Materialist’, *Mind*, **90**.
- Farrell, R.P. [2000] ‘Rival Theories and Empirical Content Revisited’, *Studies in History and Philosophy of Science*, **31**.
- Farrell, R.P. [2000] ‘Will the Popperian Feyerabend please step forward: pluralistic, Popperian themes in the Philosophy of Paul Feyerabend’, *International Studies in the Philosophy of Science*, **14**.
- Farrell, R.P. [2001] ‘Feyerabend's Metaphysics: Process-Realism or Voluntarist-Idealism?’, *Journal for General Philosophy of Science*, **32**.
- Finocchiaro, M.A. [1978]: ‘Rhetoric and Scientific Rationality’, in P.D.Asquith & I.Hacking (eds.), *PSA 1978, Volume 1*. (East Lansing, Mi: Philosophy of Science Association).
- Fuller, S. [1995]: ‘Paul Feyerabend: An Appreciation’, *Vest*, **8**, 1995.
- Gellner, E. [1975]: ‘Beyond Truth and Falsehood (Review of *Against Method*)’, *British Journal for the Philosophy of Science*, **26**.
- Giedymin, J. [1970]: ‘The Paradox of Meaning Variance’, *British Journal for the Philosophy of Science*, **21**.
- Giedymin, J. [1971]: ‘Consolations for the Irrationalist?’, *British Journal for the Philosophy of Science*, **22**.
- Giedymin, J. [1976]: ‘Instrumentalism and its Critique: A Reappraisal’, in R.S.Cohen, P.K.Feyerabend & M.Wartofsky (eds.), *Essays in Memory of Imre Lakatos*. Dordrecht: D.Reidel.
- Goldman, M. [1980]: ‘The Material Basis for Progress in Science’, in P.T.Durbin (ed.), *Research in Philosophy & Technology*. (Greenwich, CT.: JAI Press).
- Goldman, M. [1982]: ‘Science and Play’, in P.D.Asquith & T.Nickles (eds.), *PSA 1982, Volume 1*. (East Lansing, MI.: Philosophy of Science Association).
- Gunaratne, R.D. [1980]: *Science, Understanding and Truth*. (Sri Lanka: Ministry of Higher Education Publications).
- Hacking, I. [1975]: *Why Does Language Matter to Philosophy?* (Cambridge: Cambridge University Press).
- Hacking, I. [1983]: *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. (Cambridge: Cambridge University Press).
- Hacking, I. [1991]: Review of P.K.Feyerabend, *Against Method*, and *Farewell to Reason*, *Journal of Philosophy*, **88**.
- Hannay, A. [1989]: ‘Politics and Feyerabend's Anarchist’, in M.Dascal & O.Gruengard (eds.), *Knowledge and Politics*. (Colorado: Westview Press, 1989).
- Hanson, N.R. [1959]: ‘Five Cautions for the Copenhagen Interpretation's Critics’, *Philosophy of Science*, **26**.
- Harré, R. [1977]: Review of P.K.Feyerabend's *Against Method*. *Mind,* **86**.
- Harré, R. [1984]: ‘For Method: A Response to Feyerabend’, *New Ideas in Psychology*, **2**.
- Hentschel, K. [1985] ‘On Feyerabend's Version of ‘Mach's Theory of Research and its Relation to

Einstein', *Studies in History and Philosophy of Science*, **16**.

- Hesse, M.B. [1974]: *The Structure of Scientific Inference*. London: MacMillan.
- Hesse, M.B. [1980]: *Revolutions and Reconstructions in the Philosophy of Science*. Sussex: Harvester Press.
- Hollis, M. & Lukes, S., eds. [1982]: *Rationality and Relativism*. Oxford: Basil Blackwell.
- Hooker, C. [1972a]: Critical Notice of M.Radner & S.Winokur, *Analyses of Theories and Methods of Physics and Psychology, Minnesota Studies in the Philosophy of Science,* **4**. *Canadian Journal of Philosophy*, **1**.
- Horgan, J. [1993]: 'Profile: Paul Karl Feyerabend: The Worst Enemy of Science', *Scientific American*, May 1993.
- Hoyningen-Huene, P. [1994]: 'Obituary of Paul K.Feyerabend (1924-1994)', *Erkenntnis,* **40**.
- Hoyningen-Huene, P. [1997]: 'Paul K. Feyerabend', *Journal for General Philosophy of Science* **28**.
- Hoyningen-Huene, P. [1999]: 'Paul K. Feyerabend', in: J.Nida-Rümelin (ed.), *Philosophie der Gegenwart in Einzeldarstellungen*. Stuttgart: Kröner.
- Hoyningen-Huene, P. [1999]: 'Feyerabends Kritik an Kuhns normaler Wissenschaft', in J.Nida-Rümelin (ed.), *Rationality, Realism, Revision: Proceedings of the 3rd international congress of the Society for Analytical Philosophy*. Berlin: de Gruyter.
- Hoyningen-Huene, P. [2002]: 'Paul Feyerabend - ein postmoderner Philosoph? Ein Portrait', *Information Philosophie* März.
- Hull, R.T. [1972]: 'Feyerabend's Attack on Observation Sentences', *Synthese*, **23**.
- Hung, H-C.E. [1987]: 'Incommensurability and Inconsistency of Languages', *Erkenntnis*, **27**.
- Jones, W.B. [1978]: 'Theory-Ladenness and Theory Comparison', in P.D.Asquith & I.Hacking (eds.), *PSA 1978, Volume 1*. East Lansing, Mi: Philosophy of Science Association.
- Kadvany, J. [1996]: 'Reason in History: Paul Feyerabend's Autobiography', *Inquiry*, **39**.
- Kleiner, S.A. [1979]: 'Feyerabend, Galileo and Darwin: How to Make the Best out of what you have - or think you can get', *Studies in History and Philosophy of Science*, **10**.
- Koertge, N. [1972]: 'For and Against Method' (Review of Radner & Winokur), *British Journal for the Philosophy of Science*, **23**.
- Koertge, N. [1980]: Review of P.K.Feyerabend's *Science in a Free Society*, *British Journal for the Philosophy of Science*, **31**.
- Kresge, S. [1996]: 'Feyerabend Unbound', (Review of *Killing Time*), *Philosophy of the Social Sciences*, **26**.
- Krige, J. [1980]: *Science, Revolution and Discontinuity*. (Sussex: Harvester Press).
- Lakatos, I. [1978]: *The Methodology of Scientific Research Programmes: Philosophical Papers, Volume 1*. Cambridge: Cambridge University Press.
- Lamb, D, Munévar, G. & Preston, J.M. (eds.), *'The Worst Enemy of Science'?: Essays on the Philosophy of Paul Feyerabend*. (Forthcoming during 1998).
- Laudan, L. [1989]: 'For Method: or, Against Feyerabend', in J.R.Brown & J.Mittelstrass (eds.), *An Intimate Relation*. (Dordrecht: Kluwer, 1989).
- Laymon, R. [1977]: 'Feyerabend, Brownian Motion, and the Hiddenness of Refuting Facts', *Philosophy of Science*, **44**.
- Leplin, J. [1969]: 'Meaning Variance and the Comparability of Theories', *British Journal for the*

Philosophy of Science, **20**.

- Machamer, P.K. [1973]: ‘Feyerabend and Galileo: the Interaction of Theories, and the Reinterpretation of Experience’, *Studies in History and Philosophy of Science*, **4**.
- Maia Neto, J.R. [1991]: ‘Feyerabend's Scepticism’, *Studies in History and Philosophy of Science*, **22**.
- Malolo Dissakè, E. [2001] *Feyerabend: Épistémologie, anarchisme, et société libre*, (Paris: Presses Universitaires de France).
- Margolis, J. [1970a]: ‘Notes on Feyerabend and Hanson’, in M.Radner & S.Winokur (eds.), *Analyses of Theories and Methods in Physics and Psychology, Minnesota Studies in the Philosophy of Science, Volume 4*. Minneapolis: University of Minnesota Press.
- Martin, M. [1984]: ‘How to be a Good Philosopher of Science: A Plea for Empiricism in Matters Methodological’, in R.S.Cohen & M.Wartofsky (eds.), *Methodology, Metaphysics and the History of Science: in Memory of Benjamin Nelson*. (Dordrecht: Reidel).
- McEvoy, J.G. [1975]: ‘A ‘Revolutionary’ Philosophy of Science: Feyerabend and the Degeneration of Critical Rationalism into Sceptical Fallibilism’, *Philosophy of Science*, **42**.
- Mellor, D.H. [1969]: ‘Physics and Furniture’, in N.Rescher (ed.), *Studies in the Philosophy of Science*, American Philosophical Quarterly Monograph Series, No.3. Oxford: Basil Blackwell.
- Moberg, D.W. [1979]: ‘Are there Rival, Incommensurable Theories?’, *Philosophy of Science*, **46**.
- Motterlini, M. (ed.), [1999] *For and Against Method, including Lakatos's Lectures on Scientific Method, and the Lakatos-Feyerabend Correspondence*, (Chicago: University of Chicago Press).
- Munévar, G., (ed.), [1991] *Beyond Reason: Essays on the Philosophy of Paul Feyerabend* (Dordrecht: Kluwer).
- Munévar, G. [1998] *Evolution and the Naked Truth: A Darwinian Approach to Philosophy*, (Aldershot: Avebury).
- Musgrave, A. [1976]: ‘Method or Madness? Can the Methodology of Research Programmes be Rescued from Epistemological Anarchism?’, in R.S.Cohen, P.K.Feyerabend & M.Wartofsky (eds.), *Essays in Memory of Imre Lakatos*. Dordrecht: Reidel.
- Musgrave, A. [1978]: ‘How to Avoid Incommensurability’, in I.Niiniluoto & R.Tuomela [1979].
- Nagel, E. [1979]: *Teleology Revisited, and Other Essays in the Philosophy and History of Science*. (New York: Columbia University Press, 1979).
- Newton-Smith, W.H.: *The Rationality of Science* (London: Routledge and Kegan Paul, 1981).
- Nordmann, A. [1990]: ‘Goodbye and Farewell: Siegel vs. Feyerabend’, *Inquiry*, **33**.
- Oberdan, T. [1990]: ‘Positivism and the Pragmatic Theory of Observation’, in A.Fine, M.Forbes & L.Wessels (eds.), *PSA 1990, Volume 1*. (East Lansing, MI: Philosophy of Science Association).
- Oberheim, E., & Hoyningen-Huene, P. [2000] ‘Feyerabend's Early Philosophy’ (Review of J.Preston, *Feyerabend: Philosophy, Science and Society*), *Studies in History and Philosophy of Science*, **31**.
- O'Gorman, F. [1989]: *Rationality and Relativity: The Quest for Objective Knowledge*. (Aldershot: Avebury Press), ch.3.
- Papineau, D. [1979]: *Theory and Meaning*. Oxford: Clarendon Press.
- Pasternak, G.P. [1984]: ‘Interview mit Paul Feyerabend’, *Unter dem Pflaster liegt der Strand*, **13**.
- Pera, M. [1994]: *The Discourses of Science*. (Chicago: University of Chicago Press).

- Post, H. [1971]: ‘Correspondence, Invariance and Heuristics’, *Studies in History and Philosophy of Science*, **2**.
- Preston, J.M. [1995a]: ‘Frictionless Philosophy: Paul Feyerabend and Relativism’ *History of European Ideas*, **20**.
- Preston, J.M., [1997a]: *Feyerabend: Philosophy, Science and Society* (Cambridge: Polity Press, 1997).
- Preston, J.M., [1997b]: ‘Feyerabend's Retreat from Realism’, *Philosophy of Science*, **64**.
- Preston, J.M., [1997c]: ‘Feyerabend's Final Relativism’ *The European Legacy*, **2**.
- Preston, J.M., [1998] ‘Science as Supermarket: ‘Post-Modern’ Themes in Paul Feyerabend's Later Philosophy of Science’, *Studies in History and Philosophy of Science*, **29**.
- Preston, J.M., Munévar, G. & Lamb, D. (eds.), [2000] *The Worst Enemy of Science? Essays in Memory of Paul Feyerabend*, (New York: Oxford University Press).
- Putnam, H. [1965]: ‘How Not to Talk about Meaning’, in R.Cohen & M.W.Wartofsky (eds.) *Boston Studies in the Philosophy of Science, Volume 2, In Honor of Philipp Frank*, New York: Humanities Press.
- Putnam, H. [1978]: *Meaning and the Moral Sciences*. London: Routledge and Kegan Paul.
- Putnam, H. [1981]: *Reason, Truth, and History*. Cambridge: Cambridge University Press.
- Sankey, H. [1994]: *The Incommensurability Thesis*, Aldershot: Avebury Press.
- Scheffler, I. [1966]: *Science and Subjectivity*. Indianapolis: Hackett Publishing Co.
- Scheibe, E. [1988]: ‘Paul Feyerabend un die rationalen Rekonstruktionen’, in P.Hoyningen-Huene & G.Hirsch (eds.), *Wozu Wissenschaftsphilosophie? Positionen und Fragen zur heutigen Wissenschaftsphilosophie*. (Berlin: De Gruyter).
- Shapere, D. [1966]: ‘Meaning and Scientific Change’, in R.G.Colodny (ed.), *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, Pittsburgh: University of Pittsburgh Press.
- Siegel, H. [1989]: ‘Farewell to Feyerabend’, *Inquiry*, **32**.
- Suppe, F., ed. [1977]: *The Structure of Scientific Theories*. (Urbana: University of Illinois Press).
- Suppe, F. [1989]: *The Semantic Conception of Theories and Scientific Realism*. (Urbana: University of Illinois Press).
- Szumilewicz, I. [1977]: ‘Incommensurability and the Rationality of the Development of Science’, *British Journal for the Philosophy of Science*, **28**.
- Theocharis, T. & Psimopoulos, M. [1987]: ‘Where Science Has Gone Wrong’, *Nature*, **329**.
- Thomason, N. [1994]: ‘The Power of ARCHED Hypotheses: Feyerabend's Galileo as a Closet Rationalist’, *British Journal for the Philosophy of Science*, **45**.
- Tibbetts, P. [1976]: ‘Feyerabend on Ideology, Human Happiness, and the Good Life’, *Man and World*, **9**.
- Townsend, B. [1971]: ‘Feyerabend's Pragmatic Theory of Observation and the Comparability of Alternative Theories’, in R.C.Buck & R.S.Cohen (eds.), *PSA 1970*, Boston Studies in the Philosophy of Science, **8**. Dordrecht: D.Reidel.
- Trigg, R.[1973]: *Reason and Commitment*. Cambridge: Cambridge University Press.
- Van Fraassen, B.C. [1980]: *The Scientific Image*. Oxford: Clarendon Press.
- Watkins, J.W.N. [1984]: *Science and Scepticism*. New Jersey: Princeton University Press.
- Weckert, J. [1986]: ‘The Theory-Ladenness of Observations’, *Studies in History and Philosophy of Science*, **17**.

- Weimer, W.B. [1980]: ‘For and Against Method: Reflections on Feyerabend and the Foibles of Philosophy’, *Pre/Text*, **1-2**.
- Werth, R. [1980]: ‘On the Theory-Dependence of Observations’, *Studies in History and Philosophy of Science*, **11**.
- Wisdom, J.O. [1974]: ‘The Incommensurability Thesis’, *Philosophical Studies*, **25**.
- Worrall, J. [1978a]: ‘Against Too Much Method’, *Erkenntnis*, **13**.
- Worrall, J. [1978b]: ‘Is the Empirical Content of a Theory Dependent On Its Rivals?’ in I.Niiniluoto and R.Tuomela (eds.), *The Logic and Epistemology of Scientific Change, Acta Philosophica Fennica*, **30**.
- Worrall, J. [1991]: ‘Feyerabend and the Facts’, in Munévar (ed.), [1991].
- Yates, S. [1984]: ‘Feyerabend's Democratic Relativism’, *Inquiry*, **27**.
- Yates, S. [1985]: ‘More on Democratic Relativism: A Response to Alford’, *Inquiry*, **28**.
- Zahar, E. [1977]: ‘Mach, Einstein, and the Rise of Modern Science’, *British Journal for the Philosophy of Science*, **28**.
- Zahar, E. [1978a]: ‘Theorienkonflikt und die Steurerung der Erkenntnis’, in O.Molden, (ed.), *Konflikt und Ordnung: Europäisches Forum Alpbach*, 1977.
- Zahar, E. [1978b]: ‘Conflict and Order in Science and Methodology’, (Unpublished English translation of Zahar [1978a], typescript).
- Zahar, E. [1981]: ‘Second Thoughts About Machian Positivism: A Reply to Feyerabend’, *British Journal for the Philosophy of Science*, **32**.
- Zahar, E. [1982]: ‘Feyerabend on Observation and Empirical Content’, *British Journal for the Philosophy of Science*, **33**.

Other Internet Resources

- The Feyerabend Discussion List Archive
- Pictures of Paul Feyerabend
- The University of Chicago Press, for information on Feyerabend's Autobiography

Related Entries

analytic philosophy | anarchism | essential vs. accidental properties | Frege, Gottlob | Galileo Galilei | Kuhn, Thomas | Lakatos, Imre | liberalism | logic: inductive | logical positivism | Mach, Ernst | Marxism | meaning | Mill, John Stuart | Nietzsche, Friedrich | paradox: of analysis | Popper, Karl | postmodernism | *Principia Mathematica* | quantum mechanics | rationalism vs. empiricism | realism | relativism | scientific method | scientific realism | social democracy | Vienna Circle | Wittgenstein, Ludwig

John Preston
j.m.preston@reading.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 26, 1997
Content last modified: May 15, 2002

John Stuart Mill

John Stuart Mill (1806-1873), British philosopher, economist, moral and political theorist, and administrator, was the most influential English-speaking philosopher of the nineteenth century. His views are of continuing significance, and are generally recognized to be among the deepest and certainly the most effective defenses of empiricism and of a liberal political view of society and culture. The overall aim of his philosophy is to develop a positive view of the universe and the place of humans in it, one which contributes to the progress of human knowledge, individual freedom and human well-being. His views are not entirely original, having their roots in the British empiricism of John Locke, George Berkeley and David Hume, and in the utilitarianism of Jeremy Bentham. But he gave them a new depth, and his formulations were sufficiently articulate to gain for them a continuing influence among a broad public.

1. Life

John Stuart Mill was born in Petonville, then a suburb of London. He was the eldest son of James Mill, a Scotsman who had come to London and become a leading figure in the group of philosophical radicals which aimed to further the utilitarian philosophy of Jeremy Bentham. John Stuart Mill's mother was Harriet Barrow, who seems to have had very little influence upon him. James Mill's income was at first slight, as he struggled to make his living as a reviewer. But his *History of India* secured him a position in the East India Company and he rose to the post of Chief Examiner, in effect the chief administrator of the Company. In spite of these duties and the work they entailed James spent considerable time on the education of his eldest son. The latter began to learn Greek at three and Latin at eight. By the age of fourteen he had read most of the Greek and Latin classics, had made a wide survey of history, had done extensive work in logic and mathematics, and had mastered the basics of economic theory. This education was undertaken according to the principle of Bentham's associationist psychology, and aimed to make of the younger Mill a leader in views of the philosophical radicals.

At fifteen John Stuart Mill undertook the study of Bentham's various fragments on the theory of legal evidence. These had an inspiring influence on him, fixing in him his life-long goal of reforming the world in the interest of human well-being. At eighteen he spent considerable time and effort at editing these manuscripts into the long coherent treatise that they became in his hands. Guided by his father he threw himself into the work of the philosophical radicals, and began an active literary career. Shortly thereafter, in 1823, his father secured him a junior position in the East India Company. He rose in the ranks, eventually to occupy his father's position of Chief Examiner. A visit to France in 1820 had made Mill thoroughly fluent in the language, and he became a life-long student of French thought and history.

In 1826, Mill suffered a sudden attack of intense depression. This lasted for many months. He continued his work, but internally felt that his former goals were without worth. He came to believe that his capacity for emotion had been severely weakened by his father's rigorous training in analysis. His intellect had been educated but not his feelings. In the reading of Wordsworth's poetry he found something of the cure that he needed, and the depression gradually disappeared.

Mill met Gustave d'Eichtahl in 1828. D'Eichtahl was a follower of St. Simon, and introduced Mill to the latter and to works of Auguste Comte. Mill also met John Sterling who was a disciple of Coleridge. Though these thinkers Mill came to appreciate the role of social and cultural institutions in the historical development of human beings. He became convinced of the Comtean view that social change proceeds through "critical periods," in which old institutions are overthrown, followed "organic periods," in which new forms of social cohesion emerge and are consolidated. He came to believe with these French thinkers that in his own time society was emerging from a critical period. From Coleridge he came to view the educated class as the vehicle for ensuring social cohesion in the emerging organic period.

Mill now saw his task to be that of helping British society to emerge into the coming organic period. The

merely negative polemics of Bentham and his father now seemed very limited. It became necessary not merely to criticize older forms of social organization but to work towards replacing them with something better. Moreover, the defenders of older forms of life should no longer be dismissed as representatives of vested interests. The very fact that the older forms had so long survived meant that they had some good in them, and their defenders should be seen not as reactionaries but as those who continue to recognize that good. If that good is now limited, it still must be acknowledged, and not merely dismissed.

Tactically, the social reformer in critical periods cannot proceed by formulating grand philosophic schemes, however correct they may be in principle. Rather he or she must work for piece- meal reform. Only gradually should general principles be proposed, so that the appearance of radical novelty will be avoided. Mill never abandoned his earlier acceptance of the principle of utility, but now used it positively, not just critically and destructively; he emphasized how it could be deployed constructively, enabling new forms of society to emerge, but ones which incorporate the best of the older forms. He now came to think that the democratic demands of the older radicals had to be tempered with a concern for the dangers which it posed for individualism.

In 1830 Mill was introduced to Harriet Taylor. Her husband was a druggist whose grandfather had once been a neighbor of James Mill. The younger Mill rapidly became intimate with Mrs. Taylor, who came to influence profoundly the rest of his life. She was an invalid who lived apart from her husband. The latter, while he lived, remained remarkably tolerant of the Platonic but very close relationship that Mill and his wife maintained. Mill's father highly disapproved of the connection. When Mill married Mrs. Taylor in 1851 two years after her husband's death there was a complete estrangement from his mother and sisters. Mill reports in his *Autobiography* that Harriet was of crucial significance to his intellectual and moral development. During a trip to Europe in 1858 she died at Avignon, where she is buried. For the rest of his life, Mill spent half a year at a house in Avignon so that he could be near to her grave.

In 1823 Mill had entered the employ of the East India Company as a clerk. India was governed by the Company through correspondence between the Court of Directors in London and the Governors on the subcontinent. This was supervised by the Office of the Examiner of Indian Correspondence. Mill rose through the ranks, and for many years was in charge of the correspondence dealing with the princely states not under the direct rule of the Company. In 1856 he became Chief Examiner, in charge of all correspondence, succeeding another utilitarian, Thomas Love Peacock, who had succeed James Mill. After the Indian Mutiny, the British parliament proposed the dissolution of the Company. Mill prepared a vigorous defense of the Company and of the government that it provided, but it was unsuccessful, and he retired on a reasonable pension in 1858.

In 1865 Mill was elected to the House of Commons. Given his reputation and his previous seclusion, his work was subject to immense attention. His performance was generally acclaimed, but he failed in his attempt at re-election in 1868. He continued to work, as he had earlier in his life, for many radical causes. He was particularly concerned for the status of women. His later work was made easier by the cooperation of Mrs. Taylor's daughter, Helen, who in many respects took the latter's place in Mill's life. A number of his important works were published posthumously by Helen Taylor. Mill died in 1873 at Avignon, where he is buried next to his wife.

Mill made his philosophical reputation with his *System of Logic*, which he published in 1843; this work re-vitalized the study of logic, and provided for the remainder of the century the definitive account of the philosophy of science and social science. This was followed by *The Principles of Political Economy* in 1848; this defined the orthodox form of liberal principles for the next quarter century. In 1861 he published his only systematic treatise in first philosophy. This was his *Examination of Sir William Hamilton's Philosophy*, a comprehensive critique of the latter's rationalism and intuitionism. So effective was Mill's critique that this work effectively dated itself and is now unfortunately neglected. His two best-known works in moral philosophy were *On Liberty* and *Utilitarianism*, which appeared in 1859 and 1861 respectively. These are of continuing significance. His *Considerations on Representative Government*, published in 1851, is perhaps now less important than his essay on *The Subjection of Women* (1869). Mill's partially finished *Autobiography* was published, with additions by Helen Taylor, in 1873. She also saw for the posthumous publication in 1874 of his *Three Essays on Religion*.

2. Language and Logic

In the *System of Logic,* Mill accepts the traditional doctrine that propositions as used to describe the world are divided into subject and predicate terms, or, as he would say, names, joined by a copula, either affirmative or negative. Among names there are singular and general. All names denote, either individuals or the attributes of individuals. Names are either singular or general. A general name connotes an attribute and denotes all individuals which have that attribute. Thus, 'white' connotes the attribute whiteness, and denotes all things that have that attribute. Some singular names only denote; they have no connotation. 'Caesar' is such a name. However, many singular names not only denote, they also have a connotation. Thus, 'the conqueror of Gaul' is a singular name. It denotes the same individual as is denoted by 'Caesar', but unlike 'Caesar' it also connotes; it connotes the attribute of being a conqueror of Gaul. In terms of logic, Mill is less concerned that later thinkers would be about the uniqueness implied by the connotation: Caesar is not only *a* conqueror of Gaul but *the* conqueror.

In a proposition, names are joined by a copula, either affirmative or negative. The meaning of a proposition -- its "import", as Mill says -- is determined by the connotation of its parts, the sole exception being given in the case of proper names, where the meaning is determined by the denotation.

Where the import of a proposition is given by connotation, truth or falsity is determined by denotation. An affirmative proposition is true just in case that the thing or things denoted by the subject term are in the class of things denoted by the predicate term; otherwise it is false. Similarly, a negative proposition is true just in case that no thing denoted by the subject term is a member of the class of things denoted by the predicate term. Things and attributes are always such that any proposition is either true or false and not both. This states the Principles of Non-contradiction and of Excluded Middle. No thing or attribute is such that is can be said to be both wholly itself but also necessarily connected to something other than itself: each thing or attribute is logically and ontologically independent of every other thing or attribute.

Since the holding of these principles depends upon the systematic nature of things and their attributes, it

follows that the truth of these principles is in the end a fact about the world and the things in the world -- the deepest, perhaps, of all metaphysical facts about the world and the entities in it.

The world we talk about in our propositions is the world that we come to know it in our ordinary sense experience or inner awareness. The ontology of the world as reflected in language and logic is the ontology of the world as we know it to be. Knowing the meaning of terms and therefore of the import of propositions is knowing the individuals and attributes which they denote and connote. As we know them in our ordinary experience these individuals and attributes are logically self-contained. The rationalists and the Aristotelians argued that beyond our ordinary experience of things we have intuitions -- "rational" intuitions -- of ontological connections that structure things in ways not apparent in our ordinary sense experience of the world. Empiricism is the claim that there is no such rational intuition and nothing in the ontology of the world beyond what we know in ordinary experience. The world of the empiricist is one without necessary connections among individuals and attributes. It is this deep fact about the world that is the metaphysical basis of the truth of the Principles of Non-contradiction and of Excluded Middle.

There is no metaphysical necessity. All necessity is verbal, a matter of the import of propositions. A proposition is necessarily true in the case of connotative names just in case that the connotation of the names is by convention the same, as in ‘Bachelors are unmarried’. Since the connotations are the same, the set of individuals denoted by the subject term is identical with the set of individuals denoted by the predicate term, and the proposition is true, simply by virtue of the verbal conventions. In the case of proper names, where the terms do not connote, as in ‘Cicero is Cicero’, the individual denoted by the one term is precisely that denoted by the other, and so its truth is again verbal.

This account of the meaning of propositions is not complete. Mill does not consider the contrasting cases of ‘Cicero is Cicero’ and ‘Cicero is Tully’. Since ‘Cicero’ and ‘Tully’ are both proper names and therefore purely denotative, it would seem to follow on Mill’s account that, contrary to fact, the two propositions have the same import. Later logicians would work hard to solve some of these difficulties in Mill’s semantics. They would also have to work on problems arising from connotative proper names ("definite descriptions").

Mill’s empiricist thesis, that all necessity is verbal, has important consequences for his account of logic.

Consider the syllogism:

> Man is mortal
> Socrates is a man
> Ergo, Socrates is mortal

It had been argued by Aristotelean and rationalist defenders of logic as a system of necessary science that it is a science that contributed to the growth of knowledge. The inference from ‘Socrates is a man’ to ‘Socrates is mortal’ is mediated, on this view, by the major premise ‘Man is mortal’ which establishes a

real necessary connection between the minor premise and the conclusion by virtue of itself recording a necessary connection among the attributes connoted by the terms which it contains. This traditional account presupposes that there are necessary connections among attributes, that, in other words, attributes are not logically self-contained. Given Mill's claim that attributes are logically independent, this is wrong and the truth of the major adds nothing to the truth of the particular propositions, 'this man is mortal,' 'that man is mortal', etc., whose conjunction it records. Accepting the major as true is simply a way, on the one hand, of accepting that particulars one already knows share the attributes in question, and, on the other hand, a determination that one will continue to affirm this connection of hitherto unexamined particulars.

On Mill's semantics, then, the major premise 'Men are mortal' of the syllogism is, in its import, semantically equivalent to the extended conjunction:

> Peter is a man and Peter is mortal, & Caius is a man and Caius is mortal, & Cicero is a man and Cicero is mortal, & etc.

It is, in other words a way of asserting an indefinitely long conjunction. Now, an inference

> Peter is a man and Peter is mortal, & Caius a a man and Caius is mortal
>
> Hence, Peter is a man and Peter is mortal

is not a genuine inference, one that is ampliative, yielding an increase of truth. It is only an apparent inference. Given the import as a conjunction of the major premise in the syllogism,

> Man is mortal
> Socrates is a man
> Hence, Socrates is mortal

it follows that this too is only an apparent inference. Thus, logic -- that is, deductive logic or syllogistic -- adds nothing to our knowledge; its rules merely reflect our determination to reason consistently with the ways in which we have reasoned in the past: the rules of formal logic, of syllogistic, are the rules of a logic of consistency.

In contrast, logic as ampliative involves a passage from the knowledge of particulars summarized in the major, or universal premise, to an application of the same rule to a new case. In terms of knowledge, the major premise of a syllogism provides no knowledge beyond what it summarizes about particulars already known. Ampliative inference, that is, inference as a part of a logic of truth, is thus always a passage from particulars to particulars.

3. Induction

Since there are no objective necessary connections among the attributes of phenomena given in ordinary experience, it follows that the only grounds that we have for inferring from a sample to a population or from the past to the future are given by present experience or memory: all ampliative inference is inductive. It follows that such inference can never achieve apodictic certainty. This does not imply, as some have suggested, a scepticism about all events beyond present experience; it does not imply that all such judgments are somehow unreasonable or unjustified. It does imply, however, that our notion of rational justification ought to be adapted to this fact.

Humans find themselves as embodied creatures in the world, making "spontaneous" and "unscientific" inductions about specific unconnected and natural phenomena or aspects of experience. Examples are "fire burns" and "food nourishes". The satisfaction of our desires, and indeed our very survival, depends upon our coming to ascertain, so far as we can, the truth about the natural world in which we find ourselves and about ourselves too. Ascertaining such truth as best we can is the cognitive means we have available for meeting those ends. No judgment aiming at such truth can, however, ever attain apodictic certainty, and so we ought, as reasonable beings at least, to be satisfied with less than that. In the absence of infallible knowledge, we ought as reasonable persons be satisfied with fallible knowledge. And within that framework we ought to find as best we can those rules of inference that on the basis of past experience form the best -- though fallible -- guide to matter of fact truth.

Deductive logic keeps us consistent in our search after matter-of-fact truth; the logic of science -- inductive logic -- provides a set of rules that form the best, though fallible, guide that we can discover for the discovery of new truth. The rules of logic, both deductive and inductive, are rules of the art that has as its end, its cognitive end, the search after truth. All inferences are matters of psychological fact. For this Mill was later criticized by some later philosophers such as Husserl; he was accused of the sin of psychologism. But this is unfair to Mill. The latter is not claiming that the laws of logic are part of the subject-matter of the empirical science of psychology. He is arguing, rather, that the laws of logic, of both deductive logic and inductive logic, are normative, rules or standards about how we ought to reason, or, at least, about how we ought to reason given that we have a concern for matter-of-fact truth.

4. Mill’s Empiricism: The Relativity of Knowledge

Mill argues that the apparatus of logic permits us to define predicates such as ‘unicorn’ that connote attributes that are not present in things given in ordinary experience. Such predicates have no denotation, and any proposition such as ‘this is a unicorn’ is false. More difficult are subject terms which connote but do not denote, e.g., ‘the present king of Mexico’. Mill was not alone in having difficulty with the logic of such terms; it was only with the work of Bertrand Russell on definite descriptions that these problems were solved in a way that could fit with an empiricist account of logic and of meaning.

But Mill’s basic point is clear enough. Language aims to state matters of fact about the world. There are logical terms, such as ‘is’ or ‘is not’ or ‘and’, and non-logical terms. The latter are the subject terms and predicate terms of propositions. There are certain subject terms and predicate terms which are primitive

to this language. All others are somehow defined on the basis of these primitive terms -- leaving aside the details, to be worked out by later logicians, how these rules for introducing non-primitive terms are to be specified. Meanings are a matter of subject terms and predicate terms being hooked as it were to things and their attributes. (Mill is hazy, unfortunately, on relations, but in this he is hardly to be distinguished from any logicians prior to Peirce and Russell.) It is Mill's empiricism that these things and attributes to which the primitive terms of language are hooked are presented to us in ordinary experience, either sensory experience or inner awareness of our own states of consciousness.

Subject terms and predicate terms provide the content of propositions. Assuming that thought is propositional, they thereby define the limits of thought. Mill's empiricism thus determines the limits of what is thinkable. In this sense, all knowledge is relative to us, to our consciousness. We can of course have beliefs and even knowledge of things of which we are not conscious; there are parts of the world that we have never experienced. There are parts of things too small to observe by means of unaided sense; there are things too far to be seen by unaided sense; and there are things, such as the inside of unopened oranges, that we do not see but which we would see were we to do certain things, e.g., cut open the orange. The grounds of such knowledge or beliefs are not direct experience, but rather inference from direct experience. But our knowledge of those things, our beliefs about them, are still relative to us in the sense that we cannot think of them except as similar to or resembling things or attributes of which we are conscious in sense experience or inner awareness.

This much Mill takes for granted in his *System of Logic*. The empiricist framework he defends in detail in the *Examination of Sir William Hamilton's Philosophy*.

Some philosophers, for example, George Berkeley, have argued that the only grounds that can justify beliefs about ordinary things is direct consciousness. Mill rejects such an idealism: there is nothing in the being of attributes of things that ontologically determines that such things when they exist must be sensed.

Mill also rejects the view of other philosophers, e.g., Kant or Sir William Hamiliton, that there are things or entities beyond phenomena. According to these philosophers, phenomena are in fact the effects of such things but those things, as trans-phenomenal or noumenal, are wholly different from those phenomena: they or it constitute the Unknowable cause or causes of the phenomena of which we are ordinarily aware. Among these philosophers, some such as Kant held that all that we can know is that the entities exist as causes of phenomena. Others such as Sir William Hamilton argued that we can know these not only as causes but also as being in themselves characterized by attributes, e..g., the primary qualities, which are also given phenomenally. Finally others, such as Hamilton's follower H. E. Mansel, argued that we can know the Unknowable as having attributes that at once resemble those that are had by ordinary things, e.g., some of the human virtues, but in ways which exceed our human powers to conceive -- the suggestion here is that God, the Unknowable, is just but in a way that exceeds all humanly conceivable forms of justice.

Mill objects to these views in the name of logic: they are simply not consistent with the logic that is appropriate to the claim accepted by these philosophers that all knowledge is relative. In the first place,

the concept of cause is, contrary to Kant and Hamilton, not an *a priori* concept. The concept of cause in its basic sense is acquired through our experience of matter-of-fact regularity: it is one that relates phenomena to other phenomena and not phenomena to noumena. The relativity of knowledge includes the relativity of our knowledge of causal relations, and these philosophers therefore have no grounds for supposing the existence of noumenal entities. As for Hamilton's claim that we have in our phenomenal experience knowledge of attributes of noumenal entities as these are in themselves, Mill has no trouble in showing the logical confusion in this claim. Hamilton wants to have it both ways -- all our knowledge is relative or phenomenal and within this phenomenal realm we can discover concepts which apply absolutely or non-relatively to the noumenal things-in-themselves. Mansel's views receive Mill's particular scorn: if terms are not to be used in their ordinary sense then they ought not be used at all. A being, no matter how powerful, whose acts cannot be described in terms characteristic of human morality, is not one that we can reasonably worship. Mill made his well-known proclamation that "I will call no being good who is not what I mean when I apply that epithet to my fellow creatures, and if such a being can sentence me to hell for not so calling him, then to hell I will go." (*Examination of Sir William Hamilton's Philosophy*, p. 103)

5. Scientific Method

Mill works out the basic principles of experimental science in the *System of Logic*, Book III.

He argues that the rules of scientific method evolve out of the spontaneous inductions about the world that we make as embodied creatures. As we investigate the world to find the best means to satisfy our natural needs and aims, some patterns maintain themselves, others turn out to be false leads. The former guide us in our anticipations of nature, and in our plans; they enable us to infer what will be and what would be if we were to do certain things or if other things were to happen. These patterns that we accept as guides we come to think of as laws: a law is a regularity that we accept for purposes of prediction and contrary-to-fact inference. Out of these ways of experiencing and coming to understand the world grows our account of explanation: to explain a fact is to locate a law under which it can be subsumed.

As we proceed in our efforts to understand and explain the world in which we find ourselves, the generalizations we accept begin to accumulate and interweave. We find, moreover, that there are not only generalizations but also at a more generic level patterns among these generalizations. We discover that as we search out such patterns we are often successful in discovering them. This itself is a pattern, a pattern about patterns, or, as Mill put it, a law about laws, a law to the effect that for all sorts of event there are laws, there to be discovered, which explain those sorts of event. This is a generic law to the effect that for every specific sort of event there is another specific sort to which it is regularly or lawfully connected. As we progress in science we generalize this law about laws to include all sorts of events: it is the Law of Universal Causation. It assures us that for any sort of event there are laws, there in the world, which, if we search diligently enough, we will be able to discover.

Guided by this principle, we also discover -- fallibly, to be sure -- that various rules of inference are more effective than others in generating acceptable causal beliefs. At the specific level, for any sort of event

there are a wide variety of alternative possible determining attributes. We make assumptions -- revisable -- about the possible relevant causes, and identify the actual sort of cause by a process of elimination. Mill provides a detailed study of the various rules of eliminative inference in his well-known "Methods of Experimental Inference" (*System of Logic*, Bk. 2, ch. 9).

At the level of specific sorts of event the rule of enumerative induction -- "from all observed A's are B's infer that all A's are B's" -- is unreliable, often leading us to accept as true generalizations that later turn out to be false. Inferences in conformity to this rule often overlook relevant factors that are the real causes or mistake necessary conditions for sufficient conditions. Because of the variety of possible factors at the specific level, the eliminative methods lead to judgments that are more secure than the judgements which mere enumerative judgments would yield.

For any event, however, if one takes it generically then there are at that level only two alternatives, caused or uncaused. Here there are not a variety of alternatives to be eliminated, and the rule of enumerative induction turns out to be more reliable than at the specific level: we have regularly been successful in finding causes at the specific level and this provides grounds for accepting the generic claim that for all events there are causes to be discovered.

But then, as we extend our researches to a new specific area, the thesis that there are causes to be discovered provides inductive support for the lower-level claim that the result of the eliminative inference has really isolated a cause. There is thus an interplay as it were between eliminative and enumerative methods, with inferences at the specific level providing support for generic-level inferences, and the latter in turn providing support for specific-level inferences. Confirmational support rises up a hierarchy from specific laws to laws about laws, and then down from the laws about laws to the specific laws. Inferences in different specific areas mutually support one another through their joint support of the Law of Universal Causation which provides the grounding principle for inductive inference.

Mill's great opponent about the logic of scientific inference was William Whewell, whose *History of the Inductive Sciences* (1837) Mill had read with considerable care. Whewell argued that acceptance of scientific hypotheses depends first upon their capacity to explain observed phenomena and more specifically upon their capacity to explain phenomena in diverse areas (the "consilience of inductions"). Mill can accept the point about the consilience of inductions, given what he argues about the interplay of enumerative and eliminative reasoning. Further, Mill could accept the importance of hypotheses as working assumptions -- "heuristic devices," in Whewell's terms -- , but he could not accept Whewell's claim that the mere fact that an hypothesis accounts for the data provides safe grounds for accepting it as a true statement of law. To be sure, eliminative methods may often show that a working hypothesis is in fact the only one consistent with the facts, and that it is therefore acceptable as true. But Whewell argued on the basis of the history of science that there are cases of hypotheses where the supposed causes have not been observed and which yet seem to yield explanations of observable phenomena. Such hypotheses involving unobserved causes can be found in inferences about areas too small or too distant to be observed -- Whewell instances the undulatory theory of light. Mill agrees that there are such cases, and even allows that such hypotheses provide useful analogies for the guidance of future research. But so long as the data do not determine a unique hypothesis, such hypotheses cannot be accepted as yielding a

new truth.

For Whewell, consilience is effected by generic hypotheses subsuming under themselves more specific hypotheses. These hypotheses involve generic concepts that, in Whewell's terms, "colligate" the more specific concepts that appear in hypotheses further down the ladder. Genuine progress in science depends not so much upon simple generalization from observed data as from the locating by inventive genius new colligating concepts. Mill does not disagree, but argues, contrary to Whewell, that colligation by itself is no test of truth.

It is Whewell's contention that as new colligating ideas emerge in the history of science, the principles in which they are embedded become necessary. The concepts in these axioms, such as 'cause' or 'force', are *a priori*, and research consists in gradually articulating these concepts into principles the necessity of which becomes more evident over time. What Mill would allow to be the free action of creative genius, Whewell construes as the uncovering by the mind of the divine ideas that provide the formal structure in conformity to which the Unknowable Creator constructs the world of phenomena. It is this necessity deriving from the Divine Creator that guarantees the truth of the basic axioms that organize scientific theories and which ensures the consilience of inductions. Needless to say, Mill rejects this account. The claim that some concepts have their origin *a priori* is inconsistent with the guiding thesis of the relativity of knowledge. Mill does not deny that in the process of scientific investigation, basic axioms become indubitable in the sense that their contraries become inconceivable. But such indubitability is psychological and does not derive from some sort of conformity to divine necessity. The truth of such axioms, if it really does obtain, is a matter of their conformity to the way the phenomenal world is, and mere fact that they are psychologically indubitable to the human mind does not guarantee that: given the relativity of knowledge, even indubitable judgements are fallible.

6. The Science of Psychology: Associationism

Central to both Mill's account of human reason and also to his social projects is his account, deriving from Bentham and his father, of the science of the human mind. This theory of how the human mind derives originally from Aristotle's discussion of associative memory. In Mill's hands it becomes a systematic hypothesis about which regularities govern human learning. Mill himself never wrote a systematic treatise on psychology, but late in his life he reprinted his father's *Analysis of the Phaenomena of the Human Mind* (1829, second edition, ed. J. S. Mill, 1869), with extensive notes revising and correcting his father's work.

The theory proposes that if f and g are regularly presented in experience as standing in relation R, then the habit forms in the mind, that if we have an impression or idea of f then it is accompanied with the idea of g. If R is the relation of spatio-temporal contiguity, then the ideas are joined to form a judgment of regularity, a causal judgment. If R is the relation of resemblance, then the ideas associated in the mind according to resemblance classes. A few experiences of connection will produce a loose connection in the mind. An increased frequency of experienced connection will produce a stronger association in the mind. And so attributes that are logically separate in experience through being repeatedly experienced in

conjunction come to be inseparably connected in the mind.

Mill argues that this theory can account, however sketchily, for our use of language and how it becomes meaningful. If a word ‘w’ comes causally to be associated with a kind f, and f is associated with the resembling kind g, then the presence of f to the mind will call up both the associated kind g and the word w, so that w will come to be associated not only with f but also with g, and in fact with all the kinds that stand in the resemblance relation R to f. In this way words become general, applying to all members of a resemblance class. The mechanisms of association and the relations of resemblance thus come, for Mill, to play the role that abstract ideas played for earlier philosophers.

The habits of causal inference provide ways of anticipating what will occur in the world; that is how we learn what to expect. It is through processes of this sort that we form our spontaneous generalizations such as "fire burns" of "food nourishes." Cognitively, these are inferences that conform to the rules of induction by simple enumeration. And in terms of our adapting to the world, these inferences are acquired purely passively. If this was all there is to Mill’s account to the human mind then the criticism often level by the idealists that it ignores the active element in the human intellect, and presents a simplistic view of human reason.

But Mill’s psychology also includes an account of motivation and action. On this theory, pleasure is the prime motivator, the primary end in itself, and the anticipation of pleasure serves as an immediate cause of bodily motions which in turn bring about that pleasure. Through regular success in attaining pleasure, anticipations of pleasure become associated with the sorts of action that bring about that pleasure. When Mill asserts that people seek pleasure, what he is to be taken to mean is that people seek things other than pleasure but that they seek it because pleasure has become associated with it, and that when the desire is fulfilled they experience the pleasure of satisfied desire. In this sense human welfare consists in satisfied desire.

There is one important feature of Mill’s psychology in which he differed from his father. On his father’s view, a complex idea produced by association is simply a collection is its associated parts. On Mill’s view, however, there is a sort of mental chemistry in which the parts fuse, as it were, into a new sort of mental whole. These new sorts of mental unity emerge from associational processes and have properties which are not had by the properties that appear in the genetic antecedents.

Given the account of association and of action, it is evident that various means to pleasure will become associated with feelings of pleasure. But on Mill’s view, this will not be a mere conjunction; to the contrary, as the association becomes strong enough the two parts will fuse into a new sort of emergent whole. The means will not simply be conjoined to pleasure but will become part of pleasure. And so money, for the miser, becomes not just a means to pleasure but for him part of pleasure, a end in itself.

This account of human action presupposes the acceptance of determinism, which Mill vigorously defends in the *System of Logic*, where he outlines the idea of a naturalistic science of human being. Freedom, Mill argues in Book Six, Ch. 2, which he thought the best in the work, is not the absence of causation but

rather the absence of coercion. In fact the whole point of education is to determine the future free actions of the individual: it aims through the associative processes to determine the person's motives and actions.

This much Mill takes over from earlier thinkers such as Hume. But leaving it here has seemed by some again to render the person passive. Mill takes up this point in detail in *Examination of Sir William Hamilton's Philosophy*. The argument of the critics is that if character and motive are determined by earlier causes, how could a person be said him- or herself to be responsible for his or her actions? Hamilton so argued: the view makes the person a creature of his or her environment. This notion seemed convincing to Mill himself until he came to recognize that among the motives that one could acquire is the motive of self-improvement or self-realization. There are irresistible motives; for these we are not as persons responsible. But there are also resistible motives, and these we can shape and determine. That is, we can shape and determine them provided that we have the desire so to do. One is free if one could have resisted the motives on which one did in fact act, provided there had been good reasons so to do. A motive impairs freedom only if it is irresistible, only if it cannot be blocked by a strong reason against it. The free person is one who is sensitive to good reasons for behaving as he or she does. The second-order ends that lead one to shape one's motives and to develop as an individual became the central feature of Mill's social thinking, and this marks a major break in detail, though, to be sure, not in principle, with the utilitarianism of Bentham and his father. In the *Examination of Hamilton's Philosophy*, Mill vigorously defends the notion of human beings as active in their own self-determination.

This account of human being also provides an answer to those who argue that Mill's picture of human reason makes persons purely passive rather than active as thinkers. For, among the ends that can come to be associated with pleasure is the end of truth; in this way curiosity becomes an end in itself. And the motive for self-improvement will lead us to find, so far as we can, better ways to satisfy that end. We will so educate ourselves that our reasoning will conform, not to the simplistic rules of induction by simple enumeration, but to the more reliable rules of eliminative induction. The charge that Mill fails to take into account the active side of human reason is thus mistaken, resting on a failure to recognize those parts of the psychological theory that deal with motivation.

7. Geometry and Arithmetic

It is within this context that one must place the account of the necessity of geometrical and arithmetical truth that Mill develops in the *System of Logic*.

The truths of geometry and arithmetic had traditionally been taken to be necessary. But they clearly have more than verbal import. They are therefore not necessary truths, given Mill's argument that the only necessity is verbal necessity: on Mill's metaphysics, therefore, they depend for their truth upon the individuals and their attributes of the world as we experience those entities.

The propositions of geometry are empirical. The theorems are deduced from premises which are themselves inductively established. These premises are inexact descriptions of objects in physical space. Insofar as the premises are exact descriptions -- referring, e.g., to exactly straight lines -- they describe

material attributes taken to their limits. Thus, all smooth lines resemble one another with regard to different degrees of curvature, and taking a line to be exactly straight is to neglect the degree of curvature. The proposition, "two straight lines cannot enclose a space", when taken as applying literally, is only inexactly true; taken as exactly true it means something to the effect that "The more closely two smooth lines approach absolute breathlessness and straightness, the smaller the space that they enclose." Mill's views on geometry are close to those of the logical positivists in the twentieth century.

His views on arithmetic are more controversial. These were later to be vehemently disputed by the logician Gottlob Frege, not without good grounds. Mill disagreed with those whom he called Conceptualists, who held that arithmetical truths were truths about psychological states. Mill also agreed with Kant against Nominalists such as Hobbes that the propositions of arithmetic are not true by definition; they are, in Kantian terms, synthetic. But that implies, for Mill, against Kant, that they are *a posteriori*, inductive rather than *a priori*. The only way that Mill could see one holding that they are both synthetic and *a priori*, is to hold that they are truths about rationally intuited forms not presented in ordinary experience. This was the solution that Frege was later to adopt. But Mill on empiricist grounds rejected this sort of Realism. This makes Mill in more recent terminology a nominalist. The problem is that arithmetic seems to have a necessity which is at once more than verbal, as Mill correctly held, but also more than that which attaches to the inductive truths of, say, physics or botany. Mill's ontology of things and attributes is simply not sophisticated enough to permit a solution to this problem.

Mill argues that a number is an attribute of an aggregate of units. This brings him close to Frege's idea that the number of a given class is the class of all classes equinumerous to that given class. But he does not clearly distinguish an aggregate from a class, nor the sum of two numbers from the (Boolean) sum of two classes. Moreover, he takes measurement to be the empirical counting of units, rather than a matter of relations among the members of an ordered dimension. In both cases a more sophisticated account of relational form is necessary, but this was developed only by later logicians. Mill is certainly confused from the point of view of later thinkers such as Frege or Russell. Certainly, the view of the later positivists that mathematical truths are a matter of logical form would fit more comfortably with his empiricism.

What Mill does argue about the necessity of geometry and arithmetic, and, for that matter, the basic axioms of other sciences such as physics and chemistry, is that these principles, while from the point of view of their truth are inductive generalizations, are from the point of view of the thinker matters of psychological necessity. The appeal is to the principles of association. The propositions of geometry and arithmetic record matters of fact that are very deep and invariable in our experience. Our repeated experience of these facts creates in the mind invariable associations. These inseparable connections create in the mind of the knower a sense of the necessity of these propositions. The necessity is there, as Whewell and others insist. But the necessity is one of thought rather than one in the ontological structure of things.

8. Perception and Material Things

In his *Examination of Sir William Hamilton's Philosophy*, Mill applies empiricist principles to the ontology of material things and his associationist principles to their perception.

When we cut open an orange we are presented with certain sensory impressions, shapes, colors and textures with which we were not previously presented. However, we also firmly believe that those parts of the orange were there even when we were not perceiving them. Our experience has so formed our habits of expectation that we not only form the conception of those things as existing when they were not being perceived but firmly believe them so to exist. These are things, parts of the orange, existing unperceived; they are possible sensations, which, through our expectations, have become conditional certainties. Mill refers to these possibilities which are conditional certainties as "permanent possibilities", thus distinguishing them from mere vague possibilities which experience gives us no warrant for reckoning upon.

It is important to note that, while we do not experience these permanent possibilities, they are not mere fictions. To the contrary, as just indicated, Mill carefully distinguishes between the permanent possibilities that constitute ordinary things from the mere or "vague" possibilities that we conjure up in our imagination. The acceptance of these possibilities is a matter of certainty, though, to be sure, a certainty that is conditional, based on inference from what we do actually experience. With regard to the ontology of ordinary objects, Mill is a phenomenalist, but among the parts of those things are unexperienced phenomena.

(Mill's "permanent possibilities of sensation" are what later philosophers such as Bertrand Russell would refer to as "unsensed sensibilia." Mill's phenomenalism is similar to what they were to call "neutral monism." The later philosophers differ in the more sophisticated logical apparatus that they could bring to bear.)

Ordinary things, physical objects, are clusters of sensations, actual and possible, that is, permanent, in Mill's sense. These clusters are lawfully ordered; it is our knowledge of these laws or regularities that make the permanent sensations conditional certainties. The clusters include not only visual but also tactual and other forms of sensation. Ideas of depth arise from associations of kinaesthetic sensations that arise as we move from here to there. At a certain point, here, there are visual sensations of color and shape. At another point there are different visual sensations, perhaps the same color, but a different shape -- things are seen in perspective. Also at the other point that shape is presented not only visually but tactually. Relative to the actual experience of the former the others are conditionally certain possibilities located at the appropriate distance.

When we perceive an orange we have certain visual sensations which through our expectations we refer to a collection that includes not only these actual sensations but also the permanent possibilities that are there but which we are not sensing. A perceiving is in effect an associational inference from given sensations to things taken as clusters of sensory parts, most of which are there as unperceived but permanent parts. Like all inferences those inferences are associations of ideas. But these perceivings are so ingrained as to be in effect instantaneous. The ideas which are their parts fuse into a single whole. Through the chemistry of association the perceiving of an ordinary thing is an emergent unity, a new

whole which has that thing as its cognitive or intentional object.

In experience we often find that whenever a given cluster of certified possibilities of sensation obtains, then a certain other cluster follows. In such a context through a further process of association our ideas of causation, power, and activity become connected not with sensations but with groups of possibilities of sensations. The perceptual object thus comes to be thought of as having the power of producing sensations. It becomes the permanent material source of the sensory data that we actually experience. As far as it goes, this inference is legitimate. But there is a tendency of the human mind to transform this material object into a noumenal object that thought somehow to exist apart from all sensory appearances. But it is precisely this tendency that Mill decries as illegitimate: this is his empiricism. "I assume only the tendency, but not the legitimacy of the tendency, to expand all the laws of our own experience to a sphere beyond our experience" (*Examination of Sir William Hamilton's Philosophy*, Ch. XI, p. 187n).

9. Minds

There are two cases, other minds and one's own. Mill discusses both in the *Examination of Sir William Hamilton's Philosophy* (Ch. XII, Appendix).

Among the bodies to which one refers one's sensations, there is one that is as it were peculiar. That is, one stands in a peculiar relationship to it. One is aware of it from the inside. For this body alone one is aware of kinesthetic sensations. One's perceivings locate other bodies at a distance to this one. Our motives and volitions move this body directly in ways that they can move no other body.

Now, there are regularities that connect outward actions of one's body with states of consciousness within that body. These would include patterns such as this: "Whenever my arm goes up there is a consciousness of my body from the inside that contains a willing that my arm go up." These regularities are verified in one's own case. But they can be used to infer the existence of conscious states within other bodies that exhibit the same outward actions as one's own body. Thus, whenever I observe the arm of another going up I can infer that there is a consciousness of that body from the inside that contains a volition that the arm go up. The regularities that obtain in one's own case render the existence of such conscious states in others conditional certainties.

The inference to other minds is thus perfectly reasonable. It is based on two facts, one the peculiar relationship that one's own conscious states have to one's body and the regularities that obtain in one's own experience between one's own conscious states and one' body. The former accounts for the privacy of conscious states, the latter justifies the inference to the presence of similar private states in others.

It is worth noting that many have suggested that our knowledge of other minds is based on an argument from analogy. On Mill's view this is not so. The inference is a simple causal inference. Nor is it an inference based on a single case. To be sure, the regularities are verified in one's own case, but the facts that verify them are the repeated instances that they describe. Nor is privacy a problem. When I infer

from a bodily state to the presence of another mind, the consciousness to which I infer is an awareness of that body from the inside. Since I am aware of only my own body from the inside and not that of any other, I should expect to consciousness to which I infer to be private to the other person.

As for the problem of mind in one's own case, this is more difficult. What is mind? Matter is resolved by Mill into a lawfully related bundle of sensations including many permanent possibilities of sensation. Can one's own mind similarly be resolved into a bundle of feelings with a background of permanent possibilities? The problem is that when I expect or remember a state of consciousness I do not simply believe that is has or will exist; it is also to believe that *I myself* have experienced or will experience that state of consciousness. If it is a series or bundle then it is a series or bundle in which a part of the bundle is conscious of the whole. This had been an objection to the bundle view ever since Plotinus used it against the Epicureans. Mill simply accepts the reality of such awareness. If we accept the bundle view, rejecting the common view of mind as a substance, as he thinks we must, then we are reduced to "accepting the paradox that something which *ex hypothesi* is but a series of feelings, can be aware of itself as a series" (*Examination of Sir William Hamilton's Philosophy*, Ch. XII, p. 194). He thus sees himself as driven to "ascribe a reality to the Ego -- to my own Mind -- different from that real existence as a Permanent Possibility, which is the only reality I acknowledge in Matter" (*ibid*.).

10. Moral Sciences

The Sixth and final book of the *System of Logic* is a masterful account of the methodology of the social or moral sciences, one that still repays detailed attention. Its strength derives not only from the thorough and systematic approach to the issues, but also from the fact that Mill himself practised the whole body of these sciences as they then existed, from psychology of course, through economics -- in his own day he was recognized as the leading political economist -- to history and the then emerging science of sociology. He thought as an economist as well as a moral philosopher about socialism, taxation, and democracy, and he thought not merely as a social thinker about the institutions that govern society but also as a colonial administrator and as a politician in his own country.

The basic pattern of explanation -- subsumption under matter-of-fact regularity -- applies to the realm of the social as it does to the physical and the mental. Idealists were of course to raise the possibility that the human escapes the natural casual order, and requires a form of explanation *toto caelo* different from that of the physical sciences. But with his basic arguments for the relativity of knowledge and for the idea that minds are bundles of perceptions, Mill rejects all such proposals. So within his framework, there are no problems in principle with the idea of a natural science of human being. There are problems, to be sure, but they are problems of detail not of principle.

The major problem in the social sciences is that of complexity. For single individuals, the experimental methods can be applied in much the same way that they can be applied in physics or biology: the science of psychology raises no problems. But with large groups those methods cannot be used; the fact that there are a large number of interacting variable precludes that. Other methods are needed. Mill suggests such methods. These proposals depend upon his views on social relations, his ontology of social structures, if

you wish.

According to Mill "the effect produced, in social phenomena, by any complex set of circumstances, amounts precisely to the sum of the effects of circumstances taken singly" (*Logic*, VI, ix, 1). The same principle does not hold in every science; in chemistry, for example, or in psychology itself, effects often have properties which are not reducible to the properties of the causes. In these cases the laws are said to be "chemical": the resultants have properties that are not present in the causes. But in the case of social phenomena, there is nothing in the resultant whole that is not already in the parts; the resultant whole is simply, as Mill says, the "sum" of the parts. In these cases, then, there is two fold process of inference. First, we consider each cause that is operating and use the science of psychology to infer what effect it would have. Second, we then deduce the laws for the group, that is, the social laws. The deduction is direct, since the social cause is the sum of the individual causes taken as parts, and the social effect is the sum of the individual effects taken as parts.

Thus, it is possible to discover the laws for the group phenomena simply by deducing them from the assumed conjunction of the many single causes. This is what Mill calls the "physical" or "concrete deductive" method (*Logic*, VI, ix). It can be used in political economy, where one assumes everyone is acting on the motive of preferring the greater gain to the smaller (*Logic*, VI, ix, 3). However, in other social phenomena there are many more motives, many more causes, as in sociology and history, and here it is necessary to trace over time the detailed effects of all the many causes. However, this detailed set of inferences is beyond the powers of human computation. The best that we can do is begin with the empirical laws of social phenomena and show by deduction that this was likely to result from what we know of the nature of humankind and the circumstances in which the many individuals then existed. This is the "inverse deductive" method (*Logic*, VI, x, 4).

In either case, however, it is never possible safely to assume that we have located all the causes. That means, in effect, that the laws of social phenomena that we have located are in fact gappy. Social science can therefore never be anything more than a science of *tendencies* rather than one in which positive predictions are possible. This is in fact the best that we can do, given the complexity of the phenomena; but even so, such knowledge can be useful in proposing policy (*Logic*, VI, ix, 2). After all, weather forecasting, too, is useful, even though it too is only a science of tendencies (*Logic*, III, xvii, 2).

It is evident that these methods for investigating social phenomena can work only if the deductions that Mill describes really are valid. Mill, naturally, argues that they are. In fact he holds that they occur elsewhere in science, in physics in particular (*Logic*, VI, ix, 1). Thus, for example, in mechanics it would appear to be possible to deduce the laws for a complex system from the laws for simple systems. If we have a three-body system, we can conceptually divide it into three two-body systems, and knowing the forces that would operate in the two-body systems were they isolated, we can deduce what the forces are that are operating in the three-body system. Mill holds that this deduction proceeds *a priori* (*Logic*, III, vi, xi); in these cases, as opposed to those such as chemistry and psychology where the effect is "heterogeneous" with its causes, "the joint effect of the causes is the *sum* of their separate effects" (*Logic*, III, vi, 2; italics added), and, while we know the law of the separate causes by induction, the inference to their joint effects involves no further induction but only "ratiocination" (*Logic*, III, xi, 1, 2). In fact, in

this matter Mill is simply wrong. In order for the deduction to go through one must take into account the relations by which the simpler systems are constituted into the more complex system, and there is no *a priori* reason for assuming that a given relational structure will yield one sort of law for the complex system rather than another. This means that the deduction of the law for the complex system depends not only upon the laws for the simpler systems but also upon another factual assumption that relates the laws of the complex system to *both* the laws for the simpler systems *and* the relational structure the constitutes the complex system out of the simpler systems. This factual assumption relating the laws for the complex system to both the laws for the simpler systems and the relational structure is itself a law, not a specific causal law but rather *a law about such laws*. Since it is a law the step from the causal laws for the simpler systems is not one of pure ratiocination or pure deduction but also involves an inductive feature. This law, this inductive step the existence of which Mill denies, has been referred to as a "composition law." When Mill asserts that the inference is a deduction that proceeds wholly *a priori* without any inductive step beyond those that provided the laws for the simple systems, he is neglecting to take into account this additional factual premise. In effect this amounts to neglecting the causal role of the relations which constitute the whole out of the parts. Mill, then, is wrong in his claim that in mechanics the deduction of a law for the complex system can be deduced *a priori* from the laws for simpler systems; what he calls the "deductive" method does not in fact have any place in mechanics.

Mill makes a similar mistake in the case of the social sciences. When he claims that the deduction of the laws for the complex social wholes can be deduced *a priori* from the laws for the parts, that is, from the laws for persons taken individually, he is claiming in effect that there is no need for a composition law, or, what amounts to the same, no need to take into account the *social relations* which, by virtue of holding among individuals, constitute the social whole out of those individuals. Mill suggests that in the social sciences, the individual cases act "conjunctively," in just the way that they act in mechanics (*Logic*, VI, ix, 1). But a conjunction is merely that and not a relational whole. Mill also indicates that the inference from the laws of the co-existent causes to the "aggregate" effect is something that we can "calculate *a priori*" (*ibid*.); the inference will, of course, be *a priori* if it proceeds on the basis of a *conjunction* of premises, but not if it requires additional factual premises concerning the relational structure and a composition law. He also suggests, as we have noted, that the total social effect is merely the "sum" of the individual effects. He makes the same point when he explicitly compares social phenomena to those of mechanics. For, he tells us, "in social phenomena the Composition of Causes is the universal law" (*Logic*, VI, vii, 1), where the Composition of Causes is "the principle which is exemplified in all cases in which the joint effect of several causes is identical with the *sum* of their several effects" (*Logic*, III, vi, 1; italics added). In short, when Mill proposes the "deductive" method for the social sciences he is neglecting to take into account social relations as relevant factors. It is much as if Newton failed to take into account the relative positions of the planets when he inferred the forces acting in the solar system from the assumption that gravity would act among the planets and the sun taking them pairwise; but then, Mill's account of mechanics implies that Newton did just that!

Mill was in fact sensitive in his many writings to the role of social relations. He had early in his career supported many of the ideas of Coleridge against the dogmatic social atomism earlier utilitarian radicals. From Coleridge he had learned to appreciate the role of social and cultural institutions in the historical development of human beings. Mill, like his father, was a determinist with regard to social phenomena,

but from Comte he had absorbed the idea that social change proceeds in a series of stages: there are "critical periods," in which old institutions are overthrown, and these are followed by "organic periods," in which new forms of social structure emerge and are consolidated. He felt that in his own time society was emerging from a critical period. It was from his reading of Coleridge on social institutions that Mill came to be aware of the roles that they, especially educational institutions, would play in re-establishing a new social structures, new forms of social relations.

In his work is the social sciences, then, Mill was well aware of the importance of social relations as relevant variables In this he had gone beyond the social atomism of his father. But in his proposals for a methodology of social science he quite neglected the role of those social relations. In his methodological pronouncements has had not yet freed himself from those atomistic assumptions.

It was only later that methodologists came to recognize that the way to deal with the complexity of social phenomena and therefore the inevitable gappiness of the laws that we can discover is through the use of statistical methods.

11. Political Economy

In political economy Mill built upon the foundations laid down by Ricardo, Malthus and his father. His *Principles of Political Economy and some of the applications to Social Philosophy* (1848) was the leading economics textbook for many years. Mill's reasoning generally followed that of Ricardo and Malthus, but was more realistic, allowing that beyond the motive of pecuniary gain and economic self-interest, there were other, higher motives that could play a role, and that moreover institutional forms and even sheer habit might also be relevant. These concerns, and well as his greater methodological insights, led him to challenge the claims of the classical school that wages, rent and profit are the result of immutable laws: there may be laws about wages, but there is no "iron law" of wages. These laws are, to the contrary, the result of institutional constraints, and these institutions can be changed, if the will be there. He came to regard the Malthusian principle of population not as an immutable law and a barrier to progress, but as showing the conditions under which progress can be achieved. His book is throughout governed by his belief in the possibility of great social improvements, combined with a determination to expose simplistic remedies and uncomfortable truths.

In analytical theory, Mill at first differed little from Ricardo, but in later editions of his *Principles* he came to modify those views. Thus, for example, the theory of the wages fund was modified almost to the point of rejection under the criticisms of William Thomas Thornton. Where Mill first adopted Ricardo's view that the average wage is determined by a fixed lot of capital divided by the number of workers, he came to allow that other factors play a role in determining wages, among them workers' expectations as well as various institutional factors.

Mill emphasized the distinction between production and distribution: there are laws in both cases, but these laws are different in kind. The laws of the former, he argued, "partake of the character of physical truths ... . It is not so with the Distribution of Wealth. That is a mater of human institution merely"

(*Principles*, p. 199). The way goods are distributed depends upon the rules of property, and Mill explores various sorts of property relations, from the usual form of his own country, to the ways of holding property in Ireland and India, to the various forms of socialism. The rules that obtain at any given time or place "are what the opinions of the ruling portion of the community makes them" (*ibid*., p. 200); but these in turn are "not a matter of chance" (*ibid*.). To the contrary, they have causes and these can be understood using the methods of empirical science: they are "as much a subject for scientific enquiry as any of the physical laws of nature" (*ibid*., p. 21). However, though Mill emphasizes how production and distribution differ, he holds that production too depends upon social factors. For example, security and monetary incentives are among the things that influence productivity.

From more recent perspectives in economic analysis, some of Mill's economics decidedly looks backward. Thus, Mill retains the now abandoned distinction between productive and unproductive labour. But if this has no place in pure economics, it does have a legitimate place in Mill's work. For Mill the distinction is related to his concern to eliminate vestiges of feudalism, the primitive sector of the economy in which retainers and menial servants are maintained more or less in idleness. This concern recommends the development of the more advanced industrial at the expense of the pre-industrial sector. Mill's economics should be seen as concerned as much with the economics of development as it is with pure theory. Nor is Mill's concern simply with the production of material goods: Book 4, Ch. 6 of the *Principles* ("Of the Stationary State") ends with a moving plea for the preservation of natural beauty.

On the whole Mill supported the laissez faire economic policies that had been defended by earlier economists such a his father and David Ricardo. His overall concern was here as elsewhere with self development, and laissez faire policies seemed to provide the scope needed for individual freedom. But on further reflection, moved in this by his wife, he came to the view that personal development required not just the freedom of the economic market but also political freedom, and that this is of little use to an individual who lacks economic security and opportunity. Mill was concerned, too, with motivation. He saw the system of wages that had developed in the industrial revolution as one which robbed the workers of any interest in the goods that they were producing. He came increasingly to re-examine the objections to socialism, and came to argue in later editions of the *Principles* that, as far as economic theory was concerned, there is nothing in principle in economic theory that precludes an economic order based on socialist policies. He therefore made the radical proposal that the whole wage system be abolished, and that it be replaced by a cooperative system in which the producers would act in combinations, collectively owning the capital necessary for carrying on their operations, and working under managers who would be responsible overall to them. Like Ricardo, he held that profits in the long run would tend to diminish and that the formation of new capital would thereby come to an end. This would bring industry to a halt and population to a stationary level. The result would be a relatively static form of society. In such a society, Mill hoped, people's thoughts would turn from concerns of self-interest to more socially and humanly worthy ends. In such a state many of our present problems would disappear.

Mill summed up his objective in his *Autobiography* (1873): "how to unite the greatest individual liberty of action, with a common ownership in the raw material of the globe, and an equal participation of all in the benefits of combined labour." (p. 239) In his economic theory Mill no doubt appears to the modern socialist to be a follower of Ricardo and the classical liberal economists, but to the latter, and no doubt to

himself, he was clearly a socialist.

12. Moral Philosophy: Utilitarianism

Throughout his major works and in him many essays, Mill argues that the moral worth of actions is to be judged in terms of the consequences of those actions. In this he contrasts his own view with that of those who appealed to moral intuitions. For some, these intuitions are just that, in which case they have little moral force indeed; they are simply the arbitrary feelings of approbation and disapprobation. But intuitions conflict, and we need some standard to decide which of these feelings is correct. Intuition does not supply that. There are some, however, such as William Whewell or Immanuel Kant, or, later, idealists such as T. H. Green, who claim that there are objective criteria for adjudicating conflicts. These philosophers support their intuitions by appeal to a moral order that pervades the universe, some sort of moral essence or objective demand from the noumenal or transcendental realm. However, given the basic argument that Mill offers for the relativity of all knowledge these claims do not amount to much; they are to be taken no more seriously than those who justify their moral judgments by appeal to "God said so". These opponents all appeal to no more than their private sentiment: this is what I like or this is what I dislike. That fact that it *appears* as a moral authority gives it no superior authority.

Moral intuitions are said to reveal ends which are superior to those of our worldly nature, superior to mere pleasure and self-interest. Mill of course agrees that our moral feelings often conflict with our inclinations of self-interest. But these feelings are not feelings that are contrary to our pleasure. They like all ends are sought to the extent to which they are enjoyable. It is just that different, and conflicting things, are enjoyable.

Mill can of course account for these divergent feelings and inclinations. On the psychological account of human being that he defends, pleasure and pain are the prime motivators. Other things are sought, at least initially, as means to pleasure or the avoidance of pain. But as the associative mechanisms work, things that are sought as means come to be associated with the ends for which they are means. These things come to be sought as ends in themselves, as parts of pleasure. The variety of ends that persons suggest are morally demanded by their intuitions are simply things that have come to be among those things that are for them part of pleasure, ends that are in conflict with those ends that are other parts of pleasure. The appeal to intuition does not solve the problems of moral philosophy. It is no more than a commonplace of fact, that we feel better about some ends rather than others and that we often feel that our ends are better that those that others have. The real problem is elsewhere: how to resolve the conflict.

All ends are either pleasure or parts or of pleasure. This is a matter of psychological fact. As Mill puts it, "to desire anything, except in proportion as the idea of it is pleasure, is a physical and metaphysical impossibility" (*Utilitarianism*, Ch. 4). This implies that pleasure is the end of morality:

> The sole evidence it is possible to produce that anything is desirable [= worthy of desire], is that people do actually desire it. If the end which the utilitarian doctrine proposes to itself were not, in theory and in practice, acknowledged to be an end, nothing could ever

> convince any person that it was so (*Utilitarianism*, Ch. 4).

Mill's point is often criticized as making an illegitimate inference from "is" to "ought," from "is desired" to "is worthy of desire." It does indeed make that inference, but it does so legitimately. Mill's point is that, as a matter of lawful fact about human beings, we *must* seek pleasure, it is unreasonable to suggest that anything else could be morally demanded of us. Mill is here relying on the principle that *must implies ought*, the converse of the principle that *ought implies can*. If these principles did not govern our moral attitudes, we would end up attempting the impossible, and, if the point of morality is to guide action, then that is unreasonable: any action attempting the impossible is bound to be pointless.

The maximization of pleasure or happiness is therefore the moral end. But this ought not to be taken in simplistic terms. Mill's is not a crude hedonism. In the first place, it is not crude sensual pleasure that is the aim. Rather, welfare consists in the satisfaction of desire, and the relevant pleasure is the pleasure that comes from satisfied desire. In the second place, when he insists that welfare consists in the experiencing of pleasurable states, he argues, in contrast to what Bentham implied, that quality, not simply the amount of pleasure, is to be taken into account. As Mill came to see in his own experience, reading Wordsworth is better as an experience than drinking ale. As he put it himself, better Socrates dissatisfied than a pig satisfied. Some experiences are qualitatively better than others, and in determining which line of action is better, this has to be part of the calculation. These pleasures are not merely the sum of more elementary pleasures; they are qualitatively different. These differences are a matter of the chemical nature of psychological processes. Among the qualitatively superior ends are the moral ends, and it is in this that people acquire the sense that they have moral intuitions superior to mere self-interest. And in the third place, Mill holds that it is possible to be content with life enough dissatisfied, provided that one has the proper balance of pleasure, reckoned both quantitatively and qualitatively. As he himself suggested, better Socrates dissatisfied than a pig satisfied. The pig may be satisfied, but Socrates' life, even with its dissatisfaction, is preferable. The person who has a good life has a reasonable balance of tranquility, on the one hand, and, on the other, moments o excitement and more intense pleasure.

Human beings collectively develop rules to aid them in their efforts to maximize their happiness. Each of us wants to appropriate goods to satisfy our material needs. But they are scarce, not everyone can satisfy these needs. Given this scarcity of material goods, there will be conflict. If one succeeds in appropriating goods, then others will attempt to take them away to satisfy their own needs. What one more exactly wants is not a maximum of goods but a satisfactory level of goods together with security of tenure. Since each has this an end, norms for the distribution of the scarce goods come to be established. Together with these norms of justice there will also come to be established norms for their enforcement, for the punishment of those who violate these norms. These norms with sanctions attached, that is, the norms of justice will function as means to the satisfaction of material desires, but through the associative mechanisms they will come to sought as ends, as parts of one's pleasure. Because they concern the essential of human well-being, they therefore come to felt as more morally demanding than the principle of utility itself.

The principle of utility judges these norms. Mill is therefore not an "act utilitarian" who holds that the principle of utility is used to judge the rightness or wrongness of each and every act. But neither is he a

"rule utilitarian" who holds that individual acts are judged by various moral rules which are themselves judged by the principle of utility acting as a second order principle to determine which set of rules secures the greatest amount of happiness. For the principle of utility judges not simply rules, according to Mill, but rules with sanctions attached. But Mill holds that there are some occasions on which the principle of utility must be used to judge individual acts. There are two sorts of such occasion. One is to judge when exceptions to ordinary rules are to occur or to judge which subsidiary rule applies when two come into conflict. The other is to judge actions aimed at changing the social structure of rules. It is the leaders in "the ruling portion of the community" who must think and plan in this way, those who are in positions of economic or political or moral power that enables them to sway or determine public feeling and sentiments for social change.

Different forms of such things as agricultural practice will generate different patterns of what will be accepted as norms for distribution; the legitimate ends of justice will be secured by different institutional forms. It is these forms that develop over time through periods of crisis and consolidation.

This consolidation or re-consolidation results in better social forms. Mill in fact argues that such social improvement is the overall trend of development: the direction is to maximize the general well-being. Mill argues that since each person aims to maximize his or her own happiness therefore the overall effect will be to maximize the pleasure of all. As he puts it, since "each person's happiness is a good to that person," then 'the general happiness" must be a good to the aggregate of all persons" (*Utilitarianism*, Ch. 4). It is commonly charged that Mill's inference commits the fallacy of composition -- the fallacy that since this person has a mother, that person has a mother, and that other person has a mother, therefore the aggregate of all persons has a mother. But as he elsewhere explains that "I merely meant in this particular sentence to argue that since A's happiness is a good, B's is a good, C's is a good, etc., the sum of all these goods must be a good (*Later Letters*, p. 1414).

Nor is Mill arguing that since each seeks his or her own happiness, therefore each seeks the happiness of all, though he is often accused of this fallacy. To the contrary, Mill clearly holds that it is seldom true that individuals seek the general happiness. In the best state of society this would be so. But we are clearly not in the best state. In fact, it would be contrary to the principle of utility itself to have individuals constantly seeking the general good. To seek the general good would require constant calculation of long term consequences, and that is hardly possible. If it were attempted, then mistakes would be made and time wasted. Better on utilitarian grounds to work with subsidiary and time-tested rules, with the appeal to utility itself being made only on those relatively rare occasions when subsidiary rules come into conflict or where exceptions are needed, or where whole systems of rules are called into question.

Mill's inference is nonetheless fallacious. It presupposes that the laws for the social whole, the aggregate, are simply the sum of the laws for the individual cases. This is simply an application of the inverse deductive methods that Mill advocates in the *System of Logic*. But this method is fallacious, as we have seen: it ignores the causal role of social relations. Mill in fact in other contexts recognizes this. For he holds that a greater degree of general well-being might be achieved by a different form of social organization.

There is, then, no general justification for the principle of utility. But this does not mean that after all each individual is nothing more than an egoist seeking his or her own happiness and that there is no basis in human nature for a rule capable of resolving conflicts. There is, in the first place, the forms of justice that develop; conformity to these rules which ensure that the needs of others as well as one's own are satisfied, becomes part of the pleasure of each. These systems of rules will often, as in complex modern societies, have rules for changing the rules and magistrates or others in positions of power who can determine how society will change and develop.

But further, in the second place, there is the natural sympathy of human kind, each for the other, or at least the others that are close to us. On this tendency, each is inclined to feel as others feel, so that the ends of others become naturally our own ends. This yields common rather than conflicting ends. In this way the good of all becomes part of the good of each. Each of us thus comes to move in unity with our fellows for the good of each and all.

Our natural sympathy thus works to establish a set of social relations that unite individuals into a community. In his theory Mill may ignore social relations, implicitly denying their existence, but when he comes to consider how society actually works he clearly allows for their existence. In these respects he both looks back to the social atomism of an earlier generation of utilitarian radicals but also, when he admits those relations, makes an advance on their thought.

Sympathy is thus important in insuring that each and all work for a common good. But this common good may not in fact be the best that can be achieved. Sympathy can often be constrained by forms of social order. Greater well being can be achieved only by achieving new forms of social organization. As Mill sees it, the opportunity for such improvement comes in the critical periods of social development. It is in such contexts that moral leaders such as Socrates and Jesus can and have played a crucial role. They can captivate the overall general sympathy present in society to bring about better social structures to be consolidated into improved organic periods.

Utilitarianism is not a simplistic moral principle to be mechanically applied, it is a long term social project.

13. Social and Political Philosophy

For Mill government is not a matter of natural rights or social contract, as in many forms of liberalism. Forms of government are, rather, to be judged according to "utility in the largest sense, grounded on the permanent interest of man as a progressive being" (*On Liberty*, p. 224). By this he means that forms of government are to be evaluated in terms of their capacity to enable each person to exercise and develop in his or her own way their capacities for higher forms of human happiness. Such development will be an end for each individual, but also a means for society as whole to develop and to make life better for all.

Given the centrality of self development, Mill argues that liberty is the fundamental human right. "The

sole end," he proposes, " for which mankind are warranted, individually or collectively... in interfering with the liberty of action of any of their number, is self-protection" (*On Liberty* p. 223). This will enable each to seek his or her own best; it will liberate a diversity of interests to the benefit of the individual and of all; and it will nurture moral freedom and rationality. With the latter will come creativity and the means of social and intellectual progress. Mill's *On Liberty* remains the strongest and most eloquent defense of liberalism that we have. He argues in particular for freedom of thought and discussion. "We can never be sure," he wrote, "that the opinion we are endeavoring to stifle is a false opinion, and if we were sure, stifling it would be an evil still" (*On Liberty*, p. 229).

Bentham and Mill's father had argued that democracy was the form of government that could best secure the happiness of all. The younger Mill was inclined to agree. But the end is not just well-being, as earlier utilitarians argued, though it is that. The end that recommends it is the tendency to foster self-development and individuality. Representative government, is particular, he defended as that form which best encourages individuality. It leads people to take a more active and intelligent participation in society. It provides moral training and encourages the development of natural human sympathies. The result is the habit of looking at social questions from an impersonal perspective rather than that of self-interest. But Mill's defense of democracy was much qualified. To be sure, he was, like the earlier utilitarians, sympathetic to the fall of the *ancien régime* and to the ends of the French Revolution. He had little use for the British aristocracy and criticizes it for its follies in its own country and in Ireland, and the vestiges, such as the Game Laws, of medieval privilege. He strove to liberalize the press still severely bound by an absurd libel law that excluded effective social criticism. But influenced by Coleridge he had come to see that there were virtues in social systems, even out-dated ones, else why would not have survived so long. He therefore came to appreciate the conservative arguments that unrestrained freedom is dangerous. The effort to achieve at year zero a new social order justified on *a priori* principles by means of Jacobin terror can be as great a threat to liberty and to human well-being as the most repressive tyranny.

Mill argued, reasonably on utilitarian grounds, that social institutions need to be adapted to the time and place where they operate. His work in the East India Company dealing with the governance of states in India undoubtedly had a significance influence here. Referring to the rule of Akbar in India, he allowed that despotic rule could be necessary under certain conditions for stable government. He even suggests that, since people must be properly fit if democracy is to function well, a despotic form of government, if well-run with this aim in mind, might prepare its people for the exercise of responsibilities of a free electorate. His position here had some influence on British colonial administrations.

Mill, with de Tocqueville, stressed the importance of local government. He was highly critical of the chaotic forms of local administration then present in Britain, and his influence was effective after 1871 when the central government moved to bring about reforms.

In his thinking about how best to administer a state as a whole, Mill argued that the best administration was one that relied upon professional skills. He was prepared to accept the British form of parliamentary government where the executive is responsible to an elected assembly. Naturally enough, however, he was highly critical of the unelected British House of Lords, which he saw as another vestige of a more

primitive feudal society. The best form of government could be determined by the test of experience and that experience found the Lords wanting.

Individuality and even eccentricity is better than massive social uniformity. The latter is the consequence of both terror and tyranny. But it can also be the consequence of democracy. Influenced by de Tocqueville's analysis of American culture, Mill came to think that the chief danger of democracy is that of suppressing individual differences, and of allowing no genuine development of minority opinion and of minority forms of culture. Democracy might will impoverish the culture of the community by imposing a single and inflexible set of mass values. This form of government has the virtue of fostering intelligence, common moral standards, and happiness; but where the citizens are unfit and passive it can be an instrument for tyranny, perhaps of one, as with Louis Napoleon, or perhaps of the many. In general, the only reliable safeguard can be institutions, educational institutions in particular, that can ensure the development of individuals with personalities strong enough to resist such pressures. But other forms of social order are also called for. Thus, after the rebellion in 1837 in Canada he defended Lord Durham's recommendations for internal responsible self-government in the colonies, free on the whole from interference from the colonial power, with a form of federal organization to defend the cultural interests of the French minority. Another means suggested by Mill for the protection of minorities in a democratic system was a system of proportional representation. Finally, one might also mention his acceptance of the principle of multiple votes, in which educated and more responsible persons would be made more influential by giving them more votes than the uneducated.

Mill is concerned to provide a form of government with as much education as feasible. A properly educated electorate would be willing and able to select the best as their governors. Since those elected would be better informed and wiser on specific issues that those who elected them, it would be wrong to bind the representatives to anything but a very general agreement with the beliefs and the aims of the electors. He agrees with the rejection of populism enunciated by Burke in his speech to his electors in Bristol, accepting the principle that the representative should be expected to exercise his or her own judgement, not merely to accept blindly the views of those on whose votes his or her tenure depended. It was a principle to which Mill himself adhered in his own brief term in the House of Commons.

14. Status of Women

Among the things for which Mill campaigned most strongly were women's rights, women's suffrage, and equal access to education for women. His essay on the *Subjection of Women* (1869) is an enduring defense of gender equality. His strong views in this area led to a deep serious disagreement with his father. To be sure, as one could be expect, he was not able to free himself completely from the prejudices of his age: he argued that it was undesirable that women seek employment outside the home in order to support the family. While Mill argues that not all motives are egoistic and self-interested, he nonetheless held that in most affairs of economics and government such motives are dominant. The elder Mill argued that votes for women were unnecessary, since the male could adequately represent the interests of the family and those who were parts of it. But the younger Mill points out that the interests of the male could diverge from those of the female in the family. Here he recognizes as elsewhere he does not, that the

actions of individuals in an aggregate often do not result in maximizing the general welfare of the parts. The essay can be seen in large part as a long argument on the abuse of power. A well-intentional power *might* secure the interests of the governed, but the power of egoism renders this unlikely. Since male self-interest can conflict with the self-interest of the female, the votes of women are needed to curb the pursuit of male self-interest. Here, as elsewhere, as he says in a letter to Florence Nightengale, "political power is the only security against every form of political oppression." (*Later Letters*, p. 1343-44) In fact, hitherto the interests of the family have been subservient to the interests of the dominant male partner. If the interests of the family, of the aggregate, are to be served, then gender equality is required. Changing the social relations between men and women to ones in which they play equal roles will require each to curb their self-interests and to broaden their social sympathies to include those of the other and of the whole. Mill felt this is his own life: through his relationship with Harriet Taylor, he came to the strong conviction that women's suffrage is an essential step toward the moral improvement of humankind.

15. Views on Religion

Mill was generally taken to be an atheist or an agnostic, though during his life he published little on the topic of religion, and as he made clear in his correspondence with Comte his fear of alienating his readers and losing his public influence led him to be determinedly cautious -- indeed cautious to the extent that he was criticized for this by those who otherwise sympathized with him. The latter were rather consternated, then, with the posthumous publication of Mill's *Three Essays on Religion*: in spite of the strictures that appeared in *Examination of Sir William Hamilton's Philosophy*, it turned out that Mill was rather more sympathetic to religion than were they.

In "On Nature" Mill argues that the maxim "Follow Nature" proposed equally by the ancient Stoics and the modern Romantics is a poor guide to action, certainly one contrary to the principle of utility. 'Nature' might have two meanings. On the first, 'nature' means 'whatever happens', and it recommends as right whatever happens, be it good or bad. In this case, it offers no moral guidance whatsoever. On the second meaning, 'nature' means whatever happens without human interference' -- natural as opposed to artificial in the sense of being the result of human art. In this case it is contradictory since it itself is a matter of human art. Mill argues that nature in the second sense offers us a view of as much evil as good, and so proposes more a challenge to change than an ideal for imitation. The task is not to follow nature but to improve it, especially human nature: virtue is not the consequence of nature but of nurture, of cultivation.

As for nature itself, the only rational conclusion that one can draw from contemplating the amount of ugliness and unavoidable evil that it contains is that whatever principle of good is at work in the universe, if any, cannot subdue the powers of evil: it cannot be omnipotent.

In the essay on "The Utility of Religion" Mill argues that much of the apparent social utility of religion derives not from its dogma and theology but to its inculcation of a widely accepted moral code, and to the force of public opinion guided by that code. The belief in a supernatural power may have had some utility in maintaining that code, but is no longer needed and may indeed be detrimental.

There is an unfortunate tendency in supernatural religion to hinder the development not only of our intellectual, but also our moral nature. Its appeal is to self-interest rather than to disinterested and ideal motives. As with intuitionism in ethics, it stands in the way of the critical evaluation of social norms, and thereby effectively prevents action aimed at social change for the improvement of the human lot in the community. Supernatural religion appeals to the sense of mystery about what lies outside the narrow realm of what we know. But the appeal can be made by poetry: the realm of the unknown can filled only by the imagination. "Religion and poetry address themselves, at least in one of their aspects, to the same part of the human constitution; they both supply the same want, that of ideal conceptions grander and more beautiful than we see realized in the prose of human life" ("Utility of Religion," *Three Essays*, p. 419).

The power of religion to motivate derives, Mill suggests, from the human need for some sort of ideal conception of being to move us to do our best, a standard beyond our common selfish objects of desire. But such purposes can be achieved, and better achieved, by a religion of humanity than can any supernatural religion. It would help us cultivate our feelings and develop our individual capacities, intellectual, moral and emotional, without burdening us with false views about a mysterious Unknowable. The contrast would be to a God of the sort Mansel proposed, one Just beyond all human justice, a principle of Goodness in the world whose existence requires us to deny that the palpable evil that we find really is evil. The religion of humanity would draw our attention to real evil in the world, and urge us to work to overcome it.

These first two essays had been written by 1858; the third, "Theism," was drafted more than a decade later. The first two suggest that the alternative to supernatural religion is not the acceptance of nature and the way things are but the construction of a positive religion of humanity. The third essay makes greater concessions to traditional religion. In this essay Mill evaluates the traditional arguments for the existence of God. He rejects straight out any argument based on an *a priori* causal principle. But he suggests that the order to be found in the universe, in particular the adjustments of organisms for the ends of survival and reproduction, provides grounds for tentatively accepting the existence of a creator. Even here, however, he allows only that it can be established with "no more than probability" that the cause of such order is the activity of some intelligent designer. He allows, too, that one might, contrary to Mansel, characterize the creator in a humanly relevant way as benevolent, though it could be neither omniscient nor omnipotent. For Mill the point about a world created by such a God is that it leaves room for the work of human beings in improving both that world and themselves as persons in it. "If man had not the power," he indicates, "by the exercise of his own energies for the improvement both of himself and of his outward circumstances, to do for himself and other creatures vastly more than god had in the first instance done, the Being who called them into existence would deserve something very different from thanks at his hands." ("Theism," *Three Essays*, p. 458)

Mill argues in the same essay that there is no evidence for the immortality of the soul, but equally none against it. For Mill, this means that there is room for *hope*. Some persons at least do hope, if not for eternal life, then for a life that extends beyond their death. It is possible, he suggests, that the benevolent and powerful (though not all-powerful) creator could grant that wish. Such at least one might hope.

Defenders of religion had long appealed to miracles as support for their beliefs about the supernatural. In his essay Mill is highly critical of such appeals; there is absolutely no evidence that supports such claims. He allows only that a benevolent deity might have indicated an intention to award to those who aspire to it a life after death; if there is no rational evidence in support of that, then one might at least so hope. To this extent he allows that Jesus was indeed miraculously Christ, and that He bore such a message of "glad tidings" for the hopeful.

In spite of Mill's argument that the proper rational attitude towards supernatural religion is neither belief nor disbelief, he now concludes, in his last essay, in a way that many found rather surprising, that "the whole domain of the supernatural is thus removed from the region of Belief into that of simple Hope." ("Theism," *Three Essays*, p. 483) Indeed, such hope might be reasonably be encouraged, since its indulgence might encourage in some persons both the feeling that life is important and their sympathy for others. Further, to construct for oneself or for one's community, an image of a person of high moral excellence, such as Jesus, and from the habit of seeking approval of this person for one's own acts, may aid that "real, though purely human, religion, which sometimes calls itself the Religion of Humanity, and sometimes that of Duty." ("Theism," p. 488) This develops further his concept of the moral significance of cultivating the emotions and reflects the lesson he had learned early in his life, as he recovered from his bout of depression, that human beings can flourish only with the cultivation of the feelings.

In his considerations about the existence of a cause for order in the universe, Mill mentions only in passing the work of Charles Darwin, that natural selection is the cause of apparent design in the natural world. As soon became apparent, this theory removed whatever tentative support Mill had allowed for the existence of a benevolent creator. Hope alone remained the only legitimate religious sentiment, but that hope rested on the sense that there is a creator who might fulfill it. Upon the demise with Darwin's work of any expectation for the existence of such a creator all the slim hopes of religion disappeared. The later Victorians could not share Mill's optimism. They found that all that remained was to shake a fist in rage at the heavens that disappointed and starred back silently.

16. Conclusion

Mill's thought was of a whole. He was consistently empiricist in his metaphysics epistemology, and he developed his moral thinking in this framework. At the same time that moral philosophy shaped his metaphysics. He aimed to show humans the way the world is and how they could accommodate themselves to it and to one another. His aim was the improvement of humankind. His guide was the principle of utility.

This is often missed. The principle of utility has become the object of scholastic discussion. People debate whether Mill's notion that there are some pleasures that are to be preferred to others makes good sense. The human aspect of this ignored. Mill put it well, and makes it clear that his claims on this point are solidly based.

> Few human creatures would consent to be changed into any other lower animals, for a

> promise of the fullest allowance of a beast's pleasures; no intelligent human being would consent to be a fool, not instructed person would be an ignoramus, no person of feeling and conscience would be selfish and base, even though they should be persuaded that the fool, the dunce, or the rascal is better satisfied with his lot than they are with theirs. (*Utilitarianism*, p. 211)

People also critically reject Mill's case for the principle, rejecting it on the grounds that this or that contrived counter-example shows some imperfection in Mill's formulation. But this is to miss the point of Mill's work. There may be imperfections in what was said, but the aim of the whole is clear and criticism should always try to be constructive, not merely negative. Mill himself made clear that nature of the moral imperative that he proposed.

> In the golden rule of Jesus of Nazareth, we read the complete spirit of the ethics of utility. To do as one would be done by, and to love one's neighbor as oneself, constitute the ideal perfection of utilitarian morality. As the means of making the nearest approach to this ideal, utility would enjoin, first, that laws and social arrangements should place the happiness, or (as speaking practically it may be called) the interest, of every individual as nearly as possible in harmony with the interest of the whole; and secondly, that education and opinion, which have so fast a power over human character, should so use that power as to establish in the mind of every individual an indissoluble association between his own happiness and the good of the whole ... (*Utilitarianism*, p. 218)

We, and the world, would do well to follow Mill in these principles. All would be the better for it.

Mill aimed at the improvement of humankind. For this end, he was active in many causes. He denounced the take over by the British government of the East India Company, correctly anticipating the evils consequent upon the scramble for spoils by second rate English officials. He supported reform of the Irish land tenure system in order to relieve the poverty of the peasants. During his period in Parliament, he denounced English methods in Ireland, a move which was unfortunately denounced as support for the Fenians. In 1866 and 1867, he was active along with Herbert Spencer and many other liberals in the committee for the prosecution of Governor Eyre for atrocities in suppressing a rebellion by blacks in Jamaica. He supported attempts to preserve natural beauty and was a founder and strong supporter of the Commons Preservation Society.

In his own day Mill was immensely influential. He was never one to compromise his principles, and his pursuit of those ideals was steady and often successful.

Mill's metaphysics is perhaps less influential now than it was in his own day. Certainly, for many decades it stood in the shadow of idealism. His revival of formal logic inspired the developments that now date it. In the philosophy of science, his empiricism has for the most part stood the test of time. But his lasting influence has been in the areas of political and social philosophy. His defenses of utilitarianism and of liberty shaped the views of his own generation, and they continue to this to inspire

and to guide. In thought especially but also in action he made of the world a better place.

Bibliography

Mill's Works

The standard edition of Mill's writings, in thirty-three volumes, is the following:

CW Mill, J. S., *Collected Works of John Stuart Mill*, J. M. Robson (ed.), Toronto: University of Toronto Press, 1963ff.

The introductions by various authors are in each case worth reading. Volume and page numbers given below for Mill's works are to this edition. Page references in the preceding article refer to this edition.

(1838) "Bentham," CW, v. 10, pp. 75-115

(1840) "Coleridge," CW, v. 10. pp. 117-63

(1843) System of Logic, Ratiocinative and Inductive, CW, v. 7-8

(1848) Principles of Political Economy, CW, v. 2-3

(1859) On Liberty, CW, v. 18, pp. 213-310

(1861a) Utilitarianism, CW, v. 10, pp. 203-59

(1861b) Considerations on Representative Government, CW, v. 29, pp. 371-577

(1865a) An Examination of Sir William Hamilton's Philosophy, CW, 4. 9

(1865b) Auguste Comte and Positivism, CW, v. 10, pp. 261-368

(1869a) The Subjection of Women, CW, v. 21, pp. 259-340

(1869b) Notes to James Mill, *Analysis of the Phaenomena of the Human Mind*, 2nd edition, J. S. Mill (ed.); (1st edition, 1829; 2nd edition, London: Longman, Green, Reader and Dyer, 1869; reprinted New York: A. Keley, 1967)

(1873) Autobiography, CW, v. 1, pp. 1-290

(1874) Three Essays on Religion, CW, v. 10, pp. 369-489

Chapters on Socialism, CW, v. 5, pp. 703-53

(1963) Earlier Letters, CW, v. 12-13 [Mill's correspondence through 1848]

(1972) Later Letters, CW, v 14-16 [Mill's correspondence after 1848]

Secondary Literature

- Annas, J.. "Mill and the Subjection of Women," Philosophy, 52 (1977), pp. 179-94
- Anschutz, R. P., Philosophy of John Stuart Mill (Oxford: Oxford University Press, 1953)
- Bain, A., John Stuart Mill: A Criticism (London: Longmans, 1882)
- Berger, F. R., Happiness, Justice and Freedom: The Moral and Political Philosophy of John Stuart Mill (Berkeley: University of California Press, 1984)

- Britton, K., John Stuart Mill (Harmondsworth, Middlesex: Penguin, 1953)
- Bosanquet, B., Philosophical Theory of the State (London: Macmillan, 1889)
- Carr, R. "The Religious Thought of John Stuart Mill: A Study in Religious Scepticism," Journal of the History of Ideas, 23 (1962), pp. 475-95
- Collini, S., Public Moralists: Political Thought and Intellectual Life in Great Britain 1850-1930 (Oxford: Oxford University Press, 1991)
- Crisp, R., Mill on Utilitarianism (London: Routledge, 1997
- Courtney, W. L., The Metaphysics of John Stuart Mill (London: Kegan Paul, 1879)
- Cowling, M., Mill and Liberalism (Cambridge: Cambridge University Press, 1963)
- Donner, W., "John Stuart Mill's Liberal Feminism," Philosophical Studies, 69 (1993), pp. 155-66
- Donner, W., The Liberal Self: John Stuart Mill['s Moral and Political Philosophy (Ithaca: Cornell University Press, 1991)
- Douglas, C. M., John Stuart Mill: A Study of His Philosophy (Edinburgh: Blackwood, 1895)
- Duncan, G., Marx and Mill: Two Views of Social Conflict and Social Harmony (Cambridge: Cambridge University Press, 1973)
- Gray, J., Mill on Liberty: A Defense, second edition (London: Routledge, 1996)
- Griffen, J., Well-Being: Its Meaning, Measurement and Moral Importance, (Oxford: Oxford University Press, 1986)
- Halevy, E., The Growth of Philosophic Radicalism, trans. M. Morris, with Preface by A. D. Lindsay (London: Faber and Faber, 1934)
- Hall, E. W., "The 'Proof' of Utility in Bentham and Mill,", Ethics, 9 (1949), pp. 1-18
- Hollander, S., The Economics of John Stuart Mill, 2 vols (Toronto: University of Toronto Press, 1985)
- Jackson, R., The Deductive Logic of John Stuart Mill (Oxford: Oxford University Press, 1941)
- Kahan, A. S. Aristocratic Liberalism: The Social and Political Thought of Jacob Burkhardt, John Stuart Mill, and Alexis de Toqueville (Oxford: Oxford University Press, 1992)
- Kitcher, P., The Nature of Mathematical Knowledge (Oxford: Oxford University Press, 1983)
- Kubitz, O. A., The Development of John Stuart Mill's System of Logic (Urbana, Ill.: University of Illinois Press, 1932)
- Laine, M., ed., A Cultivated Mind: Essays on John Stuart Mill Presented to John Robson (Toronto: University of Toronto Press, 1991)
- Lipkes, J., Politics, Religion, and Classical Political Economy in Britain: John Stuart Mill and His Followers (New York: St., Martin's Press, 1999)
- Lyons, D., Rights, Welfare and Mill's Moral Theory (Oxford: Oxford University Press, 1994)
- Macke, J. L., The Cement of the Universe (Oxford: Oxford University Press, 1974)
- Markus, Ingrid, Women, Politics and Reproduction: The Liberal Legacy (Toronto: University of Toronto Press, 1996)
- Oakley, A., Classical Economic Man: Human Agency and Methodology in the Political Economy of Adam Smith and J. S. Mill (Aldershot, Hants., England: E. Elgar, 1994)
- Okin, S., Women in Western Political Philosophy (Princeton: Princeton University Press, 1979)
- Packe, M. St. John, The Life of John Stuart Mill (London: Secker and Warburg, 1954)
- Pappe, H. O., John Stuart Mill and Harriet Taylor (Melbourne: University of Melbourne Press, 1960)

- Plamenatz, J., The English Utilitarians, (Oxford: Oxford University Press, 1949)
- Popper, K., The Open Society and Its Enemies, 2 vols. (Princeton: Princeton University Press, 1950)
- Pyle, A., ed., Liberty: Contemporary Responses to John Stuart Mill (Bristol: Thoemmes Press, 1994)
- Pyle, O., ed., The Subjection of Women: Contemporary Responses to John Stuart Mill (Bristol: Thoemmes Press, 1995)
- Raz, J. The Morality of Freedom (Oxford: University of Oxford Press, 1983)
- Rees, J. C., Mill and His Early Critics (Leicester: University of Leicester Press,1956)
- Ring, J., Political Theory and Contemporary Feminism (Albany: State University of New York Press, 1985)
- Ritchie, D. G., Principles of State Interference (London: Swann Sonneschein & Co,: 1902)
- Robson, J. M., The Improvement of Mankind (Toronto: University of Toronto Press, 1968)
- Robson, J. M. and Laine, M., eds., James and John Stuart Mill: Papers of the Centenary Conference (Toronto: University of Toronto Press, 1976)
- Ryan, A., J. S. Mill (London: Routledge and Kegan Paul, 1974)
- Scare, G., Logic and Reality in the Philosophy of John Stuart Mill (Dordrecht, The Netherlands: Kluwer, 1989)
- Schneewind, J. B., ed., Mill: A Collection of Critical Essays (Garden City, NY: Doubleday, 1968)
- Skorupski, J., John Stuart Mill (London: Routledge, 1989)
- Skorupski, J., ed., The Cambridge Companion to John Stuart Mill (Cambridge: Cambridge University Press, 1998)
- Stephen, J. F., Liberty, Freedom and Fraternity, second edition (London: Smith Elder, 1874)
- Stephen, Leslie, English Utilitarians, 3. vols. (London: Duckworth, 1900)
- Ten, C. L., Mill on Liberty (Oxford: Oxford University press, 1980)
- Thompson, D., John Stuart Mill and Representative Government (Princeton: Princeton University Press, 1976)
- Urmson, J. O., "The Interpretation of the Moral Philosophy of John Stuart Mill," Philosophical Quarterly, 3 (1963), pp. 33-39
- Whewell, W., History of the Inductive Sciences, 3 vols. (London: J. Parker, 1837)
- Wilson, F., "Mill’s Proof that Happiness Is the Criterion of Morality," Journal of Business Ethics, 1 (1982), pp. 59-72.
- Wilson, F., "Mill’s ’Proof’ of Utility and the Composition of Causes," Journal of Business Ethics, 3 (1983), pp. 135-58
- Wilson, F., Psychological Analysis and the Philosophy of John Stuart Mill (Toronto: University of Toronto Press, 1990)
- Winch, P., The Idea of a Social Science (London: Routledge, 1958)
- Woods, T., Poetry and Philosophy: A Study in the Thought of John Stuart Mill (London: Hutchinson: 1961)
- Zastopil, L., John Stuart Mill and India (Stanford: Stanford University Press, 1994)

Other Internet Resources

- John Stuart Mill (Internet Encyclopedia of Philosophy, James Fieser (ed.), U. Tennessee/Martin)
- John Stuart Mill (The Victorian Web, University Scholars Programme Project, National University of Singapore)
- Search Mill's Works (Great Books Concordances, William A. Williams, Jr.)
- John Stuart Mill (Utilitarianism Resources)

Related Entries

Bentham, Jeremy | causation: in science | causation: the metaphysics of | consequentialism | idealism: British | induction: problem of | introspection | liberalism | logic: history of | mathematics, philosophy of | Mill, James | moral motivation | other minds | perception | rationalism vs. empiricism | representation, political | sense-data | Whewell, William

Fred Wilson
University of Toronto
fwilson@chass.utoronto.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 3, 2002
Content last modified: January 3, 2002

William Whewell

William Whewell (1794-1866) was one of the most important and influential figures in nineteenth-century Britain. Whewell, a polymath, wrote extensively on numerous subjects, including mechanics, mineralogy, geology, astronomy, political economy, theology, and architecture, as well as the works that remain the most well-known today in philosophy of science, history of science, and moral philosophy. He was one of the founding members and an early president of the British Association for the Advancement of Science, a fellow of the Royal Society, president of the Geological Society, and longtime Master of Trinity College, Cambridge. In his own time his influence was acknowledged by the major scientists of the day, such as John Herschel, Charles Darwin, Charles Lyell and Michael Faraday, who frequently turned to Whewell for philosophical and scientific advice, and, interestingly, for terminological assistance. Whewell invented the terms "anode," "cathode," and "ion" for Faraday. Upon the request of the poet Coleridge in 1833 Whewell invented the English word "scientist;" before this time the only terms in use were "natural philosopher" and "man of science."

Whewell is most known today for his massive works on the History and Philosophy of Science. His philosophy of science was attacked by John Stuart Mill in his *System of Logic*, causing an interesting and fruitful debate between them over the nature of inductive reasoning in science. It is in the context of this debate that Whewell's philosophy was rediscovered in the 20th century by critics of logical positivism. In this entry I will focus on the most important aspects of Whewell's works: his philosophy of science, including his views of induction, confirmation, and necessary truth; his view of the relation between scientific practice, history of science, and philosophy of science; and his moral philosophy. I will spend the most time on his view of induction, as this is the most interesting part of his philosophy as well as the most misinterpreted.

1. Biography

Whewell was born in 1794, the eldest child of a master-carpenter in Lancaster. The headmaster of his local grammar school, a parish priest, recognized Whewell's intellectual abilities and persuaded his father to allow him to attend the Haversham Grammar School in Westmoreland, some twelve miles to the north, where he would be able to qualify for a closed exhibition to Trinity College, Cambridge. In the 19th century and earlier, these "closed exhibitions" or scholarships were set aside for the children of working class parents, to allow for some social mobility. Whewell studied at Haversham Grammar for two years, and received private coaching in mathematics. Although he did win the exhibition it did not provide full resources for a boy of his family's means to attend Cambridge; so money had to be raised in a public subscription to supplement the scholarship money.

He thus came up to Trinity in 1812 as a "sub-sizar" (scholarship student). In 1814 he won the Chancellor's prize for his epic poem "Boadicea," thus following in the footsteps of his mother, who had published poems in the local papers. Yet he did not neglect the mathematical side of his training; in 1816 he proved his mathematical prowess by placing as both second Wrangler and second Smith's Prize man. The following year he won a college fellowship. He was elected to the Royal Society in 1820, and ordained a priest (as required for Trinity Fellows) in 1825. He took up the Chair in Mineralogy in 1828, and resigned it in 1832. In 1838 Whewell became Professor of Moral Philosophy. Almost immediately after his marriage to Cordelia Marshall on 12 October 1841, he was named Master of Trinity College upon the recommendation of the Prime Minister Robert Peel. He was Vice-Chancellor of the University in 1842 and again in 1855. In 1848 he played a large role in establishing the Natural and Moral Sciences Triposes at the University. His first wife died in 1855, and he remarried Lady Affleck, the sister of his friend Robert Ellis; Lady Affleck died in 1865. Whewell left no descendents when he died, after being thrown from his horse, on 6 March 1866.

2. Philosophy of Science: Induction

According to Whewell, all knowledge has both an ideal, or subjective dimension, as well as an objective dimension. He calls this the "fundamental antithesis" of knowledge. Whewell explains that "in every act of knowledge ... there are two opposite elements, which we may call Ideas and Perceptions" (1860a, p. 307). He criticizes Kant and the German Idealists for their exclusive focus on the ideal or subjective element, and Locke and the "Sensationalist School" for their exclusive focus on the empirical, objective element. Gaining knowledge, Whewell claims, requires attention to both elements, to ideas as well as sensations. These ideas, which he calls "Fundamental Ideas," are "supplied by the mind itself"--they are not (as Mill protests) merely received from our observations of the world. Whewell explains that the fundamental ideas are "not a consequence of experience, but a result of the particular constitution and activity of the mind, which is independent of all experience in its origin, though constantly combined with experience in its exercise" (1858a, I, p. 91). The mind is an active participant in our attempts to gain knowledge of the world, not merely a passive recipient of sense data. Ideas such as Space, Time, Cause,

and Resemblance provide a structure or form for the multitude of sensations we experience. The Ideas provide a structure by expressing the general relations that exist between our sensations (1847, I, p. 25). Thus, the Idea of Space allows us to apprehend objects as having form, magnitude, and position. Whewell holds then that observation is "idea-laden;" all observation, he says, involves "unconscious inference" using the fundamental ideas (see 1858a, I, p. 46). Each science has a particular fundamental idea which is needed to organize the facts with which that science is concerned; thus, Space is the fundamental idea of geometry, Cause the fundamental idea of mechanics, and Substance the fundamental idea of chemistry. Moreover, Whewell explains that each fundamental idea has certain "conceptions" included within it; these conceptions are "special modifications" of the idea applied to particular types of circumstances (1858b, p. 187). For example, the conception of force is a modification of the idea of Cause, applied to the particular case of motion (see 1858a, I, pp. 184-5 and p. 236).

Thus far, this discussion of the fundamental ideas may suggest that they are similar to Kant's forms of intuition, and indeed there are some similarities. Because of this, some commentators argue that Whewell's epistemology is a type of Kantianism (see, e.g., Butts, 1973, and Buchdahl, 1991). However, this interpretation ignores several crucial differences between the two views. The ideas of Space and Time do function in Whewell's epistemology as "conditions of experience" (1858a, I, p. 268), similar to Kant's forms of intuition. However, Whewell does not follow Kant in drawing a distinction between the forms of intuition, such as Space and Time, and the categories, or forms of thought, in which Kant includes Cause and Substance. Moreover, Whewell includes as fundamental ideas many which function not as conditions of experience but as conditions for having knowledge within their respective sciences: although it is certainly possible to have knowledge about the world without having a clear idea of Chemical Affinity, we could not have any knowledge of certain chemical processes within it. Unlike Kant, Whewell does not give an exhaustive list of fundamental ideas; indeed, he believes that others will emerge in the course of the development of new sciences. Further, the type of necessity which Whewell claims is derived from the ideas is very different from Kant's notion of the synthetic a priori. Finally, and perhaps most importantly, Whewell rejects Kant's subjectivism regarding the concepts of the understanding. The fundamental ideas, on Whewell's view, reflect objective features of the world, independent of the processes of the mind (I will return to these last two points in the section on Necessary Truth below).

I turn now to a discussion of the theory of induction Whewell develops with his antithetical epistemology. Whewell's first explicit, lengthy discussion of induction is found in his *Philosophy of the Inductive Sciences, founded upon their History*, which was originally published in 1840 (a second, enlarged edition appeared in 1847, and the third edition appeared as three separate works published between 1858 and 1860). Whewell considers himself to be a follower of Francis Bacon, claiming to be "renovating" Bacon's inductive method; thus one volume of the third edition of the *Philosophy* is entitled *Novum Organon Renovatum*. He calls his induction "Discoverers' Induction" and claims that it is used to discover both phenomenal and causal laws. Whewell follows Bacon in rejecting the standard, overly-narrow notion of induction that holds induction to be merely simple enumeration of instances. Rather, Whewell explains that, in induction, "there is a New Element added to the combination [of instances] by the very act of thought by which they were combined" (1847, II, p. 48). This "act of thought" is a process Whewell calls "colligation." Colligation, Whewell explains, is the mental operation of bringing together a

number of empirical facts by "superinducing" upon them a conception which unites the facts and renders them capable of being expressed by a general law. The conception thus provides the "true bond of Unity by which the phenomena are held together" (1847, II, p. 46), by providing a property shared by the known members of a class (in the case of causal laws, the colligating property is that of sharing the same cause).

Thus the known points of the Martian orbit were colligated by Kepler using the conception of an elliptical curve. Often new discoveries are made, Whewell points out, not when new facts are discovered but when the appropriate conception is applied to the facts. In the case of Kepler's discovery, the observed points of the orbit were known to Tycho Brahe, but only when Kepler applied the ellipse conception was the true path of the orbit discovered. Kepler was the first one to apply this conception to an orbital path in part because he had, in his mind, a very clear notion of the conception of an ellipse. This is important because the fundamental ideas and conceptions are provided by our minds, but they cannot be used in their innate form. Whewell explains that "the Ideas, the germs of them at least, were in the human mind before [experience]; but by the progress of scientific thought they are unfolded into clearness and distinctness" (1860a, p. 373). Whewell refers to this "unfolding" of ideas and conceptions as the "explication of conceptions." Explication is a necessary precondition to discovery, and it consists in a partly empirical, partly rational process. Scientists first try to clarify and make explicit a conception in their minds, then attempt to apply it to the facts they have precisely examined, to determine whether the conception can colligate the facts into a law. If not, the scientist uses this experience to attempt a further refinement of the conception. Whewell claims that a large part of the history of science is the "history of scientific ideas," that is, the history of their explication and subsequent use as colligating concepts. Thus in the case of Kepler's use of the ellipse conception, Whewell notes that "to supply this conception, required a special preparation, and a special activity in the mind of the discoverer. ... To discover such a connection, the mind must be conversant with certain relations of space, and with certain kinds of figures" (1849, pp. 28-9).

Once conceptions have been explicated, it is possible to choose the appropriate conception with which to colligate phenomena. But how is the appropriate conception chosen? According to Whewell , it is not a matter of guesswork. Nor, importantly, is it merely a matter of observation. Whewell explains that "there is a special process in the mind, in addition to the mere observation of facts, which is necessary" (1849, p. 40). This "special process in the mind" is a process of inference. "We infer more than we see," Whewell claims (1858a, I, p. 46). Typically, finding the appropriate conception with which to colligate a class of phenomena requires a series of inferences, thus Whewell claims that discoverers's induction is a process involving a "train of researches" (1857/1873, I, p. 297). He allows any type of inference in the colligation, including enumerative, eliminative and analogical. Thus Kepler in his *Astronomia Nova* (1609) can be seen as using various forms of inference to reach the ellipse conception (see Snyder, 1997a). When Augustus DeMorgan complains, in his 1847 logic text, about certain writers using the term "induction" as including "the use of the whole box of [logical] tools," he is undoubtedly referring to his teacher and friend Whewell.

After the known members of a class are colligated with the use of a conception, the second step of Whewell's discoverers' induction occurs: namely, the generalization of the shared property over the

complete class, including its unknown members. Often, as Whewell admits, this is a trivially simple procedure. Once Kepler supplied the conception of an ellipse to the observed members of the class of Mars' positions, he generalized it to all members of the class, including those which were unknown (unobserved), to reach the conclusion that "all the points of Mars' orbit lie on an ellipse with the sun at one focus." He then performed a further generalization to reach his first law of planetary motion: "the orbits of all the planets lie on ellipses with the sun at one focus."

I mentioned earlier that Whewell thought of himself as renovating Bacon's inductive philosophy. His inductivism does share numerous features with Bacon's method of interpreting nature: for instance the claims that induction must involve more than merely simple enumeration of instances, that science must be proceed by successive steps of generalization, that inductive science can reach unobservables (for Bacon, the "forms," for Whewell, unobservable entities such as light waves or properties such as elliptical orbits or gravitational forces). (For more on the relation between Whewell and Bacon see Snyder 1999). Yet, surprisingly, the received view of Whewell's methodology in the 20th century has tended to describe Whewell as an anti-inductivist in the Popperian mold (see, for example, Butts 1987, Buchdahl 1991, Laudan, 1980, Niiniluoto 1977, and Ruse 1975). That is, it is claimed that Whewell endorses a "conjectures and refutations" view of scientific discovery. However, it is clear from the above discussion that his view of discoverers' induction does not resemble the view asserting that hypotheses can be and are typically arrived at by mere guesswork. Moreover, Whewell explicitly rejects the hypothetico-deductive claim that hypotheses discovered by non-rational guesswork can be confirmed by consequentialist testing. For example, in his review of his friend Herschel's *Preliminary Discourse on the Study of Natural Philosophy*, Whewell argues, against Herschel, that verification is not possible when a hypothesis has been formed non-inductively (1831, pp. 400-1). Nearly thirty years later, in the last edition of the *Philosophy*, Whewell refers to the belief that "the discovery of laws and causes of phenomena is a loose hap-hazard sort of guessing," and claims that this type of view "appears to me to be a misapprehension of the whole nature of science" (1860a, p. 274). In other mature works he notes that discoveries are made "not by any capricious conjecture of arbitrary selection" (1858a, I, p. 29) and explains that new hypotheses are properly "collected from the facts" (1849, p. 17).

Why has Whewell been misinterpreted by so many commentators? One reason has to do with the error of reading certain terms used by Whewell in the 19th century as if they held the same meaning they have in the 20th and 21st. Thus, since Whewell does at times speak of "conjectures" and "guesses," we are told that he shares Popper's methodology. He speaks, for instance, of the "happy guesses" made by scientists (1858b, p. 64) and claims that "advances in knowledge" often follow "the previous exercise of some boldness and license in guessing" (1847, II, p. 55). But Whewell often uses these terms in a way which connotes a conclusion which is simply not conclusively confirmed. The *Oxford English Dictionary* tells us that prior to the 20th century the term "conjecture" was used to connote not a hypothesis reached by non-rational means, but rather one which is "unverified," or which is "a conclusion as to what is likely or probable" (as opposed to the results of demonstration). The term is used this way by Bacon, Kepler, Newton, and Dugald Stewart, writers whose work was well-known to Whewell. In other places where Whewell uses the term "conjecture" he suggests that what appears to be the result of guesswork is actually what we might call an "educated guess," i.e., a conclusion drawn by (weak) inference. Whewell describes Kepler's discovery, which seems so "capricious and fanciful" as actually being "regulated" by

his "clear scientific ideas" (1857/1873, I, pp. 291-2). Finally Whewell's use of the terminology of guessing sometimes occurs in the context of a distinction he draws between the generation of a number of possible conceptions, and the selection of one to superinduce upon the facts. Before the appropriate conception is found, the scientist must be able to call up in his mind a number of possible ones (see 1858b, p. 79). Whewell notes that this calling up of many possibilities "is, in some measure, a process of conjecture." However, selecting the appropriate conception with which to colligate the data is not conjectural (1858b, p. 78). Thus Whewell claims that the selection of the conception is often "*preluded* by guesses" (1858b, p. xix); he does not, that is, claim that the selection *consists* in guesswork. When inference is not used to select the appropriate conception, the resulting theory is not an "induction," but rather a "hasty and imperfect hypothesis." He draws such a distinction between Copernicus' heliocentric theory, which he calls an induction, and the heliocentric system proposed by Aristarchus in the third century b.c., to which he refers as a hasty and imperfect hypothesis (1857/1873, I, p. 258).

Thus Whewell's philosophy of science cannot be described as the hypothetico-deductive view. It is an inductive method; yet it clearly differs from the more narrow inductivism of Mill. Whewell's view of induction has the advantage over Mill's of allowing the inference to unobservable properties and entities (for more on this topic see Snyder 1997a and 1997b).

3. Philosophy of Science: Confirmation

On Whewell's view, once a theory is invented by discoverers' induction, it must pass a variety of deductive tests before it can be considered confirmed as an empirical truth. These tests are prediction, consilience, and coherence (see 1858b, pp. 83-96). These are characterized by Whewell as, first, that "our hypotheses ought to *fortel* [sic] phenomena which have not yet been observed" (1858b, p. 86); second, that they should "explain and determine cases of a *kind different* from those which were contemplated in the formation" of those hypotheses (1858b, p. 88); and third that hypotheses must "become more coherent" over time (1858b, p. 91).

I start by discussing the criterion of prediction. Our hypotheses ought to foretell phenomena, "at least all phenomena of the same kind," Whewell explains, because "our assent to the hypothesis implies that it is held to be true of all particular instances. That these cases belong to past or to future times, that they have or have not already occurred, makes no difference in the applicability of the rule to them. Because the rule prevails, it includes all cases" (1858b, p. 86). Whewell's point here is simply that since our hypotheses are in universal form, a true hypothesis will cover all particular instances of the rule, including past, present, and future cases. But he also makes the stronger claim that successful predictions of unknown facts provide greater confirmatory value than explanations of already-known facts. Thus he holds the historical claim that "new evidence" is more valuable than "old evidence." He argues that "to predict unknown facts found afterwards to be true is ... a confirmation of a theory which in impressiveness and value goes beyond any explanation of known facts" (1857/1873, II, p. 557). Whewell claims that the agreement of the prediction with what occurs (i.e., the fact that the prediction turns out to be correct), is "nothing strange, if the theory be true, but quite unaccountable, if it be not" (1860a, pp. 273-4). For example, if Newtonian theory were not true, he argues, the fact that from the theory we could

correctly predict the existence, location and mass of a new planet, Neptune (as did happen in 1846), would be bewildering, and indeed miraculous.

An even more valuable confirmation criterion, according to Whewell, is that of "consilience." Whewell explains that "the evidence in favour of our induction is of a much higher and more forcible character when it enables us to explain and determine [i.e., predict] cases of a *kind different* from those which were contemplated in the formation of our hypothesis. The instances in which this have occurred, indeed, impress us with a conviction that the truth of our hypothesis is certain" (1858b, pp. 87-8). Whewell calls this type of evidence a "jumping together" or "consilience" of inductions. An induction, which results from the colligation of one class of facts, is found also to colligate successfully facts belonging to another class. This is especially powerful as confirmation when the second class of phenomena had appeared to be unrelated to the first. For instance, consilience occurred when "the force of Universal Gravitation, which had been inferred from the Perturbations of the moon and planets by the sun and by each other, also accounted for the fact, apparently altogether dissimilar and remote, of the *Precession of the equinoxes*" (1847, II, p. 66).

Whewell discusses a further, related test of a theory's truth: namely, "coherence." In the case of true theories, Whewell claims, "the system becomes more coherent as it is further extended. The elements which we require for explaining a new class of facts are already contained in our system....In false theories, the contrary is the case" (1858b, p. 91). Coherence occurs when we are able to extend our hypothesis to colligate a new class of phenomena without ad hoc modification of the hypothesis. When Newton extended his theory regarding an inverse-square attractive force, which colligated facts of planetary motion and lunar motion, to the class of "tidal activity," he did not need to add any new suppositions to the theory in order to colligate correctly the facts about particular tides. On the other hand, Whewell claims, when phlogiston theory, which colligated facts about the class of phenomena "chemical combination," was extended to colligate the class of phenomena "weight of bodies," it was *unable* to do so without an ad hoc and implausible modification (namely, the assumption that phlogiston has "negative weight") (see 1858b, pp. 92-3). Thus coherence can be seen as a type of consilience that happens over time; indeed, Whewell remarks that these two criteria--consilience and coherence--"are, in fact, hardly different" (1858b, p. 95).

4. Philosophy of Science: Necessary Truth

A particularly intriguing aspect of Whewell's philosophy of science is his claim that empirical science can reach necessary truths. Explaining this apparently contradictory claim was considered by Whewell and Herschel to be the "ultimate problem" of philosophy (see Morrison 1997). Whewell explains it by reference to his antithetical epistemology. Necessary truths are truths which can be known a priori; they can be known in this way because they are necessary consequences of ideas which are a priori. They are necessary consequences in the sense of being analytic consequences; Whewell explicitly rejects Kant's claim that necessary truths are synthetic. Using the example "7 + 8 = 15," Whewell claims that "we refer to our conceptions of seven, of eight, and of addition, and as soon as we possess the conceptions distinctly, we see that the sum must be 15." Merely by knowing the *meanings* of "seven," and "eight," and

"addition," we see that it follows necessarily that "7 + 8 = 15" (1848, p. 471).

Once the Ideas and conceptions are explicated, so that we understand their meanings, the necessary truths which follow from them are seen as being necessarily true. Thus, once the Idea of Space is explicated, it is seen to be necessarily true that "two straight lines cannot enclose a space." Whewell suggests that the first law of motion is also a necessary truth, which was knowable a priori once the idea of Cause and the associated conception of force were explicated. This is why empirical science is needed to see necessary truths—because, as we saw above, empirical science is needed in order to explicate the ideas. Thus Whewell also claims that, in the course of science, truths which at first required experiment to be known are seen to be capable of being known independently of experiment. That is, once the idea is clarified, the necessary connection between the idea and an empirical truth becomes apparent. Whewell claims that "though the discovery of the First Law of Motion was made, historically speaking, by means of experiment, we have now attained a point of view in which we see that it might have been certainly known to be true independently of experience" (1847, I, p. 221). Science, then, consists in the "idealization of facts," the transferring of truths from the empirical to the ideal side of the fundamental antithesis. He describes this by noting that there is a "progressive intuition of necessary truths."

Although they follow analytically from the meanings of ideas our minds supply, necessary truths are nevertheless informative statements about the physical world outside us; they have empirical content. Whewell's justification for this claim is a theological one. Whewell notes that God created the universe in accordance with certain "Divine Ideas." That is, all objects and events in the world were created by God to conform to certain of his ideas. For example, God made the world such that it corresponds to the idea of Cause partially expressed by the axiom "every event has a cause." Hence in the universe every event conforms to this idea, not only by having a cause but by being such that it could not occur without a cause. On Whewell's view, we are able to have knowledge of the world because the fundamental ideas which are used to organize our sciences resemble the ideas used by God in his creation of the physical world. The fact that this is so is no coincidence: God has created our minds such that they contain these same ideas. That is, God has given us our ideas so that "they can and must agree with the world" (1860a, p. 359). God intends that we can have knowledge of the physical world, and this is possible only through the use of ideas which resemble those that were used in creating the world. Hence with our ideas we can colligate correctly the facts of the world and form true theories. And when these ideas are distinct, we can know a priori the axioms which express their meaning.

An interesting consequence of this interpretation of Whewell's view of necessity is that every law of nature is a necessary truth, in virtue of following analytically from some idea used by God in creating the world. Whewell draws no distinction between truths which can be idealized and those which cannot; thus, potentially, any empirical truth can be seen to be a necessary truth, once the ideas and conceptions are explicated sufficiently. For example, Whewell suggests that experiential truths such as "salt is soluble" may be necessary truths, even if we do not recognize this necessity (i.e., even if it is not yet knowable a priori): (1860b, p. 483). Whewell's view thus destroys the line traditionally drawn between laws of nature and the axiomatic propositions of the pure sciences of mathematics; mathematical truth is granted no special status.

Whewell thus suggests a view of scientific understanding which is, not surprisingly, grounded in his conception of natural theology. Since our ideas are "shadows" of the Divine Ideas, to see a law as a necessary consequence of our ideas is to see it as a consequence of the Divine Ideas exemplified in the world. Understanding involves seeing a law as being not an arbitrary "accident on the cosmic scale," but as a necessary consequence of the ideas God used in creating the universe. Hence the more we idealize the facts, the more difficult it will be to deny God's existence. We will come to see more and more truths as the intelligible result of intentional design. This view is related to the claim Whewell makes in his Bridgewater Treatise (1833), that the more we study the laws of nature the more convinced we will be in the existence of a Divine Law-giver

5. Relation of Science to Philosophy of Science

An issue of interest to philosophers of science today is the relation between knowledge of the actual practice and history of science and writing philosophy of science. Whewell is interesting to examine in relation to this issue because he claimed to be inferring his philosophy of science from his study of the history and practice of science. His large-scale *History of the Inductive Sciences* (first edition published 1837) was a survey of science from ancient to modern times. Besides knowing about science by studying its history, Whewell had first-hand knowledge: he was actively involved in science in several important ways. In 1825 he traveled to Berlin, Freiburg and Vienna to study mineralogy and crystallography with Mohs and other acknowledged masters of the field (see Becher, 1986). He published numerous papers in the field, as well as a monograph, and is still credited with making important contributions to giving a mathematical foundation to crystallography. He also made contributions to the science of tidal research, pushing for a large-scale world-wide project of tidal observations; he won a Royal Society gold medal for this accomplishment (see Ruse, 1991). Whewell acted as a terminological consultant for Faraday and other scientists, who wrote to him asking for new words. Whewell only provides terminology when he feels he is fully knowledgeable about the science involved. In his section on the "Language of Science" in the *Philosophy*, Whewell makes clear this position (see 1858b, p. 293). Another interesting aspect of his intercourse with scientists becomes clear in reading his correspondence with them: namely, that Whewell constantly pushes Faraday, Forbes, Lubbock and others to perform certain experiments, make specific observations, and to try to connect their findings in ways of interest to Whewell (see Whewell Papers). In all these ways, Whewell indicates that he has a deep understanding of the activity of science.

So how is this important for his work on the philosophy of science? Some commentators have claimed that Whewell developed an a priori philosophy of science and then shaped his *History* to conform to his own view (see Stoll 1929 and Strong 1955). It is true that he starts out, from his undergraduate days, with the project of reforming the inductive philosophy of Bacon; indeed this early inductivism leads him to the view that learning about scientific method must be inductive (i.e., that it requires the study of the history of science). Yet he also refuses to complete his *Philosophy* before he has written the *History*; he even sends proof-sheets of the *History* to his many scientist-friends to ensure the accuracy of his account. Ultimately, he criticizes Mill's view of induction developed in the *System of Logic* on the grounds that Mill has not found a large number of appropriate examples illustrating the use of his "Methods of Experimental Inquiry." Thus it appears that what is important to Whewell is not whether a philosophy of

science is, in fact, inferred from a study of the history of science, but rather, whether a philosophy of science is *inferable from* it. That is, regardless of how a philosopher came to invent her theory, she must be able to show it to be exemplified in the actual scientific practice used throughout history.

6. Moral Philosophy

Whewell's moral philosophy was criticized by Mill as being "intuitionist" (see Mill, 1852). Mill meant two claims about his moral philosophy by the use of this term. The first characteristic of an intuitionist morality is the claim that there are primary principles of morality which are known by intuition. Whewell does claim this, but what he means by "intuition" is not the non-rational process of which Mill accuses him. For Whewell the contemplation of moral principles is a rational process; thus he refers to moral rules as "principles of reason" (1864, p. 3), and describes the discovery of these rules as an activity of reason (1864, pp. 23-4). Unlike other intuitionist moral philosophers, Whewell holds that intuition plays only a small role in our moral decision-making; reason leads to common decisions about the correct way to act (see 1864, p. 43). Thus his position is more accurately described as a form of rationalism, or "rational intuitionism," in the sense later used by Rawls (see Singer, 1992).

The second central aspect of intuitionism is the claim that moral rules are necessary truths which are self-evident. This is often taken, as it is by Mill, to lead to the conclusion that there can be no progress in morality—what is self-evident must always remain so—and thus to the further conclusion that the intuitionist considers the current rules of society to be necessary truths. Such a view would tend to support the status quo, as Mill rightly complains. (Thus he accuses Whewell of justifying evil practices such as slavery, forced marriages, and cruelty to animals.) But Mill is wrong to attribute such a view to Whewell. Whewell does claim that moral rules are necessary truths, and invests them with the epistemological status of self-evident "axioms" (see 1864, p. 58). However, as noted above, Whewell's view of necessary truth is a progressive one. The realm of morality, like the realm of physical science, is structured by certain Fundamental Ideas: Benevolence, Justice, Truth, Purity, and Order (see 1852, p. xxiii). These moral ideas are conditions of our moral experience; they enable us to perceive actions as being in accordance with the demands of morality. Like the ideas of the physical sciences, the ideas of morality must be explicated before the moral rules can be derived from them (see 1860a, p. 388). There is a progressive intuition of necessary truth in morality as well as in science. Hence it does not follow that because the moral truths are axiomatic and self-evident that we currently know them (see 1846, pp. 38-9). Whewell thus claims that "to test self-evidence by the casual opinion of individual men, is a self-contradiction" (1846, p. 35). Nevertheless, Whewell does claim that we can look to the dictates of positive law of the most morally advanced societies as a starting point in our explication of the moral ideas. But he is not therefore suggesting that these laws are the standard of morality. Just as we examine the phenomena of the physical world in order to explicate our scientific conceptions, we can examine the facts of positive law and the history of moral philosophy in order to explicate our moral conceptions. Mill is therefore wrong to interpret Whewell's moral philosophy as a justification of the status quo or as constituting a "vicious circle." Rather, Whewell's view shares some features of Rawls's later use of the notion of "reflective equilibrium."

Bibliography

Whewell's letters and papers, mostly unpublished, are found in the Whewell Collection, Trinity College Library, Cambridge. A selection of letters was published by I. Todhunter in *William Whewell, An account of his Writings*, Vol. II (London, 1876) and by J. Stair-Douglas in *The Life, and Selections from the Correspondence of William Whewell* (London, 1882).

During his lifetime Whewell published approximately 150 books, articles, scientific papers, society reports, reviews, and translations. In the list which follows I mention only his most important philosophical works. More complete bibliographies can be found in Yeo (1993) and Fisch and Schaffer (1991).

Major Philosophical Works by Whewell

- (1831) "Review of J. Herschel's *Preliminary Discourse on the Study of Natural Philosophy (1830)," Quarterly Review 90*: 374-407.
- (1833) *Astronomy and General Physics Considered With Reference to Natural Theology* (Bridgewater Treatise), Cambridge.
- (1840) *The Philosophy of the Inductive Sciences, Founded Upon Their History*, in two volumes, London.
- (1844) "On the Fundamental Antithesis of Philosophy," Transactions of the Cambridge Philosophical Society 7, part 2, 170-81.
- (1845) *The Elements of Morality, including Polity*, in two volumes, London.
- (1846) *Lectures on Systematic Morality*, London.
- (1847) *The Philosophy of the Inductive Sciences, Founded Upon Their History*, 2nd edition, in two volumes, London.
- (1848) "Second Memoir on the Fundamental Antithesis of Philosophy," *Transactions of the Cambridge Philosophical Society* 8, part five, pp. 614-20.
- (1849) *Of Induction, With Especial Reference to Mr. J. Stuart Mill's System of Logic*, London.
- (1850) "Mathematical Exposition of Some Doctrines of Political Economy: Second Memoir," *Transactions of the Cambridge Philosophical Society* 9:128-49.
- (1853) *Of the Plurality of Worlds. An Essay*, London.
- (1857) "Spedding's Complete Edition of the Works of Bacon," *Edinburgh Review* 106:287-322.
- (1857/1873) *History of the Inductive Sciences, from the Earliest to the Present Time*, 3rd edition, in two volumes, New York.
- (1858a) *The History of Scientific Ideas*, in two volumes, London.
- (1858b) *Novum Organon Renovatum*, London.
- (1860a) *On the Philosophy of Discovery: Chapters Historical and Critical*, London.
- (1860b) "Remarks on a Review of the Philosophy of the Inductive Sciences," letter to John Herschel, 11 April 1844; published as essay F in 1860a.
- (1861) Plato's *Republic* (translation). Cambridge.
- (1862) *Six Lectures on Political Economy*, Cambridge.

- (1864) *The Elements of Morality, Including Polity*. 4th edition, with Supplement, Cambridge.
- (1866a) "Comte and Positivism," *Macmillan's Magazine* 13:353-62.
- (1866b) "Grote's Plato," *Fraser's Magazine* 73:411-23.

Selected Works on Whewell

- Becher, H. (1981) "William Whewell and Cambridge Mathematics," *Historical Studies in the Physical Sciences* 11:1-48.
- ----------(1986) "Voluntary Science in Nineteenth-Century Cambridge University to the 1850s," *British Journal for the History of Science* 19: 57-87.
- Blanche, R. (1935) *Le Rationalisme de Whewell*, Paris.
- Brooke, J.H. (1977) "Natural Theology and the Plurality of Worlds: Observations on the Brewster-Whewell Debate," *Annals of Science* 34: 221-86.
- Buchdahl, G. (1991) "Deductivist versus Inductivist Approaches in the Philosophy of Science as Illustrated by Some Controversies Between Whewell and Mill," in Fisch and Schaffer (eds.), pp. 311-44.
- Butts, R. (1973) "Whewell's Logic of Induction," in R.N. Giere and R.S. Westfall (eds.), *Foundations of Scientific Method: The Nineteenth Century*, Bloomington, IN., pp. 53-85.
- ----------(1987) "Pragmatism in Theories of Induction in the Victorian Era: Herschel, Whewell, Mach and Mill," in H. Stachowiak (ed.) *Pragmatik: Handbuch Pragmatishchen Denkens* (Hamburg), pp. 40-58.
- Donagan, A. (1992) "Sidgwick and Whewellian Intuitionism: Some Enigmas," in Schultz (ed.) pp. 123-42.
- Fisch, M. (1991) *William Whewell, Philosopher of Science*, Oxford.
- Fisch, M. and S. Schaffer (eds.) (1991) *William Whewell: A Composite Portrait*, Oxford.
- Harper, W. (1989) "Consilience and Natural Kind Reasoning," in J.R. Brown and J. Mittelstrass (eds.) *An Intimate Relation*, Dordrecht, pp. 115-52.
- Herschel, J. (1841) "Whewell on Inductive Sciences," *Quarterly Review* 68:177-238.
- Laudan, L. (1971) "William Whewell on the Consilience of Inductions," *Monist* 55:368-91.
- ----------(1980) "Why was the Logic of Discovery Abandoned?" in T. Nickles (ed.). *Scientific Discovery, Logic, and Rationality*, Dordrecht, pp. 173-183.
- Losee, J. (1983) "Whewell and Mill on the Relation between Science and Philosophy of Science," *Studies in History and Philosophy of Science* 14:113-26.
- Marcucci, S. (1963) *L'idealismo scientifico di William Whewell*, Pisa.
- Mill, J.S. (1836) "Dr. Whewell on Moral Philosophy," *Westminster Review* 58:349-85.
- Morrison, M. (1997) "Whewell on the Ultimate Problem of Philosophy," *Studies in History and Philosophy of Science*.
- Niiniluoto, I. (1977) "Notes on Popper as a Follower of Whewell and Peirce," *Ajatus* 37:272-327.
- Peirce, C.S. (1865 [1982]) "Lecture on the Theories of Whewell, Mill and Comte," in M. Fisch (ed.) *Writings of Charles S. Peirce: Chronological Edition*, Bloomington Ind., pp. 205-23.
- Ruse, M. (1975) "Darwin's Debt to Philosophy: An Examination of the Influence of the Philosophical Ideas of John F.W. Herschel and William Whewell on the Development of Charles Darwin's Theory of Evolution," *Studies in History and Philosophy of Science* 6: 159-81.

- ----------(1976) "The Scientific Methodology of William Whewell," *Centaurus* 20:227-57.
- ----------(1977) "William Whewell and the Argument from Design," *Monist* 60:244-68.
- ---------- (1991) "William Whewell: Omniscientist," in Fisch and Schaffer (eds.), pp. 87-116.
- Schultz, B. (1992) *Essays on Henry Sidgwick*. Cambridge.
- Singer, M. (1992) "Sidgwick and 19th century Ethical Thought," in Schultz (ed.), pp. 65-91.
- Snyder, L.J. (1994) "It's All Necessarily So: William Whewell on Scientific Truth," *Studies in History and Philosophy of Science* 25: 785-807.
- ----------(1997a) "Discoverers' Induction," *Philosophy of Science* 64:580-604.
- ----------(1997b) "The Mill-Whewell Debate: Much Ado About Induction," *Perspectives on Science* 5: 159-198.
- ----------(1998) "Is Evidence Historical?" in *Philosophy of Science*, The Central Issues, ed. M. Curd and J.A. Cover (New York), pp. 460-80.
- ----------(1999) "Renovating the *Novum Organum*: Bacon, Whewell and Induction," *Studies in History and Philosophy of Science* 30:531-557.
- Stoll, M.R. (1929) *Whewell's Philosophy of Induction*, Lancaster, PA.
- Strong, E.W. (1955) "William Whewell and John Stuart Mill: Their Controversy over Scientific Knowledge," *Journal of the History of Ideas* 16:209-31.
- Yeo, R. (1979) "William Whewell, Natural Theology and the Philosophy of Science in mid-19th century Britain," *Annals of Science* 36:493-512.
- ----------(1993) *Defining Science: William Whewell, Natural Knowledge, and Public Debate in Early Victorian Britain*, Cambridge, UK.

Other Internet Resources

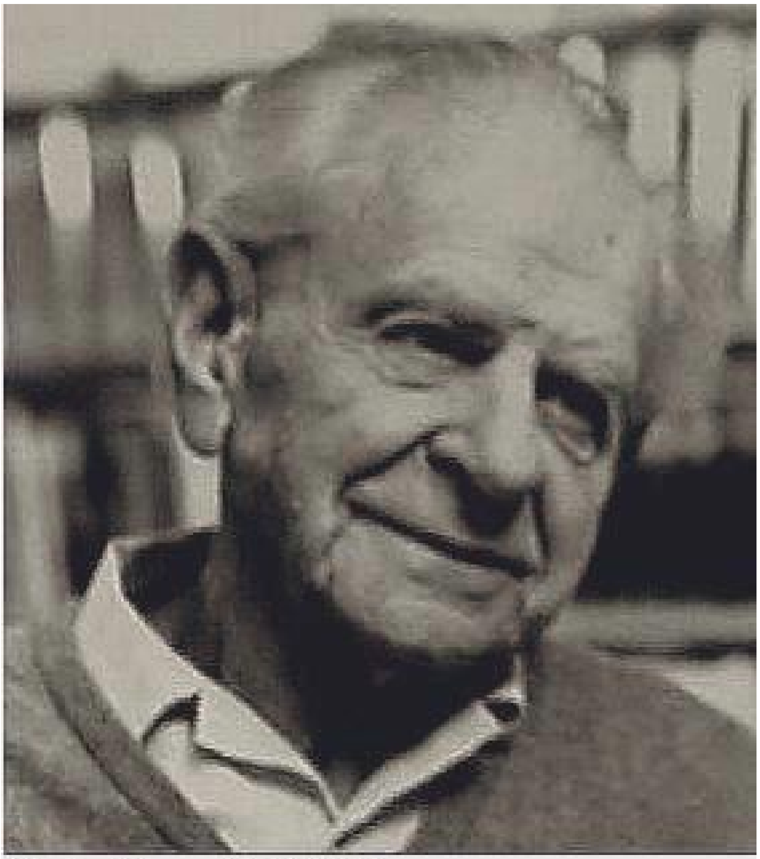
[Please contact the author with suggestion.]

Related Entries

Bacon, Francis | confirmation | induction: problem of | Kant, Immanuel | Mill, John Stuart | Popper, Karl

Laura J. Snyder
St. Johns University
snyderl@stjohns.edu

First published: December 22, 2000
Content last modified: December 22, 2000



Sir Karl Popper (1902-1994)

Karl Popper

Karl Popper is generally regarded as one of the greatest philosophers of science of this century. He was also a social and political philosopher of considerable stature, a self-professed ‘critical-rationalist’, a dedicated opponent of all forms of scepticism, conventionalism, and relativism in science and in human affairs generally, a committed advocate and staunch defender of the ‘Open Society’, and an implacable critic of totalitarianism in all of its forms. One of the many remarkable features of Popper's thought is the scope of his intellectual influence. In the modern technological and highly-specialised world scientists are rarely aware of the work of philosophers; it is virtually unprecedented to find them queuing up, as they have done in Popper's case, to testify to the enormously practical beneficial impact which that philosophical work has had upon their own. But notwithstanding the fact that he wrote on even the most technical matters with consummate clarity, the scope of Popper's work is such that it is commonplace by now to find that commentators tend to deal with the epistemological, scientific and social elements of his thought as if they were quite disparate and unconnected, and thus the fundamental unity of his philosophical vision and method has to a large degree been dissipated. Here we will try to trace the threads which interconnect the various elements of his philosophy, and which give it its fundamental

unity.

Section Headings:

Life

Karl Raimund Popper was born on 28 July 1902 in Vienna, which at that time could make some claim to be the cultural epicentre of the western world. His parents, who were of Jewish origin, brought him up in an atmosphere which he was later to describe as ‘decidedly bookish’. His father was a lawyer by profession, but he also took a keen interest in the classics and in philosophy, and communicated to his son an interest in social and political issues which he was to never lose. His mother inculcated in him such a passion for music that for a time he seriously contemplated taking it up as a career, and indeed he initially chose the history of music as a second subject for his Ph.D examination. Subsequently, his love for music became one of the inspirational forces in the development of his thought, and manifested itself in his highly original interpretation of the relationship between dogmatic and critical thinking, in his account of the distinction between objectivity and subjectivity, and, most importantly, in the growth of his hostility towards all forms of historicism, including historicist ideas about the nature of the ‘progressive’ in music. The young Karl attended the local *Realgymnasium*, where he was unhappy with the standards of the teaching, and, after an illness which kept him at home for a number of months, he left to attend the University of Vienna in 1918. However, he did not formally enrol at the University by taking the matriculation examination for another four years. 1919 was in many respects the most important formative year of his intellectual life. In that year he became heavily involved in left-wing politics, joined the Association of Socialist School Students, and became for a time a Marxist. However, he was quickly disillusioned with the doctrinaire character of the latter, and soon abandoned it entirely. He also discovered the psychoanalytic theories of Freud and Adler (under whose aegis he engaged

briefly in social work with deprived children), and listened entranced to a lecture which Einstein gave in Vienna on relativity theory. The dominance of the critical spirit in Einstein, and its total absence in Marx, Freud and Adler, struck Popper as being of fundamental importance: the latter, he came to think, couched their theories in terms which made them amenable only to confirmation, while Einstein's theory, crucially, had testable implications which, if false, would have falsified the theory itself.

Popper obtained a primary school teaching diploma in 1925, took a Ph.D. in philosophy in 1928, and qualified to teach mathematics and physics in secondary school in 1929. The dominant philosophical group in Vienna from its inception in 1928 was the *Wiener Kreis*, the circle of ‘scientifically-minded’ intellectuals who gathered around the figure of Moritz Schlick. This included Rudolf Carnap, Otto Neurath, Viktor Kraft, Hans Hahn, and Herbert Feigl. The principal objective of the members of the Circle was to unify the sciences, which carried with it, in their view, the need to eliminate metaphysics once and for all by showing that metaphysical propositions are meaningless. Thus was born the movement in philosophy known as logical positivism, and its chief tool became the verification principle. Although he was friendly with some of the Circle's members - especially Feigl, who encouraged him to write his first book - and shared their esteem for science, Popper was heavily critical of the main tenets of logical positivism, especially of what he considered to be its misplaced focus on the theory of meaning in philosophy and upon verification in scientific methodology. He articulated his own view of science, and his criticisms of the positivists, in his first work, published under the title *Logik der Forschung* in 1934. The book - which he was later to claim rang the death knell for logical positivism - attracted more attention than Popper had anticipated, and he was invited to lecture in England in 1935. He spent the next few years working productively on science and philosophy, but storm clouds were gathering - the growth of Nazism in Germany and Austria compelled him, like many other intellectuals who shared his Jewish origins, to leave his native country.

In 1937 Popper took up a position teaching philosophy at the University of Canterbury in New Zealand, where he was to remain for the duration of the Second World War. The annexation of Austria in 1938 became the catalyst which prompted him to refocus his writings on social and political philosophy. In 1946 he moved to England to teach at the London School of Economics, and became professor of logic and scientific method at the University of London in 1949. From this point on Popper's reputation and stature as a philosopher of science and social thinker grew enormously, and he continued to write prolifically - a number of his works, particularly *The Logic of Scientific Discovery* (1959), are now universally recognised as classics in the field. He was knighted in 1965, and retired from the University of London in 1969, though he remained active as a writer, broadcaster and lecturer until his death in 1994. (For more detail on Popper's life, cf. his *Unended Quest*).
[Return to Section Headings]

Backdrop to his Thought

A number of biographical features may be identified as having a particular influence upon Popper's thought. In the first place, his teenage flirtation with Marxism left him thoroughly familiar with the Marxist view of economics, class-war, and history. Secondly, he was appalled by the failure of the

democratic parties to stem the rising tide of fascism in his native Austria in the 1920s and 1930s, and the effective welcome extended to it by the Marxists. The latter acted on the ideological grounds that it constituted what they believed to be a necessary dialectical step towards the implosion of capitalism and the ultimate revolutionary victory of communism. This was one factor which led to the much feared *Anschluss*, the annexation of Austria by the German Reich, the anticipation of which forced Popper into permanent exile from his native country. *The Poverty of Historicism* (1944) and *The Open Society and Its Enemies* (1945), his most impassioned and brilliant social works, are as a consequence a powerful defence of democratic liberalism as a social and political philosophy, and a devastating critique of the principal philosophical presuppositions underpinning all forms of totalitarianism. Thirdly, as we have seen, Popper was profoundly impressed by the differences between the allegedly ‘scientific’ theories of Freud and Adler and the revolution effected by Einstein's theory of relativity in physics in the first two decades of this century. The main difference between them, as Popper saw it, was that while Einstein's theory was highly ‘risky’, in the sense that it was possible to deduce consequences from it which were, in the light of the then dominant Newtonian physics, highly improbable (e.g. that light is deflected towards solid bodies - confirmed by Eddington's experiments in 1919), and which would, if they turned out to be false, falsify the whole theory, nothing could, even *in principle*, falsify psychoanalytic theories. These latter, Popper came to feel, have more in common with primitive myths than with genuine science. That is to say, he saw that what is apparently the chief source of strength of psychoanalysis, and the principal basis on which its claim to scientific status is grounded, viz. its capability to accommodate, and explain, every possible form of human behaviour, is in fact a critical weakness, for it entails that it is not, and could not be, genuinely predictive. Psychoanalytic theories by their nature are insufficiently precise to have negative implications, and so are immunised from experiential falsification.

The Marxist account of history too, Popper held, is not scientific, although it differs in certain crucial respects from psychoanalysis. For Marxism, Popper believed, had been initially scientific, in that Marx had postulated a theory which was genuinely predictive. However, when these predictions were not in fact borne out, the theory was saved from falsification by the addition of *ad hoc* hypotheses which made it compatible with the facts. By this means, Popper asserted, a theory which was initially genuinely scientific degenerated into pseudo-scientific dogma.

These factors combined to make Popper take *falsifiability* as his criterion for demarcating science from non-science: if a theory is incompatible with possible empirical observations it is scientific; conversely, a theory which is compatible with all such observations, either because, as in the case of Marxism, it has been modified solely to accommodate such observations, or because, as in the case of psychoanalytic theories, it is consistent with all possible observations, is unscientific. For Popper, however, to assert that a theory is unscientific, is not necessarily to hold that it is unenlightening, still less that it is meaningless, for it sometimes happens that a theory which is unscientific (because it is unfalsifiable) at a given time may become falsifiable, and thus scientific, with the development of technology, or with the further articulation and refinement of the theory. Further, even purely mythogenic explanations have performed a valuable function in the past in expediting our understanding of the nature of reality.
[Return to Section Headings]

The Problem of Demarcation

As Popper represents it, the central problem in the philosophy of science is that of demarcation, i.e. of distinguishing between science and what he terms 'non-science', under which heading he ranks, amongst others, logic, metaphysics, psychoanalysis, and Adler's individual psychology. Popper is unusual amongst contemporary philosophers in that he *accepts* the validity of the Humean critique of induction, and indeed, goes beyond it in arguing that induction is never actually used by the scientist. However, he does not concede that this entails the scepticism which is associated with Hume, and argues that the Baconian/Newtonian insistence on the primacy of 'pure' observation, as the initial step in the formation of theories, is completely misguided: all observation is selective and theory-laden - there are no pure or theory-free observations. In this way he destabilises the traditional view that science can be distinguished from non-science on the basis of its inductive methodology; in contradistinction to this, Popper holds that there is no unique methodology specific to science. Science, like virtually every other human, and indeed organic, activity, Popper believes, consists largely of problem-solving.

Popper, then, repudiates induction, and rejects the view that it is the characteristic method of scientific investigation and inference, and substitutes *falsifiability* in its place. It is easy, he argues, to obtain evidence in favour of virtually any theory, and he consequently holds that such 'corroboration', as he terms it, should count scientifically only if it is the positive result of a genuinely 'risky' prediction, which might conceivably have been false. For Popper, a theory is scientific only if it is refutable by a conceivable event. Every genuine test of a scientific theory, then, is logically an attempt to refute or to falsify it, and one genuine counter-instance falsifies the whole theory. In a critical sense, Popper's theory of demarcation is based upon his perception of the logical asymmetry which holds between verification and falsification: it is logically impossible to conclusively verify a universal proposition by reference to experience (as Hume saw clearly), but a single counter-instance conclusively falsifies the corresponding universal law. In a word, an exception, far from 'proving' a rule, conclusively refutes it.

Every genuine scientific theory then, in Popper's view, is *prohibitive*, in the sense that it forbids, by implication, particular events or occurrences. As such it can be tested and falsified, but never logically verified. Thus Popper stresses that it should not be inferred from the fact that a theory has withstood the most rigorous testing, for however long a period of time, that it has been verified; rather we should recognise that such a theory has received a high measure of corroboration. and may be provisionally retained as the best available theory until it is finally falsified (if indeed it is ever falsified), and/or is superseded by a better theory.

Popper has always drawn a clear distinction between the *logic* of falsifiability and its *applied methodology*. The logic of his theory is utterly simple: if a single ferrous metal is unaffected by a magnetic field it cannot be the case that all ferrous metals are affected by magnetic fields. Logically speaking, a scientific law is conclusively falsifiable although it is not conclusively verifiable. Methodologically, however, the situation is much more complex: no observation is free from the possibility of error - consequently we may question whether our experimental result was what it appeared to be.

Thus, while advocating falsifiability as the criterion of demarcation for science, Popper explicitly allows for the fact that in practice a single conflicting or counter-instance is never sufficient methodologically to falsify a theory, and that scientific theories are often retained even though much of the available evidence conflicts with them, or is anomalous with respect to them. Scientific theories may, and do, arise genetically in many different ways, and the manner in which a particular scientist comes to formulate a particular theory may be of biographical interest, but it is of no consequence as far as the philosophy of science is concerned. Popper stresses in particular that there is no unique way, no single method such as induction, which functions as the route to scientific theory, a view which Einstein personally endorsed with his affirmation that ‘There is no logical path leading to [the highly universal laws of science]. They can only be reached by intuition, based upon something like an intellectual love of the objects of experience’. Science, in Popper's view, starts with problems rather than with observations - it is, indeed, precisely in the context of grappling with a problem that the scientist makes observations in the first instance: his observations are selectively designed to test the extent to which a given theory functions as a satisfactory solution to a given problem.

On this criterion of demarcation physics, chemistry, and (non-introspective) psychology, amongst others, are sciences, psychoanalysis is a pre-science (i.e. it undoubtedly contains useful and informative truths, but until such time as psychoanalytical theories can be formulated in such a manner as to be falsifiable, they will not attain the status of scientific theories), and astrology and phrenology are pseudo-sciences. Formally, then, Popper's theory of demarcation may be articulated as follows: where a ‘basic statement’ is to be understood as a particular observation-report, then we may say that a theory is scientific if and only if it divides the class of basic statements into the following two non-empty sub-classes: (a) the class of all those basic statements with which it is inconsistent, or which it prohibits - this is the class of its *potential falsifiers* (i.e. those statements which, if true, falsify the whole theory), and (b) the class of those basic statements with which it is consistent, or which it permits (i.e. those statements which, if true, corroborate it, or bear it out).
[Return to Section Headings]

The Growth of Human Knowledge

For Popper accordingly, the growth of human knowledge proceeds from our problems and from our attempts to solve them. These attempts involve the formulation of theories which, if they are to explain anomalies which exist with respect to earlier theories, must go beyond existing knowledge and therefore require a leap of the imagination. For this reason, Popper places special emphasis on the role played by the independent creative imagination in the formulation of theory. The centrality and priority of *problems* in Popper's account of science is paramount, and it is this which leads him to characterise scientists as ‘problem-solvers’. Further, since the scientist begins with problems rather than with observations or ‘bare facts’, Popper argues that the only logical technique which is an integral part of scientific method is that of the deductive testing of theories which are not themselves the product of any logical operation. In this deductive procedure conclusions are inferred from a tentative hypothesis. These conclusions are then compared with one another and with other relevant statements to determine whether they falsify or

corroborate the hypothesis. Such conclusions are not directly compared with the facts, Popper stresses, simply because there are no ‘pure’ facts available; all observation-statements are theory-laden, and are as much a function of purely subjective factors (interests, expectations, wishes, etc.) as they are a function of what is objectively real.

How then does the deductive procedure work? Popper specifies four steps:

> (a) The first is *formal,* a testing of the internal consistency of the theoretical system to see if it involves any contradictions.
>
> (b) The second step is *semi-formal,* the axiomatising of the theory to distinguish between its empirical and its logical elements. In performing this step the scientist makes the logical form of the theory explicit. Failure to do this can lead to category-mistakes - the scientist ends up asking the wrong questions, and searches for empirical data where none are available. Most scientific theories contain analytic (i.e. *a priori*) and synthetic elements, and it is necessary to axiomatise them in order to distinguish the two clearly.
>
> (c) The third step is the comparing of the new theory with existing ones to determine whether it constitutes an advance upon them. If it does not constitute such an advance, it will not be adopted. If, on the other hand, its explanatory success matches that of the existing theories, and additionally, it explains some hitherto anomalous phenomenon, or solves some hitherto unsolvable problems, it will be deemed to constitute an advance upon the existing theories, and will be adopted. Thus science involves theoretical progress. However, Popper stresses that we ascertain whether one theory is better than another by deductively testing both theories, rather than by induction. For this reason, he argues that a theory is deemed to be better than another if (while unfalsified) it has greater empirical content, and therefore greater predictive power than its rival. The classic illustration of this in physics was the replacement of Newton's theory of universal gravitation by Einstein's theory of relativity. This elucidates the nature of science as Popper sees it: at any given time there will be a number of conflicting theories or conjectures, some of which will explain more than others. The latter will consequently be provisionally adopted. In short, for Popper any theory X is better than a ‘rival’ theory Y if X has *greater empirical content*, and hence *greater predictive power*, than Y.
>
> (d) The fourth and final step is the testing of a theory by the empirical application of the conclusions derived from it. If such conclusions are shown to be true, the theory is corroborated (but never verified). If the conclusion is shown to be false, then this is taken as a signal that the theory cannot be completely correct (logically the theory is falsified), and the scientist begins his quest for a better theory. He does not, however, *abandon* the present theory until such time as he has a better one to substitute for it. More precisely, the method of theory-testing is as follows: certain singular propositions are deduced from the new theory - these are predictions, and of special interest are those predictions which are

> 'risky' (in the sense of being intuitively implausible or of being startlingly novel) and experimentally testable. From amongst the latter the scientist next selects those which are not derivable from the current or existing theory - of particular importance are those which contradict the current or existing theory. He then seeks a decision as regards these and other derived statements by comparing them with the results of practical applications and experimentation. If the new predictions are borne out, then the new theory is *corroborated* (and the old one falsified), and is adopted as a working hypothesis. If the predictions are not borne out, then they falsify the theory from which they are derived. Thus Popper retains an element of empiricism: for him scientific method does involve making an appeal to experience. But unlike traditional empiricists, Popper holds that experience cannot *determine* theory (i.e. we do not argue or infer from observation to theory), it rather *delimits* it: it shows which theories are false, not which theories are true. Moreover, Popper also rejects the empiricist doctrine that empirical observations are, or can be, infallible, in view of the fact that they are themselves theory-laden.

The general picture of Popper's philosophy of science, then is this: Hume's philosophy demonstrates that there is a contradiction implicit in traditional empiricism, which holds both that all knowledge is derived from experience *and* that universal propositions (including scientific laws) are verifiable by reference to experience. The contradiction, which Hume himself saw clearly, derives from the attempt to show that, notwithstanding the open-ended nature of experience, scientific laws may be construed as empirical generalisations which are in some way finally confirmable by a 'positive' experience. Popper eliminates the contradiction by rejecting the first of these principles and removing the demand for empirical verification in favour of empirical falsification in the second. Scientific theories, for him, are not inductively inferred from experience, nor is scientific experimentation carried out with a view to verifying or finally establishing the truth of theories; rather, *all knowledge is provisional, conjectural, hypothetical* - we can never finally prove our scientific theories, we can merely (provisionally) confirm or (conclusively) refute them; hence at any given time we have to choose between the potentially infinite number of theories which will explain the set of phenomena under investigation. Faced with this choice, we can only eliminate those theories which are demonstrably false, and rationally choose between the remaining, unfalsified theories. Hence Popper's emphasis on the importance of the critical spirit to science - for him critical thinking is the very essence of rationality. For it is only by critical thought that we can eliminate false theories, and determine which of the remaining theories is the best available one, in the sense of possessing the highest level of explanatory force and predictive power. It is precisely this kind of critical thinking which is conspicuous by its absence in contemporary Marxism and in psychoanalysis.
[Return to Section Headings]

Probability, Knowledge and Verisimilitude

In the view of many social scientists, the more probable a theory is, the *better* it is, and if we have to choose between two theories which are equally strong in terms of their explanatory power, and differ only in that one is probable and the other is improbable, then we should choose the former. Popper

rejects this. Science, or to be precise, the working scientist, is interested, in Popper's view, in theories with a high informative content, because such theories possess a high predictive power and are consequently highly testable. But if this is true, Popper argues, then, paradoxical as it may sound, the more *improbable* a theory is the better it is scientifically, because the probability and informative content of a theory vary inversely - the higher the informative content of a theory the lower will be its probability, for the more information a statement contains, the greater will be the number of ways in which it may turn out to be false. Thus the statements which are of special interest to the scientist are those with a high informative content and (consequentially) a low probability, which nevertheless come close to the truth. Informative content, which is in inverse proportion to probability, is in direct proportion to testability. Consequently the severity of the test to which a theory can be subjected, and by means of which it is falsified or corroborated, is all-important.

For Popper, all scientific criticism must be piecemeal, i.e. he holds that it is not possible to question every aspect of a theory at once. More precisely, while attempting to resolve a particular problem a scientist of necessity accepts all kinds of things as unproblematic. These things constitute what Popper terms the ‘background knowledge’. However, he stresses that the background knowledge is *not* knowledge in the sense of being conclusively established; it may be challenged at any time, especially if it is suspected that its uncritical acceptance may be responsible for difficulties which are subsequently encountered. Nevertheless, it is clearly not possible to question both the theory and the background knowledge at the same time (e.g. in conducting an experiment the scientist of necessity assumes that the apparatus used is in working order).

How then can one be certain that one is questioning the right thing? The Popperian answer is that we cannot have absolute certainty here, but repeated tests usually show where the trouble lies. Even observation statements, Popper maintains, are fallible, and science in his view is not a quest for certain knowledge, but an evolutionary process in which hypotheses or conjectures are imaginatively proposed and tested in order to explain facts or to solve problems. Popper emphasises both the importance of questioning the background knowledge when the need arises, and the significance of the fact that observation-statements are theory-laden, and hence fallible. For while falsifiability is simple as a logical principle, in practice it is exceedingly complicated - no single observation can ever be taken to falsify a theory, for there is always the possibility (a) that the observation itself is mistaken, or (b) that the assumed background knowledge is faulty or defective.

Popper was initially uneasy with the concept of truth, and in his earliest writings he avoided asserting that a theory which is corroborated is true - for clearly if every theory is an open-ended hypothesis, as he maintains, then *ipso facto* it has to be at least potentially false. For this reason Popper restricted himself to the contention that a theory which is falsified is false and is known to be such, and that a theory which replaces a falsified theory (because it has a higher empirical content than the latter, and explains what has falsified it) is a ‘better theory’ than its predecessor. However, he came to accept Tarski's reformulation of the correspondence theory of truth, and in *Conjectures and Refutations* (1963) he integrated the concepts of truth and content to frame the metalogical concept of ‘truthlikeness’ or ‘*verisimilitude*’. A ‘good’ scientific theory, Popper thus argued, has a higher level of verisimilitude than its rivals, and he explicated this concept by reference to the logical consequences of theories. A theory's content is the totality of its

logical consequences, which can be divided into two classes: there is the '*truth-content*' of a theory, which is the class of true propositions which may be derived from it, on the one hand, and the '*falsity-content*' of a theory, on the other hand, which is the class of the theory's false consequences (this latter class may of course be empty, and in the case of a theory which is true is necessarily empty).

Popper offered two methods of comparing theories in terms of verisimilitude, the qualitative and quantitative definitions. On the qualitative account, Popper asserted:

> Assuming that the truth-content and the falsity-content of two theories t_1 and t_2 are comparable, we can say that t_2 is more closely similar to the truth, or corresponds better to the facts, than t_1, if and only if either:
>
> (a) the truth-content but not the falsity-content of t_2 exceeds that of t_1, or
>
> (b) the falsity-content of t_1, but not its truth-content, exceeds that of t_2. (*Conjectures and Refutations*, 233).

Here, verisimilitude is defined in terms of subclass relationships: t_2 has a higher level of verisimilitude than t_1 if and only if their truth- and falsity-contents are comparable through subclass relationships, and *either* (a) t_2's truth-content includes t_1's and t_2's falsity-content, if it exists, is included in, or is the same as, t_1's, *or* (b) t_2's truth-content includes or is the same as t_1's and t_2's falsity-content, if it exists, is included in t_1's.

On the quantitative account, verisimilitude is defined by assigning quantities to contents, where the index of the content of a given theory is its logical improbability (given again that content and probability vary inversely). Formally, then, Popper defines the quantitative verisimilitude which a statement 'a' possesses by means of a formula:

$$Vs(a)=Ct_T(a) - Ct_F(a),$$

where Vs(a) represents the verisimilitude of 'a', $Ct_T(a)$ is a measure of the truth-content of 'a', and $Ct_F(a)$ is a measure of its falsity-content.

The utilisation of either method of computing verisimilitude shows, Popper held, that even if a theory t_2 with a higher content than a rival theory t_1 is subsequently falsified, it can still legitimately be regarded as a better theory than t_1, and 'better' is here now understood to mean t_2 is *closer to the truth* than t_1. Thus scientific progress involves, on this view, the abandonment of partially true, but falsified, theories, for theories with a higher level of verisimilitude, i.e., which approach more closely to the truth. In this way, verisimilitude allowed Popper to mitigate what many saw as the pessimism of an anti-inductivist

philosophy of science which held that most, if not all scientific theories are false, and that a true theory, even if discovered, could not be *known* to be such. With the introduction of the new concept, Popper was able to represent this as an essentially optimistic position in terms of which we can legitimately be said to have reason to believe that science makes progress towards the truth through the falsification and corroboration of theories. Scientific progress, in other words, could now be represented as progress *towards* the truth, and experimental corroboration could be seen an *indicator* of verisimilitude.

However, in the 1970's a series of papers published by researchers such as Miller, Tichý, and Grünbaum in particular revealed fundamental defects in Popper's formal definitions of verisimilitude. The significance of this work was that verisimilitude is largely important in Popper's system because of its application to theories which are known to be *false*. In this connection, Popper had written:

> Ultimately, the idea of verisimilitude is most important in cases where we know that we have to work with theories which are *at best* approximations—that is to say, theories of which we know that they cannot be true. (This is often the case in the social sciences). In these cases we can still speak of better or worse approximations to the truth (and we therefore do not need to interpret these cases in an instrumentalist sense). (*Conjectures and Refutations*, 235).

For these reasons, the deficiencies discovered by the critics in Popper's formal definitions were seen by many as devastating, precisely because the most significant of these related to the levels of verisimilitude of *false* theories. In 1974, Miller and Tichý, working independently of each other, demonstrated that the conditions specified by Popper in his accounts of both qualitative and quantitative verisimilitude for comparing the truth- and falsity-contents of theories can be satisfied only when the theories are *true*. In the crucially important case of false theories, however, Popper's definitions are formally defective. For while Popper had believed that verisimilitude intersected positively with his account of corroboration, in the sense that he viewed an improbable theory which had withstood critical testing as one the truth-content of which is great relative to rival theories, while its falsity-content (if it exists) would be relatively low, Miller and Tichý proved, on the contrary, that in the case of a false theory $t_2$which has excess content over a rival theory false $t_1$both the truth-content *and* the falsity-content of $t_2$will exceed that of t_1. With respect to theories which are false, therefore, Popper's conditions for comparing levels of verisimilitude, whether in quantitative and qualitative terms, can never be met.

Commentators on Popper, with few exceptions, had initially attached little importance to his theory of verisimilitude. However, after the failure of Popper's definitions in 1974, some critics came to see it as central to his philosophy of science, and consequentially held that the whole edifice of the latter had been subverted. For his part, Popper's response was two-fold. In the first place, while acknowledging the deficiencies in his own formal account ("my main mistake was my failure to see at once that … if the content of a false statement *a* exceeds that of a statement *b*, then the truth-content of *a* exceeds the truth-content of *b*, and the same holds of their falsity-contents", *Objective Knowledge*, 371), Popper argued that "I do think that we should not conclude from the failure of my attempts to solve the problem [of defining verisimilitude] that the problem cannot be solved" (*Objective Knowledge*, 372), a point of view

which was to precipitate more than two decades of important technical research in this field. At another, more fundamental level, he moved the task of formally defining the concept from centre-stage in his philosophy of science, by protesting that he had never intended to imply "that degrees of verisimilitude ... can ever be numerically determined, except in certain limiting cases" (*Objective Knowledge*, 59), and arguing instead that the chief value of the concept is heuristic and intuitive, in which the absence of an adequate formal definition is not an insuperable impediment to its utilisation in the actual appraisal of theories relativised to problems in which we have an interest. The thrust of the latter strategy seems to many to genuinely reflect the significance of the concept of verisimilitude in Popper's system, but it has not satisfied all of his critics.
[Return to Section Headings]

Social and Political Thought -- The Critique of Historicism and Holism

Given Popper's personal history and background, it is hardly surprising that he developed a deep and abiding interest in social and political philosophy. However, it is worth emphasising that his angle of approach to these fields is through a consideration of the nature of the social sciences which seek to describe and explicate them systematically, particularly history. It is in this context that he offers an account of the nature of scientific prediction, which in turn allows him a point of departure for his attack upon totalitarianism and all its intellectual supports, especially holism and historicism. In this context holism is to be understood as the view that human social groupings are greater than the sum of their members, that such groupings are 'organic' entities in their own right, that they act on their human members and shape their destinies, and that they are subject to their own independent laws of development. Historicism, which is closely associated with holism, is the belief that history develops inexorably and necessarily according to certain principles or rules towards a determinate end (as for example in the dialectic of Hegel, which was adopted and implemented by Marx). The link between holism and historicism is that the holist believes that individuals are essentially formed by the social groupings to which they belong, while the historicist - who is usually also a holist - holds that we can understand such a social grouping only in terms of the internal principles which determine its development.

These beliefs lead to what Popper calls 'The Historicist Doctrine of the Social Sciences', the views (a) that the principal task of the social sciences is to make predictions about the social and political development of man, and (b) that the task of politics, once the key predictions have been made, is, in Marx's words, to lessen the 'birth pangs' of future social and political developments. Popper thinks that this view of the social sciences is both theoretically misconceived (in the sense of being based upon a view of natural science and its methodology which is totally wrong), and socially dangerous, as it leads inevitably to totalitarianism and authoritarianism - to centralised governmental control of the individual and the attempted imposition of large-scale social planning. Against this Popper strongly advances the view that any human social grouping is no more (or less) than the sum of its individual members, that what happens in history is the (largely unplanned and unforeseeable) result of the actions of such

individuals, and that large scale social planning to an antecedently conceived blueprint is inherently misconceived - and inevitably disastrous - precisely because human actions have consequences which cannot be foreseen. Popper, then, is an historical *indeterminist*, insofar as he holds that history does not evolve in accordance with intrinsic laws or principles, that in the absence of such laws and principles unconditional prediction in the social sciences is an impossibility, and that there is no such thing as historical necessity.

The link between Popper's theory of knowledge and his social philosophy is his fallibilism - just as we make theoretical progress in science by deliberately subjecting our theories to critical scrutiny, and abandoning those which have been falsified, so too, Popper holds, the critical spirit can and should be sustained at the social level. More specifically, the open society can be brought about only if it is possible for the individual citizen to evaluate critically the consequences of the implementation of government policies, which can then be abandoned or modified in the light of such critical scrutiny - in such a society, the rights of the individual to criticise administrative policies will be formally safeguarded and upheld, undesirable policies will be eliminated in a manner analogous to the elimination of falsified scientific theories, and differences between people on social policy will be resolved by critical discussion and argument rather than by force. The open society as thus conceived of by Popper may be defined as ‘an association of free individuals respecting each other's rights within the framework of mutual protection supplied by the state, and achieving, through the making of responsible, rational decisions, a growing measure of humane and enlightened life’ (Levinson, R.B. *In Defense of Plato*, 17). As such, Popper holds, it is not a utopian ideal, but an empirically realised form of social organisation which, he argues, is in every respect superior to its (real or potential) totalitarian rivals. But he does not engage in a moral defence of the ideology of liberalism; rather his strategy is the much deeper one of showing that totalitarianism is typically based upon historicist and holist presuppositions, and of demonstrating that these presuppositions are fundamentally incoherent.
[Return to Section Headings]

Scientific Knowledge, History, and Prediction

At a very general level, Popper argues that historicism and holism have their origins in what he terms ‘one of the oldest dreams of mankind - the dream of prophecy, the idea that we can know what the future has in store for us, and that we can profit from such knowledge by adjusting our policy to it.’ (*Conjectures and Refutations*, 338). This dream was given further impetus, he speculates, by the emergence of a genuine predictive capability regarding such events as solar and lunar eclipses at an early stage in human civilisation, which has of course become increasingly refined with the development of the natural sciences and their concomitant technologies. The kind of reasoning which has made, and continues to make, historicism plausible may, on this account, be reconstructed as follows: if the application of the laws of the natural sciences can lead to the successful prediction of such future events as eclipses, then surely it is reasonable to infer that knowledge of the laws of history as yielded by a social science or sciences (assuming that such laws exist) would lead to the successful prediction of such future social phenomena as revolutions? Why should it be possible to predict an eclipse, but not a revolution? Why can we not conceive of a social science which could and would function as the

theoretical natural sciences function, and yield precise unconditional predictions in the appropriate sphere of application? These are amongst the questions which Popper seeks to answer, and in doing so, to show that they are based upon a series of misconceptions about the nature of science, and about the relationship between scientific laws and scientific prediction.

His first argument may be summarised as follows: in relation to the critically important concept of prediction, Popper makes a distinction between what he terms ‘conditional scientific predictions’, which have the form ‘If X takes place, then Y will take place’, and ‘unconditional scientific prophecies’, which have the form ‘Y will take place’. Contrary to popular belief, it is the former rather than the latter which are typical of the natural sciences, which means that typically prediction in natural science is conditional and limited in scope - it takes the form of hypothetical assertions stating that certain specified changes will come about if particular specified events antecedently take place. This is not to deny that ‘unconditional scientific prophecies’, such as the prediction of eclipses, for example, do take place in science, and that the theoretical natural sciences make them possible. However, Popper argues that (a) these unconditional prophecies are not *characteristic* of the natural sciences, and (b) that the mechanism whereby they occur, in the very limited way in which they do, is not understood by the historicist.

What is the mechanism which makes unconditional scientific prophecies possible? The answer is that such prophecies can sometimes be derived from a combination of conditional predictions (themselves derived from scientific laws) *and* existential statements specifying that the conditions in relation to the system being investigated are fulfilled. Schematically, this can be represented as follows:

[C.P. + E.S.]=U.P.

where C.P.=Conditional Prediction; E.S.=Existential Statement; U.P.=Unconditional Prophecy. The most common examples of unconditional scientific prophecies in science relate to the prediction of such phenomena as lunar and solar eclipses and comets.

Given, then, that this is the mechanism which generates unconditional scientific prophecies, Popper makes two related claims about historicism: (a) That the historicist does not in fact derive his unconditional scientific prophecies in this manner from conditional predictions, and (b) the historicist *cannot* do so because long-term unconditional scientific prophecies can be derived from conditional predictions only if they apply to systems which are well-isolated, stationary, and recurrent (like our solar system). Such systems are quite rare in nature, and human society is most emphatically not one of them.

This, then, Popper argues, is the reason why it is a fundamental mistake for the historicist to take the unconditional scientific prophecies of eclipses as being typical and characteristic of the predictions of natural science - in fact such predictions are possible only because our solar system is a stationary and repetitive system which is isolated from other such systems by immense expanses of empty space. The solar system aside, there are very few such systems around for scientific investigation - most of the others are confined to the field of biology, where unconditional prophecies about the life-cycles of organisms are made possible by the existence of precisely the same factors. Thus one of the fallacies

committed by the historicist is to take the (relatively rare) instances of unconditional prophecies in the natural science as constituting the essence of what scientific prediction is, to fail to see that such prophecies apply only to systems which are isolated, stationary, and repetitive, and to seek to apply the method of scientific prophecy to human society and human history. The latter, of course, is *not* an isolated system (in fact it's not a system at all), it is constantly changing, and it continually undergoes rapid, non-repetitive development. In the most fundamental sense possible, every event in human history is discrete, novel, quite unique, and ontologically distinct from every other historical event. For this reason, it is impossible in principle that unconditional scientific prophecies could be made in relation to human history - the idea that the successful unconditional prediction of eclipses provides us with reasonable grounds for the hope of successful unconditional prediction regarding the evolution of human history turns out to be based upon a gross misconception, and is quite false. As Popper himself concludes, "The fact that we predict eclipses does not, therefore, provide a valid reason for expecting that we can predict revolutions." (*Conjectures and Refutations*, 340).
[Return to Section Headings]

Immutable Laws and Contingent Trends

This argument is one of the strongest that has ever been brought against historicism, cutting, as it does, right to the heart of one of its main theoretical presuppositions. However, it is not Popper's only argument against it. An additional mistake which he detects in historicism is the failure of the historicist to distinguish between scientific *laws* and *trends*, which is also frequently accompanied by a simple logical fallacy. The fallacy is that of inferring from the fact that our understanding of any (past) historical event - such as, for example, the French Revolution - is in direct proportion to our knowledge of the antecedent conditions which led to that event, that knowledge of all the antecedent conditions of some future event is possible, and that such knowledge would make that future event precisely predictable. For the truth is that the number of factors which predate and lead to the occurrence of any event, past, present, or future, is indefinitely large, and therefore knowledge of all of these factors is impossible, even in principle. What gives rise to the fallacy is the manner in which the historian (necessarily) selectively isolates a finite number of the antecedent conditions of some past event as being of particular importance, which are then somewhat misleadingly termed ‘the causes’ of that event, when in fact what this means is that they are the specific conditions which a particular historian or group of historians take to be more *relevant* than any other of the indefinitely large number of such conditions (for this reason, most historical debates range over the question as to whether the conditions thus specified are the *right* ones). While this kind of selectivity may be justifiable in relation to the treatment of any past event, it has no basis whatsoever in relation to the future - if we now select, as Marx did, the ‘relevant’ antecedent conditions for some future event, the likelihood is that we will select wrongly.

The historicist's failure to distinguish between scientific laws and trends is equally destructive of his cause. This failure makes him think it possible to explain change by discovering trends running through past history, and to anticipate and predict future occurrences on the basis of such observations. Here Popper points out that there is a critical difference between a trend and a scientific law, the failure to observe which is fatal. For a scientific law is universal in form, while a trend can be expressed only as a

singular existential statement. This logical difference is crucial because unconditional predictions, as we have already seen, can be based only upon conditional ones, which themselves must be derived from scientific laws. Neither conditional nor unconditional predictions can be based upon trends, because these may change or be reversed with a change in the conditions which gave rise to them in the first instance. As Popper puts it, there can be no doubt that "the habit of confusing trends with laws, together with the intuitive observation of trends such as technical progress, inspired the central doctrines of ... historicism." (*The Poverty of Historicism*, 116). Popper does not, of course, dispute the existence of trends, nor does he deny that the observation of trends can be of practical utility value - but the essential point is that a trend is something which *itself* ultimately stands in need of scientific explanation, and it cannot therefore function as the frame of reference in terms of which anything else can be scientifically explained or predicted.

A point which connects with this has to do with the role which the evolution of human knowledge has played in the historical development of human society. It is incontestable that, as Marx himself observed, there has been a causal link between the two, in the sense that advances in scientific and technological knowledge have given rise to widespread global changes in patterns of human social organisation and social interaction, which in turn have led to social structures (e.g. educational systems) which further growth in human knowledge. In short, the evolution of human history has been strongly influenced by *the growth of human knowledge*, and it is extremely likely that this will continue to be the case - all the empirical evidence suggests that the link between the two is progressively consolidating. However, this gives rise to further problems for the historicist. In the first place, the statement that ‘if there is such a thing as growing human knowledge, then we cannot anticipate today what we shall know only tomorrow’ is, Popper holds, intuitively highly plausible. Moreover, he argues, it is logically demonstrable by a consideration of the implications of the fact that no scientific predictor, human or otherwise, can possibly predict, by scientific methods, its own future results. From this it follows, he holds, that ‘no society can predict, scientifically, its own future states of knowledge’. (*The Poverty of Historicism*, vii). Thus, while the future evolution of human history is extremely likely to be influenced by new developments in human knowledge, as it always has in the past, we cannot now scientifically determine what such knowledge will be. From this it follows that if the future holds any new discoveries or any new developments in the growth of our knowledge (and given the fallible nature of the latter, it is inconceivable that it does not), then it is impossible for us to predict them now, and it is therefore impossible for us to predict the future development of human history now, given that the latter will, at least in part, be determined by the future growth of our knowledge. Thus once again historicism collapses - the dream of a theoretical, predictive science of history is unrealisable, because it is an impossible dream.

Popper's arguments against holism, and in particular his arguments against the propriety of large-scale planning of social structures, are interconnected with his demonstration of the logical shortcomings of the presuppositions of historicism. Such planning (which actually took place, of course, in the USSR, in China, and in Cambodia, for example, under totalitarian regimes which accepted forms of historicism and holism), Popper points out, is necessarily structured in the light of the predictions which have been made about future history on the basis of the so-called ‘laws’ which historicists such as Marx and Mao claimed to have discovered in relation to human history. Accordingly, recognition that there are no such

laws, and that unconditional predictions about future history are based, at best, upon nothing more substantial than the observation of contingent trends, shows that, from a purely theoretical as well as a practical point of view, large-scale social planning is indeed a recipe for disaster. In summary, unconditional large-scale planning for the future is theoretically as well as practically misguided, because, again, part of what we are planning for is our future knowledge, and our future knowledge is not something which we can in principle now possess - we cannot adequately plan for unexpected advances in our future knowledge, or for the effects which such advances will have upon society as a whole. The acceptance of historical indeterminism, then, as the only philosophy of history which is commensurate with a proper understanding of the nature of scientific knowledge, fatally undermines both historicism and holism.

Popper's critique of both historicism and holism is balanced, on the positive side, by his strong defence of the open society, the view, again, that a society is equivalent to the sum of its members, that the actions of the members of society serve to fashion and to shape it, not conversely, and that the social consequences of intentional actions are very often, and very largely, unintentional. This is why Popper himself advocates what he (rather unfortunately) terms ‘piecemeal social engineering’ as the central mechanism for social planning - for in utilising this mechanism intentional actions are directed to the achievement of one specific goal at a time, which makes it possible to monitor the situation to determine whether adverse unintended effects of intentional actions occur, in order to correct and readjust when this proves necessary. This, of course, parallels precisely the critical testing of theories in scientific investigation. This approach to social planning (which is explicitly based upon the premise that we do not, because we cannot, know what the future will be like) encourages attempts to put right what is problematic in society - generally-acknowledged social ills - rather than attempts to impose some preconceived idea of the ‘good’ upon society as a whole. For this reason, in a genuinely open society piecemeal social engineering goes hand-in-hand for Popper with *negative* utilitarianism (the attempt to minimise the amount of misery, rather than, as with positive utilitarianism, the attempt to maximise the amount of happiness). The state, he holds, should concern itself with the task of progressively formulating and implementing policies designed to deal with the social problems which actually confront it, with the goal of eliminating human misery and suffering to the highest possible degree. The positive task of increasing social and personal happiness, by contrast, can and should be should be left to individual citizens (who may, of course, act collectively to this end), who, unlike the state, have at least a chance of achieving this goal, but who in a free society are rarely in a position to systematically subvert the rights of others in the pursuit of idealised objectives. Thus in the final analysis for Popper the activity of problem-solving is as definitive of our humanity at the level of social and political organisation as it is at the level of science, and it is this key insight which unifies and integrates the broad spectrum of his thought.
[Return to Section Headings]

Critical Evaluation

While it cannot be said that Popper was a modest man, he took criticism of his theories very seriously, and spent much of his time in his later years endeavouring to show that such criticisms were either based

upon misunderstandings, or that his theories could, without loss of integrity, be made compatible with new and important insights (such as Kuhn's distinction between normal and revolutionary science). The following is a summary of some of the main criticisms which he has had to address.

1. Popper professes to be anti-conventionalist, and his commitment to the correspondence theory of truth places him firmly within the realist's camp. Yet, following Kant, he strongly repudiates the positivist/empiricist view that basic statements (i.e. present-tense observation statements about sense-data) are infallible, and argues convincingly that such basic statements are not mere ‘reports’ of passively registered sensations. Rather they are descriptions of what is observed as interpreted by the observer with reference to a determinate theoretical framework. This is why Popper repeatedly emphasises that basic statements are not infallible, and it indicates what he means when he says that they are ‘theory laden’ - perception itself is an active process, in which the mind assimilates data by reference to an assumed theoretical backdrop. He accordingly asserts that basic statements themselves are open-ended hypotheses: they have a certain causal relationship with experience, but they are not *determined* by experience, and they cannot be verified or confirmed by experience. However, this poses a difficulty regarding the consistency of Popper's theory: if a theory X is to be genuinely testable (and so scientific) it must be possible to determine whether or not the basic propositions which would, if true, falsify it, are *actually* true or false (i.e. whether its potential falsifiers are actual falsifiers). But how can this be known, if such basic statements cannot be verified by experience? Popper's answer is that ‘basic statements are not justifiable by our immediate experiences, but are .... accepted by an act, a free decision’. (*Logic of Scientific Discovery*, 109). However, and notwithstanding Popper's claims to the contrary, this itself seems to be a refined form of conventionalism - it implies that it is almost entirely an arbitrary matter whether it is accepted that a potential falsifier is an actual one, and consequently that the falsification of a theory is itself the function of a ‘free’ and arbitrary act. It also seems very difficult to reconcile this with Popper's view that science progressively moves closer to the truth, conceived of in terms of the correspondence theory, for this kind of conventionalism is inimical to this (classical) conception of truth.

2. As Lakatos has pointed out, Popper's theory of demarcation hinges quite fundamentally on the assumption that there are such things as critical tests, which either conclusively falsify a theory, or give it a strong measure of corroboration. Popper himself is fond of citing, as an example of such a critical test, the resolution, by Adams and Leverrier, of the problem which the anomalous orbit of Uranus posed for nineteenth century astronomers. Both men independently came to the conclusion that, assuming Newtonian mechanics to be precisely correct, the observed divergence in the elliptical orbit of Uranus could be explained if the existence of a seventh, as yet unobserved outer planet was posited. Further, they were able, again within the framework of Newtonian mechanics, to calculate the precise position of the ‘new’ planet. Thus when subsequent research by Galle at the Berlin observatory revealed that such a planet (Neptune) did in fact exist, and was situated precisely where Adams and Leverrier had calculated, this was hailed as by all and sundry as a magnificent triumph for Newtonian physics: in Popperian terms, Newton's theory had been subjected to a critical test, and had passed with flying colours. Popper himself refers to this strong corroboration of Newtonian physics as ‘the most startling and convincing success of any human intellectual achievement’. Yet Lakatos flatly denies that there are critical tests, in the Popperian sense, in science, and argues the point convincingly by turning the above example of an alleged critical test on its head. What, he asks, would have happened if Galle had *not* found the planet

Neptune? Would Newtonian physics have been abandoned, or would Newton's theory have been falsified? The answer is clearly not, for Galle's failure could have been attributed to any number of causes other than the falsity of Newtonian physics (e.g. the interference of the earth's atmosphere with the telescope, the existence of an asteroid belt which hides the new planet from the earth, etc). The point here is that the ‘falsification/corroboration’ disjunction offered by Popper is far too logically neat: non-corroboration is *not necessarily* falsification, and falsification of a high-level scientific theory is never brought about by an isolated observation or set of observations. Such theories are, it is now generally accepted, highly resistant to falsification. They are falsified, if at all, Lakatos argues, not by Popperian critical tests, but rather within the elaborate context of the research programmes associated with them gradually grinding to a halt, with the result that an ever-widening gap opens up between the facts to be explained, and the research programmes themselves. (Lakatos, I. *The Methodology of Scientific Research Programmes*, passim). Popper's distinction between the logic of falsifiability and its applied methodology does not in the end do full justice to the fact that all high-level theories grow and live despite the existence of anomalies (i.e. events/phenomena which are incompatible with the theories). The existence of such anomalies is not usually taken by the working scientist as an indication that the theory in question is false; on the contrary, he will usually, and necessarily, assume that the auxiliary hypotheses which are associated with the theory can be modified to incorporate, and explain, existing anomalies.

3. Scientific laws are expressed by universal statements (i.e. they take the logical form ‘All A's are X’, or some equivalent) which are therefore concealed conditionals - they have to be understood as hypothetical statements asserting what would be the case under certain ideal conditions. In themselves they are not *existential* in nature. Thus ‘All A's are X’ means ‘If anything is an A, then it is X’. Since scientific laws are non-existential in nature, they logically cannot imply any basic statements, since the latter are explicitly existential. The question arises, then, as to how any basic statement can falsify a scientific law, given that basic statements are not deducible from scientific laws in themselves? Popper answers that scientific laws are always taken in *conjunction with* statements outlining the ‘initial conditions’ of the system under investigation; these latter, which are singular existential statements, do, when combined with the scientific law, yield hard and fast implications. Thus, the law ‘All A's are X’, together with the initial condition statement ‘There is an A at Y’, yields the implication ‘The A at Y is X’, which, if false, falsifies the original law.

This reply is adequate only if it is true, as Popper assumes, that *singular* existential statements will always do the work of bridging the gap between a universal theory and a prediction. Hilary Putnam in particular has argued that this assumption is false, in that in some cases at least the statements required to bridge this gap (which he calls ‘auxiliary hypotheses’) are general rather than particular, and consequently that when the prediction turns out to be false we have no way of knowing whether this is due to the falsity of the scientific law *or* the falsity of the auxiliary hypotheses. The working scientist, Putnam argues, always initially assumes that it is the latter, which shows not only that scientific laws are, *contra* Popper, highly resistant to falsification, but also why they are so highly resistant to falsification.

Popper's final position is that he acknowledges that it is impossible to discriminate science from non-science on the basis of the falsifiability of the scientific statements *alone*; he recognizes that scientific theories are predictive, and consequently prohibitive, *only* when taken in conjunction with auxiliary

hypotheses, and he also recognizes that readjustment or modification of the latter is an integral part of scientific practice. Hence his final concern is to outline conditions which indicate when such modification is genuinely scientific, and when it is merely *ad hoc*. This is itself clearly a major alteration in his position, and arguably represents a substantial retraction on his part: Marxism can no longer be dismissed as ‘unscientific’ simply because its advocates preserved the theory from falsification by modifying it (for in general terms, such a procedure, it now transpires, is perfectly respectable scientific practice). It is now condemned as unscientific by Popper because the only *rationale* for the modifications which were made to the original theory was to ensure that it evaded falsification, and so such modifications were *ad hoc*, rather than scientific. This contention - though not at all implausible - has, to hostile eyes, a somewhat contrived air about it, and is unlikely to worry the convinced Marxist. On the other hand, the shift in Popper's own basic position is taken by some critics as an indicator that falsificationism, for all its apparent merits, fares no better in the final analysis than verificationism.
[Return to Section Headings]

Bibliography

Works By Popper

- *Logik der Forschung*. Julius Springer Verlag, Vienna, 1935.
- *The Open Society and Its Enemies*. (2 Vols). Routledge, London, 1945.
- *The Logic of Scientific Discovery*. (translation of *Logik der Forschung*). Hutchinson, London, 1959.
- *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, London, 1963.
- *The Poverty of Historicism* (2nd. ed). Routledge, London, 1961.
- *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford, 1972.
- *Unended Quest; An Intellectual Autobiography*. Fontana, London, 1976.
- ‘A Note on Verisimilitude’, *The British Journal for the Philosophy of Science* **27**, 1976, 147-159.
- *The Self and Its Brain: An Argument for Interactionism* (with J.C. Eccles). Springer International, London, 1977.
- *The Open Universe: An Argument for Indeterminism*. (ed. W.W. Bartley 111). Hutchinson, London, 1982.
- *Realism and the Aim of Science*, Hutchinson, London, 1982.
- *The Myth of the Framework: In Defence of Science and Rationality*. Routledge, London, 1994.
- *Knowledge and the Mind-Body Problem: In Defence of Interactionism*. (ed. M.A. Notturno). Routledge, London, 1994.

Works by Other Authors

- Ackermann, R. *The Philosophy of Karl Popper*. University of Massachusetts Press, Amherst, 1976.
- Bambrough, R. (ed). *Plato, Popper, and Politics: Some Contributions to a Modern Controversy*.

Barnes and Noble, New York, 1967.

- Baudoin, J. *Karl Popper*. PUF, Paris, 1989.
- Brink, C. & Heidema, J. 'A Verisimilar Ordering of Theories Phrased in a Propositional Language', *British Journal for the Philosophy of Science* **38**, 1987, 533-549.
- Brink, C. 'Verisimilitude: Views and Reviews', *History and Philosophy of Logic* **10**, 1989, 181-201.
- Brink, C. & Britz, K. 'Computing Verisimilitude', *Notre Dame Journal of Formal Logic* **36**, 1, 1995, 31-43.
- Bunge, M. (ed). *The Critical Approach to Science and Philosophy*. The Free Press, London & New York, 1964.
- Burke, T.E. *The Philosophy of Popper*. Manchester University Press, Manchester, 1983.
- Carr, E.H. *What is History?* Macmillan, London, 1962.
- Cornforth, M. *The Open Philosophy and the Open Society: A Reply to Dr. Popper's Refutations of Marxism*. Lawrence & Wishart, London, 1968.
- Corvi, R. *An Introduction to the Thought of Karl Popper*. (trans. P. Camiller). Routledge, London & New York, 1997.
- Currie, G. & Musgrave, A. (eds). *Popper and the Human Sciences*. Nijhoff, Dordrecht, 1985.
- Feyerabend, P. *Against Method*. New Left Books, London, 1975.
- Grünbaum, A. 'Is the Method of Bold Conjectures and Attempted Refutations Justifiably the Method of Science?', *British Journal for the Philosophy of Science* **27**, 1976, 105-136.
- Hume, D. *A Treatise of Human Nature*, in *The Philosophical Works* (ed. T.H. Green & T.H. Grose), 4 vols (reprint of 1886 edition). Scientia Verlag Aalen, Darmstadt, 1964.
- Jacobs, S. *Science and British Liberalism: Locke, Bentham, Mill and Popper*. Avebury, Aldershot, 1991.
- James, R. *Return to Reason: Popper's Thought in Public Life*. Open Books, Shepton Mallet, 1980.
- Johannson, I. *A Critique of Karl Popper's Methodology*. Scandinavian University Books, Stockholm, 1975.
- Kekes, J. 'Popper in Perspective', *Metaphilosophy* 8 (1977), pp. 36-61.
- Keuth, H. 'Verisimilitude or the Approach to the Whole Truth', *Philosophy of Science* 1976, 311-336.
- Kuipers, T. A. F. 'Approaching Descriptive and Theoretical Truth', *Erkenntnis* **18**, 1982, 343-378.
- Kuipers, T. A. F., (ed). *What is Closer-to-the-Truth?*, Rodopi, Amsterdam, 1987.
- Kuhn, T.S. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago and London, 1962.
- Lakatos, I. 'Falsification and the Methodology of Scientific Research Programmes', in Lakatos, I & Musgrove, A. (eds). *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, 1970.
- Lakatos, I. *The Methodology of Scientific Research Programmes,* (ed. J. Worrall & G. Currie). Cambridge University Press, 1978.
- Lakatos, I & Musgrove, A. (eds). *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, 1970.
- Laudan, L. *Progress and Its Problems: Towards a Theory of Scientific Growth*. Routledge,

London, 1977.

- Levinson, P. (ed.). *In Pursuit of Truth. Essays in Honour of Karl Popper on the Occasion of his 80th Birthday*. Humanities Press, Atlantic Highlands, 1982.
- Levinson, R.B. *In Defense of Plato*. Cambridge University Press, Cambridge, 1957.
- Magee, B. *Popper*. Fontana, London, 1977.
- Mellor, D.H. 'The Popper Phenomenon', *Philosophy* 52 (1977), pp. 195-202.
- Miller, D. 'On the Comparison of False Theories by their Bases', *The British Journal for the Philosophy of Science* **25**, 1974, 178-188.
- Miller, D. 'Popper's Qualitative Theory of Verisimilitude', *The British Journal for the Philosophy of Science* **25**, 1974, 166-177.
- Miller, D. *Critical Rationalism: A Restatement and Defence*, Open Court, Chicago, 1994.
- Munz, P. *Our Knowledge of the Growth of Knowledge: Popper or Wittgenstein?* Routledge, London, 1985.
- Naydler, J. 'The Poverty of Popperism', *Thomist* 46 (1982), pp. 92-107.
- Niiniluoto, I. *Truthlikeness*, Reidel, Dordrecht, 1987.
- Oddie, G. *Likeness to Truth*, Reidel, Dordrecht, 1986.
- O'Hear, A. *Karl Popper*. Routledge, London, 1980.
- Putnam, H. 'The Corroboration of Theories', in *The Philosophy of Karl Popper* (ed. P.A. Schilpp). Open Court Press, La Salle, 1974.
- Quinton, A. 'Popper, Karl Raimund', in *Encyclopedia of Philosophy*, vol. 6 (ed. P. Edwards). Collier Macmillan, New York, 1967.
- Radnitzky, G. & Andersson, G. (eds). *Progress and Rationality in Science*. Reidel, Dordrecht, 1978.
- Radnitzky, G. & Bartley, W.W. (eds). *Evolutionary Epistemology, Rationality, and the Sociology of Knowledge*. Open Court, La Salle, 1987.
- Shearmur, J. *Political Thought of Karl Popper*. London & New York: Routledge, 1996.
- Simkin, C. *Popper's Views on Natural and Social Science*. Brill, Leiden, 1993.
- Stokes, G. *Popper: Philosophy, Politics and Scientific Method*. Polity Press, 1998.
- Stove, D. *Popper and After: Four Modern Irrationalists*. Pergamon Press, Oxford, 1982.
- Schilpp, P.A. (ed) *The Philosophy of Karl Popper*. (2 Vols). Open Court Press, La Salle, 1974.
- Tichý, P. 'On Popper's Definitions of Verisimilitude', *The British Journal for the Philosophy of Science* **25**, 1974, 155-160
- Tichý, P. ' Verisimilitude Revisited', Synthèse **38**, 1978, 175-196.
- Vetter, H. 'A New Concept of Verisimilitude', *Theory and Decision* **8**, 1977, 369-375.
- Watkins, J. *Science and Scepticism*, Princeton University Press and Hutchinson, Princeton and London, 1984.
- Watkins, J. 'Popperian Ideas on Progress and Rationality in Science', *The Critical Rationalist*, Vol. 2 No. 2, June 1997.
- Wilkins, B.T. *Has History Any Meaning? A Critique of Popper's Philosophy of History*. Hassocks/Cornell University Press/The Harvester Press, Ithaca, 1978.
- Williams, D.E. *Truth, Hope and Power: The Thought of Karl Popper*. University of Toronto Press, Toronto, 1989.
- Wuketits, F.M. *Concepts and Approaches in Evolutionary Epistemology: Towards an*

Evolutionary Theory of Knowledge. Reidel, Dordrecht, 1984.

[Return to Section Headings]

Other Internet Resources

- The Karl Popper Web
- The Japan Popper Society
- Institute Vienna Circle
- "Popper, Karl Raimund", by Peter Munz in the Dictionary of New Zealand Biography

[Return to Section Headings]

Related Entries

confirmation | Feyerabend, Paul | Hume, David | induction: problem of | Kuhn, Thomas | Lakatos, Imre | science, philosophy of | truthlikeness | Vienna Circle

Stephen Thornton
stephen.thornton@mic.ul.ie

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 13, 1997
Content last modified: April 15, 2002

Truthlikeness

While *truth* is the aim of inquiry, some falsehoods seem to realize this aim better than others. Some truths better realize the aim than other truths. And perhaps even some falsehoods realize the aim better than some truths do. The dichotomy of the class of propositions into truths and falsehoods should thus be supplemented with a more fine-grained ordering -- one which classifies propositions according to their *closeness* to the truth, their degree of truthlikeness or verisimilitude. The problem of truthlikeness is to give an adequate account of the concept and to explore its logical properties and its applications to epistemology and methodology.

In §1 we will examine the basic assumptions which generate the problem of truthlikeness, which in part explain why the problem emerged when it did. Attempted solutions to the problem quickly proliferated, but they can all be gathered together into two broad lines of attack. The first, the content approach (§2), was initiated by Popper in his ground-breaking work. However, because it treats truthlikeness as a function of just two variables, neither Popper's original proposals, nor subsequent attempts to elaborate them, can fully capture the richness of the concept. The second, the similarity approach (§3), takes the *likeness* in *truthlikeness* seriously Although it promises to catch more of the complexity of the concept than does the content approach, it faces two serious problems: whether the approach can be suitably generalized to complex examples (§5), and whether it can be developed in a way that is translation invariant (§6).

---

1. The Problem

Truth, perhaps even more than beauty and goodness, has been the target of an extraordinary amount of philosophical dissection and speculation. This is unsurprising. After all, truth is the primary aim of all inquiry and a necessary condition of knowledge. And yet, as redundancy theorists have emphasized, there is something disarmingly simple about truth. That *the number of planets is ten* is true just in case, well, … the number of planets is ten. By comparison with truth, the more complex, and perhaps more interesting, concept of truthlikeness has only recently become the subject of serious investigation. The proposition *the number of planets is ten* may be false, but closer to the truth than the proposition that *the number of planets is ten thousand*. Investigation into truthlikeness began with a tiny trickle of activity in the early nineteen sixties; became something of a torrent from the mid seventies until the late eighties; and is now a steady, albeit diminished, current.

Truthlikeness is a relative latecomer to the philosophical scene largely because it wasn't until the latter half of the twentieth century that mainstream philosophers gave up on the Cartesian goal of infallible knowledge. The idea that we are quite possibily, even probably, mistaken in our most cherished beliefs, that they might well be just *false*, was mostly considered tantamount to capitulation to the skeptic. By the middle of the twentieth century, however, it was clear that natural science postulated a very odd world behind the phenomena, one rather remote from our everyday experience, one which renders many of our commonsense beliefs, as well as previous scientific theories, strictly speaking, false. Further, the increasingly rapid turnover of scientific theories suggested that, far from being established as certain, they are ever vulnerable to refutation, and typically are eventually refuted, to be replaced by some new theory. Taking the dismal view, the history of inquiry is a history of theories shown to be false, replaced by other theories awaiting their turn at the guillotine.

Realism affirms that the primary aim of inquiry is the truth of some matter. Epistemic *optimism* affirms that the history of inquiry is one of progress with respect to its primary aim. But *fallibilism* affirms that, typically, our theories are false or very likely to be false, and when shown to be false they are replaced by other false theories. To combine all three ideas, we must affirm that some false propositions better realize the goal of truth -- are closer to the truth -- than others. So the optimistic realist who has discarded infallibilism has a problem -- the problem of truthlikeness.

While a multitude of apparently different solutions to the problem have been proposed, it is now standard to classify them into two main approaches -- the content approach and the likeness approach.

2. The First Assay: the Content Approach

Sir Karl Popper was the first philosopher to take the problem of truthlikeness seriously enough to make an assay on it. This is not surprising, since Popper was also the first prominent realist to embrace a radical fallibilism about science while trumpeting the epistemic superiority of the enterprise.

According to Popper, Hume had shown not only that we can't verify an interesting theory, we can't even render it more probable. Luckily, there is an asymmetry between verification and falsification. While no

finite amount of data can verify or probabilify an interesting scientific theory, they can falsify the theory.. According to Popper, it is the falsifiability of a theory which makes it scientific. In his early work, he implied that the only kind of progress an inquiry can make consists in falsification of theories. This is a little depressing, to say the least. What it lacks is the idea that a succession of falsehoods can constitute genuine cognitive progress. Perhaps this is why, for many years after first publishing these ideas in his 1934 *Logik der Forschung* Popper received a pretty short shrift from the philosophers. If all we can say with confidence is “Missed again!” and “A miss is as good as a mile!”, and the history of inquiry is a sequence of such misses, then epistemic pessimism follows. Popper eventually realized that this naive falsificationism is compatible with optimism provided we have an acceptable notion of verisimilitude (or truthlikeness). If some false hypotheses are closer to the truth than others, if verisimilitude admits of degrees, then the history of inquiry may turn out to be one of steady progress towards the goal of truth. Moreover, it may be reasonable, on the basis of the evidence, to conjecture that our theories are indeed making such progress even though it would be unreasonable to conjecture that they are true simpliciter.

Popper saw very clearly that the concept of truthlikeness is easily confused with the concept of epistemic probability, and that it has often been so confused. (See Popper, 1963 for a history off the confusion). Popper's insight here was undoubtedly facilitated by his deep, and largely unjustified, antipathy to epistemic probability. His starkly falsificationist account favors bold, contentful theories. Degree of informative content varies inversely with probability -- the greater the content the less likely a theory is to be true. So if you are after theories which seem, on the evidence, to be true, then you will eschew those which make bold -- that is, highly unlikely -- predictions. On this picture the quest for theories with high probability must be quite wrongheaded. Certainly we want inquiry to yield true propositions, but not any old truths will do. A tautology is a truth, and as certain as anything can be, but it is never the answer to any interesting inquiry outside mathematics and logic. What we want are deep truths, truths which capture more rather than less, of the whole truth.

What, then, is the source of the widespread confusion between probability and truthlikeness? Epistemic probability measures the degree of seeming to be true, while truthlikeness measures degree of being like the truth. Seeming and being like might at first strike one as closely related, but of course they are rather different. Seeming concerns the appearances whereas being like concerns the objective facts, facts about similarity or likeness. Even more important, there is a difference between being true and being the truth. The truth, of course, has the property of being true, but not every proposition that is true is the truth in the sense required by the aim of inquiry. The truth of a matter at which an inquiry aims has to be the complete, true answer. Thus there are two dimensions along which probability (seeming to be true) and truthlikeness (being like the truth) differ radically.

To see this distinction clearly, and to articulate it, was one of Popper's most significant contributions, not only to the debate about truthlikeness, but to philosophy of science and logic in general. As we will see, however, his deep antagonism to probability combined with his great love affair with boldness was both a blessing and a curse. The blessing: it led him to produce not only the first interesting and important account of truthlikeness, but to initiate a whole approach to the problem -- the content approach (Oddie 1978, 1981 and 1986, Zwart 2000). The curse, as is now almost universally recognized: content alone is insufficient to characterize truthlikeness.

Popper made the first assay on the problem in his famous collection *Conjectures and Refutations*. First, let a matter for investigation be circumscribed by a language L adequate for discussing it. (Popper was a great admirer of Tarski's assay on the concept of truth and strove to model his theory of truthlikeness on Tarski's theory.) The world induces a partition of sentences of L into those that are true and those that are false. The set of all true sentences is thus a complete true account of the world, as far as that investigation goes. It is aptly called the Truth, *T*. *T* is the target of the investigation couched in L. It is the theory that we are seeking, and, if truthlikeness is to make sense, theories other than *T*, even false theories, come more or less close to capturing *T*.

T, the Truth, is a theory only in the technical Tarskian sense, not in the ordinary everyday sense of that term. It is a set of sentences closed under the consequence relation: a consequence of some sentences in the set is also a sentence in the set. T may not be finitely axiomatisable, or even axiomatisable at all. Where the language involves elementary arithmetic it follows (from Gödel's theorem) that *T* won't be axiomatisable. However, it is a perfectly good set of sentences all the same. In general we will follow the Tarski-Popper usage here and call any set of sentences closed under consequence a theory, and we will assume that each proposition we deal with is identified with the theory it generates in this sense. (Note that when theories are classes of sentences, theory *A* logically entails theory *B* just in case *B* is a subset of *A*.)

The complement of *T*, the set of false sentences *F*, is not a theory even in this technical sense. Since falsehoods always entail truths, *F* is not closed under the consequence relation. (This is part of the reason we have no complementary expression like *the Falsth*. The set of false sentences does not describe a possible alternative to the actual world.) But *F* too is a perfectly good set of sentences. The consequences of any theory *A* that can be formulated in L will thus divide its consequences between *T* and *F*. Popper called the intersection of *A* and *T*, the *truth content* of *A* (A_T), and the intersection of A and F, the falsity content of *A* (A_F). Any theory *A* is thus the union of its non-overlapping truth content and falsity content. Note that since every theory entails all logical truths, these will constitute a special set, at the center of *T*, which will be included in every theory, whether true or false.

A false theory will cover some of *F*, but because every false theory has true consequences, it will also overlap with some of *T* (Diagram 1).

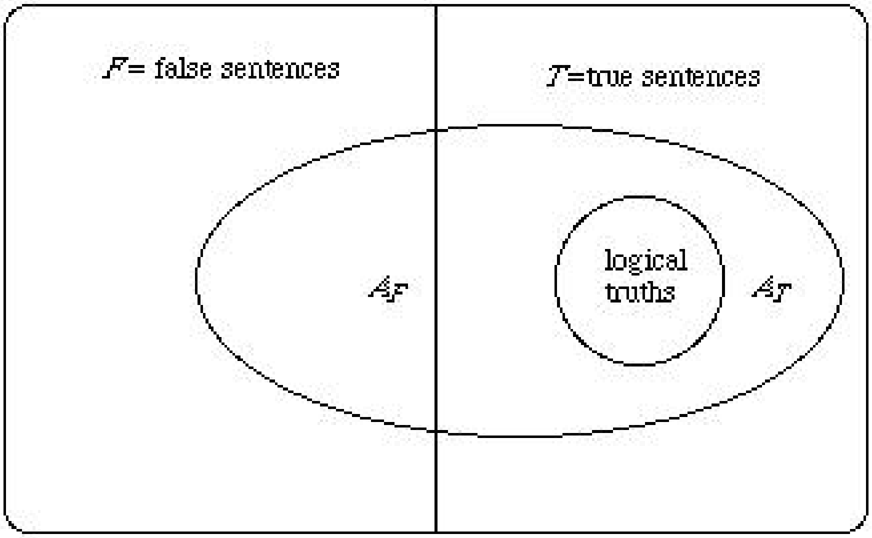


Diagram 1: Truth and falsity contents of false theory *A*

A true theory, however, will only cover *T* (Diagram 2):

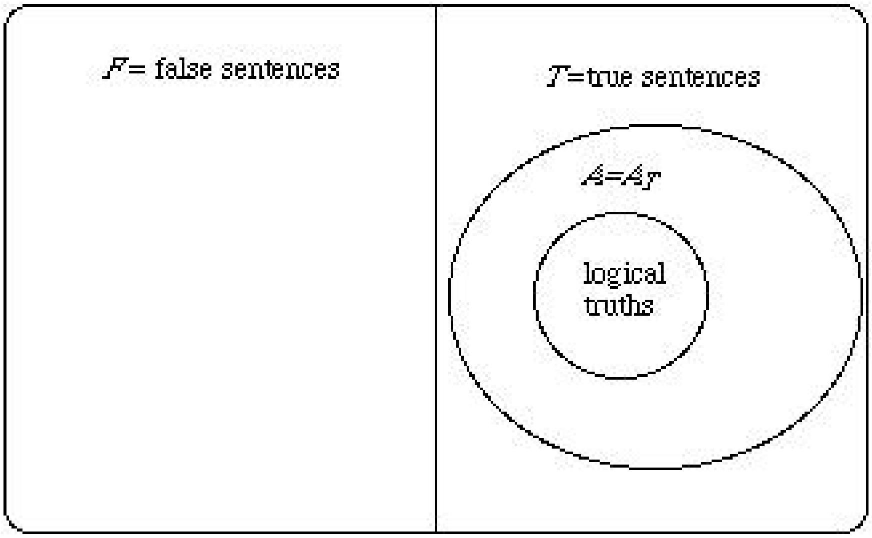


Diagram 2: True theory *A* is identical to its own truth content

Amongst true theories, then, it seems that the more true sentences entailed the closer we get to *T*, hence the more truthlike. Set theoretically that simply means that, where *A* and *B* are both true, *A* will be more truthlike than *B* just in case *B* is a subset of *A* (which for true theories means that the truth content of *B* is a subset of the truth content of *A*.

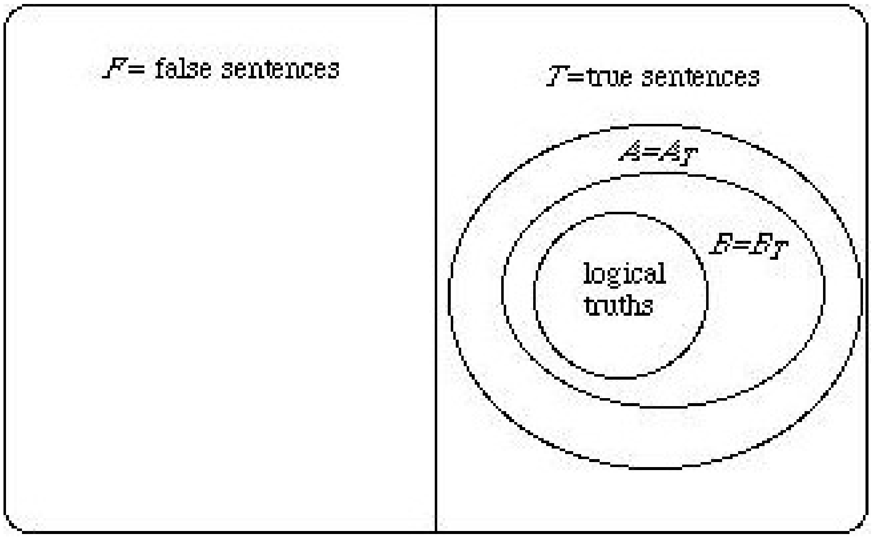


Diagram 3: True theory *A* has more truth content than true theory *B*

This account has some nice features. It induces a partial ordering of truths, with the whole Truth *T* at the top of the ordering: *T* is closer to the Truth than any other true theory. The set of logical truths is at the bottom: further from the Truth than any other true theory. In between these two extremes, true theories are ordered simply by logical strength: the more logical content, the closer to the Truth. Since epistemic probability varies inversely with logical strength, amongst truths the theory with the greatest truthlikeness (*T*) must have the smallest probability, and the theory with the largest probability (the logical truth) is the furthest from the Truth. Popper's love affair with logical strength is thus consummated in his first sketch of an account of truthlikeness.

Popper made a bold and simple generalization of this. Just as truth content (coverage of *T*) counts in favour of truthlikeness, falsity content (coverage of *F*) counts against. In general then, a theory A is closer to the truth if it has more truth content without engendering more falsity content, or has less falsity content without sacrificing truth content (diagram 4):

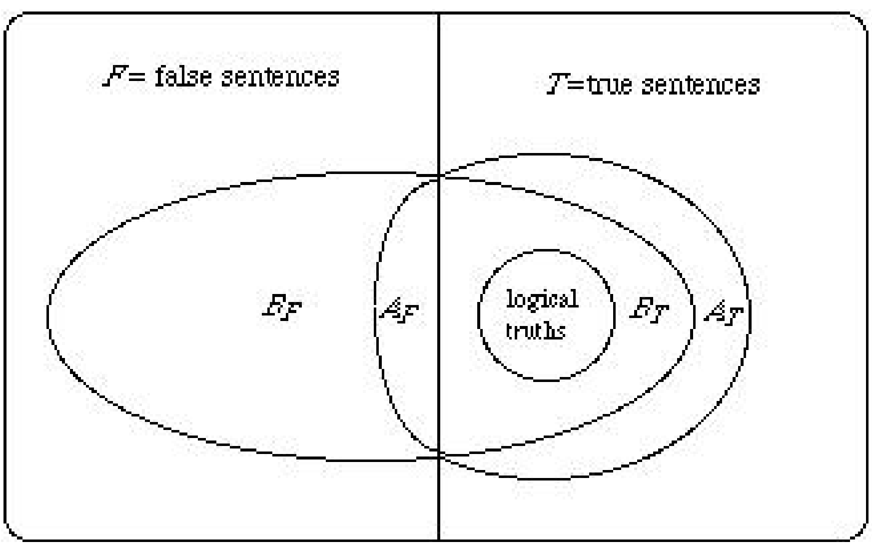


Diagram 4: False theory *A* closer to the Truth than false theory B

This account also has some nice features. It preserves the comparisons of true theories mentioned above. The truth content A_T of a false theory *A* (itself a theory) will clearly be closer to the truth than *A* (diagram

1). More generally, a true theory *A* will be closer to the truth than a false theory *B* provided A's truth content exceeds B's.

Despite these nice features the account suffers the following fatal flaw: it entails that no false theory is closer to the truth than any other. This was shown independently by Pavel Tichý and David Miller (in 1973, reported in Miller 1974, and Tichý 1974). It is instructive to see why. Let us suppose that *A* and *B* are both false, and that *A*'s truth content exceeds *B*'s. Let *a* be a true sentence entailed by *A* but not by *B*. Let *f* be any falsehood entailed by A. Since *A* entails both *a* and *f* the conjunction, *a*&*f* is a falsehood entailed by *A*, and so part of *A*'s falsity content. If *a*&*f* were also part of *B*'s falsity content *B* would entail both *a* and *f*. But then it would entail *a* contrary to the assumption. Hence *a*&*f* is in *A*'s falsity content and not in *B*'s. So *A*'s truth content cannot exceeds *B*'s without *A*'s falsity content also exceeding *B*'s. Suppose now that *B*'s falsity content exceeds *A*'s. Let *g* be some falsehood entailed by *B* but not by *A*, and let *f*, as before, be some falsehood entailed by *A*. The sentence $f \supset g$ is a truth, and since it is entailed by *g*, is in *B*'s truth content. If it were also in *A*'s then both *f* and $f \supset g$ would be consequences of *A* and hence so would *g*, contrary to the assumption. Thus *A*'s truth content lacks a sentence, $f \supset g$, which is in *B*'s. So *B*'s falsity content cannot exceeds *A*'s without *B*'s truth also content exceeding *A*'s. The relationship depicted in diagram 4 simply cannot obtain.

It is tempting to retreat to something like the comparison of truth contents alone. But then we get a result which is almost as bad: that a false theory is the closer to the truth the stonger it is. So, for example, since the false proposition that *all heavenly bodies are made of green cheese* is logically stronger than the false proposition that *all heavenly bodies orbiting the earth are made of green cheese* the former is closer to the truth. And once we know a theory is false we can be confident that tacking any old arbitrary proposition will lead us inexorably closer to the truth. Amongst false theories, brute strength becomes the only criterion of a theory's likeness to truth.

After the failure of Popper's proposal there have been two main variations on the content approach. One stays within Popper's essentially syntactic paradigm, comparing classes of true and false sentences (e.g. Schurz and Weingarter 1987, Newton Smith 1981). The other makes the switch to a more semantic paradigm, searching for a plausible theory of distance between propositions, construing these not as classes of sentences, but rather as classes of possibilities. One main variant takes the class of models of a language as a surrogate for possible states of affairs (Miller 1978a). The other utilizes a semantics of incomplete possible states like those favored by structuralist accounts of scientific theories (Kuipers 1987). The idea which these share in common is that the distance between two propositions is measured by the symmetric difference of the two classes of associated states. Roughly speaking, the larger the symmetric difference, the greater the distance between the two propositions.

If the truth is taken to be represented by a unique model, or complete possible world, then we end up with results very close to Popper's truth content account (Oddie 1978). In particular, false propositions are closer to the truth the stronger they are. However, if we take the structuralist approach then we will take the relevant states of affairs to be “small” states of affairs -- chunks of the world rather than the entire world -- and then the possibility of more fine-grained distinctions between theories opens up. The most promising recent developments exploring this idea are to be found in Volpe 1995.

The fundamental problem with the oriignal pure content approach lies not with the particular proposals but with the underlying strength assumption: that verisimilitude is a function of just two variables -- logical strength and truth value. This assumption has a number of somewhat counterintuitive consequences.

Firstly, it is clear that whatever function of strength and truth value one selects, a given theory A can have only two degrees of verisimilitude: one in case it is false and the other in case it is true. This is obviously wrong. A theory can be false in very many different ways. The proposition that *there are eight planets* is false whether there are nine planets or a thousand planets, but the degree of truthlikeness is much higher in the first case than in the latter. We will see later that the degree of verisimilitude of a true theory may also vary according to where the truth lies.

Secondly, the brute strength assumption entails that if we fix truth value, verisimilitude will vary only on strength. So, for example, two equally strong false theories will have to have the same degree of verisimilitude. That's pretty far-fetched. That there are eight planets and that there are a thousand planets are (intuitively) equally strong, and both are false in fact (assuming that Pluto is indeed a planet), but the latter is much further from the truth.

Finally, how does strength determine verisimilitude amongst false theories? There are really only two plausible candidates: that verisimilitude increases with increasing strength, or that it decreases with increasing strength. Both proposals are at odds with attractive judgements and principles. One does not necessarily make a step toward the truth by reducing the content of a false proposition. The proposition that *the moon is the only heavenly body made of green cheese* is logically stronger than the proposition that *the moon is made of green cheese*, but the latter hardly seems a step towards the truth. Nor does one necessarily make a step toward the truth by increasing the content of a false theory. The false proposition that *all heavenly bodies are made of green cheese* is logically stronger than the false proposition *all heavenly bodies orbiting the earth are made of green cheese* but doesn't seem to constitute progess towards the truth.

3. The Second Assay: the Likeness Approach

In the wake of the collapse of Popper's account two philosophers, working quite independently, suggested a radically different approach: one which takes the likeness in truthlikeness seriously (Tichý 1974, Hilpinen 1976). The shift from content to likeness is also marked by a shift from Popper's syntactic approach to one that is more semantic, one trafficking in possible worlds.

A possible world is a complete possible way for things to be. It is a complete distribution of properties, relations and magnitudes over the appropriate kinds of items. Naturally, these distributions are relativized to a certain collection of features. A proposition carves the class of possibilities into two -- those in which the proposition is true and those in which it is false. Call the class of worlds in which the proposition is true its range. Some have proposed that propositions simply be identified with their ranges, but whether

or not that identification is plausible, certainly the range of a proposition is an important aspect of it. It is the proposition's truth condition. Normal logical relations and operations correspond to well-understood set-theoretic relations and operations on ranges. The range of the conjunction of two proposition is the intersection of the ranges of the two conjuncts. Entailment corresponds to the subset relation on ranges. The actual world is a single point in logical space -- a complete specification of every matter of fact -- and a proposition is true if its range contains the actual world, false otherwise. The Truth is the complete true proposition: that proposition which entails all true propositions. It is none other than the singleton of the actual world. That singleton is the target, the bullseye, the thing at which the most comprehensive inquiry is aiming.

In addition to the set-theoretic structures which underlie the familiar logical relations, the logical space might be structured by similarity or likeness. For example, worlds might be more or less like other worlds. There might be a betweeness relation amongst worlds, or even a fully-fledged distance metric. If that's the case we can start to see how one proposition might be closer to the Truth, the target world, than another. Suppose, for example, that worlds are arranged in similarity spheres nested around the actual world, familiar from the Stalnaker-Lewis approach to counterfactuals. Consider Diagram 5:

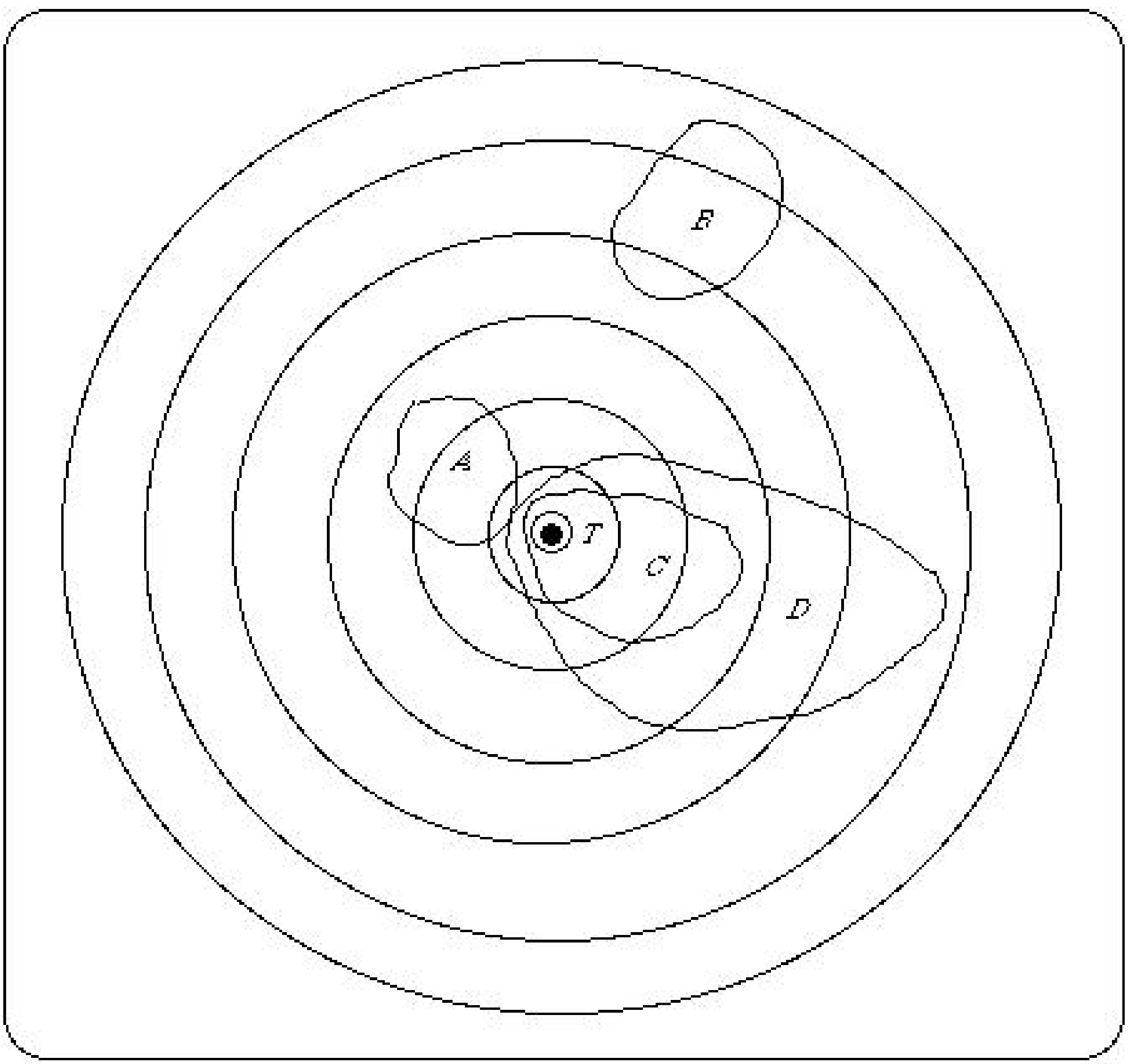


Diagram 5: Verisimilitude by similarity circles

The bullseye is the actual world and the small sphere which includes it is *T*, the Truth. The nested spheres represent likeness to the actual world. A world is less like the actual world the larger the first sphere of which it is a member. Propositions *A* and *B* are false, *C* and *D* are true. A carves out a class of worlds which are rather close to the actual world -- all within spheres two to four -- whereas *B* carves out a class rather far from the actual world -- all within spheres five to seven. Intuitively *A* is closer to the bullseye than is *B*.

The largest sphere which does not overlap at all with a proposition is plausibly a measure of how close the proposition is to being true. Call that the *truth factor*. A proposition *X* is closer to being true than *Y* if the truth factor of *X* is included in the truth factor of *Y*. The truth factor of *A*, for example, is the smallest non-empty sphere, *T* itself, whereas the truth factor of *B* is the fourth sphere, of which *T* is a proper subset.

If a proposition includes the bullseye then of course it is true simpliciter, it has the maximal truth factor (the empty set). So all true propositions are equally close to being true. But truthlikeness is not just a matter of being close to being true. The tautology, *D*, *C* and the Truth itself are equally true, but in that order they increase in their closeness to the whole truth. Taking a leaf out of Popper's book, we can regard closeness to the whole truth as in part a matter of degree of informativeness of a proposition. In the case of the true propositions, this correlates roughly with the smallest sphere which totally includes the proposition. The further out the outermost sphere, the less informative the proposition is, because the larger the area of the logical space which it covers. So, in a way which echoes Popper's account, we could take truthlikeness to be a combination of the truth factor and the content factor.

> *X* is closer to the truth than *Y* if and only if *X* does as well as *Y* on both truth factor and content factor, and better on at least one of those.

Applying this definition we capture two judgements, in addition to those already mentioned, that seem intuitively acceptable: that *C* is closer to the truth than *A*, and that *D* is closer than *B*. (Note, however, that we have here a partial ordering: *A* and *D*, for example, are not ranked). We can derive various apparently desirable features of the relation closer to the truth: for example, that the relation is transitive, asymmetric and irreflexive; that the Truth is closer to the Truth than any other theory; that the tautology is at least as far from the Truth as any other truth; that one cannot make a true theory worse by strengthening it by a truth; that a falsehood is not necessarily improved by adding another falsehood. But there are also some worrying features here. No falsehood can be closer to the truth than any truth, for example. So Newton's theory is no closer to the Truth than the tautology. That's bad.

Stating Hilpinen's account in the above fashion masks its departure from Popper's account.. The incorporation of similarity spheres marks a fundamental break with the pure content approach, and opens up a range of possible new accounts.

One objection to Hilpinen's proposal is that it simply takes as given the similarity relation. Tichý

anticipated this objection, and at the end of his 1974 paper he not only suggested the use of similarity rankings on worlds, but also provided a ranking in simple cases and indicated how to generalize this to more complex cases.

Examples and counterexamples in Tichý 1974 are very simple, framed in a language with three primitives -- *h* (for the state hot), *r* (for rainy) and *w* (for windy). The sentences of this language are taken to express propositions over a dinky little eight-membered logical space. Tichý took judgements of truthlikeness like the following to be self-evident: Suppose that in fact it is hot, raining and windy. Then the proposition that it is cold, and dry and still (expressed by the sentence ~*h*&~*r*&~*w*) is further from the truth than the proposition that it is cold, rainy and windy (expressed by the sentence ~*h*&*r*&*w*). And the proposition that it is cold, dry and windy (expressed by the sentence ~*h*&~*r*&*w*) is somewhere between the two. These kinds of judgements are taken to be core intuitions which any adequate account of truthlikeness would have to deliver, and which Popper's theory patently can not handle. Unlike Popper, Tichý is not trying to find the missing theoretical bridge to epistemic optimism in a fallibilist philosophy of science. Rather, he takes the intuitive concept of truthlikeness to be as much a standard component of the intellectual armory of the folk as is the concept of truth. Doubtless, like the concept of truth, it needs tidying up and trimming down, but he assumes that it is basically sound, and that the job of the philosopher-logician is to explicate it: to give a precise, logically perspicuous, consistent account which captures the core intuitions and excludes core counterintuitions. In the grey areas, where our intuitions are not clear, it is a case of “spoils to the victor” -- the best account of the core intuitions can legislate where the intuitions are fuzzy or contradictory.

4. Similarity in a Simple Logical Space

Consider the eight-membered logical space generated by distributions of truth values through the three basic conditions: hot; rainy; windy. There are eight of them, expressed by the eight propositional constituents, or maximal conjunctions, of which the following are a sample:

w1 *h*&*r*&*w*

w2 *h*&*r*&~*w*

w5 ~*h*&*r*&*w*

w8 ~*h*&~*r*&~*w*

Worlds differ in the distributions of these traits, and a natural, albeit simple, suggestion is to measure the likeness between two worlds by the number of agreements on traits. (This is tantamount to taking distance to be measured by the size of the symmetric difference of generating states. As is well known, this will generate a genuine metric, in particular satisfying the triangular inequality.) If w1 is the actual world this immediately induces a system of nested spheres, but one in which the spheres come with numbers attached:

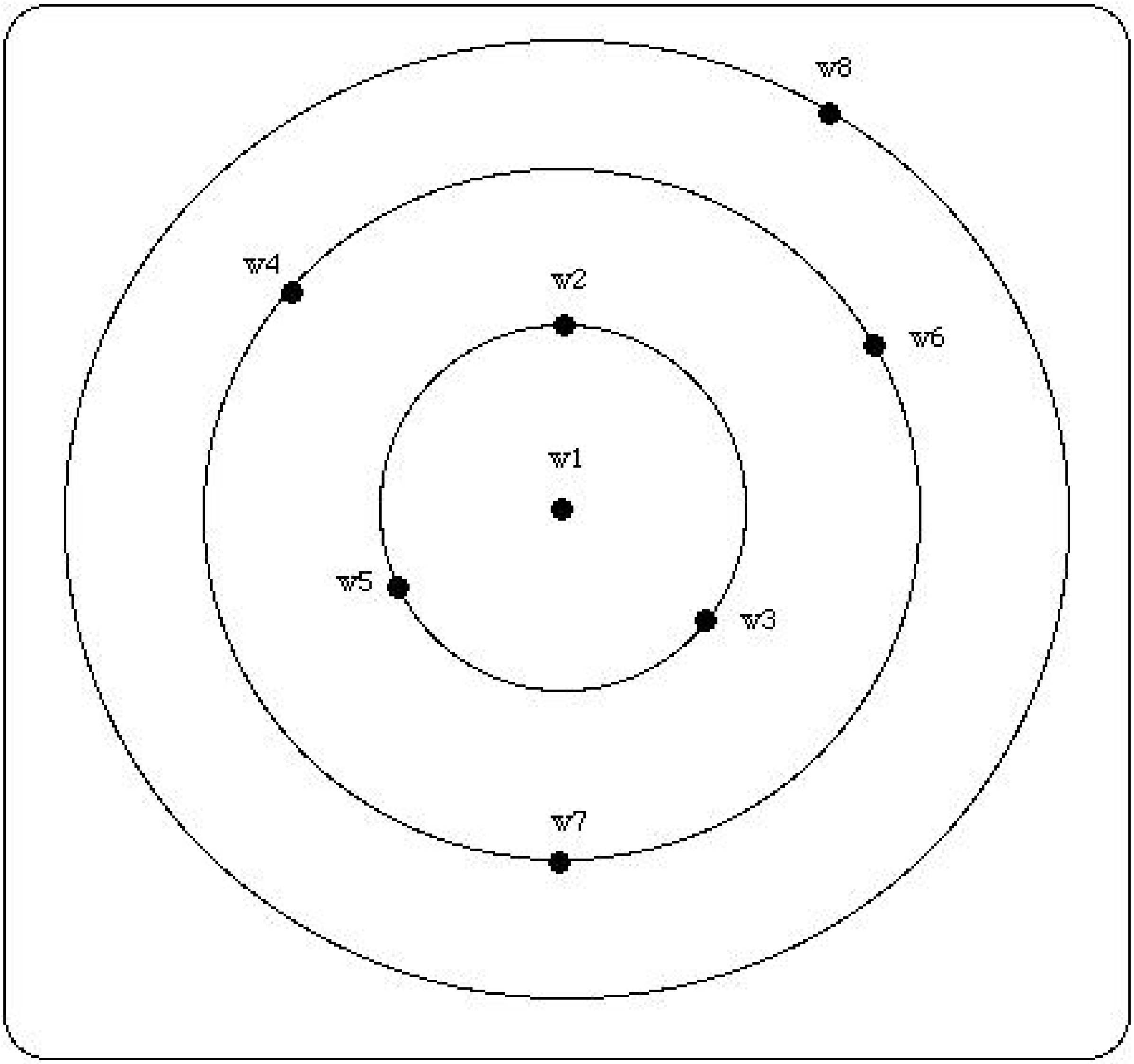


Diagram 6: Similarity circles for the weather space

Those worlds orbiting on the sphere n are of distance n from the actual world. Now that we have numbers associated we can do something a little more ambitious than the partial ordering induced by Hilpinen's proposal. A numerical measure can be defined as some function of distances, from the actual world, of worlds in the range of a proposition. One particularly simple proposal is to take the average distance of worlds from the actual world. This is tantamount to measuring the distance from actuality of the “center of gravity” of the proposition.

This idea of averaging delivers all of the particular judgements we used above to motivate Hilpinen's proposal, but in conjunction with the metric it delivers more comparisons. For example, we have the following sample propositions in descending order of truthlikeness:

Truth Value	Proposition	Distance
true	*h&r&w*	0

true	*h*&*r*	0.5
false	*h*&*r*&~w	1.0
true	*h*	1.3
false	*h*&~*r*	1.5
false	~*h*	1.7
false	~*h*&~*r*&*w*	2.0
false	~*h*&~*r*	2.5
false	~*h*&~*r*&~*w*	3.0

So far these results look quite pleasing. Propositions are closer to the truth the more they get the basic weather traits right, further away the more mistakes they make. A false proposition may be made either worse or better by strengthening (~*w* is the same distance from the Truth as ~*h*; *h*&*r*&~*w* is better than ~*w* while ~*h*&~*r*&~*w* is worse). A false proposition (like *h*&*r*&~*w*) can be closer to the truth than some true propositions (like *h*).

But a number of problems immediately arise. First, the proposal embodies presuppositions of equal weight: both for basic states (in the distance metric on worlds) and for worlds (in the averaging procedure). These simplifications can be easily relaxed. We can weight the different factors according to their importance for the purposes of similarity, and we can clearly take a weighted average rather than a straight average. More importantly, some of the apparently pleasing general features of Hilpinen's account are violated. We can find pairs of true propositions such that the stronger is further from the truth. The tautology is not the true proposition furthest from the Truth (Popper 1976).

Truth Value	**Proposition**	**Distance**
true	*h* ∨ ~*r* ∨ *w*	1.4
true	*h* ∨ ~*r*	1.5
true	*h* ∨ ~*h*	1.5

In deciding how to proceed here we confront a methodological problem. The methodology exemplified by Tichý is very much bottom-up. For the purposes of deciding between rival accounts it takes the intuitive data very seriously . Popper, and Popperians like Miller, take a far more top-down approach. They are suspicious of folk intuitions, and consider themselves to be in the business of constructing a new concept rather than explicating an existing one. They do place enormous weight on certain plausible general principles, largely those that fit in with other principles of their overall theory of science: like the

principle that strength is a virtue and that the stronger of two true theories is the closer to the Truth. A third approach, one which lies between these two extremes, is that of reflective equilibrium. This recognizes the claims of both intuitive judgements on low-level cases, and plausible high-level principles, and enjoins us to bring principle and judgement into equilibrium, possibly by tinkering with both. Neither intuitive low-level judgements nor plausible high-level principles are given advance priority. The protagonist in the truthlikeness debate who argues most consistently for this approach is Niiniluoto.

How does this impact on the current dispute? Consider a different space of possibilities, generated by a single magnitude like the number of the planets (N). For the sake of the argument let's agree that N is in fact 9 and that the further n is from 9, the further the proposition that N=n from the Truth. Consider three sets of propositions. In the left-hand column we have a sequence of false propositions which, intuitively, decrease in truthlikeness while increasing in strength. In the middle column we have a sequence of corresponding true propositions, in each case the strongest true consequence of its false counterpart on the left (Popper's “truth content”). Again members of this sequence steadily increase in strength. Finally on the right we have another column of falsehoods. These are also steadily increasing in strength, and like the left-hand falsehoods, seem also to be decreasing in truthlikeness.

Falsehood (1)	Strongest True Consequence	Falsehood (2)
11 ≤ N ≤ 20	N=9 or 11 ≤ N ≤ 20	N=10 or 11 ≤ N ≤ 20
12 ≤ N ≤ 20	N=9 or 12 ≤ N ≤ 20	N=10 or 12 ≤ N ≤ 20
……	……	……
19 ≤ N ≤ 20	N=9 or 19 ≤ N ≤ 20	N=10 or 19 ≤ N ≤ 20
N = 20	N=9 or N = 20	N=10 or N = 20

Judgements about the closeness of the true propositions to the truth may be less clear than are intuitions about their left-hand counterparts. However, it would seem highly incongruous to judge the truths to be steadily increasing in truthlikeness, while the falsehoods on the right, minimally different in content, steadily decrease in truthlikeness. So both the bottom-up approach and reflective equilibrium suggest that all three are sequences of steadily increasing strength combined with steadily decreasing truthlikeness. And that is enough to overturn Popper's claim that amongst true theories strength and truthlikeness covary. This removes an objection to averaging (or weighted averaging), but does not settle the issue in its favor, for there may still be other more plausible counterexamples to averaging that we have not considered.

5. Generalizing the Approach

The similarity approach faces two obstacles, both of which have analogues in the history of the

development of logical probability. The first is the difficulty of defining a plausible distance measure on spaces of complex worlds. Simple finite examples are all very nice for the purposes of illustration, but what we want is some indication that this is not a special case. The second has been dubbed the problem of translation-invariance. A measure of truthlikeness should be invariant under translations into essentially equivalent languages. This requirement simply reflects the fact that closeness to the truth is a semantic, not a syntactic, affair. David Miller has argue that any account which captures low-level intuitions about similarity must fail this constraint.

One fruitful way of generalizing the simple idea to complex spaces involves cutting such spaces down into finite chunks. This can be done in various ways, but one promising idea (Tichý, 1974, Niiniluoto 1976) is to make use of a certain kind of normal form -- Hintikka's *distributive normal forms* (Hintikka 1963). Constituents correspond to propositional maximal conjunctions. Hintikka defined what he called constituents, which, like maximal conjunctions, are jointly exhaustive and mutually exclusive. Constituents lay out, in a very perspicuous manner, all the different ways individuals can be related. For example, Every sentence in a first-order language comes with a certain *depth* -- the number of embedded quantifiers required to formulate it. So, for example, (1) is a depth-1 sentence; (2) is depth-2; and (3) is depth-3.

> (1) Everyone loves himself.
> (2) Everyone loves another.
> (3) Everyone who loves another loves the other's lovers.

We could call a proposition *depth-d* if the shallowest depth at which it can be expressed is *d*. What Hintikka showed is that every depth-*d* proposition can be expressed by a disjunction of depth-*d* constituents. Constituents can be represented as finite tree-structures, the nodes of which are like straightforward conjunctions of atomic states. Consequently, if we can measure distance between such trees we will be well down the path of measuring the truthlikeness of depth-d propositions: it will be some function of the distance of constituents in its normal form from the true depth-d constituent.

This program has proved flexible and fruitful, delivering a wide range of intuitively appealing results in simple first-order cases. Further, the general idea can be extended in a number of different directions: to higher-order languages and to spaces based on functions rather than properties.

6. Translation Invariance

The single most influential argument against the likeness approach is the charge that it is not translation invariant (Miller 1974a, 1975 a, 1976). Early formulations of the approach (Tichý 1974, 1976) proceeded in terms of syntactic entities -- sentences, predicates, distributive normal forms and the like. The question naturally arises, then, whether we obtain the same measures if all the sentences are translated into an essentially equivalent language -- one capable of expressing the same propositions with a different set of primitive predicates.

Take our simple weather-framework above. This trafficks in three primitives -- *hot*, *rainy*, and *windy* Suppose, however, that we define the following two new weather conditions:

minnesotan $=_{\text{df}}$ hot if and only if rainy

arizonan $=_{\text{df}}$ hot if and only if windy

Now it appears as though we can describe the same sets of weather states in an *h-m-a*-ese based on these conditions.

	***h-r-w*-ese**	***h-m-a*-ese**
T	*h&r&w*	*h&m&a*
A	*~h&r&w*	*~h&~m&~a*
B	*~h&~r&w*	*~h&m&~a*
C	*~h&~r&~w*	*~h&m&a*

If **T** is the truth about the weather then theory **A**, in *h-r-w*-ese, seems to make just one error concerning the original weather states, while **B** makes two and **C** makes three. However, if we express these two theories in *h-m-a*-ese however, then this is reversed: **A** appears to make three errors and **B** still makes two and **C** makes only one error. But that means the account makes truthlikeness, unlike truth, radically language-relative.

There are two live responses to this criticism. But before detailing them, note a dead one: the similarity theorist cannot object that *h-m-a* somehow logically different from *h-r-w*, on the grounds that the primitives of the latter are essentially biconditional whereas the primitives of the former are not. This is because there is a perfect symmetry between the two. Starting within *h-m-a*-ese we can arrive at the original primitives by exactly analogous definitions:

rainy $=_{\text{df}}$ hot if and only if minnesotan
windy $=_{\text{df}}$ hot if and only if arizonan

Thus if we are going to object to *h-m-a*-ese it will have to be on other than purely logical grounds.

Firstly, then, the similarity theorist could maintain that certain predicates (presumably “hot”, “rainy” and “windy”) are primitive in some absolute, realist, sense. Such predicates “carve reality at the joints” whereas others (like “minnesotan” and “arizonan”) are gerrymandered affairs. With the demise of predicate nominalism as a viable account of properties and relations this approach is not as unattractive

as it might have seemed in the middle of the last century. Realism about universals is certainly on the rise. While this version of realism presupposes a sparse theory of properties -- that is to say, it is not the case that to every definable predicate there corresponds a genuine universal -- such theories have been championed both by those doing traditional a priori metaphysics of properties (e.g. Bealer 1982) as well as those who favor or more empiricist, scientifically informed approach (e.g. Armstrong 1978, Tooley 1977). According to Armstrong, for example, which predicates pick out genuine universals is a matter for developed science. The primitive predicates of our best fundamental physical theory will give us our best guess at what the genuine universals in nature are. They might be predicates like electron or mass, or more likely something even more abstruse and remote from the phenomena -- like the primitives of String Theory.

One apparently powerful objection to this realist solution is that it would render the task of empirically estimating degree of truthlikeness completely hopeless. If we know a priori which primitives should be used in the computation of distances between theories it will be difficult to estimate truthlikeness, but not impossible. For example, we might compute the distance of a theory from the various possibilities for the truth, and then make a weighted average, weighting each possible true theory by its probability on the evidence. That would be the credence-mean estimate of truthlikeness. However, if we don't know which features should count towards the computation of similarities and distances then we cannot even get off first base.

To see this consider our simple weather frameworks. Suppose that all I learn is that it is rainy. Do I thereby some grounds for thinking **A** is closer to the truth than **B**? I would if I also knew that *h*-*r*-*w*-ese is the language for calculating distances. For then whatever the truth is, **A** makes one fewer mistake than **B** makes. **A** gets it right on the rain factor, while **B** doesn't, and they must score the same on the other two factors whatever the truth of the matter. But if we switch to *h*-*m*-*a*-ese then **A**'s epistemic superiority is no longer guaranteed. If, for example, **T** is the truth then **B** will be closer to the truth than **A**. That's because in the *h*-*m*-*a* framework raininess as such doesn't count in favor or against the truthlikeness of a proposition.

However, this objection fails if there can be empirical indicators of which conditions are the genuine ones, the ones that carve reality at the joints. Obviously the framework would have to contain more than just *h*, *m* and *a*. It would have to contain resources for describing the states that indicate whether these were genuine primitives. Maybe whether they enter into genuine causal relations will be important, for example. Once we can distribute probabilities over the various candidates for the real universals, then we can use those probabilities to weight the various possible distances which a hypothesis might be from any given theory.

The second live response is both more modest and more radical. It is more modest in that it is not hostage to the objective priority of a particular conceptual scheme, whether that priority is accessed a priori or a posteriori. It is more radical in that it denies a premise of the invariance argument that at first blush is apparently obvious. It denies the equivalence of the two conceptual schemes. It denies that *h*&*r*&*w*, for example, expresses the very same proposition as *h*&*m*r&*a* expresses. If we deny translatability then we can grant the invariance principle, and grant the judgements of distance in both cases, but remain

untroubled. There is no contradiction. (Tichý 1978,, Oddie 1986).

At first blush this seems truly desperate. Haven't the respective conditions been defined in such a way that they are simple equivalent by fiat? That would, of course, be the case if m and a had been introduced as defined terms into *h-r-w*. But if that were the intention then the similarity theorist could retort that the calculation of distances should proceed in terms of the primitives, not the introduced terms. However that is not the only way the argument can be read. We are asked to contemplate two partially overlapping sequences of conditions, and two spaces of possibilities generated by those two sequences. We can thus think of each possibility as a point in a simple three dimensional space. These points are ordered triples of 0s and 1s, the nth entry being a 0 if the *n*th condition is satisfied and 1 if it isn't. Thinking of possibilities in this way, we already have rudimentary geometrical features generated simply by the selection of generating conditions. Points are adjacent if the differ on only one dimension. A path is a sequence of adjacent points. A point *q* is between two points *p* and *r* if *q* lies on a shortest path from *p* to *r*. A region of possibility space is convex if it is closed under the betweeness relation -- anything between two points in the region is also in the region.

Evidently we have two spaces of possibilities, S1 and S2, and the question now arises whether a sentence interpreted over one of these spaces expresses the very same thing as any sentence interpreted over the other. Does *h*&*r*&*w* express the same thing as *h*&*m*&*a*? *h*&*r*&*w* expresses (the singleton of) u1 (which is the entity <1,1,1> in S1 or <1,1,1>$_{S1}$) and *h*&*m*&*a* expresses v1 (the entity <1,1,1>$_{S2}$). ~*h*&*r*&*w* expresses u2 (<0,1,1>$_{S1}$), a point adjacent to that expressed by *h*&*r*&*w*. However ~*h*&~*m*&~*a* expresses v8 (<0,0,0>$_{S2}$), which is not adjacent to v1 (<1,1,1>$_{S2}$). So now we can construct a simple proof that the two sentences do not express the same thing.

> u1 is adjacent to u2
>
> v1 is not adjacent to v8

therefore

> either u1 is not identical to v1 or u2 is not identical v8.

therefore

> Either *h*&*r*&*w* and *h*&*m*&*a* do not express the same thing, or
>
> ~*h*&*r*&*w* and ~*h*&~*m*&~*a* do not express the same thing.

Thus at least one of the two required intertranslatability claims fails, and *h-r-w*-ese is not intertranslatable with *h-m-a*-ese. The important point here is that a space of possibilities already comes with a structure and the points in such a space cannot be individuated without reference to rest of the space and its

structure. The identity of a possibility is bound up with its geometrical relations to other possibilities. Different relations, different possibilities.

This idea meshes well with recent work on conceptual spaces in Gärdenfors [2000]. Gärdenfors is concerned both with the semantics and the nature of genuine properties, and his bold and simple hypothesis is that properties carve out convex regions of an n-dimensional quality space. He supports this hypothesis with an impressive array of logical, linguistic and empirical data. (Looking back at our little spaces above it is not hard to see that the convex regions are those that correspond to the generating (or atomic) conditions and conjunctions of those. See Oddie 1987a.) While is dealing with properties it is not hard to see that similar considerations apply, since propositions can be regarded as 0-ary properties.

Ultimately, however, this response seems less than entirely satisfactory by itself. If the choice of a conceptual space is just a matter of taste then we may be forced to embrace a radical kind of incommensurability. Those who talk *h-r-w*-ese and conjecture ~*h*&*r*&*w* on the basis of the available evidence will be close to the truth. Those who talk *h-m-a*-ese while exposed to the "same" circumstances would presumably conjecture ~*h*&~*m*&~*a* on the basis of the "same" evidence (or the corresponding evidence that they gather). If in fact *h*&*r*&*w* is the truth (in *h-r-w*-ese) then the *h-r-w* weather researchers will be close to the truth. But the *h-m-a* researchers will be very far from the truth. This may not be an explicit contradiction, but it should be worrying. Realists started out with the ambition of defending a concept of truthlikeness which would enable them to embrace both fallibilism and optimism. But what they have ended up with is something that smacks of rather too radical a version of the incommensurability of competing conceptual frameworks. Presumably the realist will need to add that some conceptual schemes really are better than others. Some "carve reality at the joints" and others don't. But is that something the realist will be reluctant to affirm?

Bibliography

- Armstrong, D. M. *What is a Law of Nature?* (Cambridge: Cambridge University Press, 1983).
- Barnes, E. "The Language Dependence of Accuracy", *Synthese* 84 (1990), 54-95.
- ------. "Beyond Verisimilitude: A Linguistically Invariant Basis for Scientific Progress", *Synthese* 88 (1991), 309-339.
- ------. "Truthlikeness, Translation, and Approximate Causal Explanation", *Philosophy of Science* 62, 2 (1995),15-226.
- Bealer, G. *Quality and Concept*. Clarendon Library of Logic and Philosophy (Oxford: Clarendon, 1982).
- Bonilla, J.S. "Verisimilitude, Structuralism and Scientific Progress", *Erkenntniss*, January 1996;
- ------. "Truthlikeness Without Truth: A Methodological Approach", Synthese, December 1992
- Brink, C., & Heidema, J. "A verisimilar ordering of theories phrased in a propositional language", *The British Journal for the Philosophy of Science*, 38 (197), 533-549.
- Britz, K., & Brink, C. "Computing verisimilitude" *Notre Dame Journal of Formal Logic*, 36(2) (1995), 30-43.
- Cohen, L.J. "What has science to do with truth?" *Synthese* 45, (1980), 489-510.

- ------. "Verisimilitude and legisimilitude" in Kuipers 1987, 129-45.
- Gerla, G. "Distances, diameters and verisimilitude of theories" *Archive for mathematical Logic* 31 (6), (1992), 407-14.
- Gärdenfors, P. *Conceptual Spaces* (Cambridge: MIT Press, 2000).
- Harris, J. "Popper's definition of 'Verisimilitude'" *The British Journal for the Philosophy of Science* 25, (1974), 160-6..
- Hilpinen, R. "Approximate truth and truthlikeness" in Przelecki, et al (eds.) *Formal Methods in the Methodology of the Empirical Sciences* (Dordrecht: Reidel, 1976).
- Hintikka, J. "Distributive normal forms in first-order logic" in *Formal Systems and Recursive Functions. Proceedings of the Eight Logic Colloquium* eds. J.N. Crossley and M.A.E. Dummett, (Amsterdam: North-Holland Pub Co, 1963), pp. 47-90.
- Kieseppa, I.A. *Truthlikeness for Multidimensional, Quantitative Cognitive Problems* (Dordrecht: Kluwer,1996).
- ------. "Truthlikeness for Hypotheses Expressed in Terms of n Quantitative Variables", 1996, *Journal of Philosophical Logic* 25, 109-134.
- ------. "On the Aim of the Theory of Verisimilitude", *Synthese* 107 (1996), 421-438
- Kuipers, T. A. F. (ed.) What is closer-to-the-truth? A parade of approaches to truthlikeness, Poznan Studies in the Philosophy of the Sciences and the Humanities, Volume 10, (Amsterdam: Rodopi, 1987).
- ------. "A structuralist approach to truthlikeness", in Kuipers (1987), 79-99.
- ------. "Truthlikeness of stratified theories", in Kuipers (1987) 177-186.
- ------. "Naive and refined truth approximation", *Synthese*, 93 (1992), 299-341.
- Miller, D. "The Truth-likeness of Truthlikeness", *Analysis* 33 (2), (1972), 50-55.
- ------. "Popper's Qualitative Theory of Verisimilitude", *The British Journal for the Philosophy of Science* 25, (1974a), 166-177.
- ------. "On the Comparison of False Theories by Their Bases", *The British Journal for the Philosophy of Science* 25 (2), (1974b), 178-188
- ------. "The Accuracy of Predictions", *Synthese* 30 (1/2), (1975a), 159-191.
- ------. "The Accuracy of Predictions: A Reply", *Synthese* 30 (1/2), (1975b), 207-21.
- ------. "Verisimilitude Redeflated", *The British Journal for the Philosophy of Science* 27 (4), (1976), 363-381.
- ------. "Bunge's Theory of Partial Truth Is No Such Thing", *Philosophical Studies* 31 (2), (1977), 147-150.
- ------. "The Distance Between Constituents", *Synthese* 38 (2), (1978a), 197-212.
- ------. "On Distance from the Truth as a True Distance", in Hintikka et al.(eds) *Essays on Mathematical and Philosophical Logic*, (Dordrecht : Reidel, 1978b), 415-435.
- ------. "Truth, Truthlikeness, Approximate Truth", *Fundamenta Scientiae* 3 (1), (1982), 93-101
- ------. "Impartial Truth", in Skala et al. (eds) *Aspects of Vagueness*, (Dordrecht : Reidel, 1984), 75-90.
- ------. "A Geometry of Logic", Skala et al. (eds) *Aspects of Vagueness*, (Dordrecht : Reidel, 1984) 91-104.
- ------. *Critical Rationalism: A Restatement & Defence* (Chicago: Open Court , 1994).
- Newton-Smith, W.H. *The rationality of science* (Boston : Routledge & Kegan Paul, 1981)

- Niiniluoto, I. "On the Truthlikeness of Generalizations', in Butts et al (eds.) *Basic Problems in Methodology and Linguistics* (Dordrecht: Reidel, 1977), 121-147.
- ------. "Truthlikeness: Comments on Recent Discussion", *Synthese* 38 (1978), 281-329.
- ------. "Verisimilitude, Theory-Change, and Scientific Progress", . in Niiniluoto Iet al. (eds.) *The Logic and Epistemology of Scientific Change, Acta Philosophica Fennica*, vol. 30, Nos. 2-4, (Amsterdam: North-Holland, 1978,) 243-264.
- ------. "Truthlikeness in First-Order Languages", in Niiniluoto et al. (eds.) *Essays on Mathematical and Philosophical Logic*, (Dordrecht: Reidel, 1979), 437-458.
- ------. "Degrees of Truthlikeness: From Singular Sentences to Generalizations", *The British Journal for the Philosophy of Science* 30 (1979), 371-376.
- ------. "Scientific Progress" *Synthese* 45(3) (1982), 427-462.
- ------. "What Shall We Do With Verisimilitude?", *Philosophy of Science* 49 (1982), 181-197.
- ------. "Verisimilitude vs. Legisimilitude", *Studia Logica* XLII 2/3 (1983), 315-329.
- ------. "Truthlikeness and Bayesian Estimation", *Synthese* 67 (1986), 321-346.
- ------. "How to Define Verisimilitude", in Kuipers (1987), 11-23.
- ------. "Verisimilitude with Indefinite Truth", in Kuipers (1987), 187-195.
- ------. *Truthlikeness* (Dordrecht: Reidel, 1987)
- ------. "Reference Invariance and Truthlikeness", in *Philosophy of Science* 64 (1997), pp. 546-554.
- ------. "Survey Article: Verisimilitude: The Third Period", in *The British Journal for the Philosophy of Science* 49 (1998), pp. 1-29.
- ------. *Critical Scientific Realism*, (Oxford: Clarendon, 1999).
- Oddie, G. "Verisimilitude and distance in Logical Space" *Acta Philosophica Fennica* 30:2-4 (1978), 227-243.
- ------. "Verisimilitude reviewed" *The British Journal for the Philosophy of Science*, 32: 237-65, (1981).
- ------. "Cohen on verisimilitude and natural necessity" *Synthese*, 51: 355-79 (1982).
- ------. *Likeness to Truth Western Ontario Series in Philosophy of Science* (Dordrecht: Reidel, 1986).
- ------. "The poverty of the Popperian program for truthlikeness" *Philosophy of Science*, 53 (2): 163-78, (1986).
- ------. "The picture theory of truthlikeness" in Kuipers (1987a), 25-46.
- ------. "Truthlikeness and the convexity of propositions" in What is Closer-to-the-Truth? ed Theo Kuipers (Amsterdam: Rodopi, 1987b), 197-217.
- ------. "Verisimilitude by power relations" British Journal for the Philosophy of Science 41: 129-35, (1990).
- ------. "Truthlikeness" in Borchert, D.., ed.*The Encyclopedia of Philosophy Supplement* (New York: Macmillan, 1997), 574-6.
- ------. "Conditionalization, cogency and cognitive value" *The British Journal for the Philosophy of Science* 48 (1997), 533-41.
- ------. "Truth, Verification, Confirmation, Verisimilitude" in Niel J.Smelser et al eds. *International Encyclopedia of the Social and Behavioral Sciences* .(Oxford: Elselvier, 2001).
- Pearce, D. "Truthlikeness and translation: a comment on Oddie" *The British Journal for the Philosophy of Science* 34 (1983), 380-5.

- Popper, K. R. *Conjectures and Refutations* (London: Routledge, 1963).
- Rosencrantz, , R. "Truthlikeness: comments on David Miller" *Synthese* 30 (1/2), (1975), 193-7
- Ryan, M. and P.Y. Schobens "Belief revision and verisimilitude" *Notre Dame Journal of Formal Logic* 36 (1995), 15-29.
- Schurz, G., & Weingartner, P. "Verisimilitude defined by relevant consequence-elements" Kuipers, A.F.(1987), 47-77.
- Tarski, A. "The concept of truth in formalized languages" in Woodger, J. (ed.) *Logic, Semantics and Mathematics* (Oxford: Clarendon, 1969).
- Tichý, P. "On Popper's definitions of verisimilitude" *The British Journal for the Philosophy of Science* 25 (1974).
- ------. "Verisimilitude Redefined" *The British Journal for the Philosophy of Science* 27 (1976), 25-42
- ------. "Verisimilitude Revisited" *Synthese* 38 (1978), 175-196.
- Tooley, M "The nature of laws" *Canadian Journal of Philosophy* 7 (1977).
- Urbach, P. "Intimations of similarity: the shaky basis of verisimilitude" *The British Journal for the Philosophy of Science* 34 (1983), 166-75.
- Vetter, H. "A new concept of verisimilitude" in *Theory and Decision* 8 (1977), 369-75.
- Volpe, G. "A semantic approach to comparative verisimilitude" *The British Journal for the Philosophy of Science* 46 (1995), 563-82.
- Weston, T. "Approximate truth and scientific realism" *Philosophy of Science* 59, (1992), 53-74.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Popper, Karl | properties

Graham Oddie
Graham.Oddie@Colorado.edu

First published: July 10, 2001

Content last modified: July 10, 2001

Quantum Mechanics

Quantum mechanics is, at least at first glance and at least in part, a mathematical machine for predicting the behaviors of microscopic particles -- or, at least, of the measuring instruments we use to explore those behaviors -- and in that capacity, it is spectacularly successful: in terms of power and precision, head and shoulders above any theory we have ever had. Mathematically, the theory is well understood; we know what its parts are, how they are put together, and why, in the mechanical sense (i.e., in a sense that can be answered by describing the internal grinding of gear against gear), the whole thing performs the way it does, how the information that gets fed in at one end is converted into what comes out the other. The question of what kind of a world it describes, however, is controversial; there is very little agreement, among physicists and among philosophers, about what the world *is like* according to quantum mechanics. Minimally interpreted, the theory describes a set of facts about the way the microscopic world impinges on the macroscopic one, how it effects our measuring instruments, described in everyday language or the language of classical mechanics. Disagreement centers on the question of what a microscopic world, which affects our apparatuses in the prescribed manner, is, or even could be, like *intrinsically*; or how those apparatuses could themselves be built out of microscopic parts of the sort the theory describes.[1]

That is what an interpretation of the theory would provide: a proper account of what the world is like according to quantum mechanics, intrinsically and from the bottom up. The problems with giving an interpretation (not just a comforting, homey sort of interpretation, i.e., not just an interpretation according to which the world isn't too different from the familiar world of common sense, but any interpretation at all) are dealt with in other sections of this encyclopedia. Here, we are concerned only with the mathematical heart of the theory, the theory in its capacity as a mathematical machine, and -- whatever is true of the rest of it -- *this* part of the theory makes exquisitely good sense.

1. Terminology

Physical systems are divided into **types** according to their unchanging (or ‘state-independent’) properties, and the **state** of a system at a time consists of a complete specification of those of its properties that change with time (its ‘state-dependent’ properties). To give a complete description of a system, then, we need to say what type of system it is and what its state is at each moment in its history.

A **physical quantity** is a mutually exclusive and jointly exhaustive family of physical properties (for those who know this way of talking, it is a family of properties with the structure of the cells in a partition). Knowing what kinds of values a quantity takes can tell us a great deal about the relations among the properties of which it is composed. The values of a bivalent quantity, for instance, form a set with two members; the values of a real-valued quantity form a set with the structure of the real numbers. This is a special case of something we will see again and again, *viz.*, that knowing what kind of mathematical objects represent the elements in some set (here, the values of a physical quantity; later, the states that a system can assume, or the quantities pertaining to it) tells us a very great deal (indeed, arguably, all there is to know) about the relations among them.

In quantum mechanical contexts, the term ‘**observable**’ is used interchangeably with ‘physical quantity’, and should be treated as a technical term with the same meaning. It is no accident that the early developers of the theory chose the term, but the choice was made for reasons that are not, nowadays, generally accepted. The **state-space** of a system is the space formed by the set of its possible states,[2] i.e., the physically possible ways of combining the values of quantities that characterize it internally. In classical theories, a set of quantities which forms a supervenience basis for the rest is typically designated as ‘basic’ or ‘fundamental’, and, since any mathematically possible way of combining their values is a physical possibility, the state-space can be obtained by simply taking these as coordinates.[3] So, for instance, the state-space of a classical mechanical system composed of n particles, obtained by specifying the values of 6n real-valued quantities - three components of position, and three of momentum for each particle in the system - is a 6n-dimensional coordinate space. Each possible state of such a system corresponds to a point in the space, and each point in the space corresponds to a possible state of such a system. The situation is a little different in quantum mechanics, where there are mathematically describable ways of combining the values of the quantities that don't represent physically possible states. As we will see, the state-spaces of quantum mechanics are special kinds of vector spaces, known as Hilbert spaces, and they have more internal structure than their classical counterparts.

A **structure** is a set of elements on which certain operations and relations are defined, a **mathematical structure** is just a structure in which the elements are mathematical objects (numbers, sets, vectors) and the operations mathematical ones, and a **model** is a mathematical structure used to represent some physically significant structure in the world.

The heart and soul of quantum mechanics is contained in the Hilbert spaces that represent the state-spaces of quantum mechanical systems. The internal relations among states and quantities, and

everything this entails about the ways quantum mechanical systems behave, are all woven into the structure of these spaces, embodied in the relations among the mathematical objects which represent them.[4] This means that understanding what a system is like according to quantum mechanics is inseparable from familiarity with the internal structure of those spaces. Know your way around Hilbert space, and become familiar with the dynamical laws that describe the paths that vectors travel through it, and you know everything there is to know, in the terms provided by the theory, about the systems that it describes.

By ‘know your way around’ Hilbert space, I mean something more than possess a description or a map of it; anybody who has a quantum mechanics textbook on their shelf has that. I mean know your way around it in the way you know your way around the city in which you live. This is a practical kind of knowledge that comes in degrees and it is best acquired by learning to solve problems of the form: How do I get from A to B? Can I get there without passing through C? And what is the shortest route? Graduate students in physics spend long years gaining familiarity with the nooks and crannies of Hilbert space, locating familiar landmarks, treading its beaten paths, learning where secret passages and dead ends lie, and developing a sense of the overall lay of the land. They learn how to navigate Hilbert space in the way a cab driver learns to navigate his city.

How much of this kind of knowledge is needed to approach the philosophical problems associated with the theory? In the beginning, not very much: just the most general facts about the geometry of the landscape (which is, in any case, unlike that of most cities, beautifully organized), and the paths that (the vectors representing the states of) systems travel through them. That is what will be introduced here: first a bit of easy math, and then, in a nutshell, the theory.

2. Mathematics

Vectors and vector spaces

A **vector** A, written ‘|A>’, is a mathematical object characterized by a length, |A|, and a direction. A normalized vector is a vector of length 1; i.e., |A| = 1. Vectors can be added together, multiplied by constants (including complex numbers), and multiplied together. Vector addition maps any pair of vectors onto another vector, specifically, the one you get by moving the second vector so that it's tail coincides with the tip of the first, without altering the length or direction of either, and then joining the tail of the first to the tip of the second. This addition rule is known as the parallelogram law. So, for example, adding vectors |A> and |B> yields vector |C> (= |A> + |B>) as in Figure 1:

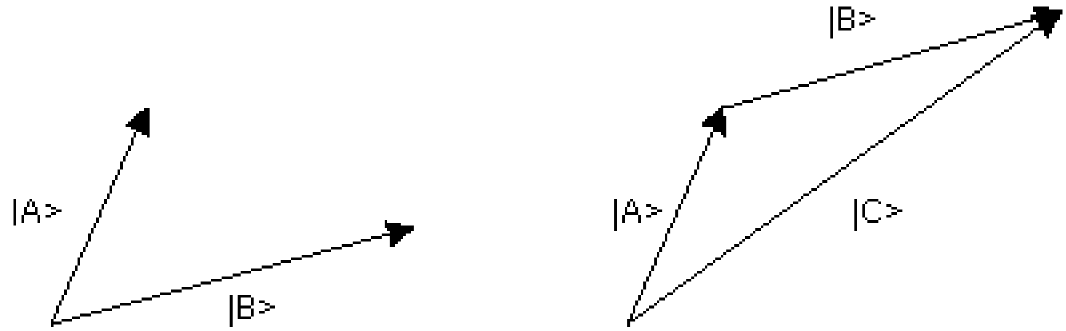


Figure 1: Vector Addition

Multiplying a vector |A> by *n*, where *n* is a constant, gives a vector which is the same direction as |A> but whose length is *n* times |A>'s length.

In a real vector space, the (inner or dot) product of a pair of vectors |A> and |B>, written ‘<A|B>’ is a scalar equal to the product of their lengths (or ‘norms’) times the cosine of the angle, θ, between them:

$$<A|B> = |A|\ |B| \cos \theta$$

Let $|A_1>$ and $|A_2>$ be vectors of length 1 ("unit vectors") such that $<A_1|A_2> = 0$. (So the angle between these two unit vectors must be 90 degrees.) Then we can represent an arbitrary vector |B> in terms of our unit vectors as follows:

$$|B> = b_1|A_1> + b_2|A_2>$$

For example, here is a graph which shows how |B> can be represented as the sum of the two unit vectors $|A_1>$ and $|A_2>$:

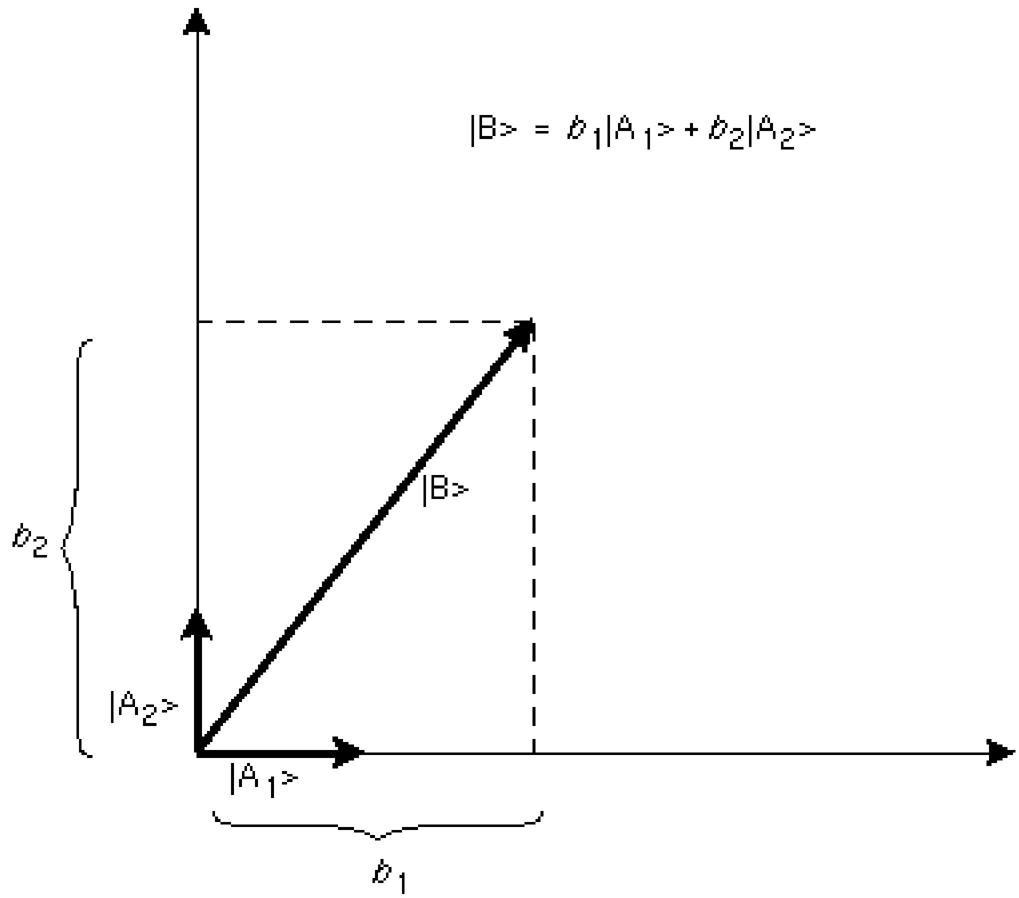


Figure 2: Representing |B> by Vector Addition of Unit Vectors

Now the definition of the inner product <A|B> has to be modified to apply to complex spaces. Let *c** be the complex conjugate of *c*. (When *c* is a complex number of the form *a* ± *bi*, then the complex conjugate *c** of *c* is defined as follows:

$[a + bi]^* = a - bi$
$[a - bi]^* = a + bi$

So, for all complex numbers *c*, [*c**]* = *c*, but *c** = *c* just in case *c* is real.) Now definition of the inner product of |A> and |B> for complex spaces can be given in terms of the conjugates of complex coefficients as follows. Where $|A_1>$ and $|A_2>$ are the unit vectors described earlier, $|A> = a_1|A_1> + a_2|A_2>$ and $|B> = b_1|A_1> + b_2|A_2>$, then

$<A|B> = (a_1^*)(b_1) + (a_2^*)(b_2)$

The most general and abstract notion of an inner product, of which we've now defined two special cases, is as follows. <A|B> is an inner product on a vector space V just in case

(i) $<A|A> = |A|^2$, and <A|A>=0 if and only if A=0

(ii) $<B|A> = <A|B>^*$

(iii) $<B|A+C> = <B|A> + <B|C>$.

It follows from this that

(i) the length of $|A>$ is the square root of inner product of $|A>$ with itself, i.e.,

$$|A| = \sqrt{<A|A>},$$

and

(ii) $|A>$ and $|B>$ are mutually perpendicular, or **orthogonal,** if, and only if, $<A|B> = 0$.

A **vector space** is a set of vectors closed under addition, and multiplication by constants, **an inner product space** is a vector space on which the operation of vector multiplication has been defined, and the **dimension** of such a space is the maximum number of nonzero, mutually orthogonal vectors it contains.

Any collection of N mutually orthogonal vectors of length 1 in an N-dimensional vector space constitutes an **orthonormal basis** for that space. Let $|A_1>, \ldots , |A_N>$ be such a collection of unit vectors. Then every vector in the space can be expressed as a sum of the form:

$$|B> = b_1|A_1> + b_2|A_2> + \ldots + b_N|A_N>,$$

where $b_i = <B|A_i>$. The b_i's here are known as B's **expansion coefficients** in the A-basis.[5]

Notice that:

(i) for all vectors A, B, and C in a given space,

$$<A|B+C> = <A|B> + <A|C>$$

(ii) for any vectors M and Q, expressed in terms of the A-basis,

$$|M> + |Q> = (m_1 + q_1)|A_1> + (m_2 + q_2)|A_2> + \ldots + (m_N + q_N)|A_N>,$$

and

$$<M|Q> = m_1q_1 + m_2q_2 + \ldots + m_nq_n$$

There is another way of writing vectors, namely by writing their expansion coefficients (relative to a given basis) in a column, like so:

$$|Q> = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$$

where $q_i = <Q|A_i>$ and the A_i are the chosen basis vectors.

When we are dealing with vector spaces of infinite dimension, since we can't write the whole column of expansion coefficients needed to pick out a vector since it would have to be infinitely long, so instead we write down the function (called the ‘wave function’ for Q, usually represented $\psi(i)$) which has those coefficients as values. We write down, that is, the function:

$$\psi(i) = q_i = <Q|A_i>$$

Given any vector in, and any basis for, a vector space, we can obtain the wave-function of the vector in that basis; and given a wave-function for a vector, in a particular basis, we can construct the vector whose wave-function it is. Since it turns out that most of the important operations on vectors correspond to simple algebraic operations on their wave-functions, this is the usual way to represent state-vectors.

When a pair of physical systems interact, they form a composite system, and, in quantum mechanics as in classical mechanics, there is a rule for constructing the state-space of a composite system from those of its components, a rule that tells us how to obtain, from the state-spaces, H_A and H_B for A and B, respectively, the state-space -- called the ‘tensor product’ of H_A and H_B, and written $H_A \otimes H_B$ -- of the pair. There are two important things about the rule; first, so long as H_A and H_B are Hilbert spaces, $H_A \otimes H_B$ will be as well, and second, there are some facts about the way $H_A \otimes H_B$ relates to H_A and H_B, that have surprising consequences for the relations between the complex system and its parts. In particular, it turns out that the state of a composite system is not uniquely defined by those of its components. What this means, or at least what it appears to mean, is that there are, according to quantum mechanics, facts about composite systems (and not just facts about their spatial configuration) that don't supervene on facts about their components; it means that there are facts about systems as wholes that don't supervene on facts about their parts and the way those parts are arranged in space. The significance of this feature of the theory cannot be overplayed; it is, in one way or another, implicated in most of its most difficult problems.

In a little more detail: if $\{v_i^A\}$ is an orthonormal basis for H_A and $\{u_j^B\}$ is an orthonormal basis for H_B, then the set of pairs (v_i^A, u_j^B) is taken to form an orthonormal basis for the tensor product space $H_A \otimes H_B$. The notation $v_i^A \otimes u_j^B$ is used for the pair (v_i^A, u_j^B), and inner product on $H_A \otimes H_B$ is defined as:[6]

$$<v_i^A \otimes u_m^B \mid v_j^A \otimes u_n^B> \; = \; <v_i^A \mid v_j^A> \; <u_m^B \mid u_n^B>$$

It is a result of this construction that although every vector in $H_A \otimes H_B$ is a linear sum of vectors expressible in the form $v^A \otimes u^B$, not every vector in the space is itself expressible in that form, and it turns out that

(i) any composite state defines uniquely the states of its components.

(ii) if the states of A and B are pure (i.e., representable by vectors v^A and u^B, respectively), then the state of (A+B) is pure and represented by $v^A \otimes u^B$, and

(iii) if the state of (A+B) is pure and expressible in the form $v^A \otimes u^B$, then the states of A and B are pure, but

(iv) if the states of A and B are not pure, i.e., if they are mixed states (these are defined below), they do not uniquely define the state of (A+B); in particular, it may be a pure state not expressible in the form $v^A \otimes u^B$.

Operators

An **operator** O is a mapping of a vector space onto itself; it takes any vector |B> in a space onto another vector |B′> also in the space; O|B> = |B′>. **Linear operators** are operators that have the following properties:

(i) O(|A> + |B>) = O|A> + O|B>, and

(ii) O(*c*|A>) = *c*(O|A>).

Just as any vector in an N-dimensional space can be represented by a column of N numbers, relative to a choice of basis for the space, any linear operator on the space can be represented in a column notation by N^2 numbers:

$$O = \begin{bmatrix} O_{11} & O_{12} \\ O_{21} & O_{22} \end{bmatrix}$$

where $O_{ij} = < A_i \mid O|A_j>$ and the $|A_N>$ are the basis vectors of the space. The effect of the linear operator O on the vector B is, then, given by

O|B>=

$$= \begin{bmatrix} O_{11} & O_{12} \\ O_{21} & O_{22} \end{bmatrix} \times \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$= \begin{bmatrix} (O_{11}b_1 + O_{12}b_2) \\ (O_{21}b_1 + O_{22}b_2) \end{bmatrix}$$

$$= (O_{11}b_1 + O_{12}b_2)|A_1\rangle + (O_{21}b_1 + O_{22}b_2)|A_2\rangle$$

$$= |B'\rangle$$

Two more definitions before we can say what Hilbert spaces are, and then we can turn to quantum mechanics. |B> is an **eigenvector** of O with eigenvalue *a* if, and only if, O|B> = *a*|B>. Different operators can have different eigenvectors, but the eigenvector/operator relation depends only on the operator and vectors in question, and not on the particular basis in which they are expressed; the eigenvector/operator relation is, that is to say, invariant under change of basis. **Hermitean operators** are linear operators, which have only real eigenvalues.

A **Hilbert space**, finally, is a vector space on which an inner product is defined, and which is complete, i.e., which is such that any Cauchy sequence of vectors in the space converges to a vector in the space. All finite-dimensional inner product spaces are complete, and I will restrict myself to these. The infinite case involves some complications that are not fruitfully entered into at this stage.

3. Quantum Mechanics

Four basic principles of quantum mechanics are:

3.1 Physical States

Every physical system is associated with a Hilbert Space, every unit vector in the space corresponds to a possible pure state of the system, and every possible pure state, to some vector in the space.[7] In standard texts on quantum mechanics, the vector is represented by a function known as the wave-function, or ψ-function.

3.2 Physical Quantities

Hermitian operators in the Hilbert space associated with a system represent physical quantities, and their eigenvalues represent the possible results of measurements of those quantities.

3.3 Composition

The Hilbert space associated with a complex system is the tensor product of those associated with the simple systems (in the standard, non-relativistic, theory: the individual particles) of which it is composed.

3.4 Dynamics

a. *Contexts of type 1*: Given the state of a system at t and the forces and constraints to which it is subject, there is an equation, '**Schrödinger's equation**', that gives the state at any other time U|v_t> → |v_t′>.[8] The important properties of U for our purposes are that it is **deterministic,** which is to say that it takes the state of a system at one time into a unique state at any other, and it is **linear**, which is to say that if it takes a state |A> onto the state |A′>, and it takes the state |B> onto the state |B′>, then it takes any state of the form α|A> + β|B> onto the state α|A′> + β|B′>.

b. *Contexts of type 2* ("Measurement Contexts"):[9] Carrying out a "measurement" of an observable B on a system in a state |A> has the effect of collapsing the system into a B-eigenstate corresponding to the eigenvalue observed. This is known as the **Collapse Postulate**. Which *particular* B-eigenstate it collapses into is a matter of probability, and the probabilities are given by a rule known as **Born's Rule**:

$$prob(b_i) = |\langle A|B=b_i\rangle|^2.$$

There are two important points to note about these two kinds of contexts:

- The distinction between contexts of type 1 and 2 remains to be made out in quantum mechanical terms; nobody has managed to say in a completely satisfactory way, in the terms provided by the theory, which contexts are measurement contexts, and
- Even if the distinction is made out, it is an open interpretive question whether there *are* contexts of type 2; i.e., it is an open interpretive question whether there are any contexts in which systems are governed by a dynamical rule *other* than Schrödinger's equation.

4. Structures on Hilbert Space

I remarked above that in the same way that all the information we have about the relations between locations in a city is embodied in the spatial relations between the points on a map which represent them, all of the information that we have about the internal relations among (and between) states and quantities in quantum mechanics is embodied in the mathematical relations among the vectors and operators which represent them.[10] From a mathematical point of view, what really distinguishes quantum mechanics from its classical predecessors is that states and quantities have a richer structure; they form families with a more interesting network of relations among their members.

All of the physically consequential features of the behaviors of quantum mechanical systems are consequences of mathematical properties of those relations, and the most important of them are easily summarized:

> (P1) Any way of adding vectors in a Hilbert space or multiplying them by scalars will yield a vector that is also in the space. In the case that the vector is normalized, it will, from (3.1), represent a possible state of the system, and in the event that it is the sum of a pair of eigenvectors of an observable B with distinct eigenvalues, it will not itself be an eigenvector of B, but will be associated, from (3.4b), with a set of probabilities for showing one or another result in B-measurements.
>
> (P2) For any Hermitian operator on a Hilbert space, there are others, on the same space, with which it doesn't share a full set of eigenvectors; indeed, it is easy to show that there are other such operators with which it has *no* eigenvectors in common.

If we make a couple of additional interpretive assumptions, we can say more. Assume, for instance, that

> (4.1) Every Hermitian operator on the Hilbert space associated with a system represents a distinct observable, and (hence) every normalized vector, a distinct state, and
>
> (4.2) A system has a value for observable A if, and only if, the vector representing its state is an eigenstate of the A-operator. The value it has, in such a case, is just the eigenvalue associated with that eigenstate.[11]

It follows from (P2), by (3.1), that no quantum mechanical state is an eigenstate of all observables (and indeed that there are observables which have *no* eigenstates in common), and so, by (3.2), that no quantum mechanical system ever has simultaneous values for all of the quantities pertaining to it (and indeed that there are pairs of quantities to which *no* state assigns simultaneous values).

There are Hermitian operators on the tensor product $H_1 \otimes H_2$ of a pair of Hilbert spaces H_1 and H_2 ... In the event that H_1 and H_2 are the state spaces of systems S1 and S2, $H_1 \otimes H_2$ is the state-space of the complex system (S1+S2). It follows from this by (4.1) that there are observables pertaining to (S1+S2) whose values are not determined by the values of observables pertaining to the two individually.

These are all straightforward consequences of taking vectors and operators in Hilbert space to represent, respectively, states and observables, and applying Born's Rule (and later (4.1) and (4.2)), to give empirical meaning to state assignments. That much is perfectly well understood; the real difficulty in understanding quantum mechanics lies in coming to grips with their implications -- physical, metaphysical, and epistemological.

There is one remaining fact about the mathematical structure of the theory that anyone trying to come to an understanding about what it says about the world has to grapple with. It is not a property of Hilbert spaces, this time, but of the dynamics, the rules that describe the trajectories that systems follow through the space. From a physical point of view, it is far more worrisome than anything that has preceded. For, it does much more than present difficulties to someone trying to provide an *interpretation* of the theory, it seems to point either to a logical inconsistency in the theory's foundations.

Suppose that we have a system S and a device S* which measures an observable A on S with values $\{a_1, a_2, a_3 ...\}$. Then there is some state of S* (the ‘ground state’), and some observable B with values $\{b_1, b_2, b_3 ...\}$ pertaining to S* (its ‘pointer observable’, so called because it is whatever plays the role of the pointer on a dial on the front of a schematic measuring instrument in registering the result of the experiment), which are such that, if S* is started in its ground state and interacts in an appropriate way with S, and if the value of A immediately before the interaction is a_1, then B's value immediately thereafter is b_1. If, however, A's value immediately before the interaction is a_2, then B's value afterwards is b_2; if the value of A immediately before the interaction is a_3, then B's value immediately after is b_3, and so on. That is just what it *means* to say that S* measures A. So, if we represent the joint, partial state of S and S* (just the part of it which specifies the value of [A on S & B on S*], the observable whose values correspond to joint assignments of values to the measured observable on S and the pointer observable on S*) by the vector $|A=a_i>_s|B=b_i>_{s*}$, and let "$\rightarrow$" stand in for the dynamical description of the interaction between the two, to say that S* is a measuring instrument for A is to say that the dynamical laws entail that,

$$|A=a_1>_s|B=\text{ground state}>_{s*} \rightarrow |A=a_1>_s|B=b_1>_{s*}$$

$$|A=a_2>_s|B=\text{ground state}>_{s*} \rightarrow |A=a_2>_s|B=b_2>_{s*}$$

$$|A=a_3>_s|B=\text{ground state}>_{s*} \rightarrow |A=a_3>_s|B=b_3>_{s*}$$

and so on.[12]

Intuitively, S* is a measuring instrument for an observable A just in case there is some observable feature of S* (it doesn't matter what, just something whose values can be ascertained by looking at the device), which is correlated with the A-values of systems fed into it in such a way that we can read those values off of S*'s observable state after the interaction. In philosophical parlance, S* is a measuring instrument for A just in case there is some observable feature of S* which *tracks* or *indicates* the A-values of systems with which it interacts in an appropriate way.

Now, it follows from (3.1), above, that there are states of S (too many to count) which are not eigenstates of A, and if we consider what Schrödinger's equation tells us about the joint evolution of S and S* when S is started out in one of these, we find that the state of the pair after interaction is a superposition of

eigenstates of [A on S & B on S*]. It doesn't matter what observable on S is being measured, and it doesn't matter what particular superposition S starts out in; when it is fed into a measuring instrument for that observable, if the interaction is correctly described by Schrödinger's equation, it follows just from the linearity of the U in that equation, the operator that effects the transformation from the earlier to the later state of the pair, that the joint state of S and the apparatus after the interaction is a superposition of eigenstates of this observable on the joint system.

Suppose, for example, that we start S* in its ground state, and S in the state

$$1/\sqrt{2}|A=a_1>_s| + 1/\sqrt{2}|A=a_2>_s$$

It is a consequence of the rules for obtaining the state-space of the composite system that the combined state of the pair is

$$1/\sqrt{2}|A=a_1>_s|B=\text{ground state}>_{s*} + 1/\sqrt{2}|A=a_2>_s|B=\text{ground state}>_{s*}$$

and it follows from the fact that S* is a measuring instrument for A, and the linearity of U that their combined state *after* interaction, is

$$1/\sqrt{2}|A=a_1>_s|B=b_1>_{s*} + 1/\sqrt{2}|A=a_2>_s|B=b_2>_{s*}$$

This, however, is inconsistent with the dynamical rule for contexts of type 2, for the dynamical rule for contexts of type 2 (and if there are any such contexts, *this* is one) entails that the state of the pair after interaction is *either*

$$|A=a_1>_s|B=b_1>_{s*}$$

or

$$|A=a_2>_s|B=b_2>_{s*}$$

Indeed, it entails that there is a precise probability of 1/2 that it will end up in the former, and a probability of 1/2 that it will end up in the latter.

We can try to restore logical consistency by giving up the dynamical rule for contexts of type 2 (or, what amounts to the same thing, by denying that there *are* any such contexts), but then we have the problem of consistency with experience. For it was no mere blunder that that rule was included in the theory; we *know* what a system looks like when it is in an eigenstate of a given observable, and we know *from looking* that the measuring apparatus after measurement is in an eigenstate of the pointer observable. And so we *know* from the outset that if a theory tells us something else about the post-measurement states of

measuring apparatuses, whatever that something else is, it is wrong.

That, in a nutshell, is the Measurement Problem in quantum mechanics; any interpretation of the theory, any detailed story about what the world is like according to quantum mechanics, and in particular those bits of the world in which measurements are going on, has to grapple with it.

Loose Ends

Mixed states are weighted sums of pure states, and they can be used to represent the states of ensembles whose components are in different pure states, or states of individual systems about which we have only partial knowledge. In the first case, the weight attached to a given pure state reflects the size of the component of the ensemble which is in that state (and hence the objective probability that an arbitrary member of the ensemble is); in the second case, they reflect the epistemic probability that the system in question to which the state is assigned is in that state.

If we don't want to lose the distinction between pure and mixed states, we need a way of representing the weighted sum of a set of pure states (equivalently, of the probability functions associated with them) that is different from adding the (suitably weighted) vectors that represent them, and that means that we need either an alternative way of representing mixed states, or a uniform way of representing both pure and mixed states that preserves the distinction between them. There is a kind of operator in Hilbert spaces, called a **density operator**, that serves well in the latter capacity, and it turns out not to be hard to restate everything that has been said about state vectors in terms of density operators. So, even though it is common to speak as though pure states are represented by vectors, the official rule is that states – pure and mixed, alike - are represented in quantum mechanics by density operators.

Although mixed states *can*, as I said, be used to represent our ignorance of the states of systems that are actually in one or another pure state, and although this has seemed to many to be an adequate way of interpreting mixtures in classical contexts, there are serious obstacles to applying it generally to quantum mechanical mixtures. These are left for detailed discussion in the other entries on quantum mechanics in the Encyclopedia.

Everything that has been said about observables, strictly speaking, applies only to the case in which the values of the observable form a discrete set; the mathematical niceties that are needed to generalize it to the case of **continuous observables** are complicated, and raise problems of a more technical nature. These, too, are best left for detailed discussion.

This should be all the initial preparation one needs to *approach* the philosophical discussion of quantum mechanics, but it is only a first step. The more one learns about the relationships among and between vectors and operators in Hilbert space, about how the spaces of simple systems relate to those of complex ones, and about the equation which describes how state-vectors move through the space, the better will be one's appreciation of both the nature and the difficulty of the problems associated with the theory. The funny backwards thing about quantum mechanics, the thing that makes it endlessly absorbing to a

philosopher, is that the more one learns, the harder the problems get.

Bibliography

- Albert, D., 1992, *Quantum Mechanics and Experience*, Cambridge, MA: Harvard University Press
- Halmos, P., 1957, *Introduction to Hilbert Space*, 2nd edition, Providence: AMS Chelsea Publishing

Other Internet Resources

- Preskill, J., 1998, Quantum Computation (Lecture Notes for Physics 219, California Institute of Technology)

Related Entries

quantum mechanics: Bohmian mechanics | quantum mechanics: collapse theories | quantum mechanics: Copenhagen interpretation of | quantum mechanics: Everett's relative-state formulation of | quantum mechanics: Kochen-Specker theorem | quantum mechanics: many-worlds interpretation of | quantum mechanics: modal interpretations of | quantum mechanics: relational | quantum mechanics: the role of decoherence in

Jenann Ismael
jtismael@u.arizona.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 28, 2000
Content last modified: November 28, 2000

Stanford Encyclopedia of Philosophy
Notes to Quantum Mechanics

Notes

1. Indeed, when pressed, we find we can't even say explicitly (in the terms provided by the theory, in terms that apply directly to the entities, quantities, and relations of which the world is, by its lights, composed) which systems count as macroscopic (or what would be just as good, which are ‘classical’, which are fit to act as measuring apparatuses, or which interactions count as measurements).

2. It is also sometimes used to refer to a mathematical model that represents that space, a mathematical model that provides a kind of map of the set of possible states.

3.There will, of course, be arbitrariness in the coordinate description of the state-space, thus obtained, precisely as much arbitrariness as there is in the choice of mathematical names for the values of basic quantities.

4. Another way to put this: if you consider the set of states associated with any quantum mechanical system, you would find that it had the structure of the set of vectors in a Hilbert space.

5. Notice the relations between bases and coordinate systems; vectors of norm 1 pointing along the perpendicular coordinate axes of an N-dimensional point space constitute an orthonormal basis for the associated N-dimensional vector space, and there are as many bases for the vector space as possible coordinatizations of the point space.

6.The reader can satisfy herself that since (v_i^A, u_j^B) spans $H_A \otimes H_B$, the equation defines, by linearity, an inner product on the whole space, and also that, since $|v| = 0$ just in case v is the zero vector, it follows that $v^A \otimes 0 = 0 = 0 \otimes u^B$.

7. The correspondence isn't unique; any vectors |A> and @|A> where @ is any complex number of absolute value 1 correspond to the same state. There may be other redundancies; some interpretations, for instance, invoke what are called superselection rules that identify states represented by distinct normalized vectors. (i) is modified below, when mixed states are defined.

8. The equation is usually expressed in the form $i\, dv/dt = \mathbf{H}v$, where v is the system's state vector and **H** the operator representing its energy, but, again, the details aren't immediately important.

9. The quotes are to recommend caution about reading too much of one's ordinary understanding of this

word into its use in quantum mechanics; one usually thinks of measurement as a way of obtaining information about a system, but the only information one takes away from an individual quantum-mechanical ‘measurement’ about the state of the measured system before the interaction is that it was not (or, at least, there is a measure zero probability that it was) in an eigenstate of the measured observable with an eigenvalue other than the one observed. Of course, if the Collapse Postulate is correct, one knows the state of the system after measurement, but if I shoot someone at close range, I can be pretty sure that they are dead afterward; that doesn't make the interaction a measurement of the state of their health.

10. The two are not independent, of course; a system's state is just a specification of the values of those quantities pertaining to it that change over time. Internal relations between a set of elements are relations that supervene on their intrinsic natures; they do not include nomological relations.

11. These are strong assumptions: the first denies the existence of superselection rules, and the second, known as the eigenstate-eigenvalue link, is the principle that defines the orthodox Copenhagen interpretation of the theory. I make them here only to illustrate some of the implications of (P2) in an interpretive context.

12. It is inessential that the reduced state of S after the interaction be what it was beforehand; that is so only in the special case of measurements that leave the value of the measured observable undisturbed.

Jenann Ismael
jtismael@u.arizona.edu

First published: November 28, 2000
Content last modified: November 28, 2000

Bohmian Mechanics

Bohmian mechanics, which is also called the de Broglie-Bohm theory, the pilot-wave model, and the causal interpretation of quantum mechanics, is a version of quantum theory discovered by Louis de Broglie in 1927 and rediscovered by David Bohm in 1952. It is the simplest example of what is often called a hidden variables interpretation of quantum mechanics. In Bohmian mechanics a system of particles is described in part by its wave function, evolving, as usual, according to Schrödinger's equation. However, the wave function provides only a partial description of the system. This description is completed by the specification of the actual positions of the particles. The latter evolve according to the "guiding equation," which expresses the velocities of the particles in terms of the wave function. Thus, in Bohmian mechanics the configuration of a system of particles evolves via a deterministic motion choreographed by the wave function. In particular, when a particle is sent into a two-slit apparatus, the slit through which it passes and where it arrives on the photographic plate are completely determined by its initial position and wave function.

Bohmian mechanics inherits and makes explicit the nonlocality implicit in the notion, common to just about all formulations and interpretations of quantum theory, of a wave function on the configuration space of a many-particle system. It accounts for all of the phenomena governed by nonrelativistic quantum mechanics, from spectral lines and scattering theory to superconductivity, the quantum Hall effect and quantum computing. In particular, the usual measurement postulates of quantum theory, including collapse of the wave function and probabilities given by the absolute square of probability amplitudes, emerge from an analysis of the two equations of motion - Schrödinger's equation and the guiding equation - without the traditional invocation of a special, and somewhat obscure, status for observation.

---

1. The Completeness of the Quantum Mechanical Description

Despite its extraordinary predictive successes, quantum mechanics has, since its inception some seventy years ago, been plagued by conceptual difficulties. The basic problem, plainly put, is this: It is not at all clear what quantum mechanics is about. What, in fact, does quantum mechanics describe?

It might seem, since it is widely agreed that any quantum mechanical system is completely described by its wave function, that quantum mechanics is fundamentally about the behavior of wave functions. Quite naturally, no physicist wanted this to be true more than did Erwin Schrödinger, the father of the wave function. Nonetheless, Schrödinger ultimately found this impossible to believe. His difficulty was not so much with the novelty of the wave function (Schrödinger 1935): "That it is an abstract, unintuitive mathematical construct is a scruple that almost always surfaces against new aids to thought and that carries no great message." Rather, it was that the "blurring" suggested by the spread out character of the wave function "affects macroscopically tangible and visible things, for which the term ‘blurring’ seems simply wrong."

For example, in the same paper Schrödinger noted that it may happen in radioactive decay that

> the emerging particle is described ... as a spherical wave ... that impinges continuously on a surrounding luminescent screen over its full expanse. The screen however does not show a more or less constant uniform surface glow, but rather lights up at *one* instant at *one* spot ....

And he observed that one can easily arrange, for example by including a cat in the system, "quite ridiculous cases" with

> the ψ-function of the entire system having in it the living and the dead cat (pardon the expression) mixed or smeared out in equal parts.

It is thus because of the "measurement problem," of macroscopic superpositions, that Schrödinger found it difficult to regard the wave function as "representing reality." But then what does? With evident disapproval, Schrödinger describes how

> the reigning doctrine rescues itself or us by having recourse to epistemology. We are told that no distinction is to be made between the state of a natural object and what I know about it, or perhaps better, what I can know about it if I go to some trouble. Actually -- so they say -- there is intrinsically only awareness, observation, measurement.

Many physicists pay lip service to the Copenhagen interpretation -- that quantum mechanics is fundamentally about observation or results of measurement. But it is becoming increasingly difficult to find any who, when pressed, will defend this interpretation. It seems clear that quantum mechanics is fundamentally about atoms and electrons, quarks and strings, not those particular macroscopic regularities associated with what we call *measurements* of the properties of these things. But if these entities are not to be somehow identified with the wave function itself -- and if talk of them is not merely shorthand for elaborate statements about measurements -- then where are they to be found in the quantum description?

There is, perhaps, a very simple reason why there has been so much difficulty discerning in the quantum description the objects we believe quantum mechanics should be describing. Perhaps the quantum mechanical description is not the whole story, a possibility most prominently associated with Albert Einstein.

In 1935 Einstein, Boris Podolsky and Nathan Rosen argued for this possibility in the famous EPR paper (Einstein et al. 1935), which they concluded with the following:

> While we have thus shown that the wave function does not provide a complete description of the physical reality, we left open the question of whether or not such a description exists. We believe, however, that such a theory is possible.

The argument given in the EPR paper for this conclusion invoked quantum correlations and an assumption of locality. (See the entry on quantum entanglement and information.)

Later, on the basis of more or less the same considerations as those of Schrödinger quoted above, Einstein again concluded that the wave function does not provide a complete description of individual systems, an idea he called "this most nearly obvious interpretation" (Einstein 1949, p. 672). In relation to a theory incorporating a more complete description, Einstein remarked that "the statistical quantum theory would ... take an approximately analogous position to the statistical mechanics within the framework of classical mechanics." It is perhaps worth noting here that Bohmian mechanics, as we shall see, exactly fits this description.

[Return to Table of Contents]

2. The Impossibility of Hidden Variables ... or the Inevitability of Nonlocality?

John von Neumann, one of the greatest mathematicians of the twentieth century, claimed to have mathematically proven that Einstein’s dream, of a deterministic completion or reinterpretation of quantum theory, was mathematically impossible. He concluded that (von Neumann 1932, p. 325 of the English translation)

> It is therefore not, as is often assumed, a question of a re-interpretation of quantum mechanics -- the present system of quantum mechanics would have to be objectively false, in order that another description of the elementary processes than the statistical one be possible.

This claim of von Neumann was almost universally accepted among physicists and philosophers of science. For example, Max Born, who formulated the statistical interpretation of the wave function, assured us that (Born 1949, p. 109)

> No concealed parameters can be introduced with the help of which the indeterministic description could be transformed into a deterministic one. Hence if a future theory should be deterministic, it cannot be a modification of the present one but must be essentially different.

Bohmian mechanics is, quite clearly, a counterexample to the claims of von Neumann, so something has to be wrong with von Neumann’s argument. In fact, according to John Bell (Mermin 1993, p. 805), von Neumann’s assumptions (about the relationships among the values of quantum observables that must be satisfied in a hidden-variables theory) are so unreasonable that the "the proof of von Neumann is not merely false but *foolish*!" Nonetheless, some physicists continue to rely on von Neumann’s proof, although in recent years it is more common to find physicists citing the Kochen-Specker Theorem and, more frequently, Bell’s inequality as the basis of this refutation. We still find, a quarter of a century after the rediscovery of Bohmian mechanics in 1952, statements such as these (Wigner 1976):

> The proof he [von Neumann] published ..., though it was made much more convincing later on by Kochen and Specker, still uses assumptions which, in my opinion, can quite reasonably be questioned. ... In my opinion, the most convincing argument against the theory of hidden variables was presented by J. S. Bell (1964).

Now there are many more statements of a similar character that could have been cited. This quotation owes its significance to the fact that Wigner was not only one of the leading physicists of his generation, but, unlike most of his contemporaries, he was also profoundly concerned with the conceptual

foundations of quantum mechanics and wrote on the subject with great clarity and insight.

There was, however, one physicist who wrote on this subject with even greater clarity and insight than Wigner himself, namely the very J. S. Bell whom Wigner praises for demonstrating the impossibility of a deterministic completion of quantum theory such as Bohmian mechanics. Here's how Bell himself reacted to Bohm's discovery (Bell 1987, p. 160):

> But in 1952 I saw the impossible done. It was in papers by David Bohm. Bohm showed explicitly how parameters could indeed be introduced, into nonrelativistic wave mechanics, with the help of which the indeterministic description could be transformed into a deterministic one. More importantly, in my opinion, the subjectivity of the orthodox version, the necessary reference to the 'observer,' could be eliminated. ...
>
> But why then had Born not told me of this 'pilot wave'? If only to point out what was wrong with it? Why did von Neumann not consider it? More extraordinarily, why did people go on producing ''impossibility'' proofs, after 1952, and as recently as 1978? ... Why is the pilot wave picture ignored in text books? Should it not be taught, not as the only way, but as an antidote to the prevailing complacency? To show us that vagueness, subjectivity, and indeterminism, are not forced on us by experimental facts, but by deliberate theoretical choice?

Wigner to the contrary notwithstanding, Bell did not establish the impossibility of a deterministic reformulation of quantum theory, nor did he ever claim to have done so. On the contrary, over the course of the past several decades, until his untimely death in 1990, Bell was the prime proponent, for a good part of this period almost the sole proponent, of the very theory, Bohmian mechanics, that he is supposed to have demolished.

Bohmian mechanics is of course as much a counterexample to the Kochen-Specker argument for the impossibility of hidden variables as it is to the one of von Neumann. It is obviously a counterexample to any such argument. However reasonable the assumptions of such an argument, some of them must fail for Bohmian mechanics.

Wigner was quite right to suggest that the assumptions of Kochen and Specker are more convincing than those of von Neumann. They appear, in fact, to be quite reasonable indeed . However, they are not. The impression that they are arises from a pervasive error, a naive realism about operators, that will be discussed below in the sections on quantum observables, on spin, and on contextuality.

One of the achievements of John Bell was to replace the "arbitrary axioms" (Bell 1987, page 11) of Kochen-Specker and others by an assumption of locality, of no action-at-a-distance. It would be hard to argue against the reasonableness of such an assumption, even if one were so bold as to doubt its inevitability. Bell showed that any hidden-variables formulation of quantum mechanics must be nonlocal, as, indeed, Bohmian mechanics is. But he showed much much more.

In a celebrated paper published in 1964, Bell showed that quantum theory itself is irreducibly nonlocal. This fact about quantum mechanics, based as it is on a short and mathematically simple analysis, could have been recognized soon after the discovery of quantum theory in the 1920's. That this did not happen is no doubt due in part to the obscurity of orthodox quantum theory and to the ambiguity of its commitments. It was, in fact, his examination of Bohmian mechanics that led Bell to his nonlocality analysis. In the course of his investigation of Bohmian mechanics he observed that (Bell 1987, p. 11):

> in this theory an explicit causal mechanism exists whereby the disposition of one piece of apparatus affects the results obtained with a distant piece.
>
> Bohm of course was well aware of these features of his scheme, and has given them much attention. However, it must be stressed that, to the present writer's knowledge, there is no *proof* that *any* hidden variable account of quantum mechanics *must* have this extraordinary character. It would therefore be interesting, perhaps, to pursue some further "impossibility proofs," replacing the arbitrary axioms objected to above by some condition of locality, or of separability of distant systems.

In a footnote, Bell added that "Since the completion of this paper such a proof has been found." This proof was published in his 1964 paper, "On the Einstein-Podolsky-Rosen Paradox," in which he derives Bell's inequality, the basis of his conclusion of quantum nonlocality. (For a discussion of how nonlocality emerges in Bohmian mechanics, see Section 13.)

It is worth stressing that Bell's analysis indeed shows that any account of quantum phenomena must be nonlocal, not just any hidden variables account. Bell showed that nonlocality is implied by the predictions of standard quantum theory itself. Thus if nature is governed by these predictions, then nature is nonlocal. [That nature is so governed, even in the crucial EPR-correlation experiments, has by now been established by a great many experiments, the most conclusive of which is perhaps that of Aspect (Aspect et al., 1982).]

Bell, too, stressed this point (by determinism Bell here means hidden variables):

> It is important to note that to the limited degree to which *determinism* plays a role in the EPR argument, it is not assumed but *inferred*. What is held sacred is the principle of 'local causality' -- or 'no action at a distance'...
>
> It is remarkably difficult to get this point across, that determinism is not a *presupposition* of the analysis. (Bell 1987, p. 143)
>
> Despite my insistence that the determinism was inferred rather than assumed, you might still suspect somehow that it is a preoccupation with determinism that creates the problem. Note well then that the following argument makes no mention whatever of determinism. ...

> Finally you might suspect that the very notion of particle, and particle orbit ... has somehow led us astray. ... So the following argument will not mention particles, nor indeed fields, nor any other particular picture of what goes on at the microscopic level. Nor will it involve any use of the words 'quantum mechanical system', which can have an unfortunate effect on the discussion. The difficulty is not created by any such picture or any such terminology. It is created by the predictions about the correlations in the visible outputs of certain conceivable experimental set-ups. (Bell 1987, p. 150)

The "problem" and "difficulty" to which Bell refers above is the conflict between the predictions of quantum theory and what can be inferred, call it C, from an assumption of locality in Bohm's version of the EPR argument, a conflict established by Bell's inequality. C happens to concern the existence of a certain kind of hidden variables, what might be called local hidden variables, but this fact is of little substantive importance. What is important is not so much the identity of C as the fact that C is incompatible with the predictions of quantum theory. The identity of C is, however, of great historical significance: It is responsible for the misconception that Bell proved that hidden variables are impossible, a belief until recently almost universally shared by physicists, as well as for the view, even now almost universally held, that what Bell's result does is to rule out local hidden variables, a view that is misleading.

Here again is Bell, expressing the logic of his *two-part* demonstration of quantum nonlocality, the first part of which is Bohm's version of the EPR argument:

> Let me summarize once again the logic that leads to the impasse. The EPRB correlations are such that the result of the experiment on one side immediately foretells that on the other, whenever the analyzers happen to be parallel. If we do not accept the intervention on one side as a causal influence on the other, we seem obliged to admit that the results on both sides are determined in advance anyway, independently of the intervention on the other side, by signals from the source and by the local magnet setting. But this has implications for non-parallel settings which conflict with those of quantum mechanics. So we *cannot* dismiss intervention on one side as a causal influence on the other. (Bell 1987, p. 149)

[Return to Table of Contents]

3. History

The pilot-wave approach to quantum theory was initiated, even before the discovery of quantum mechanics itself, by Einstein, who hoped that interference phenomena involving particle-like photons could be explained if the motion of the photons were somehow guided by the electromagnetic field -- which would thus play the role of what he called a *Führungsfeld* or guiding field (Wigner 1976, p. 262). While the notion of the electromagnetic field as guiding field turned out to be rather problematical, the possibility that for a system of electrons the wave function might play this role, of guiding field or pilot

wave, was explored by Max Born in his early paper founding quantum scattering theory (Born 1926) -- a suggestion to which Heisenberg was profoundly unsympathetic.

Not long after Schrödinger's discovery, in 1926, of wave mechanics, i.e., of Schrödinger's equation, Louis de Broglie in effect discovered Bohmian mechanics: In 1927, de Broglie found an equation of particle motion equivalent to the guiding equation for a scalar wave function (de Broglie 1928, p. 119), and he explained at the 1927 Solvay Congress how this motion could account for quantum interference phenomena. However, de Broglie responded poorly to an objection of Wolfgang Pauli (Pauli 1928) concerning inelastic scattering, no doubt making a rather bad impression on the illustrious audience gathered for the occasion.

Born and de Broglie very quickly abandoned the pilot-wave approach and became enthusiastic supporters of the rapidly developing consensus in favor of the Copenhagen interpretation. Bohmian mechanics was rediscovered in 1952 by David Bohm (Bohm 1952), the first person genuinely to understand its significance and implications. Its principal proponent during the sixties, seventies and eighties was John Bell.

[Return to Table of Contents]

4. The Defining Equations of Bohmian Mechanics

In Bohmian mechanics the wave function, obeying Schrödinger's equation, does not provide a complete description or representation of a quantum system. Rather, it governs the motion of the fundamental variables, the positions of the particles: In the Bohmian mechanical version of nonrelativistic quantum theory, quantum mechanics is fundamentally about the behavior of particles; the particles are described by their positions, and Bohmian mechanics prescribes how these change with time. In this sense, for Bohmian mechanics the particles, described by their positions, are primary, or primitive, while the wave function is secondary, or derivative.

Bohmian mechanics is the minimal completion of Schrödinger's equation, for a nonrelativistic system of particles, to a theory describing a genuine motion of particles. For Bohmian mechanics the state of a system of N particles is described by its wave function $\psi = \psi(\boldsymbol{q}_1,\dots,\boldsymbol{q}_N) = \psi(q)$, a complex (or spinor) valued function on the space of possible configurations q of the system, together with its actual configuration Q defined by the actual positions $\mathbf{Q}_1,\dots,\mathbf{Q}_N$ of its particles. The theory is then defined by two evolution equations: Schrödinger's equation

$$i\hbar(\partial\psi/\partial t) = H\psi$$

for $\psi(t)$, where H is the nonrelativistic (Schrödinger) Hamiltonian, containing the masses of the particles and a potential energy term, and a first-order evolution equation,

The Guiding Equation:
$d\mathbf{Q}_k/dt = (\hbar/m_k) \operatorname{Im} [\psi^* \partial_k \psi / \psi^* \psi] (\mathbf{Q}_1,...,\mathbf{Q}_N)$

for $Q(t)$, the simplest first-order evolution equation for the positions of the particles that is compatible with the Galilean (and time-reversal) covariance of the Schrödinger evolution (Dürr et al. 1992, pp. 852-854). Here $\hbar$ is Planck's constant divided by 2π, m_k is the mass of the k-th particle, and ∂_k is the gradient with respect to the coordinates of the k-th particle. If ψ is spinor-valued, the products in numerator and denominator should be understood as scalar products. If external magnetic fields are present, the gradient should be understood as the covariant derivative, involving the vector potential. (Since the denominator on the right hand side of the guiding equation vanishes at the nodes of ψ, global existence and uniqueness for the Bohmian dynamics is a nontrivial matter. It is proven in Berndl, Dürr, et al. 1995.)

For an N-particle system these two equations (together with the detailed specification of the Hamiltonian, including all interactions contributing to the potential energy) completely define the Bohmian mechanics. This deterministic theory of particles in motion accounts for all the phenomena of nonrelativistic quantum mechanics, from interference effects to spectral lines (Bohm 1952, pp. 175-178) to spin (Bell 1964, p. 10), and it does so in an entirely ordinary manner, as we shall explain in the following sections.

The form of the guiding equation given above is, for a scalar wave function, describing particles without spin, a little more complicated than necessary, since the complex conjugate of the wave function, appearing in the numerator and the denominator, cancels. If one looks for an evolution equation for the configuration compatible with the space-time symmetries of Schrödinger's equation, one almost immediately arrives at the guiding equation in this simpler form as the simplest possibility.

However, the form given above has two advantages: First, it makes sense for particles with spin -- and all the apparently paradoxical quantum phenomena associated with spin are, in fact, thereby accounted for by Bohmian mechanics without further ado. Secondly, and this is crucial to the fact that Bohmian mechanics is empirically equivalent to orthodox quantum theory, the right hand side of the guiding equation is J/ρ, the ratio of the quantum probability current to the quantum probability density. This shows first of all that it should require no imagination whatsoever to guess the guiding equation from Schrödinger's equation, provided one is looking for one, since the classical formula for current is density times velocity. Moreover, it follows from the quantum continuity equation $\partial\rho/\partial t + \operatorname{div} J = 0$, an immediate consequence of Schrödinger's equation, that if at some time (say the initial time) the configuration Q of our system is random, with distribution given by $|\psi|^2 = \psi^*\psi$, this will be true at all times (so long as the system does not interact with its environment).

This demonstrates that all claims to the effect that the predictions of quantum theory are incompatible with the existence of hidden variables, with an underlying deterministic model in which quantum randomness arises from averaging over ignorance, are wrong. For Bohmian mechanics provides us with

just such a model: For any quantum experiment we merely take as the relevant Bohmian system the combined system that includes the system upon which the experiment is performed as well as all the measuring instruments and other devices used in performing the experiment (together with all other systems with which these have significant interaction over the course of the experiment). The "hidden variables" model is then obtained by regarding the initial configuration of this big system as random in the usual quantum mechanical way, with distribution given by $|\psi|^2$. The initial configuration is then transformed, via the guiding equation for the big system, into the final configuration at the conclusion of the experiment. It then follows that this final configuration of the big system, including in particular the orientation of instrument pointers, will also be distributed in the quantum mechanical way, so that this deterministic Bohmian model yields the usual quantum predictions for the results of the experiment.

As the preceding paragraph suggests, and as we discuss in more detail in later sections, in Bohmian mechanics there is no need -- and, indeed, no room -- for any "measurement postulates" or *axioms* governing the behavior of other "observables": Any such axioms would be at best redundant, and would quite possibly be inconsistent.

[Return to Table of Contents]

5. The Quantum Potential

Bohmian mechanics has been presented here as a first-order theory, in which it is the velocity, the rate of change of position, that is fundamental: it is this quantity, given by the guiding equation, that is specified by the theory, directly and simply, with the second-order (Newtonian) concepts of acceleration and force, work and energy playing no fundamental role. Bohm, however, did not regard his theory in this way. He regarded it, fundamentally, as a second-order theory, describing particles moving under the influence of forces, among which, however, one must include a force stemming from a "quantum potential."

In his 1952 hidden-variables paper (Bohm 1952), Bohm arrived at his theory by writing the wave function in the polar form $\psi = R\exp(iS/\hbar)$, where S and R are real, with R nonnegative, and rewriting Schrödinger’s equation in terms of these new variables to obtain a pair of coupled evolution equations: the continuity equation for $\rho = R^2$ and a modified Hamilton-Jacobi equation for S, differing from the usual classical Hamilton-Jacobi equation only by the appearance of an extra term, the *quantum potential*

$$U = -\Sigma_k(\hbar^2/2m_k)\,(\partial_k^2 R / R),$$

alongside the classical potential energy term.

Bohm then used the modified Hamilton-Jacobi equation to define particle trajectories just as is done for the classical Hamilton-Jacobi equation, that is, by identifying $\partial_k S$ with $m_k \mathbf{v}_k$, i.e., by setting

$$d\mathbf{Q}_k/dt = \partial_k S \,/\, m_k,$$

which is equivalent to the guiding equation for particles without spin. [Notice that in this form the guiding equation is already suggested by the (pre-Schrödinger equation) de Broglie relation $\boldsymbol{p} = \hbar \boldsymbol{k}$, as well as by the eikonal equation of classical optics.] The resulting motion is precisely what would be obtained classically if the particles were acted upon by, in addition to the usual forces, the force generated by the quantum potential.

The quantum potential formulation of the de Broglie-Bohm theory is still fairly widely used. For example, the theory is presented in this way in the two existing monographs, by Bohm and Hiley and by Holland. And regardless of whether or not we regard the quantum potential as fundamental, it can in fact be quite useful. In order most simply to see that Newtonian mechanics should be expected to emerge from Bohmian mechanics in the classical limit, it is convenient to transform the theory into Bohm's Hamilton-Jacobi form. One then sees that the (size of the) quantum potential provides a measure of the deviation of Bohmian mechanics from its classical approximation. Moreover, the quantum potential can also be used to develop approximation schemes for solutions to Schrödinger's equation (Nerukh and Frederick 2000).

However, Bohm's rewriting of Schrödinger's equation in terms of variables that seem interpretable in classical terms does not come without a cost. The most obvious is increased complexity: Schrödinger's equation is rather simple, not to mention linear, whereas the modified Hamilton-Jacobi equation is somewhat complicated, and highly nonlinear -- and still requires the continuity equation for its closure. The quantum potential itself is neither simple nor natural. Even to Bohm it has seemed "rather strange and arbitrary" (Bohm 1980, p. 80). And it is not very satisfying to think of the quantum revolution as amounting to the insight that nature is classical after all, except that there is in nature what appears to be a rather ad hoc additional force term, the one arising from the quantum potential. The artificiality suggested by the quantum potential is the price one pays if one insists on casting a highly nonclassical theory into a classical mold.

Moreover, the connection between classical mechanics and Bohmian mechanics that is suggested by the quantum potential is rather misleading. Bohmian mechanics is not simply classical mechanics with an additional force term. In Bohmian mechanics the velocities are not independent of positions, as they are classically, but are constrained by the guiding equation. In classical Hamilton-Jacobi theory we also have this equation for the velocity, but there the Hamilton-Jacobi function S can be entirely eliminated and the description in terms of S simplified and reduced to a finite-dimensional description, with basic variables the positions and the (unconstrained) momenta of all the particles, given by Hamilton's or Newton's equations.

It can be argued that the most serious flaw in the quantum potential formulation of Bohmian mechanics is that it gives a completely false impression of the lengths to which we must go in order to convert orthodox quantum theory into something more rational. The quantum potential suggests, and indeed it has often been stated, that in order to transform Schrödinger's equation into a theory that can, in what are

often called "realistic" terms, account for quantum phenomena, many of which are dramatically nonlocal, we must add to the theory a complicated quantum potential of a grossly nonlocal character. It should be clear that such sentiments are inappropriate, since the quantum potential need not be mentioned in the *formulation* of Bohmian mechanics and in any case is merely a reflection of the wave function, which Bohmian mechanics does not add to but shares with orthodox quantum theory.

[Return to Table of Contents]

6. The Two-Slit Experiment

According to Richard Feynman, the two-slit experiment for electrons is (Feynman et al. 1963, p. 37-2) "a phenomenon which is impossible, *absolutely* impossible, to explain in any classical way, and which has in it the heart of quantum mechanics. In reality it contains the *only* mystery." This experiment (Feynman 1967, p. 130) "has been designed to contain all of the mystery of quantum mechanics, to put you up against the paradoxes and mysteries and peculiarities of nature one hundred per cent." As to the question (Feynman 1967, p. 145), "How does it really work? What machinery is actually producing this thing? Nobody knows any machinery. Nobody can give you a deeper explanation of this phenomenon than I have given; that is, a description of it."

But Bohmian mechanics is just such a deeper explanation. It resolves the dilemma of the appearance, in one and the same phenomenon, of both particle and wave properties in a rather straightforward manner: Bohmian mechanics is a theory of motion describing a particle (or particles) guided by a wave. Here we have a family of Bohmian trajectories for the two-slit experiment.

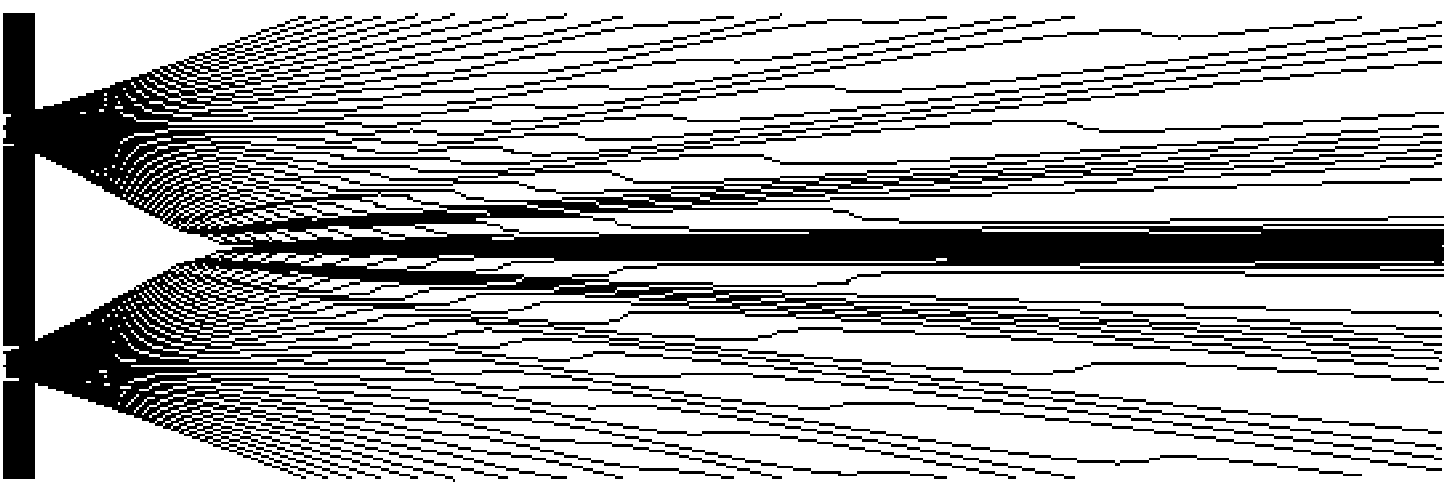


Figure 1: An ensemble of trajectories for the two-slit experiment, uniform in the slits. (Adapted by Gernot Bauer from Philippidis et al. 1979.)

While each trajectory passes through but one of the slits, the wave passes through both; the interference profile that therefore develops in the wave generates a similar pattern in the trajectories guided by this wave.

Compare Feynman's presentation with Bell's (Bell 1987, p. 191):

> Is it not clear from the smallness of the scintillation on the screen that we have to do with a particle? And is it not clear, from the diffraction and interference patterns, that the motion of the particle is directed by a wave? De Broglie showed in detail how the motion of a particle, passing through just one of two holes in screen, could be influenced by waves propagating through both holes. And so influenced that the particle does not go where the waves cancel out, but is attracted to where they cooperate. This idea seems to me so natural and simple, to resolve the wave-particle dilemma in such a clear and ordinary way, that it is a great mystery to me that it was so generally ignored.

The most puzzling aspect of the two-slit experiment is perhaps the following: If, by any means whatsoever, one is able to determine through which slit the particle passes, the interference pattern will be destroyed. This dramatic effect of observation is, in fact, a simple consequence of Bohmian mechanics. To see this one need only carefully consider what determining the slit through which the particle passes should mean. In particular, one must recognize that this must involve interaction with another system that must also be included in the Bohmian mechanical analysis. This destruction of interference is related, naturally enough, to the Bohmian mechanical analysis of quantum measurement (Bohm 1952), and it occurs via the mechanism that leads, in Bohmian mechanics, to the "collapse of the wave function."

[Return to Table of Contents]

7. The Measurement Problem

The most commonly cited of the conceptual difficulties that plague quantum mechanics is the measurement problem, or, what amounts to more or less the same thing, the paradox of Schrödinger's cat. Indeed, for many physicists the measurement problem is not merely one of the conceptual difficulties of quantum mechanics; it is *the* conceptual difficulty.

The problem is as follows. Suppose that the wave function of any individual system provides a complete description of that system. When we analyze the process of measurement in quantum mechanical terms, we find that the after-measurement wave function for system and apparatus arising from Schrödinger's equation for the composite system typically involves a superposition over terms corresponding to what we would like to regard as the various possible results of the measurement -- e.g., different pointer orientations. It is difficult to discern in this description of the after-measurement situation the actual result of the measurement -- e.g., some specific pointer orientation. But the whole point of quantum theory, and the reason we are to believe in it, is that it is supposed to provide a compelling, or at least an efficient, account of our observations, that is, of outcomes of measurements. In short, the measurement problem is this: Quantum theory implies that measurements typically fail to have outcomes of the sort the theory was created to explain.

By contrast, if, like Einstein, we regard the description provided by the wave function as incomplete, the measurement problem vanishes: With a theory or interpretation like Bohmian mechanics, in which the description of the after-measurement situation includes, in addition to the wave function, at least the values of the variables that register the result, there is no measurement problem. In Bohmian mechanics pointers always point.

The measurement problem is often expressed a little differently. It is noted that textbook quantum theory provides two rules for the evolution of the wave function of a quantum system: A deterministic dynamics given by Schrödinger’s equation for when the system is not being "measured" or observed, and a random collapse of the wave function to an *eigenstate* of the "measured observable" for when it is. However, the objection continues, textbook quantum theory does not provide a coherent account of how these two apparently incompatible rules can be reconciled.

That this formulation of the measurement problem is more or less equivalent to the previous one should be reasonably clear: If a wave function provides a complete description of the after-measurement situation, the outcome of the measurement must correspond to a wave function describing the actual result, that is, a "collapsed" wave function. Hence the collapse rule. But it is difficult to take seriously the idea that those interactions between system and apparatus that we happen to call measurements should be governed by laws different from those governing all other interactions. Hence the apparently incompatibility of the two rules.

The second formulation of the measurement problem, though basically equivalent to the first one, suggests an important question: Can Bohmian mechanics itself provide a coherent account of how the two dynamical rules might be reconciled? How does Bohmian mechanics justify the use of the "collapsed" wave function in place of the original one? This question was answered in Bohm’s first papers on Bohmian mechanics (Bohm 1952, Part I, Section 7, and Part II, Section 2). What would nowadays be called effects of decoherence, produced by interaction with the environment (air molecules, cosmic rays, internal microscopic degrees of freedom, etc.), make it extremely difficult for the component of the after-measurement wave function corresponding to the actual result of the measurement to develop significant overlap -- in the configuration space of the very large system that includes all systems with which the original system and apparatus come into interaction -- with the other components of the after-measurement wave function. But without such overlap the future evolution of the configuration of the system and apparatus is generated, to a high degree of accuracy, by that component all by itself. The replacement is thus justified as a practical matter. (See also Dürr et al. 1992, Section 5.)

It is widely believed by proponents of orthodox quantum theory that the measurement problem itself is somehow resolved by decoherence. It is not easy to understand this belief. In the first formulation of the measurement problem, nothing prevents us from including in the apparatus all sources of decoherence. But then there is no longer any room for decoherence to be in any way relevant to that argument. Be that as it may, one of the best descriptions of the mechanisms of decoherence, though not the word itself, was given by Bohm (Bohm 1952), who recognized its importance several decades before it became

fashionable. (See also the encyclopedia entry on The Role of Decoherence in Quantum Mechanics.)

[Return to Table of Contents]

8. The Collapse of the Wave Function

In the previous section it was indicated that collapse of the wave function can be regarded in Bohmian mechanics as a pragmatic affair. However, there is a sense in which the collapse of the wave function in Bohmian mechanics is more than a matter of convenience. If we focus on what should be regarded as the wave function, not of the composite of system and apparatus -- which strictly speaking remains a superposition if the composite is treated as closed during the measurement process -- but of the system itself, we find that for Bohmian mechanics this does indeed collapse, precisely as described by the quantum formalism. The key element here is the notion of the *conditional wave function* of a subsystem of a larger system, described briefly in this section and discussed in some detail, together with the related notion of the effective wave function, in Dürr et al. 1992, Section 5.

For the evolution of the wave function, Bohmian mechanics is formulated in terms of Schrödinger's equation alone. Nonetheless the textbook collapse rule is a consequence of the Bohmian dynamics. To appreciate this one should first note that, since observation implies interaction, a system under observation cannot be a closed system but rather must be a subsystem of a larger system that is closed, which we may take to be the entire universe, or any smaller more or less closed system that contains the system to be observed, *the subsystem*. The configuration Q of this larger system naturally splits into X, the configuration of the subsystem, and Y, the configuration of the *environment* of the subsystem.

Suppose the larger system has wave function $\Psi = \Psi(q) = \Psi(x, y)$. According to Bohmian mechanics, the larger system is then completely described by Ψ, evolving according to Schrödinger's equation, together with X and Y. The question then arises -- and it is a critical question -- as to what should be meant by the wave function of the subsystem.

There is a rather obvious answer for this, a natural function of x that suitably incorporates the objective structure at hand, namely the *conditional wave function*

$$\psi(x) = \Psi(x, Y)$$

obtained by plugging the actual configuration of the environment into the wave function of the larger system. (This definition is appropriate only for scalar wave functions; for particles with spin the situation would be a little more complicated.) It then follows immediately that the configuration of the subsystem obeys the guiding equation with the conditional wave function on its right hand side.

Moreover, taking into account the way that the conditional wave function depends upon time t

$$\psi_t(x) = \Psi_t(x, Y_t)$$

via the time dependence of Y as well as that of Ψ, it is not difficult to see (Dürr et al. 1992) that the conditional wave function obeys Schrödinger's equation for the subsystem when that system is suitably decoupled from its environment -- this is meant to imply, in particular, that Ψ has a special form, what might be called an effective product form (similar to but more general than the superposition produced in an "ideal quantum measurement"), in which case the conditional wave function of the subsystem is also called its *effective wave function* -- and, using the quantum equilibrium hypothesis, that it randomly collapses according to the usual quantum mechanical rules under precisely those conditions on the interaction between the subsystem and its environment that define an ideal quantum measurement.

It is perhaps worth noting that orthodox quantum theory lacks the resources, namely, the actual configuration of the environment, that make possible the definition of the conditional wave function. Indeed, from an orthodox point of view what should be meant by the wave function of a subsystem is entirely obscure.

[Return to Table of Contents]

9. Quantum Randomness

According to the quantum formalism, the probability density for finding a system whose wave function is ψ in the configuration q is $|\psi(q)|^2$. To the extent that the results of measurement are registered configurationally, at least potentially, it follows that the predictions of Bohmian mechanics for the results of measurement must agree with those of orthodox quantum theory (assuming the same Schrödinger equation for both) provided that it is somehow true for Bohmian mechanics that configurations are random, with distribution given by the *quantum equilibrium* distribution $|\psi(q)|^2$. Now the status and justification of this *quantum equilibrium hypothesis* is a rather delicate matter, one that has been explored in considerable detail (Dürr et al. 1992). Here are but a few relevant points.

It is nowadays a rather familiar fact that dynamical systems quite generally give rise to behavior of a statistical character, with the statistics given by the (or a) stationary probability distribution for the dynamics. So it is with Bohmian mechanics, except that for the Bohmian system stationarity is not quite the right concept, and it is rather the notion of *equivariance* that is relevant. A probability distribution ρ^ψ on configuration space, depending upon the wave function ψ, is *equivariant* if

$$(\rho^\psi)_t = \rho^{\psi_t}$$

where the dependence on t on the right arises from Schrödinger's equation and on the left from the evolution on probability distributions arising from the flow induced by the guiding equation. Thus

equivariance expresses the mutual compatibility, relative to ρ^{ψ}, of the Schrödinger evolution of the wave function and the Bohmian motion of the configuration. It is an immediate consequence of the guiding equation and the quantum continuity equation that $\rho^{\psi} = |\psi(q)|^2$ is equivariant.

It is perhaps helpful, in trying to understand the status in Bohmian mechanics of the quantum equilibrium distribution, to think of

quantum equilibrium, $\rho = |\psi|^2$

as roughly analogous to (classical)

thermodynamic equilibrium, $\rho = \exp(-H/kT)/Z$,

the probability distribution of the phase-space point of a system in equilibrium at temperature *T*. (*Z* is a normalization constant called the partition function and *k* is Boltzmann's constant.) This analogy has several facets: In both cases the probability distributions are naturally associated with their respective dynamical systems. In particular, these distributions are stationary or, what amounts to the same thing within the framework of Bohmian mechanics, equivariant. In both cases it appears natural to try to justify these equilibrium distributions by means of mixing-type, convergence-to-equilibrium arguments (Bohm 1953, Valentini 2001). In both cases the ultimate justification for these probability distributions must, arguably, be in terms of statistical patterns exhibited by ensembles of actual subsystems within a *typical* individual universe (Bell 1987, page 129, Dürr et al. 1992). (And in both cases the status of, and justification for, equilibrium distributions is still controversial.) It can be shown (Dürr et al. 1992) that probabilities for positions given by the quantum equilibrium distribution emerge naturally from an analysis of "equilibrium" for the deterministic dynamical system defined by Bohmian mechanics, in much the same way that the Maxwellian velocity distribution emerges from an analysis of classical thermodynamic equilibrium. (For more on the thermodynamic side of the analogy see Goldstein 2001.) Thus with Bohmian mechanics the statistical description in quantum theory indeed takes, as Einstein anticipated, "an approximately analogous position to the statistical mechanics within the framework of classical mechanics."

[Return to Table of Contents]

10. Quantum Observables

It would appear that because orthodox quantum theory supplies us with probabilities not merely for positions but for a huge class of quantum observables, it is a much richer theory than Bohmian mechanics, which seems exclusively concerned with positions. Appearances are, however, misleading. In this regard, as with so much else in the foundations of quantum mechanics, the crucial remark has been made by Bell (Bell 1987, p. 166):

> [I]n physics the only observations we must consider are position observations, if only the positions of instrument pointers. It is a great merit of the de Broglie-Bohm picture to force us to consider this fact. If you make axioms, rather than definitions and theorems, about the "measurement" of anything else, then you commit redundancy and risk inconsistency.

Consider first classical mechanics. The observables are functions on phase space, functions of the positions and momenta of the particles. The theory is defined by the axioms governing the behavior of the basic observables -- Newton's equations for the positions or Hamilton's for positions and momenta. What would be the point of making additional axioms, for other observables? After all, the behavior of any observable is entirely determined by the behavior of the basic observables. For example, for classical mechanics, the principle of the conservation of energy is a theorem, not an axiom.

The situation might seem to be different in quantum mechanics, since in quantum mechanics there are no basic observables having the property that all other observables are functions of these. This is connected with the fact that in quantum mechanics, with its positivistic orientation, no observables are taken seriously as describing objective properties, as actually having values regardless of whether they are or have been measured. Rather, all talk of observables in quantum mechanics is supposed to be understood as talk about the measurement of the observables.

But if this is so, the situation with regard to other observables in quantum mechanics is not really that different from in classical mechanics. Whatever is supposed to be meant in quantum mechanics by the measurement of (the values of) observables -- that, we are urged to believe, don't actually have values -- it must at least refer to some experiment involving interaction between the "measured" system and a "measuring" apparatus leading to a recognizable result, as given potentially by, say, a pointer orientation. But then if the axioms that we do have suffice for the behavior of pointer orientations (at least when they are observed), rules about the measurement of other observables must be theorems, following from those axioms, not additional axioms.

It should be clear from the discussion towards the end of Section 4 and at the beginning of Section 9 that, assuming the quantum equilibrium hypothesis, any analysis of the measurement of a quantum observable for orthodox quantum theory -- whatever that may be taken to mean and however the corresponding experiment may be performed -- provides ipso facto at least as adequate an account for Bohmian mechanics. The only part of orthodox quantum theory relevant to the analysis is the Schrödinger evolution, and this it *shares* with Bohmian mechanics. The main difference in the two accounts is that the orthodox one encounters the measurement problem before reaching a satisfactory conclusion while the Bohmian account does not. This difference stems of course from what Bohmian mechanics *adds* to orthodox quantum theory: actual configurations.

In the rest of this section, I wish to touch upon the significance of quantum observables for Bohmian mechanics: on how they naturally emerge and what talk of them means. (It follows from what has been said in the three preceding paragraphs that what we conclude here about quantum observables for Bohmian mechanics holds for orthodox quantum theory as well.)

It happens that Bohmian mechanics leads to a natural association between experiments and so-called *generalized observables*, given by positive-operator-valued measures (Davies 1976), or POVM's, $O(dz)$, on the value spaces for the results of the experiments (Berndl, Daumer, et al. 1995). This association is such that the probability distribution of the result Z of an experiment, when performed upon a system with wave function ψ, is given by $<\psi \mid O(dz)\psi>$ (where < | > is the usual inner product between quantum state vectors).

Moreover, this conclusion is basically an immediate consequence of the very meaning of an experiment from a Bohmian perspective: a coupling of system to apparatus leading to a result Z that is a function of the final configuration of the total system, e.g., the orientation of a pointer. Analyzed in Bohmian mechanical terms, the experiment defines a map from the initial wave function of the system to the distribution of the result. It follows directly from the structure of Bohmian mechanics, and from the fact that the quantum equilibrium distribution is quadratic in the wave function, that this map is bilinear (or, more precisely, sesquilinear). Such a map is equivalent to a POVM.

The simplest example of a POVM is a standard quantum observable, corresponding to a self-adjoint operator A on the Hilbert space of quantum states (i.e., wave functions). For Bohmian mechanics, more or less every "measurement-like" experiment is associated with this special kind of POVM, and the familiar quantum measurement axiom that the distribution of the result of the "measurement of the observable A" is given by the *spectral measure* for A relative to the wave function (in the very simplest cases just the absolute squares of the so-called *probability amplitudes*) is thus obtained.

For a variety of reasons, it quickly became almost universal, after quantum mechanics was discovered, to speak of an experiment associated with an operator A in the manner just sketched as a *measurement* of the *observable A*, as if the operator somehow corresponded to a property of the system that is in some sense measured by that experiment. There is no greater source of confusion about the meaning and implications of quantum theory than this *naive realism about operators* (Daumer et al., 1997).

[Return to Table of Contents]

11. Spin

Both the way non-configurational quantum observables are treated in Bohmian mechanics, and some of the difficulties caused by the naive realism about operators mentioned above, can be illustrated nicely with the case of spin.

Spin is the canonical quantum observable having no classical counterpart, reputed to be impossible to grasp in a nonquantum way. The source of the difficulty is not so much that spin is quantized in the sense that its allowable values form a discrete set (for a spin-1/2 particle, $\pm\hbar/2$) -- energy too may be quantized in this sense -- nor even precisely that the components of spin in the different directions fail to commute

and so cannot be simultaneously discussed, measured, imagined, or whatever it is that we are admonished not to do with noncommuting observables. Rather the difficulty is that there is no ordinary (nonquantum) quantity which, like the spin observable, is a 3-vector and which also is such that its components in all possible directions belong to the same discrete set. The problem, in other words, is that the usual vector relationships among the various components of the spin vector are incompatible with the quantization conditions on the values of these components.

For a particle of spin-1 the problem is even more severe. Since the components of spin in different directions aren't simultaneously measurable, the impossible vector relationships for the spin components of a quantum particle are not observable relationships. Simon Kochen and Ernst Specker (Kochen and Specker 1967) showed that for a spin-1 particle the squares of the spin components in the various directions satisfy, according to quantum theory, a collection of relationships, each individually observable, that taken together are impossible: the relationships are incompatible with the idea that measurements of these observables merely reveal their preexisting values rather than, as we are urged to believe in quantum theory, creating them. This Kochen-Specker Theorem continues to be regarded by many physicists and philosophers of physics as a definitive argument against the possibility of hidden variables.

We thus might naturally wonder how Bohmian mechanics manages to cope with spin. But this question has already been answered here. Bohmian mechanics makes sense for particles with spin, i.e., for particles whose wave functions are spinor-valued. When such particles are suitably directed toward Stern-Gerlach magnets, they emerge moving in more or less a discrete set of directions -- 2 possible directions for a spin-1/2 particle, having 2 spin components, 3 for spin-1 with 3 spin components, and so on. This occurs because the Stern-Gerlach magnets are so designed and so oriented that a wave packet (a localized wave function with reasonably well defined velocity) directed towards the magnet will, by virtue of the Schrödinger evolution, separate into distinct packets -- corresponding to the spin components of the wave function and moving in the discrete set of directions. The particle itself, depending upon its initial position, ends up in one of the packets moving in one of the directions.

The probability distribution for the result of such a *Stern-Gerlach experiment* is conveniently expressed in terms of the quantum mechanical spin operators -- for a spin-1/2 particle given by the Pauli spin matrices -- in the manner alluded to above. From a Bohmian perspective there is no hint of paradox in any of this unless we are seduced by naive realism about operators into insisting, despite its evident impossibility, that the spin operators correspond to genuine properties of the particles.

[Return to Table of Contents]

12. Contextuality

The Kochen-Specker Theorem, the earlier theorem of Gleason (Gleason 1957 and Bell 1966), as well as a variety of other no-hidden-variables results, including Bell's inequality (Bell 1964), show that any hidden-

variables formulation of quantum mechanics must be *contextual*. It must violate the noncontextuality assumption "that measurement of an observable must yield the same value independently of what other measurements may be made simultaneously" (Bell 1987, p. 9). To many physicists and philosophers of science contextuality has seemed too great a price to pay for the rather modest benefits -- largely psychological, so they would say -- provided by hidden variables.

Even many Bohmians suggest that contextuality marks a significant departure from classical principles. For example, Bohm and Hiley (1993) write that "The context dependence of results of measurements is a further indication of how our interpretation does not imply a simple return to the basic principles of classical physics."

However, to understand contextuality from the perspective of Bohmian mechanics is to appreciate that almost nothing needs to be explained. Consider an operator *A* that commutes with operators *B* and *C* (which however don't commute with each other). What is often called the "result for *A*" in an experiment for "measuring *A* together with *B*" usually disagrees with the "result for *A*" in an experiment for "measuring *A* together with *C*" because, even if everything else is the same, these experiments are different and different experiments usually have different results. The misleading reference to measurement, with the associated naive realism about operators, makes contextuality seem more than it is.

If we avoid naive realism about operators, contextuality amounts to little more than the rather unremarkable observation that results of experiments should depend upon how they are performed, even when the experiments considered are associated with the same operator in the manner alluded to above. David Albert (Albert 1992, p. 153) has given a particularly simple and striking example of this dependence for Stern-Gerlach experiments "measuring" the *z*-component of spin. If one reverses the polarity in a magnet for "measuring" the *z*-component of spin, keeping the same geometry, one obtains another magnet for "measuring" the *z*-component of spin. The use of one or the other of these two magnets will often lead to opposite conclusions about the "value of the *z*-component of spin" prior to the "measurement" (for the same initial value of the position of the particle).

As Bell has insisted (Bell 1987, p. 166):

> A final moral concerns terminology. Why did such serious people take so seriously axioms which now seem so arbitrary? I suspect that they were misled by the pernicious misuse of the word 'measurement' in contemporary theory. This word very strongly suggests the ascertaining of some preexisting property of some thing, any instrument involved playing a purely passive role. Quantum experiments are just not like that, as we learned especially from Bohr. The results have to be regarded as the joint product of 'system' and 'apparatus,' the complete experimental set-up. But the misuse of the word 'measurement' makes it easy to forget this and then to expect that the 'results of measurements' should obey some simple logic in which the apparatus is not mentioned. The resulting difficulties soon show that any such logic is not ordinary logic. It is my impression that the whole vast subject of

> ‘Quantum Logic’ has arisen in this way from the misuse of a word. I am convinced that the word ‘measurement’ has now been so abused that the field would be significantly advanced by banning its use altogether, in favour for example of the word ‘experiment.’

[Return to Table of Contents]

13. Nonlocality

Bohmian mechanics is manifestly nonlocal: The velocity, as expressed in the guiding equation, of any one of the particles of a many-particle system will typically depend upon the positions of the other, possibly distant, particles whenever the wave function of the system is entangled, i.e., not a product of single-particle wave functions. This is true, for example, for the EPR-Bohm wave function, describing a pair of spin-1/2 particles in the singlet state, analyzed by Bell and many others. Thus does Bohmian mechanics make explicit the most dramatic feature of quantum theory: quantum nonlocality.

It should be emphasized that the nonlocality of Bohmian mechanics derives solely from the nonlocality built into the structure of standard quantum theory, as provided by a wave function on configuration space, an abstraction which, roughly speaking, combines -- or binds -- distant particles into a single irreducible reality. As Bell (Bell 1987, p. 115) has stressed,

> That the guiding wave, in the general case, propagates not in ordinary three-space but in a multidimensional-configuration space is the origin of the notorious ‘nonlocality’ of quantum mechanics. It is a merit of the de Broglie-Bohm version to bring this out so explicitly that it cannot be ignored.

Thus the nonlocal velocity relation in the guiding equation is but one aspect of the nonlocality of Bohmian mechanics. There is also the nonlocality, or nonseparability, implicit in the wave function itself and in its propagation, a nonlocality that does not in fact assume the structure -- actual configurations -- that Bohmian mechanics adds to orthodox quantum theory. And as Bell has shown, using the connection between the wave function and the predictions of quantum theory concerning experimental results, this nonlocality cannot easily be argued away (see Section 2).

The nonlocality of Bohmian mechanics can be appreciated perhaps most efficiently, in all its aspects, by focusing on the conditional wave function. Suppose, for example, that in an EPR-Bohm experiment particle 1 passes through its Stern-Gerlach magnet before particle 2 arrives at its magnet. Then the orientation of the Stern-Gerlach magnet for particle 1 will have a significant effect upon the conditional wave function of particle 2: If the Stern-Gerlach magnet for particle 1 is so oriented as to "measure the *z*-component of spin," then after particle 1 has passed through its magnet the conditional wave function of particle 2 will be an eigenvector (or eigenstate) of the *z*-component of spin (in fact, belonging to the eigenvalue that is the negative of the one "measured" for particle 1), and the same thing is true for any other component of spin. You can dictate the *kind* of spin eigenstate produced for particle 2 by

appropriately choosing the orientation of an arbitrarily distant magnet. As to the future behavior of particle 2, in particular how it is affected by its magnet, this of course depends very much on the character of its conditional wave function and hence is very strongly influenced by the choice of orientation of the distant magnet.

This nonlocal effect upon the conditional wave function of particle 2 follows from combining the standard analysis of the evolution of the wave function in the EPR-Bohm experiment with the definition of the conditional wave function. (For simplicity, we ignore permutation symmetry.) Before any magnets have been reached the EPR-Bohm wave function is a sum of two terms, corresponding to nonvanishing values for two of the four possible joint spin components for the two particles, each term a product of an eigenstate for a component of spin in a given direction for particle 1 with the opposite eigenstate (i.e., belonging to the eigenvalue that is the negative of the eigenvalue for particle 1) for the component of spin in the same direction for particle 2. Moreover, by virtue of its symmetry under rotations, it happens that the EPR-Bohm wave function has the property that any component of spin, i.e., any direction, can be used in this decomposition. (This property is very interesting.)

Decomposing the EPR-Bohm wave function using the component of spin in the direction associated with the magnet for particle 1, the evolution of the wave function as particle 1 passes its magnet is easy to grasp: The evolution of the sum is determined (using linearity) by that of its individual terms, and the evolution of each term by that of each of its factors. The evolution of the particle-1 factor leads to a displacement along the magnetic axis in the direction determined by the (sign of the) spin component (i.e., the eigenvalue), as described in the fourth paragraph of Section 11. Once this displacement has occurred (and is large enough) the conditional wave function for particle 2 will correspond to the term in the sum selected by the actual position of particle 1. In particular, it will be an eigenstate of the component of spin "measured by" the magnet for particle 1.

The nonlocality of Bohmian mechanics has a remarkable feature: it is screened by quantum equilibrium. It is a consequence of the quantum equilibrium hypothesis that the nonlocal effects in Bohmian mechanics don't yield observable consequences that are also controllable -- we can't use them to send instantaneous messages. This follows from the fact that, given the quantum equilibrium hypothesis, the observable consequences of Bohmian mechanics are the same as those of orthodox quantum theory, for which instantaneous communication based on quantum nonlocality is impossible (see Eberhard 1978). The importance of quantum equilibrium for obscuring the nonlocality of Bohmian mechanics has been stressed by Valentini (1991).

[Return to Table of Contents]

14. Lorentz Invariance

Like nonrelativistic quantum theory, of which it is a version, Bohmian mechanics is incompatible with special relativity, a central principle of physics: it is not Lorentz invariant. Nor can Bohmian mechanics

easily be modified to become Lorentz invariant. Configurations, defined by the *simultaneous* positions of all particles, play too crucial a role in its formulation, the guiding equation defining an evolution on *configuration* space.

This difficulty with Lorentz invariance is intimately connected with the nonlocality in Bohmian mechanics. Since quantum theory itself, by virtue merely of the character of its predictions concerning EPR-Bohm correlations, is irreducibly nonlocal (see Section 2), one might expect considerable difficulty with the Lorentz invariance of orthodox quantum theory as well with Bohmian mechanics. For example, the collapse rule of textbook quantum theory blatantly violates Lorentz invariance. As a matter of fact, the intrinsic nonlocality of quantum theory presents formidable difficulties for the development of any (many-particle) Lorentz invariant formulation that avoids the vagueness of orthodox quantum theory (see Maudlin 1994).

A somewhat surprising, and I think correct, evaluation of the importance of the problem of Lorentz invariance was made by Bell in an interview with the philosopher Renée Weber, not long before he died. Referring to the paradoxes of quantum mechanics, Bell observed that "Those paradoxes are simply disposed of by the 1952 theory of Bohm, leaving as *the* question, the question of Lorentz invariance. So one of my missions in life is to get people to see that if they want to talk about the problems of quantum mechanics -- the real problems of quantum mechanics -- they must be talking about Lorentz invariance."

The most common view on the matter of Lorentz invariance and quantum nonlocality is that a detailed description of microscopic quantum processes, such as would be provided by an extension of Bohmian mechanics to the relativistic domain, must violate Lorentz invariance. In this view Lorentz invariance is an emergent symmetry obeyed by our observations -- a statistical consequence of quantum equilibrium that governs the results of quantum experiments. This is the opinion of Bohm and Hiley (1993), of Holland (1993), and of Valentini (2001).

However -- unlike nonlocality -- violation of Lorentz invariance is not inevitable. It should be possible, it seems, to construct a fully Lorentz invariant theory providing a detailed description of microscopic quantum processes. One way to do this is by means of additional Lorentz invariant dynamical structure, for example a suitable time-like 4-vector field, that permits the definition of a foliation of space-time into space-like hypersurfaces providing a Lorentz invariant notion of "evolving configuration" and along which nonlocal effects are transmitted. See Dürr et al. 1999 for a toy model. Another possibility that should not be dismissed is that a fully Lorentz invariant account of quantum nonlocality can be achieved without the invocation of additional structure, exploiting only what is already at hand, for example, light-cone structure.

Be that as it may, Lorentz invariant nonlocality would remain somewhat enigmatic. The issues are extremely subtle. For example, Bell (1987, page 155) rightly would find "disturbing ... the impossibility of 'messages' faster than light, which follows from ordinary relativistic quantum mechanics in so far as it is unambiguous and adequate for procedures *we* [emphasis added] can actually perform. The exact elucidation of concepts like 'message' and 'we', would be a formidable challenge." While quantum

equilibrium and the absolute uncertainty that it entails (Dürr et al. 1992) may be of some help here, the situation remains puzzling.

[Return to Table of Contents]

15. Objections

Anyone who has engaged in arguments with colleagues about the foundations of quantum mechanics, whatever his position, will likely agree with the following observation of Tolstoy:

> I know that most men, including those at ease with problems of the highest complexity, can seldom accept even the simplest and most obvious truth if it be such as would oblige them to admit the falsity of conclusions which they have delighted in explaining to colleagues, which they have proudly taught to others, and which they have woven, thread by thread, into the fabric of their lives.

A great many objections have been and continue to be raised against Bohmian mechanics. Here are some of them: Bohmian mechanics makes predictions about results of experiments different from those of orthodox quantum theory so it is wrong. Bohmian mechanics makes the same predictions about results of experiments as orthodox quantum theory so it is untestable and therefore meaningless. Bohmian mechanics is mathematically equivalent to orthodox quantum theory so it is not really an alternative at all. Bohmian mechanics is more complicated than orthodox quantum theory, since it involves an extra equation. (This objection is based on the surprisingly common misconception that orthodox quantum theory is defined solely by Schrödinger's equation, and does not actually need as part of its formulation any of the measurement postulates found in textbook quantum theory. It is only within a many-worlds framework that this view could begin to make sense, but I strongly doubt that it makes sense even there.) Bohmian mechanics requires the postulation of a mysterious and undetectable quantum potential. Bohmian mechanics requires the addition to quantum theory of a mysterious pilot wave. Bohmian mechanics, as von Neumann has shown, can't possibly work. Bohmian mechanics, as Kochen and Specker have shown, can't possibly work. Bohmian mechanics, as Bell has shown, can't possibly work. Bohmian mechanics is a childish regression to discredited classical modes of thought. Bohmian trajectories are crazy, since they may be curved even when no classical forces are present. Bohmian trajectories are crazy, since a Bohmian particle may be at rest in stationary quantum states. Bohmian trajectories are crazy, since a Bohmian particle may be at rest in stationary quantum states, even when these are large-energy eigenstates. Bohmian trajectories are surrealistic. Bohmian mechanics, since it is deterministic, is incompatible with quantum randomness. Bohmian mechanics is nonlocal. Bohmian mechanics is unintuitive. Bohmian mechanics is the many-worlds interpretation in disguise. (For a bit of discussion of some of these objections, see the exchange of letters on Quantum Theory Without Observers, in the February 1999 issue of *Physics Today*, particularly the last four of the eight letters. A link is provided in the Other Internet Resources section below.)

Most of these objections have little or no merit. Some arise from naive realism about operators, some from the idea that, to the extent that the concepts of classical physics apply at all, the laws of classical physics are more or less a priori, some from an inability to grasp the point of Bohmian mechanics, and some from sheer ignorance.

It is perhaps worth mentioning that despite the empirical equivalence between Bohmian mechanics and orthodox quantum theory, there are a variety of experiments and experimental issues that don't fit comfortably within the standard quantum formalism but are easily handled by Bohmian mechanics. Among these are dwell and tunneling times (Leavens 1996), escape times and escape positions (Daumer et al. 1997a), scattering theory (Dürr et al., 2000), and quantum chaos (Cushing 1994, Dürr et al., 1992a).

There is one striking feature of Bohmian mechanics that is often presented as an objection but is better regarded as an important clue about the meaning of quantum mechanics: in Bohmian mechanics the wave function acts upon the positions of the particles but, evolving as it does autonomously via Schrödinger's equation, it is not acted upon by the particles. This point is discussed in Dürr et al. 1997 and in Goldstein and Teufel 2001, where it is suggested that, from a deeper perspective than afforded by standard Bohmian mechanics or quantum theory, the wave function should be regarded as nomological, as an object for conveniently expressing the law of motion somewhat analogous to the Hamiltonian in classical mechanics, and that a time-dependent Schrödinger-type equation, from this deeper (cosmological) perspective, is merely phenomenological.

Bohmian mechanics does not account for phenomena such as pair creation and annihilation characteristic of quantum field theory. This is not an objection to Bohmian mechanics but merely a recognition that quantum field theory explains a great deal more than does nonrelativistic quantum mechanics, whether in orthodox or Bohmian form. It does, however, underline the need to find an adequate, if not compelling, Bohmian version of quantum field theory, and of gauge theories in particular, a problem that is pretty much wide open. Some rather tentative steps in this direction can be found in Bohm and Hiley 1993, Holland 1993, Bell 1987 (p. 173), and in some of the articles in Cushing et al. 1996. (For a general discussion of this issue and of the point and value of Bohmian mechanics, see the exchange of letters between Goldstein and Weinberg by following the link provided in the Other Internet Resources section below.)

[Return to Table of Contents]

Bibliography

- Albert, D. Z., 1992, *Quantum Mechanics and Experience*, Cambridge, MA: Harvard University Press
- Aspect, A., Dalibard, J., and Roger, G., 1982, "Experimental test of Bell's inequalities using time-varying analyzers," *Phys. Rev. Lett* **49**: 1804-1807
- Bell, J. S., 1964, "On the Einstein-Podolsky-Rosen Paradox," *Physics* : 195-200; reprinted in Bell 1987 and in Wheeler and Zurek 1983

- Bell, J. S., 1966, "On the Problem of Hidden Variables in Quantum Theory," *Rev. Mod. Phys.* **38**: 447-452; reprinted in Bell 1987 and in Wheeler and Zurek 1983
- Bell, J. S., 1987, *Speakable and Unspeakable in Quantum Mechanics*, Cambridge: Cambridge University Press
- Beller, M., 1999, *Quantum Dialogue: The Making of a Revolution*, Chicago: University of Chicago Press
- Berndl, K., Daumer, M., Dürr, D., Goldstein, S., and Zanghì, N., 1995, "A Survey of Bohmian Mechanics," *Il Nuovo Cimento* **110B**: 737-750.
  [Preprint (in Postscript) available online.]
- Berndl, K., Dürr, D., Goldstein, S., Peruzzi, G., and Zanghì, N., 1995, "On the Global Existence of Bohmian Mechanics," *Commun. Math. Phys.* **173**: 647-673.
  [Preprint (in Postscript) available online.]
- Bohm, D., 1952, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II," *Physical Review* **85**: 166-193
- Bohm, D., 1953, "Proof that Probability Density Approaches $|\psi|^2$ in Causal Interpretation of Quantum Theory," *Physical Review* **89**: 458-466
- Bohm, D., 1980, *Wholeness and the Implicate Order*, New York: Routledge
- Bohm, D., and Hiley, B. J., 1993, *The Undivided Universe: An Ontological Interpretation of Quantum Theory*, London: Routledge & Kegan Paul
- Born, M., 1926, *Z. Phys.* **38**: 803; English translation in Ludwig, G., ed., 1968, *Wave Mechanics*, Oxford: Pergamon Press: 206
- Born, M., 1949, *Natural Philosophy of Cause and Chance* Oxford: Oxford University Press
- de Broglie, L., 1928, in Solvay 1928
- Cushing, J. T., 1994, *Quantum Mechanics: Historical Contingency and the Copenhagen Hegemony*, Chicago: University of Chicago Press
- Cushing, J. T., Fine, A., and Goldstein, S., eds., 1996, *Bohmian Mechanics and Quantum Theory: An Appraisal; Boston Studies in the Philosophy of Science **184***, Boston: Kluwer Academic Publishers
- Daumer, M., Dürr, D., Goldstein, S., and Zanghì, N., 1997, "Naive Realism About Operators," *Erkenntnis* **45**: 379-397.
  [Preprint (in Postscript) available online.]
- Daumer, M., Dürr, D., Goldstein, S., and Zanghì, N., 1997a, "On the Quantum Probability Flux Through Surfaces," *Journal of Statistical Physics* **88**: 967-977.
  [Preprint (in Postscript) available online.]
- Davies, E. B., 1976, *Quantum Theory of Open Systems*, London: Academic Press
- Dürr, D., 2001, *Bohmsche Mechanik als Grundlage der Quantenmechanik*, Berlin: Springer-Verlag
- Dürr, D., Goldstein, S., and Zanghì, N., 1992, "Quantum Equilibrium and the Origin of Absolute Uncertainty," *Journal of Statistical Physics* **67**: 843-907.
  [Available (in Postscript) online .]
- Dürr, D., Goldstein, S., and Zanghì, N., 1992a, "Quantum Chaos, Classical Randomness, and Bohmian Mechanics," *Journal of Statistical Physics* **68**: 259-270.

[Available (in Postscript) online.]

- Dürr, D., Goldstein, S., and Zanghì, N., 1997, "Bohmian Mechanics and the Meaning of the Wave Function," in Cohen, R. S., Horne, M., and Stachel, J., eds., *Experimental Metaphysics -- Quantum Mechanical Studies for Abner Shimony, Volume One; Boston Studies in the Philosophy of Science* **193**, Boston: Kluwer Academic Publishers.
  Preprint (in Postscript) available online.]
- Dürr, D., Goldstein, S., Münch-Berndl, K., and Zanghì, N., 1999, "Hypersurface Bohm-Dirac Models," *Phys. Rev. A* **60**: 2729-2736.
  [Preprint (in Postscript) available online.]
- Dürr, D., Goldstein, S., Teufel, S., and Zanghì, N., 2000, "Scattering Theory from Microscopic First Principles," *Physica A* **279**: 416-431
- Eberhard, P. H., 1978, "Bell's Theorem and the Different Concepts of Locality," *Il Nuovo Cimento* **46B**: 392-419
- Einstein, A., 1949, "Reply to Criticisms," in Schilpp 1949
- Einstein, A., Podolsky, B., and Rosen, N., 1935, "Can Quantum-Mechanical Description of Physical Reality be Considered Complete?," *Phys. Rev.* **47**: 777-780
- Feynman, R. P., 1967, *The Character of Physical Law*, Cambridge, MA: MIT Press
- Feynman, R. P., Leighton, R. B., and Sands, M., 1963, *The Feynman Lectures on Physics, I,* New York: Addison-Wesley
- Gleason, A. M., 1957, "Measures on the Closed Subspaces of a Hilbert Space," *J. Math. and Mech.* **6**: 885-893
- Goldstein, R., 2000, *Properties of Light: A Novel of Love, Betrayal, and Quantum Physics*, Boston : Houghton Mifflin
- Goldstein, S., 2001, "Boltzmann's Approach to Statistical Mechanics," in Bricmont, J., Dürr, D., Galavotti, M. C., Ghirardi, G., Petruccione, F., Nino Zanghì, N., eds., *Chance in Physics: Foundations and Perspectives, Lecture Notes in Physics* **574**, Berlin: Springer-Verlag.
  [Preprint (in Postscript) available online.]
- Goldstein, S., and Teufel, S., 2001, "Quantum Spacetime without Observers: Ontological Clarity and the Conceptual Foundations of Quantum Gravity," in Callender, C. and Huggett, N., eds., *Physics meets Philosophy at the Planck Scale*, Cambridge: Cambridge University Press.
  [Preprint (in Postscript) available online.]
- Holland, P. R., 1993, *The Quantum Theory of Motion*, Cambridge: Cambridge University Press
- Kochen, S., and Specker, E. P., 1967, "The Problem of Hidden Variables in Quantum Mechanics," *J. Math. and Mech.* **17**: 59-87
- Leavens, C. R., 1996, "The 'Tunneling-Time Problem' for Electrons," in Cushing et al. 1996
- Maudlin, T., 1994, *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*, Cambridge, MA: Blackwell
- Mermin, N. D., 1993, "Hidden Variables and the Two Theorems of John Bell," *Rev. Mod. Phys.* **65**: 803-815
- Nerukh, D., and Frederick, J. H., 2000, "Multidimensional Quantum Dynamics with Trajectories: a Novel Numerical Implementation of Bohmian Mechanics," *Chem. Phys. Lett.* **332**: 145-153
- Pauli, W., 1928, in Solvay 1928: 280-282
- Philippidis, C., Dewdney, C., and Hiley, B. J., 1979, "Quantum Interference and the Quantum

Potential," *Il Nuovo Cimento* **52B**: 15-28

- Schilpp, P. A., ed., 1949, *Albert Einstein, Philosopher-Scientist*, Evanston, IL: Library of Living Philosophers
- Schrödinger, E., 1935, "Die gegenwärtige Situation in der Quantenmechanik," *Naturwissenschaften* **23**: 807-812, 823-828, 844-849; English translation by Trimmer, J. D., 1980, "The Present Situation in Quantum Mechanics: A Translation of Schrödinger's 'Cat Paradox' Paper, *Proceedings of the American Philosophical Society* **124**: 323-338, reprinted in Wheeler and Zurek 1983
- Solvay Congress (1927), 1928, *Electrons et Photons: Rapports et Discussions du Cinquième Conseil de Physique tenu à Bruxelles du 24 au 29 Octobre 1927 sous les Auspices de l'Institut International de Physique Solvay*, Paris: Gauthier-Villars
- Valentini, A., 1991, "Signal-Locality, Uncertainty and the Subquantum *H*-Theorem. II," *Physics Letters A* **158**: 1-8
- Valentini, A., 2001, *Pilot-Wave Theory: An Alternative Approach to Modern Physics*, Cambridge: Cambridge University Press
- von Neumann, J., 1932, *Mathematische Grundlagen der Quantenmechanik*, Berlin: Springer Verlag; English translation by Beyer, R. T., 1955, *Mathematical Foundations of Quantum Mechanics*, Princeton: Princeton University Press
- Wheeler, J. A., and Zurek, W. H., eds., 1983, *Quantum Theory and Measurement*, Princeton: Princeton University Press
- Wigner, E. P., 1976, "Interpretation of Quantum Mechanics," in Wheeler and Zurek 1983
- Wigner, E. P., 1983, "Review of Quantum Mechanical Measurement Problem," in Meystre, P., and Scully, M. O., eds., *Quantum Optics, Experimental Gravity and Measurement Theory*, New York: Plenum Press

[Return to Table of Contents]

Other Internet Resources

- Collaboration Bohmian Mechanics
  - Exchange of letters between S. Goldstein and S. Weinberg.
- The "Mt. Rushmore" of foundations of quantum theory
- Eric Dennis' discussion e-group on Bell's inequality and Bohmian mechanics
- Exchange of letters on Quantum Theory Without Observers, (reprinted from *Physics Today*, February 1999)

[Return to Table of Contents]

Related Entries

Einstein, Albert: Einstein-Bohr debates | physics: holism and nonseparability | quantum mechanics | quantum mechanics: Copenhagen interpretation of | quantum mechanics: Kochen-Specker theorem | quantum mechanics: many-worlds interpretation of | quantum mechanics: modal interpretations of | quantum mechanics: the role of decoherence in | quantum theory: measurement in | quantum theory: quantum entanglement and information | quantum theory: quantum gravity | quantum theory: quantum logic and probability theory | quantum theory: the Einstein-Podolsky-Rosen argument in | Uncertainty Principle

Acknowledgements

I am grateful to the editor, Guido Bacciagaluppi, for a very careful reading and many valuable suggestions.

Sheldon Goldstein
oldstein@math.rutgers.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 26, 2001
Content last modified: October 26, 2001

Quantum Entanglement and Information

Quantum entanglement is a physical resource, like energy, associated with the peculiar nonclassical correlations that are possible between separated quantum systems. Entanglement can be measured, transformed, and purified. A pair of quantum systems in an entangled state can be used as a quantum information channel to perform computational and cryptographic tasks that are impossible for classical systems. The general study of the information-processing capabilities of quantum systems is the subject of quantum information.

Quantum Entanglement

In 1935 and 1936, Schrödinger published a two-part article in the *Proceedings of the Cambridge Philosophical Society* in which he discussed and extended a remarkable argument by Einstein, Podolsky, and Rosen. The Einstein-Podolsky-Rosen (EPR) argument was, in many ways, the culmination of Einstein's critique of the orthodox Copenhagen interpretation of quantum mechanics, and was designed to show that the theory is incomplete. In classical mechanics the state of a system is essentially a list of the system's properties -- or, more precisely, it is the specification of a set of parameters from which the list of properties can be reconstructed: the positions and momenta of all the particles comprising the system. The dynamics of the theory specifies how properties change in terms of a law of evolution for the state. Pauli characterized this mode of description of physical systems as a ‘detached observer’ idealization (see Pauli's letter to Born in *The Born-Einstein Letters*, p. 218). On the Copenhagen interpretation, such a description is not possible for quantum systems. Instead, the quantum state of a

system should be understood as a catalogue of what an observer has done to the system and what has been observed, and the import of the state then lies in the probabilities that can be inferred (in terms of the theory) for the outcomes of possible future observations on the system. Einstein rejected this view and proposed a series of arguments to show that the quantum state is simply an incomplete characterization of the system. The missing parameters are sometimes referred to as ‘hidden parameters’ or ‘hidden variables’ (although Einstein did not use this terminology, presumably because he did not want to endorse any particular ‘hidden variable’ theory).

It should not be supposed that Einstein's definition of a complete theory included the requirement that it be deterministic. Rather, he required certain conditions of separability and locality for composite systems consisting of separated component systems: each component system separately should be characterized by its own properties (even if these properties manifest themselves stochastically), and it should be impossible to alter the properties of a distant system instantaneously (or the probabilities of these properties) by acting on a local system. In later analyses -- notably in Bell's extension of the EPR argument -- it became apparent that these conditions, suitably formulated as probability constraints, are equivalent to the requirement that statistical correlations between separated systems should be reducible to a common cause in the sense of Reichenbach.

In the original EPR article, two particles are prepared from a source in a certain quantum state and then move apart. There are ‘matching’ correlations between both the positions of the two particles and their momenta: a measurement of either position or momentum on a particular particle will allow the prediction, with certainty, of the outcome of a position measurement or momentum measurement, respectively, on the other particle. These measurements are mutually exclusive: either a position measurement can be performed, or a momentum measurement, but not both simultaneously. Either correlation can be observed, but the subsequent measurement of momentum, say, after establishing the position correlation, will no longer yield any correlation in the momenta of the two particles. It is as if the position measurement disturbs the correlation between the momentum values. The puzzle is that the quantum state of the particle pair is inconsistent with *any* assignment of precise position and momentum values to the particles separately. These values would be the common cause of the correlations, and would provide an explanation of the correlations in terms of the initial correlations between the properties of the two systems at the source. EPR concluded that the quantum state was incomplete.

Here is how Schrödinger put the puzzle in the first part of his two-part article (Schrödinger, p. 559):

> Yet since I can predict *either* x_1 or p_1 without interfering with the system No. 1 and since system No. 1, like a scholar in an examination, cannot possibly know which of the two questions I am going to ask first: it so seems that our scholar is prepared to give the right answer to the *first* question he is asked, *anyhow*. Therefore he must know both answers; which is an amazing knowledge; quite irrespective of the fact that after having given his first answer our scholar is invariably so disconcerted or tired out, that all the following answers are ‘wrong.’

What Schrödinger showed was that if two particles are prepared in a quantum state such that there is a matching correlation between two ‘canonically conjugate’ dynamical quantities -- quantities like position and momentum whose values suffice to specify all the properties of a classical system -- then there are infinitely many dynamical quantities of the two particles for which there exist similar matching correlations: every function of the canonically conjugate pair of the first particle matches with the same function of the canonically conjugate pair of the second particle. Thus (p. 559) system No. 1 ‘does not only know these two answers but a vast number of others, and that with no mnemotechnical help whatsoever, at least with none that we know of.’

Schrödinger coined the term ‘entanglement’ to describe this peculiar connection between quantum systems (Schrödinger, p. 555):

> When two systems, of which we know the states by their respective representatives, enter into temporary physical interaction due to known forces between them, and when after a time of mutual influence the systems separate again, then they can no longer be described in the same way as before, viz. by endowing each of them with a representative of its own. I would not call that *one* but rather *the* characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought. By the interaction the two representatives [the quantum states] have become entangled.

He added (Schrödinger, p. 555):

> Another way of expressing the peculiar situation is: the best possible knowledge of a *whole* does not necessarily include the best possible knowledge of all its *parts*, even though they may be entirely separate and therefore virtually capable of being ‘best possibly known,’ i.e., of possessing, each of them, a representative of its own. The lack of knowledge is by no means due to the interaction being insufficiently known -- at least not in the way that it could possibly be known more completely -- it is due to the interaction itself.
>
> Attention has recently been called to the obvious but very disconcerting fact that even though we restrict the disentangling measurements to *one* system, the representative obtained for the *other* system is by no means independent of the particular choice of observations which we select for that purpose and which by the way are *entirely* arbitrary. It is rather discomforting that the theory should allow a system to be steered or piloted into one or the other type of state at the experimenter's mercy in spite of his having no access to it.

In the second part of the paper, Schrödinger showed that, in general, a sophisticated experimenter can, by a suitable choice of operations carried out on one system, steer the second system into any ‘mixture’ of quantum states he chooses, i.e., not steer the system into any one particular state, but constrain the state into which the system evolves to lie in a given set, and at the same time fix the probabilities with which the system evolves into the states from the given set. He found this conclusion sufficiently unsettling to

suggest that the entanglement between two separating systems would persist only for distances small enough that the time taken by light to travel from one system to the other could be neglected, compared with the characteristic time periods associated with other changes in the composite system. He speculated that for longer distances each of the two systems might in fact be in a state associated with a certain mixture, determined by the precise form of the entangled state.

Most physicists dismissed the puzzling features of entangled quantum states as an artefact of Einstein's inappropriate ‘detached observer’ view of physical theory, and regarded Bohr's reply to the EPR argument as vindicating the Copenhagen interpretation. This was unfortunate, because the study of entanglement was ignored for thirty years until John Bell's reconsideration and extension of the EPR argument. Bell looked at entanglement in simpler systems than the EPR case: matching correlations between two-valued dynamical quantities, such as polarization or spin, of two separated systems in an entangled state. What Bell showed was that the statistical correlations between the measurement outcomes of suitably chosen *different* quantities on the two systems are inconsistent with an inequality derivable from Einstein's separability and locality assumptions -- in effect from the assumption that the correlations have a common cause.

Bell's investigation generated an ongoing debate on the foundations of quantum mechanics. One important feature of this debate was confirmation that entanglement can persist over long distances(see Aspect *et al*.), thus falsifying Schrödinger's supposition of the spontaneous decay of entanglement as two entangled particles separate. But it was not until the 1980s that physicists, computer scientists, and cryptographers began to regard the non-local correlations of entangled quantum states as a new kind of non-classical resource that could be exploited, rather than an embarrassment to be explained away. (For further discussion of entanglement as a physical resource, including measuring entanglement, and the manipulation and purification of entanglement by local operations, see "The Joy of Entanglement" by Popescu and Rohrlich in Lo, Popescu, and Spiller, or Nielsen and Chuang.)

Exploiting Entanglement: Quantum Teleportation

Consider again Schrödinger's realization that an entangled state could be used to steer a distant particle into one of a set of states, with a certain probability. In fact, this possibility of ‘remote steering’ is even more dramatic than Schrödinger demonstrated. Suppose Alice and Bob share an entangled state of the sort considered by Bell, say two photons in an entangled state of polarization. That is, Alice has in her possession one of the entangled photons, and Bob the other. Suppose that Alice has an additional photon in an unknown state of polarization, *u*. It is possible for Alice to perform an operation on the two photons in her possession that will transform Bob's photon into one of four states, depending on the four possible (random) outcomes of Alice's operation: either the state *u*, or a state that is related to *u* in a definite way. Alice's operation entangles the two photons in her possession, and disentangles Bob's photon, steering it into a state *u**. After Alice communicates the outcome of her operation to Bob, Bob knows either that *u** = *u*, or how to transform *u** to *u* by a local operation. This phenomenon is known as ‘quantum teleportation.’

What is extraordinary about this phenomenon is that Alice and Bob have managed to use their shared entangled state as a quantum communication channel to destroy the state *u* of a photon in Alice's part of the universe and recreate it in Bob's part of the universe. Since the state of a photon requires specifying a direction in space (essentially the value of an angle that can vary continuously), without a shared entangled state Alice would have to convey an infinite amount of classical information to Bob for Bob to be able to reconstruct the state *u* precisely. To see why this is so, consider that the decimal expansion of an angle variable represented by a real number is represented by a potentially infinite sequence of digits between 0 and 9. The binary expansion is represented by a potentially infinite sequence of 0's and 1's. Ever since Shannon formalized the notion of classical information, the amount of classical information associated with a binary alternative (represented as 0 or 1), where each alternative has equal *a priori* probability, is measured as one binary digit or ‘bit’. So to specify the value of an arbitrary angle variable requires an infinite number of bits. To specify the outcome of Alice's operation, which has four possible outcomes, with equal *a priori* probabilities, requires two bits of classical information. Remarkably, Bob can reconstruct the state *u* on the basis of just two bits of classical information communicated by Alice, apparently by exploiting the entangled state as a quantum communication channel to transfer the remaining information. (For further discussion of quantum teleportation, see Nielsen and Chuang, or Richard Josza's article "Quantum Information and its Properties" in Lo, Popescu, and Spiller.)

Quantum Information

Formally, the amount of information we gain, on average, when we learn the value of a random variable (or, equivalently, the amount of uncertainty in the value of a random variable before we learn its value) is represented by a quantity called the Shannon entropy, measured in bits. A random variable is defined by a probability distribution over a set of values. In the case of a binary random variable, with equal probability for each of the two possibilities, the Shannon entropy is 1 bit, representing maximal uncertainty. For all other probabilities -- intuitively, representing some information about which alternative is more likely -- the Shannon entropy is less than 1. For the case of maximal knowledge or zero uncertainty about the alternatives, where the probabilities are 0 and 1, the Shannon entropy is zero.

Since information is always embodied in the state of a physical system, we can also think of the Shannon entropy as quantifying the physical resources required to store classical information. Suppose Alice wishes to communicate some classical information to Bob over a classical communication channel such as a telephone line, say an email message. A relevant question concerns the extent to which the message can be compressed without loss of information, so that Bob can reconstruct the original message accurately from the compressed version. According to Shannon's noiseless coding theorem (assuming a noiseless telephone line with no loss of information), the minimal physical resources required to represent the message (effectively, a lower bound on the possibility of compression) are given by the Shannon entropy of the source.

What happens if we use the quantum states of physical systems to store information, rather than classical states? It turns out that quantum information is radically different from classical information. The unit of quantum information is the ‘qubit’, representing the amount of information that can be stored in the state

of the simplest quantum system, for example, the polarization state of a photon. (The term is due to Schumacher, who proved a quantum analogue of Shannon's noiseless coding theorem.) As we have seen, an arbitrarily large amount of classical information can be encoded in a qubit. This information can be processed and communicated but, because of the peculiarities of quantum measurement, at most one bit can be accessed! According to a theorem by Holevo, the accessible information in a probability distribution over a set of alternative qubits is limited by the von Neumann entropy, which is equal to the Shannon entropy only when the qubits are orthogonal in the space of quantum states, and is otherwise less than the Shannon entropy.

While classical information can be copied or cloned, the quantum ‘no cloning’ theorem (see Dieks, and Wootters and Zurek) asserts the impossibility of cloning an unknown quantum state. To see why, consider how we might construct a classical copying device. A NOT gate is a device that takes a bit as input and produces as output either a 1 if the input is 0, or a 0 if the input is 1. In other words, a NOT gate is a 1-bit gate that flips the input bit. A controlled-NOT gate, or CNOT gate, takes two bits as inputs, a control bit and a target bit, and flips the target bit if and only if the control bit is 1, while reproducing the control bit. (So there are two inputs, the control and target, and two outputs: the control, and either the target or the flipped target, depending on the value of the control.) A CNOT gate functions as a copying device for the control bit if the target bit is set to 0, because the output of the target bit is then a copy of the control bit (i.e., the input 00 produces output 00, and the input 10 produces output 11). Insofar as we can think of a measurement as simply a copying operation, a CNOT gate is the paradigm of a classical measuring device. (Imagine Alice equipped with such a device, with input and output control and target wires, measuring the properties of an unknown classical world. The input control wire is a probe for the presence of absence of a property, represented by a 1 or a 0. The target wire functions as the pointer, which is initially set to 0. The output of the target is a 1 or a 0, depending on the presence or absence of the property.)

Suppose we attempt to use our CNOT gate to copy an unknown qubit. Since the CNOT gate is now understood as a device for processing quantum states, the evolution from input states to output states must be effected by a physical quantum transformation. Now quantum transformations are linear on the linear state space of qubits. Linearity of the state space means that for any two qubit states (call them 0 and 1) that are orthogonal in the space of qubit states, there is a qubit state that is represented by a linear superposition or sum of these orthogonal states, say 0+1 (which will be non-orthogonal to either of these states). Linearity of the transformation means that any transformation must take a qubit state represented by the sum of two orthogonal qubits to a new qubit state that is the sum of the transformed orthogonal qubits. If the CNOT gate succeeds in copying two orthogonal qubits, it cannot succeed in copying a linear superposition of these qubits. Since the gate functions linearly, it must instead produce a state that is a linear superposition of the outputs obtained for the two orthogonal qubits. That is to say, the output of the gate will be represented by a quantum state that is a sum of two terms, where the first term represents the output of the control and target for the first orthogonal qubit, and the second term represents the output of the control and target for the second orthogonal qubit. This could be written as 00+11. This is an entangled state and not the output that would be required by a successful copying operation, where the control and target each outputs the superposed qubit (which could be written as (0+1)(0+1)).

Quantum Cryptography

Linearity prevents the possibility of cloning or measuring an unknown quantum state. Similarly, it can be shown that if Alice sends Bob one of two nonorthogonal qubits, Bob can obtain information about which of these qubits was sent only at the expense of disturbing the state. In general, for quantum information there is no information gain without disturbance. The impossibility of copying an unknown quantum state, or a state that is known to belong to a set of nonorthogonal states with a certain probability, and the existence of a trade-off relation between information gain and state disturbance, is the basis of the application of quantum information to cryptography. There are quantum protocols involving the exchange of classical and quantum information that Alice and Bob can exploit to share a secret random key, which they can then use to communicate privately. (See Lo's article "Quantum Cryptology" in Lo, Popescu, and Spiller.) Any attempt by an eavesdropper, Eve, to monitor the communication between Alice and Bob will be detectable, in principle, because Eve cannot gain any quantum information without some disturbance to the quantum communication channel. Moreover, the ‘no cloning’ theorem prohibits Eve from copying the quantum communications and processing them off-line, so to speak, after she monitors the classical communication between Alice and Bob.

While the difference between classical and quantum information can be exploited to achieve successful key distribution, there are other cryptographic protocols that are thwarted by quantum entanglement. Bit commitment is a key cryptographic protocol that can be used as a subroutine in a variety of important cryptographic tasks. In a bit commitment protocol, Alice supplies an encoded bit to Bob. The information available in the encoding should be insufficient for Bob to ascertain the value of the bit, but sufficient, together with further information supplied by Alice at a subsequent stage when she is supposed to reveal the value of the bit, for Bob to be convinced that the protocol does not allow Alice to cheat by encoding the bit in a way that leaves her free to reveal either 0 or 1 at will.

To illustrate the idea, suppose Alice claims the ability to predict advances or declines in the stock market on a daily basis. To substantiate her claim without revealing valuable information (perhaps to a potential employer, Bob) she suggests the following demonstration: She proposes to record her prediction, before the market opens, by writing a 0 (for ‘decline’) or a 1 (for ‘advance’) on a piece of paper, which she will lock in a safe. The safe will be handed to Bob, but Alice will keep the key. At the end of the day's trading, she will announce the bit she chose and prove that she in fact made the commitment at the earlier time by handing Bob the key. Of course, the key-and-safe protocol is not provably secure from cheating by Bob, because there is no principle of classical physics that that prevents Bob (if he is an ’ideal’ safe-cracker) from opening the safe and closing it again without leaving any traces. The question is whether there exists a quantum analogue of this procedure that is unconditionally secure: provably secure by the laws of physics against cheating by either Alice or Bob. Bob can cheat if he can obtain *some* information about Alice's commitment before she reveals it (which would give him an advantage in repetitions of the protocol with Alice). Alice can cheat if she can delay actually making a commitment until the final stage when she is required to reveal her commitment, or if she can change her commitment at the final stage with a very low probability of detection.

It turns out that unconditionally secure two-party bit commitment, based solely on the principles of quantum or classical mechanics (without exploiting special relativistic signalling constraints, or principles of general relativity or thermodynamics) is impossible. (See Mayers, Lo and Chau, and Lo's article "Quantum Cryptology" in Lo, Popescu, an Spiller for further discussion. Note that Kent has shown that one can implement a secure classical bit commitment protocol by exploiting relativistic signalling constraints in a timed sequence of communications between verifiably separated sites for both Alice and Bob.) Roughly, the impossibility arises because at any step in the protocol where either Alice or Bob is required to make a determinate choice (perform a measurement on a particle in the quantum channel, choose randomly and perhaps conditionally between a set of alternative actions to be implemented on the particle in the quantum channel, etc.), the choice can delayed by entangling one or more ‘ancilla’ (helper) particles with the channel particle in an appropriate way. By suitable operations on the ancillas, the channel particle can be ‘steered’ so that this cheating strategy is undetectable. In effect, if Bob can obtain no information about the bit in the safe, then entanglement will allow Alice to ‘steer’ the bit to either 0 or 1 at will.

Quantum Computation

Quantum information can be processed, but the accessibility of this information is limited by the Holevo bound. David Deutsch first showed how to exploit quantum entanglement to perform a computational task that is impossible for a classical computer. Suppose we have a black box that evaluates a function *f*. The arguments of *f* (inputs) are either 0 or 1. The values (outputs) of *f* (which are also 0 or 1) are either the same for both arguments (in which case *f* is constant), or different for the two arguments (in which case *f* is said to be ‘balanced’). We are interested in determining whether *f* is constant or balanced. Now, classically, the only way to do this is to run the black box twice, for both arguments 0 and 1, and to pass the values (outputs of *f*) to a circuit that determines whether they are the same (for ‘constant’) or different (for ‘balanced’). Deutsch showed that if we use quantum states and quantum gates to store and process information, then we can determine whether *f* is constant or balanced in one evaluation of the function *f*. The trick is to design the circuit (the sequence of gates) to produce the answer to a *global* question about the function (’constant’ or ‘balanced’) in an output qubit register that can then be read out or measured.

Consider again the quantum CNOT gate, with two orthogonal qubits 0 and 1 as possible inputs for the control, and 0 as the input for the target. Once can think of the input control and output target qubits, respectively, as the argument and associated value of a function. This CNOT function associates the value 0 with the argument 0 and the value 1 with the argument 1. For a linear superposition of the orthogonal qubits, say 0+1, as inut to the control, and the qubit representing 0 as the input to the target, the output is the entangled state 00+11, a linear superposition in which the first term represents the argument 0 and associated value (0) of the CNOT function, and the second term represents the argument 1 and associated value (1) of the CNOT function. The entangled state represents all possible arguments and corresponding values of the function as a linear superposition, but this information is not accessible. What can be shown to be accessible, by a suitable choice of quantum gates, is information about whether or not the function has certain global properties. This information is obtainable without reading out the

evaluation of any individual arguments and values. (Indeed, accessing information in the entangled state about a global property of the function will typically require losing access to all information about individual arguments and values.)

The situation is analogous for Deutsch's function *f*. Here the output of *f* can be represented as either 00 + 10 or 01 + 11 (in the ‘constant’ case), or 00 + 11 or 01 + 10 (in the ‘balanced’ case). The two entangled states in the ‘constant’ case are orthogonal in the 4-dimensional two-qubit state space and span a plane. Call this the ‘constant’ plane. Similarly, the two entangled states in the ‘balanced’ case span a plane, the ‘balanced’ plane. These planes are orthogonal in the 4-dimensional state space, except for an overlap: a line, representing a (non-entangled) two-qubit state. It is therefore possible to design a measurement to distinguish the two global properties of *f*, ‘constant’ or ‘balanced,’ with a certain probability (actually, 1/2) of failure, when the measurement yields an outcome corresponding to the overlap state, which is common to the two cases. Nevertheless, only one query of the function is required when the measurement succeeds in identifying the global property. With a judicious choice of quantum gates, it is even possible to design a quantum circuit that always succeeds in distinguishing the two cases.

Deutsch's example shows how quantum information, and quantum entanglement, can be exploited to compute a global property of a function in one step that would take two steps classically. In general, though, the potential speed-up in quantum computation relative to classical computation is exponential. Essentially, this is again due to the phenomenon of entanglement. Indeed, the amount of information required to describe a general entangled state of n qubits grows exponentially with n. The state space (Hilbert space) has 2^n dimensions, so a general entangled state is a superposition of 2^n n-qubit states. In classical mechanics there are no entangled states: a general n-bit composite system can be described with just n times the amount of information required to describe a single bit system. So the classical simulation of a quantum process would involve an exponential increase in the classical informational resource required to represent the quantum state, as the number of qubits that become entangled in the evolution grows linearly, and there would be a corresponding exponential slowdown in calculating the evolution, compared to the actual quantum computation performed naturally by the system.

While Deutsch's problem is in a sense trivial and has no interesting application, there now exist several quantum algorithms for non-trivial problems, notably Shor's factorization algorithm for factoring large composite integers in polynomial time (with direct application to ‘public key’ cryptography, a widely used classical cryptographic scheme) and Grover's database search algorithm. (See Nielsen and Chuang, or Barenco's article "Quantum Computation: An Introduction" in Lo, Popescu, and Spiller for details.)

Interpretative Remarks

The explanation favoured by Deutsch and others of how a quantum system processes information is the so-called ‘many-worlds’ interpretation. The idea, roughly, is that an entangled state of the sort that arises in the quantum computation of a function, which represents a linear superposition over all possible arguments and corresponding values of the function, should be understood as a manifestation of parallel computations in different worlds. The quantum circuit is designed to enable the computation of a global

property of the function by achieving some sort of 'interference' between these different worlds. (For an insightful critique of this idea of 'quantum parallelism' as explanatory, see Steane. It should be noted that the term 'many-worlds' can refer to a variety of interpretational ideas, some more refined than others.)

An alternative view, not much discussed in the literature in this connection, is the quantum logical interpretation, which emphasizes the non-Boolean structure of properties of quantum systems. (The properties of a classical system form a Boolean algebra, essentially the abstract characterization of a set-theoretic structure. This is reflected in the Boolean character of classical logic, and the Boolen gates in a classical computer.) A crucial difference between quantum and classical information is the possibility of computing the truth value of an exclusive disjunction -- for example, the 'constant' disjunction asserting that the value of the function (for both arguments) is either 0 or 1, or the 'balanced' disjunction asserting that the value of the function (for both arguments) is either the same as the argument or different from the argument -- without computing the truth values of the disjuncts. Classically, an exclusive disjunction is true if and only if one of the disjuncts is true. In effect, Deutsch's quantum circuit achieves its speed-up by exploiting the non-Boolean structure of quantum properties to compute the value of a disjunctive property, without computing the value of the disjuncts (representing the association of individual arguments with corresponding function values).

Bibliography

- Aspect, A., Grangier, P., Roger, G., "Experimental Tests of Bell's Inequalities Using Time-Varying Analyzers" *Physical Review Letters* **49** (1982): 1804-1807
- Bell, J.S., "On the Einstein-Podolsky-Rosen Paradox" *Physics* **1** (1964): 195-200
- Bennett, C.H., DiVicenzo, B.D., "Quantum Information and Computation" *Nature* **404** (2000): 247-255
- Bohr, N., "Can Quantum-Mechanical Description of Physical Reality be Considered Complete?" *Physical Review* **38** (1935): 696-702
- Born, M. (ed.), *The Born-Einstein Letters* (Dordrecht: Reidel, 1992)
- Cover, T.M., Thomas, J.A., *Elements of Information Theory* (New York: Wiley, 1991)
- Deutsch, D., "Quantum Theory, the Church-Turing principle and the Universal Quantum Computer" *Proceedings of the Royal Society (London)* **A400** (1985): 97-117
- Deutsch, D., *The Fabric of Reality* (London: Penguin, 1997)
- Dieks, D., "Communication by EPR Devices" *Physics Leters A* **92** (1982): 271-272
- Einstein, A., Podolsky, B., Rosen, N., "Can Quantum-Mechanical Description of Physical Reality be Considered Complete?" *Physical Review* **47** (1935): 777-780
- Holevo, A.S., "Statistical Problems in Quantum Physics" in G. Murayama and J.V. Prokhorov (eds) *Proceedings of the Second Japan-USSR Symposium on Probability Theory*, pp. 104-109 (Berlin: Springer, 1973)
- Kent, A., "Unconditionally Secure Bit Commitment" *Physical Review Letters* **83** (1999): 1447-1450
- Lo, H.-K., Chau, H.F., "Is Quantum Bit Commitment Really Possible?" *Physical Review Letters* **78** (1997): 3410--3413

- Lo, H.-K., Popescu, S., Spiller, T., *Introduction to Quantum Computation and Information* (Singapore: World Scientific, 1998)
- Mayers, D., "Unconditionally Secure Quantum Bit Commitment is Impossible" *Physical Review Letters* **78** (1997): 3414-3417
- Nielsen, M.A., Chuang, I.L., *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press, 2000)
- Schrödinger, E., "Discussion of Probability Relations Between Separated Systems," *Proceedings of the Cambridge Philosophical Society* **31** (1935): 555-563; **32** (1936): 446-451
- Schumacher, B., "Quantum Coding" *Physical Review A* **51** (1995): 2738-2747
- Shannon, C.E., Weaver, W., *The Mathematical Theory of Communication* (Urbana: University of Illinois Press, 1949)
- Steane, A.M., "A Quantum Computer Needs Only One Universe" LANL Preprint Archive for Quantum Physics, quant-ph/0003084
- van Fraassen, B., "The Charybdis of Realism: Epistemological Implications of Bell's Inequality" *Synthese* **52** (1982): 25-38
- Wootters, W.K., Zurek, W.H., "A Single Quantum Cannot be Cloned" *Nature* **299** (1982): 802-803

Other Internet Resources

- LANL Preprint Archive for Quantum Physics
- Quantum Computation Archive
- David Mermin's Home Page
- Oxford Centre for Quantum Computation
- Home Pages of Researchers on Quantum Information

Related Entries

Bell’s Theorem | physics: Reichenbach’s common cause principle | quantum mechanics: Copenhagen interpretation of | quantum mechanics: many-worlds interpretation of | quantum theory: the Einstein-Podolsky-Rosen argument in

Jeffrey Bub
jbub@carnap.umd.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 13, 2001
Content last modified: August 13, 2001

Reichenbach's Common Cause Principle

Suppose that two geysers, about one mile apart, erupt at irregular intervals, but usually erupt almost exactly at the same time. One would suspect that they come from a common source, or at least that there is a common cause of their eruptions. And this common cause surely acts before both eruptions take place. This idea, that simultaneous correlated events must have prior common causes, was first made precise by Hans Reichenbach (Reichenbach 1956). It can be used to infer the existence of unobserved and unobservable events, and to infer causal relations from statistical relations. Unfortunately it does not appear to be universally valid, nor is there agreement as to the circumstances in which it is valid.

Reichenbach's Common Cause Principle

It seems that a correlation between events A and B indicates either that A causes B, or that B causes A, or that A and B have a common cause. It also seems that causes always occur before their effects and, thus, that common causes always occur before the correlated events. Reichenbach was the first to formalize this idea rather precisely. He suggested that when Pr(A&B) > Pr(A)xPr(B) for simultaneous events A and B, there exists an earlier common cause C of A and B, such that Pr(A/C) > Pr(A/~C), Pr(B/C) > Pr(B/~C), Pr(A&B/C) = Pr(A/C)xPr(B/C) and Pr(A&B/~C) = Pr(A/~C)xPr(B/~C). (See Reichenbach 1956 pp. 158-159.) C is said to ‘screen off´ the correlation between A and B when A and B are uncorrelated conditional upon C. Thus Reichenbach´s principle can also be formulated as follows: simultaneous correlated events have a prior common cause that screens off the correlation.[1, 2]

Reichenbach's common cause principle needs to be modified. Consider, for instance, the following example. Harry normally takes the 8 a.m. train from New York to Washington. But he does not like full trains, so if the 8 a.m. train is full he sometimes takes the next train. He also likes trains that have diner cars, so if the 8 a.m. train does not have a diner car he

sometimes takes the next train. If the 8 a.m. train is both full and has no diner car, he is very likely to take the next train. Johnny, an unrelated commuter, also normally takes the 8 a.m. train from New York to Washington. Johnny, it so happens, also does not like full trains, and he also likes diner cars. Whether or not Harry and Johnny take the 8 a.m. train will therefore be correlated. But, since the probability of Harry and Johnny taking the 8 a.m. train depends on the occurrence of two distinct events (the train being full, the train having a diner car) there is no single event C, such that conditional upon C and conditional upon ~C we have independence. Thus Reichenbach's common cause principle as stated above is violated. Yet this example clearly does not violate the spirit of Reichenbach's common cause principle, for there is a partition into four possibilities such that conditional upon each of these four possibilities the correlation disappears.

More generally, we would like to have a common cause principle for cases in which the common causes and the effects are sets of quantities with continuous or discrete sets of values, rather than single events that occur or do not occur. A natural way to modify Reichenbach's common cause principle in order to deal with such types of cases is as follows. If simultaneous values of quantities A and B are correlated, then there are common causes C_1, C_2,....., C_n, such that conditional upon any combination of values of these quantities at an earlier time, the values of A and B are probabilistically independent. (For a fuller discussion of modifications like this, including cases in which there are correlations between more than two quantities, see Uffink (forthcoming)). I will continue to call this generalization ‘Reichenbach's common cause principle’, since, in spirit, it is very close to the principle that Reichenbach originally stated.

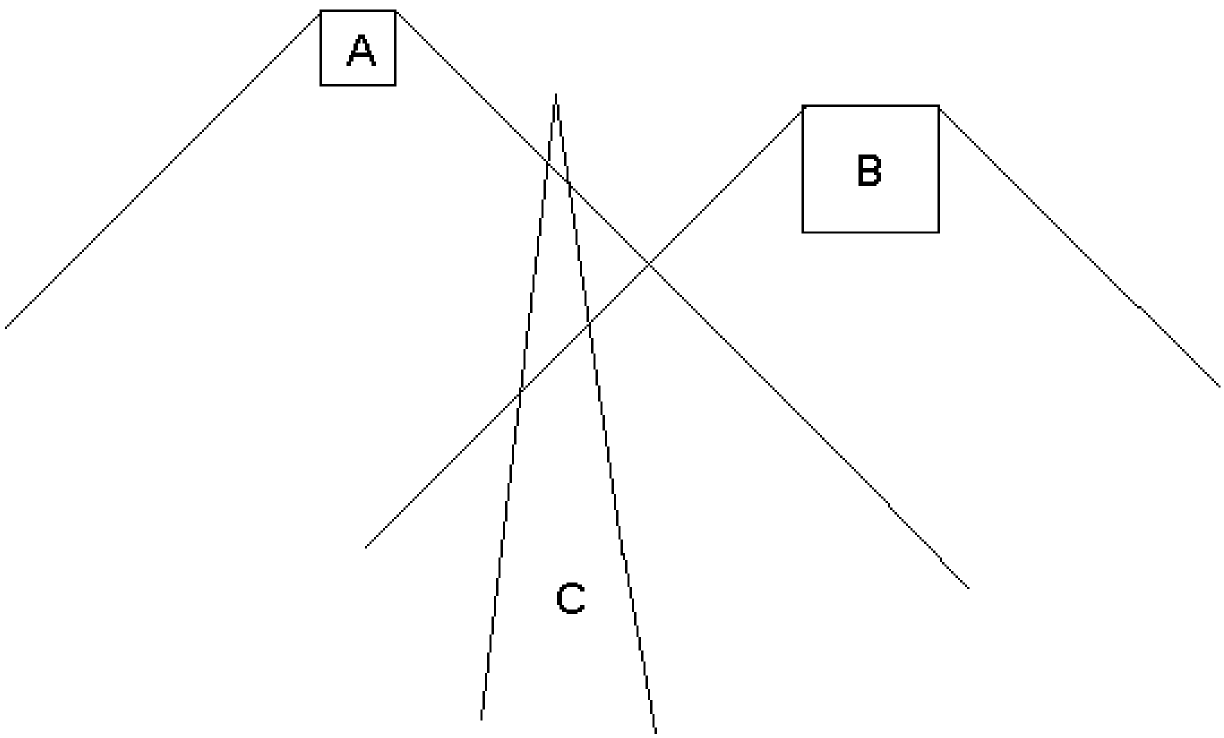
Now let me turn to two principles, the ‘causal Markov condition’ and the ‘law of conditional independence’, that are closely related to Reichenbach's common cause principle.

The Causal Markov Condition

There is a long tradition of attempts to infer causal relations among a set of quantities from probabilistic facts about the values of these quantities. In order to be able to do so, one needs principles relating causal facts and probabilistic facts. A principle that has been used to great effect in Spirtes, Glymour & Scheines 1993, is the ‘causal Markov condition’. This principle holds of a set of quantities $\{Q_1,......,Q_n\}$ if and only if the values of any quantity Q_i in that set, conditional upon the values of all the quantities in the set that are direct causes of Q_i, are probabilistically independent of the values of all quantities in the set other than Q_i's effects.[3] The causal Markov condition implies the following version of the common cause principle: If Q_i and Q_j are correlated and Q_i is not a cause of Q_j, and Q_j is not a cause of Q_i, then there are common causes of Q_i and Q_j in the set $\{Q_1,......,Q_n\}$ such that Q_i and Q_j are independent conditional upon these common causes.[4]

The Law of Conditional Independence

Penrose and Percival (1962), following Costa de Beauregard, have suggested as a general principle that the effects of interactions are felt after those interactions rather than before. In particular, they suggest that a system that has been isolated throughout the past is uncorrelated with the rest of the universe. Of course, this is almost a vacuous claim, since, other than in the case of horizons in cosmology, there would not appear to be a surfeit of systems that have been completely isolated from the rest of the universe throughout the past. Penrose and Percival, however, strengthen their principle by claiming that if one sets up a ‘statistical barrier’ that prevents any influences from acting both upon a space-time region A and a space-time region B, then states a in A and b in B will be uncorrelated. Penrose and Percival use the assumption that influences can not travel faster than the speed of light to make this idea more precise. Consider a space-time region C where there is no point P to the past of A or B such that one can travel, at a speed no faster than the speed of light, both from P to A and from P to B without entering C.



Penrose and Percival then say that one can prevent any influence from acting on both A and B by fixing the state c throughout such a region C. They therefore claim that states a in A and b in B will be uncorrelated conditional upon any state c in C. To be precise, they suggest the ‘law of conditional independence’: "If A and B are two disjoint 4-regions, and C is any 4-region which divides the union of the pasts of A and B into two parts, one containing A and the other containing B, then A and B are conditionally independent given c. That is, Pr(a&b/c) = Pr(a/c)xPr(b/c), for all a,b." (Penrose and Percival 1962, p. 611).

This is a time asymmetric principle which is clearly closely related to Reichenbach's common cause principle and the causal Markov condition. However one should not take states c in region C to be, or include, the common causes of the (unconditional) correlations that might exist between the states in regions A and B. It is merely a region such that influences from a past common source on both A and B must pass through it, assuming that such influences do not travel at speeds exceeding the speed of light. Note also that the region must stretch to the beginning of time. Thus, one cannot derive anything like Reichenbach's common cause principle or the causal Markov condition from the law of conditional independence, and one therefore would not inherit the richness of applications of these principles, especially the causal Markov condition, even if one were to accept the law of conditional independence.

Let us now turn to reasons that have been given for not believing any of these principles.

Conserved Quantities, Indeterminism and Quantum Mechanics

Suppose that a particle decays into 2 parts, that conservation of total momentum obtains, and that it is not determined by the prior state of the particle what the momentum of each part will be after the decay. By conservation, the momentum of one

part will be determined by the momentum of the other part. By indeterminism, the prior state of the particle will not determine what the momenta of each part will be after the decay. Thus there is no prior screener off. By simultaneity and symmetry, it is implausible to suppose that the momentum of the one part causes the momentum of the other part. So common cause principles fail. (This example is from van Fraassen (1980), page 29.)

More generally, suppose that there is a quantity Q, which is a function $f(q_1,....,q_n)$ of quantities q_i. Suppose that some of the quantities q_i develop indeterministically, but that quantity Q is conserved in such developments. There will then be correlations among the values of the quantities q_i which have no prior screener off. The only way that common cause principles can hold when there are conserved global quantities is when the development of each of the quantities that jointly determine the value of the global quantity is deterministic. And then it holds in the trivial sense that the prior determinants make everything else irrelevant. The results of quantum mechanical measurements are not determined by the quantum mechanical state prior to those measurements. And often there are conserved quantities during such a measurement. For instance, the total spin of 2 particles in a quantum 'singlet' state is 0. This quantity is conserved when one measures the spins of each of those 2 particles in the same direction: one will always find opposite spins during such a measurement, i.e. the spins that one finds will be perfectly anti-correlated. However what spins one will find is not determined by the prior quantum state. Thus the prior quantum state does not screen off the anti-correlations. There is no quantum common cause of such correlations.

One might think that this violation of common cause principles is a reason to believe that there must then be more to the prior state of the particles than the quantum state; there must be 'hidden variables' that screen off such correlations. (And we have seen above that such hidden variables must determine the results off the measurements if they are to screen of the correlations.) However, one can show, given some extremely plausible assumptions, that there cannot be any such hidden variables. (See, for instance, van Fraassen 1982 and Elby 1992 for more detail.)

One can also show that such correlations without a possible prior screener off are not confined to very special states, but occur generically in quantum mechanics and quantum field theory. (See Redhead (1995), Clifton, Feldman, Halvorson, Redhead & Wilce (1998), and Clifton & Ruetsche (forthcoming).)

Electromagnetism; Laws of Coexistence

Maxwell's equations not only govern the development of electromagnetic fields, they also imply simultaneous (in all frames of reference) relations between charge distributions and electromagnetic fields. In particular they imply that the electric flux through a surface which encloses some region of space must equal the total charge in that region. Thus electromagnetism implies that there is a strict and simultaneous correlation between the state of the field on such a surface and the charge distribution in the region contained by that surface. And this correlation must hold even on the space-like boundary at the beginning of the universe (if there be such). This violates all three common cause principles. (For more detail and subtlety, see Earman 1995, chapter 5).

More generally, any coexistence law, such as Newtonian gravitation, or Pauli's exclusion principle, will imply correlations which have no prior common cause conditionally upon which they disappear. Therefore, contrary to what one might hope, there are relativistic co-existence laws which violate common cause principles.

Bread and Water; Similar Laws of Evolution

The bread prices in Britain have been going up steadily over the last few centuries. The water levels in Venice have been going up steadily over the last few centuries. There is therefore a correlation between (simultaneous) bread prices in Britain and sea levels in Venice. However, there is presumably no direct causation involved, nor a common cause. More generally, Elliot Sober (see Sober 1988) has suggested that similar laws of evolution of otherwise independent quantities can lead to

correlations for which no common cause exists.

There is a way of understanding common cause principles such that this example is not a counterexample to it. Suppose that in nature there are transition chances from values of quantities at earlier times to values of quantities at later times. (For more in this idea see Arntzenius 1997). One could then state a common cause principle as follows: conditional upon the values of all the quantities upon which the transition chances to quantities X and Y depend, X and Y will be probabilistically independent. In Sober's example, there are transition chances from earlier costs of bread to later costs of bread, and there are transition chances from earlier water levels to later water levels. Conditional upon earlier costs of bread, later costs of bread are independent of later water levels. A common cause principle formulated as above thus holds in this case. Of course, if one looks at a collection of (simultaneous) data for water levels and bread prices one will see a correlation due to similar laws of development (similar transition chances). But a common cause principle, understood in terms of transition chances, does not imply that there should be a common cause of this correlation. The data (which include these correlations) should be understood as evidence for what the transition chances in nature are, and it is those transition chances that could be demanded to satisfy a common cause principle.

Markov Processes

Suppose a particular type of object has 4 possible states: S_1, S_2, S_3 and S_4. Suppose that if such an object is in state S_i at time t, and is not interfered with (isolated), then at time t+1 it has probability ½ of being in the same state S_i, and probability ½ of being in state S_i+1, where we define 4 + 1 = 1 (i.e. '+' represents addition mod 4). Now suppose we put many such objects in state S_1 at time t = 0. Then at time t = 1 approximately half of the systems will be in state S_1, and approximately half will be in state S_2. Let us define property A to be the property that obtains precisely when the system is either in state S_2 or in state S_3, and let us define property B to be the property that obtains precisely when the system is either in state S_2 or in state S_4. At time t = 1 half of the systems are in state S_1, and therefore have neither property A nor property B, and the other half are in state S_2, so that they have both property A and property B. Thus A and B are perfectly correlated at t = 1. Since these correlations remain conditional on the full prior state (S_1), there can be no quantity such that conditional upon a prior value of this quantity A and B are uncorrelated. Thus all three principles fail in this case. One can generalize this example to all generic state-space processes with indeterministic laws of developments, namely Markov processes. At least, one can do this if one allows arbitrary partitions of state-space to count as quantities. (In particular, therefore, Markov processes generically do not satisfy the causal Markov condition. The similarity of names is thus a bit misleading. See Arntzenius 1993 for more detail.)

Deterministic Systems

Suppose that the state of the world (or a system of interest) at any time determines the state of the world (that system) at any other time. It then follows that for any quantity X (of that system) at any time *t*, there will be at any other time t', in particular any later time t', a quantity X' (to be precise: a partition of state-space) such that the value of X' at t' uniquely determines the value of X at *t*. Conditional upon the value of X' at t', the value of X at *t* will be independent of any value of any quantity at any time. (For more detail see Arntzenius 1993.) Reichenbach's common cause principle thus fails in deterministic contexts. The problem is not that there will not always be earlier events conditional upon which the correlations disappear. Conditional upon the deterministic causes all correlations disappear. The problem is that there will also always be later events that determine whether the earlier correlated events occur. Reichenbach's common cause principle thus fails in so far as it claims that typically there are no later events conditional upon which earlier correlated simultaneous events are uncorrelated.

This does not imply a violation of the causal Markov condition. However, in order to be able to infer causal relations from statistical ones, Spirtes, Glymour and Scheines in effect assume that whenever (unconditionally correlated) quantities Q_i and

Q_j are independent conditional upon some quantity Q_k, then Q_k is a cause of either Q_i or Q_j. To be more precise they assume the ‘Faithfulness condition’, which states that there are no probabilistic independencies in nature other than the ones entailed by the causal Markov condition. Since the values of such quantities X' at later times t' surely are not direct causes of X at *t*, Faithfulness is violated, and with it goes our ability to infer causal relations from probabilistic relations, and much of the practical value of the causal Markov condition.[5]

A quantity like X' whose values at a later time t' are deterministically related to the values of X at *t*, will in general correspond to a non-natural, non-local, and not directly observable quantity. Perhaps a common cause principle can hold of some natural class of quantities. However, the next two sections will show that two suggestions for such a natural class do not work.

Macroscopic Quantities

Cleopatra is throwing a big party, and wants to sacrifice around fifty slaves to appease the gods. She is having a hard time convincing the slaves that this is a good idea, and decides that she ought to give them a chance at least. She has obtained a very strong poison, so strong that one molecule of it will kill a person. She puts one molecule of the poison in each of a hundred goblets of wine, which she presents to one hundred slaves. Having let the molecules of poison move around in Brownian motion for a while, she then orders the slaves to drink half a goblet of wine each. Let us now assume that if one consumes the poison, then death is preceded by an ominous reddening of the left hand and of the right hand. Then, the molecule being in the consumed half of the wine glass will be a prior screener off of the correlation between left hand reddening and right hand reddening. Assuming that death occurs exactly in the cases that the poison is swallowed, death will be a posterior screener off. If one restricts oneself to macroscopic events, there will only be a posterior screener off. If death is not strictly determined by the swallowing or non-swallowing of the poison, there will be no macroscopic screener off at any time. Thus, if microscopic events can have such macroscopic consequences, a common cause principle cannot hold of macroscopic events. More generally, this argument suggests that the common cause principle cannot hold of a class of events that has causes outside that class. This argument appears even more forceful for those who believe that the only reason that we can acquire knowledge of microscopic events and microscopic laws, is precisely the fact that microscopic events, in certain situations, have effects upon observable events.

Let us now consider another type of counterexample to the idea that a common cause principle can hold of macroscopic quantities, namely cases in which order arises out of chaos. When one lowers the temperature of certain materials, the spins of all the atoms of the material, which originally are not aligned, will line up in the same direction. Pick any two atoms in this structure. Their spins will be correlated. However, it is not the case that the one spin orientation caused the other spin orientation. Nor is there a simple or macroscopic common cause of each orientation of each spin. The lowering of the temperature determines that the orientations will be correlated, but not the direction in which they will line up. Indeed, typically, what determines the direction of alignment, in the absence of an external magnetic field, is a very complicated fact about the total microscopic prior state of the material and the microscopic influences upon the material. Thus, other than virtually the complete microscopic state of the material and its environment there is no prior screener off of the correlation between the spin alignments.

In general when chaotic developments result in ordered states there will be final correlations which have no prior screener off, other than virtually the full microscopic state of the system and its environment. (For more examples, see Prigogine 1980). In such cases the only screener off will be a horrendously complex microscopic quantity.

Local Quantities

If a common cause principle does not hold when one restricts oneself to macroscopic quantities, perhaps it holds if one restricts oneself to local quantities? Let me show that this is not so by giving a counterexample. There is a correlation between the take-off time of airplanes at airports and the time clothes take to dry on washing lines in any city near those

airports. An apparently satisfactory common cause explanation of this phenomenon is that high humidity causes both long drying times and long take-off times. However, this explanation presupposes that the humidity at the airport and at nearby houses is correlated. Now, it is not the case that the humidity in one area directly causes the humidity in other nearby areas. Moreover, there is no local common cause of the correlation among humidities in nearby areas, for there is no local earlier quantity that determines the humidity at separated locations at later times. Rather, the explanation of the correlation between the humidities in quite widely separated areas is that, when the total system is in (approximate) equilibrium then the humidity in different areas is (approximately) identical. Indeed the world is full of (approximate) equilibrium correlations, without local common causes conditional upon which these correlations disappear. (For more examples of this type of case see Forster 1986).

Next consider a flock of birds that flies, more or less, like a single unit in a rather varied trajectory through the sky. The correlation between the motions of each bird in the flock could have a rather straightforward common cause explanation: there could be a leader bird that every other bird follows. But it could also be that there is no leader bird, that each bird reacts to certain factors in the environment (presence of predator birds, insects, etc.), while at the same time constraining the distance that it will remove itself from its neighboring birds in the flock (as if tied to them by springs that pull harder the further away it gets from the other birds). In the latter case there will be a correlation of motions for which there is no local common cause. There will be an 'equilibrium' correlation that is maintained in the face of external perturbations. In 'equilibrium' the flock acts more or less as a unit, and reacts as a unit, possibly in a very complicated way, in response to its environment. The explanation of the correlation among the motions of its parts is not a common cause explanation, but the fact that in 'equilibrium' the myriad connections between its parts make it act as a unit.

In general we have learned to divide the world into systems which we regard as single units, since their parts normally (in 'equilibrium') behave in a highly correlated manner. We routinely do not regard correlations among the motions and properties of the parts of these systems as demanding a common cause explanation.

Initial Microscopic Chaos and the Common Cause Principle

Many authors have noted that there are circumstances in which the causal Markov condition, and the common cause principle that it implies, provably holds. Roughly speaking, this is the case when the world is deterministic, and the factors A and B which, in addition to the common cause C, determine whether effects D and E occur, are uncorrelated. Let me be more general and precise. Consider a deterministic world and a set of quantities S with certain causal relations holding between them. For any quantity Q, let us call the factors not in S which, when combined with the direct causes of Q that are in S, determine whether Q occurs, the 'determinants of Q outside S'. Suppose now that the determinants outside S are all independent, i.e. that the joint distribution of all determinants outside S is a product of distributions for each such determinant outside S. One can then prove that the causal Markov condition holds in S.[6]

But when should one expect such independence? P. Horwich (Horwich 1987) has suggested that such independence follows from initial microscopic chaos. (See also Papineau 1985 for a similar suggestion.) His idea is that if all the determinants outside S are microscopic, then they will all be uncorrelated since all microscopic factors will be uncorrelated when they are chaotically distributed. However, even if one has microscopic chaos (i.e. a uniform probability distribution in certain parts of state-space in a canonical coordinatization of the state-space), it is still not the case that all microscopic factors are uncorrelated. Let me give a generic counterexample.

Suppose that quantity C is a common cause of quantities A and B, that the system in question is deterministic, and that the quantities a and b which, in addition to C, determine the values of A and B are microscopic and independently distributed for each value of C. Then A and B will be uncorrelated conditional upon each value of C. Now define quantities D:A+B and E:A-B. ("+" and "-" here represent ordinary addition and subtraction of the values of quantities.) Then, generically, D and E will be correlated conditional upon each value of C. To illustrate why this is so let me give a very simple example. Suppose that for a given value of C quantities A and B are independently distributed, that A has value 1 with probability 1/2 and

value -1 with probability 1/2, and that B has value 1 with probability 1/2 and value -1 with probability 1/2. Then the possible values of D are -2, 0 and 2, with probabilities 1/4, 1/2 and 1/4 respectively. The possible values of E are also -2, 0 and 2, with probabilities 1/4, 1/2 and 1/4 respectively. But note, for instance, that if the value of D is -2, then the value of E must be 0. In general a non-zero value for D implies value 0 for E and a non-zero value for E implies value 0 for D. Thus, the values of D and E are strongly correlated for the given value of C. And it is not too hard to show that, generically, if quantities A and B are uncorrelated, then D and E are correlated. Now, since D and E are correlated conditional upon any value of C, it follows that C is not a prior common cause which screens off the correlation between D and E. And since the factors a and b which, in addition to C, determine the values of A and B, and hence those of D and E, can be microscopic and horrendously complex, there will be no screener off of the correlations between D and E other than some incredibly complex and inaccessible microscopic determinant. Thus common cause principles fail if one uses quantities D and E rather than quantities A and B to characterize the later state of the system.

One might try to save common cause principles by suggesting that in addition to C being a cause of D and of E, D is also a cause of E, or E is also a cause of D. (See Glymour and Spirtes 1994, pp 277-278 for such a suggestion). This would explain why D and E are still correlated conditional upon C. Nonetheless, this does not seem a plausible suggestion. In the first place, D and E are simultaneous. In the second place, the situation sketched is symmetric with respect to D and E, so which is supposed to cause which? It seems far more plausible to admit that common cause principles fail if one uses quantities D and E.

One might next try to defend common cause principles by suggesting that D and E are not really independent quantities, given that each is defined in terms of A and B, and that one should only expect common cause principles to be true of good, honest, independent quantities. Although this argument is along the right lines, as it stands it is too quick and simple. One cannot say that D and E are not independent because of the way they are defined in terms of A and B. For similarly A = ½(D+E) and B = ½(D-E), and unless there are reasons independent of such equations to claim that A and B are bona fide independent quantities while D and E are not, one is stuck. For now let us therefore conclude that an attempt to prove the common cause principle by assuming that all microscopic factors are uncorrelated rests on a false premise.

Nonetheless such arguments are pretty close to being correct: microscopic chaos does imply that a very large and useful class of microscopic conditions are independently distributed. For instance, assuming a uniform distribution of microscopic states in macroscopic cells, it follows that the microscopic states of two spatially separated regions will be independently distributed, given any macroscopic states in the two regions. Thus microscopic chaos and spatial separation is sufficient to provide independence of microscopic factors. This in fact covers a very large and useful class of cases. For almost all correlations that we are interested in are between factors of systems that are not exactly in the same location. Consider, for instance, an example due to Reichenbach.

Suppose that two actors almost always eat the same food. Every now and then the food will be bad. Let us assume that whether or not each of the actors become sick depends on the quality of the food that they consume and on other local factors (properties of their body etc.) at the time of consumption (and perhaps also later), which previously have developed chaotically. The values of these local factors for one of the actors will then be independent of the values of these local factors for the other actor. It then follows that there will be a correlation between their states of health, and that this correlation will disappear conditional upon the quality of the food. In general when one has a process that physically splits into two separate processes which remain separated in space, then all the ‘microscopic’ influences on those two processes will be independent from then on. Indeed there are very many cases in which two processes, whether spatially separated or not, will have a point after which microscopic influences on the processes are independent given microscopic chaos. In such cases common cause principles will be valid as long as one chooses as one's quantities the (relevant aspects of the) macroscopic states of the processes at the time of such separations (rather than the macroscopic states significantly prior to such separations) and some aspects of macroscopic states somewhere along each separate process (rather than some amalgam of quantities of the separate processes).

Conjectural and Provocative Conclusions

Reichenbach's principle of the common cause and its cousins, insofar as they hold, have the same origin as the temporal asymmetries of statistical mechanics, namely, roughly speaking, initial microscopic chaos. (I am being very rough here. There is no absolute, dynamics independent, distinction between microscopic and macroscopic factors. For more detail on exactly which quantities will behave as if they are uniformly distributed in which circumstances see, e.g., D. Albert (1999).) This explains why the three principles we have discussed sometimes fail. For the demand of initial microscopic chaos is a demand that microscopic conditions are uniformly distributed (in canonical coordinates) in the areas of state-space that are compatible with the fundamental laws of physics. If there are fundamental (equal time) laws of physics that rule out certain areas in state-space, which thus imply that there are (equal time) correlations among certain quantities, this is no violation of initial microscopic chaos. But the three common cause principles that we discussed will fail for such correlations. Similarly, quantum mechanics implies that for certain quantum states there will be correlations between the results of measurements that can have no common cause which screens all these correlations off. But this does not violate initial microscopic chaos. Initial microscopic chaos is a principle that tells one how to distribute probabilities over quantum states in certain circumstances; it does not tell one what the probabilities of values of observables given certain quantum states should be. And if they violate common cause principles, so be it. There is no fundamental law of nature that is, or implies, a common cause principle. The extent of the truth of common cause principles is approximate and derivative, not fundamental.

One should also not be interested in common cause principles which allow any conditions, no matter how microscopic, scattered and unnatural, to count as common causes. For, as we have seen, this would trivialize such principles in deterministic worlds, and would hide from view the remarkable fact that when one has a correlation among fairly natural localized quantities that are not related as cause and effect, almost always one can find a fairly natural, localized prior common cause that screens off the correlation. The explanation of this remarkable fact, which was suggested in the previous section, is that Reichenbach's common cause principle, and the causal Markov condition, must hold if the determinants, other than the causes, are independently distributed for each value of the causes. The fundamental assumptions of statistical mechanics imply that this independence will hold in a large class of cases given a judicious choice of quantities characterizing the causes and effects. In view of this, it is indeed more puzzling why common cause principles fail in cases like those described above, such as the coordinated flights of certain flocks of birds, equilibrium correlations, order arising out of chaos, etc. The answer is that in such cases the interactions between the parts of these systems are so complicated, and there are so many causes acting on the systems, that the only way one can get independence of further determinants is by specifying so many causes as to make this a practical impossibility. This, in any case, would amount to allowing just about any scattered and unnatural set of factors to count as common causes, thereby trivializing common cause principles. Thus, rather than do that, we regard such systems as single unified systems, and do not demand a common cause explanation for the correlated motions and properties of their parts. A fairly intuitive notion of what counts as a single system, after all, is a system that behaves in a unified manner, i.e. a system whose parts have a very strong correlation in their motions and/or other properties, no matter how complicated the set of influences acting on them. For instance a rigid physical object has parts whose motions are all correlated, and a biological organism has parts whose motions and properties are strongly correlated, no matter how complicated the influences acting on it. These systems therefore are naturally and usefully treated as single systems for almost any purpose. The core truth of common cause principles thus in part relies on our choice as to how to partition the world into unified and independent objects and quantities, and in part on the objective, temporally asymmetric, principles that lie at the foundation of statistical mechanics.

Bibliography

- Albert, D (1999): *Chance and Time*, Boston: Harvard University Press.
- Arntzenius, F. (1993): "The common cause principle", *PSA* 1992, **2**: 227-237.
- Arntzenius, F. (1997): "Transition chances and causation", *Pacific Philosophical Quarterly* **78** (2): 149-168.
- Clifton, R., Feldman, D., Halvorson, H., Redhead, M. & Wilce, A. (1998): "Superentangled states", *Physical Review A* **58**: 135-145
- Clifton, R. & Ruetsche, L. (forthcoming): "Changing the subject: Redei on causal dependence and screening off in algebraic quantum field theory", *Philosophy of Science (PSA 1998 Proceedings Supplement)*.

- Earman, J. (1995): *Bangs, crunches, whimpers and shrieks*, Oxford, Oxford University Press.
- Elby (1992): "Should we explain the EPR correlations causally?", *Philosophy of Science* **59** (1): 16-25.
- Forster, M. (1986), "Unification and Scientific Realism revisited", in *PSA* 1986, **1**: 394-405.
- Glymour, C. & Spirtes, P. (1994): "Selecting variables and getting to the truth", pp 273-280 in D. Stalker (ed, 1994): *Grue! The new riddle of induction*, Open Court.
- Horwich, P. (1987), *Asymmetries in Time*. Cambridge: MIT Press.
- Papineau, D. (1985), "Causal Asymmetry", *British Journal for the Philosophy of Science*, **36**: 273-289.
- Prigogine, I. (1980), *From Being to Becoming*. San Francisco: W. H. Freeman.
- Redhead, M. (1995): "More ado about nothing", *Foundations of Physics* **25**: 123-137.
- Reichenbach, H. (1956): *The Direction of Time*, Berkeley, University of Los Angeles Press.
- Sober, E. (1988), "The Principle of the Common Cause", in *Probability and Causality*, J. Fetzer (ed.). Dordrecht: Reidel, pp 211-229.
- Spirtes, P., Glymour, C. & Scheines, R. (1993), *Causation, Prediction and Search*. Springer Verlag.
- Uffink (forthcoming): "Reichenbach's common cause principle", *Philosophy of Science, Proceedings of the 1998 biennial meeting of the Philosophy of Science Association*.
- Van Fraassen, B. (1980), *The Scientific Image*. Oxford: Clarendon Press.
- Van Fraassen, B. (1982), "The Charybdis of Realism: Epistemological Implications of Bell's Inequality", *Synthese* **52**: 25-38.

Other Internet Resources

- Hans Reichenbach (Internet Encyclopedia of Philosophy)

Related Entries

causation: probabilistic | cause and effect | probability calculus: interpretations of | quantum mechanics | Reichenbach, Hans

Frank Arntzenius
arntzeni@rci.rutgers.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 22, 1999
Content last modified: September 22, 1999

Stanford Encyclopedia of Philosophy
Notes to Reichenbach's Common Cause Principle

Notes

1. Reichenbach's common cause principle is a time-asymmetric principle. One could make this asymmetry more explicit by adding the claim that, typically, there are no later events E conditional upon which A and B are uncorrelated. As with all time-asymmetric principles, one might wonder whether nature really does exhibit such an asymmetry, and if indeed it does, what its origin is, and how it relates to other temporal asymmetries. Reichenbach himself wanted to use the common cause asymmetry as part of a definition of the distinction between the future and the past. On his view, then, it is not interesting to ask why a common cause C occurs before its effects A and B; for that is true by definition of the direction of time. But on his view it remains an interesting question as to why all common causes are always placed in the same direction of time relative to their effects, and how this asymmetry relates to other temporal asymmetries such as thermodynamic asymmetries.

2. One might wish to understand the probabilistic relations stated in this principle as a definition of what a common cause is. However, this is not a plausible idea. For, consider a case in which A_1 is a common cause of A_2 and A_3, while A_2 in its turn causes A_4, and A_3 in its turn causes A_5, where A_4 and A_5 are simultaneous. In this case A_4 and A_5 will typically be uncorrelated conditional upon A_2 (assuming that A_2 is the only cause of A_4). Moreover, an occurrence of A_2 will typically make an occurrence of A_4 more likely, and, typically, it will make an occurrence of A_5 more likely. Thus A_2 would count as a common cause of A_4 and A_5 if the above probabilistic relations were a definition of what it is to be a common cause. But by assumption A_2 is not a cause of A_5, and thus not a common cause of A_4 and A_5. Thus the probabilistic relations that are used to state Reichenbach's principle of the common cause should not be regarded as a definition of what a common cause is.

3. Q_k is an indirect cause of Q_p exactly when there is a chain of direct causes starting at Q_k and ending at Q_p. Q_p is an effect of Q_k iff Q_k is a cause of Q_p iff Q_k is a direct or indirect cause of Q_p.

4. Let me explain how this follows from the Markov condition using terminology and theorems from Spirtes, Glymour and Scheines 1993, chapter 3. Let us call a quantity Q_c a ‘forking path common cause’ of quantities Q_a and Q_b if and only if there exists a path that always goes causally downstream from Q_c to Q_a, a path that always goes causally downstream from Q_c to Q_b, and these paths do not have any vertices other than Q_c in common. Let us now conditionalize on all the forking path common causes of Q_a and Q_b. Claim: Every path from Q_a to Q_b will then be inactive. (Paths here are assumed not to contain any vertex more than once.) Let me indicate how to prove this claim.

If such a path P starts at Q_a going causally downstream, then it will at some point have to switch to going causally upstream, since Q_a is not a cause of Q_b. It will thus contain a collider. This collider will be inactive since we are not conditionalizing on it nor on any quantity that is a descendent of it, since we are conditionalizing only on quantities that are causally upstream from Q_a. (I am assuming that the graphs are not cyclical.) Thus such a path P will be inactive. Now consider any path P from Q_a to Q_b that starts at Q_a going causally upstream. There must be some point Q_c at which point P first starts going causally downstream (since Q_b is not a cause of Q_a). There are two possibilities: P keeps going causally downstream all the way until it reaches Q_b, or it reaches a collider Q_d where P switches back to going causally upstream again. In the first case, Q_c is a common cause of Q_a and Q_b. So we have conditionalized on it, so P is inactive. In the second case, Q_d is a collider. There are then 2 subcases: Q_d does not lie causally upstream from a forking path common cause of Q_a and Q_b, or it does. If Q_d does not lie causally upstream from a common cause, then it does not lie upstream from a quantity that is conditionalized upon, so the collider is inactive, so P is inactive. If Q_d does lie causally upstream from a forking path common cause of Q_a and Q_b, then there must be a downstream path P' from Q_d to Q_b. There are now 2 sub-subcases to consider.

Sub-subcase i: P' has no vertices in common with the part of path P that lies between Q_a and Q_c. In that case, Q_c is a forking path common cause of Q_a and Q_b: to go downstream from Q_c to Q_a, follow the part of P that lies between Q_c and Q_a;to go downstream from Q_c to Q_b (without intersecting the path to Q_a), first take the part of P that lies between Q_c and Q_d, and then follow P' to Q_b. So in this case we have conditionalized upon Q_c on P, so P is inactive.

Sub-subcase ii: every downstream path P' from Q_d to Q_b intersects somewhere with the part of P that lies between Q_a and Q_c. Take some such P', and consider the furthest point Q_e downstream along P' at which P' intersects with P between Q_a and Q_c. Such a Q_e must be a forking path common cause of Q_a and Q_b: follow P to get to Q_a, follow P' to get to Q_b. Thus we have conditionalized upon Q_e which lies on P. Thus P is inactive.

Thus any path P from Q_a to Q_b must be inactive once we have conditionalized upon all forking path common causes. Thus Q_a and Q_b are independent conditional upon a subset of all common causes of Q_a and Q_b.

5. The law of conditional independence is not violated by this type of case.

6. Let me sketch a proof. Any probability distribution that is allowed by the independence condition can be generated as follows. Assign some probability distribution over all the determinants outside S. By assumption this must be a probability distribution that is jointly independent, i.e. a product of

distributions for each such determinant. Now first look at the set S_1 of quantities in S that have no direct causes in S. The probability distribution over these quantities will be determined by the distribution of their determinants outside S, and hence be a jointly independent distribution. Now look at the set S_2 of quantities all of whose direct causes in S are in S_1. The probability distribution over any quantity S_2 is obtained by multiplying the probability distributions of its direct causes in S_1 with the probability distribution of its determinant outside S. (At least, this is so if all distinct values of direct causes of Q in S and determinants of Q outside S, determine distinct values of Q. This may not be so, but this does not affect the independence claims that I am making here.) And let us continue in this way with S_3, ... until we have a distribution over all quantities in S. The only correlations in the joint distribution over quantities in S that will now occur will be between causes and their effects, and between the effects of a common cause. For consider any quantities Q_1 and Q_2 that are not so related. They will have no ‘ultimate inputs’ (the determinants outside S that determine the values of these quantities) in common, so the sets of ‘ultimate inputs’ for Q_1 and ‘ultimate inputs’ for Q_2 are independent, which entails that Q_1 and Q_2 are themselves independent. Moreover, the correlations between any two quantities Q_1 and Q_2 that are not related as cause and effect will disappear when one conditionalizes upon the direct causes of one of them, say Q_1. For the only remaining input into Q_1 is independent of anything other than effects of Q_1. So Q_1 is independent of anything other than effects of Q_1 conditional upon the direct causes of Q_1. Hence, the causal Markov condition holds.

Frank Arntzenius
arntzeni@rci.rutgers.edu

First published: September 22, 1999
Content last modified: September 22, 1999

The Copenhagen Interpretation of Quantum Mechanics

As the theory of the atom, quantum mechanics is perhaps the most successful theory in the history of science. It enables physicists, chemists, and technicians to calculate and predict the outcome of a vast number of experiments and to create new and advanced technology based on the insight into the behavior of atomic objects. But it is also a theory that challenges our imagination. It seems to violate some fundamental principles of classical physics, principles that eventually have become a part of western common sense since the rise of the modern worldview in the Renaissance. So the aim of any metaphysical interpretation of quantum mechanics is to account for these violations.

The Copenhagen interpretation was the first general attempt to understand the world of atoms as this is represented by quantum mechanics. The founding father was mainly the Danish physicist Niels Bohr, but also Werner Heisenberg, Max Born and other physicists made important contributions to the overall understanding of the atomic world that is associated with the name of the capital of Denmark.

In fact Bohr and Heisenberg never totally agreed on how to understand the mathematical formalism of quantum mechanics, and none of them ever used the term “the Copenhagen interpretation” as a joint name for their ideas. In fact, Bohr once distanced himself from what he considered to be Heisenberg’s more subjective interpretation (*APHK*, p.51). The term is rather a label introduced by people opposing Bohr’s idea of complementarity, to identify what they saw as the common features behind the Bohr-Heisenberg interpretation as it emerged in the late 1920s. Today the Copenhagen interpretation is mostly regarded as synonymous with indeterminism, Bohr’s correspondence principle, Born’s statistical interpretation of the wave function, and Bohr’s complementarity interpretation of certain atomic phenomena.

1. The Background

In 1900 Max Planck discovered that the radiation spectrum of black bodies occurs only with discrete energies separated by the value hv where v is the frequency and h is a new constant, the so-called Planck constant. According to classical physics the intensity of this continuous radiation would grow unlimitedly with growing frequencies, resulting in what was called the ultraviolet catastrophe. But Planck's suggestion was that if black bodies only exchange energy with the radiation field in a proportion equal to hv that problem would disappear. The fact that the absorption and the emission of energy is discontinuous is in conflict with the principles of classical physics. A few years later Albert Einstein used this discovery in his explanation of the photoelectric effect. He suggested that light waves were quantized, and that the amount of energy which each quantum of light could deliver to the electrons of the cathode, was exactly hv. The next step came in 1911 when Ernest Rutherford performed some experiments shooting alpha particles into a gold foil. Based on these results he could set up a model of the atom in which the atom consisted of a heavy nucleus with a positive charge surrounded by negatively charged electrons like a small solar system. Also this model was in conflict with the laws of classical physics. According to classical mechanics and electrodynamics one might expect that the electrons orbiting around a positively charged nucleus would continuously emit radiation so that the nucleus would quickly swallow the electrons.

At this point Niels Bohr entered the scene and soon became the leading physicist on atoms. In 1913 Bohr, visiting Rutherford in Manchester, put forward a mathematical model of the atom which provided the first theoretical support for Rutherford's model and could explain the emission spectrum of the hydrogen atom (the Balmer series). The theory was based on two postulates:

1. An atomic system is only stable in a certain set of states, called stationary states, each state being associated with a discrete energy, and every change of energy corresponds to a complete transition from one state to another.
2. The possibility for the atom to absorb and emit radiation is determined by a law according to which the energy of the radiation is given by the energy difference between two stationary states being equal to hv.

Some features of Bohr's semi-classical model were indeed very strange compared to the principles of classical physics. It introduced an element of discontinuity and indeterminism foreign to classical mechanics:

1. Apparently not every point in space was accessible to an electron moving around a hydrogen nucleus. An electron moved in classical orbits, but during its transition from one orbit to another it was at no definite place between these orbits. Thus, an electron could only be in its ground state (the orbit of lowest energy) or an excited state (if an impact of another particle had forced it to leave its ground state.)

2. It was impossible to predict when the transition would take place and how it would take place. Moreover, there were no external (or internal) causes that determined the “jump” back again. Any excited electron might in principle move spontaneously to either a lower state or down to the ground state.
3. Rutherford pointed out that if, as Bohr did, one postulates that the frequency of light v, which an electron emits in a transition, depends on the difference between the initial energy level and the final energy level, it appears as if the electron must “know” to what final energy level it is heading in order to emit light with the right frequency.
4. Einstein made another strange observation. He was curious to know in which direction the photon decided to move off from the electron.

Between 1913 and 1925 Bohr, Arnold Sommerfeld and others were able to improve Bohr’s model, and together with the introduction of spin and Wolfgang Pauli’s exclusion principle it gave a reasonably good description of the basic chemical elements. The model ran into problems, nonetheless, when one tried to apply it to spectra other than that of hydrogen. So there was a general feeling among all leading physicists that Bohr’s model had to be replaced by a more radical theory. In 1925 Werner Heisenberg, at that time Bohr’s assistant in Copenhagen, laid down the basic principles of a complete quantum mechanics. In his new matrix theory he replaced classical commuting variables with non-commuting ones. The following year, Erwin Schrödinger gave a simpler formulation of the theory in which he introduced a second-order differential equation for a wave function. He himself attempted a largely classical interpretation of the wave function. However, already the same year Max Born proposed a consistent statistical interpretation in which the square of the absolute value of this wave function expresses a probability amplitude for the outcome of a measurement.

2. Classical Physics

Bohr saw quantum mechanics as a generalization of classical physics although it violates some of the basic ontological principles on which classical physics rests. These principles are:

- *The principle of space and time*, *i.e*., physical objects (systems) exist separately in space and time in such a way that they are localizable and countable, and physical processes (the evolution of systems) take place in space and time;
- *The principle of causality*, i.e., every event has a cause;
- *The principle of determination*, i.e., every later state of a system is uniquely determined by any earlier state;
- *The principle of continuity*, i.e., all processes exhibiting a difference between the initial and the final state have to go through every intervening state; and finally
- *The principle of the conservation of energy*, i.e., the energy of a closed system can be transformed into various forms but is never gained, lost or destroyed.

Due to these principles it is possible within, say, classical mechanics, to define a state of a system at any later time with respect to a state at any earlier time. So whenever we know the initial state consisting of

the system's position and momentum, and know all external forces acting on it, we also know what will be its later states. The knowledge of the initial state is usually acquired by observing the state properties of the system at the time selected as the initial moment. Furthermore, the observation of a system does not affect its later behavior or, if observation somehow should influence this behavior, it is always possible to incorporate the effect into the prediction of the system's later state. Thus, in classical physics we can always draw a sharp distinction between the state of the measuring instrument being used on a system and the state of the physical system itself. It means that the physical description of the system is objective because the definition of any later state is not dependent on measuring conditions or other observational conditions.

Much of Kant's philosophy can be seen as an attempt to provide satisfactory philosophical grounds for the objective basis of Newton's mechanics against Humean scepticism. Kant showed that classical mechanics is in accordance with the transcendental conditions for objective knowledge. Kant's philosophy undoubtedly influenced Bohr in various ways as many scholars in recent years have noticed (Hooker 1972; Folse 1985; Honnor 1987; Faye 1991; Kaiser 1992; and Chevalley 1994). Bohr was definitely neither a subjectivist nor a positivist philosopher, as Karl Popper (1967) and Mario Bunge (1967) have claimed. He explicitly rejected the idea that the experimental outcome is due to the observer. As he said: "It is certainly not possible for the observer to influence the events which may appear under the conditions he has arranged" (*APHK*, p.51). Not unlike Kant, Bohr thought that we could have objective knowledge only in case we can distinguish between the experiential subject and the experienced object. It is a precondition for the knowledge of a phenomenon as being something distinct from the sensorial subject, that we can refer to it as an object without involving the subject's experience of the object. In order to separate the object from the subject itself, the experiential subject must be able to distinguish between the form and the content of his or her experiences. This is possible only if the subject uses causal and spatial-temporal concepts for describing the sensorial content, placing phenomena in causal connection in space and time, since it is the causal space-time description of our perceptions that constitutes the criterion of reality for them. Bohr therefore believed that what gives us the possibility of talking about an object and an objectively existing reality is the application of those necessary concepts, and that the physical equivalents of "space," "time," "causation," and "continuity" were the concepts "position," "time," "momentum," and "energy," which he referred to as *the classical concepts*. He also believed that the above basic concepts exist already as preconditions of unambiguous and meaningful communication, built in as rules of our ordinary language. So, in Bohr's opinion the conditions for an objective description of nature given by the concepts of classical physics were merely a refinement of the preconditions of human knowledge.

3. The Correspondence Rule

The guiding principle behind Bohr's and later Heisenberg's work in the development of a consistent theory of atoms was the correspondence rule. Bohr had realized that according to his theory of the hydrogen atom, the frequencies of radiation due to the electron's transition between stationary states with large quantum numbers, i.e. states far from the ground state, coincide approximately with the results of classical electrodynamics. Hence in the search for a theory of quantum mechanics it became a

methodological requirement to Bohr that any further theory of the atom should predict values in domains of large quantum numbers that should be a close approximation to the values of classical physics. The correspondence rule was a heuristic principle meant to make sure that in areas where the influence of Planck's constant could be neglected the numerical values predicted by such a theory should be the same as if they were predicted by classical radiation theory.

The correspondence rule was an important methodological principle. In the beginning it had a clear technical meaning for Bohr. It is obvious, however, that it makes no sense to compare the numerical values of the theory of atoms with those of classical physics unless the meaning of the physical terms in both theories is commensurable. The correspondence rule was based on the metaphysical idea that classical concepts were indispensable for our understanding of physical reality, and it is only when classical phenomena and quantum phenomena are described in terms of the same classical concepts that we can compare different physical experiences. It was this broader sense of the correspondence rule that Bohr often had in mind later on.

Bohr's practical methodology stands therefore in direct opposition to Thomas Kuhn and Paul Feyerabend's historical view that succeeding theories, like classical mechanics and quantum mechanics, are incommensurable. In contrast to their philosophical claims of meaning gaps and partial lack of rationality in the choice between incommensurable theories, Bohr believed not just retrospectively that quantum mechanics was a natural generalization of classical physics, but he and Heisenberg followed in practice the requirements of the correspondence rule. Thus, in the mind of Bohr, the meaning of the classical concepts did not change but their application was restricted. This was the lesson of complementarity.

4. Complementarity

After Heisenberg had managed to formulate a consistent quantum mechanics in 1925, both he and Bohr began their struggle to find a coherent interpretation for the mathematical formalism. Heisenberg and Bohr followed somewhat different approaches. Where Heisenberg looked to the formalism and developed his famous uncertainty principle or indeterminacy relation, Bohr chose to analyze concrete experimental arrangements, especially the double-slit experiment. In a way Bohr merely regarded Heisenberg's relation as an expression of his general notion that our understanding of atomic phenomena builds on complementary descriptions. At Como in 1927 he presented for the first time his ideas according to which certain different descriptions are said to be complementary.

Bohr pointed to two sets of descriptions which he took to be complementary. On the one hand there are those that attribute either kinematic or dynamic properties to the atom, that is "space-time descriptions" are complementary to "claims of causality." On the other hand there are those that ascribe either wave or particle properties to a single object. How these two kinds of complementary sets of descriptions are related is something Bohr never indicated (Murdoch 1987). Even among people, like Rosenfeld and Pais, who claim they speak on behalf of Bohr, there is no agreement. The fact is that the description of light as either particles or waves was already a classical dilemma, which not even Einstein's definition of a

photon really solved since the momentum of the photon as a particle depends on the frequency of the light as a wave. Furthermore, Bohr eventually realized that the attribution of kinematic and dynamic properties to an object is complementary because the ascription of both of these conjugate variables rests on mutually exclusive experiments. The attribution of particle and wave properties to an object may, however, occur in a single experiment; for instance, in the double-slit experiment where the interference pattern consists of single dots. So within less than ten years after his Como lecture Bohr tacitly abandoned “wave-particle complementarity” in favor of the exclusivity of “kinematic-dynamic complementarity” (Held 1994).

It was clear to Bohr that any interpretation of the atomic world had to take into account an important empirical fact. The discovery of the quantization of action meant that quantum mechanics could not fulfill the above principles of classical physics. Every time we measure, say, an electron’s position the apparatus and the electron interact in an uncontrollable way, so that we are unable to measure the electron's momentum at the same time. Until the mid-1930s when Einstein, Podolsky and Rosen published their famous thought-experiment with the intention of showing that quantum mechanics was incomplete, Bohr spoke as if the measurement apparatus disturbed the electron. This paper had a significant influence on Bohr’s line of thought. Apparently, Bohr realized that speaking of disturbance seemed to indicate—as some of his opponents may have understood him—that atomic objects were classical particles with definite inherent kinematic and dynamic properties. After the EPR paper he stated quite clearly: “the whole situation in atomic physics deprives of all meaning such inherent attributes as the idealization of classical physics would ascribe to such objects.”

Also after the EPR paper Bohr spoke about Heisenberg’s “indeterminacy relation” as indicating the ontological consequences of his claim that kinematic and dynamic variables are ill-defined unless they refer to an experimental outcome. Earlier he had often called it Heisenberg’s “uncertainty relation”, as if it were a question of a merely epistemological limitation. Furthermore, Bohr no longer mentioned descriptions as being complementary, but rather phenomena or information. He introduced the definition of a “phenomenon” as requiring a complete description of the entire experimental arrangement, and he took a phenomenon to be a measurement of the values of either kinematic or dynamic properties.

Bohr’s more mature view, i.e., his view after the EPR paper, on complementarity and the interpretation of quantum mechanics may be summarized in the following points:

1. The interpretation of a physical theory has to rely on an experimental practice.
2. The experimental practice presupposes a certain pre-scientific practice of description, which establishes the norm for experimental measurement apparatus, and consequently what counts as scientific experience.
3. Our pre-scientific practice of understanding our environment is an adaptation to the sense experience of separation, orientation, identification and reidentification over time of physical objects.
4. This pre-scientific experience is grasped in terms of common categories like thing’s position and change of position, duration and change of duration, and the relation of cause and effect, terms and principles that are now parts of our common language.

5. These common categories yield the preconditions for objective knowledge, and any description of nature has to use these concepts to be objective.
6. The concepts of classical physics are merely exact specifications of the above categories.
7. The classical concepts—and not classical physics itself—are therefore necessary in any description of physical experience in order to understand what we are doing and to be able to communicate our results to others, in particular in the description of quantum phenomena as they present themselves in experiments;
8. Planck's empirical discovery of the quantization of action requires a revision of the foundation for the use of classical concepts, because they are not all applicable at the same time. Their use is well defined only if they apply to experimental interactions in which the quantization of action can be regarded as negligible.
9. In experimental cases where the quantization of action plays a significant role, the application of a classical concept does not refer to independent properties of the object; rather the ascription of either kinematic or dynamic properties to the object as it exists independently of a specific experimental interaction is ill-defined.
10. The quantization of action demands a limitation of the use of classical concepts so that these concepts apply only to a phenomenon, which Bohr understood as the macroscopic manifestation of a measurement on the object, i.e. the uncontrollable interaction between the object and the apparatus.
11. The quantum mechanical description of the object differs from the classical description of the measuring apparatus, and this requires that the object and the measuring device should be separated in the description, but the line of separation is not the one between macroscopic instruments and microscopic objects. It has been argued in detail (Howard 1994) that Bohr pointed out that parts of the measuring device may sometimes be treated as parts of the object in the quantum mechanical description.
12. The quantum mechanical formalism does not provide physicists with a 'pictorial' representation: the ψ-function does not, as Schrödinger had hoped, represent a new kind of reality. Instead, as Born suggested, the square of the absolute value of the ψ-function expresses a probability amplitude for the outcome of a measurement. Due to the fact that the wave equation involves an imaginary quantity this equation can have only a symbolic character, but the formalism may be used to predict the outcome of a measurement that establishes the conditions under which concepts like position, momentum, time and energy apply to the phenomena.
13. The ascription of these classical concepts to the phenomena of measurements rely on the experimental context of the phenomena, so that the entire setup provides us with the defining conditions for the application of kinematic and dynamic concepts in the domain of quantum physics.
14. Such phenomena are complementary in the sense that their manifestations depend on mutually exclusive measurements, but that the information gained through these various experiments exhausts all possible objective knowledge of the object.

Earlier generations of philosophers have often accused the Copenhagen interpretation of being subjectivistic or positivistic. Today anyone who has studied Bohr's essays carefully agrees that his view is neither. There are, as many have noticed, both typically realist as well as antirealist elements involved

in it, and it has affinities to Kant or neo-Kantianism.

Bohr thought of the atom as real. Atoms are neither heuristic nor logical constructions. A couple of times he emphasized this directly using arguments from experiments in a very similar way to Ian Hacking and Nancy Cartwright much later. What he did not believe was that the quantum mechanical formalism was true in the sense that it gave us a literal ('pictorial') rather than a symbolic representation of the quantum world. It makes much sense to characterize Bohr in modern terms as an entity realist who opposes theory realism (Folse 1987). It is because of the imaginary quantities in quantum mechanics (where the commutation rule for canonically conjugate variable, p and q, introduces Planck's constant into the formalism by $pq - qp = ih/2\pi$) that quantum mechanics does not give us a 'pictorial' representation of the world. Neither does the theory of relativity, Bohr argued, provide us with a literal representation, since the velocity of light is introduced with a factor of i in the definition of the fourth coordinate in a four-dimensional manifold (*CC*, p. 86 and p. 105). Instead these theories can only be used symbolically to predict observations under well-defined conditions. Thus Bohr was an antirealist or an instrumentalist when it comes to theories.

In general, Bohr considered the demands of complementarity in quantum mechanics to be logically on a par with the requirements of relativity in the theory of relativity. He believed that both theories were a result of novel aspects of the observation problem, namely the fact that observation in physics is context-dependent. This again is due to the existence of a maximum velocity of propagation of all actions in the domain of relativity and a minimum of any action in the domain of quantum mechanics. And it is because of these universal limits that it is impossible in the theory of relativity to make an unambiguous separation between time and space without reference to the observer (the context) and impossible in quantum mechanics to make a sharp distinction between the behavior of the object and its interaction with the means of observation (*CC*, p. 105).

In emphasizing the necessity of classical concepts for the description of the quantum phenomena, Bohr was influenced by Kant or neo-Kantianism. But he was a naturalized or a pragmatized Kantian. The classical concepts are merely explications of common concepts that are already a result of our adaptation to the world. These concepts and the conditions of their application determine the conditions for objective knowledge. The discovery of the quantization of action has revealed to us, however, that we cannot apply these concepts to quantum objects as we did in classical physics. Now kinematic and dynamic properties (represented by conjugate variables) can be meaningfully ascribed to the object only in relation to some actual experimental results, whereas classical physics attributes such properties to the object regardless of whether we actually observe them or not. In other words, Bohr denied that classical concepts could be used to attribute properties to a physical world in-itself behind the phenomena, i.e. properties different from those being observed. In contrast, classical physics rests on an idealization, he said, in the sense that it assumes that the physical world has these properties in-itself, i.e. as inherent properties, independent of their actual observation.

The Copenhagen interpretation is first and foremost a semantic and epistemological reading of quantum mechanics that carries certain ontological implications. Bohr's view was, to phrase it in a modern

philosophical jargon, that the truth conditions of sentences ascribing a certain kinematic or dynamic value to an atomic object are dependent on the apparatus involved, in such a way that these truth conditions have to include reference to the experimental setup as well as the actual outcome of the experiment. Hence, those physicists who accuse this interpretation of operating with a mysterious collapse of the wave function during measurements do not understand a word of it. Bohr accepted the Born statistical interpretation because he believed that the ψ-function has only a symbolic meaning and does not represent anything real. It makes sense to talk about a collapse of the wave function only if, as Bohr put it, the ψ-function can be given a pictorial representation, something he strongly denied.

Indeed, Bohr, Heisenberg and many other physicists considered the Copenhagen Interpretation to be the only rational interpretation of the quantum world. They thought that it gave us the understanding of atomic phenomena that is in accordance with the conditions for any physical description and the possible objective knowledge of the world. Bohr believed that atoms are real, but it remains a much debated point in the recent literature what sort of reality he believed them to have, whether or not they are something beyond and different from what they are observed to be (Folse 1985 and 1994; and Faye 1991).

Bibliography

References to Work by Bohr

ATDN Bohr, N. (1934/1987), *Atomic Theory and the Description of Nature*, reprinted as *The Philosophical Writings of Niels Bohr, Vol. I*, Woodbridge: Ox Bow Press.

APHK Bohr, N. (1958/1987), *Essays 1932-1957 on Atomic Physics and Human Knowledge*, reprinted as *The Philosophical Writings of Niels Bohr, Vol. II*, Woodbridge: Ox Bow Press.

Essays Bohr, N. (1963/1987), *Essays 1958-1962 on Atomic Physics and Human Knowledge*, reprinted as *The Philosophical Writings of Niels Bohr, Vol. III*, Woodbridge: Ox Bow Press.

CC Bohr, N. (1998), *Causality and Complementarity*, supplementary papers edited by Jan Faye and Henry Folse as *The Philosophical Writings of Niels Bohr, Vol. IV*, Woodbridge: Ox Bow Press.

Other References

- Bunge, M. (1967), “The Turn of the Tide”, in Mario Bunge (ed.) *Quantum Theory and Reality*, New York: Springer, pp. 1-12.
- Chevalley, C. (1994), “Niels Bohr’s Words and the Atlantis of Kantianism”, in J. Faye and H. Folse (eds), *Niels Bohr and Contemporary Philosophy*, pp. 33-55.
- Faye, J. (1991), *Niels Bohr: His Heritage and Legacy. An Antirealist View of Quantum Mechanics*, Dordrecht: Kluwer Academic.
- Faye, J., and Folse, H. (eds) (1994), *Niels Bohr and Contemporary Philosophy*, Boston Studies in the Philosophy of Science, vol. 158, Dordrecht: Kluwer Academic.
- Folse, H. (1985), *The Philosophy of Niels Bohr. The Framework of Complementarity*,

Amsterdam: North Holland.

- Folse, H. (1986), “Niels Bohr, Complementarity, and Realism”, in A. Fine and P. Machamer (eds), *PSA 1986: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol. I,* East Lansing: PSA, pp. 96-104.
- Folse, H (1994), “Bohr’s Framework of Complementarity and the Realism Debate”, in Faye and Folse (1994), pp. 119-139.
- Held, C. (1994), “The Meaning of Complementarity”, *Studies in History and Philosophy of Science*, *25*, 871-893.
- Honner, J. (1987), *The Description of Nature: Niels Bohr and The Philosophy of Quantum Physics*, Oxford: Clarendon Press.
- Hooker, C. A. (1972), “The Nature of Quantum Mechanical Reality”, in R. G. Colodny (ed.), *Paradigms and Paradoxes*, Pittsburgh: University of Pittsburgh Press, pp. 67-305.
- Howard, D. (1994), “What Makes a Classical Concept Classical? Toward a Reconstruction of Niels Bohr’s Philosophy of Physics”, in Faye and Folse (1994), pp. 201-229.
- Kaiser, D. (1992), “More Roots of Complementarity: Kantian Aspects and Influences”, *Studies in History and Philosophy of Science*, *23*, 213-239.
- Murdoch, D. (1987), *Niels Bohr’s Philosophy of Physics*, Cambridge: Cambridge University Press.
- Petruccioli, S. (1993), *Atoms, Metaphors and Paradoxes*, Cambridge: Cambridge University Press.
- Plotnitsky, A. (1994), *Complementarity: Anti-Epistemology after Bohr and Derrida*, Durham: Duke University Press.
- Popper, K. R. (1967), “Quantum Mechanics Without ‘the Observer’”, in Mario Bunge (ed.) *Quantum Theory and Reality*, New York: Springer, pp. 1-12.

Other Internet Resources

- Entry on Niels Bohr (MacTutor History of Mathematics Archive, University of St. Andrews)

Related Entries

quantum mechanics | quantum theory: the Einstein-Podolsky-Rosen argument in | Uncertainty Principle

Jan Faye
faye@hum.ku.dk

Table of Contents

First published: May 3, 2002
Content last modified: May 3, 2002

The Uncertainty Principle

Quantum mechanics is generally regarded as the physical theory which is our best candidate yet for a universal and fundamental description of the physical world. The conceptual framework employed by this theory differs drastically from that of classical physics. Indeed, the transition from classical to quantum physics marks a genuine revolution in our understanding of the physical world.

One striking aspect of the difference between classical and quantum physics is that whereas classical mechanics presupposes that one can assign exact simultaneous values to the position and momentum of a particle, quantum mechanics denies this possibility. Instead, according to quantum mechanics, the more precisely the position of a particle is given, the less precisely one can say what its momentum is. This is (a simplistic and preliminary formulation of) the quantum mechanical uncertainty principle. This principle played an important role in many discussions on the philosophical implications of quantum mechanics and on the consistency of the interpretation endorsed by the founding fathers Heisenberg and Bohr, the so-called Copenhagen interpretation.

This, of course, should not suggest that the uncertainty principle is the only aspect in which classical and quantum physics differ conceptually. In particular the implications of quantum mechanics for notions such as (non)-locality, entanglement and identity play no less havoc with classical intuitions.

1. Introduction

The uncertainty principle is certainly one of the most famous and important aspects of quantum mechanics. Often, it has even been regarded as the most distinctive feature in which this theory differs from a classical conception of the physical world. Roughly speaking, the uncertainty principle states that one cannot assign exact simultaneous values to the position and momentum of a quantum mechanical system. Rather, we can only determine such quantities with some characteristic ‘uncertainties’, which cannot both become arbitrarily small at the same time. But what exactly is the meaning of this uncertainty principle? And indeed, is it really a principle of quantum mechanics? In particular, what does it mean that a quantity is determined only up to some uncertainty? These are the main questions we will explore in the following, focussing on the views of Heisenberg and Bohr.

In many expositions of the subject, the ‘uncertainty’ may refer sometimes to a lack of knowledge of a quantity by an observer, or to the experimental inaccuracy with which a quantity is measured, or to some ambiguity in the definition of a quantity, or to a statistical spread in some ensemble of similarly prepared systems. Corresponding to this confusing multitude of different meanings, there are many different names for these ‘uncertainties’. For example, apart from those already mentioned (inaccuracy, spread) one finds imprecision, indefiniteness, indeterminateness, indeterminacy, latitude, etc. As we shall see in the sequel, even Heisenberg and Bohr did not decide on a single terminology. Forestalling a discussion about which name is the most appropriate, we mention here that we use the name ‘uncertainty principle’ simply because it seems the most common one in the literature.

2. Heisenberg

2.1 Heisenberg's road to the uncertainty relations

Heisenberg introduced his now famous relations in an article of 1927, entitled "*Ueber den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik*". A (partial) translation of this title is: "On the *anschaulich* content of quantum theoretical kinematics and mechanics". Here, the term *anschaulich* is particularly notable. Apparently, it is one of those German words that defy an unambiguous translation into other languages. Heisenberg's title is translated as "*On the physical content* …" by Wheeler and Zurek (1983). His collected works (Heisenberg, 1984) translate it as "*On the perceptible content* …", while Cassidy's biography of Heisenberg (Cassidy, 1992), refers to the paper as "*On the perceptual content* …". Literally, the closest translation of the term *anschaulich* is ‘visualizable’. But, as in most languages, words that make reference to vision are not always intended literally. Seeing is widely used as a metaphor for understanding, especially for immediate understanding. Hence, *anschaulich* also means ‘intelligible’ or ‘intuitive’.[1]

Why was this issue of the *Anschaulichkeit* of quantum mechanics such a prominent concern to Heisenberg? This question has already been considered by a number of commentators (Jammer, 1977; Miller 1982; de Regt, 1997; Beller, 1999). For the answer, it turns out, we must go back a little in time. In 1925 Heisenberg had developed the first coherent mathematical formalism for quantum theory (Heisenberg, 1925). His leading idea was that only those quantities that are in principle observable should play a role in the theory, and that all attempts to form a picture of what goes on inside the atom should be avoided. In atomic physics the observational data were obtained from spectroscopy and associated with atomic transitions. Thus, Heisenberg was led to consider the 'transition quantities' as the basic ingredients of the theory. Max Born, later that year, realized that the transition quantities obeyed the rules of matrix calculus, a branch of mathematics that was not so well-known then as it is now. In a famous series of papers Heisenberg, Born and Jordan developed this idea into the matrix mechanics version of quantum theory.

Formally, matrix mechanics remains close to classical mechanics. The central idea is that all physical quantities must be represented by infinite self-adjoint matrices (later identified with operators on a Hilbert space). It is postulated that the matrices $\boldsymbol{q}$ and $\boldsymbol{p}$ representing the canonical position and momentum variables of a particle satisfy the so-called canonical commutation rule

$$\boldsymbol{qp} - \boldsymbol{pq} = i\hbar \qquad (1)$$

where $\hbar = h/2\pi$, h denotes Planck's constant, and boldface type is used to represent matrices. The new theory scored spectacular empirical success by encompassing nearly all spectroscopic data known at the time, especially after the concept of the electron spin was included in the theoretical framework.

It came as a great surprise, therefore, when one year later, Erwin Schrödinger presented an alternative theory, which became known as wave mechanics. Schrödinger assumed that an electron could be represented as an oscillating charge cloud, evolving continuously in space and time according to a wave equation. The discrete frequencies in the atomic spectra were not due to discontinuous transitions (quantum jumps) but to a resonance phenomenon. Further, Schrödinger argued that the two theories were equivalent.[2]

Even so, the two approaches differed greatly in interpretation and spirit. Whereas Heisenberg eschewed the use of visualizable pictures, and accepted discontinuous transitions as a primitive notion, Schrödinger claimed as an advantage of his theory that it was *anschaulich*. In Schrödinger's vocabulary, this meant that the theory represented the observational data by means of continuously evolving causal processes in space and time. He considered this condition of *Anschaulichkeit* to be an essential requirement on any acceptable physical theory. In fact, Schrödinger was not alone in appreciating this aspect of his theory. Many other leading physicists were attracted to wave mechanics for the same reason. For a while in 1926, before it emerged that wave mechanics has serious problems of its own, Schrödinger's approach seemed to gather more support in the physics community than matrix mechanics.

Understandably, Heisenberg was unhappy about this development. In a letter of 8 June 1926 to Pauli he

confessed that Schrödinger's approach disgusted him, and in particular: "What Schrödinger writes about the *Anschaulichkeit* of his theory, … I consider *Mist* (Pauli, 1979, p. 328)". Again, this last German term is translated differently by various commentators: as "junk" (Miller, 1982) "rubbish" (Beller 1999) "crap" (Cassidy, 1992), and perhaps more literally, as "bullshit" (de Regt, 1997). Nevertheless, in published writings, Heisenberg voiced a more balanced opinion. In a paper in *Die Naturwissenschaften* (1926) he summarized the peculiar situation which the simultaneous development of two competing theories had brought about. Although he argued that Schrödinger's interpretation was untenable, he admitted that matrix mechanics did not provide the *Anschaulichkeit* which made wave mechanics so attractive. He concluded: "to obtain a contradiction-free *anschaulich* interpretation, we still lack some essential feature in our image of the structure of matter". The purpose of his 1927 paper was to provide exactly this lacking feature.

2.2 Heisenberg's argument

Let us now look at the argument that led Heisenberg to his uncertainty relations. He started by redefining the notion of *Anschaulichkeit*. Whereas Schrödinger associated this term with the provision of a causal space-time picture of the phenomena, Heisenberg, by contrast, declared:

> We believe we have gained *anschaulich* understanding of a physical theory, if in all simple cases, we can grasp the experimental consequences qualitatively and see that the theory does not lead to any contradictions. Heisenberg, 1927, p. 172)

His goal was, of course, to show that, in this new sense of the word, matrix mechanics could lay the same claim to *Anschaulichkeit* as wave mechanics.

To do this, he adopted an operational assumption: terms like ‘the position of a particle’ have meaning only if one specifies a suitable experiment by which ‘the position of a particle’ can be measured. We will call this assumption the ‘measurement=meaning principle’. In general, there is no lack of such experiments, even in the domain of atomic physics. However, experiments are never completely accurate. We should be prepared to accept, therefore, that in general the meaning of these quantities is also determined only up to some characteristic inaccuracy.

As an example, he considered the measurement of the position of an electron by a microscope. The accuracy of such a measurement is limited by the wave length of the light illuminating the electron. Thus, it is possible, in principle, to make such a position measurement as accurate as one wishes, but only by using light of a very short wave length, e.g., γ-rays. But for γ-rays, the Compton effect cannot be ignored: the interaction of the electron and the illuminating light should then be considered as a collision of at least one photon with the electron. In such a collision, the electron suffers a recoil which disturbs its momentum. Moreover, the shorter the wave length, the larger is this change in momentum. Thus, at the moment when the position of the particle is accurately known, Heisenberg argued, its momentum cannot be accurately known:

> At the instant of time when the position is determined, that is, at the instant when the photon is scattered by the electron, the electron undergoes a discontinuous change in momentum. This change is the greater the smaller the wavelength of the light employed, i.e., the more exact the determination of the position. At the instant at which the position of the electron is known, its momentum therefore can be known only up to magnitudes which correspond to that discontinuous change; thus, the more precisely the position is determined, the less precisely the momentum is known, and conversely (Heisenberg, 1927, p. 174-5).

This is the first formulation of the uncertainty principle. In its present form it is an epistemological principle, since it limits what we can *know* about the electron. From "elementary formulae of the Compton effect" Heisenberg estimated the ‘imprecisions’ to be of the order

$$\delta p \delta q \sim h \qquad (2)$$

He continued: "In this circumstance we see the direct *anschaulich* content of the relation $\boldsymbol{qp} - \boldsymbol{pq} = i\hbar$".

He went on to consider other experiments, designed to measure other physical quantities and obtained analogous relations for time and energy:

$$\delta t \delta E \sim h \qquad (3)$$

and action J and angle w

$$\delta w \delta J \sim h \qquad (4)$$

which he saw as corresponding to the "well-known" relations

$$\boldsymbol{tE} - \boldsymbol{Et} = i\hbar \quad \text{or} \quad \boldsymbol{wJ} - \boldsymbol{Jw} = i\hbar \qquad (5)$$

However, we shall see in Section 2.5 that these generalisations did not turn out as straightforward as Heisenberg suggested.

He summarized his findings in a general conclusion: all concepts used in classical mechanics are also well-defined in the realm of atomic processes. But, as a pure fact of experience ("*rein erfahrungsgemäß*"), experiments that serve to provide such a definition for one quantity are subject to particular indeterminacies, obeying relations (2)-(4) which prohibit them from providing a simultaneous definition of two canonically conjugate quantities. Note that in this formulation the emphasis has slightly shifted: he now speaks of a limit on the definition of concepts, i.e. not merely on what we can *know*, but what we can meaningfully *say* about a particle. Of course, this stronger formulation follows by

application of the above measurement=meaning principle: if there are, as Heisenberg claims, no experiments that allow a simultaneous precise measurement of two conjugate quantities, then these quantities are also not simultaneously well-defined.

Heisenberg's paper has an interesting "Addition in proof" mentioning critical remarks by Bohr, who saw the paper only after it had been sent to the publisher. Among other things, Bohr pointed out that in the microscope experiment it is not the change of the momentum of the electron that is important, but rather the circumstance that this change cannot be precisely determined in the *same* experiment. An improved version of the argument, which responds to this objection, is given in Heisenberg's Chicago lectures of 1930.

Here (Heisenberg, 1930, p. 16), it is assumed that the electron is illuminated by light of wavelength λ and that the scattered light enters a microscope with aperture angle ε. According to the laws of classical optics, the accuracy of the microscope depends on both the wave length and the aperture angle; Abbe's criterium for its ‘resolving power’, i.e. the size of the smallest discernable details, gives

$$\delta q \sim \lambda/\sin\varepsilon \qquad (6)$$

On the other hand, the direction of a scattered photon, when it enters the microscope, is unknown within the angle ε, rendering the momentum change of the electron uncertain by an amount

$$\delta p \sim h \sin\varepsilon/\lambda \qquad (7)$$

leading again to the result (2).

Let us now analyse Heisenberg's argument in more detail. First note that, even in this improved version, Heisenberg's argument is incomplete. According to Heisenberg's ‘measurement=meaning principle’, one must also specify, in the given context, what the meaning is of the phrase ‘momentum of the electron’, in order to make sense of the claim that this momentum is changed by the position measurement. A solution to this problem can again be found in the Chicago lectures (Heisenberg, 1930, p. 15). Here, he assumes that initially the momentum of the electron is precisely known, e.g. it has been measured in a previous experiment with an inaccuracy δp_i, which may be arbitrarily small. Then, its position is measured with inaccuracy δq, and after this, its final momentum is measured with an inaccuracy δp_f. All three measurements can be performed with arbitrary precision. Thus, the three quantities δp_i, δq, and δp_f can be made as small as one wishes. If we assume further that the initial momentum has not changed until the position measurement, we can speak of a definite momentum until the time of the position measurement. Moreover we can give operational meaning to the idea that the momentum is changed during the position measurement: the outcome of the second momentum measurement (say p_f) will generally differ from the initial value p_i. In fact, one can also show that this change is discontinuous, by varying the time between the three measurements.

Let us now try to see, adopting this more elaborate set-up, if we can complete Heisenberg's argument. We have now been able to give empirical meaning to the ‘change of momentum’ of the electron, $p_f - p_i$. Heisenberg's argument claims that the order of magnitude of this change is at least inversely proportional to the inaccuracy of the position measurement:

$$|p_f - p_i| \delta q \sim h \quad (8)$$

However, can we now draw the conclusion that the momentum is only imprecisely defined? Certainly not. Before the position measurement, its value was p_i, after the measurement it is p_f. One might, perhaps, claim that the value at the very instant of the position measurement is not yet defined, but we could simply settle this assignment by a convention (e.g., we might assign the mean value $(p_i + p_f)/2$ to this instant). But then, the momentum is precisely determined at all instants, and Heisenberg's formulation of the uncertainty principle no longer follows. The above attempt of completing Heisenberg's argument thus overshoots its mark.

A solution to this problem can again be found in the Chicago Lectures. Heisenberg admits that position and momentum can be known exactly. He writes:

> If the velocity of the electron is at first known, and the position then exactly measured, the position of the electron for times previous to the position measurement may be calculated. For these past times, $\delta p \delta q$ is smaller than the usual bound. (Heisenberg 1930, p. 15)

Indeed, Heisenberg says: "the uncertainty relation does not hold for the past".

Apparently, when Heisenberg refers to the uncertainty or imprecision of a quantity, he means that the value of this quantity cannot be given *beforehand*. In the sequence of measurements we have considered above, the uncertainty in the momentum after the measurement of position has occurred, refers to the idea that the value in momentum is not fixed *before* the final momentum measurement takes place. Once this measurement is performed, and reveals a value p_f, the uncertainty relation no longer holds; these values then belong to the past. Clearly, then, Heisenberg is concerned with *unpredictability*: the point is not that the momentum of a particle changes, due to a position measurement, but rather that it changes by an unpredictable amount. It is, however always possible to measure, and hence define, the size of this change in a subsequent measurement of the final momentum with arbitrary precision.

Although Heisenberg admits that we can consistently attribute values of momentum and position to an electron in the past, he sees little merit in such talk. He points out that these values can never be used as initial conditions in a prediction about the future behavior of the electron, or subjected to experimental verification. Whether or not we grant them physical reality is, as he puts it, a matter of personal taste. Heisenberg's own taste is, of course, to deny their physical reality. For example, he writes, "I believe that one can formulate the emergence of the classical ‘path’ of a particle pregnantly as follows: *the ‘path’ comes into being only because we observe it*" (Heisenberg, 1927, p. 185). Apparently, in his view, a

measurement does not only serve to give meaning to a quantity, it *creates* a particular value for this quantity. This may be called the ‘measurement=creation’ principle. It is an ontological principle, for it states what is physically real.

This then leads to the following picture. First we measure the momentum of the electron very accurately. By ‘measurement= meaning’, this entails that the term "the momentum of the particle" is now well-defined. Moreover, by the ‘measurement=creation’ principle, we may say that this momentum is physically real. Next, the position is measured with inaccuracy δq. At this instant, the position of the particle becomes well-defined and, again, one can regard this as a physically real attribute of the particle. However, the momentum has now changed by an amount, which is unpredictable by an order of magnitude $| p_f - p_i | \sim h/\delta q$. The meaning and validity of this claim can be verified by a subsequent momentum measurement.

The question is then what status we shall assign to the momentum of the electron just before its final measurement. Is it real? According to Heisenberg it is not. Before the final measurement, the best we can attribute to the electron is some unsharp, or fuzzy momentum. These terms are meant here in an ontological sense, characterizing a real attribute of the electron.

2.3 The interpretation of Heisenberg's relation

The relations Heisenberg had proposed were soon considered to be a cornerstone of the Copenhagen interpretation of quantum mechanics. Just a few months later, Kennard (1927) already called them the "essential core" of the new theory. Taken together with Heisenberg's contention that they provided the intuitive content of the theory and their prominent role in later discussions on the Copenhagen interpretation, a dominant view emerged in which they were regarded as a fundamental principle of the theory.

The interpretation of these relations has often been debated. Do Heisenberg's relations express restrictions on the experiments we can perform on quantum systems, and, therefore, restrictions on the information we can gather about such systems; or do they express restrictions on the meaning of the concepts we use to describe quantum systems? Or else, are they restrictions of an ontological nature, i.e., do they assert that a quantum system simply does not possess a definite value for its position and momentum at the same time? The difference between these interpretations is partly reflected in the various names by which the relations are known, e.g. as ‘inaccuracy relations’, or: ‘uncertainty’, ‘indeterminacy’ or ‘unsharpness relations’, etc. The debate between these different views has been addressed by many authors, but it has never been settled completely. Let it suffice here to make only two general observations.

First, it is clear that in Heisenberg's own view, all the above questions stand or fall together. Indeed, we have seen that he adopted an operational "measurement=meaning" principle according to which the meaningfulness of a physical quantity was equivalent to the existence of an experiment purporting to measure that quantity. Similarly, his "measurement=creation" principle allowed him to attribute physical reality to such quantities. Hence, Heisenberg's discussions moved rather freely and quickly from talk

about experimental inaccuracies to epistemological or ontological issues and back again.

However, ontological questions seemed to be of somewhat less interest to him. For example, there is a passage (Heisenberg, 1927, p. 197), where he discusses the idea that, behind our observational data, there might still exist a hidden reality in which quantum systems have definite values for position and momentum, unaffected by the uncertainty relations. He emphatically dismisses this conception as an unfruitful and meaningless speculation, because, as he says, the aim of physics is only to describe observable data. Similarly in the Chicago Lectures (Heisenberg 1930, p. 11) he warns against the fact that the human language permits the utterance of statements which have no empirical content at all, but nevertheless produce a picture in our imagination. He notes, "One should be especially careful in using the words 'reality', 'actually', etc., since these words very often lead to statements of the type just mentioned." So, Heisenberg also endorsed an interpretation of his relations as rejecting a reality in which particles have simultaneous definite values for position and momentum.

The second observation is that although for Heisenberg experimental, informational, epistemological and ontological formulations of his relations were, so to say, just different sides of the same coin, this does not hold for those who do not share his operational principles or his view on the task of physics. Alternative points of view, in which e.g. the ontological reading of the uncertainty relations is denied, are therefore still viable. The statement, often found in the literature of the thirties, that Heisenberg had *proved* the impossibility of associating a definite position and momentum to a particle is certainly wrong. But the precise meaning one can coherently attach to Heisenberg's relations depends rather heavily on the interpretation one favors for quantum mechanics as a whole. And in view of the fact that no agreement has been reached on this latter issue, one cannot expect agreement on the meaning of the uncertainty relations either.

2.4 Uncertainty relations or uncertainty principle?

Let us now move to another question about Heisenberg's relations: do they express a *principle* of quantum theory? Probably the first influential author to call these relations a 'principle' was Eddington, who, in his Gifford Lectures of 1928 referred to them as the 'Principle of Indeterminacy'. In the English literature the name uncertainty principle became most common. It is used both by Condon and Robertson in 1929, and also in the English version of Heisenberg's Chicago Lectures (Heisenberg, 1930), although, remarkably, nowhere in the original German version of the same book (see also Cassidy, 1998). Indeed, Heisenberg never seems to have endorsed the name 'principle' for his relations. His favourite terminology was 'inaccuracy relations' (*Ungenauigkeitsrelationen*) or 'indeterminacy relations' (*Unbestimmtheitsrelationen*). We know only one passage, in Heisenberg's own Gifford lectures, delivered in 1955-56 (Heisenberg, 1958, p. 43), where he mentioned that his relations "are usually called relations of uncertainty or principle of indeterminacy". But this can well be read as his yielding to common practice rather than his own preference.

But does the relation (2) qualify as a principle of quantum mechanics? Several authors, foremost Karl Popper (1967), have contested this view. Popper argued that the uncertainty relations cannot be granted

the status of a principle, on the grounds that they are derivable from the theory, whereas one cannot obtain the theory from the uncertainty relations. (The argument being that one can never derive any equation, say, the Schrödinger equation, or the commutation relation (1), from an inequality.)

Popper's argument is, of course, correct but we think it misses the point. There are many statements in physical theories which are called principles even though they are in fact derivable from other statements in the theory in question. A more appropriate departing point for this issue is not the question of logical priority but rather Einstein's distinction between ‘constructive theories’ and ‘principle theories’.

Einstein proposed this famous classification in (Einstein, 1919). Constructive theories are theories which postulate the existence of simple entities behind the phenomena. They endeavour to reconstruct the phenomena by framing hypotheses about these entities. Principle theories, on the other hand, start from empirical principles, i.e. general statements of empirical regularities, employing no or only a bare minimum of theoretical terms. The purpose is to build up the theory from such principles. That is, one aims to show how these empirical principles provide sufficient conditions for the introduction of further theoretical concepts and structure.

The prime example of a theory of principle is thermodynamics. Here the role of the empirical principles is played by the statements of the impossibility of various kinds of perpetual motion machines. These are regarded as expressions of brute empirical fact, providing the appropriate conditions for the introduction of the concepts of energy and entropy and their properties. (There is a lot to be said about the tenability of this view, but that is not the topic of this entry.)

Now obviously, once the formal thermodynamic theory is built, one can also *derive* the impossibility of the various kinds of perpetual motion. (They would violate the laws of energy conservation and entropy increase.) But this derivation should not misguide one into thinking that they were no principles of the theory after all. The point is just that empirical principles are statements that do not rely on the theoretical concepts (in this case entropy and energy) for their meaning. They are interpretable independently of these concepts and, further, their validity on the empirical level still provides the physical content of the theory.

A similar example is provided by special relativity, another theory of principle, which Einstein deliberately designed after the ideal of thermodynamics. Here, the empirical principles are the light postulate and the relativity principle. Again, once we have built up the modern theoretical formalism of the theory (the Minkowski space-time) it is straightforward to prove the validity of these principles. But again this does not count as an argument for claiming that they were no principles after all. So the question whether the term ‘principle’ is justified for Heisenberg's relations, should, in our view, be understood as the question whether they are conceived of as empirical principles.

One can easily show that this idea was never far from Heisenberg's intentions. We have already seen that Heisenberg presented them as the result of a "pure fact of experience". A few months after his 1927 paper, he wrote a popular paper with the title "*Ueber die Grundprincipien der Quantenmechanik*" ("On the fundamental principles of quantum mechanics") where he made the point even more clearly. Here

Heisenberg described his recent break-through in the interpretation of the theory as follows: "It seems to be a general law of nature that we cannot determine position and velocity simultaneously with arbitrary accuracy". Now actually, and in spite of its title, the paper does not identify or discuss any ‘fundamental principle’ of quantum mechanics. So, it must have seemed obvious to his readers that he intended to claim that the uncertainty relation was a fundamental principle, forced upon us as an empirical law of nature, rather than a result derived from the formalism of this theory.

This reading of Heisenberg's intentions is corroborated by the fact that, even in his 1927 paper, applications of his relation frequently present the conclusion as a matter of principle. For example, he says "In a stationary state of an atom its phase is *in principle* indeterminate" (Heisenberg, 1927, p. 177, [emphasis added]). Similarly, in a paper of 1928, he described the content of his relations as: "It has turned out that it is *in principle* impossible to know, to measure the position and velocity of a piece of matter with arbitrary accuracy. (Heisenberg, 1984, p. 26, [emphasis added])"

So, although Heisenberg did not originate the tradition of calling his relations a principle, it is not implausible to attribute the view to him that the uncertainty relations represent an empirical principle that could serve as a foundation of quantum mechanics. In fact, his 1927 paper expressed this desire explicitly: "Surely, one would like to be able to deduce the quantitative laws of quantum mechanics directly from their *anschaulich* foundations, that is, essentially, relation [(2)]" (*ibid*, p. 196). This is not to say that Heisenberg was successful in reaching this goal, or that he did not express other opinions on other occasions.

Let us conclude this section with three remarks. First, if the uncertainty relation is to serve as an empirical principle, one might well ask what its direct empirical support is. In Heisenberg's analysis, no such support is mentioned. His arguments concerned thought experiments in which the validity of the theory, at least at a rudimentary level, is implicitly taken for granted. Jammer (1974, p. 82) conducted a literature search for high precision experiments that could seriously test the uncertainty relations and concluded they were still scarce in 1974. Real experimental support for the uncertainty relations in experiments in which the inaccuracies are close to the quantum limit have come about only more recently. (See Kaiser, Werner and George 1983, Uffink 1985, Nairz, Andt, and Zeilinger, 2001.)

A second point is the question whether the theoretical structure or the quantitative laws of quantum theory can indeed be derived on the basis of the uncertainty principle, as Heisenberg wished. Serious attempts to build up quantum theory as a full-fledged Theory of Principle on the basis of the uncertainty principle have never been carried out. Indeed, the most Heisenberg could and did claim in this respect was that the uncertainty relations created "room" (Heisenberg 1927, p. 180) or "freedom" (Heisenberg, 1931, p. 43) for the introduction of some non-classical mode of description of experimental data, not that they uniquely lead to the formalism of quantum mechanics. A serious proposal to construe quantum mechanics as a theory of principle was provided only recently by Bub (2000). But, remarkably, this proposal does not use the uncertainty relation as one of its fundamental principles.

Third, it is remarkable that in his later years Heisenberg put a somewhat different gloss on his relations.

In his autobiography *Der Teil und das Ganze* of 1969 he described how he had found his relations inspired by a remark by Einstein that "it is the theory which decides what one can observe" -- thus giving precedence to theory above experience, rather than the other way around. Some years later he even admitted that his famous discussions of thought experiments were actually trivial since "… if the process of observation itself is subject to the laws of quantum theory, it must be possible to represent its result in the mathematical scheme of this theory" (Heisenberg, 1975, p. 6).

2.5 Mathematical elaboration

When Heisenberg introduced his relation, his argument was based only on qualitative examples. He did not provide a general, exact derivation of his relations.[3] Indeed, he did not even give a definition of the uncertainties δq, etc., occurring in these relations. Of course, this was consistent with the announced goal of that paper, i.e. to provide some qualitative understanding of quantum mechanics for simple experiments.

The first mathematically exact formulation of the uncertainty relations is due to Kennard. He proved in 1927 the theorem that for all normalized state vectors $|\psi\rangle$ the following inequality holds:

$$\Delta_\psi \boldsymbol{p} \Delta_\psi \boldsymbol{q} \geq \frac{\hbar}{2} \tag{9}$$

Here, $\Delta_\psi \boldsymbol{p}$ and $\Delta_\psi \boldsymbol{q}$ are standard deviations of position and momentum in the state vector $|\psi\rangle$, i.e.,

$$(\Delta_\psi \boldsymbol{p})^2 = \langle \boldsymbol{p}^2 \rangle_\psi - \langle \boldsymbol{p} \rangle_\psi^2, \qquad (\Delta_\psi \boldsymbol{q})^2 = \langle \boldsymbol{q}^2 \rangle_\psi - \langle \boldsymbol{q} \rangle_\psi^2 \tag{10}$$

where $\langle \cdot \rangle_\psi = \langle \psi | \cdot | \psi \rangle$ denotes the expectation value in state $|\psi\rangle$. This inequality (10) was generalized in 1929 by Robertson who proved the result that for all observables (self-adjoint operators) $\boldsymbol{A}$ and $\boldsymbol{B}$

$$\Delta_\psi \boldsymbol{A} \Delta_\psi \boldsymbol{B} \geq \frac{1}{2} |\langle [\boldsymbol{A}, \boldsymbol{B}] \rangle_\psi| \tag{11}$$

where $[\boldsymbol{A}, \boldsymbol{B}] := \boldsymbol{AB} - \boldsymbol{BA}$ denotes the commutator. This relation was in turn strengthened by Schrödinger (1930), who obtained:

$$(\Delta_\psi \boldsymbol{A})^2 (\Delta_\psi \boldsymbol{B})^2 \geq \frac{1}{4} |\langle [\boldsymbol{A}, \boldsymbol{B}] \rangle_\psi|^2 + \frac{1}{4} |\langle \{ \boldsymbol{A} - \langle \boldsymbol{A} \rangle_\psi, \boldsymbol{B} - \langle \boldsymbol{B} \rangle_\psi \}] \rangle_\psi|^2 \tag{12}$$

where $\{A, B\} := (AB + BA)$ denotes the anti-commutator.

Since the above inequalities have the virtue of being exact and general, in contrast to Heisenberg's original semi-quantitative formulation, it is tempting to regard them as the exact counterpart of Heisenberg's relations (2)-(4). Indeed, such was Heisenberg's own view. In his Chicago Lectures (Heisenberg 1930, pp. 15-19), he presented Kennard's derivation of relation (9) and claimed that "this proof does not differ at all in mathematical content" from the semi-quantitative argument he had presented earlier, the only difference being that now "the proof is carried through exactly".

But it may be useful to point out that both in status and intended role there is a subtle difference between Kennard's inequality and Heisenberg's previous formulation (2). The inequalities discussed here are not statements of empirical fact, but theorems of the formalism. As such, they presuppose the validity of this formalism, and in particular the commutation relation (1), rather than elucidating its intuitive content or to create ‘room’ or ‘freedom’ for the validity of this relation. At best, one should see the above inequalities as showing that the formalism is consistent with Heisenberg's empirical principle.

This situation is similar to that arising in other theories of principle where, as noted in Section 2.4, one often finds that, next to an empirical principle, the formalism also provides a corresponding theorem. And similarly, this situation should not, by itself, cast doubt on the question whether Heisenberg's relation can be regarded as a principle of quantum mechanics.

There is a second notable difference between (2) and (9). Heisenberg did not give a general definition for the ‘uncertainties’ δp and δq. The most definite remark he made about them was that they could be taken as "something like the mean error". In the discussions of thought experiments, he and Bohr would always quantify uncertainties on a case-to-case basis by choosing some parameters which happened to be relevant to the experiment at hand. By contrast, the inequalities (9)-(12) employ a single specific expression as a measure for ‘uncertainty’: the standard deviation. This choice is not unnatural, given that this expression is well-known and widely used in error theory and the description of statistical fluctuations.

However, there was very little or no discussion of whether this choice was appropriate for the general formulation of the uncertainty relations. A standard deviation reflects the spread or expected fluctuations in a series of measurements of an observable on a given state. It is not at all easy to connect this idea with the concept of the ‘inaccuracy’ of a measurement, such as the resolving power of a microscope. In fact, even though Heisenberg had taken Kennard's inequality as the precise formulation of the uncertainty relation, he and Bohr never relied on standard deviations in their many discussions of thought experiments, and indeed, it has been shown (Uffink and Hilgevoord, 1985; Hilgevoord and Uffink, 1988) that these discussions cannot be framed in terms of standard deviation.

Another problem with the above elaboration is that it soon turned out that there are no analogous inequalities for time and energy, nor for action and angle. Jordan (1927) pointed out in 1927 that the ‘well-known’ relations (5) are actually false in the case of action and angle, and Pauli (1933) showed that there

is no operator canonically conjugate to the Hamiltonian, if the latter is bounded from below. This menas that for many systems a time operator does not exist. These observations have led to a quite extensive literature on time-energy and phase-action uncertainty relations, proposing many different attempts to overcome these obstacles.

3. Bohr

In spite of the fact that Heisenberg's and Bohr's views on quantum mechanics are often lumped together as (part of) ‘the Copenhagen interpretation’, there is considerable difference between their views on the uncertainty relations.

3.1 From wave-particle duality to complementarity

Long before the development of modern quantum mechanics, Bohr had been particularly concerned with the problem of particle-wave duality, i.e. the problem that experimental evidence on the behaviour of both light and matter seemed to demand a wave picture in some cases, and a particle picture in others. Yet these pictures are mutually exclusive. Whereas a particle is always localized, the very definition of the notions of wavelength and frequency requires an extension in space and in time. Moreover, the classical particle picture is incompatible with the characteristic phenomenon of interference.

His long struggle with wave-particle duality had prepared him for a radical step when the dispute between matrix and wave mechanics broke out in 1926-27. For the main contestants, Heisenberg and Schrödinger, the issue at stake was which view could claim to provide a single coherent and universal framework for the description of the observational data. The choice was, essentially between a description in terms of continuously evolving waves, or else one of particles undergoing discontinuous quantum jumps. By contrast, Bohr insisted that elements from both views were equally valid and equally needed for an exhaustive description of the data. His way out of the contradiction was to renounce the idea that the pictures refer, in a literal one-to-one correspondence, to physical reality. Instead, the applicability of these pictures was to become dependent on the experimental context. This is the gist of the viewpoint he called ‘complementarity’.

Bohr first conceived the general outline of his complementarity argument in early 1927, during a skiing holiday in Norway, at the same time when Heisenberg wrote his uncertainty paper. When he returned to Copenhagen and found Heisenberg's manuscript, they got into an intense discussion. On the one hand, Bohr was quite enthusiastic about Heisenberg's ideas which seemed to fit wonderfully with his own thinking. Indeed, in his subsequent work, Bohr always presented the uncertainty relations as the symbolic expression of his complementarity viewpoint. On the other hand, he criticized Heisenberg severely for his suggestion that these relations were due to discontinuous changes occurring during a measurement process. Rather, Bohr argued, their proper derivation should start from the indispensability of both particle and wave concepts. He pointed out that the uncertainties in the experiment did not exclusively arise from the discontinuities but also from the fact that in the experiment we need to take into account both the particle theory and the wave theory. It is not so much the unknown disturbance which renders

the momentum of the electron uncertain but rather the fact that the position and the momentum of the electron cannot be simultaneously defined in this experiment. (See the "Addition in Proof" to Heisenberg's paper.)

We shall not go too deeply into the matter of Bohr's interpretation of quantum mechanics since we are mostly interested in Bohr's view on the uncertainty principle. For a more detailed discussion of Bohr's philosophy of quantum physics we refer to Scheibe (1973), Folse (1985), Honner (1987) and Murdoch (1987). It may be useful, however, to sketch some of the main points. Central in Bohr's considerations is the *language* which we use in physics. No matter how abstract and subtle the concepts of modern physics may be, they are essentially an extension of our ordinary language and a means to communicate the results of our experiments. These results, obtained under well-defined experimental circumstances, are what Bohr calls the "phenomena". A phenomenon is "the comprehension of the effects observed under given experimental conditions" (Bohr 1939, p. 24), it is the resultant of a physical object, a measuring apparatus and the interaction between them in a concrete experimental situation. The essential difference between classical and quantum physics is that in quantum physics the interaction between the object and the apparatus cannot be made arbitrarily small; the interaction must at least comprise one quantum. This is expressed by Bohr's quantum postulate:

> [… the] essence [of the formulation of the quantum theory] may be expressed in the so-called quantum postulate, which attributes to any atomic process an essential discontinuity or rather individuality, completely foreign to classical theories and symbolized by Planck's quantum of action. (Bohr, 1928, p. 580)

A phenomenon, therefore, is an indivisible whole and the result of a measurement cannot be considered as an autonomous manifestation of the object itself independently of the measurement context. The quantum postulate forces upon us a new way of describing physical phenomena:

> In this situation, we are faced with the necessity of a radical revision of the foundation for the description and explanation of physical phenomena. Here, it must above all be recognized that, however far quantum effects transcend the scope of classical physical analysis, the account of the experimental arrangement and the record of the observations must always be expressed in common language supplemented with the terminology of classical physics. (Bohr, 1948, p. 313)

This is what Scheibe (1973) has called the "buffer postulate" because it prevents the quantum from penetrating into the classical description: A phenomenon must always be described in classical terms; Planck's constant does not occur in this description.

Together, the two postulates induce the following reasoning. In every phenomenon the interaction between the object and the apparatus comprises at least one quantum. But the description of the phenomenon must use classical notions in which the quantum of action does not occur. Hence, the interaction cannot be analysed in this description. On the other hand, the classical character of the

description allows to speak in terms of the object itself. Instead of saying: ‘the interaction between a particle and a photographic plate has resulted in a black spot in a certain place on the plate’, we are allowed to forgo mentioning the apparatus and say: ‘the particle has been found in this place’. The experimental context, rather than changing or disturbing pre-existing properties of the object, defines what can meaningfully be said about the object.

Because the interaction between object and apparatus is left out in our description of the phenomenon, we do not get the whole picture. Yet, any attempt to extend our description by performing the measurement of a different observable quantity of the object, or indeed, on the measurement apparatus, produces a new phenomenon and we are again confronted with the same situation. Because of the unanalyzable interaction in both measurements, the two descriptions cannot, generally, be united into a single picture. They are what Bohr calls complementary descriptions:

> [the quantum of action]...forces us to adopt a new mode of description designated as complementary in the sense that any given application of classical concepts precludes the simultaneous use of other classical concepts which in a different connection are equally necessary for the elucidation of the phenomena. (Bohr, 1929, p. 10)

The most important example of complementary descriptions is provided by the measurements of the position and momentum of an object. If one wants to measure the position of the object relative to a given spatial frame of reference, the measuring instrument must be rigidly fixed to the bodies which define the frame of reference. But this implies the impossibility of investigating the exchange of momentum between the object and the instrument and we are cut off from obtaining any information about the momentum of the object. If, on the other hand, one wants to measure the momentum of an object the measuring instrument must be able to move relative to the spatial reference frame. Bohr here assumes that a momentum measurement involves the registration of the recoil of some movable part of the instrument and the use of the law of momentum conservation. The looseness of the part of the instrument with which the object interacts entails that the instrument cannot serve to accurately determine the position of the object. Since a measuring instrument cannot be rigidly fixed to the spatial reference frame and, at the same time, be movable relative to it, the experiments which serve to precisely determine the position and the momentum of an object are mutually exclusive. Of course, in itself, this is not at all typical for quantum mechanics. But, because the interaction between object and instrument during the measurement can neither be neglected nor determined the two measurements cannot be combined. This means that in the description of the object one must choose between the assignment of a precise position or of a precise momentum.

Similar considerations hold with respect to the measurement of time and energy. Just as the spatial coordinate system must be fixed by means of solid bodies so must the time coordinate be fixed by means of unperturbable, synchronised clocks. But it is precisely this requirement which prevents one from taking into account of the exchange of energy with the instrument if this is to serve its purpose. Conversely, any conclusion about the object based on the conservation of energy prevents following its development in time.

The conclusion is that in quantum mechanics we are confronted with a complementarity between two descriptions which are united in the classical mode of description: the space-time description (or coordination) of a process and the description based on the applicability of the dynamical conservation laws. The quantum forces us to give up the classical mode of description (also called the ‘causal’ mode of description by Bohr[4]): it is impossible to form a classical picture of what is going on when radiation interacts with matter as, e.g., in the Compton effect.

> Any arrangement suited to study the exchange of energy and momentum between the electron and the photon must involve a latitude in the space-time description sufficient for the definition of wave-number and frequency which enter in the relation [$E = h\nu$ and $p = h\sigma$]. Conversely, any attempt of locating the collision between the photon and the electron more accurately would, on account of the unavoidable interaction with the fixed scales and clocks defining the space-time reference frame, exclude all closer account as regards the balance of momentum and energy. (Bohr, 1949, p. 210)

A causal description of the process cannot be attained; we have to content ourselves with complementary descriptions. "The viewpoint of complementarity may be regarded", according to Bohr, "as a rational generalization of the very ideal of causality".

In addition to complementary descriptions Bohr also talks about complementary phenomena and complementary quantities. Position and momentum, as well as time and energy, are complementary quantities.[5]

We have seen that Bohr's approach to quantum theory puts heavy emphasis on the language used to communicate experimental observations, which, in his opinion, must always remain classical. By comparison, he seemed to put little value on arguments starting from the mathematical formalism of quantum theory. This informal approach is typical of all of Bohr's discussions on the meaning of quantum mechanics. One might say that for Bohr the conceptual clarification of the situation has primary importance while the formalism is only a symbolic representation of this situation.

This is remarkable since, finally, it is the formalism which needs to be interpreted. This neglect of the formalism, certainly, is one of the reasons why it is so difficult to get a clear understanding of Bohr's interpretation of quantum mechanics and why it has aroused so much controversy. We close this section by citing from an article of 1948 to show how Bohr conceived the role of the formalism of quantum mechanics:

> The entire formalism is to be considered as a tool for deriving predictions, of definite or statistical character, as regards information obtainable under experimental conditions described in classical terms and specified by means of parameters entering into the algebraic or differential equations of which the matrices or the wave-functions, respectively, are solutions. These symbols themselves, as is indicated already by the use of imaginary numbers, are not susceptible to pictorial interpretation; and even derived real

> functions like densities and currents are only to be regarded as expressing the probabilities for the occurrence of individual events observable under well-defined experimental conditions. (Bohr, 1948, p. 314)

3.2 Bohr's view on the uncertainty relations

In his Como lecture, published in 1928, Bohr gave his own version of a derivation of the uncertainty relations between position and momentum and between time and energy. He started from the relations

$$E = h\nu \text{ and } p = h/\lambda \tag{13}$$

which connect the notions of energy E and momentum p from the particle picture with those of frequency ν and wavelength λ from the wave picture. He noticed that a wave packet of limited extension in space and time can only be built up by the superposition of a number of elementary waves with a large range of wave numbers and frequencies. Denoting the spatial and temporal extensions of the wave packet by Δx and Δt, and the extensions in the wave number $\sigma := 1/\lambda$ and frequency by $\Delta\sigma$ and $\Delta\nu$, it follows from Fourier analysis that in the most favorable case $\Delta x\Delta\sigma \approx \Delta t\Delta\nu \approx 1$, and, using (13), one obtains the relations

$$\Delta t\Delta E \approx \Delta x\Delta p \approx h \tag{14}$$

Note that Δx, $\Delta\sigma$, etc., are not standard deviations but unspecified measures of the size of a wave packet. (The original text has equality signs instead of approximate equality signs, but, since Bohr does not define the spreads exactly the use of approximate equality signs seems more in line with his intentions. Moreover, Bohr himself used approximate equality signs in later presentations.) These equations determine, according to Bohr: "the highest possible accuracy in the definition of the energy and momentum of the individuals associated with the wave field" (Bohr 1928, p. 571). He noted, "This circumstance may be regarded as a simple symbolic expression of the complementary nature of the space-time description and the claims of causality" (*ibid*).[6] We note a few points about Bohr's view on the uncertainty relations. First of all, Bohr does not refer to *discontinuous changes* in the relevant quantities during the measurement process. Rather, he emphasizes the possibility of *defining* these quantities. This view is markedly different from Heisenberg's. A draft version of the Como lecture is even more explicit on the difference between Bohr and Heisenberg:

> These reciprocal uncertainty relations were given in a recent paper of Heisenberg as the expression of the statistical element which, due to the feature of discontinuity implied in the quantum postulate, characterizes any interpretation of observations by means of classical concepts. It must be remembered, however, that the uncertainty in question is not simply a consequence of a discontinuous change of energy and momentum say during an interaction between radiation and material particles employed in measuring the space-time coordinates of the individuals. According to the above considerations the question is rather

> that of the impossibility of defining rigourously such a change when the space-time coordination of the individuals is also considered. (Bohr, 1985 p. 93)

Indeed, Bohr not only rejected Heisenberg's argument that these relations are due to discontinuous disturbances implied by the act of measuring, but also his view that the measurement process *creates* a definite result:

> The unaccustomed features of the situation with which we are confronted in quantum theory necessitate the greatest caution as regard all questions of terminology. Speaking, as it is often done of disturbing a phenomenon by observation, or even of creating physical attributes to objects by measuring processes is liable to be confusing, since all such sentences imply a departure from conventions of basic language which even though it can be practical for the sake of brevity, can never be unambiguous. (Bohr, 1939, p. 24)

Nor did he approve of an epistemological formulation or one in terms of experimental inaccuracies:

> […] a sentence like "we cannot know both the momentum and the position of an atomic object" raises at once questions as to the physical reality of two such attributes of the object, which can be answered only by referring to the mutual exclusive conditions for an unambiguous use of space-time concepts, on the one hand, and dynamical conservation laws on the other hand. (Bohr, 1948, p. 315; also Bohr 1949, p. 211)

> It would in particular not be out of place in this connection to warn against a misunderstanding likely to arise when one tries to express the content of Heisenberg's well-known indeterminacy relation by such a statement as 'the position and momentum of a particle cannot simultaneously be measured with arbitrary accuracy'. According to such a formulation it would appear as though we had to do with some arbitrary renunciation of the measurement of either the one or the other of two well-defined attributes of the object, which would not preclude the possibility of a future theory taking both attributes into account on the lines of the classical physics. (Bohr 1937, p. 292)

Instead, Bohr always stressed that in his point of view the uncertainty relations are foremost an expression of complementarity. At first sight, this might seem odd, since, after all, complementarity corresponds to a dichotomic relation between two types of description. The uncertainty relations "express" this dichotomy in the informal sense that if we take energy and momentum to be perfectly well-defined, i.e., symbolically $\Delta E = \Delta p = 0$, the position and time variables are completely undefined, $\Delta x = \Delta t = \infty$, and vice versa. However, by focussing on these extremes only, we leave out of consideration that the uncertainty relations also (and more properly) allow for an intermediate situation in which the mentioned uncertainties are all non-zero and finite. This more positive aspect of the uncertainty relation is mentioned in the Como lecture:

> At the same time, however, the general character of this relation makes it possible to a

> certain extent to reconcile the conservation laws with the space-time coordination of observations, the idea of a coincidence of well-defined events in space-time points being replaced by that of unsharply defined individuals within space-time regions. (Bohr 1928, p. 571)

However, Bohr never followed up on this suggestion that we might be able to strike a compromise between the two mutually exclusive modes of description in terms of unsharply defined quantities. Indeed, an attempt to do so, would take the formalism of quantum theory more seriously than the concepts of classical language, and this step Bohr refused to take. Instead, in his later writings he would be content with stating that the uncertainty relations simply defy an unambiguous interpretation in classical terms:

> These so-called indeterminacy relations explicitly bear out the limitation of causal analysis, but it is important to recognize that no unambiguous interpretation of such a relation can be given in words suited to describe a situation in which physical attributes are objectified in a classical way. (Bohr, 1948, p.315)

> It must here be remembered that even in the indeterminacy relation [$\Delta q \Delta p \approx h$] we are dealing with an implication of the formalism which defies unambiguous expression in words suited to describe classical pictures. Thus a sentence like "we cannot know both the momentum and the position of an atomic object" raises at once questions as to the physical reality of two such attributes of the object, which can be answered only by referring to the conditions for an unambiguous use of space-time concepts, on the one hand, and dynamical conservation laws on the other hand. (Bohr, 1949, p. 211)

Finally, on a more formal level, we note that Bohr's derivation does not rely on the commutation relations (1) and (5), but on Fourier analysis. To be sure, these two approaches are equivalent as far as the relationship between position and momentum is concerned. But this is not so for time and energy. This means that, for a derivation based on the commutation relations, the position-momentum and time-energy relations are not on an equal footing, which is contrary to Bohr's approach in terms of Fourier analysis (Hilgevoord, 1996 and 1998).

4. The Minimal Interpretation

In the previous two sections we have seen how both Heisenberg and Bohr attributed a far-reaching status to the uncertainty relations. They both argued that these relations place fundamental limits on the applicability of the usual classical concepts. Moreover, they both believed that these limitations were inevitable and forced upon us. However, we have also seen that they reached such conclusions by starting from radical and controversial assumptions. This entails, of course, that their radical conclusions remain unconvincing for those who reject these, or other assumptions. Indeed, the operationalist-positivist viewpoint adopted by these authors has long since lost its appeal among philosophers of physics.

So the question may be asked what alternative views of the uncertainty relations are still viable. Of course, this problem is intimately connected with that of the interpretation of the wave function, and hence of quantum mechanics as a whole. Since there is no consensus about the latter, one cannot expect consensus about the interpretation of the uncertainty relations either. Here we only describe a point of view, which we call the 'minimal interpretation', which seems to be shared by both the adherents of the Copenhagen interpretation and of other views.

In quantum mechanics a system is supposed to be described by its quantum state, also called its state vector. Given the state vector, one can derive probability distributions for all the physical quantities pertaining to the system such as its position, momentum, angular momentum, energy, etc. The operational meaning of these probability distributions is that they correspond to the distribution of the values obtained for these quantities in a long series of repetitions of the measurement. More precisely, one imagines a great number of copies of the system under consideration, all prepared in the same way. On each copy the momentum, say, is measured. Generally, the outcomes of these measurements differ and a distribution of outcomes is obtained. The theoretical momentum distribution derived from the quantum state is supposed to coincide with the hypothetical distribution of outcomes obtained in an infinite series of repetitions of the momentum measurement. The same holds, *mutatis mutandis*, for all the other physical quantities pertaining to the system. Note that no simultaneous measurements of two or more quantities are required in defining the operational meaning of the probability distributions.

Uncertainty relations can be considered as statements about the spreads of the probability distributions of the several physical quantities arising from the same state. For example, the uncertainty relation between the position and momentum of a system may be understood as the statement that the position and momentum distributions cannot both be arbitrarily narrow -- in some sense of the word "narrow" -- in any quantum state. Inequality (9) is an example of such a relation in which the standard deviation is employed as a measure of spread. From this characterization of uncertainty relations follows that a more detailed interpretation of the quantum state than the one given in the previous paragraph is not required to study uncertainty relations as such. In particular, a further ontological or linguistic interpretation of the notion of uncertainty, as limits on the applicability of our concepts given by Heisenberg or Bohr, need not be supposed.

Indeed, this minimal interpretation leaves open whether it makes sense to attribute precise values of position and momentum to an individual system. Some interpretations of quantum mechanics, e.g. Heisenberg and Bohr, deny this; while others, e.g. the interpretation of de Broglie and Bohm insist that each individual system has a definite position and momentum. The only requirement is that, as an empirical fact, it is not possible to prepare pure ensembles in which all systems have the same values for these quantities, or ensembles in which the spreads are smaller than allowed by quantum theory. Although interpretations of quantum mechanics, in which each system has a definite value for its position and momentum are still viable, this is not to say that they are without problems or, at least strange features, of their own. They do not imply a return to classical physics.

We end with a few remarks on this minimal interpretation. First, it may be noted that the minimal interpretation of the uncertainty relations is little more than filling in the empirical meaning of inequality

(9), or an inequality in terms of other measures of width, as obtained from the standard formalism of quantum mechanics. As such, this view shares many of the limitations we have noted above about this inequality. Indeed, it is not straightforward to relate the spread in a statistical distribution of measurement results with the *inaccuracy* of this measurement, such as, e.g. the resolving power of a microscope. Moreover, the minimal interpretation does not address the question whether one can make *simultaneous* accurate measurements of position and momentum. As a matter of fact, one can show that the standard formalism of quantum mechanics does not allow such simultaneous measurements. But this is not a consequence of relation (9).

If one feels that statements about inaccuracy of measurement, or the possibility of simultaneous measurements, belong to any satisfactory formulation of the uncertainty principle, the minimal interpretation may thus be too minimal.

Bibliography

- Beller M. (1999) *Quantum Dialogues* (Chicago: University of Chicago Press).
- Bohr, N. (1928) ‘The Quantum postulate and the recent development of atomic theory’ *Nature* (Supplement) **121** 580-590. Also in (Bohr, 1934), (Wheeler and Zurek, 1983), and in (Bohr, 1985).
- Bohr, N. (1929) ‘Introductory survey’ in (Bohr, 1934), pp. 1-24.
- Bohr, N. (1934) *Atomic Theory and the Description of Nature* (Cambridge: Cambridge University Press). Reissued in 1961. Appeared also as Volume I of *The Philosophical Writings of Niels Bohr* (Woodbridge Connecticut: Ox Bow Press, 1987).
- Bohr, N. (1937) ‘Causality and complementarity’ *Philosophy of Science* **4** 289-298.
- Bohr, N. (1939) ‘The causality problem in atomic physics’ in *New Theories in Physics* (Paris: International Institute of Intellectual Co-operation). Also in (Bohr, 1996), pp. 303-322.
- Bohr, N. (1948) ‘On the notions of causality and complementarity’ *Dialectica* **2** 312-319. Also in (Bohr, 1996) pp. 330-337.
- Bohr, N. (1985) *Collected Works* Volume 6, J. Kalckar (ed.) (Amsterdam: North-Holland).
- Bohr, N. (1996) *Collected Works* Volume 7, J. Kalckar (ed.) (Amsterdam: North-Holland).
- Bub, J. (2000) ‘Quantum mechanics as a principle theory’ *Studies in History and Philosophy of Modern Physics* **31B** 75-94.
- Cassidy, D.C. (1992) *Uncertainty, the Life and Science of Werner Heisenberg* (New York: Freeman).
- Cassidy, D.C. (1998) ‘Answer to the question: When did the indeterminacy principle become the uncertainty principle?’ *American Journal of Physics* **66** 278-279.
- Condon, E.U. (1929) ‘Remarks on uncertainty principles’ *Science* **69** 573-574.
- Eddington, A. (1928) *The Nature of the Physical World*, (Cambridge: Cambridge University Press).
- Einstein, A. (1919) ‘My Theory’, *The London Times*, November 28, p. 13. Reprinted as ‘What is the theory of relativity?’ in *Ideas and Opinions* (New York: Crown Publishers, 1954) pp. 227-232.
- Folse, H.J. (1985) *The Philosophy of Niels Bohr* (Amsterdam: Elsevier).

- Heisenberg, W. (1925) 'Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen' *Zeitschrift für Physik* **33** 879-893.
- Heisenberg, W. (1926) 'Quantenmechanik' *Die Naturwissenschaften* **14** 899-894.
- Heisenberg, W. (1927) 'Ueber den anschaulichen Inhalt der quantentheoretischen Kinematik and Mechanik' *Zeitschrift für Physik* **43** 172-198. English translation in (Wheeler and Zurek, 1983), pp. 62-84.
- Heisenberg, W. (1927) 'Ueber die Grundprincipien der "Quantenmechanik" ' *Forschungen und Fortschritte* **3** 83.
- Heisenberg, W. (1928) 'Erkenntnistheoretische Probleme der modernen Physik' in (Heisenberg, 1984), pp. 22-28.
- Heisenberg W. (1930) *Die Physikalischen Prinzipien der Quantenmechanik* (Leipzig: Hirzel). English translation *The Physical Principles of Quantum Theory* (Chicago: University of Chicago Press, 1930).
- Heisenberg, W. (1931) 'Die Rolle der Unbestimmtheitsrelationen in der modernen Physik' *Monatshefte für Mathematik und Physik* **38** 365-372.
- Heisenberg, W. (1958) *Physics and Philosophy* (New York: Harper).
- Heisenberg, W. (1969) *Der Teil und das Ganze* (München : Piper).
- Heisenberg, W. (1975) 'Bemerkungen über die Entstehung der Unbestimmtheitsrelation' *Physikalische Blätter* **31** 193-196. English translation in (Price and Chissick, 1977).
- Heisenberg W. (1984) *Gesammelte Werke* Volume C1, W. Blum, H.-P. Dürr and H. Rechenberg (eds) (München: Piper).
- Hilgevoord, J. and Uffink, J. (1988) 'The mathematical expression of the uncertainty principle' in *Microphysical Reality and Quantum Description*, A. van der Merwe et al. (eds.), (Dordrecht: Kluwer) pp. 91-114.
- Hilgevoord, J. (1996) 'the uncertainty principle for energy and time I' *American Journal of Physics* **64**, 1451-1456.
- Hilgevoord, J. (1998) 'the uncertainty principle for energy and time II' *American Journal of Physics* **66**, 396-402.
- Hilgevoord, J. (2001) 'Time in quantum mechanics' *American Journal of Physics* (to appear).
- Jammer, M. (1974) *The Philosophy of Quantum Mechanics* (New York: Wiley).
- Jordan, P. (1927) 'Über eine neue Begründung der Quantenmechanik II' *Zeitschrift für Physik* **44** 1-25.
- Kaiser, H., Werner, S.A., and George, E.A. (1983) 'Direct measurement of the longitudinal coherence length of a thermal neutron beam' *Physical Review Letters* **50** 560.
- Kennard E.H. (1927) 'Zur Quantenmechanik einfacher Bewegungstypen' *Zeitschrift für Physik*, **44** 326-352.
- Miller, A.I. (1982) 'Redefining Anschaulichkeit' in: A. Shimony and H.Feshbach (eds) *Physics as Natural Philosophy* (Cambridge Mass.: MIT Press).
- Muller, F.A. (1997) 'The equivalence myth of quantum mechanics' *Studies in History and Philosophy of Modern Physics* **28** 35-61, 219-247, ibid. **30** (1999) 543-545.
- Murdoch, D. (1987) *Niels Bohr's Philosophy of Physics* (Cambridge: Cambridge University Press).
- Pauli, W. (1933) 'Die allgemeinen Prinzipien der Wellenmechanik' in K. Geiger, and H. Scheel

(eds) *Handbuch der Physik* 2nd edition, Vol. 245, (Berlin: Springer).

- Pauli, W. (1979) *Wissentschaftlicher Briefwechsel mit Bohr, Einstein, Heisenberg u.a.* Volume 1 (1919-1929) A. Hermann, K. von Meyenn and V.F. Weiskopf (eds) (Berlin: Springer).
- Popper, K. (1967) 'Quantum mechanics without "the observer"' in M. Bunge (ed.) *Quantum Theory and Reality* (Berlin: Springer).
- Price, W.C. and Chissick, S.S (eds) (1977) *The Uncertainty Principle and the Foundations of Quantum Mechanics*, (New York: Wiley).
- Regt, H. de (1997) 'Erwin Schrödinger, *Anschaulichkeit*, and quantum theory' *Studies in History and Philosophy of Modern Physics* **28** 461-481.
- Robertson, H.P. (1929) 'The uncertainty principle' *Physical Review* **34** 573-574. Reprinted in Wheeler and Zurek (1983) pp. 127-128.
- Scheibe, E. (1973) *The Logical Analysis of Quantum Mechanics* (Oxford: Pergamon Press).
- Schrödinger, E. (1930) 'Zum Heisenbergschen Unschärfeprinzip' *Berliner Berichte* 296-303.
- Uffink, J. (1985) 'Verification of the uncertainty principle in neutron interferometry' *Physics Letters* **108 A** 59-62.
- Uffink, J. and Hilgevoord, J. (1985) 'Uncertainty principle and uncertainty relations' *Foundations of Physics* **15** 925-944.
- Wheeler, J.A. and Zurek, W.H. (eds) (1983) *Quantum Theory and Measurement* (Princeton NJ: Princeton University Press).

Other Internet Resources

- Hilgevoord, J. (2001) 'Time in quantum mechanics' (preprint available online)
- Nairz, O., Andt, M. and Zeilinger, A. (2001) 'Experimental confirmation of the Heisenberg uncertainty principle for hot fullerene molecules' (preprint available online)
- American Institute of Physics Exhibit on Heisenberg and the Uncertainty Principle

Related Entries

quantum mechanics

Jos Uffink
uffink@phys.uu.nl
and
Jan Hilgevoord
Institute for History and Foundations of Science
janhilgevoord@netnet.nl

Table of Contents

First published: October 8, 2001
Content last modified: October 26, 2001

Stanford Encyclopedia of Philosophy
Notes to Uncertainty Relations in Quantum Theory

Notes

1. The translation of *anschaulich* as ‘intuitive’ is obviously the best candidate, since it has a corresponding etymological root. Unfortunately, this term is often used to refer to a kind of understanding which bypasses reasoning. This is not intended here.

2. Although this claim by Schrödinger seems to have been accepted by his contemporaries, his argument was incomplete. A satisfactory proof of the equivalence of wave mechanics and matrix mechanics was only provided later, when von Neumann (1932) showed that both approaches could be seen as different realisations of an abstract Hilbert space formalism. (See Muller, 1997, for historical details.)

3. However, the first draft of his paper, which he sent as a letter to Pauli (Pauli, 1979 pp. 376-382) shows that Heisenberg did have the outlines of a general, albeit qualitative, mathematical argument.

4. Note that this usage of the term ‘causal’ by Bohr differs from his usage of that term in earlier texts, where it refers only to the applicability of dynamical conservation laws, and not to the union of a space-time description with these conservation laws. Thus, in (Bohr, 1928), he characterized complementarity as a relation between space-time description and causality.

5. Note, that while Bohr started from the duality between the particle and wave pictures, which are mutually exclusive also in classical physics, he later considered as complementary two descriptions which in the classical theory are united.

6. See footnote 4 for Bohr's usage of the term ‘causality’.

Jos Uffink
uffink@phys.uu.nl
and
Jan Hilgevoord
Institute for History and Foundations of Science
janhilgevoord@netnet.nl

First published: October 8, 2001
Content last modified: October 8, 2001

The Many-Worlds Interpretation of Quantum Mechanics

The Many-Worlds Interpretation (MWI) is an approach to quantum mechanics according to which, in addition to the world we are aware of directly, there are many other similar worlds which exist in parallel at the same space and time. The existence of the other worlds makes it possible to remove randomness and action at a distance from quantum theory and thus from all physics.

1. Introduction

The fundamental idea of the MWI, going back to Everett 1957, is that there are myriads of worlds in the Universe in addition to the world we are aware of. In particular, every time a quantum experiment with different outcomes with non-zero probability is performed, all outcomes are obtained, each in a different world, even if we are aware only of the world with the outcome we have seen. In fact, quantum experiments take place everywhere and very often, not just in physics laboratories: even the irregular blinking of an old fluorescent bulb is a quantum experiment.

There are numerous variations and reinterpretations of the original Everett proposal, most of which are briefly discussed in the entry on Everett's relative state formulation of quantum mechanics. Here, a particular approach to the MWI (which differs from the popular "actual splitting worlds" approach in De Witt 1970) will be presented in detail, followed by a discussion relevant for many variants of the MWI.

The MWI consists of two parts:

i. A mathematical theory which yields evolution in time of the quantum state of the (single) Universe.
ii. A prescription which sets up a correspondence between the quantum state of the Universe and our experiences.

Part (i) is essentially summarized by the Schrödinger equation or its relativistic generalization. It is a rigorous mathematical theory and is not problematic philosophically. Part (ii) involves "our experiences" which do not have a rigorous definition. An additional difficulty in setting up (ii) follows from the fact that human languages were developed at a time when people did not suspect the existence of parallel worlds. This, however, is only a semantic problem.[1]

2. Definitions

2.1 What is "A World"?

> *A world is the totality of (macroscopic) objects: stars, cities, people, grains of sand, etc. in a definite classically described state.*

This definition is based on the common attitude to the concept of world shared by human beings.

Another concept (considered in some approaches as the basic one, e.g., in Saunders 1995) is a relative, or perspectival, world defined for every physical system and every one of its states (provided it is a state of non-zero probability): I will call it a *centered world*. This concept is useful when a world is centered on a

perceptual state of a sentient being. In this world, all objects which the sentient being perceives have definite states, but objects that are not under her observation might be in a superposition of different (classical) states. The advantage of a centered world is that it does not split due to a quantum phenomenon in a distant galaxy, while the advantage of our definition is that we can consider a world without specifying a center, and in particular our usual language is just as useful for describing worlds at times when there were no sentient beings.

The concept of "world" in the MWI belongs to part (ii) of the theory, i.e., it is not a rigorously defined mathematical entity, but a term defined by us (sentient beings) in describing our experience. When we refer to the "definite classically described state" of, say, a cat, it means that the position and the state (alive, dead, smiling, etc.) of the cat is maximally specified according to our ability to distinguish between the alternatives and that this specification corresponds to a classical picture, e.g., no superpositions of dead and alive cats are allowed in a single world.[2]

The concept of a world in the MWI is based on the layman's conception of a world; however, several features are different:

Obviously, the definition of the world as *everything that exists* does not hold in the MWI. "Everything that exists" is the Universe, and there is only one Universe. The Universe incorporates many worlds similar to the one the layman is familiar with.

Nowadays, the layman knows that objects are made of elementary microscopic particles, and he believes that, consequently, a more precise definition of the world is the totality of all these particles. In the MWI this naive step is incorrect. Microscopic particles might be in a superposition, while objects within a world (as defined in the MWI) cannot be in a superposition. The connection between macroscopic objects defined according to our experience, and microscopic objects defined in a physical theory that aims to explain our experience, is more subtle, and will be discussed further below. The definition of a world in the MWI involves only concepts related to our experience.

A layman believes that our present world has a unique past and future. According to the MWI, a world defined at some moment of time corresponds to a unique world at a time in the past, but to a multitude of worlds at a time in the future.

2.2 Who am "I"?

"I" am an object, such as Earth, cat, etc. "I" is defined at a particular time by a complete (classical) description of the state of my body and of my brain. "I" and "Lev" do not name the same things (even though my name is Lev). At the present moment there are many different "Lev"s in different worlds (not more than one in each world), but it is meaningless to say that now there is another "I". I have a particular, well defined past: I correspond to a particular "Lev" in 2002, but I do not have a well defined future: I correspond to a multitude of "Lev"s in 2010. In the framework of the MWI it is meaningless to ask: Which Lev in 2010 will I be? I will correspond to them all. Every time I perform a quantum

experiment (with several possible results) it only seems to me that I obtain a single definite result. Indeed, Lev who obtains this particular result thinks this way. However, this Lev cannot be identified as the only Lev after the experiment. Lev before the experiment corresponds to all "Lev"s obtaining all possible results. Although this approach to the concept of personal identity seems somewhat unusual, it is plausible in the light of the critique of personal identity by Parfit 1986. Parfit considers some artificial situations in which a person splits into several copies, and argues that there is no good answer to the question: Which copy is me? He concludes that personal identity is not what matters when I divide.

3. Correspondence Between the Formalism and Our Experience

3.1 The Quantum State of an Object

The basis for the correspondence between the quantum state (the wave function) of the Universe and our experience is the description that physicists give in the framework of standard quantum theory for objects composed of elementary particles. Elementary particles of the same kind are identical. Therefore, the essence of an object is the quantum state of its particles and not the particles themselves (see the elaborate discussion in the entry on identity and individuality in quantum theory): one quantum state of a set of elementary particles might be a cat and another state of the same particles might be a small table. Clearly, we cannot now write down an exact wave function of a cat. We know with a reasonable approximation the wave function of some elementary particles that constitute a nucleon. The wave function of the electrons and the nucleons that together make up an atom is known with even better precision. The wave functions of molecules (i.e. the wave functions of the ions and electrons out of which molecules are built) are well studied. A lot is known about biological cells, so physicists can write down a rough form of the quantum state of a cell. This is difficult because there are many molecules in a cell. Out of cells we construct various tissues and then the whole body of a cat or of a table. So, let us denote the quantum state constructed in this way $|\Psi\rangle_{OBJECT}$.

In our construction $|\Psi\rangle_{OBJECT}$ is the quantum state of an object in a definite state and position.[3] According to the definition of a world we have adopted, in each world the cat is in a definite state: either alive or dead. Schrödinger's experiment with the cat leads to a splitting of worlds even before opening the box. Only in the alternative approach is Schrödinger's cat, which is in a superposition of being alive and dead, a member of the (single) centered world of the observer before she opened the sealed box with the cat (the observer perceives directly the facts related to the preparation of the experiment and she deduces that the cat is in a superposition).

3.2 The Quantum State that corresponds to a World

The wave function of all particles in the Universe corresponding to any particular world will be a product

of states of sets of particles corresponding to all objects in the world multiplied by the quantum state $|\Phi\rangle$ of all the particles that do not constitute "objects". Within a world, "objects" have definite macroscopic states by fiat:[4]

$$|\Psi_{\text{WORLD}}\rangle = |\Psi\rangle_{\text{OBJECT 1}} |\Psi\rangle_{\text{OBJECT 2}} \ldots |\Psi\rangle_{\text{OBJECT } N} \quad |\Phi\rangle \qquad (1)$$

The quantum states corresponding to centered worlds of sentient beings have exactly the same form. The only difference is that in the product there are only states of the objects perceived directly, while most of the universe is, in general, entangled; it is described by $|\Phi\rangle$.

3.3 The Quantum State of the Universe

The quantum state of the Universe can be decomposed into a superposition of terms corresponding to different worlds:

$$|\Psi_{\text{UNIVERSE}}\rangle = \sum \alpha_i |\Psi_{\text{WORLD } i}\rangle \qquad (2)$$

Different worlds correspond to different classically described states of at least one object. Different classically described states correspond to orthogonal quantum states. Therefore, different worlds correspond to orthogonal states: all states $|\Psi_{\text{WORLD } i}\rangle$ are mutually orthogonal and consequently, $\sum|\alpha_i|^2 = 1$.

3.4 FAPP

The construction of the quantum state of the Universe in terms of the quantum states of objects presented above is only approximate, it is good only *for all practical purposes* (FAPP). Indeed, the concept of an object itself has no rigorous definition: should a mouse that a cat just swallowed be considered as a part of the cat? The concept of a "definite position" is also only approximately defined: how far should a cat be displaced in order for it to be considered to be in a different position? If the displacement is much smaller than the quantum uncertainty, it must be considered to be at the same place, because in this case the quantum state of the cat is almost the same and the displacement is undetectable in principle. But this is only an absolute bound, because our ability to distinguish various locations of the cat is far from this quantum limit. Further, the state of an object (e.g. alive or dead) is meaningful only if the object is considered for a period of time. In our construction, however, the quantum state of an object is defined at a particular time. In fact, we have to ensure that the quantum state will have the shape of the object not only at that time, but for some period of time. Splitting of the world during this period of time is another source of ambiguity, in particular, due to the fact that there is no precise definition of when the splitting occurs.

The reason that I am only able to propose an approximate prescription for correspondence between the quantum state of the Universe and our experience, is essentially the same that led Bell 1990 to claim that "ordinary quantum mechanics is just fine FAPP". The concepts we use: "object", "measurement", etc. are not rigorously defined. Bell was, and many others are looking (until now in vain) for a "precise quantum mechanics". Since it is not enough for a physical theory to be just fine FAPP, a quantum mechanics needs rigorous foundations. However, in the MWI just fine FAPP is enough. Indeed, the MWI has rigorous foundations for (i), the "physics part" of the theory; only part (ii), the correspondence with our experience, is approximate (just fine FAPP). But "just fine FAPP" means that the theory explains our experience for any possible experiment, and this is the goal of (ii). See Butterfield 2001 and Wallace 2001b for more arguments why a FAPP definition of a world ("branch" in their language) is enough.

3.5 The Measure of Existence

There are many worlds existing in parallel in the Universe. Although all worlds are of the same physical size (this might not be true if we take quantum gravity into account), and in every world sentient beings feel as "real" as in any other world, in some sense some worlds are larger than others. I describe this property as the *measure of existence* of a world.[5] The measure of existence of a world quantifies its ability to interfere with other worlds in a gedanken experiment, see Vaidman 1998 (p. 256), and is the basis for introducing *probability* in the MWI. The measure of existence makes precise what is meant by the probability measure discussed in Everett 1957 and pictorially described in Lockwood 1989 (p. 230).

Given the decomposition (2), the measure of existence of the world i is $\mu_i = |\alpha_i|^2$. It also can be expressed as the expectation value of $\mathbf{P}_i$, the projection operator on the space of quantum states corresponding to the actual values of all physical variables describing the world i:

$$\mu_i \equiv \langle \Psi_{\mathrm{UNIVERSE}} | \mathbf{P}_i | \Psi_{\mathrm{UNIVERSE}} \rangle \qquad (3)$$

"I" also have a measure of existence. It is the sum of measures of existence of all different worlds in which I exist; equally, it can be defined as the measure of existence of my perception world. Note that I do not experience directly the measure of my existence. I feel the same weight, see the same brightness, etc. irrespectively of how tiny my measure of existence might be.

4. Probability in the MWI

There is a serious difficulty with the concept of probability in the context of the MWI. In a deterministic theory, such as the MWI, the only possible meaning for probability is an *ignorance* probability, but there is no relevant information that an observer who is going to perform a quantum experiment is ignorant about. The quantum state of the Universe at one time specifies the quantum state at all times. If I am going to perform a quantum experiment with two possible outcomes such that standard quantum

mechanics predicts probability 1/3 for outcome A and 2/3 for outcome B, then, according to the MWI, both the world with outcome A and the world with outcome B will exist. It is senseless to ask: "What is the probability that I will get A instead of B?" because I will correspond to both "Lev"s: the one who observes A and the other one who observes B.[6]

To solve this difficulty, Albert and Loewer 1988 proposed the Many Minds interpretation (in which the different worlds are only in the minds of sentient beings). In addition to the quantum wave of the Universe, Albert and Loewer postulate that every sentient being has a continuum of minds. Whenever the quantum wave of the Universe develops into a superposition containing states of a sentient being corresponding to different perceptions, the minds of this sentient being evolve randomly and independently to mental states corresponding to these different states of perception (with probabilities equal to the quantum probabilities for these states). In particular, whenever a measurement is performed by an observer, the observer’s minds develop mental states that correspond to perceptions of the different outcomes, i.e. corresponding to the worlds A or B in our example. Since there is a continuum of minds, there will always be an infinity of minds in any sentient being and the procedure can continue indefinitely. This resolves the difficulty: each "I" corresponds to one mind and it ends up in a state corresponding to a world with a particular outcome. However, this solution comes at the price of introducing additional structure into the theory, including a genuinely random process.

Vaidman1998 (p. 254) resolves the problem by constructing an *ignorance probability* in the framework of the MWI. It seems senseless to ask: "What is the probability that Lev in the world A will observe A?" This probability is trivially equal to 1. The task is to define the probability in such a way that we could reconstruct the prediction of the standard approach: probability 1/3 for A. It is indeed senseless *for you* to ask what is the probability that Lev in the world A will observe A, but this might be a meaningful question for Lev in the world of the outcome A. Under normal circumstances, the world A is created (i.e. measuring devices and objects which interact with measuring devices will become localized according to the outcome A) before Lev will be aware of the result A. Then, it is sensible to ask this Lev about his probability to be in world A. There is a matter of fact about which outcome this Lev will see, but he is ignorant about this fact at the time of the question. In order to make this point vivid, Vaidman proposed an experiment in which the experimenter is given a sleeping pill before the experiment. Then, while asleep, he is moved to room A or to room B depending on the results of the experiment. When the experimenter has woken up (in one of the rooms), but before he has opened his eyes, he is asked "In which room are you?" Certainly, there is a matter of fact about which room he is in (he can learn about it by opening his eyes), but he is ignorant about this fact at the time of the question. This construction provides the ignorance interpretation of probability, but the value of the probability has to be postulated (see Section 6.3 below for attempts to derive it):

> **Probability Postulate**
> The probability of an outcome of a quantum experiment is proportional to the total measure of existence of all worlds with that outcome.[7]

The question of the probability of obtaining A also makes sense for the Lev in world B before he

becomes aware of the outcome. Both "Lev"s have the same information on the basis of which they should give their answer. According to the probability postulate they will give the *same* answer: 1/3 (the relative measure of existence of the world A). Since Lev before the measurement is associated with two "Lev"s after the measurement who have identical ignorance probability concepts for the outcome of the experiment, I can define the probability of the outcome of the experiment to be performed as the ignorance probability of the successors of Lev for being in a world with a particular outcome.

The "sleeping pill" argument does not reduce the probability of an outcome of a quantum experiment to a familiar concept of probability in the classical context. The quantum situation is genuinely different. Since all outcomes of a quantum experiment are actualized, there is no probability in the usual sense. The argument explains the Behavior Principle (see below) for an experimenter according to which he should behave as if there were certain probabilities for different outcomes. The justification is particularly clear in the approach to probability as the value of a rational bet on a particular result. The results of the betting of the experimenter are relevant for his successors emerging after performing the experiment in different worlds. Since the experimenter is related to all of his successors and they all have identical rational strategies for betting, then, this should also be the strategy of the experimenter before the experiment.

Several authors justify the probability postulate without relying on the sleeping pill argument. Tappenden 2000 (p. 111) adopts a different semantics according to which "I" live in all branches and have "distinct experiences" in different "superslices", and uses "weight of a superslice" instead of measure of existence. He argues that it is intelligible to associate probabilities according to the probability postulate: "Faced with an array of weighted superslices as part of myself ... what choice do I have but to assign an array of attitudes, degrees of belief, towards the experiences associated with those superslices?". Saunders 1998, exploiting a variety of ideas in decoherence theory, the relational theory of tense and theories of identity over time, also argues for "identification of probability with the Hilbert Space norm" (which equals the measure of existence). Page 2002 promotes an approach which he has recently named *Mindless Sensationalism*. The basic concept in this approach is a conscious experience. He assigns *weights* to different experiences depending on the quantum state of the universe, as the expectation values of presently-unknown positive operators corresponding to the experiences (similar to the measures of existence of the corresponding worlds (3)). Page writes "... experiences with greater weights exist in some sense more ..." In all of these approaches, the postulate is justified by appeal to an analogy with treatments of time, e.g., the measure of existence of a world is analogous to the duration of a time interval. In a more ambitious work, Deutsch 1999 has claimed to derive the probability postulate from the quantum formalism and the classical decision theory, but it is far from clear that he achieves this (see Barnum *et al.*).

5. Tests of the MWI

Despite the name "interpretation", the MWI is a variant of quantum theory that is different from others. Experimentally, the difference is relative to collapse theories. It seems that there is no experiment

distinguishing the MWI from other no-collapse theories such as Bohmian mechanics or other variants of MWI.

The collapse leads to effects that are, in principle, observable; these effects do not exist if the MWI is the correct theory. To observe the collapse we would need a super technology, which allows "undoing" a quantum experiment, including a reversal of the detection process by macroscopic devices. See Lockwood 1989 (p. 223), Vaidman 1998 (p. 257), and other proposals in Deutsch 1986. These proposals are all for gedanken experiments that cannot be performed with current or any foreseen future technology. Indeed, in these experiments an interference of different worlds has to be observed. Worlds are different when at least one macroscopic object is in macroscopically distinguishable states. Thus, what is needed is an interference experiment with a macroscopic body. Today there are interference experiments with larger and larger objects (e.g., fullerene molecules C_{60}), but these objects are still not large enough to be considered "macroscopic". Such experiments can only refine the constraints on the boundary where the collapse might take place. A decisive experiment should involve the interference of states which differ in a macroscopic number of degrees of freedom: an impossible task for today's technology.[8]

The collapse mechanism seems to be in contradiction with basic physical principles such as relativistic covariance, but nevertheless, some ingenious concrete proposals have been made (see Pearle 1986 and the entry on collapse theories). These proposals (and Weissman's 1999 non-linear MW idea) have additional observable effects, such as a tiny energy non-conservation, that were tested in several experiments. The effects were not found and some (but not all!) of these models have been ruled out.

In most no-collapse interpretations, the evolution of the quantum state of the Universe is the same. Still, one might imagine that there is an experiment distinguishing the MWI from another no-collapse interepretation based on the difference in the correspondence between the formalism and the experience (the results of experiments).

An apparent candidate for such an experiment is a setup proposed in Englert *et al*. 1992 in which a Bohmian world is different from the worlds of the MWI (see also Aharonov and Vaidman 1996). In this example, the Bohmian trajectory of a particle in the past is contrary to the records of seemingly good measuring devices (such trajectories were named *surrealistic*). However, at present, there are no memory records that can determine unambiguously (without deduction from a particular theory) the particle trajectory in the past. Thus, this difference does not lead to an experimental way of distinguishing between the MWI and Bohmian mechanics. I believe that no other experiment can distinguish between the MWI and other no-collapse theories either, except for some perhaps exotic modifications, e.g., Bohmian mechanics with initial particle position distribution deviating from the quantum distribution.

There are other opinions about the possibility of testing the MWI. It has frequently been claimed, e.g. by De Witt 1970, that the MWI is in principle indistinguishable from the ideal collapse theory. On the other hand, Plaga 1997 claims to have a realistic proposal for testing the MWI, and Page 2000 argues that

certain cosmological observations might support the MWI.

6. Objections to the MWI

Some of the objections to the MWI follow from misinterpretations due to the multitude of various MWIs. The terminology of the MWI can be confusing: "world" is "universe" in Deutsch 1996, while "universe" is "multiverse", etc. There are two very different approaches with the same name "The Many-Minds Interpretation (MMI)". The Albert and Loewer 1988 MMI mentioned above should not be confused with Lockwood' 1996 MMI (which resembles the approach of Zeh 1981). The latter is much closer to the MWI as it is presented here, see Sec. 17 of Vaidman 1998. Further, the MWI in the Heisenberg representation (Deutsch 2001) differs significantly from the MWI presented in the Schrödinger representation (used here). The MWI presented here is very close to Everett's original proposal, but in the entry on Everett's relative state formulation of quantum mechanics, as well as in his book Barrett 1999, Barrett uses the name "MWI" for the splitting worlds view publicized by De Witt 1970. This approach has been justly criticized: it has both some kind of collapse (an irreversible splitting of worlds in a preferred basis) and the multitude of worlds. Now I consider the main objections in detail.

6.1 Ockham's Razor

It seems that the majority of the opponents of the MWI reject it because, for them, introducing a very large number of worlds that we do not see is an extreme violation of Ockham's principle: "Entities are not to be multiplied beyond necessity". However, in judging physical theories one could reasonably argue that one should not multiply physical laws beyond necessity either (such a verion of Ockham's Razor has been applied in the past), and in this respect the MWI is the most economical theory. Indeed, it has all the laws of the standard quantum theory, but without the collapse postulate, the most problematic of physical laws. The MWI is also more economic than Bohmian mechanics which has in addition the ontology of the particle trajectories and the laws which give their evolution. Tipler 1986 (p. 208) has presented an effective analogy with the criticism of Copernican theory on the grounds of Ockham's razor.

One might consider also a possible philosophical advantage of the plurality of worlds in the MWI, similar to that claimed by realists about possible worlds, such as Lewis 1986 (see the discussion of the analogy between the MWI and Lewis's theory by Skyrms 1976). However, the analogy is not complete: Lewis' theory considers all logically possible worlds, many more than all worlds incorporated in the quantum state of the Universe.

6.2 The Problem of the Preferred Basis

A common criticism of the MWI stems from the fact that the formalism of quantum theory allows infinitely many ways to decompose the quantum state of the Universe into a superposition of orthogonal

states. The question arises: "Why choose the particular decomposition (2) and not any other?" Since other decompositions might lead to a very different picture, the whole construction seems to lack predictive power.

Indeed, the mathematical structure of the theory (i) does not yield a particular basis. The basis for decomposition into worlds follows from the common concept of a world according to which it consists of objects in definite positions and states ("definite" on the scale of our ability to distinguish them). In the alternative approach, the basis of a centered world is defined directly by an observer. Therefore, given the nature of the observer and given her concepts for describing the world, the particular choice of the decomposition (2) follows (up to a precision which is good FAPP, as required). If we do not ask why we are what we are, and why the world we perceive is what it is, but only how to explain relations between the events we observe in our world, then the problem of the preferred basis does not arise: we and the concepts of our world define the preferred basis.

But a stronger response can be made to this criticism. Looking at the details of the physical world, the structure of the Hamiltonian, the value of the Planck constant, etc., one can argue why the sentient beings we know are of a particular type and why they have the particular concepts they do for describing their worlds. The main argument is that the locality of interactions yields *stability* of worlds in which objects are well localized. The small value of the Planck constant allows macroscopic objects to be well localized for a considerable period of time. Thus, such worlds (corresponding to quantum states $|\Psi_i\rangle$) can maintain their macroscopic description long enough to be perceived by sentient beings. By constrast, a "world" with macroscopic objects being in a superposition of macroscopically distinguishable states (corresponding to a quantum state $1/\sqrt{2}(|\Psi_1\rangle+|\Psi_2\rangle)$ evolves during an extremely small time, much smaller than the perception time of any feasible sentient being, into a mixture with the other "world" $1/\sqrt{2}(|\Psi_1\rangle-|\Psi_2\rangle)$ (see Zurek 1998).

This is a good argument why sentient beings perceive localized objects and not superpositions, but one cannot rely on the decoherence argument alone in order to single out the proper basis. (See some technical difficulties in Barvinsky and Kamenshchik 1995.) The fact that we can perceive only well localized objects in definite macroscopic states might not be just a physics issue: chemistry, biology, and even psychology might be needed to account for our evolution. See various attempts to construct a theory of evolution of sentient beings based on the MWI or its variants in Albert 1992, Chalmers 1996, Deutsch 1996, Donald 1990, Gell-Mann and Hartle 1990, Lehner 1997, Lockwood 1989, Page 2002, Penrose 1994, Saunders 1994, and Zeh 1981.

6.3 Derivation of the Probability Postulate from the Formalism of the MWI

Besides the question of the interpretation of the probability measure, which we have treated above, there is a separate issue about probabilities in the MWI, namely the claim that was sometimes made, e.g. by De

Witt 1970, that the probability postulate, i.e. the postulate that the probability measure is proportional to the measure of existence, can be derived from the formalism of the MWI. Several authors, e.g., Kent 1990, criticize the MWI on the grounds that this claim fails. As a matter of fact, the MWI has no advantage over other interpretations with regard to this issue. What is true instead is that one *can* derive the Probability Postulate from a weaker postulate according to which the probability is a function of the measure of existence. The derivation can be based on Gleason's 1957 theorem about the uniqueness of the probability measure. Similar results can be achieved by the analysis of the frequency operator originated by Hartle 1968 and from more general arguments by Deutsch 1999. All these results can be derived in the framework of various interpretations and thus the success or failure of these proofs cannot be an argument in favor or against the MWI. The MWI, like all other interpretations, requires a probability postulate.

Another idea for obtaining a probability law out of the formalism is to state, by analogy to the frequency interpretation of classical probability, that the probability of an outcome is proportional to the number of worlds with this outcome. This proposal immediately yields predictions that are different from what we observe in experiments. Some authors, arguing that counting is the only sensible way to introduce probability, consider this to be a fatal difficulty for the MWI, e.g., Ballentine 1975. Graham 1973 suggested that the counting of worlds *does* yield correct probabilities if one takes into account detailed splitting of the worlds in realistic experiments, but other authors have criticized the MWI because of the failure of Graham's claim. Weissman 1999 has proposed a modification of quantum theory with additional non-linear decoherence (and hence even more worlds than standard MWI), which can lead asymptotically to worlds of equal mean measure for different outcomes. Although this avoids random processes, like other MWI's, the price in the complication of the mathematical theory seems to be too high for the simplification in explaining probability. I believe that assigning equal probability to every world is unjustified. The formalism of quantum theory includes different amplitudes for quantum states corresponding to different worlds. It is a positive feature of the theory that the differences in the mathematical descriptions of worlds (different absolute values of amplitudes) are manifest in our experience. See Saunders 1998 for a detailed analysis of this issue.

From the weak probability postulate (the probability is a function of the measure of existence) follows that in case all the worlds in which a particular experiment took place have equal measures of existence, the probability of an outcome *is* proportional to the number of worlds with this outcome. If the measures of existence of these worlds are not equal, the experimenters in all the worlds can perform additional auxiliary measurements of some variables such that all the new worlds will have equal measures of existence. The experimenters should be completely indifferent to the results of these auxiliary measurements: their only purpose is to split the worlds into "equal-weight" worlds. This procedure reconstructs the standard quantum probability rule from the counting worlds approach; see Deutsch 1999 for details.

6.4 Social Behavior of a Believer in the MWI

There are claims that a believer in the MWI will behave in an irrational way. One claim is based on the naive argument described in the previous section: a believer who assigns equal probabilities to all different worlds will bet equal bets for the outcomes of quantum experiments that have unequal probabilities.

Another claim, recently discussed by Lewis 2000, is related to the strategy of a believer in the MWI who is offered to play a *quantum Russian roulette* game. The argument is that I, who would not accept an offer to play a classical Russian roulette, should agree to play the roulette any number of times if the triggering occurs according to the outcome of a quantum experiment. Indeed, at the end, there will be one world in which Lev is a multi-millionaire and all other worlds in which there will be no Lev Vaidman alive. Thus, in the future, Lev will be rich and presumably a happy man.

However, adopting the Probability Postulate leads all believers in the MWI to behave according to the following principle:

> **Behavior Principle**
> We care about all our successive worlds in proportion to their measures of existence.

With this principle our behavior will be similar to the behavior of a believer in the collapse theory who cares about possible future worlds according to the probability of their occurrence. I should not agree to play quantum Russian roulette because the measure of existence of worlds with Lev dead will be much larger than the measure of existence of the worlds with rich Lev alive.

7. Why the MWI?

The reason for adopting the MWI is that it avoids the collapse of the quantum wave. (Other non-collapse theories are not better than MWI for various reasons, e.g., nonlocality of Bohmian mechanics; and the disadvantage of all of them is that they have some additional structure.) The collapse postulate is a physical law that differs from all known physics in two aspects: it is genuinely random and it involves some kind of action at a distance. According to the collapse postulate the outcome of a quantum experiment is not determined by the initial conditions of the Universe prior to the experiment: only the probabilities are governed by the initial state. Moreover, Bell 1964 has shown that there cannot be a compatible local-variables theory that will make deterministic predictions. There is no experimental evidence in favor of collapse and against the MWI. We need not assume that Nature plays dice. The MWI is a deterministic theory for a physical Universe and it explains why a world appears to be indeterministic for human observers.

The MWI exhibits some kind of nonlocality: "world" is a nonlocal concept, but it avoids action at a distance and, therefore, it is not in conflict with the relativistic quantum mechanics; see discussions of nonlocality in Vaidman 1994, Tipler 2000, Bacciagaluppi 2002, and Hemmo and Pitowsky 2001. Although the issues of (non)locality are most transparent in the Schrödinger representation, an additional

insight can be gained through recent analysis in the framework of the Heisenberg representation, see Deutsch and Hayden 2000, Rubin 2001, and Deutsch 2001. The most celebrated example of nonlocality was given by Bell 1964 in the context of the Einstein-Podolsky-Rosen argument. However, in the framework of the MWI, Bell's argument cannot get off the ground because it requires a predetermined single outcome of a quantum experiment.

Another example of a kind of an action at a distance in a quantum theory with collapse is the *interaction-free measurement* of Elitzur and Vaidman 1993. Consider a super-sensitive bomb which explodes when *any* single particle arrives at its location. It seems that it is impossible to see this bomb, because any photon that arrives at the location of the bomb will cause an explosion. Nevertheless, using the Elitzur and Vaidman method, it is possible, at least sometimes, to find the location of the bomb without exploding it. In the case of success, a paradoxical situation arises: we obtain information about some region of space without any particle being there. Indeed, we know that no particle was in the region of the bomb because there was no explosion. The paradox disappears in the framework of the MWI. The situation is paradoxical because it contradicts physical intuition: the bomb causes an observable change in a remote region without sending or reflecting any particle. Physics is the theory of the Universe and therefore the paradox is real if this story is correct in the whole physical Universe. But it is not. There was no photon in the region of the bomb in a particular world, but there are other worlds in which a photon reaches the bomb and causes it to explode. Since the Universe incorporates all the worlds, it is not true that in the Universe no photon arrived at the location of the bomb. It is not surprising that our physical intuition leads to a paradox when we limit ourselves to a particular world: physical laws are applicable when applied to the physical universe that incorporates all of the worlds.

The MWI is not the most accepted interpretation of quantum theory among physicists, but it is becoming increasingly popular (see Tegmark 1998). The strongest proponents of the MWI can be found in the communities of quantum cosmology and quantum computing. In quantum cosmology it makes it possible to discuss the whole Universe avoiding the difficulty of the standard interpretation which requires an external observer. In quantum computing, the key issue is the parallel processing performed on the same computer; this is very similar to the basic picture of the MWI.[9]

Many physicists and philosophers believe that the most serious weakness of the MWI (and especially of its version presented here) is that it "gives up trying to explain things". In the words of Steane 1999, "It is no use to say that the [Schrödinger] cat is 'really' both alive and dead when every experimental test yields unambiguously the result that the cat is *either* alive *or* dead." (Steane dismisses the interference experiment which can reveal the presence of the superposition as unfeasible.) Indeed, if there is nothing in physics except the wave-function of the Universe, evolving according to the Schrödinger equation, then there are questions answering which requires help by other sciences. However, the advantage of the MWI is that it allows us to view quantum mechanics as a complete and consistent physical theory which agrees with all experimental results obtained to date.

Bibliography

- Aharonov, Y. and Vaidman, L. (1996) 'About Position Measurements Which Do Not Show the Bohmian Particle Position', in J. T. Cushing, A. Fine, and S. Goldstein (eds.), *Bohmian Mechanics and Quantum Theory: an Appraisal*, Netherlands: Kluwer Academic Publishers pp. 141-154. [Abstract | Preprint]
- Albert, D., (1992) *Quantum Mechanics and Experience*, Cambridge, MA: Harvard University Press.
- Albert, D., and Loewer, B. (1988) 'Interpreting the Many Worlds Interpretation', *Synthese* **77**, 195-213.
- Bacciagaluppi, G., (2002) 'Remarks on Space-Time and Locality in Everett's Interpretation', in *Modality, Probability, and Bell's Theorems* , (NATO Science Series). [Abstract | Preprint]
- Ballentine, L. E., (1975) *Measurements and Time Reversal in Objective Quantum Theory*, Oxford: Pergamon Press, pp. 50-51.
- Barnum, H., Caves, C.M., Finkelstein, J., Fuchs, C.A., and Schack, R., 1999, 'Quantum Probability from Decision Theory', *Proceedings of the Royal Society of London* (Series A), vol. 456/no. 1997 (8 May 2000), pp.1175-82. [Abstract | Preprint]
- Barrett, J. A., (1999) *The Quantum Mechanics of Minds and Worlds*, Oxford: University Press.
- Barvinsky, A. O., and Kamenshchik, A.Y., (1995) 'Preferred Basis in Quantum Theory and the Problem of Classicalization of the Quantum Universe' *Physical Review D* **52**, 743-757.
- Bell, J. S., (1990) 'Against Measurements', in A. I. Miller (ed.), *Sixty-Two Years of Uncertainty*, New York: Plenum Press, pp. 17-32.
- Bell, J. S., (1964) 'On the Einstein Podolsky Rosen Paradox', *Physics* **1**, 195-.
- Butterfield, J. N., (2001) 'Some Worlds of Quantum Theory', in R. Russell et al. (ed.), *Quantum Physics and Divine Action*, Vatican Observatory Publications. [Abstract | Preprint]
- Chalmers, D. J., (1996) *The Conscious Mind*, New York: Oxford University Press.
- Deutsch, D., (1986) 'Three experimental implications of the Everett interpretation', in R. Penrose and C.J. Isham (eds.), *Quantum Concepts of Space and Time*, Oxford: The Clarendon Press, pp. 204-214.
- Deutsch, D., (1996) *The Fabric of Reality*, New York: The Penguin Press.
- Deutsch, D., (1999) 'Quantum Theory of Probability and Decisions', *Proceedings of the Royal Society of London* **A 455**, 3129-3137. [Abstract | Preprint]
- Deutsch, D., and Hayden, P., (2000) 'Information Flow in Entangled Quantum Systems', *Proceedings of the Royal Society of London* **A 456**, 1759-1774. [Abstract | Preprint]
- De Witt, B. S. M., (1970) 'Quantum mechanics and Reality', *Physics Today* **23**, No. 9, pp. 30-35.
- Donald, M. J., (1990) 'Quantum Theory and the Brain', *Philosophical Transactions of the Royal Society of London*, **A 427** (1872), 43-93. [Abstract | Preprint]
- Elitzur, A., and Vaidman, L., (1993) 'Interaction-Free Quantum Measurements', *Foundation of Physics* **23**, 987-997. [Abstract | Preprint]
- Englert, B., Scully, M. O., Süssmann, G. and Walther, H., (1992) 'Surrealistic Bohm trajectories', *Zeitschrifft für Naturforschung* **47a**, 1175-1186.
- Everett, H., (1957) '*Relative State* Formulation of quantum mechanics', *Review of Modern Physics* **29**, pp. 454-462; see also 'The Theory of the Universal Wave Function', in B. De Witt and N. Graham (eds.), *The Many-Worlds Interpretation of Quantum Mechanics*, Princeton NJ:

Princeton University Press, 1973.

- Gell-Mann, M., and Hartle, J. B., (1990) ‘Quantum Mechanics in the Light of Quantum Cosmology’, in W. H. Zurek (ed.), *Complexity, Entropy and the Physics of Information*, Reading: Addison-Wesley, pp. 425-459.
- Gleason, A. M., (1957) ‘Measures on the Closed Subspaces of Hilbert Space’, *Journal of Mathematics and Mechanics* **6**, 885-894.
- Graham, N., (1973) ‘The Measurement of Relative Frequency’, in De Witt and N. Graham (eds.) *The Many-Words Interpretation of Quantum Mechanics*, Princeton NJ: Princeton University Press.
- Hartle, J. B., (1968) ‘Quantum Mechanics of Individual Systems’, *American Journal of Physics* **36**, 704-712.
- Hemmo, M., and Pitowsky, I., (2001) ‘Probability and Nonlocality in Many Minds Interpretations of Quantum Mechanics’, forthcoming, *British Journal for the Philosophy of Science* [Abstract | Preprint]
- Kent, A., (1990) ‘Against Many-Worlds Interpretation’, *International Journal of Modern Physics* **A5**, 1745-1762. [Abstract | Preprint]
- Lehner, C., (1997) ‘What it Feels Like to Be in a Superposition. And Why -- Consciousness and the Interpretation of Everett’s Quantum Mechanics’, *Synthese* **110**, 191-216.
- Lewis, D., (1986) *On the Plurality of Worlds*, Oxford, New York: Basil Blackwell.
- Lewis, P., (2000) ‘What is it like to be Schrödinger’s cat?’ *Analysis* **60**, 22-29.
- Lockwood, M., (1989) *Mind, Brain & the Quantum*, Oxford: Basil Blackwell.
- Lockwood, M., Brown, H. R., Butterfield, J., Deutsch, D., Loewer, B., Papineau, D., Saunders, S. (1996) ‘Symposium: The ‘Many Minds’ Interpretation of Quantum Theory’, *British Journal for the Philosophy of Science* **47**, 159-248.
- Page, D., (2002) ‘Mindless Sensationalism: a Quantum Framework for Consciousness’, in *Consciousness: New Philosophical Essays*, Q. Smith and A. Jokic (eds.), Oxford: Oxford University Press. [Abstract | Preprint]
- Parfit, D., (1986) *Reasons and Persons*, New York: Oxford University Press.
- Plaga, R., (1997) ‘Proposal for an experimental test of the many-worlds interpretation of quantum mechanics’, *Foundations of Physics* **27**, 559-577. [Abstract | Preprint]
- Pearle, P., (1986) ‘Models for Reduction’, in R. Penrose and C.J. Isham (eds.), *Quantum Concepts of Space and Time*, Oxford: Caledonia Press, pp. 204-214.
- Penrose, R., (1994) *Shadows of the Mind*, Oxford: Oxford University Press.
- Rubin, M., (2001) ‘Locality in the Everett Interpretation of Heisenberg-Picture Quantum Mechanics’, *Foundations of Physics Letters*, 14, 301-322 [Abstract | Preprint]
- Saunders, S., (1994) ‘Decoherence and Evolutionary Adaptation’, *Physics Letters* **A 184**, 1-5.
- Saunders, S., (1995) ‘Time, Quantum Mechanics, and Decoherence’, *Synthese* **102**, 235-266. [Abstract | Preprint]
- Saunders, S., (1998) ‘Time, Quantum Mechanics, and Probability’, *Synthese* **114**, 373-404. [Abstract | Preprint]
- Skyrms, B., (1976) ‘Possible Worlds, Physics and Metaphysics’, *Philosophical Studies* **30**, 323-332.

- Tappenden, P., (2000) ‘Identity and Probability in Everett’s Multiverse’, *British Journal for the Philosophy of Science* **51**, 99-114.
- Tegmark, M., (1998) ‘The Interpretation of Quantum Mechanics: Many Worlds or Many Words?’, *Fortschritte der Physik* **46**, 855-862. [Abstract | Preprint (in Postscript)]
- Tipler, D., (1986) ‘The Many-Worlds Interpretation of Quantum Mechanics in Quantum Cosmology’, in R. Penrose and C.J. Isham (eds.), *Quantum Concepts of Space and Time*, Oxford: The Clarendon Press, 1986, pp. 204-214.
- Vaidman, L., (1994) ‘On the paradoxical aspects of new quantum experiments’, *Philosophy of Science Association 1994*, pp. 211-217. [Abstract | Preprint]
- Vaidman, L., (1998) ‘On Schizophrenic Experiences of the Neutron or Why We should Believe in the Many-Worlds Interpretation of Quantum Theory’, *International Studies in the Philosophy of Science 12*, 245-261. [Abstract | Preprint]
- Wallace, D., (2001a) ‘Everett and Structure’, forthcoming, *Studies in the History and Philosophy of Modern Physics* [Abstract | Preprint]
- Wallace, D. (2001b) ‘Worlds in the Everett Interpretation’ *Studies in the History and Philosophy of Modern Physics* (to appear). [Abstract | Preprint]
- Weissman, M. B., (1999) ‘Emergent Measure-Dependent Probabilities from Modified Quantum Dynamics without State-Vector Reduction’, *Foundations of Physics Letters* **12**, 407-426. [Abstract | Preprint]
- Zeh, H. D., (1981) ‘The Problem of Conscious Observation in Quantum Mechanical Description’, *Epistemological Letters*, No. 63. [Abstract | Preprint]
- Zurek, W. H., (1998) ‘Decoherence, Einselection and the Existential Interpretation (the Rough Guide)’, *Philosophical Transactions of the Royal Society of London*, **A 356** (1743), 1793-1821. [Abstract | Preprint]

Other Internet Resources

Preprints

- Deutsch, D., (2001) ‘The Structure of the Multiverse’. [Abstract | Preprint]
- Page, D., (2000) ‘Can Quantum Cosmology Give Observational Consequences of Many-Worlds Quantum Theory?’. [Abstract | Preprint]
- Steane, A. M., (1999) ‘A quantum computer only needs one universe’. [Abstract | Preprint]
- Tipler, D., (2000) ‘Does Quantum Nonlocality Exist? Bell’s Theorem and the Many-Worlds Interpretation’. [Abstract | Preprint]

Other Resources

- Search Results at arXiv.org Preprint Archive (This is a search on the Boolean string "many+worlds or Everett".)

- Search Results at the Philosophy of Science Archives (U. Pittsburgh)
- The Everett FAQ (maintained by Michael Price)

Related Entries

quantum mechanics | quantum mechanics: Everett's relative-state formulation of | quantum theory: measurement in

Acknowledgements

I am thankful to everybody who has borne with me through endless discussions of the MWI (in this and other worlds) and, in particular, to Yakir Aharonov, David Albert, Guido Bacciagalupi, Jeremy Butterfield, Rob Clifton, David Deutsch, Simon Saunders, Philip Pearle, and David Wallace. I acknowledge partial support by grant 62/01 of the Israel Science Foundation and the EPSRC grant GR/N33058.

Lev Vaidman
vaidman@post.tau.ac.il

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 24, 2002
Content last modified: March 24, 2002

Everett's Relative-State Formulation of Quantum Mechanics

Everett's relative-state formulation of quantum mechanics is an attempt to solve the measurement problem by dropping the collapse dynamics from the standard von Neumann-Dirac theory of quantum mechanics. The main problem with Everett's theory is that it is not at all clear how it is supposed to work. In particular, while it is clear that he wanted to explain why we get determinate measurement results in the context of his theory, it is unclear how he intended to do this. There have been many attempts to reconstruct Everett's no-collapse theory in order to account for the apparent determinateness of measurement outcomes. These attempts have led to such formulations of quantum mechanics as the many-worlds, many-minds, many-histories, and relative-fact theories. Each of these captures part of what Everett claimed for his theory, but each also encounters problems.

Introduction

Everett formulated his relative-state interpretation of quantum mechanics while he was a graduate student in physics at Princeton University. His doctoral dissertation (1957a) was recommended for publication March 1957 and a paper reporting the results of his dissertation (1957b) was published in July of the

same year. He also published an extended discussion of his relative-state interpretation in the DeWitt and Graham anthology (1973). After graduating from Princeton, Everett worked as a defense analyst. He died in 1982.

Everett's no-collapse formulation of quantum mechanics was a reaction to problems that arise in the standard von Neumann-Dirac collapse theory. Everett's proposal was to drop the collapse postulate from the standard theory, then deduce the empirical predictions of the standard theory as the subjective experiences of observers who are themselves treated as physical systems described by his theory. It is, however, unclear precisely how Everett intended for this to work. Consequently, there have been many, mutually incompatible, attempts at trying to explain what he in fact had in mind. Indeed, it is probably fair to say that most no-collapse interpretations of quantum mechanics have at one time or another been attributed to Everett.

In what follows, I will describe Everett's worry about the standard collapse formulation of quantum mechanics and his proposal for solving the problem as it is presented in his 1957 paper. I will then very briefly describe a few approaches to interpreting Everett's theory.

The Measurement Problem

Everett presented his relative-state formulation of quantum mechanics as a way of avoiding the problems encountered by the standard von Neumann-Dirac theory. The main problem, according to Everett, was that the standard theory required observers always to be treated as *external* to the system described by the theory. One unfortunate consequence of this was that the standard theory could not be used to describe the universe as a whole since the universe is a system containing observers.

In order to understand what Everett was worried about, one must first understand how the standard theory works. The standard von Neumann-Dirac theory is based on the following principles (see von Neumann 1955):

> **1. Representation of States**: The possible physical states of a system *S* are represented by the unit-length vectors in a Hilbert space (which for present purposes one may regard as a vector space with an inner product). The physical state at a time is then represented by a single vector in the Hilbert space.
>
> **2. Representation of Properties**: For each physical property *P* that one might observe of a system *S* there is an operator ***P*** (on the vectors that represent the possible states of *S*) that represents the property (in a way determined by the following principle).
>
> **3. Eigenvalue-Eigenstate Link**: A system *S* determinately has physical property *P* if and only if ***P*** operating on ***S*** (the vector representing *S*'s state) yields ***S***. We say then that *S* is in an eigenstate of ***P*** with eigenvalue 1. *S* determinately does not have property *P* if and only

> if ***P*** operating on ***S*** yields 0.
>
> **4. Dynamics**: **(a)**. If no measurement is made, then a system evolves continuously according to the linear, deterministic Schroedinger dynamics, which depends only on the energy properties of the system. **(b)**. If a measurement is made, then the system instantaneously and randomly jumps to a state where it either determinately has or determinately does not have the property being measured (according to the eigenvalue-eigenstate link). The probability of each possible post-measurement state is determined by the system's initial state.

It is worth noting that according to the eigenvalue-eigenstate link (3) a system would typically neither determinately have nor determinately not have a particular given property (a specific property might be represented by a line in the state space: in order to determinately have the property the state of a system must be parallel to the line, and in order to determinately not have the property the state of a system must be orthogonal to the line, and most state vectors are neither parallel nor orthogonal to any given line). Further, the deterministic dynamics (4a), as we will see below, does nothing to guarantee that a system would either determinately have or determinately not have a particular property when one observes the system to see whether the system has that property. This is why the collapse dynamics (4b) is needed: in the standard theory it is what guarantees that a system will either determinately have or determinately not have a property whenever one observes the system to see whether or not it has the property. This is why the standard theory has two dynamical laws: a random, discontinuous one (4b) that describes what happens when a measurement is made and a deterministic, continuous one (4a) that describes what happens the rest of the time.

But what does it take for an interaction to count as a *measurement*? Unless we know this, the standard theory is at best incomplete, since we do not know when each dynamical law obtains. Moreover, and this is what worried Everett, if we suppose that observers and their measuring devices are constructed from simpler systems that each obey the deterministic dynamics, then the composite systems, the observers and their measuring devices, must evolve in a continuous deterministic way, and nothing like the random, discontinuous evolution described by rule 4b can ever occur. That is, if observers and their measuring devices are understood as being constructed of simpler systems each behaving as quantum mechanics requires (each obeying 4a), then the standard theory is logically inconsistent since it says that together the systems must obey 4b. In order to preserve the consistency of the theory, Everett concluded that the standard theory cannot be used to describe systems that contain observers--it can only be used to describe systems where all observers are *external* to the described system. This restriction on the applicability of the theory was unacceptable. Everett wanted a theory that could be applied to any physical system whatsoever, one that described observers and their measuring devices the same way that it described every other physical system.

Everett's Proposal

In order to solve the standard theory's measurement problem Everett proposed dropping the collapse

dynamics (4b) and taking the resulting theory as providing a complete and accurate description of all physical systems whatsoever. He then intended to *deduce* the statistical predictions of quantum mechanics (the predictions that depend on rule 4b in the standard theory) as the subjective experiences of observers who are themselves treated as ordinary physical systems within the new theory.

This is what Everett says: "We shall be able to introduce into [the relative-state theory] systems which represent observers. Such systems can be conceived as automatically functioning machines (servomechanisms) possessing recording devices (memory) and which are capable of responding to their environment. The behavior of these observers shall always be treated within the framework of wave mechanics. Furthermore, we shall deduce the probabilistic assertions of Process 1 [rule 4b] as *subjective* appearances to such observers, thus placing the theory in correspondence with experience. We are then led to the novel situation in which the formal theory is objectively continuous and causal, while subjectively discontinuous and probabilistic. While this point of view thus shall ultimately justify our use of the statistical assertions of the orthodox view, it enables us to do so in a logically consistent manner, allowing for the existence of other observers." (1973, 9)

Everett's goal then was to show that the memory records of an observer as described by his no-collapse theory would match those predicted by the standard theory with the collapse dynamics. The main problem in understanding what Everett had in mind is in figuring out how this correspondence between the predictions of the two theories was supposed to work.

Suppose that *J* is a good observer who measures the *x*-spin of a spin-1/2 system *S* (spin is a property of fundamental particles and other quantum-mechanical systems; if one specifies a particular axis (like *x* or *z*), then a spin-1/2 system will be found, on measurement, to be either "spin up" or "spin down" with respect to the particular axis). For Everett, being a good *x*-spin observer means that *J* has the following two dispositions (the arrows below represent the time-evolution described by the deterministic dynamics 4a):

1. $|\text{“ready”}\rangle_J\,|\text{x-spin up}\rangle_S \rightarrow |\text{“spin up”}\rangle_J\,|\text{x-spin up}\rangle_S$

2. $|\text{“ready”}\rangle_J\,|\text{x-spin down}\rangle_S \rightarrow |\text{“spin down”}\rangle_J\,|\text{x-spin down}_S\rangle$

J measures a system that is determinately *x*-spin up, then *J* will determinately record "spin up"; and if *J* measures a system that is determinately *x*-spin down, then *J* will determinately record "spin down" (and we assume for simplicity that the object system is not undisturbed).

Now consider what happens when *J* observes the *x*-spin of a system that begins in a *superposition* of *x*-spin eigenstates:

$$a|\text{x-spin up}\rangle_S + b|\text{x-spin down}\rangle_S$$

The initial state of the composite system then is:

$$|\text{“ready”}\rangle_J \, (a|\text{x-spin up}\rangle_S + b|\text{x-spin down}\rangle_S)$$

Given *J*'s two dispositions and the fact that the deterministic dynamics is linear, the state of the composite system after *J*'s *x*-spin measurement will be:

$$a|\text{“spin up”}\rangle_J \, |\text{x-spin up}\rangle_S + b|\text{“spin down”}\rangle_J \, |\text{x-spin down}\rangle_S$$

Let's call this state ***E*** for Everett. Note that on the standard eigenvalue-eigenstate link ***E*** is not a state where *J* determinately records "spin up" neither is it a state where *J* determinately records "spin down". So in what sense is *J*'s record supposed to agree with the empirical prediction made by the standard theory (which predicts that *J* will either end up with the determinate record "spin up" or the determinate record "spin down", with probabilities equal to *a*-squared and *b*-squared respectively)?

Everett confesses that such a post-measurement state is puzzling: "As a result of the interaction the state of the measuring apparatus is no longer capable of independent definition. It can be defined only *relative* to the state of the object system. In other words, there exists only a correlation between the states of the two systems. It seems as if nothing can ever be settled by such a measurement." (1957b, 318)

And he describes the problem he faces: "This indefinite behavior seems to be quite at variance with our observations, since physical objects always appear to us to have definite positions. Can we reconcile this feature of wave mechanical theory built purely on [rule 4a] with experience, or must the theory be abandoned as untenable? In order to answer this question we consider the problem of observation itself within the framework of the theory." (1957b, 318)

Then he describes his solution to the determinate-experience problem: "Let one regard an observer as a subsystem of the composite system: observer + object-system. It is then an inescapable consequence that after the interaction has taken place there will not, generally, exist a single observer state. There will, however, be a superposition of the composite system states, each element of which contains a definite observer state and a definite relative object-system state. Furthermore, as we shall see, *each* of these relative object system states will be, approximately, the eigenstates of the observation corresponding to the value obtained by the observer which is described by the same element of the superposition. Thus, each element of the resulting superposition describes an observer who perceived a definite and generally different result, and to whom it appears that the object-system state has been transformed into the corresponding eigenstate. In this sense the usual assertions of [the collapse dynamics (4b)] appear to hold on a subjective level to each observer described by an element of the superposition. We shall also see that correlation plays an important role in preserving consistency when several observers are present and allowed to interact with one another (to ‘consult’ one another) as well as with other object-systems." (1973, 10)

Everett presents a principle of the fundamental relativity of states: one can say that in state ***E***, *J* recorded "*x*-spin up" relative to *S* being in the *x*-spin up state and that *J* recorded "*x*-spin down" relative to *S* being in the *x*-spin down state. But note that this principle alone does not allow Everett to deduce anything like the experiences predicted by the standard collapse theory. The standard theory predicts that on measurement the quantum-mechanical state of the composite system will collapse to either:

$$|\text{“spin up”}\rangle_J\,|\text{x-spin up}\rangle_S \text{ or } |\text{“spin down”}\rangle_J\,|\text{x-spin down}\rangle_S$$

when a measurement is made, and that there is thus a single, simple matter of fact about which measurement result *J* recorded. On Everett's account it is unclear whether *J* ends up recording one result or the other or somehow both or perhaps neither.

The problem is that there is a gap in Everett's exposition between what he sets out to explain (why observers have precisely the same experiences as predicted by the standard theory) and what he ultimately ends up saying. Since it is unclear exactly how he intends for his theory to explain an observer's (apparently?) determinate measurement records, it is also unclear how he intends to explain why one should expect one's determinate measurement records to exhibit the standard quantum statistics. This gap in Everett's exposition has led to the many mutually incompatible reconstructions of his theory--each can be taken as presenting a different way of explaining how one's records can be determinate (or at least *appear* to be determinate) in a post-measurement state like ***E***.

The Bare Theory

Albert and Loewer's bare theory (Albert and Loewer 1988 and Albert 1992) is arguably the wildest interpretation of Everett's theory around. On this reading, one supposes that Everett intended to drop the collapse dynamics but to keep the standard eigenvalue-eigenstate link.

So how does the bare theory account for *J*'s determinate experience? The short answer is that it doesn't. Rather, on the bare theory, one tries to explain why *J* would *falsely* believe that he has an ordinary determinate measurement record. The trick is to ask the observer not what result he got, but rather whether he got *some* specific determinate result. If the post-measurement state was:

$$|\text{“spin up”}\rangle_J\,|\text{x-spin up}\rangle_S$$

then *J* would report "I got a determinate result, either spin up or spin down." And he would make precisely the same report if he ended up in the post-measurement state:

$$|\text{“spin down”}\rangle_J\,|\text{x-spin down}\rangle_S$$

So, by the linearity of the dynamics, *J* would *falsely* report "I got a determinate result, either spin up or

spin down" when in the state ***E***:

$$a|\text{"spin up"}\rangle_J \; |\text{x-spin up}\rangle_S + b|\text{"spin down"}\rangle_J \; |\text{x-spin down}\rangle_S$$

Thus, one might argue, it would *seem* to *J* that he got a perfectly determinate ordinary measurement result even when he did not (that is, he did not determinately get "spin up" and did not determinately get "spin down").

The idea is to try to account for all of *J*'s beliefs about his determinate experiences by appealing to such *illusions*. Rather than predicting the experiences that we believe that we have, a proponent of the bare theory tells us that we do not have many determinate beliefs at all and then tries to explain why we nonetheless determinately believe that we do.

While one can tell several suggestive stories about the sort of illusions that an observer would experience, there are at least two serious problems with the bare theory. One problem is that the bare theory is not empirically coherent: that is, if the bare theory were true, it would be impossible to ever have empirical evidence for accepting it as true. Another is that if the bare theory were true, one would most likely fail to have any determinate beliefs at all (since on the deterministic dynamics one would almost never expect that the global state was an eigenstate of any particular observer being sentient), which is presumably not the sort of prediction one looks for in a successful physical theory (for more details on how experience is supposed to work in the bare theory and some the problems it encounters see Bub, Clifton, and Monton 1998 and Barrett 1994 and 1998).

Many Worlds

DeWitt's many-worlds interpretation is easily the most popular reading of Everett. On this theory there is one world corresponding to each term in the expansion of ***E*** when written in the preferred basis (there are always many ways one might write the quantum-mechanical state of a system as the sum of vectors in the Hilbert space; in choosing a preferred basis, one chooses a single set of vectors that can be used to represent a state and thus one chooses a single *preferred* way of representing a state as the sum of vectors in the Hilbert space). The theory's preferred basis is chosen so that each term in the expansion of ***E*** describes a world where there is a determinate measurement record. Given the preferred basis (surreptitiously) chosen above, ***E*** describes two worlds: one where *J* (or perhaps better *J1*) determinately records the measurement result "spin up" and another where *J* (or *J2*) determinately records "spin down".

DeWitt and Graham describe their reading of Everett as follows: "[Everett's interpretation of quantum mechanics] denies the existence of a separate classical realm and asserts that it makes sense to talk about a state vector for the whole universe. This state vector never collapses and hence reality as a whole is rigorously deterministic. This reality, which is described *jointly* by the dynamical variables and the state vector, is not the reality we customarily think of, but is a reality composed of many worlds. By virtue of the temporal development of the dynamical variables the state vector decomposes naturally into

orthogonal vectors, reflecting a continual splitting of the universe into a multitude of mutually unobservable but equally real worlds, in each of which every good measurement has yielded a definite result and in most of which the familiar statistical quantum laws hold." (1973, v)

DeWitt admits that this constant splitting of worlds whenever the states of systems become correlated is counterintuitive: "I still recall vividly the shock I experienced on first encountering this multiworld concept. The idea of 10^{100} slightly imperfect copies of oneself all constantly spitting into further copies, which ultimately become unrecognizable, is not easy to reconcile with common sense. Here is schizophrenia with a vengeance." (1973, 161)

But while the theory is counterintuitive, it does (unlike the bare theory) explain why observers end up recording determinate measurement results. In the state described by ***E*** there are two observers each occupying a different world and each with a perfectly determinate measurement record. There are, however, other problems with the many-worlds theory.

A standard complaint is that the theory is ontologically extravagant. One would presumably only ever need one physical world, *our* world, to account for *our* experience. The idea behind postulating the actual existence of a different physical world corresponding to each term in the quantum-mechanical state is that is allows one to explain our determinate experiences while taking the deterministically-evolving quantum-mechanical state to be in some sense a complete and accurate description of the physical facts. But again one might wonder whether the sort of completeness one gets warrants the vast ontology of worlds.

Perhaps more serious, in order to explain our determinate measurement records, the theory requires one to choose a preferred basis so that observers have determinate records (or better, determinate *experiences*) in each term of the quantum-mechanical state as expressed in this basis. The problem is that not just any old preferred basis will do this--indeed, we presumably do not know what basis would make our experiences and beliefs determinate in every world. Indeed, the right preferred basis would presumably depend on such things as our physiology and experimental practice, so even if we did know which one to choose, this choice of a fundamental part of our most basic physical theory (the part that tells us when worlds split) would have to be contingent on accidents of biology and practice. We tend to think that our physical laws are independent of such accidental features of the world.

Another problem concerns the statistical predictions of the theory. The standard collapse theory predicts that *J* will get the result "spin up" with probability *a*-squared and "spin down" with probability *b*-squared in the above experiment. But the many-worlds theory cannot, as it stands, make any statistical predictions concerning an observer's future experiences. Indeed, the question "What is the probability that *J* will record the result ‘spin up’?" is strictly nonsense since one cannot identify which of the two future observers is *J*. The upshot is that one cannot capture the standard probabilistic predictions of quantum mechanics. And the moral is that if one does not have transtemporal identity of observers in a theory, then one cannot assign probabilities to their future experiences.

Many Minds

Everett said that on his formulation of quantum mechanics "the formal theory is objectively continuous and causal, while subjectively discontinuous and probabilistic" (1973, 9). Albert and Loewer (1988) have captured this feature in their many-minds theory by distinguishing between the time-evolution of an observer's physical state, which is continuous and causal, and the evolution of an observer's mental state, which is discontinuous and probabilistic.

Perhaps the oddest thing about this theory is that in order to get the observer's mental state in some way to supervene on his physical state, Albert and Loewer associate with each observer a continuous infinity of minds. The physical state always evolves in the usual deterministic way, but each mind evolves randomly (with probabilities determined by the particular mind's current mental state and the evolution of the global quantum-mechanical state). On the mental dynamics that they describe, one would expect *a*-squared of *J*'s minds to end up associated with the result "spin up" (the first term of ***E***) and *b*-squared of *J*'s minds to end up associated with the result "spin down" (the second term of ***E***). The mental dynamics is also stipulated to be memory preserving.

An advantage of this theory over the many-worlds theory is that there is no *physically* preferred basis. To be sure, one must choose a preferred basis in order to specify the mental dynamics completely (something that Albert and Loewer never completely specify), but as Albert and Loewer point out, this choice has absolutely nothing to do with any physical facts; rather, it can be thought of as part of the description of the relationship between physical and mental states. Another advantage of the many-minds theory is that, unlike the many-worlds theory, it really does make the usual probabilistic predictions for the future experiences of a particular mind (this, of course, requires that one take the minds to have transtemporal identities, which Albert and Loewer do as part of their unabashed commitment to a strong mind-body dualism).

The main problems with the many-minds theory concern its commitment to a strong mind-body dualism and the question of whether the sort of mental supervenience one gets is worth the trouble of postulating a continuous infinity of minds associated with each observer. Concerning the latter, one might well conclude that a *single*-mind theory would be preferable (see Barrett 1995).

Many Histories

Gell-Mann and Hartle (1990) understand Everett's theory as one that describes many, mutually decohering histories. The main difference between this approach and the many-worlds interpretation is that, instead of stipulating a preferred basis, here one relies on the physical interactions between a physical system and its environment (the way in which the quantum-mechanical states become correlated) to effectively choose what physical quantity is determinate at each time for each system.

One problem concerns whether and in what sense environmental interactions can select *a physically*

preferred basis for the entire universe, which is what we presumably need in order to make sense of Everett's formulation. After all, in order to be involved in environmental interactions a system must have an environment, and the universe, by definition, has no environment. Another problem is that it is unclear that the environment-selected determinate quantity at a time is a quantity that would explain *our* determinate measurement records and experience. Proponents who argue for this approach often appeal to biological or evolutionary arguments to justify the assumption that sentient beings must record their beliefs in terms of the environment-selected (or decohering) physical properties (see Gell-Mann and Hartle 1990 and Zurek 1991 for this sort of argument). The short story is that it is not yet clear how the account of our determinate experience is suppose to work when one relies on decoherence to select a preferred basis (see Dowker and Kent 1996 for an extended discussion of some of the problems one encounters in such an approach).

If one allows oneself the luxury of stipulating a preferred basis (a basis where every observer's measurement records are determinate), one can construct a many-histories theory from Albert and Loewer's many-minds theory. One might, for example, take the trajectory of each of an observer's minds to describe the history of a possible physical world. One might then stipulate a measure over the set of all possible histories (trajectories) that would represent the prior probability of each history being *ours*. Note that since such worlds (and everything in them) would have transtemporal identities, unlike the many-worlds theory, there would be no special problem here in talking about probabilities concerning one's future experience--the quantum probabilities in such a theory might naturally be interpreted as *epistemic* probabilities.

Relative Facts

Perhaps the approach closest to the spirit of Everett's relative-state formulation would be simply to deny that there are typically any absolute matters of fact about the properties of physical systems or the records, experiences, and beliefs of observers (see Saunders 1995 and Mermin 1997). In the experiment above, *J* does not end up believing that his result was "spin up", and he does not end up believing that his result was "spin down"; rather, on this sort of theory all facts would be relative. Not relative to a particular world, mind, or history, but relative to each other: here *J* believes that his result was "spin up" *relative to S* being *x*-spin up and *J* believes that his result was "spin down" *relative to S* being *x*-spin down. But, one might ask, what is the state of *S* then? Well, *S* is *x*-spin up *relative to J* believing that his result was "spin up", etc. Again, on this sort of theory there are typically no absolute matters of fact about the properties of individual physical systems.

So how do we account for our determinate experience? On this approach, one simply denies that there is any simple matter of fact concerning what an observer's experience is. Which means, of course, that insofar as one believes that there really is a simple matter of fact about what one's experiences is, the relative-fact theory provides no account of one's experience. Similarly, one cannot make sense of the usual statistical predictions of quantum mechanics insofar as one takes these to be predictions concerning the probability that a particular measurement outcome will in fact occur. Again, there are typically no such simple facts in such a theory.

It has been argued that since there is no simple matter of fact concerning whether a particular event does or does not occur, quantum mechanics (in fact?) concerns only the probabilistic correlations between events. It seems to me, however, that any coherent talk about the probabilistic correlations between events presupposes that there are determinate matters of fact concerning what events occur (otherwise what are the probabilities probabilities *of*?).

Summary

Such are some of the ways of understanding Everett's relative-state formulation of quantum mechanics. It will probably never be entirely clear precisely what Everett himself had in mind, but his goal of trying to make sense of quantum mechanics without the collapse postulate was heroic. And even in light of the problems one faces, puzzling over how one might reconstruct Everett's theory continues to hold promise.

Bibliography

- Albert, D. Z.: 1992, *Quantum Mechanics and Experience*, Harvard University Press, Cambridge, MA.
- Albert, D. Z. and J. A. Barrett: 1995 "On What It Takes To Be a World," *Topoi* 14: 35-37.
- Albert, D. and B. Loewer: 1988, "Interpreting the Many Worlds Interpretation," *Synthese* 77: 195-213.
- Barrett, J.: 1998, *The Quantum Mechanics of Minds and Worlds*, Oxford University Press, Oxford.
- 1998, "On the Nature of Experience in the Bare Theory", forthcoming in Dieks and Vermaas (eds).
- 1996, "Empirical Adequacy and the Availability of Reliable Records in Quantum Mechanics," *Philosophy of Science* 63: 49-64.
- 1995, "The Single-Mind and Many-Minds Formulations of Quantum Mechanics," *Erkenntnis* 42: 89-105.
- 1994, "The Suggestive Properties of Quantum Mechanics Without the Collapse Postulate," *Erkenntnis* 41: 233-252.
- Bell, J. S.: 1987, *Speakable and Unspeakable in Quantum Theory*, Cambridge University Press, Cambridge.
- Bub, Clifton, and Monton: 1998, "The Bare Theory Has No Clothes", *Minnesota Studies in the Philosophy of Science*, forthcoming.
- Butterfield, J.: 1995, "Worlds, Minds, and Quanta," invited paper for the Aristotelian Society and Mind Association Joint Session, Liverpool, July 1995.
- Clifton, R.: 1996, "On What Being a World Takes Away," *Philosophy of Science* 63: S151-S158.
- DeWitt, B. S.: 1971, "The Many-Universes Interpretation of Quantum Mechanics," in *Foundations of Quantum Mechanics*, Academic Press, New York; reprinted in DeWitt and Graham (eds), 167-218.

- DeWitt, B. S. and N. Graham (eds): 1973, *The Many-Worlds Interpretation of Quantum Mechanics*, Princeton University Press, Princeton.
- Dowker, F. A. and Kent: 1996, "On the Consistent Histories Approach to Quantum Mechanics," *Journal of Statistical Physics*, vol. 83, nos. 5-6, 1575-1646.
- Everett, H: 1957a, *On the Foundations of Quantum Mechanics*, thesis submitted to Princeton University, March, 1, 1957, in partial fulfillment of the requirements for the Ph.D. degree.
- Everett, H: 1957b, "'Relative State' Formulation of Quantum Mechanics," *Reviews of Modern Physics*, 29: 454-462.
- Everett, H: 1973, "The Theory of the Universal Wave Function," in DeWitt and Graham (eds).
- Farhi, E., J. Goldstone, S. Gutmann: 1989, "How Probability Arises in Quantum Mechanics," *Annals of Physics*, 192: 368-382.
- Gell-Mann, M. and J. B. Hartle: 1990, "Quantum Mechanics in the Light of Quantum Cosmology," in *Complexity, Entropy, and the Physics of Information*, Proceedings of the Santa Fe Institute Studies in the Sciences of Complexity, vol. VIII, W. H. Zurek (ed), Addison-Wesley, Redwood City, CA, 425-58.
- Geroch, R.: 1984, "The Everett Interpretation," *Nous* 18: 617-33.
- Healey, R.: 1984, "How Many Worlds?", *Nous* 18: 591-616.
- Hemmo, M.: 1996, *Quantum Mechanics Without Collapse: Modal Interpretations, Histories, and Many Worlds*, thesis submitted for the degree of Doctor of Philosophy University of Cambridge.
- Lockwood, M.: 1996, "Many Minds Interpretations of Quantum Mechanics," *British Journal for the Philosophy of Science*, v47, n2: 159-188.
- 1989, *Mind, Brain, and the Quantum*, Blackwell, Oxford.
- Mermin: 1997, "The Ithaca Interpretation of Quantum Mechanics", LANL Archives at xxx.lanl.gov.
- Saunders, S.: 1995, "Time, Quantum Mechanics, and Decoherence," *Synthese*, v102, n2: 235-266.
- Stein, H.: 1984, "The Everett Interpretation of Quantum Mechanics: Many Worlds or None?", *Nous* 18: 635-52.
- von Neumann, J.: 1955, *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton; translated by R. Beyer from *Mathematische Grundlagen der Quantenmechanik*, Springer, Berlin, 1932.
- Wheeler, J. A. and W. H. Zurek (eds): 1983, *Quantum Theory and Measurement*, Princeton University Press, Princeton.
- Zurek, W. H.: 1991, "Decoherence and the Transition from Quantum to Classical," *Physics Today*, October 1991, 36-44.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

quantum mechanics

Jeffrey A. Barrett
jabarret@uci.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 3, 1998
Content last modified: June 3, 1998

Stanford Encyclopedia of Philosophy
Notes to The Many-Worlds Interpretation of Quantum Mechanics

Notes

1. The mathematical part of the MWI, (i), yields less than mathematical parts of some other theories such as, e.g., Bohmian mechanics. Indeed, our experience is consistent with the MWI, but it does not follow from its mathematical part. The Schrödinger equation itself does not explain why we experience definite results in quantum measurements. In contrast, in the Bohmian mechanics the mathematical part yields almost everything, and the analog of (ii) is very simple: it is the postulate according to which only the "Bohmian positions" (and not the quantum wave) correspond to our experience. The Bohmian positions of all particles yield the familiar picture of the (single) world we are aware of. Thus, philosophically, a theory like the Bohmian mechanics achieves more than the MWI, but at the price of a significant impairment of the physical aspects of the theory, e.g., addition of the non-local dynamics of Bohmian particle positions. However, Wallace (2001a) argues that stripping the experiential content from empty waves in the Bohmian approach has significant philosophical difficulties too.

2. Wallace 2001a points out that the term “superposition of a cat” is a misnomer. I use it as a shortcut for “a superposition of states of elementary particles corresponding to different (classical) states of the cat”.

3. It corresponds to the fact that we are aware of objects like cats, tables, etc. that are well localized and are in a definite state. The position need not and must not be exact: its uncertainty should be small only relative to the precision with which we can measure it and the uncertainty must remain such for a period of time. Therefore, due to the uncertainty principle, it cannot be too small.

4. The quantum state of the world is the normalized projection of the quantum state of the Universe onto the space corresponding to the classical description of the world. It is a product state only for variables which are relevant for the macroscopic description of the objects. There might be some entanglement between weakly coupled variables like nuclear spins belonging to different objects. In order to keep the form of the quantum state of the world (1), the quantum state of such variables should belong to $|\Phi\rangle$.

5. Since there is a strong philosophical denial of a possibility to have a nondichotomic degree of existence, the name is clearly problematic, however, it seems that no other word fits better.

6. An even more severe difficulty of this kind appears in the *consistent-histories approach* considered by Gell-Mann and Hartle as an advanced MWI. Its basic concept, the *probability of a history*, seems to be meaningless since all histories exist. However, Saunders finds this approach useful for the analysis of probability.

7. This postulate is a counterpart of the collapse postulate of standard quantum mechanics according to which, after measurement, the quantum state collapses to a particular branch with probability proportional to its squared amplitude. (See the entry on quantum mechanics.) However, it differs in two aspects. First, it is the parallel of only the second part of the collapse postulate, the Born Rule, and second, it is related only to part (ii) of the MWI, the connection to our experience, and not to the mathematical part of the theory (i).

8. Proponents of the MWI might argue that, in fact, the burden of an experimental proof lies on the opponents of the MWI, because it is they who claim that there is new physics beyond the well tested Schrödinger equation.

9. Steane challenges the claim that a quantum computer performs parallel computations, but this is certainly the most natural interpretation of the operation of the first quantum algorithm which works faster than any classical one, see Experiment 2 in Deutsch (1986).

Lev Vaidman
vaidman@post.tau.ac.il

First published: March 24, 2002
Content last modified: March 24, 2002

Identity and Individuality in Quantum Theory

What are the metaphysical implications of quantum physics? One way of approaching this question is to consider the impact of the theory on our understanding of objects as individuals with well defined identity conditions. One view is that quantum theory implies that the fundamental particles of physics cannot be regarded as individual objects in this sense. Such a view has motivated the development of non-standard formal systems which are appropriate for representing such non-individual objects. However, it has also been argued that quantum physics is in fact compatible with a metaphysics of individual objects. Nevertheless, such objects are indistinguishable in a sense which leads to the violation of Leibniz's famous Principle of the Identity of Indiscernibles. Finally, this underdetermination of the metaphysics of individuality by the physics has important implications for the realism-antirealism debate.

Introduction

It is typically held that chairs, trees, rocks, people and many of the so-called ‘everyday’ objects we encounter can be regarded as individuals; the issue, then, is how this individuality is to be understood, or what constitutes the ‘principle’ of individuality. This is an issue which has a very long history in philosophy. A number of approaches to it can be broadly delineated.

We might begin by noting that a tree and rock, say, can be *distinguished* in terms of their different

properties. We might then go further and insist that this also forms the basis for ascribing *individuality* to them. Even two apparently very similar objects, such as two coins of the same denomination or so-called identical twins, will display *some* differences in their properties - a scratch here, a scar there, and so on. On this account such differences are sufficient to both distinguish and individuate the objects. This forms the basis of the so-called ‘bundle’ view of individuality, according to which an individual is *nothing but* a bundle or properties. On this view, no two individuals can be absolutely indistinguishable, or indiscernible, in the sense of possessing *exactly* the same set of properties. This last claim has been expressed as the Principle of Identity of Indiscernibles and we shall return to it below.

However, this approach has been criticised on the grounds (among others) that we can surely *conceive* of two absolutely indistinguishable objects: thinking of Star Trek, we could imagine a replicator device which precisely reproduces an object, such as a coin or even a person, giving two such objects with exactly the same set of properties. Not quite, one might respond, since these two objects do not and indeed cannot exist at the same place at the same time; that is, they do not possess the same spatio-temporal properties. In terms of *these* properties, then, the objects can still be distinguished and hence regarded as different individuals. Clearly, then, this approach to the issue of individuality must be underpinned by the assumption that individual objects are *impenetrable*.

A more thorough-going criticism of this property based approach to individuality insists that it conflates *epistemological* issues concerning how we distinguish objects, with *ontological* issues concerning the metaphysical basis of individuality. Thus, it is argued, to talk of distinguishability requires at least two objects but we can imagine a universe in which there exists only one. In such a situation, it is claimed, it would be inappropriate to say that the object is distinguishable but not that it is an individual. Although we do not actually find ourselves in such situations, of course, still, it is insisted, distinguishability and individuality should be kept *conceptually* distinct.

If this line of argument is accepted, then the principle of individuality must be sought in something over and above the properties of an object. One candidate is the notion of substance, in which properties are taken to inhere in some way. Locke famously described substance as a ‘something, we know not what’, since to describe it we would have to talk of its properties but bare substance, by its very nature, has no properties itself.

Alternatively, the individuality of an object has been expressed in terms of its ‘haecceity’ or ‘primitive thisness’ (Adams 1979). As the name suggests, this is taken to be the primitive basis of individuality, which cannot be analysed further. However, it has also been identified with the notion of self-identity, understood as a relational property (Adams *ibid*.) and expressed more formally as ‘a=a’. Each individual is understood to be identical to itself. This may seem like a form of the property based approach we started with, but self-identity is a rather peculiar kind of property.

This is just a sketch of some of the various positions that have been adopted. There has been considerable debate over which of them applies to the everyday objects mentioned above. But at least it is generally agreed that such objects should be regarded as individuals to begin with. What about the fundamental objects posited by current physical theories, such as electrons, protons, neutrons etc.? Can these be

regarded as individuals? One response is that they cannot, since they behave very differently in aggregates from 'classical' individuals.

Quantum Non-Individuality

The argument for the above conclusion -- that the fundamental objects of physics cannot be regarded as individuals -- can be summed up as follows: First of all, both 'classical' and 'quantal' objects of the same kind (e.g. electrons) can be regarded as indistinguishable in the sense of possessing the same intrinsic properties, such as rest mass, charge, spin etc. Consider now the distribution of two such indistinguishable particles over two boxes, or two states in general:

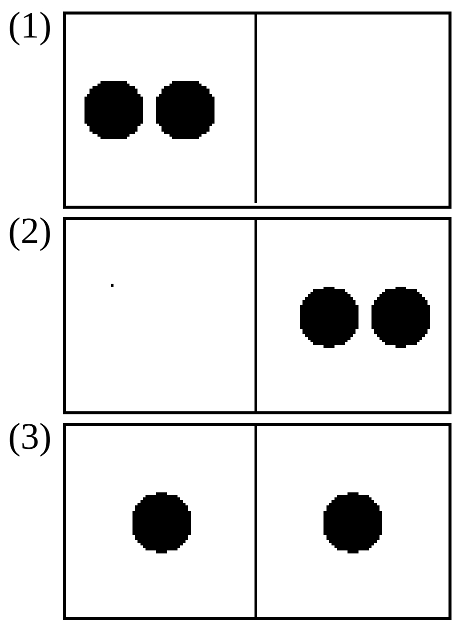


Figure 1

In classical physics, (3) is given a weight of twice that of (1) or (2), corresponding to the two ways the former can be achieved by permuting the particles. This gives us four combinations or complexions in total and hence we can conclude that the probability of finding one particle in each state, for example, is 1/2. (Note that it is assumed that none of the four combinations is regarded as privileged in any way, so each is just as likely to occur.) This is an example of the well-known 'Maxwell-Boltzmann' statistics to which, it is claimed, thermodynamics was reduced at the turn of the century.

In quantum statistical mechanics, however, there are two 'standard' possibilities: one for which there are three possible arrangements in the above situation (both particles in one box, both particles in the other, and one in each box), giving 'Bose-Einstein' statistics; and one for which there is only one arrangement (one particle in each box), giving 'Fermi-Dirac' statistics. Setting aside the differences between these two kinds of quantum statistics, the important point for the present discussion is that in the quantum case, a permutation of the particles is not regarded as giving rise to a new arrangement. This result lies at the very heart of quantum physics and putting things slightly more formally, it is expressed by the so-called

'Indistinguishability Postulate':

> If a particle permutation P is applied to any state function for an assembly of particles, then there is no way of distinguishing the resulting permuted state function from the original unpermuted one by means of any observation at any time.

(The state function of quantum mechanics determines the probability of measurement results. Hence what the Indistinguishability Postulate expresses is that a particle permutation does not lead to any difference in the probabilities for measurement outcomes.)

The argument then continues as follows: that a permutation of the particles is counted as giving a different arrangement in classical statistical mechanics implies that, although they are indistinguishable, such particles can be regarded as individuals (indeed, Boltzmann himself made this explicit in the first axiom of his 'Lectures on Mechanics'). Since this individuality resides in something over and above the intrinsic properties of the particles in terms of which they can be regarded as indistinguishable, it has been called 'Transcendental Individuality' by Post (1963). This notion can be cashed out in various well-known ways, as indicated in the Introduction above: in terms of some kind of underlying Lockean substance (French 1989a), for example, or in terms of primitive thisness (Teller 1995). More generally, one might approach it in modal fashion, through the doctrine of haecceitism: this asserts that two possible worlds may describe some individual in qualitatively the same way (that is, as possessing the same set of properties), yet represent that individual differently by ascribing a different haecceity or thisness in each world, or more generally, by ascribing some non-qualitative aspect to the individual. (Lewis 1986; Huggett 1999).

Conversely, it is argued, if such permutations are not counted in quantum statistics, it follows that quantal particles cannot be regarded as individuals in any of these senses (Post op. cit.). In other words, quantal objects are very different from most everyday objects in that they are 'non-individuals', in some sense.

This radical metaphysical conclusion can be traced back to the very earliest reflections on the foundations of quantum physics. As Weyl put it in his classic text:

> ... the possibility that one of the identical twins Mike and Ike is in the quantum state E1 and the other in the quantum state E2 does not include two differentiable cases which are permuted on permuting Mike and Ike; it is impossible for either of these individuals to retain his identity so that one of them will always be able to say 'I'm Mike' and the other 'I'm Ike.' Even in principle one cannot demand an alibi of an electron! (Weyl 1931)

Recalling the discussion sketched in the Introduction, if we were to create a twin using some kind of Star trek replicator, say, then in the classical domain such a twin could insist that 'I'm here and she's there' or, more generally, 'I'm in this state and she's in that one' and 'swapping us over makes a difference'. In the classical domain each (indistinguishable) twin has a metaphysical 'alibi' grounded in their individuality. Weyl's point is that in quantum mechanics, they do not.

Quantum Individuality

This conclusion -- that quantal objects are not individuals -- is not the whole story, however. First of all, the contrast between classical and quantum physics with regard to individuality and non-individuality is not as straightforward as it might seem. As already indicated, the above account involving permutations of particles in boxes appears to fit nicely with an understanding of individuality in terms of Lockean substance or primitive thisness. However, one can give an alternative field-theoretic account in which particles are represented as dichotomic ‘Yes/No’ fields: with such a field, the field amplitude is simply ‘Yes’ at location x if the ‘particle’ is present at x and ‘No’ if it is not (Redhead 1983). On this account, individuality is conferred via spatio-temporal location together with the assumption of impenetrability mentioned in the Introduction. Thus the above account of particle individuality in terms of either Lockean substance or primitive thisness is not necessary for classical statistical mechanics (French 1989a).

The particles-and-boxes picture above corresponds to the physicists' multidimensional ‘phase space’, which describes which individuals have which properties, whereas the field- theoretic representation corresponds to ‘distribution space’, which simply describes which properties are instantiated in what numbers. Huggett has pointed out that the former supports haecceitism, whereas the latter does not and, furthermore,that the empirical evidence provides no basis for choosing between these two spaces (Huggett 1999). Thus the claim that classical statistical mechanics is wedded to haecceitism also becomes suspect.

Secondly, the above argument from permutations can be considered from a radically different perspective. In the classical case the situations with one particle in each box are given a weight of ‘2’ in the counting of possible arrangements. In the case of quantum statistics this situation is given a weight of ‘1’. With this weighting, there are two possible statistics, however: Bose-Einstein, corresponding to a symmetric state function for the assembly of particles and Fermi-Dirac, corresponding to an anti-symmetric state function. Given the Indistinguishability Postulate, it can be shown that symmetric state functions will always remain symmetric and anti-symmetric always anti-symmetric. Thus, if the initial condition is imposed that the state of the system is either symmetric or anti-symmetric, then only one of the two possibilities -- Bose-Einstein or Fermi-Dirac -- is ever available to the system, and this explains why the weighting assigned to ‘one particle in each state’ is half the classical value. This gives us an alternative way of understanding the difference between classical and quantum statistics, not in terms of the lack of individuality of the particles, but rather in terms of which states are accessible to them (French and Redhead 1988; French 1989a). In other words, the implication of the different ‘counting’ in quantum statistics is not that the particles are non-individuals in some sense, but that there are different sets of states available to them, compared to the classical case. On this view, the particles can still be regarded as individuals - however their individuality is to be understood metaphysically.

Both of these perspectives raise interesting metaphysical issues. Let us consider, first, Leibniz's famous Principle of the Identity of Indiscernibles in the context of the particles- as-individuals package.

Quantum Physics and the Identity of Indiscernibles

It should be emphasised, first of all, that quantal particles are indistinguishable in a much stronger sense than classical particles. It is not just that two or more electrons, say, possess all intrinsic properties in common but that - on the standard understanding - no measurement whatsoever could in principle determine which one is which. If the non-intrinsic, state-dependent properties are identified with all the monadic or relational properties which can be expressed in terms of physical magnitudes associated with self-adjoint operators that can be defined for the particles, then it can be shown that two bosons or two fermions in a joint symmetric or anti-symmetric state respectively have the same monadic properties and the same relational properties one to another (French and Redhead 1988; see also Butterfield 1993). This has immediate implications for Leibniz's Principle of the Identity of Indiscernibles which, expressed crudely, insists that two things which are indiscernible, must be, in fact, identical.

Setting aside the historical issue of Leibniz's own attitude, supporters of the Principle have tended to retreat from the claim that it is necessary, to the view that it is at least contingently true. There is the further issue as to how the Principle should be characterised and, in particular, there is the question of what properties are to be included within the scope of those relevant to judgments of indiscernibility. Excluding the peculiar property of self-identity, three forms of the Principle can be distinguished according to the properties involved: the weakest form, PII(1), states that it is not possible for two individuals to possess all properties and relations in common; the next strongest, PII(2), excludes spatio-temporal properties from this description; and the strongest form, PII(3), includes only monadic, non-relational properties. Thus, for example, PII(3) is the claim that no two individuals can possess all the same monadic properties (a strong claim indeed, although it is one way of understanding Leibniz's own view).

In fact, PII(2) and PII(3) are clearly violated in classical physics, where distinct particles of the same kind are typically regarded as indistinguishable in the sense of possessing all intrinsic properties in common and such properties are regarded as non-relational in general and non-spatio-temporal in particular. (Of course, Leibniz himself would not have been perturbed by this result, since he took the Principle of Identity of Indiscernibles to ultimately apply only to ‘monads’, which were the fundamental entities of his ontology. Physical objects such as particles were regarded by him as merely ‘well founded phenomena’.) However, PII(1) is not violated classically, since classical statistical mechanics typically assumes that such particles are impenetrable, in precisely the sense that their spatio-temporal trajectories cannot overlap. Hence they can be individuated via their spatio-temporal properties, as indicated above.

The situation appears to be very different in quantum mechanics, however. If the particles are taken to possess both their intrinsic and state-dependent properties in common, as suggested above, then there is a sense in which even the weakest form of the Principle, PII(1), fails (Cortes 1976; Teller 1983; French 1989b; for an alternative view, see van Fraassen 1985 and 1991). On this understanding, the Principle of Identity of Indiscernibles is actually false. Hence it cannot be used to effectively guarantee individuation via the state-dependent properties by analogy with the classical case. If one wishes to maintain that quantum particles are individuals, then their individuality will have to be taken as conferred by Lockean

substance, primitive thisness or, in general, some form of non-qualitative haecceistic difference.

Of course, if the particles are taken to be non-individuals, in some still to be articulated sense, then the issue is simply obviated and Leibniz's Principle does not apply. However, what sense can we make of the notion of ‘non-individuality’?

Non-individuality and self-identity

Let us recall Weyl's statement that one can't ask alibis of electrons. Dalla Chiara and Toraldo di Francia refer to quantum physics as ‘the land of anonymity’, in the sense that, on this view,the particles cannot be uniquely labelled (1993). They ask, then, how can we talk about what happens in such a land? Their suggestion is that quantal particles can be regarded as ‘intensional-like entities’, where the intensions are represented by conjunctions of intrinsic properties. The extension of the natural kind, ‘electron’, say, is then given by the collection of indistinguishable elements, or a ‘quaset’. Quaset theory then gives the possibility of a semantics for quantum particles without alibis (ibid.).

Alternatively, but relatedly, non-individuality can be understood in terms of a loss of self- identity. This suggestion can be found most prominently in the philosophical reflections of Born, Schrödinger, Hesse and Post (Born 1943; Schrödinger 1952; Hesse 1963; Post 1963). It is immediately and clearly problematic, however: how can we have objects that are not self-identical? Such self-identity seems bound up with the very notion of an object in the sense that it is an essential part of what it is to be an object. This intuition is summed up in the Quinean slogan, ‘no entity without identity’ (Quine 1969), with all its attendant consequences regarding reference etc.

However, Barcan Marcus has offered an alternative perspective, insisting on ‘No identity without entity.’ (Marcus 1993) and arguing that although ‘... all terms may "refer" to objects... not all objects are things, where a thing is at least that about which it is appropriate to assert the identity relation.’ (ibid., p. 25) Object-reference then becomes a wider notion than thing-reference. Within such a framework, we can then begin to get a formal grip on the notion of objects which are not self-identical through so-called ‘Schrödinger logics’, introduced by da Costa (da Costa and Krause 1994) These are many-sorted logics in which the expression x = y is not a well-formed formula in general; it is where x and y are one sort of term, but not for the other sort corresponding to quantum objects. A semantics for such logics can be given in terms of ‘quasi-sets’ (da Costa and Krause 1997). The motivation behind such developments is the idea that collections of quantum objects cannot be considered as sets in the usual Cantorian sense of ‘... collections into a whole of definite, distinct objects of our intuition or of our thought.’ (Cantor 1955, p. 85.). Quasi-set theory incorporates two kinds of basic posits or ‘Urelemente’: m-atoms, whose intended interpretation are the quantal objects and M-atoms, which stand for the ‘everyday’ objects, and which fall within the remit of classical set theory with Ur- elements. Quasi-sets are then the collections obtained by applying the usual Zermelo- Fraenkel framework plus Urelement ZFU-like axioms to a basic domain composed of m- atoms, M-atoms and aggregates of them (Krause 1992; for a comparison of qua-set theory with quasi-set theory, see Dalla Chiara, Giuntini and Krause 1998).

These developments supply the beginnings of a categorial framework for quantum ‘non- individuality’ which can be extended into the foundations of Quantum Field Theory, where it has been argued, one has non-individual ‘quanta’ (Teller 1995). A form of quasi-set theory may offer one way of formally capturing this notion (French and Krause 1999).

Metaphysical Underdetermination

We now appear to have an interesting situation. Quantum mechanics is compatible with two distinct metaphysical ‘packages’, one in which the particles are regarded as individuals and one in which they are not. Thus, we have a form of ‘underdetermination’ of the metaphysics by the physics (see van Fraassen 1985 and 1991; French 1989a; Huggett 1997; Balousek, forthcoming). This has implications for the broader issue of realism within the philosophy of science. If asked to spell out her beliefs, the realist will point to currently accepted fundamental physics, such as quantum mechanics, and insist that the world is, at least approximately, however the physics says it is. Of course, there are the well-known problems of ontological change (giving rise to the so-called pessimistic meta-induction) and underdetermination of theories by the data. However, the above underdetermination of metaphysical packages seems to pose an even more fundamental problem, as the physics involved is well entrenched and the difference in the metaphysics seemingly as wide as it could be. These packages support dramatically different world-views: one in which quantal particles are individuals and one in which they are not. The realist must then face the question: which package corresponds to the world? The physics itself can offer no help whatsoever and any justification for choosing one package over the other which appeals to metaphysical considerations, for example, runs the risk of drastically watering down the science in scientific realism.

Faced with this situation, the anti-realist may conclude ‘so much for metaphysics’ and insist that all that theories can tell us is how the world *could* be (van Fraassen 1991). A possible alternative would be for realism to retreat from a metaphysics of objects entirely and develop an ontology of structure compatible with the physics (Ladyman 1998). An early attempt to do this in the quantum context can be seen in the work of Cassirer who noted the implications for our notion of individual objects and concluded that particles were describable only as ‘"points of intersection" of certain relations’ (1937, p. 180) However, the neo-Kantian elements in Cassirer's structuralist approach may lead one to wonder whether this suggestion actually takes us too far from realism.

Alternatively, it has been argued that the underdetermination can in fact be ‘broken’ because the package of particles-as-non-individuals meshes better with quantum field theory (QFT) where, it is claimed, talk of individuals is avoided from the word go (Post, op. cit.; Redhead and Teller 1991 and 1992; Teller 1995). The central argument for such a claim focuses on the above view that particles may be seen as individuals subject to restrictions on the sets of states they may occupy. The states that are inaccessible to the particles of a particular kind can be seen as corresponding to just so much ‘surplus structure’. In particular, if the view of particles as individuals is adopted, then it is entirely mysterious as to why a particular sub-set of these inaccessible, surplus states, namely those that are non-symmetric, are not actually realised. Applying the general methodological principle that a theory which does not contain such surplus structure is to be preferred over one that does, Redhead and Teller conclude that we have

grounds for preferring the non-individuals approach and the afore-mentioned mystery simply does not arise.

This line of argument has been criticised by Huggett on the grounds that the apparent mystery is a mere fabrication: the inaccessible non-symmetric states can be ruled out as simply not physically possible (Huggett 1995). The surplus structure, then, is a consequence of the representation chosen and has no further metaphysical significance. At issue here is the claim that a theory should tell us why a state of affairs is not possible. Consider the possible state of affairs in which a cold cup of tea spontaneously starts to boil. Statistical mechanics can explain why we never observe such a possibility, whereas the quantum-particles-as-individuals view cannot explain why we never observe non- symmetric states (Teller 1998).

However, the analogy is problematic. Statistical mechanics does not say that the above situation never occurs but only that the probability of its occurrence is extremely low. The question then reduces to that of ‘why is this probability so low?’ The answer to that is typically given in terms of the very low number of states corresponding to the tea boiling compared to the vast number of states for which it remains cold. Why, then, this disparity in the number of accessible states? Or, equivalently, why do we find ourselves in situations in which entropy increases? One answer takes us back to the initial conditions of the big bang. Perhaps a similar line can be taken in the case of quantum statistics. Why do we never observe non-symmetric states? Because that is the way the universe is and we should not expect quantum mechanics alone to have to explain why certain initial conditions obtain and not others. Here we recall that the symmetry of the Hamiltonian ensures that if a particle is in a state of a particular symmetry to begin with, it will remain in states of that symmetry. Hence, if non-symmetric states do not feature in the initial conditions which held at the beginning of the universe, they will remain forever inaccessible to the particles. The issue then turns on different views of the significance of the above ‘surplus structure’. (A detailed critique of the presuppositions of the Redhead and Teller argument can also be found in Balousek, forthcoming.)

Furthermore, even if we accept the methodological principle of ‘the less surplus structure the better’, it is not clear that QFT understood in terms of non-individual ‘quanta’ offers a significant advantage in this respect. Indeed, it has been argued that the formalism of QFT is compatible with the alternative package of metaphysically individual particles. van Fraassen has pressed this claim (1991), drawing on de Muynck's construction of state spaces for quantum field theory which involve labelled particles (1975). However, Butterfield has suggested that the existence of states that are superpositions of particle number, within QFT, undermines the equivalence (1993). Nevertheless, Huggett insists, in this case the undermining is empirical, rather than methodological (Huggett 1995). When the number is constant, it is the states for arbitrary numbers of particles which are so much surplus structure and now, if the methodological argument is applied, it is the individuals package which is to be preferred.

The exploration of these concerns in the context of quantum field theory has only just begun (see also Auyang 1995) and a collection of historical and philosophical reflections on relevant issues can be found in Cao (1999).

A further approach to this underdetermination is to reject both packages and seek a third way. Thus Lavine has suggested that quantum particles can be regarded as the smallest possible amounts of 'stuff' and, crucially, that a multi-particle state represents a further amount of stuff such that it does not contain proper parts (1991). Such a view, he claims, avoids the metaphysically problematic aspects of both the individuals and non-individuals packages. Of course, there are then the issues of the metaphysics and logic of 'stuff', but, he insists, these are familiar and not peculiar to quantum mechanics. One such issue concerns the nature of 'stuff': is it our familiar primitive substance? Substance as a fundamental metaphysical primitive faces acute difficulties and it has been suggested that it should be dropped in favour of an analysis of individual objects in terms of 'tropes', where a trope is an individual instance of a property or a relation. If this notion is broadened to include an individual whose existence depends on that of another individual which is not a part of it then, it is claimed, this notion may be flexible enough to accommodate quantum physics (Simons 1998). Another issue concerns the manner in which 'stuff' combines: how do we go from the amounts of stuff represented by two independent photons, to the amount represented by a joint two-photon state? The analogies Lavine gives are well known: drops of water, money in the bank, bumps on a rope (Teller 1983; Hesse 1963). Of course, these may also be appropriated by the non-individual objects view but, more significantly, they are suggestive of a field-theoretic approach in which the 'stuff' in question is the quantum field.

Here we return to issues concerning the metaphysics of quantum field theory and it is worth pointing out that underdetermination may arise here too. In classical physics we are faced with a choice between the view of field quantities as properties of space-time points and the view of the field as a kind of substance or stuff. In the case of quantum field theory, the field quantities are not well-defined at space-time points (because of difficulties in defining exact locational states in quantum field theory). Instead they are regarded as 'smeared' over space-time regions (see Teller 1999). This does not remove the possibility of underdetermination, of course, as it now arises between the understanding of the quantum field in terms of properties of space-time *regions* and the understanding of the field in terms of substance. However, further issues then arise with regard to the nature of space-time itself. Conceiving of a field in terms of a set of properties meshes comfortably with the approach that takes space-time to be a kind of substance or 'stuff'. This approach faces well known difficulties in the context of modern physics (see, for example, Earman 1989). Unfortunately, the above properties based account of a field is incompatible with the alternative approach to space-time, which takes it to be merely a system of relations (such as contiguity) between physical bodies: if the field quantities are properties of space-time regions and the latter are understood, ultimately, to be reducible to relations between physical objects, where the latter are conceived of in field-theoretic terms, then a circularity appears to arise. If General Relativity is understood as supporting this 'relationist' account of space-time, then we appear to have a significant incompatibility between these two fundamental theories of modern physics (Rovelli 1999). Perhaps, as Stachel has suggested, this incompatibility can be traced back to the sharp, *metaphysical* distinction between *things* and *relations* between things (Stachel 1999). A broadly 'structural realist' approach might offer a way around this incompatibility by regarding both space-time and the quantum field in structural terms (see Auyang 1995).

Such an approach can also be articulated within the particles picture. Returning to the more developed views of both Weyl and Wigner, particles can be understood as ontologically constituted, in group

theoretical terms, as sets of invariants, such as rest mass, charge or spin, for example (Castellani 1998a). From this perspective, both the individuality and non-individuality packages get off on the wrong feet, as it were, by taking it that there is something - transcendental individuality - that is present in the one case and ‘lost’ in the other. The suggestion that particles might be seen as aspects of ‘world structure’ again fits nicely with structural realism. However, in the absence of further metaphysical explication of the notion of structure itself, it is not yet clear whether or not such an approach collapses into another form of the well known conception of objects as bundles of properties, mentioned in the Introduction.

Excellent overviews of the above and related issues can be found in Huggett (1997) and Castellani (1998b).

Bibliography

- Adams, R., 1979, "Primitive Thisness and Primitive Identity", *Journal of Philosophy***76** (1979): 5-26
- Auyang, S. Y., 1995, *How is Quantum Field Theory Possible?* Oxford: Oxford University Press
- Balousek, D., forthcoming, "Statistics, Symmetry and the Conventionality of Indistinguishability in Quantum Mechanics," *Foundations of Physics*
- Born, M., 1943, *Experiment and Theory in Physics*, Cambridge: Cambridge University Press
- Butterfield, J., 1993, "Interpretation and Identity in Quantum Theory", *Studies in History and Philosophy of Science* **24** (1993): 443-476
- Cao, T.L. (ed.), 1999, *Conceptual Foundations of Quantum Field Theory*, Cambridge: Cambridge University Press
- Cassirer, E., 1937, *Determinism and Indeterminism in Modern Physics*, New Haven: Yale University, 1956; translation of *Determinismus und Indeterminismus in der modern Physik*, Goteborg: Elanders Boktryckeri Aktiebolag, 1937
- Cantor, G., 1955, *Contributions to the Founding of the Theory of Transfinite Numbers*, New York: Dover
- Castellani, E., 1998a, "Galilean Particles: An Example of Constitution of Objects", in Castellani, E. (ed.), *Interpreting Bodies: Classical and Quantum Objects in Modern Physics*, Princeton: Princeton University Press, pp. 181-194
- Castellani, E., 1988b, "Introduction", in Castellani, E. (ed.), *Interpreting Bodies: Classical and Quantum Objects in Modern Physics*, Princeton: Princeton University Press, pp. 3-17
- Cortes, A., 1976, "Leibniz's Principle of the Identity of Indiscernibles: A False Principle", *Philosophy of Science* **43** (1976): 491-505
- da Costa, N. C. A. and Krause, D., 1994, "Schr&oouml;dinger Logics", *Studia Logica* **53** (1994): 533-550
- de Muynck, W., 1975, "Distinguishable and Indistinguishable-Particle Descriptions of Systems of Identical Particles", *International Journal of Theoretical Physics* **14** (1975): 327-346
- Dalla Chiara, M. L. and Toraldo di Francia, G., 1993, "Individuals, Kinds and Names in Physics", in Corsi, G. et al. (eds.), *Bridging the Gap: Philosophy, Mathematics, Physics*, Dordrecht: Kluwer Academic Publishers, pp. 261-283

- Dalla Chiara M. L. and Toraldo di Francia, G., 1995, "Identity Questions from Quantum Theory", in Gavroglu, K. et. al. (eds.), *Physics, Philosophy and the Scientific Community*, Dordrecht: Kluwer Academic Publishers, pp. 39-46
- Dalla Chiara, M. L., Giuntini, R. and Krause, D., 1998, "Quasiset Theories for Microobjects: A Comparision", in Castellani, E. (ed.), *Interpreting Bodies: Classical and Quantum Objects in Modern Physics*, Princeton: Princeton University Press, pp. 142-152
- Earman, J., 1989, *World Enough and Space-Time*, Cambridge: MIT Press
- French, S., 1989, "Identity and Individuality in Classical and Quantum Physics", *Australasian Journal of Philosophy* **67** (1989): 432-446
- French, S., 1998, "On the Withering Away of Physical Objects", in Castellani, E. (ed.), *Interpreting Bodies: Classical and Quantum Objects in Modern Physics*, Princeton: Princeton University Press, pp. 93-113
- French, S. and Krause, D., 1999, "The Logic of Quanta", in T.L. Cao (ed.), *Conceptual Foundations of Quantum Field Theory* Cambridge: Cambridge University Press, pp. 324-342
- French, S. and Redhead, M., 1988, "Quantum Physics and the Identity of Indiscernibles", *British Journal for the Philosophy of Science* **39** (1988): 233-246
- Hesse, M., 1963, *Models and Analogies in Science*, London: Sheed and Ward; reprinted Notre Dame: University of Notre Dame Press, 1966
- Huggett, N., 1995, "What are Quanta, and Why Does it Matter?", *PSA 1994 (Proceedings of the 1994 Biennial Meeting of the Philosophy of Science Association)* Vol. 2, East Lansing: Philosophy of Science Association (1995): 69-76
- Huggett, N., 1997, "Identity, Quantum Mechanics and Common Sense", *The Monist* **80** (1997): 118-130.
- Huggett, N., 1999, "Atomic Metaphysics", *The Journal of Philosophy* **96** (1999): 5-24.
- Krause, D., 1992, "On a Quasi-set Theory", *Notre Dame Journal of Formal Logic* **33** (1992): 402-411.
- Ladyman, J., 1998, "What is Structural Realism?", *Studies in History and Philosophy of Science* **29** (1998): 409-424.
- Lavine, S., 1991, "Is Quantum Mechanics an Atomistic Theory?", *Synthese* **89** (1991): 253-271.
- Lewis, D., 1986, *On the Plurality of Worlds*, Oxford: Blackwell
- Marcus, Barcan R., 1993, *Modalities: Philosophical Essays* Oxford: Oxford University Press
- Post, H., 1963, "Individuality and Physics", *The Listener* **70** (1963): 534-537; reprinted in *Vedanta for East and West* **32**: 14-22
- Quine, W.V.O., 1969, "Speaking of Objects", *Ontological Relativity and Other Essays*, New York: Columbia University Press
- Redhead, M., 1983, "Quantum Field Theory for Philosophers", in Asquith, P.D. and Nickles, T. (eds.), *PSA 1982 (Proceedings of the 1982 Biennial Meeting of the Philosophy of Science Association)* Vol. 2, East Lansing: Philosophy of Science Association (1983): 57-99
- Redhead, M. and Teller, P., 1991, "Particles, Particle Labels, and Quanta: the Toll of Unacknowledged Metaphysics", *Foundations of Physics* **21** (1991): 43-62
- Redhead, M. and Teller, P., 1992, "Particle Labels and the Theory of Indistinguishable Particles in Quantum Mechanics", *British Journal for the Philosophy of Science* **43** (1992): 201-218
- Rovelli, C., 1999, "Localization in Quantum Field Theory: How Much of Quantum Field Theory

is Compatible With What We Know about Space-Time?", in Cao, T. (ed.), *Conceptual Foundations of Quantum Field Theory* Cambridge: Cambridge University Press, pp. 207-229
- Schrödinger, E., 1952, *Science and Humanism*, Cambridge: Cambridge University Press
- Simons, P., 1998, "Farewell to Substance: A Differentiated Leave-taking", *Ratio* **11** (1998): 235-252
- Stachel, J., 1999, "Comments", in Cao, T. (ed.), *Conceptual Foundations of Quantum Field Theory* Cambridge: Cambridge University Press, pp. 233-240
- Teller, P., 1983, "Quantum Physics, the Identity of Indiscernibles and Some Unanswered Questions", *Philosophy of Science* **50** (1983): 309-319
- Teller, P., 1995, *An Interpretative Introduction to Quantum Field Theory*, Princeton: Princeton University Press
- Teller, P., 1998, "Quantum Mechanics and Haecceities", in Castellani, E. (ed.), *Interpreting Bodies: Classical and Quantum Objects in Modern Physics*, Princeton: Princeton University Press, pp. 114-141
- Teller, P., 1999, "The Ineliminable Classical Face of Quantum Field Theory", in Cao, T. (ed.), *Conceptual Foundations of Quantum Field Theory* Cambridge: Cambridge University Press, pp. 314-323
- van Fraassen, B., 1984, "The Problem of Indistinguishable Particles", in Cushing, J. T., Delaney, C. F. and Gutting, G. M. (eds.), *Science and Reality: Recent Work in the Philosophy of Science: Essays in Honor of Erman McMullin*, Notre Dame: University Notre Dame Press, pp. 153-172
- van Fraassen, B., 1985, "Statistical Behaviour of Indistinguishable Particles: Problems of Interpretation", in Mittelstaed, P. and Stachow, E. W. (eds.), *Recent Developments in Quantum Logic* Mannheim, pp. 161-187
- van Fraassen, B., 1991, *Quantum Mechanics: An Empiricist View* Oxford: Oxford University Press
- Weyl, H., 1931, *The Theory of Groups and Quantum Mechanics* London: Methuen and Co.; English trans. 2nd ed

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

identity: of indiscernibles | physics: holism and nonseparability | quantum mechanics

Acknowledgements

Thanks to Rob Clifton, Nick Huggett, Decio Krause and James Ladyman for helpful comments.

Steven French
s.r.d.french@leeds.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 15, 2000
Content last modified: February 15, 2000

The Identity of Indiscernibles

The Identity of Indiscernibles is a principle of analytic ontology first explicitly formulated by Wilhelm Gottfried Leibniz in his *Discourse on Metaphysics*, Section 9 (Loemker 1969: 308). It states that no two distinct substances exactly resemble each other. This is often referred to as 'Leibniz's Law' and is typically understood to mean that no two objects have exactly the same properties. The Identity of Indiscernibles is of interest because it raises questions about the factors which individuate qualitatively identical objects. Recent work on the interpretation of quantum mechanics suggests that the principle fails in the quantum domain.

Formulating the Principle

The Identity of Indiscernibles (hereafter called the Principle) is usually formulated as follows: if, for every property *F*, object *x* has *F* if and only if object *y* has *F*, then *x* is identical to *y*. Or in the notation of symbolic logic:

$$(\forall F)(Fx \leftrightarrow Fy) \rightarrow x=y.$$

This formulation of the Principle is equivalent to the Dissimilarity of the Diverse as McTaggart called it, namely: if *x* and *y* are distinct then there is at least one property that *x* has and *y* does not, or vice versa.

The converse of the Principle, $x=y \rightarrow (\forall F)(Fx \leftrightarrow Fy)$, is called the Indiscernibility of Identicals. Sometimes the conjunction of both principles, rather than the Principle by itself, is known as Leibniz's

Law.

Thus formulated, the actual truth of the Principle seems unproblematic for medium-sized objects, such as rocks and trees, for they are complex enough to have distinguishing or individuating features, and hence may always be distinguished by some slight physical difference. But fundamental principles are widely held to be non-contingent. We might require, therefore, that the Principle should hold even for hypothetical cases of qualitatively identical medium sized objects (e.g., clones which, contrary to fact, really are molecule for molecule replicas). In that case, we shall need to distinguish such objects by their spatial relations to other objects (e.g., where they are on the surface of the planet). In that case the Principle is consistent with a universe in which there are three qualitatively identical spheres A, B, and C where B and C are 3 units apart, C and A are 4 units apart and A and B are 5 units apart. In such a universe, A's being 5 units from B distinguishes it from C, and A's being 4 units from C distinguishes it from B. The Principle often gets called into question, however, when we consider qualitatively identical objects in a symmetrical universe. Consider, for instance, a perfectly symmetrical universe consisting solely of three qualitatively identical spheres, A, B and C, each of which is the same distance away from the others. In this case there seems to be no property which distinguishes any of the spheres from any of the others. Some would defend the Principle even in this case by claiming that there are properties such as *being that very object A.* Call such a property a *thisness* or *haecceity*.

The possibility of resorting to thisnesses might make us query whether the usual formulation of the Principle is correct. For as initially stated the Principle told us that no two substances exactly resemble each other. Yet if A and B otherwise exactly resemble each other then, on a common intuition, the fact that A has the property *being identical to A* while B has the distinct property *being identical to B* cannot result in a respect in which A and B fail to resemble each other.

Rather than argue about these intuitions and hence argue as to which is the correct formulation of the Principle we may distinguish different formulations, and then discuss which, if any, of these are correct. To that end a distinction is commonly made between *intrinsic* and *extrinsic* properties. Here it might initially seem that extrinsic properties are those analysed in terms of some relation. But this is not correct. For the property *being composed of two concentric spheres* is intrinsic. For present purposes it suffices to have an intuitive grasp of the intrinsic/extrinsic distinction.

Another useful distinction is between the *pure* and the *impure*. A property is said to be *impure* if it is analysed in terms of a relation with some particular substance (e.g., *being within a light year of the Sun*). Otherwise it is *pure* (e.g., *being within a light year of a star*). Those two examples are both of extrinsic properties, but some intrinsic properties are impure, (e.g., *being composed of the Earth and the Moon*). According to my definitions all non-relational properties are pure.

Armed with these distinction we may ask which properties are to be considered when we formulate the Principle. Of the various possibilities two seem to be of greatest interest. The *Strong* version of the Principle restricts it to pure intrinsic properties, the *Weak* to pure properties. If we allow impure properties the Principle will be even weaker and, I would say, trivialised. For instance in the three sphere example the impure properties *being 2 units from B* and *being 2 units from C* are possessed by A and

only A, yet intuitively they do not prevent exact resemblance between A B and C. (For a different classification of principles, see Swinburne (1995.))

Suppose we take identity to be a relation and analyse thisnesses as relational properties, (So A's thisness is analysed as *being identical to A*). Then thisnesses will be impure but intrinsic. In that case the world consisting of the three qualitatively identical spheres distance apart 3, 4 and 5 units satisfies the Weak but not the Strong Principle. And the world with the three spheres each 2 units distance from the others satisfies neither version.

A further distinction is whether the Principle concerns all items in the ontology or it is restricted to just the category of *substances* (ie things which have properties and/or relations but are not themselves properties and/or relations.) It is usually thus restricted although Swinburne (1995) does consider, and defend, its application to such abstract objects as integers, times and places, without explicitly treating these as substance.

Ontological Implications

Most formulations of the Principle carry a prima facie commitment to an ontology of properties, but nominalists of various kinds should have little difficulty in providing suitable paraphrases to avoid this commitment. Most interesting in this context is the way the Principle can be stated in terms of resemblance without any mention of properties at all. Thus the Strong Principle might be formulated as denying that distinct substances ever exactly resemble, and the Weak Principle as denying that distinct states of affairs ever exactly resemble.

Russell (e.g., 1940, Chapter 6) held that a substance just is a bundle of universals themselves related by a special relation between properties, known as *compresence*. If the universals in question are taken to be intrinsic properties, then Russell's theory implies the Strong Principle.(At least it *seems* to imply it, but see O'Leary-Hawthorne 1995 and Zimmerman 1997.) And if the status of substances is non-contingent then it implies the necessity of the Strong Principle. This is important because the most vulnerable version is clearly the Strong when it is held to be non-contingent. (See also Armstrong 1989, Chapter 4.)

Recent arguments for and against the Principle

(i) The Principle appeals to empiricists. For how could we ever have empirical evidence for two indiscernible items? If we did, empiricists might say, then they would have to be differently related to us. So unless we ourselves have exact replicas, which is implausible, empirically distinguishable objects must have different pure properties. From this and the empiricist premiss that there are no things which are not empirically distinguishable, we would conclude that the Weak Principle holds. Presumably the premiss would not be proposed as anything more than contingently true. For there are possible situations in which there would be theoretical reasons for believing in indiscernible items as a consequence of a theory which best explains the empirical data. Thus we might come to hold a theory of the origins of the

physical universe which had large amounts of empirical support, and which implied that, in addition to our enormously complicated universe, various simpler ones had been generated. For some of the simplest universes this theory might imply that there were exact replicas. In that case the Weak Principle would fail.

(ii) If we ignore quantum mechanics, we might well conclude that not merely the Weak Principle is contingently correct but even the Strong Principle. For unless we take space to be discrete, the classical mechanical situation would seem to be summed up by the Poincaré recurrence theorem which tells us that typically we get arbitrarily close to an exact repetition, but never get to one. (See Earman 1986, p. 130.)

(iii) Concerning the Weak Principle there has been an interesting development of a line of argument due to Black (1952) and Ayer (1954) in which it is proposed that there could be exact symmetry in the universe, even though, once again, the probability of this occurring is infinitesimal. In Black's example it is suggested that there could be a universe containing nothing but two exactly resembling spheres. In such a completely symmetrical universe the two spheres would be indiscernible. Against this has been noted, e.g., Hacking (1975), that such a completely symmetrical situation of two spheres could be re-interpreted as one sphere in a non-Euclidean space. So what might be described as a journey from one sphere to a qualitatively identical one 2 units apart could be redescribed as a journey around space back to the very same sphere. Quite generally it might be said that we may always redescribe apparent counter-examples to the Weak Principle so that qualitatively identical objects symmetrically situated are interpreted as the very same object.

A rejoinder to this is the continuity argument, essentially due to Adams (1979). It is granted that almost perfect symmetry is possible. For there could be a space with nothing in it but two distinct spheres differing very slightly. Black's example of qualitatively identical spheres is the limiting case as the differences get less and less.

In addition to this rejoinder, it should be noted that in only slightly more complicated examples the identification strategy is rather less persuasive than in the two sphere case. Consider the example of three qualitatively identical spheres arranged in a line, with the two outside ones the same distance from the middle one. The identification strategy would first require the two outer ones to be identified. But in that case there remain two qualitatively identical spheres, so these must in turn be identified. The upshot is that it is not merely the two spheres we took to be indistinguishable that are said to be identical but all three, including the middle one which seemed clearly distinguished from the other two by means of a pure relational property.

Without an appeal to quantum mechanics we have, then, arguments that many find persuasive to show that both the Weak and the Strong Principle are contingently true but neither are necessarily so.

The Impact of Quantum Mechanics

Quantum mechanics has been taken to have implications for the Principle.

(i) Orthodox quantum mechanics tells us that the state of a system of n particles of the same kind is one in which there is nothing to distinguish the particles one from another. That is not to say that they occupy the same position, have the same momentum, and have the same spin, but rather that there is nothing in the many-particle state which says which particle is which. Although controversial as an interpretation, a useful heuristic is to think of all the particles as equally at all the positions they might be in but not determinately at any of them. And likewise for momentum and spin. In that case the particles would seem to be indiscernible, thus showing that even the Weak Principle is false, and that it is false of nomic necessity. (See French 1988, 1989.) Against this it might be argued that a hidden variable interpretation would show that there are in fact several distinct particles each with their own locations, momenta and spins, although we cannot in fact re-identify a particle from one time to another.

(ii) The issue is complicated by Teller's thesis (Teller 1995, Chapter 2) that particles are not individuals at all. The argument for this may be expressed in a way which is neutral on the topic of hidden variables. Consider someone who knows all about quantum mechanics for a single particle and is predicting what is likely for two particles of the same kind. If that person assumes the particles are genuine individuals, whether discernible or otherwise, then it is fairly likely that among all the allowed probability distributions for such quantities as position, momentum and spin there will be some in which the distributions for the two particles are probabilistically independent. But in fact such independence is (nomically) impossible and the resulting distribution, in both the boson and the fermion case, is one of those we might have predicted for particles which are not individuals. This supports Teller's thesis that particles are not individuals, and so, in one respect at least, like waves.

(iii) Teller's thesis relates nicely to a defence of the Principle which I have not yet mentioned, namely that in the prima facie counter-examples the indiscernible entities are not in fact substances. Thus if the only substances were universes, it would be hard to object to even the necessary Strong Principle. Relying on this style of defence we might well deny that the particles are substances. Perhaps the substance is the composite of all the particles of a given kind. Or we could take the regions of spacetime to be the substances and the quantum state as specifying the intrinsic properties of those regions. States with a spatial symmetry are possible but of infinitesimal probability. So we would draw the conclusion that the Strong Principle is contingently true.

The History of the Principle

Leibniz prudently restricts the Principle to substances. Moreover, Leibniz is committed to saying that the extrinsic properties of substances supervene on the intrinsic ones, which collapses the distinction between the strong and the Weak Principles.

Although the details of Leibniz's metaphysics are debatable, the Principle would seem to follow from Leibniz's thesis of the priority of possibility. (See Leibniz's remarks on possible Adams in his 1686 letter to Arnauld, in Loemker 1969, p. 333.) It does not appear to require the Principle of Sufficient Reason,

which Leibniz sometimes bases it on.(See for example Section 21 of Leibniz's fifth paper in his correspondence with Clarke (Loemker 1969, p. 699). See also Rodriguez-Pereyra 1999.) For Leibniz takes God to have created by actualising substances which already exist as possibilia. Hence there could only be indiscernible actual substances if there were indiscernible ones which were merely possible. Hence if the Principle holds for merely possible substances it holds for actual ones as well. There is, therefore, no point in speculating as to whether there might not be a sufficient reason to actualise two of a possible substance, for God cannot do that since both would have to be identical to the one possible substance. The Principle restricted to merely possible substances follows from Leibniz's identification of substances with complete concepts. For two complete concepts must differ in some conceptual respect and so be discernible.

Bibliography

- Adams, R. M., "Primitive Thisness and Primitive Identity", *Journal of Philosophy* **76** (1979).
- Armstrong, D. M., *Universals: An Opinionated Introduction*, Westview Press (1989).
- Ayer, A. J., *Philosophical Essays*, Macmillan (1954).
- Black, M. , "The Identity of Indiscernibles", *Mind* **61** (1952).
- Cross, C., "Max Black on the Identity of Indiscernibles", *Philosophical Quarterly* **45** (1995).
- Earman, J., *A Primer on Determinism*, Dordrecht: D. Reidel (1954).
- French, S. , "Quantum Physics and the Identity of Indiscernibles", *British Journal of the Philosophy of Science* **39** (1988)
- French, S., "Why the Principle of the Identity of Indiscernibles is not Contingently True Either", *Synthese* **78** (1989)
- Hacking, I., "The Identity of Indiscernibles", *Journal of Philosophy* **72** (1975) Dordrecht: D. Reidel (1969).
- Leibniz, G. W., *Philosophical Papers and Letters*, in Loemker 1969.
- Loemker, L., (ed. and trans.), Leibniz, G. W., *Philosophical Papers and Letters*, 2nd ed., Dordrecht: D. Reidel (1969).
- Morris, M. and Parkinson G. H. R., *Leibniz Philosophical Writings*, Dent (1973).
- O'Leary-Hawthorne, J., "The Bundle Theory of Substance and the Identity of Indiscernibles", *Analysis* **55** (1995)
- Rodriguez-Pereyra. G.,"Leibniz's Argument for the Identity of Indiscernibles in His Correspondence with Clarke". *Australasian Journal of Philosophy* **77** (1999).
- Russell, B., *An Inquiry into Meaning and Truth*, Allen and Unwin (1940).
- Swinburne, R. "Thisness", *Australasian Journal of Philosophy*, **73** (1995)
- Teller, P., *An Interpretive Introduction to Quantum Field Theory*, Princeton University Press (1995).
- Zimmerman, D., "Distinct Indiscernibles and the Bundle Theory", *Mind*, **106** (1997).

Other Internet Resources

- Leibniz Page (Gregory Brown, Philosophy, University of Massachusetts/Amherst)
- Entry on Leibniz, MacTutor History of Mathematics Archive (edited by John J O'Connor and Edmund F Robertson, University of St. Andrews)
- Page on Identity (Open Directory Project: Society -> Philosophy -> Philosophy of Logic -> Identity)

Related Entries

Leibniz, Gottfried Wilhelm | ontology | quantum mechanics | quantum theory: identity and individuality in

Peter Forrest
pforrest@metz.une.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 31, 1996
Content last modified: May 15, 2002

Holism and Nonseparability in Physics

It has sometimes been suggested that quantum phenomena exhibit a characteristic holism or nonseparability, and that this distinguishes quantum from classical physics. One puzzling quantum phenomenon arises when one performs measurements of spin or polarization on certain separated quantum systems. The results of these measurements exhibit patterns of statistical correlation that resist traditional causal explanation. Some have held that it is possible to understand these patterns as instances or consequences of quantum holism or nonseparability. Just what holism and nonseparability are supposed to be has not always been made clear, though, and each of these notions has been understood in different ways. Moreover, while some have taken holism and nonseparability to come to the same thing, others have thought it important to distinguish the two. Any evaluation of the significance of quantum holism and/or nonseparability must rest on a careful analysis of these notions.

Introduction

Holism has often been taken as the thesis that the whole is more than the sum of its parts. Several different interpretations of this epigram prove relevant to physics, as we shall see. Here is a correspondingly vague initial statement of nonseparability: The state of the whole is not constituted by

states of its parts. It is already apparent both that holism and nonseparability are related notions and that their exact relation needs to be clarified.

In one interpretation, holism is a methodological thesis, to the effect that the best way to study the behavior of a complex system is to treat it as a whole, and not merely to analyze the structure and behavior of its component parts. Alternatively, holism may be taken as a metaphysical thesis: There are some wholes whose natures are simply not determined by the nature of their parts. Methodological holism stands opposed to methodological reductionism, in physics as well as in other sciences. But it is a certain variety of metaphysical holism that is more closely related to nonseparability. What is at issue here is the extent to which the properties of the whole are determined by the properties of its parts: property holism denies such determination, and thereby comes very close to a thesis of nonseparability.

By and large, a system in classical physics can be analyzed into parts, whose states and properties determine those of the whole they compose. But the state of a system in quantum mechanics resists such analysis. The quantum state of a system gives a specification of it's probabilistic dispositions to display various properties on measurement. Quantum theory's most complete such specification is given by what is called a pure state. Even when a compound system has a pure state, some of its subsystems may not have their own pure states. Emphasizing this characteristic of quantum mechanics, Schrödinger described such component subsystems as "entangled". Superficially, such entanglement of systems already demonstrates nonseparability. At a deeper level, it has been maintained that the puzzling statistics that arise from measurements on entangled quantum systems either demonstrate, or are explicable in terms of, holism or nonseparability rather than any problematic action at a distance.

The Aharonov-Bohm effect also appears to exhibit action at a distance, as the behavior of electrons is modified by a magnetic field they never experience. But this effect may be understood instead as a result of the local action of nonseparable electromagnetism.

Puzzling correlations arise between distant simultaneous measurements even in the vacuum, according to quantum field theory. Perhaps these, too, are not a result of direct causal connections, but rather a manifestation of some kind of holism or nonseparability?

String theory is an ambitious research program in the framework of quantum field theory. According to string theory, all fundamental particles can be considered to be excitations of underlying non-pointlike entities in a multi-dimensional space. The particles' intrinsic charge, mass and spin may then arise as nonseparable features of the world at the deepest level.

Methodological Holism

Methodologically, holism stands opposed to reductionism, somewhat as follows.

> **Methodological Holism**: An understanding of a certain kind of complex system is best

sought at the level of principles governing the behavior of the whole system, and not at the level of the structure and behavior of its component parts.

Methodological Reductionism: An understanding of a complex system is best sought at the level of the structure and behavior of its component parts.

This seems to capture much of what is at stake in debates about holism in social and biological science. In social science, societies are the complex systems, composed of individuals; while in biology, the complex systems are organisms, composed of cells, and ultimately of proteins, DNA and other molecules. A methodological individualist maintains that the right way to approach the study of a society is to investigate the behavior of the individual people that compose it. A methodological holist, on the other hand, believes that such an investigation will fail to shed much light on the nature and development of society as a whole. There is a corresponding debate within physics. Methodological reductionists favor an approach to (say) condensed matter physics which seeks to understand the behavior of a solid or liquid by applying quantum mechanics (say) to its component molecules, atoms, ions or electrons. Methodological holists think this approach is misguided: As one condensed matter physicist put it "the most important advances in this area come about by the emergence of qualitatively new concepts at the intermediate or macroscopic levels--concepts which, one hopes, will be compatible with one's information about the microscopic constituents, but which are in no sense logically dependent on it." [Leggett (1987), p.113.]

It is surprisingly difficult to find methodological reductionists among physicists. The elementary particle physicist Steven Weinberg, for example, is an avowed reductionist. He believes that by asking any sequence of deeper and deeper why-questions one will arrive ultimately at the same fundamental laws of physics. But this explanatory reductionism is metaphysical in so far as he takes explanation to be an ontic rather than a pragmatic category. On this view, it is not physicists but the fundamental laws themselves that explain why "higher level" scientific principles are the way they are. Weinberg (1992) explicitly distinguishes his view from methodological reductionism by saying that there is no reason to suppose that the convergence of scientific explanations must lead to a convergence of scientific methods.

Metaphysical Holism

The metaphysical holist believes that the nature of some wholes is not determined by that of their parts. One may distinguish three varieties of metaphysical holism: ontological, nomological and property holism.

Ontological Holism: Some objects are not wholly composed of basic physical parts.

Property Holism: Some objects have properties that are not determined by physical properties of their basic physical parts.

> **Nomological Holism**: Some objects obey laws that are not determined by fundamental physical laws governing the structure and behavior of their basic physical parts.

All three theses require an adequate clarification of the notion of a basic physical part. One way to do this would be to consider objects as basic, relative to a given class of objects subjected only to a certain kind of process, just in case every object in that class continues to be wholly composed of a fixed set of these (basic) objects. Thus, atoms would count as basic parts of hydrogen if it is burnt to form water, but not if it is converted into helium by a thermonuclear reaction.

Weinberg's (1992) reductionism is opposed to nomological holism in science. He claims, in particular, that thermodynamics has been explained in terms of particles and forces, which could hardly be the case if thermodynamic laws were autonomous. In fact thermodynamics presents a fascinating but complex test case for the theses both of property holism and of nomological holism. One source of complexity is the variety of distinct concepts of both temperature and entropy that figure in both classical thermodynamics and statistical mechanics. Another is the large number of quite differently constituted systems to which thermodynamics can be applied, including not just gases and electromagnetic radiation but also magnets, chemical reactions, star clusters and black holes. Both sources of complexity require a careful examination of the extent to which thermodynamic properties are determined by the physical properties of the basic parts of thermodynamic systems. A third difficulty stems from the problematic status of the probability assumptions that are required in addition to the basic mechanical laws in order to recover thermodynamic principles within statistical mechanics. (An important example is the assumption that the micro-canonical ensemble is to be assigned the standard, invariant, probability distribution.) Since the basic laws of mechanics do not determine the principles of thermodynamics without some such assumptions (however weak), there may well be at least one interesting sense in which thermodynamics establishes nomological holism.

Property/Relational Holism

While some form of ontological holism has occasionally been considered, the variety of metaphysical holism most clearly at issue in quantum mechanics is property holism. But to see just what the issue is we need a more careful formulation of that thesis.

First the thesis should be contextualized to *physical* properties of composite *physical* objects. We are interested here in how far a physical object's properties are fixed by those of its parts, not in some more general determinationist physicalism. Next, to arrive at an interesting formulation of property holism we must accept that this thesis is not only concerned with properties, and not concerned with all properties. The properties of a whole will typically depend upon *relations* among its proper parts as well as on properties of the individual parts. But if we are permitted to consider *all* properties and relations among the parts, then these trivially determine the properties of the whole they compose. For one relation among the parts is what we might call the complete composition relation--that relation among the parts which holds just in case they compose this very whole with all its properties.

Let us call a canonical set of properties and relations of the parts which may or may not determine the properties and relations of the whole the supervenience basis. To avoid trivializing the theses we are trying to formulate, only certain properties and relations can be allowed in the supervenience basis. The intuition as to which these are is simple--the supervenience basis is to include just the qualitative intrinsic properties and relations of the parts, i.e. the properties and relations which these bear in and of themselves, without regard to any other objects, and irrespective of any further consequences of their bearing these properties for the properties of any wholes they might compose. Unfortunately, this simple intuition resists precise formulation. It is notoriously difficult to say precisely what is meant either by an intrinsic property or relation, or by a purely qualitative property or relation. And the other notions appealed to in expressing the simple intuition are hardly less problematic. But, imprecise as it is, this statement serves already to exclude certain unwanted properties and relations, including the complete composition relation, from the supervenience basis.

Finally, we arrive at the following opposing theses:

> **Physical Property Determination**: Every qualitative intrinsic physical property and relation of a set of physical objects from any domain **D** subject only to type **P** processes supervenes on qualitative intrinsic physical properties and relations in the supervenience basis of their basic physical parts (relative to **D** and **P**).

> **Physical Property Holism**: There is some set of physical objects from a domain **D** subject only to type **P** processes, not all of whose qualitative intrinsic physical properties and relations supervene on qualitative intrinsic physical properties and relations in the supervenience basis of their basic physical parts (relative to **D** and **P**).

There is some residual unclarity in the notion of supervenience that figures in these theses. The idea is familiar enough--that there can be no relevant difference in objects in **D** without a relevant difference in their basic physical parts. I take it that the modality involved here is not logical but broadly physical. One might try to explicate the notion of supervenience here in terms of models of a true, descriptively complete, physical theory. At issue is whether such a physical theory has two models which agree on the qualitative intrinsic physical properties and relations of the basic parts of one or more objects in **D** but disagree on some qualitative intrinsic property or relation of these objects.

Teller (1989) has introduced the related idea of what he calls relational holism.

> **Relational Holism**: There are non-supervening relations--that is, relations that do not supervene on the nonrelational properties of the relata. (p. 214)

Within physics, this specializes to a close relative of physical property holism, namely:

> **Physical Relational Holism**: There are qualitative intrinsic physical relations between some physical objects that do not supervene on their qualitative intrinsic physical

> properties.

Physical property holism entails physical relational holism, but not vice versa. Indeed, physical relational holism seems at first sight too weak to capture any distinctive feature of quantum phenomena: even in classical physics the spatiotemporal relations between physical objects seem not to supervene on their qualitative intrinsic physical properties. But when he introduced relational holism Teller(1987) maintained a view of spacetime as a quantity: On this view spatiotemporal relations do in fact supervene on qualitative intrinsic physical properties of ordinary physical objects, since these include their spatiotemporal properties.

State Nonseparability

Physics treats systems by assigning them states. The thermodynamic state of a gas specifies its pressure, volume and temperature. The state of a system of classical particles is represented in a phase space coordinatized by their positions and momenta, either as a point in that space, or in statistical mechanics by a probability distribution over the space. One expects that if a physical system is composed of physical subsystems, then both the composite system and its subsystems will be assigned states by the relevant physical theory. One further expects that the state of the whole will not be independent of those of its parts, and specifically that if a system is composed of two subsystems, A and B, then it will satisfy a principle formulated by Einstein (1935). Howard (1985, p.180) gives the following translation of this principle, which I will call the

> **Real State Separability Principle**: The real state of the pair AB consists precisely of the real state of A and the real state of B, which states have nothing to do with one another.

But the assignment of states to systems in quantum mechanics seems not to conform to these expectations. Recall that the quantum state of a system gives a specification of its probabilistic dispositions to display various properties on measurement. The mathematical representative of this state is an object defined in a Hilbert space--a kind of vector space. This is somewhat analogous to the representation of the state of a system of particles in classical mechanics in a phase space. Let us formulate a principle of

> **State Separability**: The state assigned to a compound physical system at any time is supervenient on the states then assigned to its component subsystems.

This principle could fail in one of two ways: the subsystems may simply not be assigned any states of their own, or else the states they are assigned may fail to determine the state of the system they compose. Interestingly, state assignments in quantum mechanics have been taken to violate state separability in both ways.

The quantum state of a system may be either pure or mixed. A pure state is represented by a vector in the

system's Hilbert space. On one common understanding, any entangled quantum systems violate state separability in so far as the vector representing the state of the system they compose does not factorize into a vector in the Hilbert space of each individual subsystem that could be taken to represent its pure state. Now in such a case each subsystem *1, 2, ... , n* may be uniquely assigned a what is called a mixed state (represented in its Hilbert space not by a vector but by a so-called von Neumann density operator). But then state separability fails for a different reason: the subsystem mixed states do not uniquely determine the compound system's state. A failure of state separability may not occasion much surprise if states are thought of merely in their role of specifying probabilistic dispositions. But it becomes more puzzling if a system's quantum state also has a role in specifying certain of its categorical properties. For that role may connect a failure of state nonseparability to metaphysical holism and nonseparability.

Spatial and Spatiotemporal Nonseparability

The idea is familiar (particularly to Lego enthusiasts!) that if one constructs a physical object by assembling its physical parts, then the physical properties of that object are wholly determined by the properties of the parts and the way it is put together from them. A principle of spatial separability tries to capture that idea.

> **Spatial Separability**: The qualitative intrinsic physical properties of a compound system are supervenient on those of its spatially separated component systems together with the spatial relations among these component systems.

If we identify the real state of a system with its qualitative intrinsic physical properties, then spatial separability is related to a separability principle stated by Howard (1985, p. 173) to the effect that any two spatially separated systems possess their own separate real states. It is even more closely related to Einstein's (1935) real state separability principle. Indeed, Einstein formulated this principle in the context of a pair A,B of spatially separated systems.

Spatial nonseparability -- the denial of spatial separability -- is closely related to physical property holism (the denial of physical property determination). For spatial relations are the only clear examples of qualitative intrinsic physical relations required in the supervenience basis of the relata in physical property determination: other intrinsic physical relations seem to supervene on them.

If we take a spacetime perspective, then spatial separability naturally generalizes to

> **Spatiotemporal Separability**: Any physical process occupying spacetime region *R* supervenes upon an assignment of qualitative intrinsic physical properties at spacetime points in *R*.

Spatiotemporal separability is a natural restriction to physics of David Lewis's (1986, p. x) principle of Humean supervenience. It is also closely related to another principle formulated by Einstein (1948) in the

following words: "An essential aspect of [the] arrangement of things in physics is that they lay claim, at a certain time, to an existence independent of one another, provided these objects ‘are situated in different parts of space’" (the context of the quote suggests that the "space" Einstein had in mind here was actually spacetime).

As Healey (1991, p. 411) shows, spatiotemporal separability entails spatial separability, and so spatial nonseparability entails spatiotemporal nonseparability. Because it is both more general and more consonant with a geometric spacetime viewpoint, it seems reasonable to consider spatiotemporal separability to be the primary notion, so that nonseparability is understood as its denial.

> **Nonseparability**: Some physical process occupying a region *R* of spacetime is not supervenient upon an assignment of qualitative intrinsic physical properties at spacetime points in *R*.

It is important to note that nonseparability entails neither physical property holism nor spatial nonseparability: a process may be nonseparable even though it involves objects without proper parts.

Holism and Nonseparability in Classical Physics

Classical physics presents no clear examples of either physical property holism or nonseparability. Since any example of physical property holism would give rise to nonseparability, it suffices to consider only the latter possibility. But the assumption that all physical processes are separable forms part of the metaphysical background to classical physics. In Newtonian spacetime, the kinematical behavior of a system of point particles under the action of finite forces would constitute a separable physical process, since it is supervenient upon ascriptions of particular values of position and momentum to the particles along their trajectories during the collision. This separability extends also to their dynamics if the forces arise from fields defined at each spacetime point.

The boiling of a kettle of water is an example of a more complex separable physical process. It consists in the increased kinetic energy of its constituent molecules permitting each to overcome the short range attractive forces which otherwise hold it in the liquid. It thus supervenes on the assignment, at each spacetime point on the trajectory of each molecule, of intrinsic physical properties to that molecule (such as its kinetic energy), together with intrinsic physical properties representing the magnitude and direction of the fields that give rise to the attractive force acting on that molecule at that point.

As an example of a separable process in Minkowski spacetime [the spacetime framework for Einstein's special theory of relativity], consider the propagation of an electromagnetic wave through empty space. This is supervenient upon an ascription of electric and magnetic field vectors at each point in the spacetime.

Any physical process described fully by a local spacetime theory will be separable. For such a theory

proceeds by assigning geometric objects (such as vectors or tensors) to each point in spacetime to represent physical fields, and then requiring that these satisfy certain field equations. But processes described fully by theories of other forms will also be separable. This is true not only of pure field theories, but also of many theories which assign properties to particles at each point on their trajectories. Of familiar classical theories, it is only theories involving direct action between spatially separated particles which involve nonseparability in their description of the dynamical histories of individual particles. But such processes are spatiotemporally separable within spacetime regions that are large enough to include all sources of forces acting on these particles, so that the appearance of nonseparability may be attributed to a mistakenly narrow understanding of the spacetime region these processes actually occupy.

The propagation of gravitational energy according to general relativity apparently involves nonseparable processes, since gravitational energy cannot be localized (it does not contribute to the stress-energy tensor defined at each point of spacetime as do other forms of energy). But even a non-locally-defined gravitational energy will still be supervenient upon the metric tensor defined at each point of the spacetime, and so therefore will be the process of its propagation.

The Quantum Physics of Entangled Systems

A set of entangled quantum systems compose a system whose quantum state is represented quantum mechanically by a tensor-product state-vector which does not factorize into a vector in the Hilbert space of each individual system.

$$\Psi_{1,2,\ldots,n} \neq \Psi_1 \otimes \Psi_2 \otimes \ldots \otimes \Psi_n$$

The quantum states of entangled quantum systems violate state separability. This is not surprising if a system's state merely specifies its probabilistic dispositions for the display of various possible properties on measurement. But it has metaphysical significance if a system's quantum state plays a role in specifying it's categorical properties--it's real state, so that the real state separability principle is threatened. His commitment to this principle is one reason why Einstein denied that a quantum system's real state is given by its quantum state (though it's not clear what he thought its real state consisted in).

Modal interpretations of quantum mechanics endorse Einstein's denial. But what a modal interpretation takes to be the real state of an entangled system may still be closely enough related to quantum states that entangled systems' violation of quantum state separability implies some kind of holism or nonseparability. Van Fraassen (1991, p. 294), for example, sees his modal interpretation as committed to "a strange holism" because it entails that a compound system may fail to have a property corresponding to a tensor product projection operator P×I even though its first component has a property corresponding to P. In fact, a clearer case of holism would arise in a modal interpretation that implied that the component lacked P while the compound had P×I: *ceteris paribus*, that would provide an instance of physical property holism. Other instances of physical property holism arise in the modal interpretation of

Healey (1989), whose rules for property attributions permit a compound quantum system to possess holistic properties--dynamical properties that do not supervene on those of their component quantum systems.

Some have located a kind of holism or nonseparability in the probabilities for results of measurements performed on spatially separated entangled systems. Quantum mechanics predicts the probability distributions for combinations of joint and single measurements of variables including spin and polarization on each of a pair of entangled systems, and many of these distributions have been experimentally verified. The joint probability distributions do not factorize into the product of two independent single distributions. If one thinks that quantum mechanics treats each dynamical variable by replacing a precise real value assignment by a probability distribution for the results of measurements of that dynamical variable, then one might see this already as a violation of the real state separability principle. But if one entertains a theory that supplements the quantum state by values of additional "hidden" variables, then the quantum mechanical probabilities would be taken to arise from averaging over many distinct hidden states. In that case, it would rather be the probability distribution conditional on a complete specification of the values of the hidden variables that should be taken to constitute irreducible dispositions of the system concerned. The real state would then include all these conditional probability distributions.

Such reasoning led Howard (1989,1992) to take outcome independence--the probabilistic independence of the outcomes of a given pair of measurements, one on each of a pair of entangled systems, conditional on definite values for any assumed hidden variables on the joint system--as a separability condition. Outcome independence is closely related to parameter independence--the condition that, given a definite hidden variable assignment, the outcome of a measurement on one of a pair of entangled systems is probabilistically independent of what measurement, if any, is made on the other system. Together with parameter independence, outcome independence implies so-called Bell inequalities. These inequalities constrain the patterns of statistical correlations to be expected between the results of measurements of variables including spin and polarization on a pair of entangled systems in any quantum state. They are often said to constrain the predictions of any local hidden variable theory: this is true to the extent that parameter and outcome independence succeed in expressing locality conditions. Quantum mechanics predicts, and experiment confirms, that such Bell inequalities do not always hold. Howard (1989), as well as Teller (1989), suggested that we understand this as stemming from a failure not of parameter independence but of outcome independence, and that this failure is consequently associated with holism or nonseparability. Howard (1989) blames the violation of Bell inequalities on the violation of his separability condition: Teller (1989) takes it to be a manifestation of relational holism. They both acquit parameter independence of blame because they believe that (at least when the measurement events on the entangled systems are spacelike separated) parameter independence (unlike outcome independence) is a consequence of relativity theory.

Others have questioned this line of reasoning, including the conclusion that its appeal to holism or nonseparability helps one to understand how these correlations involving entangled systems come about without any action at a distance that violates either relativity theory or the

> **Principle of Local Action**: If A and B are spatially distant things, then an external influence on A has no immediate effect on B.

Howard's (1989,1992) identification of outcome independence with a separability condition has proved controversial, as has Teller's (1989) claim that violations of Bell inequalities are no longer puzzling if one embraces (physical) relational holism [Laudisa (1995), Berkowitz (1998)]. And the view that violations of outcome independence are perfectly consistent with relativity theory, while violations of parameter independence are not, has also been criticized [Jones and Clifton (1993), Maudlin (1994)].

Healey (1989,1994) has offered a modal interpretation and used it to present a model account of the puzzling correlations which portrays them as resulting from the operation of a process that violates both spatial and spatiotemporal separability. He argues that, on this interpretation, the nonseparability of the process is a consequence of a violation of physical property holism; and that the resulting account yields genuine understanding of how the correlations come about without any violation of relativity theory or the principle of local action. But subsequent work by Clifton and Dickson (1998) has cast doubt on whether the account can be squared with relativity theory's requirement of Lorentz invariance.

Ontological Holism in Quantum Mechanics?

As applied to physics, ontological holism is the thesis that there are physical objects that are not wholly composed of basic physical parts. Views of Bohr, Bohm and others may be interpreted as endorsing some version of this thesis. In no case is it claimed that any physical object has *non*physical parts. The idea is rather that some physical entities that we take to be wholly composed of a particular set of basic physical parts are in fact not so composed.

It was Bohr's (1934) view that one can meaningfully ascribe properties such as position or momentum to a quantum system only in the context of some well-defined experimental arrangement suitable for measuring the corresponding property. He used the expression ‘quantum phenomenon’ to describe what happens in such an arrangement. In his view, then, although a quantum phenomenon is purely physical, it is not composed of distinct happenings involving independently characterizable physical objects--the quantum system on the one hand, and the classical apparatus on the other. And even if the quantum system may be taken to exist outside the context of a quantum phenomenon, little or nothing can then be meaningfully said about its properties. It would therefore be a mistake to consider a quantum object to be an independently existing component part of the apparatus-object whole.

Bohm's (1980,1993) reflections on quantum mechanics lead him to adopt a more general holism. He believed that not just quantum object and apparatus, but any collection of quantum objects by themselves, constitute an indivisible whole. This may be made precise in the context of Bohm's (1952) interpretation of quantum mechanics by noting that a complete specification of the state of the "undivided universe" requires not only a listing of all its constituent particles and their positions, but also of a field associated with the wave-function that guides their trajectories. If one assumes that the basic

physical parts of the universe are just the particles it contains, then this establishes ontological holism in the context of Bohm's interpretation.

Some [Howard (1989), Dickson (1998)] have connected the failure of a principle of separability to ontological holism in the context of violations of Bell inequalities. Howard (1989) states the following separability principle (pp. 225-6)

> The contents of any two regions of space-time separated by a nonvanishing spatiotemporal interval constitute separable physical systems, in the sense that (1) each possesses its own, distinct physical state, and (2) the joint state of the two systems is wholly determined by these separate states.

He takes Einstein to defend this as a principle of individuation of physical systems, without which physical thought "in the sense familiar to us" would not be possible. Howard himself contemplates the possible failure of this principle for entangled quantum systems, with the consequence that these could no longer be taken to be wholly composed of what are typically regarded as their subsystems. Dickson (1998), on the other hand, argues that such holism is not "a tenable scientific doctrine, much less an explanatory one" (p. 156).

One may try to avoid the conclusion that experimental violations of Bell inequalities manifest a failure of Local Action by invoking ontological holism for events. The idea would be to deny that these experiments involve distinct, spatiotemporally separate, measurement events, and to maintain instead that what we usually describe as separate measurements involving an entangled system in fact constitute one indivisible, spatiotemporally disconnected, event with no spatiotemporal parts. But such ontological holism conflicts with the criteria of individuation of events inherent in both quantum theory and experimental practice.

The Aharonov-Bohm Effect

Aharonov and Bohm (1959) drew attention to the quantum mechanical prediction that an interference pattern due to a beam of charged particles could be produced or altered by the presence of a constant magnetic field in a region from which the particles were excluded. This effect has since been experimentally demonstrated. At first sight, the Aharonov-Bohm effect seems to involve action at a distance. It seems clear that the (electro-)magnetic field acts on the particles since it affects the interference pattern they produce; and this must be action at a distance since the particles pass through a region from which that field is absent. But alternative accounts of the phenomenon are possible which portray it rather as a manifestation of nonseparability [Healey (1997)]. There need be no action at a distance if the behavior both of the charged particles and of electromagnetism are nonseparable processes. While such a treatment of electromagnetism (and other gauge theories) is increasingly common in physics, to treat the motion of the charged particles as a nonseparable process is to endorse a particular position on how quantum mechanics is to be interpreted.

An interpretation of quantum mechanics that ascribes a nonlocalized position to a charged particle on its way through the apparatus is committed to a violation of spatiotemporal separability in the Aharonov-Bohm effect, since the particle's passage constitutes a nonseparable process. To see why the electromagnetism that acts on the particles during their passage may also be taken to be nonseparable it is necessary to consider contemporary representations of electromagnetism in terms of neither fields nor potentials.

Following Wu and Yang's (1975) analysis of the Aharonov-Bohm effect, it has become common to consider electromagnetism to be completely and nonredundantly described neither by the electromagnetic field, nor by its generating potential, but rather by the so-called Dirac phase factor:

$$\exp\left[-(ie/\hbar)\oint_C A_\mu(x^\mu).dx^\mu\right]$$

where A_μ is the electromagnetic potential at spacetime point x^μ, e is the particles' charge, and the integral is taken over each closed loop C in spacetime. Applied to the Aharonov-Bohm case, this means that the constant magnetic field is accompanied by an association of a phase factor $S(C)$ with all closed curves C in space, where $S(C)$ is defined by

$$S(C) = \exp\left[-(ie/\hbar)\oint_C \mathbf{A}(\mathbf{r}).d\mathbf{r}\right]$$

This approach has the advantage that since $S(C)$ is gauge-invariant, it may readily be considered a physically real quantity. Moreover, the effects of electromagnetism in the field-free region may be attributed to the fact that $S(C)$ is nonvanishing for certain closed curves C within that region. But it is significant that, unlike the magnetic field and its potential, $S(C)$ is not defined at each point of space at each moment of time.

Can $S(C)$ at some time be taken to represent an intrinsic property of a region of space corresponding to the curve C? There are two difficulties with this suggestion. The first is that the presence of the quantity e in the definition of $S(C)$ appears to indicate that $S(C)$ rather codes the effect of electromagnetism on objects with that specific charge. If in fact *all* charges are multiples of some minimal value e, then this would no longer be a problem: the fact that $S(C)$ at some time represents an intrinsic property of a region of space corresponding to the curve C would be a natural reflection of this fact. If not, one could rather take

$$I(C) = \oint_C \mathbf{A}(\mathbf{r}).d\mathbf{r}$$

to be an intrinsic property of C. The second difficulty is that closed curves do not correspond uniquely to regions of space: e.g. circling the region in which there is a magnetic field twice on the same circle will

produce a different curve from circling it once. But this does not prevent one from taking $S(C)$ at some time to represent an intrinsic property of the region of space occupied by a nonself-intersecting closed curve C.

Once these difficulties have been handled, it is indeed possible to consider electromagnetism in the Aharonov-Bohm effect as faithfully represented at a time by a set of intrinsic properties of regions of space occupied by nonself-intersecting closed curves. But if one does so, then electromagnetism itself manifests nonseparability. For these intrinsic properties do not supervene on any assignment of qualitative intrinsic physical properties at spacetime points in the region concerned. Whether the magnetic field remains constant or changes, the associated electromagnetism constitutes a nonseparable process, and so the Aharonov-Bohm effect violates spatiotemporal separability. If the motion of the particles through the apparatus is a nonseparable process, then it is possible to account for the AB effect in terms of a purely local interaction between (nonseparable) electromagnetism and this process. For the particles effectively traverse closed curves C on their nonlocalized "trajectories", and so they interact with electromagnetism precisely where this is defined.

Even if the Aharonov-Bohm effect does exhibit such nonseparability, there is no violation of physical property holism (or, indeed, spatial separability). This makes it clear by example that holism and nonseparability are indeed distinct, though related, notions.

Quantum Field Theory

Certain phenomena that arise within quantum field theory have been taken as instances of nonseparability. As in the Aharonov-Bohm effect, this seems not to result from either physical property holism or a violation of spatial separability.

It is well known that even in the vacuum state, quantum field theory predicts statistical correlations between the results of measurements even if these occur in regions that cannot be connected by a light signal. At least in certain special cases, these correlations imply violations of Bell inequalities. No compound systems like photon pairs are involved, so it is hard to see how this can be explained by appeal to physical property holism or spatial nonseparability (though one might argue that in this context spacetime regions constitute the relevant quantum systems, with the subsystem relation corresponding to containment). But it is not unreasonable to suggest that the correlations reflect some failure of spatiotemporal separability. Whether this is true or not depends on whether it is possible to understand the results of simultaneous measurements in quantum field theory as reflecting some intrinsic physical property associated with the disconnected spacetime region occupied by the measurement events.

Wayne (1998) has suggested that quantum field theory is best interpreted as postulating extensive holism or nonseparability. On this interpretation, the fundamental quantities in quantum field theory are vacuum expectation values of products of field operators defined at various spacetime points. The field can be reconstructed out of all of these. Nonseparability supposedly arises because the vacuum expectation

value of a product of field operators defined at an *n*-tuple of distinct spacetime points does not supervene on qualitative intrinsic physical properties defined at those *n* points, together with the spatiotemporal relations among the points.

But it is not clear that vacuum expectation values of products of field operators defined at *n*-tuples of distinct spacetime points represent either qualitative intrinsic physical properties of these n-tuples or physical relations between them. Evaluation of the extent to which quantum field theory illustrates holism or nonseparability must await further progress in the interpretation of quantum field theory. (Redhead(1995) represents a relevant first step.)

String Theory

At the turn of the 21st century, string theory (or its descendant, *M*-theory) has emerged as a speculative candidate for unifying much of fundamental physics, including quantum mechanics and general relativity. Existing string theories proceed by quantizing classical theories of basic entities that are extended in one or more dimensions of a space that has 6 or 7 tiny compact dimensions in addition to the three spatial dimensions of ordinary geometry. If these additional dimensions are appropriately considered spatial, then it is natural to extend the concepts of spatial and spatiotemporal separability to encompass them. In that case, processes involving classical strings (or *p*-branes with $p > 0$) would count as (spatiotemporally) nonseparable, even though all particles and their properties conform to spatial separability.

The status of nonseparability within a quantized string field theory is not so easy to assess, because of the general problems associated with deciding what the ontology of any relativistic quantum field theory should be taken to be.

Bibliography

- Aharonov, Y. and Bohm, D. (1959) "Significance of Electromagnetic Potentials in the Quantum Theory", *Physical Review 115*: 485-91.
- Bell, John (1987) *Speakable and Unspeakable in Quantum Mechanics.* (Cambridge: Cambridge University Press).
- Berkowitz, J. (1998) "Aspects of Quantum Non-Locality I", *Studies in History and Philosophy of Modern Physics 29B*, 183-222.
- Bohm, D. (1952) "A suggested interpretation of the quantum theory in terms of "hidden variables", I and II", *Physical Review 85*: 166-193.
- Bohm, D. (1980) *Wholeness and the Implicate Order* (London: Routledge & Kegan Paul).
- Bohm, D. and Hiley, B.J. (1993) *The Undivided Universe* (New York: Routledge).
- Bohr, N. (1934) *Atomic Theory and the Description of Nature.* (Cambridge: Cambridge University Press).
- Clifton, R. and Dickson, M. (1998) "Lorentz-Invariance in Modal Interpretations", in D.Dieks and

P. Vermaas, *The Modal Interpretation of Quantum Mechanics* (Dordrecht: Kluwer Academic), 9-47.

- Cushing, J. and McMullin, E. (1989) *Philosophical Consequences of Quantum Theory: Reflections on Bell's Theorem* (Notre Dame, Indiana: University of Notre Dame Press).
- D'Espagnat, B. (1983) *In Search of Reality* (New York: Springer Verlag).
- Dickson, M. (1998) *Quantum Chance and Non-Locality*. (Cambridge: Cambridge University Press).
- Einstein, A. (1935) Letter to E. Schroedinger of June 19th: see Howard (1985).
- Einstein, A. (1948) "Quantum Mechanics and Reality", *Dialectica 2*: 320-4. (This translation from the original German by Howard in Howard(1989, pp.233-4.)
- Greene, B. (1999) *The Elegant Universe* (New York: W.W. Norton and Company)
- Healey, R.A. (1989) *The Philosophy of Quantum Mechanics: an Interactive Interpretation* (Cambridge: Cambridge University Press).
- --------- (1991) "Holism and Nonseparability", *Journal of Philosophy, 88*: 393-421.
- --------- (1994) "Nonseparability and Causal Explanation", *Studies in History and Philosophy of the Physical Sciences, 25*: 337-374.
- --------- (1997) "Nonlocality and the Aharonov-Bohm Effect", *Philosophy of Science 64*: 18-41.
- Howard, D. (1985) "Einstein on Locality and Separability", *Studies in History and Philosophy of Science 16*: 171-201.
- --------- (1989) "Holism, Separability and the Metaphysical Implications of the Bell Experiments", in Cushing and McMullin eds. (1989): 224-53.
- --------- (1992) "Locality, Separability and the Physical Implications of the Bell Experiments", in van der Merwe, A., Selleri, F., and Tarozzi, G., eds. *Bell's Theorem and the Foundations of Modern Physics*. (Singapore: World Scientific).
- Jones, M. and Clifton, R. (1993) "Against Experimental Metaphysics", in *Midwest Studies in Philosophy Volume 18*, eds. P. French et. al. (South Bend, Indiana: University of Notre Dame Press), pp.295-316.
- Laudisa, F. (1995) "Einstein, Bell, and Nonseparable Realism", *British Journal for the Philosophy of Science 46*, 309-39.
- Leggett, A. J. (1987) *The Problems of Physics* (New York: Oxford University Press).
- Lewis, D. (1986) *Philosophical Papers, Volume II* (New York: Oxford).
- Maudlin, T. (1994) *Quantum Nonlocality and Relativity*. Oxford: Basil Blackwell.
- --------- (1998) "Part and Whole in Quantum Mechanics", in *Interpreting Bodies*, E. Castellani ed., (Princeton, N.J.: Princeton University Press), 46-60.
- Redhead, M.L.G. (1987) *Incompleteness, Nonlocality and Realism* (Oxford: Clarendon Press).
- --------- (1995) "More Ado About Nothing", *Foundations of Physics 25*, 123-.
- Schroedinger, E. (1935) "Discussion of Probability Relations Between Separated Systems," *Proceedings of the Cambridge Philosophical Society 31*, 555-563.
- Teller, P. (1986) "Relational Holism and Quantum Mechanics," *British Journal for the Philosophy of Science 37*, 71-81.
- --------- (1987) "Space-Time as a Physical Quantity", in *Kelvin's Baltimore Lectures and Modern Theoretical Physics*, R. Kargon and P. Achinstein eds., (Cambridge, Mass.: the MIT Press), 425-447.

- --------- (1989) "Relativity, Relational Holism, and the Bell Inequalities," in Cushing and McMullin, eds., 208-223.
- van Fraassen, B. (1991) *Quantum Mechanics: an Empiricist View*. (Oxford: Clarendon Press, 1991).
- Wayne, A. (1998) "Locality and Separability in the Quantum World", paper read at the meetings of the American Philosophical Association, Pacific Division.
- Weinberg, S. (1992) *Dreams of a Final Theory* (New York: Vintage Books).
- Wu, T.T. and Yang, C.N. (1975) "Concept of Nonintegrable Phase Factors and Global Formulation of Gauge Fields", *Physical Review D12*: 3845.

Other Internet Resources

Two sites with some relevance are:

- James Schombert's (U. of Oregon/Physics) page on Quantum Mechanics
- David Fideler's page on Quantum Nonlocality

Related Entries

action at a distance | Bell’s Theorem | Einstein, Albert: Einstein-Bohr debates | physics: Reichenbach’s common cause principle | quantum mechanics: modal interpretations of | quantum theory: quantum entanglement and information

Richard Healey
rhealey@u.arizona.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 22, 1999
Content last modified: July 22, 1999

Collapse Theories

Quantum mechanics, with its revolutionary implications, has posed innumerable problems to philosophers of science. In particular, it has suggested reconsidering basic concepts such as the existence of a world that is, at least to some extent, independent of the observer, the possibility of getting reliable and objective knowledge about it, and the possibility of taking (under appropriate circumstances) certain properties to be objectively possessed by physical systems. It has also raised many others questions which are well known to those involved in the debate on the interpretation of this pillar of modern science. One can argue that most of the problems are not only due to the intrinsic revolutionary nature of the phenomena which have led to the development of the theory. They are also related to the fact that, in its standard formulation and interpretation, quantum mechanics is a theory which is excellent (in fact it has met with a success unprecedented in the history of science) in telling us everything about *what we observe*, but it meets with serious difficulties in telling us *what is*. We are making here specific reference to the central problem of the theory, usually referred to as *the measurement problem*, or, with a more appropriate term, as the *macro-objectification problem*. It is just one of the many attempts to overcome the difficulties posed by this problem that has led to the development of *Collapse Theories*, i.e., to the *Dynamical Reduction Program* (DRP). As we shall see, this approach consists in accepting that the dynamical equation of the standard theory should be modified by the addition of stochastic and nonlinear terms. The nice fact is that the resulting theory is capable, on the basis of a unique dynamics which is assumed to govern all natural processes, to account at the same time for all well-established facts about microscopic systems as described by the standard theory as well as for the so-called postulate of wave packet reduction (WPR). As is well known, such a postulate is assumed in the standard scheme just in order to guarantee that *measurements have outcomes* but, as we shall discuss below, it meets with insurmountable difficulties if one takes the measurement itself to be a process governed by the linear laws of the theory. Finally, the collapse theories account in a completely satisfactory way for the classical behavior of macroscopic systems.

Two specifications are necessary in order to make clear from the beginning what are the limitations and the merits of the program. The only satisfactory explicit models of this type (which are essentially variations and refinements of the one, usually referred to as the GRW theory, proposed in refs. [Ghirardi, Rimini and Weber, 1985, 1986]) are phenomenological attempts to solve a foundational problem. At present, they involve phenomenological parameters which, if the theory is taken seriously, acquire the status of new constants of nature. Moreover, up to now, all attempts to build satisfactory relativistic generalizations of these models have met with serious mathematical difficulties due to the appearance of untractable divergences, even though they elucidate some crucial points and suggest that there is no reason of principle preventing to reach this goal.

In spite of the above remarks, we think that Collapse Theories have a remarkable relevance, since they represent a new way to overcome the difficulties of the formalism, to *close the circle* in the precise sense defined by Abner Shimony [Shimony, 1989], a way which until a few years ago was considered impracticable, and which, on the contrary, has been shown to be perfectly viable. Moreover, they have allowed a clear identification of the formal features which should characterize any unified theory of micro and macro processes.

---

1. General Considerations

As stated already, a very natural question which all scientists who are concerned about the meaning and the value of science have to face, is whether one can develop a coherent worldview that can accommodate our knowledge concerning natural phenomena as it is embodied in our best theories. Such a program meets serious difficulties with quantum mechanics, essentially because of two formal aspects of the theory which are common to all of its versions, from the original nonrelativistic formulations of the 1920s, to the quantum field theories of recent years: the linear nature of the state space and of the evolution equation, i.e., the validity of the superposition principle and the related phenomenon of entanglement, which, in Schrödinger's words:

> is not one but the characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought [Schrödinger, 1935, p. 807].

These two formal features have embarrassing consequences, since they imply

- objective chance in natural processes, i.e., the nonepistemic nature of quantum probabilities;
- objective indefiniteness of physical properties both at the micro and macro level;
- objective entanglement between spatially separated and non-interacting constituents of a composite system, entailing a sort of holism and a precise kind of nonlocality.

For the sake of generality, we shall first of all present a very concise sketch of ‘the rules of the game’.

2. The Formalism: A Concise Sketch

Let us recall the axiomatic structure of quantum theory:

1. States of physical systems are associated with normalized vectors in a Hilbert space, a complex, infinite-dimensional, linear vector space equipped with a scalar product. Linearity implies that the superposition principle holds: if |*f*> is a state and |*g*> is a state, then (for *a* and *b* arbitrary complex numbers) also

   $$|K> = a|f> + b|g>$$

   is a state. Moreover, the state evolution is linear, i.e., it preserves superpositions: if |*f*,*t*> and |*g*,*t*> are the states obtained by evolving the states |*f*,0> and |*g*,0>, respectively, from the initial time *t*=0 to the time *t*, then *a*|*f*,*t*> + *b*|*g*,*t*> is the state obtained by the evolution of *a*|*f*,0> + *b*|*g*,0>. Finally, the completeness assumption is made, i.e., that the knowledge of its statevector represents, in principle, the most accurate information one can have about the state of an individual physical system.

2. The observable quantities are represented by self-adjoint operators B on the Hilbert space. The associated eigenvalue equations $B|b_k> = b_k|b_k>$ and the corresponding eigenmanifolds (the linear manifolds spanned by the eigenvectors associated to a given eigenvalue, also called eigenspaces) play a basic role for the predictive content of the theory. In fact:

   i. The eigenvalues b_k of an operator B represent the only possible outcomes in a measurement of the corresponding observable.

ii. The norm (i.e. the length) of the projection of the normalized vector (i.e. of length 1) describing the state of the system onto the eigenmanifold associated to a given eigenvalue gives the probability of obtaining the corresponding eigenvalue as the outcome of the measurement. In particular, it is useful to recall that when one is interested in the probability of finding a particle at a given place, one has to resort to the so-called configuration space representation of the statevector. In such a case the statevector becomes a square-integrable function of the position variables of the particles of the system, whose modulus squared yields the probability density for the outcomes of position measurements.

We stress that, according to the above scheme, quantum mechanics makes only conditional probabilistic predictions (conditional on the measurement being actually performed) for the outcomes of prospective (and in general incompatible) measurement processes. Only if a state belongs already before the act of measurement to an eigenmanifold of the observable which is going to be measured, can one predict the outcome with certainty. In all other cases -- if the completeness assumption is made -- one has objective nonepistemic probabilities for different outcomes.

The orthodox position gives a very simple answer to the question: what determines the outcome when different outcomes are possible? Nothing -- the theory is complete and, as a consequence, it is illegitimate to raise any question about possessed properties referring to observables for which different outcomes have non-vanishing probabilities of being obtained. Correspondingly, the referent of the theory are the results of measurement procedures. These are to be described in classical terms and involve in general mutually exclusive physical conditions.

As regards the legitimacy of attributing properties to physical systems, one could say that quantum mechanics warns us against requiring too many properties to be actually possessed by physical systems. However -- with Einstein -- one can adopt as a sufficient condition that one be able (without in any way disturbing the system) to predict with certainty the outcome of a measurement. In this case then, whenever the overall statevector factorizes into the product of a state of the Hilbert space of the physical system S and of the rest of the world, S does possess some properties (actually a complete set of properties, i.e., those associated to a maximal set of commuting observables).

Before concluding this section we must add some comments about the measurement process. Quantum theory was created to deal with microscopic phenomena. In order to obtain information about them one must be able to establish strict correlations between the states of the microscopic systems and the states of objects we can perceive. Within the formalism, this is described by considering appropriate micro-macro interactions. The fact that when the measurement is completed one can make statements about the outcome is accounted for by the already mentioned WPR postulate [Dirac, 1948]: *a measurement always causes a system to jump in an eigenstate of the observed quantity*. Correspondingly, also the statevector of the apparatus ‘jumps’ into the manifold associated to the recorded outcome.

3. The Macro-Objectification Problem

In this Section we shall clarify why the formalism we have just presented gives rise to the measurement or macro-objectification problem. To this purpose we shall, first of all, discuss the standard oversimplified argument based on the so-called von Neumann ideal measurement scheme. Then we shall discuss more recent results [Bassi and Ghirardi, 2000], which relinquish von Neumann's assumptions.

Let us begin by recalling the basic points of the standard argument:

> Suppose that a microsystem *S*, just before the measurement of an observable *B*, is in the eigenstate $|b_j>$ of the corresponding operator. The apparatus (a macrosystem) used to gain information about *B* is initially assumed to be in a precise macroscopic state, its ready state, corresponding to a definite macro property -- e.g., its pointer points at 0 on a scale. Since the apparatus *A* is made of elementary particles, atoms and so on, it must be described by quantum mechanics, which will associate to it the state vector $|A_0>$. One then assumes that there is an appropriate system-apparatus interaction lasting for a finite time, such that when the initial apparatus state is triggered by the state $|b_j>$ it ends up in a final configuration $|A_j>$, which is macroscopically distinguishable from the initial one and from the other configurations $|A_k>$ in which it would end up if triggered by a different eigenstate $|b_k>$. Moreover, one assumes that the system is left in its initial state. In brief, one assumes that one can dispose things in such a way that the system-apparatus interaction can be described as:
>
> (1) (*initial state*): $|b_k>|A_0>$
>
> (*final state*): $|b_k>|A_k>$
>
> Equation (1) and the hypothesis that the superposition principle governs all natural processes tell us that, if the initial state of the microsystem is a linear superposition of different eigenstates (for simplicity we will consider only two of them), one has:
>
> (2) (*initial state*): $(a|b_k> + b|b_j>)|A_0>$
>
> (*final state*): $(a|b_k>|A_k> + b|b_j>|A_j>)$.

Some remarks about this are in order:

- The scheme is highly idealized, both because it takes for granted that one can prepare the apparatus in a precise state, which is impossible since we cannot have control over all its degrees of freedom, and because it assumes that the apparatus registers the outcome without altering the state of the measured system. However, as we shall discuss below, these assumptions are by no means essential to derive the embarrassing conclusion we have to face, i.e., that the final state is a

linear superposition of two states corresponding to two macroscopically different states of the apparatus. Since we know that the + representing linear superpositions cannot be replaced by the logical alternative *either* ... *or*, the measurement problem arises: what meaning can one attach to a state of affairs in which two macroscopically and perceptively different states occur simultaneously?

- As already mentioned, the standard solution to this problem is given by the WPR postulate: in a measurement process reduction occurs: the final state is not the one appearing at the right hand side. of Eq.(2) but, since macro-objectification takes place, it is

(3) either $|b_k>|A_k>$ or $|b_j>|A_j>$ with probabilities $|a|^2$ and $|b|^2$, respectively.

Nowadays, there is a general consensus that this solution is absolutely unacceptable for two basic reasons:

1. It corresponds to assuming that the linear nature of the theory is broken at a certain level. Thus, quantum theory is unable to explain how it can happen that the apparata behave as required by the WPR postulate (which is one of the axioms of the theory).

2. Even if one were to accept that quantum mechanics has a limited field of applicability, so that it does not account for all natural processes and, in particular, it breaks down at the macrolevel, it is clear that the theory does not contain any precise criterion for identifying the borderline between micro and macro, linear and nonlinear, deterministic and stochastic, reversible and irreversible. To use J.S. Bell's words, there is nothing in the theory fixing such a borderline and the *split* between the two above types of processes is fundamentally *shifty*. As a matter of fact, if one looks at the historical debate on this problem, one can easily see that it is precisely by continuously resorting to this ambiguity about the split that adherents of the Copenhagen orthodoxy or *easy solvers* [Bell, 1990] of the measurement problem have rejected the criticism of the *heretics* [Gottfried, 2000]. For instance, Bohr succeeded in rejecting Einstein's criticisms at the Solvay Conferences by stressing that some macroscopic parts of the apparatus had to be treated fully quantum mechanically; von Neumann and Wigner displaced the split by locating it between the physical and the conscious (but what is a conscious being?), and so on. Also other proposed solutions to the problem, notably certain versions of many-worlds interpretations, suffer from analogous ambiguities.

It is not our task to review here the various attempts to solve the above difficulties. One can find many exhaustive treatments of this problem in the literature. On the contrary, we would like to discuss how the macro-objectification problem is indeed a consequence of very general, in fact unavoidable, assumptions on the nature of measurements, and not specifically of the assumptions of von Neumann's model. This was established in a series of theorems of increasing generality, notably the ones by Fine [1970], Shimony [1974], Brown [1986] and Busch and Shimony [1996]. Possibly the most general and direct proof is given by Bassi and Ghirardi [2000], whose results we briefly summarize. The assumptions of the theorem are:

(i) that a microsystem can be prepared in two different eigenstates of an observable (such as, e.g., the spin component along the z-axis) and in a superposition of two such states;

(ii) that one has a sufficiently reliable way of ‘measuring’ such an observable, meaning that when the measurement is triggered by each of the two above eigenstates, the process leads in the vast majority of cases to macroscopically and perceptually different situations of the universe. This requirement allows for cases in which the experimenter does not have perfect control of the apparatus, the apparatus is entangled with the rest of the universe, the apparatus makes mistakes, or the measured system is altered or even destroyed in the measurement process;

(iii) that all natural processes obey the linear laws of the theory.

From these very general assumptions one can show that, repeating the measurement on systems prepared in the superposition of the two given eigenstates, in the great majority of cases one ends up in a superposition of macroscopically and perceptually different situations of the whole universe. If one wishes to have an acceptable final situation, one mirroring the fact that we have definite perceptions, one is arguably compelled to break the linearity of the theory at an appropriate stage.

4. The Birth of Collapse Theories

The debate on the macro-objectification problem continued for many years after the early days of quantum mechanics. In the early 1950s an important step was taken by D. Bohm who presented [Bohm, 1952] a mathematically precise deterministic completion of quantum mechanics (see the entry on Bohmian Mechanics). In the area of Collapse Theories, one should mention the contribution by Bohm and Bub [1966], which was based on the interaction of the statevector with Wiener -- Siegel hidden variables. But let us come to Collapse Theories in the sense currently attached to this expression.

Various investigations during the 1970s can be considered as preliminary steps for the subsequent developments. In the years 1970-1973 L. Fonda, A. Rimini, T. Weber and myself were seriously concerned with quantum decay processes and in particular with the possibility of deriving, within a quantum context, the exponential decay law [Fonda, Ghirardi, Rimini and Weber; 1973, Fonda *et al.*, 1978]. Some features of this approach are extremely relevant for the DRP. Let us list them:

- One deals with individual physical systems;
- The statevector is supposed to undergo random processes at random times, inducing sudden changes driving it either within the linear manifold of the unstable state or within the one of the decay products;
- To make the treatment quite general (the apparatus does not know which kind of unstable system it is testing) one is led to identify the random processes with localization processes of the relative coordinates of the decay fragments. Such an assumption, combined with the peculiar resonant

dynamics characterizing an unstable system, yields, completely generally, the desired result. The 'relative position basis' is the preferred basis of this theory;

- We have applied analogous ideas to measurement processes [Fonda, Ghirardi and Rimini, 1973];
- The final equation for the evolution at the ensemble level is of the quantum dynamical semigroup type and has a structure extremely similar to the final one of the GRW theory.

Obviously, in these papers the reduction processes which are involved were not assumed to be 'spontaneous and fundamental' natural processes, but due to system-environment interactions.

Almost in the same years, P. Pearle [Pearle, 1976,1979], and subsequently N. Gisin [Gisin, 1984] and others, had entertained the idea of accounting for the reduction process in terms of a stochastic differential equation. However, they had not given any general suggestion about how to identify the states to which the dynamical equation should lead. Indeed, these states were assumed to depend on the particular measurement process one was considering. Without a clear indication on this point there was no way to identify a mechanism whose effect could be negligible for microsystems but extremely relevant for the macroscopic ones. N. Gisin gave subsequently an extremely interesting proof [Gisin, 1989] that nonlinear modifications of the standard equation without stochasticity are unacceptable since they imply the possibility of sending superluminal signals. Soon afterwards, R. Grassi and myself [Ghirardi and Grassi, 1991] showed that stochastic modifications without nonlinearity can at most induce ensemble and not individual reductions, i.e., they do not guarantee that the state vector of each individual physical system is driven in a manifold corresponding to definite properties.

5. The Original Collapse Model

As already mentioned, the Collapse Theory [Ghirardi, Rimini and Weber, 1986] we are going to describe amounts to accepting a modification of the standard evolution law of the theory such that microprocesses and macroprocesses are governed by a unique dynamics. Such a dynamics must imply that the micro-macro interaction in a measurement process leads to WPR. Bearing this in mind, recall that the characteristic feature distinguishing quantum evolution from WPR is that, while Schrödinger's equation is linear and deterministic (at the wave function level), WPR is nonlinear and stochastic. It is then natural to consider, as was suggested in the above papers, the possibility of nonlinear and stochastic modifications of the standard Schrödinger dynamics. However, the initial attempts to implement this idea were unsatisfactory for various reasons. The first, which we have already discussed, concerns the choice of the preferred basis: if one wants to have a universal mechanism leading to reductions, to which linear manifolds should the reduction mechanism drive the statevector? Or, equivalently, which of the (generally) incompatible 'potentialities' of the standard theory should we choose to make actual? The second, referred to as the trigger problem by Pearle [Pearle, 1989], is the problem of how the reduction mechanism can become more and more effective in going from the micro to the macro domain. The solution to this problem constitutes the central feature of the Collapse Theories of the GRW type. To discuss these points, let us briefly review the first consistent Collapse model [Ghirardi, Rimini and Weber, 1985] to appear in the literature.

Within such a model, originally referred to as QMSL (Quantum Mechanics with Spontaneous Localizations), the problem of the choice of the preferred basis is solved by remarking that the most embarrassing superpositions, at the macroscopic level, are those involving different spatial locations of macroscopic objects. Actually, as Einstein has stressed [Born, 1971, p. 223], this is a crucial point which has to be faced by anybody aiming to take a macro-objective position about natural phenomena: ‘A macro-body must always have a quasi-sharply defined position in the objective description of reality’. Accordingly, QMSL considers the possibility of spontaneous processes, which are assumed to occur instantaneously and at the microscopic level, which tend to suppress the linear superpositions of differently localized states. The required trigger mechanism must then follow consistently.

The key assumption of QMSL is the following: each elementary constituent of any physical system is subjected, at random times, to random and spontaneous localization processes (which we will call hittings) around appropriate positions. To have a precise mathematical model one has to be very specific about the above assumptions; in particular one has to make explicit HOW the process works, i.e. which modifications of the wave function are induced by the localizations, WHERE it occurs, i.e. what determines the occurrence of a localization at a certain position rather than at another one, and finally WHEN, i.e. at what times, it occurs. The answers to these questions are as follows.

Let us consider a system of N distinguishable particles and let us denote by $F(\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_N)$ the coordinate representation (wave function) of the state vector (we disregard spin variables since hittings are assumed not to act on them).

(a) The answer to the question HOW is then: if a hitting occurs for the i-th particle at point $\boldsymbol{x}$, the wave function is instantaneously multiplied by a Gaussian function (appropriately normalized)

$$G(\boldsymbol{q}_i, \boldsymbol{x}) = K \exp[-(1/2\, d^2)(\boldsymbol{q}_i - \boldsymbol{x})^2],$$

where d represents the localization accuracy. Let us denote as

$$L_i(\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_N ; \boldsymbol{x}) = F(\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_N)\, G(\boldsymbol{q}_i, \boldsymbol{x})$$

the wave function immediately after the localization, as yet unnormalized.

(b) As concerns the specification of WHERE the localization occurs, it is assumed that the probability density $P(\boldsymbol{x})$ of its taking place at the point $\boldsymbol{x}$ is given by the norm of the state L_i (the length, or to be more precise, the integral of the modulus squared of the function L_i over the $3N$-dimensional space). This implies that hittings occur with higher probability at those places where, in the standard quantum description, there is a higher probability of finding the particle. Note that the above prescription introduces nonlinear and stochastic elements in the dynamics. The constant K appearing in the expression of $G(\boldsymbol{q}_i,\boldsymbol{x})$ is chosen

in such a way that the integral of $P(\boldsymbol{x})$ over the whole space equals 1.

(c) Finally, the question WHEN is answered by assuming that the hittings occur at randomly distributed times, according to a Poisson distribution, with mean frequency *f*.

It is straightforward to convince oneself that the hitting process leads, when it occurs, to the suppression of the linear superpositions of states in which the same particle is well localized at different positions separated by a distance greater than *d*. As a simple example we can consider a single particle whose wavefunction is different from zero only in two small and far apart regions *h* and *t*. Suppose that a localization occurs around *h*; the state after the hitting is then appreciably different from zero only in a region around *h* itself. A completely analogous argument holds for the case in which the hitting takes place around *t*. As concerns points which are far from both *h* and *t*, one easily sees that the probability density for such hittings , according to the multiplication rule determining L_i, turns out to be practically zero, and moreover, that if such a hitting were to occur, after the wave function is normalized, the wave function of the system would remain almost unchanged.

We can now discuss the most important feature of the theory, i.e., the Trigger Mechanism. To understand the way in which the spontaneous localization mechanism is enhanced by increasing the number of particles which are in far apart spatial regions (as compared to *d*), one can consider, for simplicity, the superposition |*S*>, with equal weights, of two macroscopic pointer states |*H*> and |*T*>, corresponding to two different pointer positions *H* and *T*, respectively. Taking into account that the pointer is ‘almost rigid’ and contains a macroscopic number *N* of microscopic constituents, the state can be written, in obvious notation, as:

(4) $|S\rangle = [|1 \text{ near } h_1\rangle \ldots |N \text{ near } h_N\rangle + |1 \text{ near } t_1\rangle \ldots |N \text{ near } t_N\rangle]$,

where h_i is near *H*, and t_i is near *T*. The states appearing in first term on the right-hand side of Eq. (4) have coordinate representations which are different from zero only when their arguments (1,...,*N*) are all near *H*, while those of the second term are different from zero only when they are all near *T*. It is now evident that if any of the particles (say, the *i*-th particle) undergoes a hitting process, e.g. near the point h_i, the multiplication prescription leads practically to the suppression of the second term in (4). Thus any spontaneous localization of any of the constituents amounts to a localization of the pointer. The hitting frequency is therefore effectively amplified proportionally to the number of constituents. Notice that, for simplicity, the argument makes reference to an almost rigid body, i.e. to one for which all particles are around *H* in one of the states of the superposition and around *T* in the other. It should however be obvious that what really matters in amplifying the reductions is the number of particles which are in different positions in the two states appearing in the superposition itself.

Under these premises we can now proceed to choose the parameters *d* and *f* of the theory, i.e., the localization accuracy and the mean localization frequency. The argument just given allows one to understand how one can choose the parameters in such a way that the quantum predictions for microscopic systems remain fully valid while the embarrassing macroscopic superpositions in

measurement-like situations are suppressed in very short times. Accordingly, as a consequence of the unified dynamics governing all physical processes, individual macroscopic objects acquire definite macroscopic properties. The choice suggested in the GRW-model is:

(5) $f = 10^{-16}\ s^{-1}$

$d = 10^{-5}$ cm

It follows that a microscopic system undergoes a localization, on average, every hundred million years, while a macroscopic one undergoes a localization every 10^{-7} seconds. With reference to the challenging version of the macro-objectification problem presented by Schrödinger with the famous example of his cat, J.S. Bell comments [Bell, 1987, p.44]: [within QMSL] *the cat is not both dead and alive for more than a split second* . Besides the extremely low frequency of the hittings for microscopic systems, also the fact that the localization width is large compared to the dimensions of atoms (so that even when a localization occurs it does very little violence to the internal economy of an atom) plays an important role in guaranteeing that no violation of well-tested quantum mechanical predictions is implied by the modified dynamics.

Some remarks are appropriate. First of all, QMSL, being precisely formulated, allows to locate precisely the ‘split’ between micro and macro, reversible and irreversible, quantum and classical. The transition between the two types of ‘regimes’ is governed by the number of particles which are well localized at positions further apart than 10^{-5} cm in the two states whose coherence is going to be dynamically suppressed. Second, the model is, in principle, testable against quantum mechanics. As a matter of fact, an essential part of the program consists in proving that its predictions do not contradict any established fact about microsystems and macrosystems.

6. The Continuous Spontaneous Localization Model (CSL)

The model just presented (QMSL) has a serious drawback: it does not allow to deal with systems containing identical constituents because it does not respect the symmetry or antisymmetry requirements for such particles. A quite natural idea to overcome this difficulty would be that of relating the hitting process not to the individual particles but to the particle number density averaged over an appropriate volume. However, incorporating this idea in the QMSL scheme would require the introduction of a new constant besides the two which already appear in the model.

A more satisfactory solution to this problem (see however the remarks at the end of this subsection) can be obtained by injecting the physically appropriate principles of the GRW model within the approach of P. Pearle. This line of thought has led to a quite elegant formulation of a dynamical reduction model, usually referred to as CSL [Pearle, 1989; Ghirardi, Pearle and Rimini, 1990] in which the discontinuous jumps which characterize QMSL are replaced by a continuous stochastic evolution in the Hilbert space (a

sort of Brownian motion of the statevector).

We will not enter into the rather technical details of this interesting development of the original GRW proposal, since the basic ideas and physical implications are precisely the same as those of the original formulation. Actually, one could argue that the above idea of tackling the problem of identical particles by considering the average particle number within an appropriate volume is correct. In fact it has been proved [Ghirardi, Pearle and Rimini, 1990] that for any CSL dynamics there is a hitting dynamics which, from a physical point of view, is 'as close to it as one wants'. Instead of entering into the details of the CSL formalism, it is useful, for the discussion below, to analyze a simplified version.

7. A Simplified Version of CSL

With the aim of understanding the physical implications of the CSL model, such as the rate of suppression of coherence, we make now some simplifying assumptions. First, we assume that we are dealing with only one kind of particles (e.g., the nucleons), secondly, we disregard the standard Schrödinger term in the evolution and, finally, we divide the whole space in cells of volume d^3. We denote by $|n_1, n_2, ... >$ a state in which there are n_i particles in cell i, and we consider a superposition of two states $|n_1, n_2, ... >$ and $|m_1, m_2, ... >$ which differ in the occupation numbers of the various cells of the universe. With these assumptions it is quite easy to prove that the rate of suppression of the coherence between the two states (so that the final state is one of the two and not their superposition) is governed by the quantity:

$$(6)\ \exp\{-f[(n_1 - m_1)^2 + (n_2 - m_2)^2 + ...]t\},$$

the sum being extended to all cells in the universe. Apart from differences relating to the identity of the constituents, the overall physics is quite similar to that implied by QMSL. Obviously, there are interesting physical implications which deserve to be discussed. A detailed analysis has been presented in [Ghirardi and Rimini, 1990]. As shown there and as follows from estimates about possible effects for superconducting devices [Rae, 1990; Gallis and Fleming, 1990; Rimini, 1995], and for the excitation of atoms [Squires, 1991], it turns out not to be possible, with present technology, to perform clear-cut experiments allowing to discriminate the model from standard quantum mechanics [Benatti *et al.*, 1995].

There is however an interesting aspect which might be relevant to idea of relating the suppression of coherence to gravitational effects. Given Eq. (6), notice that the worst case scenario (from the point of view of the time necessary to suppress coherence) is the superposition of two states for which the occupation numbers of the individual cells differ only by one unit. Indeed, in this case the amplifying effect of taking the square of the differences disappears. Let us then raise the question: how many nucleons (at worst) should occupy different cells, in order for the given superposition to be dynamically suppressed within the time which characterizes human perceptual processes? Since such a time is of the order of 10^{-2} sec and $f = 10^{-16}$ sec^{-1}, the number of displaced nucleons must be of the order of 10^{18}, which corresponds, to a remarkable accuracy, to a Planck mass. This figure seems to point in the same

direction as attempts such as Penrose's attempts to relate reduction mechanisms to quantum gravitational effects [Penrose, 1989].

8. Achievements of Collapse Theories

A. Pais famously recalls in his biography of Einstein:

> We often discussed his notions on objective reality. I recall that during one walk Einstein suddenly stopped, turned to me and asked whether I really believed that the moon exists only when I look at it [Pais, 1982, p. 5].

In the context of Einstein's remarks in *Albert Einstein, Philosopher-Scientist* [Schilpp, 1949], we can regard this reference to the moon as an extreme example of 'a fact that belongs entirely within the sphere of macroscopic concepts', as is also a mark on a strip of paper that is used to register the outcome of a decay experiment, so that

> as a consequence, there is hardly likely to be anyone who would be inclined to consider seriously [...] that the existence of the location is essentially dependent upon the carrying out of an observation made on the registration strip. For, in the macroscopic sphere it simply is considered certain that one must adhere to the program of a realistic description in space and time; whereas in the sphere of microscopic situations one is more readily inclined to give up, or at least to modify, this program [p. 671].

However,

> the 'macroscopic' and the 'microscopic' are so inter-related that it appears impracticable to give up this program in the 'microscopic' alone [p. 674].

One might speculate that Einstein would not have taken the DRP seriously, given that it is a fundamentally indeterministic program. On the other hand, the DRP allows precisely for this middle ground, between giving up a 'realistic description in space and time' altogether (the moon is not there when nobody looks), and requiring that it be applicable also at the microscopic level (as some kind of 'hidden variables' theory). It would seem that the pursuit of 'realism' for Einstein was more a program that had been very successful rather than an a priori commitment, and that in principle he would have welcomed attempts to give up or weaken microrealistic requirements, provided they allowed one to adopt a macrorealist position.

In the DRP, we can say of an electron in an EPR-Bohm situation that 'when nobody looks', it has no definite spin in any direction , and in particular that when it is in a superposition of two states localised far away from each other, it cannot be thought to be at a definite place. In the macrorealm, however, objects do have definite positions and are generally describable in classical terms. That is, the DRP

program is not adding ‘hidden variables’ to the theory, but the moon is definitely there even if no sentient being has ever looked at it, or, in the words of J. S. Bell, the DRP

> allows electrons (in general microsystems) to enjoy the cloudiness of waves, while allowing tables and chairs, and ourselves, and black marks on photographs, to be rather definitely in one place rather than another, and to be described in classical terms [Bell, 1986, p. 364].

Such a program, as we have seen, is realized by assuming only the existence of wave functions, and by proposing a unified dynamics that will govern both microscopic processes and ‘measurements’. As regards the latter, no vague definitions are needed in order to apply the theory. The equations are followed exactly, and the macroscopic ambiguities that would arise from the linear evolution are theoretically possible, but only of momentary duration, and thus arguably of no practical importance and no source of embarrassment.

We have not analyzed yet the implications about locality, but since in the DRP program no hidden variables are introduced, the situation can be no worse that in ordinary quantum mechanics: *‘by adding mathematical precision to the jumps in the wave function, it simply makes precise the action at a distance of ordinary quantum mechanics’* [Bell, 1987, p. 46]. Indeed, a detailed investigation of the locality properties of the theory becomes possible, and one can investigate whether the theory represents an approximation to a relativistically invariant theory. The analysis carried on so far, however, proves that at least in the non-relativistic version a program of dynamical reduction can be consistently developed. Moreover, as it will become clear when we will discuss the interpretation of the theory in terms of mass density, the QMSL and CSL theories lead in a natural way to attach definite properties in space and time to macroscopic objects, the main objective of Einstein’s requirements.

The achievements of the DRP which are relevant for the debate about the foundations of quantum mechanics can also be concisely summarized in the words of H.P. Stapp:

> The collapse mechanisms so far proposed could, on the one hand, be viewed as ad hoc mutilations designed to force ontology to kneel to prejudice. On the other hand, these proposals show that one can certainly erect a coherent quantum ontology that generally conforms to ordinary ideas at the macroscopic level [Stapp, 1989, p. 157].

9. Relativistic Dynamical Reduction Models

When confronted with a new theoretical scheme, particularly with one which, as we have seen [Bell, 1987], ‘makes precise the action at a distance of ordinary quantum mechanics’, one is naturally led to raise the question of whether it represents an approximation to a relativistically invariant theory. In this connection it is useful to mention, first of all, some interesting recent investigations of the non-local aspects of CSL.

As is well known, [Suppes and Zanotti, 1976; van Fraassen, 1982; Jarrett, 1984; Shimony, 1983; see also the entry on Bell's Theorem], Bell's locality assumption is equivalent to the conjunction of two other assumptions, viz., in Shimony's terminology, parameter independence and outcome independence. In view of the experimental violation of Bell's inequality, one has to give up either or both of these assumptions. The above splitting of the locality requirement into two logically independent conditions is particularly useful in discussing the different status of CSL and deterministic hidden variable theories with respect to relativistic requirements. In fact, as proved by Jarrett himself, when parameter independence is violated, if one had access to the variables which specify completely the state of individual physical systems, one could send faster-than-light signals from one wing of the apparatus to the other. Moreover, in ref. [Ghirardi and Grassi, 1994, 1996] it has been shown that it is impossible to build a genuine relativistically invariant theory which, in its nonrelativistic limit, exhibits parameter dependence and does not entail backward causation. On the other hand, if locality is violated only by the occurrence of outcome dependence then faster-than-light signaling cannot be achieved.

Now, it is well known that any deterministic theory (i.e., one in which the range of all probability distributions for outcomes is the set {0,1}) that reproduces quantum predictions must exhibit parameter dependence. This fact by itself suggest that such theories will certainly meet more serious difficulties with relativity than theories like standard quantum mechanics which violate only outcome independence and which do not allow faster-than-light signaling [Eberhard, 1978; Ghirardi, Rimini and Weber, 1980; Ghirardi, Grassi, Rimini and Weber, 1988]. What about CSL? It has been possible to prove [Ghirardi, Grassi, Butterfield and Fleming, 1993; Butterfield *et al*., 1993] that it, too, violates Bell's locality only by violating outcome independence. This is to some extent encouraging; even though, as we will be led to conclude, it seems extremely difficult to build a relativistic model inducing reductions, this result shows that there are no reasons of principle making such a project unviable.

Let us be more specific. The first attempt to obtain a relativistic generalization of dynamical reduction models was presented in ref. [Pearle, 1990]. It should be stressed that having individual reductions prevents the theory from being invariant at the individual level (note that QMSL and CSL are not even Galilei invariant at the individual level). Thus one is led to introduce a generalization of the invariance requirement: the theory must be stochastically invariant. This means that, even though the individual processes may look different to different observers, any two of them will agree on the composition of the final ensemble for (subjectively) the same initial conditions. We remark that it is precisely in this sense that both QMSL and CSL turn out to be Galilei invariant.

Pearle [1990] considered a fermion field coupled to a meson field and has put forward the idea of inducing localizations for the fermions through their coupling to the mesons and a stochastic dynamical reduction mechanism acting on the meson variables. He considered Heisenberg evolution equations for the coupled fields and a Tomonaga-Schwinger CSL-type evolution equation with a skew-hermitian coupling to a c-number stochastic potential for the state vector. This approach has been systematically investigated in refs. [Ghirardi, Grassi and Pearle, 1990a, 1990b] to which we refer the reader for a detailed discussion. Here we limit ourselves to stressing that, under certain approximations, one obtains in the non-relativistic limit a CSL-type equation inducing spatial localization. However, due to the white

noise nature of the stochastic potential, novel renormalization problems arise: the increase per unit time and per unit volume in the energy of the meson field is infinite due to the fact that infinitely many mesons are created. For the reasons we have just discussed one cannot say that the possibility of generalizing CSL to the relativistic case has been established. Not even more recent attempts have succeeded in overcoming these difficulties.

Nevertheless, the efforts which have been spent on such a program have led to a better understanding of some points and have thrown light on important conceptual issues. First, they have led to a completely general and rigorous formulation of the concept of stochastic invariance [Ghirardi, Grassi and Pearle, 1990a]. Second, they have prompted a critical reconsideration, based on the discussion of smeared observables with compact support, of the problem of locality at the individual level. This analysis has brought out the necessity of reconsidering the criteria for the attribution of objective local properties to physical systems. In specific situations, one cannot attribute any local property to a microsystem: any attempt to do so gives rise to ambiguities. However, in the case of macroscopic systems, the impossibility of attributing to them local properties (or, equivalently, the ambiguity surrounding such properties) lasts only for time intervals of the order of those necessary for the dynamical reduction to take place. Moreover, no objective property corresponding to a local observable, even for microsystems, can emerge as a consequence of a measurement-like event occurring in a space-like separated region: such properties emerge only in the future light cone of the macroscopic event considered. Finally, recent investigations [Ghirardi and Grassi, 1994, 1996; Ghirardi, 1996, 2000] have shown that the very formal structure of the theory is such that it does not allow, even conceptually, to establish cause-effect relations between space-like events.

Having listed some interesting results obtained along these lines, in concluding this section it is necessary to stress once more the immense difficulties that the program of a relativistic generalization has met until now. The question of whether such a program will find a satisfactory formulation still remains ‘the big problem’ for this type of investigations.

10. Collapse Theories and Definite Perceptions

Some authors [Albert and Vaidman, 1989; Albert, 1990, 1992] have raised an interesting objection concerning the emergence of definite perceptions within Collapse Theories. The objection is based on the fact that one can easily imagine situations leading to definite perceptions, that nevertheless do not involve the displacement of a large number of particles up to the stage of the perception itself. These cases would then constitute actual measurement situations which cannot be described by QMSL, contrary to what happens for the idealized (according to the authors) situations considered in QMSL, i.e. those involving the displacement of some sort of pointer. To be more specific, the above papers considered a ‘measurement’ process whose output is the emission of a burst of photons. This can easily be devised by considering, e.g., a Stern-Gerlach set-up in which the two paths followed by the microsystem according to the value of its spin component hit a fluorescent screen and excite a small number of atoms which subsequently decay, emitting a small number of photons. The argument goes as follows: since only a few atoms are excited, since the excitations involve displacements which are smaller than the characteristic

localization distance of QMSL, since QMSL does not induce reductions on photon states and, finally, since the photon states immediately overlap, there is no way for the spontaneous localization mechanism to become effective. The superposition of the states ‘photons emerging from point *A* of the screen’ and ‘photons emerging from point *B* of the screen’ will last for a long time. On the other hand, since the visual perception threshold is quite low (about 6-7 photons), there is no doubt that the naked eye of a human observer is sufficient to detect whether the luminous spot on the screen is at *A* or at *B*. The conclusion follows: in the case under consideration no dynamical reduction can take place and as a consequence no measurement is over, no outcome is definite, up to the moment in which a conscious observer perceives the spot.

We have presented a detailed answer to this criticism [Aicardi *et al*., 1991]. The crucial points of our argument are the following: we perfectly agree that in the case considered the superposition persists for long times (actually the superposition must persist, since, the system under consideration being microscopic, one could perform interference experiments which everybody would expect to confirm quantum mechanics). However, to deal in the appropriate and correct way with such a criticism, one has to consider all the systems which enter into play (electron, screen, photons and brain) and the universal dynamics governing all relevant physical processes. A simple estimate of the number of ions which are involved in the visual perception mechanism makes perfectly plausible that, in the process, a sufficient number of particles are displaced by a sufficient spatial amount to satisfy the conditions under which, according to QMSL, the suppression of the superposition of the two nervous signals will take place within the time scale of perception.

To avoid misunderstandings, this analysis by no means amounts to attributing a special role to the conscious observer or to perception. The observer’s brain is the only system present in the set-up in which a superposition of two states involving different locations of a large number of particles occurs. As such it is the only place where the reduction can and actually must take place according to the theory. It is extremely important to stress that if in place of the eye of a human being one puts in front of the photon beams a spark chamber or a device leading to the displacement of a macroscopic pointer, or producing ink spots on a computer output, reduction will equally take place. In the given example, the human nervous system is simply a physical system, a specific assembly of particles, which performs the same function as one of these devices, if no other such device interacts with the photons before the human observer does. It follows that it is incorrect and seriously misleading to claim that QMSL requires a conscious observer to make definite the macroscopic properties of physical systems.

A further remark may be appropriate. The above analysis could be taken by the reader as indicating a very naive and oversimplified attitude towards the deep problem of the mind-brain correspondence. There is no claim and no presumption that QMSL allows a physicalist explanation of conscious perception. It is only pointed out that, for what we know about the purely physical aspects of the process, one can state that before the nervous pulses reach the higher visual cortex, the conditions guaranteeing the suppression of one of the two signals are verified. In brief, a consistent use of the dynamical reduction mechanism in the above situation accounts for the definiteness of the conscious perception, even in the extremely peculiar situation devised by Albert and Vaidman.

11. The Interpretation of the Theory

As stressed in the opening sentences of this contribution, the most serious problem of standard quantum mechanics lies in its being extremely successful in telling us about *what we observe*, but being basically silent on *what is*. This specific feature is closely related to the probabilistic interpretation of the statevector, combined with the completeness assumption of the theory. Notice that what is under discussion is the probabilistic interpretation, not the probabilistic character, of the theory. Also collapse theories have a fundamentally stochastic character, but, due to their most specific feature, i.e. that of driving the statevector of any individual physical system into appropriate and physically meaningful manifolds, they allow for a different interpretation. In fact, one could say (if one wants to avoid that they too, as the standard theory, speak only of *what we find*) that they impose a different interpretation, one that accounts for our perceptions at the appropriate, i.e. macroscopic, level.

The question of the correct interpretation of the theory has been the subject of debate, some of the principal approaches having originated with J. S. Bell. Given that the wavefunction itself is an object in the (higher-dimensional) configuration space, Bell was particularly keen to identify what could be taken as some kind of ‘local beable’, from which one could obtain a representation of the perceived reality in ordinary three-dimensional space. In the specific context of QMSL, Bell [1987, p. 45] suggested that the ‘GRW jumps’, which we called ‘hittings’ above, could play this role. Later, he suggested that the most natural interpretation for the wavefunction in the context of a collapse theory would be as describing the ‘density [...] of stuff’ in configuration space [Bell, 1990, p. 30]. The interpretation which, in the opinion of the present writer, is most appropriate for collapse theories [Ghirardi, Grassi and Benatti, 1995, Ghirardi, 1997a, 1997b] was ultimately developed from this suggestion, together with the firm conviction that an acceptable interpretation should establish precise links between our formal description of physical processes and the events taking place in the three-dimensional space we ‘see’ around us.

First of all, various investigations [Pearle and Squires, 1994] had made clear that QMSL and CSL needed a modification, i.e., the characteristic localization frequency of the elementary constituents of matter had to be made proportional to the mass characterizing the particle under consideration. In particular, the original frequency for the hitting processes $f = 10^{-16}$ sec^{-1} is the one characterizing the nucleons, while, e.g., electrons would suffer hittings with a frequency reduced by about 2000 times. Unfortunately we have no space to discuss here the physical reasons which make this choice appropriate; we refer the reader to the above paper, as well as to the recent detailed analysis by Peruzzi and Rimini [Peruzzi and Rimini, 2000]. With this modification, what the nonlinear dynamics strives to make ‘objectively definite’ is the average mass distribution in the whole universe (averaged over appropriate volumes associated with the characteristic localization accuracy of the theory). Second, a deep critical reconsideration [Ghirardi, Grassi and Benatti, 1995] has made evident how the concept of ‘distance’ that characterizes the Hilbert space is inappropriate in accounting for the similarity or difference between macroscopic situations. Just to give a convincing example, consider three states |*h*>, |*h**> and |*t*> of a macrosystem (let us say a massive macroscopic bulk of matter), the first corresponding to its being located here, the second to its having the same location but one of its atoms (or molecules) being in a state orthogonal to the corresponding state in |*h*>, and the third having exactly the same internal state of the first but being

differently located (there). Then, despite the fact that the first two states are practically indistinguishable from each other at the macrolevel, while the first and the third correspond to completely different and directly perceivable situations, the Hilbert space distance between $|h\rangle$ and $|h^*\rangle$, is equal to that between $|h\rangle$ and $|t\rangle$.

When the localization frequency is related to the mass of the constituents, then, as above and completely generally (i.e. even when one is dealing with a body which is not almost rigid, such as a gas or a cloud), the mechanism leading to the suppression of the superpositions of macroscopically different states is fundamentally governed by the sum (or the integral) of the squared differences of the mass densities associated to the two superposed states, averaged over the characteristic volume of the theory, i.e., 10^{-15} cm^3 . This suggests taking the following attitude: what the theory is about, what is real ‘out there’ at a given space point **x**, is just the average mass density in the characteristic volume around **x** :

$$(7)\ M(\boldsymbol{x},t) = \langle F,t \mid M(\boldsymbol{x}) \mid F,t \rangle,$$

where $M(\boldsymbol{x})$ is the mass density operator corresponding to the given volume around $\boldsymbol{x}$ and $|F,t\rangle$ is the statevector characterizing the system at the given time. It is obvious that within standard quantum mechanics such a function cannot be endowed with any objective physical meaning due to the occurrence of linear superpositions of macroscopically different mass distributions which give rise to values that do not correspond to what we find in a measurement process or what we perceive (typically the equal weight superposition of the states $|h\rangle$ and $|t\rangle$ will give rise to a mass density distribution which is 1/2 of the actual one both ‘here’ and ‘there’). But in a CSL model relating reductions to mass density differences, the dynamics, as we have seen, suppresses in extremely short times these embarrassing superpositions. In this case, if one considers only the states allowed by the dynamics one can give a description of the world in terms of $M(\boldsymbol{x},t)$, i.e., one recovers a physically meaningful account of physical reality in the usual 3-dimensional space and time. Resorting to the quantity (7) one can also define an appropriate ‘distance’ between two states as the integral over the whole 3-dimensional space of the square of the difference of $M(\boldsymbol{x},t)$ for the two given states, a quantity which turns out to be perfectly appropriate to ground the concept of macroscopically similar or distinguishable Hilbert space states. In turn this distance can be used as a basis to define a sensible psychophysical correspondence within the theory.

12. The Problem of the Tails of the Wave Function

In recent years, there has been a lively debate around a problem which has its origin, according to some of the authors which have raised it, in the fact that even though the localization process, which corresponds to multiplying the wave function times a Gaussian, and thus leads to wave functions strongly peaked around the position of the hitting, it leads to wave functions that are different from zero over the whole of space. The first criticism of this kind was raised by A. Shimony [Shimony, 1990] and can be summarized by his sentence,

> one should not tolerate tails in wave functions which are so broad that their different parts can be discriminated by the senses, even if very low probability amplitude is assigned to them.

After a localization of a macroscopic system, typically the pointer of the apparatus, its centre of mass will be associated to a wave function which is different from zero over the whole space. If one adopts the probabilistic interpretation of the standard theory, this means that even when the measurement process is over, there is a nonzero (even though extremely small) probability of finding its pointer in an arbitrary position, instead of the one corresponding to the registered outcome. This is taken as inacceptable, as indicating that the DRP does not actually overcome the macro-objectification problem.

Let us state immediately that the (alleged) problem arises entirely from keeping the standard interpretation of the wave function unchanged, in particular assuming that its modulus squared gives the probability density of the position variable. However, as we have discussed in the previous section, there are much more serious reasons of principle which require to abandon the probabilistic interpretation and replace it either with one of those proposed by Bell, or, more appropriately in our opinion, with the mass density interpretation have outlined above.

Before entering into a detailed discussion of this subtle point we need to focus the problem better. We cannot avoid making two remarks. Suppose one adopts, for the moment, the conventional quantum position. We agree that, within such a framework, the fact that wave functions never have strictly compact spatial support can be considered puzzling. However this is a problem arising directly from the mathematical features (spreading of wave functions) and from the probabilistic interpretation of the theory, and not at all a problem peculiar to the dynamical reduction models. Indeed, the fact that, e.g., the wave function of the centre of mass of a pointer or of a table has not a compact support has never been taken to be a problem for standard quantum mechanics. When the wave function is extremely well peaked around a given point in space, it has always been accepted that it describes a table located at a certain position, and that this corresponds in some way to our perception of it. It is obviously true that, for the given wave function, the quantum rules entail that if a measurement were performed the table could be found (with an extremely small probability) to be kilometers far away, but this *is not* the measurement or the macro-objectification problem of the standard theory. The latter concerns a completely different situation, i.e., that in which one is confronted with a superposition with comparable weights of two macroscopically separated wave functions, both of which possess tails (i.e., have non-compact support) but are appreciably different from zero only in very narrow intervals. This is the really embarrassing situation which conventional quantum mechanics is unable to make understandable. To which perception of the position of the table does this wave function correspond?

The implications for this problem of the adoption of the QMSL theory should be obvious. Within QMSL, the superposition of two states which, when considered individually, are assumed to lead to different and definite perceptions of macroscopic locations, are dynamically forbidden. If some process tends to produce such superpositions, then the reducing dynamics induces the localization of the centre of mass (the associated wave function being appreciably different from zero only in a narrow and precise interval). Correspondingly, the possibility arises of attributing to the system the property of being in a

definite place and thus of accounting for our definite perception of it. Summarizing, we stress once more that the criticism about the tails as well as the requirement that the appearance of macroscopically extended (even though extremely small) tails be strictly forbidden is exclusively motivated by uncritically committing oneself to the probabilistic interpretation of the theory, even for what concerns the psycho-physical correspondence: states assigning non-exactly vanishing probabilities to different outcomes of position measurements must correspond to ambiguous perceptions about these positions. Since neither within the standard formalism nor within the framework of dynamical reduction models a wave function can have compact support, taking such a position leads to conclude that it is just the Hilbert space description of physical systems which has to be given up.

It ought to be stressed that there is nothing in the GRW theory which would make the choice of functions with compact support problematic for the purpose of the localizations, but it also has to be noted that following this line would be totally useless: since the evolution equation contains the kinetic energy term, any function, even if it has compact support at a given time, will instantaneously spread acquiring a tail extending over the whole of space. If one sticks to the probabilistic intepretation and one accepts the completeness of the description of the states of physical systems in terms of the wave function, the tail problem cannot be avoided.

The solution to the tails problem can only derive from abandoning completely the probabilistic interpretation and from adopting a more physical and realistic intepretation relating ‘what is out there’ to, e.g., the mass density distribution over the whole universe. In this connection, the following example will be instructive [Ghirardi, Grassi and Benatti, 1995]. Take a massive sphere of normal density and mass of about 1 kg. Classically, the mass of this body would be totally concentrated within the radius of the sphere, call it r. In QMSL, after the extremely short time interval in which the collapse dynamics leads to a ‘regime’ situation, and if one considers a sphere with radius $r + 10^{-5}$ cm, the integral of the mass density over the rest of space turns out to be an incredibly small fraction (of the order of 1 over 10 to the power 10^{15}) of the mass of a single proton. In such conditions, it seems quite legitimate to claim that the macroscopic body is localised within the sphere.

However, also this quite reasonable position has been questioned and it has been claimed [Lewis, 1997], that the very existence of the tails implies that the enumeration principle (i.e. the fact that the claim ‘particle 1 is within this box & particle 2 is within this box & ... & particle *n* is within this box & no other particle is within this box’ implies the claim ‘there are *n* particles within this box’) does not hold, if one takes seriously the mass density interpretation of collapse theories. This paper has given rise to a long debate which would be inappropriate to reproduce here. We refer the reader to the following papers [Ghirardi and Bassi, 1999; Clifton and Monton, 1999a, 1999b; Bassi and Ghirardi, 1999, 2001]. Various arguments have been presented in favour and against the criticism by Lewis.

We would like to conclude this brief analysis by stressing once more that, in our opinion, all the disagreements and the misunderstandings concerning this problem have their origin in the fact that the idea that the probabilistic interpretation of the wave function must be abandoned has not been fully accepted by the authors who find some difficulties in the proposed mass density intepretation of the

Collapse Theories.

Summary

We hope to have succeeded in giving a clear picture of the ideas, the implications, the achievements and the problems of the DRP. We conclude by stressing once more our position with respect to the Collapse Theories. Their interest derives entirely from the fact that they have given some hints about a possible way out from the difficulties characterizing standard quantum mechanics, by proving that explicit and precise models can be worked out which agree with all known predictions of the theory and nevertheless allow, on the basis of a universal dynamics governing all natural processes, to overcome in a mathematically clean and precise way the basic problems of the standard theory. In particular, the Collapse Models show how one can work out a theory that makes perfectly legitimate to take a macrorealistic position about natural processes, without contradicting any of the experimentally tested predictions of standard quantum mechanics.

Bibliography

- Aicardi, F., Borsellino, A., Ghirardi, G.C., and Grassi, R. [1991], 'Dynamic models for state-vector reduction - Do they ensure that measurements have outcomes?', *Foundations of Physics Letters*, **4**, 109.
- Albert, D.Z. [1990], 'On the Collapse of the Wave Function', in *Sixty-Two Years of Uncertainty*, A. Miller (ed.), Plenum, New York.
- -----. [1992], *Quantum Mechanics and Experience*, Harvard University Press, Cambridge, Mass.
- Albert, D.Z., and Vaidman, L. [1989], 'On a proposed postulate of state reduction', , *Physics Letters*, **A139**, 1.
- Bassi, A., and Ghirardi, G.C. [1999], 'More about dynamical reduction and the enumeration principle', *British Journal for the Philosophy of Science*, **50**, 719.
- -----. [2000], 'A general argument against the universal validity of the superposition principle', *Physics Letters*, **A 275**, 373.
- -----. [2001], 'Counting marbles: Reply to Clifton and Monton', *British Journal for the Philosophy of Science*, **52**, 125.
- Bell, J.S. [1986], 'Six possible worlds of quantum mechanics', in *Proceedings of the Nobel Symposium 65: Possible Worlds in Arts and Sciences*, de Gruyter, New York.
- -----. [1987], 'Are there quantum jumps?', in *Schrödinger -- Centenary Celebration of a Polymath*, C.W. Kilmister (ed.), Cambridge University Press, Cambridge.
- -----. [1990], 'Against "measurement"', in *Sixty-Two Years of Uncertainty*, A. Miller (ed.), Plenum, New York.
- Benatti, F., Ghirardi, G.C., and Grassi, R. [1995], 'Quantum Mechanics with Spontaneous Localization and Experiments', in *Advances in quantum Phenomena*, E. Beltrametti *et al*. (eds), Plenum, New York.
- Bohm, D. [1952], 'A suggested interpretation of the quantum theory in terms of hidden variables.

I & II.' *Physical Review*, **85** , 166, *ibid*., **85**, 180.

- Bohm, D., and Bub, J. [1966], 'A proposed solution of the measurement problem in quantum mechanics by a hidden variable theory', *Reviews of Modern Physics*, **38**, 453.
- Born, M. [1971], *The Born-Einstein Letters*, Walter and Co., New York.
- Brown, H.R. [1986], 'The insolubility proof of the quantum measurement problem', *Foundations of Physics*, **16**, 857.
- Busch, P., and Shimony, A. [1996], 'Insolubility of the quantum measurement problem for unsharp observables', *Studies in History and Philosophy of Modern Physics*, **27B**, 397.
- Butterfield, J., Fleming, G.N., Ghirardi, G.C., and Grassi, R. [1993], 'Parameter dependence in dynamical models for state-vector reduction', *International Journal of Theoretical Physics*, **32**, 2287.
- Clifton, R., and Monton, B. [1999a], 'Losing your marbles in wavefunction collapse theories', *British Journal for the Philosophy of Science*, **50**, 697.
- -----. [1999b], 'Counting marbles with "accessible" mass density: A reply to Bassi and Ghirardi', *British Journal for the Philosophy of Science*, **51**, 155.
- Dirac, P.A.M. [1948], *Quantum Mechanics*, Clarendon Press, Oxford.
- Eberhard, P. [1978], 'Bell's theorem and different concepts of locality', *Nuovo Cimento*, **46B**, 392.
- Fine, A. [1970], 'Insolubility of the quantum measurement problem', *Physical Review*, **D2**, 2783.
- Fonda, L., Ghirardi, G.C., and Rimini A. [1973], 'Evolution of quantum systems subject to random measurements', *Nuovo Cimento*, **18B**, 1.
- -----. [1978], 'Decay theory of unstable quantum systems', *Reports on Progress in Physics*, **41**, 587.
- Fonda, L., Ghirardi, G.C., Rimini, A., and Weber, T. [1973], 'Quantum foundations of exponential decay law', *Nuovo Cimento*, **15A**, 689.
- Gallis, M.R., and Fleming, G.N. [1990], 'Environmental and spontaneous localization', *Physical Review*, **A42**, 38.
- Ghirardi, G.C. [1996], 'Properties and events in a relativistic context: Revisiting the dynamical reduction program', *Foundations of Physics Letters*, **9**, 313.
- -----. [1997a], 'Quantum Dynamical Reduction and Reality: Replacing Probability Densities with Densities in Real Space', *Erkenntnis*, **45** , 349.
- -----. [1997b], 'Macroscopic Reality and the Dynamical Reduction Program', in *Structures and Norms in Science*, M.L. Dalla Chiara (ed.), Kluwer, Dordrecht.
- -----. [2000], 'Local measurements of nonlocal observables and the relativistic reduction process', *Foundations of Physics*, **30**, 1337.
- Ghirardi, G.C., and Bassi, A. [1999], 'Do dynamical reduction models imply that arithmetic does not apply to ordinary macroscopic objects', *British Journal for the Philosophy of Science*, **50**, 49.
- Ghirardi, G.C., and Grassi, R. [1991], 'Dynamical Reduction Models: some General Remarks', in *Nuovi Problemi della Logica e della Filosofia della Scienza*, D. Costantini *et al*. (eds), Editrice Clueb, Bologna.
- -----. [1994], 'Outcome predictions and property attribution - The EPR argument reconsidered', *Studies in the History and Philosophy of Science*, **25** , 397.
- -----. [1996], 'Bohm's Theory versus Dynamical Reduction', in *Bohmian Mechanics and*

Quantum Theory: an Appraisal, J. Cushing *et al*. (eds), Kluwer, Dordrecht.

- Ghirardi, G.C., Grassi, R., and Benatti, F. [1995], 'Describing the macroscopic world - Closing the circle within the dynamical reduction program', *Foundations of Physics*, **25**, 5.
- Ghirardi, G.C., Grassi, R., Butterfield, J., and Fleming, G.N. [1993], 'Parameter dependence and outcome dependence in dynamic models for state-vector reduction', *Foundations of Physics*, **23**, 341.
- Ghirardi, G.C., Grassi, R., and Pearle, P. [1990a], 'Relativistic dynamic reduction models - General framework and examples', *Foundations of Physics*, **20**, 1271.
- -----. [1990b], 'Relativistic Dynamical Reduction Models and Nonlocality', in *Symposium on the Foundations of Modern Physics 1990*, P. Lahti and P. Mittelstaedt (eds), World Scientific, Singapore.
- Ghirardi, G.C., Grassi, R., Rimini, A., and Weber, T. [1988], 'Experiments of the electron-paramagnetic-res type involving CP-violation do not allow faster-than-light communication between distant observers', *Europhysics Letters*, **6**, 95.
- Ghirardi, G.C., Pearle, P., and Rimini, A. [1990], 'Markov-processes in Hilbert-space and continuous spontaneous localization of systems of identical particles', *Physical Review*, **A42**, 78.
- Ghirardi, G.C., and Rimini, A. [1990], 'Old and New Ideas in the Theory of Quantum Measurement', in *Sixty-Two Years of Uncertainty*, A. Miller (ed.), Plenum, New York .
- Ghirardi, G.C., Rimini, A., and Weber, T. [1980], 'General argument against superluminal transmission through the quantum-mechanical measurement process', *Lettere al Nuovo Cimento*, **27**, 293.
- -----. [1985], 'A Model for a Unified Quantum Description of Macroscopic and Microscopic Systems', in *Quantum Probability and Applications*, L. Accardi *et al*. (eds), Springer, Berlin.
- -----. [1986], 'Unified dynamics for microscopic and macroscopic systems', *Physical Review*, **D 34**, 470.
- Gisin, N. [1984], 'Quantum measurements and stochastic processes', *Physical Review Letters*, **52**, 1657, and 'Reply', *ibid*., **53**, 1776.
- -----. [1989], 'Stochastic quantum dynamics and relativity', *Helvetica Physica Acta*, **62**, 363.
- Gottfried, K. [2000], 'Does Quantum Mechanics Carry the Seeds of its own Destruction?', in *Quantum Reflections*, D. Amati *et al*. (eds), Cambridge University Press, Cambridge.
- Jarrett, J.P. [1984], 'On the physical significance of the locality conditions in the Bell arguments', *Nous*, **18**, 569.
- Lewis, P. [1997], 'Quantum mechanics, orthogonality and counting', *British Journal for the Philosophy of Science*, **48**, 313.
- Pais, A. [1982], *Subtle is the Lord*, Oxford University Press, Oxford.
- Pearle, P. [1976], 'Reduction of statevector by a nonlinear Schrödinger equation', *Physical Review*, **D13**, 857.
- -----. [1979], 'Toward explaining why events occur', *International Journal of Theoretical Physics*, **18**, 489 .
- -----. [1989], *Physical Review*, 'Combining stochastic dynamical state-vector reduction with spontaneous localization', **A39**, 2277.
- -----. [1990], 'Toward a Relativistic Theory of Statevector Reduction', in *Sixty-Two Years of Uncertainty*, A. Miller (ed.), Plenum, New York.

- Pearle, P., and Squires, E. [1994], ‘Bound-state excitation, nucleon decay experiments, and models of wave-function collapse’, *Physical Review Letters*, **73**, 1.
- Penrose, R. [1989], *The Emperor’s New Mind*, Oxford University Press, Oxford.
- Peruzzi, G., and Rimini, A. [2000], ‘Compoundation invariance and Bohmian mechanics’, *Foundations of Physics*, **30**, 1445.
- Rae, A.I.M. [1990], ‘Can GRW theory be tested by experiments on SQUIDs?’, *Journal of Physics*, **A23** , 57.
- Rimini, A. [1995], ‘Spontaneous Localization and Superconductivity’, in *Advances in Quantum Phenomena*, E. Beltrametti *et al*. (eds.), Plenum, New York.
- Schrödinger, E. [1935], ‘Die gegenwärtige Situation in der Quantenmechanik’, *Naturwissenschaften*, **23**, 807.
- Schilpp, P.A. (ed.) [1949], *Albert Einstein: Philosopher-Scientist*, Tudor, New York.
- Shimony, A. [1974], ‘Approximate measurement in quantum-mechanics. 2’, *Physical Review*, **D9**, 2321.
- -----. [1983], ‘Controllable and uncontrollable non-locality’, in *Proceedings of the International Symposium on the Foundations of Quantum Mechanics*, S. Kamefuchi *et al*. (eds), Physical Society of Japan, Tokyo.
- -----. [1989], ‘Search for a worldview which can accommodate our knowledge of microphysics’, in *Philosophical Consequences of Quantum Theory*, J.T. Cushing and E. McMullin (eds), University of Notre Dame Press, Notre Dame, Indiana.
- -----. [1990], ‘Desiderata for modified quantum dynamics’, in *PSA 1990*, Volume 2, A. Fine, M. Forbes and L. Wessels (eds), Philosophy of Science Association, East Lansing, Michigan.
- Squires, E. [1991], ‘Wave-function collapse and ultraviolet photons’, *Physics Letters*, **A 158**, 431.
- Stapp, H.P. [1989], ‘Quantum nonlocality and the description of nature’, in *Philosophical Consequences of Quantum Theory*, J.T. Cushing and E. McMullin (eds), University of Notre Dame Press, Notre Dame, Indiana.
- Suppes, P., and Zanotti, M. [1976], ‘On the determinism of hidden variables theories with strict correlation and conditional statistical independence of observables’, in *Logic and Probability in Quantum Mechanics*, P. Suppes (ed.), Reidel, Dordrecht.
- van Fraassen, B. [1982], ‘The Charybdis of Realism: Epistemological Implications of Bell’s Inequality’, *Synthese*, **52**, 25 (1982).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Bell’s Theorem | quantum mechanics | quantum mechanics: Bohmian mechanics

GianCarlo Ghirardi
Department of Theoretical Physics, University of Trieste
ghirardi@ts.infn.it

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 7, 2002
Content last modified: March 7, 2002

Measurement in Quantum Theory

From the inception of Quantum Mechanics (QM) the concept of measurement has proved a source of difficulty. The Einstein-Bohr debates, out of which both the Einstein Podolski Rosen paradox and Schrödinger's cat paradox developed, centered upon this difficulty. The problem of measurement in quantum mechanics arises out of the fact that several principles of the theory appear to be in conflict. In particular, the dynamic principles of quantum mechanics seem to be in conflict with the postulate of collapse. David Albert puts the problem nicely when he says:

> The dynamics and the postulate of collapse are flatly in contradiction with one another ... the postulate of collapse seems to be right about what happens when we make measurements, and the dynamics seems to be bizarrely *wrong* about what happens when we make measurements, and yet the dynamics seems to be *right* about what happens whenever we *aren't* making measurements. (Albert 1992, 79)

This has come to be known as "the measurement problem" and in what follows, we study the details and examine some of the implications of this problem.

The measurement problem is not just an interpretational problem internal to QM. It raises broader issues as well, such more general philosophical debates between, on the one hand, Cartesian and Lockean accounts of observation as the creation of "inner reflections" and, on the other, neo-Kantean conceptions of observation as a quasi-externalized physiological process. In this article I trace the history of these debates, and indicate some of the interpretative strategies that they stimulated.

The Birth of the Measurement Problem

The measurement problem in QM (Quantum Mechanics) grew out of early debates over Niels Bohr's "Copenhagen interpretation". Bohr maintained that the physical properties of quantum systems depend in a fundamental way upon experimental conditions, including conditions of measurement. This doctrine appeared explicitly in Bohr's 1935 reply to Einstein, Podolski, and Rosen: "The procedure of measurement has an essential influence on the conditions on which the very definition of the physical quantities in question rests" (Bohr 1935, 1025; see too Bohr 1929). To be specific, Bohr endorsed the following principle:

> (P) If a quantity Q is measured in system S at time *t* then Q has a particular value in S at *t*.[1]

But, instead of taking the dependence of properties upon experimental conditions to be causal in nature, he proposed an analogy with the dependence of relations of simultaneity upon frames of reference postulated by special relativity theory: "The theory of relativity reminds us of the subjective [observer dependent] character of all physical phenomena, a character which depends essentially upon the state of motion of the observer" (Bohr 1929, 73). In general terms, then, Bohr proposed that, like temporal relations in special relativity, properties in QM exhibit a hidden relationalism - "hidden", that is, from a classical, Newtonian point of view. Paul Feyerabend gave a clear exposition of this Bohrian position in his "Problems of Microphysics" essay (Feyerabend, 1962). It can also be found in earlier commentaries upon Bohr by Vladimir Fock and Philip Frank (Jammer 1974, section 6.5).

Many of Bohr's colleagues, including his young *protege* Werner Heisenberg, misunderstood or rejected the relationalist metaphysics underpinning Bohr's endorsement of (P). Instead, they favored the positivistic, anti-metaphysical approach expressed in Heisenberg's influential book, *The Physical Principles of the Quantum Theory* (Heisenberg 1930): "It seems necessary to demand that no concept enter a theory which has not been experimentally verified at least to the same degree of accuracy as the experiments to be explained by the theory" (1).[2] On this view, (P) may be strengthened to the principle (P)′:

> (P)′ It is meaningless to assign Q a value *q* for S at *t* unless Q is measured to have value *q* for S at *t*.

Heisenberg's approach, as presented in *The Physical Principles of the Quantum Theory*, quickly became a popular way of reading (or misreading, as Bohr would claim) the philosophically more forbidding complexities of the Copenhagen interpretation. As Max Jammer points out: "It would be difficult to find a textbook of the period [1930-1950] which denied that the numerical value of a physical quantity has no meaning whatsoever until an observation has been performed" (Jammer 1974, 246).

Bohr disagreed with Heisenberg's extreme positivistic gloss of the Copenhagen interpretation that reduced questions of "definability to measurability" (Jammer 1974, 69). The disagreement was no casual matter. Heisenberg reports a discussion that arose while preparing his 1927 *Zeitschrift für Physik* paper in

the following terms: "I remember that it ended with my breaking out in tears because I just couldn't stand this pressure from Bohr" (Jammer 1974, 65). Nevertheless, the two men agreed in broad terms that ways of describing quantum systems depended upon experimental conditions. This agreement was sufficient to create at least the appearance of a unified Copenhagen position.[3]

The assumptions that framed the Bohr-Heisenberg interpretation were, in turn, rejected by Albert Einstein (Jammer 1974, chap.5; see too Bohr 1949). Einstein's disagreement with the Copenhagen school came to a head in the famous exchange with Bohr at the fifth Solvay conference (1927) and in the no less famous Einstein, Podolski, Rosen paper of 1935. Arguing from what might be called a "realist" position, Einstein contended that under ideal conditions observations (and measurements more generally) function like "mirrors" (or, as Crary argues, camera obscura) reflecting an independently existing reality (Crary 1995, 48). In particular, in the Einstein, Podolski, Rosen paper, we find the following criterion for the existence of physical reality: "If without in any way disturbing a system we can predict with certainty...the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity" (Einstein et al 1935, 778). This criterion characterizes physical reality in terms of objectivity, meaning its independence from any direct measurement. By implication, then, when a direct measurement of physical reality occurs it merely passively reflects rather than actively constituting that which is observed.

Einstein's position has roots in Cartesian as well as empiricist, and specifically Lockean, notions of perception. This realist position opposes the Kantian metaphor of the "veil of perception" that pictures the apparatus of observation as like a pair of spectacles through which a highly mediated sight of the world can be glimpsed. To be specific, according to Kant, rather than simply reflecting an independently existing reality, "appearances" are constituted through the act of perception in a way that conforms them to the fundamental categories of sensible intuition. As Kant made the point in the Transcendental Aesthetic: "Not only are the drops of rain mere appearances, but...even their round shape, and even the space in which they fall, are nothing in themselves, but merely modifications of fundamental forms of our sensible intuition, and...the transcendental object remains unknown to us" (Kant 1973, 85).

By contrast, the realism that I am associating with Einstein takes the point of view that, insofar as they are real, when we observe rain drops under ideal conditions we are seeing objects "in themselves", that is, as they exist independently of being perceived. In other words, not only do the rain drops exist independently of our observations but also, in observing them, what we see reflects how they really are. In William Blake's succinct formulation, "As the eye [sees], such the object [is]" (Crary 1995, 70). According to this "realist" point of view, ideal observations not only reflect the way things are during but also immediately before and after observation.[4]

Such realism was opposed by both Bohr and Heisenberg.[5] Bohr took a position that, by taking acts of observation and measurement more generally as constitutive of phenomena, aligned him more closely with a Kantian point of view. To be specific, Bohr took it that "measurement has an essential [by which I take him to mean constitutive] influence on the conditions on which the very definition of the physical quantities in question rests" (Bohr 1935, 1025).

As Henry Folse points out, however, it is misleading to take the parallel between Bohr and Kant too far (Folse 1985, 49 and 217-221). Bohr strongly opposed the Kantian position that "space and time as well as cause and effect had to be taken as *a priori* categories for the comprehension of all knowledge" (Folse 1985, 218). This opposition between Bohr and Kant reflected a deeper division. Whereas for Kant "concepts played their role prior to experience and give form to what is experienced" (Folse, 220), for Bohr it was the other way around, that is, objective reality, in particular conditions of observation, determine the applicability of concepts. Thus, although for Bohr no less than for Kant, observation took on a role in determining the forms that structure the world of visible objects, the two men conceived the way in which that role is discharged quite differently. For Kant subjective experience was structured in terms of certain prior forms, whereas Bohr argued for a hidden relationalism in the domain of appearances, contending that the properties in terms of which a system is described are relative to the conditions of measurement.

This difference between Bohr and Kant may be seen as an aspect, indeed radicalization, of a more general shift in nineteenth century conceptions of vision, exemplified in Johannes Müller's compendious summary of current physiology, *Handbuch der Physiologie des Menschen* (1833). Müller (a mentor of the influential physicist Hermann von Helmholtz) may be seen as physiologizing the Kantian conception of observation. As Jonathon Crary makes the point:

> His [Müller's] work, in spite of his praise of Kant, implies something quite different. Far from being apodictic or universal in nature, like the ‘spectacles’ of time and space, our physiological apparatus is again and again shown to be defective, inconsistent, prey to illusion, and, in a crucial manner, susceptible to external procedures of manipulation and stimulation that have the essential capacity *to produce experience for the subject*.(Crary 1995, 92)

Crary implies here that during the nineteenth century observation, and specifically vision, were both reconceptualized not as a Kantian universal faculty but rather as physiological processes. In particular, it was assumed that observable phenomena were conditioned, not by universal forms of sensible intuition, but rather by the sorts of external physical factors that affected bodily and specifically physiological processes in general.

Bohr extended the nineteenth century concept by proposing that the "external procedures" that influence vision affect not only how we see but also the scientific concepts in terms of which what we see should be described. Even more radically, Bohr proposed that the "external procedures" that affect sensible intuitions include the processes of observation themselves. Thus Bohr stood at the end of a long historical trajectory. Both Kant and Descartes conceived the apparatus of observation as an inner mental faculty, analogous to a pair of spectacles (Kant) or a camera obscura (Descartes) mobilizing the perceptions of some inner Eye. In the nineteenth century, vision was projected outwards, reconceived as a bodily, and specifically physiological process (Müller, Helmholtz, and Johann Friedrich Herbart, Kant's successor at Königsberg). Bohr, then, completed the process of externalization by severing observation from the body, including it as one among many "external procedures" that affect accounts of what we see.[6]

Like Bohr, Heisenberg opposed Einstein's "realism". But whereas Bohr's opposition was rooted in a neo-Kantian relationalism that reversed Kant by externalizing the inner mental faculties, Heisenberg opposed Einstein from a more straightforwardly positivistic standpoint that disagreed not only with Einstein but also with Bohr.[7]

To be specific, Heisenberg took as meaningless the sorts of metaphysical speculations about the "true nature of reality" that preoccupied both Einstein and Bohr, speculations that, according to Heisenberg, betrayed their metaphysical nature by divorcing questions of truth from more concrete issues of what is observed:

> It is possible to ask whether there is still concealed behind the statistical universe of perception a ‘true’ universe in which the law of causality would be valid. But such speculation seems to us to be without value and meaningless, for physics must confine itself to the description of the relationship between perceptions.(Heisenberg 1927, 197)

The End of Copenhagen Monocracy

By embedding QM within the formal theory of Hilbert spaces, John von Neumann, a brilliant pure mathematician, provided the first rigorous axiomatic treatment of QM (von Neumann 1955 - the original German edition of this book appeared in 1932). Unlike Bohr and Einstein, he took seriously QM's formalism, not only providing the theory with rigorous mathematical foundations but also allowing a new conceptual architectonic to emerge from within the theory itself rather than following Heisenberg, Bohr, and Einstein who imposed a system of concepts *a priori*.

Von Neumann also intervened decisively into the measurement problem. Summarizing earlier work, he argued that a measurement on a quantum system involves two distinct processes that may be thought of as temporally contiguous stages (417-418).[8] In the first stage, the measured quantum system S interacts with M, a macroscopic measuring apparatus for the physical quantity Q. This interaction is governed by the linear, deterministic Schrödinger equation, and is represented in the following terms: at time t, when the measurement begins, S, the measured system, is in a state represented by a Hilbert space vector f that, like any vector in the Hilbert space of possible state vectors, is decomposable into a weighted sum - a "linear superposition" - of the set of so-called "eigenvectors" $\{f_i\}$ belonging to Q. In other words, $f = \Sigma c_i f_i$, for some set $\{c_i\}$ of complex numbers. f_i, the eigenvector of Q corresponding to possible value q_i, is that state of S at t for which, when S is in that state, there is unit probability that Q has value q_i.[9] M, the measuring apparatus, is taken to be in a "ready" state g at time t when the measurement begins. According to the laws of QM, this entails that S+M at t is in the "tensor product" state $\Sigma c_i f_i \otimes g$.

By applying the Schrödinger equation to this product state, we deduce that at time t', when the first stage of the measurement terminates, the state of S+M is $\Sigma c_i f_i \otimes g_i$, where g_i is a state in which M registers the value q_i.[10] Such states, represented by a linear combination of products of the form $f_i \otimes g_i$, have been

dubbed "entangled states".[11]

After the first stage of the measurement process, a second non-linear, indeterministic process takes place, the "reduction of the wave packet", that involves S+M "jumping" (the famous "quantum leap") from the entangled state $\Sigma c_i f_i \otimes g_i$ into the state $f_i \otimes g_i$ for some i. This, in turn (according to the laws of QM) means that S is in state f_i and M is in the state g_i, where g_i, it is assumed, is the state in which M registers the value q_i. Let t'' denote the time when this second and final stage of the measurement is finished.[12] It follows that at t'', when the measurement as a whole terminates, M registers the value q_i. Since the reduction of the wave-packet is indeterministic, there is no possibility of predicting which value M will register at t''. We can conclude only that M will register some value.

The second stage of the measurement, with its radical, non-linear discontinuities, was from its introduction the source of many of the philosophical difficulties that plagued QM, including what von Neumann referred to as its "peculiar dual nature" (417). As Schrödinger was moved to say during a visit to Bohr's institute during September 1926: "If all this damned quantum jumping [*verdamnte Quantenspringerei*] were really to stay, I should be sorry I ever got involved with quantum theory" (Jammer 1974, 57)

QM has nothing else definite to say about the measurement process. To be specific, from within the resources of QM there is no way of predicting what value of Q will be registered. QM does, however, give us some additional *statistical* information, via the so called Born statistical interpretation:

> The probability of q_i being registered is $|c_i|^2$, where $_i$ is the coefficient of f_i (the eigenvector of Q corresponding to value q_i) when the initial measured state of S is expressed as a linear superposition of eigenvectors of Q.

In short, QM does not predict what the measured value will be but does at least tells us the probability distribution over various possible measured values.

Cats in Singlets

Von Neumann's formal work enables a clear exposition of various paradoxes that have haunted QM from its inception. One of these was the famous Schrödinger cat paradox (Schrödinger 1935b). This paradox dramatizes the fact that, according to QM, the observer's intervention at the end of the first stage of the measurement process precipitates S+M from a complex entangled state, that is from a superposition or hybrid of states for which M registers different possible values for Q, into a state for which M registers a single value. In short, the act of observation creates a paradoxical shift in M: from a hybrid state for which the value that M registers is "indeterminate" to a state in which M registers precisely one such value. In the case of the cat paradox, matters are so arranged that a cat being dead or alive corresponds to the differing states of M. Thus, it seems, the act of observing the cat precipitates it from a strange zombie-

like state in which its state of morbidity is indeterminate to a state in which it is either dead or alive. "Looks can kill", as we might say.

The measurement problem was exacerbated by another paradox that arose in the context of the Einstein-Bohr debate: what has come to be called the EPR (Einstein-Podolski-Rosen) paradox (Einstein, Podolski, Rosen 1935). It should be stressed that in their original article EPR presented their argument as proof of the incompleteness rather than inconsistency of QM. Nevertheless, in the subsequent literature their argument quickly took on the role of a paradox, one that is most perspicuously presented in terms of the formalism developed by Bohm and Aharanov (Bohm and Aharanov 1957). Consider a pair of electrons S_1, S_2 at time *t* when they are in a so-called singlet state, represented by the vector

$$\{(f_{x+} \otimes g_{x-}) + (f_{x-} \otimes g_{x+})\}/\sqrt{2},$$

where f_{x+} and f_{x-} represent the two possible eigenstates of the x-directed spin of S_1 corresponding to the two possible spin values +1/2 and -1/2 respectively; g_{x-} and g_{x+} represent the corresponding eigenstates for S_2. From the Born statistical interpretation it is easy to deduce that when S_1+S_2 is in the singlet state, the x-spin values of S_1 and S_2 are anticorrelated, that is, the conditional probability of measuring the x-spin of S_1 to have value +1/2 given that the x-spin of S_2 is measured to have value -1/2 is 1, and vise versa. It is also a theorem of QM that the linear decomposition of the singlet state vector is invariant under rotation and in particular invariant under interchange of *x* and *y*:

$$(f_{x+} \otimes g_{x-}) + (f_{x-} \otimes g_{x+}) = (f_{y+} \otimes g_{y-}) + (f_{y-} \otimes g_{y+})$$

Now suppose that S_1 and S_2 have been allowed to drift out of each other's spheres of influence, so that a disturbance of S_1 can have no simultaneous effect upon S_2. Suppose too that we measure the *x*-spin of S_1 just before t, and that the value revealed by measurement is +1/2. In that case, the anti-correlation between the *x*-spin values for S_1 and S_2 makes it possible to predict with certainty that, in the event that the *x*-spin of S_2 is measured just before t, the value revealed by measurement is -1/2. The possibility of making this prediction means that the *x*-spin measurement on S_1 also counts as an *x*-spin measurement on S_2, albeit an indirect measurement since it is carried out in a region of space remote from S_2. By applying the reduction of the wave-packet postulate to this indirect measurement, we conclude that, at time t immediately after the measurement, the state of S_2 is g_{x-}, the eigenvector of *x*-spin for value -1/2.

But now assume that a second measurement has been carried out just before *t*, one that *directly* measures the spin of S_2 in the y direction. There is no difficulty in simultaneously conducting both of these measurements since, because they take place in different regions of space, they cannot interfere with each other. By applying the reduction of the wave-packet postulate to this second measurement, we conclude that the state of S_2 immediately post-measurement is either g_{y-} or g_{y+}, depending on whether the measured value for *y*-spin is -1/2 or +1/2. Thus we arrive at a direct contradiction, since the state of S_2

post-measurement cannot be both g_{x-} and one of g_{y-} or g_{y+}. Here, then, lies the nub of the EPR paradox, showing that QM is inconsistent with the reduction of the wave-packet postulate. (In its original form the EPR argument merely showed that without the reduction of the wave-packet postulate, QM is incomplete.[13])

The World of Many Interpretations

The measurement problem and its associated paradoxes have generated a multitude of responses. One such, based upon von Neumann's work with density operators, is due to Josef M. Jauch (Jauch 1968, chapter 11). The post measurement entangled state of S+M, $\Sigma c_i f_i \otimes g_i$, is a "pure state", represented by a single Hilbert space vector. There are, however, other sorts of states in QM, namely "mixed states", represented not by single vectors but rather by so called density operators. It is characteristic of S being in a mixed state that, from the point of view of statistical distributions over possible results of measurements on S, S behaves as if, for some set of vectors $\{f_i\}$ and some set of numbers $\{p_i\}$ for which $\Sigma p_i = 1$, there is probability p_i that S is in the state f_i, for i = 1,2,.... (Mathematically, such a state is represented by a so-called density operator, $\Sigma p_i \mid f_i\rangle\langle f_i \mid$, where $\mid f_i\rangle\langle f_i \mid$ is the projection operator onto the vector f_i.[14]) Von Neumann proved that when S+M is in the entangled state $\Sigma c_i f_i \otimes g_i$ then S is in such a mixed state. In particular, S behaves as if there were probability $|c_i|^2$ of being in state f_i, for i = 1,2...; similarly M is in a mixed state, behaving as if there were probability $|c_i|^2$ of being in state g_i, for i = 1,2....(von Neumann 1955, 424).

It seems, then, that a solution to the measurement problem is within easy reach. We simply interpret the state of S when S+M is in the entangled state $\Sigma c_i f_i \otimes g_i$ as a new sort of "mixed" state in which there really is probability $|c_i|^2$ that S is in the state f_i, for i = 1,2,.... The probability in question is not merely a subjective measure of ignorance (otherwise the state is really a pure state, as defined in the previous endnote) but instead is an "intrinsic" property of the system S, in particular, it may be thought of as an objective measure of a propensity of S at t to be in the state f_i (Jauch 1968, 173-174). This, in turn, means that, already at the end of the first stage of the measurement, there is probability $|c_i|^2$ that Q has value q_i in S (as above I am taking f_i as the eigenvector of Q corresponding to value q_i). By parity of reasoning, at the end of the first stage of the measurement, there is probability $|c_i|^2$ of M being in state g_i and hence of registering the value q_i for Q. Thus, it seems, the "reduction of the wave packet" is redundant, since already at the end of the first stage measurement the measuring apparatus registers the appropriate possible values with probabilities in agreement with the Born statistical interpretation. As such, those paradoxes of QM, such as EPR and Schrödinger's cat, that depend upon the reduction of the wave packet simply disappear.[15]

But a difficulty remains. The state of S+M at the end of measurement is still an entangled state for which, it seems, we cannot say that Q has value q_i with probability $|c_i|^2$. Indeed, from the perspective of that

combined state, it seems that the value of Q is indeterminate, suspended between the various possible values q_1, q_2, and so on. More seriously, it seems that the measuring apparatus suffers from a similar indeterminacy: that is, it is indeterminate which value it registers. In short, it seems that, from the point of view of the combined measuring and measured system, Schrödinger's cat paradox (although not his cat) survives unscathed.[16]

The Italian School of Daneri, Loinger, Prosperi, *et al* responded to this problem by advancing what has come to be called a "phase wash out" theory (Daneri, Loinger and Prosperi 1962). They showed that in virtue of statistical thermodynamic features of the measuring apparatus, the state of S+M at t' (the end of the first stage of measurement) approximates a mixed state - also called the "reduced state" - in which there is probability $|c_i|^2$ that S+M is in the state represented by the product vector $f_i \otimes g_i$, for all the various i = 1,2,..... In this reduced state the nagging indeterminacy effects vanish.

A serious difficulty remains, however. It may well be true that S+M is approximately in a mixed state. But this does not solve the cat paradox. That is, although it may be true that to a good approximation Schrödinger's cat is either dead or alive, the air of paradox remains if, when we examine in detail the micro-correlations between the measured and measuring systems, we see that the cat is in a zombie like dead-and-alive state.

The "phase wash out" approach and Jauch's approach more generally suffer another drawback, one they share with hidden variable interpretations of QM (for a discussion of the latter interpretations, see Belinfante 1973). In the special situation described by the EPR paradox, for which the density operator of the measured system is an identity operator, these interpretations assign determinate values to *all* physical quantities for a particular quantum system.[17] Thus they fall prey to a new generation of paradoxes that depend upon Gleason's theorem and the related Kochen and Specker theorem.[18]

The paradoxes and questions raised by the measurement problem have spawned a host of interpretations of QM, including hidden variable theories that continue Einstein's search for a "complete" account of physical reality, and the Everett-Wheeler "many worlds interpretation" (Wheeler and Zureck 1983, II.3 and III.3; Bell 1987, chapters 4 and 20). Most physicists bypass these philosophical resolutions of the interpretative difficulties of QM, and revert instead to some version of the Bohr interpretation. Often that version is related closely to the early Heisenberg's positivistic, anti-metaphysical approach. It is as if the long history of failure to resolve the acrimonious disputes surrounding the interpretation of QM has led quantum physicists to become disenchanted with the garden of metaphysical delights. As John S. Bell has made the point, despite more than seventy years of interpreting QM and resolving the measurement problem, the Bohr interpretation in its more pragmatic, less metaphysical forms remains the "working philosophy" for the average physicist (Bell 1987, 189).[19]

Bibliography

- Albert, D., 1992, *Quantum Mechanics and Experience*, Cambridge, MA: Harvard university Press

- Araki, H., and Yanase, M.M., 1960. "Measurement of quantum mechanical operators". *Physical Review*. **120**:622-626.
- Belinfante, F.J., 1973. *A Survey of Hidden-Variables Theories*. New York: Pergamon Press.
- Bell, J.S., 1987. *Speakable and unspeakable in quantum mechanics*. New York: Cambridge University Press.
- Beller, M., and Fine, A., 1994. "Bohr's Response to E-P-R". In *Niels Bohr and Contemporary Philosophy*. Edited Jan Faye and Henry Folse. Dordrecht: Kluwer.
- Beller, M. 1996. "The Rhetoric of Antirealism and the Copenhagen Spirit". *Philosophy of Science*. **63**:183-204.
- Bohm, D., and Aharanov, Y., 1957. "Discussion of experimental proof for the paradox of Einstein, Rosen and Podolski". *Physical Review*. **108**:1070-1076.
- Bohr, N., 1929. "*Die Atomtheorie und die Prinzipien der Naturschreibung*". *Die Naturwissenschaften*. **18**:73-78.
- Bohr, N., 1935. "Quantum Mechanics and Physical Reality". *Nature*. **136**:1025-1026.
- Bohr, N., 1959. "Discussion with Einstein on epistemological problems in atomic physics". in *Albert Einstein: Philosopher-Scientist*. Edited Paul A. Schilpp. New York: Harper and Row.
- Bunge, M., 1967. *Foundations of Physics*. Berlin: Springer.
- Crary, J., 1995. *Techniques of the Observer: On Vision and Modernity in the Nineteenth Century*. Cambridge Mass: October Books.
- Daneri, A., Loinger, A., and Prosperi, G.M., 1962. "Quantum theory of measurement and ergodicity requirements". *Nuclear Physics*. **33**:297-319.
- Einstein, E., Podolski, B., and Rosen, N., 1935, "Can quantum-mechanical description of physical reality be considered complete?". *Physical Review*. **47**:777-780.
- Feyerabend, P., 1962. "Problems of Microphysics". in *Frontiers of Science and Philosophy*. Edited Robert G. Colodny. Pittsburgh: Pittsburgh University Press.
- Folse, H., 1985. *The Philosophy of Niels Bohr: The Framework of Complementarity*. Amsterdam: North-Holland.
- Healey, R., and Hellman, G., (eds.), 1998, *Quantum Measurement*, Minneapolis, MN : University of Minnesota Press
- Heisenberg, W., 1927. "*Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik*". *Zeitschrift für Physik*. **43**:172-198.
- Heisenberg, W., 1930. *The Physical Principles of the Quantum Theory*. Trans. Carl Eckhart and F.C. Hoyt. New York: Dover.
- Jammer, M., 1974. *The Philosophy of Quantum Mechanics*. New York: Wiley
- Jauch, J.M., 1968. *Foundations of Quantum Mechanics*. Reading: Addison-Wesley.
- Kant, I., 1973. *Critique of Pure Reason*. Trans. Norman Kemp Smith. London: Macmillan.
- Krips, H., 1990. *The Metaphysics of Quantum Theory*. Oxford: Clarendon.
- Popper, K., 1968. *The Logic of Scientific Discovery*. London: Hutchinson.
- Popper, K., 1982. *Quantum Theory and the Schism in Physics*. London: Hutchinson.
- Redhead, M., 1987. *Incompleteness, Nonlocality, and Realism*. Oxford: Clarendon.
- Schrödinger, E., 1935a. "Discussion of probability relations between separated systems". *Proceedings of the Cambridge Philosophical Society*. **31**:555-562.
- Schrödinger, E., 1935b. "*Die Gegenwärtige Situation in der Quantenmechanik*". *Die*

Naturwissenschaften. **23**:807-812; 824-828; 844-849.

- von Neumann, J., 1955. *Mathematical Foundations of Quantum Mechanics*. Trans. Robert T. Geyer. Princeton: Princeton University Press.
- Wheeler, J.A., and Zurek, W.H., 1983. *Quantum Theory and Measurement*. Princeton: Princeton University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Einstein, Albert: Einstein-Bohr debates | physics: experiment in | physics: holism and nonseparability | quantum mechanics | quantum mechanics: Bohmian mechanics | quantum mechanics: collapse theories | quantum mechanics: Copenhagen interpretation of | quantum mechanics: Everett's relative-state formulation of | quantum mechanics: many-worlds interpretation of | quantum mechanics: modal interpretations of | quantum mechanics: the role of decoherence in | quantum theory: the Einstein-Podolsky-Rosen argument in | Uncertainty Principle

Henry Krips
krips+@pitt.edu

First published: October 28, 1999
Content last modified: October 28, 1999

Stanford Encyclopedia of Philosophy
Notes to Quantum Theory

Notes

1. The issue of whether measurement was also a necessary condition for the assignment of values to physical quantities remained a question of controversy within the Copenhagen school. Bohr seems *not* to have insisted on this. Heisenberg, by contrast, at least in his early positivistic writings, seems committed to strengthening (P) so that measurement is both necessary and sufficient for the assignment of a determinate value to a physical quantity. I discuss this difference later.

2. Heisenberg discarded this positivistic approach in his later work, indeed, as early as 1930, according to Jammer (Jammer 76).

3. The question of Bohr's and Heisenberg's attitudes to positivism is a complex one. At least initially, in the context of their debate with Einstein they took common cause in favor of positivism (Jammer 1974, 109). As indicated in the previous footnote, however, Heisenberg quickly abandoned his earlier positivism, whereas, according to Beller and Fine, Bohr moved towards a more positivistic attitude in his later work (Beller and Fine 1994). On the issue of a unified "Copenhagen Spirit", see Beller 1996.

4. In more recent times, Popper and Bunge have been exponents of the Einsteinian position in the context of QM (Popper 1982, 35-41; Bunge 1967, 274-287). This position does not rule out the possibility of observations that fail to report what exists before or after their occurrence. But such observations, it is held, are defective. As one might say pejoratively, they present a "distorted" image of what we observe.

5. Indeed, it took them somewhat by surprise in the light of Einstein's early Machian position (Jammer 1974, 109).

6. Von Neumann's awkward attempt to rescue the internal status of observation, by talking about the arbitrariness of the *Schnitt* (cut) between observer and observed, may be understood as a reluctant acceptance of this Bohrian impoverishment of the "inner" life of the observer (von Neumann 1955, 418-420).

7. But see Beller and Fine 1994.

8. Von Neumann himself did not present these processes as temporally ordered stages, referring instead to the "peculiar dual nature of the quantum mechanical procedure" (417).

9. For simplicity I am assuming physical quantities with discrete spectra and one eigenvector for each possible value, that is, I am assuming a Hilbert space representation for Q as non-degenerate.

10. g_i may be thought of as a state for which a pointer that is part of M points to the i-th interval on a scale. Araki and Yanase have shown that such interactions are subject to very strong restrictions (Araki and Yanase 1960). Their work suggests that some weakening of the idealized form of measurement interaction is required.

11. Schrödinger seems to have been the first to suggest the term "entangled" in this context - see Schrödinger 1935a.

12. Note that it is not clear whether this second stage of the measurement process is instantaneous, that is whether $t'' > t'$ or $t'' = t'$. In part this reflects an ambiguity present in von Neumann's formulation of the problem, namely whether the two parts of the measurement should be seen as taking place simultaneously or as succeeding stages.

13. The version of the EPR paradox which I give here is close to the one reported in appendix *xii to Popper's *Logic of Scientific Discovery* (Popper 1968). It is taken from a letter by Einstein to Popper written in 1935, after publication of the EPR paper.

14. The more usual sorts of states of QM - the "pure states" - may be thought as degenerate cases of mixed states, that is, as mixed states for which $\{p_i\}$ is a singleton set, consisting of the number 1. In other words, a pure state is simply a mixed state for which there is unit probability that the system in question is in the state f, for some vector f.

15. Jauch 1968, 188-191. This interpretation also has the formal advantage that the Born statistical interpretation emerges as a theorem rather than being postulated as an independent axiom. This is because after measurement M is in a mixed state for which there is probability c_i^2 of M registering value q_i.

16. Jauch attempts to address this problem in terms of his theory of equivalent states - Jauch 1968, 184.

17. This is because the identity operator is multiply diagonalizable, that is, it is equal to $f_i >< f_i/N$ for any complete orthonormal set of vectors $\{f_i\}$, where N is the dimension of the Hilbert space.

18. For a discussion of these difficulties see Krips 1990, and Redhead 1987.

19. I am grateful to the editor, Rob Clifton, for his helpful comments.

Henry Krips
krips+@pitt.edu

First published: October 28, 1999
Content last modified: October 28, 1999

Stanford Encyclopedia of Philosophy

Experiment in Physics

Physics, and natural science in general, is a reasonable enterprise based on valid experimental evidence, criticism, and rational discussion. It provides us with knowledge of the physical world and it is experiment that provides the evidence that grounds that knowledge. Experiment plays many roles in science. One of its important roles is to test theories and to provide the basis for scientific knowledge.[*/] It can also call for a new theory, either by showing that an accepted theory is incorrect, or by exhibiting a new phenomenon which needs explanation. Experiment can provide hints toward the structure or mathematical form of a theory and it can provide evidence for the existence of the entities involved in our theories. Finally, it may also have a life of its own, independent of theory. Scientists may investigate a phenomenon just because it looks interesting. This will also provide evidence for a future theory to explain. [Examples of these different roles will be presented below.] As we shall see below, a single experiment may play several of these roles at once.

If experiment is to play these important roles in science then we must have good reasons to believe experimental results, for science is a fallible enterprise. Theoretical calculations, experimental results, or the comparison between experiment and theory may all be wrong. Science is more complex than "The scientist proposes, Nature disposes." It may not always be clear what the scientist is proposing. Theories often need to be articulated and clarified. It also may not be clear how Nature is disposing. Experiments may not always give clear-cut results, and may even disagree for a time.

In what follows, the reader will find an epistemology of experiment, a set of strategies that provides reasonable belief in experimental results. Scientific knowledge can then be reasonably based on these experimental results.

I. Experimental Results

A. The Case For Learning From Experiment

1. An Epistemology of Experiment

It has been almost two decades since Ian Hacking asked, "Do we see through a microscope?" (Hacking 1981). Hacking's question really asked how do we come to believe in an experimental result obtained with a complex experimental apparatus? How do we distinguish between a valid result[1] and an artifact created by that apparatus? If experiment is to play all of the important roles in science mentioned above and to provide the evidential basis for scientific knowledge, then we must have good reasons to believe in those results. Hacking provided an extended answer in the second half of *Representing and Intervening* (1983). He pointed out that even though an experimental apparatus is laden with, at the very least, the theory of the apparatus, observations remain robust despite changes in the theory of the apparatus or in the theory of the phenomenon. His illustration was the continuous belief in microscope images despite the major change in the theory of the microscope when Abbe pointed out the importance of diffraction in its operation. One reason Hacking gave for this is that in making such observations the experimenters intervened. They manipulated the object under observation. Thus, in looking at a cell through a microscope one might inject fluid into the cell or stain the specimen. One expects the cell to change shape or color when this is done. Observing the predicted effect strengthens our belief in both the proper operation of the microscope and in the observation. This is true in general. Observing the predicted effect of an intervention strengthens our belief in both the proper operation of the experimental apparatus and in

the observations made with it.

Hacking also discussed the strengthening of one's belief in an observation by independent confirmation. The fact that the same pattern of dots, dense bodies in cells, is seen with "different" microscopes, i.e. ordinary, polarizing, phase-contrast, fluorescence, interference, electron, acoustic etc., argues for the validity of the observation. One might question whether or not "different" is theory laden. After all, it is our theory of light and of the microscope that allows us to consider these microscopes "different." Nevertheless, the argument goes through. Hacking correctly argues that it would be a preposterous coincidence if the same pattern of dots were produced in two totally different kinds of physical systems. Different apparatuses have different backgrounds and systematic errors, making the coincidence, if it is an artifact, most unlikely. If it is a correct result, and the instruments are working properly, the coincidence of results is understandable.

Hacking's answer is correct as far as it goes. It is, however, incomplete. What happens when one can perform the experiment with only one type of apparatus, such as an electron microscope or a radio telescope, or when intervention is either impossible or extremely difficult? Other strategies are needed to validate the observation.\2/ These may include:

> 1) Experimental checks and calibration, in which the experimental apparatus reproduces known phenomena. For example, if we wished to argue that the spectrum of a substance obtained with a new type of spectrometer is correct, we might check that this new spectrometer could reproduce the known Balmer Series in hydrogen. If we correctly observe the Balmer Series then we strengthen our belief that the spectrometer is working properly. This also strengthens our belief in the results obtained with that spectrometer. If the check fails then we have good reason to question the results obtained with that apparatus.
>
> 2) Reproducing artifacts that are known in advance to be present. An example of this comes from experiments to measure the infrared spectra of organic molecules (Randall et al. 1949). It was not always possible to prepare a pure sample of such material. Sometimes one had to place the substance in an oil paste or in solution. In such cases, one expects to observe, superimposed on the spectrum of the substance, the spectrum of the oil or the solvent, which one can compare with the known spectrum of the oil or the solvent. Observation of this artifact gives confidence in other measurements made with the spectrometer.
>
> 3) Elimination of plausible sources of error and alternative explanations of the result (the Sherlock Holmes strategy).\3/ Thus, when scientists claimed to have observed electric discharges in the rings of Saturn, they argued for their result by showing that it could not have been caused by defects in the telemetry, by interaction with the environment of Saturn, by lightning, or by dust. The only remaining explanation of their result was that it was due to electric discharges in the rings. There was no other plausible explanation of the

observation. In addition, the same result was observed by both Voyager 1 and Voyager 2. This provided independent confirmation. Often, several epistemological strategies are used in the same experiment.

4) Using the results themselves to argue for their validity. Consider the problem of Galileo's telescopic observations of the moons of Jupiter. Although one might very well believe that his early telescope might have created spots of light, it would have been extremely implausible that the telescope would create them so that they would appear to be a small planetary system with eclipses and other consistent motions. It would have been even more implausible to believe that the created spots would satisfy Kepler's Third Law (R^3/T^2 = constant). A similar argument was used by Robert Millikan to support his observation of the quantization of electric charge and his measurement of the charge of the electron. Millikan remarked, "The total number of changes which we have observed would be between one and two thousand, and *in not one single instance has there been any change which did not represent the advent upon the drop of one definite invariable quantity of electricity or a very small multiple of that quantity*"(Millikan 1911, p. 360). In both of these cases one is arguing that there was no plausible malfunction of the apparatus, or background, that would explain the observations.

5) Using an independently well-corroborated theory of the phenomena to explain the results. This was illustrated in the discovery of the $W^{\pm}$, the charged intermediate vector boson required by the Weinberg-Salam unified theory of electroweak interactions. Although these experiments used very complex apparatuses and used other epistemological strategies (see (Franklin 1986, pp. 170-72) for details) I believe that the agreement of the observations with the theoretical predictions of the particle properties helped to validate the experimental results. In this case the particle candidates were observed in events that contained an electron with high transverse momentum and in which there were no particle jets, just as predicted by the theory. In addition, the measured particle mass of 81 ± 5 GeV/c^2 and 80^{+10}_{-6}, GeV/c^2, found in the two experiments (note the independent confirmation also), was in good agreement with the theoretical prediction of 82 ± 2.4 GeV/c^2. It was very improbable that any background effect, which might mimic the presence of the particle, would be in agreement with theory.

6) Using an apparatus based on a well-corroborated theory. In this case the support for the theory passes on to the apparatus based on that theory. This is the case with both the electron microscope and the radio telescope, whose proper operation is based on a well-supported theory, although other strategies are also used to validate the observations.

7) Using statistical arguments. An interesting example of this arose in the 1960s when the search for new particles and resonances occupied a substantial fraction of the time and effort of those physicists working in experimental high-energy physics. The usual technique was to plot the number of events observed as a function of the invariant mass of

> the final-state particles and to look for bumps above a smooth background. The usual informal criterion for the presence of a new particle was that it resulted in a three standard-deviation effect above the background, a result that had a probability of 0.27% of occurring in a single bin. This criterion was later changed to four standard deviations, which had a probability of 0.0064% when it was pointed out that the number of graphs plotted each year by high-energy physicists made it rather probable, on statistical grounds, that a three standard-deviation effect would be observed.

These strategies along with Hacking's intervention and independent confirmation constitute an epistemology of experiment. They provide us with good reasons for belief in experimental results, They do not, however, guarantee that the results are correct. There are many experiments in which these strategies are applied, but whose results are later shown to be incorrect (examples will be presented below). Experiment is fallible.

2. Galison's Elaboration

In *How Experiments End* (1987), Peter Galison extended the discussion of experiment to more complex situations. In his histories of the measurements of the gyromagnetic ratio of the electron, of the discovery of the muon, and of the discovery of weak neutral currents, he considered a series of experiments measuring a single quantity, a set of different experiments culminating in a discovery, and two high energy physics experiments performed by large groups with complex experimental apparatus.

Galison's view is that experiments end when the experimenters believe that they have a result that will stand up in court. A result that I believe will include, and has included, the use of the epistemological strategies discussed earlier. Thus, David Cline, one of the weak neutral current experimenters remarked, "At present I don't see how to make these effects [the weak neutral current event candidates] go away" (Galison, 1987, p. 235).

Galison emphasizes that, within a large experimental group, different members of the group may find different pieces of evidence most convincing. In the Gargamelle weak neutral current experiment, several group members found the single photograph of a neutrino-electron scattering event particularly important, whereas for others the difference in spatial distribution between the observed neutral current candidates and the neutron background was decisive. Galison attributes this, in large part, to differences in experimental traditions, in which scientists develop skill in using certain types of instruments or apparatus. In particle physics, for example, there is the tradition of visual detectors, such as the cloud chamber or the bubble chamber, in contrast to the electronic tradition of Geiger and scintillation counters and spark chambers. Scientists within the visual tradition tend to prefer "golden events" that clearly demonstrate the phenomenon in question, whereas those in the electronic tradition tend to find statistical arguments more persuasive and important than individual events. (For further discussion of this issue see Galison (1997)).

Galison points out that major changes in theory and in experimental practice and instruments do not necessarily occur at the same time. This persistence of experimental results provides continuity across

these conceptual changes. The experiments on the gyromagnetic ratio spanned classical electromagnetism, Bohr's old quantum theory, and the new quantum mechanics of Heisenberg and Schrodinger. Robert Ackermann has offered a similar view in his discussion of scientific instruments.

The advantages of a scientific instrument are that it cannot change theories. Instruments embody theories, to be sure, or we wouldn't have any grasp of the significance of their operation....Instruments create an invariant relationship between their operations and the world, at least when we abstract from the expertise involved in their correct use. When our theories change, we may conceive of the significance of the instrument and the world with which it is interacting differently, and the datum of an instrument may change in significance, but the datum can nonetheless stay the same, and will typically be expected to do so. An instrument reads 2 when exposed to some phenomenon. After a change in theory,\4/ it will continue to show the same reading, even though we may take the reading to be no longer important, or to tell us something other than what we thought originally (Ackermann 1985, p. 33).

Galison also discusses other aspects of the interaction between experiment and theory. Theory may influence what is considered to be a real effect, demanding explanation, and what is considered background. In his discussion of the discovery of the muon, he argues that the calculation of Oppenheimer and Carlson, which showed that showers were to be expected in the passage of electrons through matter, left the penetrating particles, later shown to be muons, as the problem. Prior to their work, physicists thought the showering particles were the problem, whereas the penetrating particles seemed to be understood.

The role of theory as an "enabling theory," one that allows calculation or estimation of the size of the expected effect and also the size of expected backgrounds is also discussed by Galison. (See also (Franklin 1995b) and the discussion of the Stern-Gerlach experiment below). Such a theory can help to determine whether or not an experiment is feasible. He also emphasizes that elimination of background that might simulate or mask an effect is central to the experimental enterprise, and not a peripheral activity. In the case of the weak neutral current experiments the existence of the currents depended crucially on showing that the event candidates could not all be due to neutron background.\5/

There is also a danger that the design of an experiment may preclude observation of a phenomenon. Galison points out that the original design of one of the neutral current experiments, which included a muon trigger would not have allowed the observation of neutral currents. In its original form the experiment was designed to observe charged currents, which produced a high energy muon. Neutral currents do not. Therefore, having a muon trigger precluded their observation. Only after the theoretical importance of the search for neutral currents was emphasized to the experimenters was the trigger changed. Changing the design did not, of course, guarantee that neutral currents would be observed.

Galison also shows that the theoretical presuppositions of the experimenters may enter into the decision to end an experiment and report the result. Einstein and de Haas ended their search for systematic errors when their value for the gyromagnetic ratio of the electron, $g = 1$, agreed with their theoretical model of orbiting electrons. This effect of presuppositions might cause one to be skeptical of both experimental

results and their role in theory evaluation. Galison's history shows, however, that, in this case, the importance of the measurement led to many repetitions of the measurement. This resulted in an agreed upon result that disagreed with theoretical expectations. Scientists do not always find what they are looking for.

B. The Case Against Learning From Experiment

1. Collins and the Experimenters' Regress

Collins, Pickering, and others, have raised objections to the view that experimental results are accepted on the basis of epistemological arguments. They point out that "a sufficiently determined critic can always find a reason to dispute any alleged ‘result’" (MacKenzie 1989, p. 412). Harry Collins, for example, is well known for his skepticism concerning both experimental results and evidence. He develops an argument that he calls the "experimenters' regress" (Collins 1985, chapter 4, pp. 79-111): What scientists take to be a correct result is one obtained with a good, that is, properly functioning, experimental apparatus. But a good experimental apparatus is simply one that gives correct results. Collins claims that there are no formal criteria that one can apply to decide whether or not an experimental apparatus is working properly. In particular, he argues that calibrating an experimental apparatus by using a surrogate signal cannot provide an independent reason for considering the apparatus to be reliable.

In Collins' view the regress is eventually broken by negotiation within the appropriate scientific community, a process driven by factors such as the career, social, and cognitive interests of the scientists, and the perceived utility for future work, but one that is not decided by what we might call epistemological criteria, or reasoned judgment. Thus, Collins concludes that his regress raises serious questions concerning both experimental evidence and its use in the evaluation of scientific hypotheses and theories. Indeed, if no way out of the regress can be found then he has a point.

Collins strongest candidate for an example of the experimenters' regress is presented in his history of the early attempts to detect gravitational radiation, or gravity waves. (For more detailed discussion of this episode see (Collins 1985; 1994; Franklin 1994; 1997a) In this case, the physics community was forced to compare Weber's claims that he had observed gravity waves with the reports from six other experiments that failed to detect them. On the one hand, Collins argues that the decision between these conflicting experimental results could not be made on epistemological or methodological grounds. He claims that the six negative experiments could not legitimately be regarded as replications[\6/] and hence become less impressive. On the other hand, Weber's apparatus, precisely because the experiments used a new type of apparatus to try to detect a hitherto unobserved phenomenon,[\7/] could not be subjected to standard calibration techniques.

The results presented by Weber's critics were not only more numerous, but they had also been carefully cross-checked. The groups had exchanged both data and analysis programs and confirmed their results. The critics had also investigated whether or not their analysis procedure, the use of a linear algorithm,

could account for their failure to observe Weber's reported results. They had used Weber's preferred procedure, a nonlinear algorithm, to analyze their own data, and still found no sign of an effect. They had also calibrated their experimental apparatuses by inserting acoustic pulses of known energy and finding that they could detect a signal. Weber, on the other hand, as well as his critics using his analysis procedure, could not detect such calibration pulses.

There were, in addition, several other serious questions raised about Weber's analysis procedures. These included an admitted programming error that generated spurious coincidences between Weber's two detectors, possible selection bias by Weber, Weber's report of coincidences between two detectors when the data had been taken four hours apart, and whether or not Weber's experimental apparatus could produce the narrow coincidences claimed.

It seems clear that the critics' results were far more credible than Weber's. They had checked their results by independent confirmation, which included the sharing of data and analysis programs. They had also eliminated a plausible source of error, that of the pulses being longer than expected, by analyzing their results using the nonlinear algorithm and by explicitly searching for such long pulses.[\8/] They had also calibrated their apparatuses by injecting pulses of known energy and observing the output.

Contrary to Collins, I believe that the scientific community made a reasoned judgment and rejected Weber's results and accepted those of his critics. Although no formal rules were applied, i.e. if you make four errors, rather than three, your results lack credibility; or if there are five, but not six, conflicting results, your work is still credible; the procedure was reasonable.

Pickering argues that the reasons for accepting results are the future utility of such results for both theoretical and experimental practice and the agreement of such results with the existing community commitments. In discussing the discovery of weak neutral currents, Pickering states,

> Quite simply, particle physicists accepted the existence of the neutral current because they could see how to ply their trade more profitably in a world in which the neutral current was real. (1984b, p. 87)

> Scientific communities tend to reject data that conflict with group commitments and, obversely, to adjust their experimental techniques to tune in on phenomena consistent with those commitments. (1981, p. 236)

The emphasis on future utility and existing commitments is clear. These two criteria do not necessarily agree. For example, there are episodes in the history of science in which more opportunity for future work is provided by the overthrow of existing theory. (See, for example, the history of the overthrow of parity conservation and of CP symmetry discussed below and in (Franklin 1986, Ch. 1, 3)).

2. Pickering on Communal Opportunism and Plastic Resources

Pickering has recently offered a different view of experimental results. In his view the material procedure including the experimental apparatus itself along with setting it up, running it, and monitoring its operation; the theoretical model of that apparatus, and the theoretical model of the phenomena under investigation are all plastic resources that the investigator brings into relations of mutual support. (Pickering 1987; Pickering 1989). He says:

> Achieving such relations of mutual support is, I suggest, the defining characteristic of the successful experiment. (1987, p. 199)

His example is Morpurgo's search for free quarks, or fractional charges of 1/3 *e* or 2/3 *e*, where *e* is the charge of the electron. (See also (Gooding 1992)). Morpurgo used a modern Millikan-type apparatus and initially found a continuous distribution of charge values. Following some tinkering with the apparatus, Morpurgo found that if he separated the capacitor plates he obtained only integral values of charge. "After some theoretical analysis, Morpurgo concluded that he now had his apparatus working properly, and reported his failure to find any evidence for fractional charges" (Pickering 1987, p. 197).

Pickering has made the important point that experimental apparatuses rarely work properly when they are first operated, and that some adjustment, or tinkering, is required before it does. He has also correctly pointed out that the theory of the apparatus and the theory of the phenomena can, and do, form part of the argument for the validity of an experimental result. He has, I believe, overemphasized theory. It was known, from Millikan onwards, that fractional charges, if they exist at all, are very rare in comparison with integral charges. The failure of Morpurgo's apparatus to find integral charges indicated quite strongly that, despite his initial theoretical analysis, it was not an accurate charge measuring device. Only after tinkering, when the apparatus measured integral charges, and thus passed a crucial experimental check, could one legitimately trust its measurements of charge. Although the modified theoretical analysis may have helped to clarify this, it was the experimental check that was crucial. There is more to an experimental apparatus than its theoretical analysis.

3. Critical Responses to Pickering

Ackermann has offered a modification of Pickering's view. He suggests that the experimental apparatus itself is a less plastic resource then either the theoretical model of the apparatus or that of the phenomenon.

> To repeat, changes in A [the apparatus] can often be seen (in real time, without waiting for accommodation by B [the theoretical model of the apparatus]) as improvements, whereas ‘improvements’ in B don't begin to count unless A is actually altered and realizes the improvements conjectured. It's conceivable that this small asymmetry can account, ultimately, for large scale directions of scientific progress and for the objectivity and rationality of those directions. (Ackermann 1991, p. 456)

Hacking (1992) has also offered a more complex version of Pickering's later view. He suggests that the

results of mature laboratory science achieve stability and are self-vindicating when the elements of laboratory science are brought into mutual consistency and support. These are (1) ideas: questions, background knowledge, systematic theory, topical hypotheses, and modeling of the apparatus; (2) things: target, source of modification, detectors, tools, and data generators; and (3) marks and the manipulation of marks: data, data assessment, data reduction, data analysis, and interpretation.

> Stable laboratory science arises when theories and laboratory equipment evolve in such a way that they match each other and are mutually self-vindicating. (1992, p. 56)
>
> We invent devices that produce data and isolate or create phenomena, and a network of different levels of theory is true to these phenomena. Conversely we may in the end count them only as phenomena only when the data can be interpreted by theory. (pp. 57-8)

One might ask whether or not such mutual adjustment between theory and experimental results can always be achieved? What happens when an experimental result is produced by an apparatus on which several of the epistemological strategies, discussed earlier, have been successfully applied, and the result is in disagreement with our theory of the phenomenon? Accepted theories can be refuted. Several examples will be presented below.

Hacking himself worries about what happens when a laboratory science that is true to the phenomena generated in the laboratory, thanks to mutual adjustment and self-vindication, is successfully applied to the world outside the laboratory. Does this argue for the truth of the science. In Hacking's view it does not. If laboratory science does produce happy effects in the "untamed world,... it is not the truth of anything that causes or explains the happy effects" (1992, p. 60).

There is a rather severe disagreement on the reasons for the acceptance of experimental results. For some, like Galison and myself, it is because of epistemological arguments. For others, like Pickering, the reasons are utility for future practice and agreement with existing theoretical commitments. Although the history of science shows that the overthrow of a well-accepted theory leads to an enormous amount of theoretical and experimental work, proponents of this view seem to accept it as unproblematical that it is always agreement with existing theory that has more future utility. Hacking and Pickering also suggest that experimental results are accepted on the basis of the mutual adjustment of elements which includes the theory of the phenomenon.

Nevertheless, everyone seems to agree that a consensus does arise on experimental results. The question then is how are these results used?

II. The Roles of Experiment

A. A Life of Its Own

Although experiment often takes its importance from its relation to theory, Hacking pointed out that it often has a life of its own, independent of theory. He notes the pristine observations of Carolyn Herschel's discovery of comets, William Herschel's work on "radiant heat," and Davy's observation of the gas emitted by algae and the flaring of a taper in that gas. In none of these cases did the experimenter have any theory of the phenomenon under investigation. One may also note the nineteenth century measurements of atomic spectra and the work on the masses and properties on elementary particles during the 1960s. Both of these sequences were conducted without any guidance from theory.

In deciding what experimental investigation to pursue, scientists may very well be influenced by the equipment available and their own ability to use that equipment (McKinney 1992). Thus, when the Mann-O'Neill collaboration was doing high energy physics experiments at the Princeton-Pennsylvania Accelerator during the late 1960s, the sequence of experiments was (1) measurement of the K^+ decay rates, (2) measurement of the K^{+}_{e3} branching ratio and decay spectrum, (3) measurement of the K^{+}_{e2} branching ratio, and (4) measurement of the form factor in K^{+}_{e3} decay. These experiments were performed with basically the same experimental apparatus, but with relatively minor modifications for each particular experiment. By the end of the sequence the experimenters had become quite expert in the use of the apparatus and knowledgeable about the backgrounds and experimental problems. This allowed the group to successfully perform the technically more difficult experiments later in the sequence. We might refer to this as "instrumental loyalty" and the "recycling of expertise" (Franklin 1997b). This meshes nicely with Galison's view of experimental traditions. Scientists, both theorists and experimentalists, tend to pursue experiments and problems in which their training and expertise can be used.

Hacking also remarks on the "noteworthy observations" on Iceland Spar by Bartholin, on diffraction by Hooke and Grimaldi, and on the dispersion of light by Newton. "Now of course Bartholin, Grimaldi, Hooke, and Newton were not mindless empiricists without an ‘idea’ in their heads. They saw what they saw because they were curious, inquisitive, reflective people. They were attempting to form theories. But in all these cases it is clear that the observations preceded any formulation of theory" (Hacking 1983, p. 156). In all of these cases we may say that these were observations waiting for, or perhaps even calling for, a theory. The discovery of any unexpected phenomenon calls for a theoretical explanation.

B. Confirmation and Refutation

Nevertheless several of the important roles of experiment involve its relation to theory. Experiment may confirm a theory, refute a theory, or give hints to the mathematical structure of a theory.

1. The Discovery of Parity Nonconservation: A Crucial Experiment

Let us consider first an episode in which the relation between theory and experiment was clear and straightforward. This was a "crucial" experiment, one that decided unequivocally between two competing theories, or classes of theory. The episode was that of the discovery that parity, mirror-reflection symmetry or left-right symmetry, is not conserved in the weak interactions. (For details of this episode

see Franklin (1986, Ch. 1) and Appendix 1). Experiments showed that in the beta decay of nuclei the number of electrons emitted in the same direction as the nuclear spin was different from the number emitted opoosite to the spin direction. This was a clear demonstartion of parity vilation in the weak interactions.

2. The Discovery of CP Violation: A Persuasive Experiment

After the discovery of parity and charge conjugation nonconservation, and following a suggestion by Landau, physicists considered CP (combined parity and particle-antiparticle symmetry), which was still conserved in the experiments, as the appropriate symmetry. One consequence of this scheme, if CP were conserved, was that the K_1^o meson could decay into two pions, whereas the K_2^o meson could not.[9] Thus, observation of the decay of K_2^o into two pions would indicate CP violation. The decay was observed by a group at Princeton University. Although several alternative explanations were offered, experiments eliminated each of the alternatives leaving only CP violation as an explanation of the experimental result. (For details of this episode see Franklin (1986, Ch. 3) and Appendix 2.)

3. The Discovery of Bose-Einstein Condensation: Confirmation After 70 Years

In both of the episodes discussed previously, those of parity nonconservation and of CP violation, we saw a decision between two competing classes of theories. This episode, the discovery of Bose-Einstein condensation (BEC), illustrates the confirmation of a specific theoretical prediction 70 years after the theoretical prediction was first made. Bose (1924) and Einstein (1924; 1925) predicted that a gas of noninteracting bosonic atoms will, below a certain temperature, suddenly develop a macroscopic population in the lowest energy quantum state.[10] (For details of this episode see Appendix 3.)

C. Complications

In the three episodes discussed in the previous section, the relation between experiment and theory was clear. The experiments gave unequivocal results and there was no ambiguity about what theory was predicting. None of the conclusions reached has since been questioned. Parity and CP symmetry are violated in the weak interactions and Bose-Einstein condensation is an accepted phenomenon. In the practice of science things are often more complex. Experimental results may be in conflict, or may even be incorrect. Theoretical calculations may also be in error or a correct theory may be incorrectly applied. There are even cases in which both experiment and theory are wrong. As noted earlier, science is fallible. In this section I will briefly discuss several episodes which illustrate these complexities.

1. The Fall of the Fifth Force

The episode of the fifth force is the case of a refutation of an hypoothesis, but only after a disagreement between experimental results was resolved. The "Fifth Force" was a proposed modification of Newton's Law of Universal Gravitation. The initial experiments gave conflicting results: one supported the

existence of the Fifth Force whereas the other argued against it. After numerous repetitions of the experiment, the discord was resolved and a consensus reached that the Fifth Force did not exist. (For details of this episode see Appendix 4.)

2. Right Experiment, Wrong Theory: The Stern-Gerlach Experiment

\11/

The Stern-Gerlach experiment was regarded as crucial at the time it was performed, but, in fact, wasn't. In the view of the physics community it decided the issue between two theories, refuting one and supporting the other. In the light of later work, however, the refutation stood, but the confirmation was questionable. In fact, the experimental result posed problems for the theory it had seemingly confirmed. A new theory was proposed and although the Stern-Gerlach result initially also posed problems for the new theory, after a modification of that new theory, the result confirmed it. In a sense, it was crucial after all. It just took some time.

The Stern-Gerlach experiment provides evidence for the existence of electron spin. These experimental results were first published in 1922, although the idea of electron spin wasn't proposed by Goudsmit and Uhlenbeck until 1925 (1925; 1926). One might say that electron spin was discovered before it was invented. (For details of this episode see Appendix 5).

3. Sometimes Refutation Doesn't Work: The Double-Scattering of Electrons

In the last section we saw some of the difficulty inherent in experiment-theory comparison. One is sometimes faced with the question of whether the experimental apparatus satisfies the conditions required by theory, or conversely, whether the appropriate theory is being compared to the experimental result. A case in point is the history of experiments on the double-scattering of electrons by heavy nuclei (Mott scattering) during the 1930s and the relation of these results to Dirac's theory of the electron, an episode in which the question of whether or not the experiment satisfied the conditions of the theoretical calculation was central. Initially, experiments disagreed with Mott's calculation, casting doubt on the underlying Dirac theory. After more than a decade of work, both experimental and theoretical, it was realized that there was a background effect in the experiments that masked the predicted effect. When the background was eliminated experiment and theory agreed. (Appendix 6)

D. Other Roles

1. Evidence for a New Entity: J.J. Thomson and the Electron

Experiment can also provide us with evidence for the existence of the entities involved in our theories. J.J. Thomson's experiments on cathode rays provided grounds for belief in the existence of electrons. (For details of this episode see Appendix 7).

2. The Articulation of Theory: Weak Interactions

Experiment can also help to articulate a theory. Experiments on beta decay during from the 1930s to the 1950s detremined the precise mathematical form of Fermi's theory of beta decay. (For details of this episode see Appendix 8.)

III. Conclusion

In this essay varying views on the nature of experimental results have been presented. Some argue that the acceptance of experimental results is based on epistemological arguments, whereas others base acceptance on future utility, social interests, or agreement with existing community commitments. Everyone agrees , however, that for whatever reasons, a consensus is reached on experimental results. These results then play many important roles in physics and we have examined several of these roles, although certainly not all of them. We have seen experiment deciding between two competing theories, calling for a new theory, confirming a theory, refuting a theory, providing evidence that determined the mathematical form of a theory, and providing evidence for the existence of an elementary particle involved in an accepted theory. We have also seen that experiment has a life of its own, independent of theory. If, as I believe, epistemological procedures provide grounds for reasonable belief in experimental results, then experiment can legitimately play the roles I have discussed and can provide the basis for scientific knowledge.

Bibliography

Principal Works:

- Ackermann, R. 1985. *Data, Instruments and Theory*. Princeton, N.J.: Princeton University Press.
- Ackermann, R. 1991. "Allan Franklin, Right or Wrong". *PSA 1990, Volume 2*. A. Fine, M. Forbes and L. Wessels (Ed.). East Lansing, MI, Philosophy of Science Association:451-457.
- Adelberger, E.G. 1989. "High-Sensitivity Hillside Results from the Eot-Wash Experiment". *Tests of Fundamental Laws in Physics: Ninth Moriond Workshop*. O. Fackler and J. Tran Thanh Van (Ed.). Les Arcs, France, Editions Frontieres:485-499.
- Anderson, M.H., J.R. Ensher, M.R. Matthews, et al. 1995. "Observation of Bose-Einstein Condensation in a Dilute Atomic Vapor". *Science* 269: 198-201.
- Bell, J.S. and J. Perring 1964. "2pi Decay of the K_2o Meson". *Physical Review Letters* 13: 348-349.
- Bennett, W.R. 1989. "Modulated-Source Eotvos Experiment at Little Goose Lock". *Physical Review Letters* 62: 365-368.
- Bizzeti, P.G., A.M. Bizzeti-Sona, T. Fazzini, et al. 1989a. "Search for a Composition Dependent Fifth Force: Results of the Vallambrosa Experiment". *Tran Thanh Van, J.* O. Fackler (Ed.). Gif Sur Yvette, Editions Frontieres.

- Bizzeti, P.G., A.M. Bizzeti-Sona, T. Fazzini, et al. 1989b. "Search for a Composition-dependent Fifth Force". *Physical Review Letters* 62: 2901-2904.
- Bose, S. 1924. "Plancks Gesetz und Lichtquantenhypothese". *Zeitschrift fur Physik* 26(1924): 178-181.
- Burnett, K. 1995. "An Intimate Gathering of Bosons". *Science* 269: 182-183.
- Cartwright, N. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.
- Chase, C. 1929. "A Test for Polarization in a beam of Electrons by Scattering". *Physical Review* 34: 1069-1074.
- Chase, C. 1930. "The Scattering of Fast Electrons by Metals. II. Polarization by Double Scattering at Right Angles". *Physical Review* 36: 1060-1065.
- Christenson, J.H., J.W. Cronin, V.L. Fitch, et al. 1964. "Evidence for the 2pi Decay of the K^o_2 Meson". *Physical Review Letters* 13: 138-140.
- Collins, H. 1985. *Changing Order: Replication and Induction in Scientific Practice*. London: Sage Publications.
- Collins, H. 1994. "A Strong Confirmation of the Experimenters' Regress". *Studies in History and Philosophy of Modern Physics* 25(3): 493-503.
- Conan Doyle, A. 1967. "The Sign of Four". *The Annotated Sherlock Holmes*. W. S. Barrington-Gould (Ed.). New York, Clarkson N. Potter.
- Cowsik, R., N. Krishnan, S.N. Tandor, et al. 1988. "Limit on the Strength of Intermediate-Range Forces Coupling to Isospin". *Physical Review Letters* 61(2179-2181).
- Cowsik, R., N. Krishnan, S.N. Tandor, et al. 1990. "Strength of Intermediate-Range Forces Coupling to Isospin". *Physical Review Letters* 64: 336-339.
- de Groot, S.R. and H.A. Tolhoek 1950. "On the Theory of Beta-Radioactivity I: The Use of Linear Combinations of Invariants in the Interaction Hamiltonian". *Physica* 16: 456-480.
- Dymond, E.G. 1931. "Polarisation of a Beam of Electrons by Scattering". *Nature* 128: 149.
- Dymond, E.G. 1932. "On the Polarisation of Electrons by Scattering". *Proceedings of the Royal Society (London)* A136: 638-651.
- Dymond, E.G. 1934. "On the Polarization of Electrons by Scattering. II.". *Proceedings of the Royal Society (London)* A145: 657-668.
- Einstein, A. 1924. "Quantentheorie des einatomigen idealen gases". *Sitzungsberischte der Preussische Akademie der Wissenschaften, Berlin*: 261-267.
- Einstein, A. 1925. "Quantentheorie des einatomigen idealen gases". *Sitzungsberichte der Preussische Akadmie der Wissenschaften, Berlin*: 3-14.
- Everett, A.E. 1965. "Evidence on the Existence of Shadow Pions in K^+ Decay". *Physical Review Letters* 14: 615-616.
- Fermi, E. 1934. "Attempt at a Theory of Beta-Rays". *Il Nuovo Cimento* 11: 1-21.
- Feynman, R.P. and M. Gell-Mann 1958. "Theory of the Fermi Interaction". *Physical Review* 109: 193-198.
- Feynman, R.P., R.B. Leighton and M. Sands 1963. *The Feynman Lectures on Physics*. Reading, MA: Addison-Wesley Publishing Company.
- Fierz, M. 1937. "Zur Fermischen Theorie des -Zerfalls". *Zeitschrift fur Physik* 104: 553-565.
- Fischbach, E., S. Aronson, C. Talmadge, et al. 1986. "Reanalysis of the Eötvös Experiment". *Physical Review Letters* 56: 3-6.

- Fitch, V.L. 1981. "The Discovery of Charge-Conjugation Parity Asymmetry". *Science* 212: 989-993.
- Fitch, V.L., M.V. Isaila and M.A. Palmer 1988. "Limits on the Existence of a Material-dependent Intermediate-Range Force". *Physical Review Letters* 60: 1801-1804.
- Ford, K.W. 1968. *Basic Physics*. Lexington: Xerox.
- Franklin, A. 1986. *The Neglect of Experiment*. Cambridge: Cambridge University Press.
- Franklin, A. 1990. *Experiment, Right or Wrong*. Cambridge: Cambridge University Press.
- Franklin, A. 1993. *The Rise and Fall of the Fifth Force: Discovery, Pursuit, and Justification in Modern Physics*. New York: American Institute of Physics.
- Franklin, A. 1994. "How to Avoid the Experimenters' Regress". *Studies in the History and Philosophy of Science* 25: 97-121.
- Franklin, A. 1995a. "The Resolution of Discordant Results". *Perspectives on Science* 3: 346-420.
- Franklin, A. 1995b. "Laws and Experiment". *Laws of Nature*. F. Weinert (Ed.). Berlin, De Gruyter:191-207.
- Franklin, A. 1996. "There Are No Antirealists in the Laboratory". *Realism and Anti-Realism in the Philosophy of Science*. R. S. Cohen, R. Hilpinen and Q. Renzong (Ed.). Dordrecht, Kluwer Academic Publishers**:** 131-148.
- Franklin, A. 1997a. "Calibration". *Perspectives on Science* 5: 31-80.
- Franklin, A. 1997b. "Recycling Expertise and Instrumental Loyalty". *Philosophy of Science* 64 (4 (Supp.)): S42-S52.
- Franklin, A. 1997c. "Are There Really Electrons? Experiment and Reality". *Physics Today* 50(10): 26-33.
- Franklin, A. and C. Howson 1984. "Why Do Scientists Prefer to Vary Their Experiments?". *Studies in History and Philosophy of Science* 15: 51-62.
- Franklin, A. and C. Howson 1988. "It Probably is a Valid Experimental Result: A Bayesian Approach to the Epistemology of Experiment". *Studies in the History and Philosophy of Science* 19: 419-427.
- Friedman, J.L. and V.L. Telegdi 1957. "Nuclear Emulsion Evidence for Parity Nonconservation in the Decay Chain pi - mu-e". *Physical Review* 105: 1681-1682.
- Galison, P. 1987. *How Experiments End*. Chicago: University of Chicago Press.
- Galison, P. 1997. *Image and Logic*. Chicago: University of Chicago Press.
- Gamow, G. and E. Teller 1936. "Selection Rules for the -Disintegration". *Physical Review* 49: 895-899.
- Garwin, R.L., L.M. Lederman and M. Weinrich 1957. "Observation of the Failure of Conservation of Parity and Charge Conjugation in Meson Decays: The Magnetic Moment of the Free Muon". *Physical Review* 105: 1415-1417.
- Gerlach, W. and O. Stern 1922a. "Der experimentelle Nachweis der Richtungsquantelung". *Zeitschrift fur Physik* 9: 349-352.
- Gerlach, W. and O. Stern 1924. "Uber die Richtungsquantelung im Magnetfeld". *Annalen der Physik* 74: 673-699.
- Gooding, D. 1992. "Putting Agency Back Into Experiment". *Science as Practice and Culture*. A. Pickering (Ed.). Chicago, University of Chicago Press**:** 65-112.
- Hacking, I. 1981. "Do We See Through a Microscope". *Pacific Philosophical Quarterly* 63: 305-

322.

- Hacking, I. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.
- Hacking, I. 1992. "The Self-Vindication of the Laboratory Sciences". *Science as Practice and Culture*. A. Pickering (Ed.). Chicago, University of Chicago Press:29-64.
- Halpern, O. and J. Schwinger 1935. "On the Polarization of Electrons by Double Scattering". *Physical Review* 48: 109-110.
- Hamilton, D.R. 1947. "Electron-Neutrino Angular Correlation in Beta-Decay". *Physical Review* 71: 456-457.
- Hellmann, H. 1935. "Bemerkung zur Polarisierung von Elektronenwellen durch Streuung". *Zeitschrift fur Physik* 96: 247-250.
- Hermannsfeldt, W.B., R.L. Burman, P. Stahelin, et al. 1958. "Determination of the Gamow-Teller Beta-Decay Interaction from the Decay of Helium-6". *Physical Review Letters* 1: 61-63.
- Kofoed-Hansen, O. 1955. "Neutrino Recoil Experiments". *Beta- and Gamma-Ray Spectroscopy*. K. Siegbahn (Ed.). New York, Interscience:357-372.
- Konopinski, E. and G. Uhlenbeck 1935. "On the Fermi Theory of Radioactivity". *Physical Review* 48: 7-12.
- Konopinski, E.J. and L.M. Langer 1953. "The Experimental Clarification of the Theory of -Decay". *Annual Reviews of Nuclear Science* 2: 261-304.
- Konopinski, E.J. and G.E. Uhlenbeck 1941. "On the Theory of Beta-Radioactivity". *Physical Review* 60: 308-320.
- Langer, L.M., J.W. Motz and H.C. Price 1950. "Low Energy Beta-Ray Spectra: Pm^{147} S^{35}". *Physical Review* 77: 798-805.
- Langer, L.M. and H.C. Price 1949. "Shape of the Beta-Spectrum of the Forbidden Transition of Yttrium 91". *Physical Review* 75: 1109.
- Langstroth, G.O. 1932. "Electron Polarisation". *Proceedings of the Royal Society (London)* A136: 558-568.
- Lee, T.D. and C.N. Yang 1956. "Question of Parity Nonconservation in Weak Interactions". *Physical Review* 104: 254-258.
- MacKenzie, D. 1989. "From Kwajelein to Armageddon? Testing and the Social Construction of Missile Accuracy". *The Uses of Experiment*. D. Gooding, T. Pinch and S. Shaffer (Ed.). Cambridge, Cambridge University Press: 409-435.
- Mayer, M.G., S.A. Moszkowski and L.W. Nordheim 1951. "Nuclear Shell Structure and Beta Decay. I. Odd A Nuclei". *Reviews of Modern Physics* 23: 315-321.
- McKinney, W. (1992). Plausibility and Experiment: Investigations in the Context of Pursuit. History and Philosophy of Science. Bloomington, IN, Indiana.
- Mehra, J. and H. Rechenberg 1982. *The Historical Development of Quantum Theory*. New York: Springer-Verlag.
- Millikan, R.A. 1911. "The Isolation of an Ion, A Precision Measurement of Its Charge, and the Correction of Stokes's Law". *Physical Review* 32: 349-397.
- Morrison, M. 1990. "Theory, Intervention, and Realism". *Synthese* 82: 1-22.
- Mott, N.F. 1929. "Scattering of Fast Electrons by Atomic Nuclei". *Proceedings of the Royal Society (London)* A124: 425-442.
- Mott, N.F. 1931. "Polarization of a Beam of Electrons by Scattering". *Nature* 128: 454.

- Mott, N.F. 1932. "Tha Polarisation of Electrons by Double Scattering". *Proceedings of the Royal Society (London)* A135: 429-458.
- Nelson, P.G., D.M. Graham and R.D. Newman 1990. "Search for an Intermediate-Range Composition-dependent Force Coupling to N-Z". *Physical Review D* 42: 963-976.
- Newman, R., D. Graham and P. Nelson 1989. "A "Fifth Force" Search for Differential Accleration of Lead and Copper toward Lead". *Tests of Fundamental Laws in Physics: Ninth Moriond Workshop*. O. Fackler and J. Tran Thanh Van (Ed.). Gif sur Yvette, Editions Frontieres:459-472.
- Nishijima, K. and M.J. Saffouri 1965. "CP Invariance and the Shadow Universe". *Physical Review Letters* 14: 205-207.
- Pais, A. 1982. *Subtle is the Lord...* Oxford: Oxford University Press.
- Pauli, W. 1933. "Die Allgemeinen Prinzipen der Wellenmechanik". *Handbuch der Physik* 24: 83-272.
- Petschek, A.G. and R.E. Marshak 1952. "The -Decay of Radium E and the Pseusoscalar Interaction". *Physical Review* 85: 698-699.
- Pickering, A. 1981. "The Hunting of the Quark". *Isis* 72: 216-236.
- Pickering, A. 1984a. *Constructing Quarks*. Chicago: University of Chicago Press.
- Pickering, A. 1984b. "Against Putting the Phenomena First: The Discovery of the Weak Neutral Current". *Studies in the History and Philosophy of Science* 15: 85-117.
- Pickering, A. 1987. "Against Correspondence: A Constructivist View of Experiment and the Real". *PSA 1986*. A. Fine and P. Machamer (Ed.). Pittsburgh, Philsophy of Science Association. 2: 196-206.
- Pickering, A. 1989. "Living in the Material World: On Realism and Experimental Practice.". *The Uses of Experiment*. D. Gooding, T. Pinch and S. Schaffer (Ed.). Cambridge, Cambridge University Press: 275-297.
- Prentki, J. 1965. *CP Violation*. Oxford International Conference on Elementary Particles, Oxford, England.
- Pursey, D.L. 1951. "The Interaction in the Theory of Beta Decay". *Philosophical Magazine* 42: 1193-1208.
- Raab, F.J. 1987. "Search for an Intermediate-Range Interaction: Results of the Eot-Wash I Experiment". *New and Exotic Phenomena: Seventh Moriond Workshop*. O. Fackler and J. Tran Thanh Van (Ed.). Les Arcs, France, Editions Frontieres: 567-577.
- Randall, H.M., R.G. Fowler, N. Fuson, et al. 1949. *Infrared Determination of Organic Structures*. New York: Van Nostrand.
- Richter, H. 1937. "Zweimalige Streuung schneller Elektronen". *Annalen der Physik* 28: 533-554.
- Ridley, B.W. (1954). Nuclear Recoil in Beta Decay. Physics. Cambridge, Cambridge University.
- Rose, M.E. and H.A. Bethe 1939. "On the Absence of Polarization in Electron Scattering". *Physical Review* 55: 277-289.
- Rupp, E. 1929. "Versuche zur Frage nach einer Polarisation der Elektronenwelle". *Zetschrift fur Physik* 53: 548-552.
- Rupp, E. 1930a. "Ueber eine unsymmetrische Winkelverteilung zweifach reflektierter Elektronen". *Zeitschrift fur Physik* 61: 158-169.
- Rupp, E. 1930b. "Ueber eine unsymmetrische Winkelverteilung zweifach reflektierter Elektronen". *Naturwissenschaften* 18: 207.

- Rupp, E. 1931. "Direkte Photographie der Ionisierung in Isolierstoffen". *Naturwissenschaften* 19: 109.
- Rupp, E. 1932a. "Versuche zum Nachweis einer Polarisation der Elektronen". *Physickalsche Zeitschrift* 33: 158-164.
- Rupp, E. 1932b. "Neure Versuche zur Polarisation der Elektronen". *Physikalische Zeitschrift* 33: 937-940.
- Rupp, E. 1932c. "Ueber die Polarisation der Elektronen bei zweimaliger 90^o - Streuung". *Zeitschrift fur Physik* 79: 642-654.
- Rupp, E. 1934. "Polarisation der Elektronen an freien Atomen". *Zeitschrift fur Physik* 88: 242-246.
- Rustad, B.M. and S.L. Ruby 1953. "Correlation between Electron and Recoil Nucleus in He^6 Decay". *Physical Review* 89: 880-881.
- Rustad, B.M. and S.L. Ruby 1955. "Gamow-Teller Interaction in the Decay of He^6". *Physical Review* 97: 991-1002.
- Sargent, B.W. 1932. "Energy Distribution Curves of the Disintegration Electrons". *Proceedings of the Cambridge Philosophical Society* 24: 538-553.
- Sargent, B.W. 1933. "The Maximum Energy of the -Rays from Uranium X and other Bodies". *Proceddings of the Royal Society (London)* A139: 659-673.
- Sauter, F. 1933. "Ueber den Mottschen Polarisationseffekt bei der Streuun von Elektronen an Atomen". *Annalen der Physik* 18: 61-80.
- Sellars, W. 1962. *Science, Perception, and Reality*. New York: Humanities Press.
- Sherr, R. and J. Gerhart 1952. "Gamma Radiation of C^{10}". *Physical Review* 86: 619.
- Sherr, R., H.R. Muether and M.G. White 1949. "Radioactivity of C^{10} and O^{14}". *Physical Review* 75: 282-292.
- Smith, A.M. 1951. "Forbidden Beta-Ray Spectra". *Physical Review* 82: 955-956.
- Stern, O. 1921. "Ein Weg zur experimentellen Prufung Richtungsquantelung im Magnet feld". *Zeitschrift fur Physik* 7: 249-253.
- Stubbs, C.W., E.G. Adelberger, B.R. Heckel, et al. 1989. "Limits on Composition-dependent Interactions using a Laboratory Source: Is There a "Fifth Force?"". *Physical Review Letters* 62: 609-612.
- Stubbs, C.W., E.G. Adelberger, F.J. Raab, et al. 1987. "Search for an Intermediate-Range Interaction". *Physical Review Letters* 58: 1070-1073.
- Sudarshan, E.C.G. and R.E. Marshak 1958. "Chirality Invariance and the Universal Fermi Interaction". *Physical Review* 109: 1860-1862.
- Thieberger, P. 1987a. "Search for a Substance-Dependent Force with a New Differential Accelerometer". *Physical Review Letters* 58: 1066-1069.
- Thomson, G.P. 1933. "Polarisation of Electrons". *Nature* 132: 1006.
- Thomson, G.P. 1934. "Experiment on the Polarization of Electrons". *Philosophical Magazine* 17: 1058-1071.
- Thomson, J.J. 1897. "Cathode Rays". *Philosophical Magazine* 44: 293-316.
- Uhlenbeck, G.E. and S. Goudsmit 1925. "Ersetzung der Hypothese von unmechanischen Zwang durch eine Forderung bezuglich des inneren Verhaltens jedes einzelnen Elektrons". *Naturwissenschaften* 13: 953-954.
- Uhlenbeck, G.E. and S. Goudsmit 1926. "Spinning Electrons and the Structure of Spectra". *Nature*

117: 264-265.

- Weinert, F. 1995. "Wrong Theory--Right Experiment: The Significance of the Stren-Gerlach Experiments". *Studies in History and Philosophy of Modern Physics* 26B(1): 75-86.
- Winter, J. 1936. "Sur la polarisation des ondes de Dirac". *Academie des Science, Paris, Comptes rendus hebdomadaires des seances* 202: 1265-1266.
- Wu, C.S. 1955. "The Interaction in Beta-Decay". *Beta- and Gamma-Ray Spectroscopy*. K. Siegbahn (Ed.). New York, Interscience: 314-356.
- Wu, C.S., E. Ambler, R.W. Hayward, et al. 1957. "Experimental Test of Parity Nonconservation in Beta Decay". *Physical Review* 105: 1413-1415.
- Wu, C.S. and A. Schwarzschild (1958). A Critical Examination of the He^6 Recoil Experiment of Rustad and Ruby. New York, Columbia University.

Other Suggested Reading

- Ackermann, R. 1988. "Experiments as the Motor of Scientific Progress". *Social Epistemology* 2: 327-335.
- Batens, D. and J.P. Van Bendegem, Eds. 1988. *Theory and Experiment*. Dordrecht: D. Reidel Publishing Company.
- Bogen, J. and J. Woodward 1988. "Saving the Phenomena". *The Philosophical Review* 97: 303-352.
- Gooding, D. 1990. *Experiment and the Making of Meaning*. Dordrecht: Kluwer Academic Publishers.
- Gooding, D., T. Pinch and S. Schaffer, Eds. 1989. *The Uses of Experiment*. Cambridge: Cambridge University Press.
- Koertge, N., Ed. 1998. *A House Built on Sand: Exposing Postmodernist Myths About Science*. Oxford: Oxford University Press.
- Nelson, A. 1994. "How Could Scientific Facts be Socially Constructed?". *Studies in History and Philosophy of Science* 25(4): 535-547.
- Pickering, A., Ed. 1992. *Science as Practice and Culture*. Chicago: University of Chicago Press.
- Pickering, A. 1995. *The Mangle of Practice*. Chicago: University of Chicago Press.
- Pinch, T. 1986. *Confronting Nature*. Dordrecht: Reidel.
- Rasmussen, N. 1993. "Facts, Artifacts, and Mesosomes: Practicing Epistemology with the Electron Microscope". *Studies in History and Philosophy of Science* 24: 227-265.
- Shapere, D. 1982. "The Concept of Observation in Science and Philosophy". *Philosophy of Science* 49: 482-525.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

confirmation | logic: inductive | rationalism vs. empiricism | scientific method | scientific realism

Allan Franklin
Allan.Franklin@Colorado.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 5, 1998
Content last modified: October 26, 1998

Notes to Experiment in Physics

* As the late Richard Feynman, one of the leading theoretical physicists of the twentieth century, wrote:

> The principle of science, the definition, almost, is the following: *The test of all knowledge is experiment*. Experiment is the *sole judge* of scientific 'truth'.
> (Feynman, Leighton and Sands 1963, p. 1-1)

In these postmodern times this might seem to be an old-fashioned view, but it is, I believe, correct. Not everyone would agree. As Andy Pickering has remarked,

> ...there is no obligation upon anyone framing a view of the world to take account of what twentieth-century science has to say.
> (Pickering 1984a, p. 413)

1. By valid, I mean that the experimental result has been argued for in the correct way, by use of epistemological strategies such as those discussed below.

2. See Franklin (1986, Ch. 6; and, 1990, Ch. 6) and Franklin and Howson (1984; 1988) for details of these strategies, along with a discussion of how they fit into a Bayesian philosophy of science

3. As Holmes remarked to Watson, "How often have I said to you that when you have eliminated the impossible, whatever remains, *however improbable*, must be the truth." (Conan Doyle 1967, p. 638)

4. It might be useful here to distinguish between the theory of the apparatus and the theory of the phenomenon. Ackermann is talking primarily about the later. It may not always be possible to separate these two theories. The analysis of the data obtained from an instrument may very well involve the theory of the phenomenon, but that doesn't necessarily cast doubt on the validity of the experimental result.

5. For another episode in which the elimination of background was crucial see the discussion of the measurement of the K^+_{e2} branching ratio in (Franklin 1990, pp. 115-31).

6. Collins offers two arguments concerning the difficulty, if not the virtual impossibility of replication. The first is philosophical. What does it mean to replicate an experiment? In what way is the replication similar to the original experiment? A rough and ready answer is that the replication measures the same physical quantity. Whether or not it, in fact, does so can, I believe, be argued for on reasonable grounds, as discussed earlier.

Collins' second argument is pragmatic. This is the fact that in practice it is often difficult to get an experimental apparatus, even one known to be similar to another, to work properly. Collins illustrates

this with his account of Harrison's attempts to construct two versions of a TEA leaser (Transverse Excited Atmospheric) (Collins 1985, pp. 51-78). Despite the fact that Harrison had previous experience with such lasers, and had excellent contacts with experts in the field, he had great difficulty in building the lasers. Hence the difficulty of replication.

Ultimately Harrison found errors in his apparatus and once these were corrected the lasers operated properly. As Collins admits, "...in the case of the TEA laser the circle was readily broken. The ability of the laser to vaporize concrete, or whatever, comprised a universally agreed criterion of experimental quality. There was never any doubt that the laser ought to be able to work and never any doubt about when one was working and when it was not." (Collins 1985, p. 84)

Although Collins seems to regard Harrison's problems with replication as casting light on the episode of gravity waves, as support for the experimenters' regress, and as casting doubt on experimental evidence in general, it really doesn't work. As Collins admits (see quote in last paragraph), the replication was clearly demonstrable. One may wonder what role Collins thinks this episode plays in his argument.

7. In more detailed discussions of this episode, Franklin (1994, 1997a), I argued that the gravity wave experiment is not at all typical of physics experiments. In most experiments, as illustrated in those essays, the adequacy of the surrogate signal used in the calibration of the experimental apparatus is clear and unproblematical. In cases where it is questionable considerable effort is devoted to establishing the adequacy of that surrogate signal. Although Collins has chosen an atypical example I believe that the questions he raises about calibration in general and about this particular episode of gravity wave experiments should be answered.

8. Weber had suggested that the actual gravity wave pulses were longer that expected, and that the nonlinear analysis algorithm was more efficient at detecting such pulses.

9. The K_1^o and K_2^o mesons were elementary particles with the same charge, mass, and intinsic spin. They did, however, differ with respect to the CP operator. The K_1^o and K_2^o mesons were eigenstates of the CP operator with eigenvalues CP = +1 and -1, respectively.

10. Bose's paper had originally been rejected by the *Philosophical Magazine*. He then sent it, in English, to Einstein with a request that if Einstein thought the paper merited publication that he would arrange for publication in the *Zeitschrift fur Physik*. Einstein personally translated the paper and submitted it to the *Zeitschrift fur Physik*, adding a translator's note, "In my opinion, Bose's derivation of the Planck formula constitutes an important advance. The method used here also yields the quantum theory of the ideal gas, as I shall discuss elsewhere in more detail" (Pais 1982, p. 423). This discussion appeared in Einstein's own papers of 1924 and 1925. For details see Pais (1982, Ch. 23).

11. This section is based on the accounts given by Weinert (1995) and by Mehra and Rechenberg (1982). Translations from the German were provided by these authors and are indicated by initials in the text.

Notes to Appendix 2

1. I surveyed eighty such theoretical papers. Sixty accepted the Princeton result as evidence for either CP violation or apparent CP violation. Even those that offered alternative explanations of the result were not necessarily indications that the authors did not accept CP violation. One should distinguish between interesting speculations and serious suggestions. The latter are characterized by a commitment to their truth. I note that T.D. Lee was author, or co-author, of three of these theoretical papers. Two offered alternative explanations of the Princeton result and one proposed a model that *avoided* CP violation. Lee was not seriously committed to the truth of any of them. Bell and Perring, authors of one of the alternatives, remarked, "Before a more mundane explanation is found *it is amusing to speculate* that it might be a local effect due to the dyssymmetry of the environment, namely the local preponderance of matter over antimatter"(Bell and Perring 1964, p. 348, emphasis added).

2. In the modus tollens if *h* entails *e* then "not *e*" entails not *h*. Duhem and Quine pointed out that it is really *h* and *b,* where *b* is background knowledge and auxiliary hypotheses, that entails *e*. Thus "not *e*" entails "*h*" or "*b*" and one doesn't know where to place the blame.

Notes to Appendix 3

1. Bose's paper had originally been rejected by the *Philosophical Magazine*. He then sent it, in English, to Einstein with a request that if Einstein thought the paper merited publication that he would arrange for publication in the *Zeitschrift fur Physik*. Einstein personally translated the paper and submitted it to the *Zeitschrift fur Physik*, adding a translator's note, "In my opinion, Bose's derivation of the Planck formula constitutes an important advance. The method used here also yields the quantum theory of the ideal gas, as I shall discuss elsewhere in more detail" (Pais 1982, p. 423). This discussion appeared in Einstein's own papers of 1924 and 1925. For details see Pais (1982, Ch. 23).

2. One difficulty with using rubidium is that at very low temperatures rubidium should be a solid. (In fact, rubidium is a solid at room temperature). Wieman, Cornell and their collaborators avoided this difficulty by creating a system that does not reach a true equilibrium. The vapor sample created equilibrates to a thermal distribution as a spin polarized gas, but takes a very long time to reach its true equilibrium state as a solid. At the low temperatures and density of the experiment the rubidium remains as a metastable super-saturated vapor for a long time.

Notes to Appendix 4

1. The original Eötvös experiment was designed to measure the ratio of the gravitational mass to the inertial mass of different substances. Eötvös found the ratio to be one, to within approximately one part in a million. Fischbach and his collaborators reanalyzed Eötvös' data and found a composition dependent effect, which they interpreted as evidence for a Fifth Force.

2. It is a fact of experimental life that experiments rarely work when they are initially turned on and that experimental results can be wrong, even if there is no apparent error. It is not necessary to know the exact source of an error in order to discount or to distrust a particular experimental result. Its disagreement with numerous other results can, I believe, be sufficient.

Notes to Appendix 6

1. Rupp's withdrawal included a note from a psychiatrist stating that Rupp had suffered from a mental illness and could not distinguish between fantasy and reality.

2. The problem with the hydrogen spectrum was not solved until the later discovery of the anomalous magnetic moment of the electron in the 1950s.

Notes to Appendix 7

1. Morrison (1990) has argued that manipulability is not sufficient to establish belief in an entity. She discusses particle physics experiments in which particle beams were viewed not only as particles, but also as beams of quarks, the constituents of the particles, even though the physicists involved had no belief in the existence of quarks. Although I believe that Morrison's argument is correct in this particular case, I do think that manipulability can, and often does, give us good reason to believe in an entity. See, for example, the discussion of the microscope in Hacking (1983).

2. Millikan, for example, used the properties of electrons emitted in the photoelectric effect to measure *h*, Planck's constant. Stern and Gerlach, as discussed below, used the properties of the electron to investigate spatial quantization, and also discovered evidence for electron spin.

3. For more details of this episode, including a discussion of other early twentieth century experiments, see (Franklin 1997c).

4. In Cartwright's discussion of the electron track in the cloud chamber, for example, she can identify the track as an electron track rather than as a proton track only because she has made an implicit commitment to the law of ionization for charged particles, and it's dependence on the mass and velocity of the particles.

5. Thomson also demonstrated the magnetic deflection of cathode rays in a separate experiment.

6. Thomson actually investigated the conductivity of the gas in the tube under varying pressure conditions. See Thomson (1897, pp. 298-300).

7. I shall return to this when I discuss Thomson's measurement of *e/m* for the electron.

8. Thomson's argument is the "duck argument." If it looks like a duck, quacks like a duck, and waddles like a duck, then we have good reason to believe that it is a duck. One need only reconstitute the argument using "it" as cathode rays and negatively charged particles as ducks.

9. Thomson actually used two different methods to determine the charge to mass ratio. The other method used the total charge carried by a beam of cathode rays in a fixed period of time, the total energy carried by the beam in that sane time, and the radius of curvature of the particles in a known magnetic field. Thomson regarded the method discussed in the text as more reliable and this is the method shown in most modern physics textbooks.

10. Not everything Thomson concluded is in agreement with modern views. Although he believed that the electron was a constituent of atoms, he thought that it was the primordial atom from which all atoms were constructed, similar to Prout's view that all atoms were constructed from hydrogen atoms. He also suggested that the charge on the electron might be larger than that of the hydrogen ion.

Notes to Appendix 8

1. The conservation of momentum also requires that the electron and proton have equal and opposite momenta, for a neutron decay at rest. They will be emitted in opposite directions.

2. Pauli's suggestion was first made in a December 4, 1930 letter to the radioactive group at a regional meeting in Tuebingen.

> Dear Radioactive Ladies and Gentlemen:
>
> I beg you to receive graciously the bearer of this letter who will report to you in detail how I have hit on a desperate was to escape from the problems of the "wrong" statistics of the N and Li^6 nuclei and of the continuous beta spectrum in order to save the "even-odd" rule of statistics and the law of conservation of energy. Namely the possibility that electrically neutral particles, which I would like to call neutrons [the particle we call the neutron, which is about the same mass as the proton, was discovered in 1932 by Chadwick. Pauli's "neutron" is our "neutrino."] might exist inside nuclei; these would have spin 1/2, would obey the exclusion principle, and would in addition duffer from photons through the fact that they would not travel at the speed of light. The mass of the neutron ought to be about the same order of magnitude as the electron mass, and in any case could not be greater than 0.01 proton masses. The continuous beta spectrum would then become understandable by assuming that in beta decay a neutron is always emitted along with the electron, in such a way that the sum of the energies of the neutron and electron is a constant.

> Now, the question is, what forces act on the neutron? The most likely model for the neutron seems to me, on wave mechanical grounds, to be the assumption that the motionless neutron is a magnetic dipole with a certain magnetic moment μ (the bearer of this letter can supply details). The experiments demand that the ionizing power of such a neutron cannot exceed that of a gamma ray, and therefore μ probably cannot be greater than $e(10^{-13}\text{cm})$. [*e* is the charge of the electron].
>
> At the moment I do not dare to publish anything about this idea, so I first turn trustingly to you, dear radioactive friends, with the question: how could such a neutron be experimentally identified if it possessed about the same penetrating power as a gamma ray or perhaps 10 times greater penetrating power?
>
> I admit that my way out may look rather improbable at first since if the neutron existed it would have been seen long ago. But nothing ventured, nothing gained. The gravity of the situation with the continuous beta spectrum was illuminated by a remark by my distinguished predecessor in office, Mr. DeBye, who recently said to me in Brussels, "Oh, that's a problem like the new taxes; one had best not think about it at all." So one ought to discuss seriously any way that may lead to salvation. Well, dear radioactive friends, weigh it and pass sentence! Unfortunately, I cannot appear personally in Tubingen, for I cannot get away from Zurich on account of a ball which is held here on the night of December 6-7. With best regards to you and to Mr. Baek,
>
> Your most obedient servant,
> W. Pauli
>
> (Quoted in Ford 1968, p. 849.)

This was a serious suggestion and although Pauli did not get all the properties of the neutrino right his suggestion was the basis of further work.

3. With a three-body decay the electron and the proton also didn't have to come off back to back. This observation was not made until the late 1930s. Assigning the neutrino a spin, intrinsic angular momentum, of h/4 also preserved the law of conservation of angular momentum.

4. The actual history is more complex. For a time, an alternative theory of decay, proposed by Konopinski and Uhlenbeck (1935) was better supported by the experimental evidence than was Fermi's theory. It was subsequently shown that both the experimental results and the theoretical calculations were wrong and that Fermi's theory was , in fact, supported by the evidence. For details see (Franklin 1990).

5. Allowed transitions are those for which both the electron and neutrino wavefunctions could be considered constant over nuclear dimensions. Forbidden transitions are those that included higher order terms in the perturbation series expansion of the matrix element.

6. I have been unable to find a published reference to this measurement. It is cited as a private communication in the literature.

7. In a post-deadline paper presented at the January 1958 meeting of the American Physical Society, Rustad and Ruby suggested that their earlier result might be wrong. There are no abstracts of post-deadline papers, but the talk was cited in the literature. Ruby remembers the tone of the paper as *mea culpa* (private communication).

Stanford Encyclopedia of Philosophy
Supplement to Experiment in Physics

Appendix 1: The Discovery of Parity Nonconservation

Let us consider first an episode in which the relation between theory and experiment was clear and straightforward. This was a "crucial" experiment, one that decided unequivocally between two competing theories, or classes of theory. The episode was that of the discovery that parity, mirror-reflection symmetry or left-right symmetry, is not conserved in the weak interactions. (For details of this episode see Franklin (1986, Ch. 1)). Parity conservation was a well-established and strongly-believed principle of physics. As students of introductory physics learn, if we wish to determine the magnetic force between two currents we first determine the direction of the magnetic field due to the first current, and then determine the force exerted on the second current by that field. We use two Right-Hand Rules. We get exactly the same answer, however, if we use two Left-Hand Rules, This is left-right symmetry, or parity conservation, in electromagnetism.

In the early 1950s physicists were faced with a problem known as the "τ-θ" puzzle. Based on one set of criteria, that of mass and lifetime, two elementary particles (the tau and the theta) appeared to be the same, whereas on another set of criteria, that of spin and intrinsic parity, they appeared to be different. T.D. Lee and C.N. Yang (1956) realized that the problem would be solved, and that the two particles would be different decay modes of the same particle, if parity were not conserved in the decay of the particles, a weak interaction. They examined the evidence for parity conservation and found, to their surprise, that although there was strong evidence that parity was conserved in the strong (nuclear) and electromagnetic interactions, there was, in fact, no supporting evidence that it was conserved in the weak interaction. It had never been tested.

Lee and Yang suggested several experiments that would test their hypothesis that parity was not conserved in the weak interactions. One was the β decay of oriented nuclei (Figure 1). Consider a collection of radioactive nuclei, all of whose spins point in the same direction. Suppose also that the electron given off in the radioactive decay of the nucleus is always emitted in a direction opposite to the spin of the nucleus In the mirror the electron is emitted in the same direction as the spin. The mirror image of the decay is different from the real decay. This would violate parity conservation, or mirror symmetry. Parity would be conserved only if, in the decay of a collection of nuclei, equal numbers of electrons were emitted in both directions. This was the experimental test performed by C.S. Wu and her collaborators (1957). They aligned Cobalt^{60} nuclei and counted the number of decay electrons in the two directions, along the nuclear spin and opposite to the spin. Their results are shown in Figure 2 and indicate clearly that more electrons are emitted opposite to the spin than along the spin. Parity is not conserved.

Two other experiments, reported at the same time, on the sequential decay pi meson decays to mu meson decays to electron also showed parity nonconservation (Friedman and Telegdi 1957; Garwin, Lederman and Weinrich 1957). These three experiments decided between two classes of theories--that is, between those theories that conserve parity and those that do not. They refuted the theories in which parity was conserved and supported or confirmed those in which it wasn't. These experiments also demonstrated that charge conjugation, or particle-antiparticle, symmetry was violated in the weak interactions and called for a new theory of decay and the weak interactions. It is fair to say that when a physicist learned the results of these experiments they were convinced that parity was not conserved in the weak interactions.

Allan Franklin
Allan.Franklin@Colorado.edu

Return to Experiment in Physics

First published: October 5, 1998
Content last modified: October 5, 1998

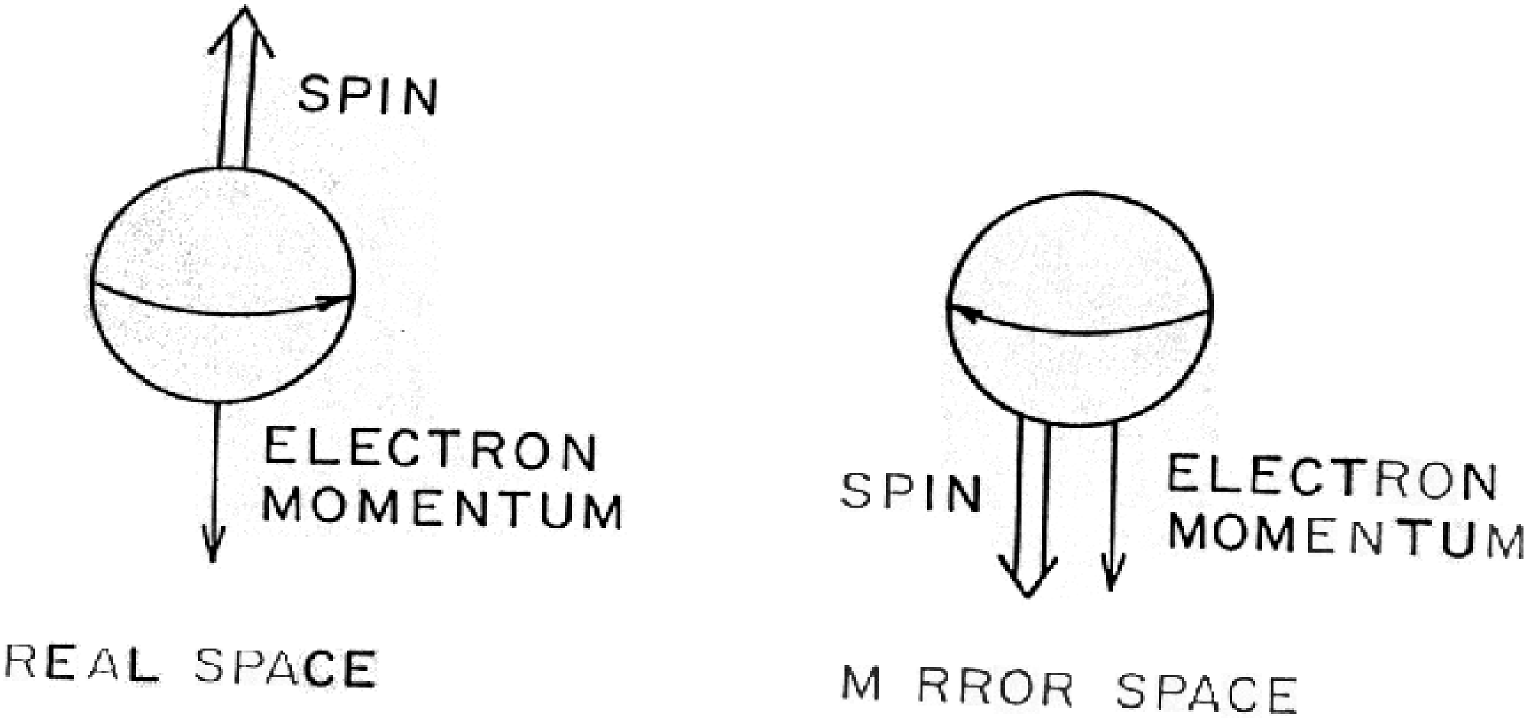
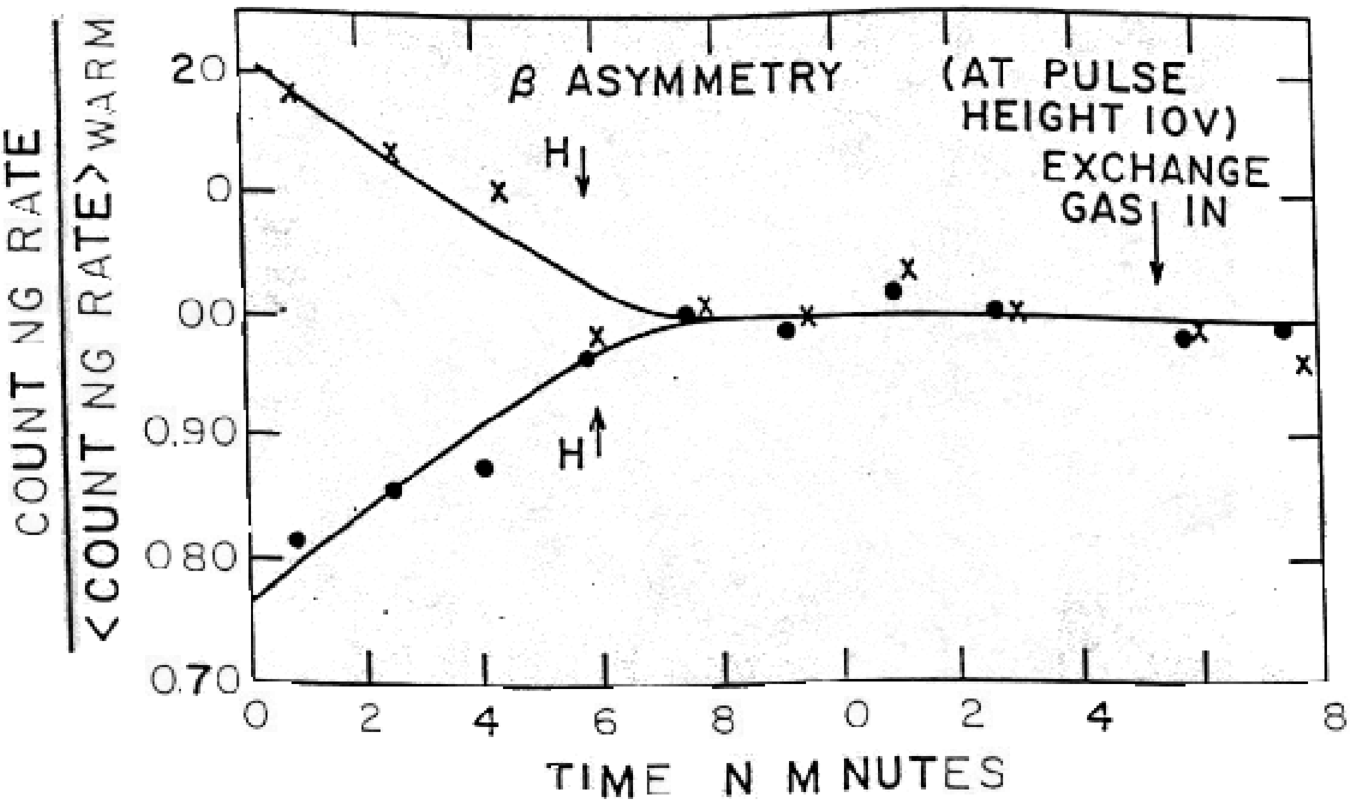


Figure 1. Nuclear spin and momentum of the decay electron in decay in both real space and in mirror space.



igu

Figure 2. Relative counting rates for particles from the decay of oriented ^{60}Co nuclei for different nuclear orientations (field directions). There is a clear asymmetry with more particles being emitted opposite to the spin direction. From Wu et al. (1957).

Stanford Encyclopedia of Philosophy
Supplement to Experiment in Physics

Appendix 2: The Discovery of CP Violation: A Persuasive Experiment

A group at Princeton University, led by Cronin and Fitch, decided to test CP conservation. The experimenters were quite aware of the relevance of their experiment to the question of CP violation, but they did not expect to observe it. As Val Fitch, one of the group leaders remarked, "Not many of our colleagues would have given credit for studying CP invariance, but we did so anyway" (Fitch 1981, p. 991). A preliminary estimate indicated that the CP phase of the experiment would detect about 7500 K_2^o decays and thus reduce the limit on CP violation from the then current limit of 1/300 (0.3%) to 1/7500 (For details of this episode see Franklin (1986, Ch. 3)).

The experimental beam contained only K_2^o mesons. (The K_1^o meson has a much shorter lifetime than the K_2^o meson, so that if we start with a beam containing both types of particles, after a time only the K_2^o mesons will remain). The experimental apparatus detected two charged particles from the decay of the K_2^o meson. The vector momentum of each of the two decay products from the K_2^o beam and the invariant mass m^* were computed assuming that each product had the mass of a pion:

$$m^* = [(E_1 + E_2)^2 - (\mathbf{p}_1 + \mathbf{p}_2)^2]^{½},$$

where E and **p** are the energy and vector momenta of the pions, respectively. If both particles were indeed pions from K_2^o decay, m^* would equal the K_2^o mass. The experimenters also computed the vector sum of the two momenta and the angle between this sum and the direction of the K_2^o beam. This angle should be zero for two-body decays, but not, in general, for three-body decays.

This was exactly what the Princeton group observed (Christenson et al. 1964). As seen clearly in Figure 3, there is a peak at the K^o mass, 498 MeV/c^2, for events with cos(θ) greater than 0.9999 (cos(θ) approximately equal to 1 means θ is approximately equal to 0). No such peak is seen in the mass regions just above or just below the K^o mass. The experimenters reported a total of 45 ± 9 two-pion K_2^o decays out of a total of 22,700 K_2^o decays. This was a branching ratio of (1.95 ± 0.2) x 10^{-3}, or approximately 0.2 percent.

The most obvious interpretation of the Princeton result was that CP symmetry was violated. This was the view taken in three out of four theoretical papers written during the period immediately following the

report of that result. The Princeton result had persuaded most of the physics community that CP symmetry was violated. The remaining theoretical papers offered alternative explanations.\1/ These alternatives relied on one or more of three arguments: (1) the Princeton results are caused by a CP asymmetry (the local preponderance of matter over antimatter) in the environment of the experiment, (2) K_2^o decay into two pions does not necessarily imply CP violation, and (3) the Princeton observations did not arise from two-pion K_2^o decay. This last argument can divided into the assertions that (3a) the decaying particle was not a K_2^o meson, (3b) the decay products were not pions, and (3c) another unobserved particle was emitted in the decay. Included in these alternatives were three suggestions that cast doubt on well-supported fundamental assumptions of modern physics. These were: (1) pions are not bosons, (2) the principle of superposition in quantum mechanics is violated, and (3) the exponential decay law fails. Although by the end of 1967 all of these alternatives had been experimentally tested and found wanting, the majority of the physics community had accepted CP violation by the end of 1965, even though all the tests had not yet been completed. As Prentki, a theoretical particle physicist, remarked, this was because in some cases "the price one has to pay in order to save CP becomes extremely high," and because other alternatives were "even more unpleasant"(Prentki 1965).

This is an example of what one might call a pragmatic solution to the Duhem-Quine problem.\2/ The alternative explanations and the auxiliary hypotheses were refuted, leaving CP violation unprotected. One might worry that other plausible alternatives were never suggested or considered. This is not a serious problem in the actual practice of physics. No fewer than ten alternative explanations of the Princeton result were offered, and not all of them were very plausible. Had others been suggested they, too, would have been considered by the physics community. Consider the model of Nishijima and Saffouri (1965). They explained two-pion K_2^o decay by the existence of a "shadow" universe in touch with our "real" universe only through the weak interactions. They attributed to the two pion decay observed to the decay of the $K^{o'}$ from the shadow universe. This implausible model was not merely considered, it was also experimentally tested. Everett (1965) noted that if the $K^{o'}$, the shadow K^o postulated by Nishijima and Saffouri existed, then a shadow pion should also exist, and the decays of the K^+ into a positive pion and a neutral pion and of the K^+ into a positive pion and a neutral shadow pion and should occur with equal rates. The presence of the shadow pion could be detected by measuring the ordinary K^+ branching ratio in two different experiments, one in which the neutral pion was detected and one in which it was not. If the shadow pion existed the two measurements would differ. They didn't. There was no shadow pion and thus, no $K^{o'}$.

What was the difference between the episodes of parity nonconservation and CP violation. In the former parity nonconservation was immediately accepted. No alternative explanations were offered. There was a convincing and decisive set of experiments. In the latter at least ten alternatives were proposed, and although CP violation was accepted rather quickly, the alternatives were tested. In both cases there are only two classes of theories, those that conserve parity or CP, and those that do not. The difference lies in the length and complexity of the derivation linking the hypothesis to the experimental result, or to the number of auxiliary hypotheses required for the derivation. In the case of parity nonconservation the experiment could be seen by inspection to violate mirror symmetry (See Figure 1). In the CP episode

what was observed was K_2^o decay into two pions. In order to connect this observation to CP conservation one had to assume (1) the principle of superposition, (2) that the exponential decay law held to 300 lifetimes, (3) that the decay particles were both "real" pions and that pions were bosons, (4) that no other particle was emitted in the decay, (5) that no other similar particle was produced, and (6) that there were no external conditions present that might regenerate K_1^o mesons. It was these auxiliary assumptions that were tested and eliminated as alternative explanations by subsequent experiments.

The discovery of CP violation called for a theoretical explanation, a call that is still unanswered.

Allan Franklin
Allan.Franklin@Colorado.edu

Return to Experiment in Physics

First published: October 5, 1998
Content last modified: October 5, 1998

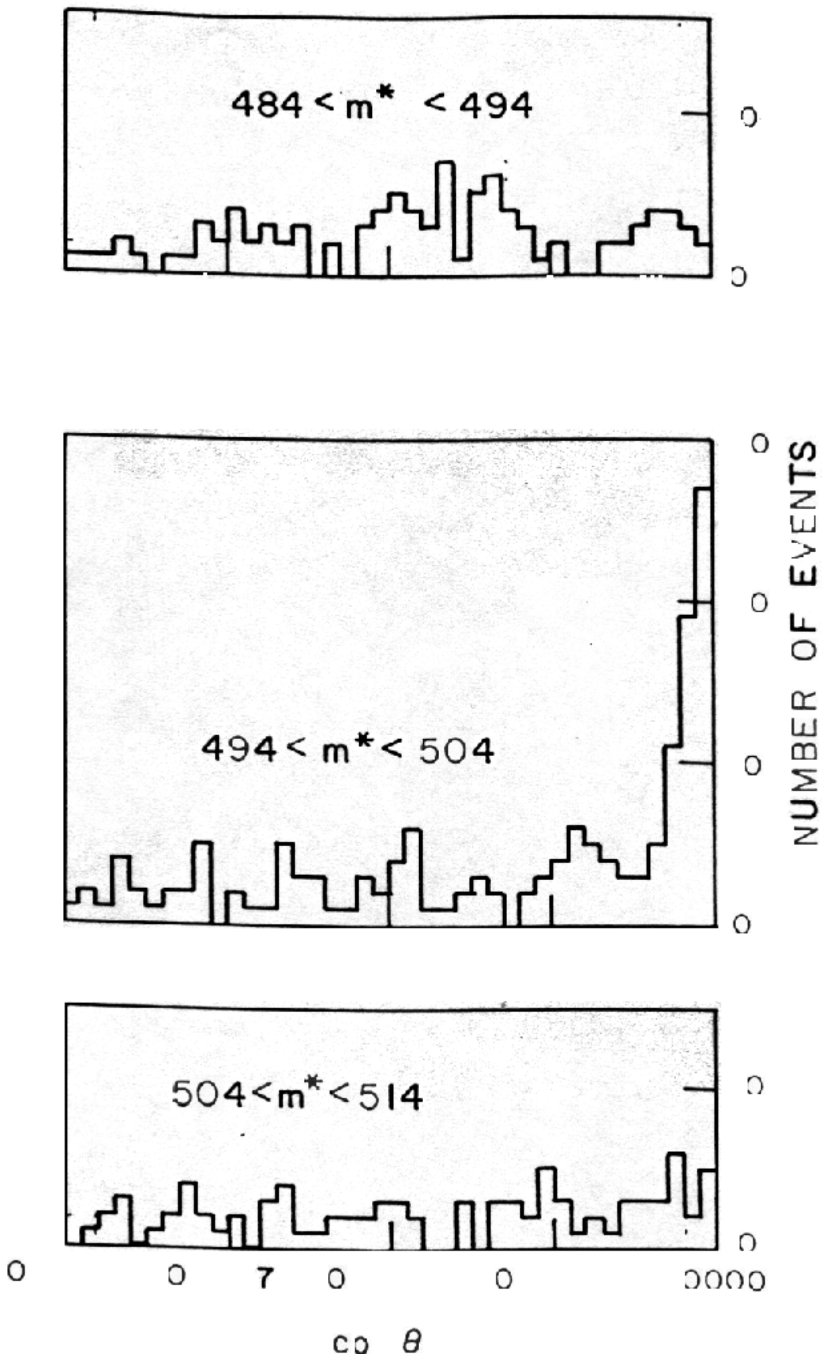


Figure 3. Angular distributions in three mass ranges for events with cos(theta) > 0.9995. From Christenson et al. (1964).

Stanford Encyclopedia of Philosophy
Supplement to Experiment in Physics

Appendix 3: The Discovery of Bose-Einstein Condensation: Confirmation After 70 Years

In both of the episodes discussed previously, those of parity nonconservation and of CP violation, we saw a decision between two competing classes of theories. This episode, the discovery of Bose-Einstein condensation (BEC), illustrates the confirmation of a specific theoretical prediction 70 years after the theoretical prediction was first made. Bose (1924) and Einstein (1924; 1925) predicted that a gas of noninteracting bosonic atoms will, below a certain temperature, suddenly develop a macroscopic population in the lowest energy quantum state.\1/ An interesting aspect of this episode is that the phenomenon in question had never been observed previously. This raises an interesting epistemological problem. How do you know you have observed something that has never been seen before?

Elementary particles can be divided onto two classes: bosons with integral spin (0, 1 ,2, ...), and fermions with half-integral spin (1/2, 3/2, 5/2, ...). Fermions, such as electrons obey the Pauli Exclusion Principle. Two fermions cannot be in the same quantum mechanical state. This explains the shell structure of electrons in atoms and the periodic table. On the other hand, any number of bosons can occupy the same state. At sufficiently low temperatures, when thermal motions are very small, there is a strong tendency for a group of bosons to all go into the same state.

The experiment that first demonstrated the existence of BEC was done by Carl Wieman, Eric Cornell, and their collaborators (Anderson et al. 1995). The experimental apparatus is shown is Figure 4. In outline the experiment was as follows. A sample of ^{87}Rb atoms was cooled in a magneto-optical trap. It was then loaded into a magnetic trap and further cooled by evaporation. The condensate was formed and the trap removed, allowing the condensate to expand. The expanded condensate was illuminated with laser light and the resulting shadow of the cloud was imaged, digitized, and stored.\2/

The experimental results are shown in Figures 5 - 7. Figure 5 shows the velocity distribution of the rubidium gas cloud (a) just before the appearance of the condensate, (b) just after, and (c) after further evaporation of the cloud has left a sample of nearly pure condensate. This figure also shows the spatial distribution of the gas. Although the measurement process destroyed the condensate sample, the entire process can be repeated so that one can measure the cloud at different stages. Figure 6 shows the peak density of the gas as a function of the RF frequency used to excite the atoms into a non-confined state and to assist the cooling by evaporation). There is a sharp increase in density at a frequency of 4.23 MHz. This indicates the appearance of Bose-Einstein condensation. As the sample is further cooled one expects to observe a two-component cloud with a dense central condensate surrounded by a diffuse non-

condensate. This is seen clearly in both Figures 5 and 7. Figure 7 shows horizontal sections of the rubidium cloud. At 4.71 MHz, above the transition temperature, one sees only a broad thermal distribution. Beginning at 4.23 MHz one sees the appearance of a sharp central peak, the Bose-Einstein condensate, above the thermal distribution. At 4.11 MHz the cloud is almost a pure condensate.

There are three clear indications of the presence of Bose-Einstein condensation: (1) the velocity distribution of the gas shows two distinct components, (2) the sudden increase in density as the temperature decreases, and (3) the elliptical shape of the velocity distribution (Figure 5). The velocity distribution should be elliptical because for the harmonic trap used, the force in the z direction was eight times larger than in the x and y directions. No phenomenon other than Bose-Einstein condensation could plausibly explain these results

This result was sufficiently credible that Keith Burnett, an atomic physicist at Oxford University remarked, in the same issue of *Science* in which Wieman and Cornell reported their result, "In short, they have observed the phenomenon called Bose-Einstein condensation (BEC) in a gas of atoms for the first time. The term Holy Grail seems quite appropriate given the singular importance of this discovery" (Burnett 1995, p. 182).

A theoretical prediction had been confirmed after 70 years.

Allan Franklin
Allan.Franklin@Colorado.edu

Return to Experiment in Physics

First published: October 5, 1998
Content last modified: October 5, 1998

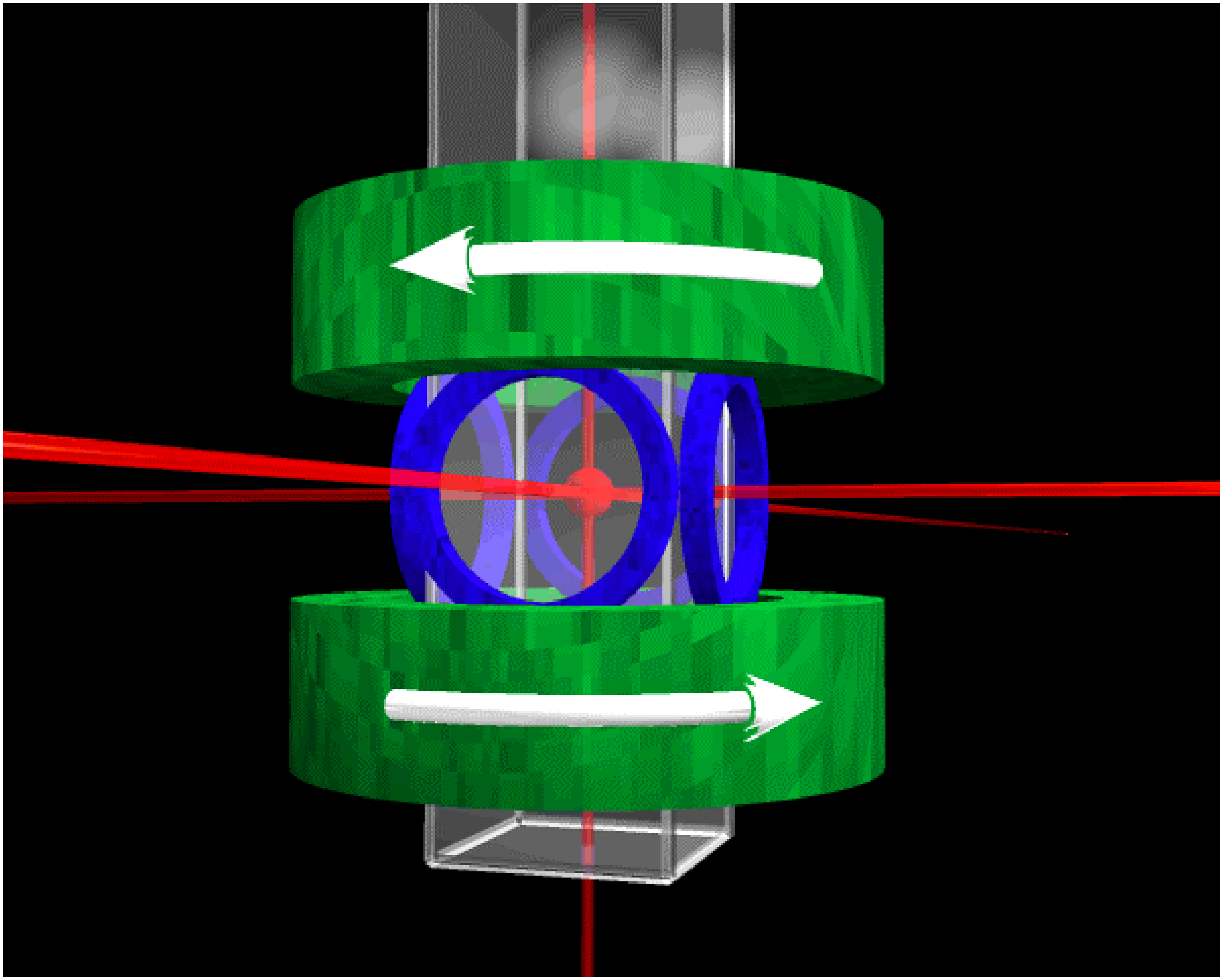


Figure 4. Schematic of the BEC apparatus. From Anderson et al. (1995).

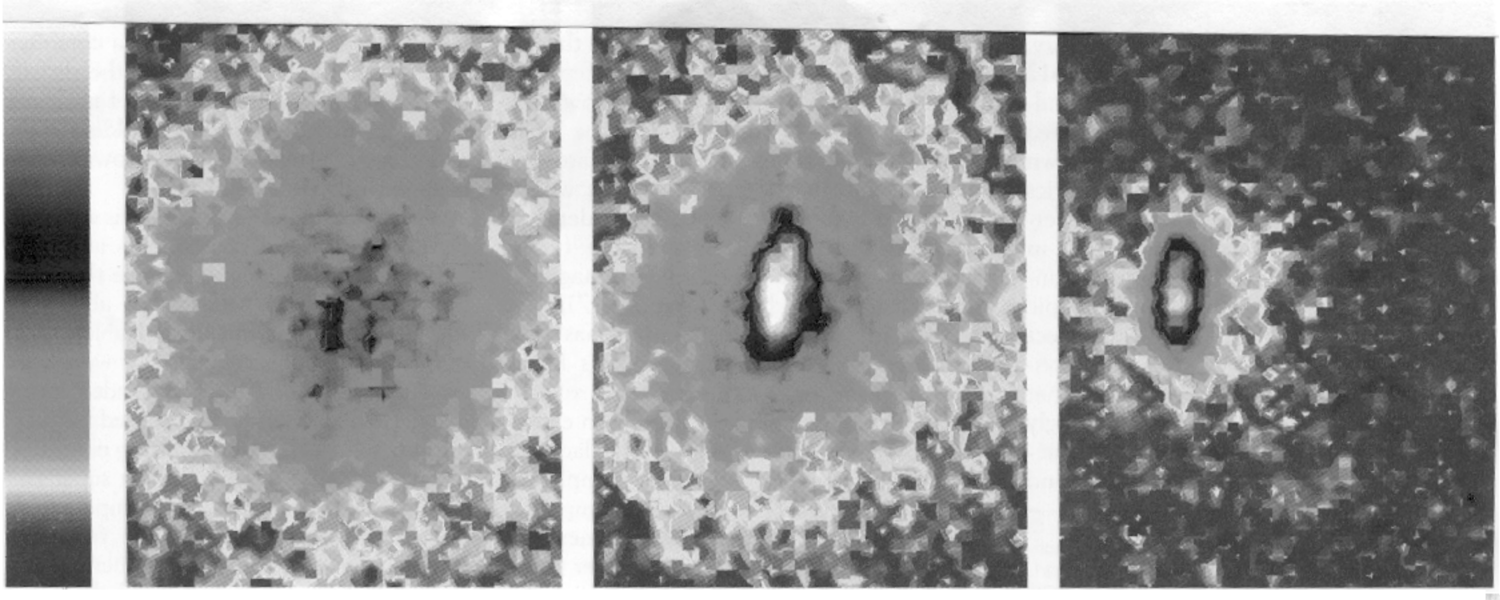


Figure 5. False color images of the velocity distribution of the rubidium BEC cloud (from the left): just before the appearance of the condensate, just after the appearance of the condensate, and after further evaporation has left a sample of nearly pure condensate. From Anderson et al. (1995).

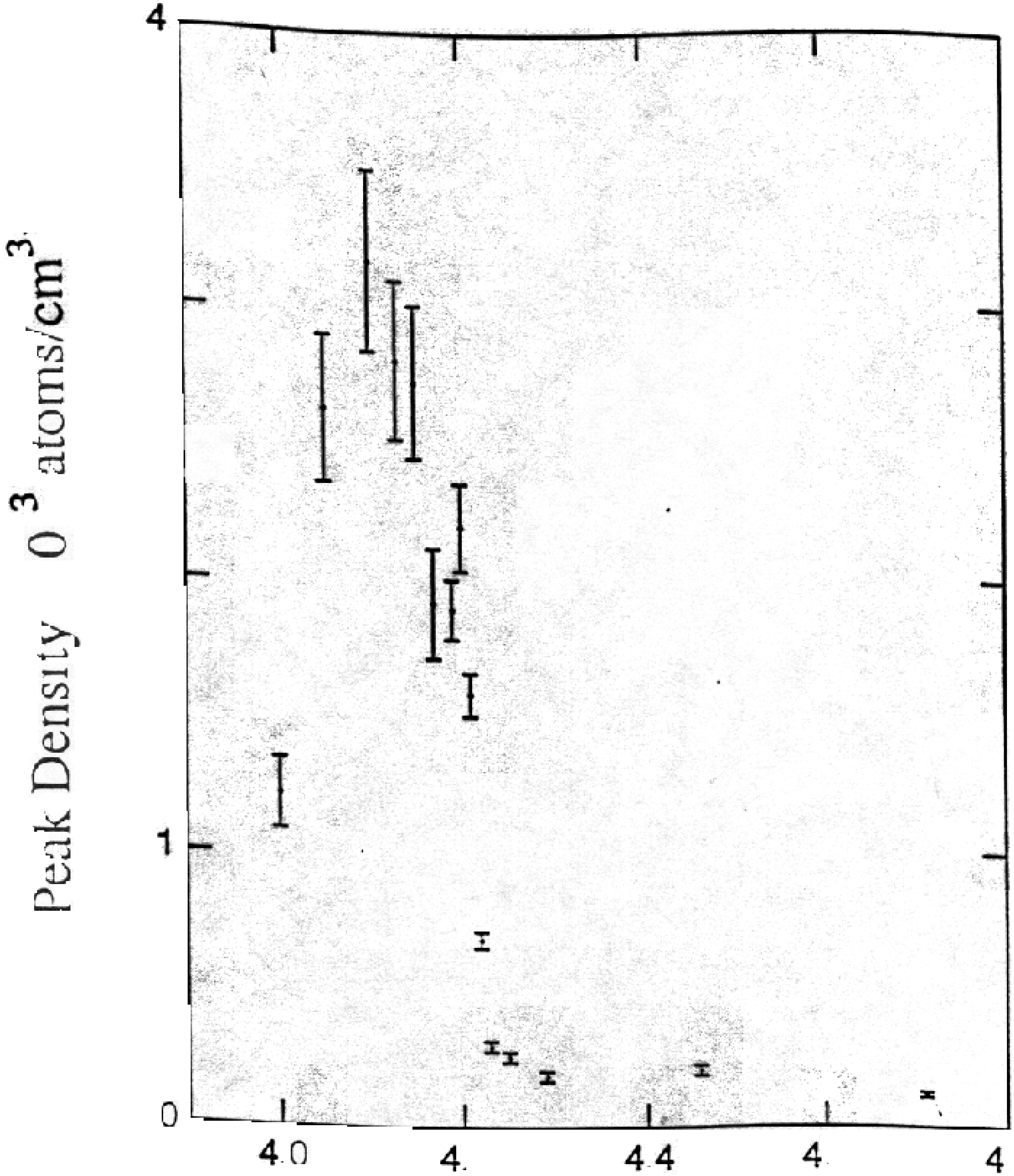


Figure 6. Peak density at the center of the sample as a function of the final depth of the evaporative cut

on the RF frequency. As evaporation progresses to smaller values of the frequency, the cloud shrinks and cools, causing a modest increase in peak density until the frequency reaches 4.23 MHz. The sudden discontinuity at 4.23 MHz indicates the first appearance of the high-density condensate as the cloud undergoes a phase transition. From Anderson et al. (1995).

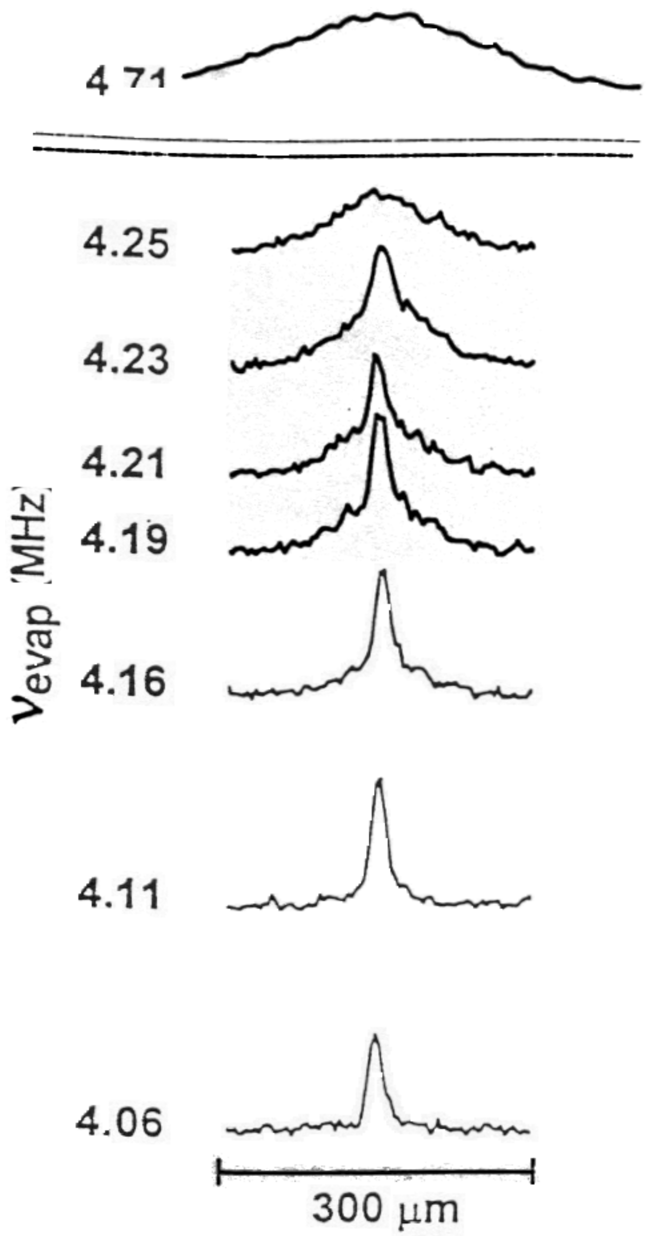


Figure 7. Horizontal sections taken through the velocity distribution at progressively lower values of the RF frequency show the appearance of the condensate fraction. From Anderson et al. (1995).

Stanford Encyclopedia of Philosophy
Supplement to Experiment in Physics

Appendix 4: The Fall of the Fifth Force

In this episode we will examine a case of the refutation of a hypothesis, but only after a disagreement between experimental results was resolved. The "Fifth Force" was a proposed modification of Newton's Law of Universal Gravitation. The initial experiments gave conflicting results: one supported the existence of the Fifth Force whereas the other regued against it. After numerous repetitions of the experiment, the discord was resolved and a consensus reached that the Fifth Force Did not exist. A reanalysis of the original Eötvös experiment[\1/] by Fischbach and his collaborators (1986) had shown a suggestive deviation from the law of gravity. The Fifth Force, in contrast to the famous Galileo experiment, depended on the composition of the objects. Thus, the Fifth Force between a copper mass and an aluminum mass would differ from that between a copper mass and a lead mass. Fischbach and collaborators also suggested modifying the gravitational potential between two masses from

$$V = -Gm_1m_2/r$$

to

$$V = -Gm_1m_2/r\ [1 + (\alpha)e^{-r/\lambda}],$$

where the second term gives the Fifth Force with strength α and range λ. The reanalysis also suggested that α was approximately 0.01 and λ was approximately 100m. (For details of this episode see (Franklin 1993)).

In this episode, we have a hitherto unobserved phenomenon along with discordant experimental results. The first two experiments on the Fifth Force gave contradictory answers. One experiment supported the existence of the Fifth Force, whereas the other found no evidence for it. The first experiment, that of Peter Thieberger (1987a) looked for a composition-dependent force using a new type of experimental apparatus, which measured the differential acceleration between copper and water. The experiment was conducted near the edge of the Palisades cliff in New Jersey to enhance the effect of an intermediate-range force. The experimental apparatus is shown in Figure 8. The horizontal acceleration of the copper sphere relative to the water can be determined by measuring the steady-state velocity of the sphere and applying Stokes' law for motion in a resistive medium. Thieberger's results are shown in Figure 9. The sphere clearly has a velocity, indicating the presence of a force. Thieberger concluded, "The present results are compatible with the existence of a medium-range, substance-dependent force" (p. 1068).

The second experiment, by the whimsically named Eöt-Wash group, was also designed to look for a substance-dependent, intermediate range force (Raab 1987; Stubbs et al. 1987). The apparatus was located on a hillside on the University of Washington campus, in Seattle (Figure 10). If the hill attracted the copper and beryllium bodies differently, then the torsion pendulum would experience a net torque. This torque could be observed by measuring shifts in the equilibrium angle of the torsion pendulum as the pendulum was moved relative to a fixed geophysical point. Their experimental results are shown in Figure 11. The theoretical curves were calculated with the assumed values of 0.01 and 100m, for the Fifth Force parameters α and λ, respectively. These were the best values for the parameters at the time. There is no evidence for such a Fifth Force in this experiment.

The problem was, however, that both experiments appeared to be carefully done, with no apparent mistakes in either experiment. Ultimately, the discord between Thieberger's result and that of the Eöt-Wash group was resolved by an overwhelming preponderance of evidence in favor of the Eöt-Wash result (The issue was actually more complex. There were also discordant results on the distance dependence of the Fifth Force. For details see Franklin (1993; 1995a)). The subsequent history is an illustration of one way in which the scientific community deals with conflicting experimental evidence. Rather than making an immediate decision as to which were the valid results, this seemed extremely difficult to do on methodological or epistemological grounds, the community chose to await further measurements and analysis before coming to any conclusion about the evidence. The torsion-balance experiments of Eöt-Wash were repeated by others including (Cowsik et al. 1988; Fitch, Isaila and Palmer 1988; Adelberger 1989; Bennett 1989; Newman, Graham and Nelson 1989; Stubbs et al. 1989; Cowsik et al. 1990; Nelson, Graham and Newman 1990). These repetitions, in different locations and using different substances, gave consistently negative results. In addition, Bizzeti and collaborators (1989a; 1989b), using a float apparatus similar to that of Thieberger, also obtained results showing no evidence of a Fifth Force. There is, in fact, no explanation of either Thieberger's original, presumably incorrect, results. The scientific community has chosen, I believe quite reasonably, to regard the preponderance of negative results as conclusive.[\2/] Experiment had shown that there is no Fifth Force.

Allan Franklin

Allan.Franklin@Colorado.edu

Return to Experiment in Physics

First published: October 5, 1998
Content last modified: October 5, 1998

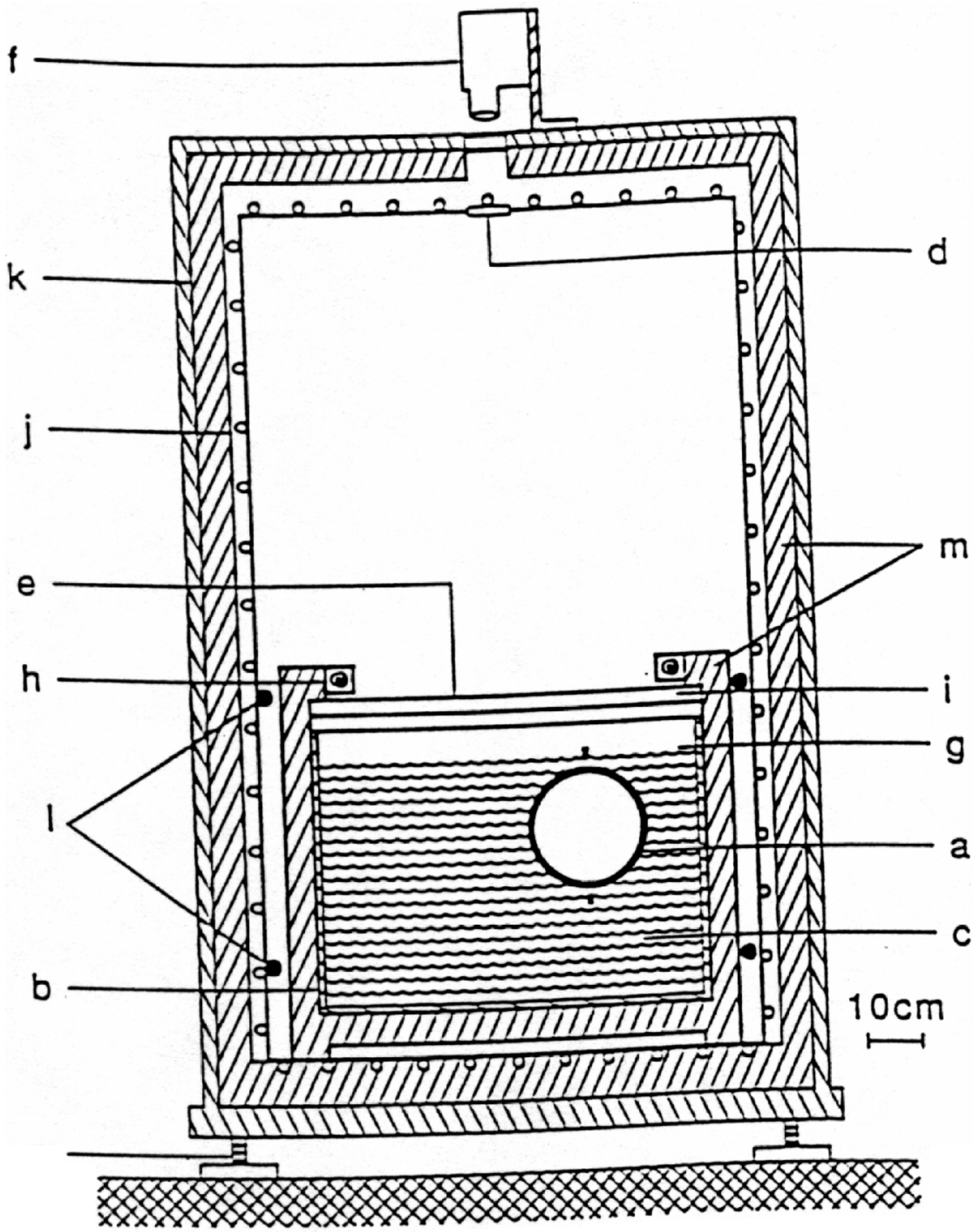


Figure 8. Schematic diagram of the differential accelerometer used in Thieberger's experiment. A precisely balanced hollow copper sphere (a) floats in a copper-lined tank (b) filled with distilled water (c). The sphere can be viewed through windows (d) and (e) by means of a television camera (f). The multiple-pane window (e) is provided with a transparent x-y coordinate grid for position determination on top with a fine copper mesh (g) on the bottom. The sphere is illuminated for one second per hour by four lamps (h) provided with infrared filters (i). Constant temperature is maintained by mea ns of a thermostatically controlled copper shield (j) surrounded by a wooden box lined with Styrofoam insulation (m). The Mumetal shield (k) reduces possible effects du e to magnetic field gradients and four circular coils (l) are used for positioning the sphere through forces due to ac-produced eddy currents, and

for dc tests. From Thieberger (1987).

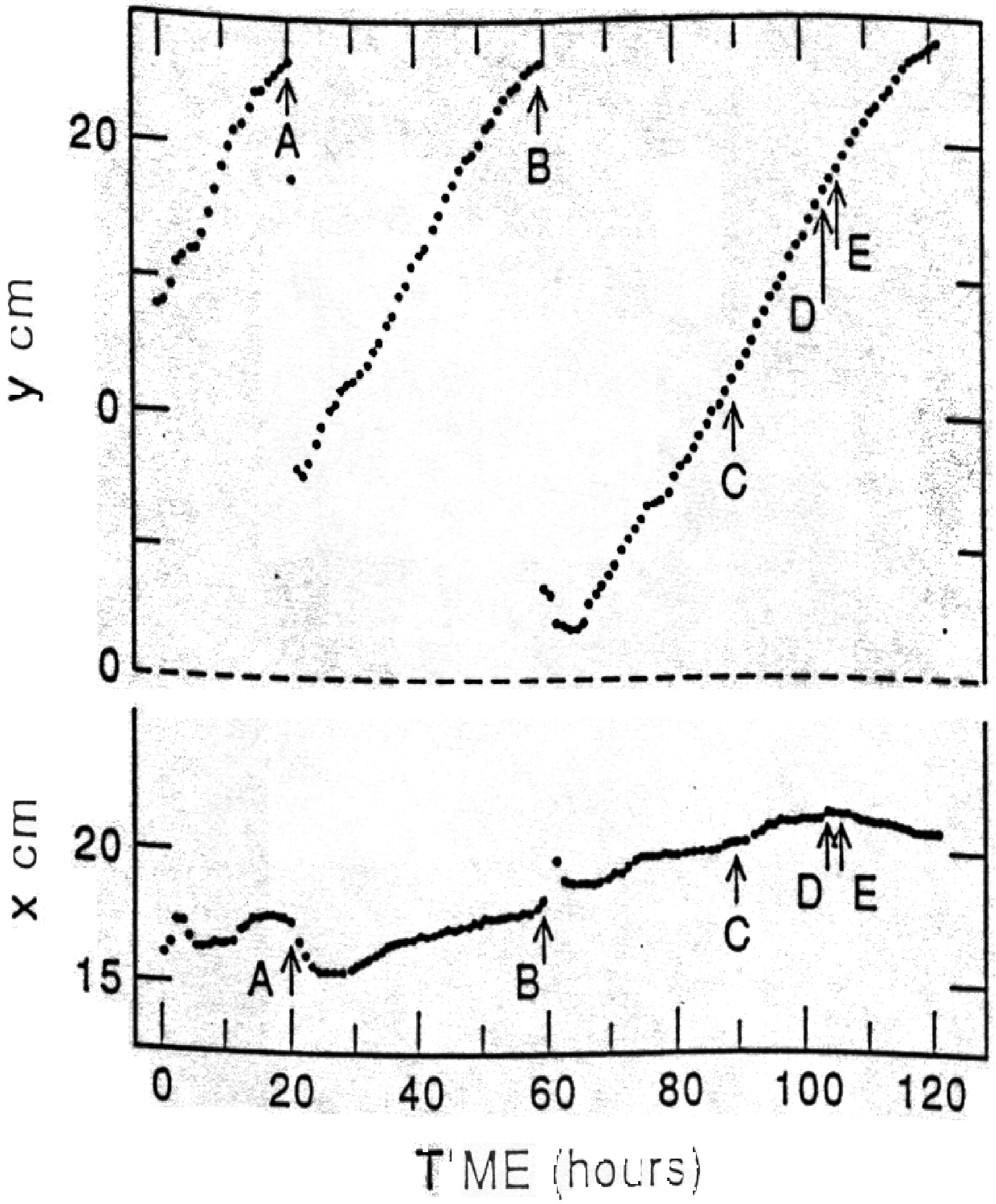


Figure 9. Position of the center of the sphere as a function of time. The y axis points away from the cliff. The position of the sphere was reset at points A and B by engaging the coils shown in Figure 21. From Thieberger (1987).

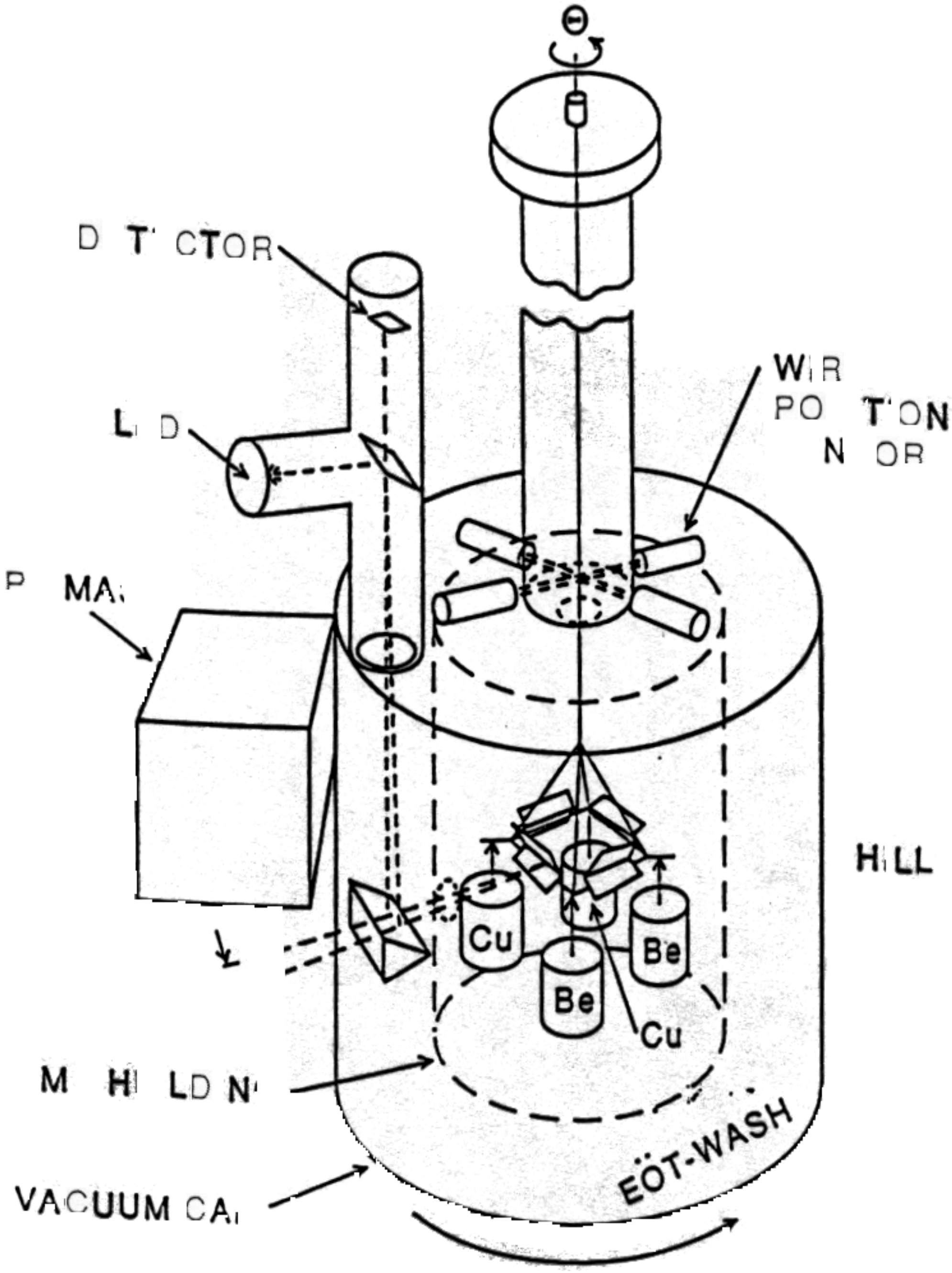


Figure 10. Schematic view of the University of Washington torsion pendulum experiment. The Helmholtz coils are not shown. From Stubbs et al. (1987).

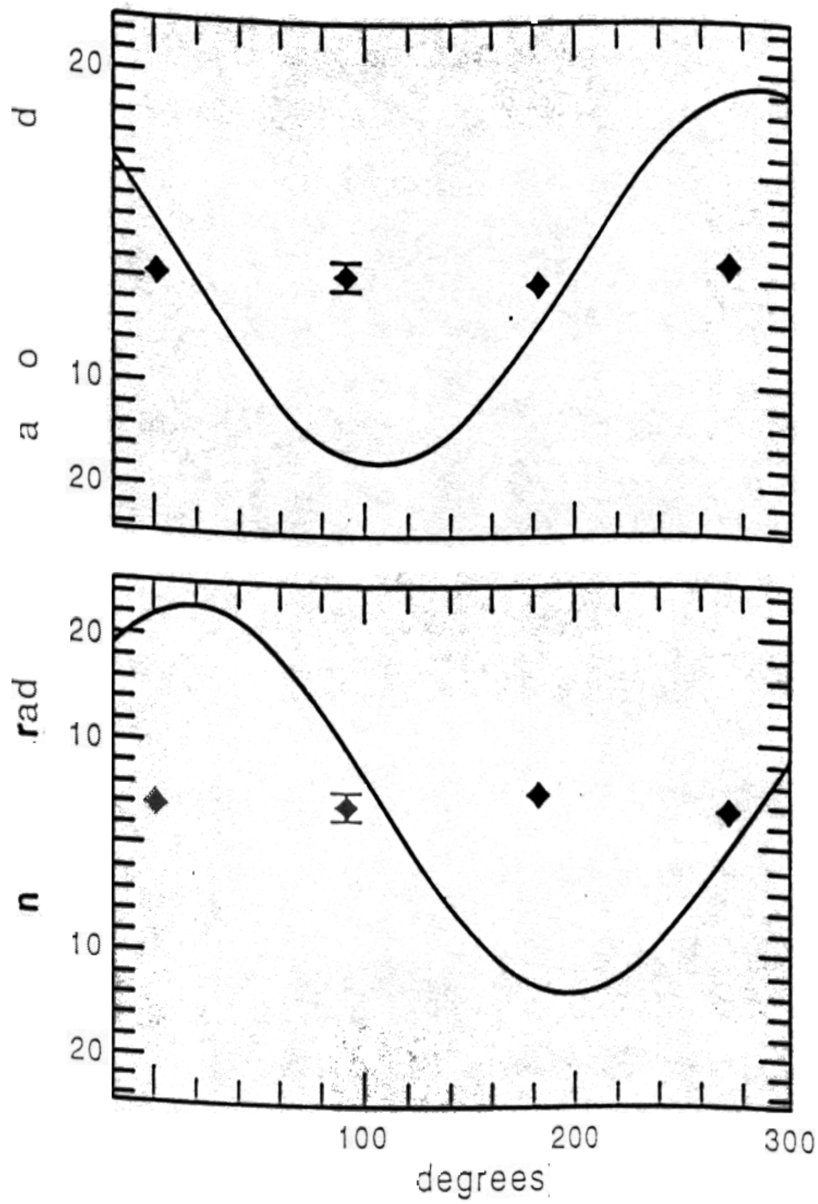


Figure 11. Deflection signal as a function of . The theoretical curves correspond to the signal expected for alpha = 0.01 and lambda = 100m. From Raab (1987).

Stanford Encyclopedia of Philosophy
Supplement to Experiment in Physics

Appendix 5: Right Experiment, Wrong Theory: The Stern-Gerlach Experiment

From the time of Ampere onward, molecular currents were regarded as giving rise to magnetic moments. In the nuclear model of the atom the electron orbits the nucleus. This circular current results in a magnetic moment. The atom behaves as if it were a tiny magnet. In the Stern-Gerlach experiment a beam of silver atoms passed through an inhomogeneous magnetic field (Figure 12). In Larmor's classical theory there was no preferential direction for the direction of the magnetic moment and so one predicted that the beam of silver atoms would show a maximum in the center of the beam. In Sommerfeld's quantum theory an atom in a state with angular momentum equal to one (L = 1) would have a magnetic moment with two components relative to the direction of the magnetic field, $\pm eh/4m_e$. (Bohr had argued that only two spatial components were allowed). In an inhomogeneous magnetic field, H, the force on the magnetic moment μ will be μ_z x (Gradient of the magnetic field in the z direction), where $\mu_z = \pm eh/4m_e$, where e is the charge of the electron, m_e is its mass, h is Planck's constant, and z is the field direction. Thus, depending on the orientation of the magnetic moment relative to the magnetic field there will be either an attractive or repulsive force and the beam will split into two components, exhibiting spatial quantization. There will be a minimum at the center of the beam. "According to quantum theory μ_z can only be $\pm (e/2m_e)(h/2\pi)$. In this case the spot on the receiving plate will therefore be split into two, each of them having the same size but half the intensity of the original spot" (Stern 1921, p. 252, JM) This difference in prediction between the Larmor and Sommerfeld theories was what Stern and Gerlach planned to use to distinguish between the two theories. Stern remarked that "the experiment, if it can be carried out, (will result) in a clear-cut decision between the quantum-theoretical and the classical view" (Stern 1921, FW).

Sommerfeld's theory also acted as an enabling theory for the experiment. It provided an estimate of the size of the magnetic moment of the atoms so that Stern could begin calculations to see if the experiment was feasible. Stern calculated, for example, that a magnetic field gradient of 10^4 Gauss per centimeter would be sufficient to produce deflections that would give detectable separations of the beam components. He asked Gerlach if he could produce such a gradient. Gerlach responded affirmatively, and said he could do even better. The experiment seemed feasible. A sketch of the apparatus is shown in Figure 12. The silver atoms pass through the inhomogeneous magnetic field. If the beam is spatially quantized, as Sommerfeld predicted, two spots should be observed on the screen. (The sketch shows the beam splitting into three components, which would be expected in modern quantum theory for an atom with angular momentum equal to one). I note that Sommerfeld's theory was incorrect, illustrating the point that an enabling theory need not be correct to be useful.

A preliminary result reported by Stern and Gerlach did not show splitting of the beam into components. It did, however, show a broadened beam spot. They concluded that although they had not demonstrated spatial quantization, they had provided "evidence that the silver atom possesses a magnetic moment." Stern and Gerlach made improvements in the apparatus, particularly in replacing a round beam slit by a rectangular one that gave a much higher intensity. The results are shown in Figure 13 (Gerlach and Stern 1922a). There is an intensity minimum in the center of the pattern, and the separation of the beam into two components is clearly seen. This result seemed to confirm Sommerfeld's quantum-theoretical prediction of spatial quantization. Pauli, a notoriously skeptical physicist, remarked, "Hopefully now even the incredulous Stern will be convinced about directional quantization" (in a letter from Pauli to Gerlach 17 February 1922). Pauli's view was shared by the physics community. Nevertheless the Stern-Gerlach result posed a problem for the Bohr-Sommerfeld theory of the atom. Stern and Gerlach had assumed that the silver atoms were in an angular momentum state with angular momentum equal to one (L = 1). In fact, the atoms are in an L = 0 state, for which no splitting of the beam would be expected in either the classical or the quantum theory. Stern and Gerlach had not considered this possibility. Had they done so they might not have done the experiment. The later, or new, quantum theory developed by Heisenberg, Schrodinger, and others, predicted that for an L = 1 state the beam should split into three components as shown in Figure 12. The magnetic moment of the atom would be either 0 or ± eh/(4π x m). Thus, if the silver atoms were in an L = 1 state as Stern and Gerlach had assumed, their result, showing two beam components, also posed a problem for the new quantum theory. This was solved when Uhlenbeck and Goudsmit (1925, 1926) proposed that the electron had an intrinsic angular momentum or spin equal to h/4π. This is analogous to the earth having orbital angular momentum about the sun and also an intrinsic angular momentum due to its rotation on its own axis. In an atom the electron will have a total angular momentum **J** = **L** + **S**, where **L** is the orbital angular momentum and **S** is the spin of the electron. For silver atoms in an L = 0 state the electron would have only its spin angular momentum and one would expect the beam to split into two components. Goudsmit and Uhlenbeck suggested the idea of electron spin to explain features in atomic spectra such as the anomalous Zeeman effect, the splitting of spectral lines in a magnetic field into more components than could be accommodated by the Bohr-Sommerfeld theory of the atom. Although the Stern-Gerlach results were known, and would certainly have provided strong support for the idea of electron spin, Goudsmit and Uhlenbeck made no mention of the result.

The Stern-Gerlach experiment was initially regarded as a crucial test between the classical theory of the atom and the Bohr-Sommerfeld theory. In a sense it was, because it showed clearly that spatial quantization existed, a phenomenon that could be accommodated only within a quantum mechanical theory. It decided between the two classes of theories, the classical and the quantum mechanical. With respect to the particular quantum theory of Bohr and Sommerfeld, however, it wasn't crucial, although it was regarded as such at the time, because that theory predicted no splitting for a beam of silver atoms in the ground state (L = 0). The theory had been wrongly applied. The two-component result was also problematic for the new quantum theory, which also predicts no splitting for an angular momentum zero state and three components for an L = 1 state. Only after the suggestion of electron spin did the Stern-Gerlach result confirm the new theory.

Although the interpretation of the experimental result was incorrect for a time, the result itself remained quite robust through the theory change from the old to the new quantum theory. It is important to remember that experimental results do not change when accepted theory changes, although certainly, as we have seen, their interpretation may change. Gerlach and Stern emphasized this point themselves.

Apart from any theory, it can be stated, as a pure result of the experiment, and as far as the exactitude of our experiments allows us to say so, that silver atoms in a magnetic field have only *two discrete* values of the component of the magnetic moment in the direction of the field strength; both have the same absolute value with each half of the atoms having a positive and a negative sign respectively (Gerlach and Stern 1924, pp. 690-691, FW)

Experimental results, as well as experiments, also have a life of their own, independent of theory.

Allan Franklin
Allan.Franklin@Colorado.edu

Return to Experiment in Physics

First published: October 5, 1998
Content last modified: October 5, 1998

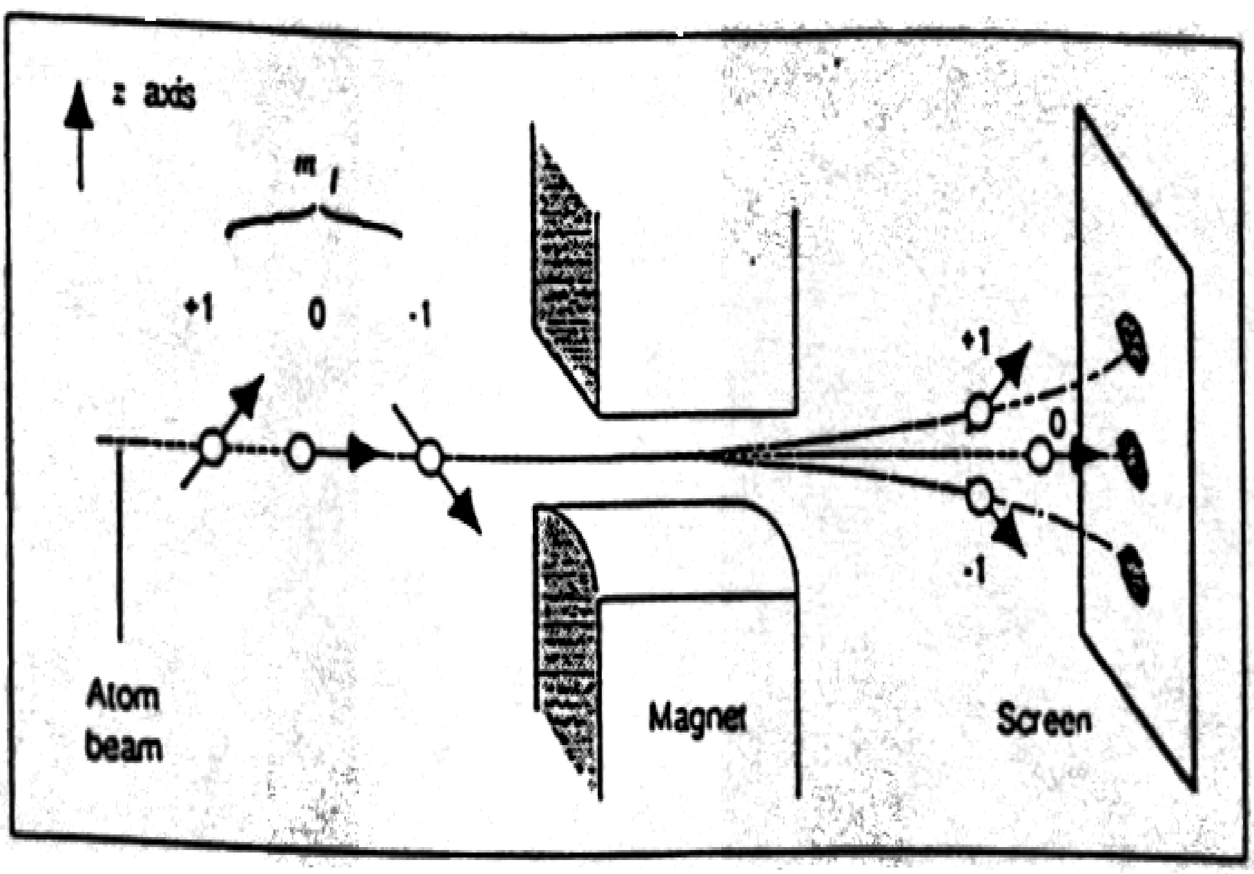

z axis
m l
+1
0
-1
+1
0
-1
Atom beam
Magnet
Screen

Figure 12. Sketch of the Stern-Gerlach experimental apparatus. The result expected for atoms in an $l = 1$ state (three components) is shown. From Weinert (1995).

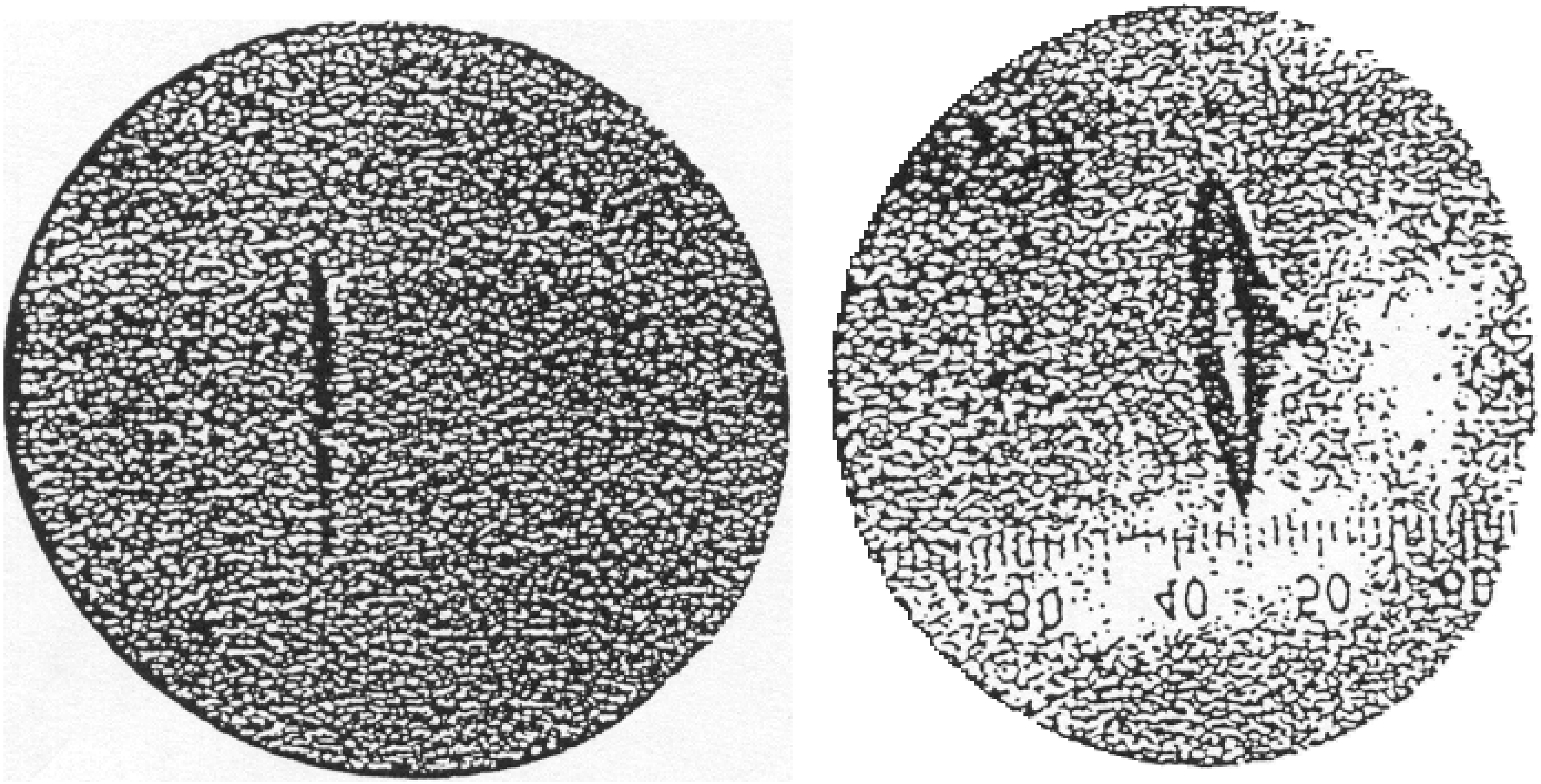


Figure 13. The experimental result of the Stern-Gerlach experiment. The beam has split into two components. From Gerlach and Stern (1922a).

Stanford Encyclopedia of Philosophy
Supplement to Experiment in Physics

Appendix 6: Sometimes Refutation Doesn't Work: The Double-Scattering of Electrons

In 1929, Mott (1929, and later 1931, 1932) calculated, on the basis of Dirac's theory of the electron, that there would be a forward-backward asymmetry of approximately 10% in the double scattering of electrons from heavy nuclei. Mott clearly specified the conditions that would have to be satisfied in order to observe this effect. One had to double scatter relativistic electrons at large angles (90º) from heavy nuclei (most calculations assumed a nuclear charge Z approximately 80). The first scatter would polarize the electrons and the second scatter would analyze the produced polarization, giving rise to an asymmetry.

The earliest experiment that discussed Mott's calculation was performed by Chase (1929). He observed a 4% asymmetry in the double scattering of electrons but attributed it to a difference in the path that the electrons followed. His subsequent experiment (Chase 1930) reported a 1.5% effect, and this time did attribute it to Mott scattering. Most experiments during the early 1930s, showed no polarization effects, although some of them did not satisfy the conditions for Mott scattering (For details see Franklin (1986, Ch. 2)). The sole positive results were provided by experiments done by Rupp (1929; 1930a; 1930b; 1931; 1932a; 1932b; 1932c). Rupp's 1932 experiment first scattered electrons at 90º from a gold foil, followed by a 90º scatter from a gold wire. He found a 3-4% asymmetry at an electron energy of 130 keV and an asymmetry of 9-10% at 250 keV. These results , although positive, were in quantitative disagreement with Mott's prediction of 15.5% at 127 keV and 14% at 204 keV (Mott 1931). Dymond (1931) also reported a positive result, but one that was five times smaller than the theoretical prediction.

Mott and the rest of the electron-scattering community were quite aware of both the confused nature of the experimental results, and of the apparent discrepancy between experiment and theory. Langstroth (1932) reviewed the situation and commented on the difficulty of experiment-theory comparison when one deals with real, as opposed to ideal, experiments. "In view of the fact that practical conditions may be immensely more complicated than those of Mott's theory, it is not surprising that it does not furnish a guide, even in a qualitative way, to all of the above experiments. This may be due to (a) the fact that a large proportion of the beam scattered from a thick target consists of electrons which have undergone more than one collision, (b) the insufficiency of the theoretical model, (c) the inclusion of extraneous effects in the experimental results" (pp. 566-67).

The situation became even more confused when Dymond(1932) published a detailed account of his experiment, which restated his positive, but discrepant, result. Adding to the confusion was the fact Dymond's experiment seemed to satisfy all the conditions for Mott scattering. Rupp (1934) continued his

work, this time using thallium vapor rather than gold targets, and again found a positive result. G.P. Thomson (1933), on the other hand, found no effect. At approximately the same time Sauter (1933) redid Mott's calculations and obtained identical results. He also considered whether or not screening by atomic electrons could cancel the predicted effect and found that it could not. If things weren't difficult enough, they got worse when Dymond (1934) published a full repudiation of his earlier results. He had found a considerable and variable experimental asymmetry in his apparatus, and concluded that he had not, in fact, observed any polarization effect. Dymond also considered possible reasons for the theory-experiment discrepancy including inelastic, stray, and plural scattering, and nuclear screening and rejected them all. He concluded, "We are driven to the conclusion that the theoretical results are wrong. There is no reason to believe that the work of Mott is incorrect;... It seems not improbable, therefore, that the divergence of theory from experiment has a more deep-seated cause, and that the Dirac wave equation needs modification in order to account successfully for the absence of polarization" (Dymond 1932, p. 666).

G.P. Thomson (1934) also published a comprehensive review of the field. He reported no effects of the type found by Rupp and he found a forward-backward ratio of (0.996 ± 0.01) in comparison to Mott's prediction of 1.15.Thomson also concluded that there was a serious discrepancy between theory and experiment.

Faced with this apparent theory-experiment discrepancy, theorists sought either to modify Dirac's theory or to propose a new theory, and thus accommodate the experimental results. Hellmann (1935), Halpern and Schwinger (1935), and Winter (1936) offered modifications of the Coulomb potential, each of which had the effect that it "annihilates the polarization effect completely." Although each of the theoretical calculations predicted null results from double scattering experiments, they were not regarded as solving the problem. One might speculate that this was because these modifications had no physical or theoretical underpinning. They seemed invented solely for the purpose of explaining the experimental results.

Experimental work also continued. The situation became even more confused when Rupp (1935) withdrew several of his results on electron scattering This eliminated the most positive results supporting Mott's theory.\1/ In 1937 Richter published what he regarded as the definitive experiment on the double scattering of electrons. He claimed to have satisfied the conditions of Mott's calculation exactly and had found no effect. He concluded that "Despite all the favorable conditions of the experiment, however, no sign of the Mott effect could be observed. *With this experimental finding, Mott's theory of the double scattering of electrons from the atomic nucleus can no longer be maintained.* It cannot be decided here how much Dirac's theory of electron spin, which is at the basis of Mott's theory, and its other applications are implicated through the denial of Mott's theory" (Richter 1937, p. 554). The discrepancy was further confirmed by the theoretical work of Rose and Bethe (1939). They examined various ways of trying to eliminate the discrepancy and concluded that "the discrepancy between theory and experiment remains -- perhaps more glaring than before" (p. 278).

Thus, at the end of 1939 there was a clear discrepancy between Dirac theory, as used by Mott, and the experimental results on the double scattering of electrons. Yet the theory was not regarded as refuted.

Why was this? The reason is that, at the time, Dirac theory, and only Dirac theory, predicted the existence of the positron (a positive electron). This particle had been discovered in 1932 and had provided very strong support for Dirac theory. In comparison with this success, the discrepancy in electron scattering, along with another small discrepancy in the spectrum of hydrogen, just did not have sufficient evidential weight. The unique, and confirmed, prediction of the positron outweighed these discrepancies. It isn't easy to refute a strongly confirmed theory. Neither is it impossible as demonstrated by the histories of both parity nonconservation and CP violation discussed earlier.

Interestingly, it was the experimental results that were wrong. In the early 1940s experimental work showed that the way in which the experiments were performed during the 1930s had precluded the possibility of observing the polarization effects predicted by Mott. In order to avoid problems with multiple scattering the experimenters had scattered the electrons from the front surface of the targets. Unfortunately this made the effects of plural scattering, a few large scatters rather than just one as required by Mott, very large. The symmetric plural scattering swamped the predicted polarization effect. When the experimental apparatuses were changed to eliminate this problem the discrepancy disappeared.\2/ Mott's theory was then supported by the experimental evidence.

We have seen here a classic case of the Duhem-Quine problem and how the physics community attempted to solve it. There was a clear discrepancy between the experimental results and the predictions of a well-confirmed theory. The experiments were redone to check the results, with careful attention to the experimental conditions required by the theory. Theorists checked on whether or not other effects might mask the predicted polarization effect. Other theorists offered competing explanations. Ultimately a solution was found.

Does the fact that Dirac theory was not regarded as refuted even though experiment clearly disagreed with its predictions mean that physicists disregard negative results whenever it suits their purposes? Do physicists really tune in on existing community commitments, as some social constructivists would have it, and overlook negative evidence? The answer is no. There is no indication in this episode that the negative evidence was disregarded. The physics community examined the theory in the light of all the available experimental evidence, weighed its importance, and then made a decision. I note that even though Dirac theory remained relatively unscathed, both experimental and theoretical work continued until the problem was solved. The discrepancy was not hidden from view, nor was it ignored.

Allan Franklin
Allan.Franklin@Colorado.edu

Return to Experiment in Physics

First published: October 5, 1998

Content last modified: October 5, 1998

Stanford Encyclopedia of Philosophy
Supplement to Experiment in Physics

Appendix 7: Evidence for a New Entity: J.J. Thomson and the Electron

In discussing the existence of electrons Ian Hacking has written, "So far as I'm concerned, if you can spray them then they are real" (Hacking 1983, p. 23). He went on to elaborate this view. "We are completely convinced of the reality of electrons when we set out to build - and often enough succeed in building - new kinds of device that use various well-understood causal properties of electrons to interfere in other more hypothetical parts of nature" (p. 265).

Hacking worried that the simple manipulation of the first quotation, the changing of the charge on an oil drop or on a superconducting niobium sphere, which involves only the charge of the electron, was insufficient grounds for belief in electrons. His second illustration, which he believed more convincing because it involved several properties of the electron, was that of Peggy II, a source of polarized electrons built at the Stanford Linear Accelerator Center in the late 1970s. Peggy II provided polarized electrons for an experiment that scattered electrons off deuterium to investigate the weak neutral current. Although I agree with Hacking that manipulability can often provide us with grounds for belief in a theoretical entity,[\1/] his illustration comes far too late. Physicists were manipulating the electron in Hacking's sense in the early twentieth century.[\2/] They believed in the existence of electrons well before Peggy II, and I will argue that they had good reasons for that belief.[\3/]

The position I adopt is one that might reasonably be called "conjectural" realism. It is conjectural because, despite having good reasons for belief in the existence of an entity or in the truth of a scientific law, we might be wrong. At one time scientists had good reason to believe in phlogiston and caloric, substances we now have good reason to believe don't exist. My position includes both Sellars' view that "to have good reason for holding a theory is *ipso facto* to have good reason for holding that the entities postulated by the theory exist" (Sellars 1962, p. 97), and the "entity realism" proposed by Cartwright (1983) and by Hacking (1983). Both Hacking, as noted above, and Cartwright emphasize the manipulability of an entity as a criterion for belief in its existence. Cartwright also stresses causal reasoning as part of her belief in entities. In her discussion of the operation of a cloud chamber she states, "...if there are no electrons in the cloud chamber, I do not know why the tracks are there" (Cartwright, 1983, p.99). In other words, if such entities don't exist then we have no plausible causal story to tell. Both Hacking and Cartwright grant existence to entities such as electrons, but do not grant "real" status to either laws or theories, which may postulate or apply to such entities.

In contrast to both Cartwright and Hacking, I suggest that we can also have good reasons for belief in the laws and theories governing the behavior of the entities, and that several of their illustrations implicitly

involve such laws.[4] I have argued elsewhere for belief in the reality of scientific laws (Franklin 1996). In this section I shall concentrate on the reality and existence of entities, in particular, the electron. I agree with both Hacking and Cartwright that we can go beyond Sellars and have good reasons for belief in entities even without laws. Hacking and Cartwright emphasize experimenting *with* entities. I will argue that experimenting *on* entities and measuring their properties can also provide grounds for belief in their existence.

In this section I will discuss the grounds for belief in the existence of the electron by examining J.J. Thomson's experiments on cathode rays. His 1897 experiment on cathode rays is generally regarded as the "discovery" of the electron.

The purpose of J.J. Thomson's experiments was clearly stated in the introduction to his 1897 paper.

The experiments discussed in this paper were undertaken in the hope of gaining some information as to the nature of Cathode Rays. The most diverse opinions are held as to these rays; according to the almost unanimous opinion of German physicists they are due to some process in the aether to which -- inasmuch as in a uniform magnetic field their course is circular and not rectilinear -- no phenomenon hitherto observed is analogous: another view of these rays is that, so far from being wholly aetherial, they are in fact wholly material, and that they mark the paths of particles of matter charged with negative electricity (Thomson 1897, p. 293).

Thomson's first order of business was to show that the cathode rays carried negative charge. This had presumably been shown previously by Perrin. Perrin placed two coaxial metal cylinders, insulated from one another, in front of a plane cathode. The cylinders each had a small hole through which the cathode rays could pass onto the inner cylinder. The outer cylinder was grounded. When cathode rays passed into the inner cylinder an electroscope attached to it showed the presence of a negative electrical charge. When the cathode rays were magnetically deflected so that they did not pass through the holes, no charge was detected. "Now the supporters of the aetherial theory do not deny that electrified particles are shot off from the cathode; they deny, however, that these charged particles have any more to do with the cathode rays than a rifle-ball has with the flash when a rifle is fired" (Thomson 1897, p. 294).

Thomson repeated the experiment, but in a form that was not open to that objection. The apparatus is shown in Figure 14. The two coaxial cylinders with holes are shown. The outer cylinder was grounded and the inner one attached to an electrometer to detect any charge. The cathode rays from A pass into the bulb, but would not enter the holes in the cylinders unless deflected by a magnetic field.

When the cathode rays (whose path was traced by the phosphorescence on the glass) did not fall on the slit, the electrical charge sent to the electrometer when the induction coil producing the rays was set in action was small and irregular; when, however, the rays were bent by a magnet so as to fall on the slit there was a large charge of negative electricity sent to the electrometer.... If the rays were so much bent by the magnet that they overshot the slits in the cylinder, the charge passing into the cylinder fell again to a very small fraction of its value when the aim was true. *Thus this experiment shows that however we*

twist and deflect the cathode rays by magnetic forces, the negative electrification follows the same path as the rays, and that this negative electrification is indissolubly connected with the cathode rays (Thomson 1897, p. 294-295, emphasis added).

This experiment also demonstrated that cathode rays were deflected by a magnetic field in exactly the way one would expect if they were negatively charged material particles.\5/

There was, however, a problem for the view that cathode rays were negatively charged particles. Several experiments, in particular those of Hertz, had failed to observe the deflection of cathode rays by an electrostatic field. Thomson proceeded to answer this objection. His apparatus is shown in Figure 15. Cathode rays from C pass through a slit in the anode A, and through another slit at B. They then passed between plates D and E and produced a narrow well-defined phosphorescent patch at the end of the tube, which also had a scale attached to measure any deflection. When Hertz had performed the experiment he had found no deflection when a potential difference was applied across D and E. He concluded that the electrostatic properties of the cathode ray are either *nil* or very feeble. Thomson admitted that when he first performed the experiment he also saw no effect. "on repeating this experiment [that of Hertz] I at first got the same result [no deflection], but subsequent experiments showed that the absence of deflexion is due to the conductivity conferred on the rarefied gas by the cathode rays.\6/ On measuring this conductivity it was found that it diminished very rapidly as the exhaustion increased; it seemed that on trying Hertz's experiment at very high exhaustion there might be a chance of detecting the deflexion of the cathode rays by an electrostatic force" (Thomson 1897, p. 296). Thomson did perform the experiment at lower pressure [higher exhaustion] and observed the deflection.\7/

Thomson concluded:

> As the cathode rays carry a charge of negative electricity, are deflected by an electrostatic force as if they were negatively electrified, and are acted on by a magnetic force in just the way in which this force would act on a negatively electrified body moving along the path of these rays, I can see no escape from the conclusion that they are charges of negative electricity carried by particles of matter. (Thomson 1897, p. 302)\8/

Having established that cathode rays were negatively charged material particles, Thomson went on to discuss what the particles were. "What are these particles? are they atoms, or molecules, or matter in a still finer state of subdivision" (p. 302). To investigate this question Thomson made measurements on the charge to mass ratio of cathode rays. Thomson's method used both the electrostatic and magnetic deflection of the cathode rays.\9/ The apparatus is shown in . It also included a magnetic field that could be created perpendicular to both the electric field and the trajectory of the cathode rays.

Let us consider a beam of particles of mass m charge e, and velocity v. Suppose the beam passes through an electric field F in the region between plates D and E, which has a length L. The time for a particle to pass through this region $t = L/v$. The electric force on the particle is Fe and its acceleration $a = Fe/m$. The deflection d at the end of the region is given by

$d = \frac{1}{2} at^2 = \frac{1}{2}(eF/m)L^2/v^2$

Now consider a situation in which the beam of cathode rays simultaneously pass through both *F* and a magnetic field *B* in the same region. Thomson adjusted *B* so that the beam was undeflected. thus the magnetic force was equal to the electrostatic force.

$evB = eF$ or $v = F/B$.

This determined the velocity of the beam. Thus, $e/m = 2dF/B^2L^2$

Each of the quantities in the above expression was measured so the *e/m* or *m/e* could be determined.

Using this method Thomson found a value of *m/e* of $(1.29 \pm 0.17) \times 10^{-7}$. This value was independent of both the gas in the tube and of the metal used in the cathode, suggesting that the particles were constituents of the atoms of all substances. It was also far smaller, by a factor of 1000, than the smallest value previously obtained, 10^{-4}, that of the hydrogen ion in electrolysis.

Thomson remarked that this might be due to the smallness of *m* or to the largeness of *e*. He argued that *m* was small citing Lenard's work on the range of cathode rays in air. The range, which is related to the mean free path for collisions, and which depends on the size of the object, was 0.5 cm. The mean free path for molecules in air was approximately 10^{-5} cm. If the cathode ray traveled so much farther than a molecule before colliding with an air molecule, Thomson argued that it must be much smaller than a molecule.[\10/]

Thomson had shown that cathode rays behave as one would expect negatively charged material particles to behave. They deposited negative charge on an electrometer, and were deflected by both electric and magnetic fields in the appropriate direction for a negative charge. In addition the value for the mass to charge ratio was far smaller than the smallest value previously obtained, that of the hydrogen ion. If the charge were the same as that on the hydrogen ion, the mass would be far less. In addition, the cathode rays traveled farther in air than did molecules, also implying that they were smaller than an atom or molecule. Thomson concluded that these negatively charged particles were constituents of atoms. In other words, Thomson's experiments had given us good reasons to believe in the existence of electrons.

Allan Franklin
Allan.Franklin@Colorado.edu

Return to Experiment in Physics

First published: October 5, 1998
Content last modified: October 5, 1998

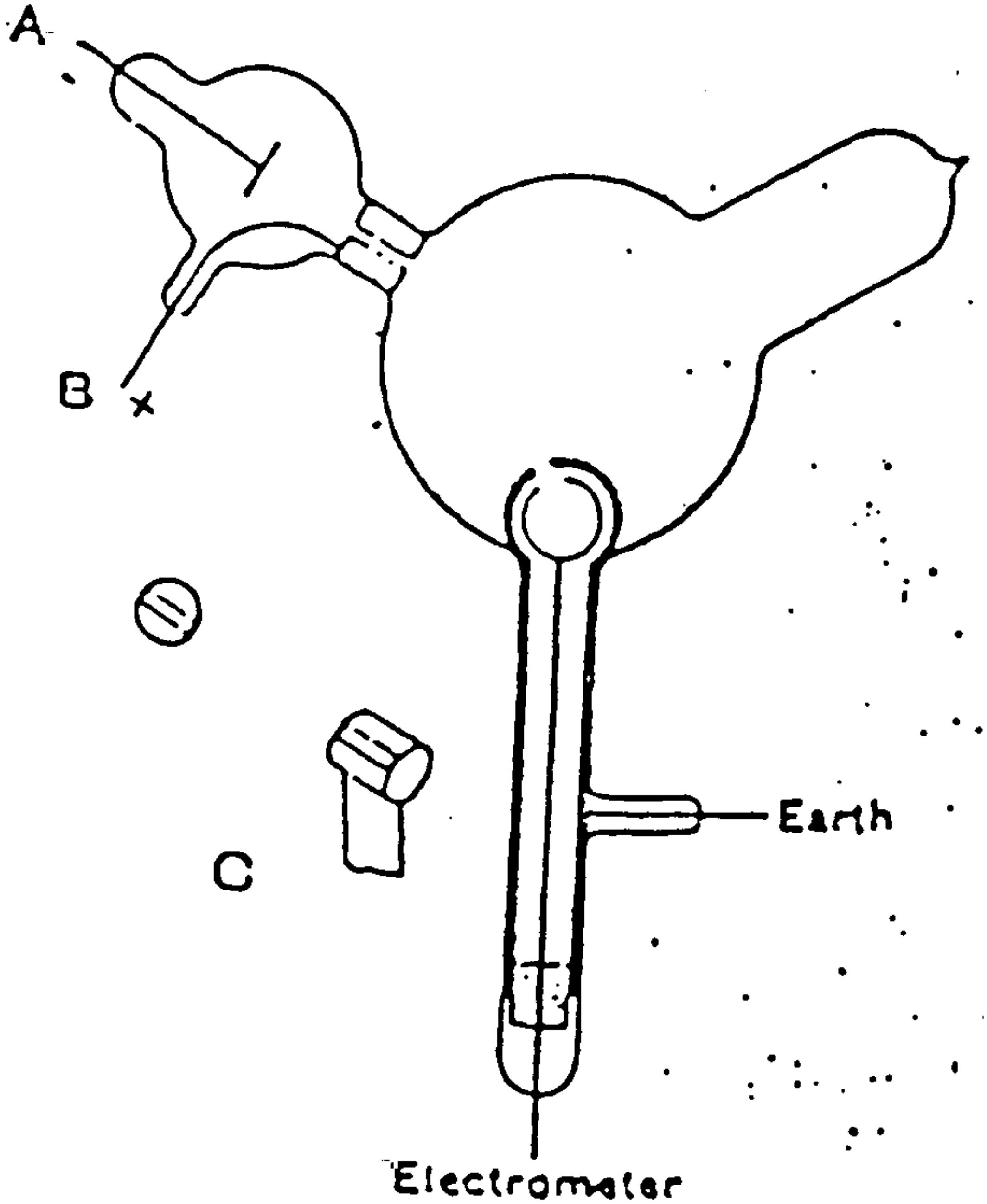


Figure 14. Thomson's apparatus for demonstrating that cathode rays have negative charge. The slits in the cylinders are shown. From Thomson (1897).

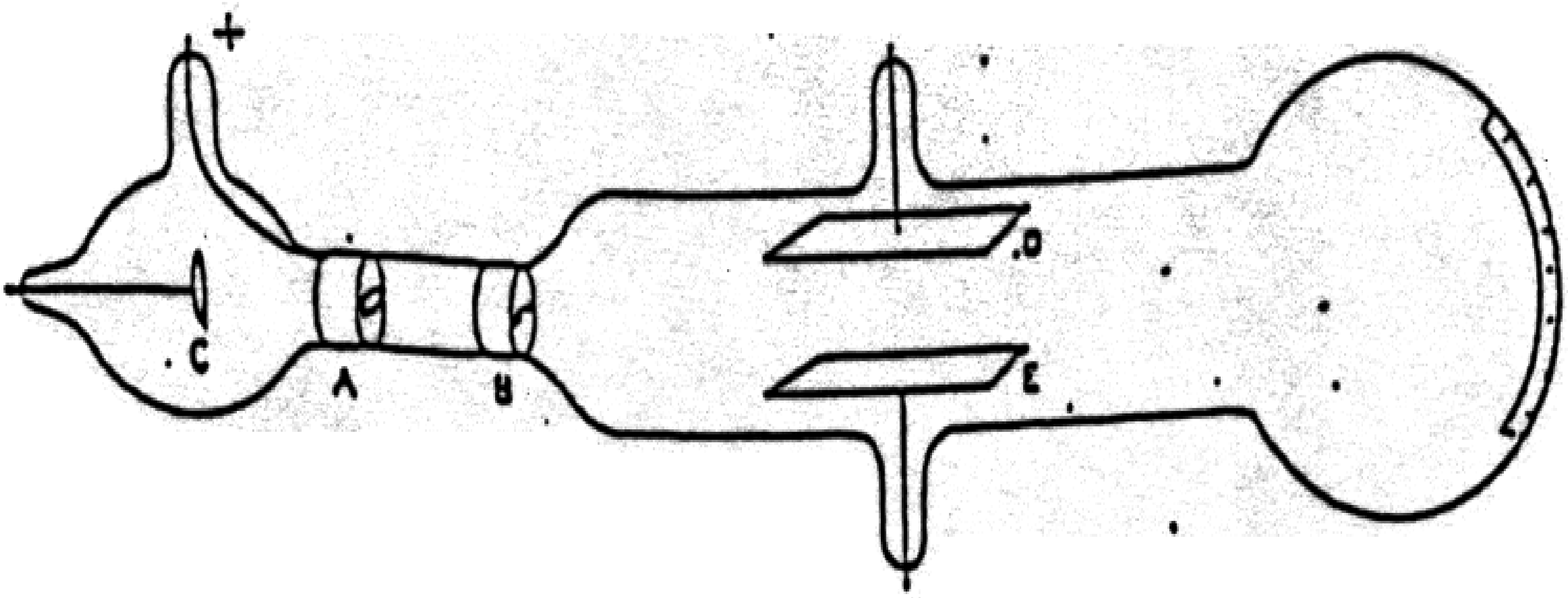


Figure 15. Thomson's apparatus for demonstrating that cathode rays are deflected by an electric field. It was also used to measure *m/e*. From Thomson (1897).

Stanford Encyclopedia of Philosophy
Supplement to Experiment in Physics

Appendix 8: The Articulation of Theory: Weak Interactions

Radioactivity, the spontaneous decay of a substance, produces alpha particles (positively charged helium nuclei), or beta particles (electrons), or gamma rays (high energy electromagnetic radiation). It was discovered in 1896 by Henri Becquerel. Experimental work on the energy of the electrons emitted in β decay began in the early twentieth century, and the observed continuous energy spectrum posed a problem. If β decay were a two-body decay (for example, neutron $\rightarrow$ proton + electron) then applying the laws of conservation of energy and of conservation of momentum requires that the energy of the electron have a unique value, not a continuous spectrum.\1/ Thus, the observed continuous energy spectrum cast doubt on both of these conservation laws. Physicists speculated that perhaps the electrons lost energy in escaping the substance, with different electrons losing different amounts of energy, thus accounting for the energy spectrum. Careful experiments showed that this was not the case so the problem remained. In the early 1930s Pauli suggested that a low-mass neutral particle, named by Fermi as the neutrino, was also emitted in β decay.\2/ This solved the problem of the continuous energy spectrum because in a three-body decay (neutron proton + electron + neutrino) the energy of the electron was no longer required to be unique. The electron could have a continuous energy spectrum and the conservation laws were saved.\3/

In 1934 Fermi proposed a new theory of β decay that incorporated this new particle (Fermi 1934). He added a perturbation energy due to the decay interaction to the Hamiltonian describing the nuclear system. Pauli (1933) had previously shown that the perturbation could have only five different forms if the Hamiltonian is to be relativistically invariant. These are S, the scalar interaction; P, pseudoscalar; V, vector; A, axial vector; and T, tensor. Fermi knew this but chose, in analogy with electromagnetic theory, to use only the vector interaction. His theory initially received support from the work of Sargent (1932; 1933) and others. There remained, however, the question of whether or not the other forms of the interaction also entered into the Hamiltonian.\4/ In this episode we shall see how experiment helped to determine the mathematical form of the weak interaction.

Gamow and Teller (1936) soon proposed a modification of Fermi's vector theory. Fermi's theory had originally required a selection rule, the change in J = 0, where J is the angular momentum of the nucleus, and did not include the effects of nuclear spin. Gamow and Teller included nuclear spin and obtained selection rules, change in J = 0, ±1 for allowed transitions, with no 0--0 transitions allowed. The Gamow-Teller modification required either a tensor or an axial vector form of the interaction. Their theory helped to solve some of the difficulties that arose in assigning nuclear spins using only the Fermi selection rule. At then end of the 1930s there was support for Fermi's theory with some preference for the Gamow-

Teller selections rules and the tensor interaction.

The work of Fierz (1937) helped to restrict the allowable forms of the interaction. He showed that if both S and V interactions were present in the allowed -decay interaction, or both A and T, then there would be an interference term of the form 1 + a/W in the allowed beta-decay spectrum, where W is the electron energy. This term vanished if the admixtures were not present. The failure to observe these interference terms showed that the decay interaction did not contain both S and V, or both A and T.

The presence of either the T or A form of the interaction in at least part of the beta-decay interaction was shown by Mayer, Moszkowski, and Nordheim (1951). They found twenty five decays for which the change in J was 0, ±1, with no parity change. These decays could only occur if the A or T forms were present. Their conclusion depended on the correct assignment of nuclear spins which, although reliable, still retained some uncertainty. Further evidence, which did not depend on knowledge of the nuclear spins, came from an examination of the spectra of unique forbidden transitions.[\5/] These were n-times forbidden transitions in which the change in nuclear spin was n + 1. These transitions require the presence of either A or T. In addition, only a single form of the interaction makes any appreciable contribution to the decay. This allows the prediction of the shape of the spectrum for such transitions. Konopinski and Uhlenbeck (1941) showed that for an n-times forbidden transition the spectrum would be that of an allowed transition multiplied by an energy dependent term $a_n(W)$. For a first-forbidden transition $a_1 = C[(W^2 - m^2c^4) + (W_o - W)^2]$. The spectrum for ^{91}Y measured by Langer and Price (1949) (Figure 16) shows the clear presence of either the A or T forms of the interaction. The spectrum requires the energy-dependent correction.

Evidence in favor of the presence of either the S or V forms of the interaction was provided by Sherr, Muether, and White (1949)and by Sherr and Gerhart (1952). They observed the decay of ^{14}O to an excited state of ^{14}N, $^{14}N^*$. They argued that both ^{14}O and $^{14}N^*$ had spin 0. This required the presence of either S or V because the decay was forbidden by A and T. (Recall that the Gamow-Teller selection rules specified no 0 to 0 transitions).

Further progress in isolating the particular forms of the interaction was made by examining the spectra of once-forbidden transitions. Here too, interference effects, similar to those predicted by Fierz, were also expected. A. Smith (1951)and Pursey (1951) found that the spectrum for these transitions would contain energy dependent terms of the form $G_V G_T/W$, $G_A G_P/W$, and $G_S G_A/W$, where the G's are the coupling constants for the various interactions, and W is the electron energy. The linear spectrum found for ^{147}Pm demonstrated the absence of these terms (Langer, Motz and Price 1950).

Let us summarize the situation. There were five allowable forms of the decay interaction; S, T, A, V, P. The failure to observe Fierz interference showed that the interaction could not contain both S and V or both A and T. Experiments showing the presence of Gamow-Teller selections rules and on unique forbidden transitions had shown that either A or T must be present. The decay of ^{14}O to $^{14}N^*$ had demonstrated that either S or V must also be present. This restricted the forms of the interaction to STP,

SAP, VTP, or VAP or doublets taken from these combinations. The absence of interference terms in the once-forbidden spectra eliminated the VT, SA, and AP combinations. VP was eliminated because it did not allow Gamow-Teller transitions. This left only the STP triplet or the VA doublet as the possible interactions.

The spectrum of RaE provided the decisive evidence. Petschek and Marshak (1952) analyzed the spectrum of RaE and found that the only interaction that would give a good fit to the spectrum was a combination of T and P. This was, in fact, the only evidence favoring the presence of the P interaction. This led Konopinski and Langer (1953), in their 1953 review article on β decay to conclude that, "As we shall interpret the evidence here, the correct law must be what is known as an STP combination (1953, p. 261)."

Unfortunately, the evidence from the RaE spectrum had led the physics community astray. Petschek and Marshak had noted that their conclusion was quite sensitive to assumptions made in their calculation. "Thus, an error in the finite radius correction of approximately 0.1 percent leads to an error of up to 25% in $C_{1(T+P)}$ [the theoretical correction term]." Further theoretical analysis cast doubt on their assumptions, but all of this became moot when K. Smith[\6/] measured the spin of RaE and found it to be one, incompatible with the Petschek-Marshak analysis.

The demise of the RaE evidence removed the necessity of including the P interaction in the theory of β decay, and left the decision between the STP and VA combinations unresolved. The dilemma was resolved by evidence provided by angular-correlation experiments, particularly that from the experiment on ^{6}He by Rustad and Ruby (1953; 1955)

(a) Angular Correlation Experiments. Angular correlation experiments are those in which both the decay electron and the recoil nucleus from β decay are detected in coincidence. The experiments measured the distribution in angle between the electron and the recoil nucleus for a fixed range of electron energy, or measured the energy spectrum of either the electron or the nucleus at a fixed angle between them. These quantities are quite sensitive to the form of the decay interaction and became decisive pieces of evidence in the search for the form of the decay interaction. Hamilton (1947) calculated the form of the angular distribution expected for both allowed and forbidden decays, assuming only one type of interaction (S, V, T, A, P) was present. He found, for allowed transitions, that the angular distributions for the specific forms of the interaction would be different. A more general treatment was given by de Groot and Tolhoek (1950). They found that the general form of the angular distribution for allowed decays depended on the combination of the particular forms of the interactions in the decay Hamiltonian. For single forms their results agreed with those of Hamilton.

The most important of the experiments performed at this time was the measurement of the angular correlation in the decay of ^{6}He. This decay was a pure Gamow-Teller transition and thus was sensitive to the amounts of A and T present in the decay interaction. The decisive experiment was that of Rustad and Ruby (1953; 1955). This experiment was regarded as establishing that the Gamow-Teller part of the interaction was predominantly tensor. This was the conclusion reached in several review papers on the

nature of β decay. (Ridley 1954; Kofoed-Hansen 1955; Wu 1955). The experimental apparatus is shown in Figure 17. The definition of the decay volume was extremely important. In order to measure the angular correlation one must know the position of the decay so that one can measure the angle between the electron and the recoil nucleus. The decay volume for the helium gas in this experiment was defined by a 180 microgram/cm^2 aluminum hemisphere and the pumping diaphragm. Rustad and Ruby (1953) presented two experimental results. The first was the coincidence rate as a function of the angle between the electron and the recoil nucleus for electrons in the energy range (2.5 - 4.0) mc^2. The second was the the energy spectrum of the decay electrons with the angle between the electron and the recoil nucleus fixed at 180^o. Both results are shown in Figure 18 along with the predicted results for A and T, respectively. The dominance of the tensor interaction is clear. This conclusion was made more emphatic in their 1955 paper which included more details of the experiment and even more data. The later results, shown in Figure 19, clearly demonstrate the superior fit of the tensor interaction.

The Rustad-Ruby result, along with several others, established that the Gamow-Teller part of the decay interaction was tensor and that the decay interaction was STP, or ST, rather than VA. We have seen clearly in this episode the fruitful interaction between experiment and theory. Theoretical predictions became more precise and were tested experimentally until the form of the weak interaction was found. Fermi's theory of β decay had been confirmed. It had also been established that the interaction was a combination of scalar, tensor, and pseudoscalar (STP).

(b) Epilogue. It would be nice to report that such a simple, satisfying story, with its happy ending was the last word. It wasn't. Work continued on angular correlation experiments and the happy agreement was soon destroyed (Franklin 1990, Chapter 3). Things became more complex with the discovery of parity nonconservation in the weak interactions, including β decay. Sudarshan and Marshak (1958) and Feynman and Gell-Mann (1958) showed that only a V-A interaction was compatible with parity nonconservation. If there was to be a single interaction describing all the weak interactions then there was a serious conflict between this work and the Rustad-Ruby result. This led Wu and Schwarzschild (1958) to reexamine and reanalyze the Rustad-Ruby experiment. They found, by calculation and by constructing a physical analogue of the gas system, that a considerable fraction of the helium gas (approximately 12%) was not in the decay volume. This changed the result for the angular correlation considerably and cast doubt on the Rustad-Ruby result.[\7/] The ^{6}He angular correlation experiment was redone, correcting the problem with the gas target, and the new result is shown in Figure 20 (Hermannsfeldt et al. 1958). It clearly favors A, the axial vector interaction. Once again, physics was both fallible and corrigible. This new result on ^{6}He combined with the discovery of parity nonconservation established that the form of the weak interaction was V-A.

Allan Franklin
Allan.Franklin@Colorado.edu

Return to Experiment in Physics

First published: October 5, 1998
Content last modified: October 5, 1998

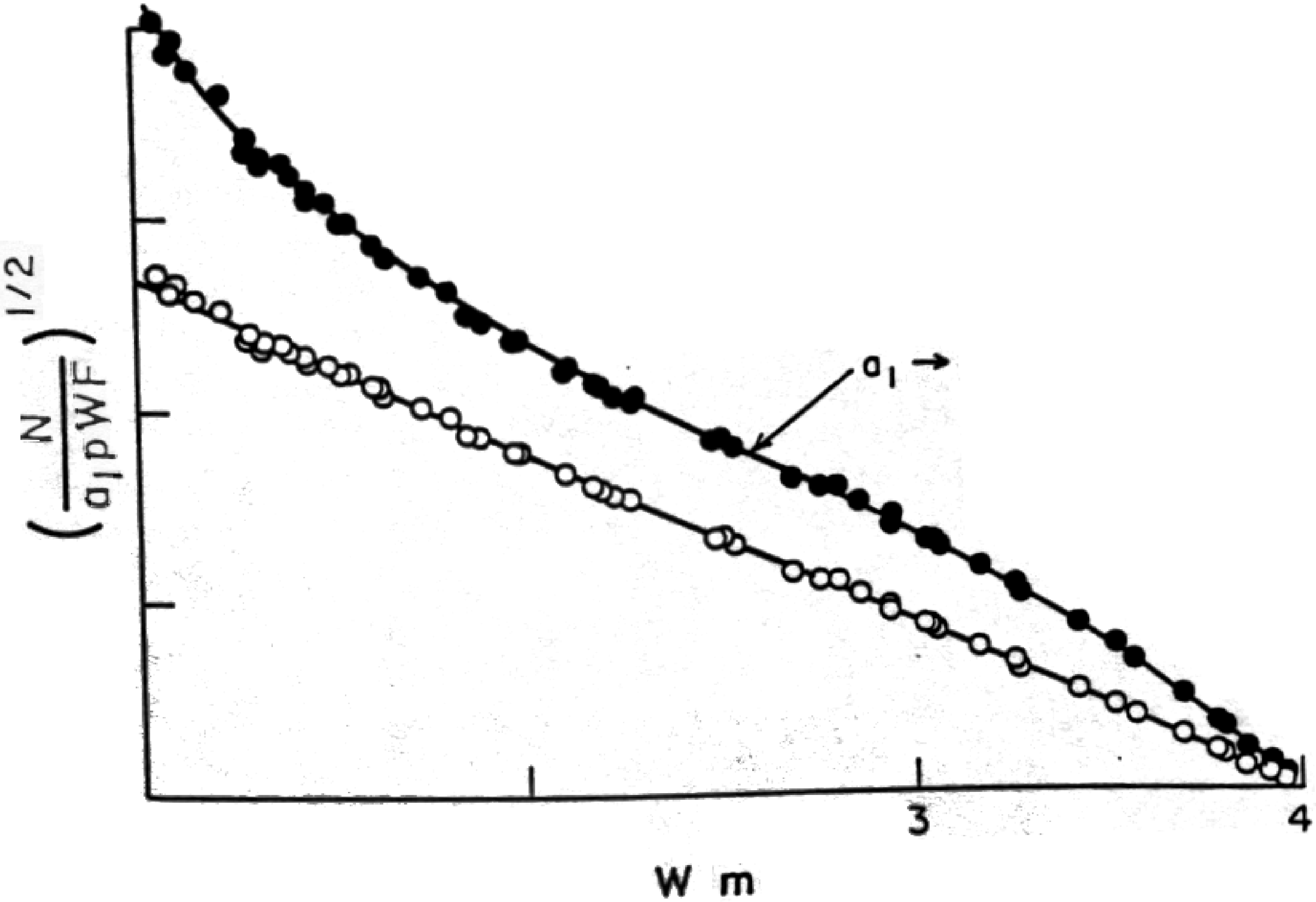


Figure 16. The unique, once-forbidden spectrum of ^{91}Y. The best theoretical fit is that which gives a straight line. The Fermi theory alone, $a_1 = 1$, does not give a straight line. The correction factor $a_1 = C[(W^2 - m_o c^2) + (W_o - W)^2]$, does give a linear plot. From Konopinski and Langer (1953)

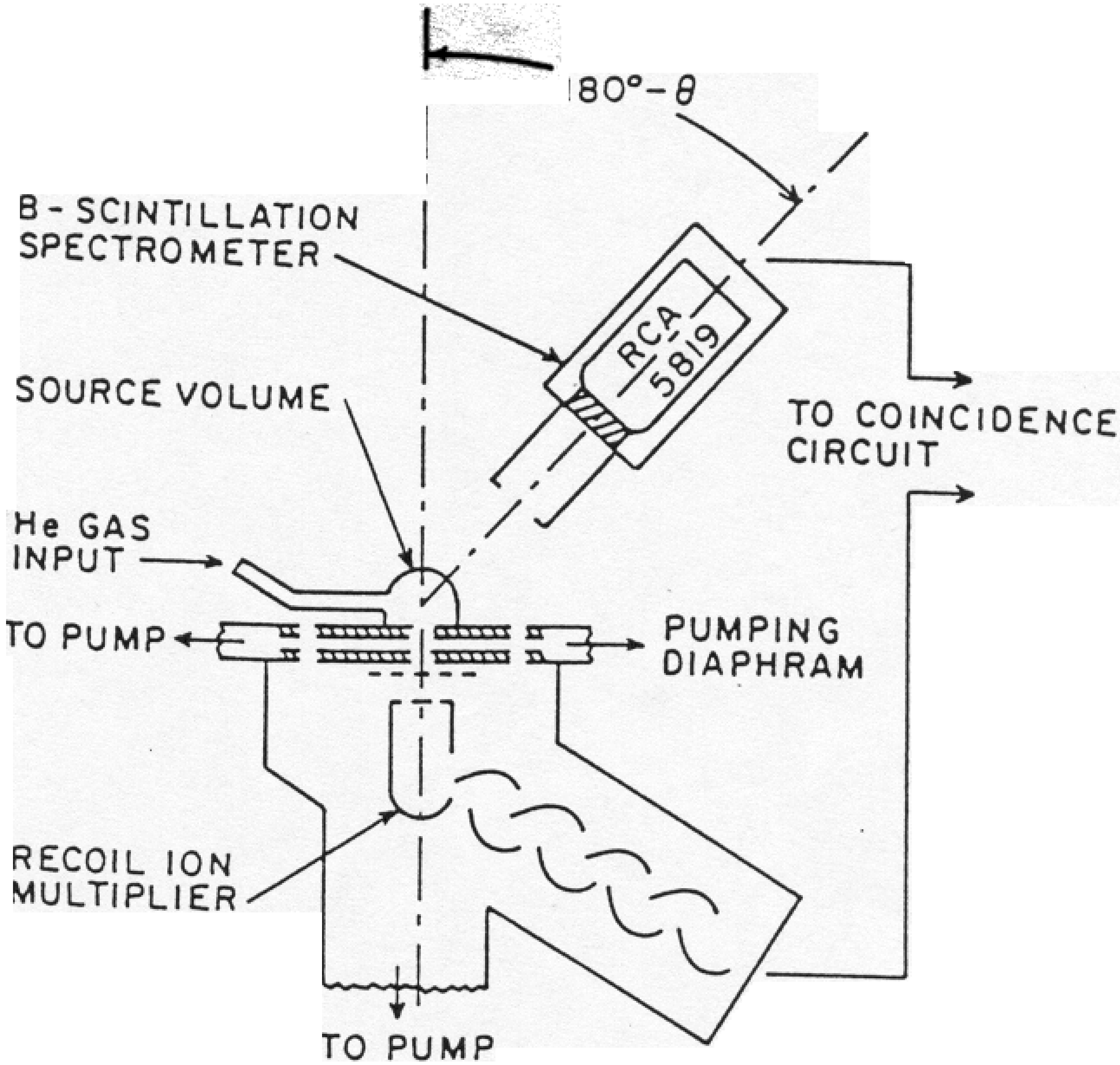


Figure 17. Schematic view of the experimental apparatus for the ^{6}He angular correlation experiment of Rustad and Ruby (1953; 1955)

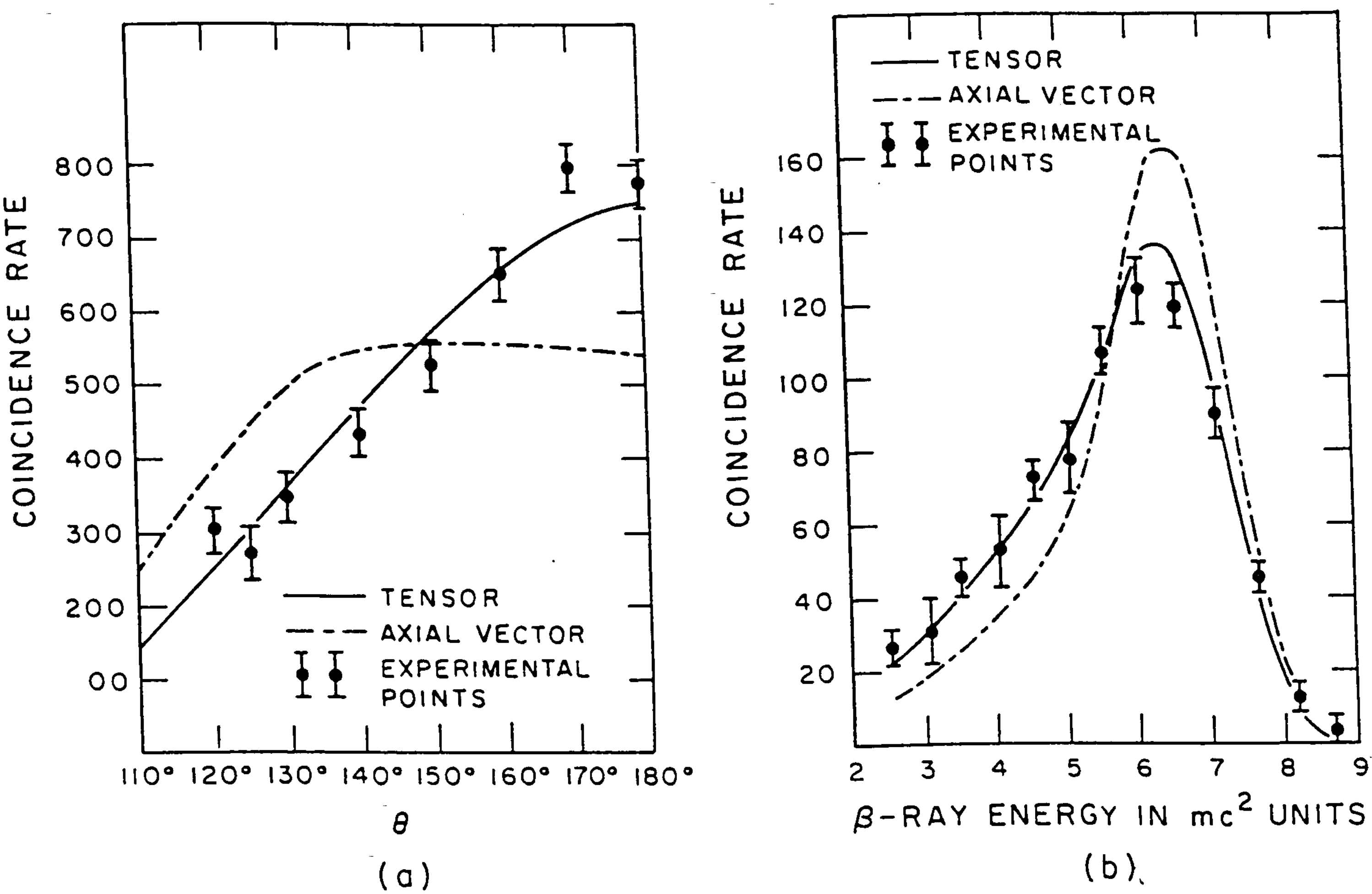


Figure 18. (a) Coincidence counting rate versus angle between the electron and the recoil nucleus, for electrons in the energy range 2.5-4.0 mc^2. (b) Coincidence counting rate versus electron energy for an angle of 180^o between the electron and the recoil nucleus. From Rustad and Ruby (1953) .

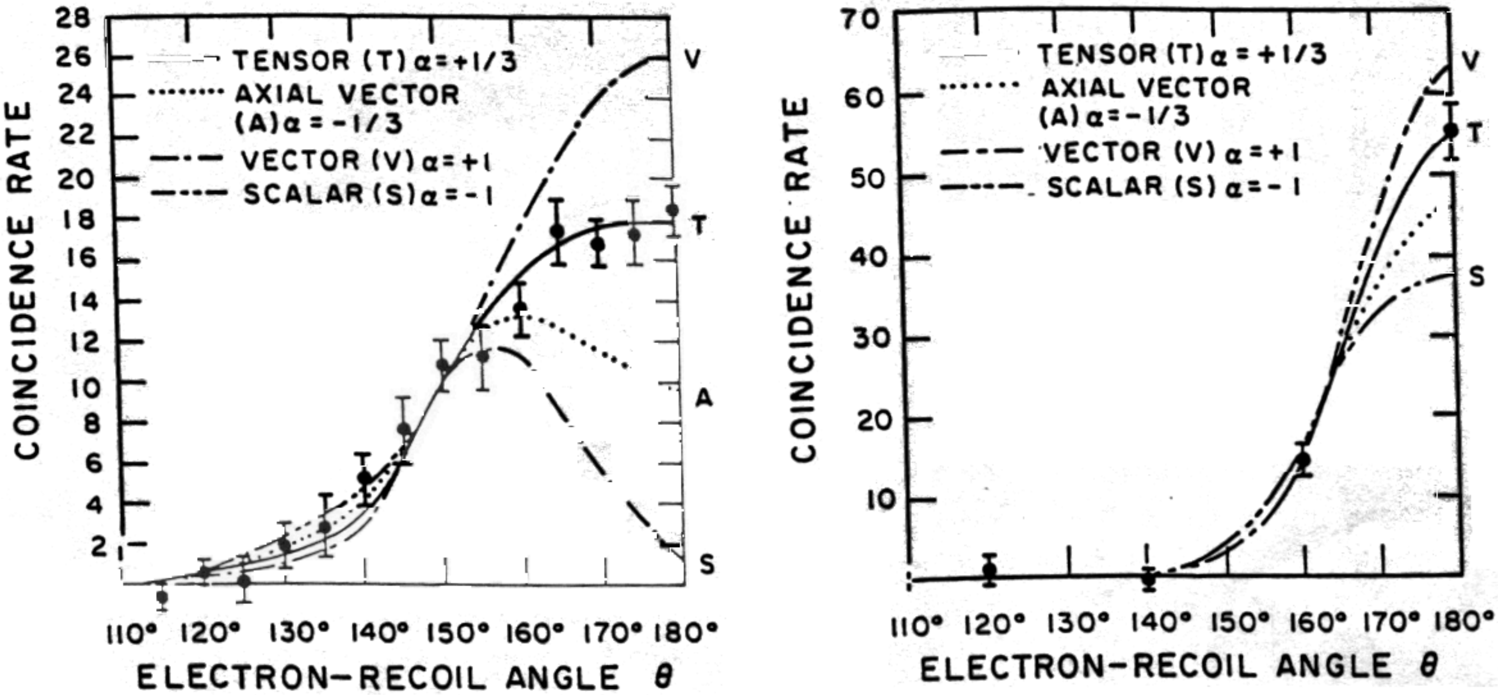


Figure 19. Coincidence counting rate versus angle between the electron and the recoil nucleus for (a) electrons in the energy range 4.5-5.5 mc^2 and (b) electrons in the energy range 5.5-7.5 mc^2. From Rustad and Ruby (1955).

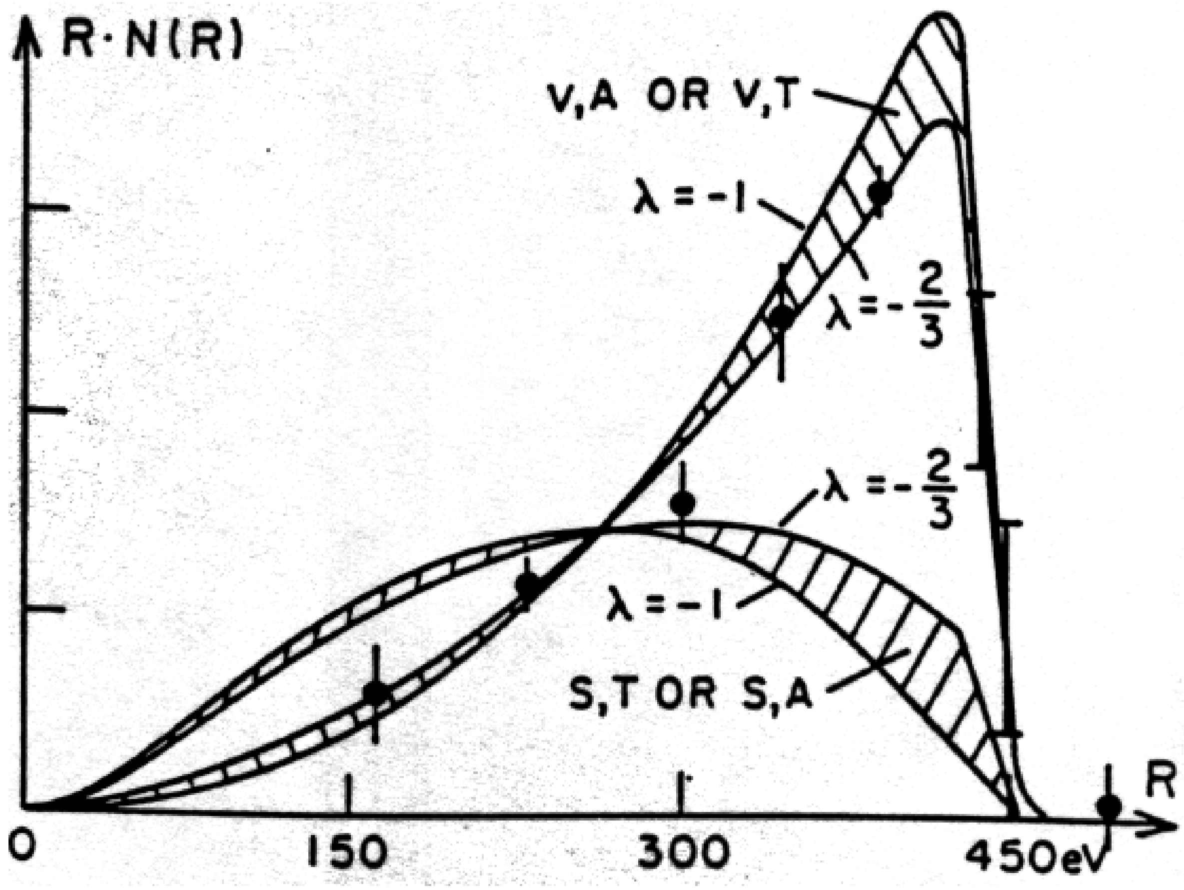


Figure 20. Energy spectrum of recoil ions from ^{35}A decay. From (Hermannsfeldt et al. 1958).

Scientific Realism

It is easier to define scientific realism than it is to identify its role as a distinctly *philosophical* doctrine. Scientific realists hold that the characteristic product of successful scientific research is knowledge of largely theory-independent phenomena and that such knowledge is possible (indeed actual) even in those cases in which the relevant phenomena are not, in any non-question-begging sense, observable. According to scientific realists, for example, if you obtain a good contemporary chemistry textbook you will have good reason to believe (because the scientists whose work the book reports had good scientific evidence for) the (approximate) truth of the claims it contains about the existence and properties of atoms, molecules, sub-atomic particles, energy levels, reaction mechanisms, etc. Moreover, you have good reason to think that such phenomena have the properties attributed to them in the textbook independently of our theoretical conceptions in chemistry. Scientific realism is thus the common sense (or common science) conception that, subject to a recognition that scientific methods are fallible and that most scientific knowledge is approximate, we are justified in accepting the most secure findings of scientists "at face value."

1. Introduction

We defined scientific realism above as the common sense (or common science) conception that, subject to a recognition that scientific methods are fallible and that most scientific knowledge is approximate, we are justified in accepting the most secure findings of scientists "at face value." What requires explanation

is why this is a philosophical position rather than just a common sense one. Consider, for example, tropical fish realism -- the doctrine that there really are tropical fish; that the little books you buy about them at pet stores tend to get it approximately right about their appearance, behavior, food and temperature requirements, etc.; and that the fish have these properties largely independently of our theories about them. That's a pretty clear doctrine, but it's so commonsensical that it doesn't seem to have any particular philosophical import. Why is the analogous doctrine about science a philosophical doctrine?

The answer is that -- setting aside skepticism about the external world -- there are no philosophical arguments against tropical fish realism, whereas important philosophical challenges have been raised against scientific realism. The dimensions of scientific realism, understood as a philosophical position, have been largely determined by the responses scientific realists have offered to these challenges. It will be conceptually useful (and approximately historically correct) to see the development of scientific realism as a response to four consecutive challenges, as follows.

1. The Empiricist Challenge: This is the challenge regarding knowledge of unobservable "theoretical" entities raised by logical empiricists and their allies and underwritten by arguments from the underdetermination of theory choice by observational data.
2. The Neo-Kantian Challenge, First Version: This is the challenge raised by Hanson (1958) and Kuhn (1970) who argue from the theory dependence of methods (and, especially, of observation) to the conclusion that a realist conception of the growth of approximate scientific knowledge cannot be sustained, given the semantic and methodological incommensurability (Kuhn's term) occasioned by revolutionary changes in science.
3. The Neo-Kantian Challenge, Second Version: This is the (somehow) related, but (somehow) less relativist conception represented by Putnam's ("internal realist") and Fine's ("natural ontological attitude") critiques of "metaphysical" versions of scientific realism.
4. The "Post-modern" Challenge: This is a challenge (to both realism and empiricism) arising from recent literary, sociological and historical studies in the emerging "science studies" tradition. It is grounded in the conception that such phenomena as science, knowledge, evidence and truth are *social constructions*, in some sense or other which implies that one should reject the idea that scientific practices achieve an approximate representational fit, of some sort or other, between the content of scientific theories and *the world* or *reality*.

We will discuss these challenges in the sections below.

2. The Empiricist Challenge: Knowledge Empiricism and the Underdetermination Argument

It is easy to characterize the basic empiricist underdetermination argument against scientific realism. Call two theories *empirically equivalent* just in case exactly the same conclusions *about observable phenomena* can be deduced from each. Let *T* be *any* theory which posits unobservable phenomena. There

will always be infinitely many theories which are empirically equivalent to *T* but which are such that each differs from *T*, and from all the rest, in what it says about *unobservable* phenomena (for formalized theories, this is an elementary theorem of mathematical logic). Evidence in favor of *T*'s conception of unobservable phenomena ("theoretical entities") would have to rule out the conceptions represented by each of those other theories. But, since *T* is empirically equivalent to each of them, they all make exactly the same predictions about the results of observations or experiments. So, no evidence could favor one of them over the others. Thus, at best, we could have evidence in favor of what all these theories have in common--their consequences about "observables"--we could confirm that they are all *empirically adequate*--but we could not have any evidence favoring *T*'s conception of unobservable theoretical entities. Since *T* was *any* theory about unobservables, knowledge of unobservable phenomena is impossible; choice between competing but empirically equivalent conceptions of theoretical entities is underdetermined by all possible observational evidence. [For an important alternative formulation of the notions of empirical adequacy and empirical equivalence, see van Fraassen 1980; see also Demopoulos 1982.]

Several points about this simple and powerful argument are important.

1. *It needs fixing up*. As it stands, the basic underdetermination argument is fatally flawed. Suppose that *T* is some ordinary middle-sized scientific theory, like, e.g., the laws of Newtonian mechanics. According to the argument as it stands if T^* is some other middle-sized theory empirically equivalent to *T*, then no evidence could favor *T* over T^*, or *vice versa*. For ordinary scientific theories this is wrong. Scientists routinely supplement theories with well established *auxiliary hypotheses* in order to obtain observational predictions from them. [In fact, no observational predictions can be deduced from Newton's laws unless they have been so supplemented, and this is true for lots of fundamental scientific theories; see Kitcher 1982 for a nice discussion.] So, even if *T* and T^* are empirically equivalent, it could still happen that they yield different observational predictions when supplemented by appropriate auxiliary hypotheses, in which case there could be observational evidence favoring one over the other.

So, it is probably best to think of the underdetermination argument as applying, not to "small" theories, but to "total sciences," large-scale conceptions of the world that might represent the total scientific conception of the world at a time. Such a conception would already contain all of the auxiliary hypotheses which were legitimate by its lights, so the problem just mentioned does not arise. In this revised form the underdetermination argument says that--whatever our best scientific conception of the world may be at any given time--we will ever have any evidence that it embodies knowledge of unobservables.

2. *It rests on (a particular interpretation of) an extremely plausible doctrine about factual knowledge*. Traditional empiricism attributed to experience or sensation two different roles: experience was the source of all of our ideas--of the raw material for thinking--and experience was the only basis we have for justifying beliefs abut matters of fact. The first of these doctrines of empiricism has fallen on hard times, but the second doctrine (called *knowledge empiricism* by Bennett 1964) enjoys widespread support. In particular, it is an epistemological doctrine to which almost all scientific realists subscribe. The logical empiricist challenge to scientific realism arises from a quite plausible interpretation of knowledge empiricism according to which what it says is that there can be no evidence which rationally distinguishes

between two empirically equivalent total sciences (call this doctrine the *evidential indistinguishability thesis*, or the *EIT*).

3. *It is part of a selectively skeptical program of anti-metaphysical "rational reconstruction."* The basic aim of the logical empiricists' project was to solve the *demarcation problem*, the problem of distinguishing science (*good*) from "metaphysics" (*bad*), by appealing to arguments like the underdetermination argument. The result was supposed to be that scientific claims are meaningful and knowable (early on, logical empiricists identified these two properties) whereas "metaphysical" claims, because they are about unobservables, are (at least) unknowable and (according to early versions of logical empiricism) meaningless.

Now almost all actual science is conducted largely in a vocabulary consisting mainly of "theoretical terms": terms apparently referring to unobservables. It was definitely not the logical empiricists' project to reject such science. They intended to be selectively skeptical: to be skeptics about "metaphysics" but not about science. So, they embarked on a project of providing "rational reconstructions" of *actual* scientific theories and methods which were designed to eliminate any apparent commitments to knowledge of unobservables while still portraying actual scientific practices as sources of knowledge (see, e.g., Carnap 1932, 1959).

In the case of scientific theories, the basic logical empiricist approaches were variations on the idea of *instrumentalism*, the view that scientific theories were predictive instruments and that the knowledge they represent is limited to what they predict about the observable properties of observables. In the case of scientific methods, strategies for rational reconstruction have not been so easy to formulate. Here's the problem. Almost all of the methods scientists actually use in conducting experimental or observational studies are *theory dependent*: they depend for their justification on knowledge reflected in previously established theories. Kuhn's (1970) discussion of "normal science" makes this point especially clearly, but all of the logical empiricists were acutely aware of it. Moreover, in sciences like physics, chemistry, molecular biology and astronomy, almost all of those methods seem *prima facie* to rest on knowledge of unobservable phenomena (just think about the presuppositions of the design of any experiment in chemistry). What the project of rational reconstruction must show is that (almost) all of these methods can be reconstructed in such a way that their application, as guides to the identification of empirically adequate theories, does not require positing knowledge of unobservables.

4. *The task of rationally reconstructing actual scientific methods has been the most significant challenge facing logical empiricism and related anti-realist approaches*. Instrumentalism and its variants provide a simple reconstruction of the content of scientific theories that pretty exactly fits the requirements of the project of rational reconstruction. The depth of the theory dependence of scientific methods, and the extent to which they seem to depend on knowledge of unobservables, has posed a deeper challenge for logical empiricists and their allies. The fate of operationalism illustrates this challenge. Operationalism was a proposal for rationally reconstructing the use of "theoretical terms" (terms that apparently refer to unobservables) in science by treating those terms as being completely defined in terms of particular operational procedures, thereby eliminating the apparent references to unobservables.

Here’s what operationalism says. For any theoretical term (say, for example, "electron density") we can "rationally reconstruct" the use of that term by treating it as having an *analytic* operational definition in terms of laboratory procedures and instrumentation. So, for example, the operational definition of "electron density" might be given by a sentence of the form

> (*) The electron density in a region, *R*, is given by the value, *x*, if and only if *E* applied to *R* yields the value *x*,

where *E* is an instrument such that -- *prior to rational reconstruction* (but not after) -- scientists would have thought of it as a procedure for measuring electron density.

The analyticity of operational definitions like (*) is essential to the project of rational reconstruction. Operationalism *is not*, for example, the idea that electron density is defined as whatever magnitude instruments of sort *E* reliably measure. On that conception (*) would represent an empirical discovery about how to measure electron density, but -- since electrons are "unobservables" -- that’s a realist conception not an empiricist one. What the project of rational reconstruction requires is that (*) be true *purely as a matter of linguistic stipulation* about how the term "electron density" is to be used.

Since (*) is supposed to be analytic, it’s supposed to be unrevisable. There is supposed to be no such thing as discovering, about *E*,, that some other instrument provides a more accurate value for electron density, or provides values for electron density under conditions where *E* doesn’t function. Here again, thinking that there could be an improvement on *E* with respect to electron density requires thinking of electron density as a real feature of the world which *E* (perhaps only approximately) measures. But, that’s the realist conception which operationalism is designed to rationally reconstruct away.

In actual, and apparently reliable, scientific practice, changes in the instrumentation associated with theoretical terms is utterly routine, and apparently crucial to the progress of science. Scientists routinely replace one instrument with another in order to achieve (as they would say) more accurate measurements of some unobservable magnitude -- often in the light of new theoretical developments -- or to permit measurement of it under conditions for which previous instrumentation was inadequate. According to an operationalist conception, these sorts of modifications would not be methodologically acceptable. Most logical empiricists were not willing to accept this conclusion. After all, they intended to rationally reconstruct the best of actual scientific practice. So most logical empiricists felt compelled to reject operationalism.

Examples such as these made it clear that -- in apparently reliable scientific practice -- scientists behave as though (1) they obtain knowledge of unobservable (as well as observable) phenomena by deploying instruments which (perhaps indirectly) *detect* them, and (2) their theory dependent methodology in these and other matters is informed by knowledge of unobservables as well as of observables. In particular, they appear to improve, or extend the range of, procedures for measuring or detecting unobservable phenomena in the light of theoretical knowledge of those phenomena.

These features of scientific practice stimulated the articulation (largely by philosophers in the empiricist

tradition) of two different but related arguments for scientific realism, to which we now turn our attention.

3. Realist Responses to the Empiricist Challenge: The Senses Extended and Explanations Rehabilitated

3.1 Extending the Senses

The fact that scientists, apparently justifiably, treat certain instruments and procedures as ways of detecting and measuring unobservable (as well as observable) phenomena led several philosophers (see especially Feigl 1956, Maxwell 1962) to adopt what amounts to a non-empiricist (or, perhaps, more flexibly empiricist) understanding of knowledge empiricism according to which (1) the special epistemic role of the senses derives from the fact that they are the only detectors we have built in to our bodies, but (2) the range of phenomena we can detect and measure can be broadened by extending the range of our senses through the use of instruments and procedures whose justification is theory dependent. Thus knowledge of phenomena "unobservable" by traditional empiricists standards is possible. This sort of focus on laboratory detection and manipulation in defense of realism finds perhaps its most energetic expression in the writings of Hacking (see, for example, Hacking 1982).

3.2 Explanation Rehabilitated

The conception that instruments, designed with the help of theoretical understanding, can extend the range of the senses so as to provide information about unobservable phenomena surely has to be a component of any even remotely plausible defense of scientific realism. Still, by itself the idea that instruments can extend the senses is inadequate as a rebuttal to the basic underdetermination argument. Here’s why. The basic idea behind the extending-the-senses approach to defending scientific realism is that -- as scientists’ knowledge of unobservable phenomena improves and as instrumental design becomes more sophisticated -- measurement and detection would become possible for phenomena hitherto beyond the reach of reliable detection and measurement; think of going from light microscopes, to electron microscopes, to x-ray crystallography devices (which can produce images of atomic structures within crystals).

That has to be the realist’s conception, but consider the effect of underdetermination arguments. Suppose that, at some stage in the process of the improvement of theories and instruments, certain phenomena, *D*, posited by existing theories are detectable by the extended senses, but others are not. Let *T* be the total science of the time, and let T^* be a theory empirically equivalent to *T* with respect, not to their observational consequences, but with respect to their consequence regarding the phenomena in *D*. The basic underdetermination argument can be repeated with respect to *T* and T^*, leading to the conclusion that *T* does not reflect any knowledge of phenomena outside *D*. Thus there is no evidential basis for any extension of measurement and detection beyond *D*. Since this argument is applicable at any stage of any

supposed extension of the senses, it challenges -- in the name of knowledge empiricism -- *any* extension of the senses.

3.3 Explanation as Evidential

Considerations such as these seem to have focused the attentions of realists on what we might call *extra-experimental* standards for theory assessment. To see what these are, let's examine the *EIT* mentioned earlier. Why would a knowledge empiricist defend the *EIT*? An obvious answer is that she might think that the only consideration which ever justifies accepting one theory, *T*, over a rival, T^*, is that some prediction about observables obtained from *T* has proven to be true, whereas a prediction from T^* about the same experiment or observation has proven to be false.

But is anything like this right? Pretty obviously -- and pretty obviously by empiricist standards -- no. Here's why. Consider any case in which observations in some set, *O*, provide us with good scientific evidence to accept some theory, *T*, such that *T* applies to an range of observable cases not represented in *O* (that is, consider any case of scientifically justified induction). In any such case there will always be infinitely many pair-wise empirically *in*equivalent theories such that (a) each of them is empirically inequivalent to *T and* (b) each of them is compatible with all the observational data ever collected. This is just the Humean point that induction is not deductively valid. If we have sufficient scientific evidence to justify our accepting *T*, that evidence must justify our rejection of each of these other theories. [Note that this conclusion must be accepted whether one is an empiricist or a scientific realist regarding the interpretation of *T* and its rivals, since the theories in question are pair-wise empirically inequivalent and are empirically inequivalent to *T*.]

Let T^* be one of these rivals to *T*. T^* is empirically inequivalent to *T*, so it would be possible in principle to run a crucial experiment to discriminate between *T* and T^*. But, rational standards for the assessment of scientific evidence dictate that we are justified in rejecting T^* even though no such experiment has been run! So, there must be rational standards for the assessment of scientific evidence *in addition to the standards which say that evidence for or against a theory can be provided by the success or failure of observational predictions derived from the theory*. Let's call these standards *extra experimental*. They solve the equation (!):

> (!) *T*'s observational predictions have been thus far confirmed + *Y* = There is good scientific evidence favoring the empirical adequacy of *T*,
>
> AND, both realists and empiricists agree, they are capable of adjudicating between competing substantive conceptions of the world (because they can adjudicate between empirically inequivalent theories).

So, realists and empiricists agree that it isn't true that rational standards for the assessment of scientific evidence dictate that choice between competing theories must always be based on the results of crucial experiments. Where does that leave the underdetermination argument against knowledge of

unobservables?

Almost all scientific realist responses to empiricist anti-realism in the last three decades can be understood as variations on the idea that the solution to (!) -- which empiricists must agree exists on pain of abandoning selective skepticism for skepticism about induction -- also solves (!!):

> (!!) *T's* observational predictions have been thus far confirmed + *Y* = There is good scientific evidence favoring the (approximate) truth of *T*, even of its claims about unobservables.

Defenses of realism along these lines (see, e.g, Boyd 1983; Byerly and Lazara 1973; Lipton 1993; Miller 1987; McMullin 1984; Psillos 1999; Putnam 1972, 1975a, 1975b) deploy somewhat different resources, but one thing they have in common is that they reflect, and participate in, what might be called the rehabilitation of explanation in recent philosophy of science. An obvious reply to the *EIT* is that it ignores the role of explanation as an evidential standard: perhaps one, among a family of empirically equivalent theories, is to be preferred because it explains observable phenomena better than the others, even though it makes the same observational predictions. The standard logical empiricist treatment of explanation, the deductive-nomological account (see Hempel 1942, 1965; Hempel and Oppenheim 1948), responds by identifying the explanatory power of a theory with its predictive power.

Over the last several decades a great many philosophers have been critical of some aspects or other of this reduction of explanation to prediction (see, e.g., Boyd 1985; Kitcher 1981; Lipton 1991; Kitcher and Salmon 1989; McMullin 1984, 1987; Miller 1987; Salmon 1984, 1989). In the context created by this critical work, the notion of explanation, as an independent component of rational scientific methodology, has been to some extent rehabilitated.

A closely related development is also important. Goodman 1954 drew the attention of philosophers of science to the important point that only some hypotheses, the *projectible* ones, are in the running for confirmation by observations, and that projectibility judgments are in some way or other *a posteriori* judgments informed by previously established theories and practices. What has become pretty clear is that, however they are to be philosophically analyzed, projectibility judgments are in fact judgments of plausibility in the light of previously established theories (Boyd 1999; Lipton 1991, 1993), and that plausibility of the relevant sort is a matter of the sort of unification with those theories which has explanatory import. So, explanation, in its own right, and as an aspect of projectibility judgments, appears to play a crucial role in the assessment of observational evidence for scientific theories.

To a good first approximation, the following characterize the conditions under which observations, *O*, substantially confirm a theory *T*:

1. *T* is projectible (that is, theoretically plausible in the light of the best established science).
2. The observations in *O* either confirm predictions obtained from *T*, or validate explanations based on *T* or both.
3. For each of the projectible alternatives, *A*, to *T* which address the same questions, the observations

in *O* provide evidence against the predictions and/or explanations underwritten by *A*.
4. The observations in *O* were obtained under conditions which embody controls for each of the experimental artifacts or errors of sampling which are suggested by projectible conceptions of the relevant observational or experimental conditions.

The basic strategy of defenses of realism which argue that the solution to (!) -- which empiricists accept -- also solves (!!) involves arguing that the considerations of explanatory power of the sort indicated in characterizations like 1.-4. successfully adjudicate between empirically equivalent theories, so that knowledge of unobservables is sometimes obtained.

3.4 Two Explanationist Strategies for Defending Realism

There is a (very) rough division between two versions of the strategy in question. One strategy, let's call it *local explanationism*, (perhaps reflected in McMullin 1980, 1987; Miller 1987; and Lipton 1993) involves arguing that the relevant explanation-involving, extra-experimental criteria do, in some cases, successfully adjudicate between empirically equivalent theories, so that some scientific knowledge of unobservable phenomena is actual. An alternative approach, the *abductive strategy*, (see, e.g., Boyd 1983, Psillos 1999) treats scientific realism *itself* as a scientific hypothesis which is supported by the fact that it provides the only viable explanation for the such success as methodological principles like 1.-4., above, have as guides to the identification of *empirically adequate* theories. The justification of inductive methods in science is, therefore, provided by scientific realism, understood as itself an *a posteriori* scientific hypothesis.

There are interesting differences between these approaches, and between the various different versions of each, but certain empiricist challenges can be raised against all or most of them. Fine (1984, 1986a) has offered two very significant, and closely related, criticisms of the abductive strategy. First, Fine argues, the strategy begs the question against anti-realist positions by treating scientific explanatory power as carrying evidential weight in philosophy. After all, the dispute between empiricist anti-realists and realists is, in the first instance, a dispute about whether a theory's explanatory power can count in favor of the claims it makes about unobservables. [van Fraassen 1980 makes similar criticisms; he and Laudan 1981 each also challenges the claim that scientific realism provides the best explanation for the reliability of scientific methods in identifying empirically adequate theories.]

Fine's second criticisms is more abstractly epistemological. He points out that, according to the realist who adopts the abductive strategy, the methods of science are to be philosophically justified by appeal to *a posteriori scientific* findings, i.e., by appeal to the scientific realist's scientific explanation for their reliability. This approach, he argues, violates the philosophical requirement that the justification for the methods in a domain of inquiry should be grounded in methods more secure than the methods being justified.

Plainly these criticisms represent serious challenges to the abductive strategy. Importantly, they also challenge *any* version of the local explanationist strategy unless it incorporates an *a priori* (as opposed to

an empirical scientific) defense of the evidential relevance of the explanatory power of theoretical claims about unobservables. There are two reasons to doubt that such an *a priori* defense is available.

In the first place, philosophical defenses of epistemological positions almost always rest, at least in part, on appeals to philosophical "intuitions" regarding particular cases. Although many philosophers regard the deliverances of philosophical intuitions as justified *a priori*, in fact epistemic intuitions about particular cases deliver to us the results of our trained (or, in some cases, untrained) judgments regarding the domain of inquiry in question. They are reliable guides to matters epistemological just in case -- and to the extent that -- the training in question has itself been relevantly reliable (Boyd 1999).

For philosophers of science, the relevant training centrally includes training in the methods and findings of the relevant sciences. Since, "pre-analytically" at least, those methods countenance inductive inferences to explanations involving unobservables, and since the most celebrated findings often incorporate the results of such inductions, a very significant burden of proof would rest on someone who maintained that her philosophical arguments in favor of accepting inductive inferences to explanatory theories about unobservables did not, at least tacitly, rest on intuitions which beg the question against empiricist anti-realism.

Moreover, there are independent reasons to doubt that there could be *an a priori* defense of accepting the results of inductive inferences to the best explanation, *whether or not that explanation posits unobservables*. To a good first approximation, typical scientific explanations offer accounts of the causal mechanisms or processes by which some phenomena are brought about, and scientists evaluate the explanatory power of a theory by trying to assess the likelihood that mechanisms or processes posited by the theory operate to produce the relevant effects. Their judgments in these matters are, almost always, informed by experiments and observations but they are nevertheless highly theory dependent, ordinarily relying heavily on previously established "background" theories concerning the relevant sorts of causal mechanisms and processes (for accounts with this flavor see, e.g., Lipton 1993, Psillos 1999, Boyd 1985). Such judgments are reliable only to the extent that those background theories are relevantly approximately accurate.

In consequence, any defense of the practice of counting the explanatory power of a particular theory as providing evidence in its favor would appear to require a defense of the proposition that the findings of the relevant background sciences are relevantly approximately accurate. While, in some cases, this may be a justified conclusion, its justification could hardly be *a priori* (for an account somewhat more sympathetic to *a prioricity* for certain cases, see Miller 1987). Exactly similar arguments regarding theory dependent judgments of projectibility provide additional *prima facie* support for the same anti-*aprioristic* conclusion (Boyd 1999).

In the light of these challenges, there is a strong case to be made that any defense of scientific realism must rest on a conception according to which both scientific methods *and* methods in the philosophy of science, must lack *a priori* justifications. Such a conception of science, and of the relevant parts of philosophy, would thus be non-foundational and, presumably, naturalistic (see Psillos 1999). [For a somewhat different naturalistic conception, see Kitcher 1993. For an excellent discussion of competing

metaphilosophical conceptions in the philosophy of science and their relation to debates about realism see Wylie 1986.]

3.5 Realism and Approximate Truth

Whether or not the defense of scientific realism requires the adoption of a non-foundationalist conception of knowledge, it almost certainly requires the articulation of a conception of approximate truth. It is central to any plausible realist conception that, at least sometimes, the historical development of scientific theories reflect progress by successive approximation to the truth -- about unobservables as well as about observables, and it is central to arguments for realism that involve the rehabilitation of explanation as an epistemic notion that relevant improvements in approximate knowledge are typically reflected in improvements of method. So, realist philosophy of science relies heavily on the notion of approximate truth.

Laudan 1981 raises against scientific realism (and especially against abductive arguments for realism) the "pessimistic meta-induction." He points out that there are lots of real historical cases in which scientific theories which have been predictively successful and have contributed positively to scientific methodology have not been true, so that the truth of scientific theories need not be posited in order to explain the successes of scientific practice.

The obvious realist reply is that what must be posited is the *approximate* truth the relevant theories (see Hardin and Rosenberg 1982 and Laudan 1984). Articulation of this reply raises important issues, since *any* consistent theory can be represented as approximately true in certain respects, Moreover, as Laudan points out, many of the historically important and methodologically significant theories are, by our current lights, deeply false in some important respects. Efforts to develop an appropriate account of approximate truth in science include Niiniluoto 1987, Oldie 1986, Weston 1992, Boyd 1990.

One novel approach to the problem of approximation is provided by Worrall’s structural realism (Worrall 1994; for a critique see Psillos 1995, 1999). The basic idea here is that the most serious departures from the truth in scientific theories tend to be errors about the *natures* of the basic phenomena rather than about their *structural relations*. In the light of this generalization, the structural realist proposes accepting the claims about causal *structures* (even unobservable ones) posited by well confirmed theories while withholding acceptance from what those theories say about the *natures* of the phenomena so related. To a good first approximation, one might think of structural realism as the view that, for any well established scientific theory, *T*, one should accept the Ramsey sentence obtained from *T* by replacing each theoretical term in *T* by a new variable, and then prefixing, to the resulting open sentence, existential quantifiers over those variables, where the quantification is understood to range over causal structures in nature.

Aside from its importance as a contribution to the literature on approximate truth, structural realism is significant in two other ways. In the first place, it reflects a general tendency in the literature on scientific realism to worry about the extent to which scientific realists must portray scientific knowledge as potentially resolving genuinely *metaphysical* questions. Putnam’s internal realism and Fine’s natural

ontological attitude (discussed below) may be seen as important ontologically deflationary approaches to this question.

The other significance of structural realism lies in the fact that the distinction upon which it relies -- that between causal structures and natures -- may have been, in a certain sense, challenged by philosophers like Shoemaker (1980) who hold that properties, magnitudes, states and the like are defined by their contributions to the causal powers of things. It is an interesting question whether approaches to metaphysics like Shoemaker's are compatible with the approaches to approximation informed by structural realism.

4. The Neo-Kantian Challenge: First Version

Hanson (1958) and, especially, Kuhn (1970 -- first published 1962; see Scheffler 1967, Shapere 1964 for early discussions) raised significant challenges to scientific realism, arguing from the theory dependence of methods (and, especially, of observation) to the conclusion that a realist conception of the growth of approximate scientific knowledge cannot be sustained The intellectual impact of their work in the philosophy of science has been very different from the impact it has had in the rest of the humanities and in many of the social sciences. In the later disciplines the impact of Kuhn, especially, has been to underwrite the sort of anti-realist "postmodernism" discussed later in this essay. In the philosophy of science, by contrast, the impact of Hanson and Kuhn has been mainly to stimulate the articulation of *naturalistic* or *causal* conceptions of reference and *essentialist* conceptions of the definitions of scientific kinds and properties. [I am here presenting what might be thought of as the "standard" conception of Kuhn's position and of responses to it. There has been a recent revival of interest in Kuhn among analytic philosophers and others, and alternative readings of Kuhn are possible (see, e.g., Hoyningen-Huene 1993 and the papers collected in Hoyningen-Huene and Sankey 2001). Whatever the merits of less standard interpretations of Kuhn, it was the standard conception of his arguments that occasioned the realist responses discussed here.]

That arguments proceeding from the theory dependence of scientific methods and of measurement should have been deployed *against* realism is initially surprising. After all, most of the significant arguments *for* scientific realism emphasize theory dependence. Moreover, Kuhn's discussion of what he calls *normal science* seem to have exactly anticipated the abductive argument for realism discussed above. He insists that the success of research in normal science is explained, in significant part, because scientific practitioners have, as a result of their understanding of the paradigmatic theory, a quasi-metaphysical knowledge of the basic (and often unobservable) causal factors involved in the phenomena they study.

Where Kuhn's account departs from a realist conception of the growth of approximate knowledge is in his treatment of what he calls *scientific revolutions*. Although most empiricist philosophers of science had recognized the theory dependence of scientific methods even before the work of Hanson and Kuhn, it was Hanson's and Kuhn's work which made it clear that accepting the theory dependence of scientific methods raise the possibility of *incommensurability* between competing scientific theories (or *paradigms*): the possibility that in science there might be disagreements between theoretical perspectives

such that there do not exist methods for their resolutions which are both *rational* and *fair* (to the competing positions).

What each author claimed was that this situation had actually obtained in important historical cases where, according to a realist perspective, one might think that the rational application of scientific methodology had resulted in the replacement of one theory by a more nearly accurate one. What was especially striking -- and challenging to a realist conception -- was Kuhn's claim that among the "scientific revolutions" where this had occurred was the transition from Newtonian mechanics to special relativity at the beginning of the 20th century.

What is important in understanding the realist response to Kuhn's claim about this particular historical case is that there are lots of experimental results (like, e.g., those which are ordinarily understood to reflect the increase of mass of particles in a cyclotron) such that they certainly look like cases in which a methodology -- including measurement procedures -- which is acceptable by both Newtonian and relativistic standards adjudicates in favor of the relativistic conception. Lots and lots of relativistic effects are such that they can be, apparently, detected and measured using instruments whose design begs no questions against either of the competing "paradigms." The transition from Newtonian mechanics to special relativity certainly looks like a textbook case of rational progression from, one theory to an even more accurate one.

Against this picture Kuhn argues that no such successive approximation occurred because Newtonian mechanics and relativity theory do not share a common subject matter regarding which the latter is a better approximation than the former. For example -- he argues -- the term "mass" as it occurs in Newtonian mechanics does not refer to the same magnitude as does the term "mass" in relativistic mechanics because "Newtonian mass is conserved; Einsteinian is convertible with energy. Only at low relative velocities may the two be measured in the same way, and even then they must not be conceived to be the same (102)."

In giving this remarkable argument Kuhn was tacitly relying on a conception of the referential semantics of scientific terms probably derived from the work of Carnap (see Carnap 1950; there are important controversies about the proper interpretation of Carnap -- see, e.g., Friedman 1987, 1991 -- but they are irrelevant to our story). The conception in question is a version of the standard empiricist "descriptivist" conception that the referent of a term is picked out by a description which constitutes the analytic definition of the term in question. According to the version Kuhn relies on, the analytic definition of a scientific term is provided by the most basic laws containing the term. Thus, as the example of "mass" illustrates, any change in the fundamental laws involving a scientific term must involve a change in referent (or reference *failure*, a possibility Kuhn 1970 does not discuss).

4.1 "Causal" and Naturalistic Theories of Reference

What was important for the development of realist philosophy of science was the fact that most philosophers of science were, at least tacitly, themselves inclined to some version of analytic

descriptivism. The anti-realist consequences which Kuhn (and Hanson) derived from descriptivist conceptions let to the articulation by realists of alternative theories of reference. Characteristically, these theories followed the lead of Kripke (1971, 1972), whose work was mainly concerned with the semantics of modality, and Putnam (1972, 1975a, 1975b), whose work was mainly concerned with issues in the semantics of scientific terms. Each of them advocated a "causal" theory of reference according to which the reference relation between a term and its referent was a matter of there being the right sort of (chain of) causal relation(s) between uses of the term and (instances of) its referent. Numerous variations on this *naturalistic* theme -- some assigning importance to descriptive elements as well as causal relations in the establishment of reference -- have been proposed (see, e.g., Boyd 1999, Dretske 1981, Enç 1976, Field 1973, Kitcher 1992, Miller 1987, Papineau 1987, 1993, Psillos 1999). It is by now pretty well accepted that some departure from analytic descriptivism, involving some causal elements, is a crucial component of a realist approach to scientific knowledge.

4.2 Realism and the Revival of Essentialism

Kuhn's analytic descriptivism assigns to the analytic definition of a scientific term the role of fixing its referent. Once that role is assigned to other ("causal," "naturalistic") features of term use, it becomes possible to explore the issue of non-analytic *a posteriori* definitions of the kinds, magnitudes, etc. to which scientific terms refer. The work of Kripke and Putnam just cited gave rise to a class of theories according to which scientific kinds, etc. have real rather than nominal definitions ("real essences" rather than "nominal essences" in the sense of Locke 1689). The paradigm example is that the real definition, or essence, of water is described by the formula "H_2O". It is by now a standard feature of realist conceptions of science that they incorporate some version or other of the idea that scientific kinds, categories, etc. (*natural kinds*) possess such real definitions (for interesting discussions of the development of this realist conception with special reference to biological kinds see Wilson 1999a, 1999b).

The idea that natural kinds possess such definitions has been consistently linked, in the realist literature, to discussions of the *projectibility* of predicates and hypotheses (Goodman 1954, Quine 1969). Only by reference to kinds (etc.) with real rather than nominal definitions -- only by, *in some sense or other*, "cutting the world at its (*a posteriori* defined) joints" -- are we able to fit our language use to the world in such a way as to make reliable induction and explanation possible (Boyd 1999; Psillos 1999; Putnam 1972, 1975a, 1975b; Sismondo 1996; Wilson 1999b).

One further point about real essences is important. The stock example of a real definition (H_2O for water) might suggest that real definitions of scientific kinds (etc.) must, like logical empiricists' ideal nominal definitions, specify necessary and sufficient conditions for kind membership. In fact, examination of cases in those sciences which study complex phenomena indicate that some natural definitions may consist of families of imperfectly "clustered" properties, with the result that the kinds they define do not have precisely determinate boundaries (Boyd 1999, Wilson 1999b; but see also Hacking 1991a, 1991b). Realism may imply that there is, in that sense, vagueness in nature (contrast Putnam 1983).

4.3 The Metaphysics of Social Construction

Kuhn tacitly adopts a semantic conception according to which the most basic laws in a paradigm are *exactly* true by linguistic convention. He also claims that such laws provide quasi-metaphysical knowledge of basic causal factors. His claim that these laws are exactly true is what leads him to conclusions about the (semantic) history of recent physics which are *prima facie* implausible, and it is this feature of his semantic conception against which causal or naturalistic theories of reference are mainly directed.

The example of the semantics of the names of fictional characters indicates that the linguistic conventions operating in fiction make it possible to establish it *by convention* that certain claims about a character are *approximately* true without thereby establishing their *exact* truth. Versions of Kuhn's social constructivist position could, therefore, be formulated according to which the establishment of a paradigm imposes *by convention*, on the phenomena scientists study, a quasi-metaphysical structure which makes the central laws of the paradigm approximately (but not necessarily exactly) true.

Although Kuhn never considered this version of constructivism, it fits well with the tradition of anthropological relativism to which Kuhn's position is often assimilated. It is not refuted by arguments for causal or naturalistic theories of reference, nor does it entail wildly implausible claims about incommensurability in recent science. It is, however, pretty clearly an anti-realist position -- one which has resonances with the sorts of "postmodern" anti-realism discussed later in this essay. A realist rebuttal to it is available if one makes explicit, and defends, a piece of common, and philosophical, sense about the metaphysics of conventionality: the no non-causal contribution thesis (2*N*2*C*). According to 2*N*2*C*, human social practices make no non-causal contribution to the causal structure of the world, and are in that way metaphysically innocent (see Boyd 1999).

5. The Neo-Kantian Challenge: Second Version

Structural realism represents one attempt to defend scientific realism while being modest about its metaphysical implications. Putnam's "internal realism" and Fine's closely related "natural ontological attitude" (*NOA*) represent other attempts to follow scientific realists in taking the findings of science at "face value" while avoiding realism's excessively metaphysical understanding of those results (Putnam 1978, 1981, 1983a; Fine 1984, 1986a, 1986b, 1991; for a nice exposition see the Introduction to Papineau 1996; for critiques see Glymour 1982; Millikan 1986; Newton-Smith 1989a, 1989b; McMullin 1991; Papineau 1987).

"Internal realism" and *NOA* are not easy to explicate and are, almost certainly, not the same position. Nevertheless they share some important elements in common.

- "*Thin*" *Truth*: Both Putnam and Fine assert that one can (and should) accept the well established theories of science (even about unobservable) as (probably) *true*, but that this should not be understood as accepting the "metaphysical realist" (Putnam's term) view that the statements which constitute those theories *correspond to reality*. They advocate a "thin" conception of truth rather

than a correspondence conception. In Putnam's early papers defending internal realism he adopts a pragmatic conception of truth according to which truth of sentences is a matter of being epistemically acceptable in the limit of ideal inquiry. In the case of Fine, it's less clear exactly which thin conception of truth is at work.

- *De-Natured Naturalism*: Naturalistic conceptions of reference have it that reference of scientific terms is a matter of certain causal patterns which relate the use of terms to instances of their referents. Relations of measurement and detection are supposed to be centrally involved in the establishment of reference, at least in paradigm cases. It is explicit in Putnam, and surely implied by Fine's position, that *if* the causal theory of reference, and related causal theories of detection and measurement, are understood as *scientific* theories (in linguistics, say, and -- for theories of measurement and detection -- theories in the relevant sciences) then they might, for all the internal realist or *NOA*er says, be well confirmed. What is to be denied is that such conceptions underwrite a correspondence conception of truth. They are bits of science, but not (also) bits of realist naturalist philosophy.

5.1 An Analogy with Phenomenalism

An analogy with issues regarding knowledge of the external world may be helpful here. One classical early logical empiricist response to questions about our knowledge of (observable) external objects was the phenomenalist strategy of representing external objects as "logical constructions" analytically defined in sense-datum terms (see, e.g., Carnap 1928). That certain experience patterns constituted experiences of, e.g., chairs was supposed to reflect, not a discovery about some epistemically important metaphysical relation between chairs and those patterns, but, instead, the implication of the analytic definition of "chair" in the sense-datum language.

Nevertheless, nothing in the phenomenalist project was supposed to preclude the possibility that psychologists studying perception might discover that those very experience patterns are caused by light reflected off chairs and stimulating the retina is particular ways. This would be unobjectionable as a bit of empirical science, but it was not to be understood as positing an epistemically relevant relation of detection and representation between the experiential pattern and *chairs*, understood as experience-independent features of the world. It could not be understood as a component in *philosophical* justification of the claim that we know about, and "chair" refers to, experience-independent chairs.

By contrast, non-foundationalist "causal" or "reliabilist" conceptions of perceptual knowledge in the tradition initiated by Goldman (1967, 1976) would treat the relevant discovery *both* as an empirical scientific discovery *and* as a component of a (naturalistic) *philosophical* (and *epistemically relevant*) explanation of why our chair beliefs sometimes represent knowledge about (experience independent) chairs. Similarly, if the psychological findings in question were incorporated into a suitable empirical theory of language use they could, on a causal or naturalistic conception of reference, underwrite the *philosophical* conception that "chair" refers to (experience independent) chairs.

5.2 Non-Foundationalist Epistemology Again

What this suggests is that a defense of realism against internal realism or NOA must follow the lead of non-foundational causal theories of knowledge, and of perception in particular, in insisting that scientific findings about, e.g., the measurement and detection of theoretical entities, and about the reference of scientific terms have philosophical as well as scientific relevance (see, e.g., Boyd 1999, Byerly and Lazara 1973,Hacking 1982, Psillos 1999).

5.3 Challenges Regarding the Epistemology and Metaphysics of Reference and Kinds

The arguments Putnam offers in defense of internal realism are complex, and (as the critiques cited indicate) both controversial and sometimes hard to explicate. Nevertheless, it seems pretty clear that Putnam attributes to "metaphysical realism" something like the following commitments:

1. Reference is a relation between linguistic entities and entirely *extra-linguistic* (and in that sense *independently existing*) natural kinds. Natural kinds (magnitudes, etc.) are, somehow or other, out in the world, and available for discovery and naming. There is a single set of such natural entities somehow given by the structure of the world itself, independently of human practice.
2. The reference relation between natural kind (magnitude, etc.) terms and their referents is a *purely* causal matter, where the purity in question is a matter of the reference relation's being definable without significantly acknowledging descriptive elements or human intentions.

If one accepts this picture of scientific realism, understood metaphysically, then it is natural to think that what makes the associated conception of truth a *correspondence* conception is that reference is seen as a relation between terms and such *independently existing* kinds. The realist correspondence conception, so conceived, is subject to two important challenges.

First, if we think of natural kinds as things somehow independent of linguistic and methodological practices, then there are lots of natural kinds out there, and it is difficult to see how the causal conception of reference fixing could explain how a natural kind term could ever have a unique referent. This problem is exacerbated if one thinks of reference as being purely causal in the way just indicated, since intentional and descriptive factors, which might otherwise be thought to reduce the ambiguity of the reference relation, are set aside. Such considerations seem to be the basis of the "model theoretic" arguments in Putnam 1978, 1980) against "metaphysical realism."

Secondly, reference to natural kinds is supposed to explain the inductive successes of scientific practice, so there must be some quite intimate connection between natural kinds and the conceptual machinery of the sciences. If one thinks of realist theories as entailing that natural kinds are independent of that machinery, it is hard to see how the explanation could work unless it rested on something like a *objective idealist* theory according to which natural kinds are somehow metaphysically "fitted" for explanation and induction independently of the relevant practices. Such an assumption is profoundly at odds with the philosophical naturalism and metaphysical materialism ordinarily associated with scientific realism. This

sort of consideration appears to underwrite aspects of Putnam's criticism of materialist metaphysical realism in "Why There Isn't a Ready Made World" (Putnam 1983a).

5.4 Realist responses

Of these two challenges, the first has received much more attention from scientific realists. There has been widespread acceptance of the view that descriptive and/or intentional factors must figure in any scientific realist account of reference (e.g., Boyd 1999, Enç 1976, Kitcher 1992, Papineau 1979, Psillos 1999).

Much less has been said by realists about the sense, if any, in which scientific realism is committed to there being natural kinds (etc.) which are independent of us. Psillos (1999), for example, discusses problems with pure causal theories of reference extensively, but takes it to be a basic posit of scientific realism that "…the world has a definite and mind-independent natural-kind structure" (xix). Boyd (1999) offers an alternative approach according to which, like natural kind *terms* and classificatory *practices*, natural kinds themselves should be thought of as social *artifacts* deployed in achieving an appropriate fit or accommodation between inductive and explanatory practices and relevant causal structures.

Whether the intrusion of descriptive and intentional notions into realist accounts of reference, or the treatment of natural kinds as social artifacts, is compatible with the main spirit of scientific realism depends on the sense(s) in which scientific realism should be understood as entailing that the phenomena scientists study are "mind independent." A possible response to this question, compatible with the proposals just mentioned, is that the relevant sense of mind independence is fully captured by the no non-causal contribution doctrine discussed earlier.

6. The "Post-modern" Challenge

Most recent work in the relatively new discipline of science studies (see, e.g, Biagioli 1999; Galison 1987; Latour and Wolgar 1979; Latour 1987; Pickering 1984, 1995; Pinch 1985; Shapin 1982, 1994; Shapin and Schaffer 1985) and a significant body of work in feminist philosophy of science or feminist aproaches to particular science (see, e.g., Alcoff and Potter 1993; Antony 1993; Antony and Witt 1993; Conkey and Spector 1984; Fuss 1989; Gero and Conkey 1991; Harding 1986, 1987, 1991; Harding and Hintikka 1983; Harding and O'Barr 1987; Hartsock 1987; Haslanger 1993; Keller 1983: Longino 1989, 1990; Tuana 1989; Wright 1996; Wylie 1991, 1993, 2000; Wylie and Okruhlik 1987) has been to some extent influenced by, or has engaged with, anti-realist "postmodern" conceptions according to which such phenomena as science, knowledge, evidence and truth are *social constructions*, in some sense or other which implies that one should reject the idea that scientific practices achieve an approximate representational fit of some sort or other between the content of scientific theories and *the world* or *reality*.

Although serious interchanges between scientific realism and these approaches have not developed to the level of exchanges between, e.g., scientific realist approaches and logical empiricist or neo-Kantian ones,

a number of philosophers of science have defended a realist approach against post modern relativism and skepticism (see, e.g., Boyd 1999; Kitcher 1993; Papineau 1998; Pettit 1998; Sismondo 1993a, 1993b, 1996). Several factors are probably important in determining the dimensions of the dispute between realists and postmodernists.

6.1 Boundary Work

Sociologists of science have identified a feature of scientific work which is especially important when new (sub)disciplines are being established. Practitioners of emerging disciplines devote considerable effort to distinguishing the approach of their new disciplines from the approaches of more fully established disciplines, often by adopting a substantially adversarial stance towards them. This phenomenon has been played out in the establishment of science studies and (to a somewhat lesser extent)feminist philosophy. In each of these cases, mainstream realist and empiricist approaches to epistemology have been special targets of such adversarial stance taking. In the case of science studies, to a very good first approximation, the "boundary work" foundations of the emerging discipline rest on a critique of epistemology and of the correspondence conception of truth. This perfectly ordinary, and non-culpable, boundary work resistance to traditional epistemology and the correspondence theory has proven to be a barrier to communication between mainstream philosophers of science and others in science studies.

6.2 Postmodern Responses to Naive Empiricist Conceptions of Objectivity

There is a prevalent conception of scientific objectivity which is historically associated with empiricist conceptions of science, even though it is sufficiently naive that probably no professional empiricist philosopher of science ever defended all of its components. According to it, the objects of scientific study are natural kinds (etc.) which are

1. independent of human practices,
2. defined by
    a. eternal,
    b. unchanging,
    c. ahistorical, and
    d. intrinsic

    necessary and sufficient membership conditions;
3. referred to in
    a. fundamental,
    b. exceptionless,
    c. eternal, and
    d. ahistorical

    laws; and
4. discovered by the deployment of

a. eternal,
b. ahistorical,
c. politically and culturally neutral, and
d. foundational

scientific methods.

To a significant extent, anti-realist postmodern conceptions of science take these components of naive empiricism to be definitive of the notion of scientific objectivity. Postmodern students of science hold -- correctly (Boyd 1999; Sismondo 1993a, 1993b, 1996; Knorr Cetina 1993) -- that nothing in actual scientific practice even remotely fits these criteria for objectivity. On this basis they often reach the anti-realist conclusion that scientific research never achieves objective knowledge. It is characteristic of defenses of realism against postmodern anti-realism that they deny, about one or more of the components mention, that they are necessary for objective knowledge.

6.3 "Quantum Superposition" of Conceptions of Social Construction

There are, in the literature and in intellectual discourse, roughly three versions of "social constructivism," the view that science is the "social construction of reality."

1. *Neo-Kantian social constructivism*. This is the view discussed earlier according to which the adoption of a scientific paradigm *successfully* imposes a quasi-metaphysical causal structure on the phenomena scientists study.
2. *Science-as-social-process social constructivism*. This is the view that the production of scientific findings is a social process subject to the same sorts of influences -- cultural, economic, political, sociological, etc. -- which affect any other social process.
3. *Debunking social constructivism*. This is the skeptical position according to which the findings of work in the sciences are determined exclusively, or in large measure, not by the "facts," but instead by relations of social power within the scientific community and the broader community within which research is conducted.

These are quite distinct positions. For example, 1. and 3. are mutually inconsistent, and 2. is compatible with either 1., or 3., or with standard logical empiricist and scientific realist conceptions. Nevertheless, in science studies and in other disciplines influenced by postmodernism they tend to become conflated.

In the first place, many practitioners in such disciples, for reasons rehearsed above, take 2. to imply that traditional realist and empiricist conceptions are mistaken. Moreover, having adopted 2., they tend to adopt a position which looks like a quantum superposition of 1. and 3., oscillating between thinking of scientific practice as (really) constructing the (quasi-metaphysical) truth and denying that it leads to truth in any metaphysically interesting sense.

The inconsistencies involved are made clearest in cases in which scientific theories of race and gender are said to be "social constructions." Often the intent here is to engage in scientific and political criticism but,

in so far as the neo-Kantian, and the fully debunking conceptions of social construction are simultaneously operative, authors often have a difficult time finding the resources for saying that such theories are really (*really*!) false. [For discussions of these conflations and their impact on methodological and political criticism see Sismondo 1993a, 1993b, 1996; Knorr Cetina 1993; Boyd 1999.]

6.4 Naturalism and the symmetry thesis

In one of the founding documents of contemporary science studies, Barnes and Bloor (1982) criticize a tendency in the history, philosophy, and sociology of science to treat true and false scientific theories asymmetrically: explaining the acceptance of true theories as the ordinary and to-be-expected result of applying the scientific method, but explaining the acceptance of false theories by appealing instead to the operation of "social factors." They propose that explanations for the acceptance of scientific theories should be symmetrical, appealing to the same sorts of factors in explaining the acceptance of true and false theories.

In science studies, it has been nearly universal to accept the symmetry thesis *and* to interpret it as requiring that *truth* or *the facts* not be treated as among the factors involved in explaining the adoption of scientific theories. Almost certainly, a defense of scientific realism in the light of the symmetry thesis will require insisting that a *naturalistic* scientific realism *does*, by considering facts of all sorts potentially relevant to the explanation of the acceptance of scientific theories, satisfy the requirements of the symmetry thesis. The *locus classicus* for this approach is Antony 1993; it is developed in Sismondo 1999.

6.5 Essentialism

One of the most important sources of resistance to scientific realism among feminist philosophers has been the conception that realism underwrites essentialism and that essentialism is a central component of racist and sexist ideology (see Fuss 1989 for a discussion). A naturalistic version of scientific realism *does* entail a sort of essentialism about natural kinds (etc.) but that sort of essentialism need not have the form suggested by the stereotype of scientific objectivity discussed above, and need not be inimical to critiques of scientific racism or sexism (Boyd 1999, Sismondo 1996). In particular, it is compatible with the sort of realist naturalism discussed here that social categories like race and gender might have *as their essences* a certain role in the stabilization or justification of particular sorts of historically situated oppression and exploitation. Similarly, realist naturalism is compatible with the view that some social categories (like races and genders) or psychological categories (like mental illnesses) are real, but are in some respects artifacts of classificatory (and other) social practices (see, e.g., Hacking 1986a, 1986b). All that is required by naturalistic realism is that the contribution of social practices not violate 2*N*2*C*.

6.6 Concluding Remark

Scientific realism is, by the lights of most of its defenders, the sciences' own philosophy of science. Considerations of the significant philosophical challenges which it faces indicate that it can be effectively defended only by the adoption of a *meta*philosophical approach which is also closely tied to the science,

viz., some version or other of philosophical naturalism.

Bibliography

- Alcoff, L., and Potter, E., 1993. *Feminist Epistemologies*, New York: Routledge.
- Antony, L., 1993. "Quine as a Feminist: The Radical Import of Naturalized Epistemology" in Antony and Witt, eds. *A Mind of One's Own* Boulder: Westview Press.
- Antony, L., and Witt, C., eds., 1993. *A Mind of One's Own* Boulder: Westview Press.
- Barnes, B., and Bloor, D., 1982. "Relativism, Rationalism and the Sociology of Knowledge" in Hollis and Lukes, eds. *Rationality and Relativism* Cambridge: MIT Press.
- Bennett, J., 1964. *Rationality*. London: Routledge & Kegan Paul (reprinted 1989, Indianapolis: Hackett).
- Biagioli, M., 1999. *The Science Studies Reader*. New York: Routledge.
- Boyd, R., 1983. "On the Current Status of the Issue of Scientific Realism." *Erkenntnis* 19: 45-90.
- -----, 1985. "Observations, Explanatory Power, and Simplicity." In P Achinstein and O. Hannaway (eds.) *Observation, Experiment, and Hypothesis In Modern Physical Science*. Cambridge: MIT Press.
- -----, 1990. "Realism, Approximate Truth and Philosophical Method" in Wade Savage, ed. *Scientific Theories*, Minnesota Studies in the Philosophy of Science vol. 14. Minneapolis: University of Minnesota Press
- -----, 1999. "Kinds as the "Workmanship of Men": Realism, Constructivism, and Natural Kinds." in Julian Nida-Rümelin, ed. *Rationalität*, *Realismus*, *Revision*: *Proceedings of the Third International Congress*, *Gesellschaft für Analytische Philosophie*. Berlin: de Gruyter.
- Byerly and Lazara, 1973. ""Realist Foundations of Measurement.", *Philosophy of Science* (40): 10-28.
- Carnap, R., 1928. *Der Logische Aufbau der Welt*. Berlin.
- -----, 1932. Überwindung der Metaphysik durch Logische Analyse der Sprache." *Erkenntnis*, vol II.
- -----, 1950. "Empiricism, Semantics and Ontology." *Revue internationale de philosophie*, 4th year.
- -----, 1959. "The Elimination of Metaphysics Through Logical Analysis of Language." in Ayer, A.J. 1959 (translation of Carnap 1932).
- Conkey, M., and Spector, J., 1984. "Archaeology and the Study of Gender." in M. Schiffer, ed. *Advances in Archaeological Method and Theory* 7:1-38. New York: Academic Press.
- Demopoulos, W., 1982. "Review of *The Scientific Image* by Bas C. van Fraassen," *Philosophical Review* 91: 603-7.
- Dretske, F., 1981. *Knowledge and the Flow of Information*. Cambridge: MIT Press.
- Enç, B., 1976. "Reference of Theoretical Terms." *Nous* 10: 261-282.
- Feigl, H., 1956. "Some Major Issues and Developments in the Philosophy of Science of Logical Empiricism." in H. Feigl and M. Scriven (eds.) *Minnesota Studies in the Philosophy of Science*, vol. 1. Minneapolis: University of Minnesota Press.
- Field, H., 1973. "Theory Change and the Indeterminacy of Reference." *Journal of Philosophy* 70: 462-481.

- Fine, A., 1984. "The Natural Ontological Attitude." in J. Leplin (ed.) *Scientific Realism*. Berkeley: University of California Press.
- -----, 1986a. *The Shaky Game*. Chicago: University of Chicago Press.
- -----, 1986b. "Unnatural Attitudes: Realist and Instrumentalist Attachments to Science." *Mind* 95: 149-79.
- Friedman, M., 1987. "Carnap's Aufbau Reconsidered." *Nous* 21: 521-45.
- -----, 1991. The Re-Evaluation of Logical Positivism." *The Journal of Philosophy* 88: 505-19.
- Fuss, D., 1989, Essentially *Speaking*, New York: Routledge.
- Galison, P., 1987. *How Experiments End*. Chicago: University of Chicago Press.
- Gero, J., and Conkey, M., 1991. *Engendering Archaeology*: *Women and Prehistory*. Oxford: Basil Blackwell.
- Glymour, C., 1982. "Conceptual Scheming, or Confessions of a Metaphysical Realist." *Synthese* 51: 169-180.
- Goldman, A., 1967. "A Causal Theory of Knowing.'' *Journal of Philosophy* (LXIV): 357-372.
- -----, 1976. ""Discrimination and Perceptual Knowledge.'' *Journal of Philosophy* (LXXIII): 771-791.
- Goodman, N., 1954. *Fact Fiction and Forecast*. Cambridge: Harvard University Press.
- Hacking, I., 1982. "Experimentation and Scientific Realism." *Philosophical Topics* 13 (71-87).
- -----, 1991a. "A Tradition of Natural Kinds" Philosophical *Studies* 61: 109-126.
- -----, 1991b. "On Boyd" *Philosophical Studies* 61: 149-154.
- -----, 1986a. "The Invention of Split Personalities" in Donagan, Perovich and Wedin, eds. *Human Nature and Natural Knowledge*, Dordrecht: D. Reidel.
- -----, 1986b. "Making up People." in T. Heller, M. Sosna, and D. Wellbery, eds. *Reconstructing Individualism*. Stanford: Stanford University Press.
- Hanson, N.R., 1958. *Patterns of Discovery*. Cambridge: Cambridge University Press.
- Hardin, C., and Rosenberg, A., 1982. "In Defense of Convergent Realism" *Philosophy of Science* 49: 604-615.
- Harding, S., 1986. *The Science Questiopn in Feminism*. Ithaca: Cornell University Press.
- -----, 1991 *Whose Science, Whose Knowledge*? Ithaca: Cornell University Press.
- -----, (ed.), 1987, Feminism *and Methodology*. Bloomington and Indianapolis: Indiana University Press.
- Harding, S., and Hintikka, M. 1983. *Discovering Reality*. Dordrecht: D. Reidel.
- Harding, S., and O'Barr, J. 1987. *Sex and Scientific Inquiry*. Chicago: University of Chicago Press.
- Hartsock, N., 1987. "The Feminist Standpoint" in S. Harding, ed. *Feminism and Methodology*. Bloomington and Indianapolis: Indiana University Press.
- Haslanger, S., 1993. "On Being Objective and Being Objectified" in Antony and Witt, eds. *A Mind of One's Own*. Boulder: Westview Press.
- Hempel, C., 1942. "The Function of General Laws in History." *Journal of Philosophy* 39: 35-48.
- -----, 1965. *The Philosophy of Natural Science*. Englewood Cliffs, New Jersey: Prentice-Hall.
- Hempel, C. and Oppenheim, P., 1948. "Studies in the Logic of Explanation." *Philosophy of Science* 15: 135-75.
- Hoyningen-Huene, P., 1993. *Reconstructing Scientific Revolutions*: *Thomas S. Kuhn's Philosophy of Science*. Chicago: University of Chicago Press.

- Hoyningen-Huene, P. and Sankey, H. eds. 2001. *Incommensurability and Related Matters*. Dordrecht: Kluwer.
- Keller, E., 1983. "Gender and Science" in Sandra Harding and Merrill B. Hintikka, eds. *Discovering Reality*. Dordrecht: D. Reidel.
- Kitcher, P., 1981. "Explanatory Unification." *Philosophy of Science* 48, 507-531.
- -----, 1982. *Abusing Science*. Cambridge: MIT press. Theories *of Truth*: *A Critical Introduction*. Cambridge: MIT Press.
- -----, 1993. *The Advancement of Science* New York: Oxford University Press.
- Kitcher, P., and Salmom, W., 1989. eds. *Scientific Explanation*. Minneapolis: University of Minnesota Press.
- Knorr Cetina, Karin, 1993. "Strong Constructivism -- from a Sociologist’s Point of View: A Personal Addendum to Sismondo’s Paper," *Social Studies of Science* 23, 555-63.
- Kripke, S., 1971. "Identity and Necessity.’’ in M.K. Munitz (ed.) *Identity and Individuation*. New York: New York University Press.
- -----, 1972. ""*Naming and Necessity*.’’ in D. Davidson and G. Harman (eds.) *The Semantics of Natural Language*. Dordrecht: D. Reidel.
- Kuhn, T., 1970. *The Structure of Scientific Revolutions*, 2nd edition. Chicago: University of Chicago Press.
- Latour, B., and Wolgar, S., 1986 *Laboratory Life*: *The Social Construction of Scientific Facts*. Princeton: Princeton University Press.
- -----, 1987. *Science in Action*. Cambridge: Harvard University Press.
- Laudan, L., 1981. "A Confutation of Convergent Realism" *Philosophy of Science* 48: 218-249.
- ------, 1984. "Discussion: Realism without the Real." *Philosophy of Science* 51: 156-62.
- Lipton, P., 1991. *Inference to the Best Explanation*. London: Routledge and Kegan Paul.
- -----, 1993. "Is the Best Good Enough?" *Proceedings of the Aristotelian Society* 93/2: 89-104.
- Locke, J., 1689/1975. *An Essay Concerning Human Understanding*. Oxford: Oxford University Press.
- Longino, H., 1989 "Can there be a Feminist Science?" in N. Tuana, ed. *Feminism and Science*. Bloomington and Indianapolis: Indiana University Press.
- -----, 1990. *Science as Social Knowledge*. Princeton: Princeton University press.
- Maxwell, Grover, 1962. "On the Ontological Status of Theoretical Entities" in Feigl, Herbert and Maxwell, Grover, eds. *Scientific Explanation*, *Space*, *and Time*, *Minnesota Studies in the Philosophy of Science*, *Volume III* (1962). Minneapolis: University of Minnesota Press, 3-27.
- McMullin, E., 1984. "A Case for Scientific Realism." in J. Leplin (ed.) *Scientific Realism*. Berkeley: University of California Press.
- -----, 1987. "Explanatory Successes and the Theory of Truth." in N. Rescher, ed., *Scientific Inquiry in Philosophical Perspective*. Lanham: University Press of America.
- -----, 1991. "Comment: Selective Anti-Realism." *Philosophical Studies* 61: 97-1080
- Miller, R., 1987. *Fact and Method*. Princeton: Princeton University Press.
- Millikan, R., 1986. "Metaphysical Anti-Realism?" *Mind* 95: 417-431.
- Newton-Smith, W.H., 1989a. "Modest Realism." in A. Fine and J. Leplin, eds. *PSA 1988*, vol. 2. East Lansing: Philosophy of Science Association.
- -----, 1989b. "The Truth in Realism." *Dialectica* 43: 31-45.

- Niiniluoto, I., 1987. *Truthlikeness*. Dordrecht: Reidel.
- Oddie, G., 1986. *Likeness to Truth*. Dordrecht: Reidel.
- Papineau, D., 1987. *Reality and Representation*. Oxford: Blackwell.
- -----, 1988. "Does the Sociology of Science Discredit Science?" In R. Nola, ed. *Relativism and Realism in Science*. Dordrecht: Kluwer.
- -----, 1993. *Philosophical Naturalism*. Oxford: Blackwell.
- -----, 1996. *The Philosophy of Science*. Oxford: Oxford University Press.
- Pettit, P., 1988. "The Strong Sociology of Knowledge without Relativism." In R. Nola, ed. *Relativism and Realism in Science*. Dordrecht: Kluwer.
- Pickering, A., 1984. *Constructing Quarks*. Edinburgh: Edinburgh University Press.
- -----, 1995. *The Mangle of Practice*. Chicago: University of Chicago Press.
- Pinch, Trevor, 1985. "Towards an Analysis of Scientific Observation: The Externality and Evidential Significance of Observational Reports in Physics", *Social Studies of Science* 15: 3-36.
- Psillos, S., 1995. "Is Structural Realism the Best of Both Worlds?" *Dialectica* 49: 15-64.
- -----, 1999. *Scientific Realism*: *How Science Tracks Truth*. New York and London: Routledge.
- Putnam, H., 1972. "Explanation and Reference." in G. Pearce and P. Maynard, eds. *Conceptual Change*. Dordrecht: Reidel.
- -----, 1975a. "The Meaning of "Meaning'." in H. Putnam, *Mind*, *Language and Reality*. Cambridge: Cambridge University Press.
- -----, 1975b. "Language and Reality." in H. Putnam, *Mind*, *Language and Reality*. Cambridge: Cambridge University Press. University Press.
- Putnam, H. 1978. *Meaning and the Moral Sciences*. London: Routledge and Kegan Paul.
- -----, 1981. *Reason*, *Truth and History*. Cambridge: Cambridge University Press.
- -----, 1983. ""Vagueness and Alternative Logic.'' in H. Putnam, *Realism and Reason*. Cambridge: Cambridge University Press.
- -----, 1983a. "Why There Isn't a Ready Made World" in H. Putnam, *Realism and Reason*. Cambridge: Cambridge University Press.
- Quine, W.V.O., 1969. ""Natural Kinds.'' in W.V.0. Quine, *Ontological Relativity and Other Essays*. New York: Columbia University Press.
- Salmon, W., 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
- -----, 1989. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.
- Scheffler, I., 1967. *Science and Subjectivity*. Indianapolis: Hackett.
- Shapere, D., 1964. "The Structure of Scientific Revolutions," *Philosophical Review*, LXXIII, 383-94.
- Shapin, S., 1982. "History of Science and its Sociological Reconstructions" *History of Science* xx.
- -----, 1994. *A Social History of Truth*: *Civility and Science in Seventeenth-Century England*. Chicago: University of Chicago Press.
- Shapin, S., and Schaffer, S., 1985. *Leviathan and the Air Pump*. Princeton: Princeton University Press.
- Shoemaker, S., 1980. "Causality and Properties.'' in P. van Inwagen (ed.) *Time and Cause*. Dordrecht: D. Reidel.
- Sismondo, S., 1993a. "Some Social Constructions," *Social Studies of Science* 23, 515-53.

- -----, 1993b. "Response to Knorr Cetina," *Social Studies of Science* 23, 563-69.
- -----, 1996. *Science without Myth*. Albany: State University of New York Press.
- Tuana, N., 1989 *Feminism and Science*. Bloomington and Indianapolis: Indiana University Press.
- van Fraassen, B., 1980. *The Scientific Image*. Oxford: Oxford University Press.
- Weston, T., 1992. "Approximate Truth and Scientific Realism." *Philosophy of Science* 59: 53-74.
- Wilson, R., 1999a. *Species*: *New Interdisciplinary Essays*. Cambridge: MIT Press.
- -----, 1999b. "Realism, Essence, and Kind: Resuscitating Species Essentialism." in R. Wilson, ed. *Species*: *New Interdisciplinary Essays*. Cambridge: MIT Press.
- Worrall, J., 1994. "How to Remain (Reasonably) Optimistic: Scientific Realism and the 'Luminiferous Ether'." In D. Hull and M. Forbes, eds. *PSA 1994*, vol 1: 334-44. East Lansing: Philosophy of Science Association.
- Wright, R., 1996. *Gender and Archaeology*. Philadelphia: University of Pennsylvania Press.
- Wylie, A., 1986. "Arguments for Scientific Realism: The Ascending Spiral." *American Philosophical Quarterly* 23: 287-97.
- -----, 1991. "Gender Theory and the Archaeological Record: Why is there no Archaeology of Gender?" in Gero and Conkey, eds. *Engendering Archaeology*. Oxford: Basil Blackwell.
- -----, 2000. "Feminism in Philosophy of Science: Making Sense of Contingency and Constraint." in Fricker, Miranda and Hornsby, Jennifer, eds. *The Cambridge Companion to Feminism in Philosophy*. Cambridge: Cambridge University Press.
- -----, 1993. "Gender Archaeology/Feminist Archaeology" in W. A. Bacus, *et al*, eds. *A Gendered Past*. Ann Arbor: University of Michigan Technical Report No. 25: vii-xiii.
- Wylie, A. and Okruhlik, K. 1987. "Philosophical Feminism: Challenges to Science." *Resources for Feminist Research* 16 (3): 12-16.

Other Internet Resources

- Preprints in the Philosophy of Science (sponsored by thr Philosophy of Science Association and the Center for Philosophy of Science, University of Pittsburgh)
- Links to Philosophy (of Science) Resources (Johns Hopkins University Philosophy Department)
- A Guide to Resources in Science Studies (Science Studies Program at the University of Missouri/Kansas City)
- The Science Wars Homepage (sponsored by the Science Wars Monitor International; contains references to the controversies around "social constructivism" in science studies)

[Note: Many of the authors mentioned in this entry who are still active have websites which often contain information about their most recent work.]

Related Entries

Feyerabend, Paul | Kuhn, Thomas | scientific explanation | scientific knowledge: social dimensions of

Richard Boyd
rnb1@cornell.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 12, 2002
Content last modified: June 12, 2002

Social Dimensions of Scientific Knowledge

Study of the social dimensions of scientific knowledge encompasses the effects of scientific research on human life and social relations, the effects of social relations and values on scientific research, and the social aspects of inquiry itself. Several factors have combined to make these questions salient to contemporary philosophy of science. These factors include the emergence of social movements, like environmentalism and feminism, critical of mainstream science; concerns about the social effects of science-based technologies; epistemological questions made salient by big science; new trends in the history of science, especially the move away from internalist historiography; anti-normative approaches in the sociology of science; turns in philosophy to naturalism and pragmatism. This entry reviews the historical background to current research in this area, features of contemporary science which invite philosophical attention, the challenge to normative philosophy from social, cultural, and feminist studies of science, and the principal philosophical models of the social character of scientific knowledge.

---

1. Historical Background

Philosophers who study the social character of scientific knowledge can trace their lineage at least as far as John Stuart Mill. Mill, Charles Sanders Peirce, and Karl Popper all took some type of critical interaction as central to the validation of knowledge claims.

Mill's arguments occur in his well-known political essay On Liberty, (Mill 1859) rather than in the

context of his logical and methodological writings, but he makes it clear that they are to apply to any kind of knowledge or truth claim. Mill argues from the fallibility of human knowers to the necessity of unobstructed opportunity for and practice of the critical discussion of ideas. Only such critical discussion can assure us of the justifiability of the (true) beliefs we do have and can help us avoid falsity or the partiality of belief or opinion framed in the context of just one point of view. The achievement of knowledge, then, is a social or collective, not an individual, matter.

Peirce's contribution to the social epistemology of science is commonly taken to be his consensual theory of truth: "The opinion which is fated to be ultimately agreed to by all who investigate is what we mean by truth, and the object represented is the real." (Peirce 1878, 133) While often read as meaning that the truth is whatever the community of inquirers converges on in the long run, the notion in turn is interpretable as meaning more precisely either that truth (and "the real") depends on the agreement of the community of inquirers or that it is the effect of the real that it will in the end produce agreement among inquirers. Whatever the correct reading of this particular statement, Peirce elsewhere makes it clear that, in his view, truth is both attainable and beyond the reach of any individual. "We individually cannot hope to attain the ultimate philosophy which we pursue; we can only seek it for the community of philosophers." (Peirce 1868, 40). Peirce puts great stock in instigating doubt and critical interaction as means to knowledge. Thus, whether his theory of truth is consensual or realist, his view of the practices by which we attain it grants a central place to dialogue and social interaction.

Popper is often treated as a precursor of social epistemology because of his emphasis on the importance of criticism in the development of scientific knowledge. Two concepts of criticism are found in his works (Popper 1963, 1972) and these can be related to logical and practical senses of falsification. The logical sense of falsification is just the structure of a modus tollens argument, in which a hypothesis is falsified by the demonstration that one of its logical consequences is false. This is one notion of criticism, but it is a matter of formal relations between statements. The practical sense of falsification refers to the efforts of scientists to demonstrate the inadequacies of one another's theories by demonstrating observational shortcomings or conceptual inconsistencies. This is a social activity. For Popper the methodology of science is falsificationist, and science progresses through the demonstration by falsification of the untenability of theories and hypotheses. Popper's falsificationism is part of an effort to demarcate genuine science from pseudo science, and has lost its plausibility as a description of scientific methodology as the demarcation project has come under challenge from naturalist and historicist approaches in philosophy of science. While criticism does play an important role in some current approaches in social epistemology, Popper's own views are more closely approximated by evolutionary epistemology, especially that version that treats cognitive progress as the effect of selection against incorrect theories and hypotheses.

The work of Mill, Peirce, and Popper is a resource for philosophers presently exploring the social dimensions of scientific knowledge. However, the current debates are framed in the context of developments in both philosophy of science and in history and social studies of science following the collapse of the logical empiricist consensus. The philosophers of the Vienna Circle are conventionally associated with an uncritical form of positivism and with the logical empiricism that replaced American pragmatism in the 1940s and 1950s. They saw natural science, however, as a potent force for progressive

social change. (Cartwright, Cat, and Chang 1996; Giere and Richardson, eds., 1996) With its grounding in observation and public forms of verification, science for them constituted a superior alternative to what they saw as metaphysical obscurantism, an obscurantism that led not only to bad thinking but to bad politics. While one development of this point of view leads to scientism, the view that any meaningful question can be answered by the methods of science; another development leads to inquiry into what social conditions promote the growth of scientific knowledge. Logical empiricism, the version of Vienna Circle philosophy that developed in the United States, focused on logical, internal aspects of scientific knowledge and discouraged philosophical inquiry into the social dimensions of science. These came into prominence again after the publication of Thomas Kuhn's *Structure of Scientific Revolutions* (Kuhn 1962). A new generation of sociologists of science took Kuhn's emphasis on the role of non-evidential community factors in scientific change even further than he had and argued that scientific judgment was determined by social factors, such as professional interests and political ideologies. This family of positions has provoked a counter-response among philosophers. These responses are marked by an effort to grant some social character to scientific knowledge while at the same time maintaining its epistemological legitimacy, which they take to be undermined by the new sociology. At the same time, features of the organization of scientific inquiry compel philosophers to consider their implications for the normative analysis of scientific practices.

2. Big Science, Trust, and Authority

The second half of the twentieth century saw the emergence of what has come to be known as Big Science: the organization of large numbers of scientists bringing different bodies of expertise to a common research project. The original model was the Manhattan Project, undertaken during the Second World War to develop an atomic weapon. Theoretical and experimental physicists located at various sites across the country, though principally at Los Alamos, New Mexico, worked on sub-problems of the project under the overall direction of J. Robert Oppenheimer. While academic and military research have since been to some degree separated, much experimental research in physics, especially high energy particle physics, continues to be pursued by large teams of researchers. Research in other areas of science as well, for example the work comprehended under the umbrella of the Human Genome Project, has taken on some of the properties of Big Science, requiring multiple forms of expertise. In addition, the dependence of research on central funding bodies prompts questions about the degree of independence of contemporary scientific knowledge from its social and economic context.

John Hardwig (1985) articulated one philosophical dilemma posed by such large teams of researchers. Each member or subgroup participating in such a project is required because each has a crucial bit of expertise not possessed by any other member or subgroup. This may be knowledge of a part of the instrumentation, the ability to perform a certain kind of calculation, the ability to make a certain kind of measurement or observation. The other members are not in a position to evaluate the results of other members' work, and hence, all must take one anothers' results on trust. The consequence is an experimental result, (for example, the measurement of a property such as the decay rate or spin of a given particle) the evidence for which is not fully understood by any single participant in the experiment. This leads Hardwig to ask two questions, one about the evidential status of testimony, and one about the

nature of the knowing subject in these cases. With respect to the latter, Hardwig says that either the group as a whole, but no single member, knows or it is possible to know vicariously. Neither of these is palatable to him. Talking about the group or the community knowing smacks of superorganisms and transcendent entities and Hardwig shrinks from that solution. Vicarious knowledge, knowing without oneself possessing the evidence for the truth of what one knows, requires, according to Hardwig, too much of a departure from our ordinary concepts of knowledge.

The first question is, as Hardwig notes, part of a more general discussion about the epistemic value of testimony. Much of what passes for common knowledge is acquired from others. We depend on experts to tell us what is wrong with our appliances, our cars, our bodies. Indeed, much of what we later come to know depends on what we previously learned as children from our parents. We acquire knowledge of the world through the institutions of education, journalism, and scientific inquiry. Philosophers disagree about the status of beliefs acquired in this way. Here is the question: If A knows that p on the basis of evidence e, B has reason to think A trustworthy and B believes p on the basis of A's testimony that p, does B know that p? Some philosophers, like Locke and Hume, argued that only what one has observed oneself could count as a good reason for belief, and that the testimony of another is, therefore, never sufficient warrant for belief. Thus, B does not know simply on the basis of A's testimony. While this result is consistent with traditional philosophical empiricism and rationalism, which emphasized the individual's sense experience or rational apprehension as foundations of knowledge, it does have the consequence that we do not know most of what we think we know.

A number of philosophers have recently offered alternative analyses focusing on one or another element in the problem. Some argue that testimony by a qualified expert is itself evidential, (Schmitt 1988), others that the expert's evidence constitutes good reason for, but is not itself evidential for the recipient of testimony (Hardwig 1985, 1988), others that what is transmitted in testimony is knowledge and not just propositional content and thus the question of the kind of reason a recipient of testimony has is not to the point (Welbourne 1981).

However this dispute is resolved, questions of trust and authority arise in a particularly pointed way in the sciences, and Hardwig's dilemma for the physics experiment is also a specific version of a more general phenomenon. A popular conception of science, fed partly by Popper's falsificationism, is that it is epistemically reliable because the results of experiments and studies are checked by independent repetition. In practice, however, only some results are so checked and many are simply accepted on trust. Thus, just as in the non-scientific world information is accepted on trust, so in science, knowledge grows by depending on the testimony of others. What are the implications of accepting this fact for our conceptions of the reliability of scientific knowledge?

David Hull, in his (1988) argues that because the overall structure of reward and punishment in the sciences is a powerful incentive not to cheat, further epistemological analysis of the sciences is unnecessary. But some celebrated recent episodes, such as the purported production of "cold fusion" were characterized by the failure of replication attempts to produce the same phenomenon. And, while the advocates of cold fusion were convinced that their experiments had produced the phenomenon, there have also been cases of outright fraud. Thus, even if the structure of reward and punishment is an

incentive not to cheat, it does not guarantee the veridicality of every research report.

The reward individual scientists seek is credit. That is, they seek recognition, to have their work cited as important and as necessary to further scientific progress. The scientific community seeks true theories or adequate models. Credit, or recognition, accrues to individuals to the extent they are perceived as having contributed to that community goal. Without strong community policing structures, there is a strong incentive to cheat, to try to obtain credit without necessarily having done the work. Communities and individuals are then faced with the question: when is it appropriate to trust and when not?

Both Alvin Goldman (Goldman and Cox 1994) and Philip Kitcher (1993) treat this as a question to be answered by means of decision theoretic models. The decision theoretic approach to problems of trust and authority treats both credit and truth as utilities. The challenge then is to devise formulas that show that actions designed to maximize credit also maximize truth. Kitcher, in particular, develops formulas intended to show that even in situations peopled by non-epistemically motivated individuals (that is, individuals motivated more by a desire for credit than by a desire for truth), the reward structure of the community can be organized in such a way as to foster scientific progress. Kitcher also applies this approach to problems in the division of cognitive labor, i.e. to the questions whether (and when) to pursue research that calls a community consensus into question or to pursue research that extends the models and theories upon which a community agrees.

Steve Fuller and Joseph Rouse are both concerned with political dimensions of cognitive authority. Rouse in his (1987) integrated analytic and continental philosophy of science and technology to develop what might be called a critical pragmatism. This perspective facilitated an analysis of the transformative impact of science on human life and social relations. Fuller (1988) partially accepts the empirical sociologists' claim that traditional normative accounts of scientific knowledge fail to get a purchase on actual scientific practices, but takes this as a challenge to relocate the normative concerns of philosophers. These should include the distribution and circulation of knowledge claims. The task of social epistemology of science is regulation of the production of knowledge by regulating the rhetorical, technological, and administrative means of its communication.

3. Social, Cultural, and Feminist Studies of Science

Kuhn's critique of logical empiricism included a strong naturalism. Scientific rationality was to be understood by studying actual episodes in the history of science, not by formal analyses developed from a priori concepts of knowledge and reason (Kuhn 1962, 1977). Sociologists and sociologically inclined historians of science took this as a mandate for the examination of the full spectrum of scientists' practices without any prior prejudice as to which were epistemically legitimate and which not. That very distinction came under suspicion from the new social scholars, often labeled "social constructivists." They urged that understanding the production of scientific knowledge required looking at all the factors causally relevant to the acceptance of a scientific idea, not just at those the researcher thinks should be relevant.

A wide range of approaches in social and cultural studies of science has come under the umbrella label of "social constructivism." Both terms in the label are understood differently in different programs of research. While constructivists agree in holding that those factors treated as evidential, or as rationally justifying acceptance, should not be privileged at the expense of other causally relevant factors, they differ in their view of which factors are causal or worth examination. Macro-analytic approaches, such as those associated with the so-called Strong Programme in the Sociology of Scientific Knowledge, treat social relations as an external, independent variable and scientific judgment and content as a dependent variable. Micro-analyses or laboratory studies, on the other hand, abjure the implied separation of social context and scientific practice and focus on the social relations within scientific research programs and communities and on those that bind research-productive and research-receptive communities together.

Researchers also differ in the degree to which they treat the social and the cognitive dimensions of inquiry as independent or interactive. The researchers associated with the macro-analytic Strong Programme in the Sociology of Scientific Knowledge (Barry Barnes, David Bloor, Harry Collins, Donald MacKenzie, Andrew Pickering, Steve Shapin) were particularly interested in the role of large scale social phenomena, whether widely held social/political ideologies or group professional interests, on the settlement of scientific controversies. Some landmark studies in this genre include Andrew Pickering's (1984) study of competing professional interests in the interpretation of high energy particle physics experiments, and Steven Shapin and Simon Shaffer's (1985) study of the controversy between Robert Boyle and Thomas Hobbes about the proper interpretation of experiments with vacuum pumps.

The micro-sociological or laboratory studies approach features ethnographic study of particular research groups, tracing the myriad activities and interactions that eventuate in the production and acceptance of a scientific fact or datum. Karin Knorr Cetina's (1981) reports her year-long study of a plant science laboratory at UC Berkeley. Bruno Latour and Steven Woolgar's (1986) study of Roger Guillemin's neuroendocrinology laboratory at the Salk Institute is another classic in this genre. These scholars argued in subsequent work that their form of study showed that philosophical analyses of rationality, of evidence, of truth and knowledge, were irrelevant to understanding scientific knowledge. Sharon Traweek's (1988) comparative study of the cultures of Japanese and North American high energy physics communities pointed to the parallels between cosmology and social organization without making such extravagant and provocative epistemological claims. The efforts of philosophers of science to articulate norms of scientific reasoning and judgment were, to all these scholars, misdirected, because actual scientists relied on quite different kinds of considerations in the practice of science.

Until recently, apart from a few anomalous figures like Caroline Herschel, Barbara Mcclintock, and Marie Curie, the sciences were a male preserve. Feminist scholars have asked what bearing the masculinity of the scientific profession has had on the content of science and on conceptions of scientific knowledge and practice. Drawing on work by feminist scientists, exposing and critiquing gender biased science, and on theories of gender, feminist historians and philosophers of science have offered a variety of models of scientific knowledge and reasoning intended to accommodate the critique of accepted science and the concomitant proposal and advocacy of alternatives. Evelyn Keller (1985) proposed a psycho-dynamic model of knowledge and objectivity, arguing that a certain psychological profile, facilitated by typical patterns of masculine psychological development, associated knowledge and

objectivity with domination. The association of knowledge and control continues to be a topic of concern for feminist thinkers as it is also for environmentally concerned critics of the sciences. Other feminists turned to Marxist models of social relations and developed versions of standpoint theory, which holds that the beliefs held by a group reflect the social interests of that group. As a consequence, the scientific theories accepted in a context marked by divisions of power such as gender will reflect the interests of those in power. Alternative theoretical perspectives can be expected from those systematically excluded from power. (Rose 1983; Haraway 1978).

Still other feminists have argued that some standard philosophical approaches to the sciences can be used to express feminist concerns. Nelson (1990) adopts Quine's holism and naturalism to analyze debates in recent biology. Elizabeth Potter (2001) adapts Mary Hesse's network theory of scientific inference to analyse gendered aspects of 17th century physics. Helen Longino (1990) develops a contextual empiricism to analyze research in human evolution and in neuroendocrinology. In addition to the direct role played by gender bias, scholars have attended to the ways shared values in the context of reception can confer an a priori implausibility on certain ideas. Keller (1983) argued that this was the fate of Barbara McClintock's unorthodox proposals of genetic transposition. Stephen Kellert (1993) makes a similar suggestion regarding the resistance to so-called chaos theory.

What the feminist and empirical sociological analyses have in common is the view that the social organization of the scientific community has a bearing on the knowledge produced by that community. There are deep differences, however, in their views as to what features of that social organization are deemed relevant and how they are expressed in the theories and models accepted by a given community. The gender relations focused on by feminists went unrecognized by sociologists pursuing macro- or microsociological research programs. The feminist scientists and scholars further differ from the scholars in empirical social and cultural studies of science in their call for alternative theories and approaches in the sciences. These calls imply that philosophical concerns with truth and justification are not only legitimate but useful tools in advancing feminist transformative goals for the sciences. As can be seen in their varying treatments of objectivity, however, philosophical concepts are often reworked in order to be made applicable to the content or episodes of interest (Cf. Haraway 1988, Harding 1993, Keller 1985, Longino 1990, Nelson 1990)

4. Models of the Social Character of Knowledge

Since 1980, interest in developing philosophical accounts of scientific knowledge that incorporate the social dimensions of scientific practice has been on the increase. Some philosophers see attention to the social as a straightforward extension of already developed approaches in epistemology. Others, inclined toward some form of naturalism, have taken the work in empirical social studies of science discussed above seriously. They have, however, diverged quite considerably in their treatment of the social. Some, understand the social as biasing or distorting, and hence see the social as opposed to or competing with the cognitive or epistemic. They see the sociologists' disdain for normative philosophical concerns as part of a general debunking of science that demands a response. They attempt either to rebut the claims of the sociologists or to reconcile the demonstration of the role of interests in science with its ultimate

rationality. Others treat the social as instead constitutive of rationality. This division parallels to some degree the division between macro-analyses and micro-analyses in the sociology of science described above.

Philosophers who treat the social as biasing or distorting tend to focus on the constructivists' view that there are no universal principles of rationality or principles of evidence that can be used to identify in any context-independent way which factors are evidential and which not. They can be divided into roughly two camps: defenders of rationality and reconciliationists who seek to disarm the sociologists' analyses by incorporating them into a broader rational framework.

Philosophers concerned to defend the rationality of science against sociological misrepresentations include Larry Laudan (1984) James Brown (1989, 1994), Alvin Goldman (1987, 1995) and Susan Haack (1996). The details of these philosophers' approaches differ, but they agree in holding that scientists are persuaded by what they regard as the best evidence or argument, the evidence most indicative of the truth by their lights, and in holding that arguments and evidence are the appropriate focus of attention for understanding the production of scientific knowledge. When evidential considerations have not trumped non-evidential considerations, we have an instance of bad science. They read the sociologists as arguing that a principled distinction between evidential and nonevidential considerations cannot be drawn and devote their efforts to refuting those arguments. The social character of science is understood as a matter of the aggregation of individuals, not their interactions, and public knowledge as simply the additive outcome of many individuals making sound epistemic judgments. Individual rationality and individual knowledge are thus the proper focus of philosophers of science. Exhibiting principles of rationality applicable to individual reasoning is sufficient to demonstrate the rationality of science, at least in its ideal form.

Reconciliationists include Ronald Giere, Mary Hesse, and Philip Kitcher. Giere (1988) models scientific judgment using decision theory. This permits incorporating scientists' interests as one of the parameters of the decision matrix. Mary Hesse (1980) employs a network model of scientific inference that resembles W.V.O. Quine's web of belief in that its constituents are heterogeneous in character, but all subject to revision in relation to changes elsewhere in the network. She understands the social factors as coherence conditions operating in tandem with logical constraints to determine the relative plausibility of beliefs in the network.

The most elaborate reconciliationist position is that developed in Philip Kitcher's (1993). In addition to modeling relations of authority and the division of cognitive labor as described above, he offers what he terms a compromise between extreme rationalists and sociological debunkers. The compromise model appeals to a principle of rationality, which Kitcher calls the External Standard. It is deemed external because it is proposed as holding independently of any particular historical, cultural or social context. Thus, not only is it is external, but it is also universal. The principle applies to change of belief (or shift from one practice to another, in Kitcher's broader locution), not to belief. It treats a shift (in practice or belief) as rational if and only "the process through which the shift was made has a success ratio at least as high as that of any other process used by human beings (ever) ..." (Kitcher 1993, 303). Kitcher's compromise proposes that scientific ideas develop over time and benefit from the contributions of many

differently motivated researchers. This is the concession to the sociologically oriented scholars. In the end, however, those theories that get accepted are those that satisfy Kitcher's External Standard. Kitcher thus joins Goldman, Haack, and Laudan in the view that it is possible to articulate a priori conditions of rationality or of epistemic warrant that operate independently of, or, perhaps one might say, orthogonally to, the social relations of science.

A third set of models is integrationist in character. Nelson (1990) uses Quine's arguments against the independently foundational status of observation statements as the basis for what she calls a feminist empiricism. According to Nelson, no principled distinction can be made between the theories, observations, or values of a community. What counts as evidence, in her view, is fixed by the entire complex of a community's theories, value commitments, and observations. There is neither knowledge nor evidence apart from such a shared complex. The community is the primary knower on this view and individual knowledge is dependent on the knowledge and values of the community.

Miriam Solomon's social empiricism is focused on scientific rationality (Solomon 1992, 1994a, 1994b). It, too, involves denying a universal principled distinction among the causes of belief. Solomon draws on contemporary cognitive science literature to argue that biases are simply any factors that influence belief. They are not necessarily distorting, and can be productive of insight and rational belief. Salience and availability (of data, of measurement technologies) are biases as much as social ideologies. The distinctive feature of Solomon's social empiricism is her contrast between individual and community rationality. The theory or belief that it is rational to accept is that which has the greatest amount of empirical success. Individuals can persist in beliefs that are less rational than others on this view, if the totality of available evidence (or empirical data) is not available to them. What matters to science, however, is that community judgments be rational. A community is rational when the theories it accepts are those with all or the most empirical successes. Thus, the community can be rational even when its members are irrational. Indeed, individual irrationality can contribute to community rationality in that individuals committed to a theory that accounts for their data keep that data in the range of phenomena any theory accepted by the entire community must eventually explain. In order that the totality of relevant constraints on theory acceptance remain available to the entire community, biases must be appropriately distributed. Thus Solomon proposes appropriate distribution of biases as a normative condition on the structure of scientific communities.

Finally, in Longino's critical contextual empiricism, the cognitive processes that eventuate in scientific knowledge are themselves social (Longino 1990). Longino's starting point is a version of the underdetermination argument: the semantic gap between statements describing data and statements expressing hypotheses or theories to be confirmed or disconfirmed by that data means that evidential relations cannot be formally specified and that data cannot support one theory or hypothesis to the exclusion of all alternatives. Instead, such relations are mediated by background assumptions. Eventually, in the chain of justification, one reaches assumptions for which no evidence is available. If these are the context in which evidential relations are constituted, questions arise concerning how the acceptance of such assumptions can be legitimated. According to Longino, the only check against the arbitrary dominance of subjective (metaphysical, political, aesthetic) preference in such cases is critical interaction among the members of the scientific community or among members of different

communities. Longino takes the underdetermination argument to express in logical terms the point made by the sociologically oriented researchers: the individuals participating in the production of scientific knowledge are historically, geographically, and socially situated and their observations and reasoning reflect their situations. This fact does not undermine the normative enterprise of philosophy, but requires its expansion to include within its scope the social interactions within and between scientific communities. What counts as knowledge is determined by such interactions. Longino claims that scientific communities do institutionalize some critical practices (for example, peer review), but argues that such practices and institutions must satisfy conditions of effectiveness in order to qualify as objective.

5. Conclusion

Philosophical study of the social dimensions of scientific knowledge has been intensifying in the decades since 1970. Scholars continue to investigate the myriad social relations within scientific communities and between them and their social, economic, and institutional contexts. Philosophers have been focused on what might be termed narrowly epistemic concerns in their response to this work. As these stabilize, we can expect that this branch of philosophy of science will expand to include attention to the ethical and political questions its analyses make salient.

Bibliography

Works Cited

- Barnes, Barry and David Bloor. 1982. "Relativism, Rationalism, and the Sociology of Knowledge." In *Rationality and Relativism*, eds. Martin Hollis and Steven Lukes, pp.21-47. Oxford: B. Blackwell.
- Brown, James. 1989.*The Rational and the Social*. London: Routledge.
- -----. 1994. *Smoke and Mirrors: How Science Reflects Reality*.New York: Routledge.
- Cartwright, Nancy, Jordy Cat, and Hasok Chang. 1996. *Otto Neurath: Philosophy Between Science and Politics*. New York, NY: Cambridge University Press.
- Fuller, Steve. 1988. *Social Epistemology*.Bloomington, IN: Indiana University Press.
- Giere, Ronald, and Alan Richardson, eds. 1996. *Origins of Logical Empiricism, Minnesota Studies in the Philosophy of Science, Vol. XVI*. Minneapolis, MN: University of Minnesota Press.
- Giere, Ronald. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.
- Goldman, Alvin. 1987. “The Foundations of Social Epistemics.” *Synthese*, **73**, 1: 109-144.
- -----. 1995. “Psychological, Social and Epistemic Factors in the Theory of Science.” In *PSA 1994: Proceedings of the 1994 Biennial Meeting of the Philosophy of Science Association*, eds. Richard Burian, Mickey Forbes, and David Hull, pp. 277-286. East Lansing, MI: Philosophy of Science Association.
- Haack, Susan. 1996. “Science as Social: Yes and No.” In *Feminism, Science, and the Philosophy*

of Science, eds. Lynn Hankinson Nelson and Jack Nelson, pp. 79-94. Dordrecht, Holland: Kluwer Academic Publishers.

- Haraway, Donna. 1978. "Animal Sociology and a natural Economy of the Body Politic, Pt. II" *Signs* **4**,1:37-60.
- -----. 1988. "Situated Knowledges" *Feminist Studies* **14**, 3: 575-600.
- Harding, Sandra. 1993. "Rethinking Standpoint Epistemology" in Alcoff and Potter, eds., *Feminist Epistemologies*. New York, NY: Routledge. pp. 49-82.
- Hardwig, John. 1985. “Epistemic Dependence.” *Journal of Philosophy***82**, no. 7: 335-349.
- -----. 1988. "Evidence, Testimony, and the Problem of Individualism." *Social Epistemology***2**,4: 309-21.
- Hesse, Mary. 1980. *Revolutions and Reconstructions in the Philosophy of Science*. Bloomington, IN: Indiana University Press.
- Hull, David. 1988. *Science As a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. Chicago: University of Chicago Press.
- Keller, Evelyn Fox. 1983. *A Feeling for the Organism: The Life and Work of Barbara McClintock.*San Francisco: W.H. Freeman.
- -----. 1985. *Reflections on Gender and Science*. New Haven, CT: Yale University Press.
- Kellert, Stephen. 1993. *In the Wake of Chaos*. Chicago, IL: University of Chicago Press.
- Kitcher, Phillip. 1993. *The Advancement of Science: Science Without Legend, Objectivity Without Illusions.*Oxford: Oxford University Press.
- Knorr-Cetina, Karin. 1981. *The Manufacture of Knowledge*. Oxford: Pergamon Press.
- Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- -----. 1977. *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.
- Latour, Bruno and Steven Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts*. 2d ed. Princeton, NJ: Princeton University Press.
- Laudan, Larry. 1984a. “The Pseudo-Science of Science?” In *Scientific Rationality: The Sociological Turn*, ed. James Brown, pp. 41-74. Dordrecht, Holland: Reidel.
- Longino, Helen. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*.Princeton, NJ: Princeton University Press.
- Mill, John Stuart. 1859. *On Liberty*. London: John W. Parker and Son. Reprinted 1974, 1982, ed. by Gertrude Himmelfarb. Harmondsworth: Penguin.
- Nelson, Lynn Hankinson. 1990. *Who Knows: From Quine to Feminist Empiricism*. Philadelphia, PA: Temple University Press.
- Peirce, Charles S. 1868. “Some Consequences of Four Incapacities.” *Journal of Speculative Philosophy* **2**: 140-157. Reprinted in Peirce, Charles S. 1958. *Selected Writings*. ed. by Philip Wiener. New York: Dover Publications. pp. 39-72.
- -----. 1878. “How to Make Our Ideas Clear.” *Popular Science Monthly* **12**: 286-302. Reprinted in Peirce, Charles S. 1958. *Selected Writings*. ed. by Philip Wiener. New York: Dover Publications. pp. 114-136.
- Pickering, Andrew. 1984. *Constructing Quarks: A Sociological History of Particle Physics*. Edinburgh: Edinburgh University Press.

- Popper, Karl. 1963. *Conjectures and Refutations*. London: Routledge and Kegan Paul.
- -----. 1972. *Objective Knowledge*. Oxford: Oxford University Press.
- Potter, Elizabeth. 2001. *Gender and Boyle's Law of Gases*. Bloomington, IN: Indiana University Press.
- Rose, Hilary. 1983. "Hand, Brain, and Heart" *Signs* **9**,1:73-96.
- Rouse, Joseph. 1987. *Knowledge and Power: Toward a Political Philosophy of Science*. Ithaca, NY: Cornell University Press.
- Schmitt, Frederick. 1988. "On the Road to Social Epistemic Interdependence." *Social Epistemology* **2**: 297-307.
- Shapin, Steven and Simon Schaffer. 1985. *Leviathan and the Air Pump*.Princeton, NJ: Princeton University Press.
- Solomon. Miriam. 1992. "Scientific Rationality and Human Reasoning." *Philosophy of Science* **59**,3: 439-54.
- -----. 1994a. "Social Empiricism." *Nous* **28**, no.3: 323-343.
- -----. 1994b. "A More Social Epistemology." In *Socializing Epistemology: The Social Dimensions of Knowledge*, ed. Frederick Schmitt, pp. 217-233. Lanham, MD: Rowman and Littlefield Publishers.
- Traweek, Sharon. 1988. *Beamtimes and Lifetimes: The World of High Energy Physicists*. Cambridge: Harvard University Press.
- Welbourne, Michael. 1981. "The Community of Knowledge." *Philosophical Quarterly* **31**, no. 125: 302-314.

For Further Reading

- Goldman, Alvin. 1999. *Knowledge in a Social World*. New York, NY: Oxford University Press.
- Hacking, Ian. 1999. *The Social Construction of What?* Cambridge, MA. Harvard University Press.
- Kitcher, Philip. 2001. *Science, Truth, and Democracy*. New York, NY: Oxford University Press.
- Longino, Helen E. 2002. *The Fate of* Knowledge. Princeton, NJ: Princeton University Press.
- McMullin, Ernan, ed. 1992. *Social Dimensions of Scientific Knowledge*. South Bend, IN: Notre Dame University Press.
- Sismondo, Sergio. 1996. *Science Without Myth.* Albany, NY: State University of New York Press.
- Solomon, Miriam. 2001. *Social Empiricism*. Cambridge,MA: Massachusetts Institute of Technology Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

epistemology: evolutionary | epistemology: social | ethics: environmental | feminism, interventions: feminist epistemology and philosophy of science | Kuhn, Thomas | logical empiricism | Mill, John Stuart | naturalism | Peirce, Charles Sanders | Popper, Karl | pragmatism | rationality: historicist theories of

Helen E. Longino
hlongino@umn.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 12, 2002
Content last modified: April 12, 2002

Evolutionary Epistemology

Evolutionary Epistemology is a naturalistic approach to epistemology, which emphasizes the importance of natural selection in two primary roles. In the first role, selection is the generator and maintainer of the reliability of our senses and cognitive mechanisms, as well as the "fit" between those mechanisms and the world. In the second role, trial and error learning and the evolution of scientific theories are construed as selection processes.

1. History, Problems, and Issues

Traditional epistemology has its roots in Plato and the ancient skeptics. One strand emerges from Plato's interest in the problem of distinguishing between knowledge and true belief. His solution was to suggest that knowledge differs from true belief in being justified. Ancient skeptics complained that all attempts to provide any such justification were hopelessly flawed. Another strand emerges from the attempt to provide a reconstruction of human knowledge showing how the pieces of human knowledge fit together in a structure of mutual support. This project got its modern stamp from Descartes and comes in empiricist as well as rationalist versions which in turn can be given either a foundational or coherentist

twist. The two strands are woven together by a common theme. The bonds that hold the reconstruction of human knowledge together are the justificational and evidential relations which enable us to distinguish knowledge from true belief.

The traditional approach is predicated on the assumption that epistemological questions have to be answered in ways which do not presuppose any particular knowledge. The argument is that any such appeal would obviously be question begging. Such approaches may be appropriately labeled "transcendental."

The Darwinian revolution of the nineteenth century suggested an alternative approach first explored by Dewey and the pragmatists. Human beings, as the products of evolutionary development, are natural beings. Their capacities for knowledge and belief are also the products of a natural evolutionary development. As such, there is some reason to suspect that knowing, as a natural activity, could and should be treated and analyzed along lines compatible with its status, i. e., by the methods of natural science. On this view, there is no sharp division of labor between science and epistemology. In particular, the results of particular sciences such as evolutionary biology and psychology are not ruled *a priori* irrelevant to the solution of epistemological problems. Such approaches, in general, are called naturalistic epistemologies, whether they are directly motivated by evolutionary considerations or not. Those which are directly motivated by evolutionary considerations and which argue that the growth of knowledge follows the pattern of evolution in biology are called "evolutionary epistemologies."

Evolutionary epistemology is the attempt to address questions in the theory of knowledge from an evolutionary point of view. Evolutionary epistemology involves, in part, deploying models and metaphors drawn from evolutionary biology in the attempt to characterize and resolve issues arising in epistemology and conceptual change. As disciplines co-evolve, models are traded back and forth. Thus, evolutionary epistemology also involves attempts to understand how biological evolution proceeds by interpreting it through models drawn from our understanding of conceptual change and the development of theories. The term "evolutionary epistemology" was coined by Donald Campbell (1974).

1.1 The Evolution of Epistemological Mechanisms (EEM) versus The Evolutionary Epistemology of Theories (EET)

There are two interrelated but distinct programs which go by the name "evolutionary epistemology." One focuses on the development of cognitive mechanisms in animals and humans. This involves a straightforward extension of the biological theory of evolution to those aspects or traits of animals which are the biological substrates of cognitive activity, e. g., their brains, sensory systems, motor systems, etc. The other program attempts to account for the evolution of ideas, scientific theories, epistemic norms and culture in general by using models and metaphors drawn from evolutionary biology. Both programs have their roots in 19th century biology and social philosophy, in the work of Darwin, Spencer, James and others. There have been a number of attempts in the intervening years to develop the programs in detail (see Campbell 1974, Bradie 1986, Cziko 1995). Much of the contemporary work in evolutionary epistemology derives from the work of Konrad Lorenz (1977), Donald Campbell (1974, et. al.), Karl

Popper (1972, 1984) and Stephen Toulmin (1967, 1972).

The two programs have been labeled EEM and EET. (Bradie, 1986) EEM is the label for the program which attempts to provide an evolutionary account of the development of cognitive structures. EET is the label for the program which attempts to analyze the development of human knowledge and epistemological norms by appealing to relevant biological considerations. Some of these attempts involve analyzing the growth of human knowledge in terms of selectionist models and metaphors (e. g., Popper 1972, Toulmin 1972, Hull 1988). Others argue for a biological grounding of epistemological norms and methodologies but eschew *selectionist* models of the growth of human knowledge as such (e. g., Ruse 1986, Rescher 1990).

The EEM and EET programs are interconnected but distinct. A successful EEM selectionist explanation of the development of cognitive brain structures provides no warrant, in itself, for extrapolating such models to understand the development of human knowledge systems. Similarly, endorsing an EET selectionist account of how human knowledge systems grow does not, in itself, warrant concluding that specific or general brain structures involved in cognition are the result of natural selection for enhanced cognitive capacities. The two programs, though similar in design and drawing upon the same models and metaphors, do not stand or fall together.

1.2 Ontogeny versus Phylogeny

Biological development involves both ontogenetic and phylogenetic considerations. Thus, the development of specific traits, such as the opposable thumb in humans, can be viewed both from the point of view of the development of that trait in individual organisms (ontogeny) and the development of that trait in the human lineage (phylogeny). The development of knowledge and knowing mechanisms exhibits a parallel distinction. We can consider the growth of an individual’s corpus of knowledge and epistemological norms or his or her brain (ontogeny) or the growth of human knowledge and establishment of epistemological norms across generations or the development of brains in the human lineage (phylogeny). The EEM/EET distinction cuts across this distinction since we may be concerned either with the ontogenetic or phylogenetic development of, e. g., the brain or the ontogenetic or phylogenetic development of norms and knowledge corpora. One might expect that since current orthodoxy maintains that biological processes of ontogenesis proceed differently from the selectionist processes of phylogenesis, evolutionary epistemologies would reflect this difference. Curiously enough, however, for the most part they do not. For example, the theory of "neural Darwinism" as put forth by Edelman (1987) and Changeaux (1985) offers a selectionist account of the ontogenetic development of the neural structures of the brain. Karl Popper’s conjectures and refutations model of the development of human knowledge is a well known example of a selectionist account which has been applied both to the ontogenetic growth of knowledge in individuals as well as the trans-generational (phylogenetic) evolution of scientific knowledge. B. F. Skinner’s theory of operant conditioning, which deals with the ontogenesis of individual behavior, is explicitly based upon the Darwinian selectionist model. (Skinner 1981)

1.3 Descriptive versus Prescriptive Approaches

A third distinction concerns descriptive versus prescriptive approaches to epistemology and the growth of human knowledge. Traditionally, epistemology has been construed as a normative project whose aim is to clarify and defend conceptions of knowledge, foundations, evidential warrant and justification. Many have argued that neither the EEM programs nor the EET programs have anything at all to do with epistemology properly (i. e., traditionally) understood. The basis for this contention is that epistemology, properly understood, is a normative discipline, whereas the EEM and EET programs are concerned with the construction of causal and genetic (i.e., descriptive) models of the evolution of cognitive capacities or knowledge systems. No such models, it is alleged, can have anything important to contribute to normative epistemology (e.g., Kim 1988). The force of this complaint depends upon how one construes the relationship between evolutionary epistemology and the tradition.

There are three possible configurations of the relationship between descriptive and traditional epistemologies. (1) Descriptive epistemologies can be construed as competitors to traditional normative epistemologies. On this view, both are trying to address the same concerns and offering competing solutions. Riedl (1984) defends this position. A standard objection to such approaches is that descriptive accounts are not adequate to do justice to the prescriptive elements of normative methodologies. The extent to which an evolutionary approach contributes to the resolution of traditional epistemological and philosophical problems is a function of which approach one adopts (cf. Dretske 1971, Bradie 1986, Ruse 1986, Radnitsky and Bartley 1987, Kim 1988). (2) Descriptive epistemology might be seen as a successor discipline to traditional epistemology. On this reading, descriptive epistemology does not address the questions of traditional epistemology because it deems them irrelevant or unanswerable or uninteresting. Many defenders of naturalized epistemologies fall into this camp (e.g., Munz 1993). (3) Descriptive epistemology might be seen as complementary to traditional epistemology. This appears to be Campbell's view. On this analysis, the function of the evolutionary approach is to provide a descriptive account of knowing mechanisms while leaving the prescriptive aspects of epistemology to more traditional approaches. At best, the evolutionary analyses serve to rule out normative approaches which are either implausible or inconsistent with an evolutionary origin of human understanding.

1.4 Future Prospects

EEM programs are saddled with the typical uncertainties of phylogenetic reconstructions. Is this or that organ or structure an adaptation and if so, for what? In addition, there are the uncertainties which result from the necessarily sparse fossil record of brain and sensory organ development. The EET programs are even more problematic. While it is plausible enough to think that the evolutionary imprint on our organs of thought influences what and how we do think, it is not at all clear that the influence is direct, significant or detectible. Selectionist epistemologies which endorse a "trial and error" methodology as an appropriate model for understanding scientific change are not analytic consequences of accepting that the brain and other ancillary organs are adaptations which have evolved primarily under the influence of natural selection. The viability of such selectionist models is an empirical question which rests on the development of adequate models. Hull's (1988) is, as he himself admits, but the first step in that

direction. Cziko (1995) is a manifesto urging the development of such models (cf. also the evolutionary game theory modeling approach of Harms 1997). Much hard empirical work needs to be done to sustain this line of research. Non-selectionist evolutionary epistemologies, along the lines of Ruse (1986), face a different range of difficulties. It remains to be shown that any biological considerations are sufficiently restrictive to narrow down the range of potential methodologies in any meaningful way.

Nevertheless, the emergence in the latter quarter of the twentieth century of serious efforts to provide an evolutionary account of human understanding has potentially radical consequences. The application of selectionist models to the development of human knowledge, for example, creates an immediate tension. Standard traditional accounts of the emergence and growth of scientific knowledge see science as a progressive enterprise which, under the appropriate conditions of rational and free inquiry, generates a body of knowledge which progressively converges on the truth. Selectionist models of biological evolution, on the other hand, are generally construed to be non-progressive or, at most, locally so. Rather than generating convergence, biological evolution produces diversity. Popper's evolutionary epistemology attempts to embrace both but does so uneasily. Kuhn's "scientific revolutions" account draws tentatively upon a Darwinian model, but when criticized, Kuhn retreated. (cf Kuhn 1972, pp 172f with Lakatos and Musgrave 1970, p. 264) Toulmin (1972) is a noteworthy exception. On his account, concepts of rationality are purely "local" and are themselves subject to evolution. This, in turn, seems to entail the need to abandon any sense of "goal directedness" in scientific inquiry. This is a radical consequence which few have embraced. Pursuing an evolutionary approach to epistemology raises fundamental questions about the concepts of knowledge, truth, realism, justification and rationality.

The interested reader should consult the extensive bibliography originally developed by Donald Campbell and maintained by Gary Cziko at <http://faculty.ed.uiuc.edu/g-cziko/stb/>.

2. Formal Models

Every scientific enterprise requires formal and semi-formal models which allow the quantitative characterization of its objects of study. The attempt to transform the philosophical study of knowledge into a scientific discipline which approaches knowledge as a biological phenomenon is no different. Much of the evolutionary epistemology literature has been concerned with how to conceive of knowledge as a natural phenomenon, what difference this would make to our understanding of our place in the world, and with answering objections to the project. There are, as well, a number of more technical projects which attempt to provide the theoretical tools necessary for a naturalistic epistemology.

2.1 Static Optimization Models

In the simplest sort of model, an organism has to deal with an environment that has two states, S_1 and S_2, and has two possible responses R_1 and R_2. We suppose that what the organism does in each state makes a difference to its fitness. Fitnesses are usually written characterized by a matrix W.

The individual elements of the matrix W_{ij} are the fitness consequences of response *i* in state *j*. So, for instance, W_{21} denotes the fitness consequences of R_2 in S_1. If we let W_{11} and W_{22} equal one and W_{12} and W_{21} equal zero, then there is a clear evolutionary advantage to performing R_1 in S_1 and R_2 in S_2.

However, the organism must first detect the state of the environment, and detectors are not in general perfectly reliable. If the organism responds automatically to the detector, we can use the probabilities of responses given states to characterize the reliability of the detector. We write the probability of R_1 given S_1 as $\Pr(R_1|S_1)$. This allows us to calculate that responding to the detector rather than always choosing R_1 or R_2 will be advantageous just in case the following inequality holds (cf. Godfrey-Smith 1996):

$$\Pr(R_2|S_2) /(1 - \Pr(R_1|S_1)) > \Pr(S_1)(W_{11} - W_{21})/(1-\Pr(S_1))(W_{22} - W_{12})$$

This simple model demonstrates that whether or not flexible responses are adaptive depends on the particular characteristics of the fitness differences that the responses make, the probability of the various states of the environment, and the reliability of the detector. The particular result is calculated assuming that detecting the environmental state and the flexible response system is free in evolutionary terms. More complete analyses would include the costs of these factors.

Static optimization models like the one outlined above can be extended in several ways. Most obviously, the number of environmental states and organismic responses can be increased, but there are other modifications that are more interesting. Signal detection theory, for instance, models the detectors and cues in more detail. In one example, a species of "sea moss" detects the presence of predatory sea slugs via a chemical cue. They respond by growing spines, which is costly. The cue in this case, the water-borne chemical, comes in a variety of concentrations, which indicate various levels of danger. Signal detection theory allows us to calculate the best threshold value of the detector for the growing of spines.

Static models depict evolutionary processes in terms of fitness costs and benefits. They are static in the sense that they model no actual process, but merely calculate the direction of change for different situations. If fitness is high, a type will increase, if low it will decrease. When fitnesses are equal, population proportions remain at stable equilibrium. Dynamic models typically employ the kinds of calculations involved in static models to depict actual change over time in population proportions. Instead of calculating whether change will occur and in what direction, dynamic models follow change.

2.2 Population Dynamics

Population dynamics, sometimes referred to as "replicator dynamics", offers a tractable way to model the evolution of populations over time under the kinds of selective pressures that can be characterized by static optimization models. This is often necessary, since the dynamics of such populations are often difficult to predict purely on the basis of static considerations of payoff differences. The so-called "replicator dynamics" were named by Taylor and Jonker (1978) and generalized by Schuster and Sigmund (1983) and Hofbauer and Sigmund (1988). They trace their source back to the seminal work of

R.A. Fisher in the 1920's and 30's. The generalization covers evolutionary models used in population genetics, evolutionary game theory, ecology, and the study of prebiotic evolution. The models can be implemented either mathematically or computationally, and can model either stepwise (discrete) or continuous evolutionary change.

Population dynamics models the evolution of populations. A population is a collection of individuals, which are categorized according to type. The types in genetics are genes, in evolutionary game theory, strategies. The types of interest in epistemological models would be types of cognitive apparatuses, or cognitive strategies -- ways of responding to environmental cues, ways of manipulating representations, and so forth. Roughly, EEM models focus on the inherited and EET models focus on the learned. The evolution of the population consists in changes of the relative frequency of the different types within the population. Selection, typified by differential reproductive success, is represented as follows. Each type has a growth rate or "fitness", designated by *w*, and a frequency designated by *p*. The frequency of type *i* at the next generation p_i' is simply the old frequency multiplied by the fitness and divided by the mean fitness of the population "$\overline{w}$".

$$p_i' = p_i \cdot w_i / \overline{w}$$

Division by $\overline{w}$ has the effect of "normalizing" the frequencies, so that they add up to one after each is multiplied by its fitness. It also makes evident that the frequency of a type will increase just in case its fitness is higher than the current population average.

Fitness

Fitnesses, which should be understood simply as the aggregation of probable-growth factors that drive the dynamics of large populations, may depend on a variety of factors. Fitness components differ from variation components in that they affect population frequencies proportionally to those frequencies, that is to say, multiplicatively. Fitness component in biological evolution include mortality and reproductive rate. In cultural evolution, they include transmission probability and rejection probability. Within either sort of model, what matters is how fitnesses change as a result of other changing factors within the model. In the simplest cases, fitnesses are fixed and the type with the highest fitness inevitably dominates the population. In more complex cases, fitnesses may depend on variable factors like who one plays against, or the state of a variable environment. Most commonly, variable fitnesses are calculated using a payoff matrix like the one above. In general, to calculate the expected fitness of a type, one multiplies the fitness a type would have in each situation times the likelihood that individuals in the population will confront that situation and adds the resulting products.

$$w_i = \mathrm{S}_A \, \Pr(A) \cdot W_{iA}$$

where W_{iA} is type i's fitness in situation *A*. This sort of calculation assumes that the effects of the various situations are additive. More complex situations can be modeled, of course, but additive matrices are the standard. It should be noted, however, that matrix-driven evolution can exhibit quite complex behavior.

For instance, chaotic behavior is possible with as few as four strategies (Skyrms 1992).

Some relationships may be represented without a matrix. Boyd and Richerson (1985), for instance, were interested in a special kind of frequency dependent transmission bias in culture, where being common conferred an advantage due to imitators "doing as the Romans do." In such a case, the operative fitness of the type is just the fitness as calculated according to the usual factors, and then modified as a function of the frequency of the type.

Continuity and Computation

The conceptual bases of replicator dynamics are quite straightforward. Getting results typically requires one of two approaches. In order to prove more than rudimentary mathematical results, one typically needs to derive a continuous version of the dynamics. The basic form is

$$dp_i/dt = p(w_i - \overline{w})$$

with fitnesses calculated as usual. Mathematical approaches have been quite productive, though the bulk of theoretical results apply primarily to population genetics. See Hofbauer and Sigmund (1988) for a compendium of such results, as well as a reasonable graduate-level introduction to the mathematical study of evolutionary processes.

The second approach is computational. With the increase in power of personal computers, computational implementation of evolutionary models become increasingly attractive. They require only rudimentary programming skills, and are in general much more flexible in the assumptions they require. The general strategy is to create an array to hold population frequencies and fitnesses, and then a series of procedures (or methods or functions) which

1. calculate fitnesses,
2. update frequencies with the new fitnesses, and
3. manage interface details like outputting the new state of the population to a file or the screen.

A loop then runs the routines in sequence, over and over again. Most modelers are happy to put their source code on the internet, which is probably the best place to find it.

Modeling Cultural Evolution

Part of the difficulty in understanding cognitive behavior as the product of evolution is that there are at least three very different evolutionary processes involved. First, there is the biological evolution of cognitive and perceptual mechanisms via genetic inheritance. Second, there is the cultural evolution of languages and concepts. Third, there is the trial-and-error learning process that occurs during an individual’s lifetime. Moreover, there is some reason to agree with Donald T. Campbell that

understanding human knowledge fully will require understanding the interaction between these processes. This requires that we be able to model both processes of biological and cultural evolution. There are by now a number of well-established models of biological evolution. Cultural evolution presents more novelty.

Perhaps the most popular attempt to understand cultural evolution is Richard Dawkins' (1976) invention of the "meme." Dawkins observed that what lies at the heart of biological evolution is differential reproduction. Evolution in general was then the competitive dynamics of lineages of self-replicating entities. If culture was to evolve, on this view, there had to be cultural "replicators", or entities whose differential replication in culture constituted the cultural evolutionary process. Dawkins dubbed these entities "memes", and they were characterized as informational entities which infect our brains, "leaping from head to head" via what we ordinarily call imitation. Common examples include infectious tunes, and religious ideologies. The main difficulty with this approach has been with providing specifications for the basic entities. The identity conditions of genes can be given, in theory, in terms of sequences of base pairs in chromosomes. There appears to be no such fundamental "alphabet" for the items of cultural transmission. Consequently, the project of "memetics" as contending basis for evolutionary epistemology is on hold pending an adequate understanding of its basic ontology. [See the online Journal of Memetics for more information.]

Population models have been used to good effect in modeling cultural transmission processes. Evolutionary game theory models are frequently claimed to cover both processes in which strategies are inherited and those in which they are imitated. This application is possible in the absence of any specification of the underlying nature of strategies, for instance, whether they are to be thought of as "things" which are replicated, or whether they are properties or states of the individuals whose strategies they are. This is sometimes referred to as the "epidemiological approach", though again, the comparison to infection is due to the quantitative tools used in analysis rather than to any presupposition regarding the underlying ontology of cultural transmission.

2.3 Multi-Level Evolution

The kind of levels involved in evolutionary epistemology are quite different than the kind of levels of selection which are discussed much more often in the "levels of selection" debate in evolutionary biology. In evolutionary biology, the "levels" of selection under discussion are levels of scale. The debate concerns whether genes are always the "units" or "targets" of selection, or whether selection can occur on higher levels, like organisms, groups, and species. The levels involved in evolutionary epistemology, on the other hand, are levels of the regulatory hierarchy involved in the control of behavior. These include the genetic bases of cognitive and perceptual hardware, concepts, languages, techniques, beliefs, preferences, and so forth. Note that in the case of evolutionary epistemology, the terms "levels" and "hierarchy" may be impressionistic. There is often no clear arrangement of levels at all.

There are at least two different approaches that have been taken to modeling multi-level evolution.

1. Dual Transmission Models: Boyd and Richerson (1985) adapted models from genetics to model a case in which a trait (cooperation) was affected both by genetic and cultural evolution. It was first shown that a genetically determined bias on cultural transmission could be selected for in a migratory population. The bias made it easier to pick up local customs, increasing the likelihood of imitation beyond that determined by the frequency and perceived value of the behavior. Once this bias was in place, its effect was strong enough to overcome the perceived costs involved in cooperative behavior. The model yielded two important results. First, it provided a novel mechanism according to which cooperative behavior can stabilize in migratory populations. But more importantly, it demonstrated that cultural evolution cannot be predicted purely on the basis of genetic fitnesses.
2. Multiple Population Models: Harms (1997) constructed a multi-level dynamic population model of bumblebee learning. Mutual information between distributions of sensor types, overt foraging behaviors, and internal foraging preferences on the one hand and environmental states on the other was assessed and compared to average fitness of the population states. It was shown that information present in overt behaviors may be underutilized, and that exaptation of sensor mechanisms for preference formation can bring about the utilization of that information.

2.4 Meaning

Full descriptive accounts of truth and justification both demand a theory of meaning. Until a sign has meaning, it cannot be true or false. Moreover, determining the meaning of justificatory claims may provide a descriptive theory of justification. Presumably, what makes a claim of justification true is the basis of that justification. If meaning is conventional, then the evolution of meaning becomes an instance of the evolution of conventions.

Models of the evolution of conventions have in one case been extended to apply to meaning conventions. Skyrms (1996, chapter 5) gave an evolutionary interpretation of David Lewis' (1969) model of rational selection of meaning conventions. Skyrms was able to show that there is strong selection on the formation of "signaling systems" in mixed populations with a full set of coordinated, countercoordinated, and uncoordinated strategies. It is significant that the structure of the model and the selective process by which meaning conventions emerge and are stabilized largely parallels the account of the evolution of meaning given by Ruth Millikan (1984).

In the simplest version, the model is constructed as follows: We imagine that there are two states of affairs T, two acts A, and two signals M. Players have an equal chance of being in either the position of sender, or receiver. Receivers must decide what to do based purely on what the sender tells them. In this purely cooperative version, each player gets one point if the receiver does A_1 if the state is T_1 or A_2 if the state is T_2.

Since players will be both sender and receiver, they must have a strategy for each situation. There are sixteen such strategies, and we suppose them to be either inherited (or learned) from biological parents, or imitated on the basis of perceived success in terms of points earned. Strategies I_1 and I_2 are signaling

systems, in that if both players play the same one of these two strategies they will always get their payoff. I_3 and I_4 are anti-signaling strategies, which result in consistent miscoordination, though they do well against each other. All of the other strategies involve S_3, S_4, R_3, or R_4, which results in the same act being performed no matter what the external state is.

Sender Strategies:

S_1: Send M_1 if T_1; M_2 if T_2

S_2: Send M_2 if T_1;M_1 if T_2

S_3: Send M_1 if T_1 or T_2

S_4: Send M_2 if T_1 or T_2

Receiver Strategies:

R_1: Do A_1 if M_1; A_2 if M_2

R_2: Do A_2 if M_1; A_1 if M_2

R_3: Do A_1 for M_1 or M_2

R_4: Do A_2 for M_1 or M_2

Complete Strategies:

I_1: S_1,R_1

I_2: S_2,R_2,

I_3: S_1,R_2

I_4: S_2,R_2

I_5: S_1,R_3

I_6: S_2,R_3

I_7: S_1,R_4

I_8: S_2,R_4

I_9: S_3,R_1

I_{10}: S_3,R_2

I_{11}: S_3,R_3

I_{12}: S_3,R_4

I_{13}: S_4,R_1

I_{14}: S_4,R_2

I_{15}: S_4,R_3

I_{16}: S_4,R_4

Simulation results showed that virtually all initial population distributions become dominated by one or the other of the two signaling system strategies. The situation becomes more complex when more realistic payoffs are introduced, for instance, that the sender incurs a cost rather than automatically sharing the benefit that the receiver gets from correct behavior for the environment. Even in such situations, however, the most likely course of evolution is domination by a signaling system.

Bibliography

- Bradie, Michael(1986), "Assessing Evolutionary Epistemology," in *Biology & Philosophy* **1**, 401-459.
- Bradie, Michael(1994), "Epistemology from an Evolutionary Point of View." In *Conceptual Issues in Evolutionary Biology*, edited by Elliott Sober, 453- 475. Cambridge, MA: The MIT Press.
- Bradie, Michael. (1989), "Evolutionary Epistemology as Naturalized Epistemology." In *Issues in Evolutionary Epistemology*, edited by K. Hahlweg and C. A. Hooker, 393-412. Albany, NY: SUNY Press.
- Boyd, Robert, and Peter J. Richerson (1985), *Culture and the Evolutionary Process*, Chicago: The University of Chicago Press, 299 pp.
- Campbell, Donald T. (1956a), "Adaptive Behavior from Random Response." *Behavioral Science* 1, no. 2: 105-110.
- Campbell, Donald T. (1956b), "Perception as Substitute Trial and Error." *Psychological Review* 63, no. 5: 331- 342.
- Campbell, Donald T. (1959), "Methodological Suggestions from a Comparative Psychology of Knowledge processes." *Inquiry* 2: 152-182.
- Campbell, Donald T. (1960), "Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes." *Psychological Review* 67, no. 6: 380-400.
- Campbell, Donald T. (1974), "Evolutionary Epistemology." In *The philosophy of Karl R. Popper*, edited by P. A. Schilpp, 412-463. LaSalle, IL: Open Court.
- Campbell, Donald T.(1974b), "Unjustified Variation and Selective Retention in Scientific Discovery." In *Studies in the philosophy of biology*, edited by F J. Ayala and T. Dobzhansky, 139-161. London: Macmillan.
- Campbell, Donald T. (1982), "The "Blind-Variation-and-Selective-Retention" Theme." In *The cognitive-developmental psychology of James Mark Baldwin: Current theory and research in*

genetic epistemology, edited by J. M. Broughton and D. J. Freeman-Moir, 87-97. Norwood, NJ: Ablex.

- Campbell, D. T. (1985), "Pattern Matching as an Essential in Distal Knowing." In *Naturalizing Epistemology*, edited by H. Kornblith, 49-70. Cambridge, MA: MIT Press.
- Campbell, Donald T. (1988), "Popper and Selection Theory." *Social Epistemology* 2, no. 4: 371-377.
- Campbell, Donald T., and Paller, Bonnie T. (1989), "Extending Evolutionary Epistemology to"Fustifying" Scientific Beliefs (A sociological rapprochement with a fallibilist perceptual foundationalism?)." In *Issues in evolutionary epistemology*, edited by K. Hahlweg and C. A. Hooker, 231-257. Albany: State University of New York Press.
- Changeux, Jean-Pierre (1985), *Neuronal Man*, New York: Pantheon.
- Dawkins, Richard (1976), *The Selfish Gene*, New York: Oxford University Press.
- Dretske, Fred (1971), "Perception From an Evolutionary Point of View." *Journal of Philosophy*, LXVIII/19: 584-591
- Edelman, G. M. (1987), *Neural Darwinism: The Theory of Neuronal Group Selection*, Basic Books: New York.
- Godfrey-Smith, Peter (1996), *Complexity and the Function of Mind in Nature*, Cambridge University Press.
- Harms, William F. (1997), "Reliability and Novelty: Information Gain in Multi-Level Selection Systems," *Erkenntnis*, Vol. 46, pp. 335-363.
- Hofbauer, Josef, and Karl Sigmund (1988), *The Theory of Evolution and Dynamical Systems*, Cambridge University Press.
- Lewis, David (1969), *Convention*, Cambridge University Press.
- Hull, D. (1988), *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*, The University of Chicago Press.
- Kim, Jagwon. (1988) "What is 'Naturalized Epistemology'?", *Philosophical Perspectives 2. Epistemology*, Atascadero: Ridgeview, 381-405.
- Kuhn, Thomas (1962), *The Structure of Scientific Revolutions*, The University of Chicago Press.
- Lakatos, I. and Musgrave, A. (eds.) (1970), *Criticism and the Growth of Knowledge*, Cambridge University Press.
- Lorenz, Konrad (1977), *Behind the Mirror*, London: Methuen.
- Millikan, Ruth (1984), *Language, Thought, and other Biological Categories*, Cambridge, MA: MIT Press.
- Munz, Peter (1993), *Philosophical Darwinism: On the Origin of Knowledge by Means of Natural Selection*, London: Routledge.
- Plotkin, H. C. (ed.). (1982), *Learning, Development, and Culture: Essays in Evolutionary Epistemology*, New York: John Wiley & Sons.
- Popper, KarlR. (1968), *The Logic of Scientific Discovery*, New York: Harper.
- Popper, Karl R. (1972), *Objective Knowledge: An Evolutionary Approach*, Oxford: The Clarendon Press.
- Popper, Karl R. (1984), "Evolutionary Epistemology," in *Evolutionary Theory: Paths into the Future*, (ed.) J. W. Pollard, London: John Wiley & Sons Ltd.
- Radnitzky, G. and Bartley, W. W. (1987), *Evolutionary Epistemology, Theory of Rationality and*

the Sociology of Knowledge, LaSalle, Ill: Open Court.

- Rescher, Nicholas (1978), *Scientific Progress: A Philosophical Essay on the Economics of Research in Natural Science*, Oxford: Basil Blackwell.
- Rescher, Nicholas (1989), *Cognitive Economy: The Economic Dimension of the Theory of Knowledge*, University of Pittsburgh Press.
- Rescher, Nicholas (1990), *A Useful Inheritance: Evolutionary Aspects of the Theory of Knowledge*, Lanham, MD: Rowman.
- Riedl, Rupert (1984), *Biology of Knowledge: The Evolutionary Basis of Reason*, Chichester: John Wiley & Sons.
- Ruse, Michael (1986), *Taking Darwin Seriously: A Naturalistic Approach to Philosophy*, Oxford: Blackwell.
- Schuster, Peter, and Karl Sigmund (1983), ‘Replicator Dynamics,’ *Journal of Theoretical Biology*, Vol. 100, pp. 533-538.
- Skinner, B. F. (1981), "Selection by Consequences," *Science* **213**, 501-504.
- Skyrms, Brian (1992), "Chaos and the Explanatory Significance of Equilibrium: Strange Attractors in Evolutionary Game Dynamics," PSA 1992 <2> (Philosophy of Science Association), pp. 374-394.
- Skyrms, Brian (1996), *Evolution of the Social Contract*, Cambridge University Press.
- Taylor, P., and L. Jonker (1978), "Evolutionary Stable Strategies and Game Dynamics," *Mathematical Biosciences*, Vol. 40:145-56.
- Toulmin, Stephen (1967), "The Evolutionary Development of Natural Science," *American Scientist* **55**, 4.
- Toulmin, Stephen (1972), *Human Understanding: The Collective Use and Evolution of Concepts*, Princeton University Press.

Other Internet Resources

- Selection Theory Bibliography, maintained by Gary Cziko (Educational Psychology, University of Illinois)
- Evolving Artificial Moral Ecologies
  , Centre for Applied Ethics, University of British Columbia
- The Journal of Memetics
  , sponsored by the Centre for Policy Modeling (Manchester Metropolitan University), the Principia Cybernetica Project, and Systems Engineering, Policy Analysis and Management (Delft University of Technology)

Related Entries

epistemology: naturalized | game theory: evolutionary | information | teleology: teleological notions in biology

Michael Bradie
mbradie@bgnet.bgsu.edu
and
William Harms
billharms@billharms.com

First published: January 10, 2001
Content last modified: January 10, 2001

Stanford Encyclopedia of Philosophy

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Game Theory

Game theory is the study of the ways in which *strategic interactions* among *rational players* produce *outcomes* with respect to the *preferences* (or *utilities*) of those players, none of which might have been intended by any of them. The meaning of this statement will not be clear to the non-expert until each of the italicized words and phrases has been explained and featured in some examples. Doing this will be the main business of this article. First, however, we provide some historical and philosophical context in order to motivate the reader for all of this technical work ahead.

1. Philosophical and Historical Motivation

The mathematical theory of games was invented by John von Neumann and Oskar Morgenstern (1944). For reasons to be discussed later, limitations in their mathematical framework initially made the theory applicable only under special and limited conditions. This situation has gradually changed, in ways we will examine as we go along, over the past six decades, as the framework was deepened and generalized. Refinements are still being made, and we will review a few outstanding philosophical problems that lie along the advancing front edge of these developments towards the end of the article. However, since at least the late 1970s it has been possible to say with confidence that game theory is the most important and useful tool in the analyst's kit whenever she confronts situations in which one agent's rational decision-making depends on her expectations about what one or more other agents will do, and theirs similarly depend on expectations about her.

Despite the fact that game theory has been rendered mathematically and logically systematic only recently, however, game-theoretic insights can be found among philosophers and political commentators going back to ancient times. For example, Plato, in the *Republic*, at one point has Socrates worry about the following situation. Consider a soldier at the front, waiting with his comrades to repulse an enemy attack. It may occur to him that if the defense is likely to be successful, then it isn't very probable that his own personal contribution will be essential. But if he stays, he runs the risk of being killed or wounded -- apparently for no point. On the other hand, if the enemy is going to win the battle, then his chances of death or injury are higher still, and now quite clearly to no point, since the line will be overwhelmed anyway. Based on this reasoning, it would appear that the soldier is better off running away regardless of who is going to win the battle. Of course, if all of the soldiers reason this way -- as they all apparently *should*, since they're all in identical situations -- then this will certainly *bring about* the outcome in which the battle is lost. Of course, this point, since it has occurred to us as analysts, can occur to the soldiers too. Does this give them a reason for staying at their posts? Just the contrary: the greater the soldiers' fear that the battle will be lost, the greater their incentive to get themselves out of harm's way. And the greater the soldiers' belief that the battle will be won, without the need of any particular individual's contributions, the less reason they have to stay and fight. If each soldier *anticipates* this sort of reasoning on the part of the others, all will quickly reason themselves into a panic, and their horrified commander will have a rout on his hands before the enemy has even fired a shot!

Long before game theory had come along to show people how to think about this sort of problem systematically, it had occurred to some actual military leaders and influenced their strategies. Thus the Spanish conqueror Cortez, when landing in Mexico with a small force who had good reason to fear their capacity to repel attack from the far more numerous Aztecs, removed the risk that his troops might think their way into a retreat by burning the ships on which they had landed. With retreat having thus been rendered physically impossible, the Spanish soldiers had no better course of action but to stand and fight -- and, furthermore, to fight with as much determination as they could muster. Better still, from Cortez's point of view, his action had a discouraging effect on the motivation of the Aztecs. He took care to burn his ships very visibly, so that the Aztecs would be sure to see what he had done. They then reasoned as follows: Any commander who could be so confident as to willfully destroy his own option to be prudent if the battle went badly for him must have good reasons for such extreme optimism. It cannot be wise to attack an opponent who has a good reason (whatever, exactly, it might be) for being sure that he can't lose. The Aztecs therefore retreated into the surrounding hills, and Cortez had his victory bloodlessly.

This situation, as imagined by Plato and as vividly acted upon by Cortez, has a deep and interesting logic. Notice that the soldiers are not motivated to retreat *just*, or even mainly, by their rational assessment of the dangers of battle and by their self-interest. Rather, they discover a sound reason to run away by realizing that what it makes sense for them to do depends on what it will make sense for others to do, and that all of the others can notice this too. Even a quite brave soldier may prefer to run rather than heroically, but pointlessly, die trying to stem the oncoming tide all by himself. Thus we could imagine, without contradiction, a circumstance in which an army, all of whose members are brave, flees at top speed before the enemy makes a move. If the soldiers really *are* brave, then this surely isn't the outcome any of them wanted; each would have preferred that all stand and fight. What we have here, then, is a case in which the *interaction* of many individually rational decision-making processes -- one process per soldier -- produces an outcome intended by no one. (Most armies try to avoid this problem just as Cortez did. Since they can't usually make retreat *physically* impossible, they make it *economically* impossible: they shoot deserters. Then standing and fighting is each soldier's individually rational course of action after all, because the cost of running is sure to be at least as high as the cost of staying.)

Another classic source that reproduces exactly this sequence of reasoning is Shakespeare. In *Henry V*, he has the victor of Agincourt explain his decision to slaughter French prisoners, in full view of the enemy, as follows. His own troops observe that the prisoners have been killed, and observe that the enemy has observed this. Therefore, they know what fate will await them at the enemy's hand if they don't win. Metaphorically, but very effectively, their boats have been burnt. The French prisoners died as a means by which Henry sent a signal to *his own troops*, thereby changing their incentives.

These examples might seem to be relevant only for those who find themselves in sordid situations of cut-throat competition. Perhaps, one might think, it is important for generals, politicians, businesspeople and others whose jobs involve manipulation of others, but the philosopher should only deplore its horrid morality. Such a conclusion would be highly premature, however. The study of the *logic* that governs the interrelationships amongst incentives, strategic interactions and outcomes has been fundamental in modern political philosophy, since centuries before anyone had an explicit name for this sort of logic.

Hobbes's *Leviathan* is often regarded as the founding work in modern political philosophy, the text that began the continuing round of analyses of the function and justification of the state and its restrictions on individual liberties. The core of Hobbes's reasoning can be given quite straightforwardly as follows. The best situation for all people is one in which each is free to do as she pleases. Often, such free people will wish to cooperate with one another in order to carry out projects that would be impossible for an individual acting alone. But if there are any immoral or amoral agents around, they will notice that their interests are best served by getting the benefits from cooperation and not returning them. Suppose, for example, that you agree to help me build my house in return for my promise to help you build yours. After my house is finished, I can make your labour free to me simply by reneging on my promise. I then realize, however, that if this leaves you with no house, you will have an incentive to take mine. This will put me in constant fear of you, and force me to spend valuable time and resources guarding myself against you. I can best minimize these costs by striking first and killing you at the first opportunity. Of course, you can anticipate all of this reasoning by me, and so have good reason to try to beat me to the

punch. Since I can anticipate *this* reasoning by *you*, my original fear of you was not paranoid; nor was yours of me. In fact, neither of us actually needs to be immoral to get this chain of mutual reasoning going; we need only think that there is some *possibility* that the other might try to cheat on bargains. Once a small wedge of doubt enters any one mind, the incentive induced by fear of the consequences of being *preempted* -- hit before hitting first -- quickly becomes overwhelming on both sides. If either of us has any resources of our own that the other might want, this murderous logic will take hold long before we are so silly as to imagine that we could ever actually get so far as making deals to help one another build houses in the first place. Left to their own devices, rational agents will never derive the benefits of cooperation, and will instead live from the outset in a state of 'war of all against all', in Hobbes's words. In these circumstances, all human life, as he vividly and famously put it, will be "solitary, poor, nasty, brutish and short."

Hobbes's proposed solution to this problem was tyranny. The people can hire an agent -- a government -- whose job is to punish anyone who breaks any promise. So long as the threatened punishment is sufficiently dire -- Hobbes thought decapitation generally appropriate -- then the cost of reneging on promises will exceed the cost of keeping them. The logic here is identical to that used by an army when it threatens to shoot deserters. If all people know that these incentives hold for most others, then cooperation will not only be possible, but will be the expected norm, and the war of all against all becomes a general peace.

Hobbes pushes the logic of this argument to a very strong conclusion, arguing that it implies not only a government with the right and the power to enforce cooperation, but an 'undivided' government in which the arbitrary will of a single ruler must impose absolute obligation on all. Few contemporary political theorists think that the particular steps by which Hobbes reasons his way to this conclusion are both sound and valid. Working through these issues here, however, would carry us away from our topic into complex details of contractarian political philosophy. What is important in the present context is that these details, as they are in fact pursued in the contemporary debates, all involve sophisticated interpretation of the issues using the resources of modern game theory. Furthermore, Hobbes's most basic point, that the fundamental justification for the coercive authority and practices of governments is peoples' own need to protect themselves from what game theorists call 'social dilemmas', is accepted by many, if not most, political theorists. Notice that Hobbes has *not* argued that tyranny is a desirable thing in itself. The structure of his argument is that the logic of strategic interaction leaves only two general political outcomes possible: tyranny and anarchy. Rational agents then choose tyranny as the lesser of two evils.

The reasoning of Cortez, of Henry V and of Hobbes's political agents has a common logic, one derived from their situations. In each case, the aspect of the environment that is most important to the agents' achievement of their preferred outcomes is the set of expectations and possible reactions to their strategies by other agents. The distinction between acting *parametrically* on a passive world and acting *non-parametrically* on a world that tries to act in anticipation of these actions is fundamental. If you wish to kick a rock down a hill, you need only concern yourself with the rock's mass relative to the force of your blow, the extent to which it is bonded with its supporting surface, the slope of the ground on the other side of the rock, and the expected impact of the collision on your foot. The values of all of these variables are independent of your plans and intentions, since the rock has no interests of its own and takes no actions to

attempt to assist or thwart you. By contrast, if you wish to kick a person down the hill, then unless that person is unconscious, bound or otherwise incapacitated, you will likely not succeed unless you can disguise your plans until it's too late for him to take either evasive or forestalling action. The logical issues associated with the second sort of situation are typically much more complicated, as a simple hypothetical example will illustrate.

Suppose first that you wish to cross a river that is spanned by three bridges. (Assume that swimming, wading or boating across are impossible.) The first bridge is known to be safe and free of obstacles; if you try to cross there, you will succeed. The second bridge lies beneath a cliff from which large rocks sometimes fall. The third is inhabited by deadly cobras. Now suppose you wish to rank-order the three bridges with respect to their preferability as crossing-points. Your task here is quite straightforward. The first bridge is obviously best, since it is safest. To rank-order the other two bridges, you require information about their relative levels of danger. If you can study the frequency of rock-falls and the movements of the cobras for awhile, you might be able to calculate that the cobra bridge is four times more dangerous than the rocky bridge, and the rocky bridge 20% more dangerous than the unobstructed bridge. Your reasoning here is strictly parametric because neither the rocks nor the cobras are trying to influence your actions, by, for example, concealing their typical patterns of behaviour because they know you are studying them. It is quite obvious what you should do here: cross at the safe bridge. Now let us complicate the situation a bit. Suppose that the bridge with the rocks was immediately before you, while the safe bridge was a day's difficult hike upstream. Your decision-making situation here is slightly more complicated, but it is still strictly parametric. You would have to decide whether the cost of the long hike was worth exchanging for the penalty of a 20% chance of being hit by a rock. However, this is all you must decide, and your probability of a successful crossing is entirely up to you; the environment is not interested in your plans.

However, if we now complicate the situation in the direction of non-parametricity, it becomes much more puzzling. Suppose that you are a fugitive of some sort, and waiting on the other side of the river with a gun is your pursuer. She will catch and shoot you, let us suppose, only if she waits at the bridge you try to cross; otherwise, you will escape. As you reason through your choice of bridge, it occurs to you that she is over there trying to anticipate your reasoning. It will seem that, surely, choosing the safe bridge straight away would be a mistake, since that is just where she will expect you, and your chances of death rise to certainty. So perhaps you should risk the rocks, since these odds are much better. But wait ... if you can reach this conclusion, your pursuer, who is just as rational and well-informed as you are, can anticipate that you will reach it, and will be waiting for you if you evade the rocks. So perhaps you must take your chances with the cobras; that is what she must least expect. But, then, no ... if she expects that you will expect that she will least expect this, then she will most expect it. This dilemma, you realize with dread, is general: you must do what your pursuer least expects; but whatever you most expect her to least expect is automatically what she will most expect. You appear to be trapped in indecision. All that might console you a bit here is that, on the other side of the river, your pursuer is trapped in exactly the same quandary, unable to decide which bridge to wait at because as soon as she imagines committing to one, she will notice that if she can find a best reason to pick a bridge, you can anticipate that same reason and then avoid her.

We know from experience that, in situations such as this, people do not usually stand and dither in circles forever. As we'll see later, there *is* a rational solution -- that is, a best rational action -- available to both players. However, until the 1940s neither philosophers nor economists knew how to find it mathematically. As a result, economists were forced to treat non-parametric influences as if they were complications on parametric ones. This is likely to strike the reader as odd, since, as our example of the bridge-crossing problem was meant to show, non-parametric features are often fundamental features of decision-making problems. Part of the explanation for game theory's relatively late entry into the field lies in the problems with which economists had historically been concerned. Classical economists, such as Adam Smith and David Ricardo, were mainly interested in the question of how agents in very large markets -- whole nations -- could interact so as to bring about maximum monetary wealth for themselves. Smith's basic insight, that efficiency is best maximized by agents freely seeking mutually advantageous bargains, was mathematically verified in the twentieth century. However, the demonstration of this fact applies only in conditions of ‘perfect competition,’ that is, when firms face no costs of entry or exit into markets, when there are no economies of scale, and when no agents' actions have unintended side-effects on other agents' well-being. Economists always recognized that this set of assumptions is purely an idealization for purposes of analysis, not a possible state of affairs anyone could try (or should want to try) to attain. But until the mathematics of game theory matured near the end of the 1970s, economists had to hope that the more closely a market *approximates* perfect competition, the more efficient it will be. No such hope, however, can be mathematically or logically justified in general; indeed, as a strict generalization the assumption can be shown to be false.

This article is not about the foundations of economics, but it is important for understanding the origins and scope of game theory to know that perfectly competitive markets have built into them a feature that renders them susceptible to parametric analysis. Because agents face no entry costs to markets, they will open shop in any given market until competition drives all profits to zero. This implies that if costs and demand are fixed, then agents have no options about how much to produce if they are trying to maximize the differences between their costs and their revenues. These production levels can be determined separately for each agent, so none need pay attention to what the others are doing; each agent treats her counterparts as passive features of the environment. The other kind of situation to which classical economic analysis can be applied without recourse to game theory is that of monopoly. Here, quite obviously, non-parametric considerations drop out, since there is only one agent under study. However, both perfect and monopolistic competition are very special and unusual market arrangements. Prior to the advent of game theory, therefore, economists were severely limited in the class of circumstances to which they could neatly apply their models.

Philosophers share with economists a professional interest in the conditions and techniques for the maximization of human welfare. In addition, philosophers have a special concern with the logical justification of actions, and often actions must be justified by reference to their expected outcomes. Without game theory, both of these problems resist analysis wherever non-parametric aspects are relevant. We will demonstrate this shortly by reference to the most famous (though not the most typical) game, the so-called *Prisoner's Dilemma*, and to other, more typical, games. In doing this, we will need to introduce, define and illustrate the basic elements and techniques of game theory. To this job we therefore now turn.

2. Basic Elements and Assumptions of Game Theory

2.1 Utility

An agent is, by definition, an entity with *goals and preferences*. Game theorists, like economists and philosophers studying rational decision-making, describe these by means of an abstract concept called *utility*. This refers to the amount of satisfaction an agent derives from an object or an event, so that an agent who, for example, adores the taste of pickles would be said to associate high utility with them, while an agent who can take or leave them derives a lower level of utility from them. Examples of this kind suggest that 'utility' denotes a measure of subjective *psychological* fulfillment, and this is indeed how the concept was generally (though not always) interpreted prior to the 1930s. During that decade, however, economists and philosophers under the influence of behaviourism objected to the theoretical use of such unobservable entities as 'psychological fulfillment quotients.' The economist Paul Samuelson (1938) therefore set out to define utility in such a way that it becomes a purely technical concept. That is, when we say that an agent acts so as to maximize her utility, we mean by 'utility' simply whatever it is that the agent's behavior suggests her to consistently desire. If this looks circular to you, it should: theorists who follow Samuelson *intend* the statement 'agents act so as to maximize their utility' as a tautology. Like other tautologies occurring in the foundations of scientific theories, it is useful not in itself, but because it helps to fix our contexts of inquiry.

Though we might no longer be moved by scruples derived from *psychological* behaviorism, many theorists continue to follow Samuelson's way of understanding utility because they think it important that game theory apply to *any* kind of agent -- a person, a bear, a bee, a firm or a country -- and not just to agents with human minds. When such theorists say that agents act so as to maximize their utility, they want this to be part of the *definition* of what it is to be an agent, not an empirical claim about possible inner states and motivations. Samuelson's conception of utility, defined by way of *Revealed Preference Theory* (RPT) introduced in his classic paper (Samuelson (1938)) satisfies this demand.

Some other theorists understand the point of game theory differently. They view game theory as providing an explanatory account of strategic reasoning. For this idea to be applicable, we must suppose that agents at least sometimes do what they do in non-parametric settings *because* game-theoretic logic recommends certain actions as the rational ones. Still other theorists interpret game theory *normatively*, as advising agents on what to do in strategic contexts in order to maximize their utility. Fortunately for our purposes, all of these ways of thinking about the possible uses of game theory are compatible with the tautological interpretation of utility maximization. The philosophical differences are not idle from the perspective of the working game theorist, however. As we will see in a later section, those who hope to use game theory to explain strategic *reasoning*, as opposed to merely strategic *behavior*, face some special philosophical and practical problems.

Since game theory involves formal reasoning, we must have a device for thinking of utility maximization in mathematical terms. Such a device is called a *utility function*. The utility-map for an agent is called a 'function' because it maps *ordered preferences* onto the real numbers. Suppose that agent *x* prefers bundle *a* to bundle *b* and bundle *b* to bundle *c*. We then map these onto a list of numbers, where the function maps the highest-ranked bundle onto the largest number in the list, the second-highest-ranked bundle onto the next-largest number in the list, and so on, thus:

bundle *a* » 3

bundle *b* » 2

bundle *c* » 1

The only property mapped by this function is *order*. The magnitudes of the numbers are irrelevant; that is, it must not be inferred that *x* gets 3 times as much utility from bundle *a* as she gets from bundle *c*. Thus we could represent *exactly the same* utility function as that above by

bundle *a* » 7,326

bundle *b* » 12.6

bundle *c* » -1,000,000

The numbers featuring in an ordinal utility function are thus not measuring any *quantity* of anything. A utility-function in which magnitudes *do* matter is called 'cardinal'. Whenever someone refers to a utility function without specifying which kind is meant, you should assume that it's ordinal. These are the sorts we'll need for the first set of games we'll examine. Later, when we come to seeing how to solve games that involve *randomization* -- our river-crossing game from Part 1 above, for example -- we'll need to build cardinal utility functions. The technique for doing this was given by von Neumann & Morgenstern (1947), and was an essential aspect of their invention of game theory. For the moment, however, we will need only ordinal functions.

2.2 Games and Information

All situations in which at least one agent can only act to maximize his utility through anticipating the responses to his actions by one or more other agents is called a *game*. Agents involved in games are referred to as *players*. If all agents have optimal actions regardless of what the others do, as in purely parametric situations or conditions of monopoly or perfect competition (see Section 1 above) we can model this without appeal to game theory; otherwise, we need it.

We assume that players are economically rational. That is, a player can (i) assess outcomes; (ii) calculate

paths to outcomes; and (iii) choose actions that yield their most-preferred outcomes, given the actions of the other players. This rationality might in some cases be internally computed by the agent. In other cases, it might simply be embodied in behavioral dispositions built by natural, cultural or economic selection.

Each player in a game faces a choice among two or more possible *strategies*. A strategy is a predetermined 'programme of play' that tells her what actions to take in response to *every possible strategy other players might use*. The significance of the italicized phrase here will become clear when we take up some sample games below.

A crucial aspect of the specification of a game involves the information that players have when they choose strategies. The simplest games (from the perspective of logical structure) are those in which agents have *perfect information*, meaning that at every point where each agent's strategy tells her to take an action, she knows everything that has happened in the game up to that point. A board-game of sequential moves in which in which both players watch all the action (and know the rules in common), such as chess, is an instance of such a game. By contrast, the example of the bridge-crossing game from Section 1 above illustrates a game of *imperfect information*, since the fugitive must choose a bridge to cross without knowing the bridge at which the pursuer has chosen to wait, and the pursuer similarly makes her decision in ignorance of the actions of her quarry. Since game theory is about rational action given the strategically significant actions of others, it should not surprise you to be told that what agents in games know, or fail to know, about each others' actions makes a considerable difference to the logic of our analyses, as we will see.

2.3 Trees and Matrices

The difference between games of perfect and of imperfect information is closely related to (though certainly not identical with!) a distinction between *ways of representing* games that is based on *order of play*. Let us begin by distinguishing between sequential-move and simultaneous-move games in terms of information. It is natural, as a first approximation, to think of sequential-move games as being ones in which players choose their strategies one after the other, and of simultaneous-move games as ones in which players choose their strategies at the same time. This isn't quite right, however, because what is of strategic importance is not the temporal *order* of events per se, but whether and when players *know about* other players' actions relative to having to choose their own. For example, if two competing businesses are both planning marketing campaigns, one might commit to its strategy months before the other does; but if neither knows what the other has committed to or will commit to when they make their decisions, this is a simultaneous-move game. Chess, by contrast, is normally played as a sequential-move game: you see what your opponent has done before choosing your own next action. (Chess *can* be turned into a simultaneous-move game if the players each call moves while isolated from the common board; but this is a very different game from conventional chess.)

It was said above that the distinction between sequential-move and simultaneous-move games is not identical to the distinction between perfect-information and imperfect-information games. Explaining why this is so is a good way of establishing full understanding of both sets of concepts. As simultaneous-

move games were characterized in the previous paragraph, it must be true that all simultaneous-move games are games of imperfect information. However, some games may contain mixes of sequential and simultaneous moves. For example, two firms might commit to their marketing strategies independently and in secrecy from one another, but thereafter engage in pricing competition in full view of one another. If the optimal marketing strategies were partially or wholly dependent on what was expected to happen in the subsequent pricing game, then the two stages would need to be analyzed as a single game, in which a stage of sequential play followed a stage of simultaneous play. Whole games that involve mixed stages of this sort are games of imperfect information, however temporally staged they might be. Games of perfect information (as the name implies) denote cases where *no* moves are simultaneous (and where no player ever forgets what has gone before).

It was said above that games of perfect information are the (logically) simplest sorts of games. This is so because in such games (as long as the games are finite, that is, terminate after a known number of actions) players and analysts can use a straightforward procedure for predicting outcomes. A rational player in such a game chooses her first action by considering each series of responses and counter-responses that will result from each action open to her. She then asks herself which of the available final outcomes brings her the highest utility, and chooses the action that starts the chain leading to this outcome. This process is called *backward induction* (because the reasoning works backwards from eventual outcomes to present decision problems).

We will have much more to say about backward induction and its properties in a later section (when we come to discuss equilibrium and equilibrium selection). For now, we have described it just in order to use it to introduce one of the two types of mathematical objects used to represent games: *game-trees*. A game-tree is an example of what mathematicians call a *directed graph*. That is, it is a set of connected nodes in which the overall graph has a direction. We can draw trees from the top of the page to the bottom, or from left to right. In the first case, nodes at the top of the page are interpreted as coming earlier in the sequence of actions. In the case of a tree drawn from left to right, leftward nodes are prior in the sequence to rightward ones. An unlabelled tree has a structure of the following sort:

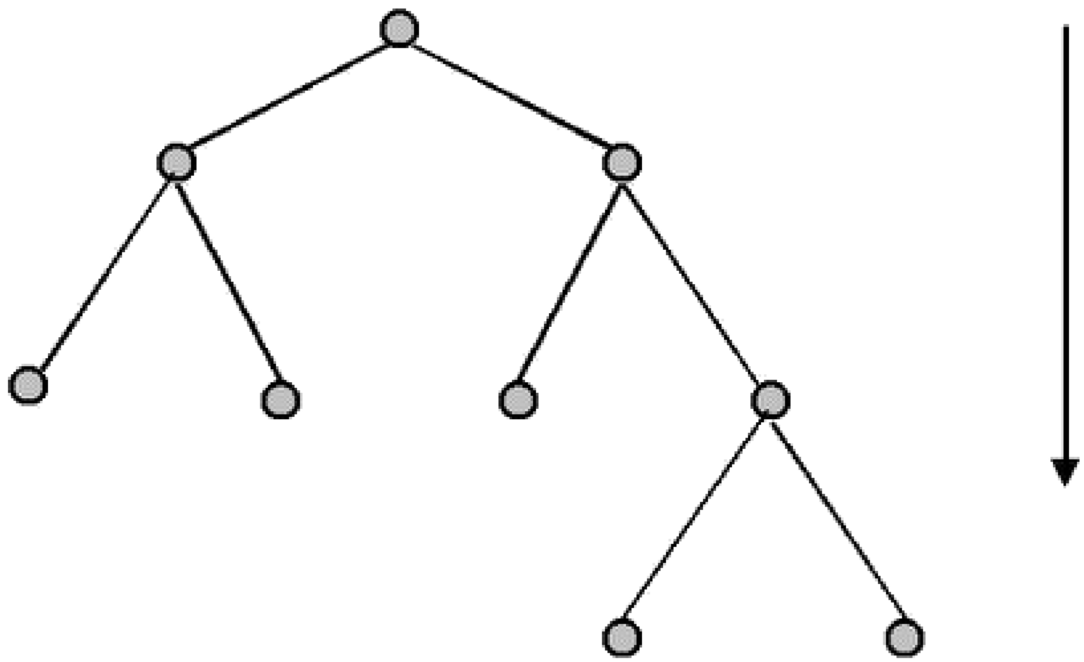


Figure 1

The point of representing games using trees can best be grasped by visualizing the use of them in supporting backward-induction reasoning. Just imagine the player (or analyst) beginning at the end of the tree, where outcomes are displayed, and then working backwards from these, looking for sets of strategies that describe paths leading to them. Since a player's utility function indicates which outcomes she prefers to which, we also know which paths she will prefer. Of course, not all paths will be possible because the other player has a role in selecting paths too, and won't take actions that lead to less preferred outcomes for him. We will present some examples of this interactive path-selection, and detailed techniques for reasoning through them, after we have described a situation we can use a tree to depict.

Trees are used to represent *sequential* games, because they show the order in which actions are taken by the players. However, games are sometimes represented on *matrices* rather than trees. This is the second type of mathematical object used to represent games. Matrices, unlike trees, simply show the outcomes, represented in terms of the players' utility functions, for every possible combination of strategies the players might use. For example, it makes sense to display the river-crossing game from Section 1 on a matrix, since in that game both the fugitive and the hunter have just one move each, and each chooses their move in ignorance of what the other has decided to do. Here, then, is *part of* the matrix:

		Hunter		
		Safe Bridge	Rocky Bridge	Cobra Bridge
Fugitive	Safe Bridge	0,1	1,0	1,0
	Rocky Bridge	?	0,1	?
	Cobra Bridge	?	?	0,1

Figure 2

The fugitive's three possible strategies -- cross at the safe bridge, risk the rocks and risk the cobras -- form the rows of the matrix. Similarly, the hunter's three possible strategies -- waiting at the safe bridge, waiting at the rocky bridge and waiting at the cobra bridge -- form the columns of the matrix. Each cell of the matrix shows -- or, rather *would* show if our matrix was complete -- an *outcome* defined in terms of the players' *payoffs*. A player's payoff is simply the number assigned by her ordinal utility function to the state of affairs corresponding to the outcome in question. For each outcome, Row's payoff is always listed first, followed by Column's. Thus, for example, the upper right-hand corner above shows that when the fugitive crosses at the safe bridge and the hunter is waiting there, the fugitive gets a payoff of 0 and the hunter gets a payoff of 1. We interpret these by reference to their utility functions, which in this game are very simple. If the fugitive gets safely across the river he receives a payoff of 1; if he doesn't he gets 0. If the fugitive doesn't make it, either because he's shot by the hunter or hit by a rock or struck by a cobra, then the hunter gets a payoff of 1 and the fugitive gets a payoff of 0.

We'll briefly explain the parts of the matrix that have been filled in, and then say why we can't yet complete the rest. Whenever the hunter waits at the bridge chosen by the fugitive, the fugitive is shot. These outcomes all deliver the payoff vector (0, 1). You can find them descending diagonally across the matrix above from the upper right-hand corner. Whenever the fugitive chooses the safe bridge but the hunter waits at another, the fugitive gets safely across, yielding the payoff vector (1, 0). These two outcomes are shown in the second two cells of the top row. All of the other cells are marked, *for now*, with question marks. Why? The problem here is that if the fugitive crosses at either the rocky bridge or the cobra bridge, he introduces parametric factors into the game. In these cases, he takes on some risk of getting killed, and so producing the payoff vector (0, 1), that is independent of anything the hunter does. We don't yet have enough concepts introduced to be able to show how to represent these outcomes in terms of utility functions -- but by the time we're finished we will, and this will provide the key to solving our puzzle from Section 1.

Matrix games are referred to as ‘normal-form’ or ‘strategic-form’ games, and games as trees are referred to as ‘extensive-form’ games. The two sorts of games are not equivalent, because extensive-form games contain information -- about sequences of play and players' levels of information about the game structure -- that strategic-form games do not. In general, a strategic-form game could represent any one of several extensive-form games, so a strategic-form game is best thought of as being a *set* of extensive-form games. When order of play is irrelevant to a game's outcome, then you should study its strategic form, since it's the whole set you want to know about. Where order of play *is* relevant, the extensive form *must* be specified or your conclusions will be unreliable.

2.4 The Prisoner's Dilemma as an Example of Strategic-Form vs. Extensive-Form Representation

The distinctions described above are difficult to fully grasp if all one has to go on are abstract descriptions. They're best illustrated by means of an example. For this purpose, we'll use the most famous game: the Prisoner's Dilemma. It in fact gives the logic of the problem faced by Cortez's and Henry V's soldiers (see Section 1 above), and by Hobbes's agents before they empower the tyrant. However, for reasons which will become clear a bit later, you should not take the PD as a *typical* game; it isn't. We use it as an extended example here only because it's particularly helpful for illustrating the *relationship* between strategic-form and extensive-form games (and later, for illustrating the relationships between one-shot and repeated games; see Section 4 below).

The name of the Prisoner's Dilemma game is derived from the following situation typically used to exemplify it. Suppose that the police have arrested two people whom they know have committed an armed robbery together. Unfortunately, they lack enough admissible evidence to get a jury to convict. They *do*, however, have enough evidence to send each prisoner away for two years for theft of the getaway car. The chief inspector now makes the following offer to each prisoner: If you will confess to the robbery, implicating your partner, and she does not also confess, then you'll go free and she'll get ten years. If you both confess, you'll each get 5 years. If neither of you confess, then you'll each get two years for the auto theft.

Our first step in modeling your situation as a game is to represent it in terms of utility functions. Both you and your partner's utility functions are identical:

Go free » 4

2 years » 3

5 years » 2

10 years » 0

The numbers in the function above are now used to express your and your partner's *payoffs* in the various outcomes possible in your situation. We will refer to you as 'Player I' and to your partner as 'Player II'. Now we can represent your entire situation on a matrix; this is the strategic form of your game:

Figure 3

Figure 3

Each cell of the matrix gives the payoffs to both players for each combination of actions. Player I's payoff appears as the first number of each pair, Player II's as the second. So, if both of you confess you each get a payoff of 2 (5 years in prison each). This appears in the upper-left cell. If neither of you confess, you each get a payoff of 3 (2 years in prison each). This appears as the lower-right cell. If you confess and your partner doesn't you get a payoff of 4 (going free) and she gets a payoff of 0 (ten years in prison). This appears in the upper-right cell. The reverse situation, in which she confesses and you refuse, appears in the lower-left cell.

You evaluate your two possible actions here by comparing your payoffs in each column, since this shows you which of your actions is preferable for each possible action by your partner. So, observe: If your partner confesses than you get a payoff of 2 by confessing and a payoff of 0 by refusing. If your partner refuses, you get a payoff of 4 by confessing and a payoff of 3 by refusing. Therefore, you're better off confessing regardless of what she does. Your partner, meanwhile, evaluates her actions by comparing her payoffs down each row, and she comes to exactly the same conclusion that you do. Wherever one action for a player is superior to her other actions for each possible action by the opponent, we say that the first action *strictly dominates* the second one. In the PD, then, confessing strictly dominates refusing for both players. Both players know this about each other, thus entirely eliminating any temptation to depart from the strictly dominated path. Thus both players will confess, and both will go to prison for 5 years.

The players, and analysts, can predict this outcome using a mechanical procedure, known as iterated elimination of strictly dominated strategies. You, as Player 1, can see by examining the matrix that your payoffs in each cell of the top row are higher than your payoffs in each corresponding cell of the bottom row. Therefore, it can never be rational for you to play your bottom-row strategy, viz., refusing to confess, *regardless of what your opponent does*. Since your bottom-row strategy will never be played, we can simply *delete* the bottom row from the matrix. Now it is obvious that Player II will not refuse to confess, since his payoff from confessing in the two cells that remain is higher than his payoff from refusing. So, once again, we can delete the one-cell column on the left from the game. We now have only one cell remaining, that corresponding to the outcome brought about by mutual confession. Since the reasoning that led us to delete all other possible outcomes depended at each step only on the premise that both players are economically rational - that is, prefer higher payoffs to lower ones - there is very strong grounds for viewing joint confession as the *solution* to the game, the outcome on which its play *must* converge. You should note that the order in which strictly dominated rows and columns are deleted doesn't matter. Had we begun by deleting the right-hand column and then deleted the bottom row, we would have arrived at the same solution.

It's been said a couple of times that the PD is not a typical game in many respects. One of these respects is

that all its rows and columns are either strictly dominated or strictly dominant. In any strategic-form game where this is true, iterated elimination of strictly dominated strategies is guaranteed to yield a unique solution. Later, however, we will see that for many games this condition does not apply, and then our analytic task is less straightforward.

You will probably have noticed something disturbing about the outcome of the PD. Had you both refused to confess, you'd have arrived at the lower-right outcome in which you each go to prison for only 2 years, thereby *both* earning higher utility than you receive when you confess. This is the most important fact about the PD, and its significance for game theory is quite general. We'll therefore return to it below when we discuss equilibrium concepts in game theory. For now, however, let us stay with our use of this particular game to illustrate the difference between strategic and extensive forms.

When people introduce the PD into popular discussions, you will sometimes hear them say that the police inspector must lock his prisoners into separate rooms so that they can't communicate with one another. The reasoning behind this idea seems obvious: if you could communicate, you'd surely see that you're both better off refusing, and could make an agreement to do so, no? This, one presumes, would remove your conviction that you must confess because you'll otherwise be sold up the river by your partner. In fact, however, this intuition is misleading and its conclusion is false.

When we represent the PD as a strategic-form game, we implicitly assume that the prisoners can't attempt collusive agreement since they choose their actions simultaneously. In this case, agreement before the fact can't help. If you are convinced that your partner will stick to the bargain then you can seize the opportunity to go scot-free by confessing. Of course, you realize that the same temptation will occur to her; but in that case you again want to make sure you confess, as this is your only means of avoiding your worst outcome. Your agreement comes to naught because you have no way of enforcing it; it constitutes what game theorists call ‘cheap talk’.

But now suppose that you do *not* move simultaneously. That is, suppose that one of you can choose *after* observing the other's action. This is the sort of situation that people who think non-communication important must have in mind. Now you can see that your partner has remained steadfast when it comes to your choice, and you need not be concerned about being suckered. However, this doesn't change anything, a point that is best made by re-representing the game in extensive form. This gives us our opportunity to introduce game-trees and the method of analysis appropriate to them.

First, however, here are definitions of some concepts that will be helpful in analyzing game-trees:

> *Node*: A point at which a player takes an action.
>
> *Initial node*: The point at which the first action in the game occurs.
>
> *Terminal node*: Any node which, if reached, ends the game. Each terminal node corresponds to an *outcome*.

Subgame: Any set of nodes and branches descending uniquely from one node.

Payoff: an ordinal utility number assigned to a player at an outcome.

Outcome: an assignment of a set of payoffs, one to each player in the game.

Strategy: a program instructing a player which action to take at every node in the tree where she could possibly be called on to make a choice.

These quick definitions may not mean very much to you until you follow them being put to use in our analyses of trees below. It will probably be best if you scroll back and forth between them and the examples as we work through them. By the time you understand each example, you'll find the concepts and their definitions quite natural and intuitive.

To make this exercise maximally instructive, let's suppose that you and your partner have studied the matrix above and, seeing that you're both better off in the outcome represented by the lower-right cell, have formed an agreement to cooperate. You are to commit to refusal first, at which point she will reciprocate. We will refer to a strategy of keeping the agreement as ‘cooperation’, and will denote it in the tree below with ‘C’. We will refer to a strategy of breaking the agreement as ‘defection’, and will denote it on the tree below with ‘D’. As before, you are I and your partner is II. Each node is numbered 1, 2, 3, ... , from top to bottom, for ease of reference in discussion. Here, then, is the tree:

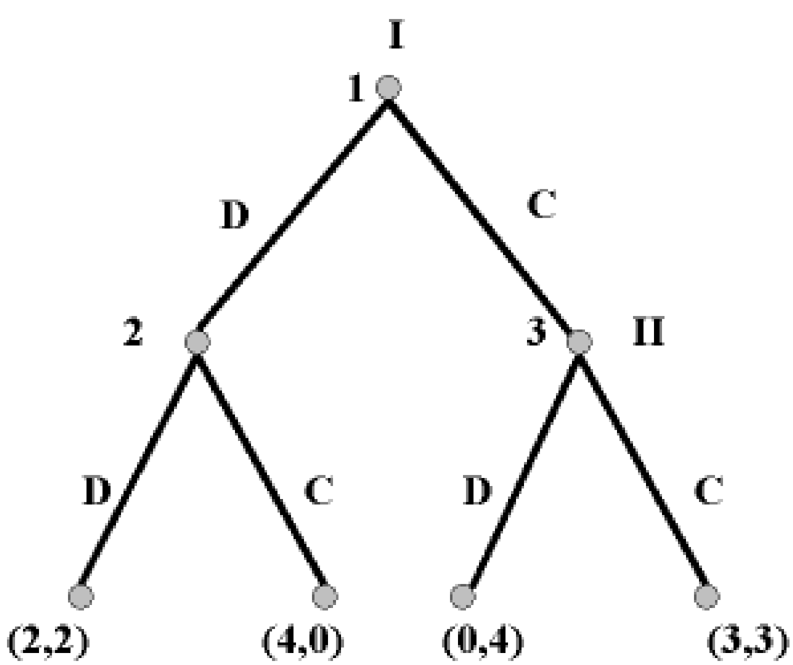


Figure 4

Look first at each of the terminal nodes (those along the bottom). These represent possible outcomes. Each is identified with a assignment of payoffs, just as in the strategic-form game, with I's payoff

appearing first in each set and II's appearing second. Each of the structures descending from the nodes 1, 2 and 3 respectively is a sub-game. We begin our backward-induction analysis -- using a technique called *Zermelo's algorithm* -- with the sub-games that arise last in the sequence of play. If the subgame descending from node 3 is played, then Player II will face a choice between a payoff of 4 and a payoff of 3. (Consult the second number, representing her payoff, in each set at a terminal node descending from node 3.) II earns her higher payoff by playing D. We may therefore replace the entire subgame with an assignment of the payoff (0, 4) directly to node 3, since this is the outcome that will be realized if the game reaches that node. Now consider the subgame descending from node 2. Here, II faces a choice between a payoff of 2 and one of 0. She obtains her higher one, 2, by playing D. We may therefore assign the payoff (2,2) directly to node 2. Now we move to the subgame descending from node 1. (This subgame is, of course, identical to the whole game; all games are subgames of themselves.) You (Player I) now face a choice between outcomes (2,2) and (0,4). Consulting the first numbers in each of these sets, you see that you get your higher payoff -- 2 -- by playing D. D is, of course, the option of confessing. So you confess, and then your partner also confesses, yielding the same outcome as in the strategic-form representation.

What has happened here is that you realize that if you play C (refuse to confess) at node 1, then your partner will be able to maximize her utility by suckering you and playing D. (On the tree, this happens at node 3.) This leaves you with a payoff of 0 (ten years in prison), which you can avoid only by playing D to begin with. You therefore defect from the agreement.

We have thus seen that in the case of the Prisoner's Dilemma, the simultaneous and sequential versions yield the same outcome. This will often not be true, however. In particular, only finite extensive-form (sequential) games of perfect information can be solved using Zermelo's algorithm.

As noted earlier in this section, sometimes we must represent simultaneous moves *within* games that are otherwise sequential. (As we said above, in all such cases the game as a whole will be one of imperfect information, so we won't be able to solve it using Zermelo's algorithm.) We represent such games using the device of *information sets*. Consider the following tree:

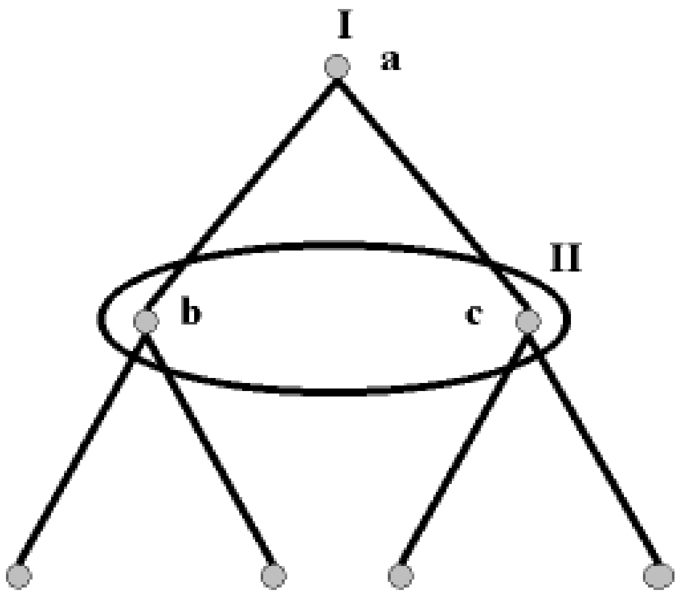


Figure 5

The oval drawn around nodes *b* and *c* indicates that they lie within a common information set. This means that at these nodes players cannot infer back up the path from whence they came; II does not know, in choosing her strategy, whether she is at *b* or *c*. (For this reason, what properly bear numbers in extensive-form games are information sets, conceived as ‘action points’, rather than nodes themselves; this is why the nodes inside the oval are labelled with letters rather than numbers.) Put another way, II, when choosing, does not know what I has done at node *a*. But you will recall from earlier in this section that this is just what defines two moves as simultaneous. We can thus see that the method of representing games as trees is entirely general. If no node after the initial node is alone in an information set on its tree, so that the game has only one subgame (itself), then the whole game is one of simultaneous play. If at least one node shares its information set with another, while others are alone, the game involves both simultaneous and sequential play, and so is still a game of imperfect information. Only if all information sets are inhabited by just one node do we have a game of perfect information.

2.5 Solution Concepts and Equilibria

In the Prisoner's Dilemma, the outcome we've represented as (2, 2), indicating mutual defection, was said to be the ‘solution’ to the game. Following the general practice in economics, game theorists refer to the solutions of games as *equilibria*. Philosophically minded readers will want to pose a conceptual question right here: What is ‘equilibrated’ about some game outcomes such that we are motivated to call them ‘solutions’? When we say that a physical system is in equilibrium, we mean that it is in a *stable* state, one in which all the causal forces internal to the system balance each other out and so leave it ‘at rest’ until and unless it is perturbed by the intervention of some exogenous (that is, ‘external’) force. This what economists have traditionally meant in talking about ‘equilibria’; they read economic systems as being networks of causal relations, just like physical systems, and the equilibria of such systems are then their endogenously stable states. As we will see after discussing evolutionary game theory in a later section, it is possible to maintain this understanding of equilibria in the case of game theory. However, as we noted

in Section 2.1, some people interpret game theory as being an explanatory theory of strategic reasoning. For them, a solution to a game must be an outcome that a rational agent would predict *using the mechanisms of rational computation alone*. Such theorists face some puzzles about solution concepts that aren't so important for the behaviorist. We will be visiting such puzzles and their possible solutions throughout the rest of this article.

It's useful to start the discussion here from the case of the Prisoner's Dilemma because it's unusually simple from the perspective of these puzzles. What we referred to as its 'solution' is the unique *Nash equilibrium* of the game. (The 'Nash' here refers to John Nash, the Nobel Prize-winning mathematician who in Nash (1950) did most to extend and generalize von Neumann & Morgenstern's pioneering work.) Nash equilibrium (henceforth 'NE') applies (or fails to apply, as the case may be) to whole *sets* of strategies, one for each player in a game. A set of strategies is a NE just in case no player could improve her payoff, given the strategies of all other players in the game, by changing her strategy. Notice how closely this idea is related to the idea of strict dominance: no strategy could be a NE strategy if it is strictly dominated. Therefore, if iterative elimination of strictly dominated strategies takes us to a unique outcome, we know we have found the game's unique NE. Now, almost all theorists agree that avoidance of strictly dominated strategies is a *minimum* requirement of rationality. This implies that *if* a game has an outcome that is a unique NE, as in the case of joint confession in the PD, that must be its unique solution. This is one of the most important respects in which the PD is an 'easy' (and atypical) game.

We can specify one class of games in which NE is always not only necessary but *sufficient* as a solution concept. These are finite perfect-information games that are also *zero-sum*. A zero-sum game (in the case of a game involving just two players) is one in which one player can only be made better off by making the other player worse off. (Tic-tac-toe is a simple example of such a game: any move that brings me closer to winning brings you closer to losing, and vice-versa.) We can determine whether a game is zero-sum by examining players' utility functions: in zero-sum games these will be mirror-images of each other, with one player's highly ranked outcomes being low-ranked for the other and vice-versa. In such a game, if I am playing a strategy such that, given your strategy, I can't do any better, and if you are *also* playing such a strategy, then, since any change of strategy by me would have to make you worse off and vice-versa, it follows that our game can have no solution compatible with our mutual rationality other than its unique NE. We can put this another way: in a zero-sum game, my playing a strategy that maximizes my minimum payoff if you play the best you can, and your simultaneously doing the same thing, is just *equivalent* to our both playing our best strategies, so this pair of so-called 'maximin' procedures is guaranteed to find the unique solution to the game, which is its unique NE. (In tic-tac-toe, this is a draw. You can't do any better than drawing, and neither can I, if both of us are trying to win and trying not to lose.)

However, most games do not have this property. It won't be possible, in this one article, to enumerate *all* of the ways in which games can be problematic from the perspective of their possible solutions. (For one thing, it is highly unlikely that theorists have yet discovered all of the possible problems!) However, we can try to generalize the issues a bit.

First, there is the problem that in most non-zero-sum games, there is more than one NE, but not all NE

look equally plausible as the solutions upon which strategically rational players would hit. Consider the strategic-form game below (taken from Kreps (1990), p. 403):

Figure 6

Figure 6

This game has two NE: s1-t1 and s2-t2. (Note that no rows or columns are strictly dominated here. But if Player I is playing s1 then Player II can do no better than t1, and vice-versa; and similarly for the s2-t2 pair.) If NE is our only solution concept, then we shall be forced to say that either of these outcomes is equally persuasive as a solution. However, if game theory is regarded as an explanatory and/or normative theory of strategic reasoning, this seems to be leaving something out: surely rational players with perfect information would converge on s1-t1? (Note that this is *not* like the situation in the PD, where the socially superior situation is unachievable because it is not a NE. In the case of the game above, both players have every reason to try to converge on the NE in which they are better off.)

This illustrates the fact that NE is a relatively (logically) *weak* solution concept, often failing to predict intuitively sensible solutions because, if applied alone, it refuses to allow players to use principles of equilibrium selection that, if not *demanded* by rationality, are at least not *irrational*. Consider another example from Kreps (1990), p. 397:

Figure 7

Figure 7

Here, no strategy strictly dominates another. However, Player I's top row, s1, *weakly* dominates s2, since I does *at least as well* using s1 as s2 for any reply by Player II, and on one reply by II (t2), I does better. So should not the players (and the analyst) delete the weakly dominated row s2? When they do so, column t1 is then strictly dominated, and the NE s1-t2 is selected as the unique solution. However, as Kreps goes on to show using this example, the idea that weakly dominated strategies should be deleted just like strict ones has odd consequences. Suppose we change the payoffs of the game just a bit, as follows:

Figure 8

Figure 8

s2 is still weakly dominated as before; but of our two NE, s2-t1 is now the most attractive for both players; so why should the analyst eliminate its possibility? (Note that this game, again, does *not* replicate the logic of the PD. There, it makes sense to eliminate the most attractive outcome, joint refusal to confess, because both players have incentives to unilaterally deviate from it, so it is not an NE. This is not true of s2-t1 in the present game. You should be starting to clearly see why we called the PD game 'atypical'.) The argument *for* eliminating weakly dominated strategies is that Player 1 may be nervous, fearing that Player II is not completely *sure* to be rational (or that Player II fears that Player I isn't completely rational, or that Player II fears that Player I fears that Player II isn't completely rational, and so

on ad infinitum) and so might play t2 with some positive probability. If the possibility of departures from rationality is taken seriously, then we have an argument for eliminating weakly dominated strategies: Player I thereby insures herself against her worst outcome, s2-t2. Of course, she pays a cost for this insurance, reducing her expected payoff from 10 to 5. On the other hand, we might imagine that the players could communicate before playing the game and agree play *correlated strategies* so as to *coordinate* on s2-t1, thereby removing some, most or all of the uncertainty that encourages elimination of the weakly dominated row s1, and eliminating s1-t2 as a viable NE instead!

Any proposed principle for solving games that may have the effect of eliminating one or more NE from consideration is referred to as a *refinement* of NE. In the case just discussed, elimination of weakly dominated strategies is one possible refinement, since it refines away the NE s2-t1, and correlation is another, since it refines away the other NE, s1-t2, instead. So which refinement is more appropriate as a solution concept? People who think of game theory as an explanatory and/or normative theory of strategic rationality have generated a substantial literature in which the merits and drawbacks of a large number of refinements are debated. In principle, there seems to be no limit on the number of refinements that could be considered, since there may also be no limits on the set of philosophical intuitions about what principles a rational agent might or might not see fit to follow or to fear or hope that other players are following.

Behaviorists take a dim view of much of this activity. They see the job of game theory as being to predict outcomes *given* some distribution of strategic dispositions, and some distribution of expectations about the strategic dispositions of others, that have been shaped by institutional processes and / or evolutionary selection. (See Section 7 for further discussion.) On this view, which NE are viable in a game is determined by the underlying dynamics that equipped players with dispositions *prior to* the commencement of a game. The strategic natures of players are thereby treated as a set of exogenous inputs to the game, just as utility functions are. Behaviorists are thus less inclined to seek *general* refinements of the equilibrium concept itself, at least insofar as these involve the modeling of *more sophisticated* expressions of rationality over and above merely consistent maximization of utility. Behaviorists are often inclined to doubt that the goal of seeking a *general* theory of rationality makes sense as a project. Institutions and evolutionary processes build many environments, and what counts as rational procedure in one environment may not be favoured in another. Economic rationality requires only that agents have consistent preferences, that is, that they not prefer *a* to *b* and *b* to *c* **and** *c* to *a*. A great many arrangements of strategic dispositions are compatible with this minimal requirement, and evolutionary or institutional processes might generate games in any of them. On this view, NE is a robust equilibrium concept because if players evolve their strategic dispositions in settings that are competitive, those who don't do what's optimal given the strategies of others *in that specific environment* will be outcompeted, and so selection will either eliminate them or encourage the learning of new dispositions. There is no more ‘refined’ concept of rationality of which this can be argued to be true *in general*; and so, according to behaviorists, refinements of NE based on refinements of rationality are likely to be of merely occasional interest.

This does not imply that behaviorists abjure all ways of restricting sets of NE to plausible subsets. In particular, they tend to be sympathetic to approaches that shift emphasis from rationality itself onto

considerations of the informational dynamics of games. We should perhaps not be surprised that NE analysis alone often fails to tell us much of interest about strategic-form games (e.g., Figure 6 above), in which informational structure is suppressed. Equilibrium selection issues are often more fruitfully addressed in the context of extensive-form games.

2.6 Modular Rationality and Subgame Perfection

In order to deepen our understanding of extensive-form games, we need an example with more interesting structure than the PD offers.

Consider the game described by this tree:

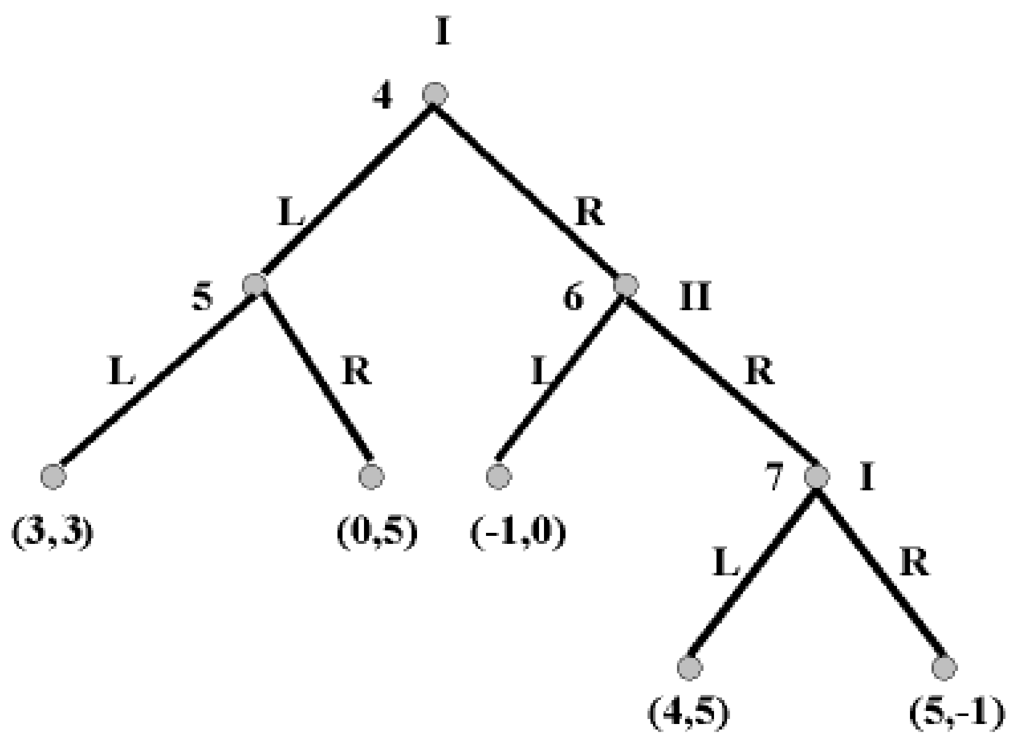


Figure 9

This game is not intended to fit any preconceived situation; it is simply a mathematical object in search of an application. (L and R here just denote ‘left’ and ‘right’ respectively.)

Now consider the strategic form of this game:

		II			
		LL	LR	RL	RR
I	LL	3,3	3,3	0,5	0,5
	LR	3,3	3,3	0,5	0,5
	RL	-1,0	4,5	-1,0	4,5
	RR	-1,0	5,-1	-1,0	5,-1

Figure 10

(If you are confused by this, remember that a strategy must tell a player what to do at *every* information set where that player has an action. Since each player chooses between two actions at each of two information sets here, each player has four strategies in total. The first letter in each strategy designation tells each player what to do if he or she reaches their first information set, the second what to do if their second information set is reached. I.e., LR for Player II tells II to play L if information set 5 is reached and R if information set 6 is reached.) If you examine this matrix, you will discover that (LL, RL) is among the NE. This is a bit puzzling, since if Player I reaches her second information set (7) in the extensive-form game, I would hardly wish to play L there; she earns a higher payoff by playing R at node 7. Mere NE analysis doesn't notice this because NE is insensitive to what happens *off the path of play*. Player I, in choosing L at node 4, ensures that node 7 will not be reached; this is what is meant by saying that it is ‘off the path of play’. In analyzing extensive-form games, however, we *should* care what happens off the path of play, because consideration of this is crucial to what happens *on* the path. For example, it is the fact that Player I *would* play R if node 7 were reached that *would* cause Player II to play L if node 6 were reached, and this is why Player I won't choose R at node 4. We are throwing away information relevant to game solutions if we ignore off-path outcomes, as mere NE analysis does. Notice that this reason for doubting that NE is a wholly satisfactory equilibrium concept in itself has nothing to do with intuitions about rationality, as in the case of the refinement concepts discussed in Section 2.5.

Now apply Zermelo's algorithm to the extensive form of our current example. Begin, again, with the last subgame, that descending from node 7. This is Player I's move, and she would choose R because she prefers her payoff of 5 to the payoff of 4 she gets by playing L. Therefore, we assign the payoff (5, -1) to node 7. Thus at node 6 II faces a choice between (-1, 0) and (5, -1). He chooses L. At node 5 II chooses R. At node 4 I is thus choosing between (0, 5) and (-1, 0), and so plays L. Note that, as in the PD, an outcome appears at a terminal node -- (4, 5) from node 7 -- that is Pareto superior to the NE. Again, however, the dynamics of the game prevent it from being reached.

The fact that Zermelo's algorithm picks out the strategy vector (LR, RL) as the unique solution to the game shows that it's yielding something other than just an NE. In fact, it is generating the game's *subgame*

perfect equilibrium (SPE). It gives an outcome that yields a NE not just in the *whole* game but in every subgame as well. This is an extremely persuasive solution concept because, again unlike the refinements of Section 2.5, it does not demand 'more' rationality of agents, but *less*. The agents, at every node, simply choose the path that brings them the highest payoff *in the subgame emanating from that node*; and, then, in solving the game, they foresee that they will all do that. Agents who proceed in this way are said to be *modular rational*, that is, short-run rational at each step. They do not imagine themselves, by some fancy processes of hyper-rationality, acting against their local preferences for the sake of some wider goal. Note that, as in the PD, this can lead to outcomes which might be regretted from the social point of view. In our current example, Player I would be better off, and Player II no worse off, at the left-hand node emanating from node 7 than at the SPE outcome. But Player I's very modular rationality, and Player II's awareness of this, blocks the socially efficient outcome. If our players wish to bring about the more equitable outcome (4,5) here, they must do so by redesigning their institutions so as to change the structures of the games they play. Merely wishing that they could be hyper-rational in some way does not seem altogether coherent as an approach.

2.7 On Interpreting Payoffs: Morality and Efficiency in Games

Many readers might suppose that the conclusion of the previous section has been asserted on the basis of no adequate defense. Surely, the players might be able to just *see* that outcome (4,5) is socially and morally superior; and since we know they can also see the path of actions that leads to it, who is the game theorist to announce that, within the game they're playing, it's unattainable? In fact, to suggest that hyper-rationality is a will o' the wisp *is* philosophically tendentious, though it is indeed what behaviorists about game theory believe. The reader who seeks a thorough justification for this belief is referred to Binmore (1994, 1998). However, before we just leave matters at a stand-off (here), we must be careful not to confuse what is controversial with the consequences of a simple technical mistake. Consider the Prisoner's Dilemma again. We have seen that in the unique NE of the PD, both players get less utility than they could have through mutual cooperation. This may struck you (as it has struck many commentators) as perverse. Surely, you may think, it simply results from a combination of selfishness and paranoia on the part of the players. To begin with they have no regard for the social good, and then they shoot themselves in the feet by being too untrustworthy to respect agreements.

This way of thinking leads to serious misunderstandings of game theory, and so must be dispelled. Let us first introduce some terminology for talking about outcomes. Welfare economists typically measure social good in terms of *Pareto efficiency*. A distribution of utility β is said to be *Pareto dominant* over another distribution δ just in case from state δ there is a possible redistribution of utility to β such that at least one player is better off in β than in δ and no player is worse off. Failure to move to a Pareto-dominant redistribution is *inefficient* because the existence of β as a logical possibility shows that in δ some utility is being wasted. Now, the outcome (3,3) that represents mutual cooperation in our model of the PD is clearly Pareto dominant over mutual defection; at (3,3) *both* players are better off than at (2,2). So it is true that PDs lead to inefficient outcomes. This was true of our example in Section 2.6 as well.

However, inefficiency should not be associated with immorality. A utility function for a player is

supposed to represent *everything that player cares about*, which may be anything at all. As we have described the situation of our prisoners they do indeed care only about their own relative prison sentences, but there is nothing essential in this. What makes a game an instance of the PD is strictly and only its payoff structure. Thus we could have two Mother Theresa types here, both of whom care little for themselves and wish only to feed starving children. But suppose the original Mother Theresa wishes to feed the children of Calcutta while Mother Juanita wishes to feed the children of Bogota. And suppose that the international aid agency will maximize its donation if the two saints nominate the same city, will give the second-highest amount if they nominate each others' cities, and the lowest amount if they each nominate their own city. Our saints are in a PD here, though hardly selfish or unconcerned with the social good.

To return to our prisoners, suppose that, contrary to our assumptions, they *do* value each other's well-being as well as their own. In that case, this must be reflected in their utility functions, and hence in their payoffs. If their payoff structures are changed, they will no longer be in a PD. But all this shows is that not every possible situation is a PD; it does *not* show that the threat of inefficient outcomes is a special artifact of selfishness. It is the *logic* of the prisoners' situation, not their psychology, that traps them in the inefficient outcome, and if that really *is* their situation then they are stuck in it (barring further complications to be discussed below). Agents who wish to avoid inefficient outcomes are best advised to prevent certain games from arising; the defender of the possibility of hyper-rationality is really proposing that they try to dig themselves out of such games by turning themselves into different kinds of agents.

In general, then, a game is partly *defined* by the payoffs assigned to the players. If a proposed solution involves tacitly changing these payoffs, then this ‘solution’ is in fact a disguised way of changing the subject.

2.8 Trembling Hands

Our last point above opens the way to a philosophical puzzle, one of several that still preoccupy those concerned with the logical foundations of game theory. It can be raised with respect to any number of examples, but we will borrow an elegant one from C. Bicchieri (1993), who also provides the most extensive treatment of the problem found in the literature. Consider the following game:

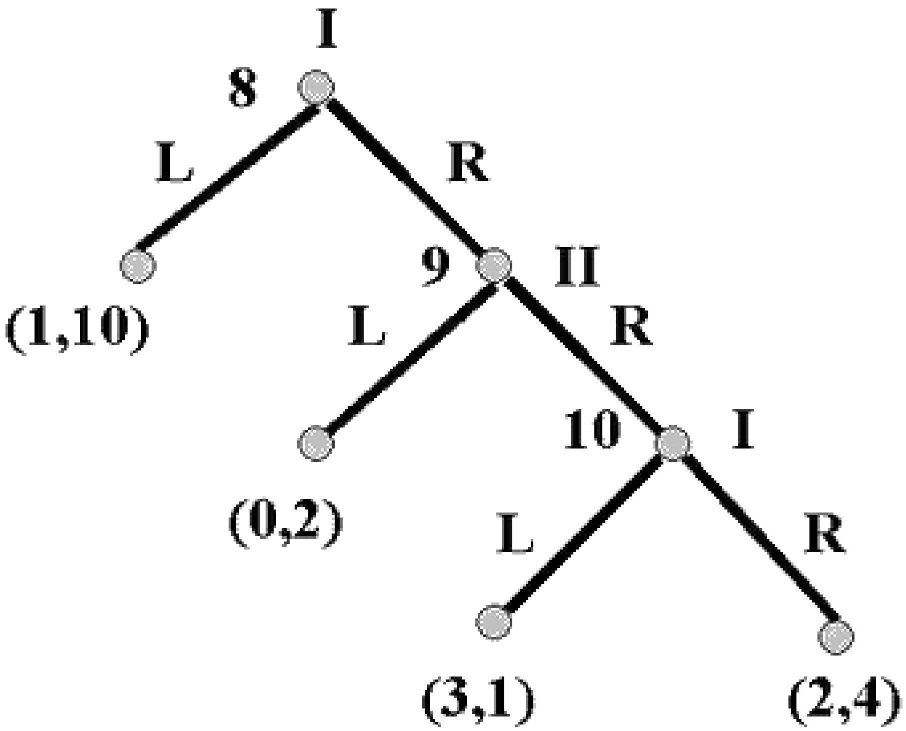


Figure 11

The NE outcome here is at the single leftmost node descending from node 8. To see this, backward induct again. At node 10, I would play L for a payoff of 3, giving II a payoff of 1. II can do better than this by playing L at node 9, giving I a payoff of 0. I can do better than this by playing L at node 8; so that is what I does, and the game terminates without II getting to move. But, now, notice the reasoning required to support this prediction. I plays L at node 8 because she knows that II is rational, and so would, at node 9, play L because II knows that I is rational and so would, at node 10, play L. But now we have the following paradox: I must suppose that II, at node 9, would predict I's rational play at node 10 despite having arrived at a node (9) that could only be reached if I is not rational! If I is not rational then II is not justified in predicting that I will not play R at node 10, in which case it is not clear that II shouldn't play R at 9; and if II plays R at 9, then I is guaranteed of a better payoff then she gets if she plays L at node 8. Both players must use backward induction to solve the game; backward induction requires that I know that II knows that I is rational; but II can solve the game only by using a backward induction argument that takes as a premise the irrationality of I. This is the *paradox of backward induction*.

A standard way around this paradox in the literature is to invoke the so-called ‘trembling hand’ due to Selten (1975). The idea here is that a decision and its consequent act may ‘come apart’ with some nonzero probability, however small. That is, a player might intend to take an action but then slip up in the execution and send the game down some other path instead. If there is even a remote possibility that a player may make a mistake -- that her ‘hand may tremble’ -- then no contradiction is introduced by a player's using a backward induction argument that requires the hypothetical assumption that another player has taken a path that a rational player could not choose. In our example, II could reason about what to do at node 9 conditional on the assumption that I rationally chose L at node 8 but then slipped.

There is a substantial technical literature on this backward-induction paradox, of which Bicchieri (1993)

is the most comprehensive source. (Bicchieri, it should be noted, does *not* endorse an appeal to trembling hands as the appropriate solution. Discussing her particular proposal here would, however, take us too far afield into technicalities. The interested reader should study her book.) The puzzle has been introduced here just in order to point out that refinements of the type discussed in Section 2.6 can be encouraged by more than mere intuitions about the concept of rationality. For if hands may tremble than merely economically rational players *will* be motivated to worry about the probabilities with which apparent departures from rational play will be observed. For example, if my opponent's hand may tremble, then this gives me good reason to avoid the weakly dominated strategy s2 in the third example from Section 2.5. After all, my opponent might promise to play t1 in that game, and I may believe his promise; but if his hand then trembles and a play of t2 results, I get my worst payoff. If I'm risk-averse, then in such situations it would seem that I should stick to weakly dominant strategies.

The reader won't be surprised to hear that the behaviorist has a reply to this. He argues that when a player considers what would happen at nodes reachable only along out-of-equilibrium paths, he need merely refer to *possible worlds* in which the subgames beginning at the nodes in question exist by themselves (i.e., detached from the remainders of the games in which they occur). This might seem like an ad hoc resort; but it is not clear that it is any *less* so than is appeal to trembling hands, which find no independent motivation from elsewhere in economic theory or the theory of rational choice. In any case, appeal to possible worlds is a common strategy in science according to many philosophers. The issue is raised here not with any hope in mind of saying something conclusive about it, but just to give the reader some further sense of the philosophical issues that generate controversy in the foundations of game theory.

3. Uncertainty, Risk and Sequential Equilibria

The games we've modeled to this point have all involved players choosing from amongst *pure strategies*, in which each seeks a single optimal course of action at each node that constitutes a best reply to the rational actions of others. Often, however, a player's utility is optimized through use of a *mixed* strategy, in which she flips a weighted coin amongst several possible actions. (We will see later that there is an alternative interpretation of mixing, not involving randomization at a particular information set; but we will start here from the coin-flipping interpretation and then build on it in Section 3.1.) Mixing is necessary whenever maximization of the player's utility depends on creating *uncertainty* in the expectations of her opponent. Our river-crossing game from Section 1 exemplifies this. As we saw, the puzzle in that game consists in the fact that if the fugitive's reasoning selects a particular bridge as optimal, his pursuer must be assumed to be able to duplicate that reasoning. Thus the fugitive can escape only if his pursuer is surprised. But symmetry of logical reasoning power on the part of the two players ensures that the fugitive can surprise the pursuer only if he succeeds in surprising *himself*.

Suppose that we ignore rocks and cobras for a moment, and imagine that the bridges are equally safe. In this case, the fugitive's best course is to roll a three-sided die, in which each side represents a different bridge (or, more conventionally, a six-sided die in which each bridge is represented by two sides). He must then pre-commit himself to using whichever bridge is selected by this *randomizing device*. Against this strategy, the pursuer's best reply is to also use a three-sided die of her own. The fugitive now has a

2/3 probability of escaping and the pursuer a 1/3 probability of catching him. The fugitive cannot improve on these odds if the pursuer is randomizing, since to favour one bridge merely provides the pursuer with a pattern that can be exploited. Identical reasoning applies to the pursuer. Therefore, the two randomizing strategies are best replies to one another, and are therefore in Nash equilibrium.

Now let us re-introduce the parametric factors, that is, the falling rocks at bridge #2 and the cobras at bridge #3. Again, suppose that bridge #3 (cobras) is four times more dangerous for the fugitive than bridge #2 (rocks), while bridge #2 is 20% more dangerous than bridge #1 (unobstructed). We can solve this new game *if* we make certain assumptions about the two players' utility functions. Suppose that Player 1, the fugitive, cares only about living or dying (preferring life to death) while the pursuer simply wishes to be able to report that the fugitive is dead, preferring this to having to report that he got away. (In other words, neither player cares about *how* the fugitive lives or dies.) In this case, the fugitive simply takes his original randomizing formula and weights it according to the different levels of parametric danger at the three bridges. Each bridge should be thought of as a *lottery* over the fugitive's possible outcomes, in which each lottery has a different *expected payoff* in terms of the items in his utility function.

Consider matters from the pursuer's point of view. She will be using her NE strategy when she chooses the mix of probabilities over the three bridges that makes the fugitive indifferent amongst them as crossing points. The bridge with rocks is 1.2 times more dangerous for him than the safe bridge. Therefore, he will be indifferent between the two when the pursuer is 1.2 times more likely to be waiting at the rocky bridge than the safe bridge. The cobra bridge is 4 times more dangerous for the fugitive than the rocky bridge. Therefore, he will be indifferent between these two bridges when the pursuer's probability of waiting at the rocky bridge is 4 times higher than the probability that she is at the cobra bridge. This gives us an indifference ratio amongst the bridges of 4.8:4:1. Suppose that no mixed strategy involving use of the cobra bridge is dominated by a mixed strategy involving use of only the other two bridges (i.e., that the cobra bridge is not so dangerous that the hunter can't make the fugitive indifferent between its use and that of the other at any value). Since the probabilities must sum to 100, we find the probability with which the pursuer waits at each bridge by solving the following equation:

$$4.8x + 4x + x = 100$$

x here equals 10.2. So at Nash equilibrium the pursuer must appear at the cobra bridge with a probability of .102, at the rocky bridge with a probability of .408, and the safe bridge with a probability of .489. These probabilities tell her how to weight her die before throwing it. The fugitive's NE strategy is, of course, as before: knowing that the pursuer will randomize so as to leave him indifferent, he will choose each bridge with equal probability.

We were able to solve this game straightforwardly because we set the utility functions in such a way as to make it *zero-sum*, or *strictly competitive*. That is, every gain in expected utility by one player represents a precisely symmetrical loss by the other. However, this condition may often not hold. Suppose now that the utility functions are more complicated. The pursuer most prefers an outcome in which she shoots the fugitive and so claims credit for his apprehension to one in which he dies of rockfall or snakebite; and she

prefers this second outcome to his escape. The fugitive prefers a quick death by gunshot to the pain of being crushed or the terror of an encounter with a cobra. Most of all, of course, he prefers to escape. We cannot solve this game, as before, simply on the basis of knowing the players' ordinal utility functions, since the *intensities* of their respective preferences will now be relevant to their strategies.

Prior to the work of von Neumann & Morgenstern (1947), situations of this sort were inherently baffling to analysts. This is because utility does not denote a hidden psychological variable such as *pleasure*. As we discussed in Section 2.1, utility is merely a measure of relative behavioural dispositions given certain consistency assumptions about relations between preferences and choices. It therefore makes no sense to imagine comparing our players' *cardinal* -- that is, intensity-sensitive -- preferences with one another's, since there is no independent, interpersonally constant yardstick we could use. How, then, can we model games in which cardinal information is relevant? After all, modeling games requires that all players' utilities be taken simultaneously into account, as we've seen.

A crucial aspect of von Neumann & Morgenstern's (1947) work was the solution to this problem. Here, we will provide a brief outline of their ingenious technique for building cardinal utility functions out of ordinal ones. It is emphasized that what follows is merely an *outline*, so as to make cardinal utility non-mysterious to you as a student who is interested in knowing about the philosophical foundations of game theory, and about the range of problems to which it can be applied. Providing a manual you could follow in *building* your own cardinal utility functions would require many pages. Fortunately, such manuals are available in many textbooks. In any case, if you are a philosophy student you may not wish to attempt this until you've taken a course in probability theory.

Suppose we have an agent whose ordinal utility function is known. Indeed, suppose that it's our river-crossing fugitive. Let's assign him the following ordinal utility function:

Escape » 4

Death by shooting » 3

Death by rockfall » 2

Death by snakebite » 1

Now, we know that his preference for escape over *any* form of death is likely to be stronger than his preference for, say, shooting over snakebite. This should be reflected in his choice behaviour in the following way. In a situation such as the river-crossing game, he should be willing to run greater risks to increase the relative probability of escape over shooting than he is to increase the relative probability of shooting over snakebite. This bit of logic is the crucial insight behind von Neumann & Morgenstern's (1947) solution to the cardinalization problem.

Begin by asking our agent to pick, from the available set of outcomes, a *best* one and a *worst* one. ‘Best’

and ‘worst’ are defined in terms of rational choice: a rational agent always chooses so as to maximize the probability of the best outcome -- call this **W** -- and to minimize the probability of the worst outcome -- call this **L**. Now consider prizes intermediate between **W** and **L**. We find, for a set of outcomes containing such prizes, a lottery over them such that our agent is indifferent between that lottery and a lottery including only **W** and **L**. In our example, this would be a lottery having shooting and rockfall as its possible outcomes. Call this lottery **T** . We define a utility function $q = u(\mathbf{T})$ such that if q is the expected prize in **T** , the agent is indifferent between winning **T** and winning a lottery in which **W** occurs with probability $u(\mathbf{T})$ and **L** occurs with probability $1 - u(\mathbf{T})$.

We now construct a *compound lottery* **T*** over the outcome set {**W**, **L**} such that the agent is indifferent between **T** and **T***. A compound lottery is one in which the prize in the lottery is another lottery. This makes sense because, after all, it is still **W** and **L** that are at stake for our agent in both cases; so we can then analyze **T*** into a simple lottery over **W** and **L**. Call this lottery **r**. It follows from transitivity that **T** is equivalent to **r**. (Note that this presupposes that our agent does not gain utility from the complexity of her gambles.) The rational agent will now choose the action that maximizes the probability of winning **W**. The mapping from the set of outcomes to $u(\mathbf{r})$ is a *von Neumann-Morgenstern utility function* (VNMuf).

What exactly have we done here? We've simply given our agent choices over lotteries, instead of over prizes directly, and observed how much extra risk he's willing to run to increase the chances of winning escape over snakebite relative to getting shot or clobbered with a rock. A VNMuf yields a *cardinal*, rather than an ordinal, measure of utility. Our choice of endpoint-values, **W** and **L**, is arbitrary, as before; but once these are fixed the values of the intermediate points are determined. Therefore, the VNMuf *does* measure the relative preference intensities of a single agent. However, since our assignment of utility values to **W** and **L** *is* arbitrary, we can't use VNMufs to compare the cardinal preferences of one agent with those of another. Furthermore, since we are using a *risk-metric* as our measuring instrument, the construction of the new utility function depends on assuming that our agent's *attitude to risk itself* stays constant from one comparison of lotteries to another. This seems reasonable for a single agent in a single game-situation. However, two agents in one game, or one agent under different sorts of circumstances, may display very different attitudes to risk. Perhaps in the river-crossing game the pursuer, whose life is not at stake, will enjoy gambling with her glory while our fugitive is cautious. In general, a *risk-averse* agent prefers a guaranteed prize to its equivalent expected value in a lottery. A *risk-loving* agent has the reverse preference. A *risk-neutral* agent is indifferent between these options. In analyzing the river-crossing game, however, we don't *have to* be able to compare the pursuer's cardinal utilities with the fugitive's. Both agents, after all, can find their NE strategies if they can estimate the probabilities each will assign to the actions of the other. This means that each must know both VNMufs; but neither need try to comparatively value the outcomes over which they're gambling.

We can now fill in the rest of the matrix for the bridge-crossing game that we started to draw in Section 2. If all that the fugitive cares about is life and death, but not the manner of death, and if all the hunter cares about is preventing the fugitive from escaping, then we can now interpret both utility functions cardinally. This permits us to assign expected utilities, expressed by multiplying the original payoffs by the relevant probabilities, as outcomes in the matrix. Suppose that the hunter waits at the cobra bridge with probability x and at the rocky bridge with probability y. Since her probabilities across the three bridges must sum to

1, this implies that she must wait at the safe bridge with probability $1 - (x + y)$. Then, continuing to assign the fugitive a payoff of 0 if he dies and 1 if he escapes, and the hunter the reverse payoffs, our complete matrix is as follows:

Figure 12

Figure 12

We can now read the following facts about the game directly from the matrix. No rows or columns strictly or weakly dominate any others. Therefore, the game's NE must be in mixed strategies.

3.1 Beliefs

Here is an odd feature of our analysis of the river-crossing game above. In the situation as we have imagined it, the fugitive knows that the hunter is least likely to cross at the cobra bridge. Yet we have told her to select a bridge at which to wait by flipping a coin. This process gives a non-zero probability of selecting the cobra bridge. Yet, surely, if the coin selects the cobra bridge it cannot be rational for the hunter to do as it directs. Won't she feel like a bit of an idiot standing there, knowing that the fugitive is probably waltzing across the safe bridge, or the relatively safe rocky bridge? Note that she will *not* feel torn in this situation if she plays this game with fugitives on a regular basis (and they know this). In that case, it is perfectly reasonable that she must *sometimes* wait at the cobra bridge, lest all fugitives be able to do better than their odds at NE by using the cobra bridge more often. Now, as far as the behaviorist is concerned this ends the matter. If the hunter and the fugitive have regularly played games that structurally *resemble* this river-crossing game, then selection pressures will have encouraged habits in them that lead them both to play its NE strategies *and to sincerely rationalize doing so* by means of some satisfying story or other. If neither party has ever been in a situation like this, and if their biological and/or cultural ancestors haven't either, and if neither is concerned with revealing information to opponents in expected future situations of this sort, then their behavior should be predicted not by a game theorist but by friends of theirs who are familiar with their personal idiosyncracies. Behaviorists are happy to recognize that game theory isn't useful for every decision problem, or even every strategic decision problem, that comes along.

However, the philosopher who wants game theory to serve as a descriptive and/or normative theory of strategic rationality cannot rest content with this answer. He must find a satisfying line of advice for the players even when their game is alone in the universe of strategic problems. No such advice can be given that is *uncontroversially* satisfactory -- behaviorists, after all, are often behaviorists *because* they aren't satisfied by any available approach here -= but there is a way of handling the matter that many game theorists have found worthy of detailed pursuit. This involves the computation of *equilibria in beliefs*.

In fact, the behaviorist needs the concept of equilibrium in beliefs too, but for different purposes. As we've seen, the concept of NE sometimes doesn't go deep enough as an analytical instrument to tell us all that we think might be important in a game. Thus even behaviorists who aren't impressed with the project of refinements make frequent use of the concept of subgame-perfect equilibrium (SPE) as discussed in

Section 2.6. But now consider the three-player imperfect-information game below (taken from Kreps (1990), p. 426):

Figure 13

Figure 13

One of the NE of this game is Lr_2l_3. This is because if Player I plays L, then Player II playing r_2 has no incentive to change strategies because her only node of action, 12, is off the path of play. But this NE seems to be purely technical; it makes little sense as a solution. This reveals itself in the fact that if the game beginning at node 14 could be treated as a subgame, Lr_2l_3 would not be an SPE. Whenever she *does* get a move, Player II should play l_2. But if Player II is playing l_2 then Player I should switch to R. In that case Player III should switch to r_3, sending Player II back to r_2. And here's a new, ‘sensible’, NE: Rr_2r_3. I and II in effect play ‘keepaway’ from III; and so that's what we'll name this game.

This NE is ‘sensible’ in just the same way that a SPE outcome in a perfect-information game is more sensible than other non-SPE NE. However, we can't select it by applying Zermelo's algorithm. Because nodes 13 and 14 fall inside a common information set, Keepaway has only one subgame (namely, the whole game). We need a ‘cousin’ concept to SPE that we can apply in cases of imperfect information, and we need a new solution procedure to replace Zermelo's algorithm for such games.

Notice what Player III in Keepaway is wondering about as he selects his strategy. "Given that I get a move," he asks himself, "was my action node reached from node 11 or from node 12?" What, in other words, are the *conditional probabilities* that III is at node 13 or 14 given that he has a move? Now, if conditional probabilities are what III wonders about, then what Players I and II must make conjectures about when they select *their* strategies are III's *beliefs* about these conditional probabilities. In that case, I must conjecture about II's beliefs about III's beliefs, and III's beliefs about II's beliefs and so on. The relevant beliefs here are not merely strategic, as before, since they are not just about what players will *do* given a set of payoffs and game structures, but about what they think makes sense given some understanding or other of conditional probability.

What beliefs about conditional probability is it reasonable for players to expect from each other? The normative theorist might insist on whatever the best mathematicians have discovered about the subject. Clearly, however, if this is applied then a theory of games that incorporated it would not be descriptively true of most people. The behaviorist will insist on imposing only behavioral habits that a process of natural selection might build into its products. Perhaps some actual or possible creatures might observe habits that respect *Bayes's rule*, which is the minimal true generalization about conditional probability that an agent could know if it knows any such generalizations at all. Adding more sophisticated knowledge about conditional probability amounts to refining the concept of equilibrium-in-belief, just as some game theorists like to refine NE. You can imagine what behaviorists think of *that* project!

Here, we will restrict our attention to the least refined equilibrium-in-belief concept, that obtained when

we require players to reason in accordance with Bayes's rule. Bayes's rule tells us how to compute the probability of an event F given information E (written 'pr(F/E)'):

pr(F/E) = [pr(E/F) X pr(F)] / pr(E)

We will henceforth assume that players do not hold beliefs inconsistent with this equality.

We may now define a *sequential equilibrium*. A SE has two parts: (1) a strategy profile § for each player, as before, and (2) a *system of beliefs* μ for each player. μ assigns to each information set *h* a probability distribution over the nodes *x* in *h*, with the interpretation that these are the beliefs of player *i(h)* about where in his information set he is, given that information set *h* has been reached. Then a sequential equilibrium is a profile of strategies § and a system of beliefs μ consistent with Bayes's rule such that starting from every information set *h* in the tree player *i(h)* plays optimally from then on, given that what he believes to have transpired previously is given by μ(*h*) and what will transpire at subsequent moves is given by §.

We now demonstrate the concept by application to Keepaway. Consider again the uninteresting NE Lr_2l_3. Suppose that Player III assigns pr(1) to her belief that if she gets a move she is at node 13. Then Player II, given a consistent μ(II), must believe that III will play l_3, in which case her only SE strategy is l_2. So although Lr_2l_3 is a NE, it is not a SE. This is of course what we want.

The use of the consistency requirement in this example is somewhat trivial, so consider now a second case (also taken from Kreps (1990), p. 429):

Figure 14

Figure 14

Suppose that I plays L, II plays l_2 and III plays l_3. Suppose also that μ(II) assigns pr(.3) to node 16. In that case, l_2 is not a SE strategy for II, since l_2 returns an expected payoff of .3(4) + .7(2) = 2.6, while r_2 brings an expected payoff of 3.1. Notice that if we fiddle the strategy profile for player III while leaving everything else fixed, l_2 could *become* a SE strategy for II. If §(III) yielded a play of l_3 with pr(.5) and r_3 with pr(.5), then if II plays r_2 his expected payoff would now be 2.2, so Ll_2l_3 would be a SE. Now imagine setting μ(III) back as it was, but change μ(II) so that II thinks the conditional probability of being at node 16 is greater than .5; in that case, l_2 is again not a SE strategy.

The idea of SE is hopefully now clear. We can apply it to the river-crossing game in a way that avoids the necessity for the hunter to flip any coins of we modify the game a bit. Suppose now that II can change bridges twice during the fugitive's passage, and will catch him just in case she meets him as he leaves the bridge. Then the hunter's SE strategy is to divide her time at the three bridges in accordance with the proportion given by the equation in the third paragraph of Section 3 above.

It must be noted that since Bayes's rule cannot be applied to events with probability 0, its application to SE requires that players assign non-zero probabilities to all actions available in trees. This requirement is captured by supposing that all strategy profiles be *strictly mixed*, that is, that every action at every information set be taken with positive probability. You will see that this is just equivalent to supposing that all hands sometimes tremble. A SE is said to be *trembling-hand perfect* if all strategies played at equilibrium are best replies to strategies that are strictly mixed. You should also not be surprised to be told that no weakly dominated strategy can be trembling-hand perfect, since the possibility of trembling hands gives players the most persuasive reason for avoiding such strategies.

4. Repeated Games and Coordination

So far we've restricted our attention to *one-shot* games, that is, games in which players' strategic concerns extend no further than the terminal nodes of their single interaction. However, games are often played with *future* games in mind, and this can significantly alter their outcomes and equilibrium strategies. Our topic in this section is *repeated games*, that is, games in which sets of players expect to face each other in similar situations on multiple occasions. We approach these first through the limited context of repeated prisoner's dilemmas.

We've seen that in the one-shot PD the only NE is mutual defection. This may no longer hold, however, if the players expect to meet each other again in future PDs. Imagine that four firms, all making widgets, agree to maintain high prices by jointly restricting supply. (That is, they form a cartel.) This will only work if each firm maintains its agreed production quota. Typically, each firm can maximize its profit by departing from its quota while the others observe theirs, since it then sells more units at the higher market price brought about by the almost-intact cartel. In the one-shot case, all firms would share this incentive to defect and the cartel would immediately collapse. However, the firms expect to face each other in competition for a long period. In this case, each firm knows that if it breaks the cartel agreement, the others can punish it by underpricing it for a period long enough to more than eliminate its short-term gain. Of course, the punishing firms will take short-term losses too during their period of underpricing. But these losses may be worth taking if they serve to reestablish the cartel and bring about maximum long-term prices.

One simple, and famous (but *not*, contrary to widespread myth, necessarily optimal) strategy for preserving cooperation in repeated PDs is called *tit-for-tat*. This strategy tells each player to behave as follows:

i. Always cooperate in the first round.
ii. Thereafter, take whatever action your opponent took in the previous round.

A group of players *all* playing tit-for-tat will never see any defections. Since, in a population where others play tit-for-tat, tit-for-tat is the rational response for each player, everyone playing tit-for-tat is a NE. You may frequently hear people who know a *little* (but not enough) game theory talk as if this is the end of the story. It is not.

There are two complications. First, the players must be uncertain as to when their interaction ends. Suppose the players know when the last round comes. In that round, it will be rational for players to defect, since no punishment will be possible. Now consider the second-last round. In this round, players also face no punishment for defection, since they know they will defect in the last round anyway. So they defect in the second-last round. But this means they face no threat of punishment in the third-last round, and defect there too. We can simply iterate this backwards through the game tree until we reach the first round. Since cooperation is not rational in that round, tit-for-tat is no longer a rational strategy, and we get the same outcome -- mutual defection -- as in the one-shot PD. Therefore, cooperation is only possible in repeated PDs where the expected number of repetitions is indeterminate. (Of course, this does apply to many real-life games.)

But now we introduce a second complication. Suppose that players' ability to distinguish defection from cooperation is imperfect. Consider our case of the widget cartel. Suppose the players observe a fall in the market price of widgets. Perhaps this is because a cartel member cheated. Or perhaps it has resulted from an exogenous drop in demand. If tit-for-tat players mistake the second case for the first, they will defect, thereby setting off a chain-reaction of mutual defections from which they can *never* recover, since every player will reply to the first encountered defection with defection, thereby begetting further defections, and so on.

If players know that such miscommunication is possible, they must resort to more sophisticated strategies. In particular, they must be prepared to sometimes risk following defections with cooperation in order to test their inferences. However, they mustn't be *too* forgiving, lest other players find it rationally optimal to exploit them through deliberate defections. In general, sophisticated strategies have a problem. Because they are more difficult for other players to infer, their use increases the probability of miscommunication. But miscommunication is what causes repeated-game cooperative equilibria to unravel in the first place! The moral of this is that PDs, even repeated ones, are very difficult to escape from. Rational players do best trying to *avoid* situations that are PDs, rather than relying on cunning strategems for trying to get out of them.

Real, complex, social and political dramas are seldom straightforward instantiations of simple games such as PDs. Russell Hardin (1995) offers an analysis of two recent, very real (and very tragic) political cases, the Yugoslavian civil war of 1991-95, and the 1994 Rwandan genocide, as PDs that were nested inside *coordination games*. A coordination game occurs whenever the utility of two or more players is maximized by their doing the same thing, and where such correspondence is more important to them than *what*, in particular, they both do. A standard example arises with rules of the road: ‘All drive on the left’ and ‘All drive on the right’ are both outcomes that are NEs, and neither is more efficient than the other. In games of ‘pure’ coordination, it doesn't even help to use more selective equilibrium criteria. For example, suppose that we require our players to reason in accordance with Bayes's rule (see Section 3 above). In these circumstances, any strategy that is a best reply to any vector of mixed strategies available in NE is said to be *rationalizable*. That is, a player can find a set of systems of beliefs for the other players such that any history of the game along an equilibrium path is consistent with that set of systems. Pure coordination games are characterized by non-unique vectors of rationalizable strategies. In such

situations, players may try to predict equilibria by searching for *focal points*, that is, features of some strategies that they believe will be salient to other players, and that they believe other players will believe to be salient to them. Unfortunately, in many of the social and political games played by people (and some other animals), the biologically shallow properties by which people sort themselves into racial and ethnic groups serve highly efficiently as such features. Hardin's analysis of recent genocides relies on this fact.

According to Hardin, neither the Yugolslavian nor the Rwandan disasters were PDs to begin with. That is, in neither situation, on either side, did most people begin by preferring the destruction of the other to mutual cooperation. However, the deadly logic of coordination, deliberately abetted by self-serving politicians, dynamically *created* PDs. Some individual Serbs (Hutus) were encouraged to perceive their individual interests as best served through identification with Serbian (Hutu) group-interests. That is, they found that some of their circumstances, such as those involving competition for jobs, had the form of coordination games. They thus acted so as to create situations in which this was true for other Serbs (Hutus) as well. Eventually, once enough Serbs (Hutus) identified self-interest with group-interest, the identification became almost universally *correct*, because (1) the most important goal for each Serb (Hutu) was to do roughly what every other Serb (Hutu) would, and (2) the most distinctively *Serbian* thing to do, the doing of which permitted coordination, was to exclude Croats (Tutsi). That is, strategies involving such exclusionary behavior were selected as a result of having efficient focal points. This situation made it the case that an individual -- and individually threatened -- Croat's (Tutsi's) self-interest was best maximized by coordinating on assertive Croat (Tutsi) group-identity, which further increased pressures on Serbs (Hutus) to coordinate, and so on. Note that it is not an aspect of this analysis to suggest that Serbs or Hutus started things; the process could have been (even if it wasn't in fact) perfectly reciprocal. But the outcome is ghastly: Serbs and Croats (Hutus and Tutsis) seem progressively more threatening to each other as they rally together for self-defense, until both see it as imperative to preempt their rivals and strike before being struck. If Hardin is right -- and the point here is not to claim that he *is*, but rather to point out the worldly importance of determining which games agents are in fact playing -- then the mere presence of an external enforcer (NATO?) would not have changed the game, pace the Hobbesian analysis, since the enforcer could not have threatened either side with anything worse than what each feared from the other. What was needed was recalibration of evaluations of interests, which (arguably) happened in Yugoslavia when the Croatian army began to decisively win, at which point Bosnian Serbs decided that their self/group interests were best served by the arrival of NATO peacekeepers. The Rwandan conflict, meanwhile, drags on in the neighbouring country (the Congo) to which military and political developments have shifted it.

Of course, it is not the case that most repeated games lead to disasters. The biological basis of friendship in people and other animals is probably *partly* a function of the logic of repeated games. The importance of payoffs achievable through cooperation in future games leads those who expect to interact in them to be less selfish than temptation would suggest in present games. Furthermore, cultivating shared interests and sentiments provides networks of focal points around which coordination can be facilitated.

5. Commitment

In some games, players can improve their outcomes by taking actions that makes it impossible for them to take what would be her best actions in the corresponding simultaneous-move games. Such actions are referred to as *commitments,* and they can serve as alternatives to external enforcement in games which would otherwise settle on Pareto-inefficient equilibria.

Consider the following hypothetical example (which is *not* a PD). Suppose you own a piece of land adjacent to mine, and I'd like to buy it so as to expand my lot. Unfortunately, you don't want to sell at the price I'm willing to pay. If we move simultaneously -- you post a selling price and I independently give my agent an asking price -- there will be no sale. So I might try to change your incentives by playing an opening move in which I announce that I'll build a putrid-smelling sewage disposal plant on my land beside yours unless you sell, thereby lowering your price. I've now turned this into a sequential-move game. However, this move so far changes nothing. If you refuse to sell in the face of my threat, it is then not in my interest to carry it out, because in damaging you I also damage myself. Since you know this you should ignore my threat. My threat is *incredible,* a case of cheap talk.

However, I could make my threat credible by *committing* myself. I could sign a contract with some farmers promising to supply them with treated sewage (fertilizer) from my plant, but including an escape clause in the contract releasing me from my obligation only if I can double my lot size and so put it to some other use. Now my threat is credible: if you don't sell, I'm committed to building the sewage plant. Since you know this, you now have an incentive to sell me your land in order to escape its ruination.

This sort of case exposes one of many fundamental differences between the logic of non-parametric and parametric maximization. In parametric situations, an agent can never be made worse off by having more options. But where circumstances are non-parametric, one agent's strategy can be influenced in another's favour if options are visibly restricted. Cortez's burning of his boats (see Section 1) is, of course, an instance of this, one which serves to make the usual metaphor literal.

Another example will illustrate this, as well as the applicability of principles across game-types. Here we will build an imaginary situation that is not a PD -- since only one player has an incentive to defect -- but which is a social dilemma insofar as its NE in the absence of commitment is Pareto-inferior to an outcome that is achievable *with* a commitment device. Suppose that two of us wish to poach a rare antelope from a national park in order to sell the trophy. One of us must flush the animal down towards the second person, who waits in a blind to shoot it and load it onto a truck. You promise, of course, to share the proceeds with me. However, your promise is not credible. Once you've got the buck, you have no reason not to drive it away and pocket the full value from it. After all, I can't very well complain to the police without getting myself arrested too. But now suppose I add the following opening move to the game. Before our hunt, I rig out the truck with an alarm that can be turned off only by punching in a code. Only I know the code. If you try to drive off without me, the alarm will sound and we'll both get caught. You, knowing this, now have an incentive to wait for me. What is crucial to notice here is that you *prefer* that I rig up the alarm, since this makes your promise to give me my share credible. If I don't do this, leaving your promise *in*credible, we'll be unable to agree to try the crime in the first place, and both of us will lose our shot at the profit from selling the trophy. Thus, you benefit from my binding you.

We may now combine our analysis of PDs and commitment devices in discussion of the application that first made game theory famous outside of the academic community. The nuclear stand-off between the superpowers during the Cold War was exhaustively studied by the first generation of game theorists, many of whom worked for the US military. (See [Poundstone 1992] for historical details.) Both the USA and the USSR maintained the following policy. If one side launched a first strike, the other threatened to answer with a devastating counter-strike. This pair of reciprocal strategies, which by the late 1960s would effectively have meant blowing up the world, was known as ‘Mutually Assured Destruction’, or ‘MAD’. Game theorists objected that MAD was mad, because it set up a Prisoner's Dilemma as a result of the fact that the reciprocal threats were incredible. Suppose the USSR launches a first strike against the USA. At that point, the American President faces the following situation. His country is already destroyed. He doesn't bring it back to life by now blowing up the world, so he has no incentive to carry out his threat, which has now manifestly failed to achieve its point. Since the Russians know this, they should ignore the threat and strike first! Of course, the Americans are in exactly the same position. Each power will recognize this incentive on the part of the other, and so will anticipate an attack if they don't preempt it. What we should therefore expect, because it is the only NE of the game, is a race between the two powers to be the first to attack.

This game-theoretic analysis caused genuine consternation and fear on both sides during the Cold War, and produced some rather bizarre attempts at setting up strategic commitment devices. President Nixon, for example, had the CIA try to convince the Russians that he was insane, so that they'd believe that he'd launch a retaliatory strike even when it was no longer in his interest to do so. Similarly, the Soviet KGB leaked fabricated medical reports exaggerating Brezhnev's senility with the same end in mind. Ultimately, the Americans broke this deadly symmetry by using a ‘doomsday device’. They equipped a worldwide fleet of submarines with enough missiles to destroy the USSR, and arranged their communications technology in such a way that the President could not be *sure* to be able to reach the submarines and cancel their orders to attack if any Soviet missile crossed the radar ‘trigger line’. Of course, this strategy depended on making sure that the Russians were aware of the device. In Stanley Kubrick's classic film *Dr. Strangelove,* the world is destroyed by accident because the Russians build a doomsday machine *but then keep it a secret*! As a result, when a mad American colonel launches missiles at Russia on his own accord, and the American President tries to convince his Soviet counterpart that the attack was unintended, the Russian President sheepishly tells him about the secret doomsday machine. Now the two Presidents can do nothing but watch in dismay as the world is blown up -- due to a game-theoretic mistake.

Commitment can sometimes be secured through the value to a player of her own *reputation*. For example, a government tempted to negotiate with terrorists to secure the release of hostages on a particular occasion may commit to a ‘line in the sand’ strategy for the sake of maintaining a reputation for toughness intended to reduce terrorists' incentives to launch future attacks. A different sort of example is provided by Qantas Airlines of Australia. Qantas has never suffered an accident, and makes much of this in its advertising. This means that its planes probably *are* safer than average even if the initial advantage was merely a statistical artifact, because the value of its ability to claim a perfect record rises the longer it lasts, and so gives the airline continuous incentives to incur greater costs in safety assurance.

Certain conditions must hold if reputation effects are to underwrite commitment. First, the game must be repeated, with uncertainty as to which round is the last one. The repeated PD can be used to illustrate the importance of this principle. Cooperation can be the dominant strategy in a repeated PD because a player can gain more from his reputation for cooperation, through inducing expectations of cooperation in others, than he can gain through defection in a single round. However, if the players know in advance which round will be their last, this equilibrium unravels. In the last round reputation no longer has a value, and so both players will defect. In the second-last round, the players know they will defect in the last round, so reputation becomes worthless here too and they will again defect. This makes reputation worthless in the third-last round, and so on. The process iterates back to the first round, so no cooperation ever occurs. This point can be generalized to state the most basic condition on the possibility for using reputation effects as commitment devices: the value of the reputation must be greater to its cultivator than the value to him of sacrificing it in *any* particular round. Thus players may establish commitment by reducing the value of each round so that the temptation to defect in any round never gets high enough to make it rational. For example, parties to a contract may exchange their obligations in small increments to reduce incentives on both sides to renege. Thus builders in construction projects may be paid in weekly or monthly installments. Similarly, the International Monetary Fund often dispenses loans to governments in small tranches, thereby reducing governments' incentives to violate loan conditions once the money is in hand; and governments may actually prefer such arrangements in order to remove domestic political pressure for non-compliant use.

6. Evolutionary Game Theory

Gintis (2000) has recently felt justified in stating baldly that "game theory is a universal language for the unification of the behavioral sciences." This may seem an extraordinary thing to say, but it is entirely plausible. Binmore (1998) has modeled social history as a series of convergences on increasingly efficient equilibria in commonly encountered transaction games, interrupted by episodes in which some people try to shift to new equilibria by moving off stable equilibrium paths, resulting in periodic catastrophes. (Stalin, for example, tried to shift his society to a set of equilibria in which people cared more about the future industrial power and equitability of their society than they cared about their own lives. He was not successful; however, his efforts certainly created a situation in which, for a few decades, many Soviet people attached far less importance to *other people's* lives than usual.) Furthermore, applications of game theory to behavioral topics extend well beyond the political arena.

In 1969, for example, the philosopher David Lewis (1969) published *Convention*, in which the conceptual framework of game-theory was applied to one of the fundamental issues of twentieth-century epistemology, the nature and extent of conventions governing semantics and their relationship to the justification of propositional beliefs. This book stands as one of the classics of analytic philosophy, and its stock is presently rising still further as we become more aware of the significance of the trail it blazed. The basic insight can be captured using a simple example. The word ‘chicken’ denotes chickens and ‘ostrich’ denotes ostriches. We would not be better or worse off if ‘chicken’ denoted ostriches and ‘ostrich’ denoted chickens; however, we *would* be worse off if half of us used the pair of words the first way and half the second, or if all of us randomized between them to refer to flightless birds generally.

This insight, of course, well preceded Lewis; but what he recognized is that this situation has the logical form of a coordination game. Thus, while particular conventions may be arbitrary, the interactive structures that stabilize and maintain them are not. Furthermore, the equilibria involved in coordinating on noun-meanings appear to have an arbitrary element only because we cannot Pareto-rank them; but Millikan (1984) shows implicitly that in this respect they are atypical of linguistic coordinations. In general, the various NE in coordination games can very often be ranked. Ross & LaCasse (1995) present the following example. In a city, drivers must coordinate on one of two NE with respect to their behaviour at traffic lights. Either all must rush yellows and pause on shifts to green, or slow down on yellows and jump forward on shifts to green. Both patterns are NE, in that once a community has coordinated on one of them no individual has an incentive to deviate: those who slow down on yellows while others are rushing them will get rear-ended, while those who rush yellows in the other equilibrium will risk collision with those who jump forward quickly on greens. Therefore, once a city's traffic pattern settles on one of these equilibria it will tend to stay there. However, the two states are not Pareto-indifferent, since the second NE allows more cars to turn left on each cycle (in a right-hand-drive jurisdiction), which reduces the extent of bottlenecks and allows all drivers to expect greater efficiency in getting about. Conventions on standards of evidence and rationality are likely to be of this character. While various arrangements might be NE in the social game of science, as followers of Thomas Kuhn like to remind us, it is highly improbable that all of these lie on a single Pareto-indifference curve. These themes, strongly represented in contemporary epistemology, philosophy of science and philosophy of language, are all bequests of game theory by way (at least indirectly) of Lewis. (The reader can find a broad sample of applications, and references to the large literature, in Nozick (1998).) However, Lewis restricted his attention to static game theory, in which agents *choose* strategies given exogenously fixed utility-functions. As a result of this restriction, his account is able to show us why conventions are important and stable, but it invites a difficult and perhaps ultimately fruitless quest for a general theory of rationality. This is because, as we saw in Section 3 above, in coordination (and other) games with multiple NE, what counts as a solution is highly sensitive to conjectures made by players about one another's beliefs and computational ability. This has excited a good deal of attention, especially from philosophers, on the implications of many subtle variations in the norms of strategic rationality. However, if game theory is to explain actual, natural behavior and its history in the way suggested by Gintis (2000) above, then we need some account of what is attractive about equilibria in games even when no analysts or rational calculators are around to identify them. To make reference again to Lewis's topic, when human language developed there was no external referee to care about and arrange for Pareto-efficiency. In order to understand Gintis's optimism about the reach of game theory, we must therefore extend our attention to *evolutionary* games.

Game theory has been fruitfully applied in evolutionary biology, where species and/or genes are treated as players, since pioneering work by Maynard Smith (1982) and his collaborators. Evolutionary (or *dynamic*) game theory now constitutes a significant new mathematical extension applicable to many settings apart from the biological. Thus Skyrms (1996) uses evolutionary game theory to try to answer questions Lewis could not even ask, about the conditions under which language, concepts of justice, the notion of private property, and other non-designed, general phenomena of interest to philosophers would be likely to arise. What is novel about evolutionary game theory is that moves are not chosen by rational agents. Instead, agents are typically hard-wired with particular strategies, and success for a strategy is defined in terms of the number of copies that a strategy will leave of itself to play in the games of

succeeding generations. The strategies themselves are therefore the players, and the games they play are dynamic rather than static.

The discussion here will closely follow Skyrms's. We begin by introducing *the replicator dynamics*. Consider first how natural selection works to change lineages of animals, modifying, creating and destroying species. The basic mechanism is *differential reproduction*. Any animal with *heritable* features that increase its *expected number of offspring* in a given environment will tend to leave more offspring than others so long as the environment remains relatively stable. These offspring will be more likely to inherit the features in question. Therefore, the proportion of these features in the population will gradually increase as generations pass. Some of these features may *go to fixation*, that is, eventually take over the entire population (until the environment changes).

How does game theory enter into this? Often, one of the most important aspects of an organism's environment will be the behavioural tendencies of other organisms. We can think of each lineage as 'trying' to maximize its reproductive fitness (= expected number of grandchildren) through finding strategies that are optimal given the strategies of other lineages. So evolutionary theory is another domain of application for non-parametric analysis.

In dynamic game theory, we no longer think of individuals as choosing strategies as they move from one game to another. This is because our interests are different. We're now concerned less with finding the equilibria of single games than with discovering which equilibria are stable, and how they will change over time. So we now model *the strategies themselves* as playing against each other. One strategy is 'better' than another if it is likely to leave more copies of itself in the next generation, when the game will be played again. We study the changes in distribution of strategies in the population as the sequence of games unfolds.

For dynamic game theory, we introduce a new equilibrium concept, due to Maynard Smith (1982). A set of strategies, in some particular proportion (e.g., 1/3:2/3, 1/2:1/2, 1/9:8/9, 1/3:1/3:1/6:1/6 -- always summing to 1) is at an *ESS* (Evolutionary Stable Strategy) equilibrium just in case (1) no individual playing one strategy could improve its reproductive fitness by switching to one of the other strategies in the proportion, and (2) no mutant playing a different strategy altogether could establish itself ('invade') in the population.

The principles of evolutionary game theory are best explained through examples. Skyrms begins by investigating the conditions under which a sense of justice -- understood as a disposition to view equal divisions of resources as fair unless efficiency considerations suggest otherwise in special cases -- might arise. He asks us to consider a population in which individuals regularly meet each other and must bargain over resources. Begin with three types of individuals:

a. *Fairmen* always demand exactly half the resource.
b. *Greedies* always demand more than half the resource. When a greedy encounters another greedy, they waste the resource in fighting over it.

c. *Modests* always demand less than half the resource. When a modest encounters another modest, they take less than all of the available resource and waste some.

Each *single* encounter where the total demands sum to 100% is a NE of that individual game. Similarly, there can be many dynamic equilibria. Suppose that Greedies demand 2/3 of the resource and Modests demand 1/3. Then the following two proportions are ESS's:

i. Half the population is greedy and half is modest. We can calculate the average payoff here. Modest gets 1/3 of the resource in every encounter. Greedy gets 2/3 when she meets Modest, but nothing when she meets another Greedy. So her average payoff is also 1/3. This is an ESS because Fairman can't invade. When Fairman meets Modest he gets 1/2. But when Fairman meets Greedy he gets nothing. So his average payoff is only 1/4. No Modest has an incentive to change strategies, and neither does any Greedy. A mutant Fairman arising in the population would do worst of all, and so selection will not encourage the propagation of any such mutants.
ii. All players are Fairmen. Everyone always gets half the resource, and no one can do better by switching to another strategy. Greedies entering this population encounter Fairmen and get an average payoff of 0. Modests get 1/3 as before, but this is less than Fairman's payoff of 1/2.

Notice that equilibrium (i) is inefficient, since the average payoff across the whole population is smaller. However, just as inefficient outcomes can be NE of static games, so they can be ESS's of dynamic ones.

We refer to equilibria in which more than one strategy occurs as *polymorphisms*. In general, in Skyrms's game, any polymorphism in which Greedy demands x and Modest demands 1-x is an ESS. The question that interests the student of justice concerns the relative likelihood with which these different equilibria arise.

This depends entirely on the proportions of strategies in the original population state. If the population begins with more than one Fairman, then there is some probability that Fairmen will encounter each other, and get the highest possible average payoff. Modests by themselves do not inhibit the spread of Fairmen; only Greedies do. But Greedies themselves depend on having Modests around in order to be viable. So the more Fairmen there are in the population relative to *pairs* of Greedies and Modests, the better Fairmen do on average. This implies a threshold effect. If the proportion of Fairmen drops below 33%, then the tendency will be for them to fall to extinction because they don't meet each other often enough. If the population of Fairmen rises above 33%, then the tendency will be for them to rise to fixation because their extra gains when they meet each other compensates for their losses when they meet Greedies. You can see this by noticing that when each strategy is used by 33% of the population, all have an expected average payoff of 1/3. Therefore, any rise above this threshold on the part of Fairmen will tend to push them towards fixation.

This result shows that and how, given certain relatively general conditions, justice as we have defined it *can* arise dynamically. The news for the fans of justice gets more cheerful still if we introduce *correlated play*.

The model we just considered assumes that strategies are not *correlated*, that is, that the probability with which every strategy meets every other strategy is a simple function of their relative frequencies in the population. We now examine what happens in our dynamic resource-division game when we introduce correlation. Suppose that Fairmen have a slight ability to distinguish and seek out other Fairmen as interaction partners. In that case, Fairmen on average do better, and this must have the effect of lowering their threshold for going to fixation.

A dynamic-game modeler studies the effects of correlation and other parametric constraints by means of running large computer simulations in which the strategies compete with one another, round after round, in the virtual environment. The starting proportions of strategies, and any chosen degree of correlation, can simply be set in the programme. One can then watch its dynamics unfold over time, and measure the proportion of time it stays in any one equilibrium. These proportions are represented by the relative sizes of the *basins of attraction* for different possible equilibria. Equilibria are attractor points in a dynamic space; a basin of attraction for each such point is then the set of points in the space from which the population will converge to the equilibrium in question.

In introducing correlation into his model, Skyrms first sets the degree of correlation at a very small .1. This causes the basin of attraction for equilibrium (i) to shrink by half. When the degree of correlation is set to .2, the polymorphic basin reduces to the point at which the population starts in the polymorphism. Thus very small increases in correlation produce large proportionate increases in the stability of the equilibrium where everyone plays Fairman. A small amount of correlation is a reasonable assumption in most populations, given that neighbours tend to interact with one another and to mimic one another (either genetically or because of tendencies to deliberately copy each other), and because genetically similar animals are more likely to live in common environments. Thus if justice can arise at all it will tend to be dominant and stable.

Much of political philosophy consists in attempts to produce deductive normative arguments intended to convince an unjust agent that she has reasons to act justly. Skyrms's analysis suggests a quite different approach. Fairman will do best of all in the dynamic game if he takes active steps to preserve correlation. Therefore, there is evolutionary pressure for both *moral approval of justice* and *just institutions* to arise. Most people may think that 50-50 splits are ‘fair’, and worth maintaining by moral and institutional reward and sanction, *because* we are the products of a dynamic game that promoted our tendency to think this way.

The topic that has received most attention from evolutionary game theorists is *altruism*, defined as any behaviour by an organism that decreases its own expected fitness in a single interaction but increases that of the other interactor. It is common in nature. How can it arise, however, given Darwinian competition?

Skyrms studies this question using the dynamic Prisoner's Dilemma as his example. This is simply a series of PD games played in a population, some of whose members are defectors and some of whom are cooperators. Payoffs, as always in dynamic games, are measured in terms of expected numbers of copies of each strategy in future generations.

Let **U**(*A*) be the average fitness of strategy *A* in the population. Let **U** be the average fitness of the whole population. Then the proportion of strategy *A* in the next generation is just the ratio **U**(*A*)/**U**. So if *A* has greater fitness than the population average *A* increases. If *A* has lower fitness than the population average then *A* decreases.

In the dynamic PD where interaction is random (i.e., there's no correlation), defectors do better than the population average as long as there are cooperators around. This follows from the fact that, as we saw in Section 2.4, defection is always the dominant strategy in a single game. 100% defection is therefore the ESS in the dynamic game without correlation, corresponding to the NE in the one-shot static PD.

However, introducing the possibility of correlation radically changes the picture. We now need to compute the average fitness of a strategy *given its probability of meeting each other possible strategy*. In the dynamic PD, cooperators whose probability of meeting other cooperators is high do better than defectors whose probability of meeting other defectors is high. Correlation thus favours cooperation.

In order to be able to say something more precise about this relationship between correlation and cooperation (and in order to be able to relate evolutionary game theory to issues in decision theory, a matter falling outside the scope of this article), Skyrms introduces a new technical concept. He calls a strategy *adaptively ratifiable* if there is a region around its fixation point in the dynamic space such that from anywhere within that region it will go to fixation. In the dynamic PD, both defection and cooperation are adaptively ratifiable. The relative sizes of basins of attraction are highly sensitive to the particular mechanisms by which correlation is achieved. To illustrate this point, Skyrms builds several examples.

One of Skyrms's models introduces correlation by means of a *filter* on pairing for interaction. Suppose that in round 1 of a dynamic PD individuals inspect each other and interact, or not, depending on what they find. In the second and subsequent rounds, all individuals who didn't pair in round 1 are randomly paired. In this game, the basin of attraction for defection is large *unless* there is a high proportion of cooperators in round one. In this case, defectors fail to pair in round 1, then get paired mostly with each other in round 2 and drive each other to extinction. A model which is more interesting, because its mechanism is less artificial, does not allow individuals to choose their partners, but requires them to interact with those closest to them. Because of genetic relatedness (or cultural learning by copying) individuals are more likely to resemble their neighbours than not. If this (finite) population is arrayed along one dimension (i.e., along a line), and both cooperators and defectors are introduced into positions along it at random, then we get the following dynamics. Isolated cooperators have lower expected fitness than the surrounding defectors and are driven locally to extinction. Members of groups of two cooperators have a 50% probability of interacting with each other, and a 50% probability of each interacting with a defector. As a result, their average expected fitness remains smaller than that of their neighbouring defectors, and they too face probable extinction. Groups of three cooperators form an unstable point from which both extinction and expansion are equally likely. However, in groups of four or more cooperators at least one encounter of a cooperator with a cooperator sufficient to at least replace the original group is guaranteed. Under this circumstance, the cooperators as a group do better than the surrounding defectors and increase at their expense. Eventually cooperators go *almost* to fixation -- but nor quite. Single

defectors on the periphery of the population prey on the cooperators at the ends and survive as little 'criminal communities'. We thus see that altruism can not only be maintained by the dynamics of evolutionary games, but, with correlation, can even spread and colonize originally non-altruistic populations.

Darwinian dynamics thus offers qualified good news for cooperation. Notice, however, that this holds only so long as individuals are stuck with their natural or cultural programming and can't re-evaluate their utilities for themselves. If our agents get too smart and flexible, they may notice that they're in PDs and would each be best off defecting. In that case, they'll eventually drive themselves to extinction - unless they develop stable, and effective, moral norms that work to reinforce cooperation. But, of course, these are just what we would expect to evolve in populations of animals whose average fitness levels are closely linked to their capacities for successful social cooperation. Even given this, these populations will go extinct unless they care about future generations for some reason. But there's no rational reason as to why agents *should* care about future generations if each new generation wholly replaces the preceding one at each change of cohorts. For this reason, economists use 'overlapping generations' models when modeling distribution games. Individuals in generation 1 who will last until generation 5 save resources for the generation 3 individuals with whom they'll want to cooperate; and by generation 3 the new individuals care about generation 6; and so on.

An enormous range of further applications of both static and dynamic game theory have been developed, but we have perhaps now provided enough to convince the reader of the tremendous utility of this analytical tool. The reader whose appetite for more has been thoroughly aroused should find that she now has sufficient grasp of fundamentals to be able to work through the large literature, of which some highlights are listed below.

Bibliography

Annotations

In the following section, books which no philosopher seriously interested in game theory can afford to miss are marked with (**).

Game theory has countless applications, of which this article has been able to suggest only a few. Readers in search of more, but not wishing to immerse themselves in mathematics, can find a number of good sources. Dixit and Nalebuff (1991) is especially strong on political and social examples. McMillan (1991) emphasizes business applications. The great historical breakthrough is von Neumann and Morgenstern (1947), which those with scholarly interest in game theory should read with classic papers of John Nash (1950a, 1950b, 1951). For a contemporary mathematical treatment that is unusually philosophically sophisticated, Binmore (1992) (**) is in a class by itself. The second half of Kreps (1990) (**) is the best available starting point for a tour of the philosophical worries surrounding equilibrium selection for non-behaviorists. Koons (1992) takes these issues further. Fudenberg and Tirole (1991) is the most thorough

and complete mathematical text available. Gintis (2000) (**) has provided a new text crammed with terrific problem exercises, which is also unique in that it treats evolutionary game theory as providing the foundational basis for game theory *in general*. This likely represents the wave of the future. Recent developments in fundamental theory are well represented in Binmore, Kirman and Tani (1993).

The philosophical foundations of the basic game-theoretic concepts as economists understand them are presented in LaCasse and Ross (1994). Ross and LaCasse (1995) outline the relationships between games and the axiomatic assumptions of microeconomics and macroeconomics. Philosophical puzzles at this foundational level are critically discussed in Bicchieri (1993) (**). Lewis (1969) (**) puts game-theoretic equilibrium concepts to wider application in philosophy, a program that is carried a good deal further in Skyrms (1996) (**). (See also Nozick [1998].) Gauthier (1986) launches a literature not surveyed in this article, in which the possibility of game-theoretic foundations for contractarian ethics is investigated. This work is critically surveyed in Vallentyne (1991), and extended into a dynamic setting in Danielson (1992). Binmore (1994, 1998) (**), however, effectively demolishes this project. Philosophers will also find Hollis (1998) to be of interest.

Hardin (1995) is one of many examples of the application of game theory to problems in applied political theory. Baird, Gertner and Picker (1994) review uses of game theory in legal theory and jurisprudence. Mueller (1997) surveys applications in political economy. Ghemawat (1997) does the same in business strategy. Poundstone (1992) provides a lively history of the Prisoner's Dilemma and its use by Cold War strategists. Durlauf and Young (2001) is a good collection on applications to social structures and social change.

Evolutionary game theory owes its explicit genesis to Maynard Smith (1982) (**). For a text that integrates game theory directly with biology, see Sigmund (1993). The most exciting applications of evolutionary game theory to a range of philosophical issues, on which this article has drawn heavily, is Skyrms (1996) (**). These issues and others are critically discussed from various angles in Danielson (1998). Mathematical foundations for dynamic games are presented in Weibull (1995), and pursued further in Samuelson (1997) and Fudenberg and Levine (1998). As noted above, Gintis (2000) (**) now provides an introductory textbook that takes evolutionary modeling to be foundational to all of game theory. Many philosophers will also be interested in Binmore (1994, 1998) (**), which shows that application of game-theoretic analysis can underwrite a loosely Rawlsian theory of justice that does not require recourse to Kantian presuppositions about what rational agents would desire behind a veil of ignorance concerning their identities and social roles. (In addition, Binmore offers excusions into a vast range of other issues both central and peripheral to both the foundations and the frontiers of game theory; these books are a tour de force.) And almost everyone will be interested in Frank (1988) (**), where evolutionary game theory is used to illuminate basic features of human nature and emotion.

References

- Baird, D., Gertner, R., and Picker, R. (1994). *Game Theory and the Law*. Cambridge, MA: Harvard University Press.
- Bicchieri, C. (1993). *Rationality and Coordination*. Cambridge: Cambridge University Press.
- Binmore, K. (1992). *Fun and Games*. Lexington, MA: D. C. Heath.
- Binmore, K., Kirman, A., and Tani, P. (eds.) (1993). *Frontiers of Game Theory*. Cambridge, MA: MIT Press
- Binmore, K. (1994). *Game Theory and the Social Contract* (v. 1): *Playing Fair*. Cambridge, MA: MIT Press.
- Binmore, K. (1998). *Game Theory and the Social Contract* (v. 2): *Just Playing*. Cambridge, MA: MIT Press.
- Danielson, P. (1992). *Artificial Morality*. London: Routledge
- Danielson, P. (ed.) (1998). *Modelling Rationality, Morality and Evolution*. Oxford: Oxford University Press.
- Dixit, A., and Nalebuff, B. (1991). *Thinking Strategically*. New York: Norton.
- Durlauf, S., and Young, H.P., eds. (2001). *Social Dynamics*. Cambridge, MA: MIT Press.
- Frank, R. (1988). *Passions Within Reason*. New York: Norton.
- Fudenberg, D., and Levine, D. (1998). *The Theory of Learning in Games*. Cambridge, MA: MIT Press.
- Fudenberg, D., and Tirole, J. (1991). *Game Theory*. Cambridge, MA: MIT Press.
- Gauthier, D. (1986). *Morals By Agreement*. Oxford: Oxford University Press.
- Ghemawat, P. (1997). *Games Businesses Play*. Cambridge, MA: MIT Press.
- Ginits, H. (2000). *Game Theory Evolving.* Princeton: Princeton University Press.
- Hardin, R. (1995). *One For All*. Princeton: Princeton University Press.
- Hollis, M. (1998). *Trust Within Reason*. Cambridge: Cambridge University Press.
- Koons, R. (1992). *Paradoxes of Belief and Strategic Rationality*. Cambridge: Cambridge University Press.
- Kreps, D. (1990). *A Course in Microeconomic Theory*. Princeton: Princeton University Press.
- LaCasse, C., and Ross, D. (1994). ‘The Microeconomic Interpretation of Games’. *PSA 1994, v. 1*. D. Hull, S. Forbes and R. Burien, eds.. East Lansing, MI: Philosophy of Science Association. Pages 479-387.
- Lewis, D. (1969). *Convention*. Cambridge, MA: Harvard University Press.
- Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
- McMillan, J. (1991). *Games, Strategies and Managers*. Oxford: Oxford University Press.
- Millikan, R. (1984). *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.
- Mueller, D. (1997). *Perspectives on Public Choice*. Cambridge: Cambridge University Press.
- Nash, J. (1950a). ‘Equilibrium Points in *n*-Person Games.’ *PNAS* 36:48-49.
- Nash, J. (1950b). ‘The Bargaining Problem.’ *Econometrica* 18:155-162.
- Nash, J. (1951). ‘Non-cooperative Games.’ *Annals of Mathematics Journal* 54:286-295.
- Nozick, R. (1998). *Socratic Puzzles*. Cambridge, MA: Harvard University Press.
- Poundstone, W. (1992). *Prisoner's Dilemma*. New York: Doubleday.
- Robbins, L. (1931). *An Essay on the Nature and Significance of Economic Science*. London:

Macmillan.
- Ross, D., and LaCasse, C. (1995). ‘Towards a New Philosophy of Positive Economics’. *Dialogue* 34: 467-493.
- Samuelson, L. (1997). *Evolutionary Games and Equilibrium Selection*. Cambridge, MA: MIT Press.
- Samuelson, P. (1938). ‘A Note on the Pure Theory of Consumers' Behaviour.’ *Econimica* 5:61-71.
- Selten, R. (1975). ‘Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games.’ *International Journal of Game Theory* 4:22-55.
- Sigmund, K. (1993). *Games of Life*. Oxford: Oxford University Press.
- Skyrms, B.(1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
- Vallentyne, P. (ed.). (1991). *Contractarianism and Rational Choice*. Cambridge: Cambridge University Press.
- von Neumann, J., and Morgenstern, O., (1947). *The Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 2nd edition.
- Weibull, J. (1995). *Evolutionary Game Theory*. Cambridge, MA: MIT Press.

Other Internet Resources

- History of Game Theory
- Principia Cybernetica entry: Game Theory
- University of Rochester Economics Department: Game Theory
- TU Wroclaw IMath -- Game Theory
- What is Game Theory?
- Al Roth's Game Theory and Experimental Economics Page

Related Entries

game theory: evolutionary | prisoner’s dilemma | rationality

Acknowledgments

I would like to thank James Joyce and Edward Zalta for their comments on the original version and on the first draft of the revised version of this entry. I would also like to thank Anthony Boting, for noticing that my solution to one of the examples rested on equivocating between relative-frequency and objective-chance interpretations of probability. Finally, thanks go to Colin Allen for all his technical support behind the scenes (in the effort to deal with bandwidth problems to South Africa) prior to publication of the revised version of this entry.

University of Cape Town
dross@commerce.uct.ac.za

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 25, 1997
Content last modified: September 11, 2001

Prisoner's Dilemma

Tanya and Cinque have been arrested for robbing the Hibernia Savings Bank and placed in separate isolation cells. Both care much more about their personal freedom than about the welfare of their accomplice. A clever prosecutor makes the following offer to each. "You may choose to confess or remain silent. If you confess and your accomplice remains silent I will drop all charges against you and use your testimony to ensure that your accomplice does serious time. Likewise, if your accomplice confesses while you remain silent, they will go free while you do the time. If you both confess I get two convictions, but I'll see to it that you both get early parole. If you both remain silent, I'll have to settle for token sentences on firearms possession charges. If you wish to confess, you must leave a note with the jailer before my return tomorrow morning."

The "dilemma" faced by the prisoners here is that, whatever the other does, each is better off confessing than remaining silent. But the outcome obtained when both confess is worse for each than the outcome they would have obtained had both remained silent. A common view is that the puzzle illustrates a conflict between individual and group rationality. A group whose members pursue rational self-interest may all end up worse off than a group whose members act contrary to rational self-interest. More generally, if the payoffs are not assumed to represent self-interest, a group whose members rationally pursue any goals may all meet less success than if they had not rationally pursued their goals individually. Puzzles with this structure were devised and discussed by Merrill Flood and Melvin Dresher in 1950, as part of the Rand Corporation's investigations into game theory (which Rand pursued because of possible applications to global nuclear strategy). The title "prisoner's dilemma" and the version with prison sentences as payoffs are due to Albert Tucker, who wanted to make Flood and Dresher's ideas more accessible to an audience of Stanford psychologists. Although Flood and Dresher didn't themselves rush to publicize their ideas in external journal articles, the puzzle attracted widespread attention in a variety of disciplines. Christian Donninger reports that "more than a thousand articles" about it were published in the sixties and seventies. A bibliography (Axelrod and D'Ambrosio) of writings between 1988 and 1994 that pertain to Robert Axelrod's research on the subject lists 209 entries. Since then the flow has shown no signs of abating.

The sections below provide a variety of more precise characterizations of the prisoner's dilemma, beginning with the narrowest. ‘Prisoner's dilemma’ is abbreviated as ‘PD’. Future editions of the entry will also survey some applications in philosophy, and attempts to "solve" the PD by showing that remaining silent is individually rational after all.

Symmetric 2x2 PD With Ordinal Payoffs

In its simplest form the PD is a game described by the payoff matrix:

	C	**D**
C	R,R	S,T
D	T,S	P,P

satisfying the following chain of inequalities:

PD1) T>R>P>S

There are two players, Row and Column. Each has two possible moves, "cooperate" or "defect," corresponding, respectively, to the options of remaining silent or confessing in the illustrative anecdote above. For each possible pair of moves, the payoffs to Row and Column (in that order) are listed in the appropriate cell. R is the "reward" payoff that each player receives if both cooperate. P is the "punishment" that each receives if both defect. T is the "temptation" that each receives if he alone defects and S is the "sucker" payoff that he receives if he alone cooperates. We assume here that the game is symmetric, i.e., that the reward, punishment, temptation or sucker payoff is the same for each player, and payoffs have only ordinal significance, i.e., they indicate whether one payoff is better than another, but tell us nothing about how much better. It is now easy to see that we have the structure of a dilemma like the one in the story. Suppose Column cooperates. Then Row gets R for cooperating and T for defecting, and so is better off defecting. Suppose Column defects. Then Row gets S for cooperating and P for defecting, and so is again better off defecting. The move **D** for Row is said to *strictly dominate* the move **C**: whatever his opponent

does, he is better off choosing **D** than **C**. By symmetry **D** also strictly dominates **C** for Column. Thus two "rational" players will defect and receive a payoff of P, while two "irrational" players can cooperate and receive greater payoff R. In standard treatments, game theory assumes rationality and common knowledge. Each player is rational, knows the other is rational, knows that the other knows he is rational, etc. Each player also knows how the other values the outcomes. But since **D** strictly dominates C for both players, the argument for dilemma here requires only that each player knows his own payoffs. (The argument remains valid, of course, under the stronger standard assumptions.) It is also worth noting that the outcome (**D**,**D**) of both players defecting is the game's only strong nash equilibrium, i.e., it is the only outcome from which each player could only do worse by unilaterally changing its move. Flood and Dresher's interest in their dilemma seems to have stemmed from their view that it provided a counterexample to the claim that the nash equilibria of a game constitute its natural "solutions".

If there can be "ties" in rankings of the payoffs, condition PD1 can be weakened without destroying the nature of the dilemma. For suppose that one of the following conditions obtain:

PD2) $T>R>P\geq S$, or
$T\geq R>P>S$

Then, for each player, although **D** does not strictly dominate **C**, it still *weakly dominates* in the sense that each player always does at least as well, and sometimes better, by playing **C**. Under these conditions it still seems rational to play **D**, which again results in the payoff that neither player prefers. Let us call a game that meets PD2 a *weak PD*. Note that in a weak PD that does not satisfy PD1 mutual defection is no longer a nash equilibrium in the strong sense defined above. It is still, however, the only nash equilibrium in the weaker sense, that neither player can improve its position by unilaterally changing its move. Again, one might suppose that if there is a unique nash equilibrium of this weaker variety, rational self-interested players would reach it.

Asymmetry

Without assuming symmetry, the PD can be represented by using subscripts r and c for the payoffs to Row and Column.

	C	**D**
C	R_r,R_c	S_r,T_c
D	T_r,S_c	P_r,P_c

If we assume that the payoffs are ordered as before for each player, i.e., that $T_i>R_i>P_i>S_i$ when i=r,c, then, as before, **D** is the strictly dominant move for both players, but the outcome (**D**,**D**) of both players making this move is worse for each than (**C**,**C**). The force of the dilemma can now also be felt under weaker conditions, however. Consider the following three pairs of inequalities:

PD3) a. $T_r > R_r$ and $P_r > S_r$

b. $T_c > R_c$ and $P_c > S_c$

c. $R_r > P_r$ and $R_c > P_c$

If these conditions all obtain the argument for dilemma goes through as before. Defection strictly dominates cooperation for each player, and (**C**,**C**) is strictly preferred by each to (**D**,**D**). If one of the two > signs in each of the conditions a - c is replaced by a weak inequality sign $\geq$ we have a weak PD. **D** weakly dominates **C** for each player (i.e., **D** is as good as **C** in all cases and better in some) and (**C**,**C**) weakly better than (**D**,**D**) (i.e., it is at least as good for both players and better for one). Since none of the clauses requires comparisons between r's payoffs and c's, we need not assume that > has any "interpersonal" significance.

Now suppose we drop the first inequality of either a or b (but not both). A game that meets the resulting conditions might be termed a *common knowledge PD*. As long as each player knows that the other is rational and each knows the other's ordering of payoffs, we still feel the force of the dilemma. For suppose b holds. Then **D** is the dominant move for Row. Column, knowing that Row is rational, knows that Row will defect, and so, by the remaining inequality in b, will defect himself. Similarly, if b holds Column will defect, and Row, realizing this, will defect herself. By c, the resulting (**D**,**D**) is again worse for both than (**C**,**C**).

Multiple Moves

Speaking generally, one might say that a PD is a game in which a "cooperative" outcome obtainable only when every player violates rational self-interest is unanimously preferred to the "selfish" outcome obtained when every player adheres to rational self-interest. We can characterize the selfish outcome either as the result of each player pursuing its dominant (strongly dominant) strategy, or as the unique weak (strong) nash equilibrium. In a two move game the two characterizations come to the same thing **--** a dominant move pair is a unique equilibrium and a unique equilibrium is a dominant move pair. As the payoff matrix below shows, however, the two notions diverge in a game with more than two moves.

	C	**D**	**N**
C	R,R	S,T	T,S
D	T,S	P,P	R,S
N	S,T	S,R	S,S

Here each player can choose "cooperate", "defect" or "neither" and the payoffs are ordered as before. Defection is no longer dominant, because each player is better off choosing **C** than **D** when the other chooses N. Nevertheless (**D**,**D**) is still the unique equilibrium. Let us label a game like this in which the selfish outcome is the unique equilibrium an *equilibrium PD*, and one in which the selfish outcome is a pair of dominant moves a *dominance PD*. As will be seen below, attempts to "solve" the PD by allowing

conditional strategies can create multiple-move games that are themselves equilibrium PDs.

Multiple Players

Most of those who maintain that the PD illustrates something important about morality seem to believe that the basic structure of the game is reflected in situations that larger groups, perhaps entire societies, face. The most obvious generalization from the two-player to the many-player game would pay each player R if all cooperate, P if all defect, and, if some cooperate and some defect, it would pay the cooperators S and the defectors T. But it is unlikely that we face many situations of this structure.

A common view is that a multi-player PD structure is reflected in what Garret Hardin popularized as "the tragedy of the commons." Each member of a group of neighboring farmers prefers to allow his cow to graze on the commons, rather than keeping it on his own inadequate land, but the commons will be rendered unsuitable for grazing if it is used by more than some threshold number use it. More generally, there is some social benefit B that each member can achieve if sufficiently many pay a cost C. We might represent the payoff matrix as follows:

	***n* or fewer choose C**	**more than *n* choose C**
C	C	C+B
D	0	B

The cost C is assumed to be a negative number. The "temptation" here is to get the benefit without the cost, the reward is the benefit with the cost, the punishment is to get neither and the sucker payoff is to pay the cost without realizing the benefit. So the payoffs are ordered B > (B+C) > 0 > C. As in the two-player game, it appears that **D** weakly dominates **C** for all players, and so rational players would choose **D** and achieve 0, while preferring that everyone would choose **C** and obtain C+B.

Unlike the more straightforward generalization, this matrix does reflect common social choices -- between depleting and conserving a scarce resource, between using polluting and non-polluting means of manufacture or disposal, and between participating and not participating in a group effort towards some common goal. When *n* is small, it represents a version of what has been called the "volunteer dilemma". A group needs a few volunteers, but each member is better off if others volunteer. (Notice, however, that in a true volunteer dilemma, where only one volunteer is needed, *n* is zero and the top left outcome is impossible. Under these conditions D no longer dominates **C** and the game loses its PD flavor.) A particularly vexing manifestation of this game occurs when a vaccination known to have serious risks is needed to prevent the outbreak of a fatal disease. If enough of her neighbors get the vaccine, each person may be protected without assuming the risks.

The tragedy of the commons game diagramed above has a somewhat different character than the two-player PD. First, even if each player's moves are entirely independent of the others, the alternatives represented by the columns in the commons matrix above are no longer independent of the alternatives represented by the

rows. My choosing **C** necessarily increases the chances that more than *n* people will choose **C**. To ensure independence we should really redraw the matrix as follows:

	fewer than *n* others choose C	***n* others choose C**	**more than *n* others choose C**
C	C	C+B	C+B
D	0	0	B

But now we see that move **D** does *not* dominate **C**. When we are at the threshold of adequate cooperation, I am better off cooperating. Provided that *n* is large, however, it would seem that this effect could be ignored and we could assume, for practical purposes, that the payoff matrix is like the previous one.

Similarly, whereas we saw in the original PD that mutual defection was the only nash equilibrium, this game has two equilibria. One is universal defection, since any player unilaterally departing from that outcome will move from payoff 0 to C. But a second is the state of *minimally effective* cooperation, where the number of cooperators just exceeds the threshold n. A defector who unilaterally departs from that outcome will move from B to B+C and a cooperator who unilaterally departs will move from B+C to 0. This might suggest that the tragedy of the commons is less tragic than the PD, but in real life situations, it would seem unlikely that the participants would know when they are at the equilibrium point of minimally effective cooperation.

Furthermore, in the ordinary PD, universal cooperation is a pareto-optimal outcome, i.e., there is no outcome in which each player is at least as well off and some are better off. But in the commons game the only pareto optimal outcomes are those of minimally effective cooperation. Whether universal cooperation is nevertheless desirable may depend on the nature of the choices involved. In the medical example it may seem best to vaccinate everyone. In the agricultural example, however it seems foolish to stipulate that nobody use the commons. Someone who avoids vaccination in the former case is seen as a "free rider". An underused commons in the latter seems to exemplify "surplus cooperation."

The two-person version of the tragedy of the commons game (with threshold of one) produces a matrix presenting little or no dilemma.

	C	**D**
C	B+C,B+C	C,0
D	0,C	0,0

This game captures David Hume's example of a boat with one oarsman on the port side and another on the starboard. Mutual cooperation is identical to minimally effective cooperation and therefore is both an equilibrium outcome and a pareto optimal outcome.

The above representations of the tragedy of the commons make the simplifying assumptions that the costs and benefits to each player are the same, and that these costs and benefits are independent of the number of players who cooperate. A somewhat more general account would replace C and B by functions C(i,j) and B(i,j) representing the cost of cooperation to player i when he is one of exactly j players who cooperate and the benefit to player i when exactly j players cooperate. For each player i, we suppose that there is some threshold t_i such that B(i,j) is not defined unless $j>t_i$. We may assume additional cooperation never reduces the benefit i gets from general cooperation, i.e., $B(i,j+1) \geq B(i,j)$ and that additional defection never reduces the cost i bears in cooperating, i.e, $C(i,j+1) \geq C(i,j)$. Now suppose that, for all players i, B(i,j)>(B(i,j+1)+C(i,j+1)) when j is greater than the threshold for i but less than the total number of players, and 0>C(i,j) when j is less than the threshold for i. Suppose additionally that, for all i and all j greater than the threshold for i, B(i,j)+C(i,j)>0. We then have a tragedy of the commons game, which (if we ignore the outcome of minimally effective cooperation) presents the familiar dilemma: defection is individually rational, but general cooperation benefits everyone.

Phillip Pettit has pointed out that examples that might be represented as many-player PD's come in two flavors. The examples discussed above might be classified as free-rider problems. My temptation is to enjoy some benefits brought about by burdens shouldered by others. The other flavor is what Pettit calls "foul dealer" problems. My temptation is to benefit myself by hurting others. Suppose, for example, that a group of people are applying for a single job, for which they are equally qualified. If all fill out their applications honestly, they all have an equal chance of being hired. If one lies, however, he can ensure that he is hired while, let us say, incurring a small risk of being exposed later. If everyone lies, they again have an equal chance for the job, but now they all incur the risk of exposure. Thus a lone liar, by reducing the others' chances of employment from slim to none, raises his own chances from slim to sure. As Pettit points out, when the minimally effective level of cooperation is the same as the size of the population, there is no opportunity for free-riding (everyone's cooperation is needed), and so the PD must be of the foul-dealing variety. But (Pettit's contrary claim notwithstanding) not all foul-dealing PDs seem to have this feature. Suppose, for example, that two applicants in the story above will be hired. Then everyone gets the benefit (a chance of employment without risk of exposure) unless two or more players lie. Nevertheless, the liars seem to be foul dealers rather than free riders. A better characterization of the foul-dealing dilemma might be that every defection from a generally cooperative state strictly reduces the payoffs to the cooperators, i.e., for every player i and every j greater than i's threshold, either B(i,j+1)>B(i,j) or C(i,j+1)>C(i,j). A free-rider's defection benefits himself but does not, by itself, hurt the cooperators. A foul-dealer's defection benefits himself and hurts the cooperators.

The game labeled a many-person PD in Schelling, Molander 1992 and elsewhere requires that the payoff to each co-operator and defector increases strictly with the number of cooperators and that the sum of the payoffs to all parties increases with the number of cooperators (so that one party's switching from defection to cooperation always raises the sum). Neither of these conditions is met by the formulation and the examples discussed above. They may, however, hold "locally," i.e., for j close to the threshold of minimally effective cooperation, it may be reasonable to assume B(i,j+1)+C(i,j+1)>B(i,j)+C(i,j), B(i,j+1)>B(i,j) and B(1,j+1)+C(1,j+1)+...+B(j+1,j+1)+C(j+1,j+1)+B(j+2,j+1))+...+B(n,j+1) > B(1,j)+C(1,j)+...+B(j,j)+C(j,j)+B(j+1,j)+...+B(n,j).

Single Person Interpretation

The PD is usually thought to illustrate conflict between individual and collective rationally, but the multiple player form (or something very similar) has also been interpreted as demonstrating problems within standard conceptions of individual rationality. This interpretation, elucidated in Quinn, derives from an example of Parfit's. A medical device enables electric current to be applied to a patient's body in increments so tiny that there is no perceivable difference between adjacent settings. You are attached to the device and given the following choice every day for ten years: advance the device one setting and collect a thousand dollars, or leave it where it is and get nothing. Since there is no perceivable difference between adjacent settings, it is apparently rational to advance the setting each day. But at the end of ten years the pain is so great that a rational person would sacrifice all his wealth to return to the first setting.

We can view the situation here as a multi-player PD in which each "player" is the temporal stage of a single person. So viewed, it has at least two features that were not discussed in connection with the multi-player examples. First, the moves of the players are sequential rather than simultaneous (and each player has knowledge of preceding moves). Second, there is the matter of gradation. Increases in electric current between adjacent settings are imperceptible, and therefore irrelevant to rational decision-making, but sums of a number such increases are noticeable and highly relevant. Neither of these features, however, is peculiar to one-person examples. Consider, for example, the choice between a polluting and non-polluting means of waste disposal. Each resident of a lakeside community may dump his or her garbage in the lake or use a less convenient landfill. It is reasonable to suppose that each acts in the knowledge of how others have acted before. (See "Asynchronous Moves" below.) It is also reasonable to suppose that addition of one can of garbage to the lake has no perceptible effect on water quality, and therefore no effect on the welfare of the residents. The fact that the dilemma remains suggests that PD-like situations sometimes involve something more than a conflict between individual and collective rationality. In the one-person example, our understanding that we care more about our overall well-being than that of our temporal stages does not (by itself) eliminate the argument that it is rational to continue to adjust the setting. Similarly, in the pollution example, a decision to let collective rationality override individual rationality may not eliminate the argument for excessive dumping. It seems appropriate, however, to separate this issue from that raised in the standard PD. Gradations that are imperceptible individually, but weighty en masse give rise to intransitive preferences. This is a challenge to standard accounts of rationality whether or not it arises in a PD-like setting.

Cardinal Payoffs

If the game specifies absolute (as opposed to relative) payoffs, then universal cooperation may not be a pareto optimal outcome even in the two person PD. For under some conditions both players do better by adopting a *mixed* strategy of cooperating with probability p and defecting with probability (1-p). This point is illustrated in the graphs below.

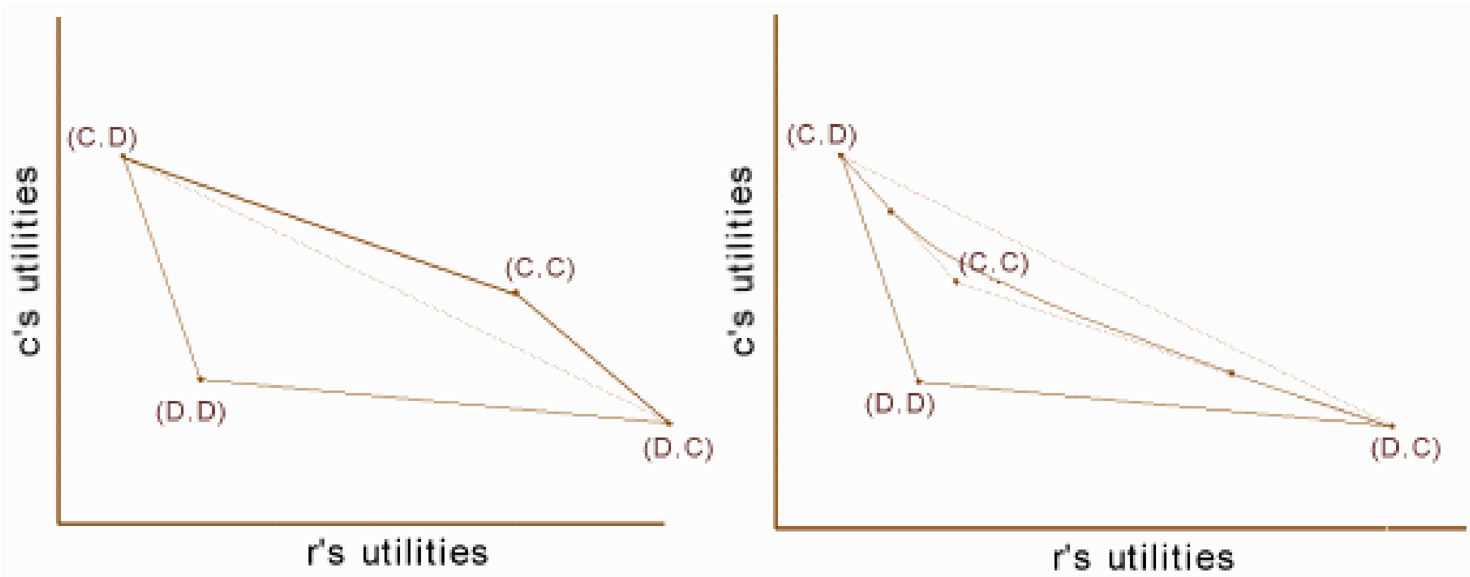


Figure 1

Here the x and y axes represent the utilities of Row and Column. The four outcomes entered in the matrix of the second section are represented by the labeled dots. Conditions PD3a and PD3b ensure that (**C**,**D**) and (**D**,**C**) lie northwest and southeast of (**D**,**D**), and PD3c is reflected in the fact that (**C**,**C**) lies northeast of (**D**,**D**). Suppose first that (**D**,**D**) and (**C**,**C**) lie on opposite sides of the line between (**C**,**D**) and (**D**,**C**), as in the graph on the left. Then the four points form a convex quadrilateral, and the payoffs of the feasible outcomes of mixed strategies are represented by all the points on or within this quadrilateral. Of course a player can really only get one of four possible payoffs each time the game is played, but the points in the quadrilateral represent the *expected values* of the payoffs to the two players. If Row and Column cooperate with probabilities p and q (and defect with probabilities p*=1-p and q*=1-q), for example, then the expected value of the payoff to Row is p*qT+pqR+p*q*P+pq*S. A rational self-interested player, according to a standard view, should prefer a higher expected payoff to a lower one. In the graph on the left the payoff for universal cooperation (with probability one) is pareto optimal among the payoffs for all mixed strategies. In the graph on the right, however, where both (**D**,**D**) and (**C**,**C**) lie southwest of the line between (**C**,**D**) and (**D**,**C**), the story is more complicated. Here the payoffs of the feasible outcome lie within a figure bounded on the northeast by three distinct curve segments, two linear and one concave. Notice that (**C**,**C**) is now in the interior of the region bounded by solid lines, indicating that there are mixed strategies that provide both players a higher expected payoff than (**C**,**C**). It is important to note that we are talking about independent mixed strategies here. Row and Column use private randomizing devices and have no communication. If they were able to correlate their mixed strategies, so as to ensure, say (**C**,**D**) with probability p and (**D**,**C**) with probability p*, the set of feasible solutions would extend up to (and include) the dotted line between (**C**,**D**) and (**D**,**C**). The point here is that, even confined to independent strategies, there are some games satisfying PD3 in which both players can both do better than they do with universal cooperation. A PD in which universal cooperation is pareto optimal may be called a pure PD. A pure PD is characterized by adding to PD3 the following condition.

P) $(T_r-R_r)(T_c-R_c) \leq (R_r-S_r)(R_c-S_c)$

In a symmetric game P reduces to the simpler condition

RCA) $R \geq 1/2(T+S)$

(named after the authors Rapoport, Chammah and Axelrod who employed it).

Asynchronous Moves

It has often been argued that rational self-interested players can obtain the cooperative outcome by making their moves conditional on the moves of the other player. Peter Danielson, for example, favors a strategy of *reciprocal cooperation*: if the other player would cooperate if you cooperate and would defect if you don't, then cooperate, but otherwise defect. Conditional strategies like this are ruled out in the versions of the game described above, but they may be possible in versions that more accurately model real world situations. In this section and the next, we consider two such versions. Here we eliminate the requirement that the two players move simultaneously. Consider the situation of a firm whose sole competitor has just lowered prices. Or suppose the buyer of a car has just paid the agreed purchase price and the seller has not yet handed over the title. We can think of these as situations in which one player has to choose to cooperate or defect after the other player has already made a similar choice. The corresponding game is an *asynchronous* or *extended* PD.

Careful discussion of an asynchronous PD example, as Skyrms (1998) and Vanderschraaf recently note, occurs in the writings of David Hume, well before Flood and Dresher's formulation of the ordinary PD. Hume writes about two neighboring grain farmers:

> Your corn is ripe today; mine will be so tomorrow. ’Tis profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains on your account; and should I labour with you upon my own account, in expectation of a return, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security.

In deference to Hume, Skyrms and Vanderschraaf refer to this kind of asynchronous PD as the "farmer's dilemma." It is instructive to picture it in a tree diagram.

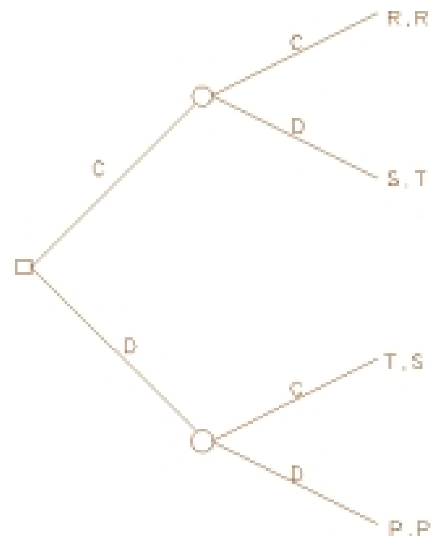


Figure 2

Here, time flows to the right. The node marked by a square indicates Player One's choice point, those marked by circles indicate when Player Two's. The moves and the payoffs to each player are exactly as in the ordinary PD, but here Player Two can choose his move according to what Player One does. Tree diagrams like Figure 2 are said to be *extensive-form* game representations, whereas the payoff matrices given previously are *normal-form* representations. As Hume's analysis indicates, making the game asynchronous does not remove the dilemma. Player One knows that if he were to choose C on the first move, Player Two would choose D on the second move (since she prefers the temptation to the reward), so he would himself end up with the sucker payoff. If Player One were to choose D, Player Two would still choose D (since she prefers the punishment to the sucker payoff), and he would end up with the punishment payoff. Since he prefers the punishment payoff to the sucker payoff, Player One will choose D on the first move and both players will end up with the payoff. This kind of "backwards" reasoning, in which the players first evaluate what would happen on the last move if various game histories were realized, and use this to determine what would happen on preceding moves applies quite broadly to games in extensive form, and a more general version of it will be discussed under iterated games below.

The farmer's dilemma can be represented in normal form by understanding Player One to be choosing between **C** and **D** and Player Two to be (simultaneously) choosing among four conditional moves: cooperate unconditionally (**Cu**), defect unconditionally (**Du**), imitate Player One's move (**I**), and do the opposite of Player One's move (**O**). The result is a two player game with the following matrix.

	Cu	**Du**	**I**	**O**
C	R,R	S,T	R,R	S,T
D	T,S	P,P	P,P	T,S

The reader may note that this game is a (multiple-move) equilibrium dilemma. The sole nash equilibrium results when Player One chooses **D** and Player Two chooses **Du**, thereby achieving for themselves the inferior payoffs of P and P. The game is not, however, a dominance dilemma. Indeed, there is no dominant move for either player. It is commonly believed that rational self-interested players will reach a nash

equilibrium even when neither player has a dominant move. If so, the farmer's dilemma is still a dilemma.

To preserve the symmetry between the players that characterizes the ordinary PD, we may wish to modify the asynchronous game. Let us take extended PD to be played in stages. First each player chooses a first move (**C** or **D**) and a second move (**Cu**, **Du**, **I** or **O**). Next a referee determines who moves first, giving each player an equal chance. Finally the outcome is computed in the appropriate way. For example, suppose Row plays (**D**,**O**) (meaning that he will defect if he moves first and do the opposite of his opponent if he moves second) and Column plays (**C**,**Du**). Then Row will get P if he goes first and T if he goes second, which implies that his expected payoff is 1/2(P+T). Column will get S if she goes first and P if she goes second, giving her an expected payoff of 1/2(P+S). It is straightforward, but tedious, to calculate the entire eight by eight payoff matrix. After doing so, the reader may observe that, like the farmer's dilemma, the symmetric form of the extended PD is an equilibrium dilemma, but not a dominance dilemma. The sole nash equilibrium occurs when both players adopt the strategy (**D**, **Du**), thereby achieving the inferior payoffs of (P,P).

Transparency

Another way that conditional moves can be introduced into the PD is by assuming that players have the property that David Gauthier has labeled *transparency*. A fully transparent player is one whose intentions are completely visible to others. Nobody holds that we humans are fully transparent, but the observation that we can often successfully predict what others will do suggests that we are at least "translucent." Furthermore agents of larger scale, like firms or countries, may be more transparent than we are. Thus there may be some theoretical interest in investigations of PDs with transparent players. Such players could presumably execute conditional strategies more sophisticated than those of the (non**-** transparent)extended game players, strategies, for example that are conditional on the conditional strategies employed by others. There is some difficulty, however, in determining exactly what strategies are feasible for such players. Suppose Row adopted the strategy "do the same as Column" and Column adopted the strategy "do the opposite of Row". There is no way that both these strategies could be satisfied. On the other hand, if each adopted the strategy "imitate the other player", there are two ways the strategies could be satisfied, and there is no way to determine which of the two they would adopt. Nigel Howard, who was probably the first to study such conditional strategies systematically, avoided this difficulty by insisting on a rigidly typed hierarchy of games. At the base level we have the ordinary PD game, where each player chooses between **C** and **D**. For any game G in the hierarchy we can generate two new games RG and CG. In RG, Column has the same moves as in game G and Row can choose any function that assigns **C** or **D** to each of Column's possible moves. Similarly in CG, Row has the same moves as in G and Column has a new set of conditional moves. For example, if [PD] is the base level game, then C[PD] is the game in which Column can choose from among the strategies **Cu**, **Du**, **I** and **O** mentioned above. Howard observed that in the two third level games RC[PD] and CR[PD] (and in every higher level game) there is an equilibrium outcome giving each player R. In particular, such an equilibrium is reached when one player plays **I** and the other cooperates when his opponent plays **I** and defects when his opponent plays **Cu**, **Du** or **O**. Notice that this last strategy is tantamount to Danielson's *reciprocal cooperation*.

The lesson of all this for rational action is not clear. Suppose two players in a PD were sufficiently

transparent to employ the conditional strategies of higher level games. How do they decide what level game to play? Who chooses the imitation move and who chooses reciprocal cooperation? To make a move in a higher level game is presumably to form an intention observable by the other player. But why should either player expect the intention to be carried out if there is benefit in ignoring it?

Conditional strategies have a more convincing application when we take our inquiry as directed, not towards playing the PD, but as designing agents who would play it well with a variety of likely opponents. This is the viewpoint of Danielson. A conditional strategy is not an intention that a player forms as a move in a game, but a deterministic algorithm defining a kind of player. Indeed, one of the lessons of the PD may be that transparent agents are better off if they can form irrevocable "action protocols" rather than always following the intentions they may form at the time of action. Danielson does not limit himself *a priori* to strategies within Howard's hierarchy. An agent is simply a computer program, which can contain lines permitting other programs to read and execute it. We could easily write two such programs, each designed to determine whether its opponent plays C or **D** and to do the opposite. What happens when these two play a PD depends on the details of implementation, but it is likely that they will be "incoherent," i.e., they will enter endless loops and be unable to make any move at all. To be successful a program *should* be able to move when paired with a variety of other programs, including copies of itself, and it should be able to get valuable outcomes. Programs implementing **I** and **O** in a straightforward way are not likely to succeed because when paired with each other they will be incoherent. Programs implementing **Du** are not likely to succeed because they get only P when paired with their clones. Those implementing **Cu** are not likely to succeed because they get only S when paired with programs that recognize and exploit their unconditionally cooperative nature. There is some vagueness in the criteria of success. In Howard's scheme we could compare a conditional strategy with all the possible alternatives of that level. Here, where any two programs can be paired, that approach is senseless. Nevertheless, certain programs seem to do well when paired with a wide variety of players. One is a version of the strategy that Gauthier has advocated as *constrained maximization*. The idea is that a player j should cooperate if the other would cooperate if j did, and defect otherwise. As stated, this appears to be a strategy for the RC[PD] or CR[PD] games. It is not clear how a program implementing it would move (if indeed it does move) when paired with itself. Danielson is able to construct an approximation to *constrained maximization*, however, that does cooperate with itself. Danielson's program (and other implementations of *constrained maximization*) cannot be coherently paired with everything. Nevertheless it does move and score well against familiar strategies. It cooperates with **Cu** and itself and it defects against **Du**. If it is coherently paired it seems guaranteed a payoff no worse than P.

A second successful program models Danielson's *reciprocal cooperation*. Again, it is not clear that the strategy (as formulated above) allows it to cooperate (or make any move) with itself, but Danielson is able to construct an approximation that does. The (approximate) *reciprocal cooperation* does as well as (approximate) *constrained maximization* against itself, **Du** and *constrained maximization*. Against **Cu** it does even better, getting T where *constrained maximization* got only R.

Iteration

Many of the situations that are alleged to have the structure of the PD, like defense appropriations of

military rivals or price setting for duopolistic firms are better modeled by an iterated version of the game in which players play the PD repeatedly, retaining access at each round to the results of all previous rounds. In these iterated PDs (hence forth IPDs) players who defect in one round can be "punished" by defections in subsequent rounds and those who cooperate can be rewarded by cooperation. Thus the appropriate strategy for rationally self-interested players is no longer obvious. The theoretical answer to this question, it turns out, depends strongly on the definition of IPD employed and the knowledge attributed to rational players.

An IPD can be represented in extensive form by a tree diagram like the one for the farmer's dilemma above.

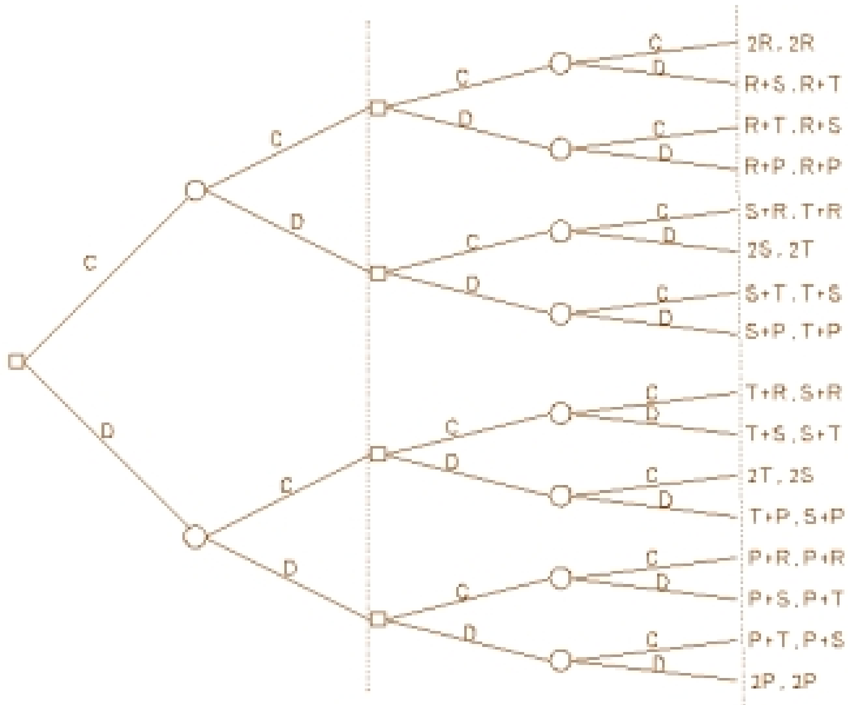


Figure 3

Here we have an IPD of length two. The end of each of the two rounds of the game is marked by a dotted vertical line. The payoffs to each of the two players (obtained by adding their payoffs for the two rounds) are listed at the end of each path through the tree. The representation differs from the previous one in that the two nodes on each branch within the same division mark simultaneous choices by the two players. Since neither player knows the move of the other at the same round, the IPD does not qualify as one of the game theorist's standard "games of perfect information." If the players move in succession rather than simultaneously (which we might indicate by removing the dotted vertical lines), the resulting game is an iterated farmer's dilemma, which does meet the game theorist's definition and which shares many of the features that make the IPD interesting.

Like the farmer's dilemma, an IPD can, in theory, be represented in normal form by taking the players' moves to be *strategies* telling them how to move if they should reach any node at the end of a round of the game tree. Like the farmer's dilemma, an IPD can, in theory, be represented in normal form by taking the players' moves to be strategies telling them how to move at each node of the game tree. The number of strategies increases very rapidly with the length of the game so that it is impossible in practice to write out the normal form for all but the shortest IPD's. Every pair of strategies determines a "play" of the game, i.e.,

a path through the extensive-form tree.

In a game like this, the notion of nash equilibrium loses some of its privileged status. Recall that a pair of moves is a nash equilibrium if each is a best reply to the other. Let us extend the notation used in the discussion of the asynchronous PD and let **Du** be the strategy that calls for defection at every node of an IPD. It is easy to see that **Du** and **Du** form a nash equilibrium. But against **Du**, a strategy that calls for defection unless the other player cooperated at, say, the fifteenth node, would determine the same play (and therefore the same payoffs) as **Du** itself does. The components that call for cooperation never come into play, because the other player does not cooperate on the fifteenth (or any other) move. Similarly, a strategy calling for cooperation after the second cooperation by itself does equally well. Thus these strategies and many others form nash equilibria with **Du**. There is a sense in which these strategies are clearly not equally rational. Although they yield the same payoffs at the nodes along the path representing the actual play, they would not yield the same payoffs if other nodes had been reached. If Player One *had* cooperated in the past, that would still provide no good reason for him to cooperate now. A nash equilibrium requires only that the two strategies are best replies to each other as the game actually develops. A stronger solution concept for extensive-form games requires that the two strategies would still be best replies to each other no matter what node on the game tree were reached. This notion of *subgame-perfect equilibrium* is defined and defended in Selten, 1975. It can be expressed by saying that the strategy-pair is a nash equilibrium for every subgame of the original game, where a subgame is the result of taking a node of the original game tree as the root, pruning away everything that does not descend from it.

Given this new, stronger solution concept, we can ask about the solutions to the IPD. There is a significant theoretical difference on this matter between IPDs of fixed, finite length, like the one pictured above, and those of infinite or indefinitely finite length. In games of the first kind, one can prove by an argument known as *backwards induction* that **Du**,**Du** is the only subgame perfect equilibrium. Suppose the players know the game will last exactly *n* rounds. Then, no matter what node have been reached, at round n-1 the players face an ordinary ("one-shot") PD, and they will defect. At round n-2 the players know that, whatever they do now, they will both defect at the next round. Thus it is rational for them to defect now as well. By repeating this argument sufficiently many times, the rational players deduce that they should defect at every node on the tree. Indeed, since at every node defection is a best response to any move, there can be no other subgame-perfect equilibria.

In practice, there is not a great difference between how people behave in long fixed-length IPDs (except in the final few rounds) and those of indeterminate length. This suggests that some of the rationality and common knowledge assumptions used in the backwards induction argument (and elsewhere in game theory) are unrealistic. There is a considerable literature attempting to formulate the argument carefully, examine its assumptions, and to see how relaxing unrealistic assumptions might change the rationally acceptable strategies in the PD and other games of fixed length. (For a small sample, see Bovens, Kreps *et al*, Kreps and Wilson, Pettit and Sugden, Selten 1978, Rabinowicz, and Binmore 1997).

Player One's belief that there is a slight chance that Two might pursue an "irrational" strategy other than continual defection could make it rational for her to cooperate frequently herself. Indeed, even if One were certain of Two's rationality, One's belief that there was some chance that Two believed she harbored such doubts could have the same effect. Thus the argument for continual defection in the IPD of fixed length

depends on complex iterated claims of certain knowledge of rationality. An even more unrealistic assumption, noted by Rabinowicz and others, is that each player continue to believe that the other will choose rationally on the next move even after evidence of irrational play on previous moves. For example, it is assumed that, at the node reached after a long series of moves (C,C), ...,(C,C), Player One will choose D despite never having done so before.

Some have used these kinds of observation to argue that the backwards induction argument shows that standard assumptions about rationality (with other plausible assumptions) are inconsistent or self-defeating. For (with plausible assumptions) one way to ensure that a rational player will doubt one's own rationality is to behave irrationally. In the fixed-length IPD, for example, Player One may be able to deduce that, if she were to follow an appropriate "irrational" strategy, Player Two would rationally react so that they can achieve mutual cooperation in almost all rounds. So our assumptions seem to imply both that Player One should continually defect and that she would do better if she didn't. (See Skyrms 1990, pp. 125-139 and Bicchieri 1989.)

Infinite Iteration

One way to avoid the dubious conclusion of the backwards induction argument without delving too deeply into conditions of knowledge and rationality is to consider infinitely repeated PD's. No human agents can actually play an infinitely repeated game, of course, but the infinite IPD has been considered an appropriate way to model a series of interactions in which the participants never have reason to think the current interaction is their last. In this setting a pair of strategies determines an infinite path through of the game tree. If the payoffs of the one-shot game are positive, their total along any such path is infinite. This makes it somewhat awkward to compare strategies. If we confine ourselves to those strategies that can be implemented by mechanical devices (with finite memories and speeds of computation), however, it turns out that the sequence of payoffs to each player will always, after a finite number of rounds, cycle repeatedly through a particular finite sequence of payoffs. The relative value of such infinite sequences of payoffs can then be identified with the average value of the payoffs in one cycle. This value reflects the limit of average payoff per round as the number of rounds increases. (See Binmore 1992, page 365 for further justification.) Since there is no last round, it is obvious that backwards induction does not apply to the infinite IPD.

Indefinite Iteration

Most contemporary investigations the IPD take it to be neither infinite nor of fixed finite length but rather of indeterminate length. This is accomplished by including in the game specification a probability p (the "shadow of the future") such that at each round the game will continue with probability p. Alternatively, a "discount factor" p is applied to the payoffs after each round so that present payoffs are valued more highly than future. Mathematically, it makes little difference whether p is regarded as a probability of continuation or a discount on payoffs. The value of cooperation at a given stage in an IPD clearly depends on the odds of meeting one's opponent in later rounds. (This has been said to explain why the level of courtesy is higher in a village than a metropolis and why customers tend leave better tips in local restaurants than distant ones.) As p approaches zero, the IPD becomes a one-shot PD, and the value of defection increases. As p approaches one the IPD becomes an infinite IPD, and the value of defection decreases. It is also customary

to insist that the game has the property labeled RCA above, so that (in the symmetric game) players do better by cooperating on every round than they would do by "taking turns" -- you cooperate while I defect and then I cooperate while you defect.

There is a controversial claim, originating in Kavka and Sobel and still repeated, that the backwards induction argument applies even to indefinitely repeated IPD's, as long as an upper bound to the length of the game is common knowledge. For if b is such an upper bound, then if the players were to get to stage b they would know that it was the last round and they would defect; if they were to get to stage b-1 they would know that their behavior on this round cannot affect the decision to defect on the next, and so they would defect; and so on. It is an easy matter to compute upper bounds on the number of interactions in real-life situations. For example, since shopkeeper Jones cannot make more than one sale a second and since he will live less than a thousand years, he and customer Smith can calculate that they cannot possibly conduct more than 10^{12} transactions. The Kavka-Sobel argument, however, seems to rest not just on common knowledge of an upper bound, but common knowledge of a "sharp" upper bound. Although there is zero probability of 10^{12}+1 interactions between Jones and Smith, the probability of the counterfactual "If there were 10^{12} interactions, there would be another" is not zero. Any world odd enough to have permitted 10^{12} interactions is very likely to permit 10^{12}+1. So the backwards induction cannot get started. As we have seen, the in standard treatment of the indefinite IPD it is assumed that there is a finite probability p of the game's continuing at each stage t. Becker and Cudd suggest that the constant p be replaced by a function p(t) that decreases at each round of the game. If p(t) gets sufficiently low then the Kavka-Sobel backwards induction argument does go through. For each IPD, however, there is a threshold p such that as long p(t) always remains above p, the game will admit cooperative solutions. (It is important in this context to give p(t) the counterfactual interpretation suggested above. To take it--as Becker and Cudd appear to do--as the conditional probability of the game's reaching t+1 given that it has reached t is to deny the Kavka-Sobel assumption that there is a stage that can be reached with probability zero.)

The iterated version of the PD was discussed from the time the game was devised, but interest accelerated after influential publications of Robert Axelrod in the early eighties. Axelrod invited professional game theorists to submit computer programs for playing IPDs. All the programs were entered into a tournament in which each played every other (as well as a clone of itself and a strategy that cooperated and defected at random) hundreds of times. It is easy to see that in a game like this no strategy is "best" in the sense that its score would be highest among any group of competitors. If the other strategies never consider the previous history of interaction in choosing their next move, it would be best to defect unconditionally. If the other strategies all begin by cooperating and then "punish" any defection against themselves by defecting on all subsequent rounds, then a policy of unconditional cooperation is better. Nevertheless, as in the transparent game, some strategies have features that seem to allow them to do well in a variety of environments. The strategy that scored highest in Axelrod's initial tournament, Tit for Tat (henceforth **TFT**), simply cooperates on the first round and imitates its opponent's previous move thereafter. More significant than **TFT**'s initial victory, perhaps is the fact that it won Axelrod's second tournament, whose sixty three entrants were all given the results of the first tournament. In analyzing the his second tournament, Axelrod noted that each of the entrants could be assigned one of five "representative" strategies in such a way that a strategy's success against a set of others can be accurately predicted by its success against their representative. As a further demonstration of the strength of **TFT**, he calculated the scores each strategy would have received in tournaments in which one of the representative strategies was five times as common as in the original

tournament. **TFT** received the highest score in all but one of these hypothetical tournaments.

Axelrod attributed the success of **TFT** to four properties. It is *nice*, meaning that it is never the first to defect. The eight nice entries in Axelrod's tournament were the eight highest ranking strategies. It is *retaliatory*, making it difficult for it to be exploited by the rules that were not nice. It is *forgiving*, in the sense of being willing to cooperate even with those who have defected against it (provided their defection wasn't in the immediately preceding round). An unforgiving rule is incapable of ever getting the reward payoff after its opponent has defected once. And it is *clear*, presumably making it easier for other strategies to predict its behavior so as to facilitate mutually beneficial interaction.

Suggestive as Axelrod's discussion is, it is worth noting that the ideas are not formulated precisely enough to permit a rigorous demonstration of the supremacy of **TFT**. One doesn't know, for example, the extent of the class of strategies that might have the four properties outlined, or what success criteria might be implied by having them. It is true that if one's opponent is playing **TFT** (and the shadow of the future is sufficiently large) then one's maximum payoff is obtained by a strategy that results in mutual cooperation on every round. Since **TFT** is itself one such strategy this implies that **TFT** forms a nash equilibrium with itself in the space of all strategies. But that does not particularly distinguish **TFT**, for **Du**, **Du** is also a nash equilibrium. Indeed, a "folk theorem" of iterated game theory (now widely published -- see, for example, Binmore 1992, pp. 373-377) implies that, for any p, $0 \leq p \leq 1$ there is a nash equilibrium in which p is the fraction of times that mutual cooperation occurs. Indeed **TFT** is, in some respects, *worse* than many of these other equilibrium strategies, because the folk theorem can be sharpened to a similar result about subgame perfect equilibria. And **TFT**, **TFT** is, in general, *not* subgame perfect. For, were one **TFT** player (*per impossible*) to defect against another, the second would have done better as an unconditional cooperator.

Iteration With Error

In a survey of the field several years after the publication of the results reported above, Axelrod and Dion, chronicle several successes of **TFT** and modifications of it. They conclude that "research has shown that many of Axelrod's findings...can be generalized to settings that are quite different from the original two-player iterated Prisoner's Dilemma game." But in several reasonable settings **TFT** has serious drawbacks. One such case, noted in the Axelrod and Dion survey, is when attempts are made to incorporate the plausible assumption that players are subject to errors of execution and perception. There are a number of ways this can be done. Bendor, for example, considers "noisy payoffs." When a player cooperates while its opponent defects its payoff is S+e, where e is a random value whose expected value is 0. Each player infers the other's move from its own payoff, and so if e is sufficiently high its inference may be mistaken. Sugden (pp 112-115) considers players who have a certain probability of making an error of execution that is apparent to them but not their opponents. Such players can adopt strategies by which they "atone" for mistaken defections by being more cooperative on later rounds than they would be after intended defection. Assuming that players themselves cannot distinguish a mistaken move or observation from a real one, however, the simplest way to model the inevitability of error is simply to forbid completely deterministic strategies like **TFT**, replacing them with "imperfect" counterparts, like "imitate the other player's last move with 99% probability and oppose it with 1% probability." **Imperfect TFT** is much less attractive than its

deterministic sibling, because when two **imperfect TFT** strategies play each other, an "error" by either one will set off a long chain of moves in which the players take turns defecting. In a long iterated game between two **imperfect TFT**'s with any probability *p* of error, $0 < p < 1/2$, players will approach the same average payoffs as in a game between two strategies that choose randomly between cooperation and defection, namely 1/4(R+P+S+T). That is considerably worse than the payoff of R, that results when *p*=0.

The predominant view seems to be that, when imperfection is inevitable, successful strategies will have to be more forgiving of defections by their opponents (since those defections might well be unintended). Molander 1985 demonstrates that strategies that mix **TFT** with **Cu** do approach a payoff of R as the probability of error approaches zero. When these mixes play each other, they benefit from higher ratios of **Cu** to **TFT**, but if they become too generous, they risk exploitation by "stingy" strategies that mix **TFT** with defection. Molander calculates that when the mix is set so that, following a defection, one cooperates with probability g(R,P,T,S)=min{1-(T-R)/(R-S), (R-P)/(T-P)}, the generous strategies will get the highest score with each other that is possible without allowing stingy strategies to do better against them than TFT does. Following Nowak and Sigmund, we label this strategy **generous TFT**, or **GTFT**.

The idea that the presence of imperfection induces greater forgiveness or generosity is only plausible for low levels of imperfection. As the level of imperfection approaches 1/2, **Imperfect TFT** becomes indistinguishable from the random strategy, for which the very ungenerous **Du** is the best reply. A simulation by Kollock seems to confirm that at high levels of imperfection, more stinginess is better policy than more forgiveness. But Bendor, Kramer and Swistak note that the strategies employed in the Kollock simulation are not representative and so the results must be interpreted with caution.

A second idea is that an imperfect environment encourages strategies to observe their opponent's play more carefully. In a tournament similar to Axelrod's (Donninger) in which each player's moves were subject to a 10% chance of alteration, **TFT** finished sixth out twenty-one strategies. As might have predicted on the dominant view, it was beaten by the more generous **Tit-for-Two-Tats** (which cooperates unless defected against twice in a row). It was also beaten, however, by two versions of **Downing**, a program that bases each new move on its best estimate how responsive its opponent has been to its previous moves. In Axelrod's two original tournaments, **Downing** had ranked near the bottom third of the programs submitted. Bendor (1987) demonstrates deductively that against imperfect strategies there are advantages to basing one's probability of defection on longer histories than does **TFT**.

One clever implementation of the idea that a strategies in an imperfect environment should pay attention to their previous interactions is the family of "Pavlovian" strategies investigated by Kraines and Kraines. For each natural number *n*, *n*-Pavlov, or $\mathbf{P}_n$, adjusts its probability of cooperation in units of 1/*n*, according to how well it fared on the previous round. More precisely, if $\mathbf{P}_n$ was cooperating with probability p on the last round, then on this round it will cooperate with probability p[+](1/*n*) if it received the reward payoff on the previous round, p[-](1/*n*) if it received the punishment payoff, p[+](2/*n*)if it received the temptation payoff, and p[-](2/*n*) if it received the sucker payoff. [+] and [-] are bounded addition and subtraction, i.e., x[+]y is the sum x+y unless that number exceeds one, in which case it is one (or as close to one as the possibility of error allows), and x[-]y is similarly either x-y or close to zero. Strictly speaking, $\mathbf{P}_n$ is not fully specified until an initial probability of cooperation is given, but for most purposes the value of that parameter

becomes insignificant in sufficiently long games and can be safely ignored. It may appear that $\mathbf{P}_n$ requires far more computational resources to implement than, say, **TFT**. Each move for the latter depends on only on its opponent's last move, whereas each move for $\mathbf{P}_n$ is a function of the entire history of previous moves of both players. $\mathbf{P}_n$, however, can always calculate its next move by tracking only its current probability of cooperation and its last payoff. As its authors maintain, this seems like "a natural strategy in the animal world." One can calculate that for n>1, $\mathbf{P}_n$ does better against the random strategy than does **TFT**. More generally, $\mathbf{P}_n$ does as well or better than **TFT** against the generous unresponsive strategies **Cp** that always cooperate with fixed probability p $\geq$ 1/2 (because an occasional temptation payoff can teach it to exploit the unresponsive strategies.) In these cases the "slow learner" versions of Pavlov with higher values of *n* do slightly better than the "fast learners" with low values. Against responsive strategies, like other Pavlovian strategies and **TFT**, $\mathbf{P}_n$ and its opponent eventually reach a state of (almost) constant cooperation. The total payoff is then inversely related to the "training time," i.e., the number of rounds required to reach that state. Since training time of $\mathbf{P}_n$ varies exponentially with n, Kraines and Kraines maintain that $\mathbf{P}_3$ or $\mathbf{P}_4$ are to be preferred to other Pavlovian strategies, and are close to "ideal" IPD strategies. It should be noted, however, that when (deterministic) **TFT** plays itself, no training time at all is required, whereas when a Pavlovian strategy plays **TFT** or another Pavlov, the training time can be large. Thus the cogency of the argument for the superiority of Pavlov over **TFT** depends on the observation that its performance shows less degradation when subject to imperfections. It is also worth remembering that no strategy is best in every environment, and the criteria used in defense of various strategies in the IPD are vague and heterogeneous. One advantage of the evolutionary versions of the IPD discussed in the next section is that they permit more careful formulation and evaluation of success criteria.

Evolution

Perhaps the most active area of research on the PD concerns evolutionary versions of the game. A population of players employing various strategies play IPDs among themselves. The lower scoring strategies decrease in number, the higher scoring increase, and the process is repeated. Thus success in an evolutionary PD (henceforth EPD), requires doing well with other successful strategies, rather than doing well with a wide range of strategies.

The initial population in an EPD can be represented by a set of pairs $\{(p_1,s_1),...(p_n,s_n)\}$ where $p_1,...,p_n$ are the proportions of the population playing strategies $s_1,...,s_n$, respectively. The description of EPD's given above does not specify exactly how the population of strategies is to be reconstituted after each IPD. The usual assumption, and the most sensible one for biological applications, is that a score in any round indicates the relative number of "offspring" in the next. It is assumed that the size of the entire population stays fixed, so that births of more successful strategies are exactly offset by deaths of less successful ones. This amounts to the condition that the proportion p_i* of each strategy s_i in the successor population is determined by the equation $p_i^*=p_i(V_i/V)$, where V_i is the score of s_i in the previous round and V is the average of all scores in the population. Thus every strategy that scores above the population average will increase in number and every one that scores below the average will decrease. This kind of evolution is referred to as "replication dynamics" or evolution according to the "proportional fitness" rule. Other rules of

evolution are possible. Bendor and Swistak argue that, for social applications, it makes more sense to think of the players as switching from one strategy to another rather than as coming into and of existence. Since rational players would presumably switch only to strategies that received the highest payoff in previous rounds, only the highest scoring strategies would increase in numbers. Discussion here, however will primarily concern EPDs with the proportional fitness rule.

Axelrod, borrowing from Trivers and Maynard Smith, includes a description of the EPD with proportional fitness, and a brief analysis of the evolutionary version of his IPD tournament. For Axelrod, the EPD provides one more piece of evidence in favor of **TFT**:

> TIT FOR TAT had a very slight lead in the original tournament, and never lost this lead in simulated generations. By the one-thousandth generation it was the most successful rule and still growing at a faster rate than any other rule.

Axelrod's EPD tournament, however, incorporated several features that might be deemed artificial. First, it permitted deterministic strategies in a noise-free environment. As noted above, **TFT** can be expected to do worse under conditions that model the inevitability of error. Second, it began with only the 63 strategies from the original IPD tournament. Success against strategies concocted in the ivory tower may not imply success against all those that might be found in nature. Third, the only strategies permitted to compete at a given stage were the survivors from the previous stage. A more realistic model, one might argue, would allow new "mutant" strategies to enter the game at any stage. Changing this third feature might well be expected to hurt **TFT**. For a large growth in the **TFT** population would make it possible for mutants employing more naive strategies like **Cu** to regain a foothold, and the presence of these naifs in the population might favor nastier strategies like **Du** over **TFT**.

Nowak and Sigmund simulated two kinds of tournaments that avoid the three questionable features. The first examined the family of "reactive" strategies. For any probabilities y, p, and q, **R**(y,p,q) is the strategy of cooperating with probability y in the first round and thereafter with probability p if the other player has cooperated in the previous round, and with probability q if she has defected. This is a broad family, including many of the strategies already considered. **Cu**, **Du**, **TFT**, and **Cp** are **R**(1,1,1), **R**(0,0,0), **R**(1,1,0), and **R**(p,p,p). To capture the inevitability of error, Nowak and Sigmund exclude the deterministic strategies, where p and q are exactly 1 or 0, from their tournaments. As before, if the game is sufficiently long (and p and q are not integers), the first move can be ignored and a reactive strategy can be identified with its p and q values. Particular attention is paid to the strategies close to Molander's **GTFT** described above, where p=1 and q=min{1-(T-R)/(R-S), (R-P)/(T-P). The first series of Nowak and Sigmund's EPD tournaments begin with representative samples of reactive strategies. For most such tournaments, they found that evolution led irreversibly to **Du**. Those strategies **R**(p,q) closest to **R**(0,0) thrived while the others perished. When one of the initial strategies is very close to **TFT**, however, the outcome changes.

> TFT and all other reciprocating strategies (near (1,0)) seem to have disappeared. But an embattled minority remains and fights back. The tide turns when 'suckers' are so decimated that exploiters can no longer feed on them. Slowly at first, but gathering momentum, the reciprocators come back, and the exploiters now wane. But the **TFT**-like strategy that caused

> this reversal of fortune is not going to profit from it: having eliminated the exploiters, it is robbed of its mission and superseded by the strategy closest to GTFT. Evolution then stops. Even if we introduce occasionally 1% of another strategy it will vanish.

On the basis of their tournaments among reactive strategies, Nowak and Sigmund conjectured that, while **TFT** is essential for the emergence of cooperation, the strategy that actually underlies persistent patterns of cooperation in the biological world is more likely to be **GTFT**.

A second series of simulations with a wider class of strategies, however, forced them to revise their opinion. The strategies considered in the second series allowed each player to base its probability of cooperation on its own previous move as well as its opponent's. A strategy can now be represented as $\mathbf{S(p_1,p_2,p_3,p_4)}$ where p_1, p_2, p_3, p_4 are the probabilities of defecting after outcomes (**C**,**C**), (**C**,**D**), (**D**,**C**), and (**D**,**D**), respectively i.e., after receiving the reward, sucker, temptation and punishment payoffs. (Again, we can ignore the probability of defecting on the first move as long as the p_i's are not zero or one. The initial population in these tournaments all play the random strategy **S**(.5,.5,.5,.5) and after every 100 generations a small amount of a randomly chosen (non-deterministic) mutant is introduced, and the population evolves by proportional fitness. The results are quite different than before. After 10^7 generations, a state of steady mutual cooperation was reached in 90% of the simulation trials. But less than 8.3% of these states were populated by players using **TFT** or **GTFT**. The remaining 91.7% were dominated by strategies close to **S**(1,0,0,1). This is the just the Pavlovian strategy $\mathbf{P}_1$ of Kraines and Kraines, which cooperates after receiving R or T and defects after receiving P or S. Kraines and Kraines had been somewhat dismissive of $\mathbf{P}_1$. They recall that Rapoport and Chammah, who identified it early in the history of game theory had labeled it "simpleton" and remark that "the appellation is well deserved". Indeed, $\mathbf{P}_1$ has the unfortunate characteristic of trying to cooperate with **Du** on every other turn, and against **TFT** it can get locked into the inferior repeating series of payoffs T,P,S,T,P,S,... . But Nowak and Sigmund's simulations suggest that these defects do not matter very much in evolutionary contexts. One reason may be that $\mathbf{P}_1$ helps to make its environment unsuitable for its enemies. **Du** does well in an environment with generous strategies, like **Cu** or **GTFT**. **TFT**, as we have seen, allows these strategies to flourish, which could pave the way for **Du** Thus, although **TFT** fares less badly against **Du** than $\mathbf{P}_1$ does, $\mathbf{P}_1$ is better at keeping its environment free of **Du**.

Simulations in a universe of deterministic strategies yield results quite different than those of Nowak and Sigmund. Bruce Linster (1992 and 1994) suggests that natural classes of strategies and realistic mechanisms of evolution can be defined by representing strategies as simple *Moore machines*. For example, P1 is represented by the machine pictured below.

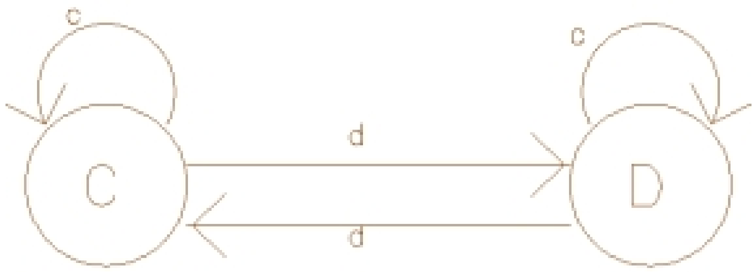


Figure 4

This machine has two states, indicated by circles. It begins in the leftmost state. The C in the left circle means that the machine cooperates on the first move. The arrow leading from the left to the right circle indicates that machine defects (enters the *D*) after it has cooperated (been in the *C* state) and its opponent has defected (the arrow is labeled by *d*). Linster has conducted simulations of evolutionary PD's among the strategies that can be represented by two-state Moore machines. It turns out that these are exactly the deterministic versions of the S strategies of Nowak and Sigmund. Since the strategies are deterministic, we must distinguish between the versions that cooperate on the first round and those that defect on the first round. Among the first round cooperators, S(0,0,0,0), S(0,0,0,1), S(0,0,1,0) and S(0,0,0,1) all represent the strategy **Cu** of unconditional cooperation. Similarly, four of the first-round defectors all represent **Du**. Each of the other **S**(p_1,p_2,p_3,p_4) where p_1, p_2, p_3, p_4 are either zero or one represent unique strategies. By deleting the six duplicates from the thirty-two deterministic versions of Nowak and Sigmund's strategies, we obtain the twenty-six "two-state" strategies considered by Linster.

Linster simulated a variety of EPD tournaments among the two-state strategies. Some used "uniform mutation" in which each strategy in the population has an equal probability *m* of mutating into any of the other strategies. Some used "stylized mutation" in which the only mutations permitted are those that can be understood as the result of a single "broken link" in the Moore machine diagrams. In some, mutations were assumed to occur to a tiny proportion of the population at each generation; in others the "mutants" represented an invading force amounting to one percent of the original population. In some, a penalty was levied for increased complexity in the form of reduced payoffs for machines requiring more states or more links. As one might expect, results vary somewhat depending on conditions. There are some striking differences, however, between all of Linster's results and those of Nowak and Sigmund. In Linster's tournaments, no single strategy ever dominated the surviving populations in the way that $\mathbf{P}_1$ and **GTFT** did in Nowak and Sigmund's. The one strategy that did generally come to comprise over fifty percent of the population was the initially-cooperating version of S(0,1,1,1). This is a strategy whose imperfect variants seem to have been remarkably uncompetitive for Nowak and Sigmund. It has been frequently discussed in the game theory literature under the label **GRIM** or **TRIGGER**. It cooperates until its opponent has defected once, and then defects for the rest of the game. According to Skyrms (1998) and Vanderschraaf, both Hobbes and Hume identified it as the strategy that underlies our cooperative behavior in important PD-

like situations. The explanation for the discrepancy between **GRIM's** performance for Linster and for Nowak and Sigmund probably has to do with its sharp deterioration in the presence of error. In a match between two imperfect **GRIM**s, an "erronious" defection by either leads to a long string of mutual defections. Thus, in the long run imperfect **GRIM** does poorly against itself. The other strategies that survived (in lesser numbers) Linster's tournaments are **TFT**, $\mathbf{P}_1$, **Cu**, and the initially-cooperative S(0,1,1,0). (Note that imperfect **GRIM** is also likely to do poorly against imperfect versions of these.) The observation that evolution might lead to a stable mix of strategies (perhaps each serving to protect others against particular types of invaders) rather than a single dominant strategy is suggestive. Equally suggestive is the result obtained under a few special conditions in which evolution leads to a recurring cycle of population mixes.

One might expect it to be possible to predict the strategies that will prevail in EPDs meeting various conditions, and to justify such predictions by formal proofs. Until recently, however, mathematical analyses of the EPD have been plagued by conceptual confusions about "evolutionary stability," the condition under which, as Nowak and Sigmund say, "evolution stops". Axelrod and Axelrod & Hamilton claim to show that **TFT** is evolutionarily stable. Selten 1983, includes an example of a game with no evolutionarily stable strategy, and Selten's argument that there is no such strategy clearly applies to the **EPD** and other evolutionary games. Boyd and Lorberbaum and Farrell and Ware present still different proofs demonstrating that no strategies for the **EPD** are evolutionarily stable. Unsurprisingly, the paradox is resolved by observing that the three groups of authors each employ slightly different conceptions of evolutionary stability. The conceptual tangle is unraveled in a series of papers by Bendor and Swistak. Two stability concepts are described and applied to the EPD below. Readers who wish to compare these with some others that appear in the literature may consult the following brief guide:

> Concepts of Stability in Evolutionary Games.

A strategy s for an evolutionary game has *universal strong narrow stability* ("usn-stability") if a population playing strategy s will, under any rule of evolution, drive to extinction any sufficiently small group of invaders all of which play the same strategy. An evolutionary game has usn-stability just in case it meets a simple condition on payoffs identified by Maynard Smith:

> MS) For all strategies j, V(i,i) > V(j,i) or both V(i,i)=V(j,i) and V(i,j)>V(j,j)

(Here, and in what follows, the notation V(i,j) indicates the payoff to strategy i when i plays j.) MS says that any invaders do strictly worse against the natives than the natives themselves do against the natives or else they get exactly the same payoff against the natives as the natives themselves do, but the native does better against the invader than the invader himself does.

For any strategy i in the IPD (or indeed in any iterated finite game), however, there are strategies j different from i such that j mimics the way i plays when it plays against i or j. The existence of these "neutral mutants" implies that MS cannot be satisfied and so no EPD has usn-stability. This argument, of course, uses the assumption that any strategy in the iterated game is a possible invader. There may be good reason to restrict the available strategies. For example, if the players are assumed to have no knowledge of

previous interactions, then it may be appropriate to restrict available strategies to the unconditional ones. Since a pair of players then get the same payoffs in every round of an iterated game, we may as well take each round of the evolutionary game to be one-shot games between every pair of players rather than iterated games. Indeed, this is the kind of evolutionary game that Maynard Smith himself considered. In this framework, any strategy S such that (S,S) is a strict nash equilibrium in the underlying one-shot game (including unconditional defection in the PD) meets the MS condition. Thus MS and usn-stability are non-trivial conditions in some contexts.

A strategy s has *restricted weak broad stability)* (rwb-stability) if, when evolution proceeds according to the proportional fitness rule and the native population is playing s, any (possibly heterogeneous) group of invaders of sufficiently small size will fail to drive the natives to extinction. This condition turns out to be equivalent to a weakened version of MS identified by Bendor and Swistak.

> BS) For all strategies j, $V(i,i) > V(j,i)$ or both $V(i,i)=V(j,i)$ and $V(i,j) \geq V(j,j)$

BS and rwb-stability are non-trivial conditions in the more general evolutionary framework: strategies for the EPD that satisfy rwb-stability do exist. This does not particularly vindicate any of the strategies discussed above, however. Bendor and Swistak prove a result analogous to the folk theorem mentioned previously: If the shadow of the future is sufficiently large, there are rwb-stable strategies supporting any degree of cooperation from zero to one. One way to distinguish among the strategies that meet BS is by the size of the invasion required to overturn the natives, or, equivalently, by the proportion of natives required to maintain stability. Bendor and Swistak show that this number, the *minimal stabilizing frequency*, never exceeds 1/2: no population can resist every invading group as large as itself. They maintain that this result does allow them to begin to provide a theoretical justification for Axelrod's claims. They are able to show that, as the shadow of the future approaches one, any strategy that is nice (meaning that it is never first to defect) and retaliatory (meaning that it always defects immediately after it has been defected against) has a minimal stabilizing frequency approaching one half. **TFT** has both these properties and, in fact, they are the first two of the four properties Axelrod cited as instrumental to **TFT**'s success. There are, of course, many other nice and retaliatory strategies, and there are strategies (like $\mathbf{P}_1$) that are not retaliatory but still satisfy rwb-stability. But Bendor and Swistak are at least able to show that any "maximally robust" strategy, i.e., any strategy whose minimum stabilizing frequency approaches one half, chooses cooperation on all but a finite number of moves in an infinitely repeated PD.

Bendor and Swistak's results must be interpreted with some care. First, one should keep in mind that no probabilistic or noise- sensitive strategies can fit the definitions of either "nice" or "retaliatory" strategies. Furthermore, imperfect versions of **TFT** do not satisfy rwb-stability. They can be overthrown by arbitrarily small invasions of deterministic **TFT** or, indeed, by arbitrary small invasions of any less imperfect **TFT**. Second, one must remember that the results about minimal stabilizing frequencies only concern weak stability. If the number of generations is large compared with the original population (as it often is in biological applications), a population that is initially composed entirely of players employing the same maximally robust strategy, could well admit a sequence of small invading groups that eventually reduces the original strategy to less than half of the population. At that point the original strategy could be overthrown.

It is likely that both of these caveats play some role in explaining an apparent discrepancy between the Bendor/Swistak results and the Nowak/Sigmund simulations. One would expect Bendor/Swistak's minimal stabilizing frequency to provide some indication of the length of time that a population plays a particular strategy. A strategy requiring a large invasion to overturn is likely to prevail longer than a strategy requiring only a small invasion. A straightforward calculation reveals that $\mathbf{P}_1$ has a relatively low minimum stabilizing frequency. It is overturned by invasions of unconditional defectors exceeding 10% of the population. Yet in the Nowak/Sigmund simulations, $\mathbf{P}_1$-like strategies predominate over **TFT**-like strategies. Since the simulations required imperfection and since they generated a sequence of mutants vastly larger than the original population, there is no real contradiction here. Nevertheless the discrepancy suggests that we do not yet have a theoretical understanding of EPD's sufficient to predict the strategies that will emerge under various plausible conditions.

Like usn-stability, the concept of rwb-stability can be more discriminating if it is relativized to a particular set of strategies. Molander's 1992 investigation of Schelling's many-person version of the PD, for example, restricts attention to the family $\{\mathbf{S}_1,...,\mathbf{S}_n\}$ of **TFT**-like stratgies. A player adopting $\mathbf{S}_i$ cooperates on the first round and on every subsequent round after at least i others cooperate. By construing stability as resistance to invasions by other family members, Molander is able to show that there are conditions under which a particular mix of two of the $\mathbf{S}_i$'s (one equivalent to **Du**) is uniquely stable. The significance of results like these, however, depends on the plausibility of such limitations on the set of permissible strategies.

Spatial PD's

In many social and biological situations said to be modeled by an EPD, individuals interact only with those within some geographical proximity. Areas inhabited by less successful individuals might be taken over by more successful neighbors. Alternatively, less successful individuals might adopt the strategies of their more successful neighbors. A spatial PD (SPD) attempts to add this geographical feature to the game. A small set of strategies is distributed among the cells of a grid (which may have a boundary, or be unbounded, like the surface of a sphere). At a given round the strategy in each cell plays PDs against the strategies in each of the "neighbor" cells (which may be stipulated to include the cell itself). The cell's score in that round is the sum of its payoffs in these interactions. The strategy in the cell is then replaced by the strategy of the neighbor cell with highest score, and the next round begins.

As usual, the impetus here seems to come from Axelrod. Axelrod randomly distributed four copies of each of the 63 strategies submitted to his tournament on a grid arranged so that each cell had four neighbors. For every initial random distribution, the resulting SPD eventually reached a state where the strategy in every cell was cooperating with all its neighbors, at which point no further evolution is possible. Only about ten of the 63 original strategies remained in these end-states, and they were no longer randomly distributed, but segregated into clumps of various sizes. Axelrod also showed that under special conditions evolution in an SPD can create successions of complex symmetrical patterns that do not appear to reach any steady-state equilibrium.

Nowak and May have investigated in greater detail SPD's in which the only permitted strategies are **Cu** and **Du**. (These are the strategies appropriate among individuals lacking memory or recognition skills.) They find that, for a variety of spatial configurations, and distributions of strategies evolution depends on relative payoffs in a uniform way. When the temptation payoff is sufficiently high clusters of **Du** grow and those of **Cu** shrink and when it is sufficiently low the **Du** clusters shrink and the **Cu** ones grow. For a narrow range of intermediate values, we get more of the complicated patterns noted above. The evolving patterns exhibit great variety, but for a given spatial configuration the ratio of the two strategies seems to approach the same constant value for all initial distributions of strategies and all payoffs within the special range. More recently, Nowak, Boenhoeffer and May have observed similar phenomena under a variety of error-conditions identified by Mukherji and Rajan, although cooperators seem to require lower relative temptation values to thrive under these conditions, and the level of error must be sufficiently low.

Grim, Mar and St Denis report a number of SPD simulations with a greater variety of initial strategies. In general their observations suggest that cooperative outcomes are more common in SPD's than ordinary EPD's. Simulations starting with all of the pure reactive strategies of Nowak and Sigmund (i.e., all of the strategies **R**(y,p,q) described above where y,p and q are either 0 or 1.), all ended with **TFT** (i.e., with **R**(1,1,0)) as the only survivor (though other outcomes--including one in which **Du** is the sole survivor and ones in which **Cu** and **TFT** are intermixed--are clearly possible.) Simulations of the 64 possible pure strategies in which a move may depend on the opponent's previous two moves, ended with mixed populations of survivors all of whom defect after two defections, cooperate after two cooperations and cooperate in the second round of the game. (Again, other outcomes are possible.) Simulations with many (viz., 100) evenly distributed samples of Nowak and Sigmunds mixed reactive strategies, tended to be taken over by **R**(.99,.1), which is a version of generous **TFT** with slightly less generosity than **GTFT**. Simulations beginning with a random selection of a few (viz., 8) of these strategies tended to evolve to a mixed stable or cyclic pattern dominated by a single version of generous **TFT** with considerably more generosity than **GTFT**. **R**(.99,.6) seems to have been a frequent victor.

SPD's differ from ordinary EPD's in at least two different ways. First, the "spatial" aspect of the games implies that (after the first round) players are more likely to face opponents playing the same strategies as themselves. Second, the "local winner imitation" dynamic can quickly drive to extinction strategies that would survive (and eventually dominate) under the proportional fitness rule commonly employed in EPD's. It is likely that both of these features are relevant to explain the differences between the results of SPD and EPD simulations.

At present there seems to be little theoretical understanding of the evolution of SPD's, and they have not made much contribution to biological or social theory. They have, however, given us some pretty pictures to contemplate. Several examples are accessible through the links at the end of this entry.

Bibliography

- Axelrod, Robert, "The Emergence of Cooperation Among Egoists," *The American Political Science Review*, **75** (1981): 306-318.
- Axelrod, Robert, *The Evolution of Cooperation* , New York: Basic Books, 1984.

- Axelrod, Robert and Douglas Dion, "The Further Evolution of Cooperation"*Science*, **242**(9 December 1988): 1385-1390.
- Axelrod, Robert and William Hamilton, "The Evolution of Cooperation"*Science*, **211**(27 March 1981): 1390-1396.
- Becker, Neal and Ann Cudd, "Indefinitely Repeated Games: A Response to Carroll," *Theory and Decision*, **28**(1990): 189-195.
- Bendor, Jonathan, "In Good Times and Bad: Reciprocity in an Uncertain World," *American Journal of Political Science*, **31** (1987): 531-558.
- Bendor, Jonathan, "Uncertainty and the Evolution of Cooperation," *Journal of Conflict Resolution*, **37** (1993): 709-733.
- Bendor, Jonathan, and Piotr Swistak, "Types of Evolutionary Stability and the Problem of cooperations," *Proceedings of the National Academy of Sciences*, **92** (April 1995): 3596-3600.
- Bendor, Jonathan, and Piotr Swistak, "The Controversy about the Evolution of Cooperation and the Evolutionary Roots of Social Institutions," in Gasparski, Wojciech et al (eds), *Social Agency*, New Brunswick, N.J.: Transaction Publishers, 1996.
- Bendor, Jonathan, and Piotr Swistak, "Evolutionary Equilibria: Characterization Theorems and Their Implications," *Theory and Decision*, **3** (forthcoming).
- Bendor, Jonathan, and Piotr Swistak, "The Evolutionary Stability of Cooperations," *American Political Science Review*, **91**/2 (June 1997): 290-307.
- Bendor, Jonathan, Roderick Kramer and Piotr Swistak, "Cooperation Under Uncertainty: What is New, What is True and What is Important?" *American Sociological Review*, **61** (April 1996): 333-338.
- Bicchieri, Cristina, "Self-refuting Theories of Strategic Interaction," *Erkenntinis*, **30** (1989): 69-85.
- Binmore, Kenneth, *Fun and Games*, Lexington, MA: D.C. Heath and Company (1992).
- Binmore, Kenneth, "Rationality and Backward Induction," *Journal of Economic Methodology*, **4** (1997): 23-41.
- Boyd, Robert and Jeffrey Lorberbaum, "No Pure Strategy is Evolutionarily Stable in the repeated Prisoner's Dilemma Game," *Nature*, **327** (7 May 1987): 58-59.
- Danielson, Peter, *Artificial Morality: Virtual Robots for Virtual Games*, London: Routledge (1992).
- Donninger, Christian "Is It Always Efficient to be Nice?" in Dickman and Mitter (eds) *Paradoxical Effects of Social Behavior* Heidelberg: Physica Verlag (1986): 123-134.
- Farrell, Joseph, and Roger Ware, "Evolutionary Stability in the Repeated Prisoner's Dilemma," *Theoretical Population Biology*, **36** (1989): 161-167.
- Gauthier, David, *Morals by Agreement*, Oxford: Clarendon Press (1986).
- Grim, Patrick, Gary Mar and Paul St. Denis, *The Philosophical Computer*, Cambrige, Mass: MIT Press (1998).
- Hardin, Garret, "The Tragedy of the Commons," *Science*, **162** (13 December 1968): 1243-1248.
- Howard, Nigel, *Paradoxes of Rationality*, Cambridge, MA: MIT Press (1971).
- Kavka, Gregory, "Hobbes War of All Against All," *Ethics*, **93** (1983): 291-310.
- Kavka, Gregory, *Hobbesean Moral and Political Theory*, Princeton: Princeton University Press (1986).
- Kollock, Peter, "An Eye For an Eye Leaves Everybody Blind: Cooperation and Accounting Systems," *American Sociological Review*, **58** (1993): 768-786.
- Kraines, David and Vivian Kraines, "Pavlov and the Prisoner's Dilemma," *Theory and Decision*, **26**

(1989): 47-79.

- Kraines, David and Vivian Kraines, "Learning to Cooperate with Pavlov: an Adaptive Strategy for the Iterated Prisoner's Dilemma with Noise," *Theory and Decision*, **35** (1993): 107-150.
- Kreps, David, Paul Milgrom, John Roberts and Robert Wilson, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory*, **27** (1982): 245-252.
- Kuhn Steven, and Serge Moresi, "Pure and Utilitarian Prisoner's Dilemmas" *Economics and Philosophy*, **11** (1995): 123-133.
- Maynard Smith, John, "The Evolution of Behavior," *Scientific American*, **239** (1978): 176-192.
- Molander, Per, "The Optimal Level of Generosity in a Selfish, Uncertain Environment," *Journal of Conflict Resolution*, **29**, (December 1985): 611-619.
- Molander, Per, "The Prevalence of Free Riding," *Journal of Conflict Resolution*, **36**, (December 1992): 756-771.
- Mukherji, Arijit, Vijay Rajan and James Slagle, "Robustness of Cooperation," *Nature*, **379** (11 January 1996): 125-126.
- Nowak, Martin, and Robert May, "Evolutionary Games and Spatial Chaos,"*Nature*, **359** (29 October 1992): 826-829.
- Nowak, Martin and Karl Sigmund, "Tit for Tat in Heterogeneous Populations," *Nature*, **355** (16 January 1992): 250-253.
- Nowak, Martin and Karl Sigmund, "A Strategy of Win-stay, Lose-shift that Outperforms Tit-for-tat in the Prisoner's Dilemma Game," *Nature*, **364** (1 July 1993): 56-58.
- Nowak, Martin, Robert May, and Karl Sigmund, "The Arithmetics of Mutual Help," *Scientific American* (June 1995): 76-81.
- Pettit, Phillip, "Free Riding and Foul Dealing," *Journal of Philosophy*, **83** 1986: 361-379.
- Pettit, Phillip and Robert Sugden, "The Backward Induction Paradox," *Journal of Philosophy*, **86** 1989: 169-182.
- Poundstone, William, *Prisoner's Dilemma* New York: Doubleday (1992).
- Rabinowicz, Wlodek, "Grappling with the Centipede: Defense of Backward Induction for BI-Terminating Games," *Economics and Philosophy*,, **14** (1998): 95-126.
- Schelling, Thomas, *Micromotives and Macrobehavior* New York: Norton (1978).
- Selten, Reinhard, "Reexamination of the Perfectness Concept of Equilibrium in Extensive Games," *International Journal of Game Theory*,, **4** (1975): 25-55.
- Selten, Reinhard, "Evolutionary Stability in Extensive Two-person Games," *Mathematical Social Sciences*, **5** (1983): 269-363.
- Selten, Reinhard, "The Chain-Store Paradox," *Theory and Decision*, **9** (1978): 127-159.
- Sigmund, Karl, *Games of Life: Explorations in Ecology Evolution and Behavior*, Oxford: Oxford University Press (1993).
- Skyrms, Brian, *The Dynamics of Rational Deliberation*, Cambridge, MA Harvard University Press (1990).
- Skyrms, Brian, "The Shadow of the Future," in Coleman and Morris (eds.)*Rational Commitment and Social Justice: Essays for Gregory Kavka,"*, New York, Cambridge University Press (1998).
- Sobel, J.H., "Reexamination of the Perfectness Concept of Equilibrium in Extensive Games," *International Journal of Game Theory*,, **4** (1975): 25-55.
- Sobel, J.H., "Utility Maximization in Iterated Prisoner's Dilemmas," *Dialogue*,, **15** (1976): 38-53.
- Sugden *The Economics of Rights, Cooperation and Welfare*, New York, Basil Blackwell (1986).

- Trivers, Robert, "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology*, **46** (1971): 35-57.
- Vanderschraaf, Peter, "The Informal Game Theory in Hume's Account of Convention," *Economics and Philosophy*, **14**(1998):215-247.
- Williamson, Timothy, "Inexact Knowledge," *Mind*, **101** (1992): 217-242.

Other Internet Resources

- "Comprehsive Repository" for information about the iterated PD
  (compiled by group at Laboratoire d'Informatique Fondamentale de Lille).
- Axelrod and D'Ambrosio's annotated bibliography (1988-1994).
- Spatial IPDs by Norman Siebrasse
- Spatial IPDs by Peter Gibbins.
- Site designed to accompany Axelrod's *Complexity of Cooperation.*
- Miscellaneous PD Resources (compiled by Constitution Society).
- Interactive Prisoner's Dilemma (at the Serendip pages at Bryn Mawr).

Related Entries

game theory | rationality

Steven T. Kuhn
kuhns@georgetown.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 4, 1997
Content last modified: October 12, 2001

Stanford Encyclopedia of Philosophy
Supplement to Prisoner's Dilemma

Stability Concepts in Evolutionary Games

Logical relations among concepts of stability used in discussions of the EPD and other evolutionary games are established in a series of papers by Bendor and Swistak. Some conditions on game payoffs and some conditions on the course of evolution are described below. Logical relations among these conditions are represented in the diagram that follows.

Conditions on payoffs

($V(i,j)$ is the total payoff that i gets playing against j.)

CS	Axelrod ("Collective Stability")	$\forall j[V(i,i) \geq V(j,i)]$
MS	Maynard Smith	$\forall j[V(i,i) > V(j,i)$ or $(V(i,i)=V(j,i)$ & $V(i,j) > V(j,j))]$
BL	Boyd and Lorberbaum	$\forall j[V(i,i) > V(j,i)$ or $(V(i,i)=V(j,i)$ & $\forall k(V(i,k) \geq V(j,k)))]$
BS	Bendor and Swistak	$\forall j[V(i,i) > V(j,i)$ or $(V(i,i)=V(j,i)$ & $V(i,j) \geq V(j,j))]$

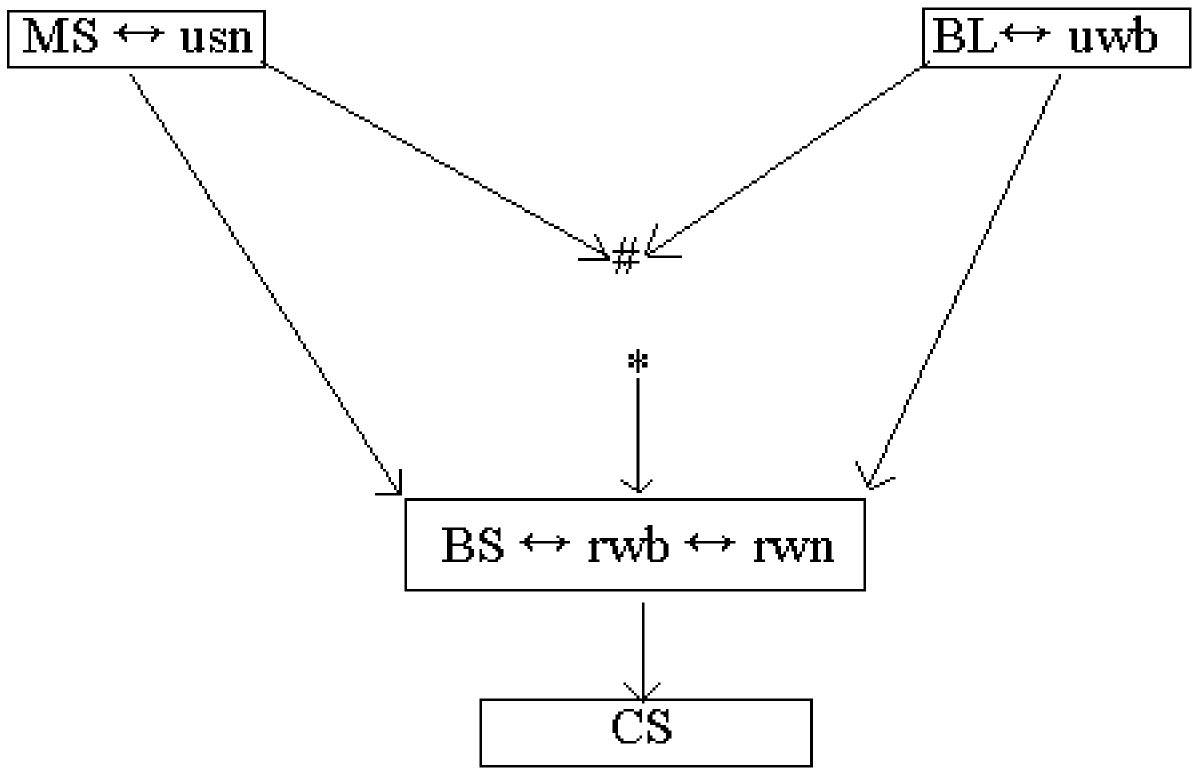
Conditions on the course of evolution

u and r ("universal" and "restricted") indicates that the condition obtains under any rule of evolution or merely under the replication dynamics. s and w ("strong" and "weak") indicate that the strategy eradicates invasions or merely survives them. n and b ("narrow" and "broad") indicate that the invaders are homogeneous (i.e., they all employ the same strategy), or heterogeneous. For example i has uwb (universal weak broad) stability if, under any rule of evolution, it survives sufficiently small heterogeneous invasions.

Relations among stability conditions

Logical implications are indicated by chains of arrows (and by relative vertical position). Conditions stronger than # cannot be satisfied by any EPD and conditions weaker than * are satisfied by EPDs with

all levels of cooperation from 0% to 100%.



Steven T. Kuhn
kuhns@georgetown.edu

Return to Prisoner's Dilemma Entry

First published: October 12, 2001
Content last modified: October 12, 2001

Stanford Encyclopedia of Philosophy
Notes to Evolutionary Game Theory

Notes

1. In a mixed strategy, a player assigns a probability to each pure strategy, and chooses which strategy to play using a randomization device. For the Hawk-Dove game, one mixed strategy would assign equal probabilities to playing Hawk or Dove, and decide which to play in a given case by flipping a fair coin.

2. In Nowak and May's model, each individual on the lattice plays the prisoner's dilemma with their eight nearest neighbors. At the end of each game round, an individual compares her score with that of her neighbors. If one of her neighbors earned a higher score, that player will adopt the strategy used by her most successful neighbor (presumably using some kind of randomization process to break ties). If no neighbor earned a higher score, that player will continue using the same strategy for the next round of play. All individuals switch strategies at the same time, and all have the same payoff structure.

3. Of course, Nowak and May were speaking somewhat loosely when they referred to this behavior as "chaotic." Since there are only finitely many states of the population, it must be the case that this dynamical system will eventually settle into a cycle (although it may not repeat itself for a very long time). However, the point is clear enough: in this particular case we are incapable of predicting the state of the model after *N* generations (for a large *N*) without running the model for that length of time. In both of the previous cases, it is easy to predict the future state of the population given its current state.

4. Both agents would clearly do better if both agents elected to cooperate, but the point where both individuals choose cooperate is not a Nash equilibrium: if both agents initially cooperate, the first individual can increase her payoff by choosing to defect (and vice versa).

5. This is a problem only when we consider games other than two-person zero sum games, since all Nash equilibria in these sort of games have equivalent payoffs and are interchangable.

6. One "beauty pageant" game is as follows: a group of subjects is told to choose any number in the interval [0,100]. The person whose choice is closest to half of the mean of the group wins the game. Assuming that the rules of the game and the rationality of the players are common knowledge, the predicted outcome of the game is that all members of the group choose 0. Since 50 is the largest possible winning choice (if each subject chose 100, half of the group mean would be 50), no rational individual would choose a number more than 50. However, since no rational individual would choose a number greater than 50, and all subjects know this, no subject will choose a number greater than 25. In the limit, all players will converge upon 0. Tests with real human subjects (other than game theorists or people who have been exposed to this game before) demonstrate that the modal offer is significantly greater than

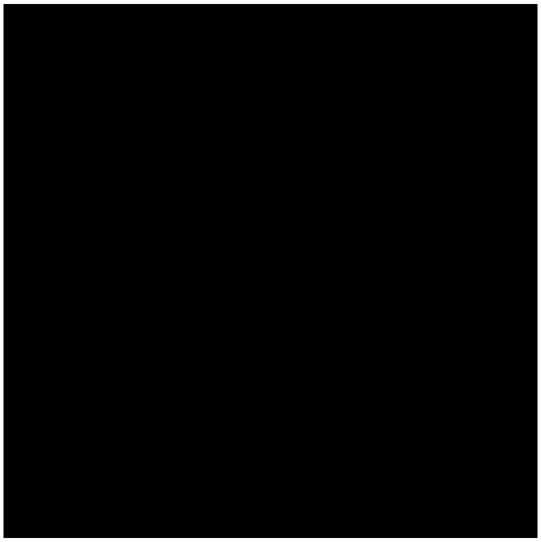
zero. If the game is repeated, though, subjects do approach the game theoretic prediction.

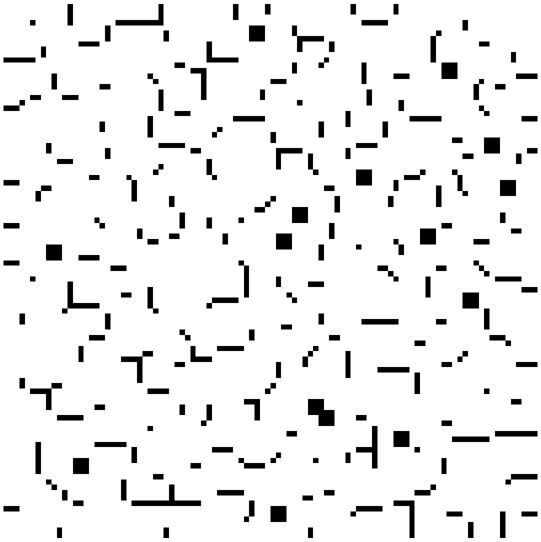
7. In the Hawk-Dove game discussed in section 2, the quantities in the payoff matrix represent the change in Darwinian fitness of the two individuals.

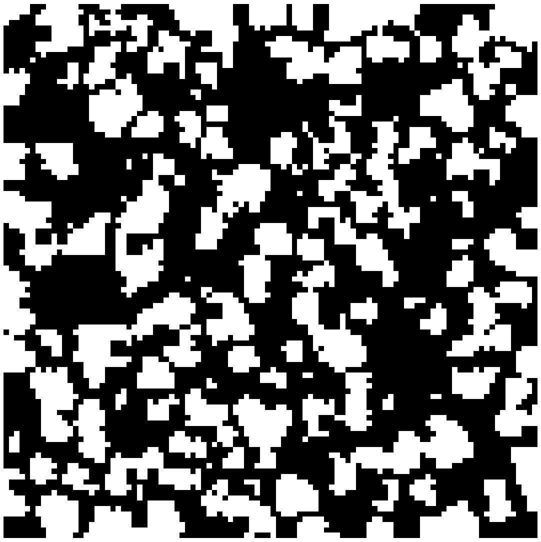
8. A more appropriate name would be "Hume's fallacy." Moore's "naturalistic fallacy" discussed in *Principia Ethica* differs significantly from the mistaken inference of an ought-statement from an is-statement. Using the same label for such different fallacies invites confusion.

J. McKenzie Alexander
jalex@lse.ac.uk

First published: January 14, 2002
Content last modified: January 14, 2002







Teleological Notions in Biology

Teleological terms such as "function" and "design" appear frequently in the biological sciences. Examples of teleological claims include:

- A (biological) function of stotting by antelopes is to communicate to predators that they have been detected.
- Eagles' wings are (naturally) designed for soaring.

Teleological notions were commonly associated with the pre-Darwinian view that the biological realm provides evidence of conscious design by a supernatural creator. Even after creationist viewpoints were rejected by most biologists there remained various grounds for concern about the role of teleology in biology, including whether such terms are:

1. vitalistic (positing some special "life-force");
2. requiring backwards causation (because future outcomes explain present traits);
3. incompatible with mechanistic explanation (because of 1 and 2);
4. mentalistic (attributing the action of mind where there is none);
5. empirically untestable (for all the above reasons).

Opinions divide over whether Darwin's theory of evolution provides a means of eliminating teleology from biology, or whether it provides a naturalistic account of the role of teleological notions in the science. Many contemporary biologists and philosophers of biology believe that teleological notions are a distinctive and ineliminable feature of biological explanations but that it is possible to provide a naturalistic account of their role that avoids the concerns above. Terminological issues sometimes serve to obscure some widely-accepted distinctions.

Teleomentalism

Teleomentalists regard the teleology of psychological intentions, goals, and purposes as the primary model for understanding teleology in biology. Aside from creationism, the most common form of teleomentalist view is that teleological claims in biology are mere metaphor---describing and explaining biological phenomena on the basis of more or less loose comparisons to psychological teleology. Those who hold teleology in biology to be metaphorical in nature typically regard it as eliminable; i.e., they believe that the science of biology would not be essentially altered if all references to teleology were eschewed.

Teleonaturalism

Those who reject teleomentalism typically seek naturalistic truth conditions for teleological claims in biology that do not refer to the intentions, goals, or purposes of psychological agents. Some *teleonaturalists* seek to reduce teleological language to forms of description and explanation that are found in other parts of science. One class of such views defines teleological notions cybernetically and maintains that teleology in biology is appropriate insofar as biological systems are cybernetic systems. Another, more widely-accepted approach treats functional claims in biology as part of the analysis of the capacities of a complex system into various component capacities.

Other forms of teleonaturalism regard the teleological aspects of biology as unique and ineliminable. One class of such views maintains that teleological claims in biology depend on natural values that apply to biological entities (such as what is good for an organism or species). A different approach, that avoids normative notions, is to define biological teleology explicitly in terms of natural selection and the theory of evolution.

Several theorists have argued for the pluralistic idea that biology may incorporate two notions of function, one to explain the presence of traits and the other to explain how those traits contribute to the complex capacities of organisms. Others have argued that these two apparently distinct notions of function can be unified by regarding the target of explanation as the biological fitness of a whole organism. Nonetheless, the mainstream view among philosophers of biology is that natural selection accounts best explain the majority of uses of teleological notions in biology.

Natural Selection Analyses of *Function*

Accounts of biological function which refer to natural selection typically have the form that *a trait's function or functions causally explain the existence or maintenance of that trait in a given population via*

the mechanism of natural selection. Three components of this view can be usefully separated:

1. Functional claims in biology are intended to explain the existence or maintenance of a trait in a given population;
2. Biological functions are causally relevant to the existence or maintenance of traits via the mechanism of natural selection;
3. Functional claims in biology are fully grounded in natural selection and are not derivative of psychological uses of notions such as design, intention, and purpose.

Variations on this account mostly center on the first two points.

1. Some theorists maintain a distinction between the initial spread of a new phenotypic trait in a population from the maintenance of traits in populations.
2. Some theorists adopt an *etiological* or backward-looking approach that analyzes the function of a trait only in terms of those effects of the trait which have in the past contributed to the selection of organisms with that trait. Others adopt a *dispositional* or forward-looking approach that analyzes function in terms of those effects it is disposed to produce that tend to contribute to the present or future maintenance of the trait in a population of organisms.

Function and Design

In the debate about biological teleology, relatively little attention has been paid to the notion of natural design. It is common for authors to slide between claims about function and design as if they accept this principle:

> A trait *T*is naturally designed for *X*if and only if *X*is a biological function of *T*.

Collapsing the notions of design and function in this manner has the advantage that if the notion of biological function is successfully naturalized then so is the notion of natural design.

The biological notion of design seems, however, to imply more than mere usefulness. Female turtles use their flippers to dig nests in sand, and doing so surely accounts for the maintenance of the trait in the population. So, on an etiological account, digging in sand is a function of the flippers. Yet it seems wrong to say that they are designed for that purpose. This suggests that function and design should be analyzed separately. One way to do this is as follows:

> *Trait*T *is naturally designed to do*X means that
> 1. *X* is a biological function of *T* and
> 2. *T* is the result of a process of change of (anatomical or behavioral) structure due to natural selection that has resulted in *T* being more optimal (or better adapted) for *X* than ancestral versions of *T*.

With respect to this analysis, to say that an eagle's wings are designed for soaring is to claim, first, that the ability to soar (as opposed to other kinds of flying) explains why some ancestral eagles had higher reproductive fitness than others and, second, that eagles' wings are better adapted for soaring than were ancestral versions of the wings. This second part is an historical claim that might be checked against the fossil record.

Adaptation, Exaptation and Co-opted Use

The notion of adaptation is controversial among biologists because it suggests the Panglossian belief that this is the best of all possible worlds. However comparative judgments about traits of organisms, e.g., that the traits of present organisms are better at producing some effect than the corresponding traits of ancestral organisms, do not require the Panglossian assumption. This is because the claim that A is more optimal or better adapted than B with respect to some function does not entail that A is optimal or even good with respect to that function.

Gould & Vrba (1982) would deny that sand-digging is a function of turtle flippers and prefer instead to label it an "exaptation". They recommend the use of "function" only when natural selection has "shaped" a trait for some use -- i.e. the trait has undergone some modification in form that makes it more suited to the use. This recommendation, however, seeks to change ordinary biological usage rather than to elucidate it. Because it conflates the notions of design and function, it becomes necessary to mark the distinction between cases of selection with modification (function/design) and cases where a trait of an organism is coopted for a use for which it is not modified (exaptation). Even if the flippers of turtles are not specially modified for burying eggs in sand, the fact that they were so used helps to explain why turtles with flippers were selected over those without. Whether one prefers to call this a function or an exaptation is a terminological issue perhaps to be settled by one's taste for neologisms.

Bibliography

- Allen, C. and Bekoff, M. "Function, natural design, and animal behavior: philosophical and ethological considerations," in N.S. Thompson (ed.) *Perspectives in Ethology, Volume 11: Behavioral Design.* (1995) NY: Plenum Press, pp.1-47.
- Allen, C. and Bekoff, M. "Biological function, adaptation, and natural design," *Philosophy of Science*: **62** (1995): 609-622.
- Gould, S.J. and Vrba, E.S. "Exaptation - a missing term in the science of form," *Paleobiology* **8** (1982): 4-15.
- Mayr, E. "The multiple meanings of teleological," in *Towards a New Philosophy of Biology*. Harvard University Press (1988), Cambridge, MA: 38-66.

Other Internet Resources

- Please contact the author with suggestions.

Related Entries

biology, philosophy of | natural selection

Colin Allen
colin.allen@tamu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 20, 1996
Content last modified: June 17, 1999

Environmental Ethics

Environmental ethics is the discipline that studies the moral relationship of human beings to, and also the value and moral status of, the environment and its nonhuman contents. This entry covers: (1) the challenge of environmental ethics to the anthropocentrism (i.e., human-centeredness) embedded in traditional western ethical thinking; (2) the early development of the discipline in the 1960s and 1970s; (3) the connection of deep ecology, feminist environmental ethics, and social ecology to politics; (4) the attempt to apply traditional ethical theories, including consequentialism, deontology, and virtue ethics, to support contemporary environmental concerns; and (5) the focus of environmental literature on wilderness, and possible future developments of the discipline.

1. Introduction: The Challenge of Environmental Ethics

Suppose that putting out natural fires, culling feral animals or destroying some individual members of overpopulated indigenous species is necessary for the protection of the integrity of a certain ecosystem. Will these actions be morally permissible or even required? Is it morally acceptable for farmers in non-industrial countries to practise slash and burn techniques to clear areas for agriculture? Consider a mining company which has performed open pit mining in some previously unspoiled area. Does the company have a moral obligation to restore the landform and surface ecology? And what is the value of a humanly restored environment compared with the originally natural environment? It is often said to be morally wrong for human beings to pollute and destroy parts of the natural environment and to consume a huge proportion of the planet's natural resources. If that is wrong, is it simply because a sustainable

environment is essential to (present and future) human well-being? Or is such behaviour also wrong because the natural environment and/or its various contents have certain values in their own right so that these values ought to be respected and protected in any case?

These are among the questions investigated by environmental ethics. Some of them are specific questions faced by individuals in particular circumstances, while others are more global questions faced by groups and communities. Yet others are more abstract questions concerning the value and moral standing of the natural environment and its nonhuman components.

In the literature on environmental ethics the distinction between *instrumental value* and *intrinsic value* (i.e., non-instrumental value) has been of considerable importance. The former is the value of things as *means* to further some other ends, whereas the latter is the value of things as *ends in themselves* regardless of whether they are also useful as means to other ends. For instance, certain fruits have instrumental value for bats who feed on them, since feeding on the fruits is a means to survival for the bats. However, it is not widely agreed that fruits have value as ends in themselves. We can likewise think of a person who teaches others as having instrumental value for those who want to acquire knowledge. Yet, in addition to any such value, it is normally said that a person, as a person, has intrinsic value, i.e., value in his or her own right independently of his or her prospects for serving the ends of others. For another example, a certain wild plant may have instrumental value because it provides the ingredients for some medicine or as an aesthetic object for human observers. But if the plant also has some value in itself independently of its prospects for furthering some other ends such as human health, or the pleasure from aesthetic experience, then the plant also has intrinsic value. Because the intrinsically valuable is that which is good as an end in itself, it is commonly agreed that something's possession of intrinsic value generates a *prima facie* direct moral duty on the part of moral agents to protect it or at least refrain from damaging it.

Many traditional western ethical perspectives, however, are *anthropocentric* or human-centered in that either they assign intrinsic value to human beings alone (i.e., what we might call anthropocentric in an *absolute* sense) or they assign a significantly greater amount of intrinsic value to human beings than to any nonhuman things such that the protection or promotion of human interests or well-being at the expense of nonhuman things turns out to be nearly always justified (i.e., what we might call anthropocentric in a *relative* sense). Aristotle (*Politics*, Bk. 1, Ch. 8) maintains that ‘nature has made all things specifically for the sake of man' and that the value of nonhuman things in nature is merely instrumental. The Bible (*Genesis* 1:27-8) says: “God created man in his own image, in the image of God created he him; male and female created he them. And God blessed them, and God said unto them, Be fruitful, and multiply, and replenish the earth, and subdue it: and have dominion over fish of the sea, and over fowl of the air, and over every living thing that moveth upon the earth.” Thomas Aquinas (*Summa Contra Gentiles*, Bk. 3, Pt 2, Ch 112) argues that because nonhuman animals are ‘ordered to man's use’, he can kill them or use them in any way he wishes without any injustice. Generally, anthropocentric positions find it problematic to articulate what is wrong with the cruel treatment of nonhuman animals, except to the extent that such treatment may lead to bad consequences for human beings. Immanuel Kant (‘Duties to Animals and Spirits’, in *Lectures on Ethics*), for instance, suggested that cruelty towards a dog might encourage a person to develop a character which would be desensitized to cruelty towards

humans. From this standpoint, cruelty towards nonhuman animals would be instrumentally, rather than intrinsically, wrong. Likewise, anthropocentrism often recognizes some non-intrinsic wrongness of anthropogenic (i.e. human-caused) environmental devastation. Such destruction might damage the well-being of human beings (now and in the future), since our well-being is essentially dependent on a sustainable environment (see Passmore 1974, Bookchin 1990, Norton, Hutchins, Stevens, and Maple (eds.) 1995).

When environmental ethics emerged as a new sub-discipline of philosophy in the early 1970s, it did so by posing a challenge to traditional anthropocentrism. In the first place, it questioned the assumed moral superiority of human beings to members of other species on earth. In the second place, it investigated the possibility of rational arguments for assigning intrinsic value to the natural environment and its nonhuman contents.

It should be noted, however, that some theorists working in the field see no need to develop new, non-anthropocentric theories. Instead, they advocate what may be called *enlightened anthropocentrism* (or, perhaps more appropriately called, *prudential anthropocentrism*). Briefly, this is the view that all the moral duties we have towards the environment are derived from our direct duties to its human inhabitants. The practical purpose of environmental ethics, they maintain, is to provide moral grounds for social policies aimed at protecting the earth's environment and remedying environmental degradation. Enlightened anthropocentrism, they argue, is sufficient for that practical purpose, and perhaps even more effective in delivering pragmatic outcomes, in terms of policy-making, than non-anthropocentric theories given the theoretical burden on the latter to provide sound arguments for its more radical view that the nonhuman environment has intrinsic value (cf. Norton 1991, de Shalit 1994, Light and Katz 1996). Furthermore, some prudential anthropocentrists may hold what might be called *cynical anthropocentrism*, which says that we have a higher-level anthropocentric reason to be non-anthropocentric in our day-to-day thinking. The reason for this is that a day-to-day non-anthropocentrist tends to act more benignly towards the nonhuman environment on which human well-being depends. In order to be an effective cynical anthropocentrist, unfortunately, one may need to hide one's cynical anthropocentrism from others and even from oneself.

2. The Early Development of Environmental Ethics

Although nature was the focus of much nineteenth and twentieth century philosophy, contemporary environmental ethics only emerged as an academic discipline in the 1970s. The questioning and rethinking of the relationship of human beings with the natural environment over the last thirty years reflected an already widespread perception in the 1960s that the late twentieth century faced a 'population time bomb' and a serious environmental crisis.

Among the accessible work that drew attention to a sense of crisis was Rachel Carson's *Silent Spring* (1963), which consisted of a number of essays earlier published in the *New Yorker* magazine detailing how pesticides such as DDT, aldrin and deildrin concentrated through the food web. Commercial farming practices aimed at maximizing crop yields and profits, Carson speculated, were capable of impacting

simultaneously on environmental and public health. On the other hand, historian Lynn White jr., in a much-cited essay published in 1967 (White 1967) on the historical roots of the environmental crisis, argued that the main strands of Christian thinking had encouraged the overexploitation of nature by maintaining the superiority of humans over all other forms of life, and by depicting all of nature as created for the use of humans. Nevertheless, White argued that some minority traditions within Christianity (e.g., the views of St. Francis) might provide an antidote to the 'arrogance' of a mainstream tradition steeped in anthropocentrism. Two years later, the Stanford ecologist, Paul Ehrlich, published *The Population Bomb* (1968), warning that the growth of human population threatened the viability of planetary life-support systems. The sense of environmental crisis stimulated by those and other popular works was intensified by NASA's production and wide dissemination of a particularly potent image of earth from space taken at Christmas 1968 and featured in the *Scientific American* in September 1970. Here, plain to see, was a living, shining planet voyaging through space and shared by all of humanity, a precious vessel vulnerable to pollution and to the overuse of its limited capacities. In 1972 a team of researchers at MIT led by Dennis Meadows produced the *Limits to Growth* study, a work that summed up in many ways the emerging concerns of the previous decade and the sense of vulnerability triggered by the view of the earth from space. In §10 of the commentary to the study, the researchers wrote:

> We affirm finally that any deliberate attempt to reach a rational and enduring state of equilibrium by planned measures, rather than by chance or catastrophe, must ultimately be founded on a basic change of values and goals at individual, national and world levels.

The call for a 'basic change of values' in connection to the environment (a call that could be interpreted in terms of either instrumental or intrinsic values) reflected a need for the development of environmental ethics as a new sub-discipline of philosophy.

The new field emerged almost simultaneously in three countries -- the United States, Australia, and Norway. In the first two of these countries, direction and inspiration largely came from the earlier twentieth century American literature of the environment. For instance, the Scottish emigrant John Muir (founder of the Sierra Club and 'father of American conservation') and subsequently the forester Aldo Leopold had advocated an appreciation and conservation of things 'natural, wild and free'. Their concerns were motivated by a combination of ethical and aesthetic responses to nature as well as a rejection of crudely economic approaches to the value of natural objects (a historical survey of the confrontation between Muir's reverentialism and the utilitarian conservationism of Gifford Pinchot (one of the major influences on the development of the US Forest Service) is provided in Norton 1991; also see Cohen 1984 and Nash (ed) 1990). Leopold's *A Sand County Almanac* (1949), in particular, advocated the adoption of a 'land ethic':

> That land is a community is the basic concept of ecology, but that land is to be loved and respected is an extension of ethics. (vii-ix)
> A thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends otherwise. (224-5)

However, Leopold himself provided no systematic ethical theory or framework to support these ethical

ideas concerning the environment. His views therefore presented a challenge and opportunity for moral theorists: could some ethical theory be devised to justify the injunction to preserve the integrity, stability and beauty of the biosphere?

The land ethic sketched by Leopold, attempting to extend our moral concern to cover the natural environment and its nonhuman contents, was drawn on explicitly by the Australian philosopher Richard Routley (later Sylvan). According to Routley (1973 (cf. Routley and Routley 1980)), the anthropocentrism imbedded in what he called the ‘dominant western view’, or ‘the western superethic’, is in effect ‘human chauvinism’. This view, he argued, is just another form of class chauvinism, which is simply based on blind class ‘loyalty’ or prejudice, and unjustifiably discriminates against those outside the privileged class. Furthermore, in his ‘last man’ (and ‘last people’) arguments, Routley asked us to imagine the hypothetical situation in which the last person, surviving a world catastrophe, acted to ensure the elimination of all other living things and the destruction of all the landscapes after his demise. From the human-chauvinistic (or absolutely anthropocentric) perspective, the last person would do nothing morally wrong, since his or her destructive act in question would not cause any damage to the interest and well-being of humans, who would have by then disappeared. Nevertheless, Routley believed that that the imagined last act would be morally wrong. An explanation for this judgment, he suggested, is that those nonhuman objects in the environment, whose destruction is ensured by the last person, have intrinsic value, a kind of value independent of their usefulness for humans. From his critique, Routley concluded that the main approaches in traditional western moral thinking were unable to allow the recognition that natural things have intrinsic value, and that the tradition required overhaul of a significant kind.

Leopold's idea that the ‘land’ as a whole is an object of our moral concern also stimulated writers to argue for certain moral obligations toward ecological wholes, such as species, communities, and ecosystems, not just their individual constituents. American environmental philosopher Holmes Rolston III (1975), for instance, argued that species protection was a moral duty. It would be wrong, Rolston maintained, to eliminate a rare butterfly species simply to increase the monetary value of specimens already held by collectors. Like Routley's ‘last man’ arguments, Rolston's example is meant to draw attention to a kind of action that seems morally dubious and yet is not clearly ruled out or condemned by traditional anthropocentric ethical views. Species, Rolston went on to argue, are intrinsically valuable and are usually more valuable than individual specimens, since the loss of a species is a loss of genetic possibilities and the deliberate destruction of a species would show disrespect for the very biological processes which make possible the emergence of individual living things (also see Rolston 1989, Ch 10).

Meanwhile, the work of Christopher Stone (a professor of law at the University of Southern California) had become widely discussed. Stone (1972) proposed that trees and other natural objects should have at least the same standing in law as corporations. This suggestion was inspired by a particular case in which the Sierra Club had mounted a challenge against the permit granted by the U.S.Forest Service to Walt Disney Enterprises for surveys preparatory to the development of the Mineral King Valley, which was at the time a relatively remote game refuge, but not designated as a national park or protected wilderness area. The Disney proposal was to develop a major resort complex serving 14000 visitors daily to be accessed by a purpose-built highway through Sequoia National Park. The Sierra Club,as a body with a

general concern for conservation, challenged the development on the grounds that the valley should be kept in its original state for its own sake.

Stone reasoned that if trees, forests and mountains could be given standing in law then they could be represented in their own right in the courts by groups such as the Sierra Club. Moreover, like any other *legal person*, these natural things could become beneficiaries of compensation if it could be shown that they had suffered compensatable injury through human activity. When the case went to the U.S. Supreme Court, it was determined by a narrow majority that the Sierra Club did not even meet the condition for bringing a case to court, for the Club was unable and unwilling to prove the likelihood of injury to the interest of the Club or its members. In a dissenting minority judgment, however, justices Douglas, Blackmun and Brennan mentioned Stone's argument: his proposal to give legal standing to natural things, they said, would allow conservation interests, community needs and business interests to be represented, debated and settled in court.

Reacting to Stone's proposal, Joel Feinberg (1974) raised a serious problem. Only items that have interests, Feinberg argued, can be regarded as having legal standing and, likewise, moral standing. For it is interests which are capable of being represented in legal proceedings and moral debates. This same point would also seem to apply to political debates. For instance, the movement for ‘animal liberation’, which also emerged strongly in the 1970s, can be thought of as a political movement aimed at representing the previously neglected interests of some animals (see Regan and Singer (eds.) 1976, and Clark 1977). Granted that some animals have interests that can be represented in this way, would it also make sense to speak of trees, forests, rivers, barnacles, or termites as having interests of a morally relevant kind? This issue was hotly contested in the years that followed. Meanwhile, John Passmore (1974) argued, like White, that the Judeo-Christian tradition of thought about nature, despite being predominantly ‘despotic’, contained resources for regarding humans as ‘stewards’ or ‘perfectors’ of God's creation. Skeptical of the prospects for any radically new ethic, Passmore cautioned that traditions of thought could not be abruptly overhauled. Any change in attitudes to our natural surroundings which stood the chance of widespread acceptance, he argued, would have to resonate and have some continuities with the very tradition which had legitimized our destructive practices. In sum, then, Leopold's land ethic, the historical analyses of White and Passmore, the pioneering work of Routley, Stone and Rolston, and the warnings of scientists, had by the late 1970s focused the attention of philosophers and political theorists firmly on the environment.

The confluence of ethical, political and legal debates about the environment, the emergence of philosophies to underpin animal rights activism and the puzzles over whether an environmental ethic would be something new rather than a modification or extension of existing ethical theories were reflected in wider social and political movements. The rise of environmental or ‘green’ parties in Europe in the 1980s was accompanied by almost immediate schisms between groups known as ‘realists’ versus ‘fundamentalists’ (see Dobson 1992). The ‘realists’ stood for reform environmentalism, working with business and government to soften the impact of pollution and resource depletion especially on fragile ecosystems or endangered species. The ‘fundies’ argued for radical change, the setting of stringent new priorities, and even the overthrow of capitalism and liberal individualism, which were taken as the major ideological causes of anthropogenic environmental devastation. Underlying these disagreements was the

distinction between 'shallow' and 'deep' environmental movements, a distinction introduced in the early 1970s by another major influence on contemporary environmental ethics, the Norwegian philosopher and climber Arne Næss. Since the work of Næss has been significant in environmental politics, the discussion of his position is given in a separate section below.

3. Environmental Ethics and Politics

3.1 Deep Ecology

'Deep ecology' was born in Scandinavia, the result of discussions between Næss and his colleagues Sigmund Kvaløy and Nils Faarlund (see Næss 1973 and 1989; also see Witoszek and Brennan (eds.) 1999 for a historical survey and commentary on the development of deep ecology). All three shared a passion for the great mountains. On a visit to the Himalayas, they became impressed with aspects of 'Sherpa culture' particularly when they found that their Sherpa guides regarded certain mountains as sacred and accordingly would not venture onto them. Næss decided to formulate a position which extended the reverence he and the Sherpas felt for mountains to other natural things in general.

The 'shallow ecology movement', as Næss (1973) calls it, is the 'fight against pollution and resource depletion', the central objective of which is 'the health and affluence of people in the developed countries.' The 'deep ecology movement', in contrast, endorses 'biospheric egalitarianism', the view that all living things are alike in having value in their own right, independent of their usefulness to human purposes. The deep ecologist respects this intrinsic value, taking care, for example, when walking on the mountainside not to cause unnecessary damage to the plants. Furthermore, deep ecology also endorses what Næss calls the 'relational, total-field image', understanding organisms (human or otherwise) as 'knots' in the biospherical net, the identities of which are defined in terms of their ecological relations to each other. Næss maintains that the 'deep' satisfaction that we receive from close partnership with other forms of life in nature contributes significantly to our life quality. As developed by Næss and others, the position also came to focus on the possibility of the 'identification' of the human ego with nature. The idea is briefly that by identifying with nature we can enlarge the boundaries of the self. Self-respect thus extends beyond the boundaries of my skin. My larger -- ecological -- Self (the capital 'S' emphasizes that I am an individual some of whose parts are found outside of my skin), according to Næss, deserves respect as well. And to respect and to care for myself is also to respect and to care for the natural environment with which I identify myself. 'Self-realization', in other words, is the reconnection of the shriveled human individual with the wider natural environment. One clear historical antecedent to this view is the romanticism of Jean-Jacques Rousseau as expressed in his last work, the *Reveries of the Solitary Walker*, though Næss himself cites Spinoza as the major historic inspiration of his deep ecology.

When Næss's view crossed the Atlantic, it was sometimes merged with ideas emerging from Leopold's land ethic (see Devall and Sessions 1985; also see Sessions (ed) 1995). But Næss -- wary of the apparent totalitarian political implications of Leopold's position that individual interests and well-being should be subordinated to the holistic good of the earth's biotic community (see section 4 below) -- has always taken care to distance himself from advocating any sort of 'land ethic'. Some later critics have argued

that Næss's deep ecology is no more than an extended social-democratic version of utilitarianism, which counts human interests in the same calculation alongside the ‘interests’ of all natural things (e.g., trees, wolves, bears, rivers, forests and mountains) in the natural environment (see Witoszek 1997). However, Næss failed to explain in any detail how to make sense of the idea that oysters or barnacles, termites or bacteria could have interests of any morally relevant sort at all. Without an account of this, Næss 's early ‘biospheric egalitarianism’ - that all living things whatsoever had a similar right to live and flourish - was an indeterminate principle. It also remains unclear how rivers, mountains and forests can be regarded as possessors of any kind of interests at all. This is an issue on which Næss has always remained elusive.

Biospheric egalitarianism was modified in the 1980s to the weaker claim that the flourishing of both human and non-human life have value in themselves. At the same time, Næss declared that his own favoured ecological philosophy -- ’Ecosophy T’, as he called it after his Tvergastein mountain cabin-- was only one of several possible foundations for an environmental ethic. Deep ecology ceased to be a specific doctrine, but instead became a ‘platform’, of eight points, on which Næss hoped all deep green thinkers could agree. The platform was conceived as establishing a middle ground, between underlying philosophical orientations, whether Christian, Buddhist, Taoist, process philosophy, or whatever, and the practical principles determining action in specific situations. Thus the deep ecological movement became explicitly pluralist (see Brennan 1999; c.f. Light 1996).

While Næss's Ecosophy T sees human Self-realization as a solution to the environmental crises resulting from human selfishness and exploitation of nature, some of the American and Australian followers of the deep ecology platform further argue that the expansion of the human self to include nonhuman nature is supported by the Copenhagen interpretation of quantum theory, which is said to have dissolved the boundaries between the observer and the observed (see Fox 1984, 1990, and Devall and Sessions 1985; cf. Callicott 1985). These developments of deep ecology were, however, criticized by some feminist theorists, who argued that the theory of the expanded self is in effect a disguised form of human -- indeed masculine -- egoism, unable to give nature its due respect as a genuine ‘other’ independent of human interest and well-being (see Plumwood 1993, Ch. 7, 1999, and Warren 1999). Meanwhile, some third-world critics have accused deep ecology of being elitist in its attempts to preserve wilderness experiences for only a select group of economically and socio-politically well-off people. The Indian writer Ramachandra Guha (1989, 1999) for instance, depicts the activities of many western-based conservation groups as a new form of cultural imperialism, aimed at securing converts to conservationism (cf. Bookchin 1987 and Brennan 1998a). ‘Green missionaries’, as Guha calls them, represent a movement aimed at further dispossessing the world's poor and indigenous people. “Putting deep ecology in its place,” he writes, “is to recognize that the trends it derides as "shallow" ecology might in fact be varieties of environmentalism that are more apposite, more representative and more popular in the countries of the South.” Although Næss himself repudiates suggestions that deep ecology is committed to any imperialism (see Witoszek and Brennan (eds.) 1999, Ch. 36-7 and 41), Guha's criticism raises important questions about the application of deep ecological principles in different social, economic and cultural contexts. In other critiques, deep ecology has been portrayed as having an inconsistent utopian vision (see Anker and Witoszek 1998).

3.2 Feminism and the Environment

Broadly speaking, a ‘feminist issue’ is any issue that contributes in some way to understanding the oppression of women. Feminist theories attempt to analyze women's oppression, its causes and consequences, and suggest strategies and directions for women's liberation. By the mid 1970s, feminist writers had raised the issue of whether patriarchal modes of thinking encouraged not only widespread inferiorizing and colonizing of women, but also of coloured people, animals and nature. Sheila Collins (1974), for instance, argued that male-dominated culture or patriarchy is supported by four interlocking pillars: sexism, racism, class exploitation, and ecological destruction.

Emphasizing the importance of feminism to the environmental movement and various other liberation movements, some writers, such as Ynestra King (1989a and 1989b), argue that the domination of women by men is the original form of domination in human society, from which all other hierarchies -- of rank, class, and political power -- flow. For instance, human domination of nature, it has been argued, is a manifestation and extension of the oppression of women, in that it is the result of associating nature with the female, which had been already inferiorized and oppressed by the male-dominating human culture. But within the plurality of feminist positions, other writers, such as Val Plumwood (1993), understand the oppression of women as only one of the many parallel forms of oppressions sharing and supported by a common structure, in which one party (the colonizer) uses a number of conceptual and rhetorical devices to privilege its interests over that of the other party (the colonized). It is argued that male-centered (androcentric) and human-centered (anthropocentric) thinking have some common characteristics, such as ‘dualism’ and the ‘logic of domination’, which are also manifested in the oppressions of many other social groups, and that in being facilitated by a common ideological structure, diverse forms of oppression often mutually-reinforce each other (Warren 1987, 1990, 1994, Cheney 1989, and Plumwood 1993). A central target of feminist analysis are those patterns of ‘dualism’ that lie deep in patriarchal thought. Examples are polar opposites, such as male/female, human/nonhuman, culture/nature, mind/body, reason/emotion, freedom/necessity. These dualisms are not just descriptive dichotomies, according to many feminists, but involve a prescriptive privileging of one side of the opposed items over the other, which is often rationalized by alleged ‘discovery’ of some qualities of the dominating groups that are meant to justify the domination that the privileged wields over the subjugated. For instance, the male may be said to excel in rationality over the emotional female; the active Cartesian mind, being free from physical constrains, may be seen as superior to the mechanical passive body; the civilized and progressive human culture may be deemed superior to the primitive nonhuman nature.

The insight of feminism, however, is not just that the dominating party often falsely sees the dominated party as lacking (or possessing) the allegedly superior (or inferior) qualities. More important, according to feminist analyses, the very premise of prescriptive dualism -- the valuing of attributes of one polarized side and the devaluing of those of the other, the idea that domination and oppression can be justified by appealing to attributes like masculinity, rationality, being civilized or developed, etc. -- is itself problematic.

Feminism represents a radical challenge for environmental thinking, politics, and traditional social

ethical perspectives. It promises to link environmental questions with wider social problems concerning various kinds of discrimination and exploitation, and fundamental investigations of human psychology. However, whether there are conceptual or merely contingent connections among the different forms of oppression and liberation remains a contested issue (see Green 1994). The term ‘ecofeminism’ (first coined by Françoise d'Eaubonne in 1974) is now generally applied to any view that combines environmental advocacy with feminist analysis. However, because of the varieties of, and disagreements among, feminist theories, the label may be too wide to be informative. Some feminist writers on environmental issues are wary of calling themselves ‘ecofeminists’.

3.3 Social Ecology

Apart from feminist-environmentalist theories and Næss's deep ecology, Murray Bookchin's ‘social ecology’ has also claimed to be radical, subversive, or countercultural (see Bookchin 1980, 1987, 1990).

One major influence on Bookchin's social ecology is the Frankfurt School of ‘critical theory’. Classical Marxists regarded Nature as a resource to be transformed by human labour and utilized. Members of the Frankfurt School such as Max Horkheimer and Theodore Adorno interpret Marx himself as representative of the problem of human alienation from nature. At the root of this alienation, they argue, is a narrow positivist conception of rationality -- which sees rationality as an instrument for pursuing power, technological control and progress, and takes observation, measurement and the application of purely quantitative methods to be capable of solving all problems. This conception, Horkheimer and Adorno (1969) argue, requires revision. Their project is to replace the narrow positivistic model of rationality with the so-called ‘Romantic’ values of the aesthetic, moral, sensual and expressive aspects of human nature, and bring these into harmony with our rational faculties. The oppression of what they call ‘outer nature’ (i.e., the natural environment) through science and technology, they argue, is bought at a very high price: the project of domination requires the suppression of our ‘inner nature’, the world of manifold needs and longings at the center of human life and its vulnerability (also see Eckersley 1992 and Vogel 1996, for a review of the Frankfurt School's thinking about nature).

Bookchin's version of critical theory takes the ‘outer’ physical world as constituting what he calls ‘first nature’, from which culture or ‘second nature’ has evolved. Environmentalism, on his view, is a social movement, and the problems it confronts are social problems. While Bookchin is prepared, like Horkheimer and Adorno, to regard (first) nature as an aesthetic and sensuous marvel, he regards our intervention in it as necessary. He suggests that we can choose to put ourselves at the service of natural evolution, to help maintain complexity and diversity, diminish suffering and reduce pollution. In this way, we can also to some extent overcome the kind of alienation that so worried the Frankfurt School. Bookchin's social ecology recommends that we use our gifts of sociability, communication and intelligence as if we were ‘nature rendered conscious’, instead of turning them against the very source and origin from which such gifts derive. Oppression of nature should be replaced by a richer form of life devoted to nature's preservation.

Deep ecology, feminism, and social ecology have had a considerable impact on the development of

political positions in regard to the environment. Feminist analyses have often been welcomed for the psychological insight they bring to several social, moral and political problems. There is, however, considerable unease about the implications of critical theory, social ecology and some varieties of deep ecology. Some recent writers have argued, for example, that critical theory is bound to be ethically anthropocentric, with nature as no more than a ‘social construction’ whose value ultimately depends on human determinations (see Vogel 1996). Others have argued that the demands of ‘deep’ green theorists and activists cannot be accommodated within contemporary theories of liberal politics and social justice (see Ferry 1998). A further suggestion is that there is a need to reassess traditional theories such as virtue ethics, which has its origins in ancient Greek philosophy (see the following section) within the context of a form of stewardship similar to that earlier endorsed by Passmore (see Barry 1999). If this last claim is correct, then the radical activist need not, after all, look for philosophical support in radical, or countercultural, theories of the sort deep ecology, feminism and social ecology claim to be.

4. Traditional Ethical Theories and Contemporary Environment Ethics

Although environmental ethicists often try to distance themselves from the anthropocentrism embedded in traditional ethical views (Passmore 1974, Norton 1991 are exceptions), they also quite often draw their theoretical resources from traditional ethical systems and theories. Consider the following two basic moral questions: (1) What kinds of thing are intrinsically valuable, good or bad? (2) What makes an action right or wrong?

Consequentialist ethical theories consider intrinsic ‘value’/‘disvalue’ or ‘goodness’/‘badness’ to be more fundamental moral notions than ‘rightness’/‘wrongness’, and maintain that whether an action is right/wrong is determined by whether its consequences are good/bad. From this perspective, answers to question (2) are informed by answers to question (1). For instance, utilitarianism, a paradigm case of consequentialism, regards pleasure (or, more broadly construed, the satisfaction of interest, desire, and/or preference) as the only intrinsic value in the world, whereas pain (or the frustration of desire, interest, and/or preference) the only intrinsic disvalue, and maintains that right actions are those that would produce the greatest balance of pleasure over pain.

As the utilitarian focus is the balance of pleasure and pain as such, the question of to whom a pleasure or pain belongs is irrelevant to the calculation and assessment of the rightness or wrongness of actions. Hence, the eighteenth century utilitarian Jeremy Bentham (1789), and now Peter Singer (1993), have argued that the interests of all the sentient beings (i.e., beings who are capable of experiencing pleasure or pain) -- including nonhuman ones -- affected by an action should be taken equally into consideration in assessing the action. Furthermore, rather like Routley (see section 2 above), Singer argues that the anthropocentric privileging of members of the species *Homo sapiens* is arbitrary, and that it is a kind of ‘speciesism’ as unjustifiable as sexism and racism. Singer regards the animal liberation movement as comparable to the liberation movements of women and people of colour. Unlike the environmental philosophers who attribute intrinsic value to the natural environment and its inhabitants, Singer and

utilitarians in general attribute intrinsic value to the experience of pleasure or interest satisfaction as such, not to the beings who have the experience. Similarly, for the utilitarian, non-sentient objects in the environment such as plant species, rivers, mountains, and landscapes, all of which are the objects of moral concern for environmentalists, are of no intrinsic but at most instrumental value to the satisfaction of sentient beings (see Singer 1993, Ch. 10). Furthermore, because right actions, for the utilitarian, are those that maximize the overall balance of interest satisfaction over frustration, practices such as whale-hunting and the killing of an elephant for ivory, which cause suffering to nonhuman animals, might turn out to be right after all: such practices might produce considerable amounts of interest-satisfaction for human beings, which, on the utilitarian calculation, outweigh the nonhuman interest-frustration involved. As the result of all the above considerations, it is unclear to what extent a utilitarian ethic can also be an environmental ethic. This point may not so readily apply to a wider consequentialist approach, which attributes intrinsic value not only to pleasure or satisfaction, but also to various objects and processes in the natural environment.

Deontological ethical theories, in contrast, maintain that whether an action is right or wrong is for the most part independent of whether its consequences are good or bad. From the deontologist perspective, there are several distinct moral rules or duties (e.g., ‘not to kill or otherwise harm the innocent’, ‘not to lie’, ‘to respect the rights of others’, ‘to keep promises’), the observance/violation of which is intrinsically right/wrong; i.e., right/wrong in itself regardless of consequences. When asked to justify an alleged moral rule, duty or its corresponding right, deontologists may appeal to the intrinsic value of those beings to whom it applies. For instance, ‘animal rights’ advocate Tom Regan (1983) argues that those animals with intrinsic value (or what he calls ‘inherent value’) have the moral right to respectful treatment, which then generates a general moral duty on our part not to treat them as mere means to other ends. We have, in particular, a *prima facie* moral duty not to harm them. Regan maintains that certain practices (such as sport or commercial hunting, and experimentation on animals) violate the moral right of intrinsically valuable animals to respectful treatment. Such practices, he argues, are intrinsically wrong regardless of whether or not some better consequences ever flow from them. Exactly which animals have intrinsic value and therefore the moral right to respectful treatment? Regan's answer is: those that meet the criterion of being the ‘subject-of-a-life’. To be such a subject is a sufficient (though not necessary) condition for having intrinsic value, and to be a subject-of-a-life involves, among other things, having sense-perceptions, beliefs, desires, motives, memory, a sense of the future, and a psychological identity over time.

Some authors have extended concern for individual well-being further, arguing for the intrinsic value of organisms achieving their own good, whether those organisms are capable of consciousness or not. Paul Taylor's version of this view (1981 and 1986), which we might call *biocentrism*, is a deontological example. He argues that each individual living thing in nature -- whether it is an animal, a plant, or a micro-organism -- is a ‘teleological-center-of-life’ having a good or well-being of its own which can be enhanced or damaged, and that all individuals who are teleological-centers-of life have equal intrinsic value (or what he calls ‘inherent worth’) which entitles them to moral respect. Furthermore, Taylor maintains that the intrinsic value of wild living things generates a *prima facie* moral duty on our part to preserve or promote their goods as ends in themselves, and that any practices which treat those beings as mere means and thus display a lack of respect for them are intrinsically wrong. Unlike Taylor's

egalitarian and deontological biocentrism, Robin Attfield (1987) argues for a hierarchical view that while all beings having a good of their own have intrinsic value, some of them (e.g., persons) have intrinsic value to a greater extent. Attfield also endorses a form of consequentialism which takes into consideration, and attempts to balance, the many and possibly conflicting goods of different living things (also see Varner 1998 for a more recent defense of what he calls *biocentric individualism* with affinities to both consequentialist and deontological approaches). However, some critics have pointed out that the notion of biological good or well-being is only descriptive not prescriptive (see Williams 1992 and O'Neill 1993, Ch. 2). For instance, the fact that HIV has a good of its own does not mean that we ought to assign any positive moral weight to the realization of that good.

Note that the ethics of animal liberation or animal rights and biocentrism are both *individualistic* in that their various moral concerns are directed towards individuals only -- not ecological wholes such as species, populations, biotic communities, and ecosystems. None of these is sentient, a subject-of-a-life, or a teleological-center-of-life, but the preservation of these collective entities is a major concern for many environmentalists. Moreover, the goals of animal liberationists, such as the reduction of animal suffering and death, may conflict with the goals of environmentalists. For example, the preservation of the integrity of an ecosystem may require the culling of feral animals or of some indigenous populations that threaten to destroy fragile habitats. So there are disputes about whether the ethics of animal liberation is a proper branch of environmental ethics (see Callicott 1980, 1988, Sagoff 1984, Jamieson 1998, Crisp 1998 and Varner 2000).

Criticizing the individualistic approach in general for failing to accommodate conservation concerns for ecological wholes, J. Baird Callicott (1980) has advocated a version of land-ethical *holism* which takes Leopold's statement “A thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends otherwise” to be the supreme deontological principle. In this theory, the earth's biotic community per se is the sole locus of intrinsic value, whereas the value of its individual members is merely instrumental and dependent on their contribution to the ‘integrity, stability, and beauty’ of the larger community. A straightforward implication of this version of the land ethic is that an individual member of the biotic community ought to be sacrificed whenever that is needed for the protection of the holistic good of the community. For instance, Callicott maintains that if culling a white-tailed deer is necessary for the protection of the holistic biotic good, then it is a land-ethical requirement to do so. But, to be consistent, the same point also applies to human individuals because they are also members of the biotic community. Not surprisingly, the misanthropy implied by Callicott's land-ethical holism has been widely criticized and regarded as a reductio of the position (see Aiken (1984), Kheel (1985), Ferré (1996), and Shrader-Frechette (1996)). Tom Regan (1983, p.362), for example, has condemned the holistic land ethic's disregard of the rights of the individual as ‘environmental fascism’. Under the pressure from the charge of ecofascism and misanthropy, Callicott (1989 Ch. 5, and 1999, Ch. 4) has later revised his position and now maintains that the biotic community (indeed, any community to which we belong) as well as its individual members (indeed, any individual who shares with us membership in some common community) all have intrinsic value. The controversy surrounding Callicott's original position, however, has inspired efforts in environment ethics to investigate possibilities of attributing intrinsic value to ecological wholes, not just their individual constituent parts (see Lo 2001 for an overview and critique of Callicott's changing position over the last two decades; also

see Ouderkirk and Hill (eds.) 2002 for debates between Callicott and others concerning the metaethical and metaphysical foundations for the land ethic and also its historical antecedents).

Individual natural entities (whether sentient or not, living or not), Andrew Brennan (1984) argues, are not designed by anyone to fulfill any purpose and therefore lack ‘intrinsic function’ (i.e., the function of a thing that constitutes part of its essence or identity conditions). This, he proposes, is a reason for thinking that individual natural entities should not be treated as mere instruments, and thus a reason for assigning them intrinsic value. Furthermore, he argues that the same moral point applies to the case of natural ecosystems, to the extent that they lack intrinsic function. In the light of Brennan's proposal, Eric Katz (1991 and 1997) argues that all natural entities, whether individuals or wholes, have intrinsic value in virtue of their ontological independence from human purpose, activity, and interest, and maintains the deontological principle that nature as a whole is an ‘autonomous subject’ which deserves moral respect and must not be treated as a mere means to human ends. Carrying the project of attributing intrinsic value to nature to its ultimate form, Robert Elliot (1997) argues that naturalness itself is a property in virtue of possessing which all natural things, events, and states of affairs, attain intrinsic value. Furthermore, Elliot argues that even a consequentialist, who in principle allows the possibility of trading off intrinsic value from naturalness for intrinsic value from other sources, could no longer justify such kind of trade-off in reality. This is because the reduction of intrinsic value due to the depletion of naturalness on earth, according to him, has reached such a level that any further reduction of it could not be compensated by any amount of intrinsic value generated in other ways, no matter how great it is.

As the notion of ‘natural’ is understood in terms of the lack of human contrivance and is often opposed to the notion of ‘artifactual’, one much contested issue is about the value of those parts of nature that have been interfered with by human artifice -- for instance, previously degraded natural environments which have been humanly restored. Based on the premise that the properties of being naturally evolved and having a natural continuity with the remote past are ‘value adding’ (i.e., adding intrinsic value to those things which possess those two properties), Elliot argues that even a perfectly restored environment would necessarily lack those two value-adding properties and therefore be less valuable than the originally undegraded natural environment. Katz, on the other hand, argues that a restored nature is really just an artifact designed and created for the satisfaction of human ends, and that the value of restored environments is merely instrumental. However, some critics have pointed out that advocates of moral dualism between the natural and the artifactual run the risk of diminishing the value of human life and culture, and fail to recognize that the natural environments interfered with by humans may still have morally relevant qualities other than pure naturalness (see Lo 1999). Two other issues central to this debate are that the key concept ‘natural’ seems ambiguous in many different ways (see Hume 1751, App. 3, and Brennan 1988, Ch. 6, Elliot 1997, Ch. 4), and that those who argue that human interference reduces the intrinsic value of nature seem to have simply assumed the crucial premise that naturalness is a source of intrinsic value. Some thinkers maintain that the natural, or the ‘wild’ construed as that which ‘is not humanized’ (Hettinger and Throop 1999, p. 12) or to some degree ‘not under human control’ (*ibid*., p. 13) is intrinsically valuable. Yet, as Bernard Williams points out (Williams 1992), we may, paradoxically, need to use our technological powers to retain a sense of something not being in our power. The retention of wild areas may thus involve planetary and ecological management to maintain, or even ‘imprison’ such areas (Birch 1990), raising a question over the extent to which national parks and

wilderness areas are free from our control. An important message underlying the debate, perhaps, is that even if ecological restoration is achievable, it might have been better to have left nature intact in the first place.

As an alternative to consequentialism and deontology both of which consider ‘thin’ concepts such as ‘goodness’ and ‘rightness’ as essential to morality, virtue ethics proposes to understand morality -- and assess the ethical quality of actions -- in terms of ‘thick’ concepts such as ‘kindness’, ‘honesty’, ‘sincerity’ and ‘justice’. As virtue ethics speaks quite a different language from the other two kinds of ethical theory, its theoretical focus is not so much on what kinds of things are good/bad, or what makes an action right/wrong. One question central to virtue ethics is what the moral reasons are for acting one way or another. For instance, from the perspective of virtue ethics, kindness and loyalty would be moral reasons for helping a friend in hardship. These are quite different from the deontologist's reason (that the action is demanded by a moral rule) or the consequentialist reason (that the action will lead to a better over-all balance of good over evil in the world). From the perspective of virtue ethics, the motivation and justification of actions are both inseparable from the character traits of the acting agent. Furthermore, unlike deontology or consequentialism the moral focus of which is other people or states of the world, one central issue for virtue ethics is how to live a flourishing human life, this being a central concern of the moral agent himself or herself. ‘Living virtuously’ is Aristotle's recipe for flourishing. Versions of virtue ethics advocating virtues such as ‘benevolence’, ‘piety’, ‘filiality’, and ‘courage’, have also been held by thinkers in the Chinese Confucian tradition. The connection between morality and psychology is another core subject of investigation for virtue ethics. It is sometimes suggested that human virtues, which constitute an important aspect of a flourishing human life, must be compatible with human needs and desires, and perhaps also sensitive to individual affection and temperaments. As its central focus is human flourishing as such, virtue ethics may seem unavoidably anthropocentric and unable to support a genuine moral concern for the nonhuman environment. But just as Aristotle has argued that a flourishing human life requires friendships and one can have genuine friendships only if one genuinely values, loves, respects, and cares for one's friends for their own sake, not merely for the benefits that they may bring to oneself, some have argued that a flourishing human life requires the moral capacities to value, love, respect, and care for the nonhuman natural world as an end in itself (see O'Neill 1992, O'Neill 1993, Barry 1999).

5. Wilderness and Developing Trends

Despite the variety of positions in environmental ethics developed over the last thirty years, they have largely focused on issues concerned with ‘wilderness’ and the reasons for its preservation. Little attention was paid to the built environment, although this is the one in which most people spend most of their time. In post-war Britain, for example, cheaply constructed new housing developments were often poor replacements for traditional communities. They have been associated with lower amounts of social interaction and increased crime compared with the earlier situation. The destruction of highly functional high-density traditional housing, indeed, might be compared with the destruction of highly diverse ecosystems and biotic communities. Likewise, the loss of the world's huge diversity of natural languages has been mourned by many, not just professionals with an interest in linguistics. Urban and linguistic

environments are just two of the many ‘places’ inhabited by humans. Perhaps the philosophical literature on the natural environment can be extended to cover environments of those other sorts as well (see King 2000, Light forthcoming, and Palmer forthcoming, for such an attempt).

The importance of wilderness experience to the human psyche has been emphasized by many environmental philosophers. Næss, for instance, urges us to ensure we spend time dwelling in situations of intrinsic value, whereas Rolston meditates in the wilderness on the importance of ‘re-creation’ afforded by experiences of wild nature. However, lifestyles in which enthusiasms for mountaineering, nature rambles and woodland meditations can be indulged demand a standard of living that is far beyond the dreams of most of the world's population. Hugh Stretton (1976) has characterized those environmentalists ‘driven chiefly by love of the wilderness’ as ‘natural aristocrats’. Mass access to wild places would likely destroy the very values held in high esteem by Stretton's ‘aristocrats’. So a question hangs over how to reconcile limiting the access of people to wilderness while maintaining the individual freedoms central to liberal democracies. Furthermore, lovers of wilderness sometimes consider the high human populations in some developing countries as a key problem underlying the environmental crisis. Rolston (1996), for instance, claims that humans are a kind of planetary ‘cancer’. He maintains that while “feeding people always seems humane, ... when we face up to what is really going on, by just feeding people, without attention to the larger social results, we could be feeding a kind of cancer.” This remark is meant to justify the view that saving nature should, in some circumstances, have a higher priority than feeding people. But such a view has been criticized for seeming to reveal a degree of misanthropy, directed at those human beings least able to protect and defend themselves (see Attfield 1998, Brennan 1998a). Guha's worries about the elitist and ‘missionary’ tendencies of some kinds of ‘deep’ green environmentalism in certain rich western countries can be quite readily extended to theorists such as Rolston. Can such elitism of the environmental sort ever be democratized? How can the psychically-reviving power of the wild become available to those living in the slums of Calcutta or Sao Paolo? These questions so far lack convincing answers.

The economic conditions which support the kind of enjoyment of wilderness by Stretton's ‘aristocrats’, and more generally the lifestyles of many people in the affluent countries, seem implicated in the destruction and pollution which has provoked the environmental turn in the first place. For those in the rich countries, for instance, engaging in outdoor recreations usually involves the motor car. Car dependency, however, is at the heart of many environmental problems, a key factor in urban pollution, while at the same time central to the economic and military activities of many nations and corporations, for example those activities associated with securing and exploiting oil resources. A range of new moral and political problems may open up for us. These include the problems of the environmental cost of tourist access to wilderness areas, and the fair ways in which limited access could be arranged to areas of natural diversity and beauty. International problems concern the exploitation of people in the world's economically poor nations as an aspect of an economic system supporting the lifestyles of the wealthy, and also oppression and war associated with securing oil and other resources of significance to the industrialized countries. In an increasingly crowded world, the answers to such problems will not be obvious and may require academic co-operation between philosophers and workers in other disciplines in an attempt to solve them.

Connections between environmental destruction, unequal resource consumption, poverty and the global economic order have been discussed by political scientists, development theorists, geographers and economists as well as by philosophers. Links between economics and environmental ethics are particularly well established. Work by Mark Sagoff (1988), for instance, has played a major part in bringing the two fields together (also see Shrader-Frechette 1987, O'Neill 1993, and Brennan 1995). Sagoff argued forcefully against confusing values with matters of preference (even considered preference), and claimed that ¡V as citizens rather than consumers ¡V people are concerned about values that cannot plausibly be monetized. The potentially misleading appeal to economic reason used to justify the expansion of the corporate sector has also come under critical scrutiny (see Korten 1999, Ch. 2, Plumwood (formerly V. Routley) 2002). These critiques do not aim to eliminate economics from environmental thinking; rather, they resist any reductive, and strongly anthropocentric, tendency to believe that all social problems are 'fundamentally' or 'essentially' economic. Interestingly, many of the assessments of issues concerned with biodiversity, ecosystem health, poverty, environmental justice and sustainability look at both human and environmental issues, eschewing in the process commitment either to a purely anthropocentric or purely non-anthropocentric ethic (Hayward and O'Neill 1997, and Dobson 1999 for collections of essays looking at the links between sustainability, justice, welfare and the distribution of environmental goods).

Other interdisciplinary approaches link environmental ethics with biology, policy studies, public administration, political theory, cultural history, post-colonial theory, literature, geography, and human ecology (for some examples, see Norton, Hutchins, Stevens, Maple 1995, Shrader-Frechette 1984, Gruen and Jamieson (eds.) 1994, Karliner 1997, Diesendorf and Hamilton 1997). The newest collection on environmental ethics (Schmidtz and Willott 2002) contains, in addition to readings on the questions discussed in the present article, sections devoted explicitly to cost-benefit analysis and environmental policy, the impact of cities on resource consumption, poverty as an environmental problem, sustainability and the growth of human population. This is in keeping with the developing philosophical interest in ethical analysis of policy fundamentals. The future development of environmental ethics may depend on these, and other interdisciplinary synergies, as much as on its anchorage within philosophy.

Bibliography

- Aquinas, T. *Summa Contra Gentiles*, trans. V. J. Bourke, London: University of Notre Dame Press, 1975.
- Aristotle. *Politics*, trans. E. Barker, Oxford: Oxford University Press, 1948.
- Aiken, W. 1984. 'Ethical Issues in Agriculture', in T. Regan (ed) *Earthbound: New Introductory Essays in Environmental Ethics*, New York: Random House, pp. 274-88.
- Anker, P. and Witoszek, N. 1998. 'The Dream of the Biocentric Community and the Structure of Utopias', *Worldviews* 2: 239-56.
- Attfield, R. 1987. *A Theory of Value and Obligation*, London: Croom Helm.
- -----. 1998. 'Saving Nature, Feeding People, and Ethics', *Environmental Values* 7: 291-304.
- Barry, J. 1999. *Rethinking Green Politic*s, London: Sage.
- Bentham, J. 1789. I*ntroduction to the Principles of Morals and Legislation*, Oxford: Basil

Blackwell, 1948.

- Birch, T. 1990 ’The Incarceration of Wilderness: Wilderness Areas as Prisons’, *Environmental Ethics* 12:3-26.
- Bookchin, M. 1980. *Toward an Ecological Society*, Montreal: Black Rose Books.
- -----. 1987. ‘Social Ecology Versus Deep Ecology’, *Green Perspectives: Newsletter of the Green Program Project*, numbers 4, 5 reprinted in Witoszek and Brennan 1999, pp. 281-301.
- -----. 1990. *The Philosophy of Social Ecology*, Montreal: Black Rose Books.
- Brennan, A. 1984. ‘The Moral Standing of Natural Objects’, *Environmental Ethics* 6: 35-56
- -----. 1988. *Thinking About Nature*, London Routledge.
- -----. 1995. ‘Ethics, Ecology and Economics’, *Biodiversity and Conservation* 4: 798-811.
- -----. 1998a. ‘Poverty, Puritanism and Environmental Conflict’, *Environmental Values* 7: 305-31.
- -----. 1998b. ‘Bioregionalism -- a Misplaced Project?’, *Worldviews* 2: 215-37.
- -----. 1999 ‘Comment: Pluralism and Deep Ecology’, in Witoszek and Brennan 1999
- Callicott, J. B. 1980. ‘Animal Liberation, A Triangular Affair’, reprinted in Callicott 1989, pp. 15-38.
- -----. 1985. ‘Intrinsic Value, Quantum Theory, and Environmental Ethics’, reprinted in Callicott 1989, pp. 157-74.
- -----. 1988. ‘Animal liberation and Environmental Ethics: Back Together Again’, reprinted in Callicott 1989, pp. 49-59.
- -----. 1989. *In Defense of the Land Ethic: Essays in Environmental Philosophy*, Albany: SUNY Press.
- -----. 1999. *Beyond the Land Ethic: More Essays in Environmental Philosophy*, Albany: SUNY Press.
- Carson, R. 1963. *Silent Spring*, London: Hamish Hamilton.
- Cheney, J. 1989. ‘Postmodern Environmental Ethics: Ethics as Bioregional Narrative’, *Environmental Ethics* 11: 117-34.
- Clark, S. R. L. 1977. *The Moral Status of Animals*, Oxford: Oxford University Press.
- Cohen, M. P. 1984. *The Pathless Way: John Muir and American Wilderness*, Madison: University of Wisconsin Press.
- Collins, S. 1974. *A Different Heaven and Earth*, Valley Forge: Judson Press.
- Crisp, R. 1998. ‘Animal Liberation is not an Environmental Ethic: A Response to Dale Jamieson’, *Environmental Values* 7: 476-8.
- d'Eaubonne , F. 1974. *Le Feminisme ou la Mort*, Paris: P. Horay
- Devall and Sessions 1985. *Deep Ecology: Living as if Nature Mattered*, Salt Lake City: Peregrine Smith.
- de Shalit, A. 1994. *Why Does Posterity Matter?* London: Routledge.
- -----. 1996. ‘Ruralism or Environmentalism?’ *Environmental Values* 5: 47-58.
- Diesendorf, M. and Hamilton, C. 1997. *Human Ecology, Human Economy*, St Leonards, NSW: Allen and Unwin.
- Dobson, A. 1990. *Green Political Thought*, London: Harper Collins.
- Dobson, A. (ed.) 1999 *Fairness and Futurity: Essays on Environmental Sustainability and Social Justice*, Oxford: Oxford University Press
- Eckersley, R. 1992. *Environmentalism and Political Theory*, London: UCL Press.

- Elliot, R. 1982. ‘Faking Nature’, *Inquiry* 25: 81-93.
- -----. 1997. *Faking Nature*, London: Routledge.
- Elliot, R. and Gare, A. (eds) 1983. *Environmental Philosophy: A Collection of Readings*, Milton Keynes: Open University Press.
- Feinberg, J. 1974. ‘The Rights of Animals and Unborn Generations’, in W. T. Blackstone (ed.), *Philosophy and Environmental Crisis*, Athens: University of Georgia Press, pp. 43-68.
- Ferré, F. 1996. ‘Persons in Nature: Toward an Applicable and Unified Environmental Ethics’, *Ethics and the Environment* 1: 15-25.
- Ferry, L. 1995. *The New Ecological Order*, translated C. Volk, Chicago: Chicago University Press.
- Fox, W. 1984. ‘Deep Ecology: A New Philosophy of Our Time?’ *The Ecologist* 14: 194-200.
- Green, K. 1994. ‘Freud, Wollstonecraft and Ecofeminism’, *Environmental Ethics* 16: 117-34.
- Grosz, E. 1989. *Sexual Subversions*, London: Allen and Unwin.
- Gruen, L. and Jamieson, D. (eds) 1994. *Reflecting on Nature*, New York: Oxford University Press.
- Guha, R. 1989. ‘Radical American Environmentalism and Wilderness Preservation: A Third World Critique’, *Environmental Ethics* 11: 71-83.
- -----. 1999. ‘Radical American Environmentalism Revisited’, in Witoszek and Brennan (eds.) 1999, pp. 473-9
- Hayward, Tim, and O'Neill, John, (eds.) 1997 *Justice, Property and the Environment: Social and Legal Perspectives*, Aldershot: Ashgate Publishing Co., 1997.
- Hettinger, N and Throop, B. 1999. ’Refocusing Ecocentrism’, *Environmental Ethics*, 21: 3-21
- Horkheiner, M. and Adorno, T. 1969. *Dialectic of Enlightenment*, trans. Cumming, J., New York: Seabury Press 1972.
- Hume, D. 1751. *An Enquiry Concerning the Principles of Morals*, ed. T. L. Beauchamp, Oxford: Oxford University Press, 1998.
- Jamieson, D. 1998. ‘Animal Liberation is an Environmental Ethic’, *Environmental Values* 7: 41-57.
- Kant, I. ‘Duties to Animals and Spirits’, in Louis Infield trans., *Lectures on Ethics*, New York: Harper and Row, 1963.
- Karliner, J. 1997 *The Corporate Planet*, San Francisco: Sierra Club Books
- Katz, E. 1991. ‘Restoration and Redesign: The Ethical Significance of Human Intervention in Nature’, *Restoration and Management Notes* 9: 90-6.
- -----. 1997. *Nature as Subject*, New York: Rowman and Littlefield.
- Kheel, M. 1985. ‘The Liberation of Nature: A Circular Affair’, *Environmental Ethics* 7: 135-49
- King, R. 2000. ‘Environmental Ethics and the Built Environment’, *Environmental Ethics* 22: 115-31
- King, Y. 1989a. ‘The Ecology of Feminism and the Feminism of Ecology’, in J. Plant (ed.), *Healing the Wounds*, Philadelphia: New Society Publishers: 18-28.
- King, Y. 1989b. ‘Healing the Wounds: Feminism, Ecology, and Nature/Culture Dualism’, in A. M. Jaggar and S. R. Bordo (eds.) *Gender/Body/Knowledge: Feminist Reconstruction of Being and Knowing*, New Brunswick: Rutgers University Press, pp. 115-41.
- Korten, D 1999 *The Post-CorporateWorld*, Hartford: Kumarian Press

- Leopold, A. 1949. *A Sand County Almanac*, Oxford: Oxford University Press.
- Light, A. 1996. ‘Callicott and Naess on Pluralism’, *Inquiry* 39: 273-294.
- -----. 2001. ‘The Urban Blindspot in Environmental Ethics’, *Environmental Politics* 10: 7-35.
- Light, A. and Katz, E. 1996. *Environmental Pragmatism*, London: Routledge.
- List, P. C. 1993. *Radical Environmentalism*, Belmont: Wadsworth.
- Lo, Y. S. 1999. ‘Natural and Artifactual: Restored Nature as Subject’, *Environmental Ethics* 21: 247-66.
- -----. 2001. ‘The Land Ethic and Callicott's Ethical System (1980-2001): An Overview and Critique’, *Inquiry* 44: 331-58.
- Meadows, D. H., Meadows, D. L., Randers, J., and Behrens, W. W. 1972. *The Limits to Growth*, New York: New American Library.
- Mill, J. S. 1874. ‘Nature’, in *Three Essays on Religion*, London: Longmans, Green, Reader and Dyer.
- Montaigne, M. de 1991. *The Complete Essays*, trans. M. A. Screech, Harmondworth: Penguin.
- Mumford, L. 1934. *Technics and Civilization*, London: Secker and Warburg.
- Mumford, L. 1961. *The City in History*, New York: Harcourt, Brace, Jovanovich
- Næss, A. 1973. ‘The Shallow and the Deep, Long-Range Ecology Movement’, *Inquiry* 16, reprinted in Sessions 1995, pp. 151-5.
- -----. 1989. *Ecology, Community, Lifestyle*, trans. and ed. D. Rothenberg, Cambridge: Cambridge University Press.
- Nash, R, (ed) 1990. *American Environmentalism: Readings in Conservation History*, New York: McGraw-Hill.
- Norton, B. 1991. *Toward Unity Among Environmentalists*, New York: Oxford University Press.
- Norton, B., Hutchins, M., Stevens, E. and Maple, T. L. (eds) 1995. *Ethics on the Ark*, Washington: Smithsonian Institution Press.
- O'Neill, J. 1992. ‘The Varieties of Intrinsic Value’, *Monist* 75: 119-137.
- -----. 1993. *Ecology, Policy and Politics*, London: Routledge.
- Ouderkirk, W. and Hill, J. (eds.) 2002. *Land, Value, Community: Callicott and Environmental* , Albany: State University of New York.
- Palmer, C. forthcoming. *‘Placing Animals in Urban Environmental Ethics’*; to appear in *Journal of Social Philosophy* 2003.
- Passmore, J. 1974. *Man's Responsibility for Nature*, London: Duckworth, 2nd ed., 1980.
- Plumwood, V. 1993. *Feminism and the Mastery of Nature*, London: Routledge.
- -----. 1999. ‘Comments: Self-Realization and Man Apart? The Reed-Næss Debate’, in Witoszek and Brennan (eds.) 1999, pp. 206-10.
- -----. 2002 *Environmental Culture*, London: Routledge
- Porter, G. and Welsh Brown, J. 1991. *Global Environmental Politics*, Boulder: Westview Press.
- Regan, T. 1983. *The Case for Animal Rights*, London: Routledge & Kegan Paul.
- Regan, T. and Singer, P. (eds.) 1976. *Animal Rights and Human Obligations*, Englewood Cliffs: Prentice Hall.
- Rolston, H. 1975. ‘Is There an Ecological Ethic?’, *Ethics* 85: 93-109.
- -----. 1989. *Philosophy Gone Wild*, New York: Prometheus Books.
- -----. 1996. ‘Feeding People versus Savng Nature?’, in W. Aiken and H. LaFollette (eds.) *World*

Hunger and Morality, Englewood Cliffs: Prentice Hall, pp. 248-67

- Rousseau, J. J. 1782. *Reveries of the Solitary Walker*, trans. P. France, Penguin Books, 1979.
- Routley, R. 1973. ‘Is there a need for a new, an environmental ethic?’ *Proceedings of the 15th World congress of Philosophy*, vol. 1 pp. 205-10, Sophia: Sophia Press (see also Sylvan, R.).
- Routley, R. and Routley, V. 1980. ‘Human Chauvinism and Environmental Ethics’ in Mannison, D., McRobbie, M. A., and Routley, R. (eds.) *Environmental Philosophy*, Canberra: Australian National University, Research School of Social Sciences, pp. 96-189.
- Sagoff, M. 1984. ‘Animal Liberation and Environmental Ethics: Bad Marriage, Quick Divorce’, *Osgoode Hall Law Journal* 22:297-307.
- -----. 1988. *The Economy of the Earth*, Cambridge: Cambridge University Press.
- Schmidtz, D. and Willott, E. 2002 *Environmental Ethics: What Really Matters, What Really Works*, New York: Oxford University Press.
- Sessions, G. (ed) 1995. *Deep Ecology for the 21st Century*, Boston: Shambhala 1995.
- Shrader-Frechette, K. 1984. *Science Policy, Ethics and Economic Methodology*, Dordrecht: D Reidel
- -----. 1987. ‘The real risks of risk-cost-benefit analysis’, in P. T. Durbin (ed.), *Technology and Responsibility*, Dordrecht: D Reidel, pp. 343-57.
- -----. 1996. ‘Individualism, Holism, and Environmental Ethics’, *Ethics and the Environment* 1: 55-69.
- Singer, P. 1975. *Animal Liberation*, New York: Random House.
- -----. 1993. *Practical Ethics*, Cambridge: Cambridge University Press, 2nd ed.
- Stone, C. D. 1972. ‘Should Trees Have Standing?’, *Southern California Law Review* 45:450-501 ; later published with a descriptive introduction as *Should Trees Have Standing?*, Los Angeles: Kaufmann, 1974, and reprinted in Schmidtz and Willott 2002.
- Stretton, H. 1976. *Capitalism, Socialism and the Environment*, Cambridge: Cambridge University Press.
- Taylor, P. 1981. ‘The Ethics of Respect for Nature’, *Environmental Ethics* 3: 197-218.
- -----. 1986. *Respect for Nature*, Princeton: Princeton University Press.
- Varner, G. 1998. *In Nature's Interests? Interests, Animal Rights, and Environmental Ethics*, Oxford: Oxford University Press
- -----. 2000. ‘Sentientism’, in D. Jamieson (ed.) *A Companion to Environmental Philosophy*, Oxford: Blackwell, pp.192-203.
- Vogel, S. 1996. *Against Nature: The Concept of Nature in Critical Theory*, Albany: State University of New York Press.
- Warren, K. J. 1987. ‘Feminism and Ecology: Making Connections’, *Environmental Ethics* 9: 3-21.
- -----. 1990. ‘The Power and Promise of Ecological Feminism’, *Environmental Ethics* 12: 125-46.
- -----. 1999. ‘Ecofeminist Philosophy and Deep Ecology’, in Witoszek and Brennan (eds.) 1999, pp. 255-69.
- Warren, K. J. (ed) 1994. *Ecological Feminism*, London: Routledge.
- White, L. 1967. ‘The Historical Roots of Our Ecological Crisis’, *Science*, 55:1203-1207 ; reprinted in Schmidtz and Willott 2002.
- Williams, B. 1992. ‘Must a Concern for the Environment be Centred on Human Beings?’,

reprinted in his *Making Sense of Humanity and Other Philosophical Papers*, Cambridge: Cambridge University Press, 1995: 233-40.
- Witoszek, N. 1997. 'Arne Næss and the Norwegian Nature Tradition', *Worldviews* 1: 57-73.
- Witoszek, N. and Brennan, A. (eds) 1999. *Philosophical Dialogues: Arne Næss and the Progress of Eco-Philosophy*, New York: Rowan and Littlefield.

Other Internet Resources

- The International Society for Environmental Ethics
  Has several useful links, including ones to discussion groups on environmental ethics, the quarterly newsletter of the society and a comprehensive searchable bibliography on environmental ethics.
- Centre for Applied Ethics (U. British Columbia) Contains several links to environmental ethics resources.
- International Association for Environmental Philosophy (Site maintained by Robert Frodeman, Colorado School of Mines)
  A recently formed group focuses on environmental philosophy from the phenomenological and 'continental' perspective.

Related Entries

common good | communitarianism | consequentialism | ecology | environmentalism | ethics: deontological | ethics: virtue | feminism, interventions: feminist ethics | rights

Acknowledgements

The authors are deeply grateful to the following people who gave generously of their time and advice to help shape the final structure of this entry: Clare Palmer, Mauro Grün, Lori Gruen, Gary Varner, William Throop, and Edward N. Zalta.

Andrew Brennan
abrennan@cyllene.uwa.edu.au
and
Yeuk-Sze Lo
University of Western Australia
ynorvas@cyllene.uwa.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 3, 2002
Content last modified: June 3, 2002

Communitarianism

Modern-day communitarianism began in the upper reaches of Anglo-American academia in the form of a critical reaction to John Rawls landmark 1971 book *A Theory of Justice*.[1] Drawing primarily upon the insights of Aristotle and Hegel, political philosophers such as Alasdair MacIntyre, Michael Sandel, Charles Taylor and Michael Walzer disputed Rawls assumption that the principal task of government is to secure and distribute fairly the liberties and economic resources individuals need to lead freely chosen lives. These critics of liberal theory never did identify themselves with the communitarian movement (the communitarian label was pinned on them by others, usually critics)[2], much less offer a grand communitarian theory as a systematic alternative to liberalism. Nonetheless, certain core arguments meant to contrast with liberalisms devaluation of community recur in the works of the four theorists named above,[3] and for purposes of clarity one can distinguish between claims of three sorts: methodological claims about the importance of tradition and social context for moral and political reasoning, ontological or metaphysical claims about the social nature of the self, and normative claims about the value of community.[4]

This essay is therefore divided in three parts, and for each part I present the main communitarian claims, followed by an argument (in each part) that philosophical concerns in the 1980s have largely given way to the political concerns that motivated much of the communitarian critique in the first place.

1. Universalism Versus Particularism

Communitarians have sought to deflate the universal pretensions of liberal theory. The main target has been Rawls description of the original position as an ‘Archemedian point’ from which the structure of a social system can be appraised, a position whose special virtue is that it allows us to regard the human

condition 'from the perspective of eternity',[5] from all social and temporal points of view. Whereas Rawls seemed to present his theory of justice as universally true, communitarians argued that the standards of justice must be found in forms of life and traditions of particular societies and hence can vary from context to context. Alasdair MacIntyre and Charles Taylor argued that moral and political judgment will depend on the language of reasons and the interpretive framework within which agents view their world, hence that it makes no sense to begin the political enterprise by abstracting from the interpretive dimensions of human beliefs, practices, and institutions.[6] Michael Walzer developed the additional argument that effective social criticism must derive from and resonate with the habits and traditions of actual people living in specific times and places. Even is there is nothing problematic about a formal procedure of universalizability meant to yield a determinate set of human goods and values, 'any such set would have to be considered in terms so abstract that they would be of little use in thinking about particular distributions.'[7] In short, liberals who ask what is just by abstracting from particular social contexts are doomed to philosophical incoherence and liberal theorists who adopt this method to persuade people to do the just thing are doomed to political irrelevance.

Rawls has since tried to eliminate the universalist presuppositions from his theory. In *Political Liberalism*,[8] he argues in a communitarian vein that his conception of the person as impartial citizen provides the best account of liberal-democratic political culture and that his political aim is only to work out the rules for consensus in political communities where people are willing to try for consensus. In the *Law of Peoples*,[9] he explicitly allows for the possibility that liberalism may not be exportable at all times and places, sketching a vision of a 'decent, well-ordered society' that liberal societies must tolerate in the international realm. Such a society, he argues, need not be democratic, but it must be non-aggressive towards other communities, and internally it must have a 'common good conception of justice', a 'reasonable consultation hierarchy', and it must secure basic human rights. Having said that, one still gets the sense that the liberal vision laid out in *A Theory of Justice* is the best possible political ideal, one that all rational individuals would want if they were able to choose between the available political alternatives. There may be justifiable non-liberal regimes, but these should be regarded as second best to be tolerated and perhaps respected, not idealized or emulated.

Other liberal theorists have taken a harder line against communitarian concessions, arguing that liberal theory can and should present itself as a universally valid ideal. Brian Barry, for one, opens his widely cited book *Justice as Impartiality* by boldly affirming the universality of his theory: 'I continue to believe in the possibility of putting forward a universally valid case in favor of liberal egalitarian principles'.[10] Barry does recognize that a theory of justice must be anchored in substantive moral considerations, but his normative vision appears to be limited to the values and practices of liberal Western societies. He seems distinctly uninterested in learning anything worthwhile from non-Western political traditions for example, his discussion of things Chinese is confined to brief criticisms of the Cultural Revolution and the traditional practice of foot-binding. One might consider the reaction to a Chinese intellectual who puts forward a universal theory of justice that draws on the Chinese political tradition for inspiration and completely ignores the history and moral argumentation in Western societies, except for brief criticisms of slavery and imperialism.

Still, it must be conceded that 1980s communitarian theorists were less-than-successful at putting forward attractive visions of non-liberal societies. The communitarian case for pluralism for the need to respect and perhaps learn from non-liberal societies that may be as good as, if not better than, the liberal societies of the West may have been unintentionally undermined by their own use of (counter) examples. In *After Virtue*, Alasdair MacIntyre defended the Aristotelian ideal of the intimate, reciprocating local community bound by shared ends, where people simply assume and fulfill socially given roles.[11] But this pre-modern Gemeinschaft conception of an all-encompassing community that members unreflectively endorse seemed distinctly ill-suited for complex and conflict-ridden large-scale industrialized societies. In *Spheres of Justice,* Michael Walzer pointed to the Indian caste system, ‘where the social meanings are integrated and hierarchical’[12] as an example of a non-liberal society that may be just according to its own standards. Not surprisingly, few readers were inspired by this example of non-liberal justice (not to mention the fact that many contemporary Indian thinkers view the caste system as an unfortunate legacy of the past that Indians should strive hard to overcome). In short, this use of ill-informed examples may have unintentionally reinforced the view that there are few if any justifiable alternatives to liberalism in modern societies. Communitarians could score some theoretical points by urging liberal thinkers to be cautious about developing universal arguments founded exclusively on the moral argumentation and political experience of Western liberal societies, but few thinkers would really contemplate the possibility of non-liberal practices appropriate for the modern world so long as the alternatives to liberalism consisted of Golden Ages, caste societies, fascism, or actually-existing communism. For the communitarian critique of liberal universalism to have any lasting credibility, thinkers need to provide compelling counter-examples to modern-day liberal-democratic regimes and 1980s communitarians came up short.

By the 1990s, fairly abstract methodological disputes over universalism versus particularism faded from academic prominence, and the debate now centers on the theory and practice of universal human rights. This is largely due to the increased political salience of human rights since the collapse of communism in the former Soviet bloc. On the liberal side, the new, more political voices for liberal universalism have been represented by the likes of Francis Fukuyama, who famously argued that liberal democracys triumph over its rivals signifies the end of history.[13] This view also revived (and provoked) the second wave communitarian critique of liberal universalism and the debate became much more concrete and political in orientation.

Needless to say, the brief moment of liberal euphoria that followed the collapse of the communism in the Soviet bloc has given way to a sober assessment of the difficulties of implementing liberal practices outside the Western world. It is now widely recognized that brutal ethnic warfare, crippling poverty, environmental degradation, and pervasive corruption, to name some of the more obvious troubles afflicting the developing world, pose serious obstacles to the successful establishment and consolidation of liberal democratic political arrangements. But these were seen as unfortunate (hopefully temporary) afflictions that may delay the end of history when liberal democracy has finally triumphed over its rivals. They were not meant to pose a challenge to the *ideal* of liberal democracy. It was widely assumed that liberal democracy is something that all rational individuals *would* want if they could get it.

The deeper challenge to Western liberal democracy has emerged from the East Asian region.[14] In the 1990s, the debate revolved around the notion of Asian values, a term devised by several Asian officials and their supporters for the purpose of challenging Western-style civil and political freedoms. Asians, they claim, place special emphasis upon family and social harmony, with the implication that those in the chaotic and crumbling societies of the West should think twice about intervening in Asia for the sake of promoting human rights and democracy. As Singapores Lee Kuan Yew put it, Asians have ‘little doubt that a society with communitarian values where the interests of society take precedence over that of the individual suits them better than the individualism of America’.[15] Such claims attracted international attention primarily because East Asian leaders seemed to be presiding over what a recent U.N. human development report called ‘the most sustained and widespread development miracle of the twentieth century, perhaps all history’.[16] In 1997-98, however, the East Asian miracle seemed to have collapsed. And it looks like Asian values was one casualty of the crisis.

The political factors that focused attention on the East Asian challenge remain in place, however. East Asian economies (with the notable exception of Indonesia) have been slowly recovering. China in particular looks set to become an economic and political heavyweight with the power to seriously challenge the hegemony of Western liberal democratic values in international fora. Thus, one hears frequent calls for cross-cultural dialogue between the West and the East designed to understand and perhaps learn from the other side. Failing to take seriously East Asian political perspectives risks widening misunderstandings and setting the stage for hostilities that could have been avoided.

From a theoretical point of view, however, it must be conceded that the official debate on Asian values has not provided much of a challenge to dominant Western political outlooks. The main problem is that the debate has been led by Asian leaders who seem to be motivated primarily by political considerations, rather than by a sincere desire to make a constructive contribution to the debate on universalism versus particularism. Thus, it was easy to dismiss rightly so, in most cases the Asian challenge as nothing but a self-serving ploy by government leaders to justify their authoritarian rule in the face of increasing demands for democracy at home and abroad.

Still, it would be a mistake to assume that nothing of theoretical significance has emerged from East Asia. The debate on Asian values has also prompted critical intellectuals in the region to reflect on how they can locate themselves in a debate on human rights and democracy in which they had not previously played a substantial part. Neither wholly rejecting nor wholly endorsing the values and practices ordinarily realized through a liberal democratic political regime, these intellectuals are drawing on their own cultural traditions and exploring areas of commonality and difference with the West. Though often less provocative than the views of their governments in the sense that few argue for the wholesale rejection of Western-style liberal democracy with an East Asian alternative these unofficial East Asian viewpoints may offer more lasting contributions to the debate. Let me (briefly) note three relatively persuasive East Asian arguments for cultural particularism that contrast with traditional Western arguments for liberal universalism:[17]

1. Cultural factors can affect the *prioritizing* of rights, and this matters when rights conflict and it

must be decided which one to sacrifice. In other words, different societies may rank rights differently, and even if they face a similar set of disagreeable circumstances they may come to different conclusions about the right that needs to be curtailed. For example, U.S. citizens may be more willing to sacrifice a social or economic right in cases of conflict with a civil or political right: if neither the constitution nor a majority of democratically elected representatives support universal access to health care, then the right to health care regardless of income can be curtailed. In contrast, the Chinese may be more willing to sacrifice a civil or political liberty in cases of conflict with a social or economic right: there may be wide support for restrictions on the internal movement of farmers if these are necessary to guarantee the right of subsistence. Different priorities assigned to rights can also matter when it must be decided how to spend scarce resources. For example, East Asian societies with a Confucian heritage will place great emphasis upon the value of education, and they may help to explain the large amount of spending on education compared to other societies with similar levels of economic development.

2. Cultural factors can affect the *justification* of rights. In line with the arguments of 1980s communitarians such as Michael Walzer, it is argued that justifications for particular practices valued by Western-style liberal democrats should not be made by relying on the abstract and unhistorical universalism that often disables Western liberal democrats. Rather, they should be made from the inside, from specific examples and argumentative strategies that East Asians themselves use in everyday moral and political debate. For example, the moral language (shared even by some local critics of authoritarianism) tends to appeal to the value of community in East Asia, and this is relevant for social critics concerned with practical effect. One such communitarian argument is that democratic rights in Singapore can be justified on the grounds that they contribute to strengthening ties to such communities as the family and the nation (see below, section III).
3. Cultural factors can provide moral foundations for *distinctive* political practices and institutions (or at least different from those found in Western-style liberal democracies). In East Asian societies influenced by Confucianism, for example, it is widely held that children have a profound duty to care for elderly parents, a duty to be forsaken only in the most exceptional circumstances.[18] In political practice, it means that East Asian governments have an obligation to provide the social and economic conditions that facilitate the realization of this duty. Political debate tends to center on the question of whether the right to filial piety is best realized by means of a law that makes it mandatory for children to provide financial support for elderly parents as in mainland China, Japan, and Singapore or whether the state should rely more on indirect methods such as tax breaks and housing benefits that simply make at-home care for the elderly easier, as in Korea and Hong Kong. But the argument that there is a pressing need to secure this duty in East Asia is not a matter of political controversy.

Thinkers influenced by East Asian cultural traditions such Confucianism have also argued for distinctive as-yet-unrealized political practices and institutions that draw on widely-held cultural values for inspiration. For example, the Korean scholar Hahm Chaihark argues for the need to revive and adapt for the contemporary era the Confucian censorate, a traditional institution that played the role of monitoring the dealings of the Emperor.[19]

In contrast to 1980s communitarian thinkers, East Asian critics of liberal universalism have succeeded in pointing to particular non-liberal practices and institutions that may be appropriate for the contemporary world. Some of these may be appropriate only for societies with a Confucian heritage, others may also offer insights for mitigating the excesses of liberal modernity in the West. What cannot be denied is that they have carried forward the debate beyond the implausible alternatives to liberalism offered by 1980s communitarian thinkers.

It is worth emphasizing, however, that contemporary communitarians have not been merely defending parochial attachments to particular non-liberal moralities. Far from arguing that the universalist discourse on human rights should be entirely displaced with particular, tradition-sensitive political language, they have criticized liberals for not taking universality seriously enough, for failing to do what must be done to make human rights a truly universal ideal. These communitarians -- let us label them the cosmopolitan critics of liberal universalism -- have suggested various means of improving the philosophical coherence and political appeal of human rights.

In fact, there is little debate over the desirability of a core set of human rights, such as prohibitions against slavery, genocide, murder, torture, prolonged arbitrary detention, and systematic racial discrimination. These rights have become part of international customary law, and they are not contested in the public rhetoric of the international arena. Of course many gross violations occur off the record, and human rights groups such as Amnesty International have the task of exposing the gap between public allegiance to rights and the sad reality of ongoing abuse. This is largely practical work, however. There is not much point writing about or deliberating about the desirability of practices that everyone condemns at the level of principle.

But political thinkers and activists around the world can and do take different sides on many pressing human rights concerns that fall outside what Walzer terms the ‘minimal and universal moral code’.[20] This gray area of debate includes criminal law, family law, womens rights, social and economic rights, the rights of indigenous peoples, and the attempt to universalize Western-style democratic practices. The question is: how can the current thin list of universal human rights be expanded to include some of these contested rights?

Charles Taylor has put forward the following proposal.[21] He imagines a cross-cultural dialogue between representatives of different traditions. Rather than argue for the universal validity of their views, however, he suggests that participants should allow for the possibility that their own beliefs may be mistaken. This way, participants can learn from each others ‘moral universe’. There will come a point, however, when differences cannot be reconciled. Taylor explicitly recognizes that different groups, countries, religious communities, and civilizations hold incompatible fundamental views on theology, metaphysics, and human nature. In response, Taylor argues that a ‘genuine, unforced consensus’ on human rights norms is possible only if we allow for disagreement on the ultimate justifications of those norms. Instead of defending contested foundational values when we encounter points of resistance (and thus condemning the values we do not like in other societies), we should try to abstract from those beliefs for the purpose of working out an ‘overlapping consensus’ of human rights norms. As Taylor puts it, ‘we

would agree on the norms while disagreeing on why they were the right norms, and we would be content to live in this consensus, undisturbed by the differences of profound underlying belief'.[22]

While Taylors proposal moves the debate on universal human rights forward, it still faces certain difficulties. For one thing, it may not be realistic to expect that people will be willing to abstract from the values they care deeply about during the course of a global dialogue on human rights. Even if people agree to abstract from culturally specific ways of justifying and implementing norms, the likely outcome is a withdrawal to a highly general, abstract realm of agreement that fails to resolve actual disputes over contested rights. For example, participants in a cross-cultural dialogue can agree on the right not to be subject to cruel and unusual punishment while radically disagreeing upon what this means in practice a committed Muslim can argue that theft can justifiably be punished by amputation of the right hand,[23] whereas a Western liberal will want to label this an example of cruel and unusual punishment.

As we have seen, the debate on universalism versus particularism has moved from fairly abstract methodological disputes between Anglo-American philosophers to relatively concrete international political disputes between philosophers, social scientists, government officials, and NGO activists. The distinctive communitarian contribution has been to cast doubt on universal theories grounded exclusively in the liberal moralities of the Western world, on the grounds that cultural particularity should both make one sensitive to the possibility of justifiable areas of difference between the West and the rest and to the need for more cross-cultural dialogue for the purpose of improving the current thin human rights regime. Various contributions from East Asia and elsewhere have given some meat to these challenges to liberal universalism. In any case, let us now turn to the second main area of controversy between liberals and communitarians -- the debate over the self that has similarly moved from philosophy to politics.

2. The Debate Over the Self

Communitarian thinkers in the 1980s such as Michael Sandel and Charles Taylor argued that Rawlsian liberalism rests on an overly individualistic conception of the self. Whereas Rawls argues that we have a supreme interest in shaping, pursuing, and revising our own life-plans, he neglects the fact that our selves tend to be defined or constituted by various communal attachments (eg, ties to the family or to a religious tradition) so close to us that they can only be set aside at great cost, if at all. This insight led to the view that politics should not be concerned solely with securing the conditions for individuals to exercise their powers of autonomous choice, as we also need to sustain and promote the social attachments crucial to our sense of well-being and respect, many of which have been involuntarily picked up during the course of our upbringing. First, however, let us review the ontological or metaphysical debate over the self that led to this political conclusion.

In an influential essay titled 'Atomism', Charles Taylor objected to the liberal view that 'men are self-sufficient outside of society'.[24] Instead, Taylor defends the Aristotelian view that 'Man is a social animal, indeed a political animal, because he is not self-sufficient alone, and in an important sense is not self-sufficient outside a polis'.[25] Moreover, this atomistic view of the self can undermine liberal

society, because it fails to grasp the extent to which liberalism presumes a context where individuals are members of, and committed to, a society that promotes particular values such as freedom and individual diversity. Fortunately, most people in liberal societies do not really view themselves as atomistic selves.

But do liberal thinkers actually defend the idea that the self is created *ex-nihilo*, outside of any social context and that humans can exist (and flourish) independently of all social contexts? In fact, Taylors essay was directed at the libertarian thinker Robert Nozick. As it turns out, the communitarian critique of the atomistic self does not apply to Rawslian liberalism: in Part III of *Theory of Justice*, Rawls pays close attention to the psychological and social conditions that facilitate the formation of liberal selves committed to justice. But few readers ever got to Part III of Rawls massive tome, so communitarians got quite a bit of mileage from their critique of liberal atomism. This charge didnt stick, however.

While liberals may not have been arguing that individuals can *completely* extricate themselves from their social context, the liberal valuation of choice still seemed to suggest an image of a subject who impinges his will on the world.[26] Drawing on the insights of Heidegger and Wittgenstein, communitarians argued that this view neglects the extent to which individuals are embodied agents in the world. Far from acting in ways designed to realize an autonomously arrived-at life-plan, vast areas of our lives are in fact governed by unchosen routines and habits that lie in the background. More often than not we act in ways specified by our social background when we walk, dress, play games, speak, and so on without having formulated any goals or made any choices. It is only when things break down from the normal, everyday, unchosen mode of existence that we think of ourselves as subjects dealing with an external world, having the experience of formulating various ways of executing our goals, choosing from among those ways, and accepting responsibility for the outcomes of our actions. In other words, traditional intentionality is introduced at the point that our ordinary way of coping with things is insufficient. Yet this breakdown mode is what we tend to notice, and philosophers have therefore argued that most of our actions are occasioned by processes of reflection. Liberals have picked up this mistaken assumption, positing the idea of a subject who seeks to realize an autonomously arrived-at life-plan, losing sight of the fact that critical reflection upon ones ends is nothing more than one possibility that arises when our ordinary ways of coping with things is insufficient to get things done.

Some liberals have replied by recognizing the point that vast areas of our lives are governed by unchosen habits and routines, that the deliberate, effortful, choosing subject mode may be the exception rather than the rule. They emphasize, however, that the main justification for a liberal politics concerned primarily with securing the conditions for individuals to lead autonomous lives rests on the possibility and desirability of *normative* self-determination, that is,on the importance of making choices with respect to things that we *value*.[27] While it may be true that certain communal practices often, or even mostly, guide our behavior behind our backs, it doesnt follow that those practices ought to be valued, or reflectively endorsed in non-ordinary moments of existence, much less that the government ought somehow to promote these practices. And what liberals care about ultimately is the provision of the rights, powers, and opportunities that individuals need to develop and implement their own conceptions of the good life.

This qualified version of the liberal self, however, still seems to imply is that moral outlooks are, or should be, the product of individual choice. Ones social world, communitarians can reply, provides more than non-moral social practices like table manners and pronunciation norms it also provides some sort of orientation in *moral* space. We cannot make sense of our moral experience unless we situate ourselves within this given moral space, within the authoritative moral horizons. What Charles Taylor calls ‘higher, strongly evaluated goods’[28] -- the goods we should feel committed to, those that generate moral obligations on us regardless of our actual preferences are not somehow invented by individuals, but rather they are located within the social world which provides ones framework of the lower and the higher. Thus, the liberal ideal of a self who freely invents her own moral outlook, or private conception of the good, cannot do justice to our actual moral experience.

But once again, liberals need not deny the assumption that our social world provides a framework of the higher and the lower nor need it be presumed that we must regard our own moral outlook as freely invented. Will Kymlicka, for example, explicitly recognizes that things have worth for us in so far as they are granted significance by our culture, in so far as they fit into a pattern of activities which is recognized by those sharing a certain form of life as a way of leading a good life.[29] That ones social world provides the range of things worth doing, achieving, or being does not, however, undermine the liberal emphasis on autonomy, for there is still substantial room for individual choice to be made within this set. The best life is still the one where the individual *chooses* what is worth doing, achieving, or being, though it may be that this choice has to be made within certain framework which is itself unchosen.

Communitarians can reply by casting doubt on the view that choice is *intrinsically* valuable, that a certain moral principle or communal attachment is more valuable simply because it has been chosen following deliberation among alternatives by an individual subject. If we have a highest-order interest in choosing our central projects and life-plans, regardless of what is chosen, it ought to follow that there is something fundamentally wrong with unchosen attachments and projects. But this view violates our actual self-understandings. We ordinarily think of ourselves, Michael Sandel says, ‘as members of this family or community or nation or people, as bearers of this history, as sons or daughters of that revolution, as citizens of this republic’,[30] social attachments that more often than not are involuntarily picked up during the course of our upbringing, rational choice having played no role whatsoever. I didnt choose to love my mother and father, to care about the neighborhood in which I grew up, to have special feelings for the people of my country, and it is difficult to understand why anyone would think I have chosen these attachments, or that I ought to have done so. In fact, there may even be something distasteful about someone who questions the things he or she deeply cares about certainly no marriage could survive too long if fundamental understandings regarding love and trust were constantly thrown open for discussion! Nor is it obvious that, say, someone who performs a good deed following prolonged calculation of pros and cons is morally superior than a Mother-Teresa type who unreflectively, spontaneously acts on behalf of other peoples interests.

Liberals can reply that the real issue is not the desirability of choice, but rather the *possibility* of choice. There may well be some unchosen attachments that need not be critically reflected upon and endorsed,

and it may even be the case that excessive deliberation about the things we care about can occasionally be counter-productive. But some of our ends may be problematic and that is why we have a fundamental interest in being able to question and revise them. Most important is not choosing our own life-plans; rather, liberalism founded on the value of self-determination requires only that we be able to critically evaluate our ends *if need be*, hence that 'no end or goal is exempt from possible re-examination'.[31] For example, an oppressed woman has a fundamental interest in being able to critically reflect upon traditional understandings of what it means to be a good wife and mother, and it would be unjust to foreclose her freedom to radically revise her plans.

This response, however, still leaves open the possibility of a deep challenge to liberal foundations. Even we are indeed able to reexamine some attachments, the problem for liberalism arises if there are others so fundamental to our identity that they cannot be set aside, and that any attempt to do so will result in serious and perhaps irreparable psychological damage. In fact, this challenge to liberalism would only require that communitarians be able to identify one end one communal attachment, for example so constitutive to ones identity that it cannot be revised and rejected. A psychoanalyst, for example, may want to argue that (at least in some cases) it is impossible to choose to shed the attachment one feels for ones mother, and that an attempt may lead to perverse and unintended consequences. A feminist theorist may point to the mother-child relationship as an example of a constitutive feature of ones identity and argue that any attempt to deny this fails to be sensitive to womens special needs and experiences.[32] An anthropologist may argue on the basis of field observations that it is impossible for an Inuit person from Canadas far north to suddenly decide to stop being an Inuit and that the only sensible response is to recognize and accept this constitutive feature of his identity. Or a gay liberation activist may claim that it is both impossible and undesirable for gays to repress their biologically-given sexual identity. These arguments are not implausible, and they seem to challenge the liberal view that no particular end or commitment should be beyond critical reflection and open to revision.

Let us assume, for the sake of argument, that we can identify one particular attachment so deeply-embedded that it is impossible to really bring to conscious awareness and so significant for ones well-being that an individual can only forsake commitment to its good at the cost of being seriously psychologically disturbed. This end is beyond willed change and one loses a commitment to it at the price of being thrown into a state of disorientation where one is unable to take a stand on many things of significance.[33] Does this really threaten liberal politics? It may, if liberal politics really rests on the liberal self. Fortunately, that is not the case. Rereading some of the communitarian texts from the 1980s, there seems to have been an assumption that once you expose faulty foundations regarding the liberal self, the whole liberal edifice will come tumbling down. The task is to criticize the underlying philosophy of the self, win people on your side, and then we can move on to a brand new communitarian society that owes nothing to the liberal tradition. This must have been an exhilarating time for would-be revolutionaries, but more level-headed communitarians soon realized that overthrowing liberal rights was never part of the agenda. Even if liberals are wrong to deny the existence of constitutive ends even if the philosophical justifications for a liberal form of social organization founded on the value of reflective choice are rotten to the core -- there are still many, relatively pragmatic reasons for caring about rights in the modern world. To name some of the more obvious benefits, liberal rights contribute to security, political stability and economic modernization.

In short, the whole debate about the self appears to have been somewhat misconceived. Liberals were wrong to think they needed to provide iron-clad philosophies of the self to justify liberal politics, and communitarians were wrong to think that challenging those foundations was sufficient to undermine liberal politics. Not surprisingly, both sides soon got tired of debating the pros and cons of the liberal self. By the early 1990s, this liberal-communitarian debate over the self had effectively faded from view in Anglo-American philosophy.[34]

So what remains of the communitarian conception of the self? What may be distinctive about communitarians is that they are more inclined to argue that individuals have a vital interest in leading decent communal lives, with the political implication that there may be a need to sustain and promote the communal attachments crucial to our sense of well-being. This is not necessarily meant to challenge the liberal view that some of our communal attachments can be problematic and may need to be changed, thus that the state needs to protect our powers to shape, pursue, and revise our own life-plans. But our interest in community may occasionally conflict with our other vital interest in leading freely chosen lives, and the communitarian view is that the latter does not automatically trump the former in cases of conflict. On the continuum between freedom and community, communitarians are more inclined to draw the line towards the latter.

But these conflicts cannot be resolved in the abstract. Much turns on empirical analyses of actual politics to what extent our interest in community is indeed threatened by excess liberal politics, to what extent the state can play a role in remedying the situation, to what extent the nourishment of communal ties should be left to civil society, and so on. This is where the political communitarians of the last decade have shed some light. Let us now turn to the politics of community, the third major strand of the communitarian thought.

3. The Politics of Community

In retrospect, it seems obvious that communitarian critics of liberalism may have been motivated not so much by philosophical concerns as by certain pressing political concerns, namely, the negative social and psychological effects related to the atomistic tendencies of modern liberal societies. Whatever the soundness of liberal principles, in other words, the fact remains that many communitarians seems worried by a perception that traditional liberal institutions and practices have contributed to, or at least do not seem up to the task of dealing with, such modern phenomena as alienation from the political process, unbridled greed, loneliness, urban crime, and high divorce rates. And given the seriousness of these problems in the United States, it was perhaps inevitable that a second wave of 1990s communitarians such as Amitai Etzioni and William Galston would turn to the more practical political terrain of emphasizing social responsibility and promoting policies meant to stem the erosion of communal life in an increasingly fragmented world.[35] Much of this thinking has been carried out in the flagship communitarian periodical, *The Responsive Community*, which is edited by Amitai Etzioni and includes contributions by an eclectic group of philosophers, social scientists, and public policy makers. Etzioni is also the director of a think-tank, *Institute for Communitarian Policy Studies*, that produces working

papers and advises government officials in Washington.[36]

Such political communitarians blame both the left and the right for our current malaise.[37] The political left is chastised not just for supporting welfare rights economically unsustainable in an era of slow growth and aging populations, but also for shifting power away from local communities and democratic institutions and towards centralized bureaucratic structures better equipped to administer the fair and equal distribution of benefits, thus leading to a growing sense of powerlessness and alienation from the political process. Moreover, the modern welfare state with its universalizing logic of rights and entitlements has undermined family and social ties in civil society by rendering superfluous obligations to communities, by actively discouraging private efforts to help others (eg, union rules and strict regulations in Sweden prevent parents from participating voluntarily in the governance of day care centers to which they send their children), and even by providing incentives that discourage the formation of families (eg, welfare payments are cut off in most American states if a recipient marries a working person) and encourage the break-up of families (eg, no-fault divorce in the US is often financially rewarding for the non custodial parent, usually the father).

Libertarian solutions favored by the political right have contributed even more directly to the erosion of social responsibilities and valued forms of communal life, particularly in the UK and the US. Far from producing beneficial communal consequences, the invisible hand of unregulated free-market capitalism undermines the family (eg, few corporations provide enough leave to parents of newborn children), disrupts local communities (eg, following plant closings or the shifting of corporate headquarters), and corrupts the political process (eg, since the mid-seventies special economic interests in the US have gained more power by drawing on political action committees to fund political representatives, with the consequence that representatives dependent on PAC money for their political survival no longer represent the community at large). Moreover, the valorization of greed in the Thatcher/Reagan era justified the extension of instrumental considerations governing relationships in the marketplace into spheres previously informed by a sense of uncalculated reciprocity and civil obligation. This trend has been reinforced by increasing globalization, which pressures states into conforming to the dictates of the international marketplace.

More specifically in the American context, communitarian thinkers such as Mary Ann Glendon indict a new version of rights discourse that has achieved dominance of late.[38] Whereas the assertion of rights was once confined to matters of essential human interest, a strident rights rhetoric has colonized contemporary political discourse, thus leaving little room for reasoned discussion and compromise, justifying the neglect of social responsibilities without which a society could not function, and ultimately weakening all appeals to rights by devaluing the really important ones.

To remedy this imbalance between rights and responsibilities in the US, political communitarians propose a moratorium on the manufacture of new rights and changes to our habits of the heart away from exclusive focus on personal fulfillment and towards concern with bolstering families, schools, neighborhoods, and national political life, changes to be supported by certain public policies. Notice that this proposal takes for granted basic civil and political liberties already in place, thus alleviating the

concern that communitarians are embarking on a slippery slope to authoritarianism. Still, there may be a concern that marginalized groups demanding new rights, eg, homosexual couples seeking the right to legally sanctioned marriage, will be paying the price for the excesses of others if the communitarian proposal to declare a moratorium on the minting of new rights is put into effect.

More serious from the standpoint of those generally sympathetic to communitarian aspirations, however, is the question of what exactly this has to do with community. For one thing, Etzioni himself seeks to justify his policies with reference to need to maintain a balance between social *order* and freedom,[39] as opposed to appealing to the importance of community. But there is nothing distinctively communitarian about the preoccupation with social order; both liberals such as John Stuart Mill and Confucian conservatives affirm the need for order. And when the term community is employed by political communitarians, it seems to mean anything they want it to mean. Worse, as Elizabeth Frazer has argued, it has often been used to justify hierarchical arrangements and delegitimize areas of conflict and contestation in modern societies.[40]

Still, it is possible to make sense of the term community as a normative ideal.[41] Communitarians begin by positing a need to experience our lives as bound up with the good of the communities out of which our identity has been constituted. This excludes contingent attachments such as golf-club memberships, that do not usually bear on ones sense of identity and well-being (the co-authors of *Habits of the Heart*[42] employ the term ‘lifestyle enclaves’ to describe these attachments). Unlike pre-modern defenders of *Gemeinshaft*, however, it is assumed that there are *many* valued forms of communal life in the modern world. So the distinctive communitarian political project is to identify valued forms of community and to devise policies designed to protect and promote them, without sacrificing too much freedom. Typically, communitarians would invoke the following types of communities:

1. Communities of place, or communities based on geographical location. This is perhaps the most common meaning associated with the word community. In this sense, community is linked to locality, in the physical, geographical sense of a community that is located somewhere. It can refer to a small village or a big city. A community of place also has an affective component it refers to the place one calls home, often the place where one is born and bred and the place where one would like to end ones days even if home is left as an adult. At the very least, communitarians posit an interest in identifying with familiar surroundings.

   In terms of political implications, it means that, for example, political authorities ought to consider the existent character of the local community when considering plans for development (Jane Jacobs famously documented the negative effects of razing, instead of renovating, run-down tenements that are replaced by functionally adequate but characterless low-income housing blocs[43]). Other suggestions to protect communities of place include: granting community councils veto power over building projects that fail to respect existent architectural styles; implementing laws regulating plant closures so as to protect local communities from the effects of rapid capital mobility and sudden industrial change; promoting local-ownership of corporations;[44] and imposing restrictions on large-scale discount outlets such as Wal-Mart that

threaten to displace small, fragmented, and diverse family and locally owned stores.[45]

2. Communities of memory, or groups of strangers who share a morally-significant history. This term first employed by the co-authors of *Habits of the Heart* refers to imagined communities that have a shared history going back several generations. Besides tying us to the past, such communities turn us towards the future members strive to realize the ideals and aspirations embedded in past experiences of those communities, seeing their efforts as being, in part, contributions to a common good. They provide a source of meaning and hope in peoples lives. Typical examples include the nation and language-based ethnocultural groups.

   In Western liberal democracies, this typically translates into various nation-building exercises meant to nourish the bonds of commonality that tie people to their nations, such as national service and national history lessons in school textbooks. Self-described republicans such as Michael Sandel place special emphasis upon the national political community and argue for measures that increase civic engagement and public-spiritedness.[46] However, there is increased recognition of the multi-national nature of contemporary states, and modern Western states must also try to make room for the political rights of minority groups. These political measures have been widely discussed in the recent literature on nationalism, citizenship, and multiculturalism.[47]
3. Psychological communities, or communities of face-to-face personal interaction governed by sentiments of trust, co-operation, and altruism. This refers to a group of persons who participate in common activity and experience a psychological sense of togetherness as shared ends are sought. Such communities, based on face-to-face interaction, are governed by sentiments of trust, cooperation, and altruism in the sense that constituent members have the good of the community in mind and act on behalf of the communitys interest. They differ from communities of place by not being necessarily defined by locality and proximity. The differ from communities of memory in the sense that they are more real, they are typically based on face to face social interaction at one point in time and consequently tend to be restricted in size.[48] The family is the prototypical example. Other examples include small-scale work or school settings founded on trust and social cooperation.

   Communitarians tend to favor policies designed to protect and promote ties to the family and family-like groups. This would include such measures as encouraging marriage and increasing the difficulty of legal marriage dissolution. These policies are supported by empirical evidence that points to the psychological and social benefits of marriage.[49] Communitarians also favor political legislation that can help to restructure education in such a way that peoples deepest needs in membership and participation in psychological communities are tapped at a young age. The primary school system in Japan, where students learn about group cooperation and benefits and rewards are assigned to the classroom as a whole rather than to individual students, could be a useful model.[50]

What makes the political project of communitarianism distinctive is that it involves the promotion all three forms of valued communal life. This leads, however, to the worry that seeking the goods of various communities may conflict in practice. Etzioni, for example, argues for a whole host of pro-family

measures: mothers and fathers should devote more time and energy to parenting (in view of the fact that most childcare centers do a poor job of caring for children), labor unions and employers ought to make it easier for parents to work at home, and the government should force corporations to provide six months of paid leave and another year of unpaid leave.[51] The combined effect of these changes of the heart and public policies in all likelihood would be to make citizens into largely private, family-centered persons.

Yet Etzioni also argues that the American political system is corrupt to the core, concluding that only extensive involvement in public affairs by virtuous citizens can remedy the situation: ‘once citizens are informed, they must make it their civic duty to *organize others* locally, regionally, and nationally to act on their understanding of what it takes to clean up public life in America.’[52] But few can afford sufficient time and energy to devote themselves fully to both family life and public affairs, and favoring one ideal is most likely to erode the other. Surely it is no coincidence that republican America in Jeffersons day relied on active, public-spirited male citizens largely freed from family responsibilities. Conversely, societies composed of persons leading rich and fulfilling family lives (such as contemporary Singapore) tend to be ruled by paternalistic despots who can rely on a compliant, politically apathetic populace.

Communitarians who advocate both increased commitment to public affairs and strengthened ties to the workplace (to the point that it becomes a psychological community) also face the problem of conflicting commitments. Michael Sandel, for example, speaks favorable of ‘proud craftsmen’ in the Jacksonian ear and of Louis Brandeiss idea of ‘industrial democracy, in which workers participated in management and shared responsibilities for running the business.’[53] Identification with the workplace and industrial democracy are said to improve workers civic capacities, but that may not be the case. In the same way that extensive involvement in family life can conflict with commitments to public life, few persons will have sufficient time and energy for extensive participation in both workplace and public affairs. Recall that the republican society of ancient Athens relied on active, public-spirited males freed from the need to work (slaves did most of the drudge labor).

It is also worth noting that devotion to the workplace can undermine family life. As Professor Tatsuo Inoue of Tokyo University argues, Japanese-style communitarianism strong communal identity based on the workplace, with extensive worker participation in management sometimes leads to *karoshi* (death from overwork) and frequently deprives workers of ‘the right to sit down at the dinner table with their families’.[54] Just as liberals (pace Ronald Dworkin) sometimes have to choose between ideals (eg, freedom and equality) that come into conflict with one another if a serious effort is made to realize any one of them fully, so communitarians may have to make some hard choices between valued forms of communal life.

Still, there may be some actual or potential win-win scenarios cases where promoting a particular form of communal life can promote, rather than undermine, other forms -- and political communitarians will of course favor change of this sort. For example, critics have objected to residential community associations, or walled communities, on the grounds that they undermine attachment to the polity at large and erode the social cohesion and trust needed to promote social justice and sustain the democratic

process.[55] Might it then be possible to reform urban planning so that people can nurture strong local communities without undermining attachment to the national community, perhaps even strengthening broader forms of public-spiritedness? Many practical suggestions along these lines have been raised. Architects and urban planners in the US known as the New Urbanists, for example, have proposed various measures to strengthen community building affordable housing, public transport, pedestrian focused environments, and public space as an integral part of neighborhoods that would not have the privatizing consequences of gated communities. The problem, as Gerald Frug points out, is that ‘virtually everything they want to do is now illegal. To promote the new urbanist version of urban design, cities would have revise municipal zoning laws and development policy from top to bottom.’[56] This points to the need for public policy recommendations explicitly designed to favor complementing forms of communal attachments.

Just as it would be wrong to assume that communitarian goals always conflict, so one should allow for the possibility that individual rights and communitarian goals can co-exist and complement each other.[57] In Singapore, for example, it can be argued that more secure democratic rights would have the effect of strengthening commitment to the common national good.[58] The Singapore government does not hide the fact that it makes life difficult for many who aim to enter the political arena on the side of opposition parties: Between 1971 and 1993, according to Attorney General Chan Sek Keong, eleven opposition politicians have been made bankrupt (and hence ineligible to run in elections).[59] Whether intended or not, such actions send an unpatriotic message to the community at large: Politics is a dangerous game for those who havent been specially anointed by the top leadership of the ruling party, so you should stick to your own private affairs. As Singaporean journalist Cherian George puts it, one can hardly blame people for ignoring their social and political obligations ‘when they hear so many cautionary tales: Of Singaporeans whose careers came to a premature end after they voiced dissent; of critics who found themselves under investigation; of individuals who were detained without trial even though they seemed not to pose any real threat; of tapped phones and opened letters’. The moral of these stories: In Singapore, better to mind your own business, make money, and leave politics to the politicians.’[60] Put positively, if the aim is to secure attachment to the community at large, then implementing genuinely competitive elections, including the freedom to run for the opposition without fear of retaliation,[61] is an essential first step.

The Singapore case, however, points to another dimension of the politics of community that brings us back to the communitarian defense of cultural particularism. Democratic reformers in Singapore typically think of democracy in terms of free and fair competitive elections what Western analysts often label minimal democracy. In Hong Kong, the situation is similar the aspiration to full democracy put forward by social critics turns out to mean (nothing more than) an elected legislature and Chief Executive. Put differently, it is quite striking that the republican tradition in communitarian thought with its vision of strong democracy supported by active, public-spirited citizens who participate in political decision-making and held shape the future direction of their society though political debate seems completely absent from political discourse in Singapore and Hong Kong, and perhaps East Asia more generally. Many East Asians are clamoring for secure democratic rights, but this doesnt translate into the demand that all citizens should be committed to politics on an ongoing basis or the view that, as David

Miller puts it, 'politics is indeed a necessary part of the good life'[62]. At one level, this can be explained by the fact that there are no equivalents of Aristotle and Jean-Jacques Rousseau in East Asian philosophy. It can also be argued that republican ideas fail to resonate because East Asians typically place more emphasis on other forms of communal life the family in particular has been important theme in Confucian ethical theory and practice (relative to Western philosophy). To the extent that different forms of communal life do conflict in practice, in short, it may the case that different cultures will draw the line in different places and they may be justified in doing so, if this conforms to the views shared by both defenders and critics of the political *status quo*.

Bibliography

- An-Naim, A., 1992, 'Toward a Cross-Cultural Approach to Defining International Standards of Human Rights: The Meaning of Cruel, Inhuman, or Degrading Treatment or Punishment', in *Human Rights in Cross-Cultural Perspectives: A Quest for Consensus*, A. An-Naim, (ed.), Philadelphia: University of Pennsylvania Press
- Avineri, S., and de-Shalit, A., (eds), 1992, *Communitarianism and Individualism*, Oxford: Clarendon Press
- Bell, D., 2000, *East Meets West: Human Rights and Democracy in East Asia*, Princeton: Princeton University Press
- ------, 1997, 'Communitarianism', in *The Blackwell Encyclopedic Dictionary of Business Ethics*, P. H. Werhane and R. E. Freeman, (eds.), Malden: Blackwell
- ------, 1997b, 'Review of Sandel 1996', in *The Responsive Community* (Spring), pp.61-68.
- ------, 1995, 'Residential Community Associations: Community or Disunity', *The Responsive Community* (Fall), pp.25-36.
- ------, 1995b, 'A Communitarian Critique of Authoritarianism', *Society* (July/August), pp.38-43
- ------, 1993, *Communitarianism and Its Critics*, Oxford: Clarendon Press
- Bellah, R., *et al.*, 1985, *Habits of the Heart*, Berkeley: University of California Press
- Beng-Huat, C., 1995, *Communitarian Ideology and Democracy in Singapore*, London: Routledge
- Benhabib, S., 1992, *Situating the Self: Gender, Community and Postmodernism in Contemporary Ethics*, Cambridge: Polity Press
- Berten, A., da Silveira, P., and Pourtois, H., (eds), 1997, *Liberaux et Communautariens*, Paris: PUF
- Caney, S., 1992, 'Liberalism and Communitarianism: A Misconceived Debate', *Political Studies* (June), pp.273-90.
- Chaihark, H., forthcoming, 'Constitutionalism, Confucian Civic Virtue, and Ritual Propriety', in *Confucianism for the Modern World*, D. Bell and H. Chaibong, (eds.), New York: Cambridge University Press, forthcoming.
- Chan, J., 1999, 'A Confucian Perspective on Human Rights for Contemporary China', in *The East Asian Challenge for Human Rights*, J. R. Bauer and D. Bell, (eds.), New York: Cambridge University Press
- D'Antonio, M., 1994, 'The High-Rise Village: Public Housing Creates a Community in Harlem', Washington: The Communitarian Network

- Doppelt, G, 1989, 'Is Rawls Kantian Liberalism Coherent and Defensible?', *Ethics* (July 1989), pp.820-21
- Dworkin, R., 1989, 'Liberal Community', *California Law Review*, 77
- Ehrenhalt, A., 1999, 'Community and the Corner Store: Retrieving Human-Scale Commerce', *The Responsive Community* (Fall), pp. 30-39.
- Etzioni, A., 2001, *The Monochrome Society*, Princeton: Princeton University Press
- ------, 1998, *The Essential Communitarian Reader*, Lanham: Rowman & Littlefield
- ------, 1996, *The New Golden Rule*, New York: Basic Books
- ------, 1995a, *New Communitarian Thinking*, Charlottesville: University of Virginia Press
- ------, 1995b, *Rights and the Common Good: The Communitarian Perspective*, New York: St. Martins Press
- ------, 1993, *The Spirit of Community*, New York: Crown Publishers
- Frazer, E., 1999, *The Problems of Communitarian Politics*, Oxford: Oxford University Press
- Frazer, E., and Lacey, N., 1993, *The Politics of Community: A feminist critique of the liberal-communitarian debate*, Hemel Hempstead: Harvester Wheatsheaf
- Fukuyama, F., 1992, *The End of History and the Last Man*, New York: Free Press
- Frug, G., 1999, *City Making: Building Communities Without Building Walls*, Princeton: Princeton University Press
- Glendon, M.-A., 1991, *Rights Talk: The Impoverishment of Political Discourse*, New York: The Free Press
- Gutmann, A. (ed.), 1992, *Multiculturalism and 'The Politics of Recognition'*, Princeton: Princeton University Press
- Inoue, T., 1993, 'The Poverty of Rights-Blind Communality: Looking Through the Window of Japan', *Brigham Young University Law Review* (January), p.534.
- Jacobs, J., 1965, *The Death and Life of American Cities*, New York: Random House
- Kymlicka, W., 1995, *Multicultural Citizenship*, Oxford: Clarendon Press
- ------, 1989, *Liberalism, Community and Culture*, Oxford: Clarendon Press
- MacIntyre, A., 1991, 'Letter', in *The Responsive Community*, Summer 1991
- ------, 1988, *Whose Justice? Which Rationality?*, Notre Dame: University of Notre Dame Press
- ------, 1984, *After Virtue*, Notre-Dame: University of Notre Dame Press, 2nd edition
- ------, 1978, *Against the Self-Images of the Age*, Notre Dame: University of Notre Dame Press
- Macedo, S., 2000, *Diversity and Distrust*, Cambridge: Harvard University Press
- ------, 1990, *Liberal Virtues: Citizenship, Virtue and Community in Liberal Constitutionalism*, Oxford: Clarendon Press
- Mason, A., 2000, *Community, Solidarity and Belonging: Levels of Community and their Normative Significance*, Cambridge: Cambridge University Press
- McKenzie, E., 1994, *Privatopia*, New Haven: Yale University Press
- Miller, D., 2000, *Citizenship and National Identity*, Cambridge: Polity Press
- Mulhall, S., and Swift, A., 1996, *Liberals and Communitarians*, Oxford: Blackwell, 2nd edition
- Qiang, L., 1998, *Ziyouzhuyi* (Liberalism), Beijing: *Zhonguo Shehui Kexue Chubanshe*
- Rawls, J., 1999, *The Law of Peoples; with The Idea of Public Reason Revisited*, Cambridge: Harvard University Press
- ------, 1993, *Political Liberalism*, New York: Columbia University Press

- ------, 1971, *A Theory of Justice*, Cambridge, Mass: Harvard University Press
- Rasmussen, D., (ed.), 1990, *Universalism vs. Communitarianism*, Cambridge: MIT Press
- Reid, T. R., 1999, *Confucius Lives Next Door*, New York: Random House
- Rosenblum, N., 1998, *Membership and Morals*, Princeton: Princeton University Press
- Sandel, M., 1998, *Liberalism and the Limits of Justice*, Cambridge: Cambridge University Press, 2nd edition
- ------, 1996, *Democracy's Discontent*, Cambridge: Harvard University Press
- ------, 1981, *Liberalism and the Limits of Justice*, Cambridge: Cambridge University Press
- Shuman, M. H., 1999, ‘Community Corporations: Engines for a New Place-Based Economics’, *The Responsive Community* (Summer), pp.48-57.
- Tamir, Y., 1993, *Liberal Nationalism*, Princeton: Princeton University Press
- Tams, H., 1998, *Communitarianism: A New Agenda for Politics and Citizenship*, Basingstocke: Macmillan
- Taylor, C., 1999, ‘Conditions of an Unforced Consensus on Human Rights’, in *The East Asian Challenge for Human Rights*, J. R. Bauer and D. Bell, (eds.), New York: Cambridge University Press
- ------, 1989, *Sources of the Self: The Making of the Modern Identity*, Cambridge: Cambridge University Press
- ------, 1985, *Philosophy and the Human Sciences: Philosophical Papers 2*, Cambridge: Cambridge University Press
- Waite, L., 1996, "Social Science Finds: ‘Marriage Matters’", *Twenty-First Century Series in Communitarian Studies*, Washington: Institute for Communitarian Policy Studies
- Walzer, M., 1994, *Thick and Thin*, Notre-Dame: University of Notre Dame Press
- ------, 1987, *Interpretation and Social Criticism*, Cambridge: Harvard University Press
- ------, 1983, *Spheres of Justice*, Oxford: Blackwell
- Young, I. M., 1990, *Justice and the Politics of Difference*, Princeton: Princeton University Press

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

citizenship | liberalism | nationalism

Daniel Bell
sadaniel@cityu.edu.hk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 3, 2001
Content last modified: October 3, 2001

Stanford Encyclopedia of Philosophy
Notes to Communitarianism

Notes

1. See the Bibliographic entry for Rawls 1971.

2. Both Taylor and Walzer identify themselves as liberals in Gutmann 1992. MacIntyre (1991) says 'In spite of rumors to the contrary, I am not and never have been a communitarian'. Sandel (1998) uses the label republican rather than communitarian.

3. For relevant references, see the bibliographies in Avineri and de-Shalit 1992, Bell 1993, Berten *et al*. 1997, Mulhall & Swift 1996, and Rasmussen 1990.

4. This essay draws on the threefold distinction in Bell 1993. For a similar threefold distinction, see Caney 1992. But for an expanded fivefold classification of arguments, see Mulhall & Swift 1996.

5. This is the language Rawls employs on the last page of the first edition of Rawls 1971.

6. See, e.g., Taylor 1985, ch.1; MacIntyre 1978, chs.18-22 and 1988, ch.1. See also the discussion in Benhabib 1992, pp.23-38, 89n4.

7. Walzer 1983, p.8. See also Young 1990, p.4.

8. Rawls 1993.

9. Rawls 1999.

10. Brian Barry, *Justice as Impartiality* (Oxford: Clarendon Press, 1995), p.3.

11. MacIntyre 1984.

12. Walzer 1983, p.313.

13. Fukuama 1992.

14. This section draws on the introduction to Bell 2000. Another challenge to Western-style liberal-

democracy has of course been mounted by Islamic civilization, though Islamic countries have not been as economically and politically successful (compared to East Asia) and therefore fail to pose as significant a challenge the claims of Western liberal-democrats that only capitalism and liberal democracy can cope with the requirements of modernity.

15. Quoted in the *International Herald Tribune*, 9-10 November 1991.

16. Quoted in Barbara Crossette, 'U.N. Survey Finds Rich-Poor Gap Widening', *New York Times*, 15 July 1996.

17. See Bell 2000, ch.1.

18. Interestingly, this moral outlook still seems to inform the practices of Asian immigrants to other societies. According to the *New York Times* (11 August 2001), fewer than one in five whites in the US help care or provide financial support for their parents, in-laws or other relatives, compared with 28% of African-Americans, 34% of Hispanic-Americans and 42% of Asian-Americans. Those who provide the most care also feel the most guilt that they are not doing enough. Almost three-quarters of Asian-Americans say they should do more for their parents, compared with two-thirds of Hispanics, slightly more than half the African-Americans and fewer than half the whites.

19. See Chaihark forthcoming. See also Bell 2000, ch.5.

20. Walzer 1987, p.24. See also Walzer 1994.

21. Taylor 1999.

22. Taylor 1999, p.124.

23. According to Abdullahi An-Naim (1992, p.34), however, the prerequisite conditions for the enforcement of this punishment are extremely difficult to realize in practice and are unlikely to materialize in any Muslim country in the foreseeable future.

24. 'Atomism', in Taylor 1985, p.200.

25. Taylor 1985, p.190.

26. This section draws on Acts I and III of Bell 1993.

27. See, e.g., Doppelt 1989.

28. See, e.g., Taylor 1989, esp. part I.

29. Kymlicka 1989, p.166.

30. Sandel 1981, p.179.

31. Will Kymlicka 1989, p. 52. See also Dworkin 1989, p. 489, and Macedo 1990, p. 247.

32. See, e.g., Frazer & Lacey 1993, pp.53-60.

33. See Taylor 1989, pp. 26-7.

34. The liberal-communitarian debate over the self has been prominent in non-anglophone publications, however, see e.g., Qiang 1998, chs.5-6. It is also interesting to note that adherents of Confucianism have recently advanced arguments against liberal foundations similar to the claims of 1980s communitarians, also with the apparent aim of undermining the foundations of liberal rights. Joseph Chan (1999) reviews these arguments and finds them wanting, with the proviso that Confucianisms understanding of the scope and justification of rights would differ from Western, rights-based perspectives.

35. For book-length treatments of communitarian politics in the US, see, Etzioni 1993, 1996, and 2001. For a book that derives largely from the UK context see Tams 1998. See also Etzioni's edited books, 1995a, 1995b, and 1998.

36. Both Democrats and Republicans seem to be receptive to communitarian political ideas. The political theorist William Galston, a co-editor of *The Responsive Community* and author of *Liberal Purposes* (Cambridge: Cambridge University Press, 1991), was President Clinton's Domestic Policy Adviser. President Bush has recently unveiled a four-year Communities of Character project that was developed following consultations with Etzioni (*Washington Post*, 29 July 2001). See also Dana Milbank, ‘Is Bush a Communitarian?’, *The Responsive Community* (Spring 2001), pp.4-7.

37. This section draws on Bell 1997.

38. Glendon 1991.

39. See especially Etzioni 1996.

40. Frazer 1999.

41. This section draws on Acts III to V Bell 1993. See also Mason 2000. Mason usefully distinguishes between different levels and kinds of communities, though one can question his argument that the ideal

of global *community* is coherent in principle and useful in practice (in my view, communities are particularistic in nature and presume an inside/outside distinction, and even if this ideal is coherent it is unclear to what extent the ideal of community does much work for defenders of universal liberal principles and global institutions).

42. Bellah *et al*. 1985.

43. Jacobs 1965.

44. See Shuman 1999.

45. See Ehrenhalt 1999.

46. Sandel 1996.

47. See, e.g., Kymlicka 1995, Macedo 2000, and Tamir 1993.

48. Though conceptions of the family can also include an imagined component for example, the widespread practice of ancestor worship in East Asian societies with a Confucian heritage suggests that (deceased) ancestors are considered as ongoing participants in the good of the family.

49. See Waite 1996.

50. See Reid 1999, esp. ch.5.

51. See Etzioni 1993, ch.2 and Etzioni 1996, ch.6.

52. Etzioni 1993, p.244.

53. Sandel 1996, pp.170, 213. See also Bell 1997b.

54. Inoue 1993.

55. See McKenzie 1994, and Bell 1995. For a contrasting account, see Rosenblum 1998, ch.4.

56. Frug 1999, pp.152-53. For an account of an actual example of diverse, mixed-income and mixed race urban housing project that contrasts with homogenous, upper-class walled communities, see D'Antonio 1994.

57. See Bell 1995b, for a critique of Etzioni's apparent assumption that rights and particularistic communal commitments always conflict.

58. This argument is developed at length in Bell 2000, ch.4. In the same vein, see Beng-Huat 1995, esp. ch.9.

59. In the latest case (mid-2001), J.B. Jeyaratnam has been declared bankrupt and has had to forfeit his Parliamentary seat.

60. *Straits Times* (Singapore), 11 July 1993.

61. The Singapore state, it must be said, resorts to endlessly creative tactics to curb opposition attempts to reach out to the electorate and communicate alternative ideas and policies. The opposition Singapore Democratic Party was recently informed by the Singapore police that it needed to engage 13 officers for crowd control purposes for a planned national day rally on 26 August 2001, amounting to several thousand dollars. One wonders if the ruling Peoples Action Party needs to pay for its own security for its rallies (not to mention the question of who pays for the undercover officers at opposition rallies).

62. Miller 2000.

Daniel Bell
sadaniel@cityu.edu.hk

First published: October 3, 2001
Content last modified: October 3, 2001

Nationalism

The term ‘nationalism’ is generally used to describe two phenomena: (1) the attitude that the members of a nation have when they care about their national identity, and (2) the actions that the members of a nation take when seeking to achieve (or sustain) some form of political sovereignty. (1) raises questions about the concept of nation (or national identity), which is often defined in terms of common origin, ethnicity, or cultural ties, and while an individual's membership in a nation is often regarded as involuntary, it is sometimes regarded as voluntary. (2) raises questions about the whether sovereignty must be understood as the acquisition of full statehood with complete authority for domestic and international affairs, or whether something less is required.

It is traditional, therefore, to distinguish nations from states -- whereas a nation often consists of an ethnic or cultural community, a state is a political entity with a high degree of sovereignty. While many states are nations in some sense, there are many nations which are not fully sovereign states. As an example, the Native American Iroquois constitute a nation but not a state, since they do not possess the requisite political authority over their internal or external affairs. If the members of the Iroquois nation were to strive to form a sovereign state in the effort to preserve their identity as a people, they would be exhibiting a kind of nationalism.

Nationalism has long been ignored as a topic in political philosophy, written off as a relic from bygone times. It has only recently come into the focus of philosophical debate, partly in consequence of rather spectacular and troubling nationalist clashes, like those in Rwanda, former Yugoslavia and former Soviet republics. The surge of nationalism usually presents a morally ambivalent, and for this reason often fascinating, picture. ‘National awakening’ and struggle for political independence are often both heroic and inhumanly cruel; the formation of a recognizably national state often responds to deep popular sentiment, but can and does sometimes bring in its wake inhuman consequences, including violent expulsion and ‘cleansing’ of non-nationals, all the way to organized mass murder. The moral debate on nationalism reflects a deep moral tension between solidarity with oppressed ethnic national groups on the one hand and the repulsion people feel in the face of crimes perpetrated in the name of nationalism on the other. Moreover, the issue of nationalism points to the wider domain of problems, having to do with the treatment of ethnic and cultural differences within democratic polity, which are arguably among the most pressing problems of contemporary political theory.

In this entry we shall first present conceptual issues of definition and classification (Sections 1 and 2), and then the arguments put forward in the moral debate (Section 3), dedicating more space to the arguments in favor of nationalism, than to those against it, in order to give the philosophical nationalist a

proper hearing.

---

1. What is a Nation?

1.1 The Basic Concept of Nationalism

Although the term ‘nationalism’ has a variety of meanings, it centrally encompasses the two phenomena noted at the outset: (1) the attitude that the members of a nation have when they care about their identity as members of that nation and (2) the actions that the members of a nation take in seeking to achieve (or sustain) some form of political sovereignty. (See, for example, Nielsen 1998-9, 9.) Each of these aspects requires elaboration. (1) raises questions about the concept of nation or national identity, about what it is to belong to a nation, and about the degree of care about a nation required. Nations and national identity may be defined in terms of common origin, ethnicity, or cultural ties, and while an individual's membership in the nation is often regarded as involuntary, it is sometimes regarded as voluntary. The degree of care for one's nation that is required by nationalists is often, but not always, taken to be very high: on such views one's nation's claims have supremacy in competition with rival contenders for authority and loyalty (see Berlin 1979, Smith 1991, Levy 2000).

(2) raises questions about the whether sovereignty entails the acquisition of full statehood with complete authority for domestic and international affairs, or whether something less than statehood would suffice. Although sovereignty is often taken to mean full statehood (Gellner 1983, ch. 1), more recently possible exceptions have been recognized (Miller 1992, 87).

Despite these definitional worries, there is a fair amount of agreement about what is historically the most typical, paradigmatic form of nationalism. It is the one which features the supremacy of the nation's claims over other claims to individual allegiance, and which features full sovereignty as the persistent aim of its political program. Nationalists often see the state as a political unit centrally `belonging' to one ethnic-cultural group, and actively charged with protecting and promulgating its traditions. This form is exemplified by the classical, ‘revivalist’ nationalism, most prominent in the 19th century in Europe and Latin America. This classical nationalism later spread across the world, and in present days still marks many contemporary nationalisms.

Thus, in its general form, the concept of nationalism concerns the relationship between the ethno-cultural domain (featuring etho-cultural groups or ‘nations’) and the domain of political organization. In our preliminary analysis of the concept, we noted that nationalism involves the attitude that the members of a nation have when they care about their national identity. We can divide the questions raised above in connection with this analysis into two sorts. First, the descriptive ones:

> (1a) What is a nation and national identity?
> (1b) What is it to belong to a nation?
> (1c) What is the nature of a pro-national attitude?
> (1d) Is membership in a nation voluntary or non-voluntary?

Second, the normative ones:

> (1e) Is the attitude of caring about national identity always appropriate?
> (1f) How much should one to care?

In the remainder of this section, we discuss the descriptive questions (1a) -- (1d). We leave questions (1e) and (1f) for Section 3, which concerns the moral debate.

1.2 The Concept of Nation

If one wants to enjoin people to struggle for their national interests, one must have some idea about what a nation is and what it is to belong to a nation. So, in order to formulate and ground their evaluations, claims, and directives for action, pro-nationalist thinkers have been elaborating theories of ethnicity, culture, nation and state. Their opponents have in their turn challenged these theories. Now some presuppositions about ethnic groups and nations are essential for the nationalist, while others are theoretical elaborations designed to support the essential ones. The former concern the definition and status of the target or social group which is the beneficiary of the nationalist program, variously called ‘nation’, ‘ethno-nation’ or ‘ethnic-group’. Since nationalism is particularly prominent with groups that don't yet have a state, one can't define belonging to a nation in terms of belonging to a state.

Indeed, purely ‘civic’ loyalties are often put into the separate category under the title ‘patriotism’, or ‘constitutional patriotism’ (Habermas 1996). This yields a spectrum of definitions for the concept of

‘nation’, with lots of intermediate positions. On one end of the spectrum, we find a small but distinguished band of theorists, including E. Renan (1882) and M. Weber (1970). Recall that on their definition, a nation is any group of people who voluntarily aspire to a common political state-like organization. Recall also that if such a group of people succeeds in forming a state, the loyalties of the group members would be ‘civic’ (as opposed to ‘ethnic’) in nature. The other end of the spectrum is more typical, for nationalist claims are focused upon a community of common origin, language, tradition and culture, membership in which is non-voluntary. Thus, on this more typical view, an ethno-nation is a community of origin and culture, including prominently a language and customs.

The distinction between these two distinct concepts of nationhood is related (although not identical) to the one drawn by older schools of social and political science between ‘civic’ and ‘ethnic’ nationalism, the first being allegedly Western European and the later more Central and Eastern European originating in Germany (a very prominent proponent of the view is H. Kohn 1965). Philosophical discussions centered around nationalism tend to concern the ethnic-cultural variants only and this habit will be followed here. A group aspiring to nationhood on this basis will be called here an ‘ethno-nation’ in order to underscore its ethno-cultural rather than purely civic underpinnings. For the ethno-(cultural) nationalist it is one’s ethnic-cultural background which determines one’s membership in the community. One can't choose to be a member; instead, membership depends on the accident of origin and early socializing. It may be, however, that commonality of origin is an almost mythical notion for most contemporary candidate groups: ethnic groups have been mixing for millennia.

Therefore, sophisticated pro-nationalists tend to stress cultural membership only, and speak of ‘nationality’, omitting the ‘ethno-’ part (Miller 1992, 87; Tamir 1993). M. Seymour in his recent proposal of a ‘socio-cultural definition’ adds a political dimension to the purely cultural one. A nation is a cultural group, possibly but not necessarily united by a common descent, endowed with some kind of civic ties (Seymour 2000). This is the kind of definition that would be accepted by most parties in the debate today. So defined, nation is a somewhat mixed, both ethno-cultural and civic category, but still closer to the purely ethno-cultural than to the purely civic extreme.

The wider descriptive underpinning of nationalist claims has varied across the last two centuries. The early German discussions feature talk about ‘the spirit of people’, while somewhat later ones, mainly of French extraction, about ‘collective mentality’, ascribing to them specific and significant causal powers. A later descendent of this notion is the idea of a ‘national character’ peculiar to each nation, which partly survives today under the guise of national ‘forms of life’ and of feeling (Margalit 1997, see below). For almost a century, up to the end of the Second World War, it was customary to link nationalistic views with the metaphor of a society being something ‘organic’. Isaiah Berlin, writing as late as the early seventies, proposed as a part of his definition of nationalism that it consists of the conviction that people belong to a particular human group, and that "...the characters of the individuals who compose the group are shaped by, and cannot be understood apart from, those of the group ..." (first published in 1972, reprinted in Berlin, 1979: 341). The nationalist claims, according to Berlin, that "the pattern of life in a society is similar to that of a biological organism" (*ibid*.), and that the needs of this ‘organism’ determine the supreme goal of all of its members.

Most contemporary defenders of nationalism, especially philosophers, avoid such language. The metaphor of organism and talk about character have been replaced by one master-metaphor -- that of national identity. It is centered upon cultural membership, and used both for the identity of a group and for the socially-based identity of its members, e.g., the national identity of George in so far as he is English or British. Various authors unpack the metaphor in various ways -- some stress involuntary membership in the community, others the strength with which one identifies with the community, yet others link it to the personal identity of each member of the community. Addressing these issues, the nationally minded philosophers, like MacIntyre (1994), Taylor (1989), Seymour and others have significantly contributed to introducing and maintaining important topics such as community, membership, tradition and social identity in contemporary philosophical debate.

Let us now turn to the issue of the origin and 'authenticity' of ethno-cultural groups or ethno-nations. In social and political science one usually distinguishes two kinds of views. The first are the 'primordialist' views. According to them, actual ethno-cultural nations have either existed 'since time immemorial' (an extreme, somewhat caricatured version, corresponding to nineteenth century nationalist rhetoric), or at least for a long time during the pre-modern period (a more moderate version championed by A. Smith (1991)). The second are the modernist views, placing the origin of nations in modern times. They can be further classified according to their answer to a further question: how real is the ethno-cultural nation? The modernist realist view is that nations are real but distinctly modern creations, instrumental in the genesis of capitalism (Gellner 1983, and Hobsbawn 1990). On the opposite side of the fence one finds anti-realist views. According to one such view nations are merely 'imagined' but somehow still powerful entities; what is meant is that belief in them holds sway over the believers (Anderson 1965). The extreme anti-realist view claims that they are pure 'constructions'. These divergent views seem to support rather divergent moral claims involving the concept of a nation.

Indeed, older authors -- from great thinkers like Herder and Bauer, to the propagandists who followed their footsteps -- have been at great pains to ground normative claims upon firm ontological realism about nations: nations are real, *bona fide* entities. However, the contemporary moral debate has tried to diminish the importance of the imagined/real divide. Prominent contemporary philosophers have claimed that normative-evaluative nationalist claims are compatible with the 'imagined' nature of nation. (See, for instance, Miller 1992, Tamir 1993, and MacCormick 1982.) They point out that common imaginings can tie people together, and that actual interaction resulting from togetherness can engender important moral obligations.

Let us now turn to question (1c), concerning the nature of pro-national attitudes. The explanatory issue that has interested political and social scientists concerns ethno-nationalist sentiment, the paradigm case of a pro-national attitude. Is it as irrational, romantic and indifferent to self-interest as it might seem on the surface? The issue has divided authors into two camps -- those who see nationalism as basically irrational and those who try to explain it as being rational at least in some sense. Authors in the first camp propose various explanations of why people assent to irrational views. Some say, critically, that nationalism is based on 'false consciousness'. But where does such false consciousness come from? The most simplistic view is that it is a result of direct manipulation of 'masses' by 'elites'. On the opposite side, the famous critic of nationalism, E. Kedourie (1960) sees this irrationality as being spontaneous.

Authors relying upon the Marxist tradition offer various deeper explanations. To mention one, the French structuralist E. Balibar sees it as a result of 'production' of ideology effectuated by mechanisms which have nothing to do with spontaneous credulity of individuals, but with impersonal, structural social factors (Balibar and Wallerstein 1992).

Consider next the other camp of those who see nationalistic sentiments as being rational at least in some weak sense. Some authors claim that it is often rational for individuals to become nationalists (Hardin 1985). On the one hand, identification and cohesion within the ethno-national group has to do with inter-group cooperation, and cooperation is easier for those who are part of the same ethno-national group. To take an example of ethnic ties in a multiethnic state, a Vietnamese newcomer to the States will do well to rely on his co-nationals: common language, customs and expectations might help him a lot in finding his way in new surroundings. Once the ties are established and he has become part of a network, it is rational to go on cooperating, and ethnic sentiment does secure the trust and the firm bond needed for smooth cooperation. A further issue is when it is rational to switch sides; to stay with our example, when does it become profitable for our Vietnamese to develop an all-American patriotism. This has received a detailed elaboration in Laitin (1998), who uses material from the ex-Soviet Union. On the other hand, there is conflict between various ethno-nations which leads to non-cooperation with outsiders. Can one rationally explain the extremes of ethno-national conflict? Authors like Hardin propose to do it in terms of a general view of when a hostile behavior is rational: most typically, if you have no reason to trust someone, it is reasonable to take precautions against him. If both sides take precautions, however, each will tend to see the other as being seriously inimical. It then becomes rational to start treating the other as an enemy. Mere suspicion can thus lead by small, individually rational steps, to a situation of conflict. (Such negative development is often presented as a variant of the so-called Prisoner's Dilemma.) Now, it is relatively easy to spot the circumstances in which this general pattern applies to national solidarities and conflicts. The line of thought just sketched is often called 'rational choice approach'. It has enabled the application of conceptual tools from game-theoretic and economic theories of cooperative and non-cooperative behavior for explanation of ethno-nationalism.

It is worth mentioning, however, that the individualist rational-choice approach, centered upon personal rationality, has serious competitors. A tradition in social psychology, initiated by Tajfel (1981), shows that individuals may identify with a randomly selected group, even when membership in the group brings no tangible rewards. Does rationality of any kind underlie this tendency of identification? Some authors (Sober & Wilson 1998) answer in the affirmative. They propose that it is a non-personal, evolutionary rationality: individuals who develop a sentiment of identification and sense of belonging end up better off in the evolutionary race, and that is why we have inherited such propensities. The initial sentiments were reserved for one's own kin, thus supporting the spreading of one's own genes. Cultural evolution has taken over the mechanisms of identification that initially developed within biological evolution. As a result, we project the sentiment originally reserved for kinship to our cultural group. Further, detailed explanations in such socio-biological perspective differ a lot among themselves, and constitute a wide research program.

Finally, concerning question (1d), a nation is typically seen as essentially a non-voluntary community to which one belongs by birth and early nurture, whereby the relation of belonging is somehow enhanced

and perhaps taken to a higher level, becoming more conscious and more complete by one's additional endorsement. (There are exceptions to this basically non-voluntaristic view, for instance, theoretical nationalists who accept voluntary changes of nationality. See also E. Renan's (1882, 19) famous definition of a nation as constituted by 'everyday plebiscite'.)

2. Varieties of Nationalism

2.1 Concepts of Nationalism: Strict and Wide

Recall again our basic analysis, namely, that nationalism involves (1) the attitude that the members of a nation have when they care about their national identity, and (2) the actions that the members of a nation take when seeking to achieve (or sustain) some form of political sovereignty. The politically central point is (2), the actions enjoined by the nationalist. This point raises three important questions:

(2a) Does political sovereignty require statehood or something weaker?
(2b) What actions are appropriate to bringing sovereignty about and maintain it?
(2c) Under what conditions is it appropriate to take actions of this kind?

To these we now turn.

The classic (and conservative) answer to (2a) is that political sovereignty requires statehood, while the more liberal answer is that some form of political autonomy is required, though not necessarily full statehood. Consider the former, classic answer to (2a).

On this view, political sovereignty requires a state 'rightfully owned' by the ethno-nation (Oldenquist, 1997, who credits the expression to the writer C. Milosz). However, this classic form of nationalism is not only concerned with the creation of a state but also with its maintenance and strengthening. Once the state is there, further options are opened for nationalists. They sometimes promote claims for its expansion (even at the cost of wars) and sometimes opt for isolationist policies. The expansion is often justified by appeal to the unfinished business of bringing literally *all* members of the nation under one state, and sometimes justified simply as an interest the nation has in gaining more territory and resources. In doing this they often imply specific answers to (2b) and (2c), i.e., that in national independence struggle the use of force is almost always a legitimate means for bringing about sovereignty. As for maintenance of sovereignty by peaceful and merely ideological means, political nationalism is closely tied to nationalism in culture. The latter insists upon the preservation and transmission of a given culture, more accurately, of recognizably ethno-national traits of the culture in its pure form, dedicating artistic creation, education and research to this goal. Of course, the ethno-national traits can be actual or invented, partly or fully so. Again, in the classical variant the relevant norm claims that one has both a right and an obligation ('a sacred duty') to promote such traditions. Its force is that of a trump that wins over other interests and even over human rights (and this trump is often needed in order to carry on

national independence struggle).

As a consequence, this classic form of nationalism has something to say about the level of attitudes as well. It answers question (1e) affirmatively, and sees caring for one's nation as a fundamental duty of each of its members. It is also prone to give, in its answer to (1f), an unlimited scope. We can now summarize this classic form of nationalism as follows:

> **Classical nationalism** is the political program that seeks the creation and maintenance of a fully sovereign state owned by a given ethno-national group ('people' or 'nation') and that sees the creation and maintenance of this state as a primary duty of each member of the group. Starting from the assumption that the appropriate (or 'natural') unit of culture is ethno-nation, classical nationalism involves the claim that a primary duty of each member is to abide in cultural matters by one's recognizably ethno-national culture.

Classical nationalists are usually quite watchful about the kind of culture they protect and promote, and about the kind of attitude people have to their nation-state. This watchful attitude carries some potential dangers -- many elements of a given culture which are universalist or simply not recognizably national might, and will sometimes, fall prey to such nationalist enthusiasms. Classical nationalism in everyday life puts various additional demands on individuals, from buying more expensive home-produced goods in preference to the cheaper imported ones, to procreating as many future members of the nation as one can manage. (See Yuval-Davies 1997.)

The more liberal answer to question (2a), that some form of political automony less than full statehood is required to achieve nationalist goals, yields more moderate forms of nationalism. Indeed, the philosophical discussion has shifted to these moderate or even ultra-moderate forms, and most philosophers who describe themselves as nationalists propose very moderate nationalist programs. These more moderate forms involve weaker claims than the classic form, along a variety of dimensions. Indeed, a wider concept of nationalism is needed to accomodate the wide variety more moderate forms. Let us therefore define:

> **Nationalism in the wide sense** is any complex of attitudes, claims and directives for action which ascribe a fundamental political, moral and cultural value to nation and nationality and which produce obligations (for individual members of the nation, and for any involved third parties, individual or collective) on the basis of this ascribed value.

Different forms of nationalism in this larger sense result when we vary the concept of nation involved, vary the ground and degree of its value, vary the scope of claims for action and of the prescribed obligations. (This wider concept of nationalism can also be applied to other cases not covered by classical nationalism, for instance, the hypothetical pre-state political forms that an ethnic identity might take). Moderate nationalism is a form of nationalism in the wider sense which is less demanding than classical nationalism. It sometimes goes under the name of 'patriotism'. (A different usage, though, reserves 'patriotism' for valuing the civic community and loyalty to the state, in contrast to nationalism,

which is centered around ethnic-cultural communities.) The various forms of nationalism most relevant for philosophy are those that influence the moral standing of claims for action and of recommended nationalist practices. The elaborate philosophical views put forward in favor of nationalism will be referred to here to as ‘theoretical nationalism’, the adjective serving to distinguish such views from the less sophisticated and more practical nationalist discourse. The central theoretical nationalist evaluative claims can usefully be put on the map of possible positions within political theory in the following somewhat simplified and schematic way.

Nationalistic claims featuring the centrality of nation for political action provide an answer to two crucial general questions. First, is there one kind of large social group (smaller than the whole of mankind) that is morally of central importance or not? The nationalist answer is that there is just one, namely, the nation. When an ultimate choice is to be made, nation has priority. (This answer is implied by rather standard definitions offered by Berlin and Smith.) Second, what is the ground of obligation that the individual has to the morally central group? Is it voluntary or involuntary membership in the group? The typical contemporary nationalist thinker answers by the latter, while admitting that voluntary endorsement of one’s national identity is a morally important achievement. On the philosophical map, the pro-nationalist normative tastes fit nicely with the communitarian stance in general: most pro-nationalist philosophers are communitarians who choose nation as the preferred community (in contrast to those of their fellow-communitarians who prefer more far-ranging communities, such as those defined by global religious traditions). However, some recent writers, e.g., Kymlicka (2001), who describe themselves as liberal nationalists, reject the communitarian underpinning.

2.2 Moral Claims: The Centrality of Nation

We now prepare the ground for the discussion of the normative dimension of nationalism. We shall first describe the very heart of the nationalist program, i.e., sketch and classify the typical normative and evaluative nationalist claims. These claims can be seen as answers to the normative subset of our initial questions about (1) pro-national attitudes and (2) actions.

The claims thus recommend various courses of action, centrally those meant to secure and sustain the political organization -- preferably a state -- for the given ethno-cultural national community, thereby making more specific the answers to our normative questions (1e), (1f), (2b) and (2c). Further, they enjoin the members of the community to promulgate recognizable ethno-cultural values, artifacts, and traditions as central features of the cultural life within such a state. Finally, we shall discuss various lines of pro-nationalist thought that have been put forward in defense of these claims. Let us begin with the claims which concern the furthering of the national state and culture. They are proposed by the nationalist as a guide and a norm of conduct. Philosophically the most important variations concern three aspects of such normative claims. They are as follows:

(i) *The normative nature and strength of the claim*. Does it promote just a right (say, to having and maintaining a form of political self-government, preferably and typically a state, or having the cultural life centered upon recognizably ethno-national culture), or a moral obligation (to get and maintain one),

or a legal and political obligation? These differences can be described in professional terminology as variations in the ‘non-comparative deontic status’ of central claims. The strongest claim is typical for the classical nationalism: its norms are both moral imperatives and, once the nation-state is in place, legally enforceable obligations in regard to all parties concerned, including the individual members of the ethno-nation. A weaker, but still quite demanding version speaks only of moral obligation (‘sacred duty’). A more liberal version is satisfied simply with a right to having a state that would be ‘owned’ by the ethno-nation.

(ii) *The strength of the nationalistic claim in relation to various external interests and rights*. To give a real example, is the use of the domestic language so important that even international conferences should be held in it, at the cost of losing the most interesting participants from abroad? Since the force of the nationalist claim is here being weighed against the force of other claims, those of individual or group interests, or rights, one can call this dimension the ‘comparative deontic status’ of the claim. Variations in comparative deontic status take place on a continuum between two extremes. On one rather nasty extreme the nation-focused claims are seen as trumps that take precedence over any other claims, even over human rights. Further towards the center is the classical nationalism that gives nation-centered claims precedence over the interests and needs of the individual (including pragmatic collective utility), but not necessarily over general human rights. (See, for example, McIntyre 1994, Oldenquist, 1997.) On the opposite end, which is mild, humane and liberal, the central nationalistic claims are accorded *prima facie* status only (see Tamir, 1993).

(iii) *For which groups are the nationalist claims meant to be valid, what is their scope?* First, they can be valid for every ethno-nation and thereby universal. An example would be the claim "every ethno-nation should have its own state". To put it more formally, we have:

> **Universalist nationalism** is the political program that claims that *every* ethno-nation should have its state which it should rightfully own, and whose interests it should promote.

Alternatively, a claim may be more particular, such as the claim "Group *X* ought to have a state", where this implies nothing about any other group:

> **Particularist nationalism** is the political program that claims that *some* ethno-nation should have its state, without extending the claim to all ethno-nations. It does this either
>
> > (A) by omission (unreflective particularist nationalism), or
> >
> > (B) by explicitly specifying who is excluded: "Group *X* ought to have a state, but group *Y* should not" (invidious nationalism).

We have called the most difficult (and indeed chauvinistic) sub-case of particularism, i.e., (B), ‘invidious’ since it explicitly denies the privilege of having a state to some peoples. T. Pogge (1997) proposes a further division of (B) into the ‘high’ stance which denies statehood to some types of peoples,

and the ‘low’ one which denies it to some particular peoples. Serious theoretical nationalists usually defend only the universalist variety, whereas the nationalist-in-the-street most often defends the particularist one (‘Some nations should have a state, above all mine!’). Classical nationalism comes both in particularist and universalist varieties.

Although the three dimensions of variation described above -- (i) internal deontic status, (ii) external deontic status, and (iii) scope -- are logically independent, they are psychologically and politically intertwined. People who are radical in one respect on the nationality issue tend also to be radical in other respects. In other words, attitudes tend to cluster together in stable clusters, so that extreme (or moderate) attitudes on one dimension psychologically and politically belong with extreme (or moderate) ones on others. The hybrids of extreme attitude on one dimension with moderate on the others are psychologically and socially unstable.

3. The Moral Debate

3.1 Classical and Liberal Nationalisms

Recall our initial analysis, centered around (1) attitudes and (2) actions. Is national partiality justified and to what extent? What actions are appropriate to bringing sovereignty about? In particular, are ethno-national states and institutionally protected (ethno-) national cultures goods independent from the individual will of the members, and how far may one go in protecting them? The philosophical debate for and against nationalism is the debate about the moral validity of its central claims. In particular the ultimate moral issue is the following: is any form of nationalism morally permissible or justified, and, if not, how bad are particular forms of it?

Why do nationalist claims require a defense? In some situations they seem very plausible; for instance the plight of some stateless national groups -- the history of Jews and Armenians, the misfortunes of Kurds -- makes one spontaneously endorse the idea that if these groups had had their own state, serious problems would have been avoided. Still, there are good reasons to examine the general nationalist claims. The most general reason is that it should first be shown that the nation as such has some value and that claims in its favor have some normative validity. Once this is established, a further defense is needed. Some classical nationalist claims appear to clash -- at least under normal circumstances of contemporary life -- with various values that people tend to accept. Some of these values are considered essential to liberal-democratic societies, while others are important specifically for the flourishing of culture and creativity. The main values in the first set are individual autonomy and benevolent impartiality (most prominently towards members of groups culturally different from one's own). The alleged special duties towards one's ethno-national culture can interfere, and often do interfere, with individuals’ right to autonomy. Also, if these duties are construed very strictly they can interfere with other individual rights, e.g., the right to privacy. Many feminist authors have noted that a suggestion typically offered by the nationalist, namely that women have a moral obligation to give birth to new members of the nation and to nurture them for the sake of the nation, clashes both with autonomy and

privacy of these women (Yuval-Davis 1997). Another endangered value is diversity within ethno-national community, which can also be thwarted by the homogeneity of a central national culture.

The nation-oriented duties also interfere with the value of unconstrained creativity, e.g., telling writers or musicians or philosophers that they have a special duty to promote national heritage interferes with the freedom of creation. (The question here is not whether these individuals have the right to promote their national heritage, but whether they have a duty to do so.)

In between these two sets of endangered values, the autonomy-centered and creativity-centered ones, are the values that seem to arise from ordinary needs of people living under ordinary circumstances. In many modern states citizens of different ethnic background live together, and very often value this kind of life. This very fact of cohabitation seems to be a good that should be upheld. Nationalism does not tend to foster this kind of multiculturalism and pluralism, judging from both theory (especially the classical nationalist one) and experience. But the problems get worse. In practice, a rather widespread variant of nationalism is the invidious particularist form claiming rights for one's own people and denying them to others, for reasons that seem to be far from accidental. The source of the problem is the competition for scarce resources; as Gellner (1983) has famously pointed out, there is too little territory for all candidate ethnic groups to have a state and the same goes for other goods demanded by nationalists for the exclusive use of their co-nationals. According to some authors (McCabe 1997) the invidious variant is more coherent than any other form of nationalist; if one values highly one's own ethnic group the simplest way is to value it *tout court*. If one definitely prefers one's own culture in all respects to any foreign one, it is a waste of time and attention to bother about others. The universalist, non-invidious variant introduces enormous psychological and political complications. They arise from a tension between spontaneous attachment to one's own community and the demand to regard all communities with an equal eye. This tension might make the humane, non-invidious position psychologically unstable, and hard to uphold in situations of conflict and crisis. This psychological weakness renders it politically less efficient.

The philosophical authors sympathetic to nationalism are aware of the evils that historical nationalism has produced, and usually distance themselves from them. They usually speak of "various accretions that have given nationalism a bad name," and they are eager to "separate the idea of nationality itself from these excesses" (Miller 1992, 87). Such thoughtful pro-nationalist writers have put forward several lines of thought in defense of such a nationalism, thereby initiating an ongoing philosophical dialogue between the proponents and the opponents of the claim (see the anthologies McKim & McMahan 1997, Couture, Nielsen, & Seymour 1998, and Miscevic 2000). In order to help the reader find his or her way through the rather involved debate, we shall briefly summarize the considerations which are open to the ethno-nationalist to defend his or her case. (Compare the useful overview in Lichtenberg 1997.) The considerations and lines of thought built upon them can be used to defend very different varieties of nationalism, from radical to very moderate ones.

It is important to offer a warning concerning the key assumptions and premises which figure within each of the lines of thought summarized below, namely, that the assumptions often live an independent life in the philosophical literature. Some of them figure in the proposed defenses of various traditional views

which have little to do with the concept of nation in particular.

For brevity, I shall reduce each line of thought to a brief argument (these arguments will be listed and discussed in Sections 3.2 and 3.3). The actual debate is, however, more involved than one can represent in a sketch. I shall sometimes indicate, in brackets, some prominent lines of criticism that have been put forward in the debate. (These are discussed in greater detail in Miscevic 2001.) The main arguments in favor of nationalism, which purport to establish its fundamental claims about state and culture, will be divided into two sets. The first set of arguments (listed in Section 3.2) defend the claim that national communities have a high value, often seen as non-instrumental and independent of the wishes and choices of their individual members, and argue that therefore they should be protected by means of state and official statist policies. The second set (listed in Section 3.3) is less deeply ‘philosophical’ (or ‘comprehensive’), and encompasses arguments from requirements of justice which are rather independent of substantial assumptions about culture and cultural values.

The first set will be presented here in more detail, since it has formed the center of the debate. The arguments in this set depict the community as the source of value or as the unique transmission device that connects the members to some important values. In this sense, the arguments from this set are communitarian in the ‘deep’ sense of being grounded in basic features of the human condition. Here is a characterization.

> The **deep communitarian perspective** is a theoretical perspective on political issues which justifies a given political arrangement (e.g., a nation-state) by appeal to philosophical assumptions about human nature, language, community ties and identity (in the philosophical sense).

The general form of the arguments in the first set is the following. First, the deep communitarian premise: there is some uncontroversial good (e.g., a person's ethnic identity), and some kind of community is essential for acquisition and preservation of it. Then comes the claim that ethno-cultural nation is the kind of community ideally suited for this task. Unfortunately, this crucial claim is rarely defended in detail in the literature. But here is a sample from Avishai Margalit:

> The idea is that people make use of different styles to express their humanity. The styles are generally determined by the communities to which they belong. There are people who express themselves ‘Frenchly’, while others have forms of life that are expressed ‘Koreanly’ or ... ‘Icelandicly’. (Margalit 1997, 80)

The argument ends with the statist conclusion: in order that such a community should preserve its own identity and support the identity of its members, it has to assume (always or at least normally) the political form of a state. The conclusion of this type of argument is that the ethno-national community has the right, in respect to any third party and to its own members, to have an ethno-national state, and the citizens of the state have the right and obligation to favor their own ethnic culture in relation to any other.

Although the philosophical assumptions in the arguments stem from the communitarian tradition, weakened forms have also been proposed by more liberally-minded philosophers. The original communitarian lines of thought in favor of nationalism suggest that there is some value in preserving ethno-national cultural traditions, in feelings of belonging to a common nation and in solidarity between its members. A liberal nationalist might accept that these may not be the central values of political life, but claim that they are values nevertheless. Moreover, the diametrically opposite views, pure individualism and cosmopolitanism, do seem arid and abstract, and may appear unmotivated. Compare, for example, the notion of cosmopolitanism:

> **Cosmopolitanism** is the moral and political doctrine which asserts that (a) one's primary moral obligations are directed to all human beings (regardless of geographical or cultural distance), and (b) political arrangements should faithfully reflect this universal moral obligation (in the form of supra-statist arrangements that take precedence over nation-states).

Confronted with these opposite pulls, many philosophers opt for a mixture of liberalism-cosmopolitanism and patriotism-nationalism. B. Barber in his writings glorifies "a remarkable mixture of cosmopolitanism and parochialism" which, in his view characterizes American national identity (in Cohen 1996, 31). Charles Taylor claims that "we have no choice but to be cosmopolitan and patriots" (*ibid*., 121). Hilary Putnam proposes loyalty to what is best in the multiple traditions that each of us participates in; apparently a middle way between a narrow-minded patriotism and a too abstract cosmopolitanism (*ibid*., 114). The compromise has been foreshadowed by Berlin (1979) and Taylor (1989, 1993), and its various versions worked out in considerable detail by authors such as Tamir (1993), Miller (1995), Nielsen (1998) and Seymour. In recent years it occupies the center stage of the debate. Most liberal nationalist authors accept various weakened versions of the arguments we list below, taking them to support moderate or ultra-moderate nationalist claims.

Here are the main weakenings of classical ethno-nationalism that the liberal, limited-liberal and cosmopolitan nationalists propose. First, ethno-national claims have only *prima facie* strength, and cannot trump individual rights. Second, legitimate ethno-national claims do not, in themselves and automatically, amount to a right to having a state, but rather a right to have a certain level of cultural autonomy. Third, ethno-nationalism is subordinate to civic patriotism, and this has little or nothing to do with ethnic criteria. Fourth, ethno-national mythologies and similar ‘important falsehoods’ are to be tolerated only if benign and inoffensive, in which case they are morally permissible in spite of their falsity. Finally, any legitimacy that ethno-national claims may have is to be derived from the choices the concerned individuals should be free to make.

3.2 Arguments in Favor of Nationalism: The Deep Need for Community

Consider now the particular arguments from the first set. The first argument depends on assumptions that

also appear in the subsequent ones, but whereas it ascribes to the community an intrinsic value, the following ones point more towards an instrumental value of a nation which is derived from the value of individual flourishing, moral understanding, firm identity and the like.

(1) *The Argument From Intrinsic Value*. Each ethno-national community is valuable in and of itself, since it is only within the natural encompassing framework of various cultural traditions that important meanings and values are produced and transmitted. The members of such communities share a special cultural proximity to each other. By speaking the same language and sharing customs and traditions, the members of these communities are typically closer to one another in various ways than they are to those who don't share the culture. The community thereby becomes a network of morally connected agents, i.e., a moral community, with special, very strong ties of obligation. A prominent obligation of each individual concerns the underlying traits of the ethnic community, above all language and customs: they ought to be cherished, protected, preserved and reinforced. (The general assumption that moral obligations increase with cultural proximity is often criticized as problematic. Moreover, even if we grant this general assumption in theory, it breaks down in practice. Nationalist activism is most often turned against close (and, in many respects, similar) neighbors rather than against distant strangers, so that in many important contexts the appeal to proximity will not work. It might retain its potential force against culturally distant groups, however.)

(2) *The Argument From Flourishing*. The ethno-national community is essential for each of its members to flourish. In particular, it is only within such a community that an individual can acquire concepts and values crucial for understanding the community’s cultural life in general and one’s own life in particular. There has been a lot of debate on the pro-nationalist side about whether divergence of values is essential for separateness of national groups. Canadian liberal nationalists, Seymour (1999), Taylor, and Kymlicka, pointed out that the ‘divergences of values between different regions of Canada’ that aspire to separate nationhood are ‘minimal’. Taylor (1993, 155) concluded that it is not separateness of value that matters. This result is still compatible with the argument from flourishing, if ‘concepts and values’ are not taken to be specifically national, as communitarian nationalists (MacIntyre 1994, Margalit 1997) have claimed.

(3) *The Argument From Moral Understanding*. A particularly important variety of value is moral value. Some values are universal, but they are too abstract and ‘thin’. The rich, ‘thick’ moral values are discernible only within particular traditions, to those who have wholeheartedly endorsed the norms and standards of the given tradition. As Charles Taylor puts it, ‘the language we have come to accept articulates the issues of the good for us’ (1989, 35). The nation offers a natural framework for moral traditions, and thereby for moral understanding; it is the primary school of morals. (I note in fairness that Taylor himself is ambivalent about the national format of morality. An often noticed problem for this line of thought is that particular nations do not each have a special morality of their own. Also, the detailed, ‘thick’ morality may vary more across other divisions, such as class or gender divisions, than across ethno-national groups.)

(4) *The Argument from Identity*. Communitarian philosophers emphasize nurture over nature as the principal force determining our identity as persons -- we come to be the persons we are because of the

social settings and contexts in which we mature. The claim certainly has some plausibility. The very identity of each person depends upon his/her participation in communal life (see MacIntyre 1994, Nielsen, 1998, and Lagerspetz 2000). For example, Nielsen writes:

> We are, to put it crudely, lost if we cannot identify ourselves with some part of an objective social reality: a nation, though not necessarily a state, with its distinctive traditions. What we find in people -- and as deeply embedded as the need to develop their talents -- is the need not only to be able to say what they can do but to say who they are. This is found, not created, and is found in the identification with others in a shared culture based on nationality or race or religion or some slice or amalgam thereof. ... Under modern conditions, this securing and nourishing of a national consciousness can only be achieved with a nation-state that corresponds to that national consciousness. (1993, 32)

Given that an individual's morality depends upon their having a mature and stable personal identity, the communal conditions which foster the development of such personal identity have to be preserved and encouraged. The philosophical nationalists claim that the national format is the right format for preserving and encouraging such identity-providing communities. Therefore, communal life should be organized around particular national cultures. The classical nationalist proposes that cultures should be given their states, while the liberal nationalist proposes that cultures should get at least some form of political protection.

(5) *The Argument from Diversity*. Each national culture contributes in a unique way to the diversity of human cultures. The most famous contemporary proponent of the idea, Isaiah Berlin (interpreting Herder, who first saw this idea as significant) writes:

> The ‘physiognomies’ of cultures are unique: each presents a wonderful exfoliation of human potentialities in its own time and place and environment. We are forbidden to make judgments of comparative value, for that is measuring the incommensurable. (1976, 206)

The carrier of basic value is thus the totality of cultures, from which each national culture and style of life that contributes to the totality derives its own value. The plurality of styles can be preserved and enhanced by tying the styles to ethno-national ‘forms of life’. The argument from diversity is therefore pluralistic: it ascribes value to each particular culture from the viewpoint of the totality of cultures available. Assuming that the (ethno-) nation is the natural unit of culture, the preservation of cultural diversity amounts to institutionally protecting the purity of (ethno-) national culture. (A pragmatic inconsistency might be threatening this argument. The issue is who can legitimately propose nationhood: the nationalist is much too tied to his or her own culture to do it, while the cosmopolitan is too eager to preserve intercultural links that go beyond the idea of having a single nation-state. Moreover, is diversity a value such that it deserves to be protected whenever it exists?)

The lines of thought in the set of arguments just presented are all linked to the importance of community life in relation to the individual. They emerged from the perspective of ‘deep’ communitarian thought,

and a recurrent theme is the importance of the fact that membership in the community is not chosen but rather involuntary. As noted previously, each argument involves a general communitarian premise (a community, to which one has no choice whether or not to belong, is crucial for one's identity, or for flourishing or for some other important good). This premise is coupled with the more narrow nation-centered descriptive claim that the ethno-nation is precisely the kind of community ideally suited for the task. However, liberal nationalists do not find these arguments completely persuasive. In their view, the premises of the arguments may not support the full package of nationalistic ambitions, and may not be unconditionally valid. Still, there is a lot to these arguments, and they might support liberal nationalism and a more modest stance in favor of national cultures.

The liberal nationalist stance is mild and civil, and there is lot to be said in its favor. It strives to reconcile our intuitions in favor of some sort of political protection of cultural communities, with a liberal political morality. Of course, this raises issues of compatibility between liberal universal principles and the particular attachments to one’s ethno-cultural nation. Very liberal nationalists such as Tamir divorce ethno-cultural nationhood from statehood. Also, the kind of love for country they suggest is tempered by all kinds of universalist considerations, which in the last instance trump national interest (Tamir 1993, 115). There is an ongoing debate among philosophical nationalists about how much weakening and compromising is still compatible with a stance’s being nationalist at all. (For example, Canovan 1996 (ch. 10) presents Tamir as abandoning the ideal of ‘nation-state’, and thereby nationhood as such; Seymour (1999) criticizes Taylor and Kymlicka for turning their backs on genuine nationalistic programs, and proposing multiculturalism instead of nationalism.) There is also a streak of cosmopolitan interest present in the work of some liberal nationalists (Nielsen (1998/9).

3.3 Arguments in Favor of Nationalism: Issues of Justice

The arguments in the second set concern political justice and do not rely on metaphysical claims about identity, flourishing and cultural values. They appeal to (actual or alleged) circumstances which would make nationalistic policies reasonable (or permissible or even mandatory), such as (a) the fact that a large part of the world is organized into nation states (so that each new group aspiring to create a nation-state just follows an established pattern), or (b) the circumstances of group self-defense or of redress of past injustice which might justify nationalist policies (to take a special case). Some of them also present nationhood as conducive to important political goods, such as equality.

(1) *The Argument From the Right to Collective Self-determination*. A group of people of a sufficient size has a *prima face* right to govern itself and decide its future membership, if the members of the group so wish. It is fundamentally the democratic will of the members themselves that grounds the right to an ethno-national state and to ethno-centric cultural institutions and practices. This argument presents the justification of (ethno-) national claims as deriving from the will of the members of the nation. It is therefore highly suitable for liberal nationalism and not very interesting for a deep communitarian, who sees the demands of the nation as being independent from, and prior to, the choices of particular individuals. (For extended discussion of this argument, see Moore 1998.)

(2) *The Argument From the Right to Self-defense and to Redress Past Injustices*. Oppression and injustice give a victimized group a just cause and the right to secede. If a minority group is oppressed by the majority, so that almost every minority member is worse off than most majority ones, simply in virtue of belonging to the minority, then the nationalist minority claims are morally plausible, and may even be compelling. The argument implies a very restrictive answer to our questions (2b) and (2c): the use of force in order to achieve sovereignty is legitimate only in the cases of self-defense and redress. It establishes a typical remedial right, which is acceptable from a liberal standpoint. (See a recent discussion in Kukathas and Poole 2000.)

(3) *The Argument From Equality*. Members of a minority group are often disadvantaged in relation to a dominant culture because they have to rely on those with the same language and culture to conduct the affairs of daily life. Since freedom to conduct one's daily life is a primary good, and it is difficult to change or give up reliance on one's minority culture to attain that good, this reliance can lead to certain inequalities if special measures are not taken. Spontaneous nation-building by the majority has to be moderated. Therefore, liberal neutrality itself requires that the majority provide certain basic cultural goods, i.e., granting differential rights. (See Kymlicka 1995b.) Institutional protections and the right to the minority group's own institutional structure are remedies that restore equality and turn the resulting nation-state into a more moderate multicultural one. (See Kymlicka 2001.)

(4) *The Argument From Success*. The nation-state has been successful in the past, promoting equality and democracy. The ethno-national solidarity is a powerful motive for a more egalitarian distribution of goods (Miller 1995, Canovan, 1996). The nation-state also seems to be essential to safeguard the moral life of communities in the future, since it is the only form of political institution capable of protecting communities from the threats of globalization and assimilationism. (For a detailed critical discussion of this argument see Mason 1999.)

These political arguments can be combined with deep communitarian ones. However, taken in isolation, they offer the more interesting perspective of a ‘liberal culturalism’ that is more suitable for ethnically and culturally plural societies. It is more remote from classical nationalism than the liberal nationalism of Tamir and Nielsen, since it eschews any communitarian philosophical underpinning (see the detailed presentation and defense in Kymlicka 2001, who still occasionally calls such culturalism ‘nationalistic’). The idea of moderate nation-building points to an open multi-culturalism, in which every group receives its share of remedial rights, but instead of walling itself up against others, participates in a common, overlapping civic culture in open communication with other sub-communities. Given the variety of pluralistic societies, and intense trans-national interactions, such openness seems to many to be the only guarantee of stable social and political life (see the debate in Shapiro and Kymlicka 1997). The dialectics of moderating nationalist claims in the context of pluralistic societies might thus lead to a stance which is respectful of cultural differences, but liberal and potentially cosmopolitan in its ultimate goals.

4. Conclusion

The philosophy of nationalism nowadays does not concern itself very much with the aggressive and

dangerous form of invidious nationalism that often occupies center stage in the news and in sociological research. Although this pernicious form can be of significant instrumental value mobilizing oppressed people and giving them a sense of dignity, its moral costs are usually taken by philosophers to outweigh its benefits. Nationalistic-minded philosophers distance themselves from such aggressive nationalism, and mainly seek to construct and defend very moderate versions and these have therefore come to be the main focus of recent philosophical debate.

In presenting the claims that nationalists defend, we have started from more radical ones and have moved towards liberal nationalistic alternatives. In examining the argument for these claims, we have first presented metaphysically demanding communitarian arguments, resting upon deep communitarian assumptions about culture, such as the premise that the ethno-cultural nation is universally the central and most important community for each human individual. This is an interesting and respectable claim, but its plausibility has not yet been established. The moral debate about nationalism has resulted in various weakenings of the cultural arguments, proposed by liberal nationalist, which render the arguments less ambitious but much more plausible. Having abandoned the old nationalistic ideal of a state owned by its ethno-cultural dominant group, liberal nationalists have become receptive to the idea that identification with a multitude of cultures and communities is important for a person's social identity. They have equally become sensitive to trans-national issues, and more willing to embrace a partly cosmopolitan perspective.

Liberal nationalism has also brought to the fore more modest, and less philosophically charged, arguments which are grounded in the concerns of justice and which stress the practical importance of ethno-cultural membership, various rights to redress injustice, democratic rights of political association, and the role that ethno-cultural ties and associations can play in promoting just social arrangements. Liberal culturalists such as Kymlicka have proposed minimal and pluralistic versions of nationalism built around such arguments. In these minimal versions, the project of building classical nation-states is moderated or abandoned, and replaced by a more sensitive form of national identity which can thrive in a multicultural society. This new project, however, might demand a further widening of moral perspectives. Given the experiences of the twentieth century, one can safely assume that culturally-plural states divided into isolated and closed sub-communities glued together only by arrangements of mere *modus vivendi* are inherently unstable. Stability might therefore require that the plural society envisioned by liberal culturalists promote quite intense interaction between cultural groups in order to forestall mistrust, reduce prejudice and create a solid basis for cohabitation. On the other hand, once membership in multiple cultures and communities is admitted as legitimate, social groups will spread beyond the borders of a single state (like many religious or racial ties) as well as within them, thus creating an opening for at least a minimal cosmopolitan perspective. The internal dialectic of the concern for ethno-cultural identity might thus lead to pluralistic and potentially cosmopolitan political arrangements that are rather distant from what was classically understood as nationalism.

Bibliography

References

- Anderson, B., (1965), *Imagined Communities*, London: Verso.
- Avineri, Shlomo and de-Shalit, Avner (eds), 1992, *Communitarianism and Individualism*, Oxford: Oxford University Press.
- Balibar, E., and Wallerstein, I., 1992, *Class, Race Nation*, London-New York: Verso.
- Barber, B., 1996, 'Constitutional Faith', in Cohen 1996.
- Berlin, I., 1976, *Vico and Herder*, Oxford: Clarendon Press.
- Berlin, I, 1979, 'Nationalism: Past neglect and Present Power', in *Against the Current*, New York: Penguin
- Billig, M., 1995, *Banal Nationalism*, Thousand Oaks-London-New Delhi: Sage Publications.
- Canovan, M., 1996, *Nationhood and Political Theory*, Cheltenham: Edward Elgar.
- Cohen, J. (ed), 1996, Martha Nussbaum and respondents, *For Love of Country: Debating the Limits of Patriotism*, Boston: Beacon Press.
- Couture, J., Nielsen, K. and Seymour, M. (eds.), 1998, *Rethinking Nationalism*, Canadian Journal of Philosophy, Supplement Volume 22.
- Crowley, B.I., 1987, *The Self, the Individual and the Community*, Oxford: Clarendon Press.
- Gellner, E., 1983, *Nations and Nationalism*, Oxford: Blackwell.
- Habermas, J., 1996 *Between Facts and Norms: Contribution to a Discourse Theory of Law and Democracy*, Cambridge: Polity Press.
- Hardin, Russell, 1985, *One for All, The Logic of Group Conflict*. Princeton, NJ: Princeton University Press.
- Hobsbawn, E. J., 1990, *Nations and Nationalism since 1780: Programme, Myth, Reality*, Cambridge: Cambridge University Press.
- Kedourie, E., 1960, *Nationalism*, London: Hutchison.
- Kohn, H., 1965, *Nationalism: its meaning and history*, New York: Van Nostrand Reinhold Company.
- Kuran Burcoglu, N. (ed), 1997, *Multiculturalism: Identity and Otherness*, Istanbul: Bogazici University Press.
- Kukathas, C., and Poole, R. (eds.), 2000, Special Issue on Indigenous Rights, *Australasion Journal of Philosophy*, v.78
- Kymlicka, W. (ed), 1995a, *The Rights of Minority Cultures*, Oxford: Oxford University Press.
- Kymlicka, W., 1995b, *Multicultural Citizenship*, Oxford: Oxford University Press.
- Kymlicka, W., 2001, *Politics in the vernacular*, Oxford: Oxford University Press.
- Lagerspetz, O., 2000, 'On National Belonging' in Miscevic (ed), *Nationalism and ethnic conflict*, La Salle and Chicago: Open Court.
- Laitin, D., 1998, *Identity in Formation: The Russian-Speaking Populations in the Near Abroad*, Ithaca, NY: Cornell University Press.
- Levy, J., 2000, *Multiculturalism of Fear*, Oxford: Oxford University Press
- Lichtenberg, J., 1997, 'Nationalism, For and (Mainly) Against', in McKim & McMahan 1997.
- MacCormick, N., 1982, *Legal Right and Social Democracy*, Oxford: Clarendon Press.
- MacIntyre, A., 1994, 'Is Patriotism a Virtue', in *Communitarianism*, ed. M. Daly. Belmont, CA: Wadsworth.
- Margalit, A., 1997, 'The Moral Psychology of Nationalism', in McKim and McMahan 1997.

- Mason, A., 1999, 'Political Community, Liberal-Nationalism and the Ethics of Assimilation', *Ethics*, v. 109, 261-286
- McCabe, D., 1997, 'Patriotic Gore Again', *The Southern Journal of Philosophy*, 35: 203-223.
- McKim, R. and McMahan, J. (eds), 1997, *The Morality of Nationalism*, Oxford: Oxford University Press.
- Miller, D., 1992, 'Community and Citizenship', in Avineri and de Shalit: *Communitarianism and individualism*, Oxford: Oxford University Press (originally from his *Market, State and Community*).
- Miller, D., 1995, *On Nationality*, Oxford: Oxford University Press.
- Miscevic, N. (ed), 2000, *Nationalism and Ethnic Conflict. Philosophical Perspectives*, La Salle and Chicago: Open Court.
- Miscevic, N., 2001, *Nationalism and Beyond*, Budapest, New York: Central European University Press
- Moore, M. (ed.), 1998, *National Self-Determination and Secession*, Oxford: Oxford University Press
- Nielsen, K., 1998, 'Liberal Nationalism, Liberal Democracies and Secession', *University of Toronto Law Journal*, vol. 48, pp. 253-295.
- Nielsen, K., 1998-99, 'Cosmopolitanism, Universalism and Particularism in the age of Nationalism and Multiculturalism', *Philosophical Exchange*, 29: 3-34.
- Nyiri, J. C. (ed), 1994, *Nationalism and Social Science*, issue of *Studies in East European Thought*, vol. 46, Nos.1-2.
- Oldenquist, A., 1997, 'Who Are the Rightful Owners of the State?', in Kohler, P. and Puhl, K. 1997 (eds.) *Proceedings of the 19th International Wittgenstein Symposium*, Vienna: Holder-Pichler-Tempsky.
- Pogge, T., 1997, 'Group Rights and Ethnicity', in I. Shapiro and W. Kymlicka (eds.), *Ethnicity and Group Rights, Nomos* XXXIX, New York: New York University Press
- Putnam, H., 1996, 'Must we choose between patriotism and universal reason?', in Cohen 1996.
- Renan, E., 1882, 'What is a nation?', in *Nation and Narration*, H. Bhabha (ed.), Routledge, London; also in *Nationalisms*, J. Hutchinson and A. Smith (eds.), Oxford: Oxford University Press.
- Seymour, M., 2000, 'On Redefining the Nation', in Miscevic (ed) (2000), *Nationalism and Ethnic Conflict. Philosophical Perspectives*, La Salle and Chicago: Open Court.
- Seymour, M., 1999, *La nation en question*, Montreal: L'Hexagone
- Shapiro, I., and Kymlicka, W. (eds.), 1997, *Ethnicity and Group Rights*, *Nomos* XXXIX, New York: New York University Press
- Smith, A. D., 1991, *National Identity*, Harmondsworth: Penguin
- Sober, E., and Wilson, D.S., 1998, *Unto Others*, Cambridge, MA: Harvard University Press.
- Tajfel, H., 1981, *Human Groups and Social Categories*, Cambridge: Cambridge University Press.
- Tamir, Y., 1993, *Liberal Nationalism*, Princeton, NJ: Princeton University Press
- Taylor, C., 1989, *Sources of the Self*, Cambridge: Cambridge University Press.
- Taylor, C., 1993, *Reconciling the Solitudes*, Montreal: McGill-Queen's University Press
- Twining, W. (ed), 1991, *Issues of Self-determination*. Aberdeen: Aberdeen University Press.
- Weber, M., 1970, *From Max Weber* (selections translated by H. H. Gerth and C. Wright Mills),

London: Routledge.
- Yuval-Davis, N., 1997, *Gender & Nation*, Thousand Oaks-London-New Delhi: Sage Publications

A Beginner's Guide to the Literature

This is a short list of books on nationalism that are readable and useful as introductions to the literature. First, the two opposing social science contemporary classics:

- Gellner, E., 1983, *Nations and Nationalism*, Oxford: Blackwell.
- Smith, A. D., 1991, *National Identity*, Harmondsworth: Penguin.

The two best recent anthologies of high-quality philosophical papers on the morality of nationalism, are:

- McKim, R. and McMahan, J. (eds), 1997, *The Morality of Nationalism*, Oxford: Oxford University Press.
- Couture, J., Nielsen, K. and Seymour, M. (eds.), 1998, *Rethinking Nationalism*, Canadian Journal of Philosophy, Supplement Volume 22.

The debate continues in:

- Miscevic, N. (ed), 2000, *Nationalism and Ethnic Conflict. Philosophical Perspectives*, La Salle and Chicago: Open Court.

A good sociological introduction to the gender-inspired criticism of nationalism is:

- Yuval-Davis, N., 1997, *Gender & Nation*, Thousand Oaks-London-New Delhi: Sage Publications.

The best general introduction to the communitarian-individualist debate is still:

- Avineri, Shlomo and de-Shalit, Avner (eds), 1992, *Communitarianism and Individualism*, Oxford: Oxford University Press.

For a non-nationalist defense of culturalist claims see

- Kymlicka, W. (ed), 1995a, *The Rights of Minority Cultures*, Oxford: Oxford University Press.

Two very readable philosophical defenses of very moderate nationalism are:

- Miller, D., 1995, *On Nationality*, Oxford: Oxford University Press.
- Tamir, Y., 1993, *Liberal Nationalism*, Princeton, NJ: Princeton University Press.

An influential critical analysis of group solidarity in general and nationalism in particular, written in the tradition of rational choice theory is:

- Hardin, R., 1985, *One for All, The Logic of Group Conflict*, Princeton, NJ: Princeton University Press

There is a wide offer of interesting sociological and political science work on nationalism, which is beginning to be summarized in

- Motyl, A. (Ed.) 2000, *Encyclopedia of Nationalism, v. I*, New York: Academic Press.

A detailed sociological study of life under nationalist rule is:

- Billig, M., 1995, *Banal Nationalism*, Thousand Oaks-London-New Delhi: Sage Publications.

The most readable short anthology of brief papers for and against cosmopolitanism (and nationalism) by leading authors in the field is:

- Cohen, J. (ed), 1996, Martha Nussbaum and respondents, *For Love of Country: Debating the Limits of Patriotism*, Boston: Beacon Press.

Other Internet Resources

- Nationalism Links (maintained by Peter Rasmussen)
  A good collection of links and bibliographies.
- The Warwick Debates (London School of Economics)
  Debate between E. Gellner and A. Smith.
- ARENA Research Program on the Evolution of Systems of European Governance (U. of Norway)
  A good selection of papers on ethics of international relations.
- Global Policy Forum's pages on Nations and States
  Papers on the future of nation-states.
- European University Institute European perspectives on nations, nationalism and nation-states.

Related Entries

communitarianism | cosmopolitanism | identity | liberalism | prisoner's dilemma | secession | self-determination, collective | war

Nenad Miscevic
University of Maribor (Slovenia)
vera.gambar-miscevic@ri.tel.hr

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 29, 2001
Content last modified: November 29, 2001

Cosmopolitanism

The word ‘cosmopolitan’, which derives from the Greek word *kosmopolitês* (‘citizen of the world’), has been used to describe a wide variety of important views in moral and socio-political philosophy. The nebulous core shared by all cosmopolitan views is the idea that all human beings, regardless of their political affiliation, do (or at least can) belong to a single community, and that this community should be cultivated. Different versions of cosmopolitanism envision this community in different ways, some focusing on political institutions, others on moral norms or relationships, and still others focusing on shared markets or forms of cultural expression. The philosophical interest in cosmopolitanism lies in its challenge to commonly recognized attachments to fellow-citizens, the local state, parochially shared cultures, and the like.

1. History of Cosmopolitanisms

1.1 Greek and Roman Cosmopolitanism

The political culture that is idealized in the writings of Plato and Aristotle is not cosmopolitan. In this culture, a man identifies himself first and foremost as a citizen of a particular polis or city, and in doing so, he signals which institutions and which body of people hold his allegiance. He would then be counted on for help in defending the city from attacks, sustaining its institutions of justice, and contributing to its common good. In this way, his own pursuit of a good life is inextricably bound to the fate of the city and to the similar pursuit carried out by other inhabitants of the city. By contrast, the good person would not be expected to share with or serve any foreigners who live outside the city. Any cosmopolitan expectations on a good Athenian extended only to concern for those foreigners who happen to reside in Athens.

It would, however, be wrong to assume that Classical Greek thought was uniformly *anti-cosmopolitan*.

Actively excluding foreigners from any ethical consideration or actively targeting foreigners for heinous treatment goes one step beyond focusing one's service and concern on compatriots, and in fact, the targeting of 'barbarians' is historically linked with the rise of panhellenism and not with the more narrow emphasis on the polis. It would be more accurate to call the Classical emphasis on the polis *uncosmopolitan*.

Yet even as Plato and Aristotle were writing, other Greeks were issuing cosmopolitan challenges. Perhaps the most obvious challenges came from the traveling intellectuals who insisted on the contrast between the conventional ties of politics and the natural ties of humanity. Notice, for example, the way Plato has the Sophist Hippias address the motley crew of Athenians and foreigners present at Callias' house in Plato's *Protagoras* (337c7-d3):

> Gentlemen present ... I regard you all as kinsmen, familiars, and fellow-citizens -- by nature and not by convention; for like is by nature akin to like, while convention, which is a tyrant over human beings, forces many things contrary to nature.

Socrates, too, it can be argued, was sensitive to this more cosmopolitan identification with human beings as such. At least as Plato characterizes him, Socrates avoids traditional political engagement as much as he can, in favor of an extraordinary career of examining himself and others, and he insists that these examinations are both genuinely political (*Gorg* 521d6-8) and extended to all, Athenians and foreigners alike (*Apol* 23b4-6). Of course, Socrates chose not to travel widely, but this decision could well have been consistent with cosmopolitan ideals, for he may have thought that his best bet for serving human beings generally lay in staying at home, on account, ironically, of Athens' superior freedom of speech (*Gorg* 461e1-3; cf. *Apol* 37c5-e2 and *Meno* 80b4-7). Whether Socrates was self-consciously cosmopolitan in this way or not, there is no doubt that his ideas accelerated the development of cosmopolitanism and that he was in later antiquity embraced as a citizen of the world. In fact, the first philosopher in the West to give perfectly explicit expression to cosmopolitanism was the Socratically inspired Cynic Diogenes in the fourth century bce. It is said that "when he was asked where he came from, he replied, 'I am a citizen of the world [*kosmopolitês*]'" (Diogenes Laertius VI 63). By identifying himself not as a citizen of Sinope but as a citizen of the world, Diogenes was refusing to agree that he owed special service to Sinope and the Sinopeans. So understood, 'I am a citizen of the cosmos' is a negative claim, and we might wonder if there is any positive content to the Cynic's world-citizenship. The most natural suggestion would be that a world-citizen should serve the world-state, helping to bring it about in order to enable the later work of sustaining its institutions and contributing to its common good. But the historical record does not suggest that Diogenes the Cynic favored the introduction of a world-state. In fact, the historical record does not unambiguously provide Diogenes any positive commitments that we can readily understand as cosmopolitan. The best we can do to find positive cosmopolitanism in Diogenes is to insist that the whole Cynic way of life is supposed to be cosmopolitan: by living in accordance with nature and rejecting what is conventional, the Cynic sets an example of high-minded virtue to all other human beings.

A fuller exploration of positively committed philosophical cosmopolitanism arrives only with the Socratizing and Cynic-influenced Stoics of the third century ce. These Stoics are fond of saying that the cosmos is, as it were, a polis, because the cosmos is put in perfect order by law, which is divine reason.

They also embrace the negative implication of their high standards for order and law: conventional polises do not, strictly speaking, deserve the name. But the Stoics do not believe that living in agreement with the cosmos -- as a citizen of the cosmos -- requires maintaining critical distance from conventional polises. Rather, as the traces of Chrysippus' *On Lives* make clear, the Stoics believe that goodness requires serving other human beings as best one can given the circumstances, that serving all human beings equally well is impossible, and that the best service one can give typically requires political engagement. Of course, the Stoics recognize that political engagement will not be possible for everyone, and that some people will best be able to help other human beings as private teachers of virtue rather than as politicians. But in no case, the Stoics insist, is consideration of political engagement to be limited to one's own polis. The motivating idea is, after all, to help human beings as such, and sometimes the best way to do that is to serve as a teacher or as a political advisor in some foreign place. In this fashion, the Stoics introduce clear, practical content to their metaphor of the cosmopolis: a cosmopolitan considers moving away in order to serve, whereas a non-cosmopolitan does not.

This content admits of a strict and a more moderate interpretation. On the strict view, when one considers whether to emigrate, one recognizes prima facie no special or stronger reason to serve compatriots than to serve a set of human beings abroad. On the moderate view, one does introduce into one's deliberations extra reason to serve compatriots, although one might still, all things considered, make the best choice by emigrating. The evidence does not permit a decisive attribution of one or the other of these interpretations to any of the earliest Stoics. But if we think of a Stoic like Chrysippus as deeply attracted to the Cynics' rejection of what is merely conventional, then we will find it easy to think of Chrysippus as a strict cosmopolitan.

Things are a bit different for at least some of the Stoics at Rome. On the one hand, the cosmopolis becomes less demanding. Whereas Chrysippus limits citizenship in the cosmos to those who in fact live in agreement with the cosmos and its law, Roman Stoics extend citizenship to all human beings by virtue of their rationality. On the other hand, local citizenship becomes more demanding. There is no doubt that the Stoicism of Cicero's *De Officiis* or of Seneca's varied corpus explicitly acknowledges obligations to the *patria*. This is a moderate Stoic cosmopolitanism, and empire made the doctrine very easy for many Romans by identifying the Roman patria with the cosmopolis itself. But neither imperialism nor a literal interpretation of world-citizenship is required for the philosophical point. The maximally committed cosmopolitan looks around to determine whom he can best help and how, knowing full well that he cannot help all people in just the same way, and his decision to help some people far more than others is justified by cosmopolitan lights if it is the best he can do to help human beings as such.

Stoic cosmopolitanism in its various guises was enormously persuasive throughout the Greco-Roman world. In part, this success can be explained by noting how cosmopolitan the world at that time was. Alexander the Great's conquests and the subsequent division of his empire into successor kingdoms sapped local cities of much of their traditional authority and fostered increased contacts between cities, and later, the rise of the Roman Empire united the whole of the Mediterranean under one political power. But it is wrong to say what has frequently been said, that cosmopolitanism arose as a *response* to the fall of the polis or to the rise of the Roman empire. First, the polis' fall has been greatly exaggerated. Under the successor kingdoms and even -- though to a lesser degree -- under Rome, there remained substantial

room for important political engagement locally. Second, and more decisively, the cosmopolitanism that was so persuasive during the so-called Hellenistic Age and under the Roman Empire was in fact rooted in intellectual developments that *predate* Alexander's conquests. Still, there is no doubting that the empires under which Stoicism developed and flourished made many people more receptive to the cosmopolitan ideal and thus contributed greatly to the widespread influence of Stoic cosmopolitanism.

Nowhere was Stoic cosmopolitanism itself more influential than in early Christianity. Early Christians took the later Stoic recognition of two cities as independent sources of obligation and added a twist. For the Stoics, the work of the polis and its citizens and the work of the cosmopolis and its citizens are the same: both aim to improve the lives of the citizens. The Christians respond to a different call: "Render therefore unto Caesar the things which are Caesar's; and unto God the things that are God's" (Matthew 22:21). On this view, the local city may have divine authority (*John* 19:11; cf. *Romans* 13:1,4,7), but the most important work for human goodness is removed from traditional politics, set aside in a sphere in which people of all nations can become "fellow-citizens with the saints" (*Ephesians* 2:20).

This development has two important and long-lasting consequences, which are canonized by Augustine. First, the cosmopolis again becomes a community for certain people only. Augustine makes this point most explicitly by limiting the citizenship in the city of God to those who love God. All others are relegated to the inferior -- though still universal -- earthly city by their love of self. These two cities of the world, which are doomed to coexist intertwined until the Final Judgment, divide the world's inhabitants. Second, the work of politics is severed from the task of building good human lives, lives of righteousness and justice. While Augustine can stress that this allows citizens in the city of God to obey local laws concerning "the necessaries for the maintenance of life," he must also acknowledge that it sets up a potential conflict over the laws of religion and the concerns of righteousness and justice (e.g., *Civitas Dei* XIX 17).

For hundreds of years to come, debates in political philosophy would surround the relation between 'temporal' political authority and the 'universal Church.' But emphasis on the cosmopolitan aspect of the Church waned, despite its ideal of a religious community comprising all humans. In a nutshell, the debate now opposed the secular and the religious, and not the local and the cosmopolitan. Though cosmopolitanism might have featured prominently in medieval political thought, it did not.

1.2 Early Modern and Enlightenment Cosmopolitanism

Cosmopolitanism slowly began to come to the fore again with the renewed study of more ancient texts, but during the humanist era cosmopolitanism still remained the exception. Despite the fact that ancient cosmopolitan sources were well-known and that many humanists emphasized the essential unity of all religions, they did not develop this idea in cosmopolitan terms. A few authors, however, most notably Erasmus of Rotterdam, explicitly drew on ancient cosmopolitanism to advocate the ideal of a world-wide peace. Emphasizing the unity of humankind over its division into different states and peoples, by arguing that humans are destined by Nature to be sociable and live in harmony, Erasmus pleaded for national and

religious tolerance and regarded like-minded people as his compatriots (*Querela Pacis*).

Early modern natural law theory might seem a likely candidate for spawning philosophical cosmopolitanism. Its secularizing tendencies and the widespread individualist view among its defenders that all humans share certain fundamental characteristics would seem to suggest a point of unification for humankind as a whole. However, according to many early modern theorists, what all individuals share is a fundamental striving for self-preservation, and the universality of this striving does not amount to a fundamental bond that unites (or should unite) all humans in a universal community.

Still, there are two factors that do sometimes push modern natural law theory in a cosmopolitan direction: first, the fact that some natural law theorists assume that nature implanted in humans, in addition to the tendency to self-preservation, *also* a fellow-feeling, a form of sociability that unites all humans at a fundamental level into a kind of world community. The appeal to such a shared human bond was very thin, however, and by no means does it necessarily lead to cosmopolitanism. In fact, the very notion of a natural sociability was sometimes used instead to legitimate war against peoples elsewhere in the world who were said to have violated this common bond in an ‘unnatural’ way, or who were easily said to have placed themselves outside of the domain of common human morality by their ‘barbaric’ customs. Second, early modern natural law theory was often connected with social contract theory, and although most social contract theorists worked out their views mostly, if not solely, for the level of the state and not for that of international relations, the very idea behind social contract theory lends itself for application to this second level. Grotius, Pufendorf, and others did draw out these implications and thereby laid the foundation for international law. Grotius envisioned a “great society of states” that is bound by a “law of nations” that holds “between all states” (*De Iure Belli ac Paci*, 1625, Prolegomena par. 17; Pufendorf, *De Iure Naturae et Gentium*, 1672).

The historical context of the philosophical resurgence of cosmopolitanism during the Enlightenment is made up of many factors: The increasing rise of capitalism and world-wide trade and its theoretical reflections; the reality of ever expanding empires whose reach extended across the globe; the voyages around the world and the anthropological so-called ‘discoveries’ facilitated through these; the renewed interest in Hellenistic philosophy; and the emergence of a notion of human rights and a philosophical focus on human reason. Many intellectuals of the time regarded their membership in the transnational ‘republic of letters’ as more significant than their membership in the particular political states they found themselves in, all the more so because their relationship with their government was often strained because of censorship issues. This prepared them to think in terms other than those of states and peoples and adopt a cosmopolitan perspective. Under the influence of the American Revolution, and especially during the first years of the French Revolution, cosmopolitanism received its strongest impulse. The 1789 declaration of ‘human’ rights had grown out of cosmopolitan modes of thinking and reinforced them in turn.

In the eighteenth century, the terms ‘cosmopolitanism’ and ‘world citizenship’ were often used not as labels for determinate philosophical theories, but rather to indicate an attitude of open-mindedness and impartiality. A cosmopolitan was someone who was not subservient to a particular religious or political authority, someone who was not biased by particular loyalties or cultural prejudice. Furthermore, the term

was sometimes used to indicate a person who led an urbane life-style, or who was fond of traveling, cherished a network of international contacts, or felt at home everywhere. In this sense the Encyclopédie mentioned that ‘cosmopolitan’ was often used to signify a “man of no fixed abode, or a man who is nowhere a stranger.” Though philosophical authors such as Montesquieu, Voltaire, Diderot, Addison, Hume, and Jefferson identified themselves as cosmopolitans in one or more of these senses, these usages are not of much philosophical interest.

Especially in the second half of the century, however, the term was increasingly also used to indicate particular philosophical convictions. Some authors revived the Cynic tradition. Fougeret de Montbron in his 1753 autobiographical report, *Le Cosmopolite*, calls himself a cosmopolitan, describes how he travels everywhere without being committed to anywhere, declaring “All the countries are the same to me” and “[I am] changing my places of residence according to my whim” (p. 130).

Despite the fact that there were only few authors who committed themselves to this kind of cosmopolitanism, this was the version that critics of cosmopolitanism took as their target. For example, Rousseau complains that cosmopolitans “boast that they love everyone [*tout le monde*, which also means ‘the whole world’], to have the right to love no one” (Geneva Manuscript version of *The Social Contract*, 158). Johann Georg Schlosser, in the critical poem ‘Der Kosmopolit’ writes, “It is better to be proud of one’s nation than to have none,” obviously assuming that cosmopolitanism implies the latter.

Yet most eighteenth-century defenders of cosmopolitanism did not recognize their own view in these critical descriptions. They understood cosmopolitanism not as a form of ultra-individualism, but rather, drawing on the Stoic tradition, as implying the positive moral ideal of a universal human community, and they did not regard this ideal as inimical to more particular attachments such as patriotism. Some, like the German author Christoph Martin Wieland, stayed quite close to Stoic views. Others developed a cosmopolitan moral theory that was distinctively new. According to Kant, all rational beings are members in a single moral community. They are analogous to citizens in the political (republican) sense in that they share the characteristics of freedom, equality, and independence, and that they give themselves the law. Their common laws, however, are the laws of morality, grounded in reason. Early utilitarian cosmopolitans like Jeremy Bentham, by contrast, defended their cosmopolitanism by pointing to the “common and equal utility of all nations.” Moral cosmopolitanism could be grounded in human reason, or in some other characteristic universally shared among humans (and in some cases other kinds of beings) such as the capacity to experience pleasure or pain, a moral sense, or the aesthetic imagination. Moral cosmopolitans regarded all humans as ‘brothers’ (though with obvious gender bias) -- an analogy with which they aimed to indicate the fundamental equality of rank of all humans, which precluded slavery, colonial exploitation, feudal hierarchy, and tutelage of various sorts.

Some cosmopolitans developed their view into a political theory about international relations. The most radical of eighteenth-century political cosmopolitans was no doubt Anarcharsis Cloots (Jean-Baptiste du Val-de-Grace, baron de Cloots, 1755-1794). Cloots advocated the abolition of all existing states and the establishment of a single world state under which all human individuals would be directly subsumed. His arguments drew first of all on the general structure of social contract theory. If it is in the general interest for everyone to submit to the authority of a state that enforces laws that provide security, then this

argument applies world-wide and justifies the establishment of a world-wide “republic of united individuals,” not a plurality of states that find themselves in the state of nature vis-à-vis each other. Second, he argues that sovereignty should reside with the people, and that the concept of sovereignty itself, because it involves indivisibility, implies that there can be but one sovereign body in the world, namely, the human race as a whole (*La république universelle ou adresse aux tyrannicides,* 1792; *Bases constitutionelles de la république du genre humain,* 1793).

Most other political cosmopolitans did not go as far as Cloots. Immanuel Kant, most famously, advocated a much weaker form of international legal order, namely, that of a ‘league of nations.’ In *Perpetual Peace* (1795) Kant argues that true and world-wide peace is possible only when states are organized internally according to ‘republican’ principles, when they are organized externally in a voluntary league for the sake of keeping peace, and when they respect the human rights not only of their citizens but also of foreigners. He argues that the league of states should not have coercive military powers because that would violate the internal sovereignty of states, constitute a potential danger to individual freedoms already established within those states (if the federal authority were less respectful of human rights than some of the member states) and reduce the chances that states would actually join.

Some critics argued in response that Kant’s position was inconsistent, on the grounds that the only way to fully overcome the state of nature among states was for them to enter into a federative unity of states with coercive powers. They transformed the concept of sovereignty in the process, by conceiving it as layered, and this enabled them to argue that states ought to transfer part of their sovereignty to the federal level, but only that part that concerns their external relations to other states, while retaining the sovereignty of the states concerning their internal affairs (the early Fichte). Romantic authors, on the other hand, felt that the ideal state should not have to involve coercion at all, and hence also that the cosmopolitan ideal should be that of a world-wide republic of ‘fraternal’ non-authoritarian republics (the young Friedrich Schlegel).

Kant also introduced the concept of "cosmopolitan law," suggesting a third sphere of public law -- in addition to constitutional law and international law -- in which both states and individuals have rights, and where individuals have these rights as "citizens of the earth" rather than as citizens of particular states.

In addition to moral and political forms of cosmopolitanism, there emerged an economic form of cosmopolitan theory. The freer trade advocated by eighteenth-century anti-mercantilists like Adam Smith and Dietrich Hermann Hegewisch took greater and greater hold. They sought to diminish the role of politics in the economic realm. Their ideal was a world in which tariffs and other restrictions on foreign trade are abolished, a world in which the market, not the government, takes care of the needs of the people. Against mercantilism, they argue that it is more advantageous for everyone involved if a nation imports those goods which are more expensive to produce domestically, and that the assumption that one’s own state will profit if other states are unable to export their goods is false. They argue that the situation is quite the contrary: the abolition of protectionism would benefit everyone, because other states would gain from their exports, reach a higher standard of living and then become even better trading partners, because they could then import more, too. On their view, after trade will have been liberalized

world-wide, the importance of national governments will diminish dramatically. As national governments currently focus on the national economy and defense, their future role will be at most auxiliary. In the ideal global market, war is in no one's interest. The freer the global market becomes, the more the role of the states will become negligible.

1.3 Cosmopolitanism in the 19th and 20th Centuries

Enlightenment cosmopolitanism has continued to be a source of debate in the subsequent two centuries. First, in the nineteenth century, economic globalization provoked fierce reactions. Marx and Engels tagged cosmopolitanism as an ideological reflection of capitalism. They regard market capitalism as inherently expansive, breaking the bounds of the nation-state system, as evidenced by the fact that production and consumption had become attuned to faraway lands. In their hands, the word 'cosmopolitan' is tied to the effects of capitalist globalization, including especially the bourgeois ideology which legitimatizes 'free' trade in terms of the freedom of individuals and mutual benefit, although this very capitalist order is the cause of the misery of millions, indeed the cause of the very existence of the proletariat. At the same time, however, Marx and Engels also hold that the proletariat in every country shares essential features and has common interests, and the Communist movement aims to convince proletarians everywhere of these common interests. Most famously, the *Communist Manifesto* ends with the call, "Proletarians of all countries, unite!" This, combined with the ideal of the class-less society and the expected withering away of the state after the revolution, implies a form of cosmopolitanism of its own.

Debates about global capitalism and about an international workers' movement have persisted. Frequently economic cosmopolitanism can be found in the advocacy of open markets, in the tradition from Adam Smith to Friedrich von Hayek and Milton Friedman. Communist versions of cosmopolitanism also developed further, although the Leninist-Stalinist tradition kept using 'cosmopolitan' itself as a derogatory term.

The second inheritance from eighteenth century cosmopolitanism is found in the two centuries' worth of attempts to create peace. It has often been noted that there are parallels between Kant's peace proposal in *Perpetual Peace* and the structure of the League of Nations as it existed in the early part of the 20th century as well as the structure of the current United Nations, although it should also be pointed out that essential features of Kant's plan were not implemented, such as the abolition of standing armies. Now, after the end of the cold war, there is again a resurgence of the discussion about the most appropriate world order to promote peace, just as there was after the first and second world wars.

The International Criminal Court should be mentioned here as an innovative form of cosmopolitanism, going much beyond Kant's conception of 'cosmopolitan law.' The ICC itself represents an extension of the long trend, in international law, to do away with the principle of the absolute subjection of individuals to the state and develop the status of individuals under international law. Individuals are now the bearers of certain rights under international law, and they can be held responsible for crimes under international

law in ways that cut through the shield of state sovereignty.

Third, moral philosophers and moralists in the wake of eighteenth-century cosmopolitanisms have insisted that we human beings have a duty to aid fellow humans in need, regardless of their citizenship status. There is a history of international relief efforts (International Red Cross and Red Crescent Societies, famine relief organizations, and the like) in the name of the reduction of human suffering and without regard to the nationality of those affected.

Cosmopolitan duty is not restricted to duties of beneficence and also requires justice and respect, and cosmopolitan morality has often been invoked as a motivation to oppose slavery and apartheid, and to defend the emancipation of women, or, in the utilitarian tradition, to demand better treatment of animals.

Most past cosmopolitan authors did not fully live up to the literal interpretation of their cosmopolitan theories, and one can find misogynist, racist, nationalist, religious, or class-based biases and inconsistencies in their accounts. These shortcomings have often been used as arguments against cosmopolitanism, but they are not as easily used for that purpose as it may seem. Because the universalist potential in the discourse of 'world citizenship' can itself be used as a basis for exposing these shortcomings as problematic, one should say that they stem from too little, rather than too much, cosmopolitanism.

2. Taxonomy of Contemporary Cosmopolitanisms

Even this brief glance backwards reveals a wide variety of views that can be called cosmopolitan. Every cosmopolitan argues for some community among all human beings, regardless of social and political affiliation. For some, what should be shared is simply moral community, which means only that living a good human life requires serving the universal community by helping human beings as such, and/or by promoting the realization of justice and the guarantee of human rights. Others conceptualize the universal community in terms of political institutions to be shared by all, in terms of cultural expressions to be appreciated by all, or in terms of economic markets that should be open to all.

The most common cosmopolitanism -- *moral* cosmopolitanism -- does not always call itself such. But just as ancient cosmopolitanism was fundamentally a 'moral' commitment to helping human beings as such, much contemporary moral philosophy insists on the duty to aid foreigners who are starving or otherwise suffering, and/or on the duty to respect and promote basic human rights and justice. One can here distinguish between strict and moderate forms of cosmopolitanism. The *strict* cosmopolitans in this sphere operate sometimes from utilitarian assumptions (e.g., Singer, Unger), sometimes from Kantian assumptions (e.g., O'Neill), and sometimes from more ancient assumptions (e.g., Nussbaum), but always with the claim that the duty to provide aid neither gets weighed against any extra duty to help locals or compatriots nor increases in strength when locals or compatriots are in question. Among these strict cosmopolitans some will say that it is permissible, at least in some situations, to concentrate one's charitable efforts on one's compatriots, while others deny this -- their position will depend on the details of their moral theory. Other philosophers whom we may call *moderate* cosmopolitans (including, e.g.,

Scheffler) acknowledge the cosmopolitan scope of a duty to provide aid, but insist that we also have special duties to compatriots. Among the moderate cosmopolitans, many further distinctions can be drawn, depending on the reasons that are admitted for recognizing special responsibilities to compatriots and depending on how the special responsibilities are balanced with the cosmopolitan duties to human beings generally. *Anti-cosmopolitanism* in the moral sphere best describes the position of those communitarians (e.g., MacIntyre) who believe either that our obligations to compatriots and more local people crowd out any obligations to benefit human beings as such or that there are no obligations except where there are close, communal relationships.

This debate about moral cosmopolitanism has sometimes led to an inquiry into *political* cosmopolitanism. Again, we can draw useful distinctions among the political cosmopolitans. Some advocate a centralized world state, some favor a federal system with a comprehensive global body of limited power, some would prefer more limited international political institutions that focus on particular concerns (e.g., war crimes, environmental preservation), and some defend a different alternative altogether. Prominent philosophical discussions of international political arrangements have recently clustered around the self-conscious heirs of Kant (e.g., Habermas, Rawls, Beitz, and Pogge) and around advocates of 'cosmopolitan democracy' (e.g., Held, Bohman). Again, there are anti-cosmopolitans, who are skeptical of all international political entanglements.

Perhaps the most common invocations of the label 'cosmopolitan' in recent philosophical literature have been in the disputes over *cultural* cosmopolitanism. Especially with disputes over multiculturalism in educational curricula and with resurgent nationalisms, cultural claims and counter-claims have received much attention. The cosmopolitan position in both of these kinds of disputes rejects exclusive attachments to parochial culture. So on the one hand, the cosmopolitan encourages cultural diversity and appreciates a multicultural mélange, and on the other hand, the cosmopolitan rejects a strong nationalism. In staking out these claims, the cosmopolitan must be wary about very strong 'rights to culture,' respecting the rights of minority cultures while rebuffing the right to unconditional national self-determination. Hence, recent advocates of 'liberal nationalism' (e.g., Margalit and Raz, Tamir) or of the rights of minority cultures (e.g., Kymlicka) generally seem to be anti-cosmopolitan. But the cosmopolitan's wariness towards very strong rights to culture and towards national self-determination need not be grounded in a wholesale skepticism about the importance of parochial cultural attachments. Cosmopolitanism can acknowledge the importance of (at least some kinds of) cultural attachments for the good human life (at least within certain limits), while denying that this implies that a person's cultural identity should be defined by any bounded or homogeneous subset of the cultural resources available in the world (e.g., Waldron).

Economic cosmopolitanism is perhaps less often defended among philosophers and more often among economists (e.g., Hayek, Friedman) and certain politicians, especially in the richer countries of this world. It is the view that one ought to cultivate a single global economic market with free trade and minimal political involvement. It tends to be criticized rather than advanced by philosophical cosmopolitans, as many of them regard it as at least a partial cause of the problem of vast international economic inequality. These debates about the desirability of a fully globalized market have intensified in recent years, as a result of the end of the Cold War and the increasing reach of the market economy.

3. Objections to Cosmopolitanism

One of the most common objections to cosmopolitanism attacks a position that is in fact made of straw. Often it is said that cosmopolitanism is meaningless without the context of a world-state or that cosmopolitanism necessarily involves the commitment to a world state. These claims are historically uninformed, because cosmopolitanism as a concept arose in the first instance as a metaphor for a way of life and not in literal guise. Ever since, there have been cosmopolitans who do not touch on the issue of international political organization, and of those who do, very few defend the ideal of a world state. Furthermore, even those cosmopolitans who do favor a world-state tend to support something more sophisticated that cannot be dismissed out of hand: a thin conception of world government with layered sovereignty.

The more serious and philosophically interesting challenges to cosmopolitanism come in two main forms. The first calls into question the possibility of realizing the cosmopolitan ideal, while the second queries its desirability. We discuss these two challenges to the different forms of cosmopolitanism in turn.

3.1 Political cosmopolitanism

It is often argued that it is impossible to change the current nation-state system and to form a world-state or a global federation of states. This claim is hard to maintain, however, in the face of the existence of the United Nations, the existence of states with more than a billion people of heterogeneous backgrounds, and the experience with the USA and the EU. So in order to be taken seriously, the objection must instead be that it is impossible to form a *good* state or federation of that magnitude, i.e., that it is impossible to realize or even approximate the cosmopolitan ideal in a way that makes it worth pursuing and that does not carry prohibitive risks. Here political cosmopolitans disagree among themselves. On one end of the spectrum we find those who argue in favor of a strong world-state, on the other end we find the defenders of a loose and voluntary federation, or a different system altogether.

The defenders of the loose, voluntary and noncoercive federation warn that a world state easily becomes despotic without there being any competing power left to break the hold of despotism, and the defenders of the world-state reply that a stronger form of federation, or even merger, is the only way to truly exit the state of nature between states. Other authors have argued that the focus among many political cosmopolitans on only these two alternatives overlooks a third, and that a concern for human rights should lead one to focus instead on institutional reform that disperses sovereignty vertically, rather than concentrating it in all-encompassing international institutions. On this view, peace, democracy, prosperity, and the environment would be better served by a system in which the political allegiance and loyalties of persons are widely dispersed over a number of political units of various sizes, without any one unit being dominant and thus occupying the traditional role of the state (Pogge).

Of the objections brought up by non- or anticosmopolitans, two deserve special mention. First, some authors argue that the (partial or whole) surrender of state sovereignty required by the cosmopolitan

scheme is an undue violation of the principle of the autonomy of states or the principle of democratic self-determination of their citizens. Second, so-called ‘realists’ argue that states are in a Hobbesian state of nature as far as the relations among them are concerned, and that it is as inappropriate as it is futile to subject states to normative constraints. To these objections cosmopolitans have various kinds of response, ranging from developing their alternative normative theory (e.g., by arguing that global democracy increases rather than diminishes the democratic control of individual world citizens) to pointing out, as has been done at least since Grotius, that states have good reasons even on Hobbesian grounds to submit to certain forms of international legal arrangements.

3.2 Economic cosmopolitanism

Various arguments have been used to show that economic cosmopolitanism is not a viable option. Marx and later Marxists have argued that capitalism is self-destructive in the long run, because the exploitation, alienation, and poverty that it inflicts on the proletariat will provoke a world-wide revolution that will bring about the end of capitalism. In the twentieth century, when nationalist tendencies proved to be stronger (or in any case more easily mobilized) than international solidarity, and when the position of workers was strengthened to the point of making them unwilling to risk a revolution, this forced the left to reconsider this view.

Critics of the economic cosmopolitan ideal have also started to emphasize another way in which capitalism bears the seeds of its own destruction within itself, namely, insofar as it is said to lead to a global environmental disaster that might spell the end of the human species, or in any event the end of capitalism as we know it. The effects of excessive consumption (in some parts of the world) and the exploitation of nature would make the earth inhospitable to future human generations.

Even if one does not think that these first two problems are so serious as to make economic cosmopolitanism unviable, they can still make it seem *undesirable* in the eyes of those who are concerned with poverty and environmental destruction.

Moreover, there are several other concerns that lead critics to regard economic cosmopolitanism as undesirable. First among these is the lack of effective democratic control by the vast majority of the world’s population, as large multinationals are able to impose demands on states that are in a weak economic position and their populations, demands that they cannot reasonably refuse to meet, although this does not mean that they meet them fully voluntarily. This concerns, for example, labor conditions or the use of raw materials in so-called Third World countries.

Second, economic cosmopolitans are accused of failing to pay attention to a number of probable side-effects of a global free market. In particular, they are criticized for neglecting or downplaying issues such as (a) the presupposition of large-scale migration or re-schooling when jobs disappear in one area (the loss of ties to friends and family, language, culture, etc., and the monetary costs of moving or re-tooling), (b) the lack of a guarantee that there will be a sufficient supply of living-wage jobs for all world citizens (especially given increasing automation), and (c) the problem of the detrimental effects of income

disparities. They are similarly accused of failing to take seriously the fact that there might be circumstances under which it would be profitable for some states to be protectionist or wage war, such as wars about markets or raw materials and energy (e.g., oil).

3.3 Moral cosmopolitanism

Another version of the criticism that cosmopolitanism is impossible targets the psychological assumptions of moral cosmopolitanism. Here it is said that human beings must have stronger attachments toward members of their own state or nation, and that attempts to disperse attachments to fellow-citizens in order to honor a moral community with human beings as such will cripple our sensibilities. If this is a *viability* claim and not simply a *desirability* claim, then it must be supposed that moral cosmopolitanism would literally leave large numbers of people unable to function. So it is claimed that people need a particular sense of national identity in order to be agents, and that a particular sense of national identity requires attachment to particular others perceived to have a similar identity. This argument seems plausible if it is assumed that cosmopolitanism requires the same attitudes towards all other human beings, but moderate cosmopolitanism does not make that assumption. Rather, the moderate cosmopolitan has to insist only that there is some favorable, motivating attitude toward all human beings as such; this leaves room for some special attitudes towards fellow-citizens. Of course, the strict moral cosmopolitan will go further and will deny that fellow-citizens deserve any special attitudes, and it might be thought that this denial is what flouts the limits of human psychology. But this does not seem to be true as an empirical generalization. The cosmopolitan does not need to deny that some people do happen to have the need for national allegiance, so long as it is true that not all people do; and insofar as some people do, the strict cosmopolitan will say that perhaps it does not need to be that way and that cosmopolitan education might lead to a different result. The historical record gives even the strict cosmopolitan some cause for cheer, as human psychology and the forms of political organization have proven to be quite plastic.

In fact, some cosmopolitans have adopted a developmental psychology according to which patriotism is a step on the way to cosmopolitanism: as human individuals mature they develop ever wider loyalties and allegiances, starting with attachments to their caregivers and ending with allegiance to humanity at large. These different attachments are not necessarily in competition with each other. Just as little as loyalty to one's family is generally seen as a problematic feature of citizens, so the argument goes, loyalty to one's state is not a necessarily problematic feature in the eyes of cosmopolitans. Thus, cosmopolitanism is regarded as an extension of a developmental process that also includes the development of patriotism. This claim is just as much in need of empirical support, however, as the opposite claim discussed in the previous paragraph.

Often, though, the critic's arguments about psychological possibility are actually run together with *desirability* claims. The critic says that the elimination of a special motivating attachment to fellow-citizens is not possible, but the critic means that the elimination of special motivating attachments to fellow-citizens will make a certain desirable form of political life impossible. To respond to this sort of argument, the cosmopolitan has two routes open. First, she can deny the claim itself. Perhaps the viability of politics as usual depends not upon certain beliefs that fellow-citizens deserve more of one's service,

but upon commitments to the polity itself. If strictly cosmopolitan patriotism is a possibility, it lives in a commitment to a universal set of principles embodied in a particular political constitution and a particular set of political institutions. If such commitment is enough for desirable politics, then the anti-cosmopolitan is disarmed. But second, the cosmopolitan can of course also deny the value of the form of political life that is posited as desirable. At this point, moral commitments run over into a discussion of political theory.

Occasionally it is said that cosmopolitans are treasonous or at least unreliable citizens. But many recognizably cosmopolitan theses (that is, the moderate ones) are consistent with loyalty to fellow-citizens, and even the strictest cosmopolitan can justify some forms of service to fellow-citizens when they are an optimal way to do good for human beings (who happen to be fellow-citizens, and not *because* they are fellow-citizens).

This last criticism can be developed further, however, and tailored specifically to target the strict cosmopolitan. If the strict cosmopolitan can justify only some forms of service to fellow-citizens, under some conditions, it might be said that she is blind to other morally required forms or conditions of service to fellow-citizens. At this point, the critic offers reasons why a person has special obligations to compatriots, which are missed by the strict cosmopolitan. Many critics who introduce these reasons are themselves moderate cosmopolitans, wishing to demonstrate that there are special obligations to fellow-citizens in addition to general duties to the community of all human beings. But if these reasons are demanding enough, then there may be no room left for any community with all human beings, and so these objections to strict cosmopolitanism can also provide some impetus toward an anti-cosmopolitan stance. Because there are several such reasons that are frequently proposed, there are, in effect, several objections to the strictly cosmopolitan position, and they should be considered one-by-one.

The first narrow objection to strict cosmopolitanism is that it neglects the obligations of reciprocity. According to this argument, we have obligations to give benefits in return for benefits received, and we receive benefits from our fellow-citizens. The best strictly cosmopolitan response to this argument will insist on a distinction between the state and fellow-citizens and will question exactly who provides which benefits and what is owed in return. On grounds of reciprocity the state may be owed certain things -- cooperative obedience -- and these things may in fact generally benefit fellow-citizens. But the state is not owed these things *because* one owes the fellow-citizens benefits. One does not appropriately signal gratitude for benefits received from the state by, say, giving more to local charities than to charities abroad because charity like this does not address the full agent responsible for the benefits one has received, and does not even seem to be the sort of thing that is commensurate with the benefits received. In assessing this exchange of arguments, there are some significantly difficult questions to answer concerning exactly how the receipt of benefits obliges one to make a return and concerning how the benefits one receives from one’s state affect the acceptability of emigration.

A second objection to strict moral cosmopolitanism gives contractarian grounds for our obligations to fellow-citizens. Because actual agreements to prioritize fellow-citizens as beneficiaries are difficult to find, the contractarians generally rely upon an implicit agreement that expresses the interests or values of the fellow-citizens themselves. So the contractarian argument turns on identifying interests or values that

obligate fellow-citizens to benefit each other. Perhaps, then, it will be argued that citizens have deep interests in what a successful civil society and state can offer them, and that these interests commit the citizens to an implicit agreement to benefit fellow-citizens. The strict cosmopolitan will reply to such an argument with skepticism about what is required for the civil society. Why is more than cooperative obedience required by our interests in what a successful state and civil society can provide? Surely some citizens have to dedicate themselves to working on behalf of this particular society, but why can they not do so on the grounds that this is the best way to benefit human beings as such? Perhaps an intermediate position here is the (Kantian) view that it is morally necessary to establish just democratic states and that just democratic states need some special commitment on the part of their citizens in order to function as democracies, a special commitment that goes beyond mere cooperative obedience but that can still be defended in universalist cosmopolitan terms. The acceptability of this type of view, however, will depend on whether one finds convincing the underlying Kantian political theory.

The final argument for recognizing obligations to benefit fellow-citizens appeals to what David Miller has called ‘relational facts.’ Here the general thought is that certain relationships are constituted by reciprocal obligations: one cannot be a friend or a brother without having certain friendship-obligations or sibling-obligations, respectively. If fellow-citizenship is like these other relations, then we would seem to have special obligations to fellow-citizens. But this argument, which can be found in Cicero’s *De Officiis*, depends upon our intuitions that fellow-citizenship is like friendship or brotherhood and that friendship and brotherhood do come with special obligations, and both intuitions require more argument. Frequently, these arguments appeal to alleged facts about human nature or about human psychology, but these appeals generally raise still further questions.

In sum, a range of interesting and difficult philosophical issues is raised by the disputes between cosmopolitans of various stripes and their critics. As the world becomes a smaller place through increased social, political, and economic contacts, these disputes and the issues they raise will only become more pressing.

Bibliography

Historical works

- Augustine. *De Civitate Dei Libri XXII*. Ed. A. Kalb. Leipzig: B.G. Teubner, 1929. Translated as *The City of God against the Pagans*. Ed. and trans. R.W. Dyson. Cambridge: Cambridge University Press, 1998.
- Bentham, Jeremy. *Principles of International Law*. In *The Works of Jeremy Bentham*, ed. John Bowring, vol. 2, 535-560. New York: Russell & Russell, 1962.
- Chrysippus. See (Stoics).
- Cicero. *De Officiis*. Ed. M. Winterbottom. Oxford: Clarendon Press, 1994. Translated as *On Duties*. Ed. and trans. M.T. Griffin and E.M. Atkins. Cambridge: Cambridge University Press, 1991.
- Cloots, Anacharsis. *Oeuvres*. München: Kraus Reprint, 1980.

- (Cynics). In *Socratis et Socraticorum Reliquiae*, Ed. G. Giannantoni, vol. 2. Naples: Bibliopolis, 1990. For an English translation of the most important source of fragments and testimonia, see Book Six of *Diogenes Laertius, Lives of the Eminent Philosophers*. Trans. R.D. Hicks. 2 vols. Harvard: Harvard University Press, 1925.
- Diogenes the Cynic. See (Cynics).
- *Encyclopédie; ou Dictionnaire raisonné des sciences, des arts et des métiers, par une société des gens de lettres.* Ed. Denis Diderot and Jean Le Rond d'Alembert. Vol. IV, p. 297 (Paris: Briasson, et al., 1754).
- Erasmus, Desiderius. *A Complaint of Peace Spurned and Rejected by the Whole World*. In: Desiderius Erasmus, *Works.* Trans. Betty Radice. Vol. 27, pp. 289-322. Toronto: University of Toronto Press, 1986.
- Fichte, Johann Gottlieb. *Foundations of Natural Right.* Ed. Frederick Neuhouser, trans. Michael Baur. Cambridge: Cambridge University Press, 2000.
- Fougeret de Montbron, *Le Cosmopolite ou le Citoyen du Monde*. Ducros, Paris: 1970 [London, 1750].
- Grotius, Hugo. *The Law of War and Peace. De Iure Belli ac Paci Libri Tres*. [1625] Trans. Francis W. Kelsey. New York: Bobbs-Merrill, 1925.
- Hegewisch, Dietrich Hermann. *Historische und litterarische Aufsätze*. Kiel: Neue akademische Buchhandlung, 1801.
- Kant, Immanuel. *Political Writings*. Ed. by Hans Reiss, trans. H.B. Nisbet, second edition. Cambridge: Cambridge University Press, 1991.
- *Marx: Early Political Writings.* Ed. and trans. Joseph O'Malley with Richard A. Davis. Cambridge: Cambridge University Press, 1994.
- Plato. *Opera Omnia*. 5 vols. Ed. J. Burnet. Oxford: Clarendon Press, 1900-1907. Translations in *Plato: Complete Works*. Ed. John M. Cooper. Indianapolis: Hackett, 1997.
- Pufendorf, Samuel. *De iure naturae et gentium libri octo*. Ed. Walter Simons. Buffalo: Hein, 1995.
- Rousseau, Jean-Jacques. *The Social Contract and Other Later Political Writings*. Ed. and trans. Victor Gourevitch. Cambridge: Cambridge University Press, 1997.
- Schlegel. "Essay on the Concept of Republicanism occasioned by the Kantian tract 'Perpetual Peace'." In *The Early Political Writings of the German Romantics*, ed. and trans. Frederick C. Beiser, 93-112. Cambridge: Cambridge University Press, 1996.
- Schlosser, Johann Georg. "Politische Fragmente." *Deutsches Museum*, February 1777.
- Seneca. *L. Annaei Senecae Dialogorum Libri Duodecim*. Ed. L.D. Reynolds. Oxford: Clarendon, 1977. Partially translated as *Moral and Political Esays*. Ed. and Trans. J.M. Cooper and J.F. Procopé. Cambridge: Cambridge University Press, 1995.
- Seneca. *L. Annaei Senecae Epistulae Morales ad Lucilium*. Ed. L.D. Reynolds. Oxford: Clarendon, 1965. A translation appears with the earlier edition of *Seneca: Epistles*. Ed. and trans. R. Gummere. 3 vols. Cambridge: Harvard University Press, 1917-1925.
- Smith, Adam. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Eds. R. H. Campbell and A. S. Skinner, textual ed. W. B. Todd. Indianapolis: Liberty Classics, 1976.
- (Stoics). *Stoicorum Veterum Fragmenta*. Ed. H. von Arnim (vols. 1-3) and M. Adler (vol. 4). Leipzig: Teubner, 1903-1905, 1924. Some of the fragments and testimonia are translated in A.A.

Long and D.N. Sedley, *The Hellenistic Philosophers, Volume One: Translations of the Principal Sources, with Philosophical Commentary*. Cambridge: Cambridge University Press, 1987. Some of the fragments and testimonia are also translated in *Hellenistic Philosophy: Introductory Readings*. Trans. Brad Inwood and L.P. Gerson. 2nd ed. Indianapolis: Hackett, 1997. For translations of more of the relevant fragments and testimonia, see the secondary literature listed below.

On the History of Cosmopolitanism

- Baldry, H.C. *The Unity of Mankind in Greek Thought*. Cambridge: Cambridge University Press, 1965.
- Brown, Eric. "Hellenistic Cosmopolitanism." in *A Companion to Ancient Philosophy*, ed. Mary Louise Gill and Pierre Pellegrin (Oxford: Blackwell, forthcoming.)
- -----. "Socrates the Cosmopolitan." in *Stanford Agora: An Online Journal of Legal Perspectives* 1 (2000).
- -----. *Stoic Cosmopolitanism*. forthcoming.
- Heater, Derek. *World Citizenship and Government: Cosmopolitan Ideas in the History of Western Political Thought*. New York: St. Martin's, 1996.
- Heuvel, Gerd van den. "Cosmopolite, Cosmopolitisme." In *Handbuch politisch-sozialer Grundbegriffe in Frankreich 1680-1820*, ed. Rolf Reichardt and Eberhard Schmidt, 41-55. München: Oldenbourg, 1986.
- Meinecke, Friedrich. *Cosmopolitanism and the National State*. Trans. Robert B. Kimber. Princeton: Princeton University Press, 1970.
- Moles, J.L. "Cynic Cosmopolitanism." In *The Cynics: The Cynic Movement in Antiquity and Its Legacy*, ed. R. Bracht Branham and Marie-Odile Goulet-Cazé, 105-120. Berkeley and Los Angeles: University of California Press, 1996.
- -----. "The Cynics." In *The Cambridge History of Greek and Roman Political Thought*, ed. Christopher Rowe and Malcolm Schofield, 415-434. Cambridge: Cambridge University Press, 2000.
- -----. "The Cynics and Politics." In *Justice and Generosity*, ed. André Laks and Malcolm Schofield, *Justice and Generosity: Studies in Hellenistic Social and Political Philosophy*, 129-158. Cambridge: Cambridge University Press, 1995.
- Nussbaum, Martha C. "Kant and Cosmopolitanism." In *Perpetual Peace: Essays on Kant's Cosmopolitan Ideal*, ed. James Bohman and Matthias Lutz-Bachmann, 25-57. Cambridge: MIT, 1997.
- Schlereth, Thomas J. *The Cosmopolitan Ideal in Enlightenment Thought: Its Form and Function in the Ideas of Franklin, Hume, and Voltaire, 1694-1790*. Notre Dame: University of Notre Dame Press, 1977.
- Schofield, Malcolm. *The Stoic Idea of the City*. Cambridge: Cambridge University Press, 1991.

On the Taxonomy of Cosmopolitanisms

- Kleingeld, Pauline. “Six Varieties of Cosmopolitanism in Late Eighteenth-Century Germany.” *Journal of the History of Ideas* 60 (1999): 505-524.
- Scheffler, Samuel. “Conceptions of Cosmopolitanism.” *Utilitas* 11 (1999): 255-276.

On Contemporary Cosmopolitanisms, For and Against

- Beitz, Charles R. “Cosmopolitan Ideals and National Sentiment.” *Journal of Philosophy* 80 (1983): 591-600.
- -----. *Political Theory and International Relations*. Princeton: Princeton University Press, 1979.
- Bohman, James. “Cosmopolitan Republicanism.” *The Monist* 84 (2001): 3-22.
- Cheah, Pheng, and Bruce Robbins, eds. *Cosmopolitics: Thinking and Feeling Beyond the Nation*. Minneapolis: University of Minnesota Press, 1998.
- Couture, Jocelyne, et al., eds. *Rethinking Nationalism*. *Canadian Journal of Philosophy* s.v. 22 (1996).
- Gewirth, Alan. “Ethical Universalism and Particularism.” *Journal of Philosophy* 85 (1988): 283-302.
- Gilbert, Margaret. “Group Membership and Political Obligation.” *Monist* 76 (1993): 119-131.
- Goodin, R.E. *Protecting the Vulnerable: A Reanalysis of Our Social Responsibilities*. Chicago, IL: University of Chicago Press, 1985.
- -----. “What is So Special about Our Fellow Countrymen?” *Ethics* 98 (1988): 663-687.
- Habermas, Jürgen. “Kant’s Idea of Perpetual Peace, with the Benefit of Two Hundred Years’ Hindsight.” In *Perpetual Peace: Essays on Kant’s Cosmopolitan Ideal*, ed. James Bohman and Matthias Lutz-Bachmann, 113-53. Cambridge: MIT Press, 1997.
- Held, David. *Democracy and the Global Order: From the Modern State to Cosmopolitan Governance*. Stanford: Stanford University Press, 1995.
- Kleingeld, Pauline. “Kantian Patriotism.” *Philosophy & Public Affairs* 29 (2000): 313-341.
- Kymlicka, Will. *Multicultural Citizenship: A Liberal Theory of Minority Rights*. Oxford: Oxford University Press, 1995.
- MacIntyre, Alasdair. “Is Patriotism a Virtue?” In *Theorizing Citizenship*, ed. Ronald Beiner, 209-228. Albany: State University of New York Press, 1995.
- McKim, Robert, and Jeff McMahan, eds. *The Morality of Nationalism*. Oxford: Oxford University Press, 1997.
- Margalit, Avishai, and Joseph Raz. “National Self-Determination.” *Journal of Philosophy* 87 (1990): 439-61.
- Mason, Andrew. “Special Obligations to Compatriots,” Ethics 107 (1997): 427-447.Miller, David. *On Nationality*. Oxford: Oxford University Press, 1995.
- Miller, Richard W. “Cosmopolitan Respect and Patriotic Concern.” *Philosophy & Public Affairs* 27 (1998): 202-224.
- Nathanson, Stephen. *Patriotism, Morality, and Peace*. Lanham, Maryland: Rowman & Littlefield, 1993.
- O’Neill, Onora. *Bounds of Justice*. Cambridge: Cambridge University Press, 2000.
- Nussbaum, Martha C., et al. *For Love of Country: Debating the Limits of Patriotism*. Ed. Joshua Cohen. Boston: Beacon Press, 1996. revised version of *The Boston Review* 19,5 (Oct/Nov 1994).

- Pogge, Thomas W. “Cosmopolitanism and Sovereignty.” *Ethics* 103 (1992): 48-75
- -----. *Realizing Rawls*. Ithaca: Cornell University Press, 1989.
- Rawls, John. *The Law of Peoples.* Cambridge: Harvard University Press, 1999.
- Scheffler, Samuel. “The Conflict between Justice and Responsibility.” In *NOMOS XLI: Global Justice*, ed. L. Brilmayer and I. Shapiro. New York: NYU Press, forthcoming.
- -----. “Liberalism, Nationalism, and Egalitarianism.” In *The Morality of Nationalism*, ed. McKim and McMahan, 191-208.
- -----. “Relationships and Responsibilities.” *Philosophy & Public Affairs* 26 (1997): 189-209.
- Shue, Henry. *Basic Rights: Subsistence, Affluence, and U.S. Foreign Policy*. 2nd ed. Princeton: Princeton University Press, 1996.
- Singer, Peter. *Practical Ethics*. 2nd ed. Cambridge: Cambridge University Press, 1993.
- Tamir, Yael. *Liberal Nationalism*. Princeton: Princeton University Press, 1993.
- Unger, Peter. *Living High and Letting Die: Our Illusion of Innocence*. Oxford: Oxford University Press, 1996.
- Waldron, Jeremy. “Minority Cultures and the Cosmopolitan Alternative.” *University of Michigan Journal of Law Reform* 25 (1992): 751-93.
- -----. “Special Ties and Natural Duties.” *Philosophy & Public Affairs* 22 (1993): 3-30.

Other Internet Resources

[Please contact the authors with suggestions.]

Related Entries

Augustine, Saint | Chrysippus | Cicero | citizenship | communitarianism | Cynics, ancient | Diogenes of Sinope | Grotius, Hugo | justice: international | Kant, Immanuel | Marxism | nationalism | obligations: special | political philosophy: history of | rights: human | Socrates | world government/state

Pauline Kleingeld
pkleinge@artsci.wustl.edu
and
Eric Brown
eabrown@twinearth.wustl.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 22, 2002
Content last modified: June 28, 2002

Stanford Encyclopedia of Philosophy

War

One apt definition of war is this: war is an actual, intentional and widespread armed conflict between political communities. Thus, a fisticuffs between individual persons does not count as a war, nor does a gang fight, nor does a feud on the order of the Hatfields versus the McCoys. War is a phenomenon which occurs only between political communities, defined as those entities which either are states or intend to become states (in order to allow for civil war). Similarly, the mere threat of war and the presence of mutual disdain between these communities do not suffice as indicators of war. The conflict of arms must be actual and not merely latent. Further, the actual armed conflict must be both intentional and widespread: isolated clashes between rogue officers, or border patrols, do not count as actions of war. The onset of war requires a conscious commitment, and a significant mobilization, on the part of the belligerent communities in question.

Perhaps it would be appropriate here to cite the legendary Carl von Clausewitz. Clausewitz famously suggested that "war is the continuation of policy by other means." As a description, this conception is both powerful and plausible. It fits in nicely with his own general definition of war as "an act of violence intended to compel our opponent to fulfil our will." War, he says, is like a duel, but on "an extensive scale." As Michael Gelven has written recently, in an elegant monograph on how we ought to conceive of the essence of war, war is intrinsically vast, communal (or political) and violent. It is a widespread and deliberate armed conflict between political communities. It is the entity or phenomenon falling under such a description which is the primary focus of this entry.

The Ethics of War and Peace

One of the most enduring, and difficult, philosophical questions with regard to war focuses on the ethics of getting involved with it in the first place. It is most helpful, in ordering one's thoughts about this issue, to realize that there are three traditions of thought which dominate the philosophical treatment of this topic. This is not necessarily to imply that these three traditions exhaust all possible options for thinking about the ethics of war and peace, merely to note that they are hegemonic and importantly different from each other. But very few theories on the ethics of war succeed in resisting ultimate classification into one of them. The three traditions are: JUST WAR THEORY; REALISM; and PACIFISM.

Before spending some time discussing the core aspects of each tradition, let's declare, right from the start, the core conceptual differences between "the big three" perspectives. The core, and controversial, proposition of just war theory is that, sometimes, states can have moral justification for resorting to armed force in the international system. War is sometimes, but of course not all the time, morally right. The idea here is not that the war in question is merely politically shrewd, or prudent, or bold and daring, but fully moral, just. It is an ethically appropriate use of mass political violence. Realism, by contrast, sports a profound skepticism about the application of moral concepts, such as justice, to the key problems of foreign policy. Power and national security, realists claim, motivate states during wartime and thus moral appeals are strictly wishful thinking. Talk of the morality of warfare is pure bunk: ethics has got nothing to do with the rough-and-tumble world of global politics, where only the strong and cunning survive. Pacifism does not share realism's moral skepticism. For the pacifist, moral concepts can indeed be applied fruitfully to international affairs. It does make sense to ask whether a war is just. But the result of such normative application, in the case of war, is always that war should not be resorted to. Where just war theory is sometimes permissive with regard to war, pacifism is always prohibitive. For the pacifist, war is always wrong. Now let's turn to each of these three traditions.

Just War Theory

Just war theory is probably the most influential perspective on the ethics of war and peace. The just war tradition has enjoyed a long and distinguished pedigree, including such notables as Augustine, Aquinas, Grotius, Suarez, Vattel and Vitoria. Hugo Grotius probably deserves credit for being the most comprehensive and formidable member of the tradition; and James T. Johnson is the authoritative historian of this tradition. Many of the rules developed by the just war tradition have since been codified into contemporary international laws governing armed conflict, such as The Hague and Geneva Conventions. The tradition has thus been doubly influential, dominating both moral and legal discourse surrounding war. It sets the tone, and the parameters, for the great debate.

Just war theory can be meaningfully divided into three parts, which in the literature are referred to, for the sake of convenience, in Latin. These parts are: 1) *jus ad bellum*, which concerns the justice of resorting to war in the first place; 2) *jus in bello*, which concerns the justice of conduct within war, after it has begun; and 3) *jus post bellum*, which concerns the justice of peace agreements and the termination phase of war.

Jus ad bellum

The rules of *jus ad bellum* are addressed, first and foremost, to heads of state. Since political leaders are the ones who inaugurate wars, setting their armed forces in motion, they are to be held accountable to *jus ad bellum* principles. If they fail in that responsibility, then they commit war crimes. In the language of the Nuremberg prosecutors, aggressive leaders who launch unjust wars commit "crimes against peace." What constitutes a just or unjust resort to armed force is disclosed to us by the rules of *jus ad bellum*. Just war theory contends that, for any resort to war to be justified, a political community, or state, must fulfil each and every one of the following six requirements:

1. Just cause. A state may launch a war only for the right reason. The just causes most frequently mentioned include: self-defence from external attack; the protection of innocents; and punishment for wrongdoing. Vitoria suggested that all of the proffered just causes be subsumed under the one category of "a wrong received." Walzer, and most modern just war theorists, speak of the one just cause for resorting to war being the resistance of aggression. Aggression, simply put, is unjustified and harmful violence.

The key principle underlying just cause, and just war theory more broadly, is the vindication of fundamental rights and the protection of those who have such rights from serious, standard threats to them, such as aggression. Self-defence, and other-defence, from rights violating aggression are thus prime just causes for resorting to war. These rights are traditionally understood as the rights of states to political sovereignty and territorial integrity: states have the right to make their own political decisions for their own people, within their own borders. Only if these rights are violated - for instance, through an armed invasion across the border - is a country justified in resorting to a war of self-defence in response. Other countries may join the war on the victim's side, since the aggressor forfeits its state rights when it violates the victim's.

But what grounds the importance of these state rights? States have state rights, to things like sovereignty and integrity, only because their individual citizens have human rights. People create, and adhere to, state structures in order to secure the objects of their human rights. Human rights are elemental entitlements we all have to basic human dignity and to the objects of vital human need. The human rights most broadly endorsed are those to life, liberty and subsistence, for instance as enshrined in the United Nation's Universal Declaration and subsequent International Covenants.

Following John Rawls, we might establish criteria of minimal justice (MJ) which a state must fulfil if it is to be entitled to state rights: MJ 1) it is able to rule its people in accord with law and order; MJ 2) it provides its people with secure access to the objects of their human rights; and MJ 3) it adheres to basic norms of international justice, notably respect for the rights of persons and other minimally just states. Thus, a state which commits aggression against the people of another country violates principle MJ 3, and thus fails to be minimally just. A minimally just state forfeits its right not to be dealt with harshly, as a matter of appropriate punishment and rectification.

2. Right intention. A state must intend to fight the war only for the sake of a just cause. Having the right reason for launching a war is not enough: the actual motivation behind the resort to war must also be morally appropriate. Ulterior motives, such as a power or land grab, or irrational motives, such as

revenge or ethnic hatred, are ruled out. The only right intention allowed is to see the just cause for resorting to war secured and consolidated. If another intentions crowd in, moral corruption sets in.

3. Proper authority and public declaration. A state may go to war only if the decision has been made by the appropriate authorities, according to the proper process, and made public, notably to its own citizens and to the enemy state(s).

4. Last Resort. A state may resort to war only if it has exhausted all plausible, peaceful alternatives to resolving the conflict in question, in particular diplomatic negotiation. One wants to make sure something as momentous and serious as war is declared only when it seems the only reasonable alternative to effectively punish aggression.

5. Probability of Success. A state may not resort to war if it can foresee that doing so will have no measurable impact on the situation. The aim here is to block mass violence which is going to be futile.

6. (Macro-) Proportionality. A state must, prior to initiating a war, weigh the universal goods expected to result from it, such as securing the just cause, against the universal evils expected to result, notably casualties. Only if the benefits are proportional to, or "worth", the costs may the war action proceed.

Just war theory insists all six criteria must each be fulfilled for a particular declaration of war to be justified: it's all or no justification, so to speak. It is important to note that the first three of these six rules are what we might call deontological requirements, otherwise known as duty-based requirements or first-principle requirements. For a war to be just, some core duty must be violated: in this case, the duty not to commit aggression. A war in punishment of this violated duty must itself respect further duties: it must be appropriately motivated, and must be publicly declared by (only) the proper authority for doing so. The next three requirements are consequentialist: given that these first principle requirements have been met, we must also consider the expected consequences of launching a war which seems justified according to first principles. Thus, just war theory attempts to provide a common sensical combination of both deontology and consequentialism as applied to the issue of war.

Jus in bello

Jus in bello refers to justice in war, to right conduct in the midst of battle. Responsibility for state adherence to *jus in bello* norms falls primarily on the shoulders of those military commanders, officers and soldiers who formulate and execute the war policy of a particular state. They are to be held responsible for any breach of the principles which follow below. Such accountability may involve being put on trial for war crimes.

Just war theorists insist that *jus in bello* is a category separate from *jus ad bellum*. For even if a state has resorted to war justly, it may be prosecuting that war in an unjustified manner. It may be deploying decrepit means in pursuit of its otherwise justified end. Just war theory insists on a fundamental moral consistency between means and ends with regard to wartime behaviour: justified ends may only be

pursued through justified means.

Concern with consistency, however, is not the only, or even the main, reason behind the endorsement of separate rules regulating wartime conduct. Such rules are also required to limit warfare, to prevent it from spilling over into an ever-escalating, and increasingly destructive, experiment in total warfare. If just wars are limited wars, designed to secure their just causes with only proportionate force, the need for rules on wartime restraint is clear. Even though modern warfare has displayed a disturbing tendency towards totality - particularly during the two World Wars - it does not follow that the death of old-time military chivalry marks the end of moral judgment. We still hold soldiers to certain standards of conduct.

There are three widely recognized rules of *jus in bello*.

1. Discrimination. Soldiers are only entitled to target those who are, in Walzer's words, "engaged in harm." Thus, when they take aim, soldiers must discriminate between the civilian population, which is morally immune from direct and intentional attack, and those legitimate military, political and industrial targets involved in rights-violating harm. While some collateral civilian casualties are excusable, it is wrong to take deliberate aim at civilian targets. An example would be saturation bombing of residential areas.

2. (Micro-) Proportionality. Soldiers may only use force proportional to the end they seek. Weapons of mass destruction, for example, are usually seen as being out of proportion to legitimate military ends.

3. No Means *Mala in Se*. Soldiers may not use weapons or methods which are "evil in themselves." These include: mass rape campaigns; genocide or ethnic cleansing; torturing captured enemy soldiers; and using weapons whose effects cannot be controlled, like chemical or biological agents.

Jus post bellum

Jus post bellum refers to justice during the third and final stage of war: that of war termination. It seeks to regulate the ending of wars, and ease the transition from war back to peace. It is one of the most recent, and topical, issues in just war theory. See Orend's works in the bibliography below for more. Orend proposes the following rules for *jus post bellum*:

1. Just cause for termination. A state has just cause to seek termination of the just war in question if there has been a reasonable vindication of those rights whose violation grounded the resort to war in the first place. Not only have most, if not all, unjust gains from aggression been eliminated and the objects of the victim's rights been reasonably restored, but the aggressor is now willing to accept terms of surrender which include not only the cessation of hostilities, a formal apology and its renouncing the gains of its aggression but also its submission to reasonable principles of punishment, including compensation, war crimes trials, and perhaps rehabilitation.

2. Right intention. A state must intend to carry out the process of war termination only in terms of those

principles contained in the other *jus post bellum* rules. Revenge is strictly ruled out as an animating force. Moreover, the just state in question must commit itself to symmetry and equal application with regard to the investigation and prosecution of any war crimes its own armed forces may have committed on the battlefield.

3. Public declaration and legitimate authority. The terms of the peace must be publicly proclaimed by a legitimate authority, which is to say the national government of the state victimized by the initial aggression, or perhaps an authorized international body.

4. Discrimination. In setting the terms of the peace, the just and victorious state is to differentiate between the political and military leaders, the soldiers and the civilian population within the aggressor. Undue and unfair hardship is not to be brought upon the civilian population in particular: punitive measures are to be focused upon those elites most responsible for the aggression.

5. Proportionality. Any terms of peace must be proportional to the end of reasonable rights vindication. Absolutist crusades against, and/or draconian punishments for, aggression are especially to be avoided. The people of the defeated aggressor never forfeit their human rights, and so are entitled not to be "blotted out" from the community of nations. There is thus no such thing as a morally-mandated unconditional surrender.

Any serious defection from these principles of *jus post bellum*, on the part either of the victim or the aggressor, is a violation of the rules of just war and so should be punished. At the very least, such violation of *jus post bellum* mandates a new round of good-faith diplomatic negotiations - perhaps even binding international arbitration - between the relevant parties to the dispute. At the very most, such violation gives the aggrieved party a just cause - but no more than a just cause - for resuming hostilities. Full recourse to the resumption of hostilities may be made only if all the other criteria of *jus ad bellum* are satisfied in addition to just cause.

Just war theory thus offers rules to guide decision-makers on the appropriateness of their conduct during the resort to war, conduct during war and the termination phase of the conflict. Its over-all aim is to try and ensure that wars are begun only for a very narrow set of truly defensible reasons, that when wars break out they are fought in a responsibly controlled and targeted manner, and that the parties to the dispute bring their war to an end in a speedy and responsible fashion that respects the requirements of justice.

Realism

Realism is most influential amongst political scientists, as well as scholars and practitioners of international relations. While realism is a complex and often sophisticated doctrine, its core propositions express a strong suspicion about applying moral concepts, like justice, to the conduct of international affairs. Realists believe that moral concepts should be employed neither as descriptions of, nor as prescriptions for, state behaviour on the international plane. Realists emphasize power and security

issues, the need for a state to maximize its expected self-interest and, above all, their view of the international arena as a kind of anarchy, in which the will to power enjoys primacy.

Referring specifically to war, realists believe that it is an intractable part of an anarchical world system; that it ought to be resorted to only if it makes sense in terms of national self-interest; and that, once war has begun, a state ought to do whatever it can to win. In other words, "all's fair in love and war." During the grim circumstances of war, "anything goes." So if adhering to the rules of just war theory, or international law, hinders a state during wartime, it should disregard them and stick steadfastly to its fundamental interests in power and security. Prominent classical realists often mentioned include Thucydides, Machiavelli and Hobbes. Modern realists include Hans Morgenthau, George Kennan, Reinhold Niebuhr and Henry Kissinger, as well as so-called neo-realists, such as Kenneth Waltz.

It is important to distinguish between descriptive and prescriptive realism. Descriptive realism is the claim that states, as a matter of fact, either do not (for reasons of motivation) or cannot (for reasons of competitive struggle) behave morally, and thus moral discourse surrounding interstate conflict is empty, the product of a category mistake. States are simply not animated in terms of morality and justice: it's all about power, security and national interest for them. States are not like "big persons": they are creations of an utterly different kind, and we cannot expect them to live by the same rules and principles we require of individual persons. States inhabit a violent international arena, and they've got to be able to get in that game and win, if they are to serve and protect their citizens in an effective way over time. Morality is simply not on the radar screen for creations such as states, given their defensive function and the brutal environment in which they subsist.

Walzer offers arguments against this kind of realism, contending that states are in fact responsive to moral concerns, even when they fail to live up to them. States, because they are the creation of individual persons, want to act morally and justly. Walzer goes so far as to say that any state which was motivated by nothing more than the struggle to survive and win power could not over time sustain the support from its own population, which demands a deeper sense of community and justice. He also argues that all the pretence regarding "the necessity" of state conduct in terms of pursuing power is exaggerated and rhetorical, ignoring the clear reality of foreign policy choice enjoyed by states in the global arena. States are not frequently forced into some kind of dramatic, do-or-die struggle: the choice to go to war is a deliberate one, freely entered into and often hotly debated and agonized over before the decision is made. And this is leaving unspoken the argument regarding the defiant, Machiavellian amorality behind certain kinds of realism, and the moral calibre of the actions it might recommend on this basis. For example, if it's all about power and winning in the competitive struggle, does that make it alright to unleash weapons of mass destruction? Or to launch a mass rape campaign? Just war theory suggests not, and just war theorists like Walzer want to claim that the rest of us agree.

Prescriptive realism, though, need not be rooted in any form of descriptive realism. Prescriptive realism is the claim that a state ought (prudential "ought") to behave amorally in the international arena. A state should, for prudence's sake, adhere to an amoral policy of smart self-regard in international affairs. A smart state will leave its morality at home when considering what to do on the international stage. It's important to note that a prescriptive realist might, in the end, actually endorse rules for the regulation of

warfare, much like those offered by just war theory. These rules include: "Wars should only be fought in response to aggression"; and "During war, non-combatants should not be directly targeted with lethal violence." Of course, the reason why a prescriptive realist might endorse such rules would be very different from the reasons offered by the just war theorist: the latter would talk about abiding moral values whereas the former would refer to useful rules which help establish expectations of behaviour, solve coordination problems and to which prudent bargainers would consent. Just war rules, the prescriptive realist might claim, do not have independent moral purchase on the attention of states. These rules are what Douglas Lackey calls "salient equilibria", stable conventions limiting war's destructiveness which all prudent states can agree on, assuming general compliance. There might even be some room for overlap between this kind of realism and just war theory.

Pacifism

It seems best to rely on Jenny Teichman's definition of pacifism as "anti-war-ism." Literally and straightforwardly, a pacifist rejects war in favour of peace. It is not violence in all its forms that the most challenging kind of pacifist objects to; rather, it is the specific kind and degree of violence that war involves which the pacifist objects to. A pacifist objects to killing (not just violence) in general and, in particular, she objects to the mass killing, for political reasons, which is part and parcel of the wartime experience. So, a pacifist rejects war; she believes that there are no moral grounds which can justify resorting to war. War, for the pacifist, is always wrong.

Mention should straight away be made of a very popular just war criticism of pacifism which will not be used here. This criticism is that pacifism amounts to an indefensible "clean hands policy." The pacifist, it is said, refuses to take the brutal measures necessary for the defense of himself and his country, for the sake of maintaining his own inner moral purity. It is contended that the pacifist is thus a kind of free-rider, gathering all the benefits of citizenship while not sharing all its burdens. Another inference drawn is that the pacifist himself constitutes a kind of internal threat to the over-all security of his state.

This "clean hands" argument is easily, and frequently, over-stated. It is important to note that, to the extent to which any moral stance will commend a certain set of actions or intentions deemed morally worthy, and condemn others as being reprehensible, the "clean hands" criticism can be so malleable as to apply to nearly any substantive doctrine. Every moral and political theory stipulates that one ought to do what it deems good or just and to avoid what it deems bad or unjust. So this popular just war criticism of pacifism is not strong. The very idea of a selfish pacifist simply does not ring true: many pacifists have, historically, paid a very high price for their pacifism during wartime (through severe ostracism and even jail time) and their pacifism seems less rooted in regard for inner moral purity than it is in regard for constructing a less violent and more humane world order. So, this argument against pacifism fails; but what of others?

Walzer, the just war theorist, contends that pacifism's idealism is excessively optimistic. In other words, pacifism lacks realism. More precisely, the nonviolent world imagined by the pacifist is not actually attainable, at least for the foreseeable future. Since "ought implies can", the set of "oughts" we are

committed to must express a moral outlook on war less utopian in nature. While we are committed to morality in wartime, we are forced to concede that, sometimes in the real world, resorting to war can be morally justified.

Another objection to pacifism is that, by failing to resist international aggression with effective means, it ends up rewarding aggression and failing to protect people who need it. Pacifists reply to this argument by contending that we do not need to resort to war in order to protect people and punish aggression effectively. In the event of an armed invasion by an aggressor state, an organized and committed campaign of non-violent civil disobedience - perhaps combined with international diplomatic and economic sanctions - would be just as effective as war in expelling the aggressor, with much less destruction of lives and property. After all, the pacifist might say, no invader could possibly maintain its grip on the conquered nation in light of such systematic isolation, non-cooperation and non-violent resistance. How could it work the factories, harvest the fields, or run the stores, when everyone would be striking? How could it maintain the will to keep the country in the face of crippling economic sanctions and diplomatic censure from the international community? And so on.

Though one cannot exactly disprove this pacifist proposition - since it is a counter-factual thesis - there are powerful reasons to agree with John Rawls that such is "an unworldly view" to hold. For, as Walzer points out, the effectiveness of this campaign of civil disobedience relies on the scruples of the invading aggressor. But what if the aggressor is brutal, ruthless? What if, faced with civil disobedience, the invader "cleanses" the area of the native population, and then imports its own people from back home? What if, faced with economic sanctions and diplomatic censure from a neighbouring country, the invader decides to invade it, too? We have some indication from history, particularly that of Nazi Germany, that such pitiless tactics are effective at breaking the will of people to resist. The defence of our lives and rights may well, against such invaders, require the use of political violence. Under such conditions, Walzer says, adherence to pacifism would amount to "a disguised form of surrender."

Pacifists respond to this accusation of "unworldliness" by citing what they believe are real world examples of effective non-violent resistance to aggression. Examples mentioned include Mahatma Ghandi's campaign to drive the British Imperial regime out of India in the late 1940s and Martin Luther King Jr.'s civil rights crusade in the 1960s on behalf of African-Americans. Walzer replies curtly that there is no evidence that non-violent resistance has ever, of itself, succeeded. This may be rash on his part, though it is clear that Britain's own exhaustion after WWII, for example, had much to do with the evaporation of its Empire. Walzer's main counter-argument against these pacifist counter-examples is that they only underline his main point: that effective non-violent resistance depends upon the scruples of those it is aimed against. It was only because the British and the Americans had some scruples, and were moved by the determined idealism of the non-violent protesters, that they acquiesced to their demands. But aggressors will not always be so moved. A tyrant like Hitler, for example, might interpret non-violent resistance as weakness, deserving contemptuous crushing. "Non-violent defense", Walzer suggests, "is no defense at all against tyrants or conquerors ready to adopt such measures."

As sensible as Walzer's remarks might seem, they remain quite narrow, by no means constituting an all-things-considered refutation of pacifism. Generally, there are two kinds of modern secular pacifism to

consider: 1) a more consequentialist form of pacifism (or CP), which maintains that the benefits accruing from war can never outweigh the costs of fighting it; and 2) a more deontological form of pacifism (or DP), which contends that the very activity of war is intrinsically wrong, since it violates foremost duties of justice, such as not killing human beings. Most common amongst contemporary secular pacifists, such as Robert Holmes, is a doctrine which attempts to combine both CP and DP. (I might add, parenthetically, that no discussion will be made here as to religious forms of pacifism. While they have been very influential historically, especially their Christian variants, as theoretical propositions I believe they rest on core premises which are too contentious and exclusionary. But the Christian pacifist literature is a very rich source of information for those interested.)

What arguments might a just war theorist employ to overcome CP and DP? A just war theorist might, for starters, focus on the relationship in CP between consequentialism and the denial of killing. Pacifism in either form places overriding value on respecting human life, notably through its injunction against killing. But this value seems to rest uneasily with consequentialism, for there is nothing inherent to consequentialism which bans killing as such. There is no absolute rule, or side-constraint, that one ought never to kill another person, or that nations ought never to deploy lethal armed force in war. With consequentialism, it's always a matter of considering the latest costs and benefits, of choosing the best option amongst feasible alternatives. Consequentialism therefore leaves conceptual space open to the claim that under these conditions, at this time and place, and given these alternatives, killing and/or war appears permissible. After all, what if killing x people (say, soldiers in an aggressive army) appears the best option if we are to save the lives of x + n people (say, fellow citizens who would perish under the brutal heel of an unchecked aggressor)? It is at least conceivable that a quick and decisive resort to war could prevent even greater killing and devastation in the future. So it seems problematic for the consequentialist pacifist, whose principles exhibit a profound abhorrence for killing people, to be willing in such a scenario to allow an even greater number of people to be killed by acquiescing to the violence of others less scrupulous. These are two telling points: CP does not, of itself, ground the categorical rejection of killing and war which is the essence of pacifism; and CP is open to counter-examples which question whether consequentialism would reject killing and war at all under certain conditions. Consequentialism might even, in a particular case, go so far as to recommend war under certain conditions.

Casting doubt on DP is a complicated procedure. Only a sketch of plausible just war theory arguments can here be offered. The first question to ask is: which foremost duty does DP understand being violated by warfare? If the DP response is the duty not to kill another human being, then contention can be made that this is by no means uncontroversial. Consider the most obvious counter-example: aggressor A attacks B for no defensible reason, posing a serious threat to B's life. Some would suggest, in good faith, that B is not duty-bound not to kill A if such seems necessary to stop A's aggression. Indeed, they would argue that B may kill A in legitimate self-defence. The DP pacifist, however, might reply that extending B moral permission to kill A, even in self-defence, violates the human rights of A. He might contend that just war theory merely compounds the wrongness of the situation by paradoxically permitting lethal force to stop lethal force.

One just war theory rejoinder to this DP contention is this: B does no wrong whatsoever - violates no

human rights - by responding to A's aggression with lethal force if required. Why does B do nothing wrong? First, it is A who is responsible for forcing B to choose between her own life and rights and those of A. We can hardly blame B for choosing her own. For if she does not choose her own, she loses an enormous amount, perhaps everything. And it is patently unreasonable to expect creatures like us to suffer catastrophic loss by default. Consider also the issue of fairness: if B is not allowed to use lethal force, if necessary, against A in the event of A's aggression, then B loses everything while A loses nothing. Indeed, A gains whatever object he desired in violating or killing B. Such would seem an unfair reward of awful behaviour. Finally, B's having rights at all provides her with an implicit entitlement to use those means necessary to secure her rights, including the use of force in the face of a serious physical threat. These powerful considerations of responsibility, reasonableness, fairness and implicit entitlement come together in support of the just war claims that: B may respond with needed lethal force to A's initial aggression; B does no wrong in doing so; it would be wrong to prohibit B's doing so; and that A bears all of the blame for the situation.

DP pacifists are not, at this point, out of options. Holmes, for example, suggests that the foremost duty of justice violated by war is not the duty not to kill aggressors, but rather the duty not to kill innocent, non-aggressive human beings. To be innocent here means to have done nothing which would justify being harmed or killed; in particular, it means not constituting a serious threat to the lives and rights of other people. It is this sense of innocence that just war theory invokes when it claims that civilians should not be directly attacked during wartime. Even if civilians support the war effort politically, or even in terms of their personal attitudes towards the war, they clearly do not pose serious threats to others. Only armed forces, and the political-industrial-technological complexes which guide them, constitute serious threats against which threatened communities may respond in kind. Civilian populations, just war theory surmises, are morally off-limits as targets. Holmes contends that this just war rule of non-combatant immunity can never be satisfied. For all possible wars in this world - given the nature of military technology and tactics, the heat of battle, and the limits of human knowledge and self-discipline - involve the killing of innocents, thus defined. We know this to be true from history and have no good reason for expecting otherwise in the future. But the killing of innocents, Holmes says, is always unjust. So no war can ever be fought justly, regardless of the nature of the goal sought after, such as national defence from an aggressor's attack. The very activities needed to fight wars are intrinsically corrupt, and cannot be redeemed by the putative justice of the ends they are aimed at. How is a just war theorist to respond to this DP challenge?

Some respond by casting doubt on the concept of innocence in wartime. But a just war theorist subscribing to the rule of non-combatant immunity will neither want, nor logically be at liberty, to argue in this fashion. It is hard to see, for example, how infants could be anything other than innocent during a war, and as such entitled not to be made the object of direct and intentional attack. It is only those who, in Walzer's phrase, are "involved in harming us" - i.e. those who pose serious threats to our lives and rights - that we can justly target in a direct and intentional fashion during wartime.

The more appropriate just war response invokes, alongside Walzer, the doctrine of double effect (or DDE). The DDE, invented by Aquinas, is a complex idea. In spite of its apparent technicality, though, the DDE is closely related to our ordinary ways of thinking about moral life. The DDE assumes the

following scenario: agent X is considering performing an action T, which X foresees will produce both good/moral/just effects J and bad/immoral/unjust effects U. The DDE permits X to perform T only if: 1) T is otherwise permissible; 2) X only intends J and not U; 3) U is not a means to J; and 4) the goodness of J is worth, or is proportionately greater than, the badness of U. Assume now that X is a country and T is war. The government of X, contemplating war in response to an attack by aggressor country Y, foresees that, should it embark on war to defend itself, civilian casualties will result, probably in both X and Y. The DDE stipulates that X may launch into this defensive (and thus otherwise permissible) war only if: 1) X does not intend the resulting civilian casualties but rather aims only at defending itself and its people; 2) such casualties are not themselves the means whereby X's end is achieved; and 3) the importance of X's defending itself and its people from Y's aggression is proportionately greater than the badness of the resulting civilian casualties. The DDE, in making these claims, refers to common shared principles regarding the moral importance of intent, of appealing to better expected consequences, and insisting that bad not be done so that good may follow from it.

Just war theorists claim that civilians are not entitled to absolute immunity from attack during wartime. Civilians are owed neither more nor less than what Walzer calls "due care" from the belligerent governments that they not be made casualties of the war action in question. "Due care" involves fighting only in certain ways, applying limited force to specific targets. But does this just war claim simply beg the question against the latest DP principle? DPs insist on absolute immunity for civilians, which in our world would result in banning warfare, whereas just war theorists, acknowledging the threat, seem to dodge it by re-defining the immunity to which civilians are entitled, demoting it to mere "due care." Despite appearances, it is not question-begging but principled disagreement which roots the difference. Just war theorists will argue that civilians cannot be entitled to absolute immunity because that would outlaw all warfare. But outlawing all warfare would ignore both the responsibility for interstate aggression and the implicit entitlement of a state to use necessary means (including armed force) to secure the lives and rights of its citizens from serious and standard threats to them. In the real world, it is neither reasonable nor fair to require a political community not to avail itself of the most effective means available for resisting an aggressive invasion which threatens the lives and rights of its citizens. It is simply not reasonable to require a state to stand down while the aggression of another state wreaks havoc - murder and mayhem - upon its people.

This is not a complete defeat for DP, merely a suggestion of how such defeat might be sought. In my view, DP constitutes the most formidable moral challenge to just war theory (whereas prescriptive realism constitutes the most formidable prudential challenge to just war theory). Suffice it for our purposes to say that the DDE is the just war principle most frequently employed to defeat the DP pacifist's assertion that it is always wrong to kill innocent human beings. Just war theorists prefer to substitute, for this DP claim, the following proposition: what is always wrong, both in peace and war, is to kill innocent human beings intentionally and deliberately. Unintended, collateral civilian casualties can be excused during the prosecution of an otherwise just war, wherein the end is the repulsion of aggression and the means are aimed at legitimate military targets.

Conclusion

This entry provides a sample of the rich and controversial argumentation surrounding philosophical discourse on war. This discourse is dominated by three major traditions of thought: just war theory; realism; and pacifism. The interaction between these three traditions structures the contemporary discussion of wartime issues, at the same time as it fuels fascinating debate about them. While just war theory occupies an especially large and influential space within the discourse, its realist and pacifist alternatives endure as provocative challenges to the philosophical mainstream which it represents.

Bibliography

All the works cited in this entry, plus relevant other works, are listed below. It may be helpful to first locate and emphasize some of the major and most influential sources.

Just war theory is the dominant tradition on the ethics of war and peace. For scholarship on the history and development of just war theory, consult the works of James T. Johnson. Hugo Grotius is often cited as the most formidable classical just war theorist. A translation of his works can be found in J. Scott's edition of *Classics of International Law*. The major contemporary statement of just war theory remains Michael Walzer's *Just and Unjust Wars*. For other contemporary statements, see the works of R. Regan; W.V. O'Brien; J.B. Elshtain; and B. Orend. Works critical of just war theory can be found in the pacifist and realist tracts below.

Other important articles on particular aspects of just war theory include: on *jus ad bellum*, D. Luban, "Just War and Human Rights"; on *jus in bello*, T. Nagel's "War and Massacre" and R. Fullinwinder's "War and Innocence"; and on *jus post bellum*, Kant's "Perpetual Peace", in his *Political Writings* and B. Orend's "Terminating War and Establishing Global Governance".

Hans Morgenthau's *Politics Among Nations* remains an often-cited defense of realism, as does G. Kennan's *Realities of American Foreign Policy*. Henry Kissinger's *Diplomacy* provides the same outlook in perhaps more accessible form. Two of the most focused and effective criticisms of the realist approach to war occur at: Chapter 1 of Walzer's *Just and Unjust Wars*; and Chapters 1-3 of R. Holmes' *On War and Morality*.

The three best contemporary, secular works defending pacifism are: R. Holmes, *On War and Morality*; J. Teichman, *Pacifism and the Just War*; and R. Norman, *Ethics, Killing and War*. Two renowned critical essays on pacifism, both reprinted in R. Wasserstrom, ed. *War and Morality* are G.E.M. Anscombe's "War and Murder" and Jan Narveson's "Pacifism: A Philosophical Analysis".

One prominent writer on the philosophy of war who resists easy classifcation into any of these categories is Carl von Clausewitz. Clausewitz wrote *On War*, one of the most influential general sources, cited by soldiers and statesmen as often as by philosophers or international lawyers. M. Gelven's *War and Existence* is an interesting contemporary piece on the meaning and experience of war, with a Clausewitzian flavor to it.

- Arendt, H. *On Violence*. New York: Harcourt Brace Jovanovich, 1970.
- Axinn, S. *A Moral Military*. Philadelphia, PA: Temple University Press, 1989.
- Bailey, S. *Prohibitions and Restraints in War*. Oxford: Oxford University Press, 1972.
- Barry, J. *The Sword of Justice: Ethics and Coercion in International Politics*. London: Praeger, 1998.
- Beitz, C. "Cosmopolitan Ideals and National Sentiment", *The Journal of Philosophy* (1983), 591-600.
- ---------, "Nonintervention and Communal Integrity", *Philosophy and Public Affairs* (1979/80), 385-91
- ---------, *Political Theory and International Relations*. Princeton, NJ: Princeton University Press, 1979.
- Beitz, C. *et al.*, eds. *International Ethics*. Princeton, NJ: Princeton University Press, 1985.
- Best, G. *Humanity in Warfare*. New York: Columbia University Press, 1980.
- --------, *War and law since 1945*. Oxford: Clarendon, 1994.
- Boucher, D. *Political Theories of International Relations.* Oxford: Oxford University Press, 1998.
- Brady, James and N. Garver, eds. *Justice, Law and Violence*. Philadelphia: Temple University Press, 1991.
- Brandt, R.B. "Utilitarianism and the Rules of War". *Philosophy and Public Affairs* (1971/72), 145-65.
- Brierly, J.L. *The Law of Nations*. New York: Waldock, 6th ed., 1963.
- Brilmayer, L. *Justifying International Acts*. Ithaca, NY: Cornell University Press, 1989.
- Brown, C. *International Relations Theory: New Normative Approaches.* London: Harvester Wheatsheaf, 1992.
- Brown, P. and H. Shue, eds. *Boundaries: National Autonomy and Its Limits*. Totowa, NJ: Rowman Littlefield, 1981.
- Brownlie, I. *International Law and the Use of Force by States*. Oxford: Clarendon Press, 1963.
- ------------, *System of the Law of Nations*. Oxford: Oxford University Press, 1983.
- Bull, H. *The Anarchical Society: A Study of Order in World Politics*. New York: Columbia University Press, 1977.
- Cady, D. *From Warism to Pacifism: A Moral Continuum*. Philadelphia, PA: Temple University Press, 1989.
- Cahill, L.S. *Love Your Enemies: Discipleship, Pacifism and Just War Theory*. Minneapolis: Fortress, 1994.
- Campbell, D. and M. Dillon, eds. *The political subject of violence*. Manchester: Manchester University Press, 1993.
- Ceadel, M. *Thinking about Peace and War*. Oxford: Oxford University Press, 1987.
- Childress, J. "Just-War Theories" in *Theological Studies* 39 (1978), 427-45.
- Christopher, P. *The Ethics of War and Peace: An Introduction to Legal and Moral Issues*. Englewood Cliffs, NJ: Prentice Hall, 1994.
- Cimbala, S. and K. Dunn, eds. *Conflict termination and military strategy: coercion, persuasion and war*. Boulder, CO: Westview, 1987.
- Clausewitz, Carl von. *On War*, trans. by A. Rapoport. Harmondsworth, UK: Penguin, 1995.

- Coates, A.J. *The ethics of war*. Manchester, UK: University of Manchester Press, 1997.
- Cohen, M. "Moral Skepticism and International Relations" in Beitz, ed. *International Ethics*, 3-50.
- Damrosch, L. "The Collective Enforcement of International Norms Through Economic Sanctions", *Ethics and International Affairs* (1994), 60-80.
- Damrosch, L. ed. *Enforcing Restraint: Collective Intervention in Internal Conflicts*. New York: Council on Foreign relations, 1993.
- Davis, G. *Warcraft and the Fragility of Virtue*. Lincoln, NA: University of Nebraska Press, 1992.
- Davis, G.S. *Religion and Justice in the War over Bosnia*. New York: Routledge, 1996.
- Detter Delupis, I. *The law of war*. Cambridge: Cambridge University Press, 1987.
- Dinstein, Y. *War, aggression and self-defence*. Cambridge: Cambridge University Press, 1995.
- Dombrowski, D. *Christian Pacifism*. Philadelphia, PA: Temple University Press, 1991.
- Doppelt, G. "Statism without Foundations", *Philosophy and Public Affairs* (1979-80), 398-403.
- ----------- "Walzer's Theory of Morality in International Relations." *Philosophy and Public Affairs* (1978/79), 3-26.
- Doyle, M. "Kant, Liberal Legacies and Foreign Affairs", *Philosophy and Public Affairs* (1984), 204-35 and 323-53.
- Dubik, J. "Human Rights, Command Responsibility and Walzer's Just War Theory". *Philosophy and Public Affairs* (1982), 354-71.
- Dyer, G. *War*. New York: Crown, 1985.
- Ellis, A. ed. *Ethics and International Relations*. Manchester: Manchester University Press, 1986.
- Elshtain, J.B. ed. *Just War Theory*. Oxford: Basil Blackwell, 1992.
- Elshtain, J.B. *et al.*, *But was it just? Reflections on the Morality of the Persian Gulf War* (ed. by D. DeCosse). New York: Doubleday, 1992.
- Filice, C. "Pacifism: A Philosophical Exploration", *Journal of Philosophical Research* (1992), 119-53.
- Forsythe, D. *Human Rights and Peace*. Lincoln, NA: University of Nebraska Press, 1993.
- Franck, T. *The Power of Legitimacy Among Nations*. Princeton: Princeton University Press, 1990.
- Freedman, L and E. Karsh, eds. *The Gulf Conflict (1990-91): Diplomacy and War in the New World Order*. Princeton, NJ: Princeton University Press, 1993.
- Frey, R.G. and C.W. Morris, eds. *Violence, Terrorism, and Justice*. Cambridge: Cambridge University Press, 1991.
- Frost, M. *Ethics in International Relations.* Cambridge: Cambridge University Press, 1996.
- Fullinwinder, R. "War and Innocence", *Philosophy and Public Affairs* (1975), 90-7.
- Gallie, W.B. *Understanding War.* London: Routledge, 1991.
- Gelven, M. *War and Existence*. Philadelphia, PA: Pennsylvania State University Press, 1994.
- Geyer, A. *Lines in the Sand: Justice and the Gulf War*. Louisville, KY: John Knox Press, 1992.
- Glossop, R.J. *Confronting War: An Examination of Humanity's Most Pressing Problem*. London: McFarland, 3rd ed., 1994.
- ------------- *World Federation? A Critical Analysis of Federal World Government*. London: McFarland, 1993.
- Gowa, J. *Ballots and Bullets: The Elusive Democratic Peace*. Princeton, NJ: Princeton University Press, 1999.
- Gray, J.G. *The Warriors: Reflections of Men in Battle*. New York: Harper and Row, 1970.

- Griffiths, M. *Realism, Idealism and International Politics.* London: Routledge, 1992.
- Haber, J., ed. *Absolutism and its Consequentialist Critics*. Lanham, Maryland: Rowman Littlefield, 1994.
- Hallett, B. ed. *Engulfed in War: Just War and the Persian Gulf*. Honolulu: University of Hawaii Press, 1991.
- Hampson, F.O. *Nurturing Peace: Why Peace Settlements Succeed or Fail*. Washington, DC: United States Institute for Peace, 1996.
- Hare, R.M. "Rules of War and Moral Reasoning". *Philosophy and Public Affairs* (1971/72), 166-81.
- Hauerwas, S. *Should War be Eliminated? Philosophical and Theological Investigations*. Milwaukee, WI: Marquette University Press, 1984.
- Hehir, J.B. "Intervention: From Theories to Cases," *Ethics and International Affairs* 9 (1995), 1-13.
- Held, V., S. Morgenbesser and T. Nagel, eds. *Philosophy, Morality and International Affairs*. New York: Oxford University Press, 1974.
- Henkin, L. *How Nations Behave: Law and Foreign Policy*. New York: Columbia University Press, 2nd ed., 1979.
- Henkin, L. and J.L. Hargrove, eds. *Human Rights: An Agenda for the Next Century*. Washington, DC: American Society of International Law, 1994.
- Hoffman, S. *Duties Beyond Borders*. Syracuse, NY: Syracuse University Press, 1981.
- ----------- *The Ethics and Politics of Humanitarian Intervention.* Notre Dame, IN: University of Notre Dame Press, 1996.
- Holmes, R. *On War and Morality*. Princeton, NJ: Princeton University Press, 1989.
- Holsti, K. *Peace and War: Armed Conflicts and International order, 1648-1989.* Cambridge: Cambridge University Press, 1991.
- ---------- *The State, War and The State of War*. Cambridge: Cambridge University Press, 1996
- Howard, M. *The laws of war: constraints on warfare in the Western world*. New Haven, CT: Yale University Press, 1994.
- Human Rights Watch, *Slaughter Among Neighbours*. New Haven: Yale University Press, 1995.
- Huntington, S. *The Soldier and the State*. Cambridge, MA: Harvard University Press, 1957.
- Johnson, J. T. *Can Modern War Be Just?*. New Haven, CT: Yale University Press, 1984.
- -------------- *Ideology, Reason and Limitation of War: Religious and Secular Concepts, 1200-1740*. Princeton, NJ: Princeton University Press, 1981.
- -------------- *Morality and Contemporary Warfare*. New Haven, CT: Yale University Press, 1999.
- -------------- *The Just War Tradition and the Restraint of War*. Princeton, NJ: Princeton University Press, 1981.
- -------------- *The Quest for Peace*. Princeton, NJ: Princeton University Press, 1987.
- -------------- "Towards Reconstructing the *Jus ad Bellum*", *Monist* (1973), 461-88.
- Johnson, J.T. and G. Weigel, eds. *Just War and Gulf War*. Washington, DC: Ethics and Public Policy Center, 1991.
- Kane, B. *Just War and the Common Good: Jus ad Bellum Principles in 20th century Papal Thought*. San Francisco: Catholic Scholars Press, 1997.
- Kant, I. *Political Writings*, trans. H. Nisbit and ed. H. Reiss. Cambridge: Cambridge University

Press, 1970.

- -------- *The Metaphysics of Morals*, trans. M. Gregor. Cambridge: Cambridge University Press, 1995.
- Kelsay, J. *Islam and War: A Study in Comparative Ethics*. Louisville, KY: Knox, 1992.
- Kelsay, J. and J.T. Johnson, eds. *Just War and Jihad*. New York: Greenwood, 1991.
- Kelsen, Hans. *Principles of International Law*. New York: Holt, Rinehart and Winston, 1966.
- Kennan, G. *Realities of American Foreign Policy*. Princeton, NJ: Princeton University Press, 1954.
- Keohane, R., ed. *Neorealism and Its Critics*. New York: Columbia University Press, 1986.
- Laberge, P. "Humanitarian Intervention: Three Ethical Positions", *Ethics and International Affairs* 9 (1995), 15-35.
- Lackey, D. "A Modern Theory of Just War", *Ethics* (April 1982), 540-6.
- ---------- *The Ethics of War and Peace*. Englewood Cliffs, NJ: Prentice Hall, 1989.
- Lauterpacht, H. *International Law*, Vols. 3 and 4: *The Law of Peace*. Cambridge: Cambridge University Press, 1977-78.
- --------------- *International Law and Human Rights*. New York: Archon Books, 1968.
- Levinson, S. "Responsibility for the Crimes of War", *Philosophy and Public Affairs* (1972-73), 244-73.
- Little, D. "The ‘Just War’ Doctrine and U.S. Policy in Vietnam", in S. Albert and E. Luck, eds. *On the Endings of Wars* (London: Kennikat Press, 1980), 157-71.
- Luban, D. "Just War and Human Rights". *Philosophy and Public Affairs* (1979/80), 160-81.
- --------- "The Romance of the Nation-State", *Philosophy and Public Affairs* (1979-80), 392-97.
- Luper-Foy, S., ed. *Problems of International Justice*. Boulder, CO: Westview, 1988.
- Mavrodes, G. "Conventions and the Laws of War", *Philosophy and Public Affairs* (1975), 117-31.
- Miller, R., ed. *The Law of War*. Lexington, MA: Lexington Books, 1975.
- Miller, R.B. *Interpretations of Conflict: Ethics, Pacifism and the Just War Tradition*. Chicago: University of Chicago Press, 1991.
- Miller, S. "Killing in Self-Defence", *Public Affairs Quarterly* (1993), 325-39.
- Montague, P. "Self-Defence and Choosing Between Lives", *Philosophical Studies* (1981), 207-20.
- Moore, J.N. *Crisis in the Gulf: Enforcing the Rule of Law*. Dobbs Ferry, NY: Oceana, 1992.
- Moore, J.N. ed. *Law and Civil War in the Modern World*. Baltimore, MD: Johns Hopkins University Press, 1974.
- Morgenthau, H. *Politics Among Nations*. New York: Knopf, 5th ed., 1973.
- Myers, R.J. "Notes on the Just War Theory: Whose Justice, Which Wars?", *Ethics and International Affairs* (1996), 116-30.
- Nagel, T. "Ruthlessness in Public Life", in his *Mortal Questions*. Cambridge: Cambridge University Press, 1979. Pp. 75-90.
- --------- "War and Massacre". *Philosophy and Public Affairs* (1971/72), 123-43.
- Nardin, T. *Law, Morality and the Relations of States*. Princeton, NJ: Princeton University Press, 1983.
- Nardin, T, ed. *The Ethics of War and Peace: Religious and Secular Perspectives*. Princeton: Princeton University Press, 1996.
- Nardin, T and D. Mapel, eds. *Traditions of International Ethics*. Cambridge: Cambridge

University Press, 1992.
- Narveson, J. "Violence and War" in T. Regan, ed. *Matters of Life and Death*. Philadelphia, PA: Temple University Press, 1980. Pp. 109-47.
- Norman, R. *Ethics, killing and war*. Cambridge: Cambridge University Press, 1995.
- O'Brien, W. *The Conduct of Just and Limited War*. New York: Praeger, 1981.
- ----------- *The Law of Limited Armed Conflict*. Washington, Dc: Georgetown University Press, 1965.
- ----------- *U.S. Military Intervention: Law and Morality*. Beverly Hills, CA: 1979.
- O'Connell, Robert L. *Of Arms and Men: A History of War, Weapons, and Aggression*. Oxford: Oxford University Press, 1989.
- Oren, N. *Termination of War: Processes, Procedures and Aftermaths*. Jerusalem: Hebrew University Press, 1982.
- Orend, Brian. "A Just War Critique of Realism and Pacifism", *Journal of Philosophical Research* (Forthcoming, Winter 2000).
- ------------- "Armed Intervention: Principles and Cases", *Flinders Journal of History and Politics* (March 1998), 63-80.
- ------------ "Crisis in Kosovo: A Just Use of Force?", *Politics* (September 1999), 125-30.
- ------------ "Evaluating Pacifism", *Dialogue: Canadian Philosophical Reviews* (Forthcoming).
- ------------- "*Jus Post Bellum*", *Journal of Social Philosophy* (Forthcoming, Spring 2000).
- ------------- "Kant on International Law and Armed Conflict", *Canadian Journal of Law and Jurisprudence* (July 1998), 329-81).
- ------------- "Kant's Just War Theory", *Journal of the History of Philosophy* (April 1999), 323-55.
- ------------- *Michael Walzer on War and Justice*. Cardiff, UK: University of Wales Press. Forthcoming.
- ------------- "Terminating War and Establishing Global Governance", *Canadian Journal of Law and Jurisprudence* (July 1999), 253-95.
- ----------- *War and International Justice: A Kantian Perspective*. Waterloo, ONT: Wilfrid Laurier University Press, 2000.
- Osgood, Robert and Robert Tucker, eds. *Force, Order and Justice*. Baltimore, MD: Johns Hopkins University Press, 1967.
- Otsuka, M. "Killing the Innocent in Self-Defense", *Philosophy and Public Affairs* (1992), 74-94.
- Paskins, B. and M. Dockrill, *The Ethics of War*. Minneapolis, MN: University of Minnesota Press, 1979.
- Paulson, S.L. "Classical Legal Positivism at Nuremberg", *Philosophy and Public Affairs* (174/75), 132-58
- Peirce, A. "Just War Principles and Economic Sanctions", *Ethics and International Affairs* (1996), 99-113.
- Peppers, D.P. "War Crimes and Induction: A Case for Selective Nonconscientious Objection", *Philosophy and Public Affairs* (1973/74), 129-66.
- Phillips, R. *War and Justice*. Oklahoma City, OK: University of Oklahoma Press, 1984.
- Pogge, T. "An Institutional Approach to Humanitarian Intervention", *Public Affairs Quarterly* (1992), 89-103.
- --------- "Cosmopolitanism and Sovereignty", *Ethics* (1992), 48-75.

- --------- "Creating Supra-National Institutions Democratically", *Journal of Political Philosophy* (1997), 163-82.
- --------- "How Should Human Rights be Conceived?" *Jahrbuch fur Recht und Ethik* 3 (1995), 103-20.
- --------- "Liberalism and Global Justice: Hoffman and Nardin on Morality in International Affairs", *Philosophy and Public Affairs* (1987), 67-81.
- --------- *Realizing Rawls*. Ithaca, NY: Cornell University Press, 1989.
- --------- "The Bounds of Nationalism" in J. Couture, *et al*, eds. *Rethinking Nationalism*. Calgary, AB: University of Calgary Press, 1998. Pp. 463-504.
- Porter, B. *War and the Rise of the Modern State*. New York: Macmillan, 1994.
- Ramsbotham, O. and T. Woodhouse. *Humanitarian Intervention in Contemporary Conflict.* Cambridge: Polity, 1996.
- Ramsey, Paul. *The Just War: Force and Political Responsibility*. New York: Charles Scribner's Sons, 1968.
- Rawls, J. *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971).
- ---------, Untitled, in *Dissent*'s Summer 1995 symposium on the bombing of Hiroshima.
- --------- "The Law of Peoples", in Shute and Hurley, eds. *On Human Rights*, 41-82.
- Regan, R. *The Moral Dimensions of Politics*. New York: Oxford University Press, 1986.
- --------- *Just War: Principles and Cases*. Washington, DC: Catholic University of America Press, 1996.
- Reiff, D. *Slaughterhouse: Bosnia and The Failure of The West*. New York: Simon and Schuster, 1995.
- Reisman, M. "International Law after the Cold War", *American Journal of International Law* (1990), 859-76.
- Reisman, M. and C. Antoniou, eds. *The Laws of War*. New York: Vintage, 1994.
- Ryan, C. "Self-Defence, Pacifism and Killing", *Ethics* (1983).
- Scott, James Brown, ed. *Classics of International Law*. Washington, DC: Carnegie Institute, 1917. (Contains Grotius, Vattel, Vitoria and Suarez).
- Shue, Henry. *Basic Rights: Subsistence, Affluence and U.S. Foreign Policy*. Princeton, NJ: Princeton University Press, 2nd ed., 1996.
- Shute, S. and S. Hurley, eds. *On Human Rights*. New York: Basic Books, 1993.
- Suganami, H. *The domestic analogy and world order proposals*. Cambridge: Cambridge University Press, 1989.
- Symposium on the Future of State Sovereignty, *Harvard International Review* (Summer 1995).
- Symposium on Gulf War and International Law in *American Journal of International Law*, July 1991.
- Symposium on War, the UN Charter and the War Powers in the U.S. Constitution in *American Journal of International Law*, January 1991.
- Taylor, A.J.P. *How Wars End*. London: Hamilton, 1985.
- ------------- *Origins of The Second World War*. London: Hamilton, 1961.
- Teichman, J. *Pacifism and the Just War*. Oxford: Basil Blackwell, 1986.
- Thompson, K.W. *Political Realism and the Crisis of World Politics*. Princeton, NJ: Princeton University Press, 1960.

- Tucker, Robert. *The Just War: A Study in Contemporary American Doctrine*. Baltimore, MD: Johns Hopkins University Press, 1960.
- Uniacke, S. *Permissible Killing: The Self-Defence Justification of Homicide*. Cambridge: Cambridge University Press, 1995.
- Van Glahen, G. *Law Among Nations*. New York: Macmillan, 1986.
- Vincent, R.J. *Human Rights and International Relations.* Cambridge: Cambridge University Press, 1986.
- Vaux, K. *Ethics and the Gulf War*. Boulder, CO: Westview, 1992.
- Waltz, K. *Man, The State and War*. Princeton, NJ: Princeton University Press, 1978.
- Walzer, M. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. New York: Basic Books, 2nd ed., 1992.
- ---------- "Justice and Injustice in the Gulf War", in D. DeCosse, ed. *Reflections*, 2-25.
- ---------- *Obligations: Citizenship, War and Disobedience*. Harvard: Harvard University Press, 1970.
- ---------- "Moral Judgment in Time of War" in R. Wasserstrom, ed. *War and Morality*, 54-62.
- ---------- "Nation and Universe" in G.B. Peterson, ed. *The Tanner Lecture on Human Values.* Salt Lake City, Utah: Utah University Press, 1990. Pp. 507-56.
- ---------- "Political Action: The Problem of Dirty Hands", *Philosophy and Public Affairs* (1972/73), 160-80.
- ---------- "Response to Lackey", *Ethics* (April 1982), 547-48.
- ---------- "The Moral Standing of States: A Response to Four Critics", *Philosophy and Public Affairs* (1979/80), 209-29.
- ---------- "The Reform of the International System" in O. Osterud, ed. *Studies of War and Peace* (Oslo: Norwegian University Press, 1986), 227-50.
- ---------- *Thick and Thin: Moral Argument at Home and Abroad*. Notre Dame, Ind.: Notre Dame University Press, 1994.
- ---------- "Untitled" contribution to the summer symposium on the 50th anniversary of the bombing of Hiroshima, *Dissent* (1995), 330-1.
- ---------- "War and Peace in the Jewish Tradition" in T. Nardin, ed. *The Ethics of War and Peace: Religious and Secular Perspectives*. Princeton: Princeton University Press, 1996. Pp. 95-112.
- ---------- "World War II: Why Was This War Different?", *Philosophy and Public Affairs* (1971/72), 3-21.
- Walzer, M. and D. Miller, eds. *Pluralism, Justice and Equality*. Oxford: Oxford University Press, 1995.
- Wasserman, D. "Justifying Self-Defense", *Philosophy and Public Affairs* (1987), 356-78.
- Wasserstrom, R. "The Relevance of Nuremberg", *Philosophy and Public Affairs* (1971-72), 22-46.
- Wasserstrom, R. ed. *War and Morality*. Belmont, CA: Wadsworth, 1970.
- Wells, D.A. *An Encyclopedia of war and ethics*. Westport, CT: Greenwood, 1996.
- Williams, H. *International relations and the limits of political theory*. New York: St. Martin's, 1996.
- ------------ *International relations in political theory*. Philadelphia, PA: Open University Press, 1992.
- Yoder, John. *When War is Unjust: Being Honest in Just-War Thinking*. Minneapolis, Augsburg

Press, 1984.
- Zohar, N. "Collective War and Individualistic Ethics: Against the Conscription of 'Self-Defence'", *Political Theory* 21 (1993), 606-22.

Other Internet Resources

- Resources on Just War Theory (Ethics Updates, U. of San Diego)

Related Entries

cosmopolitanism | justice: international | Kant, Immanuel: social and political philosophy | nationalism | responsibility: collective | rights: human | secession | self-determination, collective | sovereignty | world government/state

Brian D. Orend
bdorend@watarts.uwaterloo.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 4, 2000
Content last modified: May 2, 2002

Feminist Ethics

Feminist Ethics is an attempt to revise, reformulate, or rethink those aspects of traditional western ethics that depreciate or devalue women's moral experience. Among others, feminist philosopher Alison Jaggar faults traditional western ethics for failing women in five related ways. First, it shows little concern for women's as opposed to men's interests and rights. Second, it dismisses as morally uninteresting the problems that arise in the so-called private world, the realm in which women cook, clean, and care for the young, the old, and the sick. Third, it suggests that, on the average, women are not as morally developed as men. Fourth, it overvalues culturally masculine traits like independence, autonomy, separation, mind, reason, culture, transcendence, war, and death, and undervalues culturally feminine traits like interdependence, community, connection, body, emotion, nature, immanence, peace, and life. Fifth, and finally, it favors culturally masculine ways of moral reasoning that emphasize rules, universality, and impartiality over culturally feminine ways of moral reasoning that emphasize relationships, particularity, and partiality (Jaggar, "Feminist Ethics," 1992).

Feminists have developed a wide variety of women-centered approaches to ethics, including those labeled "feminine," "maternal," and "lesbian." Each of these approaches to ethics highlights the differences between men's and women's respective situations in life- biological and social; provides strategies for dealing with issues that arise in private as well as public life; and offers action guides intended to undermine rather than bolster the present systematic subordination of women (Jaggar, "Feminist Ethics," 1992). Considered together the overall aim of all feminist approaches to ethics, irrespective of their specific labels, is to create a gender-equal ethics, a moral theory that generates non-sexist moral principles, policies, and practices.

Feminist Ethics: Historical Background

Feminist approaches to ethics, as well as debates about the allegedly gendered nature of morality, are not contemporary developments. A variety of eighteenth and nineteenth-century thinkers like Mary Wollstonecraft, John Stuart Mill, Catherine Beecher, Charlotte Perkins Gilman, and Elizabeth Cady Stanton all discussed what is probably best termed "women's morality." Each of these thinkers pondered questions such as: Are women's psychological feminine traits all natural? Or is it only women's *positive* psychological feminine traits that are natural, their *negative* ones being somehow socially-constructed? Is there a gender neutral standard available to separate women's good or positive traits from women's bad or negative traits? If moral virtues as well as psychological traits are connected with one's affective as well as cognitive dimensions, indeed with one's physiology as Aristotle and Aquinas suggested, shouldn't we expect men and women to manifest different moral virtues as well as different psychological traits? Should all individuals be urged to cultivate precisely the same set of psychological traits and moral virtues, or should there be room for trait and virtue specialization, provided that this specialization does not split down gender lines? Even if this specialization is split down gender lines?

With respect to the kind of questions about "women's morality" posed above, the eighteenth-century thinker Mary Wollstonecraft answered that women's and men's moralities are fundamentally the same. Although she did not use terms such as "socially-constructed gender roles," Wollstonecraft denied that women are *by nature* more pleasure seeking than men. She reasoned that if men were confined to the same cages women find themselves locked in, as are low-ranking military men, for example, they would develop the same kind of weak characters women develop. Denied the chance to develop their rational powers, to become moral persons who have concerns, causes, and commitments over and beyond their own physical and psychological pleasure, men as well as women would become overly "emotional," a condition Wollstonecraft associated with hypersensitivity, extreme narcissism, and excessive self indulgence (Wollstonecraft, A Vindication of the Rights of Women, 1988).

Because she regarded the ability to reason rather than the capacity to feel as the characteristic that distinguishes humans from brutes, Wollstonecraft contrasted manners, such as any automaton might master, with morals, which requires an educated understanding. Whereas society teaches men morals, it teaches women manners. More specifically, society encourages women to cultivate negative psychological traits like "cunning," "vanity," and "immaturity," all of which impede the development of more positive psychological traits. Even worse, society twists what could be woman's genuine virtue into vices. Wollstonecraft specifically claimed that when strong women practice gentleness, it is a grand, even godly, virtue; but when weak women practice it, it is a demeaning, even subhuman, vice. The positive psychological trait of gentleness is transformed into the negative psychological trait of obsequiousness "when it is the submissive demeanor of dependence, the support of weakness that loves, because it wants protection; and is forebearing because it must silently endure injuries; smiling under the lash at which it dare not snarl" (Wollstonecraft, A Vindication of the Rights of Women, p. 117).

Distressed by her female contemporaries' negative psychological traits, Wollstonecraft concluded that the quickest way for women to be regarded as moral is for them to become like men--that is, for women to

display the psychological traits usually associated with men. Yet, just because Wollstonecraft lamented women's moral deficiencies does not mean that she totally blamed women for not being as good as they possibly could be. On the contrary, she claimed that because women are politically and economically oppressed, they do not have the material means necessary to develop their moral potential.

Debates about what makes a character good and a personality socially acceptable did not end with Mary Wollstonecraft. The passing of time can certainly result in changes, for by the nineteenth-century women were regarded as more moral (though also as less intellectual) than men, a view that disturbed utilitarian philosopher John Stuart Mill. As he saw it virtue (as well as intellect) has nothing to do with gender. Society is wrong, he said to set up an ethical double standard according to which women's morality is to be assessed differently than men's morality. Reflecting further on women's alleged moral superiority, Mill concluded that women's "moral nature" is not the result of innate female propensities but of systematic social conditioning. To praise women on account of their great "virtue" is merely to compliment patriarchal society for having inculcated in women those psychological traits that serve to maintain it. Women are taught to live for and sacrifice for others; to always give and never receive; to submit, yield and obey; to be long-suffering. Their "virtue" is not of their own doing; society imposes it upon them (J.S. Mill, The Subjection of Women, 1970).

In contrast to Wollstonecraft and Mill, other eighteenth-and nineteenth-century thinkers denied that virtue is one. Instead, they forwarded either a separate-but-equal theory of virtue according to which male and female virtues are simply different, or a separate-and unequal theory of virtue according to which female virtue is ultimately better than male virtue. Significantly, this diverse group of thinkers disagreed among themselves whether the characteristics typically associated with women (nurturance, empathy, compassion, self sacrifice, kindness) are (1) full-fledged moral virtues to be developed by men as well as by women; (2) *positive* psychological traits to be developed by women alone; (3) or *negative* psychological traits to be developed neither by women nor men.

Catherine Beecher belonged to this group of thinkers. Even though she believed that women's place is in the home, she did not believe that women's work is unimportant. On the contrary, she believed that women's work--the creation and maintenance of homes in which moral virtue thrives--is absolutely essential for society's well-being. She reasoned that men would lose their *raison d'etre* for working if they lacked loving families and well-ordered homes. In an effort to make certain that society would indeed value women's work *at least* as much as men's work, Beecher and her sister Harriet created the discipline of "domestic science." They stressed that women's work requires much intelligence and skill; that it is not easy to manage a household properly. They also emphasized that women's most important work is to make society Christlike--that is, submissive, self-sacrificial, and benevolent. Sheltered safely in the private realm, where they are largely insulated from the siren calls of wealth, power, and prestige that pervade the public sphere of politics and economics, women are supposedly better situated to cultivate what Beecher and her sister termed the Christlike virtue of "self-denying benevolence." The better that women are, the better that everyone else will be. Convinced that women were somehow responsible for the moral rectitude of men and children, it never occurred to Beecher to ask herself why Christ, a man, had selected women rather than men to specialize in the virtue of self-denying benevolence (Beecher and Stowe, The American Woman's Home, 1971).

Writing around the time Beecher wrote, Elizabeth Cady Stanton also found differences between women's and men's moralities. Stanton's discussion of this already knotty topic is complicated by her apparent inability to decide whether female and male virtue and vice are more the product of nurture or nature. But whether her final view is that men's and women's diverging virtues and vices are the result of social manipulation or biological imperative, Stanton consistently maintained that what she regarded as men's inferior set of morals have set the standard for behavior in the public world. Women's set of morals, which Stanton regarded as superior to men's, have been either suppressed or ignored to the detriment of the public world. The solution to this unfortunate state of affairs, said Stanton, is a relatively simple one: permit, indeed require, women to enter the public world. Humankind cannot afford to leave women, as Beecher would, in the private world, struggling to exert their good influence there and only there. (Buhle and Buhle, eds., The Concise History of Women's Suffrage, 1988).

Yet, despite the fact that like Beecher, Stanton valued women's self-denying benevolence, she believed there was a higher virtue for women to develop: namely, self-development. In the course of interpreting a biblical passage in which Jesus praises a poor widow for her charitable actions, Stanton observed that an oppressed person cannot always afford to be so giving--not without destroying herself. Agreeing that the poor widow's charitable actions were indeed laudatory, Stanton nonetheless cautioned women that women's self-sacrifice may effectively perpetuate women's second-class status. Although the duty of self-sacrifice is morally required in the abstract, "ought" implies can in the concrete. Because few women in a patriarchal society have the political and economic means to practice benevolence without men taking advantage of them, they cannot always afford to be other-directed; sometimes they have to be self-centered.

Although she probably did not think of herself as extending Stanton's line of reasoning, Charlotte Perkins Gilman portrayed on all-female society in which the women are able to serve each other and their daughters (produced through parthenogenesis) without anyone being "sacrificed" in the process. Herland is a child-centered society of mothers in which the lines between the so-called private and public realms have been radically redrawn. The women of Herland are at ease in the halls of justice and centers of trade as well as in the nurseries and schools. Competitive individualistic approaches to life, with their hostility toward connectedness, disappear in Herland, and its women are able to relate to each other without dominating each other.

No wonder that the three American explorers--Terry, Jeff and Van --who stumble on Herland are shocked and confused. Before they arrive, they joke about the mythical land, assuming that there must be men it, since women could not possibly cooperate well enough, or be competent enough, to run a country. When they see how successfully Herland is run, only one of them, Van, praises its all female population as a group of exemplary human beings whose behavior all persons, male as well as female, should seek to emulate. As he sees it, the women of Herland exhibit virtues that are neither feminine nor masculine, but simply fully human (Gilman, Herland, 1979).
Of course, Herland is a fictional, ideal-world in which imagined social, economic, and ideological conditions permit women to develop in morally good as well as psychological healthy ways. Conditions are quite different for women in our nonfictional, real-world. In Women and Economics (1966), Gilman

wrote that to the degree women are dependent on men for support, women will be known for their blind faith, complete submission, and servile self-sacrifice, and men will be known for their stubborn opinions, dominating actions, and arrogant selfishness. Only when women are men's economic equals will women and men both be able to develop truly human moral virtue, the perfect blend of pride and humility: namely, self-respect.

Feminine Approaches to Ethics

Clearly, women-centered thinkers in the eighteenth and nineteenth centuries tended to think of morality as gendered. Since women-centered thinkers in the twentieth century also tend to think of morality as gendered, it is important to determine whether a gendered conception of morality is indeed correct. Such a determination cannot be made, however, unless the ontological and epistemological assumptions of those who advocate feminine, maternal, feminist, and/or lesbian approaches to ethics are first articulated. Apparently, most of the thinkers who have forwarded a woman-centered approach to ethics have rejected the ontological assumptions that the more separate the self is from others, the more fully-developed that self is; and the epistemological assumption that the more universal, abstract, impartial, and rational knowledge is, the more closely it mirrors reality. In place of these assumptions, they have instead embraced the ontological assumption that the more connected the self is to others, the better that self is, and the epistemological assumption that the more particular, concrete, partial, and emotional knowledge is, the more likely it represents the world as it truly is. Thus, it is not surprising that "communal woman" rather than "autonomous man" appears in almost every woman-centered approach to ethics, but that she stresses a different message about "women's morality" depending on her particular guise: feminine, maternal, feminist, or lesbian.

Proponents of feminine approaches to ethics like Carol Gilligan and Nel Noddings stress that traditional western moral theories, principles, practices, and policies are deficient to the degree that they lack, ignore, trivialize, or demean those traits of personality and virtues of character that are culturally associated with women. Gilligan presents her work as a response to the Freudian notion that whereas men have a well-developed moral sense, women do not. Freud attributed women's supposed moral inferiority to girls' psychosexual development. Whereas boys break their attachment to their mothers for fear of being castrated by their fathers if they don't, girls remain emotionally tied to their mothers since castration threats have no power over them. As a result of this state of affairs, girls are supposedly much slower than boys to develop a sense of themselves as autonomous moral agents personally responsible for the consequences of their actions or inactions; as persons who must obey society's rules or face its punishments.

According to Gilligan, Freud is simply one of many western psychologists and philosophers who have seen women's moral inferiority where, in Gilligan's estimation, they should have instead seen simply women's moral difference from men. Gilligan singles out her former mentor, educational psychologist and moralist Lawrence Kohlberg for particular criticism. Kohlberg claimed that moral development is a six-stage process. Stage One is the punishment and obedience orientation. To avoid the "stick" of punishment and/or to receive the "carrot" of a reward, children do as they are told. Stage Two is "the

instrumental relativist orientation." Based on a limited principle of reciprocity- You scratch my back and I'll scratch yours--children meet others' needs only if others meet their needs. Stage Three is the "good boy-nice girl" orientation. Adolescents conform to prevailing norms to secure others' approval and love. Stage Four is the law and order orientation. Adolescents begin to do their duty, show respect for authority, and maintain the given social order to secure others' admiration and respect for them as honorable, law abiding citizens. Stage Five is the social-contract legalistic orientation. Adults adopt an essentially utilitarian moral point of view according to which individuals are permitted to do as they please, provided they refrain from harming other people in the process. Stage Six is the universal ethical principle orientation. Adults adopt an essentially Kantian moral perspective that seeks to transcend and judge all conventional moralities. Adults are no longer ruled by self-interest, the opinion of others, or the fear of legal punishment, but by self-legislated and self-imposed universal principles such as those of justice, reciprocity, and respect for the dignity of human persons (Kohlberg in Mischel, ed., Cognitive Development and Epistemology, 1971).

Although Gilligan concedes that Kohlberg's six-stage scale appeals to many people schooled in traditional western ethics, she insists that the popularity of a theory of moral development is not an index of its truth. She asks whether Kohlberg's six stages of moral developments are indeed: (1) universal, (2) invariant (a always precedes b, b always precedes c, etc.), and (3) hierarchical (b is "more adequate" than a, c is "more adequate" than b, etc.). In particular, she asks why, in the Kohlbergian work with which she is most familiar, women rarely climb past Stage Three, whereas men routinely ascend to Stages Four and even Five? Does this gender difference mean that women are less morally developed than men are? Or does it instead mean that there is something wrong with Kohlberg's methodology--some bias that permits men to achieve higher moral development scores than women?

Gilligan answers that Kohlberg's methodology is male-biased. Its "ears" are tuned to male, not female, moral voices. Thus, it fails to register the different voice Gilligan claims to have first heard in her study of twenty-nine women reflecting on their abortion decisions. This moral voice, insists Gilligan, speaks a language of care stressing relationships and responsibilities, a language that is largely unintelligible to Kohlbergian researchers, who speak the dominant moral language of western ethical tradition--namely, a language of justice emphasizing rights and rules.

Although Gilligan emphasizes that the languages of care and justice are not gender correlated in any iron-clad way, with all women speaking only the language of care and all men speaking only the language of justice, the examples she uses tend to belie her important disclaimer. In her foundational abortion study, she shows only women moving in and out of the three moral frames of reference that constitute her relational ethics: Level One in which women overemphasize the interests of their selves; Level Two in which women overemphasize others' interests; and Level Three in which women weave their own interests together with those of others. Thus, a woman at Level One would make her abortion decision in terms of what is best for herself, at Level Two in terms of what is best for others, and at Level Three in terms of what is best for herself and others considered as a relational unit (Gilligan, In a Different Voice, 1982).

As described so far, Gilligan's Levels seem no more an account of human moral development than

Kohlberg's Stages, with Kohlberg focusing on men's moral experience, and Gilligan on women's. Openly admitting this point, Gilligan has begun to study men's as well as women's moral experience. Her central aim is to expose the ways in which the U.S. society continues to muffle boys' and mens' sensitivity, encouraging them to be less than caring and fully nurturant human persons. Gilligan stresses that unlike today's women who speak the moral language of justice and rights nearly as fluently as the moral language of care and relationship, today's boys and men remain largely unable to articulate their moral concerns in anything other than the moral language of justice and rights.

One index of the importance of Gilligan's work is not only the number of thinkers who have applied her insights to their areas of expertise but also the number of thinkers who have taken her work seriously enough to critique it. To date Gilligan's critics have focused either on the relationship between justice and care, considered as two, gender-neutral perspectives on morality, or on the fact, that women are culturally associated with care and men are culturally associated with justice.

Critics who adopt the first strategy are primarily non-feminist critics. Some of them argue that even if care is a moral virtue and not simply a pleasing psychological trait that some people happen to have, it is a less essential moral virtue than justice. Among the statements such non feminist critics make is that it is better to act out of a general moral principle like "aid the needy" than a particular caring feeling like human heartedness because principles are more reliable and less ephemeral than feelings; and (2) that, when justice and care conflict, considerations of impartiality must trump considerations of partiality: my children's fundamental rights and basic needs are neither more nor less important than anyone else's children's.

Other non-feminist critics fault Gilligan not for claiming that care is a genuine moral virtue equal in value to justice, but for suggesting that this is ethical "news." These critics stress that two, not one, basic principles of prima facie obligation, benevolence and justice, have always ruled traditional western ethics. From benevolence flow the principle of utility, the principle of not harming anyone, and the principle of not interfering with another's liberty. From justice flows the principles of equality of respect and consideration and equality before the law. But in defense of Gilligan, what some traditional western philosophers mean by "benevolence" may not be what Gilligan means by "care." Philosopher Lawrence A. Blum invites us to consider the specific principle "Protect one's children from harm," a principle that flows from the general principle of benevolence. As Blum sees it, all sorts of parents subscribe to this specific principle, but only those parents who are caring--that is, sensitive to and aware of their children's unique interests and needs--will not only know when and how to meet the terms of the principle, but actually be motivated to do so. Although most traditional western philosophers agree with Blum that caring parents are more likely to actually act benevolently than uncaring parents are, they do not agree with him that only caring parents are capable of so acting. Instead they insist that a formal sense of duty, whether or not it is accompanied by caring feelings, is sufficient to generate moral action. Like many ethicists who are developing feminine approaches to ethics, however, Blum believes that the person who would be moral must do more than merely obey the letter of the law. He or she must also be infused with the proper spirit--the appropriate emotions, sentiments, feelings--to perform an entirely morally worthy action (Blum, Friendship, Altruism, and Morality, 1980).

In addition to the non-feminist criticisms that have been raised against Gilligan, several feminist criticisms have been directed against her work. Of these criticisms, the most powerful worries that even if women are better carers than men, it may still be epistemically, ethically, and politically imprudent to associate women with the value of care. To link women with caring is to promote the view that because women can care, they should care no matter the cost to themselves.

In Femininity and Domination (1990), feminist critic Sandra Lee Bartky argues that women's experience of feeding men's egos and tending men's wounds ultimately disempowers women. She notes that the kind of emotional work practiced by women in some service-oriented occupations often causes these women to lose touch with their own emotional base. For example, to pay a person to be "relentlessly cheerful"--to smile at even the most verbally-abusive and unreasonably demanding customer--means paying a person to feign a certain set of emotions. Yet, a person can pretend to be happy only so many times before that person forgets how it feels to be genuinely or authentically happy.

Bartky concedes that women insist that, far from draining them; the emotional work they do energizes them. Indeed many wives and mothers claim the experience of caring for their husbands and children is meaning-giving and self-validating. The better carers they become, they more they view themselves as the family's or marriage's indispensable backbone. Yet subjective feelings of empowerment are not the same as the objective reality of actually having power, says Bartky. She explains how women's androcentric emotional work can work against women distorting women's moral integrity. Bartky points to Teresa Stangl, wife of Fritz Stangl, Kommandant of Treblinka. Despite the fact that her husband's monstrous activities horrified her, she continued to "feed" and "tend" him dutifully, even lovingly. In doing so, however, she played footloose and fancy free with her own soul, for a woman cannot remain silent about evil and still expect to keep her goodness entirely intact. Since horror perpetrated by a loved one is still horror, women need to analyze "the pitfalls and temptations of caregiving itself" before they embrace as ethics of care wholeheartedly (Bartky, Femininity and Domination, 1990).

Mullet reinforces Bartky's fears about a feminine ethics of care. She distinguishes between "distortion of caring" on the one hand and "undistorted caring" one the other. According to Mullet, a person cannot truly care for someone if she is economically, socially, and/or psychologically coerced to do so. Thus, genuine, or fully authentic caring cannot occur, for example, under conditions characterized by male domination and female subordination. Only under conditions of sexual equality and freedom can women care for men without men diminishing, disempowering, and/or disregarding them. As long as men demand and expect more caring from women than women demand and expect from men, both sexes will remain morally impoverished. Neither men nor women will be able to authentically care.

Bartky's and Mullet's interpretation of care are far too pessimistic in the opinion of the thinkers--and there are more than Gilligan--who favor "feminine" approaches to ethics. They stress that even if it is dangerous for women to care in a patriarchal society, care remains part of the solution as well as part of the problem. Care's conflicted status calls for the development of a more robust ethics of care, not for the abandonment of care.

In response to the summons for a sound and complete ethics of care, Nel Noddings has developed a

feminine, relational ethics. For Noddings, ethics is about particular relationships between two parties, the "one-caring" and the "cared-for." Caring is not simply a matter of feeling favorably disposed towards humankind in general, of being concerned about people with whom one has no concrete connection. I can't be said to care about the children in Somalia as much as I care for my own two sons. Real care requires actual encounters with specific individuals; it cannot be accomplished through good intentions alone.

Noddings claims that as children we act from a natural caring that moves us to help others simply because we want to. Later, when society distorts our wants and makes it harder for us to care, the deliberateness of ethical caring supplements the spontaneity of natural caring. Nevertheless, says Noddings, natural caring remains somehow better than ethical caring--and certainly the condition of its possibility.

Although Noddings insists that men as well as women can and should be carers, most of her examples of caring involve women, many of whom seem to care too much--that is, to the point of imperiling their own identity, integrity, and even survival. Although Noddings protests that, in her estimation, it is moral for the one-caring to care for herself, she conveys the impression that the one caring should care for herself only insomuch as her self-caring enables her to care for others better. Thus, the one-caring's self-caring is actually a disguised form of other-directed care.

Maternal Approaches to Ethics

Closely related to feminine approaches to ethics are so-called maternal approaches to ethics. Maternal thinkers like Sara Ruddick, Virginia Held, and Caroline Whitbeck affirm the feminine psychological traits and moral virtues that society associates with women. As they see it, a truly gender-equal ethics would not favor paradigms, such as the contract model, that speak much more to men's experience in the public world than anyone's, but especially women's and children's experience in the private world. Most of our relationships, say Ruddick, Held, and Whitbeck, are not between equally-informed and equally powerful persons, but between unequal persons. When a parent relates to a child, or a physician to a patient, or a self-confident adolescent to a depressed and distraught friend, they do not relate as two tycoons do during a business negotiation, but as two individuals with very different and amorphous strengths and weaknesses. Ethics should be built on a model that fits life as most people live it on an everyday basis. Not the concepts, metaphors, and images associated with the practice of contracting, but those associated with the practice of parenting (especially mothering) better express the dynamics of moral life.

In her maternal approach to ethics, Ruddick claims that society should not trivialize what she terms "maternal practice." Like any human practice, maternal practice has its own form of thinking with a vocabulary and logic peculiar to it, and its own aims and goals. In the case of maternal thinking, these aims and goals consist in the preservation, growth, and acceptability of one's children (Ruddick, Maternal Thinking, 1989).

Preserving the life of a child is the "constitutive maternal act." Infants are totally vulnerable. They simply will not survive unless their caretakers feed, clothe, and shelter them. Ruddick gives the example of Julie,

an exhausted young mother with a very demanding infant. Having reached her physical and psychological limits, Julie pictures herself killing her baby daughter. Horrified by her own thought, Julie spends the night riding a city bus, her baby in her arms. She reasons that, as long as they remain in the public eye, she will not harm her baby.

Ruddick tells Julie's story to stress how difficult it is for a mother to meet her child's material needs. Not every mother grows so run-down and desperate that she has to take steps to ensure that she will not slaughter her child. However, even under relatively ideal circumstances, most mothers do have days when they wish they were not mothers. In order to be able to "preserve" their children on these bad days, says Ruddick, mothers should cultivate the intellectual virtue of scrutiny and the moral virtues of humility and cheerfulness. Armed with these three virtues, mothers will be able to roll with the punches that life delivers to children and adults alike.

The second dimension of Ruddick's maternal-practice is *fostering* children's growth. Whatever fostering a child's growth may mean, it does not mean the act of imposing an already written script on one's child. A mother should not hand book entitled The Tale of the Perfect Child to her daughter insisting that she enact in her imperfect life scenes of the "perfect child's" *perfect* life. Rather, a mother should tell her daughter, realistic, compassionate, and delightful "maternal stories"--the kind of stories that will enable her daughter to realize that she is lovable despite her weaknesses.

The third and final dimension of Ruddick's maternal practice is *training*. For the most part, mothers work hard to socialize their children--to help them become committed and concerned citizens as opposed to members of either armed gangs or opium dens. On occasion, however, mothers will demur from "properly" socializing their children. For example, they will refuse to fit their children's vulnerable bodies into military uniforms, or diet them into designer jeans, or dress them for success in the so-called "dog-eat-dog" world. In a patriarchal society--that is, an overly competitive, hierarchical, and individualistic society-mothers may find themselves caught between the external demands of patriarchy on the one hand and their own inner conviction that these external demands are dehumanizing ones on the other. If a mother trains her son to dress for success, he may become both the chief executive officer of a large firm and a very mean-spirited human being. In contrast, if she refuses to teach her son the lessons of conformity, he may become both an exemplary human being and someone whom patriarchy labels "a loser." Like all mothers, says Ruddick, this mother must decide whether her maternal values or those of the larger society should guide her child-training practices.

Grounding the work of preserving, fostering the growth of, and training children, says Ruddick, is the metavirtue of "attentive love." This metavirtue, which is at once cognitive and affective, rational and emotional, enables mothers to look their children in the eyes and not be shocked, horrified, or appalled by what they see. Indeed, among the several characteristics that distinguish maternal thinkers from nonmaternal thinkers is their utter realism. Mothers who love their children *inattentively* let their fantasies blind them. They do not see their children as they actually are. Rather they see their children as they could perhaps be: the fulfillment of their dreams. In contrast to these mothers, mothers who love their children attentively accept their children for whom they are, working within their physical and psychological limits.

Ruddick's ultimate goal is not simply to develop a phenomenology of maternal practice. Rather, she wants to demonstrate that anyone, male or female, who engages in maternal practice will come to think like a mother in the public world as well as the private world. If men spent as much time rearing children as women do, men as well as women would come to think and see what mothers think and see. People who think and see like mothers make connections, for example, between war in the abstract and war in the concrete. For them, war is not about winning, defending one's way of life, and establishing one's position of power. Instead, it is about destroying that boy or girl whom one has spent years preserving, nurturing, and training: a unique human person who cannot be replaced. In sum, for a maternal thinker, war is about death--about canceling out the "product(s)" of maternal practice.

Held approaches maternal practice from a somewhat different perspective than Ruddick does. She suggests that because women have spent so much of their time mothering, they should develop moral theories that fit the kind of relationships and activities that characterize the *private* rather than the *public* domain. Although Held knows that not all women live in the private world, and although she does not believe that all women are determined by nature to have a distinctive set of moral experiences, she nonetheless claims that a sizable gap exists between women's and men's moral experience. It concerns her that traditional western ethics not only discounts *women's* morality but presents what amounts to *men's* morality as *gender neutral* morality. She claims that if traditional western ethics really gender-neutral, however, it would not favor paradigms--for example, the contract model --that speak much more to men's experience than to women's. In Held's estimation, too many traditional western ethicists bless a human relationship as moral to the degree that it serves the separate interests of individual rational contractors. Yet life is about more than conflict, competition, and controversy--about getting what one wants. It is, as mothering persons know, also about cooperation, consensus, and community- about meeting other people's needs. Held speculates that were the relationship between a mothering person and a child, rather than the relationship between two rational contractors, the paradigm for good human relationships, society might look very different (Held, "Feminism and Moral Theory," 1987).

Held concedes, however, that the kind of relationships that exists between mothering persons and children can be just as oppressive--indeed, even *more* oppressive--than the relationship that exist between two rational contractors. For example, it is sometimes harder to recognize abuses of power in a father-son relationship than in an employer-employee relationship. A father's subtle pressure that his artistic son give up the theater and go to law school may not be as evident an abuse of power as the executive who steals his assistant's ideas and presents them as his own, but both situations exploit and undercut the autonomy of the two relatively powerless agents involved.

Held also admits that, in their attempt to celebrate the positive features of maternal ethics, some maternal thinkers unnecessarily reject the valuable features of traditional ethics. Just because a maternal ethics can handle issues that exceed the "moral minimum" of taking everyone's rights seriously does not mean that it should dispense with this "moral minimum." Mothering persons must be fair as well as compassionate; rational as well as emotional; able to make generalizations about human relations as well as to articulate their unique features. A maternal thinker who says that no two human relationships are ever alike invites moral chaos. Like principles, relationships can be qualified as good, better, or best (bad, worse, or worst),

and that which is subject to *qualification* is also subject to *evaluation*. In the same way that we can ask what makes a principle good or bad, we can ask what makes a relationship good or bad.

Unlike some maternal thinkers, Held believes that men as well as women can be mothering persons. Just because men cannot bear children does not mean that they cannot rear children. Men as well as women can, indeed should, appropriate the moral outlook of those who care for others. Leaving caregiving to women alone produces *boys* with relatively combative and insensitive personalities. Because bellicose, unfeeling boys usually mature into bellicose, unfeeling *men* in positions of power, Held claims that human survival may depend on our ability to reorganize the way we parent. Equal parenting, based on men's and women's equal respect and consideration for each other's equal rights of self-determination must become the order of the day (Held, "The Obligations of Mothers and Fathers," 1984).

Despite the fact that Held believes that men as well as women can mother, she still indicates that there may be a *qualitative* difference between female mothering and male mothering. The fact that women can *bear* children as well as *rear* children may signal that women are, afterall, more responsible than men for the existence of new persons. She notes that although women need men to begin a pregnancy, they do not need men to end a pregnancy. Through abortion or suicide, women can say a definite "no" to life.

By stressing women's ultimate responsibility for bringing (or not bringing) new persons into existence, Held does not wish to negate her previous point that fathers as well as mothers are obligated to rear children. Because men participate in the creation of life, they should participate in its maintenance. Nevertheless, men's direct role in procreation lasts but a few moments, whereas women's lasts for nine months. The experiences of pregnancy make women especially aware of their procreative role. For example, when a pregnant woman eats, she can focus on the fact that she is, as they old saying goes, "eating for two." If she fails to eat a healthy diet, both she and the fetus will suffer. Likewise, when a pregnant women finally gives birth, she can say to herself, "Through this pain, I bring life into this world." These kinds of experiences are ones that a man can never have. The well-being of the fetus his sperm helped create depends on him only indirectly, and he will probably never be called to physically suffer for his son or daughter as much as his wife (or lover) did on the day she gave birth to their child.

Even if the daily toil of bringing up a child will eventually take a greater toll on a parent than the momentary suffering of giving birth, Held asserts that we should not trivialize the birthing act as if it had no effect whatsoever on subsequent parent-child relationships. In suggesting that biological experiences may influence the attitudes of the mother and father toward the preciousness of a particular child, Held wants to explore the relationship between the kind of feelings women and men have for their children on the one hand and the kind of obligations they have to them on the other. If ethicists assume that "natural" male tendencies like the desire to pursue their own self-interest competively play a role in determining men's moral rights and responsibilities, then they should make the same assumption about "natural" female tendencies like the tendency to center their lives on the well-being of their children. Thus, the fact that mothers shudder at the Biblical account of Abraham, who was willing to kill his son, Isaac, in order to honor God's command, is to be expected. Women who birth children--who preserve, nurture, and train them--are not likely to believe that obeying a command, even a command of God, is more valuable than preserving the very lives of these children. To be sure, from the standpoint of traditional western ethics, a

mother's refusal to subordinate the concrete life of *her* child to the abstract commands of duty or higher law indicates her underdevelopment as a moral agent. Yet, in an age where a blindness to interconnection has led to the destruction of the environment and a perilous buildup of the nuclear arsenal, a focus on connection rather than individualistic rights may indeed not only suggest a higher morality but offer a saving grace in an increasingly chaotic world.

In general Whitbeck offers an interpretation of mothering that, more than Ruddick's or Held's, emphasizes the *biological* facts of motherhood. In fact, Whitbeck suggests that women's "maternal instinct" causes them to notice things about their babies that men do not. Specifically, a mother often feels as vulnerable as an infant throughout her pregnancy, but especially during labor. It is her helplessness--in the sense of losing control of her body during pregnancy, of suffering pain during labor, of feeling weak during the postpartum period- that enables a mother to understand just how dependent her infant is on her. No matter how hard a father tries, he can never experience either a mother's or an infant's helplessness. All he can do is to intellectualize about this experience, sympathizing with it the best he can.

Despite the fact that Whitbeck concedes that men can learn how to mother their children and that women can choose not to mother their children, she nonetheless believes that men's and women's different biological experiences typically affect the *intensity* of their respective attachments to their offspring. The bodily experiences that women have simply because they are women tend to deepen those feelings and attitudes which cause people generally to care for their infants. To the degree that human beings are mind-body unities rather than mind-body dualities, a mother's physical experiences will affect the way she thinks about her children (Whitbeck, "The Maternal Instinct," 1984).

The critics of maternal approaches to ethics, like the critics of feminine approaches to ethics, are of two kinds: nonfeminist and feminist. Nonfeminist critics doubt that any one human relationship either can, or should, serve as the paradigm for all human relationships. As they see it, any human relationship--be it one of husband-wife, parent-child, sibling-sibling, friend-friend, or ruler subject--is simply too specific to provide a general model for how people should treat each and every person with equal respect and consideration. Certainly, relationships between unequals should not serve as the model for relationships between equals or vice versa.

Feminist critics express reservations about maternal approaches to ethics quite similar to those of non-feminist critics. They note that Ruddick herself has some doubts even about her own version of maternal ethics. She worries that she might be over-idealizing mothers, unnecessarily excluding men and nonbiological mothers from maternal work, and underemphasizing the differences that exist among mothers, some of whom find themselves "mothering" under extremely oppressive circumstances.

Stressing that some mothers abuse and neglect their children, that many men and nonbiological mothers are just as good or even better parents than many biological mothers, other feminist critics add the point that the mother-child relationship is a particularly problematic choice for a moral paradigm, freighted as it is with enough patriarchal baggage to weigh down even the strongest of women. Although these feminists critics of maternal approaches to ethics concede that the mother-child relationship is a better model for human and humanizing relationships than the traditional rational-contractor model, they believe that even

better models are available. After all, the mother-child relationship is not the only human relationship that is based more on need than on desire, more on love than on obligation, and more on trust than on fear. Friendship relationships, especially ones based on shared goals and aspirations as well as on having a good time and/or on providing emotional and economic support, offer all that the mother-child relationship offers and more. Held together by tears, laughter, and sweat rather than waivers, subpoenas, and depositions, friendship relations seem to hold out more possibilities for moral development than contractual relations. They are also less imbalanced than mother-child relationships in that the parties to them are equals in the sense that they can give to each other approximately as much as they take from each other (Friedman, What Are Friends For?:Feminist Perspectives on Personal Relationships and Moral Theory, 1990).

Feminist Approaches to Ethics

Given that some feminine and maternal approaches to ethics have feminist aspects, and that most of the thinkers who are developing feminine and/or maternal approaches to ethics regard themselves as feminists, it is challenging to specify what makes an approach to ethics "feminist" as opposed to simply "feminine" and/or "maternal." Clearly, the focus of feminist approaches to ethics does not make them distinctive. Feminine, maternal, and feminist approaches to ethics are all women-centered; they all speak primarily to women about *women's* moral experiences. Rather, feminist approaches to ethics are distinctive because they, far more than their feminine and/or maternal counterparts, are "political" in the sense that *fully* feminist ethicists are committed, first and foremost, to the elimination of women's subordination--and that of other oppressed persons--in all of its manifestations." A feminist approach to ethics asks questions about *power*- that is, about domination and subordination--even before it asks questions about *good* and *evil, care* and *justice*, or maternal and paternal thinking.

Focused as they are on questions about power, those developing fully feminist approaches to ethics offer action guides aimed at subverting rather than reinforcing the present systematic subordination of women. Liberal, Marxist, radical, socialist, multicultural, global, and ecological feminists have each offered a different set of explanations and solutions for this state of affairs. So too have existentialist, psychoanalytic, cultural, and postmodern feminists. Proponents of these varied schools of feminist thought maintain that the destruction of all systems, structures, institutions, and practices that create or maintain invidious power differentials between men and women is the necessary prerequisite for the creation of gender equality.

Liberal feminists charge that the main cause of female subordination is a set of informal rules and formal laws that block women's entrance and/or success in the public world. Excluded from places such as the academy, the forum, the marketplace, and the operating room, women cannot reach their potential. Women cannot become men's full equals until society grants women the same educational opportunities and political rights it grants men.

Marxist feminists disagree with liberal feminists. They argue that it is impossible for any oppressed person, especially a female one, to prosper personally and professionally in a class society. The only

effective way to end women's subordination to men is to replace the capitalist system with a socialist system in which both women and men are paid fair wages for their work. Women must be men's economic as well as educational and political equals before they can be as powerful as men.

Disagreeing with both Marxist and liberal feminists, *radical feminists* claim that the primary causes of women's subordination to men are women's sexual and reproductive roles and responsibilities. Radical feminists demand an end to all systems and structures that in any way restrict women's sexual preferences and procreative choices. Unless women become truly free to have or not have children, to love or not love men, women will remain men's subordinates.

Seeing wisdom in both radical and Marxist feminist ideas, *socialist feminists* attempt to weave these separate streams of thought into a coherent whole. For example, in Women's Estate, Juliet Mitchell argues that four structures *overdetermine* women's condition: production, reproduction, sexuality, and the socialization of children. A woman's status and function in *all* of these structures must change if she is to be a man's equal. Furthermore, as Mitchell adds in Psychoanalysis and Feminism, a woman's interior world, her psyche, must also be transformed;

for unless a woman is convinced of her own value, no change in her exterior world can totally liberate her.

Multicultural feminists generally affirm socialist feminist thought, but they believe it is inattentive to issues of race and ethnicity. They note, for example, that U.S. "white" culture does not praise the physical attractiveness of African American women in a way that validates the natural arrangement of black facial features and bodies, but only insofar as they look white with straightened hair, very light brown skin and thin figures. Thus, African-American women are doubly oppressed. Not only are they subject to gender discrimination in its many forms, but racial discrimination as well.

Although *global feminists* praise the ways in which multiculturalist feminists have amplified socialist feminist thought, they nonetheless regard even this enriched discussion of women's oppression as incomplete. All too often, feminists focus in a nearly exclusive manner on the gender politics of their own nation. Thus, while U.S. feminists struggle to formulate laws to prevent sexual harassment and date rape, thousands of women in Central America, for example, are sexually tortured on account of their own, their fathers', their husbands', or their sons' political beliefs. Similarly, while U.S. feminists debate the extent to which contraceptives ought to be funded by the government or distributed in public schools, women in many Asian and African countries have no access to contraception or family planning services from any source.

Ecofeminists agree with global feminists that it is important for women to understand how women's interests can diverge as well as converse. When a wealthy U.S. woman seeks to adopt a child, for example, her desire might prompt profiteering middlemen to prey on indigent Asian or African women, desperate to give their yet-to-be-born children a life better than their own. Ecofeminists add another concern to this analysis: In wanting to give her adopted child the best that money can buy, an affluent woman might not realize how her spending habits negatively affect not only less fortunate women and

their families, but also many members of the greater animal community and the environment in general.

Departing from these inclusionary ways of understanding women's oppression, *existentialist* feminists stress how, in the final analysis, all selves are lonely and in fundamental conflict. In The Second Sex, Simone de Beauvoir writes that, from the beginning, man has named himself the Self and woman the Other. If the Other is a threat to the Self, then woman is a threat to man; and if men wish to remain free, they must not only economically, politically, and sexually subordinate women to themselves, but also convince women they deserve no better treatment. Thus, if women are to become true Selves, they must recognize themselves as free and responsible moral agents who possess the capacity to perform excellently in the public as well as the private world.

Like existentialist feminists, *psychoanalytic* and *cultural feminists* seek an explanation of women's oppression in the inner recesses of women's psyche. As they see it, because children are reared almost exclusively by women, boys and girls are psychosocialized in radically different ways. Boys grow up wanting to separate themselves from others and from the values culturally linked to their mothers and sisters. In contrast, girls grow up copying their mothers' behavior and wanting to remain connected to them and others. Moreover, because of the patriarchal cues they receive both in and outside the home, boys and girls-come to think that such "masculine" values as justice and conscientiousness, which they associate with culture and the public world, are more fully *human* than such "feminine" values as caring and kindness, which they associate with nature and the private world.

In the estimation of many *psychoanalytic and cultural* feminists, the solution to this dichotomous, women-demeaning state of affairs rests in some type of dual-parenting arrangement. Were men to spend as much time fathering as women presently spend mothering, and were women to play as active a role in the world of enterprise as men currently do, then children would cease to associate authority, autonomy, and universalism with men and love, dependence, and particularism with women. Rather, they would identify all of these ways of being and thinking as ones that full persons incorporate in their daily lives.

Finally, as *postmodern* feminists see it, all attempts to provide a single explanation for women's oppression not only will fail but should also fail. They *will* fail because there is no one entity, "Woman," upon whom a label may be fixed. Women are individuals, each with a unique story to tell about a particular self. Moreover, any single explanation for "Woman's" oppression *should* fail from a feminist point of view, for it would be yet another instance of so-call "phallogocentric" thought: that is, the kind of "male thinking" that insists on telling as absolute truth *one* and *only* one story about reality. Women must, in the estimation of postmortem feminists reveal their differences to each other so that they can better resist the patriarchal tendency to center, congeal, and cement thought into a rigid "truth" that always was, is, and forever will be.

Because feminist approaches to ethics focus on how power is used to oppress *women* in particular, nonfeminist critics of them have complained that these approaches are "female-biased." Ethics, insist these critics, cannot proceed from a specific standpoint--in this case, from the standpoint of women--and still be regarded as an ethics. Indeed, traditional western ethics has proceeded on the assumption that its values and rules apply to all rational persons equally. Yet, any number of the "Great Philosophers'" moral

theories seem to be based on the *moral experience of men*--usually powerful ones--as opposed to women. For example, Aristotle's ethics reflects the values of Athenian *citizens*: that is, property-owning Greek males. It does not reflect the values of Greek females or of slaves/foreigners--be they male or female. Nevertheless, traditional western ethicists have tried to make the case that, properly interpreted, Aristotle's ethics applies equally well to both women and men, to both non-Greeks and Greeks; and that it would be misguided to *deliberately*--as opposed to *nonreflectively*- construct an ethics that focuses on a specific group of people.

Related to the above controversy are similar controversies about women's history and literature courses, for example. A person developing a feminist approach to ethics could argue, for example, that she is simply doing what Aristotle, Mill, and Kant should have done in the first place--namely, paying as much attention to women's moral experience as men's. In the same way that historians have ignored the stresses, strains, and struggles of the private world of children, church, and kitchen to focus on the economic revolutions, political upheavals, and military conquests of the public world, traditional western ethicists have focused on men's moral interests, issues, and values, failing to notice just how significant and interesting women's moral issues and values are. Therefore, when a proponent of feminist ethics insists on highlighting "women's morality," she may be doing little more than some corrective surgery--adding women's moral experiences to a male-biased ethical tradition sorely in need of them.

However, she may be doing more than this. She may be suggesting that it is not enough for traditional western ethics to incorporate women's interests and issues, and to recognize women as moral agents who must be taken seriously. On the contrary, she may be urging the "Tradition" to rethink all of the ontological and epistemological assumptions upon which it is based; and even to consider the possibility, that far from being sources of human liberation, its principles, rules, regulations, norms, and criteria actually serve to support patterns of domination and subordination that "demoralize" everyone.

If its focus on women-oppressive system and structures is indeed what makes an ethics feminist, as opposed to simply feminine or maternal, then Alison Jaggar's summary of the fourfold function of feminist ethics cannot be improved upon in any significant way. According to Jaggar, all fully feminist approaches to ethics seek to (1) articulate moral critiques of actions and practices that perpetuate women's subordination; (2) prescribe morally justifiable ways of resisting such actions and practices; (3) envision morally desirable alternatives for such actions and practices; and (4) take women's moral experience seriously, though not uncritically (Jaggar, "Feminist Ethics" 1992). Women should not focus on making the world a better place for *everyone* in general; rather, their primary aim should be to make the world a better place for *women* in particular -and perhaps also for other vulnerable people like children, the elderly, the infirm, the disabled, minorities, etc. In Jaggar's estimation, encouraging women with supportive thoughts, kind words, and benign actions is not enough. A feminist approach to ethics entails women resisting and overcoming their continuing oppression under patriarchy.

Lesbian Approaches to Ethics

That feminist approaches to ethics should be so bold as to focus on women is part of what makes them

unique and controversial. In a similar vein, lesbian approaches to ethics dare to focus only on lesbians in what represents a further assault on the Tradition. First, by speaking to lesbians primarily or even exclusively, those developing lesbian approaches to ethics carry "particularity" to what even some feminists believe is a fault. There is concern, for example, that lesbian ethics privileges lesbians' moral concerns over those of heterosexual women. But here the question to ask is if there really is that much difference between an ethics that focuses on lesbians in particular and one that focuses on women in general. In both instances, the purpose of the narrower focus is to identify and overcome structures of domination and subordination.

Second, lesbian ethics threatens the Tradition because thinkers like Mary Daly recommend a transvaluation of values equal to the one Nietzsche offered nineteenth-century western moralists. Nietzsche's disenchantment with western civilization--and its good-naturedness, mediocrity, egalitarianism, softness--led him to redefine and counter prevailing notions of good and bad. Virtue, said Nietzsche, does not consist in what the Jews, Christians, democrats and socialists believe it consists; namely, kindness, humility, and sympathy. Rather it consists in what noble aristocrats, or *übermenschen*, regard as good; namely, assertiveness, aloofness, and pridefulness. What western civilization has come to accept as "good" is in actuality very bad, insisted Nietzsche. Value must be transvalued. What is praised as "virtue" must be exposed as *vice*; and what is condemned as "vice" must be revealed as *virtue*.

Daly is Nietzschean not because she posits two types of morality--a superior female morality and an inferior male morality--but because she insists that when it comes to women, she whom the patriarch calls "evil," is in fact good, whereas she whom the patriarch calls "good" is in fact bad. If a woman is to escape the traps men have laid for her--if she is to assert her power, to be all that she can--she must realize that it is not *good* for her to sacrifice, deny, and deprive herself for the sake of the men and children in her life. What *is* actually good for women, observes Daly, is precisely what patriarchy identifies as *evil* for women; namely, becoming her own person (Daly, Gyn/Ecology, 1978).

Finally and continuing Daly's revolution, Sarah Lucia Hoagland releases in her Lesbian Ethics (1989) the too-long suppressed female freedom to question and choose. Hoagland says women must replace the questions "Am I good?" and "Is this good" with the question "Does this contribute to my self-creation, freedom, and liberation?" Of course, from the perspective of traditional western ethics, Hoagland's question is not the "right" question to ask.

As Hoagland's supporter Marilyn Frye sees it, however, the need to ask the "right" question tends to arise among people who have a vested interest not only in being good but also in making others be good. For example, a white/Christian/middle-class/ heterosexual American bases his conception of himself as a judge, teacher, preacher, director, administrator, manager, leader and in this mode, as an authority, upon his conviction that he is in the right--that he knows what is good for others as well as himself. So long as women continue to accept this "male" conception of moral agency, says Frye, they will have only two choices: (1) to become like men so that they can exert *men's* moral authority over others; or (2) to become female moral authorities who then make it their business to exert *women's* moral authority over others.

That Frye should regard both of these options as unacceptable is not surprising. The first forces women to

negate themselves; and the second sends women down the same moral blind-alley men have supposedly gone. If ethics is about some people not only proclaiming to other people what is "good" for them, but imposing that "good" upon them, then Frye welcomes the criticism that lesbian approaches to ethics are not really about ethics. Only those who have a vested interest in the status quo, in the powerful remaining powerful, require certitude about their righteousness and their warrant to "direct" and "administer" everything. But because lesbians are some of society's least powerful members, says Frye, they have no way to impose their conception of the good on anyone but themselves nor do they have any desire to do so (Frye, "A Response to Lesbian Ethics: Why Ethics?", 1991).

Claiming the particularity rather than universality of their conception of the good, lesbian ethicists contrast their self-created values with the Tradition's other-imposed values. They imply that insofar as lesbians are concerned, the act of choosing, in and of itself, makes what is chosen somehow "good." More than any other feature of lesbian ethics, it is this one that provokes comment from its nonfeminist critics. These critics ask thinkers like Frye and Hoagland if it makes no moral difference to them what a lesbian chooses, provided that she *freely* choose it.

To this query Hoagland has a response. She notes that so many limitations and boundaries have been imposed on lesbian choice that perhaps for now choice *is* of more moral importance to lesbians than the things chosen. Still, Hoagland does not mean to insist there are no limits on lesbian choice. At one point in her analysis of what constitutes moral agency and interaction, Hoagland observes that in choosing for herself, a lesbian chooses for other lesbians, who in turn choose for her. Lesbians do not weave value in isolation from each other; they weave value together. Ethics is not an individualistic quest; moral value does not emerge from somewhere deep within one's self or from far outside of one's self. On the contrary, moral value--that is, meaning--emerges from what Hoagland terms "lesbian context," or "an energy field capable of resisting oppression." A lesbian approach to ethics is about lesbians becoming the kind of human beings who refuse to participate in anything other than egalitarian relationships (Hoagland, Lesbian Ethics, 1989)..

Insofar as relationships of domination and subordination are a very bad thing indeed, what heterosexual women and men can learn from a lesbian approach to ethics, then, is what Hoagland and other lesbian ethicists--Maria Lugones in particular--term "playfulness;" that is, "the ability to travel in and out of each other's worlds" (Lugones, "Playfulness, ‘World’-travelling, and Loving Perception," 1987). In fact, Hoagland believes that playfulness is the essence of a lesbian approach to ethics--a welcome relief, to what she perceives as the deadly and deadening seriousness of traditional western approaches to ethics. An emphasis on playfulness--on adventure, curiosity, desire--enables lesbians to create moral meanings and values for lesbians only, leaving it to nonlesbians to create their own moral meanings and values.

Conclusion

Although feminine, maternal, feminist, and lesbian approaches to ethics are all women-centered, they do not impose a single normative standard on women. Rather they offer to women multiple standards that validate women's different moral experiences in ways that points to the weaknesses as well as the

strengths of the values and virtues culture has traditionally labeled "feminine." In addition, they suggest to women several paths, all of which lead toward the one goal that is essential to the project of any women-centered ethics; namely, the elimination of gender inequality.

Although feminists' different interpretations of what constitutes a voluntary and intentional choice, an illegitimate or legitimate exercise of control, and a healthy or a pathological relationship reassure the intellectual and moral community that, after all, feminism is not a monolithic ideology that prescribes one and only one way for *all* women to be, this variety of thought is also the occasion of considerable political fragmentation among feminists. Asked to come to the policy table to express *the* feminist perspective on a moral issue, all that an honest feminist ethicist can say is that there is *no* such perspective. Yet, if feminists have no clear, cogent, and unified position on a key moral issue, then a perspective less appealing to women may fill the gap. Although it is crucial for feminist ethicists to emphasize, for example, how a policy that benefits one group of women might at the same time harm another group of women, it is probably a mistake for feminist ethicists to leave the policy table without suggesting policies that are able to serve the *most important* interests of the *widest* range of women. For this reason, many feminist ethicists believe that, over and beyond their commitment to eliminating gender inequality, feminists need to develop a mutually-agreeable methodology that will permit them to achieve a consensus position on many, if not all, the moral issues related to women. Feminist ethicists have a moral duty first to listen to each others' differing points of view, and then to develop a theory and practice that, despite their shortcomings, will nevertheless help inch as many women as possible towards the goal of gender equality with men.

Bibliography

- Bartky, S.L. (ed.), (1990). Femininity and Domination. New York: Routledge.
- Bartlett, K.T. (1991). Feminist legal methods. In K.T. Bartlett & R. Kennedy (eds.), Feminist Legal Theory: Readings in Law and Gender. Boulder, Colo.: Westview Press, 370-403.
- Beecher, C.E. and Stowe, H.B. (1971). The American Woman's Home: Principle of Domestic Science. New York: Aeno Press and The New York Times.
- Blum, C.P. (1966). Friendship, Altruism, and Morality. London: Routledge and Kegan Paul.
- Buhle, M.J., Buhle, P. (eds.) (1978). The Concise History of Women's Suffrage. Urbana: University of Illinois Press.
- Daly, M. (1978). Gyn/ecology: The Metaethics of Radical Feminism. Boston: Beacon Press.
- Daly, M. (1984). Pure Lust: Elemental Feminist Philosophy. Boston: Beacon Press.
- Frye. M. (1991). A response to Lesbian Ethics: Why ethics? In C. Card (ed.), Feminist Ethics. Lawrence, Kans.: University Press of Kansas, 53.
- Gilligan, C. (1982) In A Different Voice: Psychological Theory and Women's Development. Cambridge, Mass.: Harvard University Press.
- Gilman, C.P. (1979). Herland: A Lost Feminist Utopian Novel. New York: Pantheon.
- Gilman, C.P. (1966). Women and Economics. New York: Harper and Row.
- Held, V. (1983). The obligations of mothers and fathers. In J. Trebilcot (ed.) Mothering: Essays in Feminist Theory, Totowa, NJ: Rowman and Allanheld, 7.

- Held, V. (1987). Feminism and moral theory. In E. Kittay and D. Meyers (eds.), Women and Moral Theory. Savage, Md.: Rowman and Littlefield.
- Held, V. (1993). Feminist Morality: Transforming Culture, Society, and Politics. Chicago: University of Chicago Press.
- Hoagland, S.L. (1988). Lesbian Ethics. Palo Alto, Calif.: Institute of Lesbian Studies.
- Jaggar, A.M. (1983). Feminist Politics and Human Nature. Totowa, NJ.: Allenheld.
- Jaggar, A.M. (1991). Feminist ethics: Projects, problems, prospects. In C. Card (ed.), Feminist Ethics. (Lawrence, Kan.: University Press of Kansas).
- Jaggar, A.M. (1992). Feminist ethics. In L. Becker and C. Becker (eds), Encyclopedia of Ethics. New York: Garland Press, 363-4.
- Kohlberg, L. (1971). From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development. In T. Mischel (ed), Cognitive Development and Epistemology. New York: Academic Press, 164-5.
- Kourany, J., Sterba, P., and Tong, R. (eds.), Feminist Philosophies: Problems, Theories, and Applications. Englewood Cliffs, NJ.: Prentice Hall Lugones, M. (1987). Playfulness, 'world'-traveling, and loving perception. Hypatia., **2**, 13.
- Lugones, M. and Spelman, M. (1992). Have we got a theory for you! Feminist theory, cultural imperialism, and the demand for "the woman's voice."
- Mill, J.S. (1970). The subjection of women. In A.S. Rossi (ed), Essays on Sex Equality. Chicago: University of Chicago Press. 125 56.
- Mullet, S. (1988). Shifting perspectives: A new approach to ethics. In L. Code, S. Mullet, and C. Overall (eds), Feminist Perspectives: Philosophical Essays on Method and Morals. Toronto: University of Toronto Press.
- Noddings, N. (1984). Caring: A Feminine Approach to Ethics and Moral Education. Berkeley: University of California Press.
- Ruddick, S. (1983). Maternal thinking. In J. Trebilcot (ed), Mothering: Essays in Feminist Theory. Totowa, NJ: Rowman and Allanheld, 213-30.
- Sichel, B.A. (1991). Different strains and strands: Feminist contributions to ethical theory, Newsletter on Feminism, 90.
- Taylor Mill, H. (1970). Enfranchisement of women. In A.S. Rossi (ed.) Essays on Sex Equality. Chicago: University of Chicago Press.
- Tong, R. (1993). Feminine and Feminist Ethics. Belmont, Calif.: Wadsworth,Whitbeck, C. (1983). The maternal instinct. In J. Trebilcott (ed.), Mothering: Essays in Feminist Theory. Totowa, NJ: Rowman and Allanheld, 185-198.
- Wollstonecraft, M. (1988). A Vindication of the Rights of Women, ed. M. Brody. (London: Penguin).

Other Internet Resources

- Hypatia (Feminist journal at Indiana University)
- Larry Hinman's Ethics Updates (University of San Diego)
- Starting Points for Internet Women's Studies Research (Library at University of Minnesota,

Morris)

- Women's Liberation Research network (by Ginny Daley, Duke University)

Related Entries

feminism, topics: feminist perspectives on class and work | feminism, topics: feminist perspectives on reproduction and the family

Rosemarie Tong
rotong@email.uncc.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 16, 1998
Content last modified: August 4, 2000

Feminist Epistemology and Philosophy of Science

Feminist epistemology and philosophy of science studies the ways in which gender does and ought to influence our conceptions of knowledge, the knowing subject, and practices of inquiry and justification. It identifies ways in which dominant conceptions and practices of knowledge attribution, acquisition, and justification systematically disadvantage women and other subordinated groups, and strives to reform these conceptions and practices so that they serve the interests of these groups. Various practitioners of feminist epistemology and philosophy of science argue that dominant knowledge practices disadvantage women by (1) excluding them from inquiry, (2) denying them epistemic authority, (3) denigrating their "feminine" cognitive styles and modes of knowledge, (4) producing theories of women that represent them as inferior, deviant, or significant only in the ways they serve male interests, (5) producing theories of social phenomena that render women's activities and interests, or gendered power relations, invisible, and (6) producing knowledge (science and technology) that is not useful for people in subordinate positions, or that reinforces gender and other social hierarchies. Feminist epistemologists trace these failures to flawed conceptions of knowledge, knowers, objectivity, and scientific methodology. They offer diverse accounts of how to overcome these failures. They also aim to (1) explain why the entry of women and feminist scholars into different academic disciplines, especially in biology and the social sciences, has generated new questions, theories, and methods, (2) show how gender has played a causal role in these transformations, and (3) defend these changes as cognitive, not just social, advances.

The central concept of feminist epistemology is that of a situated knower, and hence of situated knowledge: knowledge that reflects the particular perspectives of the subject. Feminist philosophers are interested in how gender situates knowing subjects. They have articulated three main approaches to this question: feminist standpoint theory, feminist postmodernism, and feminist empiricism. Different conceptions of how gender situates knowers also inform feminist approaches to the central problems of the field: grounding feminist criticisms of science and feminist science, defining the proper roles of social and political values in inquiry, evaluating ideals of objectivity and rationality, and reforming structures of epistemic authority.

Situated Knowers

Feminist epistemology conceives of knowers as situated in particular relations to what is known and to other knowers. What is known, and the way that it is known, thereby reflects the situation or perspective of the knower. Here we are concerned with *claims* to know, temporarily bracketing the question of which claims are true or warranted.

Situated knowledge in general. Consider how people may understand the same object in different ways that reflect the distinct relations in which they stand to it.

Embodiment. People experience the world by using their bodies, which have different constitutions and are differently located in space and time. In virtue of their different physical locations, observers who stand in front of an object have different information about it than observers who have a distant but bird's eye view of it.

First-person vs. third-person knowledge. People have first-personal access to some of their own bodily and mental states, yielding direct knowledge of phenomenological facts about what it is like for them to be in these states. Third parties may know these states only by interpreting external symptoms, imaginative projection, or obtaining their testimony. People also have knowledge *de se* about themselves, expressed in the form "*I* am F *here, now*." This is distinct in character and inferential role from propositional knowledge having the same content, which does not use indexicals.

Emotions, attitudes, interests, and values. People often represent objects in relation to their emotions, attitudes and interests. A thief represents a lock as a frustrating obstacle while its owner represents the lock as a comforting source of security.

Personal knowledge of others. People have different knowledge of others, in virtue of their different personal relationships to them. Such knowledge is often tacit, incompletely articulated, and intuitive. Like the knowledge it takes to get a joke, it is more an interpretive skill in making sense of a person than a set of propositions. (The German language usefully marks this as the distinction between *Erkenntnis*

and *Wissenschaft*.) Because people behave differently toward others, and others interpret their behavior differently, depending on their personal relationships, what others know of them depends on these relationships. *Know-how.* People have different skills, which may also be a source of different propositional knowledge. An expert dog handler knows how to elicit more interesting behavior from an a dog than a novice does. Such know-how expresses a more sophisticated understanding of dogs on the part of the expert, and also generates new phenomena about dogs for investigation.

Cognitive Styles. People have different styles of investigation and representation. What looks like one phenomenon to a lumper may look like three to a splitter. *Background beliefs and worldviews.* People form different beliefs about an object, in virtue of different background beliefs. In virtue of the different background beliefs against which they interpret a patient's symptoms, a patient may think he is having a heart attack while his doctor believes he just has heartburn. Differences in global metaphysical or political worldviews (naturalism, theism, liberalism, marxism) may also generate different beliefs about particulars on a more comprehensive scale.

Relations to other inquirers. People may stand in different epistemic relations to other inquirers -- for example, as informants, interlocutors, students -- which affects their access to relevant information and their ability to convey their beliefs to others.

These kinds of situatedness affect knowledge in several ways. They influence knowers' access to information and the terms in which they represent what they know. They bear on the form of their knowledge (articulate/implicit, formal/informal, by acquaintance or description, and so forth). They affect their attitudes toward their beliefs (certainty/doubt, dogmatic/open to revision), their standards of justification (relative weights they give to different epistemic values such as predictive power and consilience, amount, sources, and kinds of evidence they require before they accept a claim, etc.), and the authority with which they lay claim to their beliefs and can offer them to others. Finally, they affect knowers' assessment of which claims are significant or important.

Social situation. Many of these ways in which knowers' physical and psychological relations to the world affects what and how they know are familiar and extensively studied by cognitive psychology, naturalized epistemology, and philosophy of science. Feminist epistemology takes such studies a further step by considering how the social location of the knower affects what and how she knows. It can thus be seen as a branch of social epistemology. An individual's social locations consists of her ascribed social identities (gender, race, sexual orientation, ethnicity, caste, kinship status, etc.) and social roles and relationships (occupation, political party membership, etc.). Partly in virtue of their different ascribed identities, individuals occupy different social roles that accord them different powers, duties, and role-given goals and interests. They are subject to different norms that prescribe different virtues, habits, emotions, and skills that are thought to be appropriate for these roles. They also acquire different subjective identities. Subjective identification with one's social groups can take several forms. One may simply know oneself to have certain ascribed identities. One may accept or endorse these identities, actively affirming the norms and roles associated with them. Or one may regard one's social identities as oppressive (if, say, one's identity is cast by society as evil, contemptable, or disgusting), yet see one's fate as tied with the groups with which one is identified, and commit oneself to collective action with other

members of those groups to overcome that oppression.

Gender as a mode of social situation. Most feminist theorists distinguish between sex and gender. Sex comprises the biological differences between males and females. Gender is what societies make of sexual differences: the different roles, norms, and meanings they assign to men and women and the things associated with them on account of their real or imagined sexual characteristics. Gender thus has several dimensions (Haslanger 2000).

Gender roles. Men and women are assigned to distinct social roles. For example, most societies reserve political and military offices mostly for men, and assign women most childrearing responsibilities.

Gender norms. Men and women are expected to comply with different norms of behavior and bodily comportment. For example, men are expected to be assertive and athletic; women, deferential and modest. Gender norms are tailored to gender roles: men and women are expected to conform to those norms that make them fit for their gender roles (whether or not they actually occupy those roles).

Gendered traits and virtues. Psychological traits are considered "masculine" and "feminine" if they dispose their bearers to comply with the gender norms assigned to men and women, respectively. "Masculine" traits are therefore regarded as virtues in men and (often) vices in women, while "feminine" traits are regarded as vices in men and virtues in women.

Gendered performance/behavior. Many feminist theorists, often influenced by postmodernism, have come to stress the contextual and performative aspects of gender (West and Zimmerman 1987; Butler 1990). Rather than viewing masculinity and femininity as fixed traits, expressed in every social context, these theorists represent human beings as more flexible and disposed to enact both "masculine" and "feminine" behaviors in different contexts. The man who avoids tenderly comforting a crying baby in the presence of women may do so when alone. Rather than viewing masculinity and femininity as manifested only in behavior within fixed, distinct gender roles, they can be seen as contrasting styles of performance in almost any role. Female body builders strive to show off their muscles in a "feminine" way.

Gender identity. A person's ascribed gender identity -- how others identify him or her -- may not match his or her subjective gender identity -- the sense that one is "really" a man or a woman. Subjective gender identity includes all of the ways one might understand oneself to be a man or a woman. One could identify with any subset of gender norms, roles, and traits ascribed to the gender of which one sees oneself as a member, while repudiating others. One could even repudiate them all, but still identify oneself as a man or a women in terms of what one sees as distinct roles men and women ought to play in bringing about a just future (one that may or may not include gender distinctions). One could, as many feminists do, understand one's gender identity as a predicament shared by all with the same ascribed identity, and thus as a basis for collective action to change the very basis of one's gender identity. One could embrace an "androcentric" identity, including both "feminine" and "masculine" roles, norms, and traits, decline to view oneself in gender polarized terms at all, or play with gender identities in a postmodernist spirit.

Gender symbolism. Animals and inanimate objects may be placed in a gendered field of representation through conventional association, imaginative projection, and metaphorical thinking. Thus, the garage is regarded as "male" space, the kitchen, "female"; male deer are said to have "harems"; pears are seen as "womanly", assault rifles as "manly."

Gendered knowledge. By bringing together the general account of situated knowledge with the account of gender as a kind of social situation, we can now generate a catalogue of ways in which what people know, or think they know, can be influenced by their own gender (roles, norms, traits, peformance, identities), other people's genders, or by ideas about gender (symbolism). Each mode of gendered knowledge raises new questionsfor epistemology.

The phenomenology of gendered bodies. People's bodies are not just differently sexed; they are differently gendered. Early child socialization trains boys' and girls' bodies to different norms of bodily comportment. In the U.S., these norms stress physical freedom, aggressive play, large motor skills, informal and relaxed posture, and indifference to clothing, neatness and appearance in boys; physical constraint, subdued play, small motor skills, formal and modest posture, and self-consciousness about clothing, neatness and appearance for girls. Once internalized, such norms profoundly affect the phenomenology of embodiment. They inform men's and women's distinct first-personal knowledge of what it is like to inhabit a body, to express capacities unique to one sex or another (e.g., breast feeding), and to have experiences that are manifested through different body parts in differently sexed bodies (e.g., orgasm). They also cause men's and women's experiences of gendered behaviors that both can peform to differ -- in comfort, fluidity, feelings of "naturalness" or novelty, self-consciousness, confidence, awkwardness, shame, and so forth. One question these facts raise for feminist epistemology is to what extent dominant models of the world, especially of the relation between minds and bodies, have seemed compelling because they conform to a male or masculine phenomenlogy (Bordo 1987; Young 1990).

Gendered first-personal knowledge de se. It is one thing to know what sexual harassment is, and how to identify it in a case described in third-personal terms. It is another to come to the recognition "*I* have been sexually harassed." Many women who are able to see that women in general are disadvantaged have difficulty recognizing themselves as sharing women's predicament (Clayton and Crosby 1992). The problems of *de se* knowledge are particularly pressing for feminist theory, because it is committed to theorizing in ways that women can use to improve their lives. This entails that women be able to recognize themselves and their lives in feminist accounts of women's predicament. Feminist epistemology is therefore particularly concerned with investigating the conditions of feminist self-understanding and the social settings in which it may arise -- feminist consciousness-raising sessions, women's studies classes, and so forth (MacKinnon 1989).

Gendered emotions, attitudes, interests, and values. Feminist theory defines a representation as *androcentric* if it depicts the world in relation to male or masculine interests, emotions, attitudes or values. A "male" interest is an interest a man has, in virtue of the goals given to him by social roles that are designated as especially appropriate for men to occupy, or in virtue of his subjective gender identity. A "masculine" interest is an interest a man has in virtue of attitudes or psychological dispositions that are

thought specifically appropriate to men. Such attitudes and interests structure the cognition of those who have them. For example, a representational scheme that classifies women as either "babes," "dogs," "whores," or (grand)mothers reflects the androcentric attitudes, interests, and values of single heterosexual adolescent men who view women in terms of their fantasized eligibility for sexual intercourse with them. A representation is *gynocentric* if it depicts the world in relation to female or feminine interests, emotions, attitudes or values. When a man is described as an "eligible bachelor," this reflects the gynocentric perspective of a heterosexual, single woman interested in marriage. An interest, emotion, attitude, or value might be symbolically gendered even if men and women do not manifest it differently. For example the ethics of care represents moral problems in terms of symbolically feminine values -- values culturally associated with women's gender roles (Gilligan 1982). It thus can qualify as a symbolically gynocentric perspective, even if men and women do not differ in their propensity to represent moral problems in its terms, and are equally able to act accordingly. From a performative perspective, this shows that men can behave in "feminine" ways, too. Feminist epistemology raises numerous questions about these phenomena. Can situated emotional responses to things be a valid source of knowledge about them (Diamond 1991, Jaggar 1989, Keller 1983)? Do dominant practices and conceptions of science and scientific method reflect an androcentric perspective, or a perspective that reflects other dominant positions, as of race and colonial rule (Merchant 1980; Harding 1986, 1991, 1993, 1998)? Do mainstream philosophical conceptions of objectivity, knowledge, and reason reflect an androcentric perspective (Bordo 1987; Code 1991; Flax 1983; Rooney 1991)? How would the conceptual frameworks of particular sciences change if they reflected women's interests (Anderson 1995b, Waring 1990)?

Knowledge of others in gendered relationships. Gender norms differentially structure the social spaces to which men and women are admitted, as well as the presentation of self to others. As performative theories of gender stress, men manifest their male identity, and women their female identity, differently alone than in mixed company, and differently in these settings than in gender-segregated contexts. Male and female inquirers therefore have access to different information about others. Male and female ethnographers may be admitted to different social spaces. Even when admitted to the same social spaces, their presence has different effects on those being observed, because they do not stand in the same social relationships to their subjects. Physical objects do not behave differently depending on whether a man or a woman is observing them. But human beings do behave differently according to their beliefs about the gender of who is observing them. Research that elicits information about others through personal contact between the researchers and the research subjects therefore raises the question of how findings might be influenced by the gendered relations between researchers and subjects, and whether gender-inclusive research teams are in a better position to detect this. Ethnography, which derives propositional knowledge of others from personal knowledge of native informants in long-term, often intimate relationships, raises these issues most acutely (Bell et al 1993; Leacocke 1981). Similar issues arise in survey research, clinical research, and human experimentation (Sherif 1987).

Gendered skills. Some skills are labelled masculine or feminine because men and women need them specifically to perform their respective gender roles, and they are not generically useful for almost any role (as walking, talking, and seeing are). It takes a particular knowledge of small children to know how to comfort them, a particular knowledge of soldiers to know how to whip up their morale. Although men

and women alike may acquire and exercise these skills, they are considered the peculiar responsibility of one or the other gender. Men and women may therefore have differential access to such skill-based knowledge. To the extent that the skill is perceived by the agent as the proper province of the "other" gender, he or she may have a difficult time seeing himself or herself perform it confidently and fluidly, and this inability to self-identify with the task can impair performance. The feedback effects of the phenomenology of gendered embodiment and *de se* knowledge of one's own subjective gender identity can therefore influence the exercise of gendered skills. To the extent that a skill is perceived by others as the proper province of one gender, others may grant or withhold acknowledgment of an agent's expertise. If the successful exercise of the skill requires that others be willing to accept it as a competent performance -- as in the cases of comforting children or raising soldiers' morale -- others' gender-based readiness or refusal to grant expertise to an agent in exercising that skill can be a self-fulfilling prophecy. These phenomena raise various questions for epistemology. Does the "masculine" symbolism of certain scientific skills, such as of assuming an "objective" stance toward nature, interfere with the integration of women into science? Do actually or symbolically "feminine" skills aid the acquisition of scientific knowledge (Keller 1983, 1985a; Rose 1987; Smith 1974)?

Gendered cognitive styles. Some theorists believe that men and women have different cognitive styles (Belenky et al 1986; Gilligan 1982). Whether or not this is true, cognitive styles are gender symbolized (Rooney 1991). Deductive, analytic, atomistic, acontextual, and quantitative cognitive styles are labelled "masculine," while intuitive, synthetic, holistic, contextual and qualitative cognitive styles are labelled "feminine." Such associations are not wholly arbitrary, the way blue is gendered male and pink, female. For example, it is seen as masculine to make one's point by means of argument, feminine to make one's point by means of narrative. Argument is commonly cast as an adversarial mode of discourse, in which one side claims vindication by vanquishing the opposition. Such pursuit of dominance follows the competitive pattern of male gender roles in combat, athletics, and business. Narrative is a seductive mode of discourse, persuading by an enticing invitation to take up the perspective of the narrator, which excites one's imagination and feeling. Its operations are more like love than war, and thereby follows a mode of persuasion thought more suitable for women. These phenomena raise numerous epistemological questions: does the quest for "masculine" prestige by using "masculine" methods distort practices of knowledge acquisition (Addelson 1983; Moulton, 1983)? Are some kinds of sound research unfairly ignored because of their association with "feminine" cognitive styles (Keller 1983, 1985b)? Do "feminine" cognitive styles yield knowledge that is inaccessible or harder to achieve by "masculine" means (Duran 1991, Rose 1987, Smith 1974)?

Gendered background beliefs and worldviews. We have seen above how men and women have access to different phenomenological knowledge, *de se* knowledge, know-how, and personal knowledge of others, in virtue of their gender. They also tend to represent the world in different terms, in virtue of their gendered interests, attitudes, emotions and values, and perhaps also (although this is a matter of controversy among feminist theorists) in virtue of different cognitive styles. These differences create different background webs of belief against which information to which men and women have in principle equal access may be processed. Representational schemes that are functional for different gender roles and gendered attitudes make different kinds of information salient. In traditional domestic settings, women tend to notice dirt that men don't. This is not because women have a specially sensitive

sensory apparatus. It is because they have a role which designates the females of the household as the ones who have to clean up. Male surgeons have no difficulty maintaining much higher degrees of vigilance about contamination in an operating room than would ever be warranted in housecleaning. Besides making different kinds of information salient to men and women, their different background knowledge may lead them to interpret commonly accessed information differently. A man might read a woman's demure smile as a coy come-on, where another woman may interpret it as her polite and defensive reaction to unwanted attention from him. Such differences can spring from differential access to phenomenological knowledge. The male and female observers imaginatively project themselves into her situation, inferring her feelings from the feelings they think underlie her body language. Because men's and women's phenomenologies of embodiment are different -- most men are not in the habit of smiling as a defense against unwanted attention from women -- the man may narcissistically imagine the smile as relaxed and spontaneous, whereas the woman may suspect it is forced. Here are a few epistemological questions raised by these phenomena. Are there epistemic obtacles to men's ability to know when they are raping or sexually harassing women, or to legal institutions recognizing this, insofar as they confine their thinking within a "masculine" perspective (MacKinnon 1989)? More generally, do the unexamined sexist or androcentric background beliefs of scientists cause them to generate sexist theories about women, despite their adherence to ostensibly objective scientific methods (Harding 1986; Harding and O'Barr, 1987; Hubbard 1990)? More generally still, how might the social practices of science be organized so that variations in background beliefs of inquirers function as a resource rather than an obstacle to scientific success (Longino 1990; Solomon 1994)?

Relations to other inquirers. Gender differences in knowledge and background beliefs can be reduced if men and women participate in inquiry together. Each gender can take on testimony what the other can acquire through direct experience. Each may also learn how to exercise imaginative projection more effectively, and to take up the perspective of the other gender. However, gender norms influence the terms on which men and women communicate (Kalbfleisch 1995). In many contexts, women are not allowed to speak or even show up, or their questions, comments, and challenges are ignored, interrupted, and systematically distorted, or they aren't accepted as experts. Gendered norms of conversational interaction and epistemic authority thus influence the ability of knowledge practices to incorporate the knowledge and experience of men and women into their processes of discovery and justification. Feminist epistemologists are therefore interested in exploring how gender norms distort the dissemination of testimony and relations of cognitive authority among inquirers (Addelson 1983; Code 1991) and how the social relations of inquirers could be reformed, especially with regard to the allocation of epistemic authority, so as to enable more successful practices of inquiry (Longino 1990; Nelson 1990, 1993).

Problems of and Approaches to Gendered Situated Knowledge. Mainstream epistemology takes as paradigms of knowledge simple propositional knowledge about matters in principle equally accessible to anyone with basic cognitive and sensory apparatus: "2 + 2=4"; "grass is green"; "water quenches thirst." Feminist epistemology does not claim that such knowledge is gendered. But examination of such examples is not particularly helpful for answering the epistemological problems that arise specifically in feminist theory and practice. What is it to know that I am a woman? What is it like to be sexually objectified? Why is it that men and women so often have dramatically divergent understandings of what happened in their sexual encounters? How can we arrange scientific practices so that science and

technology serve women's interests? These kinds of questions make other kinds of knowledge salient for feminist epistemology: phenomenological knowledge, *de se* knowledge, knowledge of persons, know-how, moral knowledge, knowledge informed by emotions, attitudes, and interests. These kinds of knowledge are often gendered, and they can influence the propositional claims people are disposed to form and accept. This has critical implications for mainstream epistemological conceptions of knowledge, insofar as the latter are based on false generalizations drawing only from examples of ungendered knowledge.

Feminist epistemologists stress the situatedness or perspective-relativity of much knowledge. They do not thereby embrace epistemological relativism. To regard some knowledge claim or form of understanding as situated in a perspective is not to claim that the perspective yields true beliefs or satisfactory understandings (not even "for" those taking up the perspective). It is not to claim that perspectives can only be judged in their own terms, nor that no perspectives are better than others, nor that one cannot take a more objective view of the phenomena than that taken up in one or another perspective. It is not to claim that all knowledge necessarily reflects some peculiar non-universalizable relation of a subset of knowers to the object of knowledge. What attention to situated knowledge does do is enable questions to be raised and addressed that are difficult even to frame in epistemologies that simply assume that gender, and the social situation of the knower more generally, is irrelevant to knowledge. How are the knowledge claims generated by gendered perspectives related to one another? Can men take up a gynocentric perspective, and women, an androcentric perspective? Or are there epistemological barriers to such perspective crossing? Are certain perspectives epistemically privileged? Is there any way to construct a more objective perspective out of differently gendered perspectives? What is the relation of an objective perspective, if one is possible, to gendered perspectives? What would be the point of achieving such a perspective? Would the achievement of such an objective perspective make possible or desirable the elimination of gendered perspectives? Feminist epistemology does not rule out in advance the possibility or desirability of objective knowledge. It does raise new questions about objectivity.

Feminist epistemologists have developed their approaches to the situatedness of knowledge within three broad epistemological traditions: standpoint theory, postmodernism, and empiricism. Standpoint theory identifies one particular social situation as epistemically privileged. Postmodernism rejects claims of epistemic privilege, emphasizing instead the contingency and instability of the social identity of knowers, and consequently of their representations. Empiricism seeks standards, within a naturalized framework, for differentiating the circumstances in which situatedness generates error and in which it constitutes a resource that can be harnessed to advance knowledge. It advances a conception of objectivity constituted by critical and cooperative relations among a plurality of differently situated inquirers.

Feminist Standpoint Theory

Standpoint Epistemology in General. Standpoint theories claim to represent the world from a particular socially situated perspective that can lay a claim to epistemic privilege or authority. A complete standpoint theory must specify (i) the *social location* of the privileged perspective, (ii) the *scope* of its privilege: what questions or subject matters it can claim a privilege over, (iii) the *aspect* of the social

location that generates superior knowledge: for example, social role, or subjective identity; (iv) the *ground* of its privilege: what it is about that aspect that justifies a claim to privilege; (v) the *type* of epistemic superiority it claims: for example, greater accuracy, or greater ability to represent fundamental truths; (vi) the *other perspectives* relative to which it claims epistemic superiority and (vii) modes of access to that perspective: is occupying the social location necessary or sufficient for getting access to the perspective? Many claims to epistemic privilege on behalf of particular perspectives with respect to certain questions are commonplace and uncontroversial. Auto mechanics are generally in a better position than auto consumers to know what is wrong with their cars. Practical experience in fulfilling the social role of the mechanic grounds the mechanic's epistemic privilege, which lays a claim to greater reliability than the judgments of auto consumers.

Standpoint theories become controversial when they claim epistemic privilege over socially and politically contested topics on behalf of the perspectives of systematically disadvantaged social groups, relative to the perspectives of the groups that dominate them. The scope of the claimed privilege includes the character, causes, and consequences of the social inequalities that define the groups in question. This type of standpoint theory *classically* claims three types of epistemic privilege over the standpoint of dominant groups: First, it claims to offer deep over surface knowledge of society: the standpoint of the disadvantaged reveals the fundamental regularities that drive the phenomena in question, whereas the standpoint of the privileged captures only surface regularities. Second, in virtue of this, it claims to offer superior knowledge of the modality of surface regularities, and thus superior knowledge of human potentialities. Where the standpoint of the privileged tends to represent existing social inequalities as natural and necessary, the standpoint of the disadvantaged correctly represents them as socially contingent, and shows how they could be overcome. Third, it claims to offer a representation of the social world in relation to universal human interests. By contrast, the standpoint of the privileged represents social phemonena only in relation to the interests of the privileged class, but ideologically misrepresents these interests as coinciding with universal human interests.

Marxist Standpoint Theory. Marxism offers the classic model of a standpoint theory, claiming an epistemic privilege over fundamental questions of economics, sociology, and history on behalf of the standpoint of the proletariat (Marx 1964, Lukács 1971). Workers do not have this standpoint to begin with. They attain it by gaining collective consciousness of their role in the capitalist system and in history. Several aspects of workers' social situation enable them to attain an epistemically privileged perspective on society. Workers are oppressed, central to the capitalist mode of production, endowed with a cognitive style based on their practical productive material interaction with nature, and collectively self-conscious agents of a potentially universal class. Oppression gives them an objective interest in the truth about whose interests really get served by the capitalist system. Centrality gives them experiential access to the fundamental relations of capitalist production. Because, under capitalism, the standing of all other classes is defined in relation to them, in coming to know themselves and their class position, workers come to know their society as a totality (Lukács 1971). Practical productive interaction with the world is the fundamental mode by which people come to know it, in a materialist epistemology. It leads workers to represent their world in terms of use values, whereas capitalists represent it in terms of exchange values. The workers' representation is more fundamental, because the basic laws of economics and history are expressed in terms of the struggle over the appropriation of surplus (use-) value, not in terms

of superficial money (exchange) values. The necessary and transhistorical character of this practical, instrumental mode of knowing also gives it an objective validity for all societies, which must come to grips with accounting for surplus value in terms of ultimate use-values. Universality -- the workers' standing as the agents for the future universal class they will become under communism (where everyone has the same class status, standing in a common relation to the means of production as both workers and collective rulers over the surplus) -- entails that workers represent the social world in relation to universal human interests, rather than in relation to class-specific interests (as is true of capitalist perspectives). This gives their representations of society greater objectivity than capitalist representations. Finally, the collective self-consciousness of the workers has, as all successful intentional action does, the character of a self-fulfilling prophecy. Workers' collective insight into their common predicament and the need to overcome it through collective revolutionary action generates a self-understanding which, when acted upon, gets realized. Workers become the universal class, the primary agent of history, by acting on that self-understanding. The epistemic privilege of the standpoint of the proletariat, therefore, is also grounded in the epistemic privilege that autonomous agents have over what they are doing.

Grounds of Feminist Standpoint Theory. Feminist standpoint theory claims an epistemic privilege over the character of gender relations, and of social and psychological phenomena in which gender is implicated, on behalf of the standpoint of women. The privilege is relative to theories that justify patriarchy or reflect sexist assumptions. Various feminist standpoint theories ground the claim to epistemic privilege in different features of women's social situation. Each can be seen as drawing an analogy with one or more strands of Marxist epistemology.

Centrality. According to marxist feminists, such as Hartsock (1987) and Rose (1987) women are central to the system of reproduction -- of socializing children and caring for bodies -- as workers are central to the system of commodity production. Because women are in charge of tending to the needs of everyone else in the household, they are in a better position than men to see how patriarchy fails to meet people's needs. Men, in virtue of their dominant position, have the privilege of ignoring how their actions undermine the interests of subordinates. The epistemic privilege of women therefore rests on the fact that women as a class have superior access to information about whose needs get better served under patriarchy.

Collective self-consciousness. According to MacKinnon (1999) male dominance is based on sexual objectification, a process involving epistemic mystification. In objectification, dominant groups project their desires onto subordinate groups and, in virtue of their power, make subordinate groups conform to the way dominants want them to be. It represents as given, natural, and necessary the group differences that are caused by dominant group desires. Gender is the mode of objectification constituted by erotic desire, the eroticization of domination. Men constitute women as women by representing their natures as essentially sexually subordinate to men and treating them accordingly. Women can unmask these ideological misrepresentations by achieving and acting on a shared understanding of themselves as women -- that is, as a social group unjustly constituted by sexual objectification. Women act collectively on this shared understanding in resisting the sexist representations made of them, through campaigns against sexual harassment, pornography, restrictions on reproductive freedom, and so forth. Through these feminist actions, in which women refuse to act as sexual objects, women show that representations

of women as sexual objects are not natural or necessary. Their privileged knowledge is agent self-knowledge, made true by being put into action.

Cognitive style. Many versions of standpoint theory (including Flax 1983, Hartsock 1987, Rose 1987, and Smith 1974) accept feminist object relations theory, which explains the development of stereotypical feminine and masculine traits in terms of the different problems of identity-formation faced by male and female children who are raised by female caregivers (Chodorow 1978). Object relations theory postulates that male children form their distinctive masculine identities by separating themselves from their mothers, a task that psychologically involves an anxious rejection of the feminine and a continuous need to maintain distance and boundaries by controlling and denigrating the feminine. Female children gain a sense of their gender identity through identification with their mothers, and so are more comfortable with a blurring of boundaries between self and other. The development of gender identities leads males and females to acquire distinctively masculine and feminine cognitive styles. The masculine cognitive style is abstract, theoretical, disembodied, emotionally detached, analytical, deductive, quantitative, atomistic, and oriented toward values of control or domination. The feminine cognitive style is concrete, practical, embodied, emotionally engaged, synthetic, intuitive, qualitative, relational, and oriented toward values of care. These cognitive styles are reinforced through the distinctive types of labor assigned to men and women -- men having a near monopoly on the theoretical sciences, warmaking, and on positions of political and economic power calling for detachment and control; and women being assigned to hands-on emotional care for others. The feminine cognitive style is said to be epistemically superior because it overcomes the dichtomy between the subject and object of knowing and because an ethics of care is superior to an ethics of domination. Ways of knowing informed by the motive of caring for everyone's needs will produce more valuable representations than ways of knowing informed by the interests of the dominant (Hartsock 1987). They will produce representations of the world in relation to universal human interests, rather than in terms of the interests of dominant classes, ideologically misrepresented as universal interests. To institutionalize the feminine way of knowing, however, would require overcoming the division of mental, manual, and caring labor that characterizes capitalist patriarchy (Rose 1987).

Oppression. Women are oppressed, and therefore have an interest in representing social phenomena in ways that reveal rather than mask this truth. They also have direct experience of their oppression, unlike men, whose privilege enables them to ignore how their actions affect women as a class. The logic of an epistemology that grounds epistemic privilege in oppression is to identify the multiply oppressed as multiply epistemically privileged. Within feminist theory, this logic has led to the development of black feminist epistemology. Collins (1990) grounds black feminist epistemology in black women's personal experiences of racism and sexism, and in cognitive styles associated with black women. She uses this epistemology to supply black women with self-representations that enable them to resist the demeaning racist and sexist images of black women in the wider world, and to take pride in their identities. The epistemic privilege of the oppressed is sometimes cast, following W.E.B. DuBois, in terms of "bifurcated consciousness": the ability to see things both from the perspective of the dominant and from the perspective of the oppressed, and therefore to comparatively evaluate both perspectives (Harding 1991, Smith 1974, Collins 1990). Black women are "outsiders within," having enough personal experience as insiders to know their social order, but enough critical distance to empower critique.

Access to the Feminist Standpoint. Every standpoint theory must offer an account of how one gains access to its situated knowledge. This depends on whether membership in the group whose perspective is privileged is defined objectively, in terms of one's position in a social structure, or subjectively, in terms of one's subjective identification as a member of the group. When group membership is defined objectively, it is neither necessary nor sufficient for gaining access to the privileged perspective. It is not sufficient, because one might be unaware of the fact or objective significance of being a member of the group. Members become aware of their objective group identity only by achieving a shared understanding of their predicament with other group members. This is the function of consciousness-raising groups in feminist practice (MacKinnon 1999). It is not necessary, because when a group is defined objectively, the facts that constitute the group as such and its interests are publicly accessible, so anyone can theorize phenomena in relation to the interests of that group. Thus, Marx theorized from the standpoint of the proletariat, even though he was not a worker. However, to the extent that the ground of epistemic privilege lies in the self-knowledge of autonomous agents, only those who participate in that agency can have first-personal agent knowledge. At this point, the site of epistemic privilege shifts from the group as defined objectively to the group defining itself as a collective political agent. The privileged standpoint is not that of women, but of feminists. Men can participate in the feminist movement, too. But they cannot assume a dominant role in defining (hence knowing) the aims of the feminist movement without defeating that movement, given that a constitutive aim of feminism is overcoming male dominance. When group membership is defined subjectively, then membership in the group is both necessary and sufficient to gain access to the perspective of the group. If subjectively identifying as a woman is necessary and sufficient to have a feminine cognitive style, as object-relations theory postulates, then all and only self-identified woman have access to the epistemically privileged standpoint. Similarly, Collins' (1990) version of black feminist epistemology rests on identity politics.

Goals of Feminist Standpoint Theory. Feminist standpoint theory is a type of *critical theory*, as this term was understood by the Frankfurt school of critical social theorists, from Adorno to Habermas. Critical theories aim to empower the oppressed to improve their situation. They therefore incorporate pragmatic constraints on theories of the social world. To serve their critical aim, social theories must (a) represent the social world in relation to the interests of the oppressed -- i.e., those who are the subjects of study; (b) supply an account of that world which is accessible to the subjects of study, which enables them to understand their problems; and (c) supply an account of the world which is usable by the subjects to study to improve their condition. Critical theory is theory of, by, and for the subjects of study. These pragmatic features of critical theory raise the possibility that claims of superiority for particular theories might be based more on pragmatic than epistemological virtues (Harding 1991, Hartsock 1996). Even if a particular feminist theory cannot make good on the claim that it has *privileged* access to reality, it may offer true representations that are more *useful* to women than other truthful representations.

Criticisms of Feminist Standpoint Theory. Longino (1993b) argues that standpoint theory cannot provide a noncircular basis for deciding which standpoints have epistemic privilege. Stevens (2000) argues that in Marxist theory, the transhistorical necessity of relations of production, and the universalizability of these relations, grounds the epistemic privilege of workers as the future universal class. But gender relations cannot be universalized, and race relations also lack transhistorical necessity, so neither the standpoint of women nor of black women can claim epistemic privilege. Bar-On (1993)

argues against grounding women's epistemic privilege in their oppression, via feminine cognitive styles. If the feminine ethics of care provides the epistemically privileged perspective on morality, then our access to moral knowledge is predicated on the continuation of existing gender relations, which produce this ethic. Grounding epistemic privilege in feminine cognitive styles therefore forces a choice between having ethical knowledge and living in a nonsexist society. Bar-On also claims that the center-periphery model that underwrites the epistemic privilege of workers does not apply to women. Marx held that class conflict is the central phenomenon that drives all other forms of group conflict, including sexism, racism, imperialism, and national and religious conflict. So understanding class could yield an understanding of other dimensions of inequality. It is no longer plausible to hold that *any* group inequality is central to all the others; they intersect in complex ways (Crenshaw 1999). This entails that women cannot even have privileged access to understanding their own oppression, since this takes different forms for different women, depending on their race, sexual orientation, and so forth. This critique has been forcefully developed by feminist postmodernists, who question the very possibility of a unified standpoint of women, and see, behind the assertion of a universal woman's viewpoint, only the perspective of relatively privileged white women (Lugones and Spelman 1983).

Feminist Postmodernism

General Postmodernist Themes. Postmodernism as a North American intellectual movement draws inspiration from a variety of French poststructuralist and postmodernist theorists, including Foucault, Lacan, Derrida, Lyotard, and Irigaray. It embodies a skeptical sensibility that questions attempts to transcend our situatedness by appeal to such ideas as universality, necessity, objectivity, rationality, essence, unity, totality, foundations, and ultimate Truth and Reality. It stresses the locality, partiality, contingency, instability, uncertainty, ambiguity and essential contestability of any particular account of the world, the self, and the good. Politically, the postmodernist emphasis on revealing the situatedness and contestability of any particular claim or system of thought is supposed to serve both critical and liberatory functions. It delegitimizes ideas that dominate and exclude by undermining their claims to transcendent justification. And it opens up space for imagining alternative possibilities that were obscured by those claims.

Although postmodernist themes are often expressed in an obscure jargon, they can be cast in terms more familiar to analytic philosophers. Postmodernists begin with ideas about language and systems of thought. They claim that (what we think of as) reality is "discursively constructed." This is the linguistic version of the now inescapable (!) Kantian thought that our minds grasp things not as they are "in themselves" but only through concepts, signified by words. "The linguistic sign acts reflexively, not referentially" in a "discursive field." This is a version of radical meaning holism: signs get their meaning not from their reference to external things but from their relations to all of the other signs in a system of discourse. Meaning holism entails that the introduction of new signs (or elimination of old ones) will change the meanings of the signs that were already in use. Signs therefore do not have a fixed meaning over time. This is a Heraclitean version of historicism: we cannot step into the same stream of thought twice. Together, these ideas support the "rejection of totalizing metanarratives." There can be no complete, unified theory of the world that captures the whole truth about it. Any such theory will contain

a definite set of terms. This entails that it cannot express all conceptual possibilities. For a discourse that contained different terms would contain meanings not available in the discursive field of the theory that claims completeness. Thus, the selection of any particular theory or narrative is an exercise of "power" -- to exclude certain possibilities from thought and to authorize others.

Postmodernism extends these ideas about language to social practices more generally. The key idea underwriting this extension is that actions and practices are linguistic signs. Like words, they signify things beyond themselves by means of linguistic devices such as metaphor and metonymy. For example, the elevation of the judge's bench metaphorically signifies his superior authority over everyone else in the courtroom. This permits an analysis of social practices and behaviors as exhibiting the same structure and dynamics as language itself. Just as words get their meaning from their relations to other words rather than from their relation to some external reality, so do actions get their meaning from their relations to other actions, rather than from their relation to some pre-linguistic realm of human nature or natural law. Thus, the superior authority of the judge consists in the conventions of deference others manifest in their actions toward him. It is not underwritten by a supposed natural tendency of humans to obey authority, or by an underlying normatively objective authority. The latter thoughts express essentialist and objectivist power plays, attempts to foreclose contests over practices by fixing them in a supposedly extra-linguistic reality. Such attempts are not only objectionable but futile, because the meanings of actions are constantly being subverted by other actions that, in changing the context of the former actions, changes their meanings. This is why postmodernists celebrate ironic, parodic, and campy renditions of conventional behaviors as politically liberating (Butler 1993). If Marx lamented that history repeats itself twice -- first as tragedy, second as farce -- postmodernists revel in the same process.

Postmodernists view the self as likewise constituted by signs that have meaning only in relation to other signs. There is no unified self that underlies the play of a stream of signifiers. This is a linguistic version of Hume's fragmented stream-of-consciousness account of the self, but with a social twist. Signs, unlike Hume's simple ideas, form language, which is socially constructed. Thus, although subjectivity is constituted through the production of signs, the self is not free to make of these whatever it wants, but finds itself entangled in a web of meanings not of its own creation. Our identities are socially imposed, not autonomously created. However, this does not foreclose the possibility of agency, because we occupy multiple social identities (e.g., a woman might be a worker, a mother, lesbian, Mexican, and so forth). The tensions among these conflicting identities open up spaces for disrupting the discursive systems that construct us.

Because, in its philosophy of language, words refer to concepts rather than things in the world, postmodernism reproduces in linguistic terms some of the same epistemological conundrums posed in the history of modern philosophy by the veil of ideas. This generates a tendency toward idealism in both traditions. However, given the constant flux of meanings generated by holism, these tendencies cannot secure the certainty or stability that empiricists thought they could attain by resorting to idealism. The more careful practitioners of postmodernism resist wholesale idealism. Claims that bodies, matter, or the objects investigated by the natural sciences are "discursively constructed" or "socially constructed" do not assert that the external world would disappear if people stopped talking about it. Rather, they assert a kind of nominalism: that the world does not dictate the categories we use to describe it, that innumerable

incompatible ways of classifying the world are available to us, and therefore that the selection of any one theory is a choice that cannot be justified by appeal to "objective" truth or reality. Even the ways we draw our distinctions between mind and body, ideas and objects, discourse and reality, are contestable.

Feminist Postmodernism. Within feminism, postmodernist ideas have been deployed against theories that purport to justify sexist practices -- notably, ideologies that claim that observed differences between men and women are natural and necessary, or that women have an essence that explains and justifies their subordination. The oft-cited claim that gender is socially or discursively constructed -- that it is an effect of social practices and systems of meaning that can be disrupted -- finds one of its homes in postmodernism (Butler 1990). However, postmodernism has figured more prominantly in internal critiques of feminist theories. One of the most important trends in feminist thinking in the past twenty years has been exposing and responding to exclusionary tendencies within feminism itself. Women of color and lesbian women have argued that mainstream feminist theories have ignored their distinct problems and perspectives (Collins 1990; Hull, Scott and Smith, 1982; Lorde 1984). Feminist postmodernism represents both a vehicle for and response to these critiques. It underwrites a critique of the concept "woman" -- the central analytical category of feminist theory. And it proposes perspective-shifting as a strategy for negotiating the proliferation of theories produced by differently situated women.

The critique of the concept "woman." Feminist postmodernists have criticized many of the leading feminist theories of gender and patriarchy as essentialist (Butler 1990, Flax 1990, Spelman 1988). Essentialism here refers to any theory that claims to identify a universal, transhistorical, necessary cause or constitution of gender identity or patriarchy. The objection to essentialism is fundamentally political: in claiming that gender identity is one thing or has one cause, such theories convert discursively constructed facts into norms, difference into deviance. They either exclude women who don't conform to the theory from the class of true "women," or else represent them as inferior. The critiques of feminist theories by lesbian women and women of color have reinforced skepticism about the unity presumed in the category "woman" by highlighting the intersectionality of identities of gender, race, class, and sexual orientation. The chief faultlines for the fragmentation of the category "woman" have thus been the other identity formations along which social inequalities are constructed.

This critique of "woman" as a unified *object* of theorizing entails that "woman" also cannot constitute a unified *subject* of knowing (Lugones and Spelman 1983). The theories of universal gender identity under attack are ones in which the authors, all white middle class heterosexual women, could see themselves. Critics claim that the authors fail to acknowledge their own situatedness and hence the ways they are implicated in and reproduce power relations -- in this case, the presumptuous authority of white middle class heterosexual women to define "the standpoint of women" -- to speak for all other women and define who they are. Feminist standpoint theorists, who claim an epistemic privilege on behalf of their standpoint, are thereby unmasked as asserting a race and class privilege over other women.

Feminist postmodernists draw two lessons from this critique. First, universal claims about women, gender, and patriarchy should be avoided. Second, feminist standpoint theory's project of identifying a single epistemically privileged perspective is fundamentally flawed, an unjustified assertion of power in the name of an unattainable objectivity. This lesson applies to subaltern feminist standpoints as well. The

assertion of a black feminist standpoint, for example, objectionably essentializes black women. Once the postmodernist critique of essentialism is granted, there is no logical stopping point in the proliferation of perspectives.

Perspective shifting. Feminist postmodernism thus envisions our epistemic situation as characterized by a permanent plurality of perspectives, none of which can claim objectivity -- that is, transcendence of situatedness to a "view from nowhere." This position has sometimes been characterized as relativist. Haraway (1991) replies that it rejects both objectivism and relativism for the ways they let knowers escape responsibility for the representations they construct. To claim objectivity for a representation is to claim that "the world made me represent things this way." To claim relativism is to claim that "my identity (my situation) made me represent things this way (and my identity/situation is not inferior to yours)." Both positions disclaim the active participation of the knower in constructing her representations. Even a photograph, the paradigm of an "objective" representation, reflects the photographer's choice of film, lenses, frames, exposure, and so forth. But the resort to a relativism of identity is no better. In asserting the equality of all perspectives, it claims immunity from the critiques of differently positioned others, and complacency in one's own position. Although it acknowledges the dependence of a knower's representations on the particulars of her situation, it claims that she had no choice about that. Postmodernists, however, reject the fixity and unity of personal identity on which relativism rests. People are not epistemically trapped inside their cultures, their gender, their race, or any other identity. They can choose to think from other perspectives. Thus, although we will always have a plurality of perspectives, their constitution is constantly shifting rather than static, and there is no stable correspondence between individuals and perspectives.

Negotiating the bewildering array of situated knowledges therefore involves two types of epistemic practice. One is acceptance of responsibility, which involves acknowledging the choices of situation that entered into the construction of one's representations (Haraway 1991), and considering how one's situation affects the content of one's representations (Harding 1993). The second is "world traveling" (Lugones 1987) or "mobile positioning" -- trying to see things from many other perspectives. Mobile positioning can never be transparent or innocent. Imagining oneself in another's situation is full of risks. It requires sensitive engagement with and sympathy for the others who occupy those positions. Both transform situated knowing into a critical and responsible practice.

Criticisms of Feminist Postmodernism. Both key features of feminist postmodernism -- the rejection of "woman" as a category of analysis, and the infinite fragmentation of perspectives -- are controversial within feminist theory. A wholesale opposition to large-scale generalizations about women seems to arbitrarily preclude a critical analysis of large-scale social forces that critically affect women (Benhabib 1995). That women in different social positions may experience sexism differently does not entail that they have nothing in common -- they still suffer from sexism (MacKinnon 2000). Intersectionality, rather than being a basis for dissolving the category "woman," may be accommodated through a structural analysis of gender that allows for racialized and otherwise particularized modes of sexist oppression (Haslanger 2000). The postmodernist alternative of fragmentation and multiplicity threatens both the possibility of analytical focus (it is impossible to keep all axes of difference in play at once) and of politically effective coalition building among women with different identities. Carried to its logical

conclusion, feminist postmodernism dissolves all groups, thereby reproducing the individualism of the Enlightenment epistemology it claims to repudiate. And the idea of mobile positioning may simply reproduce the objectivism and ideas of autonomy that postmodernists claim to reject, only now in the guise of "the view from everywhere" rather than "the view from nowhere" (Bordo 1990). Critics argue that feminists would do better if they forthrightly appropriated ideals of human rights and autonomy, rather than embracing "the death of the subject" in the fragmentation of the self (Benhabib 1995). Despite these difficulties, postmodernism remains a powerful current in feminist epistemology, due to the acknowledgment by all feminists that a plurality of situated knowledges appears to be an inescapable consequence of social differentiation and embodiment.

Feminist Empiricism

Relations of Feminist Empiricism to Empiricism in general. Empiricism is the view that experience provides the sole, or at least the primary, justification for all knowledge. From the classical empiricists to some early twentieth-century theorists, empiricists held that the content of experience could be described in fixed, basic, theory-neutral terms -- for example, in terms of sense-data. Most also regarded philosophy as a discipline that could provide a transcendent or external justification for empirical or scientific methods. Quine revolutionized empiricism by rejecting both of these ideas. For Quine, observation is thoroughly theory-laden. It is cast in terms of complex concepts that cannot be immediately given in experience, all of which are potentially subject to revision in light of further experience (Quine 1963). And epistemology, far from providing an extrascientific vindication of natural science, is simply another project within science, in which we empirically investigate our own practices of inquiry (Quine 1969). In these two respects, feminist empiricists are the daughters of Quine. However, Quine accepted a sharp division between facts and values that feminist empiricists argue cannot be sustained within a thoroughly naturalized empiricism. Feminist empiricists are deeply engaged in considering how feminist values can legitimately inform empirical inquiry, and how scientific methods can be improved in light of feminist demonstrations of sex bias in currently accepted methods. Their version of naturalized epistemology therefore does not follow Quine in reducing epistemology to nonnormative psychological investigations, but rather upholds the roles of value judgments in rigorous empirical inquiry (Campbell 1998, Nelson 1990). Quine also presupposes an individualist account of inquiry; his preferred reduction basis for naturalized epistemology is behavioral and neuro- psychology. Feminist empiricists are concerned with the impact on inquiry of social practices relating to gender, race, class and other bases of inequality. They therefore take sociology, history, and science studies seriously. Most also advocate a socialized epistemology, in which inquiry is treated as a fundamentally social process and the basic subjects of knowledge may even be communities or networks of individuals.

The Paradoxes of Bias and Social Construction. The central problematics of feminist empiricism can be captured in two apparent paradoxes. First, much feminist science criticism consists in exposing the androcentric and sexist biases in scientific research, especially in theories about women, sexuality, and gender differences. The force of this criticism seems to rest on a prior empiricist commitment to the view that bias is epistemically bad -- that it leads to false theories. Yet, advocates of feminist science urge that feminist values inform scientific inquiry. This amounts to a recommendation that science incorporate

certain biases into its operations. Feminist empiricists need to reconcile these conflicting claims. This is known as the paradox of bias. Second, and relatedly, much feminist science criticism is devoted to exposing the influence of social and political factors on scientific inquiry. Scientists advocate androcentric and sexist theories because they are influenced by the sexist values of the wider society. This would seem to imply that, to eliminate these social biases, feminists adopt an individualist epistemology. Instead, feminist epistemologists stress the social construction of knowledge. They urge, not that inquirers insulate themselves from social influcences, but that they restructure scientific practices to be open to *different* social influences. This can be called the paradox of social construction.

Feminist empiricists argue that the key to dissolving both paradoxes is to undermine the assumptions that underlie them: that biases, political values, and social factors can influence inquiry only by *displacing* the influence of evidence, logic, and whatever other purely cognitive factors tend to lead to true theories. Not all bias is epistemically bad (Antony 1993). There are three general strategies for showing this, which may be called pragmatic, procedural, and moral realist. The pragmatic approach stresses the plurality of aims that inquiry serves. Inquiry seeks truths, or at least empirically adequate representations, but *which* truths any particular inquiry seeks depends on the uses to which those representations will be put, many of which are practical and derived from social interests. The paradoxes are dissolved by showing how responsible inquiry respects a division of labor between the functions of evidence and social values -- the evidence helping inquirers track the truth, the social values helping inquirers construct representations out of those truths that serve the pragmatic aims of inquiry (Anderson 1995b). This view may be joined with a view of nature as rich, complex, and messy. No single theory captures the whole structure of reality, since different ways of classifying phenomena will reveal different patterns useful to different practical interests (Longino 2001). The procedural approach argues that epistemically bad biases can be kept in check through an appropriate social organization of inquiry. A social organization that holds people with different biases accountable to one another will be able to weed out bad biases, even if no individual on her own can be free of bias (Longino 1990). This view may be joined with the idea that the subject of knowledge (Nelson 1993), epistemic rationality (Solomon 1994) or objectivity (Longino 1990, 2001) is the epistemic community, not the individual. The moral realist approach argues that moral, social and political value judgments have truth-values, and that feminist values are true. Inquiry informed by feminist values therefore does not displace attention to the evidence, because the evidence vindicates these values (Campbell 1998).

Feminist empiricists appeal to the pragmatist tradition to undermine the sharp dichotomy between fact and value (Antony 1993; Nelson 1993). They argue (compatibly with other pragmatists, such as Hilary Putnam), that Quine's arguments about the underdetermination of theory by evidence lead to a view of facts as partially constituted by values, and values by facts. In the absence of a sharp distinction between facts and values, it cannot be argued that inquiry explicitly motivated by feminist values is *in principle* opposed to the truth. Whether any particular feminist, or sexist, theory is true or false will depend on empirical investigation informed by epistemic norms -- norms which may themselves be reformed in light of the merits of the theories they generate. This is the project of naturalized epistemology, whereby the vindication of norms of inquiry is sought not outside, but within, ordinary empirical investigation. Feminist empiricist investigations of the interaction of facts and values are further discussed below. Feminist empiricist explorations of how norms of inquiry should be constituted to enhance objectivity are

also discussed below.

Criticisms of Feminist Empiricism. Within feminist theory, the intellectual traditions and training of standpoint and postmodernist epistemologists have not kept track of the radical changes in the empiricist tradition inspired by Quine and further developed by feminist empiricists. Consequently, some criticisms of what is called "feminist empiricism" by other feminist theorists do not fit what feminists who call themselves "feminist empiricists" believe. For example, feminist postmodernists criticize feminist empiricists for presuming the existence of an individual, transhistorical subject of knowledge outside of social determination (Harding 1990), even though the naturalized epistemology that feminist empiricists adopt has long since abandoned that conception of knowers in favor of viewing knowers as socially situated. Feminist empiricists are also criticized for accepting an uncritical concept of experience (Scott 1991), even though feminist empiricists accept the theory- and value-laden character of evidence and hence the critical revisability of descriptions of experience in light of new evidence, theoretical, and normative reflections. Feminist empiricists have also been criticized for naively holding that that science will correct the errors and biases in its theories about women and other subordinated groups all by itself, without the aid of feminist values or insights (Harding 1986, 1991). This contrasts with the actual position of those who call themselves feminist empiricists, who argue that science cannot claim to attain objective knowledge of gendered beings or our gendered social world without actively including feminist inquirers as equals in the collective project of inquiry (Longino 1993a, 1993b). More pointedly, the standpoint theorist Hundleby (1997) criticizes feminist empiricism for overlooking the vital role of feminist political *activity*, in particular, the development of oppositional consciousness, as a superior source of hypotheses and evidence for challenging sexist and androcentric theories.

Feminist Science Criticism and Feminist Science

The history of feminist interventions into most disciplines follows a common pattern. Feminist inquiry begins as a critique of accepted disciplinary methods, assumptions, and canons. As it matures, it develops constructive projects of its own. The history of feminism and science follows this pattern. In the empirical sciences, the pattern helps us see how feminist epistemology negotiates the tension between the two poles in the paradox of bias that lies at the core of the feminist empiricist project. Feminist science critics focus on identifying androcentric and sexist biases in the actual practice of science. This practice began by representing bias as a source of error. But as philosophers and historians of science joined the practice of feminist science criticism, they developed a more sophisticated way of understanding some biases as epistemic resources. Advocates of feminist science develop this theme in seeking to practice science in light of and in the service of feminist aims and values. They thereby represent feminist biases as epistemic resources.

Feminist Science Criticism: Bias as Error. Feminist science criticism originated in the critiques that working biologists, psychologists, and other scientists made of the androcentric and sexist biases and practices in their own disciplines -- especially of theories about women and gender differences that legitimate sexist practices. Exemplary works in this tradition include Bleier (1984), Fausto-Sterling (1985), Hrdy (1981), Leacock (1981), Sherif (1987), and Tavris (1992). The criticism takes many forms.

(1) Studies of how the exclusion or marginalization of women scientists impair scientific progress. For example, the failure to provide Barbara McClintock with professional standing, resources, and access to graduate students delayed incorporation of her pioneering discoveries of genetic transposition into mainstream biology (Keller 1983). (2) Studies of how the applications of science and technology disadvantage women and other vulnerable groups, treat their interests as less important, or express contempt for them. Examples include eugenics (Hubbard 1990), and economic development policies that reinforce gender hierarchy by offering training and resources to men, but not women, in developing countries (Waring 1990). Such practical ill-effects of science applications can be traced in part to epistemic defects in the underlying science -- to bogus concepts of race in the case of eugenics, and to failures to recognize women's work as contributing to the "economy" in the case of sexist development policies. (3) Studies of how science has ignored women and gender, and how turning attention to these issues may require revisions of accepted theories. Hays-Gilpin and Whitley (1998) document particularly dramatic examples of this in the field of archaeology. (4) Studies of how biases toward working with "masculine" cognitive styles -- for example, toward centralized, hierarchical control models of causation as opposed to "feminine" (contextual, interactive, diffused) models -- have impaired scientific understanding, for example, in studies of slime-mold (Keller 1985b) and molecular biology (Spanier 1995). (5) Studies of how research into sex differences and women's and men's "natures" that reinforces sex stereotypes and sexist practices fail to live up to standards of good science -- for example, in drawing inferences on the basis of miniscule sample sizes or correlations not tested against an appropriately designed control group, or in ignoring disconfirming data (Fausto-Sterling 1985, Tavris 1992). Gender bias may also be revealed in the conceptual framework of the theory in question -- for example, in representing subjective gender identification as a dichotomous variable, thereby eliminating other possibilities, such as androgyny, from consideration (Bem 1993).

In all of these cases, gender bias is represented as a cause of error, or at least delay in recognizing the truth. But, as philosophers and historians of science joined the practice of feminist science criticism, alternative models of gender bias were developed, sometimes in cooperation with working scientists. Exemplary works of feminist science criticism by philosophers and historians of science include Haraway (1989), Harding (1986, 1991, 1993, 1998), Lloyd (1993), Longino and Doell (1983), Schiebinger (1989), and Wylie (1996). Although some of this work is devoted to exposing errors caused by sexist and androcentric bias, some of it is devoted rather to showing how the interests in technological control that underlie the modern practice of science limit its scope and what it takes to be significant knowledge (Lacey 1999, Merchant 1980, Tiles 1987). Another core project of feminist science criticism is demonstrating that the evidence assembled on behalf of the theories under study does not compel assent to the theories. The theories go well beyond the data that support them, with the gap being filled by sexist and androcentric assumptions. Thus, Haraway (1989) uses the tools of literary theory to demonstrate how hypotheses in primatology and evolutionary theory depend on narrative conventions (for example, casting the transition from ape to hominid as a heroic drama) and tropes (for example, casting primates as mirrors of human nature). While these narrative conventions and tropes have considerable persuasive power, their appeal is rhetorical, and the evidence does not compel their selection. Beyond this negative critique, feminist science critics are interested in uncovering and defending the viability of alternative nonsexist and feminist theories of the phenomena in question. When they operate in this mode, the critics are not claiming that sexist and androcentric theories are false, but rather that they are not proven or

established, because at this stage in the development of the evidence, legitimate and at least equally viable rivals exist.

To sort out these different accounts of the cognitive role of gender bias, it is helpful to distinguish four dimensions in the evaluation of research programs: (1) conceptual criticism; (2) methodological criticism; (3) evaluating the relation of the available evidence to the program's hypotheses (does the evidence tend to confirm or disconfirm them?); (4) *comparing* the program's theory to rival theories in terms of their empirical adequacy and other epistemic values. Bias in a research program is revealed as *error* to the extent that it is shown to generate or rest on (1) confused or nonreferring concepts that purport to refer (for example, the concept of "race" as biological subspecies of human beings); (2) violation of valid methodological principles; (3) belief in a theory in the face of a lack of evidential support for it, or strong evidence against it; or (4) continued commitment to a theory with some evidential support, even when some rival theories dominate it with respect to *all* epistemic values, including empirical adequacy. Biases shown to generate error in this way should be stopped, through better training of scientists or the adoption and enforcement of methodological principles designed to check their influence. Feminist science criticism in the bias-as-error mode parallels the heuristics-and-biases tradition of psychology (Kahneman, Slovic and Tversky 1982), a tradition which has already been taken up in naturalized epistemology and philosophy of science (e.g., Solomon 1994). On a normative level, it generates methodological principles for engaging in nonsexist science. Exemplary normative (methodological) works generated by feminist science criticism include Altmann (1974) and Eichler (1988).

Bias in a research program is shown to be *limiting* or *partial*, but not necessarily erroneous, to the extent that avoids clear error and generates (1) a limited range of concepts and/or (2) uses a limited range of methods, (3) has some empirical successes, while (4) rival theories, depending on different concepts and/or methods, can also claim to avoid clear error and to possess some empirical successes or other epistemic virtues not possessed by the research program in question. Such biases are *legitimate*: it is rationally acceptable to conduct scientific inquiry under the influence of such biases. Indeed, empirical investigations into the workings of the human mind strongly suggest that we have no choice but to think in accordance with some biases (Chomsky has shown, for example, that we have innate ideas of deep grammatical structures). Moreover, the underdetermination of theory by data implies that without some biases, we would be unable to make sense of our world at all (Antony 1993). When biases are partial but not clearly erroneous, they serve a positive *generative* function: they produce new concepts, methods, and hypotheses that open up new aspects of the world for understanding. They are *resources* for enhancing our grasp of the world. From a normative point of view, feminist philosophers of science argue that we have an epistemic interest in ensuring that certain limiting biases do not dominate research *to the exclusion of other generative biases that would generate rival theories possessing a different range of important empirical successes.* The point in exposing the androcentric and sexist biases lying behind certain research theories is not to show that they are false (they might in the end be empirically vindicated), but to make salient the room for alternative programs not based on such biases.

Feminist Science: Bias as Resource. Most advocates of *feminist science* argue, in this vein, that scientific inquiries informed by feminist values are based on legitimate, generative limiting biases. They argue not that feminist sciences should exclude other ways of doing science, but that feminist sciences

should be included as among the legitimate choices available to investigators. This picture of science is pluralistic, compatible with the postmodern rejection of "totalizing narratives," but more inclined than postmodernists to explain the persistence of pluralism in the social and applied sciences in scientific realist terms: science is disunified because the world is rich with a multitude of cross-cutting structures, which no single theoretical vocabulary can capture. Different communities have interests in different aspects of reality, so leaving them free to follow their interests will reveal different patterns and structures in the world (Harding 1998; Longino 2001).

Against this pluralistic view, some advocates of feminist science define it in terms of adherence to specific ontologies and methodologies expressing a "feminine" cognitive style (Duran 1991, Keller 1983, 1985a). On this conception, for example, the content of any feminist theory should have a relational rather than an atomistic ontology, favor the concrete over the abstract, avoid generalizations about women in favor of exposing the richness and particularity of different women's lives and perspectives, and so forth. Its methods should encompass intuition, emotional engagment, and other cognitive styles associated with a feminine sensibility. This view has had perhaps its greatest impact in feminist works attacking quantitative methods in the social sciences. For example, Stanley and Wise (1983) argue that only qualitative methods that accept women's reports of their experiences in their own terms, refusing to generalize, can uphold feminist values of respecting differences among women and avoiding the replication of power differences between researchers and research subjects.

Pluralist feminist scientists and philosophers of science have vigorously contested these attempts to define feminist science in terms preferred content and "feminine" method. They argue that many questions of interest to feminists are best answered with quantiative methods (Jayaratne and Stewart, 1991), and indeed that feminists may properly make use of a wide range of methods (Harding 1987, Nielsen 1990, Reinharz 1992). Feminist science is not defined by its content, but rather by the pragmatic interests that generate the questions it asks. (Sometimes this distinction is cast as one between "feminist science" and "doing science as a feminist.") Feminists are interested in uncovering the causes of women's oppression, revealing the dynamics of gender in society, and producing knowledge that women can use to overcome the disadvantages to which they are subject. Forms of knowledge that simply valorize the "feminine" may not be helpful to women who would be better off not having norms of femininity imposed on them. In any event, feminist pluralists argue that advocates of "feminine" science have not shown that feminine cognitive styles and ontologies are, as a general matter, better able to track the truth (Longino 1989).

If feminist science amounts to "doing science as a feminist" -- that is, using science to answer questions generated by feminist interests -- one may ask whether it differs in any substantive respect from the science that is already practiced by nonfeminists. Feminist pluralists reply that scientific practice is already highly disunified; philosophies of the special sciences reveal great variations in methods, background assumptions, sources of evidence, cognitive values, and interpretive strategies (Longino 2001). So the dichotomy between feminist and mainstream science presupposed by the question is false. Doing biology, primatology, anthropology, archaeology, psychology, economics, history or any other special science as a feminist -- that is, with the aim of answering feminist questions -- has resulted in many and various *local* methodological innovations, discoveries of new sources of evidence, and

developments of alternative theories (see, for example, Bell, Caplan and Karim 1993; Haraway 1989; Hays-Gilpin and Whitley 1998; Nielsen 1990). These are then made available to inquirers asking other, nonfeminist questions. Thus, there is no presumption that certain methods, evidence, etc. are uniquely available to serve feminist cognitive interests.

Nevertheless, there are some common threads in "doing science as a feminist" that *tend* to favor certain types of representation over others (Longino 1994). Feminists are interested in epistemic practices that reveal the operations of gender in the world, and opportunities for women to resist and transform these operations. One way gender bias operates to reinforce sexism is through the perpetuation of categorical, dichotomous thinking which represents masculinity and femininity as "opposites," femininity as inferiority, and nonconformity to gender norms as deviant. This gives feminists an interest in the value of "ontological heterogeneity" -- using categories that permit the observation of within-group variation and that resist the representation of difference from the group mean as a form of deviance. Gender bias also reinforces sexism through single-factor causal models that attribute seemingly intrinsic powers to men by neglecting their wider context. The value of "complexity of relationship" favors the development of causal models that facilitate the representation of features of the social context that support male power, including female participation and complicity. Other feminist cognitive values involve the *accessibility* of knowledge: feminist favor knowledge that "diffuses power" in being cast in a form usable to people in subordinate positions, who usally lack technical expertise and access to expensive equipment. This interest underlies the appropriate technology movement in developing countries. For similar reasons, feminists are more interested in knowledge applicable to meeting human needs than in research programs with little prospect of advancing these interests. These values are feminist in the sense of advancing feminist interests, but their usefulness is not confined to feminism. None of these feminist cognitive values displace or compete with the search for truth, because doing science as a feminist, like doing science with any other interest in mind (for example, medical or military interests) involves commitment to the cognitive value of producing empirically adequate theories.

Feminist Defenses of Value-Laden Inquiry

The Challenge of Value-Neutrality. The theory and practice of feminist science raises the question of how any inquiry shaped by moral, social, and political interests can simultaneously be faithful to the fundamental epistemic interest in truth. Against the project of feminist science, many philosophers hold that true science is neutral among social, moral, and political values. Lacey (1999) usefully distinguishes the following claims of value-neutrality: (1) *Autonomy*: science progresses best when uninfluenced by social/political movements and values. (2) *Neutrality*: scientific theories do not imply or presuppose any judgments about noncognitive values, nor do scientific theories serve any particular noncognitive values more fully than others. (3) *Impartiality*: The only grounds for accepting a theory are its relations to the evidence. These grounds are impartial among rival noncognitive values.

Of these claims, neutrality is the most dubious, because it depicts the grounds for accepting social, political and moral values as utterly detached from evidence about human potentialities and about what happens when people try to realize these values in practice. If this were true, then the defenders of

keeping mathematics a male preserve would not have bothered arguing that women were not intellectually capable of doing mathematics and that their uteri would wander if they tried to do it -- and feminists would not have bothered disputing those claims. Neutrality is less a claim about the character of science than about the justification of social and political values. As a categorical claim about the latter, it is false. Taylor (1985) and Tiles and Oberdiek (1995) show, in detailed case studies, how scientific theories do serve some social and political values more than others.

The core claim of value-neutrality, however, is impartiality. The thought that underwrites impartiality is that scientific theories aim at the truth, at what *is* the case, whereas value judgments deal with what *ought* to be the case. Even if neutrality is false, because facts constitute part of the warrant for value judgments, the converse is not true. Only facts can supply the warrant for other facts. Autonomy, in turn, is defended as a means to ensure that science satisfies the demands of impartiality. Social and political movements are thought to threaten autonomy because their primary influence on science is thought to consist in pressuring scientists to ignore the facts and validate their worldviews. Defenders of impartiality object to the very idea of feminist science because they view it as threatening impartiality.

The Basic Underdetermination Argument. Feminist empiricists reply to the challenge of value-neutrality by extending Quine's argument that theory is underdetermined by evidence (Longino 1990, Nelson 1993). Any body of observations counts as evidence for particular hypotheses only in conjunction with certain background assumptions. Vary the background assumptions, and the same observations will support quite different hypotheses. For example, the failure to observe stellar parallax in the 17th century was taken as evidence that the Earth stands still by geocentrists, and as evidence that the stars are very far away by heliocentrists. No logical principle stops scientists from choosing different background assumptions against which to interpret their observations. In practice, scientists face some constraints in the selection of background assumptions, based on cognitive values such as simplicity and conservatism (resistance to revising deeply entrenched assumptions on which many other beliefs depend). But these cognitive values rarely reduce the scope for choice down to one option, and their interpretation and weights are contestable in any event (geocentrism was overturned only by overriding conservatism). Feminist empiricists conclude that, given the scope for choice in background assumptions, nothing stops scientists from selecting their background assumptions on account of their fit with social and political values, or indeed any other preference or interest. It follows that feminist scientists may select their background assumptions on account of their fit with feminist values.

Putnam (1981) has advanced a similar argument, carried to feminist conclusions by Nelson (1993). Value judgments operate like factual judgments in the web of belief, such that values judgments figure in the background assumptions that support factual judgments, and vice-versa. If the web of belief integrates judgments of fact and of value, then there is no clear distinction between these two judgment types. So there is no good argument against permitting feminist values to shape scientific judgments.

The underdetermination argument pries open a *potential* space for social values in science. But it is not sufficient to demonstrate the legitimacy of any *particular* ways of introducing feminist values into science. Feminist science critics and feminist scientists agree that there are cognitively illegitimate as well as cognitively legitimate ways for social values to influence science. That is the basis for

distinguishing error-generating biases from biases that serve as cognitive resources, a distinction required to dissolve the paradox of bias. Standing alone, the underdetermination argument does not help us discriminate one from the other. Additional criteria are needed.

One lesson about what to look for can be drawn from earlier debates over the theory-ladenness of observation. It is now generally agreed that the theory-laden character of observations does not threaten their status as evidence for a theory, provided that the theories presupposed in those observations do not immediately include the very theory being tested by those observations. Circularity, at least of a narrow sort, should be avoided. Similarly, the chief danger of value-laden inquiry is a kind of circularity of wishful thinking. The value-laden character of the background assumptions linking evidence to theories should not foreclose the possibility of discovering that one's values are mistaken, because based on erroneous beliefs about human potentialities and the consequences of putting certain values into practice. (Notice that it makes sense to worry about the danger of wishful thinking only if the neutrality thesis is false.) If women really can't do math, or their uteri really do migrate when they try (causing hysteria, as the sexist theory held), the values incorporated into feminist science should not close off this possibility in advance. Although, in setting out to test these sexist hypotheses, women scientists presuppose their own mathematical competence, this does not preclude their discovering otherwise. They need only open their calculations to public criticism to keep this possibility alive. If their evidence disconfirms the sexist hypotheses, and their calculations survive public scrutiny, then they have not run a vicious circle.

The Basic Pragmatic Strategy. The above reflections provide a standard for determining when socially value-laden inquiry has gone wrong. But they do not explain what positive epistemic influence they could have. How could they function as an epistemic *resource?* Feminist epistemologists at this point stress the pragmatic functions of inquiry (Anderson 1995b). All inquiry begins with a question. Questions may be motivated not only by the purely cognitive interest of curiosity, but by various practical interests in understanding the nature and causes of situations one judges to be problematic, and in finding out how to improve those situations. The resulting product of inquiry -- a theory or set of systematically connected beliefs -- should therefore be shaped to these practical-cum-cognitive interests. The pragmatic aspects of inquiry introduce new dimensions of evaluation to theories. We can ask not only whether the theories are backed by sufficient evidence to warrant their acceptance, but whether they are cast in forms that are cognitively accessible to the situated knowers who want to use these theories, whether they are useful to these knowers (help them solve their problems), and whether they answer the questions they were designed to answer. A set of statements can be true, yet fail these pragmatic tests.

Even the staunchest defenders of the value-neutrality of science acknowledge that pragmatic factors legitimately influence the choice of objects of study. In this function, then, pragmatic interests, including social and political values, are epistemic resources: inquirers with different interests will study and make discoveries about different aspects of the world. But the defenders of value-neutral science contend that once inquirers decide where to cast their flashlight, what gets lit up is determined entirely by the nature of the world. Feminist epistemologists argue that the light of practical interests penetrates more deeply into what is discovered than this. Knowers (subjects) play a more active role in constituting the object of knowledge than the flashlight metaphor suggests. (This is one thing feminist epistemologists mean when they say they reject "the subject-object dichotomy".) "Constitution" has two senses, representational and

causal. In the representational sense, knowers constitute the object of knowledge in choosing the terms in which they represent it, and in defining the context in which it is represented as operating. If knowing is like seeing, all seeing is a form of "seeing as" -- and different interests will make us see the "same" things differently (Longino 1990). This is a straightforward implication of the fact of situated knowing. In the causal sense, some representations have a causal impact on what is represented. When what we are representing is *ourselves*, uptake of our self-representations will change who we are and what we do. This follows from our agency, which is the determination to govern ourselves by our self-understandings. This is sometimes what is meant by the claim that subjects, or their identities, are "socially constructed."

The basic pragmatic strategy for defending feminist science, or any inquiry shaped by social and political values, is to show how the pragmatic interests of that inquiry license or require a particular mode of influence of values on the process, product, and uptake of the product of inquiry, while at the same time leaving appropriate room for evidence to play its warranting role. Values and evidence play different, cooperative roles in properly conducted inquiry; values do not *compete* with evidence for the determination of belief (Anderson 1995b).

A Catalogue of Types of Legitimate Influence of Social Values in Science. One can examine the actual operation of particular values in particular scientific investigations and judge, on a case by case basis, whether the values are closing off the possibility of discovering unwelcome facts, leading scientists to reason in a vicious circle, or insulating their findings from critical scrutiny, or whether the values are enabling the discovery of new facts -- whether they are, in short, obstructing or facilitating the search for knowledge. Such judgments are contextual and subject to revision in light of new evidence. What follows is a catalogue of *types* of influence of social values that feminist epistemologists and philosophers of science have argued may *in principle* legitimately influence theory choice (although whether their influence is epistemically good or bad in a particular case requires further investigation).

Selection and weighting of cognitive values. Kuhn (1977) argued that scientists need to appeal to cognitive values to take up the slack between theory and evidence. His list of cognitive values included accuracy (empirical adequacy or truth), scope, simplicity, fruitfulness, internal consistency and consistency with other beliefs (conservatism). As we have seen above, Longino (1994) argues that feminists have reason to prefer theories that manifest other cognitive values, such as diffusion of power. Diffusion of power, like simplicity, is not a truth-oriented cognitive value. Both count as cognitive values because they make theories cognitively accessible, comprehensible to our finite minds. Diffusion of power recognizes that cognitive accessibility is relative to the situation of the knower. (Longino's characterization of other values of feminist science, such as ontological heterogeneity and complexity of relationship, as "cognitive" values is something of a misnomer -- these fit better under the rubrics of classification and models, below.) Both simplification and diffusion of power stand in tension with truth, in that theories that embody them not only ignore many complex, messy truths, but may even make false claims. Whether this is bad depends on whether the truths ignored or the inaccuracies embraced are *important*, and this can be judged only in relation to the interests of the investigator. All legitimate research programs must seek to incorporate the value of empirical adequacy, which requires at least that theories try to *approximate* the truth. But how much accuracy this requires depends on how much the expected usefulness of the knowledge will be compromised by larger margins of error. The situation and

pragmatic interests of the inquirer or of the potential users of a theory may therefore legitimately affect the selection and weighting of cognitive values in theory choice.

Standards of Proof. By convention, social scientists reject the null hypothesis (that observed results in a statistical study reflect mere chance variation in the sample) only for P-values < 5%, an arbitrary level of statistical significance. Bayesians and others argue that the level of statistical significance should vary, depending on the relative costs of type I error (believing something false) and type II error (failing to believe something true). In medicine, clinical trials are routinely stopped and results accepted as genuine notwithstanding much higher P-values, if the results are dramatic enough and the estimated costs to patients of not acting on them are considered high enough. (The cost of not providing a potentially effective treatment may be death, while the cost of providing a useless treatment may be small.) This practice explicitly incorporates social value judgments in the standard of proof required before results are accepted. Hare-Mustin and Maracek (1994) argue, by parallel reasoning, that whether studies that find gender differences, or that fail to find them, should be accepted depends on the relative costs of Alpha Bias (exaggerating differences) and Beta Bias (neglecting differences) in the context at hand.

Classification. The ways observed phenomena are classified may legitimately depend on the values of the researcher. In medicine, the distinction between health and disease reflects moral judgments about human welfare and appropriate ways of dealing with problems, as well as judgments about causation. A condition regarded as bad for human beings is not classified as a disease unless some kind of medical therapy is considered both an appropriate and a potentially effective way to deal with it. Feminist inquiries, too, raise questions about the causes of women's oppression that require classifying phenomena as instances of rape, sexual objectification, sex discrimination, and so forth -- classifications all tied to their meeting both empirical and evaluative criteria (Anderson 1995a, 1995b).

Methods. The methods selected for investigating phenomena depend on the questions one asks and the kinds of knowledge one seeks, both of which may reflect the social interests of the investigator. Experimental methods in social science may be good for discovering factors that can be used to control people's behavior in similar settings. But to grasp their behavior as *action* -- that is, as attempts by agents to govern their behavior through their self-understandings -- requires different empirical methods, including qualitative interviews (which allow subjects to delineate their own systems of meaning) and participant observation. Standpoint theories, as critical theories, aim as well at empowering the subjects of study by helping them forge liberatory self-understandings, and these, too, may require different methods of inquiry -- for example, consciousness-raising (MacKinnon 1999).

Causal Explanations; Explanations of Meaning; Narratives. For most phenomena, the number of factors that have a causal impact on their occurrence is vast -- too large to comprehend or test in a single model. Investigators must therefore select a subset of causal factors to include in the models they test. This selection may be based on considerations of cost or availability -- some types of data are hard or expensive to get; cheap and accessible methods may be better suited to testing the causal influence of some variables than others. The selection of causal variables may also be based on fit with the social or personal interests of the investigator (Longino 1990, 2001). These interests often reflect background social and moral judgments of blame, responsibility, and acceptability of change. To take an innocuous

case, in most contexts, what is singled out as a cause of dangerous fires is a spark, flame, or flammable material, not the presence of oxygen. The items judged possible to change, or worth changing, are the focus of causal explanation. To take a more controversial case, conservatives are more likely to study divorce and out-of-wedlock birth as causes of women's poverty, whereas feminists are more likely to focus on other causes -- for example, the exclusion of women from better-paid jobs, the weak bargaining power of women in marriage, and norms of masculinity that induce fathers to avoid significant participation in child-rearing, thereby forcing women to forego earnings in taking up the slack. Notice that these causal explanations are not incompatible. All the causal factors cited may contribute to the feminization of poverty.

Often what inquirers seek is not merely a set of facts, but what the facts mean. Meaning holism implies that the meaning or signficance of facts depends on their relations to other facts. Even if two inquirers agree on the causal facts, they may still disagree about their meaning because they relate the facts in different ways, reflecting their background values. Feminists may agree with conservatives that divorce is a cause of the feminization of poverty, but deny that this means that women are better off married. They argue that marriage itself, with its gendered division of domestic and market labor, constitutes one of the major structural disadvantages women face, setting them up for worse outcomes in the event of divorce (Okin 1989). Conservatives, viewing marriage as an indispensible condition of the good life, are no more willing to view marriage in this light than most people would be willing to blame oxygen for the occurrence of house fires. It might be thought that scientists should stick to the facts and avoid judgments of meaning. But most of the questions we ask demand answers that fit facts into larger, meaningful patterns. Scientists therefore cannot help but tell stories, which require the selection of narrative frameworks that necessarily go beyond the facts (Haraway 1989). This selection may depend both on their fit with the facts and on their fit with the background values of the storyteller.

Framework Assumptions. As we ascend to higher levels of abstraction, very general framework assumptions come into play in constituting the object of study. Some of these are disciplinary. Economics studies humans as self-interested, instrumentally rational choosers. Social psychology studies humans as responding to socially meaningful situations. Behaviorism studies humans as controlled by objectively defined environmental variables. Behavioral genetics studies humans as controlled by their genes. These are all forms of "seeing as." Longino (1990) and Tiles (1987) argue that the selection of framework assumptions may depend on their fit with the interests of the inquirer. Feminists are interested in promoting women's agency, so they tend to prefer frameworks that permit the representation of women as agents. This selection does not guarantee that they will confirm the background assumption that women are agents. Causal models that include only agentic variables may not explain much of the variation in women's behavior; models that include both agentic and nonagentic variables may find that the latter explain all of the variation. The value-laden selection of framework assumptions need not lead to a vicious circle of reasoning, because it is still left up to the evidence to determine how successful the assumptions are in explaining the phenomena of interest.

Pluralism as the upshot of value-laden inquiry. Because the interests and values of inquirers vary, and inquirers select background assumptions in part for their fit with their interests and values, their background assumptions will also vary. Rather than lamenting this fact, feminist epistemologists urge us

to embrace it (Haraway 1991, Harding 1998, Longino 2001). A pluralism of theories and research programs should be accepted as a normal feature of science -- as it is, certainly, in the human sciences. As long as the different research programs are producing empirical successes not produced by the others, and avoiding clear error and viciously circular reasoning, there is good reason to treat the value-biases animating them as epistemic resources, helping us discover and understand new aspects of the world and see them in new perspectives, rather than as obstacles to the search for truth. Feminist science takes its place as one set of legitimate research programs among others, rather than as something that replaces the others. The price to be paid for this is the disunity of science. This does not imply relativism. Value-laden research programs are still open to internal and external critique. A naturalized epistemology that rejects neutrality allows that observations may undermine any background assumptions, including value judgments.

Trends in Feminist Epistemology

When Harding (1986) proposed her classification of feminist epistemologies into empiricism, standpoint theory, and postmodernism, she cast them as offering three fundamentally contrasting frameworks. Empiricism was thought to presuppose an unsituated, politically neutral subject of knowledge, whereas standpoint theory and postmodernism offered different approaches to the problem of situated knowledge -- the first upholding an epistemic privilege of one situation over others, the other embracing a relativism of standpoints. Trends in feminist epistemology in the last ten years have blurred the distinctions among feminist empiricism, standpoint theory, and feminist postmodernism -- trends Harding herself both predicted and advanced (1990, 1991, 1998). Most importantly, all three approaches to feminist epistemology embrace pluralism and reject totalizing theories. All three approaches also reject the traditional epistemological project of validating epistemic norms from a transcendent viewpoint, because they deny that there is any such viewpoint to be had.

Feminist standpoint theory. The postmodernist critique of standpoint theory, in conjunction with the proliferation of subaltern women's standpoints (black, Latina, lesbian, postcolonial, etc.) has led most standpoint theorists to abandon the search for a *single* feminist standpoint that can claim overarching epistemic superiority. Feminist standpoint theory has therefore moved in a pluralistic direction, acknowledging a multiplicity of epistemically informative situated standpoints (Harding 1991, 1998; Collins 1990). They claim that there are important things to learn from taking seriously the perspectives of all marginalized groups -- not just of various groups of women, but men and women in postcolonial societies, men and women of color, gay men, and so forth. A system of knowledge that draws on their insights and starts from their prediaments will be richer than one that draws only on the insights and starts from the predicaments of privileged groups alone (Harding 1993, 1998). This shifts the privilege claimed on behalf of subaltern standpoints from the context of justification to the context of discovery: thinking from subaltern standpoints is more *fruitful* than confining one's thinking to dominant perspectives. And the fruitfulness in question is judged more often on pragmatic grounds -- thinking from these standpoints enables us to envision and realize more just social relations (Hartsock 1996). At the same time, many standpoint theorists have focused more sharply on the epistemic value of the *experiences* of subordinated people (as opposed to making categorical claims about group differences in cognitive style). The shift to

pluralism represents a convergence with feminist postmodernism; the shift to pragmatism and experience, a convergence with feminist empiricism.

Feminist postmodernism. Wariness of the fractionating and centrifugal forces in postmodernism has led some feminists sympathetic to postmodernism to seek middle, more stable grounds that feminist empiricists, standpoint theorists, and postmodernists can share. Haraway (1989) stands out among feminist postmodernists for the tributes she pays to the achievements of feminist scientists working within empiricist standards of evaluation. She also seeks to reconstruct ideas of objectivity and epistemic responsibility consistent with situated knowledge (1991). Fraser (1995) and Fraser and Nicholson (1990) also urge a reformulation of the lessons of postmodernism, toward pragmatism, fallibilism, and contextualization of knowledge claims -- all features fully compatible with naturalized feminist empiricism -- as against categorical rejections of large-scale social theory, history, normative philosophy, and even humanist values. While it remains to be seen whether feminist postmodernists will actually take up these calls, they signal directions in which postmodernism could be taken.

Feminist empiricism. While early, nonphilosophical feminist science criticism by working scientists may have presupposed a naive version of empiricism, attempts by feminist epistemologists to make sense of feminist science criticism have, following Quine, incorporated explicitly pragmatist and naturalizing themes into feminist empiricism. Thus, feminist empiricists today stress the centrality of situated knowledge, the interplay of facts and values, the absence of transcendental standpoints, and the plurality of theories. These themes converge with those of postmodernism.

Remaining differences. The differences that remain among feminist postmodernists, empiricists, and standpoint theorists partially reflect different choices of tools. Feminist postmodernists use the tools of poststructuralism and literary theory. Feminist empiricists prefer the tools of analytic philosophy of science. Some versions of standpoint theory, such as Collins' (1990), rest on an identity politics alien to both postmodernists and empiricists. (To the extent that standpoint theory remains tied to a materialist epistemology, as in Hartsock (1996) and MacKinnon (1999), it is fully compatible with feminist empiricist naturalized epistemology.)

Other differences reflect different attitudes toward and conceptions of objectivity. Although feminist postmodernism has relativist tendencies, its skepticism and stress on instability undermines both the purportedly all-encompassing stance of objectivity and the self-contained, complacent parochiality of relativism. What's missing is not the thought that critique is possible, but any form of critique that enables one to build and synthesize rather than tear down and deconstruct claims to know. Although Haraway reconceives objectivity in terms of epistemic responsibility, it is hard to hold knowers accountable for their claims if they never stick to any claims for very long (Bordo 1990). Harding (1993, 1998) continues standpoint theory's project of trying to identify an epistemically superior approach that can claim "strong objectivity." Feminist empiricism offers an alternative procedural account of objectivity (Longino 2001). Both envision critical theory as a constructive, not just a deconstructive project. But standpoint theory remains residually attached to claims of epistemic superiority for feminist science, whereas feminist empiricism is concerned rather simply to make room at the pluralist table for feminist approaches to science, without asserting epistemic superiority on their behalf.

Bibliography

- Addelson, Kathryn, 1983. "The Man of Professional Wisdom." In Harding and Hintikka, 165-86.
- Alcoff, Linda, and Elizabeth Potter, (eds.) 1993. *Feminist Epistemologies*. New York: Routledge.
- Altmann, Jeanne, 1974. "Observational study of behavior: Sampling methods". *Behavior* 49:227-267
- Anderson, Elizabeth. 1995a. Feminist Epistemology: An Interpretation and Defense. *Hypatia* 10:50-84.
- Anderson, Elizabeth. 1995b. Knowledge, Human Interests, and Objectivity in Feminist Epistemology. *Philosophical Topics* 23:27-58.
- Antony, Louise M., 1993. "Quine as Feminist: The Radical Import of Naturalized Epistemology". In Antony and Witt.
- Antony, Louise M., and Charlotte Witt, (eds.) 1993. *A Mind of One's Own.* Boulder: Westview Press.
- Bar On, Bat-Ami. 1993. "Marginality and Epistemic Privilege". In Alcoff and Potter.
- Belenky, Mary Field, et al, 1986. *Women's Ways of Knowing*. New York: Basic Books.
- Bell, Diane, Patricia Caplan, and Karim Wazir-Jahan Begum, 1993. *Gendered Fields : Women, Men, and Ethnography*. London/New York: Routledge.
- Bem, Sandra, 1993. *The Lenses of Gender*. New Haven: Yale University Press.
- Benhabib, Seyla, 1995. "Feminism and Postmodernism". In Benhabib, Butler, Cornell and Fraser.
- Benhabib, Seyla, Judith Butler, Drucilla Cornell and Nancy Fraser, 1995. *Feminist Contentions*. New York: Routledge.
- Bleier, Ruth, 1984. *Science and Gender: A Critique of Biology and its Theories on Women.* New York: Pergamon.
- Bordo, Susan, 1987. *The Flight to Objectivity : Essays on Cartesianism and Culture.* Albany: State University of New York Press.
- Bordo, Susan. 1990. "Feminism, Postmodernism, and Gender Skepticism." In Nicholson.
- Butler, Judith, 1990. *Gender Trouble*. New York: Routledge.
- Butler, Judith, 1993. *Bodies that Matter*. New York: Routledge.
- Campbell, Richmond, 1998. *Illusions of Paradox*. Lanham, Md.: Rowman & Littlefield.
- Clayton, Susan and Faye Crosby, 1992. *Justice, Gender, and Affirmative Action.* Ann Arbor: University of Michigan Press.
- Code, Lorraine, 1991. *What Can She Know?* Ithaca, New York: Cornell University Press.
- Collins, Patricia Hill, 1990. *Black Feminist Thought*. Boston: Unwin Hyman.
- Crenshaw, Kimberlé, 1989. "Demarginalizing the Intersection of Race and Sex". *University of Chicago Legal Forum,* pp. 139-167.
- Diamond, Cora, 1991. "Knowing Tornadoes and Other Things". *New Literary History* 22:1001-15.
- Duran, Jane, 1991. *Toward a Feminist Epistemology*. Savage, Md.: Rowman & Littlefield.
- Eichler, Margrit, 1988. *Nonsexist Research Methods: A Practical Guide.* Winchester, Mass.: Allen & Unwin.

- Fausto-Sterling, 1985. *Myths of Gender*. New York: Basic Books.
- Flax, Jane, 1983. "Political Philosophy and the Patriarchal Unconscious". In Harding and Hintikka.
- Fraser, Nancy, 1995. "False Antitheses." In Benhabib, Butler, Cornell and Fraser.
- Fraser, Nancy and Linda Nicholson, 1990. "Social Criticism without Philosophy." In Nicholson.
- Garry, Ann, and Marilyn Pearsall, (eds.) 1989. *Women, Knowledge, and Reality*. Boston: Unwin Hyman.
- Gilligan, Carol, 1982. *In a Different Voice.* Cambridge, Mass.: Harvard University Press.
- Haraway, Donna, 1989. *Primate Visions.* New York: Routledge.
- Haraway, Donna, 1991. "Situated Knowledges". In *Simians, Cyborgs, and Women*. New York: Routledge.
- Harding, Sandra, 1986. *The Science Question in Feminism*. Ithaca: Cornell University Press.
- Harding, Sandra, (ed.) 1987. *Feminism and Methodology : Social Science Issues*. Bloomington: Indiana University Press.
- Harding, Sandra, 1990. "Feminism, Science, and the Anti-Enlightenment Critiques" In Nicholson.
- Harding, Sandra, 1991. *Whose science? Whose knowledge?* Ithaca, N.Y.: Cornell University Press.
- Harding, Sandra. 1993. "Rethinking Standpoint Epistemology: ‘What is Strong Objectivity?’" In Alcoff and Potter.
- Harding, Sandra, (ed.) 1993. *The "Racial" Economy of Science.* Bloomington: Indiana University Press.
- Harding, Sandra, 1998. *Is science multicultural?: Postcolonialisms, feminisms, and epistemologies.* Bloomington, Ind.: Indiana University Press.
- Harding, Sandra, and Merrill Hintikka, (eds.) 1983. *Discovering Reality*. Dordrecht, Holland: D. Reidel; Boston: Kluwer.
- Harding, Sandra, and Jean O'Barr, (eds.) 1987. *Sex and Scientific Inquiry*. Chicago: University of Chicago Press.
- Hare-Mustin, Rachel and Jeanne Maracek, 1994. "Gender and the Meaning of Difference: Postmodernism and Psychology". In Anne Herrmann and Abigail Stewart, eds. *Theorizing Feminism*. Boulder, Col.: Westview.
- Hartsock, Nancy. 1987. "The Feminist Standpoint: Developing the Ground for a Specifically Feminist Historical Materialism." In Harding 1987.
- Hartsock, Nancy. 1996. "Comment on Hekman's ‘Truth and Method’: Truth or Justice" *Signs* 22:367-73.
- Haslanger, Sally, 2000. "Gender and Race: (What) Are They? (What) Do We Want Them To Be?" *Nous* 34 (1). [Available online]
- Hays-Gilpin, Kelley; Whitley, David, 1998, eds. *Reader in Gender Archaeology*. New York: Routledge.
- Hrdy, Sarah, 1981. *The Woman that Never Evolved.* Cambridge, Mass.: Harvard University Press.
- Hubbard, Ruth, 1990. *The Politics of Women's Biology*. New Brunswick, [N.J.]: Rutgers University Press.
- Hundleby, Catherine, 1997. "Where Standpoint Stands Now," *Women and Politics* 18:25-.
- Hull, Gloria, Patricia Scott, and Barbara Smith, eds. 1982. *All the Women Are White, All the*

Blacks Are Men, But Some of Us Are Brave. Old Westbury, N.Y.: Feminist Press.

- Kahneman, Daniel, Paul Slovic and Amos Tversky, 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
- Keller, Evelyn Fox, 1983. *A Feeling for the Organism.* San Francisco: W.H. Freeman.
- Keller, Evelyn Fox, 1985a. *Reflections on Gender and Scienc*e. New Haven: Yale University Press.
- Keller, Evelyn Fox, 1985b. "The Force of the Pacemaker Concept in Theories of Aggregation in Cellular Slime Mold". In Keller 1985a.
- Keller, Evelyn Fox, 1992. *Secrets of Life, Secrets of Death : Essays on Language, Gender, and Science*. New York: Routledge.
- Jaggar, Alison, 1989. "Love and Knowledge: Emotion in Feminist Epistemology". In Garry and Pearsall.
- Jayaratne, Toby and Abigail Stewart, 1991. "Quantitative and Qualitative Methods in the Social Sciences: Current Feminist Issues and Practical Strategies" in Mary Fonow and Judith Cook, eds. *Beyond Methodology*. Bloomington, Ind.: University of Indiana Press.
- Kalbfleisch, Pamela, (ed.) 1995. *Gender, Power, and Communication in Human Relationships.* Hillsdale, N.J.: Erlbaum.
- Kuhn, Thomas, 1977. "Objectivity, Value Judgment and Theory Choice." In *The Essential Tension.* Chicago: University of Chicago Press.
- Lacey, Hugh, 1999. *Is Science Value Free?* New York: Routledge.
- Leacock, Eleanor Burke, 1981. *Myths of Male Dominance.* New York: Monthly Review Press.
- Lloyd, Elisabeth, 1993. "Pre-Theoretical Assumptions in Evolutionary Explanations of Female Sexuality" *Philosophical Studies* 69:139-153.
- Longino, Helen, and Doell, Ruth, 1983. "Body, Bias, and Behavior." *Signs* 9.
- Longino, Helen, 1989. "Can there Be a Feminist Science?". In Garry and Pearsall.
- Longino, Helen, 1990. *Science as Social Knowledge.* Princeton, N.J.: Princeton University Press.
- Longino, Helen, 1993a. "Essential Tensions -- Phase Two: Feminist, Philosophical, and Social Studies of Science". In Antony and Witt.
- Longino, Helen, 1993b. "Subjects, Power, and Knowledge: Description and Prescription in Feminist Philosophy of Science". In Alcoff and Potter.
- Longino, Helen, 1994. "In Search of Feminist Epistemology," *Monist* 77:472-485.
- Longino, Helen, 2001. *The Fate of Knowledge*. Princeton: Princeton University Press.
- Lorde, Audre. 1984. *Sister Outsider*. Trumansburg, NY: Crossing Press.
- Lugones, Maria. 1987. "Playfulness, 'World'-Traveling, and Loving Perception." *Hypatia* 2:3-19.
- Lugones, Maria, and Elizabeth Spelman. 1983. "Have We Got a Theory for You! Feminist Theory, Cultural Imperialism, and the Demand for 'The Woman's Voice'?", *Women's Studies International Forum* 6:573-581.
- Lukács, Georg. 1971. "Reification and the Consciousness of the Proletariat". In *History and Class Consciousness,* tr. Rodney Livingstone. Cambridge, Mass.: MIT Press
- MacKinnon, Catherine, 1999. *Toward a Feminist Theory of the State.* Cambridge, Mass.: Harvard University Press.
- Marx, Karl. 1964. *The Eighteenth Brumaire of Louis Bonaparte*. New York: International Publishers.

- Merchant, Carolyn, 1980. *The Death of Nature: Women, Ecology, and the Scientific Revolution.* New York: Harper and Row.
- Moulton, Janice, 1983. "A Paradigm of Philosophy: The Adversary Method." In Harding and Hintikka.
- Nelson, Lynn Hankinson, 1990. *Who Knows : From Quine to a Feminist Empiricism*. Philadelphia, Pa.: Temple University Press.
- Nelson, Lynn, 1993. "Epistemological Communities". In Alcoff and Potter.
- Nicholson, Linda, ed. 1990. *Feminism/Postmodernism*. New York and London: Routledge.
- Nielsen, Joyce, ed., 1990. *Feminist Research Methods*. Boulder, Col.: Westview.
- Okin, Susan, 1989. *Justice, Gender and the Family*. New York: Basic Books.
- Potter, Elizabeth, 1993. "Gender and Epistemic Negotiation". In Alcoff and Potter.
- Putnam, Hilary, 1981. *Reason, Truth, and History*. Cambridge: Cambridge University Press.
- Quine, W. V. O., [Rev. ed.] 1963. "Two Dogmas of Empiricism." In *From a Logical Point of View*. New York: Harper & Row.
- Quine, W. V. O., 1969. "Epistemology Naturalized." In *Ontological Relativity and Other Essays*. New York: Columbia University Press.
- Reinharz, Shulamit, 1992. *Feminist Methods in Social Research*. Oxford: Oxford University Press.
- Rooney, Phyllis, 1991. "Gendered Reason: Sex Metaphor and Conceptions of Reason". *Hypatia* 6:77-103.
- Rose, Hilary, 1987. "Hand, Brain, and Heart: A Feminist Epistemology for the Natural Sciences". In Harding and O'Barr.
- Rose, Suzanna, 1989. "Women Biologists and the ‘Old Boy’ Network". *Women's Studies International Forum* 12(3), pp. 349-54.
- Schiebinger, Londa, 1989. *The Mind Has No Sex?* Cambridge, Mass.: Harvard University Press.
- Scott, Joan, 1991. "The Evidence of Experience" *Critical Inquiry* 17:773-97.
- Sherif, Carol, 1987. "Bias in Psychology." In Harding 1987.
- Smith, Dorothy, 1974. "Women's Perspective as a Radical Critique of Sociology". *Sociological Inquiry* 44:7-13.
- Solomon, Miriam, 1994. "Social Epistemology". *Noûs* 28:325-343.
- Spanier, Bonnie, 1995. *Im/partial Science: Gender Ideology in Molecular Biology*. Bloomington, Ind.: Indiana University Press.
- Spelman, Elizabeth, 1988. *Inessential Woman*. Boston: Beacon Press.
- Stanley, Liz and Sue Wise, 1983. *Breaking Out: Feminist Consciousness and Feminist Research.* London: Routledge and Kegan Paul.
- Tavris, Carol, 1992. *The Mismeasure of Women*. New York: Simon and Schuster.
- Taylor, Charles, 1985. "Neutrality in Political Science" in *Philosophy and the Human Sciences*. Cambridge: Cambridge University Press.
- Tiles, Mary, 1987. "A Science of Mars or of Venus?". *Philosophy* 62:293-306.
- Tiles, Mary and Hans Oberdiek, 1995. *Living in a Technological Culture*. London and New York: Routledge.
- Tuana, Nancy, (ed.) 1989. *Feminism & Science*. Bloomington: Indiana University Press.
- Waring, Marilyn, 1990. *If Women Counted.* San Francisco: HarperCollins.

- West, Candace and Don Zimmerman, 1987. "Doing Gender". *Gender and Society* 1:125-51.
- Wylie, Alison, 1996. "The Constitution of Archaeological Evidence: Gender Politics and Science" in P. Galison and D. Stump, eds., *The Disunity of Science*. Stanford: Stanford University Press.
- Young, I. M. 1990. *Throwing Like a Girl and Other Essays in Feminist Political Theory*. Bloomington: Indiana University Press.

Other Internet Resources

- Bibliography: Feminist Epistemology (by Heidi Grasswick and Nathan Anderson)
- CTRL -- Postmodernism, Social Constructivism, and Feminist Epistemology (extensive critique of these movements)
- Feminist Methodology (lecture slides on feminist methodology, with references)
- Kelley Hays-Gilpin Webpage (resource on gender and archaeology)
- Research Methods -- Doing Feminist Research (annotated bibliography by Ann Hall)
- Theory of Science (includes links to articles, bibliographies on gender & science, postmodernism, etc.)

Related Entries

feminism, interventions: feminist ethics

Elizabeth Anderson
eandersn@umich.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 9, 2000
Content last modified: November 13, 2001

Charles Sanders Peirce

Charles Sanders Peirce (1839-1914) was the founder of American pragmatism (later called by Peirce "pragmaticism"), an extender of the Scotistic theory of signs (called by Peirce "semeiotic"), an extraordinarily prolific logician and mathematician, and a developer of an evolutionary, psycho-physically monistic metaphysical system. A practicing chemist and geodesist by profession, he nevertheless considered scientific philosophy, and especially logic, to be his vocation. In the course of his polymathic researches, he wrote on a wide range of topics, ranging from mathematical logic to psychology.

Brief Biography

Charles Sanders Peirce was born on September 10, 1839 in Cambridge, Massachusetts, and died on April 19, 1914 in Milford, Pennsylvania. His writings extend from about 1857 until near his death, a period of approximately 57 years. His published works run to about 12,000 printed pages and his known unpublished manuscripts run to about 80,000 handwritten pages. The topics on which he wrote have an immense range, from mathematics and the hard sciences at one extreme, to economics, psychology, anthroplogy, history of science, and the theory of signs, at the other extreme.

Peirce's father Benjamin was Professor of Mathematics at Harvard University and was one of the founders of the U. S. Coast and Geodetic Survey as well as one of the founders of the Smithsonian Institution. The department of mathematics at Harvard was essentially built by Benjamin. From Bejamin Peirce Charles Sanders Peirce received most of the substance of his early education as well as a good deal of intellectual encouragement and stimulation. Benjamin's didactic technique mostly took the form of setting interesting problems and checking Charles's solutions of them, and in this instructional atmosphere Charles learned his lifelong habit of thinking through problems entirely on his own. To this habit, perhaps, is to be attributed Charles Peirce's originality.

Peirce graduated from Harvard in 1859 and received the bachelor of science degree in chemistry in 1863. From 1859 until late 1891 he was employed by the U. S. Coast and Geodetic Survey, mainly doing geodetic investigations. From 1879 until 1884, Peirce maintained a second job teaching logic in the Department of Mathematics at Johns Hopkins University. At the time the Department of Mathematics was headed by the famous mathematician J. J. Sylvester. This job suddenly evaporated for reasons apparently connected with the fact that Peirce's second wife was a gypsy, and a gypsy moreover with whom he had apparently cohabited before marriage. This was Peirce's only academic employment, and after losing it Peirce worked thereafter only for the U. S. Coast and Geodetic Survey. This employment was lost in late 1891 because of funding worries generated in Congress. Thereafter, Peirce eked out a living doing odd-jobs and consulting work (mainly in chemical engineering and analysis). For the remainder of his life Peirce was often in dire financial straits, and sometimes he managed to survive only because of the charity of friends, for example William James.

At age 12 Charles read a standard textbook on logic by Bishop Richard Whately, and he began reading Immanuel Kant's *Critique of Pure Reason* at age 13. After three years of careful study of Kant, Peirce concluded that Kant's system was vitiated by what he called its "puerile logic," and about the age of 16 he formed the fixed intention of devoting his life to the study of and research in logic. Although it was impossible to earn a living as a research logician, his adoption of the profession of chemistry and his practice of the profession of geodesy allowed Charles to continue for many years to engage in researches on logic. One of his logical systems was the basis for Ernst Schroeder's three-volume treatise on logic, the *Vorlesungen ueber die Algebra der Logik*, and Peirce became widely regarded as the greatest logician of his day. By all who are familiar with his work he is considered one of the greatest logicians who ever lived.

Despite Peirce's early disagreements with Kant's position, Peirce continued to respect and read the first *Critique* throughout his life, and his ultimate philosophical position has much in common with the transcendental idealism of Kant, as well as with the objective idealism of Hegel. Like Kant, Peirce even

developed a set of ultimate categories (more on which later).

Difficulty of Access to Peirce's Writings

Peirce's extensive publications are scattered among various publication media, and have been difficult to collect. Shortly after his death in 1914, his widow Juliette sold his unpublished manuscripts to the Department of Philosophy at Harvard University. Initially they were under the care of Josiah Royce, but after Royce's death in 1916, and especially after the end of the First World War, the papers were poorly cared for. Many of them were misplaced, lost, given away, scrambled, and the like. Carolyn Eisele, one of several heroes in a great effort to locate and assemble Peirce's writings, reports having found a lost trunk of Peirce's papers only in the mid-1950's; it had, apparently for decades, been secreted in an unlit, obscure part of the basement in Harvard's Widener Library.

In the 1930's volumes of *The Collected Papers of Charles Sanders Peirce* began to appear, and for almost three decades these volumes, and collections of entries culled from them were the only generally available source for Peirce's thoughts. Unfortunately, many of the entries in the *Collected Papers* are not integral pieces of Peirce's own design, but rather pieces that were cobbled together by the editors from different Peircean sources. Often a single entry will consist of stretches of writing from very different periods of Peirce's intellectual life, and the stretches may be in tension or outright contradition with each other. Such entries not only make very difficult reading if one tries to regard them as consistent sustained passages of argument, but also tend to give a false picture of Peirce as unsystematic, desultory, and much more obscure than he really is.

The only intelligent way to publish the works Peirce is chronologically and with extremely careful editing. In such a fashion, entire Peircean works can be presented, and presented in their natural temporal setting. At last, but not until the 1980's, there began to appear such a chronological edition of carefully selected works of Peirce: the *Writings of Charles S. Peirce: a Chronological Edition*. Though currently somewhat slowed by lack of proper funding, the *Chronological Edition* has succeeded in covering quite well in five volumes the major writings from 1857 to 1886. This achievement is finally making it posssible to assess the real Peirce, and in particular the development of his thinking from its earliest to its later stages. Questions long vexed in Peirce scholarship are finally beginning to be debated usefully by Peirce scholars: whether there is genuine systematic unity in Peirce's thought, whether his ideas changed or developed over time and in what particulars his thought did change, when certain notions were developed by Peirce, whether there were definite "periods" in Peirce's intellectual development, what exactly Peirce meant by some of his more obscure notions such as his universal categories (on which see below).

Deduction, Induction, and Abduction

Prior to about 1865, thinkers on logic commonly had divided arguments into two subclasses: the class of deductive arguments (a.k.a. necessary inferences) and the class of inductive arguments (a.k.a. probable

inferences). About this time, Peirce began to hold that there were two utterly distinct classes of probable inferences, which he referred to as inductive inferences and abductive inferences (which he also called hypotheses and retroductive inferences). Peirce reached this conclusion by entertaining what would happen if one were to interchange propositions in the syllogism AAA-1 (Barbara): All M's are P's; All S's are M's; therefore, All S's are P's. This valid syllogism Peirce accepted as representative of deduction. But also seemed typically to regard it in connection with a problem of sampling theory. Let us regard being an M as being a member of a population of some sort, say being a ball of the population of balls in some urn. Let us regard P as being some property a member of this population can have, say being red. And, finally, let us regard being an S as being a member of a random sample taken from the population. Then our syllogism in Barbara becomes: All balls in this urn are red; All balls in this particular random sample are taken from this urn; therefore, All balls in this particular random sample are red. Peirce regarded the major premise here as being the Rule, the minor premise as being the particular Case, and the conclusion as being the Result of the argument. The argument is a piece of deduction (necessary inference): an argument from population to random sample.

Let us now see what happens if we form a new argument by interchanging the conclusion (the Result) with the major premise (the Rule). The resultant argument becomes: All S's are P's (Result); All S's are M's (Case); therefore, All M's are P's (Rule). This is the invalid syllogism AAA-3. But let us now regard it as pertaining to sampling theory. The argument becomes: All balls in this particular random sample are red; All balls in this particular random sample are taken from this urn; therefore, All balls in this urn are red. What we have here is an argument from sample to population, and this is what Peirce understood to be the core meaning of induction: argument from random sample to population.

Let us now go further and see what happens if, from the deduction AAA-1, we form a new argument by interchanging the conclusion (the Result) with the minor premise (the Case). The resultant argument becomes: All M's are P's (Rule); All S's are P's (Result); therefore, All S's are M's (Case). This is the invalid syllogism AAA-2. But let us now regard it as pertaining to sampling theory. The argument becomes: All balls in this urn are red; All balls in this particular random sample are red; therefore, All balls in this particular random sample are taken from this urn. What we have here is nothing at all like an argument from population to sample or an argument from sample to population: it is a form of probable argument entirely different from both deduction and induction. This new type of argument Peirce called abduction (also, retroduction, and also, hypothesis).

Over many years Peirce modified his views on the three types of arguments, sometimes changing his views but mostly just extending them. He seemed to have some hesitation, for example, about whether arguments from analogy were inductions (on properties of things) or abductions. The main extension of his earliest views involved integrating the argument forms into his view of the scientific method. In fact, his most mature position seemed almost to equate the three types of argument with three phases of the scientific method.

Pragmatism, Pragmaticism, and the Scientific

Method

Probably Peirce's best-known works are the first two articles in a series of six that were collectively entitled *Illustrations of the Logic of Science* and published in *Popular Science Monthly* in 1877 and 1878. The first is entitled "The Fixation of Belief" and the second is entitled "How to Make Our Ideas Clear." In the first Peirce defended, in a manner consistent with idealism, the superiority of the scientific method over other methods of overcoming doubt by "fixing belief." In the second Peirce defended the pragmatic/pragmaticistic notion of clear concepts.

Perhaps the single most important fact to keep in mind in understanding Peirce's philosophy is that Peirce was a practicing physical scientist all his life, and that as he understood them, philosophy and logic were themselves also sciences. Moreover, he understood philosophy to be the philosophy of science and logic to be the logic of science (where "science" has its broadest sense, which is best captured by the German word *Wissenschaft*).

It is in this light that his specifications of the nature of pragmatism (which he later called "pragmaticism" in order to distinguish his own scientific philosophy from other conceptions and theories that were trafficked under the title "pragmatism") are to be understood. When he said that the whole meaning of a (clear) conception consists in the entire set of its practical consequences, he had in mind that a meaningful conception must have some experiential "cash value," capable of being specified as some sort of collection of possible empirical observations under specifiable conditions. Peirce insisted that the entire meaning of a meaningful conception consisted in the totality of such specifications of possible observations. For example, Peirce tended to spell out the meaning of dispositional properties such as "hard" or "heavy" by using the same sort of counterfactual constructions that, say, Hempel would use. Peirce was not a simple operationalist or verificationist, but his views were akin to operationalism and verificationism.

The previous point is related to the fact that Peirce was always a philosopher who had broad and deep affinities with idealism. Indeed, having rejected a great deal in Kant, Peirce nevertheless shared with Kant the view that the *Ding an sich* plays no role in philosophy or science other than that of a *Grenzbegriff*, or, perhaps a bit more accuratedly, a limiting concept. The notion can play no direct role in the sciences; science can deal only with phenomena, and all concepts must somehow be traceable back to phenomenological roots. Toward the end of his life Peirce began to regard himself as a thinker somewhat akin to Hegel, one of the avowed philosophical enemies of his youth. Peirce's brand of idealism is of the Kantian "transcendental" sort; and, even when Peirce called himself or is called by others a "realist," it must be kept in mind that Peirce was always a realist of the Kantian "empirical" sort. His realism is similar to what Hilary Putnam has called "internal realism." (Peirce was also a realist in another sense: anti-nominalist. More on this is given below.)

From his earliest to his latest writings Peirce attacked all forms of epistemological foundationalism and in particular all form of Cartesianism. Philosophy must begin wherever it happens to be at the moment, he thought, and not at some ideal foundation, for example in private references. The only important thing

in thinking scientifically is applying the scientific method itself. This method he held to be essentially public and reproducible in its activities, as well as self-correcting in following sense: No matter where different researchers may begin, as long as they follow the scientific method, their results will eventually converge toward the same result. (The pragmatic conception of meaning implies that two theories with exactly the same empirical content must have, despite superficial appearances, the same meaning.) This ideal point of convergence is what Peirce means by "the truth," and "reality" is simply what is meant by "the truth." That these notions of reality and truth are inherently idealist rather than realist in nature should need no special pleading.

Connected with Peirce's anti-foundationalism is his insistence on the fallibility of particular achievements in science. Although the scientific method will eventually converge to something, nevertheless at any temporal point in inquiry we are only at a provisional stage of it and cannot ascertain how far off we may be from the limit to which we are somehow converging. This insistence on the fallibilism of human inquiry is connected with several other important ideas of Peirce's, such as his tychism, his evolutionism, and his anti-determinism. (These will be discussed below.) Despite Peirce's insistence on fallibilism, he is not an epistemological pessimist: indeed, quite the opposite: he tends to hold that every genuine question (that is, every question whose possible answers have empirical content) can be answered in principle, or at least should not be assumed to be unanswerable. For this reason, one his most important dicta is "Do not block the path of inquiry!"

Peirce described the scientific method as consisting of abduction, deduction, and induction, plus the economics of research. His understanding of the scientific method is not far different from the standard idea of the scientific method (which perhaps derives historically from William Whewell and Peirce) as being the method of constructing hypotheses, deriving consequences from these hypotheses, and then experimentally testing these hypotheses. (The main Peircean factor left out is Peirce's notion of the economics of research.) Conversely, he increasingly came to understand the three types of inference as being the stages of the scientific method. For example, as Peirce came to extend and generalize his notion of abduction, abduction became defined as inference to and provisional acceptance of an explanatory hypothesis for the purposes of testing it. Abduction is not always inference to the best explanation, but it is always inference to some explanation or at least to something that clarifies or makes routine some information that has previously been "surprising," in the sense that we would not have routinely expected it, given our then-current state of knowledge. Deduction came to mean for Peirce the drawing of conclusions as to what observable phenomena should be expected if the hypothesis is correct. Induction came for him to mean the entire process of experimentation performed in service of hypothesis testing.

Peirce's idea of the economy (or: the economics) of research is an ineliminable part of his idea of the scientific method. He understood that science always operates in some given historical and socio-economic context, in which context certain problems are paramount and other problems trivial or frivolous. He understood that in such a context some experiments may be crucial and others insignificant. He understood that the economic resources of the scientist are severely limited, while the "great ocean of truth" that lies undiscovered before us is infinite. Research resources, such as personnel, time, and apparatus, are costly; and it is irrational to squander them. He proposed, therefore, that careful

consideration be paid to the problem of how to obtain the biggest epistemological "bang for the buck." In effect, the economics of research is akin to a cost/benefit analysis in connection with states of knowledge. Although this idea has been only little explored by Peirce scholars, Peirce himself regarded it as central to the scientific method and to the idea of rational behavior. It is connected with what he called "speculative rhetoric" or "methodeutic" (which will be discussed below).

Anti-determinism, Tychism, and Evolutionism

Against powerful currents of determinism that derived from the Enlightenment philosophy of the eighteenth century, Peirce urged that there was not the slightest scientific evidence for determinism and that there was considerable scientific evidence against it. Always by the words "science" and "scientific" Peirce understood reference to actual practice by scientists in the laboratory and the field, and not reference to entries in scientific textbooks. In attacking determinism, therefore, Peirce appealed to the evidence of the actual phenomena in laboratories and fields. Here, what is obtained as the actual observations (e.g. measurements) does not conform to some exact point or smooth function. For example, if we take, however carefully we may do so, a thousand measurements of some physical quantity, we will not obtain a thousand equal results, but rather only a distribution (usually a normal or Gaussian distribution of hundreds) of different results. Again, if we measure the value of an independent variable that we assume to depend on some given parameter, and if we let the parameter vary while we take successive measurements, the result will never be a smooth function (for example, a straight line or an ellipse); rather it will be a "jagged" result, to which we can at best *fit* a smooth function by using some clever method (for example, least-squares fitting). Moreover, the variation and inexactness of measurements become, Peirce maintained, the more pronounced and obtrusive the more refined and microscopic are our measurements. (Obviously, Peirce would not have been the least surprised by the results of measurements at the quantum level.)

What the facts of scientific practice tell us, then, is that, although the universe displays varying degrees of *habit*, that is to say of partial, varying, approximate, and statistical regularity, the universe does not display deterministic *law*, that is to say total, exact, non-statistical regularity. Moreover, the habits that nature displays appear in varying degrees of entrenchment: from the almost pure freedom and spontaneity of some processes of thought, at one end of the spectrum, to the nearly law-like behavior of large physical objects like planets, at the other end of the spectrum.

Science shows, then, that not everything is fixed by exact law (even if everything should be constrained to some extent by habit) and that spontaneity has an objective place in the universe. Peirce called this doctrine "tychism," a word taken from the Greek word for "chance" or "luck" or "what the gods choose to lay on one." Tychism is a fundamental part of Peirce's view, and reference to his tychism provides an added reason for Peirce's insisting on the irreducible fallibilism of inquiry. For nature is not a static world of law but rather a dynamic world that manifests considerable spontaneity. (Peirce would have regarded the irreducibility of quantum mechanics to some "hidden-variables" theory as being a mere matter of course.)

Three figures from the history of culture loomed exceedingly large in the intellectual atmosphere of the period in which Peirce was most active: Hegel in philosophy, Lyell in geology, and Darwin (along with Watson) in biology. These thinkers have a single theme in common: evolution. Hegel described an evolution of ideas, Lyell an evolution of geological structures, and Darwin an evolution of biological species and varieties. Peirce's thinking is deeply permeated with the evolutionary idea, which he extended beyond the confines of any particular subject matter. For Peirce, the entire universe is an evolutionary product; indeed, he conceived that even the most firmly entrenched of nature's habits (for example, even those habits typically called "natural laws") have themselves evolved, and accordingly should be subjects of inquiry. One can sensibly seek evolutionary explanations of the existence of particular natural laws.

One possible path along which nature acquires its habits was explored by Peirce using statistical analysis in situations of non-Bernoullian trials. Peirce showed that, if we posit a primal habit in nature, viz. the tendency however slight to take on habits, then the result is often a high degree of regularity in the long run. For this reason, Peirce suggested that in the remote past nature was considerably more spontaneous than it later became, and that in general the habits nature has come to exhibit have evolved, just like ideas, geological formations, and biological species have evolved.

In this evolutionary notion of nature and natural law we have an additional support of Peirce's insistence on the inherent fallibilism of scientific inquiry. Nature may simply change sometimes, even in its most entrenched fundamentals. Thus, even if scientists were at one point in time to have accurate conceptions about nature, this fact would not ensure that at some later point in time these same concpetions would remain accurate.

An especially intriguing and curious twist that Peirce's evolutionism takes is what is called its "agapeism." According to Peirce, the most fundamental engine of the evolutionary process is not struggle, strife, greed, or competition. Rather it is nurturing love, in which an entity is prepared to sacrifice its own perfection for the sake of the wellbeing of its neighbor. This doctrine had both a social significance for Peirce, who apparently had the intention of arguing against the popular socio-economic Darwinism of the late nineteenth century, and a cosmic significance, which Peirce associated with the doctrine of the Gospel of John and with the mystical ideas of Swedenborg and Henry James. Peirce even argued that logicality in some sense presupposes the ethics of self-sacrifice.

Synechism, the Continuum, Infinites, and Infinitesimals

Peirce was the first scientific thinker, or at least one of the first scientific thinkers, to argue in favor of the actual existence of infinite sets. His criterion of the difference between finite and infinite sets was that the so-called "syllogism of transposed quantity" introduced by de Morgan applied only to finite sets and not to infinite ones. The syllogism of transposed quantity runs as follows. We have a binary relation *R* defined on a set *S*, such that the relation satisfies the following two properties (where the quantifications

are taken over the set *S*). For all *x* there is a *y* such that *Rxy*. And for all *x*, *y*, *z*, *Rxz* and *Ryz* implies that $x = y$. The conclusion (of the syllogism of transposed quantity) is that for all *x* there exists a *y* such that *Ryx*. One of Peirce's favorite examples helps elucidate the idea, even if it perhaps be not perfectly politically correct: Every Texan kills some Texan; no Texan is killed by more than one Texan; therefore every Texan is killed by some Texan. The argument's conclusion follows only if the set of Texans is finite.

If for the relation *R* in question we take $f(x) = y$, where the function is defined on and has values in the set *S*, we can easily see that the syllogism of transposed quantity then says that no one-one function can map a set to a proper subset of itself. This assertion holds, of course, only for finite sets. So, as it turns out, Peirce's definition of the difference between finite and infinite sets is (pretty close to) equivalent to the standard one.

Peirce held that the continuity of space, time, ideation, feeling, and perception is an irreducible deliverance of science, and that an adequate conception of the continuous is an extremely important part of all the sciences. This doctrine he called "synechism," a word deriving from the Greek preposition that means "(together) with." In mid-1892, somewhat under the influence of reading Cantor's works, Peirce defined a (linear) continuum to be a linearly-ordered infinite set *C* such that (1) for any two distinct members of *C* there exists a third member of *C* that is strictly between these; and (2) every countably infinite subset of *C* that has an upper (lower) bound in *C* has a least upper bound (greatest lower bound) in *C*. The first property he called "Kanticity" and the second "Aristotelicity." (Today we would likely call these properties "density" and "closedness," respectively.) The second condition has the corollary that a continuum contains all its limit points, and sometimes Peirce used this property in conjunction with "Kanticity" to define a continuum. Toward the end of the nineteenth century Peirce remarked that he had framed an updated conception of continua by loosening his attachment to Cantor's ideas, but what this new approach is has not yet been explored by Peirce scholars.

Not only did Peirce defend infinite magnitudes, but also he defended infinitesimal magnitudes. Moreover, he argued for the consistency of introducing infinitesimals into the number system, and he wanted to use infinitesimals to justify the traditional pre-Gaussian definitions and underpinnings of the differential calculus. He also made a number of remarks that suggest that, in connection with the foregoing enterprise, he had a novel conception of the topology of the real numbers. All these remarks he connected with his notion of the continuum and his previous defenses of infinite sets. For these reasons some Peirce scholars have suggested that his ideas were an anticipation of Abraham Robinson's non-standard analysis. Whether this be so or not is, however, at the present time far from clear: so far no commentator has provided anything close to being a careful and detailed exposition of this point, and most of Peirce's published writing on this topic is extremely obscure. The entire analysis of Peirce's notion of an infinitesimal, as well as the exact bearing this notion has on his concept of continuity and on his idea of the topology of the real numbers, still awaits meticulous mathematical discussion.

Theory of Probability

In light of Peirce's tychism and his view that statistical information is often the most exact information we can have about phenomena, it should not be surprising that Peirce devoted close attention to probability theory and statistical analysis. Indeed, Peirce not only extensively used the concept of probability but also offered a pragmatistic account of the notion of probability itself.

Peirce vigorously attacked the view of de Morgan that probability was a measure of our confidence or degree of belief: a view known today as the subjectivist theory of probability. Along with this attack, Peirce ridiculed various Bayesian-type analyses of the problem of induction (for example, the work of Quételet), on the grounds that the relevant Bayesian "prior probabilities" cannot be assigned unless one first assumes a subjectivist view and thus equates complete lack of information about something with that something's having a probability of 1/2, which equation is an egregious error in Peirce's estimation.

Rather than holding probability to be a measure of degree of confidence or belief, Peirce adopted an objectivist notion of probability that he likened to the doctrine of John Venn. Indeed, he held that probability is actually a notion with clear empirical content and clear empirical procedures for ascertaining that content, as follows. First, what is assigned a probability, insofar as the notion is used scientifically, is neither a proposition nor an event nor a type of event. Rather, what is assigned a probability is an argument, with premisses (Peirce insisted on this spelling rather than the spelling "premises") and a conclusion. Second, in order to ascertain the probability of a particular argument, the observer notes all occasions on which all of its premisses are true, case by case, just as they come under observation. For each of these occasions the observer notes whether the conclusion is true or not. The observer keeps an ongoing ratio whose numerator is the number of occasions so far observed on which the conclusion as well as the premisses are true and whose denominator is the number of occasions so far observed on which simply the premisses are true (irrespective of whether the conclusion is also true). At each observation the observer computes this ratio, which then encompasses all the observer's past observations of occasions on which the premisses are true. The probability of the argument in question is defined by Peirce to be the limit of the crucial ratio as the number of observations tends to grow infinitely large (if this limit exists).

It might be thought that, when Peirce adopted the view of objective spontaneity and more or less entrenched objective habit in the universe (tychism), he perforce gave up the foregoing account of probability. Such a thought, however, would be a mistake; it rests failing to realize that objective attributions of probability for Peirce are consequent upon rather than inconsistent with his commitment to the pragmatistic account of probability that was given above.

Psycho-physical Monism and Anti-nominalism

Peirce held that science suggests that the universe has evolved from a condition of maximum freedom and spontaneity into its present condition, in which it has taken on a number of more and less entrenched habits. With pure freedom and spontaneity Peirce tended to associate mind, and with entrenched habits he tended to associate matter (or, more generally, the physical). Thus he tended to see the universe as the end-product-so-far of a process in which mind has acquired habits and has "congealed" (this is the very

word Peirce used) into matter.

This notion of all things as being evolved psycho-physical unities of some sort places Peirce well within the sphere of what might be called "the grand old-fashioned metaphysicians," along with such thinkers as Plato, Aristotle, Aquinas, Spinoza, Leibnitz, Hegel, Schopenhauer, Whitehead, *et al.* Some contemporary philosophers might be inclined to reject Peirce out of hand upon discovering this fact. Others might find his notion of psycho-physical unities not so offputting or indeed even attractive. What is crucial is that Peirce argued that mind pervades all of nature in varying degrees, and is not found merely in its most advanced animal species.

This pan-psychistic view, combined with synechism, meant for Peirce that mind is extended in some sort of continuum throughout the universe. Peirce tended to think of ideas as existing in mind in somewhat the same way as physical forms exist in physically extended things, and he even spoke of ideas as "spreading" out through the same continuum in which mind is extended. This set of conceptions is part of what Peirce regarded as (his version of) Scotistic realism, which he opposed to nominalism. He tended to blame what he regarded as the errors of much of the philosophy of his contemporaries as owing to the nominalistic disregard for the objective existence of form.

Triadism and the Universal Categories

Merely to say that Peirce was extremely fond of placing things into groups of three, of trichotomies, and of triadic relations, would fail miserably to do justice to the overwhelming obtrusiveness in his philosophy of the number three. Indeed, he made the most fundamental categories of all "things" of any sort whatsoever the categories of "Firstness," "Secondness," and "Thirdness," and he often described "things" as being "firsts" or "seconds" or "thirds." For example, with regard to the trichotomy "possibility," "actuality," and "necessity," possibility he called a first, actuality he called a second, and necessity he called a third. Again: quality was a first, fact was a second, and habit (or rule or law) was a third. Again: entity was a first, relation was a second, and representation was a third. Again: rheme (by which Peirce meant a relation of arbitrary adicity or arity) was a first, proposition was a second, and argument was a third. The list goes on and on. Let us refer to Peirce's penchant for describing things in terms of trichotomies and triadic realtions as Peirce's "triadism."

If Peirce had a general rationale for his triadism, Peirce scholars have not yet made it clear what this rationale might be. He seemed to base his triadism on what he called "phaneroscopy," by which word he meant the mere observation of phenomenal appearances. He regularly commented that the phenomena just *do* fall into three groups and that they just *do* display irreducibly triadic relations.

Although there are many examples of phenomena that do seem more or less naturally to divide into three groups, Peirce seems to have been driven by something more than mere examples in his insistence on applying his categories to almost everything imaginable. Perhaps it was the influence of Kant, whose twelve categories divide into four groups of three each. Perhaps it was the triadic structure of the stages of thought as described by Hegel, or even the triune commitments of orthodox Christianity (to which

Peirce seemed to subscribe). Certainly involved was Peirce's commitment to the ineliminability of mind in nature, for Peirce closely associated the activities of mind with a particular triadic relation that he called the "sign" relation. (More on this topic appears below.) Also involved was Peirce's so-called "reduction thesis" in logic (on which more will given below), to which Peirce had concluded as early as 1870.

It is difficult to imagine even the most fervently devout of the passionate admirers of Peirce, of which there are many, saying that his account (or, more accurately, his various accounts) of the three universal categories is (or are) clear and compelling. Yet, in almost everything Peirce wrote from the time the categories were first introduced, they found a place. Their analysis and an account of their general rationale, if there be such, constitute chief problems in Peirce scholarship.

Mind and Semeiotic

Connected with Peirce's insistence on the ubiquity of mind in the cosmos is the importance he attached to what he called "semeiotic," the theory of signs in the most general sense. Peircean semeiotic is almost totally different what has come to be called "semiotics," and which hales not so much from Peirce as from Saussure and Charles W. Morris. Peircean semeiotic derives ultimately from the theory of signs of Duns Scotus and its later development by John of St. Thomas (John Poirot). In Peirce's theory the sign relation is a triadic relation, a special species of the genus: the representing relation. Whenever the representing relation has an instance, we find one thing (the "object") being represented by another thing (the "representamen") to (or: in) a third thing (the "interpretant"), and represented in such a way that the interpretant is thereby determined to be also a representamen of the object to yet another interpretant in the representation relation. Obviously, Peirce's definition entails that we have an infinite sequence of representamens of an object whenever we have any one representamen.

The sign relation is the representing relation whenever the first interpretant (and consequently each member of the whole infinite sequence of interpretants) is a cognition of a mind. In any instance of the sign relation an object is signified by a sign to a mind. One of Peirce's central tasks was that of analyzing all possible kinds of signs.

Semeiotic and Logic

Peirce's settled opinion was that logic in the broadest sense is to be equated with semeiotic, and that logic in a much narrower sense (which he typically called "logical critic") is one of three major divisions or parts of semeiotic. Thus, in his later writings, he divided semeiotic into speculative grammar, logical critic, and speculative rhetoric (also called "methodeutic"). Peirce's word "speculative" is his Latinate version of the Greek-derived word "theoretical," and should be understood to mean exactly the word "theoretical." Peirce's tripartite division of semeiotic is not to be confused with Charles W. Morris's division: syntax, semantics, and pragmatics (although there may be some commonalities in the two trichotomies).

By speculative grammar Peirce understood the analysis of the kinds of signs there are and the ways that they can be combined significantly. For example, under this heading he introduced three trichotomies of signs and argued for the real possibility of only certain kinds of signs. Signs are qualisigns, sinsigns, or legisigns, accordingly as they are mere qualities, individual events and states, or habits (or laws), respectively. Signs are icons, indices (also called "semes"), or symbols (sometimes called "tokens"), accordingly as they derive their significance from resemblance to their objects, a real relation (for example, of causation) with their objects, or are connected only by convention to their objects, respectively. Signs are rhematic signs (also called "sumisigns" and "rhemes"), dicisigns (also called "quasi-propositions"), or arguments (also called "suadisigns"), accordingly as they are predicational/relational in character, propositional in character, or argumentative in character. Because the three trichotomies are orthogonal to each other, together they yield the abstract possibility that there are 27 distinct kinds of signs. Peirce argued, however, that 17 of these are logically impossible, so that finally only 10 kinds of signs are genuinely possible. In terms of these 10 kinds of signs, Peirce endeavored to construct a theory of all possible natural and conventional signs, whether simple or complex.

What Peirce meant by "logical critic" is pretty much logic in the ordinary, accepted sense of "logic" from Aristotle's logic to present-day mathematical logic. As might be expected, a crucial concern of logical critic is to characterize the difference between correct and incorrect reasoning. Some of Peirce's accomplishments in this area will be discussed below.

By "speculative rhetoric" or "methodeutic" Peirce understood all inquiry into the principles of the effective use of signs for producing valuable courses of research and giving valuable expositions. Methodeutic studies the methods that researchers should use in investigating, giving expositions of, and creating applications of the truth. This idea may overlap to some small extent Morris's notion of "pragmatics," but the spirit of Peirce's notion is much wider than that of Morris's. Moreover, Peirce handled the notion of indexical reference under the heading of speculative grammar and not speculative rhetoric, whereas the topic certainly belongs to Morris's pragmatics. So far as is known, Peirce did not develop details of the topic of speculative rhetoric to any great extent, but it is clear that the important topic of the economy of research is closely affiliated with it.

The Classification of the Sciences

Peirce maintained an interest in the topic of classification or taxonomy in general, and he considered biology and geology the foremost sciences to have made progress in developing genuinely useful classifications of things. In his own theory of classification, he seemed to regard some sort of cluster analysis as holding the key to creating really useful classifications. He regularly strove to create a classification of all the sciences that would be as useful to logic as the systems of the biologists and geologists were to these scientists.

As with many of Peirce's divisions, his classification of the sciences is a taxonomy consisting of triads.

For example he classifies all the sciences into those of discovery, review, and practicality. Sciences of discovery he divides into mathematics, philosophy, and what he calls "idioscopy" (by which he seems to mean the class of all the particular or special sciences like physics, psychology, and so forth). Mathematics he divides into mathematics of logic, of discrete series, and of continua and pseudo-continua. Philosphy divides into phenomenology, normative science, and metaphysics. Normative science divides into aesthetics, ethics, and logic. And so on and on. Occasionally there is found a division into two: for example, he divides idioscopy into the physical sciences and the psychical (or human) sciences. But mostly the division is into threes.

Peirce scholars have found the topic of Peirce's classification of the sciences a fertile ground for assertions about what is most basic in all thinking, in Peirce's view. Whether or not such assertions run afoul of Peirce's anti-foundationalism is itself a topic for further study.

Logic

In the extensiveness and originality of his contributions to mathematical logic, Peirce is almost without equal. His writings and original ideas are so numerous that there is no way to do them justice in a small article such as the present one. Accordingly, only a few of his achievements will be mentioned here.

Peirce's special strength lay not so much in theorem-proving as rather in the invention and developmental elaboration of novel systems of logical syntax and fundamental logical concepts. He invented dozens of different systems of logical syntax, including a syntax based on a generalization of de Morgan's relative product operation, an algebraic syntax that mirrored Boolean algebra to some extent, a quantifier-and-variable syntax that (except for specific symbols) is identical to the much later Russell-Whitehead syntax, and even two systems of two-dimensional syntax: the entitative graphs and the existential graphs, the latter being a syntax for logic using the mathematical apparatus of topological graph theory.

In 1870 Peirce published a long paper "Description of a Notation for the Logic of Relatives" in which he introduced for the first time in history, two years before Frege's *Begriffschrift* a complete syntax for the logic of relations of arbitrary adicity (or arity). In this paper the notion of the variable (though not under the name "variable") was invented, and Peirce provided devices for negating, combining relations, and quantifying. By 1883, along with his student O. H. Mitchell, he had developed a full syntax for quantificational logic, only a little different in specific symbols (as was mentioned just above) from the standard Russell-Whitehed syntax that did not appear until 1910.

Peirce introduced the material-conditional operator into logic, developed the Shaffer stroke and daggar operators 40 years before Shaffer, and developed a full logical system based only on the stroke function. As Garret Birkhoff notes in his *Lattice Theory* it was in fact Peirce who invented the concept of a lattice (around 1883). (Quite possibly, it is Peirce's lattice theory that holds the key to his technical theory of infinitesimals and the continuum.)

During his years teaching at Johns Hopkins University, Peirce began to research the four-color map conjecture and developed extensive connections between logic and topology, especially topological graph theory. Ultimately these researches bore fruit in his existential graphs, but probably his writings in this area contain a considerable number of other valuable ideas and results. He hinted that he had made great progress in the theory of provability and unprovability by exploring the connections between logic and topology.

Peirce's Reduction Thesis

Peirce's so-called "Reduction Thesis" is the thesis that all relations of arbitrary adicity may be constructed from triadic relations alone, whereas monadic and dyadic relations alone are not sufficient to allow the construction of even a single "non-degenerate" (that is: non-Cartesian-factorable) triadic relation. Although the germ of his argument for the Reduction Thesis lay in his 1870 paper "Description of a Notation for the Logic of Relatives," the Thesis was for over a century doubted by many, especially after the publication of a proof by Willard Van Orman Quine that all relations could be constructed exclusively from dyadic ones. As it turns out, both Peirce and Quine were correct: the issue all depends on exactly what constructive resources are to be allowed to be used in building relations out of other relations. (Obviously, the more extensive and powerful are the constructive resources, the more likely it is that all relations can be constructed from dyadic ones alone by using them.) An exact exposition and proof of Peirce's Reduction Thesis was finally accomplished in 1988, and it makes clear that Peirce's constructive resources are to be understood to include only negation, a generalization of de Morgan's relative product operation, and the use of a particular triadic relation that Peirce called "the teridentity relation" and that we might today write as $x = y = z$.

Peirce felt that the teridentity relation was in some way much more fundamental than the usual dyadic identity relation $x = y$. He also felt that relative product was a much more fundamental operation than, say, Boolean product or Boolean sum. The full philosophical import of his Reduction Thesis, and the philosophical importance of his triadism insofar as this triadism rests on his Reduction Thesis, cannot be ascertained without a prior understanding of his theory of identity and his view of the nature of the relative product operation.

Contemporary Practical Applications of Peirce's Ideas

Currently, considerable interest is being taken in Peirce's ideas from outside the arena of academic philosophy. The interest comes from industry, business, technology, and the military; and it has resulted in the existence of a number of agencies, institutes, and laboratories in which ongoing research into and development of Peircean concepts is being undertaken.

This interest has arisen, apparently, in two ways. First, some two decades agao in the former Soviet

Union interest in Peirce and Karl Popper led logicians and computer scientists like Victor Finn and Dmitri Pospelov to try to find ways in which computer programs could generate Peircean hypotheses (Popperian "conjectures") in semeiotic contexts (non-numerical contexts). Under the guide in particular of Finn's intelligent systems laboratory in VINITI-RAN (the All-Russian Institute of Scientific and Technical Information of the Russian Academy of Sciences), elaborate techniques for automatic generation of hypotheses have been found and extensively utilized for many practical purposes. Among these are sociological prediction, pharmacological discovery, and the analysis of processes of industrial production. Interest in Finn's work, and through it the practical application of Peirce's philosophy, has spread to France, Germany, Denmark, and ultimately the United States.

Second, as the limits of expert systems in artificial intelligence contexts have become increasingly clear to computer scientists, they have begun to search for methods beyond those of the production rules of expert systems. One promising line of research has been in automating Peirce's concept of the scientific method, complete with techniques for hypothesis-generation and making assessments of the costs and benefits of exploring hypotheses. In some areas of research added impetus has been provided by the similarity of Peircean techniques to techniques that have already proven useful. For example, in the field of automated multi-track radar, the similarity of Peircean scientific method to the so-called "Kalman filter" has been noted by many systems analysts. Again, those interested in military command-and-control often note the similarity of Peircean scientific method to the classic OODA loop ("observe, orient, decide, act") of comand-and-control-theory. The areospace industry, especially in France and the United States, is currently investigating Peircean ideas in connection with avionics systems that monitor aircraft "health."

Such practical applications of Peircean ideas may seem surprising to many philosophers, but they surely would not have surprised Peirce. Indeed, given his lifelong goals as a scientist-philosopher, he probably would have found the current situation entirely in accord with his expectations.

Bibliography

Primary Sources

- *Collected Papers of Charles Sanders Peirce*, 8 vols. Edited by Charles Hartshorne, Paul Weiss, and Arthur Burks (Harvard University Press, Cambridge, Massachusetts, 1931-1958).
- *The Essential Peirce*, 2 vols. Edited by Nathan Houser, Christian Kloesel, and the Peirce Edition Project (Indiana University Press, Bloomington, Indiana, 1992, 1998).
- *The New Elements of Mathematics by Charles S. Peirce*, Volume I Arithmetic, Volume II Algebra and Geometry, Volume III/1 and III/2 Mathematical Miscellanea, Volume IV Mathematical Philosophy. Edited by Carolyn Eisele (Mouton Publishers, The Hague, 1976).
- *Reasoning and the Logic of Things: the Cambridge Conferences Lectures of 1898*. Edited by Kenneth Laine Ketner (Harvard University Press, Cambridge, Massachusetts, 1992).
- *Writings of Charles S. Peirce: a Chronological Edition*, Volume I 1857-1866, Volume II 1867-

1871, Volume III 1872-1878, Volume IV 1879-1884, Volume V 1884-1886. Edited by the Peirce Edition Project (Indiana University Press, Bloomington, Indiana, 1982, 1984, 1986, 1989, 1993).

Secondary Sources

- *Studies in the Scientific and Mathematical Philosophy of Charles S. Peirce: Essays by Carolyn Eisele*. Edited by Richard M. Martin (Mouton, The Hague, 1979).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle: logic | Hegel, Georg Wilhelm Friedrich | James, William | Kant, Immanuel | logic: classical | panpsychism | Peirce, Benjamin | Popper, Karl | Russell, Bertrand | universals: the medieval problem of | Whewell, William | Whitehead, Alfred North

Robert W. Burch
rwb@beth-chayyim.tamu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 22, 2001
Content last modified: June 22, 2001

Aristotle's Logic

Aristotle's logic, especially his theory of the syllogism, has had an unparalleled influence on the history of Western thought. It did not always hold this position: in the Hellenistic period, Stoic logic, and in particular the work of Chrysippus, was much more celebrated. However, in later antiquity, following the work of Aristotelian Commentators, Aristotle's logic became dominant, and Aristotelian logic was what was transmitted to the Arabic and the Latin medieval traditions, while the works of Chrysippus have not survived.

This unique historical position has not always contributed to the understanding of Aristotle's logical works. Kant thought that Aristotle had discovered everything there was to know about logic, and the historian of logic Prantl drew the corollary that any logician after Aristotle who said anything new was confused, stupid, or perverse. During the rise of modern formal logic following Frege and Peirce, adherents of Traditional Logic (seen as the descendant of Aristotelian Logic) and the new mathematical logic tended to see one another as rivals, with incompatible notions of logic. More recent scholarship has often applied the very techniques of mathematical logic to Aristotle's theories, revealing (in the opinion of many) a number of similarities of approach and interest between Aristotle and modern logicians.

This article is written from the latter perspective. As such, it is about Aristotle's logic, which is not always the same thing as what has been called "Aristotelian" logic.

---

§1: Introduction

Aristotle's logical works contain the earliest formal study of logic that we have. It is therefore all the more remarkable that together they comprise a highly developed logical theory, one that was able to command immense respect for many centuries: Kant, who was ten times more distant from Aristotle than we are from him, even held that nothing significant had been added to Aristotle's views in the intervening two millennia.

In the last century, Aristotle's reputation as a logician has undergone two remarkable reversals. The rise of modern formal logic following the work of Frege and Russell brought with it a recognition of the many serious limitations of Aristotle's logic; today, very few would try to maintain that it is adequate as a basis for understanding science, mathematics, or even everyday reasoning. At the same time, scholars trained in modern formal techniques have come to view Aristotle with new respect, not so much for the correctness of his results as for the remarkable similarity in spirit between much of his work and modern logic. As Jonathan Lear has put it, "Aristotle shares with modern logicians a fundamental interest in metatheory": his primary goal is not to offer a practical guide to argumentation but to study the properties of inferential systems themselves.

§2: Aristotle's Logical Works: The *Organon*

The ancient commentators grouped together several of Aristotle's treatises under the title *Organon* ("Instrument") and regarded them as comprising his logical works:

1. *Categories*
2. *On Interpretation*
3. *Prior Analytics*
4. *Posterior Analytics*
5. *Topics*
6. *On Sophistical Refutations*

In fact, the title *Organon* reflects a much later controversy about whether logic is a part of philosophy (as the Stoics maintained) or merely a tool used by philosophy (as the later Peripatetics thought); calling the logical works "The Instrument" is a way of taking sides on this point. Aristotle himself never uses this term, nor does he give much indication that these particular treatises form some kind of group, though there are frequent cross-references between the *Topics* and the *Analytics*. On the other hand, Aristotle

treats the *Prior* and *Posterior Analytics* as one work, and *On Sophistical Refutations* is a final section, or an appendix, to the *Topics*). To these works should be added the *Rhetoric*, which explicitly declares its reliance on the *Topics*.

§3: The Subject of Logic: "Syllogisms"

All Aristotle's logic revolves around one notion: the **deduction** (*sullogismos*). A thorough explanation of what a deduction is, and what they are composed of, will necessarily lead us through the whole of his theory. What, then, is a deduction? Aristotle says:

> A deduction is speech (*logos*) in which, certain things having been supposed, something different from those supposed results of necessity because of their being so. (*Prior Analytics* I.2, 24b18-20)

Each of the "things supposed" is a **premise** (*protasis*) of the argument, and what "results of necessity" is the **conclusion** (*sumperasma*).

The core of this definition is the notion of "resulting of necessity" (*ex anankês sumbainein*). This corresponds to a modern notion of logical consequence: X results of necessity from Y and Z if it would be impossible for X to be false when Y and Z are true. We could therefore take this to be a general definition of "valid argument".

A. Induction and Deduction

Deductions are one of two species of argument recognized by Aristotle. The other species is **induction** (*epagôgê*). He has far less to say about this than deduction, doing little more than characterize it as "argument from the particular to the universal". However, induction (or something very much like it) plays a crucial role in the theory of scientific knowledge in the *Posterior Analytics*: it is induction, or at any rate a cognitive process that moves from particulars to their generalizations, that is the basis of knowledge of the indemonstrable first principles of sciences.

B. Aristotelian Deductions and Modern Valid Arguments

Despite its wide generality, Aristotle's definition of deduction is not a precise match for a modern definition of validity. Some of the differences may have important consequences:

1. Aristotle explicitly says that what results of necessity must be different from what is supposed. This would rule out arguments in which the conclusion is identical to one of the premises. Modern notions of validity regard such arguments as valid, though trivially so.
2. The plural "certain things having been supposed" was taken by some ancient commentators to rule out arguments with only one premise.

3. The force of the qualification "because of their being so" has sometimes been seen as ruling out arguments in which the conclusion is not ‘relevant’ to the premises, e.g., arguments in which the premises are inconsistent, arguments with conclusions that would follow from any premises whatsoever, or arguments with superfluous premises.

Of these three possible restrictions, the most interesting would be the third. This could be (and has been) interpreted as committing Aristotle to something like a relevance logic. In fact, there are passages that appear to confirm this. However, this is too complex a matter to discuss here.

However the definition is interpreted, it is clear that Aristotle does not mean to restrict it only to a subset of the valid arguments. This is why I have translated *sullogismos* with ‘deduction’ rather than its English cognate. In modern usage, ‘syllogism’ means an argument of a very specific form. Moreover, modern usage distinguishes between valid syllogisms (the conclusions of which follow from their premises) and invalid syllogisms (the conclusions of which do not follow from their premises). The second of these is inconsistent with Aristotle's use: since he defines a *sullogismos* as an argument in which the conclusion results of necessity from the premises, "invalid *sullogismos*" is a contradiction in terms. The first is also at least highly misleading, since Aristotle does not appear to think that the *sullogismoi* are simply an interesting subset of the valid arguments. Moreover (see below), Aristotle expends great efforts to argue that every valid argument, in a broad sense, can be "reduced" to an argument, or series of arguments, in something like one of the forms traditionally called a syllogism. If we translate *sullogismos* as "syllogism, ", this becomes the trivial claim "Every syllogism is a syllogism",

§4: Premises: The Structures of Assertions

Syllogisms are structures of sentences each of which can meaningfully be called true or false: **assertions** (*apophanseis*), in Aristotle's terminology. According to Aristotle, every such sentence must have the same structure: it must contain a **subject** (*hupokeimenon*) and a **predicate** and must either affirm or deny the predicate of the subject. Thus, every assertion is either the **affirmation** *kataphasis* or the **denial** (*apophasis*) of a single predicate of a single subject.

In *On Interpretation*, Aristotle argues that a single assertion must always either affirm or deny a single predicate of a single subject. Thus, he does not recognize sentential compounds, such as conjunctions and disjunctions, as single assertions. This appears to be a deliberate choice on his part: he argues, for instance, that a conjunction is simply a collection of assertions, with no more intrinsic unity than the sequence of sentences in a lengthy account (e.g. the entire *Iliad*, to take Aristotle's own example). Since he also treats denials as one of the two basic species of assertion, he does not view negations as sentential compounds. His treatment of conditional sentences and disjunctions is more difficult to appraise, but it is at any rate clear that Aristotle made no efforts to develop a sentential logic. Some of the consequences of this for his theory of demonstration are important.

A. Terms

Subjects and predicates of assertions are **terms**. A term (*horos*) can be either individual, e.g. *Socrates*, *Plato* or universal, e.g. *human*, *horse*, *animal*, *white*. Subjects may be either individual or universal, but predicates can only be universals: *Socrates is human*, *Plato is not a horse*, *horses are animals*, *humans are not horses*.

The word **universal** (*katholou*)appears to be an Aristotelian coinage. Literally, it means "of a whole"; its opposite is therefore "of a particular" (*kath' hekaston*). Universal terms are those which can properly serve as predicates, while particular terms are those which cannot.

This distinction is not simply a matter of grammatical function. We can readily enough construct a sentence with "Socrates" as its grammatical predicate: "The person sitting down is Socrates". Aristotle, however, does not consider this a genuine predication. He calls it instead a merely **accidental** or **incidental** (*kata sumbebêkos*) predication. Such sentences are, for him, dependent for their truth values on other genuine predications (in this case, "Socrates is sitting down").

Consequently, predication for Aristotle is as much a matter of metaphysics as a matter of grammar. The reason that the term *Socrates* is an individual term and not a universal is that the entity which it designates is an individual, not a universal. What makes *white* and *human* universal terms is that they designate universals.

Further discussion of these issues can be found in the entry on Aristotle's metaphysics.

B. Affirmations, Denials, and Contradictions

Aristotle takes some pains in *On Interpretation* to argue that to every affirmation there corresponds exactly one denial such that that denial denies exactly what that affirmation affirms. The pair consisting of an affirmation and its corresponding denial is a **contradiction** (*antiphasis*). In general, Aristotle holds, exactly one member of any contradiction is true and one false: they cannot both be true, and they cannot both be false. However, he appears to make an exception for propositions about future events, though interpreters have debated extensively what this exception might be (see further discussion below). The principle that contradictories cannot both be true has fundamental importance in Aristotle's metaphysics (see further discussion below).

C. All, Some, and None

One major difference between Aristotle's understanding of predication and modern (i.e. post-Fregean) logic is that Aristotle treats individual predications and general predications as similar in logical form: he gives the same analysis to "Socrates is an animal" and "Humans are animals". However, he notes that when the subject is a universal, predication takes on two forms: it can be either **universal** or **particular**. These expressions are parallel to those with which Aristotle distinguishes universal and particular terms, and Aristotle is aware of that, explicitly distinguishing between a term being a universal and a term being

universally predicated of another.

Whatever is affirmed or denied of a universal subject may be affirmed or denied of it it **universally** (*katholou* or "of all", *kata pantos*), **in part** (*kata meros*, *en merei*). or **indefinitely** (*adihoristos*).

	Affirmations		**Denials**	
Universal	P affirmed of all of S	Every S is P, All S is (are) P	P denied of all of S	No S is P
Particular	P affirmed of some of S	Some S is (are) P	P denied of some of S	Some S is not P, Not every S is P
Indefinite	P affirmed of S	S is P	P denied of S	S is not P

The "Square of Opposition"

In *On Interpretation,* Aristotle spells out the relationships of contradiction for sentences with universal subjects as follows:

	Affirmation	**Denial**
Universal	Every A is B	No A is B
Universal	Some A is B	Not every A is B

Simple as it appears, this table raises important difficulties of interpretation (for a thorough discussion, see the entry on the Square of Opposition).

In the *Prior Analytics*, Aristotle adopts a somewhat artificial way of expressing predications: instead of saying "X is predicated of Y" he says "Y belongs (*huparchei*) to X". This should really be regarded as a technical expression. The verb *huparchein* usually means either "begin" or "exist, be present", and Aristotle's usage appears to be a development of this latter use.

Some Convenient Abbreviations

For clarity and brevity, I will use the following semi-traditional abbreviations for Aristotelian categorical sentences (note that the predicate term comes *first* and the subject term *second*):

Abbreviation	**Sentence**
Aab	a belongs to all b (Every b is a)
Eab	a belongs to no b (No b is a)
Iab	a belongs to some b (Some b is a)
Oab	a does not belong to all b (Some b is not a)

§5: The Syllogistic

Aristotle's most famous achievement as logician is his theory of inference, traditionally called the **syllogistic** (though not by Aristotle). That theory is in fact the theory of inferences of a very specific sort: inferences with two premises, each of which is a categorical sentence, having exactly one term in common, and having as conclusion a categorical sentence the terms of which are just those two terms not shared by the premises. Aristotle calls the term shared by the premises the **middle term** (*meson*) and each of the other two terms in the premises an **extreme** (*akron*). The middle term must be either subject or predicate of each premise, and this can occur in three ways: the middle term can be the subject of one premise and the predicate of the other, the predicate of both premises, or the subject of both premises. Aristotle refers to these term arrangements as **figures** (*schêmata*):

A. The Figures

	First Figure		**Second Figure**		**Third Figure**	
	Predicate	**Subject**	**Predicate**	**Subject**	**Predicate**	**Subject**
Premise	a	b	a	b	a	c
Premise	b	c	a	c	b	c
Conclusion	a	c	b	c	a	b

Aristotle calls the term which is the predicate of the conclusion the **major** term and the term which is the subject of the conclusion the **minor** term. The premise containing the major term is the **major premise**, and the premise containing the minor term is the **minor premise**.

Aristotle's procedure is then a systematic investigation of the possible combinations of premises in each of the three figures. For each combination, he seeks either to demonstrate that some conclusion necessarily follows or to demonstrate that no conclusion follows. The results he states are exactly correct.

B. Methods of Proof: Conversion and Reduction

Aristotle shows each valid form to be valid by showing how to construct a deduction of its conclusion from its premises. These deductions, in turn, can take one of two forms: **direct** or **probative** (*deiktikos*) deductions and deductions **through the impossible** (*dia to adunaton*).

A direct deduction is a series of steps leading from the premises to the conclusion, each of which is either a **conversion** of a previous step or an inference from two previous steps relying on a first-figure deduction. Conversion, in turn, is inferring from a proposition another which has the subject and predicate interchanged. Specifically, Aristotle argues that three such conversions are sound:

- Eab → Eba
- Iab → Iba
- Aab → Iba

He undertakes to justify these in *An. Pr.* I.2. From a modern standpoint, the third is sometimes regarded with suspicion. Using it we can get *Some monsters are chimeras* from the apparently true *All chimeras are monsters*; but the former is often construed as implying in turn *There is something which is a monster and a chimera*, and thus that there are monsters and there are chimeras. In fact, this simply points up something about Aristotle's system: Aristotle in effect supposes that *all* terms in syllogisms are non-empty. (For further discussion of this point, see the entry on the Square of Opposition).

As an example of the procedure, we may take Aristotle's proof of *Camestres*. He says:

> If M belongs to every N but to no X, then neither will N belong to any X. For if M belongs to no X, then neither does X belong to any M; but M belonged to every N; therefore, N will belong to no M (for the first figure has come about). And since the privative converts, neither will N belong to any X. (*An. Pr.* I.5, 27a9-12)

From this text, we can extract an exact formal proof, as follows:

Step	**Justification**	**Aristotle's Text**
1. MaN		*If M belongs to every N*
2. MeX		*but to no X,*
To prove: NeX		*then neither will N belong to any X.*
3. MeX	(2, premise)	*For if M belongs to no X,*
4. XeM	(3, conversion of *e*)	*then neither does X belong to any M;*
5. MaN	(1, premise)	*but M belonged to every N;*
6. XeN	(4, 5, *Celarent*)	*therefore, X will belong to no N (for the first figure has come about).*
7. NeX	(6, conversion of *e*)	*And since the privative converts, neither will N belong to any X.*

C. Methods of Disproof: Counterexamples and Terms

Aristotle proves invalidity by constructing counterexamples. This is very much in the spirit of modern logical theory: all that it takes to show that a certain *form* is invalid is a single *instance* of that form with true premises and a false conclusion. However, Aristotle states his results not by saying that certain premise-conclusion combinations are invalid but by saying that certain premise pairs do not "syllogize":

that is, that, given the pair in question, examples can be constructed in which premises of that form are true and a conclusion of any of the four possible forms is false.

When possible, he does this by a clever and economical method: he gives two triplets of terms, one of which makes the premises true and a universal affirmative "conclusion" true, and the other of which makes the premises true and a universal negative "conclusion" true. The first is a counterexample for an argument with either an E or an O conclusion, and the second is a counterexample for an argument with either an A or an I conclusion.

D. The Deductions in the Figures ("Moods")

In *Prior Analytics* I.4-6, Aristotle shows that the premise combinations given in the following table yield deductions and that all other premise combinations fail to yield a deduction. in the terminology traditional since the middle ages, each of these combinations is known as a **mood** (from Latin *modus*, "way", which in turn is a translation of Greek *tropos*). Aristotle, however, does not use this expression and instead refers to "the arguments in the figures".

In this table, "⊢" separates premises from conclusion; it may be read "therefore". The second column lists the medieval mnemonic name associated with the inference (these are still widely used, and each is actually a mnemonic for Aristotle's proof of the mood in question). The third column briefly summarizes Aristotle's procedure for demonstrating the deduction.

Table of the Deductions in the Figures

Form	**Mnemonic**	**Proof**
Aab, Abc ⊢ Aac	*Barbara*	Perfect
Eab, Abc ⊢ Eac	*Celarent*	Perfect
Aab, Ibc ⊢ Iac	*Darii*	Perfect; also by impossibility, from *Camestres*
Eab, Ibc ⊢ Oac	*Ferio*	Perfect; also by impossibility, from *Cesare*
SECOND FIGURE		
Eab, Aac ⊢ Ebc	*Cesare*	(Eab, Aac)→(Eba, Aac)$\vdash_{Cel}$Ebc
Aab, Eac ⊢ Ebc	*Camestres*	(Aab, Eac)→(Aab, Eca)=(Eca, Aab)$\vdash_{Cel}$Ecb→Ebc
Eab, Iac ⊢ Obc	*Festino*	(Eab, Iac)→(Eba, Iac)$\vdash_{Fer}$Obc
Aab, Oac ⊢ Obc	*Baroco*	(Aab, Oac +Abc)$\vdash_{Bar}$(Aac, Oac)$\vdash_{Imp}$Obc
THIRD FIGURE		
Aac, Abc ⊢ Iab	*Darapti*	(Aac, Abc)→(Aac, Icb)$\vdash_{Dar}$Iab

Eac, Abc ⊢ Oab	*Felapton*	(Eac, Abc)→(Eac, Icb)$\vdash_{Fer}$Oab
Iac, Abc ⊢ Iab	*Disamis*	(Iac, Abc)→(Ica, Abc)=(Abc, Ica)$\vdash_{Dar}$Iba→Iab
Aac, Ibc ⊢ Iab	*Datisi*	(Aac, Ibc)→(Aac, Icb)$\vdash_{Dar}$Iab
Oac, Abc ⊢ Oab	*Bocardo*	(Oac, +Aab, Abc)$\vdash_{Bar}$(Aac, Oac)$\vdash_{Imp}$Oab
Eac, Ibc ⊢ Oab	*Ferison*	(Eac, Ibc)→(Eac, Icb)$\vdash_{Fer}$Oab

E. Metatheoretical Results

Having established which deductions in the figures are possible, Aristotle draws a number of metatheoretical conclusions, including:

1. No deduction has two negative premises
2. No deduction has two particular premises
3. A deduction with an affirmative conclusion must have two affirmative premises
4. A deduction with a negative conclusion must have one negative premise.
5. A deduction with a universal conclusion must have two universal premises

He also proves the following metatheorem:

> All deductions can be reduced to the two universal deductions in the first figure.

His proof of this is elegant. First, he shows that the two particular deductions of the first figure can be reduced, by proof through impossibility, to the universal deductions in the second figure:

> (*Darii*)(Aab, Ibc, +Eac)$\vdash_{Camestres}$(Ebc, Ibc)$\vdash_{Imp}$Iac
> (*Feri*)(Eab, Ibc, +Aac)$\vdash_{Cesare}$(Ebc, Ibc)$\vdash_{Imp}$Oac

He then observes that since he has already shown how to reduce all the particular deductions in the other figures except Baroco and Bocardo to *Darii* and *Ferio*, these deductions can thus be reduced to *Barbara* and *Celarent*. This proof is strikingly similar both in structure and in subject to modern proofs of the redundancy of axioms in a system.

Many more metatheoretical results, some of them quite sophisticated, are proved in *Prior Analytics* I.45 and in *Prior Analytics* II. As noted below, some of Aristotle's metatheoretical results are appealed to in the epistemological arguments of the *Posterior Analytics*.

F. Syllogisms with Modalities

Aristotle follows his treatment of "arguments in the figures" with a much longer, and much more problematic, discussion of what happens to these figured arguments when we add the qualifications "necessarily" and "possibly" to their premises in various ways. In contrast to the syllogistic itself (or, as commentators like to call it, the *assertoric* syllogistic), this *modal* syllogistic appears to be much less satisfactory and is certainly far more difficult to interpret. Here, I only outline Aristotle's treatment of this subject and note some of the principal points of interpretive controversy.

The Definitions of the Modalities

Modern modal logic treats necessity and possibility as interdefinable: "necessarily P" is equivalent to "not possibly not P", and "possibly P" to "not necessarily not P". Aristotle gives these same equivalences in *On Interpretation*. However, in *Prior Analytics*, he makes a distinction between two notions of possibility. On the first, which he takes as his preferred notion, "possibly P" is equivalent to "not necessarily P and not necessarily not P". He then acknowledges an alternative definition of possibility according to the modern equivalence, but this plays only a secondary role in his system.

Aristotle's General Approach

Aristotle builds his treatment of modal syllogisms on his account of non-modal (**assertoric**) syllogisms: he works his way through the syllogisms he has already proved and considers the consequences of adding a modal qualification to one or both premises. Most often, then, the questions he explores have the form: "Here is an assertoric syllogism; if I add these modal qualifications to the premises, then what modally qualified form of the conclusion (if any) follows?". A premise can have one of three modalities: it can be necessary, possible, or assertoric. Aristotle works through the combinations of these in order:

- Two necessary premises
- One necessary and one assertoric premise
- Two possible premises
- One assertoric and one possible premise
- One necessary and one possible premise

Though he generally considers only premise combinations which syllogize in their assertoric forms, he does sometimes extend this; similarly, he sometimes considers conclusions in addition to those which would follow from purely assertoric premises.

Since this is his procedure, it is convenient to describe modal syllogisms in terms of the corresponding non-modal syllogism plus a triplet of letters indicating the modalities of premises and conclusion: N = "necessary", P = "possible", A = "assertoric". Thus, "Barbara NAN" would mean "The form *Barbara* with necessary major premise, assertoric minor premise, and necessary conclusion". I use the letters "N" and "P" as prefixes for premises as well; a premise with no prefix is assertoric. Thus, *Barbara* NAN would be NAab, Abc, ⊢ NAac.

Modal Conversions

As in the case of assertoric syllogisms, Aristotle makes use of conversion rules to prove validity. The conversion rules for necessary premises are exactly analogous to those for assertoric premises:

- NEab→NEba
- NIab→NIba
- NAab→NIba

Possible premises behave differently, however. Since he defines "possible" as "neither necessary nor impossible", it turns out that *x is possibly F* entails, and is entailed by, *x is possibly not F*. Aristotle generalizes this to the case of categorical sentences as follows:

- PAab→PEab
- PEab→PAab
- PIab→POab
- POab→PIab

In addition, Aristotle uses the intermodal principle N→A: that is, a necessary premise entails the corresponding assertoric one. However, because of his definition of possibility, the principle A→P does not generally hold: if it did, then N→P would hold, but on his definition "necessarily P" and "possibly P" are actually inconsistent ("possibly P" entails "possibly not P").

This leads to a further complication. The denial of "possibly P" for Aristotle is "either necessarily P or necessarily not P". The denial of "necessarily P" is still more difficult to express in terms of a combination of modalities: "either possibly P (and thus possibly not P) or necessarily not P" This is important because of Aristotle's proof procedures, which include proof through impossibility. If we give a proof through impossibility in which we assume a necessary premise, then the conclusion we ultimately establish is simply the denial of that necessary premise, not a "possible" conclusion in Aristotle's sense. Such propositions do occur in his system, but only in exactly this way, i.e. as conclusions established by proof through impossiblity from necessary assumptions. Somewhat confusingly, Aristotle calls such propositions "possible" but immediately adds " not in the sense defined": in this sense, "possibly Oab" is simply the denial of "necessarily Aab". Such propositions appear only as premises, never as conclusions.

Syllogisms with Necessary Premises

Aristotle holds that an assertoric syllogism remains valid if "necessarily" is added to its premises and its conclusion: the modal pattern NNN is always valid. He does not treat this as a trivial consequence but instead offers proofs; in all but two cases, these are parallel to those offered for the assertoric case. The exceptions are *Baroco* and *Bocardo*, which he proved in the assertoric case through impossibility: attempting to use that method here would require him to take the denial of a necessary O proposition as hypothesis, raising the complication noted above, and he must resort to a different form of proof instead.

NA/AN Combinations: The Problem of the "Two Barbaras" and Other Difficulties

Since a necessary premise entails an assertoric premise, every AN or NA combination of premises will entail the corresponding AA pair, and thus the corresponding A conclusion. Thus, ANA and NAA syllogisms are always valid. However, Aristotle holds that some, but not all, ANN and NAN combinations are valid. Specifically, he accepts *Barbara* NAN but rejects *Barbara* ANN. Almost from Aristotle's own time, interpreters have found his reasons for this distinction obscure, or unpersuasive, or both. Theophrastus, for instance, adopted the simpler rule that the modality of the conclusion of a syllogism was always the "weakest" modality found in either premise, where N is stronger than A and A is stronger than P (and where P probably has to be defined as "not necessarily not"). Other difficulties follow from the problem of the "Two Barbaras", as it is often called, and it has often been maintained that the modal syllogistic is inconsistent.

This subject quickly becomes too complex for summarizing in this brief article. For further discussion, see Becker, McCall, Patterson, van Rijen, Striker, Nortmann, Thom, and Thomason.

§6: Demonstrations and Demonstrative Sciences

A **demonstration** (*apodeixis*) is "a deduction that produces knowledge". Aristotle's *Posterior Analytics* contains his account of demonstrations and their role in knowledge. From a modern perspective, we might think that this subject moves outside of logic to epistemology. From Aristotle's perspective, however, the connection of the theory of *sullogismoi* with the theory of knowledge is especially close.

A. Aristotelian Sciences

The subject of the *Posterior Analytics* is *epistêmê*. This is one of several Greek words that can reasonably be translated "knowledge", but Aristotle is concerned only with knowledge of a certain type (as will be explained below). There is a long tradition of translating *epistêmê* in this technical sense as **science**, and I shall follow that tradition here. However, readers should not be misled by the use of that word. In particular, Aristotle's theory of science cannot be considered a counterpart to modern philosophy of science, at least not without substantial qualifications.

We have scientific knowledge, according to Aristotle, when we know:

> the cause why the thing is, that it is the cause of this, and that this cannot be otherwise. (*Posterior Analytics* I.2)

This implies two strong conditions on what can be the object of scientific knowledge:

1. Only what is necessarily the case can be known scientifically

2. Scientific knowledge is knowledge of causes

He then proceeds to consider what science so defined will consist in, beginning with the observation that at any rate one form of science consists in the possession of a **demonstration** (*apodeixis*), which he defines as a "scientific deduction":

> by "scientific" (*epistêmonikon*), I mean that in virtue of possessing it, we have knowledge.

The remainder of *Posterior Analytics* I is largely concerned with two tasks: spelling out the nature of demonstration and demonstrative science and answering an important challenge to its very possibility. Aristotle first tells us that a demonstration is a deduction in which the premises are:

1. true
2. **primary** (*prota*)
3. **immediate** (*amesa*, "without a middle")
4. **better known** or **more familiar** (*gnôrimôtera*) than the conclusion
5. **prior** to the conclusion
6. **causes** (*aitia*) of the conclusion

The interpretation of all these conditions except the first has been the subject of much controversy. Aristotle clearly thinks that science is knowledge of causes and that in a demonstration, knowledge of the premises is what brings about knowledge of the conclusion. The fourth condition shows that the knower of a demonstration must be in some better epistemic condition towards them, and so modern interpreters often suppose that Aristotle has defined a kind of epistemic justification here. However, as noted above, Aristotle is defining a special variety of knowledge. Comparisons with discussions of justification in modern epistemology may therefore be misleading.

The same can be said of the terms "primary", "immediate" and "better known". Modern interpreters sometimes take "immediate" to mean "self-evident"; Aristotle does say that an immediate proposition is one "to which no other is prior", but (as I suggest in the next section) the notion of priority involved is likely a notion of logical priority that it is hard to detach from Aristotle's own logical theories. "Better known" has sometimes been interpreted simply as "previously known to the knower of the demonstration" (i.e. already known in advance of the demonstration). However, Aristotle explicitly distinguishes between what is "better known for us" with what is "better known in itself" or "in nature" and says that he means the latter in his definition. In fact, he says that the process of acquiring scientific knowledge is a process of *changing* what is better known "for us", until we arrive at that condition in which what is better known in itself is also better known for us.

B. The Regress Problem

In *Posterior Analytics* I.2, Aristotle considers two challenges to the possibility of science. One party (dubbed the "agnostics" by Jonathan Barnes) began with the following two premises:

1. Whatever is scientifically known must be demonstrated.
2. The premises of a demonstration must be scientifically known.

They then argued that demonstration is impossible with the following dilemma:

1. If the premises of a demonstration are scientifically known, then they must be demonstrated.
2. The premises from which each premise are demonstrated must be scientifically known.
3. Either this process continues forever, creating an infinite regress of premises, or it comes to a stop at some point.
4. If it continues forever, then there are no first premises from which the subsequent ones are demonstrated, and so nothing is demonstrated.
5. On the other hand, if it comes to a stop at some point, then the premises at which it comes to a stop are undemonstrated and therefore not scientifically known; consequently, neither are any of the others deduced from them.
6. Therefore, nothing can be demonstrated.

A second group accepted the agnostics' view that scientific knowledge comes only from demonstration but rejected their conclusion by rejecting the dilemma. Instead, they maintained:

- Demonstration "in a circle" is possible, so that it is possible for all premises also to be conclusions and therefore demonstrated.

Aristotle does not give us much information about how circular demonstration was supposed to work, but the most plausible interpretation would be supposing that at least for some set of fundamental principles, each principle could be deduced from the others. (Some modern interpreters have compared this position to a coherence theory of knowledge.) However their position worked, the circular demonstrators claimed to have a third alternative avoiding the agnostics' dilemma, since circular demonstration gives us a regress that is both unending (in the sense that we never reach premises at which it comes to a stop) and finite (because it works its way round the finite circle of premises).

C. Aristotle's Solution: "It Eventually Comes to a Stop"

Aristotle rejects circular demonstration as an incoherent notion on the grounds that the premises of any demonstration must be prior (in an appropriate sense) to the conclusion, whereas a circular demosntration would make the same premises both prior and posterior to one another (and indeed every premise prior and posterior to itself). He agrees with the agnostics' analysis of the regress problem: the only plausible options are that it continues indefinitely or that it "comes to a stop" at some point. However, he thinks both the agnostics and the circular demosntrators are wrong in maintaining that scientific knowledge is only possible by demonstration from premises scientifically known: instead, he claims, there is another form of knowledge possible for the first premises, and this provides the starting points for demonstrations.

To solve this problem, Aristotle needs to do something quite specific. It will not be enough for him to establish that we can have knowledge of *some propositions* without demonstrating them: unless it is in turn possible to deduce all the other propositions of a science from them, we shall not have solved the regress problem. Moreover (and obviously), it is no solution to this problem for Aristotle simply to *assert* that we have knowledge without demonstration of some appropriate starting points. He does indeed say that it is his position that we have such knowledge (*An. Post.* I.2,), but he owes us an account of why that should be so.

D. Knowledge of First Principles: *Nous*

Aristotle's account of knowledge of the indemonstrable first premises of sciences is found in *Posterior Analytics* II.19, long regarded as a difficult text to interpret. Briefly, what he says there is that it is another cognitive state, *nous* (translated variously as "insight", "intuition", "intelligence"), which knows them. There is wide disagreement among commentators about the interpretation of his account of how this state is reached; I will offer one possible interpretation. First, Aristotle identifies his problem as explaining how the principles can "become familiar to us", using the same term "familiar" (*gnôrimos*) that he used in presenting the regress problem. What he is presenting, then, is not a method of discovery but a process of becoming wise. Second, he says that in order for knowledge of immediate premises to be possible, we must have a kind of knowledge of them without having learned it, but this knowledge must not be as "precise" as the knowledge that a possessor of science must have. The kind of knowledge in question turns out to be a capacity or power (*dunamis*) which Aristotle compares to the the capacity for sense-perception: since our senses are innate, i.e. develop naturally, it is in a way correct to say that we know what e.g. all the colors look like before we have seen them: we have the capacity to see them by nature, and when we first see a color we exercise this capacity without having to learn how to do so first. Likewise, Aristotle holds, our minds have by nature the capacity to recognize the starting points of the sciences.

In the case of sensation, the capacity for perception in the sense organ is actualized by the operation on it of the perceptible object. Similarly, Aristotle holds that coming to know first premises is a matter of a potentiality in the mind being actualized by experience of its proper objects: "The soul is of such a nature as to be capable of undergoing this". So, although we cannot come to know the first premises without the necessary experience, just as we cannot see colors without the presence of colored objects, our minds are already so constituted as to be able to recognize the right objects, just as our eyes are already so constituted as to be able to perceive the colors that exist.

It is considerably less clear what these objects are and how it is that experience actualizes the relevant potentialities in the soul. Aristotle describes a series of stages of cognition. First is what is common to all animals: perception of what is present. Next is memory, which he regards as a retention of a sensation: only some animals have this capacity. Even fewer have the next capacity, the capacity to form a single experience (*empeiria*) from many repetitions of the same memory. Finally, many experiences repeated give rise to knowledge of a single universal (*katholou*). This last capacity is present only in humans.

See the article on Aristotle's psychology for more on his views about mind.

§7: Definitions

The **definition** (*horos*, *horismos*) was an important matter for Plato and for the Early Academy. Concern with answering the question "What is so-and-so?" are at the center of the majority of Plato's dialogues, some of which (most elaborately the *Sophist*) propound methods for finding definitions. External sources (sometimes the satirical remarks of comedians) also reflect this Academic concern with definitions. Aristotle himself traces the quest for definitions back to Socrates.

A. Definitions and Essences

For Aristotle, a definition is "an account which signifies what it is to be for something" (*logos ho to ti ên einai sêmainei*). The phrase "what it is to be" and its variants are crucial: giving a definition is saying, of some existent thing, what it is, not simply specifying the meaning of a word (Aristotle does recognize definitions of the latter sort, but he has little interest in them).

The notion of "what it is to be" for a thing is so pervasive in Aristotle that it becomes formulaic: what a definition expresses is "the what-it-is-to-be" (*to ti ên einai*). Roman translators, vexed by this odd Greek phrase, devised a word for it, *essentia*, from which our "essence" descends. So, an Aristotelian definition is an account of the essence of something.

B. Species, Genus, and Differentia

Since a definition defines an essence, only what has an essence can be defined. What has an essence, then? That is one of the central questions of Aristotle's metaphysics; once again, we must leave the details to another article. In general, however, it is not individuals but rather **species** (*eidos*: the word is one of those Plato uses for "Form") that have essences. A species is defined by giving its **genus** (*genos*) and its **differentia** (*diaphora*): the genus is the kind under which the species falls, and the differentia tells what characterizes the species within that genus. As an example, *human* might be defined as *animal* (the genus) *having the capacity to reason* (the differentia).

Essential Predication and the Predicables

Underlying Aristotle's concept of a definition is the concept of **essential predication** (*katêgoreisthai en tôi ti esti*, predication in the what it is). In any true affirmative predication, the predicate either does or does not "say what the subject is", i.e., the predicate either is or is not an acceptable answer to the question "What is it?" asked of the subject. Bucephalus is a horse, and a horse is an animal; so, "Bucephalus is a horse" and "Bucephalus is an animal" are essential predications. However, "Bucephalus is brown", though true, does not state what Bucephalus us but only says something about him.

Since a thing's definition says what it is, definitions are essentially predicated. However, not everything essentially predicated is a definition. Since Bucephalus is a horse, and horses are a kind of mammal, and mammals are a kind of animal, "horse" "mammal" and "animal" are all essential predicates of Bucephalus. Moreover, since what a horse is is a kind of mammal, "mammal" is an essential predicate of horse. When predicate X is an essential predicate of Y but also of other things, then X is a **genus** (*genos*) of Y.

A definition of X must not only be essentially predicated of it but must also be predicated only of it: to use a term from Aristotle's *Topics*, a definition and what it defines must "counterpredicate" (*antikatêgoreisthai*) with one another. X counterpredicates with Y if X applies to what Y applies to and conversely. Though X's definition must counterpredicate with X, not everything that counterpredicates with X is its definition. "Capable of laughing", for example, counterpredicates with "human" but fails to be its definition. Such a predicate (non-essential but counterpredicating) is a **peculiar property** or **proprium** (*idion*).

Finally, if X is predicated of Y but is neither essential nor counterpredicates, then X is an **accident** (*sumbebêkos*) of Y.

Aristotle sometimes treats genus, peculiar property, definition, and accident as including all possible predications (e.g. *Topics* I). Later commentators listed these four and the differentia as the five **predicables**, and as such they were of great importance to late ancient and to medieval philosophy (e.g., Porphyry).

C. The Categories

The notion of essential predication is connected to what are traditionally called the **categories** (*katêgoriai*). In a word, Aristotle is famous for having held a "doctrine of categories". Just what that doctrine was, and indeed just what a category is, are considerably more vexing questions. They also quickly take us outside his logic and into his metaphysics. Here, I will try to give a very general overview, beginning with the somewhat simpler question "What categories are there?"

We can answer this question by listing the categories. Here are two passages containing such lists:

> We should distinguish the kinds of predication (*ta genê tôn katêgoriôn*) in which the four predications mentioned are found. These are ten in number: what-it-is, quantity, quality, relative, where, when, being-in-a-position, having, doing, undergoing. An accident, a genus, a peculiar property and a definition will always be in one of these categories. (*Topics* I.9, 103b20-25)

> Of things said without any combination, each signifies either substance or quantity or quality or a relative or where or when or being-in-a-position or having or doing or undergoing. To give a rough idea, examples of substance are man, horse; of quantity: four-

> foot, five-foot; of quality: white, literate; of a relative: double, half, larger; of where: in the Lyceum, in the market-place; of when: yesterday, last year; of being-in-a-position: is-lying, is-sitting; of having: has-shoes-on, has-armor-on; of doing: cutting, burning; of undergoing: being-cut, being-burned. (*Categories* 4, 1b25-2a4, tr. Ackrill, slightly modified)

These two passages give ten-item lists, identical except for their first members. What are they lists *of*? Here are three ways they might be interpreted:

The word "category" (*katêgoria*) means "predication". Aristotle holds that predications and predicates can be grouped into several largest "kinds of predication" (*genê tôn katêgoriôn*). He refers to this classification frequently, often calling the "kinds of predication" simply "the predications", and this (by way of Latin) leads to our word "category".

- First, the categories may be *kinds of predicate*: predicates (or, more precisely, predicate expressions) can be divided into ten separate classes, with each expression belonging to just one class. This comports well with the root meaning of the word *katêgoria* ("predication"). On this interpretation, the categories arise out of considering the most general types of question that can be asked about something: "*What* is it?"; "*How much* is it?"; "*What sort* is it?"; "*Where* is it?"; "*What* is it *doing*?" Answers appropriate to one of these questions are nonsensical in response to another ("When is it?" "A horse"). Thus, the categories may rule out certain kinds of question as ill-formed or confused. This plays an important role in Aristotle's metaphysics.
- Second, the categories may be seen as classifications of *predications*, that is, kinds of relation that may hold between the predicate and the subject of a predication. To say of Socrates that he is human is to say what he *is*, whereas to say that he is literate is not to say what he is but rather to give a quality that he *has*. For Aristotle, the relation of predicate to subject in these two sentences is quite different (in this respect he differs both from Plato and from modern logicians). The categories may be interpreted as ten different ways in which a predicate may be related to its subject. This last division has importance for Aristotle's logic as well as his metaphysics.
- Third, the categories may be seen as *kinds of entity*, as highest genera or kinds of thing that are. A given thing can be classified under a series of progressively wider genera: Socrates is a human, a mammal, an animal, a living being. The categories are the highest such genera. Each falls under no other genus, and each is completely separate from the others. This distinction is of critical importance to Aristotle's metaphysics.

Which of these interpretations fits best with the two passages above? The answer appears to be different in the two cases. This is most evident if we take note of point in which they differ: the *Categories* lists **substance** (*ousia*) in first place, while the *Topics* list **what-it-is** (*ti esti*). A substance, for Aristotle, is a type of entity, suggesting that the *Categories* list is a list of types of entity.

On the other hand, the expression "what-it-is" suggests most strongly a type of predication. Indeed, the *Topics* confirms this by telling us that we can "say what it is" of an *entity* falling under any of the categories:

> an expression signifying what-it-is will sometimes signify a substance, sometimes a quantity, sometimes a quality, and sometimes one of the other categories.

As Aristotle explains, if I say that Socrates is a man, then I have said what Socrates is and signified a substance; if I say that white is a color, then I have said what white is and signified a quality; if I say that some length is a foot long, then I have said what it is and signified a quantity; and so on for the other categories. What-it-is, then, here designates a kind of predication, not a kind of entity.

This might lead us to conclude that the categories in the *Topics* are only to be interpreted as kinds of predicate or predication, those in the *Categories* as kinds of being. Even so, we would still want to ask what the relationship is between these two nearly-identical lists of terms, given these distinct interpretations. However, the situation is much more complicated. First, there are dozens of other passages in which the categories appear. Nowhere else do we find a list of ten, but we do find shorter lists containing eight, or six, or five, or four of them (with substance/what-it-is, quality, quantity, and relative the most common). Aristotle describes what these lists are lists of in different ways: they tell us "how being is divided", or "how many ways being is said", or "the figures of predication" (ta schêmata tês katêgorias). The designation of the first category also varies: we find not only "substance" and "what it is" but also the expressions "this" or "the this" (*tode ti, to tode, to ti*). These latter expressions are closely associated with, but not synonymous with, substance. He even combines the latter with "what-it-is" (*Metaphysics* Z 1, 1028a10: ". . . one sense signifies what it is and the this, one signifies quality . . .").

Moreover, substances are for Aristotle fundamental for predication as well as metaphysically fundamental. He tells us that everything that exists exists because substances exist: if there were no substances, there would not be anything else. He also conceives of predication as reflecting a metaphysical relationship (or perhaps more than one, depending on the type of predication). The sentence "Socrates is pale" gets its truth from a state of affairs consisting of a substance (Socrates) and a quality (whiteness) which is in that substance. At this point we have gone far outside the realm of Aristotle's logic into his metaphysics, the fundamental question of which, according to Aristotle, is "What is a substance?". (For further discussion of this topic, see the entry on Aristotle's metaphysics, and in particular, Section 2 on the categories.)

See Frede 1981, Ebert 1985 for additional discussion of Aristotle's lists of categories.

For convenience of reference, I include a table of the categories, along with Aristotle's examples and the traditional names often used for them. For reasons explained above, I have treated the first item in the list quite differently, since an example of a substance and an example of a what-it-is are necessarily (as one might put it) in different categories.

Traditional name	Literally	Greek	Examples

(Substance)	substance "this" what-it-is	*ousia* *tode ti* *ti esti*	man, horse Socrates "Socrates is a man"
Quantity	How much	*poson*	four-foot, five-foot
Quality	What sort	*poion*	white, literate
Relation	related to what	*pros ti*	double, half, greater
Location	Where	*pou*	in the Lyceum, in the marketplace
Time	when	*pote*	yesterday, last year
Position	being situated	*keisthai*	lies, sits
Habit	having, possession	*echein*	is shod, is armed
Action	doing	*poiein*	cuts, burns
Passion	undergoing	*paschein*	is cut, is burned

D. The Method of Division

In the *Sophist*, Plato introduces a procedure of "Division" as a method for discovering definitions. To find a definition of X, first locate the largest kind of thing under which X falls; then, divide that kind into two parts, and decide which of the two X falls into. Repeat this method with the part until X has been fully located.

This method is part of Aristotle's Platonic legacy. His attitude towards it, however, is complex. He adopts a view of the proper structure of definitions that is closely allied to it: a correct definition of X should give the **genus** (*genos*: kind or family) of X, which tells what kind of thing X is, and the **differentia** (*diaphora*: difference) which uniquely identifies X within that genus. Something defined in this way is a **species** (*eidos*: the term is one of Plato's terms for "Form"), and the differentia is thus the "difference that makes a species" (*eidopoios diaphora*, "specific difference"). In *Posterior Analytics* II.13, he gives his own account of the use of Division in finding definitions.

However, Aristotle is strongly critical of the Platonic view of Division as a method for *establishing* definitions. In *Prior Analytics* I.31, he contrasts Division with the syllogistic method he has just presented, arguing that Division cannot actually prove anything but rather assumes the very thing it is supposed to be proving. He also charges that the partisans of Division failed to understand what their own method was capable of proving.

E. Definition and Demonstration

Closely related to this is the discussion, in *Posterior Analytics* II.3-10, of the question whether there can be both definition and demonstration of the same thing. Since the definitions Aristotle is interested in are

statements of essences, knowing a definition is knowing, of some existing thing, what it is. Consequently, Aristotle's question amounts to a question whether defining and demonstrating can be alternative ways of acquiring the same knowledge. His reply is complex:

1. Not everything demonstrable can be known by finding definitions, since all definitions are universal and affirmative whereas some demonstrable propositions are negative.
2. If a thing is demonstrable, then to know it just is to possess its demonstration; therefore, it cannot be known just by definition.
3. Nevertheless, some definitions can be understood as demonstrations differently arranged.

As an example of case 3, Aristotle considers the definition "Thunder is the extinction of fire in the clouds". He sees this as a compressed and rearranged form of this demonstration:

- Sound accompanies the extinguishing of fire.
- Fire is extinguished in the clouds.
- Therefore, a sound occurs in the clouds.

We can see the connection by considering the answers to two questions: "What is thunder?" "The extinction of fire in the clouds" (definition). "Why does it thunder?" "Because fire is extinguished in the clouds" (demonstration).

As with his criticisms of Division, Aristotle is arguing for the superiority of his own concept of science to the Platonic concept. Knowledge is composed of demonstrations, even if it may also include definitions; the method of science is demonstrative, even if it may also include the process of defining.

§8: Dialectical Argument and the Art of Dialectic

Aristotle often contrasts *dialectical arguments* with demonstrations. The difference, he tells us, is in the character of their premises, not in their logical structure: whether an argument is a *sullogismos* is only a matter of whether its conclusion results of necessity from its premises. The premises of demonstrations must be *true and primary*, that is, not only true but also prior to their conclusions in the way explained in the *Posterior Analytics*. The premises of dialectical deductions, by contrast, must be **accepted** (*endoxos*).

A. Dialectical Premises: The Meaning of *Endoxos*

Recent scholars have proposed different interpretations of the term *endoxos*. Aristotle often uses this adjective as a substantive: *ta endoxa*, "accepted things", "accepted opinions". On one understanding, descended from the work of G. E. L. Owen and developed more fully by Jonathan Barnes and especially Terence Irwin, the *endoxa* are a compilation of views held by various people with some form or other of standing: "the views of fairly reflective people after some reflection", in Irwin's phrase. Dialectic is then simply "a method of argument from [the] common beliefs [held by these people]". For Irwin, then,

endoxa are "common beliefs". Jonathan Barnes, noting that *endoxa* are opinions with a certain standing, translates with "reputable".

My own view is that Aristotle's texts support a somewhat different understanding. He also tells us that dialectical premises differ from demonstrative ones in that the former are *questions*, whereas the latter are *assumptions* or *assertions*: "the demonstrator does not ask, but takes", he says. This fits most naturally with a view of dialectic as argument directed at another person by question and answer and consequently taking as premises that other person's concessions. Anyone arguing in this manner will, in order to be successful, have to ask for premises which the interlocutor is liable to accept, and the best way to be successful at that is to have an inventory of acceptable premises, i.e. premises that are in fact acceptable to people of different types.

In fact, we can discern in the *Topics* (and the *Rhetoric*, which Aristotle says depends on the art explained in the *Topics*) an art of dialectic for use in such arguments. My reconstruction of this art (which would not be accepted by all scholars) is as follows.

B. The Two Elements of the Art of Dialectic

Given the above picture of dialectical argument, the dialectical art will consist of two elements. One will be a method for discovering premises from which a given conclusion follows, while the other will be a method for determining which premises a given interlocutor will be likely to concede. The first task is accomplished by developing a system for classifying premises according to their logical structure. We might expect Aristotle to avail himself here of the syllogistic, but in fact he develops quite another approach, one that seems less systematic and rests on various "common" terms. The second task is accomplished by developing lists of the premises which are acceptable to various types of interlocutor. Then, once one knows what sort of person one is dealing with, one can choose premises accordingly. Aristotle stresses that, as in all arts, the dialectician must study, not what is acceptable to this or that specific person, but what is acceptable to this or that type of person, just as the doctor studies what is healthful for different types of person: "art is of the universal".

The "Logical System" of the *Topics*

The method presented in the *Topics* for classifying arguments relies on the presence in the conclusion of certain "common" terms (*koina*) -- common in the sense that they are not peculiar to any subject matter but may play a role in arguments about anything whatever. We find enumerations of arguments involving these terms in a similar order several times. Typically, they include:

I. Opposites (*antikeimena*, *antitheseis*)
    1. Contraries (*enantia*)
    2. Contradictories (*apophaseis*)
    3. Possession and Privation (*hexis kai sterêsis*)
    4. Relatives (*pros ti*)

II. Cases (*ptôseis*)
III. "More and Less and Likewise"

The four types of **opposites** are the best represented. Each designates a type of term pair, i.e. a way two terms can be opposed to one another. **Contraries** are polar opposites or opposed extremes such as hot and cold, dry and wet, good and bad. A pair of **contradictories** consists of a term and its negation: good, not good. A **possession** (or condition) and **privation** are illustrated by sight and blindness.**Relatives** are relative terms in the modern sense: a pair consists of a term and its correlative, e.g. large and small, parent and child.

The argumentative patterns Aristotle associated with **cases** generally involve inferring a sentence contaning adverbial or declined forms from another sentence containing different forms of the same word stem: "if what is useful is good, then what is done usefully is done well and the useful person is good". In Hellenistic grammatical usage, *ptôsis* meant "case" (e.g. nominative, dative, accusative); Aristotle's use here is obviously an early form of that.

Under the heading **more and less and likewise**, Aristotle groups a somewhat motley assortment of argument patterns all involving, in some way or other, the terms "more", "less", and "likewise". Examples: "If whatever is A is B, then whatever is more (less) A is more (less) B"; "If A is more likely B than C is, and A is not B, then neither is C"; "If A is more likely than B and B is the case, then A is the case".

The *Topoi*

At the heart of the *Topics* is a collection of what Aristotle calls *topoi*, "places" or "locations". Unfortunately, though it is clear that he intends most of the *Topics* (Books II-VI) as a collection of these, he never explicitly defines this term. Interpreters have consequently disagreed considerably about just what a *topos* is. A discussion may be found in Brunschwig 1967, Slomkowski 1996, Primavesi 1997, and Smith 1997.

C. The Uses of Dialectic and Dialectical Argument

An *art* of dialectic will be useful wherever dialectical argument is useful. Aristotle mentions three such uses; each merits some comment.

Gymnastic Dialectic

First, there appears to have been a form of stylized argumentative exchange practiced in the Academy in Aristotle's time. The main evidence for this is simply Aristotle's *Topics*, especially Book VIII, which makes frequent reference to rule-governed procedures, apparently taking it for granted that the audience will understand them. In these exchanges, one participant took the role of answerer, the other the role of questioner. The answerer began by asserting some proposition (a *thesis*: "position" or "acceptance"). The

questioner then asked questions of the answerer in an attempt to secure concessions from which a contradiction could be deduced: that is, to **refute** (*elenchein*) the answerer's position. The questioner was limited to questions that could be answered by yes or no; generally, the answerer could only respond with yes or no, though in some cases answeres could object to the form of a question. Answerers might undertake to answer in accordance with the views of a particular type of person or a particular person (e.g. a famous philosopher), or they might answer according to their own beliefs. There appear to have been judges or scorekeepers for the process. Gymnastic dialectical contests were sometimes, as the name suggests, for the sake of exercise in developing argumentative skill, but they may also have been pursued as a part of a process of inquiry.

Dialectic That Puts to the Test

Aristotle also mentions an "art of making trial", or a variety of dialectical argument that "puts to the test" (the Greek word is the adjective *peirastikê*, in the feminine: such expressions often designate arts or skills, e.g. *rhêtorikê*, "the art of rhetoric"). Its function is to examine the claims of those who say they have some knowledge, and it can practiced by someone who does not possess the knowledge in question. The examination is a matter of refutation, based on the principle that whoever knows a subject must have consistent beliefs about it: so, if you can show me that my beliefs about something lead to a contradiction, then you have shown that I do not have knowledge about it.

This is strongly reminiscent of Socrates' style of interrogation, from which it is almost certainly descended. In fact, Aristotle often indicates that dialectical argument is by nature refutative.

Dialectic and Philosophy

Dialectical refutation cannot of itself establish any proposition (except perhaps the proposition that some set of propositions is inconsistent). More to the point, though deducing a contradiction from my beliefs may show that they do not constitute knowledge, failure to deduce a contradiction from them is no proof that they are true. Not surprisingly, then, Aristotle often insists that "dialectic does not prove anything" and that the dialectical art is not some sort of universal knowledge.

In *Topics* I.2, however, Aristotle says that the art of dialectic is useful in connection with "the philosophical sciences". One reason he gives for this follows closely on the refutative function: if we have subjected our opinions (and the opinions of our fellows, and of the wise) to a thorough refutative examination, we will be in a much better position to judge what is most likely true and false. In fact, we find just such a procedure at the start of many of Aristotle's treatises: an enumeration of the opinions current about the subject together with a compilation of "puzzles" raised by these opinions. Aristotle has a special term for this kind of review: a *diaporia*, a "puzzling through".

He adds a second use that is both more difficult to understand and more intriguing. The *Posterior Analytics* argues that if anything can be proved, then not everything that is known is known as a result of proof. What alternative means is there whereby the first principles of sciences are known? Aristotle's own

answer as found in *Posterior Analytics* II.19 is difficult to interpret, and recent philosophers have often found it unsatisfying since (as often construed) it appears to commit Aristotle to a form of apriorism or rationalism both indefensible in itself and not consonant with his own insistence on the indispensability of empirical inquiry in natural science.

Against this background, the following passage in *Topics* I.2 may have special importance:

> It is also useful in connection with the first things concerning each of the sciences. For it is impossible to say anything about the science under consideration on the basis of its own principles, since the principles are first of all, and we must work our way through about these by means of what is generally accepted about each. But this is peculiar, or most proper, to dialectic: for since it is examinative with respect to the principles of all the sciences, it has a way to proceed.

A number of interpreters (beginning with Owen 1961) have built on this passage and others to find dialectic at the heart of Aristotle's philosophical method. Further discussion of this issue would take us far beyond the subject of this article (the fullest development is in Irwin 1988; see also Nussbaum 1986 and, for criticism, Hamlyn 1990, Smith 1997).

§9: Dialectic and Rhetoric

Aristotle says that rhetoric, i.e. the study of persuasive speech, is a "counterpart" (*antistrophos*) of dialectic and that the rhetorical art is a kind of "outgrowth" (*paraphues ti*) of dialectic and the study of character types. The correspondence with dialectical method is straightforward: rhetorical speeches, like dialectical arguments, seek to persuade others to accept certain conclusions on the basis of premises they already accept. Therefore, the same measures useful in dialectical contexts will, mutatis mutandis, be useful here: knowing what premises an audience of a given type is likely to believe, and knowing how to find premises from which the desired conclusion follows.

The *Rhetoric* does fit this general description: Aristotle includes both discussions of types of person or audience (with generalizations about what each type tends to believe) and a summary version (in II.23) of the argument patterns discussed in the *Topics*. For further discussion of his rhetoric see Aristotle's rhetoric.

§10: Sophistical Arguments

Demonstrations and dialectical arguments are both forms of valid argument, for Aristotle. However, he also studies what he calls **contentious** (*eristikos*) or **sophistical** arguments: these he defines as arguments which only apparently establish their conclusions. In fact, Aristotle defines these as apparent (but not genuine) *dialectical sullogismoi*. They may have this appearance in either of two ways:

1. Arguments in which the conclusion only appears to follow of necessity from the premises (apparent, but not genuine, *sullogismoi*).
2. Genuine *sullogismois* the premises of which are merely apparently, but not genuinely, acceptable.

Arguments of the first type in modern terms, appear to be valid but are really invalid. Arguments of the second type are at first more perplexing: given that acceptability is a matter of what people believe, it might seem that whatever appears to be *endoxos* must actually be *endoxos*. However, Aristotle probably has in mind arguments with premises that may at first glance *seem* to be acceptable but which, upon a moment's reflection, we immediately realize we don not actually accept. Consider this example from Aristotle's time:

- Whatever you have not lost, you still have.
- You have not lost horns.
- Therefore, you still have horns

This is transparently bad, but the problem is not that it is invalid: the problem is rather that the first premise, though superficially plausible, is false. In fact, anyone with a little ability to follow an argument will realize that at once upon seeing this very argument.

Aristotle's study of sophistical arguments is contained in *On Sophistical Refutations,* which is actually a sort of appendix to the *Topics*.

To a remarkable extent, contemporary discussions of fallacies reproduce Aristotle's own classifications. See Dorion 1995 for further discussion.

§11: Non-Contradiction and Metaphysics

Two frequent themes of Aristotle's account of science are (1) that the first principles of sciences are not demonstrable and (2) that there is no single universal science including all other sciences as its parts. "All things are not in a single genus", he says, "and even if they were, all beings could not fall under the same principles" (*On Sophistical Refutations* 11). Thus, it is exactly the universal applicability of dialectic that leads him to deny it the status of a science.

In *Metaphysics* IV (Γ), however, Aristotle takes what appears to be a different view. First, he argues that there is, in a way, a science that takes being as its genus (his name for it is "first philosophy"). Second, he argues that the principles of this science will be, in a way, the first principles of all (though he does not claim that the principles of other sciences can be demonstrated from them). Third, he identifies one of its first principles as the "most secure" of all principles: the principle of non-contradiction. As he states it,

> It is impossible for the same thing to belong and not belong simultaneously to the same thing in the same respect (*Met.*)

This is the most secure of all principles, Aristotle tells us, because "it is impossible to be in error about it". Since it is a first principle, it cannot be demonstrated; those who think otherwise are "uneducated in analytics". However, Aristotle then proceeds to give what he calls a "refutative demonstration" (*apodeixai elenktikôs*) of this principle.

Further discussion of this principle and Aristotle's arguments concerning it belong to a treatment of his metaphysics (see Aristotle: Metaphysics). However, it should be noted that: (1) these arguments draw on Aristotle's views about logic to a greater extent than any treatise outside the logical works themselves; (2) in the logical works, the principle of non-contradiction is one of Aristotle's favorite illustrations of the "common principles" (*koinai archai*) that underlie the art of dialectic.

See Aristotle's Metaphysics, Dancy 1975, Code 1986 for further discussion.

§12: Time and Necessity: The Sea-Battle

The passage in Aristotle's logical works which has received perhaps the most intense discussion in recent decades is *On Interpretation* 9, where Aristotle discusses the question whether every proposition about the future must be either true or false. Though something of a side issue in its context, the passage raises a problem of great importance to Aristotle's near contemporaries (and perhaps contemporaries).

A **contradiction** (*antiphasis*) is a pair of propositions one of which asserts what the other denies. A major goal of *On Interpretation* is to discuss the thesis that, of every such contradiction, one member must be true and the other false. In the course of his discussion, Aristotle allows for some exceptions. One case is what he calls **indefinite** propositions such as "A man is walking": nothing prevents both this proposition and "A man is not walking" being simultaneously true. This exception can be explained on relatively simple grounds.

A different exception arises for more complex reasons. Consider these two propositions:

1. There will be a sea-battle tomorrow
2. There will not be a sea-battle tomorrow

It seems that exactly one of these must be true and the other false. But if (1) is *now* true, then there *must* be a sea-battle tomorrow, and there *cannot* fail to be a sea-battle tomorrow. The result, according to this puzzle, is that nothing is possible except what actually happens: there are no unactualized possibilities.

Such a conclusion is, as Aristotle is quick to note, a problem both for his own metaphysical views about potentialities and for the commonsense notion that some things are up to us. He therefore proposes another exception to the general thesis concerning contradictory pairs.

This much would probably be accepted by most interpreters. What the restriction is, however, and just what motivates it are matters of wide disagreement. It has been proposed, for instance, that Aristotle adopted, or at least flirted with, a three-valued logic for future propositions, or that he countenanced truth-value gaps, or that his solution includes still more abstruse reasoning. The literature is much too complex to summarize: see Anscombe, Hintikka, D. Frede, Whitaker, Waterlow.

Historically, at least, it is likely that Aristotle is responding to an argument originating in the Megarian School. He ascribes the view that only that which happens is possible to the Megarians in *Metaphysics* IX (Θ). The puzzle with which he is concerned strongly recalls the "Master Argument" of Diodorus Cronus, especially in certain further details. For instance, Aristotle imagines the statement about tomorrow's sea battle having been uttered ten thousand years ago. If it was true, then its truth was a fact about the past; if the past is now unchangeable, then so is the truth value of that past utterance. This recalls the Master Argument's premise that "what is past is necessary". Diodorus Cronus was active a little after Aristotle, and he was a Megarian (see Dorion 1995 for criticism of David Sedley's attempt to reject this). It seems to me reasonable to conclude that Aristotle's target here is some Megarian argument, perhaps an earlier version of the Master.

§13: Glossary of Aristotelian Terminology

- Accept: *tithenai* (in a dialectical argument)
- Accepted: *endoxos* (also 'reputable' 'common belief')
- Accident: *sumbebêkos* (see *incidental*)
- Accidental: *kata sumbebêkos*
- Affirmation: *kataphasis*
- Affirmative: *kataphatikos*
- Assertion: *apophansis* (sentence with a truth value, declarative sentence)
- Assumption: *hupothesis*
- Belong: *huparchein*
- Category: *katêgoria* See the discussion in §C.
- Contradict: *antiphanai*
- Contradiction: *antiphasis* (in the sense "contradictory pair of propositions" and also in the sense "denial of a proposition")
- Contrary: *enantion*
- Deduction: *sullogismos*
- Definition: *horos*, *horismos*
- Demonstration: *apodeixis*
- Denial (of a proposition): *apophasis*
- Dialectic: *dialektikê* (the *art* of dialectic)
- Differentia: *diaphora*; specific difference, *eidopoios diaphora*.
- Direct: *deiktikos* (of proofs; opposed to "through the impossible")
- Essence: *to ti esti*, *to ti ên einai*
- Essential: *en tôi ti esti* (of predications)

- Extreme: *akron* (of the major and minor terms of a deduction)
- Figure: *schêma*
- Form: *eidos* (see also Species)
- Genus: *genos*
- Immediate: *amesos* ("without a middle")
- Impossible: *adunaton*; "through the impossible" (*dia tou adunatou*), of some proofs.
- Incidental: see Accidental
- Induction: *epagôgê*
- Middle, middle term (of a deduction): *meson*
- Negation (of a term): *apophasis*
- Objection: *enstasis*
- Particular: *en merei*, *epi meros* (of a proposition); *kath'hekaston* (of individuals)
- Peculiar, Peculiar Property: *idios, idion*
- Possible: *dunaton, endechomenon*; *endechesthai*(verb: "be possible")
- Predicate: *katêgorein* (verb); *katêegoroumenon*("what is predicated")
- Predication: *katêgoria* (act or instance of predicating, type of predication)
- Primary: *prôton*
- Principle: *archê* (starting point of a demonstration)
- Quality: *poion*
- Reduce, Reduction: *anagein, anagôgê*
- Refute: *elenchein*; refutation, *elenchos*
- Science: *epistêmê*
- Species: *eidos*
- Specific: *eidopoios* (of a differentia that "makes a species",*eidopoios diaphora*)
- Subject: *hupokeimenon*
- Substance: *ousia*
- Term: *horos*
- Universal: *katholou* (both of propositions and of individuals)

Bibliography

- Ackrill, J. L. 1961. *Aristotle's Categories and De Interpretatione*. Clarendon Aristotle Series. Oxford: Clarendon Press, 1961.
- Barnes, Jonathan. 1981. "Proof and the Syllogism". 17-59 in Berti, 1981.
- ----, trans. 1975, 1994. *Aristotle, Posterior Analytics* (translation with commentary). Clarendon Aristotle Series. Oxford: Clarendon Press. Second edition 1996.
- Becker, Albrecht. 1933. *Die Aristotelische Theorie der Möglichkeitsschlüsse*. Berlin: Junker und Dunnhaupt.
- Berti, Enrico, ed. 1981. *Aristotle on Science: The Posterior Analytics* Padua: Antenore.
- Broadie, Sarah [Waterlow]. 1982. *Passage and Possibility*. Oxford: Clarendon Press.
- Brunschwig, Jacques, ed. & trans. 1967. *Aristote, Topiques I-IV*. Paris.
- -----. 1984. 31-40 in "Aristotle on Arguments without Winners or Losers". *Wissenschaftskolleg* -

Jahrbuch 1984/85. Berlin.
- Burnyeat, Myles. 1981. "Aristotle on Understanding Knowledge". 97-139 in Berti 1981.
- Clark, Michael. 1980. *The Place of Syllogistic in Logical Theory*. Nottingham : Nottingham University Press.
- Code, Alan. 1986. "Aristotle's Investigation of a Basic Logical Principle: Which Science Investigates the Principle of Non-Contradiction?". *Canadian Journal of Philosophy* **16**: 341-358.
- Corcoran, John. 1972. "Completeness of an Ancient Logic". *Journal of Symbolic Logic* **37**: 696-705.
- ----. 1973. "A Mathematical Model of Aristotle's Syllogistic". *Archiv für Geschichte der Philosophie* **55**:191-219.
- Dancy, Russell. 1975. *Sense and Contradiction*. Dordrecht: D. Reidel,
- Dorion, Louis-André, tr. & comm. 1995. *Les Réfutations sophistiques*. Paris: J. Vrin.
- Ebert, Theodor. 1985. "Gattungen der Prädikate und Gattungen des Seienden bei Aristoteles: Zum Verhältnis von Kat. 4 und Top. I.9". *Archiv füur Geschichte der Philosophie* **67**: 113-138.
- Evans, J. D. G. 1977. *Aristotle's Concept of Dialectic*. Cambridge: Cambridge University Press,
- Ferejohn, Michael. 1980. *The Origins of Aristotelian Science*. New Haven: Yale University Press,
- Frede, Dorothea. 1970. *Aristoteles und die Seeschlacht*. Goettingen.
- Frede, Michael. 1975. "Stoic vs. Peripatetic Syllogistic". *Archiv für Geschichte der Philosophie* **56**.
- ----. 1981. "Categories in Aristotle.". 29-48 in M. Frede, *Essays in Ancient Philosophy* (University of Minnesota Press, 1987).
- Hambruch, Ernst. 1904. *Logische Regeln der Platonischen Schule in der Aristotelischen Topik*. Berlin: Weidemann.
- Hamlyn, D. W. 1990. "Aristotle on Dialectic". *Philosophy* **65**: 465-476.
- Hintikka, Jaakko. 1973. *Time and Necessity; Studies in Aristotle's Theory of Modality*. Oxford: Clarendon Press.
- Irwin, Terence. 1988. *Aristotle's First Principles*. Oxford: Clarendon Press.
- Johnson, Fred. "Apodictic Syllogisms: Deductions and Decision Procedures." *History and Philosophy of Logic* **16** (1994): 1-18.
- Le Blond, J. M. 1939. *Logique et méthode chez Aristote*. Paris: J. Vrin.
- Lear, Jonathan. 1980. *Aristotle and Logical Theory*. Cambridge University Press.
- Lukasiewicz, Jan. 1957. *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic*. Second ed. Oxford: Clarendon Press.
- ----. "Aristotle on the Principle of Non-Contradiction"
- McCall, Storrs. 1963. *Aristotle's Modal Syllogistic*. Amsterdam: North-Holland.
- McKirahan, Richard. 1992. *Principles and Proofs*. Princeton: Princeton University Press.
- Moraux, Paul. 1968. "La joute dialectique d'après le huitième livre des *Topiques*". In Owen 1968.
- Nortmann, Ulrich. 1996. *Modale Syllogismen, mögliche Welten, Essentialismus: eine Analyse der aristotelischen Modallogik.*. Berlin: De Gruyter.
- Nussbaum, Martha. 1986. *The Fragility of Goodness*. Cambridge: Cambridge University Press.
- Owen, G. E. L., 1961. "*Tithenai ta phainomena*." In S. Mansion, ed., *Aristote et les problèmes de méthode* (Louvain: Presses Universitaires de Louvain).
- -----, ed. 1968. *Aristotle on Dialectic: The Topics*. Proceedings of the Third Symposium

Aristotelicum. Cambridge: Cambridge University Press.

- Patterson, Richard. 1995. *Aristotle's Modal Logic: Essence and Entailment in the Organon*. Cambridge University Press.
- Patzig, Günther. 1969. *Aristotle's Theory of the Syllogism*. Tr. Jonathan Barnes. Dordrecht: D. Reidel.
- Primavesi, Oliver. 1996. *Die aristotelische Topik*. Munich: C. H. Beck.
- Ross, W. D., ed. 1951. *Aristotle's Prior and Posterior Analytics*. Oxford: Clarendon Press.
- Slomkowski, Paul. 1997. *Aristotle's Topics*. Leiden: Brill.
- Smiley, Timothy. 1974. "What Is a Syllogism?" *Journal of Philosophical Logic* **1**: 136-154.
- ----. 1994."Aristotle's Completeness Proof". *Ancient Philosophy*, special issue (1994):
- Smith, Robin, tr. & comm. 1989. *Aristotle's Prior Analytics*. Indianapolis: Hackett.
- ----. 1997. *Aristotle, Topics I, VIII, and Selections*. (Clarendon Aristotle Series). Oxford: Clarendon Press.
- Solmsen, Friedrich. 1929. *Die Entwicklung der aristotelischen Logik und Rhetorik*. Berlin.
- Striker, Gisela. 1985. "Notwendigkeit mit Lücken". *Neue Hefte für Philosophie* **24/25**: 146-164.
- ----. 1994. "Modal vs. Assertoric Syllogistic." *Ancient Philosophy* (special issue): 39-51.
- Thom, Paul. 1981. *The Syllogism*. Munich.
- ----. 1996. *The Logic of Essentialism: An Interpretation of Aristotle's Modal Syllogistic*. Dordrecht: Kluwer.
- Thomason, Steven K. 1993. "Semantic Analysis of the Modal Syllogistic". *Journal of Philosophical Logic* **22**: 111-128.
- van Rijen, Jeroen. 1989. *Aspects of Aristotle's Logic of Modalities*. Dordrecht: Reidel.
- Weidemann, Hermann, tr. & comm. 1994. *Aristoteles, Peri Hermeneias*. Berlin: Akademie Verlag.
- Whitaker, C. W. A. 1996. *Aristotle's De Interpretatione: Contradiction and Dialectic*. Oxford: Clarendon Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle: mathematics | Aristotle: metaphysics | Aristotle: poetics | Aristotle: rhetoric | Chrysippus | Diodorus Cronus | future contingents | logic: ancient | logic: relevance | Megarian School | square of opposition | Stoicism

Acknowledgements

I am indebted to Alan Code, Marc Cohen, and Theodor Ebert for helpful criticisms of earlier versions of

this article

Robin Smith
rasmith@aristotle.tamu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 18, 2000
Content last modified: October 5, 2000

Aristotle's Logic: Detailed Table of Contents

Return to Section 1: Introduction

The Traditional Square of Opposition

This entry traces the historical development of the Square of Opposition, a collection of logical relationships traditionally embodied in a square diagram. This body of doctrine provided a foundation for work in logic for over two millenia. For most of this history, logicians assumed that negative particular propositions ("Some S is not P") are vacuously true if their subjects are empty. This validates the logical laws embodied in the diagram, and preserves the doctrine against modern criticisms. Certain additional principles ("contraposition" and "obversion") were sometimes adopted along with the Square, and they genuinely yielded inconsistency. By the nineteenth century an inconsistent set of doctrines was widely adopted. Strawson's 1952 attempt to rehabilitate the Square does not apply to the traditional doctrine; it does salvage the nineteenth century version but at the cost of yielding inferences that lead from truth to falsity when strung together.

Introduction

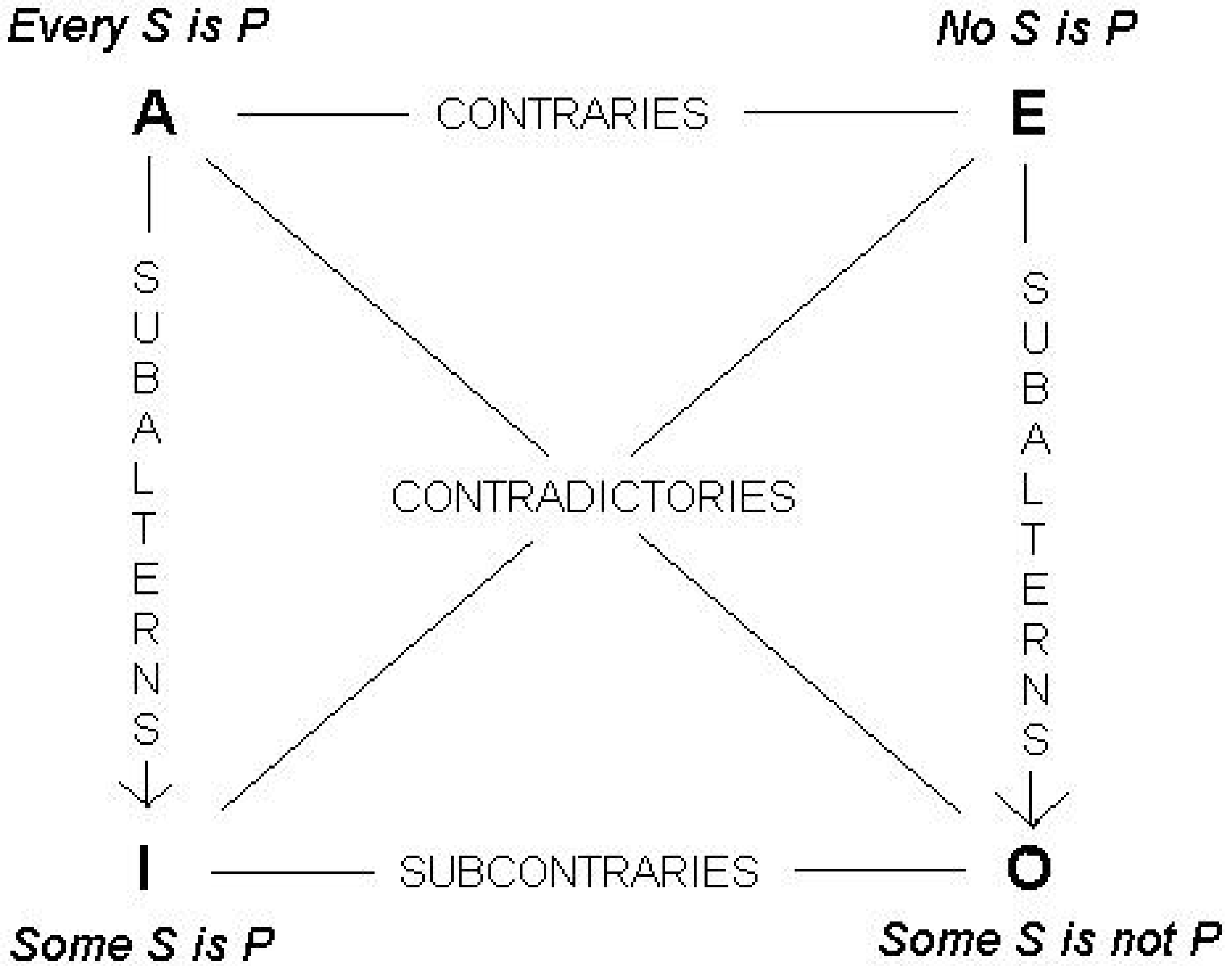
The doctrine of the square of opposition originated with Aristotle in the fourth century BC and has occurred in logic texts ever since. Although severely criticized in recent decades, it is still regularly referred to. The point of this entry is to trace its history from the vantage point of the end of the twentieth century, along with closely related doctrines bearing on empty terms.

The square of opposition is a group of theses embodied in a diagram. The diagram is not essential to the theses; it is just a useful way to keep them straight. The theses concern logical relations among four

logical forms:

NAME	*FORM*	*TITLE*
A	Every S is P	Universal Affirmative
E	No S is P	Universal Negative
I	Some S is P	Particular Affirmative
O	Some S is not P	Particular Negative

The diagram for the traditional square of opposition is:



The theses embodied in this diagram I call 'SQUARE'. They are:

SQUARE

- 'Every S is P' and 'Some S is not P' are contradictories.
- 'No S is P' and 'Some S is P' are contradictories.
- 'Every S is P' and 'No S is P' are contraries.

- ‘Some S is P’ and ‘Some S is not P’ are subcontraries.
- ‘Some S is P’ is a subaltern of ‘Every S is P’.
- ‘Some S is not P’ is a subaltern of ‘No S is P’.

These theses were supplemented with the following explanations:

- Two propositions are contradictory iff they cannot both be true and they cannot both be false.
- Two propositions are contraries iff they cannot both be true.
- Two propositions are subcontraries iff they cannot both be false.
- A proposition is a subaltern of another iff it must be true if its superaltern is true, and the superaltern must be false if the subaltern is false.

Probably nobody before the twentieth century every held exactly these views without holding certain closely linked ones as well. The most common closely linked view that is associated with the traditional diagram is that the E and I propositions *convert simply*; that is, ‘No S is P’ is equivalent in truth value to ‘No P is S’, and ‘Some S is P’ is equivalent in truth value to ‘Some P is S’. The traditional doctrine supplemented with simple conversion is a very natural view to discuss. It is Aristotle's view, and it was apparently endorsed (or at least not challenged) by everyone who wrote on this topic before the 19th century. I call this total body of doctrine ‘[SQUARE]’:

> [SQUARE] $=_{df}$ SQUARE + "the E and I forms convert simply"

where:

> A proposition *converts simply* iff it is necessarily equivalent in truth value to the proposition you get by interchanging its terms.

So [SQUARE] includes the relations illustrated in the diagram plus the view that ‘No S is P’ is equivalent to ‘No P is S’, and the view that ‘Some S is P’ is equivalent to ‘Some P is S’.

The Modern Revision of the Square

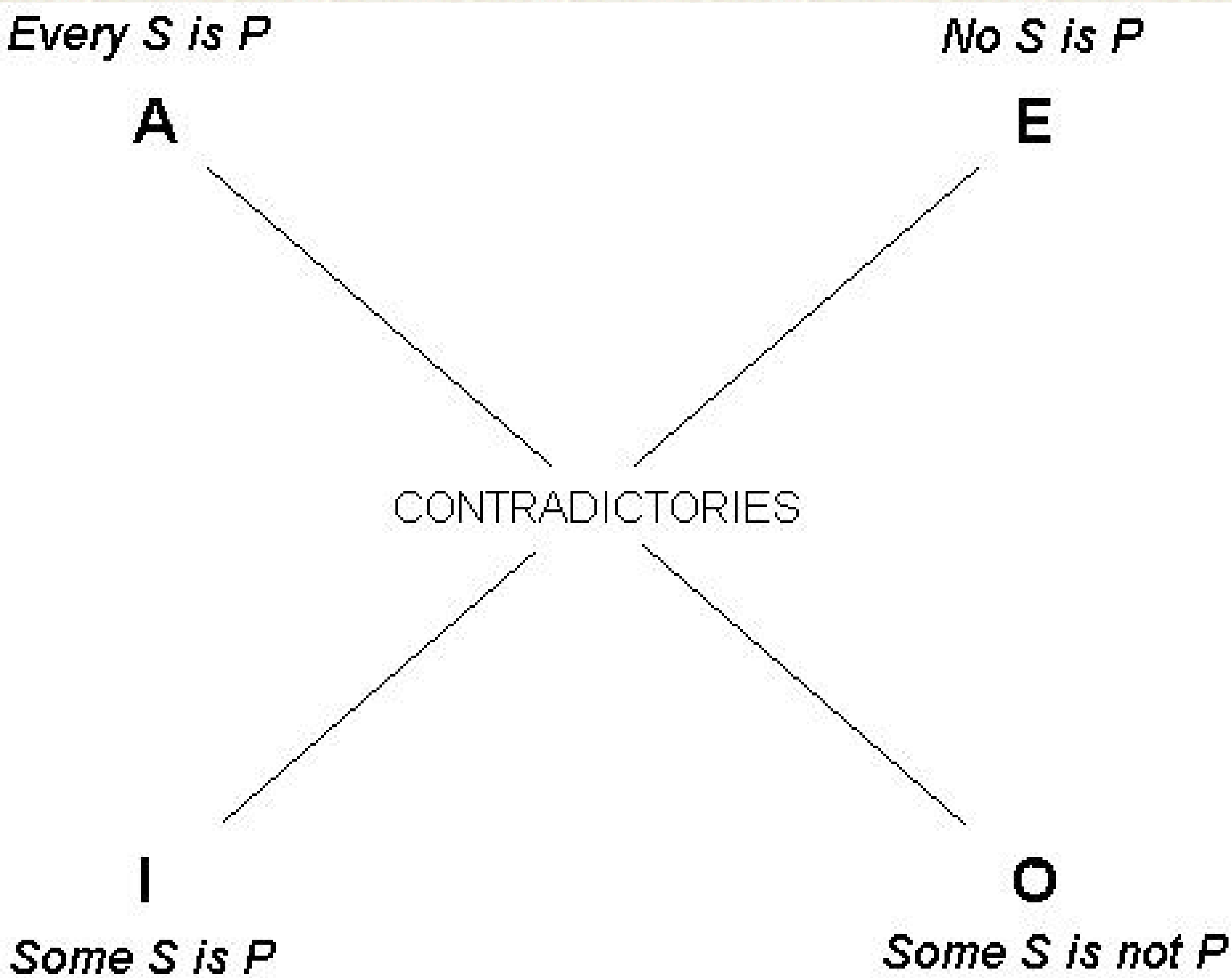
Most contemporary logic texts symbolize the traditional forms as follows:

Every S is P	$(x)(Sx \rightarrow Px)$
No S is P	$(x)(Sx \rightarrow -Px)$
Some S is P	$(\exists x)(Sx \& Px)$

Some S is not P (∃x)(Sx & -Px)

If this symbolization is adopted along with standard views about the logic of connectives and quantifiers, the relations embodied in the traditional square mostly disappear. The modern diagram looks like this:

THE MODERN REVISED SQUARE:



This has too little structure to be particularly useful, and so it is not commonly used. According to Alonzo Church, this modern view probably originated sometime in the late nineteenth century.[1] This representation of the four forms is now generally accepted, except for qualms about the loss of subalternation in the left-hand column. Most English speakers tend to understand ‘Every S is P’ as requiring for its truth that there be some S's, and if that requirement is imposed, then subalternation holds for affirmative propositions. Every modern logic text must address the apparent implausibility of letting ‘Every S is P’ be true when there are no S's. The common defense of this is usually that this is a logical notation devised for purposes of logic, and it does not claim to capture every nuance of the natural language forms that the symbols resemble. So perhaps ‘(x)(Sx → Px)’ does fail to do complete justice to ordinary usage of ‘Every S is P’, but this is not a problem with the logic. If you think that ‘Every S is P’ requires for its truth that there be S's, then you can have that result simply and easily: just represent the recalcitrant uses of ‘Every S is P’ in symbolic notation by adding an extra conjunct to the symbolization,

like this: (x)(Sx → Px) & (∃x) Sx.

This defense leaves logic intact and also meets the objection, which is not a logical objection, but merely a reservation about the representation of natural language.

Authors typically go on to explain that we often wish to make generalizations in science when we are unsure of whether or not they have instances, and sometimes even when we know they do not, and they sometimes use this as a defense of symbolizing the A form so as to allow it to be vacuously true. This is, however, an argument from convenience of notation, and does not bear on logical coherence.

The Argument Against the Traditional Square

Why does the traditional square need revising at all? The argument is a simple one:[2]

> Suppose that ‘S’ is an empty term; it is true of nothing. Then the I form: ‘Some S is P’ is false. But then its contradictory E form: ‘No S is P’ must be true. But then the subaltern O form: ‘Some S is not P’ must be true. But that is wrong, since there aren't any S's.

The puzzle about this argument is why the doctrine of the traditional square was maintained for well over 20 centuries in the face of this consideration. Were 20 centuries of logicians so obtuse as not to have noticed this apparently fatal flaw? Or is there some other explanation?

One possibility is that logicians previous to the 20th century must have thought that no terms are empty. You see this view referred to frequently as one that others held.[3] But with a few very special exceptions (discussed below) I have been unable to find anyone who held such a view before the nineteenth century. Many authors do not discuss empty terms, but those who do typically take their presence for granted. Explicitly rejecting empty terms was never a mainstream option, even in the nineteenth century.

In fact, the traditional doctrine of [SQUARE] is completely coherent in the presence of empty terms. This is because on the traditional interpretation, the O form lacks existential import. The O form is (vacuously) true if its subject term is empty, not false, and thus the logical interrelations of [SQUARE] are unobjectionable. In what follows, I trace the development of this view.

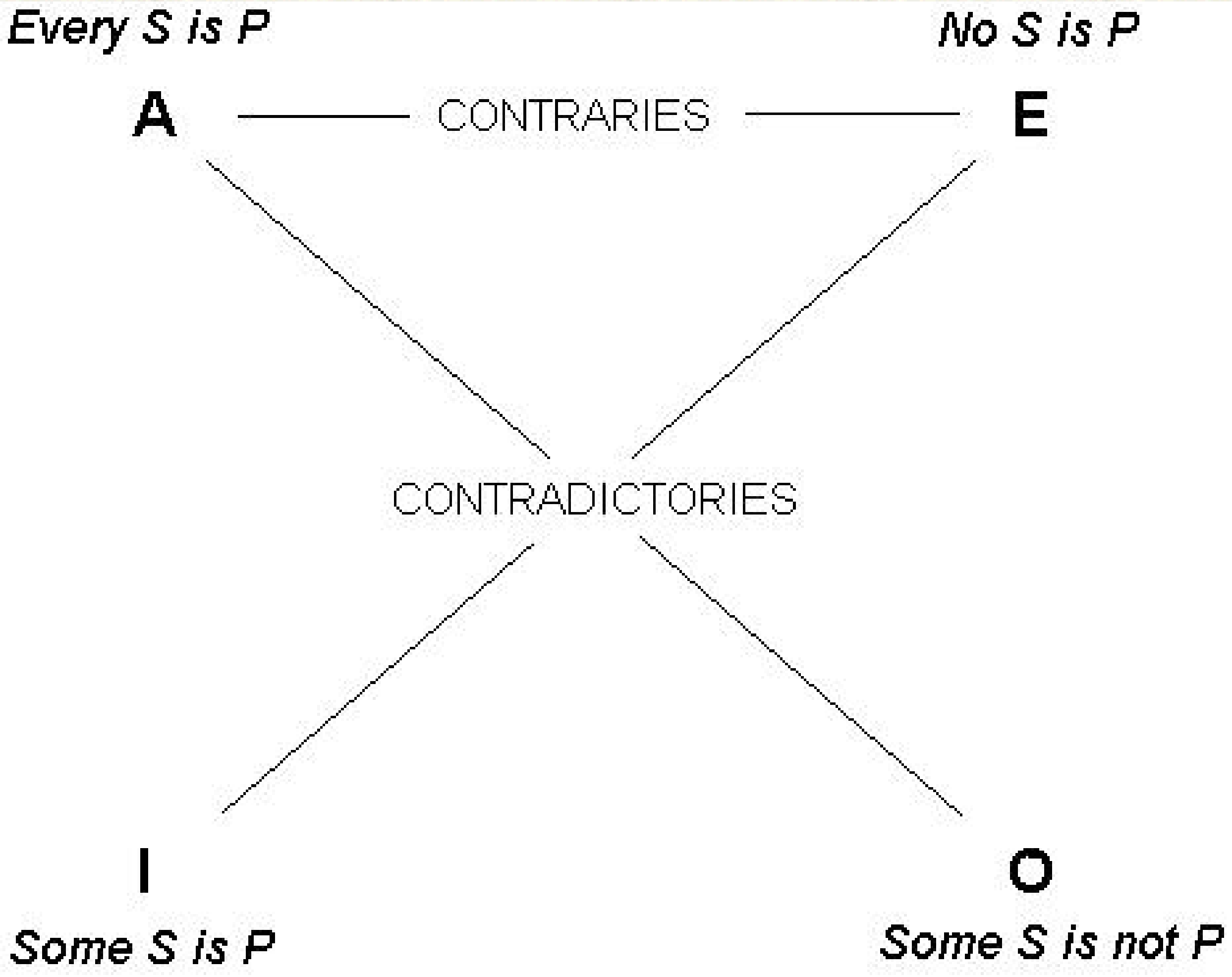
Origin of the Square of Opposition

The *doctrine* that I call [SQUARE], occurs in Aristotle. It begins in *De Interpretatione* 6-7, which contains three claims: that A and O are contradictories, that E and I are contradictories, and that A and E are contraries (17b.17-26):

> "I call an affirmation and a negation contradictory opposites when what one signifies universally the other signifies not universally, e.g. every man is white -- not every man is

> white, no man is white -- some man is white. But I call the universal affirmation and the universal negation contrary opposites, e.g. every man is just -- no man is just. So these cannot be true together, but their opposites may both be true with respect to the same thing, e.g. not every man is white -- some man is white."

This gives us the following fragment of the square:



But the rest is there by implication. For example, there is enough to show that I and O are subcontraries: they cannot both be false. For suppose that I is false. Then its contradictory, E, is true. So E's contrary, A, is false. So A's contradictory, O, is true. This refutes the possibility that I and O are both false, and thus fills in the bottom relation of subcontraries. Subalternation also follows. Suppose that the A form is true. Then its contrary E form must be false. But then the E form's contradictory, I, must be true. Thus if the A form is true, so must be the I form. A parallel argument establishes subalternation from I to O as well. The result is SQUARE.

In *Prior Analytics* I.2, 25a.1-25 we get the additional claims that the E and I propositions convert simply. Putting this together with the doctrine of *De Interpretatione* we have the full [SQUARE].[4]

The Diagram

The diagram accompanying and illustrating the doctrine shows up already in the second century A.D. Boethius incorporated it into his writing, and it passed down through the dark ages to the high medieval period, and from thence to today. Diagrams of this sort were popular among late classical and medieval authors, who used them for a variety of purposes. (Similar diagrams for modal propositions were especially popular.)

Aristotle's Formulation of the O Form

Ackrill's translation contains something a bit remarkable: Aristotle's articulation of the O form is *not* the familiar ‘Some S is not P’ or one of its variants; it is rather ‘Not every S is P’. With this wording, Aristotle's doctrine automatically escapes the modern criticism. (This holds for his views throughout *De Interpretatione*.[5]) For assume again that ‘S’ is an empty term, and suppose that this makes the I form ‘Some S is P’ false. Its contradictory, the E form: ‘No S is P’, is thus true, and this entails the O form in Aristotle's formulation: ‘Not every S is P’, which must therefore be true. When the O form was worded ‘Some S is not P’ this bothered us, but with it worded ‘Not every S is P’ it seems plainly right. Recall that we are granting that ‘Every S is P’ has existential import, and so if ‘S’ is empty the A form must be false. But then ‘Not every S is P’ *should* be true, as Aristotle's square requires.

On this view *affirmatives* have existential import, and *negatives* do not -- a point that became elevated to a general principle in late medieval times.[6] The ancients thus did not see the incoherence of the square as formulated by Aristotle because there was no incoherence to see.

The Rewording of the O Form

Aristotle's work was made available to the Latin west principally via Boethius's translations and commentaries, written a bit after 500 AD. In his translation of *De interpretatione*, Boethius preserves Aristotle's wording of the O form as "Not every man is white." But when Boethius comments on this text he illustrates Aristotle's doctrine with the now-famous diagram, and he uses the wording ‘Some man is not just’.[7] So this must have seemed to him to be a natural equivalent in Latin. It looks odd to us in English, but he wasn't bothered by it.

Early in the twelfth century Abelard objected to Boethius's rewording of the O form,[8] but Abelard's writing was not widely influential, and except for him and some of his followers people regularly used ‘Some S is not P’ for the O form in the diagram that represents the square. Did they allow the O form to be vacuously true? Perhaps we can get some clues to how medieval writers interpreted these forms by looking at other doctrines they endorsed. These are the theory of the syllogism and the doctrines of contraposition and obversion.

The (Ir)relevance of Syllogistic

One central concern of the Aristotelian tradition in logic is the theory of the categorical syllogism. This is the theory of two-premised arguments in which the premises and conclusion share three terms among them, with each proposition containing two of them. It is distinctive of this enterprise that everybody agrees on which syllogisms are valid. The theory of the syllogism partly constrains the interpretation of the forms. For example, it determines that the A form has existential import, at least if the O form does. For one of the valid patterns (Barbari[9]) is:

Every B is C

Every A is B

So, some A is C

This is invalid if the A form lacks existential import, and valid if it has existential import. It is held to be valid, and so we know how the A form is to be interpreted. One then naturally asks about the O form; what do the syllogisms tell us about it? The answer is that they tell us nothing beyond the fact that the O form is entailed by the E form. This lack of an answer is a fluke, resulting from the restrictions on what counts as a categorical syllogism; cases that would clearly decide the issue are not well-formed categorical syllogisms. At most, the theory confirms what we already know from SQUARE: that truth of the E form entails truth of the O form. And so if you are sure that the E form lacks such import, the O form must also lack it. Syllogistic gets us no further than this, though it does confirm that this body of theory nicely coheres with the lack of existential import of the O form.

The Principles of Contraposition and Obversion

One other piece of subject-matter bears on the interpretation of the O form. People were interested in Aristotle's discussion of "infinite" negation,[10] which is the use of negation to form a term from a term instead of a proposition from a proposition. In modern English we use ‘non’ for this; we make ‘non-horse’, which is true of exactly those things that are not horses. In medieval Latin ‘non-’ and ‘not’ are the same word, and so the distinction required special discussion. It became common to use infinite negation, and logicians pondered its logic. Some writers in the twelfth and thirteenth centuries adopted a principle called "conversion by contraposition." It states that

- ‘Every S is P’ is equivalent to ‘Every non-P is non-S’
- ‘Some S is not P’ is equivalent to ‘Some non-P is not non-S’

Unfortunately, this principle (which is not endorsed by Aristotle[11]) conflicts with the idea that there may be empty terms. For in the universal case it leads directly from the truth:

Every man is a being

to the falsehood:

Every non-being is a non-man

(which is false because the universal affirmative has existential import, and there are no non-beings). And in the particular case it leads from the truth (remember that the O form has no existential import):

A chimera is not a man.

to the falsehood:

A non-man is not a non-chimera.

These are Buridan's examples, used in the fourteenth century to show the invalidity of contraposition. Unfortunately, by Buridan's time the principle of contraposition had been advocated by a number of authors. The doctrine is already present in several twelfth century tracts,[12] and it is endorsed in the thirteenth century by Peter of Spain,[13] whose work was republished for centuries. William of Sherwood gives a counterexample to it, but then states confusingly that it holds in "every other case."[14] By the fourteenth century, problems associated with contraposition seem to be well-known, and authors generally cite the principle and note that it is not valid, but that it becomes valid with an additional assumption of existence of things falling under the subject term. For example, Paul of Venice in his eclectic and widely published *Logica Parva* from the end of the fourteenth century gives the traditional square with simple conversion[15] but rejects conversion by contraposition, essentially for Buridan's reason.

A similar thing happened with the principle of obversion. This is the principle that states that you can change a proposition from affirmative to negative, or vice versa, if you change the predicate term from finite to infinite (or infinite to finite). Some examples are:

Every S is P	=	No S is non-P
No S is P	=	Every S is non-P
Some S is P	=	Some S is not non-P
Some S is not P	=	Some S is non-P

Aristotle discussed some instances of obversion in *De Interpretatione*. It is apparent, given the truth conditions for the forms, that these inferences are valid when moving from affirmative to negative, but not in the reverse direction when the terms may be empty, as Buridan makes clear.[16] Some medieval writers before Buridan accepted the fallacious versions, and some did not.[17]

Later Developments

In Paul of Venice's other major work, the *Logica Magna* (*circa* 1400), he gives some pertinent examples of particular negative propositions that follow from true universal negatives. His examples of true particular negatives with patently empty subject terms are these:[18]

> *Some man who is a donkey is not a donkey.*
>
> *What is different from being is not.*
>
> *Some thing willed against by a chimera is not willed against by a chimera.*
>
> *A chimera does not exist.*
>
> *Some man whom a donkey has begotten is not his son.*

So by the end of the 14th century the problem of empty terms was clearly recognized. They were permitted in the theory, the O form definitely did not have existential import, and the logical theory, stripped of the incorrect special cases of contraposition and obversion, was coherent and immune to 20th century criticism.

Work on logic continued for the next couple of centuries, though most of it was lost and had little influence. But the topic of empty terms was squarely faced, and all solutions that were given within the Medieval tradition were consistent with [SQUARE]. I rely here on Ashworth 1974, 201-02, who reports the most common themes in the context of post-medieval discussions of contraposition. One theme is that contraposition is invalid when applied to universal or empty terms, for the sorts of reasons given by Buridan. The O form is explicitly held to lack existential import. A second theme, which Ashworth says was the most usual thing to say, is also found in Buridan: additional inferences, such as contraposition, become valid when supplemented by an additional premise asserting that the terms in question are non-empty.

An Oddity

There is one odd view that occurs at least twice, which may have as a consequence that there are no empty terms. In the thirteenth century, Lambert of Auxerre proposed that a term such as ‘chimera’ which stands for no existing thing must "revert to nonexistent things". So if we suppose that there no roses exist, then the term ‘rose’ stands for nonexistent things.[19] A related view also occurs much later; Ashworth reports that Menghus Blanchellus Faventinus held that term negations such as ‘nonman’ are true of non-beings, and he concluded from this that ‘A nonman is a chimera’ is true (apparently assuming that ‘chimera’ is also true of nonbeings).[20] However, neither of these views seems to have been clearly developed, and neither was widely adopted (or even widely discussed).[21] Nor is it even clear that either of them is supposed to have the consequence that there are no empty terms.

Modern, Renaissance, and Nineteenth Centuries

According to Ashworth,[22] serious and sophisticated investigation of logic ended at about the third decade of the sixteenth century. The *Port Royal Logic* of the following (seventeenth) century seems typical in its approach: its authors frequently suggest that logic is trivial and unimportant. Its doctrine includes that of the square of opposition, but the discussion of the O form is so vague that nobody could pin down its exact truth conditions, and there is certainly no awareness indicated of problems of existential import. This seems to typify popular texts for the next while. In the nineteenth century, the apparently most widely used textbook in Britain and America was Whately's *Elements of Logic*. Whately gives the traditional doctrine of the square, without any discussion of issues of existential import or of empty terms. He includes the problematic principles of contraposition (which he calls "conversion by negation"):

Every S is P	=	Every not-P is not-S
No S is P	=	No not-P is not-S

He also endorses obversion:[23]

- Some A is not B is equivalent to Some A is not-B, and thus it converts to Some not-B is A.

He says that this principle is "not found in Aldrich," but that it is "in frequent use."[24] This "frequent use" continued; later nineteenth and early twentieth century text books in England and America continued to endorse obversion (also called "infinitation" or "permutation"), and contraposition (also called "illative conversion").[25] This full nineteenth century tradition is consistent only on the assumption that empty (and universal) terms are prohibited, but authors seem unaware of this; Keynes 1928, 126, says generously "This assumption appears to have been made implicitly in the traditional treatment of logic." De Morgan is atypical in making the assumption explicit: in his 1847 text (64) he forbids universal terms (empty terms disappear by implication because if A is empty, non-A will be universal), and in his technical 1860 work he justifies this limitation on the grounds that universal terms are "needless" (16).[26]

In the twentieth century Lukasiewicz also developed a version of syllogistic that depends explicitly on the absence of empty terms; he attributed the system to Aristotle, thus helping to foster the tradition according to which the ancients were unaware of empty terms.

Today, logic texts divide between those based on contemporary logic and those from the Aristotelian tradition or the nineteenth century tradition, but even many texts that teach syllogistic teach it with the forms interpreted in the modern way, so that e.g. subalternation is lost. So the traditional square, as traditionally interpreted, is now mostly abandoned.

Strawson's Defense

In the twentieth century there were many creative uses of logical tools and techniques in reassessing past doctrines. One might naturally wonder if there is some ingenious interpretation of the square that attributes existential import to the O form *and* makes sense of it all without forbidding empty or universal terms, thus reconciling traditional doctrine with modern views. This is very close to what Strawson, 1952, 176-78, claims. Strawson's idea was to justify the square by adopting a nonclassical view of truth of statements, and by redefining the logical relation of validity. First, he suggested, we need to suppose that a proposition whose subject term is empty is neither true nor false, but lacks truth value altogether. Then we say that Q entails R just in case there are no instances of Q and R such that the instance of Q is true and the instance of R is false. For example, the A form ‘Every S is P’ entails the I form ‘Some S is P’ because there is no instance of the A form that is true when the corresponding instance of the I form is false. The troublesome cases involving empty terms turn out to be instances in which one or both forms lack truth value, and these are irrelevant so far as entailment is concerned. With this revised account of entailment, all of the "traditional" logical relations result, if they are worded as follows:

Contradictories:	The A and O forms entail each other's negations, as do the E and I forms. The negation of the A form entails the (unnegated) O form, and *vice versa*; likewise for the E and I forms.
Contraries: .	The A and E forms entail each other's negations
Subcontraries:	The negation of the I form entails the (unnegated) E form, and *vice versa*.
Subalternation:	The A form entails the I form, and the E form entails the O form.
Converses:	The E and I forms each entail their own converses.
Contraposition:	The A and O forms each entail their own contrapositives.
Obverses:	Each form entails its own obverse.

These doctrines are not, however, the doctrines of [SQUARE]. The doctrines of [SQUARE] are worded entirely in terms of the possibilities of truth values, not in terms of entailment. So "entailment" is irrelevant to [SQUARE]. It turns out that Strawson's revision of truth conditions *does* preserve the principles of SQUARE (these can easily be checked by cases),[27] but not the additional conversion principles of [SQUARE], and also not the traditional principles of contraposition or obversion. For example, Strawson's reinterpreted version of conversion holds for the I form because any I form proposition entails its own converse: if ‘Some A is B’ and ‘Some B is A’ both have truth value, then neither has an empty subject term, and so if neither lack truth value and if either is true the other will be true as well. But the original doctrine of conversion says that an I form and its converse always have the same truth value, and that is false on Strawson's account; if there are A's but no B's, then ‘Some A is B’ is false and ‘Some B is A’ has no truth value at all. Similar results follow for contraposition and obversion.

The "traditional logic" that Strawson discusses is much closer to that of nineteenth century logic texts than it is to the version that held sway for two millennia before that.[28] But even though he literally salvages a version of nineteenth century logic, the view he saves is unable to serve the purposes for which logical principles are formulated, as was pointed out by Timothy Smiley in a short note in *Mind* in 1967.[29] People have always taken the square to embody principles by which one can reason, and by which one can construct extended chains of reasoning. But if you string together Strawson's entailments you can infer falsehoods from truths, something that nobody in any tradition would consider legitimate. For example, begin with this truth (the subject term is non-empty):

> *No man is a chimera.*

By conversion, we get:

> *No chimera is a man.*

By obversion:

> *Every chimera is a non-man.*

By subalternation:

> *Some chimera is a non-man.*

By conversion:

> *Some non-man is a chimera.*

Since there are non-men, the conclusion is not truth-valueless, and since there are no chimeras it is false. Thus we have passed from a true claim to a false one. (The example does not even involve the problematic O form.) All steps are validated by Strawson's doctrine. So Strawson reaches his goal of preserving certain patterns commonly identified as constituting traditional logic, but at the cost of sacrificing the application of logic to extended reasoning.[30]

Bibliography

- Abelardus, Petrus. 11th-12th century. *Dialectica*. Ed. by L. M. de Rijk. Van Gorcum & Co., Assen, 1970.
- Aldrich, Henry. 1692. *Artis Logicæ Compendium*. E. Theatro Sheldoniano, Oxonii.
- Aristotle. 4th century B.C. *De Interpretatione* and *Prior Analytics* in Barnes, Jonathan (ed) *The Complete Works of Aristotle*. Princeton University Press, Princeton, 1984.
- Ashworth, E. J. 1974. *Logic and Language in the Post-Medieval Period*. Reidel, Dordrecht.

- Ashworth, E. J. 1978. "Existential Assumptions in Late Medieval Logic," *American Philosophical Quarterly* 10, 1978, 141-47.
- Brentano, Franz. 1874 *Psychologie vom Empirischen Standpunkte*, Dunker & Humbolt, Leipzig.
- Buridan, John. 14th century. *Tractatus de Suppositionibus*. In Reina, Maria Elena, "Giovanni Buridano: *Tractatus de Suppositionibus*," *Rivista critica di storia della filosofia* (1957), 175-208. Translated in King 1985.
- Buridan, John. 14th century. *Tractatus de Consequentiis*. In Hubien, Hubert, *Iohannis Buridani tractatus de consequentiis: Édition critique*, Volume XVI of *Philosophes médiévaux*, Publications universitaires, Louvain, 1976. Translated in King 1985.
- Burley, Walter. 14th century. "De Suppositionibus," in Brown, Stephen, "Walter Burleigh's Treatise *De Suppositionibus* and Its Influence on William of Ockham," *Franciscan Studies* 32, 1972, 15-64. Translated (part) in Spade 1997.
- Burley, Walter. 14th century. *Walter Burleigh: De puritate artis logicae tractatus longior, with a Revised Edition of the Tractatus brevior*, Philotheus Boehner (ed), The Franciscan Institute, St Bonaventure, NY, 1955. Translated in Spade [forthcoming].
- Cayley, Arthur. 1871 "Note on the Calculus of Logic." *The Quarterly Journal of Pure and Applied Mathematics* 11, 282-83.
- Church, Alonzo. 1965 "The History of the Question of Existential Import of Categorical Propositions," in Bar-Hillel, Yehoshua (ed.) *Logic, Methodology, and Philosophy of Science: Proceedings of the 1964 International Congress*. North-Holland, Amsterdam, 417-24.
- Coppée, Henry. 1882. *Elements of Logic*. American Book Co, New York.
- Davis, Noah. 1894. *Elements of Deductive Logic*. Harper, New York.
- De Rijk, L. M. 1967 *Logica Modernorum*, Volume II Part 2. Koninklijke Van Gorcum & Company N.V.; Assen, The Netherlands.
- De Morgan, Augustus. 1847. *Formal Logic*. Open Court, London.
- De Morgan, Augustus. 1860. *Syllabus of a Proposed System of Logic*. Reprinted in De Morgan, Augustus, *On the Syllogism and Other Logical Writings*. Yale University Press, New Haven, 1966.
- Dinneen, Francis P. 1990. *Peter of Spain: Language in Dispute*. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Freddoso, Alfred J, and Schuurman, Henry. 1980. *Ockham's Theory of Propositions: Part II of the Summa Logicae.* University of Notre Dame Press, Notre Dame.
- Jevons, W. Stanley. 1888. *Elementary Lessons in Logic*. Macmillan, London and New York.
- Joseph, H. W. B. 1916. *An Introduction to Logic*. Oxford University Press, London.
- Keynes, John Neville. 1928. *Studies and Exercises in Formal Logic*. Macmillan, London.
- King, Peter. 1985. *Jean Buridan's Logic: The Treatise on Supposition, The Treatise on Consequences*. D. Reidel, Dordrecht.
- Kneale, William and Kneale, Martha. 1962. *The Development of Logic*. Oxford University Press, Oxford.
- Kretzmann, Norman. 1966. *William of Sherwood's Introduction to Logic*. University of Minnesota Press, Minneapolis.
- Kretzmann, Norman. 1979. Pauli Veneti, *Logica Magna, Prima Pars: Tractatus de Terminis*. Oxford University Press, 1979.

- Kretzmann, Norman, Kenny, Anthony, and Pinborg, Jan. 1982. *The Cambridge history of Later Medieval Philosophy*. Cambridge University Press, Cambridge.
- Kretzmann, Norman and Stump, Eleonore. 1988. *The Cambridge Translations of Medieval Philosophical Texts*. Cambridge University Press, Cambridge.
- Londey, David and Johanson, Carmen. 1984. "Apuleius and the Square of Opposition," *Phronesis* 29, 165-73.
- Loux, Michael. 1974. *Ockham's Theory of Terms: Part I of the Summa Logicae*. University of Notre Dame Press, Notre Dame.
- Lukasiewicz, J. 1929. *Elementy Logiki Matematycznej*. Nakl. Komisji wydawniczeij Kola matematyczno-fizycznego sluchaczow uniwersytetu warszawskiego, Warsaw, and 1951 *Aristotle's Syllogistic*, Clarendon Press, Oxford.
- Ockham, William. 14th Century. *Summa Logicae*. Opera philosophica, vol. 2, The Franciscan Institute, St. Bonaventure, NY, 1974.
- Paul of Venice. 14th century. *Logica Parva* (1472 edition), Venice. Reprinted by Georg Olms Verlag, Hildesheim/New York, 1970. Translated in Perreiah 1984.
- Peirce, Charles. 1880. "On the Algebra of Logic." *American Journal of Mathematics* 3, 15-57.
- Perreiah, Alan. 1984. *Logica Parva: Translation of the 1472 Edition*. Philosophia Verlag, München.
- Peter of Spain. 13th century. *Tractatus: Summule Logicales*. De Rijk, L. M. (ed.). Van Gorcum, Assen, 1972. Translated in Dinneen 1990.
- Sellars, Roy Wood. 1925. *The Essentials of Logic*. Houghton Mifflin, New York.
- Smiley, Timothy. 1967. "Mr. Strawson on the Traditional Logic," *Mind* 76, 118-20.
- Spade, Paul Vincent. 1997. Translation of the beginning of Walter Burley's *Treatise on the Kinds of Supposition (De Suppositionibus)*, translated from Brown, Stephen, "Walter Burleigh's Treatise *De Suppositionibus* and Its Influence on William of Ockham," *Franciscan Studies* 32, 1972, 15-64.
- Spade, Paul Vincent. [forthcoming] *Walter Burley: on the Purity of the Art of Logic, the Shorter and the Longer Treatises.*, Yale University Press, New Haven. Translation of the second Burley item above.
- Strawson, Peter. 1952. *Introduction to Logical Theory*. Methuen, London.
- Wedin, Michael. 1990. "Negation and Quantification in Aristotle," *History and Philosophy of Logic* 11, 131-150.
- Whately, Richard. 1827. Reproduced in *Elements of Logic*, Scholar's Facsimiles & Reprints, Delmar, N.Y., 1975.
- William of Sherwood. 13th century. Charles H. Lohr, Peter Kunze and Bernhard Mussler, ed., "William of Sherwood, 'Introductiones in logicam': Critical Text," *Traditio* 39 (1983), 219-299. An earlier edition is translated in Kretzmann 1966: *Introductiones in logicam*. Grabman, Martin (ed.). *Sitzungsberichte der Bayersichen Akademie der Wissenschaften*, Philosophisch-historische Abteilung, Jahrgang 1937, H. 10. Munich, 1937.

Other Internet Resources

- Mediaeval Logic and Philosophy

Related Entries

logic: history of

Terence Parsons
tparsons@ucla.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 8, 1997
Content last modified: March 10, 1999

Stanford Encyclopedia of Philosophy
Notes to The Square of Opposition

Notes

1. Church 1965, 422 finds it first (implicitly) in Cayley 1871, explicitly in Brentano 1874, and in Peirce 1880, 15-57.

2. This argument is given e.g. in Kneale & Kneale 1962, 55-60.

3. Cf. Kneale & Kneale 1962, 60, and Keynes 1928, 126 note 1.

4. *Prior Analytics* also contains the doctrine of conversion *per accidens*: that you can interchange the subject and predicate terms of either universal form if you also turn it into a particular. This is not an additional principle, since it follows from the rules of subalternation together with simple conversion.

5. Cf. Wedin 1990.

6. Ockham (SL I.72): "In affirmative propositions a term is always asserted to supposit for something. Thus, if it supposits for nothing the proposition is false. However, in negative propositions the assertion is either that the term does not supposit for something or that it supposits for something of which the predicate is truly denied. Thus a negative proposition has two causes of truth." Loux 1974, 206.

7. Paul Spade called these facts to my attention.

8. Abelardus Petrus 1970, 186: "And thus in categorical propositions, the only proper and rightly destroying contradiction to any affirmation seems to be a proposition that with a negation appended in front destroys the entire meaning, so that ‘not every human is human’, but not ‘some human is not human’ is the contradictory of ‘every human is human’. For the latter may happen to be false simultaneously with that. For a thing of ‘human’ not existing now, neither will this be true, namely ‘every human is human’ nor that which says ‘some human is not human’ . . ."

9. Oddly, Aristotle does not mention this form, though it was accepted from the medieval period onward. (It is common to speculate that this form was too trivial to mention.) The same point can be made with a form he did discuss (Darapti): If every A is C and every A is B then some B is C.

10. Aristotle, *De Interpretatione* 2 and 3.

11. Church 1965, 419 quotes some apparent instances of contraposition endorsed by Aristotle. But the wordings he gives are not actually contraposition; their premises are apparently meant to be universal affirmatives (Every A is B) and their conclusions seem to be of a "nonstandard" negative form: Every non-B is not A. (Such conditionals, though not their converses, are valid.) These examples, however, might easily have been interpreted as contraposition by later commentators, just as Church did.

12. Contraposition is discussed in several twelfth and early thirteenth century anonymous texts edited in De Rijk 1967. It is endorsed in *Excerpta Norimbergenses* (138-39), *Ars Burana* (190), *Introductiones Parisienses* (362), *Logica "Ut dicit"* (385), *Logica "Cum sit nostra"* (426), *Dialectica Monacensis* (478). One text, *Tractatus Anagnini* (238) states contrapositions without making clear whether they are endorsed or not, though the context seems to suggest acceptance. But they are illustrated with what others took to be a counterexample to contraposition: "Every non-animal is a non-phoenix; so every phoenix is an animal". Another text, *Ars Emmerana* (157) endorses contraposition, but then says that it does not hold for the particular negative unless understood with "constancy," a term that had to do with assuming instances of the terms in question. One other text, *Introductiones Montane Minores* (34-35) straightforwardly objects to contraposition because of how it handles empty terms.

13. I.15 in Dineen 1990, 8.

14. 3.3 in Kretzmann 1966, 59.

15. 1.8-11 in Perreiah 1984, 127-130.

16. "From any affirmative there follows a negative by changing the predicate according to finite and infinite, keeping the rest the same, but there is no formal consequence from a negative to an affirmative, although there is a consequence under the assumption that all of the terms supposit for something." [*Tractatus de Consequentiis* I.8.107: King 1985, 226.]

The good direction he gives as:

> Every B is A; therefore no B is non-A.

The fallacious direction is illustrated by

> A chimera is not a man; therefore a chimera is a non-man.

17. Obversion is discussed in several twelfth and early thirteenth century anonymous texts edited in De Rijk 1967, where it is seen as a type of *equipollence*. It is endorsed in four of these: *Excerpta Norimbergenses* (138-39), *Logica "Ut dicit"* (385), *Logica "Cum sit nostra"* (426), *Dialectica Monacensis* (478). Three texts, *Ars Burana*, *Introductiones Parisienses*, *Tractatus Anagnini* omit it (while including a discussion of contraposition). One text, *Introductiones Montane Minores* (37-38)

objects to it because it mishandles empty terms.

18. Chapter 4 in Kretzmann 1979, 233.

19. Kretzmann & Stump 1988, 118.

20. Ashworth 1974, 201-02.

21. These odd views should not be confused with a widely adopted theory which held that in certain circumstances a term is "ampliated" so as to stand for not presently existing things. For example, the past tense of ‘A donkey ran’ makes ‘donkey’ stand for both present and past donkeys, the ‘can’ in ‘A rose can be red’ makes ‘rose’ stand for both actual and possible roses, and the special nature of the predicate in ‘A chimera is imaginable’ makes ‘chimera’ stand for (impossible but) merely intelligible chimeras. But this widely held doctrine does not apply to ‘A chimera is an animal’ since there is no "cause" of ampliation in that sentence; so the doctrine does not prohibit empty terms.

22. Ashworth 1978, 147.

23. Whately 1826, 84-85.

24. Whately 1827, 85. He apparently refers there to an earlier edition of Aldrich 1849.

25. Coppée 1882, 76, Jevons, 1888, 83-5, Davis, 1894, 91, Joseph, 1916, 237-38, R. W. Sellars, 1925, 107.

26. The system of logic including obversion and contraposition is consistent if empty and universal terms are forbidden; see Miller 1938. Miller states that if the system is supplemented by all of the traditional nineteenth century rules for the syllogism, including rules of distribution, the system is inconsistent. But his proof involves applying the rules of distribution to syllogisms that are not in standard form. Miller thinks that authors intended this, but I see no evidence of it.

27. The wording of SQUARE does not rule out the possibility of truth-valueless sentences -- so Strawson's view that empty subject terms lead to lack of truth value does not conflict with SQUARE. However, hardly anyone exploited the possibility of sentences without truth value. A few medieval writers appeared to have tried to solve the semantic paradoxes by holding that paradoxical sentences "say nothing", but the idea was not taken seriously by most theorists. (Cf. Spade's summary in Kretzmann, Kenny, and Pinborg 1982, 245-6.) Nor does it occur in Renaissance or Modern or nineteenth century texts.

Paul Spade points out (personal communication) that a very few authors held that future contingent

propositions are now without truth-value (and will stay that way until the future event comes out one way or the other). The most prominent mediaeval author to hold this was the Frenchman Peter Auriol (early 14th century). But the Englishman Roger Swyneshed (fl. 1330s) also appears to accept it in an offhand comment in one passage. Many authors (e.g., Ockham) thought it was Aristotle's view of future contingents (*De interpretation*, ch. 7), but did not accept it themselves.

28. Strawson 1952, 152 calls what he is discussing "traditional formal logic." His source is Miller 1938, which bears the title *The Structure of Aristotelian Logic*, which might lead one to think he is discussing the ancient and medieval doctrine. But Miller makes explicit (pages 11-12) that the subject matter under discussion is the version of logic that existed when he wrote: in particular, "the system of deductive logic expounded in the principal introductory textbooks of logic; for example, the well-known manuals by Whately, Jevons, Joseph, Wolf, Creighton, Hibbons, and Sellars." Miller's logical work is excellent, but his historical remarks should not be trusted. For example, he sees the use of the notion of distribution in testing for the validity of syllogisms as part of a longstanding tradition (6), followed by the introduction of negative terms in the nineteenth century (3). In fact negative terms were in the original theory (they were in Aristotle), and the use of rules of distribution to test for validity were probably first used seventeen centuries later (by Buridan in the fourteenth century; perhaps earlier, but certainly no earlier than the twelfth century).

29. Smiley 1967. Church 1965 also criticizes Strawson's proposal, but without actually saying what is wrong with it.

30. This entry is condensed from a longer version to appear in *Acta Analytica*.

Terence Parsons
tparsons@ucla.edu

First published: August 8, 1997
Content last modified: August 8, 1997

Stanford Encyclopedia of Philosophy
Notes to Aristotle's Metaphysics

Notes

1. This crucial idea is put forward at *Posterior Analytics* 71b32; *Prior Analytics* 68b35-7; *Physics* A.1, 184a16-20; *Metaphysics* Z.3, 1029b3-12; *Topics* Z.4, 141b2-142a12.)

2. This inverse tree-like structure was first noticed in the 3rd century C.E. by Porphyry: “Substance is itself a genus, under this is body, and under body is living body, under which is animal. Under animal is rational animal, under which is man. Under man are Socrates and Plato and individual (*kata meros*) men” (*Isagoge* 4, 21-25). This so-called “tree of Porphyry” later found its way, with illustrations, into medieval discussions of Aristotle.

3. Although Aristotle himself never puts it this way, one might think of each category itself as a genus. This is certainly what Porphyry thought (see note 2). See also Owen 1965b. Note, however, that if a category is a genus, it is a maximally general one — it cannot be a species of some higher genus. For the union of all the categories contains everything that there is — i.e., all of the beings — and Aristotle insists that being is not a genus (*Posterior Analytics* 92b14, *Metaphysics* B.3, 998b22).

S. Marc Cohen
smcohen@u.washington.edu

First posted: October 8, 2000
Last modified: September 29, 2001

Aristotle's Rhetoric

Aristotle's rhetoric has had an enormous influence on the development of the art of rhetoric. Not only authors writing in the peripatetic tradition, but also the famous Roman teachers of rhetoric, such as Cicero and Quintilian, frequently used elements stemming from the Aristotelian doctrine. Nevertheless, these authors were neither interested in an authentic interpretation of the Aristotelian works nor in the philosophical sources and backgrounds of the vocabulary that Aristotle had introduced into rhetorical theory. Thus, for two millennia the interpretation of Aristotelian rhetoric has become a matter of the history of rhetoric, not of philosophy. In the most influential manuscripts and editions, Aristotle's *Rhetoric* was surrounded by rhetorical works and even written speeches of other Greek and Latin authors, and was seldom interpreted in the context of the whole Corpus Aristotelicum. It was not until the last few decades that the philosophically salient features of the Aristotelian rhetoric were rediscovered: in construing a general theory of the persuasive, Aristotle applies numerous concepts and arguments which are also treated in his logical, ethical, and psychological writings. His theory of rhetorical arguments, for example, is only one further application of his general doctrine of the *sullogismos*, which also forms the basis of dialectic, logic and his theory of demonstration. Another example is the concept of emotions: though emotions are one of the most important topics in the Aristotelian ethics, he nowhere offers such an illuminating account of single emotions as in the *Rhetoric*. Finally, it is the *Rhetoric* too which informs us about the cognitive features of language and style.

1. Works on Rhetoric

According to ancient testimonies, Aristotle wrote an early dialogue on rhetoric entitled '*Grullos*', in which he put forward the argument that rhetoric cannot be an art (*technê*); and since this is precisely the position of Plato's *Gorgias*, the lost dialogue *Grullos* has traditionally been regarded as a sign of Aristotle's (alleged) early Platonism. But the evidence for the position of this dialogue is too tenuous to support such strong conclusions: it also could have been a 'dialectical' dialogue, which listed the Pros and Cons of the thesis that rhetoric is an art. We do not know much more about the so-called '*Technê Sunagogê*', a collection of previous theories of rhetoric which is also ascribed to Aristotle. Cicero seems to use this collection itself, or at least a secondary source relying on it, as his main historical source when he gives a short survey of the history of pre-Aristotelian rhetoric in his *Brutus* 46-48. Finally, Aristotle once mentions a work called '*Theodecteia*' which has also been supposed to be Aristotelian; but more probably he meant the rhetorical handbook of his follower Theodectes, who was a former pupil of Isocrates.

What has come down to us are just the three books on rhetoric, which we know as *The Rhetoric*, though the ancient catalogue of the Aristotelian works, reported by Diogenes Laertius, mentions only two books on rhetoric (perhaps our *Rhetoric* I & II), and two further books on style (perhaps our *Rhetoric* III?). Whereas most modern authors agree that at least the core of *Rhet.* I & II presents a coherent rhetorical theory, the two themes of *Rhet.* III are not mentioned in the agenda of *Rhet.* I & II. The conceptual link between *Rhet.* I & II and *Rhet.* III is not given until the very last sentence of the second book. It is quite understandable that the authenticity of this *ad hoc* composition has been questioned: we cannot exclude the possibility that these two parts of the *Rhetoric* had not been put together until the first edition of Aristotle's works completed by Andronicus in the first century. In the *Poetics* (1456a33) we find a cross-reference to a work called '*Rhetoric*' which obviously refers to *Rhet.* I & II, but excludes *Rhet.* III.

Regardless of such doubts, the systematic idea which links the two heterogeneous parts of the *Rhetoric* together does not at all seem to be unreasonable : it is not enough to have a supply of things to say (the so-called “thought”), the theorist of rhetoric must also inform us about the right way to say those things (the so-called “style”).

The chronological fixing of the *Rhetoric* has turned out to be a delicate matter. At least the core of *Rhet.* I & II seems to be an early work, written during Aristotle's first stay in Athens (it is unclear, however, which chapters belong to that core; regularly mentioned are the chapters I.4-15 and II.1-17). It is true that the *Rhetoric* gives references to historical events which fall in the time of Aristotle's exile and his second stay in Athens, but most of them can be found in the chapters II.23-24, and besides this, examples could have been updated, which is especially plausible if we assume that the *Rhetoric* formed the basis of a lecture held several times. Most striking are the affinities to the (also early) *Topics*; if, as it is widely agreed, the *Topics* represents a pre-syllogistic state of Aristotelian logic, the same is true of the *Rhetoric*: we actually find no hints of syllogistic inventory in it.

2. The Agenda of the *Rhetoric*

The structure of *Rhet.* I & II is determined by two tripartite divisions. The first division consists in the distinction of the three means of persuasion: The speech can produce persuasion either through the character of the speaker, the emotional state of the listener, or the argument (*logos*) itself (see below §5). The second tripartite division concerns the three species of public speech. The speech that takes place in the assembly is defined as the deliberative species. In this rhetorical species the speaker either advises the audience to do something or warns against doing something. Accordingly, the audience has to judge things that are going to happen in the future, and they have to decide whether these future events are good or bad for the polis, whether they will cause advantage or harm. The speech that takes place before a court is defined as the judicial species. The speaker either accuses somebody or defends herself or someone else. Naturally, this kind of speech treats things that happened in the past. The audience or rather jury has to judge whether a past event was just or unjust, i.e. whether it was according to the law or contrary to the law. While the deliberative and judicial species have their context in a controversial situation in which the listener has to decide in favor of one of two opposing parties, the third species does not aim at such a decision: the epideictic speech praises or blames somebody, it tries to describe things or deeds of the respective person as honorable or shameful.

The first book of the *Rhetoric* treats the three species in succession. *Rhet.* I.4-8 deals with the deliberative, I.9 with the epideictic, I.10-14 the judicial species. These chapters are understood as contributing to the argumentative mode of persuasion or—more precisely—to that part of argumentative persuasion which is specific to the several species of persuasion. The second part of the argumentative persuasion which is common to all three species of rhetorical speech is treated in the chapters II.19-26. The second means of persuasion, which works by evoking the emotions of the audience, is described in the chapters II.2-11. Though the following chapters II.12-17 treat of different types of character these chapters do not, as is often assumed, develop the third means of persuasion, which depends on the character of the speaker. The underlying theory of this means of persuasion is elaborated in a few lines of chapter II.1. The

aforementioned chapters II.12-17 give information about different types of character and their disposition to emotional response, which can be useful for those speakers who want to arouse the emotions of the audience. Why the chapters on the argumentative means of persuasion are separated by the treatment of emotions and character (in II.2-17) remains a riddle, especially since the chapter II.18 tries to give a link between the specific and the common aspects of argumentative persuasion. *Rhetoric* III.1-12 discusses several questions of style (see below §8.1), Rhetoric III.13-19 is on the several parts of a speech.

3. Rhetoric as a Counterpart to Dialectic

Aristotle stresses that rhetoric is closely related to dialectic. He offers several formulas to describe this affinity between the two disciplines: first of all, rhetoric is said to be a “counterpart” (*antistrophos*) to dialectic (*Rhet.* I.1, 1354a1); (ii) it is also called an “outgrowth” (*paraphues ti*) of dialectic and the study of character (*Rhet.* I.2, 1356a25f.); finally, Aristotle says that rhetoric is part of dialectic and resembles it (*Rhet.* I.2, 1356a30f.). In saying that rhetoric is a counterpart to dialectic Aristotle obviously alludes to Plato's *Gorgias* (464bff.) where rhetoric is ironically defined as a counterpart to cookery in the soul. Since, in this passage, Plato uses the word ‘*antistrophos*’ to designate an analogy, it is likely that Aristotle wants to express a kind of analogy too: what dialectic is for the (private or academic) practice of attacking and maintaining an argument, rhetoric is for the (public) practice of defending oneself or accusing an opponent.

This analogy between rhetoric and dialectic can be substantiated by several common features of both disciplines:

- Rhetoric and dialectic are concerned with things that do not belong to a definite genus or are not the object of a specific science.
- Rhetoric and dialectic rely on accepted sentences (*endoxa*).
- Rhetoric and dialectic are not dependent on the principles of certain sciences.
- Rhetoric and dialectic are concerned with both sides of an opposition.
- Rhetoric and dialectic rely on the same theory of deduction and induction.
- Rhetoric and dialectic similarly apply the so-called *topoi*.

The analogy to dialectic has important implications for the status of rhetoric. Plato argued in his *Gorgias* that rhetoric cannot be an art (*technê*), since it is not related to a definite subject, while real arts are defined by their specific subjects, as e.g. medicine or shoemaking are defined by their products, i.e. health and shoes. However, though dialectic has no definite subject, it is easy to see that it nevertheless rests on a method, because dialectic has to grasp the reason why some arguments are valid and others are not. Now, if rhetoric is nothing but the counterpart to dialectic in the domain of public speech, it must be grounded on an investigation of what is persuasive and what is not, and this, in turn, qualifies rhetoric as an art.

Further, it is central for both disciplines that they deal with arguments from accepted premises. Hence the rhetorician who wants to persuade by arguments or (rhetorical) proofs can adapt most of the dialectical equipment. Nevertheless, persuasion which takes place before a public audience is *not only* a matter of

arguments and proofs, but also of credibility and emotional attitudes. This is why there are remarkable differences between the two disciplines too:

- Dialectic can be applied to every object whatsoever, rhetoric is useful especially in practical and public matters.
- Dialectic proceeds by questioning and answering, while rhetoric for the most part proceeds in continuous form.
- Dialectic is concerned with general questions, while rhetoric is concerned for the most part with particular topics (i.e. things about which we cannot gain real knowledge).
- Certain uses of dialectic apply qualified *endoxa*, i.e. *endoxa* which are approved by experts, while rhetoric aims at *endoxa* which are popular.
- Rhetoric must take into account that its target group has only restricted intellectual resources, whereas such concerns are totally absent from dialectic.
- While dialectic tries to test the consistency of a set of sentences, rhetoric tries to achieve the persuasion of a given audience.
- Non-argumentative methods are absent from dialectic, while rhetoric uses non-argumentative means of persuasion.

4. The Purpose of Rhetoric

4.1 The Definition of Rhetoric

Aristotle defines the rhetorician as someone who is always able to see what is persuasive (*Topics* VI.12, 149b25). Correspondingly, rhetoric is defined as the ability to see what is possibly persuasive in every given case (*Rhet.* I.2, 1355b26f.). This is not to say that the rhetorician will be able to convince under all circumstances. Rather he is in a similar situation as the physician: the latter has a complete grasp of his art only if he neglects nothing which might heal his patient, though he is not able to heal *every* patient. Similarly, the rhetorician has a complete grasp of his method, if he discovers the available means of persuasion, though he is not able to convince *everybody*.

4.2 The Neutrality of Aristotelian Rhetoric

Aristotelian rhetoric as such is a neutral tool that can be used by persons of virtuous or depraved character. This capacity can be used for good or bad purposes, it can cause great benefits as well as great harms. There is no doubt that Aristotle himself regards his system of rhetoric as something useful, but the good purposes for which rhetoric is useful do not define the rhetorical capacity as such. Thus, Aristotle does not hesitate to concede on the one hand that his art of rhetoric can be misused. But on the other hand he tones down the risk of misuse by stressing several factors: Generally, it is true of all goods, except virtue, that they can be misused. Secondly, using rhetoric of the Aristotelian style it is easier to convince of the just and good than of their opposites. Finally, the risk of misuse is compensated by the benefits which can be accomplished by rhetoric of the Aristotelian style.

4.3 Why We Need Rhetoric

It could still be objected that rhetoric is only useful for those who want to outwit their audience and conceal their real aims, since someone who just wants to communicate the truth could be straightforward and would not need rhetorical tools. This, however, is not Aristotle's point of view: Even those who just try to establish what is just and true need the help of rhetoric when they are faced with a public audience. Aristotle tells us, that it is impossible to teach such an audience, even if the speaker had the most exact knowledge of the subject. Obviously he thinks that the audience of a public speech consists of ordinary people who are not able to follow an exact proof based on the principles of a science. Further, such an audience can easily be distracted by factors which do not pertain to the subject at all; sometimes they are receptive to flattery or just try to increase their own advantage. And this situation even becomes worse if the constitution, the laws, and the rhetorical habits in a city are bad. Finally, most of the topics that are usually discussed in public speeches do not allow of exact knowledge, but leave room for doubt; especially in such cases it is important that the speaker seems to be a credible person and that the audience is in a sympathetic mood. For all those reasons it is a matter of persuasiveness, not of knowledge, to affect the decisions of juries and assemblies. It is true that some people manage to be persuasive either at random or by habit, but it is rhetoric which gives us a method to discover *all* means of persuasion on *any* topic whatsoever.

4.4 Aristotelian Rhetoric as Proof-centered and Pertinent

Aristotle joins Plato in criticizing contemporary manuals of rhetoric. But how does he manage to distinguish his own project over and against the criticized manuals? The general idea seems to be this: Previous theorists of rhetoric gave most of their attention to methods outside the subject; they taught how to slander, how to arouse emotions in the audience, or how to distract the attention of the hearers from the subject. This style of rhetoric promotes a situation in which juries and assemblies no longer form rational judgments about the given issues but surrender to the litigants. Aristotelian rhetoric is different in this respect: it is centered around the rhetorical kind of proof, the enthymeme (see below §6), which is called the most important means of persuasion. Since people are most strongly convinced when they suppose that something has been proven (*Rhet.* I.1, 1355a5f.), there is no need for the orator to confuse or distract the audience by the use of emotional appeals etc. In Aristotle's view an orator will be even more successful when he just picks up the convincing aspects of a given issue, thereby using commonly held opinions as premises. Since people have a natural disposition for the true (*Rhet.* I.1, 1355a15f.) and every man has some contribution to make to the truth (*Eudemian Ethics* I.6, 1216b31) there is no unbridgeable gap between the commonly held opinions and what is true. This alleged affinity between the true and the persuasive justifies Aristotle's project of a rhetoric which essentially relies on the persuasiveness of pertinent argumentation; and it is just this argumentative character of Aristotelian rhetoric that explains the close affinity between rhetoric and dialectic (see above §3).

4.5 Is There an Inconsistency in Aristotle's Rhetorical Theory?

Of course, Aristotle's rhetoric covers non-argumentative tools of persuasion as well. He tells the orator how to stimulate emotions and how to make himself credible (see below §5); his art of rhetoric includes considerations about delivery and style (see below §8.1) and the parts of a speech. It is understandable that several interpreters found an insoluble tension between the argumentative means of pertinent rhetoric and non-argumentative tools which aims at what is outside the subject. It does not seem, however, that Aristotle himself saw a major conflict between these diverse tools of persuasion. Presumably, for the following reasons: (i) He leaves no doubt that the subject that is treated in a speech has the highest priority (e.g. *Rhet.* III.1, 1403b18-27). Thus, it is not surprising that there are even passages which regard the non-argumentative tools as a sort of accidental contribution to the process of persuasion which essentially proceeds in the manner of dialectic (cp. *Rhet.* I.1, 1354a15). (ii) There are, he says (III.1, 1404a2f.) methods which are not right, but necessary because of certain deficiencies of the audience. His point seems to be that the argumentative method becomes less effective, the worse the condition of the audience is. This again is to say that it is due to the badness of the audience when his rhetoric includes aspects which are not in line with the idea of argumentative and pertinent rhetoric. (iii) In dealing with methods of traditional rhetoric Aristotle obviously assumes that even methods which have traditionally been used instead of argumentation can be refined so that they support the aim of an argumentative style of rhetoric. The prologue of a speech, for example, was traditionally used for appeals to the hearer, but it can also be used to set out the issue of the speech, thus contributing to its clearness. Similarly, the epilogue has traditionally been used to arouse emotions like pity or anger; but as soon as the epilogue recalls the conclusions reached, it will make the speech more understandable.

5. The Three Means of Persuasion

The systematical core of Aristotle's *Rhetoric* is the doctrine that there are three technical means of persuasion. The attribute “technical” implies two characteristics: (i) Technical persuasion must rest on a method, and this, in turn, is to say that we must know the reason why some things are persuasive and some are not. Further, methodical persuasion must rest on a complete analysis of what it means to be persuasive. (ii) Technical means of persuasion must be provided by the speaker himself, whereas preexisting facts, such as oaths, witnesses, testimonies, etc. are non-technical, since they cannot be prepared by the speaker.

A speech consists of three things: the speaker, the subject which is treated in the speech, and the hearer to whom the speech is addressed (*Rhet.* I.3, 1358a37ff.). It seems that this reason why only three technical means of persuasion are possible: Technical means of persuasion are either (a) in the character of the speaker, or (b) in the emotional state of the hearer, or (c) in the argument (*logos*) itself.

(a) The persuasion is accomplished by character whenever the speech is held in such a way as to render the speaker worthy of credence. If the speaker appears to be credible, the audience will form the second order judgment that propositions put forward by the credible speaker are true or acceptable. This is especially important in cases where there is no exact knowledge but room for doubt. But how does the speaker manage to appear as a credible person? He must display (i) practical intelligence (*phronêsis*), (ii) a virtuous character, and (iii) good will (*Rhet.* II.1, 1378a6ff.); for, if he displayed none of them, the

audience would doubt that he is able to give good advices at all. Again, if he displayed (i) without (ii) and (iii), the audience could doubt whether the aims of the speaker are good. Finally, if he displayed (i) and (ii) without (iii), the audience could still doubt whether the speaker gives the best suggestion, though he knows what it is. But if he displays all of them, Aristotle concludes, it cannot rationally be doubted that his suggestions are credible. It must be stressed that the speaker must accomplish these effects by *what* he says; it is not necessary that he is actually virtuous: on the contrary, a preexisting good character cannot be part of the technical means of persuasion.

(b) The success of the persuasive efforts depends on the emotional dispositions of the audience; for we do not judge in the same way when we grieve and rejoice or when we are friendly and hostile. Thus, the orator has to arouse emotions exactly because emotions have the power to modify our judgments: to a judge who is in a friendly mood, the person about whom he is going to judge seems not to do wrong or only in a small way; but to the judge who is in an angry mood, the same person will seem to do the opposite (cp. *Rhet.* II.1, 1378a1ff.). Many interpreters writing on the rhetorical emotions were misled by the role of the emotions in Aristotle's ethics: they suggested that the orator has to arouse the emotions in order (i) to motivate the audience or (ii) to make them better persons (since Aristotle requires that virtuous persons do the right things together with the right emotions). Thesis (i) is false for the simple reason that the aim of rhetorical persuasion is a certain judgment (*krisis*), not an action or practical decision (*prohairesis*). Thesis (ii) is false, because moral education is not the purpose of rhetoric (see above §4), nor could it be effected by a public speech: “Now if speeches were in themselves enough to make men good, they would justly, as Theognis says, have won very great rewards, and such rewards should have been provided; but as things are …. they are not able to encourage the many to nobility and goodness.” (EN X.9. 1179b4-10)

How is it possible for the orator to bring the audience to a certain emotion? Aristotle's technique essentially rests on the knowledge of the definition of every significant emotion. Let, for example, anger be defined as “desire, accompanied with pain, for conspicuous revenge for a conspicuous slight that was directed against oneself or those near to one, when such a slight is undeserved.” (Rhet. II.2 1378a31-33). According to such definitions, someone who believes that he has suffered a slight from a person, who is not entitled to do so, etc., will become angry. If we take such a definition for granted, it is possible to deduce circumstances in which a person will most probably be angry; for example, we can deduce (i) in what state of mind people are angry and (ii) against whom they are angry and (iii) for what sorts of reason. Aristotle deduces these three factors for several emotions in the chapters II.2-11. With this equipment the orator will be able, for example, to highlight such characteristics of a case which are likely to provoke anger in the audience. In comparison with the tricks of former rhetoricians this method of arousing emotions has a striking advantage: The orator who wants to arouse emotions must not even speak outside the subject; it is sufficient to detect aspects of a given subject which are causally connected with the intended emotion.

(c) We persuade by the argument itself when we demonstrate or seem to demonstrate that something is the case. For Aristotle, there are two species of arguments: inductions and deductions (*Posterior Analytics* I.1, 71a5ff.). Induction (*epagôgê*) is defined as the proceeding from particulars up to a universal (*Topics* I.12, 105a13ff.). A deduction (*sullogismos*) is an argument in which, certain things having been supposed,

something different from the suppositions results of necessity through them (*Topics* I.1, 100a25ff.) or because of their being true (*Prior Analytics* I.2, 24b18-20). The inductive argument in rhetoric is the example (*paradeigma*); as opposed to other inductive arguments it does not proceed from many particular cases to one universal case, but from one particular to a similar particular if both particulars fall under the same genus (*Rhet.* I.2, 1357b25ff.). The deductive argument in rhetoric is the enthymeme (see below §6):

> but when, certain things being the case, something different results, beside them because of their being true, *either universally or for the most part*, it is called deduction here (in dialectic) and enthymeme there (in rhetoric).

It is remarkable that Aristotle uses the qualification "either universally or for the most part": obviously, he wants to say that in some cases the conclusion follows universally, i.e. by necessity, while in other cases it follows only *for the most part*. At first glance, this seems to be inconsistent, since a non-necessary inference is no longer a deduction. However, it has been disputed whether in arguments from probable premises the formula "for the most part" qualifies the inference itself ("If for the most part such and such is the case *it follows for the most part* that something different is the case"), or only the conclusion ("If for the most part such and such is the case *it follows by necessity* that *for the most part* something different is the case"). If the former interpretation is true, then Aristotle concedes in the very definition of the enthymeme that some enthymemes are not deductive. But if the latter interpretation (which has a parallel in *An. post.* 87b23-25) is correct, an enthymeme whose premises and conclusion are for the most part true would still be a valid deduction.

6. The Enthymeme

6.1 The Concept of Enthymeme

For Aristotle, an enthymeme is what has the function of a proof or demonstration in the domain of public speech. Since a demonstration is a kind of *sullogismos*, and the enthymeme is said to be a *sullogismos* too. The word '*enthymeme*' (from '*enthumeisthai* - to consider') had already been coined by Aristotle's predecessors and originally designated clever sayings, bon mots and short arguments involving a paradox or contradiction. The concepts 'proof' (*apodeixis*) and '*sullogismos*' play a crucial role in Aristotle's logical-dialectical theory. In applying them to a term of conventional rhetoric Aristotle appeals to a well known rhetorical technique, but, at the same time, restricts and codifies the original meaning of 'enthymeme': properly understood, what people call 'enthymeme' *should* have the form of a *sullogismos*, i.e. a deductive argument.

6.2 Formal Requirements

In general, Aristotle regards deductive arguments as a set of sentences in which some sentences are premises and one is the conclusion, and the inference from the premises to the conclusion is guaranteed by the premises alone. Since enthymemes in the proper sense are expected to be deductive arguments, the

minimal requirement for the formulation of enthymemes is that they have to display the premise-conclusion-structure of deductive arguments. This is why enthymemes have to include a statement as well as a kind of reason for the given statement. Typically this reason is given in a conditional ‘if’-clause or a causal ‘since’- or ‘for’-clause. Examples of the former , conditional type are: “If not even the gods know everything, human beings can hardly do so.” “If the war is the cause of present evils, things should be set right by making peace.” Examples of the latter, causal type are: “One should not be educated, for one ought not be envied (and educated people are usually envied).” “She has given birth, for she has milk.” Aristotle stresses that the sentence “There is no man among us who is free” taken for itself is a maxim, but becomes an enthymeme as soon as it is used together with a reason such as “for all are slaves of money or of chance (and no slave of money or chance is free).” Sometimes the required reason may even be implicit, as e.g. in the sentence “As a mortal do not cherish immortal anger” the reason why one should not cherish mortal anger is implicitly given in the phrase “immortal,” which alludes to the rule that is not appropriate for mortal beings to have such an attitude.

6.3 Enthymemes as Dialectical Arguments

Aristotle calls the enthymeme the “body of persuasion,” implying that everything else is only an addition or accident to the core of the persuasive process. The reason why the enthymeme as the rhetorical kind of proof or demonstration should be regarded as central for the rhetorical process of persuasion is that we are most easily persuaded when we think that something has been demonstrated. Hence, the basic idea of a rhetorical demonstration seems to be this: In order to make a target group believe that q, the orator must first select a sentence p or some sentences $p_1 \ldots p_n$ that are already accepted by the target group, secondly she has to show that q can be derived from p or $p_1 \ldots p_n$, using p or $p_1 \ldots p_n$ as premises. Given that the target persons form their beliefs in accordance with rational standards, they will accept q as soon as they understand that q can be demonstrated on the basis of their own opinions.

Consequently, the construction of enthymemes is primarily a matter of deducing from accepted opinions (*endoxa*). Of course, it is also possible to use premises which are not commonly accepted by themselves, but can be derived from commonly accepted opinions; other premises are only accepted since the speaker is held to be credible; still other enthymemes are built from signs: see §6.5. That a deduction is made from accepted opinions—as opposed to deductions from first and true sentences or principles—is the defining feature of dialectical argumentation in the Aristotelian sense. Thus, the formulation of enthymemes is a matter of dialectic, and the dialectician has the competence that is needed for the construction of enthymemes. If enthymemes are a subclass of dialectical arguments then, it is natural to expect a specific difference by which one can tell enthymemes apart from all other kinds of dialectical arguments (traditionally, commentators regarded logical incompleteness as such a difference; for some objections against the traditional view see §6.4). Nevertheless, this expectation is somehow misled: The enthymeme is different from other kinds of dialectical arguments, insofar as it is used in the rhetorical context of public speech (and rhetorical arguments are called ‘enthymemes’); thus, no further formal or qualitative differences are needed.

However, in the rhetorical context there are two factors that the dialectician has to keep in mind if she

wants to become a rhetorician too, and if the dialectical argument is to become a successful enthymeme. Firstly, the typical subjects of public speech do not - as the subject of dialectic and theoretical philosophy - belong to the things that are necessarily the case, but are among those things which are the goal of practical deliberation and can also be otherwise. Secondly, as opposed to well trained dialecticians the audience of public speech is characterized by an intellectual insufficiency; above all, the member of a jury or assembly are not accustomed to follow a longer chain of inferences. Therefore enthymemes must not be as precise as a scientific demonstration and should be shorter than ordinary dialectical arguments. This, however, is not to say that the enthymeme is defined by incompleteness and brevity. Rather, it is a sign of a well executed enthymeme that the content and the number of its premises are adjusted to the intellectual capacities of the public audience; but even an enthymeme which failed to incorporate these qualities would still be enthymeme.

6.4 The Brevity of the Enthymeme

In a well known passage (*Rhet.* I.2, 1357a7-18; similar: *Rhet.* II.22, 1395b24-26) Aristotle says that the enthymeme often has few or even fewer premises than some other deductions, (*sullogismoi)*. Since most interpreters refer the word '*sullogismos*' to the syllogistic theory (see the entry on Aristotle's logic) according to which a proper deduction has exactly two premises, those lines have led to the wide spread understanding that Aristotle defines the enthymeme as a *sullogismos* in which one of two premises has been suppressed, i.e. as an abbreviated, incomplete syllogism. But certainly the mentioned passages do not attempt to give a definition of the enthymeme, nor does the word '*sullogismos*' necessarily refer to deductions with exactly two premises. Properly understood, both passages are about the selection of appropriate premises, not about logical incompleteness. The remark that enthymemes often have few or less premises concludes the discussion of two possible mistakes the orator could make (Rhet. I.2, 1357a7-10): One can draw conclusions from things that have previously been deduced or from things that have not been deduced yet. The latter method is unpersuasive, for the premises are not accepted nor have they been introduced. The former method is problematic too: if the orator has to introduce the needed premises by another deduction, and the premises of this pre-deduction too, etc., one will end up with a long chain of deductions. Arguments with several deductive steps are common in dialectical practice, but one cannot expect the audience of a public speech to follow such long arguments. This is why Aristotle says that the enthymeme is and should be from fewer premises.

Supplement on The Brevity of the Enthymeme

6.5 Different Types of Enthymemes

Just as there is a difference between real and apparent or fallacious deductions in dialectic, we have to distinguish between real and apparent or fallacious enthymemes in rhetoric. The *topoi* for real enthymemes are given in chapter II.23, for fallacious enthymemes in chapter II.24. The fallacious enthymeme pretends to include a valid deduction, while it actually rests on a fallacious inference.

Further, Aristotle distinguishes between enthymemes taken from probable (*eikos*) premises and

enthymemes taken from signs (*sêmeia*). (Rhet. I.2, 1357a32-33). In a different context he says that enthymemes are based on probabilities, examples, *tekmêria* (i.e. proofs, evidences), and signs (Rhet. II.25, 1402b12-14). Since the so-called *tekmêria* are a subclass of signs and the examples are used to establish general premises, this is only an extension of the former classification. (Note that both classifications do not interfere with the idea that premises have to be accepted opinions: with respect to the signs the audience must *believe* that they exist and *accept* that they indicate the existence of something else, and with respect to the probabilities people must *accept* that something is likely to happen.)However, it is not clear whether this is meant to be an exhaustive typology. That most of the rhetorical arguments are taken from probable premises (“For the most part it is true that …,” “It is likely that …”), is due to the typical subjects of public speech, which are rarely necessary. When using a sign-argument or sign-enthymeme we do not try to explain a given fact; we just indicate, *that* something exists or is the case: “… anything such that when it is another thing is, or when it has come into being the other has come into being before or after, is a sign of the other's being or having come into being.” (*Prior Analytics* II.27, 70a7ff.). But there are several types of sign-arguments too; Aristotle offers the following examples:

	Rhetoric **I.2**	***Prior Analytics*** **II.27**
(i)	Wise men are just, since Socrates is just.	Wise men are good, since Pittacus is good.
(ii)	He is ill, since he has fever.	This man has fever, since he breathes rapidly.
(iii)	She has given birth, since she has milk.	This woman has a child, since she has milk.
		She is pregnant, since she is pale.

Sign-arguments of type (i) and (iii) can always be refuted, even if the premises are true; that is to say that they do not include a valid deduction (*sullogismos*); Aristotle calls them *asullogistos* (non-deductive). Sign-arguments of type (ii) can never be refuted if the premise is true, since, for example, it is not possible that someone has fever without being ill, or that someone has milk without having given birth, etc. This latter type of sign-enthymemes is necessary and is also called *tekmêrion* (proof, evidence). Now, if some sign-enthymemes are valid deductions and some are not, it is tempting to ask whether Aristotle regarded the non-necessary sign-enthymemes as apparent or fallacious arguments. However, there seems to be a more attractive reading: We accept a fallacious argument only if we are deceived about its logical form. But we could regard, for example, the inference “She is pregnant, since she is pale.” as a good and informative argument, even if we know that it does not include a logically necessary inference. So it seems as if Aristotle didn't regard all non-necessary sign-arguments as fallacious or deceptive; but even if this is true, it is difficult for Aristotle to determine the sense in which non-necessary sign-enthymemes are valid arguments, since he is bound to the alternative of deduction and induction, and neither class seems appropriate for non-necessary sign-arguments.

7. The *Topoi*

Generally speaking, an Aristotelian *topos* ('place', 'location') is an argumentative scheme which enables a dialectician or rhetorician to construe an argument for a given conclusion. The use of so-called *topoi* or '*loci communes*' can be traced back to early rhetoricians such as Protagoras, Gorgias (cp. Cicero, *Brutus* 46-48) and Isocrates. But, while in earlier rhetoric a *topos* was understood as a complete pattern or formula that can be mentioned at a certain stage of the speech to produce a certain effect, most of the Aristotelian *topoi* are general instructions saying that a conclusion of a certain form can be derived from premises of a certain form; and because of this 'formal' or 'semi-formal' character of Aristotelian *topoi*, one *topos* can be used to construe several different arguments. —Aristotle's book *Topics* lists some hundred *topoi* for the construction of dialectical arguments. These lists of *topoi* form the core of the method by which the dialectician should be able to formulate deductions on any problem that could be proposed. Most of the instructions that the *Rhetoric* gives for the composition of enthymemes are also organized as lists of *topoi*; especially the first book of the *Rhetoric* essentially consists of *topoi* concerning the subjects of the three species of public speech.

7.1 The Definition of 'Topos'

It is striking that the work which is almost exclusively dedicated to the collection of *topoi*, the book *Topics*, does not even make an attempt to define the concept of *topos*. At any rate the *Rhetoric* gives a sort of defining characterization: "I call the same thing element and *topos*; for an element or a *topos* is a heading under which many enthymemes fall" (*Rhet.* 1403a18-19). By 'element' Aristotle does not mean a proper part of the enthymeme, but a general form under which many concrete enthymemes of the same type can be subsumed. According to this definition the *topos* is a general argumentative form or pattern, and the concrete arguments are instantiations of the general *topos*. That the *topos* is a general instruction from which several arguments can be derived, is crucial for Aristotle's understanding of an artful method of argumentation; for a teacher of rhetoric who makes his pupils learn ready samples of arguments would not impart the art itself to them, but only the products of this art, just as if someone pretending to teach the art of shoe-making only gave samples of already made shoes to his pupils (see *Sophistical Refutations* 183b36ff.).

7.2 The Word '*Topos*' and the Technique of Places

The word '*topos*' (place, location) most probably is derived from an ancient method of memorizing a great number of items on a list by associating them with successive places, say the houses along a street one is acquainted with. By recalling the houses along the street we can also remember the associated items. Full descriptions of this technique can be found in Cicero, *De Oratore* II 86-88, 351--360, *Auctor ad Herennium* III 16-24, 29-40, and in Quintilian, *Institutio* XI 2, 11-33). In *Topics* 163b28--32 Aristotle seems to allude to this technique: "For just as in the art of remembering, the mere mention of the places instantly makes us recall the things, so these will make us more apt at deductions through looking to these defined premises in order of enumeration." Aristotle also alludes to this technique in *On the soul* 427b18-

20, *On Memory* 452a12-16, and *On Dreams* 458b20-22.

But though the name ‘*topos*’ may be derived from this mnemotechnical context, Aristotle's use of *topoi* does not rely on the technique of places. At least within the system of the book *Topics*, every given problem must be analyzed in terms of some formal criteria: Does the predicate of the sentence in question ascribe a genus or a definition or peculiar or accidental properties to the subject? Does the sentence express a sort of opposition, either contradiction or contrariety etc.? Does the sentence express that something is more or less the case? Does it maintain identity or diversity? Are the words used linguistically derived from words that are part of an accepted premise? Depending such formal criteria of the analyzed sentence one has to refer to a fitting *topos*. For this reason the succession of *topoi* in the book *Topics* is organized in accordance with their salient formal criteria; and this, again, makes a further mnemotechnique superfluous. More or less the same is true of the *Rhetoric*—except that most of its *topoi* are structured by material and not by formal criteria as we shall see in section 7.4.—Besides all this, there is at least one passage in which the use of the word ‘*topos*’ can be explained without referring to the previously mentioned mnemotechnique: In *Topics* VIII.1, 155b4-5 Aristotle says: “we must find the location (*topos*) from which to attack,” where the word ‘*topos*’ is obviously used to mean a starting point for attacking the theses of the opponents.

7.3 The Elements of a *Topos*

A typical Aristotelian *topos* runs as follows: “Again, if the accident of a thing has a contrary, see whether it belongs to the subject to which the accident in question has been declared to belong: for if the latter belongs, the former could not belong; for it is impossible that contrary predicates should belong at the same time to the same thing.” (*Topics* 113a20-24). As most *topoi* it includes (i) a sort of general instruction (“see, whether …”); further it mentions (ii) an argumentative scheme—in the given example the scheme ‘if the accidental predicate *p* belongs to the subject *s*, then the opposed *P** cannot belong to *s* too’. Finally, the *topos* refers to (iii) a general rule or principle (“for it is impossible, …”) which justifies the given scheme. Other *topoi* often include the discussion of (iv) examples; still other *topoi* suggest (v) how to apply the given schemes.—Though these are elements that regularly occur in Aristotelian *topoi*, there is nothing like a standard form with which all *topoi* comply. Often Aristotle is very brief and leaves it to the reader to add the missing elements.

7.4 The Function of a *Topos*

In a nutshell, the function of a *topos* can be explained as follows. First of all one has to select an apt *topos* for a given conclusion. The conclusion is either a thesis of our opponent which we want to refute, or our own assertion we want to establish or defend. Accordingly, there are two uses of *topoi*: they can either prove or disprove a given sentence; some can be used for both purposes, others for only one of them. Most *topoi* are selected by certain formal features of the given conclusion; if, for example, the conclusion maintains a definition, we have to select our *topos* from a list of *topoi* pertaining to definitions, etc. When it comes to the so-called ‘material’ *topoi* of the *Rhetoric* the appropriate *topos* must be selected not by formal criteria, but in accordance with the content of the conclusion—whether, for example, something is

said to be useful or honorable or just, etc. Once we have selected a *topos* which is appropriate for a given conclusion, the *topos* can be used to construe a premise from which the given conclusion can be derived. If for example the argumentative scheme is 'If a predicate is generally true of a genus, then the predicate is also true of any species of that genus', we can derive the conclusion 'the capacity of nutrition belongs to plants' using the premise 'the capacity of nutrition belongs to all living things', since 'living thing' is the genus of the species 'plants'. If the construed premise is accepted, either by the opponent in a dialectical debate or by the audience in public speech, we can draw the intended conclusion.

It has been disputed whether the *topos* (or, more precisely, the 'if …, then …' scheme that is included in a *topos*) which we use to construe an argument must itself be regarded as a further premise of the argument. It could be a premise either, as some say, as the premise of a propositional scheme such as the modus ponens, or, as others assume, as the conditional premise of a hypothetical syllogism. Aristotle himself does not favor one of these interpretations explicitly. But even if he regarded the *topoi* as additional premises in a dialectical or rhetorical argument, it is beyond any doubt that he did not use them as premises which must explicitly be mentioned or even approved by the opponent or audience.

Supplement on the *Topoi* of the Rhetoric

8. Style: How to Say Things with Words

8.1 The Virtue of Style

[Not yet available]

8.2 Aristotelian Metaphors

According to Aristotle *Poetics* 21, 1457b9-16 and 20-22 a metaphor is "the application of an alien name by transference either from genus to species, or from species to genus, or from species to species, or by analogy, that is, proportion." These four types are exemplified as follows:

	Type	Example	Explanation
(i)	From genus to species	There lies my ship	Lying at anchor is a species of the genus "lying"
(ii)	From species to genus	Verily ten thousand noble deeds hath Odysseus wrought	Ten thousand is a species of the genus "large number"
(iii)	From species to species.	(a) With blade of bronze drew away the life	(a) "To draw away" is used for "to cleave"

	(b) Cleft the water with the vessel of unyielding bronze	(b) "To cleave" is used for "to draw away." Both, to draw away and to cleave, are species of "taking away"
(iv) From analogy.	(a) To call the cup "the shield of Dionysus"	(a) The cup is to Dionysus as the shield to Ares
	(b) To call the shield "the cup of Ares"	(b) The shield is to Ares as the cup to Dionysus

Most of the examples Aristotle offers for types (i) to (iii) would not be regarded as metaphors in the modern sense; rather they would fall under the headings of metonomy or synecdoche. The examples offered for type (iv) are more like modern metaphors. Aristotle himself regards the metaphors of group (iv), which are built from analogy, as the most important type of enthymemes. An analogy is given if the second term is to the first as the fourth to the third. Correspondingly, an analogous metaphor use the fourth term for the second, or the second for the fourth. This principle can be illustrated by the following Aristotelian examples:

	Analogy	Metaphor
(a)	The cup to Dionysus as shield to Ares.	To call the cup "the shield of Dionysus" or the shield "the cup of Ares" is a metaphor.
(b)	The old age to life as the evening to day	To call the old age "evening of the life" or the evening "old age of the day" is a metaphor
(c)	Sowing to seed as *X* to sun rays, while the action of the sun in scattering his rays is nameless; still this process bears to the sun the same relation as sowing to the seed.	To call (a nameless) *X* "sowing of sun rays" is a metaphor by analogy
(d)	= (a)	To call the shield "a cup without wine" is also a metaphor by analogy.

Examples (a) and (b) obey to the optional instruction that metaphors can be qualified by adding the term to which the proper word is relative (cp. "the shield *of Ares*," "the evening *of life*"). In example (c) there is no proper name for the thing which is referred to by the metaphor. In example (d) the relation of analogy is not, as in the other cases, indicated by the domain to which an item is referred to, but by a certain negation (for example "without name"); the negations make clear that the term is not used in its usual sense.

Metaphors are closely related to similes; but as opposed to the later tradition, Aristotle does not define the metaphor as a abbreviated simile, but, the other way around, the simile as metaphor. The simile differs from the metaphor in the form of expression: while in the metaphor something is identified or substituted, the simile compares two things with each other, using words as "like," "as" etc. For example, "He rushed

as a lion," is, according to Aristotle, a simile, but "The lion rushed," is a metaphor.

While in the later tradition the use of metaphors has been seen as a matter of mere decoration, which has to delight the hearer, Aristotle stresses the cognitive function of metaphors. Metaphors, he says, bring about learning (*Rhet.* III.10, 1410b14f.). In order to understand a metaphor, the hearer has to find something common between the metaphor and the thing which the metaphor is referred to. For example, if someone calls the old age "stubble," we have to find a common genus to which old age and stubble belong; we do not grasp the very sense of the metaphor until we find that both, old age and stubble, have lost their bloom. Thus, a metaphor does not only refer to a thing, but simultaneously describes the respective thing in a certain respect. This is why Aristotle says that the metaphor brings about learning: as soon as we understand why someone uses the metaphor "stubble" to refer to old age, we have learned at least one characteristic of old age.

Glossary of Selected Terms

- Accepted opinions: *endoxa*
- Argument: *logos*
- Art: *technê*
- Character: *êthos*
- Counterpart: *antistrophos*
- Credible: *axiopistos*
- Decision (practical): *prohairesis*
- Deduction: *sullogismos*
- Emotions: *pathê*
- Enthymeme: *enthumêma*
- Example: *paradeigma*
- For the most part: *hôs epi to polu*
- Induction (*epagôgê*)
- Judgement: *krisis*
- Location: *topos* (an argumentative scheme)
- Maxim: *gnômê*
- Means of persuasion: *pistis* (in pre-Aristotelian use this word also designates a certain part of the speech)
- Metaphor: *metaphora*
- Persuasive: *pithanon*
- Place: *topos* (an argumentative scheme)
- Practical intelligence: *phronêsis*
- Premise: *protasis* (can also mean 'sentence', statement')
- Probable: *eikos*
- Proof: *apodeixis* (in the sense of 'demonstrative argument, demonstration')
- Proof: *tekmêrion* (i.e. a necessary sign or sign argument)
- Sign: *sêmeion* (can also mean 'sign argument')

- Style: *lexis*
- Specific *topoi*: *idioi topoi* (Aristotle refers to them also by '*idiai protaseis*' or '*eidê*')

Bibliography

- Barnes, Jonathan. 1981. "Proof and the Syllogism." In E. Berti (ed.), *Aristotle on Science: The Posterior Analytics*. Padua: Antenore. 17-59.
- Bitzer, L. F. 1959. "Aristotle's Enthymeme Revisited." In *Quarterly Journal of Speech* 45: 399-408.
- Burnyeat, Myles. 1994. "Enthymeme: The Logic of Persuasion." In D. J. Furley and A. Nehamas (eds.), *Aristotle's Rhetoric*. Princeton: Princeton University Press. 3-55.
- Cooper, John M. 1993. "Rhetoric, Dialectic, and the Passions." In *Oxford Studies in Ancient Philosophy* 11: 175-198.
- Cope, Edward Meredith. 1867 [1970]. *An Introduction to Aristotle's Rhetoric*. Cambridge 1867. Repr. Hildesheim: Olms.
- -----. 1877 [1970]. *The Rhetoric of Aristotle, with a Commentary*. Revised and edited by John Edwin Sandys. 3 vols. Cambridge: Cambridge University Press. Repr. Hildesheim: Olms.
- Cronkhite, Garry L. 1966. "The Enthymeme as Deductive Rhetorical Argument." In *Western Speech Journal* 30: 129-134.
- Dufour, Médéric and Wartelle, André. 1960-73. *Aristote, Rhétorique*. Texte établi et traduit. 3 vols. Paris: Les Belles Lettres.
- Erickson, Keith V. (ed.). 1974. *Aristotle: The Classical Heritage of Rhetoric*, Metuchen, N.J.
- Fortenbaugh, William W. 1970. "Aristotle's *Rhetoric* on Emotions." In *Archiv fuer Geschichte der Philosophie* 52: 40-70.
- ----- and Mirhady, David C. (eds.). 1994. *Peripatetic Rhetoric after Aristotle*. Rutgers University Studies in Classical Humanities 6, New Brunswick/London: Transaction Publishers.
- Freese, John Henry. 1926. *Aristotle, The 'Art' of Rhetoric*. London and Cambridge, Mass.: Loeb Classical Library. Harvard University Press.
- Furley, David J. and Nehamas, Alexander (eds.). 1994. *Aristotle's Rhetoric*. Princeton: Princeton University Press.
- Garver, Eugene. 1994. *Aristotle's Rhetoric. An Art of Character*, Chicago/London: The University of Chicago Press .
- Grimaldi, William M. A. 1957. "A Note on the *PISTEIS* in Aristotle's Rhetoric 1354-1356." In *American Journal of Philology* 78: 188-192.
- -----. 1980/1988. *Aristotle, Rhetoric I-II. A Commentary*. New York: Fordham University Press.
- Halliwell, Stephen. 1993. "Style and Sense in Aristotle's *Rhetoric* Book 3." In *Revue Internationale de Philosophie* 47: 50-69.
- Kassel, Rudolf. 1976. *Aristotelis Ars Rhetorica*. Berlin and New York: De Gruyter.
- Kennedy, George A. 1991. *Aristotle, On Rhetoric. A Theory of Civic Discourse*, Newly Translated, with Introduction, Notes and Appendices, New York/Oxford: Oxford University Press.
- Leighton, Stephen. 1982. "Aristotle and the Emotions." In *Phronesis* 27: 144-174.
- Madden, Edward H. 1952. "The Enthymeme. Crossroads of Logic, Rhetoric and Metaphysics." In

Philosophical Review 61: 368-376.
- McBurney, James H. 1936. “The Place of the Enthymeme in Rhetorical Theory.” In *Speech Monographs* 3: 49-74.
- Miller, Arthur B./Bee, John D. 1972. “Enthymemes: Body and Soul.” In *Philosophy and Rhetoric* 5: 201-214.
- Natali, Carlo 1990. “Due modi di trattare le opinioni notevole. La nozione di felicità in Aristotele, *Retorica* I 5.” In *Methexis* 3: 51-63.
- Primavesi, Oliver. 1996. *Die aristotelische Topik*. Munich: C. H. Beck.
- Raphael, Sally. 1974. “Rhetoric, Dialectic and Syllogistic Argument: Aristotle's Position in Rhetoric I-II.” In *Phronesis* 19: 153-167.
- Rapp, Christof. 1996. “Aristoteles ueber die Rationalitaet rhetorischer Argumente.” In *Zeitschrift für philosophische Forschung* 50: 197-222.
- -----. 2002. *Aristoteles, Rhetorik*. Translation, Introduction, and Commentary, 2 Vol. Berlin: Akademie Verlag.
- Roberts, W. Rhys. 1924 [1984]. *Rhetorica*. In W. D. Ross (ed.), *The Works of Aristotle Translated into English*, Oxford: Clarendon Press. Repr. in Jonathan Barnes (ed.), *The Works of Aristotle*. Princeton: Princeton University Press. II 2152-2269.
- Ryan, Eugene E. 1984. *Aristotle's Theory of Rhetorical Argumentation*, Montreal: Les Éditions Bellarmin.
- Seaton, R. C. 1914. “The Aristotelian Enthymeme.” In *Classical Review* 28: 113-119.
- Rorty, Amelie O. (ed.). 1996. *Essays on Aristotle's Rhetoric.* Berkeley/Los Angeles/London: University of California Press.
- Ross, W. D. (ed.). 1959. *Aristotelis ars rhetorica*. Oxford: Clarendon Press.
- Solmsen, Friedrich. 1929. *Die Entwicklung der aristotelischen Logik und Rhetorik*. Berlin: Weidmann.
- -----. 1938. “Aristotle and Cicero on the Orator's Playing upon the Feelings.” In *Classical Philology* 33: 390-404.
- Sprute, Juergen. 1982. *Die Enthymemtheorie der aristotelischen Rhetorik*. Goettingen: Vandenhoeck & Ruprecht .
- Thompson, W. H.. 1972. “Stasis in Aristotle's *Rhetoric*.” In *Quarterly Journal of Speech* 58: 134-141 .
- Weidemann, Hermann. 1989. “Aristotle on Inferences from Signs (*Rhetoric* I 2, 1357b1-25).” In *Phronesis* 34: 343-351.
- Woerner, Markus. 1990. *Das Ethische in der Rhetorik des Aristoteles*. Freiburg/Munich: Alber.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle | Cicero

Christof Rapp
RappC@philosophie.HU-Berlin.de

First published: May 2, 2002
Content last modified: May 2, 2002

Stanford Encyclopedia of Philosophy
Supplement to Aristotle's Rhetoric

The Brevity of the Enthymeme

In Rhet. II.22, 1395b24-26, a passage which is parallel to Rhet. I.2, 1357a7-18, Aristotle says that the orator should avoid two tendencies: in formulating enthymemes one should neither (a) deduce from far away nor (b) take up everything. Traditionally, this description has been associated with the alleged logical incompleteness of the enthymeme. But again one cannot infer that the enthymeme has to leave out logically required premises: (a) Deducing from too far away means that the orator has chosen premises that are so remote that it is difficult to see the connection with the intended conclusion (for example, if the premise is a rather general principle and the conclusion pertains to a concrete decision). But this is a question which does not even affect the number or completeness of the used premises (it affects the length of the argument only if one tries to bridge the span between the original premise and the conclusion by several intermediate steps). Whether or not one (b) takes up everything, can, but must not affect the logical completeness of an argument: Let us assume that someone has chosen a very remote premise as in the argument that the Scythians have no flute players, since they have no wine plants. Obviously, the respective arguments can be construed in different ways. On the one hand one could put it in the straight form:

> (P_1): Since there is no wine in the land of the Scythians and
> (P_2): since flute-players can only be found where there is wine,
> (C): the Scythians have no flute-players.

But most probably this argument would not be persuasive, since the premise (P_2) can hardly be taken as a generally accepted and evident opinion. On the other hand one could take up everything to elucidate the connection to the intended conclusion; for example: “Since the art of flute-playing only flourishes where there are glittering parties, and glittering parties can only be where the guests get drunk, and the guests only get drunk where there is a sufficient supply of wine, and a sufficient supply of wine is only where … etc.” This latter method takes up everything, even things that are already evident. This is exactly what the orator has to avoid. But either way, whether one takes up everything or not, the argument can be a complete deduction. Therefore, the advice that enthymemes should be short and should not have redundant premises can, but need not affect the logical completeness of an argument. The fallacious tendencies of deducing from far away and taking up everything can be avoided if one starts with apt premises.

Christof Rapp
RappC@philosophie.HU-Berlin.de

Return to Aristotle's Rhetoric

First published: May 2, 2002
Content last modified: May 2, 2002

Stanford Encyclopedia of Philosophy
Supplement to Aristotle's Rhetoric

The *topoi* of the Rhetoric

Interpreters are faced with the problem that the use of the word ‘*topos*’ in Aristotle’s *Rhetoric* is much more heterogeneous than in the *Topics*. Beside *topoi* which do perfectly comply with the description given in the *Topics*, there is an important group of *topoi* in the *Rhetoric* which contain instructions for arguments not of a certain form, but with a certain predicate (for example, that something is good, or honorable, or just, or contributes to happiness etc.). While those material *topoi* are still used to build arguments, there are also uses of ‘*topos*’ in the context of the non-argumentative means of persuasion.

In I.2, 1358a2-35 Aristotle distinguishes between general/common *topoi* on the one hand and specific *topoi* on the other hand. In chapter I.2 he explains the sense of ‘specific’ by pointing out that some things are specific to physics, others to ethics, etc. But from chapter I.3 on he makes us think that ‘specific’ refers to the different species of rhetoric, so that some *topoi* are specific to deliberative, other to epideictic, and still others to juridical speech. While he is inclined to call the general or common *topoi* simply ‘*topoi*’, he uses several names for the specific *topoi* (*idiai protaseis*, *eidê*, *idioi topoi*). Roughly, it is clear that the specific *topoi* can be found in the first and the common *topoi* in the second book of the *Rhetoric*. Most interpreters assumed that all common *topoi* are listed in chapters II.23-24 (for real enthymemes in II.23, for fallacious enthymemes in II.24), but failed to notice that more common *topoi* can be found in II.19. Further, it may be tempting to call, as some do, the specific *topoi* ‘material’ and the common *topoi* ‘formal’; but in so doing interpreters often neglected that some of the common *topoi* in chapters II.23-24 are not all based on those formal categories on which the *topoi* of the *Topics* rely. Most of them are ‘common’ only in the sense that they are not specific to one single species of speech. Some of them only offer strategic advice, as, for example, to turn what has been said against oneself upon the one who said it. For this reason, it is completely misleading to say that the functions of specific and common *topoi* are complementary, insofar as the common *topoi* allegedly offer the logical form to a content that has been provided by the specific ones.

The specific *topoi* of the Rhetoric

Since Aristotle sometimes calls the specific *topoi* ‘*protaseis*’, and ‘*protasis*’ is at the same time the Greek word for ‘premise’ and ‘statement, sentence’, his treatment of specific *topoi* gave rise to serious confusions. Several authors subscribed to the view of Friedrich Solmsen that there are two types of enthymemes, insofar as some are taken from *topoi* and some are built from premises, not from *topoi*. According to this view the specific *topoi* given in the first book of the *Rhetoric* are the premises of the latter type of enthymemes ,and the enthymemes of the former type are taken only from common *topoi*. From this point of view only common *topoi* would be *topoi* in the proper sense, while specific *topoi*

would be, strictly speaking, nothing else but premises. Accordingly, one would expect to find sentences of the form “All *F* are just/noble/good” in the first book of the *Rhetoric*; with such sentences one could construe syllogisms like “All *F* are just/noble/good—This particular *x* is *F*—This particular *x* is just/noble/good.” But what we actually find in the first book hardly fits Solmsen’s model. In some sense we find more than the required premises, insofar as Aristotle not only gives us isolated sentences, but also certain sentences together with a reason or a justification. Further Solmsen can hardly make sense of the fact that Aristotle calls these alleged premises ‘*topoi*’. And above all, chapters I.6-7 of the *Rhetoric* offer *topoi* which can also be found in the third book of *Topics*; in the *Topics* they are clearly called ‘*topoi*’, so that there is no reason to assume that they are premises rather than *topoi*.

The general idea by which the specific *topoi* can be characterized is rather this: Every specific *topos* gives us a general (but not formal) description of things that are supposed to be good, noble, just etc. It also gives us a reason with which we are enabled to argue that the things described are good, noble, just etc. Typically, this reason refers the given description back to a generally held definition of what is good, noble, just etc., which is provided at the outset the several chapters. In some cases the reason is directly, in some cases it is indirectly linked with the initial definition. Example: The specific *topos* is: “what is pleasant is good, since it is desirable.” The phrase “what is pleasant” provides the general description, the phrase “since it is desirable” provides the reason. Now, at the beginning of the chapter the good has been defined as “what is desirable.” Another specific *topos* is “honor is good, since it is pleasant”; here the reason in question applies the previous *topos* that what is pleasant is good, so that the current *topos* is indirectly linked with the initial definition of what is good. The general description included in those *topoi* enables us to identify cases which the orator can present as good, noble, just etc., the added reason shows us how to argue for the goodness etc. of the selected things. Thus, the specific *topoi* do not only provide premises but complete argumentative patterns.

Several types of rhetorical topoi

The several topoi that can be found under the headings ‘specific’ and ‘common’ do not at all make up two homogeneous classes. Some of them have only a vague affinity to the standard form of topoi which prevails in the book *Topics*. Some so-called topoi of the Rhetoric neither belong to the specific nor to the common class, but are just instructions or patterns which are somehow useful in public speech. At least the following groups must be distinguished:

	Place	Description	Examples
(i)	I.5-14 (without I.6b—I.7)	specific *topoi* of the three species of speech	“Further, health, beauty, and the like are goods, for being bodily excellences and productive of many other good things.” “It is noble to avenge oneself on one's enemies and not to come to terms with them; for requital is just, and the just is noble.”

(ii)	I.6b	*topoi* on controversial goods	"That which most people seek after, and which is obviously an object of contention, is also a good; for, as has been shown, that is good which is sought after by everybody, and 'most people' is taken to be equivalent to 'everybody'."
(iii)	I.7	*topoi* on the greater good (the better)	"Again, where one good is always accompanied by another, but does not always accompany it, it is greater than the other, for the use of the second thing is implied in the use of the first."
(iv)	I.15	*topoi* of non-technical means of persuasion	"We shall argue that justice indeed is true and profitable, but that sham justice is not, and that consequently the written law is not, because it does not fulfill the true purpose of law."
(v)	II.2-11, II.12-17	*topoi* to arouse emotions	"Again we are angry if something is not in line with what we expected, since what is not in line with what we expect provides more pain."
(vi)	II.19	*topoi* about the possible, the past, the future	"If the beginning of a thing can occur, so can the end; for nothing impossible occurs or begins to occur."
(vii)	II.23-24	common *topoi* (type 1)	"If a quality does not in fact exist where it is more likely to exist, it clearly does not exist where it is less likely."
(viii)	II.23-24	common *topoi* (type 2)	"Another line of argument is common to forensic and deliberative oratory, namely, to consider inducements and deterrents, and the motives people have for doing or avoiding the actions in question."
(ix)	II.23-24	common *topoi* (type 3)	"Another line is to apply to the other speaker what he has said against yourself."

(x)	III.15	topoi for slandering	"Another method is to denounce calumny, by saying what an enormity it is, and in particular that it raises false issues, and that it means a lack of confidence in the merits of his case."

Typical examples of group (i) can be found in chapters I.5-6 (first half), 9 and others. Starting form a definition of what is happiness, good, honorableand just etc., those *topoi* instruct arguments to the effect that items of a certain type are part of happiness, are good, honorable, just etc. The groups (ii) und (iii) have been inserted to indicate that the so-called specific *topoi* include various kinds of instructions: while the *topoi* of the first group offer determinate premises from which one can deduce that items of a certain type are good, just etc., the *topoi* of groups (ii) to (iii) help to construe diverse premises for items of several kinds. The *topoi* of group (iv) tell the orator what to say if one is using non-technical (i.e. artless) means of persuasion such as contracts, laws, witnesses etc. Just as the *topoi* of group (x), which offers the orators formulas for slandering, the underlying concept of *topos* in this group is essentially the same as in the pre-Aristotelian usage. The items mentioned in group (v) by which the orator should be enabled to arouse certain emotions in different contexts are also called '*topoi*', though they do not contribute to argumentation in the strictest sense. With respect to their formal character the topoi of group (vi) are quite familiar to the topoi of the *Topics*. The topoi of groups (vii) to (ix) are common insofar as they are not connected with a certain species of rhetoric. But while the topoi of group (vii) are roughly of the same type as the topoi of the *Topics*, other so-called 'common' topoi are exclusively suited for rhetorical purposes; the *topoi* of group (ix) only offer strategic advices.

Christof Rapp
RappC@philosophie.HU-Berlin.de

Return to Aristotle's Rhetoric

First published: May 2, 2002
Content last modified: May 2, 2002

Georg Wilhelm Friedrich Hegel

Along with J. G. Fichte and F. W. J. von Schelling, Hegel (1770-1831) belongs to the period of “German idealism” in the decades following Kant. The most systematic of the post-Kantian idealists, Hegel attempted, throughout his published writings as well as in his lectures, to elaborate a comprehensive and systematic ontology from a “logical” starting point. He is perhaps most well-known for his teleological account of history, an account which was later taken over by Marx and “inverted” into a materialist theory of an historical development culminating in communism. For most of the twentieth century, the “logical” side of Hegel's thought had been largely forgotten, but his political and social philosophy continued to find interest and support. However, since the 1970s, a degree of more general philosophical interest in Hegel's systematic thought has also been revived.

1. Life, Work, and Influence

Born in 1770 in Stuttgart, Hegel spent the years 1788-1793 as a theology student in nearby Tübingen, forming friendships there with fellow students, the future great romantic poet Friedrich Hölderlin (1770-1843) and Friedrich W. J. von Schelling (1775-1854), who, like Hegel, would become one of the major figures of the German philosophical scene in the first half of the nineteenth century. These friendships clearly had a major influence on Hegel's philosophical development, and for a while the intellectual lives of the three were closely intertwined.

After graduation Hegel worked as a tutor for families in Bern and then Frankfurt, where he was reunited with Hölderlin. Until around 1800, Hegel devoted himself to developing his ideas on religious and social themes, and seemed to have envisaged a future for himself as a type of modernising and reforming educator, in the image of figures of the German Enlightenment such as Lessing and Schiller. Around the turn of the century, however, possibly under the influence of Hölderlin, his interests turned more to the issues in the "critical" philosophy of Immanuel Kant (1724-1804) that had enthused Hölderlin, Schelling, and many others, and in 1801 he moved to the University of Jena to join Schelling. In the 1790s Jena had become a centre of both "Kantian" philosophy and the early romantic movement and by the time of Hegel's arrival Schelling had already become an established figure, taking the approach of J. G. Fichte (1762-1814), the most important of the new Kantian-styled philosophers, in novel directions. In late 1801, Hegel published his first philosophical work, *The Difference between Fichte's and Schelling's System of Philosophy*, and up until 1803 worked closely with Schelling, with whom he edited the *Critical Journal of Philosophy*. In his "*Difference*" essay Hegel had argued that Schelling's approach succeeded where Fichte's failed in the project of systematising and thereby completing Kant's transcendental idealism, and on the basis of this type of advocacy was dogged for many years by the reputation of being a "mere" follower of Schelling (who was five years his junior).

By late 1806 Hegel had completed his first major work, the *Phenomenology of Spirit* (published 1807), which showed a divergence from his earlier, seemingly more Schellingian, approach. Schelling, who had left Jena in 1803, interpreted a barbed criticism in the *Phenomenology*'s preface as aimed at him, and their friendship abruptly ended. The occupation of Jena by Napoleon's troops as Hegel was completing the manuscript closed the university and Hegel left the town. Now without a university appointment he worked for a short time, apparently very successfully, as an editor of a newspaper in Bamberg, and then from 1808-1815 as the headmaster and philosophy teacher at a "gymnasium" in Nuremberg. During his time at Nuremberg he married and started a family, and wrote and published his *Science of Logic*. In 1816 he managed to return to his university career by being appointed to a chair in philosophy at the University of Heidelberg. Then in 1818, he was offered and took up the chair of philosophy at the University of Berlin, the most prestigious position in the German philosophical world. While in Heidelberg he published the *Encyclopaedia of the Philosophical Sciences*, a systematic work in which an abbreviated version of the earlier *Science of Logic* (the "Encyclopaedia Logic" or "Lesser Logic") was followed by the application of its principles to the *Philosophy of Nature* and the *Philosophy of Spirit*. In 1821 in Berlin Hegel published his major work in political philosophy, *Elements of the Philosophy of Right*, based on lectures given at Heidelberg but ultimately grounded in the section of the Encyclopaedia *Philosophy of Spirit* dealing with "objective spirit." During the following ten years up to his death in 1831 Hegel enjoyed celebrity at Berlin, and published subsequent versions of the *Encyclopaedia*. After his death versions of his lectures on philosophy of history, philosophy of religion, aesthetics, and the history of philosophy were published.

After Hegel's death, Schelling, whose reputation had long since been eclipsed by that of Hegel, was invited to take up the chair at Berlin, reputedly because the government of the day had wanted to counter the influence that Hegelian philosophy had developed among a generation of students. Since the early period of his collaboration with Hegel, Schelling had become more religious in his philosophising and

criticised the “rationalism” of Hegel's philosophy. During this time of Schelling's tenure at Berlin, important forms of later critical reaction to Hegelian philosophy developed. Hegel himself had been a supporter of progressive but non-revolutionary politics, but his followers divided into “left-” and “right-wing” factions; from out of the former circle, Karl Marx was to develop his own “scientific” approach to society and history which appropriated many Hegelian ideas into Marx's materialistic outlook. (Later, especially in reaction to orthodox Soviet versions of Marxism, many “Western Marxists” re-incorporated further Hegelian elements back into their forms of Marxist philosophy.) Many of Schelling's own criticisms of Hegel's rationalism found their way into subsequent “existentialist” thought, especially via the writings of Kierkegaard, who had attended Schelling's lectures. Furthermore, the interpretation Schelling offered of Hegel during these years itself helped to shape subsequent generations' understanding of Hegel, contributing to the orthodox or traditional understanding of Hegel as a “metaphysical” thinker in the pre-Kantian “dogmatic” sense.

In academic philosophy, Hegelian idealism underwent a revival in both Great Britain and the United States in the last decades of the nineteenth century. In Britain, where philosophers such as T. H Green and F. H. Bradley had developed metaphysical ideas which they related back to Hegel's thought, Hegel came to be one of the main targets of attack by the founders of the emerging “analytic” movement, Bertrand Russell and G. E. Moore. For most of the twentieth century, interest in Hegel became limited to the context of his relation to other more popular philosophical movements like existentialism or Marxism, or to his social and political thought. In France, a version of Hegelianism came to influence a generation of thinkers, including Jean-Paul Sartre and the psychoanalyst, Jacques Lacan, largely through the lectures of Alexandre Kojève, an important precursor to the later “post-modern” movement. A later generation of French philosophers coming to prominence in the late 1960s and after, however, tended to react against Hegel in ways analogous to those in which early analytic philosophers had reacted against the Hegel who had influenced their predecessors. In Germany, interest in Hegel was revived early in the century with the historical work of Wilhelm Dilthey, and important Hegelian elements were incorporated into the approach of thinkers of the Frankfurt School, such as Theodor Adorno, and later, Jürgen Habermas, as well as the “hermeneutic” approach of H.-G. Gadamer. In Hungary, similar Hegelian themes were developed by Georg Lukács and later thinkers of the “Budapest School.” In the 1960s the German philosopher Klaus Hartmann developed what was termed a “non-metaphysical” interpretation of Hegel which, together with the work of Dieter Henrich and others, played an important role in the revival of interest in Hegel in academic philosophy in the second half of the century. Within English-speaking philosophy, the final quarter of the twentieth century saw something of a revival of serious interest in Hegel's philosophy, especially in North America, with important works appearing such as those by H. S. Harris, Charles Taylor, Robert Pippin and Terry Pinkard.

2. Hegel's Philosophy

Hegel's own pithy account of the nature of philosophy given in the “Preface” to his *Elements of the Philosophy of Right* captures a characteristic tension in his philosophical approach and, in particular, in his approach to the nature and limits of human cognition. “Philosophy,” he says there, “is its own time raised to the level of thought.”

On the one hand we can clearly see in the phrase “its own time” the suggestion of an historical or cultural conditionedness and variability which applies even to the highest form of human cognition, philosophy itself -- the contents of philosophical knowledge, we might suspect, will come from the historically changing contents of contemporary culture. On the other, there is the hint of such contents being “raised” to some higher level, presumably higher than other levels of cognitive functioning -- those based in everyday perceptual experience, for example, or those characteristic of other areas of culture such as art and religion. This higher level takes the form of “thought” -- a type of cognition commonly taken as capable of having “eternal” contents (think of Plato and Frege, for example).

This antithetical combination within human cognition of the temporally-conditioned and the eternal, a combination which reflects a broader conception of the human being as what Hegel describes elsewhere as a “finite-infinite,” has led to Hegel being regarded in different ways by different types of philosophical readers. For example, an historically-minded pragmatist like Richard Rorty, distrustful of all claims or aspirations to the “God's-eye view,” could praise Hegel as a philosopher who had introduced this historically reflective dimension into philosophy (and setting it on the characteristically “hermeneutic” path which has predominated in modern continental philosophy) but who had unfortunately still remained bogged down in the remnants of the Platonistic idea of the search for ahistorical truths. Those adopting such an approach to Hegel tend to have in mind the (relatively) young author of the *Phenomenology of Spirit* and have tended to dismiss as “metaphysical” later and more systematic works like the *Science of Logic*. In contrast, the British Hegelian movement at the end of the nineteenth century, for example, tended to ignore the *Phenomenology* and the more historicist dimensions of his thought, and found in Hegel a systematic metaphysician whose *Logic* provided a systematic and definitive philosophical ontology of an idealist type. This latter traditional, “metaphysical” view of Hegel dominated Hegel reception for most of the twentieth century, but has over the last few decades been contested by many Hegel scholars who have offered an alternative, “post-Kantian” view of Hegel.

2.1 The traditional “metaphysical” view of Hegel's philosophy

Given the understanding of Hegel that predominated at the time of the birth of analytic philosophy together with the fact that early analytic philosophers were rebelling precisely against “Hegelianism” so understood, the “Hegel” encountered in discussions within analytic philosophy is often that of the late nineteenth-century interpretation. In this picture, Hegel is seen as offering a metaphysico-religious view of “Absolute Spirit” which draws on pantheistic ideas of the identity of the universe and God, together with theistic ideas concerning the necessary “self-consciousness” of God. The peculiarity of Hegel's view, on this account, lies in his idea that the mind of God becomes actual only via the minds of his creatures, who serve as its vehicle. It is as distributed bearers of this developing self-consciousness of God that those finitely-embodied inhabitants of the universe -- we humans -- can be such “finite-infinites.”

An important consequence of Hegel's metaphysics, so understood, concerns history and the idea of historical development or progress, and it is as an advocate of an idea concerning the logically-necessitated teleological course of history that Hegel is most often decried. To many critics Hegel not

only was an advocate of a disastrous political conception of the state and the relation of its citizens to it, a conception prefiguring twentieth-century totalitarianism, but had tried to underpin such advocacy with dubious logico-metaphysical speculations. With his idea of the development of “spirit” in history, Hegel is seen as literalising a way of talking about different cultures in terms of their “spirits,” of constructing a developmental sequence of epochs typical of nineteenth-century ideas of linear historical progress, and then enveloping this story of human progress in terms of one about the developing self-conscious of the cosmos-God itself.

As the bottom line of such an account concerned the evolution of states of a mind (God's), such an account is clearly an idealist one, but not in the sense, say, of Berkeley. The pantheistic legacy inherited by Hegel meant that he had no problem in considering an objective *outer* world beyond any particular subjective mind. But this objective world itself had to be understood as conceptually informed, as it were -- it was *objectified* spirit. Thus in contrast to Berkeleian “subjective idealism” it became common to talk of Hegel as incorporating the “objective idealism” of views, especially common among German historians, in which social life and thought were understood in terms of the conceptual or “spiritual” structures that informed them. But in contrast to both forms of idealism, Hegel, according to this reading, postulated a form of *absolute* idealism by including both subjective life and the objective cultural practices on which subjective life depended within the dynamics of the development of the self-consciousness and self-actualisation of God, the “Absolute Spirit.”

It is hardly surprising, given the more secular character of much twentieth-century philosophy, that Hegel, so understood, would be generally regarded as of merely historical interest. Nevertheless, Hegel was still seen by many as an important precursor of other more characteristically modern strands of thought such as existentialism and Marxist materialism. Existentialists were thought of as taking the idea of the finitude and historical and cultural dependence of individual subjects from Hegel and leaving out all pretensions to the “absolute,” while Marxists were thought of as taking the historical dynamics of the Hegelian picture but understanding this in materialist rather than idealist categories. But while the traditional view of Hegel remained a commonplace throughout the twentieth century it has come to be increasingly questioned as an accurate account of Hegel's philosophy within Hegel scholarship itself. In the last quarter of the century, an increasing number of Hegel interpreters argued that such an understanding was seriously flawed, and while various quite different philosophical interpretations of Hegel have emerged which attempt to acquit him of implausible metaphysico-theological views, one common tendency has been to stress the continuity of his ideas with the “critical philosophy” of Immanuel Kant.

2.2 The non-traditional or “post-Kantian” view of Hegel

Least controversially, it has been claimed that either particular works such as the *Phenomenology of Spirit*, or particular areas of Hegel's philosophy, especially his ethical and political philosophy, can be understood as standing independently of the type of unacceptable metaphysical system sketched above. Somewhat more controversially, it has also been argued that the traditional picture is simply wrong at a more general “metaphysical” level and that Hegel is in no way committed to the bizarre “spirit monism”

that has been traditionally attributed to him. While these latter views often differ among themselves and continue to take exception to various aspects of Hegel's actual work, they commonly agree in regarding Hegel as being a “post-Kantian” philosopher who had accepted that aspect of Kant's critical philosophy which has been the most influential, his critique of traditional “dogmatic” metaphysics. Thus while the traditional view sees Hegel as exemplifying the very type of metaphysical speculation that Kant successfully criticised, the post-Kantian view of Hegel sees him as both accepting and extending Kant's critique, even of turning it against the residual “dogmatically metaphysical” aspects of Kant's own philosophy.

To see Hegel as a post-Kantian is to regard him as extending that “critical” turn that Kant saw as setting his philosophy on a scientific footing in a way analogous to the work of Copernicus in cosmology. With his Copernican analogy Kant had compared the way that the positions of the sun and earth were reversed in Copernicus' transformation of cosmology to the way that the positions of knowing subject and known object were reversed in his own transcendental idealism. Objectivity could no longer be thought as a matter of mental representations “corresponding” to an object “in itself” . Having posed the question of the ground of the relation of a representation to an object, Kant had answered that where a representation was not made possible by the process of sensory affection, it could be justified as objective only if through it it became possible to *cognise* something *as an object*.

No sooner had Kant's philosophy appeared then many objections were raised, among which were complaints about the apparently irreducible gap between the mind qua universal discursive intelligence and the mind as individual psychological reality. Kantian ideas were quickly integrated by Schelling with extant Spinozist ideas concerning mind and body as different aspects of an underlying substance to yield a type of philosophical biology. Others, such as Wilhelm von Humboldt and Friedrich D. E. Schleiermacher joined Kantian ideas about the mind with philological ideas linking thought to the structures of historically variable languages. Other critics pointed to internal inconsistencies in Kant's picture in which the world in itself seemed to be thought of on the one hand as the cause of its appearance, and on the other, as beyond knowledge and its constituent categories such as “cause.” Among the ambitions of many of Kant's successors, including Hegel, was that of somehow “completing” Kant. In Hegel especially, many argue, one can see the ambition to bring together the universalist dimensions of Kant's transcendental program with the culturally particularist conceptions of his more historically and relativistically-minded contemporaries. This resulted in his controversial conception of “spirit,” as developed in his *Phenomenology of Spirit*. With this notion, it has been argued, Hegel was pursing the Kantian question of the conditions of rational human “mindedness” rather than being concerned with giving an account of the developing self-consciousness of God. But while Kant had limited such conditions to “formal” structures of the mind, Hegel extended them to include aspects of historically and socially determined forms of embodied existence.

3. Hegel's Works

3.1 Phenomenology of Spirit

The term “phenomenology” had been coined by the German scientist and mathematician (and Kant correspondent) J. H. Lambert (1728 -- 1777), and in a letter to Lambert, sent to accompany a copy of his “Inaugural Dissertation” (1770), Kant had proposed a “general phenomenology” as a necessary “propaedeutic” presupposed by the science of metaphysics. Such a phenomenology was meant to determine the “validity and limitations” of what he called the “principles of sensibility,” principles he had (he thought) shown in the accompanying work to be importantly different to those of conceptual thought. The term clearly suited Kant as he had distinguished the “phenomena” known through the faculty of sensibility from the “noumena” known conceptually. This envisioned “phenomenology” seems to coincide roughly with what he was to eventually entitle a “critique of pure reason,” although Kant's thought had gone through important changes by the time that he came to publish the work of that name (1781, second edition 1787). Perhaps because of this he never again used the term “phenomenology” for quite this purpose.

There is clearly some continuity between this Kantian notion and Hegel's project. In a sense Hegel's phenomenology is a study of “phenomena” (although this is not a realm he would contrast with that of “noumena”) and Hegel's *Phenomenology of Spirit* is likewise to be regarded as a type of “propaedeutic” to philosophy rather than an exercise in it -- a type of induction or education of the reader to the “standpoint” of purely conceptual thought of philosophy itself. As such, its structure has been compared to that of an “educational novel,” having an abstractly conceived protagonist -- the bearer of an evolving series of “shapes of consciousness” or the inhabitant of a series of successive phenomenal worlds -- whose progress and set-backs the reader follows and learns from. Or at least this is how the work sets out: in the later sections the earlier series of “shapes of consciousness” becomes replaced with what seem more like configurations of human social existence, and the work comes to look more like an account of interlinked forms of social existence and thought, the series of which maps onto the history of western European civilization from the Greeks to Hegel's own time. The fact that it ends in the attainment of “Absolute Knowing,” the standpoint from which real philosophy gets done, seems to support the traditionalist reading in which a “triumphalist” narrative of the growth of western civilization is combined with the theological interpretation of God's self-manifestation and self-comprehension. When Kant had broached the idea of a phenomenological propaedeutic to Lambert, he himself had still believed in the project of a purely conceptual metaphysics, but this was a project that in his later critical philosophy he came to disavow. Traditional readers of Hegel thus see the *Phenomenology*'s telos as attesting to Hegel's “pre-Kantian” (that is, “pre-critical”) outlook and his embrace of the metaphysical project that Kant famously came to dismiss as illusory. Supporters of the non-metaphysical Hegel obviously interpret this work and its telos differently. For example, some have argued that what this history tracks is the development of a type of social existence which enables a unique form of rationality, in that in such a society all dogmatic bases of thought have been gradually replaced by a system in which all claims become open to rational self-correction, by becoming exposed to demands for conceptually-articulated justifications.

Something of Hegel's phenomenological method may be conveyed by the first few chapters, which are perhaps among the more conventionally philosophical parts. Chapters 1 to 3 effectively follow a developmental series of “shapes of consciousness” or conscious attitudes which seem to be based upon distinct criteria for epistemic certainty. Chapter 1, “Sense-certainty” considers an epistemological attitude

involving an appeal to some immediately given perceptual contents -- the sort of role played by “sense data” in some early twentieth-century approaches to epistemology, for example. By following the protagonist's attempts to make these implicit criteria *explicit* we are meant to appreciate that any such contents, even the apparently most “immediate,” in fact contain implicit conceptually articulated presuppositions, and so, in Hegel's terminology, are “mediated.” One might compare Hegel's point here to that expressed by Kant in his well known claim that without concepts, those singular and immediate mental representations he calls “intuitions” are “blind.” In more recent terminology one might talk of the “concept-” or “theory-ladenness” of all experience, and the lessons of this chapter have been likened to that of Wilfrid Sellars's famous criticism of the “myth of the given.”

By the end of this chapter our protagonist consciousness (and by implication, we the audience to this drama) has learnt that the nature of consciousness cannot be as originally thought, rather its contents must have some implicit universal (conceptual) aspect to them. Consciousness thus now commences anew with its new implicit epistemic criterion -- the assumption that since the contents of consciousness are “universal” they must be publicly graspable by others as well. Hegel's name for this type of perceptual realism in which any individual's idiosyncratic private apprehension will always be in principle correctable by the experience of others is “perception” (*Wahrnehmung* -- in German this term having the connotations of *taking* (*nehmen*) to be *true* (*wahr*)). As with the case for “sense-certainty,” here again, by following the protagonist consciousness's efforts to make this implicit criterion explicit, we see how the criterion generates contradictions which eventually undermine it *as* a criterion for certainty. In fact, such collapse into a type of self-generated scepticism is typical of all the “shapes” we follow in the work, and there seems something inherently skeptical about such reflexive cognitive processes. But Hegel's point is equally that there has always been something *positive* that has been learned in such processes, and this learning is more than that which consists in the mere elimination of epistemological dead-ends. Rather, as in the way that the internal contradictions that emerged from sense-certainty had generated a new shape, perception, the collapse of any given attitude always involves the emergence of some new implicit criterion which will be the basis of a new emergent attitude. In the case of “perception,” the emergent new shape of consciousness Hegel calls “understanding” -- a shape which he identifies with scientific cognition rather than that of everyday “perception.”

The transition from Chapter 3 to 4, “The Truth of Self-Certainty,” also marks a more general transition from “consciousness” to “self-consciousness.” It is in the course of chapter 4 that we find what is perhaps the most well-known part of the *Phenomenology*, the account of the “struggle of recognition” in which Hegel examines the intersubjective conditions which he sees as necessary for any form of “consciousness“.

Like Kant, Hegel thinks that one's capacity to be “conscious” of some external object as something *distinct* from oneself requires the reflexivity of “self-consciousness,” that is, it requires one's awareness *of oneself* as a subject *for whom* something distinct, the object, is presented *as known*. Hegel goes beyond Kant, however, in making this requirement dependent on one's recognition (or acknowledgment -- *Anerkennung*) *as a subject* by *other* self-consciousnesses whom one recognises in turn. In short, one's *self*-consciousness is in no sense direct, as it was for Descartes, for example. It comes about only indirectly via one's recognising other conscious subjects' *recognition* of oneself! It is in this way that the

Phenomenology can change course, the earlier tracking of “shapes of consciousness” being effectively replaced by the tracking of distinct patterns of “mutual recognition” between subjects.

It is thus that Hegel has effected the transition from a phenomenology of “subjective mind,” as it were, to one of “objective spirit,” thought of as culturally distinct patterns of social interaction analysed in terms of the patterns of reciprocal recognition they embody. (“*Geist*” can be translated as either “mind” or “spirit,” but the latter, allowing a more cultural sense, as in the phrase “spirit of the age” (“*Zeitgeist*”), seems a more suitable rendering for the title.) But this is only worked out in the text gradually. We -- the reading, “phenomenological” we -- can see how particular shapes of self-consciousness, such as that of the other-worldly religious self-consciousness (“unhappy consciousness”) with which chapter 4 ends, depend on certain institutionalised forms of mutual recognition. But *we* are seeing this from the “outside” as it were, we still have to learn how real *in situ* self-consciousnesses could learn this of *themselves*. So we have to see how the protagonist self-consciousness could achieve this insight. It is to this end that we further trace the learning path of self-consciousness through the processes of “reason” (in chapter 5) before “objective spirit” can become the *explicit* subject matter of chapter 6, (*Spirit*).

Hegel's discussion of spirit starts from what he calls “*Sittlichkeit*” (translated as “ethical order” or “ethical substance”), “*Sittlichkeit*” being a nominalisation from the adjectival (or adverbial) form “*sittlich*,” “customary,” from the stem “*Sitte*” -- “custom” or “convention.” Thus Hegel might be seen as adopting the viewpoint that since social life is ordered by customs we can approach the lives of those living in it in terms of the patterns of those customs or conventions themselves -- the conventional practices, as it were, constituting specific *forms of life*. It is not surprising then that his account of spirit here starts with a discussion of religious and civic law. Undoubtedly it is Hegel's tendency to nominalise such abstract concepts as “customary” in his attempt to capture the *concrete* nature of such as patterns of conventional life, together with the tendency to then *personify* them (as in talking about “spirit” becoming “self-conscious”) that lends plausibility to the traditionalist understanding of Hegel. But for non-traditionalists it is not obvious that Hegel is in any way committed to any metaphysical supra-individual conscious beings with such usages. To take an example, in the second section of the chapter “Spirit” Hegel discusses “culture” as the “world of self-alienated spirit.” The idea seems to be that humans in society not only interact, but that they collectively create relatively enduring cultural products (stories, dramas, and so forth) within which they can recognise their own patterns of life reflected. We might find intelligible the idea that such products “hold up a mirror to society” within which “the society can regard itself,” without thinking we are thereby committed to some supra-individual social “mind” achieving self-consciousness. Furthermore, such cultural products themselves provide conditions allowing individuals to adopt particular cognitive attitudes. Thus, for example, the capacity to adopt the type of objective viewpoint demanded by Kantian morality (discussed in the final section of Spirit) -- the capacity to see things, as it were, from a “universal” point of view -- is bound up with the attitude implicitly adopted in engaging with spirit's “alienations.”

We might think that if *Kant* had written the *Phenomenology*, he would have ended it at chapter 6 with the modern moral subject as the telos of the story. For Kant, the practical knowledge of morality, orienting one within the *noumenal* world, exceeds the scope of theoretical knowledge which had been limited to phenomena. Hegel, however, thought that philosophy had to unify theoretical and practical knowledge,

and so the *Phenomenology* has further to go. Again, this is seen differently by traditionalists and revisionists. For traditionalists, Chapters 7, “Religion” and 8, “Absolute Knowing,” testify to Hegel's disregard for Kant's critical limitation of theoretical knowledge to empirical experience. Revisionists, on the other hand, tend to see Hegel as furthering the Kantian critique into the very coherence of a conception of an “in-itself” reality which is beyond the limits of our theoretical (but not practical) cognition. Rather than understand “absolute knowing” as the achievement of some ultimate “God's-eye view” of everything, the philosophical analogue to the connection with God sought in religion, revisionists see it as the accession to a mode of self-critical thought that has finally abandoned all non-questionable mythical “givens,” and which will only countenance reason-giving argument as justification. However we understand this, absolute knowing is the standpoint to which Hegel has hoped to bring the reader in this complex work. This is the “standpoint of science,” the standpoint from which philosophy proper commences, and it commences in Hegel's next book, the *Science of Logic*.

3.2 Science of Logic

Hegel's *Science of Logic*, the three constituent “books” of which appeared in 1812, 1813, and 1816 respectively, is a work that few contemporary logicians would recognise as a work of *logic*, but it is not meant as a treatise in formal (or “general”) logic. Rather, its provenance is to be found in what Kant had called “transcendental logic,” and which is more akin to what now is termed “epistemic” logic. In this sense it stands as a successor to Kant's “transcendental deduction of the categories” in the *Critique of Pure Reason* in which Kant attempted to “deduce” a list of those non-empirical concepts, the “categories,” which he believed to be presupposed by the empirical judgments of finite, discursive knowers like ourselves.

A glance at the table of contents of *Science of Logic* reveals the same triadic structuring noted in the *Phenomenology*. At the highest level of its branching structure there are three “books,” devoted to the doctrines of “being,” “essence,” and “concept” respectively. In turn, each book has three sections, each section containing three chapters, and so on. In general each of these nodes deals with some particular category or “thought determination,” sometimes the first subheading under a node having the same name as the node itself. To some extent, the treatment of the syllogism found in Book 3 (and following Aristotle's three-termed schematism of the syllogistic structure) might be seen as providing a retrospective justification for this structuring, Hegel's idea being that all rigorous thought about anything must grasp it in terms of the fundamental thought determinations of “singularity,” “particularity,” and “universality.” (This combination may, in fact, reflect the post-Kantians' re-interpretation of Kant's taxonomy of the basic components of cognition -- the division of mental representation into “singular” intuitions and “general” concepts. Fichte had understood that Kant equivocates over the relation of “sensation” to “intuition” : sometimes Kant treats sensations as *parts of* intuitive representations (their “matter”) and sometimes as *non-representational* states of the subject somehow “corresponding to” such matter. Kant's two-termed account therefore gets rearticulated as a three-termed account. In the later nineteenth century, no less a logician than Charles Sanders Peirce came to a similar idea about the fundamentally *trinary* structure of the categories of thought.)

Reading into the first chapter of Book 1, “Being,” it is quickly seen that the *Logic* repeats the movements of the first chapters of the *Phenomenology*, now, however, at the level of “thought” rather than conscious experience. Thus “being” is the thought determination with which the work commences because it at first seems to be the most “immediate,” fundamental determination characterising any possible thought content at all. It apparently has no internal structure (in the way that “bachelor,” say, has a structure containing further concepts “male” and “unmarried"). Again parallel to the *Phenomenology*, it is the effort of thought to make such contents explicit that both undermines them and brings about a new contents. “Being” *seems* “immediate” but reflection reveals that it itself is, in fact, only meaningful in opposition to another concept, “nothing.” In fact, the attempt to think “being” *as* immediate, and so as not *mediated* by its opposing concept “nothing,” has so deprived it of any determinacy or meaning at all that it effectively *becomes* nothing. That is, on reflection it is grasped as *having passed over* into its “negation” . Thus, while “being” and “nothing” *seem* both absolutely distinct and opposed, from another point of view they appear *the same* as no criterion can be invoked which differentiates them. The only way out of this paradox is to posit a third category, “becoming,” which seems to save thinking from paralysis because it accommodates both concepts: “becoming” contains “being” and “nothing” since when something “becomes” it passes, as it were, between nothingness and being. That is, when something *becomes* it seems to posses aspects of both *being* and *nothingness*.

In general this is how the *Logic* proceeds: seeking its most basic and universal determination, thought posits a category to be reflected upon, finds then that this collapses due to a contradiction generated, but then seeks a further category with which to make retrospective sense of that contradiction. This new category is more complex as it has internal structure in the way that “becoming” contains “being” and “nothing” as *moments*. But in turn the new category will generate some further contradictory negation and again the demand will arise for a further concept which will reconcile these opposed concepts by incorporating *them* as moments.

In this way the categorical infrastructure to thought becomes unpacked with only the use of those resources available to thought itself, its capacity to make its contents determinate (i.e., clear and distinct) and its refusal to tolerate contradiction. As has been mentioned, Hegel's logic might best be considered as a “transcendental” not a “formal” logic. Rather than treating the pure “form” of thought that has been abstracted from any possible content, transcendental logic treats thought that already possesses a certain type of content that Kant had called (predictably) “transcendental content.” This was that non-empirical but nevertheless intuitive element of “content” that was implicit in our thought, given that it was the thought of a particular *kind* of thinker, whose cognition about the world was restricted to the capacity to apply general concepts to singular and immediate empirical “intuitions.” It would seem to be this difference to traditional formal logic that underlies the contrast between the conceptual structure generated here, and that of the traditional “Tree of Porphyry” that results from the Platonic “method of division.” In the traditional structure, a more general concept is divided into more specific ones by means of some differentiating characteristic, in the way, for example, that the more general concept “animal” can be differentiated into “vertebrates” and “invertebrates.” In such a structure, the direction of conceptual *specificity*, and conceptual *containment* are reversed: a concept at any level will “contain,” as sub-concepts, all members of the chain of more abstract concepts standing “above” it. Thus if the concept “animal” is divided into the contraries “vertebrate” and “invertebrate,” each will in turn “contain” the

superordinating concept “animal” and thereby in turn contain every concept that is contained within (and stands above) “animal.” In contrast, in Hegel's conceptual structure, reflection on a concept produces its negation in a type of *internal* division, and then both concept and negation become contained as “moments” in the more specific concept that is posited to resolve the paradox of that internal negation.

If Hegel's is a transcendental logic, however, it is clearly different from that of Kant's. For Kant, transcendental logic was the logic governing the thought of *finite* thinkers like ourselves, whose cognition was constrained by the necessity of applying general discursive *concepts* to the singular contents given in sensory *intuitions*, and he kept open the possibility that there could be a kind of thinker not so constrained -- God, for example, whose thought could apply directly to the world in a type of “intellectual” intuition. Again, opinions divide as to how Hegel's approach to logic relates to that of Kant. Traditionalists see Hegel as treating the finite thought of individual human discursive intellects as a type of “distributed” vehicle for the classically conceived *infinite* and *intuitive* thought of God. Non-traditionalists, in contrast, see the post-Kantians as removing the last residual remnant of the mythical idea of transcendent godly thought from Kant's approach. On their account, the very opposition that Kant has between finite human thought and infinite godly thought is suspect, and the removal of this mythical obstacle allows an expanded role for “transcendental content.” Regardless of how we interpret this however, it is important to grasp that for Hegel logic is not simply a science of the *form* of our thoughts but is also a science of actual “content” as well, and as such is a type of *ontology*. Thus it is not just about the concepts “being,” “nothing,” “becoming” and so on, but about *being*, *nothing*, *becoming* and so on, *themselves*. This in turn is linked to Hegel's radically non-representationalist (and in some sense “direct realist”) understanding of thought. The world is not “represented” in thought by a type of “proxy” standing for it, but rather is presented, exhibited, or made manifest in it. (In recent analytic philosophy, John McDowell has presented an account of thought with this type of character, and has explicitly drawn a parallel to the approach of Hegel.)

The thought determinations of Book 1 lead eventually into those of Book 2, “The Doctrine of Essence.” Naturally the structures implicit in “essence” thinking are more developed than those of “being” thinking. Crucially, the contrasting pair “essence” and “appearance” allow the thought of some underlying reality which manifests itself through a different overlying appearance, a relation not able to be captured in the simpler “being” structures. Given the ontological dimension of Hegel's logic, its various stages are meant to coincide roughly with actual ontologies encountered in a history of metaphysics. Thus the metaphysics of Parmenides and Heraclitus, for example, line up with the thought determinations “being” and “becoming” at the beginning of Being-logic while Essence-logic culminates in concepts bound up with modern forms of substance metaphysics as found in Spinoza and Leibniz.

Book 3, “The Doctrine of Concept” effects a shift from the “Objective Logic” of Books 1 and 2, to “Subjective Logic,” and metaphysically coincides with a shift to the modern subject-based ontology of Kant. Just as Kantian philosophy is founded on a conception of objectivity secured by conceptual coherence, Concept-logic commences with the concept of “concept” *itself*! While in the two books of objective logic, the movement had been between particular concepts, “being,” “nothing,” “becoming” etc., in the subjective logic, the conceptual relations are grasped at a meta-level, such that the concept “concept” treated in Chapter 1 of section 1 (“Subjectivity”) passes over into that of “judgment” in

Chapter 2, as judgments are the larger wholes within which concepts themselves get related to each other. When the anti-foundationalism and holism of the *Phenomenology* is recalled, it will come as no surprise that the concept of judgment passes over into that of “syllogism”: for Hegel just as a concept gains its determinacy in the context of the judgments within which it is applied, so too do judgements gain their determinacy within larger patterns of *inference*. When Hegel declares the syllogism to be “the truth” of the judgment, he might be thought, as has been suggested by Robert Brandom, to be advocating a view somewhat akin to contemporary “inferentialist” approaches to semantics. On these approaches, an utterance gains its semantic content not from any combination of its already meaningful sub-sentential components, but from the particular inferential “commitments and entitlements” acquired when it is offered to others in practices presupposing the *asking for* and *giving of reasons*. Thought of in terms of the framework of Kant's “transcendental logic,” Hegel's position would be akin to allowing *inferences* -- “syllogisms” -- a role in the determination of “transcendental content,” a role which inference definitely does *not* have in Kant.

We might see then how the different ways of approaching Hegel's logic will be reflected in the interpretation given to the puzzling claim in Book 3 concerning the syllogism becoming “concrete” and “pregnant with” a content that has necessary existence. In contrast with Kant, Hegel seems to go beyond a “transcendental deduction” of the *formal* conditions of experience and thought and to a deduction of their *material* conditions. Traditionalists will see here something akin to the “ontological argument” of medieval theology in which the *existence* of something seems to have been necessitated by its concept -- an argument undermined by Kant's criticism of the treatment of *existence* as a predicate. In Hegel's version, it would be said, the objective existence that God achieves in the world has been necessitated by his essential self-consciousness. The revisionist reading, in contrast, would have to interpret this aspect of Hegel's logic differently.

As already noted, for Hegel, the logic of *inference* has a “transcendental content” in a way analogous to that possessed by the logic of *judgment* in Kant's transcendental logic. It is this which is behind the idea that the treatment of the formal syllogisms of inference will lead to a consideration of those syllogisms as “pregnant with content.” But for logic to be truly ontological a further step “beyond” Kant is necessary. For the post-Kantians, Kant had been mistaken in restricting the conditions of experience and thought to a “subjective” status. Kant's idea of our knowledge as restricted to the world as it is *for us* requires us to have a concept of the noumenal as that which cannot be known, the concept “noumenon” playing the purely *negative* role of giving a determinate sense to “phenomenon” by specifying its limits. That is, for Kant we need to be able to think of our experience and knowledge *as* finite and conditioned, and this is achieved in terms of a concept of a realm *we cannot know*. But, the post-Kantian objection goes, if the concept “noumenon” is to provide some sort of boundary to that of “phenomenon,” then it cannot be the *empty* concept that Kant supposed. Only a concept *with a content* can determine the limits of the content of some another concept (as when our empirical concept of “river,” for example, is made determinate by opposing empirical concepts like “stream” or “creek”). The positing of a noumenal realm must be the positing of a realm about which we can have *some* understanding.

This need felt by the post-Kantians for having a contentful concept of the “noumenal” or the “in itself” can also be seen from the inverse perspective. For Kant, sensation testifies to the existence of an

objective noumenal world beyond us, but *this* world cannot be known *as such*; we can only know that world as it appears to us from within the constraints of the subjective conditions of our experience and thought. But for Hegel this is to attribute to a wholly inadequate form of knowledge -- sensation or feeling -- a power that is being denied to a much *better* form of knowledge -- that articulated by *concepts*. To think that our inarticulate sensations or feelings give us a *truer* account of reality than that of which we are capable via the scientific exercise of conceptualised thought indicates a type of irrationalist potential within Kantian thought, a potential that Hegel thought was being realised by the approach of his romantic contemporaries. The rational kernel of Kant's approach, then, had to be carried beyond the limits of a method in which the conditions of thought and experience were regarded as *merely* subjective. Rather than restrict its scope to “formal” conditions of experience and thought, it had to be understood as capable of revealing the *objective* or material conditions. Transcendental logic must thereby become ontological. It may be significant here that, as some recent studies of Kant's own later work (the *Opus Postumum*) suggest, Kant himself seems to have revised his own approach such that something like a deduction of the material conditions of thought was now considered as the proper province of transcendental philosophy.

3.3 Philosophy of Right

Like the *Science of Logic*, the *Encyclopaedia of the Philosophical Sciences* is itself divided into three parts: a *Logic*; a *Philosophy of Nature*; and a *Philosophy of Spirit*. The same triadic pattern in the *Philosophy of Spirit* results in the philosophies of *subjective* spirit, *objective* spirit, and *absolute* spirit. The first of these constitutes Hegel's philosophy of mind, the last, his philosophy of art, religion, and philosophy itself. The philosophy of *objective spirit* concerns the objective patterns of social interaction and the cultural institutions within which “spirit” is objectified. The book entitled *Elements of the Philosophy of Right* which Hegel published as a textbook for his lectures at Berlin essentially corresponds to a more developed version of the section on “Objective Spirit” in the *Philosophy of Spirit*.

The *Philosophy of Right* (as it is more commonly called) can, and has been, read as a political philosophy which stands independently of the system, but it is clear that Hegel intended it to be read against the background of the developing conceptual determinations of the *Logic*. The text proper starts from the conception of a singular willing subject (grasped from its own first-person point of view) as the bearer of “abstract right.” While this conception of the individual willing subject with some kind of fundamental right is in fact the starting point of many modern political philosophies (such as that of Locke, for example) the fact that Hegel commences here does not testify to any ontological assumption that the consciously willing and right-bearing individual is the basic *atom* from which all society can be understood as constructed -- an idea at the heart of standard “social contract” theories. Rather, this is merely the most “immediate” starting point of Hegel's presentation and corresponds to analogous starting places of the *Logic*. Just as the categories of the *Logic* develop in a way meant to demonstrate that what had at the start been conceived as simple is in fact only made determinate in virtue of its being part of some larger structure or process, here too it is meant to be shown that any simple willing and right-bearing subject only gains *its* determinacy in virtue of a place it finds for itself in a larger *social*, and ultimately *historical*, structure or process. Thus even a contractual exchange (the minimal social interaction for contract theorists) is not to be thought simply as an occurrence consequent upon the

existence of two beings with natural wants and some natural calculative rationality; rather, the system of interaction within which individual exchanges take place (the economy) will be treated holistically as a culturally-shaped form of social life within which the actual wants of individuals as well as their reasoning powers are given determinate forms.

Here too it becomes apparent in Hegel's treatment of property and the exchange contract that the notion of *recognition* plays a crucial role in his general conception of the relation of individuals to each other and to society as a whole. A contractual exchange of commodities between two individuals itself involves an implicit act of recognition in as much as each, in giving something to the other in exchange for what they want, are thereby recognizing them *as* a proprietor of that thing, or, more properly, of the inalienable *value* attaching to it. By contrast, such proprietorship would be *denied* rather than recognised in fraud or theft -- forms of “wrong” (*Unrecht*) in which right is negated rather than acknowledged or posited. Thus what differentiates property from mere *possession* is that it is grounded in a relation of reciprocal recognition between two willing subjects. Moreover, it is in the exchange relation that we can see what it means for Hegel for individual subjects to share a “common will” -- an idea which will have important implications with respect to the difference of Hegel's conception of the state from that of Rousseau. Such an interactive constitution of the *common will* means that for Hegel such an identity of will is achieved *because of* not *in spite of* a co-existing *difference* between the particular wills of the subjects involved: while contracting individuals both “will” *the same* exchange, at a more concrete level, they do with different ends in mind. Each wants something different *from* the exchange.

Hegel passes from the abstract individualism of “Abstract Right” to the social determinacies of *“Sittlichkeit”* or “Ethical Life” via considerations first of “wrong” (the negation of right) and its punishment (the negation of wrong, and hence the “negation of the negation” of the original right), and then of “morality,” conceived more or less as an internalisation of the external legal relations. Consideration of Hegel's version of the retributivist approach to punishment affords a good example of his use of the logic of “negation.” In punishing the criminal the state makes it clear to its members that it is the acknowledgment of right *per se* that is essential to developed social life: the significance of “acknowledging another's right” in the contractual exchange cannot be, as it at first might have appeared to the participants, simply that of being a way of each getting what he or she wants *from the other*. Hegel's treatment of punishment also brings out the continuity of his way of conceiving of the structure and dynamics of the social world with that of Kant, as Kant too, in his *Metaphysics of Morals* had employed the idea of the state's punitive action as a *negating* of the original criminal act. Kant's idea, conceived on the model of the physical principle of action and reaction, was structured by the category of “community” or reciprocal interaction, and was conceived as involving what he called “real opposition.” Such an idea of opposed dynamic forces seems to form something of a model for Hegel's idea of contradiction and the starting point for his conception of reciprocal *recognition*. Nevertheless, clearly Hegel articulates the structures of recognition in more complex ways than those derivable from Kant's category of community.

First of all, in Hegel's analysis of *Sittlichkeit* the type of sociality found in the market-based “civil society” is to be understood as dependent upon and in contrastive opposition with the more immediate form found in the institution of the family -- a form of sociality mediated by a quasi-natural inter-

subjective recognition rooted in sentiment and feeling: love. In the family the particularity of each individual tends to be absorbed into the social unit, giving this manifestation of *Sittlichkeit* a one-sidedness that is the inverse of that found in market relations in which participants grasp themselves in the first instance as separate individuals who then enter into relationships that are external to them.

These two opposite but interlocking principles of social existence provide the basic structures in terms of which the component parts of the modern state are articulated and understood. As both contribute particular characteristics to the subjects involved in them, part of the problem for the rational state will be to ensure that each of these two principles mediate the other, each thereby mitigating the *one-sidedness* of the other. Thus, individuals who encounter each other in the “external” relations of the market place and who have their subjectivity shaped by such relations also belong to families where they are subject to opposed influences. Moreover, even within the ensemble of production and exchange mechanisms of civil society individuals will belong to particular “estates” (the agricultural estate, that of trade and industry, and the “universal estate” of civil servants), whose internal forms of sociality will show family-like features.

Although the actual details of Hegel's “mapping” of the categorical structures of the *Logic* onto the *Philosophy of Right* are far from clear, the general motivation is apparent. As has been mentioned above, Hegel's logical categories can be read as an attempt to provide a schematic account of the material (rather than formal) conditions required for developed self-consciousness. Thus we might regard the various “syllogisms” of Hegel's *Subjective Logic* as attempts to chart the skeletal structures of those different types of recognitive inter-subjectivity necessary to sustain various aspects of rational cognitive and conative functioning (“self-consciousness”). From this perspective, we might see his “logical” schematisation of the modern “rational” state as a way of displaying just those sorts of institutions that a state must provide if it is to answer Rousseau's question of the form of association needed for the formation and expression of the “general will.”

Concretely, for Hegel it is representation of the estates within the legislative bodies that is to achieve this. As the estates of civil society group their members according to their common interests, and as the deputies elected from the estates to the legislative bodies give voice to those interests within the deliberative processes of legislation, we might see how the outcome of this process might be considered to give expression to the general interest. But Hegel's “republicanism” is here cut short by his invocation of the familial principle: such representative bodies can only provide the *content* of the legislation to a *constitutional monarch* who must add to it the form of the royal decree -- an individual “I will ….” To declare that for Hegel the monarch plays only a “symbolic” role here is to miss the fundamentally idealist complexion of his political philosophy. The expression of the general will in legislation cannot be thought of as an outcome of some quasi-mechanical process: it must be “willed.” If legislation is to express the general will, citizens must recognize it as expressing *their* wills; and this means, recognising it *as* willed. The monarch's explicit “I will” is thus needed to close this recognitive circle, lest legislation look like a mechanical compromise resulting from a clash of interests, and so as actively *willed* by nobody. Thus while Hegel is critical of standard “social contract” theories, his own conception of the state is still clearly a complicated transformation of those of Rousseau and Kant.

Perhaps one of the most influential parts of Hegel's *Philosophy of Right* concerns his analysis of the contradictions of the unfettered capitalist economy. On the one hand, Hegel agreed with Adam Smith that the interlinking of productive activities allowed by the modern market meant that “subjective selfishness” turned into a ”contribution towards the satisfaction of the needs of everyone else.” But this did not mean that he accepted Smith's idea that this “general plenty” produced thereby *diffused* (or “trickled down”) though the rest of society. From within the type of consciousness generated within civil society, in which individuals are grasped as “bearers of rights” abstracted from the particular concrete relationships to which they belong, Smithean optimism may seem justified. But this simply attests to the one-sidedness of this type of abstract thought, and the need for it to be mediated by the type of consciousness based in the family in which individuals are grasped in terms of the way they *belong to* the social body. In fact, the unfettered operations of the market *produces* a class caught in a spiral of poverty. Starting from this analysis, Marx later used it as evidence of the need to abolish the individual proprietorial rights at the heart of Hegel's “civil society” and socialise the means of production. Hegel, however, did not draw this conclusion. His conception of the exchange contract as a form of recognition that played an essential role within the state's capacity to provide the conditions for the existence of rational and free-willing subjects would certainly prevent such a move. Rather, the economy was to be contained within an over-arching institutional framework of the state, and its social effects offset by welfarist state intervention.

Bibliography

Collected Works:

- *Gesammelte Werke*, Rheinisch-Westfälischen Akademie der Wissenschaften, ed., (Hamburg: Felix Meiner Verlag, 1968-).
- *Werke in zwanzig Bänden*, Moldenhauer, Eva and Michel, Karl Markus, ed., (Frankfurt am Main: Suhrkamp Verlag, 1971).

English Translations of Key Texts:

- *Early Theological Writings*, trans. T. M. Knox, (Chicago: Chicago University Press, 1948).
- *The Difference Between Fichte's and Schelling's System of Philosophy*, trans. H. S. Harris and W. Cerf, (Albany: State University of New York Press, 1977).
- *Phenomenology of Spirit*, trans. A. V. Miller, (Oxford: Oxford University Press, 1977).
- *Hegel's Science of Logic*, trans. A. V. Miller, (London: Allen and Unwin, 1969).
- *The Encyclopedia Logic: Part 1 of the Encyclopaedia of Philosophical Sciences*, trans. T. F. Geraets, W. A. Suchting, and H. S. Harris, (Indianapolis: Hackett, 1991).
- *Philosophy of Nature (Part Three of the Encyclopaedia of Philosophical Sciences*), trans. Michael John Perry, 3 vols, (London: George Allen and Unwin, 1970).
- *Hegel's Philosophy of Mind: Being Part Three of the Encyclopaedia of Philosophical Sciences*, trans. William Wallace, (Oxford: Clarendon Press, 1971).
- *Elements of the Philosophy of Right*, ed. Allen W. Wood, trans. H. B. Nisbet, (Cambridge:

Cambridge University Press, 1991).
- *Political Writings*, ed. Laurence Dickey and H. B. Nisbet, trans. H. B Nisbet, (Cambridge: Cambridge University Press, 1999).

Secondary Literature

- Avineri, Shlomo, *Hegel's Theory of the Modern State*, (Cambridge: Cambridge University Press, 1972).
- Beiser, Frederick C., *The Cambridge Companion to Hegel*, (Cambridge: Cambridge University Press, 1993).
- Brandom, Robert B., *Making It Explicit* (Cambridge, Mass.: Harvard University Press, 1994).
- Crites, Stephen, *Dialectic and Gospel in the Development of Hegel's Thinking*, (University Park: Pennsylvania State University Press, 1998).
- Forster, Michael N., *Hegel and Skepticism*, (Cambridge, Mass.: Harvard University Press, 1989).
- Forster, Michael N., *Hegel's Idea of a Phenomenology of Spirit*, (Chicago: University of Chicago Press, 1998).
- Gadamer, Hans-Georg, *Hegel's Dialectic: Five Hermeneutical Studies*, trans. P. Christopher Smith, (New Haven: Yale University Press, 1976).
- Harris, H. S., *Hegel's Development: Toward the Sunlight 1770-1801*, (Oxford: Clarendon Press, 1972).
- Harris, H. S., *Hegel's Development II: Night Thoughts (Jena 1801-6)*, (Oxford: Oxford University Press, 1983).
- Harris, H. S., *Hegel's Ladder*, 2 vols, (Indianapolis: Hackett, 1997).
- Horstmann, Rolf-Peter, *Ontologie und Relationen: Hegel, Bradley, Russell und die Kontroverse über interne und externe Beziehungen*, (Hain: Athenäum, 1984).
- Hösle, Vittorio, *Hegels System: Der Idealismus der Subjectivität und das Problem der Intersubjectivität*, 2 vols, (Hamburg: Meiner Verlag, 1987).
- Houlgate, Stephen, *Freedom, Truth and History: An Introduction to Hegel's Philosophy*, (London and New York: Routledge, 1991).
- Kant, Immanuel. *Critique of Pure Reason*, trans. N. Kemp-Smith, (London: Macmillan, 1929).
- Kojève, Alexandre, *Introduction to the Reading of Hegel*, ed. Allan Bloom, trans. J. H. Nichols, Jr, (New York: Basic Books, 1969).
- Lukács, Georg, *The Young Hegel*, trans. R. Livingston, (London: Merlin Press, 1975).
- McDowell, John, *Mind and World*, (Cambridge, Mass.: Harvard University Press, 1994).
- Neuhouser, Frederick, *Foundations of Hegel's Social Theory: Actualizing Freedom*, (Cambridge, Mass.: Harvard University Press, 2000).
- Pelczynski, Z. A. (ed.), *The State and Civil Society: Studies in Hegel's Political Philosophy*, (Cambridge: Cambridge University Press, 1984).
- Pinkard, Terry, *Hegel's Phenomenology: The Sociality of Reason*, (Cambridge: Cambridge University Press, 1994).
- Pinkard, Terry, *Hegel: A Biography*, (Cambridge: Cambridge University Press, 2000).
- Pippin, Robert B., *Hegel's Idealism: The Satisfactions of Self-Consciousness*, (Cambridge: Cambridge University Press, 1989).

- Pippin, Robert B., *Idealism as Modernism: Hegelian Variations*, (Cambridge: Cambridge University Press, 1997).
- Redding, Paul, *Hegel's Hermeneutics*, (Ithaca: Cornell University Press, 1996).
- Siep, Ludwig, *Anerkennung als Prinzip der praktische Philosophie: Untersuchungen zu Hegels Jenaer Philosophie des Geistes*, (Freiburg: Karl Alber Verlag, 1979).
- Stern, Robert, *Hegel, Kant and the Structure of the Object*, (London: Routledge, 1990).
- Stern, Robert, ed., *G. W. F. Hegel: Critical Assessments*, 4 vols, (London: Routledge, 1993).
- Stern, Robert, *Routledge Philosophy Guidebook to Hegel and the Phenomenology of Spirit*, (London: Routledge, 2002).
- Taylor, Charles, *Hegel*, (Cambridge: Cambridge University Press, 1975).
- Westphal, Kenneth R., *Hegel's Epistemological Realism: A Study of the Aim and Method of Hegel's 'Phenomenology of Spirit'*, (Dordrecht: Kluwer, 1989).
- Williams, Robert R., *Recognition: Fichte and Hegel on the Other* , (Albany: State University of New York Press, 1992).
- Williams, Robert R., *Hegel's Ethics of Recognition*, (Berkeley: University of California Press, 1997).
- Wood, Allen W., *Hegel's Ethical Thought*, (Cambridge: Cambridge University Press, 1990).

Other Internet Resources

- Hegel Society of America Home Page
- Extensive Bibliography by Andrew Chitty

Related Entries

Fichte, Johann Gottlieb | Hölderlin, Johann Christian Friedrich | Jacobi, Friedrich Heinrich | Kant, Immanuel | Marxism | Schelling, Friedrich Wilhelm Joseph von

Paul Redding
paul.redding@philosophy.usyd.edu.au

First published: February 13, 1997
Content last modified: May 20, 2002

Johann Gottlieb Fichte

Inspired by his reading of Kant, Johann Gottlieb Fichte (1762 - 1814) developed during the final decade of the eighteenth century a radically revised and rigorously systematic version of transcendental idealism, which he called *Wissenschaftslehre* of "Doctrine of Scientific Knowledge." Perhaps the most characteristic, as well as most controversial, feature of the *Wissenschaftslehre* (at least in its earlier and most influential version) is Fichte's effort to ground his entire system upon the bare concept of subjectivity, or, as Fichte expressed it, the "pure I." During his career at the University of Jena (1794-1799) Fichte erected upon this foundation an elaborate transcendental system that embraced the philosophy of science, ethics, philosophy of law or "right." and philosophy of religion.

1. Life and Work

Fichte was born May 19, 1762 in the village of Rammenau in the Oberlausitz area of Saxony. He was the eldest son in a family of poor and pious ribbon weavers. His extraordinary intellectual talent soon

brought him to the attention of a local baron, who sponsored his education, first in the home of a local pastor, then at the famous Pforta boarding school, and finally at the universities of Jena and Leipzig. With the death of his patron, Fichte was forced to discontinue his studies and seek his livelihood as a private tutor, a profession he quickly came to detest.

Following a lengthy sojourn in Zurich, were he met his future wife, Johanna Rahn, Fichte returned to Leipzig with the intention of pursuing a literary career. When his projects failed, he was again forced to survive as a tutor. It was in this capacity that he began giving lessons on the Kantian philosophy in the summer of 1790. This first encounter with Kant's writings produced what Fichte himself described as a "revolution" in his manner of thinking. Whereas he had formally been torn between, on the one hand, a practical commitment to the moral improvement of humanity and, on the other, a theoretical commitment to "intelligible fatalism," he found in the Critical philosophy a way of reconciling his "head" and "heart" in a system that could meet the highest intellectual standards without requiring him to sacrifice his belief in human freedom.

Fichte eventually made his way to Königsberg, where he lived for a few months. After a disappointing interview with Kant, he resolved to demonstrate his mastery of the latter's philosophy by writing a treatise on a theme as yet unaddressed by Kant: namely, the question of the compatibility of the Critical philosophy with any concept of divine revelation. In a few weeks Fichte composed a remarkable manuscript in which he concluded that the only revelation consistent with the Critical philosophy is the moral law itself. Kant was sufficiently impressed by the talent of this unknown and impoverished young man to offer to arrange for the publication of Fichte's manuscript, which was published by Kant's own publisher in 1792 under the title *Attempt at a Critique of All Revelation*. The first edition of this work, however, for reasons that have never been satisfactorily explained, appeared without the author's name and preface and was quickly and widely hailed as a work by Kant himself. When the true identity of its author was revealed, Fichte was immediately catapulted from total obscurity to philosophical celebrity.

Meanwhile, Fichte was once again employed as a private tutor, this time on an estate near Danzig, where he wrote several, anonymously published political tracts. The first of these was published in 1793 with the provocative title *Reclamation of the Freedom of Thought from the Princes of Europe, who have hitherto Suppressed it*. In the summer of 1793 Fichte returned to Zurich where he married his fiancé and oversaw the publication of the first two installments of his spirited *Contribution to the Rectification of the Public's Judgment of the French Revolution* (1793 and 1794). In this work he not only defended the principles (if not all the practices) of the French revolutionaries, but also attempted to outline his own democratic view of legitimate state authority and insisted on the right of revolution. Despite the fact that these political writings were published anonymously, the author's identity was widely known, and Fichte thereby acquired a reputation, not wholly deserved, as a radical "Jacobin."

Following the completion of these projects, Fichte devoted his time in Zurich to concentrated efforts to rethink and to revise his own philosophical position. While maintaining his allegiance to the new Critical or Kantian philosophy, Fichte was powerfully impressed by the efforts of K. L. Reinhold to provide the Critical philosophy with a new, more secure "foundation" and to base the entire system upon a single "first principle." At the same time, he became acquainted with the works of two authors who were

engaged in skeptical attacks upon the philosophies of both Kant and Reinhold: Solomon Maimon and G. E. Schulze ("Aenesidemus"). It was the need to respond to the sharp criticisms of these authors that eventually led Fichte to construct his own, unique version of transcendental idealism, for which, in the spring of 1794, he eventually coined the name *Wissenschaftslehre* ("Doctrine of Science" or "Theory of Scientific Knowledge").

It was at precisely this moment that he received an invitation to assume the recently vacated chair of Critical Philosophy at the University of Jena, which was rapidly emerging as the capital of the new German philosophy. Fichte arrived in Jena in May of 1794, and enjoyed tremendous popular success there for the next six years, during which time he laid the foundations and developed the first systematic articulations of his new system. Even as he was engaged in this immense theoretical labor, he also tried to address a larger, popular audience and also threw himself into various practical efforts to reform university life. As one bemused colleague observed, "his is a restless spirit; he thirsts for some opportunity to act in the world. Fichte wants to employ his philosophy to guide the spirit of his age." Indeed, a passionate desire to "have an effect" upon his own age remained a central feature of Fichte's character, most notably expressed a decade later in his celebrated *Addresses to the German Nation,* delivered in Berlin in 1806 during the French occupation. In Jena, this same desire is reflected in the enormously popular series of public lectures on "Morality for Scholars," which he began to deliver immediately upon his arrival in Jena. The first five of these lectures were published in 1794 under the title *Some Lectures concerning the Scholar's Vocation*.

Though Fichte has already hinted at his new philosophical position in his 1794 review of G. E,. Schulze's *Aenesidemus,* the first full-scale announcement of the same came in a short manifesto that he published as a means of introducing himself to his students and colleagues and attracting listeners to his lectures. (As an "extraordinary professor," Fichte was largely dependent upon fees paid by students attending his "private" lectures.) This manifesto, *Concerning the Concept of the Wissenschaftslehre* (1794), articulated some of the basic ideas of the new philosophy, but it mainly focused upon questions of systematic form and the foundations of the same.

Fichte's first truly systematic work was his *Foundation of the Entire Wissenschaftslehre* (1794/95). As the title implies, this work, which remains to this day Fichte's best-known philosophical treatise, was not meant to be a presentation of his entire system, but only of the rudiments or first principles of the same. In fact, Fichte had not originally intended to publish this work at all, which was written less than a year after his first tentative efforts to articulate for himself his new conception of transcendental philosophy. The *Foundation* was originally intended to be distributed, in fascicles, to students attending his private lectures during his first two semesters at Jena, where the printed sheets could be subjected to analysis and questions and supplemented with oral explanations. Because of the great interest in Fichte's new philosophy, however, he soon authorized a public edition of the same, in two volumes. Parts I and II of the *Foundation* were published in 1794 and Part II in 1795. In 1795 he also published a substantial supplement to the *Foundation,* under the title *Outline of the Distinctive Character of the Wissenschaftslehre with Respect to the Theoretical Faculty*. The title pages of all three of these publications, however, still stipulated that they were intended only as "a manuscript for the use of his listeners." (When, in 1802, Fichte issued a second, one-volume edition of the *Foundation* and *Outline*, in

1801, this subtitle was dropped.)

Dissatisfied with many features of his initial presentation of the "foundational" portion of his system and surprised and shocked by the virtually universal misunderstanding of his published *Foundation*, Fichte immediately set to work on an entirely new exposition of the same, which he repeated three times in his private lectures on "The Foundations of Transcendental Philosophy (*Wissenschaftslehre) nova methodo*" (1796/76, 1797/98, 1798/99). Though he intended to revise these lectures for serial publication under the title *An Attempt a New Presentation of the Wissenschaftslehre* in the *Philosophisches Journal einer Gesellschaft Teutscher Gelehrten*, of which he himself was by then co-editor, only the two Introductions to and the first chapter of this "New Presentation" ever appeared (1797/98).

Even as he was thoroughly revising his presentation of the foundational portion of his system, Fichte was simultaneously engaged in elaborating the various subdivisions or systematic branches of the same. As was his custom, he did this first in his private lectures and then in published texts based upon the same. The first such extension was into the realm of philosophy of law and social philosophy, which resulted in the publication of *Foundation of Natural Right in accordance with the Principles of the Wissenschaftslehre* (published in two volumes in 1796 and 1797). The second was into the realm of moral philosophy, which resulted in the publication of the *System of Ethical Theory in accordance with the Principles of the Wissenschaftslehre* (1798). At this point, Fichte's own plans called for him to extend his system into the realm of philosophy of religion. He announced lectures on this topic for the Spring Semester of 1799, but before he could commence these lectures, his career at Jena had come to an abrupt and unhappy conclusion in the wake of the so-called "Atheism Controversy" of 1798/99.

In 1798 Fichte published in his *Philosophical Journal* a brief essay "On the Basis of Our Belief in a Divine Governance of the World," in which he attempted to sketch some of his preliminary ideas on the topic indicated in the title and simultaneously to give the first clear public hint of the character of a philosophy of religion "in accordance with the principles of the *Wissenschaftslehre*." The occasion for this essay was another essay, published in the same issue of the *Philosophical Journal*, by K. L. Forberg. As it happened, these two essays provoked an anonymous author to publish a pamphlet charging the authors with atheism and demanding Fichte's dismissal from his post at Jena. The matter quickly escalated into a major public controversy which eventually led to the official suppression of the offending issue of the journal and to public threats by various German princes to prevent their students from enrolling at the University of Jena. The crisis produced by these actions and the growing number of publications for and against Fichte -- which included an intemperate *Appeal to the Public* by Fichte himself (1799), as well as a more thoughtful response entitled "From a Private Letter" (1799) - eventually provoked F. H. Jacobi to publish his famous "open letter" to Fichte, in which he equated philosophy in general and Fichte's transcendental philosophy in particular with "nihilism." As the public controversy unfolded, Fichte badly miscalculated his own position and was finally forced to resign his position at Jena and to flee to Berlin, where he arrived in the summer of 1799.

At this point, the Prussian capital had no university of its own, and Fichte was forced to support himself by giving private tutorials and lectures on the *Wissenschaftslehre* and by a new flurry of literary production, increasingly aimed at a large, popular audience. The first of these "popular" writings was a

brilliant presentation of some of the characteristic doctrines and conclusions of Fichte's system, with a strong emphasis upon the moral and religious character of the same. This work, *The Vocation of Man* (1800), which is perhaps Fichte's greatest literary achievements, was intended as an indirect response to Jacobi's public repudiation of the *Wissenschaftslehre*. That same year also saw the publication of a typically bold foray into political economy, *The Closed Commercial State*, in which Fichte propounds a curious blend of socialist political ideas and autarkic economic principles. Defending his philosophy against misunderstanding remained, however, Fichte's chief concern during this period, as is evidenced by the more direct response to Jacobi contained in his poignantly titled *Sun-Clear Report to the Public at Large concerning the Actual Character of the latest Philosophy: An Attempt to Force the Reader to Understand* (1801).

At the same time that he was addressing the public in this manner, Fichte was becoming ever more deeply engrossed in efforts to rethink and to rearticulate the very foundations of his system, beginning with his private lectures on the *Wissenschaftslehre* of 1801/2, and culminating in the three, radically new versions of the same produced during the year 1804. Indeed, he continued to produce new versions of the *Wissenschaftslehre* right up until his death. However, with the single exception of the extraordinarily condensed (and extraordinarily opaque) *Presentation of the General Outlines of the Wissenschaftslehre* (1810), none of these later versions of the *Wissenschaftslehre* was published during Fichte's lifetime. Some of them appeared, in severely edited form, in the collection of Fichte's *Works* published by his son several decades following his death, but most of them are only now being published for the first time in the critical edition of Fichte's writings produced by the Bavarian Academy of the Sciences. It appears that Fichte was so discouraged by the public reception of the first, 1794/95 presentation of the foundation of his system that he concluded that it was prudent to limit future new presentations of the same to the lecture hall and seminar room, where he could elicit reactions and objections from his listeners and respond immediately with the requisite corrections and clarifications. Be that as it may, Fichte never stopped trying to refine his philosophical insights and to revise his systematic presentation of the same. Their are more than a dozen different full-scale presentations or versions of the *Wissenschaftslehre*, most of which were written after his departure from Jena. "The *Wissenschaftslehre*" is not the name of a book; it is the name of a *system of philosophy*, one capable of being expounded in a variety of different ways.

In 1805 Fichte spend a semester as a professor at the University of Erlangen, but returned to Berlin in the fall of that year. The next year, 1806, he published in rapid succession three popular and well-received books, all of which were based upon earlier series of public lectures that he had delivered in Berlin: *On the Essence of the Scholar* (a reworking of some of the same themes first addressed in the similarly titled lectures of 1794); *The Characteristics of the Present Age* (an attempt to show the implications of his "system of freedom" for a speculative philosophy of history); and *Guide to the Blessed Life* (an eloquent and rather mystical treatise on philosophy and religion). Taken together, these three "popular" works are remarkable blends of speculative profundity and rhetorical eloquence.

With the entry of the French army of occupation into Berlin in 1806, Fichte joined the Prussian government in exile in Königsberg, where he delivered yet another course of lectures on the *Wissenschaftslehre* and wrote an important short book on *Machiavelli as Author* (1807), which defends a form of *Realpolitik* that at least appears to contrast quite starkly with the liberalism and political idealism

of Fichte's earlier political writings. Fichte soon returned to occupied Berlin, however, where, in the winter of 1807/8, he delivered his celebrated *Addresses to the German Nation* (published in 1808). Though these lectures later obtained a place of dubious honor as founding documents in the history of German nationalism, they are mainly concerned with the issue of national identity (and particularly with the relationship between language and nationality) and the question of national education (which is the main topic of the work) -- both understood as means toward a larger, cosmopolitan end.

Fichte had always had a lively interest in pedagogical issues and assumed a leading role in planning the new Prussian university to be established in Berlin (though his own detailed plans for the same were eventually rejected in favor of those put forward by Wilhelm von Humboldt). When the new university finally opened in 1810, Fichte was the first head of the philosophical faculty as well as the first elected rector of the university. His final years saw no diminishment in the pace either of his public activity or of his philosophical efforts. He continued to produce new lectures on the foundations and first principles of his system, as well as new introductory lectures on philosophy in general ("Logic and Philosophy" [1812] and "The Facts of Consciousness" [1813]), political philosophy ("System of the Theory of Right" [1812] and ‘Theory of the State’ [1813]) and ethics ("System of Ethical Theory" [1812]). As presaged perhaps by his earlier book on Machiavelli, these late forays into the domain of practical philosophy betray a far darker view of human nature and defend a more authoritarian view of the state than anything to be found in Fichte's earlier, published writings on these subject.

In 1813 Fichte canceled his lectures so that his students could enlist in the "War of Liberation" against Napoleon, of which Fichte himself proved to be an indirect casualty. From his wife, who was serving as a volunteer nurse in a Berlin military hospital, he contracted a fatal infection of which he died on January 29, 1814.

2. Fichte's Philosophical Project

The primary task of Fichte's system of philosophy (the *Wissenschaftslehre*) is to reconcile freedom with necessity, or, more specifically, to explain how freely willing, morally responsible agents can at the same time be considered part of a world of causally conditioned material objects in space and time. Fichte's strategy for answering this question -- at least in his early writings, which are the ones upon which his historical reputation as a philosopher has (at least until recently) been grounded and hence are the ones to be expounded here -- was to begin simply with the ungrounded assertion of the subjective spontaneity and freedom (infinity) of the I and then to proceed to a transcendental derivation of objective necessity and limitation (finitude) as a condition necessary for the possibility of the former. This is the meaning of his description, in his "First Introduction to the *Wissenschaftslehre*," of philosophy's task as that of "displaying the foundation of experience" or "explaining the basis of the system of representations accompanied by a feeling of necessity." Fichte derived this conception of the task and strategy of philosophy from his study of Kant, and no matter how far his own system seemed to diverge from "the letter" of the Critical philosophy, Fichte always maintained that it remained true to "the spirit" of the same. Central to this "spirit," for Fichte, is an uncompromising insistence upon the practical certainty of human freedom and a thoroughgoing commitment to the task of providing a transcendental account of

ordinary experience that could explain the objectivity and necessity of theoretical reason (cognition) in a manner consistent with the practical affirmation of human liberty. Though Fichte attributed the discovery of this task to Kant, he believed that it was first accomplished successfully only in the *Wissenschaftslehre*, which he therefore described as the first "system of human freedom."

In an effort to clarify the task and method of transcendental philosophy, Fichte insisted upon the sharp distinction between the "standpoint" of natural consciousness (which it is the task of philosophy to "derive," and hence to "explain") and that of transcendental reflection, which is the standpoint required of the philosopher. He thus insisted that there is no conflict between transcendental idealism and the commonsense realism of everyday life. On the contrary, the whole point of the former is to demonstrate the necessity and unavailability of the latter.

However "Kantian" in spirit Fichte's enterprise might have been, he was at the same time all too keenly aware of what he considered to be certain glaring weaknesses and inadequacies in Kant's own execution of this project. Taking to heart the criticisms of such contemporaries as F. H. Jacobi, Salomon Maimon, and G. E. Schulze, Fichte propounded a radically revised version of the Critical philosophy. First of all, he argued that the very concept of a "thing in itself," understood as a mind-independent, external "cause" of sensations, is indefensible on Critical grounds. In addition, he maintained that Kant's denial of the possibility of "intellectual intuition," though certainly justified as a denial of the possibility of any non-sensory awareness of external objects, is nevertheless difficult to reconcile with certain other Kantian doctrines regarding the I's immediate presence to itself both as a (theoretically) cognizing subject (the doctrine of the transcendental apperception) and as a (practically) striving moral agent (the doctrine of the categorical imperative).

His study of the writings of K. L. Reinhold convinced Fichte that the systematic unity of the Critical philosophy -- specifically, the unity of theoretical and practical reason, of the First and Second *Critiques* -- was insufficiently evident in Kant's own presentation of his philosophy and that the most promising way to display the unity in question would be to provide both theoretical and practical philosophy with a common foundation. The first task for philosophy, Fichte therefore concluded, is to discover a single, self-evident starting point or first principle from which one could then somehow "derive" both theoretical and practical philosophy, which is to say, our experience of ourselves as finite cognizers and as finite agents. Not only would such a strategy guarantee the systematic unity of philosophy itself, but, more importantly, it would also display what Kant hinted at but never demonstrated: viz., the underlying unity of reason itself.

Since it is a central task of philosophy, so construed, to establish the very possibility of any knowledge or science (*Wissenschaft*) whatsoever, Fichte proposed to replace the disputed term "philosophy" (or "love of wisdom") with the new term *Wissenschaftslehre* or ‘Theory or Science’ - a name intended to highlight the distinctively "second order" character of philosophical reflection. Though Fichte's proposal never caught on as a general name for what was once called "philosophy," it did become the universally acknowledged name for his own distinctive version of transcendental idealism. Here again, it important to keep in mind that "*Wissenschaftslehre"* is not the name of any particular Fichtean treatise, but is instead the general name for his entire system or project - an allegedly all-encompassing system that

consists of a number of interrelated parts or systematic subdisciplines and an overarching project that could and would be expounded in a series of radically different presentations, employing a bewildering variety of systematic vocabularies.

3. The Starting Point of the Jena *Wissenschaftslehre*

In order to construct any genuine philosophy of freedom, maintained Fichte, the reality of freedom itself must simply be presupposed and thus treated as an incontrovertible "fact of reason" in the Kantian sense. This, of course, is not to deny the possibility of raising skeptical, theoretically grounded objections to such claims; on the contrary, it was the very impossibility of any theoretically satisfactory refutation of skepticism concerning the reality of freedom that led Fichte to affirm the inescapable "primacy of the practical" with respect to the selection of one's philosophical starting point.

To the extent that any proposed first principle of philosophy is supposed to be the first principle of all knowledge and hence of all argument, it clearly cannot be derived from any higher principle and hence cannot be established by any sort of reasoning. Furthermore, Fichte maintained that there are two and only two possible starting points for the philosophical project of "explaining" experience: namely, the concept of pure selfhood (which Fichte associated with pure freedom) and that of pure thinghood (which Fichte associated with utter necessity) -- neither of which can be warranted, qua philosophical starting point, by a direct appeal to experience, and each of which can be arrived at only by a self-conscious act of philosophical *abstraction* from ordinary experience (within which freedom and necessity, subject and object, are invariably joined as well as distinguished).

The two rival philosophical strategies made possible by these opposed starting points are unforgettably limned by Fichte in his two 1797 "Introductions to the *Wissenschaftslehre*," in which he characterizes the sort of philosophy that begins with the pure I as "idealism" and that which begins with the thing in itself as "dogmatism." Since, according to Fichte's earlier argument in *Concerning the Concept of the Wissenschaftslehre*, a unified system of philosophy can have one and only one first principle, and since there are two and only two possible first principles, then it follows that no "mixed" system of idealism/dogmatism is possible. Moreover, since dogmatism, as understood by Fichte, unavoidably implies a strict form of determinism or "intelligible fatalism," whereas idealism is, from the start, committed to the reality of human freedom, it is also practically impossible to reach any sort of "compromise" between two such radically opposed systems.

Though Fichte conceded that neither dogmatism nor idealism could directly refute its opposite and thus recognized that the choice between philosophical starting points could never be resolved on purely theoretical grounds, he nevertheless denied that any dogmatic system, that is to say, any system that commences with the concept of sheer objectivity, could ever succeed in accomplishing what was required of all philosophy. Dogmatism, he argued, could never provide a transcendental deduction of ordinary consciousness, for, in order to accomplish this, it would have to make an illicit leap from the realm of

"things" to that of mental events or "representations" [*Vorstellungen*]. Idealism, in contrast, at least when correctly understood as the kind of Critical idealism that demonstrates that the intellect itself most operate in accordance with certain necessary laws, can -- at least in principle -- accomplish the prescribed task of philosophy and explain our experience of objects ("representations accompanied by a feeling of necessity") in terms of the necessary operations of the intellect itself, and thus without having to make an illicit appeal to things in themselves. To be sure, one cannot decide in advance whether or not any such deduction of experience from the mere concept of free self-consciousness is *actually* possible. This, Fichte conceded, is something that can be decided only after the construction of the system in question. Until then, it remains a mere *hypothesis* that the principle of human freedom, for all of its practical certainty, is also the proper starting point for a transcendental account of objective experience.

It must be granted that the truth of the *Wissenschaftslehre's* starting point cannot be established by any philosophical means, including its utility as a philosophical first principle. On the contrary -- and this is one of Fichte's most characteristic and controversial claims -- one already has to be convinced, on wholly extra-philosophical grounds, of the reality of one's own freedom *before* one can enter into the chain of deductions and arguments that constitute the *Wissenschaftslehre*. This is the meaning of Fichte's oft-cited assertion that "the kind of philosophy one chooses depends upon the kind of person one is." The only compelling reason why the transcendental idealist comes to a stop with -- and thus begins his system with -- the proposition that "the I freely posits itself" is therefore not because he is unable to entertain theoretical doubts on this score nor because he is simply unable to continue the process of reflective abstraction. Instead, he appeals to a principle eloquently expressed by Fichte in his essay "On the Basis of Our Belief in a Moral Governance of the World," namely: "I cannot go beyond this standpoint because I am not permitted to do so." It is precisely because the categorical imperative is in this way invoked to secure the first principle of his entire system that Fichte felt entitled to make the rather startling claim that the *Wissenschaftslehre* is the only system of philosophy that "accords with duty."

4. Systematic Overview of the Jena *Wissenschaftslehre*

4.1 The "Foundation"

The published presentation of the first principles of the Jena *Wissenschaftslehre* commences with the proposition, "the I posits itself"; more specifically, "the I posits itself as an I." Since this activity of "self-positing" is taken to be the fundamental feature of I-hood in general, the first principle asserts that "the I posits itself as self-positing." Unfortunately, this starting point is somewhat obscured in Part I of the *Foundation of the Entire Wissenschaftslehre* by a difficult and somewhat forced attempt on Fichte's part to connect this starting point to the logical law of identity, as well as by the introduction of two additional "first principles," corresponding to the logical laws of non-contradiction and sufficient reason. (Significantly, this distraction is eliminated entirely in the 1796/99 *Wissenschaftslehre nova methodo*, which begins with the simple "postulate" or "summons" to the reader: "think the I, and observe what is involved in doing this.")

"To posit" (*setzen*) means simply "to be aware of," "to reflect upon," or "to be conscious of"; this term does not imply that the I must simply "creates" its objects of consciousness. The principle in question simply states that the essence of I-hood lies in the assertion of ones own self-identity, i.e., that consciousness presupposes self-consciousness (the Kantian "I think," which must, at least in principle, be able to accompany all our representations). Such immediate self-identify, however, cannot be understood as a psychological "fact," no matter how privileged, nor as an "action" or "accident" of some previously existing substance or being. To be sure, it is an "action" of the I, but one that is identical with the very existence of the same. In Fichte's technical terminology, the original unity of self-consciousness is to be understood as both an action and as the product of the same: as a *Tathandlung* or "fact/act," a unity that is presupposed by and contained within every fact and every act of empirical consciousness, though it never appears as such therein.

This same "identity in difference" of original self-consciousness might also be described as an "intellectual intuition," inasmuch as it involves the *immediate* presence of the I to itself, prior to and independently of any sensory content. To be sure, such an "intellectual intuition" never occurs, as such, within empirical consciousness; instead, it must simply be presupposed (that is, "posited") in order to explain the possibility of actual consciousness, within which subject and object are always already distinguished. The occurrence of such an original intellectual intuition is itself inferred, not intuited.

Unfortunately, Fichte confuses matters by sometimes using the term "inner" or "intellectual intuition" to designate something else entirely: namely, the act of *philosophical reflection* or purified self-observation through which the philosopher becomes conscious of the transcendental conditions for the possibility of ordinary experience -- among which, of course, is the occurrence of the "original" intellectual intuition as a *Tathandlung*. On other occasions, he employs the term "intellectual intuition" in yet another sense: namely, to designate our direct, practical awareness within everyday life of our moral obligations (categorical imperative qua "*real* intellectual intuition"). Given the subsequent abuse of this term by Schelling and the romantics, as well as the confusion that one sometimes finds among expositors of Fichte on this issue, it is crucial to recognize systematic ambiguity of the term "intellectual intuition" in Fichte's own writings.

A fundamental corollary of Fichte's understanding of I-hood (*Ichheit*) as a kind of *fact/act* is his denial that the I is originally any sort of "thing" or "substance." Instead, the I is simply what it posits itself to be, and thus its "being" is, so to speak, a consequence of its self-positing, or rather, is co-terminus with the same. The first principle of the Jena *Wissenschaftslehre* is thus equally "practical" and "theoretical," insofar as the act described by this principle is a "doing" as well as a "knowing," a deed as well as a cognition. Thus the problematic unity of theoretical and practical reason is guaranteed from the start, inasmuch as this very unity is a condition for the possibility of self-consciousness.

After establishing the first principle and conceiving the act expressed therein, the philosophical task is then to discover what other acts must necessarily occur as conditions for the possibility of the original, "simply posited," first act and then to do the same for each of these successively discovered acts (or the theorems in which they are formulated). By continuing in this manner, one will, according to Fichte,

finally arrive at a complete deduction of the a priori structure of ordinary experience or, what amount to the same thing, a complete inventory of the "original acts of the mind." This is precisely the task of the first or "foundational" portion of the Jena system.

Just as we are never directly aware of the original act of self-positing with which the system commences, so are we also unaware -- except, of course, from the artificial standpoint of philosophical reflection -- of each of these additional "necessary but unconscious" acts that are derived as conditions necessary for the possibility of the originally posited act of self-positing. Furthermore, though we must, due to the discursive character of reflection itself, distinguish each of these acts from the others that it is conditioned by and that are, in turn, conditioned by it, none of these individual acts actually occurs in isolation from all of the others. Transcendental philosophy is thus an effort to *analyze* what is in fact the single, *synthetic* act through which the I posits for itself both itself and its world, thereby becoming aware in a single moment of both its freedom and its limitations, its infinity and its finitude. The result of such an analysis is the recognition that, although "the I simply posits itself," its freedom is never "absolute" or "unlimited"; instead, freedom proves to be conceivable -- and hence the I itself proves to be possible -- only as limited and finite. Despite widespread misunderstanding of this point, the *Wissenschaftslehre* is not a theory of the absolute I. Instead, the conclusion of both the *Foundation of the Entire Wissenschaftslehre* and of the *Wissenschaftslehre nova methodo* is that the "absolute I" is a mere abstraction and that the only sort of I that can actually exist or act is a *finite, empirical, embodied, individual self*.

The I must posit itself in order to be an I at all; but it can posit itself only insofar as it posits itself as limited (and hence divided against itself, inasmuch as it also posits itself as unlimited or "absolute"). Moreover, it cannot even posit for itself its own limitations, in the sense of producing or creating these limits. The finite I (the intellect) cannot be the ground of its own passivity. Instead, according to Fichte's analysis, if the I is to posit itself at all, it must simply *discover* itself to be limited, a discovery that Fichte characterizes as a ‘check’ or *Anstoß* to the free, practical activity of the I. Such an original limitation of the I is, however, a limit for the I only insofar as the I posits it as such. I does this, according to Fichte's analysis, by positing its own limitation, first, as a mere "feeling," then as a "sensation," then as an "intuition" of a thing, and finally as a "concept." The *Anstoß* thus provides the essential occasion or impetus that first sets in motion the entire complex train of activities that finally result in our conscious experience both of ourselves as empirical individuals and of a world of spatio-temporal material objects.

Though this doctrine of the *Anstoß* may seem to play a role in Fichte's philosophy not unlike that which has sometimes been assigned to the thing in itself in the Kantian system, the fundamental difference is this: the *Anstoß* is not something foreign to the I. Instead, it denotes the I's original encounter with its own finitude. Rather than claim that the Not-I is the cause or ground of the *Anstoß*, Fichte argues that the former is posited by the I precisely in order to "explain" to itself the latter, that is, in order to become conscious of the same. Though the *Wissenschaftslehre* demonstrates that such an *Anstoß* must occur if self-consciousness is to be actual, transcendental philosophy itself is quite unable to deduce or to explain the actual occurrence of such an *Anstoß* -- except as a condition for the possibility of consciousness. Accordingly, there are strict limits to what can be expected from any a priori deduction of experience. According to Fichte, transcendental philosophy can explain, for example, why the world has a spatio-

temporal character and a causal structure, but it can never explain why objects have the particular sensible properties they happen to have or why I am this determinate individual rather than another. This is something that the I simply has to discover at the same time that it discovers its own freedom, and indeed, as a condition for the latter. (It must be admitted, however, that Fichte's own ambitious descriptions of his project sometimes obscure the essential limits of the same and that he sometimes gives his readers the false impression that the *Wissenschaftslehre* proposes to provide a complete a priori deduction of all the empirical details of experience. This however is certainly not the case.)

Despite this important stricture on the scope of transcendental philosophy, there remains much that can be demonstrated within the foundational portion of the *Wissenschaftslehre*. For example, it can be show that the I could not become conscious of its own limits in the manner required for the possibility of any self-consciousness unless it also possessed an original and spontaneous ability to synthesize the finite and the infinite. In this sense, the *Wissenschaftslehre* deduces the power of productive imagination as an original power of the mind. Similarly, it can be shown that the I could not be "checked" in the manner required for the possibility of consciousness unless it possessed, in addition to its original "theoretical" power of productive imaginative, an equally original "practical" power of sheer willing, which, once "checked," is immediately converted into a capacity for endless striving. The foundational portion of the *Wissenschaftslehre* thus also includes a deduction of the categorical imperative (albeit in a particularly abstract and morally empty form) and of the practical power of the I. For Fichte, therefore, "the primacy of the practical" means not simply that philosophy must recognize a certain autonomous sphere within which practical reason is efficacious and practical considerations are appropriate; instead, it implies something much stronger: namely, the recognition that, as Fichte puts it, "the practical power is the innermost root of the I" and thus that "our freedom itself is a theoretical determining principle of our world." The *Wissenschaftslehre* as a whole can therefore be described as a massive effort to demonstrate that reason could not be theoretical if it were not also practical -- at the same time, to be sure, that also demonstrates that reason could not be practical if were not also theoretical.

Freedom, according to Fichte's argument, is possible and actual only within the context of limitation and necessity, and thus it is never "absolute," but always limited and finite. On the other hand, just as surely as a free subject must posit its freedom "absolutely" -- that is to say, ‘purely and simply’ (*schlechthin*) and "for no reason" whatsoever -- so must it never identify itself with any determinate or limited state of its own being. On the contrary, a finite free self must constantly strive to transform both the natural and the human worlds in accordance with its own freely-posited goals. The sheer unity of the self, which was posited as the starting point of the *Foundations*, is thereby transformed into an *idea* of reason in the Kantian sense: the actual I is always finite and divided against itself, and hence it is always striving for a sheer self-determinacy that it never achieves. Between the original abstraction of pure selfhood as sheer *Tathandlung* and the concluding (necessary) idea of a self that is only what it determines itself to be, in which "is" and "ought" wholly coincide, lies the entire realm of actual consciousness and real human experience.

4.2 Philosophy of Nature

Having established the foundation of his new system, Fichte then turned to the task of constructing upon this foundation a fully-articulated transcendental system, the overall structure of which is most clearly outlined in the concluding section of the transcripts of his lectures on *Wissenschaftslehre nova methodo*. According to this plan, which has no analog in Fichte's later writings, the *Entire Wissenschaftslehre* is to consisting of four, systematically interrelated parts:: (1) first philosophy, which corresponds to the "foundational" portion of the system, as presented in the *Foundation of the entire Wissenschaftslehre* and revised in the lectures on *Wissenschaftslehre nova methodo*; (2)"theoretical philosophy" or "philosophy of nature," (3) "practical philosophy" or ethics (corresponding to the content of the *System of Ethics*); and (4) "philosophy of the postulates," which includes the subdisciplines of "natural law" or "theory of right" (as expounded in the *Foundation of Natural Right*) and philosophy of religion.

By "philosophy of nature," Fichte seems to have had in mind something similar to Kant's *Metaphysical First Principles of Nature*, though Fichte himself devoted very little attention to the execution of such a project. The closest he ever came to developing a philosophy of nature according to transcendental principles is the compressed account of space, time, and matter presented in the *Outline of the Distinctive Character of the Wissenschaftslehre with Respect to the Theoretical Faculty* and the lectures on *Wissenschaftslehre nova methodo*. In neither of these works, however, does he make any effort to distinguish rigorously between the "theoretical" aspect of the foundational portion of his system and a distinctively "theoretical" subdivision of the same ("philosophy of nature"). In fact, a "philosophy of nature in accordance with the principles of the *Wissenschaftslehre*" turns out to me even more modest than Kant's and more closely resembles what later came to be called the philosophy of (natural) science than it does the speculative *Naturphilosophie* of Schelling and Hegel. Indeed, disagreement concerning the compatibility of a rigorously transcendental philosophy with a speculative, a priori "philosophy of nature" was the very issue that precipitated the rift between Fichte and his erstwhile disciple, Schelling. The popular picture of Fichte's attitude toward nature, namely, that he viewed the latter almost entirely from the perspective of human projects, that is, as the necessary realm for moral striving, is therefore very close to the truth.

4.3 Ethics

In contrast to Fichte's rather cursory treatment of purely theoretical philosophy, ethics or "practical philosophy," which analyzes the determinate ways in which willing and acting are determinable by principles of pure reason, constitutes a major portion of the Jena system, and the *System of Ethics* is Fichte's longest single book. Whereas theoretical philosophy explains how the world necessarily *is*, practical philosophy explains how the world *ought* to be, which is to say, how it ought to be *altered* by rational beings. Ethics thus considers the object of consciousness not as something given or even as something constructed by necessary laws of consciousness, but rather as something to be produced by a freely acting subject, consciously striving to establish and to accomplish its own goals and guided only by its *own* self-legislated laws. The specific task of Fichte's ethics is therefore, first of all, to deduce the categorical imperative (in its distinctively moral sense) from the general obligation to determine oneself freely, and, second, to deduce from this the particular obligations that apply to every free and finite rational being.

Like all of Fichte's systematic treatises of the Jena period, *The System of Ethics* begins with a detailed analysis of what is involved in the self-positing of the I. In this case, the focus is upon the necessity that the I posit for itself its own activity or "*efficacy*," and upon a detailed analysis of the conditions for doing this. In this manner Fichte deduces what he calls "the principle of all practical philosophy," viz., that something objective (a being) follows from something subjective (a concept), and hence that the I must ascribe to itself a power of free purposiveness or causality in the sensible world. The I must posit itself as an *embodied will*, and only as such does it "discover" itself at all. From this starting point Fichte then proceeds to a deduction of the principle of morality: namely, that I must think of my freedom as standing under a certain necessary *law* or *categorical imperative*, which Fichte calls "the law of self-sufficiency" or "autonomy," and that I *ought* always to determine my freedom in accordance with this law. This, therefore, is the task of the philosophical science of "ethics," as understood by Fichte: to provide an a priori deduction of our moral nature in general and of our specific duties as human beings.

Viewed from the perspective of practical philosophy, the world really is nothing more than what Fichte once described as "the material of our duty made sensible," which is precisely the viewpoint adopted by the morally engaged, practically striving subject. On the other hand, this is not the only way the world can be viewed, and, more specifically, it is not the only way in which it is construed by transcendental philosophy. For this reason it is somewhat misleading to characterized the *Wissenschaftslehre* as a whole as a system of "ethical idealism." As noted above, Fichte certainly does succeed in constructing an account of consciousness that fully integrates the imperatives and activities of practical reason into the very structure of the latter, but this integration is always balanced by a recognition of the constitutive role of theoretical reason and of the sheer, contingent "giveness" of the I's original determinacy (doctrine of the *Anstoß*).

4.4 Philosophy of Law (*Recht*)

The final portion of the Jena system is devoted to "the philosophy of the postulates," a discipline that Fichte conceived of as occupying the middle ground between purely theoretical and purely practical philosophy. In this portion of the system the world is considered neither as it simply is nor as it simply ought to be; instead, the moral world is itself considered from the perspective of the natural world (that is, one considers the postulates that theoretical reason addresses to the practical realm) or else, alternatively, the natural world is considered from the perspective of the moral law (that is, one considers the postulates that practical reason addresses to the realm of theory). The first of these perspectives is that of juridical philosophy or philosophy of law, or what Fichte calls the "doctrine of right" (*Rechtslehre*); the latter is that of the philosophy of religion.

Fichte's philosophy of right (or of "natural law"), as expounded in his *Foundation of Natural Right*, is one of the most original and influential portions of the Jena *Wissenschaftslehre*. Written prior to Kant's treatment of the same topic (in Part One of the *Metaphysics of* Morals), Fichte's philosophy of right is notable, first of all, because of the way in which it distinguishes sharply between the realm of ethics and that of "right" and tries to develop a complete theory of the latter (a "theory of justice") without appealing to the categorical imperative or the moral law, and secondly, because of the inclusion within this theory

of a thoroughly original "deduction" of the social character of human beings.

Fichte's transcendental account of natural right proceeds from the general principle that the I must posit itself as an *individual* in order to posit itself at all, and that in order to posit itself as an individual it must recognize itself as "summoned" or "solicited" by *another* free individual -- summoned, that is, to limit its own freedom out of respect for that of the freedom of the other. The same condition applies, of course, to the other; hence, mutual recognition of rational individuals turns out to be condition necessary for the possibility of I-hood in general. This a priori deduction of intersubjectivity is so central to the conception of selfhood developed in the Jena *Wissenschaftslehre* that Fichte, in his lectures on *Wissenschaftslehre nova methodo*, incorporated it into his revised presentation of the very foundations of his system, where the "summons" takes its place alongside "original feeling" (which takes the place of the earlier "check") as both a *limit* upon the absolute freedom of the I and a condition for the positing of the same.

The specific task of Fichte's theory of right is to consider the specific ways in which the freedom of each individual must be restricted in order that several individuals can live together with the maximum amount of mutual freedom, and it derives its a priori concepts of the laws of social interaction entirely from the sheer concept of an individual I, as conditions for the possibility of the latter. Fichte's concept of right therefore obtains its binding force not from the ethical law, but rather from the general laws of thinking and from enlightened self-interest, and the force of such considerations is hypothetical rather than categorical. The theory of right examines how the freedom of each individual must be externally limited <u>if</u> a free society of free and equal individuals is to be possible.

Unlike Kant, Fichte does not treat political philosophy merely as a subdivision of moral theory. On the contrary, it is an independent philosophical discipline with a topic and a priori principles of its own. Whereas ethics analyzes the concept of what is *demanded* of a freely willing subject, the theory of right describes what such a subject is *permitted* to do (as well as what he can rightfully be *coerced* to do). Whereas ethics is concerned with the inner world of conscience, the theory of right is concerned only with the external, public realm, though only insofar as the latter can be viewed as an embodiment of freedom.

Having established the general, albeit hypothetical concept of right, Fichte then turns to an investigation of the conditions necessary for the realization or "application" of the same: that is, for the actual coexistence of free individuals, or the existence of a free society. The sum of these "conditions" constitute the sum of our "natural rights" as human beings, rights that can be instantiated and guaranteed only within a deliberately constructed free society. On purely a priori grounds, therefore, Fichte purports to be able to determine the general requirements of such a community and the sole justification for legitimate political coercion and obligation.

The precise relationship of Fichte's theory of right to the social contract tradition is complex, but the general outline is as follows: Fichte presents an a priori argument for the fundamentally *social* character of human beings, an argument grounded upon an analysis of the very structure of self-consciousness and the requirements for self-posting. Only after this "deduction" of the concept of right and of the applicability of the same does he explicitly introduce the notion of what he calls the *Staatsbürgervertrag*

or "citizens' contract," a notion that he goes on to analyze into a series of distinguishable moments, including the "civil contract" proper (or "property contract"), the "protection contract," and the "contract of unification," all of which must be supplemented by the contracts of "subjection" and "expiation." Fichte thus propounds what one might call a "contract theory of the state," but not of human community.

As numerous commentators, beginning with Hegel, whose own *Philosophy of Right* was strongly influenced, both positively and negatively, by Fichte's *Foundations of Natural Right*, have pointed out, the actual theory of the state that Fichte himself, in Part Two of that work, erected upon what would appear to be a rather "liberal" theoretical foundation contains many elements that are not usually associated with the individualistic, liberal tradition -- including a general indifference to questions of constitutional structure, public participation in government, etc., and a strong emphasis upon the "police" functions of the state (functions which, for Fichte, were not limited to concerns of security, but also included those of social welfare). This, however, is not particularly surprising, since the function of the state in Fichte's system is primarily to employ coercion to guarantee that the parties to the contract will, in fact, do what they have promised to do and to insure that every citizen will have an opportunity to realize his own (limited) freedom. One of the more remarkable features of Fichte's conception of right is that every citizen is entitled to the full and productive employment of his labor, and hence that the state has a duty to manage the economy accordingly. The truth is that Fichte's social and political theory is very difficulty to fit into the usual categories, but combines certain elements usually associated with liberal individualism with others more commonly associated with communitarian statism.

4.5 Philosophy of Religion

In addition to the postulates addressed by theoretical to practical reason, there are also those addressed by practical reason to nature itself. The latter is the domain of the transcendental philosophy of religion, which is concerned solely with the question of the extent to which the realm of nature can be said to accommodate itself to the aims of morality. The questions dealt with within such a philosophy of religion are those concerning the nature, limits, and legitimacy of our belief in divine providence. The philosophy of religion, as conceived by Fichte, his nothing to do with the historical claims of revealed religion or with particular religious traditions and practices. Indeed, this is precisely the distinction between philosophy of religion and "theology."

As noted above, Fichte never had a chance to develop this final subdivision of his Jena system, beyond the tentative foray into this domain represented by his controversial essay "Concerning the Basis of Our Belief in a Divine Governance of the World" and the works he contributed to the ensuing "atheism controversy." In "Concerning the Basis of our Belief" he certainly seems to contend that, so far as philosophy is concerned, the realm of the divine is that of this world, albeit viewed in terms of the requirements of the moral law, in which case it is transformed from the natural to the "the moral world order," and that no further inference to a transcendent "moral lawgiver" is theoretically or practically required or warranted. In this same essay Fichte also sought to draw a sharp distinction between religion and philosophy (a distinction parallel to the crucial distinction between the "ordinary" and "transcendental" standpoints) and to defend philosophy's right to postulate, on purely a priori grounds,

something like a "moral world order." Philosophy of religion thus includes a deduction of the postulate that our moral actions really do make some difference in the world. But this is about as far as it can go.

With respect to the existence of God, the argument of Fichte's essay is primarily negative, inasmuch as it explicitly denies that any postulate of the existence of a God independent of the moral law is justifiable on philosophical grounds. In the wake of the atheism controversy, Fichte returned to this subject and, in his "From a Private Letter" and in Part Three of *The Vocation of Man*, attempted to restate his position in a manner that at least appeared to be more compatible with the claims of theism.

5. The later *Wissenschaftslehre* and the Reception of Fichte's Philosophy

For much of the nineteenth century, beginning with Hegel's self-serving interpretation of the history of modern philosophy, Fichte's *Wissenschaftslehre* was generally assimilated into the larger history of Germany idealism. Criticized by both Schelling and Hegel as a one-sided, "subjective" idealism and a prime instance of the "philosophy of reflection," Fichte's *Wissenschaftslehre* was almost universally treated as a superceded rung on the ladder "from Kant to Hegel" and thus assigned a purely historical significance. Neglected as the *Wissenschaftslehre* may have been during this period, Fichte was not entirely forgotten, but remained influential as the author of the *Addresses to the German Nation* and was alternately hailed and vilified as one of the founders of modern pan-German nationalism.

This same situation prevailed throughout much of the twentieth century as well, during which Fichte's fortune seemed closely tied to that of Germany. Particularly during the long periods preceding, during, and following the two World Wars, Fichte was discussed almost exclusively in the context of German politics and national identity, and his technical philosophy tended to be dismissed as a monstrous or comical speculative aberration of no relevance whatsoever to contemporary philosophy. There were, to be sure, isolated exceptions and authors such as Fritz Medicus, Martial Gueroult, Xavier Léon, and Max Wundt who, during the first half of the twentieth century, took Fichte seriously as a philosopher and made lasting contributions to the study of his thought. But the real boom in Fichte studies has come only in the past four decades, during which the *Wissenschaftslehre* has once again become the object of intense philosophical scrutiny and lively, world-wide discussion -- as is evidenced by the establishment of large and active professional societies devoted to Fichte in Europe, Japan, and North America.

J. H. Stirling once quipped that "Fichte had *two* philosophical epochs; and if both belong to biography, only one belongs to history," and until quite recently there was a great deal of truth to this observation. Indeed, even today, Fichte's technical writings of the post-Jena period remain little-known to the vast majority of philosophers. Admittedly, it is hard to recognize these late texts -- which drop the strategy of beginning with an analysis of the self, along with the strong emphasis upon the "primacy of the practical," and which include unembarrassed references to "the absolute" and even to "absolute being" -- as the work of the same author who wrote the *Foundations of the Entire Wissenschaftslehre*. Though Fichte himself always insisted that his basic philosophy remained the same, no matter how much his

presentation thereof may have altered over the years, many sympathetic readers and not a few well-informed scholars have found it impossible to reconcile this claim with what at least appear to be the profound systematic and doctrinal differences between the "early" and "late" *Wissenschaftslehre*. It is therefore not surprising that the problematic "unity" of Fichte's thought continues to be vigorously debated by experts in the field.

Whatever one may conclude concerning the relationship between Fichte's earlier and later writings, it is certainly the case that, with the publication of numerous, faithfully edited "new" manuscripts of later versions of the *Wissenschaftslehre*, the focus of much of the best contemporary Fichte scholarship has shifted to his later texts, most of which were entirely unknown to earlier generations of readers. Seldom has a new edition of a philosopher's literary corpus had a greater effect upon the contemporary reputation of the thinker in question or a more stimulating effect upon contemporary scholarship than in the case of the monumental new critical edition of Fichte's works begun in the early nineteen-sixties under the auspices of the Bavarian Academy of the Sciences and the general editorship of Reinhard Lauth and others. Now nearing completion, this edition has contributed directly and enormously to the contemporary revival of interest in Fichte's philosophy in general, and in the later versions of the *Wissenschaftslehre* in particular. Much of the best recent work on Fichte, particularly in Germany, Italy, and Japan, has been devoted exclusively to his later thought. Stirling's observation is thus no longer true, inasmuch as the work of Fichte's "second epoch" has, however belatedly, now become the object of genuine and lively philosophical discussion and dispute.

In contrast, anglophone Fichte scholarship, which has also experienced quite a renaissance of its own over the past few decades, has remained largely focused upon the "classical" texts of the Jena period. This is no doubt due, in large part at least, to the appearance, during these same decades, of new, reliable translations of almost all of Fichte's early writings and the lack of translations of his later, unpublished texts. But it is also a reflection of the relatively anemic tradition of Fichte scholarship in England and North America, where even the early *Wissenschaftslehre* has long been neglected and underappreciated. Until quite recently, virtually no scholars writing in English were interested in examining the *Wissenschaftslehre* in its own right, but were mainly concerned to determine Fichte's position in relationship to Kant's or Hegel's. This situation, however, has fundamentally altered, and some of the most insightful and original current work on Fichte is being done in English.

Bibliography

Editions of Fichte's Complete Works in German

- *Johann Gottlieb Fichtes nachgelassene Werke*, 3 vols., ed. I. H. Fichte (Bonn: Adolph-Marcus, 1834-35).
- *Johann Gottlieb Fichtes sämmtliche Werke*, 8 vols., ed. I. H. Fichte (Berlin: Veit, 1845-46). [Taken together, these 11 volumes, edited by Fichte's son, constituted the first attempt at a complete edition of his works and are still widely cited and reprinted, most recently by de Gruyter, under the title *Fichtes Werke*.]

- *J. G. Fichte: Gesamtausgabe der Bayerischen Akademie der Wissenschaften*, ed. Reinhard Lauth, Hans Jacobs, Hans Gliwitzky, and Erich Fuchs (Stuttgart-Bad Cannstatt: Frommann, 1964 --) 32 vols to date. [Organized into four separate series -- writings published by Fichte, unpublished writings, correspondence, and student lecture transcripts -- this monumental critical edition, which should be complete by 2010, supersedes all earlier editions.]

Individual Works and English translations

- *Versuch einer Kritik aller Offenbarung* (1792; 2nd ed., 1793). *Attempt at a Critique of All Revelation*, trans. Garrett Green (New York: Cambridge University Press, 1978).
- *[Rezension:] Aenesidemus* (1794). "*Aenesidemus* Review." In *Fichte: Early Philosophical Writings* (henceforth = *EPW*), trans. and ed. Daniel Breazeale (Ithaca: Cornell University Press, 1988, 2nd ed., 1993).
- *Ueber den Begriff der Wissenschaftslehre* (1794, 2nd ed. 1798). *Concerning the Concept of the Wissenschaftslehre*, trans. Breazeale, in *EPW*.
- *Einige Vorlesungen über die Bestimmung des Gelehrten* (1794). *Some Lectures Concerning the Scholar's Vocation*, trans. Breazeale, in *EPW*.
- *Grundlage der gesamten Wissenschaftslehre* (1794/95; 2nd edn. 1802). *Foundations of the Entire Science of Knowledge*, trans. Peter Heath. In *Fichte: Science of Knowledge (Wissenschaftslehre)*, ed. Peter Heath and John Lachs (New York: Appleton-Century-Crofts, 1970; 2nd ed., Cambridge: Cambridge University Press, 1982).
- *Grundriß des Eigenthümlichen der Wissenschaftslehre in Rücksicht auf das theoretische Vermögen* (1795). *Outline of the Distinctive Character of the Wissenschaftslehre with respect to the Theoretical Faculty*, trans. Breazeale, in *EPW*.
- *Wissenschaftslehre nova methodo* (student lecture transcripts, 1796 - 1799). *Foundations of Transcendental Philosophy (Wissenschaftslehre) nova methodo*, trans. and ed. Daniel Breazeale (Ithaca: Cornell University Press, 1992).
- *Grundlage des Naturrechts nach Principien der Wissenschaftslehre* (1796/97). *Foundations of Natural Right*, ed. Frederick Neuhouser, trans. Michael Baur (Cambridge: Cambridge University Press, 2000).
- *Versuch einer neuen Darstellung der Wissenschaftslehre* ("Erste" und "Zweite Einleitung," 1797; "Erste Capitel," 1798). *Attempt at a New Presentation of the Wissenschaftslehre*. In *Introductions to the Wissenschaftslehre and Other Writings (1797 - 1800)* [henceforth = *IWL*], ed. and trans. Daniel Breazeale (Indianapolis: Hackett, 1994).
- *Das System der Sittenlehre nach den Principien der Wissenschaftslehre* (1798). *The System of Ethics in accordance with the Principles of the Wissenschaftslehre*, ed. and trans. Günter Zöller and Daniel Breazeale (Cambridge: Cambridge University Press, forthcoming).
- *Ueber den Grund unsers Glaubens an eine göttliche Weltregierung* (1798). "On the Basis of our Belief in a Divine Governance of the World," trans. Breazeale, in *IWL*.
- *Aus einem Privatschreiben* (1800). "From a Private Letter," trans. Breazeale, in IWL.
- *Die Bestimmung des Menschen* (1800). *The Vocation of Man*, trans. Peter Preuss (Indianapolis: Hackett, 1987).
- *Sonnenklarer Bericht an das größere Publikum über das eigentliche Wesen der neuesten*

Philosophie. Ein Versuch, die Leser zum Verstehen zu zwingen (1801). *A Crystal Clear Report to the General Public Concerning the Actual Essence of the Newest Philosophy: An Attempt to Force the Reader to Understand*, trans. John Botterman and William Rash, in *Philosophy of German Idealism*, ed. Ernst Behler (New York: Continuum, 1987). [Unreliable translation.]
- *Der Grundzüge des gegewärtigen Zeitalters* (1806). *The Characteristics of the Present Age*, in *The Popular Works of Johann Gottlieb Fichte* [henceforth = *PWF*], trans. William Smith, ed. and with an introduction by Daniel Breazeale (Bristol, England: Thoemes Press, 1999). [These translations were original published between 1848 and 1889.]
- *Über des Wesen des Gelehrten, und seine Erscheinungen im Gebiete der Freiheit* (1806). *On The Nature of the Scholar and Its Manifestations*, trans. Smith, in *PWF*.
- *Die Anweisung zum sieligen Leben, oder auch die Religionslehre* (1806). *The Way Towards the Blessed Life; or, the Doctrine of Religion*, trans. Smith, in *PWF*.
- *Reden an die deutsche Nation* (1808). *Addresses to the German Nation*, trans. R. F. Jones and G. H. Turnbull, ed. George Armstrong Kelly (New York: Harper & Row, 1968).
- *Die Wissenschaftslehre, in ihrem allgemeinen Umrisse dargestellt* (1810). "The Science of Knowledge in its General Outline," trans. Walter E. Wright, *Idealistic Studies* 6 (1976) 106-117.

Secondary Literature about Fichte and the *Wissenschaftslehre*

- Adamson, Robert. *Fichte* (Edinburgh: Blackwood, 1881). [Though woefully out of date, this remains the only full-scale treatment of Fichte in English.]
- Baumanns, Peter. *Fichtes Wissenschaftslehre. Probleme ihres Anfangs* (Bonn: Bouvier, 1974). [Useful exposition of various ways of interpreting the starting point of the first *Wissenschaftslehre*.]
- Baumanns, Peter. *J. G. Fichte: Kritische Gesammtdarstellung seiner Philosophie* (Freiburg: Alber, 1990). [A critical overview of Fichte's philosophical development and analysis of his system, concentrating upon the Jena period.]
- Baumgartner, Michael and Jacobs, Wilhelm G. *J. G. Fichte: Bibliographie* (Stuttgart-Bad Cannstatt: Frommann, 1968). [A complete bibliography. Supplemented by Doyé's bibliography.]
- Breazeale, Daniel and Rockmore, Tom, eds. *Fichte: Historical Context/Contemporary Controversies* (Atlantic Highlands: Humanities, 1993.) [A collection of essays on various aspects of Fichte's philosophy. Includes a complete bibliography of works in English by and about Fichte.]
- Breazeale, Daniel and Tom Rockmore, eds. *New Essays on Fichte's Foundation of the Entire Doctrine of Scientific Knowledge (*Amherst, NY: Humanity Books, 2001).
- Breazeale, Daniel and Tom Rockmore, eds. *New Essays on Fichte's Later Jena Wissenschaftslehre* (Evanston, IL: Northwestern University Press, 2002). [A collection of essays on the *Wissenschaftslehre nova methodo, Foundation of Natural Right*, *System of Ethics*, and Atheism Controversy.]
- Breazeale, Daniel and Rockmore, Tom, eds. *New Perspectives on Fichte* (Atlantic Highlands, N.J.: Humanities Press, 1996). [More essays on various themes from Fichte's early philosophy.]
- Doyé, Sabine, ed. *J. G. Fichte-Bibliographie (1969-1991)* (Amsterdam and Atlanta: Editions Rodopi, 1993. [An essential supplement to the bibliography by Baumgartner and Jacobs.]

- Gueroult, Martial. *L'evolution et la structure de la doctrine de la science chez Fichte,* 2 vols. (Paris: Société de l'édition, 1930). [A pioneering developmental study of the *Wissenschaftslehre*.]
- Everett, Charles Caroll. *Fichte's Science of Knowledge: A Critical Exposition* (Chicago: Griggs, 1884).
- "Fichte and Contemporary Philosophy," special issue of *Philosophical Forum* 19, nos. 2 and 3 (1988).
- Fuchs, Erich, ed. (1978 - 1992) *J. G. Fichte im Gespräch: Berichte der Zeitgenossen*, 6 vols. (Stuttgart-Bad Cannstatt: Frommann, 1978-92). [An encyclopedic collection of contemporary reports on Fichte and his writings. An invaluable research tool.]
- Henrich, Dieter. *Fichtes ursprüngliche Einsicht* (Frankfurt am Main: Klostermann, 1967); trans. David Lachterman, "Fichte's Original Insight," Contemporary German Philosophy 1 (1982) 15-52. [An influential reading of Fichte's alleged movement beyond a "reflective theory of consciousness."]
- Janke, Wolfgang. *Fichte: Sein und Reflexion -- Grundlagen der kritischen Vernunft* (Berlin: de Gruyter, 1970). [A hermeneutic-Heideggerian reading of Fichte, with an emphasis upon the later writings.]
- Lauth, Reinhard. *Hegel vor der Wissenschaftslehre* (Stuttgart: Steiner-Verlag, 1987). [A vigorous and convincing defense of Fichte against Hegel's criticisms.]
- ------ *Die transzendentale Naturlehre Fichtes nach den Prinzipien der Wissenschaftslehre* (Hamburg: Meiner, 1984). [A masterful exposition of Fichte's transcendental approach to the philosophy of nature.]
- ------ *Transzendentale Entwicklungslinien von Descartes bis zu Marx und Dostojewski*, Hamburg: Meiner, 1989. [A collection of essays, most of them on Fichte, by the leading Fichte scholar of the age.]
- La Vopa, Anthony J. *Fichte. The Self and the Calling of Philosophy, 1762-1799* (Cambridge: Cambridge University Press, 2001). [Sophisticated partial biography which tries to connect Fichte's philosophy to his life.]
- Léon, Xavier. *Fichte et son temps*, 3 vols. Paris: Armand Colin, 1922-27.)
- Martin, Wayne M. *Idealism and Objectivity: Understanding Fichte's Jena Project* (Stanford, CA: Stanford University Press, 1997). [An bracing, critical interpretation of the project of the early *Wissenschaftslehre*.]
- Neuhouser, Frederick. *Fichte's Theory of Subjectivity* (Cambridge: Cambridge University Press, 1990). [A fine example of a contemporary appropriation of Fichte's thought and of an analytically sensitive exposition of the same.]
- "New Studies in the Philosophy of Fichte," special issue of *Idealistic Studies* 6, no. 2 (1979). [A collection of essays on Fichte in English.]
- Pareyson, Luigi. *Fichte. Il sistema della libertà* , 2nd ed. (Milan: Mursia, 1976). [Along with Philonenko's similarly titled book, Pareyson's exposition of the early system as a "system of freedom" is one of the most influential works on Fichte of the post-war period.]
- Philonenko, Alexis. *La liberté humaine dans la philosophie de Fichte,* Paris: Vrin, 1966; 2nd ed., 1980). [A very original and influential study of Fichte's early philosophy, interpreted as a "philosophy of human finitude." Essential.]
- Radrizzani, Ives (1993) *Vers la fondation de l'intersubjectivité chez Fichte: des Prinzipes à la*

Nova Methodo (Paris: Vrin, 1993). [Argues for the continuity of Fichte's development within the Jena period and for the centrality therein of the *Wissenschaftslehre nova methodo*.]

- Renaut, Alain. *Le système du droit: Philosophie et droit dans la pensée de Fichte* (Paris: Presses Universitaires de France, 1986). [A powerful reading of the *Foundation of Natural Right*, which argues that political philosophy is the keystone of the Jena *Wissenschaftslehre*.]
- Rockmore, Tom. *Fichte, Marx, and the German Philosophical Tradition* (Carbondale: Southern Illinois University Press, 1980. [One of the first successful effort in English to liberate Fichte's philosophy from the shadow of Hegel.]
- Seidel, George J. *Fichte's Wissenschaftslehre on 1794: A Commentary on Part I* (Lafayette: Purdue University Press, 1993). [An elementary introduction to the early system. Written with the beginning student in mind.]
- Wundt, Max. *Fichte-Forschungen* (Stuttgart: Frommann, 1929). [Another pioneering study of the development of Fichte's thought, with an emphasis upon the different "spirits" of the various versions of the *Wissenschaftslehre*.]
- Zöller, Günter. *Fichte's Transcendental Philosophy: The Original Duplicity of Intelligence and Will* (Cambridge: Cambridge University Press, 1998). [A well-informed and stimulating analysis of several central themes from the early *Wissenschaftslehre*.]
- See too the journal *Fichte-Studien* (Amsterdam and Atlanta: Editions Rodopi, 1990 ff.), which appears roughly once a year and publishes papers, most of them in German, on every aspect of Fichte's life and thought, as well as the occasional newsletter, "Fichteana," published by the North American Fichte Society (and available at their website).

Other Internet Resources

- North American Fichte Society website
- Internationale-Fichte-Gesellschaft website
- Bavarian Academy of the Sciences Fichte website

Related Entries

Kant, Immanuel | Maimon, Salomon | Reinhold, Karl Leonhard

Daniel Breazeale
breazeal@pop.uky.edu

Table of Contents

First published: August 30, 2001
Content last modified: August 30, 2001

Salomon Maimon

Salomon Maimon (1753-1800), the man whom Kant described as the best of his critics, stands as one of the most acute and complicated thinkers -- and certainly one of the most fascinating personalities -- of the 1780s and 1790s. By granting the principle of sufficient reason unlimited validity Maimon embraces a radical form of rationalism. By presenting strong criteria for the validity of knowledge Maimon suggests that even Kant's attempt to limit epistemological claims to the realm of possible experience cannot be secured without a substantial ontological commitment. By doing this Maimon tries to present Kant with the dilemma of either adopting elements from the dogmatic metaphysics Kant set to challenge, or having his system undermined by skepticism. In revealing what he sees as the skeptical implications of rationalism, Maimon raises important objections to Kant's critical idealism, as well as develops deep insights into the problems of experience and givenness. His ‘skeptically rationalist’ claims about the nature and limits of human cognition present a distinctive perspective on the Kantian project of transcendental idealism, as well as on central epistemological issues concerning the relation between thought and the world.

1. Intellectual Biography

“*Scholars of Wisdom have no rest in this world or in the world to come*.”

This Talmudic saying, with which Maimon concludes his first philosophical work, applies strikingly well to the life story of Salomon Maimon. Maimon was born in 1753 in Suchowyborg, a village next to the town of Mir in Lithuania. His family, which originally was rather wealthy, fell into poverty due to poor management of its properties. Thus, Maimon's father became a children's teacher, an example that was followed later by his son Salomon. Maimon received a traditional religious education, which concentrated mainly on the study of the *Talmud*. At the age of 11, shortly after the death of his mother and following a comedy of errors involving the mothers of two young girls, Maimon was married in an arranged ceremony, and three years later his first son, David, was born. During his early adulthood, Maimon developed a keen curiosity about the sciences and philosophy. The most crucial influence was that of Maimonides' *Guide of the Perplexed*, through which Maimon became acquainted with Aristotelian philosophy in its medieval cloth. Maimon's attachment to Maimonides -- both personally and philosophically -- ran throughout his life. Even Maimon's own name was adopted as an expression of respect towards this teacher (At the time, few Jews adopted family names. Before taking the surname ‘Maimon’, Maimon used to be called after his father: Salomon *ben* ("*son of*") *Joshua*). Maimon also developed an interest in Kabbalistic texts, which in spite of his relatively young age, he attempted to study and interpret according to the bodies of knowledge he had already acquired (i.e., Maimonidean philosophy). In his early twenties Maimon went to visit the *Maggid* of Mezrich, the contemporary leader and one of the founders of *Hassidism*. Though in his *Lebensgeschichte* Maimon passes a strong negative judgment against the *Hassidic* movement, it seems that several main themes of his later philosophy developed through a careful dialogue with its teachings. In his mid-twenties, hoping to widen his knowledge of philosophy and the sciences, Maimon left his family and went to Berlin (under the pretense of attempting to study medicine there). This first visit to Berlin ended shortly and grimly. Having confided to one of the officers of the Jewish community that the purpose of his visit was to study philosophy and that he intended to publish a new commentary on Maimonides' *Guide of the Perplexed*, Maimon was asked to pack his belongings and leave the shelter of the Jewish community -- and thus the city -- at once. The following half year Maimon spent as a wandering beggar, till he reached Posen, where he was offered a shelter and a position as a tutor at the house of one of the city's Jews. During his stay in Posen Maimon wrote one of his most fascinating works, *Hesheq Shelomo* (“Solomon's Desire.” See the Bibliography). In 1780 Maimon went again to Berlin. This journey was much more successful and Maimon established a close connection with Moses Mendelssohn and entered the circles of the *Haskala* (the Jewish Enlightenment movement) in Berlin. Yet neither Berlin nor its enlightened Jews provided a real home for Maimon. For the cultured Jews of Berlin, Maimon was a rude *Ostjude* (East-European Jew), who spoke awful German accompanied by a variety of wild gesticulations. They did, however, acknowledge the genius of this person, who could, for example, read a difficult book of mathematics for the first time, and then explain it -- in his savage manner -- a short while later. Similarly, Maimon had little appreciation for these fine bourgeoises who, free from any barriers which would impede their ability to study the sciences, only contented themselves with a shallow acquaintance with what a civilized person *should* know. Furthermore, they lacked the sharpness of mind of his fellow Talmudists in Poland. Maimon does seem to have had a genuine appreciation for Mendelssohn, both because of his kindheartedness and because, unlike most others of the circles of the Jewish Enlightenment, Mendelssohn had a reasonable grasp of the *Talmud* and rabbinic literature.

In 1783, after a journey to Hamburg, Amsterdam and then back to Hamburg, Maimon entered a

gymnasium in Altona, where he stayed for two years. During this period Maimon studied several European languages and improved his knowledge of the natural sciences and his command of German. In 1785, Maimon left Altona for Berlin, where he met Mendelssohn for the last time. Later that year Maimon moved to Dessau, where he wrote a Hebrew textbook on mathematics, and subsequently settled in Breslau. Here, after a failed attempt to study medicine, Maimon again took up the position of a tutor. While staying in Breslau, Maimon translated Mendelssohn's *Morgenstunden* into Hebrew and wrote a Hebrew textbook on Newtonian physics -- *Ta'alumoth Hochma* (“*Mysteries of Wisdom*”). After more than a decade of separation, Maimon's wife, accompanied by their eldest son, succeeded in locating him in Breslau, and demanded that he either return to Lithuania or get a divorce. Reluctant to sever his ties with his past, Maimon tried to postpone the decision, but after his wife insisted that he must make the choice, he finally agreed to the divorce.

Towards the end of the 1780s Maimon traveled again to Berlin. There he heard about Kant's new philosophy, and for a couple of months dedicated himself to a careful study of the *Critique of Pure Reason*. In his *Autobiography*, Maimon alludes to his “rather curious” methods for understanding this text:

> In the first reading I reached a vague sense of each section, which through subsequent readings I then sought to make determinate, and by this to penetrate the meaning of the author. This is what is properly meant when one thinks oneself into a system. Since I had already used this method in mastering the systems of Spinoza, D. Hume, and Leibniz, it was natural that I would be led to think of them as a ‘Coalition-System.’ This I actually discovered, and by and by set it out in the form of notes and observations on the *Critique of Pure Reason*(GW I, 557 | LB 253).

Maimon put his thoughts about the *Critique* in a letter he sent to Kant through a common friend, Markus Herz. Kant responded in a letter full of praise for Maimon, describing him as “having an acumen for such deep investigation that very few men have” and claiming that “none of my critics understood me and the main questions as well as Herr Maimon does.” (Ak. 11:48)

Kant's recognition opened for Maimon the salons of Berlin as well as the contemporary leading journals, in which Maimon started publishing. The story of the crude genius who came from the East and penetrated the heart of German philosophy became a common topic of small talk in these circles. In 1790, Maimon published an expanded version of his comments of Kant's first *Critique* as the *Versuch über die Transcendentalphilosophie* (“An Essay on Transcendental Philosophy”). A year later, Maimon collaborated with the members of the Jewish Enlightenment and prepared a Hebrew commentary on Maimonides' *Guide of the Perplexed.* (The publisher decided to publish only the first part of this work, since he found Maimon's commentary much too deep philosophically, and thus unfitting for the political purposes of propagating the ideas of the enlightenment among the Jews.)

During the early 1790s, Maimon struck a close friendship with Karl Philipp Moritz (the author of *Anton Reiser*). Maimon became a frequent contributor and later the co-editor of Moritz's *Magazin zur Erfahrungsseelenkunde*, which was in fact the first journal dedicated to the study of psychology.

Following Moritz's death in 1793 Maimon attempted to find a new patron. He established a connection with Goethe, who invited him to Weimar, yet because of reasons that remain unclear, this relationship did not succeed. Maimon's material life in that period was quite miserable. He lived in extreme poverty, and spent the little money he earned on alcohol -- for the price of a drink one could buy his conversation in a tavern. In 1795, Maimon accepted a generous offer by a young Silesian nobleman, Graf Kalkreuth, and moved to the latter's estate in Siegersdorf (currently: Kozuchow) in Lower Silesia. From this time until his death in November 22, 1800, Maimon led a quiet, though lonely and melancholic, life on Kalkreuth's estate. In May 1800, Maimon wrote to one of his Jewish friends in Berlin, attempting to arrange his return to Berlin -- and to find financial support for that purpose -- but this plan never materialized.

In the last decade of his life, Maimon wrote ten books, and numerous articles. The most important among these books (apart from the *Transcendentalphilosophie*) are his philosophical works, *Versuch einer neuen Logik, oder Theorie des Denkens (An Attempt at a New Logic, or a Theory of Thinking*) (1793), *Kritische Untersuchungen über den menschlichen Geist, oder das höhere Erkentniß und Willensvermögen* (*Critical Investigations on the Human Mind, or the Highest Faculty of Knowledge and Will*) (1797), and his *Lebensgeschichte* (*Autobiography*) (1792/3), which is his only work which achieved public recognition and popularity.

2. Maimon's Critique of Kant

In the highly polarized German philosophical community in the 1790s, Maimon's intellectual allegiances remained rather ambiguous. In a letter of 1791, Maimon had written to Kant that while he found the skeptical part of the *Critique of Pure Reason* wholly convincing, he harbored doubts about the more dogmatic aspects of Kant's system. Kant, of course, viewed Maimon not as an ally, but as the best of his critics -- it remains an open question whether Maimon saw himself as fundamentally a friend or foe of the critical philosophy. At the very least, however, Maimon's criticisms of the *Critique* cut to the core of Kant's transcendental idealism, in particular because they engage with what Maimon sees as an *internal* problem for Kant's system.

Perhaps the most obvious problem -- and certainly one of the earliest -- that Kant faces concerns the issue of the thing in itself. In order to account for the content of cognition, Kant notoriously posits a thing in itself that stands outside of the realm of possible experience yet nevertheless serves as the causal source of cognitive content. This mysterious entity was attacked most famously by Jacobi, with his accusation that one could not enter into Kant's system without the assumption of the thing in itself, but that with such an assumption one could not remain within it. Maimon too was quick to point out that the thing in itself could not be understood realistically. Rather, Maimon argues, the thing in itself must be understood as completely conceptual determination of an object, which can be approached only asymptotically. Against Kant's claims about a noumenal realm of things as they are in themselves, Maimon holds that the thing in itself stands only as an object of inquiry, rather than an independent, noumenal entity. As such, while Maimon agrees with Kant that since we are finite beings the thing in itself is in fact beyond the realm of possible experience, this does not imply that the thing in itself cannot *in principle* be an object of cognition.

Maimon's criticisms of Kant's account of the thing in itself, however, are connected to a deeper concern about the notions of experience and cognition that stand at the heart of Kant's critical idealism. Kant posits the thing in itself in large part because of his commitment to a type of cognitive dualism, in which human experience is taken to involve both a faculty of thought (the understanding), as well as a faculty of receptivity (sensibility). In order to have any content, experience requires data, which must be given to the subject through the senses. Moreover, the data of experience cannot be produced by the faculty of thought itself. The given content of sensibility plays an ineliminable role in cognition, and its source must ultimately be traced to the subject's being affected by something distinct from itself, a role played by the thing in itself. But while the content of thought is provided by the affection of an object on the faculty of sensibility, the way in which this content is cognized remains the purview of the faculty of the understanding. Mere affection, in other words, is not equivalent to cognition. In Kant's terms, the “understanding is not capable of intuiting anything, and the senses are not capable of thinking anything.” (A52/B56)

For Maimon, Kant's cognitive dualism -- which begins with wholly distinct faculties of cognition -- fails to explain how the various elements can come together in a way that makes experience possible. On this objection, Kant cannot justify his assumption that concepts and intuitions necessarily unite in cognition. In more Kantian terms, Maimon calls into question Kant's answers to both the *quid facti* and the *quid juris* that begin the Transcendental Deduction in the first *Critique*. According to Maimon, while the *quid facti* -- the question of the fact of our use of a priori concepts in experience -- is taken by Kant to be an unproblematic statement about the nature of human experience, the very assumption that we in fact do possess the type of experience Kant ascribes to us can be called into doubt. Kant's central argument in the Transcendental Deduction begins by assuming that experience exhibits a ‘dualistic’ structure, but according to Maimon, this position is not warranted, since the presumed constituent elements of experience (bare intuitions or concepts) are never themselves objects of experience. For Maimon, Kant's transcendental arguments remain mere ‘castles in the air’: while they might be valid, they fail to provide the ‘fact of experience’ that would make them sound. From the beginning, then, Maimon views Kant's transcendental project with suspicion.

This suspicion becomes an explicit problem, Maimon claims, when the specific structure of Kant's system is examined. Here, Maimon argues, the *quid juris* of the deduction -- the question about the legitimate use of the categories -- can likewise be cast into doubt. While Kant claims in the Deduction that the *a priori* concepts of the understanding -- the categories -- are necessary conditions on the unity of the manifold of sensible data given in intuition, Maimon argues that such a position leads to serious problems for the later claims Kant makes in the ‘Schematism’ and ‘Principles’ sections of the Critique. According to Maimon, Kant cannot explain how different categories are able to discriminate between different intutive content. Using the example of causality, Maimon argues that Kant has no way of explaining why some orders of perceptions represent causal connections, and why other orders of perceptions are merely associative. The reason for this claim is found in the fact that while Kant grounds the category of causality in a necessary order of perceptions, his justification for this claim appeals not to the content of the perceptions, but only to their formal ‘rule-governed’ connection -- but, as Maimon argues, *any* order of perceptions can meet this formal requirement. Yet since the content of intuitions

does not contain any temporal ordering, Kant explicitly bars an appeal to such content in applying the category of causality. As such, Kant's account of cognition faces a dilemma: either it must appeal to the content of experience and so violate Kant's own strictures, or the application of categories is merely arbitrary. Kant's central epistemological commitment -- his ‘cognitive dualism’ -- then leads, Maimon claims, to insuperable problems.

While Maimon's objections to Kant focus on specific issues that arise in the *Critique*, they all rest on Maimon's commitment to a type of ‘skeptical rationalism.’ Maimon notes that in the *Versuch über die Transcendentalphilosophie* the important problem of the *quid juris* was presented in a much wider sense than that in which Kant takes it, and thereby [such a position] allowed a place for Hume's skepticism in its complete power. On the other hand, the complete solution to this problem necessarily leads to Spinozistic or Leibnizian dogmatism (GW I, 558 | LB 254).

Moreover, in the face of the question of how the understanding can subsume or comprehend a given object, Maimon notes that “for the Kantian system, [which claims] that our sensibility and understanding are two entirely different sources of our cognition, this question is … unanswerable; on the other hand, for the Leibizian-Wolffian system, in which both [sensibility and understanding] flow from the same cognitive source (their difference consisting only in the grades of completeness of this cognition), the question can easily be answered.” (GW II, 63-4 | VT 63-4) But while Maimon sees in the dogmatism of Spinoza, Leibniz and Wolff a means of avoiding the problems that attend Kant's ‘cognitive dualism,’ he remains a skeptic about whether this solution can ever be demonstrated.

3. Content, Givenness, and Space and Time

The rejection of cognitive dualism raises a rather vexing problem for Maimon: if the content of cognition is not to be found in the affection of the merely passive faculty of sensibility, where does the content come from? On the face of it, of course, it seems clear that humans are not wholly responsible for the world of experience, but rather encounter it; Maimon must provide some explanation of the ‘given’ content of experience without appealing to something like the Kantian cognitive dualism he finds so problematic.

Although the details of Maimon's answer to this problem remain obscure, the kernel of his position can be found in his analysis of what it is to be a finite cognizer. While Kant moves from the fact of our human finitude to the need for a ‘given’ element of cognition, Maimon claims that such a step need not be taken. Instead of characterizing finitude in terms of the need for a passive faculty of receptivity, Maimon insists that finitude only implies incompleteness in our cognition -- but this incompleteness does not warrant any conclusions about the provenance of the matter of cognition. The content of sensibility is simply that which is passive in cognition -- namely, that upon which the understanding operates. The expression that content is given from ‘outside of us,’ Maimon writes, means only “something in a representation of which we are conscious of no spontaneity, that is, (in view of our consciousness) a mere passivity without activity.” And, he continues, the word ‘given’ signifies not “something in us that has its cause outside of us … rather, [the given] means merely a representation, whose manner of origin in us is

unknown to us.” (*GW* II, 203 | VT 203) What we take to be merely given to us in experience can in fact be explained in terms of a productive -- and hence active -- capacity of the mind, although the procedures of this activity remain unknown to us.

In this respect Maimon revives the Leibnizian notion that there is not a difference in kind, but only in degree, between a finite and an infinite intellect. Maimon argues that for an infinite intellect, all of the content of thought is consciously produced through the mind's own activity -- by virtue of its infinity, nothing needs to be given to such an intellect. By the same token, Maimon holds that we can think of finite cognizers in the same terms, but with the crucial difference that finite minds are not aware of the productive capacities that create the matter of experience. The supposedly given content provided by sensibility, in other words, can in fact be explained in terms of the ‘subconscious’ productive capacities of the active mind. In this respect, Maimon argues that our minds are limitations of the divine or infinite mind; our active powers are conscious, he claims, in mathematics, where we display a ‘god-like’ ability to create content according to rules of thought. In the case of empirical content, however, the creative process remains uncognized.

These ‘subconscious’ products become conscious to the finite mind, Maimon claims, by being represented in space and time. The contrast with Kant is again important, for while Kant claims that space and time are forms of human intuition, Maimon maintains that space and time are in fact the ways in which humans represent conceptual differences between thoughts. Space and time, that is, “are both concepts and intuitions, and the latter presupposes the former.” (GW II, 18 | VT 18) Space and time are concepts as representations of the differences of things in general, but they are intuitions when they represent a particular sensible object in relation to other sensible objects. As finite cognizers, we represent in space and time what we have not completely conceptualized. The fact that we represent content spatially and temporally indicates only that there is some incompleteness in our conception of the world, and not that this content is provided by a realm of wholly independent objects. Maimon claims that the representations of space and time as intuitions arise as the result of the faculty of imagination, which is, as he describes it, the faculty of fictions (GW III, 61 | PW 37). Space and time are then taken to be fictions in that they add properties to objects that are not present in the conceptual determination of these objects. As such, they serve as “negative criteria” for the incompleteness of our knowledge of objects (GW V, 192 | VnL 134). Although we never have complete determination, we do get *nearer* to the complete concept of an object. The fact that we represent objects in space and time points to the fact that something *remains to be* determined -- spatial or temporal diversities, that is, must have their ground in some conceptual differences. The representations of space and time provide indications that “the concepts of experience, and consequently also the determined relations of objects of experience, are incomplete.” (*GW* V, 192 | VnL 134)

This emphasis on the fictional nature of spatial and temporal properties again echoes the Leibnizian explanation of space and time as the representation of conceptual differences, but where Leibniz claims that space and time are derived from monadic relations, Maimon argues that the intuitions of space and time are in fact *a priori* human forms of representation, or, in Maimon's terms, ‘forms of difference’. In order to represent an object in space and time, Maimon maintains, the conceptual content that grounds such a representation must contain a diversity in order to be represented spatially and temporally. An

intuited visual field of homogeneous red, for example, would not be spatially represented, since there would not be any diversity present. Spatiality would arise only with the introduction of some distinct content -- a spot of green in the red field, for example.

For Maimon, the *formal* nature of space and time suffices to yield mathematical and geometric necessity. In mathematics, Maimon claims, space and time are *given a priori* to the faculty of cognition; the objects of mathematics are “nothing but space and time in all their possible modifications.” (*GW* V, 184 | VnL 126) Mathematics, that is, relates to an object given a priori, or rather “itself determines these objects *a priori*.” (*GW* V, 183 | VnL 125) The objects of mathematics and geometry are then directly created or determined according to the understanding's *a priori* rules of production. For Maimon, as for Kant, the ground of the a priority and necessity of geometry and arithmetic lies in the need for the construction and exhibition of concepts in intuition.

Space and time are then presented as *a priori* ‘forms of difference,’ but a question remains about the content of experience: the supposedly independent world of objects, in all of its variety and multiplicity, still needs to be explained. On this question, unfortunately, Maimon's position is particularly obscure. In the *Versuch über die Transcendentalphilosophie*, Maimon develops a theory about the content of experience in terms of what he calls ‘infinitesimals of perception.’ Here the invocation of the calculus is deliberate, for he argues that the content of experience can be explained in terms analogous to the way in which infinitesimals are treated in mathematics. A line, for example, can be understood as composed of an infinite number of points, each of which stands in relation to the others; moreover, these points are densely ordered, for there are an infinite number of points between any two points. But while the differences between the points of the line are themselves infinitely small, the relation between them is a determinate value -- the slope, which can be calculated for any point on the line. Likewise, Maimon claims, the content of experience can be understood as analogous to a ‘perceptible line’ that is composed of an infinite number of smaller components, none of which is ontologically distinct from the experience itself. On this view, “sensibility provides the differentials to a determined consciousness; the imagination produces from these [differentials] finite (determinate) objects of intuition; the understanding produces from the relations of these different differentials, which are its objects, the relation from which arise sensible objects.” (*GW* II, 31-2 | VT 31-2) In this sense, there is -- at least in principle -- no need to appeal to a content given from outside of experience; instead, experience is itself composed or ‘integrated’ from the posited infinitesimal elements of thought. As finite cognizers, we *represent* in space and time the purely conceptual differences that are simply presented in thought. (A similar strategy for dealing with the problem of the thing-in-itself was developed a century later by the Marburg Neo-Kantians, especially Hermann Cohen.)

While the theory of the infinitesimals of perception is complex, it points toward Maimon's deeper rationalism. Both his objections to Kant's account of cognition as well as his own positive project turn on a rejection of what might be called ‘brute givens.’ For Maimon, givenness stands in opposition to cognition, for he holds that no explanation can be provided for how merely given content can be taken up in thought -- how, that is, the active faculties of thought can legitimately apply to a passively received given. By attempting to explain givenness within a larger framework of an active consciousness, Maimon presents a position that avoids -- at least in principle -- the problems that Maimon sees with Kant's

cognitive dualism.

4. Logic and the Law of Determinability

In the *Critique of Pure Reason* Kant contemplates the possibility of a law that would govern the *content* of synthetic judgements. This law is supposed to be a complement to the principles of non-contradiction and excluded middle, which govern the logical *form* of both synthetic and analytic propositions. The law, which Kant terms “The Principle of Thoroughgoing Determinability” (*Grundsatz der durchgngigen Bestimmung*), states that “of every pair of possible [and opposite] predicates, one of them must apply” to every single thing [A573/B601]. As a result, every single thing would be thoroughly determined with regard to any pair of opposite predicates. This law seems to necessitate the idea of the sum total of all possibility, and as a further result, the concept of an *ens realismus*. Kant argues, however, that this derivation is not valid insofar as it attempts to apply a principle that is limited to possible experience to all things (including things in themselves) [A583/B661]. Like Kant, Maimon too suggests a transcendental principle that will govern synthetic propositions with regard to their content and not their form (which is governed by the law on non-contradiction). In spite of these similarities, Maimon's law the Law of Determinability (*Satz der Bestimmbarkeit*) is significantly different from the one offered by Kant. The two laws not only differ with regard to their contents, but they also serve different functions, and have different weight in the two systems. While the Kantian law has a relatively marginal place within his systems (and is considered by many scholars as a mere residue of the metaphysical inheritance of Baumgarten and Wolff), Maimon's law seems to be the crucial axis of his positive philosophy (and seems to be influenced by Spinozistic metaphysics). The laws also differ with regard to their domains of applicability. While Kant restricts his law to possible experience, Maimon argues that the demands of his law are satisfied only within the domain of *a priori* mathematical thinking, and that empirical cognition fails to pass its test.

The main aim of Maimon's Law of Determinability is to provide a *criterion* that would distinguish between syntheses that reflect a *real* connection between concepts, and *arbitrary* syntheses, which result from the activity of the imagination. Real syntheses are of crucial importance for Maimon, since through this kind of syntheses we can create new concepts, and, in the case of *a priori* syntheses, even create genuine objects. According to Maimon in any real synthesis of a subject and a predicate, the following two principles must be observed:

> (1) A principle for the *subject* in general: every subject must be a possible object of consciousness, not only as a subject but also in itself;
>
> (2) A principle for *predicates*: every predicate must be a possible object of consciousness, not in itself, but only as a predicate (in connection with the subject) What does not conform to this principle can be a merely *formal*, or *arbitrary*, but not a *real* thought (GW V, 78 | VnL 20).

A synthetic judgment then accords with the law of determinability when its predicate is a real

determination of the subject (i.e., when its predicate is asymmetrically dependent upon its subject). Thus, for example, in the synthesis straight line, the predicate straight is a real determination of the subject line, since one can think of the subject without the concept of the predicate, while one can conceive of straightness only through the concept of a line. In an empirical synthesis, such as, the red line, the predicate is merely an arbitrary determination insofar as our intellect does not decipher any internal connection and dependence between the subject and the predicate (GW II, 92-3 | VT 92-3). Another paradigmatic example, which Maimon uses to explain the notion of real synthesis, is that of a right angle. Here too, Maimon argues, the subject (angle) can be thought in itself without relation to the predicate, while the predicate (right) cannot be thought without the subject. The use of mathematical examples is not coincidental, since Maimon argues that it is only in mathematics that we can find real syntheses, namely, syntheses that pass the test of the law of determinability. This special advantage of mathematics is due to the role of construction in mathematics. Thus, in the case of the straight line the intellect commands a construction of a line in pure intuition according to the concept of straight. In such a way, the connection between the subject and the predicate, while synthetic, is nevertheless necessary. By contrast, a judgment such as The cup is green fails to accord with the law of determinability, for the connection between the subject and the predicate remains merely problematic. For our intellect, the greenness of the cup is something that is merely encountered, rather than consciously constructed, and thus it fails to express any internal connection between the subject and its predicate. By using the Law of Determinability Maimon thought to provide a way both to generate new concepts as well as to decipher the basic categories of thought. Mathematics provides us with an example of how should these derivations work. However, this law also seems to point out the unreliability of empirical judgments.

Determinability then provides a standard of synthetic judgments: it tells us not merely the form that such judgments must take, but also specifies what counts as a legitimate content of such a judgment. In this sense, determinability provides a certain cohesion between the products of the mind. But, while the Law of Determinability presents the standard which real thought must meet, it is important to note that Maimon remains dubious about the possibility of ever achieving real thought except in the realm of mathematics. Only when the determinable relation between the subject and predicate of a judgment is *constructed* can real thought be reached; in empirical judgments about the world, no such determinable relation can be proved. And it is just this concern that leads to Maimon's skepticism.

5. From Dogmatism to Skepticism (and Back?)

Maimon describes his position as “Humean dogmatism,” (or, alternately, as “Leibnizian skepticism”), and this characterization is apt. His commitment to both camps, however, makes the question of his ultimate allegiance a difficult one. While Maimon agrees with the rationalists about the standards provided by reason, he claims to follow Hume in denying that there can ever be a proof of the applicability of reason to the world of experience. In this respect, the status of the principle of determinability encapsulates Maimon's position: while we can comprehend what the standard of real thought involves (since we have an example in mathematics), we can never be sure about its application to empirical matters (since we cannot guarantee the requisite determinable relation between subject and predicate). Maimon's skepticism arises from the lack of the required ground for this use [of the

categories], namely, the insight into the relation of determinability (that the subject, as the determinable, can be an object of consciousness in itself, while the predicate cannot be so in itself, but only as a determination [of the subject]). The categories are therefore, according to me, determined not for empirical use, but only for the objects of mathematics determined *a priori* (*GW* V, 495-6 | VnL 437-8).

Maimon here exposes what seem to be the skeptical implications of his rationalism, for while the standards of real thought are available to us as finite cognizers, the satisfaction of these criteria remains beyond our power.

But while Maimon's skepticism commits him to the view that human knowledge remains incomplete, he does not abandon the notion of intellectual progress. While we can despair of ever reaching complete knowledge, the rationalist project at least holds out the hope of advancing us in our conceptual grasp of the world. In this sense, while Maimon is led by his skeptical conclusions to see human cognition as essentially antinomial, he retains the prospect of a solution:

> Thought in general consists in a relation of a form (a rule of the understanding) to matter (the given subsumed by [the form]). Without matter one cannot attain consciousness of the form, and consequently the matter is a necessary condition of thought; that is, for the real thought of a form or rule of the understanding, there must necessarily be given a matter to which this form relates. On the other hand, however, the completeness of the thought of an object requires that nothing be given in [this completeness], but rather that everything must be thought. Since we cannot deny either of these demands, we must therefore try to satisfy both, in that we make our thought ever more complete, whereby the matter always approaches the form, through infinity -- and this is the solution of this antinomy (*GW* III, 186-7 | PW 162-3).

This relation between skepticism and rationalism is nicely captured in a biblical metaphor Maimon offers:

> Reason, which in its theoretical employment is conditioned by given objects, and is thus very limited, is now, in its practical use in relation to the will, absolute. The principle which it presents is both determined in itself, and in its application is capable of no illegitimacy. This highly pleasant prospect is certainly doubted by the skeptic, who reduces any law, as an original fact, into many. An uplifting and at once humbling voice calls out to him: “You should see the promised land from afar, but you may not enter it!” [Deut 34:4] Still, fortunately the seeing and the entering are the same: for those who boast of being able to enter can, for their legitimation, do no more than show the distant view (*GW* VII, 554).

The vision of the promised land is an especially apt characterization of Maimon's skepticism: we can see what conditions would need to be satisfied for real cognition, but we are barred from ever knowing if these conditions are met. When pressed for a warrant for his claims to certainty, the dogmatist can do no more than simply point to the far-off land of philosophical milk and honey.

6. Ethics

Unlike Kant and most of the German Idealists, Maimon denies that practical reason enjoys a primacy over theoretical reason. For Maimon, the force of both morality and epistemology resides in the notion of universal validity: as rational beings we are bound by both philosophical truth and moral duty. As Maimon puts the matter, “the moral good is good only because it is true.” (*GW* II, 405 | VT 409) As such, the theoretical and the practical go hand in hand; for Maimon, it makes no sense to follow Kant's claim to have denied knowledge to make room for faith. Rather, Maimon revives something like the Aristotelian notion that the highest virtue and pleasure are found in philosophical contemplation. This results, Maimon claims, because both practical and theoretical cognition follow from the same notion of freedom:

> Just as I have produced the principle of practical cognition from the mere widening of the theoretical, so I find practical freedom from the mere widening of what is theoretically given as a fact, and this concept of freedom first makes possible the use of this principle (*GW* VII, 275 | KU 273).

Our cognitive situations in the theoretical and moral arenas are then identical; in each, the conflicts that characterize our cognitive capacities can be resolved only by assuming that our cognition is a ‘schema’ of an infinite intellect, which is to embrace a type of dogmatism that echoes Spinoza, Leibniz and Wolff.

The similarities between the theoretical and the practical realms allow for an explanation of the possibility of freedom, even though “this concept [of freedom] is capable of no empirical presentation.” (*GW* VII 241 | KU 241) Thought, Maimon argues, is an absolutely free activity of the faculty of cognition, which is not determined *a priori* by natural laws, but rather according to the laws of the faculty of cognition itself. The will that is related to the faculty of cognition (the will to think) is likewise not determined through objects of thought, but rather through the *a priori* form of thought (which precedes the actual thought of these objects). We thus have an instance of a free will in general (*GW* VII 242 | KU 240).

Although Maimon's account of freedom is related to Kant's position, Maimon diverges from the critical philosophy by insisting that since the moral law provides only the form of the determination of the will, the actualization of this form “must be connected to an originally agreeable feeling (which does not arise from habit).” (GW VII 243 | KU 241) This feeling, Maimon argues, must be understood not in terms of a particular sensuous desire, but rather as “abstracted from all individuality.” (*GW* VII 245 | KU 243) Here the close connection between the theoretical and the practical spheres is again important, for Maimon argues that the ‘abstract feeling’ is best understood in terms of a ‘drive for cognition’:

> Man considers himself as an object of nature, and consequently as a limited being, and yet, since his faculty of cognition extends to all possible objects, he finds himself in a position to strive to infinity, and to get ever closer to the infinite faculty of cognition (divinity). Can a greater worth for a being be thought than to get closer to divinity? And must not all other

> motives vanish in the face of the motives of cognition and morality (whereby all lofty preferences extend to outer actions)? Here we have the prevailing motive of morality, whose power no one who has considered the case can doubt (*GW* VII 246-7 | KU 244-5).

The role played by the pleasure of the striving after knowledge however, recedes in importance as Maimon becomes more skeptical about the possibility of explaining the motivations that lead to moral actions. Perhaps the clearest formulation of Maimon's later position is in ‘The Moral Skeptic,’ a late work from 1800, where a sketch of the difference between moral dogmatism and skepticism is provided. As in the theoretical realm, Maimon focuses on the problem of the legitimate application of universal rules to particular cases (a problem that Kant in the *Groundwork* is well aware of). Just as we can be skeptical of whether the category of causality is legitimately applied to particular intuitions, so too can the connection between the moral law and particular actions be called into doubt. While the moral law presents the rule of conduct that I ought to follow, there is no way, Maimon claims, to determine whether in fact I act only according to it, or whether other motivations have insinuated themselves into the action. Kant, of course, agrees with this point, but Maimon attempts to draw a more skeptical conclusion from it than Kant does. On Maimon's view, one cannot determine whether an action simply accords with the moral law “and hence possesses mere legality” or whether it in fact follows from the motive of duty itself, and as such lays claim to morality. This uncertainty arises, Maimon argues, because the moral character of an action is not immediately present to cognition. Rather, one can ascribe a moral character to a person or action only after excluding all other possible motives as insufficient to explain the action -- but, Maimon notes, such a strategy would require an ‘infinite faculty of cognition’ in order to accomplish this task. Given the uncertainty about motivations, Maimon argues that the moral law can at best be viewed as an “idea that provides only a regulative use (for legality).” (*GW* VII 547 | MS 285) As such, while the moral law presents a universal command, it cannot be shown to be the ground of human actions, since -- to employ a term from Maimon's theoretical philosophy -- there is no determinate relation between the moral law and particular actions. For Maimon, the moral law stands, as do the Kantian categories in the theoretical realm, as a ‘castle in the air,’ incapable of reaching the solid ground of particular actions.

7. Maimon's Influence on the Formation of German Idealism

Kant's high regard for Maimon was shared by a number of the important figures of the movements of German Enlightenment and Idealism. Mendelssohn, of course, had initially recognized the genius beneath Maimon's course exterior, and Reinhold too took Maimon seriously, although their relation soured after Maimon published a volume of their increasingly acrimonious correspondence without Reinhold's permission. On a much more positive note, Fichte wrote that his admiration for Maimon's talent “[k]nows no limit,” and he continues that “Maimon has completely overturned the entire Kantian philosophy as it has been understood by everyone until now.” (*Gesamtausgabe* III, 2: 275)

Although Maimon's positive system was of interest to the German philosophical community in the 1790s, his greatest influence was as a skeptic, and in particular as a critic of the ‘cognitive dualism’ that

characterized Kant's critical philosophy. The force of Maimon's objections to Kant was felt most directly by Fichte, who in the *Wissenschaftslehre* devotes considerable attention to Maimon's skeptical challenge to Kant. For Fichte, Maimon pointed out the fatal shortcomings of Kant's system, but failed to move beyond a skepticism about the prospects for knowledge. This is reflected in Fichte's claim that Maimon's position "would ground a skepticism that taught us to doubt our own existence." (*GA* I, 2: 369) Fichte sought to remedy this situation by developing a more thoroughgoing idealistic account of experience, one which placed the notion of 'positing' at the center of the system. Only in so doing, Fichte claims, can both Kant's dualism and Maimon's skepticism be avoided.

In the *Wissenschaftslehre*, Fichte expresses the central themes of Absolute Idealism: the centrality of a productive faculty of mind, as well as an embrace of a rationalist methodology and a concomitant rejection of any type of dualism. In this respect, the project of Absolute Idealism owes much to Maimon, who himself had developed just these themes in his various works. But where Fichte -- and Schelling and Hegel -- stand as optimists about the prospects for this type of philosophical inquiry, Maimon remains much less sanguine about the ability of finite minds to reach a final 'System' of philosophy. It is perhaps testament to the force of Maimon's skepticism that while the Systems of German Idealism have collapsed, the skeptical challenge Maimon poses for accounts of experience remains vital and forceful -- if unduly neglected -- today. Maimon, that is, anticipates key points of the contemporary debates about the nature of Givenness, and his skeptical position offers and interesting and novel perspective on discussions of the issue.

Maimon also seems to have played an important role in the reception of Spinoza in German Idealism. Maimon was not only the first to try to create a synthesis between Kantian idealism and Spinozistic pantheism (*GW* III 455), but he was also apparently the first to suggest that Spinoza's philosophy was not atheistic, but rather a strong though unorthodox religious view. Instead of characterizing Spinoza's philosophy as atheism, a view which denies Gods existence, Maimon argues that it should rather be called acosmism, insofar as it denies the reality of the diverse world and affirms the sole reality of God. This characterization of Spinoza was later adopted by Hegel (See Hegel's *Encyclopedia Logic*, sections 50 and 151, and his discussion of Spinoza in the *Lectures on the History of Philosophy*) and seems to play a central role in Hegel's perception of Spinoza.

8. Jewish Philosophy and Culture

Being raised in an East-European Jewish surrounding Maimon's thought was influenced by the major intellectual movements of this arena. Talmud, *Kabbalah*, contemporary *Hassidism* and medieval Jewish philosophy played a crucial role in the formation of Maimon's philosophy as well as the style of his writings. The strongest influence was that of Maimonides, whose unbiased and strict rationalism Maimon took as a guiding example throughout his life (even when he no longer adhered to Maimonidean metaphysics). Following Maimonides, Maimon held intellectual perfection as the ultimate human end, and saw moral perfection merely as means for achieving this end. Like Maimonides, Maimon argued that God's image in humanity is the intellect and that to the extent that we activate and develop our intellectual capacities we become closer and more similar to God.

Maimon's relation to the *Kabbalah* was a bit more ambivalent. While he had no sympathy for the anthropomorphic teachings of some of the major Kabbalistic works, Maimon attempted to disclose the rationalistic core of the *Kabbala,* which he identified with Spinoza's pantheistic teachings. In his early Hebrew writings Maimon develops the view that God is also the *material cause* of the world (i.e., that all things are merely predicates of God, who is their substratum). Since Maimon conceives God to be a pure intellect, the result is a genuine form of radical -- and pantheistic -- idealism. This form of idealism plays a significant role in Maimon's thought in the 1790s, and possibly also in the development of German Idealism in general.

Maimon also seems to borrow from Kabbalistic and contemporary Hassidic writings the idea of an infinite process through which one "strives to turn matter into form," though he will interpret this formula in metaphysical rather than ethical terms.

In his early Hebrew writings Maimon expresses a deep interest and curiosity about astrology and magic, an interest that he returns to as a contributor and editor of the *Magazin zur Erfahrungsseelenkunde*.

After his migration from Lithuania to Germany, Maimon entered the circles of the *Haskala* (the Jewish Enlightenment movement) in Berlin. Maimon shared with this circle the idea that there is a need to propagate the enlightenment and scientific education among traditional Jews; yet, he seemed to have a very different understanding of what Enlightenment is. While for the Berlin *Haskala,* 'Enlightenment' was primarily the attempt to acculturate the Jewish masses in order to allow their acceptance into modern German society, Maimon's idea of Enlightenment was that of propagating science and philosophy. This understanding of Enlightenment was deeply imbedded in Maimon's inheritance of Maimonidean philosophy, which took philosophy and the sciences to be the highest stages of religious work, through which one comes to know God in the deepest sense. This attitude is clearly demonstrated in Maimon's 1791 commentary on the *Guide of the Perplexed* -- *Giva'ath ha-Moreh* (Hebrew: The Hill of the Guide). In this work Maimon frequently interprets the claims of Maimonides according to 18th century science and philosophy (especially Kant). While this form of intentional anachronism reveals Maimon's view of philosophy as a perennial discourse, it was also designed to serve the propagation of modern science and thought among its readers (Maimon himself explains in a similar way Maimonides' decision to open his legal codex, the *Mishne Torah*, with a summary of Aristotelian first philosophy).

Like other members of the Jewish enlightenment, Maimon criticized traditional Jewish society, and primarily the Talmudists, for their prejudices and idleness. Yet, along with this straightforward criticism, Maimon also expressed a deep appreciation for the sharpness, devotion and moral character of the Talmudists. In his *Autobiography*, Maimon writes that he "would have to write a book, had I wished to answer all the unjust charges and ridicule brought against the Talmud by both Christian authors as well as wishing-to-be-enlightened Jews." (GW I 172 | LB I 172) Here Maimon gives a detailed picture of various streams and aspects of the Jewish culture. In most cases his account is a masterpiece of a thoughtful, thoroughly informed, and unbiased exploration of one's own culture.

Maimon's reception by both traditional and enlightened Jews was quite poor. In a few texts Maimon was

lumped together with Spinoza and Acosta to form "the great chain of Jewish heretics," but mostly Maimon's writings and philosophy were ignored. The traditional community could not forgive him his infidelity and his desertion of their ranks (A certain literary source tells us that in Maimon's funeral the children of the nearby Jewish community of Glogau ran after his coffin and threw stones on it. Maimon's corpse was buried at the margin of the Jewish cemetery in Glogau. When Maimon's friend, Graf Kalkreuth, asked why he was treated with such disrespect, he was told that "the edge of the cemetery is an honorary place designated traditionally for philosophers and their like"). For the enlightened Jews of Germany, Maimon was too much of an *Ostjude* and had too much sympathy for and similarity to the Talmudists. Furthermore, Maimon never took part in the attempt to define the "essence" of Judaism and by this to provide a theology that would imitate and be able to compete with modern Protestant theology. Being thoroughly knowledgeable about the variety of aspects and streams of Judaism, Maimon simply *could not* participate in this reductive project which was quite central to modern Jewish philosophy. Thus, in spite of the fact that Maimon's erudition in Jewish literature was hardly equaled by any other modern Jewish thinker, Maimon's name is omitted in many, if not most, 20th century surveys of modern Jewish philosophy.

Bibliography

Original editions of Maimon's main Works:

- *Versuchüber die Transcendentalphilosophie (=VT)* (Berlin: Christian Friedrich Voé und Sohn, 1790).
- *Gibeath Hamore* (Hebrew: *The Hill of the Guide*), (Berlin: 1791).
- *Philosophisches Wörterbuch, oder Beleuchtung der wichtigsten Gegenstände der Philosophie, in alphabetischer Ordnung (=PW)* (Berlin: Johann Friedrich Unger, 1791).
- *Salomon Maimon's Lebensgeschichte. Von ihm selbst geschrieben und herausgegeben von K. P. Moritz (=LB)*. (Berlin: Friedrich Vieweg, 1792-3).
- *Ankündigung und Aufforderung zu einer allgemeinen Revision der Wissenschaften: Einer königl. Akademie der Wissenschaften zu Berlin vorgelegt von Salomon Maimon*. (Berlin: Johann Georg Langhoff, 1792).
- *Bacons von Verulam neues Organon*, edited with notes by Maimon, 2 volumes (Berlin: Gottfried Carl Nauck, 1793).
- *Über die Progressen der Philosophie (veranlat durch die Preisfrage der königl. Akademie zu Berlin für das Jahr 1792: Was hat die Metaphisik seit Leibniz und Wolf für Progressen gemacht?)*, (Berlin: Wilhelm Vieweg, 1793).
- *Salomon Maimon's Streifereien im Gebiete der Philosophie*, (Berlin: Wilhelm Vieweg, 1793).
- *Die Kathegorien des Aristoteles ("With notes and as a propadeutic to a new thoery of thinking")*, (Berlin: Ernst Felisch, 1794).
- *Anfangsgründe der Newtonischen Philosophie von Dr. Pemberton*. (translated from the English with introduction and notes by Maimon), First Part (Berlin: Friedrich Maurer, 1793).
- *Versuch einer neuen Logik oder Theorie des Denkens. Nebst angehängten Briefen des Philaletes an Änesidemus (=VnL)*, (Berlin: Ernst Felisch,1794).

- *Kritische Untersuchungen über den menschlichen Geist oder das höhere Erkenntnié und Willensvermögen (=KU)*, (Leipzig: Gerhard Fleischer, 1797).

Modern editions and translations of Maimon's Works:

- Maimon, Salomon. *Gesammelte Werke (=GW)*, edited by Valerio Verra, 7 volumes (Hildsheim: Olms, 1965-1976).
- -----. *Salomon Maimons Lebensgeschichte*, edited by Zwi Batsch, (Frankfurt a. M: Insel Verlag, 1984).
- -----. Giva'ath ha-Moreh, edited by S.H. Bergmann and N. Rotherstreich (Jerusalem: Israeli Academy of Science, 1965 [Reprint 2000]). [Maimon's Hebrew commentary on the first part of Maimonides' *Guide of the Perplexed*].
- -----. *Letters of Philaletes to Aenesidemus*, trans. by George di Giovanni in di Giovanni and H.S. Harris (eds.*) Between Kant and Hegel: Texts in the Development of Post-Kantian Idealism* (Indianapolis: Hackett, 2001) [The only available English translation of any of Maimon's philosophical works].
- -----. *The Autobiography of Solomon Maimon*, translated by J. Clark Murray (Urbana and Chicago: University of Illinois Press, 2001)
- -----. *Commentaires de Maïmonide*, edited and translated into French by Maurice-Ruben Hayoun (Paris: Cerf, 1999). [A French translation of two major texts of Maimon dealing with Maimonides' *Guide of the Perplexed*: Chapters 1-10 of the second part of Maimon's *Lebensgeschichte*, and Maimon's Hebrew commentary on the *Guide*, *Giva'at ha-Moreh*]

Unpublished Manuscripts

- *Heshek Shlomo* (Hebrew: *Solomon's Desire*), Posen 1778. Currently held by the National and University Library, Jerusalem (MS 806426). [This 300 pages long manuscript is comprised of five -- rather independent -- works. The first text -- *Ma'ase Nissim* -- is a commentary on the homiletic book of the medieval talmudist, Nissim of Gerondi. The second text -- *Eved Avraham* -- is a super-commentary on Avraham Ibn Ezra's commentary on the Pentateuch and Psalms. The third text -- *Ma'ase Livnat ha-Sapir* -- is an attempt to harmonize some main Kabbalistic doctrines with the trachings of Maimoides. The fourth section -- *Ma'ase Hoshev* -- is an algebra textbook The fifth and last section -- *Avarchecha Bahya* -- is a short exposition of a biblical commentary by the 11th century Jewish philosopher, Bahya Ibn Pakuda].
- *Ta'alumoth Hochma* (Hebrew: *Mysteries of Wisdom*), Breslau 1786. Currently held by the Bodelian Library, Oxford (MS Mich.186).

Missing Manuscripts

Throughout his wanderings Maimon kept manuscripts of several works which he never published. After his death, Maimon's patron, Graf Kalkreuth, gave the manuscripts to Benjamin Fraenkel from the nearby Jewish community in Glogau. During the 19th century Maimon's manuscripts and letters were dispersed

and held by several prominent Jewish scholars and bibliophiles such as Abraham Geiger, Leopold Zunz and Heimann Michael. The collection of the latter -- which included Maimon's *Ta'alumoth Hochma* -- was purchased by the Bodleian library in 1848. Several other manuscripts found different ways to the collection of the Berlin *Hochschule für die Wissenschaft des Judentums* where they were held till World War II. Before and during the war the *Hochschule's* collection was smuggled out of Germany. The most important Maimonian MS held by the *Hochschule* arrived, after a long odyssey, at the National and University Library in Jerusalem. Other manuscripts of Maimon, held before the war by various individuals and institutions, are still missing. The authors of this entry participate in the search after these missing manuscripts, the most important of which are:

1. The Early Commentary on Maimonides' Guide. [A text which Maimon wrote before his first arrival in Berlin].
2. Maimon's translation of Mendelssohn's *Morgenstunden*.
3. Parts two and three of *Giva'ath ha-Moreh*. [In his *Autobiography* Maimon writes that these parts were ready to print and were supposed to be published soon thereafter. Apparently, the publishers were not interested in the work due to its lack of popular appeal].
4. A fragmentary commentary on Aristotle's Ethics.
5. *Die Mysterien der Philosophie* ("The Mysteries of Philosophy"). [Apparently, a complete book, ready for print].
6. *Über Logik* ("On Logic"). [Mainly remarks on Kiesewetter's book on logic].

Selected Secondary Literature

- Atlas, Samuel. *From Critical to Speculative Idealism: The Philosophy of Solomon Maimon* (The Hague: Nijhoff, 1964).
- Baumgardt, David. "The Ethics of Salomon Maimon" *Journal of the History of Philosophy* 1 (1963), 199-210.
- Beiser, Fredrick C.. *The Fate of Reason* -- *German Philosophy from Kant to Fichte*, (Cambridge Mass.: Harvard University Press, 1987), 285-323.
- Bergman, Shmuel Hugo. *The Philosophy of Solomon Maimon*, translated from the Hebrew by Noah J. Jacobs (Jerusalem: Magnes Press, 1967).
- Bransen, Jan. *The Antinomy of Thought: Maimonian Skepticism and the Relation between Thoughts and Objects* (Dordrecht: Kluwer, 1991).
- Buzaglo, Meir. *Solomon Maimon: Monism, Skepticism and Mathematics* (Pittsburgh: Pittsburgh University Press, 2002).
- Engstler, Achim. *Untersuchungen zum Idealismus Salomon Maimons* (Stutgart-Bad Cannstatt: Fromann-Holzboog, 1990).
- Franks, Paul. "All or nothing: systematicity and nihilism in Jacobi, Reinhold and Maimon" in Karl Ameriks (ed.) *The Cambridge Companion to German Idealism* (Cambridge: Canbridge University Press, 2000), 95-116.
- Guéroult, Martial. *La philosophie transcendentale de Salomon Maimon* (Paris: Societe d'edition, 1930).

- Jacobs, Noah J.. “Schrifttum über Salomon Maimon: Eine Bibliographie mit Anmerkungen” in *Wolfenbütteler Studien zur Aufklärung* 4 (Wolfenbüttel: 1977), 353-395. [A comprehensive bibliography of Maimon’s writings and secondary literature. Originally appeared in the Hebrew journal *Kiryat Sefer* 41 (1966). No updated bibligraphy is available].
- Kunze, Friedrich. *Die Philosophie Salomon Maimons* (Heidelberg: Carl Winter, 1912).
- Lachterman, David. “Mathematical Construction, Symbolic Cognition and the Infinite Intellect: Reflections on Maimon and Maimonides” *Journal of the History of Philosophy* 30 (1992), 497-522.
- Thielke, Peter. “Getting Maimon’s Goad: Discursivity, Skepticism, and Fichte’s Idealism,” *Journal of the History of Philosophy* (2001).
- Weissberg, Liliane. “Salomon Maimon writes his Lebensgeschichte” in Sander L. Gilman and Jack Zipes(eds.), *The Yale Companion to Jewish Writing and Thought in German Culture* (New Haven: Yale University Press, 1997), 108-15.
- Zac, Sylvain. *Salomon Maimon -- Critique de Kant*, (Paris: Cerf, 1988).

Other Internet Resources

- *Berlinische Monatsschrift* (U. Bielefeld Library)
  [Maimon published a number of articles in the *Berlinische Monatsschrift*. You can use the site's search tools to locate these articles as well as some contemporary responses].
- ‘Studying the Kabbala’
  [Excerpts from Maimon's discussion of the *Kabbala* in his *Autobiography*, 1792 (in English translation)].
- Maimon's *Autobiography*
  [Almost the complete text of Murray's English translation].
- Salomon Maimon website
  [An excellent site developed by Florian Ehrensperger (Philosophy/ University of Munich)].

Related Entries

Fichte, Johann Gottlieb | Judaic Philosophy | Kant, Immanuel | Leibniz, Gottfried Wilhelm | Maimonides [Moses ben Maimon] | Mendelssohn, Moses | Spinoza, Baruch [Benedict]

Peter Thielke
peter.thielke@pomona.edu
and
Yitzhak Melamed
yitzhak.melamed@yale.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 28, 2002
Content last modified: July 2, 2002

Friedrich Heinrich Jacobi

Polemicist, socialite, and literary figure, Jacobi was an outspoken critic, first of the rationalism of German late Enlightenment philosophy, then of Kant's Transcendental Idealism, especially in the form that the early Fichte gave to it, and finally of the Romantic Idealism of the late Schelling. In all cases, his opposition to the philosophers was based on his belief that their passion for explanation unwittingly led them to confuse conditions of conceptualization with conditions of existence, thereby denying all room for individual freedom or for a personal God. Jacobi made this point, in defence of individualism and personalistic values, in a number of public controversies, in the course of which he put in circulation expressions and themes that resonate to this day. He was the one who invited Lessing, who he thought was walking on his head in the manner of all philosophers, to perform a *salto mortale* (a jump heels over head) that would redress his position and thus allow him to move again on the ground of common sense. He was also responsible for forging the concept of 'nihilism' -- a condition of which he accused the philosophers -- and thereby initiating the discourse associated with it. His battle cry, which he first directed at the defenders of Enlightenment rationalism and then at Kant and his successors, was that 'consistent philosophy is Spinozist, hence pantheist, fatalist and atheist'. The formula had the effect of bringing Spinoza to the centre of the philosophical discussion of the day. In the face Kant and his idealistic successors, Jacobi complained that they had subverted the language of the 'I' by reintroducing it on the basis of abstractions that in fact negated its original value. They had thus replaced real selfhood with the mere illusion of one.

But perhaps the most influential of Jacobi's formulas was the claim that there is no 'I' without a 'Thou', and that the two can recognize and respect one another only in the presence of a transcendent and personal God. Because of his defence of the individual and the 'exception', Jacobi is sometime taken as a proto-existentialist. This view must be balanced by the consideration that Jacobi was a defender of conservative values that he felt threatened by the culture of the day; that he never considered himself an irrationalist; on the contrary, that he thought his 'faith' to be essentially and truly rational; and that he tried more than once to develop a positive theory of reason. As a literary figure, he criticized the *Sturm und Drang* movement and dramatized in two novels the problem of reconciling individualism with social obligation. An exponent of British economic and political liberalism, Jacobi was an early critic of the French revolution, the destructiveness of which he considered the practical counterpart of the speculative nihilism of the philosophers.

1. Life and Intellectual Carrier

Like his junior contemporary Goethe (1749-1832), Friedrich Heinrich Jacobi was blessed with a long life, at least as measured by the standards of the time, and had the good fortune of witnessing in its course events that radically altered the cultural and political face of Western Europe.[1] Jacobi saw the Enlightenment (*die Aufklärung*) take hold in the German lands, and reach maturity in a peculiarly German cultural form. He saw it as it nurtured ideas of subjectivity and nature, and an interest in history, that were the precursors of later Romanticism. He saw it also as it was violently, even brutally, disrupted in the last decade of the eighteenth century by the events surrounding the French Revolution; and, finally, as it gave way to the new order of the nineteenth century. Despite long bouts of illness, Jacobi was deeply involved in these world shaking events, and actively contributed to them. He may not have had the political and military engagement of a Goethe, who could combine vast literary activities with weighty state responsibilities, but he was certainly an influential cultural and political commentator. Jacobi contributed enormously, throughout his active life, to the shaping of educated German public opinion. He carried on a most extensive correspondence with just about all the literary and political personalities of the day, and for many years also provided a meeting place for them at his country estate in Pempelfort, by hosting there a very popular literary salon. He was ably aided in this enterprise by his much admired wife Elisabeth von Clermont (universally known as 'Betty'). The list of names who frequented the salon, or with whom Jacobi entered in correspondence on different occasions of his life, reads as a Who's Who of the age. Heinse, Wieland, Goethe, Lavater, Herder, the Humboldt brothers, Diderot, Hemsterhuis, Hamann, Dohm, Georg Forster, the duchess Anna Amalia (Sachsen-Weimar), F. L. Stolberg, Fürstenberg, Princess Gallitzin and Sophie La Roche, counted among them, to mention only a representative segment.

In an age, moreover, in which friendship had been raised to the status of cult, Jacobi nurtured a number of intimate personal relations, all of which influenced his literary production in many ways. There was, for instance, his life long tempestuous relationship with Goethe -- punctuated as it was by periods of extreme intimacy and extreme alienation, and culminating in an irreparable break in the final period of Jacobi's career. (Nicolai, 1965) There was also his warm friendship with Hamann -- for the most part at a distance, since the two men managed to meet physically only once, in what were to be Hamann's last days, and this much anticipated meeting unfortunately devolved into a huge quarrel. As long as it lasted in the medium of letters, the friendship produced a body of correspondence widely recognized as one of the finest contributions to the German literature of the time. (Jacobi/Hamann, 1955-1979) There was then Jacobi's relationship with Princess Galitzin, much cultivated by Jacobi but overshadowed by the Princess's conversion to Catholicism and by her adoption of a quasi monastic style of religiosity

distasteful to Jacobi (Brachin, 1952; Trunk, 1955); his relationship with the Stolbergs, also embittered at some point by the Stolbergs'entrance in the Catholic Church (Brachin, 1952); with Lavater, whose pietism seem to have had a special attraction for Jacobi; with Basedow and Claudius,to both of whom Jacobi entrusted the education of his two surviving legitimate sons; and many other as well. These friendships, while engaging Jacobi emotionally, were perceived by him also as integral part of his intellectual life, since his fundamental belief, from beginning to end, was that thought must resist the lures of abstraction and remain throughout personal. Jacobi can perhaps best be characterized as a philosophical polemicist, but only in the sense that polemics is the natural medium of expression for a philosopher whose thought is essentially personalist, and for whom, therefore, (just as for Kierkegaard roughly half a century later) there is no serious thinking unless it is directed to someone in particular.

Jacobi also had the good fortune of belonging to a privileged segment of society. The Enlightenment as well as the French revolution presented him with more than just an intellectual challenge. In their different ways (which, as will be seen in Section 3, Jacobi however thought to have much in common), they both threatened the system of values that justified his social status. In the case of the French revolution, the threat was physical as well, since the events in France spilled over across the Rhine and were the cause for a time of much dislocation in his life. There were, in other words, interests other than purely intellectual that must have motivated his reaction to philosophy and philosophers and which would have to come under close scrutiny in a full study of Jacobi's life. They can however be abstracted from when the main interest, as here, is with the internal logic, or lack of it, of his view of things. A quick review of the main events and the more influential circumstances of his life is none the less in order, at least as signposts of his intellectual career.

Jacobi was born in Düsseldorf, on January 25, 1743, the second son of a well-to-do merchant. His older brother, Johann Georg, was the one selected by the family for an intellectual career, and he did achieve indeed a certain notoriety as a poet in the sentimental style of Gleim. The task of pursuing the commercial activities of the family was reserved for the younger Jacobi. There also were two daughters in the family, half-sisters to Jacobi. Neither of them married, and they eventually took over the management of Jacobi's household. According to his own account, (Jacobi, 1785: 8; 1787, 67ff; 1789, 328ff) Jacobi was temperamentally given to extremes of piety from his early youth, and was constantly troubled by questions concerning the existence of God and the endlessness of time. After confirmation, he joined a society of pietists in whose company these tendencies were reinforced. At sixteen, after a brief and disappointing apprenticeship in a Frankfurt commercial house, he was sent in 1759 to Geneva to develop there the social skills required for his appointed vocation. It is in that city that Jacobi, exposed to French ideas, as he matured socially, at the same time also formulated the basic beliefs that were to guide his intellectual career to the end. Under the tutelage of the famous mathematician G.L. Lesage, he became acquainted with traditional scholastic metaphysics. But he also worked on the writings of Charles de Bonnet, in which he found elements of psychological sensualism combined (allegedly even harmonized) with such Christian beliefs as the after-life, and read Rousseau's *Emile*(1761). Appended to the latter was the so titled 'Profession of Faith of a Savoyard Vicar', a manifesto upholding the rights of the 'heart'against the religion of reason popular among the rationalist philosophers of the day. Jacobi's resolution, at the time, was never to accept a system of thought unless it could be tested against actual existence and did not contradict his yearnings for God.

Back in Germany in 1762, Jacobi continued his philosophical studies. He eagerly read the essays submitted by Moses Mendelssohn and Kant in response to the competition sponsored that year by the Berlin Academy on the theme, 'Concerning Evidence in the Metaphysical Sciences'. According to his own account, Jacobi found Kant's essay much more convincing than Mendelssohn's, though first prize was awarded to the latter. (Jacobi, 1787: 74-75) He apparently also engaged in the study of Spinoza, and was impressed by Kant's essay on 'The Only Possible Ground for a Proof of God's Existence'. (Jacobi, 1787: 78-88) In 1764 he took charge of his father's business and -- after an affair with a housemaid with whom he conceived a son, both of whom (housemaid and son) he treated very shabbily by modern standards (Booy/Mortier, 1966) -- married the richly endowed Betty von Clermont. Together with her, he established the already mentioned literary salon at Pempelfort.

In 1772 Jacobi entrusted all his business affairs in the hands of his capable brother-in-law and accepted an invitation to join the governing body of the Duchy of Julich and Berg. His commission was to supervise the rationalization of local manufacturing and taxation practices. In 1779, he was also invited to Munich to conduct a similar reform in the Bavarian lands. In the same year, however, these official activities also came abruptly to an end. Jacobi's liberal views quickly came under attack in the Court at Munich, and, embroiled in bitter controversies, Jacobi returned to his native Düsseldorf. The literary fall out of this episode was the publication (Jacobi, 1779 (1) &(2)) of two essays in which Jacobi defended Adam Smith's economic views. The first essay was published while Jacobi was still in office in Munich, and was itself a major cause of his troubles there.

While engaged in these wordily affair, Jacobi had not abandoned his more humanistic interests. He had already ventured in some minor literary publications in association with his brother Georg while still busy with his commercial practice in Düsseldorf. In 1772, together with Wieland, whom Jacobi had come to know through his brother, he laid out the plans for a German journal patterned after the *Mercure de France*. The *Deutcher Merkur*made its appearance soon after, under Wieland's editorship. Jacobi used the journal as the vehicle of some of his occasional writings until 1777, when his collaboration with Wieland came to an end because of a bitter disagreement between the two men on political issues. (Jacobi, 1781, 1782) More significant for Jacobi's future were two other events that took place in the following years. The first was Goethe's visit to Pempelfort in the company of Basedow and Lavater. The encounter between the already famous poet and Jacobi that followed was, according to the reports left by both, of a highly emotionally charged nature. (Prantl, 1881: 579) Each man made a deep impression on the other. Jacobi's two novels, *Allwill's Briefsammlung*and *Woldemar*, owe their origin to that event. They were Jacobi's response to Goethe's ardent request, made during the visit, that he put in writing all that was closest to his heart. As things turned out, Goethe quickly cooled towards Jacobi, much to the latter's disappointment, and was later (1779) to prove especially cruel towards him by nailing to a tree ('crucifying'), to the great amusement of a large company, a copy of a just published new edition of the *Woldemar*(Stockum, 1957; Sudhof, 1959; di Giovanni, 1994: 52-53).

The other event occurred on the occasion of a long journey that took Jacobi, on July 5, 1780, to Wolfenbüttel, where Lessing was librarian at the Herzog August's library. Jacobi immediately paid a visit to the famous man, and it was there, while a guest at Lessing's house, that Jacobi engaged his host in that

now famous conversation in which, according to Jacobi's report, Lessing declared himself to be a Spinozist. (Jacobi, 1785: 10-45) This alleged revelation, made only a few months before Lessing's sudden death (1781), was to be the occasion of an exchange of letters between Jacobi and Moses Mendelssohn on the nature of philosophy in general and Spinozism in particular. Mendelssohn was well known at the time as Lessing's great friend, and as himself an artificer of the German Enlightenment. The exchange began in 1783 (Jacobi, 1785: 1), and continued in a period of time when Jacobi was deeply in sorrow because of the premature death of his beloved Betty (1784) and of a younger son, and was also suffering from bad health. As the result of an intricate and unpleasant set of circumstances -- one that was to provide fodder to opposing parties, in the dispute that soon followed, for questioning the integrity of both Mendelssohn and Jacobi -- the latter published the letters in 1785, complete with commentary, under the title of *Concerning the Doctrine of Spinoza in Letters to Herr Moses Mendelssohn*(Spinoza Letters).

Nobody could have predicted the extent and the fury of the controversy that the publication of the book caused. The controversy goes in the literature under the name of the Spinozism or the Pantheism Dispute. (Jacobi, 1916) Mendelssohn responded to Jacobi in writing (1786) but did not survive to see his reply in print. Already in bad health, he died. (4 January 1786) Jacobi responded to the reply. (1786) Both publications were notable for their bitterly personal tone. In retrospect, the controversy itself and the personal tone it assumed made sense, since the issues that Jacobi had raised had put into question the nature and the value of the new humanism being sponsored by the Enlightenment. As Goethe was to remark many years later, the controversy touched everyone in their deepest convictions.[2]Mendelssohn's death caused the debate to become even more personal, for his defenders (notably the Berlin's *Aufklärer*led by Nicolai) used it as an occasion to raise their hero to the status of martyr. One can gather a detailed account of the events, seen from Jacobi's point of view, in the latter's thick correspondence with Hamann of those years. (Jacobi/Hamann, 1955-1979) Hamann sided with Jacobi, though his irrepressibly eccentric self could not refrain from repeatedly poking fun at his friend. Two unforseen results of the episode were that Jacobi, until then a peripheral author, was propelled to the centre of public attention; and Spinoza's philosophy, already influential among the literary circles of the *Sturm und Drang*but otherwise traditionally rejected by the philosophers of the schools, became an object of philosophical discussion.

In 1787 Jacobi published *David Hume on Faith, or Idealism and Realism, a Dialogue*. In it he tried to clear himself of the charge of irrationalism brought against him because of his defence of the primacy of *Glaube*('faith'or 'belief') over reason in the still on-going controversy. Though appealing to the authority of Hume in this defence, Jacobi however also forcefully distanced himself from the latter's scepticism by declaring the sense in which he was rather a realist. The dialogue appeared in the wake of the publication of the second edition of Kant's *Critique of Pure Reason*(1787), and included an appendix in which Kant's transcendental idealism was severely criticized. The appendix was to become a *locus classicus*of anti-Kantianism. It eloquently restated all the main objections that had already been raised against the Critique since its first publication, and were to be repeated again against it, in a variety of forms, in the years to come. It is there that Jacobi made the famous quip: 'Without the presupposition [of the "thing in itself,"] I was unable to enter into [Kant's] system, but with it I was unable to stay within it'. (Jacobi, 1787: 223) His point was that, in presupposing the allegedly unknown 'thing in itself', yet by assigning to it the many functions that it played in his system, Kant was in fact demonstrating knowledge of it, thereby

contradicting his assumption of critical ignorance. The negative, even antagonistic, attitude that Jacobi assumed in this appendix overshadowed the fact, however, that he had always felt close in spirit to Kant. He had felt deeply hurt when Kant has sided with Mendelssohn[3]in the still on-going dispute. More important, it also overshadowed the fact that Jacobi had already developed a much more constructive critique of transcendental idealism, and had even offered a sophisticated alternative to it, in the body of the dialogue itself, while expounding his own realism. (See Section 2.2) This is an aspect of the dialogue that was not lost on some of Kant's own disciples, and was a factor in the development of post-Kantian idealism.

There followed a number of polemical writings directed at the Berlin *Aufklärer*. They reflected Jacobi's engagement in the famous 'Starck affair'-- a defamation suit, eventually lost by the plaintiff, brought by a certain J. A. Starck against the leaders of the Berlin Enlightenment.[4](Blum, 1912; di Giovanni, 1995) The case brought into question the legitimacy of freedom of the press. It also unfolded in the context of wide-spread conspiracy theories regarding the alleged efforts by secret societies, working under the manipulation of Jesuits or Papists in general, to undermine the rule of enlightened states. (See Section 4) In 1789 a second, much enlarged, edition of the Spinoza-Letters went to press. It reflected the recent publication of Kant's *Critique of Practical Reason*(1788). Among many additions, it included two prefatory sets of theses, the first denying human freedom, the second asserting it, together suggesting a resolution of the antinomy between nature and freedom quite different than the one proposed by Kant. It also included a number of new appendices, one of them sharply criticizing Herder, whose idea of God amounted, in Jacobi's opinion, to a half-baked pantheism.[5](Jacobi, 1789: 349-357) Jacobi's prefatory theses on human freedom were to be highly influential in the debate on the subject soon to break out between K.L Reinhold and C. C. E. Schmid. (di Giovanni, 2001) The debate quickly implicated many other personalities, Fichte among them.

The following years were marked by the visits to Pempelfort of Hamann, Stolberg, Herder, and Goethe (November 1792). Soon, however, political events were to overtake philosophical discourse. The beheading in Paris of Louis XVI (1793) shocked German public opinion, and occasioned Jacobi's lament in 'Accidental Outpourings of a Lonely Thinker'. (Jacobi, 1795) In 1794, as the French revolutionary armies crossed the Rhine and began bombarding Düsseldorf, Jacobi abandoned Pempelfort and undertook a long peregrination in the North, the guest of many friends in whose company he found comfort. (Prantl, 1881: 582) He settled for a time in Eutin. At this time he began the composition of 'Of Divine Things', a manuscript that he completed and published only later. In 1799 he became involved in the so called 'Atheism dispute'. This was a controversy surrounding the alleged atheism of Fichte which Jacobi himself had not initiated but in which he none the less played a leading role. At the request of Lavater, he wrote an open letter to Fichte (Jacobi, 1799) in which he restated all his objections to philosophy in general and the new idealism in particular, and reaffirmed his commitment to the primacy of faith. He also singled out Fichte as one who had realized all his misgivings about philosophy, since he had succeeded in restating Spinozism even from the standpoint of subjectivity. In a supplement (Supplement II) to the letter, Jacobi included a long comment on the nature of reason that he reprinted in the second volume of his collected works (1815)as an independent essay under the title of 'On the Inseparability of the Concept of Freedom and Providence from the Concept of Reason'. (Jacobi, 1812-1825) In 1802, he published 'On the Undertaking of Critique to Reduce Reason to Understanding'(Jacobi, 1802 (1)) -- this

latter directed particularly at Kant.

In 1804, Jacobi suffered a drastic reversal of financial fortunes but was spared a much reduced life-style by the call from Munich, late in the same year, to work there at the reorganization of the Academy of the Sciences. Jacobi accepted the invitation and in the following year moved to Munich. He remained there until the end of his life. Elected president of the reformed Academy in 1807, he unfortunately fed the animosity already felt in Munich against the foreigners recently invited to the Academy by implying in his inaugural speech a criticism of Southern Germany culture. (Jacobi, 1807) He also antagonized Schelling, at that time Director of the Munich Academy of the Arts. In 1809, in response to a writing of Schelling on human freedom that contained pointed criticisms of his understanding of Spinoza, Jacobi resurrected his unfinished manuscript ‘Of Divine Things’, completed it as an obvious polemic against the new nature philosophy that Schelling was promoting, and published it in 1811 under the title of ‘Of Divine Things and Their Revelation’. (Jacobi, 1811) The essay provoked a reply from Schelling which, in turn, provoked a counter reply by Jacobi. Thus arose the third of the major controversies in which Jacobi was involved -- one even more personal and bitter in tone than the earlier ones.

In 1812 Jacobi retired and, aided by his disciples J. F. Köppen and C. J. F. Roth, began the preparation of his collected works. (Jacobi, 1812-25) Since the turn of the century, he had begun to attract to his side such younger figures as F. D. E. Schleiermacher, J. F. Fries, and F. Bouterwek -- newly established academics whose thought he had inspired but who, in turn, influenced his final definition of the nature of reason. They all shared a scientific positivism which they combined with a contrasting moral/religious positivism -- this latter based on the assumption that higher values and the reality of divine things can be intuitively assessed by way of feeling. This younger generation helped to popularize some of Jacobi’s ideas and attitudes in the nineteenth century. Ironically, despite his life long opposition to Roman Catholicism, Jacobi exercised considerable influence in some German Catholic theological circles. (Weindel, 1950)

In 1817, Jacobi made reference in a letter to his friend Johann Neeb to a favourable review of the second volume of his collected works by Hegel. (Jacobi, 1825-27: vol. 2, 467-68; di Giovanni, 1994: 165-166) There was no apparent trace left in his comments of the anger that Hegel’s essay, ‘Glauben und Wissen’(‘Faith and Knowledge’), had aroused in him in 1801. (Jacobi, 1803: 221) Jacobi allowed that Hegel’s interpretation of his thought might be the correct one, and conceded that the two men might come to a meeting of minds, were it not for old age that prevented him from studying Hegel’s philosophy.

Jacobi’s extensive travels took him to London in 1786 and to Paris in 1801. (Prantl, 1881: 581, 582) Though first exposed to philosophy in a French cultural milieu, Jacobi was an Anglophile throughout his life and favourably compared British philosophers with the French *philosophes*. He credited the British for never having denied that virtue has value on its own; for never having made it just an instrument of happiness. The French, on the contrary, the moment their philosophizing progressed beyond mere common sense, were always all too prone in moral matters of falling into materialism. (Jacobi, 1812-125: vol. 5, 73-74)

Jacobi never saw the completion of the planned edition of his works. He died on March 10, 1819.

2. Main Philosophical Works

2.1 The Spinoza Letters (1st edition, 1785)

The Spinoza Letters is a cumbersome work, its disjointed composition indicative of the haste with which it was put together. The first edition opens with the text of a poem by Goethe that was however dropped in the second edition. The poem carried the name of its author but had been mechanically edited by Jacobi to bring out, by printing certain key phrases in relief, images especially important to him. Thus edited, the poem conveyed the idea that the pagan gods are Man writ large, and are therefore to be praised, since they reveal what is most noble in the human being. Brute nature is without feelings or discernment, whereas humans can judge, draw distinctions, and dare the impossible. They transcend nature. This was a humanistic message with which Jacobi agreed wholeheartedly. After a brief explanation of how the exchange of letters with Mendelssohn had originated, the main text goes on with a somewhat abbreviated version of the correspondence itself. The first letter gave an account of the conversation Jacobi had had with Lessing on the subject of Spinoza during his visit to Wolfenbüttel, and included the text of another poem of Goethe -- the *Prometheus*, one that had been the occasion of the conversation in the first place. This poem, hitherto still unpublished, was printed without the author's name (some thought it was Jacobi's own) and conveyed quite a different image of the gods than was found in the other poem. It suggested that they, no less than mortal men, are subject to the blind 'almighty time'that rules over all. In a universe thus ruled by 'eternal fate', they are therefore even more wretched than human beings, for the latter have at least the power to rebel against the order of things and thereby retain their dignity. At least as typified by Prometheus, human beings can rejoice in their own sufferings, and by this defiance uphold their individuality in the face of Fate.

In the rest of the letters we see Mendelssohn trying to play down the seriousness of Lessing's alleged admission to being a Spinozist. Since Mendelssohn thought at first that Jacobi was himself a Spinozist, we see him also attacking the basic doctrines of Spinozism as he understood them, while at the same time suggesting how they could be reformulated in an internally more consistent form to have them rejoin the accepted teachings of school metaphysics. Jacobi, for his part, was intent on displaying his scholarly (and indeed more accurate) knowledge of Spinoza's philosophy. In response to Mendelssohn's criticisms of Spinoza, he argued that Spinozism represented, on the contrary, a perfectly self-consistent position. It was school metaphysics that would logically lead to Spinozism, if its implications were just fully understood, not the other way around. Jacobi also gave an explanation for his sudden decision to publish the letters. Since Mendelssohn had announced the impending publication of his *Morgenstunden*,[6]a book in which, as Mendelssohn said, he would take up the issue of Spinozism, Jacobi thought that his opponent was unfairly trying to make a public head start on him in a controversy that was agreed would have to remain private before being brought to public attention by mutual consent. Jacobi then concluded the book with rambling extensive quotations from sundry theologians, and 'sealed'it (his expression) with an excerpt from Lavater.

Mendelssohn was bitterly to complain to Kant that Jacobi's volume was like a monster that sported Goethe for head, Spinoza as torso, and Lavater for feet. There was a point to his complaint. Yet, despite the many divagations and the pervading preachy tone of the text, there was a definite philosophical message that Jacobi was conveying. Philosophers are temperamentally inclined to reconstruct reality according to the requirements of explanation, in total disregard of the requirements of existence. They are possessed, as it were, by a logical fanaticism that leads them to mistake the abstractive principles of explanation for principles of existence. Since the individuality of things is the first victim of this confusion, yet individuality is the necessary condition of all existence, it follows that in the world as reconstructed by philosophers there is no room left for truly existing subjects -- least of all for agents who can seriously take responsibility for their actions and relate to one another as person to person, an 'I'to a 'Thou'. All that is most noble in human nature (as depicted in Goethe's first poem, Jacobi implies), notably the ability to make decisions and transcend brute nature, is thus denied, and there is no alternative left for human individuals, if they care to preserve at least a modicum of dignity, except to take pleasure in their avowedly only delusory sense of freedom (according to Goethe's second poem, as Jacobi also seems to imply). The much vaunted Enlightenment humanism, based as it was on the ideal of pure rationality was only a sham. It led in fact to 'nihilism'(a term, incidentally, that Jacobi made popular).

According to Jacobi's report of his conversation with Lessing, after a disquisition on his part regarding the devastating effects of a thought based on the demands of reason alone, and upon Lessing's apparent declaration of sympathy for Spinoza, Jacobi had urged upon the latter to perform a *salto mortale*(Jacobi, 1785: 17), i.e., a kind of jump, heels over head, that would redress his position. Enlightened philosopher that he was, Lessing had been walking on his head. The jump -- which Lessing humorously declined to execute citing old age as excuse (Jacobi: 1785, 33-34) -- would have brought him to his feet, back on the solid soil of common sense. Jacobi's obvious agenda, in other words, both at the time of this meeting with Lessing and later, in the exchange of views with Mendelssohn, was to attack the rationalism of the Enlightenment. Also implied, however, was a criticism of Goethe. The latter had spearheaded the *Sturm und Drang*reaction against the same rationalism that Jacobi was now condemning. Jacobi's veiled message was that the adepts of this new cultural phenomenon had failed to escape the rationalism of the philosophers, since the rebellious new humanism they advocated made sense only on the presupposition that the philosophers's conception of reality was the right one. It is the desire to convey this message that alone explains the otherwise puzzling presence in Jacobi's volume of the two poems by Goethe. To Jacobi's contemporaries, who lacked the background knowledge of Jacobi's special relationship to Goethe, and were even unaware of Goethe's authorship of the second poem, their inclusion in the volume remained a mystery to the end.

At any rate, whether Jacobi's agenda was directed at the Enlightenment or the *Sturm und Drang*, the presence of Spinoza both in his conversation with Lessing and the later controversy with Mendelssohn was clear. Quite apart from the fact that he had been the central object of discussion in Jacobi's first encounter with Goethe -- another circumstance that Jacobi's contemporaries could not know -- Spinoza was now being portrayed by Jacobi as the one philosopher who had had the courage to press the logicism of the philosophers to its ultimate limit, and to draw from it its inevitable conclusions. Spinoza had subsumed all reality under the one highest abstraction of 'substance', with the net result that any real distinction between one presumed individual thing and any other, and between all things and God, was

being denied. True generation, or anything connected with temporality, equally disappeared, except as mere illusionary phenomena. This was as nihilistic a conclusion as, in Jacobi's opinion, one could possible draw. In his eyes, therefore, Spinoza stood as the philosopher *par excellence*(an attribute, however, which he later conferred on Fichte). According to the formula that he drew, philosophy equals Spinozism, and Spinozism in turn equals atheism -- this last part of the equation being based on the assumption that, since Spinoza's God, viz. Substance, lacked the attributes of a person, it he could not satisfy the requirements of true religiosity.

One can understand, therefore, why Jacobi's attitude towards Spinoza could have been a source of confusion for Mendelssohn and later critics as well -- to the point that Jacobi himself was often taken to be a Spinozist. On the one hand, Jacobi pointed to Spinoza as an object lesson of all that is wrong with philosophical reason. On the other hand, he was praising him for being the most consistent of all philosophers (Jacobi, 1785: 27-29), and even defended him against those -- Mendelssohn included -- who, as he thought, detracted from his philosophy by their misguided attempts at 'saving'it from the consequences of its unmitigated rationalism. And there was another reason as well for Jacobi's otherwise puzzling predilection for Spinoza -- the Benedictus, as he once calls him, the 'Blessed One'. (Jacobi, 1799: 41) Spinoza knew that truth is its own criterion; that ultimately, therefore, it is not amenable to discursive reason but must be apprehended immediately on its own, intuitively. This was also Jacobi's position. His objection against the rationalism of the philosophers boiled down to precisely this -- that it was the product of a reason that had lost touch with its intuitive sources, and was therefore given to mistaking its own productions for the real thing. On this score, Jacobi thought that he had found a strange ally in the most rationalist of all philosophers.

Yet there was a serious problem in the Spinoza Letters. Jacobi never quite gave to understand what *he*,as contrasted with Spinoza, meant by 'intuition'. He says that philosophy must be 'historical'-- presumably, that it must take its starting point from the records of human actions and human events; that it must never abandon experience, or common sense. These, though vague and hardly the basis for a well defined position, were in themselves perfectly acceptable claims. But Jacobi then obscured them with his pious perorations at the end -- by citing, among others, the theologian Jerusalem, who thought that the task of philosophy is simply to elucidate the content of revealed faith, and Lavater, who believed in the ever present witness of miracles. He could easily give the impression, as he in fact did, that, by attacking philosophy and the philosophers, he was advocating a return to the blind acceptance of Biblical revelation; that, by 'intuition', he meant religiously inspired 'faith'. Lessing had suspected that much at the end of his conversation with Jacobi, according to the latter's own report. After the publication of the Spinoza Letters, his friends and admirers had no doubt about it. Led by Nicolai in Berlin, they undertook their campaign against Jacobi, branding him with obscurantism and religious enthusiasm -- these, in the eyes of the *Aufklärer*, the greatest of all possible sins.

2.2 The 'David Hume'(1st edition, 1787)

The dialogue that Jacobi published in 1787 carried the full title of *David Hume on Faith, or Idealism and Realism, A Dialogue*. It had originally been intended as three separate dialogues, as the structure of the

final product still betrays. It clearly falls into three parts, with an autobiographical interlude connecting the first and second which is the source of much of our knowledge about Jacobi's early philosophical education. Jacobi uses it as a vehicle for documenting how much, from the beginning, he had been temperamentally driven always to give priority to existence, as immediately apprehended, over any conceptual representation of it, and also for putting on record that, when still a young man, he had found this irresistible tendency of his best satisfied in Kant's two early essays, one on evidence in metaphysics and the other on the proof of God's existence. (Jacobi, 1787: 67ff) These autobiographical remarks come at the heel of a defence, in the first part of the Dialogue, against the accusation of irrationalism brought against him chiefly because of his use of the term 'faith'(*Glaube*, in German) in the Spinoza Letters. Jacobi's many polemical divagations apart, the main line of the defence is very simple. Hume also had used the term 'belief'(also translated in German as *Glaube)*for the kind of assent which is based on experience and is expressed in the form of judgement. Such an assent is immediate, a matter of feeling rather than ratiocination, and all the more unimpeachable precisely because reason alone would never induce it on its own. Now, Jacobi argues, in the Spinoza Letters he had used the term 'faith'(also *Glaube*in German) in the same sense, in order to claim that judgement of existence must be immediate. No process of ratiocination could produce the certainty that accompanies it. There was no 'irrationalism', therefore, either intended or implied. On the contrary, Jacobi had been forced to use the term, and to oppose it to reason, only because the philosophers had pre-empted the latter term, and had unduly restricted it to mean the kind of discursive conceptualization that abstracts from real things and is ultimately irrelevant to judgments of existence. But nobody had ever accused Hume of irrationalism. Why was the charge, then, being laid against him? (Jacobi, 1787: 29ff)

Jacobi's critics, in this case Hamann included, were quick to point out that the English 'belief'does not have quite the same meaning as the German *Glaube;*that the latter carries religious connotations best expressed in English by 'faith', even though there is only one German word *(Glaube)*to translate both 'belief'and 'faith'. Jacobi's point, however, was that he, as a matter of fact, had used the term to mean 'belief', in Hume's sense. Whether he was being disingenuous in pressing it, granted the profuse use of religious rhetoric with which he had padded his earlier book, is of course a good question. None the less, once Jacobi had thus appealed to Hume's authority for his defence, he was faced by the task of having to distance himself from the scepticism which was entailed in Hume's use of belief but did not in any way fall within his own purview. Jacobi thought of himself as a realist. (Jacobi, 1787: vi) The second part of the Dialogue is concerned precisely with this issue. Inasmuch as Jacobi had a well formulated theory of knowledge at this stage of his development, it is to be found there.

Distilled from the many twists and turns of dialogue style argumentation, the theory can be summed up as follows:

> (1) The starting point is the denial, because contrary to fact, of the fundamental assumption of classical empiricism -- namely, that experience begins with purely subjective representations, and that belief in external objects is arrived at only by way of an inference based on the passivity of some of these representations. This assumption inevitably leads to Hume's scepticism. Jacobi rejects it off-hand on the ground that, as a matter of fact, a subject could not be aware of himself -- aware also, therefore, of the alleged subjectivity of

some of his representations -- without defining his 'self'in opposition to some admittedly external object, i.e. without immediately referring his representations to something other than himself. The very possibility of subjectivity entails the possibility of objectivity. Jacobi's classical formula for this position is that there is no 'I'without a 'Thou'. (Jacobi, 1787: 63-65)

(2) Hume had of course denied that we have consciousness of any definite 'self'. Such a denial, however, is only possible according to Jacobi if one assumes a purely theoretical standpoint with respect to the 'self'. We act, and we become aware of ourselves precisely in action. (Jacobi, 1787: 102ff) Self-awareness originates in a subject's feeling of power that accompanies action. Jacobi cites Hume himself in support this claim, even though he then uses the claim precisely in order to overcome Hume's theoretical scepticism. The alleged feeling of power immediately implicates the presence to the subject of an external something that exists in itself and interferes with the felt power, but, in so doing, also provides a reality check for it.

(3) Representation is called into play as the reflective attempt on the part of the subject to sort out the differences between his own self and the external thing resisting his power. This is a formula that brought together in an original unity the three components of consciousness, namely feeling, sense representation, and reflective conceptualization, that Hume as well as Kant had instead sought to synthesize externally. Through a series of arguments, Jacobi then shows how it is possible to arrive at all the categories that Kant had proposed *a priori*by descriptively identifying, rather, the basic conditions that define the distance between the self and its 'other'. (Jacobi, 1787:111-121)

(4) It follows that 'reason'is not a faculty that supervenes to the 'senses'*a priori*, but a more refined form of otherwise fundamentally sensible representations. As Jacobi puts it, the greater the sensitivity of a subject, the greater his rationality also. The two, 'senses'and 'reason', are inextricably bound together. (Jacobi, 1787: 125-34)

In making this last point, Jacobi paraphrases Thomas Reid, without however mentioning him explicitly in this immediate context.[7]He apparently thought that he was deploying against Hume the same line of argument already advanced against him by his Scottish 'common sense'critic. He was also thereby clarifying what he meant by 'intuition'-- that supposed immediate apprehension of truth that gives rise to belief. It now appears that the intuition in question is a product of the 'senses'. It is not, however, anything blind, for -- as understood by Jacobi in the spirit of Reid, and certainly in contrast to both Hume's and Kant's notion of sensibility -- 'sensations'are not passive impressions of the mind but exhibit rather from the start a complex relation between subject and object that can then be developed into even more reflective (i.e. representational) forms. Jacobi actually says that he is dependent on Spinoza for the seminal idea of his present method of deriving reflective representations from the senses. (Jacobi, 1787: 120, note 25) More significant, however, than any reference to Hume, Reid, or Spinoza, is that he also refers to Kant and his transcendental method deliberately, thus betraying that *he*(Kant), rather than any of them, had been his main concern all along. (Jacobi, 1787: 119-20, &120, note 25) Jacobi had been

advancing a theory of knowledge that, in his opinion, could compete with Kant's. For it explained the basic facts of experience, and even explained the factual necessity of certain fundamental concepts arising in the course of it, without incurring the kind of formalism that, in his opinion and many of Kant's contemporary critics, affected Kant's own transcendental method. Jacobi's first representations were allegedly drawn from experience directly. They were not applied to it *a priori*, in the manner of Kant's categories which Jacobi regarded rather as nothing more than ‘prejudices of the intellect’.

But there was a difficulty inherent in Jacobi's proposed theory. It came out most poignantly in the third part of the Dialogue -- a part dedicated, for the most part, to an exposition of Leibniz's metaphysics. (Jacobi, 1787: 144ff) Here Jacobi portrays the German Leibniz as the champion of individuality, and also tries to show how it is possible to accept his Monadology if duly modified. The transition in the Dialogue between second and third part is performed rhetorically. There was, however, a conceptual basis for it. And it was in this, in the obvious logical connection between Jacobi's just suggested theory of experience and Leibniz's Monadology, that the difficulty lay. For the organic conception of rationality that that theory implied -- reason being nothing more than a more developed, more reflective, form of sensibility -- fitted well indeed with Leibniz's system. But it also clearly led to Leibnizian naturalism. And the idea that the ‘self’cannot be defined except in terms of an ‘other’, while also well fitted to the Monadology, also led to the equally Leibnizian position that everything in the universe is itself by reflecting everything else. Jacobi had however made his philosophical debut precisely by combating the assumption that everything can be explained by reference to everything else -- a position that he thought reflected the philosophers's logical enthusiasm for explanation and which he opposed because it ended up undermining human subjectivity.

Jacobi expressed hesitations about the Dialogue from the beginning. He eventually criticized it openly, attributing what he later thought was the confusion marring it to his lack, at the time of composition, of a clear conception of reason. Whether his later conception was to be any better, or whether the original text was truly confused, is of course a question open to discussion. At any rate, the presentment of an unholy alliance against Kant that Jacobi was unwittingly forging with schlastic metaphysics -- his otherwise natural foe -- is perhaps the reason for the abrupt, even puzzling, conclusion that he gave the Dialogue. Jacobi ends it with an unexpected renewed effusion of the same pious rhetoric we find at the conclusion of the Spinoza Letters. Indications are given that one can have direct experiential evidence of things supernatural. (Jacobi, 1787: 195ff) Indeed, on the naturalistic conception of reason earlier developed in the Dialogue, any evidence of the Divine would have to be obtained directly, on the evidence of miracles, as Lavater believed. This was perhaps a position dear to Jacobi's own heart, but it embarrassingly made him vulnerable again -- this is the puzzling part about this production of Jacobi -- to the charge of obscurantism that had been the purpose of the Dialogue to fend off in the first place.

2.3 The Spinoza Letters (2ndedition, 1789)

The much augmented second edition of Jacobi's first major work included a fuller documentation of the correspondence with Mendelssohn, and eight new supplements in which, among other things, Jacobi offered a more detailed exposition of Spinoza's philosophy (Supplements VI, VII) and a lengthy criticism

of Herder's version of Spinozism (Supplement IV). It was also prefaced with fifty-two propositions -- twenty-three purporting to defend the thesis 'Man does not have Freedom'; the remaining twenty-nine, the opposite thesis 'Man has Freedom'. The argument in defence of the first thesis is based on the assumption that human affairs are organized mechanistically, i.e. according to a concatenation of purely efficient causes that would make the being of each totally dependent on the external being of everything else. In defending the thesis Jacobi follows Spinoza closely. The mechanism at issue applies to the cognitive side of man just as much as to his bodily side. Consciousness, or representational being, is only a mirror of extended being. Whatever happens in the latter is repeated on the side of thought, though according to thought's own modality of existence. Thus, according to Jacobi, syllogistic thinking is mechanistic in nature, being driven by principles *ab ante*no less than any corporeal sequence of events. So far as the opposing thesis is concerned, Jacobi argues for it by retrieving a theme from the *David Hume*.If it is the case that a finite 'I'acquires identity only when confronted by an equally finite 'Thou', then, just as for the required interplay between the two one must assume on each side a source of 'passivity'that allows each to be limited by the other, so one must also equally assume in them a countervailing source of 'activity'. This source is irreducible. It follows that 'activity'and 'passivity'must be assumed to run across the world we know in every part of it. In some parts, the one element might well display a greater degree of intensity than the other; nowhere, however, can either element crowd out the other totally. It also follows that, though mechanism is both possible and necessary, a totally mechanistic organization of things that would *eo ipso*allow no room for individual freedom, i.e. one that would reduce all things to external relations, is an impossibility, for it would in fact reduce things to mere nothingness. (Jacobi, 1789: xxxv-xxxvi) It would make them totally passive. 'Passivity', however, has no meaning except in relation to 'activity'. The conceptual basis for denying the reality of individually attributable human act is thus removed.

The two theses, appearing as they did one year after the publication of Kant's *Critique of Practical Reason*, may be taken as a tacit criticism of Kant's strategy of defining the problem of freedom in the form of an antinomy. Jacobi's two theses do not constitute an antinomy because, from their perspective, any apparent opposition between mechanistic explanation and moral demands is already resolved in the propositions detailing the second thesis. Presupposed by both sets of propositions is the assumption, which Jacobi accepts, that beings (at least, created beings) exist in limiting relationships, and that such relationships entail an irreducible element of passivity as well as activity on both the sides entering into them. The difference between the two sets of propositions is that the second (the one defending freedom) respects this necessary condition of any relationship, whereas the first does not. Only the philosopher, because of his passion for abstractions, is tempted to conceive, on the one hand, a purely mechanistic relation, one which in fact negates activity altogether and thus removes the basis for the desired relationships; on the other hand, one which is purely active, exclusively spontaneous, and thus attains the same result by the opposite route. For this last point, Jacobi could have easily connected Kant's notion of freedom as purely spontaneous activity with Spinoza's *causa sui*. Only then, when one has thus given way to the philosopher's passion for abstractions, is one faced with the Kantian kind of antinomy -- a conflict of opposing yet conceptually valid views that can be controlled, though never truly resolved, only by appeal to the unknown.

In other words, the choice with which Jacobi confronted his readers at the opening of his new edition of

the Spinoza Letters was not between two conceptually valid yet contradictory claims, but between the anti-humanism of the philosophers -- based at it was on an abstraction -- and a humanism that on the contrary stays close to the facts of experience. In this respect, Jacobi was reaffirming the point he had already made to Lessing ten years earlier, namely, that the only way to deal with the irrational results of philosophizing is to know when to stop to philosophize. Ironically, however, he had used the same general paradigm of two interacting forces to define, in his two sets of propositions, both the system of mechanical causes and the system of freedom. This was a point that did not go lost. Fichte was soon to use the paradigm as his basis for resolving Kant's antinomy, by arguing that freedom requires as its counterpart the kind of mechanistic organization that scientific explanation demands. Where Jacobi had tried to circumvent the antinomy altogether by suspending philosophical reflection at the right moment, Fichte was to try to resolve it by prompting that same reflection to a higher level of abstraction instead. So far as Jacobi was concerned, the results of any such move could only be disastrous. And it was indeed against Fichte that, ten years later, he felt compelled to repeat in stronger terms than ever before his original interdiction against philosophers.

2.4 The Open Letter to Fichte (1799)

Many were the circumstances that precipitated the charge of atheism against Fichte and eventually led to his departure from Jena, not all of them of a philosophical nature. His early philosophical system did nothing, however, to help him.[8]In brief, Fichte's position was that the resistance that nature *de facto*poses to the exercise of human freedom should be interpreted as a product of freedom itself, inasmuch as the human self, by raising itself through abstractive reflection to the status of a pure 'I', thereby introduces new possibilities of being, in the face of which given nature necessarily assumes the appearance of a purely contingent, foreign, quantity. It has no meaning in itself except as a material on which the self can exercise its freedom by shaping it according to its intentions. The freer the self, the more insubstantial nature will appear to it, and the greater the possibility of rationalizing it through external means. Even if nature were not in itself a mechanical organization of mere appearances, it would have to be assumed by the free self to be such for the sake of freedom. This is of course an oversimplification. It conveys none the less an aspect of Fichte's early thought that could easily have given the impression that his system constituted the ideological justification for the kind of social engineering that was being done across the Rhine, in revolutionary France, in the name of Freedom. Many thought, at the time, that the *philosophes*had been the ones to hatch the revolution in the first place. And Fichte now seemed to confirm their worse suspicions. This fear of the revolutionary potential of his thought -- a fear, it must be said, not altogether ungrounded -- is what gave the debate surrounding his suspected atheism its special emotional tone.

In part it also explains the stridency of Jacobi's public reaction to Fichte -- even though privately, as person to person, Jacobi was concerned about the man's safety, and even considered possible venues of refuge for him. He declared Fichte as the true disciple of Kant -- the one who had brought the premises of Transcendental Idealism to their logical conclusions. He also revised his prior estimate of Spinoza. He had always portrayed the latter as the most consistent of all philosophers. He now recognized that the attribute belonged in truth to Fichte. *He*was in fact the philosopher *par excellance*-- the Messias of

Reason, as Jacobi now loudly proclaims him. (Jacobi, 1799: 2) For Fichte had extended the reach of philosophical abstraction, with which Spinoza had gone only so far as to exclude the possibility of subjectivity, to retrieve this subjectivity itself within its compass, as if it could be excogitated *a priori*out of pure thought. Fichte's idealism was a case of inverted Spinozism, of spiritual materialism. (Jacobi, 1799: 2-5) For Spinoza's substance, which explains all things by explaining none of them in particular, hence relegates the determination of individuals to the endless mechanism of external causes, was now being reintroduced as the product of pure reflection, as its first objectification. It was as if reflection dissolved all things in the ether of pure thought, out of sheer freedom, and then reassembled them again -- but now as a game, according to self-concocted rules. (Jacobi, 1799: 24-27)

All this infuriated Jacobi. No piece of his is as overflowing with religious language as this letter to Fichte. But the rhetoric exhibited here has none of the pious triumphalism, smacking of Lavater, that had vitiated his earlier productions. It seems born, rather, out of genuine fear in the face of what he thought was the ultimate nihilism of reason. Jacobi conceded that there was no argument against Fichte, no way of refuting him on his own conceptual grounds. He stood before him, therefore, as one who gave witness against the philosophers, forcing the audience to a choice between him and them. Against the universal norms of either Kant's or Fichte's morality, he stood rather as the champion of the exception -- the one who, for the sake of the individual, would dare profane the sacred altars, indeed to lie and even murder. (Jacobi, 1799: 32-33) This was a powerful individualistic manifesto that Jacobi was promoting, the kind not be heard again until Kierkegaard more than half a century later. But it had nothing to do with the asocial individualism typical of the *Sturm und Drang*movement that had been the cause of Jacobi's concern in his younger years. On the contrary, even before this commotion with Fichte, and probably with Goethe in mind, Jacobi had long been meditating on the place of the individual in society, and had formulated for himself a definite social theory. To gather an idea of it, however, we must turn to his more literary works, as we shall do in due place. (Section 3)

2.5 After 1800

The final period of Jacobi's life saw, among other things, two major publications and the undertaking of the editions of his collected work. We shall consider this last at the end, in the section 'Retrospect'(Section 5). As for the other two pieces, the first was a renewed and ever sharper attack on Kant's Transcendental Idealism, as the title indicated: 'On Critique's Attempt to Reduce Reason to the Understanding'. It was completed by Jacobi's disciple Köppen, as clearly indicated in the text itself; and published in 1802. Reduced to its bare minimum, the attack was based on three arguments, as follows:

(1) Kant had tried to establish the possibility of synthetic *a priori*judgements on the assumption of an object (= x) and of subject (= x) that we *ex hypothesi*cannot ever know but upon which the system of experience based on the said judgements none the less depends. The two (viz., the subject and object) constitute transcendental conditions of the possibility of the system of experience. Kant's attempt, however, entails an irresoluble difficulty. For although the assumed subject and object are factors that *ex hypothesi*fall *outside*the system of the *a priori*judgements to be validated (i.e., they are 'transcendental'), they can only be defined in terms of the concepts and distinctions that have meaning exclusively

*inside*that system. Kant thus finds himself in the embarrassing situation of negating, while at the same time affirming, the transcendality of the conditions of his system, thereby undermining the stability that the latter should derive from them. (Jacobi, 1812-25: vol. 2, 85-91)

(2) The same objection can be generalized in terms of the relation that obtains in Kant's Critique between 'reason'and 'understanding'. It is the task of reason, according to Kant, to define the extra-systemic conditions that make the system of experience possible. But, *ex hypothesi*, reason has no knowledge of these conditions. As defined by it, they are empty conceptual constructs that acquire meaning only to the extent that they are used by the understanding in its endless work of systematizing experience. To this extent, reason is totally subordinated to the understanding. Yet the latter *needs*reason. To the extent, therefore, that reason does satisfy this need with its ideas, it cannot help creating the illusion that, through them, it yields genuine knowledge -- that, logically, its conceptions are prior to those of the understanding. Now, Jacobi objected to what he took to be the existentially impossible requirement thus being imposed on reason, namely, that it should generate illusions for the sake of the understanding, yet be fully aware that such illusions are just *that*. (Jacobi, 1812-25: vol. 2, 81, 100-01, 115)

(3) The third objection, though repeated in a variety of ways, consists essentially in a criticism of Kant's analysis of the function of the understanding in experience. Even granted all the elements of Kant's transcendental psychology, notably, that sensations are *per se*amorphous mental impressions, that space and time can be at once intuitions and objects of intuition, that the categories are *a priori*concepts of the understanding -- assumptions these, all of which Jacobi in fact rejects, and against which he strenuously argues in the body of the essay -- even granted these, there still is an internal defect in Kant's system. The problem is that the two terms to be synthesized in any presumed *a priori*judgement of experience, namely, thought and sensations, are on their own too indeterminate to provide any clue as to how the object that should result from the synthesis would look like. So far as sensibility is concerned, such indetermination applies whether we take it according to its material content or its *a priori*forms of intuition, for 'space'and 'time', as defined by Kant, lack *per se*definite limits. (Jacobi, 1812-25: vol. 2, 77-79, 122ff, 134-35, 136-39) It follows that the determination, in actual experience, must be provided by the intervention of the imagination, and whether we take the latter in a psychological or transcendental sense, its contribution will necessarily entail an element of arbitrariness, and it will thus render any synthesis thereby attained vulnerable to sceptical doubt. Kant, in other words, fails to deliver on his promise of a knowledge, limited indeed to experience, but necessary within such limits. (Jacobi, 1812-25: vol. 2, 95, 97, 115ff, 118ff, 135, 150, 154ff)

The other major work of Jacobi in this last period,*Of Divine Things and Their Revelation*, was published in 1811. Jacobi had actually begun writing the piece ten years earlier, as the review of a recently published volume by a pietist writer, but had later picked it up again and completed in a totally different cast as a critique of Schelling's 'System of Identity'and of the new Philosophy of Nature connected with it. Evidence of its earlier purpose is still visible in the final text, beginning as it does with the discussion of a moral issue, namely, whether morality is to be based on virtue alone, or on natural happiness, or a combination of both. But then the text suddenly shifts to a long discussion of Kant's critical philosophy, the burden of which is however to demonstrate that Schelling's recent idealistic productions were but the logical consequences of precisely that philosophy. The treatment of Kant in this context is a strangely

qualified one. On the one hand, Jacobi praises him highly, showing him all the respect due to someone who had become an icon of German philosophical culture. He praises him because he had unequivocally recognized that a true God (presumably, such as morality demands) must be a person; also because he acknowledged that we have an immediate belief in God, freedom, and immortality, and 'belief'for Jacobi of course meant 'knowledge'; finally, because, just like Jacobi himself, he had tried to respect and harmonize the interests of reason, which are directed to God, and those of the understanding, which are directed to the senses. (Jacobi, 1812-25: vol. 3, 341-44; 351-52) Such praises, however, turn out to be hollow -- even a masked form of condemnation. For Jacobi's main thesis throughout the otherwise rambling work is that Kant, while upholding personalistic values, had in fact subverted them conceptually, the net product being Fichte's system of ideal materialism and now the openly naturalistic system of Schelling.

The kernel of Jacobi's argument was in fact still the same as the one presented in his earlier Kant essay (1802). The point there was that Kantian reason, though transcending the understanding, was none the less restricted in its knowledge to the limits of the latter and even subservient to it in the exercise of its powers. As a result, reason found itself in the strange situation of spontaneously generating belief that we have knowledge of transcendent realities, while at the same time having to recognize that any such knowledge is illusionary. Jacobi now further develops this point by arguing that Kant's critical system was explicitly intended from the start to serve the interests of science -- where by 'science'Kant meant first and foremost a cognition of the understanding directed at the objects of the senses. The understanding, however, despite its earth-bound nature, cannot help falling under the influence of the higher faculty of reason. It is caught up in reason's desire to behold all things under a transcendent single principle. And reason -- as a matter of fact, according to Jacobi -- does have an intuitive knowledge of things transcendent. But since the understanding cannot allow this knowledge, yet still needs reason as a higher regulative principle, it substitutes for reason's knowledge ideal constructs that are indeed also the products of reason, but are made up, however, of abstractions derived from the understanding. These abstractions then hide from the understanding and reason itself the transcendent knowledge that the latter enjoys *de facto*. For since reason naturally begins with the assumption of such a knowledge, yet cannot recognize it in the abstractions construed for the sake of the understanding, it ends up treating it as mere illusion. (Jacobi, 1812-25: vol. 3, 372-94)

Fichte's next obvious step was to remove the ground of the illusion by declaring the ideal constructs of reason to be themselves the objects of the transcendent knowledge that reason desires -- in effect, of identifying God with the logical order of things. And Fichte, according to Jacobi, quite consistently also took as the test of true personality the capacity on the part of a self to reconstruct the whole of nature*a priori*from the principles of that logic, thus reenacting ideally God's creative deed. In fact, however, such a logic was from the start the product of understanding's abstractions. It has no content of its own. The idealizing activities of a self cannot in fact be realized, therefore, except by attending to the material details of nature. And since, even when idealized, nature still remains 'nature', it follows according to Jacobi that in practice Fichte's idealism is but a form of materialism. Schelling's naturalism only made this fact explicit. In sum, the great hoax that according to Jacobi idealism had perpetrated -- starting with Kant and concluding with Schelling -- was to give to believe, not indeed that the human spirit can dispense with Things Divine, but that these things can be saved and attained in the impersonal medium of

nature.

Spinoza figures prominently in this last major work of Jacobi, just as much as he had in his first. Jacobi actually turns to him in order to clinch his own overall argument historically, and also, since Schelling now had no compunction about declaring the affinity of his system with Spinoza's, in order to vindicate the validity of his original claim -- the one with which he had made his philosophical public debut -- that philosophy equals Spinozism, and Spinozism equals atheism. Post-Kantian developments, in his opinion, had proved him right. His special concern in this late work was to establish the materialism of Spinoza. According to him, the great philosophical move with which which Spinoza founded his system was to distinguish extended substance from thinking substance while at the same time asserting the identity of the two in the one substance which is God. For Spinoza, therefore, extended substance was all that there is of objective being ('formal being', according to his terminology), i.e. of substantial and efficient being. Thinking substance is only the representation of it, with no content *qua* thought except that of substance, i.e. extended substance. It followed that, according to Spinoza, in order for thought to think God, or to think thought itself, it has to think extended substance. There is absolutely no other object available to it. And this, according to Jacobi, constituted an extreme form of materialism. Malbranche, Leibniz, and Berkeley -- again, according to Jacobi -- had recognized this materialistic implication of Spinoza's thought, and, in order to counteract it, they had all taken the obvious step of declaring 'extension'itself to be just an appearance. But then Kant came along, and he took the further step of declaring the 'self'itself an appearance, a series of representations with no content of their own and with no intrinsic substantial unity -- hence dependent for both on their phenomenal, spatially extended, objects. Through a circuitous route, through the language of subjectivity and appearance, Kant had thus reached the same materialistic position as Spinoza's. Jacobi had therefore been right when he had suspected from the beginning that Kantianism was but a form of inverted Spinozism. Fichte and Schelling had only worked out the implications of Kant's position, thereby rendering its Spinozist essence all too obvious. (Jacobi, 1812-25: vol. 3, 345-56, 431-32)

Important to note is the new terminology that Jacobi was using in this last work, and had already begun to use in the previous essay on Kant. He was now operating on the assumption that there is a distinction between 'reason'and 'understanding'-- where by 'reason'he meant the capacity for transcendent, unifying, principles; by 'understanding', the faculty of abstracting general representations from more particular ones. (Jacobi, 1812-25: vol. 3, 395ff, 400, 434-35) The distinction was, of course, Kantian in origin. The claim, moreover, that Kant had held 'reason'hostage to the abstractive requirements of the 'understanding'was one widely debated among his idealistic followers. Jacobi, however, was now using 'reason'to mean a faculty of intuitive knowledge of transcendent things. He was giving to it a positive noetic value. As used by him, therefore, the distinction between 'reason'and 'understanding'allowed him to draw what he thought was the crucial difference between himself and Kant. (Jacobi, 1812-25: vol. 3, 369-72) Whereas Kant had tried to harmonize the two faculties negatively, simply by keeping each within its appointed limits, Jacobi had revealed their harmony positively, by bringing to light the knowledge that reason actually contributes to experience, thus providing the positive matrix within which the understanding can operate.

In another way as well the distinction helped Jacobi -- or so he thought it did. From the beginning Jacobi

had been faced by the difficulty that, in order to oppose philosophical reflection and the nihilism that it brought in train, he had had to attack ‘reason’and to summon ‘faith’in its stead. He had thereby made himself vulnerable to the charge of ‘irrationalism’and ‘blind fideism’, a charge that he always strenuously rejected. So far as Jacobi was concerned, the true irrationalists were the philosophers. With the newly adopted distinction, Jacobi could now specifically direct to the ‘understanding’, or more accurately, to a ‘reason bound to the understanding’, the kind of criticism earlier directed quite generally to ‘reason’. He could then summon ‘reason’, with the new meaning attributed to it, to do the job previously performed by ‘faith’. In this way, he could more easily dodge the charge of ‘irrationalism’or ‘fideism’, for it was for the sake of ‘reason’-- that is, ‘true reason’-- that he could now claim he was waging his campaign against the philosophers. The problem, of course, was that Jacobi never quite clarified this new meaning of ‘reason’. He was apparently using the concept in the same way as he had previously used ‘faith’, namely, as denoting an intuitive faculty akin to feeling but at the same time sporting the clarity of a concept. And this was a combination of notes just as obscure when it went under the name of ‘faith’as it did when called ‘reason’. Whether the new distinction, and the new language that it made possible, in fact ever clarified Jacobi’s position is a question, therefore, open to debate.

3. Literary Works

Jacobi’s earliest publications, some done in collaboration with his brother Georg, were of a purely literary nature. And Jacobi continued to publish the occasional piece in this vein also later, when engrossed in philosophical debate. His reputation as a literary figure is mostly based, however, on his two novels, *Edward Allwill’s Collection of Letters*, and *Woldemar*. Both productions underwent a long process of development, and were published in different forms at different times. (David, 1913; Jacobi, 1957) The central character of the first, Edward Allwill, very likely underwent radical changes in the course of formation. We are only interested here with the substantially definitive form that Jacobi gave to them in the editions of, respectively, 1792 and 1794 (the latter as reproduced in 1820). Both, but the *Woldemar*especially, came in for devastating criticisms upon publication.[9]But they both also enjoyed a certain popularity. To call them ‘novels’in any modern sense of the word would be misleading. Both lack in serious dramatic action, and the characters, in each case, lack credibility. They are one dimensional figures, mouthpieces of Jacobi’s moral and social ideas rather than fully bodied personalities. In the case of the *Woldemar*especially, Jacobi seems to be talking about his characters rather than letting them act on their own. (David, 1913: 92 &notes) As narrations, the two pieces are in fact philosophical arguments thinly disguised by a veneer of imagination. Yet this feature, which might count as a fatal flaw from a purely artistic point of view, is precisely what makes them an indispensable part of Jacobi’s philosophical output.

As literary productions, the two pieces belong to the popular Enlightenment genre of *Erziehungsroman*, ‘education-novel’. In both cases the education at issue, as one would expect from the genre, is that of a *Herzensmensch*. This is the kind of human being who presumes to live by natural feelings alone but who, when confronted by actual social situations, learns (or fails to learn, to his ultimate destruction) that such feelings are insufficient for coping with the complexities of real life. Nature must rather be subjected to social discipline. Allwill and Woldemar, the main characters in Jacobi’s two novels, are both this type of

Mensch. Both behave as young men, though the biological age of Allwill is an object of discussion among the other characters of the novel, and he might well be taken to be young by vocation only.[10]Both seem to have come out of nowhere but to be at the moment deeply involved in the affairs of a family circle. Both are temperamentally gifted with strong feelings yet also given to highly abstract reflections -- both, sensitive products of nature yet refined products of thought at the same time. And it is as such juxtapositions of extremes that they play an influential and at times also disturbing role in the societies within which each has been admitted as an eccentric but also interesting permanent visitor.

The education that each of these ‘young men’undergoes (or should undergo) is shaped within the framework of these societies. From what we learn about Allwill, since early age, out of sheer natural impetuosity, he was given to some most extraordinary behaviour -- the kind that would have appeared to an outside observer as indicative, at times, of a noble and compassionate frame of mind; at other times, of sheer foolishness. So far as Allwill himself was concerned, however, such distinctions would have seemed immaterial, even meaningless. Allwill (whose name in German, as in English, means just *that*, ‘all will’) did whatever he did out of sheer impulse -- his only norm of action, the action itself. And now that he has grown older, he displays this character trait in the arena of speculative dispute as well. For the impetuous young man has also become a skilled dialectician, as we learn from a scene in which we see him deliver a long disquisition on the nature of knowledge in the course of a social gathering. (Jacobi, 1792: 143ff) Anyone familiar with Jacobi’s position on the subject can easily recognize that Allwill’s expressed views are also Jacobi’s. But Allwill apparently defends them only as an exercise in intellectual virtuosity, simply for the sake of winning a point. In a completely different mood, as he displays his skill at the clavier moments after his disquisition, we see him flippantly rejecting them all, in sharp contrast to the eagerness with which he has just defended them. (Jacobi, 1792: 156ff) Allwill thus moves in his social circle as an insubstantial being, one who appears now under one form, now under another, all the time giving the impression of an inner core that he in fact does not possess. Precisely for this reason, by holding out possibilities that are all the more interesting just because they are indefinite -- the illusionary products of a non-self -- he exercises in the eyes of some of the women in the circle the fascination of a seducer.

Allwill, who might have been taken as a beautiful product of nature in his first literary appearances, became an ever darker character as he developed at his creator’s hand. This is especially true in the 1792 edition of Jacobi’s novel, because of the singularly ominous shadow retrospectively cast upon his figure by a letter addressed to a certain ‘Erhard O**’, and carrying Jacobi’s own signature, added to it at the end. (Jacobi, 1799: 13) There are dark references in the letter to the demonic nature of this Erhard, and even more disturbing hints at the connection between his behaviour and the social destruction being wrought by the French revolution. It is significant that in the body of the novel itself, the one character who has penetrated through to the true nature of Allwill and recognizes the destructiveness of his character is a woman who, like him, does not belong to the family circle that provides the social context of the narration but, like Allwill again, preys upon the emotions of its members. She is a widow (Sylli is her name) who has lost her only child besides her husband, and now, left to her own, is tempted to despair. As a woman she stands according to the convention of the time for mother earth, and thus reveals another aspect of nature that might be lost in the brilliant figure of Allwill -- the one who would be the innocent first child of it. By itself, nature has no determinations. It is a shapeless, dark power, that

absorbs its products just as much as it gives birth to them. Sylli, if not saved by the intervention of her friends, is on the verge of collapsing into an inner maelstrom of painful feelings and distracted passions. Among all the women of the family circle, she understands and fears Allwill because, in her more elemental ways, she is just as seductive and destructive as he. Together, the two convey the message that Jacobi intends from the beginning of the narration. The cult of nature is just as misguided as the cult of reason; each is the reflection of the other. For nature as well as reason, when taken in abstraction from the social relations such as we see realized among the inner members the family circle in the narration, have no structure, no centre of gravity, and therefore only lead to a dispersion of existence. Those family members might not display the emotional power of a Sylli, or the brilliance of an Allwill. They actually look rather dull, and they certainly lack the intellectual acumen of Allwill. Yet, it is in the virtues they have developed by living in communion that Jacobi pins his hopes for a true humanity.

True rationality is social rationality. This is Jacobi's lesson in *Allwill*, and it is repeated in the *Woldemar*--though the central character there, Woldemar, while exhibiting the same propensity to abstract feeling and abstract thought as Allwill, is treated by Jacobi with much warmth. The character might well have been intended by him as a self-portrait. The novel itself, in its final form, was put together from two pieces originally published independently. (Jacobi, 1779 (3) (4)) The synthesis never quite worked. The result is that the narration is both convoluted and disjointed -- the overall theme, inasmuch as there is one, never transparent. There is, however, a sort of clarification at the end, in the emotional resolution of an intricate social situation that has been brewing from the beginning. Woldemar, unrealistic man of feelings that he is, has construed his whole world around the friendship he has developed with one of the women in the circle of friends he frequents. He has invested his whole existence in that friendship. And the woman has colluded with him. She has however secretly made an oath to her dying father, who has always distrusted Woldemar, never to marry him. And this secret, even though neither she nor Woldemar have ever considered marriage, none the less stands in the way of complete transparency in her relationship with Woldemar -- a circumstance for which she feels guilty. Woldemar learns about the secret by accident. His whole existence, built as it had been on the assumption of total reciprocity with the woman, is shattered. He totters on the verge of insanity. The woman, for her part, is distraught by the strange behaviour of Woldemar. But fortunately she learns about Woldemar's knowledge of her secret, and Woldemar in turn finds out about her knowledge of his knowledge. And this newly shared wisdom provides the basis for a renewed and healthier relationship between the two. Both acknowledge their guilt -- Woldemar, for having expected the kind of total transparency on the part of his woman friend that would have dissolved her individuality; the woman, for having used her individuality as an occasion for deception. The two are now ready to begin sharing their existence in friendship, fully aware of the dependence of each upon the other, yet also of the limits that such a sharing must respect.[11](Jacobi, 1812-25: vol. 5, 457ff)

The story might sound improbable to the modern ear, and Jacobi's lack of artistry does nothing to help it. Yet, despite its many flaws, the ethical lesson it was supposed to convey is interesting as well as significant. Jacobi was actually harking back to the notion of reason he had adumbrated even as early as the first edition of the Spinoza Letters. Rationality is constituted in the relationship between an &145;I'and a 'Thou', a relationship that respects the conditions of both its terms. Within the context of the *Woldemar*, this conclusion also brings resolution to a question discussed in a dialogue (the

*Waldgespräch)*introduced in the first part of the narration -- the question, namely, of the extent to which virtue depends on either nature or art. The answer is that it depends on both, provided that each is tested by the limits of actual human relations. At the conclusion of the novel, just as in the earlier dialogue, it is also clearly implied that the mutual respect that such relations demand from the individuals entering into them is possible only on the assumption of a 'Thou'greater than any human 'I'-- a 'Thou'whose transcendent pull forces the human 'I'outside its otherwise purely natural limitations.[12]Jacobi's last word, in his novels just as much as in his philosophical productions, always belongs to God.

4. Polemical Works

All of Jacobi's publications are polemical in nature. Some, however, were occasional pieces written with a particular event or situation in mind. Abstraction made of the perversely polemical tracts surrounding the Spinoza dispute and the controversy with Schelling (Jacobi, 1916, 1967), all of them are historically as well as conceptually interesting. Among them, we can single out a few.

In 1777, Jacobi's friend and literary collaborator C.M. Wieland (1733-1813) had published an essay arguing that power is the only source of legitimacy for political authority, hence that the only adequate form of government is a strictly autocratic one.[13]Right follows upon the physical ability to enforce obedience. Jacobi replied with a point-by-point rebuttal that was published only four years later. (Jacobi, 1781) For one thing, Wieland was historically wrong by misconstruing the actual origin of societies. Even more important, Wieland was conceptually wrong by not recognizing that 'moral law'and 'natural necessity'can be subsumed under the one concept of 'natural right'only in a very broad sense. Moral rights derive their force from the freedom of individuals, not from any consideration of natural laws. There is an irreducible difference between the domain of nature and that of freedom. On Wieland's assumption, it would follow that any human action is morally right by the very fact it is performed, just as a natural event gives evidence of its natural necessity by the very fact that it occurs. Thus the attempt on a monarch's life would be justified, provided it is successful. And even if, *per impossibile*, Wieland's assumption were right, and 'moral necessity'were identical with 'natural necessity', his conclusions regarding the state, Jacobi argues, would still not follow necessarily. A philosopher such as Spinoza, though a naturalist in moral matters, had arrived at directly opposite results so far as the organization of the state is concerned.

Jacobi renewed his political polemic in 1782, this time on the occasion of the publication by the historian Johann von Müller of a pamphlet entitled *The Travels of the Popes*,[14]in which, contrary to current 'enlightened'views, a positive revaluation of the role of the Popes in the Middles Ages was offered. Jacobi responded with a pamphlet of his own (Jacobi, 1782) in which he defended Müller's position -- not because he had any sympathy for Catholicism, or because he was opposed to secularism, but because he thought that the Popes's spiritual despotism was much to be preferred to the secular, allegedly enlightened, despotism of the princes. Especially interesting about the pamphlet is the essentially classical, even Platonic, conception of reason that Jacobi advances there, in defence, however, of a theory of individual liberalism which, in the eighteenth-century, was being defended rather on the basis of quite an opposite mechanistic conception of rationality.

In 1788, Jacobi published a piece in dialogue form in the *Deutsches Museum*. (Jacobi, 1788)He did it in the midst of the notorious Starck affair, and in response to the campaign being waged at the time by the Berlin *Aufklärer*against the pious religiosity of people like Lavater -- a religiosity which they took as a form of crypto-Catholicism, and as an attack on the universal religion of reason which they promoted. (Blum, 1912) The dialogue is between a pious believer and an enlightened philosopher. The believer's main point is that the philosopher has the right indeed to criticize faith. But the believer has just as much right not to accept the philosopher's portrayal of his faith. For, at the abstract level of conceptualization at which the philosopher operates, he is in no position even to recognize the true nature of the object of his attack. He cannot understand that there is nothing in what he says abstractly about God and the world that has not already been known by the believer from time immemorial in the medium of the latter's historical faith. The philosopher, moreover, forgets that his philosophy has a history as well; that its past is shrouded in a faith on which it still depends for the meaning of its abstract conceptions. By attacking faith, therefore, the philosopher risks undermining his own world of meaning.[15]

In 1802, Jacobi published a short piece (Jacobi, 1802 (2)) in response to a prophesy made by G. C. Lichtenberg, a well known science popularizer of the day,[16]to the effect that some day, as science progresses in its efforts at reducing matter to the laws governing it, our world will become so refined in our eyes that it will be just as laughable to believe in God as is now to believe in ghosts. Jacobi replied with a prophesy of his own, imitating the Gospels's account of the last days. The day prophesized by Lichtenberg will come indeed. But, after a while yet, as the world becomes even more refined, the sages will reverse their judgements. And then -- that will be the end of all things -- they will believe in nothing but ghosts. They will be like God, in the knowledge that being and substance are everywhere but ghosts. (Jacobi, 1811: 3-4) Jacobi reproduced the piece at the opening of his *Of Divine Things and their Revelation*in 1811, in obvious parody of Schelling's philosophy of nature.

5. Retrospect

The first volume of Jacobi's collected works was published in 1812; the second, in 1815. This last is especially important because it contains, added to the text of the *David Hume,*a long new Preface intended by Jacobi, as the title indicates, also as Introduction to his life long philosophical production. ('Preface and also Introduction to the Author's Philosophical Collected Works', Jacobi, 1812-25, vol. 2) In the piece, Jacobi tried to sum up his intellectual odyssey by articulating the interest that had motivated it from beginning to end, and thereby also to bring some systematic unity to what might otherwise have appeared a scattered philosophical production. Jacobi was obviously sensitive to the charge of irrationalism that had repeatedly been brought against him over the years, and anxious to disarm it. He appealed to the distinction between 'reason'and 'understanding'that he had adopted from around 1800 to argue, as he had already done in *Of Divine Things*, that the kind of knowledge he had earlier presented under the rubric of 'faith'should be understood rather as a product of 'reason'-- a 'reason'properly understood, of course, as an intuitive faculty for the immediate apprehension of such eternal verities as the Good, the True, and the Beautiful. (Jacobi, 1812-1825, vol. 2, 1815: 59-63 and *passim)*He accordingly edited the 1787 text of the *David Hume*, attributing what he thought was the

inconclusiveness of the dialogue to his yet unclarified concept of 'reason'at the time of the first edition. He added long footnotes to it, and even modified some crucial passages of the text itself -- a circumstance, incidentally, for the most part ignored by later commentators -- in an obvious effort to dispel the naturalism otherwise clearly implied in the original text.[17]While thus distancing himself from any possible evidence of naturalism, Jacobi however also tried to deflect the accusation, sometime brought against him even by friends,[18]that he was against science. He tried to compensate for his past repeatedly made remark that science is but a game of abstractions (Jacobi, 1799: 22-27) by reasserting now, as he had already done in *Of Divine Things*, that, without the understanding's power to synthesize in the medium of abstract representations the content of sensations, reason, just like the senses, would have no form and hence no cognition of itself. (Jacobi, 1812-25: vol. 2, 58, 110)

How successful these defensive tactics were is open to discussion. The additions and modification made to the *David Hume*only succeeded in disrupting Jacobi's earlier theory without offering a credible new resolution to it. Moreover, as already remarked, it made little difference replacing 'faith'with 'reason'when the meaning of the latter remained just as unclarified as that of the earlier term. And Jacobi's positive remarks about science did not come unqualified. Jacobi also stressed that, however necessary the functions of the understanding, the latter is none the less still naturally prone to naturalism and to the atheism consequent upon it. The question could naturally be asked how, on Jacobi's premises, reason was *both*dependent on the understanding for its form yet naturally exposed to falsification at its hand. One could legitimately doubt, as Friedrich Schlegel had done earlier in a review *of Of Divine Things*, whether Jacobi had truly made peace with reason.[19]None the less, despite all ambivalence, there was no doubt about Jacobi's motivation throughout. Jacobi had always perceived himself as the champion of personalism, of human individuality and of human transcendence over nature -- values these, that he had always thought threatened by the rationalism of the Enlightenment. At the end of his life, reviewing his long struggle against Kant's Critique and its idealistic aftermath, he judged the struggle justified. For, as he thought events had demonstrated amply, that idealism was but a more sophisticated form of traditional metaphysics, and had indeed led to the same naturalism.

A measure of the great influence that Jacobi had in his own lifetime, and continued to have in the rest of the nineteenth century, is that he was the first to put in circulation the term 'nihilism', and to inaugurate the discourse associated wit it. Ironically, however, more often than not that influence did not work itself out in ways Jacobi himself would have wanted. He had been the one to bring Spinoza to the centre of philosophical discussion, and many were to be the young philosophers (Schleiermacher among them) who were first exposed to his pantheism through Jacobi's intermediary. Rather than rejecting it, however, as Jacobi would have expected, they often embraced it enthusiastically. Jacobi's influence on Fichte can also not be overestimated. And his crisp formulation of how Kantian idealism stood with respect to Spinoza's philosophy of substance -- namely, that it repeated the latter in subjective terms, the result being that, while it reintroduced the language of personalism, it also subverted it by changing its meaning -- caught indeed the imagination of many nineteenth century philosophers. But these took it to mean that the values of the old morality had run its historical course, and that it was high time to reestablish humanism on a new foundation. And this was a conclusion that Jacobi would have found just as abhorrent as he had found the French revolution.

Because of Jacobi's insistence on the primacy of immediate existence over reflective conceptualization, and of the rights of the 'exception', the possibility is there to interpret his position as case of proto-existentialism, and to treat him, just as Kierkegaard, as an essentially religious thinker. (Beiser, 1987) Indeed, some of the language Jacobi uses, and the themes he explores, are to be found in Kierkegaard again. (Whether the latter was himself an existentialist is, of course, itself an open question.) One must however keep in mind that the language of the 'leap of faith'does not belong to Jacobi. The *salto mortale*he had proposed to Lessing was no leap into the unknown but, according to his explicit testimony, a jump that would have brought Lessing, who had been walking on his head in the manner of philosophers, back to his feet. (Jacobi, 1787: 62; 1789: 353) And as for the religious outpourings that pervade his writings, and often mar them, they must be measured against what Jacobi himself had to say about his religiosity when confiding to Reinhold late in life. As he said, his problem, the source of his many ambiguities, was that, though temperamentally endowed with a Christian heart, his mind was just as temperamentally pagan. (Jacobi, 1825-27: vol. 2, 478) And there are testimonies to the effect that he always kept himself at a psychological distance from Christian believers. (di Giovanni, 1994: 42) At the end, he identified with 'reason'what he had earlier referred to as 'faith'. Also to be kept in mind is that the personalist values that Jacobi championed were Enlightenment values as well. Jacobi belongs very much to his times. To label him 'an anti-Enlightenment figure', as is routinely done, is perhaps misleading.

In sum, Jacobi's figure, including its place within the Enlightenment, is much more complex than usually assumed, and still in need of discussion. It might well be that the secret of this complexity is that Jacobi, just like Kierkegaard after him, was motivated by deeply conservative beliefs which he saw threatened by the culture of the day; but, again like Kierkegaard, in trying to reassert them, developed a language that was later to be used, contrary to anything he would have ever imagined, to undermine them instead.

Bibliography

Selected Original Sources

- Jacobi, F. H., 2000. 'On Transcendental Idealism', appendix to the 2nd Edition of *David Hume über den Glauben oder Idealismus und Realismus*, tr. B. Sassen, in *Kant's Early Critics, The Empiricist Critique of Theoretical Philosophy*, pp. 169-175. Cambridge: Cambridge University Press.
- Jacobi, F. H., 1994. *The Main Philosophical Writings and the Novel*Allwill, tr. ed. G. di Giovanni, includes complete major texts from original editions and with original pagination, historical and critical notes, extensive bibliography that includes a complete list of Jacobi's publications. Montréal &Kingston: McGill-Queen's University Press.
- Jacobi, F. H., 1988. *The Spinoza Conversations between Lessing and Jacobi: Text with Excerpts from the Ensuing Controversy*. Introduced by Gérard Vallée, trs G.Vallée, J.B.Lawson, C.G.Chapple. Lanham, New York: University Press of America.
- Jacobi, F. H., 1987. 'Open Letter to Fichte, 1799', tr. Diana I. Behler, in *Philosophy of German Idealism,* Ernst Behler (ed.), New York: Continuum, 119-141.

- Jacobi, F. H., 1981--. *Friedrich Heinrich Jacobi, Briefwechsel, Gesammtausgabe,*eds. Brüggen, M. &Sudhof, S. Stuttgart-Bad Constatt: Fromann-Holzboog, 1981--. This is the beginning of a critical edition. Volumes published to date, dedicated to early Jacobi's correspondence: Series 1, vols 1-3; Series 2, vols 1-2.
- Jacobi, F. H., 1967. *Streit um die göttlichen Dinge. Die Auseinandersetzung zwischen Jacobi und Schelling,* ed. Weischedel, W., includes texts from the dispute between Jacobi and Schelling Darmstadt: Wissenschaftliche Buchgesellschaft.
- Jacobi, F. H., 1957. *Friedrich Heinrich Jacobis*"*Allwill*", critical edition with introduction and notes by J. U. Terpstra. Groningen: Djacarta.
- Jacobi. F. H., 1946. *Oeuvres philosophiques de F.-H. Jacobi,* tr. in French Anstett, J-J. Paris: Aubier.
- Jacobi, F. H., 1916. *Die Hauptschriften zum Pantheismusstreit zwischen Jacobi und Mendelssohn,* ed. Scholz, H., includes all the relevant texts in the dispute between Jacobi and Mendelssohn. Berlin: Reuther &Reichard.
- Jacobi, F. H. 1869.*Aus F.H. Jacobi's Nachlaß. Ungedruckte Briefe von und an Jacobi und andere. Nebst ungedruckten Gedichten von Goethe und Lenz.,*2 vols, ed. Zoeppritz, R. Leipzig: Engelmann.
- Jacobi, F. H., 1854. *F. H. Jacobi's ausgewählte Werke*, vols. I-III. Leipzig: G. Fleischer, 1854.
- Jacobi, F. H. 1825-27. *Friedrich Heinrich Jacobi's auserlesener Briefwechsel,* 2 vols., ed. Roth, F. Leipzig: Fleischer.
- Jacobi, F. H., 1812-1825. *Friedrich Heinrich Jacobi's Werke*, eds Köppen, J.F. and Roth, C.J.F., vols. I-VI. Leipzig: Gerhard Fleischer. Reprinted, Darmstadt: Wissenschaftliche Buchgesellschaft, 1968. This is the edition that Jacobi himself supervised before his death and the one still most easily availabe. It must however be treated with care, since editorial comments and emendations tend to impose Jacobi's later views on the earlier texts.
- Jacobi, F. H., 1811. *Von den Göttlichen Dingen und ihrer Offenbarung*. Leipzig: G.Fleischer.
- Jacobi, F. H., 1807. *Über gelehrte Gesellschaften, ihren Geist und Zweck. Eine Abhandlung, vorgelesen bey der feyerlichen Erneuung der Königlichen Akademie der Wissenschaften zu München von dem Präsidenten der Akademie*. München: E. A. Fleischmann.
- Jacobi, F. H., 1803. [Three letters in] Köppen, Friedrich. *Schellings Lehre oder das Ganze der Philosophie des absoluten Nichts, Nebst drey Briefen verwandten Inhalts von Friedr. Heinr. Jacobi*. Hamburg: 207-278.
- Jacobi, F. H., 1802 (1). 'Über das Unternehmen des Kriticismus, die Vernunft zu Verstande zu bringen, und der Philosophie überhaupt eine neue Absicht zu geben.'*Beyträge zur leichtern Uebersicht des Zustandes der Philosophie beym Anfange des 19. Jahrhunderts*, ed. Reinhold, C.L. Hamburg, 3: 1-110.
- Jacobi, F.H., 1802 (2). 'Über eine Weissagung Lichtenbergs.'*Taschenbuch für das Jahr 1802*, ed. Jacobi, J. Georg. Hamburg: Perthes: 3-46.
- Jacobi, F. H., 1799. *Jacobi an Fichte*. Hamburg: Perthes. Supplement II was republished in *Werke*II, 1815 (Jacobi,1812-1825) under the title of 'Über die Unzertrennlichkeit der Freiheit und Vorsehung von dem Begriffe der Vernunft'.
- Jacobi, F. H., 1795. 'Zufällige Ergießungen eines einsamen Denkers in Briefen an vertraute Freunde.'*Die Horen,* ed. Schiller, 3.8: 1-34.

- Jacobi, F. H., 1794. *Woldemar*, 2 Theile. Königsberg: Nicolovius.
- Jacobi, F. H., 1792.*Eduard Allwills Briefsammlung, herausgegeben von Friedrich Heinrich Jacobi, mit einer Zugabe von eigenen Briefen*. Königsberg: Nicolovius. 1957, critical edition, ed. Terpstra, J.U. Groningen: Djakarta, 1957.
- Jacobi, F. H., 1789. *Über die Lehre des Spinoza in Briefen an den Herrn Moses Mendelssohn. Neue vermehrte Ausgabe*. Breslau: Gottlieb Löwe.
- Jacobi, F. H. 1788. 'Einige Betrachtungen über den frommen Betrug und über eine Vernunft, welche nicht die Vernunft ist, von Friedrich Heinrich Jacobi in einem Briefe an den Herrn geheimen Hofrath Schlosser.'*Deutsches Museum*, 1.2: 153-184.
- Jacobi, F. H., 1787.*David Hume über den Glauben, oder Idealismus und Realismus. Ein Gespräch*. Breslau: Gottlieb Löwe. 1983 facsimile reproduction that includes the *Vorrede*of 1815, New York and London: Garland.
- Jacobi, F. H., 1786. *Friedrich Heinrich Jacobi wider Mendelssohns Beschuldigungen betreffend die Briefe über die Lehre des Spinoza*. Leipzig Georg Joachim Goeschen.
- Jacobi, F. H., 1785. *Über die Lehre des Spinoza in Briefen an den Herrn Moses Mendelssohn.*Breslau: Gottlieb Löwe.
- Jacobi, F. H., 1782. *Etwas das Leßing gesagt hat. Ein Commentar zu den Reisen der Päpste nebst Betrachtungen von einem Dritten.*Berlin: George Jacob Decker.
- Jacobi, F. H., 1781. 'Über Recht und Gewalt, oder philosophische Erwägung eines Aufsazes von dem Herrn Hofrath Wieland, über das göttliche Recht der Obrigkeit, im deutschen Merkur, November 1777'*Deutsches Museum*, 1: 522-554.
- Jacobi, F. H., 1779 (4). 'Ein Stück Philosophie des Lebens und der Menschheit: Aus dem zweiten Bande von Woldemar.'*Deutsches Museum*, 1: 307-348; 393-427.
- Jacobi, F. H., 1779 (3). *Woldemar. Eine Seltenheit aus der Naturgeschichte*, vol. 1. Flensburg and Leipzig.
- Jacobi, F. H., 1779 (2). 'Noch eine politische Rhapsodie, worinn sich verschiedene Plagia befinden; betitelt: Es ist nicht recht, und es ist nicht klug.'*Baierische Beyträge zur schönen und nützlichen Litteratur*,1.5: 418-458.
- Jacobi, F. H., 1779 (1). 'Eine politische Rhapsodie. Aus einem Aktenstock entwendet. Ein eingesandtes Stück.'*Baierische Beyträge zur schönen und nützlichen Litteratur*, 1.5: 407-418.
- Jaocobi/Baggasens, Reinhold, 1831. *Jens Baggasens Briefwechsel mit Karl Leonhard Reinhold und Friedrich Heinrich Jacobi*, 2 vols, ed. Baggasen, K. &A. Leipzig: Brockhaus.
- Jacobi/Bouterwek, 1868. *Fried. Heinr. Jacobi's Briefe an Friedr. Bouterwek aus dem Jahren 1800 bis 1819*, ed. Meyer, W. Göttingen: Deuer.
- Jacobi/Hamann, 1955-1979. *Johann Georg Hamann, Briefwechsel*, ed. Henkel, A. vols 6-7. Wiesbaden &Frankfurt am Main: Insel-Verlag.
- Jacobi/Goethe, 1846. *Briefwechsel zwischen Goethe und Fr.H.Jacobi*, ed. Jacobi, M. Leipzig: Weidmannsche Buchhandlung.
- Jacobi/Humboldt, 1892. *Briefe von Wilhelm von Humboldt an R. H. Jacobi*, ed Leitzman, A. Halle: Niemeyer.
- Jacobi/Mendelssohn, 1929.*Mendelssohn-Briefwechsel*, ed. Altmann, A.. In *Moses Mendelssohn: Gesammelte Schriften*, Jubiläumsausgabe. Berlin: Akademie-Verlag. Reprint, 1971: Stuttgart-Bad Cannstatt: Frommann-Holzboog

Selected Secondary Literature

- Baum, Günther, 1971. 'Über das Verhältnis von Erkenntnisgewißheit und Anschauungsgewißheit in F. H. Jacobis Interpretation der Vernunft.'*Friedrich Heinrich Jacobi. Philosoph und Literat der Goethezeit. Beiträge einer Tagung in Düsseldorf (16.-19. 10. 1969) aus Anlaßseines 150. Todestages und Berichte*, ed. Hammacher, K. Frankfurt am Main: Vittorio Klostermann: 7-26.
- -----,1969. *Vernunft und Erkenntnis. Die Philosophie F. H. Jacobis.*Bonn: Bouvier.
- Beiser, Frederick C., 1992. *Enlightenment, Revolution and Romanticism: The Genesis of Modern Political Thought, 1790-1800.*Cambridge, Mass.: Harvard University Press.
- -----, 1987. *The Fate of Reason.*Cambridge, Mass.: Harvard.
- Blum, Jean, 1912.*J.A. Starck et la querelle du crypto-catholicisme en Allemagne, 1785-1789.*Paris: Alcan.
- Bollnow, Otto Friedrich, 1933. *Die Lebensphilosophie F. H. Jacobis*. Stuttgart.
- Booy, J.Th. de, &Mortier, Roland, 1966.*Les années de formation de F.H.Jacobi, d'après ses lettres inédites àM.M.Rey (1763-1771), avec "Le Noble" de Madame de Charrière*. Geneva: Institut et Musée Voltaire.
- Brachin, Pierre, 1952. *Le Cercle de Münster (1779-1806) et la Pensée religieuse de F.L.Stolberg*. Paris: IAC.
- Brüggen, Michael, 1971. 'Jacobi, Schelling und Hegel,'*Friedrich Heinrich Jacobi. Philosoph und Literat der Goethezeit. Beiträge einer Tagung in Düsseldorf (16.-19. 10. 1969) aus Anlaßseines 150. Todestages und Berichte*, ed.Hammacher, K. Frankfurt am Main: Vittorio Klostermann: 209-232.
- -----, 1967/68. 'Jacobi und Schelling,'*Philosophisches Jahrbuch der Görresgesellschaft*, 75: 419-429.
- ----- , 1967. 'La critique de Jacobi par Hegel dans "Foi et Savoir",'*Archives de Philosophie*, 30: 187-198.
- Brüggen, Michael; Gockel, Heinz; Schneider, P.-P., eds, 1989. *Friedrich Heinrich Jacobi, Dokumente zu Leben und Werk*, vol. 1, *Die Bibliothek Friedrich Heinrich Jacobis*,2 parts. Stuttgart-Bad Canstatt: Frommann-Holzboog.
- David, Frida, 1913.*Friedrich Heinrich Jacobis 'Woldemar'in seinen verschiedenen Fassungen*. Leipzig: Voigtländer.
- David, Frieda, 1913. *Friedrich Heinrich Jacobis 'Woldemar'"in seinen verschiedenen Fassungen*. Leipzig: Voigtländer.
- di Giovanni, George, 2001. 'Rehberg, Reinhold und C.C.E. Schmid über Kant und moralische Freiheit', in *Vernunftkritik und Aufklärung*,eds. Oberhausen, M., Delfosse, D. P., Pozzo, R. Suttgart-Bad Canstaat: Fromann-Holzboog: 93-113.
- -----, 1997. 'Hume, Jacobi, and Common Sense: An Episode in the Reception of Hume in Germany at the Time of Kant', *Kant-Studien*, 88: 44-58
- -----, 1995. 'Hegel, Jacobi, and Crypto-Catholicism, or, Hegel in Dialogue with the Enlightenment', *Hegel on the Modern World*, ed. Collins, A. Albany: SUNY: 53-72.
- -----, 1994. 'The Unfinished Philosophy of Friedrich Heinrich Jacobi,'*The Main Philosophical Writings and the Novel* Allwill. Montréal &Kingston: McGill-Queen's University Press: 1-167.

- -----, 1989. ‘From Jacobi’s Philosophical Novel to Fichte’s Idealism: Some Comments on the 1798-99 “Atheism Dispute”,’ *Journal of the History of Philosophy*, 27: 75-100.
- Ford, Lewis S., 1965. ‘The Controversy Between Schelling and Jacobi.’ *Journal of the History of Philosophy*, 3: 75-89.
- Göres, Jörn, 1977. ‘Veränderungen 1774-1794. Goethe, Jacobi und der Kreis von Münster.’ *Staat und Gesellschaft im Zeitalter Goethes. Festschrift für Hans Timmler zu seinem 70. Geburtstag*, ed. Berglar, P. Köln, Wien: Böhlau: 273-284.
- Hammacher, Klaus, ed., 1998. *Fichte und Jacobi.Fichte-Studien*, vol. 14. Amsterdam - Atlanta GA: Rodopoi.
- -----1985. ‘Über Friedrich Heinrich Jacobis Beziehungen zu Lessing im Zusammenhang mit dem Streit um Spinoza,’ *Lessing und der Kreis seiner Freunde*, ed. Schulz, G. Heidelberg: L. Schneider: 51-74.
- -----, 1969. *Kritik und Leben II. Die Philosophie Friedrich Heinrich Jacobis*. München: Wilhelm Fink Verlag.
- Hammacher, Klaus; Hirsch, Hans,1993. *Die Wirtschaftspolitik des Philosophen Jacobi*. Amsterdam - Atlanta GA: Rodopoi.
- Höhn, Gerhard, 1970. ‘F. H. Jacobi et G. W. Hegel ou la naissance du nihilisme et la renaissance du “Logos”,’ *Revue de Métaphysique et de Morale*, 75: 129-150.
- Kirscher, Gilbert, 1970. ‘Hegel et Jacobi critiques de Kant,’ *Archives de Philosophie*, 33: 801-828.
- -----, 1969. ‘Hegel et la philosophie de F. H. Jacobi,’ *Hegel-Studien*, Beiheft 4. Bonn: Bouvier: 181-191.
- Knoll, Renate, 1979. ‘Hamanns Kritik an Jacobi mit Jacobis Briefen vom 1., 6. und 30. 4. 1787 und Hamanns Briefen vom 17., 22. und 27. 4. 1787,’ *Johann Georg Hamann, Acta des Internationalen Hamann-Colloquiums in Lüneburg 1976*, ed. Gajek, B. Frankfurt: Klostermann: 214-234; 234-236; 236-276.
- -----, 1963. *Johann Georg Hamann und Friedrich Heinrich Jacobi*. Heidelberg: Winter.
- Lauth, Reinhard, 1971 (1). ‘Fichtes Verhältnis zu Jacobi unter besonderer Berücksichtigung der Rolle Friedrich Schlegels in dieser Sache,’ *Friedrich Heinrich Jacobi. Philosoph und Literat der Goethezeit. Beiträge einer Tagung in Düsseldorf (16.-19. 10. 1969) aus Anlaßseines 150. Todestages und Berichte*, ed. Hammacher, K. Frankfurt am Main: Vittorio Klostermann: 165-197.
- -----, 1971 (2). ‘Nouvelles recherches sur Jacobi I et II,’ *Archives de Philosophie*, 34: 281-286; 495-502.
- Lévy-Bruhl, 1894. *La philosophie de Jacobi*.Paris: Alcan.
- Nicolai, Heinz, 1965. *Goethe und Jacobi, Studien zur Geschichte ihrer Freundschaft.*Stuttgart: Metzler.
- Prantl, Karl, 1881. ‘Jacobi: Friedrich Heinrich,’ *Allgemeine Deutsche Biographie*, vol. 13. Leipzig: Duncker&Humblot: 577-584.
- Sandkaulen, Birgit, 2000. *Grund und Ursache: Die Vernunftkritik Jacobis*. München: Fink.
- Snow, Dale Evarts, 1987. ‘F.H. Jacobi and the Development of German Idealism,’ *Journal of the History of Philosophy*, 25.3: 397-415.
- Stockum, Theodor Cornelius van, 1957. ‘Goethe, Jacobi und die Ettersburger “Woldemar-Kreuzigung”(1779),’ *Neophilologus*, 41: 106-116.
- Sudhof, Siegfried, 1959. ‘Friedrich Heinrich Jacobi und die “Kreuzigung”seines

Woldemar,'*Neophilologus*, 43: 42-49.

- Süß, Theobald, 1951. 'Der Nihilismus bei F. H. Jacobi,'*Theol. Literatur-Zeitung*, 76: 193-200.
- Timm, Hermann, 1974.*Gott und die Freiheit. Studien zur Religionphilosophie der Goethezeit: I, Die Spinozarenaissance.*Frankfurt/Main: Klostermann.
- -----, 1971. 'Die Bedeutung der Spinozabriefe Jacobis für die Entwicklung der idealistischen Religionsphilosophie,'*Friedrich Heinrich Jacobi. Philosoph und Literat der Goethezeit. Beiträge einer Tagung in Düsseldorf (16.-19. 10. 1969) aus Anlaßseines 150. Todestages und Berichte*, ed. Hammacher, K Frankfurt am Main: Vittorio Klostermann: 35-81.
- Trunk, Erich, ed., 1955. *Fürstenberg, Fürstin Gallitzin, und ihr Kreis*. Münster: Aschendorff.
- Verra, Valerio, 1963. *F. H. Jacobi. Dall'Illuminismo all'Idealismo*. Torino: Edizioni di Filosofia.
- Weindel, Philipp, 1950. 'Fr. H. Jacobis Einwirkung auf die Glaubenswissenschaft der katholischen Tübinger Schule,'*Aus Theologie und Philosophie. Festschrift für Fritz Tillmann zu seinem 75. Geburtstag*, eds Steinbüchel, Th. &Müncker, Th. Düsseldorf: Patmos-Verlag: 573-596.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Fichte, Johann Gottlieb| Hegel, Georg Wilhelm Friedrich| Hume, David| Kant, Immanuel | Leibniz, Gottfried Wilhelm | Mendelssohn, Moses | Schelling, Friedrich Wilhelm Joseph von| Spinoza, Baruch [Benedict]

George di Giovanni
McGill University
george.di_giovanni@mcgill.ca

A| B| C| D| E| F| G| H| I| J| K| L| M| N| O| P| Q| R| S| T| U| V| W| X| Y| Z

Table of Contents

First published: December 6, 2001
Content last modified: December 6, 2001

Stanford Encyclopedia of Philosophy
Notes to Friedrich Heinrich Jacobi

Notes

1. For a detailed sketch of Jacobi's life, the article in the *Allgemeine Deutsche Biographie* (Prantl, 1881) is still the best but must be supplemented with more recent discoveries. (Booy/Mortier, 1966). Basic information is drawn from Jacobi's autobiographical comments scattered throughout his works (Jacobi, 1812-1825, 1787, 1785) and enriched from his massive correspondence and the correspondence of third parties relating to him.

2. Johann Wolfang von Goethe, *Dichtung und Wahrheit*, III, p. 681. In *Sämmtliche Werke*, eds Richter, K. *et al.*, vol. 16. München: Hauser, 1985-.

3. Kant reacted to the dispute with his essay 'Was heißt sich im Denken Orientieren?' ('What Does It Mean to Orient Oneself in Thought,' *Berlinische Monatsschrift*, 8(1786): 304-330). Without endorsing Mendelssohn, whose metaphysics he could not have accepted, he none the less showed that his sympathy lay on his side. See also, di Giovanni, 1994: 32, note 70.

4. Johann August Starck, theologian and in his days renowned preacher at the court of Darmstadt, became at some point deeply involved in the shadowy side of Masonic internal politics. He apparently zealously conspired to establish within the order an 'ecclesiastical branch' that would incorporate in its ceremonials older rituals allegedly going back to the Mediaeval Templars. For this intriguing, he was repeatedly criticized on the pages of the Berlinmonatsschrift and of the Allgemeine deutsche Bibliothek, and accused of obscurantism. He was connected by the two journals with what at the time was widely feared to be a plot by the Jesuits to undermine Protestant enlightened society. Starck reacted by suing the two editors of the Berlinmonatsschrift (F. Gedike and J. Biester) for defamation of character. The courts found in favour of the two editors, on the ground that it was in the interest of the press to expose perceived dangers to society.

5. As recently expressed in *Gott, Einige Gespräche.* Gotha: Ettinger, 1787.

6. *Morgenstunden, oder Vorlesungen über das Daseyn Gottes, Erster Theil* (*Morning Hours, or Lectures Concerning the Existence of God.* Berlin: Voß, 1785)

7.See the record of Jacobi's 1788 oral comments regarding Reid, in *Wilhelm von Humboldts Tagebücher, 1788-1789,* ed. Albert Leitzmann (Berlin: Bher, 1916), 58, 61. See, also, ; di Giovanni, 1997.

8. Anthony J. La Vopa, *Fichte: The Self and the Calling of Philosophy, 1762-1799* (Cambridge: University Press, 2001, Chapters12-13.

9. Friedrich Schlegel's review of the 1796 edition amounted to a scathing attack. *Deutschland,* 3.8(1796): 185-213.

10. There are remarkable similarities between Allwill and Kierkegaard's seducer in *Either/Or*.

11. Hegel paraphrases some of the language in the scene in the *Phenomenology of Spirit,* tr. A. V. Miller (Oxford: Clarendon), p. 409.

12. Jacobi reproduced the passage in an appendix to his *Open Letter to Fichte*. (Jacobi, 1799: 101)

13.145;Über das göttliche Recht der Obrigkeit' ('On the Divine Right of Authority'), *Der Teutscher Merkur,* 20(1777): 119-45.

14. Originally published in French; German translation, *Die Reisen der Päbste*, 1783.

15. For a detailed account of the context of Jacobi's essay, and its possible connection with Hegel's *Phenomenology of Spirit,* see di Giovanni, 1995.

16. The Review Focus of volume 25 of the *Lessing Yearbook* (1993) is dedicated to Lichtenberg.

17. See the alteration made in the 1815 edition to the text on p. 123 of the original, and the new note entered by Jacobi.

18. See, for instance, J. F. Fries, *Von Deutscher Philosophie, Art, und Kunst. Ein Votum für Friedrich Heinrich Jacobi gegen F. W. J. Schelling* (Heidelberg: Mohr und Zimmer, 1812), pp. 40-49, especially 40-41, 44-48.

19. *Deutsches Museum,* 1.1(1812): 79-98, especially pp. 89ff.

George di Giovanni
McGill University
george.di_giovanni@mcgill.ca

First published: December 6, 2001

Content last modified: December 7, 2001

Friedrich Wilhelm Joseph von Schelling

Friedrich Wilhelm Joseph von Schelling (1775-1854) is, along with J.G. Fichte and G.W.F. Hegel, one of the three most influential thinkers in the tradition of ‘German Idealism’. Although he is often regarded as a philosophical Proteus who changed his conception so radically and so often that it is hard to attribute a clear philosophy to him, Schelling was in fact often an impressively rigorous logical thinker. In the era during which Schelling was writing, so much was changing in philosophy that a stable, fixed point of view was as likely to lead to a failure to grasp important new developments as it was to lead to a defensible philosophical system. Schelling’s continuing importance today relates mainly to three aspects of his work. The first is his Naturphilosophie, which, although its empirical claims are largely indefensible, opens up the possibility of a modern hermeneutic view of nature that does not restrict nature’s significance to what can be established about it in scientific terms. The second is his anti-Cartesian account of subjectivity, which prefigures some of the best ideas of thinkers like Nietzsche and Jacques Lacan, in showing how the thinking subject cannot be fully transparent to itself. The third is his later critique of Hegelian Idealism, which influenced Kierkegaard, Marx, Nietzsche, Heidegger, and others, and aspects of which are still echoed in contemporary thought by thinkers like Jacques Derrida.

1. Career

Schelling was born in Leonberg near Stuttgart on 27 January 1775. He attended a Protestant seminary in Tübingen from 1790 to 1795, where he was close friends with both Hegel and the poet and philosopher Friedrich Hölderlin. He moved to Leipzig in 1797, then to Jena, where he came into contact with the early Romantic thinkers, Friedrich Schlegel and Novalis, and, via Goethe’s influence, took up his first

professorship from 1798 to 1803. From 1803 to 1806 he lived in Würzburg, whence he left for Munich, where he mainly lived from 1806 onwards, with an interruption from 1820 to 1827, when he lived in Erlangen. He moved to Berlin in 1841 to take up what had, until Hegel's death in 1831, been Hegel's chair of philosophy. Although his lectures in Berlin were initially attended by such luminaries as Kierkegaard, Engels, Bakunin, Ranke, Burkhardt, and Alexander von Humboldt, he soon came to be largely ignored by most of the leading thinkers of the day. It is clear, however, that his philosophical thought still influenced many who rejected him on mainly political grounds. He died on 20 August 1854 in Bad Ragaz, Switzerland. Schelling's influence on many directions in modern philosophy has been seriously underestimated in the English-speaking world, though this underestimation is now beginning to be countered by renewed attention to his work.

2. Transcendental Philosophy and *Naturphilosophie*

The significance of the work of the early Schelling (1795-1800) lies in its attempts to give a new account of nature which, while taking account of the fact that Kant has irrevocably changed the status of nature in modern philosophy, avoids some of the problematic consequences of Kant's theory. For the Kant of the *Critique of Pure Reason* (1781, 1787) nature is largely seen in the 'formal' sense: nature is that which is subject to necessary laws. These laws are accessible to us, Kant argues, because cognition depends on the subject bringing necessary forms of thought, the categories, to bear on what it perceives. The problem this leads to is how the subject could fit into a nature conceived of in deterministic terms, given that the subject's ability to know is dependent upon its 'spontaneous' self-caused ability to judge in terms of the categories. Kant's response to this dilemma is to split the 'sensuous' realm of nature as law-bound appearance from the 'intelligible' realm of the subject's cognitive and ethical self-determination. However, if the subject is part of nature there would seem to be no way of explaining how a nature which we can only *know* as deterministic can give rise to a subject which seems to transcend determinism in its knowing and in its ethical doings. Kant himself sought to bridge the realms of necessity and spontaneity in the *Critique of Judgement* (1790), by suggesting that nature itself could be seen in more than formal terms: it also produces self-determining organisms and can give rise to disinterested aesthetic pleasure in the subject that contemplates its forms. The essential problems remained, however, that 1) Kant gave no account of the genesis of the subject that transcends its status as a piece of determined nature, and 2) such an account would have to be able to bridge the divide between nature and freedom.

The tensions in Schelling's philosophy of this period, which set the agenda for most of his subsequent work, derive, then, from the need to overcome the perceived lack in Kant's philosophy of a substantial account of how nature and freedom come to co-exist. Two ways out of Kantian dualism immediately suggested themselves to thinkers in the 1780s and 90s. On the one hand, Kant's arguments about the division between appearances and things in themselves, which gave rise to the problem of how something 'in itself' could give rise to appearances for the subject, might be overcome by rejecting the notion of the thing in itself altogether. If what we know of the object is the product of the spontaneity of the I, an Idealist could argue that the whole of the world's intelligibility is therefore the result of the activity of the subject, and that a new account of subjectivity is required which would achieve what Kant had failed to achieve. On the other hand, the fact that nature gives rise to self-determining subjectivity would seem to

suggest that a monist account of a nature which was more than a concatenation of laws, and was in some sense inherently ‘subjective’, would offer a different way of accounting for what Kant’s conception did not provide. Schelling seeks answers to the Kantian problems in terms that relate to both these conceptions. Indeed, it is possible to argue that the conceptions are in one sense potentially identical: if the essence of nature is that it produces the subjectivity which enables it to understand itself, nature itself could be construed as a kind of ‘super-subject’. The main thinkers whose work establishes these alternatives are J.G. Fichte, and Spinoza.

The source of Schelling’s concern with Spinoza is the ‘Pantheism controversy’, which brought Spinoza’s monism into the mainstream of German philosophy. In 1783 the writer and philosopher F.H. Jacobi became involved in an influential dispute with the Berlin Enlightenment philosopher Moses Mendelssohn over the claim that G.E. Lessing had admitted to being a Spinozist, an admission which at that time was tantamount to the admission of atheism, with all the dangerous political and other consequences that entailed. In his *On the Doctrine of Spinoza in Letters to Herr Moses Mendelssohn*, (1785, second edition 1789), which was influenced by his reading of Kant’s first *Critique*, Jacobi revealed a problem which would recur in differing ways throughout Schelling’s work. Jacobi’s interpretation of Spinozism was concerned with the relationship between what he termed the ‘unconditioned’ and the ‘conditioned’, between God as the ground of which the laws of nature are the consequent, and the linked chains of the deterministic laws of nature. Cognitive explanation relies, as Kant suggested, upon finding a thing’s ‘condition’. Jacobi’s question is how finding a thing’s condition can finally ground its explanation, given that each explanation leads to a regress in which each condition depends upon another condition *ad infinitum*. Any philosophical system that would ground the explanation of a part of nature thus ‘necessarily ends by having to discover *conditions* of the *unconditioned*’. For Jacobi this led to the need for a theological leap of faith, as the world’s intelligibility otherwise threatened to become a mere illusion, in which nothing was finally grounded at all. In the 1787 Introduction to the first *Critique* Kant maintains this problem of cognitive grounding can be overcome by acknowledging that, while reason must postulate the ‘unconditioned (...) in all things in themselves for everything conditioned, so that the series of conditions should thus become complete’, by restricting knowledge to appearances, rather than allowing it to be of ‘things in themselves’, the contradiction of seeking conditions of the unconditioned can be avoided. As we have already seen, though, this gives rise precisely to the problem of how a subject which is not conditioned like the nature it comes to know can emerge as the ground of knowledge from nature.

The condition of the knowledge of appearances for Kant is the ‘transcendental subject’, but what sort of ‘condition’ is the transcendental subject? The perception that Kant has no proper answer to this problem initially unites Schelling and Fichte. Fichte insists in the *Wissenschaftslehre* (1794) that the unconditioned status of the I has to be established if Kant’s system is to legitimate itself. He asserts that ‘It is (...) the ground of explanation of all facts of empirical consciousness that before all positing in the I the I itself must previously be posited’, thereby giving the I the founding role which he thought Kant had failed adequately to explicate. Fichte does this by extending the consequences of Kant’s claim that the cognitive activity of the I, via which it can reflect upon *itself*, cannot be understood as part of the causal world of appearances, and must therefore be part of the noumenal realm, the realm of the ‘unconditioned’. For Fichte the very fact of philosophy’s existence depends upon the free act of the I

which initiates the reflective questioning of its own activity by the I.

Schelling takes up the issues raised by Jacobi and Fichte in two texts of 1795: *Of the I as Principle of Philosophy or on the Unconditional in Human Knowledge*, and *Philosophical Letters on Dogmatism and Criticism*. In a move which prefigures aspects of Heidegger's questioning of the notion of being, he reinterprets Kant's question as to the condition of possibility of synthetic judgements a priori as a question about why there is a realm of judgements, a manifest world requiring syntheses by the subject for knowledge to be produced, at all. In *Of the I* Schelling puts Kant's question in Fichtean terms: 'how is it that the absolute I goes out of itself and opposes a Not-I to itself?'. He maintains that the condition of knowledge, the 'positing' by the I of that which is opposed to it, must have a different status from the determined realm which it posits: 'nothing can be posited by itself as a thing, i.e. an absolute/unconditioned thing (*unbedingtes Ding*) is a contradiction'. However, his key worry about Fichte's position already becomes apparent in the *Philosophical Letters*, where he drops the Fichtean terminology: '*How is it that I step at all out of the absolute and move towards something opposed (auf ein Entgegengesetztes)?*'. The problem Schelling confronts was identified by his friend Hölderlin, in the light of Jacobi's formulation of the problem of the 'unconditioned'. Fichte wished to understand the absolute as an I in order to avoid the problem of nature 'in itself' which creates Kantian dualism. For something to be an I, though, it must be conscious of an other, and thus in a relationship to that other. The overall structure of the relationship could not, therefore, be described from only one side of that relationship. Hölderlin argued that one has to understand the structure of the relationship of subject to object in consciousness as grounded in 'a whole of which subject and object are the parts', which he termed 'being'. This idea will be vital to Schelling at various times in his philosophy.

In the 1790s, then, Schelling is seeking a way of coming to terms with the ground of the subject's relationship to the object world. His aim is to avoid the fatalist consequences of Spinoza's system by taking on key aspects of Kant's and Fichte's transcendental philosophy, and yet not to fall into the trap Hölderlin identified in Fichte's conception of an absolute I. In his *Naturphilosophie* (philosophy of nature), which emerges in 1797 and develops in the succeeding years, and in the *System of Transcendental Idealism* of 1800, Schelling wavers between a Spinozist and a Fichtean approach to the 'unconditioned'. In the *Naturphilosophie* the Kantian division between appearing nature and nature in itself is seen as resulting from the fact that the nature theorised in cognitive judgements is objectified in opposition to the knowing subject. This objectification, the result of the natural sciences' search for fixed laws, fails to account for the living dynamic forces in nature, including those in our own organism, with which Kant himself became concerned in the third *Critique* and other late work, and which had played a role in Leibniz's account of nature. Nature in itself is thought of by Schelling as a 'productivity': 'As the object [*qua* 'conditioned condition'] is never absolute/unconditioned (*unbedingt*) then something per se non-objective must be posited in nature; this absolutely non-objective postulate is precisely the original productivity of nature'. The Kantian dualism between things in themselves and appearances is a result of the fact that the productivity can never appear as itself and can only appear in the form of 'products', which are the productivity 'inhibiting' itself. The products are never complete in themselves: they are like the eddies in a stream, which temporarily keep their shape via the resistance of the movement of the fluid to itself that creates them, despite the changing material flowing through them.

Schelling next tries to use the insights of transcendental philosophy, while still avoiding Kant's dualism, to explain our knowledge of nature. The vital point is that things in themselves and 'representations' cannot be absolutely different because we know a world which exists independently of our will which can yet be affected by our will:

> one can push as many transitory materials as one wants, which become finer and finer, between mind and matter, but sometime the point must come where mind and matter are One, or where the great leap that we so long wished to avoid becomes inevitable.

The *Naturphilosophie* includes ourselves within nature, as part of an interrelated whole, which is structured in an ascending series of 'potentials' that contain a polar opposition within themselves. The model is a magnet, whose opposing poles are inseparable from each other, even though they are opposites. As productivity nature cannot be conceived of as an object, since it is the subject of all possible real 'predicates', of the 'eddies' of which transient, objective nature consists. However, nature's 'inhibiting' itself in order to become something determinate means that the 'principle of all explanation of nature' is 'universal duality', an inherent difference of subject and object which prevents nature ever finally reaching stasis. At the same time this difference of subject and object must be grounded in an identity which links them together, otherwise all the problems of dualism would just reappear. In a decisive move for German Idealism, Schelling parallels the idea of nature as an absolute producing subject, whose predicates are appearing objective nature, with the spontaneity of the thinking subject, which is the condition of the syntheses required for the constitution of objectivity, thus for the possibility of predication in judgements. The problem for Schelling lies in explicating how these two subjects relate to each other.

In the *System of Transcendental Idealism* Schelling goes back to Fichtean terminology, though he will soon abandon most of it. He endeavours to explain the emergence of the thinking subject from nature in terms of an 'absolute I' coming retrospectively to know itself in a 'history of self-consciousness' that forms the material of the system. The *System* recounts the history of which the transcendental subject is the result. A version of the model Schelling establishes will be adopted by Hegel in the *Phenomenology of Mind*. Schelling presents the process in terms of the initially undivided I splitting itself in order to articulate itself in the syntheses, the 'products', which constitute the world of knowable nature. The founding stages of this process, which bring the world of material nature into being, are 'unconscious'. These stages then lead to organic nature, and thence to consciousness and self-consciousness. Schelling claims, in the wake of Fichte, that the resistance of the noumenal realm to theoretical knowledge results from the fact that 'the [practical] act [of the absolute I] via which all limitation is posited, as condition of all consciousness, does not itself come to consciousness'. He prophetically attempts to articulate a theory which comes to terms with the idea that thought is driven by forces which are not finally transparent to it, of the kind later to become familiar in psychoanalysis. How, though, does one gain access by thought to what cannot be an object of consciousness? This access is crucial to the whole project because without it there can be no understanding of why the move from determined nature to the freedom of self-determining thinking takes place at all.

Schelling adopts the idea from the early Romantic thinkers Friedrich Schlegel and Novalis, whom he

knew in Jena at this time, that art is the route to an understanding of what cannot appear as an object of knowledge. Philosophy cannot represent nature in itself because access to the sphere of the unconscious must be via what appears to consciousness in the realm of theoretical knowledge. The work of art is evidently an empirical, appearing object like any other, but if it is not more than what it is *qua* determinable object it cannot be a work of art, because this requires both the free judgement of the subject and the object's conveying of something beyond its objective nature. Although the *System*'s own very existence depends upon the transition from theoretical to practical philosophy, which requires the breaking-off of Jacobi's chain of 'conditions' by something unconditioned, Schelling is concerned to understand how the highest insight must be into reality as a product of the interrelation of both the 'conscious' and the 'unconscious'. Reality is not, therefore, essentially captured by a re-presentation of the objective by the subjective. Whereas in the *System* nature begins unconsciously and ends in conscious philosophical and scientific knowledge, in the art work: 'the I is conscious according to the production, unconscious with regard to the product'. The product cannot be understood via the intentions of its producer, as this would mean that it became a 'conditioned' object, something produced in terms of a pre-existing rule, and would therefore lack what makes mere craft into art. Art is, then, 'the only true and eternal organ and document of philosophy, which always and continuously documents what philosophy cannot represent externally'. The particular sciences can only follow the chain of conditions, via the principle of sufficient reason, and must determine any object via its place in that chain, a process which has no necessary end. The art object, on the other hand, manifests what cannot be understood in terms of its knowable conditions, because an account of the materials of which it is made or of its status as object in the world does not constitute it as art. Art shows what cannot be said. Philosophy cannot positively represent the absolute because 'conscious' thinking operates from the position where the 'absolute identity' of the subjective and the objective has always already been lost in the emergence of consciousness.

Although Schelling's early work did not fully satisfy either himself, or anybody else, it manages to address, in a cogent and illuminating fashion, a great deal of topics which affect subsequent philosophy. The model presented in the *System* impresses not least because, at the same time as establishing the notion of the history of self-consciousness that would be decisive for Hegel, it offers, in a manner which goes beyond its sources in Fichte, a model of the relationship between the subject and its conceptually inaccessible motivating forces which would affect significant parts of nineteenth century thought from Schopenhauer, to Nietzsche, to Freud.

3. Identity Philosophy

Although the period of Schelling's 'identity philosophy' is usually dated from the 1801 *Presentation of My System of Philosophy* until sometime before the 1809 *On the Essence of Human Freedom*, the project of that philosophy can be said to be carried on in differing ways throughout his work. The identity philosophy derives from Schelling's conviction that the self-conscious I must be seen as a result, rather than as the originating act it is in Fichte, and thus that the I cannot be seen as the generative matrix of the whole system. This takes him more in the direction of Spinoza, but the problem is still that of articulating the relationship between the I and the world of material nature, without either reverting to Kantian

dualism or failing to explain how a purely objective nature could give rise to subjectivity.

Schelling's mature identity philosophy, which is contained in the *System of the Whole of Philosophy and of Naturphilosophie in Particular*, written in Würzburg in 1804, and in other texts between 1804 and 1807, breaks with the model of truth as correspondence. It does so because:

> It is clear that in every explanation of the truth as a correspondence (*Übereinstimmung*) of subjectivity and objectivity in knowledge, both, subject and object, are already presupposed as separate, for only what is different can agree, what is not different is in itself one.

The crucial problem is how to explain the *link* between the subject and object world that makes judgements possible, and this cannot be achieved in terms of how a subject can have thoughts which correspond to an object essentially separate from it. For there to be judgements at all what is split and then synthesised in the judgement must, Schelling contends, in some way already be the same. This has often been understood as leading Schelling to a philosophy in which, as Hegel puts it in the *Phenomenology*, the absolute is the 'night in which all cows are black', because it swallows all differentiated knowledge in the assertion that everything is ultimately the same, namely an absolute which excludes all relativity from itself and thus becomes inarticulable. This is not a valid interpretation of Schelling's argument, and Hegel's remark seems, incidentally, not to have been directed against Schelling anyway.

In order to try to get over the problem in monism of how the One is also the many, Schelling, following the idea outlined above from Hölderlin, introduces a notion of 'transitive' being, which links mind and matter as predicates of itself. Schelling explains this 'transitivity' via the metaphor of the earth:

> you recognise its [the earth's] true essence only in the link by which it eternally posits its unity as the multiplicity of its things and again posits this multiplicity as its unity. You also do not imagine that, apart from this infinity of things which are in it, there is another earth which is the unity of these things, rather *the same* which is the multiplicity is also unity, and *what* the unity is, is also the multiplicity, and this necessary and indissoluble One of unity and multiplicity in it is what you call its existence (...) Existence is the link of a being (*Wesen*) as One, with itself as a multiplicity.

'Absolute identity' is, then, the *link* of the two aspects of being, which, on the one hand, is the *uni*verse, and, on the other, is the changing *multi*plicity which the knowable universe also is. Schelling insists now that 'The *I* think, *I* am, is, since Descartes, the basic mistake of all knowledge; thinking is not my thinking, and being is not my being, for everything is only of God or the totality', so the I is 'affirmed' as a predicate of the being by which it is preceded. In consequence he already begins to move away, albeit inconsistently, from the German Idealist model in which the intelligibility of being is regarded as a result of its having an essentially mind-like structure.

Schelling is led to this view by his understanding of the changing and relative status of theoretical knowledge. It is the inherent incompleteness of all finite determinations which reveals the nature of the absolute. His description of time makes clear what he means: ‘time is itself nothing but *the totality appearing in opposition to the particular life of things*’, so that the totality ‘posits or intuits itself, by not positing, not intuiting the particular’. The particular is determined in judgements, but the truth of claims about the totality cannot be proven because judgements are necessarily conditioned, whereas the totality is not. Given the relative status of the particular there must, though, be a ground which enables us to be *aware* of that relativity, and this ground must have a different status from the knowable world of finite particulars. At the same time, if the ground were wholly different from the world of relative particulars the problems of dualism would recur. As such the absolute *is* the finite, but we do not *know* this in the manner we know the finite. Without the *presupposition* of ‘absolute identity’, therefore, the evident relativity of particular knowledge becomes inexplicable, since there would be no reason to claim that a revised judgement is predicated of the same world as the preceding -- now false -- judgement.

Schelling summarises his theory of identity as follows:

> for being, actual, real being is precisely self-disclosure/revelation (*Selbstoffenbarung*). If it is to be as One then it must disclose/reveal itself in itself; but it does not disclose/reveal itself in itself if it is not an other in itself, and is *in* this other the One for itself, thus if it is not absolutely the living link of itself and an other.

The link between the ‘real’ and the ‘ideal’ cannot be regarded as a causal link. Although there cannot be mental events without physical events, the former cannot be reduced to being the causal results of the latter: ‘For real and ideal are only different views of one and the same substance’. Schelling wavers at this time between a ‘reflexive’ position of the kind which Hegel will soon try to articulate, in which, in Schelling’s terms, ‘the sameness of the subjective and the objective is made the same as itself, knows itself, and is the subject and object of itself’, in the ‘identity of identity and difference’, and the sense that this position cannot finally circumscribe the structure of the absolute. The structure of reflection, where each aspect reflects itself and then is reflected in the other, upon which this account of the identity of subject and object relies, must be grounded in a being which carries it:

> reflection (...) only knows the universal and the particular as two relative negations, the universal as relative negation of the particular, which is, as such, without reality, the particular, on the other hand, as a relative negation of the universal.(...) something independent of the concept must be added to posit the substance as such.

Without this independent basis subject and object would merely be, as Schelling thinks they are in Fichte, relative negations of each other, leading to a circle ‘inside which a nothing gains reality by the relation to another nothing’. Schelling prophetically distinguishes between the cognitive -- reflexive -- ground of finite knowledge and the real -- non-reflexive -- ground that sustains the movement of negation from one finite determination to another. As a two-sided relationship reflection alone always entails the problem that the subject and the object in a case of reflection can only be *known* to be the same via that which cannot appear in the reflection. If I am to recognise myself *as* myself in a mirror, rather than see a

random object in the world, I must *already* be familiar with myself *before* the reflection, in a way which is not part of the reflection. This means a complete system based on reflection is impossible, because, in order for the system to be grounded, it must presuppose as external to itself what it claims is part of itself. Schelling will, in his philosophy from the 1820s onwards, raise this objection against Hegel's system of 'absolute reflection'.

Schelling's own dissatisfaction with his early versions of identity theory derives from his rejection of Spinozism. Spinoza regards the move from God to the world of 'conditions' as a logical consequence of the nature of God. Schelling becomes convinced that such a theory gives no reason why the absolute, the 'unconditioned', should manifest itself in a world of negative 'conditions' at all. Schelling is therefore confronted with explaining why there is a transition from the absolute to the finite world. In *Philosophy and Religion*, of 1804, he claims, like Jacobi, that there is no way of mediating between conditioned and unconditioned, and already makes the distinction between 'negative' and 'positive' philosophy, which will form the heart of his late work. Explicating the structure of the finite world leads to 'negative philosophy, but much has already been gained by the fact that the negative, the realm of nothingness, has been separated by a sharp limit from the realm of reality and of what alone is positive'. The question which comes to concern Schelling is how philosophy can come to terms with a ground which cannot be regarded as the rational explanation of the finite world.

4. The 'Ages of the World'

Schelling's work from his middle period (1809-1827) is usually referred to as the philosophy of the *Ages of the World* (WA = *Weltalter*), after the title of the unfinished work of that name he worked on in the period 1809-1827. The work characteristic of this period begins with the 1809 *On the Essence of Human Freedom* (FS = *Freiheitsschrift*) (written in Stuttgart). The WA philosophy is an attempt to explain the emergence of an intelligible world at the same time as coming to terms with mind's inextricable relation to matter. The initial concern is to avoid Spinoza's fatalism, which renders the human freedom to do good *and* evil incomprehensible. Schelling's crucial objection is to the idea that evil should be understood as merely another form of negativity which can be comprehended by insight into the inherent lack in all finite parts of a totality, rather than as a positive fact relating to the nature of human freedom. He now sees the fundamental contradictions of the *Naturphilosophie* in terms of the relationship of the intelligibility of nature and ourselves to a ground without which there could be no intelligibility, but which is not the explicable cause of intelligibility. In an attempt to get to grips with the problem of the ground of the finite world Schelling introduces a Kant-derived conception of 'willing' in the FS which will be influential for Schopenhauer's conception of the 'Will': 'In the last and highest instance there is no other being but willing. Willing is primal being, and all the predicates of primal being only fit willing: groundlessness, eternity, being independent of time, self-affirmation'. Schelling now establishes a more conflictual version of the structure of the identity philosophy. The 'ground' is 'groundless' -- in the sense of 'uncaused' -- and it must be understood in terms of freedom if a Spinozist determinism is to be avoided. This means there cannot be an explanation of the finite world, because that would entail taking the ground as a cause and thus rendering freedom non-existent.

At the same time Schelling insists there must be that against which freedom can be manifest -- a being which is not free and is therefore necessitated -- for it to be meaningful freedom at all. The theory is based on the antagonisms between opposing forces which constitute the ‘ages of the world’, the past, present, and future. He argues that the world whose origins the WA wishes to understand must entail the *same* conflicting forces which still act, though not necessarily in the same *form*, in this world, of which the mind is an aspect: ‘Poured from the source of things and the same as the source, the human soul has a co-knowledge/con-science (*Mitwissenschaft*) of creation’. Schelling suggests that there are two principles in us: ‘an unconscious, dark principle and a conscious principle’, which must yet in some way be identical. The same structure applies to what Schelling means by ‘God’. At this point his account of the ground is not consistent, but this inconsistency points to the essential issue Schelling is trying to understand, namely whether philosophy can give a rational account of the fact of the manifest world. As that which makes the world intelligible, God relates to the ground in such a way that the ‘real’, which takes the form of material nature, is ‘in God’ but ‘is not God seen absolutely, i.e. insofar as He exists; for it is only the ground of His existence, it is *nature* in God; an essence which is inseparable from God, but different from Him’. The point is that God would be meaningless if there were not that which He transcends: without opposition, Schelling argues, there is no life and no sense of development, which are the highest aspects of reality. The aim of the move away from Spinoza is to avoid the sense of a world complete in itself which would render freedom illusory because freedom’s goal would already be determined as part of the totality. Schelling starts to confront the idea that the reconciliation of freedom and necessity that had been sought by Kant in the acknowledgement of the necessity of the law, and which was the aim of German Idealism’s attempt to reconcile mind and nature, might be intrinsically unattainable.

Wolfram Hogrebe has convincingly claimed that the WA philosophy is an ontological theory of predication. Being, as initially One and enclosed within itself, is not manifest, and has no reason to be manifest. Hogrebe terms this ‘pronominal being’. The *same* being must also, given that there is now a manifest world, be ‘predicative being’ (ibid.), which ‘flows out, spreads, gives itself’. The contradiction between the two kinds of being is only apparent. Schelling maintains, in line with the identity philosophy, that the ‘properly understood law of contradiction really only says that the same cannot be *as the same* something and also the opposite thereof, but this does not prevent the same, which is A, being able, as an other, to be not A’. One aspect of being, the dark force, which he sometimes terms ‘gravity’, is contractive, the other expansive, which he terms ‘light’. Dynamic processes are the result of the interchange between these ultimately identical forces: if they were wholly separate there would either be no manifest universe or the universe would dissipate at infinite speed. If something is to be *as* something it must both be, in the positive sense in which everything else is, which makes it indeterminately positive, pronominal, and it must have a relationship to what it is not, in order to be determinate, which brings it into the realm of predication by taking it beyond itself. In the WA the One comes into contradiction with itself and the two forces constantly vie with each other. Differences must, however, be grounded in unity, as otherwise they could not be *manifest* at all as differences. The ground is now increasingly regarded as the source of the transitory nature of everything particular, and less and less as the source of tranquil insight into how we can be reconciled to finite existence. The mood of the WA is summed up in Schelling’s reference to the ‘veil of melancholy which is spread over the whole of nature, the deep indestructible melancholy of all life’. The source of this melancholy is that everything finite must ‘go to

ground' and that we are *aware* of this.

The abandonment of his residual Spinozism leads Schelling to a growing concern with the tensions which result from contradictions that are also embodied in human beings. The ages of the world are constituted by the development of forms and structures in the material and the mental world. This development depends upon the expanding force's interaction with the contracting force's slowing of any expansion, which allows transient but determinate forms to develop. This process also gives rise to language, which Schelling regards as the model for the development of the whole world because it manifests how expansion and the release of tension can lead to intelligibility, rather than mere dissipation:

> It seems universal that every creature which cannot contain itself or draw itself together in its own fullness, draws itself together outside itself, whence e.g. the elevated miracle of the formation of the word in the mouth belongs, which is a true creation of the full inside when it can no longer remain in itself.

Language as 'contracted' material signifier, and 'expanding' ideal meaning repeats the basic structure of the WA, and Schelling insists that, like the material world without the 'ideal' capacity for expansion, language can become 'congealed'. This interaction between what is contained in itself and what draws something beyond itself is also what gives rise to consciousness, and thus to an inherent tension within consciousness, which can only be itself by its relation to an other. Hegel uses a related model of subjectivity, but Schelling will come to reject Hegel's model for its conjuring away of the ultimately irresolvable tension in all subjectivity. Schelling's later philosophy will present a subject whose origin prevents it from ever achieving the 'self-presence' Hegel thinks he can explicate via the completed structure of 'self-reflection' in the other. Schelling's WA philosophy is never completed: its Idealist aim of systematically unifying subject and object by comprehending the real development of history from the very origins of being founders on problems concerning the relationship between philosophical system and historical contingency which do not admit of solutions. Furthermore, the structures he develops lead him to ideas which take him beyond Idealism and make him one of the crucial precursors of existential and other non-Idealist forms of modern philosophy.

5. Positive and Negative Philosophy, and the Critique of Hegel

Schelling has usually been understood as providing the transitional 'objective idealist' link between Fichte and Hegel. By regarding Hegel's system as the culmination of German Idealism this interpretation fails to do justice to Schelling's real philosophical insights. Many of these insights, particularly in the later philosophy (1827-1854), directly and indirectly influenced the ideas of thinkers, like Feuerbach, Kierkegaard, Nietzsche, and Heidegger, who were critical of Hegel's claim to articulate a complete philosophical system.

The differences between Hegel and Schelling derive from their respective approaches to understanding

the absolute. For Hegel the absolute is the *result* of the self-cancellation of the finite. It can therefore be presented in the form of the successive overcoming of finite determinations, the 'negation of the negation', in a system whose end comprehends its beginning. For Hegel the result becomes known when the beginning moves from being 'in itself' to being 'for itself' at the end of the system, thus in a process in which it reflects itself to itself. Schelling already becomes publicly critical of Hegel while working on a later version of the WA philosophy in Erlangen in the 1820s, but makes his criticisms fully public in lectures given in Munich in the 1830s, and in the 1840s and 1850s as professor in Berlin. The aim of the Idealist systems was for thought to reflect what it is not -- being -- as really itself, even as it appears not to be itself, thereby avoiding Kant's dualism. The issue between Schelling and Hegel is whether the grounding of reason by itself is not in fact a sort of philosophical narcissism, in which reason admires its reflection in being without being able fully to articulate its relationship to that reflection. Like Hegel, Schelling argues that it is not the particular manifestation of knowledge which tells me the truth about the world, but rather the necessity of moving from one piece of knowledge to the next. However, a logical reconstruction of the process of knowledge can, for Schelling, only be a reflection of thought by itself. The real process cannot be described in philosophy, because the cognitive ground of knowledge and the real ground, although they are inseparable from each other, cannot be shown to *reflect* each other.

Dieter Henrich characterises Hegel's conception of the absolute as follows: 'The absolute is the finite to the extent to which the finite is nothing at all but negative relation to itself'. Hegel's system depends upon showing how each particular way of conceiving of the world has an internal contradiction. This necessarily leads thought to more comprehensive ways of grasping the world, until the point where there can be no more comprehensive way because there is no longer any contradiction to give rise to it. The very fact of the finite limitations of empirical thought therefore becomes what gives rise to the infinite, which, in Hegel's terms, is thought that is bounded by itself and by nothing else.

Schelling accepts such a conception, to which he substantially contributed in his early philosophy, as the way to construct a 'negative' system of philosophy, because it explains the logic of change, once there is a world to be explained. The conception does not, though, explain why there is a developing world at all, but merely reconstructs in thought the necessary structure of development on the basis of necessities in thought. Schelling's own attempt at explaining the world's ontological and historical facticity will lead him to a 'philosophical theology' which traces the development of mythology and then of Christian revelation in his *Philosophy of Mythology* and *Philosophy of Revelation*, which, like all his substantial works after 1811, are not published in his lifetime. The failure of his philosophical theology does not, though, necessarily invalidate his philosophical arguments against Hegel. His alternative to the 'common mistake of every philosophy that has existed up to now' -- the 'merely logical relationship of God to the world' (ibid.) -- Schelling terms 'positive philosophy'. The 'merely logical relationship' entails a reflexivity, in which the world necessarily follows from the nature of God, and God and the world are therefore the 'other of themselves'. Hegel's system tries to obviate the facticity of the world by understanding reason as the world's immanent self-articulation. Schelling, in contrast, insists that human reason cannot explain its *own* existence, and therefore cannot encompass itself and its other within a system of philosophy. We cannot, he maintains, make sense of the manifest world by beginning with reason, but must instead begin with the contingency of being and try to make sense of it with the reason which is only one aspect of it and which cannot be explained in terms of its being a reflection of the true

nature of being.

Schelling contends that the identity of thought and being cannot be articulated *within* thought, because thought must *presuppose* that they are identical in a way which thought, as one side of a relation, cannot comprehend. By redefining the ‘concept’ in such a way that it is always already both subject and object, Hegel aims to avoid any presuppositions on either the subject or the object side, allowing the system to complete itself as the ‘self-determination of the concept’. Schelling presents the basic alternative as follows:

> For either the concept would have to go first, and being would have to be the consequence of the concept, which would mean it was no longer absolute being; or the concept is the consequence of being, then we must begin with being without the concept.

Hegel attempts to merge concept and being by making being part of a structure of self-reflection, rather than the basis of the interrelation between subject and object. He invalidly assumes that ‘essence’, which is one side of the relationship between being and essence, can articulate its identity with the other side in the ‘concept’, because the other side is revealed as being ‘nothing’ until it has entered into a relationship which makes it determinate as a knowable moment of the whole process.

The problem which Hegel does not overcome is that this identity cannot be *known*, because, as Schelling claims of his concept of being, ‘existing is not here the consequence of the concept or of essence, but rather existence is here itself the concept and itself the essence’. The problem of reflection cannot be overcome in Hegel’s manner: identifying one’s reflection in a mirror as oneself (understood now as a metaphor for essence) entails, as we saw above, a prior non-reflexive moment if one is to know that the reflection *is* oneself, rather than a random reflected object. How far Schelling moves from any reflexive version of identity philosophy is evident in the following from the *Introduction to the Philosophy of Revelation or Foundation of the Positive Philosophy* of 1842-3:

> our self-consciousness is not at all the consciousness of that nature which has passed through everything, it is precisely just *our* consciousness (...) for the consciousness of man is not = the consciousness of nature (...) Far from man and his activity making the world comprehensible, man himself is that which is most incomprehensible.

Schelling refuses to allow that reason can confirm its status via its reflection in being:

> what we call the world, which is *so completely contingent* both as a whole and in its parts, cannot possibly be the impression of something which has arisen by the *necessity of reason* (...) it contains a *preponderant* mass of *unreason*.

Schelling is, then, one of the first philosophers seriously to begin the destruction of the model of metaphysics based on the idea of representation, a destruction which can be seen as one of the key aspects of modern philosophy from Heidegger to the later Wittgenstein and beyond. He is, at the same

time, unlike some of his successors, committed to an account of human reason which does not assume that reason's incapacity to ground itself should lead to the Nietzschean abandonment of rationality. This is one of the respects in which Schelling has again become part of contemporary debate, where the need to seek means of legitimation which do not rely on the notion of a rationality inherent in the world remains a major challenge. Schelling's account of mind and world, particularly his insistence on the need not to limit our conception of nature to what is accessible to objectifying forms, is, in the light of the ecological crisis, proving to be more durable than his reception might until recently have suggested.

Bibliography

Editions of Schelling

- *Friedrich Wilhelm Joseph Schelling's Sämmtliche Werke*, ed. K.F.A. Schelling, I Abtheilung Vols. 1-10, II Abtheilung Vols. 1-4, Stuttgart: Cotta, 1856-61. An easily accessible substantial selection of the complete works has been published, ed. M. Frank, as *Friedrich Wilhelm Joseph von Schelling, Ausgewählte Schriften*, 6 Vols., Frankfurt: Suhrkamp 1985.
- *Die Weltalter*, ed. M. Schröter, Munich: Biederstein 1946, which has other versions than the version from 1813 printed in the *Sämmtliche Werke*.
- The *Historisch-kritische Ausgabe, im Auftrag der Schelling-Kommission der Bayerischen Akademie der Wissenschaften*, edited by H. M. Baumgartner, W.G. Jacobs, H. Krings, Stuttgart 1976- is still a long way from completion, but will become the new standard edition.
- *Über die Möglichkeit einer Form der Philosophie überhaupt* (1794) (On the Possibility of an Absolute Form of Philosophy), *Vom Ich als Prinzip der Philosophie oder über das Unbedingte im menschlichen Wissen* (1795) (Of the I as the Principle of Philosophy or on the Unconditional in Human Knowledge), *Philosophische Briefe über Dogmatismus und Kriticismus* (1795) (Philosophical Letters on Dogmatism and Criticism) in *The Unconditional in Human Knowledge*: Four early essays 1794-6 (1980) translation and commentary by F. Marti, Lewisburg: Bucknell University Press.
- *Abhandlungen zur Erläuterung des Idealismus der Wissenschaftslehre* (1796-7) (Essays in Explanation of the Idealism of the Doctrine of Science).
- *Ideen zu einer Philosophie der Natur als Einleitung in das Studium dieser Wissenschaft* (1797) *Ideas for a Philosophy of Nature: as Introduction to the Study of this Science* (1988) translated by E.E. Harris and P. Heath, introduction R. Stern, Cambridge: Cambridge University Press.
- *Erster Entwurf eines Systems der Naturphilosophie* (1799) (First Plan of a System of the Philosophy of Nature).
- *System des transcendentalen Idealismus* (1800) *System of Transcendental Idealism* (1978) translated by P. Heath, introduction M. Vater, Charlottesville: University Press of Virginia.
- *Über den wahren Begriff der Naturphilosophie und die richtige Art, ihre Probleme zu lösen* (1801) (On the True Concept of the Philosophy of Nature and the Right Way to Solve its Problems).
- *Darstellung meines Systems der Philosophie* (1801) (Presentation of My System of Philosophy).
- *Fernere Darstellungen aus dem System der Philosophie* (1802) (Further Presentations from the

System of Philosophy).

- *Bruno oder über das göttliche und natürliche Prinzip der Dinge* (1802) *Bruno, or On the Natural and the Divine Principle of Things* (1984) translated with an introduction by M. Vater, Albany: State University of New York Press.
- *Philosophie der Kunst* (1802-3) *The Philosophy of Art* (1989) Minnesota: Minnesota University Press.
- *Vorlesungen über die Methode des akademischen Studiums* (1803) *On University Studies* (1966) translated E.S. Morgan, edited N. Guterman, Athens, Ohio: Ohio University Press.
- *Philosophie und Religion* (1804) (Philosophy and Religion).
- *System der gesammten Philosophie und der Naturphilosophie insbesondere* (1804) (System of the Whole of Philosophy and the Philosophy of Nature in Particular).
- *Aphorismen zur Einleitung in die Naturphilosophie* (1806) (Aphorisms as an Introduction to the Philosophy of Nature).
- *Aphorismen über die Naturphilosophie* (1806) (Aphorisms on the Philosophy of Nature).
- *Über das Verhältnis der bildenden Künste zur Natur* (1807) (On the Relationship of the Fine Arts to Nature).
- *Philosophische Untersuchungen über das Wesen der menschlichen Freiheit und die damit zusammenhängenden Gegenstände* (1809) *Of Human Freedom* (1936) a translation with critical introduction and notes by J. Gutmann, Chicago: Open Court.
- *Briefwechsel mit Eschenmayer* (1810) (Correspondence with Eschenmayer).
- *Stuttgarter Privatvorlesungen* (1810) (Stuttgart Private Lectures).
- *Die Weltalter* (1811-15). *The Ages of the World* (1967) translated with introduction and notes by F. de W. Bolman, jr., New York: Columbia University Press. *The Abyss of Freedom/Ages of the World* (1997), trans. Judith Norman, with an essay by Slavoj Zizek, Anne Arbor: The University of Michigan Press
- *Über die Gottheiten von Samothrake* (1815) *Schelling's Treatise on 'The Deities of Samothrace'* (1977) a translation and introduction by R.F. Brown, Missoula, Mont.: Scholars Press.
- *Initia Philosophiae Universae* (1820-1), (1969) ed. H. Fuhrmans, Bonn: Bouvier.
- *Über die Nature der Philosophie als Wissenschaft* (1821) (On the Nature of Philosophy as a Science).
- *System der Weltalter* (1827-8) (System of the Ages of the World) (1990) ed. S. Peetz, Frankfurt: Klostermann.
- *Einleitung in die Philosophie* (1830) (Introduction to Philosophy) (1989) ed. W. E. Ehrhardt (Schellingiana 11), Stuttgart: Frommann-Holzboog.
- *Grundlegung der positiven Philosophie* (1832-3) (Foundations of the Positive Philosophy) (1972) ed. H. Fuhrmans Turin: Bottega d'Erasmo.
- *Zur Geschichte der neueren Philosophie* (probably 1833-4) *On the History of Modern Philosophy* (1994) translation and introduction by A. Bowie, Cambridge: Cambridge University Press.
- *Philosophie der Offenbarung* (1841-2) (Philosophy of Revelation) (1977) ed. M. Frank, Frankfurt: Suhrkamp.
- *Philosophie der Mythologie* (1842) (Philosophy of Mythology).
- *Philosophie der Offenbarung* (1842-3) (Philosophy of Revelation).
- *Philosophische Einleitung in die Philosophie der Mythologie oder Darstellung der reinrationalen*

Philosophie (Between 1847 and 1852) (Philosophical Introduction to the Philosophy of Mythology or Presentation of the Purely Rational Philosophy).

References and Further Reading

- Beach, Edward A. (1994) *The Potencies of the God(s): Schelling's Philosophy of Mythology*, Albany: SUNY Press (Account of the late philosophy.)
- Bowie, A. (1990) *Aesthetics and Subjectivity: from Kant to Nietzsche*, Manchester: Manchester University Press, reprinted 1993, completely revised edition 2002. (Chapter on Schelling which characterises him in relation to Hölderlin and to Romantic and post-Romantic theories of aesthetics, and as a theorist of subjectivity who does not rely on the idea of self-presence).
- ------ (1993) *Schelling and Modern European Philosophy: An Introduction*, London: Routledge. (The first full-length account of Schelling in English to consider him as a major philosopher in his own right, rather than as a pendant to Hegel. Connects Schelling to issues in contemporary analytical and European philosophy).
- Fichte, J.G. (1971) *Werke* I, Berlin: de Gruyter. (See § 1).
- Frank, M. (1975) *Der unendliche Mangel an Sein*, Frankfurt: Suhrkamp. (The classic modern account of Schelling's critique of Hegel: a dense and very difficult, but indispensable work).
- ------ (1985) *Eine Einführung in Schellings Philosophie*, Frankfurt: Suhrkamp. (A detailed account of Schelling's early work until the end of the identity philosophy: see §2).
- ------ (1991) *Selbstbewußtsein und Selbsterkenntnis,* Stuttgart: Reclam. (Contains a vital essay on Schelling's identity theory, 'Identität und Subjektivität', which sees the theory as a major event in Western philosophy).
- ------ed. (1975a) with Kurz, G., *Materialien zu Schellings philosophischen Anfängen*, Frankfurt: Suhrkamp. (Essays on various aspects of Schelling's philosophy between 1795 and 1804, with accompanying historical material).
- Heidegger, M. (1971) *Schellings Abhandlung über das Wesen der menschlichen Freiheit*, Tübingen: Niemeyer. (Dense and difficult, but essential commentary on Schelling's *On the Essence of Human Freedom*, with material from later lectures by Heidegger. See §3).
- ------ (1991) *Die Metaphysik des deutschen Idealismus (Schelling)*, Frankfurt: Klostermann. (After the positive account in Heidegger (1971) the claim here is that Schelling is, after all, another example of the 'Western metaphysics' which culminates in Nietzsche's 'will to power'. Difficult and clearly flawed, because it ignores the late work altogether).
- Henrich, D. (1982) *Selbstverhältnisse*, Stuttgart: Reclam. (Important essays on Schelling, Hegel and modern philosophy).
- Heuser-Kessler, M.-L. (1986) '*Die Produktivität der Natur', Schellings Naturphilosophie und das neue Paradigma der Selbstorganisation in den Naturwissenschaften*, Berlin: de Gruyter. (Claims that Schelling's philosophy of nature can be linked to developments in non-linear dynamics and to the theory of self-organising systems).
- Hogrebe, W. (1989) *Prädikation und Genesis. Metaphysik als Fundamentalheuristik im Ausgang von Schellings 'Die Weltalter'*, Frankfurt: Suhrkamp. (A brilliant, but demanding account of the WA as a theory of predication, which uses the tools of analytical philosophy to show how consistent much of Schelling's position is).

- Jähnig, D. (1966, 1969) *Schelling. Die Kunst in der Philosophie* Two Vols. Pfullingen: Neske. (Detailed and impressive account of the importance of art for Schelling's philosophy as a whole).
- Jaspers, K. (1955) *Schelling: Größe und Verhängnis*, Munich: Piper. (An interesting, if outdated, account of Schelling's life and work, which sees Schelling as failing to achieve his philosophical goals).
- Marx, W. (1984) *The Philosophy of F.W.J. Schelling: History, System, Freedom*, Bloomington: Indiana University Press. (General and fairly accessible account, mainly of earlier work by Schelling, as far as *On the Essence of Human Freedom*).
- Sandkaulen-Bock, B. (1990) *Ausgang vom Unbedingten. Über den Anfang in der Philosophie Schellings*, Göttingen: Vandenhoeck and Ruprecht. (Excellent account of Schelling's response to questions posed in particular by Jacobi concerning the grounding of philosophy in the absolute: historically detailed and very thorough on the early work).
- Sandkühler, H. J. (1970) *Friedrich Wilhelm Joseph Schelling,* Stuttgart: Metzler. (Contains bibliography, which compliments that of Schneeberger -- see below).
- ------ ed. (1984) *Natur und geschichtlicher Prozeß*, Frankfurt: Suhrkamp. (Selection of essays on the philosophy of nature with useful bibliography of writings on that philosophy).
- Schneeberger, G. (1954) *Friedrich Wilhelm Joseph von Schelling. Eine Bibliographie*, Bern: Franke. (The standard bibliography, to be complimented by those cited above).
- Scholz, H., ed. (1916) *Die Hauptschriften zum Pantheismusstreit zwischen Jacobi und Mendelssohn*, Berlin: Reuther and Reichard. (Contains most of the key texts by Jacobi in the Pantheism controversy).
- Schulz, W. (1975) *Die Vollendung des deutschen Idealismus in der Spätphilosophie Schellings*, Pfullingen: Neske. (The book which reoriented the study of Schelling after World War 2 towards the study of the later work, particularly the Hegel-critique, and linked Schelling to Kierkegaard and Heidegger. Difficult but thought-provoking).
- Snow, Dale E. (1996) *Schelling and the End of Idealism*, Albany: SUNY Press. (Excellent, very lucid, account of the early and middle Schelling in particular.)
- Tilliette, X. (1970) *Schelling une philosophie en devenir*, Two Volumes, Paris: Vrin. (Encyclopedic historical account of the development of Schelling's work: stronger on general exposition and on theology than on Schelling's philosophical arguments).
- White, A. (1983a) *Absolute Knowledge: Hegel and the Problem of Metaphysics,* Ohio: Ohio University Press. (Defends Hegel against Schelling's critique, but does not take account of the arguments of Frank on the failure of reflection in Hegel).
- ------ (1983b) *Schelling: Introduction to the System of Freedom*, New Haven and London: Yale University Press. (Good introduction to Schelling's work as a whole, which tends to focus, though, on its undoubted weaknesses, at the expense of its strengths).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Fichte, Johann Gottlieb | Hegel, Georg Wilhelm Friedrich | Heidegger, Martin | Jacobi, Friedrich Heinrich | Kant, Immanuel | Kierkegaard, Søren | Leibniz, Gottfried Wilhelm | Nietzsche, Friedrich | Novalis [Friedrich Leopold, Baron von Hardenberg] | Schleiermacher, Friedrich Daniel | Spinoza, Baruch [Benedict]

Andrew Bowie
a.bowie@rhul.ac.uk

First published: October 22, 2001
Content last modified: October 22, 2001

Søren Kierkegaard

Søren Aabye Kierkegaard (b.1813, d. 1855) was a profound and prolific writer in the Danish "golden age" of intellectual and artistic activity. His work crosses the boundaries of philosophy, theology, psychology, literary criticism, devotional literature and fiction. Kierkegaard brought this potent mixture of discourses to bear as social critique and for the purpose of renewing Christian faith within Christendom. At the same time he made many original conceptual contributions to each of the disciplines he employed. He is known as the "father of existentialism", but at least as important are his critiques of Hegel and of the German romantics, his contributions to the development of modernism, his literary experimentation, his vivid re-presentation of biblical figures to bring out their modern relevance, his invention of key concepts which have been explored and redeployed by thinkers ever since, his interventions in contemporary Danish church politics, and his fervent attempts to analyse and revitalise Christian faith. Kierkegaard burned with the passion of a religious poet, was armed with extraordinary dialectical talent, and drew on vast resources of erudition.

Kierkegaard's Life

Kierkegaard led a somewhat uneventful life. He rarely left his hometown of Copenhagen, and travelled abroad only three times -- to Berlin. His prime recreational activities were attending the theatre, walking the streets of Copenhagen to chat with ordinary people, and taking brief carriage jaunts into the surrounding countryside. He was educated at a prestigious boys' school (*Borgedydskolen*), then attended

Copenhagen University where he studied philosophy and theology. His teachers at the university included F.C. Sibbern, Poul Martin Møller, and H.L. Martensen.

Sibbern and Møller were both philosophers who also wrote fiction. The latter in particular had a great influence on Kierkegaard's philosophico-literary development. Martensen also had a profound effect on Kierkegaard, but largely in a negative manner. Martensen was a champion of Hegelianism, and when he became Bishop Primate of the Danish People's Church, Kierkegaard published a vitriolic attack on Martensen's theological views. Kierkegaard's brother Peter, on the other hand, was an adherent of Martensen and himself became a bishop in the church.

Another very important figure in Kierkegaard's life was J.L. Heiberg, the doyen of Copenhagen's literati. Heiberg, more than any other person, was responsible for introducing Hegelianism into Denmark. Kierkegaard spent a good deal of energy trying to break into the Heiberg literary circle, but desisted once he had found his own voice in *The Concept of Irony*. Kierkegaard's first major publication, *From the Papers of One Still Living*, is largely an attempt to articulate a Heibergian aesthetics - which is a modified version of Hegel's aesthetics. In *From the Papers of One Still Living*, which is a critical review of Hans Christian Andersen's novel *Only A Fiddler*, Kierkegaard attacks Andersen for lacking life-development (*Livs-Udvikling*) and a life-view (*Livs-Anskuelse*) both of which Kierkegaard deemed necessary for someone to be a genuine novelist (*Romandigter*).

Kierkegaard's life is more relevant to his work than is the case for many writers. Much of the thrust of his critique of Hegelianism is that its system of thought is abstracted from the everyday lives of its proponents. This existential critique consists in demonstrating how the life and work of a philosopher contradict one another. Kierkegaard derived this form of critique from the Greek notion of judging philosophers by their lives rather than simply by their intellectual artefacts. The Christian ideal, according to Kierkegaard, is even more exacting since the totality of an individual's existence is the artefact on the basis of which s/he is judged by God for h/er eternal validity. Of course a writer's work is an important part of h/er existence, but for the purpose of judgement we should focus on the whole life not just on one part.

In a less abstract manner, an understanding of Kierkegaard's biography is important for an understanding of his writing because his life was the source of many of the preoccupations and repetitions within his *oeuvre*. Because of his existentialist orientation, most of his interventions in contemporary theory do double duty as means of working through events from his own life. In particular Kierkegaard's relations to his mother, his father, and his fiancée Regine Olsen pervade his work.

Kierkegaard's relation to his mother is the least frequently commented upon since it is invisible in his work. His mother does not rate a direct mention in his published works, or in his diaries -- not even on the day she died. However, for a writer who places so much emphasis on indirect communication, and on the semiotics of invisibility, we should regard this absence as significant. Johannes Climacus in *Concluding Unscientific Postscript* remarks, "... how deceptive then, that an omnipresent being should be recognisable precisely by being invisible." Kierkegaard's mother, who was not well educated, is

represented in his writings by the mother-tongue (Danish). Kierkegaard was deeply enamoured of the Danish language and worked throughout his writings to assert the strengths of his mother-tongue over the invasive, imperialistic influences of Latin and German. With respect to the former, Kierkegaard had to petition the king to be allowed to write his philosophy dissertation *On the Concept of Irony with constant reference to Socrates* in Danish. Even though permission was granted he was still required to defend his dissertation publicly in Latin. Latin had been the pan-European language of science and scholarship. In Denmark, in Kierkegaard's time, German language and culture were at least as dominant as Latin in the production of knowledge. In defiance of this, Kierkegaard revelled in his mother-tongue and created some of the most beautifully poetic prose in the Danish language -- including a paeon to his mother-tongue in *Stages On Life's Way*. In *Repetition* Constantin Constantius congratulates the Danish language on providing the word for an important new philosophical concept, viz. *Gjentagelse* (repetition), to replace the foreign word "mediation". In general, the Danish language is Kierkegaard's umbilical attachment to the mother whereas Latin and German represent the law of the father, especially when employed in systematic scholarship (*Videnskab*).

The influence of Kierkegaard's father on his work has been frequently noted. Not only did Kierkegaard inherit his father's melancholy, his sense of guilt and anxiety, and his pietistic emphasis on the dour aspects of Christian faith, but he also inherited his talents for philosophical argument and creative imagination. In addition Kierkegaard inherited enough of his father's wealth to allow him to pursue his life as a freelance writer. The themes of sacrificial father/son relationships, of inherited sin, of the burden of history, and of the centrality of the "individual, human existence relationship, the old text, well known, handed down from the fathers" (*Postscript*) are repeated many times in Kierkegaard's oeuvre. The father's sense of guilt was so great (for having cursed God? for having impregnated Kierkegaard's mother out of wedlock?) that he thought God would punish him by taking the lives of all seven of his children before they reached the age of 34 (the age of Jesus Christ at his crucifixion). This was born out for all but two of the children, Søren and his older brother Peter, both of whom were astonished to survive beyond that age. This may explain the sense of urgency that drove Kierkegaard to write so prolifically in the years leading up to his 34th birthday.

Kierkegaard's (broken) engagement to Regine Olsen has also been the focus of much scholarly attention. The theme of a young woman being the occasion for a young man to become "poeticized" recurs in Kierkegaard's writings, as does the theme of the sacrifice of worldly happiness for a higher (religious) purpose. Kierkegaard's infatuation with Regine, and the sublimated libidinal energy it lent to his poetic production, were crucial for setting his life course. The breaking of the engagement allowed Kierkegaard to devote himself monastically to his religious purpose, as well as to establish his outsider status (outside the norm of married bourgeois life). It also freed him from close personal entanglements with women, thereby leading him to objectify them as ideal creatures, and to reproduce the patriarchal values of his church and father.

Kierkegaard's Rhetoric

Kierkegaard's central problematic was *how to become a Christian in Christendom*. The task was most

difficult for the well-educated, since prevailing educational and cultural institutions tended to produce stereotyped members of "the crowd" rather than to allow individuals to discover their own unique identities. This problem was compounded by the fact that Denmark had recently and very rapidly been transformed from a feudal society into a capitalist society. Universal elementary education, large-scale migration from rural areas into cities, and greatly increased social mobility meant that the social structure changed from a rigidly hierarchical one to a relatively "horizontal" one. In this context it became increasingly difficult to "become who you are" for two reasons: (i) social identities were unusually fluid; and (ii) there was a proliferation of normalizing institutions which produced pseudo-individuals.

Given this problematic in this social context Kierkegaard perceived a need to invent a form of communication which would not produce stereotyped identities. On the contrary, he needed a form of rhetoric which would force people back onto their own resources, to take responsibility for their own existential choices, and to become who they are beyond their socially imposed identities. In this undertaking Kierkegaard was inspired by the figure of Socrates, whose incessant irony undermined all knowledge claims that were taken for granted or unreflectively inherited from traditional culture. In his dissertation *On the Concept of Irony with constant reference to Socrates* Kierkegaard argued that the historical Socrates used his irony in order to facilitate the birth of subjectivity in his interlocutors. Because they were constantly forced to abandon their pat answers to Socrates' annoying questions, they had to begin to think for themselves and to take individual responsibility for their claims about knowledge and value.

Kierkegaard sought to provide a similar service for his own contemporaries. He used irony, parody, satire, humor, and deconstructive techniques in order to make conventionally accepted forms of knowledge and value untenable. He was a gadfly -- constantly irritating his contemporaries with discomforting thoughts. He was also a midwife -- assisting at the birth of individual subjectivity by forcing his contemporaries to think for themselves. His art of communication became "the art of *taking away*" since he thought his audience suffered from too much knowledge rather than too little.

Hegelianism promised to make absolute knowledge available by virtue of a science of logic. Anyone with the capacity to follow the dialectical progression of the purportedly transparent concepts of Hegel's logic would have access to the mind of God (which for Hegel was equivalent to the logical structure of the universe). Kierkegaard thought this to be the hubristic attempt to build a new tower of Babel, or a *scala paradisi* -- a dialectical ladder by which humans can climb with ease up to heaven. Kierkegaard's strategy was to invert this dialectic by seeking to make everything more difficult. Instead of seeing scientific knowledge as the means of human redemption, he regarded it as the greatest obstacle to redemption. Instead of seeking to give people more knowledge he sought to take away what passed for knowledge. Instead of seeking to make God and Christian faith perfectly intelligible he sought to emphasize the absolute transcendence by God of all human categories. Instead of setting himself up as a religious authority, Kierkegaard used a vast array of textual devices to undermine his authority as an author and to place responsibility for the existential significance to be derived from his texts squarely on the reader.

Kierkegaard distanced himself from his texts by a variety of devices which served to problematize the

authorial voice for the reader. He used pseudonyms in many of his works (both overtly aesthetic ones and overtly religious ones). He partitioned the texts into prefaces, forewords, interludes, postscripts, appendices. He assigned the "authorship" of parts of texts to different pseudonyms, and invented further pseudonyms to be the editors or compilers of these pseudonymous writings. Sometimes Kierkegaard appended his name as author, sometimes as the person responsible for publication, sometimes not at all. Sometimes Kierkegaard would publish more than one book on the same day. These simultaneous books embodied strikingly contrasting perspectives. He also published whole *series* of works simultaneously, viz. the pseudonymous works on the one hand and on the other hand the *Edifying Discourses* published under his own name.

All of this play with narrative point of view, with contrasting works, and with contrasting internal partitions within individual works leaves the reader very disoriented. In combination with the incessant play of irony and Kierkegaard's predilection for paradox and semantic opacity, the text becomes a polished surface for the reader in which the prime meaning to be discerned is the reader's own reflection. Christian faith, for Kierkegaard, is not a matter of learning dogma by rote. It is a matter of the individual repeatedly renewing h/er passionate subjective relationship to an object which can never be known, but only believed in. This belief is offensive to reason, since it only exists in the face of the absurd (the paradox of the eternal, immortal, infinite God being incarnated in time as a finite mortal).

Kierkegaard's "method of indirect communication" was designed to sever the reliance of the reader on the authority of the author and on the received wisdom of the community. The reader was to be forced to take individual responsibility for knowing who s/he is and for knowing where s/he stands on the existential, ethical and religious issues raised in the texts.

Kierkegaard's "inverted Christian dialectic" was designed not to make the word of God easier to assimilate, but to establish more clearly the absolute distance that separates human beings from God. This was in order to emphasize that human beings are absolutely reliant on God's grace for salvation.

Kierkegaard's Aesthetics

Kierkegaard presents his pseudonymous authorship as a dialectical progression of existential stages. The first is the aesthetic, which gives way to the ethical, which gives way to the religious. The aesthetic stage of existence is characterized by the following: immersion in sensuous experience; valorization of possibility over actuality; egotism; fragmentation of the subject of experience; nihilistic wielding of irony and scepticism; and flight from boredom.

The figure of the aesthete in the first volume of *Either-Or* is an ironic portrayal of German romanticism, but it also draws on medieval characters as diverse as Don Juan, Ahasverus (the wandering Jew), and Faust. It finds its most sophisticated form in the author of "The Seducer's Diary", the final section of *Either-Or*. Johannes the seducer is a *reflective aesthete*, who gains sensuous delight not so much from the act of seduction but from engineering the possibility of seduction. His real aim is the manipulation of people and situations in ways which generate interesting reflections in his own voyeuristic mind. The

aesthetic perspective transforms quotidian dullness into a richly poetic world by whatever means it can. Sometimes the reflective aesthete will inject interest into a book by reading only the last third, or into a conversation by provoking a bore into an apoplectic fit so that he can see a bead of sweat form between the bore's eyes and run down his nose. That is, the aesthete uses artifice, arbitrariness, irony, and wilful imagination to recreate the world in his own image. The prime motivation for the aesthete is the transformation of the boring into the interesting.

This type of aestheticism is criticized from the point of view of ethics. It is seen to be emptily self-serving and escapist. It is a despairing means of avoiding commitment and responsibility. It fails to acknowledge one's social debt and communal existence. And it is self-deceiving insofar as it substitutes fantasies for actual states of affairs.

But Kierkegaard did not want to abandon aesthetics altogether in favor of the ethical and the religious. A key concept in the Hegelian dialectic, which Kierkegaard's pseudonymous authorship parodies, is *Aufhebung* (sublation). In Hegel's dialectic, when contradictory positions are reconciled in a higher unity (synthesis) they are both annulled and preserved (*aufgehoben*). Similarly with Kierkegaard's pseudo-dialectic: the aesthetic and the ethical are both annulled and preserved in their synthesis in the religious stage. As far as the aesthetic stage of existence is concerned what is preserved in the higher religious stage is the sense of infinite possibility made available through the imagination. But this no longer excludes what is actual. Nor is it employed for egotistic ends. Aesthetic irony is transformed into religious humor, and the aesthetic *transfiguration* of the actual world into the ideal is transformed into the religious *transubstantiation* of the finite world into an actual reconciliation with the infinite.

But the dialectic of the pseudonymous authorship never quite reaches the truly religious. We stop short at the *representation* of the religious by a self-confessed humorist (Johannes Climacus) in a medium which, according to Climacus's own account, necessarily alienates the reader from true (Christian) faith. For faith is a matter of lived experience, of constant striving within an individual's existence. According to Climacus's metaphysics, the world is divided dualistically into the actual and the ideal. Language (and all other media of representation) belong to the realm of the ideal. No matter how eloquent or evocative language is it can never *be* the actual. Therefore, any representation of faith is always suspended in the realm of ideality and can never *be* actual faith.

So the whole dialectic of the pseudonymous authorship is recuperated by the aesthetic by virtue of its medium of representation. In fact Johannes Climacus acknowledges this implicitly when at the end of *Concluding Unscientific Postscript* he *revokes* everything he has said, with the important rider that to say something then to revoke it is not the same as never having said it in the first place. His presentation of religious faith in an aesthetic medium at least provides an opportunity for his readers to make their own leap of faith, by appropriating with inward passion the paradoxical religion of Christianity into their own lives.

As a poet of the religious Kierkegaard was always preoccupied with aesthetics. In fact, contrary to popular misconceptions of Kierkegaard which represent him as becoming increasingly hostile to poetry,

he referred increasingly to himself as a poet in his later years (all but one of over ninety references to himself as a poet in his journals date from after 1847). Kierkegaard never claimed to write with religious authority, as an apostle. His works represent both less religiously enlightened and more religiously enlightened positions than he thought he had attained in his own existence. Such representations were only possible in an aesthetic medium of imagined possibilities like poetry.

Kierkegaard's Ethics

Like the terms "aesthetic" and "religious", the term "ethics" in Kierkegaard's work has more than one meaning. It is used to denote both: (i) a limited existential sphere, or stage, which is superseded by the higher stage of the religious life; and (ii) an aspect of life which is retained even within the religious life. In the first sense "ethics" is synonymous with the Hegelian notion of *Sittlichkeit*, or customary mores. In this sense "ethics" represents "the universal", or more accurately the prevailing social norms. The social norms are seen to be the highest court of appeal for judging human affairs -- nothing outranks them for this sort of ethicist. Even human sacrifice is justified in terms of how it serves the community, so that when Agamemnon sacrifices his daughter Iphigenia he is regarded as a tragic hero since the sacrifice is required for the success of the Greek expedition to Troy (*Fear and Trembling*).

Kierkegaard, however, does recognize duties to a power higher than social norms. Much of *Fear and Trembling* turns on the notion that Abraham's would-be sacrifice of his son Isaac is not for the sake of social norms, but is the result of a "teleological suspension of the ethical". That is, Abraham recognizes a duty to something higher than both his social duty not to kill an innocent person and his personal commitment to his beloved son, viz. his duty to obey God's commands.

But in order to arrive at a position of religious faith, which might entail a "teleological suspension of the ethical", the individual must first embrace the ethical (in the first sense). In order to raise oneself beyond the merely aesthetic life, which is a life of drifting in imagination, possibility and sensation, one needs to make a commitment. That is, the aesthete needs to choose the ethical, which entails a commitment to communication and decision procedures.

The ethical position advocated by Judge Wilhelm in "Equilibrium Between the Aesthetic and the Ethical in the Composition of Personality" (*Either-Or* II) is a peculiar mix of cognitivism and noncognitivism. The metaethics or normative ethics are cognitivist, laying down various necessary conditions for ethically correct action. These conditions include: the necessity of choosing seriously and inwardly; commitment to the belief that predications of good and evil of our actions have a truth-value; the necessity of choosing what one is actually doing, rather than just responding to a situation; actions are to be in accordance with rules; and these rules are universally applicable to moral agents.

The choice of metaethics, however, is noncognitive. There is no adequate proof of the truth of metaethics. The choice of normative ethics is motivated, but in a noncognitive way. The Judge seeks to motivate the choice of his normative ethics through the avoidance of despair. Here despair (*Fortvivlelse*) is to let one's life depend on conditions outside one's control (and later, more radically, despair is the very

possibility of despair in this first sense). For Judge Wilhelm, the choice of normative ethics is a noncognitive choice of cognitivism, and thereby an acceptance of the applicability of the conceptual distinction between good and evil.

From Kierkegaard's religious perspective, however, the conceptual distinction between good and evil is ultimately dependent not on social norms but on God. Therefore it is possible, as Johannes de Silentio argues was the case for Abraham (the father of faith), that God demand a suspension of the ethical (in the sense of the socially prescribed norms). This is still ethical in the second sense, since ultimately God's definition of the distinction between good and evil outranks any human society's definition. The requirement of communicability and clear decision procedures can also be suspended by God's fiat. This renders cases such as Abraham's extremely problematic, since we have no recourse to public reason to decide whether he is legitimately obeying God's command or whether he is a deluded would-be murderer. Since public reason cannot decide the issue for us, we must decide for ourselves as a matter of religious faith.

Kierkegaard's Religion

Kierkegaard styled himself above all as a religious poet. The religion to which he sought to relate his readers is Christianity. The type of Christianity that underlies his writings is a very serious strain of Lutheran pietism informed by the dour values of sin, guilt, suffering, and individual responsibility. Kierkegaard was immersed in these values in the family home through his father, whose own childhood was lived in the shadow of the severe *Indre Mission* (Inner Mission) -- a pietistic cult from Jutland. Kierkegaard's father subsequently became a member of the Moravian Brethren congregation in Copenhagen.

For Kierkegaard Christian faith is not a matter of regurgitating church dogma. It is a matter of individual subjective passion, which cannot be mediated by the clergy or by human artefacts. Faith is the most important task to be achieved by a human being, because only on the basis of faith does an individual have a chance to become a true self. This self is the life-work which God judges for eternity.

The individual is thereby subject to an enormous burden of responsibility, for upon h/er existential choices hangs h/er eternal salvation or damnation. Anxiety or dread (*Angest*) is the presentiment of this terrible responsibility when the individual stands at the threshold of momentous existential choice. Anxiety is a two-sided emotion: on one side is the dread burden of choosing for eternity; on the other side is the exhilaration of freedom in choosing oneself. Choice occurs in the instant (*Øjeblikket*), which is the point at which time and eternity intersect -- for the individual creates through temporal choice a self which will be judged for eternity.

But the choice of faith is not made once and for all. It is essential that faith be constantly renewed by means of repeated avowals of faith. One's very selfhood depends upon this repetition, for according to Anti-Climacus, the self "is a relation which relates itself to itself" (*The Sickness Unto Death*). But unless this self acknowledges a "power which constituted it," it falls into a despair which undoes its selfhood.

Therefore, in order to maintain itself as a relation which relates itself to itself, the self must constantly renew its faith in "the power which posited it." There is no *mediation* between the individual self and God by priest or by logical system (*contra* Catholicism and Hegelianism respectively). There is only the individual's own *repetition* of faith. This repetition of faith is the way the self relates itself to itself and to the power which constituted it, i.e. the repetition of faith *is* the self.

Christian dogma, according to Kierkegaard, embodies paradoxes which are offensive to reason. The central paradox is the assertion that the eternal, infinite, transcendent God simultaneously became incarnated as a temporal, finite, human being (Jesus). There are two possible attitudes we can adopt to this assertion, viz. we can have faith, or we can take offense. What we cannot do, according to Kierkegaard, is believe by virtue of reason. If we choose faith we must suspend our reason in order to believe in something higher than reason. In fact we must believe *by virtue of the absurd*.

Much of Kierkegaard's authorship explores the notion of the absurd: Job gets everything back again by virtue of the absurd (*Repetition*); Abraham gets a reprieve from having to sacrifice Isaac, by virtue of the absurd (*Fear and Trembling*); Kierkegaard hoped to get Regine back again after breaking off their engagement, by virtue of the absurd (*Journals*); Climacus hopes to deceive readers into the truth of Christianity by virtue of an absurd representation of Christianity's ineffability; the Christian God is represented as absolutely transcendent of human categories yet is absurdly presented as a personal God with the human capacities to love, judge, forgive, teach, etc. Kierkegaard's notion of the absurd subsequently became an important category for twentieth century existentialists, though usually devoid of its religious associations.

According to Johannes Climacus, faith is a miracle, a gift from God whereby eternal truth enters time in the instant. This Christian conception of the relation between (eternal) truth and time is distinct from the Socratic notion that (eternal) truth is always already within us -- it just needs to be recovered by means of recollection (*anamnesis*). The condition for realizing (eternal) truth for the Christian is a gift (*Gave*) from God, but its realization is a task (*Opgave*) which must be repeatedly performed by the individual believer. Whereas Socratic recollection is a recuperation of the past, Christian repetition is a "recollection forwards" -- so that the eternal (future) truth is captured in time.

Crucial to the miracle of Christian faith is the realization that over against God we are always in the wrong. That is, we must realize that we are always in sin. This is the condition for faith, and must be given by God. The idea of sin cannot evolve from purely human origins. Rather, it must have been introduced into the world from a transcendent source. Once we understand that we are in sin, we can understand that there is some being over against which we are always in the wrong. On this basis we can have faith that, by virtue of the absurd, we can ultimately be atoned with this being.

Kierkegaard's Politics

Kierkegaard is sometimes regarded as an apolitical thinker, but in fact he intervened stridently in church politics, cultural politics, and in the turbulent social changes of his time. His earliest published essay, for

example, was a polemic against women's liberation. It is a reactionary apologetic for the prevailing patriarchal values, and was motivated largely by Kierkegaard's desire to ingratiate himself with factions within Copenhagen's intellectual circles. This latter desire gradually left him, but his relation to women remained highly questionable.

One of Kierkegaard's main interventions in cultural politics was his sustained attack on Hegelianism. Hegel's philosophy had been introduced into Denmark with religious zeal by J.L. Heiberg, and was taken up enthusiastically within the theology faculty of Copenhagen University and by Copenhagen's literati. Kierkegaard, too, was induced to make a serious study of Hegel's work. While Kierkegaard greatly admired Hegel, he had grave reservations about Hegelianism and its bombastic promises. Hegel would have been the greatest thinker who ever lived, said Kierkegaard, if only he had regarded his system as a thought-experiment. Instead he took himself seriously to have reached the truth, and so rendered himself comical.

Kierkegaard's tactic in undermining Hegelianism was to produce an elaborate parody of Hegel's entire system. The pseudonymous authorship, from *Either-Or* to *Concluding Unscientific Postscript*, presents an inverted Hegelian dialectic which is designed to lead readers away from knowledge rather than towards it. This authorship simultaneously snipes at German romanticism and contemporary Danish literati (with J.L. Heiberg receiving much acerbic comment).

This intriguing pseudonymous authorship received little popular attention, aimed as it was at the literary elite. So it had little immediate effect as discursive action. Kierkegaard sought to remedy this by provoking an attack on himself in the popular satirical review *The Corsair*. Kierkegaard succeeded in having himself mercilessly lampooned in this publication, largely on personal grounds rather than in terms of the substance of his writings. The suffering incurred by these attacks sparked Kierkegaard into another highly productive phase of authorship, but this time his focus was the creation of positive Christian discourses rather than satire or parody.

Eventually Kierkegaard became more and more worried about the direction taken by the Danish People's Church, especially after the death of the Bishop Primate J.P. Mynster. He realized he could no longer indulge himself in the painstakingly erudite and poetically meticulous writing he had practised hitherto. He had to intervene decisively in a popular medium, so he published his own pamphlet under the title *The Instant*. This addressed church politics directly and increasingly shrilly.

There were two main foci of Kierkegaard's concern in church politics. One was the influence of Hegel, largely through the teachings of H.L. Martensen; the other was the popularity of N.F.S. Grundtvig, a theologian, educator and poet who composed most of the pieces in the Danish hymn book. Grundtvig's theology was diametrically opposed to Kierkegaard's in tone. Grundtvig emphasized the light, joyous, celebratory and communal aspects of Christianity, whereas Kierkegaard emphasized seriousness, suffering, sin, guilt, and individual isolation. Kierkegaard's intervention failed miserably with respect to the Danish People's Church, which became predominantly Grundtvigian. His intervention with respect to Hegelianism also failed, with Martensen succeeding Mynster as Bishop Primate. Hegelianism in the

church went on to die of natural causes.

Kierkegaard also provided critical commentary on social change. He was an untiring champion of "the single individual" as opposed to "the crowd". He feared that the opportunity of achieving geniune selfhood was diminished by the social production of stereotypes. He lived in an age when mass society was emerging from a highly stratified feudal order and was contemptuous of the mediocrity the new social order generated. One symptom of the change was that mass society substitutes detached reflection for engaged passionate commitment. Yet the latter is crucial for Christian faith and for authentic selfhood according to Kierkegaard.

Kierkegaard's real value as a social and political thinker was not realized until after his death. His pamphleteering achieved little immediate impact, but his substantial philosophical, literary, psychological and theological writings have had a lasting effect. Much of Heidegger's very influential work, *Being And Time*, is indebted to Kierkegaard's writings (though this goes unacknowledged by Heidegger). Kierkegaard's social realism, his deep psychological and philosophical analyses of contemporary problems, and his concern to address "the present age" were taken up by fellow Scandinavians Henrik Ibsen and August Strindberg. Ibsen and Strindberg, together with Friedrich Nietzsche, became central icons of the modernism movement in Berlin in the 1890s. The Danish literary critic Georg Brandes was instrumental in conjoining these intellectual figures: he had given the first university lectures on Kierkegaard and on Nietzsche; he had promoted Kierkegaard's work to Nietzsche and to Strindberg; and he had put Strindberg in correspondence with Nietzsche. Taking his cue from Brandes, the Swedish literary critic Ola Hansson subsequently promoted this conjunction of writers in Berlin itself. Berlin modernism self-consciously sought to use art as a means of political and social change. It continued Kierkegaard's concern to use discursive action for social transformation.

Chronology of Kierkegaard's Life and Works

1813	born May 5 in Copenhagen (Denmark)
1830	matriculated to the university of Copenhagen
1834	mother died
1837	met Regine Olsen
1838	father died
-	*From the Papers of One Still Living. Published against his Will by S. Kierkegaard (Af en endnu Levendes Papirer -- Udgivet mod hans Villie af S. Kierkegaard)*
1840	passed final theological examination
-	proposed to Regine Olsen, who accepted him
1841	broke off his engagement to Regine Olsen
-	defended his dissertation *On the Concept of Irony with constant reference to Socrates (Om Begrebet Ironi med stadigt Hensyn til Socrates)*

- trip to Berlin, where he attended lectures by Schelling

1842 returned from Berlin

1843 *Either-Or: A Fragment of Life edited by Victor Eremita (Enten-Eller. Et Livs-Fragment, udgivet af Victor Eremita)*

- second trip to Berlin
- *Two Edifying Discourses by S. Kierkegaard (To opbyggelige Taler)*
- *Fear and Trembling: A Dialectical Lyric by Johannes de Silentio (Frygt og Bæven. Dialektisk Lyrik af Johannes de Silentio)*
- *Repetition: A Venture in Experimenting Psychology by Constantin Constantius (Gjentagelsen. Et Forsøg i den experimenterende Psychologi af Constantin Constantius)* (published the same day as *Fear and Trembling*)
- *Three Edifying Discourses by S. Kierkegaard (Tre opbyggelige Taler)*
- *Four Edifying Discourses by S. Kierkegaard (Fire opbyggelige Taler)*

1844 *Two Edifying Discourses by S. Kierkegaard (To opbyggelige Taler)*

- *Three Edifying Discourses by S. Kierkegaard (Tre opbyggelige Taler)*
- *Philosophical Fragments or a Fragment of Philosophy by Johannes Climacus, published by S. Kierkegaard (Philosophiske Smuler eller En Smule Philosophie. Af Johannes Climacus. Udgivet af S. Kierkegaard)*
- *The Concept of Anxiety: A Simple Psychologically-Oriented Reflection on the Dogmatic Problem of Original Sin by Vigilius Haufniensis (Begrebet Angest. En simpel psychologisk-paapegende Overveielse i Retning of det dogmatiske Problem om Arvesynden af Vigilius Haufniensis)*
- *Prefaces: Light Reading for Certain Classes as the Occasion may Require by Nicolaus Notabene (Forord. Morskabslæsning for enkelte Stænder efter Tid og Lejlighed, af Nicolaus Notabene)* (published on the same day as *The Concept of Anxiety*)
- *Four Edifying Discourses by S. Kierkegaard (Fire opbyggelige Taler)*

1845 *Three Addresses on Imagined Occasions by S. Kierkegaard (Tre Taler ved tænkte Leiligheder)*

- *Stages On Life's Way: Studies by Various Persons, compiled, forwarded to the press, and published by Hilarious Bookbinder (Stadier paa Livets Vej. Studier af Forskjellige. Sammenbragte, befordrede til Trykken og udgivne af Hilarius Bogbinder)*
- third trip to Berlin
- *Eighteen Edifying Discourses by S. Kierkegaard* (a collection of the remaindered *Edifying Discourses* from 1843 and 1844)
- in an article in *Fædrelandet* Frater Taciturnus (a character from *Stages on Life's Way*) asked to be lambasted in *The Corsair*

1846 Kierkegaard lampooned in *The Corsair*

- *Concluding Unscientific Postscript to Philosophical Fragments: A Mimetic-Pathetic-Dialectic Compilation, An Existential Plea, by Johannes Climacus, published by S. Kierkegaard (Afsluttende uvidenskabelig Efterskrift til de philosophiske Smuler. -- Mimisk-pathetisk-dialektisk Sammenskrift, Existentielt Indlæg, af Johannes Climacus. Udgiven af S. Kierkegaard)*
- *A Literary Review: "Two Ages" -- novella by the author of "An Everyday Story" -- reviewed by S. Kierkegaard (En literair Anmeldelse af S. Kierkegaard)*

1847 *Edifying Discourses in Different Spirits by S. Kierkegaard (Opbyggelige Taler i forskjellig Aand af S. Kierkegaard)*

- *Works of Love: Some Christian Reflections in the Form of Discourses by S. Kierkegaard (Kjerlighedens Gjerninger. Nogle christelige Overveielser i Talers Form, af S. Kierkegaard)*
- Regine marries Fritz Schlegel

1848 *Christian Discourses by S. Kierkegaard (Christelige Taler, af S. Kierkegaard)*

- *The Crisis and a Crisis in the Life of an Actress by Inter et Inter (Krisen og en Krise i en Skuespillerindes Liv af Inter et Inter)*
- *The Point of View for my Work as an Author: A Direct Communication, A Report to History (Synspunktet for min Forfatter-Virksomhed. En ligefrem Meddelelse, Rapport til Historien, af S. Kierkegaard)* (unpublished)

1849 second edition of *Either-Or*

- *The Lilies of the Field and the Birds of the Air: Three devotional discourses by S. Kierkegaard (Lilien paa Marken og Fuglen under Himlen. Tre gudelige Taler af S. Kierkegaard)*
- *Two Ethico-Religious Treatises by H.H. (Tvende ethisk-religieuse Smaa-Afhandlinger. Af H.H.)*
- *The Sickness Unto Death: A Christian psychological exposition for edification and awakening by Anti-Climacus, edited by S. Kierkegaard (Sygdommen til Døden. En christelig psychologisk Udvikling til Opvækkelse. Af Anticlimacus. Udgivet af S. Kierkegaard)*
- *"The High Priest" -- "The Publican" -- and "The Woman taken in Sin": three addresses at Holy Communion on Fridays by S. Kierkegaard („Ypperstepræsten" -- „Tolderen" -- „Synderinden", tre Taler ved Altergangen om Fredagen. Af S. Kierkegaard)*

1850 *Training in Christianity by Anti-Climacus, Nos. I, II, III, edited by S. Kierkegaard (Indøvelse i Christendom. Af Anti-Climacus -- Udgivet af S. Kierkegaard)*

- *An Edifying Discourse by S. Kierkegaard (En opbyggelig Tale. Af S. Kierkegaard)*

1851 *On My Activity As A Writer by S. Kierkegaard (Om min Forfatter-Virksomhed. Af S. Kierkegaard)*

- *Two Discourses at Holy Communion on Fridays by S. Kierkegaard (To Taler ved Altergangen om Fredagen)*
- *For Self-Examination: Recommended to the Contemporary Age by S. Kierkegaard (Til Selvprøvelse, Samtiden anbefalet. Af S. Kierkegaard)*
- *Judge For Yourselves! Recommended to the present time for Self-Examination. Second series, by S. Kierkegaard (Dømmer Selv! Til Selvprøvelse Samtiden anbefalet. Anden Række, af S. Kierkegaard)* (published posthumously 1876)

1854 Bishop Mynster died

- Martensen appointed bishop
- "Was Bishop Mynster ‘a witness to the truth,’ one of ‘the true witnesses to the truth’ -- *is this the truth?*" by S. Kierkegaard in *Fædrelandet* („Var Biskop Mynster et "Sandhedsvidne", et af "de rette Sandhedsvidner", *er dette Sandhed?"* Af S. Kierkegaard) (the first of 21 articles in *Fædrelandet*)

1855 *This Must Be Said, So Let It Be Said, by S. Kierkegaard (Dette skal siges; saa være det da sagt. Af S. Kierkegaard)*

- *The Instant by S. Kierkegaard (Øjeblikket. Af S. Kierkegaard)*
- *Christ's Judgement on Official Christianity by S. Kierkegaard (Hvad Christus dømmer om officiel Christendom. Af S. Kierkegaard)*
- *God's Unchangeability: A Discourse by S. Kierkegaard (Guds Uforanderlighed. En Tale -- Af S. Kierkegaard)*
- Kierkegaard died November 11.

Bibliography

- Adorno, Theordor W., *Kierkegaard: Construction of the Aesthetic*,trans. Robert Hullot-Kentor, Minneapolis: University of Minnesota Press, 1989.
- Agacinski, Sylviane, *Aparté: Conceptions and Deaths of Søren Kierkegaard*, trans. Kevin Newmark, Tallahassee: Florida State University Press, 1988.
- Bigelow, Pat, *Kierkegaard & The Problem Of Writing*, Tallahassee: Florida State University Press, 1987.
- Billeskov Jansen, F.J., *Studier i Søren Kierkegaards litterære Kunst*, Copenhagen: Rosenkilde & Bagger, 1951.
- Bloom, Harold (ed.), *Søren Kierkegaard*, New York: Chelsea House Publishers, 1989.
- Brandes, Georg, *Søren Kierkegaard. En kritisk Fremstilling i Grundrids*, Copenhagen: Gyldendal, 1877.
- Derrida, Jacques, *The Gift of Death*, trans. David Wills, Chicago & London: University of Chicago Press, 1995.
- Diderichsen, Adam, *Den Sårede Odysseus: Kierkegaard og subjektivitetens genese*, København: Hans Reitzels Forlag, 1998.
- Dooley, Mark, *The Politics of Exodus: Kierkegaard's ethics of responsibility*, New York: Fordham University Press, 2001.
- Ferguson, Harvie, *Melancholy and the Critique of Modernity: Søren Kierkegaard's Religious Psychology*, London & New York: Routledge, 1995.
- Ferreira, M. Jamie, *Transforming Vision: Imagination and Will in Kierkegaardian Faith*, Oxford: Clarendon Press, 1991.
- Ferreira, M. Jamie, *Love's Grateful Striving: A Commentary on Kierkegaard's Works of Love*, Oxford University Press, 2001.

- Garff, Joachim, *Den søvnløse: Kierkegaard læst æstetisk/biografisk*, Copenhagen: C.A.Reitzels Forlag, 1995.
- Garff, Joachim, *SAK: Søren Aabye Kierkegaard: en biografi*, København: Gad, 2000.
- Hall, Ronald L., *Word and Spirit: A Kierkegaardian Critique of the Modern Age*, Bloomington: Indiana State University Press, 1993.
- Hannay, Alastair, *Kierkegaard*, London: Routledge & Kegan Paul, 1982.
- Henriksen, Aage, *Kierkegaards Romaner*, Copenhagen: Gyldendal, 1954.
- Houe, Poul, Gordon D. Marino, & Sven Hakon Rossel (eds), *Anthropology and Authority: Essays on Søren Kierkegaard* , Amsterdam-Atlanta, GA: Editions Rodopi B.V., 2000.
- Kirmmse, Bruce H., *Kierkegaard in Golden Age Denmark*, Bloomington: Indiana State University Press, 1990.
- Kirmmse, Bruce H.(ed.), *Encounters with Kierkegaard: A Life as Seen by His Contemporaries* , trans. Bruce H. Kirmmse &n Virginia R. Laursen, Princeton, N.J.: Princeton University Press, 1996.
- Lippitt, John, *Humour and Irony in Kierkegaard's Thought* , London: Macmillan & New York: St. Martin's Press, 2000.
- Lowrie, Walter, *Kierkegaard*, 2 volumes, New York: Harper & Brothers, 1962.
- Mackey, Louis, *Points of View: Readings of Kierkegaard*, Tallahassee: Florida State University Press, 1986.
- Malantschuk, Gregor, *Frihed og Eksistens. Studier i Søren Kierkegaards tænkning*, Copenhagen: C.A.Reitzels Forlag, 1980.
- Malik, Habib C., *Receiving Søren Kierkegaard: The Early Impact and Transmission of His Thought* , Washington, D.C.: The Catholic University of America Press, 1997.
- Matustík, Martin J., *Postnational Identity: Critical Theory and Existential Philosophy in Habermas, Kierkegaard, and Havel*, New York & London: The Guilford Press, 1993.
- Nordentoft, Kresten, *Kierkegaard's Psychology*, trans. Bruce H. Kirmmse, Pittsburgh, Pa.: Duquesne University Press, 1978.
- Pattison, George, *Kierkegaard: The Aesthetic and the Religious*, London: Macmillan, 1992.
- Pattison, George, & Stephen Shakespeare (eds), *Kierkegaard: the self in society*, New York: St. Martin's Press, 1998.
- Pojman, Louis, *The Logic of Subjectivity*, University of Alabama Press, 1984.
- Rée, Jonathan, & Jane Chamberlain (eds), *Kierkegaard: A Critical Reader* , Oxford: Blackwell, 1998.
- Roos, Carl, *Kierkegaard og Goethe*, Copenhagen: Gads Forlag, 1955.
- Rudd, Anthony, *Kierkegaard and the Limits of the Ethical* , Oxford: Clarendon Press, 1993.
- Schleifer, Ronald, & Robert Markley(eds), *Kierkegaard and Literature: Irony, Repetition, and Criticism*, Norman: University of Oklahoma Press, 1984.
- Scopetea, Sophia, *Kierkegaard og græciteten: en kamp med ironi* , Kœbenhavn: C.A. Reitzel, 1995.
- Taylor, Mark C., *Journeys to Selfhood: Hegel & Kierkegaard*, Berkeley, Los Angeles, London: University of California Press, 1980.
- Viallaneix, Nelly, *Écoute, Kierkegaard: Essai sur la communication de la parole*, 2 volumes, Paris: Éditions du Cerf, 1979.

- Walsh, Sylvia, *Living Poetically: Kierkegaard's Existential Aesthetics*, University Park, Pennsylvania: Pennsylvania State University Press, 1994.
- Watkin, Julia, *Historical Dictionary of Kierkegaard's Philosophy* , Lanham, Maryland & London: The Scarecrow Press, 2001.
- Weston, Michael, *Kierkegaard and Modern Continental Philosophy*, London: Routledge, 1994.
- Westphal, Merold & Martin J. Matustik, *Kierkegaard in Post/Modernity*, Bloomington & Indianapolis: Indiana University Press, 1995.

Other Internet Resources

- Kierkegaard on the Internet
- Søren Kierkegaard Forskningscenteret
- International Kierkegaard Information
- The Howard V. Hong and Edna H. Hong Kierkegaard Library
- Royal Library Denmark Kierkegaard Manuscripts

Related Entries

aesthetics | Hegel, Georg Wilhelm Friedrich | individual | personal identity | Socrates

William McDonald
wmcdonal@metz.une.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 3, 1996
Content last modified: November 28, 2001

Friedrich Daniel Ernst Schleiermacher

Friedrich Daniel Ernst Schleiermacher (1768-1834) probably cannot be ranked as one of the greatest German philosophers of the eighteenth and nineteenth centuries (like Kant, Herder, Hegel, Marx, or Nietzsche). But he is certainly one of the most interesting of the second-tier philosophers of the period. Nor was he only a philosopher; he was also an eminent classicist and theologian. Much of his philosophical work was in the philosophy of religion, but from a modern philosophical point of view it is probably his hermeneutics (i.e. theory of interpretation) and his theory of translation that deserve the most attention. This article will attempt to provide a fairly broad overview of his philosophical thought. One thing which will emerge when this is done is that although he has important philosophical debts to many predecessors and contemporaries (including Spinoza, Kant, Friedrich Schlegel, and Schelling), he was above all following in the philosophical footsteps of one predecessor in particular: Herder.

1. Life and Works

Friedrich Daniel Ernst Schleiermacher (1768-1834) was born 1768 in Breslau as son of a reformed clergyman. His earlier education took place in institutions of the Moravian Brethren (Herrnhuter), a strict pietist sect, where he also pursued broader humanistic interests however. Largely as a result of skepticism

about certain Christian doctrines taught there, in 1787 he moved to the more liberal University of Halle. However, he continued in theology (with philosophy and classical philology as minor fields). He passed his theological examinations in Berlin in 1790. This was followed by a period as a private tutor, which ended in 1793, partly it seems due to friction caused by his sympathy with the French Revolution (to which his employer was opposed).

During the periods just mentioned he was heavily occupied with the study and criticism of Kant's philosophy. This work culminated in several unpublished essays -- *On the Highest Good* (1789), *On What Gives Value to Life* (1792-3), and *On Freedom* (1790-3) -- which rejected Kant's conception of the "*summum bonum* [highest good]" as requiring an apportioning of happiness to moral desert, rejected Kant's connected doctrine of the "postulates" of an afterlife of the soul and God, and developed an anti-Kantian theory of the thoroughgoing causal determination of human action but of the compatibility of this with moral responsibility. In 1793-4 he wrote two essays about Spinoza: *Spinozism* and *Brief Presentation of the Spinozistic System*. The main catalyst of these essays was Jacobi's 1785 work *On the Doctrine of Spinoza, in Letters to Mr. Moses Mendelssohn*, which was highly critical of Spinozism, but they also show the influence of Herder's 1787 work *God. Some Conversations*, which championed a modified form of Spinozism. In his two essays Schleiermacher himself embraces a modified form of Spinozist monism similar in character to Herder's (in particular, like Herder, he inclines to substitute for Spinoza's single substance the more active principle of a single fundamental force). He also attempts to defend this position by showing it to be reconcilable with central features of Kant's theoretical philosophy (notably, Kant's doctrine of things in themselves). This neo-Spinozistic position would subsequently be fundamental to Schleiermacher's most important work in the philosophy of religion, *On Religion: Speeches to Its Cultured Despisers* (1799). However, in thus rejecting Jacobi's anti-Spinozism, Schleiermacher seems also to have absorbed something from Jacobi which would be no less important for his future philosophy of religion: the idea (for which his pietist background no doubt made him receptive) that we enjoy a sort of immediate intuition or feeling of God.

During the period 1794-6 Schleiermacher served as a pastor in Landsberg. In 1796 he moved to Berlin, where he became chaplain to a hospital. In Berlin he met Friedrich and August Wilhelm Schlegel and other romantics, became deeply engaged in the romantic movement, and collaborated with the Schlegel brothers on the short-lived but important literary journal *Athenaeum* (1798-1800). Among Schleiermacher's contributions to this journal was a short proto-feminist piece *Idea for a Catechism of Reason for Noble Ladies*. During 1797-9 he shared a house with Friedrich Schlegel. Encouraged by the romantic circle to write a statement of his religious views, in 1799 he published his most important and radical work in the philosophy of religion, *On Religion: Speeches to its Cultured Despisers* (revised editions: 1806, 1821, the latter including significant "explanations," and 1831). This work sought to save religion in the eyes of its cultured despisers (prominent among them some of the romantics) by, inter alia, arguing that human immortality and even God are inessential to religion, diagnosing current religion's more off-putting features in terms of its corruption by worldly bourgeois culture and state-interference, and arguing that there are an endless multiplicity of valid forms of religion. The book won Schleiermacher a national reputation. In the same year (1799) he also published an essay on the situation of the Jews in Prussia, *Letters on the Occasion of the Political-Theological Task and the Open Letter of Jewish Householders*. In this work he rejected a proposed expedient of effecting the Jews' civil

assimilation through baptism (which would, he argues, harm both Judaism and Christianity) and instead advocated full civil rights for them (on certain reasonable conditions). The same year also saw Schleiermacher's composition of the interesting short essay *Toward a Theory of Sociable Conduct*, which is important as his first significant discussion of the art of conversation (an art which would later be central to his dialectics lectures). Finally, 1799 also saw his publication of a highly critical review of Kant's *Anthropology*. This review in particular takes Kant to task for his dualistic philosophy of mind, and his superficial, disparaging attitude towards women and other peoples.

During the following several years Schleiermacher complemented *On Religion* with two substantial publications which were more ethical in orientation: the especially important *Soliloquies* (1800; second edition 1810) and the *Outlines of a Critique of Previous Ethical Theory* (1803). In 1800 he also defended his friend Friedrich Schlegel's controversial and (it is widely agreed) pornographic novel *Lucinde* of the same year in his *Confidential Letters Concerning Friedrich Schlegel's Lucinde* -- a shared proto-feminism constituting a large part of his reason for sympathy with Schlegel's book. During the period 1799-1804 Schleiermacher developed with Schlegel the project of translating Plato's dialogues. As time went on, however, Schlegel left this work to Schleiermacher (which contributed to increasingly difficult relations between the two men after 1800). Schleiermacher's translations appeared during the period 1804-28 (though not all of the dialogues were translated in the end), and are still widely used and admired today.

While in Berlin Schleiermacher developed romantic attachments to two married women, Henriette Herz and Eleonore Grunow -- the latter of which attachments led to scandal and unhappiness, eventually encouraging Schleiermacher to leave the city. He spent the years 1802-4 in Stolpe. By 1804 he was teaching at Halle University. In the period 1804-5 he began lecturing on ethics (as he would do repeatedly until 1832). In 1805 he also began his famous and important lectures on hermeneutics (which he delivered repeatedly until 1833). In 1806 he published the short book *Christmas Eve*, a literary work which explores the meaning of Christian love by depicting a German family's celebration of Christmas Eve (in keeping with *On Religion*'s ideal of (Christian) religion as family- rather than state-centered). In 1806-7 he left Halle as a result of the French occupation, and moved to Berlin. From this time on he began actively promoting German resistance to the French occupation, and the cause of German unity. In 1808 he married Henriette von Willich (a young widow), with whom he had several children. In 1808-9 he became preacher at the Dreifaltigkeitskirche, in 1810 professor of theology at the University of Berlin, and by 1811 he was also a member of the Berlin Academy of Sciences.

After becoming a member of the Academy he often delivered addresses before it, among which several on ethics and one from 1831 on Leibniz's idea of a universal language are especially significant. In 1811 he lectured on dialectics for the first time (as he would do repeatedly until his death, at which time he was in the early stages of preparing a version for publication). In 1813 he published the shortish essay *On the Different Methods of Translation* -- a very important work in translation theory deeply informed by his experience as a translator. In 1818 he lectured on psychology for the first time (as he would do repeatedly until 1833-4). In 1819 he lectured on aesthetics for the first time (as he subsequently did on two further occasions, the last of them in 1832-3). In the same year he also began lecturing on the life of Jesus (as he did again on four further occasions over the following twelve years), thereby inaugurating an important genre of literature on this subject in the nineteenth century. In 1821-2 he published his major work of

systematic theology, *The Christian Faith* (revised edition 1830-1). In 1829 he published two open letters on this work (nominally addressed to his friend Lücke), in which he discusses it and central issues in the philosophy of religion and theology relating to it in a concise and lucid way. He died in 1834.

As can be seen even from this brief sketch of his life and works, a large proportion of Schleiermacher's career was taken up with the philosophy of religion and theology. However, from the secular standpoint of modern philosophy it is his work in such areas as hermeneutics (i.e. the theory of interpretation) and the theory of translation that is more interesting. Accordingly, this article will begin with these more interesting areas of his thought, only turning briefly to his philosophy of religion at the end.

2. Philosophy of Language

Since the topics of language and psychology are central to Schleiermacher's hermeneutics and theory of translation, it may be appropriate to begin with some discussion of his philosophies of language and mind. Schleiermacher nowhere presents his philosophy of language separately; instead, it is found scattered through such works as his lectures on psychology, dialectics, and hermeneutics. The following positions -- all but the last of which are heavily indebted to Herder -- are especially worth noting:

> (1) In his psychology lectures, Schleiermacher takes the following stance on the question of the origin of language (virtually identical to Herder's stance in his *Treatise on the Origin of Language* (1772)): Language's origin is not to be explained in terms of a divine source. Nor is it to be explained in terms of the primitive expression of feelings. Rather, the use of inner language is simply fundamental to human nature. It is the foundation of, and indeed identical with, thought; and it is also the foundation of other distinctively human mental features, in particular self-consciousness and a clear distinguishing of perception from both feeling and desire.
>
> (2) Language (and hence thought) are fundamentally social in nature. More precisely, although inner language is not dependent on a social stimulus (so that even in the absence of that children would develop their own languages), it does already involve a tendency or implicit directedness towards social communication.
>
> (3) Language and thought are not mere additions on top of other mental processes which human beings share with the animals. Rather, they are infused throughout, and lend a distinctive character to, all human mental processes. In particular, they structure human beings' sensory images in distinctive ways.

The next five positions are especially important for Schleiermacher's hermeneutics and theory of translation (to be discussed below):

> (4) Schleiermacher in his early work postulates an identity of thought with linguistic expression. He often equates thought more specifically with *inner* language (he already

does so in his 1812-13 ethics lectures). His main motive behind such a refinement can be seen from the lectures on psychology, which discuss cases in which thought occurs but without arriving at any outward linguistic expression. In later work he seems to retreat somewhat from such identity claims, though in his psychology lectures of 1830 we still find him writing of "the activity of thought in its identity with language."

(5) Schleiermacher adopts a view of meaning which equates it -- not with such items, in principle independent of language, as the referents involved, Platonic forms, or the mentalistic “ideas” favored by the British empiricists and others -- but with word usages, or rules for the use of words. For example, in the hermeneutics lectures he says that “the … meaning of a term is to be derived from the unity of the word-sphere and from the rules governing the presupposition of this unity.”

(6) In his psychology lectures, Schleiermacher argues that thought and conceptualization are not *reducible* to the occurrence of sensuous images (since that would conflict with his position that they require or are identical with language), but he also argues that the latter are an essential *foundation* for the former. This position is also reflected in his strong attraction in some of his later hermeneutics and dialectics lectures to Kant's theory of empirical schemata -- according to which empirical concepts are grounded, or consist, in unconscious rules for the generation of sensuous images -- and to turning it into an account of the nature of all concepts. (This invites the question whether there do not also exist strictly a priori concepts. In his psychology lectures Schleiermacher vacillates in his answer to this question: sometimes implying so, but at other points instead implying -- more consistently with the position just described -- that it is merely that some concepts are *more* distantly abstracted from sensory images than others. The latter is his normal answer in his dialectics lectures as well.)

(7) Human beings exhibit, not only significant linguistic and conceptual-intellectual similarities, but also striking linguistic and conceptual-intellectual diversities, especially between different historical periods and cultures, but even to some extent between individuals within a single period and culture. (Schleiermacher argues, plausibly, that the phenomenon of the linguistic and conceptual-intellectual development of cultures over time is only explicable in terms of linguistic and conceptual-intellectual innovations performed by individuals, which get taken over by the broader culture, becoming part of its common stock.)

(8) Schleiermacher, importantly, develops a much more *holistic* conception of meaning than was yet found in his predecessors (this is the one major respect in which his philosophy of language goes beyond Herder's). At least three aspects of his semantic holism can be distinguished: (a) (As can be seen from a passage quoted above,) he espouses a doctrine of “the unity of the word-sphere.” This doctrine in effect says that the various specific senses which a single word will typically bear and which will normally be distinguished by any good dictionary entry (e.g. the different senses of “impression” in “He

made an impression in the clay," "My impression is that he is reluctant," and "He made a big impression at the party") always form a larger semantic unity to which they each essentially belong (so that any loss, addition, or alteration among them entails an alteration in each of them, albeit possibly a subtle one). (b) He holds that the nature of any particular concept is partly defined by its relations to a "system of concepts." In this connection, the dialectics lectures emphasize a concept's relations as a species-concept to superordinate genus-concepts, relations as a genus-concept to subordinate species-concepts, and relations of contrast to coordinate concepts falling under the same genus-concepts. However, other types of conceptual relationships would be included here as well (e.g. those between "to work," "worker," and "a work"). (c) He holds that the distinctive nature of a language's *grammatical* system (e.g. its system of declensions) is also partly constitutive of the character of the concepts expressed within it. (This last position was also developed at about the same time by Friedrich Schlegel, who has a strong claim to be considered the real founder of modern linguistics, and for whom it constituted one of the main rationales for a new discipline of "comparative grammar" (see his *On the Language and Wisdom of the Indians* (1808)). It was shortly afterwards taken over and used to similar effect by another of the founders of modern linguistics, Wilhelm von Humboldt.)

As was mentioned earlier, with the sole exception of this final feature (semantic holism), this entire Schleiermacherian philosophy of language is heavily indebted to Herder's. However, it arguably also weakens Herder's in certain respects. For example, whereas Herder's version of doctrine (4) normally restricted itself to a claim that thought is essentially dependent on and bounded by thought, Schleiermacher, as we saw, turns it into a doctrine of the outright *identity* of thought with language, or with inner language. But such a strong version of the doctrine seems philosophically problematic -- vulnerable to counterexamples in which thought occurs without any corresponding (inner) language use, and vice versa. Again, as we saw, in later works Schleiermacher tends to add to the Herderian doctrine (5) a thesis that concepts are empirical schemata à la Kant (see (6)). What is problematic about this is arguably not, as it might seem to be, the inclusion of sensory images in meaning per se; Herder had included them as well, doing so is probably reconcilable with a (suitably understood) doctrine of meanings as word usages, and the currently popular Fregean-Wittgensteinian attack on such "psychologism" is probably misguided. What is problematic about it is rather that Kant's theory of empirical schemata had implied a sharp distinction between meanings, conceived as something purely psychological, and word usages, so that Schleiermacher's unmodified reintroduction of the theory implies the same (and hence conflicts with doctrine (5), the doctrine that meaning is word usage). Again, whereas for Herder doctrine (7) was merely an empirically established rule of thumb and admitted of exceptions, Schleiermacher in his ethics and dialectics lectures attempts to give a sort of *a priori proof* of linguistic and conceptual-intellectual diversity even at the individual level as a *universal* fact -- a proof which is not only dubious in itself (both in its (quasi-)a priori status and in its specific details), but also implies the extremely counterintuitive consequence (often explicitly asserted by Schleiermacher) that, strictly speaking, no one can ever understand another person.

3. Philosophy of Mind

Schleiermacher's philosophy of mind is found mainly in his lectures on psychology. It is too extensive to be presented in detail here. But the following four central principles -- all of which have their roots in Herder, and especially in Herder's main work in the philosophy of mind, *On the Cognition and Sensation of the Human Soul* (1778) - are especially striking and important:

> (a) Schleiermacher argues for a strong dependence of the soul (or mind) on the body, and indeed for their identity. However, he resists reductionism in either direction, arguing that both what he calls “spiritualism” (i.e. the reduction of the body to the mind) and “materialism” (i.e. the reduction of the mind to the body) are errors. He refers to the sort of non-reductive unity of mind and body that he instead champions as “life.”
>
> (b) Schleiermacher also identifies the soul (or mind) with ”force.” Thus already in *On Freedom* (1790-3) he writes that the soul is “a force or a composite of forces.”
>
> (c) Schleiermacher argues strongly for the unity of the soul (or mind) within itself; the soul is not composed of separate faculties (e.g. sensation, understanding, imagination, reason, desire). (He himself often works with a twofold distinction between what he refers to as the mind's “organic” (i.e. sensory) and ”intellectual” functions, but he holds these too to be at bottom identical.)
>
> (d) Schleiermacher argues that human minds, while they certainly share similarities, are also deeply different from each other -- not only across social groupings such as peoples and genders, but also at the level of individuals who belong to the same groupings. The deep distinctiveness of individual minds periodically exercises an important influence on the development of society at large -- both in the political-ethical sphere (where Schleiermacher calls the individuals in question “heroes”) and in the sphere of thought and art (where he calls them “geniuses”). The distinctiveness of individual minds cannot be explained by any process of calculation (in particular, it is a mistake to suppose that all minds begin the same and that they only come to differ due to the impact of different causal influences on their development, which might in principle be calculated). It can, however, be understood by means of “divination” (on which more anon).

Finally, one feature of Schleiermacher's philosophy of mind which *distinguishes* it from that of Herder and other predecessors is also worth noting: Schleiermacher says relatively little about *unconscious* mental processes, and when he does mention them often seems skeptical about them. For example, he argues that thought cannot be unconscious, and that so-called “obscure representations” are in fact merely sensuous images which do not involve thoughts.

4. Hermeneutics (i.e. Theory of Interpretation)

Some of Schleiermacher's most important philosophical work concerns the theories of interpretation (“hermeneutics”) and translation. Friedrich Schlegel was an immediate influence on his thought here.

Their ideas on these subjects began to take shape in the late 1790s, when they lived together in the same house for a time. Many of their ideas are shared, and it is often unclear which of the two men was the (more) original source of a given idea. But since Schlegel's surviving treatments are far less detailed and systematic than Schleiermacher's, the latter take on prime importance.

Schleiermacher's theories of interpretation and translation rest squarely on three of the Herder-inspired doctrines in the philosophy of language which were described earlier: (4) thought is essentially dependent on and bounded by, or even identical with, language; (5) meaning is word usage; and (7) there are deep linguistic and conceptual-intellectual differences between people. Doctrine (7) poses a severe challenge to both interpretation and translation, and it is the main task of Schleiermacher's theories to cope with this challenge. Schleiermacher's most original doctrine in the philosophy of language, (8) (semantic holism), is also highly relevant in this connection, for, as Schleiermacher perceives, semantic holism greatly exacerbates the challenge to interpretation and translation posed by (7).

Schleiermacher lectured on hermeneutics frequently between 1805 and 1833. The following are his main principles:

> (a) Hermeneutics is strictly the art of *understanding* verbal communication -- as contrasted, not equated, with explicating, applying, or translating it.
>
> (b) Hermeneutics should be a *universal* discipline -- i.e. one which applies equally to all subjects-areas (e.g. the bible, law, and literature), to oral as well as to written language, to modern texts as well as to ancient, to works in one's own language as well as to works in foreign languages, and so forth.
>
> (c) In particular, the interpretation of sacred texts such as the bible is included within it -- this may not rely on *special* principles, such as divine inspiration (of either the author or the interpreter).
>
> (d) Interpretation is a much more difficult task than is generally realized: contrary to a common misconception that “understanding occurs as a matter of course,” “misunderstanding occurs as a matter of course, and so understanding must be willed and sought at every point.” (This position derives from Schleiermacher's version of principle (7): deep linguistic and conceptual-intellectual diversity.) How, then, is interpretation to be accomplished?
>
> (e) Before the interpretation proper of a text can even begin, the interpreter must acquire a good knowledge of the text's historical context. (The suggestion found in some of the secondary literature that Schleiermacher thinks historical context *irrelevant* to interpretation is absurd.)
>
> (f) Interpretation proper always has two sides: one linguistic, the other psychological.

Linguistic interpretation's task (which rests on principle (5)) consists in inferring from the evidence consisting in particular actual uses of words to the rules that are governing them, i.e. to their usages and thus to their meanings; psychological interpretation instead focuses on an author's psychology. Linguistic interpretation is mainly concerned with what is common or shared in a language; psychological interpretation mainly with what is distinctive to a particular author.

(g) Schleiermacher implies several reasons why an interpreter needs to complement linguistic interpretation with psychological in this way. First, he sees this need as arising from the deep linguistic and conceptual-intellectual distinctiveness of individuals. Such distinctiveness at the individual level leads to the problem for linguistic interpretation that the actual uses of words which are available to serve as evidence from which to infer an author's exact usage or meaning will usually be relatively few in number and poor in contextual variety -- a problem which an appeal to authorial psychology is supposed to help solve by providing additional clues. Second, an appeal to authorial psychology is also required in order to resolve ambiguities at the level of linguistic meaning which occur in particular contexts (i.e. even after the range of meanings available to the author for the word(s) in question is known). Third, in order fully to understand a linguistic act one needs to know not only its linguistic meaning but also what some more recent philosophers have called its “illocutionary” force or intention. For example, if I encounter a stranger by a frozen lake who says to me, “The ice is thin over there,” in order fully to understand this utterance I need to know not only its linguistic meaning (which in this case is clear) but also whether it is being made merely as a factual statement, as a threat, as a joke … (Schleiermacher emphasizes the first of these three considerations most. However, if, as he does, one wants to argue that interpretation needs to invoke psychology *generally*, and if, as I hinted earlier, linguistic and conceptual-intellectual distinctiveness is not the pervasive phenomenon that he normally takes it to be, then it is arguably the latter two considerations that one should consider the more fundamental ones.)

(i) Interpretation also requires two different methods: a “comparative” method (i.e. roughly, a method of plain induction), which Schleiermacher sees as predominating on the linguistic side of interpretation (where it takes the interpreter from the particular uses of a word to the rule for use governing them all), and a “divinatory” method (i.e. roughly, a method of tentative and fallible hypothesis based on but also going well beyond available empirical evidence -- the etymology to keep in mind here is not Latin *divinus* but French *deviner*, to guess or conjecture), which he sees as predominating on the psychological side of interpretation. (The widespread idea in the secondary literature that “divination” is for Schleiermacher a process of psychological self-projection into texts contains a small grain of truth -- in that it is his view that interpretation requires some measure of psychological commonality between interpreter and interpreted -- but is basically mistaken.)

(j) Ideal interpretation is of its nature a holistic activity. (This principle in part rests on but also goes well beyond Schleiermacher's semantic holism.) In particular, any given piece of

text needs to be interpreted in light of the whole text to which it belongs, and both need to be interpreted in light of the broader language in which they are written, their larger historical context, a broader preexisting genre, the author's whole corpus, and the author's overall psychology. Such holism introduces a pervasive circularity into interpretation, for, ultimately, interpreting these broader items in its turn depends on interpreting such pieces of text. Schleiermacher does not see this circle as vicious, however. Why not? His solution is not that all of these tasks can and should be accomplished simultaneously -- something which would be beyond human capacities. Rather, it lies in the (very plausible) thought that understanding is not an all-or-nothing matter but something that comes in *degrees*, so that it is possible to make progress towards full understanding in a piecemeal way. For example, concerning the relation between a piece of text and its whole text, Schleiermacher recommends that we first read through and interpret as best we can each of the parts of the text in turn in order thereby to arrive at an approximate overall interpretation of the text, and that we then apply this approximate overall interpretation in order to refine our initial interpretations of the particular parts (which in turn gives us an improved overall interpretation, which can then be re-applied towards further refinement of the interpretation of the parts, and so on indefinitely).

Schleiermacher's indebtedness to Herder in this theory of interpretation extends far beyond the framework-principles (4), (5), and (7) mentioned earlier. Indeed, Schleiermacher's theory as it has just been described is almost identical to Herder's. Some of this commonality is admittedly due to shared influences (especially Ernesti). But Schleiermacher's theory owes exclusively to Herder the two central moves (often wrongly thought to have been original with Schleiermacher) of supplementing “linguistic” with “psychological” interpretation and of identifying “divination” as the predominant method of the latter. (Herder had introduced these two moves mainly in *On Thomas Abbt's Writings* (1768) and *On the Cognition and Sensation of the Human Soul* (1778).) Schleiermacher's theory as it has just been described in the main merely draws together and systematizes ideas which already lay scattered through a number of Herder's works.

There are two significant exceptions to that rule, however. First, as was mentioned, Schleiermacher exacerbates the challenge to interpretation which principle (7) already poses by introducing principle (8), semantic holism. Second, Schleiermacher's theory explicitly introduces principle (b), the *universality* of hermeneutics. This principle is very much in the spirit of Herder's theory, but it does go beyond its letter. (There were, however, other more explicit precedents -- e.g. van der Hardt, Chladenius, Pfeiffer, Grosch, and Meier.)

Schleiermacher's theory of interpretation also departs from Herder's in certain further respects not yet described. However, it is precisely here that it tends to become most problematic. This point has already been made above in relation to the exact force of the three principles in the philosophy of language which underpin it, (4), (5), and (7). But in addition: Unlike Herder, Schleiermacher, especially in his later work, more closely specifies psychological interpretation as a process of identifying, and tracing the necessary development of, a single authorial “seminal decision [Keimentschluß]” that lies behind a work and unfolds itself as the work in a necessary fashion -- which seems a very unhelpful move to make, for how

many works are composed, and hence properly interpretable, in such a way? Again, whereas Herder includes not only an author's linguistic behavior but also his non-linguistic behavior among the evidence relevant to psychological interpretation, Schleiermacher normally insists on a restriction to the former -- which seems misguided (e.g. the Marquis de Sade's recorded *acts* of cruelty seem no less potentially relevant to establishing the sadistic side of his psychological make-up, and hence to interpreting his texts in light of this, than his cruel statements). Again, unlike Herder, Schleiermacher regards the central role of “divination,” or hypothesis, in interpretation as a ground for sharply distinguishing interpretation from natural science, and hence for classifying interpretation as an art rather than a science -- whereas he should arguably instead have seen it as a ground for thinking interpretation and natural science *similar*. (This mistake was apparently caused by his false assumption that natural science works by something like plain induction -- i.e. roughly: this first *a* is *F*, this second *a* is *F*, this third *a* is *F*, … therefore all *a*s are *F* -- rather than by hypothesis.)

Schleiermacher's theory also tends to play down, obscure, or miss certain arguably important points relating to interpretation that Friedrich Schlegel had already made. Schlegel's treatment of hermeneutical matters, found mainly in his *Philosophy of Philology* (1797) and *Athenaeum Fragments* (1798-1800), largely resembles Schleiermacher's, but it also includes the following three points which are less radical, obscured, or altogether missing in Schleiermacher: (i) Schlegel stresses that (superior or “classical”) texts often express *unconscious* meanings (“Every excellent work … aims at more than it knows” - *Wilhelm Meister* (1798)). Schleiermacher sometimes implies a similar view (most famously in his doctrine that the interpreter should aim to understand an author better than the author understood himself), but Schlegel's version of it is more radical, envisaging indeed an “infinite depth” of meaning largely unknown to the author. (ii) Schlegel emphasizes that a work often expresses some of its important meanings, not explicitly in its parts, but rather through the way in which these are put together to form a whole. This is a very valuable point. Schleiermacher *in a way* makes this point as well, but only as incorporated into and obscured by his more dubious doctrine of the “seminal decision.” (iii) Unlike Schleiermacher, Schlegel, especially in his essay *On Unintelligibility* (1798), stresses that works typically contain confusions, and that it is vitally important for an interpreter to identify and explain these: “It is not enough that one understand the actual sense of a confused work better than the author understood it. One must also oneself be able to know, *characterize*, and even *construe* the confusion even down to its very principles.” This is arguably another very important point.

Be these shortcomings in the details of Schleiermacher's hermeneutics as they may, his pupil Böckh, an eminent classical philologist, subsequently gave a faithful and even more systematic articulation of Schleiermacher's hermeneutics in lectures which were eventually published as Böckh's *Encyclopaedia and Methodology of the Philological Sciences* (1877), and through the combined influence of Schleiermacher's and Böckh's treatments it achieved something like the status of the official hermeneutical theory of nineteenth-century classical scholarship.

5. Theory of Translation

Once again, Schleiermacher develops his theory of translation on the foundation of the Herder-influenced

principles (4), (5), and (7), together with (8), his own holism about meaning, which exacerbates the challenge to translation already posed by (7).

Schleiermacher was himself a masterful translator, whose German translations of Plato are still widely used and admired today, nearly two hundred years after they were done. So his views on translation carry a certain prima facie authority. He explains his theory of translation mainly in the brilliant little essay *On the Different Methods of Translation* (1813). The following are his most important points:

> (a) Translation typically faces the problem of a conceptual gulf between the language of the text to be translated and the translator's home language (as the latter currently exists). (This is an application of principle (7).)
>
> (b) This situation makes translation an extremely difficult task, posing a major obstacle to the attainment of translation's primary goal, the faithful reproduction of meaning. In this connection, Schleiermacher in particular notes the following problem, which one might dub the paradox of paraphrase: If, faced with the task of translating an alien concept, a translator attempts to reproduce its intension by reproducing its extension by means of an elaborate paraphrase in his own language, he will generally find that as he gets closer to the original extension he undermines the original intension in other ways. For example, faced with Homer's word *chlôros*, which Homer sometimes applies to things that we would classify as green (e.g. healthy foliage) but at other times to things that we would classify as yellow (e.g. honey), a translator might attempt to reproduce the extension correctly by translating the word as “green or yellow.” But in doing so he would be sacrificing the original intension in other ways -- for Homer did not *have* the concept green (only *chlôros*), and in addition for Homer *chlôros* was not a *disjunctive* concept. (Schleiermacher also identifies a number of further challenges which often exacerbate the difficulty of translation. For example, he notes that in the case of poetry it is necessary to reproduce not only the semantical but also the musical aspects of the original, such as meter and rhyme -- and this not only as a desideratum over and above the main task of reproducing meaning, but also as an essential *part* of that task, because in poetry such musical features serve as essential vehicles for the precise expression of meaning. Also, he argues that in addition to reproducing meaning a translation should attempt to convey to its readership where an author was being conceptually conventional and where conceptually original, e.g. by using older vocabulary in the former cases and relative neologisms in the latter. Also, he notes that in both of these connections the added requirement or desideratum involved will frequently stand in tension with that of finding the closest semantical fit -- e.g. it will turn out that the word which would best reproduce a rhyme or best reflect a concept's vintage is not the one that is closest in meaning to the word in the original.)
>
> (c) Because of this daunting difficulty the translator needs to possess hermeneutical expertise and to be an “artist” if he is to cope with the task of translation at all adequately.
>
> (d) The conceptual gulf which poses the central challenge here might in principle be tackled

in one of two broad ways: either by bringing the author's linguistic-conceptual world closer to that of the reader of the translation or vice versa. The former approach had in fact been championed by Luther in his classic essay *On Translating: An Open Letter* (1530) and practiced by him in his translation of the bible (he called it *Verdeutschung*, “Germanizing”). However, Schleiermacher finds it unacceptable, mainly because it inevitably distorts the author's concepts and thoughts. Schleiermacher therefore champions the alternative approach of bringing the reader towards the linguistic-conceptual world of the author as the only acceptable one. But how can this possibly be accomplished?

(e) According to Schleiermacher, the key to the solution lies in the plasticity of language. Because of this plasticity, even if the usages of words and hence the concepts expressed by the language into which the translation is to be done *as it currently exists* are incommensurable with the author's, it is still possible for a translator to “bend the language of the translation as far possible towards that of the original in order to communicate as far as possible an impression of the system of concepts developed in it.” (Note that this solution presupposes principle (5).) Consider, for example, a translator facing the challenge of translating Homer's word *aretê* into English. The translator will recognize that nothing in existing English exactly expresses this concept. He will therefore judge that the best way to convey it in English is to modify existing English usage in a systematic way for the course of the translation in order thereby to mimic Greek usage and hence meaning. He will begin by taking the word from existing English which comes closest to *aretê* in meaning, say the word *virtue*. However, he will recognize that the rule for use which governs this word in existing English is still very different from that which governs Homer's word *aretê* so that the two words are still quite sharply different in meaning - that, for example, the descriptive component of the rule which governs the word *virtue* in existing English makes it a solecism to ascribe *virtue* to a habitual liar or a pirate, but quite proper under certain circumstances to ascribe it to a physically weak man, whereas exactly the converse rule governs the word *aretê* in Homer. What therefore will the translator do? He will not simply resign himself to living with this discrepancy. Instead, for the duration of his translation he will modify the rule which governs the word *virtue* in order to make this rule conform (or at least more closely conform) to that which governs Homer's word *aretê* -- for example, he will drop the descriptive rule governing the word *virtue* which was just mentioned and switch to its converse instead, consequently for the duration of his translation writing quite happily of certain habitual liars and pirates as having *virtue* (e.g. Odysseus and Achilles, respectively), but scrupulously avoiding describing any physically weak man as having it. He will thereby succeed in expressing -- or at least come close to expressing -- in English the meaning of Homer's word *aretê*.

(f) This approach entails a strong preference for translating any given word in the original in a uniform way throughout the translation rather than switching between two or more different ways of translating it in different contexts.

(g) This approach also makes for translations which are considerably less easy to read than

those which can be achieved by the alternative approach (*Verdeutschung*). However, this is an acceptable price to pay given that the only alternative is a failure to convey the author's meaning at all accurately. Moreover, the offending peculiarities have a positive value in that they constantly remind the reader of the conceptual unfamiliarity of the material that is being translated and of the “bending” approach that is being employed.

(h) In order to work at all effectively, though, this approach requires that large amounts of relevant material be translated, so that the reader of the translation acquires enough examples of a word's unfamiliar use in enough different contexts to enable him to infer the unfamiliar rule for use involved.

(i) Even this optimal approach to translation has severe limitations, however. In particular, it will often be impossible to reproduce the holistic aspects of meaning -- the several related usages of a given word, the systems of related words / concepts, and the distinctive grammar of the language. And since these holistic features are internal to a word's meaning, this will entail a shortfall in the communication of its meaning by the translation. Reading a translation therefore inevitably remains only a poor second best to reading the original.

(j) Translation is still justified, though -- not only by the obvious consideration that it makes works available to people who want to read them but who are not in the fortunate position of knowing the original languages, but also by the less obvious one that through its “bending” approach it effects a conceptual enrichment of their language.

(k) Nor (Schleiermacher adds in answer to a worry which Herder had expressed) need we fear that this enrichment will deprive our language of its authentic character. For in cases where a real conflict with that character arises, the enrichments in question will soon wither from the language.

Once again, not only the framework principles (4), (5), and (7), but most of these ideas about translation come from Herder. In particular, Schleiermacher's central strategy of “bending” the pregiven language in order to cope with conceptual incommensurability, and his point that it is important to convey the musical aspects of an original (poetic) text in order to convey its meaning accurately, both do so. (Relevant Herderian sources in this connection are the *Fragments on Recent German Literature* (1767-8) and the prefaces of the *Volkslieder* (1774).) However, unlike Schleiermacher's theory of interpretation, which as was mentioned often worsens Herder's, this theory of translation tends to refine Herder's in some modest but significant ways. Among the ideas just adumbrated, examples of this occur in (b), where Schleiermacher's paradox of paraphrase and his ideal of making clear in the translation at which points the author is being conceptually conventional and at which points conceptually original are both novel; (g), where Schleiermacher's point that the resulting peculiarities are not only unavoidable but can actually serve a positive function of reminding readers of the conceptual incommensurability involved and the “bending” approach being employed to cope with it is novel; (h), which is a novel point; (i), where Schleiermacher's point about the in-principle limitation on the successfulness of translations posed by semantic holism is a novel one; and (k), which plausibly contradicts Herder.

6. Aesthetics

Schleiermacher was generally quite self-deprecating about his sensitivity to and knowledge of art (e.g. in *On Religion* and the *Soliloquies*, where he is clearly rather in awe of the greater expertise in this area enjoyed by such romantic friends as the Schlegel brothers), and accordingly he tended to shy away from discussing it in detail in his earlier work. However, he did eventually bring himself to confront the subject systematically, namely in his lectures on aesthetics (first given in 1819, and then again in 1825 and 1832-3).

Part of his motivation behind this eventual confrontation with the subject -- and part of why it remains interesting today -- derives from the fact that the phenomenon of art, and in particular the phenomenon of non-linguistic art (e.g. painting, sculpture, and music), provokes a certain theoretical question which is of fundamental importance, not only for the philosophy of art itself, but also for hermeneutics or the theory of interpretation, and for the philosophy of language which underlies the theory of interpretation: Do non-linguistic arts such as painting, sculpture, and music express meanings and thoughts, and if so how? This question is important for the theory of interpretation because it brings in its train such further questions as whether the theory of interpretation should not include among its proper objects further forms of expression in addition to the linguistic ones treated by Schleiermacher's own hermeneutics, what the appropriate methods of interpretation might be in such further cases, and how such cases and their interpretive methods might relate to the linguistic ones. Moreover, this question is equally important for the philosophy of language which underpins Schleiermacher's theory of interpretation, as embodied in principles (4) and (5). For a positive answer to this question might seem to threaten these two principles, or at least to show that they need radical revision if they are to be defensible.

In his last cycle of aesthetics lectures (1832-3) Schleiermacher initially pursues a very simple strategy for dealing with these issues arising in relation to non-linguistic art; however, he eventually realizes that the strategy in question is untenable, and abandons it for a more promising but also more ambiguous position.

His whole train of thought closely follows one that Herder had already pursued in the *Critical Forests* (1769), and so it may be useful to begin with a brief sketch of the latter. By the time of writing the *Critical Forests* Herder was already committed to his own versions of principles (4) and (5). Accordingly, in reaction to the phenomenon of the non-linguistic arts the book initially set out to argue for a theory of their nature which would preserve consistency with those principles, and it did so in a very straightforward way, denying the non-linguistic arts the ability to express thoughts or meanings *autonomously* of language by denying them the ability to express thoughts or meanings *at all*: whereas poetry has a sense, a soul, a force, music is a mere succession of objects in time, and sculpture and painting are merely spatial; whereas poetry not only depends on the senses but also relates to the imagination, music, sculpture, and painting belong solely to the senses (to hearing, feeling, and vision, respectively); whereas poetry uses *voluntary and conventional* signs, music, sculpture, and painting employ only *natural* ones. However, as Herder proceeded with his book he eventually realized that this simplistic solution was untenable: in the third part of the work he came across the awkward case of

ancient coins, which, though normally non-linguistic, clearly do nonetheless often express meanings and thoughts in pictorial ways. This did not lead him to abandon his versions of principles (4) and (5), however. Instead, it brought him to a more refined account of the non-linguistic arts which was still consistent with those principles: the non-linguistic arts do sometimes express meanings and thoughts, but the meanings and thoughts in question are ones which are *parasitic on a prior linguistic expression or expressibility of them possessed by the artist*. In the fourth part of the book (not published until the mid-nineteenth century, and hence unknown to Schleiermacher) Herder extended this solution from coins to painting, and in subsequent works to sculpture and music as well.

Schleiermacher's aesthetics lectures follow a very similar course. He initially sets out to develop a version of the theory that Herder had initially developed in the *Critical Forests*, correlating the several non-linguistic arts with the different senses as Herder's theory had done (his only significant revision consists in modifying Herder's correlation of sculpture with the sense of touch to include vision as well as touch). Like Herder's initial theory, Schleiermacher's is largely motivated by his prior commitment to principles (4) and (5), which, again like Herder's initial theory, it seeks to vindicate in a naive way: non-linguistic arts, such as music and sculpture, do not express meanings or thoughts *autonomously* of language because they do not express them *at all*. (For example, Schleiermacher argues that music merely expresses physiologically based “life-conditions [Lebenszustände],” not representations or thoughts.) However, rather like Herder with his ancient coins, in the course of developing this naive solution Schleiermacher abruptly confronts a case which forces him to the realization that it is untenable. He develops his naive solution smoothly enough for the cases of music and painting. But in the middle of his discussion of sculpture he suddenly recalls Pausanias's account that the very earliest Greek sculptures were merely rough blocks whose function was to serve, precisely, as symbols of religious *ideas* (oops!). He subsequently goes on to note that an analogous point in fact holds for other non-linguistic arts such as painting as well. Accordingly, at this stage in his lectures he changes tack: He now acknowledges that non-linguistic arts *do* (at least sometimes) express meanings and thoughts after all. And he goes on to vacillate between two new, and conflicting, accounts of that fact: (a) The arts in question do so in such a way that the meanings and thoughts involved are at least sometimes not (yet) linguistically articulable. (In particular, Schleiermacher suggests that the early Greek sculpture just mentioned expressed religious ideas which only *later* got expressed linguistically.) This account would entail abandoning or at least severely revising principles (4) and (5). (b) The arts in question do so in virtue of a pre-existing linguistic articulation or articulability of the same meanings and thoughts in the artist. (Schleiermacher actually only says in virtue of “something universal,” “a representation,” but a dependence on language seems clearly implied.) This account is similar to Herder's final account, and would preserve principles (4) and (5). In the end, then, having renounced his initial -- clearly untenable -- position, Schleiermacher is left torn between these two more plausible-looking positions, which, however, contradict each other.

The eighteenth- and nineteenth-century German hermeneutical tradition as a whole was similarly torn between these two positions. As has already been mentioned, (b) was the considered position at which Herder eventually arrived. But (a) had strong champions as well -- in particular, Hamann, Hegel (concerning architecture and sculpture), and the later Dilthey. The choice between these two positions is a genuinely difficult one, philosophically speaking.

Where does this leave Schleiermacher in relation to the several issues bearing on his theory of interpretation and philosophy of language which, I suggested, encouraged him to undertake this investigation of non-linguistic art in the first place? Concerning the primary question, whether the non-linguistic arts express meanings and thoughts and if so how, he has now realized that they do indeed (sometimes) express meanings and thoughts, but he remains torn on exactly how they do so. Concerning his theory of interpretation, that realization is itself important, because it has shown him that interpretation theory must indeed extend its objects beyond linguistic ones to include at least some that are non-linguistic. But he remains torn on the further questions in this area -- in particular whether, as (a) implies, there will be cases in which the interpretation of non-linguistic art will transcend the interpretation of any associated language or, as (b) implies, it will always be dependent on and restricted by the interpretation of associated language. Finally, concerning the philosophy of language which underpins his theory of interpretation, he remains torn about whether the meanings and thoughts expressed by non-linguistic art are always parasitic on language (position (b)), so that principles (4) and (5) can be retained without qualification or modification, or instead sometimes independent of language (position (a)), so that principles (4) and (5) will either have to be abandoned or (as Hamann had already done in his *Metacritique* (1784)) (re)construed in a way that stretches their reference to “language” and “words” to include not-strictly-linguistic-or-verbal symbol-use by the non-linguistic arts.

Another important motive behind Schleiermacher's treatment of art in his late aesthetics lectures concerns its cultural status, and in particular its cultural status relative to religion. It was an abiding concern of Schleiermacher's from very early in his career until the very end of it to subordinate art to religion. The final cycle of aesthetics lectures is merely the last in a long line of attempts to achieve this goal. It seems to me, however, that, partly for reasons already touched on, this last attempt turns out to be oddly and interestingly self-subverting.

I shall briefly review Schleiermacher's series of attempts to subordinate art, and then explain how this last one proves self-subverting. It was already one of the early Schleiermacher's primary goals to turn contemporary culture, and especially the romantic movement, away from the then fashionable idea that art was the highest possible form of insight towards the idea that religion was. This is an important part of the project of *On Religion* (1799). In this work Schleiermacher criticizes the sort of elevation of art above religion that Goethe and Schiller had begun and the romantics had developed, complains of the trivial nature of modern art, and argues that art ought to subserve religion as Plato had thought. (The early Schleiermacher was in a sense strikingly successful in achieving his goal; after 1799 the leading romantics did indeed increasingly turn away from art towards religion, and to some extent the same was also true of German culture more broadly.)

The ethics lectures of 1812-13 continue the same project in a certain way. There Schleiermacher represents art as of its very nature a collective expression of religious feeling (one which differs in accordance with the differences between religions). In other words, he represents art as only true to its own nature when it subserves religion.

The 1830 psychology lectures play an interesting variation on this theme. There Schleiermacher argues that the perception of beauty is a feeling but one which has a sort of deep cognitive content in that it

expresses the relation of intelligence to Being. This makes it sound very much like religious feeling, and indeed in these lectures it is treated as a sort of close second-in-command to religious feeling. It might seem as though, from Schleiermacher's viewpoint, there was a danger here of art acquiring too independent and exalted a status. However, this danger is partly averted by the fact that he is here talking primarily about *natural* beauty, and only secondarily about artistic.

The 1832-3 aesthetics lectures continue this sort of demoting project, but in a different way. Schleiermacher's initial intention there, it seems, was to demote art (in comparison with religion) in two ways: First, as we saw, the lectures initially set out aiming to give an account of non-linguistic arts (music, painting, and sculpture) which represents them as merely expressive of sensuous feelings and non-cognitive in character. Second, the lectures give an account of poetry which represents it as merely national and indeed merely individual in nature (not universal). Thus the lectures argue that art generally, and therefore poetry in particular, is national in nature, not universal like science and (in a sense) religion, and more radically that poetry has the function of expressing individuality, of resisting even the commonality of a national language (thereby making explicit a potential which is also present, though less realized, in normal language use).

However, this two-fold strategy for demoting art turns out to be curiously self-subverting. For one thing, as we saw, the model of non-linguistic art as merely sensuous and non-cognitive in the end proves unsustainable. Moreover, not only does non-linguistic art turn out to have a cognitive content after all, but in addition this becomes clear from a case (the earliest Greek sculpture) in which the content in question is not trivial but deeply religious in character. Also, this self-subversion would be even more extreme if position (a) won out over position (b). For another thing, poetry's function of expressing individuality implies that it represents Schleiermacher's *highest ethical value.* In short, what was intended as a demotion of art turns willy-nilly into a sort of cognitive-religious and ethical exaltation of it.

7. Dialectics

Most of Schleiermacher's earliest philosophical work was in areas of the subject which might reasonably be described as peripheral in comparison with such central areas as metaphysics and epistemology (in particular, ethics, philosophy of religion, and hermeneutics). This fact, together no doubt with the imposing presence of several competitors who had recently made or were making contributions in those central areas (including Kant, Fichte, Schelling, and Hegel), seems to have spurred Schleiermacher to develop his own treatment of them. The result was his “dialectics,” which he began to present in lecture-form in 1811. (The subject-title calls to mind relevant thought not only in Plato and Aristotle, but also in Kant and Hegel.)

Accordingly, Schleiermacher's dialectics in certain ways carries the marks of a discipline which he felt forced to develop, rather than one for which he had a clear, compelling vision (as he did for his philosophy of religion and hermeneutics for instance). For one thing, the nature of the discipline undergoes a striking shift between its two earliest versions (the lectures of 1811 and 1814-15) -- which have the character of fairly conventional treatments of metaphysical and epistemological issues, already

concerned indeed with resolving disagreements to some extent, but in a purely theoretical way -- and its two main later versions (the lectures of 1822 and the book-fragment from 1833), which make the art of actually resolving disagreements through conversation the main core of the discipline (albeit that “conversation” is here understood in a broad sense so as to include not only the paradigm case of oral communication but also written communication and even dialogue internal to a single person's mind). (With some qualification, this shift might be described as one from a more Aristotelian to a more Socratic-Platonic conception of “dialectics.”)

For another thing, in all of its versions Schleiermacher's dialectics has an oddly rag-bag appearance, including as it does not only material that would traditionally be classified as metaphysics and epistemology, but also large components of philosophy of mind, logic (especially the logic of concepts and judgments; Schleiermacher treats the logic of syllogism in a reductive and rather deprecatory way), philosophy of science, and philosophy of religion.

In its final versions (on which I will focus), the discipline has roughly the following character: Its concern is with what Schleiermacher calls “pure thought,” as distinguished from the thought of everday affairs or art -- i.e. with thought which aims at truth, rather than merely at achieving practical ends or inventing fictions. (Schleiermacher denies, though, that the former is sharply divorced from the latter; rather, it is to some degree implicit in them and vice versa.)

According to Schleiermacher, genuine knowledge of its very nature requires, not only (1) correspondence to reality, but also (2) systematic coherence with *all* knowledge and (3) univeral agreement among people. (The main motive behind this elaborate position seems to derive from the thought that there is in principle no way of determining the fulfillment of condition (1) *directly*, so that believers need to rely on guidance by the fulfillment of conditions (2) and (3).)

(In a well-known early interpretation of Schleiermacher's dialectics the German scholar Manfred Frank accentuated condition (3), attributing to Schleiermacher on this basis a consensus-theory of truth. However, in his recent edition of Schleiermacher's dialectics lectures Frank rightly admits that this interpretation overlooked the realism implied by condition (1). This revision of Frank's early reading of Schleiermacher's dialectics also undercuts Frank's equally well-known early reading of Schleiermacher's hermeneutics, which built upon this ascription to Schleiermacher of a consensus-theory of truth an ascription to him also of a (roughly Gadamerian) conception of interpretation as an ongoing construction of facts about meaning through the development of interpretations.)

Not surprisingly given the strength of the three conditions just mentioned, Schleiermacher considers genuine knowledge to be only an “idea” towards which we can make progress, not something that we can ever actually achieve. (His position here resembles his official position in hermeneutics that the genuine understanding of another person is only something to which we can approximate, not something we can ever actually achieve.)

Schleiermacher's dialectics is largely conceived as a methodology for making such progress. This project proceeds relatively smoothly in connection with conditions (1) and (2). For example, in connection with

(1), Schleiermacher develops certain principles concerning how to form concepts correctly rather than incorrectly (i.e. in such a way that -- to borrow a more recent idiom -- they, their superordinate genus-concepts, their subordinate species-concepts, and their contrasting coordinate concepts “carve nature at the joints”). And in connection with (2), while he acknowledges that the task of forming a totality of knowledge is of its nature incompletable, he nonetheless prescribes what he calls ”heuristic” and “architectonic” procedures for, respectively, amassing the pieces of knowledge and forming them together into a coherent whole.

However, the project runs into deeper difficulties in relation to condition (3). There are two main problems here. First, in addition to the obvious and avowed impossibility of actually accessing all people in order to come to agreement with them, Schleiermacher also identifies a further obstacle in the way of reaching, or even making significant progress towards, agreement with them: the deep differences which occur between different languages and modes of thought. The dialectics lectures themselves fail to find any very promising way of coping with this problem. The 1822 version attempts two ways, but neither looks hopeful. Its first approach consists in hypothesizing a domain of “innate concepts” common to everyone (with certain qualifications, e.g. that these concepts require sensations in order to be actualized). This would certainly solve the problem, but only by contradicting Schleiermacher's normal, and surely philosophically superior, position, from which the problem arose in the first place, that there *is no* such conceptual commonality across all different languages (or even, Schleiermacher would normally add, between all individuals who in some sense share a language). The second approach attempted is an argument that we need to develop a complete history of the differences in question and of how they arose. However, this seems beside the point -- a distraction from the problem rather than a solution to it. In the 1833 book-fragment Schleiermacher in places seems close to giving up on this problem, saying at one point that because of it dialectics must restrict itself to a specific “linguistic sphere,” but at other points he evidently still clings to the hope of finding common ground uniting different “linguistic spheres.” What sort of solution does he have in mind? The sort of solution he may have in mind can perhaps best be seen from an 1831 lecture which he gave on Leibniz's idea of a universal language. In this lecture he in effect argues that it was a mistake on Leibniz's part to suppose that there was *already* a conceptual common ground shared by everyone which could be captured in a universal language (this also amounts to a rejection of his own dubious idea in the 1822 lectures of common “innate concepts”), but that the sort of conceptual common ground that Leibniz had thus wrongly envisaged as *already* existing can nonetheless be *achieved* (or at least approached) for the sciences, namely by cultivating an attitude of openness to the borrowing of conceptual resources from other languages as such resources prove themselves useful for the sciences (a process which, according to Schleiermacher, is in fact already heavily underway, and which is realizable either through outright borrowing of the foreign words in question or through translation of them into one's language in the sort of sensitive way that his theory of translation advocates). Schleiermacher points out that this solution requires an (in any case healthy) shedding of prejudices about the superiority of one's own language, mode of thought, and people over others. This looks like Schleiermacher's most promising solution to the problem in question. He did not, however, live long enough to develop it in detail or to build on it towards a more complete method for resolving interlinguistic disagreements.

Second, and more surprisingly, Schleiermacher's dialectics lectures do not even develop any substantive

account of how to resolve disagreements through conversation *within* a “linguistic sphere.” However, here again it is fortunately possible to supplement the dialectics lectures with extraneous material which goes further in such a direction. One important text in this connection is Schleiermacher's early essay *Toward a Theory of Sociable Conduct* (1799), which is precisely concerned with the art of conversation within a linguistic sphere. This early essay emphasizes the importance of finding a (conceptual) “content” that one shares with one's interlocutor(s), and restricting one's conversation to this. Schleiermacher accordingly recommends there that one begin a conversation guided by a sort of minimal estimate of such content arrived at from one's knowledge of such things as the profession, the educational background, and the class of one's interlocutor(s), but that one thence tentatively and experimentally work outwards towards identifying and exploiting further shared content -- a process which he recommends one should undertake, not by the heavy-handed method of introducing doubtfully shared content directly, but rather by the subtler method of introducing it indirectly in the form of a dimension of allusion and satire which one adds to one's treatment of already established shared content (after which, if the response is positive, it can join the previously established shared content as a proper subject matter for direct treatment). Another important text in this connection is Schleiermacher's hermeneutics lectures, which implicitly revise the earlier account just described in two respects: (a) In that account conversation was to be restricted to conceptual content that was already shared between interlocutors. But as we saw previously, by the time Schleiermacher writes the hermeneutics lectures he is skeptical that people *ever* really share conceptual content. Consequently, he would presumably now set the bar for fruitful conversation somewhat lower than strict sharing. (b) Also, it seems reasonable to infer from his conception of hermeneutics that he would now place less emphasis on discovering preexisting commonalities, or near-commonalities, and more on refining those found and establishing further ones -- namely, in both cases, through adept use of the art of hermeneutics.

Finally, Schleiermacher's hermeneutics lectures also supply a further part of his seemingly missing solution to the problem of reaching agreement through conversation, both for inter- and for intra-linguistic contexts. Clearly, any art of reaching agreement through conversation is going to depend on an art of interpreting interlocutors. Accordingly, the dialectics lectures explicitly assert the dependence of dialectics on hermeneutics (as well as vice versa), Schleiermacher's conception of hermeneutics as a universal discipline ensures its applicability to conversations, and Schleiermacher mentions in the hermeneutics lectures that he sometimes applies his own hermeneutical principles in conversational contexts. In short, Schleiermacher's hermeneutics itself constitutes an important component of his art of reaching agreement through conversation.

In sum, whereas Schleiermacher's final conception of dialectics as a discipline leads one to expect it to provide a fairly detailed set of procedures for resolving both inter- and intra-linguistic disagreements in conversation (analogous to the detailed set of procedures for interpretation which one finds in his hermeneutics), this expectation is largely disappointed by the dialectics lectures themselves. However, one can supplement the dialectics lectures from other texts in order to see how Schleiermacher might have envisaged a fuller solution to this task.

One last point which also deserves mention in this connection is the following. Schleiermacher's most prominent motive for developing such an art of conversation is the epistemological one mentioned earlier.

That may or may not be a good motive in the end. However, Schleiermacher also has further and independent motives behind this art which are more obviously attractive. Thus, the 1831 lecture on Leibniz implies two additional motives behind the intercultural side of the art: first, Schleiermacher's cosmopolitan concern for humanity as a whole in all its diversity constitutes a moral reason for promoting fruitful intercultural dialogue; and second, his sense that insight, far from being our monopoly, is dispersed among many cultures constitutes another reason for us to engage in such dialogue. (Schleiermacher would presumably say that analogous considerations help to justify the intracultural side of the art as well.) Again, the essay *Toward a Theory of Sociable Conduct* emphasizes an additional motive behind the intracultural side of the art. In this essay Schleiermacher does not mention his later epistemological motive at all, but instead focuses on more direct benefits which he sees accruing from and depending on fruitful conversation between members of society, in particular an enrichment of the individual's own limited perspective through his incorporation of the different perspectives of other people. (Schleiermacher would presumably say that an analogous consideration helps to justify the intercultural side of the art as well.) In short, even if it were to turn out that Schleiermacher's predominant epistemological motive for developing an art of inter- and intra-cultural conversation were unpersuasive, such an art might still be highly valuable for other reasons that he has in mind such as these.

Finally, a few further positions from Schleiermacher's dialectics may also be worth mentioning briefly. One striking position is a denial that any concepts, thoughts, or cognitions are either purely a priori in nature or purely empirical, either the product of the “intellectual” function alone or the product of the “organic” function alone. All are the product of *both* functions -- though the *proportions* in which they are involved vary from case to case.

More specifically, as Schleiermacher conceives matters, all are located on a continuum which stretches between the maximally “intellectual” ideas of Being or God and the maximally “organic” chaos of sensations. These two extremes do not themselves involve mixture, Being or God being purely intellectual, while the chaos of sensations is purely organic. However, they do not for that reason constitute counterexamples to the position just mentioned, because they are not themselves strictly speaking concepts, thoughts, or cognitions.

As was previously mentioned, Schleiermacher's theory of concepts also says that they are in each case defined by relations of subsumption under higher concepts, contrast with correlative concepts similarly subsumed, and subsumption of further concepts under them. Subsumption under the non-concept Being and the subsumption of a class of primitive judgments about sensations constitute special cases at the two extremes of this conceptual hierarchy.

Another position which Schleiermacher holds is that the distinction between analytic and synthetic judgments is a merely “relative” one. One reason for this position is probably his view that all judgments are partly empirical in nature (a consideration which anticipates Quine). But what he mainly seems to have in mind is that it is always in some sense up to us to decide how many and which characteristic marks to build into any given subject concept, and therefore how many and which judgments in which that subject concept features will count as analytic or as synthetic.

A last feature of Schleiermacher's dialectics is more puzzling. Schleiermacher notes at one point that he wants to chart a sort of middle course between ancient dialectics, which had the virtue of openness but the vice of courting skepticism, and the dogmatism of the scholastics, for whom everything of importance was pre-decided in an assumed religious principle. His concession to the former position has in effect been described above. But what about his concession to the latter? This takes the form of positing a “transcendental ground” or God which is (1) beyond all oppositions, including those of thought/reality, thought/volition, and concept/judgment, (2) beyond Being (even though Being is itself beyond such oppositions), (3) an essential impulse behind, and accompaniment of, all attempts to know, and (4) not thinkable or linguistically expressible but instead felt. This is all rather mysterious. For example, the philosophical rationale for positing this “transcendental ground” or God as beyond rather than identical with Being is obscure, and so too is the exact way in which it is supposed to be the impulse behind and accompaniment of all attempts to know.

8. Ethics

Schleiermacher's ethical thought divides into two overlapping chronological phases: The first phase -- which stretched from the late 1780's until about 1803 -- was mainly critical in character. Early in this phase, the three unpublished essays *On the Highest Good* (1789), *On What Gives Value to Life* (1792-3), and *On Freedom* (1790-3) mounted a thorough attack on Kant's ethical theory, and at the end of this phase the longer published work *Outlines of a Critique of Previous Ethical Theory* (1803) developed that attack into a more comprehensive and systematic critique of predecessors' ethical theories. The second phase -- which began around 1800 -- was mainly constructive in character. To this phase belong the *Soliloquies* (1800), the *Draft of an Ethics* (1805-6), and Schleiermacher's mature ethics lectures (including the complete draft from 1812-13, as well as a number of later partial drafts).

The three early essays *On the Highest Good*, *On What Gives Value to Life*, and *On Freedom* criticize and reject central tenets of Kant's moral philosophy: in particular, Kant's inclusion in the “*summum bonum* [highest good]” of an apportioning of happiness to moral desert, his position that this must be believed in as a presupposition of morality, so that its own implicit presuppositions, an afterlife of the soul and a God, must be so too (the doctrine of the “postulates”), and his incompatibilism concerning causal determinism and the freedom required for moral responsibility and consequent recourse to the causally indeterministic noumenal realm as the locus for freedom (Schleiermacher's *On Freedom* argues for the causal determination of all human actions, but for the compatibility of this with the freedom required for moral responsibility).

A further area of disagreement with Kant forms the hinge on which Schleiermacher's development of his own constructive ethical theory turns. Kant's fundamental moral principle, the “categorical imperative,” consisted in a requirement of the consistency of an agent's moral maxim (or intention) when universalized, and was conceived by Kant to apply uniformly to *all* human beings. Schleiermacher rejects this position in two ways. First, already in *On What Gives Value to Life*, and then especially in the *Soliloquies*, he argues against the (latter) idea of uniformity in ethics -- instead asserting, in the spirit of Herder (and others influenced by Herder, such as Goethe, Schiller, and the romantics), the value of

diversity or individuality even in the moral sphere. In this connection, Schleiermacher champions not only a (moral) distinctiveness of different human societies vis-à-vis the human species as a whole (this had been Herder's pet cause), but also a (moral) distinctiveness of the individual vis-à-vis his society. (In *On Religion* he makes an analogous case for both societal and individual diversity in religion. His positive valuing of societal and individual diversity naturally also extends beyond morals and religion.)

Second, Schleiermacher also rejects the content of Kant's "categorical imperative," in *On Religion* and the *Soliloquies* championing against this an -- again Herderian -- commitment to *humanity* (in the sense of an ideal of the welfare of all human beings). (In *On Religion* Schleiermacher discusses the historical dimension of this principle of humanity in a Herderian vein, like Herder in his *Ideas for the Philosophy of History of Humanity* (1784-91) stressing the important role of (Christian) religion in advancing it, and interpreting history as its progressive realization.)

This double position of Schleiermacher's might seem to court the following sort of problem: What if the moral values of a society or an individual conflict with the ideal of humanity? (What, for example, if the society is Nazi Germany or the individual Hitler?) In the *Soliloquies* Schleiermacher forestalls this sort of problem by limiting the forms of moral distinctiveness and individuality that he supports to those which are compatible with or even promotive of the ideal of humanity. Thus he expresses his commitment to moral distinctiveness or individuality in such formulas as that a person should be an individual "without violating the laws of humanity," that "each human being should represent humanity in his own way," and that what is valuable is a person's "distinctive being and its relation to humanity."

Similarly, Schleiermacher's championing of (moral) diversity or individuality is always combined with requirements of a measure of conformity with a broader species-wide or societal whole.

This constructive tension between "distinctive [eigentümlich]" and "universal" sides of ethics survives to constitute the central principle of Schleiermacher's mature ethics lectures. There he begins by arguing that very general forms or analogues of such a constructive tension are universal facts of nature -- that all finite beings exhibit such a tension, more specifically that all life does so in the more specific form of a tension between autonomy and social commonality, and more specifically still that all human mental life does so in this same form. He then goes on to derive a moral duty to realize such a tension in one's own person.

This provokes certain questions, to which the answers are not entirely clear. First, is Schleiermacher not here guilty of the so-called "naturalistic fallacy," of attempting to deduce an "ought" from an "is"? The answer to this question would depend on the exact nature of his derivation of the moral duty from the universal facts of nature, which is obscure. Second, how can a synthesis of commonality with individuality both be an unavoidable fact about human nature (e.g. because we can never quite share *any* concepts, we also can never quite share any moral concepts in particular) and be a moral duty? There are two possible answers to this puzzle. One would appeal to Schleiermacher's determinism and compatibilism; that a mode of existence or behavior is inevitable does not for him exclude its moral obligatoriness. The other would instead appeal to the fact that the sort of synthesis in question can come in varying degrees; it might be that *some* degree of moral individuality is indeed inevitable for the reason

mentioned but that the degree which is morally required is greater.

In addition to the central principle just discussed, three further aspects of the ethics lectures are worth mentioning briefly: (a) (As was reflected in the structure of the argument just sketched) Schleiermacher's mature conception of ethics is that it is fundamentally ontological rather than merely prescriptive in character: it concerns the immanence of “reason” in “nature,” and is hence more fundamentally a matter of an ”is” than of a “should.” (b) Accordingly (with an eye to the role of “reason” just mentioned), for Schleiermacher ethics is not fundamentally a matter of sentiments (these, he says, simply vary), but instead of cognitions, or more exactly, of something that grounds both ethical sentiments and ethical cognitions. (Here Schleiermacher is for once close to agreement with Kant.) (c) Accordingly again (but this time with an eye to the predominance of ontology over prescription just mentioned), Schleiermacher divides his ethics into a Doctrine of Goods, a Doctrine of Virtue, and a Doctrine of Duties, treating them in this sequence in order to reflect what he takes to be the greater fundamentalness of goods over virtues and of virtues over duties.

Generally speaking, though, Schleiermacher's ethics lectures are not a great success. They form an unholy mixture of, not only ethics in the usual sense, but also political philosophy, metaphysics, epistemology, and philosophy of mind; lurch back and forth between claims of startling dubiousness and claims of startling banality (with too little in between); and stick all this together with a thick stain of obscurantism and a thin varnish of systematicity. One is left with the impression that, having put the critical phase of his work in ethics behind him, Schleiermacher found that he did not really have enough constructive to say about ethics to fill up the hours in the lecture hall.

9. Political and Social Philosophy

Schleiermacher's political and social philosophy is scattered through a considerable number of works from different periods. Its most systematic, but not necessarily most interesting, statement is found in his lectures on the theory of the state, delivered between 1808-9 and 1833.

Concerning international politics, Schleiermacher's fundamental position is a thoroughly Herderian one: a cosmopolitan commitment to equal moral respect for all peoples in their diversity. This position is already articulated in *On What Gives Value to Life* (1792-3); it is central to *On Religion* (1799) and the *Soliloquies* (1800), in the form of a commitment to the Herderian ideal of “humanity”; and it survives in later works (e.g. in the 1831 lecture on Leibniz's idea of a universal language).

Concerning domestic politics: Schleiermacher was always somewhat reticent about fundamental constitutional questions. To judge from his early enthusiasm for the French Revolution, and his republican-democratic model of an ideal church in *On Religion*, the early Schleiermacher was strongly attracted to republicanism and democracy (like Herder). However, his later position -- while it does still make consent a conditio sine qua non of any genuine state -- is more sympathetic to aristocratic and monarchical forms of government. Thus in his lectures on the theory of the state from 1829-33 he argues that whereas smaller and “lower” states are naturally democratic, large and “higher” ones are naturally aristocratic or

monarchical.

However, Schleiermacher's domestic politics is more consistently radical in another respect: liberalism. (Here again he is indebted to Herder.) Already in 1799 the essay *Toward a Theory of Sociable Conduct* argues that there should be a sphere of free (by which Schleiermacher means especially: state-free) social interaction, in order to make possible the development and communication of individuality, and *On Religion* argues strongly against state-interference in religion, making the liberation of religion from such interference a fundamental part of a program for developing individualism in religion, and diagnosing some of the worst vices of current churches and religion in terms of such interference. This liberalism remains prominent in the ethics lectures of 1812-13, which add to the positions just mentioned a proscription of state interference in the universities. And it is still central to Schleiermacher's political thought in his (otherwise more conservative) late lectures on the theory of the state from 1829-33, in which he argues that the three spheres of sociality, religion, and science (e.g. the universities) lie beyond the legitimate power of the state, and critically notes that the current (Prussian) state falls short of this ideal. Schleiermacher's reasons for his broad liberalism are severalfold, but a fundamental one is the need to free up a domain in which the basic good of *individuality* can develop.

Schleiermacher devotes especially close attention to the question of religion's proper relation to the state (and to other socio-political institutions). As was mentioned, in *On Religion* he has two main reasons for wanting to see religion freed from state interference: first, because he values individualism in religion, the free development of a multiplicity of forms of religion; and second, because he sees state-interference as corrupting the nature of religion by, for example, attracting the wrong sorts of people into leadership positions within the church (men with worldly skills and motives rather than religious ones) and foisting alien political functions onto religious mysteries such as baptism and marriage. He argues that the true socio-political center of religion should instead be the family (a position which he later goes on to illustrate in *Christmas Eve* (1806), a work which depicts in a literary way a sort of ideal interweaving of (Christian) religion with family life).

One especially interesting case to which he applied his general insistence on the freedom of religion from state-interference was that of Prussia's Jews. (Once again, Herder had set the tone here -- both by developing a very sympathetic interpretation of ancient Judaism and by forcefully criticizing modern anti-semitism.) In an early work on the subject of Jewish emancipation in Prussia, *Letters on the Occasion of the Political-Theological Task and the Open Letter of Jewish Householders* (1799), Schleiermacher argues that Jews should receive full citizenship and civil rights, provided only that they compromise in their religious observances to a point enabling them to meet their duties to the state, and that they give up such politically threatening commitments as those to a coming messiah and to their status as a separate nation. He argues that Jews should not have to resort to the expedient of baptism as a means for achieving citizenship and civil rights (as some (Jewish) contemporaries had proposed), on the grounds that this expedient would be detrimental both to the Jews and their religion and to Christianity. In the latter connection his main expressed concern is that it would further water down an already rather watery church. But another concern is clearly that it would in effect amount to yet more interference by the state in a religious mystery (baptism). It is significant to note that Schleiermacher takes this strikingly liberal position concerning the Jews despite himself being rather critical of Judaism as a religion: in *On Religion*

he argues that Reimarus's conception that there are deep continuities between Judaism and Christianity is mistaken, and that although Judaism was a beautiful religion in its day it has long since become corrupted and is now effectively moribund (unlike vibrant Christianity).

A further important aspect of Schleiermacher's socio-political philosophy, especially in its earlier phases, is his proto-feminism (in which he is strongly influenced by Friedrich Schlegel, but also by Herder, who was arguably the real pioneer in this area). This proto-feminism has several sides. First, Schleiermacher encourages women to strive for goods which have traditionally been the monopoly of men. For example, in his short *Idea for a Catechism of Reason for Noble Ladies* (published in the *Athenaeum*) he enjoins women, “Let yourself covet men's culture, art, wisdom, and honor.” Second, as a special case of this, he encourages women to seek sexual fulfillment, and to free themselves from inhibitions about discussing sex. This is one of the central themes of his *Confidential Letters Concerning Friedrich Schlegel's Lucinde* (1800). Third, he sees women as a source of valuable moral and intellectual resources for the benefit and improvement of society as a whole. One example of this is their natural aversion to the sorts of insensitivity and violence to which men are commonly prone, and their potential ability to restrain instead of permitting or even encouraging these. In this vein the *Idea for a Catechism* enjoins women, “You should not bear false witness for men. You should not beautify their barbarism with words and works.” Another (more historically specific) example, discussed in *Toward a Theory of Sociable Conduct*, is the ability of women, due to their broad educations but their freedom from the narrow confines of the professions, to direct social conversation away from limited professional concerns towards deeper and more broadly shared ones (Schleiermacher is thinking especially of the hostesses of salons of the period). Another example can be seen in an argument which Schleiermacher develops especially in his ethics lectures to the effect that women are by nature more attuned to recognizing and respecting individuality, whereas men are more attuned to recognizing and respecting abstract generalizations, and that accordingly one of the key functions of marriage is to bring about a valuable blending of these (equally important) intellectual-moral qualities in each partner. (It should be noted, however, that Schleiermacher later on tended to be more conservative in his views about women.)

It is perhaps worth underscoring that in its broad cosmopolitan concern for other peoples, for Jews, and for women Schleiermacher's socio-political philosophy was continuing a paradigm which was above all the achievement of a single predecessor: Herder.

One final feature of Schleiermacher's socio-political philosophy, especially prominent in the works from 1799-1800, is a broad critique of some central modern socio-economic institutions and a set of proposals for remedying their negative effects. (The *Soliloquies* casts this critique in the form of an attack on the self-satisfaction of the Enlightenment which is very reminiscent of Herder's attack in *This Too a Philosophy of History*.) Three parts of Schleiermacher's case are particularly noteworthy: First, in *Toward a Theory of Sociable Conduct* he implicitly criticizes modern division of labor for the way it blinkers people, inhibiting their development of their own individuality and their sense for the individuality of others. His solution here is the development of a sphere of “sociability,” i.e. a sphere of free conversation and social intercourse, in which such one-sidedness can be overcome. Second, in *On Religion* he criticizes the deadening repetitive labor typical of modern economies as an obstacle to spiritual, and in particular religious, self-development. His solution here is mainly a hope that advances in technology will free

people from the sort of labor in question. Third, in *On Religion* and the *Soliloquies* he criticizes the hedonism, utilitarianism, and materialism of the modern age for preventing people's spiritual and religious self-development. His main solution here is the sort of revival of a vibrant religious and moral life for which *On Religion* and the *Soliloquies* argue.

10. Philosophy of Religion

Schleiermacher's most important and radical work in the philosophy of religion is his *On Religion: Speeches to Its Cultured Despisers* of 1799. (Later editions of this work and his later theological treatise *The Christian Faith* strive for greater Christian orthodoxy, and are consequently as a rule less interesting from a philosophical point of view.)

As its title implies, the project of *On Religion* is to save religion from the contempt of enlightenment and especially romantic skeptics about religion, “its cultured despisers.” At least where the romantics were concerned, the work was strikingly successful in this regard, in the sense that several of them, including Friedrich Schlegel, did in fact turn to religion in the years following the book's publication (though admittedly not to quite the sort of religion that Schleiermacher had envisaged). Schleiermacher's later philosophy of religion is similarly motivated. In his 1829 open letters to Lücke he especially stresses the pressing need to defend religion against the twin threats posed to it by modern natural science and modern historical-philological scholarship.

This project of defending religion against educated skeptics is reminiscent of Kant's similarly motivated critical philosophy. Schleiermacher is also sympathetic to Kant's general strategy of “deny[ing] knowledge in order to make room for faith” in connection with religious matters (*Critique of Pure Reason*, Bxxx), and in particular to Kant's attack on traditional proofs of the existence of God; Schleiermacher himself denies that religion is a form of knowledge or can be based on metaphysics or science. However, as can already be seen from his early unpublished essays *On the Highest Good* (1789) and *On What Gives Value to Life* (1792-3), Schleiermacher's strategy is in other respects defined more by opposition to than by agreement with Kant's. In particular, Schleiermacher sharply rejects Kant's alternative *moral* proof of an otherworldly God and human immortality (Kant's proof of them by showing them to be necessary presuppositions of morality); for Schleiermacher religion can no more be based on morality than on metaphysics or science.

As this stance already suggests, Schleiermacher has a large measure of sympathy with the skeptics about religion whom he means to answer. But the early Schleiermacher's sympathy with them also goes far deeper than this. In *On Religion* he is skeptical about the ideas of God and human immortality altogether, arguing that the former is merely optional (to be included in one's religion or not depending on the nature of one's imagination), and that the latter is positively unacceptable. Moreover, he diagnoses the modern prevalence of such religious ideas in terms of the deadening influence exerted by modern bourgeois society and state-interference in religion. He reconciles this rather startling concession to the skeptics with his ultimate goal of defending religion by claiming that such ideas are inessential to religion. This stance strikingly anticipates such more recent radical religious positions as Mauthner's “godless mysticism.”

(Schleiermacher's later religious thought tended to backtrack on this radicalism, however, restoring God and even human immortality to a central place in religion.)

This naturally leaves one wondering what the content and the epistemological basis of religion *are* for Schleiermacher. As can already be seen from the 1793-4 essays *Spinozism* and *Brief Presentation of the Spinozistic System*, and then again from *On Religion*, the early Schleiermacher follows Spinoza in believing in a monistic principle that encompasses everything, a “one and all.” However, he also modifies Spinoza's conception in certain ways, partly under the influence of Herder (who is mentioned by name in the essays on Spinoza). In particular, whereas Spinoza had conceived his monistic principle as a substance, Schleiermacher follows Herder in thinking of it rather as an original force and the unifying source of a multiplicity of more mundane forces. (Later on Schleiermacher distanced himself from this neo-Spinozistic position. He explicitly denied that he was a follower of Spinoza. And accordingly, in the dialectics lectures he argued that there was an even higher “transcendental ground” *beyond* the Spinozist *natura naturans* or the Herderian highest force. His main motive behind this change of position seems to have been a desire to avoid the heavily charged accusations of Spinozism and pantheism -- which is hardly an impressive motive philosophically speaking.)

So much for the content of religion, as Schleiermacher envisages it. What about its epistemological basis? As was mentioned, for Schleiermacher religion is based neither on theoretical knowledge nor on morality. According to *On Religion*, it is instead based on an intuition or feeling of the universe: “Religion's essence is neither thinking nor acting, but intuition and feeling. It wishes to intuit the universe.”

The term “intuition” here is both revealing and problematic. As Kant had defined it, “*intuition* is that through which [a mode of knowledge] is in immediate relation to [objects]” (*Critique of Pure Reason*, A19). So part of what Schleiermacher means to convey here is some sort of immediate cognitive relation to some sort of object, namely the universe as a single whole. On the other hand, the term “intuition” also imported certain implications which Schleiermacher in fact wanted to avoid. In particular, Kantian pure or empirical intuition required the addition of concepts in order to constitute any sort of insight (“intuitions without concepts are blind” -- *Critique of Pure Reason*, A51), whereas Schleiermacher had in mind a sort of insight unmediated by concepts. In the later editions of *On Religion* he therefore retreated from speaking of “intuition” in connection with religion (instead reserving the term for science), and instead spoke simply of “feeling.” In accordance with this change, *The Christian Faith* then went on to define religion more specifically as a feeling of absolute dependence, or what Schleiermacher also described in his open letters to Lücke as the immediate consciousness of “an immediate existence-relationship.”

This epistemological position looks suspiciously like philosophical sleight-of-hand, however. “Feelings” can be of two very different sorts: on the one hand, non-cognitive “feelings,” such as physical pains and pleasures; on the other hand, “feelings” which incorporate cognitions or beliefs, for example a feeling that such and such is the case. Whereas the possession and awareness of non-cognitive feelings such as pains and pleasures may indeed be conceptually unmediated, beyond mediation by reasons for or against, and in a sense infallible, the possession and awareness of feelings which incorporate cognitions or beliefs, for instance the feeling that such and such is the case, does require conceptual mediation, is subject to reasoning for or against, and is fallible. As can be seen from the neo-Spinozistic content that

Schleiermacher's religious intuition or feeling was originally supposed to have, from his original characterization of it as an intuition in the Kantian sense of an immediate cognitive relation to an object, from his later characterization of it as representing “an immediate existence-relationship,” and so forth, he does not mean religious feeling to be merely non-cognitive, but to incorporate some sort of cognition or belief. However, he also helps himself to the apparent epistemological advantages which belong only to non-cognitive feelings: non-mediation by concepts, transcendence of reasons for or against, and infallibility. In short, it looks as though his epistemological grounding of religion in “feeling” depends on a systematic confusion of these two crucially different sorts of cases.

One further, and less problematic, aspect of the “feeling” on which Schleiermacher bases religion should also be mentioned: its inclusion of *motivating* force, its self-manifestation in *actions*. The wish to include this aspect was one of Schleiermacher's reasons for supplementing religious “intuition” with “feeling” in the first edition of *On Religion*. And his later work stresses this dimension of religious “feeling” as well.

Turning to some further features of Schleiermacher's philosophy of religion in *On Religion*: He recognizes a potentially endless multiplicity of valid religions, and strongly advocates religious toleration. However, he also arranges the various types of religion in a hierarchy, with animism at the bottom, polytheism in the middle, and monotheistic or otherwise monistic religions at the top. This hierarchy makes reasonably good sense given his fundamental neo-Spinozism.

More problematic, however, is a further elaboration of this hierarchy which he introduces: he identifies Christianity as the highest among the monotheistic or monistic religions, and in particular as higher than Judaism. His rationale for this is that Christianity introduces “the idea that everything finite requires higher mediation in order to be connected with the divine” (i.e. the higher mediation of Christ). But this looks contrived. Even if one granted that “higher mediation” was a good thing, do not other monotheistic religions such as Judaism share this putative advantage as well, namely in the form of their prophets? And if the answer is No because prophets are not themselves divine, then why is the mediator's divinity supposed to be such a great advantage?

Moreover, Schleiermacher remarks on the distinctively *polemical* nature of Christianity, the extraordinary extent to which Christianity's religious and moral standpoint is defined by a hostile opposition to other standpoints, and even to dissenting positions within Christianity itself. This is an extremely insightful observation. (For example, one thinks in this connection of Nietzsche's brilliant explanation of Christian values as a deliberate inversion of Greco-Roman values, and of the revealing fact that the Christian word “demon” had earlier been the Greeks' most generic word for a deity; also of the bloody early internal history of Christianity, the Crusades, the Inquisition's treatment of Jews and witches, and similar horrors -- all of which only stopped (or receded) when Christianity became politically impotent in the modern period.) But then, how can a proponent of religious pluralism and toleration like Schleiermacher consistently see this striking trait of Christianity as anything but a serious vice?

On the (flimsy) basis of this perception of Christianity's superiority, Schleiermacher is moved to attempt to reconcile his neo-Spinozism with traditional Christian doctrines as far as possible. This project begins in a modest way in *On Religion*, where for example he works to salvage the Christian doctrine of miracles

in the modified form of a doctrine which includes *all* events as miracles (insofar as viewed from a religious perspective). A similar project is pursued in a much more elaborate (and tedious) way in *The Christian Faith*.

Bibliography

Primary texts

There are two main editions of Schleiermacher's works:

- *Gesamtausgabe der Werke Schleiermachers in drei Abteilungen* (Berlin, 1838-)
- *Kritische Gesamtausgabe* (Berlin / New York, 1984-)
  The latter edition will eventually supersede the former, but is still far from complete.

In addition to the above, the following editions are especially important for philosophers:

- *F. Schleiermacher: Hermeneutik. Nach den Handschriften neu herausgegeben und eingeleitet von Heinz Kimmerle* (second edition, Heidelberg, 1974)
- *Schleiermacher: Hermeneutik und Kritik. Herausgegeben und eingeleitet von Manfred Frank* (seventh edition, Frankfurt am Main, 1999)
- *Friedrich Schleiermacher: Dialektik. Herausgegeben und eingeleitet von Manfred Frank*, 2 vols. (Frankfurt am Main, 2001)
  Note that the above two editions of Schleiermacher's hermeneutics (like their respective translations below) contain different material and should be used together.

Since the publication of the old *Gesamtausgabe* a number of improved editions of other parts of Schleiermacher's corpus have also been published (but not yet superseded by the new *Kritische Gesamtausgabe*).

One further significant resource:

- *Aus Schleiermachers Leben in Briefen*, 4 vols., ed. L. Jonas and W. Dilthey (Berlin, 1858-)

Translations

- *Hermeneutics: The Handwritten Manuscripts*, ed. J. Duke and J. Forstman (Atlanta, 1986)
- *Hermeneutics and Criticism*, ed. A. Bowie (Cambridge, 1998)
- ”On the Different Methods of Translation,” in *German Romantic Criticism*, ed. A.L. Willson (New York, 1982)
- *Dialectic or The Art of Doing Philosophy*, ed. T.N. Tice (Atlanta, 1996) [This contains Schleiermacher's first lecture notes on dialectics from 1811.]

- *Friedrich Schleiermacher's "Toward a Theory of Sociable Conduct" and Essays on Its Intellectual-Cultural Context*, ed. R.D. Richardson (Lewiston, New York, 1995)
- *On the Highest Good*, ed. H.V. Froese (Lewiston, New York, 1992)
- *On What Gives Value to Life*, ed. E. Lawler and T.N. Tice (Lewiston, New York, 1995)
- *On Freedom*, ed. A.L. Blackwell (Lewiston, New York, 1992)
- *Soliloquies*, ed. H. Friess (Chicago, 1957)
- *On Religion: Speeches to Its Cultured Despisers*, ed. R. Crouter (Cambridge, 1988)
- *Christmas Eve: Dialogue on the Incarnation*, ed. T.N. Tice (San Fransisco, 1990)
- *The Life of Jesus*, ed. J. Verheyden (Philadelphia, 1975)
- *The Christian Faith*, ed. H.R. Mackintosh and J.S. Stewart (Edinburgh, 1999)
- *On the "Glaubenslehre,"* ed. J. Duke and F. Fiorenza (Atlanta, 1981) [This is a translation of Schleiermacher's two 1829 open letters to Lücke.]
- *Schleiermacher on Workings of the Knowing Mind*, ed. R.D. Richardson (Lewiston, New York, 1998) [This includes a translation of Schleiermacher's 1799 review of Kant's *Anthropology*.]

Secondary Literature in German

Life and Works

- W. Dilthey, *Leben Schleiermachers*, 2 vols., ed. M. Redeker (Berlin, 1966)
- R. Haym, *Die Romantische Schule* (Berlin, 1914), ch. 3
- G. Scholtz, *Die Philosophie Schleiermachers* (Darmstadt, 1984)

Philosophy of Mind

- C. von Sigwart, *Schleiermachers psychologische Voraussetzungen, insbesondere die Begriffe des Gefühls und der Individualität* (reprint, Darmstadt, 1974)

Hermeneutics

- A. Böckh, *Enzyklopädie und Methodologie der philologischen Wissenschaften* (reprint, Darmstadt, 1966)
- M. Frank, *Das individuelle Allgemeine: Textstrukturierung und -interpretation nach Schleiermacher* (Frankfurt am Main, 1985)
- M. Frank, *Das Sagbare und das Unsagbare: Studien zur deutsch-französischen Hermeneutik und Texttheorie* (Frankfurt am Main, 1990)
- H. Kimmerle, *Die Hermeneutik Schleiermachers im Zusammenhang seines spekulativen Denkens* (dissertation, Heidelberg, 1957)
- H. Patsch, "Friedrich Schlegels 'Philosophie der Philologie' und Schleiermachers frühe Entwürfe zur Hermeneutik," in *Zeitschrift für Theologie und Kirche*, no. 63 (1966)
- J. Wach, *Das Verstehen. Grundzüge einer Geschichte der hermeneutischen Theorie im 19. Jahrhundert* (Tübingen, 1926)

- [See also the editors' introductions to the two German editions of the hermeneutics lectures cited above.]

Aesthetics

- T. Lehnerer, *Die Kunsttheorie Friedrich Schleiermachers* (Stuttgart, 1987)
- R. Odebrecht, *Schleiermachers System der Ästhetik* (Berlin, 1932)

Dialectics

- D. Burdorf and R. Schmücker eds., *Dialogische Wissenschaft* (Paderborn, 1998)
- F. Kaulbach, “Schleiermachers Idee der Dialektik,” in *Neue Zeitschrift für systematische Theologie und Religionsphilosophie*, 1968, Band 10, Heft 3
- F. Wagner, *Schleiermachers Dialektik. Eine kritische Interpretation* (Gütersloh, 1974)
- G. Wehrung, *Die Dialektik Schleiermachers* (Tübingen, 1920)
- [M. Frank's introduction to his edition of the dialectics lectures cited above is also very helpful.]

Secondary Literature in French

Life and Works

- C. Berner, *La philosophie de Schleiermacher* (Paris, 1995)

Hermeneutics

- P. Szondi, “L'herméneutique de Schleiermacher,” in *Poétique*, no. 2 (1970)

Secondary Literature in English

Life and Works

- M. Redeker, *Schleiermacher: Life and Thought* (Philadelphia, 1973)

Hermeneutics

- A. Bowie, *From Romanticism to Critical Theory. The Philosophy of German Literary Theory* (London, 1997)
- W. Dilthey, *Hermeneutics and the Study of History* (Princeton, 1985), esp. for “Schleiermacher's Hermeneutical System in Relation to Earlier Protestant Hermeneutics” and “The Rise of Hermeneutics”
- M.N. Forster, “Schleiermacher's Hermeneutics: Some Problems and Solutions,” (forthcoming,

proceedings of the 2000 international Schleiermacher conference held at Drew University)

- H.G. Gadamer, *Truth and Method* (New York, 1982), part 2
- E.D. Hirsch Jr., *Validity in Interpretation* (New Haven, 1967)
- R.E. Palmer, *Hermeneutics. Interpretation Theory in Schleiermacher, Dilthey, Heidegger, and Gadamer* (Evanston, 1969)
- P. Ricouer, “The Task of Hermeneutics,” *Philosophy Today*, no. 17 (1973)
- P. Ricouer, “Schleiermacher's Hermeneutics,” *The Monist*, no. 60 (1977)
- [The introductory materials in the two English-language editions of Schleiermacher's hermeneutics cited above are also helpful.]

Political and Social Philosophy

- K.M. Faull, “Beyond Confrontation? The Early Schleiermacher and Feminist Moral Theory,” in *Friedrich Schleiermacher's “Toward a Theory of Sociable Conduct” and Essays on Its Intellectual-Cultural Context*
- P.E. Guenther-Gleason, *On Schleiermacher and Gender Politics* (Harrisburg, 1994)

Philosophy of Religion

- R.B. Brandt, *The Philosophy of Schleiermacher: The Development of His Theory of Scientific and Religious Knowledge* (Westport, 1968)
- J.A. Lamm, *The Living God: Schleiermacher's Theological Appropriation of Spinoza* (University Park, Pennsylvania, 1996)
- R.R. Niebuhr, *Schleiermacher on Christ and Religion* (New York, 1964)
- [Also helpful is R. Crouter's introduction to the translation *On Religion: Speeches to Its Cultured Despisers* cited above.]

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

aesthetics | dialectic | Herder, Johann Gottfried von | hermeneutics | Kant, Immanuel | Kant, Immanuel: moral philosophy | language: philosophy of | mind: philosophy of | religion: philosophy of | Schlegel, Friedrich | Spinoza, Baruch [Benedict]

Michael N. Forster
mnforste@uchicago.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 16, 2002
Content last modified: April 16, 2002

Johann Gottfried von Herder

Johann Gottfried von Herder (1744-1803) is a philosopher of the first importance. This claim depends largely on the intrinsic quality of his ideas (of which this article will try to give an impression). But another aspect of it is his intellectual influence. This has been immense both within philosophy and beyond (much greater than is usually realized). For example, Hegel's philosophy turns out to be an elaborate systematic development of Herderian ideas (especially concerning God, the mind, and history); so too does Schleiermacher's (concerning God, the mind, interpretation, translation, and art); Nietzsche is deeply influenced by Herder (concerning the mind, history, and values); so too is Dilthey (in his theory of the human sciences); even J.S. Mill has important debts to Herder (in political philosophy); and beyond philosophy, Goethe was transformed from being merely a clever but conventional poet into a great artist largely through the early impact on him of Herder's ideas.

Indeed, Herder can claim to have virtually established whole *disciplines* which we now take for granted. For example, it was mainly Herder (not, as is often claimed, Hamann) who established fundamental ideas about an intimate dependence of thought on language which underpin modern philosophy of language. It was Herder who, through the same ideas, his broad empirical approach to languages, his recognition of deep variations in language and thought across historical periods and cultures, and in other ways, inspired W. von Humboldt to found modern linguistics. It was Herder who developed modern hermeneutics, or interpretation-theory, in a form that would subsequently be taken over by Schleiermacher and then more systematically formulated by Schleiermacher's pupil Böckh. It was Herder who, in doing so, also established the methodological foundations of nineteenth-century German classical scholarship (which rested on the Schleiermacher-Böckh methodology), and hence of modern classical scholarship generally. It was arguably Herder who did more than anyone else to establish the general conception and the interpretive methodology of our modern discipline of anthropology. Finally, Herder also made vital contributions to the progress of modern biblical scholarship.

---

1. Life and Works

Johann Gottfried von Herder (1744-1803) was born in Mohrungen in East Prussia. His father was a school teacher and he grew up in humble circumstances. In 1762 he enrolled at the University of Königsberg, where he studied with Kant, who accorded him special privileges because of his unusual intellectual abilities. At this period he also began a lifelong friendship with the irrationalist philosopher Hamann. In 1764 he left Königsberg to take up a school-teaching position in Riga. There he wrote the programmatic essay *How Philosophy Can Become More Universal and Useful for the Benefit of the People* (1765); published his first major work, on the philosophy of language and literature, the *Fragments on Recent German Literature* (1767-8); and also an important work in aesthetics, the *Critical Forests* (1769). In 1769 he resigned his position and travelled -- first to France, and then to Strasbourg, where he met, and had a powerful impact on, the young Goethe. In 1771 Herder won a prize from the Berlin Academy for his best-known work in the philosophy of language, the *Treatise on the Origin of Language* (published 1772). From 1771-6 he served as court preacher to the ruling house in Bückeburg. The most important work from this period is his first major essay on the philosophy of history, *This Too a Philosophy of History for the Formation of Humanity* (1774). In 1776, partly through Goethe's influence, he was appointed General Superintendant of the Lutheran clergy in Weimar, a post he kept for the rest of his life. During this period he published an important essay in the philosophy of mind, *On the Cognition and Sensation of the Human Soul* (1778); a seminal work concerning the Old Testament, *On the Spirit of Hebrew Poetry* (1782-3); his well-known longer work on the philosophy of history, the *Ideas for the Philosophy of History of Humanity* (1784-91); an influential essay in the philosophy of religion, *God. Some Conversations* (1787); a work largely on political philosophy, written in response to the French Revolution, the *Letters for the Advancement of Humanity* (1793-7); a series of *Christian Writings* (1794-8) concerned with the gospels of the New Testament; and two works opposing Kant's critical philosophy, the *Metacritique* (1799) (against the theoretical philosophy of the *Critique of Pure Reason*) and the *Calligone* (1800) (against the aesthetics of the *Critique of Judgment*). In addition to the works mentioned, Herder wrote many others during his career as well.

2. Philosophical Style

In certain ways Herder's philosophical texts are easier to read than others from the period. For example, he avoids technical jargon, his writing is lively and rich in examples rather than dry and abstract, and he has no large, complex system for the reader to keep track of. But his texts also have certain peculiarities which can impede a proper understanding and appreciation of his thought, and it is important to be

alerted to these.

To begin with, Herder's writing often seems emotional and grammatically undisciplined in a way that might perhaps be expected in casual speech but not in philosophical texts. This is intentional. Indeed, Herder sometimes deliberately "roughed up" material in this direction between drafts. When writing in this way he is actually often using grammatical-rhetorical figures which can easily look like mere carelessness to an untutored eye but receive high literary sanction from classical sources and are employed artfully (e.g. anacoluthon). Moreover, he has serious philosophical reasons for writing in this way rather than in the manner of conventional academic prose, including: (1) This promises to make his writing more broadly accessible and interesting to people -- a decidedly non-trivial goal for him, since he believes it to be an essential part of philosophy's vocation to have a broad social impact. (2) One of his central theses in the philosophy of mind holds that thought is not and should not be separate from volition, or affect, that types of thinking which aspire to exclude affect are inherently distorting and inferior. Standard academic writing has this vice, whereas spontaneous speech, and writing which imitates it, do not. (3) Herder is opposed to any grammatical or lexical straightjacketing of language, any slavish obedience to grammar books and dictionaries. In Herder's view, such straightjacketing is inimical, not only to linguistic creativity and inventiveness, but also (much worse), because thought is essentially dependent on and confined in its scope by language, thereby to creativity and inventiveness in thought itself.

Another peculiarity of Herder's philosophy is its *unsystematic* nature. This is again deliberate. For Herder is largely hostile towards systematicity in philosophy (a fact reflected both in explicit remarks and in many of his titles: *Fragments* . . . , *Ideas* . . . , etc.). He is in particular hostile to the ambitious sort of systematicity aspired to in the tradition of Spinoza, Wolff, Kant, Fichte, Schelling, and Hegel: the ideal of a theory whose parts form and exhaust some sort of strict overall pattern of derivation. He has compelling reasons for this hostility: (1) He is very skeptical that such systematic designs can be made to work (as opposed to creating, through illicit means, an *illusion* that they do so). (2) He believes that such system-building leads to a premature closure of inquiry, and in particular to the disregarding or distorting of new empirical evidence. Scrutiny of such systems amply bears out these concerns. Herder's well-grounded hostility to this type of systematicity established an important countertradition in German philosophy (which subsequently included e.g. F. Schlegel, Nietzsche, and Wittgenstein).

On the other hand, unlike Hamann, Herder is in *favor* of "systematicity" in a more modest sense: the ideal of a theory which is self-consistent and maximally supported by argument. He by no means always achieves this ideal (so that interpreting him requires more selectivity and reconstruction than is the case with some philosophers). But his failures to do so are often more apparent than real: First, often when he may seem to be guilty of inconsistency he really is not. For he is often developing philosophical dialogues between two or more opposing viewpoints, in which cases it would clearly be a mistake to accuse him of inconsistency in any usual or pejorative sense; and (less obviously) in many other cases he is in effect still working in this dialogue-mode, only without bothering to distribute the positions among different interlocutors explicitly, and so is again really innocent of inconsistency (examples occur in *How Philosophy* and *This Too*). Moreover, he has serious motives for this method of (implicit) dialogue: (1) Sometimes his motive is simply that when dealing with religiously or politically delicate matters it

permits him to state his views but without quite stating them as his own and therefore without inviting trouble. But there are also philosophically deeper motives: (2) He takes over from the precritical Kant an idea (inspired by ancient skepticism) that the best way for the philosopher to pursue the truth is by setting contrary views on a subject into opposition with one another in order to advance towards, and hopefully attain, the truth through their mutual testing and modification. (3) Also, he develops an original variant of that idea on the socio-historical plane: analogously, the way for humankind as a whole to attain the elusive goal of truth is through an ongoing contest between opposing positions, in the course of which the best ones will eventually win out (this idea anticipates, and inspired, a central thesis of J.S. Mill's *On Liberty*). This yields a further motive for the dialogue-method (even where it does not lead Herder himself to any definite conclusion), in effect warranting the rhetorical question, And what does it matter to the cause of humankind and its discovery of truth whether those various opposing positions are advanced by different people or by *the same* person? Second, Herder's appearance of neglecting to give arguments is often, rather, a principled rejection of arguments *of certain sorts*. For example, he has a general commitment to empiricism and against apriorism in philosophy which leads him to avoid familiar sorts of apriorist arguments in philosophy; and a commitment to non-cognitivism in ethics which leads him to refrain from familiar sorts of cognitivist arguments in ethics.

3. General Program in Philosophy

Hamann's influence on Herder's best thought has been greatly exaggerated. But Kant's was early, fundamental, and enduring. However, the Kant who influenced Herder in this way was the *pre*critical Kant of the early and middle 1760's, not the critical Kant (against whom Herder later engaged in the -- distracting and rather ineffective -- public polemics mentioned above). Some of Kant's key positions in the 1760's, sharply contrasting with those which he would later adopt in the critical period, were: a (Pyrrhonist-influenced) skepticism about metaphysics; a form of empiricism; and a (Hume-influenced) non-cognitivism in ethics. Herder took over these positions in the 1760's and retained them throughout his career. It should by no means be assumed that this Herderian debt to the *early* Kant is a debt to a philosophically *inferior* Kant; a good case could be made for the very opposite.

Herder's 1765 essay *How Philosophy* is a key text for understanding both his debt to Kant and the broad orientation of his philosophy. The essay was written under strong influence from Kant, especially, it seems, Kant's 1766 essay *Dreams of a Spirit Seer*, which Kant sent Herder before its publication.

Herder's essay answers a prize question set by a society in Bern: "How can the truths of philosophy become more universal and useful for the benefit of the people?" This question is conceived in the spirit of the *Popularphilosophie* that was competing with school-philosophy at the time. Kant himself tended to identify with *Popularphilosophie* at this period, and Herder's selection of this question shows him doing so as well, though in his case the identification would last a lifetime. Philosophy should become relevant and useful for the people as a whole -- this is a basic ideal of Herder's philosophy.

Largely in the service of this ideal, Herder's essay argues for two sharp turns in philosophy, turns which would again remain fundamental throughout the rest of his career. The first involves a rejection of

traditional metaphysics, and closely follows an argument of Kant's in *Dreams of a Spirit Seer*. Herder's case is roughly this: (1) Traditional metaphysics, by undertaking to transcend experience (or strictly, and a little more broadly, "healthy understanding," which includes, in addition to empirical knowledge, also ordinary morality, intuitive logic, and mathematics), succumbs to unresolvable contradictions between claims, and hence to the Pyrrhonian skeptical problem of an equal plausibility on both sides requiring a suspension of judgment. Moreover (Herder adds in the *Fragments*), given the truth of a broadly empiricist theory of concepts, much of the terminology of traditional metaphysics turns out to lack the basis in experience that is required in order even to be meaningful, and hence is meaningless (the illusion of meaningfulness arising through the role of *language*, which spins on, creating illusions of meaning, even after the empirical conditions of meaning have been left behind). (2) Traditional metaphysics is not only, for these reasons, useless; it is also *harmful*, because it distracts its participants from the matters which should be their focus: empirical nature and human society. (3) By contrast, empirical knowledge (or strictly, and a bit more broadly, "healthy understanding") is free of these problems. Philosophy should therefore be based on and continuous with this.

Herder's second sharp turn concerns ethics. Here he remains indebted to Kant, but also goes further beyond him. Herder's basic claims are these: (1) Morality is fundamentally more a matter of sentiments than of cognitions (Herder's sentimentalism is not crude, however; in subsequent works he stresses that cognition plays a large role in morality as well). (2) Cognitivist theories of morality -- of the sort espoused in this period by Rationalists like Wolff, but also by many other philosophers before and since (e.g. Plato and the critical Kant) -- are therefore based on a mistake, and so useless as means of moral enlightenment or improvement. (3) But (and here Herder's theory moves beyond Kant's), worse than that, they are actually *harmful* to morality, because they *weaken* the moral sentiments on which it really rests. In *This Too* and *On the Cognition* Herder suggests several reasons why: (a) Abstract theorizing weakens sentiments *generally*, and hence moral sentiments in particular. (b) The cognitivists' theories turn out to be so *strikingly* implausible that they bring morality itself into disrepute, people reacting to them roughly along the lines: If this is the best that even the *experts* can say in explanation and justification of morality, then morality must certainly be a sham, and I may as well ignore it and do as I please. (c) Such theories distract people from recognizing, and working to reinforce, the *real* foundations of morality: not an imaginary theoretical insight of some sort, but a set of causal mechanisms for inculcating the moral sentiments. (4) More positively, Herder accordingly turns instead to determining theoretically and promoting in practice just such a set of causal mechanisms. In *How Philosophy* he mainly stresses forms of education and an emotive type of preaching. But he elsewhere also identifies and promotes a much broader set of mechanisms, including: the influence of morally exemplary individuals; morally relevant laws; and literature (along with other art forms). Literature is a special focus of Herder's theory and practice here. He sees it as exerting moral influence in various ways -- e.g. not only through fairly direct moral instruction, but also through the literary perpetuation (or creation) of morally exemplary individuals (e.g. Jesus in the New Testament), and the exposure of readers to other people's inner lives and a consequent enhancement of their sympathies for them (a motive behind Herder's publication of *Volkslieder*, or popular songs, from peoples around the world). Herder's development of this theory and practice of moral pedagogy was lifelong and tireless.

4. Philosophy of Language and Interpretation

On the Origin is Herder's best known work in the philosophy of language, but it is in certain respects unrepresentative and inferior in comparison with other works such as the *Fragments* and should not monopolize attention. *On the Origin* is primarily concerned with the question whether the origin of language can be explained in purely natural, human terms or (as Süßmilch had recently argued) only in terms of a divine source. Herder argues for the former position and against the latter. His argument is quite persuasive, especially when supplemented on its positive side from the *Fragments*. But this argument is unlikely to constitute a modern philosopher's main reason for interest in Herder's ideas about language (deriving its zest, as it does, from a religious background that is no longer ours).

Of much greater modern relevance is Herder's theory of interpretation, including his account of the relation between thought and language. This theory is scattered through a large number of works. The following are its main features:

Herder's theory rests on, but also in turn supports, an epoch-making insight: (1) Such eminent Enlightenment philosopher-historians as Hume and Voltaire had still believed that, as Hume put it, "mankind are so much the same in all times and places that history informs us of nothing new or strange." What Herder discovered, or at least saw more clearly and fully than anyone before him, was that this was false, that peoples from different historical periods and cultures often vary *tremendously* in their concepts, beliefs, (perceptual and affective) sensations, and so forth. He also noted that similar, albeit usually less dramatic, variations occur even between individuals within a single culture and period. (These two positions are prominent in many works, including e.g. *On the Change of Taste* (1766) and *On the Cognition*.) Let us call this twofold principle the principle of radical difference.

(2) Given such radical difference, and the gulf that consequently often divides an interpreter's own thought from that of the person he wants to interpret, interpretation is often an extremely *difficult* task, requiring extraordinary efforts from the interpreter. (Herder does not draw the more extreme -- and misguided -- conclusion to which some recent philosophers have been tempted that it would be *impossible*.)

(3) In particular, the interpreter often faces, and needs to resist, a temptation falsely to *assimilate* the thought which he is interpreting to someone else's, especially his own. (This theme is prominent in *This Too*, for example.)

How is the interpreter to meet the challenge? Herder advances three theses concerning thought and language which underpin the rest of his theory of interpretation (and the first two of which in addition founded the philosophy of language as we know it today):

(4) Thought is essentially dependent on, and bounded in scope by, language -- i.e. one can only think if one has a language, and one can only think what one can express linguistically. (Herder, to his credit,

normally refrains from a more extreme, but philosophically untenable, version of this thesis, favored by some of his successors, which *identifies* thought with language, or with inner language.) One consequence of this thesis for interpretation is that an interpreted subject's language is a reliable indicator of the scope of his thought.

(5) Meanings or concepts are not to be equated with the sorts of items, in principle autonomous of language, with which much of the philosophical tradition has equated them -- e.g. the referents involved, Platonic forms, or "ideas" of the sort championed by the British empiricists and others. Instead, they consist in *usages of words*. Consequently, interpretation will essentially involve pinning down word-usages. (Positions (4) and (5) are already embraced by Herder in the 1760's, e.g. in the *Fragments*.)

(6) Conceptualization is intimately bound up with (perceptual and affective) sensation. More specifically, Herder develops a quasi-empiricist theory of concepts according to which sensation is the source and basis of all our concepts, though we are able to achieve non-empirical concepts by means of a sort of metaphorical extension from the empirical ones -- so that all of our concepts ultimately depend in one way or another on sensation. This position carries the important consequence for interpretation that any understanding of a concept must somehow capture its basis in sensation. (For this position, see e.g. *On the Cognition*.)

Herder also has two further basic principles in interpretation-theory:

(7) A principle of *secularism* in interpretation: the interpretation of texts must never rely on religious assumptions or means, even when the texts are sacred ones. (This principle is already prominent in works from the 1760's.)

(8) A principle of *methodological empiricism* in interpretation: interpretation must always be based on, and kept strictly faithful to, exact observations of linguistic (and other relevant) evidence. (This principle is again already prominent in the 1760's, e.g. in the *Fragments* and *On Thomas Abbt's Writings* (1768).)

Beyond this, Herder also advances a further set of interpretive principles which can easily sound much more "touchy-feely" at first hearing (the first of them rather literally so!), but which are in fact on the contrary quite "hard-nosed":

(9) Herder proposes (prominently in *This Too*) that the way to bridge radical difference when interpreting is through *Einfühlung*, "feeling one's way in." This proposal has often been thought (e.g. by Meinecke) to mean that the interpreter should perform some sort of psychological self-projection onto texts. But that is emphatically *not* Herder's idea -- for that would amount to exactly the sort of assimilation of the thought in a text to one's own which he is above all concerned to *avoid*. As can be seen from *This Too*, what he has in mind is instead an arduous process of historical-philological inquiry. What, though, more specifically, is the cash value of the metaphor of *Einfühlung*? It has at least five components: (a) Note, first, that the metaphor implies (once again) that there typically exists radical difference, a gulf, between an interpreter's mentality and that of the interpreted subject, making interpretation a difficult, laborious

task (it implies that there is an "in" there which one must carefully and laboriously "feel one's way into"). (b) It also implies (*This Too* shows) that the "feeling one's way in" should include thorough research not only into a text's use of language but also into its historical, geographical, and social context. (c) It also implies a claim - based on Herder's quasi-empiricist theory of concepts -- that in order to interpret a subject's language one must achieve an imaginative reproduction of his (perceptual and affective) sensations. (d) It also implies that hostility in an interpreter towards the people he interprets will generally distort his interpretation, and must therefore be avoided (though Herder is equally opposed to excessive *identification* with them for the same reason). (e) Finally, it also implies that the interpreter should strive to develop his grasp of linguistic usage, contextual facts, and relevant sensations to the point where this achieves something like the same immediacy and automaticness that it had for a text's original audience when *they* understood the text in light of such things (so that it acquires for him, as it had for them, the phenomenology more of a *feeling* than a cognition).

(10) In addition, Herder insists (e.g. in the *Critical Forests*) on a principle of *holism* in interpretation. This principle rests on several motives, including: (a) Pieces of text taken in isolation are typically ambiguous in various ways (in relation to background linguistic possibilities). In order to resolve such ambiguities, one needs the guidance provided by surrounding text. (b) That problem arises *once* a range of possible linguistic meanings, etc. is established for a piece of text. But in the case of a text separated from the interpreter by radical difference, knowledge of such a range itself presents a problem. How, for example, is he to pin down the range of possible meanings, i.e. possible usages, for a word? This requires collation of the word's actual uses and inference from these to the rules that govern them, i.e. to their usages, a collation which in turn requires looking to remoter contexts in which the same word occurs (other parts of the text, other works in the author's corpus, works by other contemporaries, etc.), or in short: holism. (c) Authors typically write a work *as* a whole, conveying ideas not only in its particular parts but also through the way in which these fit together to make up a whole (either in instantiation of a general genre or in a manner more specific to the particular work). Consequently, readings which fail to interpret the work as a whole will miss essential aspects of its meaning -- not only the ideas in question themselves but also meanings of the particular parts on which they shed important light.

(11) In *On Thomas Abbt's Writings*, *On the Cognition*, and elsewhere Herder makes one of his most important innovations: interpretation must supplement its focus on word-usage with attention to authorial *psychology*. Herder implies several reasons for this: (a) As already mentioned, Herder embraces a quasi-empiricist theory of concepts which implies that in order to understand an author's concepts an interpreter must recapture his relevant sensations. (b) As Quentin Skinner has recently stressed, understanding the linguistic meaning of an utterance or text is only a necessary, not a sufficient, condition for understanding it *tout court* -- one needs, in addition, to establish the author's illocutionary *intentions*. For example, a stranger tells me, "The ice is thin over there"; I understand his linguistic meaning perfectly; but is he simply informing me?, warning me?, threatening me?, joking? . . . (c) Skinner implies that one can determine linguistic meanings prior to establishing authorial intentions. That may *sometimes* be so (e.g. in the example just given). But is it *generally*? Herder implies not. And this seems right, because commonly the linguistic meaning of a formula is ambiguous (in terms of background linguistic possibilities), and in order to identify the relevant meaning one must turn, not only (as previously mentioned) to larger bodies of text, but also to hypotheses, largely derived therefrom,

about the author's intentions (e.g. about the subject-matter he intends to treat). This is a further reason why interpreters must invoke psychology. (d) Herder also (as recently mentioned) stresses that an author often conveys ideas in his work, not explicitly in its verbal expressions, but rather via these and the way in which they are put together to form a textual whole (either in instantiation of a general genre or in a manner more specific to the particular text). It is necessary for the interpreter to capture these ideas both for their own sakes and because doing so is often essential for resolving ambiguities at the level of particular verbal expressions. (e) Herder also refers to the second limb of his doctrine of radical difference -- *individual* variations in mode of thought even within a single culture and period -- as a source of the need for psychological interpretation. Why does any special need arise here? Part of the answer seems to be that when one is dealing, for example, with a concept that is distinctive of a particular author rather than common to a whole culture, one typically faces a problem of *relative paucity and lack of contextual variety* in the actual uses of the word available as empirical evidence from which to infer the rule for use, or usage, constitutive of its meaning. Hence one needs extra help -- and the author's general psychology may provide this.

(12) In the same works Herder also indicates that interpretation, especially in its psychological aspect, requires the use of *divination*. This is another principle which can sound disturbingly "touchy-feely" at first hearing -- in particular, it can sound as though Herder means some sort of prophetic process that has a religious basis and is perhaps even infallible. However, what he really has in mind is (far more sensibly) a process of hypothesis, based on meager empirical evidence, but also going well beyond it, and therefore vulnerable to subsequent falsification, and abandonment or revision if falsified.

(13) Finally, a point concerning the general nature of interpretation and its subject-matter: After Herder, the question arose whether interpretation was a science or an art. Herder does not really address this question. But his inclination would clearly be to say that it is *like* rather than unlike natural science (pace a reading in the German secondary literature which makes him out to be a sort of proto-Gadamer). There are several reasons for this: (a) He assumes (as did virtually everyone at this period) that the meaning of an author's text is as much an *objective* matter as the subjects addressed by the natural scientist. (b) The *difficulty* of interpretation that results from radical difference, and the consequent need for a *methodologically subtle* and *laborious* approach to it in many cases, make for another point of similarity between interpretation and natural science. (c) The essential role of "divination" qua *hypothesis* in interpretation constitutes a further point of similarity between it and natural science. Moreover, (d) even the subject-matter of interpretation is not, in Herder's view, sharply different from that dealt with by natural science: the latter investigates physical processes in nature in order to determine the forces that underly them, but similarly interpretation investigates human verbal (and non-verbal) physical behavior in order to determine the forces that underly *it* (Herder explicitly identifying mental conditions, including conceptual understanding, as "forces").

Herder's theory owes many debts to predecessors. Hamann has commonly been credited with introducing the revolutionary doctrines (4) and (5). But that seems a mistake; Herder was already committed to them in the 1760's, Hamann only later. Herder's debts are rather to a group of authors influenced by Wolff including Abbt and Süßmilch (for (4)) and especially Ernesti (for (1), (2), (5), (7), (8), and (10)). However, Herder's borrowings incorporate important refinements, and his overall

contribution is enormous.

Herder’s theory was taken over almost in its entirety by Schleiermacher in his hermeneutics lectures. Schleiermacher’s theory is also directly influenced by sources shared with Herder, especially Ernesti. But such fundamental and famous positions in it as Schleiermacher’s supplementing of "linguistic" with "psychological" interpretation and identification of "divination" as the method especially of the latter are due entirely to Herder. Moreover, where Herder and Schleiermacher *do* occasionally disagree, Herder’s position is almost always philosophically superior.

5. Philosophy of Mind

Herder in *On the Cognition* and elsewhere also develops an extremely interesting and influential position in the philosophy of mind. The following are some of its central features.

Herder’s position is uncompromisingly *naturalistic* and *anti-dualistic* in intent. In *On the Cognition* he tries to erase the division between the mental and the physical in two specific ways: First, he advances a theory that minds consist in *forces* [Kräfte] which manifest themselves in people’s bodily behavior -- just as physical nature contains forces which manifest themselves in the behavior of bodies. (Note that the general notion of mental "forces" can already be found before Herder in Rationalists such as Wolff and Süßmilch.) He is officially agnostic on the question of what force is, except for conceiving it as something apt to produce a type of bodily behavior, and as a real source thereof (not something reducible thereto). This, strictly speaking, frees the theory from some common characterizations and objections (e.g. vitalism). But it also leaves the theory with enough content to have great virtues over rival theories: (1) The theory ties mental states conceptually to corresponding types of bodily behavior -- which seems correct, and therefore marks a point of superiority over dualistic theories, and indeed over mind-brain identity theories as well. (2) On the other hand, it also avoids *reducing* mental states to bodily behavior -- which again seems correct, in view of such obvious facts as that we can be, and indeed often are, in mental states which happen to receive no behavioral manifestation, and hence marks a point of superiority over outright behaviorist theories.

Second, Herder also tries to explain the mind in terms of the phenomenon of *irritation* [Reiz], a phenomenon recently identified by Haller and exemplified by muscle fibers contracting in response to direct physical stimuli and relaxing upon their removal -- in other words, a phenomenon which, while basically physiological, also seems to exhibit a transition to mental characteristics. There is an ambiguity in Herder’s position here: usually, he wants to resist physicalist reductionism, and so would resist saying that irritation is purely physiological and fully constitutes mental states. But in the 1775 draft of *On the Cognition* and even in parts of the published version this *is* his position. And from a modern standpoint, this is a further virtue of his account (though we would certainly today want to recast it in terms of different, and more complex, physiological processes than irritation). This line of thought might seem at odds with his first one (forces). But it need not be. For, given his official agnosticism about what forces are, it could, so to speak, fill in the "black box" of the hypothesized real forces, namely in physicalist terms. In other words, it turns out (not as a conceptual matter, but as a contingent one) that the real forces

in question consist in physiological processes.

Herder’s philosophy of mind also holds that the mind is a *unity*, that there is no real division between its faculties. This position contradicts theorists such as Sulzer and Kant. However, it is not in itself new with Herder (or Hamann), having already been central to Rationalism, especially Wolff. Where Herder (with Hamann) *is* more original is in rejecting the Rationalists’ reduction of sensation and volition to cognition; establishing the unity thesis in an empirical rather than apriorist way; and adding a normative dimension to it -- this is not only how the mind *is* but also how it *ought* to be. This last feature can sound incoherent, since if the mind is this way by its very nature, what sense is there in prescribing to people that it should be so rather than otherwise? But Herder’s idea is in fact the coherent one that, while the mind is indeed this way by its very nature, people sometimes behave as though one faculty could be abstracted from another, and try to effect that, and this then leads to various malfunctions, and should therefore be avoided.

Herder’s whole position on the mind’s unity rests on three more specific doctrines of intimate mutual involvements between mental faculties, and of malfunctions that arise from striving against these, doctrines which are in large part empirically motivated and hence lend the overall position a sort of empirical basis:

A first concerns the relation between thought and language: Not only does language of its very nature express thought (an uncontroversial point), but also (as noted earlier) for Herder thought is dependent on and bounded by language. Herder bases this further claim largely on empirical grounds (e.g. about how children’s thought develops with language acquisition). The normative aspect of his position here is that attempts (in the manner of some metaphysics) to cut language free from the constraints of thought or (a more original point) vice versa lead to nonsense.

A second area of intimate mutual involvement concerns cognition and volition, or affects. The claim that volition is and should be based on cognition is not particularly controversial. But Herder also argues the converse, that all cognition is and should be based on volition, on affects (and not only on such relatively anemic ones as the impulse to know the truth, but also on less anemic ones). He is especially concerned to combat the idea that *theoretical work* is or should be detached from volition, from affects. In his view, it never really is even when it purports to be, and attempts to make it so merely impoverish and weaken it. His grounds for this whole position are again mainly empirical.

A third area of intimate mutual involvement concerns thought and sensation. Conceptualization and belief, on the one hand, and sensation, on the other, are intimately connected according to Herder. Thus he advances the quasi-empiricist theory of concepts mentioned earlier, which entails that all our concepts (and hence also all our beliefs) ultimately depend in one way or another on sensation. And conversely, he argues -- anticipating much recent work in philosophy -- that there is a dependence in the other direction as well, that the character of our sensations depends on our concepts and beliefs. Normatively, he sees attempts to violate this interdependence as inevitably leading to intellectual malfunction -- e.g., as already mentioned, metaphysicians’ attempts to cut entirely free from the empirical origin of our

concepts lead to meaninglessness. His grounds for this whole position are again largely empirical.

In a further seminal move Herder also argues that (linguistic) meaning is fundamentally social -- so that thought and other aspects of human mental life (as essentially articulated in terms of meanings), and therefore also the very self (as essentially dependent on thought and other aspects of human mental life, and defined in its specific identity by theirs), are so too. Herder's version of this position seems meant only as an empirically-based causal claim. It has since fathered attempts to generate more ambitious arguments for stronger versions of the claim that meaning -- and hence also thought and the very self -- is socially constituted (e.g. by Hegel, Wittgenstein, Kripke, and Burge). However, it may well be that these more ambitious arguments do not work, and that Herder's version is exactly what should be accepted.

Herder also, in tension though not contradiction with this, holds that (even within a single culture and period) human minds are as a rule deeply *individual*, deeply different from each other -- so that in addition to a generalizing psychology we also need a psychology oriented to individuality. This is an important idea which has strongly influenced subsequent thinkers (e.g. Schleiermacher, Nietzsche, Proust, Sartre, and Manfred Frank). Herder advances it only as an empirical rule of thumb. By contrast, a prominent strand in Schleiermacher and Frank purports to make it an a priori universal truth. But Herder's position is again arguably the more plausible one.

Finally, like predecessors in the Rationalist tradition and Kant, Herder sharply rejects the Cartesian idea of the mind's self-transparency -- instead insisting that much of what occurs in the mind is unconscious, so that self-knowledge is often deeply problematic. This is another compelling position which has had a strong influence on subsequent thinkers.

This whole Herderian philosophy of mind owes much to predecessors, especially in the Rationalist tradition. But it is also in many ways original. The theory is important in its own right. And it also exercised enormous influence on successors (e.g. on Hegel in connection with anti-dualism, the role of physical behavior in mental conditions, faculty-unity, and the sociality of meaning, thought, and self; on Schleiermacher in connection with anti-dualism and faculty-unity; and on Nietzsche in connection with the interdependence of cognition and volition, or affects, the individuality of the mind and the need for an individualistic psychology, and the mind's lack of self-transparency).

6. Aesthetics

In the *Critical Forests* (1769, though the important fourth part was not published until the middle of the nineteenth century) Herder sets out to argue for the following aesthetic theory: whereas music is a mere succession of objects in time, and sculpture and painting are merely spatial, poetry has a sense, a soul, a force; whereas music, sculpture, and painting belong solely to the senses (to hearing, feeling, and vision, respectively), poetry not only depends on the senses but also relates to the imagination; whereas music, sculpture, and painting employ only *natural* signs, poetry uses *voluntary and conventional* signs. This theory was taken over (with minor modifications) by Schleiermacher in his aesthetics lectures, and it has sometimes been touted as Herder's main achievement in aesthetics. But it is a naive theory, and his real

achievements in aesthetics are other than and contrary to it.

As noted earlier, Herder's philosophy of language is committed to the two doctrines that thought is essentially dependent on and bounded by language, and that meaning is word-usage. This invites certain questions: These doctrines plausibly break with an Enlightenment assumption that thought and meaning are in principle autonomous of whatever material, perceptible expressions they may happen to receive. Following Charles Taylor, we might call such a move one to "expressivism." But what form should expressivism take exactly? Is the dependence of thought and meaning on external symbols strictly one on *language* (in the usual sense of "language")? Or is it not rather a dependence on a broader range of symbolic media including, besides language, also such things as painting, sculpture, and music, so that a person might be able to entertain thoughts which he was not able to express in language but only in some other symbolic medium? Let us call the former position *narrow expressivism* and the latter *broad expressivism*.

Also, is Herder's own position narrow expressivism or broad expressivism? It might seem at first sight that his two doctrines themselves already answer this question in favor of narrow expressivism because of their reference to "language" and "words." However, matters are not quite so simple. For one thing, such terms easily lend themselves to broadened uses which might include media beyond language in the usual sense. For another, precisely such a broadening actually occurs in a philosopher closely connected with Herder: Hamann. In his *Metacritique* (1784), Hamann is just as much verbally committed to the two doctrines in question as Herder. But he embraces broad expressivism. And he does so quite consistently, because he understands the terms "language" and "word" as they occur in the doctrines in unusually broad senses -- for example, he explicitly includes as forms of the "language" on which he says thought depends not only language in the usual sense but also painting, drawing, and music.

Nonetheless, Herder's considered position *is* in fact the narrow expressivism that his two doctrines initially seem to suggest (so that his verbal sharing of these doctrines with Hamann masks an important difference of philosophical position between them).

Moreover, after much wrestling with the subject, Herder eventually developed a particularly compelling version of narrow expressivism. The key work in this connection is the *Critical Forests*. By the time of writing this work, Herder was already committed to the two doctrines mentioned, and, as this would suggest, from the start in the *Critical Forests* he is committed to narrow expressivism. However, his commitment to it is initially unsatisfactory and inconsistent. For one thing, it initially takes the extreme and implausible form of denying to the non-linguistic arts any capacity to express thoughts *autonomously* of language by denying that they can express thoughts *at all*. This is the force of the naive theory described earlier which the work sets out to develop. Adding inconsistency to this unsatisfactoriness, Herder is from the start in the work also committed to saying (more plausibly) that visual art *does* express thoughts -- e.g. he intervenes in a quarrel between Lessing and Winckelmann on the question of whether linguistic art (especially poetry) or visual art (especially sculpture) is expressively superior *in ways which tend to support Winckelmann's case for visual art*. This unsatisfactoriness and inconsistency result from the fact that Herder has not yet realized that it is perfectly possible to reconcile narrow expressivism with the attribution of thoughts to non-linguistic art, namely *by insisting that the thoughts*

expressed by non-linguistic art must be derivative from and bounded by the artist's capacity for linguistic expression. However, by the time Herder writes the later parts of the *Critical Forests*, he has found this solution. Thus in the third part, focusing on a particularly instructive example, he notes that the pictorial representations on Greek coins are typically allegorical. And by the fourth part he is prepared to say something similar about much painting as well, writing, for example, of "the sense, the allegory, the story / history which is put into the whole of a painting." By 1778 he extends this account to sculpture as well. Thus in the *Plastic* of 1778 he abandons the merely sensualistic conception of sculpture that dominated the *Critical Forests* and instead argues that sculpture is essentially expressive of, and therefore needs to be interpreted by, a *soul*, but this no longer forces him into unfaithfulness to his principle that thought is dependent on, and bounded by, language, for he now conceives the thoughts expressed by sculpture to have a linguistic source: "The sculptor stands in the dark of night and gropes towards the forms of gods. *The stories of the poets are before and in him*." Subsequently, in the *Theological Letters* (1780-1) and the *Letters for the Advancement of Humanity*, Herder extends the same solution to music as well.

Herder also in his considered position implies that "non-linguistic" art is dependent on thought and language in another way: In the fourth part of the *Critical Forests* he develops the point (mentioned earlier) that human perception is of its nature infused with concepts and beliefs, and consequently with language -- which of course implies that the same is true of the perception of "non-linguistic" artworks in particular. So "non-linguistic" art is really doubly dependent on thought and language: not only for the thoughts which it *expresses* but also for those which it *presupposes* in perception.

With Herder's achievement of this refined form of narrow expressivism and Hamann's articulation of broad expressivism, there were two plausible but competing theories available. Nineteenth-century theorists (e.g. Hegel, Schleiermacher, and Dilthey) would subsequently be deeply torn between them, and the dispute remains a live and important one today.

Because for Herder thought and language play important roles not only in linguistic but also in "non-linguistic" art, both for him present similar interpretive challenges, requiring similar interpretive solutions. One aspect of this which deserves special emphasis is *genre*.

Herder believes, plausibly, that a work of art is always written or made to exemplify a certain genre, and that it is vital for the interpreter to identify its genre in order to understand it. Herder's basic conception of genre is that it consists in an overall purpose together with certain rules of composition dictated thereby. For Herder, genres are in large measure socially pregiven, but they always play their role in a work via authorial intention (not autonomously thereof), and are not something the individual artist is inexorably locked into but something he can and often does modify.

Why does Herder believe that it is vital to define a work's genre-conception correctly in order to understand the work properly? He has two main reasons (both good ones): First, because an author intends his work to exemplify a certain genre, there will normally be aspects of his meaning in the work which are expressed, not explicitly in any particular part or parts of it, but rather through its intended

exemplification of the genre. For instance, Lessing had argued that the function of Aesop's fables as a genre was to illustrate through a concrete example a universal moral principle, whereas Herder argues that it was rather to illustrate general rules of life, experience, or prudence -- so the full interpretation of any particular fable must include either a universal moral principle (if Lessing is right) or a general rule of life, etc. (if Herder is right). Or to cite a "non-linguistic" case, Herder argues that Egyptian sculpture (unlike Greek) had a function as a genre of expressing certain ideas about death and eternity -- so that the full interpretation of a piece of Egyptian sculpture must include this aspect of meaning deriving from the general genre. Second, correctly identifying the genre is also vitally important for correctly interpreting things which *are* expressed explicitly in parts of a work. Hence, for example, in the *Critical Forests* Herder argues that in order to achieve a proper understanding of "ridiculous" passages in Homer (such as the Thersites episode in *Iliad*, book 2) it is essential to understand them in light of the nature of the whole text and their contribution thereto.

Just as Herder insists on a scrupulous methodological empiricism in interpretation generally, so he insists on it in connection with defining genres in particular. He therefore sharply rejects apriorism here -- both the absolute apriorism of refusing in one's definition of a genre to be guided by the observation of examples at all, and the more seductive relative apriorism of allowing oneself to be guided by the observation of examples but excluding from these particular cases, or even whole classes of cases, to which the resulting genre-conception is to be applied in interpretation. The latter procedure is still disastrous, in Herder's view, because the superficial appearance of a similar genre shared by different historical periods or cultures, or even by different authors within one period and culture, or even by a single author in one work and the same author in another, usually masks vitally important differences. Herder identifies a misguided apriorism in the definition of genres in many areas of interpretation. For example, the essay *Shakespeare* (1773) finds it in the French critics' approach to tragedy, an approach which assumes the universal validity of Aristotelian genre-rules which were originally derived exclusively from ancient tragedies (sometimes even overlooking this empirical derivation), and consequently assumes that they provide an appropriate yardstick for interpreting Shakespearean tragedy, whose genre-conception is in fact quite different. And *This Too* and other pieces find it in Winckelmann's treatment of Egyptian sculpture: Winckelmann implicitly assumes the universal validity of a genre-conception for sculpture which he has derived from the Greeks, namely one dominated by the genre-purpose of a this-worldly portrayal of life and beauty, and he then applies this in the interpretation of Egyptian sculpture, where the genre-conception is in fact quite different, in particular involving a contrary genre-purpose of conveying ideas of death and eternity.

Moreover, Herder stresses that getting questions of genre right is vitally important not only for the correct *interpretation* of artworks, but also for their correct critical *evaluation*. The French critics not only make an *interpretive* mistake when they go to Shakespeare with a genre dogmatically in mind that was not his, but they also, on this basis, make an *evaluative* one: because they falsely assume that he somehow must be aspiring to realize the genre-purpose and -rules which Aristotle found in ancient tragedy, they fault him for failing to realize them, while at the same time they overlook the quite different genre-purpose and -rules to which he really aspires and his success in realizing these. Similarly, Winckelmann not only makes an *interpretive* mistake when he implicitly imputes to the Egyptians a Greek genre-purpose and -rules for sculpture that were not theirs, but he also, on this basis, makes an

evaluative one: because he falsely assumes that the Egyptians somehow must be aspiring to realize the Greek genre-purpose and -rules, he faults them for failing to realize them, and at the same time he overlooks their success in realizing the very different genre-purpose and -rules which they really do aspire to realize.

Nothing has yet been said about *beauty*, which philosophers often think of as the central concern of the philosophy of art. Herder strikingly, and plausibly, argues that, on the contrary, beauty is not nearly as essential to art as it is often taken to be. He makes this point in the *Calligone*, for example, where he argues that art is much more essentially a matter of *Bildung* -- i.e., roughly, cultural formation or education (especially in moral respects).

A further claim which he makes about beauty (both in art and more generally) is that standards of beauty vary greatly from one historical period and culture to another. This is his usual position, from early works such as *On the Change of Taste* to late works such as *Calligone* (where he invokes it against Kant's *Critique of Judgment*). There is also an occasional counterstrand in which he argues for a deeper unity in standards of beauty across historical periods and cultures (e.g. in the *Critical Forests*). However, the former position is his considered one, and seems much the more plausible one.

Finally, in close connection with the point mentioned above that the fundamental role of the arts is one of *Bildung*, Herder in *On the Effect of Poetic Art on the Ethics of Peoples in Ancient and Modern Times* (1778) and in *Calligone* argues more specifically that the fundamental role of the arts both has been historically and moreover should be one of moral character formation.

Herder has a fairly nuanced account of how the arts do and should perform this function. For example, *On the Influence of the Beautiful Sciences on the Higher Sciences* (1781) specifies three ways in which poetry and literature promote moral character formation: First, they do so "through light rules," in other words through conveying ethical principles in explicit or implicit ways. Second, and more important, they do so by presenting in an attractive light good moral examples for people to emulate: "still better, through good examples." Third, they also convey a broad range of practical experience relevant to the formation of moral character which would otherwise have to be acquired, if at all, by the more arduous route of first-hand experience. In *Calligone* Herder also notes the power that music has to affect moral character for good or ill depending on the principles with which it is associated, and the power of visual art to make moral ideals attractive by presenting them blended with physical beauty.

Herder's conception that it should be the primary function of art to form moral character also serves him as a criterion for evaluating artworks. Thus when he observes in *On the Effect* that in contrast to earlier poetry modern poetry has typically lost this function, he means this as a serious criticism of modern poetry. He even applies this criterion as a ground for criticizing certain works by his friends Goethe and Schiller which he considers amoral or immoral in content.

7. Philosophy of History

Herder's philosophy of history appears mainly in two works, *This Too* and the later *Ideas*. His fundamental achievement in this area lies in his development of the thesis mentioned earlier -- contradicting such Enlightenment philosopher-historians as Hume and Voltaire -- that there exist radical mental differences between historical periods, that people's concepts, beliefs, sensations, etc. differ in important ways from one period to another. This thesis is already prominent in *On the Change of Taste* (1766). It had an enormous influence on successors such as Hegel and Nietzsche.

Herder makes the empirical exploration of the realm of mental diversity posited by this thesis the very core of the discipline of history. For, as has often been noted, he takes little interest in the so-called "great" political and military deeds and events of history, focusing instead on the "innerness" of history's participants. This choice is deliberate and self-conscious. Because of it, *psychology and interpretation* inevitably take center-stage in the discipline of history for Herder.

Herder has deep philosophical reasons for this choice, and hence for assigning psychology and interpretation a central role in history. To begin with, he has *negative* reasons directed against traditional political-military history. Why *should* history focus on the "great" political and military deeds and events of the past? There are several possible answers: (1) A first would be that they are fascinating or morally edifying. But Herder will not accept this. For one thing, he denies that mere fascination or curiosity is a sufficiently serious motive for doing history. For another, his antiauthoritarianism, antimilitarism, and borderless humanitarianism cause him to find the acts of political domination, war, and empire which make up the vast bulk of these "great" deeds and events not morally edifying but morally *repugnant*.

This leaves two other types of motivation which might be appealed to for doing the sort of history in question: (2) because examining the course of such deeds and events reveals some sort of overall *meaning* in history, or (3) because it leads to *efficient causal insights which enable us to explain the past and perhaps also predict or control the future*. Herder is again skeptical about these rationales, however. This skepticism is clearest in the *Older Critical Forestlet* (1767-8) where, in criticism of rationale (2), he consigns the task of "the whole ordering together of many occurrences into a plan" not to the historian but to the "creator, . . . painter, and artist," and in criticism of rationale (3), he goes as far as to assert (on the basis of a Hume- and Kant-influenced general skepticism about causal knowledge) that with the search for efficient causes in history "historical seeing stops and prophecy begins." His later writings depart from this early position in some obvious ways, but also in less obvious ways remain faithful to it. They by no means *officially* stay loyal to the view that history has no discernible meaning; famously, *This Too* insists that history does have an overall purpose, and that this fact (though not the *nature* of the purpose) is discernible from the cumulative way in which cultures have built upon one another, and the *Ideas* then tells a long story to the effect that history's purpose consists in its steady realization of "humanity" and "reason." However, Herder clearly still harbors grave doubts just below the surface. This is visible in *This Too* from the work's ironically self-deprecating title; Pyrrhonian-spirited motto; vacillations between several incompatible models of history's direction (progressive?, progressive and cyclical?, merely cyclical?, even regressive?); and morbid dwelling on, and unpersuasive attempt to rebut, the "skeptical" view of history as meaningless "Penelope-work." (A few years later Herder would write that history is "a textbook of the nullity of all human things.") It is also visible in the *Ideas* from the fact that Herder's official account of the purposiveness of history gets contradicted by passages which

insist on the *inappropriateness* of teleological (as contrasted with efficient causal) explanations in history. Herder's official position certainly had a powerful influence on some successors (especially Hegel), but it is this quieter counterstrand of skepticism that represents his better philosophical judgment. Concerning efficient causal insights, Herder's later works again in a sense stay faithful to his skeptical position in the *Older Critical Forestlet* -- but they also modify it, and this time for the better philosophically speaking. The mature Herder does not, like the Herder of that work, rest his case on a *general* skepticism about the role or discernibility of efficient causation in history. On the contrary, he insists that history *is* governed by efficient causation and that we should try to discover as far as possible the specific ways in which it is so. But he remains highly skeptical about the *extent* to which such an undertaking can be successful, and hence about how far it can take us towards real explanations of the past, and towards predicting or controlling the future. His main reason for this skepticism is that major historical deeds and events are not the products of some one or few readily identifiable causal factors (as political and military historians tend to assume), but rather of chance confluences of huge numbers of different causal factors, many of which, moreover, are individually unknown and unknowable by the historian (e.g. because in themselves too trivial to have been recorded, or, in the case of psychological factors, because the historical agent failed to make them public, deliberately misrepresented them, or was himself unaware of them due to the hidden depths of his mind).

Complementing this *negative* case against the claims of traditional political-military history to be of overriding importance, Herder also has *positive* reasons for focusing instead on the "innerness" of human life in history. One reason is certainly just the sheer interest of this subject-matter -- though, as was mentioned, that would not be a sufficient reason in his eyes. Another reason is that his discovery of radical diversity in human mentality has shown there to be a much broader, less explored, and more intellectually challenging field for investigation here than previous generations of historians have realized. Two further reasons are moral in nature: (1) He believes, and plausibly so, that studying people's minds through their literature, visual art, etc. generally exposes one to them at their moral best (in sharp contrast to studying their political-military history), so that there are benefits of moral edification to be gleaned here. (2) He has cosmopolitan and egalitarian moral motives for studying people's minds through their literature, visual art, etc.: (in sharp contrast to studying unedifying and elite-focused political-military history) this promises to enhance our sympathies for peoples and for peoples at all social levels, including lower ones. Finally, doing "inner" history is also an important instrument for our *non*-moral self-improvement: (1) It serves to enhance our self-understanding. One reason for this is that it is by, and only by, contrasting one's own outlook with the outlooks of other peoples that one recognizes what is universal and invariant in it and what by contrast distinctive and variable. Another very important reason is that in order fully to understand one's own outlook one needs to identify its origins and how they developed into it (this is Herder's famous "genetic method," which subsequently became fundamental to the work of Hegel, Nietzsche, and Foucault). (2) Herder believes that an accurate investigation of the (non-moral) ideals of past ages can serve to enrich our own ideals and happiness. This motive finds broad application in Herder. An example is his exploration of past literatures in the *Fragments* largely with a view to drawing from them lessons about how better to develop modern German literature.

Herder's decision to focus on the "innerness" of history's participants, and his consequent emphasis on

psychology and interpretation as historical methods, strikingly anticipated and influenced Dilthey. So too did Herder's rationale for this, as described above, which is indeed arguably superior to Dilthey's, especially on its positive side.

Finally, Herder is also impressive for having recognized, and, though not solved, at least grappled with, a problem that flows from his picture of history (and intercultural comparisons) as an arena of deep variations in human mentality. This is the problem of *skepticism*. He tends to run together *two* problems here: (1) the problem of whether there is any *meaning* to the seemingly endless and bewildering series of changes from epoch to epoch (or culture to culture); (2) the problem that the multiplication of conflicting viewpoints on virtually all subjects that is found in history (or in intercultural comparisons) causes, or at least exacerbates, the ancient skeptic's difficulty of unresolvable disputes forcing one to suspend belief on virtually all subjects. Problem (1) has been discussed. Here it is problem (2) that concerns us. This is a problem that Troeltsch would make much of in the twentieth century. But Herder had already clearly seen it.

Herder is determined to avoid this sort of skepticism. He has two main strategies for doing so, but they are inconsistent with each other, and neither in the end works: His first is to try to defuse the problem at source by arguing that, on closer inspection, there is much more common ground between different periods and cultures than it allows. This strategy plays a central role in the *Ideas*, where in particular "humanity" is presented as a shared ethical value; and it is also present in the *Critical Forests*, where (as mentioned earlier) Herder argues that standards of beauty have an underlying unity. Herder's second strategy is rather to acknowledge the problem fully and to respond with relativism: especially in *This Too* he argues that -- at least where questions of moral, aesthetic, and prudential value are concerned -- the different positions taken by different periods and cultures are equally valid, namely for the periods and cultures to which they belong, and that there can be no question of any preferential ranking between them. The later *Letters* vacillates between these two strategies.

Neither of these strategies is satisfactory in the end. The first, that of asserting deep commonalities, is hopeless (notwithstanding its seemingly eternal appeal to empirically underinformed Anglophone philosophers). It flies in the face of the empirical evidence -- e.g. Herder in this mode sentimentally praises Homer for his "humanity," and thereby lays himself open to Nietzsche's just retort in *Homer's Contest* that what is striking about Homer and his culture is rather their *cruelty*. Moreover, it flies in the face of Herder's own better interpretive judgments about the empirical evidence -- e.g. his observation in *On the Change of Taste* that basic values have not only changed through history but in certain cases actually been inverted (an observation which strikingly anticipates a brilliant insight of Nietzsche's concerning an inversion of ethical values that occurred in antiquity).

Herder's alternative, relativist, strategy, is more interesting, but is not in the end satisfactory either (even concerning values, where its prospects seem best). There are several potential problems with it. One, which is of historical interest but probably not in the end fatal, is this: Hegel in the *Phenomenology of Spirit* and then Nietzsche in his treatment of Christian moral values saw the possibility that one might accept Herder's insight that there were basic differences in values but nonetheless avoid his relativism by subjecting others' values to an *internal* critique, a demonstration that they were internally inconsistent.

For example, Nietzsche (whose version of this idea is the more plausible) traced back such Christian values as forgiveness to a contrary underlying motive of resentment [ressentiment]. However, in order to work, such a response would need to show that the inconsistency was *essential* to the values in question, not merely something contingent that could disappear leaving the values consistently held -- and this it probably cannot do. A more serious problem with the strategy is rather a twofold one, which Nietzsche again saw: First, we cannot in fact sustain such a relativist indifference vis-à-vis others' values. Do we, for example, *really* think that a moral rule requiring the forcible burning of dead men's wives is no better and no worse than one forbidding it? Second, nor does the phenomenon of fundamental value variations *require* us to adopt such an indifference. For, while it may indeed show there to be no universal values, it leaves us with a better alternative to indifference: continuing to hold our values and to judge others' values in light of them *only now in a self-consciously non-universal way*. (As Nietzsche puts it, "My judgment is *my* judgment." Or if we reject Nietzsche's extreme individualism, "Our judgment is *our* judgment," for some less-than-universal *us*.)

8. Political Philosophy

Herder is not usually thought of as a political philosopher. But he was one, and moreover one whose political ideals are more admirable, theoretical stances more defensible, and thematic focus of more enduring relevance than those of any other German philosopher of the period. His most developed treatment of political philosophy occurs late, in a work prompted by the French Revolution of 1789: the *Letters* (including the early draft of 1792, important for its frank statement of his views about domestic politics).

What are the main features of Herder's political philosophy? We should begin with his political *ideals*, first in domestic and then in international politics: In domestic politics, the mature Herder is a liberal, a republican, a democrat, and an egalitarian (this in circumstances where such positions were by no means commonplace, and were embraced at a personal cost). His *liberalism* is especially radical in advocating virtually unrestricted freedom of thought and expression (including freedom of worship). He has several reasons for this position: (1) He feels that such freedom belongs to people's moral dignity. (2) He believes that it is essential for individuals' self-realization. (3) As mentioned earlier, he believes that human beings' capacities for discerning the truth are limited and that it is through, and only through, an ongoing contest between opposing viewpoints that the cause of truth gets advanced. (J.S. Mill would later borrow these considerations -- partly via intermediaries such as von Humboldt -- to form the core of his case for freedom of thought and expression in *On Liberty*.) Herder is also committed to *republicanism and democracy* (advocating a much broader franchise than Kant, for example). He has several reasons for this position, ultimately deriving from an egalitarian concern for the interests of all members of society: (1) He thinks it intrinsically right that the mass of people should share in their government, rather than having it imposed upon them. (2) He believes that this will better serve their *other* interests as well, since government *by* also tends to be government *for*. (3) He in particular believes that it will diminish the warfare that is pervasive under the prevailing autocratic political régimes of Europe, where it benefits the few rulers who decide on it but costs the mass of people dearly. Finally, Herder's *egalitarianism* also extends further. He does not reject class differences, property, or

inequalities of property outright. But he opposes all hierarchical oppression; argues that all people in society have capacities for self-realization, and must receive the opportunity to realize them; and insists that government must intervene to ensure that they do, e.g. by guaranteeing education and a minimum standard of living for the poor.

Concerning international politics, Herder often gets classified as a "nationalist" or (even worse) a "German nationalist." Some other philosophers from the period deserve this slur (e.g. Fichte). But where Herder is concerned it is deeply misleading and unjust. On the contrary, his fundamental position in international politics is a committed *cosmopolitanism*, an impartial concern for *all* human beings. This is a large part of the force of his ideal of "humanity." Hence, for example, in the *Letters* he approvingly quotes Fénelon’s remark, "I love my family more than myself; more than my family my fatherland; more than my fatherland humankind." Moreover, unlike Kant’s cosmopolitanism, Herder’s is genuine. Kant’s cosmopolitanism is vitiated by a set of empirically ignorant and morally inexcusable prejudices which he harbors -- in particular, racism, antisemitism, and misogyny. By contrast, Herder’s is entirely free of these prejudices, which he indeed works tirelessly to combat.

Herder does *also* insist on respecting, preserving, and advancing national groupings. But this is unalarming, for the following reasons: (1) For Herder, this is emphatically something that must be done for *all* national groupings *equally* (not just or especially Germany!). (2) The "nation" in question is not racial but linguistic and cultural (Herder rejects the very concept of race). (3) Nor does it involve a centralized or militaristic state (Herder advocates the disappearance of such a state and its replacement by loosely federated local governments with minimal instruments of force). (4) In addition, Herder’s insistence on respecting national groupings is accompanied by the strongest denunciations of military conflict, colonial exploitation, and all other forms of harm between nations; a demand that nations instead peacefully cooperate and compete in trade and intellectual endeavors for their mutual benefit; and a plea that they should indeed actively work to help each other.

Moreover, Herder has compelling reasons for his insistence on respecting national groupings: (1) The deep diversity of values between nations entails that homogenization is ultimately impracticable, only a fantasy. (2) Such diversity also entails that, to the extent that it *is* practicable, it cannot occur voluntarily but only through external coercion. (3) In practice, attempts to achieve it, e.g. by European colonialism, are moreover coercive from, and subserve, ulterior motives of domination and exploitation. (4) Real national variety is moreover positively valuable, both as affording individuals a vital sense of local belonging and in itself.

It might be objected that all this does not yet really amount to a political *theory* -- such as other philosophers have given, including some of Herder’s contemporaries in Germany. In a sense that is true, but philosophically defensible; in another sense it is false. It is true in this sense: There is indeed no grand metaphysical theory underpinning Herder’s position -- no Platonic theory of forms, no correlation of political institutions with "moments" in a Hegelian Logic, no "deduction" of political institutions from the nature of the self or the will à la Fichte and Hegel, etc. But that is deliberate, given Herder’s skepticism about such metaphysics. And is it not indeed philosophically a good thing? Nor does Herder have any elaborate account purporting to justify the moral intuitions at work in his political position as a

sort of theoretical insight (in the manner of Kant's theory of the "categorical imperative" or Rawls's theory of the "original position," for example). But that is again quite deliberate, given his non-cognitivism in ethics, and his rejection of such theories as both false and harmful. And is he not again right about this, and the absence of such an account therefore again a good thing? Nor is Herder sympathetic with such tired staples of political theory as the state of nature, the social contract, natural rights, the general will, and utopias for the future. But again, he has good specific reasons for skepticism about these things. This, then, is the sense in which the objection is correct; Herder does indeed lack a "political theory" of *these* sorts. But he lacks it *on principle,* and is arguably quite right to do so.

On the other hand, he *does* have a "political theory" of another, and arguably more valuable, sort. For one thing, consistently with his general empiricism, his position in political philosophy is deeply empirically informed. For instance, as can be seen from the *Dissertation on the Reciprocal Influence of Government and the Sciences* (1780), his thesis about the importance of freedom of thought and expression, and the competition between views which it makes possible, for producing intellectual progress is largely based on the historical example of ancient Greece and in particular Athens (as contrasted with societies which have lacked the freedom and competition in question). And in the 1792 draft of the *Letters* he even describes the French Revolution and its attempts to establish a modern democracy as a sort of "experiment" from which we can learn (e.g. whether democracy can be successfully extended to nations much larger than ancient Athens). For another thing, conformably with his general non-cognitivism about morals, he is acutely aware that his political position ultimately rests on moral sentiments -- his own and, for its success, other people's as well. For example, the 10th Collection of the *Letters* stresses the fundamental role of moral "dispositions" or "feelings" as required supports for his political position's realization. As was mentioned, this standpoint absolves him of the need to do certain sorts of theorizing. However, it also leads him to engage in theorizing of another sort, namely theorizing about how, and by what means, people's moral sentiments should be molded in order to realize the ideals of his political position. His discussion of moral "dispositions" in the 10th Collection is an example of such theorizing (concerning the *how* rather than the means; his theorizing about causal means has been sketched earlier in this article). *These* two sorts of theorizing *are* deeply developed in Herder. And they are arguably much more pointful than the sorts which are not.

In short, to the extent that Herder's political philosophy really is theoretically superficial, it is arguably, to borrow a phrase of Nietzsche's, "superficial -- *out of profundity*" (whereas more familiar forms of political philosophy are profound out of superficiality). And in another, more important, sense it is not theoretically superficial at all.

9. Philosophy of Religion

In Herder's day German philosophy was deeply committed to a game of trying to reconcile the insights of the Enlightenment, especially those of modern science, with religion, and indeed with Christianity. Leibniz, Kant, Hegel, Schleiermacher, and many others played this game -- each proposing some new reconciliation or other. Herder was part of this game as well. This was not a good game for philosophers to be playing. But it was only in the nineteenth century that German philosophy found the courage to cut

the Gordian knot and turn from apologetics for religion and Christianity to thoroughgoing criticism of them (prime examples being Marx and Nietzsche). This situation imposes limits on the interest of Herder's philosophy of religion, as on that of the other reconciling philosophers mentioned.

Also, while Herder's philosophy of religion was extremely enlightened and progressive in both his early and his late periods, there was a spell in the middle, the years 1771-6 in Bückeburg, during which he fell into the sort of religious irrationalism more characteristic of his friend Hamann. This happened as the result of what we would today classify as a mild nervous breakdown (documentable from his correspondence at the time), and should be discounted.

Despite these qualifications, Herder did make important contributions to the philosophy of religion -- i.e. important in terms of their influence, their intrinsic value, or both. One of these (important for its influence) is his neo-Spinozism, expounded in *God. Some Conversations* of 1787. In this work he develops a version of "Spinozism" which consciously modifies the original in important respects. He shares with Spinoza the basic thesis of *monism*, and like Spinoza equates the single, all-encompassing principle in question with God. But whereas Spinoza characterized it as *substance*, Herder prefers to characterize it as *force*, or *primal force*. Moreover, this modification involves further ones which Herder finds attractive, including: (1) Spinoza's theory had rejected conceptions of God as a mind, as a being who thinks or has purposes. Given Herder's general philosophy of mind and its identification of the mind with force, his identification of God with force imports a claim that God *is* in fact a mind -- hence in works such as *On the Spirit of Christianity* (1798) he describes God as a *Geist*, a mind or spirit. Accordingly, Herder claims that God *does* think, and even have purposes. (2) Herder believes that Spinoza's original theory contains a residue of objectionable dualism, inherited from Descartes, in its conception of the relation between God's two known attributes, thought and extension (and similarly, in its conception of the relation between finite minds and bodies). By contrast, the conception of God as a force (and of finite minds as likewise forces) overcomes this residual dualism. For forces are of their very nature expressed in extended bodies. From around the time of *God. Some Conversations* until well into the nineteenth century a wave of neo-Spinozism swept through German philosophy: Goethe, Schelling, Hegel, Schleiermacher, and lesser figures such as Hölderlin, Novalis, and F. Schlegel. This wave was mainly the result of Herder's embrace of neo-Spinozism in that work, and took over his modifications of Spinoza's position.

However, Herder's most intrinsically valuable contribution to the philosophy of religion concerns the interpretation of the bible. In this connection, as previously mentioned, he champions a strict *secularism*. This was already his position in the 1760's. At that period he argued vigorously, in the spirit of Galileo, for disallowing revelation any jurisdiction over natural science -- though he did so not in an anti-religious spirit but in the hope and expectation that an autonomous natural science would confirm religion. And he made a parallel case for the autonomy of *interpretation*: Religious assumptions and means have no business interfering in the interpretation of texts either, even when the texts are sacred ones. Instead, biblical texts must be interpreted as the works of human beings, and by means of the same sorts of rigorous hermeneutical methods that are employed for interpreting other ancient texts -- any religious enlightenment coming as a *result* of such interpretation, not entering into the process itself. This whole position remained Herder's considered stance in his later period as well.

The general idea that the bible should be interpreted in the same way as other texts was by no means the commonplace in Herder’s day that it has since become, but nor was it new with him. In adopting this principle he was self-consciously following the lead of several recent bible scholars -- in particular, Ernesti, Michaelis, and Semler. However, Herder’s secularism is more consistent and radical than theirs.

This can be illustrated by a comparison with Ernesti (the most important of the scholars just mentioned, and the one most consistently admired by Herder). Ernesti’s great work, *Institutio interpretis Novi Testamenti* (1761), which Herder singles out for special praise, is a key statement of the sort of secularism in question. Initially, this work seems to advocate a secularism identical in spirit to Herder’s, arguing that we must interpret biblical books in the same way as profane texts, and *thereby* learn whatever religious truth they contain. However, as the work develops, matters become much cloudier. In this connection, it is important to distinguish two questions which can be asked concerning divine inspiration and interpretation: (1) May readers of sacred texts rely on a divine inspiration of *themselves* (e.g. by the Holy Spirit) bringing them to a correct interpretation rather than on more usual interpretive means? (2) May they assume in interpretation that because *the texts’ authors* are divinely inspired the texts must be completely true and therefore also (a fortiori) completely self-consistent? When Ernesti develops the details of his position it becomes clear that he has really only advanced as far towards secularism as consistently answering question (1) in the negative, *not question (2)*. His failure to give a consistently negative answer to question (2) lands him in flat contradiction with his official commitment to interpreting sacred texts in exactly the same way as profane texts (for of course, as he indeed himself implies, in interpreting profane texts we may *not* assume that the texts are throughout true and therefore also self-consistent). It also seems intellectually indefensible in itself -- merely a rather transparent refusal to stop, so to speak, "cooking the books" in favor of the bible when interpreting it. By contrast, the young Herder advances in his secularism beyond Ernesti because he consistently answers *both* questions in the negative, and thereby, unlike Ernesti, achieves a position which is both self-consistent and otherwise intellectually defensible. Moreover, Herder’s actual interpretations of the bible admirably conform to this theoretical position, not only refraining from any reliance on divine inspiration and instead employing normal interpretive techniques, but also frequently attributing false and even inconsistent positions to the bible (both to the Old and to the New Testaments).

Another noteworthy feature of Herder’s secularism is his insistence that interpreters of the bible must resist the temptation to read the bible as *allegory* (except in those few cases -- e.g. the parables of the New Testament -- where there is clear textual evidence of a biblical author’s intention to convey an allegorical meaning). Herder gives a perceptive general diagnosis of the temptation to allegorical interpretation: over the course of history people’s beliefs and values change, leading to discrepancies between the claims made by their traditional texts and their own beliefs and values, but they expect and want to find their traditional texts correct, and so they try to effect a reconciliation with their own beliefs and values by means of allegorical readings.

Herder’s theoretical commitment to strict secularism in biblical interpretation led him to interpretive discoveries concerning the bible which were in themselves of epoch-making importance. For example, concerning the Old Testament, his commitment to applying normal hermeneutical methods enabled him

to distinguish and define the different genres of poetry in the Old Testament in a way that was superior to anything done before him. Also, that commitment, and in particular his consequent readiness to find falsehood and even inconsistency in the bible, allowed him to make such important interpretive observations as that the ancient Jews' conceptions about death, afterlife, mind, and body, had changed dramatically over time. (For these two achievements, see especially *On the Spirit of Hebrew Poetry*.) Again, that commitment, and in particular Herder's consequent rejection of unwarranted allegorical interpretations, allowed him to substitute for the prevailing interpretation of the *Song of Solomon* as religious allegory an interpretation of it as simple erotic love poetry which is today generally accepted as correct. Similarly concerning the New Testament, Herder's commitment to applying normal hermeneutical methods, including his consequent readiness to recognize falsehood and inconsistency, enabled him to treat the authors of the four gospels as individual human authors instead of as mere mouthpieces of the deity, to perceive inconsistencies between their accounts, to establish the relative dates of the gospels correctly for the first time (Mark first, Matthew and Luke in the middle, John last and late), and to give a broadly correct account of their genesis in oral sermon and their likely relations to each other -- achievements attained above all in two late works from 1796-97, *On the Savior of Mankind* and *On God's Son, the World's Savior*.

Herder's strict secularism in interpretation would later be replicated by Schleiermacher, who similarly embraces the principle that the interpretation of sacred texts must treat them as the works of human authors and by means of exactly the same interpretive methods as are applied to profane texts, and similarly follows through on this commitment, in particular finding not only falsehoods but also inconsistencies in the bible.

Herder's achievements in this area have something of the character of the early acts of an inexorable tragedy, however. As was mentioned, he by no means intended his championing of the cause of intellectual conscience in insisting on the autonomy of natural science and interpretation to undermine religion in general or Christianity in particular; on the contrary, his hope and expectation was that both sorts of autonomy would in the end support religion and Christianity. However, this hope has been sorely disappointed. Autonomous natural science has increasingly made religion generally and Christianity in particular look untenable. And Herder's policy of reading the bible as a collection of human texts, with all the foibles of human texts, has increasingly led to an undermining of the bible's claims to intellectual authority. Much of what Herder has ultimately achieved in this area would therefore be deeply unwelcome to him.

Bibliography

Primary Texts

There are two main German editions of Herder's works:

- *Johann Gottfried Herder Sämtliche Werke*, B. Suphan, *et al*. (eds.), Berlin, 1887-.

- *Johann Gottfried Herder Werke*, U. Gaier, *et al*. (eds.), Frankfurt am Main, 1985-.

The latter edition includes very helpful notes.

Translations

- Adler, H., and Menze, E.A., *On World History,* Armonk, 1996.
  (Contains short excerpts on history from a variety of works, prominently including the *Ideas*.)
- Barnard, F.M., *J.G. Herder on Social and Political Culture*, Cambridge, 1969.
  (Includes (partial) translations of Herder's 1769 *Journal*, *On the Origin*, *This Too*, the *Dissertation on the Reciprocal Influence of Government and the Sciences*, and the *Ideas*, plus a very good introduction.)
- Burkhardt, F.H., *God. Some Conversations*, New York, 1940.
- Churchill, T., *Outlines of a Philosophy of the History of Man*, London, 1800.
  (A translation of the *Ideas*.)
- Forster, M.N., *J.G. Herder: Philosophical Writings*, Cambridge, 2001/2.
  (Contains full translations of *How Philosophy, On the Origin*, *On the Cognition*, and *This Too*, as well as other pieces.)
- Marsh, J., *The Spirit of Hebrew Poetry*, Burlington, Vt., 1833.
- Menze, E.A., Menges, K., and Palma, M., *Johann Gottfried Herder: Selected Early Works, 1764-7*, Pennsylvania, 1992.
  (Contains some early essays and selections from the *Fragments*.)
- Moran, J.H., and Gode, A., *On the Origin of Language*, Chicago, 1986.
  (Contains a partial translation of *On the Origin*.)
- Nisbet, H.B., *German Aesthetics and Literary Criticism: Winckelmann, Lessing, Hamann, Herder, Schiller, Goethe*, Cambridge, 1985.
  (Contains two pieces of Herder's in aesthetics, including his important essay *Shakespeare*.)

Secondary Literature in German

By far the most helpful single item remains:

- Haym, R., *Herder nach seinem Leben und seinen Werken*, Berlin, 1880.
  (A classic, detailed intellectual biography.)

Two useful recent collections of essays covering a broad range of topics are:

- Sauder, G. (ed.), *Johann Gottfried Herder 1744-1803*, Hamburg, 1987.
- Bollacher, M. (ed.), *Johann Gottfried Herder: Geschichte und Kultur*, Würzburg, 1994.

H.D. Irmischer has written several important articles on topics covered here, including:

- Irmischer, H., "Grundzüge der Hermeneutik Herders," in *Bückeburger Gespräche über J.G. Herder 1971*, Bückeburg, 1973.
- Irmischer, H., "Grundfragen der Geschichtsphilosophie Herders bis 1774," in *Bückeburger Gespräche über J.G. Herder 1983*, Bückeburg, 1984.

A helpful treatment of Herder’s interest in world literature, and in particular his theory and practice of translation, is:

- Kelletat, A.F., *Herder und die Weltliteratur*, Frankfurt am Main, 1984.

A good treatment of Herder’s approach to the Old Testament:

- Willi, T., *Herders Beitrag zum Verstehen des Alten Testaments*, Tübingen, 1971.

Secondary Literature in English

General treatments:

- Berlin, I., *Vico and Herder*, New York, 1976.
  (Concise and excellent.)
- Clark Jr., R.T., *Herder: His Life and Thought*, Berkeley, 1955.
  (Detailed and useful but unimaginative.)
- Beiser, F.C., *The Fate of Reason*, Cambridge, Mass., 1987.
  (Ch. 5 covers several topics helpfully, including Herder’s philosophies of language, mind, and religion.)

Herder’s general program and debts to the precritical Kant:

- Zammito, J.H., *Kant, Herder, and the Birth of Anthropology*, Chicago, 2001.

Philosophy of language:

- Forster, M. N., "Herder’s Philosophy of Language, Interpretation, and Translation: Three Fundamental Principles," forthcoming. *The Review of Metaphysics*.
- Taylor, C., "The Importance of Herder," in E. and A. Margalit eds., *Isaiah Berlin: A Celebration*, Chicago, 1991.
- Taylor, C., "Language and Human Nature," in C. Taylor, *Human Agency and Language: Philosophical Papers 1*, Cambridge, 1996.

Aesthetics:

- Norton, R.E., *Herder's Aesthetics and the European Enlightenment*, Ithaca, 1991.
  (Helpful both on aspects of Herder's aesthetic theory and on Herder's general relation to the Enlightenment.)

Philosophy of history:

- Lovejoy, A.O., "Herder and the Enlightenment Philosophy of History," in *Essays on the History of Ideas*, Baltimore, 1948.
  (A helpful short treatment.)
- Meinecke, F., *Historism: The Rise of a New Historical Outlook*, London, 1972.
  (Ch. 9 is very helpful.)

Political Philosophy:

- Ergang, R., *Herder and the Foundations of German Nationalism*, New York, 1931.
  (Helpful both on Herder's political thought and on his general intellectual influence.)
- Barnard, F.M., *Herder's Social and Political Thought: From Enlightenment to Nationalism*, Oxford, 1965.
  (Chs. 3-5 deal with Herder's political thought.)
- Beiser, F.C., *Enlightenment, Revolution, and Romanticism*, Cambridge, Mass., 1992.
  (Ch. 8 on Herder's political philosophy is excellent.)

Other Subjects:

- Nisbet, H.B., *Herder and the Philosophy and History of Science*, Cambridge, Mass., 1970.
  (An excellent account of Herder's stance towards science.)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

aesthetics | cosmopolitanism | Dilthey, Wilhelm | dualism | egalitarianism | Hamann, Johann Georg | Hegel, Georg Wilhelm Friedrich | hermeneutics | history, philosophy of | Humboldt, Wilhelm von | Kant, Immanuel | language: philosophy of | liberalism | Mill, John Stuart | mind: philosophy of | nationalism | naturalism | Nietzsche, Friedrich | physicalism | rationalism vs. empiricism | relativism | religion: philosophy of | Schleiermacher, Friedrich Daniel | skepticism | Spinoza, Baruch [Benedict] | Wolff, Christian

Michael N. Forster
University of Chicago
mnforste@uchicago.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 23, 2001
Content last modified: October 23, 2001

Johann Georg Hamann

Johann Georg Hamann (1730-1788) lived and worked in Prussia, in the context of the late German Enlightenment. Although he remained outside 'professional' philosophical circles, in that he never held a University post, he was respected in his time for his scholarship and breadth of learning. His writings were notorious even in his own time for the challenges they threw down to the reader. These challenges to interpretation and understanding are only heightened today.

Nevertheless an increasing number of scholars from philosophy, theology, aesthetics and German studies are finding his ideas and insights of value to contemporary concerns. His central preoccupations are still pertinent: language, knowledge, the nature of the human person, sexuality and gender and the relationship of humanity to God. Meanwhile, his views, which in many respects anticipate later challenges to the Enlightenment project and to modernity, are still relevant and even provocative.

---

1. Life

Johann Georg Hamann was born in Königsberg in 1730, the son of a midwife and a barber-surgeon. He

began study in philosophy and theology at the age of 16, changed to law but mainly read literature, philology, and rhetoric, but also mathematics and science. He left university without completing his studies and became the governor to a wealthy family on a Baltic estate. During this time he continued his extraordinarily broad reading and private research. He took up a job in the family firm of a friend from his Königsberg days, Christoph Berens, and was sent on an obscure mission to London, in which he evidently failed. He then led a high life until he ran out of friends, money and support. In a garret, depressed and impoverished, he read the Bible cover to cover and experienced a religious conversion.

He returned to the House of Berens in Riga, where they evidently forgave him his failure. He fell in love with Christoph Berens' sister, Katharina, but was refused permission to marry her by his friend, on the grounds of his religious conversion; Berens was an enthusiastic follower of the Enlightenment and was nauseated by the more pious manifestations of Hamann's new-found religiosity. Smarting from this blow and its motivations, Hamann returned to his father's house in Königsberg, where he lived for the rest of his life until his final months.

In Königsberg, he never held an official academic post, nor an ecclesiastical one; this may in part have been due to his pronounced speech impediment, which inhibited him from either lecturing or preaching. Eventually, through the intercession of his acquaintance, Immanuel Kant, he found employment as a low-level civil servant working in the tax office of Frederick the Great; a ruler Hamann in fact despised. Nevertheless his principal activity was as an editor and a writer; he was considered one of the most widely-read scholars of his time (greatly aided by his fluency in many languages), as well as a notorious author. During this time, despite his committed Christianity, he lived with a woman whom he never married but to whom he remained devoted and faithful, having four children on whom he doted, and who occasionally feature in his writings (principally as unruly distractions to the author's scholarship).

At the end of his life he accepted an invitation to Münster from one of his admirers, Princess Gallitzin. He died in Münster in 1788.

Hamann had a profound influence on the German ‘Storm and Stress’ movement, and on other contemporaries such as Herder and Jacobi; he impressed Hegel and Goethe (who called him the brightest head of his time) and was a major influence on Kierkegaard. His influence continued on twentieth century German thinkers, particularly those interested in language. His popularity has increased dramatically in the last few decades amongst philosophers, theologians, and German studies scholars around the world.

2. Writings

Hamann's writings are all short; he was not given to extensive treatises. They are also usually motivated by something very specific: someone else's publication, or particular circumstances and events. When responding to these, he presupposes considerable knowledge on the part of the reader; typically his responses to the work of others involves adoption of their terminology and style, blending into mimicry and parody as a rhetorical and argumentative device. Moreover, woven into these writings is an extraordinary breadth and quantity of citations and allusions; and by no means are these all clear and

obvious. Thus, when he chooses, his essays are a tapestry of multicolored threads of the ideas, language, and imagery of thinkers, be they ancient, biblical, or contemporary. These are woven across a woof of a love of irony, which as ever adds a layer of interpretative complexity.

Hamann's writings also frequently appear under the name of various fanciful characters: Aristobolus, the Knight of the Rose-Cross, the Sibyl, Adelgunde. It cannot be assumed that such characters faithfully represent Hamann's own views; The opinions of Aristobolus, for example, are a device to deconstruct and drive to absurdity a number of views which Hamann opposes.

These factors combine to make Hamann's writings notoriously difficult to understand and interpret. Goethe observed that when reading Hamann, “one must completely rule out what one normally means by understanding.” [Goethe, 550] Even if one has – or painfully acquires, or borrows from other current scholars – the breadth of reference to understand the source of an allusion or citation, it is not always clear what the citation is doing there, and what Hamann means to suggest by referring to it. He delighted in sporting with his reader; preferring to present a balled fist and leave it to the reader to unroll it into a flat hand (to borrow one of his own images; cf. the end of the “Metacritique of the Purism of Reason”, 1780.)

This impenetrability has been taken by some to signify an incapacity for clear expression on Hamann's part. Hegel remarked: "The French have a saying: *Le stile c'est l'homme meme* (“The style is the man himself”); Hamann's writings do not *have* a particular style but rather *are* style through and through." [Hegel, 209.] Others, however, have pointed to the clarity and concision of his letters, in contrast to the puzzling style he uses to address the public. They suggest that Hamann's challenging style forms are an important part of Hamann's hermeneutics and his understanding of the relationship between the writer and reader: two halves of a whole who must relate themselves to one another and unite for a common goal (“Reader and Critic”, 1762.) The reader cannot be passive and must work to reconstruct Hamann's meaning. As Hamann observed, “A writer who is in a hurry to be understood today or tomorrow runs the danger of being misunderstood the day after tomorrow.”

Consequently, Hamann is highly susceptible to misinterpretation. For example, one can find in recent English-language treatments the understanding of Hamann as an ‘irrationalist’, and one who simply opposed the Enlightenment with all his might; this however is not supported by the majority of Hamann scholars and is seen as a failure to understand the complexity of his thinking.

His principal writings include: *Biblische Betrachtungen* [Biblical Reflections], *Gedanken über meinen Lebenslauf* [Thoughts on the Course of my Life], *Brocken* [Fragments], *Sokratische Denkwürdigkeiten* [Socratic Memorabilia], *Wolken* [Clouds], *Kreuzzüge des Philologen* [Crusades of the Philologian], a collection of essays including Aesthetica in Nuce, *Versuch über eine akademische Frage* [Essay on an academic question], and *Kleeblatt Hellenistischer Briefe* [Cloverleaf of Hellenistic Letters]; *Schriftsteller und Kunstrichter* [Author and Critic], *Leser und Kunstrichter* [Reader and Critic], *Fünf Hirtenbriefe* [Five Pastoral Letters], *Des Ritters von Rosencreuz letzte Willensmeynung über den göttlichen und menschlichen Urprung der Sprache* [The Knight of the Rose-Cross' Last Will and Testament on the divine and human origin of language], *Philologische Einfälle und Zweifel* [Philological Ideas and Doubts], *Hierophantische Briefe* [Hierophantic Letters], *Versuch einer Sibylle über die Ehe* [Essay of a

Sibyl on Marriage], *Konxompax*, *Metakritik über den Purismum der Vernunft* [Metacritique of the Purism of Reason], *Golgotha und Scheblimini* [Golgotha and Scheblimini], *Fliegender Brief* [Flying Letter].

3. Metacritique

At the end of his life, Hamann chose to designate his authorship as “Metacritique”, a word he coined for his engagement with Kant's *Critique of Pure Reason*. Instead of creating a systematic theology, or an epistemology, he seems to have seen his work as one that examines the foundations and nature of philosophical and theological critique itself. Rather like the late Wittgenstein, his work was deconstructive; he belongs in the camp of philosophers whom Richard Rorty has described as “edifying and therapeutic” rather than “constructive and systematic”. (Rorty, 5-6.) He brings to any issue in philosophy not a constructive account, but an approach, a set of convictions, something akin to ethical principles. He anticipated Rorty's emphasis on the curative aspects of this task; at the end of his life, he wanted his collected works to be published under the title “Curative Baths” (“*Saalbadereyen*”&151;a reference to healing practices of the time and an allusion to his father's profession.) Each volume was to be called a ‘Tub’. This project was sadly never realized, not even under a more conventional title.

One abiding characteristic of Hamann's many responses to the philosophy of his time, therefore, is this ‘metacritical’ instinct: not to construct a rival account, but to go for the jugular; not to set up a rival critique, but to insist that critique itself must be subject to meta-critique, which concerns itself with the issues that must be attended to in relation to the act of philosophical reflection itself. It consists of attention to the fundamental stance or position on issues or insights that must underlie any work on philosophy or theology. According to Hamann, often “the difficulties lie in womb of the concepts”. (N III, 31, 21) Hamann's writing therefore is not so much ‘unsystematic’ as is sometimes said, but ‘presystematic’. He addresses issues that must be recognized in any self-critical reflection, matters that must be presuppositions for any system. Thus one of the most salient features of his “Metacritique of the Purism of Reason”, his unpublished response to Kant's *Critique of Pure Reason*, was to focus on the question of language. Through a tissue of imagery, he suggests that a proper view of language implies that the alleged ‘purity’ of *a priori* reason is untenable.

These metacritical issues, for Hamann, principally include language, knowledge, and the nature of the human person. Hamann also, most urgently and most controversially (then as now), did not believe that any of these issues can be answered outside a theological perspective; that is, without reference to God as humanity's creator and dialogue partner.

4. Relation

A second feature of Hamann's approach is a tendency which Goethe saw as holism. This is perhaps not the best way to describe Hamann's insight, as Hamann characteristically emphasised the brokenness of human experience, and fragmentariness of human knowledge: “Gaps and lacks&151;is the highest and deepest knowledge of human nature, through which we must climb our way up to the ideal—ideas and

doubts&151;the *summum bonum* of our reason.” (ZH 3, 34:33-35) Hamann essentially disliked attempts to isolate the phenomenon under consideration from other aspects with which he felt it to be intimately connected; this precludes a deep and true understanding of our existence. Taken as far as he did, this means that philosophy of language must include a discussion of God, and a discussion of God must make reference to sexuality and *vice versa*.

Thus in “Essay of a Sibyl on Marriage”, which he takes as an opportunity to write about sex, a proper understanding of human sexuality and erotic enjoyment cannot be understood without seeing humanity as the creature of God, made in God's image. He plays with the Christian idea of God as a Trinity to depict a trinity of woman-man-God in the moment of lovemaking; and reworks the account of the creation of Adam in Genesis to describe the act of coitus itself. The woman on perceiving her lover in his excitement sees ‘that rib’ and cries out in enthusiastic appropriation, ‘That is bone of my bone and flesh of my flesh!’ The man then ‘fills the hole of the place with flesh’ (as Genesis describes God doing with Adam after the creation of Eve). In doing so the lover also acknowledges that the origin of a man is in woman's body: the ‘Sibyl’ describes this moment of lovemaking as ”he entered in whence he once came forth.” Indeed, as Christ was born of a woman, the salvation of humanity proceeds from a woman's sexual body; the vagina is also described as the place that the Saviour came forth as the body's healer (the German language permits Hamann here to pun on ‘healer’ and ‘saviour’). This inclination to combine topics more often kept separate (such as ‘the concept of God’ and ‘having sex’) is salient throughout his work.

5. The Union of Opposites

Hamann's tool for conceiving the interrelation of these dimensions of human life increasingly was the Principle of the Union of Opposites. He writes approvingly of this principle to his friends; particularly after his encounter with Kant's new epistemology, claiming to value it more than the principles of contradiction and of sufficient reason, and indeed more than the whole Kantian Critique. (ZH 5, 327:12ff; ZH 4, 462:7-8) Contradictions and apparent oppositions fill our experience:

> Yes, daily at home I have the experience that one must always contradict oneself from two viewpoints, [which] never can agree, and that it is impossible to change these viewpoints into the other without doing the greatest violence to them. Our knowledge is piecemeal — no dogmatist is in a position to feel this great truth, if he is to play his role and play it well; and through a vicious circle of pure reason skepsis itself becomes dogma. (ZH 5, 432:29-36) (This is in the context of a discussion of Kant.)

Far from being a pre-condition for truth, the absence of contradiction is in Hamann's eyes a pre-condition for dogmatism. Knowledge must not proceed on the basis of unanimity and the absence of contradiction, but must proceed through the dialogue and relation of these different voices. (Hamann does not think in terms of Hegel's later dialectical synthesis.) When Hamann speaks of ‘opposition’ and contradictions, however, he does so in an ironic tone; for it is clearly his conviction that there is a fundamental unity in things, and the oppositions and contradictions that we perceive are chiefly of our own making. He insists that his perception is ‘without Manichaeism’. (ZH 5, 327:16-17) Body and mind, senses and reason,

reason and passion are not truly opposed. These are contrasting elements of the same unified – unified but not homogenous&151;reality. Hamann tries to steer a course between Scylla and Charybdis: between the dogmatic, even tyrannical extermination of opposition and contradiction; and the elimination of contradiction through a false synthesis or fusion achieved by an apparent acceptance of antithetical realities.

The Principle of the Union of Opposites as a tactical tool, therefore, does not imply that Hamann sees the world in terms of divisions and dualism. It is his strategy for coping with the schematic antitheses abundant in Enlightenment philosophy.

> Nothing seems easier than the leap from one extreme to the other, and nothing so difficult as the union to a center. … [TheUnion of Opposites] always seems to me to be the one sufficient reason of all contradictions&151;and the true process of their resolution and mediation, that makes an end to all feuds of healthy reason and pure unreason.(ZH 4, 287:5-17)

6. ‘Prosopopoeia’

Hamann used the notion of ‘Prosopopoeia’, or personification, as an image of what can happen in philosophical reflection. In a medieval morality or mystery play, the experience of being chaste or being lustful is transformed from a way of acting or feeling into a dramatic character who then speaks and acts as a personification of that quality. So too in philosophy, Hamann suggests. The philosopher distinguishes differing aspects in the phenomenon under scrutiny and exaggerates their difference. These aspects are ennobled into faculties, and through ‘prosopopoeia’ are hypostasized into entities. Thus in the act of reflecting on something, ‘reasoning’ is distinguished from ‘feeling’, and turned from a verb or gerund into a noun — ‘reason’—which is then named as a constituent of our being. Reason then becomes a thing to which we can ascribe properties. (This shows perhaps a streak bordering on nominalism in Hamann.)

The best example of this, and of Hamann's response, is his treatment of the word ‘reason’. Since he handles it with a kind of skepticism or even distaste, he is often called an ‘irrationalist’. It is clear however that Hamann puts a high value on certain ways of being reasonable and of reasoning activity. “Without *language* we would have no reason, without reason no religion, and without these three essential aspects of our nature, neither mind *[Geist]* nor bond of society”. (N III, 231, 10-12)

Hamann's treatment of reason instead is a deconstruction, both of the prosopopoeic use of the word and the Enlightenment valuation of it. There is no such thing as reason — there is only reasoning. Reasoning, as something we do, is as fallible as we are, and as such is subject to our position in history, or own personality, or the circumstances of the moment. ‘It’ is therefore not a universal, healthy and infallible ‘faculty’ as Hamann's Enlightenment contemporaries often maintained:

> *Being*, *belief and reason* are pure relations, which cannot be dealt with absolutely, and are not things but pure scholastic concepts, *signs* for understanding, not for worshipping, aids

to awaken our attention, not to fetter it. (ZH 7, 165:7-11)

7. Enlightenment

Hamann is sometimes portrayed simply as an opponent of ‘the Enlightenment’. This presupposes of course that ‘the Enlightenment’ constitutes a unified stance on a number of philosophical issues, an assumption which is questionable. The majority of Hamann scholars today see his position in a more complex way. Hamann opposed many of the popular convictions of his time. However, Hamann fought his contemporaries on many fronts; often with areas of considerable agreement with some of his opponents. One example would be the way that he deployed Hume as a weapon against Enlightenment rationalism, not least against Kant (although Hamann was the one who introduced Kant to Hume's writings in the first place). Although Hamann, as a Christian, had profound disagreements with Hume's thought in its atheistic aspects, nevertheless he used Humean skepticism in his own deconstructive writings. Hume's doubts about the reliability and self-sufficiency of reason were grist to Hamann's mill. Hume's insistence that ‘belief’ underlies much of our thinking and reasoning was adopted and deployed by Hamann, often with a linguistic sleight of hand. By using the word ‘*Glaube*’ (which in German includes both ‘belief’ in an epistemic sense and ‘faith’ in a religious sense), Hamann could assert that ‘faith’, not rational grounds, underlies his contemporaries' high valuation of reason. Thus even the enthusiastic advocates of impartiality and ‘reason’, who are also skeptics about ‘blind faith’, have ultimately only faith as the ground for their convictions.

In one sense, however, Hamann can certainly be seen as a critic—or metacritic—of the Enlightenment. The question of what ‘Enlightenment’ consists in was a challenge Hamann through down to his contemporaries, from his debut with Socratic Memorabilia (1759) to the end of his life. It is instructive to juxtapose Kant's famous essay “An Answer to the Question: What is Enlightenment?” (1784) with Hamann's response in a letter to his acquaintance Christian Jacob Kraus. Kant defines enlightenment as the exit from ‘self-incurred minority’ (or ‘immaturity’ or ‘tutelage’), which arises from laziness and cowardice. (The ‘entire fairer sex’ in particular is said by Kant to regard the transition to maturity and thinking for oneself as difficult and dangerous). ‘*Sapere aude*!’ (‘Dare to know!’), Kant instructs the reader. However, while Kant urges the ‘public’ use of reason (use of reason as a scholar), he nevertheless claims that ‘private’ (we would perhaps say, ‘professional’) use of reason must be circumscribed, for example, for the clergyman, soldier, or taxpayer; they must simply obey. Moreover, Kant heaps praise on their monarch, Frederick the Great, whom Hamann deemed immoral and despotic.

The irony of being instructed to think for oneself, and being told to have the courage to know, was not lost on Hamann. More painful is the irony in being told to appreciate the freedom to think, but “believe, march, pay if the devil is not to take you” (Hamann's depiction of Kant's insistence that clergymen, soldiers and taxpayers must just obey orders). “What good to me is the festive garment of freedom when I am in a slave's smock at home?” Hamann asks, referring to Kant's approval of public use of reason but ‘private’ requirement to obey. In Hamann's view, the scholarly freedom to reflect, which Kant commends, is a luxury compared to the ethical imperative to question and debate in the professional and political sphere, which Kant restricts. “Thus the public use of reason and freedom is nothing but a dessert, a

sumptuous dessert. The private use is the *daily bread* that we should give up for its sake. The *self-incurred immaturity* is just such a sneer as he makes at the whole fair sex, and which my three daughters will not put up with.”

Most urgently, therefore, Hamann objects to the allegation that this immaturity is ‘self-incurred’, rather than imposed on the people firstly by a despotic monarch, and secondly by intellectuals like Kant, with the ‘prattle and reasoning of those emancipated immature ones, who set themselves up as guardians’. ‘*True enlightenment*,’ Hamann concludes sarcastically, with an eye to the likes of Kant and Frederick, “consists in an emergence of the immature person from a supremely *self-incurred guardianship*.”

8. Language

Language is one of Hamann's most abiding philosophical concerns. From the beginning of his work, Hamann championed the priority which expression and communication, passion and symbol possess over abstraction, analysis and logic in matters of language. Neither logic nor even representation (in Rorty's sense) possesses the rights of the first-born. Representation is secondary and derivative rather than the whole function of language. Symbolism, imagery, metaphor have primacy; “Poetry is the mother-tongue of the human race.” (N II, 197) To think that language is essentially a passive system of signs for communicating thoughts is to deal a deathblow to true language.

For all Hamann's emphasis in his earlier writings on passion and emotion, he does not equate language with emotional expression. This became clear in his engagement with the writing of his younger friend Herder on the origin of language. Language has a mediating relationship between our reflection, one another, and our world; and as it is not simply the cries of emotion of an animal, so too it is not a smothering curtain between us and the rest of reality. Language also has a mediating role between God and us. Hamann's answer to a debate of his time, the origin of language — divine or human?—is that its origin is found in the relationship between God and humanity. Typically he has the ‘Knight of the Rose-Cross’ express this in the form of a ‘myth’, rather than attempting to work out such a claim logically and systematically. Rewriting the story of the Garden of Eden, he describes this paradise as:

> Every phenomenon of nature was a word,—the sign, symbol and pledge of a new, mysterious, inexpressible but all the more intimate union, participation and community of divine energies and ideas. Everything the human being heard from the beginning, saw with its eyes, looked upon and touched with its hands was a living word; for God was the word.(NIII, 32: 21-30)

This makes the origin of language as easy and natural as child's play.

By the end of his life, because of his engagement with Kant, the most urgent question among the relationships that constitute language is the relationship of language to thinking or ‘reason’. In his view, the central question of Kant's first *Critique*, the very possibility of *a priori* knowledge and of pure reason, depends on the nature of language. In a passage full of subtle allusions to Kantian passages and terms, he

writes:

> Indeed, if a chief question does remain: *how is the power to think possible?*—The power to think *right* and *left, before* and *without, with* and *above* experience? then it does not take a deduction to prove the genealogical priority of language…. Not only the entire ability to think rests on language... but language is also the *crux of the misunderstanding of reason with itself*. (N III, 286:1-10)

Critique creates.

For Hamann, in contrast to Kant, the question is therefore not so much 'what is reason?' as 'what is language?', as he writes in a letter. This is the ground of the paralogisms and antinomies that Kant raises in his Critique. Sharing Hume's empiricism and Berkeley's suspicion of universals and abstract terms, he concludes: "Hence it happens that one takes *words* for *concepts*, and *concepts* for the *things themselves*." (ZH 5, 264:34-265:1) Language then has a fundamental role to play in unmasking the philosopher's tendency to 'prosopopoeia'. The relation of language to reason he certainly did not feel had solved, however, as he wrote to a friend:

> If only I was as eloquent as Demosthenes, I would have to do no more than repeat a single word three times. Reason is language—Logos; I gnaw on this marrowbone and will gnaw myself to death over it. It is still always dark over these depths for me: I am still always awaiting an apocalyptic angel with a key to this abyss. (ZH 5, 177:16-21)

9. Knowledge

For Hamann, knowledge is inseparable from self-knowledge, and self-knowledge inseparable from knowledge of the other. We are visible, as in a mirror, in each other; 'God and my neighbor are therefore a part of my self-knowledge, my self-love.' (N I, 302:16-23) He writes in a letter: "Self knowledge begins with the neighbor, the mirror, and just the same with true self-love; that goes from the mirror to the matter." (ZH 6, 281:16-17) Sometimes this exploration of self-knowledge through interpersonal intimacy takes a sexual form, as in the Sibyl's Essay on Marriage (already discussed).

All forms of knowledge, of learning and development even of the most natural functions, require the help of another. (The 'Knight of the Rose Cross', while jesting with Hume, tells us ironically that even eating and drinking, and indeed excretion, are not instinctual or innate but require teaching.) (N III, 28:26-28; N III, 29:7-10) This is conceived not only in such immediately interpersonal ways, but also more widely in the context of the community. The indispensability of 'the other' for knowledge is also the reason that Hamann gives the importance he does to tradition in the formation of knowledge. "Our reason arises, at the very least, from this twofold lesson of sensuous revelations and human testimonies". (N III 29:28-30) Years before Kant's first *Critique*, Hamann attempts to relate the senses and the understanding and their roles in knowledge, using a characteristically concrete metaphor: the senses are like the stomach, the understanding like blood vessels. Not only do the blood vessels need the stomach to receive the

nourishment that they distribute; the stomach also needs the blood vessels to function. This insistence on the mutual dependency and interrelation of sense experience and understanding (as opposed to many Enlightenment views that plumped for either reason or the senses as the dominant party) was refined in his engagement with Kant's critique. Throughout his life, he was neither materialist, purely empiricist or positivist, nor idealist, rationalist or intellectualist in his epistemology. Rather, he is firmly against dividing knowledge or ways of knowing into different kinds. “The philosophers have always given truth a bill of divorce, by separating what nature has joined together and vice versa”. (N III, 40: 3-5) With Kant's critique, the problem becomes still more urgent; Kant's treatment of the issue is a “violent, unwarranted, obstinate divorce of what nature has joined together”. Revising Kant's metaphor of sensibility and understanding as two stems of human knowledge, he suggests we see knowledge as a single stem with two roots. Thus he rejects the division of what can be known *a priori* from what can be known a posteriori, and many of the consequences of such a stance. Here again he reaches for the ‘Principle of the Union of Opposites’ in his deployment of imagery to suggest a different approach. The relation of the senses and understanding is a ‘hypostatic union’, a ‘*communicatio idiomatum*’ (phrases borrowed from Christian theological discussion of how the two natures of divine and human are united in Christ). This mysterious union can only be revealed and understood by ‘ordinary language’. (In suggesting that the problems in philosophy can be cured by attending to ‘ordinary language’, he clearly anticipates the late Wittgenstein).

In this engagement with Kant, Hamann returns and deepens the lesson he had learnt much earlier in his reading of Hume: that belief or faith is an essential precursor for knowledge. Everything is dependent or grounded on faith; there is no privileged position for any kind or form of knowledge (a priori, scientific, etc.) In Hamann's epistemology, the hard division between ‘knowledge’ and ‘belief’ or ‘faith’ becomes eroded. Both knowledge and faith rest on a foundation of trust; neither rest on a foundation of indubitability. “Every philosophy consists of certain and uncertain knowledge, of idealism and realism, of sensuousness and deductions. Why should only the uncertain be called belief? What then are—*rational grounds?*” (ZH 7, 165:33-37) ‘Sensuousness’ translates *Sinnlichkeit* (Kant's ‘sensibility’). Belief and reason both need each other; idealism and realism are a fantasized opposition, of which the authentic use of reason knows nothing. The unity that lies in the nature of things should lie at the foundation of all our concepts and reflection. (ZH 7, 165:7-17)

10. Interpretation and Understanding

From his ‘debut’ work, *Socratic Memorabilia*, Hamann began to promulgate a particular view of what it means to understand something. From the beginning of that essay he emphasized the importance of passion and commitment in interpretation; undermining the more conventional assumption that objectivity and detachment are prerequisites of philosophical reflection and understanding. In *Aesthetica in Nuce*, wearing the authorial mask of the ‘kabbalistic philologian’, he provocatively maintained that initiation into orgies were necessary before the interpreter could safely begin the hermeneutical act. The idea that one must rid oneself of presuppositions, prejudices, and predilections in order to do justice to the subject matter he characterizes as ‘monastic rules’—i.e. an excessive asceticism and abstinence. He goes so far as to compare such individuals to self-castrating eunuchs. (N II, 207:10-20)

Hamann's skepticism about neutrality and objectivity does not make him a ‘subjectivist’, however. The stance and disposition of the interpreter is integral, helpful, indeed, indispensable and must be acknowledged; but limitless subjectivity arouses Hamann's scorn. Those who ‘flood the text’ with glosses and marginalia, “dreaming up one's own inspiration and interpretation,” Hamann likens to the blind leading the blind. (N II, 208:3ff.)

The constraints which Hamann places on the interpreter's subjectivity are not those usually advocated, therefore: an avoidance of prejudice and pre-conceptions; an amnesia for one's own history, tradition and culture; an obedience to exegetical rules. The first restraint on subjective distortion is the interpreter's own common sense; the last is the reaction of the text itself: “what are you trying to make of me?” The interpreter's freedom is inextricable from the interpreter's respect and responsibility, in Hamann's view.

The responsible interpreter is conscious of standing within something larger than oneself: a tradition. The wise interpreter is a kabbalist, one who interprets an ancient text, and a rhapsodist—the original meaning of the latter being one who stitches something together from pre-existing materials. (In Ancient Greece the ‘rhapsodist’ was one who recited poems cobbled together from prior sources, usually bits of Homer.) In creating interpretations, the interpreter enjoys the freedom to create anew, as Hamann created his characteristic prose from pre-existing texts, while creating a new meaningful piece. Hamann's use of this genre itself makes the point: the demand that only one meaning may exist for a text arises from an impoverished notion of meaning and creativity; one that misunderstands the nature of composition and the nature of interpretation alike.

Hamann's rejected both exegetical ‘materialism’ and ‘idealism’, as he called them — literalism and excessive flights of fancy. In thus insisting on the integrity of ‘the letter *and* the spirit’, he means to preserve the place of author, text and reader alike. Both meaning and interpretation rest in a three-way relationship.

For Hamann, the depth and meaning of a text go beyond the author's own contribution, and are the responsibility of the interpreter: “Few authors understand themselves, and a proper reader must not only *understand* his author but also be able to *see beyond him*.” (ZH 6, 22:10-12) And yet this recognition that the author's opinions and intentions do not exhaust the possibilities of the text does not annihilate the place of the author. At the very least, the ‘beyond’ of the text includes a territory which, if unknown to the author, is not unrelated: that is, the author's own unconscious workings and meanings. It is not accidental that Hamann observes *not* that few authors understand their own *text,* but that few authors understand *themselves*. This suggests a picture of creation in which more of the author is expressed in a text and entrusted to the interpreter than the author's conscious intentions and opinions. This in turn suggests a picture of interpretation—of “understanding one's hero” as Hamann put it when writing about Socrates—in which greater sensitivity, insight, and fidelity is demanded of the interpreter than would otherwise be the case. Above all, the interpreter must have the courage to be a kabbalist; that is, to say more than the text does, not to express oneself but to say what the author left unsaid. The fruits of such faithful creativity may be impossible to ‘justify’ or ‘verify’ to the demands of the objectivist, however.

Fundamentally, for Hamann hermeneutics consists in perceiving the underlying relationship beneath the phenomenon in question; at the least, of course, the relationship between the author and the interpreter which requires such fidelity. Given Hamann's religious views, this at once introduces a theological dimension. Ultimately, this means that for Hamann proper hermeneutics rests on one thing: perceiving God revealed within the phenomenon, whether that be nature or history (cf. *Socratic Memorabilia* and *Aesthetica in Nuce* for examples). Even the interpretation of ourselves is a revelation of God; a recognition of whose image solves all the most complicated knots and riddles of our nature. (N II, 206:32-207:2; 198:3-5)

11. Humanity

The topics examined so far all have their anthropological implications. Hamann's critique of the socio-political implications of Kant's vision of ‘enlightenment’ rests on a conviction about our social and political destiny. Hamann sees our socio-political vocation as consisting firstly in ‘criticism’ (or ‘critique’)—recognizing and appropriating, or hating and rejecting, the true vs. the false, good vs. evil, beautiful vs. ugly; and secondly in ‘politics’, which is increasing or reducing them. This is not the prerogative of the ruler; every one is at once their own ‘king’, their own ‘legislator’; but also the ‘first-born of their subjects’. It is our “republican privilege” to contribute to this destiny, “the critical and magisterial office of a political animal.” (N III, 38-39)

The concepts of knowledge and language and their many facets also imply a particular anthropology: the diversity yet integration of the human being. For Hamann, the truest picture of humanity is of diversity in unity; a number of different, often contrasting aspects and features together composing the human person. Hamann consequently did not confine his attention to epistemology and reason when considering what human beings are, and passion, the thirst for vengeance, and sexual ecstasy form a part of his picture as well. (In response to the Enlightenment aesthetic of art as the imitation of ‘beautiful nature’, Hamann's ironic observation was: “The thirst for vengeance was the beautiful nature which Homer imitated.” ZH 2, 157:12)

The theme of interdependence between human beings, which was emphasized in his epistemology, also has its roots in his understanding of what it is to be human. We are not self-sufficient; but for Hamann, even our lacks and failings have a positive thrust, this signifier of dependency making us all the more suited for the enjoyment of nature and one another.

If there is a fundamental key to his thinking on humanity, it is the idea that the human being is the image of God. This is admittedly more theological than philosophical, but is essential for understanding Hamann's philosophical anthropology. Hamann's treatment of this perennial theme is hardly conventional in the history of Christian thinking. While the experience of sinfulness and wickedness is a powerful theme, particularly in his earlier, post-conversion writing, the fundamental thrust of his thinking is the easy exchange between the human and the divine. In this exchange, language is “the sign, symbol and pledge of a new, mysterious, inexpressible but all the more intimate union, participation and community of divine energies and ideas.“ (N III, 32:21-24) Despite his reputation for being an irrationalist, reasoning

too relates us to God; God, nature and reason are described as having the same relation as light, the eye and what we see, or as author, text and reader. (ZH 5, 272:14-16)

One must also remember that Hamann confessed that he could not conceive of a Creative Spirit without genitalia; indeed, he was quite happy to assert that the genitals are the unique bond between creature and Creator. So sexuality in divine-human relations has two aspects. First, as paradigm of creativity, it is the way in which our God-likeness can most strikingly be seen. Secondly, as the point of the most profound unity, it is the locus for our union both with another human being and with the divine. Provocatively, Hamann sees original sin and its rebellion as embodied not in sexuality, but in reason. Overweening reason is our attempt to be like God; meanwhile, prudery is the rejection of God's image, while trying to be like God in the wrong sense (bodilessness). (See Essay of a Sibyl on Marriage and Konxompax.) One should therefore distinguish ‘likeness to God’ from ‘being equal to God’. In the Sibyl's essay, the male version of grasping at equality with God (cf. Phil. 2:6) is the attempt to be self-sufficient, to be the God of monotheism: the sole ruler, who possesses self-existence. Instead, the encounter with the opposite sex should engender in the man an attitude of profound respect towards the woman's body, as the source of his own existence, from his mother. As the source of his own joy, lovemaking also is an acknowledgement of his own dependence, his lack of self-sufficiency and autonomy. But this dependence on another paradoxically is the Godlikeness of the Creator, the father, the one who humbles himself in self-giving (a favourite Hamannian theme in his discussion of God). Meanwhile, the woman's temptation is to an artificial innocence; a secret envy of God's incorporeality and impassibility. The defence of one's virginity is another cryptic attempt at self-sufficiency. Instead, the woman must brave the ‘tongues of fire’ in a ‘sacrifical offering of innocence’, in order to realize her Godlikeness; which is not to be found in bodilessness and the absence of passion, but in passionate creativity; in the willingness to be incarnate. Thus, if human beings are in the image of God, it is a trinitarian image of God, a mutual relation of love of ‘Father’, ‘Son’ and ‘Spirit’; found in creating, in saving, and in tongues of fire.

Bibliography

Hamann's writings

Hamann's works, including those unpublished in his lifetime, are reprinted in the collection edited by Josef Nadler:

- Hamann, Johann Georg. *Sämtliche Werken*, edited by Josef Nadler. 6 volumes. Vienna: Verlag Herder, 1949-1957). This was reprinted recently by Brockhaus in Wuppertal, 1999.

Citations from this source, in conformity with common practice for Hamann references, are given above as: N II, 13:10. This means “Nadler's edition, volume two, page thirteen, line ten.” This same mode of reference also applies to the translations in Gwen Griffith Dickson (see below), wherein the pages are laid out in as close an approximation as possible to Nadler's edition.

Hamann's Letters

- Hamann, Johann Georg. *Briefwechsel*, edited by Walther Ziesemer and Arthur Henkel (from volume 4 on, edited by Henkel alone). 8 volumes. Wiesbaden/ Frankfurt: Insel Verlag, 1955-1975.

Citations from this source are given as: ZH 4 etc. as above.

All translations from Hamann's works and letters in the above article are from Gwen Griffith Dickson (see below) except for the translation from the letter to Kraus, which is cited in the translation by Garrett Green, in Schmidt, James (ed.). What is Enlightenment? Eighteenth-Century Answers and Twentieth-Century Questions. Berkeley and Los Angeles: University of California Press, 1996.

Other selections

- Johann Georg Hamann. *Schriften zur Sprache. Einleitung und Anmerkungen von Josef Simon*. Frankfurt a.M.: Suhrkamp Verlag 1967. (suhrkamp theorie 1).
- -----. *Eine Auswahl aus seinen Schriften. Entkleidung und Verklärung*. Hg. von Martin Seils. Wuppertal: R. Brockhaus Verlag 1987.
- -----. *Vom Magus im Norden und der Verwegenheit des Geistes. Ausgewählte Schriften*. Hg. von Stefan Majetschak. Düsseldorf: Parerga Verlag 1993.
- -----. *Ausgewählte Schriften*. Hg. von Hans Eichner. Berlin: Nicolaische Verlagsbuchhandlung 1994.
- -----. *Ausgewählt, eingeleitet und mit Anmerkungen versehen von Arthur Henkel*. Frankfurt a.M.: Insel Verlag 1988.

Other editions and commentaries, in German

- *Daphne. Nachdruck der von Johann Georg Hamann, Johann Gotthelf Lindner u.a. herausgegebenen Königsberger Zeitschrift (1749-1750). Mit einem Nachwort von Joseph Kohnen*. Frankfurt a.M.: Peter Lang 1991. (Regensburger Beiträge zur deutschen Sprach- und Literaturwissenschaft. Reihe A: Quellen, Bd.5).
- Johann Georg Hamann, *Londoner Schriften. Historisch-kritische Neuedition von Oswald Bayer und Bernd Weißenborn*. München: C.H. Beck 1993.
- -----. *Sokratische Denkwürdigkeiten. Aesthetica in nuce*. Mit einem Kommentar hg. von Sven-Aage Jörgensen. Stuttgart: Reclam Verlag 1968. (Reclams Universalbibliothek 926/26a).
- -----. *Kleeblatt Hellenistischer Briefe*. Text mit Wiedergabe des Erstdruckes, hg. und kommentiert von Karlheinz Löhrer. Frankfurt a.M.: Peter Lang 1994. (Regensburger Beiträge zur deutschen Sprach- und Literaturwissenschaft. Reihe A: Bd. 8).
- Wild, Reiner. “Metacriticus bonae spei”. Johann Georg Hamann's “Fliegender Brief”. Einführung, Text und Kommentar. Frankfurt a.M.: Peter Lang 1975. (Regensburger Beiträge zur deutschen Sprach- und Literaturwissenschaft. Reihe B: Untersuchungen, Bd.6).
- Bayer, Oswald und Christian Knudsen (Hg.), *Kreuz und Kritik*. Johann Georg Hamanns Letztes

Blatt. Text und Interpretation. Tübingen: Mohr 1983.

- Blanke, Fritz and Karlfried Gründer. *Johann Georg Hamanns Hauptschriften Erklärt*. 8 volumes were projected, the following appeared: Gütersloh: Mohn, 1962f.

- I: *Die Hamann-Forschung* edited by Fritz Blanke and Lothar Schreiner (1956).
- II: *Sokratische Denkwürdigkeiten* edited by Fritz Blanke and Karlfried Gründer, *erklärt von* Fritz Blanke (1959).
- IV: *Über den Ursprung der Sprache* edited by Fritz Blanke und Karlfried Gründer, *erklärt von* Elfriede Büchsel (1963).
- V: *Mysterienschriften*. edited by Fritz Blanke und Karlfried Gründer, *erklärt von* Evert Jansen Schoonhoven and Martin Seils (1962).
- VII: *Golgotha und Scheblimini*. edited by Fritz Blanke und Lothar Schreiner, *erklärt von* Lothar Schreiner (1956).
- Bayer, Oswald. *Vernunft ist Sprache. Hamanns Metakritik Kants*. Stuttgart: Frommann-Holzboog, 2002.

- Manegold, Ingemarie. *Johann Georg Hamanns Schrift Konxompax*. Heidelberg: Carl Winter Universitätsverlag, 1963.

English translations and commentaries

- O'Flaherty, James. *Hamann's Socratic Memorabilia. A Translation and a Commentary*. Baltimore: The Johns Hopkins Press, 1967. [*SM*] (Socratic Memorabilia)
- Nisbet, H. B., ed. *German Aesthetic and Literary Criticism: Winckelmann, Lessing, Hamann, Herder, Schiller, Goethe*. Cambridge: Cambridge University Press, 1985. (*Aesthetica in NuceM*, translated by Joyce Crick)
- Smith, Ronald Gregor. *J.G. Hamann, 1730-1788. A Study in Christian Existence. With Selections from his Writings*. London: Collins, 1960. (A selection of short passages)
- Dickson, Gwen Griffith [Gwen Griffith-Dickson]. *Johann Georg Hamann's Relational Metacriticism*. Berlin: de Gruyter 1995. (*Socratic Memorabilia*, *Aesthetica in Nuce*, a selection of essays on language, “Essay of a Sibyl on Marriage,” *Metacritique of the Purism of Reason*)
- Schmidt, James (ed.). *What is Enlightenment? Eighteenth-Century Answers and Twentieth-Century Questions*. Berkeley and Los Angeles: University of California Press, 1996. (“Letter to Kraus,” translated by Garrett Green and *Metacritique of the Purism of Reason*, translated by Kenneth Haynes)

Monographs and studies in English

- Alexander, W.M. *Johann Georg Haman. Philosophy and Faith*. The Hague 1966.
- Beiser, Frederick C. *The Fate of Reason. German Philosophy from Kant to Fichte*. Cambridge, Mass.: Harvard University Press, 1987.

- Berlin, Isaiah. *The Magus of North*. London: John Murray 1993.
- Dickson, Gwen Griffith [Gwen Griffith-Dickson] *Johann Georg Hamann´s Relational Metacriticism*. Berlin: de Gruyter 1995.
- Dunning, Stephen. *The Tongues of Men: Hegel and Hamann on Religious Language and History*. Missoula: Scholars Press, 1979.
- German, Terence J. *Hamann on Language and Religion*. Oxford: Oxford University Press, 1981.
- Leibrecht, Walter. *God and Man in the Thought of Hamann*. Translated by James H. Stam and Martin H. Bertram. Philadelphia: Fortress Press 1966.
- Lowrie, Walter. *Johann Georg Hamann, an Existentialist*. Princeton 1950.
- O'Flaherty, James C. *Unity and Language: A Study in the Philosophy of Johann Georg Hamann*. Chapel Hill: University of No. Carolina Press, 1952.
- -----. *Hamann's Socratic Memorabilia. A Translation and a Commentary*. Baltimore: The Johns Hopkins Press, 1967.
- -----. *Johann Georg Hamann*. Boston: 1979.
- -----. *The Quarrel of Reason with Itself. Essays on Hamann, Michaelis, Lessing, Nietzsche*. Columbia: Camden House 1988.
- Vaughan, Larry. *Johann Georg Hamann: Metaphysics of Language and Vision of History*. Frankfurt a.M.: Peter Lang 1989. (American University Studies. Series I. Germanic Languages and Literatur. Vol. 60).

Literature reviews

- Büchsel, Elfriede. “Geschärfte Aufmerksamkeit - Hamannliteratur seit 1972.” In: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte*. 60. Jg., H.3, 1986, S.375-425.
- Büchsel, Elfriede. “Weitgefächertes Interess’. Hamannliteratur 1986-1995. In: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte*. 71. Jg., H.2, 1997, S.288-356.

The International Hamann-Colloquium

The International Hamann-Colloquium meets every few years. Collections of its papers are some of the most important contributions to Hamann scholarship:

- *Acta des Internationalen Hamann-Colloquiums. Johann Georg Hamann*, ed. by Bernhard Gajek. Frankfurt am Main: Vittorio Klostermann, 1979.
- *Acta des zweiten Internationalen Hamann-Colloquiums im Herder-Institut zu Marburg/Lahn*, ed. by Bernhard Gajek. Marburg: N. G. Elwert Verlag: 1983.
- *Acta des dritten Internationalen Hamann-Colloquiums. Hamann und Frankreich*, ed. by Bernhard Gajek. Marburg: Elwert Verlag, 1987.
- *Acta des vierten Internationalen Hamann-Colloquiums. Hamann-Kant-Herder*.
- *Acta des fünften Internationalen Hamann-Colloquiums. Johann Georg Hamann und die Krise der Aufklärung* in Münster. 1990.
- *Acta des Siebten Internationalen Hamann-Kolloquiums* - Johann Georg Hamann und England:

Hamann und die englischsprachige Aufklärung ; zu Marburg/Lahn 1996. Frankfurt am Main (u.a.): Lang, 1999.

Other selected works

- Altenhöner, Ingrid. *Die Sibylle als literarische Chiffre bei Johann Georg Hamann - Friedrich Schlegel - Johann Wolfgang Goethe.* Frankfurt a.M: Peter Lang 1997. (Europäische Hochschulschriften. Reihe 1, Deutsche Sprache und Literatur, Bd. 1646).
- Bayer, Oswald. *Zeitgenosse im Widerspruch. Johann Georg Hamann als radikaler Aufklärer*. München: Piper 1988. (Serie Piper 918).
- Bayer, Oswald. *Johann Georg Hamann: Der hellste Kopf seiner Zeit.* Tübingen: Attempo, 1998.
- Cloeren, Hermann J. 'Language and Thought'. In: *German Approaches to Analytic philosophy in the 18th and 19th Centuries* (1988), S. 21-26.
- Dahlstrom, Daniel. 'The Aesthetic Holism of Hamann, Herder and Schiller'. In: Karl Ameriks (ed.), *The Cambridge Companion to German Idealism.* Cambridge 2000.
- Fischer, Rainer. *Die Kunst des Bibellesens. Theologische Ästhetik am Beispiel des Schriftverständnisses*. Frankfurt a.M.: Peter Lang 1996 (Beiträge zur theologischen Urteilsbildung, Bd.1).
- Fritsch, Friedemann. 'Wirklichkeit als göttlich und menschlich zugleich. Überlegungen zur Verallgemeinerung einer christologischen Bestimmung in Hamanns Denken.' In: Oswald Bayer (Hg.), *Der hellste Kopf seiner Zeit*, S. 52-79
- Kleffmann, Tom. *Die Erbsündenlehre in sprachtheologischem Horizont: eine Interpretation Augustins, Luthers und Hamanns*. Tübingen: Mohr, 1994.
- Hoffmann, Volker. *Johann Georg Hamanns Philologie: Hamanns Philologie zwischen enzyklopädischer Mikrologie und Hermeneutik*. Stuttgart (u.a.): Kohlhammer, 1972.
- Merlan, Philip. 'From Hume to Hamann', *The Personalist* 321(1951): 11-18.
- -----. "'Parva Hamanniana': J. G. Hamann as Spokesman of the Middle Class." *Journal of the History of Ideas* 9 (1948): 380-384.
- -----. "'Parva Hamannianad'" (II) : Hamann and Schmohl". *Journal of the History of Ideas* 10 (1949): 567-574.
- -----. "'Parva Hamanniana': Hamann and Galiani." *Journal of the History of Ideas* 11 (1950): 486-489.
- Nadler, Josef. *Johann Georg Hamann. Der Zeuge des Corpus mysticum*. Salzburg: Otto Müller 1949.
- Nebel, Gerhard. *Hamann*. 1973.
- Piske, Irmgard. *Offenbarung, Sprache, Vernunft. Zur Auseinandersetzung Hamanns mit Kant*. Frankfurt am Main: Regensburg, 1989.
- Redmond, M. 'The Hamann-Hume Connection', *Religious Studies* 23, 95-107.
- Salmony, H.A. *Johann Georg Hamanns metakritische Philosophie*. Erster Band: Einführung in die metakritische Philosophie J. G. Hamanns. Basel 1958.
- Seils, Martin. *Theologische Aspekte zur gegenwärtigen Hamann-Deutung*. Göttingen: Vandenhoeck & Ruprecht, 1957
- Swain, Charles W. 'Hamann and the Philosophy of Hume', *Journal of the History of Philosophy* 5

(Oct 1967): 343-351.

- Unger, Rudolf. *Hamann und die Aufklärung. Studien zur Vorgeschichte des romantischen Geistes im 18. Jahrhundert.* Bd.1: Text. Bd. 2: Anmerkungen und Beilagen. 2. Aufl. Halle an der Saale: Max Niemeyer 1925. Nachdruck: Darmstadt: Wissenschaftliche Buchgesellschaft 1963.
- Weishoff, Axel. *Wider den Purismus der Vernunft. J.G. Hamanns sakralrhetorischer Ansatz zu einer Metakritik des Kantischen Kritizismus*. Wiesbaden: Westdeutscher Verlag 1998.
- Wessel, Leonard P. 'Hamann's Philosophy of Aesthetics: its meaning for the Storm & Stress Period', *Journal of Aesthetics and Art Criticism* 27 (Summer 1969): 433-443.
- Wohlfart, Günter. *Denken der Sprache: Sprache und Kunst bei Vico, Hamann, Humboldt und Hegel*. Freiburg (u.a.): Alber 1984.
- Wühr, Paul *Ob der Magus in Norden. Selbstgespräch eines Autors mit Johann Georg Hamann.* München: Renner, 1995.

Citations in the article from other sources

- Hegel, G. W. F. 'Über Hamanns Schriften', [1928] in: *Sämtliche Werke*, edited by Hermann Glockner, vol. 20. Stuttgart: Frommann Verlag, 1958.
- Goethe, Johann. *Sämtliche Werke nach Epochen seines Schaffens, Münchener Ausgabe* edited by Gerhard Sauder. 21 volumes. München: Carl Hanser Verlag, 1985-1991.
- Rorty, Richard. *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press, 1979.

Other Internet Sources

- The Hamann-Homepage. This extensive site in German contains much background information, research updates, and digital versions of Hamann's writings along with an extremely thorough bibliography. The website is maintained by those involved in the International Hamann Colloquium, under the leadership of Professor Bernhard Gajek (University of Regensburg, Germany) and with the website maintained by Andre Rudolph (Leipzig).li>
- Mauthner-Gesellschaft, Sprachkritiker website. This site contains an article by Professor Josef Simon on Hamann, in particular in relation to the philosophy of language.
- Internationales Hamann-Kolloquium. The International Hamann Colloquium website, under the directorship of Professor Bernhard Gajek, contains bibliography and information about the colloquium meetings and publications.

Related entries

Berkeley, George | Hegel, Georg Wilhelm Friedrich | Herder, Johann Gottfried von | Hume, David | Jacobi, Friedrich Heinrich | Kant, Immanuel | Kierkegaard, Søren | Rorty, Richard

Gwen Griffith-Dickson
gcgriffithdickson@blueyonder.co.uk

First published: June 29, 2002
Content last modified: June 29, 2002

Richard Rorty

Richard Rorty's distinctive and controversial brand of pragmatism expresses itself along two main axes. One is negative---a critical diagnosis of what Rorty takes to be defining projects of modern philosophy. The other is positive---an attempt to show what intellectual culture might look like, once we free ourselves from the governing metaphors of mind and knowledge in which the traditional problems of epistemology and metaphysics (and indeed, in Rorty's view, the self-conception of modern philosophy) are rooted. The centerpiece of Rorty's critique is the provocative account offered in *Philosophy and the Mirror of Nature* (1979, hereafter PMN). In this book, and in the closely related essays collected in *Consequences of Pragmatism* (1982, hereafter CP), Rorty's principal target is the philosophical idea of knowledge as representation, as a mental mirroring of a mind-external world. Providing a contrasting image of philosophy, Rorty has sought to integrate and apply the milestone achievements of Dewey, Hegel and Darwin in a pragmatist synthesis of historicism and naturalism. Characterizations and illustrations of a post-epistemological intellectual culture, present in both PMN (part III) and CP (xxxvii-xliv), are more richly developed in later works, such as *Contingency, Irony, and Solidarity* (1989, hereafter CIS) and in the three volumes of philosophical papers, *Objectivity, Relativism, and Truth* (1991, hereafter ORT); *Essays on Heidegger and Others* (1991, hereafter EHO); and *Truth and Progress* (1998, hereafter TP). In these writings, ranging over an unusually wide intellectual territory, Rorty offers a highly integrated, multifaceted view of thought, culture, and politics which has made him one of the most widely discussed philosophers writing today.

---

1. Biographical Sketch

Richard Rorty was born on October 4th, 1931, in New York City. He grew up, as he recounts in *Achieving Our Country* (1998, hereafter AC), "on the anti-communist reformist Left in mid-century" (AC

59), within a circle combining anti-Stalinism with leftist social activism. "In that circle," Rorty tells us, "American patriotism, redistributionist economics, anticommunism, and Deweyan pragmatism went together easily and naturally." (AC 61) In 1946 Rorty went to the University of Chicago, to a philosophy department which at that time included Rudolph Carnap, Charles Hartshorne, and Richard McKeon, all of whom were Rorty's teachers. After receiving his BA in 1949, Rorty stayed on at Chicago to complete an M.A. (1952) with a thesis on Whitehead supervised by Hartshorne. From 1952 to 1956 Rorty was at Yale, where he wrote a dissertation entitled "The Concept of Potentiality." His supervisor was Paul Weiss. After the completion of his Ph.D., followed by two years in the army, Rorty received his first academic appointment, at Wellesley College. In 1961, after three years at Wellesley, Rorty moved to Princeton University where he stayed until he went to the University of Virginia, in 1982, as Kenan Professor of the Humanities. Rorty left the University of Virginia in 1998, accepting an appointment in the Department of Comparative Literature at Stanford University. In the course of his career, Rorty has received several academic awards and honours, including a Guggenheim Fellowship (1973-74) and a MacArthur Fellowship (1981-1986). He has held a number of prestigious lectureships, giving, among others, the Northcliffe Lectures at University College, London (1986), the Clark Lectures at Trinity College, Cambridge (1987), and the Massey Lectures at Harvard (1997).

2. Against Epistemology

On Rorty's account, modern epistemology is not only an attempt to legitimate our claim to knowledge of what is real, but also an attempt to legitimate philosophical reflection itself--a pressing task, on many accounts, once the advent of the new science gradually gave content to a notion of knowledge obtained by the methodological interrogation of nature herself. Because the result of this kind of interrogation, theoretical empirical knowledge, is so obviously fruitful, and also carries with it seemingly uncontentious norms of progress, its mere presence poses a legitimation challenge to a form of thought, and claim to knowledge, that is distinct from it. Cartesian epistemology, in Rorty's picture, is designed to meet this challenge. It is sceptical in a fundamental way; sceptical doubts of a Cartesian sort, that is, doubts that can be raised about any set of empirical claims whatever, and so cannot be alleviated by experience, are tailor-made to preserve at once a domain and a job for philosophical reflection. Rorty's aim in PMN is to undermine the assumptions in light of which this double legitimation project makes sense.

2.1 Epistemological Behaviorism

That any vocabulary is optional and mutable is the basic conviction behind Rorty's attack on representationalist epistemology carried out in PMN. It informs, for instance, the genealogy (chapter one) and deconstruction (chapter two) of the concept of mind offered in the book's first part, "Our Glassy Essence." This historicist conviction, however, is not itself a central theme of PMN, and it emerges for

explicit discussion only in the final section of the book, "Philosophy," which is the shortest and in some ways least developed of its three parts. The argumentative core of PMN is found in its second part, "Mirroring". Here Rorty develops and extends a diverse lot of arguments--notably from Wilfrid Sellars, Willard Van Orman Quine, Thomas Kuhn, Ludwig Wittgenstein, and Donald Davidson--into a general critique of the defining project of modern epistemology, viz. the conceptions of mind, of knowledge and of philosophy bequeathed by the 17th and 18th centuries. Rorty's key claim is that "the Kantian picture of concepts and intuitions getting together to produce knowledge is needed to give sense to the idea of 'theory of knowledge' as a specifically philosophical discipline, distinct from psychology." (PMN 168). According to Rorty,

> This is equivalent to saying that if we do not have the distinction between what is "given" and what is "added by the mind," or that between the "contingent" (because influenced by what is given) and the necessary (because entirely "within" the mind and under its control), then we will not know what would count as a "rational reconstruction" of our knowledge. We will not know what epistemology's goal or method could be. (PMN 168-9)

Epistemology, in Rorty's account, is wedded to a picture of mind's structure working on empirical content to produce in itself items--thoughts, representations--which, when things go well, correctly mirror reality. To loosen the grip of this picture on our thinking is to challenge the idea that epistemology--whether traditional Cartesian or 20th century linguistic--is the essence of philosophy. To this end, Rorty combines a reading of Quine's attack on a version of the structure-content distinction in "Two Dogmas of Empiricism" (1952), with a reading of Sellars' attack on the idea of givenness in "Empiricism and the Philosophy of Mind" (1956/1997). On Rorty's reading, though neither Sellars nor Quine is able fully to take in the liberating influence of the other, they are really attacking the same distinction, or set of distinctions. While Quine casts doubt on the notion of structure or meaning which linguistically-turned epistemology had instated in place of mental entities, Sellars, questioning the very idea of givenness, came at the distinction from the other side:

> ...Sellars and Quine invoke the same argument, one which bears equally against the given-versus-nongiven and the necessary-versus-contingent distinctions. The crucial premise of this argument is that we understand knowledge when we understand the social justification of belief, and thus have no need to view it as accuracy of representation. (PMN 170)

The upshot of Quine's and Sellars' criticisms of the myths and dogmas of epistemology is, Rorty suggests, that "we see knowledge as a matter of conversation and of social practice, rather than as an attempt to mirror nature." (PMN 171) Rorty provides this view with a label: "Explaining rationality and epistemic authority by reference to what society lets us say, rather than the latter by the former, is the essence of what I shall call 'epistemological behaviorism,' an attitude common to Dewey and Wittgenstein." (PMN 174)

Epistemological behaviorism leaves no room for the kind of practice-transcending legitimation that Rorty identifies as the defining aspiration of modern epistemology. Assuming that epistemic practices do, or at least can, diverge, it is not surprising that Rorty's commitment to epistemological behaviorism should lead

to charges of relativism or subjectivism. Indeed, many who share Rorty's historicist scepticism toward the transcending ambitions of epistemology--friendly critics like Hilary Putnam, John McDowell and Daniel Dennett--balk at the idea that there are no constraints on knowledge save conversational ones. Yet this is a central part of Rorty's position, repeated and elaborated as recently as in TP. Indeed, he invokes it there precisely in order to deflect this sort of criticism. In "Hilary Putnam and the Relativist Menace," Rorty says:

> In short, my strategy for escaping the self-referential difficulties into which "the Relativist" keeps getting himself is to move everything over from epistemology and metaphysics into cultural politics, from claims to knowledge and appeals to self-evidence to suggestions about what we should try. (TP 57)

That epistemological behaviorism differs from traditional forms of relativism and subjectivism is easier to see in light of Rorty's criticism of the notion of representation, and the cluster of philosophical images which surround it.

2.2 Antirepresentationalism

Rorty's enduring attitude to relativism and subjectivism is that both are products of the representationalist paradigm. Though the theme is explicit in PM and CP ("Pragmatism, Relativism, Irrationalism"), it is with Rorty's later and further appropriation of Davidson that his criticism of the idea of knowledge as representation becomes fully elaborated (ORT "Introduction" and Part II). Drawing on Davidson's criticism of the scheme-content distinction ("On the Very Idea of a Conceptual Scheme") and of the correspondence theory of truth ("The Structure and Content of Truth"), Rorty is able to back up his rejection of any philosophical position or project which attempts to draw a general line between what is made and what is found, what is subjective and what is objective, what is mere appearance and what is real. Rorty's position is not that these conceptual contrasts never have application, but that such application is always context and interest bound and that there is, as in the case of the related notion of truth, nothing to be said about them in general. Rorty's commitment to the conversationalist view of knowledge must therefore be distinguished from subjectivism or relativism, which, Rorty argues, presuppose the very distinctions he seeks to reject. Equally, Rorty's epistemological behaviorism must not be confused with an idealism that asserts a primacy of thought or language with respect to the unmediated world, since this, too, is a position that is undercut by Rorty's Davidsonian position. In light of the view of truth and of meaning that Rorty appropriates from Davidson, his conversationalism is not a matter of giving priority to the subjective over the objective, or to mind's power over world's constraint. Rather it is the other side of his anti-representationalism, which denies that we are related to the world in something other than causal terms. Differently put, Rorty argues that we can give no useful content to the notion that the world, by its very nature, rationally constrains choices of vocabulary with which to cope with it. (TP "The Very Idea of Human Answerability to the World: John McDowell's Version of Empiricism").

2.3 Rationality, Science, and Truth

Attacking the idea that we must acknowledge the world's normative constraint on our belief-systems if we are to be rational subjects, Rorty has drawn a great deal of criticism that takes science, particularly natural science, as its chief reference point. Two general kinds of criticisms are often raised. The first insists that science consists precisely in the effort to learn the truth about how things are by methodically allowing us to be constrained in our beliefs by the world. On this view, Rorty is simply denying the very idea of science. The other kind of criticism seeks to be internal: if Rorty's view of science were to prevail, scientists would no longer be motivated to carry on as they are; science would cease to be the useful sort of thing that Rorty also thinks it is (see, eg., Bernard Williams, "Auto da Fe" in Malachowski). However, Rorty's view of science is more complicated than he himself sometimes implies. He says: "I tend to view natural science as in the business of controlling and predicting things, and as largely useless for philosophical purposes." ("Reply to Hartshorne," Saatkamp 32) Yet he spends a good deal of time drawing an alternative picture of the intellectual virtues that good science embodies (ORT Part I). This is a picture which eschews the notion that science succeeds, when it does, in virtue of being in touch with reality in a special way, the sort of way that epistemologists, when successful, can clarify. It is in this sense specifically that Rorty disavows science as philosophically significant. Good science may nevertheless be a model of rationality, in Rorty's view, exactly in so far as scientific practice has succeeded in establishing institutions conducive to democratic exchange of view.

The provocative and counterintuitive force of Rorty's treatment of rationality and science in terms of conversational ethics is undeniable. It is important to realize, though, that Rorty is not denying that there is any bona fide use of notions like truth, knowledge, or objectivity. Rather his point is that our ordinary uses of these notions always trade for their content and point on particular features of their varying contexts of application. His further point is that when we abstract away from these different contexts and practices, in search of general notions, we are left with pure abstract hypostatizations incapable of providing us with any guide to action at all. The upshot, Rorty holds, is that we simply do not have a concept of objective reality which can be invoked either to explain the success of some set of norms of warrant, or to justify some set of standards over against others. This is perhaps clearest in Rorty's treatment of the concept of truth. With regard to truth, Rorty's rhetoric and philosophical strategy has indeed shifted over the last three decades. As late as in 1982 (in CP) he still attempts to articulate his view of truth by drawing on William James's famous definition in terms of what is good in the way of belief. Soon after this, however, Rorty comes to doubt the point of any theory of truth, and, following Davidson's lead explicitly rejects all attempts to explicate the notion of truth in terms of other concepts. Rorty's mature view of of the point and significance of the concept of truth is first elaborated in "Davidson, Pragmatism and Truth," in ORT. Recent expressions are found in the first of the two Spinoza Lectures given at the University of Amsterdam in 1997, "Is it Desirable to Love truth?", in the paper, "Is Truth a Goal of Inquiry? Donald Davidson versus Crispin Wright" (TP), as well as in the introductions to, respectively, TP and PSH. In these writings Rorty argues that while "truth" has various important uses, it does not itself name a goal towards which we can strive, over and above warrant or justification. His argument is not that truth is reducible to warrant, but that the concept has no deep or substantive criterial content at all. That is, there are only semantic explanations to be offered for why it is the case that a given sentence is true just when its truth conditions are satisfied. So aiming for truth, as opposed to warrant, does not point to a possible line of action, just as we have no measure of our approximation to truth other than increasing warrant. Indeed, for Rorty, this is part of what makes the concept so useful, in a manner

not coincidentally analogous with goodness; it ensures that no sentence can ever be analytically certified as true by virtue of its possession of some other property. Rorty's attitude to the concept of truth has been much criticized, often on the ground that the very notion of warrant, indeed the concept of belief in general, presupposes the notion of truth. However, it may be that we can do justice to these connections without supposing that the notion of truth thus involved backs up the notions of belief and warrant with any substantive normative content of its own. Indeed, that neither the concept of truth, nor those of objectivity and of reality, can be invoked to explain or legitimate our inferential practices and our standards of warrant, is the essence of Rorty's conversationalism, or epistemological behaviorism.

3. Pragmatized Culture

Taking epistemological behaviorism to heart, Rorty urges, means that we can no longer construe the authority of science in terms of ontological claims. Though many disagree, this is not, for Rorty, to denigrate or weaken the authority of science. Indeed, a prominent feature of Rorty's post-metaphysical, post-epistemological culture, is a thoroughgoing Darwinian naturalism.

3.1 Naturalism

To be a naturalist in Rorty's sense,

> is to be the kind of antiessentialist who, like Dewey, sees no breaks in the hierarchy of increasingly complex adjustments to novel stimulation--the hierarchy which has amoeba adjusting themselves to changed water temperature at the bottom, bees dancing and chess players check-mating in the middle, and people fomenting scientific, artistic, and political revolutions at the top. (ORT 109)

In Rorty's view, both Dewey's pragmatism and Darwinism encourage us to see vocabularies as tools, to be assessed in terms of the particular purposes they may serve. Our vocabularies, Rorty suggests, "have no more of a *representational* relation to an intrinsic nature of things than does the anteater's snout or the bowerbird's skill at weaving." (TP 48)

Pragmatic evaluation of various linguistically infused practices requires a degree of specificity. From Rorty's perspective, to suggest that we might evaluate vocabularies with respect to their ability to uncover the truth, would be like claiming to evaluate tools for their ability to help us get what we want--full stop. Is the hammer or the saw or the scissors better--in general? Questions about usefulness can only be answered, Rorty points out, once we give substance to our purposes.

Rorty's pragmatist appropriation of Darwin also defuses the significance of reduction. He rejects as representationalist the sort of naturalism that implies a program of nomological or conceptual reduction to terms at home in a basic science. Rorty's naturalism echoes Nietzsche's perspectivism; a descriptive vocabulary is useful insofar as the patterns it highlights are usefully attended to by creatures with needs and interests like ours. Darwinian naturalism, for Rorty, implies that there is no one privileged vocabulary whose purpose is to serve as a critical touchstone for our various descriptive practices.

For Rorty, then, any vocabulary, even that of evolutionary explanation, is a tool for a purpose, and therefore subject to teleological assessment. Typically, Rorty justifies his own commitment to Darwinian naturalism by suggesting that this vocabulary is suited to further the secularization and democratization of society that Rorty thinks we should aim for. Accordingly, there is a close tie between Rorty's construal of the naturalism he endorses and his most basic political convictions.

3.2 Liberalism

Rorty is a self-proclaimed romantic bourgeois liberal, a believer in piecemeal reforms advancing economic justice and increasing the freedoms citizens are able to enjoy. The key imperative in Rorty's political agenda is the deepening and widening of solidarity. Rorty is sceptical toward radicalism; political thought purporting to uncover hidden, systematic causes for injustice and exploitation, and on that basis proposing sweeping changes to set things right. (ORT Part III; EHO; CIS Part II; AC) The task of the intellectual, with respect to social justice, is not to provide refinements of social theory, but to sensitize us to the suffering of others, and refine, deepen and expand our ability to identify with others, to think of others as like ourselves in morally relevant ways. (EHO Part III; CIS Part III) Reformist liberalism with its commitment to the expansion of democratic freedoms in ever wider political solidarities is, on Rorty's view, an historical contingency which has no philosophical foundation, and needs none. Recognizing the contingency of these values and the vocabulary in which they are expressed, while retaining the commitments, is the attitude of the liberal ironist. (CIS essays 3,4) Liberal ironists have the ability to combine the consciousness of the contingency of their own evaluative vocabulary with a commitment to reducing suffering--in particular, with a commitment to combatting cruelty. (CIS essay 4, ORT Part III) They promote their cause through redescriptions, rather than arguments. The distinction between argumentative discourse and redescription corresponds to that between propositions and vocabularies. Change in belief may result from convincing argument. A change in what we perceive as interesting truth value candidates results from acquiring new vocabularies. Rorty identifies romanticism as the view that the latter sort of change is the more significant. (CIS "Introduction", essay 1).

Rorty's romantic version of liberalism is expressed also in the distinction he draws between the private and the public. (CIS) This distinction is often misinterpreted to imply that certain domains of interaction or behaviour should be exempted from evaluation in moral or political or social terms. The distinction Rorty draws, however, has little to do with traditional attempts to draw lines of demarcation of this sort between a private and a public domain--to determine which aspects of our lives we do and which we don't have to answer for publically. Rorty's distinction, rather, goes to the purposes of theoretical vocabularies. We should, Rorty urges, be "content to treat the demands of self-creation and of human solidarity as equally valid, yet forever incommensurable." (CIS xv) Rorty's view is that we should treat vocabularies

for deliberation about public goods and social and political arrangements, on the one hand, and vocabularies developed or created in pursuit of personal fulfilment, self-creation, and self-realization, on the other, as distinct tools.

3.3 Ethnocentrism

Rorty's liberal ironist, recognizing--indeed, affirming--the contingency of her own commitments, is explicitly ethnocentric. (ORT "Solidarity or Objectivity") For the liberal ironist,

> ...one consequence of antirepresentationalism is the recognition that no description of how things are from a God's-eye point of view, no skyhook provided by some contemporary or yet-to-be-developed science, is going to free us from the contingency of having been acculturated as we were. Our acculturation is what makes certain options live, or momentous, or forced, while leaving others dead, or trivial, or optional. (ORT 13)

So the liberal ironist accepts that bourgeois liberalism has no universality other than the transient and unstable one which time, luck, and discursive effort might win for it. This view looks to many readers like a version of cultural relativism. True, Rorty does not say that what is true, what is good, and what is right is relative to some particular ethnos, and so in that sense he is no relativist. But the worry about relativism, that it leaves us with no rational way to adjudicate conflict, seems to apply equally to Rorty's ethnocentric view. Rorty's answer is to say that in one sense of "rational" that is true, but that in another sense it is not, and to recommend that we drop the former. Rorty's position is that we have no notion of rational warrant that exceeds, or transcends, or grounds, the norms that liberal intellectuals take to define thorough, open-minded, reflective discussion. It is chimerical, Rorty holds, to think that the force or attractiveness of these norms can be enhanced by argument that does not presuppose them. It is pointless, equally, to look for ways of convicting those who pay them no heed of irrationality. Persuasion across such fundamental differences is achieved, if at all, by concrete comparisons of particular alternatives, by elaborate description and redescription of the kinds of life to which different practices conduce.

4. Rorty and Philosophy

The broad scope of Rorty's metaphilosophical deconstruction, together with a penchant for uncashed metaphor and swift, broad-stroke historical narrative, has gained Rorty a sturdy reputation as an anti-philosopher's philosopher. While his writing enjoys an unusual degree of popularity beyond the confines of the profession, Rorty's work is often regarded with suspicion and scepticism within academic philosophy.

4.1 Critical Responses

As we have seen in connection with Rorty's attitude to science, it is particularly Rorty's treatment of truth and knowledge that has drawn fire from philosophers. While a great variety of philosophers have criticized Rorty on this general score in a great variety of ways, it is not very difficult to discern a common concern; Rorty's conversationalist view of truth and knowledge leaves us entirely unable to account for the notion that a reasonable view of how things are is a view suitably constrained by how the world actually is. This criticism is levelled against Rorty not only from the standpoint of metaphysical and scientific realist views of the sort that Rorty hopes will soon be extinct. It is expressed also by thinkers who have some sympathy with Rorty's historicist view of intellectual progress, and his critique of Kantian and Platonist features of modern philosophy. Frank B. Farrell, for instance, argues that Rorty fails to appreciate Davidson's view on just this point, and claims that Rorty's conversationalist view of belief-constraint is a distorted, worldless, version of Davidson's picture of how communication between agents occurs. Similarly, John McDowell, while also critical of Davidson's epistmeological views, claims that Rorty's view of the relation between agent and world as merely causal runs foul of the notion that our very concept of a creature with beliefs involves the idea of a rational constraint of the world on our epistemic states.

However, critics are concerned not only with what they see as a misguided view of belief, truth, and knowledge, whether relativist, subjectivist, or idealist in nature. An important reason for the high temperature of much of the debate that Rorty has inspired is that he appears to some to rejects the very values that are the basis for any articulation of a philosophical view of truth and knowledge at all. Rorty is critical of the role of argument in intellectual progress, and dismissive of the very idea of theories of truth, knowledge, rationality, and the like. Philosophers such as Hilary Putnam and Susan Haack have increasingly focussed on this aspect of Rorty's views. Haack, in particular, frames criticism of Rorty along these lines in moral terms; to her mind, Rorty's efforts to abandon the basic concepts of traditional epistemology are symptoms of a vulgar cynicism, which contributes to the decline of reason and intellectual integrity that Haack and others find to be characteristic of much contemporary thought. The charge of intellectual irresponsibility is sometimes raised, or at last implied, in connection with Rorty's use of historical figures. Rorty's reading of Descartes and of Kant i PMN have often been challenged, as has his more constructive uses of Hegel, Nietzsche, Heidegger, and Wittgenstein. The kind of appropriation of other writers and thinkers that Rorty performs will at times seem to do violence to the views and intentions of the protagonists. Rorty, however, is quite clear about the rhetorical point and scholarly limits of these kinds of redescriptions, as he explains in "The Historiography of Philosophy: Four Genres."

4.2 Claim to Pragmatism

One particularly contentious issue has arisen in connection with Rorty's appropriation of earlier philosophers; prominent readers of the classical American pragmatists have expressed deep reservations about Rorty's interpretation of Dewey and Peirce, in particular, and the pragmatist movement in general. Consequently, Rorty's entitlement to the label "pragmatist" has been challenged. Recently, Susan Haack's

strong claims on this score have received much attention, but there are many others. (See, for example, the discussions of Rorty in Thomas M. Alexander, 1987; Gary Brodsky, 1982; James Campbell, 1984; Abraham Edel, 1985; James Gouinlock, 1995; Lavine 1995; R.W: Sleeper, 1986; as well as the essays in Lenore Langdorf and Andrew R. Smith, 1995.) For Rorty, the key figure in the American pragmatist movement is John Dewey, to whom he attributes many of his own central doctrines. In particular, Rorty finds in Dewey an anticipation of his own view of philosophy as the hand-maiden of a humanist politics, of a non-ontological view of the virtues of enquiry, of a holistic conception of human intellectual life, and of an anti-essentialist, historicist conception of philosophical thought. To read Dewey his way, however, Rorty explicitly sets about separating the "good" from the "bad" Dewey. (See "Dewey's Metaphysics," CP, 72-89, and "Dewey between Hegel and Darwin, in Saatkamp, 1-15.) He is critical of what he takes to be Dewey's backsliding into metaphysics in *Experience and Nature*, and has no patience for the constructive attempt of *Logic: The Theory of Inquiry*. Rorty thus scheme of evaluation on Dewey's works which many scholars object to. Lavine, for instance, claims that "scientific method" is Dewey's central concept (Lavine 1995, 44). R.W. Sleeper holds that reform rather than elimination of metaphysics and epistemology is Dewey's aim (Sleeper 1986, 2, chapter 6).

Rorty's least favourite pragmatist is Peirce, whom he regards as subject to both scheme-content dualism and to a degree of scientism. So it is not surprising that Haack, whose own pragmatism draws inspiration from Peirce, finds Rorty's recasting of pragmatism literally unworthy of the name. Rorty's key break with the pragmatists is a fundamental one; to Haack's mind, by situating himself in opposition to the epistemological orientation of modern philosophy, Rorty ends up dismissing the very project that gave direction to the works of the American pragmatists. While classical pragmatism is an attempt to understand and work out a novel legitimating framework for scientific enquiry, Haack maintains, Rorty's "pragmatism" (Haack consistently uses quotes) is simply an abandonment of the very attempt to learn more about the nature and adequacy conditions of enquiry. Instead of aiding us in our aspiration to be govern ourselves through rational thought, Rorty weakens our intellectual resilience and leaves us even more vulnerable to rhetorical seduction. To Haack and her sympathisers, Rorty's pragmatism is dangerous, performing an end-run on reason, and therefore on philosophy.

4.3 Analytic Philosophy

Nevertheless, the founding impulses of Western philosophy clearly express themselves in Rorty's fundamental concern with the connection between philosophical thinking and the pursuit of human happiness. Rorty's relationship to the traditions of Western philosophy is more nuanced than his reputation might suggest. So, too, is Rorty's relation to analytical philosophy in particular. Rorty is sometimes portrayed as a renegade, as someone who went through a transformation from bona fide analytical philosopher to something else, and then lived to tell a tale of liberation from youthful enchantment. This portrayal, however, distorts both Rorty's view of analytical philosophy and the trajectory of his thinking.

In the mid nineteen sixties, Rorty gained attention for his articulation of eliminative materialism (cf., "Mind-Body Identity, Privacy and Categories," 1965). Around that time, he also edited, and wrote a lengthy introduction to, a volume entitled *The Linguistic Turn* (1967, reissued with a new introduction in

1992). Though the introduction to the 1967 volume and the early papers in philosophy of mind show Rorty adopting frameworks for philosophical problems he has since dispensed with, these writings at the same time clearly bear the mark of the fundamental metaphilosophical attitude which becomes explicit in the next decade. In the "Preface" to PMN, referring to Hartshorne, McKeon, Carnap, Robert Brumbaugh, Carl Hempel, and Paul Weiss, Rorty says,

> I was very fortunate in having these men as my teachers, but, for better or worse, I treated them all as saying the same thing: that a "philosophical problem" was a product of the unconscious adoption of a set of assumptions built into the vocabulary in which the problem was stated--assumptions which were to be questioned before the problem itself was taken seriously. (PMN xiii)

This way of stating the lesson, however, appears to leave open the possibility that certain philosophical problems eventually may legitimately be taken seriously--that is, at face value in the sense that they require constructive solutions--provided the assumptions which sustain their formulation stand up to proper critical scrutiny. Taken this way, the attitude Rorty here expresses would be more or less the same as that of all those philosophers who have diagnosed their predecessors' work as mixtures of pseudo-questions and genuine problems dimly glimpsed, problems which now, with the proper frame of questioning fully clarified, may be productively addressed. But the full force of the lesson Rorty learned emerges only with the view that this notion of proper critical scrutiny is illusory. For Rorty, to legitimate the assumptions on which a philosophical problem is based, would be to establish that the terms we require to pose it are mandatory, that the vocabulary in which we encounter it is in principle inescapable. But Rorty's construal of the linguistic turn, as well as his proposals for eliminating the vocabulary of the mental, are really at odds with the idea that we might hope to construct a definitive vocabulary for philosophy. Even in his early days, Rorty's approach to philosophy is shaped by the historicist conviction that no vocabularies are inescapable in principle. This means that progress in philosophy is gained less from constructive solutions to problems than through therapeutic dissolution of their causes, that is, through the invention of new vocabularies by the launch of new and fruitful metaphors. (PMN "Introduction"; ORT "Unfamiliar Noises: Hesse and Davidson on Metaphor")

To hold that no vocabulary is final is also to hold that no vocabulary can be free of unthematized yet optional assumptions. Hence any effort to circumvent a philosophical problem by making such assumptions visible is subject to its own circumvention. Accordingly, the fact that Rorty often distances himself from the terms in which he earlier framed arguments and made diagnoses is in itself no reason to impose on him, as some have done, a temporal dichotomy. It may be that Rorty's early work, inspired by a less critical, less dialectical reading of Quine and Sellars than that offered in PMN, is more constructive than therapeutic in tone and jargon, and therefore from Rorty's later perspective in an important sense misguided. However, what ties together all Rorty's work, over time and across themes, is his complete lack of faith in the idea that there is an ideal vocabulary, one which contains all genuine discursive options. Rorty designates this faith Platonism (an important theme in CIS). That there are no inescapable forms of description is a thought which permeates Rorty's work from the 1960s right through his later therapeutic articulations of pragmatism. These characterizations of pragmatism in terms of anti-foundationalism (PMN), of anti-representationalism (ORT), of anti-essentialism (TP) are explicitly

parasitic on constructive efforts in epistemology and metaphysics, and are intended to high-light the various ways that these efforts remain under the spell of a Platonic faith in ideal concepts and mandatory forms of descriptions.

Rorty's use of Quine and Sellars to make his fundamental points against the idea of philosophy as a knowledge legitimation project, as well as his articulation of his critique in terms of typically "analytical" philosophical problems, has contributed to an impression of PMN as an internal indictment of analytic philosophy as such. Many--some gleeful, some chagrined--have read PMN as a purported demonstration of the bankruptcy of one of the two contemporary main streams of Western philosophy. Such readers draw support for this view also from the fact that much of Rorty's writings since PMN has been concerned to show the virtues in thinkers like Heidegger and Derrida. (EHO) Twenty years later, however, it seems better not to superimpose the analytic-continental divide onto the message of PMN, or on Rorty. In PMN, his central point is that philosophy needs to break free from the metaphor of mind as a medium of appearances, appearances that philosophy must help us sort into the mere and the reality-corresponding ones. Rorty made this point in a vocabulary that was developed by Anglo-American (whether by birth, naturalization, or late adoption) philosophers in the course of the preceding half-century. It is not necessary, and probably misleading, to see Rorty's criticism of epistemology and the assumptions that make it appear worthwhile as a criticism of a particular philosophical style of philosophy or set of methodological habits. Reading PMN in conjunction with the essays in CP (see particularly essay 4, "Professionalized Philosophy and Transcendentalist Culture", essay 12, "Philosophy in America Today", and also "Introduction"), one quickly sees that the target of PMN cannot be a putative school or branch of the subject called "Analytic Philosophy". Because Rorty thinks philosophy has no essence, has no defining historical task, fails to mark out a special domain of knowledge, and is not, in short, a natural kind (CP 226), he leaves no ground from which to level that sort of critique. Nor is it his intention to do so. Around the time of the publication of PMN, Rorty's view of the matter was ‘that "analytic philosophy" now has only a stylistic and sociological unity’ (CP 217). He then qualifies this point as follows: "In saying....[this], I am not suggesting that analytic philosophy is a bad thing, or is in bad shape. The analytic style is, I think, a *good* style. The *esprit de corps* among analytic philosophers is healthy and useful." (CP 217) However, while Rorty apparently bears no prejudice against analytic philosophy in particular, the very reason for his tolerance--his antiessentialist, historicist view of philosophy and its problems--has for many critics been a point of objection. After his faint praise, Rorty goes on:

> All I am saying is that analytic philosophy has become, whether it likes it or not, the same sort of discipline as we find in the other "humanities" departments--departments where pretensions to "rigor" and to "scientific" status are less evident. The normal form of life in the humanities is the same as that in the arts and in belles-lettres; a genius does something new and interesting and persuasive, and his or her admirers begin to form a school or movement. (CP 217-218)

This is perfectly consonant with the attitude to the notion of philosophical method Rorty expresses 20 years later: "So-called methods are simply descriptions of the activities engaged in by the enthusiastic imitators of one or another original mind--what Kuhn would call the "research programs" to which their

works gave rise." (TP 10) Rorty's metaphilosophical critique, then, is directed not at particular techniques or styles or vocabularies, but toward the idea that philosophical problems are anything other than transient tensions in the dynamics of evolving, contingent vocabularies. If his critique has bite specifically against analytic philosophy, this may be because of a lingering faith in philosophical problems as lasting intellectual challenges that any honest thinker has to acknowledge, and which may be met by making progress in methodology. Rorty himself, however, nowhere says that this faith is part of the essence of analytical philosophy. On the contrary, it would seem that analytical philosophers, people like Sellars, Quine, and Davidson, have provided Rorty with indispensable critical tools in his attack on the epistemological legitimation-project which has been a central concern in philosophy since Descartes.

Bibliography

Works by Rorty

Abbreviations

Objectivity, Relativism, and Truth: Philosophical Papers, Volume 1.Cambridge: Cambridge University Press, 1991.

[PMN] *Philosophy and the Mirror of Nature*. Princeton, NJ: Princeton University Press, 1979.

[CP] *Consequences of Pragmatism*. Minneapolis: University of Minnesota Press, 1982.

[CIS] *Contingency, Irony, and Solidarity*. Cambridge: Cambridge University Press, 1989.

[ORT]

[EHO] *Essays on Heidegger and Others: Philosophical Papers, Volume 2.* [EHO] Cambridge: Cambridge University Press, 1991.

[TP] *Truth and Progress: Philosophical Papers, Volume 3.* Cambridge: Cambridge University Press, 1998.

[AC] *Achieving Our Country: Leftist Thought in Twentieth Century America*. Cambridge, MA: Harvard University Press,1998.

Other Works by Rorty

- "Pragmatism, Categories and Language." *Philosophical Review* 70, April 1961.
- "The Limits of Reductionism." In *Experience, Existence and the Good*, ed. Irwin C. Lieb, Southern Illinois University Press, 1961.
- "Empiricism, Extensionalism and Reductionalism." *Mind* 72, April 1963.
- "Mind-Body Identity, Privacy, and Categories." *Review of Metaphysics* 19, September 1965.
- (ed.), *The Linguistic Turn*. Chicago: University of Chicago Press, 1967. Second, enlarged, edition l992.

- "Incorrigibility as the Mark of the Mental." *Journal of Philosophy* 67, June 1970.
- "In Defence of Eliminative Materialism." *Review of Metaphysics* 24, September 1970.
- "Verificationism and Transcendental Arguments." *Nous* 5, Fall 1971.
- "Indeterminacy of Translation and of Truth." *Synthese* 23, 1972.
- "Criteria and Necessity." *Nous* 7, November 1973.
- with Edward Lee and Alexander Mourelatos, (eds.), *Exegesis and Argument: Essays in Greek Philosophy presented to Gregory Vlastos*. Amsterdam: Van Gorcum, 1973.
- "Transcendental Arguments, Self-Reference and Pragmatism." In *Transcendental Arguments and Science*, ed. Peter Bieri, Rolf P. Hortsman, and Lorentz Kruger. Dordrecht: D. Reidel, 1979.
- "Contemporary Philosophy of Mind." *Synthese* 53, November 1982.
- "The Historiography of Philosophy: Four Genres." In Richard Rorty, J. B. Schneewind and Quentin Skinner, editors, *Philosophy in History*. Cambridge: Cambridge University Press, 1984.
- "Beyond Realism and Anti-Realism." In *Wo steht die Analytische Philosophie heute*? ed. Ludwig Nagl and Richard Heinrich. Vienna: R. Oldenbourg Verlag, Munich, 1986.
- *Hoffnung statt Erkenntnis: Einleitung in die pragmatische Philosophie*. Vienna: Passagen Verlag, 1994. [This volume contains three lectures delivered in Vienna and Paris in 1993, and not published in English. The French version appeared as *L'Espoir au lieu de savoir: introduction au pragmatisme*, Paris: Albin Michel, 1995.
- "Responses." In *Rorty and Pragmatism: The Philosopher Responds to his Critics*, ed. Herman J. Saatkamp, Jr.. Nashville and London: Vanderbilt University Press, 1995.
- "Responses." In *Debating the State of Philosophy: Habermas, Rorty and Kolakowski*, eds. Jozef Niznik and John T. Sanders. Westport: Praeger Publishers, 1996.
- "Responses." In *Deconstruction and Pragmatism*, ed. Chantal Mouffe. London and New York: Routledge, 1996.
- "Introduction." In *Empiricism and the Philosophy of Mind*, by Wilfrid Sellars. Cambridge, Mass. and London: Harvard University Press, 1997.
- *Truth, Politics and 'Post-Modernism.' The 1997 Spinoza Lectures.* Amsterdam: Van Gorcum, 1997.
- *Philosophy and Social Hope*. Penguin, 2000.
- "Responses." In *Richard Rorty: The Philosopher Meets His Critics*, ed. Robert Brandom. Oxford and Cambridge, MA: Blackwell, 2000.

Secondary Literature

- Alexander, Thomas M., *John Dewey's Theory of Art, Experience, and Nature: The Horizons of Feeling*. Albany: State University of New York Press, 1987.
- Brodsky, Gary, "Rorty's Interpretation of Pragmatism." *Transactions of the Charles S. Peirce Society*, 17, 1982.
- Brandom, Robert, ed., *Rorty and His Critics*. Oxford and Cambridge, Mass.: Blackwell, 2000.
- Campbell, James, "Rorty's Use of Dewey." *Southern Journal of Philosophy*, 22, 1984.
- Davidson, Donald, *Inquiries Into Truth and Interpretation*. Oxford and New York: Oxford University Press, 1984.
- Davidson, Donald, "The Structure and Content of Truth." *Journal of Philosophy* 87, June 1990.

- Dewey, John, *Experience and Nature*. In *Later Works of John Dewey*, Vol. 1, Jo Ann Boydston, ed.. Carbondale: Southern Illinois University Press, 1981.
- Dewey, John, *Logic: The Theory of Inquiry*. In *Later Works of John Dewey*, Vol. 12, Jo Ann Boydston, ed.. Carbondale: Southern Illinois University Press, 1986.
- Edel, Abraham, "A Missing Dimension in Rorty's Use of Pragmatism." *Transactions of the Charles S. Peirce Society*. 21, 1985.
- Farrell, Frank B., *Subjectivity, Realism and Postmodernism: The Recovery of the World in Recent Philosophy*. Cambridge: Cambridge University Press, 1994.
- Gouinlock, James, "What is the Legacy of Instrumentalism? Rorty's Interpretation of Dewey." In Herman J. Saatkamp, ed., *Rorty and Pragmatism*. Nashville, TN: Vanderbilt University Press, 1995.
- Haack, Susan, *Evidence and Enquiry: Towards Reconstruction in Epistemology*.. Oxford and Cambridge, MA: Blackwell, 1993.
- Haack, Susan, *Manifesto of a Passionate Moderate*. Chicago and London: The University of Chicago Press, 1998.
- Hall, David L., *Richard Rorty: Poet and Prophet of the New Pragmatism*. Albany, NY: State University of New York Press, 1994.
- Kolenda, Konstantin, *Rorty's Humanistic Pragmatism: Philosophy Democratized*. Tampa: University of South Florida Press, 1990.
- Kulp, Christopher B., ed., *Realism/Antirealism and Epistemology*. Lanham, MD: Rowman and Littlefield Publishers, Inc., 1997.
- Langsdorf, Lenore and Smith, Andrew R., eds., *Recovering Pragmatism's Voice: The Classical Tradition, Rorty, and the Philosophy of Communication*. Albany: State University of New York Press, 1995.
- Lavine, Thelma Z.,"America & the Contestations of Modernity: Bentley, Dewey, Rorty." In Herman J. Saatkamp, ed., *Rorty and Pragmatism*. Nashville, TN: Vanderbilt University Press, 1995.
- McDowell, John, *Mind and World*. Cambridge, MA: Harvard University Press, 1994.
- Malachowsky, Alan R., ed., *Reading Rorty*. Oxford and Cambridge, MA: Blackwell, 1991.
- Mouffe, Chantal, ed., *Deconstruction and Pragmatism*. London and New York: Routledge, 1996.
- Niznik, Jozef and Sanders, John T., eds., *Debating the State of Philosophy: Habermas, Rorty and Kolakowski*. Westport: Praeger Publishers, 1996.
- Nystrom, Derek and Puckett, Kent, *Against Bosses, Against Oligarchies: A Conversation with Richard Rorty*. Charlottesville, VA: Prickly Pear Pamphlets (North America), 1998.
- Peters, Michael and Ghiraldelli, Paulo, eds., *Richard Rorty: Education, Philosophy, Politics*. Lanham, MD: Rowman and Littlefield Publishers, Inc., 2002.
- Pettegrew, John, ed., *A Pragmatist's Progress? Richard Rorty and American Intellectual History*. Lanham, MD: Rowman and Littlefield Publishers, Inc., 2002.
- Prado, C.G., *The Limits of Pragmatism*. Atlantic Highlands, NJ: Humanities Press, 1987.
- Quine, Willard Van Orman, "Two Dogmas of Empiricism." In *From a Logical Point of View*. Cambridge, MA: Harvard University Press, 1953.
- Saatkamp, Herman J., ed., *Rorty and Pragmatism*. Nashville, TN: Vanderbilt University Press, 1995.

- Sleeper, R.W., *The Necessity of Pragmatism*. New Haven and London: Yale University Press, 1986.

Other Internet Resources

- Richard Rorty's Home Page
- American Philosophers (Scott Moore, Baylor University) -->
- Postmodern Thought (University of Colorado, Denver, School of Education)

Related Entries

Davidson, Donald | Dewey, John | liberalism | naturalism | postmodernism | pragmatism | Sellars, Wilfrid

Bjørn Ramberg
b.t.ramberg@filosofi.uio.no

First published: February 3, 2001
Content last modified: February 3, 2001

Stanford Encyclopedia of Philosophy

(ca. 1895, in *The Letters of William James*, ed. by Henry James, Boston, 1920)

William James

William James was an original thinker in and between the disciplines of physiology, psychology and philosophy. His twelve-hundred page masterwork, *The Principles of Psychology* (1890), is a rich blend of physiology, psychology, philosophy, and personal reflection that has given us such ideas as "the stream of thought" and the baby's impression of the world "as one great blooming, buzzing confusion" (PP 462). It contains seeds of pragmatism and phenomenology, and influenced generations of thinkers in Europe and America, including Edmund Husserl, Bertrand Russell, John Dewey, and Ludwig Wittgenstein. James studied at Harvard's Lawrence Scientific School and the School of Medicine, but his writings were from the first as much philosophical as scientific. "Some Remarks on Spencer's Notion of Mind as Correspondence" (1878) and "The Sentiment of Rationality" (1879, 1882) presage his future pragmatism and pluralism, and contain the first statements of his view that philosophical theories are reflections of a philosopher's temperament or vision.

James hints at his religious concerns in his earliest essays and in *The Principles*, but they become more explicit in *The Will to Believe and Other Essays in Popular Philosophy* (1897), *Human Immortality: Two Supposed Objections to the Doctrine* (1898), *The Varieties of Religious Experience* (1902) and *A Pluralistic Universe* (1909). James oscillated between thinking that a "study in human nature" such as *Varieties* could contribute to a "Science of Religion" and the belief that religious experience involves an altogether supernatural domain, somehow inaccessible to science but accessible to the individual human subject. James made some of his most important philosophical contributions in the last decade of his life.

In a burst of writing in 1904-5 (collected in *Essays in Radical Empiricism* (1912)) he set out the metaphysical view most commonly known as "neutral monism," according to which there is one fundamental "stuff" which is neither material nor mental. He also published *Pragmatism* (1907), the culminating expression of a set of views permeating his writings.

Chronology of James's Life

- 1842. Born in New York City, first child of Henry James and Mary Walsh. James. Educated by tutors and at private schools in New York.
- 1843. Brother Henry born.
- 1848. Sister Alice born.
- 1855-8. Family moves to Europe. William attends school in Geneva, Paris, and Boulogne-sur-Mer; develops interests in painting and science.
- 1858. Family settles in Newport, Rhode Island, where James studies painting with William Hunt.
- 1859-60. Family settles in Geneva, where William studies science at Geneva Academy; then returns to Newport when William decides he wishes to resume his study of painting.
- 1861. William abandons painting and enters Lawrence Scientific School at Harvard.
- 1864. Enters Harvard School of Medicine.
- 1865. Joins Amazon expedition of his teacher Louis Agassiz, contracts a mild form of smallpox, recovers and travels up the Amazon, collecting specimens for Agassiz's zoological museum at Harvard.
- 1866. Returns to medical school. Suffers eye strain, back problems, and suicidal depression in the fall.
- 1867-8. Travels to Europe for health and education: Dresden, Bad Teplitz, Berlin, Geneva, Paris. Studies physiology at Berlin University, reads philosophy, psychology and physiology (Wundt, Kant, Lessing, Goethe, Schiller, Renan, Renouvier).
- 1869. Receives M. D. degree, but never practices. Severe depression in the fall.
- 1870-1. Depression and poor health continue.
- 1872. Accepts offer from President Eliot of Harvard to teach undergraduate course in comparative

physiology.

- 1873. Accepts an appointment to teach full year of anatomy and physiology, but postpones teaching for a year to travel in Europe.
- 1874-5. Begins teaching psychology; establishes first American psychology laboratory.
- 1878. Marries Alice Howe Gibbens. Publishes "Remarks on Spencer's Definition of Mind as Correspondence" in *Journal of Speculative Philosophy*.
- 1879. Publishes "The Sentiment of Rationality" in *Mind*.
- 1880. Appointed Assistant Professor of Philosophy at Harvard. Continues to teach psychology.
- 1882. Travels to Europe. Meets with Ewald Hering, Carl Stumpf, Ernst Mach, Wilhelm Wundt, Joseph Delboeuf, Jean Charcot, George Croom Robertson, Shadworth Hodgson, Leslie Stephen.
- 1884. Lectures on "The Dilemma of Determinism" and publishes "On Some Omissions of Introspective Psychology" in *Mind*.
- 1885-92. Teaches psychology and philosophy at Harvard: logic, ethics, English empirical philosophy, psychological research.
- 1890. Publishes *The Principles of Psychology* with Henry Holt of Boston, twelve years after agreeing to write it.
- 1897. Publishes *The Will to Believe and Other Essays in Popular Philosophy*. Lectures on "Human Immortality" (published in 1898).
- 1898. Identifies himself as a pragmatist in "Philosophical Conceptions and Practical Results," given at the University of California, Berkeley. Develops heart problems.
- 1899. Publishes *Talks to Teachers on Psychology: and to Students on Some of Life's Ideals* (including "On a Certain Blindness in Human Beings" and "What Makes Life Worth Living?"). Becomes active member of the Anti-Imperialist League, opposing U. S. policy in Philippines.
- 1901-2. Delivers Gifford lectures on "The Varieties of Religious Experience" in Edinburgh (published in 1902).
- 1904-5 Publishes "Does 'Consciousness' Exist?," "A World of Pure Experience," "How Two Minds Can Know the Same Thing," "Is Radical Empiricism Solipsistic?" and "The Place of Affectional Facts in a World of Pure Experience" in *Journal of Philosophy, Psychology and Scientific Methods*. All were reprinted in *Essays in Radical Empiricism* (1912).
- 1907. Resigns Harvard professorship. Publishes *Pragmatism: A New Name for Some Old Ways of Thinking*, based on lectures given in Boston and at Columbia.
- 1909. Publishes *A Pluralistic Universe*, based on Hibbert Lectures delivered in England and at Harvard the previous year.
- 1910. Publishes "A Pluralistic Mystic" in *Hibbert Journal*. Abandons attempt to complete a "system" of philosophy. (His partially completed manuscript published posthumously as *Some Problems of Philosophy*). Dies of heart failure at summer home in Chocorua, New Hampshire.

Early Writings

"Remarks on Spencer's Definition of Mind as Correspondence" (1878)

James's discussion of Herbert Spencer broaches what will turn out to be characteristic Jamesean themes:

the importance of religion and the passions, the variety of human responses to life, and the idea that we help to "create" the truths that we "register" (E, 21). Taking up Spencer's view that the adjustment of the organism to the environment is the basic feature of mental evolution, James charges that Spencer projects his own vision of what ought to be onto the phenomena he claims to describe. Survival, James asserts, is merely one of many interests human beings have: "The social affections, all the various forms of play, the thrilling intimations of art, the delights of philosophic contemplation, the rest of religious emotion, the joy of moral self-approbation, the charm of fancy and of wit -- some or all of these are absolutely required to make the notion of mere existence tolerable;..." (E, 13). We are all teleological creatures at base, James holds, each with his or her or own *a priori* values and categories. Spencer, then, "merely takes sides with the *telos* he happens to prefer" (E, 18).

James's fundamental empiricism is expressed in the claim that these values and categories fight it out in the course of human experience, that their conflicts "can only be solved *ambulando*, and not by any *a priori* definition." The "formula which proves to have the most massive destiny," he concludes, "will be the true one" (E, 17). Yet James wishes to defend his sense that any such formulation will be determined as much by a freely-acting human mind as by the world, a position he would later (in *Pragmatism*) call "humanism": "there belongs to mind, from its birth upward, a spontaneity, a vote. It is in the game, and not a mere looker-on; and its judgments of the should-be, its ideals, cannot be peeled off from the body of the cogitandum as if they were excrescences..." (E, 21).

"The Sentiment of Rationality" (1879, 82)

This essay was first published in *Mind* in 1879; a second part appeared in the *Princeton Review* in 1882. Both parts were used to construct the essay of this title published in *The Will to Believe and Other Essays in Popular Philosophy* (1897). Although he never quite says that rationality is a sentiment, James holds that a sentiment -- really a set of sentiments -- is a "mark" of rationality. The philosopher, James writes, will recognize the rationality of a conception "as he recognizes everything else, by certain subjective marks with which it affects him. When he gets the marks, he may know that he has got the rationality." These marks include a "strong feeling of ease, peace, rest" (WB 57), and a "feeling of the sufficiency of the present moment, of its absoluteness" (WB 58). "Fluid" thinking, whether logical or cosmological, is said to produce the sentiment, as when we learn that the "the balloon rises by the same law whereby the stone sinks" (WB 59). However, balancing the "passion for parsimony" (WB 58) such unifications satisfy is a passion for distinguishing: a "loyalty to clearness and integrity of perception, dislike of blurred outlines, of vague identifications" (WB 59). The ideal philosopher is a blend of these two passions of rationality, and James thinks that even some great philosophers go too far in one direction or another. Spinoza's unity of all things in one substance is "barren,"; and so is Hume's "'looseness and separateness' of everything..." (WB 60).

Sentiments of rationality operate not just in logic or science, but in ordinary life. When we first move into a room, for example, "we do not know what draughts may blow in upon our back, what doors may open, what forms may enter, what interesting objects may be found in cupboards and corners." These uncertainties, minor as they may be, act as "mental irritant[s]" (WB 67-8), which disappear when we know our way around the room and come to "feel at home" there. These feelings of confident

expectation, of knowing how certain things will turn out, are another form of the sentiment of rationality.

James begins the second part of his essay by considering the case when "two conceptions [are] equally fit to satisfy the logical demand" for fluency or unification. In such a case, one must consider a "practical" component of rationality: "the one which awakens the active impulses, or satisfies other aesthetic demands better than the other, will be accounted the more rational conception, and will deservedly prevail" (WB 66). James here puts the point as one of psychology -- a prediction of what will prevail -- but he also evaluates it, for it will prevail "deservedly." James rejects reductive materialism because it denies to "our most intimate powers...all relevancy in universal affairs" (WB 71), and hence fails to activate these impulses or satisfy these demands.

As in his essay on Spencer, James continues to explore the relations between the temperament that forms the philosopher's outlook and the outlook itself: "Idealism will be chosen by a man of one emotional constitution, materialism by another." James's empathetic understanding extends to both: idealism offers a sense of intimacy with the universe, the feeling that ultimately I "am all." But people of contrasting temperaments find in idealism "a narrow, close, sick-room air," leaving out an element of danger, contingency and wildness -- "the rough, harsh, sea-wave, north-wind element" (WB 75). Both the intimacy and the wildness answer to propensities, passions, and powers in human beings. Although James has his criticisms of reductive materialisms, he understands -- from the inside as it were -- some of the materialistic passion. Those with a materialistic temperament, he writes, "sicken at a life wholly constituted of intimacy," and have a desire "to escape personality, to revel in the action of forces that have no respect for our ego, to let the tides flow, even though they flow over us." The "strife" of these two forms of "mental temper," James predicts, will always be seen in philosophy (WB 76). Certainly they are always seen in the philosophy of William James.

The Principles of Psychology

James's masterwork officially follows the psychological method of introspection, defined as follows: "it means, of course, the looking into our own minds and reporting what we there discover" (PP 185). James is thinking in part of the experiments his contemporaries Wundt, Stumpf and Fechner were performing in their laboratories, which led them to results such as that "sounds are less delicately discriminated in intensity than lights" (PP 513). Although James does discuss these results in detail, many of his most important and memorable introspective observations come from his everyday experience. For example:

> The rhythm of a lost word may be there without a sound to clothe it....Everyone must know the tantalizing effect of the blank rhythm of some forgotten verse, restlessly dancing in one's mind, striving to be filled out with words (PP 244).
>
> Our father and mother, our wife and babes, are bone of our bone and flesh of our flesh. When they die, a part of our very selves is gone. If they do anything wrong, it is our shame. If they are insulted, our anger flashes forth as readily as if we stood in their place. (PP 280).

> There is an excitement during the crying fit which is not without a certain pungent pleasure of its own; but it would take a genius for felicity to discover any dash of redeeming quality in the feeling of dry and shrunken sorrow (PP, p. 1061).

> "*Will you or won't you have it so?*" is the most probing question we are ever asked; we are asked it every hour of the day, and about the largest as well as the smallest, the most theoretical as well as the most practical, things. We answer by *consents or non-consents* and not by words. What wonder that these dumb responses should seem our deepest organs of communication with the nature of things! (PP, p. 1182).

In this last quotation, James tackles a philosophical problem from a psychological perspective. Although he refrains from answering the question of whether these "responses" are in fact deep organs of communication with the nature of things -- reporting only that they seem to us to be so -- in his later writings, such as *Varieties of Religious Experience*, and *A Pluralistic Universe*, he confesses, and to some degree defends, his belief that the question should be answered affirmatively.

In the deservedly famous chapter on "The Stream of Thought" James takes himself to be offering a richer account of experience than those of traditional empiricists such as Hume. He believes relations, vague fringes, and tendencies are as experienced directly (he would later label this fact part of his "radical empiricism.") Rather than a succession of "ideas," James finds a stream, the waters of which blend; and where, because of its position in the flow, each situation is unique. Our consciousness -- or, as he prefers to call it sometimes, our "sciousness" -- , is "steeped and dyed" in the waters of sciousness or thought that surround it. Our psychic life not only has edges, it has rhythm: it is a series of transitions and resting-places, of "flights and perchings" (PP 236). We rest when we remember the name we have been searching for; and we are off again when we hear a noise that might be the baby waking from her nap.

Interest -- and its close relative, attention -- is a major component not only of James's psychology, but of the epistemology and metaphysics that seep into his discussion. A thing, James states in "The Stream of Thought," is a group of qualities "which happen practically or aesthetically to interest us, to which we therefore give substantive names.... (PP 274). And reality "*means simply relation to our emotional and active life...whatever excites and stimulates our interest is real* " (PP 924).

Our capacity for attention to one thing rather than another is for James the sign of an "*active* element in all consciousness,...a spiritual something...which seems to go out to meet these qualities and contents, whilst they seem to come in to be received by it." (PP 285). This "spiritual something" is the target of another passage from James's chapter on "Attention," where he speaks of "a star performer" or "original psychic force" which takes the form not of mere attention, but of "*the effort to attend*." According to James's revisionary account of freedom, we are mostly not free, but we achieve freedom in cases which settle, for a while, the direction or orientation of our lives. In these cases we feel the "sting and excitement of our voluntary life," and we sense that "things are really being decided..." (PP, 429). Faced with the tension between scientific determinism and our belief in our own freedom or autonomy, James

restricts the claims of science, which "must be constantly reminded that her purposes are not the only purposes, and that the order of uniform causation which she has use for, and is therefore right in postulating, may be enveloped in a wider order, on which she has no claims at all" (PP, 1179).

The Principles thus encompasses several metaphysical and methodological stances. James often writes as a scientist, as befits his education in biology and medicine. The second and third chapters are entitled "The Functions of the Brain" and "On Some General Conditions of Brain Activity." James alleges that habit, the subject of the fourth chapter, is "at bottom a physical principle" (PP 110). Yet, James presents himself -- as in his title -- as a psychologist, not as a physiologist. The method of the psychologist "first last and always" is introspection, which sometimes reveals a personal, traditionally subjective stream of thought, but which, as practiced by James, anticipates phenomenology in its pursuit of a more "pure" description of the stream of thought that does not presuppose it to be mental or material. One such passage concerns a

> child newly born in Boston, who gets a sensation from the candle-flame which lights the bedroom, or from his diaper-pin [who] does not feel either of these objects to be situated in longitude 71 W. and latitude 42 N.....The flame fills its own place, the pain fills its own place; but as yet these places are neither identified with, nor discriminated from, any other places. That comes later. For the places thus first sensibly known are elements of the child's space-world which remain with him all his life" (PP 681-2).

This passage, taking off from sensations, is rooted in James's empiricism, but it operates as a counter to traditional empiricism, for which "all our sensations at first appear to us as subjective or internal, and are afterwards and by a special act on our part ‘extradited’ or ‘projected’ so as to appear located in an outer world" (PP 678). In contrast, James's view is that the outer and inner worlds are later constructions from the original data of consciousness -- which are neither objective nor subjective. This psychological view anticipates James's late metaphysical "pure experience" account, published in *Essays in Radical Empiricism*.

Two noteworthy chapters late in *The Principles* are entitled "The Emotions" and "Will." The first sets out the theory -- also enunciated by the Danish physiologist Carl Lange -- that emotion follows, rather than causes, its bodily expression: "Common-sense says, we lose our fortune, are sorry and weep; we meet a bear, are frightened and run; we are insulted by a rival, are angry and strike. The hypothesis here to be defended says that this order of sequence is incorrect...that we feel sorry because we cry, angry because we strike, afraid because we tremble....." (PP, pp. 1065-6). The significance of this view, according to James, is that our emotions are tied in with our bodily expressions. What, he asks, would grief be "without its tears, its sobs, its suffocation of the heart, its pang in the breast-bone?" Not an emotion, James answers, for a "purely disembodied human emotion is a nonentity" (PP 1068). As Wittgenstein was to suggest, the human body may be "the best picture of the human soul" (Ludwig Wittgenstein, *Philosophical Investigations*, New York, Macmillan, 1968, p. 178).

In his chapter on "Will" James opposes the theory of his contemporary Wilhelm Wundt that there is one

special feeling -- a "feeling of innervation" -- present in all intentional action. In his survey of a range of cases, James finds that some actions involve an act of resolve or of outgoing nervous energy, but others do not. For example:

> I sit at table after dinner and find myself from time to time taking nuts or raisins out of the dish and eating them. My dinner properly is over, and in the heat of the conversation I am hardly aware of what I do; but the perception of the fruit, and the fleeting notion that I may eat it, seem fatally to bring the act about. There is certainly no express fiat here;..." (PP 1131).

The chapter on "Will" also contains striking passages that anticipate the concerns of *The Varieties of Religious Experience*: about moods, "changes of heart," and "awakenings of conscience." These, James observes, may affect the "whole scale of values of our motives and impulses (PP 1140).

The Will to Believe and Other Essays in Popular Philosophy

This popular and influential collection includes "The Sentiment of Rationality" (discussed above), and essays on "The Dilemma of Determinism," "Great Men and Their Environment" and "Is Life Worth Living?" Its most famous essay gives the collection its title, although James later wrote that he should have called the essay "the *right* to believe," indicating his justificatory intent.

In science, James notes, we can afford to await the outcome of investigation before coming to a belief. The cases of belief formation and justification to which James draws attention in "The Will to Believe" are, in contrast, "forced" -- we must come to some belief even if all the relevant evidence is not in. If I am on a mountain trail, faced with an icy ledge to cross, and do not know whether I can make it, I may be forced to consider the question whether I can or should believe that I can cross the ledge. This is a "momentous" question: if I am wrong I may fall to my death, and if I believe rightly that I can cross the ledge, my holding of the belief may itself contribute to my success. In such a case, James asserts, I have the "right to believe" -- precisely because such a belief may help bring about the fact believed in. This is a case "where a fact cannot come at all unless a preliminary faith exists in its coming" (WB 25).

James applies his analysis to religious belief, particularly to the possible case in which one's salvation depends on believing in God in advance of any proof that God exists. In such a case the belief may be justified by the outcome to which having the belief leads. James also takes his analysis outside of the religious domain, to a wide range of secular human life:

> A social organism of any sort is what it is because each member proceeds to his own duty with a trust that the other members will simultaneously do theirs.... A government, an army, a commercial system, a ship, a college, an athletic team, all exist on this condition, without which not only is nothing achieved, but nothing is even attempted (WB, 24).

James tries to justify the systems of "trust" on which society's beliefs and actions are based. Yet at the same time there is an intensely personal, even existential, tone to his essay, epitomized by James's appeal near the end to "respect one another's mental freedom," and by his concluding citation of Fitz James Stephen's statement that "‘In all important transactions of life we have to take a leap in the dark....’"(WB 31).

Another essay in the collection, "Reflex Action and Theism," attempts a reconciliation of science and religion. By "reflex action" James understands the biological picture of the organism as responding to sensations with a series of actions. Human beings and other higher animals have evolved a theoretical or thinking stage between the sensation and the action, and it is here that God comes up, at least for human beings. James holds that the thought of God is a natural human response to the universe, regardless of whether we can prove that God exists. God will be, as James puts it, the "centre of gravity of all attempts to solve the riddle of life" (WB, 116). James ends the essay by advocating a "theism" that posits "an ultimate opacity in things, a dimension of being which escapes our theoretic control" (WB, 143).

The Will to Believe also contains James's most developed account of morality, "The Moral Philosopher and the Moral Life." Morality for James rests on sentience -- without it there are no moral claims and no moral obligations. But once sentience exists, a claim is made, and morality gets "a foothold in the universe" (WB 198). Although James insists that there is no common essence to morality, he does find a guiding principle for ethical philosophy in the principle that we "satisfy at all times as many demands as we can" (WB, 205). This satisfaction is to be arrived at by working towards a "richer universe...the good which seems most organizable, most fit to enter into complex combinations, most apt to be a member of a more inclusive whole" (WB, 210). We arrive at laws and moral formulations by a kind of "experiment," James holds. (WB, 206). By such experiments, we have liberated ourselves for the most part from slavery and "arbitrary royal power" (WB, 205). But there is nothing final about the results: "as our present laws and customs have fought and conquered other past ones, so they will in their turn be overthrown by any newly discovered order which will hush up the complaints that they still give rise to, without producing others louder still" (WB, 206).

The Varieties of Religious Experience

Like *The Principles of Psychology*, *Varieties* is "A Study in Human Nature," as its subtitle says. But at some five hundred pages it is only half the length of *The Principles of Psychology*, befitting its more restricted, if still immense, scope. For James studies that part of human nature that is, or is related to, religious experience. His interest is not in religious institutions, ritual, or, even for the most part, religious ideas, but in "the feelings, acts, and experiences of individual men in their solitude, so far as they apprehend themselves to stand in relation to whatever they may consider the divine" (V, 31).

James sets out a central distinction of the book in early chapters on "The Religion of Healthy-Mindedness" and "The Sick Soul." The healthy-minded religious person -- Walt Whitman is one of James's main examples -- has a deep sense of "the goodness of life," (79) and a soul of "sky-blue tint"

(80). Healthy-mindedness can be involuntary, just natural to someone, but often comes in more willful forms. Liberal Christianity, for example, represents the triumph of a resolute devotion to healthy-mindedness over a morbid "old hell-fire theology" (91). James also cites the "mind-cure movement" of Mary Baker Eddy, for whom "evil is simply a lie, and any one who mentions it is a liar" (107). This remark allows us to draw the contrast with the religion of "The Sick Soul," for whom evil cannot be eliminated. From the perspective of the sick soul, "radical evil gets its innings" (163). No matter how secure one may feel, the sick soul finds that "[u]nsuspectedly from the bottom of every fountain of pleasure, as the old poet said, something bitter rises up: a touch of nausea, a falling dead of the delight, a whiff of melancholy...." These states are not simply unpleasant sensations, for they bring "a feeling of coming from a deeper region and often have an appalling convincingness" (136). James's main examples here are Leo Tolstoy's "My Confession," John Bunyan's autobiography, and a report of terrifying "dread" -- allegedly from a French correspondent but actually from James himself. Some sick souls never get well, while others recover or even triumph: these are "twice-born." In chapters on "The Divided Self, and the Process of Its Unification" and on "Conversion," James discusses St. Augustine, Henry Alline, Bunyan, Tolstoy, and a range of popular evangelists, focusing on what he calls "the state of assurance" (241) they achieve. Central to this state is "the loss of all the worry, the sense that all is ultimately well with one, the peace, the harmony, the *willingness to be*, even though the outer conditions should remain the same" (248).

Varieties' classic chapter on "Mysticism" offers "four marks which, when an experience has them, may justify us in calling it mystical..." (380). The first is ineffability: "it defies expression...its quality must be directly experienced; it cannot be imparted or transferred to others." Second is a "noetic quality": mystical states present themselves as states of knowledge. Thirdly, mystical states are transient; and, fourth, subjects are passive with respect to them: they cannot control their coming and going. Are these states, James ends the chapter by asking, "windows through which the mind looks out upon a more extensive and inclusive world[?]" (428).

In chapters entitled "Philosophy" -- devoted in large part to pragmatism -- and "Conclusions," James finds that religious experience is on the whole useful, even "amongst the most important biological functions of mankind," but he concedes that this does not make it true. James articulates his own belief -- which he does not claim to prove -- that religious experiences connect us with a greater, or further, reality not accessible in our normal cognitive relations to the world: "The further limits of our being plunge, it seems to me, into an altogether other dimension of existence from the sensible and merely ‘understandable’ world" (515).

Late Writings

Pragmatism (1907)

James first announced his commitment to pragmatism in a lecture given at Berkeley in 1898, entitled "Philosophical Conceptions and Practical Results." Other sources for *Pragmatism* were lectures given at Wellesley College in 1905, and at the Lowell Institute and Columbia University in 1906 and 1907.

Pragmatism emerges in James's book as five things: a philosophical temperament, a theory of truth, a theory of meaning, a holistic account of knowledge, and a method of resolving philosophical disputes.

The pragmatic temperament appears in the book's opening chapter, where James classifies philosophers according to their tough-or tender-mindedness. The pragmatist is a mediator like James himself, someone with "scientific loyalty to facts" as well as "the old confidence in human values and the resultant spontaneity, whether of the religious or romantic type" (P, 17). The method of resolving disputes and the theory of meaning are on display in James's discussion of an argument about whether a man chasing a squirrel around a tree goes around the squirrel too. Taking meaning as the "conceivable effects of a practical kind the object may involve," the pragmatist philosopher finds that two "practical" meanings of "go around" are in play: either the man goes North, East, South, and West of the squirrel, or he faces first the squirrel's head, then one of his sides, then his tail, then his other side. "Make the distinction," James writes, "and there is no occasion for any further dispute."

The pragmatic theory of truth is the subject of the book's sixth (and to some degree its second) chapter. Truth, James holds, is "a species of the good," like health. Truths are goods because we can "ride" on them into the future without being unpleasantly surprised. They "lead us into useful verbal and conceptual quarters as well as directly up to useful sensible termini. They lead to consistency, stability and flowing human intercourse. They lead away from excentricity and isolation, from foiled and barren thinking" (103). James holds that truths are "*made*" (104) in the course of human experience; yet although they live for the most part "on a credit system" in that they are not currently being verified by most of those who have them, "beliefs verified concretely by *somebody* are the posts of the whole superstructure" (P, 100).

James's chapter on "Pragmatism and Humanism" sets out James's voluntaristic epistemology. "We carve out everything," he states, "just as we carve out constellations, to serve our human purposes" (P, 100). Nevertheless, he recognizes "resisting factors in every experience of truth-making" (P, 117), including not only our present sensations or experiences but the whole body of our prior beliefs. James holds neither that we create our truths out of nothing, nor that truth is entirely independent of humanity. He embraces "the humanistic principle: you can't weed out the human contribution" (P, 122). *Pragmatism's* final chapter on "Pragmatism and Religion" follows James's line in *Varieties* in attacking "transcendental absolutism" for its unverifiable account of God, and in defending a "pluralistic and moralistic religion" (144) based on human experience. "On pragmatistic principles," James writes, "if the hypothesis of God works satisfactorily in the widest sense of the word, it is true" (143).

A Pluralistic Universe (1909)

Originally delivered in Oxford as a set of lectures "On the Present Situation in Philosophy," James's book begin with a discussion of philosophic temperament. He condemns the "over-technicality and consequent dreariness of the younger disciples at our American universities..." (PU 13), and holds that a philosopher's "vision" is "the important thing" about him (PU 3). Passing from critical discussions of Royce's idealism and the "vicious intellectualism" of Hegel, James comes, in chapters four through six,

to philosophers whose vision he admires. He praises Gustav Fechner for holding that "the whole universe in its different spans and wave-lengths, exclusions and developments, is everywhere alive and conscious" (PU, 70), and he seeks to refine and justify Fechner's idea that separate human, animal and vegetable consciousnesses meet or merge in a "consciousness of still wider scope" (72). James employs Henri Bergson's critique of "intellectualism" in this project, for Bergson shows that the "concrete pulses of experience appear pent in by no such definite limits as our conceptual substitutes are confined by. They run into one another continuously and seem to interpenetrate" (PU 127). James concludes by embracing a position that he had more tentatively set forth in *The Varieties of Religious Experience*: that religious experiences "point with reasonable probability to the continuity of our consciousness with a wider spiritual environment from which the ordinary prudential man (who is the only man that scientific psychology, so called, takes cognizance of) is shut off" (PU, 135). Whereas in *Pragmatism* James subsumes the religious within the pragmatic (as yet another way of successfully making one's way through the world), in *A Pluralistic Universe*, James distinguishes the religious from the "prudential" or "practical."

Essays in Radical Empiricism (1912)

James's radical empiricism, which is basically an epistemological doctrine, has been confused with the metaphysical doctrine of "pure experience," largely because the latter is set forth in most of the essays collected after James's death in *Essays in Radical Empiricism*. Radical empiricism is never precisely defined in the *Essays*, being best explicated by a passage from *The Meaning of Truth* in which James states that it consists of a postulate, a statement of fact, and a conclusion. The postulate is that "the only things that shall be debatable among philosophers shall be things definable in terms drawn from experience," the fact is that relations are just as directly experienced as the things they relate, and the conclusion is that "the parts of experience hold together from next to next by relations that are themselves parts of experience" (MT, 6-7).

James's "pure experience" doctrine is the view that Bertrand Russell -- giving full credit to James in *The Analysis of Mind* -- calls "neutral monism." James holds that mind and matter are both aspects of, or structures formed from, a more fundamental stuff -- pure experience -- that (despite being called "experience") is neither mental nor physical. Pure experience, James explains, is "the immediate flux of life which furnishes the material to our later reflection with its conceptual categories... a *that* which is not yet any definite *what*, tho' ready to be all sorts of whats..." (ERE, 46). That "whats" pure experience may be include minds and bodies, people and material objects. The "what" of pure experience depends not on a fundamental ontological difference among experiences, but on the *relations* into which pure experiences enter. Certain sequences of pure experiences constitute physical objects, and others constitute persons; but one pure experience (say the perception of a chair) may be part both of the sequence constituting the chair and of the sequence constituting a person.

Bibliography

- Works by William James

- Selected Writings on James

Works by William James

- *The Works of William James*, Cambridge, MA and London: Harvard University Press, 17 vol., 1975--.
- *William James: Writings 1878-1899*. New York: Library of America, 1992.
- *William James: Writings 1902-1910*. New York: Library of America, 1987
- "Remarks on Spencer's Definition of Mind as Correspondence," first published in *The Journal of Speculative Philosophy*, 1878. Contained in *Essays in Philosophy*, pp. 7-22.
- *The Principles of Psychology*, Cambridge, MA: Harvard University Press, 1981. Originally published in 1890 [PP].
- *The Will to Believe and Other Essays in Popular Philosophy*, Cambridge, MA and London: Harvard University Press, 1979; first published in 1897 [WB].
- "Philosophical Conceptions and Practical Results," 1898. Contained in *Pragmatism*, in *The Works of William James*, pp. 255-70.
- *Pragmatism*. Cambridge, MA: Harvard University Press, 1979. Originally published in 1907 [P].
- *A Pluralistic Universe*. Cambridge, MA: Harvard University Press, 1977. Originally published in 1909 [PU].
- *The Meaning of Truth*, Cambridge, MA and London: Harvard University Press, 1979 [MT]. Originally published in 1909.
- *Essays in Philosophy*. Cambridge, MA and London: Harvard University Press, 1978 [E].
- *Some Problems of Philosophy*. Cambridge, MA and London: Harvard University Press, 1979. Originally published in 1911.
- *The Letters of William James*, ed. Henry James, Boston: Little Brown, 1926.
- *The Correspondence of William James*, ed. Ignas K. Skrupskelis and Elizabeth M. Berkeley, 12 volumes. Charlottesville and London, University Press of Virginia, 1992--. (Volumes 1-3, William and Henry; Volume 4 1856-77.)
- *Selected Letters of William and Henry James*, Charlottesville and London, University Press of Virginia, 1997.

Selected Writings on James

- Barzun, Jacques, *A Stroll with William James*. New York: Harper and Row, 1983.
- Bird, Graham, *William James* (The Arguments of the Philosophers). London: Routledge and Kegan Paul, 1986.
- Edie, James, *William James and Phenomenology*, Indianapolis: Indiana University Press, 1987.
- Feinstein, Howard M., *Becoming William James*. Ithaca, NY: Cornell University Press, 1984.
- Fontinell, Eugene, *Self, God, and Immortality: A Jamesian Investigation*. Philadelphia: Temple University Press, 1986.
- Gale, Richard M., *The Divided Self of William James*, Cambridge: Cambridge University Press, 1999.

- Goodman, Russell B., *American Philosophy and the Romantic Tradition*. Cambridge: Cambridge University Press, 1990, Chapter 3.
- -------------. "What Wittgenstein Learned From William James," *History of Philosophy Quarterly*, 11: 3, 1994, pp. 339-54.
- Levinson, Henry S., *The Religious Investigations of William James*. Chapel Hill: University of North Carolina Press, 1981.
- Matthiessen, F. O., *The James Family*. New York: Knopf, 1947.
- McDermott, John, *Streams of Experience: Reflections on the History and Philosophy of American Culture*. Amherst: University of Massachusetts Press, 1986.
- Moore, G. E., "William James' ‘Pragmatism’", in *Philosophical Studies*, London: Routledge and Kegan Paul, 1922, pp. 138.
- Myers, Gerald, *William James: His Life and Thought*. New Haven: Yale University Press, 1986.
- Perry, Ralph Barton, *The Thought and Character of William James*. Boston: Little, Brown, 1935, 2 vols.
- Poirier, Richard, *Poetry and Pragmatism*. Cambridge, MA: Harvard University Press, 1992.
- Putnam, Hilary, *The Many Faces of Realism*. La Salle, IL: Open Court, 1987.
- -------------, (with Ruth Anna Putnam), "William James's Ideas," in Putnam, Hilary, *Realism with a Human Face* (Cambridge, MA: Harvard University Press, 1990, pp. 217-231.
- Putnam, Ruth Anna, *The Cambridge Companion to William James*. Cambridge: Cambridge University Press, 1997.
- Russell, B., *The Collected Papers of Bertrand Russell*, Vol. 6, London: George Allen and Unwin, 1986, pp. 257-306.
- Simon, Linda, *Genuine Reality: a life of William James* , New York: Harcourt Brace, 1998.
- Seigfried, Charlene Haddock, *Chaos and Context*. Athens: Ohio University Press, 1978.
- ----------, *William James's Radical Reconstruction of Philosophy*. Albany: State University of New York Press, 1990.
- Sprigge, T. L. S., *James and Bradley: American Truth and British Reality*. Chicago: Open Court, 1993.
- Suckiel, Ellen Kappy, *The Pragmatic Philosophy of William James*, Notre Dame, IN and London: University of Notre Dame Press, 1982.
- ----------, *Heaven's Champion*. Notre Dame, IN and London: University of Notre Dame Press, 1996.
- Taylor, Eugene, *William James on Consciousness Beyond the Fringe*. Princeton, NJ: Princeton University Press, 1996.
- Wilshire, Bruce, *William James and Phenomenology: A Study of "The Principles of Psychology"*. New York: AMS Press, 1979.

Other Internet Resources

- Professor Frank Pajares' web page on William James (Emory University)

Related Entries

Dewey, John | Husserl, Edmund | pluralism | pragmatism | Russell, Bertrand | Wittgenstein, Ludwig

Russell Goodman
rgoodman@unm.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 7, 2000
Content last modified: September 7, 2000

Panpsychism

Panpsychism is the doctrine that mind is a fundamental feature of the world which exists throughout the universe. Unsurprisingly, each of the key terms, "mind", "fundamental" and "throughout the universe" is subject to a variety of interpretations by panpsychists, leading to a range of possible philosophical positions. For example, an important distinction is that between conscious and unconscious mental states, and appeal to it allows a panpsychism which asserts the ubiquity of the mental while denying that consciousness is similarly widespread. Interpretations of "fundamental" range from the inexplicability of mentality in other, and non-mentalistic, terms to the idealist view that in some sense everything that exists is, and is only, a mental entity. And, although the omnipresence of the mental would seem to be the hallmark feature of panpsychism, there have been versions of the doctrine that make mind a relatively rare and exceptional feature of the universe.

Against the backdrop of our immense scientific knowledge of the physical world, and the corresponding widespread desire to explain everything ultimately in physical terms, panpsychism has come to seem an implausible view. Nonetheless, the doctrine retains some attractive and interesting features. The recalcitrance of the mind, and especially consciousness, to fit smoothly into the scientific picture recommends our consideration of them.

1. Panpsychism and the Scientific World View

It is salutary to remember that not so very long ago anti-physicalism was orthodox philosophical opinion. The nature of mind or consciousness was believed to be entirely distinct from physical nature. It was sometimes allowed, by Cartesian dualists, that mind interacted with the physical world causally under very rare conditions, but the mode of interaction was impossible to understand and seemed to conflict with elementary principles of physics. These worries led to such bizarre views as Malebranche's occasionalism, in which God had to intervene between volition and action, between stimulus and sensation. A yet more extreme position, idealism, which became widespread in the nineteenth century and retained much support well into the twentieth, held that mind stood as the sole ontological foundation of reality, supporting a physical world conceived of as entirely constructed -- somehow -- out of mental phenomena. Such positions have lost much of their philosophical attractiveness and there can be little doubt that the primary reason for this radical change of philosophical opinion is that a principled separation of mind and matter precludes any deep integration of mind with the ever expanding and ever more powerful scientific picture of the physical world. For most philosophers the mind-body problem has become this problem of integration and a plethora of theories, which can be gathered together under the title of physicalism, have arisen in the attempt at its solution.

Broadly speaking, there are, at bottom, only two positions that can promise the desired integration: panpsychism and emergentism. If one believes that the most fundamental physical entities (quarks, leptons, bosons, or whatever physics will ultimately settle upon) are devoid of any mental attributes, and if one also believes that some systems of these entities, such as human brains, *do* possess mental attributes, one is espousing some kind of doctrine of the *emergence* of mind. All the currently popular physicalist theories (such as behaviorism, central state identity theories, functionalism) are theories which attempt to provide an account of how the mental emerges from the physical. Other, more radical, forms of emergentism are possible. These are theories that deny that there is any explanatory account of how emergence works: it is a brute fact of natural law that certain configurations of physical entities underpin certain mental states. This fact is, or has to be accepted, as Samuel Alexander said, with "natural piety". Of course, an inexplicable or brute emergence is still a form of emergence. It is possible to mark the distinction between the forms of emergence in terms of the *explanatory* role of the mental. Modern materialists do not regard the emergent features of the mind as either ontologically or explanatorily fundamental. The more radical emergentists would regard mind as explanatorily fundamental, but not ontologically basic in the sense that material conditions (or in general some system of non-mentalistic conditions) are required for the existence of mental features. The panpsychist naturally regards mind as both explanatorily and ontologically fundamental in the sense just mentioned.

The panpsychist alternative disputes the intelligibility of emergence whether based on the claim that the nature of emergence is simply inexplicable or the claim that the mental can be reduced to a set of relations amongst purely physical entities, and thus must opt instead for the attribution of mentalistic properties to the physically fundamental entities.[1]

It always remains possible to give up the job of integrating mind into our scientific picture of the physical world in favor of accepting some more or less remote relation between independent domains of matter and mind. Cartesian dualism can be seen as involving such a refusal which in the case of Descartes was

entirely self-conscious. But integrators seem to be stuck with the dilemma of emergence versus panpsychism. By and large, the twentieth century witnessed the victory of emergentism, articulated in a bewildering, and still expanding, variety of forms.

Since panpsychism is, by definition, the doctrine that *mind*, in some sense of the term, is *everywhere*, in some sense of that term, it is worth mentioning a complication which is a possible source of confusion at the outset. There have been some panpsychists who, while being much more liberal than most in their willingness to ascribe mind, seem to have been unwilling to extend mind right down into the roots of the world. Both Gustav Fechner (1801-1887) and Josiah Royce (1855-1916) developed panpsychist accounts of nature that did not necessarily attribute mental properties to the ultimate constituents of mentalistic "systems". It would seem to be intuitively clear that if one does not place mind at the very foundation, and in fact regards mentality to be a feature of systems of non-mentalistic entities, then one is an emergentist. Crudely put, someone who believes that amoebas have experiences, but that quarks and electrons, which ultimately constitute amoebas, do not is no panpsychist. However, this simplifying view contains an implicit assumption about the nature of the fundamental physical constituents of the world, namely that the unobservable and hypothetical entities postulated by physics are entirely real and are indeed the ontological foundation of the world. In the nineteenth century, the time of panpsychism's greatest flourishing, this was a rather more daring assumption than it is today. Furthermore, underlying metaphysical assumptions, in particular various forms of idealism, can provide an overarching argument for panpsychism, and Royce (among many other panpsychists of the nineteenth century, if indeed not most thinkers of the period) was an idealist. Fechner's panpsychism was also distinctive in its endorsement of a "world-soul" or "world-mind" of which everything is a part (there are obvious echoes of Spinoza in such a view). This rather top-down view of the place of mind in the world does seem to be a legitimate sort of panpsychism, and it is one that does not require that everything in the world be itself enminded. Nonetheless, without a basis in an explicitly idealist philosophy, it leaves entirely open the question of how the word-soul comes into being. If, for example, the world-mind is conditioned by the structure of its non-mentalistic parts we would seem to be returned to some form of emergentism. In any case, the arguments of such "synechological" panpsychists (as Hartshorne (1950) labels them in contrast with "atomistic" or "monadological" panpsychists) are arguments for panpsychism in general and can be examined as such.

Panpsychism's assertion that mind suffuses the universe presents a fundamental and sharp contrast with its basic rival, emergentism, which asserts that mind appears only at certain times, in certain places under certain -- probably very special and very rare -- conditions. But trying to explicate a little more precisely the key terms of this vague characterization of panpsychism results in several different versions of it. A cardinal distinction within the realm of the mind, though one that still carries more than a whiff of controversy, is that between conscious and unconscious mental states, and thus we could wonder whether panpsychism claims that consciousness is everywhere or merely that some unconscious form of mentality (often labelled *proto-mentality*) lurks throughout the universe. With regard to the ubiquity of the mental, we might wonder whether every *thing* has a mind (or associated mental attributes) or whether there is, even from within a panpsychist view of the world, a viable distinction between things with minds and things lacking minds (as we have seen, the world-mind form of panpsychism may have the resources to fund such a distinction). We might go so far as to wonder whether mind is to be thought of as some kind

of *field-like* entity or in analogy with something as fundamental as energy[2], spread out over the universe and not connected directly with or dependent upon any particular things. Although seldom clearly distinguished, the history of panpsychism reveals that all of these variants have been tried out. So before turning to the virtues and vices of panpsychism, let's look to its past.

2. Early History of Panpsychism

Panpsychism seems to be such an ancient doctrine that its origins long precede any records of systematic philosophy. Some form of animism, which, insofar as it is any kind of *doctrine* at all, is very closely related to panpsychism, seems to be an almost universal feature of pre-literate societies, and studies of human development suggest that children pass through an animist phase, in which mental states are attributed to a wide variety of objects quite naturally (see Piaget 1929).[3] It is tempting to speculate that the basic idea of panpsychism arose in what is a common process of explanatory extension based upon the existence of what is nowadays called "folk psychology". It would have been difficult for our ancestors, in the face of a perplexing and complex world, to resist applying one of the few systematic, and highly successful, modes of explanation in their possession.

In any event, clear indications of panpsychist doctrines are evident in early Greek thought. One of the first presocratic philosophers of ancient Greece, Thales (c. 624-545 BC) deployed an analogical argument for the attribution of mind that tends towards panpsychism. The argument depends upon the idea that enminded beings are self-movers. Thales notes that magnets and, under certain circumstances, amber, can move themselves and concludes that they therefore possess minds. It is claimed that Thales went much beyond such particular attributions and endorsed a true panpsychism and pantheism. For example, as reported by Barnes (1982, pp. 96-7), Diogenes claimed that Thales believed that "the universe is alive and full of spirits" but this remark is derived from an earlier claim of Aristotle: "some say a soul is mingled in the whole universe -- which is perhaps why Thales thought that everything is full of gods". While Barnes disputes the pantheistic reading of Thales, he allows that Thales believed in the "ubiquity of animation" and hence by the above argument accepted a true panpsychism.[4]

Of greater interest is the role of ancient panpsychism in the much wider debate between panpsychism and emergentism. This basic conflict is not merely a reflection of a problem about the mind's place in the world but rather represents a fundamental distinction within our schemes for understanding the world. We like to break the world down into bits and pieces, and then face the problem of retrieving the explanatory target properties from the simpler properties of the bits. This mode of explanation began to be codified with the Presocratics who to their everlasting credit strove to produce comprehensive and refreshingly naturalistic accounts of the world. Their accounts are obscure and less than rigorously scientific, and littered with oracular pronouncements notoriously difficult to interpret, but they nonetheless do mark the beginning of the attempt to give a scientific account of the world.

The Presocratics immediately recognized the basic dilemma: either mind (or, more generally, whatever the apparently "macroscopic", "high-level", or non-fundamental property at issue) is an elemental feature of the world or it somehow emerges from, or is conditioned by, such features. If one opts for emergence,

it is incumbent upon one to at least sketch the means by which new features emerge. If one opts for panpsychism (thus broadly construed for now) then one must account for the all too obviously apparent total *lack* of certain features at the fundamental level. For example, Anaxagoras (c. 500-425 BC) flatly denied that emergence was possible and instead advanced the view that "everything is in everything". Anaxagoras explained the obvious contrary appearance by a "principle of dominance and latency" (see Mourelatos 1986) which asserted that some qualities were dominant in their contribution to the behavior and appearance of things. However, Anaxagoras's views on mind are complex since he apparently regarded mind as uniquely *not* containing any measure of other things and thus not fully complying with his mixing principles. Perhaps this can be interpreted as the assertion that mind is ontologically fundamental in a special way; Anaxagoras did seem to believe that everything has some portion of mind in it while refraining from the assertion that everything has a mind (even this is controversial, see Barnes 1982, p. 405 ff.).

On the other hand, Empedocles, an almost exact contemporary of Anaxagoras, favored an emergentist account based upon the famous doctrine of the four elements: earth, air, fire and water. All qualities were to be explicated in terms of ratios of these elements. The overall distribution of the elements, which were themselves eternal and unchangeable, was controlled by "love and strife", whose operations are curiously reminiscent of some doctrines of modern thermodynamics, in a grand cyclically dynamic universe.[5] The purest form of emergentism was propounded by the famed atomist Democritus (c. 460-370 BC). His principle of emergence was based upon the possibility of multi-shaped atoms "interlocking" to form an infinity of more complex shapes. But Democritus, in a way echoing Anaxagoras, had to admit that the *qualities* of experience (what we nowadays called "qualia") could not be accounted for in this way and chose, ultimately unsatisfactorily, to relegate them to non-existence: "sweet exists by convention, bitter by convention, in truth only atoms and the void". Although Democritus provides a remarkable anticipation of the modern doctrine of eliminativist materialism, we sorely miss his account of how *conventions* themselves -- the consciously agreed upon means of common reference -- emerge from the dancing atoms. Thus the core difficulty of the problem of consciousness remains unresolved.

What is striking about these early struggles about the proper "form" of a scientific understanding of the world, is that the mind and particularly consciousness keep rising as special problems. It is sometimes said that the mind-body problem is not an ancient philosophical worry (see Matson 1966), but it does seem that the problem of consciousness was vexing philosophers 2500 years ago, and in a form redolent of contemporary worries. Also critically important is the way that the problem of consciousness, and its origin, inescapably arises within the context of developing an integrated scientific view of the world.

3. Modern history of panpsychism

It is this that explains the lack of interest in panpsychism, emergentism etc. that sets in after the Presocratics and lasts until the scientific revolution of the seventeenth century with its renewed interest in comprehensive naturalistic accounts of the world. The modern "mechanistic" picture of the world inaugurated by Galileo, Descartes and Newton put the problem of the mind at center stage while paradoxically sweeping it under the rug.[6] The whole problem-space was severely distorted by what was

virtually a stipulated separation of matter from mind, so that what could have been merely a useful conceptual distinction was transformed into an ontological gulf. Thus, everything that could not be accounted for in terms of the interactions of simple material components was conveniently labelled a "secondary quality" inhabiting not the "real" world but merely the conscious mind. For instance, in a maneuver reminiscent of Democritus, colors were banished from the world of matter, replaced with the "causal powers" of physical things to produce "in the mind" the experience we call color. Thus the world was made safe for physics.

But the problem of the relation of the physical world to conscious minds was unavoidable and became ever more pressing. As Newton himself drolly pointed out in a letter to Henry Oldenburg: " ... to determine by what modes or actions light produceth in our minds the phantasm of colour is not so easie." One option was simply to give up - remove the mind from the expanding scientific picture of the world, and such was the motivation for René Descartes's infamous dualism of mind and body. But this leaves us with an untidy, perhaps incoherent, and certainly disintegrated view of the world. Another approach was to question the underlying definitional move of the scientific revolution, which was to *stipulate* that science was to study a "purely physical" world, voided of mentality by fiat. For one can wonder whether there is such a world. This question exacts its own price, however, which is our familiar dilemma, to which many thinkers responded with an endorsement of panpsychism.

Baruch Spinoza (1632-77) and Gottfried Wilhelm Leibniz (1646-1716) provide examples of two distinct and formatively important versions of panpsychism. Spinoza regarded both mind and matter as simply aspects (or attributes) of the eternal, infinite and unique *substance* he identified with God Himself. In the illustrative scholium to proposition seven of book two of the *Ethics* (1677/1985) Spinoza writes: "a circle existing in nature and the idea of the existing circle, which is also in God, are one and the same thing ... therefore, whether we conceive nature under the attribute of Extension, or under the attribute of Thought ... we shall find one and the same order, *or* one and the same connection of causes ...". We might say that, for Spinoza, physical science is a way of studying the psychology of God. There is nothing in nature that does not have a mental aspect -- the proper appreciation of matter itself reveals it to be the other side of a mentalistic coin.

Leibniz's view is sometimes caricatured as: Spinoza with infinitely many rather than just one substance. These substances Leibniz called *monads* (see Leibniz 1714/1989). Since they are true substances, and hence can exist independently of any other thing, and since they are absolutely *simple*, they cannot interact with each other in any way (nonetheless they are created by God, who is one of them -- here Spinoza seems rather more consistent than Leibniz). Yet each monad carries within it complete information about the entire universe. What we call space and time are in reality sets of relations amongst these monads (or, better, the information which they contain) which are in themselves radically non-spatial and perhaps even non-temporal (Leibniz's vision of space and time emerging from some more elementary systems of relations has always been tempting, if hard to fathom, and now fuels some of the most advanced physics on the planet).

Leibniz's monads are fundamentally to be conceived mentalistically -- they are in a way mentalistic automatons moving from one perceptual state (some conscious and some not) to another, all exactly

according to a God imposed pre-defined rule. It is highly significant for the development of later forms of panpsychism that Leibniz could find no intrinsic nature for his basic elements other than a mentalistic nature -- the only model he found adequate to describe his monads was one of perception and spontaneous activity. The physical world is, so to speak, an aspect of these perceptual states (so Leibniz's panpsychism is one that favors the mental realm, that is, it is at bottom a kind of idealism as opposed to Spinoza's "neutral monism"). What is of special interest is that unlike Spinoza, Leibniz can maintain a distinction between things that have minds or mental attributes from those that do not, despite his panpsychism. This crucial distinction hinges on the difference between a "mere aggregate" and what Leibniz sometimes calls an "organic unity" or an *organism*. Each monad represents the world -- in all its infinite detail -- from a unique point of view (both literally in the sense of having a perceptual perspective but also in terms of clarity and "significance"). Consider a heap of sand. It corresponds to a set of monads, but there is no monad which represents anything like a "point of view" of the heap. By contrast, your body also corresponds to a set of monads but one of these monads -- the so called *dominant* monad -- represents the point of view of the system which is your living body. (There presumably are also sub-unities within you, corresponding to organized and functionally unified physiological, and hence also psychological, sub-systems.) Organisms correspond to a hierarchically ordered set of monads, mere aggregates do not. This means that there is no mental aspect to heaps of sand as such, even though at the most fundamental level mind pervades the universe. In fact, for Leibniz minds are only rarely associated with physical systems and he explicitly denied that the world-system had a corresponding monad. In sharp contrast with Spinoza's views, Leibniz's universe is a mere aggregate. One last point: you might wonder why you, a monad that represents every detail of the entire universe, seem so relatively ignorant. The answer depends upon another important aspect of the conception of mentality. Leibniz allows that there are unconscious mental states. In fact, almost all mental states are unconscious and low-level monads never aspire to consciousness (or what Leibniz calls *apperception*). You are aware, of course, only of your conscious mental states and these represent a literally infinitesimal fraction of the life of your mind, the most of which is composed of consciously imperceptible *petite perceptions* (it is galling to think that the answers to such questions as whether there are advanced civilizations in the Andromeda galaxy lie hidden within each of our minds, but there it is).

The growth of idealist philosophy through the eighteenth and nineteenth centuries meant that panpsychism became, in effect, the default philosophy, but of course with a decided bias resulting from the positioning of mentality as the primary component of reality. However, this had little effect upon science (though it may well have contributed to the growth of positivism and "radical empiricism" in the *philosophy* of science) which throughout this period continued to rapidly expand in every direction, laying the groundwork for the overthrow of a philosophy-based mentalistic metaphysics with the physics-based scientific/materialistic metaphysical structure within which we still reside today.

The nineteenth century was the heyday of panpsychism. Even a partial list of panpsychists of that period reveals how many of the best minds of the time gravitated towards this doctrine. Prominent exponents of distinctive forms of panpsychism include Gustav Fechner, one of the founders of scientific psychology, Wilhelm Wundt (1832-1920), another famous early psychologist who established the first psychological research laboratory, Rudolf Hermann Lotze (1817-1881), a polymath who also figured in the creation of psychology as an empirical science, William James (1842-1910), the brilliant American philosopher and

psychologist who co-founded the philosophy of pragmatism, Josiah Royce (1855-1916), famed teacher and defender of a monistic idealism (in this respect, Royce had a philosophical role in America similar to that of F. H. Bradley in Britain), and William Clifford (1845-1879), a tragically short lived mathematical and philosophical genius whose work on of the nature of space and time prefigured Einstein's general relativity.

Other notable panpsychist thinkers of this period include Arthur Schopenhauer (1788-1860), who held the curious dual doctrine that everything is conscious, but that *not* everything is alive, Friedrich Paulsen (1846-1908), a student of Fechner's who extended his teacher's version of panpsychism, Morton Prince (1854-1929), psychologist and physician who advocated a panpsychism which emphasized that it is matter that must be "psychologized" or imbued with mentalistic attributes (Prince regarded this as a form of materialism and there are affinities here with some recent views of Galen Strawson as we shall see below). Also to be mentioned are Eduard von Hartmann (1842-1906) who extended his famous doctrine of the unconscious down to the level of atoms, Ferdinand C. S. Schiller (1864-1937) who provided a pragmatist defense of panpsychism as a doctrine which by various analogical arguments yields otherwise unattainable insights into nature and Ernst Häckel (1834-1919), an early and avid proponent of Darwinism in Germany who Clifford credits with the evolutionary continuity argument for panpsychism (for which see below) and Häckel was certainly willing to ascribe mental properties to living cells.

Royce and Lotze represent what may be called "idealist panpsychism". That is, the primary motivation for the ascription of mental attributes to matter is that matter is, in essence, a "form" of mind and thus panpsychism is a kind of *theorem* which follows from this more fundamental philosophical view. Royce believed that reality was a "world self", a conscious being that comprised absolutely everything and of which we, as well as everything else of course, were but parts. But Royce's panpsychism was of the synechological variety. Although every thing participates in the conscious life of the world self, not every "object" which one might nominate within the world of experience need itself be conscious, for these things are but thoughts of the world self and do not necessarily correspond to a being (or a sub-being) with its own mental life or consciousness. Yet some aspects of the world self do have a conscious life of their own. We are obvious examples, but Royce also believed that the range of such conscious beings went far beyond what we normally allow. Planets, stars and galaxies and even *species* are themselves conscious beings.[7] To the complaint that such things exhibit no sign of conscious life or thought, Royce had an interesting reply that raises intriguing philosophical issues. The reply was that the time scale of a conscious mind could vary tremendously -- the scale of the processes of consciousness in a galaxy are billions of time slower than the scale of human conscious processes (and mayhap the consciousness of subatomic particles, if they be conscious at all, runs billions of times faster).

Fechner, Wundt and perhaps James are "parallelist panpsychists". Their metaphysics endorses a thorough going, Spinozistic, parallelism between mind and matter, so that every physical entity has mental attributes, and vice versa. In fact, it was this metaphysical parallelism that suggested to Fechner the idea that there should be a lawful relation between the mental and the physical, which led to the birth of psycho-physics and the discovery of his famous law relating the strength of a sensation, S, to the strength of the physical stimulus, P: $S = k \log(P)$ (still one of the very few psycho-physical laws with any claim to validity). This is an interesting, if minor, illustration of the general point that scientific advance is often

contingent upon more or less explicit background metaphysical views. Although a clear scientific thinker Fechner was given to some rather mystical flights of fancy in defense of what he called the "day-view" (that is, the vibrant, open, panpsychist understanding of the world) as opposed to the dark and dead "night-view" of materialism. In the rather curiously entitled *The Little Book of Life After Death*, which was highly popular, boasted an introduction by no less than William James, and was last reprinted in 1977, Fechner asserts that "the plant thinks it is in its place ... to play with beetles and bees". Charles Hartshorne, a panpsychist himself, drily remarks of Fechner's ascription of consciousness to plants: "whatever can be said for this view must, it seems, have been said by Fechner" (1950, p. 447).

As mentioned above, Fechner's panpsychism is usually regarded to be of the synechological variety which withholds mental attributes from some of the simple constituents of larger, enminded systems (up to and including the world-soul itself). This view of Fechner stems from the extreme reliance Fechner placed upon analogical arguments for the existence of mental qualities, and which he regarded as the sole ground for the attribution of mind to anything other than oneself. Thus plants are like animals which are *asleep*, but not thereby mindless nor even unconscious, insofar as they possess no less than animals a complex set of teleological mechanisms serving their perseverance. Fechner was much taken to task for his failure to endorse a thorough going panpsychism by, among others, Lotze, who wrote (referring to Fechner's (1848) *Nanna, or On the Mental Life of the Plants*) "one cannot search for the mind arbitrarily in the plants, the darlings of our fantasy, and remain satisfied with the existence of dead matter in the rocks" (Lotze 1852, p. 133). This is worth mentioning since there is evidence that Fechner, in *Zend-Avista* (first published in 1851), did take the seemingly logical step of extending his panpsychism to all of nature, in line with the dual aspect metaphysics which he officially advocated (see Woodward 1972, from which the above translation).

William James's panpsychism grew out of his "neutral monism" -- the view that reality is *neither* mental nor physical but has a distinct, and seemingly intrinsically mysterious, basic character which can be regarded as either mental or physical from certain viewpoints. To the extent that a neutral monism can be regarded as a dual-aspect view (as in Spinoza's philosophy), it might be regarded as a kind of panpsychism in its own right, but James's view developed beyond this, to incorporate mind like elements into the basic structure of reality. In a notebook of 1909 he wrote: "the constitution of reality which I am making for is of the psychic type" (see Cooper 1990). James's commitment to panpsychism remains somewhat controversial, since he also advanced a cogent set of objections against a version of the view, which he somewhat derisively labelled the "mind dust" theory, in chapter six of *The Principles of Psychology* (1890/1950). But in the end his commitment is quite clear (see James (1911), and for an excellent analysis of James's views on mind see Cooper (1990)).

The most significant development and defense of a panpsychist philosophy in the twentieth century was undoubtably that of Alfred North Whitehead (1861-1847). Exploration of the details of Whitehead's philosophy would require an article of its own, and would be fraught with interpretive difficulties in any case since Whitehead's own presentation is forbiddingly complex, full of idiosyncratic technical terms and sometimes of dubious intelligibility. But roughly speaking Whitehead proposed a radical reform of our conception of the fundamental nature of the world, placing *events* (or items that are more event-like than thing-like) and the ongoing *processes* of their creation and extinction as the core feature of the

world, rather than the traditional triad of matter, space and time. His panpsychism arises from the idea that the elementary events that make up the world (which he called *occasions*) partake of mentality in some -- often extremely attenuated -- sense, metaphorically expressed in terms of the mentalistic notions of creativity, spontaneity and perception. The echoes of Leibniz are not accidental here, and Whitehead also has a form of Leibniz's distinction between *unities* and mere *aggregates*, which he explains in these terms: "... in bodies that are obviously living, a coordination has been achieved that raises into prominence some functions inherent in the ultimate occasions. For lifeless matter these functionings thwart each other, and average out so as to produce a negligible total effect. In the case of living bodies the coordination intervenes, and the average effect of these intimate functionings has to be taken into account" (1933, p. 207; lest it seem that Whitehead is only discussing *life*, he is clear that this depends upon a sort of mental functioning). Unavoidably, if perhaps unfortunately, Whitehead's panpsychism stands or falls with his entire metaphysical system which entails a more radical revision of our current scientifically based picture of the world than even panpsychism necessitates. In very general terms, Whitehead's panpsychism faces the same objections as any other version, and stems from the same basic anti-emergentist intuition (for a clear introduction to, and defense of, Whitehead's panpsychism see Griffin 1998; another interpretation, and pantheistic reworking, can be found in the writings of Charles Hartshorne (1897-2000), for example, in Hartshorne 1972).

With his emphasis on the vitality and spontaneity of nature, Whitehead represents a culmination of nineteenth century panpsychist thinking, and probably not coincidentally its presentation was pretty much simultaneous with the culminating development of a robust and serious emergentism (as worked out by, for example, C. Lloyd Morgan (1852-1936) and C. D. Broad (1887-1971)). It may have seemed that, for a moment, the ground was prepared for another great battle between the two basic conflicting ideas about mind's place in the natural world. But history moved in another direction. Big science took center stage, and metaphysics became a bit player in a new kind of philosophical drama. The kind of radical emergentism espoused by thinkers such as Broad was doomed by the huge technological advances and theoretical successes of physical science, in particular quantum mechanics' victory in explaining how chemical complexity arises from purely physical principles, along with the rise of a logical positivist philosophy that derided any philosophical idea that was not cleanly rooted in empirical science. But all this also had the predictable effect of relegating panpsychism, which also required a philosophical extension of scientific belief, to the limbo of unwarranted philosophical intercession into domains beyond its expertise.

Thus for some fifty years after the 1929 publication of Whitehead's panpsychist *Process and Reality* and the 1925 publication of C. D. Broad's emergentist *Mind and Its Place in Nature* there was relatively little interest in either doctrine. There are few explicit defenders of panpsychism at the present time. The two most prominent are Timothy Sprigge and David Griffin, who also represent two of the main versions of panpsychism. Sprigge, in *A Vindication of Absolute Idealism* (1983), defends an idealist based panpsychism somewhat akin to that of Royce, while Griffin, in *Unsnarling the World Knot* (1998), espouses an atomistic panpsychism in the form of an explicit interpretation, extension and defense of Whitehead's version of the doctrine. Although not providing full scale defences of panpsychism, several other writers have recently approached the problem of consciousness in ways sympathetic to panpsychism. See for example chapter eight of Chalmers (1996), or the articles by Piet Hut and Roger

Shepard, Gregg Rosenberg, and William Seager, all in Shear (1997).

The current burst of scientific and philosophical studies of mind sparked by the "cognitive revolution" has rekindled debate about the perennial dilemma of emergentism versus panpsychism. The recently renewed and once again influential claim of some philosophers, especially David Chalmers, that the explanation of consciousness presents a uniquely difficult problem for science has forced the reexamination of the metaphysical foundations of the scientific world view (see *The Conscious Mind* 1996). Chalmers calls this problem the "hard problem of consciousness"; it is also sometimes called the "explanatory gap" or the "generation problem". The key difficulty is how to explain in naturalistic terms the generation of consciousness by "mere matter". Once again it seems imperative to decide whether and how mind *emerges* upon, or exists only under, some specifiable and non-universal natural and non-mentalistic conditions or whether mind itself forms a part of the fundamental structure of the world, perhaps in some of the ways panpsychists have suggested.

4. Arguments For Panpsychism

In an excellent article on panpsychism and its history, Paul Edwards (1967) divided the arguments for panpsychism into two broad categories: genetic and analogical. The division is incomplete but makes a fine start. Genetic arguments assert that the best account of the genesis of mind lies in panpsychism; the analogical arguments seek to find analogies between clearly enminded entities and the rest of nature which are strong enough to warrant the extension of mental attributes throughout nature.

4.1 Genetic Arguments

There exist both *a priori* and empirical genetic arguments. The claim that emergence is strictly impossible has a metaphysical root in the ancient dictum "ex nihlio, nihil fit" to which Wundt, for example, explicitly appealed (see Wundt 1892/1894, p. 443). A much more recent version of this argument can be found in Nagel's article "Panpsychism" (1979). Nagel explicitly links panpsychism to a necessary failure of emergentism, namely that emergentism cannot rise to the status of a *metaphysical* relation. Nagel says: "there are no truly emergent properties of complex systems. All properties of complex systems that are not relations between it and something else derive from the properties of its constituents and their effects on each other when so combined" (p. 182). Thus the only coherent form of emergentism is an epistemological doctrine about the limits of our understanding of the behavior of complex systems. The link to panpsychism appears with Nagel's denial of reductionism, which precludes simply identifying mental properties with complex physical properties. Then, since, as Nagel says, we can build an enminded system out of "any matter", mind must be associated with matter in general and in its most fundamental forms (whatever these may be as eventually revealed by physics).[8] The argument appears to suffer from the lack of a clear proof that a more radical form of emergentism than the epistemological variety countenanced by Nagel is impossible. Although there are philosophical questions about the coherence of such a radical emergentism, exactly such a doctrine was developed in some detail by the likes of Morgan and Broad (see above). So this is a serious defect in Nagel's argument.

Nonetheless, the epistemological form of emergentism is highly congenial to common interpretations of complexity in modern science and is usually what is meant in modern discussions of emergence. Thus the anti-emergence argument can retain some force within that context, if now in an empirical form.

The empirically based forms of the genetic argument have been traditionally more popular. Wundt himself makes an "inference to the best explanation" in defense of panpsychism. He states that panpsychism is "a theory, it is true; but it is the only theory which can explain the phenomena of movement displayed by these primitive creatures" (1892/1894, p. 443). Wundt found it literally incredible that the apparent purposiveness and appropriateness of the behavior of even simply micro-organisms -- which he thought lent themselves naturally to mentalistic explanation -- could spring, suddenly and arbitrarily, into existence through the mere conglomeration, via elementary physical forces, of material particles into complex systems.[9]

But by far the most popular empirical ground for the genetic argument stems from Darwinism, whose ascension in the mid-nineteenth century transformed debate about life and mind. This form of the genetic argument turns on the assumption that evolution is a continuous process that moulds pre-existing properties into more complex forms but which can not produce "entirely novel" properties. An important proponent of this argument was William Clifford. Clifford puts the argument thus: "... we cannot suppose that so enormous a jump from one creature to another should have occurred at any point in the process of evolution as the introduction of a fact entirely different and absolutely separate from the physical fact. It is impossible for anybody to point out the particular place in the line of descent where that event can be supposed to have taken place. The only thing that we can come to, if we accept the doctrine of evolution at all, is that even in the very lowest organism, even in the Amoeba which swims about in our own blood, there is something or other, inconceivably simple to us, which is of the same nature with our own consciousness ..." (1874/1886, p. 266). Another extremely influential figure whose panpsychism rests in part on this idea is William James, who writes that "we ought ... to try every possible mode of conceiving of consciousness so that it may *not* appear equivalent to the irruption into the universe of a new nature non-existent to then" (1890/1950, p. 148).The argument has drawn supporters throughout the twentieth century (see for example Drake (1925), Wright (1953), Waddington (1961) and of course Nagel (1979).

It is difficult to assess this argument, since, for example, the existence of such an obvious example as *wings* seems on the face of it to present a perfectly clear case of the evolutionary development of novel features (as compared, say, to the aviational equipment of the distant single celled ancestors of the birds). It would then be claimed on the other side that wings are nothing but a more complex configuration of matter itself, the possibilities of which configurations are implicit in the pure physics of the DNA based phylogeny of all living things. Wings, and all other materially embodied biological organs, seem clearly to fall under the kind of merely epistemological emergence discussed above. Mind, and especially consciousness, certainly does not seem to be merely a new kind of material organ nor a new kind of behavioral propensity, so there is indeed some cogency in this reply. The panpsychist position would clearly fail if there was a clear and uncontroversial conception of how consciousness emerges, in an ontological rather than epistemological sense, from entirely non-mentalistic physical features, but at present we simply do not possess such a conception, although many controversial suggestions are in play. It is the burden of the emergentist to provide one, or convince us to be content with the brute fact that

mental properties are conditioned by certain physically complex states in a fundamentally inexplicable way. Either task is decidedly non-trivial.

It is also worth noting as an historical point that the empirical version of the above "argument from continuity" was bolstered for some time in the late nineteenth and early twentieth centuries by laboratory research. For example, work of Hans Driesch (1867-1941), famous as one of the last serious defenders of vitalism, a doctrine which had very close connections to panpsychism, and R. Lotze, who was a determined foe of vitalism despite his advocacy of panpsychism (his brand of idealism left the "material world" explicable by mechanical laws) was taken to support panpsychist claims. Both men had taken to dividing up certain creatures, to discover that whole organisms could develop from the parts. Driesch's experiments on sea urchin embryos -- very demanding and cutting edge work at the time -- suggested that every cell of the developing urchin was capable of forming a new embryo. This was taken as evidence that there was some "principle of life" inherent within each cell. In less rigorous experiments, Lotze showed that parts of polyps could grow into complete, new, polyps. A kind of analogical extension suggested that mental properties might be similarly inherent in the basic structures of the world. However, as mysterious and suggestive as such findings might have been in their time, they would seem now to be entirely explicable, albeit only in principle, by modern reductionist DNA based biology; any analogical support they might have offered to panpsychism has thus entirely evaporated. But perhaps other, more direct, analogical arguments might fare better.

4.2 Analogical Arguments

The most straightforward argument from analogy goes like this: if we look closely, with an open mind, we see that even the simplest forms of matter actually exhibit behavior which is akin to that we associate with mentality in animals and human beings. Unfortunately, in general, this seems quite preposterous, and some panpsychists have written some pretty silly things in its defense. For example, Ferdinand Schiller attempted to "explain" catalysis in terms of mentalistic relations: "is not this [that is, catalysis of a reaction between A and B by the catalyst C] strangely suggestive of the idea that A and B did not know each other until they were introduced by C, and then liked each other so well that C was left out in the cold" (as quoted by Edwards (1967) in an acidly humorous paragraph, from Schiller (1907)). Strange? Certainly, but not really very suggestive at all compared to the physical chemists' intricately worked out, mathematical and empirically testable tale of energy reducing reaction pathways. There has always been a strain of mysticism in many panpsychists, who like to imagine they can "sense" that the world is alive and thinking, or find that panpsychism provides a more "satisfying" picture of the world, liberating them from the arid barrenness of materialism and perhaps this leads them to seek analogies somewhat too assiduously (as noted above, Fechner was the most poetical advocate of the mystical appeal of panpsychism and also a fervent advocate of analogical arguments for panpsychism).

A more intriguing hope for an analogical defense of panpsychism springs from the overthrow of determinism in physics occasioned by the birth of quantum mechanics. There have been occasional attempts by some modern panpsychists, starting with Whitehead, to see this indeterminacy as an expression not of blind chance but spontaneous *freedom* in response to a kind of *informational*

inclination rather than mechanical causation. This updated version of the analogy argument has the advantage that the property at issue, freedom, modelled as spontaneity and grounded in indeterminacy, can be found at the most fundamental level of the physical world. As in any analogical argument, the crux of the issue is whether the phenomena cited on the one side are *sufficiently* analogous to the target phenomena to warrant the conclusion that the attributes in question can be extended from the one domain to the other. In this case, we have to ask whether the indeterminacy found at the micro-level genuinely corresponds to what we take freedom to be, and this is doubtful. The indeterminacy of modern physics seems to be a pure randomness quite remote from deliberation, decision and indecision.

But still another analogical argument which draws upon quantum physics is much more promising. The analogy in this case involves the relation between consciousness and information. It is natural to think that among the functions of consciousness is the integration of diverse fields of information and the monitoring of various external and internal states. The consciousness of pain, for example, at least involves the monitoring and processing of information about significant states of the body.[10] In a recent work on consciousness which emphasizes the informational and monitoring functions of consciousness, William Lycan comes surprisingly close to a form of panpsychism when he states that "one little monitor does make for a little bit of consciousness. More monitors and better integration and control make for more and fuller consciousness" (1996, p. 40). This is only intended by Lycan to be part of an account of how consciousness *emerges* which is then forced to allow that consciousness is rather more ubiquitous than untutored intuition might expect. But it follows from this view that *if* information monitoring is a fundamental and pervasive feature of the world at even the most basic levels, then consciousness too should appear at those levels.

It is then highly suggestive that one of the central features of quantum mechanics is the existence of informational but non-causal relations between elements of systems. These relations are non-causal insofar as they are modulated instantaneously over any distance and do not involve the transfer of energy between the parts of the system. But they are informational in the sense that the changes of state of one part of the system seems in some way to be communicated to the other. There is no doubt whatsoever that such quantum systems can exist (they have been created in the laboratory) although the interpretation of them in terms of information exchange is contentious. For example, it is possible to create pairs of photons with correlated polarization states, such that, while neither photon is in a definite state of polarization prior to measurement, they must be discovered to be in opposite polarization states when a measurement takes place, no matter how far apart they are when the measurements occur. Such correlated particles are said to be "entangled". It does not seem unreasonable to regard two such entangled photons as effectively monitoring each other's state of polarization. We can then use a theory of consciousness such as Lycan's to argue that a little monitoring makes for a little bit of consciousness. Furthermore, while entangled states are normally very delicate and susceptible to "decoherence" caused by environmental disturbance, there might be certain systems that can resist decoherence and it has been conjectured that these systems are the physical foundation of more complex states of consciousness (see Hameroff and Penrose 1996; Hameroff, at least, is willing to entertain a panpsychist interpretation of this work). To follow this line of thought even further, the decoherence argument evidently collapses for the universe as a whole, which by definition cannot be disturbed by any outside force, so presumably the total universe is in one immensely complex entangled state. Given a link between consciousness,

monitoring and information exchange, this leads to a view highly reminiscent of Leibniz's monadology, with centres of (perhaps rudimentary) consciousness, or at least mind, at the foundation of the world. Michael Lockwood has developed a highly interesting and well worked out version of this panpsychist view combining quantum mechanical considerations with the intrinsic nature argument, to be considered below, which endorses "a conception of the world as ... a sum of perspectives" (1991, p. 177).

4.3 Intrinsic Nature Arguments

Another possible argument for panpsychism is neither genetic nor analogical but instead depends on the idea that every actual thing, or kind of thing, must have an intrinsic nature. The objects studied by physics, it is claimed, are described in purely dispositional terms. That is, while an electron, for example, is said to possess "spin", all this amounts to is that the electron has certain dispositions to behave in certain ways under certain circumstances. It is arguable that dispositions must be grounded in some intrinsic, non-dispositional attributes, but we have no conception whatsoever of what the intrinsic nature of matter might be. In fact, the only intrinsic nature with which we are familiar is consciousness itself. The qualities of conscious experience (to take simply sensory experience: the smell of a rose, the taste of a strawberry, etc.) seem not to be reducible to relations amongst non-experiential states nor entirely specifiable without remainder in terms of their causal powers to produce behavior (and other mental states). They seem instead to possess (or be) intrinsic and irreducible characteristics. If this is the only idea of intrinsic nature we possess, and matter must be assigned some intrinsic nature, it seems that matter must be granted a mentalistic intrinsic nature. The core idea of this argument can be traced back to Leibniz who felt forced to ascribe mentalistic attributes to his monads as the only possible feature which could make intelligible the active forces that seemed to be required in an adequate physics, and which finally laid to rest the dream of a purely mechanical world view. In his discussion of this difficulty, Whitehead describes all "modern cosmologies" as having to admit a "mysterious reality in the background, intrinsically unknowable" (1933/1967, p. 133) and notes that Leibniz "explained what it must be like to be an atom" (1933/1967, p. 132). See Sprigge (1983) for a defense of this argument within an extended discussion of the virtues of panpsychism (for another brief summary of the argument see Sprigge 1999). Another, less idealist, version of the argument is developed in Lockwood (1991), based upon ideas taken from Russell's later philosophy, married to an interpretation of quantum physics. Although far from demonstrative this is, in the words of Thomas Sprigge (1999), "a hypothesis worth exploring as the only alternative to saying that matter is unknowable in its inner essence, and as likely also to cast light on the mind-body or mind-brain relationship."

Still, one obvious reply to this argument is to bite the bullet of unknowability and accept that the intrinsic nature of matter is either unknown or even essentially unknowable. Belief in such irremediable ignorance would seem neither to entail panpsychism nor to be incoherent, and many might prefer it to panpsychism.

However, recently several philosophers have made remarks somewhat reminiscent of this argument. For example, Galen Strawson has argued for a revised conception of materialism and remarks that "the experiential considered specifically as such -- the portion of reality we have to do with when we consider experiences specifically and solely in respect of the experiential character they have for those who have

them as they have them- that "just is" physical" (1997/1999, p. 7) . Strawson hints that only a "revolutionary development" in physics would allow consciousness to be "discerned and described" by that science. The idea that a revolutionary change in physics may be necessitated by the problem of consciousness is endorsed, suggested or at least hinted at by several distinguished thinkers, including Roger Penrose (1989), John Searle (1991, pp. 123-4), Thomas Nagel (1979, 1986, 1999) and Noam Chomsky (1999; see the remarks about unification and revision on p. 82 for example). Suggestive as these thoughts may be, it only leaves a gap into which the wedge of panpsychism might be inserted. What reason have we to suppose that the hoped for revolution in our understanding of matter at the most fundamental level will involve ascribing essentially *mentalistic* properties to it? The panpsychist's hope lies in the thought that any modification of our conception of the physical that does not incorporate mind will leave us in an essentially unchanged position, with no explanation of how consciousness emerges from the radically non-mental physical elements of the world. We have seen that this argument has been bruited since at least the time of the Presocratics and it has often led emergentists to reconsider their position when the problem of consciousness is directly considered (it is this worry that probably explains why Morgan, a radical emergentist, retreated into a Spinozistic parallelism of mind and matter; see Morgan 1923, p. 32).

This leads to the final consideration in favor of panpsychism to be considered here, which is a sort of methodological argument. Panpsychism enjoys a metaphysical advantage in that it avoids the difficulties of emergentism, which are greater than is generally thought. Not only is there a problem simply in accounting for the emergence of something so distinctive as consciousness from mere matter, it is surprisingly difficult to articulate a form of emergentism that does not threaten to make the emergent features causally impotent or epiphenomenal. This is not the place to discuss the difficulties of all the varieties of emergentism, but they seem serious.

5. Arguments Against Panpsychism

This article would be incomplete without a consideration of some of the objections against panpsychism, but it will also serve to sharpen our understanding of the doctrine to consider possible replies available to the panpsychist.

Perhaps the initially most obvious problem with panpsychism is simply the apparent lack of evidence that the fundamental entities of the physical world possess any mentalistic characteristics. Protons, electrons, photons (to say nothing of rocks, planets, bridges etc.) exhibit nothing justifying the ascription of psychological attributes and thus Occam's razor, if nothing else, encourages withholding any such ascriptions. Furthermore, it is argued, since we now have scientific explanations (or modes of explanation at least) which have no need to ascribe mental properties very widely (it is tempting to interject: not even to *people*!) panpsychism can be seen as merely a vestige of primitive pre-scientific beliefs. At one time, perhaps, panpsychism or animism may have been the conclusions of successful inferences to the best explanation, but that time has long passed.

As we examine ever smaller, more basic units of the physical world, it seems harder and harder even to

imagine that such things have *any* properties that go beyond those ascribed to them by the physical theories which are, after all, the only reason we have to believe in them. In any case, there seems no reason to assign any intrinsic nature to the theoretically postulated entities of physics that goes beyond providing for the causal powers they are presumed to possess according to the theories which posit them. Even granting the need to assign *some* intrinsic nature to matter, it remains far from clear that *mentality* is the intrinsic character required for possession of these causal powers. Some such argument likely accounts for the general sense of implausibility with which many people greet panpsychism nowadays. For example, perhaps a somewhat extreme one, John Searle describes panpsychism as an "absurd view" (Colin McGinn uses the word "ludicrous") and argues that thermostats do not have "enough structure even to be a remote candidate for consciousness" (1997, p. 48).

Another possible, and closely related, diagnosis for the sense of implausibility which panpsychism engenders in many may stem from a certain methodological ideal. The job of philosophers, it may well be thought, is to show how the mind, and especially consciousness, is to be integrated into the scientific world view, or, to use the current term, *naturalized*. It is against the implicit rules of this game to demand that science be changed to accommodate consciousness -- the point is to take science as it is and show that consciousness can be fitted into *that* kind of conceptual structure. This assumes of course that science is, as it stands, or near enough, already true and complete. Thus it is curious to find Searle both advocating that consciousness is a "biological property" whose conditions of emergence are no stranger than those of the liquidity of water, and also hinting that a revolution in our understanding of the physical world will be needed to accommodate consciousness.

Such remarks as Searle's betray emergentist presuppositions as well as assumptions about the nature of consciousness. After all, on the view of Lycan canvassed above it would be difficult to withhold attributing a "little bit" of consciousness to thermostats. In any case, this "no evidence" argument can be weakened by noting that we should not necessarily expect to see signs of complex mentality at the simplest level or perhaps any sign at all. After all, the effects of gravitation are invisible at the level of extremely small sizes and masses but this does not mean that gravitation is insignificant in the universe, nor that it is not a ubiquitous and fundamental feature of the world, of which every existing thing partakes.

A problem very closely related to this difficulty of lack of evidence is this: even if there was a need to revolutionize fundamental physics in order to give an account of consciousness, why would the new features of the transformed physics be *mental* features? One should not ascribe anything more to these new features than what is necessary for them to solve whatever problem in *physics* that prompted their postulation.

Leaving aside how the argument from intrinsic nature impinges on this point, what is crucial here is just how these hypothetical revolutionary features would operate. If they involved basic operations on *informational* states and, for example, the cross monitoring of and by fundamental physical entities (as discussed above) then there might well be some reason to connect them with mentality. It seems that physics *already* has posited something like these informational operations and with them something at least somewhat analogous to aspects of the psychological domain.

This reply, so far as it goes, can also serve to deflect another objection, which is that the mental attributes assigned to the fundamental physical entities by the panpsychist must lack all causal efficacy, that is be entirely epiphenomenal, since the physical world, as described by physics, is *causally closed*. (Something like this argument is advanced by McGinn in his discussion of panpsychism in *The Mysterious Flame* (1999, pp. 95-101), though McGinn tendentiously ignores the distinction frequently drawn by panpsychist's and discussed above between "mere aggregates" and "unities".) That is, for every physical event there is a purely physical explanation for its occurrence and these explanations make no reference to mental properties. This argument suffers from an intentional fallacy. It is possible that some of the properties referred to in physical description of an event and its causes are identical to mental properties. The dispositional aspect of the properties of remote connectedness via informational states that we have been discussing are a part of basic physics but the panpsychist may urge that they also represent the primitive consciousness of the basic entities involved in these interactions. Physics has *described* them in the physical terms appropriate for physical theory, that is, purely in terms of their dispositions to interact with other physical entities in certain ways; this does not preclude their *being* mental properties. Of course, we need some independent argument that these properties ought to be regarded as *mental*, but that is provided, to the extent it can be, by the informational and mutual monitoring aspects of them.

There is another form of the argument from the causal closure of the physical world. One might expect that a fundamental feature as significant as consciousness appears to be should take some part in the world's causal commerce. But if it does play such a role, then we should expect it to turn up in our investigation of the physical world; we should expect, that is, to see *physically* indistinguishable systems at least occasionally diverge in their behavior because of the lurking causal powers of their mental dimension. The argument proceeds with the claim that it is very doubtful that there is any such evidence and, if it were supposed to exist, our physical picture of the world would then be radically *causally* incomplete in contradiction with the world's presumed causal closure at the physical level. The reply in the text serves against this objection as well. But it is also worth noting that, as a matter of fact, physically indistinguishable systems do behave differently from each other; no one knows if some "hidden variables" can be invoked to account for this indeterminism. Perhaps the hidden feature is in some way related to mentality and consciousness -- such is the core notion of panpsychism. If one regards the problem of free will as especially pressing one may be drawn to this second line of reply against the argument from causal closure (one panpsychist who takes this line is Griffin 1998).

We noted above that a common distinction within the field of psychological attributes is that between conscious and unconscious mental states, and several panpsychists have appealed to this distinction in setting forth their doctrines. But there is a danger here that threatens to undercut one of the prime virtues of panpsychism which is the avoidance of emergence. For if the mental attributes which the panpsychist ascribes to the fundamental or simplest entities in the world are all unconscious mental properties, then the question of how *conscious* mental states arise will be inescapable. And this in turn means that we now need a theory of the emergence of consciousness from the merely unconscious mental states licensed by this cautious panpsychism.[11] It is best to cut this line of objection off at the start. The panpsychist needs to bite the bullet here and *postulate* that it is indeed states of consciousness, although presumably with a very impoverished degree and kind of content, which are to be assigned to the most simple elements of

nature.

However, the postulate of primitive consciousness still leaves open a line of objection, call it the "combination problem," which was first raised by William James, who in the following passage argues that panpsychism will still face its own problem of emergence:

> Take a sentence of a dozen words, and take twelve men and tell to each one word. Then stand the men in a row or jam them in a bunch, and let each think of his word as intently as he will; nowhere will there be a consciousness of the whole sentence ... Where the elemental units are supposed to be feelings, the case is in no wise altered. Take a hundred of them, shuffle them and pack them as close together as you can (whatever that might mean); still each remains the same feeling it always was, shut in its own skin, windowless, ignorant of what the other feelings are and mean. There would be a hundred-and-first feeling there, if, when a group or series of such feeling were set up, a consciousness *belonging to the group as such* should emerge. And this 101st feeling would be a totally new fact; the 100 original feelings might, by a curious physical law, be a signal for its *creation*, when they came together; but they would have no substantial identity with it, nor it with them, and one could never deduce the one from the others, or (in any intelligible sense) say that they *evolved* it (1890/1950, p. 160, original emphasis).

This is a powerful objection since if panpsychism must allow for the emergence of states of consciousness then what prevents an emergence doctrine which avoids the implausible and indiscriminate broadcasting of mental characteristics throughout the world?

Note first that a form of panpsychism such as Leibniz's entirely escapes this objection. For Leibniz, minds are not formed out of combinations of parts (whether sub-minds or non-mental entities). Each mind is complete in itself, and in fact totally causally isolated from all other minds; there is no way that the combination problem could arise. However, the cost to Leibniz is the downgrading of the physical world to a kind of "consensual illusion"; matter, space and time are essentially constructs of mental phenomena.

If we wish to retain a robust conception of matter, which is extended but not mutilated by panpsychism, there seems little doubt that we will require a theory of emergence. But it does not follow, at least not directly, that a *non-mentalistic* emergentism is therefore to be favored. That depends upon the nature of the emergence in question. Whitehead, for example, embraced the need for a kind of emergence with no diminishment in his support for panpsychism. As Hartshorne explains, "it is the destiny of the many to enter into a novel unity, an additional reality" which means that Whitehead makes the "admission not merely of emergence, but of emergent or creative synthesis as the very principle of process and reality" (1972, p. 162).

It is clear from the way that James develops his version of the combination problem that he is presupposing a metaphysics of part-whole reductionism such that the properties of the whole are no more

than the sum or combined effect of the properties of the parts, in which the parts entirely retain their identities. For example, he says "... in the parallelogram of forces, the "forces" themselves do not combine into the diagonal resultant; a *body* is needed on which they may impinge, to exhibit their resultant effect" (1890/1950, p. 159). Such a view undoubtedly has a certain attractiveness; it seems no more than a reasonable generalization of the mereological reductionism of which the world provides so much evidence.

But we know that this view is inadequate. Quantum mechanics has made it abundantly clear that systems are not simply the sum of their parts in James's sense but can exhibit properties that go beyond those of the parts and which cannot be detected by examining the parts in isolation. It is impossible to tell if an electron, for example, possesses an entangled partner positron by looking only at the electron and the positron (they individually look identical to non-entangled particles). Yet the system of entangled particles exhibits properties quite distinct from the properties of pairs of non-entangled particles. Thus there *is* a mode of combination which goes far beyond what James allows and which we know is actually at work in the world. This mode of combination also seems to have some intimate connection with information and some sort of non-causal information exchange which, as noted above, has some affinity with psychological notions. (A more detailed examination of this argument can be found in Seager 1999, ch. 9.)

Nevertheless, it obviously remains far from clear that quantum mechanics necessarily leads to panpsychism and one might wish to deploy the powerful theory of emergence which quantum mechanics provides in the service of a more traditional emergentism which sees mind developing from non-mental aspects of nature (such an approach is taken by Silberstein and McGeever 1999). Assessing such a strategy would require consideration of the plausibility of the claim that mind and consciousness can be explicated solely in terms of the physical properties and entities postulated by quantum mechanics, a difficult task beyond the scope of this article, but about which one might harbor some doubts. The point here is simply that the combination problem can be addressed from within a panpsychist framework.

The existence of possible replies to a set of objections does not, in itself, provide positive grounds for endorsing a theory. At present, the predominance of the scientific view of the world, and a general disinclination towards dualistic as well as idealistic metaphysics, brings with it the triumph of emergentism, and the key issue becomes that of assessing the prospects of theories of emergent mentality. All modern physicalistic theories of mind implicitly rest upon a theory of emergence (which is seldom articulated in any detail), but, thus far, none of these has dealt with consciousness in a fully satisfactory way (that is, the problem of the emergence of consciousness has not gone the way of the problem of the emergence of chemistry). Unless and until we have such a satisfactory account, panpsychism remains an open possibility.

Panpsychism is an abstract metaphysical doctrine which as such has no direct bearing on any scientific work; there is no empirical test that could decisively confirm or refute panpsychism. One might complain about this remoteness, as Thomas Nagel does with the remark that panpsychism has "the faintly sickening odor of something put together in the metaphysical laboratory" (1986, p. 49). Nonetheless, metaphysical views form an indispensable background to all science. They integrate our world views and

allow us to situate our scientific endeavors within a larger vista and can suggest fruitful new lines of empirical enquiry (as the example of Fechner's psycho-physics illustrates). In particular, panpsychism accords with an approach that rejects physicalist reductionism at the same time as enjoining the search for neural correlates of consciousness, and it sees, or wants to see, a fundamental unity in the world which emergentism denies. Thus it is not a doctrine at odds with current empirical research.

It has always been and remains impossible to resist metaphysical speculation about the fundamental nature of the world. As long has there been science, science has informed this speculation and in return metaphysics has both helped to tell us what the point of science is and paved the way for new science. Panpsychism remains an active player in this endless speculative interchange.

Bibliography

- Binet, A. (1889). *Études de psychologie exprimentale: Le fétichisme dans l'amour, La vie psychique des micro-organismes, L'intensité des images mentales, etc*. Paris: O. Doin. The work was translated by Thomas McCormack as *The Psychic Life Microorganisms*, 1891, New York: Open Court.
- Broad, C. (1925). *The Mind and Its Place in Nature*, New York: Harcourt, Brace.
- Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.
- Chalmers, D. (1996). *The Conscious Mind*. Oxford: University of Oxford Press.
- Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*, Cambridge: Cambridge University Press.
- Clifford, W. (1874/1886). "Body and Mind", in *Fortnightly Review*, December. Reprinted in *Lectures and Essays*, Leslie Stephen and Frederick Pollock (eds.), London: Macmillan. (Page references are to the 1886 reprint.)
- Cooper, W. E. (1990). "William James's Theory of Mind", *Journal of the History of Philosophy*, 28, 4, pp. 571-93.
- DeSousa, R. (1989). "Kinds of Kinds: Individuality and Biological Species", in *International Studies in the Philosophy of Science*, 3, pp. 119-35.
- Drake, D. (1925). *Mind and Its Place in Nature*, New York: Macmillan.
- Dretske, F. (1995). *Naturalizing the Mind*, Cambridge, MA: MIT Press.
- Edwards, P. (1967). "Panpsychism", in *The Encyclopedia of Philosophy*, vol. 5, P. Edwards (ed.), New York: Macmillan.
- Fechner, G. (1848). *Nanna, oder, Über das Seelenleben der Pflanzen*, Leipzig : L. Voss.
- Fechner, G. (1906). *Zend-Avista: oder über die Dinge des Jenseits vom Standpunkt der Naturbetrachtung*, 3rd edition, Hamburg: L. Voss.
- Fechner, G. (1946). *The Religion of a Scientist*, (selections of Fechner's writing in English translation), W. Lowrie (ed. trans.), New York: Pantheon.
- Griffin, D. (1998). *Unsnarling the World Knot: Consciousness, Freedom and the Mind-Body Problem*, Berkeley: University of California Press.
- Hameroff, S. and Penrose, R. (1996). "Conscious Events as Orchestrated Spacetime Selections", in *The Journal of Consciousness Studies*, 3(1), pp. 36-53.

- Hartshorne, C. (1950) "Panpsychism", in *A History of Philosophical Systems*, V. Ferm (ed.), New York: Rider and Company, pp 442-453.
- Hartshorne, C. (1972). *Whitehead's Philosophy: Selected Essays 1935-1970*, Lincoln: University of Nebraska Press.
- Hull, D. (1976). "Are Species Really Individuals?", in *Systematic Zoology* 25, pp. 174-91.
- James, W. (1890/1950). *The Principles of Psychology*, v. 1, New York: Henry Holt and Co. Reprinted in 1950, New York: Dover.
- James, W. (1911) "Novelty and Causation: The Perceptual View", in *Some Problems of Philosophy*, ch. 13, New York: Longmans, Green & Co.
- Kim, J. (1999). *Mind in a Physical World*, Cambridge, MA: MIT press.
- Leibniz, G. (1714/1989). *Monadology*, in *G. W. Leibniz: Philosophical Essays*, R. Ariew and D. Garber (eds. and trans.), Indianapolis: Hackett Publishing Company.
- Lockwood, M. (1991). *Mind, Brain and the Quantum: The Compound "I"*, Oxford: Blackwell.
- Lotze, R. (1852). *Medicinische Psychologie, oder Physiologie der Seele*, Leipzig: Weidmann.
- Lycan, W. (1996). *Consciousness and Experience*, Cambridge, MA: MIT Press.
- Matson, W. (1966). "Why Isn't the Mind-Body Problem Ancient?", in *Mind, Matter and Method*, P. Feyerabend and G. Maxwell (eds.), Minneapolis: University of Minnesota Press.
- McGinn, C. (1999). *The Mysterious Flame: Conscious Minds in a Material World*, New York: Basic Books.
- Morgan, C. (1923). *Emergent Evolution*, London: Williams and Norgate.
- Mourelatos, A. (1986). "Quality, Structure, and Emergence in Later Pre-Socratic Philosophy", in *Proceedings of the Boston Colloquium in Ancient Philosophy*, 2, pp. 127-194.
- Nagel, T. (1979). "Panpsychism" in Nagel's *Mortal Questions*, Cambridge: Cambridge University Press.
- Nagel, T. (1986). *The View from Nowhere*, Oxford: Oxford University Press.
- Nagel, T. (1999). "Conceiving the Impossible and the Mind-Body Problem", in *Philosophy*, 73 (285), pp. 337-352.
- Paulsen, F. (1904). *Einleitung in die Philosophie*, 13th edition, Stuttgart: Cotta. (An English translation of an earlier edition is *An Introduction to Philosophy*, F. Thilly (trans.), New York: Holt, 1895.)
- Penrose, R. (1989). *The Emperor's New Mind*, Oxford: Oxford University Press.
- Piaget, J. (1932/1973). *The Language and Thought of the Child*, London: Routledge and Kegan Paul.
- Prince, M. (1885). *The Nature of Mind and Human Automatism*, Philadelphia: Lippincott.
- Royce, J. (1901). *The World and the Individual*, New York: Macmillan.
- Schiller, F. (1907). *Studies in Humanism*, London: Macmillan.
- Schopenhauer, A. (1818). *Die Welt als Wille und Vorstellung*, Leipzig.
- Seager, W. (1999). *Theories of Consciousness*, London: Routledge.
- Searle, J. (1992). *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.
- Searle, J. (1997). "Consciousness and the Philosophers", in *The New York Review of Books*, 44, 4, pp. 43-40).
- Shear, J. (1997). *Explaining Consciousness: The "Hard Problem"*, Cambridge, MA: MIT Press.
- Silberstein, M. and McGeever, J. (1999). "The Search for Ontological Emergence", in *The*

Philosophical Quarterly, 49:195, pp. 182-200.
- Spinoza, B. (1677/1985). *Ethics*, in *The Collected Works of Spinoza*, E. Curley (ed. and trans.), Princeton: Princeton University Press.
- Sprigge, T. (1983). *A Vindication of Absolute Idealism*, London: Routledge and Kegan Paul.
- Sprigge, T. (1999). "Panpsychism", in the *Routledge Encyclopedia of Philosophy*, London: Routledge.
- Strawson, G. (1997/1999). "The Self", in *Journal of Consciousness Studies*, 4, no. 5/6, pp. 405-28. Reprinted in S. Gallagher and J. Shear (eds.) *Models of the Self*, Thorverton: Imprint Academic, 1999 (my page references are to Gallagher and Shear).
- Tye, M. (1995). *Ten Problems of Consciousness*, Cambridge, MA: MIT Press.
- Waddington, C. (1961). *The Nature of Life*, London: Allen and Unwin.
- Wellman, H. (1992). *The Child's Theory of Mind*, Cambridge, MA: MIT Press.
- Whitehead, A. (1929). *Process and Reality: an Essay in Cosmology*, New York : Macmillan.
- Whitehead, A. (1933). *Adventures of Ideas*, New York: Macmillan. (My page references are to the 1961 Free Press (New York) edition.)
- Woodward, W. (1972). "Fechner's Panpsychism: A Scientific Solution to the Mind-Body Problem", in Journal of the History of the Behavioral Sciences, 8, pp. 367-86.
- Wright, Sewall (1953). "Gene and Organism", in *The American Naturalist*, v. 87.
- Wundt, W. (1892/1894). *Vorlesungen über die Menschen- und Thierseele*, Hamburg: L. Voss. An English translation is *Lectures on Human and Animal Psychology*, J.E. Creighton and E.B. Titchener (trans.), London: S. Sonnenschein.

Other Internet Resources

- Chalmers, D., Bibliography on Panpsychism, Part 1.4g of Contemporary Philosophy of Mind: An Annotated Bibliography
- More, T., Panpsychism, *The Catholic Encyclopedia*, Volume XI, New York, Robert Appelton, 1911

Related Entries

consciousness | Descartes, René | dualism | emergent properties | epiphenomenalism | Hartshorne, Charles | James, William | Leibniz, Gottfried Wilhelm | mereology | mind: philosophy of | monism | pantheism | physicalism | qualia | quantum theory: and free will | Royce, Josiah | Spinoza, Baruch [Benedict]

William Seager
seager@scar.utoronto.ca
and
Sean Allen-Hermanson

Univerity of Toronto
sean.allen.hermanson@utoronto.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 23, 2001
Content last modified: May 23, 2001

Stanford Encyclopedia of Philosophy
Notes to Panpsychism

Notes

1. A possible variant is, then, a doctrine which asserts that only *some* of the basic physical constituents of the world possess mental attributes (for example, as it might be, only electrons). Since there seems to be nothing upon which to base excluding certain of the fundamental physical entities from the realm of the mind, and since the panpsychist would regard the position that the physically fundamental entities should be conceived of as potentially mindless as wrongheaded, such a position has never been developed.

2. The appropriateness of the analogy is suspect since energy, although certainly ubiquitous, may not be as fundamental as the panpsychist would wish. The energy of systems is usually reducible, as in the common Hamiltonian description of a system, to velocities, masses and positions (within fields that yield potential energy).

3. Piaget's conclusions may be overstated although animistic tendencies in children are obviously present (for views contrary to Piaget's see Carey 1985, Wellman 1992). It is impossible to resist pointing out that the apparent decline in children's willingness to make widespread mentalistic attributions from 1929 to 1992 may reflect the increasing cultural ascendancy of scientific and materialistic modes of thought over that some time span.

4. Another Presocratic philosopher who has been said to espouse panpsychism for reasons similar in form to those of Thales (that is, via analogy and indeed an analogy with motion production) is Anaximenes (whose dates within the sixth century BC are uncertain), who in some way identified "air" (or "breath") with soul or mind, thus making mind ubiquitous.

5. Empedocles is sometimes regarded as a panpsychist because of the universal role of love and strife (see Edwards 1967 for example) but there seems little of the mental in Empedocles's conceptions, which are rather more like forces of aggregation and dis-aggregation respectively (see Barnes 1982, pp. 308 ff.).

6. It is important to note that although we, overall rightly, regard Newton as one of the founders of the materialist world view he was reluctantly forced to imbue matter with a mysterious power that transcended the pure mechanics of Descartes (hence my scare quotes above). The postulation of a "force" of gravity added an intrinsic power to matter that many, including Newton himself, regarded as inconsistent with a scientific understanding of the world. Some philosophers have seen the notion of force as a transmuted form of the concept of *spirit*, and thus as a limited and covert importation of mentalistic features into matter.

7. The notion that biological species should be thought of as individuals rather than collections of individuals has received some attention in contemporary philosophy of biology (see Hull 1976, DeSousa 1989). It is, of course, another matter whether species, even if they are granted status as individuals, ought to be granted minds but it presumably cannot be ruled out *a priori*.

8. Interestingly, Jaegwon Kim (1999) has recently endorsed the basic cogency of Nagel's argument, though in a back-handed way (Kim, so to speak, favors *modus tollens* over *modus ponens*). Rather than accept any form of panpsychism -- indeed he does not consider the option at all -- Kim prefers to defend physicalism by returning us to the days of reductive materialism; he sees no other way to avoid the charge that mental properties are epiphenomenal. Panpsychism appears to avoid both epiphenomenalism and reductionism, though at the cost of rejecting traditional forms of physicalism.

9. The mental lives of protozoa was apparently a lively topic at the time. Alfred Binet, the pioneer in intelligence testing, wrote the intriguingly entitled *Études de psychologie expérimentale: le fétichisme dans l'amour, la vie psychique des micro-organismes, l'intensité des images mentales, etc.* in 1891 (an English translation -- *The Psychic Life of Microorganisms* -- appeared in 1889).

10. Recent defenders of the so-called representational thedefense of consciousness (see for example Dretske 1995 or Tye 1995) contend that the informational aspect of consciousness exhausts its nature. We need not go so far here; it is enough that consciousness clearly does possess such an aspect.

11. Someone with preexisting emergentist leanings will tend to see this as trivialising panpsychism or rendering it empty. As McGinn says: "the weak version [of panpsychism] says that matter has some properties or other, to be labeled "proto-mental", that account for the emergence of consciousness from brains. But of course *that* is true" (1999, p. 99).

William Seager
seager@scar.utoronto.ca
and
Sean Allen-Hermanson
Univerity of Toronto
sean.allen.hermanson@utoronto.ca

First published: May 23, 2001
Content last modified: May 23, 2001

Pantheism

Pantheism is a metaphysical and religious position. Broadly defined it is the view that (1) "God is everything and everything is God ... the world is either identical with God or in some way a self-expression of his nature" (Owen 1971: 74). Similarly, it is the view that (2) everything that exists constitutes a "unity" and this all-inclusive unity is in some sense divine (MacIntyre 1967: 34). A slightly more specific definition is given by Owen (1971: 65) who says (3) "‘Pantheism’ ... signifies the belief that every existing entity is, only one Being; and that all other forms of reality are either modes (or appearances) of it or identical with it." Even with these definitions there is dispute as to just how pantheism is to be understood and who is and is not a pantheist. Aside from Spinoza, other possible pantheists include some of the Presocratics; Plato; Lao Tzu; Plotinus; Schelling; Hegel; Bruno, Eriugena and Tillich. Possible pantheists among literary figures include Emerson; Walt Whitman, D.H. Lawrence, and Robinson Jeffers. Beethoven (Crabbe 1982) and Martha Graham (Kisselgoff 1987) have also been thought to be pantheistic in some of their work-if not pantheists.

The book recognized as containing the most complete attempt at explaining and defending pantheism from a philosophical perspective is Spinoza's *Ethics*, finished in 1675 two years before his death. In 1720 John Toland wrote the *Pantheisticon: or The Form of Celebrating the Socratic-Society* in Latin. He (possibly) coined the term "pantheist" and used it as a synonym for "Spinozist." However, aside from some interesting pantheistic sounding slogans (like "Every Thing is to All, as All is to Every Thing"), and despite promising "A short Dissertation upon a Two-fold philosophy of the Pantheists" Toland's work has little to do with pantheism.

Pantheism and Theism

Where pantheism is considered as an *alternative* to theism it involves a denial of at least one, and usually both, central theistic claims. Theism is the belief in a "personal" God which in some sense is separate from (transcends) the world. Pantheists usually deny the existence of a personal God. They deny the existence of a "minded" Being that possesses the characteristic properties of a "person," such as having intentional states, and the associated capacities like the ability to make decisions. Taken as an alternative to, and denial of, theism and atheism, pantheists deny that what they mean by God (i.e. an all-inclusive divine Unity) is completely transcendent. They deny that God is "totally other" than the world or ontologically distinct from it. The dichotomy between transcendence and immanence has been a principal source of philosophical and religious concern in Western and non-Western traditions; and all major traditions have at times turned to pantheism as a way of resolving difficulties associated with the theistic notion of a transcendent deity or reality.

Not all of the problems generated by the theistic notion of God are also problems for pantheism. But given a suitable reformulation, some of them will be. And, as expected, pantheism will also generate some difficulties peculiar to itself. Thus, although evil and creation do not present identical problems for pantheism and the theism, and may even be inherent to theism; it *may* also be possible to reformulate them in a way that makes them applicable to pantheism. There may be pantheistic counterparts to the problem of evil and other classical theistic problems, and perhaps they can be resolved by pantheism.

There are probably more (grass-root) pantheists than Protestants, or theists in general, and pantheism continues to be the traditional religious alternative to theism for those who reject the classical theistic notion of God. Not only is pantheism not antithetical to religion, but certain religions are better understood as pantheistic rather than theistic when their doctrines are examined. Philosophical Taoism is the most pantheistic, but Advaita Vedanta, certain forms of Buddhism and some mystical strands in monotheistic traditions are also pantheistic. But even apart from any religious tradition many people profess pantheistic beliefs-though somewhat obscurely. Pantheism remains a much neglected topic of inquiry. Given their prevalence, non-theistic notions of deity have not received the kind of careful philosophical attention they deserve. Certainly the central claims of pantheism are *prima facie* no more "fantastic" than the central claims of theism-and probably a great deal less so.

Is Pantheism Atheistic?

Like "atheism" the term "pantheism" was used in the eighteenth century as a term of "theological abuse," and it often still is (Tapper 1987). A.H. Armstrong says the term "pantheistic" is a "large, vague term of theological abuse," (Armstrong 1976: 187). With some exceptions, pantheism is non-theistic, but it is not atheistic. It is a form of non-theistic monotheism, or even non-personal theism. It is the belief in one God, a God identical to the all-inclusive unity, but pantheists (generally) do not believe God is a person or anything like a person. The fact that pantheism *clearly* is not atheistic, and is an explicit denial of atheism, is disputed by its critics. The primary reason for equating pantheism with atheism is the assumption that belief in any kind of "God" must be belief in a personalistic God, because God must be a person.

In his non-pantheistic phase, Coleridge claimed that "every thing God, and no God, are identical positions" (McFarland 1969: 228). Owen (1971: 69-70) says, "if ‘God’ (theos) is identical with the Universe (to pan) it is merely another name for the Universe. It is therefore bereft of any distinctive meaning; so that pantheism is equivalent to atheism." Similarly, Schopenhauer (1951: 40) said that "to call the world ‘God’ is not to explain it; it is only to enrich our language with a superfluous synonym for the word ‘world’." The charge that pantheism is atheistic is as old as pantheism itself. Christopher Rowe (1980: 54-5) says, "When Cicero's Velleius describes Speusippus' pantheism as an attempt to ‘root out the notion of gods from our minds’, he is echoing a charge which was commonly made against the pantheism of the earlier Greek natural philosophers ... like Anaximander or Heraclitus. These tended to be identified as atheists in the popular mind; and indeed Plato himself implies a similar view ... the opponents who classify them as atheists are in reality attacking them for undermining traditional beliefs about the gods-or, to borrow a phrase from the indictment against Socrates, ‘for not believing in the gods the city believes in’."

At most, what Schopenhauer, Coleridge, Owen etc. can show, and probably all they intend, is that the pantheistic Unity can be explained in terms that would either eliminate the notion of deity from pantheism altogether, or that it is incoherent. They want to show that believing in a pantheistic God is a convoluted and confused way of believing in something that can adequately be described apart from any notion of deity-and in this they are mistaken.

Unity

Different versions of pantheism offer different accounts of the meaning of "unity," and "divinity." There is no one meaning in all forms of pantheism, and within some forms several types are found. Often, the meaning of unity present is vague and indeterminate. Because of this, the central problem of pantheism, unlike theism, is to determine just what pantheism means. For example, philosophical Taoism is one of the best articulated and thoroughly pantheistic positions there is. The *Tao* is the central unifying feature. What kind of unity is (or should be) claimed by pantheists and which, if any, is plausible? After dealing with these fundamental questions, the philosophical and religious consequences of analyzing unity in some particular way can be examined. There may be acceptable alternative criteria of Unity. But even if

there are alternatively acceptable criteria, some may be more acceptable to the pantheist than others-given criteria of adequacy in addition to those necessary. Among those that are acceptable, they need not be equally acceptable. However, just as there are alternative theisms, one would expect that there are alternative pantheisms. Pantheism need not be, any more than theism needs to be, a univocal view.

Misunderstandings

Schopenhauer criticized pantheism's identification of "the world" with "God," on the basis of what he took to be the meanings of both for the pantheist. He said calling the world "God," or God "the world," is "superfluous," and redundant. He also ridiculed the idea that the world could be called God given our general notions of what God and the world are like. Schopenhauer's criticism fails because he equivocates on the terms central to his argument. The meanings of both Unity and divinity involved in the pantheistic claim that there exists an all-inclusive divine Unity are different than the senses Schopenhauer attributes to the world and God in his criticism. The pantheist does not mean what Schopenhauer means by God, and the "all-inclusive unity" in pantheism is not another word for the "world" as he uses it (i.e. everything). The interpretation of "world" Schopenhauer attributes to pantheists is not what they mean when they describe it as a Unity.

For the pantheist, however Unity is interpreted, the world is not simply an all-inclusive Unity in the sense that the world, understood to be everything, is the "unity" composed of everything. This would be to interpret it as asserting that everything that exists simply is everything that exists; or to put it another way, everything is (of course) all-inclusively everything. This is true but vacuous, and it trivialises pantheism at the outset.

Attributing Unity simply on the basis of all-inclusiveness is irrelevant to pantheism. Formal unity can always be attributed to the world on this basis alone. To understand the world as "everything" is to attribute a sense of unity to the world, but there is no reason to suppose this sense of all-inclusiveness is the pantheistically relevant Unity. Similarly, unity as mere numerical, class or categorical unity is irrelevant, since just about anything (and everything) can be "one" or a "unity" in these senses. Suppose "formal unity" to be "the sense in which things are only in virtue of the fact that they are members of one and the same class ... the same universal" (Demos 1945-6: 534-545). Then clearly formal unity is not pantheistic Unity. Furthermore, formal unity neither entails or is entailed by types of unity (e.g. substantial unity) sometimes taken to be Unity. Hegel's *Geist*, Lao Tzu's *Tao*, Plotinus' "One," and arguably Spinoza's "substance," are independent of this kind of formal unity.

Unity is explained in various ways that are often interrelated. These connections range from mutual entailment, to different types of causal and contingent relations. Roughly, Unity is interpreted 1) ontologically; 2) naturalistically-in terms of ordering principle(s), force(s) or plans; 3) substantively-where this is distinguished from "ontologically"; and 4) genealogically-in terms of origin. Christopher Rowe (1980: 57) calls 4 a "genealogical model of explanation" of unity. "Thales, Anaximander, and Anaximenes, the Milesian monists appear to have claimed that what unifies the world is that it sprang from a single undifferentiated substance.

Unity may have to be explained partly in terms of divinity. The all-inclusive whole may be a Unity because it is divine-either in itself (Spinoza's substance), or because of a divine power informing the whole-as with the Presocratics. The Presocratics give an account of why they think the unifying principle is divine. It is immortal and indestructible. But this does not satisfactorily explain the relation between Unity and divinity, or why divinity might be seen as a basis of Unity. Similarly, though less naturally, the question arises as to whether the all-inclusive whole is divine because it is a Unity. Can Unity be a basis for attributing divinity to the whole? If divinity is the basis for Unity, as it may be for the Presocratics; or alternatively if Unity is the basis for divinity; then there is something of a redundancy in the definition of pantheism as the belief that everything that exists constitutes a divine Unity. A simpler non-redundant definition would be that pantheism holds that "everything is divine".

Divinity

"Divine" is defined as pertaining to God ("of, from, or like a god"), but also as "sacred" or "holy." Either definition suits the present purpose, since determining why pantheists regard the Unity as divine, or god, is equivalent to determining why they regard the Unity to be sacred or holy. The idea of "divinity" in pantheism is similar in some respects to its theistic meaning.

Why do pantheists ascribe divinity to the Unity? The reason is similar to why theists describe God as holy. They experience it as such. In Otto's (1950) experiential account, what is divine is what evokes the numinous experience. This can be a theistic god, but it can also be a pantheistic Unity. And, when looked at from socio-scientific perspectives in terms of how the concept of divinity functions intellectually and affectively (e.g., its ethical, soteriological and explanatory roles), its application in theism and pantheism is much the same.

There is no reason to suppose the idea of "divinity" relevant to pantheism should be modelled after a specific tradition's concept of divinity-like Christianity. At best, this tradition-dependent concept would be relevant to Christian/pantheist and other theist/pantheist hybrids (e.g. panentheism). It is too specific for any general analysis of pantheism, and it refers to the theistic variants of pantheism which are most inconsequential for pantheistic practice.

Whatever criteria are decided upon as necessary for attributing divinity to something, one cannot decide *a priori* that the possession of divinity requires personhood without ruling out the possibility of the most typical types of pantheism (i.e. non-personal types). After all, theism is what pantheism is most of all trying to distance itself from. I am not sure the reverse is true-but theism does ordinarily strongly oppose itself to pantheism. In any case, Spinoza's God and Lao Tzu's *Tao*, for example, are distinctly non-personal, as are the governing principles of the Presocratics. It seems unwarranted, therefore, to suppose that a necessary condition of something's being divine is that it be personal on the grounds that "Of all the modes of creaturely existence, personality is the highest and so the fittest to serve as an analogy of divine being" (Macquarrie 1984: 42). At least to do so begs the question against Spinoza, some of the Presocratics, Lao Tzu, probably Plotinus, as well as against experiential and socio-scientific accounts of

divinity.

Monism

Following a long and still current tradition H.P. Owen (1971: 65) claimed that "Pantheists are 'monists'...they believe that there is only one Being, and that all other forms of reality are either modes (or appearances) of it or identical with it." Although, like Spinoza, some pantheists may also be monists, and monism may even be essential to some versions of pantheism (like Spinoza's), pantheist's are not monists. Like most people they are pluralists. They believe, quite plausibly, that there are many things and kinds of things and many different kinds of value. Even in Spinoza's case, explaining his pantheism in terms of his substance monism glosses the far more significant, pantheistically speaking, evaluative implications he sees as entailed by that monism for his pantheistic metaphysic and his concept of Unity. The *Ethics* is not about monism, but about what it entails. Why Spinoza sees things as a Unity cannot be explained wholly or even primarily in terms of his monism.

Whether or not substance monism is ontologically necessary for Unity, an explanation of its relevance requires something extra-ontological to be cited. The same is true of any factual ground for Unity. Delineating metaphysical or modal properties of a substance, or anything else, does not make their relevance to Unity obvious. So what if everything is made from one self-subsistent immutable substance? So what if everything is really a single organism when considered macrocosmically? Why would this be pantheistically, rather than merely metaphysically significant? What is the evaluative or religious significance of natural features of the totality that pantheism claims is central to Unity? Because value must be partly constitutive of Unity, it must be explained in partly evaluative terms. This is a necessary condition for an adequate criterion of Unity. Without it one is left only with this or that fact as a basis for positing Unity, but no adequate account of the relevance of the basis, and so no account of Unity itself.

There may be ways of conceiving of the monistic "One" such that it is taken both as a unity and as "divine"-yet still not as a pantheistic Unity. The monistic unity (the "One") may not be regarded as a "Unity" (i.e. unity in some relevant pantheistic sense). Not just any monistic unity (e.g. mere substance monism) suffices for pantheism, whether or not it is also regarded as divine. Thus, although Hegel conceived of Reality as unified and rational in terms of the Absolute (Geist), and in a manner that I take it would qualify Geist as divine, he denies he was a pantheist. Similarly, Sankara's Brahman is ontologically all-inclusive and is part of a metaphysical account of the nature of Reality that is religiously significant (i.e. "Reality" is divine in some sense). However, it may be denied that advaita Vedanta, although monistic, is pantheistic. "Unity" is seen as absent from, or even antithetical to, essential aspects of advaita Vedanta such as its monism.

Monists, like pantheists, believe that Reality, or an aspect of it, is "One" or unified. Of course they also deny it is "One" or a "unity" in most other senses. Whatever similarities there are in this regard, there is insufficient reason for attributing pantheism to monists, because the oneness of Reality is neither a necessary nor sufficient condition of pantheism. It is at most a necessary condition if monistic "oneness" is construed in a unitive sense that is constitutive of some particular pantheistic account of the divine

Unity. An alleged entailment between pantheism and monism is even less likely since pantheists, like everyone else, are generally pluralistic. Any appearance to the contrary has been fostered by simply conflating Unity with monism, or by considering the few pantheists who were also monists and taking them as the norm. The connection between Spinoza's monism and his pantheism, does not rest on an identification of the two positions, but is instead the result of the wider metaphysical position constructed in his *Ethics*.

Substance monism need not have any implications concerning God or an Absolute in either a theistic or pantheistic sense. Differences among substance monists may be greater than differences between monists who deny and theists who affirm that God and creation are substantially distinct. For example, a substance monist (e.g. &Sankara-interpreted atheistically) need not identify substance with God, or recognize any God, at all. In this case it is plausible to hold that the difference between such an atheistic monist and a theistic or pantheistic monist is far greater than that between the theistic monist who perhaps holds that creatures and creator are co-substantial (though the theistic monist need not hold this view), and the theistic non-monist who believes that all creatures are substantially distinct from the creator. The latter two have their theism in common, while the former two have their monism in common. The latter two are "closer" in kind than the former, if (and so far as) one assumes that theism is a more significant common denominator than monism.

Transcendence

Like the notions of "Unity" and "Divinity," understanding transcendence and immanence is essential to any account of pantheism. A defining feature of pantheism is allegedly that God is wholly immanent. However, what is actually (or mostly) involved in this claim is that pantheism denies the theistic view that God transcends the world. Pantheism clearly does not claim that God in the theistic sense is immanent in the world since it denies such a God -- transcendent or immanent-exists. According to pantheism it is (of course) the pantheistic "God" (i.e. the all-inclusive divine Unity) that is immanent, not the theistic one. Theists and pantheists do not differ as to whether the theistic God is immanent or transcendent, but whether the theistic God exists. So to differentiate between them on the basis of one's affirming and the other denying immanence is utterly confused.

Many of the difficulties associated with theistic transcendence are not dissipated for the pantheist when relevantly adjusted. For example, theistic transcendence presents *prima facie* difficulties concerning knowledge of and relations with God. The pantheist is part of the Unity, but both the nature of Unity, and its practical implications must be determined. In the *Meditations* of Marcus Aurelius this appears as much a problem for pantheists, if Aurelius is one, as knowing and relating to God is for theists.

In a sense, the Unity in pantheism is wholly immanent, but this is bare ontological immanence that follows from the Unity's all-inclusiveness (i.e. there is nothing else). Yet even this overstates the pantheistic commitment to immanence. Aspects of the Unity or the unifying principle often have a transcendent aspect to them. Unity is "all-inclusive" but with the possible exception of Spinoza, pantheists generally deny complete immanence. Thus, the metaphysical *Tao* informs everything and is part of the all-

inclusive Unity, but it does have a transcendent aspect to it. It does transcend the phenomenal world of "myriad things." The same is true of Hegel's *Geist*, the Plotinian "One," and Presocratic unifying principles as well. So the claim that pantheists deny "God's" transcendence is altogether misleading on several counts unless taken to mean what it usually does mean when asserted by theists-which is that pantheists deny the transcendence of a theistic God.

If pantheism is seen as the quintessential expression of divine immanence, then it is not difficult to see why it might be combined with panpsychism or animism. Like pantheism, both of these express a kind of pervasive immanence-"mind" in the former case and "living soul," "spirit," or "animal life" in the latter. But however consonant or combined with pantheism these may be, they should be distinguished from both from each other and from pantheism. None of these three views entail one another, and the suggestion that pantheism and panpsychism naturally go together is vague apart from specific accounts of the two positions.

What immediately sets panpsychism apart from pantheism is its belief that mental activity, usually of a kind we can only at times be mildly aware of, is all-pervasive. Although such a supposition is not necessarily inconsistent with pantheism, it is not part of pantheism. Pantheism does not imply that the material/immaterial, or organic/inorganic dichotomies must be rejected. It does not reject these distinctions, but implies that Unity ranges over such divisions. There are other major differences between the two positions as well. Pantheism is a much broader theory. It has implications beyond the scope of panpsychism where the latter is seen as an account of the origin of mind and the relation between mind and matter.

Animism, panpsychism, and especially the doctrine of a world-soul as embodied in the macrocosm/microcosm distinction, have at times been equated with pantheism. These positions may be intrinsic to particular versions of pantheism, but pantheism as such is broader than these and distinct from them.

Creation

"Why is there something rather than nothing?" Pantheism rejects the theistic response that God exists necessarily and freely creates the universe from nothing. But does pantheism require an alternative doctrine of creation? What might such a doctrine be? For pantheism, creation remains problematic and even mysterious. However, difficulties associated with the theistic doctrine of creation *ex nihilo* (i.e. God creating the world out of nothing) vanish. If pantheism requires a creation doctrine, some type of emanationism seems most plausible. This is the type usually associated with, and probably most congenial to pantheism (e.g. Taoism, the Stoics, Plotinus) -- although pantheists can also eschew such doctrines.

Assuming pantheism does require a doctrine or view about creation, what can be said positively about it? Pantheism has a range of options unavailable to theism since the theistic doctrine is extrapolated from scripture. A pantheist might be a kind of existentialist with regard to questions like "Why is there

anything at all?" They could believe existence is a brute fact, with no explanation possible. This might be seen as a refusal to deal with the issue of creation-as rejecting the idea that pantheism requires a theory of creation suited to the notion of a divine Unity. But this is not necessarily so. For all its seeming negativity, this is a positive position and not one that simply denies other views. It is a theory of origin or creation that could be acceptable to some pantheists.

One reason any account of origin, including the view of existence as a brute fact, might be rejected as being especially relevant to pantheism, is that the account is not thought to be intrinsically connected to the notion of Unity. Indeed, pantheists might reject the idea that they require an account of creation intrinsic to their idea of Unity. Instead, *any* account that does not conflict with the way in which Unity is conceived of might be accepted.

In distinguishing between creation *ex nihilo* and emanationism as he does, Macquarrie (1984: 34-5) makes it easy to see why emanationism is often closely associated with pantheism. Emanationism is the view that "creation" is not a "making," but in some sense a "flowing forth" from God or its origin, as Macquarrie puts it. And, what "flows forth" "maintains a closer relation to [its] origin. It participates in the origin, and the origin participates in it." He says, "...emanationism does not necessarily lead to pantheism, but it does imply that in some sense God is in the world and the world is in God."

Even though doctrines of creation *ex nihilo* do not necessarily conflict with that central pantheistic claim, they are usually seen as doing do-partly because they are associated with other incompatible theistic elements (e.g. the creator is a person). On the other hand, emanationism appears to provide a doctrine which-if not an explicit ground on which to base pantheism-is at least one that is seen as congenial. As a doctrine of creation, it may even provide a partial basis for pantheism-as it has (arguably) for Plotinus, Eriugena, and even for Spinoza where "God" is the immanent cause of all things. The view that God is the "immanent cause" of things is a kind of creation doctrine for Spinoza and a basis for Unity. So far as Lao Tzu has a doctrine of creation it too is emanationist. "The Tao engenders one, One engenders two, Two engenders three, And three engenders the myriad things" (*Tao Te Ching*, XLII) (Ku-ying 1981: 49). The *Tao* is "the primordial natural force, possessing an infinite supply of power and creativity" (Ku-ying 1981: 6). Not only does the *Tao* create things-it is responsible for, or makes possible, their growth. "It nourishes them and develops them ... provides for them and shelters them"(*Tao Te Ching,* LI).

Emanationism tends to affirm rather than deny a common ontological, substantial, and evaluative base among everything that exists (e.g., whatever it is which creatively emanates, it is "Good"). It is therefore seen as in keeping with the central tenets of pantheism, and where pantheists adhere to a doctrine of creation it tends to be emanationist. Since Unity must partly be explained evaluatively, the fact that emanationism is often linked to the "Good" provides further reason for supposing it consonant with pantheism. Thus, although Macquarrie is right in claiming that the emanationist view of creation "does not necessarily lead to pantheism," the implication is that it often does.

Evil

The problem of evil is basically a theistic one that is not directly pertinent to pantheism. It is not, as Owen (1971: 72) claims, "an embarrassment" intellectually speaking, to pantheists, nor can it be. The "problem of evil," as it appears in classical theism, cannot be relevant to pantheism since pantheism rejects all of the aspects of theism that are essential to generating the problem. The "problem of evil" is peculiar to theism. This conflicts with the common view among Spinoza's earliest critics that pantheism, unlike theism, can neither account for evil nor offer any resolution to the problem of evil. The reason for claiming pantheism cannot account for evil usually rests on an unwarranted conflation of pantheism with monism, and on the even more untoward supposition that the pantheist's "God" is "theistic" in important respects.

It is not the case that pantheism need not address the existence of evil and associated moral issues. It offers both its own formulation(s) of a "problem of evil" and its own responses. However, the very idea of evil may be something the pantheist wishes to eschew. "Evil" is essentially a metaphysical rather than a moral concept; or it is moral concept with a particular theistic metaphysical commitment. The pantheist may prefer, as most contemporary ethical theorists do, to talk of what is morally or ethically right and wrong. The term "evil" could be retained and applied to particular (usually extreme) instances of moral wrongness, but it would be understood in a sense that divorces it from its original theological and metaphysical context.

Given the classical argument from evil in either its logical or empirical versions it is surprising that anyone should think evil presents any problem whatsoever for the pantheist; for example, that evil counts against the existence of the pantheistic Unity in a way similar to the way in which it counts against the existence of the theistic God. Evil might be taken to be indicative of a lack of pantheistic Unity, as evidence of some kind of chaos instead. But it cannot count against the existence of a pantheistic Unity in the way it can count against the existence of a theistic God. The argument from evil states that given the following propositions it is either impossible that God exists, or it improbable that God exists. 1) God is omnipotent, omniscient, and perfectly good. 2) God would prevent all preventable evil. 3) The world contains preventable evil. The pantheist rejects the proposition needed to generate the problem to begin with. The pantheist accepts (3) "The world contains preventable evil." The pantheist also accepts that if there was a theistic God, which for the pantheist *ex hypothesi* there is not, then (2) "God would prevent preventable evil." But the pantheist rejects (1) "God is omnipotent, omniscient, and perfectly good." Undeniably there is evil in the world that could be prevented, and supposing there was a theistic God one would assume that he would prevent it. But since there is no such God why suppose that proposition (3) requires some kind of special explanation or is cause for any "unease" on the part of the pantheist? The existence of preventable evil, for all that has been, does not even constitute a *prima facie* reason for rejecting the coherence of a pantheistic notion of Unity, or the probability of the existence of Unity. (3) is not incompatible with anything the pantheist believes to be true. Certainly it is not incompatible with (1) since the pantheist denies the truth of (1), and it is not incompatible with (2) which is only hypothetically true for the pantheist. The pantheist has no need to explain evil, or to explain evil away-at least not in any way resembling theism's need to do so.

Evil may be a problem for the pantheist, but it is not the kind of problem that it is for the theist. It does not even conflict, *prima facie* with the existence of a divine Unity. Pantheism does not claim that its divine Unity is a "perfect being" or being at all (generally), or that it is omniscient etc. Surely it is mistaken to

interpret Spinoza's "God" as "perfect" and "omniscient" etc. in anything like the way these predicates are interpreted theistically as applying to God. It might be supposed that the existence of evil is inconsistent or incongruous with the "divinity" of the Unity. But this would have to argued. In theism it is assumed that what is divine cannot also be (in part) evil. But why assume this is the case with pantheism? Even in Otto's account of the "holy" the holy has a demonic aspect. There seems little reason to suppose that what is divine cannot also, in part, be evil. At any rate, there is little reason for the pantheist to argue that what is divine can also be evil, since they can deny that evil falls within the purview of the divine Unity. To say that everything that exists constitutes a divine Unity (i.e. pantheism's essential claim) need not be interpreted in such a way so that it entails that all parts and every aspect of the Unity is divine or good. There can be a Unity and it can be divine without everything about it always, or even sometimes, being divine.

Ethics

Pantheists, like theists, tend to be "moral realists." They believe it is an objective fact that some kinds of actions are ethically right and others wrong, and what is right and wrong is independent of what any person thinks is right and wrong. With the exception of religious ethics, moral realism has not been a widely accepted philosophical position in recent times. However, the pantheist, like the theist, is not troubled by the fact that her moral realism is based on metaphysical assumptions that some regard as otiose. Furthermore, pantheists, like theists, generally think that moral judgements, and value judgments generally, are not empirically verifiable-at least not in the way one verifies matters of fact generally.

"Natural properties" are properties such as being a certain colour, shape, temperature or height, causing pain, "producing the greatest good for the greatest number" etc. They are properties that one can, in principle, verify that an object or action has or lacks. Some ethical "naturalists" (e.g., some Utilitarians) claim that moral properties are identical with natural properties. For example, a morally right action is sometimes equated with the action which "produces the greatest good for the greatest number." Others claim that moral properties are entailed by natural properties. Pantheists, however, generally believe that moral properties are both distinct from natural properties and are not entailed by them. Thus, they are usually "nonnaturalists.

Despite their nonnaturalism, pantheists, like theists, reject G.E. Moore's contention that these properties (i.e. goodness and badness) are ultimate and irreducible. For the theist the fact that "X is wrong" will be explained, and partially analysed, in terms of (even if not reducible to) nonnatural facts about God's will and nature. And, for the pantheist the fact that "X is wrong" will be explained, and partially analysed, in terms of (even if not reducible to) nonnatural facts about the divine Unity. Nonnaturalism is the position most congenial to pantheism, but a pantheist could make a case for being an ethical naturalist just as one could argue for a naturalistic theistic ethics.

Pantheism leaves the option between ethical naturalism and ethical nonnaturalism open. For the pantheist, though perhaps not for the theist, value-properties and predicates may be empirical or natural, or supervene upon natural properties, even if they are not entailed by such properties. So pantheists may be

ethical naturalists. This may be the case even if assertions containing value predicates are not taken to be empirically verifiable in any straightforward way as they often are for naturalism. Such value-predicates are not "empirical" in a narrow sense in which facts in the physical or even psychological sciences are empirical; but neither are they facts about some transcendent reality. Pantheism may, in a sense, deny the existence of any properties that are not "natural." It depends on how much one is willing to broaden one's notion of "natural." Of course, classifications such as "objectivist" and "nonnaturalist," are only a partial explanation of pantheists' ethical views.

Ecology

It is not accidental that pantheism is often taken to be a view inherently sympathetic to ecological concerns. This makes a decision to deal with ecology alongside pantheistic ethics less artificial than it might be if I were discussing, for example, theism and ethics-or a particular normative theory of ethics. There is a tendency to picture pantheists (i.e. pantheists other than Spinoza), outdoors and in pastoral settings. This has roots in the Stoics' veneration of nature, and in the much later nature mysticism, and perhaps pantheism, of some of the nineteenth century poets such as Wordsworth and Whitman. It has been fostered in the twentieth century by pantheists such as John Muir, Robinson Jeffers, D.H. Lawrence and Gary Snyder who explicitly "identify" with and extol nature, and claim people's close association and identification with "nature" and the "natural" is necessary to well-being. The belief in a divine Unity, and some kind of identification with that Unity, is seen as the basis for an ethical framework (and "way of life") that extends beyond the human to non-human and non-living things. The divine Unity is, after all, "all-inclusive."

A pantheistic ecological ethic will not be anthropocentric. This rules out the notion of man as a "steward of nature," whether his own or God's, who is responsible for nature. It also rules out utilitarian, contractarian, and Kantian approaches as providing an ultimate basis since they are anthropocentric. It does not, however, rule out contractarian etc. principles as useful guides to making and justifying environmental decisions. Applying anthropocentrically conceived principles to environmental issues would suffice in many cases, but not all, to sound reasoning about the environment. (The practical problem environmentally speaking has been that almost no principles have been applied until recently. Selfish economic "forces," i.e. people, have ruled without restraint.) The situation here is no different than with respect to theism. For the theist, ultimate justification of ethics resides in a view about the nature of God. But the theist is not prevented, *qua* theist, from invoking less ultimate ethical principles.

The pantheist's ethic, her environmental ethic and her ethics more generally, will be metaphysically based in terms of the divine Unity. It will be based on the Unifying principle which accounts for an important commonality, and it will be the grounds for extending one's notion of the moral community to other living and non-living things. Everything that is part of the divine Unity (as everything is) is also part of the moral community. Aldo Leopold (1949: 219, 240) says, "The land ethic simply enlarges the boundaries of the community to include soils, waters, plants, and animals, or collectively, the land ... A thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends otherwise." Looking towards pantheism as a metaphysical justification of, for example, Leopold's

"land ethic" is not unreasonable-or no more unreasonable than pantheism itself is.

An anthropocentric view of morality can at best make the non-human and non-living world an object of moral consideration. But it cannot, according to some, provide a basis for regarding those things as having a "good" of their own or as being non-human members of a moral community. Pantheists (and theists) will generally reject any environmental ethic as unsound if it fails to regard the non-human world as a full-fledged member of the moral community. In their view, to do otherwise is ultimately to rest the prospects of environmental well-being on the good will of the only members of the moral community there are-humans. This is seen like resting the welfare of colonies on the goodwill of the colonisers. In order to enlarge our understanding of the moral community in the appropriate ways a metaphysical basis for an environmental ethic is needed which limits the significance of the anthropocentric view.

Furthermore, it is clear that those, like deep ecologists, who argue that our notion of the moral community must be enlarged to include the "good" of the non-human and non-living, and that it is metaphysically correct to do so, also claim that practical consequences *are* involved. The issue is not merely one of providing a rational basis for an environmental ethic.

It may seem that pantheists can claim that ethics and an approach to ecology should be kept separate from, or that they are separate from, the more general pantheistic view that asserts the existence of a divine Unity. A kind of "separation between church and environment" might be proposed. But I doubt that such a separation is possible. The pantheist, like the theist or atheist takes the nature of reality as determinative of ethical requirements. Since Unity is predicated upon some evaluative consideration (e.g. the divine Unity being constituted on the basis of "goodness"), value is a focal point for the pantheist and a principle concern. This situation in regard to pantheism is not too different than the one for theism. For the theist, ethical requirements and evaluative concerns of all sorts are connected to God's alleged goodness, and overall nature.

Salvation and Immortality

Like the term "evil," "salvation" may be rejected pantheists as being too integral to the theistic world view they reject. It is a term borrowed from theism and one not consonant with pantheism. I use the term "salvation" with this in mind.

Pantheistic ethics are, in some ways, Aristotelian. For pantheism the notion of "the good life" as a regulative ideal-a *telos* or end to be strived for-is an aspect of salvation. This can be explained by examining some similarities between pantheistic ethics and Aristotelianism. The pantheist has what Paul Taylor (1975: 132) calls "an essentialist conception of happiness." Like the Aristotelian; Platonist; and theist-the pantheist's conception of happiness "presupposes that there is such a thing as an essential human nature." They all disagree as to what that essential nature is. The pantheist's conception of human nature, her philosophical anthropology, is generally broader and less specific than the others. When goals are stipulated that man *qua* man should achieve this indicates an essentialist conception of human nature.

Furthermore, in an essentialist conception of happiness (one which presupposes that there is such a thing as an essential human nature), "happiness" is largely a function of how well one fulfils one's essential nature. Pantheism's wide conception of human nature allows for a broad range of ways for people to achieve happiness. There are fewer ways for the Aristotelian or theist to achieve happiness than there are for the pantheist. To the extent that a human being is able to achieve "happiness" by actualising the properties that "define the good of man as such"--they will be leading an intrinsically good life. "Happiness" is then the standard by which to judge the non-derivative (intrinsic) value of a person's life.

Pantheism has a nonanthropocentric conception of human well-being. The human good is characterised partly in terms of relational properties. One must have a certain kind of relation to the Unity in order to live "properly." The set of properties common and unique to humans, which also define the good for humans as such, include relational properties. When a person exemplifies their essential human nature in this way-and it can only be exemplified in this relational way-they are living the "Good" life and can thereby achieve well-being and happiness. This nonanthropocentric conception of human well-being constitutes pantheism's standard of human perfection and virtue. It is a standard of intrinsic value.

As in the case of Aristotle's essentialist conception of the nature of things, the Human Good (defined as it is in terms of human nature) will be different than an animal's good or a plant's good. For the pantheist, the Good of these other things must also be understood partly in terms of their relation to the Unity. Furthermore, the Good associated with various things (humans, plants, etc.) is incommensurable. There is no standard external to each kind of thing by which all things can be measured in terms of perfection, or virtue, or intrinsic value. There is no such thing as intrinsic value *per se* given an essentialist account of the nature's of things which includes essentialist standards of perfection. It is not just wrong to say that a human being is intrinsically more valuable than a tree. It is also nonsense. Of course this does not mean trees should not be used by people.

Taylor (1975) claims that according to the essentialist conception of human nature, the value achieved in human life by fulfilling the standard of intrinsic value is independent its consequences in the lives of others. If this is right then the pantheist will reject any unqualified account of the essentialist's standard of human perfection and virtue. (Indeed, an Aristotelian need not hold such an absolute non-consequentialist account either.) Intrinsic value is, of course, value that is non-derivative. But, what determines the intrinsic goodness in a person's life will, for the pantheist, rely on that person's relationship to the Unity. A person's "good" is partially constituted by the divine Unity of which everything is a part. In pantheistic terms it makes little sense to speak of the intrinsic value of a human life as measured against a standard independent of how that life affects others, since for the pantheist all such value, even so-called "intrinsic value," is partly derivative. The standard of intrinsic value and perfection cannot be determined without reference to the divine Unity. The essential nature and well-being of a person, or anything else, cannot be analysed apart from its context in relation to the Unity and everything it includes.

Although both theism and pantheism have essentialist conceptions of human nature, well-being on either of those accounts cannot be achieved apart from one's relation to others, or the consequences of one's actions for others. And, the pantheist and theist are not the only kind of essentialists for whom consequences and relations matter. For the Aristotelian, in order to achieve well-being it is necessary to

develop a certain kind of character. This requires, in part, certain virtues (e.g. courage, temperance etc.). Since the development and display of character and virtue is connected in significant ways with the consequences of an individual's actions in relation to other people-the concept of one life having "intrinsic value" apart from how it affects any other life is vacuous. Aristotle's account of the virtues makes a practical impossibility of living a "good life" that is fundamentally bad for others. Plato too claims that the virtuous life has its rewards for all. Thus, essentialist conceptions of human nature and the Good need not preclude, and may even entail, an account of persons in relation to other things. For the pantheist, "realising the good for man as man" must be interpreted in terms of the Unity. For pantheism, an essentialist account of human nature does not suggest that there is necessarily only one kind of ideal person or way to achieve happiness.

An essentialist conception of human nature may recognise a range of human natures compatible with "Human Nature" as such. Just as various plants are constituted in such a way that their different requirements must be met if they are to thrive and flourish (i.e. what constitutes their well-being varies), so too will conditions for a person's "well-being" vary from person to person. The pantheist maintains that there is no such thing as an (i.e. one) essential human nature-although some properties are shared. Yet given various human natures, well-being can only be achieved to the extent that the individual satisfies their own nature--achieve their own potential--in their particular circumstances in relation to the Unity. Pantheists eschew hierarchies that have as a criterion for the "good life" any particular intrinsic feature that certain human beings may have which others lack. A good mind used in a good way may help one lead a better life, but so will good looks and a good job.

Pantheists deny personal immortality. There is no life after death in the sense that it is "they" who survive. Historically, the denial of personal immortality is one of pantheism's most distinctive features. This is partly because it is in clear opposition to the theistic view. But, it is primarily significant because it is constitutive of the pantheist's world-view and ethos, and so has implications for pantheistic practice. Believing that one is not going to live again after one dies, just as believing one will live again, has implications for one's choices in this life. There, is of course, nothing like a direct correlation in terms of what one believes concerning immortality and how they choose to live. But for some people, seeing death as the permanent end of one's existence, or alternatively as a prolegomenon to another life, will be a constitutive factor of the ultimate context in which to live. The goals they choose to pursue, the relationships they have, their vocations, may to varying degrees be affected by their belief that death is or is not the permanent end of the individual. The pantheist need not believe that it would be tedious to live forever. They just claim that no one does. This fact is not so much something to be lived with-as to be lived in terms of. The denial of personal immortality is as determinative of how the pantheist lives as the belief in an afterlife is for the theist.

The fact that pantheists (e.g. Spinoza) deny personal immortality is at times given as reason why pantheism is atheistic. The doctrine of immortality is so central to classical Christian theism, that rejecting the former is taken as entailing the denial of the latter. Yet, denying personal immortality can hardly be regarded as grounds for atheism unless theism, with its insistence on personal immortality, is taken to be the only position asserting the existence of a "God" that is not atheistic. The doctrine of personal immortality is not even essential to all forms of theism. Since many theists, e.g. many Jewish theists, deny

immortality, it would seem this denial is neither a necessary nor sufficient condition of atheism.

People who are interested in personal immortality, like people who are not interested (perhaps because they do not believe people survive death) may nevertheless be concerned with their continued existence in an impersonal sense. Impersonal forms of "immortality," or surviving death, can include "surviving" in people's memories, being remembered for one's work, a bone in a reliquary, or becoming another part of the matter/energy cycle once again. One may want to remembered for what one has accomplished, or for the person one was. Impersonal "immortality" may seem to pale next to the theists' insistence on personal immortality and the meeting again of people known in this life. Nevertheless, people's notions of impersonal immortality may be important in various ways. Whether or not they believe in personal immortality, it matters to some people how they will be thought of. No doubt, people who believe in personal immortality are also generally concerned with the impersonal forms. Some may even value being remembered for something they produced as more important than personal survival after death. But typically, the person who believes in personal immortality regards it with a concern that they do not have for various impersonal types of survival.

Some pantheists believe in various types of non-personal immortality (e.g. Spinoza and Robinson Jeffers), and they regard this as significant for reasons other than, or in addition to, the reasons non-pantheists give. They reject the view that personal immortality is more valuable than impersonal immortality. This is not to say that if pantheists believed there was personal immortality they could not regard it as desirable. Perhaps the could even though the idea is anthropocentric and uncongenial to pantheism. But pantheists do not believe in personal immortality, and they regard some types of impersonal immortality as important on distinctively pantheistic grounds.

Robinson Jeffers suggests that what may be important to the pantheist, and regarded as "a kind of salvation," is neither the realisation of the theist's hope for personal immortality, nor the atheists' (or theists') desire to be remembered in certain ways-although the pantheist can desire this as well. Instead, what is distinctively significant is the recognition of the individual as a part of the Unity-what Jeffers calls the "one organic whole ... this one God." The "parts change and pass, or die, people and races and rocks and stars," but the whole remains. He says, "... all its parts are different expressions of the same energy, and they are all in communication with each other, influencing each other, [and are] therefore parts of one organic whole." (See, George Sessions 1977: 481-528). Part of what Jeffers is suggesting is that "salvation" (or immortality) it is not so much a matter of the fact of one's survival in some form; rather, "salvation" consists in the recognition of the "oneness" or Unity of everything. "[T]his whole alone is worthy of the deeper sort of love; and that there is peace, freedom, I might say a kind of salvation, in turning one's affections outward toward this one God, rather than inwards on one's self, or on humanity." This is impersonal rather than personal immortality or salvation, but it is different from the kinds of impersonal survival discussed above. It may even be regarded as a kind of personal salvation, since Jeffers suggests that salvation can be experienced for oneself while alive-and only when alive. Such salvation resembles neither impersonal forms of immortality, nor theists' personal life after death.

Pantheism in Practice

Can pantheists employ traditional modes of theistic and non-theistic practice such as worship, prayer, and meditation? What form might a distinctively pantheistic type of practice take? Pantheists believe in a divine Unity. Yet, in pantheism there is no apparent community of believers organised around their common (though not identical) beliefs by an established body of religious teaching and scripture. Without these traditional constituents of religion pantheists may find themselves wanting to practice their faith-seeking to relate their actions to their beliefs-and yet wondering how to go about it. Pantheists have to ask themselves what they should do given what they believe.

Even if we are not at all clear about what pantheists should do, it may seem we are relatively clear about what they believe. However, if theorists who claim that action sometimes explains belief, or that action and belief must be understood together are right, then it follows that we do not yet know what pantheists believe. Insofar as pantheists lack a distinctive practice, they may be taken not to believe anything (pantheistically) at all. Such theorists claim that systems of belief and practice, if not individual beliefs and practices, are intrinsically related so as to define one another-and they develop together. Therefore, it may not be possible to keep the question of what pantheists believe distinct from the question of what they do. One need not accept such theories to believe, as a matter of commonsense, that belief and practice are connected in such a way that they cannot be adequately understood apart from one another.

In attempting to construct an account of what the contemporary pantheist should do, it would be useful to examine practices that pantheists have traditionally undertaken. But it makes no sense to suppose contemporary pantheists could replicate the representation of social relations that pantheistic rituals might be analysed as having by a symbolist account. If a pantheistic ritual symbolically represents social relations, it represents those of its own society. At any rate, the point is largely moot since the practice of pantheism has never been associated with ritual practice but with a way of life. Thus, Lao Tzu, explicitly eschewed ritual, and Spinoza thought that while ordinary religious practice, ritual etc., was a good idea for the common people since it inculcated valuable ideals, it was beside the point for him. The fact that pantheistic practice has never been associated with ritual may partly explain why pantheism has not been practiced communally in a church.

In literalist or Geertzian terms it makes sense to ask what to do, given certain beliefs, in a way it does not for a symbolist. The kinds of practice suitable to pantheism are explicable in terms of beliefs literally and symbolically understood; and especially (in Geertz's account) in terms of a world view (e.g. belief in a divine Unity) and corresponding ethos. Thus, Lao Tzu describes the *Tao* as a metaphysical reality; as natural law or system of self-regulated principles; and also as a principle, pattern and standard for human conduct. One emulates the *Tao* after discerning its manifest characteristics in the phenomenal world, and to emulate the *Tao* is to practice *Taoism*. In "Song of Myself" Whitman articulates a world view, and evokes the connected ethos he envisages. For Spinoza, examining the nature and implications of Unity (substance) in the *Ethics*, and trying to live in accord with that account, was itself a form of pantheistic practice. Similarly, in writing and living as depicted in "Song of Myself," Whitman practiced the pantheism he preached. The relationship between the thought and practice of Hegel, Plotinus, Bruno etc. is less apparent, but should be of interest to pantheists. If pantheists find any of the various world views and 'ethos' described as consonant with their own, they may pattern their practices after those associated

with such views. In having a particular pantheistic view of the nature of things, certain practices and a way of life must, to an extent, follow.

The idea of looking to religions with pantheistic practices for examples of what to do may seem promising in a literalist or Geertzian approach. Similar kinds of practice should follow similar beliefs. The difficulty is that there seem to be no pantheistic traditions to examine-not even *Taoism*, since as practiced, it is not pantheistic. In traditional religions, practices that might be identifiable as pantheistic are always seen in the context of wider religious (e.g. theistic) practice. In traditions that are partly pantheistic like some native American Indian religions, it is difficult to discern how practices relating to pantheistic beliefs can be distinguished from various kinds of god and spirit worship. Since pantheism has largely been non-communal, individual pantheists, not traditions, must examined.

Religious practice is usually prescribed by teachings and doctrine, and informed by other beliefs widely held among the community of believers. Since there is no widely recognised body of scriptural or other religious teaching in pantheism and never has been (there is little doctrine and no church), there should be little in the way of prescribed practice. As already noted, the philosophical *Taoism* of the *Tao Te Ching* is pantheistic, but it has never been widely practiced and there is no body of ritual associated with it.

The kind of activity undertaken by a believer ideally reflects (i.e. is explainable in terms of) the way in which the religious object, and one's relation to it, is conceived. Differences in practice are the products of varying views on the nature of God and the world-set in the context of a more comprehensive world view. Since pantheistic and theistic accounts of God and the world are best regarded as mutually exclusive, it is likely that the practices of each would be dissimilar. Theistic practice, the intent and so forth, is inappropriate for the pantheist, and *vice versa*. Pantheists will not want to practice in a way that reflects beliefs they do not hold.

If specific pantheistic practices could be identified, these might be adapted to modern pantheism. Yet, to talk of adapting practices in this way is artificial. As a whole, practice neither precedes nor follows the body of beliefs formulated and codified by a religious community. It develops along with them. Even where religious beliefs are taken (e.g. Durkheim) to be rationalisations of practices that precede them, practice occurs in a context of shared conceptions, beliefs and concerns, and are-whether literally or symbolically-expressions of these. Ritual, and religious practice generally, is a product of conscious and unconscious, literal and symbolic, communal religious reflection. Given (and one wonders why) that there has been little structured pantheistic communal reflection, despite the fact that there are many pantheists, there is no identifiable pantheistic practice. There are only identifiable pantheistic world views and beliefs. This does not explain why individual pantheists have not developed recognisable rituals, unless a community of believers (i.e. a church) is necessary for such practices. The practice of pantheism seems confined to individuals acting in ways they see as according with the nature of things.

If contemporary pantheistic ritual exists, it is scarce. (Is the solstice gathering at Stonehenge pantheistic?) The extent to which one can self-consciously set out to construct a ritual is, for reasons already given, suspect. But, given that one can consciously construct symbols that address a community's concerns, there seems no reason why pantheistic rituals cannot be formulated. Indeed, various theistic rituals are self-

consciously created. Furthermore, ritual is only one aspect of religious practice, and pantheists may develop other ways to express their beliefs in action. Since belief and practice are interdependent and evolve together, if some future pantheistic communal reflection results in doctrines, then it is likely to result in practices of various sorts as well. Other than the fact they have lacked what seems to be requisite in terms of a community of pantheists, there may be additional or alternative explanations of why pantheists have not developed rituals. Maybe the lack of community can just as easily be explained by the lack of a developed mode of practice as *vice versa*.

Pantheists basically lack scripture and an established body of doctrine and discourse that could help establish the nature of pantheistic practice. However, it this is not entirely true. The pantheist can, to some extent, rely on traditional religious scripture that is recognisably pantheistic; for example, some *taoist* texts, and some Western and non-Western theistic scripture. Pantheists also have recourse to numerous philosophical sources-Spinoza etc. But, the pantheist is not without alternatives to the scripture and discourse theists have at their disposal. To some extent, the pantheist too will know what to do to practice pantheism. Art, music, literature and poetry, fulfil the same kinds of roles in pantheism as they do in theism. As representations of cultural patterns they reflect and sustain a world view and ethos. In Geertz's (1973: 93) terms they symbolically function as both a model *of* reality and a model *for* reality. "Culture patterns are ‘models’ ... they are sets of symbols whose relations to one another ‘model’ relations among entities, processes or what-have-you ... they give meaning, that is objective conceptual form, to social and psychological reality both by shaping themselves to it and by shaping it to themselves." Pantheists recognise cultural patterns and symbolic representations that "model" their beliefs. Given such beliefs, and the efficacy of symbolic representations of those beliefs; certain other beliefs, actions, and attitudes will be regarded, cognitively and affectively, as appropriate and correct.

In theistic traditions, prayer-which is a type of worship-and sometimes meditation, are the principle forms of religious practice. They are often set in the context of ritual. Theism gives a variety of reasons why prayer and worship are appropriate and necessary forms of theistic practice. But, what about for the pantheist? In principle, pantheists will not do things that literally conflict with the beliefs they express. They will not worship if worship implies the recognition of an independent and superior god, since this theistic belief is antithetical to a central tenet of pantheism. Are prayer and worship appropriate kinds of practice for the pantheist? Given that the pantheist should not pray to or worship a theistic God, can she worship the pantheistic Unity?

Worship and Prayer

Worship and prayer are not suitable to pantheism. It has often been claimed by theists and atheists that pantheistic worship (e.g. worshipping the Unity) is idolatrous. It is worshipping a false god. Unlike the theist or atheist, however, the pantheist believes a divine Unity exists-a kind of god. So pantheists, if they do worship the Unity, reject the idea that they are worshipping a false god. What is wrong with pantheistic worship is not that it is idolatrous, but something more basic having to do with both the nature of worship and Unity. Even if the Unity exists, worshipping it would not be proper pantheistic practice.

Pantheistic worship might naively be thought to be a kind of self-worship; worshipping something of which one is a part or identified with. This too is a mistake. As we have seen, pantheism is not the view that "everything that exists," including oneself, is god; and it is not the view that every particular thing or person is equally god. If worship is not acceptable religious practice for pantheists, it is for reasons other than that such practice involves adoring and venerating (i.e. worshipping) oneself. Worship and prayer are not consonant with pantheism. Like "evil" and "salvation," they are connected to the theistic world-view that pantheists reject. Therefore, except in a highly derivative sense (i.e., derivative from theism) worship and prayer are types of practice that are not acceptable to pantheists. Devotion to the universe, artistic expression, nature observation, etc., are not types of worship as theistically understood-though they may be ways of respecting, honouring, and revering.

What makes worship and prayer inappropriate for the pantheist is not the lack of ontological separation from the Unity that theism claims God has from the world. If there is a sense in which pantheists are ontologically, or in other ways, distinct from the divine Unity, worship and prayer are still inappropriate. If a necessary condition of worship is that it has to be in some significant sense "other regarding," then worship would not on that account be inappropriate to pantheism. What makes it unsuitable is that worship, and especially prayer, are basically directed at "persons"-or at a being with personal characteristics separate and superior to oneself. Whether one's reasons for worship are petitionary or devotional is irrelevant; and so is one's motivation-whether a Freudian way of coping with guilt, or a rationally-based sense of duty. Objects of worship are not oneself, and perhaps not even ontologically distinct from oneself as theism claims, but they are generally taken to be conscious, personal and superior.

Given the nature and goal of worship objects of worship must have a personal character. It might be thought that showing the pantheistic Unity should not, on conceptual grounds, be worshipped is rather uninteresting. That may be right. The implications of this result, however, are anything but insignificant. For the pantheist, the practical consequences of worship and prayer being unavailable as forms of religious practice are enormous.

In the theistic view, worship and prayer are practically synonymous with religious practice. And, even in (theoretically) non-theistic religious traditions such as Buddhism and Taoism, worship and prayer are frequent if not prevalent. Yet, the pantheist is faced with the problem of finding a way to practice pantheism that is consistent with the finding that worship and prayer make sense only in a theistic context. As a result, one of the defining and most noticeable characteristics of pantheism will be the type of practice it takes up. The practices involved, whatever they are, will be different not only from those in theistic traditions, but also from those in non-theistic ones in which theistic practice is so much a part.

Goal: Relationship or State?

Do pantheists seek a relationship with the impersonal Unity rather than a "state"? The choices of the religious objective for the pantheist are either a relationship with the Unity or a state of some kind. The kind of religious practice pantheism (like theism) engenders is a function of the kind of goal sought. What then is the religious objective of pantheism? If there were no such objective to pursue through practice,

the question of how to practice pantheism becomes superfluous. That there is a goal to pursue is intrinsic to the nature of religion.

In pantheistic systems such as Spinoza's or philosophical Taoism, the objective is best described as a state rather than a relation. However, just as theism correctly claims that although the principle goal of theistic practice is a relation to God, this also involves a "state of the individual"; so the pantheist claims that although pantheistic practice is principally concerned with a "state of an individual," a crucial and intrinsic aspect of this state is one's relation to the divine Unity. However, granted that a dichotomy between the objective as "relationship" or "state" is not firm; the principal form of practice-contemplative and meditative on the one hand, or worship on the other-follows from the objective emphasised. In theism it is on a personal relationship to God. In pantheism, the emphasis is on an individual state resulting from an understanding of, and a right relation to, the Unity. Practice will be contemplative and meditative rather than devotional. As in the case of theism, pantheistic practices-like the beliefs they are related to-are meant to have practical consequences both in terms of what one does, and more generally, the way one lives.

The question, of course, is how the pantheist is to arrive at "the right relation" to the Unity thereby achieving their objective. Answering this is the principal focus of both Spinoza's *Ethics*, the *Tao Te Ching* and most other pantheistic literature (e.g. Whitman's "Song of Myself"). What one actually does depends partly on the individual (i.e. Spinoza is no Whitman), and also on the particulars of the state sought. Since the pantheistic conception(s) of reality is ultimately very different from, for example, that of the Theravada Buddhist, there is no reason to suppose the pantheistic objective to be like *nirvana*. The pantheist's relation to the divine Unity does not entail the obliteration of self or liberation that a Buddhist's identification with Brahman does; nor is it like the theistic mystic's union with God. There may be aspects of the state pantheists seek that are similar to Buddhist goals, and even to theistic ones-though to a far lesser extent. But, even if the pantheist's objective is as different from what the Buddhist seeks as it is from what the theist seeks, the means for achieving it remain contemplative or meditative, rather than devotional.

For Spinoza, acquiring the happiness described in the *Ethics*is largely an intellectual achievement. It is difficult to see how one can attain the understanding and identification with "God" that Spinoza claims leads to peace of mind and "blessedness" (i.e., the highest achievement of the individual) without addressing the problem discursively rather than affectively by intuition and meditation-although discursive thinking and these other methods are by no means inconsistent. But even though Spinoza's approach involves little that is not discursive; it is contemplative, and the objective remains primarily a state rather than relation. Worship is not a mode of practice conducive to achieving the state Spinoza seeks. Granted that Spinoza's method is intellectualistic; other approaches are possible-especially where the objective itself is conceived of differently (i.e. less intellectualistic). Spinoza of course recognises that his own method is not suited to most people and acknowledges that ordinary practice such as worship and prayer may at times engender ends he describes. Just as theists use various methods to pursue their objective-some more intellectualistic than others-so in pantheism certain kinds of practices are suited to certain kinds of people. As in other religions, the means by which pantheists pursue their objectives are generally not overtly or overly intellectualised. To do so can undermine practice by upsetting the balance

between the affective and intellectual aspects of their belief system.

The pantheist is likely to view the kinds of goals that most religious traditions envision as excessive and grandiose-as neither believable nor desirable. What is more, although they are not humanists, like humanists pantheists are likely to view those objectives and related beliefs much as theistic traditions viewed those of "primitive religions" and of each other: as superstitiously anthropocentric and so capable of being naturalistically explained. The state sought by the pantheist supervenes (as in Taoism) on establishing the right relation with the Unity by means of cultivating a life suited to the nature of the Unity and of oneself. But for the pantheist this is a goal in itself, a this-worldly happiness. The pantheist eschews a notion of their being further goals; for example, the theist's beatific vision; personal immortality; *nirvana;* and even Spinoza's "blessedness," interpreted as something other-worldly.

The pantheist's happiness is nevertheless a special "state" that is difficult to achieve. Being a kind of utopian ideal it too is perhaps grandiose. Ordinary happiness is part of it but should not be conflated with the kind of thoroughgoing happiness the pantheist thinks it is possible to attain now and again. Much as Kierkegaard denied that "truth," "subjectivity" and even "immortality" are attainable once and for all, the pantheists denies their objective is a once and for all achievement. It is a state of well-being that involves a sustained peace of mind and the kind of happiness that comes from, or is identical with, such a state of mind. Since one's own state of mind and relation to the all-inclusive Unity are partly dependent upon other people and things, the state the pantheist seeks is not something achievable in isolation. Pantheism involves a this-worldly utopian vision based on individual's relations to, and identification with, the Unity.

"Nature"-which appears to be equated with the "Great Outdoors"-has pride of place in a pantheistic world view and ethos. It is assumed that pantheists are nature lovers, if not nature mystics. This view of pantheists as naturalists and rural outdoor people as opposed to city dwellers, is common. A reason for the pantheist's stress on Nature is that anthropocentricism is seen as incompatible with a proper recognition of Unity. It is seen as undermining the cosmocentric perspective required by pantheistic ethics, and a pantheistic way of life; as antithetical to the pantheist world view and ethos. Involvement in nature de-emphasises the anthropocentrism pantheism believes endemic to theism and detrimental to well-being and Unity.

This characterisation of pantheists as loving nature and as having to establish a relationship to things natural is what principally informs vague views as to how pantheism is to be practiced-especially among contemporary pantheists. Practice becomes an expression of a love of nature-usually by "communing" with it. It is no wonder pantheism is often regarded as little more than a type of nature mysticism. But for the pantheist, "love" of nature is expressed primarily in ethical rather than in mystical or quasi-mystical terms. Pantheistic ethics focuses on how to live and on the individual's relation to the natural order-an order of which others are a part. One's own well-being and that of others depends on it. Since nature is taken as intrinsically valuable, and because relating appropriately to nature presupposes its preservation and protection; nature in general and environmental issues in particular, are important to the pantheist. Like many others, pantheists see their well-being as intrinsically connected to the wider environment as well as to things more immediate (e.g. employment).

Is the urban person at a *religious* disadvantage from a pantheistic perspective? Without denying the significance Nature has for the pantheist (e.g., as a standard of behaviour, and as an object of meditation conducive to a "right" state of mind), is there reason to believe a pantheist who prefers an urban to a pastoral setting, and who likes technology, is risking spiritual depravity? Does the pantheist have a duty to spend time in natural settings if they prefer the city? Technology is associated with the Urban, and the pantheist may see much of it, or too much of it, as inimical to Unity and well-being. Technology is devalued when it is taken as undermining the kinds of value pantheism seeks to promote. Technology (people using it) despoils the environment. At any rate, since the world is increasingly urban, for pantheism to be viable it will have to be possible to practice it in cities.

A person who prefers city street life may claim there is a bias towards the non-human in a pantheist's exclusive insistence on Nature. Why cannot cities-themselves "natural" in a way-also be conducive to the practice of pantheism? Perhaps cities could be if they and many of their people were not as neglected and abused as much as some wilderness areas (if the comparison makes any sense). "God's country" for the pantheist denotes urban as well as pastoral settings-indeed it extends to the suburbs. Given the existence of a divine Unity one should not regard all personal preferences (e.g., for a garden), as cosmically endorsed. If the goal of pantheism is a way of life and a kind of "state," then any locale that is generally conducive to promoting those goals is acceptable. This may have more to do with the kind of urban or rural setting one lives in rather than just whether the setting is urban or rural.

In terms of its practice, one of the striking things about pantheism is that it has not produced a church or any kind of organisation engaged in overseeing its practice. Apparently a community of pantheists is not necessary for the practice of pantheism. Either this is an historical accident, or it has to do with structural features of pantheism. Pantheists, like many theists, tend to regard Churches and religious leaders with suspicion. The kind of orientation that the pantheist seeks vis a vis the Unity is not taken to be something a church can facilitate. The mediation churches provide is seen as superfluous or harmful-just as it has been by many mystics. Organised religions are seen as divisive and exclusivist, and churches perhaps are seen as essentially anthropocentric. It is for these kinds of reasons that there never has been a pantheistic church and probably never will be.

Pantheism remains a fertile subject for natural theology. Natural theologians have hardly approached it. Pantheism should be of interest to those in the philosophy of religion who seek a way out of the constrictions (often institutional ones) put upon them by working within the confines of classical theism; especially as the issues relating to classical theism have been taken up by the contemporary christian conservative analytic philosophers of religion. Perhaps pantheism will be of most interest to those who do not believe in a theistic God, yet are concerned with many of the traditional questions that natural theologians have always asked, and that religious traditions necessarily address.

Pantheism's lack of "success" in worldly terms on the religion market may have to do with the fact that it is antithetical to any power structure; the kind, for example, found in the Catholic church. If so, then even though pantheism may be more profoundly religious than institutionalised religions, it may be doomed to ineffectiveness because it cannot manipulate power-it cannot "play the game." Wielding various kinds of power has been a feature of religion from its most "primitive" to its most sophisticated levels-a feature

churches can generally not control. Pantheism negates the power struggle through its emphasis on Unity. It refuses to see religion in political and hierarchical terms. Pantheism is the religion that tries most completely to escape the limitations created by anthropocentric models of religion that create god in man's image.

Bibliography

- Armstrong, A.H. *The Architecture of the Intelligible Universe in the Philosophy of Plotinus.* Cambridge: Cambridge University Press, 1940.
- Armstrong, A.H. "The Apprehension of Divinity in the Self and Cosmos in Plotinus." In *The Significance of Neoplatonism.* Ed. R. Baine Harris. Norfolk:International Society for Neoplatonic Studies, 1976, pp.187-198.
- Aurelius, Marcus. *The Meditations*. Translated by G.M.A. Grube. Indianapolis: Hackett Publishing, 1983.
- Barnard, Frederick, M. "Spinozism." In *Encyclopedia of Philosophy.* Ed. Paul Edwards. New York: Macmillan and Free Press, 1967.
- Bayle, Pierre. *Historical and Critical Dictionary: Selections*. Translated by Richard Popkin. Indianapolis: BobbsMerrill, 1965.
- Bedell, Gary. "Theistic Realism and Monistic Idealism." *Thomist*, 35 (1971), pp. 661-683.
- Bedell, Gary. "Bradley's Monistic Idealism." *Thomist*, 34 (1970), pp. 568-579.
- Bedford, David. "God, Nature and the End of History." *History of European Ideas*, 19 (1994), pp. 371-376.
- Bennett, Jonathan. *A Study of Spinoza's Ethics.* Indiana: Hackett, 1984.
- Berman, David. *A History of Atheism in Britain: From Hobbes to Russell.* London: Croom Helm, 1988.
- Bradley, F.H. *Appearance and Reality*. Oxford: Clarendon Press, 1930.
- Bronson, Bertrand. *Walking Stewart*. University of California Publications in English, xiv. Berkeley and Los Angeles, 1943.
- Burch, George Bosworth. "Principles and Problems of Monistic Vedanta." *Philosophy East and West*, 11-12 (1961-63), pp. 231-237.
- Carman, John. *The Theology of Ramanuja*. New Haven: Yale University Press, 1974.
- Charlton, William. "Spinoza's Monism." *Philosophical Review*, 90 (1981), pp. 503-529.
- Chryssides, George D. "Subject and Object in Worship." *Religious Studies* 23 (1987), pp. 367-375.
- Copleston, F. C. *Religion and the One: Philosophies East and West*. London and Tunbridge Wells: Search Press, 1982.
- Copleston, F.C. "Pantheism in Spinoza and The German Idealists." *Philosophy*, 21 (1946), pp. 42-56.
- Copleston, F.C. "Man, Transcendence, and the Absence of God." *Thought*, 43 (1968).
- Crabbe, John. *Beethoven's Empire of the Mind*. Newbury, England: Lovell Baines, 1982.
- Crombie, I.M. *An Examination of Plato's Doctrines*, Vol. I. London: RKP, 1962.
- Dauenhauer, Bernard P. "Some Aspects of Language and Time in Ritual Worship." *International Journal for Philosophy of Religion*, 6 (1975), pp. 54-62.

- Demos, Raphael. "Types of Unity According to Plato and Aristotle." *Philosophy and Phenomenological Research*, 6 (1945-6), pp. 534-545.
- Deutsch, Eliot. *Advaita Vedanta: A Philosophical Reconstruction.* Honolulu: East West Center Press, 1969.
- Donagan, Alan. *Spinoza.* London: Harvester Wheatsheaf, 1988.
- Edwards, Paul. "Panpsychism." In *Encyclopedia of Philosophy*. Ed. Paul Edwards. New York: Macmillan and Free Press, 1967.
- Emerson, Ralph Waldo. *Nature*. 1836.
- EvansPritchard, E. E. *Theories of Primitive Religion*. Oxford: Oxford University Press, 1965.
- Farber, Marvin. "Types of Unity and the Problem of Monism." *Philosophy* and Phenomenological Research, 4 (1943-4), pp. 37-58.
- Ferre, Nels F.S. *Living God of Nowhere and Nothing*. Philadephia: Westminster Press, 1966.
- Feuerbach, Ludwig. *Principles of the Philosophy of the Future*. Translated by Manfred Vogel. Indiana: Hackett Publishing, 1986.
- Ford, Marcus. "Pluralistic Pantheism?" *Southern Journal of Philosophy*, 17 (1979), pp. 155-161.
- Ford, Marcus P. "William James: Panpsychist and Metaphysical Realist." Transactions of the Charles Peirce Society, 17 (1981) pp. 158-170.
- Forrest, Peter. "Some Varieties of Monism." In *Indian Philosophy of Religion*. Ed. R. W. Perrett. Dordrecht: Kluwer Academic Publishers, 1989, pp.75-91.
- Francks, Richard. "Omniscience, Omnipotence and Pantheism." *Philosophy*, 54 (1979) pp. 395-399.
- Frankenberry, Nancy. "Classical Theism, Panentheism and Pantheism: On the Relation between God Construction and Gender Construction." *Zygon*, 28 (1993).
- Funkenstein, Amos. *Theology and The Scientific Imagination from the Middle Ages to the Seventeenth Century*. Princeton: Princeton University Press, 1986.
- Furley, D. and Allen, R. *Studies in Presocratic Philosophy* vol. I. London: RKP, 1970.
- Geertz, Clifford. "Religion as a Cultural System." Chapter 4 in *The Interpretation of Cultures*. New York: Basic Books, 1973, pp. 87-125.
- Guthrie, W.K.C. *A History of Greek Philosophy, Vol. I.* Cambridge: Cambridge University Press, 1962.
- Harris, R. Baine., ed. *The Significance of Neoplatonism*. Norfolk: International Society for Neoplatonic Studies, 1976.
- Hastings, James, ed. *Encyclopedia of Religion and Ethics*. Edinburgh: Clark 1908-1926.
- Hegel, G.W.F. *Lectures on the Philosophy of Religion,* Vol. I. Berkeley: University of California Press, 1984.
- Henderson, E.H. "Theistic Reductionism and the Practice of Worship." *International Journal for Philosophy of Religion*, X (1979), pp. 25-40.
- Hick, John. *An Interpretation of Religion*. New Haven: Yale University Press, 1989.
- Hudson, W. Donald. "The Concept of Divine Transcendence." *Religious Studies*, 15 (1979), 197-210.
- Hume, David. *The Natural History of Religion and Dialogues Concerning Natural Religion.* Ed. A. Wayne Colver and John V. Price. Oxford: Clarendon Press, 1976.
- Hunt, Murray, W. "Some Remarks About the Embodiment of God." Religious *Studies*, 17 (1981),

pp. 105-108.

- Huxley, Aldous. *The Perennial Philosophy*. New York: Harper & Brothers, 1945.
- Inge, W.R. *The Philosophy of Plotinus vol II*. London: Longmans, 1928.
- Jaeger, Werner. *The Theology of The Early Greek Philosophers*. Oxford: Oxford University Press, 1947.
- Jantzen, Grace. "On Worshipping an Embodied God." *Canadian Journal of Philosophy*, 8 (1978), pp. 511-519.
- Jantzen, Grace. *God's World, God's Body*. Philadelphia: Westminster, 1984.
- Jantzen, Grace. "'Where Two are to Become One': Mysticism and Monism." In *The Philosophy in Christianity*. Ed. Godfrey Vesey. Cambridge: Cambridge University Press, 1989, pp. 147-166.
- Joad, C.E.M. "Monism in the Light of Recent Developments of Philosophy." *Proceedings of the Aristotelian Society*, 17 (1916-17), pp. 95-116.
- Kesarcodi Watson, Ian. "Is Hinduism Pantheistic?" *Sophia*, 15 (1976), pp. 26-36.
- Kisselgoff, Anna. "Dance: Graham's 'Canticle' Revived." Review of Martha Graham's "Canticle for Innocent Comedians," New York Times, October 9, 1987, p. C3.
- Kim, Chin-Tai. "Transcendence and Immanence." *Journal of The American Academy of Religion*, 55 (1987), pp. 537-549.
- Kochumuttam, Thomas. "Limits of Worship in Indian Religion." *Journal of Dharma*, 3 (1978), pp. 364-372.
- Kuying, Ch'en. *Lao Tzu, Text, Notes, and Comments*. Translated and adapted by Rhett Y.W. Young and Roger T. Ames. Republic of China: Chinese Materials Center, 1981.
- Kvastad, Nils Bjorn. "Pantheism and Mysticism." Part 1. *Sophia*, 14, no. 2, (1975), pp.115. Part II. *Sophia*, 14, no. 3, (1975), pp. 1930. Leopold, Aldo. "The Land Ethic." In *A Sand County Almanac*. New York: Oxford University Press, 1949.
- Leslie, John. *Value and Existence*. Oxford: Basil Blackwell, 1979.
- Leslie, John. *Physical Cosmology and Philosophy*. New York and London: Macmillan, 1990.
- Levine, Michael P. (ed.) *The Monist*. Issue on *Pantheism*, 80:2 (1997).
- Levine, Michael P. *Pantheism: A Non-Theistic Concept of Deity*. London and New York: Routledge, 1994.
- Levine, Michael P. "Pantheism, Ethics and Ecology," *Environmental Values*, 3 (1994), pp. 121-138.
- Levy, Donald. "Macrocosm and Microcosm." In *Encyclopedia of Philosophy*. Ed. Paul Edwards. New York: Macmillan and Free Press, 1967.
- Lipner, J.J. "The World as God's 'Body': In Pursuit of Dialogue With
- Ramanuja." *Religious Studies*, 20 (1984), pp. 145-161.
- Lipner, J.J. Ramanuja: *The Face of Truth*. London: Macmillan, 1985.
- Liu, Shuhsien. "The Confucian Approach to the Problem of Transcendence
- and Immanence." *Philosophy East and West*, 22 (1972), pp. 45-52.
- Liu, Shuhsien. "Commentary: Theism from a Chinese Perspective." *Philosophy East and West*, 28 (1978), 413-418.
- Lovejoy, Arthur. *The Great Chain of Being*. New York: Harper and Row, 1960.
- Lloyd, Genevieve. "Spinoza's Environmental Ethics." *Inquiry*, 23 (1980), pp. 293-311.
- MacIntyre, Alasdair. "Pantheism." In *Encyclopedia of Philosophy*. Ed. Paul Edwards. New York:

Macmillan and Free Press, 1967.

- Macquarrie, John. *Thinking about God*. London: SCM, 1975.
- Macquarrie, John. *In Search of Deity*. The Gifford Lectures 19834. London: SCM, 1984.
- Mamo, Plato. "Is Plotinian Mysticism Monistic?" In *The Significance of Neoplatonism*. Ed. R. Baine Harris. Norfolk: International Society for Neoplatonic Studies, 1976, pp. 199-215.
- Mathews, Freya. *The Ecological Self*. London and New York: Routledge, 1990.
- McFarland, Thomas. *Coleridge and the Pantheist Tradition*. Oxford: Oxford University Press, 1969.
- Michel, Paul Henri. *The Cosmology of Giordano Bruno*. Translated by R.E.W. Maddison. Paris: Hermann; London: Methuen; Ithaca, New York: Cornell, 1962.
- Min, Anselm K. "Hegel's Absolute: Transcendent or Immanent?" *Journal of Religion*, 56 (1976), pp. 61-87.
- Naess, Arne. "Environmental Ethics and Spinoza's Ethics." *Inquiry*, 23 (1980) pp. 313-325.
- Naess, Arne. "The Shallow and the Deep, Longrange Ecology Movement." *Inquiry*, 16 (1973), pp. 95-100.
- Naess, Arne. "Identification as Source of Deep Ecological Attitudes." In *Deep Ecology*. Ed. M. Tobias. San Diego: Avant Books, 1983.
- Naess, Arne. "The Deep Ecological Movement: Some Philosophical Aspects," *Philosophical Inquiry*, 8 (1986), pp. 10-29.
- Naess, Arne. "Spinoza and Ecology." *Philosophia*, 7 (1977), pp. 45-54.
- Oakes, Robert. "Does Traditional Theism Entail Pantheism?" *American Philosophical Quarterly*, 20 (1983), pp. 105-112. Reprinted in *The Concept of God*. Ed. Thomas V. Morris. Oxford: Oxford University Press, 1987.
- Oakes, Robert. "Theism and Pantheism Again," *Sophia*, 24 (1985), pp. 3237.
- Oakes, Robert. "Classical Theism and Pantheism: A Victory for Process Theism?" *Religious Studies*, 13 (1977), pp. 167-173.
- O'Connor, D.J. "Substance and Attribute." In *Encyclopedia of Philosophy*. Ed. Paul Edwards. New York: Macmillan and Free Press, 1967.
- Otto, Rudolf. *The Idea of the Holy*. Second edition. Oxford: Oxford University Press, 1950.
- Owen, H. P. *Concepts of Deity*. London: Macmillan, 1971.
- Parkinson, G.H.R. "Hegel, Pantheism, And Spinoza." *Journal of the History of Ideas*, 38 (1977), pp. 449-459.
- Quinn, Philip. "Divine Conservation and Spinozistic Pantheism." *Religious Studies*, 15 (1979), pp. 289-302.
- Rees, D.A. "Greek Views of Nature and Mind." *Philosophy*, 29 (1954), pp. 99-111.
- Rensch, Bernhard. "Panpsychistic Identism and its Meaning for a Universal Evolutionary Picture." *Scientia*, 112 (1977), pp. 337-347.
- Rist, J. *Plotinus: The Road to Reality*. Cambridge: Cambridge University Press, 1967.
- Rowe, Christopher. "One and Many in Greek Religion." In *Oneness and Variety*. Ed. Adolf Portman and Rudolf Ritsema. Leiden: E.J. Brill, 1980.
- Russell, Paul. "Epigram, Pantheists, and Free Thought in Hume's 'Treatise': A Study in Esoteric Communication." *Journal of the History of Ideas*, 54 (1993), pp. 659-673.
- Schopenahuer, Arthur. "A Few Words On Pantheism." In *Essays from the Parerga and*

Paralipomena. Translated by T. Bailey Saunders. London: George Allen & Unwin, 1951.

- Sessions, George. "Anthropocentrism and the Environmental Crisis." *Humbolt Journal of Social Relations*, 2 (1974), pp. 71-81.
- Sessions, George. "Spinoza and Jeffers on Man in Nature." *Inquiry*, 20 (1977), pp. 481-528.
- Siwek, Paul. "How Pantheism Resolves the Enigma of Evil." *Laval Theologique et Philosophique*, 1112 (195556), pp. 213-221.
- Smart, Ninian. "Myth and Transcendence." *Monist*, 50 (1966), pp. 475-487.
- Smart, Ninian. *The Concept of Worship*. London: Macmillan, 1972.
- Smart, Ninian. "God's Body." *Union Seminary Quarterly Review*, 37 (1981-2), pp. 51-59.
- Smart, Ninian. "Our Experience of the Ultimate." *Religious Studies*, 20 (1984), pp. 19-26.
- Smith, J.A. "The Issue Between Monism and Pluralism." *Proceedings of the Aristotelian Society*, 26 (192526), pp. 124.
- Spinoza, Baruch. *Ethics*. Ed. James Gutmann. New York: Hafner, 1949.
- Spinoza, Baruch. *The Collected Works of Spinoza*. Vol. 1. Edwin Curly, translator and editor. Princeton: Princeton University Press, 1985.
- Stokes, Michael C. *One and Many in Presocratic Philosophy*. Washington D.C.: Center For Hellenic Studies, 1971.
- Taylor, Paul. *Principles of Ethics*. Encino, Calif.: Dickenson Publishers, 1975.
- Tapper, Alan. *Priestley's Metaphysics*. Ph.D. Dissertation, University of Western Australia, 1987.
- Tillich, Paul. *Systematic Theology Vols. I-III*. Chicago and London: University of Chicago Press and SCM, 1963.
- Toland, John. *Pantheisticon*. New York and London: Garland, 1976. Reprint of 1751 ed.
- Vlastos, Gregory. "Theology and Philosophy in Greek Thought." *Philosophical Quarterly*, 2 (1952).
- Wainwright, William. "God's Body." *Journal of the American Academy of Religion*, 42 (1974), pp. 470-481.
- Whitman, Walt. *Leaves of Grass*. The First (1855) Edition. Edited, with an Introduction by Malcolm Cowley. New York: Viking Press, 1959.
- Wolfson, Abraham. *Spinoza: A Life of Reason*. New York, 1932.
- Wood, Harold, W. Jr. "Modern Pantheism as an Approach to Environmental Ethics." *Environmental Ethics*, 7 (1985), pp. 151-163.

Other Internet Resources

- Einstein's Pantheism

Related Entries

monism | Spinoza, Baruch [Benedict] | theism

Michael P. Levine
mlevine@arts.uwa.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

A to Z Table of Contents

First published: June 4, 1996
Content last modified: August 14, 1997

Benjamin Peirce

Benjamin Peirce (b. April 4, 1809, d. October 6, 1880) was a professor at Harvard with interests in celestial mechanics, applications of plane and spherical trigonometry to navigation, number theory and algebra. In mechanics, he helped to establish the (effects of the) orbit of Neptune (in relation to Uranus). In number theory, he proved that there is no odd perfect number with fewer than four distinct prime factors. In algebra, he published a comprehensive book on complex associative algebras. Peirce is also of interest to philosophers because of his remarks about the nature and necessity of mathematics.

1. Career

Born in 1809, Peirce became a major figure in mathematics and the physical sciences during a period when the U.S. was still a minor country in these areas (Hogan 1991). A student at Harvard College, he was appointed tutor there in 1829. Two years later he became Professor of Mathematics in the University, a post which was changed in 1842 to cover astronomy also; he held it until his death in 1880. He played a prominent role in the development of the science curriculum of the university, and also acted as College librarian for a time. However, he was not a successful teacher, being impatient with students lacking strong gifts; but he wrote some introductory textbooks in mathematics, and also a more advanced one in mechanics (Peirce 1855). Among his other appointments, the most important one was Director of the U.S. Coast Survey from 1867 to 1874. Peirce also exercised influence through his children. By far the most prominent was Charles Sanders Peirce (1839-1914), who became a remarkable though maverick polymath, as mathematician, chemist, logician, historian, and many other activities. In addition, James Mills (1834-1906) became in turn professor of mathematics at Harvard, Benjamin Mills (1844-1870) a mining engineer, and Herbert Henry Davis (1849-1916) a diplomat. However, Harvard professor

Benjamin Osgood Peirce (1854-1914), mathematician and physicist, was not a relative. Benjamin Peirce did not think of himself as a philosopher in any academic sense, yet his work manifests interests of this kind, in two different ways. The first was related to his teaching.

2. Mathematics, mechanics and God

To a degree unusually explicit in a mathematician of that time Peirce affirmed his Christianity, seeing mathematics as study of God's work by God's creatures. He rarely committed such sentiments to print; but a short passage occurs in the textbook on mechanics previously mentioned, when considering the idea that the occurrence of perpetual motion in nature

> would have proved destructive to human belief, in the spiritual origin of force and the necessity of a First Cause superior to matter, and would have subjected the grand plans of Divine benevolence to the will and caprice of man (Peirce 1855, 31).

Peirce was more direct in a course of Lowell Lectures on ‘Ideality in the physical sciences’ delivered at Harvard in 1879, which James Peirce edited for posthumous publication (Peirce 1881b). ‘Ideality’ connoted ‘ideal-ism’ as evident in certain knowledge, ‘pre-eminently the foundation of the mathematics’. His detailed account concentrated almost entirely upon cosmology and cosmogony with some geology (Petersen 1955). He did not argue for his stance beyond some claims for existence by design.

3. Algebras and their philosophy

Peirce was primarily an algebraist in his mathematical style; for example, he was enthusiastic for the cause of quaternions in mechanics after their introduction by W. R. Hamilton in the mid 1840s, and of the various traditions in mechanics he showed some favour for the ‘analytical’ approach, where this adjective refers to the links to algebra. His best remembered publication was a treatment of ‘linear associative algebras’, that is, all algebras in which the associative law $x(yz)=(xy)z$ was upheld. ‘Linear’ did not carry the connotation of matrix theory, which was still being born in others' hands, but referred to the form of linear combination, such as:

$$q = a + bi + cj + dk$$

in the case of a quaternion q. Peirce wrote an extensive survey (Peirce 1870), determining the numbers of all algebras with from two to six elements obeying also various other laws (Walsh 2000, ch. 2). To two of those he gave names which have become durable: ‘idempotent’, the law $x^m = x$ (for $m \geq 2$) which George Boole had introduced in this form in his algebra of logic in 1847; and ‘nilpotent’, when $x^m = 0$, for some m. The history of the publication of this work is very unusual (Grattan-Guinness 1997). Peirce had presented some of his results from 1867 onwards to the National Academy of Sciences, of which he had been appointed a founder member four years earlier; but they could not afford to print it. Thus, in an

initiative taken by Coast Survey staff, a lady without mathematical training but possessing a fine hand was found who could both read his ghastly script and write out the entire text 12 pages at a time on lithograph stones. 100 copies were printed (Peirce 1870), and distributed world-wide to major mathematicians and professional colleagues. Eleven years later Charles, then at Johns Hopkins University, had the lithograph reprinted posthumously, with some additional notes of his own, as a long paper in American journal of mathematics, which J.J. Sylvester had recently launched (Peirce 1881a); it also came out in book form in the next year. This study helped mathematicians to recognise an aspect of the wide variety of algebras which could be examined; it also played a role in the development of model theory in the U.S. in the early 1900s. Enough work on it had been done by then for a book-length study to be written (Shaw 1907).

4. The philosophy of necessity

Peirce seems to have upheld his theological stance for all mathematics, and a little sign is evident in the dedication at its head:

> To my friends This work has been the pleasantest mathematical effort of my life. In no other have I seemed to myself to have received so full a reward for my mental labor in the novelty and breadth of the results. I presume that to the uninitiated the formulae will appear cold and cheerless. But let it be remembered that, like other mathematical formulae, they find their origin in the divine source of all geometry. Whether I shall have the satisfaction of taking part in their exposition, or whether that will remain for some more profound expositer, will be seen in the future (Peirce 1870, 1).

Peirce began with a philosophical statement of a different kind about mathematics which has become his best remembered single statement "Mathematics is the science that draws necessary conclusions" (Peirce 1870, p. 1). What does ‘necessary’ denote? Perhaps he was following a tradition in algebra, upheld especially by Britons such as George Peacock and Augustus De Morgan (a recipient of the lithograph), of distinguishing the ‘form’ of an algebra from its ‘matter’ (that is, an interpretation or application to a given mathematical and/or physical situation) and claiming that its form alone would deliver the consequences from the premises. In his first draft of his text he wrote the rather more comprehensible "Mathematics is the science that draws inferences", and in the second draft "Mathematics is the science that draws consequences", though the last word was altered to yield the enigmatic form involving ‘necessary’ used in the book. The change is not just verbal; he must have realised that the earlier forms were not sufficient (they are satisfied by other sciences, for example), and so added the crucial adjective. Certainly no whiff of modal logic was in his air. His statement appears in the mathematical literature fairly often, but usually without explanation. One feature is clear, but often is not stressed. In all versions Peirce always used the active verb ‘draws’: mathematics was concerned with the act of drawing conclusions, not with the theory of so acting, which belonged in disciplines such as logic. He continued:

> Mathematics, as here defined, belongs to every enquiry; moral as well as physical. Even the rules of logic, by which it is rigidly bound could not be deduced without its aid (Peirce

> 1870, 3).

In a lecture of the late 1870s he described his definition as

> wider than the ordinary definitions. It is subjective; they are objective. This will include knowledge in all lines of research. Under this definition mathematics applies to every mode of enquiry (Peirce 1880, 377).

Thus Peirce maintained the position asserted by Boole that mathematics could be used to analyse logic, not the vice versa relationship between the two disciplines that Gottlob Frege was about to put forward for arithmetic, and which Bertrand Russell was optimistically to claim for *all* mathematics during the 1900s. Curiously, the third draft of the lithograph contains this contrary stance in "Mathematics, as here defined, belongs to every enquiry; it is even a portion of deductive logic, to the laws of which it is rigidly subject"; but by completion he had changed his mind. Peirce's son Charles claimed to have influenced his father in forming his definitive position, and fiercely upheld it himself; thereby he helped to forge a wide division between the algebraic logic which he was developing from the early 1870s with his father, Boole and de Morgan as chief formative influences, and the logicism (as it became called later) of Frege and Russell and also the ‘mathematical logic’ of Giuseppe Peano and his school in Turin (Grattan-Guinness 1988).

Bibliography

This list includes some valuable items not cited in the text.

Primary Sources

- Peirce Manuscripts: Houghton Library, Harvard University.
- 1855. *Physical and celestial mathematics*, Boston: Little, Brown.
- 1861. *An elementary treatise on plane and spherical trigonometry, with their applications to navigation, surveying, heights, and distances, and spherical astronomy, and particularly adapted to explaining the construction of Bowditch's navigator, and the nautical almanac*, rev. ed., Boston: J. Munroe.
- 1870. *Linear associative algebra*, Washington (lithograph).
- 1880. ‘The impossible in mathematics’, in Mrs. J. T. Sargent (ed.), *Sketches and reminiscences of the Radical Club of Chestnut St. Boston*, Boston : James R. Osgood, 376-379.
- 1881a. ‘Linear associative algebra’, *Amer. j. math.*, 4, 97-215. Also (C.S. Peirce, ed.)in book form, New York, 1882. [Printed version of Peirce 1870.]
- 1881b. *Ideality in the physical sciences*, (J. M. Peirce, ed.), Boston: Little, Brown.
- 1980. Benjamin Peirce: "Father of Pure Mathematics" in America, (I. Bernard Cohen, ed.), New York: Arno Press. [Photoreprints, including that of (Peirce 1881a).]

Secondary Sources

- Archibald, R.C. 1925. [ed.], ‘Benjamin Peirce’, *American mathematical monthly*, 32, 1-30; repr. Oberlin, Ohio.: Mathematical Association of America.
- Archibald, R.C. 1927. ‘Benjamin Peirce's linear associative algebra and C.S. Peirce’, *American mathematical monthly*, 34, 525-527.
- Grattan-Guinness, I. 1988. ‘Living together and living apart: on the interactions between mathematics and logics from the French Revolution to the First World War’, *South African journal of philosophy*, 7, no. 2, 73-82.
- Grattan-Guinness, I. 1997. ‘Benjamin Peirce's Linear associative algebra (1870): new light on its preparation and "publication"’, *Annals of science*, 54, 597-606.
- Hogan, E. 1991. ‘ "A proper spirit is abroad": Peirce, Sylvester, Ward, and American mathematics’, *Historia mathematica*, 18, 158-172.
- King, M. 1881. (Ed.), *Benjamin Peirce. A memorial collection*, Cambridge, Mass.: Rand, Avery. [Obituaries.]
- Novy, L. ‘Benjamin Peirce's concept of linear algebra’, *Acta historiae rerum naturalium necnon technicarum* (Special Issue) 7 (1974), 211-230.
- Peterson, S. R. 1955. ‘Benjamin Peirce: mathematician and philosopher’, *Journal of the history of ideas*, 16, 89-112.
- Pycior, H. 1979. ‘Benjamin Peirce's linear associative algebra’, *Isis*, 70, 537-551.
- Schlote, K.-H. 1983. ‘Zur Geschichte der Algebrentheorie in Peirces "Linear Associative Algebra"’, *Schriftenreihe der Geschichte der Naturwissenschaften, Technik und Medizin*, 20, no. 1, 1-20.
- Shaw, J. B. 1907. *Synopsis of linear associative algebra. A report on its natural development and results reached to the present time*, Washington.
- Walsh, A. 2000. ‘Relationships between logic and mathematics in the works of Benjamin and Charles S. Peirce’, Ph. D. thesis, Middlesex University.

Other Internet Resources

- The MacTutor History of Mathematics Archive entry on Peirce
- Photos of Peirce at the MacTutor Archive

Related Entries

Peirce, Charles Sanders

I. Grattan-Guinness
Middlesex University at Enfield

ivor2@mdx.ac.uk
and
Alison Walsh
Cambridge Regional College
awalsh@crc.tcom.co.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 3, 2001
Content last modified: February 3, 2001

Stanford Encyclopedia of Philosophy

Historicist Theories of Rationality

Of those philosophers who have attempted to characterize scientific rationality, most have attended in some way to the history of science. Even Karl Popper, who is hardly a historicist by anyone's standards, frequently employs the history of science as an illustrative and polemical device. However, relatively few theorists have offered theories according to which data drawn from the history of science somehow *constitute* or *are evidential for* the concept of rationality. Let us call such theories historicist theories.

Roughly put, the idea behind historicist theories of rationality is that a good theory of rationality should somehow fit the history of science. According to a minimal reading of "fit", a good theory of rationality will label as rational most of the major episodes in the history of science. A more demanding reading asserts that the best theory of rationality is the one that maximizes the number of rational episodes in the history of science (subject to some filtering out of some sociologically infected episodes). However, it is unclear whether (i)historicism is a conceptual claim according to which it is an analytic or at least necessary truth that rationality fit history, or (ii)whether historicism is an epistemological claim according to which the best way to find out about rationality is to consult the history of science. Historicism (i) seems unmotivated, while historicism (ii) might descend into triviality. For instance, in the case of instrumental rules which tell us the best way to achieve certain goals, philosophers of all stripes would say that looking at historical attempts to achieve those goals will help us evaluate our current proposals for achieving them. Another ambiguity in historicism concerns its scope. Does historicism become a good idea only once one has established that science is basically successful, or should historicism be endorsed within every scientific community and possible world?

In order to understand historicism, one must also understand the distinction between methodology and meta-methodology. In the parlance of the history and philosophy of science, a methodology for scientific rationality is a theory of rationality: it tells us what is rational and what is not in specific cases. Thus, the rule "Always accept the theory with the greatest degree of confirmation" would count as (part of) a methodology. On the other hand, a meta-methodology provides us with the standards by which be evaluate the theories of rationality which constitute our methodologies. To be a historicist about rationality is to accept a meta-methodological claim: a good theory of rationality must fit the history of science. Thus although historicists might agree on a general meta-methodology, they can and do vary widely in the sort of theory that they produce using that meta- methodology. To acquire a feel for the general approach, let us first review the work of the three major historicists, Thomas Kuhn, Imre Lakatos and Larry Laudan.

- Kuhn

Kuhn

Historicism in the philosophy of science is a fairly recent development. It can perhaps be dated to the publication of Kuhn's influential *The Structure of Scientific Revolutions* in 1962, Before that point, the two dominant theories of rationality were confirmationism (scientists should accept theories that are probably true, given the evidence) and falsificationism (scientists should reject theories that make false predictions about observables and replace them with theories that conform to all available evidence). Each of these theories springs from purely logical roots, confirmationism from Carnap's work on inductive logic, and falsificationism from Popper's rejection of inductive logic coupled with his assertion that universals can be falsified by a single counter-instance. Neither of these theories was accountable to the history of science in the following important sense:. If it turned out that the history of science exemplified few or no decisions in accordance with, say, Carnap's confirmationism, then so much the worse for the history of science. Such a discovery would merely show that scientists were largely irrational. It would not challenge confirmationism. Rather, confirmationism was mainly challenged on conceptual, ahistorical grounds, such as its inability to generate plausible yet non-arbitrary levels of confirmation for moderately sized samples, the difficulties encountered in devising a suitable criterion for evidential relevance, and so on.

Kuhn's work effected three major transformations in the study of scientific rationality. First, and most importantly, it brought history to the fore. The implicit (if not explicit) message of *The Structure of Scientific Revolutions* is that a respectable theory of *rational* scientific procedure must conform to the greater part of *actual* scientific procedure. Second, instead of focussing on the theory as the unit of rational exchange, *Structure* was based on a unit that could persist through minor theoretical changes. In so doing, it could distinguish between revisions and wholesale rejection. Kuhn called this unit "the paradigm", and its ancestors live on as the research programme, the research tradition, the global theoretical unit, and so on. Third, Kuhn's work highlighted the real problems that historically aware accounts of rationality face: when all is said and done, there may be no trans-historical rule for rational scientific procedure.

Before dealing with specifics, it is important to stress that one should approach the exegetical details of Kuhn's account with great trepidation. Many different interpretations exist, perhaps because the internal consistency of Kuhn's positions still stands in some doubt. Thus what follows is no doubt too quick and

would be regarded by some as fundamentally inaccurate. Those who are especially interested in Kuhnian hermeneutics are encouraged to consult Hoyningen-Huene (1993). As an example of the exegetical problems which confront the reader at every turn, there is considerable disagreement over the proper interpretation of the word "paradigm", At one extreme is a very narrow interpretation according to which a paradigm consists of a set of exemplars, where an exemplar can be a famous solved problem, a textbook, or a famous experiment. At the other extreme a paradigm consists of an entire theoretical worldview, represented by an ontology, a set of laws, a list of methodological prescriptions, and a set of fundamental values for science. According to a third reading, which is orthogonal to the others, a paradigm is a fundamentally sociological entity , individuated and constituted by patterns of education and alliances. To maximize the sense of the following account, it is best to choose a fairly wide sense of the term.

According to Kuhn, scientific practice is divided into two phases, called *normal* science and *revolutionary* science. During normal science, the dominant paradigm is neither questioned nor seriously tested. Rather, the members of the scientific community employ the paradigm as a tool for solving outstanding problems. Occasionally, the community will encounter especially resistant problems, or anomalies, but a small number of these problems will occasion little anxiety. Only as the anomalies accumulate will the community pass into a state of crisis, which may push it into the phase of revolutionary science.

During a period of revolutionary science, the scientific community actively debates the underlying principles of the dominant paradigm and its rivals. Thus, the business- as-usual of routine problem solving is suspended until a new paradigm (or perhaps the old one) establishes dominance. The way in which dominance is established is perhaps the most important locus of disagreement concerning Kuhn's work. The most influential interpretation (one which Kuhn has spent much time disavowing) paints Kuhn as an arationalist. This interpretation makes much of Kuhn's use of the theory-ladenness of observation and various sorts of incommensurability. The supposed result of these features is that the proponents of different paradigms will often be unable to communicate with each other, and that even when they can communicate, their standards of assessment will always favor their own paradigms. Thus, there is no rational basis for choosing between paradigms: the switch from one worldview to another is not so much a reasoned matter as the scientific equivalent of a perceptual gestalt shift. On this view, the transition between paradigms is best explained sociologically, in terms of institutional might, polemics and perhaps generational replacement.

The previous argument requires very strong, possibly unrealistically strong senses of incommensurability and theory- ladenness. The more moderate view of revolutionary science does not presuppose that paradigms are separated by unbridgeable linguistic gaps. However, it does retain incommensurability about values. On this view, scientists are trained to reach a rational consensus *in the absence of rules for doing so*. Rationality can no longer be procedurally flow-charted. This interpretation of Kuhn is often coupled with the claim that science has progressed in light of its increasing ability to solve problems. Again, though, there is an important qualification: while we can claim that, e.g., the Newtonian paradigm solved more problems than the Aristotelian one, we cannot claim that Aristotelian set of solved problems is included in the Newtonian one. The transitions from one paradigm to another involves some losses as

well as gains, but on balance, there is a net gain in problem solving ability.

Although this interpretation of Kuhn paints him as a rationalist, it posits a form of rationalism that rejects two claims that many rationalists had thought essential to their enterprise: (i)that rationality is a rule-governed process, and (ii)that scientific progress is cumulative. The reasons for these two claims are not so much historical as conceptual. For instance, if we suppose that the choice between paradigms is made in the absence of rules, and that we should trust it as rational simply because the people making the choices are properly trained, then might we not wonder whether a purely sociological explanation is in order? Similarly, if one paradigm is held to solve more problems than another, even in the presence of important problems solved by the second paradigm but not by the first, then might we not wonder whether the apparent progress is no more than a case of history re-written by the victors. What solid philosophical grounds are there for holding that the gains achieved by the victorious paradigm outweigh the losses? Among others, Brown (1985) addresses the first worry and Laudan (1977) the second (as will be discussed later in this entry), but, to date there has been no satisfactory answer to any of these questions. Thus, Kuhn the rationalist seems to stand on shaky conceptual ground.

Specific worries aside, Kuhn is unsatisfactory for our purposes because he provides us with neither a specific account of rationality nor an explicit account of historicist meta- methodology. Because they are specific where Kuhn is not, Kuhn's main successors, Imre Lakatos and Larry Laudan deserve our special attention.

Lakatos

Lakatos's theory of rationality is based on the idea of the research programme, which is a sequence of theories characterized by a *hard-core* (the features of the theories that are essential for membership in the research programme), the *protective belt* (the features that may be altered), the *negative heuristic* (an injunction not to change the hard core), and the *positive heuristic* (a plan for modifying the protective belt). The protective belt is altered for two reasons. In its early stages, a research programme will make unrealistic assumptions (i.e. Newton's early assumption that the sun and the earth are point masses). The protective belt is altered in order to make the programme more realistic. It becomes testable only when it has achieved a sufficient degree of realism. Once the programme has reached the phase of testability, the protective belt is altered when the programme makes false experimental predictions.

However, not all alterations to the protective belt are equal. If an alteration not only fixes the problem at hand but also allows the research programme to make a *novel prediction*, then the alteration is said to be progressive. If the alteration is no more than an ad hoc manoeuvre, that is, if it does not lead to any novel predictions, then it is regarded as *degenerate*. Initially Lakatos classifies a prediction as novel if and only if the phenomenon being predicted has never been observed prior to the prediction. Later Lakatos (Lakatos and Zahar, 1976) extends the definition to cover phenomena that may have been observed before the time of prediction but which were not among the problems which the alteration was designed to solve.

A research programme is in good health as long as a sufficient number of the alterations to it are progressive. Its troubles multiply to the extent that these alterations are degenerate. Once a research programme is sufficiently degenerate, and once there is a progressive research programme to take its place, the degenerate programme should be jettisoned. However, Lakatos does not provide us with details concerning ways to measure degeneracy, nor does he locate the point at which degeneracy can prove fatal to a research programme.

Lakatos's meta-methodology is interesting precisely because it matches his methodology: a meta-methodological research programme in the philosophy of science is progressive as long as it continues to make novel predictions. This may seem puzzling. What predictions can a theory of rationality make? Lakatos's answer is that the predictions concern basic value judgments made by scientists at the time concerning the rationality and irrationality of certain episodes. To see this, suppose that, according to Lakatos's theory, a certain research programme of the past became unacceptably degenerate at a certain time. Subsequent historical investigation might uncover documents which attest to the attitudes of the scientific community at the time. Suppose that these documents show that the community was preparing to reject the research programme in question. In this case, we would say that Lakatos's theory had made a successful novel prediction.

One could easily question the weight that Lakatos places on novel predictions, both at the methodological and meta-methodological levels. Lakatos faces the following dilemma. The attainment of novel predictions is either valuable in and of itself or it is valuable as a way of achieving some other goal. If Lakatos claims that novel predictions are especially valuable in and of themselves then he faces a quite justified charge of arbitrariness. If he says that they are valuable for their use in achieving other goals then he must say what those goals are and demonstrate that novel predictions are expecially useful in achieving them. For instance, suppose Lakatos were to say that the pursuit of novel predictions provides us with the best and fastest way of increasing the observable content of our theories. Were he to say this he would need to provide us with a viable notion of and metric for observable content. In particular he would have to tell us what it is for one theory to have more observable content than another. If he presupposes some sort of cumulativity principle (i.e. that the better theory says everything true about observables that the worse one did plus a little bit more) than his theory is historically implausible. If he denies cumulativity, then the problem he faces, i.e. that of providing a sound basis for observational content, has foiled all who have tried to solve it.

Laudan

In *Progress and Its Problems,* Larry Laudan presents both an explicit meta-methodology and a normative theory of rationality. According to his meta-methodology, a successful theory of rationality should respect "our preferred pre-analytic intuitions about scientific rationality". (Laudan 1977 160) These intuitions consist of judgments of the rationality of certain explicit cases, (e.g., "it was rational to accept Newtonian mechanics and to reject Aristotelian mechanics by, say, 1800", and "it was irrational after 1830 to accept the biblical chronology as a literal account of earth history"). Thus, although not every episode in the history of science is represented in Laudan's meta-methodology, a subset of it is, where

this subset consists of the "obvious" cases.

The theory of rationality supposedly generated by Laudan's methodology is centered on the notion of the research tradition. Laudan's research traditions somewhat resemble Kuhn's paradigms and Lakatos's research programmes. Like Kuhn's paradigms (on the wider reading of the term) research traditions contain both metaphysical and methodological elements. However, Laudan downplays the sociological and pedagogical elements (e.g. training networks and exemplars) that are so important to Kuhn. Like Lakatos's paradigms, the theories generated by a research tradition will change through time, but, where Lakatos's research programmes are defined as a sequence of theories, the theories do not themselves constitute the research tradition. Laudan also claims that the research tradition is a much less rigid concept than the Lakatosian research programme, which is based on an inflexible hard-core.

However, Laudan radically differentiates himself from Kuhn and Lakatos in his accounts of scientific progress and rationality. He claims that there are two sorts of problems that face every research tradition, empirical problems (akin to Kuhnian anomalies) and conceptual problems (i.e. problems of consistency, either internal or with dominant traditions in other fields). We should *accept* the research tradition that has solved the most problems and *pursue* the tradition that is currently solving problems at the greatest rate. Science progresses by solving more problems. However, Laudan does not presume cumulativity: although a given current research tradition will have solved more problems than its predecessors, there may be particular problems that have become "unsolved" by the current tradition. Thus, unlike Kuhn, Laudan believes that there is a simple concept which serves as a basis for both progress and rationality. Unlike Lakatos, Laudan (i)rejects both the idea of empirical content and the cumulative growth of theories and (ii)places no extra value on the concept of a novel prediction, and no great disvalue on ad hocness.

As appealing as it may seem, Laudan's theory of rationality faces some potentially fatal criticisms. First, how do we determine which research tradition has solved the most problems. Is the "problem of the planets" to be counted as one problem or nine? There is no reason to believe that the enumeration and/or weighting of problems is not relative to research tradition. Laudan offers us no grounds for faith here, since he wants his theory to be robust even in the presence of a strong degree of incommensurability. Without a common scheme of enumeration and/or weighting Laudan's theory may lead to ambiguous results, according to which the rational tradition to pursue depends on who is doing the counting. Second, although Laudan takes some pains in differentiating research traditions from paradigms and research programmes, the notion of a research tradition is still somewhat fuzzy. As with paradigms and programmes, the fuzziness is especially apparent at the level of historical application.

An independent set of problems concerns Laudan's meta-methodology and its link to his theory of rationality. First, since Laudan takes his theory of rationality to apply to all spheres of intellectual endeavor, including the philosophy of science, we should expect his meta-methodology (i.e., his criterion regulating the rational choice of a theory of scientific rationality) to be identical with his theory of rationality. Yet the two are very different. His meta-methodology is a foundational fit-the-data affair, while the ground-level criterion rejects the existence of data in he foundationalist sense. Now, Laudan could retract the claim that his theory of rationality has applicability outside of science, but as we shall

see later, that would lead him into serious problems. Second, Laudan's list of 7 pre-analytic intuitions is fairly uncontroversial. But, we may ask why we believe it to be uncontroversial. Is it because we have all been socially conditioned in the same way.? Or, is it because we have a prior criterion of rationality, according to which we judge the "intuitive" cases? If the former, then there is no reason to privilege our pre-analytic intuitions. If the latter, then, rather than consulting the history of science, we should merely try to explicate our prior criterion. Either way, an approach based on intuitions faces severe difficulties. Third, even if we could provide a firm philosophical basis for this approach, we would have very little data to go on. Laudan cites only seven data points. Without doubt, many theories of rationality, some plausible and some not, would fit these data points. For instance, consider the following criterion:

> An episode in the history of science is rational if and only if it is one of the following episodes: {here follows the list of paradigmatically rational episodes}; and an episode in the history of science is irrational if and only if it is one of the following episodes: {here follows the list of paradigmatically irrational episodes}. All other episodes are neither rational nor irrational.

Clearly this is a silly criterion, but it meets Laudan's meta- methodological constraints. Laudan differentiates his methodology from his meta-methodology to avoid circular and/or self- supporting means of testing a methodology. Circularity is probably not a worry. Laudan probably would do better by equating his methodology with his meta-methodology. At any rate, Laudan himself has disavowed historicism (e.g., Laudan 1986), although for somewhat suspect reasons.

General Criticisms

Specifics aside, there are a number of important issues that have yet to be satisfactorily addressed by historicist theories of rationality. (1) What precisely is a historicist theory of rationality supposed to accomplish? According to Lakatos, one is rational as long as one avoids ad hocness as much as possible; according to Laudan, one is rational as long as one accepts the research tradition that has solved the most problems and pursues the one that is solving them at the greatest rate. Yet neither writer stipulates that rational agents must have the avoidance of ad hocness or the maximization of solved problems in mind as they go about their scientific business. As long as their theoretical behavior is in accord with the Lakatosian/Laudanian dicta, they are rational, regardless of their conscious motivations.

Let us call theories of rationality which evaluate agents on the basis of their theoretical choices and not on the basis of the reasons for the choices *externalist* theories. Externalist theories are wider than internalist (motive based) ones in an important way: the right choice made for the wrong reasons is rational according to externalism. Since Lakatos and Laudan want their theories of rationality to cover most of the history of science, and since the conscious motivations of scientists do seem to have changed over time, it seems that these writers are locked into externalism.

However, upon further examination externalist theories of rationality are very puzzling. Let us compare them with another form of epistemic externalism, an externalist theory of perception. According to such

theories, whether one is justified depends only on whether one's perceptual belief was produced by a reliable mechanism or process. One need not be conscious of a description or justification of that process. Now, in the perceptual case, we have a general idea of the nature of the process and every reason to trust in its reliability (dreamer arguments aside). The problem with externalist theories of rationality, on the other hand, is that we have no idea of the mechanism that would make a scientist, act in such a way that she minimized ad hocness even though her actual intentions were directed towards some other cognitive goal. Where externalist theories of perception depend on tangible information provided by the psychology of perception, externalist theories of rationality depend on a very mysterious invisible hand. Until the workings of that hand are made visible, we should be very suspicious of externalist theories of rationality.

(2)Historicist theories of rationality are also much more difficult to *apply* than their proponents let on. Because the historicist unit of exchange (the paradigm, research programme, research tradition) has much looser conditions of individuation than the single theory, the question of how to group theories into their respective paradigms, etc. can be a difficult one. For instance, Copernicus's theory shared much of Aristotle's physics, Aristotle's commitment to spherical motion and his use of aethereal spheres, Kepler's geo-centrism (almost) and Ptolemy's use of epicycles. In grouping Copernicus with Kepler and Newton, we say that his geocentrism is more important than his beliefs about the way in which things in the heaven moved. There may be reasons for deciding upon this grouping, but the choice is not an automatic one. Since the actual scientists at the time did not think in terms of paradigms, etc., we are not going to be able to make the choice on the basis of historical information. More needs to be said about the standards for such decisions.

A related problem concerns the notion of the *acceptance* of a paradigm, research programme or research tradition. Does the acceptance of a programme involve the literal belief in its truth by every single person in the scientific community? Does it require a general belief in its usefulness? These questions have practical correlates. Was the Copernican system accepted by the time that most astronomer used the Copernican tables, despite their explicit allegiance to an Aristotelian/Ptolemaic cosmology? When it was widely taught in universities? Is quantum mechanics now accepted despite the fact that very few physicists think that it can paint an accurate picture of micro-reality? The question of acceptance has two dimensions. The first concerns what it is for a single person to accept a paradigm, etc. The second concerns the weight of individual acceptance required for community acceptance. Since the data for historicist theories consists of matters of acceptance and rejection at the community level, historicists must provide a great deal more information here before their theories can be applied to the historical record.

(3)What *motivates* us to adopt a historicist theory? One possible motivation comes from our faith in science. To reject historicism is "to claim..., that it is entirely possible that all actual scientific practice, past and present, is irrational and ‘unscientific’, which is in turn to accept the (I think) absurd further consequence that scientists might be bad at doing science". (Brown 1989, 98) However, there are several problems with this motivation. First, our faith in the rationality of science may be more an *a posteriori* matter than an a priori. That is, our faith in science is not blind. We have faith in our science because we have seen what it has accomplished: given our evidence from the history of science, it would be absurd

to conclude that science was not rational. Here, we see that the history of science is rational because it meets our (proto) criteria for progress and rationality. However, in other, counterfactual cases, we would not immediately conclude that scientific practice was rational. For instance, it is not true at every possible world that there is a conceptual link between scientific practice and scientific rationality. Thus, on this view, the history of science is illustrative (and not constitutive) of rationality.

Another problem with the faith-in-science motivation that it is much too weak for many forms of historicism. Our faith in science might lead us to believe that science is not completely irrational, or that it is more rational than not. However, some historicist theories (e.g., some readings of Lakatos, Brown (1989)) claim that the best theory of rationality is the one that, subject to certain conditions, *maximizes* the number of rational episodes in the history of science. General faith in science cannot prop up these maximizing theories.

The second motivation for historicism is due to a form of naturalism. If we reject the idea that epistemology is an *a priori* enterprise and accept that it is merely a form of science, as naturalists tend to do, then historicism might seem tempting. Scientific theories succeed insofar as they fit the data. The data for a *scientific* theory of scientific rationality, if it is to be found anywhere, should be drawn from the history of science. Hence historicism. Leave aside the sloppiness of the preceding argument. Even taken on its own terms, it depends on a simplistic view of the role of theoretical concepts within naturalism. Suppose we endorse naturalism. We can consequently treat rationality as a theoretical posit, much like electrons, viruses and the other theoretical posits of science. Those posits are not justified on the basis of a simple fit-the- data approach. It's difficult to see how they could even play such a role. Rather, they are accepted because they are essential parts of our best explanations for relevant phenomena. The test for a theory of rationality, then, should be this: the best theory of rationality is the one that best explains the data in the history and current practice of science. Simple matters of curve-fitting and maximization should be left aside.

However, if we take this explanationist approach to rationality, then most historicist theories seem irrelevant. None of them explain very much at all. Indeed, the Strong Programme (Bloor 1976) in the sociology of knowledge has argued that rationality plays no explanatory role whatsoever. No doubt, the arguments of the Strong Programme are at least slightly overblown, but they do show that once one moves to an explanationist viewpoint, there is no *guaranteed* role for rationality within naturalism. In the end, one might be left with no more than the kernel of instrumental rationality.

Bibliography

- Bloor, D. (1976), *Knowledge and Social Imagery*, London: Routledge & Kegan Paul
- Brown, H., (1988), *Rationality*, London: Routledge
- Brown, J.R. (1989), *The Rational and The Social*, London: Routledge
- Hoyningen-Huene, P. (1993), *Reconstructing Scientific Revolutions: Thomas S. Kuhn's Philosophy of Science* (translated by A. Levine), Chicago: University of Chicago Press
- Kuhn, T.S. (1962), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press

(2nd edition published in 1970)

- Kuhn, T.S. (1977), *The Essential Tension*, Chicago: The University of Chicago Press
- Lakatos, I. (1970), "Falsification and the Methodology of Scientific Research Programmes" in Lakatos and I. Musgrave (ed.) *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press.
- Lakatos, I. and E.G. Zahar (1976), "Why Did Copernicus's Programme Supersede Ptolemy's?", in R. Westman (ed.) *The Copernican Achievement*, Los Angeles: University of California Press.
- Laudan, L. (1977), *Progress and its Problems*, Berkeley: University of California Press

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

rationality | science, philosophy of

Carl Matheson
University of Manitoba
matheso@cc.umanitoba.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 12, 1996
Content last modified: August 12, 1996

The Kochen-Specker Theorem

The Kochen-Specker theorem is an important, but subtle, topic in the foundations of quantum mechanics (QM). The theorem provides a powerful argument against the possibility of interpreting QM in terms of hidden variables (HV). We here present the theorem/argument and the foundational discussion surrounding it at different levels. The reader looking for a quick overview should read the following sections and subsections: 1, 2, 3.1, 3.2, 4, and 6. Those who read the whole entry will find proofs of some non-trivial claims in supplementary documents.

§1: Introduction

QM has the peculiar property that quantum-mechanical states imply, in general, only statistical restrictions on the results of measurements. The natural conclusion to be drawn is that these states are incomplete descriptions of quantum systems. QM, thus, would be incomplete in the sense that a typical

QM state description of an individual system could be supplemented with a more complete description in terms of an HV theory. In an HV description of the system the QM probabilities would be naturally interpreted as epistemic probabilities of the sort that arise in ordinary statistical mechanics. Such an HV description might not be practically useful, but one is tempted to think that it should at least be possible in principle. There are, however, two powerful theorems to the effect that such description is impossible even in principle: QM, given certain extremely plausible premises, cannot be supplemented by an HV theory. The more famous of these two theorems is Bell's theorem which states that, given a premise of locality, an HV model cannot match the statistical predictions of QM. The second important no-go theorem against HV theories is the theorem of Kochen and Specker (KS) which states that, given a premise of noncontextuality (to be explained presently) certain sets of QM observables cannot consistently be assigned values at all (even before the question of their statistical distributions arises).

Before seeing the workings of the KS theorem in some detail, we must clarify why it is of importance to philosophers of science. The explicit premise of HV interpretations is one of *value definiteness*:

> (VD) All observables defined for a QM system have definite values at all times.

VD, however, is motivated by a more basic principle, an apparently innocuous realism about physical measurement which, initially, seems an indispensable tenet of natural science. This realism consists in the assumption that whatever exists in the physical world is causally independent of our measurements which serve to give us information about it. Now, since measurements of all QM observables, typically, yield more or less precise values, there is good reason to think that such values exist independently of any measurements - which leads us to assume VD. (Note that we do not need to assume here that the values are faithfully revealed by measurement, but only that they exist!) We can concretize our innocuous realism in a second assumption of *noncontextuality*:

> (NC) If a QM system possesses a property (value of an observable), then it does so independently of any measurement context, i.e. independently of *how* that value is eventually measured.

This means that if a system possesses a given property, it does so independently of possessing other values pertaining to other arrangements. So, both our assumptions incorporate the basic idea of an independence of physical reality from its being measured.

The KS theorem establishes a contradiction between VD + NC and QM; thus, acceptance of QM logically forces us to renounce either VD or NC. However, the situation is more dramatic than it would initially seem. VD is the key motivating assumption of the HV programme in the sense that, if feasible, it would most naturally explain the statistical character of QM and most elegantly explain away the infamous measurement problem haunting all interpreters of QM [see the entries on quantum mechanics and measurement in quantum theory for details]. But, as we just saw, the second assumption NC is motivated by the same innocuous realism which embodies a standard of scientific rationality, and it is far from obvious what an interpretation obeying this standard only partly, i.e. endorsing only VD but rejecting NC,

should look like. This complex of issues -- namely, (1) VD + NC contradict QM; (2) the conceptual difficulties of interpreting QM provide a strong motivation for VD; (3) it is not obvious how to come up with a plausible story about QM containing VD, but not NC -- is what fuels philosophical interest in the KS theorem.

Return to Table of Contents

§2: Background to the KS Theorem

In the following, we will presuppose some familiarity with elementary QM notions like 'state', 'observable', 'value' and their mathematical representatives 'vector', '(self-adjoint) operator' and 'eigenvalue' [see the entry on quantum mechanics for details]. We will usually identify observables and the operators on an appropriate Hilbert space which represent them; if there is a need to distinguish operators and observables, we write the operators underlined and in boldface. (Thus an operator $\underline{\mathbf{A}}$ represents an observable A.)

The present section states some elements of the historical and systematic background of the KS theorem. Most importantly, an argument by von Neumann (1932), a theorem by Gleason (1957), and a critical discussion of both plus a later argument by Bell (1966) have to be considered. Von Neumann, in his famous 1932 book *Die mathematischen Grundlagen der Quantenmechanik*, disputed the possibility of providing QM with an HV underpinning. He gave an argument which boils down to the following: Consider the mathematical fact that, if $\underline{\mathbf{A}}$ and $\underline{\mathbf{B}}$ are self-adjoint operators, then any linear combination of them (any $\underline{\mathbf{C}} = \alpha\underline{\mathbf{A}} + \beta\underline{\mathbf{B}}$, where α, β are arbitrary real numbers) is also a self-adjoint operator. QM further dictates that for any QM state:

> (1) If A and B (represented by self-adjoint operators $\underline{\mathbf{A}}$ and $\underline{\mathbf{B}}$) are observables on a system, then there also is an observable C (represented by self-adjoint operator $\underline{\mathbf{C}}$ defined as before) on the same system.
>
> (2) If the expectation values of A and B are given by $\langle A \rangle$ and $\langle B \rangle$, then C's expectation value is given by $\langle C \rangle = \alpha\langle A \rangle + \beta\langle B \rangle$.

Now consider A, B, C, as above, and let their values be *v*(A), *v*(B), *v*(C). Consider a 'hidden state' V which determines *v*(A), *v*(B), *v*(C). We can then derive from V trivial 'expectation values' which are just the possessed values themselves: $\langle A \rangle_V = v(A)$, and so on.[1] Of course, these 'expectation values' do not, in general, equal the QM ones: $\langle A \rangle_V \neq \langle A \rangle$ (we would indeed think of the latter as averages over the former for different hidden states V!), but we require that the $\langle A \rangle_V$, like the $\langle A \rangle$, conform to (2). This automatically entails that the values themselves must conform to a condition parallel to (2), i.e.:

> (3) $v(C) = \alpha v(A) + \beta v(B)$.

This, however, is impossible, in general. An example very easily shows how (3) is violated, but because of its simplicity it also shows the argument's inadequacy. (This example is not due to von Neumann himself, but to Bell![2]) Let A = σx and B = σy, then operator C = $(\sigma x + \sigma y)/\sqrt{2}$ corresponds to the observable of the spin component along the direction bisecting *x* and *y*. Now all spin components have (in suitable units) possible values ±1 only, thus, the HV proponent is forced to ascribe ±1 to A, B, C as values, and thus as ‘expectation values’. This, in turn, implies (3) which obviously cannot be fulfilled, since $\pm 1 \neq (\pm 1 + \pm 1)/\sqrt{2}$.

The example illustrates why von Neumann's argument is unsatisfying. Nobody disputes the move from (2) to (3) for compatible observables, i.e. those which, according to QM, are jointly measurable in one arrangement. The above choice of A, B, C, however, is such that any two of them are incompatible, i.e. are not jointly observable. For these we will not want to require any HV interpretation to meet (3), but only (2). The hidden values need not conform with (3) in general, only the averages of their values in a series of tests must conform with (2). The authority of von Neumann's argument comes from the fact that requirements (1) and (2), *for QM states*, are consequences of the QM formalism, but this does not in itself justify extending these requirements to the hypothetical *hidden states*. Indeed, if (3) were unrestrictedly true, this would nicely explain, in the presence of hidden values, why (2) is. Von Neumann apparently thought that the HV proponent is committed to this explanation, but this seems an implausible restriction.

The KS theorem remedies this defect, spotted by Bell in von Neumann's argument, and thus strengthens the case against HV theories insofar as KS assume (3) only for sets of observables {A, B, C} which are all mutually compatible. The theorem requires that only for compatible observables assumption (3) must hold, which is something the HV theorist cannot reasonably deny.

A second, independent line of thought leading to the KS theorem is provided by Gleason's theorem (Gleason 1957). The theorem states that on a Hilbert space of dimension greater than or equal to 3, the only possible probability measures are the measures $\mu(P_\alpha) = Tr(P_\alpha W)$, where P_α is a projection operator, W is the statistical operator characterizing the system's actual state and Tr is the trace operation. The P_α can be understood as representing yes-no observables, i.e. questions concerning whether a QM system represented by a Hilbert space of dimension greater than or equal to 3 has a property α or not, and every possible property α is associated uniquely with a vector $|\alpha\rangle$ in the Hilbert space -- so, the task is to unambiguously assign probabilities to all vectors in the space. Now, the QM measure μ is continuous, so Gleason's theorem in effect proves *that every probability assignment to all the possible properties in a three-dimensional Hilbert space must be continuous*, i.e. must map all vectors in the space continuously into the interval [0, 1]. On the other hand, an HV theory (if characterized by VD + NC) would imply that of every property we can say whether the system has it or not. This yields a trivial probability function which maps all the P_i to either 1 or 0, and, provided that values 1 and 0 both occur (which follows trivially from interpreting the numbers as probabilities), this function must clearly be discontinuous (cf. Redhead 1987: 28).

This is the easiest argument against the possibility of an HV interpretation afforded by Gleason's theorem. Bell (1966: 6-8) offers a variant with a particular twist which later is repeated as the crucial step in the KS

theorem. (This explains why some authors (like Mermin 1990b) call the KS theorem the *Bell*-Kochen-Specker theorem; they think that the decisive idea of the KS theorem is due to Bell.[3]) He proves that the mapping μ dictates that two vectors $|\alpha\rangle$ and $|\alpha'\rangle$ mapped into 1 and 0 cannot be arbitrarily close, but must have a minimal angular separation, while the HV mapping, on the other hand, requires that they must be arbitrarily close.

After having offered his variant of the argument against HV theories from Gleason's theorem, Bell proceeds again to criticise it. The strategy parallels the one directed against von Neumann. Bell points out that his own Gleason-type argument against arbitrary closeness of two opposite-valued points presupposes non-trivial relations between values of non-commuting observables, which are only justified, given an assumption of noncontextuality (NC). He proposes as an analysis of what went wrong that his own argument "tacitly assumed that measurement of an observable must yield the same value independently of what other measurements may be made simultaneously" (1966: 9). In opposition to von Neumann, the Gleason-type argument derives restrictions on value assignments like (3) only for sets of compatible observables; but still one and the same observable can be a member of different commuting sets, and it is essential to the arguments that the observable gets assigned the same value in both sets, i.e. that the value assignment is not sensitive to a measurement context.

The KS theorem improves on the argument from Gleason's theorem. First, the authors repeat, in effect, Bell's proof that two vectors in the Hilbert space having values 1 and 0 cannot be arbitrarily close. However, while the Gleason argument and Bell's variant assume value assignments for a continuum of vectors in the Hilbert space, KS are able to explicitly present a discrete, even finite set of observables in the space for which an HV value assignment would lead to inconsistency. Obviously, the assumptions needed for the step of establishing that two opposite-valued points cannot be arbitrarily close are still in play in KS's improvement -- especially NC is! -- so Bell's criticism of his own Gleason-type argument survives that improvement.

Despite Bell's reasoning, the KS argument is of crucial importance in the HV discussions for two reasons: (1) It involves only a finite set of discrete observables. It thus avoids a possible objection to Bell's Gleason-type argument, namely that "it is not meaningful to assume that there are a continuum number of quantum mechanical propositions [viz. experiments]" (Kochen and Specker 1967: 70/307). So the KS theorem closes a loophole which a HV proponent might spot in Bell's argument. (2) KS propose a one-particle system as a physical realization of their argument. Thus, the argument trivially involves no separability or locality assumptions. Indeed, Bell first pointed out the tacit noncontextuality premise, but did so only in passing, and then, in the final section discussed an example of a two-particle system. Here, an eventual contextuality returns as nonseparability of the two particles, but Bell does not state the connection explicitly. Nor does he point out that the issue about the possibility of HV interpretations is, at bottom, not one about (non)separability or (non)locality, but rather one about (non)contextuality.[4] (After all, Bell's own argument against HV interpretations involves separability and/or locality assumptions!) This fact, however, is clearly illustrated by KS-type arguments.

Return to Table of Contents

§3: Statement and Proof of the KS Theorem

§3.1: Statement of the KS Theorem

An explicit statement of the KS theorem runs thus:

> Let H be a Hilbert space of QM state vectors of dimension $x \geq 3$. Let M be a set containing *y* observables, defined by operators on H. Then, for specific values of *x* and *y*, the following two assumptions are contradictory:
>
> (KS1) All *y* members of M simultaneously have values, i.e. are unambiguously mapped onto real unique numbers (designated, for observables A, B, C, ... by *v*(A), *v*(B), *v*(C), ...).
>
> (KS2) Values of observables conform to the following constraints:
>
> > (a) If A, B, C are all compatible and C = A+B, then *v*(C) = *v*(A)+*v*(B);
> >
> > (b) if A, B, C are all compatible and C = A·B, then *v*(C) = *v*(A)·*v*(B).

Assumption KS1 of the theorem obviously is an equivalent of VD. Assumptions KS2 (a) and (b) are called the *Sum Rule* and the *Product Rule*, respectively, in the literature. (The reader should again note that, in opposition to von Neumann's implicit premise, these rules non-trivially relate the values of *compatible* observables only.) Both are consequences of a deeper principle called the functional composition principle (FUNC), which in turn is a consequence of (among other assumptions) NC. The connection between NC, FUNC, Sum Rule and Product Rule will be made explicit in §4.

In the original KS proof *x*=3 and *y*=117. More recently proofs involving less observables have been given by (among many others) Peres (1991, 1995) for *x*=3 and *y*=33 and by Kernaghan (1994) for *x*=4 and *y*=20. The KS proof is notoriously complex, and we will only sketch it in §3.4. The Peres proof establishes the KS result in full strength, with great simplicity, and, moreover, in an intuitively accessible way, since it operates in three dimensions; we refer the reader to Peres (1995: 197-99). The Kernaghan proof establishes a contradiction in four dimensions. This is a weaker result, of course, than the KS theorem (since every contradiction in 3 dimensions is also a contradiction in higher dimensions, but not conversely). However, the proof is so much simpler that we present it for starters in §3.2. Finally, in §3.5, we explain an argument by Clifton (1993) where *x*=3 and *y*=8 and an additional statistical assumption yields an easy and instructive KS argument.

Return to Table of Contents

§3.2: A Quick KS Argument in Four Dimensions

A particularly easy KS argument proceeds on a four-dimensional Hilbert space H4. In order to get the gist of it quickly, the reader has to accept the following two facts on faith:

(1) From KS2 we can derive a constraint on value assignments to projection operators, namely that for every set of projection operators P_1, P_2, P_3, P_4, corresponding to the four distinct eigenvalues q_1, q_2, q_3, q_4 of an observable Q on H4 the following holds:

$$(\mathrm{VC1}') \quad v(P_1) + v(P_2) + v(P_3) + v(P_4) = 1, \text{ where } v(P_i) = 1 \text{ or } 0, \text{ for } i = 1, 2, 3, 4.$$

((VC1′) is a variant of (VC1) which we prove explicitly in the next section.) This means in effect that of every set of four orthogonal rays in H4 exactly one is assigned the number 1, the others 0.

(2) Although the Hilbert space mentioned in the theorem, in order to be suited for QM, must be *complex*, it is enough, in order to show the inconsistency of claims KS1 and KS2, to consider a *real* Hilbert space of the same dimension. So, instead of H4 we consider a real Hilbert space R4 and translate VC1′ into the requirement: Within every set of orthogonal rays in R4, exactly one is assigned the number 1 and the others 0. As usual in the literature, we translate all this into the following colouring problem: *Within every set of orthogonal rays in R4, exactly one must be coloured white and the others black.* This, however, is impossible, as is shown immediately by the following table (Kernaghan 1994):

1,0,0,0	1,0,0,0	1,0,0,0	1,0,0,0	-1,1,1,1	-1,1,1,1	1,-1,1,1	1,1,-1,1	0,1,-1,0	0,0,1,-1	1,0,1,0
0,1,0,0	0,1,0,0	0,0,1,0	0,0,0,1	1,-1,1,1	1,1,-1,1	1,1,-1,1	1,1,1,-1	1,0,0,-1	1,-1,0,0	0,1,0,1
0,0,1,0	0,0,1,1	0,1,0,1	0,1,1,0	1,1,-1,1	1,0,1,0	0,1,1,0	0,0,1,1	1,1,1,1	1,1,1,1	1,1,-1,-1
0,0,0,1	0,0,1,-1	0,1,0,-1	0,1,-1,0	1,1,1,-1	0,1,0,-1	1,0,0,-1	1,-1,0,0	1,-1,-1,1	1,1,-1,-1	1,-1,-1,1

There are 4 x 11 = 44 entries in this table. These entries are taken from a set of 20 rays (so we allow for repeats). [Recall that to specify a ray or line through the origin in four dimensions, it suffices to give the four coordinates of any single point (apart from the origin) that the line contains. For example, "1,0,0,0" denotes the unique line containing the points with coordinates "0,0,0,0" and "1,0,0,0", which line is, of course, just the "x-axis".] It is easy to verify that every column in the table represents a set of four *orthogonal* rays (simply calculate the dot products between the vectors within each column --- they are always zero). Since the number of columns is 11, we must end up with an *odd* number of the table's entries coloured white. On the other hand, it can be checked that each of the 20 rays appears either twice or four times in the table. So any time we designate a particular one of those rays as white, we commit ourselves to colouring an even number of the entries white. It follows that the total number of table entries coloured white must be even, *not* odd. Thus, a colouring of the 20 ray set in accordance with VC1′ is impossible. (Note for future reference that the first part of the argument -- the argument for ‘odd’ -- uses only VC1′, while the second -- the argument for ‘even’ -- relies essentially on NC, by assuming that occurrences of the same ray in different columns get assigned the same number!)

Return to Table of Contents

§3.3: The Original KS Argument. Technical Preliminaries.

The original KS proof operates on a three-dimensional complex Hilbert space H3. It requires two things: (1) sets of triples of rays which are orthogonal in H3; (2) a constraint to the effect that of every orthogonal triple one ray gets assigned the number 1, the two others 0. Both things are achieved thus:

We consider an arbitrary operator Q on H3 with three distinct eigenvalues q_1, q_2, q_3, its eigenvectors $|q_1>$, $|q_2>$, $|q_3>$, and projection operators P_1, P_2, P_3 projecting on the rays spanned by these vectors. Now, P_1, P_2, P_3 are themselves observables (namely, P_i is a 'yes-no observable' corresponding to the question 'Does the system have value q_i for Q?'). Moreover, P_1, P_2, P_3 are mutually compatible, so we can apply the Sum Rule and Product Rule, and thereby derive a constraint on the assignment of values (Proof):

(VC1) $v(P_1) + v(P_2) + v(P_3) = 1$, where $v(P_i) = 1$ or 0, for i = 1, 2, 3.

The arbitrary choice of an observable Q defines new observables P_1, P_2, P_3 which, in turn, select rays in H3. So, to impose that observables P_1, P_2, P_3 all have values means to assign numbers to rays in H3, and VC1, in particular, means that of an arbitrary triple of orthogonal rays, specified by choice of an arbitrary Q (briefly: an orthogonal triple in H3), exactly one of its rays is assigned 1, the others 0. Now, if we introduce different incompatible observables Q, Q', Q'', ... these observables select different orthogonal triples in H3. Assumption (1) of the KS theorem (which, effectively, is VD) now tells us that every one of these triples has three values, and VC1 tells us that these values must be for every triple, exactly {1, 0, 0}. What KS now shows is that, *for a specific set of orthogonal triples in H3, an assignment of numbers* {1, 0, 0} *to every one of them is impossible*. Further reflection yields that while H3 is complex, it is in fact enough to consider a real three-dimensional Hilbert space R3. For, we can show that if an assignment of values according to VC1 is possible on H3, then it is possible on R3. Contrapositively, if the assignment is impossible on R3, then it is impossible on H3. So we can fulfill the conditions necessary to get the KS proof started and at the same time reduce the problem to one on R3. Now, the equivalent in R3, of an arbitrary orthogonal triple in H3, is, again, an arbitrary triple of orthogonal rays (briefly: an orthogonal triple in R3). So, if KS want to show that, for a specific set of n orthogonal triples in H3 (where n is a natural number), an assignment of numbers {1, 0, 0} to every one of them is impossible, it is enough for them to show that, *for a specific set of n orthogonal triples in R3, an assignment of numbers* {1, 0, 0} *to every one of them is impossible*. And this is exactly what they do.

It should be stressed, however, that at this point there is no direct connection between R3 and physical space. KS wish to show that for an arbitrary QM system requiring a representation in a Hilbert space of at least three dimensions, the ascription of values in conjunction with condition (KS2) (Sum Rule and Product Rule) is impossible, and in order to do this it is sufficient to consider the space R3. This space R3, however, does not represent physical space for the quantum system at issue. In particular, orthogonality in R3 is not to be confused with orthogonality in physical space. This becomes obvious, if we move to an

example of a QM system sitting in physical space and at the same time requiring a QM representation in H3, e.g. a one particle spin-1 system measured for spin. Given one arbitrary direction α in physical space and an operator $S\alpha$ representing the observable of a spin component in direction α, H3 is spanned by the eigenvectors of $S\alpha$, namely $|S\alpha=1>$, $|S\alpha=0>$, $|S\alpha=-1>$, which are mutually orthogonal in H3. The fact that these three vectors corresponding to three possible results of measurement in one spatial direction are mutually orthogonal illustrates the different senses of orthogonality in H3 and in physical space. (The reason lies, of course, in the structure of QM which represents different values of an observable by different directions in H3.) Now, if orthogonality in H3 differs from orthogonality in physical space, and we just use R3 to prove a result about H3, then certainly orthogonality in R3 bears no direct connection with physical space.

KS themselves, in the abstract, proceed in exactly the same way, but they illustrate with an example that *does* establish a direct connection with physical space. It is important to see this connection, but also to be clear that it is produced by KS's example and is not inherent in their mathematical result. KS propose to consider a one-particle spin 1 system and the measurement of the squared components of orthogonal directions of spin in physical space S_x^2, S_y^2, S_z^2 which are compatible (while S_x, S_y, S_z themselves are not).[5] Measurement of a squared component of spin determines its absolute magnitude, but not its direction. Here, we derive a slightly different constraint on value assignments, again using the Sum Rule and the Product Rule (Proof):

(VC2) $v(S_x^2) + v(S_y^2) + v(S_z^2) = 2$, where $v(S\alpha^2) = 1$ or 0, for $\alpha = x, y, z$.

Now, since S_x^2, S_y^2, S_z^2 are compatible, there is an observable O such that S_x^2, S_y^2, S_z^2 are all functions of O. So, the choice of an arbitrary O fixes S_x^2, S_y^2, S_z^2 and, since these latter can be directly associated with mutually orthogonal rays in H3, again fixes the choice of an orthogonal triple in H3. The resulting problem here is to assign numbers {1, 1, 0} to an orthogonal triple in H3 specified by the choice of O or, more directly, S_x^2, S_y^2, S_z^2. This is, of course, the mirror-image of our previous problem of assigning numbers {1, 0, 0} to such a triple, and we need not consider it separately.

However, the choice of a specific O which selects observables S_x^2, S_y^2, S_z^2 at the same time selects three orthogonal rays in physical space, namely by fixing a coordinate system $\pm x$, $\pm y$, $\pm z$ (which defines along which orthogonal rays the squared spin components are to be measured) *in physical space*. So now, by choice of an observable O, there *is* a direct connection of with directions in H3: orthogonality in H3 now *does* correspond to orthogonality in physical space. The same holds for R3, if, in order to give an argument for H3, we consider R3. Orthogonality in R3 now corresponds to orthogonality in physical space. It is important to notice that this correspondence is not necessary to give the argument, even if we insist that the pure mathematical facts should be supplemented by a physical interpretation - since we have, just before, seen an example without any correspondence. The point is only that we *can* devise an example such that there is a correspondence. In particular, we can now follow the proof in R3 and all along imagine a system sitting in physical space, namely a spin 1 particle, returning three values upon measurement of three physical magnitudes, associated directly with orthogonal directions in physical

space, namely $v(S_x^2)$, $v(S_y^2)$, $v(S_z^2)$, for arbitrary choices of *x*, *y*, *z*. The KS proof then shows that it is impossible (given its premises, of course) to assign to the spin 1 particle values for all these arbitrary choices. That is, the KS argument shows that (given the premises) a spin 1 particle cannot possess all the properties at once which it displays in different measurement arrangements.

Three further features which have become customary in KS arguments need to be mentioned:

(1) Obviously, we can unambiguously specify any ray in R3 through the origin by just giving one point contained in it. KS thus identify rays with points on the unit sphere E. KS do not need to refer to concrete coordinates of a certain point, since their argument is ‘coordinate-free’. We will, however, for illustration sometimes mention concrete points and then (a) use Cartesian coordinates to check orthogonality relations and (b) specify rays by points not lying on E. (Thus, e.g., the triple of points (0, 0, 1), (4, 1, 0), (1, -4, 0) is used to specify a triple of orthogonal rays.) Both usages conform with the recent literature (see e.g. Peres (1991) and Clifton (1993)).

(2) We translate the constraints (VC1) and (VC2) on value ascriptions into constraints for colouring the points. We can, operating under (VC1) colour the points white (for "1") and black (for "0"), or, operating under (VC2) colour the points white (for "0") and black (for "1"). In either case the constraints translate into the same colouring problem.

(3) KS illustrate orthogonality relations of rays by graphs which have come to be called *KS diagrams*. In such a diagram each ray (or point specifying a ray) is represented by a vertex. Vertices joined by a straight line represent orthogonal rays. The colouring problem then translates into the problem of colouring the vertices of the diagram white or black such that joined vertices cannot be both white and triangles have exactly one white vertex.

Return to Table of Contents

§3.4: The Original KS Argument. Sketch of the Proof.

KS proceed in two steps.

(1) In the first (and decisive) step they show *that two rays with opposite colours cannot be arbitrarily close*. They show that the diagram Γ_1 depicted in Fig. 1 which consists of ten vertices including a_0 and a_9 is constructible, if a_0 and a_9 are separated by an angle θ with $0 \leq \theta \leq \sin^{-1}(1/3)$ (Proof).

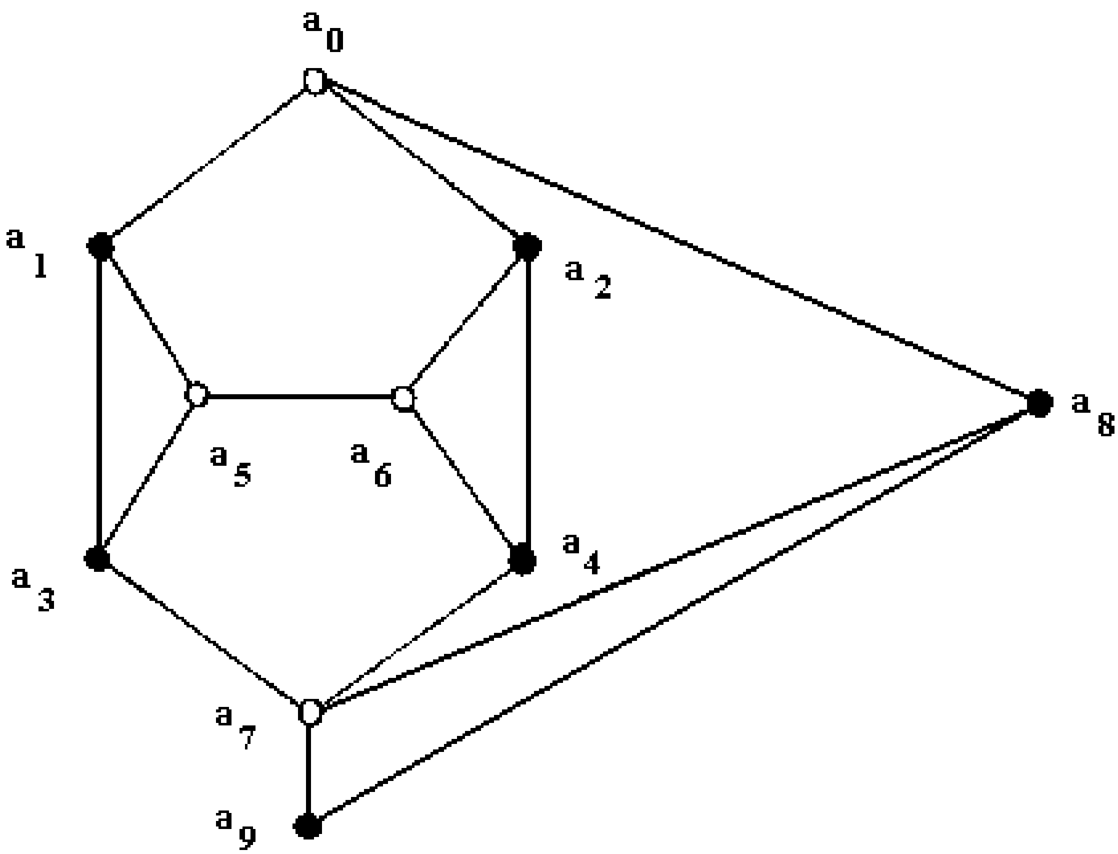


Figure 1: Ten-point KS graph Γ_1 with inconsistent colouring.

What this step shows is the following: It is possible to construct this KS diagram, i.e. to specify ten rays in R3 with orthogonality relations as specified in the diagram, but only if rays a_0 and a_9 are closer than $\sin^{-1}(1/3)$. Consider now (for a *reductio ad absurdum*) that a_0 and a_9 have different colours. We arbitrarily colour a_0 white and a_9 black. The colouring constraints then force us to colour the rest of the diagram as is done in fig.1, but this forces that a_5 and a_6 are orthogonal and both white -- which is forbidden. Hence, two points closer than $\sin^{-1}(1/3)$ cannot have different colour. Contrapositively, two points of different colour cannot be closer than $\sin^{-1}(1/3)$.

(2) KS now construct another quite complicated KS diagram Γ_2 in the following way. They consider a realization of Γ_1 for an angle θ=18° < $\sin^{-1}(1/3)$. Now they choose three orthogonal points p_0, q_0, r_0 and space interlocking copies of Γ_1 between them such that every instance of point a_9 of one copy of Γ_1 is identified with the instance of a_0 of the next copy. In this way five interlocking copies of Γ_1 are spaced between p_0 and q_0 and all five instances of a_8 are identified with r_0 (likewise for q_0, r_0, and p_0, and for r_0, p_0, and q_0). That Γ_2 is constructible is borne out directly by the construction itself. Spacing out five copies of with angles θ=18° between instances of a_0 will space out an angle of 5x18° = 90° which is exactly what is required. Moreover, wandering from one copy of Γ_1 to the next between, say, p_0 and q_0 is equivalent to a rotation of the copy about the axis through the origin and r_0 of 18° which evidently

conserves the orthogonality between the points a_0 and a_9 of the copy and r_0.

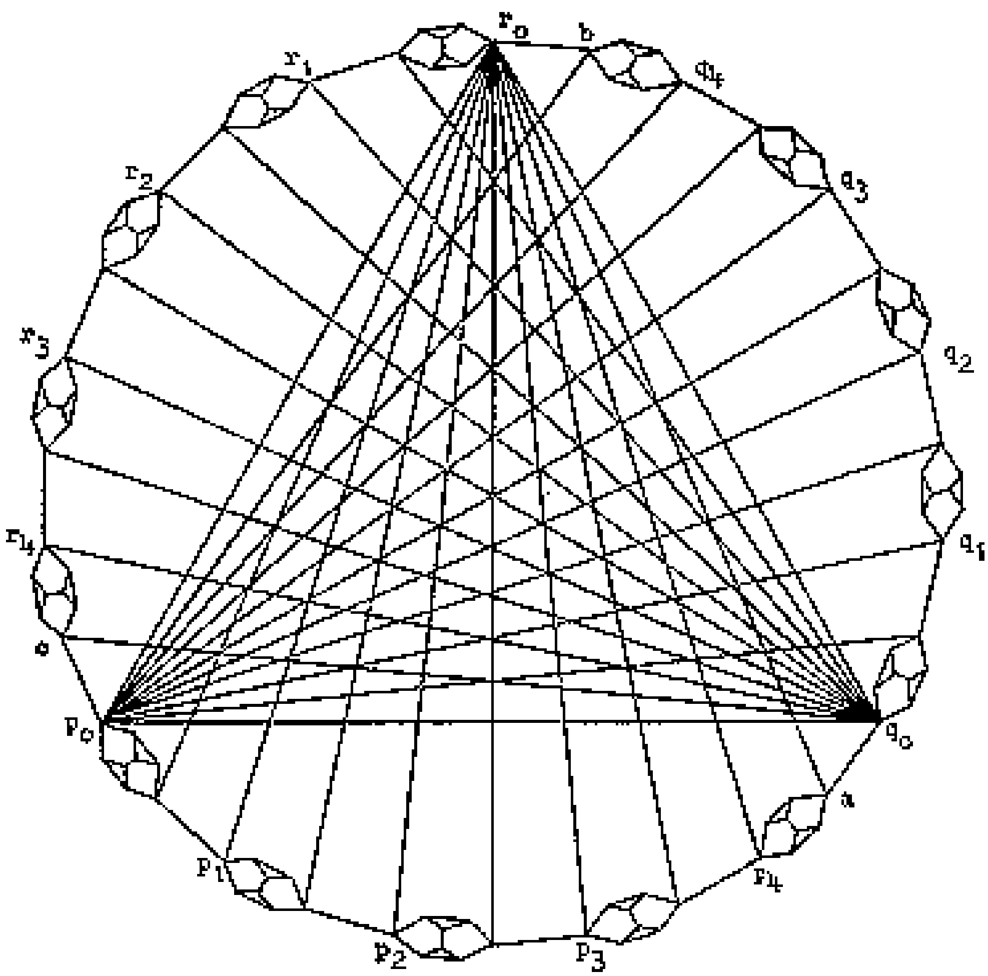


Figure 2: 117-point KS graph Γ_2
(*From* Kochen and Specker 1967, 69; *by permission of the Indiana University Mathematics Journal*)

However, although Γ_2 is constructible it is *not* consistently colourable. From the first step we know that a copy of Γ_1 with ϑ=18° forces that points a_0 and a_9 have equal colour. Now, since a_9 in one copy of Γ_1 is identical to a_0 in the next copy, a_9 in the second copy must have the same colour as a_0 in the first. Indeed, by repetition of this argument all instances of a_0 must have the same colour. Now, p_0, q_0, r_0 are identified with points a_0, so they must be either all white or all black - both of which are inconsistent with the colouring constraint that exactly one of them be white.

If from the 15 copies of Γ_1 used in the process of constructing Γ_2 we subtract those points that were identified with each other, we end up with 117 different points. So what KS have shown is that a set of 117 observables cannot consistently be assigned values in accordance with VC1 (or, equivalently, VC2).

Note that in the construction of Γ_1, i.e. the set of 10 points forming 22 interlocking triples, all points

except a_9 appear in more than one triple. In Γ_2 every point appears in a multiplicity of triples. It is here that the noncontextuality premise is crucial to the argument: We assume that an arbitrary point keeps its value 1 or 0 as we move from one orthogonal triple to the next (i.e. from one maximal set of compatible observables to another).

Return to Table of Contents

§3.5: A Statistical KS Argument in Three Dimensions (Clifton)

Finally, we return to R3. Recall KS's first step which establishes that two points with opposite colour cannot be arbitrarily close. It is this first step which carries the whole force of the argument. Bell had established it in a different way and had then argued that in a noncontextual HV interpretation points with opposite colour must be arbitrarily close. It is this first step which Clifton exploits in an argument that combines Bell's and KS's ideas.

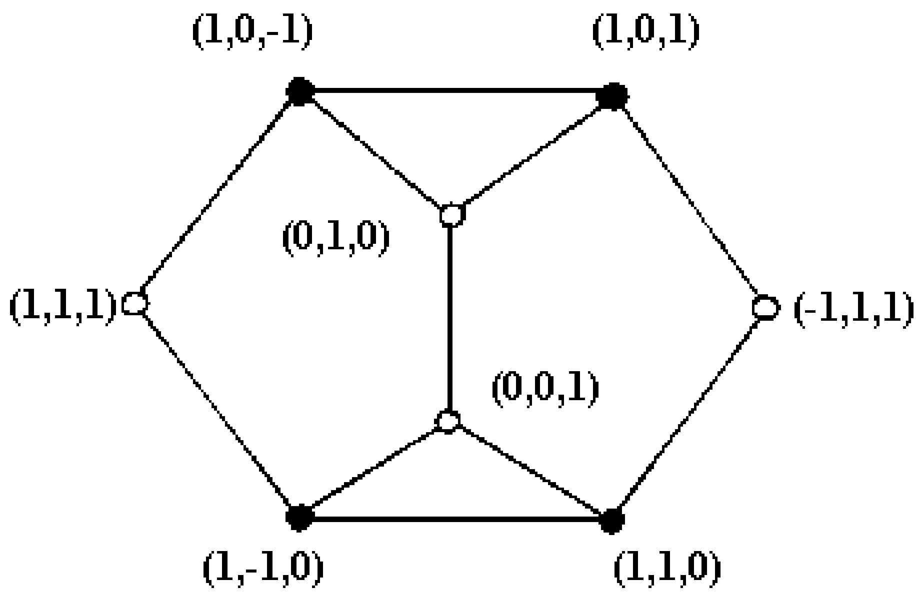


Figure 3: 8-point KS-Clifton graph Γ_3 with inconsistent colouring.

Consider the KS diagram Γ_3 shown in Figure 3 which obviously is a part of KS's Γ_1, but which has additional concrete assignments of eight points satisfying the orthogonality relations (and thus proving directly that Γ_3 is constructible). From our previous colouring constraints (joined points are not both white and a triangle has exactly one white point) we see immediately that Γ_3 is colourable only if the outermost points are not both white (which would force, as shown in fig. 3, that two joined points are white - contrary to the constraints). Moreover, we easily calculate the angle between the two outermost points to be $\cos^{-1}(1/3)$.[6] So we conclude that if one wants to colour all eight points and wants to colour white one of the outer ones, then the other must be black. Taking into account that we can insert a

diagram between any two points in R3 which are separated by exactly the angle $\cos^{-1}(1/3)$ and translating our problem back from a colouring problem into KS's example (constraint VC2), we end with a constraint VC2′:

> (VC2′) If, for a spin-1 system, a certain direction *x* of spin in space is assigned value 0, then any other direction *x*′ which lies away from *x* by an angle $\cos^{-1}(1/3)$ must be assigned value 1, or, in symbols: If $v(S_x)=0$, then $v(S_{x'})=1$.

The argument so far has made use of the original KS conditions KS1 and KS2. We now assume, in addition, that any constraint on value assignments will show up in the measurement statistics. In particular: A value assignment dictated by a constraint entails that this assigned value with certainty is the result of any measurement respecting the constraint. Or in symbols:

> (3) If prob[v(A)=a] = 1, and v(A)=a implies v(B)=b, then prob[v(B)=b] = 1.

Despite the use of statistics, this reasoning crucially differs from von Neumann's argument. Von Neumann had argued that algebraic relations between values should transfer into the statistics of the measured values, therefore the QM constraints on these statistics should have value constraints as their exact mirror images - which reasoning leads us to derive value constraints from statistical constraints (for arbitrary observables). Here, on the contrary, we derive a value constraint independently from any statistical reasoning, and then conclude that this constraint should transfer into the measurement statistics.[7]

Now, VC2′ and the statistical condition (3) entail: If prob[$v(S_x)=0$]=1, then prob[$v(S_{x'})=1$]=1. This, however, contradicts the statistics derived from QM for a state where prob[$v(S_x)=0$] = 1.[8] In fact, there is a probability of 1/17 that $v(S_{x'}=0)$. So, in a long-run test 1/17 of the spin-1 particles will violate the constraint.

1/17 may not seem a terribly impressive number, but if we accept Clifton's statistical reasoning, we have an entirely valid KS argument establishing a contradiction between an HV interpretation of QM and the very predictions of QM. Moreover, Clifton presents a slightly more complex set of 13 observables yielding, along the same lines, a statistical contradiction of 1/3.

Return to Table of Contents

§4: The Functional Composition Principle

The key ingredients of the KS theorem are the constraints on value assignments spelled out in (2): the Sum Rule and Product Rule. They derive from a more general principle, called the *Functional Composition Principle* (FUNC).[9] The principle trades on the mathematical fact that for a self-adjoint

operator **A** operating on a Hilbert space, and an arbitrary function f: **R**→**R** (where **R** is the set of the real numbers), we can define f(**A**) and show that it also is a self-adjoint operator (hence, we write **f(A)**). If we further assume that to every self-adjoint operator there corresponds a QM observable, then the principle can be formulated thus:

> **FUNC:** Let **A** be a self-adjoint operator associated with observable A, let f: **R** → **R** be an arbitrary function, such that **f(A)** is another self-adjoint operator, and let $|\varphi\rangle$ be an arbitrary state; then **f(A)** is associated uniquely with an observable f(A) such that:
>
> $v(\mathrm{f(A)})^{|\varphi\rangle} = \mathrm{f}(v(\mathrm{A}))^{|\varphi\rangle}$

(We introduced the state superscript above to allow for a possible dependence of values on the particular quantum state the system is prepared in.) The Sum Rule and the Product Rule are straightforward consequences of FUNC (Proof). FUNC itself is not derivable from the formalism of QM, but a statistical version of it (called STAT FUNC) is [Proof]:

> **STAT FUNC:** Given A, f, $|\varphi\rangle$ as defined in FUNC, then, for an arbitrary real number *b*:
>
> $\mathrm{prob}[v(\mathrm{f(A)})^{|\varphi\rangle}=b] = \mathrm{prob}[\mathrm{f}(v(\mathrm{A}))^{|\varphi\rangle}=b]$

But STAT FUNC cannot only be derived from the QM formalism, it also follows from FUNC [Proof]. This can be seen as providing "a plausibility argument for FUNC" (Redhead 1987: 132): STAT FUNC is true, as a matter of the mathematics of QM. Now, if FUNC were true, we could derive STAT FUNC, and thus understand part of the QM mathematics as a consequence of FUNC.

But how can we derive FUNC itself, if not from STAT FUNC? It is a direct consequence of STAT FUNC and three assumptions (two of which are familiar from the introduction):

> **Value Realism (VR):** If there is an operationally defined real number α, associated with a self-adjoint operator **A** and distributed probabilistically according to the statistical algorithm of QM for **A**, i.e. if there exists a real number β with β=prob (v(**A**)=α), then there exists an observable A with value α.
>
> **Value Definiteness (VD):** All observables defined for a QM system have definite values at all times.
>
> **Noncontextuality (NC):** If a QM system possesses a property (value of an observable), then it does so independently of any measurement context.

Some comments on these conditions are in order. First, we need to explain the content of VR. The statistical algorithm of QM tells us how to calculate a probability from a given state, a given observable

and its value. Here we understand it as a mere mathematical device without any physical interpretation: Given a Hilbert space vector, an operator and its eigenvalues, the algorithm tells us how to calculate new numbers (which have the properties of probabilities). In addition, by ‘operationally defined’ we here simply mean ‘made up from a number which we know to denote a real property’. So, VR, in effect, says that, if we have a real property γ (value γ of an observable G), and we are able to construct from γ a new number α and find an operator **A** such that α is an eigenvalue of **A**, then (we have fulfilled everything necessary to apply the statistical algorithm; thus) **A** represents an observable A and its value α is a real property.

Secondly, concerning NC: A failure of NC could be understood in two ways. Either, the value of an observable might be context-dependent, although the observable itself is not; or, the value of an observable might be context-dependent, because the observable itself is. There are, however, good grounds to think that both options are equivalent. We will indeed assume that, if NC holds, this means that the *observable* -- and *thereby* also its value -- is independent of the measurement context, i.e. is independent of how it is measured. In particular, the independence from context of an observable implies that there is a 1:1 correspondence of observables and operators. This implication of NC is what we will use presently in the derivation of FUNC. *Conversely*, failure of NC will be construed solely as failure of the 1:1 correspondence.

From VR, VD, NC and STAT FUNC, we can derive FUNC as follows. Consider an arbitrary state of a system and an arbitrary observable Q. By VD, Q possesses a value $v(Q)=a$. Thus, we can form the number $f(v(Q))=b$ for an arbitrary function f. For this number, by STAT FUNC, prob[$f(v(Q))=b$] = prob[$v(f(Q))=b$]. Hence, we have, by transforming probabilities according to STAT FUNC, created a new self-adjoint operator **f(Q)**, and associated it with the two real numbers b and prob[$f(v(Q))=b$]. Thus, by VR, there is an observable corresponding to **f(Q)** with value b, hence $f(v(Q))=v(f(Q))$. By NC, that observable is unique, hence FUNC follows.

Return to Table of Contents

§5: Escaping the KS Argument

The previous section clarifies which possibilities the HV theorist has to escape the KS argument: denying one of the three premises which together entail FUNC (hence the Sum Rule and Product Rule).

§5.1: No General Value Definiteness

VD, we recall, was the fundamental presupposition of HV interpretations. So, if, in order to escape a powerful argument against the possibility of HV interpretations, these interpretations drop their fundamental motivation, this seems not to make much sense. But some interpreters point out that, between holding that only those observables which QM prescribes to have values[10] and holding that all of them have values, there is some leeway, namely, to propose that more observables, than prescribed in

QM, but not all, have values (‘partial value definiteness’). This option of partial value definiteness has been taken by various modal interpretations and also has been explored by John Bell in his ‘beable approach’ to QM (1987: ch.7).

The rocks and shoals of modal interpretations are beyond the scope of this article (see the entry on modal interpretations). We just note that it is by no means clear how these interpretations can manage to always pick out the right set of observables assumed to have values. ‘Right set’ here means that the observable actually measured must always be included (in order to avoid the measurement problem) and must always recover the QM statistics. We also mention two important results which cast doubt on the feasibility of modal interpretations: First, it can be shown that either partial value definiteness collapses into total value definiteness (i.e., VD) or classical reasoning about physical properties must be abandoned. (Clifton 1995). Second, it is possible to derive a kind of KS theorem even in certain modal interpretations (Bacciagaluppi 1995, Clifton 1996).

Return to Table of Contents

§5.2: Denial of Value Realism

The derivation of FUNC basically consists in the construction of an observable (i.e. f(Q)) via an operator (i.e. **f(Q)**) from the probability distribution of a number (i.e. f(*v*(Q)) which number in turn is constructed from another number, (i.e. *v*(Q)). Now, instead of denying that *v*(Q) exists in all cases (like the first option would have it), we can reject that the existence of a number α and the construction of **f(Q)** automatically lead to an observable, i.e. we reject VR. This amounts to rejecting that for every self-adjoint operator, there is a well-defined observable.

Now, in order to formulate VR we had to give a very reduced reading to the statistical algorithm, i.e. that it is a mere mathematical device for calculating numbers from vectors, operators and numbers. (What if we had done otherwise? Well, if we say: ‘Whatever fulfills the statistical algorithm is an observable’, we cannot very well suppose that an operator, in order to fulfill the algorithm, must be understood as an observable, since this would make the condition a trivial consequence of the algorithm.) This reading is very artificial and presupposes that a minimal interpretational apparatus required to make physical sense of some operators (like **Q**) can be withheld for others (like **f(Q)**).

Moreover, it seems entirely implausible to assume that some operators - sums and products of operators that are associated with well-defined observables - are themselves not associated with well-defined observables, even if they mathematically inherit exact values from their summands or factors. Put in a crude example, this would amount to saying that to ask for a system's energy is a well-defined question, while to ask for the square of the system's energy is not, even if, from our answer to the first question and trivial mathematics, we have a well-defined answer at hand. There seems no good a priori reason to justify this restriction. So, to make rejecting VR plausible at all, an additional proposal is made: It is crucial to the KS argument that one and the same operator is constructed from different maximal ones which are incompatible: **f(Q)** is identical to **g(P)**, where $\mathbf{PQ} - \mathbf{QP} \neq 0$. We now assume that only the

construction of **f(Q)** via Q, but not the one via P, leads to a well-defined observable.[11]

This move however, automatically makes some observables context-sensitive. So, this way of motivating the denial of VR amounts to a kind of contextualism, which we might come by cheaper, by directly rejecting NC, and without any tampering with the statistical algorithm. (This fact explains why we did not mention denial of VR as a separate option in the introduction.).

Return to Table of Contents

§5.3: Contextuality

Finally, we might accept VD and VR, but deny that our construction of an observable f(Q) is unambiguous. Thus, we accept that **f(Q)** and **g(P)** are mathematically identical, but physically they correspond to different observables, since an actual determination of *v*(f(Q)) must proceed via measuring Q, but the determination of *v*(g(P)) involves measuring P which is incompatible with Q. Since *v*(f(Q)) and *v*(g(P)) are outcomes of different measurement situations, there is no reason to assume that *v*(f(Q)) = *v*(g(P)). This way to block the KS proof comes to understanding f(Q) and g(P) as different observables (because of sensitivity to context), thus it amounts to rejecting NC. There are mainly two ways, in the literature, to further motivate this step. Accordingly, there are two important brands of contextuality to be discussed -- causal and ontological contextuality.

The KS argument has been presented for possessed values of a QM system - independently of considerations about measurement. Indeed, in the argument measurement was mentioned only once and in the negative - in NC. However, since now we consider the rejection of NC, we must also take into account measurement and its complications. An additional manifestation of our innocuous realism (see the introduction above) is a principle of *faithful measurement* (FM): QM measurement of an observable faithfully delivers the value which that observable had immediately prior to the measurement interaction. FM also is an extremely plausible presupposition of natural science. Moreover, FM entails VD (therefore we could have, using the stronger principle, given a KS argument for possible measurement results). Consider now the motivation, for the HV proponent, to reject NC. Obviously, the aim is to save other presuppositions, especially VD. Now, VD and NC are independent realist convictions, but NC and FM are not quite so independent. Indeed, we will see that rejection of NC entails the rejection of FM in one version of contextuality, and strongly suggests it in the other. (This makes more precise the somewhat cryptical remark from the introduction that it is not obvious what an interpretation endorsing the realist principle VD, but rejecting the realist principle NC, should look like. Such an interpretation would have to violate a third realist principle, i.e. FM.)

Causal Contextuality

An observable might be *causally* context-dependent in the sense that it is causally sensitive to how it is measured. The basic idea is that the observed value comes about as the effect of the system-apparatus interaction. Hence, measuring a system via interaction with a P-measuring apparatus might yield a value

v(g(P)), measuring the same system via interaction with a Q-measuring apparatus a different value v(f(Q)), although both observables are represented by the same operator **f(Q)** = **g(P)**. The difference in values is explained in terms of a context-dependence of the observables: The latter are context-dependent, since the different ways to physically realize them causally influence the system in different ways and thereby change the observed values.

If an interpreter wanted to defend causal contextuality, this would entail abandoning FM, at least for observables of the type f(Q) (non-maximal observables): Since their values causally depend on the presence of certain measurement arrangements, these arrangements are causally necessary for the values to come about, thus the values cannot be present before the system-apparatus interaction, and FM is violated. As an advantage of causal contextualism the following might be pointed out. It does not imply that the ontological status of the physical properties involved must change, i.e. does not imply that they become relational. If the property in an object is brought about via interaction with another one, it can still be one which the object has for itself after the interaction. However, the idea of causal contextuality is sometimes discussed critically, since there is reason to think that it may be empirically inadequate (see Shimony 1984, Stairs 1992).

Ontological Contextuality

An observable might be *ontologically* context-dependent in the sense that in order for it to be well-defined the specification of the observable it 'comes from' is necessary. Thus, in order to construct a well-defined observable from operator **f(Q)** = **g(P)**, we need to know whether it is physically realized via observable P or observable Q. This way out of the KS problem, was first proposed (but not advocated) by van Fraassen (1973). There are, then, as many observables and kinds of physical properties for an operator **f(Q)** as there are ways to construct **f(Q)** from maximal operators. Without further explanation, however, this idea just amounts to an ad hoc proliferation of physical magnitudes. A defender of ontological contextuality certainly owes us a more explicit story about the dependence of observable f(Q) on observable Q. Two possibilities come to mind:

(a) We might think that v(f(Q)) just is not a self-sustained physical property, but one which ontologically depends on the presence of another property v(Q). (Recall that in the proof of FUNC v(f(Q)) *is* constructed from v(Q).) But, since the position does not reject questions about values of f(Q) in a P-measurement situation as illegitimate (because it does not trade on a notion of an observable being well-defined in one context only!), this seems to lead to new and pressing questions, to say the least. As an attempt to defend a contextualist *hidden variables* interpretation, this position must concede that not only does the system have, in the Q-measurement situation, a value v(Q), but also, in a P-measurement situation, it has a value v'(Q), although perhaps $v'(Q) \neq v(Q)$. Now, questions for values of f(Q) in this situation at least are legitimate. Does v'(Q) install another $v'(f(Q)) \neq v(f(Q))$? Or does v'(Q), in opposition to v(Q), not lead to a value of f(Q), at all? Neither option seems plausible, for couldn't we, just by switching for a certain prepared system between a P- and Q-measurement situation either switch v(f(Q)) in and out of existence or switch between v(f(Q)) and v'(f(Q))? . (b) We might think that, in order for f(Q) to be well-defined, one measurement arrangement rather than the other is necessary. The idea is strongly reminiscent of Bohr's 1935 argument against EPR, and indeed may be viewed as the appropriate

extension of Bohr's views on QM to the modern HV discussion (see Held 1998, ch.7). In this version of ontological contextualism the property *v*(f(Q)), rather than depending on the presence of another property *v*(Q), is dependent on the presence of a Q-measuring apparatus. This amounts to a holistic position: For some properties it only makes sense to speak of them as pertaining to the system, if that system is part of a certain system-apparatus whole. Here, the question for values of f(Q) in a P-measurement situation *does* become illegitimate, since f(Q)'s being well-defined is tied to a Q-measurement situation. But again reservations apply. Does the position hold that, in opposition to f(Q), Q itself is well-defined in a P-measurement situation? If it does not, Q hardly can have a value (since not being well-defined was the reason to deny f(Q) a value) which means that we are not considering an HV interpretation any longer, and that there is no need to block the KS argument, at all. If it does, what explains that, in the P-measurement situation, Q remains well-defined, but f(Q) loses this status?

What becomes of FM in both versions of ontological contextualism? Well, if we remain agnostic about how the position could be made plausible, we can save FM, while, if we choose version (a) or (b) to make it plausible, we lose it. Consider first an agnostic denial of NC. FM said that every QM *observable* is faithfully measured. Now, contextualism splits an operator which can be constructed from two different noncommuting operators into two observables, and ontological contextualism does not try to give us a causal story which would ruin the causal independence of the measured value from the measurement interaction embodied in FM. We simply introduce a more fine-grained conception of observables, but for these new contextual observables still can impose FM.

However, the concrete versions of ontological contextualism, by attempting to motivate the contextual feature, ruin FM. Version (a) allows f(Q) to switch ‘on and off’ or to switch between different values upon the change between P- and Q-measurement situations - which is a flagrant violation of FM. Version (b) fares no better. It introduces the ontological dependence on the measuring arrangement. It is hard to see what else this should be, but the same causal dependence pushed to a higher, ‘ontological’ key. Again, couldn't we, just by flipping back and forth the measurement arrangement, change back and forth that f(Q) is well-defined, thus flip in and out of existence *v*(f(Q))?

Finally, we note that both types of ontological contextualism, in opposition to the causal version, *do* entail that system properties which we earlier thought to be intrinsic, become relational in the sense that a system can only have these properties either if it has certain others, or if it is related to a certain measurement arrangement.

Return to Table of Contents

§6: The Question of Empirical Testing

Famously, the violation of Bell's inequalities, prescribed by QM, has been confirmed experimentally. Is something similar possible for the KS theorem? We distinguish three questions: (1) Is it possible to realize the experiment proposed by KS as a motivation of their theorem? (2) Is it possible to test the principles leading to the theorem: the Sum Rule and Product Rule, FUNC, or NC? (3) Is it possible to test

the theorem itself?

(1) KS themselves describe a concrete experimental arrangement to measure S_x^2, S_y^2, S_z^2 on a one-particle spin 1 system as functions of one maximal observable. An orthohelium atom in the lowest triplet state is placed in a small electric field E of rhombic symmetry. The three observables in question then can be measured as functions of one single observable, the perturbation Hamiltonian H_s. H_s, by the geometry of E, has three distinct possible values measurement of which reveals which two observables of S_x^2, S_y^2, S_z^2 have value 1, which one has value 0 (see Kochen and Specker 1967: 72/311). This is, of course, a proposal to realize an experiment exemplifying our above value constraint (VC2). Could we also realize a (VC1) experiment, i.e. measure a set of commuting projectors projecting on eigenstates of one maximal observable? Peres (1995: 200) answers the question in the affirmative, discusses such an experiment, and refers to Swift and Wright (1980) for details about the technical feasibility. It seems, however, that, despite being possible in principle, no such experiment has been actually carried out (see Cabello and Garcìa-Alcaine (1998) for more discussion and another experimental proposal).

(2) In conjunction with manifestations of FUNC, i.e. the Sum Rule and the Product Rule, QM yields constraints like VC1 or VC2 that contradict VD. So providing concrete physical examples that could, given the Sum Rule and the Product Rule, instantiate VC1 or VC2, as just outlined, is not enough. We must ask whether these rules themselves can be empirically supported. There was considerable discussion in the early 80s about this question --- explicitly about whether the Sum Rule is empirically testable --- and there was general agreement that it is not.[12]

The reason is the following: Recall that the derivation of FUNC established uniqueness of the new observable f(Q) only in its final step (via NC). It is this uniqueness which guarantees that one operator represents exactly one observable such that observables (and thereby their values) in different contexts can be equated. This allows to establish indirect connections between different incompatible observables. Without this final step, FUNC must be viewed as holding relative to different contexts, the connection is broken and FUNC is restricted to one set of observables which are all mutually compatible. Then indeed FUNC, the Sum Rule and the Product Rule become trivial, and empirical testing in these cases would be a pointless question.[13] It is NC which does all the work and which deserves to be tested via checking whether for incompatible P, Q such that f(Q)=g(P) it is true that v(f(Q))=v(g(P)). Testing this, however, is impossible, due to the impossibility of simultaneously measuring P and Q.

(3) Very recently, it has been argued that the (physically reasonable) assumption of finitely precise measurement creates a decisive loophole in the KS argument (see Meyer 1999, Kent 1999, Clifton and Kent 1999; briefly MKC).[14] Indeed, if we consider a KS argument for measured values, infinite precision is crucial to the argument in two different ways: (1) It is necessary to the argument that the measured components of one triple (or quadruple) are exactly orthogonal. (2) It is necessary (to install NC) that two measurements intended to pick out the same observable as member of two different maximal sets, pick out exactly the same direction. If we relax this assumption of infinite precision, noncontextual HV models can be constructed. In these models, it is not exactly the sets of observables specified in the KS argument (or related arguments) by points in R3, but sets specified by points with

rational components (which approximate the former arbitrarily closely) that are colourable, i.e. that can consistently be assigned noncontextual values. So the argument ultimately trades on the fact that we cannot empirically distinguish between a ‘real point’ and its ‘rational’ approximation.

The MKC argument is hotly debated and the question whether it is relevant or even destructive to the KS argument is unsettled, so we shall just record part of the discussion. One quite obvious objection is that the original KS argument works for possessed values, not measured values, so the MKC argument, which turns on the finite precision of measurements, misses the mark. We might not be able to test observables which are exactly orthogonal or exactly alike in different tests, but it would be a strange HV interpretation that asserts that such components do not exist (see Cabello 1999). Of course, such a noncontextual HV proposal would be immune to the KS argument, but it would be forced to either deny that for every one of the continuously many directions in physical space there is an observable, or else deny that there are continuously many directions -- and neither denial seems very attractive.

In addition, the MKC argument is dissatisfying, since it exploits the finite precision of real measurements only in one of the above senses, but presupposes infinite precision in the other. MKC assume, for measured observables, that there is finite precision in the choice of different orthogonal triples, such that we cannot, in general, have exactly the same observable twice, as a member of two different triples. However, MKC still assume infinite precision, i.e. exact orthogonality, within the triple (otherwise the colouring constraints could find no application, at all). It has been claimed that this feature can be exploited to rebut the argument and to re-install contextualism (see Mermin 1999, Appleby 2000).

Finally, it can be shown that quantum probabilities vary continuously as we change directions in R3, so small imperfections of selection of observables that block the argument (but only for measured values!) in the single case will wash out in the long run (see Mermin 1999). This in itself does not constitute an argument, since in the colourable sets of observables in MKC's constructions probabilities also vary (in a sense) continuously.[15] We might, however, exploit Mermin's reasoning in the following way. Reconsider Clifton's set of eight directions (in Figure 3) leading to a colouring constraint for the outermost points which statistically contradicts the QM statistics by a fraction of 1/17. Now, starting from the colourable subset of directions constructed by MKC, we are unable to derive the constraint for the eight points, since these eight points do not lie in that set; i.e., as we move, in the colourable subset, from one mutually orthogonal triple of rays to the next, we never hit upon exactly the same ray again, but only to one approximating it arbitrarily closely. However, consider the following response. Assume that observables corresponding to the eight directions, though not lying in the colourable subset, exist and, according to the HV premise, all have values. Then we can derive Clifton's constraint for the outermost points. For these outermost points it is irrelevant whether, in an eventual empirical test we hit them exactly, for the Mermin argument says that, even if, in every single imperfect measurement, we only measure points nearby, we will, in the long-run better and better approximate the QM statistics for exactly the points in question - which means that we will better and better approach 1/17, while the HV assumption requires that we will better and better approach 0. (Recall also that this number can be pushed up to 1/3 by choosing a set of 13 directions!)

So, in sum it seems that, as long as we assume that there are continuously many QM observables

(corresponding to the continuum of directions in physical space), statistical tests building, e.g., on the Clifton 1993 or the Cabello/Alcaìne 1998 proposal remain entirely valid as empirical confirmations of the KS theorem. Since these statistical violations of the HV programme come about as contradictions of results of QM, VD, VR, and NC on the one hand, and QM and experiment on the other, the experimental data still force upon us the trilemma of giving up either VD or VR or NC. As we have seen, denial of value realism in the end becomes identical to a kind of contextualism, hence we really have only two options: (1) Giving up VD, either for all observables forbidden to have values in the orthodox interpretation (thus giving up the HV programme), or for a subset of these observables (as modal interpretations do). (2) Endorse a kind of contextualism. Moreover, as things presently stand, the choice between these two options seems not to be a matter of empirical testing, but one of pure philosophical argument.

Return to Table of Contents

Bibliography

- Appleby, D. M. (2000): "Contextuality of Approximate Measurements" (Can be downloaded from arXig.org E-print Archives.)
- Bacciagaluppi, Guido (1995): "Kochen-Specker Theorem in the Modal Interpretation", *International Journal of Theoretical Physics* 34: 1205-15.
- Bell, John S. (1966): "On the Problem of Hidden Variables in Quantum Mechanics", *Reviews of Modern Physics* 38: 447-52; reprinted in his (1987) (page references are to the reprint).
- Bell, John S. (1987): *Speakable and Unspeakable in Quantum Mechanics* (Cambridge; Cambridge University Press)
- Bohr, Niels (1935): "Can Quantum Mechanical Description of Physical Reality be Considered Complete?" *Physical Review* 48: 696-702 reprinted in J. Kalckar (ed.) (1996): *Niels Bohr. Collected Works. Vol. 7* (Amsterdam: Elsevier): 292-98.
- Cabello, Adàn (1999): "Comment on ‘Non-Contextual Hidden Variables and Physical Measurements’". (Can be downloaded from arXig.org E-print Archives.)
- Cabello, Adàn and Garcìa-Alcaine, Guillermo (1998): "Proposed Experimental Test of the Bell-Kochen-Specker Theorem", *Physical Review Letters* 80: 1797-99.
- Clifton, Robert K. (1993): "Getting Contextual and Nonlocal Elements-of-Reality the Easy Way", *American Journal of Physics* 61: 443-47.
- Clifton, Robert K. (1995): "Why Modal Interpretations of Quantum Mechanics Must Abandon Classical Reasoning About Physical Properties," *The International Journal of Theoretical Physics* 34 (1995): 1302-1312.
- Clifton, Robert K. (1996): "The Properties of Modal Interpretations of Quantum Mechanics," *British Journal for Philosophy of Science* 47 (1996): 371-398.
- Clifton, Robert K. and Kent, Adrian (1999): "Simulating Quantum Mechanics by Non-Contextual Hidden Variables", *Proceedings of the Royal Society of London* (to appear). (Can be downloaded from arXig.org E-print Archives.)
- Cooke, R.M., Keane, M., and Moran, W. (1985): "An Elementary Proof of Gleason's Theorem",

Mathematical Proceedings of the Cambridge Philosophical Society 98: 117-28; reprinted in Hughes (1989): 321-46.

- Fine, Arthur (1973): "Probability and the Interpretation of Quantum Mechanics", *British Journal for the Philosophy of Science* 24: 1-37.
- Fine, Arthur (1974): "On the Completeness of Quantum Mechanics", *Synthese* 29: 257-89. Reprinted in P. Suppes (ed.) (1976): *Logic and Probability in Quantum Mechanics* (Dordrecht; Reidel): 249-81.
- Fine, Arthur and Teller, Paul (1978): "Algebraic Constraints on Hidden Variables", *Foundations of Physics* 8: 629-36.
- Gleason, A.M. (1957): "Measures on the Closed Subspaces of a Hilbert Space", *Journal of Mathematics and Mechanics* 6: 885-93; reprinted in Hooker (1975): 123-34.
- Held, Carsten (1998): *Die Bohr-Einstein-Debatte. Quantenmechanik und physikalische Wirklichkeit* (Paderborn; Schöningh)
- Hooker, Clifford (ed.) (1975): *The Logico-Algebraic Approach to Quantum Mechanics* (Dordrecht; Reidel)
- Hughes, R.I.G. (1989): *The Structure and Interpretation of Quantum Mechanics* (Cambridge, MA; Harvard University Press)
- Kent, Adrian (1999): "Noncontextual Hidden Variables and Physical Measurements", *Physical Review Letters*, 83: 3755-3757.
- Kernaghan, M. (1994): "Bell-Kochen-Specker Theorem for 20 Vectors", *Journal of Physics* A 27: L829-L830.
- Kochen, Simon and Specker, Ernst (1967): "The Problem of Hidden Variables in Quantum Mechanics", *Journal of Mathematics and Mechanics* 17: 59-87; reprinted in Hooker (1975): 293-328 (page references to original and reprint).
- Meyer, David A.: "Finite Precision Measurement Nullifies the Kochen-Specker Theorem", *Physical Review Letters* 83 (1999) 3751-3754.
- Mermin, N. David (1990a): "Quantum Mysteries Revisited", *American Journal of Physics* 58: 731-34.
- Mermin, N. David (1990b): "Simple Unified Form of the Major No-Hidden Variables Theorems", *Physical Review Letters* 65: 3373-76.
- Mermin, N. David (1999): "A Kochen-Specker Theorem for Imprecisely Specified Measurements" (Can be downloaded from arXig.org E-print Archives.)
- Peres, Asher (1990): "", *Physics Letters* A 51: 107-8.
- Peres, Asher (1991): "Two Simple Proofs of the Kochen-Specker Theorem", *Journal of Physics* A 24: L175-L178.
- Peres, Asher (1995): *Quantum Theory: Concepts and Methods* (Dordrecht; Kluwer)
- Redhead, Michael (1987): *Incompleteness, Nonlocality, and Realism. A Prolegomenon to the Philosophy of Quantum Mechanics* (Oxford; Clarendon Press)
- Redhead, Michael (1995): *From Physics to Metaphysics* (Cambridge; Cambridge University Press)
- Shimony, Abner (1984): "Contextual Hidden Variables Theories and Bell's Inequalities", *British Journal for the Philosophy of Science* 35: 25-45.
- Specker, Ernst (1960): "Die Logik nicht gleichzeitig entscheidbarer Aussagen" *Dialectica* 14: 239-46.

- Stairs, Allen (1992): "Value Definiteness and Contextualism: Cut and Paste with Hilbert Space", *PSA 1992* (vol.1): 91-103.
- Swift, Arthur R. and Wright, Ron (1980): "Generalized Stern-Gerlach Experiments and the Observability of Arbitrary Spin Operators", *Journal of Mathematical Physics* 21: 77-82.
- van Fraassen, B.C. (1973): "Semantic Analysis of Quantum Logic", in C.A. Hooker (ed.): *Contemporary Research in the Foundations and Philosophy of Quantum Theory* (Dordrecht; Reidel): 80-113.
- von Neumann, John (1955): *Mathematical Foundations of Quantum Mechanics* (German edition 1932) (Princeton; Princeton University Press)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

quantum mechanics: modal interpretations of | quantum theory: measurement in

Acknowledgements

I am grateful to the editor, Rob Clifton, for many helpful comments and to Richard Sembera for technically setting up this entry.

Carsten Held
cheld@uni-freiburg.de

First published: September 10, 2000
Content last modified: September 10, 2000

Stanford Encyclopedia of Philosophy
Notes to The Kochen-Specker Theorem

Notes

[1] Note that if we understand the 'expectation' values deriving from the hidden states as what someone knowing such a state should expect as the average measurement result, then this claim is correct only given an assumption of faithful measurement (FM). FM is another typical assumption of a realist (or noncontextualist HV) interpretation of QM. The KS argument can be given without FM, since it makes claims about possessed values, not measured values. Only in statistical arguments, like von Neumann's and Clifton's, must FM be assumed. We suppress FM in the main text with the exception of Contextuality (see Section 5.3).

[2] See Bell (1966: 4) and Jammer (1974: 274, 304); see also Kochen and Specker (1967: 82/322, theorem 3) for a parallel example and criticism of von Neumann.

[3] Mermin is right insofar as Bell (1966) appeared before Kochen and Specker (1967). However, KS establish even the first step in the argument (two points being assigned different numbers (1 or 0) cannot be arbitrarily close) by an argument differing from Bell's proposal and refer to Specker (1960) which effectively establishes that step as a consequence of Gleason's theorem. So it is probably right to say that Specker and Bell independently saw and exploited that theorem.

[4] For a discussion of the relationship between contextuality and nonlocality, see Mermin (1990a) and Clifton (1993).

[5] See Kochen and Specker (1967: 71-72/310-11).

[6] By elementary vector algebra, viz.: $\mathbf{a}\cdot\mathbf{b}=|a|\ |b| \cos \varphi$, where $\mathbf{a}\cdot\mathbf{b}$ is the inner product of vectors **a** and **b** and φ is the angle between them.

[7] Both arguments, i.e., von Neumann's and Clifton's, presuppose faithful measurement (FM) (see also fn 1). Indeed, an eventual failure of FM would be the natural way to explain why a certain constraint on possessed values does not show up in the measurement statistics.

[8] A state of a spin-1 particle where prob($v(S_x^2)=0$)=1 is $|S_x=0\rangle$. Expanding this state in the eigenvectors of $S_{x'}$ yields

$|S_x=0\rangle=$

$$\cos \wp\ (1 + \sin^2 \wp)^{-1/2}\ |S_x{}' = 0\rangle + \sin \wp\ (1 + \sin^2 \wp)^{-1/2}\ (|S_x{}' = -1\rangle - |S_x{}' = +1\rangle)$$

Now, with $\cos \wp = 1/3$, we get:

$$\begin{aligned} \text{prob}(v(S_x{}'^2) = 0) &= \text{prob}(v(S_x{}') = 0) \\ &= (\cos \wp)^2 / (1 + \sin^2 \wp) \\ &= 1/9\ (1 + 8/9)^{-1} \\ &= 1/17 \end{aligned}$$

[9] See Redhead (1987: 121). See also Kochen and Specker (1967: 64/299, eq. 4) and Fine and Teller (1978: 631) where the principle appears under the name "functional relation condition".

[10] It is not exactly true that the QM formalism prescribes *any* observable to have a value. In fact, the formalism itself is entirely silent about values, apart from the statistical predictions it entails via the statistical algorithm (see Redhead 1987: 8). But a crucial assumption of orthodox interpretations is the eigenstate-eigenvalue link (see Fine 1973: 20): Observable A on a system has value a_k iff the system is in state $|a_k\rangle$. The 'if' direction of this principle, which leads to a minimum of value ascriptions, is endorsed by modal interpreters (this direction is equivalent to the Eigenvector Rule in Redhead 1987: 73, 120). What they usually deny is the 'only if' direction, which thereby allows them to attribute values to more observables than allowed by orthodox interpretations.

[11] This is Redhead's (1987: 135-36) construal of a proposal by Fine (1974).

[12] See Fine and Teller (1978: 636), Redhead (1987: 138) and references therein.

[13] What we mean here is the following. A definition of observable f(A) (or observable A + B, or observable A▪B, both constructed from observables A and B) would be that it is an observable which takes a value calculated by measuring v(A) (or measuring v(A) and v(B)) and applying f to the result (or calculating $v(A) + v(B)$, or calculating $v(A)\ v(B)$). FUNC, the Sum Rule and the Product Rule, as restricted to one measurement context, trivially repeat these definitions, and there obviously is no point in testing, e.g., whether $v(A \cdot B)$ really equals $v(A)\ v(B)$, if the former expression is defined by the latter.

[14] It should be noted here that Meyer and Clifton / Kent have quite different intentions. Meyer really criticizes the KS argument against HV as being invalid. Clifton and Kent, however, by presenting HV constructions along Meyer's lines do not mean to bring forth a physically plausible HV theory, but only intend to show that QM statistics for KS-type sets of observables can be *simulated* by a noncontextual HV model, thus in an entirely classical way.

[15] The proviso 'in a sense' comes from the fact that a colourable subset of the set of observables

corresponding to all directions in R3, since it is a proper subset of the latter, is not itself continuous in the intuitive sense. We can, however, define a probability function from such a colourable subset into [0, 1] which obeys the usual continuity definition of elementary calculus.

Carsten Held
cheld@uni-freiburg.de

First posted: September 10, 2000
Last modified: September 10, 2000

Stanford Encyclopedia of Philosophy
Supplement to The Kochen-Specker Theorem

Proof of VC1

We exploit two mathematical facts about projection operators P_i:

(A) $P_i^2 = P_i$ (the P_i are 'idempotent');

(B) If H is a Hilbert space of denumerable dimension, and if the P_i are operators projecting on q_i, where the set $\{q_i\}$ forms an orthonormal basis of H, then $\Sigma_i P_i = I$ (where I is the identity operator) (the P_i form 'a resolution of the identity').

Consider now an arbitrary state $|\psi>$ and an arbitrary nondegenerate operator Q on H3, its eigenvectors $|q_1>$, $|q_2>$, $|q_3>$, and projection operators P_1, P_2, P_3 whose ranges are the rays spanned by these vectors. The eigenvectors form an orthonormal basis, thus, by (B):

$$P_1 + P_2 + P_3 = I$$

Now, P_1, P_2, P_3 are compatible, so from assumption KS2 (a) (Sum Rule):

$$(P_1) + v(P_2) + v(P_3) = v(I)$$

Now, from KS2 (b) (Product Rule) and (A):

$$v(P_i)^2 = v(P_i^2) = v(P_i)$$
$$\Rightarrow v(P_i) = 1 \text{ or } 0$$

Now, assume an observable R such that $v(R) \neq 0$ in state $|\psi>$. From this assumption and KS2 (b) (Product Rule):

$$v(R) = v(I\,R) = v(I)\,v(R)$$
$$\Rightarrow v(I) = 1$$

Hence:

$$\text{(VC1)}\quad v(P_1) + v(P_2) + v(P_3) = 1$$

where $v(P_i) = 1$ or 0, for $i = 1, 2, 3$.

Carsten Held
cheld@uni-freiburg.de

Return to The Kochen-Specker Theorem

First posted: September 10, 2000
Last modified: September 10, 2000

Stanford Encyclopedia of Philosophy
Supplement to The Kochen-Specker Theorem

Proof of VC2

Let S_x, S_y, S_z be the usual angular momentum operators satisfying $[S_x, S_y] = i\, S_z$, and define $S^2 := S_x^2 + S_y^2 + S_z^2$. It can be shown that the eigenvalues of S^2 are $s(s + 1)$ where s is an integer or half-integer.

Now let s=1. Then it follows (see e.g. Kochen and Specker 1967: 308, Redhead 1987: 37-38) that S_x^2, S_y^2, S_z^2 are all mutually commuting and that:

$$S_x^2 + S_y^2 + S_z^2 = 2I,$$

where I is the identity operator. Now, from KS2 (a) (Sum Rule):

$$v(S_x^2) + v(S_y^2) + v(S_z^2) = 2v(I)$$

Now, assume an observable R such that $v(R) \neq 0$ in state $|\psi\rangle$. From this assumption and KS2 (b) (Product Rule):

$$v(R) = v(I\, R) = v(I)\, v(R)$$
$$\Rightarrow v(I) = 1$$

Hence:

$$\text{(VC2)} \quad v(S_x^2) + v(S_y^2) + v(S_z^2) = 2$$

where $v(S_i^2) = 1$ or 0, for $i = x, y, z$.

Carsten Held
cheld@uni-freiburg.de

Return to The Kochen-Specker Theorem

First posted: September 10, 2000
Last modified: September 10, 2000

Stanford Encyclopedia of Philosophy
Supplement to The Kochen-Specker Theorem

Proof of Step 1

A given point a_0 on the unit sphere E uniquely picks out a unit vector from the origin to a_0 which in turn uniquely picks out a ray in R3 through the origin and a_0. We here work with unit vectors, since this involves no loss of generality. We write a_0, a_1, ... for points and $u(a_0)$, $u(a_1)$, ... for the corresponding unit vectors. We call a KS diagram realizable on E, if there is a 1:1 mapping of points of E, and thus of vectors in R3, to vertices of the diagram such that the orthogonality relations in the diagram -- namely, vertices joined by a straight line represent mutually orthogonal points -- are satisfied by the corresponding vectors.

We now show (see Kochen and Specker 1967: , Redhead 1987: 126):

> If vectors $u(a_0)$ and $u(a_9)$, corresponding to points a_0 and a_9 of the following ten-point KS graph Γ_1 are separated by an angle θ with $0 \leq \theta \leq \sin^{-1}(1/3)$, then Γ_1 is realizable.

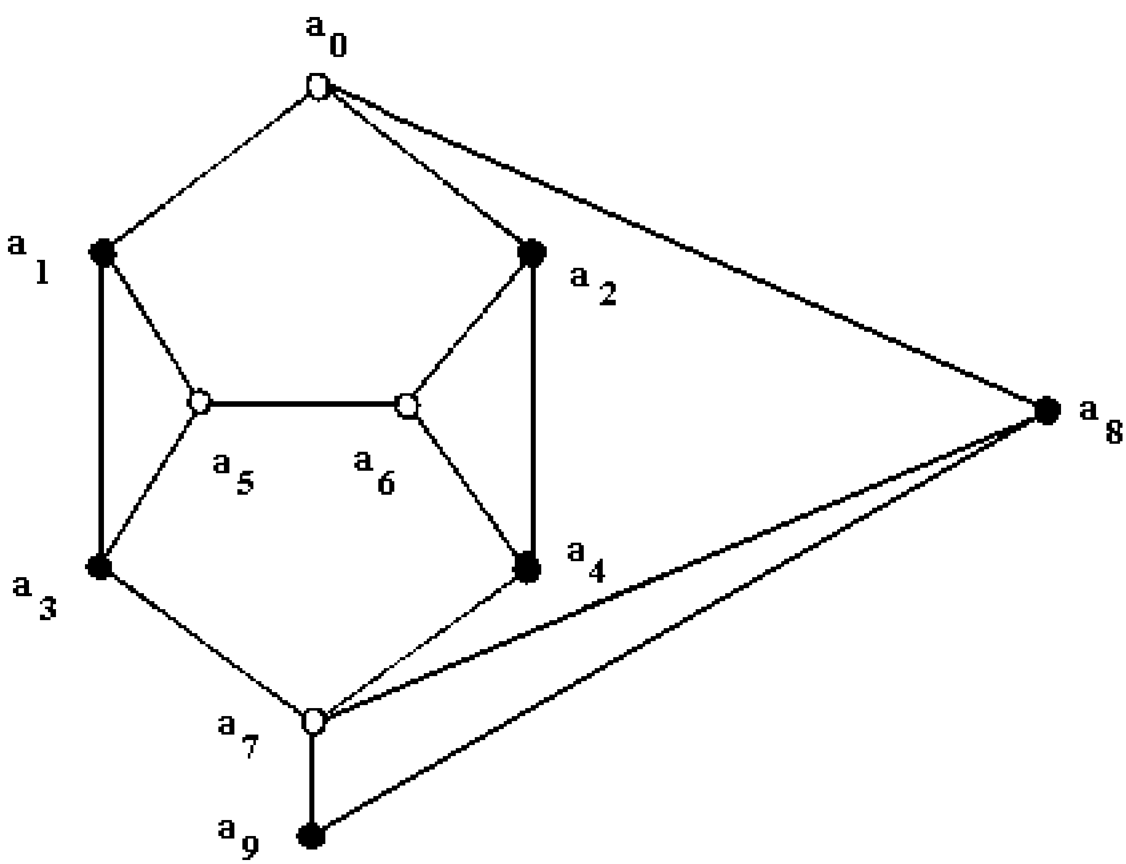


Figure 4: Ten-point KS graph Γ_1

Suppose that θ, the angle between $u(a_0)$ and $u(a_9)$, is any acute angle. Since $u(a_8)$ is orthogonal to $u(a_0)$ and $u(a_9)$, and $u(a_7)$ also is orthogonal to $u(a_9)$, $u(a_7)$ must lie in the plane defined by $u(a_0)$ and $u(a_9)$. Moreover, the direction of $u(a_7)$ can be chosen such that, if φ is the angle between $u(a_0)$ and $u(a_7)$, then $\varphi = \pi/2 - \theta$.

Now, let $u(a_5) = \mathbf{i}$ and $u(a_6) = \mathbf{k}$ and choose a third vector $\mathbf{j}$ such that $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ form a complete set of orthonormal vectors. Then $u(a_1)$, being orthogonal to $\mathbf{i}$, may be written as:

$$u(a_1) = (\mathbf{j} + x\mathbf{k})\,(1 + x^2)^{-1/2}$$

for a suitable real number x, and similarly $u(a_2)$, being orthogonal to $\mathbf{k}$, may be written as:

$$u(a_2) = (\mathbf{i} + y\mathbf{j})\,(1 + y^2)^{-1/2}$$

for a suitable real number y. But now the orthogonality relations in the diagram yield:

$$u(a_3) = u(a_5) \times u(a_1) = (-x\mathbf{j} + \mathbf{k})\,(1 + x^2)^{-1/2}$$

$$u(a_4) = u(a_2) \times u(a_6) = (y\mathbf{i} - \mathbf{j})\,(1 + y^2)^{-1/2}$$

Now, $u(a_0)$ is orthogonal to $u(a_1)$ and $u(a_2)$, so:

$$u(a_0) = u(a_1) \times u(a_2) / (\,|\,u(a_1) \times u(a_2)\,|\,) = (-xy\mathbf{i} + x\mathbf{j} - \mathbf{k})\,(1 + x^2 + x^2 y^2)^{-1/2}$$

Similarly, $u(a_7)$ is orthogonal to $u(a_3)$ and $u(a_4)$, so:

$$u(a_7) = u(a_4) \times u(a_3) / (\,|\,u(a_4) \times u(a_3)\,|\,) = (-\mathbf{i} - y\mathbf{j} - xy\mathbf{k})\,(1 + y^2 + x^2 y^2)^{-1/2}$$

Recalling now that the inner product of two unit vectors just equals the cosine of the angle between them, we get:

$$u(a_0)\,u(a_7) = \cos\varphi = xy[(1 + x^2 + x^2 y^2)\,(1 + y^2 + x^2 y^2)]^{-1/2}$$

Thus:

$$\sin\theta = xy[(1 + x^2 + x^2 y^2)\,(1 + y^2 + x^2 y^2)]^{-1/2}$$

This expression achieves a maxium value of 1/3 for $x = y = \pm 1$. Hence, the diagram is realizable, if $0 \leq$

$\theta \leq \sin^{-1}(1/3)$, or, equivalently if $0 \leq \sin\theta \leq 1/3$.

Carsten Held
cheld@uni-freiburg.de

Return to The Kochen-Specker Theorem

First posted: September 10, 2000
Last modified: September 10, 2000

Derivation of Sum Rule and Product Rule from FUNC

The three principles, in full detail, are:

> **FUNC:** Let **A** be a self-adjoint operator associated with observable A, let f: **R**→**R** be an arbitrary function, such that **f(A)** is self-adjoint operator, and let | f > be an arbitrary state; then **f(A)** is associated uniquely with an observable f(A) such that:
>
> $v(f(A))_{|\varphi\rangle} = f(v(A))_{|\varphi\rangle}$
>
> **Sum Rule:** If **A** and **B** are commuting self-adjoint operators corresponding to observables A and B, respectively, then A + B is the unique observable corresponding to the self-adjoint operator **A + B** and
>
> $v(A + B)_{|\varphi\rangle} = v(A)_{|\varphi\rangle} + v(B)_{|\varphi\rangle}$
>
> **Product Rule:** If **A** and **B** are commuting self-adjoint operators corresponding to observables A and B, respectively, then if A · B is the unique observable corresponding to the self-adjoint operator **A** · **B** and
>
> $v(AB)_{|\varphi\rangle} = v(A)_{|\varphi\rangle} \cdot v(B)_{|\varphi\rangle}$

In order to derive Sum Rule and Product Rule from FUNC, we use the following mathematical fact: Let **A** and **B** be commuting operators, then there is a maximal operator **C** and there are functions f, g such that **A** = f(**C**) and **B** = g(**C**).

So, for two commuting operators **A**, **B**:

> Since **A** = f(**C**) and **B** = g(**C**), there is a function h = f+g, such that **A + B** = h(**C**).

Therefore:

$$
\begin{aligned}
v(A + B)_{|\varphi\rangle} &= h(v(C)_{|\varphi\rangle}) && \text{(by FUNC)} \\
&= f(v(C)_{|\varphi\rangle}) + g(v(C)_{|\varphi\rangle}) \\
&= v(f(C))_{|\varphi\rangle} + v(g(C))_{|\varphi\rangle} && \text{(by FUNC)}
\end{aligned}
$$

$= \; v(A)_{|\varphi\rangle} + v(B)_{|\varphi\rangle}$ (Sum Rule)

Similarly:

Since $\mathbf{\underline{A}} = f(\mathbf{\underline{C}})$ and $\mathbf{\underline{B}} = g(\mathbf{\underline{C}})$, there is a function $k = f \cdot g$, such that $\mathbf{\underline{A}} \cdot \mathbf{\underline{B}} = k(\mathbf{\underline{C}})$.

Therefore:

$$
\begin{aligned}
v(A \cdot B)_{|\varphi\rangle} &= k(v(C)_{|\varphi\rangle}) && \text{(by FUNC)} \\
&= f(v(C)_{|\varphi\rangle}) \cdot g(v(C)_{|\varphi\rangle}) && \\
&= v(f(C))_{|\varphi\rangle} \cdot v(g(C))_{|\varphi\rangle} && \text{(by FUNC)} \\
&= v(A)_{|\varphi\rangle} \cdot v(B)_{|\varphi\rangle} && \text{(Product Rule)}
\end{aligned}
$$

Carsten Held
cheld@uni-freiburg.de

Return to The Kochen-Specker Theorem

First posted: September 10, 2000
Last modified: September 10, 2000

Stanford Encyclopedia of Philosophy
Supplement to The Kochen-Specker Theorem

Derivation of STAT FUNC

The result is proved for a pure state and a non-degenerate discrete observable A with eigenvalues a_i.

We first rewrite the statistical algorithm for projection operators:

$$\textbf{(1)}\quad \text{prob}\,(v\,(A)^{|\psi>} = a_k) = \text{Tr}\,(P_{|a_k>} \cdot P_{|\psi>})$$

For an arbitrary function f: **R** → **R** (where **R** is the set of real numbers) we define the function of an observable A as:

$$f(A) =_{\text{Def}} \Sigma_i\, f(a_i)\, P_{|a_i>}$$

Moreover, we introduce the characteristic function χ_a of a number a as:

$$\begin{aligned} \chi_a(x) &= 1 \text{ for } x = a \\ &= 0 \text{ for } x \neq a \end{aligned}$$

As a result, we can rewrite a project operator $P_{|a_k>}$ as:

$$\textbf{(2)}\quad P_{|a_k>} = \chi_{a_k}(A)$$

and thus the statistical algorithm as:

$$\text{prob}\,(v\,(A)^{|\psi>} = a_k) = \text{Tr}\,(\chi_{a_k}(A) \cdot P_{|\psi>})$$

We also use a simple mathematical property of characteristic functions:

$$\chi_a(f(x)) = \chi_{f^{-1}(a)}(x)$$

whence we can also write:

(3) $\chi_a(f(A)) = \chi_{f^{-1}(a)}(A)$

Then:

$$\begin{aligned}
\text{prob}\,(v(f(A))^{|\phi>} = b) &= \text{Tr}\,(P_{|b>} \cdot P_{|\phi>}) && \text{(by (1))} \\
&= \text{Tr}\,(\chi_b\,(f(A)) \cdot P_{|\phi>}) && \text{(by (2))} \\
&= \text{Tr}\,((\chi_{f^{-1}(b)}\,(A) \cdot P_{|\phi>}) && \text{(by (3))} \\
&= \text{Tr}\,(P_{|f^{-1}(b)>} \cdot P_{|\phi>}) && \text{(by (2))} \\
&= \text{prob}\,(v\,(A)^{|\phi>} = f^{-1}(b)) && \text{(by (1))}
\end{aligned}$$

Hence:

$$\text{prob}\,(v(f(A))^{|\phi>} = b) = \text{prob}\,(v\,(A)^{|\phi>} = f^{-1}(b))$$

Now since

$$v\,(A) = f^{-1}(b) \quad \Leftrightarrow \quad f\,(v\,(A)) = b$$

$$\text{prob}\,(v(f(A))^{|\phi>} = b) = \text{prob}\,(f(v\,(A))^{|\phi>} = b)$$

which is STAT FUNC.

Carsten Held
cheld@uni-freiburg.de

Return to The Kochen-Specker Theorem

First posted: September 10, 2000
Last modified: September 10, 2000

STAT FUNC from FUNC

Indeed, STAT FUNC follows immediately from FUNC. Given A, f, and $|\varphi\rangle$ as defined in FUNC, we have:

$$v(f(A))_{|\varphi\rangle} = f(v(A))_{|\varphi\rangle},$$

where f(A) is a new observable. Now, by the statistical algorithm:

> **STAT FUNC:** Given A, f, and $|\varphi\rangle$ as defined in FUNC, then, for an arbitrary real number b:
>
> $$\text{prob}[v(f(A))_{|\varphi\rangle} = b] = \text{prob}[f(v(A))_{|\varphi\rangle} = b]$$

Carsten Held
cheld@uni-freiburg.de

Return to The Kochen-Specker Theorem

First posted: September 10, 2000
Last modified: September 10, 2000

Quantum Logic and Quantum Probability

At its core, quantum mechanics can be regarded as a non-classical probability calculus resting upon a non-classical propositional logic. More specifically, in quantum mechanics each probability-bearing proposition of the form "the value of physical quantity *A* lies in the range *B*" is represented by a projection operator on a Hilbert space **H**. These form a non-Boolean -- in particular, non-distributive -- orthocomplemented lattice. Quantum-mechanical states correspond exactly to probability measures (suitably defined) on this lattice.

What are we to make of this? Some have argued that the empirical success of quantum mechanics calls for a revolution in logic itself. This view is associated with the demand for a realistic interpretation of quantum mechanics, i.e., one not grounded in any primitive notion of measurement. Against this, there is a long tradition of interpreting quantum mechanics operationally, that is, as being precisely a theory of measurement. On this latter view, it is not surprising that a "logic" of measurement-outcomes, in a setting where not all measurements are compatible, should prove not to be Boolean. Rather, the mystery is why it should have the *particular* non-Boolean structure that it does in quantum mechanics. A substantial literature has grown up around the programme of giving some independent motivation for this structure -- ideally, by deriving it from more primitive and plausible axioms governing a generalized probability theory.

1. Quantum Mechanics as a Probability Calculus

It is uncontroversial (though remarkable) that the formal apparatus of quantum mechanics reduces neatly to a generalization of classical probability in which the role played by a Boolean algebra of events in the latter is taken over by the "quantum logic" of projection operators on a Hilbert space.[1] Moreover, the usual statistical interpretation of quantum mechanics asks us to take this generalized quantum probability theory quite literally -- that is, not as merely a formal analogue of its classical counterpart, but as a genuine doctrine of chances. In this section, I survey this quantum probability theory and its supporting quantum logic.[2]

> [For further background on Hilbert spaces, see Supplement 1: The Basic Theory of Hilbert Spaces. For further background on ordered sets and lattices, see Supplement 2: The Basic Theory of Ordering Relations. Concepts and results explained these supplements will be used freely in what follows.]

Quantum Probability in a Nutshell

The quantum-probabilistic formalism, as developed by von Neumann [1932], assumes that each physical system is associated with a (separable) Hilbert space **H**, the unit vectors of which correspond to possible physical *states* of the system. Each "observable" real-valued random quantity is represented by a self-adjoint operator A on **H**, the spectrum of which is the set of possible values of A. If u is a unit vector in the domain of A, representing a state, then the expected value of the observable represented by A in this state is given by the inner product $\langle Au,u \rangle$. The observables represented by two operators A and B are commensurable iff A and B commute, i.e., $AB = BA$. (For further discussion, see the entry on Quantum Mechanics.)

The "Logic" of Projections

As stressed by von Neumann, the {0,1}-valued observables may be regarded as encoding propositions about -- or, to use his phrasing, properties of -- the state of the system. It is not difficult to show that a self-adjoint operator P with spectrum contained in the two-point set {0,1} must be a projection; i.e., $P^2 = P$. Such operators are in one-to-one correspondence with the closed subspaces of **H**. Indeed, if P is a projection, its range is closed, and any closed subspace is the range of a unique projection. If u is any unit vector, then $\langle Pu,u \rangle = ||Pu||^2$ is the expected value of the corresponding observable in the state represented by u. Since this is {0,1}-valued, we can interpret this as the *probability* that a measurement of the observable will produce the "affirmative" answer 1. In particular, the affirmative answer will have probability 1 if and only if $Pu = u$; that is, u lies in the range of P. Von Neumann concludes that

> … the relation between the properties of a physical system on the one hand, and the projections on the other, makes possible a sort of logical calculus with these. However, in

contrast to the concepts of ordinary logic, this system is extended by the concept of "simultaneous decidability" which is characteristic for quantum mechanics [1932, p. 253].

Let's examine this "logical calculus" of projections. Ordered by set-inclusion, the closed subspaces of **H** form a complete lattice, in which the meet (greatest lower bound) of a set of subspaces is their intersection, while their join (least upper bound) is the closed span of their union. Since a typical closed subspace has infinitely many complementary closed subspaces, this lattice is not distributive; however, it is orthocomplemented by the mapping

$$\mathbf{M} \rightarrow \mathbf{M}\perp = \{v \in \mathbf{H} \mid \forall u \in \mathbf{M}(\langle v,u \rangle = 0)\}.$$

In view of the above-mentioned one-one correspondence between closed subspaces and projections, we may impose upon the set $L(\mathbf{H})$ the structure of a complete orthocomplemented lattice, defining $P \leq Q$, where $ran(P) \subseteq ran(Q)$ and $P' = 1 - P$ (so that $ran(P') = ran(P)\perp$). It is straightforward that $P \leq Q$ just in case $PQ = QP = P$. More generally, if $PQ = QP$, then $PQ = P \wedge Q$, the meet of P and Q in $L(\mathbf{H})$; also in this case their join is given by $P \vee Q = P+Q - PQ$.

1.1 Lemma:
Let P and Q be projection operators on the Hilbert space **H**. The following are equivalent:

a. $PQ = QP$
b. The sublattice of $L(\mathbf{H})$ generated by P, Q, P' and Q' is Boolean
c. P, Q lie in a common Boolean sub-ortholattice of $L(\mathbf{H})$.

Adhering to the idea that commuting observables -- in particular, projections -- are simultaneously measurable, we conclude that the members of a Boolean "block" (that is, a Boolean sub-ortholattice) of $L(\mathbf{H})$ are simultaneously testable. This suggests that we can maintain a classical logical interpretation of the meet, join and orthocomplement as applied to commuting projections.

Probability Measures and Gleason's Theorem

The foregoing discussion motivates the following. Call projections P and Q *orthogonal*, and write $P \perp Q$ iff $P \leq Q'$. Note that $P \perp Q$ iff $PQ = QP = 0$. If P and Q are orthogonal projections, then their join is simply their sum; traditionally, this is denoted $P \oplus Q$. We denote the identity mapping on **H** by **1**.

1.2 Definition:
A (countably additive) *probability measure* on $L(\mathbf{H})$ is a mapping $\mu : L \rightarrow [0,1]$ such that $\mu(\mathbf{1}) = 1$ and, for any sequence of pair-wise orthogonal projections P_i, $i = 1,2,...$

$$\mu(\oplus_i P_i) = \Sigma_i \mu(P_i)$$

Here is one way in which we can manufacture a probability measure on $L(\mathbf{H})$. Let u be a unit vector of $\mathbf{H}$, and set $\mu_u(P) = \langle Pu,u \rangle$. This gives the usual quantum-mechanical recipe for the probability that P will have value 1 in the state u. Note that we can also express μ_u as $\mu_u(P) = Tr(P\,P_u)$, where P_u is the one-dimensional projection associated with the unit vector u.

More generally, if μ_i, i=1,2,…, are probability measures on $L(\mathbf{H})$, then so is any "mixture", or convex combination $\mu = \Sigma_i\, t_i \mu_i$ where $0 \leq t_i \leq 1$ and $\Sigma_i\, t_i = 1$. Given any sequence $u_1, u_2,\ldots$, of unit vectors, let $\mu_i = \mu_{u_i}$ and let $P_i = P_{u_i}$. Forming the operator

$$W = t_1 P_1 + t_2 P_2 + \ldots ,$$

one sees that

$$\mu(P) = t_1 Tr(P\,P_1) + t_2 Tr(P\,P_2) + \ldots = Tr(WP)$$

An operator expressible in this way as a convex combination of one-dimensional projections in is called a *density operator*. Thus, every density operator W gives rise to a countably additive probability measure on $L(\mathbf{H})$. The following striking converse, due to A. Gleason [1957], shows that the theory of probability measures on $L(\mathbf{H})$ is co-extensive with the theory of (mixed) quantum mechanical states on $\mathbf{H}$:

> **1.3 Gleason's Theorem**:
> Let $\mathbf{H}$ have dimension > 2. Then every countably additive probability measure on $L(\mathbf{H})$ has the form $\mu(P) = Tr(WP)$, for a density operator W on $\mathbf{H}$.

An important consequence of Gleason's Theorem is that $L(\mathbf{H})$ does not admit any probability measures having only the values 0 and 1. To see this, note that for any density operator W, the mapping $u \rightarrow \langle Wu,u \rangle$ is continuous on the unit sphere of $\mathbf{H}$. But since the latter is connected, no continuous function on it can take only the two values 0 and 1. This result is often taken to rule out the possibility of 'hidden variables' -- an issue taken up in more detail in section 6.

The Reconstruction of QM

From the single premise that the "experimental propositions" associated with a physical system are encoded by projections in the way indicated above, one can reconstruct the rest of the formal apparatus of quantum mechanics. The first step, of course, is Gleason's theorem, which tells us that probability measures on $L(\mathbf{H})$ correspond to density operators. There remains to recover, e.g., the representation of "observables" by self-adjoint operators, and the dynamics (unitary evolution). The former can be recovered with the help of the Spectral theorem and the latter with the aid of a deep theorem of E. Wigner on the projective representation of groups. See also R. Wright [1980]. A detailed outline of this reconstruction (which involves some distinctly non-trivial mathematics) can be found in the book of

Varadarajan [1985]. The point to bear in mind is that, once the quantum-logical skeleton $L(\mathbf{H})$ is in place, the remaining statistical and dynamical apparatus of quantum mechanics is essentially fixed. In this sense, then, quantum mechanics -- or, at any rate, its mathematical framework -- *reduces to* quantum logic and its attendant probability theory.

2. Interpretations of Quantum Logic

The reduction of QM to probability theory based on $L(\mathbf{H})$ is mathematically compelling, but what does it tell us about QM --- or, assuming QM to be a correct and complete physical theory, about the world? How, in other words, are we to interpret the quantum logic $L(\mathbf{H})$? The answer will turn on how we unpack the phrase, freely used above,

> (*) The value of the observable A lies in the range B.

One possible reading of (*) is *operational*: "measurement of the observable A would yield (or will yield, or has yielded) a value in the set B ". On this view, projections represent statements about the possible results of measurements. This sits badly with realists of a certain stripe, who, shunning reference to 'measurement', prefer to understand (*) as a *property ascription*: "the system has a certain categorical property, which corresponds to the observable A having, independently of any measurement, a value in the set B". (One must be careful in how one understands this last phrase, however: construed incautiously, it seems to posit a hidden-variables interpretation of quantum mechanics of just the sort ruled out by Gleason's Theorem. I will have more to say about this below.)

Realist Quantum Logic

The interpretation of projection operators as representing the properties of a physical system is already explicit in von Neumann's *Grundlagen.*. However, the logical operations discussed there apply only to commuting projections, which are identified with simultaneously decidable propositions. In [1936] von Neumann and Birkhoff took a step further, proposing to interpret the lattice-theoretic meet and join of projections as their conjunction and disjunction, *whether or not* they commute. Immediately this proposal faces the problem that the lattice $L(\mathbf{H})$ is not distributive, making it impossible to give these 'quantum' connectives a truth-functional interpretation. Undaunted, von Neumann and Birkhoff suggested that the empirical success of quantum mechanics as a framework for physics casts into doubt the universal validity of the distributive laws of propositional logic. Their phrasing remains cautious:

> Whereas logicians have usually assumed that properties … of negation were the ones least able to withstand a critical analysis, the study of mechanics points to the distributive identities … as the weakest link in the algebra of logic. [1937, p. 839]

In the 1960s and early 1970s, this thesis was advanced rather more aggressively by a number of authors, including especially David Finkelstein and Hilary Putnam, who argued that quantum mechanics requires

a revolution in our understanding of logic *per se*. According to Putnam [1968], “Logic is as empirical as geometry. … We live in a world with a non-classical logic.”

For Putnam, the elements of $L(\mathbf{H})$ represent categorical properties that an object possesses, or does not, independently of whether or not we look. Inasmuch as this picture of physical properties is confirmed by the empirical success of quantum mechanics, we must, on this view, accept that the way in which physical properties actually hang together is not Boolean. Since logic is, for Putnam, very much the study of how physical properties actually hang together, he concludes that classical logic is simply mistaken: the distributive law is not universally valid.

Classically, if S is the set of states of a physical system, then *every* subset of S corresponds to a categorical property of the system, and vice versa. In quantum mechanics, the state space is the (projective) unit sphere $S = S(\mathbf{H})$ of a Hilbert space. However, not all subsets of S correspond to quantum-mechanical properties of the system. The latter correspond only to subsets of the special form $S \cap \mathbf{M}$, for $\mathbf{M}$ a closed linear subspace of $\mathbf{H}$. In particular, only subsets of this form are assigned probabilities. This leaves us with two options. One is to take only these special properties as ‘real’ (or ‘physical’, or ‘meaningful’), regarding more general subsets of S as corresponding to no real categorical properties at all. The other is to regard the ‘quantum’ properties as a small subset of the set of all physically (or at any rate, metaphysically) reasonable, but not necessarily *observable*, properties of the system. On this latter view, the set of *all* properties of a physical system is entirely classical in its logical structure, but we decline to assign probabilities to the non-observable properties.[3]

This second position, while certainly not inconsistent with realism *per se*, turns upon a distinction involving a notion of "observation", "measurement", "test", or something of this sort -- a notion that realists are often at pains to avoid in connection with fundamental physical theory. Of course, any realist account of a statistical physical theory such as quantum mechanics will ultimately have to render up some explanation of how measurements are supposed to take place. That is, it will have to give an account of which physical interactions between "object" and "probe" systems count as measurements, and of how these interactions cause the probe system to evolve into final ‘outcome-states’ that correspond to -- and have the same probabilities as -- the outcomes predicted by the theory. This is notorious *measurement problem* .

In fact, Putnam advanced his version of quantum-logical realism as offering a (radical) dissolution of the measurement problem: According to Putnam, the measurement problem (and indeed every other quantum-mechanical "paradox") arises through an improper application of the distributive law, and hence *disappears* once this is recognized. This proposal, however, is widely regarded as mistaken.

As mentioned above, realist interpretations of quantum mechanics must be careful in how they construe the phrase "the observable A has a value in the set B ". The simplest and most traditional proposal -- often dubbed the "eigenstate-eigenvalue link" (Find 1973) -- is that (*) holds if and only if a measurement of A yields a value in the set B with certainty, i.e., with (quantum-mechanical!) probability 1. While this certainly gives a realist interpretation of (*) [4], it does not provide a solution to the measurement

problem. Indeed, we can use it to give a sharp formulation of that problem: even though *A* is certain to yield a value in *B* when measured, unless the quantum state is an eigenstate of the measured observable *A*, the system does not possess any categorical property corresponding to *A*'s having a specific value in the set *B*. Putnam seems to assume that a realist interpretation of (*) should consist in assigning to *A* some unknown value within *B*, for which quantum mechanics yields a non-trivial probability. However, an attempt to make such assignments simultaneously for all observables runs afoul of Gleason's Theorem. [5]

Operational Quantum Logic

If we put aside scruples about 'measurement' as a primitive term in physical theory, and accept a principled distinction between 'testable' and non-testable properties, then the fact that *L*(**H**) is not Boolean is unremarkable, and carries no implication about logic *per se*. Quantum mechanics is, on this view, a theory about the possible statistical distributions of outcomes of certain measurements, and its non-classical 'logic' simply reflects the fact that not all observable phenomena can be observed simultaneously. Because of this, the set of probability-bearing events (or propositions) is *less* rich than it would be in classical probability theory, and the set of possible statistical distributions, accordingly, less tightly constrained. That some 'non-classical' probability distributions allowed by this theory are actually manifested in nature is perhaps surprising, but in no way requires any deep shift in our understanding of logic or, for that matter, of probability.

This is hardly the last word, however. Having accepted all of the above, there still remains the question of *why* the logic of measurement outcomes should have the very special form *L*(**H**), and never anything more general.[6] This question entertains the idea that the formal structure of quantum mechanics may be *uniquely determined* by a small number of reasonable assumptions, together perhaps with certain manifest regularities in the observed phenomena. This possibility is already contemplated in von Neumann's *Grundlagen* (and also his later work in continuous geometry), but first becomes explicit -- and programmatic -- in the work of George Mackey [1957, 1963]. Mackey presents a sequence of six axioms, framing a very conservative generalized probability theory, that underwrite the construction of a 'logic' of experimental propositions, or, in his terminology, 'questions', having the structure of a sigma-orthomodular poset. The outstanding problem, for Mackey, was to explain why this poset *ought to* be isomorphic to *L*(**H**):

> Almost all modern quantum mechanics is based implicitly or explicitly on the following assumption, which we shall state as an axiom:
>
> > Axiom VII: The partially ordered set of all questions in quantum mechanics is isomorphic to the partially ordered set of all closed subspaces of a separable, infinite dimensional Hilbert space.
>
> This axiom has rather a different character from Axioms I through VI. These all had some degree of physical naturalness and plausibility. Axiom VII seems entirely ad hoc. Why do

> we make it? Can we justify making it? … Ideally, one would like to have a list of physically plausible assumptions from which one could deduce Axiom VII. Short of this one would like a list from which one could deduce a set of possibilities for the structure … all but one of which could be shown to be inconsistent with suitably planned experiments. [19, pp. 71-72]

Since Mackey's writing there has grown up an extensive technical literature exploring variations on his axiomatic framework in an effort to supply the missing assumptions. The remainder of this article presents a brief survey of the current state of this project.

3. Generalized Probability Theory

Rather than restate Mackey's axioms verbatim, I shall paraphrase them in the context of an approach to generalized probability theory due to D. J. Foulis and C. H. Randall having -- among the many more or less homologous approaches available[7] -- certain advantages of simplicity and flexibility. References for this section are [Foulis, Greechie and Ruttimann 1992, Foulis, Piron and Randall 1983, Foulis and Randall 1982, Randall and Foulis 1983; see also Gudder 1985 and Wilce 2000b for surveys.]

Discrete Classical Probability Theory

It will be helpful to begin with a review of classical probability theory**.** In its simplest formulation, classical probability theory deals with a (discrete) set E of mutually exclusive outcomes, as of some measurement, experiment, etc., and with the various *probability weights* that can be defined thereon --- that is, with mappings $\omega : E \rightarrow [0,1]$ summing to 1 over E.[8]

Notice that the set $\Delta(E)$ of all probability weights on E is *convex*, in that, given any sequence $\omega_1, \omega_2, \ldots$ of probability weights and any sequence $t_1, t_2, \ldots$ of non-negative real numbers summing to one, the convex sum or 'mixture' $t_1\omega_1 + t_2\omega_2 + \ldots$ (taken pointwise on E) is again a probability weight. The extreme points of this convex set are exactly the "point-masses" $\delta(x)$ associated with the outcomes $x \in E$:

$$\delta(x)(y) = 1 \text{ if } x = y, \text{ and } 0 \text{ otherwise.}$$

Thus, $\Delta(E)$ is a *simplex*: each point $\omega \in \Delta(E)$ is representable in a unique way as a convex combination of extreme points, namely:

$$\omega = \sum \omega(x)\delta(x)$$

We need also to recall the concept of a *random variable*. If E is an outcome set and V, some set of 'values' (real numbers, pointer-readings, or what not), a *V-valued random variable* is simply a mapping f

: $E \to V$. The heuristic (but it need only be taken as that) is that one ‘measures’ the random variable f by ‘performing’ the experiment represented by E and, upon obtaining the outcome $x \in E$, recording $f(x)$ as the measured value. Note that if V is a set of real numbers, or, more generally, a subset of a vector space, we may define the *expected value* of f in a state $\omega \in \Delta(E)$ by:

$$E(f,\omega) = \sum_{x \in E} f(x)\omega(x).$$

Test Spaces

A very natural direction in which to generalize discrete classical probability theory is to allow for a multiplicity of outcome-sets, each representing a different ‘experiment’. To formalize this, let’s agree that a *test space* is a non-empty collection A of non-empty sets $E,F,\ldots$, each construed as a discrete outcome-set as in classical probability theory. Each set $E \in \mathcal{A}$ is called a *test*. The set $X = \bigcup\mathcal{A}$ of all outcomes of all tests belonging to $\mathcal{A}$ is called the *outcome space* of $\mathcal{A}$. Notice that we allow distinct tests to overlap, i.e., to have outcomes in common.[9]

If $\mathcal{A}$ is a test space with outcome-space X, a *state* on $\mathcal{A}$ is a mapping $\omega : X \to [0,1]$ such that $\Sigma_{x \in E}\, \omega(x) = 1$ for every test $E \in \mathcal{A}$. Thus, a state is a consistent assignment of a probability weight to each test -- consistent in that, where two distinct tests share a common outcome, the state assigns that outcome the same probability whether it is secured as a result of one test or the other. (This may be regarded as a normative requirement on the outcome-identifications implicit in the structure of $\mathcal{A}$: if outcomes of two tests are not equiprobable in all states, they ought not to be identified.) The set of all states on $\mathcal{A}$ is denoted by $\Omega(\mathcal{A})$. This is a convex set, but in contrast to the situation in discrete classical probability theory, it is generally not a simplex.

The concept of a random variable admits several generalizations to the setting of test spaces. Let us agree that a *simple (real-valued) random variable* on a test space $\mathcal{A}$ is a mapping $f : E \to \mathbf{R}$ where E is a test in $\mathcal{A}$. We define the *expected value* of f in a state $\omega \in \Omega(\mathcal{A})$ in the obvious way, namely, as the expected value of f with respect to the probability weight obtained by restricting ω to E (provided, of course, that this expected value exists). One can go on to define more general classes of random variables by taking suitable limits (for details, see [Younce, 1987]).

In classical probability theory (and especially in classical statistics) one usually focuses, not on the set of all possible probability weights, but on some designated subset of these (e.g., those belonging to a given family of distributions). Accordingly, by a *probabilistic model*, I mean pair $(\mathcal{A},\Delta)$ consisting of a test space $\mathcal{A}$ and a designated set of states $\Delta \subseteq \Omega(\mathcal{A})$ on $\mathcal{A}$. I’ll refer to $\mathcal{A}$ as the *test space* and to Δ as the *state space* of the model.

I’ll now indicate how this framework can accommodate both the usual measure-theoretic formalism of full-blown classical probability theory and the Hilbert-space formalism of quantum probability theory.

Kolmogorovian Probability Theory

Let S be a set, construed for the moment as the state-space of a physical system, and let Σ be a sigma-field of subsets of S. We can regard each partition E of S into countably many pair-wise disjoint Σ-measurable subsets as representing a 'coarse-grained' approximation to an imagined perfect experiment that would reveal the state of the system. Let $\mathcal{A}$ be the test space consisting of all such partitions. Note that the outcome set for $\mathcal{A}$ is the set $X = B - \{\varnothing\}$ of non-empty Σ-measurable subsets of S. Evidently, the probability weights on $\mathcal{A}$ correspond exactly to the countably additive probability measures on Σ.

Quantum Probability Theory

Let **H** denote a complex Hilbert space and let $\mathcal{A}$ denote the collection of (unordered) orthonormal bases of **H**. Thus, the outcome-space X of $\bigcup\mathcal{A}$ will be the unit sphere of **H**. Note that if u is any unit vector of **H** and $E \in \mathcal{A}$ is any orthonormal basis, we have

$$\sum_{x \in E} |\langle u,x\rangle|^2 = \|u\|^2 = 1$$

Thus, each unit vector of **H** determines a probability weight on $\mathcal{A}$. Quantum mechanics asks us to take this literally: any 'maximal' discrete quantum-mechanical observable is modeled by an orthonormal basis, and any pure quantum mechanical state, by a unit vector in exactly this way. Conversely, every orthonormal basis and every unit vector are understood to correspond to such a measurement and such a state.

Gleason's theorem can now be invoked to identify the states on $\mathcal{A}$ with the density operators on **H**: to each state ω in $\Omega(\mathcal{A}_{\mathbf{H}})$ there corresponds a unique density operator W such that, for every unit vector x of **H**, $\omega(x) = \langle Wx,x\rangle = Tr(WP_x)$, P_x being the one-dimensional projection associated with x. Conversely, of course, every such density operator defines a unique state by the formula above. We can also represent simple real-valued random variables operator-theoretically. Each bounded simple random variable f gives rise to a bounded self-adjoint operator $A = \Sigma_{x \in E} f(x)P_x$. The spectral theorem tells us that every self-adjoint operator on **H** can be obtained by taking suitable limits of operators of this form.

4. Logics associated with probabilistic models

Associated with any statistical model $(\mathcal{A},\Delta)$ are several partially ordered sets, each of which has some claim to the status of an 'empirical logic' associated with the model. In this section, I'll discuss two: the so-called *operational logic* $\Pi(\mathcal{A})$ and the *property lattice* $\mathbf{L}(\mathcal{A},\Delta)$. Under relatively benign conditions on $\mathcal{A}$, the former is an *orthoalgebra*. The latter is always a complete lattice, and under plausible further assumptions, atomic. Moreover, there is a natural order preserving mapping from Π to $\mathbf{L}$. This is not generally an order-isomorphism, but when it is, we obtain a complete orthomodular lattice, and thus come a step closer to the projection lattice of a Hilbert space.

Operational Logics

If $\mathcal{A}$ is a test space, an $\mathcal{A}$-*event* is a set of $\mathcal{A}$-outcomes that is contained in some test. In other words, an $\mathcal{A}$-event is simply an event in the classical sense for any one of the tests comprising $\mathcal{A}$. Now, if A and B are two $\mathcal{A}$-events, we say that A and B are *orthogonal*, and write $A \perp B$, if they are disjoint and their union is again an event. We say that two orthogonal events are *complements* of one another if their union is a test. We say that events A and B are *perspective*, and write $A \sim B$, if they share any common complement. (Notice that any two tests E and F are perspective, since they are both complementary to the empty event.)

> **4.1 Definition**:
> A test space $\mathcal{A}$ is said to be *algebraic* if for all events A, B, C of $\mathcal{A}$, $A \sim B$ and $B \perp C$ implies $A \perp C$.

While it is possible to construct perfectly plausible examples of test spaces that are not algebraic, most of test spaces that one encounters 'in nature' -- including the Borel and quantum test spaces described in the preceding section -- do seem to enjoy this property. The more important point is that, as an axiom, algebraicity is relatively benign, in the sense that many test spaces can be 'completed' to become algebraic. In particular, if every outcome has probability greater than .5 in at least one state, then $\mathcal{A}$ is contained in an algebraic test space $\mathcal{B}$ having the same outcomes and the same states as $\mathcal{A}$. (See [Gudder, 1985] for details).

Suppose now that $\mathcal{A}$ is algebraic. It is easy to see that the relation ~ of perspectivity is then an equivalence relation on the set of $\mathcal{A}$-events. More than this, if $\mathcal{A}$ is algebraic, then ~ is a *congruence* for the partial binary operation of forming unions of orthogonal events: in other words, $A \sim B$ and $B \perp C$ imply that $A \cup C \sim B \cup C$ for all $\mathcal{A}$-events A, B, and C.

Let $\Pi(\mathcal{A})$ be the set of equivalence classes of $\mathcal{A}$-events under perspectivity, and denote the equivalence class of an event A by $p(A)$; we then have a natural partial binary operation on $\Pi(\mathcal{A})$ defined by $p(A) \oplus p(B) = p(A \cup B)$ for orthogonal events A and B. Setting $0 := p(\varnothing)$ and $1 := p(E)$, E any member of $\mathcal{A}$, we obtain a partial-algebraic structure $(\Pi(\mathcal{A}), \oplus, 0, 1)$, called the *logic* of $\mathcal{A}$. This satisfies the following conditions:

a. $\oplus$ is associative and commutative:
   - If $a \oplus (b \oplus c)$ is defined, so is $(a \oplus b) \oplus c$, and the two are equal
   - If $a \oplus b$ is defined, so is $b \oplus a$, and the two are equal.

b. $0 \oplus a = a$, for every $a \in \mathbf{L}$

c. For every $a \in \mathbf{L}$, there exists a unique $a' \in \mathbf{L}$ with $a \oplus a' = 1$

d. $a \oplus a$ exists only if $a = 0$

We may now define:

4.2 Definition:
A structure (**L**,$\oplus$,0,1) satisfying conditions (a)-(d) above is called an *orthoalgebra*.

Thus, the logic of an algebraic test space is an orthoalgebra. One can show that, conversely, every orthoalgebra arises as the logic $\Pi(\mathcal{A})$ of an algebraic test space $\mathcal{A}$ (Golfin [1988]). Note that non-isomorphic test spaces can have isomorphic logics.

Orthocoherence

Any orthoalgebra **L** is partially ordered by the relation $a \leq b$ iff $b = a \oplus c$ for some $c \perp a$. Relative to this ordering, the mapping $a \rightarrow a'$ is an orthocomplementation and $a \perp b$ iff $a \leq b'$. It can be shown that $a \oplus b$ is always a minimal upper bound for *a* and *b*, but it is generally not the *least* upper bound. Indeed, we have the following [ref]:

4.3 Lemma:
For an orthoalgebra (**L**,$\oplus$,0,1), the following are equivalent:

a. $a \oplus b = a \vee b$, for all *a*, *b* in **L**
b. If $a \oplus b$, $b \oplus c$, and $c \oplus a$ all exist, then so does $a \oplus b \oplus c$
c. The orthoposet (**L**,$\leq$,$'$) is *orthomodular*, i.e., for all *a*, $b \in L$, if $a \leq b$ then $(b \wedge a') \vee a$ exists and equals *b*.

An orthoalgebra satisfying condition (b) is said to be *orthocoherent*. In other words: an orthoalgebra is ortho-coherent if and only if finite pair-wise summable subsets of **L** are jointly summable. The lemma tells us that every orthocoherent orthoalgebra is, *inter alia*, an orthomodular poset. Conversely, an orthocomplemented poset is orthomodular iff $a \oplus b = a \vee b$ is defined for all pairs with $a \leq b'$ and the resulting partial binary operation is associative -- in which case the resulting structure (**L**,$\oplus$,0,1) is an orthocoherent orthoalgebra, the canonical ordering on which agrees with the given ordering on **L**. Thus, orthomodular posets (the framework for Mackey's version of quantum logic) are equivalent to orthocoherent orthoalgebras.

Some version of orthocoherence was taken by Mackey and many of his successors as an axiom. (It appears, in an infinitary form, as Mackey's axiom V; a related but stronger condition appears in the definition of a partial Boolean algebra in the work of Kochen and Specker [1965].) However, it is quite easy to construct simple model test spaces, having perfectly straightforward -- even classical -- interpretations, the logics of which are not orthocoherent. As far as I know, there has never been given any entirely compelling reason for regarding orthocoherence as an essential feature of all reasonable physical models. Moreover, certain apparently quite well-motivated constructions that one wants to perform with test spaces tend to destroy orthocoherence (see Section 7).

Lattices of Properties

The decision to accept measurements and their outcomes as primitive concepts in our description of physical systems does not mean that we must forgo talk of the physical properties of such a system. Indeed, such talk is readily accommodated in our present formalism.[10] In the approach we have been pursuing, a physical system is represented by a probabilistic model $(\mathcal{A},\Delta)$, and the system's states are identified with the probability weights in Δ. Classically, *any* subset Γ of the state-space Δ corresponds to a categorical property of the system. However, in quantum mechanics, and indeed even classically, not every such property will be testable (or "physical"). (In quantum mechanics, only subsets of the state-space corresponding to closed subspaces of the Hilbert space are testable; in classical mechanics, one usually takes only, e.g., Borel sets to correspond to testable properties: the difference is that the testable properties in the latter case happen still to form a Boolean algebra of sets, where in the former case, they do not.)

One way to frame this distinction is as follows. The *support* of a set of states $\Gamma \subseteq \Delta$ is the set

$$S(\Gamma) = \{x \in X \mid \exists \omega \in \Gamma(\omega(x) > 0)\}$$

of outcomes that are possible when the property Γ obtains. There is a sense in which two properties are empirically indistinguishable if they have the same support: we cannot distinguish between them by means of a single execution of a single test. We might therefore wish to identify physical properties with classes of physically indistinguishable classical properties, or, equivalently, with their associated supports. However, if we wish to adhere to the programme of representing physical properties as subsets (rather than as equivalence-classes of subsets) of the state-space, we can do so, as follows. Define a mapping $F : \mathcal{P}(X) \to \mathcal{P}(\Delta)$ by $F(J) = \{\omega \in \Delta \mid S(\omega) \subseteq J\}$. The mapping $\Gamma \to F(S(\Gamma))$ is then a *closure operator* on $\mathcal{P}(\Delta)$, and the collection of closed sets (that is, the range of F) is a complete lattice of sets, closed under arbitrary intersection.[11] Evidently, classical properties -- subsets of Δ -- have the same support iff they have the same closure, so we may identify physical properties with closed subsets of the state-space:

> **4.4 Definition**:
> The *property lattice* of the model $(\mathcal{A},\Delta)$ is the complete lattice $\mathbf{L} = \mathbf{L}(\mathcal{A},\Delta)$ of all subsets of Δ of the form $F(J)$, J any set of outcomes. .[12]

We now have two different 'logics' associated with an entity $(\mathcal{A},\Delta)$ with $\mathcal{A}$ algebraic: a 'logic' $\Pi(\mathcal{A})$ of experimental propositions that is an orthoalgebra, but generally not a lattice, and a 'logic' $\mathbf{L}(\mathcal{A},\Delta)$ of properties that is a complete lattice, but rarely orthocomplemented in any natural way (Randall and Foulis, 1983). The two are connected by a natural mapping $[\ ] : \Pi \to \mathbf{L}$, given by $p \to [p] = F(J_p)$ where for each $p \in \Pi$, $J_p = \{x \in X \mid p(x) \nleq p'\}$. That is, J_p is the set of outcomes that are consistent with p, and $[p]$ is the largest (i.e., weakest) physical property making p certain to be confirmed if tested.

The mapping $p \to [p]$ is order preserving. For both the classical and quantum-mechanical models

considered above, it is in fact an order-isomorphism. Note that whenever this is the case, Π will inherit from **L** the structure of a complete lattice, which will then automatically be orthomodular by Lemma 4.3. In other words, in such cases we have only *one* logic, which is a complete orthomodular lattice. While it is surely too much to expect that every *conceivable* physical system should enjoy this property -- indeed, we can easily construct toy examples to the contrary -- the condition is at least reasonably transparent in its meaning.

5. Piron's Theorem

Suppose that the logic and property lattices of a model *are* isomorphic, so that the logic of propositions/properties is a complete orthomodular lattice. The question then arises: how close does this bring us to quantum mechanics – that is, to the projection lattice $L(\mathbf{H})$ of a Hilbert space?

The answer is: without additional assumptions, not very. The lattice $L(\mathbf{H})$ has several quite special order-theoretic features. First it is *atomic* -- every element is the join of minimal non-zero elements (i.e., one-dimensional subspaces). Second, it is *irreducible* -- it can not be expressed as a non-trivial direct product of simpler OMLs.[13] Finally, and most significantly, it satisfies the so-called *atomic covering law*: if $p \in L(\mathbf{H})$ is an atom and $p \not\leq q$, then $p \vee q$ *covers* q (no element of $L(\mathbf{H})$ lies strictly between $p \vee q$ and q).

These properties do not quite suffice to capture $L(\mathbf{H})$, but they do get us into the right ballpark. Let **V** be any inner product space over an involutive division ring D. A subspace **M** of **V** is said to be ⊥-*closed* iff $\mathbf{M} = \mathbf{M}^{\perp\perp}$, where $\mathbf{M}^{\perp} = \{v \in \mathbf{V} \mid \forall m \in \mathbf{M}(\langle v,m \rangle = 0)\}$. Ordered by set-inclusion, the collection $L(\mathbf{V})$ of all ⊥-closed subspaces of **V** forms a complete atomic lattice, orthocomplemented by the mapping $\mathbf{M} \rightarrow \mathbf{M}^{\perp}$. A theorem of Amemiya and Araki [1965] shows that a real, complex or quaternionic inner product space **V** with $L(\mathbf{V})$ orthomodular, is necessarily complete. For this reason, an inner product space **V** over an involutive division ring is called a *generalized Hilbert space* if its lattice of closed subspaces $L(\mathbf{V})$ is orthomodular. The following representation theorem is due to C. Piron [1964]:

> **5.1 Theorem**:
> Let L be a complete, atomic, irreducible orthomodular lattice satisfying the atomic covering law. If L contains at least 4 orthogonal atoms, then there exists an involutive division ring D and an inner-product space **V** over D such that L is isomorphic to $L(\mathbf{V})$.

It should be noted that generalized Hilbert spaces have been constructed over fairly exotic division rings.[14] Thus, while it brings us tantalizingly close, Piron's theorem does not quite bring us all the way back to orthodox quantum mechanics.

Conditioning and the Covering Law

Let us call a complete orthomodular lattice satisfying the hypotheses of Piron's theorem a *Piron lattice*. Can we give any general reason for supposing that the logic/property lattice of a physical system (one for

which these are isomorphic) is a Piron lattice? Or, failing this, can we at least ascribe some clear physical content to these assumptions? The atomicity of L follows if we assume that every pure state represents a "physical property". This is a strong assumption, but its content seems clear enough. Irreducibility is usually regarded as a benign assumption, in that a reducible system can be decomposed into its irreducible parts, to each of which Piron's Theorem applies.

The covering law presents a more delicate problem. While it is probably safe to say that no simple and entirely compelling argument has been given for assuming its general validity, Piron [1964, 1976] and others (e.g., Beltrametti and Cassinelli [1981] and Guz [1980]) have derived the covering law from assumptions about the way in which measurement results warrant inference from an initial state to a final state. Here is a brief sketch of how this argument goes. Suppose that there is some reasonable way to define, for an initial state q of the system, represented by an atom of the logic/property lattice L, a final state $\varphi_p(q)$ -- either another atom, or perhaps 0 -- conditional on the proposition p having been confirmed. Various arguments can be adduced suggesting that the only reasonable candidate for such a mapping is the *Sasaki projection* $\varphi_p : L \rightarrow L$, defined by $\varphi_p(q) = (q \vee p') \wedge p$.[15] It can be shown that an atomic OML satisfies the atomic covering law just in case Sasaki projections take atoms again to atoms, or to 0. Another interesting view of the covering law is developed by Cohen and Svetlichny [1987].

6. Classical Representations

The perennial question in the interpretation of quantum mechanics is that of whether or not essentially classical explanations are available, even in principle, for quantum-mechanical phenomena. Quantum logic has played a large role in shaping (and clarifying) this discussion, in particular by allowing us to be quite precise about what we *mean* by a classical explanation.

Classical Embeddings

Suppose we are given a statistical model $(\mathcal{A}, \Delta)$. A very straightforward approach to constructing a "classical interpretation" of $(\mathcal{A}, \Delta)$ would begin by trying to embed $\mathcal{A}$ in a Borel test space $\mathcal{B}$, with the hope of then accounting for the statistical states in Δ as averages over "hidden" classical -- that is, dispersion-free -- states on the latter. Thus, we'd want to find a set S and a mapping $X \rightarrow \mathcal{P}(S)$ assigning to each outcome x of $\mathcal{A}$ a *set* $x^* \subseteq S$ in such a way that, for each test $E \in \mathcal{A}$, $\{x^* \mid x \in E\}$ forms a partition of S. If this can be done, then each outcome x of $\mathcal{A}$ simply records the fact that the system is in one of a certain set of states, namely, x^*. If we let Σ be the Σ-algebra of sets generated by sets of the form $\{x^* \mid x \in X\}$, we find that each probability measure μ on Σ pulls back to a state μ^* on $\mathcal{A}$, namely, $\mu^*(x) = \mu(x^*)$. So long as every state in Δ is of this form, we may claim to have given a completely classical interpretation of the model $(\mathcal{A}, \Delta)$.

The minimal candidate for S is the set of *all* dispersion-free states on $\mathcal{A}$. Setting $x^* = \{s \in S \mid s(x) = 1\}$ gives us a classical interpretation as above, which I'll call the *classical image* of $\mathcal{A}$. Any other classical

interpretation factors through this one. Notice, however, that the mapping $x \to x^*$ is injective only if there are sufficiently many dispersion-free states to separate distinct outcomes of $\mathcal{A}$. If $\mathcal{A}$ has *no* dispersion-free states at all, then its classical image is *empty*. Gleason's theorem tells us that this is the case for quantum-mechanical models. Thus, this particular kind of classical explanation is not available for quantum mechanical models.

It is sometimes overlooked that, even if a test space $\mathcal{A}$ does have a separating set of dispersion-free states, there may exist statistical states on $\mathcal{A}$ that *can not* be realized as mixtures of these. The classical image provides no explanation for such states. For a very simple example of this sort of thing, consider the the test space:

$$\mathcal{A} = \{\{a, x, b\}, \{b, y, c\}, \{c, z, a\}\}$$

and the state $\omega(a) = \omega(b) = \omega(c) = ½$, $\omega(x) = \omega(y) = \omega(z) = 0$. It is a simple exercise to show that ω cannot be expressed as a weighted average of {0,1}-valued states on $\mathcal{A}$. For further examples and discussion of this point, see Wright [1980].]

Contextual Hidden Variables

The upshot of the foregoing discussion is that most test spaces can't be embedded into any classical test space, and that even where such an embedding exists, it typically fails to account for some of the model's states. However, there is one very important class of models for which a satisfactory classical interpretation is *always* possible. Let us call a test space $\mathcal{A}$ *semi-classical* if its tests do not overlap; i.e., if $E \cap F = \emptyset$ for $E, F \in \mathcal{A}$, with $E \neq$ F.

6.1 Lemma:
Let $\mathcal{A}$ be semi-classical. Then $\mathcal{A}$ has a separating set of dispersion-free states, and every extreme state on $\mathcal{A}$ is dispersion-free.

As long as $\mathcal{A}$ is locally countable (i.e., no test E in $\mathcal{A}$ is uncountable), every state can be represented as a convex combination, in a suitable sense, of extreme states [Wilce, 1992]. Thus, every state of a locally countable semi-classical test space has a classical interpretation.

Even though neither Borel test spaces nor quantum test spaces are semi-classical, one might argue that in any real laboratory situation, semi-classicality is the rule. Ordinarily, when one writes down in one's laboratory notebook that one has performed a given test and obtained a given outcome, one always has a record of which test was performed. Indeed, given any test space $\mathcal{A}$, we may always form a semi-classical test space simply by forming the co-product (disjoint union) of the tests in $\mathcal{A}$. More formally:

6.2 Definition:
For each test E in $\mathcal{A}$, let $E^{\sim} = \{ (x,E) \mid x \in E \}$. The *semi-classical cover* of $\mathcal{A}$ is the test space

$$\mathcal{A}^{\sim} = \{E^{\sim} \mid E \in \mathcal{A}\}.$$

We can regard $\mathcal{A}$ as arising from $\mathcal{A}^{\sim}$ by deletion of the record of which test was performed to secure a given outcome. Note that every state on $\mathcal{A}$ defines a state $\tilde{\omega}$ on $\mathcal{A}^{\sim}$ by $\tilde{\omega}(x,E) = \omega(x)$. The mapping $\omega \to \tilde{\omega}$ is plainly injective; thus, we may identify the state-space of $\mathcal{A}$ with a subset of the state-space of $\mathcal{A}^{\sim}$. Notice that there will typically be many states on $\mathcal{A}^{\sim}$ that *do not* descend to states on $\mathcal{A}$. We might wish to think of these as "non-physical", since they do not respect the (presumably, physically motivated) outcome-identifications whereby $\mathcal{A}$ is defined.

Since it is semi-classical, $\mathcal{A}^{\sim}$ admits a classical interpretation, as per Lemma 7.1. Let's examine this. An element of $S(\mathcal{A}^{\sim})$ amounts to a mapping $f : \mathcal{A}^{\sim} \to X$, assigning to each test $E \in \mathcal{A}$, an outcome $f(E) \in E$. This is a (rather brutal) example of what is meant by a *contextual (dispersion-free) hidden variable*. The construction above tells us that such contextual hidden variables will be available for statistical models quite generally. For other results to the same effect, see Kochen and Specker [1967], Gudder [1970], Holevo [1982], and, in a different direction, Pitowsky [1989].[16]

Note that the simple random variables on $\mathcal{A}$ correspond exactly to the simple random variables on $\mathcal{A}^{\sim}$, and that these, in turn, correspond to *some* of the simple random variables (in the usual sense) on the measurable space $S(\mathcal{A}^{\sim})$. Thus, we have the following picture: The model $(\mathcal{A},\Delta)$ can always be obtained from a classical model simply by omitting some random variables, and identifying outcomes that can no longer be distinguished by those that remain.

All of this might suggest that our generalized probability theory presents no significant conceptual departure from classical probability theory. On the other hand, models constructed along the foregoing lines have a distinctly ad hoc character. In particular, the set of "physical" states in one of the classical (or semi-classical) models constructed above is determined not by any independent physical principle, but only by consistency with the original, non-semiclassical model. Another objection is that the contextual hidden variables introduced in this section are badly non-local. It is by now widely recognized that this non-locality is the principal locus of non-classicality in quantum (and more general) probability models. (For more on this, see the entry on the Bell inequalities.)

7. Composite Systems

Some of the most puzzling features of quantum mechanics arise in connection with attempts to describe compound physical systems. It is in this context, for instance, that both the measurement problem and the non-locality results centered on Bell's theorem arise. It is interesting that coupled systems also present a challenge to the quantum-logical programme. I will conclude this article with a description of two results that show that the coupling of quantum-logical models tends to move us further from the realm of Hilbert space quantum mechanics.

The Foulis-Randall Example

A particularly striking result in this connection is the observation of Foulis and Randall [1981] that any reasonable (and reasonably general) tensor product of orthoalgebras will fail to preserve ortho-coherence. Let $\mathcal{A}_5$ denote the test space

$$\{\{a,x,b\}, \{b,y,c\}, \{c,z,d\}, \{d,w,e\}, \{e,v,s\}\}$$

consisting of five three-outcome tests pasted together in a loop. This test space is by no means pathological; it is both ortho-coherent and algebraic. Moreover, it admits a separating set of dispersion-free states and hence, a classical interpretation. Now consider how we might model a compound system consisting of two separated sub-systems each modeled by $\mathcal{A}_5$. We would need to construct a test space $\mathcal{B}$ and a mapping $\otimes : X \times X \to Y = \cup\mathcal{B}$ satisfying, minimally, the following;

a. For all outcomes $x, y, z \in X$, if $x \perp y$, then $x \otimes z \perp y \otimes z$ and $z \otimes x \perp z \otimes y$,
b. For each pair of states $\alpha, \beta \in \Omega(\mathcal{A}_5)$, there exists at least one state ω on $\mathcal{B}$ such that $\omega(x \otimes y) = \alpha(x)\beta(y)$, for all outcomes $x, y \in X$.

Foulis and Randall show that no such embedding exists for which $\mathcal{B}$ is orthocoherent.

Aerts' Theorem

Another result having a somewhat similar force is that of Aerts [1982]. If L_1 and L_2 are two Piron lattices, Aerts constructs in a rather natural way a lattice L representing two *separated* systems, each modeled by one of the given lattices. Here "separated" means that each pure state of the larger system L is entirely determined by the states of the two component systems L_1 and L_2. Aerts then shows that L is again a Piron lattice iff at least one of the two factors L_1 and L_2 is classical. (This result has recently been strengthened by Ischi [2000] in several ways.)

The thrust of these no-go results is that straightforward constructions of plausible models for composite systems destroy regularity conditions (ortho-coherence in the case of the Foulis-Randall result, orthomodularity and the covering law in that of Aerts' result) that have widely been used to underwrite reconstructions of the usual quantum-mechanical formalism. This puts in doubt whether any of these conditions can be regarded as having the universality that the most optimistic version of Mackey's programme asks for. Of course, this does not rule out the possibility that these conditions may yet be motivated in the case of especially *simple* physical systems.

Bibliography

- Aerts, D., *The One and the Many*, Doctoral Dissertation, Free University of Brussels. (1982)
- Amemiya, H., and Aaraki, H., "A Remark on Piron's Paper", *Publ. Res. Inst. Math. Sci., series A*, 2 (1965): 423-427.
- Beltrametti, E., and Cassinelli, G., *The Logic of Quantum Mechanics*, van Nostrand, 1981.
- Birkhoff, G., *Lattice Theory*, American Mathematical Society (Providence): 1967.
- Birkhoff, G., and von Neumann, J., "The Logic of Quantum Mechanics", *Annals of Mathematics* **37** (1936): 823-843.
- Clifton, R., and Kent, A., "Simulating Quantum Mechanics by Non-Contextual Hidden Variables", *Proc. Royal Soc. London* **A 456** (2000): 2101-2114.
- Cohen, D., and Svetlichny, G., Minimal Supports in quantum logics, International Journal of Theoretical Physics **27** (1987): 435-450.
- Cooke, R., and Hilgevoord, J., "A New Approach to Equivalence in Quantum Logic", in E. Beltrammetti and B. van Fraassen (eds.) *Current Issues in Quantum Logic*, New York: Plenum (1981).
- Davey, B., and Priestley, H., *Introduction to Lattices and Order*, Cambridge: Cambridge University Press, 1990
- Fine, A. "Probability and the Interpretation of Quantum Mechanics", *British Journal for the Philosophy of Science* **24** (1973): 1-37.
- Foulis, D. J., Greechie, R., and Ruttimann, G. T., "Filters and Supports in Orthoalgebras", *Int. J. Theor. Phys.* **31** (1992): 789-807.
- Foulis, D. J., and Randall, C.H., "What are quantum logics and what ought they to be?", in E. Beltrametti and B. C. van Fraassen, eds., *Current Issues in Quantum Logic*, New York: Plenum (1980).
- Foulis, D.J., Piron, C., and Randall, C. H., "Realism, Operationalism and Quantum Mechanics", *Foundations of Physics* **13** (1983): 813-841.
- Foulis, D. J., and Randall, C. H., "Empirical Logic and Tensor Products", in H. Neumann (ed), *Interpretations and Foundations of Quantum Mechanics*, Mannheim: Wissenchaftsverlag (1981).
- Foulis, D. J., and Randall, C. H., "Stochastic Entities" in P. Mittelstaedt and E. Stachow (eds.), Recent Developments in Quantum Logic, Mannheim: B. I. Wissenschaft (1984).
- Gleason, A., "Measures on the Closed Subspaces of a Hilbert Space", *Journal of Mathematics and Mechanics* **6** (1957): 885-893.
- Grätzer, G., *General Lattice Theory*, Basel: Birkhäuser Verlag, 1998 (2nd edition).
- Gudder, S., "On Hidden-Variable Theories", *J. Math. Phys.* **11** (1970): 431-436.
- Gudder, S., *Quantum Probability Theory*, San Diego: Academic Press (1989).
- Guz, S., "Filter Theory and Covering Law", *Ann. Inst. H. Poincare* (1980).
- Holevo, A. S., *Probabilistic and Statistical Aspects of Quantum Theory*, Amsterdam: North Holland (1982).
- Holland, S. S. Jr., "Quantum mechanics in Hilbert space: A result of M. P. Soler", *Bulletin of the American Mathematical Society* **32** (1995): 205-232.
- Ischi, B., "Endomorphisms of the Separated Product of Lattices", preprint, University of Geneva, 2001 (to appear in *Int. J. Theor. Phys.)*
- Jauch, J.M., and Piron, C., "On the Structure of Quantal Propositional Systems", *Helvetica Physica Acta* **42** (1969): 842-848.

- Kalmbach, G., *Orthomodular Lattices* , London: Academic Press (1983).
- Klay, M., "Einstein-Podolsky-Rosen Experiments: The Structure of the Sample Space I, II", *Foundations of Physics Letters* **1** (1988): 205-244.
- Kochen, S., and Specker, E. P., "Logical structures arising in quantum theory", *Symposium on the Theory of Models*, Amsterdam: North-Holland (1965).
- Kochen, S. and Specker, E. P., "The Problem of Hidden Variables in Quantum Mechanics" *Journal of Mathematics and Mechanics* **17** (1967): 59-87.
- Ludwig, G., *Mathematical Foundations of Quantum Mechanics I*, New York: Springer-Verlag (1983).
- Mackey, G., Quantum Mechanics and Hilbert Space, *American Math. Monthly* **64** (1957): 45-57.
- Mackey, G. *Foundations of Quantum Mechanics*, Reading: W. A. Benjamin (1963).
- Piron, C., "Axiomatique Quantique", *Helvetica Physica Acta* **37** (1964): 439-468.
- Piron, C., *Foundations of Quantum Physics*, Reading: W. A. Benjamin (1976).
- Pitowsky, I., *Quantum Probability - Quantum Logic*, Lecture Notes in Physics **321**, Springer-Verlag (1989).
- Putnam, H., "Is logic empirical?" in R. Cohen and M. P. Wartofski (eds.), *Boston Studies in the Philosophy of Science* **5** (Dordrecht, Holland: D. Reidel, 1968). Reprinted as "The logic of quantum mechanics" in H. Putnam, *Mathematics, Matter and Method*, Cambridge University Press (1976).
- Randall, C. H., and Foulis, D. J., "Properties and operational propositions in quantum mechanics", *Foundations of Physics* **13** (1983): 843-863.
- van Fraassen, B., *Quantum Mechanics: An Empricist View*, Oxford University Press (1992).
- von Neumann, J., *Mathematische Grundlagen der Quantenmechanik*, Berlin: Springer-Verlag (1932); English translation: *Mathematical Foundations of Quantum Mechanics*, Princeton University Press (1955).
- Wilce, A., "Tensor Products in Generalized Measure Theory", *Int. J. Theor. Phys.* **31** (1992): 1915-1928.
- Wilce, A., "Generalized Sasaki Projections", *Int. J. Theor. Physics* **39** (2000), 969-974.
- Wilce, A., "Test Spaces and Orthoalgebras", in Coecke, B., Moore, D. and Wilce, A., *Current Research in Operational Quantum Logic*, Dordrecht: Kluwer Academic Publishers (2000).
- Wright, R., "The State of the Pentagon", in A. R. Marlowe (Ed.) *Mathematical Foundations of Quantum Physics*, New York: Academic Press (1980).
- Varadarajan, V. S., *The Geometry of Quantum Mechanics*, New York: Springer-Verlag (1985).
- Younce, M., *Random Variables on Non-Boolean Structures*, Doctoral Dissertation, University of Massachusetts (1987).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

quantum mechanics | quantum mechanics: Kochen-Specker theorem | quantum theory: von Neumann vs. Dirac

Alexander Wilce
wilce@juniata.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 4, 2002
Content last modified: February 4, 2002

Stanford Encyclopedia of Philosophy
Supplement to Quantum Logic and Quantum Probability

The Basic Theory of Ordering Relations

What follows is the briefest possible summary of the order-theoretic notions used in the main text. For a good introduction to this material, see Davey & Priestley [1990]. More advanced treatments can be found in Gratzer [1998] and Birkhoff [1967].

---

1. Ordered Sets

A *partial ordering* -- henceforth, just an ordering -- on a set P is a reflexive, anti-symmetric, and transitive binary relation $\trianglelefteq$ on P. Thus, for all p, q, $r \in P$, we have

1. $p \trianglelefteq p$
2. $p \trianglelefteq q$ and $q \trianglelefteq p$ only if $p = q$.
3. if $p \trianglelefteq q$ and $q \trianglelefteq r$ then $p \trianglelefteq r$

If $p \trianglelefteq q$, we speak of p as being *less than*, or *below* q, and of q as being *greater than*, or *above* p, in the ordering.

A *partially ordered set*, or *poset*, is a pair $(P, \trianglelefteq)$ where P is a set and $\trianglelefteq$ is a specified ordering on P. It is usual to let P denote both the set and the structure, leaving $\trianglelefteq$ tacit wherever possible. Any collection of subsets of some fixed set X, ordered by set-inclusion, is a poset; in particular, the full power set $\mathcal{P}(X)$ is a poset under set inclusion.

Let P be a poset. The *meet*, or *greatest lower bound*, of p, $q \in P$, denoted by $p \wedge q$, is the greatest element of P -- if there is one -- lying below both p and q. The *join*, or *least upper bound*, of p and q, denoted by $p \vee q$, is the least element of P -- if there is one -- lying above both p and q. Thus, for any elements p, q, r of P, we have

a. if $r \trianglelefteq p \wedge q$, then $r \trianglelefteq p$ and $r \trianglelefteq q$
b. if $p \vee q \trianglelefteq r$, then $p \trianglelefteq r$ and $q \trianglelefteq r$

Note that $p \wedge p = p \vee p = p$ for all p in P. Note also that $p \trianglelefteq q$ iff $p \wedge q = p$ iff $p \vee q = q$.

Note that if the set $P = \mathcal{P}(X)$, ordered by set-inclusion, then $p \wedge q = p \cap q$ and $p \vee q = p \cup q$. However, if P is an arbitrary collection of subsets of X ordered by inclusion, this need not be true. For instance, consider the collection P of all subsets of $X = \{1,2,...,n\}$ having even cardinality. Then, for instance, $\{1,2\} \vee \{2,3\}$ does not exist in P, since there is no *smallest* set of 4 elements of X containing $\{1,2,3\}$. For a different sort of example, let X be a vector space and let P be the set of *subspaces* of X. For subspaces **M** and **N**, we have

$$\mathbf{M} \wedge \mathbf{N} = \mathbf{M} \cap \mathbf{N}, \text{ but } \mathbf{M} \vee \mathbf{N} = \text{span}(\mathbf{M} \cup \mathbf{N}).$$

The concepts of meet and join extend to infinite subsets of a poset P. Thus, if $A \subseteq P$, the meet of A is the largest element (if any) below A, while the join of A is the least element (if any) above A. We denote the meet of A by $\wedge A$ or by $\wedge_{a \in A}\, a$. Similarly, the join of A is denoted by $\vee A$ or by $\vee_{a \in A}\, a$.

2. Lattices

A *lattice* is a poset $(L, \trianglelefteq)$ in which every pair of elements has both a meet and a join. A *complete lattice* is one in which *every* subset of L has a meet and a join. Note that $\mathcal{P}(X)$ is a complete lattice with respect to set inclusion, as is the set of all subspaces of a vector space. The set of finite subsets of an infinite set X is a lattice, but not a complete lattice. The set of subsets of a finite set having an even number of elements is an example of a poset that is not a lattice.

A lattice $(L, \trianglelefteq)$ is *distributive* iff meets distribute over joins and vice versa:

$$p \wedge (q \vee r) = (p \wedge q) \vee (p \wedge r), \text{ and}$$

$$p \vee (q \wedge r) = (p \vee q) \wedge (p \vee r).$$

The power set lattice $\mathcal{P}(X)$, for instance, is distributive (as is any lattice of sets in which meet and join are given by set-theoretic intersection and union). On the other hand, the lattice of subspaces of a vector space is not distributive, for reasons that will become clear in a moment.

A lattice L is said to be bounded iff it contains a smallest element 0 and a largest element 1. Note that any complete lattice is automatically bounded. For the balance of this appendix, *all lattices are assumed to be bounded*, absent any indication to the contrary.

A *complement* for an element p of a (bounded) lattice L is another element q such that $p \wedge q = 0$ and $p \vee q = 1$.

In the lattice $\mathcal{P}(X)$, every element has exactly one complement, namely, its usual set-theoretic complement. On the other hand, in the lattice of subspaces of a vector space, an element will typically have infinitely many complements. For instance, if L is the lattice of subspaces of 3-dimensional Euclidean space, then a complement for a given plane through the origin is provided by any line through the origin not lying in that plane.

> **Proposition**:
> If L is distributive, an element of L can have at most one complement.
>
> **Proof**:
> Suppose that q and r both serve as complements for p. Then, since L is distributive, we have
>
> $$\begin{aligned} q &= q \wedge 1 \\ &= q \wedge (p \vee r) \\ &= (q \wedge p) \vee (q \wedge r) \\ &= 0 \vee (q \wedge r) \\ &= q \wedge r \end{aligned}$$
>
> Hence, $q \trianglelefteq r$. Symmetrically, we have $r \trianglelefteq q$; thus, $q = r$.

Thus, no lattice in which elements have multiple complements is distributive. In particular, the subspace lattice of a vector space (of dimension greater than 1) is not distributive.

If a lattice *is* distributive, it may be that some of its elements have a complement, while others lack a complement. A distributive lattice in which every element has a complement is called a *Boolean lattice* or a *Boolean algebra*. The basic example, of course, is the power set $\mathcal{P}(X)$ of a set X. More generally, any collection of subsets of X closed under unions, intersections and complements is a Boolean algebra; a theorem of Stone and Birkhoff tells us that, up to isomorphism, every Boolean algebra arises in this way.

3. Ortholattices

In some non-uniquely complemented (hence, non-distributive) lattices, it is possible to pick out, for each element p, a preferred complement p' in such a way that

a. if $p \trianglelefteq q$ then $q' \trianglelefteq p'$
b. $p'' = p$

When these conditions are satisfied, one calls the mapping $p \to p'$ an *orthocomplementation* on L, and the structure $(L, \trianglelefteq, ')$ an *orthocomplemented lattice*, or an *ortholattice* for short.

Note again that if a distributive lattice can be orthocomplemented at all, it is a Boolean algebra, and hence can be orthocomplemented in only one way. In the case of $L(\mathbf{H})$ the orthocomplementation one has in mind is $\mathbf{M} \to \mathbf{M}^{\perp}$ where $\mathbf{M}^{\perp}$ is defined as in Section 1 of the main text. More generally, if $\mathbf{V}$ is any inner product space (complete or not), let $L(\mathbf{V})$ denote the set of subspaces $\mathbf{M}$ of $\mathbf{V}$ such that $\mathbf{M} = \mathbf{M}^{\perp\perp}$ (such a subspace is said to be algebraically closed). This again is a complete lattice, orthocomplemented by the mapping $\mathbf{M} \to \mathbf{M}^{\perp}$.

4. Orthomodularity

There is a striking order-theoretic characterization of the lattice of closed subspaces of a Hilbert space among lattices $L(\mathbf{V})$ of closed subspaces of more general inner product spaces. An ortholattice L is said to be orthomodular iff, for any pair p, q in L with $p \trianglelefteq q$,

(OMI) $\quad (q \wedge p') \vee p = q.$

Note that this is a weakening of the distributive law. Hence, a Boolean lattice is orthomodular. It is not difficult to show that if $\mathbf{H}$ is a Hilbert space, then $L(\mathbf{H})$ is orthomodular. The striking converse of this fact is due to Amemiya and Araki [1965]:

> **Theorem**:
> Let $\mathbf{V}$ be an inner product space (over $\mathbf{R}$, $\mathbf{C}$ or the quaternions) such that $L(\mathbf{V})$ is orthomodular. Then $\mathbf{V}$ is complete, i.e., a Hilbert space.

5. Closure Operators, Interior Operators and Adjunctions

Let P and Q be posets. A mapping $f : P \to Q$ is *order preserving* iff for all $p,q \in P$, if $p \trianglelefteq q$ then $f(p) \trianglelefteq f(q)$.

A *closure operator* on a poset P is an order-preserving map $\mathbf{cl} : P \to P$ such that for all $p \in P$,

- $\mathbf{cl}(\mathbf{cl}(p)) = p$
- $p \trianglelefteq \mathbf{cl}(p)$.

Dually, an *interior operator* on P is an order-preserving mapping $\mathbf{int} : P \to P$ on P such that for all $p \in P$,

- $\mathbf{int}(\mathbf{int}(p)) = \mathbf{int}(p)$

- **int**(p) $\trianglelefteq$ p

Elements in the range of **cl** are said to be *closed*; those in the range of **int** are said to be *open*. If P is a (complete) lattice, then the set of closed, respectively open, subsets of P under a closure or interior mapping is again a (complete) lattice.

By way of illustration, suppose that $\mathcal{O}$ and $\mathcal{C}$ are collections of subsets of a set X with $\mathcal{O}$ closed under arbitrary unions and $\mathcal{C}$ under arbitrary intersections. For any set $A \subseteq X$, let

> **cl**(A) = $\bigcap\{C \in \mathcal{C} \mid A \subseteq C\}$, and
>
> **int**(A) = $\bigcup\{O \in \mathcal{O} \mid O \subseteq A\}$

Then **cl** and **int** are interior operators on $\mathcal{P}(X)$, for which the closed and open sets are precisely $\mathcal{C}$ and $\mathcal{O}$, respectively. The most familiar example, of course, is that in which $\mathcal{O}$, $\mathcal{C}$ are the open and closed subsets, respectively, of a topological space. Another important special case is that in which $\mathcal{C}$ is the set of linear subspaces of a vector space **V**; in this case, the mapping *span* : $\mathcal{P}$(**V**) $\rightarrow$ $\mathcal{P}$(**V**) sending each subset of **V** to its span is a corresponding closure.

An *adjunction* between two posets P and Q is an ordered pair (f, g) of mappings $f : P \rightarrow Q$ and $g : Q \rightarrow P$ connected by the condition that, for all $p \in P$, $q \in Q$

> $f(p) \trianglelefteq q$ if and only if $p \trianglelefteq g(q)$.

In this case, we call f a *left adjoint* for g, and call g a *right adjoint* for f. Two basic facts about adjunctions, both easily proved, are the following:

> **Proposition**:
> Let $f : L \rightarrow M$ be an order-preserving map between complete lattices L and M. Then
>
> a. f preserves arbitrary joins if and only if it has a right adjoint.
> b. f preserves arbitrary meets if and only if it has a left adjoint.

> **Proposition**:
> Let (f, g) be an adjunction between complete lattices L and M. Then
>
> a. $g \circ f : L \rightarrow L$ is a closure operator.
> b. $f \circ g : M \rightarrow M$ is an interior operator.

Alexander Wilce
wilce@juniata.edu

Return to Quantum Logic and Quantum Probability

First published: February 4, 2002
Content last modified: February 4, 2002

Stanford Encyclopedia of Philosophy
Notes to Quantum Logic and Quantum Probability

Notes

1. A few qualifications are in order already: In a more general formulation, one considers the lattice of projections of a von Neumann algebra. Only in the context of non-relativistic quantum mechanics, and then only absent superselection rules, is this algebra a type I factor. For the expository purposes of this paper, we restrict our discussion to this context.

2. Throughout this paper, I use the term "logic" rather narrowly to refer to the algebraic and order-theoretic aspect of propositional logic. There exists a substantial technical literature devoted to non-classical formal deductive systems that are intended to stand to quantum propositional logics rather as classical deductive systems stand to Boolean algebras. A good reference for this material is Kalmbach [1983].

3. It is important to note here that even in classical mechanics, only subsets of the state-space that are measurable (in the sense of measure theory) are regarded as representing observable properties of the system, and only these are assigned probabilities. The difference is that in the classical case, the observable properties form a sub-Boolean algebra of the power set of *S*, while in the quantum case, they do not.

4. The first explicit formulation of this interpretation seems to have been given by Jauch and Piron [1969].

5. So-called *modal* interpretations of quantum mechanics do not attempt to assign actual values (or ranges of values) to *all* observables. Rather, they identify (in various different ways) a privileged observable or class of observables as having "definite values", and thereby avoid the various no-go theorems for hidden variables. For references and further details, see the entry on modal interpretations of Quantum Mechanics.

6. Of course, another possible response to such a question is to dismiss it, perhaps with the observation that mathematical models of natural phenomena evolve, more or less organically, to fit the facts, and require no a priori justification. Or, to put it differently, we can insist that quantum logic has the structure that it does just because the *world* has that structure.

7. These include, besides Mackey's original formalism, that of Piron [1976], the approach based on partial Boolean algebras of Kochen and Specker [1965, 1967], and various approaches emphasizing the

convex structure of the set of (statistical) states of a system, e.g., [Holevo 1982, Ludwig 1983, Pitowsky 1989].

8. Even at this point, a couple of remarks are in order. First, notice that *E* need not be the set of outcomes of any humanly executable, let alone repeatable, measurement or experiment. Any exhaustive set of mutually exclusive alternatives will serve, at least as long as some sense can be attached to the terms “occurrence” and "realization" (even as terms of art). Secondly, notice that every standard interpretation of probability theory, whether relative-frequentist, propensity, subjective or what-have-you, represents probability weights mathematically in the same way. Thus, the framework just sketched is agnostic among these interpretations.

9. It is worth remarking that all approaches to a generalized probability theory contain some mechanism for identifying outcomes of distinct measurements or values of distinct observables, though this mechanism varies from author to author . The most common approaches are to identify outcomes that are equi-probable in every state (as in the work of Mackey), or to identify outcomes that are certain (i.e., have probability 1) in exactly the same states (as in the work of Piron). Both of these prescriptions become problematic in the context of sequential measurements --- see, e.g., [Cooke and Hilgevoord, 1979]. The Foulis-Randall theory has the advantage of remaining neutral as to the precise mechanism whereby outcomes are identified.

10. The approach taken here is modeled on, but somewhat less general than, that of Foulis, Piron and Randall [1983]. See also [Foulis and Randall, 1987]

11. Indeed, the pair (S, F) is an *adjunction* between $\mathcal{P}(X)$ and $\mathcal{P}(\Delta)$ -- that is, for any $\Gamma \in \mathcal{P}(\Delta)$ and any $J \in \mathcal{P}(X)$, we have $S(\Gamma) \subseteq J$ if and only if $\Gamma \subseteq F(J)$. It follows that the mappings

$$\Gamma \rightarrow F(S(\Gamma)) \text{ and } J \rightarrow S(F(J))$$

are respectively a closure operator on $\mathcal{P}(\Delta)$ and an interior operator on $\mathcal{P}(X)$. The collection of closed subsets of Δ and the collection of open subsets of X (which are exactly the ranges of the mappings F and S) are complete lattices, closed under intersection and union respectively, and are mapped isomorphically onto one another by S and F. (See Supplement 2 for further details.)

12. In the case where Δ is a *convex* set of states the closed subsets of Δ are *faces* of Δ; that is, $F(J)$ is itself convex, and if a convex combination of states lies in $F(J)$, then those states themselves lie in $F(J)$. The faces of any convex set Δ, ordered by inclusion, form a complete lattice, closed under intersection. Thus, in this context, the property lattice is a complete sublattice of the face lattice of the state-space. This is the starting point for a number of approaches to generalized probability theory. Note that if Δ is a simplex, as it is for classical models, then the face lattice of Δ is a Boolean algebra.

13. This condition fails for the projection lattice modeling a quantum-mechanical system with

superselection rules, but continues to hold for the lattices associated with each superselection sector.

14. The first such examples were obtained by H. Keller. Recently, M. P. Soler has shown that if a generalized Hilbert space contains any infinite orthonormal set, it must in fact be a classical Hilbert space --- i.e., the division ring D must be the field of real numbers, the field of complex numbers, or the ring of quaternions. See Holland [1995] for references and further discussion.

15. For a development of this idea in a more general context, see Wilce [2000].

16. Interesting recent work of Mayers [1999] and Clifton and Kent [2000] shows that the quantum test space *F*(**H**) contains a dense semi-classical sub-test space. A detailed hidden-variables model along these lines has not been seriously entertained (and would presumably be otherwise problematic), but these results do suggest that experiments having finite precision can not rule out non-contextual hidden variables.

Alexander Wilce
wilce@juniata.edu

First published: February 4, 2002
Content last modified: February 4, 2002

Relational Quantum Mechanics

Relational quantum mechanics is an interpretation of quantum theory which discards the notions of absolute state of a system, absolute value of its physical quantities, or absolute event. The theory describes only the way systems affect each other in the course of physical interactions. State and physical quantities refer always to the interaction, or the relation, between *two* systems. Nevertheless, the theory is assumed to be complete. The physical content of quantum theory is understood as expressing the net of relations connecting all different physical systems.

1. Introduction

Quantum theory is our current general theory of physical motion. The theory is the core component of the momentous change that our understanding of the physical world has undergone during the first

decades of the 20th century. It is one of the most successful scientific theories ever: it is supported by vast and unquestionable empirical and technological effectiveness and is today virtually unchallenged. But the interpretation of what the theory actually tells us about the physical world raises a lively debate, which has continued with alternating fortunes, from the early days of the theory in the late twenties, to nowadays. The *relational interpretations* are a number of reflections by different authors, which were independently developed, but converge in indicating an interpretation of the physical content of the theory. The core idea is to read the theory as a theoretical account of the way distinct physical systems *affect each other* when they interact (and not of the way physical systems "are"), and the idea that this account exhausts all that can be said about the physical world. The physical world is thus seen as a net of interacting components, where there is no meaning to the state of an isolated system. A physical system (or, more precisely, its contingent state) is reduced to the net of relations it entertains with the surrounding systems, and the physical structure of the world is identified as this net of relationships.

The possibility that the physical content of an empirically successful physical theory could be debated should not surprise: examples abound in the history of science. For instance, the great scientific revolution was fueled by the grand debate on whether the effectiveness of the Copernican system could be taken as an indication that the Earth was not *in fact* at the center of the universe. In more recent times, Einstein's celebrated first major theoretical success, special relativity, consisted to a large extent just in understanding the physical meaning (simultaneity is relative) of an already existing effective mathematical formalism (the Lorentz transformations). In these cases, as in the case of quantum mechanics, a very strictly empiricist position could have circumvented the problem altogether, by reducing the content of the theory to a list of predicted numbers. But perhaps science can offer us more than such a list; and certainly science needs more than such a list to find its ways.

The difficulty in the interpretation of quantum mechanics derives from the fact that the theory was first constructed for describing microscopic systems (atoms, electrons, photons) and the way these interact with macroscopic apparatuses built to measure their properties. Such interactions are denoted as "measurements". The theory consists in a mathematical formalism, which allows probabilities of alternative outcomes of such measurements to be calculated. If used just for this purpose, the theory raises no difficulty. But we expect the macroscopic apparatuses themselves -- in fact, any physical system in the world -- to obey quantum theory, and this seems to raise contradictions in the theory.

1.1 The Problem

In classical mechanics, a system S is described by a certain number of physical variables. For instance, an electron is described by its position and its spin (intrinsic angular momentum). These variables change with time and represent the contingent properties of the system. We say that their values determine, at every moment, the "state" of the system. A measurement of a system's variable is an interaction between the system S and an external system O, whose effect on O, depends on the actual value q of the variable (of S) which is measured. The characteristic feature of quantum mechanics is that it does not allow us to assume that all variables of the system have determined values at every moment (this irrespectively of whether or not we know such values). It was Werner Heisenberg who first realized the need to free

ourselves from the belief that, say, an electron has a well determined position at every time. When it is not interacting with an external system that can detect its position, the electron can be "spread out" over different positions. In the jargon of the theory, one says that the electron is in a "quantum superposition" of two (or many) different positions. It follows that the state of the system cannot be captured by giving the value of its variables. Instead, quantum theory introduces a new notion of "state" of a system, which is different from a list of values of its variables. Such a new notion of state was developed in the work of Erwin Schrödinger in the form of the "wave function" of the system, usually denoted by Ψ. Paul Adrien Maurice Dirac gave a general abstract formulation of the notion of quantum state, in terms of a vector Ψ moving in an abstract vector space. The time evolution of the state Ψ is deterministic and is governed by the Schrödinger equation. From the knowledge of the state Ψ, one can compute the probability of the different measurement outcomes q. That is, the probability of the different ways in which the system S can affect a system O in an interaction with it. The theory then prescribes that at every such ‘measurement’, one must update the value of Ψ, to take into account which of the different outcomes has happened. This sudden change of the state Ψ depends on the specific outcome of the measurement and is therefore probabilistic. It is called the "collapse of the wave function".

The problem of the interpretation of quantum mechanics takes then different forms, depending on the relative ontological weight we choose to assign to the wave function Ψ or, respectively, to the sequence of the measurement outcomes q, q', q'', …. If we take Ψ as the "real" entity which fully represents the actual state of affairs of the world, we encounter a number of difficulties. First, we have to understand how Ψ can change suddenly in the course of a measurement: if we describe the evolution of two interacting quantum systems in terms of the Schrödinger equation, no collapse happens. Furthermore, the collapse, seen as a physical process, seems to depend on arbitrary choices in our description and shows a disturbing amount of nonlocality. But even if we can circumvent the collapse problem, the more serious difficulty of this point of view is that it appears to be impossible to understand how specific observed values q, q', q'', … can emerge from the same Ψ. A better alternative is to take the observed values q, q', q'', … as the actual elements of reality, and view Ψ just as a bookkeeping device, determined by the actual values q, q', q'', … that happened in past. From this perspective, the real events of the world are the "realization" (the "coming to reality", the "actualization") of the values q, q', q'', … in the course of the interaction between physical systems. This actualization of a variable q in the course of an interaction can be denoted as the *quantum event* q. An exemple of a quantum event is the detection of an electron in a certain position. The position variable of the electron assumes a determined value in the course of the interaction between the electron and an external system and the quantum event is the "manifestation" of the electron in a certain position. Quantum events have an intrinsically discrete ("quantized") granular structure.

The difficulty of this second option is that if we take the quantum nature of all physical systems into account, the statement that a certain specific event q "has happened" (or, equivalently that a certain variable has or has not taken the value q) can be true and not-true at the same time. To clarify this key point, consider the case in which a system S interacts with another system (an apparatus) O, and exhibits a value q of one of its variables. Assume that the system O obeys the laws of quantum theory as well, and use the quantum theory of the combined system formed by O and S in order to predict the way this combined system can later interact with a *third* system O'. Then quantum mechanics forbids us to

assume that q has happened. Indeed, as far as its later behavior is concerned, the combined system $S+O$ may very well be in a quantum superposition of alternative possible values q, q', q'', …. This "second observer" situation captures the core conceptual difficulty of the interpretation of quantum mechanics: reconciling the possibility of quantum superposition with the fact that the observed world is characterized by uniquely determined events q, q', q'', …. More precisely, it shows that we cannot disentangle the two: according to the theory an observed quantity (q) can be at the same time determined and not determined. An event may have happened and at the same time may not have happened.

2. Relational view of quantum states

The way out from this dilemma suggested by the relational interpretation is that the quantum events, and thus the values of the variables of a physical system S, namely the q's, are relational. That is, they do not express properties of the system S alone, but rather refer to the relation between two systems. In particular, the central tenet of *relational quantum mechanics* (Rovelli 1996, 1997) is that there is no meaning in saying that a certain quantum event has happened or that a variable of the system S has taken the value q: rather, there is meaning in saying that the event q has happened or the variable has taken the value *q for O*, or *with respect to O*. The apparent contradiction between the two statements that a variable has or hasn't a value is resolved by indexing the statements with the different systems with which the system in question interacts. If I observe an electron at a certain position, I cannot conclude that the electron *is* there: I can only conclude that the electron *as seen by me* is there. Quantum events only happen in interactions between systems, and the fact that a quantum event has happened is only true with respect to the systems involved in the interaction. The unique account of the state of the world of the classical theory is thus fractured into a multiplicity of accounts, one for each possible "observing" physical system. In the words of (Rovelli 1996): "Quantum mechanics is a theory about the physical description of physical systems relative to other systems, and this is a complete description of the world".

This relativisation of actuality is viable thanks to a remarkable property of the formalism of quantum mechanics. John von Neumann was the first to notice that the formalism of the theory treats the measured system (S) and the measuring system (O) differently, but the theory is surprisingly flexible on the choice of where to put the boundary between the two. Different choices give different accounts of the state of the world (for instance, the collapse of the wave function happens at different times); but this does not affect the predictions on the final observations. Von Neumann only described a rather special situation, but this flexibility reflects a general structural property of quantum theory, which guarantees the consistency among all the distinct "accounts of the world" of the different observing systems. The manner in which this consistency is realized, however, is subtle.

What appears with respect to O as a measurement of the variable q (with a specific outcome), appears with respect to O' simply as the establishing of a *correlation* between S and O (without any specific outcome). As far as the observer O is concerned, a quantum event has happened and a property q of a system S has taken a certain value. As far as the second observer O' is concerned, the only relevant element of reality is that a correlation is established between S and O. This correlation will manifest itself only in any further observation that O' would perform on the $S+O$ system. Up to the time in which it

physically interacts with $S+O$, the system O' has no access to the actual outcomes of the measurements performed by O on S . This actual outcome is real only with respect to O (Rovelli 1996, pp. 1650-52). Consider for instance a two-state system O (say, a light-emitting diode, or l.e.d., which can be *on* or *off*) interacting with a two-state system S (say, the spin of an electron, which can be *up* or *down*). Assume the interaction is such that if the spin is *up* (*down*) the l.e.d. goes *on* (*off*). To start with, the electron can be in a superposition of its two states. In the account of the state of the electron that we can associate with the l.e.d., a quantum event happens in the interaction, the wave function of the electron collapses to one of two states, and the l.e.d. is then either *on* or *off*. But we can also consider the l.e.d./electron composite system as a quantum system and study the interactions of this composite system with another system O'. In the account associated to O', there is no event and no collapse at the time of the interaction, and the composite system is still in the superposition of the two states [spin *up*/l.e.d. *on*] and [spin *down*/l.e.d. *off*] after the interaction. It is necessary to assume this superposition because it accounts for measurable interference effects between the two states: if quantum mechanics is correct, these interference effects are truly observable by O'. So, we have two discordant accounts of the same events. Can the two discord accounts be compared and does the comparison lead to contradiction? They can be compared, because the information on the first account is stored in the state of the l.e.d. and O' has access to this information. Therefore O and O' can compare their accounts of the state of the world.

However, the comparison does not lead to contradiction *because the comparison is itself a physical process that must be understood in the context of quantum mechanics*. Indeed, O' can physically interact with the electron and then with the l.e.d. (or, equivalently, the other way around). If, for instance, he finds the spin of the electron *up*, quantum mechanics predicts that he will then consistently find the l.e.d. on (because in the first measurement the state of the composite system collapses on its [spin *up*/l.e.d. *on*] component). That is, the multiplicity of accounts leads to no contradiction precisely because the comparison between different accounts can only be a physical quantum interaction. This internal self-consistency of the quantum formalism is general, and it is perhaps its most remarkable aspect. This self consistency is taken in relational quantum mechanics as a strong indication of the relational nature of the world.

In fact, one may conjecture that this peculiar consistency between the observations of different observers is the missing ingredient for a reconstruction theorem of the Hilbert space formalism of quantum theory. Such a reconstruction theorem is still unavailable: On the basis of reasonable physical assumptions, one is able to derive the structure of an orthomodular lattice containing subsets that form Boolean algebras, which "almost", but not quite, implies the existence of a Hilbert space and its projectors' algebra (see the entry Quantum Logic and Quantum Probability.) Perhaps an appropriate algebraic formulation of the condition of consistency between subsystems could provide the missing hypothesis to complete the reconstruction theorem.

3. Correlations

The conceptual relevance of *correlations* in quantum mechanics, - a central aspect of relational quantum mechanics -- is emphasized by David Mermin, who analyses the statistical features of correlation

(Mermin 1998), and arrives at views close to the relational ones. Mermin points out that a theorem on correlations in Hilbert space quantum mechanics is relevant to the problem of what exactly quantum theory tells us about the physical world. Consider a quantum system *S* with internal parts *s*, s',…, that may be considered as subsystems of *S* , and define the correlations among subsystems as the expectation values of products of subsystems' observables. It can be proved that, for any resolution of *S* into subsystems, the subsystems' correlations determine *uniquely* the state of *S*. According to Mermin, this theorem highlights two major lessons that quantum mechanics teaches us: first, the relevant physics of *S* is entirely contained in the correlations both among the *s*, s',…, themselves (internal correlations) and among the s',…, and other systems (external correlations); second, correlations may be ascribed physical reality whereas, according to well-known 'no-go' theorems, the quantities that are the terms of the correlations cannot (Mermin 1998).

4. Self-reference and self-measurement

From a relational point of view, the properties of a system exists only in reference to another system. What about the properties of a system with respect to itself? Can a system measure itself? Is there any meaning of the correlations of a system with itself? Implicit in the relational point of view is the intuition that a complete self-measurement is impossible. It is this impossibility that forces all properties to be referred to another system. The issue of the self-measurement has been analyzed in details in two remarkable works, from very different perspectives, but with similar conclusions, by Marisa Dalla Chiara and by Thomas Breuer.

4.1 Logical aspect of the measurement problem

Marisa Dalla Chiara (1977) has addressed the *logical* aspect of the measurement problem. She observes that the problem of self-measurement in quantum mechanics is strictly related to the *self-reference* problem, which has an old tradition in logic. From a logical point of view the measurement problem of quantum mechanics can be described as a characteristic question of "semantical closure" of a theory. To what extent can quantum mechanics apply consistently to the objects and the concepts in terms of which its metatheory is expressed? Dalla Chiara shows that the duality in the description of state evolution, encoded in the ordinary (i.e. von Neumann's) approach to the measurement problem, can be given a purely logical interpretation: "If the apparatus observer *O* is an object of the theory, then *O* cannot realize the reduction of the wave function. This is possible only to another O', which is 'external' with respect to the universe of the theory. In other words, any apparatus, as a particular physical system, can be an object of the theory. Nevertheless, *any apparatus which realizes the reduction of the wave function is necessarily only a metatheoretical object* " (Dalla Chiara 1977, p. 340). This observation is remarkably consistent with the way in which the state vector reduction is justified within the relational interpretation of quantum mechanics. When the system *S*+*O* is considered from the point of view of O', the measurement can be seen as an interaction whose dynamics is fully unitary, whereas by the point of view of *O* the measurement breaks the unitarity of the evolution of *S*. The unitary evolution does not break down through mysterious physical jumps, due to unknown effects, but simply because *O* is not giving a full dynamical description of the interaction. *O* cannot have a full description of the interaction of *S* with

himself (*O*), because his information is correlation information and there is no meaning in being correlated with oneself. If we include the observer into the system, then the evolution is still unitary, but we are now dealing with the description of a different observer.

4.2 Impossibility of complete self-measurement

As is well known, from a purely logical point of view self-reference properties in formal systems impose limitations on the descriptive power of the systems themselves. Thomas Breuer has shown that, from a physical point of view, this feature is expressed by the existence of limitations in the universal validity of physical theories, *no matter whether classical or quantum* (Breuer 1995). Breuer studies the possibility for an apparatus *O* to measure its own state. More precisely, of measuring the state of a system *containing* an apparatus *O*. He defines a map from the space of all sets of states of the apparatus to the space of all sets of states of the system. Such a map assigns to every set of apparatus states the set of system states that is compatible with the information that -- after the measurement interaction -- the apparatus is in one of these states. Under reasonable assumptions on this map, Breuer is able to prove a theorem stating that no such map can exist that can distinguish all the states of the system. An apparatus *O* cannot distinguish all the states of a system *S* containing *O*. This conclusion holds irrespective of the classical or quantum nature of the systems involved, but in the quantum context it implies that no quantum mechanical apparatus can measure all the quantum correlations between *itself* and an external system. These correlations are only measurable by a second external apparatus, observing both the system and the first apparatus.

5. Other relational views

5.1 Quantum reference systems

A relational view of quantum mechanics has been proposed also by Gyula Bene (1997). Bene argues that quantum states are relative in the sense that they express a relation between a system to be described and a different system, containing the former as a subsystem and acting for it as a *quantum reference system* (here the system is contained in the reference system, while in Breuer's work the system contains the apparatus). Consider again a measuring system (*O*) that has become entangled with a measured system (*S*) during a measurement. Once again, the difficulty of quantum theory is that there is an apparent contradiction between the fact that the quantity *q* of the system assumes an observed value in the measurement, while the composite *S*+*O* system still has to be considered in a superposition state, if we want to properly predict the outcome of measurements on the *S*+*O* system. This apparent contradiction is resolved by Bene by relativizing the state not to an observer, as in the relational quantum mechanics sketched in Section 2, but rather to a relevant composite system. That is: there is a state of the system *S* relative to *S* alone, and a state of the system *S* relative to the *S*+*O* composite system. (Similarly, there is a state of the system *O* relative to itself alone, and a state of the system *O* relative to the *S*+*O* ensemble.) The ensemble with respect to which the state is defined is called by Bene the *quantum reference system* . The state of a system with respect to a given quantum reference system correctly predicts the probability

distributions of any measurement on the entire reference system. This dependence of the states of quantum systems from different quantum systems that act as reference systems is viewed as a fundamental property that holds no matter whether a system is observed or not.

5.2 Sigma algebra of interactive properties

Similar views have been expressed by Simon Kochen in unpublished but rather well-known notes (Kochen, 1979). In Kochen's words: "The basic change in the classical framework which we advocate lies in dropping the assumption of the absoluteness of physical properties of interacting systems... Thus quantum mechanical properties acquire an interactive or relational character." Kochen uses a σ-algebra formalism. Each quantum system has an associated Hilbert space. The properties of the system are established by its interaction with other quantum systems, and these properties are represented by the corresponding projection operators on the Hilbert space. These projectors are elements of a Boolean σ-algebra, determined by the physics of the interaction between the two systems. Suppose a quantum system S can interact with quantum systems Q, Q',…. In each case, S will acquire an interaction σ-algebra of properties $\sigma(Q)$, $\sigma(Q')$ since the interaction between S and Q may be finer grained than the interaction between S and Q'. Thus, interaction σ-algebras may have non-trivial intersections. The *family* of all Boolean σ-algebras forms a category, with the sets of the projectors of each σ-algebra as objects. In Kochen's words: "Just as the state of a composite system does not determine states of its components, conversely, the states of the... correlated systems do not determine the state of the composite system [...] We thus resolve the measurement problem by cutting the Gordian knot tying the states of component systems uniquely to the state of the combined system." This is very similar in spirit to the Bene approach. The precise relation between Kochen's approach and Rovelli's relational quantum mechanics has been analysed by Bill Curry (1999).

Further approaches at least formally related to Kochen's have been proposed by Healey (1989), who also emphasises an interactive aspect of his approach, and by Dieks (1989). See also the entry on 'Modal Interpretations of Quantum Mechanics'.

5.3 Quantum theory of the universe

Relational views on quantum theory have been defended also by Lee Smolin (1995) and by Louis Crane (1995) in a cosmological context. If one is interested in the quantum theory of the entire universe, then, by definition, an external observer is not available. Breuer's theorem shows then that a quantum state of the universe, containing all correlations between all subsystems, expresses information that is not available, not even in principle, to any observer. In order to write a meaningful quantum state, argue Crane and Smolin, we have to divide the universe in two components and consider the relative quantum state predicting the outcomes of the observations that one component can make on the other.

5.4 Relation with Everett's relative-state interpretation

Relational ideas underlie also the interpretations of quantum theory inspired by the work of Everett. Everett' original work (Everett 1975) relies on the notion of "relative state" and has a marked relational tone (see quantum mechanics: Everett's relative-state formulation of). In the context of Everettian accounts, a state may be taken as relative either (more commonly) to a "world", or "branch", or (sometimes) to the state of another system (see for instance Saunders 1996, 1998). While the first variant (relationalism with respect to branches) is far from the relational views described here, the second variant (relationalism with respect to the state of a system) is closer.

However, it is different to say that something is relative to a system or that something is relative to a state of a system. Consider for instance the situation described in the example of Section 5: According to the relational interpretation, after the first measurement the quantity q has a given value and only one for O, while in Everettian terms the quantity q has a value for one state of O and a different value for another state of O, and the two are equally real. In Everett, there is an ontological multiplicity of realities, which is absent in the relational point of view, where physisical quantities are uniquely determined, once two systems are given.

The difference derives from a very general interpretational difference between Everettian accounts and the relational point of view. Everett (at least in its widespread version) takes the state Ψ as the basis of the ontology of quantum theory. The overall state Ψ includes different possible branches and different possible outcomes. On the other hand, the relational interpretation takes the quantum events q, that is, the actualizations of values of physical quantities, as the basic elements of reality (see Section 1.1 above) and such q's are assumed to be univocal. The relational view avoids the traditional difficulties in taking the q's as univocal simply by noticing that a q does not refer to a system, but rather to a pair of systems.

For a comparison between the relational interpretation and other current interpretations of quantum mechanics, see Rovelli 1996.

6. Some consequences of the relational point of view

A number of open conceptual issues in quantum mechanics appear in a different light when seen in the context of a relational interpretation of the theory. For instance, the conventional conclusions of the Einstein-Podolsky-Rosen argument turn out to be frame-dependent, and this result supports the "peaceful coexistence" of quantum mechanics and special relativity (Laudisa 2001). In certain istances, the descriptions given in different Lorentz frames can be identified with descriptions relative to different observer systems, which are consistent in the sense of Section 2.

Also, the relational interpretation allows one to give a precise definition of the time (or, better, the probability distribution of the time) at which a measurement happens, in terms of the probability distribution of the correlation between system and apparatus, as measurable by a third observer (Rovelli 1998).

Finally, it has been suggested in (Rovelli 1997) that the relationalism at the core of quantum theory pointed out by the relational interpretations might be connected with the spatiotemporal relationalism that characterizes general relativity. Quantum mechanical relationalism is the observation that there are no absolute properties: properties of a system *S* are relative to another system *O* with which *S* is interacting. General relativistic relationalism is the well known observation that there is no absolute localization in spacetime: localization of an object *S* in spacetime is only relative to the gravitational field, or to any other object *O*, to which *S* is contiguous. There is a connection between the two, since interaction between *S* and *O* implies contiguity and contiguity between *S* and *O* can only be checked via some *quantum* interaction. However, because of the difficulty of developing a consistent and conceptually transparent theory of quantum gravity, so far this suggestion has not been developed beyond the stage of a simple intuition.

7. Conclusion

Relational interpretations of quantum mechanics propose a solution to the interpretational difficulties of quantum theory based on the idea of weakening the notions of the state of a system, event, and the idea that a system, at a certain time, may just have a certain property. The world is described as an ensemble of events ("the electron is the point *x*") which happen only *relatively to* a given observer. Accordingly, the state and the properties of a system are relative to another system only. There is a wide diversity in style, emphasis, and language in the authors that we have mentioned. Indeed, most of the works mentioned have developed independently from each other. But it is rather clear that there is a common idea underlying all these approaches, and the convergence is remarkable.

Werner Heisenberg first recognized that the electron does not have a well defined position when it is not interacting. Relational interpretations push this intuition further, by stating that, even when interacting, the position of the electron is only determined in relation to a certain observer, or to a certain quantum reference system, or similar.

In physics, the move of deepening our insight into the physical world by relativizing notions previously used as absolute has been applied repeatedly and very successfully. Here are a few examples. The notion of the velocity of an object has been recognized as meaningless, unless it is indexed with a reference body with respect to which the object is moving. With special relativity, simultaneity of two distant events has been recognized as meaningless, unless referred to a specific state of motion of something. (This something is usually denoted as "the observer" without, of course, any implication that the observer is human or has any other peculiar property besides having a state of motion. Similarly, the "observer system" *O* in quantum mechanics need not to be human or have any other property beside the possibility of interacting with the "observed" system *S*.) With general relativity, the position in space and time of an object has been recognized as meaningless, unless it is referred to the gravitational field, or to some other dynamical physical entity. The move proposed by the relational interpretations of quantum mechanics has strong analogies with these, but is, in a sense, a longer jump, since all physical events and the entirety of the contingent properties of any physical system are taken to be meaningful only as relative to a

second physical system. The claim of the relational interpretations is that this is not an arbitrary move. Rather, it is a conclusion which is difficult to escape, following from the observation -- explained above in the example of the "second observer" -- that a variable (of a system *S*) can have a well determined value *q* for one observer (*O*) and at the same time fail to have a determined value for another observer (*O*′).

This way of thinking the world has certainly heavy philosophical implications. The claim of the relational interpretations is that it is nature itself that is forcing us to this way of thinking. If we want to understand nature, our task is not to frame nature into our philosophical prejudices, but rather to learn how to adjust our philosophical prejudices to what we learn from nature.

Bibliography

- Bene, G., "Quantum reference systems: a new framework for quantum mechanics", *Physica* **A** 242 (1992): 529-560.
- Breuer, T., "The impossibility of accurate state self-measurements", *Philosophy of Science* **62** (1993): 197-214.
- Crane, L., "Clock and Category: Is Quantum Gravity Algebraic?", *Journal of Mathematical Physics* **36** (1993): 6180-6193.
- Dalla Chiara M.L., "Logical self-reference, set theoretical paradoxes and the measurement problem in quantum mechanics", *Journal of Philosophical Logic* **6** (1977): 331-347.
- Everett H., "‘Relative State’ Formulation of Quantum Mechanics," Reviews of Modern Physics, 29 (1957) 454-462.
- Curry B.,"Interactive Property versus Relational Interpretations of Quantum Mechanics", unpublished notes (1999).
- Dieks, D., "Resolution of the Measurement Problem through Decoherence of the Quantum State ", Physics letters A 142, 439-446 (1989).
- Healey, R., The Philosophy of Quantum Mechanics: An Interactive Interpretation, Cambridge University Press (1989).
- Kochen S., "The interpretation of quantum mechanics", unpublished notes (1979).
- Laudisa F.,"The EPR Argument in a Relational Interpretation of Quantum Mechanics", *Foundations of Physics Letters* **14** (2001): 119-132.
- Mermin N.D., "What is quantum mechanics trying to tell us?", *American Journal of Physics* **66** (1998): 753-767.
- Rovelli C.,"Relational quantum mechanics " *International Journal of Theoretical Physics* **35** (1996): 1637-1678.
- Rovelli C.,"Half way through the woods", in J Earman and JD Norton (eds.) *The Cosmos of Science* University of Pittsburgh Press and Universitäts-Verlag Konstanz (1997).
- Rovelli C.," ‘Incerto tempore, incertisque loci’: Can we compute the exact time at which a quantum measurement happens?", *Foundations of Physics* **28** (1998): 1031-1043.
- Saunders S., "Relativism", in R. Clifton (ed.), Perspectives on Quantum Reality, Kluwer (1996)
- Saunders S., "Time, quantum mechanics and probability", Synthese 114 (1998), pp. 373-404.

- Smolin L.,"The Bekenstein bound, topological quantum field theory and pluralistic quantum field theory", Penn State preprint CGPG-95/8-7, 1995, Los Alamos Archives gr-qc/9508064.

Other Internet Resources

[Please contact the authors with suggestions.]

Related Entries

action at a distance | properties | quantum mechanics | quantum mechanics: collapse theories | quantum mechanics: Everett's relative-state formulation of | quantum mechanics: modal interpretations of | quantum theory: measurement in | quantum theory: quantum entanglement and information | quantum theory: quantum logic and probability theory

Federico Laudisa
federico.laudisa@unimib.it
and
Carlo Rovelli
rovelli@cpt.univ-mrs.fr

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 4, 2002
Content last modified: February 4, 2002

The Identity Theory of Mind

The identity theory of mind holds that states and processes of the mind are identical to states and processes of the brain. Strictly speaking, it need not hold that the mind is identical to the brain. Idiomatically we do use ‘She has a good mind’ and ‘She has a good brain’ interchangeably but we would hardly say ‘Her mind weighs fifty ounces’. Here I take identifying mind and brain as being a matter of identifying processes and perhaps states of the mind and brain. Consider an experience of pain, or of seeing something, or of having a mental image. The identity theory of mind is to the effect that these experiences just *are* brain processes, not merely *correlated with* brain processes.

Some philosophers hold that though experiences are brain processes they nevertheless have fundamentally non-physical, psychical, properties, sometimes called ‘qualia’. Here I shall take the identity theory as denying the existence of such irreducible non-physical properties. Some identity theorists give a behaviouristic analysis of mental *states*, such as beliefs and desires, but others, sometimes called ‘central state materialists’, say that mental states are actual brain states. Identity theorists often describe themselves as ‘materialists’ but ‘physicalists’ may be a better word. That is, one might be a materialist about mind but nevertheless hold that there are entities referred to in physics that are not happily described as ‘material’.

In taking the identity theory (in its various forms) as a species of physicalism, I should say that this is an ontological, not a translational physicalism. It would be absurd to try to translate sentences containing the word ‘brain’ or the word ‘sensation’ into sentences about electrons, protons and so on. Nor can we so translate sentences containing the word ‘tree’. After all ‘tree’ is largely learned ostensively, and is not even part of botanical classification. If we were small enough a dandelion might count as a tree. Nevertheless a physicalist could say that trees are complicated physical mechanisms. The physicalist will deny strong emergence in the sense of some philosophers, such as Samuel Alexander and possibly C.D. Broad . The latter remarked (Broad 1937) that as far as was known at that time the properties of common salt cannot be deduced from the properties of sodium in isolation and of chlorine in isolation. (He put it too epistemologically: chaos theory shows that even in a deterministic theory theory physical consequences can outrun predictability.) Of course the physicalist will not deny the harmless sense of "emergence" in which an apparatus is not just a jumble of its parts (Smart 1981).

Historical Antecedents

The identity theory as I understand it here goes back to U.T. Place and Herbert Feigl in the 1950s. Historically philosophers and scientists, for example Leucippus, Hobbes, La Mettrie, and d'Holbach, as well as Karl Vogt who made the preposterous remark (perhaps not meant to be taken too seriously) that the brain secretes thought as the liver secretes bile, have embraced materialism. However, here I shall date interest in the identity theory from the pioneering papers ‘Is Consciousness a Brain Process?’ by U.T. Place (Place 1956) and H. Feigl ‘The "Mental" and the "Physical"’ (Feigl 1958). Nevertheless mention should be made of suggestions by Rudolf Carnap (1932, p. 127), H. Reichenbach (1938) and M. Schlick (1935). Reichenbach said that mental events can be identified by the corresponding stimuli and responses much as the (possibly unknown) internal state of a photo-electric cell can be identified by the stimulus (light falling on it) and response (electric current flowing) from it. In both cases the internal states can be physical states. However Carnap did regard the identity as a linguistic recommendation rather than as asserting a question of fact. See his ‘Herbert Feigl on Physicalism’ in Schilpp (1963), especially p. 886. The psychologist E.G. Boring (1933) may well have been the first to use the term ‘identity theory’. See Place (1990).

Place's very original and pioneering paper was written after discussions at the University of Adelaide with Smart and C.B. Martin. For recollections of Martin's contributions to the discussion see Place (1989) ‘Low Claim Assertions’ in Heil (1989). Smart at the time argued for a behaviourist position in which mental events were elucidated purely in terms of hypothetical propositions about behaviour, as well as first person reports of experiences which Gilbert Ryle regarded as ‘avowals’. Avowals were thought of as mere pieces of behaviour, as if saying that one had a pain was just doing a sophisticated sort of wince. Smart saw Ryle's theory as friendly to physicalism though that was not part of Ryle's motivation. Smart hoped that the hypotheticals would ultimately be explained by neuroscience and cybernetics. Being unable to refute Place, and recognizing the unsatisfactoriness of Ryle's treatment of inner experience, to some extent recognized by Ryle himself (Ryle 1949, p. 240), Smart soon became converted to Place's view (Smart 1959). In this he was also encouraged and influenced by Feigl's ‘"The Mental" and the "Physical" ’ (Feigl 1958, 1967). Feigl's wide ranging contribution covered many problems, including those connected with intentionality, and he introduced the useful term ‘nomological danglers’ for the dualists' supposed mental-physical correlations. They would dangle from the

nomological net of physical science and should strike one as implausible excrescences on the fair face of science. Feigl (1967) contains a valuable ‘Postscript’.

The Nature of the Identity Theory

Place spoke of constitution rather than of identity. One of his examples is ‘This table is an old packing case’. Another is ‘lightning is an electric discharge’. Indeed this latter was foreshadowed by Place in his earlier paper ‘The Concept of Heed’ (Place 1954), in which he took issue with Ryle's behaviourism as it applied to concepts of consciousness, sensation and imagery. Place remarked (p. 255)

> The logical objections which might be raised to the statement ‘consciousness is a process in the brain’ are no greater than the logical objections which might be raised to the statement ‘lightning is a motion of electric charges’.

It should be noticed that Place was using the word ‘logical’ in the way that it was used at Oxford at the time, not in the way that it is normally used now. One objection was that ‘sensation’ does not mean the same as ‘brain process’. Place's reply was to point out that ‘this table’ does not mean the same as ‘this old packing case’ and ‘lightning’ does not mean the same as ‘motion of electric charges’. We find out whether this is a table in a different way from the way in which we find out that it is an old packing case. We find out whether a thing is lightning by looking and that it is a motion of electric charges by theory and experiment. This does not prevent the table being identical with the old packing case and the perceived lightning being nothing other than an electric discharge. Feigl and Smart put the matter more in terms of the distinction between meaning and reference. ‘Sensation’ and ‘brain process’ may differ in meaning and yet have the same reference. ‘Very bright planet seen in the morning’ and ‘very bright planet seen in the evening’ both refer to the same entity Venus. (Of course these expressions *could* be construed as referring to different things, different sequences of temporal stages of Venus, but not necessarily or most naturally so.)

There did seem to be a tendency among philosophers to have thought that identity statements needed to be necessary and a priori truths. However identity theorists have treated ‘sensations are brain processes’ as contingent. We had to *find out* that the identity holds. Aristotle, after all, thought that the brain was for cooling the blood. Descartes thought that consciousness is immaterial.

It was sometimes objected that sensation statements are incorrigible whereas statements about brains are corrigible. The inference was made that there must be something different about sensations. Ryle and in effect Wittgenstein toyed with the attractive but quite implausible notion that ostensible reports of immediate experience are not really reports but are ‘avowals’, as if my report that I have toothache is just a sophisticated sort of wince. Place, influenced by Martin, was able to explain the relative incorrigibility of sensation statements by their low claims: ‘I see a bent oar’ makes a bigger claim than ‘It looks to me that there is a bent oar’. Nevertheless my sensation and my putative awareness of the sensation are distinct existences and so, by Hume's principle, it must be possible for one to occur without the other. One should deny anything other than a relative incorrigibility (Place 1989).

As remarked above, Place preferred to express the theory by the notion of constitution, whereas Smart preferred to make prominent the notion of identity as it occurs in the axioms of identity in logic. So Smart had to say that if sensation X is identical to brain process Y then if Y is between my ears and is straight or circular (absurdly to oversimplify) then the sensation X is between my ears and is straight or circular. Of course it is not presented to us as such in experience. Perhaps only the neuroscientist could know that it is straight or circular. The professor of anatomy might be identical with the dean of the medical school. A visitor might know that the professor hiccups in lectures but not know that the dean hiccups in lectures.

Phenomenal Properties and Topic-Neutral Analyses

Someone might object that the dean of the medical school does not *qua* dean hiccup in lectures. *Qua* dean he goes to meetings with the vice-chancellor. This is not to the point but there is a point behind it. This is that *the property* of being the professor of anatomy is not identical with *the property* of being the dean of the medical school. The question might be asked, that even if sensations are identical with brain processes, are there not introspected non-physical properties of sensations that are not identical with properties of brain processes? How would a physicalist identity theorist deal with this? The answer (Smart 1959) is that the properties of experiences are 'topic neutral'. Smart adapted the words 'topic-neutral' from Ryle, who used them to characterise words such as 'if, 'or', 'and', 'not', 'because'. If you overheard only these words in a conversation you would not be able to tell whether the conversation was one of mathematics, physics, geology, history, theology, or any other subject. Smart used the words 'topic neutral' in the narrower sense of being neutral between physicalism and dualism. For example 'going on', 'occurring', 'intermittent', 'waxing', 'waning' are topic neutral. So is 'me' in so far as it refers to the utterer of the sentence in question. Thus to say that a sensation is caused by lightning or the presence of a cabbage before my eyes leaves it open as to whether the sensation is non-physical as the dualist believes or is physical as the materialist believes. This sentence also is neutral as to whether the properties of the sensation are physical or whether some of them are irreducibly psychical. To see how this idea can be applied to the present purpose let us consider the following example.

Suppose that I have a yellow, green and purple striped mental image. We may also introduce the philosophical term 'sense datum' to cover the case of seeing or seeming to see something yellow, green and purple: we say that we have a yellow, green and purple sense datum. That is I would see or seem to see, for example, a flag or an array of lamps which is green, yellow and purple striped. Suppose also, as seems plausible, that there is nothing yellow, green and purple striped in the brain. Thus it is important for identity theorists to say (as indeed they have done) that sense data and images are not part of the furniture of the world. 'I have a green sense datum' is really just a way of saying that I see or seem to see something that really is green. This move should not be seen as merely an *ad hoc* device, since Ryle and J.L. Austin, in effect Wittgenstein, and others had provided arguments, as when Ryle argued that mental images were not a sort of ghostly picture postcard. Place characterised the fallacy of thinking that when we perceive something green we are perceiving something green in the mind as 'the phenomenological fallacy'. He characterizes this fallacy (Place 1956):

> the mistake of supposing that when the subject describes his experience, when he describes how things look, sound, smell, taste, or feel to him, he is describing the literal properties of objects and events on a peculiar sort of internal cinema or television screen, usually referred to in the modern psychological literature as the ‘phenomenal field’.

Of course, as Smart recognised, this leaves the identity theory dependent on a physicalist account of colour . His early account of colour (1961) was too behaviourist, and could not deal, for example, with the reversed spectrum problem, but he later gave a realist and objectivist account (Smart 1975). Armstrong had been realist about colour but Smart worried that if so colour would be a very idiosyncratic and disjunctive concept, of no cosmic importance, of no interest to extraterrestrials (for instance) who had different visual systems. Prompted by Lewis in conversation Smart came to realize that this was no objection to colours being objective properties.

One first gives the notion of a normal human percipient with respect to colour for which there are objective tests in terms of ability to make discriminations with respect to colour. This can be done without circularity. Thus ‘discriminate with respect to colour’ is a more primitive notion than is that of colour. (Compare the way that in set theory ‘equinumerous’ is antecedent to ‘number’.) Then Smart elucidated the notion of colour in terms of the discriminations with respect to colour of normal human percipients in normal conditions (say cloudy Scottish daylight). This account of colour may be disjunctive and idiosyncratic. (Maxwell's equations might be of interest to Alpha Centaurians but hardly our colour concepts.) Anthropocentric and disjunctive they may be, but objective none the less. David R. Hilbert (1987) identifies colours with reflectances, thus reducing the idiosyncrasy and disjunctiveness. A few epicycles are easily added to deal with radiated light, the colours of rainbows or the sun at sunset and the colours due to diffraction from feathers. John Locke was on the right track in making the secondary qualities objective as powers in the object, but erred in making these powers to be powers to produce ideas in the mind rather than to make behavioural discriminations. (Also Smart would say that if powers are dispositions we should treat the secondary qualities as the categorical bases of these powers, e.g. in the case of colours properties of the surfaces of objects.) Locke's view suggested that the ideas have mysterious qualia observed on the screen of an internal mental theatre. However to do Locke justice he does not talk in effect of ‘red ideas’ but of ‘ideas of red’. Philosophers who elucidate ‘is red’ in terms of ‘looks red’ have the matter the wrong way round (Smart 1995).

Let us return to the issue of us having a yellow, purple and green striped sense datum or mental image and yet there being no yellow, purple and green striped thing in the brain. The identity theorist (Smart 1959) can say that sense data and images are not real things in the world: they are like the average plumber. Sentences ostensibly about the average plumber can be translated into, or elucidated in terms of, sentences about plumbers. So also there is having a green sense datum or image but not sense data or images, and the having of a green sense datum or image is not itself green. So it can, so far as this goes, easily be a brain process which is not green either.

Thus Place (1956, p. 49):

> When we describe the after-image as green... we are saying that we are having the sort of experience which we normally have when, and which we have learned to described as, looking at a green patch of light.

and Smart (1959) says:

> When a person says ‘I see a yellowish-orange after-image’ he is saying something like this: "*There is something going on which is like what is going on when* I have my eyes open, am awake, and there is an orange illuminated in good light in front of me".

Quoting these passages, David Chalmers (1996, p. 360) objects that if ‘something is going on’ is construed broadly enough it is inadequate, and if it is construed narrowly enough to cover only experiential states (or processes) it is not sufficient for the conclusion. Smart would counter this by stressing the word ‘typically’. Of course a lot of things go on in me when I have a yellow after image (for example my heart is pumping blood through my brain). However they do not *typically* go on then: they go on at other times too. Against Place Chalmers says that the word ‘experience’ is unanalysed and so Place's analysis is insufficient towards establishing an identity between sensations and brain processes. As against Smart he says that leaving the word ‘experience’ out of the analysis renders it inadequate. That is, he does not accept the ‘topic-neutral’ analysis. Smart hopes, and Chalmers denies, that the account in terms of ‘typically of’ saves the topic-neutral analysis. In defence of Place one might perhaps say that it is not clear that the word ‘experience’ cannot be given a topic neutral analysis, perhaps building on Farrell (1950). If we do not need the word ‘experience’ neither do we need the word ‘mental’. Rosenthal (1994) complains (against the identity theorist) that experiences have some characteristically mental properties, and that ‘We inevitably lose the distinctively mental if we construe these properties as neither physical nor mental’. Of course to be topic neutral is to be able to be both physical and mental, just as arithmetic is. There is no need for the word ‘mental’ itself to occur in the topic neutral formula. ‘Mental’, as Ryle (1949) suggests, in its ordinary use is a rather grab-bag term, ‘mental arithmetic’, ‘mental illness’, etc. with which an identity theorist finds no trouble.

Causal Role Theories

In their accounts of mind, David Lewis and D.M. Armstrong emphasise the notion of causality. Lewis's 1966 was a particularly clear headed presentation of the identity theory in which he says (I here refer to the reprint in Lewis 1983, p. 100):

> My argument is this: The definitive characteristic of any (sort of) experience as such is its causal role, its syndrome of most typical causes and effects. But we materialists believe that these causal roles which belong by analytic necessity to experiences belong in fact to certain physical states. Since these physical states possess the definitive character of experiences, they must be experiences.

Similarly, Robert Kirk (1999) has argued for the impossibility of zombies. If the supposed zombie has all the behavioural and neural properties ascribed to it by those who argue from the possibility of zombies against materialism, then the zombie is conscious and so not a zombie.

Thus there is no need for explicit use of Ockham's Razor as in Smart (1959) though not in Place (1956). (See Place 1960.) Lewis's paper was extremely valuable and already there are hints of a marriage between the identity theory of mind and so-called ‘functionalist’ ideas that are explicit in Lewis 1972 and 1994. In his 1972 (‘Psychophysical and Theoretical Identifications’) he applies ideas in his more formal paper ‘How to Define Theoretical Terms’ (1970). Folk psychology contains words such as ‘sensation’, ‘perceive’, ‘belief, ‘desire’, ‘emotion’, etc. which we recognise as psychological. Words for colours, smells, sounds, tastes and so on also occur. One can regard common sense platitudes containing both these sorts of these words as constituting a theory and we can take them as theoretical terms of common sense psychology and thus as denoting whatever entities or sorts of entities uniquely realise the theory. Then if certain neural states do so too (as we believe) then the mental states must be these neural states. In his 1994 he allows for tact in extracting a consistent theory from common sense. One cannot uncritically collect platitudes, just as in producing a grammar, implicit in our speech patterns, one must allow for departures from what on our best theory would constitute grammaticality.

A great advantage of this approach over the early identity theory is its holism. Two features of this holism should be noted. One is that the approach is able to allow for the causal interactions between brain states and processes themselves, as well as in the case of external stimuli and responses. Another is the ability to draw on the notion of Ramseyfication of a theory. F.P. Ramsey had shown how to replace the theoretical terms of a theory such as ‘the property of being an electron’ by ‘the property X such that...’. so that when this is done for all the theoretical terms, we are left only with ‘property X such that’, ‘property Y such that’ etc. Take the terms describing behaviour as the observation terms and psychological terms as the theoretical ones of folk psychology. Then Ramseyfication shows that folk psychology is compatible with materialism. This seems right, though perhaps the earlier identity theory deals more directly with reports of immediate experience.

The causal approach was also characteristic of D.M. Armstrong's careful conceptual analysis of mental states and processes, such as perception and the secondary qualities, sensation, consciousness, belief, desire, emotion, voluntary action, in his *A Materialist Theory of the Mind* (1968a) with a second edition (1993) containing a valuable new preface. Parts I and II of this book are concerned with conceptual analysis, paving the way for a contingent identification of mental states and processes with material ones. As had Brian Medlin, in an impressive critique of Ryle and defence of materialism (Medlin 1967), Armstrong preferred to describe the identity theory as ‘Central State Materialism’. Independently of Armstrong and Lewis, Medlin's central state materialism depended, as theirs did, on a causal analysis of concepts of mental states and processes. See Medlin 1967, and 1969 (including endnote 1).

Mention should particularly be made here of two of Armstrong's other books, one on perception (1961), and one on bodily sensations, (1962). Armstrong thought of perception as *coming to believe by means of the senses* (compare also Pitcher 1971). This combines the advantages of Direct Realism with hospitality towards the scientific causal story which had been thought to have supported the earlier representative

theory of perception. Armstrong regarded bodily sensations as perceptions of states of our body. Of course the latter may be mixed up with emotional states, as an itch may include a propensity to scratch, and contrariwise in exceptional circumstances pain may be felt without distress. However, Armstrong sees the central notion here as that of perception. This suggests a terminological problem. Smart had talked of visual sensations. These were not perceptions but something which occurred in perception. So in *this* sense of ‘sensation’ there should be bodily sensation sensations. The ambiguity could perhaps be resolved by using the word ‘sensing’ in the context of ‘visual’, ‘auditory’, ‘tactile’ and ‘bodily’, so that bodily sensations would be perceivings which involved introspectible ‘sensings’. These bodily sensations are perceptions and there can be misperceptions as when a person with his foot amputated can think that he has a pain in the foot. He has a sensing ‘having a pain in the foot’ but the world does not contain a pain in the foot, just as it does not contain sense data or images but does contain havings of sense data and of images.

Armstrong's central state materialism involved identifying beliefs and desires with states of the brain (1968a). Smart came to agree with this. On the other hand Place resisted the proposal to extend the identity theory to dispositional states such as beliefs and desires. He stressed that we do not have privileged access to our beliefs and desires. Like Ryle he thought of beliefs and desires as to be elucidated by means of hypothetical statements about behaviour and gave the analogy of the horsepower of a car (Place 1967). However he holds that the dispute here is not so much about the neural basis of mental states as about the nature of dispositions. His views on dispositions are argued at length in his debate with Armstrong and Martin (Armstrong, Martin and Place, T. Crane (ed.) 1996). Perhaps we can be relaxed about whether mental states such as beliefs and desires are dispositions or are topic neutrally described neurophysiological states and return to what seems to be the more difficult issue of consciousness. Causal identity theories are closely related to Functionalism, to be discussed in the next section. Smart had been wary of the notion of causality in metaphysics believing that it had no place in theoretical physics. However even so he should have admitted it in folk psychology and also in scientific psychology and biology generally, in which physics and chemistry are applied to explain generalisations rather than strict laws. If folk psychology uses the notion of causality, it is no matter if it is what Quine has called second grade discourse, involving the very contextual notions of modality.

Functionalism and Identity Theory

It has commonly been thought that the identity theory has been superseded by a theory called ‘functionalism’. It could be argued that functionalists greatly exaggerate their difference from identity theorists. Indeed some philosophers, such as Lewis (1972 and 1994) and Jackson, Pargetter and Prior (1982), see functionalism as a route towards an identity theory.

Like Lewis and Armstrong, functionalists define mental states and processes in terms of the their causal relations to behaviour but stop short of identifying them with their neural realisations. Of course the term ‘functionalism’ has been used vaguely and in different ways, and it could be argued that even the theories of Place, Smart and Armstrong were at bottom functionalist. The word ‘functionalist’ has affinities with that of ‘function’ in mathematics and also with that of ‘function’ in biology. In mathematics a function is

a set of ordered n-tuples. Similarly if mental processes are defined directly or indirectly by sets of stimulus-response pairs the definitions could be seen as ‘functional’ in the mathematical sense. However there is probably a closer connection with the term as it is used in biology, as one might define ‘eye’ by its function even though a fly's eye and a dog's eye are anatomically and physiologically very different. Functionalism identifies mental states and processes by means of their causal roles, and as noted above in connection with Lewis, we know that the functional roles are possessed by neural states and processes. (There are teleological and homuncular forms of functionalism, which I do not consider here.) Nevertheless an interactionist dualist such as the eminent neurophysiologist Sir John Eccles would (implausibly for most of us) deny that all functional roles are so possessed. One might think of folk psychology, and indeed much of cognitive science too, as analogous to a ‘block diagram’ in electronics. A box in the diagram might be labelled (say) ‘intermediate frequency amplifier’ while remaining) neutral as to the exact circuit and whether the amplification is carried out by a thermionic valve or by a transistor. Using terminology of F. Jackson and P. Pettit (1988, pp. 381-400) the ‘role state’ would be given by ‘amplifier’, the ‘realiser state’ would be given by ‘thermionic valve’, say. So we can think of functionalism as a ‘black box’ theory. This line of thought will be pursued in the next section.

Thinking very much in causal terms about beliefs and desires fits in very well not only with folk psychology but also with Humean ideas about the motives of action. Though this point of view has been criticised by some philosophers it does seem to be right, as can be seen if we consider a possible robot aeroplane designed to find its way from Melbourne to Sydney. The designer would have to include an electronic version of something like a map of south-eastern Australia. This would provide the ‘belief’ side. One would also have to program in an electronic equivalent of ‘go to Sydney’. This program would provide the ‘desire’ side. If wind and weather pushed the aeroplane off course then negative feedback would push the aeroplane back on to the right course for Sydney. The existence of purposive mechanisms has at last (I hope) shown to philosophers that there is nothing mysterious about teleology. Nor are there any great semantic problems over intentionality (with a ‘t’). Consider the sentence ‘Joe desires a unicorn’. This is not like ‘Joe kicks a football’. For Joe to kick a football there must be a football to be kicked, but there are no unicorns. However we can say ‘Joe desires-true of himself "possesses a unicorn" ’. Or more generally ‘Joe believes-true S’ or ‘Joe desires-true S’ where S is an appropriate sentence (Quine 1960, pp. 206-16). Of course if one does not want to relativise to a language one needs to insert ‘or some samesayer of S’ or use the word ‘proposition’, and this involves the notion of proposition or intertranslatability. Even if one does not accept Quine's notion of indeterminacy of translation, there is still fuzziness in the notions of ‘belief’ and ‘desire’ arising from the fuzziness of ‘analyticity’ and ‘synonymy’. The identity theorist could say that on any occasion this fuzziness is matched by the fuzziness of the brain state that constitutes the belief or desire. Just how many interconnections are involved in a belief or desire? On a holistic account such as Lewis's one need not suppose that individuation of beliefs and desires is precise, even though good enough for folk psychology and Humean metaethics. Thus the way in which the brain represents the world might not be like a language. The representation might be like a map. A map relates every feature on it to every other feature. Nevertheless maps contain a finite amount of information. They have not infinitely many parts, still less continuum many. We can think of beliefs as expressing the different bits of information that could be extracted from the map. Thinking in this way beliefs would correspond near enough to the individualist beliefs characteristic of folk and Humean psychology.

Type and Token Identity Theories

The notion ‘type’ and ‘token’ here comes by analogy from ‘type’ and ‘token’ as applied to words. A telegram ‘love and love and love’ contains only two type words but in another sense, as the telegraph clerk would insist, it contains five words (‘token words’). Similarly a particular pain (more exactly a having a pain) according to the token identity theory is identical to a particular brain process. A functionalist could agree to this. Functionalism came to be seen as an improvement on and as inconsistent with the identity theory because of the correct assertion that a functional state can be realised by quite different brain states: thus a functional state might be realised by a silicon based brain as well as by a carbon based brain, and leaving robotics or science fiction aside, my feeling of toothache could be realised by a different neural process from what realises your toothache.

As far as this goes, at any rate, a functionalist can accept token identities. Functionalists commonly deny type identities. However Jackson, Pargetter and Prior (1982) and Braddon-Mitchell and Jackson (1996) argue that this is an over-reaction on the part of the functionalist. (Indeed they see functionalism as a route to the identity theory.) The functionalist may define mental states as having some state or other (e.g., carbon based or silicon based) which accounts for the functional properties. The functionalist second order state is a state of having some first order state or other which causes or is caused by the behaviour to which the functionalist alludes. In this way we have a second order type theory. Compare brittleness. The brittleness of glass and the brittleness of biscuits are both the state of having some property which explains their breaking, though the first order physical property may be different in the two cases. This way of looking at the matter is perhaps more plausible in relation to mental states such as beliefs and desires than it is to immediately reported experiences. When I report a toothache I do seem to be concerned with first order properties, even though topic neutral ones.

If we continue to concern ourselves with first order properties, we could say that the type-token distinction is not an all or nothing affair. We could say that human experiences are brain processes of one lot of sorts and Alpha Centaurian experiences are brain processes of another lot of sorts. We could indeed propose much finer classifications without going to the limit of mere token identities.

How restricted should be the restriction of a restricted type theory? How many hairs must a bald man have no more of? An identity theorist would expect his toothache today to be very similar to his toothache yesterday. He would expect his toothache to be quite similar to his wife's toothache. He would expect his toothache to be somewhat similar to his cat's toothache. He would not be confident about similarity to an extra-terrestrial's pain. Even here, however, he might expect some similarities of wave form or the like.

Even in the case of the similarity of my pain now to my pain ten minutes ago, there will be unimportant dissimilarities, and also between my pain and your pain. Compare topiary, making use of an analogy exploited by Quine in a different connection. In English country gardens the tops of box hedges are often cut in various shapes, for example peacock shapes. One might make generalizations about peacock

shapes on box hedges, and one might say that all the imitation peacocks on a particular hedge have the same shape. However if we approach the two imitation peacocks and peer into them to note the precise shapes of the twigs that make them up we will find differences. Whether we say that two things are similar or not is a matter of abstractness of description. If we were to go to the limit of concreteness the types would shrink to single membered types, but there would still be no ontological difference between identity theory and functionalism.

An interesting form of token identity theory is the anomalous monism of Davidson 1980. Davidson argues that causal relations occur under the neural descriptions but not under the descriptions of psychological language. The latter descriptions use intentional predicates, but because of indeterminacy of translation and of interpretation, these predicates do not occur in law statements. It follows that mind-brain identities can occur only on the level of individual (token) events. It would be beyond the scope of the present essay to consider Davidson's ingenious approach, since it differs importantly from the more usual forms of identity theory.

Consciousness

Place answered the question ‘Is Consciousness a Brain Process?’ in the affirmative. But what sort of brain process? It is natural to feel that there is something ineffable about which no mere neurophysiological process (with only physical intrinsic properties) could have. There is a challenge to the identity theorist to dispel this feeling.

Suppose that I am riding my bicycle from my home to the university. Suddenly I realise that I have crossed a bridge over a creek, gone along a twisty path for half a mile, avoided oncoming traffic, and so on, and yet have no memories of all this. In one sense I was conscious: I was perceiving, getting information about my position and speed, the state of the bicycle track and the road, the positions and speeds of approaching cars, the width of the familiar narrow bridge. But in another sense I was not conscious: I was on ‘automatic pilot’. So let me use the word ‘awareness’ for this automatic or subconscious sort of consciousness. Perhaps I am not one hundred percent on automatic pilot. For one thing I might be absent minded and thinking about philosophy. Still, this would not be relevant to my bicycle riding. One might indeed wonder whether one is ever one hundred percent on automatic pilot, and perhaps one hopes that one isn't, especially in Armstrong's example of the long distance truck driver (Armstrong 1962). Still it probably does happen, and if it does the driver is conscious only in the sense that he or she is alert to the route, of oncoming traffic etc., i.e. is perceiving in the sense of ‘coming to believe by means of the senses’. The driver gets the beliefs but is not aware of doing so. There is no suggestion of ineffability in this sense of ‘consciousness’, for which I shall reserve the term ‘awareness’.

For the full consciousness, the one that puzzles us and suggests ineffability, we need the sense elucidated by Armstrong in a debate with Norman Malcolm (Armstrong and Malcolm 1962, p. 110). Somewhat similar views have been expressed by other philosophers, such as Savage (1976), Dennett (1991), Lycan (1996), Rosenthal (1996). In the debate with Norman Malcolm, Armstrong compared consciousness with proprioception. In recent conversation he was prepared to say that it *is* proprioception. A case of

proprioception occurs when with our eyes shut and without touch we are immediately aware of the angle at which one of our elbows is bent. That is, proprioception is a special sense, different from that of bodily sensation, in which we become aware of parts of our body. Now the brain is part of our body and so perhaps immediate awareness of a process in or a state of our brain deserves to be called ‘proprioception’. Thus the proprioception which constitutes consciousness, as distinguished from mere awareness, is a higher order awareness, a perception of one part of (or configuration in) our brain by the brain itself. Some may sense circularity here. If so let them suppose that the proprioception occurs in an in practice negligible time after the process proprioceptd. Then perhaps there can be proprioceptions of proprioceptions, proprioceptions of proprioceptions of proprioceptions, and so on up, though in fact the sequence will probably not go up more than two or three steps. The last proprioception in the sequence will not be proprioceptd, and this may help to explain our sense of the ineffability of consciousness. Compare Gilbert Ryle in *The Concept of Mind* on the systematic elusiveness of ‘I’ (Ryle 1949, pp. 195-8).

Place has argued that the function of the ‘automatic pilot’, to which he refers as ‘the zombie within’, is to alert consciousness to inputs which it identifies as problematic, while it ignores non-problematic inputs or re-routes them to output without the need for conscious awareness. For this view of consciousness see Place (1999).

Later Objections to the Identity Theory

Mention should here be made of influential criticisms of the identity theory by Saul Kripke and David Chalmers respectively. It will not be possible to discuss them in great detail, partly because of the fact that Kripke's remarks rely on views about modality, possible worlds semantics, and essentialism which some philosophers would want to contest, and because Chalmers' long and rich book would deserve a lengthy answer. Kripke (1980) calls an expression a rigid designator if it refers to the same object in every possible world. Or in counterpart theory it would have an exactly similar counterpart in every possible world. It seems to me that what we count as counterparts is highly contextual. Take the example ‘water is H_20’. In another world, or in a twin earth in our world as Putnam imagines (1975), the stuff found in rivers, lakes, the sea would not be H_20 but XYZ and so would not be water. This is certainly giving preference to real chemistry over folk chemistry, and so far I applaud this. There are therefore contexts in which we say that on twin earth or the envisaged possible world the stuff found in rivers would not be water. Nevertheless there are contexts in which we could envisage a possible world (write a science fiction novel) in which being found in rivers and lakes and the sea, assuaging thirst and sustaining life was more important than the chemical composition and so XYZ would be the counterpart of H_20.

Kripke considers the identity ‘heat = molecular motion’, and holds that this is true in every possible world and so is a necessary truth. Actually the proposition is not quite true, for what about radiant heat? What about heat as defined in classical thermodynamics which is ‘topic neutral’ compared with statistical thermodynamics? Still, suppose that heat has an essence and that it is molecular motion, or at least is in

the context envisaged. Kripke says (1980, p. 151) that when we think that molecular motion might exist in the absence of heat we are confusing this with thinking that the molecular motion might have existed without being *felt* as heat. He asks whether it is analogously possible that if pain is a certain sort of brain process that it has existed without being *felt* as pain. He suggests that the answer is ‘No’. An identity theorist who accepted the account of consciousness as a higher order perception could answer ‘Yes’. We might be aware of a damaged tooth and also of being in an agitation condition (to use Ryle's term for emotional states) without being aware of our awareness. An identity theorist such as Smart would prefer talk of ‘having a pain’ rather than of ‘pain’: pain is not part of the furniture of the world any more than a sense datum or the average plumber is. Kripke concludes (p. 152) that the

> apparent contingency of the connection between the physical state and the corresponding brain state thus cannot be explained by some sort of qualitative analogue as in the case of heat.

Smart would say that there is a sense in which the connection of sensations (sensings) and brain processes is only half contingent. A complete description of the brain state or process (including causes and effects of it) would imply the report of inner experience, but the latter, being topic neutral and so very abstract would not imply the neurological description.

Chalmers (1996) in the course of his exhaustive study of consciousness developed a theory of non-physical qualia which to some extent avoids the worry about nomological danglers. The worry expressed by Smart (1959) is that if there were non-physical qualia there would, most implausibly, have to be laws relating neurophysiological processes to apparently simple properties, and the correlation laws would have to be fundamental, mere danglers from the nomological net (as Feigl called it) of science. Chalmers counters this by supposing that the qualia are not simple but unknown to us are made up of simple proto-qualia, and that the fundamental laws relating these to physical entities relate them to *fundamental* physical entities. His view comes to a rather interesting panpsychism. On the other hand if the topic neutral account is correct, then qualia are no more than points in a multidimensional similarity space, and the overwhelming plausibility will fall on the side of the identity theorist.

On Chalmers' view how are we aware of non-physical qualia? It has been suggested above that this inner awareness is proprioception of the brain by the brain. But what sort of story is possible in the case of awareness of a quale? Chalmers could have some sort of answer to this by means of his principle of coherence according to which the causal neurological story parallels the story of succession of qualia. It is not clear however that this would make us aware of the qualia. The qualia do not seem to be needed in the physiological story of how an antelope avoids a tiger.

People often think that even if a robot could scan its own perceptual processes this would not mean that the robot was conscious. This appeals to our intuitions, but perhaps we could reverse the argument and say that because the robot can be aware of its awareness the robot *is* conscious. I have given reason above to distrust intuitions, but in any case Chalmers comes some of the way in that he toys with the idea that a thermostat has a sort of proto-qualia. The dispute between identity theorists (and physicalists

generally) and Chalmers comes down to our attitude to phenomenology. Certainly walking in a forest, seeing the blue of the sky, the green of the trees, the red of the track, one may find it hard to believe that our qualia are merely points in a multidimensional similarity space. But perhaps that is what *it is like* (to use a phrase that can be distrusted) to be aware of a point in a multidimensional similarity space. One may also, as Place would suggest, be subject to 'the phenomenological fallacy'. At the end of his book Chalmers makes some speculations about the interpretation of quantum mechanics. If they succeed then perhaps we could envisage Chalmers' theory as integrated into physics and him as a physicalist after all. However it could be doubted whether we need to go down to the quantum level to understand consciousness or whether consciousness is relevant to quantum mechanics.

Bibliography

- Armstrong, D.M. 1961: *Perception and the Physical World*, London, Routledge.
- Armstrong, D.M. 1961: *Bodily Sensations*, London, Routledge.
- Armstrong, D.M. 1962: 'Consciousness and Causality', and 'Reply'. In Armstrong, D.M. and Malcolm, N. *Consciousness and Causality*, Oxford, Blackwell.
- Armstrong, D.M. 1968a: *A Materialist Theory of the Mind*, London, Routledge. Second Edition with new preface 1993.
- Armstrong, D.M. 1968b: 'The Headless Woman Illusion and the Defence of Materialism', *Analysis*, 29, 48-49.
- Armstrong, D.M. 1999: *The Mind-Body Problem: An Opinionated Introduction*, Boulder, Colorado, Westview Press.
- Armstrong, D.M., Martin, C.B. and Place, U.T. 1996: *Dispositions: A Debate*, T. Crane (ed.), London, Routledge.
- Braddon-Mitchell, D. and Jackson, F. 1996: *Philosophy of Mind and Cognition*, Oxford, Blackwell.
- Broad, C.D. 1937: *The Mind and its Place in Nature*, London, Routledge and Kegan Paul.
- Campbell, K. 1984: *Body and Mind*, Indiana, University of Notre Dame Press.
- Carnap, R. 1932: 'Psychologie in Physikalischer Sprache', *Erkenntnis*, 3, 107-142. English translation in A.J. Ayer (ed.) *Logical Positivism*, Glencoe, Illinois, Free Press 1959.
- Carnap, R. 1963: 'Herbert Feigl on Physicalism'. In Schilpp, 1963, pp. 882-886.
- Chalmers, D.M. 1996: *The Conscious Mind*, New York, Oxford University Press.
- Clark, A. 1993: *Sensory Qualities*, Oxford, Oxford University Press.
- Davidson, D. 1980: 'Mental Events', 'The Material Mind' and 'Psychology as Part of Philosophy'. In Davidson, D. *Essays on Actions and Events*, Oxford, Clarendon Press.
- Dennett, D.C. 1991: *Consciousness Explained*, Boston, Little and Brown.
- Farrell, B.A. 1950: 'Experience', *Mind* 50, 170-198.
- Feigl, H. 1958: 'The "Mental" and the "Physical" '. In Feigl, H., Scriven, M. and Maxwell, G. (eds.) *Concepts, Theories and the Mind-Body Problem*, Minneapolis, Minnesota Studies in the Philosophy of Science, Vol. 2, reprinted with a Postscript in Feigl 1967.
- Feigl, H. 1967: *The 'Mental' and the 'Physical'*, *The Essay and a Postscript*, Minneapolis, University of Minnesota Press.

- Heil, J. 1989: *Cause, Mind and Reality, Essays Honoring C.B.Martin*, Dordrecht, Kluwer Academic Publishers.
- Hilbert, D.R. 1987: *Color and Color Perception: A Study in Anthropocentric Realism*, Stanford, CSLI.
- Hill, C.S. 1991: *Sensations: A Defense of Type Materialism*, Cambridge, Cambridge University Press.
- Jackson, F. 1998: ‘What Mary didn't know’, and ‘Postscript on qualia’. In Jackson, F. *Mind, Method and Conditionals*, London, Routledge.
- Jackson, F. and Pettit, P. 1988: ‘Functionalism and Broad Content’, *Mind, 97*. 381-400.
- Jackson, F., Pargetter, R. and Prior, E. 1982: ‘Functionalism and Type-Type Identity Theories’, *Philosophical Studies*, 42, 209-225.
- Kirk, R. 1999: ‘Why There Couldn't be Zombies’, *Aristotelian Society*, Supp. Vol. 73, 1-16.
- Kripke, S. 1980: *Naming and Necessity*, Cambridge, Mass., Harvard University Press.
- Levin, M.E. 1979: *Metaphysics and the Mind-Body Problem*, Oxford, Clarendon Press.
- Lewis, D. 1966: ‘An Argument for the Identity Theory’, *Journal of Philosophy*, 63, 17-25.
- Lewis, D. 1970: ‘How to Define Theoretical Terms’, *Journal of Philosophy*, 67, 427-446.
- Lewis, D. 1972: ‘Psychophysical and Theoretical Identifications’, *Australasian Journal of Philosophy*, 50, 249-258.
- Lewis, D. 1983: ‘Mad Pain and Martian Pain’ and ‘Postscript’. In Lewis D. *Philosophical Papers, Vol.* 1, Oxford, Oxford University Press.
- Lewis, D. 1989: ‘What Experience Teaches’. In Lycan, W. (ed.) *Mind and Cognition*, Oxford, Blackwell
- Lewis, D. 1994: ‘Reduction of Mind’. In Guttenplan, S. (ed.) *A Companion to the Philosophy of Mind*, Oxford, Blackwell.
- Lycan, W.G. 1996: *Consciousness and Experience*, Cambridge, Mass., M.I.T. Press.
- Medlin, B.H. 1967: ‘Ryle and the Mechanical Hypothesis’. In Presley, C.F. (ed.) *The Identity Theory of Mind*, St. Lucia, Queensland, Queensland University Press.
- Medlin, B.H. 1969: ‘Materialism and the Argument from Distinct Existences’. In MacIntosh, J.J. and Coval, S., *The Business of Reason*, London: Routledge and Kegan Paul.
- Pitcher, G. 1971: *A Theory of Perception*, Princeton, N.J., Princeton University Press.
- Place, U.T. 1954: ‘The Concept of Heed’, *British Journal of Psychology*, 45, 243-255.
- Place, U.T. 1956: ‘Is Consciousness a Brain Process?’, *British Journal of Psychology*, 47, 44-50,
- Place, U.T. 1960: ‘Materialism as a Scientific Hypothesis’, *Philosophical Review*, 69, 101-104.
- Place, U.T. 1967: ‘Comments on Putnam's "Psychological Predicates"’. In Capitan, W.H. and Merrill, D.D. (eds) *Art, Mind and Religion*, Pittsburgh, Pittsburgh University Press.
- Place, U.T. 1988: ‘Thirty Years on--Is Consciousness still a Brain Process?’, *Australasian Journal of Philosophy*, 66, 208-219.
- Place, U.T. 1989: ‘Low Claim Assertions’. In Heil, J. (ed.) *Cause, Mind and Reality: Essays Honoring C.B. Martin*, Dordrecht, Kluwer Academic Publishers.
- Place, U.T. 1990: ‘E.G. Boring and the Mind-Brain Identity Theory’, *British Psychological Society, History and Philosophy of Science Newsletter*, 11, 20-31.
- Place, U.T. 1999: ‘Connectionism and the Problem of Consciousness’, *Acta Analytica*, 22, 197-226.

- Putnam, H. 1960: ‘Minds and Machines’. In Hook, S. (ed.) *Dimensions of Mind*, New York, New York University Press.
- Putnam, H. 1975: ‘The Meaning of "Meaning" ’. In Putnam, H. *Mind, Language and Reality*, Cambridge, Cambridge University Press.
- Quine, W.V. 1960: *Word and Object*, Cambridge, Mass., MIT Press.
- Reichenbach, H. 1938: *Experience and Prediction*, Chicago, Chicago University
- Rosenthal, D.M. 1994: ‘Identity Theories’. In Guttenplan, S. (ed.), *A Companion to the Philosophy of Mind*, Oxford, Blackwell, pp. 348-355.
- Rosenthal, D.M. 1996: ‘A Theory of Consciousness’. In Block, N., Flanagan, O. and Güzeldere, G. (eds) *The Nature of Consciousness*, Cambridge, Mass., MIT Press.
- Ryle, G. 1949: *The Concept of Mind*, London, Hutchinson.
- Savage, C.W. 1976: ‘An Old Ghost in a New Body’. In Globus, G.G., Maxwell, G. and Savodnik, I., (eds.), *Consciousness and the Brain*, New York, Plenum Press.
- Schilpp, P.A. (ed.) 1963: *The Philosophy of Rudolf Carnap*, La Salle, Illinois, Open Court.
- Schlick, M. 1935: ‘De la Relation des Notions Psychologiques et des Notions Physiques’, *Revue de Synthese*, 10, 5-26. English translation in Feigl, H. and Sellars, W., *Readings in Philosophical Analysis*, New York, Appleton-Century Crofts, 1949.
- Smart, J.J.C. 1959: ‘Sensations and Brain Processes’, *Philosophical Review*, 68, 141-156.
- Smart, J.J.C. 1961: ‘Colours’, *Philosophy*, 36, 128-142.
- Smart, J.J.C. 1963: ‘Materialism’, *Journal of Philosophy*, 60, 651-662.
- Smart, J.J.C. 1975: ‘On Some Criticisms of a Physicalist Theory of Colour’. In Chung-ying Cheng (ed.) *Philosophical Aspects of the Mind-Body Problem*, Honolulu, University of Hawaii Press.
- Smart, J.J.C. 1981: ‘Physicalism and Emergence’, *Neuroscience*, 6, 109-113.
- Smart, J.J.C. 1995: ‘"Looks Red" and Dangerous Talk’, *Philosophy*, 70, 545-554.

Other Internet Resources

- Identity Theories, an incomplete paper by U.T. Place, published in the Field Guide to Philosophy of Mind

Related Entries

consciousness | functionalism

Acknowledgements

I would like to express my thanks to David Armstrong, Frank Jackson and Ullin Place for comments on an earlier draft of this article and David Chalmers for careful editorial suggestions.

J. J. C. Smart
John.Smart@arts.monash.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 12, 2000
Content last modified: September 15, 2000

Propositional Attitude Reports

A person can be cognitively related to a proposition in many ways. These cognitive relations might be attributed in sentences like the following:

> Alicia believes that people walked on the Moon.
> Boris knows that people walked on the Moon.
> Carla fears that people walked on the Moon.
> Denzel hopes that people walked on the Moon.

Believing, knowing, fearing and hoping are four different attitudes towards the proposition that people walked on the moon.

Most propositional attitude attributions use a propositional attitude verb that is followed by a *that*-clause, a clause that includes a full sentence expressing a proposition. Attributions of cognitive relations to propositions can also take other kinds of clauses, though: *Alicia wanted to swim, Juan wanted Mary to succeed*, for example. These still attribute propositional attitudes, cognitive relations to an identifiable proposition (*Alicia will swim*, *Mary will succeed*), though the proposition is not so directly expressed.

In this discussion, we will look at some of the attempts to deal with a puzzle about propositional attitude attributions that was posed by Gottlob Frege in [1892]. Developing a semantic theory that deals satisfactorily with propositional attitude attributions has proved to be a very difficult project, and no fully satisfactory theory exists. In what follows, we will explore some of the theories developed to deal with the puzzle and note some of the problems with each theory.

- Other Internet Resources
- Related Entries

The Fregean Puzzle

The attribution of propositional attitudes raises problems in our attempts to provide a systematic semantics for language. Powerful considerations developed by Gottlob Frege [1892] suggest that words in propositional attitude ascriptions cannot function as they ordinarily do. Frege presents his puzzle as one about the relationship between the cognitive value of expressions and their ordinary reference. Developing a systematic semantics for belief attribution that responds adequately to Frege's puzzle has proved to be very difficult. Every theory seems to have problems that seriously compromise its viability.

The principle of semantic compositionality is fundamental to the Fregean puzzle. We can't learn language by learning each sentence's meaning individually, because mastering a language involves being able to recognize the meanings of an infinite variety of new sentences that we encounter. So our ability to understand language seems to require that we be able to discern the meaning of a sentence on the basis of our knowledge of the meanings of its parts and the way that they are put together. In a word, linguistic meaning must be *compositional* (or at least largely compositional).

The Fregean puzzle can be posed as a question about propositional attitude attributions. (We will use the verb ‘believe’ in the discussion of these puzzles. Similar puzzles arise with all propositional attitudes.) Consider the situation of Lois Lane, who is very familiar with Clark Kent, her fellow employee, and Superman, the hero she most admires, but who does not recognize that ‘Clark Kent’ and ‘Superman’ refer to the same person. We would ordinarily accept these claims about Lois:

(1) Lois believes that Superman is strong.

(2) Lois believes that Clark Kent is not strong.

It would even seem to be true that:

(3) Lois does not believe that Clark Kent is strong.

When we compare the belief attribution (1) with the belief attributions (2) and (3), it seems apparent that the names ‘Superman’ and ‘Clark Kent’ make different semantic contributions to the sentences in which they appear. In particular, it appears that if we replace ‘Superman’ in (1) by the ‘Clark Kent’, it will change a true sentence to a false one.

(4) Lois believes that Clark Kent is strong.

Sentence (4) seems false, even though it results from (1) by replacing one name by another that refers to the same individual. Since the names have the same reference, it seems that something other than the reference of the name must be relevant to the semantic evaluation of the belief attribution.

Without propositional attitude attributions, we might hope for a simpler situation for semantics, in which only the referent of a name is relevant to the evaluation of sentences that contain it.

> (5) Superman is strong.

> (6) Clark Kent is strong.

If (5) is true, then (6) is true as well. Even if Lois and others do not realize it, these sentences must have the same truth-value. So their objective semantics can be the same: each involves a reference to an individual and each predicates the same property of that individual.

However, if we expect semantics to account for the difference in cognitive value of (5) and (6) (Lois accepts one but not the other, and they play different roles in her reasoning), we must recognize a semantic difference in the contribution of the two different names.

So Frege calls our attention to two problems, (i) the problem of the the difference in truth-value of corresponding belief attributions (such as (1) and (4)), and (ii) the problem of the difference in the cognitive significance of sentences composed in the same way of elements with the same reference (such as (5) and (6)). If distinct belief attributions indicate differences in cognitive value of the sentences in their '*that*' -clauses, then these are the same issue, presumably with a single solution.

[For a more complete description of these problems, see the subsection on Frege's Puzzles in the entry on Gottlob Frege.]

Frege's theory

Frege held that correct belief attributions must indicate the way that individuals are represented by the believer (the believer's *mode of presentation* of the referent) and that our use of a referential expression within a belief context refers to a way of representing an object rather than to the expression's ordinary referent. [Frege 1892] When we use 'Superman' and 'Clark' in (1) and (4), we use them to refer to two different ways that Lois has of representing these individuals.

Frege also held that the ordinary sense of an expression --- the way that the expression indicates its referent --- becomes a part of the truth-conditions for a sentence in which the expression occurs, if that expression is used within a belief context.

On the face of it, these are two different theories about terms in belief contexts --- that they refer to the believer's way of representing the object, and that they refer to the ordinary sense. Frege unifies these by holding that the ordinary sense is a way of representing an object.

Thus he can explain why the truth-value of the sentence as a whole can vary in (1) and (4), even though the constituents of the sentences ordinarily make the same contribution to the truth-values of sentences. Belief contexts are not ordinary in that regard, so we must look to the usual sense, not the usual referent, in determining the truth-value of a belief attribution.

Difference in sense also explains the difference in cognitive value of sentences (5) and (6). Though they must have the same truth-value, because the constituents are co-referential, they have different senses, because the names ‘Clark Kent’ and ‘Superman’ have different senses, according to Frege. Though the truth-value depends on the referents of terms, the cognitive value depends on the sense attached to the terms involved in a sentence.

[For a more complete description of Frege's theory, see the subsection on Frege's Theory of Sense and Denotation in the entry on Gottlob Frege.]

Problems for the simple Fregean solution

Frege's solution fails when we try to extend it to all types of propositional clauses that can occur in belief contexts. In referring to individuals, one of our most common ways of doing it is to use demonstratives, such as ‘this’, ‘she’, ‘they’ and similar expressions that can get their reference from an accompanying demonstration. Another common way to refer is by the use of indexicals, such as ‘I’, ‘you’, and ‘now’, that get their reference from features of the context in which they are used. It is dificult to accommodate the use of such terms with a Fregean theory that requires that the mode of presentation is the semantic value of a singular term in a propositional attitude clause. (Kaplan 1977, Perry 1977.) When I point at someone and say:

> (7) Alice believes that he will solve an important problem in physics,

my use of ‘he’ does not indicate anyhing about the way in which Alice represents the individual indicated. Since Frege holds (i) that indicating the sense is indicating the way the believer represents the individual and (ii) that belief attributions indicate the sense, this appears to be a grave problem for his theory. This is perhaps even more evident if I say

> (8) Alice believes that I will solve an important problem in physics.

My use of ‘I’ tells us nothing about how Alice represents me, and it does not refer to anything that we would ordinarily call the *sense* of ‘I’.

Attributions of a common belief to many people also highlight the problems in the Fregean idea that a belief attribution must indicate the way in which the believer represents an individual in belief. If different people have different modes of representation associated with a name, then the belief attributer can't be responsible for the many modes of representation when he ascribes a common belief.

> (9) Many people believe that Tom McKay will solve an important problem in physics.

Someone attributing such a belief cannot be responsible for the many modes of representation that the various believers associate with ‘Tom McKay’. The theory could work in this case only if we could find a sense for ‘Tom McKay’ that does not vary from person to person and that can serve as the mode of presentation of that individual for each person. That seem unlikely. When we use indexicals in such attributions, the problem is perhaps clearest of all:

> (10) Many people believe that I will solve an important problem in physics.

(Similar problems are raised in Schiffer 1992, 507-508.) It seems quite implausible that the truth-conditions for my uterance involve either the many modes of representation that the believers have of me when I assert (10). Only the referent, not its mode of presentation, seems relevant with ordinary uses of indexical expressions like ‘I’, ‘you’, ‘now’, ‘yesterday’ and demonstratives like ‘that’ and ‘he’. [See Perry 1977, 1979].

It appears, then, that our fundamental responsibility in belief attributions is to get the referent right -- i.e., to indicate the individuals the beliefs are about and the properties that the believer attributes to them. The Fregean theory that requires only a reference to the mode of presentation seems wrong or irrelevant in many of our most common sorts of belief attribution. Yet focus on the referent alone seems wrong when we consider the contrast between (1) and (4).

The Russellian theory

Bertrand Russell proposed a view very different from Frege's, arguing that names contribute only their referents to the propositions that contain them. More recent Russellians, (McKay 1980, Salmon 1986 and many subsequent papers, and Soames 1989 and 1995, for example) have followed this much of Russell's view, arguing that the fundamental truth-conditions for belief attributions involve only the objects and properties, not the way those are represented. [Our formulation of the puzzle does not use definite descriptions --- noun phrases of the form *the F*. Russell's own approach to these issues involves, in part, his analysis of definite descriptions. That is explored in the linked discussion of definite descriptions .] According to these Russellians, we are wrong when we say that (1) is true and (4) is false. They are composed in the same way from elements with the same semantic value, and problems like those that we have just considered for the Fregean approach can be a part of arguments for the view that most people's ordinary judgments here are incorrect.

Many Russellians (McKay and Salmon, for example) go on to try to explain why people make those incorrect judgments about truth-values.

There may often be additional responsibilities for indicating the way in which the believer represents the individuals that the beliefs are about, but these go beyond the truth-conditions of the attributions. Any additional responsibilities are pragmatic, not semantic, requirements on the speaker. They may be conditions that a speaker must ordinarily fulfill when making a belief attribution, but they are not part of the semantic content of the attribution.

As we saw in the discussion of the problems in the Fregean approach, indexical cases show clearly that indicating the believer's mode of representation cannot be a general requirement of belief attribution. In addition, we often use names in ways that clearly do not fulfill that requirement. For example, I may say:

> (11) The students believe that John is a great teacher.

even if I know that they refer to him only as ‘Professor Adams’, and none know his first name. In using a name in the attribution, I am not required to use a name that indicates the students' way of representing John in order to say something true.

We sometimes have responsibilities that go beyond just avoiding falsehood. For example, it is true but inappropriate to say ‘John is sober today’ if John is always sober. That sentence is misleading because it suggests a contrast where there is none, though the sentence is true. Confounding pragmatic responsibilities with our responsibility to avoid falsehood can lead people astray concerning the fundamental semantics of belief attribution.

The Russellian must say that (4) is true (given that (1) is), but (4) may be very misleading, depending on the details of the context of attribution. Although Lois believes that Clark Kent is strong, she would never accept the sentence ‘Clark Kent is strong’. She does not accept that sentence, because she does not recognize that it expresses a proposition that she believes. There is a particular individual, who is known both as Clark Kent and as Superman, and she believes that he is strong. (1) and (4) both attribute that belief, but (4) does it in a misleading way, because it uses a sentence that she would not accept. According to this pragmatic explanation of why people go wrong here, the recognition that some utterances of (4) are very misleading and therefore inappropriate produces the feeling that those utterances of (4) are wrong. Though an utterance can feel wrong, we can't trust that feeling to tell us just how it is wrong. On this view, then, (4) is a true but misleading report of Lois's belief. Belief attribution may be guided by a pragmatic rule such as "Do not use a sentence that the believer would not accept, if possible". However, that cannot be a part of the semantics of belief attribution, according to the Russellian.

The Russellian disagrees with the Fregean about what is part of the truth-conditions or the content for belief attributions. This is also a disagreement about the elements of the objects of belief. However, there need not be a very substantive disagreement about what is involved in having a belief. For example, a

Russellian might agree with Frege that in order for Bernie to have a belief about an individual, say Carla, Bernie must have a way of representing Carla (a mode of presentation of Carla). Where they disagree: the Russellian holds that the truth-conditions for the belief sentence

> (12) Bernie believes that Carla is pretty.

do not include anything about Bernie's way of representing Carla, while the Fregean holds that the use of ‘Carla’ in that sentence refers only to that mode of presentation. Still, they can agree that Bernie must have some way of representing Carla for this to be right. They might even agree that a belief attributer should avoid referring to Carla in a way that would mislead us about how Bernie represents her (in the way that (4) is at least misleading about Lois's state of mind). The difference may just be in whether that last responsibility is a part of the truth-conditions or only the appropriateness conditions of the belief attribution.

Problems for Russellian theories

There are several problems and puzzles for this Russellian view, and a thoroughly worked out view will need to respond to all of them.

Pragmatic principles. It seems that the Russellian should have some account of what leads so many people astray in their judgments that (1) and (4) differ in truth-value. If the Russellian wishes to give a pragmatic account of people's ordinary judgments about the truth-value, the Russellian must clearly identify the pragmatic principles that make these incorrect judgments are so pervasive and stubborn.

Asymmetric relations. There are cases that are much less intuitive than even the claim that (4) is true. If names really are inter-substitutable, and if it is true that:

> Lois believes that Superman is stronger than Clark Kent.

then these must also be true:

> Lois believes that Superman is stronger than Superman.

> Lois believes that Clark Kent is stronger than Superman.

Can it really be that these are true, and that the typical strong feeling that they are false is really a feeling of pragmatic inappropriateness? Must Lois also believe that Superman is stronger than himself, or can we differentiate this from the previous claims? (See Salmon 1992, McKay 1991)

Negation. According to Russellians, the belief attribution:

(1) Lois believes that Superman is strong.

ordinarily conveys the meaning that Lois believes of that individual that he is strong. However, according to some, it is also a conversational implicature of that sentence that Lois would accept 'Superman is strong', and an ordinary use of the sentence conveys that too. On the other hand

(4) Lois believes that Clark Kent is strong.

is also true, but it implicates something false, and that leads to our judgment that the sentence is inappropriate or even false.

What about negation?

(3) Lois does not believe that Clark Kent is strong.

This is a false sentence, according to the Russellian, but it would ordinarily be judged true. In addition, it would ordinarily convey the idea that there is a particular way of representing the individual (as *Clark Kent*) such that Lois does not represent him as strong in association with that mode of representation. How does a use of a false sentence like this produce such judgments? The Russellian who wishes to provide a pragmatics-based account of how people's ordinary judgments go astray needs to explain this, and it is not so evident how to do that. (See Recanati, 1993, pp. 341-345)

Mixed theories

Mark Richard [1990], Crimmins and Perry [1989], Crimmins [1992], and many others have interpreted belief attributions in a way that holds us responsible for both the Russellian propositional content and for the believer's way of representing the content. The standards for getting the mode of representation right vary with context, and tend to fade to nothing in cases invoving attributions like (9) and (10). However, they do more than just hold us responsible --- they agree with Frege that a belief attributer will say something false if the singular term in the attribution does not appropriately represent the believer's way of representing the object. They differ from Frege in that they also regard the Russellian proposition as a part of the content of the belief attributed. (Some views of this kind have been called "hidden-indexical theories," because they make reference to a type of mode of representation without having any word that explicitly refers to that does just that job. See Schiffer 1992, p. 503.)

This can provide for results that many people find satisfying. For example, on this kind of view:

(13) Lois believes that Superman is strong

is true in ordinary contexts; but

(14) Lois believes that Clark Kent is strong

is false in ordinary contexts. On the other hand, pointing to Clark Kent naked in the sauna and saying

(15) Lois believes that he is strong

one may speak truly or falsely, depending on a variety of contextual factors. (Whether the audience knows about the Superman disguise, what kind of information about Lois is required in the context, etc..)

Objections of Bach [1993], Braun [1991], Rieber [1995], Schiffer [1992, 1994], Saul [1992, 1997] and Soames [1995] have raised difficulties for such views. The principal problem is how to incorporate the mode of presentation.

(A) If the mode of presentation must be referred to (as Frege required), then we get incorrect results concerning the possibility of belief attribution, because our attributions are too particular. For example, we often wish to say that distinct individuals share a belief even though they have different modes of presentation of the individuals the belief is about. Lois Lane and Lex Luthor share the belief that Superman flies, even though they have somewhat different modes of presentation for Superman. This is a problem that also arose for the pure Fregean theory.

(B) If the truth of a belief attribution requires only that the mode of presentation is of some general sort, without a reference to a particular mode of presentation, still that cannot be a part of the content of the belief attribution. When we attribute beliefs, we are not attributing acceptance of a claim that there are modes of presentation of a certain sort. In saying that Lois Lane believes that Superman flies, a speaker does not take on a commitment to the metaphysical claim that there exist modes of presentation. Belief attributions just don't seem to be making the kind of claim that such a view would require.

(C) Could the mode of presentation be a part of the truth-conditions without being a part of the content? Could it be like the situation with indexicals, where, for example it is part of the truth-conditions for an utterance of ‘You are bald’ that the addressee of the utterance is bald, but it is not a part of the content of that utterance that there is an addressee or that that there are utterances? That doesn't seem to work. If I point to Smith and say

(16) You are bald.

it is never a part of the content of the utterance that there is an addressee of my utterance, even though there must be an addressee if that sentence is to succeed in expressing the intended meaning. Other utterances, by other speakers (‘He is bald’, ‘I am bald’ and ‘Smith is bald’), could have the same content, (and they clearly do not have as part of their content the claim that there are utterances and addressees of those utterances). When I say

(1) Lois believes that Superman is strong

the analogue would be to say that it is never a part of the content of my utterance that Lois represents this individual in a certain way. That, though, is the Russellian theory, not the view we are currently considering.

Ambiguity theories

Ambiguity theories share many features of the mixed theories just described. Consideration of such views is useful, even if we find the views ultimately unsatisfactory.

First I want to consider the view that ‘believe’ is ambiguous. In one use, the verb is used to relate us to a Fregean proposition, a way of representing the world. On the other use, it relates us to a Russellian proposition or state of affairs, something that has individuals, properties, relations, etc. as contituents.

When there is an ambiguity, usually the most useful thing is to introduce two terms, one for each of the two different meanings involved. Thus, for a time, I will introduce ‘accept’ and ‘doxate’. Since it is natural to say that we ‘accept’ sentences and other sentence-like representations, let us extend this usage to say that we ‘accept’ Fregean thoughts, which include modes of presentation of individuals. By contrast, let us say that we ‘doxate’ Russellian propositions. For example, using quoted sentences to refer to Fregean thoughts:

(17) Lois Lane accepts ‘Superman is strong’.

(18) Lois does not accept ‘Superman is not strong’.

(19) Lois accepts ‘Clark Kent is not strong’.

(20) Lois does not accept ‘Clark Kent is strong’.

(21) Lois doxates the Russelian proposition that Clark Kent is strong. (Because of (1).)

(22) Lois doxates the Russellian proposition that Clark Kent is not strong. (Because of (3).)

It is natural to hold that there are two standards of consistency. One applies to sentences, according to which the sentences ‘Superman is strong’ and ‘Clark Kent is not strong’ are consistent. Nothing in the form of the sentences rules out the possibility that both are true. There is nothing about the use of negation that would create inconsistency between the quoted sentences within (17) and (19), in the way that it does for the quoted sentences within (17) and (18). The Russellian propositions that the quoted sentences in (17) and (19) express, on the other hand, are propositionally inconsistent. They cannot both be true. That cannot be seen just by inspecting the sentences that express those propostions; one must also know the co-reference relations among the terms used. Thus the sentences that Lois accepts are

consistent (by the first standard), but the Russellian propositions that she doxates in virtue of accepting those sentences (in the world we imagine for her story) are inconsistent. Because ‘Superman’ and ‘Clark Kent’ refer to the same individual, the two sentences that Lois accepts express Russellian propositions that cannot both be true.

The puzzles that arise when we consider the question of whether Lois has consistent *beliefs* can at least be addressed if we instead ask about what she accepts and about what she doxates, for use of these two terms helps to make it clear that two standards of consistency apply. One standard applies to the form of the sentences accepted, without taking account of co-reference relations among terms, and the other considers the content of terms and applies to Russellian objects of doxation. The sentences she accepts are consistent by the first standard -- they are the kinds of sentences that could be used to express propositions that are both true, and Lois has no reason to believe that that is not the case in the current situation. She is wrong, however, because these sentences express propositions that cannot both be true. By identifying an ambiguity, we provide resources for an answer to the vexing puzzles.

Some philosophers have connected the *de dicto -de re* distinction with such an ambiguity in ‘believe’. That view, though, has many difficulties, and the *de dicto -de re* terminology is used in other ways, to mark distinctions among beliefs or distinctions among belief attributions. Rather than trying to sort out different interpretations of that terminology, though, let's stick with the invented disambiguating verbs ‘accept’ and ‘doxate’, briefly, for heuristic purposes. [Interested readers may with to pursue a more in-depth discussion of the de re/de dicto distinction.]

The ambiguity theory of belief attribution has a certain appeal. In ordinary cases of belief attribution, the sentence attributing the belief can do both things at once. It can indicate the Russellian proposition that is the object of doxation, and it can indicate the sentence (or sentence-like representation) that is accepted by the believer. (When I say "Alice believes that George Eliot was a great novelist", I indicate the way that Alice represents the novelist and the individual her belief is about.) Those are paradigm cases of the use of ‘believe’. Treating ‘believe’ as univocal, however, would leave us with a problem in cases like Lois's, where we need to say different things about acceptance and doxation. (For example, regarding what she accepts, she is epistemically in the clear. Her view meets the appropriate standard of consistency. However, the propositions that she doxates are inconsistent. And she rejects certain sentences that express propositions that she doxates (such as (2) and (4)).)

The situation with ‘believe’, on this view, is something like the situation with ‘is heavier than’ as ordinarily used. That expression could be talking about greater weight or greater mass. Ordinarily it doesn't matter; we can express both propositions at the same time, since for the most part their truth-conditions vary together. (We can have *polysemy*, where the two meanings are simultaneously expressed, not just ambiguity.) But if we are comparing objects that are on different planets, or even at very different elevations on this planet, then we will have to make it clear whether we are concerned with mass or weight. Similarly, we can often indicate what is accepted and what is doxated, but in some situations we cannot do both, and so we must make it clear which we are concerned with. We can make it clear by employing the language of acceptance and doxation instead of the language of belief.

This first ambiguity view, then, just claims that 'believe' is ambiguous (in the way that 'is heavier than' is ambiguous) and that the ambiguity can be resolved by recognizing the two different things that are ordinarily captured in successful belief attributions. The strategy suggested here is the ordinary one for dealing with an ambiguity; we find or invent two terms and stipulate which will go with each of the meanings. Thus we might firmly stipulate that 'is heavier than' will be associated with weight comparisons and introduce 'has more mass than' for comparisons of mass.

Ultimately such a strategy will not work for 'believe' however, because the situation is more complicated. Some belief attributions involve sentences that use one singular term in a purely referential way but use other terms in a way that seems to require getting the mode of representation right. For example, it could bethat Lois spots a man walking in the corridor, and makes a height judgment that leads her to say two things:

> (23) He is taller than Superman.

> (24) He is not taller than Clark Kent.

I might recognize the man in question as Rudy Sanchez, someone known to me and the people I am speaking to, but unknown to Lois (outside of the brief sighting in the corridor). I can then make these attributions:

> (25) Lois believes that Rudy Sanchez is taller than Superman.

> (26) Lois believes that Rudy Sanchez is not taller than Clark Kent.

Is this acceptance or doxation? It seems that it must be both, because the name 'Rudy Sanchez' is not being used to indicate Lois's mode of representation of the individual, but the names 'Superman' and 'Clark Kent' are (according to a theorist who might find this ambiguity theory attractive). With the following pair, where Rudy Sanchez is the speaker, the situation is perhaps even clearer.

> (27) Lois believes that I am taller than Superman.

> (28) Lois believes that I am not taller than Clark Kent.

This cannot be an ambiguity located in the verb 'believe' after all. Our intuitions support the idea that there are two different kinds of uses of singular terms (*de re* and *de dicto*) in belief attributions, but that will not divide belief attributions into the *de re* and *de dicto*, because names being used in different ways can occur in a single attribution.

Edward N. Zalta has suggested a different approach that involves ambiguity in belief attributions. [Zalta 1983]

Non-compositional accounts

Our ability to understand the sentences of a language is open-ended. Our semantic theory must provide for that fact, and this is the basis for holding firm to the idea of semantic compositionality. We can understand an infinity of novel sentences, based on our understanding of words and how they are put together. Any account that breaches the principle of compositionality must provide an alternative account of how this understanding is possible.

Mark Crimmins [Crimmins 1998] has recently suggested an approach that gives up on compositionality, in at least a limited way. The truth conditions of complex expressions are not determined simply by the meanings of their parts and how they are composed. Instead, in some cases, the truth-value of a sentence is determined by the truth-value it would have in a certain fictional situation. This is based on a theory of fiction that says, for example, that

> Santa Claus wears a red suit

is true as we usually use it, because it is true in a certain fictional situation that serves as the background for the utterance.

Applied to belief attributions, we imagine a certain pretense that goes with sentences like these:

> (1) Lois believes that Superman is strong.
>
> (2) Lois believes that Clark Kent is not strong.

The pretense is that there are people, Superman and Clark Kent, corresponding to Lois's modes of presentation, and associated with our use of ‘Superman’ and ‘Clark Kent’. The pretense makes the differing attributions possible, even though ‘Superman’ and ‘Clark Kent’ both refer to the same person. At the same time, in order for Lois's first belief to be true and the second false, we must keep the terms firmly attached to that person as the object of his belief. Thus, this theory gives us the dual interpretation of names required in the mixed theories of belief attribution discussed earlier. It supplements the theory by providing a specific, non-compositional account of the truth-conditions for such sentences. Although this account is non-compositional, it is clear how we can understand new belief attributions, by applying our ordinary compositional semantics to interpret sentences with respect to non-actual situations. (Note: in using the Superman example throughout our discussion, we have already taken advantage of our ability to do engage in some pretense --- the pretense that Clark and Lois exist. Perhaps this is similar.)

This account provides no evident way to avoid problems like those that come up for the mixed theories, however. If we are engaging in pretenses that correspond to particular people's modes of presentation when we use names, then we cannot take agreement to be sharing a belief. Different people may share a belief even though they represent the object of belief in different ways. If many people know that ‘Clark

Kent' and 'Superman' are co-referential and many don't know that, then we cannot attribute common beliefs, like "Almost everyone believes that Superman is the most important citizen of Metropolis." [See Crimmins 1998]

William Taschek has also developed a theory that he says is non-compositional, in another recent attempt to deal with the Fregean puzzles. [Taschek 1998]

Concluding note

Ultimately, a response to Frege's puzzle must fit with a broader theory of human interaction. People communicate their beliefs, they agree when they share a belief, and beliefs play a role in motivating and explaining action. Any account of belief attributions must say how these attributions can accurately reflect these many roles.

Bibliography

- Almog, J., Perry, J., and Wettstein, H. (eds.). 1989. *Themes from Kaplan,* Oxford: Oxford University Press.
- Braun, D. 1991. "Proper Names, Cognitive Contents, and Beliefs." *Philosophical Studies* 62, 289-305.
- Boër, S. 1995. "Propositional Attitudes and Compositional Semantics," *Philosophical Perspectives* 9, 341-380.
- Boër, S. 1994. "Propositional Attitudes and Formal Ontology," *Synthese* 98, 187-242.
- Boër, S. and W. Lycan. 1980. "Who Me?" *Philosophical Review* 89, 427-466.
- Burge, T. 1977. "De Re Belief." *Journal of Philosophy* 74, 338-362.
- Crimmins, M. 1992. *Talk about Beliefs*, Cambridge, Massachusetts: MIT Press.
- Crimmins, M. 1998. "Hesperus and Phosphorus: Sense, Pretense, and Reference," *The Philosophical Review*, 107/1 (January), 1-47.
- Crimmins, M. and J. Perry. 1989. "The Prince and the Phone Booth: Reporting Puzzling Beliefs," *Journal of Philosophy* 86, 685-711.
- Davidson, D., and Harman, G., 1975. *The Logic of Grammar*, Encino: Dickenson Publishing.
- Frege, G. 1892. "On Sense and Reference." Translation of "Über Sinn und Bedeutung" in P. Geach and M. Black, *Translations from the Writings of Gottlob Frege*. Blackwell, 1952, 56-78.
- Frege, G. 1918. "Thoughts." In *Collected Papers on Mathematics, Logic, and Philosophy,* Oxford: Basil Blackwell, 1984, B. McGuinness (ed.), P. Geach and R. H. Stoothoff (trans.). Also in Salmon and Soames [1988] 33-55.
- Frege, G. 1904. Correspondence with Russell, in Salmon and Soames, [1988], p.56.
- Kaplan, D. 1972, "What is Russell's Theory of Descriptions?" In Davidson and Harman's *The Logic of Grammar* 1975, and in D. F. Pears 1972.
- Kaplan, D. 1977. "Demonstratives." In J. Almog, *et al*. [1989].
- Kaplan, D. 1969. "Quantifying in," in *Words and Objections: Essays on the Work of W. V. Quine,*

D. Davidson and J. Hintikka (eds.), Dordrecht: Reidel, 178-214. Reprinted in Linsky [1971].

- Kazmi, A. 1987. "Quantification and Opacity." *Linguistics and Philosophy* 10, 77-100.
- Kripke, S. 1979. "A Puzzle about Belief." In *Meaning and Use,* A. Margalit (ed.), Dordrecht: Reidel, 239-283.
- Lewis, D. 1981. "What Puzzling Pierre Believes." *Australasian Journal of Philosophy* 59, 283-289.
- Linsky, L. 1971, *Reference and Modality*, London: Oxford University Press.
- McKay, T. 1981. "On Proper Names in Belief Ascriptions." *Philosophical Studies* 39, 287-303.
- McKay, T. 1991. "Representing *De Re* Beliefs". *Linguistics and Philosophy* 14, 711-739.
- Pears, D., 1972, *Bertrand Russell*, Garden City, NY: Doubleday Anchor, 227-244.
- Perry, J. 1977. "Frege on Demonstratives." *Philosphical Review* 86, 474-97.
- Perry, J. 1980. "Belief and Acceptance." In *Midwest Studies in Philosophy*, Vol. 5, P. French, T. Uehling and H Wettstein (eds.), Minneapolis: University of Minnesota Press, 533-542.
- Perry, J. 1979. "The Problem of the Essential Indexical." In Salmon and Soames [1988] and in Perry [1993].
- Perry, J. 1993 *The Problem of the Essential Indexical*, (Oxford: Oxford University Press, 1993).
- Quine, W. 1961, "Reference and Modality," in *From a Logical Point of View*, New York: Harper and Row. Reprinted in Linsky [1971].
- Quine, W. 1956 "Quantifiers and Propositional Attitudes," *Journal of Philosophy* 53, 177-187. Also in *The Ways of Paradox*, New York: Random House, 1966, and in Linsky [1971].
- Recanati, F. 1993. *Direct Reference: from Language to Thought*. Oxford: Blackwell
- Richard, M. 1983. "Direct Reference and Ascriptions of Belief." In Salmon & Soames. From *Journal of Philosophical Logic* 12, 425-452.
- Richard, M. 1990. *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them*. Cambridge University Press.
- Russell, B. 1905, "On Denoting." In Davidson and Harman 1975.
- Russell, B. 1912. "Knowledge by Acquaintance and Knowledge by Description," in *The Problems of Philosophy*, Oxford: Oxford University Press, 1979. Reprinted in Salmon and Soames [1988].
- Russell, B. 1904. Correspondence with Frege, in Salmon and Soames, [1988], p. 57.
- Salmon, N. 1986. *Frege's Puzzle*. (Cambridge, Massachusetts: MIT Press).
- Salmon, N. 1992 "Reflections on Reflexivity," *Linguistics and Philosophy* 15, 53-63
- Salmon, N. and S. Soames. 1988. *Propositions and Attitudes*. Oxford: Oxford University Press.
- Saul, J. 1997. "Substitution and Simple Sentences". *Analysis* 57, 102-108.
- Schiffer, S. 1992. "Belief Ascription." *Journal of Philosophy* 89, 499-521.
- Schiffer, S. 1987. "The ‘Fido’-Fido theory of belief," *Philosophical Perspectives I*, Atascadero: Ridgeview
- Schiffer, S. 1990 "The Mode-of-Presentation Problem," in *Propositional Attitudes*, C. Anderson and J. Owens (eds.), Stanford: CSLI
- Schiffer, S. 1994. "A Paradox of Meaning." *Noûs* 28, 279-324.
- Soames, S. 1989. "Direct Reference, Propositional Attitudes and Semantic Content." *Philosophical Topics* 15, 44-87.
- Soames, S. 1995. "Beyond Singular Propositions," *Canadian Journal of Philosophy* 25, 515-549.
- Taschek, W. 1998. "On Ascribing Beliefs: Content in Context," *Journal of Philosophy*, 95, 323-

353.

- Zalta, E. 1983. *Abstract Objects: An Introduction to Axiomatic Metaphysics*, Dordrecht: D. Reidel.

Other Internet Resources

- A Puzzle about Belief Reports, a paper by Kent Bach
- Reflexivity, Indexicality, and Names, a paper by John Perry
- Fregean Senses, Modes of Presentation, and Concepts, a paper by Edward N. Zalta

Related Entries

descriptions | Frege, Gottlob | indexicals | intentionality | propositions: singular | propositions: structured | Quine, Willard van Orman | reference | Russell, Bertrand

Thomas J. McKay
tjmckay@syr.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 16, 2000
Content last modified: February 16, 2000

Chronological Catalog of Frege's Work
Supplementary Document to 'Gottlob Frege'
Stanford Encyclopedia of Philosophy
http://plato.stanford.edu/entries/frege/

Edward N. Zalta
zalta@stanford.edu

Content Last Modified: September 15, 2000

This Catalog was compiled initially by cross-checking the Bibliographies in the following works: Bynum [1972], Beaney [1997], Hermes [1969], and Angelelli [1967]. Sources were checked whenever possible and inconsistencies and omissions were rectified. The publication history of each English translation was developed. Please notify the author if you find any errors or missing items. (No citations to Frege's letters are included. Frege's correspondence may be found in Gabriel *et al.* [1976] and Janik [1989], many of which were translated in Gabriel *et al.* [1980], Kluge [1971], and Woodward [1967].)

Contents

Chronological Catalog of Frege's Work

• [**1873**] *Über eine geometrische Darstellung der imaginären Gebilde in der Ebene*, Inaugural-Dissertation der Philosophischen Fakultät zu Göttingen zur Erlangung der Doktorwürde, Jena: A. Neuenhann, 1873; reprinted in Angelelli [1967] (pp. 1-49)

> *On a Geometrical Representation of the Imaginary Forms in the Plane*, Doctoral Dissertation, School of Philosophy, U. Göttingen; translation by H. Kaal in McGuinness [1984] (pp. 1-55)

• [**1874a**] *Rechnungsmethoden, die sich auf eine Erweiterung des Grössenbegriffes gründen*, Dissertation zur Erlangung der Venia Docendi bei der Philosophischen Fakultät in Jena, Jena: Friedrich Frommann, 1874; reprinted in Angelelli [1967] (pp. 50-84)

> *Methods of Calculation based on an Extension of the Concept of Quantity*, Dissertation for the Venia Docendi (Habilitationsschrift) in the School of Philosophy of Jena; translation by H. Kaal in McGuinness [1984] (pp. 56-92)

• [**1874b**] 'Rezension von: H. Seeger, *Die Elemente der Arithmetik*', *Jenaer Literaturzeitung* **1**/46 (1874): 722; reprinted in Angelelli [1967] (pp. 85-86)

> 'Review of H. Seeger, *The Elements of Arithmetic*'; translation by H. Kaal in McGuinness [1984] (pp. 93-4)

• [**1877a**] 'Rezension von: A. v. Gall and E. Winter, *Die analytische Geometrie des Punktes und der Geraden und ihre Anwendung auf Aufgaben*', *Jenaer Literaturzeitung* **4**/9 (1877): 133-134; reprinted in Angelelli [1967] (pp. 87-88)

> Review of A. v. Gall and E. Winter, *Analytic Geometry of the Point and the Line and its Application to Problems*; translation by H. Kaal in McGuinness [1984] (pp. 95-97)

• [**1877b**] 'Review of J. Thomae, *Sammlung von Formeln welche bei Anwendung der elliptischen und Rosenhainschen Funktionen gebraucht werden*', *Jenaer Literaturzeitung* **4**/30 (1877), p. 472; reprinted in Angelelli [1967] (p. 89)

> 'Review of J. Thomae, *Collection of Formulae used in the Application of Elliptical and Rosenhain Functions*'; translation by H. Kaal in McGuinness [1984] (p. 98)

• [**1878**] 'Über eine Weise, die Gestalt eines Dreiecks als komplexe Grösse aufzufassen', Vortrag, gehalten in in der Sitzung vom 8. Februar 1878, der Jenaischen Gesellschaft für Medizin und Naturwissenschaft, in *Jenaische Zeitschrift für Naturwissenschft* **12**/Supplement (1878) (= Sitzungsberichte der Jenaischen Gesellschaft für Medizin und Naturwissenschaft für das Jahr 1878), p. xviii; reprinted in Angelelli [1967] (pp. 90-91)

'On a Way of Conceiving the Shape of a Triangle as a Complex Quantity' (Lecture at the February 8, 1878 meeting of Jena's Society for Medicine and Natural Science); translation by H. Kaal in McGuinness [1984] (pp. 99-100)

• [**1879**] *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*, Halle a. S.: Louis Nebert, 1879; reprinted in Angelelli [1964] (pp. ix-xvi, 1-88)

Concept Script, a formal language of pure thought modelled upon that of arithmetic;

- Complete translation by S. Bauer-Mengelberg in van Heijenoort [1967] (pp. 1-82)
- Complete translation by T. Bynum in Bynum [1972] (pp. 101-203)
- Partial translation (§§1-12) by P. Geach in Geach and Black [1980] (pp. 1-20)
- Partial translation (§§1-12) by M. Beaney in Beaney [1997] (pp. 47-78)

• [**1879b**] 'Anwendungen der Begriffsschrift', Vortrag, gehalten in der Sitzung vom 24. Januar 1879 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft, in *Jenaische Zeitschrift für Naturwissenschaft* **13**/Supplement II (1879) (= Sitzungsberichte der Jenaischen Gesellschaft für Medizin und Naturwissenschaft für das Jahr 1879), pp. 29-33; reprinted in Angelelli [1964] (pp. 89-93)

'Applications of the Concept Script' (Lecture at the January 24, 1879 meeting of Jena's Society for Medicine and Natural Science); translation by T. Bynum in Bynum [1972] (pp. 204-8)

• [1879?] (Logik), originally unpublished; recently published in Hermes *et al.* [1969] (pp. 1-8)

(Logic), (originally unpublished), translation by Long and White in Hermes *et al.* [1979] (pp. 1-9)

• [**1880**] 'Rezension von: Hoppe, *Lehrbuch der analytischen Geometrie I*', *Deutsche Literaturzeitung*, **1** (1880): columns 210-211; reprinted in Angelelli [1967] (pp. 92-93)

'Review of Hoppe, *Textbook of Analytic Geometry I*'; translation by H. Kaal in McGuinness [1984] (pp. 101-102)

• [1880?] 'Booles rechnende Logik und die Begriffsschrift', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 9-52)

'Boole's Logical Calculus and the Concept Script', (originally unpublished), translation by Long and White in Hermes *et al.* [1979] (pp. 9-46)

• [**1881**] ‘Über den Briefwechsel Leibnizens und Huygens mit Papin’, Vortrag, gehalten in der Sitzung vom 15. Juli 1881 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft, in *Jenaische Zeitschrift für Naturwissenschaft*, **15**/Supplement (1881-1882) (= Sitzungsberichte der Jenaischen Gesellschaft für Medizin und Naturwissenschaft für das Jahr 1881), pp. 29-32; reprinted in Angelelli [1964] (pp. 93-96)

> ‘On Leibniz’s and Huygens’ Letter Exchange with Papin’ (Lecture at the July 15, 1881 meeting of Jena’s Society for Medicine and Natural Science); no English translation available.

• [**1882a**] ‘Über die wissenschaftliche Berechtigung einer Begriffsschrift’, *Zeitschrift für Philosophie und philosophische Kritik*, **81** (1882), pp. 48-56; reprinted in Angelelli [1964] (pp. 106-114) and in Patzig [1962] (pp. 91-97)

> ‘On the Scientific Justification of a Concept Script’
>
> - Translation by J. Bartlett in Bartlett [1964]
> - Translation by T. Bynum in Bynum [1972] (pp. 83-89)

• [**1882b**] ‘Über den Zweck der Begriffsschrift’, Vortrag, gehalten in der Sitzung vom 27. Januar 1882 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft, *Jenaische Zeitschrift für Naturwissenschaft*, **16**/Supplement (1882/1883) (= Sitzungsberichte der Jenaischen Gesellschaft für Medizin und Naturwissenschaft für das Jahr 1882), pp. 1-10 reprinted in Angelelli [1964] (pp. 97-106)

> ‘On the Purpose of Concept Script (Lecture at the January 27, 1882 meeting of Jena’s Society for Medicine and Natural Science)
>
> - Translation by V. Dudman in Dudman [1968]
> - Translation by T. Bynum in Bynum [1972] (pp. 90-100)

• [1882c] ‘Booles logische Formelsprache und meine Begriffsschrift’, originally unpublished; recently published in Hermes *et al.* [1969] (pp. 53-59)

> ‘Boole’s Logical Formula Language and my Concept Script’, (originally unpublished), translation by Long and White in Hermes *et al.* [1979] (pp. 47-52)

• [1882d?] (17 Kernsätze zur Logik), originally unpublished; recently published in Hermes *et al.* [1969] (pp. 189-190) [Note: This piece was thought to have been written in 1876, but research by G. Gabriel suggests the date of 1882.]

> (17 Key Sentences on Logic), (originally unpublished), translation by Long and White in Hermes *et al.* [1979] (pp. 174-175)

• [**1883**] ‘Geometrie der Punktpaare in der Ebene’, Vortrag, gehalten in der Sitzung vom 2. November 1883 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft, in *Jenaische Zeitschrift für Naturwissenschaft*, **17**/Supplement (1884) (= Sitzungberichte der Jenaischen Gesellschaft für Medizin und Naturwissenschaft für das Jahr 1883), pp. 98-102; reprinted in Angelelli [1967] (pp. 94-98)

‘Geometry of Pairs of Points in the Plane’ (Lecture at the November 2, 1883 meeting of Jena’s Society for Medicine and Natural Science); translation by H. Kaal in McGuinness [1984] pp. 103-107

• [1883?] (Dialog mit Pünjer über Existenz), originally unpublished; recently published in Hermes *et al.* [1969] (pp. 60-75)

(Dialogue with Pünjer on Existence), (originally unpublished), translation by Long and White in Hermes *et al.* [1979] (pp. 53-67)

• [**1884**] *Die Grundlagen der Arithmetik: eine logisch-mathematische Untersuchung über den Begriff der Zahl*, Breslau: W. Koebner, 1884; reprinted Breslau: M. & H. Marcus, 1934; reprinted Darmstadt: Wissenschaftliche Buchgesellschaft and Hildesheim: Olms, 1961; reprinted in Thiel [1986]

The Foundations of Arithmetic: A logico-mathematical enquiry into the concept of number

- Complete translation by J. L. Austin in Austin [1974] (complete) and (pp. v-viii and §§21-27) in Kennick and Lazerowitz [1966] (pp. 67-74)
- Partial translation (§§55-91, §§106-109) by M. Mahoney in Benacerraf and Putnam [1983] (pp. 130-159)/[1964] (pp. 85-112)
- Partial translation (Introduction, §§1-4, §§45-69, §§87-91, §§104-9) by M. Beaney in Beaney [1997] (pp. 84-129)

• [**1885a**] ‘Rezension von: H. Cohen, *Das Prinzip der Infinitesimal-Methode und seine Geschichte*’, *Zeitschrift für Philosophie und philosophische Kritik* **87** (1885), pp. 324-329; reprinted in Angelelli [1967] (pp. 99-102)

‘Review of H. Cohen, *The Principle of the Method of Infinitesimals and its History*’; translation by H. Kaal in McGuinness [1984] (pp. 108-111)

• [**1885b**] ‘Über formale Theorien der Arithmetik’, Vortrag, gehalten in der Sitzung vom 17. July 1885 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft, in *Jenaische Zeitschrift für Naturwissenschaft*, **19**/Supplement II (1886) (= Sitzungberichte der Jenaischen Gesellschaft für Medizin und Naturwissenschaft für das Jahr 1885), pp. 94-104; reprinted in Angelelli [1967] (pp. 103-111)

‘On Formal Theories of Arithmetic’ (Lecture at the July 17, 1885 meeting of Jena’s Society for Medicine and Natural Science); translation by E.-H. W. Kluge in Kluge [1971] (pp. 141-153) and in McGuinness [1984] (pp. 112-121)

• [**1885c**] ‘Erwiderung’ (auf Cantors Rezension der *Grundlagen der Arithmetik*), *Deutsche Literaturzeitung*, **6**/28 (1885), column 1030; reprinted in Angelelli [1967] (p. 112)

'Reply', translation by H. Kaal in McGuinness [1984] (p. 122)

[Note: This is a brief reply to Cantor, G., 'Rezension der *Grundlagen der Arithmetik*', which appeared in *Deutsche Literaturzeitung*, **6**/20 (1885), columns 728-729.]

• [1890-2] (Entwurf zu einer Besprechung von Cantors Gesammelten Abhandlungen zur Lehre vom Transfiniten), originally unpublished; recently published in Hermes *et al.* [1969] (pp. 76-80)

(Draft of a review of Cantor's book *Contributions to the Theory of the Transfinite: Collected Articles from Zeitschrift für Philosophie und philosophische Kritik*), (originally unpublished), translation by Long and White in Hermes *et al.* [1979] (pp. 68-71)

• [**1891a**] 'Funktion und Begriff', Vortrag, gehalten in der Sitzung vom 9. Januar 1891 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft, Jena: Hermann Pohle, 1891; reprinted in Angelelli [1967] (pp. 125-142) and in Patzig [1962] (pp. 17-39)

'Function and Concept' (Lecture at the January 9, 1891 meeting of Jena's Society for Medicine and Natural Science); translation by P. Geach in Geach and Black [1980] (pp. 21-41), McGuinness [1984] (pp. 137-156), and (with minor revisions) in Beaney [1997] (pp. 130-148)

• [**1891b**] 'Über das Trägheitsgesetz', *Zeitschrift für Philosophie und philosophische Kritik*, **98** (1891): 145-161; reprinted in Angelelli [1967] (pp. 113-124)

'On the Law of Inertia'

- Translation by H. Kaal in McGuinness [1984] (pp. 123-136)
- Translation by R. Rand in Rand [1961]

• [1891?a] 'Über den Begriff der Zahl: 1. Eine kritische Auseinandersetzung mit Biermann', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 81-95)

'On the Concept of Number: 1. A Criticism of Biermann', (originally unpublished), translation by Long and White in Hermes *et al.* [1979] (pp. 72-86)

• [1891?b] 'Über den Begriff der Zahl: 2. Eine kritische Auseinandersetzung mit Kerry', originally unpublished draft of Frege [1892b]; recently published in Hermes *et al.* [1969] (pp. 96-127)

'On the Concept of Number: 2. A Criticism of Kerry', originally unpublished draft of Frege [1892b]; translation by Long and White in Hermes *et al.* [1979] (pp. 87-117)

• [**1892a**] 'Über Sinn und Bedeutung', in *Zeitschrift für Philosophie und philosophische Kritik*, **100** (1892), pp. 25-50; reprinted in Angelelli [1967] (pp. 143-162) and in Patzig [1962] (pp. 40-65)

'On Sense and Meaning'

- Translation by M. Black in Black [1948], Geach and Black [1980] (pp. 56-78), McGuinness [1984] (pp. 157-177), Martinich [1985] (pp. 200-212), and (with minor revisions) Beaney [1997] (pp. 151-171)
- Translation by H. Feigl in Feigl and Sellars [1949] (pp. 85-102), in Nagel and Brandt [1965] (pp. 69-78), and in Martinich [1996] (pp. 186-198)

• [**1892b**] 'Über Begriff und Gegenstand', in *Vierteljahresschrift für wissenschaftliche Philosophie*, **16** (1892): 192-205; reprinted in Angelelli [1967] (pp. 167-178) and in Patzig [1962] (pp. 66-80)

'On Concept and Object', translation by P. Geach in Geach [1951], Geach and Black [1980] (pp. 42-55), Hermes *et al.* [1979] (pp. 87-117), McGuinness [1984] (pp. 182-194), and (with minor revisions) in Beaney [1997] (pp. 181-193)

• [**1892c**] 'Rezension von: G. Cantor, *Zur Lehre vom Transfiniten: Gesammelte Abhandlungen aus der Zeitschrift für Philosophie und philosophische Kritik*', **100** (1892), pp. 269-272; reprinted in Angelelli [1967] (pp. 163-166)

'Review of Georg Cantor, *Contributions to the Theory of the Transfinite: Collected Articles from Zeitschrift für Philosophie und philosophische Kritik*', translation by H. Kaal in McGuinness [1984] (pp. 178-181)

• [1892-5] (Ausführungen über Sinn und Bedeutung), originally unpublished; recently published in Hermes *et al.* [1969] (pp. 128-136)

(Comments on Sense and Meaning), (originally unpublished), translation by Long and White in Hermes *et al.* [1979] (pp. 118-125)

• [**1893**] *Grundgesetze der Arithmetik*, Jena: Verlag Hermann Pohle, Band I; reprinted Hildesheim: Olms, 1962; reprinted in Thiel [1998]

The Fundamental Laws of Arithmetic: I

- Partial translation (Preface, Introduction, §§1-52) by M. Furth in Furth [1964]
- Partial translation (almost all of the Preface, Introduction, §§1-7) by Stachelroth and Jourdain in Stachelroth and Jourdain [1915-1917], and (Preface abridged further) in Geach and Black [1980] (pp. 137-159), and (only small part of the Preface) in Copi and Gould [1964] (pp. 205-210)

- Partial translation (most of the Preface, Introduction, §§1-6, parts of §§26-33) by M. Beaney in Beaney [1997] (pp. 194-223)

• [**1894**] 'Rezension von: E. Husserl, *Philosophie der Arithmetik I*', in *Zeitschrift für Philosophie und philosophische Kritik*, **103** (1894), pp. 313-332; reprinted in Angelelli [1967] (pp. 179-192)

'Review of E. Husserl's *Philosophy of Arithmetic I*'

- Translation by H. Kaal in McGuinness [1984] (pp. 195-209) (complete) and in Beaney [1997] (pp. 224-226) (partial)
- Partial translation by P. Geach in Geach and Black [1980] (pp. 79-85)

• [**1895a**] 'Kritische Beleuchtung einiger Punkte in E. Schröders *Vorlesungen über die Algebra der Logik*, in *Archiv für systematische Philosophie*, **I** (1895): 433-456; reprinted in Angelelli [1967] (pp. 193-210) and in Patzig [1966] (pp. 92-112)

'A Critical Elucidation of Some Points in E. Schröder, *Lectures on the Algebra of Logic*', translation by P. Geach in Geach and Black [1980] (pp. 86-106) and in McGuinness [1984] (pp. 210-228)

• [**1895b**] 'Le Nombre Entier', *Revue de Métaphysique et de Morale* **3** (1895) pp. 73-78; reprinted in Angelelli [1967] (pp. 211-219)

'Whole Numbers', (the original is in French); translation by V. H. Dudman in Dudman [1970], and in McGuinness [1984] (pp. 229-233)

• [**1896a**] 'Lettera del Sig. G. Frege all'Editore', *Rivista di Matematica*, **6** (1896-9): 53-59; reprinted in Angelelli [1967] (pp. 234-239) and in Gabriel *et al.* [1976] (pp. 181-186)

'Letter to the Editor' (original in German); translation by H. Kaal in Gabriel *et al.* [1980] (pp. 112-118). [Note: The Editor of the *Rivista di Matematica* was G. Peano.]

• [**1897**] 'Über die Begriffsschrift des Herrn Peano und meine eigene', Vortrag, gehalten in der ausserordenlichen Sitzung vom 6. Juli 1896, *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig: Mathematisch-physische Klasse*, **48** (1897): 361-378; reprinted in Angelelli [1967] (pp. 220-233)

'On Mr. Peano's Concept Script and My Own', Lecture at the special July 6, 1896 meeting of Leipzig's Science Society; translation by V. H. Dudman in Dudman [1969] and in McGuinness [1984] (pp. 234-248)

• [1897] 'Logik', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 137-163)

'Logic', (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 126-151) and (the first two sections reappear with minor revisions) in Beaney [1997] (pp. 227-250)

• [1897?] 'Begründung meiner strengeren Grundsätze des Definierens', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 164-170)

'An Argument for my Stricter Canons of Definition', originally unpublished; translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 152-156)

• [1898?] 'Logische Mängel in der Mathematik', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 171-181)

'Logical Defects in Mathematics', (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 157-166)

• [**1899**] 'Über die Zahlen des Herrn H. Schubert', Jena: Herman Pohle, 1899; reprinted in Angelelli [1967] (pp. 240-261) and in Patzig [1966] (pp. 113-138)

'On Mr. Schubert's Numbers'; translation by H. Kraal in McGuinness [1984] (pp. 249-272)

• [1899?] 'Über Euklidische Geometrie', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 182-184)

'On Euclidean Geometry', (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 167-169)

• [**1903a**] *Grundgesetze der Arithmetik*, Jena: Verlag Hermann Pohle, Band II; reprinted Hildesheim: Olms, 1962; reprinted in Thiel [1998]

The Fundamental Laws of Arithmetic: II

- Partial translation (§§56-67, §§139-144, §§146-147, and Appendix) by P. Geach in Geach and Black [1980] (pp. 159-172, 173-179, 179-181, 234-244) and in Runes [1962] (329-342) [Note: the Geach translation in Runes covers only §§56-67 ('Definitions') and the source is incorrectly labeled as Frege [1884].]
- Partial translation (§§86-137) by M. Black in Black [1950]; reprinted in Geach and Black [1980] (pp. 182-233)
- Partial translation (§§55-67, §§138-147, and Appendix) by P. Geach and M. Beaney in Beaney [1997] (pp. 258-289)
- Partial translation (Appendix only) by M. Furth in Furth [1964] (pp. 127-143)

• [**1903b**] 'Über die Grundlagen der Geometrie' (First Series), *Jahresbericht der Deutschen Mathematiker-Vereinigung* **12** (1903), pp. 319-324 (Part I), pp. 368-375 (Part II); reprinted in Angelelli [1967] (pp. 262-272)

> 'On the Foundations of Geometry' (First Series),
>
> - Translation by M. Szabo in Szabo [1960] and in Klemke [1968] (pp. 559-575)
> - Translation by E.-H. W. Kluge in Kluge [1971] (pp. 22-37) and in McGuinness [1984] (pp. 273-284)
>
> [Note: In this article, Frege criticizes Hilbert's understanding and use of the axiomatic method. There is a reply in defense of Hilbert by A. Korselt in *Jahresbericht der Deutschen Mathematiker-Vereinigung* **12** (1903), pp. 402-407. It has been translated by E.-H. W. Kluge in Kluge [1971] (pp. 38-48).]

• [1903?] (Notizen Freges zu Hilberts *Grundlagen der Geometrie*), originally unpublished; recently published in Hermes *et al.* [1969] (pp. 185-188)

> (Frege's Notes on Hilbert's *Foundations of Geometry*), (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 170-173)

• [**1904**] 'Was ist eine Funktion?', in *Festschrift Ludwig Boltzmann gewidmet zum sechzigsten Geburtstage, 20. Februar 1904*, S. Meyer (ed.), Leipzig: Barth, 1904, pp. 656-666; reprinted in Angelelli [1967] (pp. 273-280) and in Patzig [1962] (pp. 81-90)

> 'What is a Function?', translation by P. Geach in Geach and Black [1980] (pp. 107-116), and McGuinness [1984] (pp. 285-292)

• [**1906a**] 'Über die Grundlagen der Geometrie' (Second Series), *Jahresbericht der Deutschen Mathematiker-Vereinigung* **15** (1906), pp. 293-309 (Part I), 377-403 (Part II), 423-430 (Part III); reprinted in Angelelli [1967] (pp. 281-323)

> 'On the Foundations of Geometry' (Second Series), translation by E.-H. W. Kluge in Kluge [1971] (pp. 49-112) and in McGuinness [1984] (pp. 293-340); [Note: In this article, Frege replies to A. Korselt's reply to Frege [1903b].]

• [**1906b**] 'Antwort auf die Ferienplauderei des Herrn Thomae', *Jahresbericht der Deutschen Mathematiker-Vereinigung* **15** (1906), pp. 586-590; reprinted in Angelelli [1967] (pp. 324-328)

> 'Reply to Mr. Thomae's Holiday Causerie', translation by E.-H. W. Kluge in in Kluge [1971] (pp. 121-127), and in McGuinness [1984] (pp. 341-345) [Note: The 'Holiday Chat' which occasioned this piece by Frege is also translated in Kluge [1971].]

• [1906c] 'Über Schoenflies: Die logischen Paradoxien der Mengenlehre', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 191-199)

> 'On Schoenflies, *The Logical Paradoxes of Set Theory*', (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 176-183)

• [1906d] 'Was kann ich als Ergebnis meiner Arbeit ansehen?', originally unpublished fragment; recently published in Hermes *et al.* [1969] (p. 200)

> 'What May I Regard as the Result of my Work?', (originally unpublished fragment), translation by P. Long and R. White in Hermes *et al.* [1979] (p. 184)

• [1906e] 'Einleitung in die Logik', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 201-212)

> 'Introduction to Logic', (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 185-196)

• [1906f] 'Kurze Übersicht meiner logischen Lehren?', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 213-218)

> 'Brief Survey of my Logical Doctrines', (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 197-202)

• [**1908**] 'Die Unmöglichkeit der Thomaeschen formalen Arithmetik aufs neue nachgewiesen' (mit Schlussbemerkung), *Jahresbericht der Deutschen Mathematiker-Vereinigung* **17** (1908), pp. 52-55, 56; reprinted in Angelelli [1967] (pp. 329-333)

> 'Renewed Proof of the Impossibility of Mr. Thomae's Formal Arithmetic' (with Concluding Remarks), translation by E.-H. W. Kluge in Kluge [1971] (pp. 132-138) and in McGuinness [1984] (pp. 346-350)

• [**1910**] '(Anmerkungen zu) Philip E. B. Jourdain, *The Development of the Theories of Mathematical Logic and the Principles of Mathematics: Gottlob Frege*', in *The Quarterly Journal of Pure and Applied Mathematics* **43** (1912) pp. 237-269; (The German source is not extant, but this section of Jourdain's text is reprinted in full in Gabriel *et al.* [1976], pp. 275-301.)

> '(Footnotes to) Philip E. B. Jourdain, *The Development of the Theories of Mathematical Logic and the Principles of Mathematics*'. In 1910, Frege sent Jourdain comments on his manuscript. Jourdain translated Frege's comments and published them as footnotes to his paper in the *The Quarterly Journal of Pure and Applied Mathematics*. Only Jourdain's translation of these comments (i.e., as published in the *Quarterly Journal*) survives; the footnotes containing Frege's remarks are collated and reprinted in Angelelli [1967] (pp. 334-341)

• [1910-13] (Vorlesungen über Begriffsschrift), an der Universität Jena, nach der Mitschrift von Rudolf Carnap, (‘Begriffsschrift I’ aus dem Wintersemester 1910/11; ‘Begriffsschrift II’ aus dem Sommersemester 1913), originally unpublished; recently published in Gabriel [1996] (pp. 1-42)

> ‘Lectures on Concept Script’, at the University of Jena, as recorded by (Frege’s student) Rudolf Carnap, (originally unpublished). The lectures from the Winter Semester of 1910-11 are called ‘Begriffsschrift I’ and the lectures from the Summer Semester 1913 are called ‘Begriffsschrift II’. No English translation is available.

• [1914] ‘Logik in der Mathematik’, originally unpublished; recently published in Hermes *et al.* [1969] (pp. 219-270)

> ‘Logic in Mathematics’, (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 203-250)
>
> [Note: According to the Introduction to Gabriel [1996], these are Frege’s lecture notes for lectures given at the University of Jena in the Summer Semester of 1914. Carnap attended these lectures and took notes. Gabriel hopes to publish Carnap’s notes from these lectures soon.]

• [1915] ‘Meine grundlegenden logischen Einsichten’, originally unpublished fragment; recently published in Hermes *et al.* [1969] (pp. 271-272)

> ‘My Basic Logical Insights’, (originally unpublished fragment), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 251-252)

• [**1918a**] ‘Der Gedanke. Eine Logische Untersuchung’, in *Beiträge zur Philosophie des deutschen Idealismus* **I** (1918-1919), pp. 58-77; reprinted in Angelelli [1967] (pp. 342-362) and in Patzig [1966] (pp. 30-53)

> ‘Thoughts: A Logical Enquiry’,
>
> - Translation by P. Geach and R. Stoothoff in Geach [1977] (pp. 1-30), McGuinness [1984] (pp. 351-372), Salmon and Soames [1988] (pp. 33-55), and Beaney [1997] (pp. 325-345)
> - Translation by A. Quinton and M. Quinton in Quinton and Quinton [1956], Strawson [1967] (pp. 17-38), and Klemke [1968] (pp. 507-35)

• [**1918b**] ‘Die Verneinung. Eine Logische Untersuchung’, *Beiträge zur Philosophie des deutschen Idealismus* **I** (1918-1919), pp. 143-157; reprinted in Angelelli [1967] (pp. 362-378) and in Patzig [1966] (pp. 54-71)

> ‘Negation: A Logical Investigation’, translation by P. Geach in Geach and Black [1980] (pp. 117-135), in Geach [1977] (pp. 31-53), and in Beaney [1997] (pp. 346-361)

• [1918c] ‘Vorschläge für ein Wahlgesetz’, originally unpublished; recently published in Gabriel and Dathe [2000] (pp. 283-313), edited and introduced by U. Dathe and W. Kienzler.

> ‘Suggestions for an Electoral Law’, (originally unpublished); no English translation is available

• [1919] (Aufzeichnungen für Ludwig Darmstaedter), originally unpublished; recently published in Hermes *et al.* [1969] (pp. 273-277)

> (Notes for Ludwig Darmstaedter), (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 253-257)

• [**1923a]** ‘Logische Untersuchungen. Dritter Teil: Gedankengefüge’, *Beiträge zur Philosophie des deutschen Idealismus* **III** (1923-1926), pp. 36-51; reprinted in Angelelli [1967] (pp. 378-394) and in Patzig [1966] (pp. 72-91)

> ‘Logical Investigations. Third Part: Compound Thoughts’, translation by R. Stoothoof in Stoothoff [1963], in Klemke [1968] (pp. 537-558), in Geach [1977] (pp. 55-77), and in McGuinness [1984] (pp. 390-406)

• [1923b] ‘Logische Allgemeinheit’, originally unpublished fragment; recently published in Hermes *et al.* [1969] (pp. 278-281)

> ‘Logical Generality’, (originally unpublished fragment); translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 258-262)

• [1924a] (Tagebuch), originally unpublished; recently published in Gabriel and Kienzler [1994] (pp. 1067-1098); two pages (‘Tagebucheintragungen über den Begriff der Zahl’) are reproduced in Hermes *et al.* [1969] (pp. 282-283)

> (‘Diary: Written by Professor Dr. Gottlob Frege in the Time from 10 March to 9 April 1924’), (originally unpublished), translation of extant entries by R. Mendelson in Gabriel and Kienzler [1996]; partial translation (‘Diary Entries on the Concept of Number’) by P. Long and R. White in Hermes *et al.* [1979] (pp. 263-264)

• [1924b] ‘Zahl’, originally unpublished fragment; recently published in Hermes *et al.* [1969] (pp. 284-285)

> ‘Number’, (originally unpublished fragment), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 265-266)

• [1924-25a] ‘Erkenntnisquellen der Mathematik und der mathematischen Naturwissenschaften’, originally unpublished; recently published in Hermes *et al.* [1969] (pp. 286-294)

'Sources of Knowledge of Mathematics and the Mathematical Natural Sciences', (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 267-274)

• [1924-25b] 'Zahlen und Arithmetik', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 295-297)

'Numbers and Arithmetic', (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 275-277)

• [1924-25c] 'Neuer Versuch der Grundlegun der Arithmetik', originally unpublished; recently published in Hermes *et al.* [1969] (pp. 298-302)

'A New Attempt at a Foundation for Arithmetic', (originally unpublished), translation by P. Long and R. White in Hermes *et al.* [1979] (pp. 278-281)

Locations of English Translations of Frege's Writings

- Angelelli, I. (ed.), 1967, *Kleine Schriften*, Hildesheim: Olms (contains the English version of Frege [1910]; the German doesn't survive)
- Austin, J. L. (ed. and trans.), 1974, *The Foundations of Arithmetic. A logic-mathematical enquiry into the concept of number*, Oxford: Blackwell, second revised edition (second edition, 1953; first edition, 1950)
- Bartlett, J. M. (trans.), 1964, 'On the Scientific Justification of a Concept-script', *Mind*, **73**: 155-60
- Beaney, M. (ed.), 1997, *The Frege Reader*, Oxford: Blackwell
- Benacerraf, P., and Putnam, H. (eds.), 1983/1964, *Philosophy of Mathematics: Selected Readings*, Cambridge: Cambridge University Press, second edition (first edition, Prentice Hall, 1964)
- Black, M. (trans.), 1948, 'Sense and Reference', *The Philosophical Review* **57**: 207-230
- Black, M. (trans.), 1950, 'Frege Against the Formalists', *The Philosophical Review* **59**: 77-93, 202-220, 332-345 (translations of §§86-137 of Frege [1903a])
- Bynum, T. W. (ed. and trans.), 1972, *Conceptual Notation and Related Articles*, Oxford: Clarendon
- Copi, I., and Gould, J. (eds.), 1964, *Readings on Logic*, New York: Macmillan
- Dudman, V. H. (trans.), 1970, 'The Whole Number', *Mind* **79**: 481-486
- Dudman, V. H. (trans.), 1969, 'On Herr Peano's *Begriffsschrift* and My Own', *Australasian Journal of Philosophy*, **47** (1969): 1-14

- Dudman, V. H. (trans.), 1968, ‘On the Purpose of the Begriffsschrift’, *The Australasian Journal of Philosophy*, **46** (1968): 89-97
- Feigl, H. and Sellars, W., (eds.), 1949, *Readings in Philosophical Analysis*, New York: Appleton-Century-Crofts
- Furth, M. (trans.), 1964, *The Basic Laws of Arithmetic*, Berkeley: U. of California Press
- Gabriel, G., and Kienzler, W. (eds.), 1996, ‘Diary: Written by Professor Gottlob Frege in the Time from 10 March to 9 April 1924’ (translated by R. Mendelson), *Inquiry* **39**: 303-342
- Gabriel, G., Hermes, H., Kambartel, F., Thiel, C., and Veraart, A. (eds.), 1980, *Philosophical and Mathematical Correspondence*, abridged from the German collection Gabriel *et al.* [1976] by B. McGuinness and translated by H. Kaal, Chicago: U. of Chicago Press
- Geach, P., and Black, M. (eds. and trans.), 1980, *Translations from the Philosophical Writings of Gottlob Frege*, Oxford: Blackwell, third edition (second edition, 1970, 1969, 1966, 1960; first edition 1952)
- Geach, P. (ed.), 1977, *Logical Investigations*, trans. by P. Geach and R. Stoothoff, Oxford: Blackwell
- Geach, P. (trans.), 1951, ‘On Concept and Object’, *Mind* **60**: 168-180
- Heijenoort, J. van (ed.), 1967, *From Frege to Gödel, a source book in mathematical logic, 1879-1931*, Cambridge, MA: Harvard University Press
- Hermes, H., Kambartel, F., and Kaulbach, F. (eds.), 1979, *Posthumous Writings*, P. Long and R. White (trans.), Chicago: U. of Chicago Press, 1979; this is a translation by P. Long and R. White of Frege’s work in the German collection Hermes *et al.* [1969]
- Kennick, W. E., and Lazerowitz, M. (eds.), 1966, *Metaphysics: Readings and Reappraisals*, Englewood Cliffs, NJ: Prentice-Hall
- Klemke, E. D. (ed.), 1968, *Essays on Frege*, Urbana, IL: University of Illinois Press
- Kluge, E.-H. W. (trans.), 1971, *On the Foundations of Geometry and Formal Theories of Arithmetic*, New Haven: Yale University Press
- Martinich, A. P. (ed.), 1985, *The Philosophy of Language*, Oxford: Oxford University Press, first edition
- Martinich, A. P. (ed.), 1996, *The Philosophy of Language*, Oxford: Oxford University Press, third edition (second edition, 1990; the first edition of 1985 is listed separately as Martinich [1985])

- McGuinness, B. (ed.), 1984, *Collected Papers on Mathematics, Logic, and Philosophy*, translated from the German collection Angelelli [1967] by M. Black, V. H. Dudman, P. Geach, H. Kaal, E.-H. W. Kluge, B. McGuinness, and R. H. Stoothoff, Oxford: Basil Blackwell
- Nagel, E., and Brandt, R. (eds.), 1965, *Meaning and Knowledge*, New York: Harcourt, Brace and World
- Quinton, A., and Quinton, M. (trans.), 1956, 'The Thought: A Logical Enquiry', *Mind*, **65**: 289-311
- Rand, R. (trans.), 1961, 'About the Law of Inertia', *Synthese*, **13**/4 (December): 350-363
- Runes, D. (ed.), 1962, *Classics in Logic: Readings in Epistemology, Theory of Knowledge, and Dialectics*, New York: Philosophical Library
- Salmon, N., and Soames, S., (eds.), 1988, *Propositional Attitudes*, Oxford: Oxford University Press
- Stachelroth, J., and Jordan, P. E. B. (trans.), 1915 - 1917, '*The Fundamental Laws of Arithmetic*', *The Monist* **25**/4 (October 1915): 481-494; **26**/2 (April 1916): 182-199 ('The Fundamental Laws of Arithmetic: Psychological Logic'); **27**/1 (January 1917): 114-127 ('Class, Function, Concept, Relation')
- Stoothoff, R. (trans.), 1963, 'Compound Thoughts', *Mind* **72**: 1-17
- Strawson, P. (ed.), 1967, *Philosophical Logic*, London: Oxford University Press
- Szabo, M. E. (trans.), 1960, 'The Foundations of Geometry', *The Philosophical Review*, **69**: 3-17
- Woodward, B. (trans.), 1967, 'A Letter to Bertrand Russell on Russell's Paradox' (dated Jena, 22 June 1902), in van Heijenoort [1967] (pp. 126-128)

Principal German Collections and Reprints of Frege's Work

- Angelelli, I. (ed.), 1967, *Kleine Schriften*, Darmstadt: Wissenschaftliche Buchgesellschaft and Hildesheim: Olms
- Angelelli, I. (ed.), 1964, *Begriffsschrift und andere Aufsätze* (reprint of Frege [1879]), Hildesheim: Olms
- Gabriel, G., and Dathe, U. (eds.), 2000, *Gottlob Frege. Werk und Wirkung*, Paderborn: Verlag Mentis
- Gabriel, G. (ed.), 1996, 'Vorlesungen über Begriffsschrift, nach der Mitschrift von Rudolf Carnap', *History and Philosophy of Logic* **17** (1996): iii-xvi, 1-48

- Gabriel, G., and Kienzler, W. (eds.), 1994, ‘Gottlob Freges politisches Tagebuch’, *Deutsche Zeitschrift für Philosophie* **42**/6: 1057-98
- Gabriel, G., Hermes, H., Kambartel, F., Thiel, C., and Veraart, A. (eds.), 1976, *Wissenschaftlicher Briefwechsel*, Hamburg: Felix Meiner
- Hermes, H., Kambartel, F., and Kaulbach, F. (eds.), 1969, *Nachgelassene Schriften*, Hamburg: Felix Meiner; second edition, revised and expanded, 1983
- Janik, A. (ed.), 1989, ‘Briefe an Ludwig Wittgenstein’ (aus den Jahren 1914-1920), in B. McGuinness and R. Haller (eds.) *Wittgenstein in Focus – Im Brennpunkt: Wittgenstein*, (special issue of) *Grazer Philosophische Studien* **33**/**34** (1989), pp. 5-33
- Patzig, G. (ed.), 1962, *Funktion, Begriff, Bedeutung: Fünf logische Studien*, Göttingen: Vandenhoeck & Ruprecht
- Patzig, G. (ed.), 1966, *Logische Untersuchungen*, Göttingen: Vandenhoeck & Ruprecht
- Thiel, C. (ed.), 1986, *Die Grundlagen der Arithmetik* (critical edition of Frege [1884]), Hamburg: Felix Meiner
- Thiel, C. (ed.), 1998, *Grundgesetze der Arithmetik*, (critical edition of Frege [1893] and [1903a]), Hildesheim: Olms

Stanford Encyclopedia of Philosophy
Supplement to Propositional Attitude Reports

Definite Descriptions

Russell's reply to Frege employed his theory of definite descriptions as a part of the reply. See Bertrand Russell. According to this theory, sentences with definite descriptions, phrases of the form ‘the F’, are to be analyzed in a way that indicates that the definite description is not a unitary part of the sentence. In particular,

> The man wearing a beret is bald.

is to be understood as equivalent to the following:

> Some man is wearing a beret, and at most one man is wearing a beret, and every man wearing a beret is bald.

This facilitates a response to versions of the Fregean puzzle involving definite descriptions. More complex sentences with definite descriptions will be ambiguous. Applying the account of definite descriptions to belief sentences, we get these results:

> George IV believed that the person who wrote *Waverly* was famous.
>
> **A version**: George IV believed that some person wrote *Waverly*, at most one person wrote *Waverly*, and every person who wrote *Waverly* is famous.
>
> **B version**: Some person wrote *Waverly*, at most one person wrote *Waverly*, and every person who wrote *Waverly* is an *x* such that George IV believed that *x* was famous.

If George IV was in doubt about the identity of the author of *Waverly*, wondering whether it was Sir Walter Scott or someone else, then the **A version** would be true, but the **B version** could be false. George IV would be engaged in a general belief, without specific knowledge of any person's writing of *Waverly*, and that is captured by the **A version**.

Retaining the definite description, we could distinguish the two readings of the sentence as follows:

> **A version**: George believed that this was true: the author of *Waverly* was famous.

> **B version**: The author of *Waverly* was an individual *x* such that George IV believed that *x* was famous.

This is a distinction in the *scope* of the definite description. In the **A version**, the definite description has small scope, within the propositional object clause of the belief ascription. In the **B version**, the definite description has large scope, in effect "picking out" an individual and then ascribing to George IV a belief about that individual. The **A version** can also be identified as a *de dicto* ascription of belief (relating him to a *dictum*, a complete proposition), whereas the **B version** is a *de re* ascription of belief (relating him to an individual, a *res*, that his belief is about). (See the discussion of *the de re/de dicto distinction*.)

Frege's puzzle concerns questions like the following:

> How could this be true: *Albert believes that the Venus rises in the morning*, though this is not true: *Albert believes that the evening star rises in the morning*?

Since Venus is the evening star, why does substitution of one name for the other affect the correctness of the ascription? The russellian answer lies in recognizing the ambiguity in the second sentence, which involves a definite description.

> Albert believes that the evening star rises in the morning.
>
> **A version**: Albert believes that this is true: the evening star rises in the morning.
>
> **B version**: The evening star is an individual *x* such that Albert believes that *x* rises in the morning**.**

The falsity of the **A version** does not conflict with Albert's belief about Venus, that it rises in the morning, since the **A version** ascribes a belief in a complete dictum and does not relate Albert to an individual. The difference in truth-value between *Albert believes that the Venus rises in the morning* and *Albert believes that the evening star rises in the morning* is no longer puzzling. The **B version** captures what is right: that Albert has a belief about a particular individual (that we are now identifying as the evening star, even though Albert wouldn't so identify it in this context), a belief that it rises in the morning.

This solution requires that we make sense of quantification into belief contexts (the **B version**), so that we have a real distinction. (See the discussion of *the de re/de dicto distinction*.) The solution is also limited, because it applies only to versions of the puzzle that involve definite descriptions.

tjmckay@syr.edu

Return to Propositional Attitude Reports

First published: February 16, 2000
Content last modified: February 16, 2000

Stanford Encyclopedia of Philosophy
Supplement to Propositional Attitude Reports

The De Re/De Dicto Distinction

There is an important distinction in belief attributions, and, according to some, in the beliefs we attribute. W. V. Quine's example (Quine 1956) brings this out nicely through the recognition of an ambiguity in

> [1] Ortcutt believes that someone is a spy.

This could mean just that

> [2] Ortcutt believes that there are spies

or that Ortcutt has more interesting information:

> [3] Someone is an *x* such that Ortcutt believes that *x* is a spy.

The distinction here can be seen as a distinction of scope for the existential quantifier. In [2], the existential quantifier is interpreted as having small scope, within the propositional clause of the belief attribution.

> [2*] Ortcutt believes: $\exists x$, x is a spy.

In [3], however, the existential quantifier has large scope, selecting an individual and then ascribing a belief that relates Ortcutt to that particular individual.

> [3*] $\exists x$, Ortcutt believes that x is a spy.

The ambiguity in [1] and the simple way of distinguishing the two interpretations in [2*] and [3*] suggest that we are on to something. Russell's theory of definite descriptions employs just such a distinction in answering Frege's puzzles about belief. (See Definite descriptions.)

We just identified the distinction between [2*] and [3*] as a distinction in the scope of the quantifier. But [3*] also introduces something that needs a further account. On the standard semantics for quantification, the interpretation of [3*] requires that we be able to say when an individual *x* satisfies the open sentence ‘ Ortcutt believes that *x* is a spy’. Looked at another way, we no longer have a complete statement of a proposition in the propositional clause position in the sentence, since ‘that *x* is a spy’ does not express a proposition with a definite truth-value. Considerations like these have motivated the identification of a

distinction between *de dicto* belief attributions like [2] and [2*] and *de re* belief attributions like [3] and [3*]. The purely *de dicto* attribution relates Ortcutt to a *dictum*, a complete propositional content. The *de re* attribution relates him to a *res*, an individual that his belief is about. To say that an individual satisfies ‘Ortcutt believes that *x* is a spy’ is to say that Ortcutt has a *de re* belief about that individual. (For more on *de re* belief, see Quine 1956, Burge 1977.)

Although it may be tempting to think of these as an ambiguity in the verb ‘believe’, the section Ambiguity theories explains why this is not possible.

Thomas McKay
tjmckay@syr.edu

Return to Propositional Attitude Reports

First published: February 16, 2000
Content last modified: February 16, 2000

Indexicals

Indexicals are linguistic expressions whose reference shifts from utterance to utterance. ‘I’, ‘here’, ‘now’, ‘he’, ‘she’, and ‘that’ are classic examples of indexicals. Two people who utter a sentence containing an indexical may say different things, even if the sentence itself has a single linguistic meaning. For instance, the sentence ‘I am female’ has a single linguistic meaning, but Fred and Wilma say different things when they utter it, as shown by the fact that Fred says something false, while Wilma says something true.

Philosophers have several reasons for being interested in indexicals. First, they wish to describe their meanings and fit them into a more general theory of meaning. Second, they wish to understand the logic of arguments containing indexicals, including, for instance, Descartes's *Cogito* and various skeptical arguments that contain ‘I’ and other indexicals. Third, they think that reflection on indexicals may give them some insight into such matters as the nature of belief, self-knowledge, and consciousness.

---

1. Some Examples, Some Terminology, and Some Distinctions

1.1 Examples of Indexicals and Some Terminology

The indexicals that philosophers have studied most are the pronouns ‘I’, ‘he’, ‘she’, ‘this’, and ‘that’; the adverbs ‘here’, ‘now’, ‘actually’, ‘presently’, ‘today’, ‘yesterday’, and ‘tomorrow’; and the adjectives ‘actual’ and ‘present’. This list comes (more or less) from David Kaplan (1989a), whose work on indexicals is perhaps the most influential in the field.

An indexical's referent is determined, in part, by *extra-linguistic context* (for instance, the time and location of the speaker and the speaker's intentions). An indexical's referent can also vary from context to context. Thus indexicals are commonly called *context-sensitive expressions*. (The *content* of an indexical can also vary from context to context. This claim will be explained in section 3.1.)

1.2 Indexical and Non-indexical Uses of Pronouns

Some of the expressions in Kaplan's list have *non*-indexical uses. ‘He’, ‘his’, ‘she’, ‘her’, and ‘that’ are sometimes used like *bound variables* in formal languages. For example, the occurrence of ‘he’ in (1) (on the relevant understanding) functions like a variable that is bound by the quantifier phrase ‘every man’. Similarly, ‘her’ in (2) (under the appropriate reading) is bound by ‘every girl’.

1. Every man believes that he is smart.
2. Every girl loves her father.

These same pronouns are also sometimes used *anaphorically*. That is, some of their utterances seem to depend for their reference on prior linguistic context. (See the entry on anaphora.) For example, ‘he’ appears to be used anaphorically in discourse (3).

3. Johnny hit a home run. He was very happy.

Finally, there are the uses of these pronouns in which we shall be interested, the *indexical* (or

demonstrative or *deictic*) uses, as in (4) and (5).

4. He likes sardines [pointing at Fred], but he does not [pointing at Barney].
5. His car [pointing at Alfred] is dirty, but his car [pointing at Alonzo] is clean.

Most philosophers and linguists think that these different uses are closely related, and are not merely uses of distinct homonymous words. See section 5.4 for more on this topic. But in most of what follows, we shall ignore the non-indexical uses of these pronouns, and concentrate solely on their indexical uses.

1.3 Pure Indexicals and True Demonstratives

Kaplan (1989a) distinguishes between two different sorts of indexical, *pure indexicals* and *true demonstratives*. The true demonstratives include ‘he’, ‘she’, ‘his’, ‘her’, and ‘that’, while the pure indexicals include ‘I’, ‘today’, ‘tomorrow’, ‘actual’, ‘present’, and (perhaps) ‘here’ and ‘now’. (More on ‘here’ and ‘now’ below.) The two types of indexical differ in how their references are determined. The reference of an utterance of a true demonstrative is determined (in part) by the speaker's accompanying actions or intentions. For example, the reference of an utterance of ‘that’ is determined (in part) by the speaker's accompanying pointing gestures, or by the speaker's intention to refer to a particular object. (See section 2 for more on the nature of reference-fixing for true demonstratives.) The reference of a pure indexical is *not* determined by the speaker's actions or intentions in this way. For instance, an utterance of ‘I’ refers to the speaker, whether or not she points at herself, and an utterance of ‘tomorrow’ refers to the day after the day of utterance, regardless of the speaker's intention to refer to some particular day. We can say (loosely speaking) that the reference of pure indexicals is *automatic*, whereas the reference of true demonstratives requires something extra from the speaker.

From here on, we shall use the term ‘demonstrative’ to mean *true demonstrative*, in the above sense. ‘Indexical’ shall be used here as a generic term, so that it encompasses both (true) demonstratives and pure indexicals.

Kaplan includes ‘here’ and ‘now’ in his list of pure indexicals, but this seems inaccurate. Every utterance of ‘now’ automatically refers to a time interval that includes the moment of utterance, but the *extent* of the time interval surrounding the moment of utterance differs radically from utterance to utterance. Consider, for example, a typical utterance of ‘I am ready to leave now’ and a typical utterance of ‘People now ride in cars rather than horse-drawn carriages’. The extent of the time interval seems to depend on the speaker's intentions.[1] Similar remarks go for ‘here’ and the spatial extent of the location surrounding the location of utterance.[2] Thus it seems that ‘here’ and ‘now’ are demonstratives rather than pure indexicals.

1.4 Which Expressions are Indexicals?

Our initial list of indexicals in English (basically, Kaplan's list) is incomplete. A complete list would include at least the plural expressions ‘we’, ‘those’, ‘they’, and ‘theirs’ (for discussion of plural

indexicals, see Nunberg 1993). There are still more indexicals in English, but how many more is controversial.

Some philosophers and linguists (Reichenbach 1947, Partee 1973, Salmon 1989) claim that words and morphemes that indicate tense are indexicals, because (very roughly) they refer to different time intervals from context to context. Some have argued (Kratzer 1977, Lewis 1979b) that modal expressions, such as ‘necessarily’ and ‘possibly’, are indexical, because they vary in the type of modality they express from context to context (for example, nomological possibility in one context, metaphysical possibility in another). The subjunctive conditional connective "if it were the case that ... then it would be the case that ..." seems to be context-sensitive, because the range of possibilities relevant for determining the truth of such sentences seems to shift from context to context (see Lewis 1973, pp. 66-68). Utterances of ‘come’, ‘go’, ‘left’, and ‘right’ seem to invoke different points of reference, or different perspectives, in different contexts (see Fillmore 1972, 1975 and Lewis 1979b).

Some philosophers have claimed that propositional attitude verbs, like ‘believe’ and ‘know’, are indexicals. Richard (1990) claims that the sentences ‘Lois believes that Superman can fly’ and ‘Lois believes that Clark Kent can fly’ have the same truth value in some contexts, but different truth values in other contexts, though Lois undergoes no relevant changes. He argues that this occurs because, in different contexts, the verb ‘believes’ invokes different translation relations between the sentence inside the ‘that’-clause of the belief sentence and the sentences that the believer accepts; therefore, ‘believes’ expresses different relations in different contexts. (See the entry on propositional attitude reports.) Cohen (1988), DeRose (1995), and Lewis (1996) claim that the sentence ‘George knows that he has a hand’ can be false in a context in which the speaker is considering skeptical arguments, but can be true in another, more ordinary, context in which no skeptical arguments are under consideration, even though no relevant change occurs in George. They explain this by claiming that ‘know’ is an indexical that expresses different relations in different contexts, depending upon the relevant alternatives being considered in the speaker's context, or the standards of justification in force in the speaker's context.

The adjective ‘rich’ seems to be context-sensitive, for the truth value of ‘Myles is rich’ seems to vary from context to context, depending on which property or comparison class is salient in the context (for example, the class of all Americans, the class of philosophers, the class of university presidents, or the class of CEO's of large organizations). There are at least two competing accounts of this apparent context-sensitivity. On one, ‘rich’ is a unary predicate whose extension (and content) varies from context to context. On a second account, ‘rich’ is a binary predicate (*x* is rich for a *y*). Sometimes the second argument is linguistically supplied and pronounced, as in ‘Myles is rich for a philosopher’. But sometimes the second argument is unpronounced, as in ‘Myles is rich’, and in those cases, context supplies a property or comparison class to serve as the second argument of the binary predicate. The apparent context-sensitivity of other comparatives, such as ‘tall’, ‘large’, ‘heavy’, ‘hot’, and ‘fast’, might also be explained in either of these two competing ways. Some theorists (Partee 1989, Condoravdi and Gawron 1996) have argued in favor of the second sort of account for certain expressions that have indexical uses, such as ‘local’. In an utterance of ‘A local bar is selling cheap beer’, the utterance of ‘local’ refers to a location near the speaker. But in an utterance of ‘Every football fan watched the Superbowl at a local bar’, the utterance of ‘local’ does not refer to a location near the speaker. Rather, the utterance says that each

football fan *x* watched the Superbowl at some bar that is local to *x*. This can be easily explained if ‘local’ has a second argument position (*y* is local to *x*) whose value is supplied by context in the first case, but which gets bound by the quantifier ‘every football fan’ in the second.

According to Stanley and Szabo (2000), *all* nouns have a hidden argument position that can be bound or contextually filled. They use this claim to explain how different utterances of a single sentence containing a quantifier phrase can (apparently) quantify over different domains. For example, the claim made by ‘Every student is present’ can vary from context to context: in some contexts, an utterance of it claims that every student *who is in school S* is present, while in other contexts, such an utterance claims that every student *who is in class C* is present. Stanley and Szabo's theory entails that *every noun* is an indexical (and, perhaps, that every adjective, verb, and adverb is also).

On some views, the extensions (and contents) of *vague* expressions shift from context to context. ‘Bald’ is a paradigm case of a vague expression. Lewis (1979b) and Soames (1999) hold that, if David has an appropriate number and distribution of hairs on his head, then ‘David is bald’ may be true in one context and false in another. Lewis and Soames explain this by holding that the extension of ‘bald’ varies from context to context; Soames holds that it expresses different properties in different contexts. Both of them extend this same view to all vague expressions. Since nearly every expression is vague, their views imply that nearly every expression is an indexical.

Summarizing: indexicality in English extends beyond Kaplan's list, though how far beyond is controversial. In what follows, however, we shall concentrate on uncontroversial cases, and mainly on the expressions in Kaplan's list.

2. Reference-fixing for Demonstratives

One major topic of work on indexicals is reference-fixing for (true) demonstratives. As mentioned before, (true) demonstratives differ from pure indexicals, in that the reference of an utterance of a true demonstrative is not fixed "automatically" by the act of utterance alone. Something more is required to fix the reference of a demonstrative utterance. The nature of this "extra something" is controversial, but two obvious candidates are *pointing gestures* and *speaker's intentions*.

In his 1989a, Kaplan emphasizes the role of pointing gestures in fixing the reference of a demonstrative utterance. More precisely, he says that the reference of a demonstrative utterance is fixed by a *demonstration*. He describes a demonstration as "typically, though not invariably, a (visual) presentation of a local object discriminated by a pointing" (1989a, p. 490). Kaplan, however, changes his mind in his 1989b.[3] He there points out that a demonstration is "typically directed by the speaker's intention to point at a perceived individual on whom he has focused". He calls such intentions *directing intentions* and then says that he has come to "regard the directing intention, at least in the case of perceptual demonstratives, as criterial, and to regard the demonstration as a mere *externalization* of this inner intention." (1989b, p. 582)

Other theorists hold other views. Devitt (1981) says that the referent of an utterance of ‘that’ is the item that stands in a certain causal relation to the utterance. McGinn (1981) proposes that the referent of an utterance of ‘that *F*’ is the first *F* to intersect the line projected from the speaker's pointing finger. Wettstein (1984) says that the reference of ‘that’ is determined by the cues that a competent and attentive addressee would reasonably take the speaker to be exploiting. Reimer (1991a, 1991b) argues, contra Kaplan's later view, that demonstrative utterances can refer to objects that are not the targets of the speaker's directing intentions. Bach (1992a, 1992b) defends a version of Kaplan's later view from Reimer's criticisms; Bach says that the reference of a demonstrative utterance is fixed by certain of the speaker's communicative intentions.

One *prima facie* problem for demonstration-theories is that some utterances of demonstratives seem to refer even though the speaker does not produce a demonstration (for instance, a pointing gesture). If ‘here’ and ‘now’ are demonstratives (as suggested earlier), then they may present a particular difficulty for demonstration-theories, for demonstrations seem irrelevant to determining the referents of their utterances.

One *prima facie* problem for intention-theories is that speakers typically have a large number of intentions when they use demonstratives, and these intentions may conflict (as Bach recognizes in his 1992a and 1992b; see also Perry 1997). For example, a speaker who utters ‘he’ may intend to speak about Joe, and about the man that she (the speaker) sees over there, and about the man about whom others are speaking, and about the man at whom she (the speaker) is pointing. The speaker may think that these are the same person when they are not. Different intention-theories can select different of these intentions as reference-fixing, and so make different predictions about who the referent is in certain cases. Not all intention theories have been completely clear about what they take the relevant reference-fixing intention to be.

3. Kaplan's Theory of Indexicals

Thus far, we have not tried to describe the meanings of indexicals in any systematic way. In this section, we present one theory that attempts to do so, namely Kaplan's theory. We begin with Kaplan's theory because (as mentioned before) it is perhaps the most influential in the field. We first present an example that motivates some of Kaplan's distinctions. We then present Kaplan's theory. In section 4, we consider a criticism of Kaplan's theory, and some alternatives to Kaplan's theory.

3.1 An Example and Some Intuitive Distinctions

Many philosophers hold that indexicals have (at least) two different sorts of meaning. To see why, consider examples (6) and (7).

6. Fred: "I am female."
7. Wilma: "I am female."

There is a clear sense in which Fred's utterance and Wilma's utterance share a meaning, for they utter the very same unambiguous sentence. Let's say that their utterances have the same *linguistic meaning*. Nevertheless, their utterances also seem to differ in meaning, in some sense, for Fred and Wilma say different things: Fred says that *he* is female, whereas Wilma says that *she* is female. Moreover, Fred says something that is false, while Wilma says something that is true. Traditionally, this difference in truth value would be taken to show that Fred and Wilma assert different *propositions*. In view of these considerations, let's say that Fred's and Wilma's utterances differ in *content*, where the content of an utterance of a full indicative sentence is a proposition.

So far, we have an intuitive distinction between two sorts of meaning for utterances of indexical sentences: linguistic meaning and content. The theories of meaning that we shall consider below, beginning with Kaplan's, attempt to describe these (apparent) meanings more systematically.

Some theorists, however, do not find the above intuitions and distinctions compelling. Lewis (1980), for instance, thinks that these intuitions are shaky, at best; he, in any case, questions their significance for semantic theory. Philosophers who eschew meanings altogether (such as those who favor Davidsonian theories of meaning) seek semantic theories that ascribe only extensions to indexical expressions (e.g., truth values to sentences, and referents to singular terms), with respect to contexts and (perhaps) other indices. These theorists would not take the extra step of hypothesizing the existence of propositions or other sorts of meanings. See, for example, Burge (1974), Weinstein (1974), Larson and Segal (1995), and Lepore and Ludwig (2000). Nevertheless, in most of what follows, we concentrate on theories of meaning for indexicals that try to respect the above distinctions and intuitions in the most straightforward way possible, by hypothesizing (at least) two sorts of meaning.

3.2 Basics of Kaplan's Theory

In Kaplan's theory, linguistic expressions have contents *in*, or *with respect to*, *contexts*. Each context has at least an agent, time, location, and possible world associated with it.[4] The content of ‘I’ with respect to a context C is the agent of C; the content of ‘here’ is the location of C; and the content of ‘now’ is the time of C. The content of a predicate, with respect to a context, is a property or relation. The content of a sentence, with respect to a context, is a structured proposition, that is, a proposition that can have individuals, properties, and relations as *constituents*. The content of a sentence S with respect to C is made up of the contents of the words in S with respect to C.[5]

To illustrate, consider the sentence ‘I am female’. Suppose that the agent of context C is Fred. Then the content of ‘I’ in C is Fred himself, while the content of ‘is female’ in C is the property being-female. The content of the whole sentence, in C, is a proposition whose constituents are just these two items. We can represent this proposition with the following ordered pair.

<Fred, being-female>

The content of ‘I’ with respect to a context C* in which Wilma is the agent is Wilma herself, and the

content of ‘I am female’ in C* is the proposition <Wilma, being-female>. Thus the word ‘I’ and the sentence ‘I am female’ have different contents in different contexts.

The content of a sentence, with respect to a context, has a truth value *at the world of the context*. Kaplan therefore says that the content is either true or false *in the context*. For instance, the content of ‘I am female’, with respect to the above context C (in which Fred is the agent) is the proposition <Fred, being-female>. This proposition is false at the world of C (call this world ‘W’). So, Kaplan says that this proposition is false in context C. But the content of ‘I am female’ with respect to context C* (in which Wilma is the agent) is true with respect to the world of context C*, which is (also) W; so this second proposition is true in context C*.

On Kaplan's theory, we can also speak about the truth values of *sentences*, as opposed to contents (or propositions). The truth value of a *sentence* (as opposed to a proposition) depends on *two* parameters, context *and* world, on this theory. For example, the *sentence* ‘I am female’ is false with respect to C and W, but is true with respect to C* and W. (Notice that the world is the same both times, but the context is different.) Thus the sentence's truth value is *doubly-relativized*. This sort of double-relativization is often called *double-indexing*. (See Vlach 1973 and Kamp 1973 for early examples of semantic theories for indexicals that use double-indexing.)

The content of a sentence, with respect to a context, can be also evaluated for truth at a world other than the world of the context. For example, the content of ‘I am a philosopher’, with respect to C, is the proposition that Fred is a philosopher. This proposition is false at W (let's suppose). But this proposition is true at some other world, say W*, in which Fred is a philosopher. Thus, the *sentence* ‘I am a philosopher’ is false in C and W, but true at C and W*. (Notice that the context is the same both times, while the world is different.) Therefore, the sentence ‘It is *possible* that I am a philosopher’ is true with respect to C and W; and the content of the sentence, with respect to C, is true in C.

Kaplan identifies the linguistic meaning of an expression with its *character*, which is a function from contexts to contents that delivers the expression's content at each context.[6] So, for example, the character of ‘I’ is a function on contexts whose value at any context is the agent of the context; its value at a context in which Fred is the agent is just Fred himself, whereas its value at a context in which Wilma is the agent is Wilma. The character of ‘here’ is a function whose value at each context C is the location of C. The character of a simple predicate, like ‘is female’, is a function on contexts that delivers the appropriate property or relation at every context (in this case, the character delivers the same property at every context, namely being-female). The character of a sentence is a function from contexts to the structured propositional content of that sentence at each context.

Kaplan's theory can be extended to other indexicals, including demonstratives, by adding further suitable features to contexts. For instance, if each context has an associated day, then we can say that the content of ‘today’ in a context is the day of the context, and that its character is a function on contexts whose value at each context is the day of the context. The content of ‘you’ with respect to a context is the addressee of the context. The content of ‘that’ in a context is the demonstat*um* of the context. If we wish to deal with sentences that contain more than one occurrence of ‘you’ or ‘that’, then we can add

sequences of addressees and demonstrat*a* to contexts, and add subscripts to occurrences of ‘you’ and ‘that’; for instance, the content of ‘you_1’ is the first addressee of the context, the content of ‘you_2’ is the second, and so on. (See section 5.2 for more on multiple occurrences of demonstratives.)

We began in section 3.1 with some intuitions about Fred's and Wilma's *utterances* in (6) and (7). It's important to note that utterances are not the same as linguistic expressions; this is shown by the fact that in (6) and (7), Fred and Wilma produce *two* utterances of *one* linguistic expression (the sentence ‘I am female’). Rather, utterances are actions in which an agent utters an expression. Kaplan's theory does not, strictly speaking, ascribe contents or characters to utterances. It instead ascribes characters to *expressions* and contents to *expressions-in-contexts*, which we can think of as pairs of expressions and contexts. So, strictly speaking, Kaplan's theory does not directly confirm or disconfirm our initial intuitions about the meanings of Fred's and Wilma's utterances.

We can, however, *extend* Kaplan's theory to utterances in a rather natural way. (In fact, Kaplan often seems to have the following kind of extension of his theory in mind when he uses judgments about utterances to motivate his view.) Kaplan's theory ascribes a character to the *sentence* that Fred and Wilma utter. So a natural extension of Kaplan's theory would ascribe that character to their two utterances of the sentence. Fred's utterance has a certain agent (Fred himself), and occurs at a certain time, place, and world. Kaplan's theory ascribes a content to the sentence ‘I am female’ with respect to a context with Fred as its agent, and with that associated time, place, and world. So a natural extension of Kaplan's theory could assign this content to Fred's utterance. Similarly, with appropriate changes, for Wilma's utterance. This extension of Kaplan's theory does confirm our intuitions about (6) and (7). (But this extension of Kaplan's theory also has some limitations; see sections 5.2 and 5.3 below.)

3.3 Logic, Logical Truth, Validity, and Necessity

Kaplan's (1989a) formal theory contains an elaborate logic. We shall concentrate here on aspects of the logic that can be understood without going into formal details.

Recall that a sentence S is true in a context C iff the content of S in C is true in the world of C. So, for instance, if the agent of context C is Fred and the world of C is W, then ‘I am hungry’ is true in C if and only if the proposition that Fred is hungry is true in world W. Consider now the sentence ‘If I am hungry, then I am hungry’. This sentence is true in every context. Kaplan says that a *logical truth* is a sentence that is true in every context.[7] Thus Kaplan's theory validates our intuition that this sentence is a logical truth. Kaplan says that an argument is *valid* iff: for every context C, if the premises of the argument are true in C, then the conclusion is true in C. Under this definition of validity, arguments (8) and (9) are valid.

8. I think. Everything that thinks exists. Therefore, I exist.
9. This is a hand. If this is a hand, then I am not a brain in a vat. Therefore, I am not a brain in a vat.

Thus Kaplan's logic for indexicals might help us understand the logic of some philosophically interesting

arguments. (For further discussion of the logic of indexicals and its relevance to Descartes's *Cogito*, see Forbes forthcoming.)

There are logical truths that are peculiar to indexical expressions in Kaplan's system. For instance, Kaplan supposes that, for any context, the agent of the context exists in the world of the context. Thus the sentence ‘I exist’ is true in every context, and counts as a logical truth. (Therefore, argument (8) is valid simply because its conclusion is valid. This may be relevant to the interpretation of Descartes's *Cogito*. See Forbes forthcoming.) Kaplan further supposes that for every context C, the agent of C is located at the time and place of C. Thus the sentence ‘I am here now’ is true in every context, and is a logical truth. (For critical discussion, see Vision 1985 and Salmon 1991.)

However, in most contexts, the contents of these sentences will be *contingent*, that is, true but not necessary. Consider the sentence ‘I exist’. It is a logical truth, according to Kaplan's theory. But its content with respect to a context in which Fred is the agent is the proposition that Fred exists. Since Fred could have failed to exist, this proposition is contingent. Thus, on Kaplan's theory, ‘I exist’ is a logical truth whose content, in many contexts, is contingent. (Similarly for ‘I am here now’.) Now if ‘I exist’ is a logical truth, then it's reasonable to think that Fred knows *a priori* the proposition that he expresses when he utters ‘I exist’ (at least when Fred considers the proposition in that way). If he does know this proposition *a priori*, then Fred has *a priori* knowledge of a contingent proposition. Kaplan thus claims that his logic of indexicals provides examples of Kripke's (1980) contingent *a priori*. (For discussion, see Salmon 1991 and Forbes 1989, forthcoming.)

3.4 Direct Reference and Rigid Designation

Kaplan (1989a) claims that indexicals are devices of *direct reference*. By this he means that the *content* of an indexical, with respect to a context C, is simply the object to which it refers in C; its content is *not* a property (or descriptive condition) that determines the referent. For instance, the content of ‘I’ in C is just the agent of C; its content in C does *not* include the property of being-the-agent-of-C, or any other sort of property.[8]

Kaplan (1989a) also says that indexicals are *rigid designators*. The notion of a rigid designator comes from Kripke (1980), who defines a rigid designator to be an expression that has the same extension (or referent) with respect to all possible worlds.[9] When Kaplan claims that indexicals are rigid designators, he means (roughly) that, once a referent for an indexical is determined by a context, that same object is the one that is relevant for determining the truth value of a sentence containing that indexical at all worlds. For example, if Fred is the agent of context C, then *Fred's* state of hunger (and no one else's) is what is relevant for determining the truth of ‘I am hungry’ with respect to C and any world W whatsoever (whether or not W is the world of C, and whether or not Fred utters ‘I’ in W).

To state Kaplan's view more precisely, let's recall that we earlier spoke of both the *content* of a sentence with respect to a context, and the *truth value* of a sentence with respect to a context *and* a world (the truth value of the *sentence* was doubly-relativized). We can similarly speak of the *content* of a *singular term*

with respect to a context, and its *referent* with respect to a context *and* a world.[10] The definite description ‘the person who invented bifocals’ has the same *content* with respect to all contexts; this content includes the property of being-a-person, the relation of inventing, and so on. But the *referent* of ‘the person who invented bifocals’, with respect to a context *and* a world, varies from world to world, because the person who invented the bifocals varies from world to world. The situation is reversed for ‘I’, Kaplan claims. The content of ‘I’ varies from context to context: its content is Fred in one context, Wilma in another, and so on. But given a single context C, the referent of ‘I’ with respect to C and world W is the same for any world W whatsoever. For instance, if Fred is the content of ‘I’ with respect to context C, then the referent of ‘I’, with respect to C and any world W whatsoever, is Fred. Thus, for all worlds W, ‘I am hungry’ is true at C and W if and only if *Fred* is hungry in W. Therefore, Kaplan says that ‘I’ is a rigid designator.[11]

Kaplan says that all directly referential expressions are rigid designators: if the content of an expression (at a context) is an individual, then that individual is the referent of that expression at that context and any world whatsoever. However, some rigid designators are not directly referential. For instance, the referent of the expression ‘the sum of 2 and 3’ with respect to any context and any world is the number 5, and so this expression is a rigid designator. But the expression ‘the sum of 2 and 3’ is not directly referential, for its content (in a context) is not simply an individual (like the number 5), but is instead a complex object whose constituents include the numbers 2 and 3, and the relation *x*-is-a-sum-of-*y*-and-*z*.

4. A Criticism of Kaplan's Theory and Some Alternatives

4.1 Some Preliminaries Concerning Belief and Cognitive Significance

Kaplan's theory is by no means universally accepted. The most common objections to Kaplan's theory concern *belief* and *cognitive significance*.[12] To understand the apparent problems, consider Fred's utterance in (10).

10. Fred: "You are hungry" [addressing Barney].

Let's suppose that Fred *assertively* utters the sentence. Then it seems that he *asserts* the *proposition* that his utterance expresses. If this proposition is the same as its (alleged) Kaplanian content, then he asserts the Kaplanian content.[13] Furthermore, if Fred utters the sentence *sincerely*, then he *believes* the proposition that his utterance expresses. Now according to Kaplan's theory, the content of Fred's utterance is a singular proposition, that is, a proposition that contains an individual as a constituent, in this case Barney. Many philosophers think that singular propositions could not be the things that people believe, and so Kaplan's theory cannot account for the cognitive significance of indexicals. What follows is an example intended to show this.

4.2 An Example and a Criticism of Kaplan's Theory

Imagine that Fred is looking at Barney, but that Barney is turned so that Fred directly sees only his left side. Suppose that, at the same time, Fred is viewing Barney's right side indirectly, via a mirror. Suppose the right side of Barney's face is masked; suppose finally that Barney is wearing a very unusual costume, in which the left side appears to be a business suit while the right side appears to be a pair of swimming trunks. Then Fred might reasonably and sincerely utter (11), while addressing Barney and pointing at the mirror image.

11. You [addressing Barney] are wearing a business suit, but he [pointing at the mirror] is not wearing a business suit.

On Kaplan's view, the content of Fred's utterance of (11) is a proposition that contains Barney as a constituent twice over. It is the proposition that Barney is wearing a business suit, but Barney is not wearing a business suit. Thus, on Kaplan's theory, Fred believes a contradictory proposition, one whose immediate constituents are a proposition and its negation. But Fred (let us suppose) is a perfectly rational person. So he would never believe an outright contradiction. Therefore, many philosophers conclude, Kaplan's theory of indexicals is incorrect.

4.3 Some Alternatives to Kaplan's Theory

The objection is a variant on Gottlob Frege's puzzle of cognitive significance (see also the entries on the sense/reference distinction, and propositional attitude reports). Frege (1892) uses puzzles like this to motivate his semantic theory, and so we might first look to him to find a solution to the problem. Frege (1984) contains a brief discussion of indexicals. One view that can be extracted from this article, together with Frege (1892), is the following. Utterances of expressions have both a referent and a *sense*. The sense of an utterance of a full sentence is a *thought*. A person who sincerely and assertively utters a sentence asserts and believes the thought (sense) expressed by that utterance. (Thus, sense plays roughly the same role in this Fregean theory that content does in Kaplan's.) The sense of an utterance of an indexical, like ‘you’, can also be expressed by an utterance of a definite description *that contains no indexicals*. The relevant definite description is one whose utterance would be cognitively equivalent, for the speaker, to the utterance of the indexical. The referent of the utterance of the indexical is the referent of the relevant definite description.

It is controversial whether Frege really held this theory about indexicals, especially the thesis that utterances of indexicals have senses that can be expressed by non-indexical definite descriptions. Nonetheless, let us call it *Frege's theory*

On Frege's theory, the sense of Fred's utterance of ‘you’ might also be expressed (roughly) by the definite description ‘the person wearing a business suit’. The sense of Fred's utterance of ‘he’ might be expressed (roughly) by ‘the person wearing swimming trunks’. Thus the thought expressed by Fred's utterance of the full sentence in (11) might also be expressed by ‘the person wearing a business suit is wearing a

business suit, but the person wearing swimming trunks is not'. The thought expressed by the latter sentence is not contradictory. Therefore, Fred does not believe a contradiction, according to Frege's theory. Thus Frege's theory seems to solve the apparent problem with Kaplan's theory.

But Perry (1977) and Kaplan (1989a) have argued that there are problems with Frege's own theory. (For related discussion, see also Burks 1949 and Castaneda 1966, 1967.) Suppose that Fred sincerely utters 'Today is July 4, 2001' on July 3, 2001. According to Frege's theory, there is some non-indexical description that captures the sense of Fred's utterance of 'today'. But, as Perry (1977) points out, Fred may find it very difficult to produce a non-indexical description that he would be willing to substitute for his utterance of 'today' and that uniquely picks out one day. (In fact, in the previous paragraph the definite descriptions that we substituted for Fred's utterances of 'you' and 'he' almost certainly do not have unique referents.) Moreover, Perry says, it's possible that any such description that Fred would provide would not refer to the same day as his utterance of 'today'. For instance, Fred might be inclined to express the sense of his utterance of 'today' with the description 'the day in 2001 during which Americans celebrate the signing of the Declaration of Independence'. If so, then Frege's theory entails that Fred's utterance of 'today' on July 3, 2001 refers to July 4, 2001. So Frege's theory entails that Fred's utterance of 'Today is July 4, 2001' is true, and that he asserts and believes a truth. But Fred's utterance is false (no matter how he is inclined to describe the day on which he produces his utterance), and he asserts and believes a falsehood. So Frege's theory is incorrect. Kaplan raises a related *modal* problem for Frege's view. Suppose that Fred utters 'If I exist, then I am speaking', and suppose that the sense of his utterance of 'I' can be expressed by 'the person speaking'. Then on Frege's theory, Fred's utterance expresses the same sense as 'If the person speaking exists, then the person speaking is speaking'. This expresses a necessary truth, but Fred's utterance clearly does not.

One further apparent difficulty with Frege's view is that two utterances containing indexicals would rarely, if ever, express the same sense (or have the same content, to use our previous terminology). Consider, for example, (12) and (13).

12. Fred: "I am hungry."
13. Wilma: "You are hungry" [addressing Fred].

We are inclined to say that Fred's and Wilma's utterances say the same thing (in some sense). Furthermore, we think that Fred and Wilma (in some sense) assert and believe the same thing. Kaplan's theory validates these judgments, for their utterances have the same Kaplanian content, namely the singular proposition that Fred is hungry, and Fred and Wilma assert and believe that proposition. But it's highly unlikely that Fred's utterance and Wilma's utterance would express the same descriptive Fregean sense.

Schiffer (1978, 1981) avoids some of the problems with Frege's theory by allowing the relevant definite descriptions to contain the indexicals 'I' and 'now'. For instance, on Schiffer's theory, Fred's utterance of 'you' in (11) might be cognitively equivalent, for him, to the description 'the person whom I am now addressing', while his utterance of 'he' might be cognitively equivalent to 'the person whom I am now viewing in a mirror'. If so, then, on Schiffer's theory, Fred believes a singular proposition, but one that is

not contradictory. Such a theory may avoid Perry's "wrong referent" objection, for the "I-now" description that Fred would use to replace ‘today’ would probably be something like ‘the day during which I am now speaking’, which would pick out the correct day. Kaplan's objection concerning ‘I’ and modality would be avoided, because an utterance of ‘I’ would have the speaker as its content. Schiffer avoids the problem of two utterances' never having the same content by distinguishing between the semantic content of an utterance and the proposition that the speaker believes. On Schiffer's view, Fred's utterance and Wilma's utterance have the same semantic content, namely the singular proposition that Fred is hungry. But Schiffer denies that Fred and Wilma believe the same singular proposition.

Schiffer, however, does not avoid all of the apparent problems raised above. On his view, two people will only very rarely believe the same proposition containing a contingent individual as a constituent. Moreover, Schiffer denies that we typically believe what we say; that is, we often do not believe the propositions that are the semantic contents of our utterances. Finally, Schiffer's view entails that certain utterances that seemingly express the speaker's contingently true beliefs actually express necessarily true beliefs. For instance, if the belief content that Fred expresses by uttering ‘you’ in (11) can be expressed by ‘the person whom I am now addressing’, then an utterance by Fred of ‘If you exist, then you are the person whom I am now addressing’ would express his belief in a necessarily true proposition; but it seems that he is expressing his belief in a contingent proposition. Austin (1990) presents further criticisms of this view.

Lewis (1979a) and Chisholm (1981) use examples like (11) to argue that people believe *properties* rather than propositions. These properties are roughly expressible by phrases or sentences containing ‘I’ and ‘now’. For example, according to Lewis and Chisholm, Fred in example (11) believes a property which might be expressed as follows: being a thing that addresses exactly one person, who is wearing business suit, and that views through a mirror exactly one person, who is not wearing a business suit. We might use ‘I’ and ‘now’ to express this property as follows: I am now addressing exactly one person, who is wearing a business suit, and viewing exactly one person through a mirror, who is not wearing a business suit. In many respects, Lewis's and Chisholm's theories of indexical belief are very similar to Schiffer's theory. (Their theories of *meaning* might not be alike: Chisholm, unlike Schiffer, holds that such properties are not only the objects of belief, but also the contents of utterances.) Thus Chisholm's and Lewis's theories are subject to some of the same difficulties as Schiffer's, as Austin (1990) points out.[14]

Evans (1981) proposes that the content of an utterance of an indexical consists (roughly) of the speaker, or the time of utterance, together with a relation that resembles its Kaplanian character. On this view, Fred's utterance of ‘you’ in (11) is (roughly representable by) an ordered pair whose first member is Fred and whose second member is (roughly) the relation *addressing*. Together, the members of this pair gives us a property, being-a-thing-that-Fred-addresses, that picks out Barney. The content of Fred's utterance of ‘he’ in (11) consists of Fred plus (roughly) the relation *demonstrating*. This pair also determines a property that picks out Barney. Nevertheless, on Evans's view, Fred's utterance does not express a contradictory proposition. In some important respects, Evans's proposal resembles that of Schiffer, Lewis, and Chisholm, and may be subject to some of the same difficulties. For instance, on Evans's view there is no sense in which Fred and Wilma say or believe the same thing in examples (12) and (13). Evans's theory might also be vulnerable to modal objections similar to those that seemingly afflict these other theories, if

the theory is filled out in certain natural ways.

Peacocke (1992) proposes a neo-Fregean theory according to which senses, or *concepts (*as he calls them), are primitive, irreducible entities. However, each concept can be uniquely identified by describing the conditions under which a thinker grasps that concept (that is, by describing conditions under which a person can entertain a proposition that contains that concept as a constituent). The indexical concepts fall into types, e.g., first-person concepts and perceptual demonstrative concepts. For instance, in example (11), Fred grasps both a 'you' concept and a 'he' concept. (If Fred were in a perceptually identical situation, but looking at Barney's twin rather than Barney, then Fred would be grasping different 'you' and 'he' concepts, though they would be similar in type.) We can uniquely describe the concept Fred grasps by saying what it takes for a person to grasp that concept (we will not attempt to describe those conditions here). Peacocke's theory avoids many of the objections to Frege's theory. However, Peacocke's theory entails that Fred and Wilma believe different propositions in example (12)-(13). It's not entirely clear whether Peacocke's theory is vulnerable to modal objections.

4.4 Kaplanian Responses to the Criticism

The above survey of alternatives to Kaplan's theory was initially motivated by an apparent problem that Kaplan's theory has with cognitive significance and belief. Let's now consider how adherents to Kaplan's theory have responded to this problem. Consider once again the example of Fred in (11).

11. You[addressing Barney] are wearing a business suit, but he [pointing at the mirror] is not.

The difficulty was that, on Kaplan's theory, Fred seems to assert and believe a contradictory singular proposition, one that contains the singular proposition that Barney is wearing a business suit, and the negation of that very same proposition.

In reply, Kaplan (1989a) and Perry (1979) hold that a singular proposition can be entertained and believed *in different ways*. These ways of entertaining and believing a proposition correspond to characters. Thus Kaplan says that an agent can believe a proposition *under* one character, but fail to believe it under another character. For instance, Fred in (11) believes the proposition that Barney is wearing a business suit *under* the character of 'You are wearing a business suit', but fails to believe that proposition *under* the character of 'He is wearing a business suit'; in fact, Fred believes the negation of that proposition under the character of 'He is *not* wearing a business suit'. On Kaplan's view, Fred really does believe a contradictory proposition, but he believes the conjuncts of this proposition under suitably different characters, which is why he still counts as rational.

Perry (1979) elaborates on this view by distinguishing between (i) the proposition that an agent believes and (ii) the *belief state* in virtue of which the agent believes that proposition. For any one proposition, there are many belief states that would enable an agent to believe that proposition. An agent can be in one of these belief states, while failing to be in another. Belief states can be classified into types according to the characters of the sentences that they dispose the agent to assert. For instance, a belief state that would

cause an agent to sincerely assert ‘You are wearing a business suit’ is distinct from a belief state that would cause him to sincerely assert ‘He is wearing a business suit’, because the two sentences differ in character. A rational agent could be in a belief state of the first type without being in a belief state of the second type. In fact, a rational agent could be in the first sort of belief state, while also being in a second sort that causes him to utter ‘He is *not* wearing a business suit’. Thus a rational agent could believe a proposition and its negation, as long as he does so by being in suitably different belief states.

Kaplan's response leaves us with the question "What does it mean to believe a content *under* a character?" Perry's postulation of belief states is a first step towards answering the question, but his work leaves the nature of belief states relatively unclear. Both Kaplan and Perry think that ways of believing either are, or correspond one-to-one with, characters. This assumption is problematic. We can easily imagine that Fred, while viewing Barney, says ‘He [pointing directly at Barney] is wearing a business suit, but he [pointing at the mirror] is not’. If the two occurrences of ‘he’ have the same character, then, on Kaplan's and Perry's view, Fred rationally believes a proposition under one character and the negation of that same proposition under the "negation" of that same character. This is exactly the sort of situation that Kaplan's and Perry's view should disallow. For further discussion and criticisms, see Wettstein (1986), Taschek (1987), and Crimmins (1992, chapter 1). For elaborations on, and modifications of, the Kaplan-Perry theory, see Perry (1997). See also section 5.2 for related issues.

5. Other Topics Concerning Indexicals

In this section, we discuss a variety of additional issues about indexicals.

5.1 Complex Demonstratives

Complex demonstratives are expressions of the form *that N* or *this N*, where *N* is a common noun phrase. Examples include ‘this dog’, ‘that red car’, and ‘that woman who is standing by the door’. Complex demonstratives raise at least two interesting questions. First, do the common noun phrases that appear in them play some role in determining their referents? For example, must a person be a crook in order to be the referent of an utterance of ‘that crook’? Second, what do these common noun phrases contribute to the contents of complex demonstratives? For instance, does an utterance of ‘That crook is untrustworthy’ express a proposition that has the property of being-a-crook as a constituent? Or does the phrase ‘that crook’ contribute only its referent to the proposition expressed?

Many (though not all) theories of complex demonstratives can be classed into three types, according to how strong a semantic role they attribute to the common noun phrases. The first type says that the common noun phrase in a complex demonstrative plays *no* semantic role in determining the referent of the complex demonstrative; so a person could be the referent of an utterance of ‘that crook’ even if she is not a crook. Furthermore, this type of view says that the content of the common noun phrase is *not* a constituent of the content of the complex demonstrative; the content of an utterance of a complex demonstrative is just its referent. The common noun phrase serves merely as a pragmatic cue that helps the speaker's audience figure out what the speaker's intended referent is. Call theories of this sort *minimal*

theories, because they assign a minimal, or non-existent, semantic role to the common noun phrase. Larson and Segal (1995) endorse (roughly) a minimal theory (they do not, however, accept the existence of Kaplan-style contents).

Theories of the second type say that the common noun phrase *does* help determine the referent, but say that its content does *not* appear as a constituent of the content of the complex demonstrative. On such views, a person must be a crook in order to be the referent of an utterance of ‘that crook’, but the property of being-a-crook is not a constituent of the complex demonstrative's content. The content of the complex demonstrative is just its referent. Kaplan (1989a) does not explicitly endorse this type of theory, but his incidental remarks about complex demonstratives suggest that he would favor it. Braun (1994) and Borg (2000) argue explicitly for this sort of view. Call theories of this sort *intermediate theories*.

Theories of the third type say that the common noun phrase helps determine the referent; moreover, the content of the common noun phrase appears as a constituent of the content of the complex demonstrative. Richard (1993) and King (2000) endorse views of this sort. (King holds, in addition, that complex demonstratives are quantifier phrases that are comparable, in many respects, to definite descriptions.) Let's call theories of this sort *maximal theories*.

Advocates of minimal theories describe cases in which a speaker apparently refers to a person with an utterance of ‘that crook’, even though the person is not a crook. But advocates of intermediate and maximal theories reply by distinguishing between *speaker referent* and *semantic referent* (see Kripke 1977): they say that, in such cases, the *speaker* refers to the non-crookish person, but the speaker's *utterance* does not semantically refer to that person. Some advocates of maximal theories claim that sentences like (14) are logically or analytically true, whereas (they claim) this would not be so on minimal and intermediate theories, because those theories say that the content of the common noun phrase is not part of the content of the proposition expressed.

14. If that crook exists, then that crook is a crook.

Advocates of intermediate views respond by admitting that (14) is true in all contexts (ignoring reference failure). But (they claim) this is not because the content of ‘crook’ is part of the content of the complex demonstrative. Rather, (14) is true in all contexts because ‘that crook’ refers in a context to a person only if the relevant person is a crook in the world of that context. Therefore, if the antecedent of (14) is true in a context, then so is its consequent. Intermediate theorists point out that (15) seems to be false in many contexts, which seems contrary to maximal theories.

15. Necessarily: if that crook exists, then that crook is a crook.

Maximal theorists, however, sometimes add rigidifying devices to the contents of complex demonstratives to avoid this difficulty. They furthermore point to cases of apparent quantification into complex demonstratives, which seem to present a difficulty for intermediate theories.

Obviously, the issues and arguments here are quite complicated. Moreover, intuitions about the relevant cases are often unstable. For further discussion, see the works mentioned above. See also Dever (2001) and Lepore and Ludwig (2000), whose views do not fit easily into the above classification scheme.

5.2 Multiple Occurrences of Demonstratives

Kaplan's theory has an apparent problem with sentences that contain multiple occurrences of the same demonstrative. The problem can be presented as an argument from Kaplanian premises to apparently false conclusions, as follows. The demonstrative ‘that’ has a single linguistic meaning. So every occurrence of ‘that’ has the same linguistic meaning. If Kaplan's theory is correct, then the linguistic meaning of an expression is its character. Thus, if Kaplan's theory is correct, then every occurrence of ‘that’ has the same character. Thus both occurrences of ‘that’ in sentence (16) have the same character.

16. That is not identical with that.

But if the two occurrences of ‘that’ in (16) have the same character, then they have the same referent and content in every context. Thus, (16) is false in every context.[15] But this is surely incorrect. If we extend Kaplan's theory to utterances in the natural way suggested in section 3.2, we get the clearly absurd result that no utterance of (16) is true.

Kaplan seems to try to escape this consequence by, in effect, making the word ‘that’ ambiguous. On one version of his theory, the single word ‘that’ is replaced with the subscripted expressions ‘that_1’, ‘that_2’, etc. The content of ‘that_1’ in a context is the first demonstratum of the context, the content of ‘that_2’ is the second demonstratum, and so on. Thus (17) is true in contexts in which the first demonstratum is distinct from the second.

17. That_1 is not identical with that_2.

However, each subscripted ‘that’ has a different character (which is perhaps why Kaplan [1989b] sometimes speaks of the "exotic ambiguity" of demonstratives). None of these characters is a reasonable candidate for being *the* linguistic meaning of the (unsubscripted) English word ‘that’. So the subscript-maneuver undercuts Kaplan's identification of linguistic meaning with character.[16]

There are several ways in which one might try to deal with this problem. One is to allow shifts in context in mid-sentence. We could allow the content of ‘that’ in a context to be *the* demonstratum of the context. We could then suppose that the two occurrences of ‘that’ in (16) are (always or sometimes) evaluated for content in *different* contexts that may have different demonstrata. Unfortunately, this option may wreak havoc with the logic of demonstratives; it may, for instance, undermine attempts to allow argument (9) (in section 3.3) to be valid. A second option is to give up the identification of linguistic meaning with character, and suppose instead that linguistic meaning is a third sort of meaning, distinct from both character and content. A third option is to move away from a Kaplan-style theory, which assigns contents

to expressions-in-contexts, and move to a theory that assigns contents to utterances. This might help because every *utterance* of (16) contains *two utterances* of ‘that’. There are almost certainly other options. For further discussion, see Kaplan (1989b, 1999), Wettstein (1986), Braun (1996), Garcia-Carpintero (1998), Richard (1992, forthcoming), Forbes (forthcoming), and section 5.3 below.

5.3 Utterance Theories vs. Expression Theories

As mentioned earlier, an *utterance* is an action (a sort of event) in which an agent utters an expression. Kaplan's theory does not, strictly speaking, assign contents to utterances. Rather, it assigns contents to *expressions-in-contexts*, which we can think of as pairs of expressions and contexts. There are many expressions-in-contexts for which there is no corresponding utterance. For instance, consider a context in which Fred utters ‘I am wearing a business suit’, and nothing else, at a particular time, place, and world. Even though Fred does *not* utter ‘I am hungry’ at that time, place, and world, there is still an expression-in-a-context consisting of that sentence together with a context which has Fred as its agent, and which has that time, location, and world associated with it. Kaplan's theory assigns a content to this expression-in-a-context (namely, the proposition that Fred is hungry) even though Fred does not utter the expression at that time, place, and world. That content even has a truth value in that context. Thus Kaplan's theory of indexicals "abstracts away" from utterances to a remarkable degree.

Kaplan (1989a, 1989b) claims that his abstract, *expression-based*, theory of indexicals has various advantages, especially when it comes to logic. First, Kaplan claims that his expression-theory classifies as logical truths all sentences that should be logical truths, whereas utterance-theories do not. Consider, for example, sentences of the form *If P then P*. All such sentences are logical truths, Kaplan claims, and are so-classified on Kaplan's theory. But now consider a sentence of this form in which P is an extremely long sentence containing ‘today’. Then there may be utterances of this sentence in which the utterance of the antecedent is true, while the utterance of the consequent is false, simply because the first utterance of P has taken more than one day to complete. If a sentence is logically true iff all of its *utterances* are true, then this sentence is *not* a logical truth, even though it has the form *If P then P*. But this seems incorrect. Second, Kaplan claims that his approach correctly labels some sentences as *not* logically true which would be incorrectly labeled as logically true on utterance-based theories. For instance, on Kaplan's abstract approach, the sentence ‘I am uttering something’ is *not* a logical truth, because it is false with respect to contexts in which the agent of the context is silent. But all *utterances* of this sentence are true, so if logical truth were taken to be truth-of-all-utterances, then this sentence would be (incorrectly) classified as a logical truth.

However, there also seem to be disadvantages to Kaplan's expression-based theory. As Kaplan (1989b) notes, there are serious difficulties in assigning contents to *true demonstratives* with respect to contexts in which the agent does not have any relevant directing intentions and produces no demonstrations. Suppose, for example, that when Fred utters ‘I am wearing a business suit’ at time T, location L, and world W, he does *not* point at any object and does *not* intend to refer demonstratively to any object. Then it will be very difficult to assign a content to ‘That is a dog’ with respect to a context that has Fred as its agent, and which has T, L, and W as its associated time, place, and world. Moreover, Kaplan's apparent problem with multiple occurrences of demonstratives is at least partly due to his assigning contents to expressions-

in-contexts.

Other theorists prefer more utterance-oriented approaches to indexicals: see, for instance, Reichenbach (1947), Burks (1949), Weinstein (1974), and Barwise and Perry (1983). Perry (1997) argues for the advantages of theories that focus primarily on utterances. Kaplan himself (Kaplan 1999) has recently taken a more positive view of utterance-based theories. For further discussion, see Braun (1996), Garcia-Carpintero (1998) and Forbes (forthcoming).

5.4 Variables, Anaphora, and Demonstratives

As mentioned in section 1.2, many of the expressions that can be used as demonstratives, such as ‘he’ and ‘she’, can also be used anaphorically and in ways similar to bound variables in formal languages. This presents a difficult challenge to semanticists. On the one hand, there are significant semantic differences between variables, anaphora, and demonstratives: for instance, the content of a demonstrative is determined by a speaker's intentions or demonstrations, while the content of a (free) variable is determined arbitrarily by an assignment of values to variables. But, on the other hand, it seems that ‘she’ is an *un*ambiguous expression, which has a single linguistic meaning. The challenge for semanticists is to describe the linguistic meaning of ‘she’ without slighting the significant semantic differences between its various use. (Similarly for ‘he’, ‘his’, ‘her’, and so on.)

Many theorists have discussed aspects of this problem, but it is unclear whether any have fully met the challenge. Kaplan (1989a, 1989b) emphasizes the semantic similarities between (on the one hand) free variables under assignments and (on the other hand) demonstratives in contexts. But he backs away from identifying the linguistic meanings of ‘he’ and ‘she’ with the linguistic meanings of variables (with appropriately restricted ranges); he even entertains the possibility that there are two homonymous expressions ‘he’, one of which is a demonstrative and the other of which is a variable. Other theorists (for instance, Heim and Kratzer 1998, chapter 9) assimilate demonstratives to free variables. Those who work on discourse representation theory and dynamic semantics emphasize the similarities between variables and some anaphoric uses of pronouns. (See, for example, Kamp and Reyle 1993 on discourse representation theory, and Chierchia 1992 on dynamic semantics; see also the entry on anaphora.) For related discussion, see Partee (1989) and Condoravdi and Gawron (1996).

Bibliography

For useful overviews, see Forbes (1989, forthcoming), Perry (1997), and Richard (1992, forthcoming).

- Almog, Joseph, Perry, John and Wettstein, Howard. 1989. *Themes from Kaplan*. Oxford: Oxford University Press.
- Austin, David. 1990. *What's the Meaning of "This"?* Ithaca, NY: Cornell University Press.
- Bach, Kent. 1992a. "Intentions and Demonstrations." *Analysis* 52, pp. 140-146.
- Bach, Kent. 1992b. "Paving the Road to Reference." *Philosophical Studies* 67, pp. 295-300.

- Barwise, Jon and Perry, John. 1983. *Situations and Attitudes*. Cambridge, MA: MIT Press.
- Borg, Emma. 2000. "Complex Demonstratives." *Philosophical Studies* 97, pp. 225-244.
- Braun, David. 1994. "Structured Characters and Complex Demonstratives." *Philosophical Studies* 74, pp. 193-219.
- Braun, David. 1995. "What Is Character?" *Journal of Philosophical Logic* 24, pp. 227-240.
- Braun, David. 1996. "Demonstratives and Their Linguistic Meanings." *Nous* 30, pp. 145-173.
- Burge, Tyler. 1974. "Demonstrative Constructions, Reference, and Truth." *Journal of Philosophy* 71, pp. 205-223.
- Burks, Arthur. 1949. "Icon, Index, and Symbol." *Philosophy and Phenomenological Research* 9, pp. 673-89.
- Castañeda, Hector-Neri. 1966. "'He': A Study in the Logic of Self-Consciousness." *Ratio* 8, pp. 130-157.
- Castañeda, Hector-Neri. 1967. "Indicators and Quasi-Indicators." *American Philosophical Quarterly* 4, pp. 85-100.
- Chierchia, Gennaro. 1992. "Anaphora and Dynamic Binding." *Linguistics and Philosophy* 15, pp. 111-183.
- Chisholm, Roderick. 1981. *The First Person: An Essay on Reference and Intentionality*. Minneapolis: University of Minnesota Press.
- Cohen, Stewart. 1988. "How to Be a Fallibilist." *Philosophical Perspectives* 2, pp. 91-123.
- Condoravdi, Cleo and Gawron, Jean Mark. 1996. "The Context Dependency of Implicit Arguments." In Makoto Kanazawa, Christopher Pinon, and Henriette de Swart (Eds.), *Quantifiers, Deduction, and Context*. Stanford: CSLI.
- Crimmins, Mark. 1992. *Talk About Beliefs*. Cambridge, MA: MIT Press.
- DeRose, Keith. 1995. "Solving the Skeptical Problem." *Philosophical Review* 104, pp. 1-52.
- Dever, Joshua. 2001. "Complex Demonstratives." *Linguistics and Philosophy* 24, pp. 271-330.
- Devitt, Michael. 1981. *Designation*. New York: Columbia University Press.
- Evans, Gareth. 1981. "Understanding Demonstratives." In H. Parret and Jacques Bouveresse (Eds.), *Meaning and Understanding*. Berlin: de Gruyter. Reprinted in Gareth Evans, 1985, *Collected Papers*. Oxford: Oxford University Press.
- Fillmore, Charles. 1972. "How to Know Whether You Are Coming or Going." In Karl Hyldgaard-Jensen (Ed.), *Linguistik 1971*. Konigstein, Germany: Athenaum-Verlag.
- Fillmore, Charles. 1975. *Santa Cruz Lectures on Deixis*. Distributed by Indiana University Linguistics Club, Bloomington, IN.
- Forbes, Graeme. 1989. "Indexicals." In D. Gabbay and F. Guenther (Eds.), *Handbook of Philosophical Logic, Volume IV*, pp. 463-490. Dordrecht: Reidel.
- Forbes, Graeme. Forthcoming. "Indexicals" (revised version). In D. Gabbay and F. Guenther (Eds.), *Handbook of Philosophical Logic, 2nd Edition, Volume 9.* Dordrect: Kluwer.
- Frege, Gottlob. 1892. "Uber Sinn und Bedeutung." *Zeitschrfit fur Philosophie und Philosophische Kritik* 100. Translation by Herbert Feigl as "Sense and Nominatum" in Herbert Feigl and Wilfrid Sellars (Eds.), 1949, *Readings in Philosphical Analysis*, pp. 85-102. New York: Appleton-Century-Crofts. This translation is reprinted in Martinich 2001, pp. 199-211. Also translated by Peter Geach and Max Black as "Sense and Reference" in Peter Geach and Max Black (Eds.), 1952, *Translations from the Philosophical Writing of Gottlob Frege*, pp. 56-78, Oxford: Blackwell.

- Frege, Gottlob. 1984. "Thoughts." In Frege (ed. B. McGuiness, trans. P. Geach and R.H. Stoothoff), *Collected Papers on Mathematics, Logic, and Philosophy*, pp. 351-72, Oxford: Blackwell.
- Garcia-Carpintero, Manuel. 1998. "Indexicals as Token-Reflexives." *Mind* 107, pp. 529-563.
- Heim, Irene and Kratzer, Angelika. 1998. *Semantics in Generative Grammar*. Oxford: Blackwell.
- Kamp, Hans. 1971. "Formal Properties of 'Now'." *Theoria* 37, pp. 227-273.
- Kamp, Hans and Reyle, Uwe. 1993. *From Discourse to Logic*. Dordrecht: Kluwer.
- Kaplan, David. 1989a. "Demonstratives." In Almog, Perry, and Wettstein 1989, pp. 481-563.
- Kaplan, David. 1989b. "Afterthoughts." In Almog, Perry, and Wettstein 1989, pp. 565-614.
- Kaplan, David. 1999. "Reichenbach's *Elements of Symbolic Logic*." German translation in Maria Reichenbach and Andreas Kamlah (Eds.), *Hans Reichenbach, Collected Works in 9 Volumes*. Frieder.: Vieweg & Sohn.
- King, Jeffrey. 2001. *Complex Demonstratives: A Quantificational Approach*. Cambridge, MA: MIT Press.
- Kratzer, Angelika. 1977. "What 'Must' and 'Can' Must and Can Mean." *Linguistics and Philosophy* 1, pp. 337-355.
- Kripke, Saul. 1977. "Speaker's Reference and Semantic Reference." In Peter French, Theodore Uehling, Jr., and Howard Wettstein (Eds.), *Contemporary Perspectives in the Philosophy of Language*, pp. 6-27. Minneapolis: University of Minnesota Press.
- Kripke, Saul. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
- Larson, Richard and Segal, Gabriel. 1995. *Knowledge of Meaning*. Cambridge, MA: MIT Press.
- Lepore, Ernest and Ludwig, Kirk. 2000. "The Semantics and Pragmatics of Complex Demonstratives." *Mind* 109, pp. 199-240.
- Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell.
- Lewis, David. 1979a. "Attitudes *De Dicto* and *De Se*." *Philosophical Review* 88, pp. 513-43. Reprinted in Lewis1983.
- Lewis, David. 1979b. "Scorekeeping in a Language Game." *Journal of Philosophical Logic* 8, pp. 339-59. Reprinted in Lewis 1983.
- Lewis, David. 1980. "Index, Context, and Content." In Stig Kanger and Sven Ohman (Eds.), *Philosophy and Grammar*. Dordrecht: Reidel. Reprinted in David Lewis, 1998, *Papers in Philosophical Logic*, Cambridge: Cambridge University Press.
- Lewis, David. 1983. *Philosophical Papers Volume I*. Oxford: Oxford University Press.
- Lewis, David. 1996. "Elusive Knowledge." *Australasian Journal of Philosophy* 74, pp. 549-567. Reprinted in David Lewis, 1999, *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.
- Martinich, A.P. (Ed.). 2001. *Philosophy of Language, 4th Edition*. Oxford: Oxford University Press.
- McGinn, Colin. 1981. "The Mechanism of Reference." *Synthese* 49, pp. 157-186.
- Nunberg, Geoffrey. 1993. "Indexicality and Deixis." *Linguistics and Philosophy* 16, pp. 1-43.
- Partee, Barbara. 1973. "Some Structural Analogies Between Tenses and Pronouns in English." *Journal of Philosophy* 70, pp. 601-609.
- Partee, Barbara. 1989. "Binding Implicit Variables in Quantified Contexts." *Papers of the Chicago Linguistic Society* 25, pp. 342-365.

- Peacocke, Christopher. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.
- Perry, John. 1977. "Frege on Demonstratives." *Philosophical Review* 86, pp. 474-97. Reprinted in Perry 1993.
- Perry, John. 1979. "The Problem of the Essential Indexical." *Nous* 13, pp. 3-21. Reprinted in Martinich 2001 and Perry 1993.
- Perry, John. 1993. *The Problem of the Essential Indexical*. Oxford: Oxford University Press.
- Perry, John. 1997. "Indexicals and Demonstratives." In Bob Hale and Crispin Wright (Eds.), *A Companion to Philosophy of Language*, pp. 586-612. Oxford: Blackwell. [Available online in PDF.]
- Reichenbach, Hans. 1947. *Elements of Symbolic Logic*. NewYork: Macmillan.
- Reimer, Marga. 1991a. "Demonstratives, Demonstrations, and Demonstrata." *Philosophical Studies* 63, pp. 187-202.
- Reimer, Marga. 1991b. "Do Demonstrations Have Semantic Significance?" *Analysis* 51, pp. 177-183.
- Richard, Mark. 1990. *Propositional Attitudes*. Cambridge: Cambridge University Press.
- Richard, Mark. 1992. "Indexicals." In William Bright (Ed.), *International Encyclopedia of Linguistics*, pp. 200-202. Oxford: Oxford University Press.
- Richard, Mark. 1993. "Articulated Terms." *Philosophical Perspectives* 7, pp. 207-230.
- Richard, Mark. Forthcoming. "Indexicals" (revised version). In William Bright (Ed.), *International Encylopedia of Linguistics, 2nd Edition.* Oxford: Oxford University Press.
- Salmon, Nathan. 1981. *Reference and Essence*. Princeton, NJ: Princeton University Press.
- Salmon, Nathan. 1989. "Tense and Singular Propositions." In Almog, Perry, and Wettstein 1989, pp. 331-392.
- Salmon, Nathan. 1991. "How *Not* to Become a Millian Heir." *Philosophical Studies* 62, pp. 165-177.
- Schiffer, Stephen. 1978. "The Basis of Reference." *Erkenntnis* 13, pp. 171-206.
- Schiffer, Stephen. 1981. "Indexicals and the Theory of Reference." *Synthese* 49, pp. 43-100.
- Smith, Quentin. 1989. "The Multiple Uses of Indexicals." *Synthese* 78, pp. 167-191.
- Soames, Scott. 1999. *Understanding Truth*. Oxford: Oxford University Press.
- Stalnaker, Robert. 1981. "Indexical Belief." *Synthese* 49, pp. 129-151.
- Stanley, Jason and Szabo, Zoltan. 2000. "On Quantifier Domain Restriction." *Mind & Language* 15, pp. 219-261.
- Taschek, William. 1987. "Content, Character, and Cognitive Significance." *Philosophical Studies* 52, pp. 161-189.
- Vision, Gerald. 1985. "'I Am Here Now'." *Analysis* 45, pp. 198-199.
- Vlach, Frank. 1973. "'Now' and 'Then': A Formal Study in the Logic of Tense and Anaphora." Ph.D. Dissertation, UCLA.
- Weinstein, Scott. 1974. "Truth and Demonstratives." *Nous* 8, pp. 179-184.
- Wettstein, Howard. 1984. "How to Bridge the Gap Between Meaning and Reference." *Synthese* 84, pp. 63-84. Reprinted in Wettstein 1991.
- Wettstein, Howard. 1986. "Has Semantics Rested on a Mistake?" *Journal of Philosophy* 83, pp. 185-209. Reprinted in Wettstein 1991.
- Wettstein, Howard. 1991. *Has Semantics Rested on a Mistake?* Stanford: Stanford University

Press

Other Internet Resources

- Other papers by John Perry (Stanford University) on indexicals (in PDF):
    - Indexicals, Contexts and Unarticulated Constituents
    - Indexicals
    - Reflexivity, Indexicality and Names

Related Entries

anaphora | Frege, Gottlob | propositional attitude reports | propositions: singular | propositions: structured | reference | sense/reference distinction

David Braun
dbrn@troi.cc.rochester.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 14, 2001
Content last modified: September 14, 2001

Stanford Encyclopedia of Philosophy
Notes to Indexicals

Notes

1. Smith 1989 says that some uses of ‘now’ do not refer to a time interval that includes the moment of utterance. For instance, when narrating the life of George Washington, one might say ‘Washington now needed to get across the Delaware River’.

2. Kaplan 1989a attributes to Michael Bennett the observation that some uses of ‘here’ are demonstrative, as when a person points at a map and says ‘Next week, I'll be *here*’.

3. Kaplan's manuscript "Demonstratives" (Kaplan 1989a) was written in the 1970's and circulated informally for many years before being published. Kaplan's "Afterthoughts" (Kaplan 1989b) was written considerably later, even though it has the same official publication date as "Demonstratives."

4. Strictly speaking, Kaplan (1989a) associates a *circumstance of evaluation* with each context, rather than just a possible world. A circumstance of evaluation is a pair of a possible world and a time. We shall ignore this refinement in the presentation that follows.

5. In his informal presentations of his theory, Kaplan (1989a, 1989b) identifies contents with individuals, properties, relations, and structured propositions, as we have done here. But when he turns to technical matters, these contents are identified with (or represented by) *intensions*, which are functions from circumstances of evaluation to extensions. For details, see Kaplan 1989a and Forbes 1989, forthcoming. We shall concentrate here on the less formal aspects of Kaplan's theory.

6. Sometimes Kaplan describes the character of an expression as being a *rule* for associating contents of the expression with contexts. In his more technical presentations, Kaplan (1989a) seems to identify the character of an expression with a function different from that described in the main text. See Braun 1995.

7. More accurately, a logical truth is a sentence that is true in every context *in every structure* (or model). From here on, we ignore this qualification.

8. Three points: (i) Kaplan tends to restrict application of the term ‘directly referential’ to singular terms. (ii) Kaplan (1989a) holds that the notion of direct reference can be defined without appeal to the notion of a singular proposition, but he does not explicitly provide any such definition. (iii) In the preceding paragraph, we used the notion of the referent of an expression with respect to a (Kaplanian) context. See the next two paragraphs, and note 10, for a discussion of this notion.

9. In fact, it is somewhat difficult to determine exactly how Kripke (1980) wishes to define ‘rigid designator’. For discussion, see Kaplan 1989a, 1989b and Salmon 1981.

10. When we speak of the referent of a term with respect to context C, we mean its referent with respect to C *and* the world of C.

11. We can say more precisely what it means for an indexical to be a rigid designator, but we must first modify Kripke's original definition of ‘rigid designator’ so that it can be extended to context-sensitive expressions. We first need to assume (or define) the notion of the *referent of a singular term with respect to a context C and world W*. Then the following definition of rigid designation may do the trick, if we restrict our attention to singular terms. If C is a context, then expression *D* is a *rigid designator with respect to C* iff: D is a singular term, and for all worlds W and W*, the referent of D with respect to C and W is identical with the referent of D with respect to C and W*. Kaplan could then be construed as saying that all (simple, singular term) indexicals are rigid designators with respect to all contexts.

12. Not all criticisms of Kaplan's theory concern belief and cognitive significance. As mentioned in section 2, various philosophers disagree with Kaplan's account of reference-fixing for demonstratives. Salmon 1989 criticizes the theory's assumption that propositions can vary in truth value from time to time. Braun 1995 criticizes the theory's (apparent) identification of character with an extensional function. Braun 1996 criticizes the theory's handling of multiple occurrences of demonstratives (see section 5.2 below).

13. We are assuming here that Kaplan's theory assigns contents to *utterances*. So, strictly speaking, we are considering the natural extension of Kaplan's theory to utterances that was presented in section 3.2.

14. Stalnaker's theory of indexical belief (Stalnaker 1981) relies heavily on a technical apparatus that is too elaborate to present here. But it is similar to Schiffer's, Chisholm's and Lewis's theories in one respect: it tries to "reduce" indexical belief to a more restricted class of "singular belief". In particular, Stalnaker tries to reduce indexical belief to belief in propositions about particular utterances or thinking-events. If *u* is Fred's utterance of (11), then Stalnaker would say that Fred's utterance expresses his belief in (very roughly) the proposition that the person who utters *u* is addressing exactly one person, who is wearing a business suit, and viewing exactly one person through a mirror, who is not. See Stalnaker 1981 for details, and Austin 1990 for critical discussion.

15. We are ignoring here any difficulties raised by contexts in which ‘that’ fails to refer.

16. Kaplan 1989a presents a second theory of demonstratives that does not use subscripted *that*s. On this theory (which he seems to prefer), occurrences of ‘that’ in English are represented by *dthat-terms* of the form "dthat[the F]". For details, see Kaplan 1989a. It turns out, however, that two dthat-terms can have different contents in a single context only if they have different characters. So the same problem arises for the dthat-term theory as for the subscripted ‘that’ theory.

David Braun
dbrn@troi.cc.rochester.edu

First published: September 14, 2001
Content last modified: September 14, 2001

Structured Propositions

It is a truism that two speakers can say the same thing by uttering different sentences, whether in the same or different languages. For example, when a German speaker utters the sentence ‘Schnee ist weiss’ and an English speaker utters the sentence ‘Snow is white’, they have said the same thing by uttering the sentences they did. Proponents of propositions hold that, speaking strictly, when speakers say the same thing by means of different declarative sentences, there is some (non-linguistic) thing, *a proposition*, that each has said. This proposition is said to be *expressed* by both of the sentences uttered (taken in the contexts of utterance -- to accommodate contextually sensitive expressions) by the speakers, and can be thought of as the information content of the sentences (taken in those contexts). The proposition is taken to be the thing that is in the first instance true or false. A declarative sentence is true or false derivatively, in virtue of expressing (in the context in which it is uttered -- I shall henceforth ignore contextual sensitivity and so dispense with qualifications of this sort) a true or false proposition.

Propositions are thought to perform a number of other functions in addition to being the primary bearers of truth and falsity and the things expressed by declarative sentences. When a German and English speaker believe the same thing, say that the earth is round, the thing they both believe is not a sentence but a proposition. For the English speaker would express her belief by means of the sentence ‘The earth is round’ and the German speaker would express her belief by means of the *different* sentence ‘Die Erde ist rund’. Thus when people believe, doubt and know things, it is propositions that they bear these cognitive relations to. Finally, it is the proposition a sentence expresses, and not the sentence itself, that possesses modal properties such as being necessary, possible or contingent.

That propositions perform these various functions is agreed upon by virtually all advocates of propositions. [1] There is considerably less agreement concerning the nature of the things, propositions, that perform these functions.

To say that propositions are *structured* is to say something about the nature of propositions. Roughly, to say that propositions are structured is to say that they are complex entities, entities having parts or constituents, where the constituents are bound together in a certain way. Thus, particular accounts of structured propositions can (and do) differ in at least two ways: 1) they can differ as to what sorts of things are the constituents of structured propositions; and 2) they can differ as to what binds these constituents together in a proposition. We shall discuss various accounts of structured propositions that differ in these ways below.

Put this way, the view that propositions are structured is purely a metaphysical thesis about what

propositions are like and entails nothing about the relation between a sentence and the proposition it expresses. But of course structured proposition theorists do have views about the relation between a sentence and the proposition it expresses.

---

1. Setting Up the Problems

Intuitively, given that a sentence expresses a structured proposition, the proposition will have parts or constituents that are the semantic values of words or subsentential complex linguistic expressions occurring in the sentence; and the proposition will have a structure similar to the structure of the sentence. For example, assuming that the semantic value of a name is its bearer and that the semantic value of a transitive verb is a relation, a structured proposition theorist will likely hold that the sentence

(1) Jason loves Patty

expresses a proposition consisting of Jason, the loving relation and Patty, bound together in some way into a unity. Letting '*j*' stand for Jason, '*p*' for Patty and '*L*' for the loving relation, we can represent the proposition in question as follows:

(1a) $[j[L[p]]]$

Thus (1a)'s structure is very close to that of (1); and (1a) has as constituents the semantic values of the words occurring in (1). Indeed, in the case of (1) and (1a), all and only semantic values of words in the sentence are constituents of the proposition. But a given account of structured propositions may not hold that this is the case in general for any one of at least three reasons. First, one might hold that certain words as they occur in phrases in sentences do not contribute *their* semantic values to the propositions expressed by those sentences because the semantic values of these words instead partially determine the semantic values of the *phrases* in which they occur, where these latter semantic values are contributed to the proposition. For example, one might hold that in the sentence

(2) Colin is a tall young man.

the phrase ‘is a tall young man’ contributes to the proposition expressed by (2) the property of being a tall young man, which is its semantic value. Thus, though the semantic value of the word ‘tall’ partly determines the semantic value of the phrase ‘is a tall young man’, the proposition expressed by (2) contains no constituent that is the semantic value of the word ‘tall’ alone. Second, one might hold that a sentence may express a proposition (in a context), where the proposition has constituents not contributed by *any* syntactic constituent of the sentence, let alone any word in the sentence. For example, Mark Crimmins [1992] claims that an utterance of the sentence

(3) It's raining

expresses a proposition to the effect that it is raining at a particular time and place. The present tense manages to somehow contribute the time of utterance to the proposition. But no syntactic constituent of the sentence contributes the place to the propositon, though Crimmins claims it is a constituent of the proposition expressed.[2] Third, one might hold that certain words simply have no semantic values, and so make no contribution to propositions. So-called neoplanastic ‘ne’ in French might be thought to be an example of this.

But even if, for one of the above reasons or some other reason, a sentence does not express a propositon whose constituents are *precisely* the semantic values of *words* in the sentence, we can still say that structured proposition theorists hold that sentences expresses propositions, where *many* (and likely most) constituents of the proposition are semantic values of words *or phrases* occurring in the sentence. So in the case of (1) and (1a), the constituents of the proposition are precisely the semantic values of the words in (1). In the case of (2) and (2a) (given the assumptions made above), the constituents of (2a) are precisely the semantic values of the name ‘Colin’ and and the verb phrase ‘is a tall young man’. And in the case of (3), the proposition it expresses has three constituents, two of which are contributed by ‘raining’ and the present tense construction.

Thus, ignoring, or at least not dwelling on, the qualifications just made, we can say that structured proposition theorists hold that sentences express propositions that are complex entities (most of) whose constituents are the semantic values of expressions occurring in the sentence, where these constituents are bound together by some structure inducing bond that renders the structure of the proposition similar to the structure of the sentence expressing it.

This highlights an important feature of structured proposition accounts that distinguishes them from the other main competing account of propositons, namely the account of propositions as sets of possible worlds (to be discussed below). Because structured propositions have as parts the semantic values of expressions in the sentences expressing them, the semantic values of those expressions are *recoverable* from the semantic values of the sentences (i.e. the propositions).

It is perhaps worth noting that one could have a theory according to which the semantic values of

expressions in a sentence are recoverable from the proposition expressed by the sentence, even though the semantic values of the expressions are not (mereological or set theoretic) parts of the proposition. For example, George Bealer [1993] formulates what he calls an *algebraic* conception of propositions (see also Bealer [1982]). Bealer associates with each propositon a "decomposition tree." This decompositon tree shows how a given proposition is the result of the application of logical operations to individuals, properties, relations or other propositions (e.g. the application of the logical operation of negation to a proposition P yields a proposition that is true iff P is false; the application of the logical operation of singular predication to an item and a property yields a proposition that is true iff the item possesses the property, etc.). In general, a sentence will express a proposition such that the semantic values of expressions in the sentence will occur on the decomposition tree associated with the proposition. Thus, the semantic values of expressions in a sentence will be recoverable from the proposition (together with its decomposition tree) expressed by the sentence. However, Bealer denies that these semantic values are in any sense set-theoretic members or mereological parts of the proposition. Bealer appears to hold that the propositon is metaphysically simple and has no parts at all. As the term is used here then, Bealer's is not an account of *structured propositions* for this reason. We will see below that others hold similar "algebraic" conceptions of propositions, where the propositions *are* complex entities consisting of constituents bound together in certain ways and so *are* structured propositions.

Since the structured proposition expressed by a sentence has a structure similar to that of the sentence and has as constituents semantic values of expressions occurring in the sentence, the theory of structured propositions allows for *distinct* necessarily equivalent propositions. For example, the propositions expressed by ‘Bachelors are unmarried’ and ‘Brothers are male siblings’ presumably are both necessarily true and hence are necessarily equivalent. But clearly the propositions expressed by these sentences have different constituents and so are distinct. The proposition expressed by the former presumbly contains the semantic value of ‘bachelor’ (perhaps, the property of being a bachelor), whereas the proposition expressed by the latter doesn't. And the proposition expressed by the latter contains the semantic value of ‘brother’, whereas the proposition expressed by the former doesn't. This is an important virtue of the structured proposition view. The fact that it has this feature and that its main competitor, the possible worlds account of propositions (discussed below), doesn't is one of the reasons many favor the structured proposition view.

In discussing recent accounts of structured propositions below, I shall highlight how, on those accounts, sentences that are necessarily equivalent may express distinct propositions; and how the semantic values of expressions in a sentence are recoverable from the proposition expressed by the sentence.

2. From Possible Worlds to Structured Propositions

Because the structured proposition view arose in large part due to dissatisfaction with the then prevailing view of propositions, discussing this other account of propositions will help to illuminate structured proposition views. The late 1950's and 1960's saw the development of a new sort of model theory, "possible worlds semantics", for systems of modal logic. In the framework of possible world semantics, linguistic expressions are assigned extensions "at" possible worlds. For example, names, *n*-place

predicates and sentences are assigned individuals, sets of *n*-tuples of individuals, and truth values, respectively, at different possible worlds. Intuitively, possible worlds are to be thought of as "ways things could have been", and the assignment of (possibly different) extensions to expressions at different possible worlds is part of capturing this intuition. Thus there might have been more or fewer cows, and this is reflected in the fact that the extension of ‘cow’ (intuitively, the set of things that are cows) can vary from possible world to possible world.

Because we wish the extensions of expressions to vary from possible world to possible world (at least in some cases), it is natural to associate with each expression a function from possible worlds to extensions appropriate to that sort of expression. Thus we associate with names, functions from possible worlds to individuals; with *n*-place predicates functions from possible worlds to sets of *n*-tuples; and with sentences, functions from possible worlds to truth values. Such functions from possible worlds to extensions of the appropriate sort are often called *intensions* of the expressions in question, and I shall use the term ‘intension’ this way throughout the present work[3] Now since most think that the extensions of sentences are truth values, as indicated above, the intension of a sentence is a function from possible worlds to truth values. Intuitively, it maps a world to the value *true* if the sentence is true at that world. Thus the intension of a sentence can be seen as the primary bearer of truth and falsity at a world: the sentence has the truth value it has at the world in virtue of its intension mapping that world to that truth value. Further, modal operators were typically construed as operating on the intensions of the sentences they embed, and so those intensions could plausibly be thought of as possessing modal properties. Since propositions were traditionally held to be the primary bearers of truth and falsity and the bearers of modal properties, it was natural for possible world semanticists to identify propositions with functions from possible worlds to truth values (sentential intensions), or equivalently, sets of possible worlds (the set of possible worlds at which the sentence in question is true). Indeed, this identification was thought by many to vindicate the previously mysterious notion of a proposition.[4] Possible worlds were apparently needed for the model theory of modal logic anyway; why not build propositions out of them?

Current structured proposition accounts arose largely out of dissatisfaction with the idea that propositions are sets of possible worlds (or functions from worlds to truth values). In fact, there were at least two quite distinct motivations for abandoning the view of propositions as sets of worlds and adopting the structured proposition account.

The first had to do with the way propositions are individuated on a possible worlds account. The view that propositions are sets of possible worlds does not individuate propositions very finely. For example, consider any pair of sentences that express metaphysically necessary propositions, say ‘Bachelors are unmarried’ and ‘Brothers are male siblings’. Since these propositions are true in all possible worlds, each must be the set of all possible worlds. But there is only one such set. Thus there is only such proposition! Hence these two sentences express the same proposition. (The view also predicts that all true sentences of mathematics express the same (necessary) proposition, that any two necessarily equivalent sentences express the same proposition, that the conjunction of any sentence *S* with a necessarily true sentence expresses the same proposition as *S*, and so on.)

This should make clear that the account of propositions as sets of worlds is not a structured proposition account. For, as we saw, on a structured proposition account, the semantic values of expressions in a sentence are recoverable from the proposition expressed by the sentence, since those semantic values are constituents of the proposition. This is why on such an account, ‘Bachelors are unmarried males’ and ‘Brothers are male siblings’ express distinct propositons: the propositions have different constituents. But on the possible worlds account, the property of being a bachelor is in no sense recoverable from or a constituent of the proposition expressed by ‘Bachelors are unmarried’. For the latter proposition is just the set of all possible worlds. How could the property of being a bachelor be "recovered" from this set? Similarly, the property of being a brother is not recoverable from the proposition expressed by ‘Brothers are male siblings.’ Again, the latter proposition is just the set of all possible worlds.

Further, if propositions are sets of possible worlds, belief is construed as a relation between individuals and propositions and a sentences of the form ‘*A* believes that *P*’ asserts that the individual *A* stands in the belief relation to the proposition expressed by ‘*P*’, then for any necessarily equivalent sentences ‘*P*’ and ‘*Q*’, ‘*A* believes that *P*’ and ‘*A* believes that *Q*’ cannot differ in truth value. This means that, for example, if ‘*A* believes that 1+1=2’ is true, so is ‘*A* believes that there is no greatest natural number’. These consequences of the view that propositions are sets of possible worlds were appreciated early on; and theorists made a variety of attempts to make these consequences seem less unpalatable. Despite these valiant efforts, many philosophers viewed these consequences as a sign that there was something very wrong with the view that propositions are sets of possible worlds. Thus, philosophers were open to an account of propositions that individuated propositions more finely than the possible worlds account. As we have seen, the structured proposition account is just such an account.

In order to make clear the second motivation for abandoning the view of propositions as sets of worlds and adopting the structured proposition account, we must discuss the notions of *rigid designation* and *direct reference*. A rigid designator is an expression that designates the same individual in all possible circumstances or worlds. In the early 1970's, Saul Kripke argued in *Naming and Necessity* that ordinary proper names are rigid designators. Kripke claimed that when we consider a sentence containing an ordinary proper name, such as

> Aristotle was a great philosopher

and ask whether it would have been true or false in various counterfactual circumstances, it is the properties of the very same man, Aristotle, in those circumstances that are relevant to the truth of the sentence. So, ‘Aristotle’ designates the same man in these various counterfactual circumstances; it is a rigid designator.

At around the same time, David Kaplan argued that indexicals (e.g. ‘I’, ‘here’, ‘now’) and demonstratives (e.g. ‘that’, ‘you’, ‘he’) are *directly referential*. Concerning directly referential expressions, Kaplan wrote:

> For me, the intuitive idea is not that of an expression which *turns out* to designate the same

> object in all possible circumstances, but an expression whose semantical *rules* provide *directly* that the referent in all possible circumstances is fixed to be the actual referent. In typical cases the semantical rules will do this only implicitly, by providing a way of determining the *actual* referent, and no way of determining any other propositional component. ([1977], p. 493)

Thus, a directly referential expression is a rigid designator: its associated semantic rules determine the actual referent of the expression (in a context) and when evaluating what is said by the sentence containing the expression (in that context) in other possible circumstances, this same referent is always relevant. To illustrate, if I utter

> I ski.

at the present time and we want to evaluate whether what I said by means of that utterance is true or false in other possible circumstances, it is *my* properties in those other circumstances that are relevant. Thus, 'I' is rigid: when evaluating the truth or falsity of what is said by an utterance of a sentence containing 'I' in counterfactual circumstances, it is the properties of the person whom 'I' referred to in the utterance (the actual utterer) that are relevant.

Kaplan intended to contrast directly referential expressions with expressions such as definite descriptions, which, though designating particular individuals, do so by means of descriptive conditions being expressed by the description and satisfied by the designated individual. Thus Kaplan wrote that directly referential expressions "refer directly without the mediation of Fregean *Sinn* as meaning". ([1977], p. 483) The designation of definite descriptions *is* mediated by something like a Fregean sense (i.e. their associated descriptive conditions).

Of course, even if descriptions are not directly referential, some are rigid designators. For example, 'the successor of 1' designates the same individual (namely, the number 2) in all possible worlds. So, though all directly referential expressions are rigid designators, some rigid designators are not directly referential. As was mentioned above, in a possible worlds semantics linguistic expressions are associated with intensions, functions from possible worlds to appropriate extensions. In the case of expressions designating individuals, these intensions will be functions from possible worlds to individuals. Note that *all* rigid designators (whether directly referential or not) will have intensions that are constant functions: they will be functions that map all possible worlds to the same individual. Thus possible worlds semantics tends to blur the distinction between directly referential expressions and rigid non-directly referential expressions (e.g. rigid definite descriptions). To make the distinction between directly referential expressions and rigid non-directly referential expressions more vivid, Kaplan invoked the notion of structured propositions:

> If I may wax metaphysical in order to fix an image, let us think of the vehicles of evaluation -- the what-is-said in a given context -- as propositions. Don't think of propositions as sets of possible worlds, but rather as structured entities looking something

> like the sentences which express them. For each occurrence of a singular term in a sentence there will be a corresponding constituent in the proposition expressed. The constituent of the proposition determines, for each circumstance of evaluation, the object relevant to evaluating the proposition in that circumstance. In general the constituent of the proposition will be some sort of logical complex, constructed from various attributes by logical composition. But in the case of a singular term which is directly referential, the constituent of the proposition is just the object itself. Thus it is that it does not just *turn out* that the constituent determines the same object in every circumstance, the constituent (corresponding to a rigid designator) just *is* the object. *There is no determining to do at all*. On this picture -- and this is *really* a picture and not a theory -- the definite description
>
> (1) The n[(snow is slight & n^2=9) $\vee$ (~snow is slight & 2^2=n+1)]
>
> would yield a constituent which is complex although it would determine the same object in all circumstances. Thus, (1), though a rigid designator, is not directly referential from this (metaphysical) point of view. ([1977], pp. 494-495)

(Kaplan goes on to attribute this "metaphysical picture" of structured propositions to Russell.) Adopting this structured proposition account makes it simple to distinguish between directly referential expressions and other expressions, rigid or not. Directly referential expressions contribute their *referents* (in a context) to the propositions expressed (in that context) by the sentences containing them. Non-directly referential expressions contribute some complex that may or may not determine the same individual in all possible circumstances. Thus the desire to distinguish clearly between directly referential expressions and other rigid designators prompted Kaplan to re-introduce the Russellian notion of a structured proposition into the philosophical literature (see the discussion of Russell below). However, in [1977], Kaplan tends to treat the notion of a structured proposition as a heuristic device. He repeatedly calls it a picture, explicitly says that it is not part of his theory, and in his formal semantics he adopts the possible worlds account of propositions (contents of formulae), taking them to be functions from worlds (and times) to truth values.

Many current direct reference theorists take the structured proposition account much more seriously. It *is* part of *their* theory in the sense that when they say that an expression is directly referential they are *literally* saying that it contributes its referent to propositions expressed by sentences containing it, (e.g. see the discussion of Salmon and Soames below).

3. Some Recent Accounts of Structured Propositions

Having discussed structured proposition accounts in a general way, the best way to further illuminate these accounts of propositions is to discuss some recent work on structured propositions. In so doing, we shall see various respects in which accounts of structured propositions can differ. Three caveats before

proceeding. First, I mentioned above that structured proposition accounts, unlike possible world accounts of propositions, allow for distinct necessarily equivalent propositions, and thus individuate propositions more finely than possible worlds accounts. There are other accounts of propositions, or things that are intended to do the work of propositions, that are not structured proposition accounts (given the way that term is used here), but that allow for distinct necessarily equivalent propositions or things that are to do the work of propositions. I already mentioned the example of Bealer [1993]. Another example is the *interpreted logical form* account defended by Larson and Ludlow [1993]. Though such accounts will not be discussed here, the reader should be aware of them and that they are motivated by many of the considerations that motivate structured proposition theorists. In particular, they attempt to individuate propositions (or things that do the work of propositions) more finely than possible worlds accounts of propositions. Second, in discussing a sampling of recent work on the view that propositions are structured, I do not intend to exhaust the versions of structured proposition approaches that there are. Rather, I intend to highlight some of the main issues and current approaches. To that end, I shall discuss three broad approaches to structured propositions. I call these the *Neo-Russellian Approach*, the *Structured Intensions Approach*, and the *Algebraic Approach*. In discussing each approach, I shall mention a number of authors who adopt the approach, but for definiteness I shall concentrate primarily on a representative of that approach in explaining it. The reader should be aware that these groupings are somewhat loose, and that there may be important differences between authors who are grouped together. Third, though I shall here and there raise questions, I shall not try here to criticize the various approaches I discuss. I view my task as that of introducing the reader to approaches to structured propositions. The task of criticizing the various apporaches is a task for another day. I begin with the neo-Russellian approach.

The Neo-Russellian Approach

In a series of papers and a book, Scott Soames [1985, 1987, 1989] and Nathan Salmon [1986a, 1986b, 1989a, 1989b] have laid out what is probably the best known current theory of structured propositons. There are some differences of detail between Salmon and Soames, but I shall treat them here as holding the same view. Though I shall discuss some of their contributions separately, I shall follow the account of structured propositions laid out in Soames [1987].

First, Soames [1985,1987] produced what many take to be a devastating attack on the view of propositions as sets of possible worlds. Soames showed that even when one tries to get more fine-grained propositions-as-sets-of-worlds by allowing metaphysically impossible worlds (e.g. worlds in which George Bush is identical with Ronald Reagan), inconsistent worlds (in which a thing can both possess and lack a property), and incomplete worlds (where some purported "matters of fact" are simply not settled), the resulting view, when combined with other independently plausible assumptions, is riddled with overwhelming difficulties. These difficulties all stem from the fact, noted earlier, that on the worlds view, sentences with very different syntactic structures and containing words with different semantic values may express the same proposition. Soames [1987] concludes that we ought to give up the view that propositions are sets of worlds of any sort, and embrace an account of propositions according to which propositions are structured entities, with individuals, properties and relations as constituents.

Soames called these *structured Russellian propositions*. If the syntactic structures of sentences and the semantic values of words occurring in them are reflected in the structures and constituents of propositions they express, sentences with different syntactic structures and containing words with different semantic values, whether true in all the same worlds or not, may express different propositions. It is perhaps worth noting that having sentences with different syntactic structures and containing words with different semantic values express different propositions doesn't *require* one to hold that propositions themselves are structured and contain the semantic values of the words as constituents. Still, it is a natural way of accounting for why sentences with different syntactic structures and containing words with different semantic values that are true in all the same worlds express different propositions. Soames [1987] sketches a formal theory of structured propositions, including an assignment of structured propositions to the sentences of a simple formal language, and a definition of *truth relative to a circumstance* for structured propositions.

Soames and Salmon are direct reference theorists, holding that names (as well as indexicals and demonstratives) have their referents as their semantic values and so contribute them to the propositions expressed by sentences containing them. Further they hold that predicates and intransitive verbs have properties as their semantic values; and that transitive verbs have relations as their semantic values. Thus they hold that sentences such as

(4) Scott runs.

(5) Scott saw Nathan.

express the propositions

(4a) $<<o>, R>$

(5a) $<<o,o'>,S>$

where o is Scott, o' is Nathan, R is the property of running, and S is the relation of seeing. The negation of (4) expresses the proposition

(4b) $<\mathrm{NEG}, <<o>, R>>$

where NEG is the truth function for negation. And the conjunction of (4) and (5) (in that order) expresses the proposition

(5b) $<\mathrm{CONJ}< <<o>, R>, <<o,o' >,S>>>$

where CONJ is the truth function for conjunction. Similar remarks apply to sentences formed with other truth functional connectives. Further a sentence such as

Something runs.

expresses the proposition

(6a) <SOME, *g*>

where SOME is the property of being a nonempty set and *g* is the function from individuals *o*′ to the proposition <<*o*′>,*R*>, (where, as before, *R* is the property of running).

It should be easy to imagine the definition of *truth relative to a circumstance* for structured propositions of the sort mentioned above. For example, (5a) will be true at a circumstance *c* iff <*o*,*o*′> is in the extension of the relation *S* at *c*. (5b) will be true at *c* iff CONJ maps the truth values of <<*o*>, *R*> and <<,*o*,*o* ′>,*S*> at *c* to truth. And (6a) will be true at *c* iff there is an individual in *c* that *g* maps to a proposition that is true in *c*, (in that case the set of individuals in *c* that *g* maps to propositions true in *c* possesses SOME).

It should be clear that on this account of propositions sentences that are necessarily equivalent may express distinct propositions, and that the semantic values of expressions in a sentence are recoverable from the proposition it expresses. For example, ‘All bachelors are unmarried’ and ‘All brother are male’ are both true in all possible worlds. But on the present view, the former expresses a proposition that has as a constituent a propositional function mapping an object to the proposition that that object is unmarried; the latter does not. Further, e.g. the semantic value of ‘runs’ as it occurs in sentence (4) is recoverable from the proposition (4) expresses in that it is a constituent of that proposition. Similar remarks apply to 5/5a and 6/6a (except that in 6a the semantic value of ‘runs’ is encoded in *g*: that is, *g* maps individuals o to the proposition that *o* runs, which in turn has the semantic value of ‘runs’ as a constituent).

Note that this account of propositions (including the commitment to names being directly referential) entails that sentences that differ only with respect to coreferential names express the same proposition. Thus

Mark Twain is Samuel Clemens.

and

Samuel Clemens is Samuel Clemens.

express the same proposition on this view. Many have found this result incredible, since it would appear that the one sentence could be informative and the other not. Salmon and Soames also hold that ‘believes’ expresses a relation between individuals and propositions, so that ‘Scott believes that Mark Twain is Samuel Clemens’ expresses a proposition to the effect Scott stands in the believes relation to the propositon expressed by ‘Mark Twain is Samuel Clemens’. But then it follows that ‘Scott believes

that Mark Twain is Samuel Clemens' and 'Scott believes that Samuel Clemens is Samuel Clemens' express the same proposition (since the embedded sentences in both belief ascriptions express the same proposition) and so cannot diverge in truth value. Many have found this consequence of the Salmon-Soames view of propositions (and belief ascriptions) hard to swallow as well.

Salmon [1986] is largely an extended defense of these two consequences of the Salmon-Soames view. It is beyond the scope of the present work to explain Salmon's defense and the interested reader should consult that work directly.[5]

In the formal semantics offered by Salmon and Soames, propositions are ordered *n*-tuples (or concatenations of *n*-tuples), as are (4a), (5a), (5b) etc. above. But since Salmon and Soames say nothing explicitly about what holds propositions together, it is unclear whether these *n*-tuples and concatenations merely *represent* propositions in the formalism, or whether Soames and Salmon take them *to be* propositions. If the former, then the Salmon and Soames view is incomplete and we need to be told what propositions really are, and more specifically what it is that really holds propositions together (i.e. what the corner brackets in (4a), (5a) etc. stand for). If the latter, then the view at least seems to have trouble accounting for some of the properties possessed by propositions. Propositions have truth conditions: they are true or false, depending on how the world is. So if some ordered *n*-tuples are propositions, some ordered *n*-tuples have truth conditions. But ordered *n*-tuples don't seem to be the kinds of things that have truth conditions. Indeed, presumably many ordered *n*-tuples have no truth conditions, (e.g. <1,2,3>). So how/why did those *n*-tuples that are propositions come to have truth conditions? Similar remarks apply to modal properties. Propositions are necessary, contingent, and possible. These, again, don't seem to be properties of *n*-tuples. Finally, if propositions are ordered *n*-tuples, that is, set theoretic constructions, it is hard to see why a *particular* set theoretic construction is the proposition in question (and so has truth conditions, modal properties, etc.) as opposed to some other set theoretic construction that seems equally well suited to the task.[6] For example, we said that the sentence

> (4) Scott runs.

expresses the proposition/set theoretic construction

> (4a) <<*o*>, *R*>

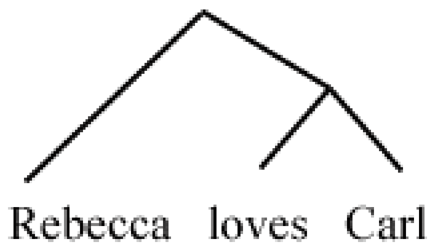
But the following set theoretic construction seems equally suited to be the proposition (4) expresses:

> (4b) <*R*, <*o*>>

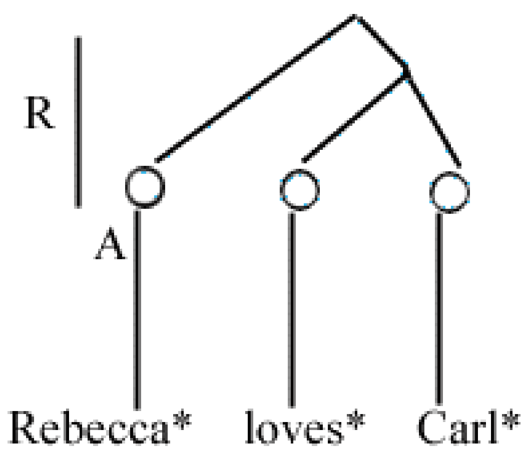
So why it is that (4a), instead of (4b), is the proposition (4) expresses, and has modal properties and truth conditions?

This question as to what holds structured propositions together and gives them their structure turns us to other recent work within the neo-Russellian approach. While adopting more or less the same view as

Salmon and Soames on the semantic values of different kind of words (though claiming to be strictly neutral on the question of what the semantic values of names and predicates are), Jeffrey C. King [1994, 1995, 1996] develops a view as to what binds the constituents of propositions together. He holds that propositions are *not n*-tuples and that a complex relation binds together the constituents of a proposition and provides the proposition with its structure. In order to explain what complex relation binds the constituents together, let us consider the sentence ‘Rebecca loves Carl’ in "tree form":



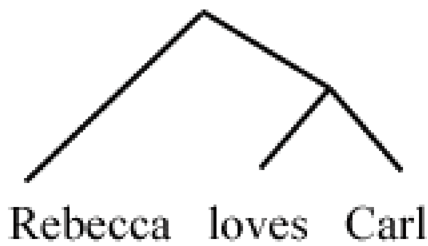
The tree represents the syntactical relation binding together the words in the sentence. We shall call this the *sentential relation* of the sentence. The words in the sentence, of course, bear semantic relations to things in the world, their semantic values (or s.v.'s): the s.v. of ‘Rebecca’ is Rebecca; the s.v. of ‘loves’ is the relation of loving; and the s.v. of ‘Carl’ is Carl. The crucial point is that since the words stand in a sentential relation in the sentence, and the words in the sentence stand in semantic relations to the s.v.'s, the s.v.'s themselves stand in the relation resulting from composing the sentential relation of the sentence with the semantic relations the words in the sentence bear to their s.v.'s, while existentially generalizing over the words. We can represent the relation the s.v.'s stand in thus:



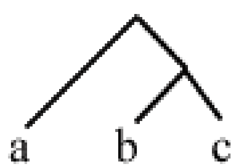
where Rebecca*, loves* and Carl* are the s.v.'s of ‘Rebecca’, ‘loves’ and ‘Carl’; the portion of the diagram labeled *R* is the sentential relation binding together the words in the sentence above; the circles at the terminal nodes of this relation represent that the words have been existentially generalized away; and the lines from the circles to the s.v.'s represent the semantic relations obtaining between ‘Rebecca’ and Rebecca* (i.e. Rebecca), etc. Thus, for example, *A* is presumably the reference relation that obtains between ‘Rebecca’ and Rebecca*. On King's view, then, the proposition expressed by ‘Rebecca loves

Carl' *is* the above: namely, the s.v.'s of the words in this sentence standing in the complex relation that is the result of composing the sentential relation of the sentence with the semantic relations the words bear to their s.v.'s, existentially generalizing away the words. Thus the relation that holds together the constituents of the proposition is literally "built out of " the sentential relation and the semantic relations words bear to their s.v.'s.

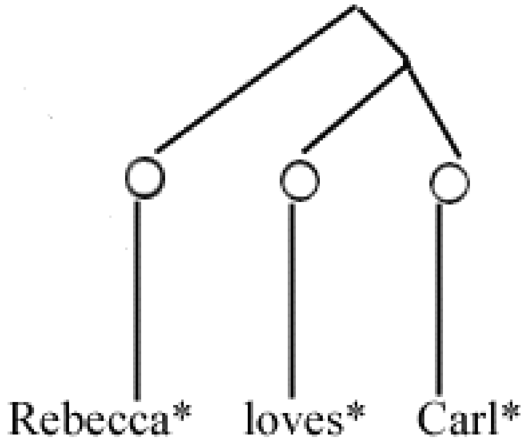
This view of what holds the constituents of propositions together has at least three virtues. First, it "builds" the relations binding together constituents of propositions out of relations (sentential relations binding together words in sentences; and semantic relations between words and their s.v.'s) that we are already committed to. Second, and related to this, it leaves little room for doubt that propositions exist. For from the fact that there is a sentence 'Rebecca loves Carl', which in tree form looks as follows:



Where 'Rebecca' bears a semantic relation to Rebecca, 'loves' bears a semantic relation to the relation of loving, and 'Carl bears a semantic relation to Carl, it follows that Rebecca, the relation of loving and Carl stand in the relation of there being lexical items *a*, *b* and *c* occurring in a sentence in which they stand in the following relation:



such that Rebecca is the s.v. of *a*, the relation of loving is the s.v. of *b*, and Carl is the s.v. of *c*. In other words, it follows that the following entity exists:



But this entity just is our proposition! Thus, given very minimal, quite uncontentious assumptions, it follows that propositions exist on this view. So even philosophers with rather strict naturalistic scruples will have to admit that propositions as they are construed on the present view exist. Third, the view can explain why propositions are the sorts of things that represent or have truth conditions. I refer the interested reader to King [1995] for this explanation.

The Structured Intensions Approach

Having seen the main features of neo-Russellian approaches, let us turn to structured meaning accounts. The roots of these accounts can be traced to Rudolf Carnap [1947], and his notion of intensional isomorphism. David Lewis [1972] and Max Cresswell [1985] have worked out similar, detailed versions of the structured meanings approach, though there are important differences in their views. I shall here discuss Cresswell's [1985] version of the view.[7] Because our concern is with conceptions of structured propositions, there are many features of Cresswell [1985] that I shall not discuss (e.g. I shall not desribe Cresswell's account of the semantics of verbs of propositional attitude).

Both Lewis [1972] and Cresswell [1985] are motivated by some of the same considerations that motivated neo-Russellians like Salmon and Soames. Lewis and Cresswell both wish to find a more fine grained "semantic value" for sentences than functions from worlds (or, as in Lewis [1972], *indices*) to truth values or, equivalently, sets of worlds. For example, Cresswell [1985] claims that verbs of propositional attitude are (sometimes) sensitive to more than the intensions (functions from worlds/indices to truth values) of the sentences they embed. Thus Cresswell claims that

> ...one might easily have two sentences α and β that are true in exactly the same worlds and yet are such that
>
> > x φs that α
>
> is true, but

x φs that β

is false.
[p. 73; φ here is a propositional attitude verb such as ‘believe’.]

Thus Cresswell wishes to associate with a sentence (or more accurately, a ‘that’ clause) some semantic value that is more fine grained than a set of worlds, so that attitude verbs may (sometimes) distinguish between sentences that are true in exactly the same worlds. Strictly, Cresswell holds that sometimes attitude verbs are sensitive only to the sets of worlds in which the sentences they embed are true (i.e. their intensions); but sometimes attitude verbs are sensitive to more than this. In these latter cases, Cresswell wishes to asscociate a semantic value more fine grained than a set of worlds with the embedded sentence (or ‘that’ clause). Cresswell accomplishes this by holding that ‘that’ in English attaches to a sentence to form a name, and that ‘that’ in this role is highly ambiguous. In one of its meanings, ‘that’ attaches to a sentence and forms a name of the intension of the sentence (i.e. the set of worlds in which it is true).[8] In such cases, the attitude verb is sensitive only to the intension named by the ‘that’ clause following it. However, ‘that’ has another meaning on which it combines with a sentence to form the name of a much more fine-grained entity (in the sense that sentences true in all the same worlds may be associated with different fine-grained entities). In such cases, verbs of attitude are sensitive to differences in these fine-grained entities. Since our concern is with structured propositions, or with semantic values of sentences more fine grained than sets of worlds, I shall henceforth focus on Cresswell's account of these fine-grained entities named by *some* ‘that’ clauses.[9] Henceforth, then, when I talk about the fine grained entity *associated with* a sentence or ‘that’ clause on Cresswell's view, I should be understood as talking about what a ‘that’ clause containing the sentence names, given that the meaning of ‘that’ in the ‘that’ clause is the one that when combined with the sentence yields a name of the most fine grained entity named by any ‘that’ clause in which the sentence occurs. This is important to bear in mind, since for Cresswell, strictly speaking, a sentence not in a ‘that’ clause does not express this fine grained entity.

Consider the sentence

(7) Max runs.

For Cresswell, the meaning of a predicate like ‘runs’ is essentially its intension: a function from individuals to sets of worlds (it maps an individual to the worlds at which she runs).[10] Let I_r be this intension. The meaning of a name like ‘Max’ (in some cases at least) is simply its referent: *o*. Thus, the fine grained entity associated with (7) is the ordered pair:

(7a) $<o,I_r>$

The negation of (7) will be associated with the following:

(7b) $<NOT,<o,I_r>>$

where NOT is the function from sets of worlds to sets of worlds that maps a set of worlds to its complement.

Finally, the sentence:

(8) Someone runs.

is associated with

(8a) $<\Sigma,I_r>$

where Σ is the function *from* functions from individuals to sets of worlds *to* sets of worlds such that $\Sigma(f) = \{w$: for some o, w is in $f(o)\}$.

It should be clear that, as on the neo-Russellian approach, sentences that are true in all the same worlds may be associated with different fine grained entities of the sort posited by Cresswell. For example, even if ‘All brothers are siblings’ and ‘All bachelors are male’ are true in exactly the same worlds, the fine grained entity associated with the latter will contain the meaning (function from individuals to sets of worlds) of ‘male’ and the fine grained entity associated with the former will not. It should also be clear that the semantic values of expressions are recoverable from the propositions expressed by sentences containing them. For example, the semantic value of ‘runs’ as it occurs in (7) is a constituent of the fine-grained entity expressed by (7) (i.e. (7a)). Similar remarks apply to (7b) and (8a).

A final remark about Cresswell's view is in order. The fine-grained entities we have been discussing are what verbs of propositional attitude are sensitive to when they are sensitive to more than the intension/set of worlds associated with a sentence they embed. However, they do not seem to be the primary bearers of truth for Cresswell. For Cresswell, each sentence is associated with a set of worlds/intension, which Cresswell calls a *proposition*; and the truth or falsity of a sentence at a world is determined by what proposition it expresses. So these seem to be the primary bearers of truth and falsity for Cresswell.[11] Thus, it appears that for Cresswell, in contrast to the neo-Russellians, the primary bearers of truth and falsity and the fine grained entities associated with the sentences (or ‘that’ clauses) that verbs of attitude embed and to which they are (sometimes) sensitive are different.

The Algebraic Approach

Finally, I turn to algebraic approaches. Above, I mentioned the view of Bealer [1982]. Though Bealer in some sense adopts an algebraic approach to propositions (see the above discussion), as mentioned above he apparently holds that propositions have no (set theoretic or mereological) parts, and so doesn't count as a structured proposition theorist in the sense in which the term is used here. Thus, I will consider here

adherents to the algebraic approach who do hold that the propositions yielded by their algebras are complex and have "parts". Aside from Bealer [1979, 1982 and 1993], work in this tradition includes Edward Zalta [1983 and 1988], and Christopher Menzel [1986 and 1993]. I shall focus here on the formulations in Zalta [1988]. Though Zalta has an extensive, axiomatized theory of propositions, and ordinary and abstract individuals, properties, and relations, we confine our attention here to his view about propositions. Still, since Zalta views propositions as zero-place relations, we will say something about his views on properties and relations.

Advocates of the algebraic approach such as Zalta, like the neo-Russellians and the advocates of structured meaning approaches, think that a good theory of propositions must allow for distinct, necessarily equivalent propositions. Thus, he writes:

> Necessarily equivalent propositions may be distinct. If the theory of propositions is not fine-grained enough to distinguish necessarily equivalent propositions, the ability to accurately represent belief is lost. [Zalta 1988, p. 57]

To appreciate how Zalta achieves the goal of having a fine-grained theory of propositions, we begin by discussing how he views relations generally, since, as just mentioned, Zalta takes propositions to be zero-place relations (and properties to be one-place relations). Zalta holds that relations are "primitive entities", by which he means that they are not to be explained or "defined" in terms of other entities/notions. But at least some relations are complex. For example, if we take a two-place relation (between objects) *Rxy* and "plug" one of its argument places with the object *b*, we get the one-place relation (property) *Rxb* ("bearing *R* to *b*").[12] This one-place relation is complex, having *b* and *R* as parts. Similarly, if we take a three-place relation *Sxyz* and "universalize" the third argument position, we get a two place relation that we might represent thus: $(\forall z)Sxyz$ ("*x* and *y* (in that order) stand in *S* to everything"). Here again, the two-place relation is complex, having *S* and (something corresponding to) "universalization" as parts. Zalta's idea is that properties, relations and individuals can be "harnessed together" to form new, complex relations. In his axiomatized theory of relations, Zalta introduces a comprehension schema for relations (see Relations, p. 46) that insures that all manner of complex relations will be available. To insure that all instances of the comprehension schema are true in all interpretations of his axiomatized theory, Zalta has these interpretations include the following group of "logical functions": PLUG_i, NEG, COND, UNIV_i, $\text{REFL}_{i,j}$ $\text{CONV}_{i,j}$, VAC_i, NEC, WAS and WILL. Roughly (and suppressing reference to worlds and times) PLUG_i is a function that maps an *n*-place relation *R* and an object *b* to the $n-1$ place relation R' such that $\langle o_1,...,o_{i-1}, o_{i+1},...,o_n\rangle$ stand in R' iff $\langle o_1,...,o_{i-1}, b, o_{i+1},..., o_n\rangle$ stand in *R*. NEG is a function that maps an *n*-place relation *R* to an *n*-place relation R' such that *n* things stand in R' iff they don't stand in *R*.[13] Thus the repeated application of these functions yields appropriate relations to make true the instances of Zalta's comprehension schema for relations. For example, consider the following two instances of his comprehension schema (where ‘F’ is a variable ranging over one-place relations; ‘*b*’ is a name of an individual; and the other predicate letters are constants and so name particular relations):[14]

$(\exists F)(Fx \text{ iff } Rxb)$

(∃*F*)(*Fx* iff ~*Px*)

If we apply PLUG_2 to the individual denoted by '*b*' and the (two-place) relation denoted by '*R*', we get a one-place relation that makes the first instance of the schema true; and if we apply NEG to the denotation of '*P*', we get a one-place relation that makes the second instance true.

As we have mentioned, propositions are zero-place relations for Zalta. Thus, PLUG_i and the rest of the "logical functions" mentioned above can be applied to various entities to yield propositions. Thus the proposition expressed by a sentence like

Ed runs.

is the result of applying the PLUG_1 function to the property of running and Ed. This proposition consists of Ed saturating the one argument place of the running property. Similarly, a sentence like

Ed does not run.

expresses the proposition that is the result of applying PLUG_1 to running and Ed as before, and then applying NEG to the output of PLUG_1. Finally, a sentence like

Everything runs.

expresses the result of applying UNIV_1 to the property of running. This idea that there is some group of "logical functions" whose repeated application to some other entities yield complex propositions (and relations) is characteristic of what I am calling algebraic approaches.

It should be clear that necessarily equivalent sentences may express distinct propositions on Zalta's view. For example, the sentences 'All brother are male siblings' and 'All bachelors are unmarried' express propositions that are true in all possible worlds. But the first expresses a proposition that results from applying COND to the properties of being a brother and being a male sibling, and applying $\text{REFL}_{1,2}$ and then UNIV_1 to the output of COND. The second does not, but instead results from applying these functions in the same order to the properties of being a bachelor and being unmarried. Further, it should be clear that the semantic values of expressions are recoverable from the propositions expressed by sentences in which they occur. E.g. both Ed and the property of running, which are the semantic values of 'Ed' and 'runs', are constituents of the proposition that Ed runs, which is expressed by 'Ed runs', since this proposition consists of Ed saturating the one argument place in the property of running.

Having sketched Zalta's view of propositions it is worth mentioning a couple points about it. At least in some cases, what binds together the constituents of a proposition is in some sense "built into" one of the

constituents of the proposition for Zalta, (as we shall see below, the same is true for Frege and Russell). Consider again the sentence:

> Ed runs.

Recall that this sentence expresses the proposition that results from applying $PLUG_1$ to Ed and the property of running. The output is Ed saturating ("plugged into") the one argument place of the property of running. There are two things to notice about this proposal. First, it is one of the constituents of the proposition, the running property, that binds the constituents of the proposition together. The proposition is held together in virtue of Ed "plugging" the one argument place in the running property. Second, the proposition *just is* Ed "plugging"/possessing that property!

Such a view immediately raises a worry about false propositions (discussed below in connection with Russell's POM account of propositions). For one might argue as follows. Suppose Ed doesn't run. Then Ed doesn't plug/saturate the one argument place in the property of running (i.e. doesn't possess this property). But Ed possessing the running property *just is* the proposition that Ed runs. Thus there is no proposition that Ed runs. Similar reasoning shows that any other false proposition fails to exist. So there are no false propositions.

Obviously, Zalta does not want to be saddled with this result, and he isn't. He holds that the even if Ed doesn't run, there is a proposition consisting of Ed saturating the argument place in the running property. That proposition simply isn't true. But the proposition consists of Ed plugging the property if running. For Ed to do this is simply different from Ed exemplifying or possessing that property, which is what happens when the proposition is true. It is in the nature of the running property to be "plugged" by things like Ed, to form the proposition that Ed runs. And it is also in its nature to allow Ed to exemplify it, in which case the result of plugging Ed into the property of running is true. This makes it sound as though the complex consisting of Ed exemplifying running makes true the proposition consisting of Ed plugging running. However, Zalta appears to deny this when he says:

> The metaphysical truth or falsity of these logical complexes [propositions] is basic. If a proposition is true, there is nothing else that "makes it true". Its being true is just the way things are (arranged). ([1988], p. 56)

Finally, though we shall not go into it here, it is worth mentioning that when Zalta gets to the semantics of verbs of propositional attitudes, it is not only the propositions considered thus far that are objects of the attitudes and make belief ascriptions true or false. Zalta ends up claiming that sentences embedded with respect to attitude verbs are ambiguous, sometimes expressing the sorts of propositions we are discussing, and sometimes expressing other propositions that instead of containing the individuals, properties and relations thus far discussed, contain *senses* of these individuals, properties and relations. These senses are "abstract" individuals and relations, instead of the ordinary individuals and relations discussed thus far. Hence Zalta ends up with a theory of belief ascriptions that invokes both fine-grained propositions and neo-Fregean senses. The interested reader should consult Zalta [1988].

4. Historical Antecedents to Current Views: Frege

As with many ideas discussed in contemporary philosophy of language, the idea that propositions are structured is present in Gottlob Frege's writings. Frege had a view both about the constituents of structured propositions and about what held these constituents together in the proposition. Frege held that simple linguistic expressions are associated with entities he called *senses*. Though there is some controversy about precisely how to understand the notion of sense, Frege explicitly distinguishes the sense of a linguistic expression from both the subjective ideas speakers associate with the expression and the thing in the world the expression "stands for". Further, the sense of an expression determines what thing in the world the expression stands for. Thus the sense of a proper name such as ‘Ronald Reagan’ must be distinguished from any subjective ideas speakers associate with the name (e.g. feelings of anger, fondness, etc.) and from Reagan himself. And the sense of the name "picks out" Reagan as the thing in the world the name stands for. It may help to think of the sense as some descriptive condition satisfied uniquely by Reagan (and not to think any more about what is meant by a *descriptive condition*). Complex linguistic expressions are also associated with senses. And Frege held that the sense of a complex expression is a function of the senses of its simple parts and how they are put together. Frege called propositions *thoughts* (*Gedanken*), and held that the thought/proposition expressed by a sentence is itself a sense. And, like the senses of other complex linguistic expressions, the proposition/thought expressed by a sentence is a function of the senses of the words in the sentence and how they are put together. Now Frege at least sometimes appears to hold the stronger view that the sense of a sentence (proposition/thought) *has as constituents* the senses of the words in the sentence. And as the following quotation shows, his account of how these sense-constituents are held together in the proposition/thought depends on different kinds of linguistic expressions having different kinds of senses:

> For not all parts of a thought can be complete; at least one must be ‘unsaturated’, or predicative; otherwise they would not hold together. For example, the sense of the phrase ‘the number 2’ does not hold together with that of the expression ‘the concept *prime number*’ without a link. We apply such a link in the sentence ‘the number 2 falls under the concept *prime number*’; it is contained in the words ‘falls under’, which need to be completed in two ways -- by a subject and an accusative; and only because their sense is thus ‘unsaturated’ are they capable of serving as a link. ([1892], p. 54)

Remarks like this suggest that Frege's view was that propositions are complex entities whose parts are other senses. The proposition is held together in virtue of the fact that at least one of the senses is unsaturated (Frege sometimes also says predicative, or in need of supplementation). The complete or saturated senses "saturate" or complete the unsaturated senses and in so doing are bound to them to form the proposition/thought. Thus, on this way of interpreting Frege, the mechanism for binding together the constituents of a structured proposition is built right into some of the constituents. (Recall that this was true of the view of Zalta [1988] as well.)

5. Historical Antecedents to Current Views: Russell

Bertrand Russell, to whom many current structured propositions theorists attribute the idea of structured propositions, held various views about the nature of propositions over the course of his career. However, the account of propositions held by Russell that is thought by many to be the progenitor of current accounts of structured propositions is the one Russell defended in *Principles of Mathematics* (Russell [1903]). We shall confine our attention to that account here. Russell differed with Frege both on what the constituents of structured propositions are and on what binds them together in the proposition. Russell uses the word ‘term’ for constituents of propositions. Thus he writes:

> Whatever may be an object of thought, or may occur in any true or false proposition ...I call a *term*. ...A man, a moment, a number, a class, a relation, a chimaera, or anything else that can be mentioned, is sure to be a term; and to deny that such and such a thing is a term must always be false. ([1903], p. 43)

Thus we already see that Russell differs from Frege on what kinds of things can be constituents of propositions. For Frege, all constituents of propositions are senses. For Russell, a man or a mountain can be a constituent of a proposition. In a now famous correspondence between Frege and Russell, Frege asserted that the sense of the name ‘Mont Blanc’, and not Mont Blanc itself "with all its snowfields", occurs in the proposition/thought that Mont Blanc is 4,000 meters high. Russell replied by saying:

> I believe that in spite of all its snowfields Mont Blanc itself is a component part of what is actually asserted in the proposition ‘Mont Blanc is more than 4,000 metres high’. (reprinted in Gabriel, *et al.*, [1980], p. 169)

Russell goes on to distinguish two kinds of terms or propositional constituents:

> Among terms, it is possible to distinguish two kinds, which I shall call respectively *things* and *concepts*. The former are the terms indicated by proper names, the latter those indicated by all other words. ([1903], p. 44)

So for Russell, all propositional constituents are things or concepts. On Russell's view, a sentence such as:

> Socrates is human

expresses a proposition with three constituents, corresponding to the three words in the sentence. Socrates himself is one of the constituents, the other two constituents being the concepts contributed by ‘is’ and ‘human’.

Having seen that Russell held different kinds of things to be constituents of propositions than did Frege, we shall turn to his views on what binds together the constituents of structured propositions. Russell appears to hold that the propositional contributions of verbs (which contributions Russell often calls

verbs) hold together the constituents of propositions. Thus he writes:

> Consider, for example, the proposition 'A differs from B'. The constituents of this proposition, if we analyze it, appear to be only *A*, difference, *B*. Yet these constituents, thus placed side by side, do not reconstitute the proposition. The difference which occurs in the proposition actually relates *A* and *B*, whereas the difference after analysis is a notion which has no connection with *A* and *B*. ([1903], p. 49)

And later Russell writes:

> Owing to the way in which the verb [propositional contribution of a verb] actually relates the terms of a proposition, every proposition has a unity which renders it distinct from the sum of its constituents. ([1903], p. 52)

Russell's idea that a proposition is something beyond the sum of its constituents certainly seems correct. The collection, or mereological sum, of *A*, difference and *B* is not the proposition that *A* differs from *B*. Different propositions may have the same Russellian constituents, as with the propositions expressed by the following sentence pair:

> Jason loves Patti.
>
> Patti loves Jason.

However, Russell's idea that the propositional contribution of the verb binds the constituents together (in different ways in the case of our pair of sentences) is hard to understand. In the next to last quotation from Russell above, he suggests that the proposition expressed by '*A* differs from *B*', whose constituents are *A*, difference and *B*, is held together by *difference* actually obtaining between *A* and *B*. If we call *difference* a relation, the proposition would consist of *A* standing in the relation of difference to *B*. Russell makes a number of remarks that suggest that this is his view. But it would seem that this cannot be correct. Intuitively, *A* standing in the relation of difference to *B* is what makes the proposition that *A* differs from *B true*. It is not the proposition itself. For if it were, it would seem that if *A* doesn't stand in the difference relation to *B*, then there is no (false) proposition to the effect that *A* differs from *B*. And quite generally, for the same reason it would appear that there are no false propositions. Perhaps Russell meant that propositional contributions of verbs hold propositional constituents together in some other manner. But it certainly is not clear what that manner would be. Or perhaps he held a view simlar to Zalta's discussed above.

Bibliography

- Bealer, George, 1979, 'Theories of Properties, Relations and Propositions', *Journal of Philosophy* 76: 634-648

- Bealer, George, 1982, *Quality and Concept*, Clarendon Press, Oxford
- Bealer, George, 1993, 'A Solution to Frege's Puzzle', in *Philosophical Perspectives, 7, Language and Logic*, Ridgeview Publishing Company, Atascadero, CA
- Carnap, Rudolf, 1947, *Meaning and Necessity*, University of Chicago Press
- Cresswell, M.J., 1985, *Structured Meanings*, MIT Press, Cambridge, MA
- Crimmins, Mark, 1992, *Talk About Beliefs*, MIT Press
- Frege, Gottlob, 1892, 'On Concept and Object', in *Translations from the Philosophical Writings of Gottlob Frege*, Geach and Black (eds.), Basil Blackwell, Oxford, 1977
- Gabriel, G., *et al.*, (eds), 1980, *Gottlob Frege: Philosophical and Mathematical Correspondence*, Chicago: University of Chicago Press
- Kaplan, David, 1977, 'Demonstratives', Draft #2, in *Themes From Kaplan*, J. Almog, H. Wettstein, and J. Perry (eds.), Oxford University Press, Oxford 1989
- King, Jeffery C., 1994, 'Can Propositions be Naturalistically Acceptable?', *Midwest Studies in Philosophy XIX*, French, Uehling, Wettstein (eds.), University of Notre Dame Press
- King, Jeffrey C., 1995, 'Structured Propositions and Complex Predicates', *Nous* 29(4), 516-535
- King, Jeffrey C., 1996, 'Structured Propositions and Sentence Structure', *Journal of Philosophical Logic* 25: 495-521
- Kripke, Saul, 1972, 1980, *Naming and Necessity*, Harvard University Press and Basil Blackwell
- Larson, R and P. Ludlow, 1993, 'Interpreted Logical Forms', *Synthese* 95 305-356
- Lewis, David, 1972, 'General Semantics', in *Semantics of Natural Language*, Davidson and Harman (eds.), Dordrecht, Reidel
- Menzel, Christopher, 1986, 'A Complete, Type-free "Second-Order" Logic and its Philosophical Foundations', Stanford: CSLI Publications Report No. CSLI-86-40.
- Menzel, Christopher, 1993, 'The Proper Treatment of Predication in Fine-Grained Intensional Logic', in *Philosophical Perspectives, 7, Language and Logic*, Ridgeview Publishing Company, Atascadero, CA
- Montague, Richard, 1960, 'On the Nature of Certain Philosophical Entites', *The Monist* 53: 159-94
- Montague, Richard, 1970, 'Pragmatics and Intensional Logic' in *Formal Philosophy: Selected Papers of Richard Montague*, Yale University Press, 1974
- Richard, Mark, 1982, 'Tense, Propositions and Meanings' *Philosophical Studies* 41, 337-351.
- Richard, Mark, 1990, *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them*, Cambridge University Press
- Russell, Bertrand, 1903, *Principles of Mathematics* , second edition, New York, Norton
- Salmon, Nathan, 1986a, *Frege's Puzzle*, MIT Press/Bradford Books
- Salmon, Nathan, 1986b, 'Reflexivity', *Notre Dame Journal of Formal Logic*, 27, 3, 401-429
- Salmon, Nathan, 1989a, 'Illogical Belief', *Philosophical Perspectives, 3, Philosophy of Mind and Action Theory*, Ridgeview Publishing Company, Atascadero, CA
- Salmon, Nathan, 1989b, 'Tense and Singular Propositions', in in *Themes From Kaplan*, Almog, Wettstein, Perry (eds.), Oxford University Press, Oxford 1989
- Soames, Scott, 1985, 'Lost Innocence', *Linguistics and Philosophy* 8, 59-71
- Soames, Scott, 1987, 'Direct Reference, Propositional Attitudes and Semantic Content', *Philosophical Topics*15, 47-87

- Soames, Scott, 1989, ‘Semantics and Semantic Competence’, *Philosophical Perspectives, 3, Philosophy of Mind and Action Theory*, Ridgeview Publishing Company, Atascadero, CA
- Stanley, Jason, 2000, ‘Context and Logical Form’, *Linguistics and Philosophy* 23, 391-434
- Zalta, Edward N., 1983, *Abstract Objects: An Introduction to Axiomatic Metaphysics*, Dordrecht, D. Reidel
- Zalta, Edward N., 1988, *Intensional Logic and the Metaphysics of Intentionality*, Cambridge, MA and London, The MIT Press

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

belief | Frege, Gottlob | language: philosophy of | meaning | mereology | possible worlds | propositions: singular | reference | Russell, Bertrand | semantics | set theory

Jeffrey C. King
jcking@ucdavis.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 22, 1997
Content last modified: August 8, 2001

Stanford Encyclopedia of Philosophy
Notes to Structured Propositions

Notes

1. Of course there are always dissenters. For example, Mark Richard [1982] argues that modal operators operate on sentence meanings (functions from contexts and times to propositions). Thus it would seem that it is sentence meanings that have modal properties for Richard. Further, there is some controversy as to whether propositions are the things that possess "tense properties" (e.g. having been true, etc.) and are the things tense operators operate on. Kaplan [1977] thinks so. Richard [1982] and Salmon [1989b] think not.

2. Strictly, Crimmins seems only to claim that sentences sometimes express propositions (in a context), where the propositions expressed contain constituents that are not supplied by any *overt* expression in the sentences. This does not rule out the possibility that such constituents are contributed by covert (phonologically and inscriptionally unrealized) expressions in the sentence. But there are those who have argued that sentences sometimes express propositions that have as constituents elements that are not supplied by any covert *or* overt expressions in the sentence. See Stanley [2000] for discussion.

3. See, for example, Montague [1960] for a discussion of an approach of this general sort.

4. Such a view is suggested in Montague [1960].

5. The interested reader should also consult Richard [1990] and Crimmins [1992] for approaches in the broadly neo-Russellian (as I use the term) tradition that try to avoid various consequences of the Salmon-Soames approach.

6. Peter Hanks presses this point against those who hold that propositions are n-tuples in an unpublished manuscript.

7. It may strike some as odd that I include the views of Cresswell [1985] in an article on structured *propositions* since he explicitly argues against what he calls "the propositional account of belief". But in so doing, he makes assumptions about propositions that modern structured proposition theorists (e.g. neo-Russellians) would reject, such as that ‘5+7=12’ and ‘12=12’ express the same proposition. And Cresswell's own views have much in common with the views of other (non-structured meaning) structured proposition theorists.

8. Actually, Cresswell eventually settles on sets of world/time pairs as sentence intensions; but I ignore

this complication here.

9. I write here as though ‘that’ for Cresswell has only two meanings: on one, it combines with a sentence to form a name of an intension/set of worlds. On the other, it combines with a sentence to form the name of the "maximally fine grained entity" that can be named by a ‘that’ clause containing the sentence. But in fact, (given sentences of sufficient complexity), there are "intermediate" meanings of ‘that’ that combine with the sentence in question to form names of entities whose grain is "between" intension/sets of worlds and the "maximally fine grained entity". This is why I say above that ‘that’ is *highly* ambiguous on Cresswell's view. Again, since our concern here is with more fine grained entities, I ignore this complication and focus on the maximally fine grained entity that can be named by a ‘that’ clause containing a sentence on Cresswell's view.

10. One can easily use this function to define the intension f of ‘runs’ (function from worlds to extensions) as follows: f is the function from worlds to sets of individuals such that o is in f(w) iff w is in I_r(o), where I_r is Cresswell's "intension" (function from individuals to sets of worlds) for ‘runs’.

11. It would seem that Cresswell needs to say this. For recall that sentences don't express fine grained entities in isolation; it is only when combined with ‘that’ (on certain of its meanings) that a fine grained entity is associated with a sentence-plus-'that’. So it appears that the sentence in isolation (without ‘that’ appended) cannot be true or false in virtue of its association with a fine grained that is in the first instance true of false. For in isolation it is not so associated!

12. The use of variables in expressions like ‘Rxy’ here is perhaps unfortunate. I remind the reader that R is a relation (not a predicate). I put variables in to indicate the "argument places" in the relation. Perhaps underlining (___) and other notation to the same effect (#####) would be better. I urge the reader to mentally substitute such things for variables that indicate argument places in relations.

13. For details about the rest of the functions mentioned, consult Zalta [1988], (especially pages 46-51; 58-61, and the Appendix containing his formal intensional logic).

14. For simplicity, I suppress the modal and tense operators that should appear in the instances of the comprehension schema.

Jeffrey C. King
jcking@ucdavis.edu

First published: August 8, 2001
Content last modified: August 8, 2001

Singular Propositions

Singular propositions (also called ‘Russellian propositions’) are propositions that are about a particular object or individual in virtue of having the object or individual as a *constituent* of the proposition. Alleged examples of singular propositions are the propositions that Mont Blanc is more than 4,000 meters high, that Socrates was wise, and that she [pointing to someone] lives in New York. Singular propositions are to be contrasted with *general* propositions and (what we can call) particularized propositions. The former are propositions that are not about any particular item (as opposed to a class or kind of item) and the latter are propositions that are about particulars or individuals but do not contain those individuals as constituents. Examples of the former are the propositions that most Americans favor a tax cut and that some music is great; examples of the latter are the propositions that the inventor of bifocals was bald and that the tallest spy is a man. The acceptance or rejection of singular propositions lies at the center of many issues in semantics, the philosophy of language, and metaphysics.

Sense and Reference

There are, roughly speaking, two kinds of theories in the philosophy of language and many versions of these theories. There are the Fregean theories that employ some version of Frege's distinction between sense and reference and the Russellian theories that employ some version of Russell's theory of scope and incomplete symbols. (One can, of course, combine the two views in various ways but generally such combinations tend to side with Frege over Russell on the nature of propositions and hence count as Fregean in the present context.) Again roughly speaking, Frege distinguished between the referent of an expression or term and the sense of that expression or term. For example, the referent of the term ‘human’ is the set of humans while the sense of the term might be described as the property of being

homo sapiens (although, strictly speaking, senses are not properties). For complex expressions, such as the inventor of bifocals, the sense of the whole expression is a function of the sense of its parts and the referent of the whole expression is a function of the referent of the parts. This is also true for complete sentences. The proposition that is expressed by a sentence is the sense of the sentence which in turn is a function of the senses of the parts of the sense. Thus for Frege objects are *not* constituents of propositions. Propositions are composed of the senses of terms and expressions that refer to objects, but not the objects themselves.

Russell, on the other hand, rejected senses and attempted to solve the problems that senses were introduced to solve by logical analysis and scope distinctions. The result was that propositions expressed by sentences such as "The inventor of bifocals was bald" did not have senses as constituents. However, they did not have objects as constituents either. Roughly speaking, Russell had propositional functions-- functions from objects to propositions-- in place of senses. Yet at the bottom level such a view seems to require singular propositions. These are the basic or atomic propositions upon which the complex propositions are built. Frege's atomic propositions are composed of senses while Russell's require individuals.

Recently, David Kaplan has argued that essential difference between Frege's and Russell's theories of language is nothing more nor less than the acceptance of singular propositions. If we were to add singular propositions to Frege's theory then with some modifications we could reduce Frege's view to Russell's view. Hence, if Kaplan is correct and there are singular propositions as well, then we need not introduce the complexities of a Fregean theory of sense and we can focus our semantical attention on the simpler Russellian theories.

Reasons for Singular Propositions

The primary reasons offered in favor of accepting singular propositions are semantical or linguistic in nature. (Although David Lewis argues that if we accept that there are individuals and properties and pairs of things we accept, then we must accept that there are singular propositions.) Perhaps the strongest of the semantical reasons for singular propositions are those associated with the development of an adequate theory of demonstratives or indexicals. David Kaplan, for example, has argued that indexical expressions (e.g., ‘I’, ‘now’) are what he calls ‘directly referential’ expressions. An expression is *directly referential* if the expression (relative to a context of use) directly refers to its object and the referent is *not* determined or fixed in virtue of the fact that the object in question is characterized by a descriptive property or a Fregean sense.

Let us assume without argument that there are propositions. We make this assumption knowing that many philosophers have serious doubts about the existence of propositions. However, our current purpose is to ask whether there are any positive reasons that can be given in favor of *singular* propositions even granting that there are propositions. We will use an argument that was originally presented by David Kaplan to show that demonstratives are directly referential expressions (and hence are used to express singular propositions). Suppose David is standing at a table with two men; Charles on

his left and Paul on his right. Paul lives in New Jersey and Charles lives in Illinois. David points to the person on his right and says (at time *t*)

(1) He lives in New Jersey.

David has expressed a proposition that we can label *p* that is about Paul. Is *p* a singular proposition; that is, is it a proposition that has Paul as a constituent? Let us begin by assuming that *p*is *not* a singular proposition. If it is not a singular proposition then it does not have Paul as a constituent, but is about Paul is another way. (If *p* is a singular proposition then it is about Paul by having Paul as a constituent.) Consider then the proposition expressed by

(2) The person on David's right (at *t*) lives in New Jersey ,

which we can label *p**. *p** is a proposition about Paul but it does not have Paul as a constituent. Perhaps *p** (or something similar to *p**) is *p*. However, consider a counterfactual circumstance where David had been tricked into thinking that Paul is to his right and that Charles is to his left. That is, suppose, contrary to the facts, Paul and Charles switched places. In such a case *p** would not be true, but *p* would be true. *p** would not be true because the person on David's right, namely Charles, does not live in New Jersey. *p* on the other hand is still true, for though in the counterfactual circumstance Paul and Charles have switched locations relative to David they still live in their respective states in the counterfactual circumstance.

It is important here to be clear about what we are supposing. We are asking what the truth value of the propositions *p* and *p** *would be* in the described counterfactual circumstance. We are *not* asking what proposition would David express by uttering (1) in the counterfactual circumstance. It is true that the proposition expressed by (1) will differ in different circumstances of utterance or different *contexts*. Paul can utter (1) pointing to Charles or Charles can utter (1) pointing to David. In these different contexts different propositions will be expressed.

Nonetheless since *p* and *p** differ in truth value in the described counterfactual circumstance, *p** cannot be *p*. (It is an axiom of propositional theory that if *p* = *p** then *p* and *p** have the same truth value in all counterfactual circumstances.) A similar argument can be presented for any proposition that is in fact about Paul, but could be about someone else in a different counterfactual circumstance. So, for example, if we replaced (2) with

(3) The person David is pointing to lives in New Jersey,

we will run into the same problem.

One might suppose that one can avoid singular propositions by claiming, along lines similar to what David Lewis has argued, that it is a mistake to place any philosophical importance on the question of what is true in a counterfactual circumstance *simplicter*. Counterfactual circumstances are merely one

parameter among many that are necessary to determine truth. What we should say is that (1) is true relative to a complex or structure that contains a *context* and a *counterfactual (or factual) circumstance* among other parameters. Viewed in this light it appears that (1) and (3) have the same truth value given the same parameters. One result with this way of looking at things is that either propositions are identified with sentences (types) or sets of items containing all the necessary parameters (such as sets of possible worlds). David Lewis accepts the latter and in so doing accepts that there are singular propositions. So Lewis's way of looking at things does not avoid commitment to singular propositions. If one accepts that propositions are sentences then one must give up the view that they can be about particulars. For on such a view the proposition David expresses when he says

(4) I am bored.

is the same proposition that Paul, Charles, or anyone else would express were they to utter (4). So it would be a mistake to say that (4) as uttered by David is about David. But then *p* would not be about Paul as we assumed. There are, of course, other reasons to deny that sentences are propositions that properly belong to a general discussion of propositions. Besides it is not completely clear that there is no philosophical significance in dividing things up the way Kaplan does. After all, there is something rather plausible in Kaplan's suggestion that context is used to help determine which proposition is expressed as opposed to being part of a complex that determines the truth of an already present proposition.

Still, given what we have said thus far, it is possible that Paul is not a constituent of *p* but the singleton set of Paul or the property of being Paul is a constituent of *p*. The reason in favor of Paul over his singleton as a constituent of *p* concerns issues of reference in the philosophy of language. Can we *directly* refer to an individual or property? Can we directly refer to the singleton of Paul? If the answer is yes, then we could run into a problem if we claimed that *p* has Paul's singleton as a constituent. Let a be an expression that directly refers to Paul's singleton. Sets, of course, do not live in states so the proposition expressed by the following is false

(5) a lives in New Jersey.

Yet, the proposition expressed by (5) appears to have the same constituents as *p* if Paul's singleton, as opposed to Paul, is a constituent of *p*. This argument is not conclusive since it depends on the assumption that one can refer directly to an item with the use of some linguistic expression and not via some representative of the item. In the end it may not matter whether Paul or some representative of Paul is a constituent of *p* provided that the representative is an essential representative of Paul (i.e., something that represents Paul in all (and only) counterfactual circumstances where Paul can represent himself).

A Modal Argument against Singular Propositions

Although there are some reasons for thinking that there are singular propositions there are also reasons for thinking that there are no singular propositions. Alvin Plantinga, for example, has raised the

following objection. Consider the proposition:

> (6) *Socrates did not exist.*

While (6) is false, it might have been true (here the number designates the proposition (if there is one) as opposed to the sentence). That is, it is possibly true. If (6) had been true, then (6) would have existed. After all, (6) cannot have the property of being true without existing, so had (6) been true then Socrates would not have existed, but (6) would have existed. But if Socrates is a constituent of (6), then (6) cannot exist without Socrates's existing. Therefore, Socrates is not a constituent of (6).

Basically, Plantinga argues that ordinary objects such as persons cannot be constituents of propositions because the propositions can exist without the individual's existing. There are different replies to Plantinga's argument one can make depending on the metaphysical position one takes with respect to modality. If one is a possibilist then one makes a distinction between something's being actual and something's existing (or subsisting or having being of some sort). So Socrates can be a constituent of a proposition even in circumstances where Socrates is not an actual object. On the other hand, if one is an actualist (and accepts singular propositions) then one must deny Plantinga's claim that (6) can exist without Socrates's existing. Here one must argue that (6)'s being possibly true does not imply that (6) can be true without Socrates's existing. As an analogy consider John's assertion "It is possible I do not exist." John can describe or represent a circumstance or world where John does not exist. So, too, (6) can represent a circumstance or world where (6) does not exist. For (6) to be possibly true is for (6) to represent a circumstance or world that could obtain. It is not required that (6) be a part of the world or circumstance that (6) represents any more than it is required that John be a part of the world that he describes with his assertions.

Temporal Problems for Singular Propositions

Another objection that one can raise with respect to singular propositions concerns a fundamental problem of combining so-called abstract objects (such as propositions) with ordinary individuals. Ordinary individuals change over time including coming into and going out of existence. Propositions, on the other hand, are usually taken to be eternal objects--things that do not change over time. Consider then the following proposition:

> (7) *Socrates exists.*

Does (7) exist? Socrates is long gone; he no longer exists. But if (7) does not exist how can it be false? Moreover, if (7) does exist then exactly what age is the constituent of (7)? 21? 37? It seems a bit absurd to say that the age of a constituent of a proposition is thus and so, yet if Socrates himself is a constituent of (7), then he must be a certain age. No human person exists without being a certain age.

Again the reply that one makes to the temporal modal objection is like the reply that one makes to the

alethic modal objection in that it depends on one's metaphysical views concerning time and individuals. If, for example, one accepts the 4D view of objects then as far as age (and other issues involving change of existing objects) goes there is no problem. The object that is a constituent of a proposition is a complete object in that all the temporal stages or parts of the object in question are involved. So part of Socrates is 21 and part is 37.

On the other hand, if one holds to a 3D view where the object is wholly present at each time that it exists, then matters become a bit more complex. There are different ways to go. For example, one could hold that at each time, *t*, at which Socrates existed there is a singular proposition involving Socrates and *t* (and perhaps the property of existence if it is a property). On one such view there is no such proposition as (7). So the question of the age (and other features of changing existing objects) will depend on the time involved in the proposition under consideration. On another view (7) expresses the conjunction of all such propositions, hence the question of age does not arise.

When we ask (on either view) how can an object that no longer exists be a constituent of a proposition, we need to consider the various metaphysical views concerning ordinary individuals. If we can think of (or refer to) Socrates even though Socrates does not exist, then Socrates can be a constituent of a proposition though Socrates does not exist. What then shall we say of the proposition that Socrates exists? Does it exist or not? If it does not exist, then that does not prevent us from thinking of it (as with Socrates). On the other hand if we are prevented from thinking of Socrates because he no longer exists, then there are no singular propositions about Socrates. But that does not prevent there from being singular propositions about currently existing objects. One must simply give up the view that propositions are eternal.

The question of whether there are singular propositions, like the general question of whether there are propositions at all, has not been settled. Different philosophers take different positions on this issue. However, there remains a major advantage for accepting some form of singular propositions, namely, on such a view is clear how propositions represent or describe the world.

Bibliography

- Arthur, Sullivan, 1998, "Singular Propositions and Singular Thoughts" *Norte Dame Journal of Formal Logic* 39 1998:114-127
- Braun, David, 1993 "Empty Names" *Nous* 25 1993:449-469
- Cartwright, Richard, 1997, "Singular Propositions" *Canadian Journal of Philosophy* Supplementary Vol. 23 1997 67-84
- Davidson, Matthew, 2000, "Direct Reference and Singular Propositions" *American Philosophical Quarterly* 37 2000:285-300
- Fitch, G. W., 1988, "The Nature of Singular Propositions," in *Philosophical Analysis*, David Austin (ed.), Dordrecht: Kluwer, 1988: 281-297
- Frege, Gottlob, 1892, "On Sense and Reference," in *Philosophical Writings,* Peter Geach and Max Black (eds.), Oxford: Blackwell, 1952: 56-78

- Hazen, A. P., 1995, "On Quantifying Out" *The Journal of Philosophical Logic* 1995 24:291-319
- Jespersen, Bjorn, 2000, "Singular Propositional Constructions" in *The Logica Yearbook 1999* 2000
- Kaplan, David, 1975, "How to Russell a Frege-Church," *The Journal of Philosphy* 72: 716-729
- Kaplan, David, 1977, "Demonstratives," in *Themes from Kaplan*, J. Almog, J. Perry, and H. Wettstein (eds.), New York: Oxford University Press, 1989: 481-564
- King, Jeff, 1996, "Structured Propositions and Sentence Structure" *Journal of Philosophical Logic* 25 1996: 155-179
- Menzel, Chris, 1993 "Singular Propositions and Modal Logic" *Phiosophical Topics* 21
- Plantinga, Alvin, 1983, "On Existentialism," *Philosophical Studies*, 44: 1-20
- Russell, Betrand, 1905, "On Denoting," in *Logic and Knowledge*, Robert Marsh (ed.), London: Allen and Unwin, 1956: 41-56
- Salmon, Nathan, 1989, "Tense and Singular Propositions," in *Themes from Kaplan*, J. Almog, J. Perry, and H. Wettstein (eds.), New York: Oxford University Press, 1989: 331-392
- Salmon, Nathan, 1998, "Nonexistance" *Nous* 32 1998:277-319
- Zalta, Edward N., 1989 "Singular Propositions, Abstract Constituents, and Propositional Attitudes" in *Themes from Kaplan*, J. Almog, J. Perry, and H. Wettstein (eds), New York: Oxford University Press, 1989

Other Internet Resources

- The Metaphysics Research Lab Web Pages

Related Entries

actualism | atomism: logical | denotation | existence | facts | Frege, Gottlob | indexicals | language: philosophy of | logic: classical | names | propositions: structured | Russell, Bertrand | semantics | sense/reference distinction | time

Greg Fitch
Arizona State University
fitch@asu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 19, 1997
Content last modified: May 13, 2002

Set Theory

Set Theory is the mathematical science of the infinite. It studies properties of sets, abstract objects that pervade the whole of modern mathematics. The language of set theory, in its simplicity, is sufficiently universal to formalize all mathematical concepts and thus set theory, along with Predicate Calculus, constitutes the true Foundations of Mathematics. As a mathematical theory, Set Theory possesses a rich internal structure, and its methods serve as a powerful tool for applications in many other fields of Mathematics. Set Theory, with its emphasis on consistency and independence proofs, provides a gauge for measuring the consistency strength of various mathematical statements. There are four main directions of current research in set theory, all intertwined and all aiming at the ultimate goal of the theory: to describe the structure of the mathematical universe. They are: inner models, independence proofs, large cardinals, and descriptive set theory. See the relevant sections in what follows.

---

1. The Essence of Set Theory

The objects of study of Set Theory are *sets*. As sets are fundamental objects that can be used to define all other concepts in mathematics, they are not defined in terms of more fundamental concepts. Rather, sets are introduced either informally, and are understood as something self-evident, or, as is now standard in modern mathematics, axiomatically, and their properties are postulated by the appropriate formal axioms.

The language of set theory is based on a single fundamental relation, called *membership*. We say that A is a member of B (in symbols $A \in B$), or that the set B contains A as its element. The understanding is that a set is determined by its elements; in other words, two sets are deemed equal if they have exactly the same elements. In practice, one considers sets of numbers, sets of points, sets of functions, sets of some other sets and so on. In theory, it is not necessary to distinguish between objects that are members and objects that contain members -- the only objects one needs for the theory are sets. See the supplement

Basic Set Theory

for further discussion.

Using the membership relation one can derive other concepts usually associated with sets, such as unions and intersections of sets. For example, a set C is the union of two sets A and B if its members are exactly those objects that are either members of A or members of B. The set C is uniquely determined, because we have specified what its elements are. There are more complicated operations on sets that can be defined in the language of set theory (i.e. using only the relation $\in$), and we shall not concern ourselves with those. Let us mention another operation: the (unordered) *pair* $\{A,B\}$ has as its elements exactly the sets $A$and B. (If it happens that $A=B$, then the "pair" has exactly one member, and is called a *singleton* $\{A\}$.) By combining the operations of union and pairing, one can produce from any finite list of sets the set that contains these sets as members: $\{A,B,C,D,...,K,L,M\}$. We also mention the *empty set*, the set that has no elements. (The empty set is uniquely determined by this property, as it is the only set that has no elements - this is a consequence of the understanding that sets are determined by their elements.)

When dealing with sets informally, such operations on sets are self-evident; with the axiomatic approach, it is postulated that such operations can be applied: for instance, one postulates that for any sets A and B, the set $\{A,B\}$ exists. In order to endow set theory with sufficient expressive power one needs to postulate more general construction principles than those alluded to above. The guiding principle is that any objects that can be singled out by means of the language can be collected into a set. For instance, it is desirable to have the "set of all integers that are divisible by number 3," the "set of all straight lines in the Euclidean plane that are parallel to a given line", the "set of all continuous real functions of two real variables" etc. Thus one is tempted to postulate that given any property P, there exists a set whose members are exactly all the sets that have property P. As we shall see below, such an assumption is logically inconsistent, and the accepted construction principles are somewhat weaker than such a postulate.

One of the basic principles of set theory is the existence of an infinite set. The concept can be formulated precisely in the language of set theory, using only the membership relation, and the definition captures the accepted meaning of "infinite". See the supplement on

Basic Set Theory

for further discussion. Using the basic construction principles, and assuming the existence of infinite sets, one can *define* numbers, including integers, real numbers and complex numbers, as well as functions, functionals, geometric and topological concepts, and all objects studied in mathematics. In this sense, set theory serves as *Foundations of Mathematics*. The significance of this is that all questions of provability (or unprovability) of mathematical statements can be in principle reduced to formal questions of formal derivability from the generally accepted axioms of Set Theory.

While the fact that all of mathematics can be reduced to a formal system of set theory is significant, it would hardly be a justification for the study of set theory. It is the internal structure of the theory that makes it worthwhile, and it turns out that this internal structure is enormously complex and interesting. Moreover, the study of this structure leads to significant questions about the nature of the mathematical universe.

The fundamental concept in the theory of infinite sets is the *cardinality* of a set. Two sets *A* and *B* have the *same cardinality* if there exists a mapping from the set *A* onto the set *B* which is *one-to-one*, that is, it assigns each element of *A* exactly one element of *B*. It is clear that when two sets are finite, then they have the same cardinality if and only if they have the same number of elements. One can extend the concept of the "number of elements" to arbitrary, even infinite, sets. It is not apparent at first that there might be infinite sets of different cardinalities, but once this becomes clear, it follows quickly that the structure so described is rich indeed.

2. Origins of Set Theory

The birth of Set Theory dates to 1873 when Georg Cantor proved the uncountability of the real line. (One could even argue that the exact birthdate is December 7, 1873, the date of Cantor's letter to Dedekind informing him of his discovery.) Until then, no one envisioned the possibility that infinities come in different sizes, and moreover, mathematicians had no use for "actual infinity." The arguments using infinity, including the Differential Calculus of Newton and Leibniz, do not require the use of infinite sets, and infinity appears only as "a manner of speaking", to paraphrase Friedrich Gauss. The fact that the set of all positive integers has a proper subset, like the set of squares {1, 4, 9, 16, 25,...} of the same cardinality (using modern terminology) was considered somewhat paradoxical (this had been discussed at length by Galileo among others). Such apparent paradoxes prevented Bernhard Bolzano in 1840s from developing set theory, even though some of his ideas are precursors of Cantor's work. (It should be mentioned that Bolzano, an accomplished mathematician himself, coined the word *Menge* (= set) that Cantor used for objects of his theory.)

Motivation for Cantor's discovery of Set Theory came from his work on Fourier series (which led him to introduce *ordinal numbers*) and on trancendental numbers. Real numbers that are solutions of polynomial equations with integer coefficients are called algebraic, and the search was on for numbers that are not algebraic. A handful of these, called transcendental numbers, was discovered around that time, and a question arose how rare such numbers are. What Cantor did was to settle this question in an unexpected way, showing in one fell swoop that transcendental numbers are plentiful indeed. His famous proof went

as follows: Let us call an infinite set *A countable*, if its elements can be enumerated; in other words, arranged in a sequence indexed by positive integers: $a(1)$, $a(2)$, $a(3)$, … , $a(n)$, … . Cantor observed that many infinite sets of numbers are countable: the set of all integers, the set of all rational numbers, and also the set of all algebraic numbers. Then he gave his ingeneous diagonal argument that proves, by contradiction, that the set of all real numbers is *not* countable. A consequence of this is that there exists a multitude of transcendental numbers, even though the proof, by contradiction, does not produce a single specific example. See the supplement on

Basic Set Theory

for further discussion.

Cantor's discovery of uncountable sets led him to the subsequent development of ordinal and cardinal numbers, with their underlying order and arithmetic, as well as to a plethora of fundamental questions that begged to be answered (such as the Continuum Hypothesis). After Cantor, mathematics has never been the same.

3. The Continuum Hypothesis

As the Continuum Hypothesis has been the most famous problem in Set Theory, let me explain what it says. The smallest infinite cardinal is the cardinality of a countable set. The set of all integers is countable, and so is the set of all rational numbers. On the other hand, the set of all real numbers is uncountable, and its cardinal is greater than the least infinite cardinal. A natural question arises: is this cardinal (*the continuum*) the very next cardinal. In other words, is it the case that there are no cardinals between the countable and the continuum? As Cantor was unable to find any set of real numbers whose cardinal lies strictly between the countable and the continuum, he conjectured that the continuum is the next cardinal: the Continuum Hypothesis. Cantor himself spent most of the rest of his life trying to prove the Continuum Hypothesis and many other mathematicians have tried too. One of these was David Hilbert, the leading mathematician of the last decades of the 19th century. At the World Congress of Mathematicians in Paris in 1900 Hilbert presented a list of major unsolved problems of the time, and the Continuum Hypothesis was the very first problem on Hilbert's list.

Despite the effort of a number of mathematicians, the problem remained unsolved until 1963, and it can be argued that in some sense the problem is still unsolved. See Section 7 on Independence Proofs.

4. Axiomatic Set Theory

In the years following Cantor's discoveries, development of Set Theory proceeded with no particular concern about how exactly sets should be defined. Cantor's informal "definition" was sufficient for proofs in the new theory, and the understanding was that the theory can be formalized by rephrasing the

informal definition as a *system of axioms*. In the early 1900s it became clear that one has to state precisely what basic assumptions are made in Set Theory; in other words, the need has arisen to axiomatize Set Theory. This was done by Ernst Zermelo, and the immediate reasons for his axioms were twofold. The first one was the discovery of a paradox in Set Theory. This paradox is referred to as *Russell's Paradox.* Consider the "set" S of all sets that are not an element of itself. If one accepts the principle that all such sets can be collected into a set, then S should be a set. It is easy to see however that this leads to a contradiction (is the set S an element of itself?)

Russell's Paradox can be avoided by a careful choice of construction principles, so that one has the expressive power needed for usual mathematical arguments while preventing the existence of paradoxical sets. See the supplement on

Zermelo-Fraenkel Set Theory

for further discussion. The price one has to pay for avoiding inconsistency is that some "sets" do not exist. For instance, there exists no "universal" set (the set of all sets), no set of all cardinal numbers, etc.

The other reason for axioms was more subtle. In the course of development of Cantor's theory of cardinal and ordinal numbers a question was raised whether every set can be provided with a certain structure, called *well-ordering* of the set. Zermelo proved that indeed every set can be well-ordered, but only after he introduced a new axiom that did not seem to follow from the other, more self-evident, principles. His *Axiom of Choice* has become a standard tool of modern mathematics, but not without numerous objections of some mathematicians and discussions in both mathematical and philosophical literature. The history of the Axiom of Choice bears strong resemblance to that of the other notorious axiom, Euclid's Fifth Postulate.

5. The Axiom of Choice

The Axiom of Choice states that for every set of mutually disjoint nonempty sets there exists a set that has exactly one member common with each of these sets. For instance, let S be a set whose members are mutually disjoint finite sets of real numbers. We can *choose* in each $X \in S$ the smallest number, and thus form a set that has exactly one member in common with each $X \in S$. What is not self-evident is whether we can make a choice every time, simultaneously for infinitely many sets X, regardless what these abstract sets are. The Axiom of Choice, which postulates the existence of a certain set (*the choice set*) without giving specific instructions how to construct such a set, is of different nature than the other axioms, which all formulate certain construction principles for sets. It was this nonconstructive nature of the Axiom of Choice that fed the controversy for years to come.

An interesting application of the Axiom of Choice is the Banach-Tarski Paradox that states that the unit ball can be partitioned into a finite number of disjoint sets which then can be rearranged to form *two* unit balls. This is of course a paradox only when we insist on visualizing abstract sets as something that exists

in the physical world. The sets used in the Banach-Tarski Paradox are not physical objects, even though they do exist in the sense that their existence is proved from the axioms of mathematics (including the Axiom of Choice).

The legitimate question is whether the Axiom of Choice is consistent, that is whether it cannot be refuted from the other axioms. (Notice the similarity with the non Euclidean geometry.) This question was answered by Gödel, and eventually the role of the Axiom of Choice has been completely clarified. See Section 7 on Independence Proofs.

6. Inner Models

In the 1930s, Gödel stunned the mathematical world by discovering that mathematics is incomplete. His Incompleteness Theorem states that every axiomatic system that purports to describe mathematics as we know it must be incomplete, in the sense that one can find a true statement expressible in the system that cannot be formally proved from the axioms. In view of this result one must consider the possibility that a mathematical conjecture that resists a proof might be an example of such an unprovable statement, and Gödel immediately embarked on the project of showing that the Continuum Hypothesis might be undecidable in the axiomatic set theory.

Several years after proving the Incompleteness Theorem, Gödel proved another groundbreaking result: he showed that both the Axiom of Choice and the Continuum Hypothesis are consistent with the axioms of set theory, that is that neither can be refuted by using those axioms. This he achieved by discovering a model of set theory in which both the Axiom of Choice and the Continuum Hypothesis are true.

Gödel’s model *L* of “constructible sets” has since served as a blueprint for building so-called *inner models*. These models form a hierarchy, corresponding to the hierarchy of large cardinals (see Section 8), and provide a glimpse into the as yet hidden structure of the mathematical universe. The advances in Inner Model Theory that have been made in the recent past owe much to the work of Ronald Jensen who introduced the study of the fine structure of constructible sets.

7. Independence Proofs

In 1963, Paul Cohen proved independence of the Axiom of Choice and of the Continuum Hypothesis. This he did by applying the *method of forcing* that he invented and constructing first a model of set theory (with the axiom of choice) in which the Continuum Hypothesis fails, and then a model of set theory in which the Axiom of Choice fails. Together with Gödel’s models, these models show that the Axiom of Choice can neither be proved nor refuted from the other axioms, and that the Continuum Hypothesis can neither be proved nor refuted from the axioms of set theory (including the Axiom of Choice).

Cohen's method proved extremely fruitful and led first to the solution of a number of outstanding problems (Suslin's Problem, the Lebesgue measurability Problem, Borel's Conjecture, Kaplansky's Conjecture, Whitehead's Problem and so on) and soon has become one of the cornerstones of modern set theory. The technique of forcing has to date been applied by hundreds of authors of numerous articles and has enormously advanced our knowledge of Foundations of Mathematics. Along with the theory of large cardinals it is used to gauge the consistency strength of mathematical statements.

8. Large Cardinals

In 1930, while working on the Measure Problem, Stanislaw Ulam discovered an important phenomenon: Assuming that a certain mathematical statement about "small sets" (such as sets of real numbers) is true, one can prove the existence of sets of enormous size (*inaccessible*). This phenomenon has become more apparent after Dana Scott's celebrated result (1961) that mesurable cardinals do not exist in *L*. Suddenly, large cardinals such as inaccessible, measurable, supercompact etc. have become the main focus of attention of set theorists. What emerged is a hierarchy of properties of infinite sets, the Large Cardinal Theory, that appears to be the basis for the structure of the set theoretical universe. Large cardinal axioms (also referred to as axioms of *strong infinity*) form a hierarchy whereby a stronger axiom not only implies a weaker axiom but also proves its consistency. To date there are scores of examples of mathematical statements whose consistency strength can be precisely calculated in terms of the hierarchy of large cardinals.(For instance, a negative solution of the Singular Cardinal Problem corresponds to a large cardinal axiom between measurabily and supercompactness.)

Since the pioneering work of Ronald Jensen, Large Cardinal Theory has been closely tied with Inner Model Theory. It turns out that for each large cardinal axiom at lower levels of the hierarchy one can find an appropriate inner model. These inner models shed additional light on the structure of the universe by employing methods of Descriptive Set Theory.

9. Descriptive Set Theory

Descriptive Set Theory traces its origins to the theory of integration by Henri Lebesgue at the beginning of 20th century. Investigations into Borel sets of real numbers led to the theory of *projective* sets, and more generally, the theory of definable sets of real numbers. Following Gödel's work, it became apparent that many natural questions in Descriptive Set Theory are undecidable in axiomatic set theory. This was further confirmed by a proliferation of independence results following Cohen's invention of the forcing method.

Modern Descriptive Set Theory revolves mostly around the powerful method using infinite games. The branch of Descriptive Set Theory known as *Determinateness*, developed by D. A. Martin, Robert Solovay and others, brought together methods of, among others, Recursion Theory and Large Cardinal Theory and has been very successful in describing the structure of definable sets. More importantly, Descriptive Set Theory provides strong evidence for the large cardinal axioms.

Bibliography

- Cantor, G., 1932, *Gesammelte Abhandlungen*, Berlin: Springer-Verlag.
- Ulam, S., 1930, 'Zur Masstheorie in der allgemeinen Mengenlehre', *Fund. Math.*, 16, 140-150.
- Gödel, K., 1940, 'The consistency of the axiom of choice and the generalized continuum hypothesis', *Ann. Math. Studies*, 3.
- Scott, D., 1961, 'Measurable cardinals and constructible sets', *Bull. Acad. Pol. Sci.*, 9, 521-524.
- Cohen, P., 1966, *Set theory and the continuum hypothesis*, New York: Benjamin.
- Jensen, R., 1972, 'The fine structure of the constructible hierarchy', *Ann. Math. Logic*, 4, 229-308.
- Martin, D. and Steel, J., 1989, 'A proof of projective determinacy', *J. Amer. Math. Soc.*, 2, 71-125.
- Hrbacek, K. and Jech, T., 1999, *Introduction to Set Theory*, New York: Marcel Dekker, Inc.

Other Internet Resources

- Set Theory, maintained by Jean Larson (Mathematics, University of Florida)
- Articles by J.J. O'Connor and E.F. Robertson, in *The MacTutor History of Mathematics* archive, (Mathematics, University of St. Andrews):
  - "A History of Set Theory"
  - "Georg Ferdinand Ludwig Philipp Cantor"
  - "Paul Joseph Cohen"
  - "Kurt Gödel"
  - "Ernst Friedrich Ferdinand Zermelo"
  - "Bernard Placidus Johann Nepomuk Bolzano"
- A Homepage for the Axiom of Choice, maintained by Eric Schechter (Mathematics, Vanderbilt University)
- Gödel's Incompleteness Theorem, maintained by Dale Myers (Mathematics, University of Hawaii)

[Please contact the author with suggestions.]

Related Entries

frege=logic | logic: classical | proof theory | Russell's paradox

Mathematical Institute of
The Academy of Sciences of Czech Republic
jech@math.cas.cz

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 10, 2002
Content last modified: July 10, 2002

Stanford Encyclopedia of Philosophy
Supplement to Set Theory

Basic Set Theory

The following basic facts are excerpted from “Introduction to Set Theory,” Third Edition, by Karel Hrbacek and Thomas Jech, published by Marcel Dekker, Inc., New York 1999.

---

1. Ordered Pairs

We begin by introducing the notion of the *ordered pair*. If a and b are sets, then the *unordered pair* $\{a, b\}$ is a set whose elements are exactly a and b. The “order” in which a and b are put together plays no role; $\{a, b\} = \{b, a\}$. For many applications, we need to pair a and b in a way making possible to “read off” which set comes “first” and which comes “second.” We denote this *ordered pair* of a and b by (a, b); a is the *first coordinate* of the pair (a, b), b is the *second coordinate*.

As any object of our study, the ordered pair has to be a set. It should be defined in such a way that two ordered pairs are equal if and only if their first coordinates are equal and their second coordinates are equal. This guarantees in particular that $(a, b) \neq (b,a)$ if $a \neq b$.

> **Definition.** $(a, b) = \{\{a\}, \{a, b\}\}$.

If $a \neq b$, (a, b) has two elements, a singleton $\{a\}$ and an unordered pair $\{a, b\}$. We find the first coordinate by looking at the element of $\{a\}$. The second coordinate is then the other element of $\{a, b\}$. If $a = b$, then $(a, a) = \{\{a\}, \{a,a\}\} = \{\{a\}\}$ has only one element. In any case, it seems obvious that both coordinates can be uniquely “read off” from the set (a, b). We make this statement precise in the

following theorem.

> **Theorem.** $(a, b) = (a', b')$ if and only if $a = a'$ and $b = b'$.

> *Proof*. If $a = a'$ and $b = b'$, then, of course, $(a, b) = \{\{a\}, \{a, b\}\} = \{\{a'\}, \{a', b'\}\} = (a', b')$. The other implication is more intricate. Let us assume that $\{\{a\}, \{a, b\}\} = \{\{a'\}, \{a', b'\}\}$. If $a \neq b$, $\{a\} = \{a'\}$ and $\{a, b\} = \{a', b'\}$. So, first, $a = a'$ and then $\{a, b\} = \{a, b'\}$ implies $b = b'$. If $a = b$, $\{\{a\}, \{a, a\}\} = \{\{a\}\}$. So $\{a\} = \{a'\}$, $\{a\} = \{a', b'\}$, and we get $a = a' = b'$, so $a = a'$ and $b = b'$ holds in this case, too. □

With ordered pairs at our disposal, we can define *ordered triples*

$$(a, b, c) = ((a, b), c),$$

ordered quadruples

$$(a, b, c, d) = ((a, b, c), d),$$

and so on. Also, we define ordered “*one-tuples*"

$$(a) = a.$$

2. Relations

A binary relation is determined by specifying all ordered pairs of objects in that relation; it does not matter by what property the set of these ordered pairs is described. We are led to the following definition.

> **Definition.** A set R is a *binary relation* if all elements of R are ordered pairs, i.e., if for any $z \in R$ there exist x and y such that $z = (x, y)$.

It is customary to write xRy instead of $(x, y) \in R$. We say that x *is in relation* R *with* y if xRy holds.

The set of all x which are in relation R with some y is called the *domain* of R and denoted by “dom R.” So dom $R = \{x \mid$ there exists y such that $xRy\}$. dom R is the set of all first coordinates of ordered pairs in R.

The set of all y such that, for some x, x is in relation R with y is called the *range* of R, denoted by “ran R.” So ran $R = \{y \mid$ there exists x such that $xRy\}$.

3. Functions

Function, as understood in mathematics, is a procedure, a rule, assigning to any object *a* from the domain of the function a unique object *b*, the value of the function at *a*. A function, therefore, represents a special type of relation, a relation where every object *a* from the domain is related to precisely one object in the range, namely, to the value of the function at *a*.

> **Definition.** A binary relation F is called a *function* (or *mapping*, *correspondence*) if aFb_1 and aFb_2 imply $b_1 = b_2$ for any a, b_1, and b_2. In other words, a binary relation F is a function if and only if for every a from dom F there is exactly one b such that aFb. This unique b is called the *value of F at a* and is denoted $F(a)$ or F_a. [$F(a)$ is not defined if $a \notin$ dom F.] If F is a function with dom $F = A$ and ran $F \subseteq B$, it is customary to use the notations $F : A \mapsto B$, $\langle F(a) \mid a \in A \rangle$, $\langle F_a \mid a \in A \rangle$, $\langle F_a \rangle_{a \in A}$ for the function F. The range of the function F can then be denoted $\{F(a) \mid a \in A\}$ or $\{F_a\}_{a \in A}$.

The Axiom of Extensionality can be applied to functions as follows.

> **Lemma.** Let F and G be functions. $F = G$ if and only if dom F = dom G and $F(x) = G(x)$ for all $x \in$ dom F.

A function f is called *one-to-one* or *injective* if $a_1 \in$ dom f, $a_2 \in$ dom f, and $a_1 \neq a_2$ implies $f(a_1) \neq f(a_2)$. In other words if $a_1 \in$ dom f, $a_2 \in$ dom f, and $f(a_1) = f(a_2)$, then $a_1 = a_2$.

4. Natural Numbers

In order to develop mathematics within the framework of the axiomatic set theory, it is necessary to define natural numbers. We all know natural numbers intuitively: 0, 1, 2, 3, ..., 17, …, 324, etc., and we can easily give examples of sets having zero, one, two, or three elements.

To define number 0, we choose a representative of all sets having no elements. But this is easy, since there is only one such set. We define $0 = \emptyset$. Let us proceed to sets having one element (singletons): $\{\emptyset\}$, $\{\{\emptyset\}\}$, $\{\{\emptyset, \{\emptyset\}\}\}$; in general, $\{x\}$. How should we choose a representative? Since we already defined one particular object, namely 0, a natural choice is $\{0\}$. So we define

$$1 = \{0\} = \{\emptyset\}.$$

Next we consider sets with two elements: $\{\emptyset, \{\emptyset\}\}$, $\{\{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$, $\{\{\emptyset\}, \{\{\emptyset\}\}\}$, etc. By now, we have defined 0 and 1, and $0 \neq 1$. We single out a particular two-element set, the set whose elements are the previously defined numbers 0 and 1:

$$2 = \{0,1\} = \{\emptyset, \{\emptyset\}\}.$$

It should begin to be obvious how the process continues:

> $3 = \{0, 1, 2\} = \{\emptyset, \{\emptyset\}, \{\emptyset,\{\emptyset\}\}\}$
> $4 = \{0, 1, 2, 3\} = \{\emptyset,\{\emptyset\}, \{\emptyset,\{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\}$
> $5 = \{0, 1, 2, 3, 4\}$etc.

The idea is simply to define a natural number n as the set of all smaller natural numbers: $\{0, 1, \ldots, n - 1\}$. In this way, n is a particular set of n elements.

This idea still has a fundamental deficiency. We have defined 0, 1, 2, 3, 4, and 5 and could easily define 17 and—not so easily—324. But no list of such definitions tells us what a natural number is in general. We need a statement of the form: A set n is a natural number if …. We cannot just say that a set n is a natural number if its elements are all the smaller natural numbers, because such a “definition” would involve the very concept being defined.

Let us observe the construction of the first few numbers again. We defined $2 = \{0, 1\}$. To get 3, we had to adjoin a third element to 2, namely, 2 itself:

> $3 = 2 \cup \{2\} = \{0, 1\} \cup \{2\}$.

Similarly,

> $4 = 3 \cup \{3\} = \{0, 1, 2\} \cup \{3\}$,
> $5 = 4 \cup \{4\}$,etc.

Given a natural number n, we get the “next” number by adjoining one more element to n, namely, n itself. The procedure works even for 1 and 2: $1 = 0 \cup \{0\}$, $2 = 1\cup\{1\}$, but, of course, not for 0, the least natural number.

These considerations suggest the following.

> **Definition.** The *successor* of a set x is the set $S(x) = x \cup \{x\}$.

Intuitively, the successor $S(n)$ of a natural number n is the “one bigger” number $n + 1$. We use the more suggestive notation n+1 for S(n) in what follows. We later define addition of natural numbers (using the notion of successor) in such a way that $n + 1$ indeed equals the sum of n and 1. Until then, it is just a notation, and no properties of addition are assumed or implied by it.

We can now summarize the intuitive understanding of natural numbers as follows:

a. 0 is a natural number.

b. If n is a natural number, then its successor $n + 1$ is also a natural number.
c. All natural numbers are obtained by application of (a) and (b), i.e., by starting with 0 and repeatedly applying the successor operation: 0, 0 + 1 = 1, 1 + 1 = 2, 2 + 1 = 3, 3 + 1 = 4, 4 + 1 = 5, …etc.

Definition. A set I is called *inductive* if

1. $0 \in I$.
2. If $n \in I$, then $(n + 1) \in I$.

An inductive set contains 0 and, with each element, also its successor. According to (c), an inductive set should contain all natural numbers. The precise meaning of (c) is that the set of natural numbers is an inductive set which contains no other elements but natural numbers, i.e., it is the *smallest* inductive set. This leads to the following definition.

Definition. *The set of all natural numbers* is the set

$$\mathbb{N} = \{x \mid x \in I \text{ for every inductive set } I\}.$$

The elements of $\mathbb{N}$ are called *natural numbers*. Thus a set x is a natural number if and only if it belongs to every inductive set.

5. Cardinality of Sets

From the point of view of pure set theory, the most basic question about a set is: How many elements does it have? It is a fundamental observation that we can define the statement “sets A and B have the same number of elements” without knowing anything about numbers.

Definition. Sets A and B have *the same cardinality* if there is a one-to-one function f with domain A and range B. We denote this by $|A| = |B|$.

Definition. The cardinality of A is less than or equal to the cardinality of B (notation: $|A| \leq |B|$) if there is a one-to-one mapping of A into B.

Notice that $|A| \leq |B|$ means that $|A| = |C|$ for some subset C of B. We also write $|A| < |B|$ to mean that $|A| \leq |B|$ and not $|A| = |B|$, i.e., that there is a one-to-one mapping of A onto a subset of B, but there is no one-to-one mapping of A onto B.

Lemma.

1. If $|A| \leq |B|$ and $|A| = |C|$, then $|C| \leq |B|$.

2. If $|A| \preceq |B|$ and $|B| = |C|$, then $|A| \preceq |C|$.
3. $|A| \preceq |A|$.
4. If $|A| \preceq |B|$ and $|B| \preceq |C|$, then $|A| \preceq |C|$.

Cantor-Bernstein Theorem. If $|X| \preceq |Y|$ and $|Y| \preceq |X|$, then $|X| = |Y|$.

6. Finite Sets

Finite sets can be defined as those sets whose size is a natural number.

> **Definition.** A set S is *finite* if it has the same cardinality as some natural number $n \in \mathbb{N}$. We then define $|S| = n$ and say that *S has n elements*. A set is *infinite* if it is not finite.

7. Countable Sets

> **Definition.** A set S is *countable* if $|S| = |\mathbb{N}|$. A set S is *at most countable* if $|S| \preceq |\mathbb{N}|$.

Thus a set S is countable if there is a one-to-one mapping of $\mathbb{N}$ onto S, that is, if S is the range of an infinite one-to-one sequence.

> **Theorem.** An infinite subset of a countable set is countable.
>
> *Proof*. Let A be a countable set, and let $B \subseteq A$ be infinite. There is an infinite one-to-one sequence $\langle a_n \rangle_{n=0}^{\infty}$, whose range is A. We let $b_0 = a_{k_0}$, where k_0 is the least k such that $a_k \in B$. Having constructed b_n, we let $b_{n+1} = a_{k_{n+1}}$, where k_{n+1} is the least k such that $a_k \in B$ and $a_k \neq b_i$ for every $i \leq n$. Such k exists since it is easily seen that $B = \{b_n \mid n \in \mathbb{N}\}$ and that $\langle b_n \rangle_{n=0}^{\infty}$ is one-to-one. Thus B is countable. □
>
> **Corollary.** A set is at most countable if and only if it is either finite or countable.

The range of an infinite one-to-one sequence is countable. If $\langle a_n \rangle_{n=0}^{\infty}$ is an infinite sequence which is not one-to-one, then the set $\{a_n\}_{n=0}^{\infty}$ may be finite (e.g., this happens if it is a constant sequence). However, if the range is infinite, then it is countable.

> **Theorem.** The range of an infinite sequence $\langle a_n \rangle_{n=0}^{\infty}$ is at most countable, i.e., either finite or countable. (In other words, the image of a countable set under any mapping is at most countable.)

Proof. By recursion, we construct a sequence $\langle b_n \rangle$ (with either finite or infinite domain) which is one-to-one and has the same range as $\langle a_n \rangle_{n=0}^{\infty}$. We let $b_0 = a_0$, and, having constructed b_n, we let $b_{n+1} = a_{k_{n+1}}$, where k_{n+1} is the least k such that $a_k \neq b_i$ for all $i \leq n$. (If no such k exists, then we consider the finite sequence $\langle b_i \mid i \leq n \rangle$.) The sequence $\langle b_i \rangle$ thus constructed is one-to-one and its range is $\{a_n\}_{n=0}^{\infty}$. □

One should realize that not all properties of size carry over from finite sets to the infinite case. For instance, a countable set S can be decomposed into two disjoint parts, A and B, such that $|A| = |B| = |S|$; that is inconceivable if S is finite (unless $S = \varnothing$).

Namely, consider the set $E = \{2k \mid k \in \mathbb{N}\}$ of all even numbers, and the set $O = \{2k + 1 \mid k \in \mathbb{N}\}$ of all odd numbers. Both E and O are infinite, hence countable; thus we have $|\mathbb{N}| = |E| = |O|$ while $\mathbb{N} = E \cup O$ and $E \cap O = \varnothing$.

We can do even better. Let p_n denote the n^{th} prime number (i.e., $p_0 = 2$, $p_1 = 3$, etc.). Let

$$S_0 = \{2^k \mid k \in \mathbb{N}\}, S_1 = \{3^k \mid k \in \mathbb{N}\}, \ldots, S_n = \{p_n^k \mid k \in \mathbb{N}\}, \ldots.$$

The sets S_n ($n \in \mathbb{N}$) are mutually disjoint countable subsets of $\mathbb{N}$. Thus we have $\mathbb{N} \supseteq \bigcup_{n=0}^{\infty} S_n$, where $|S_n| = |\mathbb{N}|$ and the S_ns are mutually disjoint.

The following two theorems show that simple operations applied to countable sets yield countable sets.

Theorem. The union of two countable sets is a countable set.

Proof. Let $A = \{a_n \mid n \in \mathbb{N}\}$ and $B = \{b_n \mid n \in \mathbb{N}\}$ be countable. We construct a sequence $\langle c_n \rangle_{n=0}^{\infty}$ as follows:

$$c_{2k} = a_k \text{and } c_{2k+1} = b_k \text{ for all } k \in \mathbb{N}.$$

Then $A \cup B = \{c_n \mid n \in \mathbb{N}\}$ and since it is infinite, it is countable. □

Corollary. The union of a finite system of countable sets is countable.

Proof. By induction (on the size of the system). □

Theorem. If A and B are countable, then $A \times B$ is countable.

Proof. It suffices to show that $|\mathbb{N} \times \mathbb{N}| = |\mathbb{N}|$, i.e., to construct either a one-to-one mapping of $\mathbb{N} \times \mathbb{N}$ onto $\mathbb{N}$ or a one-to-one sequence with range $\mathbb{N} \times \mathbb{N}$. Consider the function

$$f(k, n) = 2^k \cdot (2n + 1) - 1.$$

It is easy to verify that f is one-to-one and that the range of f is $\mathbb{N}$. □

Corollary. The cartesian product of a finite number of countable sets is countable. Consequently, $\mathbb{N}^m$ is countable, for every $m > 0$.

Theorem. Let $\langle A_n \mid n \in \mathbb{N}\rangle$ be a countable system of at most countable sets, and let $\langle a_n \mid n \in \mathbb{N}\rangle$ be a system of enumerations of A_n; i.e., for each $n \in \mathbb{N}$, $a_n = \langle a_n(k) \mid k \in \mathbb{N}\rangle$ is an infinite sequence, and $A_n = \{a_n(k) \mid k \in \mathbb{N}\}$. Then $\bigcup_{n=0}^{\infty} A_n$ is at most countable.

Proof. Define $f : \mathbb{N} \times \mathbb{N} \mapsto \bigcup_{n=0}^{\infty} A_n$ by $f(n, k) = a_n(k)$. f maps $\mathbb{N} \times \mathbb{N}$ onto $\bigcup_{n=0}^{\infty} A_n$, so the latter is at most countable. □

As a corollary of this result we can now prove

Theorem. If A is countable, then the set Seq (A) of all finite sequences of elements of A is countable.

Proof. It is enough to prove the theorem for $A = \mathbb{N}$. As Seq ($\mathbb{N}$) = $\bigcup_{n=0}^{\infty} \mathbb{N}^n$, the theorem follows if we can produce a sequence $\langle a_n \mid n \geq 1\rangle$ of enumerations of $\mathbb{N}^n$. We do that by recursion.

Let g be a one-to-one mapping of $\mathbb{N}$ onto $\mathbb{N} \times \mathbb{N}$. Define recursively

$a_1(i) = \langle i\rangle$ for all $i \in \mathbb{N}$;
$a_{n+1}(i) = \langle b_0, \ldots, b_{n-1}, i_2\rangle$ where $g(i)=(i_1, i_2)$ and $\langle b_0, \ldots, b_{n-1}\rangle = a_n(i_1)$,for all $i \in \mathbb{N}$.

The idea is to let $a_{n+1}(i)$ be the $(n + 1)$-tuple resulting from the concatenation of the $(i_1)^{\text{th}}$ n-tuple (in the previously constructed enumeration of n-tuples, a_n) with i_2. An easy proof by induction shows that a_n is onto $\mathbb{N}^n$, for all $n \geq 1$, and therefore $\bigcup_{n=1}^{\infty} \mathbb{N}^n$ is countable.

Since $\mathbb{N}^0 = \{<>\}$, $\bigcup_{n=0}^{\infty} \mathbb{N}^n$ is also countable. □

Corollary. The set of all finite subsets of a countable set is countable.

Proof. The function F defined by $F(<a_0, \ldots, a_{n-1}>) = \{a_0, \ldots, a_{n-1}\}$ maps the countable set Seq (A) onto the set of all finite subsets of A. □

Other useful results about countable sets are the following.

Theorem. The set of all integers **Z** and the set of all rational numbers **Q** are countable.

Proof. **Z** is countable because it is the union of two countable sets:

$$\mathbf{Z} = \{0, 1, 2, 3, \ldots\} \cup \{-1, -2, -3, \ldots\}.$$

Q is countable because the function $f : \mathbf{Z} \times (\mathbf{Z} - \{0\}) \mapsto \mathbf{Q}$ defined by $f(p, q) = p / q$ maps a countable set onto **Q**. □

8. Real Numbers

Definition. An ordered set $(X, <)$ is *dense* if it has at least two elements and if for all $a, b \in X$, $a < b$ implies that there exists $x \in X$ such that $a < x < b$.

Let us call the least and the greatest elements of a linearly ordered set (if they exist) the *endpoints* of the set.

The most important example of a countable dense linearly ordered set is the set **Q** of all rational numbers, ordered by size. The ordering is dense because, if r, s are rational numbers and $r < s$, then $x = (r + s) / 2$ is also a rational number, and $r < x < s$. Moreover, (**Q**,<) has no endpoints (if $r \in \mathbf{Q}$ then $r + 1$, $r - 1 \in \mathbf{Q}$ and $r - 1 < r < r + 1$).

Definition. Let $(P, <)$ be a dense linearly ordered set. P is *complete* if every non-empty $S \subseteq P$ bounded from above has a supremum.

The ordered set (**Q**, <) of rationals has a unique completion (up to isomorphism); this is the ordered set of real numbers. The completion of (**Q**, <) is denoted $(\mathbb{R}, <)$; the elements of $\mathbb{R}$ are the *real numbers*.

Theorem. $(\mathbb{R}, <)$ is the unique (up to isomorphism) complete linearly ordered set without endpoints that has a countable subset dense in it.

9. Uncountable Sets

All infinite sets whose cardinalities we have determined up to this point turned out to be countable. Naturally, a question arises whether perhaps all infinite sets are countable. If it were so, this book might end with the preceding section. It was a great discovery of Georg Cantor that uncountable sets, in fact, exist. This discovery provided an impetus for the development of set theory and became a source of its depth and richness.

Theorem The set $\mathbb{R}$ of all real numbers is uncountable.

Proof. Assume that $\mathbb{R}$ is countable, i.e., $\mathbb{R}$ is the range of some infinite sequence $\langle r_n \rangle_{n=1}^{\infty}$. Let $a_0^{(n)}.a_1^{(n)}a_2^{(n)}a_3^{(n)}\cdots$ be the decimal expansion of r_n. Let $b_n = 1$ if $a_n^{(n)} = 0$, $b_n = 0$ otherwise; and let r be the real number whose decimal expansion is $0.b_1b_2b_3\cdots$. We have $b_n \neq a_n^{(n)}$, hence $r \neq r_n$, for all $n = 1, 2, 3, \ldots$, a contradiction. □

The combinatorial heart of the diagonal argument (quite similar to Russell's Paradox, which is of later origin) becomes even clearer in the next theorem.

Theorem. The set of all sets of natural numbers is uncountable; in fact, $|\mathcal{P}(\mathbb{N})| > |\mathbb{N}|$.

Proof. The function $f : \mathbb{N} \mapsto \mathcal{P}(\mathbb{N})$ defined by $f(n) = \{n\}$ is one-to-one, so $|\mathbb{N}| \leq |\mathcal{P}(\mathbb{N})|$. We prove that for every sequence $\langle S_n \mid n \in \mathbb{N}\rangle$ of subsets of $\mathbb{N}$ there is some $S \subseteq \mathbb{N}$ such that $S \neq S_n$ for all $n \in \mathbb{N}$. This shows that there is no mapping of $\mathbb{N}$ onto $\mathcal{P}(\mathbb{N})$, and hence $|\mathbb{N}| < |\mathcal{P}(\mathbb{N})|$.

We define the set $S \subseteq \mathbb{N}$ as follows: $S = \{n \in \mathbb{N} \mid n \notin S_n\}$. The number n is used to distinguish S from S_n: If $n \in S_n$, then $n \notin S$, and if $n \notin S_n$, then $n \in S$. In either case, $S \neq S_n$, as required. □

The set $2^{\mathbb{N}} = \{0, 1\}^{\mathbb{N}}$ of all infinite sequences of 0's and 1's is also uncountable, and, in fact, has the same cardinality as $\mathcal{P}(\mathbb{N})$ and $\mathbb{R}$.

Theorem. $|\mathcal{P}(\mathbb{N})| = |2^{\mathbb{N}}| = |\mathbb{R}|$.

Proof. For each $S \subseteq \mathbb{N}$ define the *characteristic function* of S, $\chi_S : \mathbb{N} \mapsto \{0, 1\}$, as follows:

$$\chi_{S(n)} = \begin{cases} 0 \text{ if } n \in S; \\ 1 \text{ if } n \notin S. \end{cases}$$

It is easy to check that the correspondence between sets and their characteristic functions is a one-to-one mapping of $\mathcal{P}(\mathbb{N})$ onto $\{0,1\}^{\mathbb{N}}$.

To complete the proof, we show that $|\mathbb{R}| \leq |\mathcal{P}(\mathbb{N})|$ and also $|2^{\mathbb{N}}| \leq |\mathbb{R}|$ and use the Cantor-Bernstein Theorem.

a. We have constructed real numbers as cuts in the set **Q** of all rational numbers. The function that assigns to each real number $r = (A, B)$ the set $A \subseteq \mathbf{Q}$ is a one-to-one mapping of $\mathbb{R}$ into $\mathcal{P}(\mathbf{Q})$. Therefore $|\mathbb{R}| \leq |\mathcal{P}(\mathbf{Q})|$. As $|\mathbf{Q}| = |\mathbb{N}|$, we have $|\mathcal{P}(\mathbf{Q})| = |\mathcal{P}(\mathbb{N})|$. Hence $|\mathbb{R}| \leq |\mathcal{P}(\mathbb{N})|$.
b. To prove $|2^{\mathbb{N}}| \leq |\mathbb{R}|$ we use the decimal representation of real numbers. The function that assigns to each infinite sequence $\langle a_n \rangle_{n=0}^{\infty}$ of 0's and 1's the unique real number whose decimal expansion is $0.a_0a_1a_2\cdots$ is a one-to-one mapping of $2^{\mathbb{N}}$ into $\mathbb{R}$. Therefore we have $|2^{\mathbb{N}}| \leq |\mathbb{R}|$.

□

Acknowledgements

The editors would like to thank Gert-Jan Lokhorst for his assistance in using Otfried Cheong' Hyperlatex system, which helped us to convert the LaTeX source to HTML.

Thomas Jech
jech@math.cas.cz

Return to Section 1 (paragraph 2) of Set Theory
Return to Section 1 (paragraph 5) of Set Theory
Return to Section 2 of Set Theory

First published: July 10, 2002
Content last modified: July 10, 2002

Zermelo-Fraenkel Set Theory

Axioms of ZF

> **Extensionality**:
> $\forall x \forall y[\forall z(z \in x \equiv z \in y) \rightarrow x=y]$

This axiom asserts that when sets *x* and *y* have the same members, they are the same set.

The next axiom asserts the existence of the empty set:

> **Null Set**:
> $\exists x \forall y(y \notin x)$

Since it is provable from this axiom and the previous axiom that there is a unique such set, we may introduce the notation '∅' to denote it.

The next axiom asserts that if given any set *x* and *y*, there exists a pair set of *x* and *y*, i.e., a set which has only *x* and *y* as members:

> **Pairs**:
> $\forall x \forall y \exists z \forall w(w \in z \equiv w=x \lor w=y)$

Since it is provable that there is a unique pair set for each given *x* and *y*, we introduce the notation '$\{x,y\}$' to denote it.

The next axiom asserts that for any given set *x*, there is a set *y* which has as members all of the members of all of the members of *x*:

> **Unions**:
> $\forall x \exists y \forall z[z \in y \equiv \exists w(w \in x \;\&\; z \in w)]$

Since it is provable that there is a unique 'union' of any set *x*, we introduce the notation '$\bigcup x$' to denote it.

The next axiom asserts that for any set *x*, there is a set *y* which contains as members all those sets whose

members are also elements of x, i.e., y contains all of the subsets of x:

> **Power Set**:
> $\forall x \exists y \forall z[z \in y \equiv \forall w(w \in z \rightarrow w \in x)]$

Since every set provably has a unique `power set', we introduce the notation '$\wp(x)$' to denote it. Note also that we may define the notion x *is a subset of* y ('$x \subseteq y$') as: $\forall z(z \in x \rightarrow z \in y)$. Then we may simplify the statement of the Power Set Axiom as follows:

> $\forall x \exists y \forall z[z \in y \equiv z \subseteq x)$

The next axiom asserts the existence of an infinite set, i.e., a set with an infinite number of members:

> **Infinity**:
> $\exists x[\varnothing \in x \ \& \ \forall y(y \in x \rightarrow \bigcup\{y,\{y\}\} \in x)]$

We may think of this as follows. Let us define *the union of* x *and* y ('$x \cup y$') as the union of the pair set of x and y, i.e., as $\bigcup\{x,y\}$. Then the Axiom of Infinity asserts that there is a set x which contains Ø as a member and which is such that, anytime y is a member of x, then $y \cup \{y\}$ is a member of x. Consequently, this axiom guarantees the existence of a set of the following form:

> $\{\varnothing, \ \{\varnothing\}, \ \{\varnothing, \{\varnothing\}\}, \ \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}, \ \ldots \}$

Notice that the second element, {Ø}, is in this set because (1) the fact that Ø is in the set implies that $\varnothing \cup \{\varnothing\}$ is in the set and (2) $\varnothing \cup \{\varnothing\}$ just is {Ø}. Similarly, the third element, {Ø, {Ø}}, is in this set because (1) the fact that {Ø} is in the set implies that $\{\varnothing\} \cup \{\{\varnothing\}\}$ is in the set and (2) $\{\varnothing\} \cup \{\{\varnothing\}\}$ just is {Ø, {Ø}}. And so forth.

The next axiom asserts that every set is 'well-founded':

> **Regularity**:
> $\forall x[x \neq \varnothing \rightarrow \exists y(y \in x \ \& \ \forall z(z \in x \rightarrow z \notin y))]$

A member y of a set x with this property is called a 'minimal' element. This axiom rules out the existence of circular chains of sets (e.g., such as $x \in y$ & $y \in z$ & and $z \in x$) as well as infinitely descending chains of sets (such as $\ldots x_3 \in x_2 \in x_1 \in x_0$).

The final axiom of ZF is the Replacement Schema. Suppose that $\varphi(x,y,\vec{z})$ is a formula with x and y free, and which may or may not have free variables $z_1,\ldots,z_k$. Furthermore, let $\varphi_{x,y,}\vec{z}[s,r,\vec{z}]$ be the result of substituting s and r for x and y, respectively, in $\varphi(x,y,\vec{z})$. The every instance of the following schema is

an axiom:

> **Replacement Schema**:
> $\forall z_1 \ldots \forall z_k [\forall x \exists! y \varphi(x,y,\vec{z}) \rightarrow \forall u \exists v \forall r (r \in v \equiv \exists s (s \in u \,\&\, \varphi_{x,y,\vec{z}}[s,r,\vec{z}]))]$

In other words, if we know that φ is a functional formula (which relates each set x to a unique set y), then if we are given a set u, we can form a new set v as follows: collect all of the sets to which the members of u are uniquely related by φ.

Note that the Replacement Schema can take you 'out of' the set u when forming the set v. The elements of v need not be elements of u. By contrast, the well-known Separation Schema of Zermelo yields new sets consisting only of those elements of a given set u which satisfy a certain condition ψ. That is, suppose that $\psi(x,\vec{z})$ has x free and may or may not have $z_1,\ldots,z_k$ free. Then the Separation Schema asserts:

> **Separation Schema** $\forall z_1 \ldots \forall z_k [\forall u \exists v \forall r (r \in v \equiv r \in u \,\&\, \psi_{x,\vec{z}}[r,\vec{z}])]$

In other words, if given a formula ψ and a set u, there exists a set v which has as members precisely the members of u which satisfy the formula ψ.

Thomas Jech
jech@math.cas.cz

Return to Set Theory

First published: July 10, 2002
Content last modified: July 10, 2002

Automated Reasoning

Reasoning is the ability to make inferences, and automated reasoning is concerned with the building of computing systems that automate this process. Although the overall goal is to mechanize different forms of reasoning, the term has largely been identified with valid deductive reasoning as practiced in mathematics and formal logic. In this respect, automated reasoning is akin to mechanical theorem proving. Building an automated reasoning program means providing an algorithmic description to a formal calculus so that it can be implemented on a computer to prove theorems of the calculus in an efficient manner. Important aspects of this exercise involve defining the class of problems the program will be required to solve, deciding what language will be used by the program to represent the information given to it as well as new information inferred by the program, specifying the mechanism that the program will use to conduct deductive inferences, and figuring out how to perform all these computations efficiently. While basic research work continues in order to provide the necessary theoretical framework, the field has reached a point where automated reasoning programs are being used by researchers to attack open questions in mathematics and logic, and to solve problems in engineering.

The Problem Domain

A problem being presented to an automated reasoning program consists of two main items, namely a statement expressing the particular question being asked called the **problem's conclusion**, and a collection of statements expressing all the relevant information available to the program -- the **problem's assumptions**. Solving a problem means proving the conclusion from the given assumptions by the systematic application of rules of deduction embedded within the reasoning program. The problem solving process ends when one such proof is found, when the program is able to detect the non-existence of a proof, or when it simply runs out of resources.

A first important consideration in the design of an automated reasoning program is to delineate the class of problems that the program will be required to solve -- the **problem domain**. The domain can be very large, as would be the case for a general-purpose theorem prover for first-order logic, or be more restricted in scope as in a special-purpose theorem prover for Tarski's geometry, or the modal logic K. A typical approach in the design of an automated reasoning program is to provide it first with sufficient logical power (e.g., first-order logic) and then further demarcate its scope to the particular domain of interest defined by a set of **domain axioms**. To illustrate, EQP, a theorem-proving program for equational logic, was used to solve an open question in Robbins algebra (McCune 1997): *Are all Robbins algebras Boolean?* For this, the program was provided with the axioms defining a Robbins algebra:

(A1)	$x + y = y + x$	(commutativity)
(A2)	$(x + y) + z = x + (y + z)$	(associativity)
(A3)	$-(-(x + y) + -(x + -y)) = x$	(Robbins equation)

The program was then used to show that a characterization of Boolean algebra that uses Huntington's equation,

$$-(-x + y) + -(-x + -y) = x,$$

follows from the axioms. We should remark that this problem is non-trivial since deciding whether a finite set of equations provides a basis for Boolean algebra is undecidable, that is, it does not permit an algorithmic representation; also, the problem was attacked by Robbins, Huntington, Tarski and many of his students with no success. The key step was to establish that all Robbins algebras satisfy

$$(\exists x)(\exists y)(x + y = x),$$

since it was known that this formula is a sufficient condition for a Robbins algebra to be Boolean. When EQP was supplied with this piece of information, the program provided invaluable assistance by completing the proof automatically.

A special-purpose theorem prover does not draw its main benefit by restricting its attention to the domain axioms but from the fact that the domain may enjoy particular theorem-proving techniques which can be hardwired -- coded -- within the reasoning program itself and which may result in a more efficient logic implementation. Much of EQP's success at settling the Robbins question can be attributed to its built-in associative-commutative inference mechanisms.

Language Representation

A second important consideration in the building of an automated reasoning program is to decide (1) how problems in its domain will be presented to the reasoning program; (2) how they will actually be represented internally within the program; and, (3) how the solutions found -- completed proofs -- will be displayed back to the user. There are several formalisms available for this, and the choice is dependent on the problem domain and the underlying deduction calculus used by the reasoning program. The most commonly used formalisms include standard first-order logic, typed λ-calculus, and clausal logic. We take up clausal logic here and assume that the reader is familiar with the rudiments of first-order logic; for the typed λ-calculus the reader may want to check Church 1940. Clausal logic is a quantifier-free variation of first-order logic and has been the most widely used notation within the automated reasoning community. Some definitions are in order: A **term** is a constant, a variable, or a function whose arguments are themselves terms. For example, a, x, $f(x)$, and $h(c,f(z),y)$ are all terms. A **literal** is either an atomic formula, e.g. $F(x)$, or the negation of an atomic formula, e.g. $\sim R(x,f(a))$. Two literals are **complementary** if one is the negation of the other. A **clause** is a (possibly empty) finite disjunction of literals $l_1 \vee \ldots \vee l_n$ where no literal appears more than once in the clause (that is, clauses can be alternatively treated as sets of literals). **Ground** terms, ground literals, and ground clauses have no variables. The **empty clause**, [ ], is the clause having no literals and, hence, is unsatisfiable -- false under any interpretation. Some examples: $\sim R(a,b)$, and $F(a) \vee \sim R(f(x),b) \vee F(z)$ are both examples of clauses but only the former is ground. The general idea is to be able to express a problem's formulation as a set of clauses or, equivalently, as a formula in **conjunctive normal form**, that is, as a conjunction of clauses.

For formulas already expressed in standard logic notation, there is a systematic two-step procedure for transforming them into conjunctive normal form. The first step consists in re-expressing a formula into a semantically equivalent formula in **prenex normal form**, $(\Theta x_1)\ldots(\Theta x_n)\alpha(x_1,\ldots,x_n)$, consisting of a string of quantifiers $(\Theta x_1)\ldots(\Theta x_n)$ followed by a quantifier-free expression $\alpha(x_1,\ldots,x_n)$ called the **matrix**. The second step in the transformation first converts the matrix into conjunctive normal form by using well-known logical equivalences such as DeMorgan's laws, distribution, double-negation, and others; then, the quantifiers in front of the matrix, which is now in conjunctive normal form, are dropped according to certain rules. In the presence of existential quantifiers, this latter step does not always preserve equivalence and requires the introduction of **Skolem functions** whose role is to "simulate" the behaviour of existentially quantified variables. For example, applying the skolemizing process to the formula

$$(\forall x)(\exists y)(\forall z)(\exists u)(\forall v)[R(x,y,v) \vee \sim K(x,z,u,v)]$$

requires the introduction of a one-place and two-place Skolem functions, *f* and *g* respectively, resulting in the formula

$$(\forall x)(\forall z)(\forall v)\ [R(x,f(x),v) \lor \sim K(x,z,g(x,z),v)]$$

The universal quantifiers can then be removed to obtain the final clause, $R(x,f(x),v) \lor \sim K(x,z,g(x,z), v)$ in our example. The Skolemizing process may not preserve equivalence but maintains satisfiability, which is enough for clause-based automated reasoning.

Although clausal form provides a more uniform and economical notation -- there are no quantifiers and all formulas are disjunctions -- it has certain disadvantages. One drawback is the exponential increase in the size of the resulting formula when transformed from standard logic notation into clausal form. The increase in size is accompanied by an increase in cognitive complexity that makes it harder for humans to read proofs written with clauses. Another disadvantage is that the syntactic structure of a formula in standard logic notation can be used to guide the construction of a proof but this information is completely lost in the transformation into clausal form.

Deduction Calculi

A third important consideration in the building of an automated reasoning program is the selection of the actual deduction calculus that will be used by the program to perform its inferences. As indicated before, the choice is highly dependent on the nature of the problem domain and there is a fair range of options available: General-purpose theorem proving and problem solving (first-order logic, simple type theory), program verification (first-order logic), distributed and concurrent systems (modal and temporal logics), program specification (intuitionistic logic), hardware verification (higher-order logic), logic programming (Horn logic), and so on.

A deduction calculus consists of a set of logical axioms and a collection of deduction rules for deriving new formulas from previously derived formulas. Solving a problem in the program's problem domain then really means establishing a particular formula α -- the problem's conclusion -- from the extended set Γ consisting of the logical axioms, the domain axioms, and the problem assumptions. That is, the program needs to determine if $\Gamma \vDash \alpha$. How the program goes about establishing this semantic fact depends, of course, on the calculus it implements. Some programs may take a very **direct** route and attempt to establish that $\Gamma \vDash \alpha$ by actually constructing a step-by-step proof of α from Γ. If successful, this shows of course that Γ derives -- proves -- α, a fact we denote by writing $\Gamma \vdash \alpha$. Other reasoning programs may instead opt for a more **indirect** approach and try to establish that $\Gamma \vDash \alpha$ by showing that $\Gamma \cup \{\sim\alpha\}$ is inconsistent which, in turn, is shown by deriving a contradiction, $\bot$, from the set $\Gamma \cup \{\sim\alpha\}$. Automated systems that implement the former approach include natural deduction systems; the latter approach is used by systems based on resolution, sequent deduction, and matrix connection methods.

Soundness and completeness are two (metatheoretical) properties of a calculus that are particularly

important for automated deduction. **Soundness** states that the rules of the calculus are truth-preserving. For a direct calculus this means that if $\Gamma \vdash \alpha$ then $\Gamma \vDash \alpha$. For indirect calculi, soundness means that if $\Gamma \cup \{\sim\alpha\} \vdash \bot$ then $\Gamma \vDash \alpha$. **Completeness** in a direct calculus states that if $\Gamma \vDash \alpha$ then $\Gamma \vdash \alpha$. For indirect calculi, the completeness property is expressed in terms of **refutations** since one establishes that $\Gamma \vDash \alpha$ by showing the existence of a proof, not of α from Γ, but of $\bot$ from $\Gamma \cup \{\sim\alpha\}$. Thus, an indirect calculus is **refutation complete** if $\Gamma \vDash \alpha$ implies $\Gamma \cup \{\sim\alpha\} \vdash \bot$. Of the two properties, soundness is the most desirable. An incomplete calculus indicates that there are entailment relations that cannot be established within the calculus. For an automated reasoning program this means, informally, that there are true statements that the program cannot prove. Incompleteness may be an unfortunate affair but lack of soundness is a truly problematic situation since an unsound reasoning program would be able to generate false conclusions from perfectly true information.

It is important to appreciate the difference between a logical calculus and its corresponding implementation in a reasoning program. The implementation of a calculus invariably involves making some modifications to the calculus and this results, strictly speaking, in a new calculus. The most important modification to the original calculus is the "mechanization" of its deduction rules, that is, the specification of the systematic way in which the rules are to be applied. In the process of doing so, one must exercise care to preserve the metatheoretical properties of the original calculus.

Two other metatheoretical properties of importance to automated deduction are decidability and complexity. A calculus is **decidable** if it admits an algorithmic representation, that is, if there is an algorithm that, for any given Γ and α, it can determine in a finite amount of time the answer, "Yes" or "No", to the question "Does $\Gamma \vDash \alpha$?" A calculus may be undecidable in which case one needs to determine which decidable fragment to implement. The time-space **complexity** of a calculus specifies how efficient its algorithmic representation is. Automated reasoning is made the more challenging because many calculi of interest are not decidable and have poor complexity measures forcing researchers to seek tradeoffs between deductive power versus algorithmic efficiency.

Resolution

Of the many calculi used in the implementation of reasoning programs, the ones based on the **resolution** principle have been the most popular. Resolution is modeled after the chain rule (of which Modus Ponens is a special case) and essentially states that from $p \vee q$ and $\sim q \vee r$ one can infer $p \vee r$. More formally, let $C - l$ denote the clause C with the literal l removed. Assume that C_1 and C_2 are ground clauses containing, respectively, a positive literal l_1 and a negative literal $\sim l_2$ such that l_1 and $\sim l_2$ are complementary. Then, the rule of **ground resolution** states that, as a result of **resolving** C_1 and C_2, one can infer $(C_1 - l_1) \vee (C_2 - \sim l_2)$:

C_1

C_2

$$\frac{\qquad\qquad}{(C_1 - l_1) \vee (C_2 - \sim l_2)} \quad \text{(ground resolution)}$$

Herbrand's theorem (Herbrand 1930) assures us that the non-satisfiability of *any* set of clauses, ground or not, can be established by using ground resolution. This is a very significant result for automated deduction since it tells us that if a set Γ is not satisfied by any of the infinitely many interpretations, this fact can be determined in *finitely* many steps. Unfortunately, a direct implementation of ground resolution using Herbrand's theorem requires the generation of a vast number of ground terms making this approach hopelessly inefficient. This issue was effectively addressed by generalizing the ground resolution rule to **binary resolution** and by introducing the notion of unification (Robinson 1965a). Unification allows resolution proofs to be “lifted” and be conducted at a more general level; clauses only need to be instantiated at the moment where they are to be resolved. Moreover, the clauses resulting from the instantiation process do not have to be ground instances and may still contain variables. The introduction of binary resolution and unification is considered one of the most important developments in the field of automated reasoning.

Unification

A **unifier** of two expressions -- terms or clauses -- is a substitution that when applied to the expressions makes them equal. For example, the substitution

$$\{x \leftarrow b, y \leftarrow b, z \leftarrow f(a,b)\}$$

is a unifier for

$$R(x,f(a,y)) \text{ and } R(b,z)$$

since when applied to both expressions makes them equal:

$$\begin{aligned} R(x,f(a,y))\{x \leftarrow b, y \leftarrow b, z \leftarrow f(a,b)\} &= R(b,f(a,b)) \\ &= R(b,z)\{x \leftarrow b, y \leftarrow b, z \leftarrow f(a,b)\} \end{aligned}$$

A **most general unifier** (mgu) produces the most general instance shared by two unifiable expressions. In the previous example, the substitution $\{x \leftarrow b, y \leftarrow b, z \leftarrow f(a,b)\}$ is a unifier but not an mgu; however, $\{x \leftarrow b, z \leftarrow f(a,y)\}$ is an mgu. Note that unification attempts to "match" two expressions and this fundamental process has become a central component of most automated deduction programs, resolution-based and otherwise. **Theory-unification** is an extension of the unification mechanism that includes built-in inference capabilities. For example, the clauses $R(g(a,b),x)$ and $R(g(b,a),d)$ do not unify but they AC-unify, where AC-unification is unification with built-in associative and commutative rules such as $g(a,b) = g(b,a)$. Shifting inference capabilities into the unification mechanism adds power but at a

price: The existence of an mgu for two unifiable expressions may not be unique (there could actually be infinitely many), and the unification process becomes undecidable in general.

Binary resolution

Let C_1 and C_2 be two clauses containing, respectively, a positive literal l_1 and a negative literal $\sim l_2$ such that l_1 and l_2 unify with mgu θ. Then,

$$\frac{C_1 \qquad C_2}{(C_1\theta - l_1\theta) \vee (C_2\theta - \sim l_2\theta)} \quad \text{(binary resolution)}$$

by binary resolution; the clause $(C_1\theta - l_1\theta) \vee (C_2\theta - \sim l_2\theta)$ is called a **binary resolvent** of C_1 and C_2.

Factoring

If two or more literals occurring in a clause C share an mgu θ then $C\theta$ is a **factor** of C. For example, in $R(x,a) \vee \sim K(f(x),b) \vee R(c,y)$ the literals $R(x,a)$ and $R(c,y)$ unify with mgu $\{x \leftarrow c, y \leftarrow a\}$ and, hence, $R(c,a) \vee \sim K(f(c),b)$ is a factor of the original clause.

The Resolution Principle

Let $C_1$and C_2 be two clauses. Then, a **resolvent** obtained by **resolution** from C_1 and C_2 is defined as: (a) a binary resolvent of C_1 and C_2; (b) a binary resolvent of C_1 and a factor of C_2; (c) a binary resolvent of a factor of C_1 and C_2; or, (d) a binary resolvent of a factor of C_1 and a factor of C_2.

Resolution proofs, more precisely refutations, are constructed by deriving the empty clause [ ] from $\Gamma \cup \{\sim\alpha\}$ using resolution; this will always be possible if $\Gamma \cup \{\sim\alpha\}$ is unsatisfiable since resolution is refutation complete (Robinson 1965a). As an example of a resolution proof, we show that the set $\{(\forall x)(P(x) \vee Q(x)), (\forall x)(P(x) \supset R(x)),(\forall x)(Q(x) \supset R(x))\}$, denoted by Γ, entails the formula $(\exists x)R(x)$. The first step is to find the clausal form of $\Gamma \cup \{\sim(\exists x)R(x)\}$; the resulting clause set, denoted by S_0, is shown in steps 1 to 4 in the refutation below. The refutation is constructed by using a level-saturation method: Compute all the resolvents of the initial set, S_0, add them to the set and repeat the process until the empty clause is derived. (This produces the sequence of increasingly larger sets: $S_0, S_1, S_2,\ldots$) The only constraint that we impose is that we do not resolve the same two clauses more than once.

S_0	1	$P(x) \vee Q(x)$	Assumption
	2	$\sim P(x) \vee R(x)$	Assumption
	3	$\sim Q(x) \vee R(x)$	Assumption
	4	$\sim R(a)$	Negation of the conclusion
S_1	5	$Q(x) \vee R(x)$	Res 1 2
	6	$P(x) \vee R(x)$	Res 1 3
	7	$\sim P(a)$	Res 2 4
	8	$\sim Q(a)$	Res 3 4
S_2	9	$Q(a)$	Res 1 7
	10	$P(a)$	Res 1 8
	11	$R(x)$	Res 2 6
	12	$R(x)$	Res 3 5
	13	$Q(a)$	Res 4 5
	14	$P(a)$	Res 4 6
	15	$R(a)$	Res 5 8
	16	$R(a)$	Res 6 7
S_3	17	$R(a)$	Res 2 10
	18	$R(a)$	Res 2 14
	19	$R(a)$	Res 3 9
	20	$R(a)$	Res 3 13
	21	[ ]	Res 4 11

Although the resolution proof is successful in deriving [ ], it has some significant drawbacks. To start with, the refutation is too long as it takes 21 steps to reach the contradiction, [ ]. This is due to the naïve brute-force nature of the implementation. The approach not only generates too many formulas but some are clearly redundant. Note how $R(a)$ is derived six times; also, $R(x)$ has more “information content” than $R(a)$ and one should keep the former and disregard the latter. Resolution, like all other automated deduction methods, must be supplemented by strategies aimed at improving the efficiency of the deduction process. The above sample derivation has 21 steps but research-type problems command derivations with thousands or hundreds of thousands of steps.

Resolution Strategies

The successful implementation of a deduction calculus in an automated reasoning program requires the integration of search strategies that reduce the search space by pruning unnecessary deduction paths. Some strategies remove redundant clauses or tautologies as soon as they appear in a derivation. Another strategy is to remove more specific clauses in the presence of more general ones by a process known as **subsumption** (Robinson 1965a). Unrestricted subsumption, however, does not preserve the refutation completeness of resolution and, hence, there is a need to restrict its applicability (Loveland 1978). **Model elimination** (Loveland 1969) can discard a sentence by showing that it is false in some model of the axioms. The subject of model generation has recently received renewed attention as a complementary process to theorem proving. The method has been used successfully by automated reasoning programs to show the independence of axioms sets and to determine the existence of discrete mathematical structures meeting some given criteria.

Instead of removing redundant clauses, some strategies prevent the generation of useless clauses in the first place. The **set-of-support strategy** (Wos, Carson and Robinson 1965) is one of the most powerful strategies of this kind. A subset T of the set S, where S is initially $\Gamma \cup \{\sim\alpha\}$, is called a **set of support** of S iff $S - T$ is satisfiable. Set-of-support resolution dictates that the resolved clauses are not both from $S - T$. The motivation behind set-of-support is that since the set Γ is usually satisfiable it might be wise not to resolve two clauses from Γ against each other. **Hyperresolution** (Robinson 1965b) reduces the number of intermediate resolvents by combining several resolution steps into a single inference step.

Independently co-discovered, **linear resolution** (Loveland 1970, Luckham 1970) always resolves a clause against the most recently derived resolvent. This gives the deduction a simple “linear” structure affording a straightforward implementation; yet, linear resolution preserves refutation completeness. Using linear resolution we can derive the empty clause in the above example in only eight steps:

1	$P(x) \lor Q(x)$	Assumption
2	$\sim P(x) \lor R(x)$	Assumption
3	$\sim Q(x) \lor R(x)$	Assumption
4	$\sim R(a)$	Negation of the conclusion
5	$\sim P(a)$	Res 2 4
6	$Q(a)$	Res 1 5
7	$R(a)$	Res 3 6
8	[ ]	Res 4 7

With the exception of unrestricted subsumption, all the strategies mentioned so far preserve refutation completeness. Efficiency is an important consideration in automated reasoning and one may sometimes be willing to trade completeness for speed. **Unit resolution** and **input resolution** are two such refinements of linear resolution. In the former, one of the resolved clauses is always a literal; in the latter, one of the resolved clauses is always selected from the original set to be refuted. Albeit efficient, neither strategy is complete. Ordering strategies impose some form of partial ordering on the predicate symbols, terms, literals, or clauses occurring in the deduction. **Ordered resolution** treats clauses not as sets of literals but as sequences -- linear orders -- of literals. Ordered resolution is extremely efficient but, like unit and input resolution, is not refutation complete. To end, it must be noted that some strategies improve certain aspects of the deduction process at the expense of others. For instance, a strategy may reduce the size of the proof search space at the expense of increasing, say, the length of the shortest refutations.

There are several automated reasoning programs that are based on resolution, or refinements of resolution. Otter is one of the most versatile among these programs and is being used in a growing number of applications (Wos, Overbeek, Lusk and Boyle 1984). Resolution also provides the underlying logico-computational mechanism for the popular logic programming language Prolog (Clocksin and Mellish 1981).

Sequent Deduction

Hilbert-style calculi (Hilbert and Ackermann 1928) have been traditionally used to characterize logic systems. These calculi usually consist of a few axiom schemata and a small number of rules that typically include modus ponens and the rule of substitution. Although they meet the required theoretical requisites (soundness, completeness, etc.) the approach at proof construction in these calculi is difficult and does not reflect standard practice. It was Gentzen's goal “to set up a formalism that reflects as accurately as possible the actual logical reasoning involved in mathematical proofs” (Gentzen 1935). To carry out this task, Gentzen analyzed the proof-construction process and then devised two deduction calculi for classical logic: the natural deduction calculus, **NK**, and the sequent calculus, **LK**. (Gentzen actually designed NK first and then introduced LK to pursue metatheoretical investigations). The calculi met his goal to a large extent while at the same time managing to secure soundness and completeness. Both calculi are characterized by a relatively larger number of deduction rules and a simple axiom schema. Of the two calculi, LK is the one that has been most widely used in implementations of automated reasoning programs, and it is the one that we will discuss first; NK will be discussed in the next section.

Although the application of the LK rules affect logic formulas, the rules are seen as manipulating not logic formulas themselves but sequents. **Sequents** are expressions of the form $\Gamma \rightarrow \Delta$, where both Γ and Δ are (possibly empty) sets of formulas. Γ is the sequent's **antecedent** and Δ its **succedent**. Sequents can be interpreted thus: Let $\mathfrak{I}$ be an interpretation. Then,

> $\mathfrak{I}$ satisfies the sequent $\Gamma \rightarrow \Delta$ (written as: $\mathfrak{I} \models \Gamma \rightarrow \Delta$) iff

either $\mathfrak{I} \nvDash \alpha$ (for some $\alpha \in \Gamma$) or $\mathfrak{I} \vDash \beta$ (for some $\beta \in \Delta$).

In other words,

> $\mathfrak{I} \vDash \Gamma \rightarrow \Delta$ iff $\mathfrak{I} \vDash (\alpha_1 \,\&\, \ldots \,\&\, \alpha_n) \supset (\beta_1 \vee \ldots \vee \beta_n)$, where $\alpha_1 \,\&\, \ldots \,\&\, \alpha_n$ is the iterated conjunction of the formulas in Γ and $\beta_1 \vee \ldots \vee \beta_n$ is the iterated disjunction of those in Δ.

If Γ or Δ are empty then they are respectively valid or unsatisfiable. An **axiom of LK** is a sequent $\Gamma \rightarrow \Delta$ where $\Gamma \cap \Delta \neq \varnothing$. Thus, the requirement that the same formula occurs at each side of the $\rightarrow$ sign means that the axioms of LK are valid, for no interpretation can then make all the formulas in Γ true and, simultaneously, make all those in Δ false. LK has two rules per logical connective, plus one extra rule: the cut rule.

<table>
<tr><th colspan="2">Axioms</th><th colspan="2">Cut Rule</th></tr>
<tr><td colspan="2">-------
$\Gamma, \alpha \rightarrow \Delta, \alpha$</td><td colspan="2">$\Gamma \rightarrow \Delta, \alpha \quad \alpha, \Lambda \rightarrow \Sigma$

$\Gamma, \Lambda \rightarrow \Delta, \Sigma$</td></tr>
<tr><th colspan="2">Antecedent Rules ($\Theta\rightarrow$)</th><th colspan="2">Succedent Rules ($\rightarrow\Theta$)</th></tr>
<tr><td>$\&\rightarrow$</td><td>$\Gamma, \alpha, \beta \rightarrow \Delta$

$\Gamma, \alpha \,\&\, \beta \rightarrow \Delta$</td><td>$\rightarrow\&$</td><td>$\Gamma \rightarrow \Delta, \alpha \quad \Gamma \rightarrow \Delta, \beta$

$\Gamma \rightarrow \Delta, \alpha \,\&\, \beta$</td></tr>
<tr><td>$\vee\rightarrow$</td><td>$\Gamma, \alpha \rightarrow \Delta \quad \Gamma, \beta \rightarrow \Delta$

$\Gamma, \alpha \vee \beta \rightarrow \Delta$</td><td>$\rightarrow\vee$</td><td>$\Gamma \rightarrow \Delta, \alpha, \beta$

$\Gamma \rightarrow \Delta, \alpha \vee \beta$</td></tr>
<tr><td>$\supset\rightarrow$</td><td>$\Gamma \rightarrow \Delta, \alpha \quad \Gamma, \beta \rightarrow \Delta$

$\Gamma, \alpha \supset \beta \rightarrow \Delta$</td><td>$\rightarrow\supset$</td><td>$\Gamma, \alpha \rightarrow \Delta, \beta$

$\Gamma \rightarrow \Delta, \alpha \supset \beta$</td></tr>
<tr><td>$\supset\equiv$</td><td>$\Gamma, \alpha, \beta \rightarrow \Delta \quad \Gamma \rightarrow \Delta, \alpha, \beta$

$\Gamma, \alpha \equiv \beta \rightarrow \Delta$</td><td>$\equiv\supset$</td><td>$\Gamma, \alpha \rightarrow \Delta, \beta \quad \Gamma, \beta, \rightarrow \Delta, \alpha$

$\Gamma \rightarrow \Delta, \alpha \equiv \beta$</td></tr>
</table>

$\sim\rightarrow$	$\Gamma \rightarrow \Delta, \alpha$ ---- $\Gamma, \sim\alpha \rightarrow \Delta$	$\rightarrow\sim$	$\Gamma, \alpha \rightarrow \Delta$ ----- $\Gamma \rightarrow \Delta, \sim\alpha$
$\exists\rightarrow$	$\Gamma, \alpha(a/x) \rightarrow \Delta$ ------- $\Gamma, (\exists x)\alpha(x) \rightarrow \Delta$	$\exists\rightarrow$	$\Gamma \rightarrow \Delta, \alpha(t/x), (\exists x)\alpha(x)$ ---------- $\Gamma \rightarrow \Delta, (\exists x)\alpha(x)$
$\forall\rightarrow$	$\Gamma, \alpha(t/x), (\forall x)\alpha(x) \rightarrow \Delta$ ---------- $\Gamma, (\forall x)\alpha(x) \rightarrow \Delta$	$\forall\rightarrow$	$\Gamma \rightarrow \Delta, \alpha(a/x)$ ------- $\Gamma \rightarrow \Delta, (\forall x)\alpha(x)$

The sequents above a rule's line are called the **rule's premises** and the sequent below the line is the **rule's conclusion**. The quantification rules $\exists\rightarrow$ and $\rightarrow\forall$ have an eigenvariable condition that restricts their applicability, namely that a must not occur in Γ, Δ or in the quantified sentence. The purpose of this restriction is to ensure that the choice of parameter, a, used in the substitution process is completely "arbitrary".

Proofs in LK are represented as trees where each node in the tree is labeled with a sequent, and where the original sequent sits at the root of the tree. The children of a node are the premises of the rule being applied at that node. The leaves of the tree are labeled with axioms. Here is the LK-proof of $(\exists x)R(x)$ from the set $\{(\forall x)(P(x) \vee Q(x)), (\forall x)(P(x) \supset R(x)),(\forall x)(Q(x) \supset R(x))\}$. In the tree below, Γ stands for this set:

$\Gamma, P(a) \rightarrow P(a), R(a),(\exists x)R(x)$ $\quad$ $\Gamma, P(a), R(a) \rightarrow R(a), (\exists x)R(x)$ $\quad$ $\Gamma, Q(a) \rightarrow Q(a), R(a), (\exists x)R(x)$ $\quad$ $\Gamma, Q(a), R(a) \rightarrow R(a), (\exists x)R(x)$

---------------------------- $\quad$ -----------------------------

$\Gamma, P(a), P(a) \supset R(a) \rightarrow R(a), (\exists x)R(x)$ $\quad$ $\Gamma, Q(a), Q(a) \supset R(a) \rightarrow R(a), (\exists x)R(x)$

--------------- $\quad$ ---------------

$\Gamma, P(a) \rightarrow R(a), (\exists x)R(x)$ $\quad$ $\Gamma, Q(a) \rightarrow R(a), (\exists x)R(x)$

$\Gamma, P(a) \vee Q(a) \rightarrow R(a), (\exists x)R(x)$

$\Gamma, \rightarrow R(a), (\exists x)R(x)$

$\Gamma \rightarrow (\exists x)R(x)$

In our example, all the leaves in the proof tree are labeled with axioms. This establishes the validity of $\Gamma \rightarrow (\exists x)R(x)$ and, hence, the fact that $\Gamma \vDash (\exists x)R(x)$. LK takes an indirect approach at proving the conclusion and this is an important difference between LK and NK. While NK constructs an actual proof (of the conclusion from the given assumptions), LK instead constructs a proof that proves the existence

of a proof (of the conclusion from the assumptions). For instance, to prove that α is entailed by Γ, NK constructs a step-by-step proof of α from Γ (assuming that one exists); in contrast, LK first constructs the sequent $\Gamma \rightarrow \alpha$ which then attempts to prove valid by showing that it cannot be made false. This is done by searching for a counterexample that makes (all the sentences in) Γ true and makes α false: If the search fails then a counterexample does not exist and the sequent is therefore valid. In this respect, proof trees in LK are actually refutation proofs. Like resolution, LK is refutation complete: If $\Gamma \vDash \alpha$ then the sequent $\Gamma \rightarrow \alpha$ has a proof tree.

As it stands, LK is unsuitable for automated deduction and there are some obstacles that must be overcome before it can be efficiently implemented. The reason is, of course, that the statement of the completeness of LK only has to assert, for each entailment relation, the existence of a proof tree but a reasoning program has the more difficult task of actually having to construct one. Some of the main obstacles: First, LK does not specify the order in which the rules must be applied in the construction of a proof tree. Second, and as a particular case of the first problem, the premises in the rules $\forall\rightarrow$ and $\rightarrow\exists$ rules inherit the quantificational formula to which the rule is applied, meaning that the rules can be applied repeatedly to the same formula sending the proof search into an endless loop. Third, LK does not indicate which formula must be selected next in the application of a rule. Fourth, the quantifier rules provide no indication as to what terms or free variables must be used in their deployment. Fifth, and as a particular case of the previous problem, the application of a quantifier rule can lead into an infinitely long tree branch because the proper term to be used in the instantiation never gets chosen. Fortunately, as we will hint at below each of these problems can be successfully addressed.

Axiom sequents in LK are valid, and the conclusion of a rule is valid iff its premises are. This fact allows us to apply the LK rules in either direction, forwards from axioms to conclusion, or backwards from conclusion to axioms. Also, with the exception of the cut rule, all the rules' premises are subformulas of their respective conclusions. For the purposes of automated deduction this is a significant fact and we would want to dispense with the cut rule; fortunately, the cut-free version of LK preserves its refutation completeness (Gentzen 1935). These results provide a strong case for constructing proof trees in a backwards fashion; indeed, by working this way a refutation in cut-free LK gets increasingly simpler as it progresses since subformulas are simpler than their parent formulas. Moreover, and as far as propositional rules go, the new subformulas entered into the tree are completely dictated by the cut-free LK rules. Furthermore, and assuming the proof tree can be brought to completion, branches eventually end up with atoms and the presence of axioms can be quickly determined. Another reason for working backwards is that the truth-functional fragment of cut-free LK is **confluent** in the sense that the order in which the non-quantifier rules are applied is irrelevant: If there is a proof, regardless of what you do, you will run into it! To bring the quantifier rules into the picture, things can be arranged so that all rules have a fair chance of being deployed: Apply, as far as possible, all the non-quantifier rules before applying any of the quantifier rules. This takes care of the first and second obstacles, and it is no too difficult to see how the third one would now be handled. The fourth and fifth obstacles can be addressed by requiring that the terms to be used in the substitutions be suitably selected from the Herbrand universe (Herbrand 1930).

The use of sequent-type calculi in automated theorem proving was initiated by efforts to mechanize

mathematics (Wang 1960). At the time, resolution captured most of the attention of the automated reasoning community but during the 1970's some researchers started to further investigate non-resolution methods (Bledsoe 1977), prompting a frutiful and sustained effort to develop more human-oriented theorem proving systems (Bledsoe 1975, Nevins 1974). Eventually, sequent-type deduction gained momentum again, particularly in its re-incarnation as **analytic tableaux** (Fitting 1990). The method of deduction used in tableaux is essentially cut-free LK's with sets used in lieu of sequents.

Natural Deduction

Although LK and NK are both commonly labeled as “natural deduction” systems, it is the latter which better deserves the title due to its more natural, human-like, approach to proof construction. The rules of NK are typically presented as acting on standard logic formulas in an implicitly understood context, but they are also commonly given in the literature as acting more explicitly on “judgements”, that is, expressions of the form $\Gamma \vdash \alpha$ where Γ is a set of formulas and α is a formula. This form is typically understood as making the metastatement that there is a proof of α from Γ (Kleene 1962). Following Gentzen 1935 and Prawitz 1965 here we take the former approach. The system NK has no logical axioms and provides two introduction-elimination rules for each logical connective:

Introduction Rules ($\odot$I)		Elimination Rules ($\odot$E)	
&I	$\alpha \quad \beta$ ----- $\alpha \,\&\, \beta$	&E	$\alpha_1 \,\&\, \alpha_2$ ---- α_i (for $i = 1,2$)
$\vee$I	α_i (for $i = 1,2$) ---- $\alpha_1 \vee \alpha_2$	$\vee$E	$\alpha \vee \beta$ [α --- γ] [β --- γ] --------- γ
$\supset$I	[α --- β] ----- $\alpha \supset \beta$	$\supset$E	$\alpha \quad \alpha \supset \beta$ ----- β
$\equiv$I	[α --- β] [β --- α] ----- $\alpha \equiv \beta$	$\equiv$E	α_i ($i = 0,1$) $\quad \alpha_0 \equiv \alpha_1$ ---------- α_{1-i}
~I	[α --- $\bot$] ---- $\sim\alpha$	~E	[$\sim\alpha$ --- $\bot$] ---- α
$\exists$I	$\alpha(t/x)$ ---- $(\exists x)\alpha(x)$	$\exists$E	$(\exists x)\alpha(x)$ [$\alpha(a/x)$ --- β] --------- β

∀I	$\alpha(a/x)$ ---- $(\forall x)\alpha(x)$	∀E	$(\forall x)\alpha(x)$ ---- $\alpha(t/x)$

A few remarks: First, the expression [α --- γ] represents the fact that α is an auxiliary assumption in the proof of γ that eventually gets discharged, i.e. discarded. For example, ∃E tells us that if in the process of constructing a proof one has already derived $(\exists x)\alpha(x)$ and also β with $\alpha(a/x)$ as an auxiliary assumption then the inference to β is allowed. Second, the eigenparameter, *a*, in ∃E and ∀I must be foreign to the premises, undischarged -- “active” -- assumptions, to the rule's conclusion and, in the case of ∃E, to $(\exists x)\alpha(x)$. Third, ⊥ is shorthand for two contradictory formulas, β and $\sim\beta$. Finally, NK is complete: If $\Gamma \vDash \alpha$ then there is a proof of α from Γ using the rules of NK.

As in LK, proofs constructed in NK are represented as trees with the proof's conclusion sitting at the root of the tree, and the problem's assumptions sitting at the leaves. (Proofs are also typically given as sequences of judgements, $\Gamma \vdash \alpha$, running from the top to the bottom of the printed page.) Here is a natural deduction proof tree of $(\exists x)R(x)$ from $(\forall x)(P(x) \vee Q(x))$, $(\forall x)(P(x) \supset R(x))$ and $(\forall x)(Q(x) \supset R(x))$:

	$(\forall x)(P(x) \supset R(x))$ ------- $P(a) \supset R(a)$	$(\forall x)(Q(x) \supset R(x))$ ------- $Q(a) \supset R(a)$
$(\forall x)(P(x) \vee Q(x))$ ------- $P(a) \vee Q(a)$	[$P(a)$ --- $R(a)$] ---------- R(a)	[$Q(a)$ --- $R(a)$] ---------- R(a)

$R(a)$

$(\exists x)R(x)$

As in LK, a forward-chaining strategy for proof construction is not well focused. So, although proofs are *read* forwards, that is, from leaves to root or, logically speaking, from assumptions to conclusion, that is not the way in which they are typically *constructed*. A backward-chaining strategy implemented by applying the rules in reverse order is more effective. Many of the obstacles that were discussed above in the implementation of sequent deduction are applicable to natural deduction as well. These issues can be handled in a similar way, but natural deduction introduces some issues of its own. For example, as suggested by the ⊃-Introduction rule, to prove a goal of the form $\alpha \supset \beta$ one could attempt to prove β on the assumption that α. But note that although the goal $\alpha \supset \beta$ does not match the conclusion of any other introduction rule, it matches the conclusion of all *elimination* rules and the reasoning program would need to consider those routes too. Similarly to forward-chaining, here there is the risk of setting goals

that are irrelevant to the proof and that could lead the program astray. To wit: What prevents a program from entering the never-ending process of building, say, larger and larger conjunctions? Or, what is there to prevent an uncontrolled chain of backward applications of, say, ⊃-Elimination? Fortunately, NK enjoys the subformula property in the sense that each formula entering into a natural deduction proof can be restricted to being a subformula of $\Gamma \cup \Delta \cup \{\alpha\}$, where Δ is the set of auxiliary assumptions made by the ~-Elimination rule. By exploiting the subformula property a natural deduction automated theorem prover can drastically reduce its search space and bring the backward application of the elimination rules under control (Portoraro 1998, Sieg and Byrnes 1996). Further gains can be realized if one is willing to restrict the scope of NK's logic to its intuitionistic fragment where every proof has a normal form in the sense that no formula is obtained by an introduction rule and then is eliminated by an elimination rule (Prawitz 1965).

Implementations of automated theorem proving systems using NK deduction have been motivated by the desire to have the program reason with precisely the same proof format and methods employed by the human user. This has been particularly true in the area of education where the student is engaged in the interactive construction of formal proofs in an NK-like calculus working under the guidance of a theorem prover ready to provide assistance when needed (Portoraro 1994, Suppes 1981). Other, research-oriented, theorem provers true to the spirit of NK exist (Pelletier 1998) but are rare.

The Matrix Connection Method

The name of the matrix connection method (Bibel 1981) is indicative of the way it operates. The term “matrix” refers to the form in which the set of logic formulas expressing the problem is represented; the term “connection” refers to the way the method operates on these formulas. To illustrate the method at work, we will use an example from propositional logic and show that R is entailed by $P \vee Q$, $P \supset R$ and $Q \supset R$. This is done by establishing that the formula

$$(P \vee Q) \,\&\, (P \supset R) \,\&\, (Q \supset R) \,\&\, {\sim}R$$

is unsatisfiable. To do this, we begin by transforming it into conjunctive normal form:

$$(P \vee Q) \,\&\, ({\sim}P \vee R) \,\&\, ({\sim}Q \vee R) \,\&\, {\sim}R$$

This formula is then represented as a matrix, one conjunct per row and, within a row, one disjunct per column:

P	Q
$\sim P$	R
$\sim Q$	R
$\sim R$	

The idea now is to explore all the possible vertical paths running through this matrix. A **vertical path** is a set of literals selected from each row in the matrix such that each literal comes from a different row. The vertical paths:

Path 1	*P*, ~*P*, ~*Q* and ~*R*
Path 2	*P*, ~*P*, *R* and ~*R*
Path 3	*P*, *R*, ~*Q* and ~*R*
Path 4	*P*, *R*, *R* and ~*R*
Path 5	*Q*, ~*P*, ~*Q* and ~*R*
Path 6	*Q*, ~*P*, *R* and ~*R*
Path 7	*Q*, *R*, ~*Q* and ~*R*
Path 8	*Q*, *R*, *R* and ~*R*

A path is **complementary** if it contains two literals which are complementary. For example, Path 2 is complementary since it has both *P* and ~*P* but so is Path 6 since it contains both *R* and ~*R*. Note that as soon as a path includes two complementary literals there is no point in pursuing the path since it has itself become complementary. This typically allows for a large reduction in the number of paths to be inspected. In any event, all the paths in the above matrix are complementary and this fact establishes the unsatisfiability of the original formula. This is the essence of the matrix connection method. The method can be extended to predicate logic but this demands additional logical apparatus: Skolemnization, variable renaming, quantifier duplication, complementarity of paths via unification, and simultaneous substitution across all matrix paths (Bibel 1981, Andrews 1981). Variations of the method have been implemented in reasoning programs in higher-order logic (Andrews 1981) and non-classical logics (Wallen 1990).

Term Rewriting

Equality is an important logical relation whose behavior within automated deduction deserves its own separate treatment. **Equational logic** and, more generally, **term rewriting** treat equality-like equations as **rewrite rules**, also known as reduction or demodulation rules. An equality statement like $f(a)= a$ allows the simplification of a term like $g(c,f(a))$ to $g(c,a)$. However, the same equation also has the potential to generate an unboundedly large term: $g(c,f(a))$, $g(c,f(f(a)))$, $g(c,f(f(f(a))))$, … . What distinguishes term rewriting from equational logic is that in term rewriting equations are used as unidirectional reduction rules as opposed to equality which works in both directions. Rewrite rules have the form $t_1 \Rightarrow t_2$ and the basic idea is to look for terms t occurring in expressions e such that t unifies with t_1 with unifier θ so that the occurrence $t_1\theta$ in $e\theta$ can be replaced by $t_2\theta$. For example, the rewrite rule $x + 0 \Rightarrow x$ allows the rewriting of $succ(succ(0) + 0)$ as $succ(succ(0))$.

To illustrate the main ideas in term rewriting, let us explore an example involving symbolic differentiation (the example and ensuing discussion are adapted from Chapter 1 of Baader and Nipkow 1998). Let *der* denote the derivative respect to x, let y be a variable different from x, and let u and v be variables ranging over expressions. We define the rewrite system:

(R1) $der(x) \Rightarrow 1$
(R2) $der(y) \Rightarrow 0$
(R3) $der(u + v) \Rightarrow der(u) + der(v)$
(R4) $der(u \times v) \Rightarrow (u \times der(v)) + (der(u) \times v)$

Again, the symbol $\Rightarrow$ indicates that a term matching the left-hand side of a rewrite rule should be replaced by the rule's right-hand side. To see the differentiation system at work, let us compute the derivative of $x \times x$ respect to x, $der(x \times x)$:

$$\begin{aligned} der(x \times x) &\Rightarrow (x \times der(x)) + (der(x) \times x) && \text{by R4} \\ &\Rightarrow (x \times 1) + (der(x) \times x) && \text{by R1} \\ &\Rightarrow (x \times 1) + (1 \times x) && \text{by R1} \end{aligned}$$

At this point, since none of the rules (R1)-(R4) applies, no further reduction is possible and the rewriting process ends. The final expression obtained is called a **normal form**, and its existence motivates the following question: Is there an expression whose reduction process will never terminate when applying the rules (R1)-(R4)? Or, more generally: Under what conditions a set of rewrite rules will always stop, for any given expression, at a normal form after finitely many applications of the rules? This fundamental question is called the **termination** problem of a rewrite system, and we state without proof that the system (R1)-(R4) meets the termination condition.

There is the possibility that when reducing an expression, the set of rules of a rewrite system could be applied in more than one way. This is actually the case in the system (R1)-(R4) where in the reduction of $der(x \times x)$ we could have applied R1 first to the second sub-expression in $(x \times der(x)) + (der(x) \times x)$, as shown below:

$$\begin{aligned} der(x \times x) &\Rightarrow (x \times der(x)) + (der(x) \times x) && \text{by R4} \\ &\Rightarrow (x \times der(x)) + (1 \times x) && \text{by R1} \\ &\Rightarrow (x \times 1) + (1 \times x) && \text{by R1} \end{aligned}$$

Following this alternative course of action, the reduction terminates with the same normal form as in the previous case. This fact, however, should not be taken for granted: A rewriting system is said to be **(globally) confluent** if and only if independently of the order in which its rules are applied every expression always ends up being reduced to its one and only normal form. It can be shown that (R1)-(R4) is confluent and, hence, we are entitled to say: "Compute *the* derivative of an expression" (as opposed to

simply "*a*" derivative). Adding more rules to a system in an effort to make it more practical can have undesired consequences. For example, if we add the rule

(R5) $u + 0 \Rightarrow u$

to (R1)-(R4) then we will be able to further reduce certain expressions but at the price of losing confluency. The following reductions show that $der(x + 0)$ now has two normal forms:

$$\begin{aligned} der(x + 0) &\Rightarrow der(x) + der(0) && \text{by R3} \\ &\Rightarrow 1 + der(0) && \text{by R1} \end{aligned}$$

$$\begin{aligned} der(x + 0) &\Rightarrow der(x) && \text{by R5} \\ &\Rightarrow 1 && \text{by R1} \end{aligned}$$

Adding the rule, (R6) $der(0) \Rightarrow 0$, would allow the further reduction of $1 + der(0)$ to $1 + 0$ and then, by R5, to 1. Although the presence of this new rule actually increases the number of alternative paths -- $der(x + 0)$ can now be reduced in four possible ways -- they all end up with the same normal form, namely 1. This is no coincidence as it can be shown that (R6) actually restores confluency. This motivates another fundamental question: Under what conditions can a non-confluent system be made into an equivalent confluent one? The Knuth-Bendix **completion** algorithm (Knuth and Bendix 1970) gives a partial answer to this question.

Term rewriting, like any other automated deduction method, needs strategies to direct its application. Rippling (Bundy, Stevens and Harmelen 1993, Basin and Walsh 1996) is a heuristic that has its origins in inductive theorem-proving that uses annotations to selectively restrict the rewriting process.

Mathematical Induction

Mathematical induction is a very important technique of theorem proving in mathematics and computer science. Problems stated in terms of objects or structures that involve recursive definitions or some form of repetition invariably require mathematical induction for their solving. In particular, reasoning about the correctness of computer systems requires induction and an automated reasoning program that effectively implements induction will have important applications.

To illustrate the need for mathematical induction, assume that a property $\wp$ is true of the number zero and also that if true of a number then is true of its successor. Then, with our deductive systems, we can deduce that for any given number n, $\wp$ is true of it, $\wp(n)$. But we cannot deduce that $\wp$ is true of all numbers, $(\forall x)\wp(x)$; this inference step requires the rule of mathematical induction:

$$\frac{\alpha(0) \quad [\alpha(n) \text{ --- } \alpha(succ(n))]}{(\forall x)\alpha(x)} \quad \text{(mathematical induction)}$$

In other words, to prove that $(\forall x)\alpha(x)$ one proves that $\alpha(0)$ is the case, and that $\alpha(succ(n))$ follows from the assumption that $\alpha(n)$. The implementation of induction in a reasoning system presents very challenging search control problems. The most important of these is the ability to determine the particular way in which induction will be applied during the proof, that is, finding the appropriate induction schema. Related issues include selecting the proper variable of induction, and recognizing all the possible cases for the base and the inductive steps.

Nqthm (Boyer and Moore 1979) has been one of the most successful implementations of automated inductive theorem proving. In the spirit of Gentzen, Boyer and Moore were interested in how people prove theorems by induction. Their theorem prover is written in the functional programming language Lisp which is also the language in which theorems are represented. For instance, to express the commutativity of addition the user would enter the Lisp expression (EQUAL (PLUS X Y) (PLUS Y X)). Everything defined in the system is a functional term, including its basic “predicates”: T, F, EQUAL X Y, IF X Y Z, AND, NOT, etc. The program operates largely as a black box, that is, the inner working details are hidden from the user; proofs are conducted by rewriting terms that posses recursive definitions, ultimately reducing the conclusion's statement to the T predicate. The Boyer-Moore theorem prover has been used to check the proofs of some quite deep theorems (Boyer, Kaufmann, and Moore 1995). Lemma caching, problem statement generalization, and proof planning are techniques particularly useful in inductive theorem proving (Bundy, Harmelen and Hesketh 1991).

Higher-Order Logic

Higher-order logic differs from first-order logic in that quantification over functions and predicates is allowed. The statement “*Any two people are related to each other in one way or another*” can be legally expressed in higher-order logic as $(\forall x)(\forall y)(\exists R)R(x,y)$ but not in first-order logic. Higher-order logic is inherently more expressive than first-order logic and is closer in spirit to actual mathematical reasoning. For example, the notion of set finiteness cannot be expressed as a first-order concept. Due to its richer expressiveness, it should not come as a surprise that implementing an automated theorem prover for higher-order logic is more challenging than for first-order logic. This is largely due to the fact that unification in higher-order logic is more complex than in the first-order case: unifiable terms do not always posess a most general unifier, and higher-order unification is itself undecidable. Finally, given that higher-order logic is incomplete, there are always proofs that will be entirely out of reach for any automated reasoning program.

Methods used to automate first-order deduction can be adapted to higher-order logic. TPS (Andrews et

al. 1996) is a theorem proving system for higher-order logic that uses Church's typed λ-calculus as its logical representation language and is based on a connection-type deduction mechanism that incorporates Huet's unification algorithm (Huet 1975). As a sample of the capabilities of TPS, the program has proved automatically that a subset of a finite set is finite, the equivalence among several formulations of the Axiom of Choice, and Cantor's Theorem that a set has more subsets than members. The latter was proved by the program by asserting that there is no onto function from individuals to sets of individuals, with the proof proceeding by a diagonal argument. HOL (Gordon and Melham 1993) is another higher-order proof development system primarily used as an aid in the development of hardware and software safety-critical systems. HOL is based on the LCF approach to interactive theorem proving (Gordon, Milner and Wadsworth 1979), and it is built on the strongly typed functional programming language ML. HOL, like TPS, can operate in automatic and interactive mode. Availability of the latter mode is welcomed since the most useful automated reasoning systems may well be those which place an emphasis on interactive theorem proving (Farmer, Guttman and Thayer 1993) and can be used as assistants operating under human guidance. Isabelle (Paulson 1994) is a generic, higher-order, framework for rapid prototyping of deductive systems. Object logics can be formulated within Isabelle's metalogic by using its many syntactic and deductive tools. Isabelle also provides some ready-made theorem proving environments, including Isabelle/HOL, Isabelle/ZF and Isabelle/FOL, which can be used as starting points for applications and further development by the user. Isabelle/ZF has been used to prove equivalent formulations of the Axiom of Choice, formulations of the Well-Ordering Principle, as well as the key result about cardinal arithmetic that, for any infinite cardinal κ, $\kappa \circ \kappa = \kappa$ (Paulson and Grabczewski 1996).

Non-classical Logics

Non-classical logics such as modal logics, intuitionsitic logic, multi-valued logics, autoepistemic logics, non-monotonic reasoning, commonsense and default reasoning, relevant logic, paraconsistent logic, and so on, have been increasingly gaining the attention of the automated reasoning community. One of the reasons has been the natural desire to extend automated deduction techniques to new domains of logic. Another reason has been the need to mechanize non-classical logics as an attempt to provide a suitable foundation for artificial intelligence. A third reason has been the desire to attack some problems that are combinatorially too large to be handled by paper and pencil. Indeed, some of the work in automated non-classical logic provides a prime example of automated reasoning programs at work. To illustrate, the Ackerman Constant Problem asks for the number of non-equivalent formulas in the relevant logic R. There are actually 3,088 such formulas (Slaney 1984) and the number was found by “sandwiching” it between a lower and an upper limit, a task that involved constraining a vast universe of 20^{400} 20-element models in search of those models that rejected non-theorems in R. It is safe to say that this result would have been impossible to obtain without the assistance of an automated reasoning program.

There have been three basic approaches to automate the solving of problems in non-classical logic (McRobie 1991). One approach has been, of course, to try to mechanize the non-classical deductive calculi. Another has been to simply provide an equivalent formulation of the problem in first-order logic and let a classical theorem prover handle it. A third approach has been to formulate the semantics of the

non-classical logic in a first-order framework where resolution or connection-matrix methods would apply.

Modal logic. Modal logics find extensive use in computing science as logics of knowledge and belief, logics of programs, and in the specification of distributed and concurrent systems. Thus, a program that automates reasoning in a modal logic such as K, K4, T, S4, or S5 would have important applications. With the exception of S5, these logics share some of the important metatheoretical results of classical logic, such as cut-elimination, and hence cut-free (modal) sequent calculi can be provided for them, along with techniques for their automation. Connection methods (Andrews 1981, Bibel 1981) have played an important role in helping to understand the source of redundancies in the search space induced by these modal sequent calculi and have provided a unifying framework not only for modal logics but also for intuitionistic and classical logic as well (Wallen 1990).

Intutionistic logic. There are different ways in which intuitionsitic logic can be automated. One is to directly implement the intuitionistic versions of Gentzen's sequent and natural deduction calculi, LJ and NJ respectively. This approach inherits the stronger normalization results enjoyed by these calculi allowing for a more compact mechanization than their classical counterparts. Another approach at mechanizing intuitionistic logic is to exploit its semantic similarities with the modal logic S4 and piggy back on an automated implementation of S4. Automating intuitionistic logic has applications in software development since writing a program that meets a specification corresponds to the problem of proving the specification within an intuitionistic logic (Martin-Löf 1982). A system that automated the proof construction process would have important applications in algorithm design but also in constructive mathematics. Nuprl (Constable et al. 1986) is a computer system supporting a particular mathematical theory, namely constructive type theory, and whose aim is to provide assistance in the proof development process. The focus is on logic-based tools to support programming and on implementing formal computational mathematics. Over the years the scope of the Nuprl project has expanded from “proofs-as-programs” to “systems-as-theories”.

Logic Programming

Logic programming, particularly represented by the language **Prolog** (Colmerauer et al. 1973), is probably the most important and widespread application of automated theorem proving. During the early 1970s, it was discovered that logic could be used as a programming language (Kowalski 1974). What distinguishes logic programming from other traditional forms of programming is that logic programs, in order to solve a problem, do not explicitly state *how* a specific computation is to be performed; instead, a logic program states *what* the problem is and then delegates the task of actually solving it to an underlying theorem prover. In Prolog, the theorem prover is based on a refinement of resolution known as SLD-resolution. SLD-resolution is a variation of linear input resolution that incorporates a special rule for selecting the next literal to be resolved upon; SLD-resolution also takes into consideration the fact that, in the computer's memory, the literals in a clause are actually ordered, that is, they form a sequence as opposed to a set. A Prolog **program** consists of clauses stating known facts and rules. For example, the following clauses make some assertions about flight connections:

flight(*toronto*, *london*).
flight(*london*,*rome*).
flight(*chicago*,*london*).
flight(*X*,*Y*) :- *flight*(*X*,*Z*) , *flight*(*Z*,*Y*).

The clause *flight*(*toronto*, *london*) is a fact while *flight*(*X*,*Y*) :- *flight*(*X*,*Z*) , *flight*(*Z*,*Y*) is a rule, written by convention as a reversed conditional (the symbol ":-" means "if"; the comma means "and"; terms starting in uppercase are variables). The former states that there is flight connection between Toronto and London; the latter states that there is a flight between cities *X* and *Y* if, for some city *Z*, there is a flight between *X* and *Z* and one between *Z* and *Y*. Clauses in Prolog programs are a special type of Horn clauses having precisely one positive literal: **Facts** are program clauses with no negative literals while **rules** have at least one negative literal. (Note that in standard clause notation the program rule in the previous example would be written as *flight*(*X*,*Y*) ∨ ~*flight*(*X*,*Z*) ∨ ~*flight*(*Z*,*Y*).) The specific form of the program rules is to effectively express statements of the form: "I*f these conditions over here are jointly met then this other fact will follow*". Finally, a **goal** is a Horn clause with no positive literals. The idea is that, once a Prolog program Π has been written, we can then try to determine if a new clause γ, the goal, is entailed by Π, $\Pi \vDash \gamma$; the Prolog prover does this by attempting to derive a contradiction from $\Pi \cup \{\sim\gamma\}$. We should remark that program facts and rules alone cannot produce a contradiction; a goal must enter into the process. Like input resolution, SLD-resolution is not refutation complete for first-order logic but it is complete for the Horn logic of Prolog programs. The fundamental theorem: If Π is a Prolog program and γ is the goal clause then $\Pi \vDash \gamma$ iff $\Pi \cup \{\sim\gamma\} \vdash$ [ ] by SLD-resolution (Lloyd 1984).

For instance, to find out if there is a flight from Toronto to Rome one asks the Prolog prover to see if the clause *flight*(*toronto*, *rome*) follows from the given program. To do this, the prover adds ~*flight*(*toronto*, *rome*) to the program clauses and attempts to derive the empty clause, [ ], by SLD-resolution:

1	*flight*(*toronto*,*london*)	Program clause
2	*flight*(*london*,*rome*)	Program clause
3	*flight*(*chicago*,*london*)	Program clause
4	*flight*(*X*,*Y*) ∨ ~*flight*(*X*,*Z*) ∨ ~*flight*(*Z*,*Y*)	Program clause
5	~*flight*(*toronto*,*rome*)	Negation of the conclusion
6	~*flight*(*toronto*,*Z*) ∨ ~*flight*(*Z*,*rome*)	Res 5 4
7	~*flight*(*london*,*rome*)	Res 6 1
8	[ ]	Res 7 2

The conditional form of rules in Prolog programs adds to their readability and also allows reasoning about the underlying refutations in a more friendly way: To prove that there is a flight between Toronto and Rome, *flight*(*toronto*,*rome*), unify this clause with the consequent *flight*(*X*,*Y*) of the fourth clause in

the program which itself becomes provable if both *flight*(*toronto*,*Z*) and *flight*(*Z*,*rome*) can be proved. This can be seen to be the case under the substitution $\{Z \leftarrow london\}$ since both *flight*(*toronto*,*london*) and *flight*(*london*,*rome*) are themselves provable. Note that the theorem prover not only establishes that there is a flight between Toronto and Rome but it can also come up with an actual itinerary, Toronto-London-Rome, by extracting it from the unifications used in the proof.

There are at least two broad problems that Prolog must address in order to achieve the ideal of a logic programming language. Logic programs consist of facts and rules describing what is true; anything that is not provable from a program is deemed to be false. In regards to our previous example, *flight*(*toronto*, *boston*) is not true since this literal cannot be deduced from the program. The identification of falsity with non-provability is further exploited in most Prolog implementations by incorporating an operator, **not**, that allows programmers to explicitly express the negation of literals (or even subclauses) within a program. By definition, not *l* succeeds if the literal *l* itself fails to be deduced. This mechanism, known as **negation-by-failure**, has been the target of criticism. Negation-by-failure does not fully capture the standard notion of negation and there are significant logical differences between the two. Standard logic, including Horn logic, is monotonic which means that enlarging an axiom set by adding new axioms simply enlarges the set of theorems derivable from it; negation-by-failure, however, is non-monotonic and the addition of new program clauses to an existing Prolog program may cause some goals to cease from being theorems. A second issue is the **control problem**. Currently, programmers need to provide a fair amount of control information if a program is to achieve acceptable levels of efficiency. For example, a programmer must be careful with the order in which the clauses are listed within a program, or how the literals are ordered within a clause. Failure to do a proper job can result in an inefficient or, worse, non-terminating program. Programmers must also embed hints within the program clauses to prevent the prover from revisiting certain paths in the search space (by using the **cut** operator) or to prune them altogether (by using **fail**). Last but not least, in order to improve their efficiency, many implementations of Prolog do not implement unification fully and bypass a time-consuming yet critical test -- the so-called occurs-check -- responsible for checking the suitability of the unifiers being computed. This results in an unsound calculus and may cause a goal to be entailed by a Prolog program (from a computational point of view) when in fact it should not (from a logical point of view).

There are variations of Prolog intended to extend its scope. By implementing a model elimination procedure, the Prolog Technology Theorem Prover (PPTP) (Stickel 1992) extends Prolog into full first-order logic. The implementation achieves both soundness and completeness. Moving beyond first-order logic, λProlog (Miller and Nadathur 1988) bases the language on higher-order constructive logic.

Conclusion

Automated reasoning is a growing field that provides a healthy interplay between basic research and application. Automated deduction is being conducted using a multiplicity of theorem-proving methods, including resolution, sequent calculi, natural deduction, matrix connection methods, term rewriting, mathematical induction, and others. These methods are implemented using a variety of logic formalisms such as first-order logic, type theory and higher-order logic, clause and Horn logic, non-classical logics,

and so on. Automated reasoning programs are being applied to solve a growing number of problems in formal logic, mathematics and computer science, logic programming, software and hardware verification, circuit design, and many others. One of the results of this variety of formalisms and automated deduction methods has been the proliferation of a large number of theorem proving programs. To test the capabilities of these different programs, selections of problems has been proposed against which their performance can be measured (McCharen, Overbeek and Wos 1976, Pelletier 1986). The TPTP (Sutcliffe and Suttner 1998) is a library of such problems that is updated on a regular basis. There is also a competition among automated theorem provers held regularly at the CADE conference; the problems for the competition are selected from the TPTP library.

Initially, computers were used to aid scientists with their complex and often tedious numerical calculations. The power of the machines was then extended from the numeric into the symbolic domain where infinite-precision computations performed by computer algebra programs have become an everyday affair. The goal of automated reasoning has been to further extend the machine's reach into the realm of deduction where they can be used as reasoning assistants in helping their users establish truth through proof.

Bibliography

- Andrews, P. B., M. Bishop, S. Issar, D. Nesmith, F. Pfenning and H. Xi, 1996, "TPS: A Theorem-Proving System for Classical Type Theory", *Journal of Automated Reasoning*, Vol. 16, No. 3, pp.321-353.
- Andrews, P. B., 1981, "Theorem-Proving via General Matings", *JACM*, Vol. 28, No. 2, pp. 193-214.
- Bibel, W., 1981, "On Matrices with Connections", *JACM*, Vol. 28, No. 4, pp. 633-645.
- Bledsoe, W. W., 1977, "Non-resolution Theorem Proving", *Artificial Intelligence*, Vol. 9, pp. 1-35.
- Bledsoe, W. W. and M. Tyson, 1975, "The UT Interactive Prover", *Memo ATP-17A*, Dept. of Mathematics, University of Texas.
- Boyer, R. S., M. Kaufmann and J. S. Moore, 1995, "The Boyer-Moore Theorem Prover and its Interactive Enhancement", *Computers and Mathematics with Applications*, Vol. 29, pp. 27-62.
- Boyer, R.S. and J. S. Moore, 1979, *A Computational Logic*, Academic Press, New York.
- Baader, F. and T. Nipkow, 1998, *Term Rewriting and All That*, Cambridge University Press.
- Bundy, A., F. van Harmelen, J. Hesketh and A. Smaill, 1991, "Experiments with Proof Plans for Induction", *Journal of Automated Reasoning*, Vol. 7, No. 3, pp. 303-324.
- Bundy, A., A. Stevens, F. van Harmelen, A. Ireland and A. Smaill, 1993, "Rippling: A Heuristic for Guiding Inductive Proofs", *Artificial Intelligence*, Vol. 62, pp. 185-253.
- Basin, D. A. and T. Walsh, 1996, "A Calculus for and Termination of Rippling", *Journal of Automated Reasoning*, Vol. 16, No. 1/2, pp. 147-180.
- Church, A., 1940, "A Formulation of the Simple Theory of Types", *Journal of Symbolic Logic*, Vol. 5, pp. 56-68.
- Chang, C. L. and R. C. T. Lee, 1973, *Symbolic Logic and Mechanical Theorem Proving*,

Academic Press.

- Clocksin, W. F. and C. S. Mellish, 1981, *Programming in Prolog*, Springer-Verlag.
- Colmerauer, A., H. Kanoui, R. Pasero and P. Roussel, 1973, *Un Système de Communication Homme-machine en Français*, Rapport, Groupe Intelligence Artificielle, Université d'Aix Marseille.
- Constable, R. L., S. F. Allen, H. M. Bromley, W.R. Cleaveland, J. F. Cremer, R. W. Harper, D. J. Howe, T. B. Knoblock, N. P. Mendler, P. Panangaden, J. T. Sasaki and S. F. Smith, 1986, *Implementing Mathematics with the Nuprl Proof Development System*, Prentice Hall.
- Fitting, M., 1990, *First-Order Logic and Automated Theorem Proving*, Springer-Verlag.
- Farmer, W. M., J. D. Guttman and F. J. Thayer, 1993, "IMPS: An Interactive Mathematical Proof System", *Journal of Automated Reasoning*, Vol. 11, No. 2., pp. 213-248.
- Gentzen, G., 1935, "Investigations into Logical Deduction", in Szabo (1969), pp.68-131.
- Gordon, M. J. C. and T. F. Melham, eds., 1993, *Introduction to HOL: A Theorem Proving Environment for Higher Order Logic*, Cambridge University Press.
- Gordon, M. J. C., A. J. Milner and C. P. Wadsworth, 1979, *Edinburgh LCF: A Mechanised Logic of Computation*, LNCS 78, Springer-Verlag.
- Hilbert, D. and W. Ackermann, 1928, *Principles of Mathematical Logic*, Trans. from 1938 ed., Chelsea Publishing Co., New York, 1950.
- Herbrand, J., 1930, *Recherches sur la Theorie de la Demonstration*, Travaux de la Societé des Sciences at des Lettres de Varsovie, Classe III, Science Mathématique et Physique, No. 33, 128.
- Huet, G. P., 1975, "A Unification Algorithm for Typed λ-calculus", *Theoretical Computer Science*, Vol. 1, pp. 27-57.
- Knuth, D. and P. B. Bendix, 1970, "Simple Word Problems in Universal Algebras", *Computational Problems in Abstract Algebra*, J. Leech, ed., pp. 263-297, Pergamon Press.
- Kleene, S. C., 1962, *Introduction to Metamathematics*, North-Holland.
- Kowalski, R., 1974, "Predicate Logic as a Programming Language", *Proc. IFIP 74*, North Holland, pp. 569-574.
- Lloyd, J. W., 1984, *Foundations of Logic Programming*, Springer-Verlag.
- Loveland, D. W., 1969, "A Simplified Format for the Model Elimination Procedure", *JACM*, Vol. 16, pp. 349-363.
- Loveland, D. W., 1970, "A Linear Format for Resolution", *Proc. IRIA Symp. Automatic Demonstration*, Springer-Verlag, New York, pp. 147-162.
- Loveland, D. W., 1978, *Automated Theorem Proving: A Logical Basis*, North Holland.
- Luckham, D., 1970, "Refinements in Resolution Theory", *Proc. IRIA Symp. Automatic Demonstration*, Springer-Verlag, New York, pp. 163-190.
- Martin-Löf, P., 1982, "Constructive Mathematics and Computer Programming", *Logic, Methodology and Philosophy of Science*, Vol. IV, pp. 153-175, North-Holland.
- McCune, W., 1997, "Solution of the Robbins Problem", *Journal of Automated Reasoning*, Vol. 19, No. 3, pp. 263-276.
- McRobie, M. A., 1991, "Automated Reasoning and Nonclassical Logics: Introduction", *Journal of Automated Reasoning*, Vol. 7, No. 4, pp. 447-451.
- Miller, D. and G. Nadathur, 1988, "An Overview of λProlog", *Proceedings of the Fifth International Logic Programming Conference -- Fifth Symposium in Logic Programming*, R.

Bowen and R. Kowalski, ed., MIT Press.

- McCharen, J. D., R. A. Overbeek and L. A. Wos, 1976, "Problems and Experiments for and with Automated Theorem-Proving Programs", *IEEE Transactions on Computers 25 (8)*, pp. 773-782.
- Nivens, A. J., 1974, "A Human-Oriented Logic for Automatic Theorem Proving", *JACM*, Vol. 21, No. 4, pp. 606-621.
- Paulson, L. C., 1994, *Isabelle: A Generic Theorem Prover*, Lecture Notes in Computer Science, Vol. 828, Springer-Verlag.
- Paulson, L. C. and K. Grabczewski, 1996, "Mechanizing Set Theory", *Journal of Automated Reasoning*, Vol. 17, No. 3, pp.291-323.
- Pelletier, F. J., 1986, "Seventy-Five Problems for Testing Automatic Theorem Provers", *Journal of Automated Reasoning*, Vol. 2, No. 2, pp.191-216.
- Pelletier, F. J., 1998, *Studia Logica* 60, No. 1, pp. 3-43.
- Portoraro, F. D., 1994, "Symlog: Automated Advice in Fitch-style Proof Construction", *Proceedings of the 12th International Conference on Automated Deduction*, Nancy, Lecture Notes in Artificial Intelligence, A. Bundy, ed., Vol. 814, pp. 802-806, Springer-Verlag.
- Portoraro, F. D., 1998, "Strategic Construction of Fitch-style Proofs", *Studia Logica*, Vol. 60, No. 1, pp. 45-66.
- Prawitz, D., 1965, *Natural Deduction: A Proof Theoretical Study*, Almqvist & Wiksell, Stockholm.
- Robinson, J. A., 1965, "A Machine Oriented Logic Based on the Resolution Principle", *JACM*, Vol. 12, pp. 23-41.
- Robinson, J. A., 1965, "Automatic Deduction with Hyper-resolution", *Internat. J. Comput. Math.*, Vol. 1, pp. 227-234.
- Sieg, W. and J. Byrnes, 1996, *Normal Natural Deduction Proofs (in Classical Logic)*, Report CMU-PHIL 74, Department of Philosophy, Carnegie-Mellon University.
- Slaney, J. K., 1984, "3,088 Varieties: A Solution to the Ackerman Constant Problem", *Journal of Symbolic Logic* 50, pp. 487-501.
- Sutcliffe, G. and C. Suttner, 1998, "The TPTP Problem Library - CNF Release v1.2.1", *Journal of Automated Reasoning*, Vol. 21, No. 2, pp. 177-203.
- Stickel, M. E., 1992, "A Prolog Technology Theorem Prover: A New Exposition and Implementation in Prolog", *Theoretical Computer Science 104*, pp. 109-128.
- Suppes, P. et al., 1981, "Part I: Interactive Theorem Proving in CAI Courses", *University-Level Computer-Assisted Instruction at Stanford: 1968-1980*, P. Suppes, ed., IMSSS, Stanford University.
- Szabo, M. E., 1969, *The Collected Papers of Gerhard Gentzen*, M. E. Szabo ed., North-Holland.
- Wallen, L. A., 1990, *Automated Deduction in Nonclassical Logics*, MIT Press.
- Wang, H., 1960, "Toward Mechanical Mathematics", in *Automation of Reasoning*, J. Siekmann and G. Wrightson, eds., Vol. 1, pp. 244-264, Springer-Verlag, 1983.
- Wos, L., D. Carson and G. R. Robinson, 1965, "Efficiency and Completeness of the Set of Support Strategy in Theorem Proving", *JACM*, Vol. 12, pp. 698-709.
- Wos, L., R. Overbeek, E. Lusk and J. Boyle, 1984, *Automated Reasoning: Introduction and Applications*, Prentice-Hall.

Other Internet Resources

- Automated Reasoning at Argonne
- Automated Reasoning Group at ANU
- HOL Automated Reasoning Group
- Isabelle
- The Nuprl Project
- TPS Theorem Proving System

Related Entries

artifical intelligence: logic and | logic: classical | reasoning: defeasible

Frederic D. Portoraro
University of Toronto
frederic@cs.toronto.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 18, 2001
Content last modified: July 18, 2001

Giles of Rome

Giles of Rome (who died in 1316 as archbishop of Bourges) was one of the most productive and influential thinkers active at the end of the 13th century, who played a major role also in the political events of his time. Giles of Rome was an extremely prolific author and left a very large corpus of writings, encompassing commentaries on Aristotle, theological treatises, questions, and sermons. In recent years, a research group led by Francesco Del Punta (Scuola Normale Superiore, Pisa, Italy) has been devoting a lot of energy to the project of publishing his *Opera Omnia* and deepening our knowledge of his thought. Although this group has produced extremely significant results, an assessment of Giles' whole work is still in progress. For this reason, the present entry only aims at providing insight into an ongoing process of research and will focus on recent studies on Giles.

Life

Born in Rome most probably in the fifth decade of the thirteenth century, Giles was the first outstanding theologian of the relatively recently founded Order of the Augustinian Hermits. Nothing more is known about his origins: the statement that he belonged to the famous Roman family of the Colonna seems to go back to Jordan of Saxony's *Liber Vitasfratrum* (second half of the 14th century), but is completely missing from contemporary, 13th-century sources. From Giles' will we know that he was sent to Paris to study in the convent of his Order. At the beginning he must have followed the courses either of a secular master or of a theologian belonging to a different Order, as the Augustinian friars did not have a regent master at the time. Probably he was a pupil of Aquinas' in the years 1269-1272. He commented on the Sentences at the beginning of the 1270's. In following years he probably wrote also a large number of his commentaries on Aristotle.

The year 1277 marked a turning-point in his career: Giles was involved in the condemnation of the heterodox Aristotelianism issued by the Parisian bishop Etienne Tempier, although the process against him must be distinguished from the famous decree of the 7 of March 1277, as Robert Wielockx has shown. After 1277 Giles must have abandoned Paris, but his presence is attested in Italy not earlier than 1281. Before leaving Paris he completed his De regimine principum, which is dedicated to the young Philip, the future Philip the Fair.

Between 1281 and 1284 Giles played an important role in the government of his Order, taking part in various chapters held in Italy. At the provincial chapter of Tuscania (nowadays in Lazio, Italy) in 1285 he acted as vicar of the prior general of his Order, Clement of Osimo.

In 1285, Giles'doctrine was examined again; after recanting only a part of what had been previously condemned in 1277, he was allowed to teach again; by 1287 he is referred to as a master of theology. This success enhanced Giles of Rome's authority even more in his Order, whose general chapter of Florence decreed that Giles' works (even future ones) should be considered as the official doctrine of the Order, to be defended by all Augustinian bachelors and masters. In 1292, at the General Chapter of Rome he was elected prior general of his Order.

Benedict Caetani's election to the papal see marked a further radical change in his career, as Boniface VIII appointed him archbishop of Bourges in 1295. As a matter of fact, Giles was very often absent from his see, spending extended periods of time at the papal curia. In his *De renuntiatione*, he defended the legitimacy of Celestine's abdication, and, consequently, of Boniface' s election. When the contrast between Boniface VIII and Philip IV reached its most critical point, he continued to side staunchly with the pope. An important sermon defending the papal position has been recently discovered by Concetta Luna, and Giles' *De ecclesiastica potestate* undoubtedly ranks among the sources of *Unam Sanctam*.

Giles' prestige decreased after Boniface' death, and even more with the rise of Clement V to the papal throne. Before being elected pope, Betrand de Got, as archbishop of Bourdeaux, had had serious conflicts with Giles. This unfavorable change, however, did not prevent Giles from playing a significant role in the debates of his time. Around 1305-6 he took part in a commission which examined and condemned the eucharistic doctrine of John of Paris, a former adversary of Giles' during the conflict between Philip the Fair and Boniface. In the discussions concerning the Templars which eventually led to the violent suppression of the Order, Giles sided with Philip the Fair, attacking the Templars and devoting a whole tract, *Contra exemptos*, to the thesis that their exemption from episcopal jurisdiction was the cause of their abuses. During the Council of Vienne Giles was asked to draw a list of errors extracted from the works of Peter John Olivi: three of them were officially condemned by the Council. On December 22, 1316 he died at the papal Curia at Avignon (For further details see Del Punta-Donati-Luna 1993).

Logic and Rhetoric

After René Antoine Gauthier identified in the master Guillaume Arnauld the real author of the *Lectura*

supra logicam veterem attributed to Giles of Rome (Tabarroni 1988), interest in logical works focused mainly on his Commentary on the *Sophistici Elenchi*. In his treatment of the *fallacia figurae dictionis* Giles proves to be a brilliant and original representative of “modistic logic”, who bases his solutions mainly on the semantics of *intentiones* (Tabarroni 1991). Alessandro Conti (1992) investigated Giles’ Commentary on *Posterior Analytics* as an example of his theory of truth, which brings to completion Aquinas’ shift from Augustine’s influence to the Aristotelian approach . Giles of Rome also authored the most important Commentary on the Rhetoric of the Latin medieval tradition, which earned him the honorific title of *expositor* of this book and influenced all later medieval commentaries. Costantino Marmo studied Giles’ approach to the different translations to which he had access and showed how he developed Aquinas’ theory of passions in commenting on the relevant portions of Aristotle’s text (Marmo 1991), trying to solve some problems left open by Aquinas. It has been suggested that Giles considers rhetoric to be a sort of “logic” of ethics and politics: this brilliant interpretation still awaits, however, full development and articulation (Staico 1992).

Metaphysics

Traditionally, Giles was described as a “faithful” disciple of Aquinas’. As a matter of fact, his “philosophical project” tends to discuss critically Aquinas’ position in order to improve the solutions he offered, without, however, trying to discard them radically. This holds true, e. g., for one of the most famous topics of his discussion with Henry of Ghent, the distinction between “essence” and “existence”. In this case, Giles radicalizes Aquinas’ doctrine of the real distinction, asserting that existence must be conceived as a “*res addita*” to essence. Although the final result of his theory was considered closer to Avicenna’s solution than to that of Aquinas, Giles nevertheless develops it starting from Aquinas’ own position (Wippel 1981). The connection of this issue with the discussion with Henry of Ghent about the concept of “creation” was deeply investigated by Giorgio Pini (1992), who could show how Giles, while defending Aquinas and thereby the possibility of using some Aristotelian principles in an orthodox account of creation, goes beyond the positions of the Dominican master, e. g. asserting the identity of “*esse*” and “*creatio*”.

As to the debate about the unicity of substantial form, Giles’ position evolves during time. If we leave aside the *Errores philosophorum*, since the authenticity of this work has been contested with serious arguments (Bruni 1935, Koch 1944, Donati 1990b, Luna 1990), we can notice that Giles changes his position from the *Contra gradus et pluralitatem formarum* (between the end of 1277 and the beginning of 1278), where he denies plurality of forms for every compound, to later works, where he takes a more cautious stance in particular concerning man. The principle of individuation is identified, as in Aquinas, with “materia signata quantitate”, that is, matter designated by its dimensions. According to Giles, who criticizes Richard of Mediavilla on this point, matter is pure potentiality and therefore cannot be distinguished into different kinds. For this reason, he cannot accept Aquinas’ doctrine that the incorruptibility of celestial bodies derives from the peculiar nature of their matter (Donati 1986). In Giles’ opinion celestial bodies are not incorruptible because their matter is different from the matter of sublunar bodies, but rather because their *quantitas materiae* cannot change its determinate dimensions. This is but one application of Giles’ famous doctrine of “indeterminate dimensions”. Modifying

Averroes' doctrine in this respect, Giles argues that a portion of matter, in order to be able to receive a form, needs to possess already a sort of quantity. Such quantity, however, should not be identified with the determinate dimensions a body possesses, but is rather a *quantitas* which remains the same during processes such as rarefaction and condensation. Giles' notion of *quantitas materiae* , which is not only generically extension or three-dimensionality, but seems to represent an unchangeable given "amount" of matter pertaining to a body, has been considered comparable, some difficulties notwithstanding, to the modern notion of mass (Donati 1988).

After the condemnation of 1277, a significant change can be noticed in Giles' position also in his solution of the problem of the eternity of the world. At the beginning of his career he admitted the theoretical possibility of the eternity of the world, although rejecting Aristotle's arguments proving the actual eternity of the world. Later he shifted to a more " Augustinian" stance, rejecting the hypothesis of a creation "*ab aeterno*" and admitting that it is possible to prove the temporality of creation, although he finds that no conclusive argument has been advanced so far. Giles was much more steadfast in his opposition to another major tenet of "Averroistic" doctrine, that is the unicity of the possible intellect. He maintained that the possibility of actual knowledge on the part of the individual depends necessarily on the fact that each body is informed by its own intellective soul, which is its form. For the same reasons Giles also rejected the unicity of the agent intellect, a doctrine he attributed to Avicenna (Del Punta-Donati-Luna 1993).

Natural Philosophy

Recent studies concerning Giles' natural philosophy focused mainly on his treatment of some pivotal concepts of Aristotle's Physics. Cecilia Trifogli opened new perspectives in this field, devoting her attention to the notions of place and motion (especially in the void, see Trifogli 1992), underlining that "Giles' emphasis on the role of place in the description of motion seems to lead to a quantitative and relational notion of place. Giles, however, does not completely substitute the Aristotelian notion of place for that of place as distance. Place as distance is only one of the two notions of place which appear in his commentary. The other, which is related to material place, assumes an intrinsic connection between place and the located body that cannot be founded on distance alone" (Trifogli 1990a, 350). Trifogli also investigated Giles' notions of time and infinity, emphasizing that his whole approach to natural philosophy is distinguished by a tendency towards a metaphysical interpretation of Aristotelian concepts, as opposed to a physical and quantitative one (Trifogli 1990b, 1991). For example, Giles conceives of time not as a quantity pertaining to every kind of motion, but rather as a mode in which motion exists. His concept of time rests in fact essentially on a broad notion of succession, which allows him on one hand to retain the unity of the concept of time, but, on the other hand, to acknowledge the existence of different temporal forms (Trifogli 1990b). This attitude emerges also from the analysis of Giles' controversy on angelic time with Henry of Ghent (Porro 1988, Porro 1991, but see also Faes de Mottoni 1983). Both authors thought that the time in which angels exist, unlike sublunar time, is a discrete succession of instants. Giles and Henry disagreed, however, on the relationship existing between angelic and sublunar time. In particular, Henry rejected Giles' thesis that more instants of angelic time can correspond to one and the same instant of sublunar time. This difference of opinion rested in part on

diverging concepts of angelic motion, which can be istantaneous according to Henry, but not according to Giles.

Between Philosophy and Medicine

Between 1285 and 1290, Giles took a stance in the much debated question of the respective roles of male and female parents in conception. The Galenist view, going back to Hippocrates, was that both male and female contributed sperm, so that the offspring could have characteristics from both parents. On the contrary, Aristotle had held that only the male alone contributed sperm containing an active and formal principle to conception, while the female provided only the matter of the fetus. Giles was well acquainted with these different positions and with the efforts to reconcile the diverging approaches of *medici* and *philosophi*, which could be traced back to Avicenna. Leaning on Averroes' *Colliget*, however, Giles rejected any attempt to attribute a formal role to the female sperm, even if it is conceived as subordinate to the male one. On the contrary he maintained that it can contribute only in a passive way to conception, while what was called "female sperm", i.e., the vaginal secretion, has a subservient, helpful but by no means necessary function. It helps the male sperm to inseminate female matter, but does not add anything essential to the new being. In this way, Giles intended to stress the superiority of the philosophical, theoretical approach to such problems with respect to the traditions of medical learning, even when they seemed to be supported by empirical evidence (Hewson 1975; Martorelli Vico 1988). After conception the human embryo begins a development which goes through different stages. Comparing these stages to the embryos of various animals Giles, like Thomas Aquinas, supported an interpretation of the fetal development which would be exploited many centuries later by the so-called "recapitulation theory" (Hewson 1975, 99). Giles maintained, however, that "the organic fetal body is not to be called a pig, a bear, or a monkey, but something immediately disposed to becoming man" (Hewson 1975, 100). This position apparently implies that human life does not fully begin at the moment of conception. Although such a thesis can be brought to bear on the moral judgment concerning abortion, Giles does not seem interested in tackling from this point of view an issue which would become central for what nowadays is called bioethics.

Ethics and Political Theory

In the debate on the respective roles of intellect and will in the determination of human action Giles is known to have taken and intermediate position, a sort of compromise between the theory of Henry of Ghent and that of Geoffrey of Fontaines. Giles maintains, in fact, that will is a passive potency and can not "move" itself, but always needs an object, a "bonum apprehensum". This starting point however, does not rule out its freedom, because will, once "moved" by its object, can determine itself and other potencies with regard to action. This view of Giles' is consistent with his interpretation of the relationship existing between knowledge and will in the sinner. Committing a sin implies an ignorance of the real good, but this ignorance is not the primary cause of the wrong behavior, because it is an effect of the will, which, affected by *malicia*, corrupts the judgment of the reason (Macken 1977).

Giles of Rome exerted considerable influence also in other fields of ethics, such as the theory of virtues. The most developed expression of his position is not to be found in a Commentary on Aristotle, but rather in his *De regimine principum*, the most successful “mirror of princes” of medieval political thought, which is still conserved in more than 300 manuscripts in its original Latin version, to which many translations in European vernaculars must be added. Written most probably between 1277 and 1280 the *De regimine* is acknowledged to be one of the most successful attempts at mediating Aristotle’s practical philosophy, and in particular his “ethical and political language” to the Latin West. Giles was the first to structure a mirror of princes in three books along the lines of a scheme -- *ethica-oeconomica-politica* -- which played an important role in the reception of Aristotle’s moral and political philosophy in the Middle Ages (Lambertini 1988). The author takes great care to give the impression that he is mainly relying on Aristotle’s text, providing numerous quotations from the *Nichomachean Ethics*, from the *Politics* and from the *Rethoric*. Scholars should not overlook, however, that his reception of Aristotle is not as direct as it can seem and that Giles is deeply influenced by a tradition in the interpretation of Aristotle’s practical philosophy. In this tradition Aquinas plays a very important role for Giles, so that, while Aristotle is the authority who is quoted on almost every occasion, it is the unnamed Aquinas who, with his *Sententia libri Politicorum*, *De regno*, *Summa Theologiae*, exerts a really decisive influence on *De regimine*. While discussing particular topics, Giles skillfully adapts Aristotle to his own purposes. This emerges with clarity in the first book, devoted to ethics, where Giles’ classification of virtues is heavily dependent on the *Summa Theologiae* and, therefore, on Aquinas’ reinterpretation of the Aristotelian heritage. For example, Giles here defines *prudentia* as a *virtus media*, sharing the nature of moral as well as of intellectual ones, a doctrine which can by no means be traced back to the Stagirite (Lambertini 1991, 1992, 1995, 2000).

The most famous example of this selective attitude towards Aristotle’s works, however, belongs rather to the field of political theory. In the third book of *De regimine* Giles wants to prove that Monarchy is the absolutely best form of government. The first arguments he puts forward in favor of monarchy are not taken from Aristotle’s *Politics*, but from Aquinas’ *De regno*. Then some arguments against monarchy which could be read in the *Politics* are presented as objections that Aristotle puts forth for subsequent refutation. At the end, Giles states squarely that Aristotle supports monarchy as the absolutely best form of monarchy and corroborates his assertion with an argument, which, in the *Politics*, actually goes in the opposite direction (Lambertini 1990). One could provide several other examples to show that the *De regimine* succeeded in presenting itself as a simplified exposition and explanation of Aristotle’s thought in practical philosophy, but at the same time transmitted to Giles’ readers a strongly biased interpretation of the Stagirite. The fact that the *De regimine* was often used as a tool to have easier access to Aristotle’s political theory deeply influenced, therefore, the way the Latin West read and understood Aristotle’s *Politics* in the Late Middle Ages. Recent codicological studies on the diffusion of *De regimine’* manuscripts do in fact show that many possessors of the manuscripts most probably used them for study (See *Opera Omnia* I.1/11, *Catalogo dei manoscritti, De regimine*; Briggs 1999).

While in the *De regimine* Giles carefully avoids any reference to the thorny problem of the relationship between secular and ecclesiastical power, his later writings which are relevant for political theory deal first and foremost with ecclesiological problems. This holds true for his treatise *De renuntiatione papae* (1297-1298) where Giles defends the lawfulness of Celestine’s abdication against the arguments put

forward by the Colonna cardinals in their first appeal against Boniface VIII. From the point of view of the history of political thought it is relevant that Giles argues that papal power, although of divine origin, is conferred on a particular individual by a human act, namely, by the election of the cardinals. Here Giles is countering the Colonna arguments that papal dignity cannot cease to reside in a pope until he dies, because the pontificate depends on God's will, and stresses the fact that divine intention in this case becomes effective through the mediation of human agents, that is, through the consent of the electors and of the elected. A jurisdiction which is given by the consent of men, however, can also be removed by their consent through a reverse procedure. This does not amount to saying that the pope can be deposed (except in case of heresy), because, according to Giles, the pope is above the law and has no earthly authority above him. He can however, depose himself, that is, abdicate. Just as for his election the consent of his electors and of the elected was necessary, so also for the removal of the pope from office his consent is decisive (Eastman 1989, 1990, 1992). In this way Giles could dismiss arguments against the validity of Celestine's abdication without admitting the possibility that the pope can be deposed, e. g., by the Council, as Boniface's adversaries maintained.

Much better known than *De renuntiatione* is Giles' *De ecclesiastica potestate*, a treatise also composed in defense of Boniface VIII. Most probably in 1302, Giles systematically expounded in this work the views on the relationship between *regnum* and *sacerdotium* he had already put forward in a recently re-discovered sermon held at the papal curia (Luna 1992). The main tenet of his fully fledged argumentation is that the pope, supreme authority of the Church but also of the whole of mankind, is the only legitimate origin of every power on earth, be it exercised -- as jurisdiction -- on persons, or -- as property -- on things. In his plenitude of power, the pope possesses an absolute supremacy both in the ecclesiastical and in the temporal sphere, and delegates the exercise of the temporal "sword" to lay sovereigns only in order to fulfill most properly his higher religious duties. Any authority that does not recognize its dependence on the papal power is but usurpation. In Giles' view, there is no space even for a partially autonomous temporal order. Coherently, Giles maintains that no property rights are valid if they are not legitimated by papal authority. Interestingly enough, such a claim is also supported by his account of the origin of property, according to which property is not a natural institution, but only the consequence of human agreements, which lack any legitimacy unless they are recognized by the supreme religious power (Miethke 2000).

Bibliography

Primary Sources

The most complete list of Giles' works can be found in Del Punta--Donati--Luna 1993 together with the most reliable attempt at dating them (see also Donati 1990b as far as commentaries on Aristotle are concerned). The same article by Del Punta, Donati and Luna also contains the best available bibliography, which can be complemented with Lezcano 1995, 32-50.It is impossible to reproduce all that information in the present entry. Standard older editions were reprinted in Frankfurt 1967-1970. Among the texts edited in our century I would mention the following:

- *De ecclesiastica potestate*, ed. R. Scholz, Weimar 1929 (English translation in R. W. Dyson, *Giles of Rome on Ecclesiastical Power. The De ecclesiastica potestate of Aegidius Romanus*, Woodbridge 1986)
- *Theoremata de esse et essentia*, ed. H.Hocedez, Louvain 1930
- *De differentia ethicae, politicae et rhetoricae*, ed. G. Bruni, *The New Scholasticism* 6 (1932), 5-12.
- *Errores philosophorum*, ed. J. Koch, Milwaukee, Wisconsin 1944
- *Quaestio de medio demonstrationis*, ed. J.Pinborg, "Diskussionen um die Wissenschaftstheorie an der Artistenfakultät", *Die Auseinandersetzungen an der Pariser Universität im XIII. Jahrhundert*. (Miscellanea medievalia, 10), ed. A. Zimmermann, Berlin-New York 1976, 240-268.
- *Quaestio de subiecto theologiae*, ed. C. Luna, "Una nuova questione di Egidio Romano 'De subiecto theologiae'", *Freiburger Zeitschrift für Philosophie und Theologie* 37 (1990), pp.397-439.
- *Super librum I Sententiarum (reportatio)*, ed. C. Luna, "Fragments d'une reportation du commentaire de Gilles de Rome sur le premier livre des Sentences. Les extraits des mss. Clm. 8005 et Paris, B. N. Lat. 15819", *Revue des sciences philosophiques et théologiques*, 74 (1990), 205-254; 437-456.
- *Super librum III Sententiarum (reportatio)*, ed. C. Luna, "La Reportatio della lettura di Egidio Romano sul libro III delle Sentenze e il problema dell'autenticità dell'Ordinatio", *Documenti e studi sulla tradizione filosofica medievale* I (1990), 113-225, II (1991)75-146.
- *Super librum IV Sententiarum (reportatio)*, ed. C. Luna, "La lecture de Gilles de Romes sur le quatriéme livre des Sentences. Les extraits du Clm. 8005", *Recherches de Théologie ancienne et médievale* LVII (1990), 183-255.
- *De renuntiatione papae*, ed. J. R. Eastman, Lewinston-Queenston-Lampeter 1992.
- Of the planned critical edition, Aegidii Romani *Opera omnia*, Firenze 1985- have already appeared:
    - III, 1, *Apologia*, ed. R. Wielockx, Firenze 1985
- I.1/1, *Catalogo dei manoscritti, Città del Vaticano, Biblioteca Apostolica Vaticana*, a cura di B. Faes de Mottoni -- C. Luna, Firenze 1987.
- I.1/3*, *Catalogo dei Manoscritti*, *Francia (Dipartimenti*), a cura di F. Del Punta e C. Luna, Firenze 1987.
- I.1/3**, *Catalogo dei Manoscritti*, *Francia (Dipartimenti)*, a cura di C. Luna, Firenze 1988.
- I.1/2*, *Catalogo dei Manoscritti*, *Italia (Firenze, Padova, Venezia*), a cura di F. Del Punta e C. Luna, Firenze 1988.
- I.1/2** *Catalogo dei manoscritti, Italia (Assisi-Venezia)*, a cura di F. Del Punta, B. Faes de Mottoni e C.Luna, Firenze 1998.
- I.1/5*, *Catalogo dei Manoscritti, Repubblica Federale di Germania (Monaco*), a cura di B. Faes de Mottoni, Firenze 1990.
- I.1/11, *Catalogo dei manoscritti, De regimine principum (Città del Vaticano- Italia*), a cura di F. del Punta e C.Luna, Firenze 1993
- I.6 *Repertorio dei sermoni*, a cura di C.Luna, Firenze 1990.

Secondary Literature

- Briggs, Ch. F. 1999*: Giles of Rome's De regimine principum. Reading and Writing Politics at Court and University, c. 1275-c.1525*, Cambridge *et alibi*.
- Bruni, G. 1935: "Di alcune opere inedite e dubbie di Egidio Romano", *Recherches de théologie ancienne et médiévale* 7, 174-196.
- Conti, A. D. 1992: "Conoscenza e verità in Egidio Romano", *Documenti e Studi sulla tradizione filosofica medievale* III, 305-361.
- Del Punta, F. -- Donati, S. -- Luna, C. 1993: "Egidio Romano", in *Dizionario Biografico degli Italiani*, 42, Roma, 319-341.
- Donati, S. 1986: "La dottrina di Egidio Romano sulla materia dei corpi celesti. Discussioni sulla natura dei corpi celesti alla fine del tredicesimo secolo", *Medioevo* XII, 229-280.
- Donati, S. 1988: "La dottrina delle dimensioni indeterminate in Egidio Romano", *Medioevo* XIV, 149-233.
- Donati, S. 1990a: "Ancora una volta sulla nozione di quantitas materiae in Egidio Romano", *Knowledge and the Sciences in Medieval Philosophy*. Proceedings of the Eighth International Congress of Medieval Philosophy (SIEPM), II, edd. S. Knuttila- R. Työrinoja -- S. Ebbesen, Helsinki,
- Donati, S. 1990b: "Studi per una cronologia delle opere di Egidio Romano. I: Le opere prima del 1285. I commenti aristotelici ", *Documenti e Studi sulla tradizione filosofica medievale* I, 1-112.
- Eastman, J. R. 1989: *Papal Abdication in Later Medieval Thought*, Lewinston-Queenston-Lampeter.
- Eastman, J. R. 1990: "Giles of Rome and Celestine V: The Franciscan Revolution and the Theology of Abdication", *The Catholic Historical Review* 76, 195-211
- Eastman, J. R. 1992: "Giles of Rome and His Fidelity to Sources in the Context of Ecclesiological Political Thought as Exemplified in *De renuntiatione papae*", *Documenti e Studi sulla tradizione filosofica medievale* III, 145-165.
- Faes de Mottoni, B. 1983: "Mensura im Werk De mensura angelorum des Aegidius Romanus", *Mensura, Mass, Zahl, Zahlensymbolik im Mittelalter* (Miscellanea Medievalia, 16/1), ed. A. Zimmermann, Berlin-New York 1983, 86-102.
- Hewson, M. A. 1975*: Giles of Rome and the medieval Theory of Conception: a Study of the 'De formatione corporis humani in utero'*, London.
- Koch, J. 1944: Introduction to: Giles of Rome, *Errores Philosophorum*,ed. J. Koch, English translation J. O. Riedl, Milwaukee, Wisconsin.
- Lambertini, R. 1988: "A proposito della costruzione dell'Oeconomica in Egidio Romano", *Medioevo* XIV, 315-370.
- Lambertini, R. 1990: "Philosophus videtur tangere tres rationes. Egidio Romano lettore ed interprete della Politica nel terzo libro del De regimine Principum", *Documenti e studi sulla tradizione filosofica medievale*, I, 277-325.
- Lambertini, R. 1991: "Il filosofo, il principe e la virtú. Note sulla ricezione e l'uso dell'Etica Nicomachea nel De regimine principum di Egidio Romano", *Documenti e studi sulla tradizione filosofica medievale*, II (1991), 239-279.
- Lambertini, R., 1992: "Tra etica e politica: la prudentia del principe nel De regimine di Egidio Romano", *Documenti e Studi sulla tradizione filosofica medievale*, III, 77-144.

- Lambertini, R. 1995: “The Prince in the Mirror of Philosophy. About the Use of Aristotle in Giles of Rome’s De regimine principum”, *Moral and Political Philosophies in the Middle Ages*, Proceedings of the Ninth International Congress of Medieval Philosophy, Ottawa, 17-22 August 1992, edd. B. C. Bazán, E. Andújar, L. G. Sbrocchi, New York -- Ottawa -- Toronto, 1522-1534.
- Lambertini, R. 2000: “Von der iustitia generalis zur iustitia legalis. Die Politisierung des Gerechtigkeitsbegriffes im 13. Jahrhundert am Beispiel des Aegidius Romanus“, *Geistesleben im 13. Jahrhundert* (Miscellanea Mediaevalia, 27*)*, edd. J. A. Aertsen -- A. Speer, Berlin-New York, 131-145
- Lezcano, R. 1995: *Generales de la Orden de San Agustin. Biografias - Documentacion - Retratos*, Roma 1995, 30-50.
- Luna, C. 1988: “Essenza divina e relazioni trinitarie nella critica di Egidio Romano a Tommaso d’Aquino”, Medioevo XIV, 3-69.
- Luna, C. 1990: “La Reportatio della lettura di Egidio Romano sul libro III delle Sentenze e il problema dell’autenticità dell’Ordinatio”, *Documenti e studi sulla tradizione filosofica medievale* I, 113-225.
- Luna, C, 1991a: “La Reportatio della lettura di Egidio Romano sul libro III delle Sentenze e il problema dell’autenticità dell’Ordinatio”,(secona parte) *Documenti e studi sulla tradizione filosofica medievale*, II, 75-146.
- Luna, C., 1991b: “Théologie trinitaire et prédication dans les sermones de Gilles de Rome”, *Archives d’histoire doctrinale et littéraire du Moyen Age*, 58, 99-195.
- Luna, C. 1992,“Un nuovo documento del conflitto tra Bonifacio VII e Filippo il Bello: il discorso ‘De potentia domini pape’ di Egidio Romano ”, *Documenti e studi sulla tradizione filosofica medievale*, III, 167-243; 491-559.
- Macken, R. 1977, “Heinrich von Gent im Gespräch mit seinen Zeitgenossen über die menschliche Freiheit”, *Franziskanische Studien* 59, 125-182.
- Marmo, C. 1991: “*Hoc autem etsi potest tollerari*: Egidio Romano e Tommaso d’Aquino sulle passioni dell’anima”, *Documenti e studi sulla tradizione filosofica medievale*, II, 281-315
- Marmo, C. 1998: “L’utilizzazione delle traduzioni latine della Retorica nel commento di Egidio Romano (1272-1273)”, *La Rhétorique d’Aristote. Tradition et commentaires de l’antiquité au XVII^e siècle*, edd. G. Dahan- I. Rosier-Catach, Paris, 111-134.
- Martorelli Vico, R. 1988, “Il De formatione corporis humani in utero’di Egidio Romano. Indagine intorno alla metodologia scientifica”, *Medioevo* XIV, 291-313.
- Miethke, J. 2000, *De potestate papae. Die päpstliche Amtskompetenz im Widerstreit der politischen Theorie von Thomas von Aquin bis Wilhelm von Ockham*, Tübingen, 94-101.
- Pini, G. 1992: “La dottrina della creazione e la ricezione delle opere di Tommaso d’Aquino nelle Quaestiones de esse et essentia (qq. 1-7) di Egidio Romano”, *Documenti e studi sulla tradizione filosofica medievale*, III, 271-304;
- Porro, P. 1988: “ Ancora sulle polemiche tra Egidio Romano ed Enrico di Gand: due questioni sul tempo angelico”, *Medioevo* XIV, 71-105.
- Porro, P. 1991: “*Ex adiacentia temporis*: Egidio Romano e la categoria ‘quando’”, *Documenti e studi sulla tradizione filosofica medievale* II, 147-181.
- Staico, U. 1992: “Retorica e politica in Egidio Romano”, *Documenti e studi sulla tradizione filosofica medievale*, III, 1-75.

- Tabarroni, A. 1988: “Lo Pseudo Egidio (Guglielmo Arnaldi) e un’inedita continuazione del commento di Tommaso al ‘Peryermeneias’”, *Medioevo* XIV, 371-427.
- Tabarroni, A. 1991: “Figura dictionis e predicazione nel commento ai Sophistici Elenchi di Egidio Romano”, *Documenti e studi sulla tradizione filosofica medievale* II, 183-215.
- Trifogli, C. 1990a : “The Place of the Last Sphere in Late-Ancient and Medieval Commentaries”, *Knowledge and the Sciences in Medieval Philosophy*. Proceedings of the Eighth International Congress of Medieval Philosophy (SIEPM), II, edd. S. Knuttila-- R. Työrinoja -- S. Ebbesen, Helsinki, 342-350.
- Trifogli, C. 1990b: “La dottrina del tempo in Egidio Romano”, *Documenti e studi sulla tradizione filosofica medievale* I, 247-276.
- Trifogli, C. 1991: “Egidio Romano e la dottrina aristotelica dell’infinito”, *Documenti e Studi sulla tradizione filosofica medievale* II, 217-238.
- Trifogli, C. 1992: “Giles of Rome on Natural Motion in the Void”, *Medieval Studies* 54, 136-161
- Wielockx, R. 1981:“Le ms. Paris Nat. lat. 16096 et la condemnation du 7 mars 1277”, *Recherches de Théologie ancienne et médiévale* 48 (1981), 227-237.
- Wielockx, R. 1985, *Introduzione e commento* a Aegidii Romani *Opera Omnia*, III/1, *Apologia*, a c. di R. Wielockx, Firenze.
- Wippel, J. F. 1981: *The metaphysical thought of Godfrey of Fontaines. A study in late thirteenth century philosophy*, Washington

Other Internet Resources

- Entry on Egidio Colonna, Catholic Encyclopedia
- Text of *On the Errors of the Philosophers*, Giles of Rome, Medieval Studies at U. California/Davis

Related Entries

Aquinas, Saint Thomas | Henry of Ghent

Roberto Lambertini
ce29340@comune.cento.fe.it

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 20, 2001
Content last modified: December 20, 2001

Paul of Venice

Paul of Venice was the most important Italian thinker of his times, and one of the most prominent and interesting logicians of the Middle Ages. His philosophical theories (culminating in a metaphysics of essences which states the ontological and epistemological primacy of universals over any other kind of beings) are the final and highest result of the preceding realistic tradition of thought. He fully developed the new form of realism started up by Wyclif and his Oxonian followers in the last decades of the 14th century, and renewed Burley's attacks against nominalistic views. The metaphysical convictions at the basis of his philosophy are an original version of the most fundamental theses of Duns Scotus (viz. univocity of being; existence of universal forms outside the mind, which are at the same time identical with and different from their own individuals; real identity and formal distinction between essence and being; thisness as the principle of individuation; real distinction among the ten categories). But Paul puts much more stress on the ontological presuppositions and entailments of the doctrine. Simultaneously, he was open to influences from many other directions, as he held in due consideration also the positions of authors such as Albert the Great, Thomas Aquinas, and Giles of Rome, and critically discussed the doctrines of the main Nominalists of the 14th century, namely William Ockham, John Buridan, and Marsilius of Inghen, sometimes playing mutually incompatible theses against each other. This contributes to making his works stimulating and enriching from an historical point of view, but also makes it difficult to grasp his own ideas in their relationships and unity. These reflections help us to explain why for about one hundred and fifty years Paul was erroneously, but unanimously, believed to be an Ockhamist in logic and metaphysics and an Averroist in psychology and epistemology.

---

1. Life and Works

Paul of Venice (Paulus Nicolettus Venetus, Paolo Nicoletti Veneto), O.E.S.A. was born in Udine, Italy,

around 1369. He joined the Augustinian order near the age of fourteen, when he entered the convent of Santo Stefano in Venice. He studied first at Padua, but in 1390 he was assigned to Oxford, where he spent three years. He became Doctor of Arts and Theology by 1405. He taught in Padua, Siena (1420-24), and Perugia (1424-28), and lectured in Bologna (1424). At various times he held positions of leadership in his order (Pope Gregory XII designated him Prior General of the Augustinians in May 1409) and served as ambassador of the Venetian Republic. He died in Padua on 15 June 1429, while commenting the *De anima* (*On the Soul*) of Aristotle.

Paul wrote many philosophical and theological treatises (the complete list of his writings and a guide to extant manuscripts are in Perreiah 1986; for the dating of his main philosophical works see Conti 1996, pp. 9-20), including: *Logica parva* (*The Small Logic*), ca. 1393-95; *Logica magna* (*The Great Logic*), ca. 1396-99; *Sophismata aurea* (*Golden Sophisms*), ca. 1399; a commentary on Aristotle's *Posterior Analytics* (*In Post*.), A.D. 1406; *Summa philosophiae naturalis* (*Summa of Natural Philosophy -- SN*), A.D. 1408; a commentary on Aristotle's *Physics* (*In Phys*.), A.D. 1409; a commentary on Aristotle's *On the Soul* (*In De anima*), ca. 1415-20; *Quaestio de universalibus* (*On Universals -- QdU*), ca. 1420-24; a commentary on Aristotle's *Metaphysics* (*In Metaph*.), ca. 1420-24; a commentary on the *Ars Vetus*, that is, on Porphyry's *Isagoge*, Aristotle's *Categories*, and the *Liber sex principiorum* (*Expositio super Universalia Porphyrii et Artem Veterem Aristotelis -- In Porph*., *In Cat*., and *In Sex pr*. respectively), A.D. 1428.

2. Logic

The main contributions of Paul of Venice to the history of logic in the Middle Ages concern the notion of formal distinction and the analysis of predication.

2.1 Identities and Distinctions

Paul's formulation of the theory of identity and distinction is a further development of Duns Scotus' and Wyclif's doctrines on the subject. The Italian master recognizes two main types of identity: material (*secundum materiam*) and formal (*secundum formam*). There is material identity when the material cause is the same, either in number (it is a case of the same thing called in different ways) or by species (it is a case of two objects made of the same kind of stuff). There is formal identity when the formal cause is the same. This happens in two ways: if the form at issue is the singular form of the individual composite, then there is a unique object known in different ways; if the form at issue is the common essence instantiated by the singular form, then there are two distinct objects belonging to the same species or genus (*In Metaph*., book V, tr. 2, chap. 3, fol. 185ra). Correspondingly, the main types of distinction (or difference) are also two: material and formal. There is material distinction when the material cause is different, so that the objects at issue are separable entities. In general, there is formal distinction when the formal cause is different. This happens in two ways: if the material cause is also different, then it is a particular case of material distinction. If the material cause is the same, then a further analysis is necessary. If the material cause is the same by species only, then it is an improper case of formal distinction; but if the

material cause is the same in number, then there is properly formal distinction, since the forms at issue have different definite descriptions but share the same substrate of existence, so that they are one and the same thing in reality. For example, there is a proper formal distinction in the case of the two properties of being-capable-of-laughing (*risibile*) and of being-capable-of-learning (*disciplinabile*), which are connected forms instantiated by the same set of individual substances (*In Metaph*., book V, tr. 2, chap. 3, fol. 185rb).

Material distinction is a necessary and sufficient criterion for real difference, traditionally conceived, whereas there is formal distinction if and only if there is one substance in number (i.e. material identity in the strict sense) and a multiplicity of formal principles with different descriptions instantiated by it. Paul therefore inverts the terms of the question in relation to what earlier approaches had done. By means of the formal distinction Duns Scotus and John Wyclif had tried to explain how it is possible to distinguish many different real aspects internal to the same individual substance (the passage is from one to many). On the contrary, Paul is attempting to reduce multiplicity to unity (the passage is from many to one). What Paul wants to account for is the way in which many different entities of a certain kind (i.e. of an incomplete and dependent mode of existence) can constitute one and the same substance in number.

2.2 Predication

The starting point of Paul's theory of predication is his doctrine of universals. Just like Wyclif and his followers (Alyngton, Penbygull, Sharpe, Milverley, Whelpdale, Tarteys), the Augustinian master claims that

1. There are real universals, which are common essences naturally apt to be present in and predicated of many similar individuals.
2. Real universals and their individuals are really the same and only formally distinct.
3. Predication is first of all a real relation between metaphysical entities (*QdU*, fols. 124ra, 124vb, 127va, 132va).

But his analysis of predication is different from those of both Wyclif and his followers. In fact, Paul divides predication into identical predication and formal predication and defines them in a different way than his sources do.

To speak of *identical predication* it is sufficient that the form signified by the subject-term of a (true) proposition and the form signified by the predicate-term share at least one of their substrates of existence. This is the case for propositions like ‘Man is (an) animal’ and ‘The universal-man is something white’ (‘*Homo in communi est album*’). One speaks of *formal predication* in two cases:

1. When for the truth of the proposition it is necessary that the form signified by the predicate-term is present in *all* the substrates of existence of the form signified by the subject-term, in virtue of a formal principle (made clear in the proposition itself) that is in turn directly present in all the substrates of existence of the form signified by the subject-term. This is the case for propositions

like ‘Man is formally (an) animal’ and ‘Socrates *qua* man is an animal’.
2. Or else when the predicate of the proposition is a term of second intention, like ‘species’ or ‘genus’. This is the case for propositions like ‘Man is a species’ and ‘Animal is a genus’ (*SN*, part VI, chap. 2, fol. 93vab; *QdU*, fols. 124vb-125rb).

As is evident, identical predication is extensionally defined, whereas formal predication is intensionally defined, since formal predication entails a relation modally determined between the subject-thing and the predicate-thing. In fact, formal predication presupposes that there is a *necessary* connection between the subject-thing and the predicate-thing of the given proposition. For this reason, Paul denies that sentences like ‘(What is) singular is (what is) universal’ (‘*Singulare est universale*’), which Wyclif and his followers acknowledged as true, are in fact true propositions. For Wyclif and his followers, the sentence at issue is an example of predication by essence. But for Paul of Venice, it is an example of formal predication; no individual *qua* individual is an universal, or *vice versa*, as no second intention intensionally considered is any other second intention (*QdU*, fol. 133va; *In Porph*., *prooem*., fol. 3ra-b). As a consequence, Paul rewrites the preceding sentece in this form: ‘(What is) singular is *this* universal’ (‘*Singulare est hoc universale*’), where the presence of the demonstrative ‘this’ changes the kind of predication from formal to identical. So corrected, the sentence is true, since it signifies that a certain entity, in itself singular, is the substrate of existence of a universal essence (*QdU*, fol. 133va-b).

As a result, Paul builds up a mixed system, where the copula of the standard philosophical sentences he deals with can have a threefold value: it means a partial identity between the subject-thing and the predicate-thing in the case of identical predication; it means a necessary link between forms in the case of the first type of formal predication; it means that the subject-thing in virtue of itself is necessarily a member of a given class of objects, which the predicate-term of the proposition labels and refers to, in the case of the second type of formal predication -- that is, when the predicate is a term of second intention.

3. Ontology

Paul's world consists of finite beings (that is, things like men or horses) really existing outside the mind, each made up of a primary substance and a host of forms existing in it and by it. The forms of a primary substance belong to ten different types of being, or categories. Therefore a finite being cannot be totally identified with the primary substance. (In fact no primary substance contains the whole being of a finite being.) Rather it is an ordered congeries of categorial items. Primary substances are not simple items but complex objects, since they are compounded of particular matter and form -- a form that is really identical with and formally distinct from the specific nature itself that the primary substance instantiates (*SN*, part VI, chap. 1, fols. 92vb-93ra). The concepts of matter and form are relative, since their meanings are connected with each other (*In Post*., fol. 40rb). Being the form of something and being the matter of something are converse relations of three different kinds, whose arguments and values are:

1. the metaphysical constituents of the individual substance (i.e. singular matter and form);
2. the metaphysical constituents of the specific natures (i.e. genus and difference); and
3. the categorial items (i.e. individual and universal substances and accidents) considered according

to their various degrees of generality.

The specific nature (or essence) can be conceived from a twofold point of view: intensionally (*in abstracto*) and extensionally (*in concreto*). Intensionally viewed, the specific nature simply expresses the set of essential properties that compose a categorial form, without any reference to the existence of individuals which, if there are any, instantiate it. Extensionally viewed, the specific nature is that same form conceived of as instantiated by at least one singular entity. For instance, human nature intensionally considered is humanity (*humanitas*); extensionally it is man (*homo*) (*In Porph*., *prooem*., fol. 9va). Both of them are substantial forms superordinated to the whole human compound, but while humanity is properly a form, i.e. something existentially incomplete and dependent, man is an existentially autonomous and independent entity. Thus they differ from each other in the same way as a predicate (for example ‘*P*’) differs from a formula (for example, ‘*P*(*x*)’).

Because of the complexity of the metaphysical composition of the finite corporeal being, every creature has four different levels of being: real, essential, temporal, and individual. The real being is nothing but the whole reality of the finite being. The essential being is the mode of being proper to the specific nature that a certain singular directly instantiates. The temporal being is the state of affairs designated by infinitival expressions like ‘being a man’ (‘*hominem esse*’) or ‘being white’ (‘*esse album*’) -- that is, the object of the act of judging. Finally, the individual being is the actual existence of the primary substance of a finite being as distinct from the whole reality of the finite being itself (*SN*, part VI, chap. 1, fol. 92vb).

According to Paul, who follows Duns Scotus and Wyclif on this subject, being is univocally shared by everything real, since it is the stuff that the ten categories modulate according to their own essence (*In Phys*., book I, tr. 1, chap. 2, t. c. 13; *In Metaph*., book IV, tr. 1, chap. 1, fols. 122ra-125vb, *passim*; *In Porph*., chap. *De specie*, fol. 22rb). In view of this position, Paul maintains no real distinction between essence and being (*In Metaph*., book IV, tr. 1, chap. 2, fol. 127rb; book VI, chap. 1, fol. 223vb). Like Duns Scotus and Wyclif, Paul speaks of a formal difference (or difference of reason) between essence and being in creatures, as the essence and the essential being of a thing are one and the same entity considered from two distinct point of view, intensionally (the essence) and extensionally (the being) (*SN*, part VI, chap. 1, fol. 93ra).

This analysis identifies the opposition between essence and being with the opposition between universals and individuals. Like Wyclif, Paul thinks of the essence as a universal form intensionally considered, and the existence (taken in the strict sense) as the mode of being proper to primary substances. Thus, when Paul affirms that essence and being are really identical and formally different, he simply restates the thesis of the real identity and formal distinction between universals and individuals that was typical of the Realists of the late Middle Ages. Consequently, like Burley and Wyclif, Paul holds that a formal universal actually (*in actu*) exists outside our minds only if there is at least one individual that instantiates it, so that without individuals common natures (or essences) are not really universals (*SN*, part VI, chap. 2, fol. 94ra).

This means that the relationship between common natures and singulars is ultimately grounded on

individuation, since no actual universality and no instantiation is possible without individuation. On this subject Paul successfully reconciles the Scotistic approach with certain Thomistic theses. Paul claims that the principle of individuation is twofold, immanent and remote. The immanent principle is the one whose presence necessarily entails the existence of the individual it constitutes, and whose absence necessarily entails the non-existence (or disappearance) of the individual. The remote principle, on the other hand, is just what the immanent principle presupposes, but whose presence and absence alone are insufficient for causing the existence or disappearance of the individual, as it continues being after the corruption of the individual. Thisness (*haecceitas*) is the immanent principle of individuation, whereas form, matter, and quantity are the remote principle. Thisness in turn has a twofold origin, as it derives from matter and form together in the case of corporeal substances, and as it derives from the essence (*quidditas*) alone in the case of angelic intelligences (*SN*, part VI, chap. 5, fol. 95vb). Furthermore, according to Paul there is a close similarity between the thisness, which he now calls individual difference (*differentia individualis*), and the specific difference. The specific difference is what differentiates the species from the genus, since it is the determination or property which, once added to the genus, results in the species. On the other hand, the specific difference is really identical with the genus, from which it is distinct only in virtue of a formal principle. The same happens to the individual difference: it is what differentiates the individual from the species; from the ontological point of view, it is really identical with and formally distinct from the species itself; and it is the formal principle in virtue of which the individual is what it is, something particular, concrete, and perfectly determined in itself (*SN*, part VI, chap. 5, fol. 96rb; chap. 26, fol. 112rb-va; *QdU*, fol. 128va; fol. 129rb; *In Metaph*., book III, tr. 1, chap. 1, fol. 83vb).

As far as the problem of angelic individuation is concerned, the logical consequence deriving from such premisses is that it is impossible to find two angels who share the same specific nature and are numerically distinct, since only one *haecceitas* can spring up from an incorporeal species (*SN*, part VI, chap. 5, fol. 96ra). This solution is close to the inner meaning of Duns Scotus' position and contrasts with Aquinas' view, although Paul affirms that the angelic intelligences are specifically, and not numerically, different. In fact, according to St. Thomas, angels are specifically different because they are incorporeal, and without matter no individuation is possible. On the contrary, Paul of Venice thinks angels are individuated by means of thisnesses, but not multiplied by them, because of the absence of matter, so that there is only one angel per species. Since specific natures of incorporeal beings do not include any reference to matter, only a unique principle of individuation (*ratio suppositalis*) can flow from such species. As a consequence, no angel is one in number in the strict sense of the term (as being one in number necessarily implies the actual presence of a multiplicity of things of the same species), even though broadly speaking every angel is one in number, as two (or more) angels are, after all, "many things" -- but never many angels of the same kind (*SN*, part VI, chap. 5, fol. 95vb).

In his last work, the commentary on the *Ars Vetus*, Paul summarizes his position as follows:

1. The individual is the final result of a process of individuation whose starting point is a specific form.
2. The individuation is what differentiates the individual from its species.
3. The individuation is nothing but the thisness itself.
4. The thisness and the specific form are only formally distinct from the individual they make up (*In*

Porph., chap. *De specie*, fol. 60ra).

5. The principle of individuation, when causing the passage from the level of universals to that of singulars, does not play the role of form (or act) in relation to the specific nature, but the role of matter (or potency), as it is what the specific form structures (*In Cat*., chap. *De substantia*, fol. 60ra).

In this way Paul of Venice tried to solve the aporetic aspects of Duns Scotus' theory of individuation. Scotus said nothing about the problem of the relation between the thisness and the particular matter and form that constitute the individual. The Franciscan master was silent also about a possible identification of the thisness with one of the two essential forms of the individual substance, the *forma partis* (for instance, the individual soul) and the *forma totius* (the human nature). Paul identifies the principle of individuation with the informing act through which the specific nature molds its matter. This identification had been already suggested by the opposition between immanent and remote principles of individuation described in the *Summa philosphiae naturalis*. In fact, all the constituents of the individual compound (matter, form, and quantity) had been contrasted with the thisness, which for that reason could not be identified with any of them. Moreover, it is obvious that:

1. It is the union of the singular form with its matter that establishes the "birth" of the individual.
2. It is its separation from the matter that establishes the "death" of the individual.
3. The union of the singular form with its matter is the necessary and sufficient condition for the passage of the specific essence from its abstract (or intensional) mode of being to its concrete (or extensional) mode of being.

4. Psychology

Paul of Venice rejects the Augustinian conception of the relation of soul to body and follows the Aristotelian view of the soul as form of the body. But, against Aristotle and following Aquinas, Paul claims that, although it is the form of the body, the human soul is a self-subsistent form, and therefore incorruptible. However, unlike St. Thomas, he claims that the human soul is twofold, since the complete human soul derives from the close union of two distinct principles, the cogitative and the intellective ones. The former is the cause of the animality and the latter of the rationality of man; neither of them can exist in man without the other, and the cogitative soul is in potency in relation to the intellective soul (*SN*, part V, chap. 5, fol. 69ra; *In De anima*, book II, t. c. 23, fol. 48ra).

Like St. Thomas and Giles of Rome, Paul maintains that there is a real distinction between the soul and its faculties. But, in opposition to them, he holds that there is only a formal distinction (*ratione et definitione*) between the faculties themselves (*SN*, part VI, chap. 4, fol. 68ra-b). Whereas the faculties of the cogitative soul depend on bodily organs for their operations, the faculties of the intellective soul, i.e. the active intellect, the passive intellect, and the will, are independent of bodily organs, even though in the state of union with the body they need sensation for exercising their powers and no act of sensation can be produced without the concurrence of the body (*SN*, part V, chap. 10, fols. 71va-72ra). Besides the vegetative faculty (which regulates nutrition, growth, and reproduction) and the power of locomotion, the

faculties of the cogitative soul are the following: the five exterior senses, the general sense (*sensus communis*), the fantasy (*phantasia*), the power of assessment (*vis aestimativa*), and memory. Against Avicenna, Paul explicitly denies that there is a fifth internal sense, the imagination, since he thinks its presumed operations are the same as those of the fantasy (*SN*, part V, chap. 30, fol. 84ra). The general sense distinguishes and collates the data of the special exterior senses. The fantasy conserves sensible species apprehended by senses and freely combines them together to produce figments. The power of assessment recognizes those properties of things which cannot be perceived through the senses, like, for example, that something is useful for a certain purpose, or friendly, or unfriendly. The memory is the ‘warehouse’ where all the sensible species are stored, so that the cogitative soul can perform its tasks even without the presence of any sensible object (*SN*, part V, chap. 30, fol. 84ra-va).

According to Nardi 1958, Ruello 1980, and Kuksewicz 1983, Paul was an Averroist in Psychology, as he would have supported the thesis of the unicity and separate character of the passive intellect for the whole human species. But this is false. On the contrary, Paul's point of view is close to that of St. Thomas for the question of the passive intellect, and to the position of Avicenna for the question of the active intellect (Conti 1992, especially pp. 338-47). If his affirmations in the *Summa philosophiae naturalis* are ambiguous and it is therefore possible to miss their deepest meaning, in his commentaries on the *De anima* and on the *Metaphysics* he clearly rejects all the main theses of the Averroism. First of all, he maintains personal immortality (a thesis denied by genuine Averroists) and, like the medieval followers of Avicenna, identifies active intellect with God's activity of ‘illumination’ in the soul (*In De anima*, book III, t. c. 11, fol. 137rb; t. c. 19, fol. 143ra). Secondly, he claims, against Averroes, that the intellective soul is form and act of the body (*In De anima*, book II, t. c. 7, fol. 39rb-va; t. c. 8, fol. 134rb). Moreover, he asserts that:

1. The intentional species present in the exterior senses, in the internal senses, and in the intellect are of three different kinds.
2. The individual is a proper object of intellection for us.
3. The same intelligible species (*species intelligibilis*) by means of which we grasp substantial essences is the medium in virtue of which we can understand the peculiar structure of the individual which instantiates that essence. (*SN*, part V, chap. 28, fol. 83ra; *In De anima*, book III, t. c. 11, fol. 136vb; 137rb-va).

These theses are just the opposite of Averroist convictions. Finally, he explicitly argues against the unicity of the passive intellect, utilizing some arguments drawn from the *De unitate intellectus contra Averroistas* (*On the Unicity of Intellect against the Averroists*) and the *Summa theologiae* (*Summa of Theology*) of St. Thomas. Among them, the most important are the following three:

1. If the soul is the form of the body, as Aristotle states, it is impossible that the passive intellect is one in all men, since one and the same principle in number cannot be the form of a multiplicity of substances.
2. If the passive intellect is one and the same for all men, then after death nothing remains of men but this unique intellect, and in this way the bestowal of rewards and punishments is done away with.
3. One and the same intellect could hold contradictory opinions at once, in apparent violation of the

law of contradiction (*In De anima*, book III, t. c. 27, fol. 149ra; *In Metaph*., book IV, tr. 1, chap. 3, fols. 136vb-137ra).

More generally, he thinks (i) the Averroistic theses are lacking in a solid philosophical basis, since they can be maintained from the physical point of view only, according to which everything is considered *qua* affected by or connected with motion, but (ii) they are totally false from the metaphysical point of view, which is the most comprehensive of all. From this viewpoint, according to which the passive intellect has to be considered a substantial form, it is evident that it has a beginning in time, but certainly not an end, and that, like any other material substantial form, it is multiplied according to the multiplication of bodies (*In Metaph*., book XII, tr. 1, chap. 3, fol. 427ra-b).

Bibliography

Primary literature

- *Logica parva,* (Venice, 1544).
- *Logica magna,* (Venice: Albertinus Vercellensis for Octavianus Scotus, 1499).
- *Logica magna: Tractatus de suppositionibus,* A. R. Perreiah, ed. & trans., (St Bonaventure, N. Y.: The Franciscan Institute, 1971).
- *Logica magna: Part I, Fascicule 1: Tractatus de terminis,* N. Kretzmann, ed. & trans., (Oxford: Oxford University Press, 1979).
- *Logica magna: Part I, Fascicule 8: Tractatus de necessitate et contingentia futurorum*, C. J. F. Williams, ed. & trans., (Oxford: Oxford University Press, 1991).
- *Logica magna: Part II, Fascicule 3: Tractatus de hypotheticis,* A. Broadie, ed. & trans., (Oxford: Oxford University Press, 1990).
- *Logica magna: Part II, Fascicule 4: Capitula de conditionali et de rationali,* G. E. Hughes, ed. & trans., (Oxford: Oxford University Press, 1990).
- *Logica magna: Part II, Fascicule 6: Tractatus de veritate et falsistate propositionis et tractatus de significato propositionis*, F. del Punta ed., M. McCord Adams, trans., (Oxford: Oxford University Press, 1978).
- *Logica magna: Part II, Fascicule 8: Tractatus de obligationibus*, E. J. Ashworth, ed. & trans., (Oxford: Oxford University Press, 1988).
- *Sophismata aurea,* (Venice: Bonetus Locatellus for Octavianus Scotus, 1493).
- *Super I Sententiarum Johannis de Ripa lecturae abbreviatio, prologus,* F. Ruello ed., (Florence: Leo S. Olschki, 1980).
- *Expositio in libros Posteriorum Aristotelis,* (Venice, 1477). Photoreprint, Hildesheim: Olms, 1976.
- *Summa Philosophiae Naturalis,* (Venice, 1503).
- *Expositio super octo libros Physicorum necnon super commento Averrois,* (Venice, 1499).
- *Expositio super libros De generatione et corruptione,* (Venice: Bonetus Locatellus for Octavianus Scotus, 1498).
- *Scriptum super libros De anima,* (Venice: 1504).
- *Quaestio de universalibus,* extant in nine mss. There is a partial transcription from ms. Paris, BN

6433B in A. D. Conti, ed., *Johannes Sharpe: Quaestio super universalia,* (FirenzeL Leo S. Olschki, 1990), Appendix V, pp. 199-207. The ms. used here for the quotations is: Paris, BN 6433B.

- *Lectura super libros Metaphysicorum,* extant in two mss. The ms. used here for the quotations is: Pavia, Biblioteca Universitaria, fondo Aldini 324.
- *Expositio super Universalia Porphyrii et Artem Veterem Aristotelis,* (Venice, 1494).

Secondary literature

- Ashworth, E. J., "A Note on Paul of Venice and the Oxford *Logica* 1483," *Medioevo* 4 (1978), pp. 93-99.
- Bochenski, I. M., *A History of Formal Logic*, I. Thomas trans., (Notre Dame, Ind.: University of Notre Dame Press, 1961), pp. 161 ff.
- Bottin, F., "Proposizioni condizionali, *consequentiae* e paradossi dell'implicazione in Paolo Veneto," *Medioevo* 2 (1976), pp. 289-330.
- Bottin, F., *La scienza degli occamisti: La scienza tardo medievale dalle origini del paradigma nominalista alla rivoluzione scientifica*, (Rimini: Maggioli, 1982), pp. 72, 99-101, 192, 274-76, 288-89.
- Bottin, F., "Paolo Veneto e il problema degli universali," in L. Olivieri ed., *Aristotelismo veneto e scienza moderna,* (Padua: Antenore, 1983), pp. 459-68.
- Bottin, F., "Logica e filosofia naturale nelle opere di Paolo Veneto", in *Scienza e filosofia all'Università di Padova nel Quattrocento,* (Padua: Antenore, 1984), pp. 85-124.
- Conti, A. D., "Alcune note sulla *Expositio super Universalia Porphyrii et Artem Veterem Aristotelis* di Paolo Veneto: Analogie e differenze con i corrispondenti commenti di Walter Burley," in A. Maierù ed., *English Logic in Italy in the 14th and 15th Centuries,* (Naples: Bibliopolis, 1982), pp. 293-303.
- Conti, A. D., "Universali e analisi della predicazione in Paolo Veneto", *Teoria* 2.2 (1982), pp. 121-39.
- Conti, A. D., "Il problema della conoscibilità del singolare nella gnoseologia di Paolo Veneto," *Bullettino dell'Istituto Storico Italiano per il Medio Evo e Archivio muratoriano* 98 (1992), pp. 323-82.
- Conti, A. D., "Il sofisma di Paolo Veneto: *Sortes in quantum homo est animal,*" in S. Read ed., *Sophisms in Medieval Logic and Grammar,* (Dordrecht: Kluwer, 1993), pp. 304-18.
- Conti, A. D., *Esistenza e verità: forme e strutture del reale in Paolo Veneto e nel pensiero filosofico del tardo Medioevo,* (Rome: Edizioni dell'Istituto Storico Italiano per il Medio Evo, 1996).
- Garin, E., *Storia della filosofia italiana,* 3 vols., (Torino: Einaudi, 1966), vol. 1, pp. 429-45.
- Karger, E., "La supposition materielle comme suppositions significative: Paul de Venice, Paul de Pergula," in A. Maierù ed., *English Logic in Italy in the 14th and 15th Centuries,* (Naples: Bibliopolis, 1982), pp. 331-41.
- Kretzmann, N., "Medieval logicians on the Meaning of the Proposition", *The Journal of Philosophy* 67 (1970), pp. 767-87.
- Kuksewicz, Z., "Paolo Veneto e la sua teoria dell'anima", in L. Olivieri ed., *Aristotelismo veneto e*

scienza moderna, (Padua: Antenore, 1983), pp. 130-64.

- Mugnai, M., "La *expositio reduplicativarum* chez Walter Burleigh et Paulus Venetus," in A. Maierù ed., *English Logic in Italy in the 14th and 15th Centuries,* (Naples: Bibliopolis, 1982), pp. 305-20.
- Nardi, B., "Paolo Veneto e l'averroismo padovano", in *Saggi sull'averroismo padovano dal secolo XIV al XVI,* (Florence: Sansoni, 1958), pp. 75-93.
- Nuchelmans, G., *Theories of the Proposition: Ancient and Medieval Conceptions of the Bearers of Truth and Falsity,* (Amsterdam: North-Holland, 1973), pp. 266-71.
- Nuchelmans, G., "Medieval Problems concerning Substitutivity (Paul of Venice, *Logica Magna* II, 11, 7-8)," in V. M. Abrusci, E. Casari, M. Mugnai ed., *Atti del Congresso Internazionale di Storia della Logica: San Gimignano, 4-8 dicembre 1982,* (Bologna: CLUEB, 1983), pp. 69-80.
- Pagallo, G. F., "Nota sulla *Logica* di Paolo Veneto: la critica alla dottrina del *complexe significabile* di Gregorio da Rimini", in *Atti del XII Congresso Internazionale di Filosofia,* (Florence: Sansoni, 1960), vol. 9, pp. 183-91.
- Perreiah, A. R., *Paul of Venice: A Bibliographical Guide*, (Bowling Green, Ohio: Philosophy Documentation Center, 1986).
- Prantl, K., *Geschichte der Logik im Abendlande,* 4 vols., (Leipzig: S. Hirzel, 1855-70), vol. 4, pp. 118-40. Photoreprint, Graz: Akademische Druck- und Verlaganstalt, 1955.
- Ruello, F., "Paul de Venise thélogien ‘averroiste’?," in J. Jolivet ed., *Multiple Averroès,* (Paris: J. Vrin, 1978), pp. 257-72.
- Ruello, F., "Introduction," in F. Ruello, ed., *Super I Sententiarum Johannis de Ripa lecturae abbreviatio, prologus,* (Florence: Leo S. Olschki, 1980), pp. 9-69.
- Wallace, W. A., *Causality and Scientific Explanation*, 2 vols., (Ann Arbor: University of Michigan Press, 1972), pp. 121-27.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Albert the Great [= Albertus magnus] | Aquinas, Saint Thomas | Buridan, John [Jean] | divine illumination | Duns Scotus, John | Giles of Rome | Marsilius of Inghen | Ockham [Occam], William | Wyclif, John

Alessandro D. Conti
a.conti@tiscalinet.it

Table of Contents

First published: August 22, 2001
Content last modified: August 22, 2001

Marsilius of Inghen

Marsilius of Inghen, master at the Universities of Paris (1362-1378) and Heidelberg (1386-1396), wrote a number of treatises on logic and natural philosophy popular at many late medieval and early modern universities. He adopted the logico-semantic approach of William of Ockham and John Buridan while at the same time defending the traditional views of Thomas Aquinas and Bonaventure. His thinking sheds light on the discussion between nominalists and realists and allows insight into the changing interests of philosophy and theology, from the critical attitude of many fourteenth-century authors to the search for tradition which was characteristic of the fifteenth century.

---

1. Life and Works

Marsilius of Inghen was born around 1340 in Nijmegen, a city in the eastern part of the Low Countries (Netherlands). In the older literature it is often said that he came from one of the villages in the vicinity of Nijmegen (Inghen), but this view is mistaken. It was based on a confused reading of the *Oratio Funebris* held in 1396 by Nicholas Prowin at the funeral of Marsilius and published in 1499 at Mainz. From 1362 on, Marsilius was master at the Faculty of Arts at the University of Paris, where he was also rector (1367 and 1371), and a student of theology. As a teacher at Paris, Marsilius was much esteemed and his lectures drew large audiences. Among his students were many compatriots, some of whom came from Nijmegen and surrounding villages. In 1378, Marsilius found himself the University's delegate at the court of Pope Urban VI in Tivoli. In 1379, he instructed one of his colleagues, Hugh of Hervort, to look after his

interests in Paris. After 1379, Marsilius's name is no longer mentioned in the acts of the University of Paris. He probably turned away from Paris because of the imbroglio surrounding the Great Schism of 1378. Meanwhile, he kept in touch with his native city. In 1382 the town council of Nijmegen treated him to a banquet. From 1386 on, Marsilius was master at the University of Heidelberg. There, as in Paris, he held a number of administrative offices. He was one of the founders of the University of Heidelberg, of which he was rector no fewer than nine times, from 1386-1392 and also in 1396. In 1389-1390, as the University's *nuncio* together with Conrad of Soltau, he was responsible for transferring the University register to Rome (Boniface IX). At the beginning of the 1390s, Marsilius again took up the study of theology. The masters who taught theology were by then Conrad of Soltau (since 1387) and Matthew of Krakow (since 1394), both from the University of Prague. In 1395/1396 Marsilius finished his lectures on the *Sentences* and so became the first theologian to obtain the doctorate at Heidelberg. He died a short time later, on August 20, 1396.

Marsilius was a prolific writer. His work was the fruit of his teachings in Paris and Heidelberg. Many of his writings have been preserved in manuscripts or early printed editions, although recently some have appeared in modern critical editions. His most important writings include:

Works on Logic and Epistemology

1. *Exposition of the Old Logic*
2. Various *Questions* on the Old and New Logic
3. *Summary* [Abbreviationes] *of the Old and New Logic*
4. Treatises on the Properties of Terms: *On Supposition, Ampliation, Appellation, Restriction, Obligation, Insolubles, and Consequences*.

Works on Natural Philosophy and Metaphysics

1. *Summary* [Abbreviationes] *of Aristotle's 'Physics'*
2. *Questions on Aristotle's 'On Generation and Corruption'*
3. *Questions on Aristotle's 'De anima'*
4. *Questions on Aristotle's 'Metaphysics'*

Works on Theology

- *Questions on the 'Sentences' of Peter Lombard*

2. Teachings

2.1 Logic and Epistemology

In his logic and epistemology, Marsilius followed the nominalist tradition of the fourteenth century as

exemplified by William of Ockham and John Buridan. Yet Marsilius never qualified himself as a nominalist or follower of Ockham. He was an independent thinker who sometimes went back to the older tradition of the thirteenth century (e.g., in Peter of Spain), or advocated theories which were more in line with ordinary speech, as against the highly specialized views of his contemporaries.

Marsilius's nominalism comes to the fore in his views on the object of scientific knowledge, the nature of universals, and the logical doctrine of supposition. His basic assumption is that there are only individuals and no universals outside the human mind.

2.1.1 The Object of Scientific Knowledge

According to the Aristotelian standard accepted by Marsilius, the object of scientific knowledge must be universal and necessarily true. This is not the case with individual things in the external world, since they are subject to change. Only the conclusion of a true and necessary syllogism can meet the standard. Hence, for Marsilius, the object of scientific knowledge is not anything outside the mind, but the mental proposition which refers to individual things and their qualities. More specifically, the proper object of scientific knowledge is a proposition in the form of a conclusion that has been deduced from necessary premises.

2.1.2 Universals

Marsilius argued that universal concepts such as 'humanity' do not refer to real universals outside the human mind. Accordingly, there is no universal essence in singular individuals. Individuals of one genus or species do resemble each other, however, and this resemblance is the foundation of universal concepts in the human mind. The generation of universal concepts is a natural process, which Marsilius described as follows: suppose individual A of species S evokes concept X in the human mind. This concept is similar to concept Y which has been evoked by B of the same species S. By abstracting from all the differences between X and Y, the human mind is able to produce another concept, Z, which stands for both A and B. Universality is then taken to be a quality of concept Z, the product of the epistemological operation of abstraction on concepts X and Y by the human mind.

2.1.3 Supposition

In line with his account of the nature of universals, Marsilius rejected simple supposition. Logicians such as Peter of Spain had used the notion to indicate that a term stood not for an individual but for a universal or common nature in the external world, like the term 'man' in the proposition, 'Man is a species'. Since Marsilius rejected the idea of universals existing outside the mind, he eliminated simple supposition from the list of different types of supposition. He was critical of some of his contemporaries (e.g., Albert of Saxony) who likewise dismissed the concept of real universals, yet kept on using the notion of simple supposition. They had changed the meaning of the term, he said, by claiming that a written or spoken term had simple supposition if it was used to refer to a concept in the human mind. Marsilius wondered whether young students would be able to understand this new meaning of simple supposition, since they

would hardly know what concepts are. To avoid confusion, Marsilius decided not to deal with simple supposition at all in his logic.

2.1.4 Some Specific Views

Marsilius was his own person when it came to assessing the views of others. In his analysis of the proposition ‘Socrates is not a chimera’ he followed what he called ‘the Parisian method’, according to which the proposition is false because the term ‘chimera’ supposits for nothing, there being no real chimeras to which it can refer. Others, however, considered the proposition to be true.

Elsewhere, he departed from the opinions of the Parisian School (*scola Parisiensis*) and opted for the perspective of ordinary language or common way of speaking (*communis modus loquendi*). This was the case with his analysis of the proposition ‘The Antichrist is not, but he will be’. According to the Parisian School, the term ‘he’ refers to the thing referred to by the term ‘Antichrist’. Since there is no Antichrist, neither term has reference. But in ordinary language it is different, for there the term ‘he’ is meant to refer to the future Antichrist. Marsilius accepted the latter analysis as sound, despite the authority of the former.

Finally, in the definition of ampliation, Marsilius went back to logicians of the thirteenth century such as Peter of Spain, who had defined ampliation as an extension of supposition, whereas fourteenth-century logicians such as Albert of Saxony did not consider ampliation to be a kind of supposition. Marsilius reinterpreted their definition so that it fit better with the older tradition. Such efforts to harmonize older and newer positions were typical of the late fourteenth century.

2.2 Natural Philosophy and Metaphysics

In natural philosophy and metaphysics Marsilius was an empiricist, meaning that he thought all scientific knowledge must be based on either sense data or self-evident propositions, i.e., propositions in which the meaning of the predicate is included in the subject. Everyone who knows the meaning of the terms of such propositions judges them to be evidently true. This has far-reaching consequences for the relationship between philosophy and theology. Since the philosopher uses only sense data and self-evident propositions, his inquiry may come to different conclusion than that of the theologian, who has additional knowledge from scripture. The philosopher makes judgments about the world from a limited human perspective, whereas the theologian is helped by divine revelation. Yet Marsilius took the task of the philosopher seriously because he thought the human mind has a natural tendency to search for truth, which is satisfied (although not ultimately satisfied) in natural philosophy and metaphysics.

2.2.1 Creation

According to the principles of natural philosophy, creation from nothing is impossible. The senses show that things always come from other things. Because there is no serious reason to doubt the information given by the senses, the human mind legitimately jumps to the universal principle that nothing can come

from nothing, driven by the natural tendency to search for truth. Consequently, for the human mind creation from nothing is impossible. It contradicts the universal principle that nothing comes from nothing. That God has created the world from nothing is therefore only a matter of faith (*sola fide est creditum*). Revelation shows that human knowledge of creation is limited, but it cannot be aided by natural philosophy at this point.

2.2.2 Theory of the Human Soul

In the later Middle Ages the study of the soul was part of natural philosophy. Marsilius treated the human soul in his commentary on Aristotle's *De anima*, in which he followed the Parisian tradition of Buridan and Oresme concerning the particular questions addressed. Following Buridan, he argued that there is no natural proof of the immortality of the human soul. For the human natural mind, unaided by revelation, the theory of Alexander of Aphrodisias that the human soul is corruptible is the most probable. That Alexander of Aphrodisias is mistaken and that the soul continues to exist after the death of the body is known through revelation alone. Faith has more authority than human reason and must be accepted in all cases where the two conflict since the things we believe on faith come from God, who cannot err.

2.3 Metaphysics

Although metaphysics cannot surpass the limits of human knowledge, Marsilius considered it to be the entry point to theology. Natural reason is capable of forming some adequate and true concepts of God, but also of forming true propositions about God. It is able to prove that God exists and possesses knowledge and will. But it cannot demonstrate that God has free will or infinite power. This, Marsilius claimed, was also true for philosophers such as Aristotle, whose teachings equal those of natural reason itself.

From Buridan, Marsilius took the idea that God according to Aristotle and Averroes is not only the final cause of the heavens and separate substances, but also their efficient cause. On this point Buridan and Marsilius were following the view of Scotus and Ockham against that of John of Jandun, Johannes Baconis, and Gregory of Rimini. It is worth noting in this connection that in the *Puncta super libros Metaphysicae* (i.e., brief abstracts of Aristotle's *Metaphysics* for teaching purposes) attributed to Johannes de Slupcza and written in Krakow in 1433, some of the views that Marsilius adopted from Buridan, including the one just mentioned, are attributed to Marsilius instead of Buridan -- notwithstanding the fact that the author was familiar with both Marsilius's and Buridan's commentaries. This illustrates the strong influence Marsilius's work exerted on fifteenth-century students and commentators.

2.4 Theology

Marsilius expressed his theological views in a voluminous commentary on the *Sentences*. He quoted and often adopted views that were put forward by fourteenth-century theologians such as Adam Wodeham and Gregory of Rimini, but was also influenced by earlier thinkers such as Thomas Aquinas and

Bonaventure. He has serious reservations about the use of logic in theology.

2.4.1 Attributes and ideas

In his discussion of the divine attributes he followed mainly the teachings of Adam Wodeham. God is perfectly one. Divine wisdom and all other perfections attributed to God are in reality as identical to the divine essence as the divine essence is identical to itself. In the divine essence itself there is no distinction or non-identity whatsoever between the attributes of God. Any distinctions between divine attributes are necessarily of a rational (rather than real) nature and are made by us.

A similarly radical stance on the unity of God was assumed in his treatment of divine ideas. Ideas are not formally distinct in God, as some Scotists would argue, but only extrinsically and objectively distinct. Their distinction is a consequence of the differences between the creatures produced by God (which is why Marsilius spoke of extrinsic distinction), and of the fact that they are known by God as different (which accounts for their objective distinction). God knows that he is the cause of infinitely many differences between creatures. That is why his mind contains infinitely many different ideas.

Marsilius criticized Ockham's view that God's idea coincides with creation. If this were true, Marsilius argued, the idea of producing a stone must be identical with either the stone itself, or the stone insofar as it is known by God. If the former, then God must look outside of himself in his idea, which contradicts the position of Augustine, who is quoted by Ockham. If the latter, then the idea of its production is not the stone itself, but rather God's foreknowledge of the stone.

2.4.2 Theology and logic

Marsilius advanced his criticism of the use of logic in theology in his discussion of the position of Robert Holcot. Holcot had argued that logically, God can be called the cause of evil. If God is the cause of every thing (*entitas*) and moral evil (*malum culpae*) is a thing, then God is the cause of evil. Marsilius acknowledged that the argument is based on true premises, yet the conclusion should not be defended as true because it contradicts faith and therefore might cause confusion among believers. Theologians should not flaunt their personal skills in logic, but always write out of reverence for the divine. Their writings should not erode the beliefs of ordinary people, who are not skilled in logic, but rather aim to strengthen them spiritually.

Marsilius was anxious, however, to avoid the implication that God's foreknowledge is somehow dependent upon human beings. In his discussion of Adam Wodeham on the causality of the human will, he complained that Adam had not been emphatic enough on this point, since he allowed the following argument: if an event *E* will happen in the future, then God knows *E* from eternity; but if not-*E* will happen, then God knows not-*E* form eternity; since man is free, he can choose between *E* and not-*E*; therefore, he can change God's foreknowledge. This argument is logically sound, Marsilius argued, but it easily leads to the false conclusion that God's knowledge depends on the free will of man, which is absurd. The eternal cannot fall under the power of that which is created by it. Therefore, this argument

should not be used. It is better to remain on the safe side by maintaining what has always been maintained, namely that God through his absolute omniscience knows the future activities of human beings, but without being dependent on them.

2.4.3 The sacraments

In his treatment of the sacraments at the end of his commentary on the *Sentences*, Marsilius drew heavily on the writings of Thomas Aquinas and Bonaventure. He defended Thomas's view that the word 'this' pronounced by Christ at the Last Supper in uttering 'This is my body' (Mk. 14:22) refers to what the bread and body have in common. Thomas of Strasbourg had attacked this view, but Marsilius showed that the earlier Thomas was right and the later wrong.

In his discussion of the causality of the sacraments, Marsilius followed the exposition of Bonaventure, according to whom the sacraments have no causality of their own. It is God who acts whenever the sacraments are administered correctly. Only in a broad sense is it true to say that the sacraments have the power to act.

3. Influence

The influence of Marsilius has been considerable, particularly through his logical works and commentaries on Aristotle. This may be gathered not only from the large number of manuscripts that have been preserved, but also from several other considerations. Marsilius's commentary on Aristotle's *Prior analytics* was used in Prague in the 1380s. His logical works, including the *Obligationes* and the *Consequentiae,* were used as textbooks in Vienna in the 1390s. His commentaries on Aristotle's *Metaphysics* and *Physics* were read in Krakow during the first sixty years of the fifteenth century. At the universities of Heidelberg, Erfurt, Basle, and Freiburg, his works were studied throughout the fifteenth century, in particular as part of the university curriculum. In 1499, the doctors and masters of the *Via Moderna* at the University of Heidelberg published a volume that included epigrams on Marsilius by well-known humanists such as Jacob Wimpfeling, as well as a defense of Nominalism in the style of Marsilius (*Via Marsiliana*). Praise in the form of epigrams can also be found in the 1501 Strasbourg edition of Marsilius's commentary on the *Sentences*. The *Obligationes*, printed in 1489 under the name of Peter of Ailly, were used by Thomas Bricot, John Major, and Domingo de Soto. The commentary on the *Prior Analytics* was quoted by Agostino Nifo. Jodocus Trutvetter and Bartholomew of Usingen, who consolidated nominalism in Erfurt, repeatedly mentioned Marsilius in their works. Both Leonardo da Vinci and Galileo Galilei referred to Marsilius's commentary on *De Generatione et Corruptione*.

The theological views of Marsilius appear to have had some circulation as well. His commentary on the *Sentences* was known in Krakow in the first half of the fifteenth century, and was used by Thomas de Strampino in his *Principia* (1441-1442). The University of Salamanca had a theological chair (Cátedra de Nominales) for commentary on the works of Marsilius of Inghen and Gabriel Biel. His commentary on the *Sentences* was quoted by Spanish theologians such as Francisco de Vitoria, Domingo de Soto, Luis de Molina, and Francisco Suárez, usually in connection with questions about divine foreknowledge and

grace.

There are nine extant manuscripts of Marsilius's commentary on the *Sentences*. Among the former owners of these manuscripts were two libraries for preachers (Ansbach and Isny), and two libraries of faculties of arts (Erfurt and Leipzig). The education in Erfurt and Leipzig included the reading of nominalist authors. In all probability, the artists became interested in Marsilius's theological work after studying his writings on logic and physics. The presence of Marsilius's commentary on the *Sentences* in preachers' libraries bears witness to the fact that the impact of his work was felt beyond university circles.

Bibliography

Catalogue of works and bibliography

- Hoenen, M. J. F. M., "Marsilius von Inghen: Bibliographie. Appendix zu der geplanten Edition der wichtigsten Werk des Marsilius von Inghen," *Bulletin de Philosophie Médiévale* 31 (1989), 150-167.
- Hoenen, M. J. F. M., Marsilius von Inghen: "Bibliographie. Ergänzungen," *Bulletin de Philosophie Médiévale* 31 (1990), 191-195.
- Lohr, Ch. H., "Medieval Latin Aristotle Commentaries. Authors: Johannes de Kanthi-Myngodus," *Traditio* 27 (1971), 251-351.
- Markowski, M.,"Katalog dziel Marsyliusza z Inghen z ewidencja rekopisow," *Studia Mediewistyczne* 25 (1988), 39-132.

Modern editions

- Marsilius of Inghen, *Quaestiones super quattuor libros Sententiarum, vol. 1: Super primum, quaestiones 1-7*, ed. G. Wieland, M. Santos Noya, M. J. F. M. Hoenen, M. Schulze, Studies in the History of Christian Thought 87, ed. M. Santos Noya, Leiden 2000.
- Marsilius of Inghen, *Quaestiones super quattuor libros Sententiarum, vol. 2: Super primum, quaestiones 8-21*, ed. G. Wieland, M. Santos Noya, M. J. F. M. Hoenen, M. Schulze, Studies in the History of Christian Thought 88, ed. M. Santos Noya, Leiden 2000.
- Marsilius of Inghen, *Treatises on the Properties of Terms*. A First Critical Edition of the Suppositiones, Ampliationes, Appellationes, Restrictiones and Alienationes with Introduction, Translation, Notes, and Appendices, ed. E. P. Bos, Synthese Historical Library 22, Dordrecht 1983.

Secondary Literature

- Braakhuis, H. A. G., and M. J. F. M. Hoenen (eds.), *Marsilius of Inghen*, Artistarium Supplementa 7, Nijmegen 1992 [contains partial editions of works of Marsilius].

- Hoenen M. J. F. M., "Der Sentenzenkommentar des Marsilius von Inghen († 1396). Aus dem Handschriftenbestand des Tübinger Wilhelmsstifts," *Theologische Quartalschrift* 171 (1991), 114-129.
- Hoenen, M. J. F. M., and P. J. J. M. Bakker (eds.), *Philosophie und Theologie des ausgehenden Mittelalters. Marsilius von Inghen und das Denken seiner Zeit*, Leiden 2000 [contains partial editions of works of Marsilius].
- Hoenen, M. J. F. M., Marsilius of Inghen. *Divine Knowledge in Late Medieval Thought*, Studies in the History of Christian Thought 50, Leiden 1993.
- Marshall P., "Parisian Psychology in the Mid-Fourteenth Century," *Archives d'histoire doctrinale et littéraire du Moyen Age* 50 (1983), 101-193.
- Möhler W., *Die Trinitätslehre des Marsilius von Inghen. Ein Beitrag zur Geschichte der Theologie des Spätmittelalters*, Limburg/Lahn 1949.
- Reina M. E., "*Comprehensio veritatis*. Una questione di Marsilio di Inghen sulla Metafisica," *Filosofia e teologia nel trecento. Studi in ricordo di Eugenio Randi*, ed. L. Bianchi, Textes et Études du Moyen Age 1, Louvain-la-Neuve 1994, 283-335.
- Ritter, G., *Studien zur Spätscholastik I: Marsilius von Inghen und die okkamistische Schule in Deutschland*, Heidelberg 1921.
- Wielgus, S. (ed.), *Marsilius von Inghen. Werk und Wirkung. Akten des Zweiten Internationalen Marsilius-von-Inghen-Kongresses*, Lublin 1993 [contains partial editions of works of Marsilius].

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Buridan, John [Jean] | Gregory of Rimini | Ockham [Occam], William

Maarten J. F. M. Hoenen
Catholic University Nijmegen
mhoenen@phil.kun.nl

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 16, 2001

Content last modified: August 16, 2001

Gregory of Rimini

Gregory of Rimini may have been the last great scholastic theologian of the Middle Ages. He was the first thinker to incorporate substantially the developments of both the post-Ockham tradition at Oxford and the post-Auriol tradition at Paris, and his original synthesis had a long-lasting impact on European thought.

1. Life and Work

Gregory of Rimini (a.k.a. de Arimino, Ariminensis, the "Torturer of Infants," and the *doctor acutus* or *authenticus*) was born in Rimini around 1300. He joined the mendicant Order of the Hermits of Saint Augustine (OESA), studied theology at Paris from about 1323 to 1329, and then taught at various Augustinian *studia* in Italy, first at Bologna, then, after 1338, at Padua and Perugia. Almost certainly while he was in Italy, Gregory came into contact with the works of Oxford scholars from the 1320s and 1330s, most notably William of Ockham, Adam Wodeham, Richard Fitzralph, and Walter Chatton. Gregory returned to Paris in 1342 for a year of preparation for his lectures on the *Sentences*, which were given in 1343-44. Gregory probably became Master of Theology in 1345, but he continued to revise his written commentary until 1346, removing certain passages that until recently were considered later *additiones*. In 1347 he returned to Italy to teach in Padua where he stayed until 1351. After that, he taught at the recently established *studium* in Rimini until 1357, when he was elected the Augustinians' prior general, succeeding the late Thomas of Strasbourg. Gregory died in Vienna toward the end of 1358 (see V. Marcolino's chapter in Oberman 1981, 127-94).

Gregory's most important writing by far is his commentary on the first two books of the *Sentences*. Book I survives in some twenty complete manuscripts, while there are about a dozen for book II. The work was printed several times from 1482 to 1532, reprinted in 1955, and finally received a modern critical edition in six volumes in 1979-84 (Rimini 1979-84; Bermon 2001). Parts have been or are being translated into French, German, and English (Rimini, forthcoming). In addition to scriptural commentaries and his letters as prior general, smaller works have also been attributed to Gregory, including a work on usury, *De usura*, printed in Reggio Emilia in 1508 and in Rimini in 1622, and a treatise on the four cardinal virtues, *De quatuor virtutibus cardinalibus*. His tract on the intension and remission of forms, *De intensione et remissione formarum corporalium*, carries the incipit "Circa secundum partem huius distinctionis" and is, therefore, just an excerpt of the *Sentences* commentary, book I, distinction 17, part 2.

2. Position in the History of Philosophy

Although Gregory of Rimini has received considerable attention from historians of medieval thought, understanding his position in the history of philosophy has been made difficult by several problems that have plagued the historiography of fourteenth-century scholasticism. He flourished at a time that has been judged by historians as on the whole decadent, fideistic, and radically skeptical, in contrast to the period in which, for example, Thomas Aquinas worked (d. 1274); this historical viewpoint already made difficult an objective evaluation of Rimini. Historians also labeled Gregory a "nominalist," a term so broad and vague when applied to fourteenth-century thinkers that, when it was used without qualification, it tended to mislead and to obscure the differences among them, as for example between Ockham and Gregory. Finally, unlike Aquinas, Henry of Ghent, and John Duns Scotus, Gregory was active in an as yet severely understudied period, so that placing Gregory in his context is difficult and statements about Gregory's originality are precarious. The history of Gregory's own University of Paris in the quarter century before his *Sentences* lectures, in contrast to that of Oxford in the same years, is particularly unclear. Only careful diachronic studies of specific philosophical problems can provide a precise picture of Gregory's role in the history of philosophy, and few such studies have been accomplished thus far. There are some, however, and epistemology, foreknowledge, and predestination are examples of topics about which we know quite a bit (see below).

Generally, what has been learned so far is that Gregory was really the first to introduce to the University of Paris the exciting ideas developed at English schools between William of Ockham (ca. 1319) and Thomas Bradwardine (ca. 1344). Beginning with Gregory the names of Adam Wodeham, Richard Fitzralph, Walter Chatton, William Heytesbury, Thomas Buckingham, Richard Kilvington, Robert Halifax and others became common knowledge among Parisian scholars. Gregory was also deeply influenced by recent thinkers at his own university, both negatively and positively. The impact of Peter Auriol has long been recognized to be great, but recent studies have made clear that other figures, such as Francis of Marchia, Thomas of Strasbourg, Gerard Odonis, and Michael of Massa, had an influence on Gregory. The question of Gregory's relationship to his Parisian predecessors needs to be investigated more fully.

More clear is Gregory's importance in the late Middle Ages and Reformation. Many scholastics after 1350 copied large passages from his works. Prominent figures who plagiarized or borrowed from Gregory include the Cistercian James of Eltville, Pierre d'Ailly, and Henry of Langenstein, but other important thinkers such as Hugolino of Orvieto OESA, Marsilius of Inghen, and Peter of Candia OFM (Pope Alexander V) knew Gregory's works well. Few philosophers in the later fourteenth century can have been unaffected by his ideas. Gregory's impact both inside and outside the Augustinian Order continued into the fifteenth century. In the celebrated quarrel over future contingents at the University of Louvain (1465-1474), for example, several of the participants cited Gregory's position or even adopted it without attribution. Of course, the fact that only books I and II of Gregory's commentary circulated means that Gregory's direct impact is to be found in topics discussed in those books rather than in issues covered in books III and IV, such as the Immaculate Conception and the Eucharist, which had their own philosophical sub-issues.

Perhaps the most central element of Gregory of Rimini's thought and influence is his adherence to Augustine and the nature of that adherence. For one thing, Gregory simply read Augustine more carefully and extensively than most previous thinkers, and so, for example, Gregory was able to attack Peter Auriol for his faulty citations and quotations of Augustine. Gregory's interest in the works of Augustine has been seen as central to the development of a "historico-critical" method in philosophical theology, especially in the Augustinian Order, partly foreshadowing modern scholarly methods. In connection with this historico-critical method, Gregory was part of a general attempt to establish reliable texts of Augustine and to separate authentic works from the pseudo-Augustinian corpus. Quotations from Augustine, moreover, were cited with great accuracy and detail in Gregory's writings, and so his *Sentences* commentary, when not plagiarized for his own ideas, was often used as a source for Augustinian quotations (Trapp 1956).

Not surprisingly, Gregory's ideas are often Augustinian. Gregory's brand of doctrinal Augustinianism, influenced rather by the Franciscan and Oxonian tradition than the more Dominican (and Parisian) variety of Giles of Rome, soon dominated the Augustinian Hermits' philosophy and theology. Thus by the early 16th century *Aegidistae* and *Gregoriistae* schools of thought existed, and a recognized *via Gregorii* was present in many universities such as Wittenberg, the university of Gregory's fellow Augustinian Hermit Martin Luther (McGrath 1987). The fact that each book of his *Sentences* commentary was printed six times between 1482 and 1532 further helps explain why some of Gregory's ideas often resemble those of Luther and Jean Calvin. Gregory's thought may even have had a life after the reformation in Francisco Suarez.

A list of Gregory's philosophical positions would perhaps not be difficult to make, and neither would it be hard to describe his relation to Ockham on various topics (e.g. Smith 1999). In natural philosophy, for example, in agreement with Ockham, Gregory was a nominalist and employed "Ockham's" razor in denying that sudden change, motion, and time, for example, are independent entities (Brown 1998b). Gregory also claims that the world could have been eternal, and that an actual infinite is possible (Maier 1949). But in these cases one would like to know better the stances of Gregory's immediate predecessors, especially Parisians like Francis of Marchia, in order to determine the possible sources and degree of originality of Gregory's ideas. Otherwise, a list of Gregory's ideas is just that, a mere list. Consequently

the focus here shall be on issues on which the theories of Gregory and his predecessors have been investigated in some depth.

3. Foreknowledge and Contingency

In many ways Gregory was a philosopher's theologian, because he began with propositions from Scripture as premises for his arguments and proceeded deductively. In his deductive theology, Gregory devoted much time and space to defining his terms and exploring exhaustively the implications of possible solutions, a practice that makes his *Sentences* commentary a joy to read and a philosophical classic. In distinctions 38-41 of book I, Gregory tackled the general problem of divine foreknowledge and future contingents and the specific dilemma of predestination and free will. Gregory's positions on these questions have already been the subject of study for many decades, and recently historians have attempted to put Gregory into his immediate Parisian and Oxonian context. Moreover, Gregory's nickname, "the Torturer of Infants," stems in part from his stance on predestination. A discussion of Gregory's thought on these issues, therefore, provides a convenient introduction both to his noetic and to his position in history.

Gregory's treatment of divine foreknowledge and future contingents is aimed primarily at Peter Auriol and secondarily at Oxford theologians (Vignaux 1934, ch. 4; Hoenen 1993, 196-214; Schabel 2000, 264-274; Rimini, forthcoming). In order to preserve the contingency of events stemming from human free will, Auriol claimed that propositions about future contingents are neither true nor false, but rather neutral, and so God does not know that the Antichrist will exist, since "the Antichrist will exist" is neither true nor false. Although like Ockham and Rimini later, Auriol maintained that exactly how God knows the future is incomprehensible to us, he did give a sophisticated explanation and defense of God's knowledge of our future. Gregory, however, chose to focus on the above-mentioned elements in Auriol's position. Gregory recognized that Auriol's theory of future-contingent propositions relies on Aristotle's stance in chapter 9 of *On Interpretation*. Interestingly, although Gregory denied the truth of the position itself, he nevertheless held that it was in fact Aristotle's. Indeed he rejected any attempt to interpret Aristotle differently, in the way that many medieval and modern philosophers have tried to do:

> [This] is apparently a friendly excuse, but in truth it is more of an accusation, because the fact that absurdities ensue [from this position] does not convince us that [Aristotle] did not think that, but convinces us that he ought not to have thought that... Moreover, some modern theologians [i.e. Auriol], great teachers, said that the conclusion [denying determinate truth to future-contingent propositions] not only was the Philosopher's intention, but also that it is very true and even demonstrated... (Rimini 1979, 243).

So for Gregory, Auriol was correct that Aristotle denied the Principle of Bivalence when applied to propositions about future contingents.

Auriol set up two basic rules for such propositions: (1) if a proposition about the future, say, "Socrates will run," is true, it is true immutably and inevitably, since no instant can be found when it would be

false. (2) The significate of such a proposition will inevitably and necessarily be put into being. The foundation for Auriol's claim is his modal theory: immutability and necessity are the same thing. If something is immutable, it cannot be different from what it is, and so it necessarily is the way it is.

Gregory answered with a rigorous and lengthy defense of Bivalence and an alternative modal theory. His defense of Bivalence includes a detailed set of rules for propositions. It is significant that this section of Gregory's text, some seven pages, stems from Francis of Marchia's refutation of Auriol's position, a refutation adopted and extended by Gregory's own Augustinian predecessor at Paris, Michael of Massa. In short, Gregory argued that the Principle of Bivalence applies universally, and Aristotle was wrong to make an exception in the case of future-contingent propositions. Although this was his basic disagreement with Auriol, Gregory was so careful a philosopher that before he refuted Auriol on this point he corrected his Franciscan predecessor on details and in so doing made Auriol's own theory more precise.

Auriol placed greater emphasis on divine simplicity and necessity than on divine freedom and contingency when he was wrestling with one of the fundamental problems of Christian philosophical theology: given an absolutely simple and necessary God, what is the source of contingency? Auriol's own explanation lies in God's relationship with events in time, but this explanation was not of interest to Rimini, who was convinced by Scriptural prophecy that God does in fact know the future, and convinced by logic that the Principle of Bivalence holds universally. So the problem becomes, if God knows that Socrates will run, and the proposition "Socrates will run" is true, will not Socrates run necessarily?

Rimini's answer is a version of the *opinio communis*, a position with roots in Scotus and the Parisian tradition but which Ockham and later Oxford scholars refined with their focus on propositions. (It is possible that Ockham was influenced by Auriol in his concentration on future-contingent propositions, as some have held, but there is nothing specific to indicate that Ockham knew Auriol's treatment, and after Scotus it was natural for theologians to focus their attention on the truth of future-contingent propositions anyway.) The *opinio communis* relies on God's freedom to save contingency in the world: everything other than God is ultimately contingent, because God wills and acts freely and contingently in creating, and so it is logically possible for the things in the world not to have been or to have been otherwise. At the same time, the common position affirms God's immutability and determinate knowledge of such things. The upshot is that true propositions about future contingents have always been true and are immutably true, even determinately true, but that they are only contingently true and not necessarily so. So Gregory denies Auriol's equation of necessity and immutability.

Gregory's position relies on interesting uses of common logical devices and distinctions developed at Paris and Oxford over the preceding century, such as the distinction between the composite and divided senses of propositions, and that between conditional and absolute necessity. The purpose of these distinctions was to offer a way of explaining the contingency of events, but in doing so they assumed the ultimate contingency of everything except God. However, far from being an affirmation of the "radical contingency" of the world, as some historians have claimed, it was in fact the only way for most theologians to save at least some contingency from the threat of absolute logical and divine determinism. In fact, Gregory and others admitted that, assuming God's knowledge of the future, the future was

necessary *ex suppositione*, although not absolutely, because it is logically possible for immutable God to know otherwise. Peter Auriol, and later Peter de Rivo, Pietro Pomponazzi, and Martin Luther, would consider these efforts feeble and deluded. The three Peters resorted to alternative theories that others considered equally feeble and deluded, whereas Luther simply accepted the conclusion that all attempts to save meaningful contingency governed by human free will were doomed to failure.

What is interesting about Gregory's treatment, again, is not his originality, but the clarity and precision with which he presented the common position. He even pointed out problems in the discussions of those with whom he broadly agreed, such as Ockham. True, almost all of what Rimini said could be found in Marchia, Massa, Ockham, Landulph Caracciolo, Adam Wodeham, and others, but not in such an organized fashion.

One final element of Gregory's stance on modal matters that deserves our attention is the contingency or necessity of the past. The *opinio communis* maintained that the past is somehow necessary in a strong sense, even though it is not absolutely necessary. It seems that Gregory did not go so far as to say that the past is necessary (beyond the normal necessity *ex suppositione*), but he does make some sort of modal distinction between the past and the future. Thus we can say that Gregory did not think God can change the past, although there has been some disagreement on this issue (Courtenay 1972-73; Schabel 2000, 271-2). Suffice it to say that the time has come for a long and careful treatment of the modal status of the past in medieval thought, to determine whether any thinker ever really thought the past could be changed. The probable answer is negative.

4. Predestination

Predestination was the traditional subject of distinctions 40-41 of commentaries on book I of the *Sentences*. This was a more purely "theological" subset of the more "philosophical" topic of foreknowledge and future contingents treated in dd. 38-39. As in the case of foreknowledge, Gregory proceeded slowly and carefully, defining his terms and outlining the possible positions. Gregory's Augustinian bent shows through more clearly in predestination than in foreknowledge. Gregory quoted Augustine's words no less than 43 times, and cited him still more often. Frequent scriptural quotations, carefully chosen, provide the ultimate basis for his theory. From Romans 9.13, where Paul comments on Malachi 1.2, "Jacob I have loved, but Esau I have hated," Gregory took his position that from eternity, God actively elects to damn some and to save others, a theory called Double Predestination or Double Particular Election (Vignaux 1934, ch. 4; Schüler 1934; Halverson 1998, 143-157; Schabel 2001; Rimini, forthcoming).

The main issue is what the causal connection is between humans' willing and acting and their salvation or damnation, and predestination or reprobation: do humans participate in or contribute to their own salvation and damnation, or is God's will the sole cause? Traditionally the answer had been that humans are the cause of their deserved damnation, but that salvation depended solely on God's will. Although there were various interpretations of this traditional stance, Peter Auriol seems to have been the first important university scholar to provide a real alternative. Auriol had already sought to distance God from

the everyday details of the world's existence, in order to preserve divine necessity and the contingency of things. Auriol now applied his general theory to the specific issue of soteriology, and claimed that God sets up general rules by which certain sets of people will be damned and other sets saved, without actively choosing to save or damn specific individuals. This maintained divine immutability but had the added bonus of providing symmetry for reprobation and predestination: the determining factor is the presence or absence of an obstacle to grace (*obex gratiae*). For Auriol, while someone's obstacle to grace is indeed a positive cause of reprobation, the absence of such an obstacle, however, is merely a *negative* or privative cause of predestination. Thus Auriol thought he could avoid charges of Pelagianism by simply denying a *positive* cause of predestination in the elect. Ockham appears to have adopted the main elements of Auriol's stance, while Walter Chatton at Oxford and Gerard Odonis and Thomas of Strasbourg at Paris went further and posited a positive cause of predestination in the elect, which would appear to approach the condemned Pelagian doctrine.

Gregory reacted by charging that both the theory of the privative cause and the notion of the positive cause of predestination in those who are predestined are Pelagian. Instead Gregory returned to the traditional view as it concerned predestination: it stems only from God's merciful will. However, Auriol's criticism of the asymmetry of the traditional position led Gregory to claim that not only do the predestined play no causal role in their salvation, but neither do the reprobate contribute to their damnation. In short, there is no reason either for one person's salvation or for another person's damnation except the inscrutable will of God: we do not know why some are saved and others damned. This, after all, Gregory believed, was the theory of Paul and of Augustine.

One has to admire Gregory's consistency here, mirroring that of his opponent, Peter Auriol. In the case of divine foreknowledge, Auriol provided an alternative to the traditional position because he claimed that the common defense of contingency failed. Auriol's theory allowed him to preserve the causal role of humans in reprobation, at the expense perhaps of involving humans in predestination and therefore coming close to Pelagianism. There were problems with Auriol's stance, but it was consistent. Gregory, on the other hand, agreed with the common position on divine foreknowledge, but when it really counted, in soteriology, Gregory took this common position to what he (and Auriol) thought was its logical conclusion. Since God's free creation and action is really the only source of contingency in the world, then God's free will is the only real cause of salvation and damnation. Salvation and damnation are contingent like anything else, but not contingent upon human free will, but merely on God's will. No doubt for Gregory, everyone else who held the *opinio communis* should also have held to Double Predestination or Double Particular Election. Luther and Calvin agreed with Gregory, but they saw no reason for the logical devices of the *opinio communis*, which for them as for Auriol could not save the contingency of human willing.

5. Cognition

Epistemology is another subject in which Gregory's thought has received much attention (e.g. Elie 1937; Dal Pra 1956; Gál 1977; Eckermann 1978; V. Wendland's chapter in Oberman 1981, 242-300; Tachau 1988, 358-71). As in natural philosophy, Gregory maintains a non-realist position that universals are

formed by the soul and only after the mind has previous apprehensions of singular things. Thus sensory experience plays a major role in intellectual cognition. For simple cognition Gregory adopts the common terminology of the dichotomy between intuitive and abstractive cognition, although the difference between the two is based on the objects rather than the modes of cognition. For Gregory, intuitive cognition terminates immediately at the extramental object, but abstractive terminates at the object's species in the soul. Inspired by some of Ockham's successors, Gregory argues against the Venerable Inceptor's claim that via intuitive cognition one could determine whether a thing does *not* exist.

In agreement with Auriol against most contemporaries, however, Gregory also holds that one can have an intuitive cognition of an object that does not exist, as for example when we see a "broken" pencil in a glass of water, when there is only an unbroken pencil in reality. But Auriol is wrong in claiming that this is an instance of an intuitive cognition of something absent, because for Gregory the cognition is really caused by the species of some present object, although perhaps not the object that the mind thinks it is. Therefore Gregory does not adopt Auriol's definition of intuitive cognition as the cognition when the soul merely thinks that the object is present. In any case the dichotomy is different for Gregory because he maintains that abstractive cognition is also somewhat intuitive, since the species of the object is known immediately and therefore intuitively.

In the course of Gregory's long discussion of the problem of foreknowledge and future contingents, he makes frequent reference to the notion of the *complexe significabile*. When it comes to complex cognition, or scientific knowledge, Gregory's inspiration was Adam Wodeham, who built on some of Walter Chatton's ideas in developing the *complexe significabile*. Ockham held that the object of scientific knowledge is the conclusion of a syllogism, and Gregory rejects this. Chatton's alternative was that scientific knowledge has as its object things outside the mind. Gregory also denies this, because

> if this were the case, many sciences would be about contingent things that could be different than they are, whereas for strict science the object must be eternal and necessary. Every being, however, besides God is contingent and not necessary. If things outside the mind were the objects of the sciences, then many sciences, physical and geometrical, and many others, would be about things other than God, and therefore about contingent things (Rimini 1979, 6; Brown 1998a, 171).

One can see here how Gregory's stress on the overarching contingency of creation connects with his epistemology.

Gregory chooses as the object of scientific knowledge the alternative offered by Adam Wodeham. Chatton's notion of "thing" in scientific knowledge was the state of affairs signified by both the negative and the affirmative proposition. For example, "Socrates is sitting" and "Socrates is not sitting" signify for Chatton the same thing, not Socrates, not sitting, and not the propositions, but somehow the whole Socrates's-being-seated. Although Chatton had his reasons for his theory, Wodeham modified it in a useful way, differentiating between positive and negative states of affairs. Thus for Wodeham, each proposition has its own total significate that is only complexly signifiable (*complexe significabile*), so

that Socrates's-being-seated and Socrates's-not-being-seated are two different things, the objects of scientific knowledge.

Gregory adopted Wodeham's theory and tailored it where necessary to his own thought. The complexe significabile, once thought to be Gregory's invention, is neither the proposition itself (although it determines the truth or falsity of the proposition) nor individual things in the world, but rather the arrangement of things in the world. He differed from Wodeham, for example, in the way he thought about "assenting to" and "dissenting from" such *complexe significabilia*, an issue which had occupied Chatton at length. Gregory then applied the notion to a host of other philosophical problems, such as future contingents, and through him the *complexe significabile* became the common intellectual property of continental thinkers, and parallel notions are found in many important later intellectuals.

Now that Gregory's works are available in a reliable modern edition, it is to be hoped that further studies of his Parisian and Oxford predecessors on various single issues will enable us to see his innovations more clearly. Recent studies have shown that he was not always as original as was once thought, but that does not diminish in any way his important position in the history of philosophy. Moreover, Gregory sometimes did come up with new solutions to problems, and even where he did not, his treatments, because of their clarity and comprehensiveness, often became the primary source for later thinkers of the ideas he adopted from his predecessors and developed.

Bibliography

Primary Sources

- Rimini (1979-84): *Gregorii Ariminensis Lectura super primum et secundum Sententiarum*, 6 vols., eds. D. Trapp, V. Marcolino, W. Eckermann, M. Santos-Noya. W. Schulze, W. Simon, W. Urban, and V. Vendland (= *Spätmittelalter und Reformation Texte und Untersuchungen* 6-11) (Berlin/New York: De Gruyter 1979-84).
- Rimini (forthcoming): *Modality, Order, and Transcendence: Gregory of Rimini on God's Knowledge, Power, and Will. An English Translation of His Lectures on the Sentences, Book I, Dist. 35-48*, trans. R. Friedman and C. Schabel (New Haven: Yale forthcoming).

Secondary Works

- Bermon, Pascale (2001): "La Lectura sur les deux premiers livres des Sentences de Grégoire de Rimini O.E.S.A. (1300-1358)," in G.R. Evans, ed., *Medieval Commentaries on Peter Lombard's Sentences, vol. I* (Leiden: Brill 2001), 267-85.
- Brown, S.F. (1998a): "Gregory of Rimini (c. 1300-1358)," *Routledge Encyclopedia of Philosophy* X (1998), 170a-72b.
- ----- (1998b): "Walter Burley, Peter Aureoli, and Gregory of Rimini," in J. Marenbon, ed., *Medieval Philosophy* (= *Routledge History of Philosophy* III) (London: Routledge 1998), 368-85.

- Courtenay, W.J. (1972-73): "John of Mirecourt and Gregory of Rimini on Whether God Can Undo the Past," *Recherches de Théologie ancienne et médiévale* 39 (1972), 224-53, and 40 (1973), 147-74.
- ----- (1978): *Adam Wodeham: An Introduction to His Life and Writings* (= *Studies in Medieval and Reformation Thought* 21) (Leiden: Brill 1978).
- Dal Pra, M. (1956): "La teoria del `significato totale' della proposizione nel pensiero di Gregorio da Rimini," *Rivista critica di storia della filosofia* 11 (1956), 287-311.
- Eckermann, W. (1978): *Wort und Wirklichkeit: Das Sprachverständnis in der Theologie Gregors von Rimini und Sein Weiterwirken in der Augustinerschule* (= *Cassiciacum* 33) (Würzburg: Augustinus 1978).
- Elie, H. (1937): *Le complexe significabile* (Paris: Vrin 1937).
- Gál, G. (1977): "Adam Wodeham's Question on the complexe significabile as the Immediate Object of Scientific Knowledge," *Franciscan Studies* 37 (1977), 66-102.
- García Lescún, E. (1970): *La teología trinitaria de Gregorio de Rimini: Contribución a la historia de la escolástica tardía* (Burgos: Ediciones Aldecos 1970).
- Halverson, J. (1988): *Peter Aureol on Predestination: A Challenge to Late Medieval Thought* (= *Studies in the History of Christian Thought* 83) (Leiden: Brill 1998).
- Hoenen, M.J.F.M. (1993): *Marsilius of Inghen: Divine Knowledge in Late Medieval Thought* (= *Studies in the History of Christian Thought* 50) (Leiden: Brill 1993).
- Leff, G. (1961): *Gregory of Rimini: Tradition and Innovation in Fourteenth Century Thought* (New York: Manchester 1961).
- Maier, A (1949): *Die Vorläufer Galileis im 14. Jahrhundert* (Rome: Edizioni di storia e letteratura 1949).
- McGrath, A.E. (1987): *Intellectual Origins of the Reformation* (Oxford: Blackwell 1987).
- Oberman, H.A., ed. (1981): *Gregor von Rimini: Werk und Wirkung bis zur Reformation* (Berlin: De Gruyter 1981).
- Santos-Noya, M. (1990): *Die Sünden und Gnadenlehre des Gregors von Rimini* (Frankfurt: P. Lang 1990).
- Schabel, C. (2000): *Theology at Paris 1316-1345: Peter Auriol and the Problem of Divine Foreknowledge and Future Contingents* (= *Ashgate Studies in Medieval Philosophy* 1) (Aldershot: Ashgate 2000).
- Schabel, C. (2001): "Parisian Commentaries from Peter Auriol to Gregory of Rimini, and the Problem of Predestination," in G.R. Evans, ed., *Medieval Commentaries on Peter Lombard's Sentences, vol. I* (Leiden: Brill 2001), 221-65.
- Schüler, M. (1934): *Prädestination, Sünde und Freiheit bei Gregor von Rimini* (Stuttgart: Kohlhammer 1934).
- Smith, K. (1999): "Ockham's Influence on Gregory of Rimini's Natural Philosophy," in V. Syros, A. Kouris, and H. Kalokairinou, eds., *Dialexeis: Akademaiko etos 1996-7* (Nicosia: Homilos Philosophias Panepistemiou Kyprou 1999), 107-42.
- Tachau, K.H., *Vision and Certitude in the Age of Ockham: Optics, Epistemology and the Foundations of Semantics, 1250-1345* (= *Studien und Texte zur Geistesgeschichte des Mittelalters* 22) (Leiden: Brill 1988).
- Trapp, D. (1956): "Augustinian Theology of the 14th Century: Notes on Editions, Marginalia,

Opinions and Book-Lore," *Augustiniana* 6 (1956), 146-241.

- Vignaux, P. (1934): *Justification et Prédestination au XIVe siècle: Duns Scot, Pierre d'Aureole, Guillaume d'Occam, Grégoire de Rimini* (Paris: Leroux 1934).
- Würsdörfer, J. (1917): *Erkennen und Wissen bei Gregor von Rimini* (Münster i. W.: Aschendorff 1917).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Bradwardine, Thomas | Ockham [Occam], William

Christopher Schabel
University of Cyprus
schabel@ucy.ac.cy

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published:
Content last modified:

William Penbygull

Wyclif's logico-metaphysical works were very influential at Oxford at the end of the 14th century and the beginning of the 15th. Among the authors who followed his doctrines (the so called Oxford Realists), William Penbygull (+1420) was almost certainly the most faithful to the master, since his extant writings appear to be essentially devoted to a defence and/or explanation of Wyclif's main philosophical theses. Notwithstanding such an attitude, Penbygull gave an original contribution to logic by developing a new theory of identity, which solved the problems that Wyclif's analysis of predication had raised, and by refining Wyclif's theory of predication itself.

Life and Works

The information on the life and works of William Penbygull (or Penbegyll) is scanty. He was from Exeter diocese; he studied at Oxford, where he was fellow of Exeter College in 1399, and rector in 1406-07. He was licensed to preach in the diocese of Bath and Wells on 28 February 1410. He probably died at Oxford in 1420. According to Emden 1957-59, he wrote the following treateses on logic: *De universalibus* (*On Universals*), *Divisio entis* (*The Division of Being*), and *Super Porphyrii Isagogen* (*On Porphyry's Isagoge*).

Universals and Predication

Like Wyclif and many other Realists of the late Middle Ages, Penbygull lists three main kinds of universals: (i) the metaphysical causes of everything, like God and the angelic intelligences; (ii) the general concepts abstracted by our mind, or mental universals; and (iii) the common natures existing in the singulars, or real universals (*De universalibus*, p. 178). Such common natures are type-forms

naturally apt to be present-in and predicated-of a set of individuals, which therefore instantiate them. Real universals are the main metaphysical components of the individuals, but they have no being outside the being of their individuals, as universals and their individuals are really (*realiter*) the same and only formally (*formaliter*) distinct (*De universalibus*, p. 189). In fact, real universals are identical with their own individuals when considered as natures of a certain kind (for instance, man is the same thing as Socrates), but different from them when considered *qua* universals and *qua* individuals respectively, because of the opposite constitutive principles: generality for universals and thisness for individuals (*De universalibus*, p. 181).

Like Walter Burley and Wyclif, Penbygull holds that such formal universals exist in act (*in actu*) outside our minds, and not in potency (*in potentia*) only, as moderate realists (like St. Thomas) thought, since for Pengygull the necessary and sufficient condition that a thing must meet for being a universal is the existence of at least one individual in which it is present (*De universalibus*, p. 178). So the actual existence of universals depends entirely on that of their individuals; without them, common natures could not be really universals.

On the logical side, this description of the relationship between universals and individuals in terms of real identity and formal distinction, entails that not all that is predicated of individuals can be *directly* (*formaliter*) attributed to their universals and *vice versa*. In particular, the accidental forms inhering in substantial individuals (for instance, the whiteness inhering in Socrates) can be predicated of the universal forms proper to these individuals (for instance, the form of humanity or that of animality) only *indirectly* (*essentialiter*), through and in virtue of the individuals themselves. As a consequence, a redefinition of the standard kinds predication and the introduction of a new type, unknown to Aristotle, was required, in order to cover the cases of indirect inherence of an accidental form in a substantial universal, admitted by this theory.

Wyclif, whose conception of universals is the source of Penbygull's, had therefore distinguished three main types of predication: formal predication, predication by essence, and habitudinal predication, each more general than the preceding one. Since the ontological presuppositions of the most general type of predication (habitudinal predication), implied by the other types, are completely different from those of the other two, Penbygull, like other Oxford logicians of his generation, tried to improve Wyclif’ theory by excluding habitudinal predication and redefining the other two kinds in a slightly different way. Penbygull therefore divides predication (which he conceives as a real relation which holds between metaphysical objects [*De universalibus*, p. 188]) into formal predication (*praedicatio formalis*) and predication by essence (*secundum essentiam*). Predication by essence shows a partial identity between subject and predicate, which share some, but not all, metaphysical component parts, and does not require that the form connotated by the predicate-term be directly present in the essence denotated by the subject-term. Formal predication, on the contrary, requires such a direct presence. If the form connotated by the predicate-term is intrinsic to the nature of the subject, then the predication is a case of formal essential predication, while if it is extrinsic, the predication is a case of formal accidental predication. "Man is an animal" is an instance of formal essential predication; "Socrates is white" is an instance of formal accidental predication. Unlike Wyclif, who applied predication by essence to second intentions only -- since he admitted sentences like "(What is) universal is (what is) singular" (that is, *universale est*

singulare) as well-formed and true -- Penbygull thinks that it holds also when applied to first intentions. So he claims that it is possible to predicate of the universal-man (*homo in communi*) the property of being white, if at least one of its individuals is white. However he makes sure to use as a predicate-term a substantival adjective in its neuter form, because only in this way can it appear that the form connoted by the predicate-term is not directly present in the subject, but is indirectly attributed to it, through its individuals. Therefore he acknowledges the proposition "The universal-man is (something) white" (*homo in communi est album*) as a true one, if at least one of the existing men is white (*De universalibus*, pp. 186-88). According to him formal essential predication and formal accidental predication would correspond to Aristotle's essential and accidental predication. But, as a matter of fact, he agrees with Wyclif in regarding predication by essence as more general than formal predication. As a consequence, in his theory formal predication is a particular type of predication by essence. This means that he implicitly recognizes a single ontological pattern, founded on a sort of partial identity, as the basis of every kind of standard philosophical statement (subject, copula, predicate). But in this way, formal essential predication and formal accidental predication are very different from their Aristotelian models, as they express degrees of identity as well as predication by essence.

The Theory of Identity

This interpretative scheme of the nature and kinds of predication is ultimately grounded on a notion of identity, necessarily different from the standard one. According to the most common opinion the logical criteria for identity and (real) distinction were the following:

> *a* is identical with *b* iff for all *x*, it is the case that *x* is predicated of *a* iff it is predicated of *b*;
>
> *a* differs from (is [really] distinct from) *b* iff there is at least one *z* such that *a* is predicated of *z* and *b* is not, or *vice versa*, or there is at least one *w* such that *w* is predicated of *a* and not of *b*, or *vice versa*.

On this basis one can easily conclude that universals and individuals can never be the same, at least because of the forms of generality (which cannot be predicated of individuals) and of thisness (which cannot be predicated of universals). So Penbygull had to put forward new criteria for identity and distinction. First of all, he distinguishes between the notion of non-identity and that of difference (or distinction) and denies that the notion of difference implies that of non-identity (*De universalibus*, p. 190); then he affirms that the two notions of difference and (real) identity are logically compatible (*ibid.*); finally he suggests the following definitions for these three notions non-identity, difference or distinction, and (absolute) identity (*De universalibus*, pp. 190-91):

> *a* is non-identical with *b* iff there is not any form *F* such that *F* is present in the same way in *a* and *b*;
>
> *a* differs from *b* iff there is at least one form *F* such that *F* is *directly* present in *a* but not in

b or vice versa;

a is (absolutely) identical with *b* iff for all forms *F*, it is the case that *F* is present in *a* iff it is present *in the same way* in *b*.

The criterion for non-identity is stronger than the common one for real distinction: two things can be qualified as non-identical iff they belong to distinct categories. On the other side, the definition of difference does not exclude the possibility that two things which differ from each other share one or more properties (or forms). Thus, there are degrees of distinction, and what is more, the degree of distinction between two things can be read as the inverse measure of their (partial) identity. For instance, if we compare the list of the forms (both substantial and accidental) which constitute Socrates and those which make up the universal-man, it is evident that Socrates and the universal-man differ from each other, since there are forms which directly inhere in Socrates and not in the universal-man and *vice versa*; but it is also evident that the two lists are identical for a long section, that is, that Socrates and the universal-man, considered from the point of view of their metaphysical composition, are partially the same. As a result, the copula of the propositions which Penbygull deals with cannot be extensionally interpreted, as it does not mean that a given object is a member of a certain set, nor that a given set is included in another, but it always means degrees in identity between two compound entities.

Bibliography

Edited works

- *De universalibus* (*On Universals*), in A.D. Conti, "Teoria degli universali e teoria della predicazione nel trattato *De universalibus* di William Penbygull: discussione e difesa della posizione di Wyclif," *Medioevo* 8 (1982), pp. 137-203, at pp. 167-203.

Secondary literature

- A.D. Conti, "Teoria degli universali e teoria della predicazione nel trattato *De universalibus* di William Penbygull," (see "Edited works" above), pp. 137-66.
- A.D. Conti, ed., *Johannes Sharpe*, *Quaestio super universalia*, Firenze: Olschki, 1990. See the "Studio storico-critico," at pp. 309-15.
- A. de Libera, *La querelle des universaux. De Platon à la fin du Moyen Age*, Paris: Éditions du Seuil, 1996, at pp. 403-28.
- A.B. Emden, *A Biographical Register of the University of Oxford to A.D. 1500*, 3 vols., Oxford: Clarendon Press, 1957-59, at vol. iii, p. 1455.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle | Burley [Burleigh], Walter | properties | Sharpe, Johannes | Wyclif, John

Alessandro D. Conti
a.conti@tiscalinet.it

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 25, 2001
Content last modified: July 25, 2001

Johannes Sharpe

Johannes Sharpe (ca. 1360 - after 1415) is the most important and original author among the so called `Oxford Realists', a group of thinkers influenced by John Wyclif's logic and ontology. His semantic and metaphysical theories are the final output of the main preceding traditions of thought, since he developed the new form of realism started up by Wyclif, on the one hand, but was open to many Nominalistic criticisms of the traditional Realistic strategies, on the other.

---

1. Life and Works

Johannes Sharpe (Scharp, Scharpe) was from the diocese of Münster in Westphalia, where he was born presumably around 1360. He received his Bachelor of Arts from the University of Prague in 1379, but spent the greatest part of his academic life in Oxford, where he was fellow at Queen's College from 1391 to 1403, and where he became a Master of Arts and a Doctor of Theology. In 1415 he was *lector ordinarius* in Lüneburg (Saxony) (see Conti 1990, p. xvii). The date of his death is unknown.

He established a reputation as a philosopher and a theologian. The number of extant manuscripts of his works and their widespread distribution attest to his importance and notoriety throughout the 15th century. The following writings are attributed to him:

- a treatise on universals (*Quaestio super universalia* [*QsU*] -- his only edited work);
- a commentary by questions on Aristotle's *On the Soul* (*Quaestio super libros De anima* [*In De anima*] -- 8 mss.; all references are to the ms. Oxford, New College 238);
- a commentary by questions on Aristotle's *On Physics* (*Quaestio super libros Physicorum* -- 7

mss.);

- a treatise on the properties of being (*De passionibus entis* -- 3 mss.);
- a treatise on formalities (*De formalitatibus* -- one ms. only);
- an abbreviation of Duns Scotus' *Quodlibeta* (6 mss.);
- a group of six short treatises on theological subjects (one ms. only).

2. The Theory of Meaning

The basic idea of the standard Medieval Realist theories of meaning was that semantic classifications derive from ontological differences among the signified objects. So, according to this approach, the simple expressions of our language (i.e. names) are distinct from the complex expressions (i.e. sentences) by virtue of their own *significata*, that is by virtue of the different kinds of objects they refer to. In fact, the objects signified by complex expressions are compounds of (at least) two of the objects signified by simple expressions and a relation of identity (or non-identity, in the case of a true negative sentence), while a simple object is an item in a category (i.e. either a singular substance, or a substantial form, or an accidental form). Furthermore, every simple expression of our language is like a label that names just *one* object in the world, but whereas proper names and singular expressions label individuals (i.e. token-objects), general terms label common natures (i.e. type-objects), which are the main metaphysical constituents of the set of individuals that instantiate them. For instance, the general expression `man' labels and can stand for each and every man only because of its primarily signifying the universal form of manhood *qua* being present in each and every man as the main constitutive principle of his essence.

Sharpe rejects the standard Realist criteria for the generality (or universality, according to his terminology) of terms, and substantially accepts the inner sense of Nominalist criticisms. In his opinion, to be matched by a common nature really existing in the world is no longer the necessary and sufficient condition for being a general term. According to him, signifying universally (that is, signifying a unitary concept that in turn refers to a multiplicity of things displaying at least a similar mode of being [*QsU*, pp. 129-30]) is a condition for semantic universality equally as important as the previous one. He thinks that not only do terms that signify a common nature existing outside the intellect have to be viewed as common, but also those that signify universally (*ibid.*, p. 69). Thus according to Sharpe there are six different kinds of general expressions, both spoken and written:

1. those that universally singify a common nature really existing in the world (*in re*), like the term `manhood';
2. those that universally connote a common nature really existing in the world, without directly signifying it, like the term `white' (`*album*'), which refers to white things and connotes the form of whiteness;
3. those that do not refer to anything really existing in the world, but are somehow correlated with a universal concept, like the terms `void' and `chimaera';
4. those to which no common nature really existing in the world corresponds, but rather a common transcategorial negative concept, under which a multiplicity of things can be collected, like the term `individual';

5. equivocal terms as such, since they are connected with a multiplicity of different notions;
6. demonstrative pronouns, like `this (one)', when used to supposit for (refer to) a common nature, even though they can signify only in a singular manner (*discrete*) (*ibid.*, pp. 69-71).

The fourth kind of general terms deserves particular attention, since it is connected with Sharpe's solution to the question of the semantic and ontological status of terms of second intentions like `individual' or `singular' -- a question that was very controversial in Oxford at the end of the 14th century. The most common explanation was that proposed by Robert Alyngton, a fellow of Queen's College in the 1380s. According to Alyngton, terms like `individual' have to be considered singular expressions; more precisely they are "range-narrowed" expressions, like `this man', because they identify a singular referent as a member of a given (manifested) set of individuals. In fact, a term like `individual' presupposes a general concept (that of being), the range of which is narrowed to just a unique object among beings by an act of our intellect -- to one object that is not common. Sharpe argues that Alyngton's answer goes against linguistic usage as well as established facts. If Alyngton were right, then the following argument (which everybody will admit) would be formally incorrect:

> man runs (*homo currit*)
> and not the universal-man (*et non homo communis*)
> therefore an individual man runs (*ergo homo singularis currit*),

just like this other one:

> man runs (*homo currit*)
> and not the universal-man (*et non homo communis*)
> therefore Socrates runs (*ergo Sortes currit*),

since the syntagm `an individual man' (`*homo singularis*') would be a singular term standing precisely for only one individual, just like `Socrates' (`*Sortes*'). Furthermore, it is a fact that anyone can understand the sentence `an individual man runs' even without knowing who the man who is running is -- which is, on the contrary, a necessary requisite according to Alyngton's theory. Therefore, Sharpe regarded second intentions of this kind as common ones (*ibid.*, pp. 132-33).

In this way, Sharpe admits that the Nominalist explanation of the universality of signs holds in the particular context of second intentions, implicitly rejecting Alyngton's reduction of epistemology to ontology, since according to Sharpe's account the former has its own range and rules partially independent of those of the latter. Furthermore, he restores the semantic rank that intuitively would be assigned to the `individual'-like terms (something Alyngton was unable to do). On the other hand, his defence of Realism on the problem of universals is partially invalidated by the acceptance, although restricted, of the Nominalist principle of the autonomy of thought in relation to the world. In fact, it is evident that from a semantic and/or epistemological point of view he can no longer justify the extra-mental reality of universals.

Like Burley's system, Sharpe's semantic system too lists a third kind of expression between simple and complex expressions: concrete accidental terms (like `white' or `father'), whose *significata* are neither simple nor complex objects but something in between. He affirms that concrete accidental terms do not signify simple objects but aggregates composed of a substance and an accidental form. Such aggregates are lacking in numerical unity, and hence do not fall into any of the ten categories, because they are not properly beings (*entia*). For that reason concrete accidental terms, although simple expressions from a merely grammatical point of view, are not names. The two metaphysical components of such aggregates (i.e. substance and accidental form) are related to the concrete accidental term as follows: although the concrete accidental term connotes the accidental form, this latter is not its direct *significatum*, so that the concrete accidental term can supposit for the substance only. In other words, the concrete accidental terms label substances by means of the accidental forms from which they draw their names, so that they name substances only *qua* bearers (*subiecta*) of a form. This fact accounts for the difference between general names in the category of substance (like `man') and concrete accidental terms. General names in the category of substance are concrete terms as well, but the form they primarily signify is really identical with the substances they label. Therefore, in this case, the name itself of the form can be used as a name of the substance. This obviously implies a slight difference in meaning between abstract and concrete substantial terms, like `manhood' and `man'. While `manhood' is not the name of the form considered in its totality, but rather the name only of the essential principle of the form, that is, of the intensional content carried by the term `man', this latter term signifies the substantial form considered as a constitutive element of the reality (*esse*) of a certain set of individual substances that instantiate it. As a consequence, according to Sharpe, `man is manhood' (`*homo est humanitas*') is a well formed and true sentence, since both subject and predicate signify the same entity, but `white is whiteness' (`*album est albedo*') is not, since `white' does not directly signify the accidental form, but only the substrate in which it inheres, as bearer of that form, and therefore `white' cannot stand for such a form in any sentence (*ibid.*, pp. 71-73).

3. Universals and Predication

The core of Sharpe's metaphysics lies in his theory of universals. He is a Realist, since he defends the extra-mental existence of universals (*ibid.*, p. 68), but his approach to the whole matter can be defined as "analytical," since he seems to believe that (i) any ontology has to be built up in relation to the resolution of semantic problems, (ii) any philosophical explanation of reality has to be preceded by a semantic explanation of the function of our language, and (iii) that there is not a close correspondence between elements and structures of language and elements and structures of the world. So Sharpe distinguishes two main kinds of universals: universal forms, like manhood, really present in a multiplicity of things, and universal signs, both mental and extra-mental, by means of which we refer to real universals and/or signify something in a universal manner (*ibid.*, p. 50; see also p. 68). On the other hand, the theoretical framework of this division is an analysis of the various meanings of the term `universal'. According to Sharpe, they are six, since we can count the following entities universal:

1. causes that have a multiplicity of effects;
2. the ideas in God;

3. the universal quantifier (*syncategorema universaliter distributivum*);
4. universal propositions, both affirmative and negative;
5. universal forms, or real universals; and
6. universal signs (*ibid.*, pp. 49-50).

The being of real universals coincides with the being of their own individuals, so that real universals can be said to be everlasting, because of the continuous succession of their individuals, and really identical with them. On the other side, universals and individuals are formally different from each other, as they have distinct constitutive formal principles, and therefore different properties (*ibid.*, pp. 91-92). The most important among universal signs are mental universals, which are both the acts of intellection through which our mind grasps the nature of universal forms and the concepts through which it connects general names with the things to which they refer (*ibid.*, pp. 68-69). As a consequence, his position on the problem of universals can be summed up as follows:

1. Universals exist in a twofold way, as common natures *in re* and as concepts in our mind.
2. Real universals are naturally apt to be present in many things as their main metaphysical components.
3. Mental universals are partially caused in our mind by the common natures existing outside.
4. Real universals have no being outside the being of their individuals.

Sharpe's theory of universals is obviously modeled on the canons of the moderate Realism. Nevertheless an important difference divides his position from the most common moderate Realist ones (exemplified by Aquinas' doctrine): whereas according to St. Thomas universals exist *in potentia* outside the mind, and *in actu* only in the mind, according to Sharpe's account they exist *in actu* outside the mind, since their being is the same as the being of individuals, which is actual. For Sharpe a universal is *in actu* if and only if there is at least one individual in which it is present. Therefore our mind does not give actuality to universals, but only a separate mode of existence.

The description of the relationship between universals and individuals in terms of real identity and formal distinction entails (i) that not all that is predicated of individuals can be *directly* (*formaliter*) attributed to their universals and *vice versa*, but (ii) that all that is predicated of individuals has to be in some way or another attributed to universals and *vice versa*. Therefore a redefinition of the standard kinds of predication was required. Like Alyngton and Penbygull, Sharpe modifies Wyclif's theory of predication. Thus he divides real predication, which is a real relation between two entities of the world, into formal predication (*praedicatio formalis*) and predication by essence (*praedicatio essentialis vel secundum essentiam*). Predication by essence shows a partial identity between the subject thing and the predicate thing, which share some metaphysical component parts, and does not require (or even excludes) that the form connoted by the predicate term be directly present in the essence signified by the subject term. Formal predication, on the contrary, requires such a direct presence (*ibid.*, pp. 90-91).

Unlike Alyngton and Penbygull, Sharpe does not divide formal predication into formal essential predication and formal accidental predication, and, as is evident from his formulations, offers two

different readings of the distinction between formal predication and predication by essence. According to Alyngton and Penbygull, predication by essence is more general than formal predication; as a consequence, in their theories formal predication is a sub-type of predication by essence. Besides this interpretation, Sharpe admits another one, according to which the two kinds of predication at issue are complementary although not mutually exclusive. This is the case if predication by essence *excludes* that the form connoted by the predicate term be *directly* present in the essence signified by the subject term (*ibid.*, p. 91).

4. Psychology and Theory of Knowledge

The sources of Sharpe's psychological and epistemological theories are St. Thomas, Duns Scotus, and Ockham, altough this latter is chiefly a polemical source, as Kennedy 1969 pointed out (pp. 253 and 270).

Like Aquinas, Sharpe

1. maintains that the intellectual soul is the immediate form of the human body, so that the whole being of the latter totally depends on the former, although souls are individuated by bodies (*In De anima*, fols. 217v-218r), and
2. claims that each man has his own intellect, arguing against Averroes' thesis of the unicity and separate character of the passive intellect for the whole human species (*ibid.*, fols. 210r-212v).

Like Duns Scotus, Sharpe thinks there is not a real distinction between the soul and its intellective faculties (i.e. the active intellect, the passive intellect, and the will) or among the intellective faculties themselves, but only a formal distinction. On the other hand, the soul's corporeal powers (*potentiae incorporatae*), which depend on bodily organs for their operations, are really distinct from the soul and from each other (*ibid.*, fol. 236v). In this context, Sharpe defines two different kinds of formal distinction. According to the first description, which is very close to that proposed by Scotus in his *Ordinatio*,

> two entities *x* and *y* are formally distinct iff (a) both of them are constitutive elements of the same reality, but (b) neither of them can exist by itself, and (c) neither is part of the definite description of the other.

This is the distinction that holds among the intellective faculties of the soul. The second kind of formal distinction holds between the essence of the soul and its intellective faculties and between a species and its individuals. Sharpe draws the definition of this second kind of formal distinction from Wyclif's *Tractatus de universalibus*. It can be formalized as follows:

> two entities *x* and *y* are formally distinct iff (a) there is at least one *z* such that *z* is predicated of *x* and not of *y*, or *vice versa*, but (b) *x* and *y* are really identical, since one is

directly predicated of the other (*ibid.*, fol. 236r-v).

Like Aquinas and Duns Scotus, and against Ockham, Sharpe affirms that intelligible species are required for intellection (*ibid.*, fol. 244r). The main arguments he uses in favor of this thesis are the following:

1. Our mind's objects of intellection are the universal essences or common natures. But they cannot be present themselves to the mind. Therefore some sign of them, that is, the intelligible species, has to be directly present in the intellect.
2. A universal principle of intellection is necessary in order to understand a universal object, like a common nature. The phantasm is particular, since it is the mental representation of a singular object. Therefore a universal species, abstracted from the phantasm, is required.
3. If there were no species, nothing would be retained by the intellect after an act of intellection. Therefore we could not understand each other, or understand more easily the second time, since there would be no objects for our memory (*ibid.*, fols. 239v-240v).

Finally, like Duns Scotus and Ockham, and against St. Thomas, Sharpe states that our intellect can know perfectly even individual material things (*ibid.*, fol. 253r). What is more, it can know immaterial beings as well, since the most general and proper object of our intellect is being in all its amplitude (*ibid.*, fol. 253v). Sharpe here distinguishes perfect knowledge from complete knowledge. For a perfect knowledge of something it is sufficient that our intellect is able to single the object at issue among any other by a proper concept. For a complete knowledge of something it is necessary that our intellect is able to list all the properties, both substantial and accidental, of the object at issue. It is therefore evident that we can have a perfect knowledge of something without completely knowing it, as is the case with individual material things and immaterial beings (*ibid.*, fol. 254r-v).

Bibliography

Edited works

- *Quaestio super universalia*, A. D. Conti ed., Florence: Olschki, 1990, at pp. 1-145.

Secondary literature

- A. D. Conti, "Studio storico-critico," in J. Sharpe, *Quaestio super universalia* (see Edited works, above), at pp. 211-336.
- A. de Libera, "Questions de réalisme: Sur deux arguments antiockhamistes de John Sharpe," *Revue de metaphysique et de morale* 97 (1992), pp. 83-110.
- A. de Libera, *La querelle des universaux: De Platon à la fin du Moyen Age*, Paris: Éditions du Seuil, 1996, at pp. 403-28.
- A. B. Emden, *A Biographical Register of the University of Oxford to A.D. 1500*, 3 vols., Oxford: Clarendon Press, 1957-59, vol. 3, p. 1680.

- L. Kennedy, "The *De anima* of John Sharpe," *Franciscan Studies,* 29 (1969), pp. 249-70.
- Ch. H. Lohr, "Medieval Latin Aristotele Commentaries: Johannes de Kanthi – Myngodus", *Traditio*, 27 (1971), pp. 251-351, at pp. 279-80.
- H. B. Workman, *John Wyclif: A Study of the English Medieval Church*, 2 vols., Oxford: Clarendon Press, 1926, vol. 2, at pp. 124-25.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Alyngton, Robert | Burley [Burleigh], Walter | Duns Scotus, John | Penbygull, William | universals: the medieval problem of | Wyclif, John

Alessandro D. Conti
a.conti@tiscalinet.it

First published: September 24, 2001
Content last modified: September 24, 2001

Sophismata

In contrast to the meaning the word ‘sophism’ had in ancient philosophy, ‘*sophisma*’ in medieval philosophy is a technical term with no pejorative connotation: a *sophisma* is an ambiguous, puzzling or simply difficult sentence that has to be solved. As an important element of scholarly training in universities, closely related to different kinds of disputations, the *sophismata* not only served to illustrate a theory but, from a more theoretical point of view, were also used to test the limits of a theory. The so-called *sophismata*-literature assumed more and more importance during the thirteenth and fourteenth centuries, and it is not an exaggeration to claim that many important developments in philosophy (mainly in logic and natural philosophy) appeared in texts of this kind, where masters could feel free to investigate problems and develop their own views, much more than they could in more academic and strictly codified literary genres.

1. The Word ‘*Sophisma*’

Although some medieval theologians -- and Humanists even more, of course, like Vivès or Rabelais -- used the words ‘sophism’ or ‘sophist’ as a derogatory designation for quibbling philosophers, ‘*sophisma*’ in medieval philosophical literature has a very precise and technical signification. Hence, to avoid any confusion with fallacies and badly-constructed arguments, we shall here use the original term ‘*sophisma*’ rather than the word ‘sophism’ that even nowadays still has a pejorative connotation.

2. Description and Characteristics

2.1 *Sophisma*-Sentences

There are several important characteristics of *sophismata*. First of all, a *sophisma* is a *sentence* rather than an argument. In particular, a *sophisma* is a sentence that either:

1. is *odd* or has odd consequences,
2. is *ambiguous*, and can be true or false according to the interpretation we give it, or
3. has nothing special about it in itself, but becomes *puzzling* when it occurs in a definite context (or "case," *casus*).

Here are some some examples of kind (1), sentences that are odd or have odd consequences:

> This donkey is your father.
> A chimaera is a chimaera.

As examples of kind (2), *ambiguous* sentences that can be true or false according to the interpretation given to to them, consider:

> All the apostles are twelve.
> The infinite are finite.
> Every man is of necessity an animal.

As an example of kind (3), sentences that have nothing special about them in themselves, but that become puzzling when they occur in a definite context ("case," *casus*), consider:

> The sentence ‘Socrates says something false’, in the case where Socrates says nothing other than ‘Socrates says something false’.

(This is paradoxical, and is one of the forms the Liar paradox can take.)

2.2 The Aim of the Discussion

Once the odd, ambiguous or puzzling *sophisma*-sentence is set out, one should try to understand what it means, what implications it has, and how it fits into or contradicts a particular theory under consideration. This is called "solving the *sophisma*," and is the aim of the entire discussion. The way solutions are searched for and established is very similar to the highly formalized scholastic method for determining a "question":

1. First, one has to examine the arguments *pro* and *contra*.
2. Second, one has to present his own solution to the problem. (Sometimes this part of the discussion is preceded by certain theoretical remarks or clarifications that make the terminology more

precise.)
3. Third, one has to refute the arguments supporting the opposite answer.

An Example

Let us take a very simple example, from Albert of Saxony, *Sophismata*, *sophisma* xi. The *sophisma* is:

> *Omnes homines sunt asini vel homines et asini sunt asini.*
> (All men are donkeys or men and donkeys are donkeys.)

In accordance with step (1), here are the *pro* and *contra* arguments:

> Proof: The *sophisma* is a copulative sentence (in modern logical terminology, a conjunction) each part of which is true; therefore the *sophisma* is true, since its analysis becomes: [All men are donkeys or men] and [donkeys are donkeys].
>
> Disproof: The *sophisma* is a disjunctive sentence each part of which is false; therefore the *sophisma* is false, since its analysis becomes: [All men are donkeys] or [men and donkeys are donkeys].

This is a *sophisma* of the second kind above, one that rests on an ambiguity and can be read with a true interpretation or with a false interpretation. Many such *sophismata*, although not this one, resist being translated from Latin into another language without losing the ambiguity. For example, the sentence 'aliquem asinum omnis homo videt' can be translated by 'Every man sees a donkey' as well as by 'There is a donkey that every man sees'. Similarly, in solving *sophismata*, sometimes Latin word-order is used as an arbitrary code for interpreting the sentence. For example, according to William Heytesbury, when the word 'infinite' is placed at the beginning of a sentence and belongs to the subject, it has to be interpreted as a syncategorematic term; in any other case, it is usually interpreted as a categorematic term (Heytesbury, *Sophismata*, *sophisma* xviii, fol.130va). Such word-order codes might seem like reasonable regimentations of language to a Latin-speaker, but in translation they often seem quite implausible and forced. No such problems arise with this example. (For clarity, square brackets have been inserted into the proof and disproof above, in order to indicate the ambiguity of the *sophisma*.)

In accordance with step (2) above, Albert of Saxony, who discusses this *sophisma*, solves it by just saying that it is either true or false depending on which interpretation we choose. He then takes the opportunity to review the basic principles governing the truth-value of copulative and disjunctive sentences.

In accordance with step (3), we would normally be required to refute the opposite answer. In this case, however, there is nothing to refute, since Albert's solution accepts both the *pro* and the *contra* arguments (for different readings of the *sophisma*).

In general, a *sophisma* was a good occasion to discuss all the problems related to a specific issue. For example, the *sophisma* '*Album fuit disputaturum*' ('The white [thing] was going to be disputed') in thirteenth-century Parisian literature was the occasion to discuss all the problems related to the theory of reference in tensed contexts, as well as to refute the positions others held on this very controversial subject. This is why Pinborg 1977 (p. xv) says that at Paris in the thirteenth century "the *sophismata* seems -- within the faculty of arts -- to play a role analoguous to the *Quaestiones quodlibetales* [quodlibetal questions] in the faculty of theology." Note that this use is quite common. (Note also that Pinborg here uses the word '*sophismata*' to signify not only *sophisma*-sentences but the whole literature that discussed them as well.)

Syncategorematic Terms, Exponible Sentences

It is important to recognize that many *sophismata* involve syncategorematic terms that are responsible for their odd, ambiguous or puzzling character. The preceding *sophisma* can be considered quite characteristic of the genre insofar as we see that the syncategorematic terms 'or' and 'and' occur in it and are responsible for the ambiguity of the sentence.

The expression 'syncategorematic term' should be taken in a broad sense here, so that it not only includes classical syncategorematic terms like 'and', 'if', 'every', etc., but also categorematic terms like 'infinite' or 'whole' that can be used both categorematically and syncategorematically. Thus the sentence "*Infinita sunt finita*" ("The infinite are finite" -- here, incidentally is another good example of a *sophisma* that cannot be translated into English without disambiguating it) is false if 'infinite' is used categorematically, for in that case its signification is "Things that are infinite are finite." But it is true if 'infinite' is used syncategorematically, for in that case its signification is "Finite things are infinite in number" or "There are infinitely many finite things." (See Heytesbury, *Sophismata*, *sophisma* xviii, fol.130va.)

Many *sophismata* too are what medieval logicians called "exponible sentences", sentences that seem to be simple but actually imply several other sentences into which they can be decomposed. For example, the sentence "<u>A</u> differs from <u>B</u>" was said to be equivalent to "<u>A</u> exists and <u>B</u> exists and <u>A</u> is not <u>B</u>"; the sentence "<u>A</u> ceases to be white" was said to be equivalent either to "Now <u>A</u> is white and immediately after this <u>A</u> will not be white" or to "Now <u>A</u> is not white and immediately before this <u>A</u> was white", depending on the theory.

2.3 The Main Fields in Which *Sophismata* Are Discussed

Just as the scholastic method can be applied to any subject, the use of *sophismata* is to be found in logic, grammar and physics as well as in theology. Let us concentrate here on the first three.

Logical *Sophismata*

As seen above, logical *sophismata* are closely linked to the discussions of syncategoremata. The aim is

either to determine the truth-value of a sentence (including sentences involving self-reference) or to discuss subjects such as:

- The syntactic and semantic properties of terms (including the difference between meaning and reference) in sentences like "Every man sees every man," "You are a donkey," and "I promise you a horse."
- Quantification and existential import, as in the sentence "Every phoenix is."
- The theory of negation and "infinite" words, as in the sentence "Nothing and a chimaera are brothers."
- The problem of universals, as in "Man is a species."
- The composite and divided senses of a sentence and the scope of modal operators, as in "The white can be black," "Every man is of necessity an animal," etc.

We could compare these discussions to contemporary discussions of sentences like "The morning star is the evening star."

Physical *Sophismata*

The aim here is to discuss physical concepts (motion, change, velocity, intension and remission of forms, maxima and minima, etc.). But, as seen above with the *sophisma* "The infinite are finite," physical problems are treated as logical and conceptual problems. This logico-semantical approach to physical problems is quite characteristic of medieval physics and should be kept in mind when we wonder the extent to which medieval physics can be considered a precursor to modern physics.

With respect to so-called physical *sophismata*, special attention should to be paid to certain fourteenth century English authors known as the "Oxford Calculators," authors like Richard Kilvington, William Heytesbury, Thomas Bradwardine, Richard and Roger Swineshead. These people developed a peculiarly "English-style" of *sophismata.* Based on the theological dogma of the absolute power of God, the distinction between what is physically possible and what is logically possible (where non-contradiction is the only limit) allowed these authors to make use of imaginary thought experiments. For example, "Suppose that A is a distance to be traversed which Socrates cannot traverse, and that his power is increased until Socrates can traverse distance A completely, and that Socrates' power is not increased further." Is the *sophisma* "Socrates will begin to be able to traverse distance A" true or false? (Richard Kilvington, *Sophismata, sophisma* 27, in Kretzmann 1990, p.60.) Thought experiments like these led these authors to, among other things, a theorem for uniformly accelerated motion (Thomas Bradwardine's "Mean Speeed Theorem").

Grammatical *Sophismata*

Sophismata like "Love is a verb," "O Master," "It rueth me" or "I run" gave rise to very sharp discussions of grammatical categories and theories. For example, does a change of word order change the meaning of a proposition? Can a participle be a subject? How should we interpret interjections? Can ‘*est*’ ("is") be

used impersonally?

3. The Various Roles of *Sophismata*

The first and most evident role of *sophismata* is pedagogical. In theoretical treatises, *sophismata* can play various roles. They can be used to explain a given statement or rule, illustrate a distinction or an ambiguity, show what would follow if a rule were violated, or test the limits of a theory.

In addition, although some differences can be identified between the Paris and the Oxford traditions, *sophismata* are important as oral exercises (disputations) in a student's training in philosophy, especially in the first years of universitary education in the Faculty of Arts. Nevertheless, it is clear that, while Heytesbury's *Rules for Solving Sophismata* is written for undergraduate students -- at Oxford '*sophista*' was the official name given to students who had disputed "on *sophismata*" ("*de sophismatibus*") for about two years -- this is probably not the case for his *Sophismata*, the discussions in which are much more complicated.

I think it is no exaggeration to say that *sophismata* in the Faculty of Arts were as important as Biblical exegesis in the Faculty of Theology.

4. Related Kinds of Treatises

If we look at the evolution of literary genres, we note that the twelfth- and early-thirteenth century *syncategoremata*-literature came to be absorbed in the *sophisma*-literature. In thirteenth and fourteenth century philosophical literature, *sophismata* can appear within many kinds of treatises. There are collections of *sophismata* named simply *Sophismata* or *On Sophismata*, but *sophismata* are also important in works -- often by the same authors, or by different authors coming from the same milieu as the former collections -- with titles like *Abstractions*, *Distinctions*, *On Exponibles*, *On Consequences*, *Sophistry*, etc.

Even if there are technical distinctions among these types of tracts, all of them play the same roles mentioned just above -- in short: to acquire logical abilities that can be applied to any subject.

Bibliography

The medieval *sophismata*-literature is a vast and complex subject of research. Many questions are still unsolved, especially about its historical origins and development. It is of central interest for people interested in medieval logic, grammar and physics, but also for those interested in the history of universities.

The study of "sophismatic works" began around 1940 with Grabmann's *Die Sophismatalitteratur des 12.*

und 13. Jahrhunderts, and much work has been done in the last two decades. But there are still a lot of texts to read, edit and analyze.

The bibliography is organized as follows:

- Main texts (in alphabetical order by medieval authors)
- Translations (in alphabetical order by medieval authors)
- Main studies (in alphabetical order by modern authors).

Main texts

Most of the logical and grammatical texts on *sophismata* have been edited by S. Ebbesen and his collaborators in the review *Cahiers de l'Institut du Moyen Age Grec et Latin*, University of Copenhagen. We will here mention only books.

- de Libera, A. *César et le Phénix. Distinctiones et sophismata parisiens du XIIIe siècle.* Centro di cultura medievale, 4; Pisa: Scuola Normale Superiore, 1991.
- De Rijk, L. M. *Some Earlier Parisian Tracts on Distinctiones sophismatum*. Nijmegen: Ingenium Publishers, 1988.
- Scott, T. K. *Johannes Buridanus. 'Sophismata'. Critical Edition with an Introduction.* Grammatica Speculativa, 1; Stuttgart-Bad Cannstatt: Frommann-Holzboog, 1977.
- Pironet, Fabienne. *Iohanni Buridani Summularum Tractatus nonus: De practica sophismatum (Sophismata). Critical Edition and Introduction.* Nijmegen: Ingenium Publishers, forthcoming.
- Kretzmann, N., and Kretzmann, B. E. *The 'Sophismata' of Richard Kilvington.* Oxford: University Press for The British Academy, 1990. (Critical edition.)
- Pinborg, J. *Sigerus de Cortraco, 'Summa modorum significandi sophismata'; New Edition, on the Basis of G. Wallerand's editio prima, with Additions, Critical Notes, an Index of Terms and an Introduction.* Amsterdam: J. Benjamins, 1977.
- Longeway, J. *William Heytesbury: On Maxima and Minima. Chapter 5 of 'Rules for Solving Sophismata', with an Anonymous Fourteenth Century Discussion, a Translation with an Introduction and Study*. Synthese Historical Library, 26; Dordrecht: Reidel, 1984.
- Pironet, Fabienne, *Guillaume Heytesbury, Sophismata asinina. Une introduction aux disputes médiévales. Présentation, édition critique et analyse*. Collection Sic et Non; Paris: Vrin, 1994. (With texts from the *Libelli sophistarum ad usum Oxoniensis.*)

Translations

- Scott, T. K. *Sophisms on Meaning and Truth*. New York: Appleton Century Crofts, 1966. (Translation of John Buridan's *Sophismata*.)
- Biard, J. *Jean Buridan, Sophismes*. Collection Sic et Non; Paris: Vrin, 1993.
- Hughes, G. E. *John Buridan on Self-Reference. Chapter Eight of Buridan's 'Sophismata'. An*

Edition and a Translation with an Introduction and a Philosophical Commentary. Cambridge: Cambridge University Press, 1982. (The paperbound edition omits the Latin text.)

- Kretzmann, N., and Kretzmann, B.E. *The 'Sophismata' of Richard Kilvington.* Cambridge: Cambridge University Press, 1990. (English translation, historical introduction and philosophical commentary.)

Main Studies

Many important studies are to be found in the following collective work: Read, S., (ed.) *Sophisms in Medieval Logic and Grammar. Acts of the Ninth European Symposium for Medieval Logic and Semantics, St. Andrews, June 1990.* Dordrecht: Kluwer Academic Publishers, 1993.

- Biard, J. "Les sophismes du savoir: Albert de Saxe entre Jean Buridan et Guillaume Heytesbury." *Vivarium* 27 (1989), 36-50.
- Biard, J. "Verbes cognitifs et appellation de la forme selon Albert de Saxe." In S. Knuuttila, R. Työrinoja, and S. Ebbesen, ed. *Knowledge and the Sciences in Medieval Philosophy. Proceedings of the Eighth International Congress of Medieval Philosophy (S.I.E.P.M.). Helsinki, 24-29 August 1987.* Helsinki: Yliopistopaino, 1990, Vol.II, pp.427-35.
- Ebbesen, S. "The Dead Man is Alive." *Synthese* 40 (1979), 43-70.
- Grabmann, M. *Die Sophismatenliteratur des 12. und 13. Jahrhunderts mit Textausgabe eines Sophisma des Boetius von Dacien. Beiträge zur Geschichte der Philosophie une Theologie des Mittelalters. Texte und Untersuchungen.* Band 36, Heft 1; Münster i. W.: Aschendorff, 1940.
- Knuuttila, S. and Lehtinen, A. I. "Plato in infinitum remisse incipit esse albus. New Texts on the Late Medieval Discussion on the Concept of Infinity in Sophismata Literature." In E. Saarinen, R. Hilpinen, I. Niiniluoto, M. P. Hintikka, ed. *Essays in Honour of J. Hintikka.* Synthese Library, 124; Dordrecht: D. Reidel Pub. Co., 1979, 309-29.
- Kretzmann, N. "Syncategoremata, exponibilia, sophismata." In N. Kretzmann, *et al*, ed. *The Cambridge History of Later Medieval Philosophy from the Rediscovery of Aristotle to the Disintegration of Scholasticism, 1100-1600.* Cambridge: Cambridge University Press, 1982, 211-45.
- Kretzmann, N. "Continuity, Contrariety, Contradiction and Change." In N. Kretzmann, ed. *Infinity and Continuity in Ancient and Medieval Thought. Papers Presented at a Conference held at Cornell University on April 20 and 21 1979, under the Title 'Infinity, Continuity and Indivisibility in Antiquity and the Middle Ages'*. Ithaca: Cornell University Press, 1982, 322-40. (With an appendix: "Text of Walter Burleigh and the Sophisms 8 and 16 of Richard Kilvington.")
- Kretzmann, N. "Tu scis hoc esse omne quod est hoc: Richard Kilvington and the Logic of Knowledge." In N. Kretzmann, ed. *Meaning and Inference in Medieval Philosophy. Studies in Memory of Jan Pinborg.* Dordrecht: Kluwer, 1988, 225-45.
- de Libera, A. "La littérature des *Sophismata* dans la tradition terministe parisienne de la seconde moitié du XIIIe s." In M. Asztalos, ed. *The Editing of Theological and Philosophical Texts from the Middle Ages. Acts of the Conference Arranged by the Department of Classical Languages, University of Stockholm, 29-31 August 1984.* Acta universitatis Stockholmiensis. Studia Latina

Stockholmiensia, 30; Stockholm: Almqvist and Wiksell International, 1986, 213-44.
- de Libera, A. "La littérature des *Abstractiones* et la tradition logique d'Oxford." In P. O. Lewry, ed. *The Rise of British Logic. Acts of the Sixth European Symposium on Medieval Logic and Semantics. Balliol College, Oxford, 19-24 June 1983.* Papers in Mediaeval Studies, 7; Toronto, Pontifical Institute of Mediaeval Studies, 1983, 63-114.
- de Libera, A. "La problématique de l' ’instant du changement’ au XIIIe siècle: contribution à l'histoire des *sophismata physicalia*." In S. Carloti, ed. *Studies in Medieval Natural Philosophy*. Florence: Leo S. Olschki, 1989, 43-93.
- Murdoch, J. E. "Mathematics and Sophisms in Late Medieval Natural Philosophy." In *Les genres littéraires dans les sources théologiques et philosophiques médiévales. Actes du colloque international de Louvain-la-Neuve, 25-27 mai 1981*. Université Catholique de Louvain, Publications de l'Institut Supérieur d'Etudes Médiévales. Deuxième Série: Textes, Etudes, Congrès, 5; Louvain-la-Neuve: Institut d'Etudes Médiévales de L'Université Catholique de Louvain, 1982, 85-100.
- Murdoch, J. E. "Infinity and Continuity." In N. Kretzmann, *et al.*, ed. *The Cambridge History of Later Medieval Philosophy from the Rediscovery of Aristotle to the Disintegration of Scholasticism, 1100-1600.* Cambridge: Cambridge University Press, 1982, 564-91.
- Rosier, I. and Roy, B. "Grammaire et liturgie dans les sophismes du XIIIe siècle." *Vivarium* 28 (1990), 118-35.
- Rosier, I. "Les sophismes grammaticaux au XIIIe s." *Medioevo* 17 (1991), 175-230.
- Sylla, E. D. "William Heytesbury on the Sophism infinita sunt finita." In *Sprache und Erkenntnis im Mittelalter. Akten des 6. Internationalen Kongresses für mittelalterliche Philosophie der Société Internationale pour l'étude de la philosophie médiévale, 29. August-3. September 1977 im Bonn.* Miscellanea Mediaevalia, 13.1-2; Berlin: W. de Gruyter, 1981, Vol.II, 628-36.

Other Internet Resources

- Fabienne Pironet's William Heytesbury *Sophismata* site. (Working text of the Latin.)
- Paul Spade's Mediaeval Logic and Philosophy page site.

Related Entries

Buridan, John [Jean] | Burley [Burleigh], Walter | Heytesbury, William | insolubles [= *insolubilia*] | Kilvington, Richard | Richard the Sophister [*Ricardus Sophista, Magister abstractionum*] | terms, properties of: medieval theories of

Fabienne Pironet
pironetf@philo.umontreal.ca

First published: September 29, 2001
Content last modified: September 29, 2001

Richard Kilvington

Richard Kilvington (ca. 1302-1361), Master of Arts and Doctor of Theology at Oxford, was a member of the household of Richard de Bury, then Archdeacon of London, and finally Dean of Saint Paul's Cathedral in London. Along with Walter Burley and Thomas Bradwardine, he formed the first academic generation of the school known as the ‘Oxford Calculators’. Although he introduced a number of original ideas and methods in logic, natural philosophy, and theology, and influenced his contemporaries and followers, he has been little studied until recently.

Life and Works

Richard Kilvington (we know almost seventy different spellings of his name) was born at the beginning of the fourteenth century in the village of Kilvington in Yorkshire. He was the son of a priest from the diocese of York. He studied at Oxford, where he became Master of Arts (1324/1325) and then Doctor of Theology (ca. 1335). His academic career was followed by a diplomatic and ecclesiastical one. He was in the of service of Edward III and took part in diplomatic missions, culminating in his service as Dean of St. Paul's Cathedral in London. Along with Richard Fitzralph, Kilvington was involved in the battle against mendicant friars. It seems that Kilvington's argument with mendicants continued almost until his death in 1361.

Except for a few sermons, all of Kilvington's known works stem from his lectures at Oxford, and none of these uses the typical late-medieval style in question commentaries, which followed the order of books in the respective works of Aristotle. Rather, in accordance with the fourteenth century Oxford practice, Kilvington reduces the number of topics discussed to some central and probably most important subjects, each of which is constructed as a set of fully developed questions, no more than 10 in each set. This reduction in the range of topics is offset by a lengthier and much more detailed analysis of the particular questions chosen for treatment, some of which print to over 120 pages in modern editions. His

philosophical works, the *Sophismata* and *Quaestiones super De generatione et corruptione,* both composed before 1325, were the result of lectures as a Bachelor of Arts; his *Quaestiones super Physicam* (1325/26) and *Quaestiones super Libros Ethicorum* (1326/1332) come from his period as a Master of Arts; finally, he composed his ten questions on Peter Lombard's *Sentences* in the Faculty of Theology before 1334. (Only the *Sophismata* has been edited and translated, in Kretzmann 1990, 1991; for the titles of other questions and their manuscripts, see Jung-Palczewska 2000b.)

Method in Science

Like a great many Oxford thinkers of the period, Kilvington is convinced that mathematics is useful in any branch of scientific inquiry that deals with measurable subjects. His “measure mania” emerges in his special interest in the problem of change, and in the application of mathematics to determine, that is to “measure”, the phenomenon in question. He makes broad use of the most popular fourteenth-century calculative techniques to solve not only physical but also ethical and theological problems. Four types of calculations can be found in Kilvington's works. The first and most predominant measure, by limits -- i.e., by the first and last instants of the beginning and ending of a continuous process, and by the intrinsic and extrinsic limits of capacities of passive and active potencies -- is not obviously mathematical, but it does raise mathematical considerations insofar as it prescribes a measure for natural processes. The second type of calculation, by latitude of forms, covers processes in which accidental forms or qualities are intensified or diminished, e.g., in the distribution of such natural qualities as heat or whiteness. This is also applied to theology to measure the latitude of moral qualities such as love, grace, sin, will or desire, and to clarify the nature of communication between God and man in general, or between God and the blessed in particular. The third type of calculation is more properly mathematical, and employs a new calculus of compounding ratios in order to measure the speed of local motion. Finally, the fourth type of calculation involves a “rule” that allows one to compare infinities as infinite sets containing infinite subsets, and thereby to determine their relative size. Kilvington finds a way to apply these techniques of measurement in his logical, physical, ethical, and theological writings, and their presence is one of the most distinctive features of his work.

Logic

Kilvington's only logical work is the *Sophismata*. A *sophisma* is neither a standard paradox of disputation nor a sophistical argument but a sentence or proposition whose truth is at issue. The basic structure of the work involves the statement of a *sophisma* sentence, the presentation of a case or hypothesis, arguments for and against the *sophisma* sentence, and finally, the resolution or reply to the *sophisma* sentence and to the arguments on the opposing side.

Kilvington's sophisms are intended to be of logical interest, but they also pose some important physical questions. In constructing his sophisms, Kilvington sometimes appeals to observable physical motion, but just as often makes use of purely imaginary cases having no reference to external reality. Although such cases are physically impossible, they are theoretically possible since they do not involve a formal

contradiction: “and for purpose of the sophisma, that is enough [*unde licet casus idem positus sit impossibilis de facto...tamen per se possibilis est; et hoc sufficit pro sophismate*]” (S 29, p. 69; see also Kretzmann 1990 p. 249).

The first eleven sophisms use the process of whitening to consider the motion of alteration as a successive entity extrinsically limited at its beginning and end. There is no first instant of alteration, claims Kilvington, only a last instant before the alteration begins; likewise, there is no last instant of alteration, only a first instant marking the introduction of the final degree of the quality in question. Thus, motion yields no minimum degree of whiteness or speed, but rather smaller and smaller degrees *ad infinitum* down to zero, since the qualities change continuously. The set of integers is potentially infinite because one can always find a higher integer, but not actually infinite since there is no infinitely great number. Since, in Kilvington's opinion, any continuum of time, space, motion, heat, whiteness, etc., is infinitely divisible, it can be understood quantitatively and measured using infinite sets of integers. Sophisms 29-44 reveal Kilvington's special interest in the causes of local motion, i.e., active and passive potencies, and in effects such as time, distance traversed, and speed. He considers both uniform and difform motion caused by voluntary agents, and calls attention to the dubious measure of instantaneous speed through the comparison of speed in uniform and accelerated motion (see Kretzmann 1982).

The last four sophisms raise issues in epistemology and the logic of knowledge by means of propositions on knowing and doubting involving intentional contexts, e.g., in sophism 45, ‘You know this to be everything that is this’. One of the most interesting historically is sophism 47, ‘You know that the king is seated’, in which Kilvington departs from the usual format to call some of the rules of obligational disputation into question (for discussion, see Kretzmann and Stump 1985; Kretzmann 1990, pp. 330-347).

Natural Philosophy

Although Kilvington does not enjoy the reputation in natural philosophy that he does in logic, recent research has revealed that his questions on Aristotle's *De generatione et corruptione* and *Physics* influenced Thomas Bradwardine's theory of motion and rule concerning velocities in motion (see Jung-Palczewska 2000b). Both works were written before 1328, i.e., before Bradwardine's important treatise, *On the Ratios of Velocities in Motions*.

Like most medieval natural philosophers, Kilvington subscribes to the general Aristotelian principles of motion. He follows Ockham, however, in allowing substance and quality as the only two kinds of really existing thing. The reality of motion is explained in terms of the mobile subject and the places, qualities, and quantities it successively acquires. Consequently, Kilvington is mostly interested in measuring local motion in terms of its actions or causes, the distance traversed and time consumed, rather than the “intensity” of its speed. It is his analysis of local motion that places Kilvington among the 14th-century pioneers who considered the problem of motion with respect to its causes (*tamquam penes causam*), corresponding to modern dynamics, and with respect to its effects (*tamquam penes effectum*), corresponding to modern kinematics.

Unlike many later Oxford Calculators, Kilvington did not advance any clear rules concerning the different kinds of division when examining the dynamical aspect of motion in the problem of setting boundaries to capacities or potencies involved in active/passive processes. But he did articulate most of the issues which were at stake and pose the questions which influenced the solutions of later Calculators (see Wilson 1956, Jung-Palczewska 2000a). Like Heytesbury (for whom, see Longeway 1984) Kilvington pointed to two different considerations which have to be met: one which refers to the everyday use of language, describing real, physical phenomena; and another which refers to a formal, i.e. logico-mathematical, language dealing with the questions in the realm of speculative, i.e., mathematical physics.

Kilvington's belief in the power of mathematics in natural philosophy led him to a new theory of ratios. In order to produce a mathematically coherent theory, he insists (in keeping with Euclid's definition from the fifth book of the *Elements*) that a proper double proportion is the multiplication of a proportion by itself. Moreover, since, in the opinion of Averroes, "the ratio of speeds in motions follows the ratio of the power of the mover to the power of the thing moved", variations in speed must be tied to variations in the proportion of forces and resistances. Consequently, the proper way of measuring the speed of motion is to describe its variations by the double ratio of motive force (F) and resistance (R), as defined by Euclid. The speed of motion thus varies arithmetically, whereas the proportion of force to resistance determining these speeds varies geometrically. Accordingly, when the proportion of force to resistance is squared, the speed will be doubled (for a comprehensive explanation of this new theory see Murdoch and Sylla 1976). Kilvington is aware that a proper understanding of Euclid's definition requires a new interpretation of Aristotle's principles of motion, and concludes that when he is talking about a power moving half of a mobile, Aristotle means precisely the subdouble ratio of F to R, and that when he is talking about power moving a mobile twice as heavy, he means the squaring of the ratio of F to R. Kilvington's function provides values of the ratio of F to R greater than 1:1 for any speed down to zero, since any root of the ratio greater than 1:1 is always a ratio greater than 1:1. Hence, he avoids a serious weakness of Aristotle's theory, which cannot explain the mathematical relationship of F and R in very slow motions.

According to Aristotle, a temporal motion can occur only if there is some resistance playing the role of a *virtus impeditiva*, together with an acting power greater than the resistance of the medium. Since there is no resistance in a vacuum, motion is impossible there. Although he holds that vacua do not exist in nature, he considers the possible temporal motion of both mixed and simple bodies in a vacuum. Kilvington's most interesting and original idea here concerns the possible temporal motion of a simple body. Such a motion could be caused only by internal resistance brought about by the unequal distance traversed by the different parts of a simple body unequally distant from the center of the Earth, relative to which they move. Since the question on motion in a vacuum is posed *secundum imaginationem* and deals with purely imaginable cases, it seems that Kilvington has here combined Aristotelian realism with Ockham's particularist ontology.

Local motion considered in its dynamic aspect, i.e., when speed is proportional to the ratios of Fs to Rs, describes only the value of speed measured at an instant and not successive changes of speed over time. In order to characterize changes of speed in motion, one must study the problem of local motion in its

kinematic aspect. Like all of the later Calculators, Kilvington does not consider speed to be a quality, so there is no real, existential referent for instantaneous speed. Therefore, speed has to be measured by distances, i.e., latitudes of quality (formal distance) or quantity traversed, and such traversals take time unless the speed is infinitely great. In his questions on the *Physics*, Kilvington considers all sorts of motion: uniform, uniformly difform, and difformly difform local motion.

Besides the new function describing the speed of motion, which was eventually adopted and developed by Thomas Bradwardine, one of the most notable achievements of Kilvington's theory of local motion is its awareness of the different levels of abstraction involved in the problem. Although his account frequently proceeds *secundum imaginationem* in the direction of “speculative physics”, it never renounces empirical verification. Kilvington ponders questions which would never arise as a result of direct observation, since the structure of nature can only be uncovered by highly abstract analyses. Such abstractions, however, arise from genuine realities, and cannot contradict them. He saw physics and mathematics as complementary, i.e., as two different ways of describing natural phenomena. Reality provides the starting point for the more complicated mental constructions which in turn make it comprehensible. While mathematics is the proper way to solve these problems, logic remains the most convenient way to pose them. These different methods together guarantee the objective and demonstrative character of the natural sciences. Like other Oxford Calculators, Kilvington refrained from including God in his speculations on natural science, which remained focused on nature as the proper subject of physics. Since the laws of nature are a reflection of God's ordained power, he saw no need to recall this obvious fact while entertaining these laws or considering thought experiments.

On the one hand, Kilvington never abandons the realm of Aristotelian physics or rejects the principles laid down in his natural philosophy. But on the other, his tendency to combine mathematics and physics frequently led him beyond Aristotle's theories to seek solutions to the many paradoxes which resulted from his principles. Kilvington's Aristotelian principles hide essentially Ockhamist views. Despite the fact that he never explicitly mentions the name of Brother William, it is beyond doubt that he not only knew the opinions of the *Venerable Inceptor* but also accepted them as the natural way of understanding the works of the Philosopher.

Impact and Influence

Kilvington's teachings on logic were influential both in England and on the Continent. Richard Billingham, Roger Rosetus, William Heytesbury, Adam Wodeham, and Richard Swineshead were among those English scholars who benefited from Kilvington's *Sophsimata*. His *Quaestiones super De generatione et corruptione* was quoted by Richard FitzRalph, Adam Wodeham, and Blasius of Parma. His *Quaestiones super Physicam* was well known to the next generation of Oxford Calculators: John Dumbleton and Richard Swineshead. It seems that the latter text also influenced Parisian masters such as Nicole Oresme and John Buridan. Thomas Bradwardine, however, was the most famous beneficiary of Kilvington's questions on motion. In his renowned treatise on the velocities of motion, Bradwardine included most of Kilvington's fundamental arguments for a new function describing the relation of motive powers and resistance. Kilvington's other questions on the *Ethics* and the *Sentences* enjoyed a

reputation not only in Oxford but also in Paris, where they were frequently quoted by Adam Junior, John of Mirecourt, Johanes de Burgo, Thomas of Cracow (see Jung-Palczewska 2000b).

Kilvington played an important role in the diffusion of the ideas of early Oxford Calculators, even if Thomas Bradwardine eclipsed him. In his works on the philosophy of nature, he raised many fundamental questions, often solving them in an original and sophisticated manner.

Bibliography

Critical Edition and Translation

- Richard Kilvington, *The Sophismata of Richard Kilvington*: 1990, Kretzmann N. Kretzmann B.E. (eds.), Oxford/New York.
- Kretzmann, Norman and Barbara Ensign Kretzmann (eds.): 1990, The Sophismata of Richard Kilvington Introduction, Translation, and Commentary, Cambridge.

Secondary Literature

- Bottin, Francesco, 1973a: “Analisi linguistica e fisica Aristotelica nei ‘Sophysmata’ di Ricard Kilmyngton”, in C. Giacon (ed.), *Filosofia e Politica, et altri sagii*, Padua, 125-45.
- -----, 1973b: “L'*Opinio de Insolubilibus* di Richard Kilmyngon”, *Rivista critica di Storia della Filosofia* 28, 409-22.
- -----, 1974 “Un testo fondamentale nell'ambito della ‘nuova fisica’ di Oxford: I Sophismata di Richard Kilmington”, *Miscellanea Medievalia*, 9, 201-205
- Caroti, Stefano, 1995: “Da Walter Burley al. ‘Tractatus sex inconvenientium’: La Tradizione inglese della discussione medievale ‘De reactione’”, *Medioevo* 21, 279-304.
- Jung-Palczewska, Elzbieta, 1997: “Motion in a Vacuum and in a Plenum in Richard Kilvington's Question: *Utrum aliquod corpus simplex posset moveri aeque velociter in vacuo et in pleno* from the ‘Commentary on the Physics’”, *Miscellanea Medievalia*, Bd. 25 179-193.
- ----- 2000a: “The Concept of Time in Richard Kilvington”, in *Tempus, Aevum, Eternity*. *La Conzettualizzazione del tempo nel Pensiero Tardomiedievale*, L. Cova, G. Alliney (eds.) Firenze, 141-167.
- -----, 2000b: “Works by Richard Kilvington”, *Archives d'Histoire Doctrinale et Littéraire du Moyen Age*, 67, 181-223.
- Katz, Bernard, D., 1996: “On a *Sophisma* of Richard Kilvington and a Problem of Analysis”, *Medieval Philosophy and Theology*, 5, 31-38.
- Knuuttila, Simo and Anja Inkeri Lehtinen, 1979: “*Plato in infinitum remisse incipit esse albus*: New Texts on the Late Medieval Discussion on the Concept of Infinity in Sophismata Literature”, in E. Saarinen, R. Hilpinen, I. Niiniluoto, M. B. P. Hintikka (eds.), *Essays in Honor of Jaakko Hintikka*, Dordrecht, 309-29.
- Kretzmann, Norman, 1977: “Socrates is Whiter than Plato Begins to be White”, *Nous*, 11, 3- 15.

- -----, 1982: “Richard Kilvington and the Logic of Instantaneous Speed”, in A. Maieru, A. Paravicini-Bagliani (eds.), *Studi sul secolo in memoria di Annelise Maier* (Edizioni di Storia e Letteratura), Rome, 142-75.
- -----, 1988: “‘Tu scis hoc esse omne quod est hoc’: Richard Kilvington and the Logic of Knowledge”, in Kretzmann (ed.), *Meaning and Inference in Medieval Philosophy*, Dotrecht, 225-45.
- Longeway, John, 1984: *William Heytesbury on Maxima and Minima* (Translation with Introduction and Study), Dordrecht.
- Murdoch, John E., and Edith Dudley Sylla, 1976: “The Science of Motion”, in D.C. Lindberg (ed.) *Science in the Middle Ages*, 206-64.
- D'Ors, A., 1991: “‘Tu scis regem sedere’ Kilvington S 47, 48”, *Anuario Filosofico* 24, 49-74.
- Stump, Eleonore, 1982: “Obligations: From the Beginnings to the Early Fourteenth Century”, in N. Kretzmann et al. (eds.), *The Cambridge History of Later Medieval Philosophy*, 315-34.
- Sylla, Edith Dudley, 1982a: “The Oxford Calculators”, in N. Kretzmann, et al. (eds.), The Cambridge History of Later Medieval Philosophy, 540-63.
- ----, 1982b: “Infinite Indivisibles and Continuity in Fourteenth-Century Theories of Alteration”, in N. Kretzmann (ed.)., *Infinity and Continuity in Ancient and Medieval Thought*, Ithaca N.Y., 231-57.
- -----, 1991: *The Oxford Calculators and the Mathematics of Motion 1320-1350. Physics and Measurement by Latitudes*, New York & London, 435-446.

Related Entries

Albert of Saxony | Bradwardine, Thomas | Burley [Burleigh], Walter | Heytesbury, William | Ockham [Occam], William

Elzbieta Jung-Palczewska
University of Lodz
palczews@krysia.uni.lodz.pl

First published: August 6, 2001
Content last modified: August 6, 2001

Albert of Saxony

Albert of Saxony (ca. 1316-1390), Master of Arts at Paris, then Rector of the University of Vienna, and finally Bishop of Halberstadt (Germany). As a logician, he was at the forefront of the movement that expanded the analysis of language based on the properties of terms, especially their reference (in Latin: *suppositio*), but also in the exploration of new fields of logic, especially the theory of consequences. As a natural philosopher, he worked in the tradition of John Buridan, and contributed to the spread of Parisian natural philosophy throughout Italy and central Europe.

---

1. Life and Works

Albert of Saxony (*Albertus de Saxonia*), whose family name was Albert of Ricmerstop or Rickmersdorf, is sometimes called *Albertucius* (Little Albert), to distinguish him from the 13th-century theologian Albert the Great. He was born at Helmstedt in present-day Germany around 1316. After initial schooling in the region of Helmstedt, and possibly a sojourn at Erfurt, he made his way to Prague and then on to Paris, where he became a master of arts in 1351. He was Rector of the University of Paris in 1353. He remained in Paris until 1362, during which time he taught arts and studied theology at the Sorbonne, apparently without obtaining any degree in the latter discipline. His logical and philosophical works were composed during this period. After two years of apparently carrying out diplomatic missions between the Pope and the Duke of Austria, he was charged with founding the University of Vienna, of which he became the first Rector in 1365. Appointed canon of Hildesheim in 1366, he was also named Bishop of Halberstadt the same year, fulfilling that office until his death, July 8, 1390.

Not having left any theological writings or commentary on Aristotle's *Metaphysics* (at least none that we

know of), Albert is primarily known for his works on logic and natural philosophy. He also wrote commentaries on Aristotle's *Nicomachean Ethics* and *Economics*, as well as several short mathematical texts (the "Treatise on Proportions" and "Question on the Squaring of the Circle").

Albert's masterwork in logic is a *summa* entitled the *Perutilis logica* (Very Useful Logic). He also composed a voluminous collection of *Sophismata*, in which he examines numerous sentences raising difficulties of interpretation due to the presence of syncategorematic words (i.e., terms such as quantifiers and certain prepositions, which, according to medieval logicians, do not themselves have a proper and determinate signification, but which modify the signification of other terms in a proposition). He also wrote several question commentaries: *Quaestiones* on the *Ars Vetus* or Old Logic (the *Isagoge* of Porphyry and Aristotle's *Categories* and *De interpretatione*), *Quaestiones* and a Commentary on the *Posterior Analytics*, and a series of 25 *Quaestiones logicales* (Logical Questions), addressed to semantic problems and the status of logic. Of dubious authenticity are commentaries on the *Prior Analytics* (both literal and question commentaries), as well as treatises *De consequentiis* (On Consequences) and *De locis dialecticis* (On Dialectical Topics), which have been attributed to him in a Parisian manuscript.

The most renowned philosopher during the era when Albert studied and taught in the Faculty of Arts at Paris was John Buridan. Albert belonged to the first generation of masters who, in one form or another, carried on the tradition of Buridan in logic and natural philosophy. Accordingly, his commentary on the *Posterior Analytics* to a large extent reprises the work of Buridan. At the same time, however, Albert's own work, especially in logic, testifies to the influence of certain ideas and methods developed in England, particularly by William of Ockham, William Heytesbury -- as is apparent in Albert's *Sophismata* -- and Thomas Bradwardine, on the study of movement. Walter Burley was another important influence on Albert. This influence is paradoxical in view of the fact that they had opposing views on the nature of universals, but it is nonetheless in evidence in his commentary on the *Nicomachean Ethics*, and also noticeable in his theory of consequences. Finally, Albert both knew and discussed certain theses of Thomas Maulfelt, who had taught at Paris around 1330. These various influences have sometimes made Albert seem an eclectic compiler -- and not without reason. But, besides making possible several of Albert's own contributions, Albert's eclecticism confers upon him a unique place in the development of logic and philosophy at the University of Paris in the 14th century.

2. Logic

In most respects, the *Perutilis logica* exhibits the influence of Ockham's *Summa logicae*, although it develops in a more autonomous fashion the treatises on obligations, insolubles, and consequences, which had assumed greater importance during this period. As has been known for some time, this work is a remarkable handbook organized into six treatises: the first defines the elements of propositions; the second treats of the properties of terms; the third of the truth conditions of different types of proposition; the fourth of consequences (including syllogisms, and in fact adding to it the theory of topics); the fifth of fallacies; and the sixth of insolubles and obligations.

In the first part of the *Perutilis logica,* which sets out the terminology of the entire text, Albert returns to

the Ockhamist conception of the sign, and so distances himself from the position defended by Buridan. After clearly including the term (an element of the proposition) in the genus of signs -- and thereby providing, in the tradition of Ockham, a semiotic approach to logico-linguistic analysis -- he establishes signification through a referential relation to a singular thing, defining the relation of spoken to conceptual signs as a relation of subordination. He is also Ockhamist in his conception of universals, which he regards as spoken or conceptual signs, and in his theory of supposition, which essentially restates the Ockhamist divisions of supposition. In particular, he preserves the notion of simple supposition -- i.e., the reference of a term to the concept to which it is subordinated, when it signifies an extra-mental thing -- which had been challenged and criticized by Buridan. Finally, Albert is close to the *Venerabilis Inceptor* in his theory of the categories, where, in contrast to Buridan, he refuses to consider quantity as something absolutely real, reducing it instead to a disposition of substance and quality. In doing so, he contributed as much as Ockham to the spread of this model of the relation between substance and quantity in natural philosophy in Paris and Italy.

Albert's treatment of relations is, on the other hand, highly original. Although (like Ockham) he refuses to make relations into things distinct from absolute entities, he clearly ascribes them to an act of the soul by which absolute entities are compared and placed in relation to each other (an act of the referring soul [*actus animae referentis*]). This leads him to reject completely certain propositions Ockham had admitted as reasonable, even if he did not construe them in quite the same way, e.g., ‘Socrates is a relation’. Both Ockham and Buridan had allowed that the term ‘relation’ could refer to the things related (whether connoted or signified) by concrete relative terms (whether collectively or not).

So Albert was not content with merely repeating Ockhamist arguments. More often than not, he developed and deepened them, e.g., in connection with the notion of the appellation of form. This property of predicates, which had previously been used by the *Venerabilis Inceptor*, was employed by Albert in an original manner when he turned to it, in place of Buridan's appellation of reason (*appellatio rationis*), to analyze verbs expressing propositional attitudes. Every proposition following a verb such as ‘believe’ or ‘know’ appellates its form. In other words, it must be possible to designate the object of the belief via the expression understood as identical to itself in its material signification, and without reformulation. Another area in which Albert deviates from Ockham is in his rejection of the idea that any distinction with multiple senses must have an equivocal proposition as its object. According to Albert, equivocal propositions can only be conceded, rejected, or left in doubt.

Albert's semantics becomes innovative when he admits that propositions have their own proper significate, which is not identical to that of their terms (see especially his *Questions on the Posterior Analytics* I, qq. 2, 7, 33). Like syncategorematic terms (see his *Questions on the Categories*, qu. 1 ‘On Names’), propositions signify the “mode of a thing [*modus rei*]”. Nevertheless, Albert avoids hypostatizing these modes, in the final analysis explaining them as relations between the things to which the terms refer. It cannot be said here that Albert is moving towards the “complexly signifiable [*complexe significabile*]” of Gregory of Rimini, although his remarks are reminiscent of the latter theory. Still, he uses the idea of the signification of a proposition to define truth and to explain ‘insolubles’, i.e., propositions expressing paradoxes of self-reference. On Albert's view, every proposition signifies that it is true by virtue of its form. Thus, an insoluble proposition will be false because it signifies at the same

time that it is true and that it is false.

In his *Sophismata*, Albert usually follows Heytesbury. The distinction between compounded and divided senses, which is presented in a highly systematic way in Heytesbury's *Tractatus de sensu composito et diviso*, is the primary instrument (besides the appellation of form) used to resolve difficulties connected with epistemic verbs, and with propositional attitudes more generally. This is abundantly clear in his discussion of infinity. Rather than appealing to the increasingly common distinction between the categorematic and syncategorematic uses of the term ‘infinite’, and then indicating the different senses it can have depending upon where it occurs in a proposition, he treats the infinite itself as a term. Albert's approach involves giving the analysis of the logical and linguistic conditions of every proposition involving the term ‘infinite’ that is significant and capable of being true. This leads him to sketch a certain number of possible definitions (where he appears to take into account the definitions of Gregory of Rimini), but also to raise other questions, e.g., on the relation between finite and infinite beings (in propositions such as ‘infinite things are finite [*infinita sunt finita*]’), on the divisibility of the continuum, and on qualitative infinity. There are echoes in Albert not only of the approach Buridan had, for his part, systematically implemented in his *Physics*, but also of the analyses of English authors -- again, especially Heytesbury. As is often the case, the treatment proposed by Albert in the *Sophismata* is quick and a little eclectic, but it provides good evidence of the extent to which questions about the infinite were discussed at the time.

Finally, one of the fields in which Albert is considered a major contributor is the theory of consequences. In the treatise of the *Perutilis Logica* devoted to consequences, Albert often seems to follow Buridan. But whereas Buridan maintained the central role of Aristotelian syllogistic, Albert, like Burley, integrated syllogistic and the study of conversions into the theory of consequences. Consequence is defined as the impossibility of the antecedent's being true without the consequent's also being true -- truth itself being such that howsoever the proposition signifies things to be, so they are. The primary division is between formal and material consequences, the latter being subdivided into consequences *simpliciter* and *ut nunc*. A syllogistic consequence is a formal consequence whose antecedent is a conjunction of two quantified propositions and whose consequent is a third quantified proposition. Albert is thus led to present a highly systematized theory of the forms of inference, which represents a major step forward in the medieval theory of logical deduction.

3. Natural Philosophy

It is this analysis of language together with a particularist ontology that places Albert in the tradition of nominalism. This is combined with an epistemological realism that emerges, e.g., in his analysis of the vacuum. In certain respects, Albert's work is a proper extension of physical analysis to imaginary cases. Distinguishing, as Buridan did, between what is absolutely impossible (i.e., the contradictory) and what is impossible “in the common course of nature” (*Questions on De Caelo* I, qu. 15), he considers hypotheses under circumstances which are not naturally possible but imaginable given God's absolute power (e.g., the existence of a vacuum and the plurality of worlds). However, even if we can imagine a vacuum existing by divine omnipotence, no vacuum can occur naturally (*Questions on the Physics* IV,

qu. 8). Albert refuses to extend the reference of physical terms to supernatural, purely imaginary possibilities. In the same way, one can certainly use the concept of a point, although this would only be an abbreviation of a connotative and negative expression. There is no simple concept of a point, a vacuum, or the infinite, and although imaginary hypotheses provide an interesting detour, physics must in the end provide an account of the natural order of things.

Historically, Albert does not enjoy the reputation in natural philosophy that he does in logic. His commentaries on the *Physics* and on the treatise *De caelo* are close to Buridan's, and Albert appeals to the authority of his “revered masters from the Faculty of Arts at Paris” at the beginning of his questions on *De caelo*. Even so, it should be noted that his *Physics* was written before the final version of Buridan's *Questions on the Physics* (between 1355 and 1358), which means that he could not have benefited from the final version of Buridan's lectures.

We have already seen that on the question of the status of the category of quantity, then at the forefront of logic and physics, Albert followed Ockham and distanced himself from Buridan by reducing quantity to a disposition of substance or quality. This move becomes evident in certain physical questions, e.g., in the study of condensation and rarefaction, where Albert openly disagrees with his Parisian master by arguing that condensation and rarefaction are possible only through the local motion of the parts of a body, and without needing to assume some quantity that would have a distinct reality on its own. Nevertheless, he defines the concept of a “lump of matter [*materie massa*]” without giving it any autonomous reality, although it does help fill in the idea of a ‘quantity of matter’, which Giles of Rome had already distinguished from simple extension.

Similarly, Albert is sometimes seen as standing alongside Ockham on the nature of motion, rejecting the idea of motion as a flux (*fluxus*), which is the position Buridan had adopted. In contrast to his master, Albert treats locomotion in the same way as alteration (movement according to quality): in neither case is it necessary to imagine local motion as a *res successiva* distinct from permanent things, at least if the common course of nature holds and one does not take into account the possibility of divine intervention.

As far as the general principles of motion are concerned, Albert, like Nicole Oresme, follows Thomas Bradwardine on the relation of motive force and resistance. On the other hand, when he considers the movement of projectiles, gravitational acceleration, and the movement of celestial bodies, Albert adopts Buridan's major innovation, i.e., the theory of *impetus*, a quality acquired by a moving body (see Buridan's *Questions on the Physics* VIII, qu. 13, on projectile motion). Like Buridan, he extends this approach to celestial bodies in his commentary on *De caelo*, clearly following its consequences in rejecting intelligences as agents of motion and in treating celestial and terrestrial bodies using the same principles. Nevertheless, he formulates the idea of *impetus* in more classical terms as a *virtus impressa* (impressed force) and *virtus motiva* (motive force). Albert makes no pronouncements about the nature of this force. This is a question for the metaphysician. His work also mentions the mean speed theorem, a method of finding the total velocity of a uniformly accelerated (or decelerated) body, which had been stated (without being demonstrated) in Heytesbury's *Tractatus de motu*, and also adopted by Nicole Oresme. Albert was part of the general scientific trend which sought the first formulations of the principles of dynamics.

As for his commentaries on Aristotle's works of natural philosophy, Albert wrote a *Treatise on Proportions* that was devoted to the analysis of movement. This was very much inspired by Bradwardine's treatise *De proportionibus velocitatum in motibus*, and extended the work he and others had begun at Oxford. Albert's texts on the kinematic measure of both rectilinear and circular motion enjoyed a broad popularity. Finally, he explained a number of curious natural phenomena, taking particular interest in earthquakes, tidal phenomena, and geology.

Like his master Buridan, Albert was interested in certain mathematical problems. To this end, he wrote a question on the squaring of the circle as well as questions on John of Sacrobosco's *Treatise on the Sphere*. In addition to authoritative arguments and purely empirical justifications, his question on the squaring of the circle uses properly mathematical arguments that appeal to both Euclid (in the version of Campanus of Novarra) and Archimedes (translated by Gerard of Cremona). His most original contribution consists in a proposal to dispense with Euclid's proposition X.1, replacing it with a postulate stating that if A is less than B, then there exists a quantity C such that $A<B<C$.

4. Impact and Influence

Albert of Saxony's teachings on logic and metaphysics were extremely influential. Although Buridan remained the predominant figure in logic, Albert's *Perutilis logica* was destined to serve as a popular text because of its systematic form and also because of the fact that it takes up and develops essential aspects of the Ockhamist position. But it was his commentary on Aristotle's *Physics* that was especially widely read. Many manuscripts of it can be found in France and Italy, in Erfurt and Prague. Albert's *Physics*, more than Oresme's and even Buridan's, basically guaranteed the transmission of the Parisian tradition in Italy, where it was used alongside the works of Heytesbury and John Dumbleton. His commentary on Aristotle's *De caelo* was also influential, eventually eclipsing Buridan's commentary on this text. Blasius of Parma read it in Bologna between 1379 and 1382. A little later, it enjoyed a wide audience at Vienna.

Albert played an essential role in the diffusion throughout Italy and central Europe of Parisian ideas which bore the mark of Buridan's teachings, but which were also clearly shaped by Albert's comprehensive grasp of English innovations. At the same time, Albert was not merely a compiler of the work of others. He knew how to construct proofs of undeniable originality on many topics in both logic and physics.

Bibliography

Primary Texts

- A. Muñoz García, 1990: “Albert of Saxony, Bibliography”, *Bulletin de Philosophie médiévale* 32, pp.161-190. [complete listing of texts, manuscripts, and editions]

- _____, 1991: “Cinco nuevos fragmentos anónimos de Alberto de Sajonia”, *Bulletin de philosophie médiévale* 33, pp.162-176.
- *Perutilis logica*, in the incunabular edition of Venice 1522, with a Spanish translation by A. Muñoz García, Univ. del Zulia, Maracaibo, 1988.
- *Perutilis logica*, Tractatus Secundus (De proprietatibus terminorum): cf. *infra*, Kann: 1993.
- *Quaestiones in artem veterem*, ed. A. Muñoz García, Univ. del Zulia, Maracaibo, 1988

Selected Studies and Critical Discussions

- Berger, Harald, 1994: “Albert von Sachsen (1316? -1390). Bibliographie der Sekundärlitteratur”, in *Bulletin de Philosophie médiévale* 36, pp. 148-185. [exhaustive listing of the secondary literature]
- Biard, Joël, 1989: “Les sophismes du savoir: Albert de Saxe entre Jean Buridan et Guillaume Heytesbury”, *Vivarium* XXVII, pp. 36-50.
- _____, (éd.), 1991: *Paris-Vienne au XIV^e siècle. Itinéraires d'Albert de Saxe* (Actes de la table ronde internationale, Paris, 19-22 juin 1990), Vrin, Paris [21 articles representing the state of research on Albert's logic and natural philosophy]
- _____, 1993: “Albert de Saxe et les sophismes de l'infini”, in Stephen Read (ed.), *Sophisms in Medieval Logic and Grammar*, Kluwer, Dordrecht-Boston-London, pp. 288-303.
- Drake, Stillman, 1975: “Free Fall from Albert of Saxony to Honoré Fabri”, *Studies in History and Philosophy of Science* 5 (4), pp. 347-366.
- Gonzales, A., 1958: “The Theory of Assertoric Consequences in Albert of Saxony”, *Franciscan Studies* XVIII, pp. 290-354; XIX, pp. 13-114.
- Heidingsfelder, G., 1927: *Albert von Sachsen. Sein Lebensgang und sein Kommentar zur Nikomachischen Ethik des Aristoteles*, in Beiträge zur Geschichte der Philosophie des Mittelalters XXII/3-4, Münster.
- Kann, Christoph, 1993a: “Die Behandlung der dialektischen Örter bei Albert von Sachsen” in Klaus Jakobi (ed.), *Argumentationsheorie. Schoslastischen Forschungen zu den logischen und semantischen Regeln korrekten Folgerns*, Brill, Leiden-New York-Köln.
- _____, 1993b: *Die Eigenschaften der Termini*. *Eine Untersuchung zur ‘Perutilis Logica’ des Alberts von Sachsen*, Brill, Amsterdam. [study of the theory of the property of terms, including the theory of supposition, with an edition of the second treatise of the *Perutilis logica*]
- Sarnowsky, Jürgen, 1989: *Die aristotelisch-scholastische Theorie der Bewegung. Studien zum Kommentar Alberts von Sachsen zur Physik des Aristoteles*, in Beiträge zur Geschichte der Philosophie und Theologie des Mittelalters, N. F. XXXII, Aschendorff, Münster.

Other Internet Resources

[Please contact the author with suggestion.]

Related Entries

Bradwardine, Thomas | Burley [Burleigh], Walter | Heytesbury, William | Ockham [Occam], William

Acknowledgements

The author gratefully acknowledges Jack Zupko for translating this entry into English.

Joël Biard
Centre d'Études Supérieures de la Renaissance
Université de Tours
jbiard@univ-tours.fr

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 29, 2001
Content last modified: January 29, 2001

Richard the Sophister

Richard the Sophister (Richardus Sophista) was an English philosopher/logician who studied at Oxford most likely sometime during the second quarter of the thirteenth century. Richard's identity is uncertain, but he is known to be the author of a collection of logically puzzling sentences, sometimes called "sophisms", entitled *Abstractiones*. The puzzling aspect of these sophisms is variously caused by semantic or syntactic ambiguities involved in certain logical or "syncategorematic" words such as "all", "every", "or", "if...then", "and", "not", "begins", "ceases", "except", "necessary", "possible" etc.

Abstractiones

The title "*Abstractiones*" was apparently meant to capture the "excerpted" nature of this collection of sophisms that are presented in a somewhat summary fashion.[1] This collection of sophisms became a kind of logical textbook used to teach students to identify sophistical fallacies. This book was probably used in several medieval universities for several decades, from the late thirteenth well into the fourteenth centuries. Richard is often referred to as "Magister Abstractionum" or "Magister Richardus Sophista" to give emphasis to the masterful, albeit summary, treatment of a very large collection of over three hundred such sophism sentences.[2] The title "*sophista*" simply meant "logician".

The seven known manuscripts of the *Abstractiones* date from the late thirteenth to the early fourteenth centuries:[3]

Ms B = Brugge, Stedelike Bibliotheek, 497
Ms C = Oxford, Corpus Christi College, E 293B
Ms D = Oxford, Bodleian Library, Digby 24

Ms K = København, Det Kongelige Bibliotek, Fragm.1075
Ms O = Oxford, Bodleian Library, Digby 2
Ms P = Paris, Bibliothèque Nationale lat. 14069
Ms R = London, British Library, Royal 12.F.xix

Only two of these manuscripts (B, D) are complete; the remaining five are in varying degrees of fragmentation. Two manuscripts (O, R) appear to be derivative texts, or redactions, of the *Abstractiones*.

The dating of the composition of the *Abstractiones* appears to have been around the 1230's or 40's. Earliest references to the *Abstractiones* appear from the late thirteenth century in a manuscript at Worcester Cathedral Library, Q. 13, which is confidently dated no later than 1295, and probably as early as 1270.[4] Worcester Q. 13 contains four references to the "Magister Abstractionum". Another reference to the *Abstractiones* can be found in a Quodlibetal Question of William of Alnwick, "Does God know infinitely many things (*Utrum Deus cognoscat infinita)*?", q. 9 probably dating from around 1320. Walter Burley's tract, *De Exceptivis*, edited by L. M. de Rijk, contains a reference to the Magister Abstractionum.[5] An interesting reference can also be found in William of Ockham's *Summa Logicae* and additional references can be found in Ps. Richard of Campsall, John of Reading and Adam Wodeham.[6]

The fact of the existence of two derivative texts of the *Abstractiones* as well as the numerous known references by accomplished logicians some seventy years after its composition indicate that the *Abstractiones* was a quite popular book that survived the test of time. Although as compared to sophism literature of the fourteenth century Richard's analysis of sophisms appears somewhat unsophisticated, as compared to treatments of sophisms of his contemporaries Richard's analyses are qualitatively similar. Richard's greatest accomplishment lay, perhaps, in his ability to encapsulate and summarize a huge number of such logical puzzles, presenting for the student a handy reference book filled with examples of a host of logical fallacies.

Examples of Sophisms

In order to give the reader a better understanding of the content of this book of *Abstractiones*, it might be helpful to list a few examples of the sophism sentences discussed in it and to illustrate Richard's treatment of just one of these. As noted above, our book contains over three hundred sophism sentences. A random sampling includes these examples:

(1) Every man is every man (*Omnis homo est omnis homo*)

(16) Whatever exists or does not exist exists (*Quicquid est vel non est est*)

(39) Everything other than an animal which plus Socrates are two differs from Socrates (*Omne aliud quam animal quod et Sortes sunt duo differt a Sorte*)

> (79) The whole Socrates is less than Socrates (*Totus Sortes est minor Sorte*)
>
> (96) If you know that you are a stone you do not know that you are a stone (*Si tu scis te esse lapidem tu non scis te esse lapidem*)
>
> (157) Nothing is true about nothing (*De nihilo nihil est verum*)
>
> (187) You have not ceased to eat iron (*Tu non cessas comedere ferrum*)
>
> (215) All ten except one are nine (*Omnia X praeter unum sunt IX*)
>
> (278) It is impossible that you know more than you know (*Impossibile est te scire plura quam scis*)

From the above rather small list of sophisms, it can already be noticed that some sentences are more complex, and thus more logically puzzling, than others. Some sentences do not seem puzzling in themselves at all, but become puzzling only within a certain "context" that is set up within which arguments proving and disproving the sophism must be developed. Such a context was called a "*casus*".[7] It seems likely that the proving and disproving of the sophism sentence was done along the lines of a medieval oral "*disputatio*" within the classroom setting. This is the way Richard writes up his summaries: Proposal of the sophism, sometimes within a *casus*; arguments establishing the truth of the sentence; arguments establishing the falsity of the sentence; resolution of the sophism along with identification of the sophistical fallacy involved.

The very first sophism sentence treated is one of the least complicated (this sophism involves no *casus*):

> Every man is every man
>
> Proof: This man is this man, that man is that man and so forth for each individual man, therefore every man is every man.
>
> Another proof: No proposition is more true than one in which the same thing is predicated of the same thing....
>
> Disproof: This proposition is true, "some man is not every man" and, similarly, "no man is every man"; therefore it is false that every man is every man.
>
> Solution: The solution to the sophism is to note that "every man" is equivocal as to reference to every man taken singularly (as in "any man") and as to every man taken as a whole (as in "all men"). Mixing the two senses of "every" from subject to predicate causes the logical puzzle.

As noted, this sophism sentence is relatively uncomplicated and it is easy to spot the obvious fallacy involved, but Richard's text proceeds with increasingly more difficult and intricate sophism sentences, many of which engage the mind in complex mental acrobatics. The fallacies noted throughout are the standard ones discussed in Aristotle's *De Sophisticis Elenchis*: the fallacy of equivocation; the fallacy of accident; the fallacy of the composite and divided senses; the fallacy of the consequent; the fallacy of absolute and qualified senses; the fallacy of many causes of truth; amphiboly; improper supposition.

Richard's Identity

There is still no certainty as to the identity of the Magister Abstractionum. Because of the colophon appended to the two complete manuscripts of the *Abstractiones*, we know, of course, that his name was "Richard":

> *Expliciunt ista, quae tu, Ricarde Sophista,*
> *fecesti, morum flos et doctor logicorum.*
> *Dirige scribentis, Spiritus alme, manum.*
> *Expliciunt Abstractiones*.
> (Digby 24, f.90rb)
>
> (These [sophisms] are complete, which you, Richard the Sophister,
> Flower of virtue and teacher of logic, have produced.
> Turn the scribe's hand [from its task], nurturing Spirit.
> The *Abstractiones* are complete.)

De Rijk suggested the name Richard Fishacre, disagreeing with the suggestion of Richard Fitz-Ralph offered by Macray on the basis that the dating of the text within the second quarter of the thirteenth century is more consistent with Fishacre's chronology.[8] Jan Pinborg suggested the name Richard Rufus also on grounds that the dating of the *Abstractiones* is consistent with Rufus' being at Oxford.[9] It is generally assumed that the author is English. Pinborg offered one further bit of evidence for the name Richard Rufus in that there is some reason to think that certain doctrines of Richard Rufus, as criticised by Roger Bacon, are in the *Abstractiones*. The primary doctrine in question attributed to Richard Rufus and criticised by Roger Bacon is, in general terms, that the signification of a name can remain in the absence of any actual thing signified by that name, although, as Bacon suggests, the proponents of this view must supply some kind of "habitual being" for the lost actual significate of such names. It is this doctrine of the "*esse habituale*" that Bacon finds objectionable. That Richard Rufus does seem to have such a doctrine appears clear in his discussion of the question "Whether Christ while three days in the tomb was a man" ("*Utrum Christus in triduo mortis fuerit homo*") in distinction 22, book III of his *Sentences Commentary*.[10]

It would be beyond the scope of this essay to deal with the complexities of both Richard Rufus' treatment

of this question as well as Roger Bacon's voiceful criticisms of the notion of the "*esse habituale*". The important question here is whether the doctrine espoused by Richard Rufus exactly parallels anything that can be found in the *Abstractiones* of Richard the Sophister. It is important to find not only terminological similarities but also doctrinal similarities. The terminology within which Richard Rufus makes the above noted distinction is "*esse in habitu*" versus "*esse in actu*", the former he also labels "*esse simpliciter*", the latter "*esse ut nunc*".[11] Bacon seems to substitute the expression "*esse habituale*" for the expression "*esse in habitu*". Rufus does not use the expression "*esse habituale*". As referenced above, Pinborg noticed the expressions ",i>esse consequentiae sive habitudinis" and "*esse quod est operatio entis*" in the Abstractiones. Bacon also uses the expression "*esse habitudinis*", but he seems to recognize a distinction between "*esse habituale*" and "*esse habitudinis*"; the latter, he says, "is used about propositions and will be destroyed later when there is talk about propositions".[12] This remark might indicate that Bacon finds the same doctrinal problem with the "*esse habitudinis*" that he finds with the "*esse habituale*" (i.e., it introduces a foil for some kind of fictive being), but we cannot be certain of this, since he never returned to this topic in the Compendium. Richardus Sophista does not use the terminology of "*esse habituale*/*esse actuale*", "*esse in habitu/esse in actu*" that we find in Bacon and Rufus respectively.[13] Pinborg also noted that William of Ockham criticizes those who make the distinction between "*esse consequentiae sive habitudinis*" and "*esse quod est operatio entis*" as ignorant of the simple distinction between hypothetical and categorical propositions, which according to Ockham is all that the distincition amounts to.[14] Pinborg sees Ockham's criticism as possibly making the same point as Bacon, i.e., the terminolgy invites some kind of talk of fictive being. Although Ockham is probably right to want to eliminate the old terminology in favor of less problematic language, the important questions are whether Richard the Sophister's use of the distinction "*esse habitudinis*/*esse quod est operatio entis*" entails anything other than a distinction between types of propositions and whether his use of the distinction comes under the attacks of Roger Bacon on the same way in which these attacks seem relevant to a terminologically similar distinction in Richard Rufus.

A complete answer to these questions would require not only a detailed examination of the use of this distinction in the resolution of sophisms in the *Abstractiones*, but also an examination of Richard's position regarding terms that "assert non-being" and his account of negation and negatives generally.[15] These tasks are clearly beyond the scope of this essay. With respect to the first question, it will have to suffice to record here that there are only three places in the *Abstractiones* where Richard employs the distinction in the resolution of sophisms and in none of these is it clear that this is his favored solution.[16] More importantly, Richard's use of the distinction in the resolution of sophisms seems to match the use made of the same distinction (under the terminology "*esse habituale*/*esse actuale*") by William of Sherwood, that is, as essentially a way of marking a distinction between hypothetical and categorical propositions.[17] It has been successfully argued by Braakhuis that the attacks of Roger Bacon do not apply to William's use of the distinction, even though, of all the authors noted so far, only William's terminology is exactly the same terminology used by Bacon.[18] Thus, if the doctrine under attack by Bacon cannot be attributed to William, neither can it be attributed to Richard the Sophister. It appears, then, that Richard Rufus and Richard the Sophister cannot be identified, at least on the basis of these alleged doctrinal and terminological similarities.

Bibliography

- Braakhuis, H.A.G.: 1981, "English Tracts on Syncategorematic Terms from Robert Bacon to Walter Burley" in *English Logic and Semantics*, Artistarium, Supplementa 1, Nijmegen.
- Ebbesen, Sten: 1987, "Talking about what is no more. Texts by Peter of Cornwall, Richard of Clive, Simon of Faversham and Radulphus Brito," *Cahiers de l'Institut du Moyen-Âge Grec et Latin* 55, Copenhague.
- Kopp, Clemens: 1985, *Die "Fallaciae ad modum Oxoniae," Ein Fehlschlußtraktat aus dem 13 Jahrhundert*, diss., Köln.
- Kretzmann, Norman, et al., (eds.): 1982, *The Cambridge History of Later Medieval Philosophy*, Cambridge.
- Lewry, P.O. (ed.): 1985, *The Rise of British Logic: Acts of the Sixth European Symposium on Medieval Logic and Semantics* , Papers in Mediaeval Studies 7, Pontifical Institute of Mediaeval Studies, Toronto.
- de Libera, Alain: 1986, "Les *Abstractiones* d'Herve le Sophiste (Hervaeus Sophista)," *Archives d'Histoire Doctrinale et Littéraire du Moyen Âge*, Paris, 163-230.
- de Libera, Alain, "La Littérature de *Abstractiones* et la Tradition Logique d'Oxford" in Lewry 1985.
- Macray, G.D.: 1883, *Catalogi Codicum Manuscriptorum Bibliotecae Bodleianae*, Pars Nona, Oxford.
- O'Donnell, J.R.: 1941, "The Syncategoremata of William of Sherwood," *Mediaeval Studies* 3, 46-93.
- Pinborg, Jan: 1976, "Magister Abstractionum," *Cahiers de l'Institut du Moyen-Âge Grec et Latin* 18, Copenhague, 1-4.
- Raedts, Peter: 1987, *Richard Rufus of Cornwall and the Tradition of Oxford Theology*, Oxford.
- Read, Stephen (ed.): 1993, *Sophisms in Medieval Logic and Grammar*, Acts of the 8th European Symposium for Medieval Logic and Semantics, Kluwer.
- de Rijk, L. M.: 1962-67, *Logica Modernorum*, I-II, van Gorcum, Assen.
- de Rijk, L.M.: 1974, "Some Thirteenth Century Tracts on the Game of Obligation," *Vivarium* 12.
- de Rijk, L.M.: 1985, "Walter Burley's Tract '*de Exclusivis*'. An Edition," *Vivarium* 23, 23-54.
- de Rijk, L.M.: 1986, "Walter Burley's Tract '*de Exceptivis*'. An Edition," *Vivarium* 24, 22-49.
- Spade, Paul Vincent, "Obligations: Developments in the Fourteenth Century" in Kretzmann, et al. 1982.
- Stump, Eleonore, "Obligations: From the Beginning to the Early Fourteenth Century" in Kretzmann, et al. 1982.
- Streveler, Paul A., "A Comparative Analysis of the Treatment of Sophisms in MSS Digby 2 and Royal 12 of the *Magister Abstractionum*," in Read 1993.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Burley [Burleigh], Walter | Ockham [Occam], William | Simon of Faversham | William of Sherwood | Wodeham, Adam

Paul Streveler
West Chester University
pstreveler@wcupa.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 4, 1999
Content last modified: February 26, 2001

Stanford Encyclopedia of Philosophy
Notes to Richard the Sophister

Notes

[1.] In Albert of Saxony's introductory remarks regarding the exceptive term "*tantum*" in his *Perutilis Logica* we read this definition of the term "*abstractio*": "*Abstractio est dubitabilium propositionum collectio* (An abstraction is a collection of puzzling propositions)" (Erfurt ms., Amploniana, Cod. Q 242, f. 33v). This section of Albert's treatise is not edited in the 1518 Venice edition. The only other known treatise from the Middle Ages with the title "*Abstractiones*" is that of a certain Hervaeus Sophista, edited in de Libera 1986. The term "*abstractio*", then, seems to refer to a "collection" of sophisms, perhaps with some emphasis upon the "excerpted" nature of such a collection. Indeed, the collection of Hervaeus consists of some 298 sophisms within a space of a mere five folios. Richard considers some 305 sophisms within a space of almost thirty folios in the Digby 24 manuscript.

[2.] De Rijk 1962-67, Vol. II, pt 1., 62-85 and 444-447.

[3.] For a description of manuscripts B, D, O, P and R see de Rijk 1962-67. De Rijk was unaware of manuscripts C and K. For a brief description of the C manuscript, see P.O. Lewry, "Oxford Logic 1250-1275: Nicholas and Peter of Cornwall," in Lewry 1985, 22. For additional descriptions of these manuscripts, see Raedts 1987, 107-111.

[4.] Ebbesen 1987, 136 and Lewry 1985. The manuscript is dated 1295, but Lewry remarks that the author uses the example "*Henricus est rex Angliae*", and Henry died in 1272. Thus the reference to the Magister Abstractionum could be as early as the 1270's.

[5.] De Rijk 1986: 27.

[6.] William of Ockham, *Summa Logicae*, ed. Philotheus Boehner, Gideon Gál, and Stephen Brown (St. Bonaventure, N.Y. 1974), 367. Although this appears to be the only place where Ockham refers explicitly to the Magister Abstractionum, the editors point to other places where they believe he clearly has the Magister in mind (e.g., *Summa Logicae* III-1, c.16, 405; II, c. 4, 262). For references in Campsall, John of Reading and Adam Wodeham, see the Introduction to the *Summa Logicae*, 51*, n. 17.

[7.] The setting up of "contexts (*casus*)" within which proponents and opponents were obliged to develop their respective arguments gave rise to a special genre of sophism literature called "*obligationes*". We note the techniques of "*obligatio*" especially in the development of sophisms towards the end of the *Abstractiones*. For a discussion of the historical development of obligations, see Stump 1982 and Spade

1982.

[8.] De Rijk 1962-67, Vol. II, 71.

[9.] Pinborg 1976, 1-4.

[10.] The text of this question can be found in Franz Pelster, "Der Oxforder Theologie Richardus Rufus O.F.M. über die Frage, '*Utrum Christus in triduo mortis fuerit homo*'", *Recherches de Théologie Ancienne et Médiévale*, 16 (1949), 259-280. For Bacon's discussion, see Thomas S. Maloney (ed. & tr.), *Roger Bacon, Compendium of the Study of Theology*, Studien und Texte zur Geistesgeschichte des Mittelalters 20, (Leiden: E. J. Brill, 1988), par. 86-101.

[11.] Pelster 1949: p. 279, ll. 136-144.

[12.] Bacon, *Compendium*, trans. Maloney 1988: par. 103: "*Sed adhuc cavillant de esse habitudinis, sed hoc in propositione habet locum, et ideo destruetur postea, cum de propositionibus fiet sermo*". It is worth noting here that Maloney translates "*esse habituale*" and "*esse habitudinis*" identically as "habitual being", but there is warrant from Bacon himself that these are different notions and, perhaps, therefore, should be translated differently.

[13.] In connection with this terminology, Lewry (1985, p. 22) incorrectly transcribes C 283B at folio 207va as "*esse habituale sive commune*", (thus perhaps seeing the terminology found in Bacon's criticism of Rufus) whereas the scribe of the *Abstractiones* clearly writes "*esse habitudinis sive consequentiae*".

[14.] Ockham, (ed. Boehner et al.,) p. 263.

[15.] Especially relevant here would be a study of the sophism OMNIS PHOENIX EST, where Lewry (1985) p. 22, sees possible evidence of the Magister adopting notions criticized by Bacon.

[16.] These are in connection with the following sophisms: (1) OMNE COLORATUM EST: The distinction is introduced as a way of explaining a perceived mistake in one of the arguments in the disproof of the sophism, i.e., "*Omne coloratum est. Omne album est coloratum. Ergo omne album est*". It is said that this argument equivocates upon "*esse*", for in the first premiss it is "*esse operatio entis*", whereas in the second it is "*esse habitudinis sive consequentiae*". Richard goes on to interpret the former categorically, the latter hypothetically. Ockham's criticisms (noted above) seem relevant here in that equivocation is a semantic fallacy, whereas the propositional distinction is essentially syntactic. As Ockham notes, any proposition can be turned into one or the other. The important observation, however, is that Richard is not adopting the distinction as a favored solution to the sophism itself, but only as what some people say about a mistake in the disproof of an argument. (2) OMNIS HOMO DE NECESSITATE EST ANIMAL: Here too he offers multiple solutions to the sophism, the second of which is again to note the fallacy of equivocation in connection with "*esse*", as above. Here, however, it

seems clear that this is not his favored solution to the sophism. Rather he argues the sophism is best solved through application of the composite and divided senses of "*necessitas*". (3) OMNIS ASINUS EST HOMO: Here the distinction is offered again as an explanation of a kind of equivocation between two premises of the argument, as in: "*Omne animal est homo. Omnis asinus est animal*. *Ergo omnis asinus est homo*".

[17.] The distinction can be found in Norman Kretzmann (tr.), *William of Sherwood's Introduction to Logic*, (Minneapolis: University of Minnesota Press, 1966), 125-126, and Kretzmann (tr.), *William of Sherwood's Treatise on Syncategorematic Words*, (Minneapolis: University of Minnesota Press, 1968), 92-93.

[18.] Braakhuis 1981, 145-149.

Paul Streveler
West Chester University
pstreveler@wcupa.edu

First published: August 4, 1999
Content Last Modified: February 26, 2001

Medieval Theories of Properties of Terms

The theory of properties of terms (*proprietates terminorum*) was the basis of the medievals' semantic theory. It embraced those properties of linguistic expressions necessary to explain truth, fallacy and inference, the three central concepts of logical analysis. The theory evolved out of the work of Anselm and Abelard at the turn of the twelfth century, developed steadily through the thirteenth and fourteenth centuries and was still undergoing changes in the fifteenth and sixteenth centuries. It is generally agreed that its early stages were closely bound up with the theory of fallacies, but as a general semantic theory, it developed in response to a variety of needs, and one mistake of modern attempts at interpretation is to seek a unique rationale of one notion or another. Each notion evolved continually, satisfying one need at one time and another at a later date, and often several conflicting needs at the same time. Another mistake is to try to map each notion *seriatim* onto corresponding notions in contemporary semantic theory, but although one can see analogies and similarities, none of the medieval "properties" matches exactly any modern notion.

---

1. Historical Survey

William of Sherwood, writing from an Oxford perspective in the 1240s, identified four properties of terms: signification, supposition, copulation and appellation: "There are four properties of terms that we intend to distinguish now ... These properties are signification, supposition, copulation, and appellation" (tr. Kretzmann, p. 105).[1] A different tradition is represented by Peter of Spain and Lambert of Auxerre,

namely, that of Paris. Lambert identifies five properties of terms: supposition, appellation, restriction, distribution and relation: "Now, there are many properties of terms: supposition, appellation, restriction, distribution [distinction], and relation ... But because signification is, as it were, the fulfillment of a term, and the properties of terms are founded on signification, for the sake of clarity in what follows we must at the outset consider what the signification of a term is, and how it differs from supposition", tr. Kretzmann and Stump, p. 104.[2] In his treatment he includes ampliation (a correlative to restriction) and thereby matches the sections in Peter of Spain's *Tractatus*: *De suppositionibus* (On Suppositions), *De relativis* (On Anaphora), *De ampliationibus* (On Ampliations), De appellationibus (On Appellations), *De restrictionibus* (On Restrictions), *De distributionibus* (On Distribution).

To these must be added further properties of terms which were important in the theory's gestation in the twelfth century but were later no longer included as distinct properties: univocation, equivocation, adjectivation, substantivation and so on. From the fourteenth century onwards, other properties were also abandoned, so that finally the important lasting properties were signification, supposition, ampliation and restriction, and the supposition of relatives.

Given the central importance which it later attained, it is surprising how late supposition was in being identified as one of the properties of terms. In the twelfth century, the primary distinction was that between signification, or univocation, and appellation. As early as Anselm's *De Grammatico* we find a distinction between *significare per se* (signify as such) and *significare per aliud* (signify relatively). The latter was also known as *appellare* (to name or appellate). Whereas in the former (*per se*) what was signified was a form, in the latter what was signified, or appellated, was a thing. A single form is common to a number of things we call by the same name, so one crucial problem in the philosophy of language is to understand how the different uses of a name are unified in the face of the distinctness of the many individuals talked about. In this way, we can understand the importance played in the twelfth century of univocation in contrast with appellation. Proper names, naming single items, are contrasted with appellative names, naming many. Univocation is the signification of a univocal word, described in the *Fallacie Parvipontane* as "a variation in the supposition/appellation of a name while keeping the signification the same" (my translation).[3]

Throughout much of the twelfth century, 'supposition' retained the linguistic sense it had had since at least the time of Priscian (4th century CE), namely, the placing of a name as subject. The relation of that name as subject to the thing named was called its appellation. This was the ability of a univocal appellative noun to name different things. The appellation of a name was not fixed, however. It could be ampliated or restricted by the predicate. So, for example, the predicate '*opinabilis*' (credible) ampliates or extends the appellation of a name such as 'man' to cover a wider range, perhaps of no longer existing men, or of men who might have existed, or who might exist in the future. Thus, appellation came to mean the present correct application of a term (William of Sherwood, p. 74), which could be ampliated, or even restricted by, say, the apposition of an adjective: 'white man' appellates only white men, not all even presently existing men.

Later, however, 'supposition' came to replace 'appellation', and the latter term underwent its own

transformation. Supposition became the major property of an occurrence of a term in a proposition (including now predicates, and even parts of predicates), distinguishing what in particular was being spoken of on some occasion of utterance by some particular use of a word, from that word's general property (signification) of meaning. The term 'man' signifies the form of man wherever the term occurs, but each occurrence of the term 'man' supposits for (the individuals in) possibly distinct classes or groups of men (but note that it supposits for the men, not the classes).

Given the literal meaning of '*supponere*' as "acting as subject", it is unsurprising that several authors, even as late as Vincent Ferrer in the 1370s, restricted supposition to subject terms, preferring to speak instead of copulation as the corresponding property of the predicate. Those reservations aside, however, the natural similarity in function of subject and predicate in picking out varying classes of things while remaining a univocal term, led most authors to extend the notion of supposition to all terms. At the same time, talk of univocation was superseded by signification, by implication univocal signification unless equivocation was identified.

A major shift in focus occurred in the early fourteenth century, probably occasioned by William of Ockham -- certainly, he was the leading figure. On the one hand, signification came to be understood entirely extensionally, given Ockham's antipathy towards real universals. Accordingly, what was taken to be signified by a term were the things of which it can be truly predicated instead of a form or property which they share. For example, where before the term 'man' was taken to signify humanity, and 'white' whiteness, for Ockham the former signifies (all) men and the latter all white things. From this perspective, the contrast found in thirteenth century authors between (as Peter of Spain called it) natural supposition versus accidental supposition, becomes redundant. Natural supposition was the term's extension, accidental supposition that range of things it supposited for in a particular proposition. For Peter, the term signifies a form, and supposits naturally for a class of objects, whereas an occurrence of the term supposits accidentally on an occasion of use for a group of those objects. For Ockham, the term signifies that class, the form being no more than a *fictum* (a figment of the mind) or (in his later works) the mental act (of conceiving of those things) itself. Thus natural supposition (William of Sherwood calls it habitual supposition, the *Tractatus de proprietatibus sermonum*, absolute supposition) drops out as unnecessary.

From the early fourteenth century, supposition came to dominate among the properties. Ampliation and restriction were already functions of supposition (or appellation as it had been), copulation as the naming function of predication is subsumed under supposition, and *relatio*, the connection of anaphoric terms to their antecedents, becomes a discussion of the supposition of those terms. Finally, distribution is treated as a particular mode or type of supposition, confused and distributive supposition. Indeed, in the final phase, it could seriously be asked if supposition was indeed a property of terms, as does Albert of Saxony. For like others, he defines supposition as '*acceptio termini*' (the act of acceptance of a term) of various kinds, and so is as much a property of the speaker as of a term.

2. Signification

Signification contrasts with the other properties of terms in one major respect, for the other properties (perhaps with the exception of appellation in some authors, and natural supposition) are all properties of terms relative to their occurrence in particular propositions -- indeed, they are properties of occurrences of terms. Signification, however, is had by a term prior to its particular uses or occurrences: "Now signification differs from supposition in that signification is prior to supposition", tr. Kretzmann and Stump, p. 105.[4] Indeed, the other properties are dependent on the signification of the term. For example, an occurrence of a term can only supposit for certain objects in virtue of its signifying them, among others, or signifying some property they share.

A twelfth century commentary on the *Perihermeneias* says that at the time of Aristotle, there was a great debate over the principal signification of utterances: was it '*res*' (things) or incorporeal natures (Plato) or *sensus* (sensations) or *imaginationes* (representations) or *intellectus* (concepts)? In fact, medieval philosophers of language were heir to two conflicting semantic theories. According to Aristotle, the greatest authority from the ancient world, words name things by signifying concepts in the mind (Boethius translated his term as *passiones animae* -- affections of the soul) which are likenesses abstracted from them. But Augustine, the greatest of the Church fathers, had held that words signify things by means of those concepts. This led the medievals to the question: do words signify concepts or things? The question had already been asked by Alexander of Aphrodisias and his answer was transmitted to the medievals in Boethius' second commentary on Aristotle's *Perihermeneias* (*De Interpretatione*): "Alexander asked, if they are names of things, why does Aristotle say that they are primarily signs of concepts ... He says that although spoken words are names of things, we do not use them to signify things, but [to signify] those affections of the soul which are produced in us by them", my translation. [5] So words primarily signify concepts.

But the matter was not settled, other than that whatever view a medieval philosopher took, it had to be made to accord with the authority of Aristotle, perhaps *in extremis* by reinterpreting Aristotle's words. Abelard refers to a distinction between *significatio intellectuum* (signification of concepts) and *significatio rei* (signification of the thing), more properly called nomination or appellation (see De Rijk, *Logica Modernorum*, vol. II(1), pp. 190-9). Similarly, the *Tractatus de proprietatibus sermonum* asks whether words signify concepts (*intellectuum*) or things, and responds: both, but primarily a thing via a concept as medium (op.cit. II (2), p. 703).

A particular novelty of the thirteenth century, however, was to conceive of the concept itself as a sign. We find this in Lambert of Auxerre: "The signification of a term is the concept of a thing, a concept on which an utterance is imposed by the will of the person instituting that term. For, as Aristotle maintains in the first book of *De Interpretatione* (16a3-5), utterances are signs of states in the soul -- i.e., in the understanding -- but concepts are the signs of things", tr. Kretzmann and Stump, p. 104.[6] Hence by transitivity, utterances which are signs of concepts which are signs of things are themselves signs of things: "An utterance that is a sign of a sign ... [is] a sign of the concept directly but a sign of the thing indirectly", tr. Kretzmann and Stump, p. 105.[7] For example, 'man' immediately signifies the concept man, but by mediation of the concept it signifies the second substance or form of man. Accordingly, it can supposit for what fall under man, e.g., Plato and Socrates. But it does not signify Plato or Socrates.

As noted above, this last distinction was elided by Ockham, in removing the universal, the form. ‘Man’, he says, signifies Plato and Socrates and all men equally. Once signification is treated extensionally in this way, its only difference from supposition lies in its priority: a general term signifies all those things of which it can be truly predicated (Ockham, *Summa Logicae* I c. 33).

3. Supposition and Copulation

Just as signification corresponds most closely -- though not exactly -- to contemporary ideas of meaning or sense, so supposition corresponds in some ways to modern notions of reference, denotation and extension. The comparison is far from exact, however. One major difference is that the medievals distinguished many different modes (*modi*) of supposition. Despite the difference between different authors’ semantic theories, particularly as they developed over the centuries, there is a remarkable consistency in the terminology and interrelation of the different modes.

The major division is between material, simple and personal supposition. Material supposition is when a term stands for a linguistic item as such. Often, of course, this will be a case of autonymy, when it stands for itself. William Sherwood writes: "It is called material when a word itself supposits either [A] for the very utterance itself or [B] for the word itself, composed of the utterance and the signification -- as if we were to say [A] ‘man is a monosyllable’ or [B] ‘man is a name’", tr. Kretzmann, p. 107.[8] Writing in Paris nearly a hundred years later, Thomas Maulfelt put it like this: "Material supposition is a term standing for itself or for another similar to it in sound or writing suppositing in the same way or otherwise which it was not imposed to signify or for some other sound which is not inferior to it, and which it does not signify naturally and properly", my translation.[9] ‘Noun’ supposits materially for itself when we say ‘Noun has four letters’ -- or ‘"Noun" has four letters’; it doesn’t supposit materially, but personally for itself, when we say ‘A noun is a part of speech’.

Simple supposition is harder to characterize generally. A common description is to say it occurs when a term supposits for the universal or form which it signifies. However, not everyone believed that terms signified universals. (See the entry on the medieval problem of universals.) So whereas William of Sherwood wrote "It is simple when a word supposits what it signifies for what it signifies", tr. Kretzmann p. 107,[10] and Walter Burleigh likewise: "supposition is simple when a general term or a complex singular term supposits for what it signifies", my translation,[11] William of Ockham appears to characterize simple supposition quite differently: "Simple supposition occurs when a term supposits for an intention of the soul and is not functioning significatively", tr. Loux p. 190.[12] But in fact, the difference between them is not in their theories of supposition, but rather of signification. As we noted earlier, Ockham believes that a general term like ‘man’ signifies individual men like Plato and Socrates; Burleigh that it signifies a second substance (the universal), man. Burleigh appeals to the authority of Aristotle: "man signifies a second substance and does not signify a substance which is a genus, and so signifies a species".[13] Ockham, however, dismissed real universals, and (at least in his later theory represented in *Summa Logicae)* believed the only universals were words, including words of the inner mental language, mental acts. So for him, a spoken or written term has simple supposition when it stands

for the mental act to which it is subordinated by the conventions of signification, that is, the mental act abstracted from those things which the word conventionally signifies.

John Buridan famously eliminated simple supposition altogether, for this very reason, namely, that universals are words of a mental language, so terms suppositing for them are suppositing for a kind of linguistic item, and so such a case should be included under material supposition.

Material and simple supposition are contrasted with personal supposition, what we might call the standard case, where a term stands for ordinary objects -- the objects it signifies (for Ockhamists) or for its supposita, as for example, Lambert (p. 209) or Burleigh express it: "Personal supposition is when a term supposits for its suppositum or supposita or some singular of which the term is [truly and] accidentally predicable".[14] The contrast is a useful one, as is shown in the following standard fallacies:

Homo est dignissima creaturarum	(Man is the worthiest of creatures)
Sortes est homo	(Socrates is a man)
Ergo Sortes est dignissima creaturarum	(So Socrates is the worthiest of creatures).

The premises are true and the conclusion false, so wherein lies the fallacy? It is one of equivocation or "four terms": ‘homo’ (’man’) has simple supposition in the first premise and personal supposition in the second, so there is no unambiguous middle term to unite the premises. Again:

Currens est participium	(Running is a participle)
Sortes est currens	(Socrates is running)
Ergo Sortes est participium	(So Socrates is a participle).

This time, ‘currens’ (’running’) supposits differently in the two premises, materially in the first and personally in the second, explaining why putting the two truths together fallaciously leads to the false conclusion.

What determines whether a term has material, simple or personal supposition? One view might be that it depends on the intention of the speaker; another, that all propositions are ambiguous. However, the prevailing medieval view was that it was determined by the predicate, so that, e.g., a predicate like ‘is a noun’ requires material supposition for the subject, while ‘is a species’ requires simple supposition: "A subject, on the other hand, sometimes supposits a form separately and sometimes does not, depending on what the predicate demands, in accordance with the following [principle]: Subjects are of such sorts as the predicates may have allowed", tr. Kretzmann, p. 113.[15] This slogan, *talia subiecta qualia predicata permiserint* (subjects are such as predicates permit), was commonly attributed to Boethius; but Sherwood correctly points out that Boethius’ point was different, and his phrase was the converse: *"talia [predicata] qualia subiecta permiserint"* -- see De Rijk, *Logica Modernorum* II (1), p. 561.

In the fourteenth and fifteenth centuries, it became commonplace to require a *nota materialitatis* (a sign of the material use) for a term in material supposition. Such a sign was to prefix ‘iste terminus’ (’this term’) or ‘ly’ (taken from the French definite article) to the term. Without such an indication, the term was considered to have personal supposition by default.

Personal supposition is divided by most authors into discrete and common supposition, the first that of singular terms (proper names, demonstrative phrases and so on), the latter that of general terms. Common personal supposition is again divided, namely, into determinate and confused, and the latter into confused and distributive and merely confused supposition. These three modes are well illustrated by the four categorical forms (see the entry on the traditional square of opposition):

> (A) All A are B
>
> (E) No A are B
>
> (I) Some A are B
>
> (O) Some A are not B.

The subject of (I)- and (O)-, and the predicate of (I)-propositions, have determinate supposition; the subject and predicate of (E)-, the subject of (A)-, and the predicate of (O)-propositions have confused and distributive supposition; and the predicate of (A)-propositions has merely confused supposition. This is common doctrine; what varies is how these modes are characterised. "It is determinate", writes William of Sherwood, "when the locution can be expounded by means of some single thing", tr. Kretzmann, p. 108, but adds a doubt: "It seems that when I say ‘a man is running’ [i.e. an I-proposition] the term ‘man’ does not supposit determinately, since [A] the proposition is indefinite, and [B] it is uncertain for whom the term ‘man’ supposits. Therefore it supposits [A] indefinitely and [B] uncertainly; therefore indeterminately", tr. Kretzmann, pp. 115-6.[16] But this is simply a matter of terminology, he replies -- determinate supposition means suppositing for one, not for many, but for no particular one, for that would constitute discrete supposition. Note, however, that for Ockham, Burleigh and others, a term with determinate supposition supposits for everything of which it can be truly predicated. The sense in which it is true of one, rather than many, is that the proposition is true if true of one. The term still supposits for all. As Burleigh says: "It is called determinate supposition, not because a term suppositing determinately supposits for one and not for all, but it is called determinate supposition because for the truth of the proposition in which the general term supposits determinately, it is required that it be true of one determinate suppositum", my translation.[17]

All that William of Sherwood can offer to characterize confused supposition is to say that it is had when a term supposits for many, then resorting to examples for its divisions. Peter of Spain tries harder: "Confused supposition is the acceptance of a general term for many by means of a universal sign", my translation,[18] and confused and distributive when it supposits for all. But this is still unclear, and eventually a solution was found, in the doctrine of ascent and descent. Call the individual things falling

under a general term its "inferiors", and call the singular proposition resulting from a general proposition by replacing a general term (with its qualifier) by a term with discrete supposition for one of its inferiors, one of the proposition's "singulars". Then the inference from a general proposition to one of its singulars is called "descent", the converse inference, "ascent". If descent under a term is valid, or at least is valid without changing the rest of the proposition, it is called "mobile", otherwise "immobile". Walter Burleigh, William of Ockham and their followers could then define the three modes of common personal supposition as follows:

1. An occurrence of a general term in a proposition has determinate supposition when one can descend validly from the proposition to the complete disjunction of its singulars with respect to that term, and conversely can ascend validly from any singular;
2. Otherwise, it is confused, and i) is confused and distributive if one can validly descend from the proposition to the indefinite conjunction of its singulars with respect to that term; otherwise ii) it is merely confused.

In this last case, 2 ii), Ockham noted: "it is possible to descend by way of a proposition with a disjunctive predicate and it is possible to infer the original proposition from any [singular]", tr. Loux, p. 201.[19] Thus the predicate of an A-proposition has merely confused supposition since from 'Every man is an animal' one can validly infer 'Every man is this animal or that animal and so on for all animals', and conversely, from 'Every man is this animal' (were it ever true) one could infer that every man was an animal.

A final complication came at the end of the fourteenth century and subsequently, when it was asked: "Are there only three modes of common personal supposition?" Some answered, 'Yes'; but others distinguished two modes of merely confused supposition, or equivalently distinguished a fourth case, 2 iii), collective supposition, where descent was permissible to a proposition with a conjunctive predicate. The standard example was 'All the apostles of God are 12', which entails 'Matthew and Mark and so on are 12', with a conjunctive subject.

Several recent commentators have asked what the medieval theory of modes of common personal supposition was: was it a theory of inference, of quantification, of truth-conditions, of fallacies, or what? Asking these questions may help us to come to terms with the theory; but if pressed too hard, they are unhelpful. The medievals' theory was none of these things; it was their theoretical basis with which to pose and answer semantic questions.

All modes of supposition so far described fall under what was variously called proper supposition or accidental supposition or, excluding material supposition, formal supposition. Distinct from these were what Peter of Spain called natural supposition, mentioned earlier in the discussion of signification, and what Ockham and others call improper supposition, covering metaphor and other figures of speech.

Moreover, various authors before the fourteenth century, and at least one fourteenth century author, too (namely, Vincent Ferrer) differed a little from what has been said in not attributing supposition to

predicate terms, calling the corresponding property of predicates, "copulation". However, even in the early thirteenth century, this distinction was fading. William of Sherwood makes it clearly: substantive names and pronouns supposit, whereas adjectives, participles and verbs copulate. Many of the divisions of supposition are repeated for copulation, but only the modes of common personal supposition: "every copulating word signifies in adjunction to a substantive and thus is copulated personally ... [and] every copulating word is the name of an accident, but every name of an accident is common; therefore no copulation is discrete", tr. Kretzmann, p. 121.[20] However, even in Lambert and Peter of Spain, there is only an empty nod towards copulation. Lambert notes that properly speaking, supposition attaches to substantives, while copulation is appropriate for adjectival terms. But broadly speaking, he says, supposition belongs to both (*op.cit.*, pp. 208/109). The distinction is clearly unnecessary and useless, and although the term is preserved in, e.g., Walter Burleigh's *De Puritate,* his discussion headed 'On Copulation' is in fact a discussion of the uses of the copula, 'est'.

4. Ampliation and Restriction

Some words have the effect of widening or narrowing the supposition of other terms in a proposition. For example, by qualifying 'man' with the adjective 'white', we restrict the supposition of 'man' in 'A white man is running' to white men; while a verb in the past tense ampliates the supposition of the subject to what were its supposita. For example, 'A white thing was black' means that something which is now white or was white in the past was black.

Lambert of Auxerre describes the many aspects of a proposition which can produce ampliation or restriction. Some are natural, as when 'rational' restricts 'animal' to supposit only for men by being adjoined to it; other cases are '*usualis*' -- 'use-governed' in Kretzmann and Stump's translation: when we say 'The king is coming', we are taken to mean the king of the country where we are, so 'king' is restricted to supposit only for that king. Some ampliation and restriction is effected by consignification, that is, by an aspect of a word -- by the tense of the verb, or by the gender of an adjective: in 'homo est alba', the feminine ending of 'alba' restricts 'homo' to supposit only for women. Other ampliation and restriction is effected by the signification of words -- as in the case of 'rational animal' above, or in 'Socrates' donkey', where the possessive restricts 'donkey' to supposit only for Socrates' donkeys.

Peter of Spain notes that only general terms can be ampliated or restricted, and only terms having personal supposition.

Just as the past tense ampliates the subject to include past as well as present supposita, modal verbs ampliate the subject to possible supposita, as do verbs such as 'I understand', 'I believe', and indeed, notes Albert of Saxony, verbal nouns ending in '-bile': 'possible', 'audible', 'credible', 'capable of laughter' and so on. Albert realises that even 'supposit' ampliates the subject: when we say 'This term supposits for something', what it supposits for need not actually exist, but might be past, future, possible or merely intelligible. Similarly, 'must' ampliates for possible supposita, for 'A must be B' means 'It is not possible that A not be B'; but contradictories must clearly ampliate in the same way, and 'It is possible that A not be B' ampliates for possibilia. Ampliation over possibles means care is needed in

inferences from ‘is’ to ‘can’: ‘A can be B’ need not follow from ‘A is B’, for ‘A can be B’ means that what is or can be A can be B, and even if A is B, not everything which can be A might be capable of being B. For example, ‘Every planet lighting our hemisphere can be the sun’ is false even though it in fact is the sun which lights it, since the moon might light our hemisphere but the moon cannot be the sun.[21]

It is an interesting fact that almost alone among terminist logicians, Ockham does not speak of ampliation and restriction. The reason appears to be that he disagrees with the truth-condition given above for ‘A white thing was black’, and similar cases. This proposition, he says, is ambiguous. Rather than meaning that what is or was white was black, it equivocates between ‘What is white was black’ and ‘What was black was white’: "quaelibet propositio de praeterito et de futuro, in qua subicitur terminus communis vel pronomen demonstrativum cum termino communi vel terminus discretus importans aliquod compositum, est distinguenda", *Summa Logicae* II 7, p. 269 ("in the case of every proposition of the past and of the future in which a general term or a demonstrative pronoun with a general term or a discrete term referring to some composite, we must distinguish [two senses]", my translation). Whether Ockham realized that his theory improved on the ampliative theory is unclear; nonetheless, his account seems to accord better with intuition. For the ampliative account is disjunctive: it says that the proposition is true if either what was white was black or what is white was black. Then it is true if either disjunct is true; whereas on Ockham’s account it has two different senses, and can be false on one while true on the other -- true because something now white used to be black but false if nothing which used to be white was ever black. For a clearer case, take the first sex-change operation -- for the first time, say, a woman was a man. But no one who was a woman had ever been a man, so it was (also) false that a woman was a man. The proposition ‘A woman was a man’ is ambiguous and "*est distinguenda* " (i.e., different senses must be distinguished). A famous example was ‘An old man will be a boy’, referring to someone not yet born -- he will be an old man sometime, and will be a boy before that.

Ockham also eschews talk of ampliation in giving his account of modal propositions. Does ‘A white thing can be black’ mean that what is or can be white can be black, i.e., that either what is white can be black or what can be white can be black, as the ampliative account demands? -- see, e.g., Albert of Saxony: "’A white thing can be black’: this signifies that what is white or what can be white can be black", my translation.[22] No, for this loses the sense in which it is self-contradictory to suggest that it is possible that a white thing be black. The proposition is ambiguous, in one sense self-contradictory, in the other true, Ockham notes (*Summa Logicae* II 10), because ‘This is black’ can be true if uttered pointing to something white: "‘A white thing can be black’ is true, because ‘This is black’ is possible, pointing to something for which ‘white’ supposits; but ‘A white thing is black’ is impossible".[23]

5. Appellation

Perhaps the term with the most varied history is ‘appellatio’, though even so, one can discern a common thread running through it. It starts, we saw, as an equivalent of ‘nominatio’ in Anselm and Abelard, and by the thirteenth century is used to pick out the present extension of a term, that of which it can be truly predicated in the present tense: "Now it is essential to know that ‘appellation’ is used in four ways ...

Used in the fourth way, 'appellation' is the acceptance of a term for a suppositum or for supposita actually existing. And it is appellation spoken of in this fourth way that is meant at present", tr. Kretzmann and Stump, pp. 113-4.[24] In Burleigh, at the turn of the fourteenth century, it almost usurps the place of 'copulatio', being for him the relation of the predicate to its inferiors: "Just as supposition, strictly speaking, is a property of the subject in relation to the predicate, so appellation is a property of the predicate in relation to the subject or to its inferiors", my translation.[25] There are shades too, of Abelard: "a univocal general term appellates its inferiors but does not signify them".[26] But the crucial phrase which runs through the history of the term appears here too: "the predicate appellates its form", *op.cit.*, p. 48 ("praedicatum appellat suam formam"). What he says he means by this is that the predicate be truly predicable at some time, in the present tense, of the supposita of the subject.

It is ampliation and restriction which distinguish this property of the predicate from properties of the subject. For, on the standard account, 'A white thing was black' is true only if 'is black' has been truly predicable of the supposita of the subject, namely, of what is or was white. In contrast, 'A white thing is black' may never have been true -- indeed, in this case, it never will have since it is self-contradictory. So the subject does not always "appellate its form". A consequence pointed out by Albert of Saxony, among others, is that conversion needs application with care in such cases. Consider 'The just will be justly damned': this can be true, if those who are just sin in the future. But 'Justly will the just be damned' is false, for 'Justly are the just damned' will never be true. Similarly, 'Socrates approaching you know' can be true while 'You know Socrates approaching' (the "hooded man" fallacy from Aristotle's *De Sophisticis Elenchis* 179b1-3) may be false (for you know Socrates, but do not recognize him approaching). It is explained by reason of the fact that the predicate appellates its form (for 'You know Socrates approaching' requires that the predicate 'know Socrates approaching' be true of you and so is false), whereas 'Socrates approaching you know' requires only that 'Him you know' be true, referring to Socrates, and it is true. The doctrine of appellation could thus be used to diagnose familiar fallacies.

Scott, in his discussion of Buridan's sophism, 'You know the one approaching' (pp. 42-9), claims that Buridan's concept of appellation (which he misleadingly translates as 'connotation') in his diagnosis is novel, and Spade follows him in his commentary on Peter of Ailly(p.109 n.188). But the notion is clearly continuous with Burleigh's in that 'appellating its form' requires true predication via a demonstration. Indeed, it is continuous with Lambert's usage, for 'chimera' has no appellata precisely because 'hoc est chimera' ('This is a chimera') is false whatever is demonstrated. Admittedly, Buridan explicitly restricts appellation to appellative terms, that is, "every term connoting something other than that for which they supposit" (*Treatise on Suppositions*, tr. King, p. 159), and "it appellates that which it connotes as belonging to that for which it supposits". Thus 'white' connotes whiteness and supposits for white things. What appears to be novel in Buridan is the extension of the doctrine to intentional verbs, which, says Buridan (and Albert) cause the terms following them (the predicate) to appellate their *rationes*, that is, the concepts by which they signify what they do. Thus in 'You know Socrates approaching', the predicate 'know Socrates approaching' appellates its concept, the *ratio* 'Socrates approaching', so the proposition is false unless you are aware who it is; whereas in 'Socrates approaching you know', the subject 'Socrates approaching' does not appellate its concept, and so it suffices that 'Him you know' be true, where 'him' refers to Socrates.

6. Relation

Lambert of Auxerre speaks explicitly of relation as a property of terms, but most authors describe the phenomenon as the supposition of relatives. The relation in question is that between anaphoric terms and their antecedents. Most treatises consist of a repetition of a standard taxonomy: there are relatives of substance and relatives of accident, relatives of identity and relatives of diversity. For example, relatives of substance and of identity are 'who', 'he', 'his', and among these, reciprocal relatives such as 'himself'; of diversity are 'another' and 'someone else'. The latter are said to refer back (*referre*) to their antecedent but to supposit for something different, as in, e.g., 'Socrates is running and someone else is debating'. Relatives of accident include, first, of identity, 'such as', 'like', 'when', and of diversity, 'in another way', 'other'; in, say, 'Socrates is running and Plato is other', 'other' refers back to 'running'; in what way it differs will depend on the author's account of predicates.

In contrast to the extensive taxonomy, medieval discussions of relatives before the mid-fourteenth century seem rather short on theory, dealing with puzzles more by common sense and description than in any unified way. For example, the prevailing view is that relatives of identity preserve the supposition of their antecedent. Why, then, can they usually not replace the antecedent? Consider, e.g., 'Every man sees himself'. Does 'himself' have the same supposition as 'man' (or 'every man') and if so, why can the antecedent not replace it? -- for 'Every man sees every man' has a very different signification. Lambert claims that in the case of reciprocal relatives the reciprocal pronoun can replace the antecedent unless the antecedent is taken universally, as here. In this case, it still supposits for the same as its antecedent, but in a different way, namely, discretely. Ockham spells it out: "in 'Every man sees himself', 'himself' supposits for every man by means of confused and distributive mobile supposition: but it does this singularly since it is not possible to descend without altering the other extreme ... thus, 'Every man sees himself, therefore, Socrates sees Socrates'", tr. Loux, p. 218.[27] This is correct and sensible; but what is the theory to explain it?

Ockham claims that, although non-reciprocal relatives of identity always supposit for that for which their antecedent supposits, those whose antecedent is a general term occurring with personal supposition can never be replaced by their antecedent and result in an equivalent proposition. For example, 'A man runs and he disputes' is not convertible with 'A man runs and a man disputes'. Buridan has two rules to explain what is happening:

1. A relative of identity refers back to its antecedent in respect only of those of its supposita for which the categorical proposition in which its antecedent occurred is verified
2. A relative of identity supposits as its antecedent does, viz materially if materially, personally if personally, distributively if distributively, determinately if determinately, merely confusedly if merely confusedly, except as in 1.

Hence, in e.g., 'Man is a species and it is predicable of many', or 'Socrates runs and he disputes', the relative can be replaced without loss of meaning by its antecedent. But in 'A man runs and he disputes',

it cannot, for the supposition would change. ‘He’ supposits only for men who are running, whereas ‘a man’, if it replaced it, would supposit for all men.

Another issue concerned identifying the contradictory of a proposition containing a relative. In the case of a categorical containing a relative of identity such as ‘A which is B is C’, it is ‘A which is B is not C’, says Albert of Saxony (*Perutilis Logica* II 8, f. 14rb, ed. Muñoz García §§526-7). For the former is equivalent to ‘A is B and A is C’, and the latter to ‘A is not B or A is not C’, he says.

7. Conclusion

The medievals did not have a solution to every semantic puzzle with which they were faced, any more than do contemporary philosophers. But the theory of properties of terms was the basis of a rich semantic theory within which they were able to develop complete and fruitful theories which yielded significant insights -- both for them and for us -- into a broad range of semantic issues.

Bibliography

- Albert of Saxony, *Perutilis Logica*, Venice 1522, ed. Á. Muñoz García, *Alberto de Sajonia: Perutilis Logica o Lógica Muy Útil (o Utilísima)*, Ciudad Universitaria: Universidad Nacional Autonoma de Mexico, 1988.
- Anselm, *De Grammatico*, ed. and tr. in D.P. Henry, *Commentary on ‘De Grammatico’*, Dordrecht: Reidel, 1974, pp. 48-80.
- A.M.S. Boethius, *In librum Aristotelis De Interpretatione libri sex.* Editio seconda, seu majora commentaria, *Patrologia Latina*, vol. 64 cols. 393-640.
- De Rijk, L.M. *Logica Modernorum*, Assen: Van Gorcum, vol. I 1962, vol. II parts 1-2 1967.
- De Rijk, L.M. ‘The origins of the theory of properties of terms’, in *The Cambridge History of Later Medieval Philosophy*, ed. N. Kretzmann, A. Kenny and J. Pinborg, Cambridge: Cambridge University Press, 1982, pp. 161-73.
- John Buridan, *De Suppositionibus*, tr. P. King, *Jean Buridan’s Logic: the treatise on supposition, the treatise on consequences*, Dordrecht: Reidel, 1985.
- Lambert of Auxerre, *Logica (Summa Lamberti)*, ed. F. Alessio, Florence: La nuova Italia Editrice 1971, tr. N. Kretzmann and E. Stump in *Cambridge Translations of Medieval Philosophical Texts*, Cambridge: Cambridge University Press, 1988, pp. 102-62.
- Maierù, A. *Terminologia logica della tarda scolastica*, Rome: Edizione dell’Ateneo, 1972.
- Peter of Spain, *Tractatus*, ed. L.M. De Rijk, Assen: Van Gorcum, 1972.
- Scott, T.K. *John Buridan: Sophisms on Meaning and Truth*, tr. with intro., New York: Appleton-Century-Crofts, 1966.
- Spade, P.V. *Peter of Ailly: Concepts and Insolubles*, Dordrecht: Reidel, 1980.
- Priest, G. and Read, S. ‘Ockham’s rejection of ampliation’, *Mind* 90, 1981, pp. 274-9.
- Read, S. ‘Thomas of Cleves and Collective Supposition’, *Vivarium* 29, 1991, pp. 50-84.
- Thomas Maulfelt, *De Suppositionibus*, Edinburgh ms. 138, ff. 62r-72r.

- Walter Burleigh, *De Puritate Artis Logicae Tractatus Longior*, ed. P. Boehner, St Bonaventure: Franciscan Institute, 1955.
- William of Ockham, *Summa Logicae*, ed. P. Boehner, G. Gal and S. Brown, St Bonaventure: Franciscan Institute, 1974; Part 1 tr. M. Loux, *Ockham's Theory of Terms*, Notre Dame: University of Notre Dame Press, 1974; Part II tr. A. Freddoso, *Ockham's Theory of Propositions*, Notre Dame: University of Notre Dame Press, 1980.
- William of Sherwood, *Introductiones in Logicam*, ed. M. Grabmann, 'Die Introductiones in logicam des Wilhelm von Shyreswood', Munich: *Sitzungberichte der Akademie der Wissenschaften*, Philosophisch-historische Klasse 10, 1937, pp. 30-106; tr. N. Kretzmann, *William of Sherwood's 'Introduction to Logic'*, Minneapolis: University of Minnesota Press, 1966.

Other Internet Resources

- Paul Spade's Medieval Logic and Philosophy page

Related Entries

Albert of Saxony | Anselm, Saint [Anselm of Bec, Anselm of Canterbury] | Aristotle: logic | Buridan, John [Jean] | fallacies: medieval theories of | medieval philosophy | Ockham [Occam], William | semiotics: medieval | square of opposition | universals: the medieval problem of

Stephen Read
slr@st-and.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 5, 2002
Content last modified: February 5, 2002

Stanford Encyclopedia of Philosophy
Notes to Medieval Theories of Properties of Terms

Notes

1. "Quattuor sunt proprietates termini quas ad presens intendimus diversificare ... Et sunt hes proprietates significatio, suppositio, copulatio et appellatio", *Introductiones in Logicam*, pp. 74-5.

2. "Multa autem sunt proprietates termini, scilicet: suppositio, appellatio, restrictio, distributio [ed.: distinctio] et relatio ... sed quia significatio est sicut perfectio termini et proprietates termini supra significationem fundantur, ideo in principio ad evidentiam sequentium videndum est quid sit termini significatio et in quo differt a suppositione", *Logica Lamberti*, p. 205.

3. "manente significatione variata nominis suppositio", edited in De Rijk's *Logica Modernorum*, vol. I p. 562, and in the *Tractatus de univocatione monacensis* as "manente eadem significatione variata nominis appellatio", *op.cit.* vol. II(2) p. 337.

4. "Differt autem significatio a suppositione in hoc, quod prior est significatio quam suppositio", *Logica Lamberti*, p. 206.

5. "quaerit Alexander si rerum nomina sunt, quid causae est ut primo intellectuum notas esse voces dixerit Aristoteles ... inquit, quod licet voces rerum nomina sint, tamen non idcirco utimur vocibus, ut res significemus, sed ut eas quae nobis ex rebus innatae sunt animae passiones. Quocirca propter quorum significantiam voces ipsae proferuntur, recte eorum primo esse dixit notas", *Patrologia Latina*, vol. 64 col. 413A.

6. "significatio termini est intellectus rei ad quem intellectum rei vox imponitur ad voluntatem instituentis: nam, sicut vult Aristotles in primo Perihermeneias, voces sunt signa passiones que sunt in anima, id est in intellectu; intellectus autem sunt signa rerum", *Logica Lamberti*, p. 205.

7. "vox que est signum signi ... immediate est signum intellectus, mediate autem signum rei", *ibid.* p. 206.

8. "et dicitur materialis, quando ipsa dictio supponit vel pro ipsa voce absoluta vel pro ipsa dictione composita ex voce et significatione, ut sic dicamus: homo est dissillabum, homo est nomen", *Introductiones in Logicam*, p. 75.

9. "Suppositio materialis est terminus stans pro se vel pro alio sibi simili in voce vel in scripto eodem

modo vel aliter supponente cui non imponitur ad significandum vel pro aliqua alia voce que non est inferior ad ipsum, nec ipsum proprie naturaliter significat", *De Suppositionibus*, f. 62r.

10. "Et est simplex, quando dictio supponit significatum pro significato", p. 75.

11. "suppositio vero simplex est, quando terminus communis vel singulare aggregatum supponit pro eo quod significat", *De Puritate*, p. 3.

12. "Suppositio simplex est quando terminus supponit pro intentione animae, sed non tenetur significative", *Summa Logicae*, p. 196.

13. "homo significat secundam substantiam et non significat substantiam quae est genus, ergo significat speciem", p. 7.

14. "Suppositio personalis est, quando terminus supponit pro supposito vel suppositis vel aliquo singulari, de quo terminus accidentaliter praedicatur", *De Puritate*, p. 3.

15. "Subiectum autem quandoque supponit formam absolute, quandoque autem non et hoc secundum exigentiam predicati secundum illud: talia sunt subiecta, qualia permiserint predicata", Sherwood, *Introductiones in Logicam* p. 78.

16. "Et est determinata quando potest exponi loqutio per aliquod unum", p. 75; "Videtur enim, quod iste terminus homo, cum dico: homo currit non supponit determinate. Est enim indefinita. Et item incertum est, pro quo supponit. Ergo supponit incerte et indefinite. Ergo indeterminate", p. 79.

17. "Et dicitur suppositio determinata, non quia terminus sic supponens determinate supponit pro uno et non pro alio, sed dicitur suppositio determinata, quia ad veritatem propositionis, in qua terminus communis supponit determinate, requiritur quod verificetur pro aliquo supposito determinato", *De Puritate*, p. 20.

18. "Confusa suppositio est acceptio termini communis pro pluribus mediante signo universali", *Tractatus*, pp. 82-3.

19. "[contingit descendere] per propositionem de disiuncto praedicato, et contingit eam inferri ex quocumque singulari", *Summa Logicae* p. 211.

20. "quia omne copulans significat in adiacentia ad substantiam et sic copulatur personaliter. Item omne copulans est nomen accidentis. Sed omne nomen accidentis est commune. Ergo nulla copulatio est discreta", p. 81.

21. Remember that ‘planet’ included the Sun and the Moon in the middle ages, that is, any heavenly body with an irregular motion.

22. "’album potest esse nigrum’; ista significat quod illud quod est album, vel potest esse album, potest esse nigrum", *Perutilis Logica* II 10 f. 15va, ed. Muñoz García §581.

23. "haec est vera ‘album potest esse nigrum’, quia haec est possibilis ‘hoc est nigrum’, demonstrando aliquid pro quo ‘album’ supponit; et tamen haec est impossibilis ‘album est nigrum’", p. 278.

24. "Sciendum autem quod appellatio dicitur quatuor modis ... Quarto modo dicitur appellatio acceptio termini pro supposito vel pro suppositis actu existentibus, et de appellatione isto quarto modo dicta est presens intentio", *Logica Lamberti*, pp. 211-2.

25. "Unde sicut suppositio stricte accepta est proprietas subiecti, prout comparatur ad praedicatum, ita appellatio est proprietas praedicati comparati ad subiectum sive ad inferius", *De Puritate*, p. 47.

26. "terminus communis univocus appellat sua inferiora sed non significat sua inferiora", *loc.cit.*

27. "in ista ‘omnis homo videt se’ li se supponit pro omni homine confuse et distributive immobiliter et singillatim, quia non contingit descendere non variando aliud extremum ... sic dicendo ‘omnis homo videt se, igitur Sortes videt Sortem’", *Summa Logicae* I 76, p. 235.

Stephen Read
slr@st-and.ac.uk

First published: February 5, 2002
Content last modified: February 5, 2002

A Proof Concerning Bradwardine's Theory

Here is a proof that on Bradwardine's theory, every proposition signifies that it is true. Let ‘P’ name the proposition replacing *’p.’* Then:

1. *p*	Assume for conditional proof.
2. P signifies that *q*	Assume for conditional proof.
3. $p \rightarrow q$	From 2 and the thesis that what propositions signify follows from them.
4. *q*	From 1, 3, and *modus ponens*
5. (P signifies that *q*) $\rightarrow q$	From 2-4, by conditional proof.
6. P is true.	From 5 and Bradwardine's definition of truth, since 5 is general with respect to propositions replacing ‘*q*’.
7. $p \rightarrow$ P is true	From 1-6, by conditional proof.
8. P signifies that P is true.	From 7 and "Bradwardine's Principle."

Paul Vincent Spade
spade@indiana.edu

Return to note 19

First published: August 27, 2001
Content last modified: August 27, 2001

Nicholas of Autrecourt [de Altricuria; Autricuria, Ultricuria; Autricort]

The most striking feature of Autrecourt's academic career is his condemnation in 1347. In almost every history of medieval philosophy, his censure is presented as one of the most important events in fourteenth-century Paris. In the older literature, Autrecourt's views have become linked to allegedly skeptical tendencies in scholastic thought, and have been unduly shadowed by assumptions about their relation to the views of William of Ockham. Over the last two decades, however, it has become apparent that the study of Autrecourt's thought has been wrongly placed in the larger context of the battle against Ockhamism at the University of Paris in the years 1339-1347. Although Autrecourt was no skeptic -- on the contrary, he attacked the "Academics" or ancient Skeptics -- his philosophical stance challenges the prevailing Aristotelian tradition. In particular, Autrecourt rejected some of the main tenets of scholastic metaphysics and epistemology, such as the substance-accident structure of reality and the principle of causality.

1. Life

As is the case with many medieval thinkers, Autrecourt's biographical details are few. What we know about his intellectual life has to be reconstructed using dates that are attached to the handful of documents in which he is mentioned. One of these is a record from sometime between 1333-36,

indicating that he served as prior at the Collège de Sorbonne. Another important document is a papal letter of 1338 in which Benedict XII confers upon him the function of canon at Metz Cathedral, and refers to him as a master of arts and bachelor of theology and civil law. Evidently, however, Autrecourt did not claim his prebendary stipend until after his trial in 1347.

On the basis of such references, Autrecourt's date of birth can be placed sometime between 1295-98. He originated from Autrécourt in the diocese of Verdun, and was probably a student in the arts faculty at Paris, belonging to either the English, or, more likely, the French nation. His master's degree in arts can be dated around 1318-20. While a student, he must have come across such famous masters as John of Jandun, Marsilius of Padua, Thomas Wilton, Walter Burley, Bartholomew of Bruges, or Siger of Courtrai. Since his law degree was in civil rather than canon law, he must have left Paris at some point for a minimum of five years, probably to study at Orléans, Avignon, or Montpellier.

His membership in the Collège de Sorbonne places Autrecourt back in Paris in the 1330s as a student of theology. On November 21, 1340, Pope Benedict XII summoned him from Paris to Avignon to respond to allegations of false teaching. In his letter, the pope refers to Autrecourt as a *licentiatus* in theology, meaning that Autrecourt had fullfilled the formal requirements for the theology degree, e.g., lecturing on the Bible and the *Sentences*. But does it also mean that he was a full-fledged master of theology? The question is controversial. In the judgment at his trial, it was stipulated that Autrecourt could only obtain "magisterial honor and degree" after special permission from the Apostolic See, which seems to imply that he was not allowed to progress to inception in theology (the ceremony in which the magisterial honors would be conferred) until the pope decided otherwise. Moreover, there are no records referring to Autrecourt as a master in theology. This suggests that Autrecourt remained a *licentiatus* in theology when he moved on to Metz to take his position as canon (and later dean) of the cathedral chapter. He died in 1369, on either July 16 or 17.

2. Autrecourt's Trial and Conviction

Autrecourt's time of trial began in 1340 when he was first summoned to appear before the papal court in Avignon and lasted until his conviction in 1346. An extensive, though as yet incomplete, dossier of the judicial process at Avignon has been preserved in the form of an *instrumentum publicum* (actually, a draft copy thereof), which served as a model for the preparation of the official record of the process. The papal dossier contains copies of a number of records that played a role during earlier stages of Autrecourt's trial, and gives an account (*narratio*) of the judicial proceedings from the moment Cardinal Curti, the judge, took over the investigation.

The record specifies the charges and summarizes the false teachings of which Autrecourt was accused in the form of four lists, together totalling 66 erroneous propositions or "articles" (*articuli)*. The articles were culled from Autrecourt's writings and oral teachings. On the basis of this record, it would appear that the papal commissions of Pope Clement VI and Cardinal Curti used evidence from earlier proceedings at the University of Paris and Autrecourt's response to this evidence to reach their verdicts. If this scenario is correct, it raises two obvious questions: why was Autrecourt's trial transferred from

Paris to Avignon, and how did it begin in the first place? Unfortunately, the surviving historical evidence is insufficient to answer either question.

The commission of prelates and theologians, which under the chairmanship of Cardinal Curti had discussed all of the articles attributed to Autrecourt, came to the conclusion that they contained many false, dangerous, presumptuous, suspect, erroneous, and heretical statements. For this reason, Autrecourt's writings were ordered burned either at Pré-aux-Clercs or Pré-de-Saint-Germain at Paris at an unspecified future date. Moreover, Autrecourt was ordered to publicly recant several of the articles specified in the legal record. These recantations and declarations, which Autrecourt was first required to make at the palace of Cardinal Curti in Avignon, had to be repeated at the University of Paris. Autrecourt's recantation at the papal court took place before May 19, 1346. The exact date is unknown because it was left blank in the draft prepared by the notary Bernard. In addition to the recantation, Autrecourt was declared unworthy to ascend to the magisterial rank in the theological faculty. Anyone in possession of the authority to present or promote Autrecourt to the magisterium of the faculty of theology was thereafter forbidden to do so.

The Parisian part of the sentence was fulfilled the following year. On November, 20, 1347 the regent and non-regent masters of the university met at the Church of Saint-Mathurin, where papal letters and the process "concering certain articles" were read. This material had been brought from Avignon by Autrecourt himself. On November 25, Autrecourt recanted the four confessed articles and the articles from the letter "*Ve michi*" in the Church of the Dominicans, and publicly declared that the propositions contained in the other two lists were wrong. In addition, he burned these articles and a treatise, most likely the *Exigit ordo*. The public reading of the *instrumentum* and the recantation served an important purpose. It not only made the sentence effective, but also informed the scholarly community of Autrecourt's errors and of the punishments set out in the *instrumentum*, which they would incur if they were to teach the censured errors. Years later, scholars such as John Buridan, Marsilius of Inghen, and André of Neufchâteau (Andreas de Novo Castro) cited the condemned erroneous propositions as the *articuli cardinalis (albi)*.

3. Writings

Autrecourt's oeuvre is not large. There is a correspondence with the Franciscan theologian Bernard of Arezzo, and with a certain master Giles (possibly Giles of Feno), and a treatise that has come to be known as the *Exigit ordo*. Furthermore, we have a theological question dealing with the intension and remission of forms and the problem of minima and maxima (*utrum visio alicuius rei naturalis possit naturali intendi* [Could the vision of any natural thing be naturally intensified?]).

Autrecourt wrote nine letters to Bernard of Arezzo, only two of which survive. In addition, there is one letter from master Giles addressed to Autrecourt, along with a brief response by the latter, which, however, breaks off in mid-sentence. The correspondence has been preserved in two manuscript copies from the intellectual milieu of the Collège de Sorbonne. Together the letters form a small dossier, the central item of which is the letter from Master Giles. Apparently, the only reason the two letters to

Bernard were copied was because they are mentioned in the letter from Master Giles. The correspondence between Autrecourt and Bernard of Arezzo is much earlier, dating from the time when both were theology students, engaged as opponents in each other's *Principia*, i.e., inaugural lecture on the *Sentences*. They can be dated between October 1335 and June 1336, although both *Principia* are now lost. There is no evidence that Autrecourt ever actually wrote a commentary on the *Sentences*, which, in any case was not a formal requirement for obtaining a degree. The theme of the discussion in the *Principia* and the letters to Bernard of Arezzo is the validity of Aristotle's principle of non-contradiction, as presented in Book IV of the *Metaphysics*.

The *Exigit ordo* is the fruit of Autrecourt's teaching in the arts faculty. Instead of expounding his views in commentaries on Aristotle's texts, Autrecourt chose to write an independent treatise discussing issues pertaining to natural philosophy, metaphysics, epistemology, philosophical psychology, and ethics, and engaging in debate with unnamed contemporaries. The work was completed in the years 1333-35, at which time Autrecourt was preparing his commentary on the *Sentences*. For financial reasons, Autrecourt taught in the arts faculty while he was enrolled as a student in theology.

The *Exigit ordo* is also known as the *Tractatus universalis* (Universal Treatise). The latter title is actually a misreading of the first two words of the treatise, "*tractatus utilis*" (useful treatise). It has been preserved in a single manuscript copy, which, like the Giles letter, breaks off in mid-sentence. It is divided into two prologues, two treatises, and several chapters, which, unfortunately, the scribe has placed in the wrong order. The Latin edition and English translation both preserve the order of the medieval manuscript without correction.

The theological question is a report (*reportatio*) of a theological disputation in which Autrecourt served as respondent to the objections. Although the presiding master of a disputation would usually have to be considered its real author, matters may be different here. Since it is a *reportatio* -- i.e., a text which, unlike an *ordinatio*, was not subjected to later editing by the master himself -- Autrecourt's views probably appear in unadulterated form. The question was disputed between 1336-1339, and has been little studied by scholars.

4. Epistemology

Central to Autrecourt's teaching is the view that all *evident* knowledge (with the exception of the certitude of faith) must be reducible to the first principle (*primum principium*), i.e., to the principle of non-contradiction. An inference yields evident knowledge only when the affirmation of its antecedent and the negation of its consquent are contradictory. This means that the antecedent and the consequent, or, more precisely, what is signified by the antecedent and the consequent, must be identical, "because if this were not the case, it would not be immediately evident that the antecedent and the opposite of the consquent cannot stand together without contradiction." It is in the context of this theory that Autrecourt launches an attack on our claim to have certain knowledge of the existence of substances and causal relations. If A and B are two distinct entities, he says, one cannot infer with certainty (knowledge of) the existence of A from that of B or vice versa, for the affirmation of the one and the denial of the other does

not result in a contradiction. On the basis of this principle, one may not infer (knowledge of) the existence of effects from knowledge of their causes, nor (knowledge of) the existence of substances from knowledge of their accidents.

This view runs contrary to the Aristotelian position, according to which causal relations really exist and are discoverable by means of induction, so that the existence of substances can be inferred from the perceptible accidents inhering in them. The upshot of Autrecourt's view is that we do not have experience of causal relations or substances, nor does logic provide certain knowledge of them. There are no logical reasons to assume that there is an evident relation between a cause and an effect, or between a substance and an accident.

The position outlined above is developed in Autrecourt's correspondence. It has led historians of philosophy to characterize him as the most important, if not the only, "real" representative of medieval scepticism, as "the medieval Hume", to use Hastings Rashdall's epithet. On closer inspection, however, it turns out that Autrecourt's scepticism is reserved for rationalist claims about the truth of our commitments to causality and substance, concepts for which we have no empirical proof. He is not a sceptic at all when it comes to defending the reliability of sense-perception.

In his *Letter to Bernard*, Autrecourt takes on Bernard of Arezzo, who had argued that the intellect is neither certain of the existence of those things of which it has a clear intuitive cognition, nor of its own acts. Autrecourt reveals the full implications of this position by pointing out to Bernard that "you are not certain of those things which are outside of you. And so you do not know if you are in the sky or on earth, in fire or in water...Similarly, you do not know what things exist in your immediate surroundings, such as whether you have a head, a beard, hair, and the like." He concludes that Bernard's stance is even worse than that of "the Academics," that is, the ancient Skeptics.

5. Metaphysics

To Bernard's skeptical challenges Autrecourt replies that sense experience is reliable. This theme is not further developed in the letters to Bernard, however. For discussion of this topic we must turn to the *Exigit ordo*. In one section of this treatise, which is reminiscent of Aristotle's *Metaphysics* IV, 5, Autrecourt addresses one of the central issues of metaphysics, namely the relation between appearance and reality. He addresses Protagoras' view that whatever is apparent is true: *An omne illud quod apparet sit*? (Does everything that appears exist?).

Autrecourt defends the thesis that what appears, is, and that what appears true, is true. He finds this view more plausible than its opposite, viz., that the intellect cannot possess certitude. His concept of appearance plays a key role in his doctrine of certain knowledge. It is used in a phenomenological sense, to describe perceptual experiences. According to Autrecourt, the intellect is certain of everything that is evident to it in the final analysis. This is the case for everything that appears in a proper sense (*apparet proprie*), i.e., that appears clearly in an act of the external senses *(in actu sensuum exteriorum*). He identifies appearances with the objects of immediate sensory experience, which are considered evident.

In this way, he implies that sense perception is a reliable source of truth, i.e., that the apparent properties of an object are its actual properties.

But is sense perception reliable? Perceptual errors and dreams seem to indicate that things are not always as they seem. Autrecourt discusses several sceptical doubts (*dubia*), versions of what would later be called the "argument from illusion" and the "argument from dreaming". These arguments work from the common sense assumption that things often appear to be other than they are: e.g., sweet food can appear bitter, a white object can appear red, in sleep it can seem to someone that he is flying through the air or fighting the Saracens. How does Autrecourt respond to these sceptical doubts?

By distinguishing between appearance and judgment. Appearances are always veridical: experience cannot be other than it is. However, judgments made from experience can be erroneous, especially if they are based on images rather than on what is perceived "in the full light." In other words, Autrecourt denies any conflict of appearances. Those not "in the full light" are not in themselves misperceptions because the experiences themselves are not illusory. They merely fail to give us the real properties of the object perceived. Potential conflict creeps in at the level of judgment, where ontological claims are made on the basis of appearances. Only those appearances that are "in the full light" reveal the true properties of the perceived object, and only they can provide the basis for true judgments. Appearances of objects that do not come to the perceiver "in the full light" are incomplete or contaminated, as if the observer were looking into a mirror. In other words, Autrecourt carefully distinguishes between ‘x appears F’ from ‘x is F’, for even if x is not really F, it can still appear F and cause someone to believe that it is F. In this way, illusions and dreams turn into mistaken beliefs. Only clear appearances (*apparentiae clarae*) can cause veridical judgments.

A final topic taken up by Autrecourt in this context is the problem of the criterion: How can one discriminate between appearances that provide the basis for true judgments and those that do not? Like Aristotle, he holds that appearances from what we perceive under "normal" conditions cause true judgments. Also like Aristotle, he asserts that there is no further proof that the criterion on which the distinction between veridical and false judgments rests is correct. Both dismiss worries about the justification of the criterion as absurd. In the words of Autrecourt: "One must accept as true what appears in the full light. Now, concerning the minor premise of this argument, how can you have certainty? ... One way of answering this would be to say that there is no way of proving the conclusion, but that the concept of certitude which is present comes as a certain natural consequence, and not as a conclusion. An example, among others, is that white and black are different. This concept of their difference is not gotten by way of conclusion."

6. Natural Philosophy

The point of departure for Autrecourt’s physics is a thesis which strikes him as more probable than what he finds in Aristotle’s *Physics*, namely that all things are eternal. One of the implications of this thesis is that there is no generation or corruption in the universe, which seems to conflict with the way in which properties begin and cease to exist in their subjects, e.g., when something white turns black, so that the

whiteness no longer exists.

According to Autrecourt, no genuine corruption has taken place. It is just that the natural form is no longer visible, since it has been dispersed and divided into its smallest units. In other words, Autrecourt attributes the apparent generation and corruption of things to the motions of atoms. In keeping with this atomistic view, he also holds that space and time consist of indivisible units, viz., points and instants, respectively.

This atomistic explanation of generation and corruption only holds true, he states, if motion is not distinct from the mobile object. For if motion were another distinct thing, it would constitute the generation of something new, whereas rest (the destruction of motion) would constitute its corruption. When seen from this perspective, local motion would refute the eternity of the universe. For this reason, Autrecourt finds it necessary to examine more closely the ontological status of motion.

Autrecourt argues that motion is not a thing distinct from the moving object. Following Ockham, he rejects the idea that motion is a positive thing inhering in the mobile object. Thus, the loss of motion should not be described as the destruction or corruption of an entity, and the eternity doctrine is saved.

7. Semantics

Autrecourt did not leave any logical writings, nor does he discuss logic or semantics in the *Exigit ordo* or in his correspondence. However, from his theological question and a few of the censured articles it is clear that he was familiar with the logical debates of his time. According to one of the articles, Autrecourt claimed that the proposition "Man is an animal" is not necessary according to the faith, because in that sense one does not attend to the necessary connection between its terms. The article should be seen against the background of the sophism "Man is an animal", which received considerable attention in the thirteenth and fourteenth centuries. It served to clarify the relation between meaning (*significatio*) and reference (*suppositio*) by investigating the verification of propositions concerning empty classes. Would the proposition "Man is an animal" still be true if no man exists?

Five other articles that turn up in Autreourt's condemnation concern the *complexe significabile*, or what is signified by an entire proposition. According to adherents of the doctrine such as Adam Wodeham and Gregory of Rimini, the object of knowledge is not the proposition, or the things (*res*) referred to in the external world, but "that which is signified" by the proposition (*complexe significabile*). One of the problems raised by this theory concerned the ontological status of the *complexe significabile*: Is it something (*aliquid*) or nothing (*nihil)*? Echos of this and other debates can be found in these articles.

Bibliography

Editions and Translations

- Edition of the *Exigit Ordo* and the theological question "*Utrum visio alicuius rei naturalis possit naturali intendi*" in: O'Donnell, J. R., "Nicholas of Autrecourt," *Mediaeval Studies* 1 (1939), 179-280.
- English translation of the *Exigit Ordo* in: Nicholas of Autrecourt, *The Universal Treatise*, tr. Leonard A. Kennedy, Richard E. Arnold, and Arthur E. Millward, with an Introduction by Leonard A. Kennedy, Milwaukee: Marquette University Press, 1971.
- First edition of the correspondence and the condemned articles in: Lappe, J., *Nicolaus von Autrecourt, sein Leben, seine Philosophie, seine Schriften*, Münster: Aschendorff, 1908 (*Beiträge zur Geschichte der Philosophie des Mittelalters*, 6.2) (now superseded by more recent editions).
- Edition and English translation of the correspondence in: *Nicholas of Autrecourt, His Correspondence with Master Giles and Bernard of Arezzo*: A Critical Edition and English Translation by L. M. de Rijk. Leiden: E. J. Brill, 1994.
- Edition and German translation of the correspondence in: Imbach, R, and D. Perler, *Nicolaus von Autrecourt: Briefe*, Hamburg: Meiner, 1988.

Studies

- Dutton, B.D., "Nicholas of Autrecourt and William of Ockham on Atomism, Nominalism, and the Ontology of Motion, *Medieval Philosophy and Theology* 5 (1996), 63-85.
- Kaluza, Z., *Nicolas d'Autrecourt. Ami de la vérité*, in: *Histoire littéraire de la France*, vol. 42, fasc. 1. Paris, 1995.
- Rashdall, H. 1907 "Nicholas de Ultricuria, a Medieval Hume," *Proceedings of the Aristotelian Society* N.S. 8 (1907), 1-27.
- Scott, T.K., "Nicholas of Autrecourt, Buridan, and Ockhamism," *Journal of the History of Philosophy* 9 (1971), 15-41.
- Tachau, K.H., *Vision and Certitude in the Age of Ockham. Optics, Epistemology and the Foundations of Semantics, 1250-1345*. Leiden: Brill Publishers, 1988.
- Thijssen, J.M.M.H., *Censure and Heresy at the University of Paris, 1200-1400*, Philadelphia: University of Pennsylvania Press, 1998.
- -----, "John Buridan and Nicholas of Autrecourt on Causality and Induction," *Traditio* 43 (1987), 237-255.
- -----, "The 'Semantic Articles' of Autrecourt's Condemnation," *Archives d'histoire doctrinale et littéraire du moyen âge*, 65 (1990), 155-175.
- -----, "The Quest for Certain Knowledge in the Fourteenth Century: Nicholas of Autrecourt against the Academics," in: J. Sihvola (ed.), *Ancient Scepticism and the Sceptical Tradition*, Helsinki: Societas Philosophica Fennica, 2000, 199-223 (*Acta Philosophica Fennica*, 66).
- Weinberg, J. R., *Nicolaus of Autrecourt. A Study in 14th Century Thought*. New York: Greenwood Press, 1969 (reprint of Princeton: Princeton University Press, 1948).
- Zupko, J., "Buridan and Skepticism," *Journal of the History of Philosophy* 31 (1993), 191-221.
- -----, "How It Played in the rue de Fouarre: The Reception of Adam Wodeham's Theory of the *Complexe Significabile* in the Arts Faculty at Paris in the Mid-Fourteenth Century," *Franciscan Studies* 54 (1994-97): 211-225.

- -----, "On Certitude," in: J.M.M.H. Thijssen and Jack Zupko (eds.), *The Metaphysics and Natural Philosophy of John Buridan*, Leiden-Boston-Köln, Brill, 2001, 165-82.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Buridan, John [Jean] | Burley [Burleigh], Walter | Marsilius of Inghen | Ockham [Occam], William

Johannes M.M.H. Thijssen
University of Nijmegen
hthijssen@phil.kun.nl

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 14, 2001
Content last modified: October 14, 2001

Peter of Spain (Petrus Hispanus)

Peter of Spain (thirteenth century), exact identity unknown, was the author of a standard textbook on logic, the *Tractatus* (Tracts),[1] which enjoyed a high renown in Europe for many centuries. His works on logic are typical examples of the type of manuals that gradually started to emerge within the context of twelfth- and thirteenth-century teaching practices. Until recently he was also identified as the author of a number of extant works on medicine.

Life and works: Some Comments on the Historiography

Peter of Spain has been established as the medieval author of a work that became widely known as *Summule logicales magistri Petri Hispani* (Collection of Logic Matters of Master Peter of Spain). The great number of manuscripts and printed editions is evidence of the enormous success this work met with throughout European universities well into the seventeenth century. But finding out the true identity of the author of this influential *Tractatus* has proved to be a difficult task. For a long time it was assumed that he was a Portuguese who became Pope in 1276, under the name of John XXI. There is also another, earlier tradition, according to which the author of the *Tractatus* was regarded as Spanish, and a member of the Dominican order. Yet another attribution, dating from the fifteenth century, was to a Petrus Ferrandi Hispanus (d. between 1254 and 1259), which would be consistent with the idea that the work originated from the first half of the thirteenth century. According to still another attribution, the *Summule* was compiled by a Black Friar no earlier than in the late thirteenth or early fourteenth century.

The ‘Dominican-thesis’ can be divided into three traditions:

1. The general view that Peter of Spain, author of the *Tractatus*, is someone who belonged to Order of Black Friars,
2. The more specific view that the author of the *Tractatus* was a frater Petrus Alfonsi Hispanus O.P.,
3. Another specific view that the Peter of Spain who created the *Tractatus* was the same Peter of Spain as the one who wrote the *Legenda sancti Dominici* and the Office of the Saint's Feast, namely Petrus Ferrandi Hispanus O.P., who died in the 1250's.

Current research on the identity of Peter of Spain has once again taken up the idea that he must have been a member of the Dominican Order instead of Pope John XXI.[2] However, we are still in the dark about the true identity of Peter of Spain. The most recent information we have on this score is that four of the Dominican candidates recently suggested as the author of the *Tractatus* can be deleted from the list.[3] The lack of further information also makes it difficult to establish the dates and specifics of his carreer.

The work that has enjoyed such enormous success, the *Tractatus*, is believed to have been written between 1230 and 1245 and has universally been recognised as a work by Peter of Spain. Another work that has been identified as Peter of Spain's is a *Syncategoreumata* (Treatise on Syncategorematic Words), probably written sometime between 1235 and 1245.[4] Considering the fact that in all the thirteenth-century manuscripts the *Syncategoreumata* directly follow the *Tractatus*, and the number of similarities between doctrinal aspects of these two works on logic, it is almost certain that they were written by the same author. Both works seem to have originated from Southern France or Northern Spain, the region where we also find the earliest commentaries on these treatises.

Besides these works on logic, there are other works that have been written by a Peter of Spain. In the *Petrus Hispanus papa* tradition, he is the supposed author of a famous medical work *Thesaurus pauperum*, as well as fourteen other works on medicine. Other works written by (a) Peter of Spain are a *Scientia libri de anima*, and commentaries on Aristotle's *De anima*, *De morte et vita* and *De sensu et sensato*, and commentaries on works by pseudo-Denys the Areopagite. As yet there is no certainty about whether the Peter of Spain who wrote these works is the author of the *Tractatus* and the *Syncategoreumata*, or about the dates of their origin.

Another Peter of Spain, referred to as *Petrus Hispanus non-papa*, has been identified as the author of the *Summa ‘Absoluta cuiuslibet’*, a late twelfth-century handbook on syntax closely linked with Priscian's *Institutiones grammatice, libb.* XVII and XVIII, which became very popular later in the Middle Ages under the name *Priscianus minor*.[5] The chronology of this work seems to rule out that this Peter of Spain is the same author as the author of the *Tractatus*.

Origins of Peter's Works on Logic

Peter's logic has its origin in the continental tradition. The educational carreer of the *Tractatus* appears

from two commentaries,[6] which contain short lemmata and a number of questions (*questiones*) together with their solutions. The tracts as contained in these texts are very similar to the the ones in the *Tractatus*. Typical of the Paris tradition is the separate treatment of ampliation, restriction and distribution, and several other, doctrinal features.[7] Peter's Masters include Johannes Pagus (who is supposed to have been a Master of Arts in Paris in the 1220's) and Hervaeus Brito (who may have been a Master of Arts either before 1229, but possibly later, in which case he does not qualify as a teacher of Peter's). Besides these direct influences, the sources for Peter's works on logic can be traced back to Boethian-Aristotelian logic, and authorities in the field of grammar such as Priscian and Donatus.

Like the *Tractatus*, the *Syncategoreumata* also displays a continental origin, and appears to have continued along the lines of a similar work by Johannes Pagus (which has been dated between 1225 and 1235), later on further developed by Nicholas of Paris (who wrote his *Syncategoreumata* between 1240 and 1250).[8]

The *Tractatus*

The *Tractatus* can be divided into two main parts. One part deals with doctrines found in the so-called *logica antiquorum* -- i.e. the *logica vetus* (old logic) and *logica nova* (new logic) -- and the other contains doctrines covered by the *logica modernorum* -- viz. the tracts that discuss the *proprietates terminorum* (properties of terms).

The first main part of the *Tractatus* divides into five tracts. The first tract, *De introductionibus* (On introductory topics) explains the concepts used in traditional logic -- *nomen* (noun), *verbum* (verb), *oratio* (phrase), *propositio* (proposition) -- and presents the divisions of and the (logical) relationships between propositions. The second tract, *De predicabilibus* (On the predicables) covers matters dealt with in Boethius's accounts of Porphyry's *Isagoge*. It gives an account of the concept *predicabile* and the five predicables -- *genus*, *species*, *differentia*, *proprium*, *accidens* -- i.e., the common features of and differences between the predicables, as well as of the terms '*predicatio*' and '*denominativum*'. Tract three, *De predicamentis* (On the categories), discusses the ten Aristotelian categories, as well as some items already dealt with in the previous treatise. The fourth tract, *De sillogismis* (On syllogisms) mainly goes back to Boethius's *De syllogismis categoricis* (On categorical syllogisms). It gives an explanation of the basic element of the syllogism, i.e. *propositio*, and of the syllogism, and then goes into mood and figure, the proper forms of syllogisms, and briefly deals with what are called paralogisms. The fifth tract, *De locis* (On topical relationships), is derived from Boethius's *De topicis differentiis* (On different topical relationships) I and II. This tract starts off with an explanation of the notions *argumentum* and *argumentatio*, and then proceeds to deal with the species of argumentation: syllogism, induction, enthymeme, and example. Next, it gives a definition of *locus* (the Latin translation of the Greek *topos*): a *locus* is the seat of an argument (i.e., the *locus* is supposed to warrant the inference by bringing it under some generic rule.) The intrinsic *loci* (= the kind of *locus* that occurs when the argument is derived from the substance of the thing involved) are covered first, followed by the extrinsic *loci* (= the kind of *locus* that occurs when the argument is derived from something that is completely separate from the substance of the thing involved) and intermediary *loci* (= the kind of *locus* that occurs when the argument is taken

from the things that partly share in the terms of the problem and partly differ from it). Examples are: intrinsic -- the *locus* "from definition": ‘a rational animal is running; therefore a man is running’; extrinsic -- the *locus* "from opposites": ‘Socrates is black; therefore he is not white’; intermediary -- ‘the just is good; therefore justice is good’.

The second part of the *Tractatus* comprises subjects that were of major importance in the doctrine of the properties of terms. In the sixth tract, *De suppositionibus*, the theory of supposition is dealt with. The treatise begins with an exposition of *significatio*. The definition of *significatio* runs: *significatio* is the respresentation of a thing by means of a word in accordance with convention. Next it gives a definition of the related terms *suppositio* and *copulatio*, and the differences between the terms *significatio*, *suppositio* and *copulatio*. Of these three *suppositio* and *significatio* are the most important in Peter's semantics. *Suppositio* is defined as the acceptance of a substantive verb for some thing. *Suppositio* is dependent on *significatio,* because supposition can only occur via a term that already has some *significatio*. Put in other words, *significatio* pertains to a word by itself, and *supposition* to a term as actually used in some context.

The tract concludes with a division of *suppositio*. The first division is into *suppositio communis* (common supposition) and *suppositio discreta* (discrete supposition) -- e.g. the terms *homo* (man) and *Sortes* (Socrates) respectively.

The second division, *suppositio communis,* is divided into *naturalis* (natural) and *accidentalis* (coincidental). *Suppositio naturalis* is described as the acceptance of a common term for all those things that can share in the common universal nature signified by the term in question -- e.g. *homo* (‘man’) taken by itself by its very nature is able to stand for all men, whether in the past, present or future; *suppositio accidentalis* is the acceptance of a common term for those things for which the term in question requires an additional term -- e.g. in *homo est* (‘A man is’) the term *homo* stands for present men, whereas in *homo fuit* (‘A man has been’) and in *homo erit* (‘A man will be’) it stands for past men and future men respectively, owing to the additional terms *fuit* and *erit*.

The third division, *suppositio accidentalis,* is divided into *suppositio simplex* (simple supposition) and *suppositio personalis* (personal supposition). *Suppositio simplex* is the acceptance of a term for the universal ‘thing’ it signifies, as in *homo est species* (‘Man is a species’, *animal est genus* (‘Animal is a genus’), in which the substantive terms *homo* and *animal* stand for the universal man and animal, and not any one of their particulars. *Suppositio simplex* can occur both in the subject- and in the predicate-term -- e.g. *homo est species* (‘Man is a species’) and *omnis homo est animal* (‘Every man is an animal’) respectively. *Suppositio personalis* is the acceptance of a common term for one or more of its particulars, as in *homo currit* (‘A man is running’).

The fourth division, *suppositio personalis,* is subdivided into either *derterminata* (determinate = standing for a certain particular) or *confusa* (confused = standing for any individual falling under that name). *Suppositio determinata* occurs when a common term is taken indefinitely or in combination with a particular sign -- e.g. *homo currit* (‘Man is running’) or *aliquis homo currit* (‘A /some man is running’). *Suppositio confusa* occurs when a common term is taken in combination with a universal sign (’Every man is running’).

The tract on supposition winds up with the discussion of a few questions regarding the attribution of supposition in a few cases.

The seventh tract of the *Tractatus*, on fallacies, which forms part of the Aristotelian-Boethian logic, is written in the tradition of the *Fallacie maiores* (Major fallacies). The eighth tract, *De relativis* (On relatives) deals with the relative pronouns as defined by Priscian in his *Institutiones grammaticae*. The relative pronouns are devided into: relatives of substance, such as *qui* (who), *ille* (he), *alius* (another), and relatives of accident, such as *talis* (of such a kind), *qualis* (of what kind), *tantus* (so much), *quantus* (how much). The former are subdivided into relatives of identity (*qui* and *ille*) and relatives of diversity (such as *alter* and *reliquus*, both of which can be translated as ‘the other’). The relative of identity is defined in terms of supposition as what refers to and stands for the same thing. These relatives are either reciprocal or non-reciprocal. With regard to the relatives of identity, Peter adds a dicussion of a number of questions about the rationale for using demonstrative pronouns, and some problems concerning how the fallacy of a relative having two diverse referents comes about.

The tract on relatives continues with a brief discussion on the relatives of diversity, accompanied by a rule about the supposition of the relative when it is added to a superior and an inferior in a premiss and a conclusion, as in *aliud ab animali*; *ergo aliud ab homine* (‘Something other than an animal; therefore something other than a man’). With regard to relatives of identity a rule of the "ancients", who deny that a proposition introduced by a relative can have a contradictory opposite, is discussed and rejected. Another rule is given about the identity of supposition of a non-reciprocal relative and what it refers to. The tract concludes with short accounts of relatives of accident.

The ninth, tenth, eleventh, and twelfth tracts of the *Tractatus*, i.e. the short tracts *De ampliationibus* (On ampliation), *De appellationibus* (On appellation), *De restrictionibus* (On restriction) and *De distributionibus* (On distribution) are in fact elaborations of the theory of supposition. Ampliation is an extension of the supposition of a term. It occurs when an expression is combined with a modal term -- e.g. *homo potest esse Antichristus* (‘A man can be the Antichrist’), and *homo necessario est animal* (‘A man is necessarily an animal’) -- in which case the supposition of the the term ‘man’ is extended to more than just individuals existing in the present. The tract on *appellationes* is very short: appellation is considered no more than a special case of restriction, i.e. the restricted supposition brought about by a present-tense verb. In this tract the rules of appellation are in fact specific kinds of rules of restriction. The subject of restriction in general is discussed in the eleventh tract. The rules of restriction are the same ones as were presented in the early Parisian textbooks on logic.[9] The final tract, on distribution, deals with the multiplication of common terms as a result of their being combined with universal signs. These universal signs are either distributive of substance (such as *omnis*, *nullus*), or of accidents (such as *qualiscumque*, *quantuscumque*). In this description ‘substance’ is defined as substistent modes of being, and ‘accident’ as accidental modes of being. Separate attention is given to the universal sign *omnis* (‘all’ or ‘every’) along with a discussion of the common rule that the use of *omnis* requires three *appellata* (particular things). The most frequently cited example in these discussions in the thirteenth century was the sophisma *omnis phenix est* (‘Every phoenix is’). According to Peter of Spain, the use of *omnis* does not call for at least three *appellata*; an exception to this rule is found in cases in which there is only one *appellatum*, as

is the phoenix-case. The tract also pays attention to a number of tongue-twisting sophisma-sentences.

The *Syncategoreumata*

Peter's treatise on syncategorematic words forms part of a separate genre that developed from the beginning of the thirteenth century. The term *syncategorema* comes from a famous passage of Priscian in his *Institutiones grammatice* II , 15, in which a distinction is made between two types of wordclasses (*partes orationis*) distinguished by logicians, viz. nouns and verbs on the one hand, and *syncategoremata*, or *consignificantia*, on the other. The latter are defined as words that do not have a definitive meaning on their own, but acquire one only in combination with other, categorematic words.

Like the treatises of the *Tractatus* kind, the *Syncategoreumata* were developed from the (twelfth-century) theories on fallacies, as well as from grammatical doctrines (from the same period). From the second half of the twelfth century, there was a growing interest in the linguistic elements that are considered to lie at the basis of ambiguity and fallacious reasoning. Hence the increase of treatises presenting a systematic account of these terms. The connection these treatises have with Priscian's grammar can be gathered from the attention different authors pay to the *signa quantitatis* (or quantifiers), and the fact that considerable attention is given to the meaning and function of syncategorematic terms.

The list of words to be included among the *syncategoreumata* was not always the same. Generally speaking it comprised exclusive words *tantum* (only), *solus* (alone), exceptive words such as *preter* (except), *nisi* (unless), consecutive words such as *si* (if) and *nisi* (if not), the words *incipit* (begins) and *desinit* (ceases), the modal terms *necessario* (necessarily) and *contingenter* (contingently), the conjunctives *an* (or), *et* (and), *nisi* (unless), *in eo quod* (in that), and *quin* (that not). In Peter's work we also find a discussion of the terms *quanto* ('how much' or 'as much as') *quam* ('than' or 'as') and *quicquid* (whatever). Unlike some other authors (such as William of Sherwood and Robert Bacon), his list does not include the word *omnis*.

In the opening of his *Syncategoreumata*, Peter presents his rationale for this investigation, viz. that there is a close link between the use of these kinds of words in sentences and their truth-value. His idea is that the syncategoreumata must have some sort of signification, but not the same as the categorematic words. For this special kind of signification he uses the words *consignificatio* and *dispositio*.

The first two separate chapters of the *Syncategoreumata* are devoted to the words *est* and *non*, which are said to be implied in all other syncategorematic words. Peter's account of the first word focuses on the notion of *compositio* (composition), which is explained in great detail, by looking into the signification of nouns and verbs (signifying a composition of a quality with a substance, and that of an act with a substance respectively). Considerable attention is given to the composition featuring in the verb 'is', in the form of the question of whether the composition involved can be counted among beings or not, considering the fact that it can be used to express different kinds of states of affairs. The chapter on negation introduces the important distinction between an act as conceived of or in the manner of a concept (*ut concepta sive per modum conceptus*) and as carried out (*ut exercita*).[10] Among the former

type we find the noun ‘negation’ and the verb ‘to deny’, whereas the latter is what is meant by the negative particle ‘not’. The remainder of the chapter deals with the function of the negation, which is to remove the composition found in whatever it covers, and discusses some well-known sophisma-sentences which turn on the specific function of negation.

The third chapter of the *Syncategoreumata* discusses the exclusive words *solus* and *tantum*. They are called exclusives because they carry out an exclusion, not because they signify one. An exclusion, furthermore, requires four things, namely, what is excluded, what is excluded from, the respect in which it is excluded, and the act of exclusion. The kinds of exclusion are divided into general and specific: the former involves an exclusion from something generic, whereas the latter from something specific. Questions that come up in this section have to do with the results of adding an exclusive term to different kinds of words, such as to a term falling under the category of Substance: does it exclude only other substances, or does it also exclude from things listed under another category? And what if it is added to a term listed under the category of Accident (such as colour, quantity, and so on)? The next question deals with the sorts of terms that can be meaningfully associated with an exclusion. For example, is it possible to exclude something from ‘being’ (as in ‘Only being is, therefore nothing other than being is’)? The tract proceeds with the kinds of things that can qualify for an exclusion. The fourth chapter, which deals with exceptive words, is compiled in a similar manner.

The fifth chapter is about the word *si*, which is said to signify causality in or via antecedence. The chapter also contains discussions of the kinds of consecution or consequence, problems of inference connected with the referents of terms used in consecutive sentences, and also on how to contradict a conditional sentence. Special attention is given to the problem whether from the impossible anything follows.

The chapter on ‘begins’ and ‘ceases’ is a good example of the way in which extra-logical considerations found their way into medieval treatises on logic. Thus, apart from the semantics and inferential problems connected with the use of these words in propositions, the chapter also looks into the notions of motion and time. An important part of Peter's ontological views can be gathered from chapter seven, which covers issues connected with the use of modal terms. Chapter eight discusses the signification and use of connectives, and the final chapter on syncategorematic words proper is concerned with the expressions *quanto*, *quam* and *quicquid*. A very short concluding chapter of Peter's *Syncategoreumata* deals with a somewhat isolated topic, i.e. the proper modes of response in an argument. The topics looked into are solution, the quantity and quality of syllogisms, and the ways to go about proving a syllogism.

Doctrinal Elements in Peter's Logic

One of the most important elements in Peter's logic concerns the doctrine of supposition. The theory of supposition has its origins in the twelfth century, when the medievals showed a growing interest in the ways in which words function in different contexts. This way of dealing with the semantics of terms has been dubbed the "contextual approach".[11]

The primary semantic property of a word is its *significatio*, in Peter's definition, the "representation of a

thing by a word in accordance with convention". It is a natural property of a word, the presentation of some (universal) content to the mind. The *significatio* of a word depends on its imposition, i.e. the application originally given to the word in question. A word can have more than one *significatio*, if it was originally applied to two or more distinct (universal) natures.

The counterpart of *significatio,* the formal constituent of every meaning, is the word's capacity to "stand for" different things (even though its *significatio* remains the same), depending on the context in which it is used. In the early stages of the development of the theory on the properties of terms, this feature of a word was called *appellatio*. For instance, the words ‘man’ and ‘horse’ can be used to stand for different individual men or horses. But they can also stand for themselves, e.g. when they are used in sentences such as ‘man is a noun’, or ‘horse is a noun’. Moreover, their meaning can differ according as the words are used in combination with verbs of different tenses.

In the final stages of the development of the theory, the notion of supposition becomes the general label that covers all the uses of a noun (substantive or adjectival), to which other recognised properties of terms (*appellatio*, *ampliatio* and *restrictio*) are subordinated.

The theory of properties of terms shows a radical inconsistency, which has been explained as "the persistent hesitation of medieval logicians between the domains of connotation (universals) and denotation (individuals)."[12] This inconsistency runs throughout Peter's account of supposition, and comes to the fore most prominently in what he says about natural supposition (*suppositio naturalis*). The main problem is in what way the property of natural supposition is related to the term's *significatio*, which was defined as the acceptance of a word for a thing (*res*). By this definition, Peter's concept of *significatio* covers both the intension and extension of a term, the universal nature of man and the individuals that have this nature in common. *Suppositio naturalis,* on the other hand, is described as "the acceptance of a common term for all those things that can share in a common universal nature"; for example, the term ‘man’ when taken by itself by its very nature stands for all the individuals that fall under it, whether they exist in the past, present or future. From this definition and the example just presented it appears that the extensional features of *significatio* and *suppositio naturalis* overlap. The latter has been explained by interpreters as the natural capacity of a significative word to stand for something.

There is a more telling difference between *significatio* and *suppositio naturalis,* however. *Significatio* is the natural property of any significative term to represent things, owing to its original imposition, whereas a term's supposition only enters the scene when it is used. The expression "taken by itself" (*per se sumptus*) found in Peter's account of *suppositio naturalis,* does not mean that no context is required, as is the case in *significatio,* but it merely indicates that for the moment the actual context is being disregarded. The link between *significatio* and *suppositio* is the following. When some word has acquired a signification by an *impositor* (= someone who bestows a meaning upon a word), then it connotes a univeral nature or essence, and acquires a natural capacity to stand for all the actual and possible individuals that share in this common nature; it owes this capacity to its *significatio*. If, however, we disregard for a moment the actual context in which the term in question is used and look upon the term as taken by itself (*per se sumptus*), then its supposition covers its entire extension. If we take the factual context in which the term is used into consideration, then its extension becomes limited, owing to the

context. The context, or more precisely, the added significative term, can be of three kinds: the added significative term can be a predicate of a proposition in which the term at issue occurs, the added significative term can be an adjective, or the context can be of a social nature.[13]

The distinction between *significatio* and *suppositio naturalis* persisted throughout the thirteenth and fourteenth centuries. Behind it is the fundamental view that regardless of whether a word is used in some context or not, it always has a *significatum*, i.e. the universal nature or essence it signifies, which can be separated from what the word comes to mean in a specific context.

Besides *suppositio naturalis*, Peter's (and other medievals') conception of *suppositio simplex* also seems to hover between connotation and denotation. In the expression *homo est species* the term *homo* has *suppositio simplex*, but this is precisely too what the term *homo* signifies. So there scarcely seems reason to separate signification from supposition on this score. The specific use of *suppositio simplex* found in Peter of Spain and other medieval authors, as the representation of a universal nature, is rejected later on by authors such as William of Ockham. For the latter, the term *homo* in the example just given has *suppositio simplex* (for Ockham a special case of *suppositio materialis*) in that it stands for the mental concept of man.[14]

Peter of Spain's logical works reveal a realistic outlook. This can be shown from the way in which he discusses the use of the word *est* (is), his account of *suppositio simplex*, and the way he analyses the occurrence of the word ‘necessarily’ in propositions. Moreover, his conception of the consecutive expression ‘if’ clearly shows his tendency to put the domains of reality and language on a par.

In his *Syncategoreumata*, Peter analyses the significative function of the word ‘is’. To a certain extent his findings are not confined to that term alone, but cover all verbs, in which the verb ‘is’ is always understood. The most remarkable feature about his discussion of ‘is’ is his focus on the notion of composition. What he is particularly interested in is the kinds of things affirmative propositions featuring that verb can refer to, in his words, the type of composition involved in such propositions.

The notion of ‘composition’ plays a prominent role in Peter's semantics. before embarking on the specifics of the word ‘is’, he first looks into the *compositiones* involved in the noun and the verb. When it comes to the composition involved in the use of ‘is’, the starting-point for his account is the question whether the expression ‘is’ in a proposition of the form ‘S is P’ implies the ‘being’ of the composition. Whether it does or not depends on how we consider the composition. If we are talking about any composition whatsoever, in his words, the composition in general, the composition can indiscriminately be connected with beings and non-beings. This is because we can talk about both things that are and things that are not by making use of the same affirmative propositions. Hence anything expressed by a proposition of the form ‘S is P’ expresses a being in a certain sense (*ens quodammodo*). The type of composition he is referring to here is the mental content of some affirmation, which is something that only has being to a certain degree. However, the composition in general, that is, the state of affairs involved in such expressions, is primarily connected with being rather than non-being. It is when we talk about non-beings, such as chimaeras, that being in a certain sense once again enters the scene. Hence a distinction of the types of being referred to, or the types of composition involved in affirmative

propositions into being in the absolute sense (*ens simpliciter*) and being in a certain sense (*ens quodammodo*). The difference between these two types of being is illustrated by the distinction between two types of inference: from ‘A man is an animal’, in which the composition involved is a being in the absolute sense, it follows ‘Therefore a man is’, but from ‘A chimaera is a non-being’, in which the composition is a being in a certain sense only, it does not follow ‘Therefore a chimaera is’.

The counterpart of Peter's discussion of composition is the section on negation. Peter specifically goes into the question of what it is the negation denies. In his words, the negation removes the composition. The composition in this connection is identified with the affirmed state of affairs (*res affirmata*). What the negation removes is not the state of affairs, but the affirmation that goes along with it. The basis of both composition and negation turns out to be the same state of affairs, i.e. something that is formulated in the mind, to which we can either assent or deny to be the case.

The focus on matters of ontology is evidenced in other portions of Peter's logic as well. For Peter, as for Henry of Ghent (who also wrote a *Syncategoreumata*) the expression *homo* (man) in *homo est animal* (‘Man is an animal’) has simple supposition: it stands for the universal nature of humanity. Accordingly, the expression is necessarily true, even if no man should exist. The term ‘necessarily’ thus has ampliative force: it enables the subject term ‘man’ to refer to individuals not only existing in the present (which is the normal case when a present-tense verb is used), but also to those of the past and the future. This analysis runs contrary to what is found in some other *Syncategoreumata* authors, like Johannes Pagus and Nicholas of Paris, who maintain that the term ‘necessarily’ does not have ampliative force. Hence the expression *homo necessario est animal* (‘A man is necessarily an animal’) is only true on the condition that a man exists.

A similar point is made in connection with the use of modal terms. For Peter of Spain, logical necessity is based upon ontological necessity, or, the necessity of propositions has its foundation in the necessity of the things spoken about. Necessity is associated with different types of things, such as the relationships between certain concepts (such as genera and species), and the specific things the notions of which we come across in the different kinds of (scientific) knowledge (such as mathematical entities and their properties). His outlook on necessity is clearly revealed in his analysis of the inference *homo necessario est animal*; *ergo Sortes necessario est animal* (‘A man is necessarily an animal; therefore Socrates is necessarily an animal’). In his view the inference is not valid, because a transition is made from necessary being to a being at a certain time. For Peter then, the notion of necessity ultimately refers to a necessary state of affairs in reality, something that is always the case.[15]

A similar fusion of the domains of language and reality is found in Peter's account of the consecutive ‘if’, which he explains as signifying causality. Like his contemporaries he looks into the question of whether from the impossible anything follows. In his account, the notion of ‘impossibility’ can be taken in two ways, viz. impossibility as such, or absolute impossibility, which amounts to nothing, or the impossible state of affairs that is referred to when notions of things that do have a reality separately but are incompatible are combined in statements. From the latter type of impossibility, such as ‘A man is an ass’, something, but not anything can follow, e.g. ‘Therefore a man is an animal’. From impossibilities as such, e.g. ‘You know that you are a stone’, nothing can follow. The fundamental idea is that in order to be able

to have something follow, the antecedent in the consecutive relationship must be a something (*res*) of some sort.[16]

Bibliography

Primary sources:

- *Peter of Spain. Tractatus called afterwards Summule logicales*. First Critical Edition from the Manuscripts with an Introduction by L.M. de Rijk, Assen, 1972.
- *Peter of Spain. Syncategoreumata*. First Critical Edition with an Introduction and Indexes by L.M. de Rijk, with an English Translation by Joke Spruyt, Leiden/Köln/New York, 1992.

Secondary sources:

- Braakhuis, H.A.G.: *De 13de Eeuwse Tractaten over Syncategorematische Termen.* Vol. I: Inleidende Studie. Vol. II: Uitgave van Nicolaas van Parijs' *Syncategoreumata*. Nijmegen (diss.), 1979.
- Kneale, William and Kneale, Martha: *The Development of Logic*, Oxford, 1978.
- Kneepkens, C.H. Het *iudicium constructionis*. Het leerstuk van de *constructio* in de 2de helft van de 12de eeuw. Een verkennende inleidende studie gevolgd door kritische uitgaven van Robert van Parijs, *Summa 'Breve sit'* en Robert Blund, *Summa in arte grammatica* en door een werkuitgave van Petrus Hispanus *(non-papa), Summa 'Absoluta cuiuslibet'*. Nijmegen (diss.), 1987.
- de Libera, A. The Oxford and Paris Traditions in Logic. *The Cambridge History of Later Medieval Philosophy*, edited by Norman Kretzmann, Anthony Kenny and Jan Pinborg. Cambridge, 1982, 174-187
- de Libera, A. Les *Appellationes* de Jean le Page. *Archives d'histoire doctrinale et littéraire du moyen-âge* 51 (1984), 193-255.
- Nuchelmans, Gabriel: The Distinction *actus exercitus/actus significatus* in Medieval Semantics. *Meaning and Inference in Medieval Philosophy*. Studies in memory of Jan Pinborg. Edited by Norman Kretzmann. Synthese Historical Library. Studies and Texts in the History of Logic and Philosophy, Vol. 32. Dordrecht, 1988, 57-90.
- d'Ors, Angel: Petrus Hispanus O.P., Auctor Summularum. *Vivarium* 35 (1997), 21-71.
- Pinborg, Jan: *Logik und Semantik im Mittelalter. Ein Überblick*. Stuttgart-Bad Canstadt: Fromann-Holzboog, 1972.
- de Rijk, L.M.: *Logica Modernorum*. A Contribution to the History of Early Terminist Logic. Vol. I: On the Twelfth Century Theories of Fallacies. Vol. II.1: The Origin and Early Development of the Theory of Supposition. Vol. II.2: Texts and Indices. Assen, 1962-67.
- de Rijk, L.M.: On the Genuine Text of Peter of Spain's *Summule logicales* I. General problems concerning possible interpolations in the manuscripts. *Vivarium* 6 (1968), 1-34.
- de Rijk, L.M.: The Development of *Suppositio Naturalis* in Mediaeval Logic I. Natural supposition as non-contextual supposition. *Vivarium* 9 (1971), 71-107.

- de Rijk, L.M.: *La philosophie au moyen-âge*, Leiden/Köln/New York, 1985.
- de Rijk, L.M.: The Origins of the Theory of the Properties of Terms. *The Cambridge History of Later Medieval Philosophy*, edited by Norman Kretzmann, Anthony Kenny and Jan Pinborg. Cambridge, 1982, 161-173.
- Spruyt, Joke: Peter of Spain on Composition and Negation. Text. Translation. Commentary, Leiden (diss.), 1989.
- Spruyt, Joke: Thirteenth-Century Positions on the Rule '*Ex impossibili sequitur quidlibet*'. *Argumentationstheorie. Scholastische Forschungen zu den logischen und semantische regeln korrekten Folgerns*. Ed. Klaus Jacobi. Leiden/ Köln /New York, 1993, 161-193.
- Spruyt, Joke: Thirteenth-Century Discussions on Modal Terms. *Vivarium* 32 (1994), 196-226.
- Tugwell, Simon O.P.: Petrus Hispanus: Comments on Some Proposed Identifications. *Vivarium* 37 (1999), 103-113.

Other Internet resources

[Please contact the author with suggestions]

Related entries

Aristotle | Boethius, Anicius Manlius Severinus | logic: modal | medieval philosophy | truth

Joke Spruyt
University of Maastricht
j.spruyt@philosophy.unimaas.nl

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 12, 2001
Content last modified: April 12, 2001

Stanford Encyclopedia of Philosophy
Notes to Peter of Spain (Petrus Hispanus)

Notes

[1.] Peter of Spain, *Tractatus*, edited by L.M. de Rijk, 1972.

[2.] D'Ors 1997.

[3.] Tugwell 1999.

[4.] Peter of Spain, *Syncategoreumata*, edited by L.M. de Rijk with an English translation by Joke Spruyt, 1992.

[5.] See Kneepkens 1987.

[6.] For details about these works and the manuscripts see L.M. de Rijk (ed.), *Peter of Spain. Tractatus*, pp. LXXff., and de Rijk 1968, pp. 23-34.

[7.] See L. M. de Rijk (ed.), *Peter of Spain. Tractatus*, pp. LXXXIV-LXXXVIII.

[8.] See Braakhuis 1979, Vol. I, p. 248.

[9.] See de Libera 1982, pp. 176-177.

[10.] For this distinction, see Nuchelmans 1988.

[11.] See de Rijk 1962-67, Vol. II, Part I, pp. 113-117.

[12.] De Rijk 1982, pp. 167-168.

[13.] De Rijk 1971. See also de Rijk 1985, pp. 183-203.

[14.] Kneale & Kneale 1978, pp. 268-269.

[15.] For a comparison between Peter's views and his contemporaries' see Spruyt 1994.

[16.] See Spruyt 1993, pp. 161-193.

Joke Spruyt
University of Maastricht
j.spruyt@philosophy.unimaas.nl

First published: April 12, 2001
Content last modified: April 12, 2001

Moral Particularism

Moral Particularism, at its most trenchant, is the claim that there are no defensible moral principles, that moral thought does not consist in the application of moral principles to cases, and that the morally perfect person should not be conceived as the person of principle. There are more cautious versions, however. The strongest defensible version, perhaps, holds that though there may be some moral principles, still the rationality of moral thought and judgement in no way depends on a suitable provision of such things; and the perfectly moral judge would need far more than a grasp on an appropriate range of principles and the ability to apply them. Moral principles are at best crutches that a morally sensitive person would not require, and indeed the use of such crutches might even lead us into moral error.

The particularist's opponent is the generalist. Ethical generalism is the view that the rationality of moral thought and judgement depends on a suitable provision of moral principles.

1. Two Conceptions of Moral Principles

If we are going to debate the question whether there is a need for moral principles, we need some idea of what we mean by a ‘moral principle’. Unfortunately there are two radically different conceptions of what moral principles are. The first conception, the ‘absolute’ conception, takes a moral principle to be a universal claim to the effect that all actions of a certain type are overall wrong (or right). The principle

‘don't break your promises’ can be expressed in various ways: ‘it is wrong to break one's promises’; ‘all actions that involve breaking a promise are wrong’ -- and so on. On the absolute conception, these all mean that each and every action of breaking a promise is a wrong action, whatever else there may be to be said for it. Each such action is wrong overall, despite any redeeming features it may have.

There is a very different way of understanding a moral principle, as ‘contributory’ rather than as absolute. Understood in this second way, our principle maintains that if an action involves breaking a promise, that counts against it. The action is the worse for being a promise-breaking. Of course it may be the worse for being a promise-breaking but the better for some other feature that it has - that of being kindly intentioned, say. The contributory conception of moral principles allows that more than one principle can apply to the case before us, since it holds that each principle is, as it were, partial; each specifies how things are only in a certain respect. But actions have many relevant features, some counting in favour and others against. Whether the action is overall right or wrong can only be determined by the overall balance of right and wrong in it. The contributory principles do not themselves tell us how to determine that balance. They only specify contributions one by one, and leave us to work out how these add up. Some people suppose that the principles can themselves be ranked in order of importance; if that were right, it would be of some help to us in working out what matters most in a given case. Others suppose that there is no available lexical ordering of such a sort, and that the matter is left to unaided ‘judgement’.

Since there are these two quite different conceptions of what a moral principle says, our discussion will need to address both possibilities. If particularism is true, there is not much room for moral principles of either sort.

2. What the Particularist Does Not Believe

It is standard, at least in cultures informed by the Christian tradition, to think of the moral person as the person of principle. This person is the person who has learnt, or developed for herself, a sufficient range of sound moral principles (of either type), and who has sufficient skill at applying these principles to cases as they crop up. There is no need to underestimate the sort of skill that would be required for this; the matter is certainly far from mechanical. One needs judgement both to discern whether a principle applies at all and, if it does, what exactly it requires of one. Nonetheless, however difficult it may be, moral judgement is conceived here as the application of principles to cases.

If moral judgement is a rational enterprise, it must be subject to constraints of consistency. What is demanded of us when we are required to be consistent in our moral judgements? The answer is that we are required to apply our principles consistently, that is, to apply the same principle to similar cases. It is inconsistent to apply the principle ‘don't lie’ to cases involving one's friends and not to those that involve strangers. If you want to behave in that sort of way, your principle is going to have to be ‘don't lie to your friends’. What this tells us, of course, is that consistency is not the only requirement. Our moral principles are supposed to be impartial, and it is not obvious that the principle ‘don't lie to your friends’ meets this condition. But at least someone who takes it as his principle can tell the truth to his friends and

lie to strangers without inconsistency.

Why do we think of the moral person as the person of principle, and why do we think of moral judgement as subject to this sort of consistency contraint? (As we will see later, there are other forms that the consistency constraint could have taken.) The answer, I think, is that we suppose that without moral principles there could be no such thing as the difference between right and wrong. Rightness and wrongness are peculiar properties, and the only way that an action can get them is by being related to a principle in one way or another. So unless there are principles saying which sorts of actions are right and which wrong, none would be right and none wrong. If this were so, it would hardly be surprising that the good moral judge would be the person capable of following in her mind the way in which actions get to be right or wrong, which requires knowing the relevant principles and seeing that they have this effect here and that effect there. And it would be hardly surprising that consistency in judgement would amount to no more than applying similar principles to similar cases.

A rather different argument appeals not so much to a metaphysical need for principles as to an epistemological need. If there is a distinction between right and wrong actions, how are we to detect it? There must be a detectable difference between the properties of the right ones and the properties of the wrong ones. Now if an action is wrong, it is wrong because of certain other features it has -- the non-moral features that make it wrong. Those non-moral features will be detectable in the ordinary way, whatever that is. Good moral judges, having detected them, can somehow work out whether they make the action right or wrong. But if this ability is not a matter of magic, it must rest on an at least implicit knowledge of regularities connecting the non-moral features of actions and their moral properties. Moral principles specify such regularities. So if moral judgement is to be even possible, there must be a set of principles connecting moral properties to non-moral properties, contrary to what the particularist claims.

If this is our picture of the individual trying to decide what she ought to do, how are we likely to conceive of the way to resolve disagreements between two individuals? Of course there are the facts of the matter to be sorted out between them. Then presumably they have to try to agree at least on those principles which are to be taken as relevant (that is, to agree on the principles, and to agree that they are the relevant ones in the present case). Finally, they have to agree on the course of action that those principles recommend in the situation that faces them. This would be, as we might put it, a full resolution of any initial disagreement. Otherwise we are looking for a compromise of one form or another. It is possible, for instance, for a disagreement on the principles not to make any practical difference as things turn out, so that it can be left to be sorted out another day.

Overall, then, we are offered a way in which moral reasons work, and an account of the perfectly moral agent whose decision processes fit the way the reasons work, that is, fit the way in which an action can get to be right or wrong. But the way moral reasons work is probably very different from the way that other reasons work. Other reasons are not principle-driven. Morality is special, since without principles it is impossible. (Remember that the two arguments given above for the need for principles appealed to the special nature of rightness and wrongness, or of moral properties in general.)

3. What the Particularist Believes

The particularist believes, like the generalist, that the perfectly moral person is the person who is fully sensitive to the moral reasons present in the case. But the particularist paints a very different picture of what it is to be fully sensitive to those reasons. The particularist picture is one which takes moral reasons to operate in ways that are not noticeably different from the way in which other reasons function -- more ordinary reasons for action, say, or reasons for belief rather than for action. Morality may be distinguished by its subject matter, but moral thought does not have a distinctive structure.

If we are to form a view about what a full sensitivity to the reasons amounts, to, we need to have some picture of how moral reasons work. The core of particularism is its insistence on variability. Essentially the generalist demands sameness in the way in which one and the same consideration functions case by case, while the particularist sees no need for any such thing. A feature can make one moral difference in one case, and a different difference in another. Features have, as we might put it, variable relevance. Whether a feature is relevant or not in a new case, and if so what exact role it is playing there (the ‘form’ that its relevance takes there) will be sensitive to other features of the case. This claim emerges as the consequence of the core particularist doctrine, which we can call the holism of reasons. This is the doctrine that what is a reason in one case may be no reason at all in another, or even a reason on the other side. In ethics, a feature that makes one action better can make another one worse, and make no difference at all to a third.

Particularists suppose that this doctrine is true for reasons in general, so that its application to moral reasons is just part and parcel of a larger story. For an example that comes from a non-moral context, suppose that it currently seems to me that something before me is red. Normally, one might say, that is a reason (some reason, that is, not necessarily sufficient reason) for me to believe that there is something red before me. But in a case where I also believe that I have recently taken a drug that makes blue things look red and red things look blue, the appearance of a red-looking thing before me is reason for me to believe that there is a blue, not a red, thing before me. It is not as if it is some reason for me to believe that there is something red before, me, but that as such a reason it is overwhelmed by contrary reasons. It is no longer any reason at all to believe that there is something red before me; indeed it is a reason for believing the opposite.

Examples like this establish the variability of reasons for belief. Turning to reasons for action, we might point out that in some contexts the fact that something is against the law is a reason not to do it, but in others it is a reason to do it (so as to protest, let us say, against the existence of a law governing an aspect of private life with which the law should interfere). Examples of this sort can be multipled at will. They appear to establish the holism, or variability of reasons for belief and of ordinary reasons for action. The particularist suggests that there is no reason to suppose that moral reasons function in a radically different way from other reasons. Indeed, there is a sort of presumption that they don't. That presumption is partly grounded on the fact that nobody is able to say with any confidence just which reasons are moral ones and which are not. This means that providing a radical difference between the way in which reasons of the two sorts function should seem rather peculiar. But the presumption is also partly grounded in the fact

that the difference suggested by the generalist is very radical, since it affects what one might call the very logic of moral thought. To suppose that moral thought has a different logic from other thought is to adopt a bifurcated conception of rationality. Moral rationality is principle-bound, based on invariant reasons. Other forms of rationality are nothing like this at all. Particularists think that this suggestion is very strange.

These points about holism or the variability of reasons need to be expressed in different ways, according to the conception of principles that they are aimed at -- the absolute or the contributory. Principles of both sorts aim to specify invariant reasons, but the reasons they specify are rather different in style. Absolute principles, which specify a feature or combination of features that always succeed in making an action wrong (or right) wherever they occur, purport to specify an invariant overall reason, as we might put it. Counter-examples to suggested principles of this sort will consist in cases where the supposed feature or combination of features is present but the action concerned is not wrong overall (or right overall). Contributory principles are different. They purport to specify features that always make the same contribution, irrespective of context. Counter-examples to suggested contributory principles consist of cases where the feature cited is present but either does not count at all or counts the wrong way (a supposed right-making feature actually making an action worse rather than better, for instance). Particularists take their holism to be a reason to reject any invariance of reasons, of either sort -- whether at the overall or at the contributory level. Reasons as such, they say, do not need to behave in this sort of way. It is consistent with this, however, to allow that there might be some invariant reasons. What the particularist says, however, is that the possibility of morality in no way depends upon a suitable provision of invariant reasons of the sorts that principles are attempting to specify. Principle-based accounts of morality, such as those that specify ten (or some other number) of basic moral principles (e. g. Gert 1998), are left looking rather peculiar.

The picture so far is that actions get to be right or wrong in a wide variety of ways. Particularists are 'pluralists', believing that there is more than one morally relevant property. Many properties (or features) are capable of making a difference to how one ought to act, and are therefore capable of being morally relevant. But a property can be relevant on one occasion and not on another, and can count in favour of action here and against action there. Isn't this all terribly confusing? If it is all as much of a mess as this, how are we capable of keeping track of it? Are we reduced to looking at the case before us and hoping that the complex interrelations between the various features that happen to be relevant here will just strike us, somehow? Is there no such thing as general moral knowledge that one can extract from experience and bring to bear on a new case? Particularists need not deny this possibility. The question will be what form such general moral knowledge will take if it is not knowledge of the sort of invariabilities that particularism sets its face against, and that principles try to capture. I suggest that what the experienced moral judge knows is a range of ways in which a feature can contribute to determining how to act. There need be no hard core to this set of 'sorts of contribution', no common element, no limited set of paradigm cases. Instead, in understanding the practical purport of a concept such as cruelty, what one knows is the sort of difference it can make that what one proposes to do would be cruel, in a way that enables one to see new differences made in situations rather different from those one has encountered so far. Particularists may suggest that this is rather like what one knows when one knows the semantic purport of a term. In knowing the semantic purport (= the meaning) of 'and', one is in

command of a range of contributions that ‘and’ can make to sentences in which it occurs. There need be no ‘core meaning’ to ‘and’; it would be wrong to suggest that ‘and’ basically signifies conjunction. If you only know about conjunction, you are not a competent user of ‘and’ in English, for there are lots of uses that have little or nothing to do with conjunction. For example: two and two make four; ‘And what do you think you are doing? (said on discovering a child playing downstairs in the middle of the night); John and Mary lifted the boulder; the smoke rose higher and higher. Those competent with ‘and’ are not unsettled by instances such as these, but nor are they trying to understand them in terms of similarity to a supposed conjunctive paradigm or core case. Particularists in ethics will want to say the same sort of thing about what one knows when one knows the practical purport of a concept; one becomes familiar with its practical grammar. There is complexity, then, but it is manageable complexity.

This tells us how particularists will conceive of moral deliberation, when an individual tries to work out for herself how to act. There is no attempt to bring principles to bear on the situation, but there is an attempt to work out what matters here and how it matters, in ways that may involve an indirect appeal to the way things were or might be elsewhere. And when two particularists are engaged in dispute, it is not as if they are reduced to saying ‘I see it this way’. There are ways of supporting or defending the way one takes the situation to be. A particularist can perfectly well point to how things are in another perhaps simpler case, and suggest that this reveals somthing about how they are in the present more difficult one. There need be no generalist suggestion that since this feature made a certain difference there, it must make the same difference here. But our judgement can be informed, and indeed defended, by seeing the way in which a feature functions in situations that resemble the present one in various ways. What we learn is not how things must be here, but how they might very well be. Argument between two people who differ on the way to see the present case can make progress as each brings to bear other situations that are both appropriately different from and also appropriately similar to the one before them. There is no guarantee that this process will lead to agreement, any more than the generalist understanding of how disagreements get resolved leads us to suppose that all disagreements are resoluble, if treated properly. But things can happen even where there is no guarantee that they will happen.

Finally, in this section, how does the particularist understand someone who says ‘that is stealing, and therefore you should not do it’? One way of understanding what is said here is as an abbreviated argument, which fully specified reads ‘that is stealing and stealing is always wrong; therefore that is wrong’. This reading introduces silent appeal to a principle -- either absolute or contributory, according to one's way of understanding ‘that is wrong’. And it suggests that what we have here is really an inference, or argument, with premises and a conclusion. This is not how the particularist is likely to see things. Particularism is likely to think of ‘that is stealing and therefore it is wrong’ as saying ‘that is stealing and wrong for that reason’. This is not an argument, and there is nothing going on here that really merits being called inference. It is simply an account of the presence of a reason and a statement of what reason it is; that is, of what it is a reason for (or against).

4. Problems for Absolute Principles

The previous section tried to lay out the main aspects of the particularist conception of moral thought,

and of the way in which actions get to be right and wrong. Particularists do not, however, restrict themselves to expounding their own view. Of course, they are likely to say that their view is at least possible, and that generalism tends merely to assume otherwise and then to carry on blithely. The mere possibility that particularism be true is of some importance in the dialectic. But there are also reasons for doubting whether any form of generalism can really be true. Some of these have already emerged; these involved the attempt to establish a broad holism of reasons, by appeal to examples. There are replies to such attempts, which we will consider below (in Problems for Particularism); the replies amount to the claim that, despite appearances, holism must be false.

In the present section we consider reasons for thinking that morality cannot be a system of absolute principles.

The first reason is that absolute principles cannot conflict, and that if they cannot conflict a vital aspect of our moral lives (that is, conflict) has been left out of account altogether by any theory that supposes that morality is entirely governed by absolute principles.

If two supposed absolute principles conflict in a single case, one of them must be abandoned. Suppose, for instance that one principle says that all actions of type A are wrong and another says that all actions of type B are right. Suppose also that no action can be both overall wrong and overall right, and that it is possible for an action to be of both types, A and B. Things are all right so far, but if there were an action of both types, one or other of the principles would have to have abandoned. But this means that we have no room for conflict. What is meant by moral conflict here is not conflict between two individuals, but conflict between reasons for and against in a given case. There cannot be that sort of conflict, if all reasons are specified in absolute principles, because if the reasons conflicted the principles specifying them would conflict, and this would just show that one of the principles was a fraud. Conflict would, then, never be more than a product of our own misconceptions. There would be no real conflict.

What this criticism amounts to is the complaint that we need to be able to make sense of cases in which there are moral reasons on both sides, for and against. But we cannot do this effectively if all moral reasons are specified in absolute principles. Morality cannot, therefore, be just a system of absolute principles. The only way in which we could continue to think of morality as governed by absolute principles is to suppose that there is only one such principle, so that there is no possibility of conflict between principles, or to arrange things in some other way so that the principles are incapable of conflict. (Even then, of course, there would be the worry that conflict is real, and that to arrange things so that conflict is merely apparent is to erase something important.) We know of one position that offers only one principle: classical utilitarianism. The argument against this ‘monistic’ position is rather different. The argument is the direct claim that monism is false; there is more than one sort of relevant property, or more than one way in which features can get to be morally relevant. So a position with only one absolute principle is false, and one with more than one such principle cannot make proper sense of conflict.

5. Problems for Contributory Principles

The best form of generalism, therefore, probably tries to do the whole thing in terms of contributory principles -- principles that specify considerations that always count as contributory reasons. In this picture it is quite possible for there to be reasons on both sides. The classic example of such a theory is W. D. Ross's theory of Prima Facie Duties (see his 1930, ch. 2). This is just an attempt to put into good theoretical order our untutored intuitions that there are many different sorts of things that can make a difference to how we should act. There is a principle that says ‘Be just’, but this does not mean that all just actions are in fact right; it only means that the justness of an action counts in its favour, or that an action is the better for being just. Sadly, an action can be just but still wrong for other reasons. This means that it can sometimes be morally required of us that we act unjustly. If it is, there will be features of the situation that require it of us; perhaps we owe an enormous debt of gratitude, or perhaps by this unjust action we can save Holland from flooding.

The generalist who takes this line supposes, qua generalist, that a feature that makes a difference in one case will make the same sort of difference in every case, and that there will be a contributory principle specifying its regular contribution. This is what particularism is concerned to rebut. Particularists applaud Ross's insistence that there can be many features of the situation each of which makes some difference to how one should act; they merely want to say that the matter is not regular in the way that Ross, as a generalist, supposes. They have three points to make, then. The first involves producing counter-examples to suggested regular contributors. Ross supposes, for instance, in accordance with long tradition, that the fact that one has promised to do something is always some reason to do it. A counter-example to this claim would be a case where, for peculiar reasons no doubt, the fact that one has promised to do something is either no reason to do it or even a reason not to do it. Suppose, for instance, that I have promised not to keep my next three promises; what then? Again, does one always have at least some reason to tell the truth? A little bit of ingenuity enables one to come up with a case in which the fact that this is true is a reason not to say it. And so on.

The second prong of the particularist attack is to ask why we should suppose that a feature that counts in favour in one case must count the same way wherever it appears. To this question, I think, no real answer has been produced. Generalists tend to point out that if one claims that a feature counts in favour here and against there, one has something to explain. But the particularist is happy to admit this. It is true that if a feature counts in favour in one case and against in another broadly similar case, there must be an explanation of how this can be. That explanation will presumably be given by pointing to other differences between the cases. In the second case, perhaps, something that is required for the feature to count in favour is in fact absent, though it was present in the first case. Such explanations must be available, and they can be found. None of this does anything to restore a generalist conception of how reasons function.

The third prong of attack on contributory generalism involves asking for an appropriate epistemology. How are we to tell, from what we can discern case by case, that this feature will function in the same way wherever else it appears? Ross, our generalist, holds that we start with the recognition that this feature counts in favour here, but that we can immediately tell (by a process which he calls ‘intuitive induction’) that it must count in favour everywhere. The question is how this is suppposed to work. What is it that is discernible in one case and tells us that what we have here must repeat in all other cases? (Ross rightly

does not suppose that we learn our moral principles by ordinary induction.) The standard, and probably the only, answer to this question is wrong. This answer amounts to an account of what it is to make a difference in a particular case -- what it is to be relevant here. That account, understands a feature as relevant here if and only if, *in any case where it is the only relevant feature, it would decide the issue*. Now if this account of particular relevance were defensible, we would indeed have some reason to suppose that what is relevant here would be relevant in any other situation. For on each further situation it will still be true that if it were the only relevant feature, it would decide the issue. So relevance is indeed general relevance, on this showing. And this gives the generalist the epistemology he needs, for it is now easy to see how, in discerning that this feature matters here, we immediately see that it would make the same difference on every occurrence. For it is true of it on each occurrence that if it were the only relevant feature, it would decide the issue.

Sadly, the account of relevance that this all depends on is not defensible. It is, after all, true of any feature whatever that if it were the only relevant feature, it would decide the issue. The word ’relevant’ appears within this formulation, and it cannot be removed. For if we said merely that if this feature were the only feature, it would decide the issue, we would have said something that is probably both false and, worse, incoherent. It would be incoherent because the idea that a feature could be present alone, without any other features whatever, is surely nonsense. The idea that an action could be merely kind, say, without having any other features at all, makes no sense at all. Further, there may be some features that can only be relevant if some other feature is also relevant -- features that (in terms of reasons) only give us reasons if some other feature is giving us reasons as well. For instance, in the Prisoner's Dilemma one prisoner only has reasons if the other one does. If this can occur, any ‘isolation test’ for reasons must miss some reasons. Finally, trying to isolate the contribution of a feature by asking how things would have been if no other feature had made any contribution is, when one comes to think of it, a rather peculiar entreprise. It is uncomfortably like trying to determine the contribution made by one football player to his team's success today by asking how things would have been if there had been no other players on the field. So the notion of relevance that is required as a basis for generalist epistemology is unacceptable.

6. The Generalists' Reply

Generalists have two possible replies to these attacks, assuming always that they accept that many of the contributory principles that they originally suggested have been refuted by counter-example. The first thing they can do is to complicate the principles. The second thing they can do is to restrict their generalism to a limited group of reasons.

Taking the first tack, one might suggest that if the fact that one has promised is in some cases not a reason for doing what one promised to do, there will be some explanation of this. Suppose that the explanation is that what one promised to do was immoral. All one needs to do is to suck this feature into one's account of the supposedly general reason. So now the reason in ordinary cases will be that one promised to do it and it is not immoral. We might object that not even this is always a reason. What if one's promise has been extracted by duress? The response will be to suck that into the reason as well. This reason is growing all the time; now it is that one promised to do it, that it is not itself immoral, and

one's promise was not made under duress. This battle can continue; it has no obvious stopping point. Still, we might say, eventually ingenuity will give out, and we will reach a (now very complex) specification of a reason to which we can think of no appropriate counter-example.

But note what has happened here. We started from a consideration which we took to count in favour of our action, and we have ended with a complex specification of something that plays rather a different role. What we got at the end was more like an elaborate guarantee that something mentioned in the guarantee counts in favour of the action. Consider the promising example above. That I promised does, let us suppose, count in favour of my acting. But that my promise was not made under duress does not do that at all. It functions as an enabling condition, one in whose absence the first feature (that I promised) would not have been the reason it is. It is not itself a reason to do the action; that role is distinctive, and it is played here only by the fact that I promised. Note, further, that the combination of that reason and this enabling condition is not itself a (further) reason in favour of doing the action. So the distinction between ‘counting in favour’ and ‘enabling something else to count in favour’ is significant, as particularists see things. What the generalist reached, in defending her supposed reason by complication, is therefore not itself a reason at all, but only a guarantee (when finally complete) that there is a reason somewhere within it. And why should we suppose that nothing can be a reason unless we can specify a condition that guarantees its status as a reason, and that it is only a reason when present in a larger state within which it is guaranteed to serve as such? No obvious answer occurs. The whole enterprise of defending one's reason by complication begins to look strangely irrelevant, and its product unnecessary. One would have thought that there can be reasons that can function perfectly well without this sort of guarantee. And the reasons given on behalf of generalism above (in What the particularist does not believe) do nothing to show otherwise.

The second generalist line of defence involves drawing in one's horns a little. Ross distinguishes between derived and underived prima facie duties. The underived ones are the duty to do the just thing, to act for the best, not to cause harm, to keep promises, and so on. Other duties are derived from these. So there is, as we might put it, a core of invariability surrounded by a variable periphery. I might have a duty to go up to London today to see my son Hugh. But this duty is derived from a general duty to do what I have promised to do. As we might put it, that Hugh is expecting to see me today sometimes gives me a reason to go up to London and sometimes does not; it is a derived, and therefore variable, reason. If it does give me a reason, it will because it is keyed in some way into an unvariable, underived reason. So derived reasons are variable, and underived ones invariant. On this account, counter-examples will only do damage if they are aimed at the supposed underived reasons. (See McNaughton and Rawling 2000.)

A different version of this picture maintains that invariant reasons derive from the virtues (see Crisp 2000). That an action is generous, honest, just, thoughtful, or helpful is always a reason to do it. The invariant core is given by the virtues, therefore, and the variant periphery depends upon that invariant core. This last point is important, because this defence of generalism needs to show why it is that morality requires a basis of invariance. Just to come up with a few invariant reasons is nothing to the point. Those who suppose they can seriously damage particularism by specifying a few (probably fairly complex) invariant reasons do little to show that moral thought depends (as it was put in the introduction above) on a suitable provision of principles (which we are now understanding as ’invariant reasons’).

The suggestion we are dealing with now does well in this respect. We are offered an invariant core and an account of why there must be such a core if moral thought is to be possible at all.

Of course, for the suggestion to work, it must be the case that the virtues function invariantly. Particularists are likely to say, for instance, that an action can be considerate without necessarily being the better for it. It may be considerate to wipe the torturer's brow, but this fact hardly functions as a reason to wipe, or makes his sweat a reason for us to wipe it off. The torturer's other activities prevent what would ordinarily give us a reason from doing so here. Similarly, it may be that a cruel response is exactly the one called for in the circumstances; cruelty, according to particularists, need not be an invariant reason. A generalist reply to these suggestions depends on showing that similar remarks cannot be made about (a sufficient range of) the other virtues.

What is at issue between particularism and generalism is the nature of moral rationality. Particularists maintain that there can be reasons -- moral reasons -- even if the features that give us those reasons function variably rather than invariably in their reason-giving. Generalists suppose that this is not possible. They claim either that all reasons, when properly understood, must function invariably, or that there is an invariant core even if there is a variable periphery. To argue for the first claim, they often demand, for each reason, that there be a discoverable guarantee of its status as such. But until they have offered some justification for this demand, their generalism will rest on nothing. Crisp's position is a model of the second approach because it offers an account of why the variability that the particularist is so fond of pointing to must be built around an invariant core. But I would say that the supposed virtues do not in fact play the role required here.

7. Do Particularism and Generalism Differ in Practice or Only in Theory?

Particularists are fond of saying that generalists will make bad decisions. One reason for this is that generalism seems to validate certain patterns of argument that particularists would think of as invalid. For instance, a generalist might think ‘Feature F made a difference in that case; so it must make the same sort of difference here too’. If our decision in the second case was influenced by such ’reasoning’, it would have been influenced by a mistake, according to the particularist. Particularism supposes that one cannot extract from one case anything that is guaranteed to make a difference to another. They recommend keeping one's eyes firmly fixed on the case before one rather than trying to squeeze an answer to one problem out of the answer to another. This does not show that there is nothing to be learnt from other cases. Particularists can even allow that it might, on occasion, be impossible to see the right answer here if one does not work to that answer from consideration of other cases, suitably constructed or provided by experience. One can perfectly well say ‘this feature mattered there, and so it might well matter here -- I had better have a look and see whether it does or not’. What one cannot and should not do is to say ‘it mattered there and so it *must* matter here’. So particularists allow a relevance to moral experience; they are not reduced to just gazing vacantly at the case before them and coming up with an answer that somehow seems appropriate. There is a practical difference between particularism and

generalism, but it is not this.

There is another possible practical difference between the two. This comes out when we consider two pretty similar cases of which we nonetheless want to make different judgements. Nobody supposes that this is impossible. The question is rather what is rationally required of the judge in such a case. The generalist might end up demanding that one make the same judgement in both cases unless one can provide a principle that distinguishes them. The particularist, by contrast, might demand only that one make the same judgement in both cases unless one can offer some reason for not doing so. Some, however, would not even demand that. All agree that there must be some relevant difference between any two cases of which one wants to make different judgements. Might it be enough to allow that there is some such difference, even though one has no idea what it is? Or is one rationally required to be able to make some suggestion about what it is? Or is one's suggestion to be formulated as a possible principle governing all similar cases? Particularists might be distinguished from generalists by their answer to these questions.

8. Problems for Particularism

People reject the persuasive charms of particularism for, broadly, two sorts of reasons: reasons to do with rationality, and reasons to do with motivation. I take rationality first. Three points are made. The first and most direct is that thinking rationally requires at least that one think consistently, and in ethics this just means taking the same feature to be the same reason wherever it occurs. Particularism, therefore, denies the rationality of moral thought. Second, what is the difference between moral choice and choosing chocolates? The difference is that when choosing morally we are required to make similar choices in similar circumstances; not so for the choice between rum truffles and peppermint creams. Third, what account can the particularist give of our ability to learn from our moral experience? Such moral self-education is certainly possible. An adolescent who has so far refused to accept that tact is a virtue can be brought to see the importance of being tactful in a particular case, and is then in a position to apply this knowledge more generally. The generalist can understand this as the extraction of a principle from an earlier case, which we then apply to later ones. What can the particularist offer as an alternative account?

Of these three points, the third is the hardest. The answer to the first is that, when we are thinking of reasons for belief, the sort of consistency required of us is merely that we do not adopt beliefs that cannot all be true together. Why should we understand the consistency requirement in a different way when we turn to moral reasons? Simply to insist that this is so must be to beg the question against particularism.

The second question asks us to justify a distinction between matters of whim, such as the choosing of chocolates, and matters of weighty reasons, such as those involved in moral choice. But this need not be a problem. Moral reasons as the particularist understands them occur in the one case and not in the other. Nothing at all like them applies to the choosing of chocolates (normally). This does nothing to show that in morality, unlike in the area of whim, we are required to make similar choices in similar situations. There are quite enough other differences between morality and whim.

The third question asks us what relevance other cases do have to a new case, if not the sort of relevance that the generalist supposes. The answer to this is that experience of similar cases can tell us what sort of thing to look out for, and the sort of relevance that a certain feature can have; in this way our judgement in a new case can be informed, though it is not forced or constrained, by our experience of similar cases in the past. There is no need to suppose that the way in which this works is by the extraction of principles from the earlier cases, which we then impose on the new case.

So much for one sort of complaint. I now turn to questions which focus on motivation. The general idea here is that a particularist morality is a lax morality: without principles, anything goes. But there are various ways in which this thought can be built up. The first is just to say that morality is in the business of imposing constraints on our choices. For there to be constraints, there needs to be regulation, and regulation means rules, and rules mean principles. This, however, is just wrong. There can be fully particular constraints on action, and the judgement that this action would be wrong is surely just such a thing. Constraints do not need to be general constraints, any more than reasons need to be general reasons.

Another line is that the person of principle will be unbudgeable; having taken a stand on an issue, he will not be moved from it. A particularist will not be like this. But here I have two things to say. First, nothing prevents a particularist from being of firm conviction case by case; an unbudgeable conviction need not be founded on principle, but simply on the nature of the case. Unbudgeability and principle have nothing essentially in common. Second, even if it were true that a principled person will on some points be unbudgeable, the question is whether those points are the right points. The worrying thought is that they might not be -- that in being driven by principle, our principled person will distort the relevance of relevant features by insisting on filtering them through principles, in a way that is at odds with the falsehood of generalism. In my view, unbudgeability and principles go very badly together. Unbudgeability may be a virtue in its place, but to be unbudgeably involved in a distortion is not a great triumph. If you are going to be incorrigible you had better always be right; incorrigible error is the worst of all worlds.

A different suggestion is that morality has the sort of authority over us that can only be provided by a rule. Here, however, I think that particularists should simply dig their heels in, and insist that moral reasons have all the authority they need already. She needs medical help, and I am the only person around to summon it. This situation demands a certain response from me, in a way that has authority over me because there is nothing that I can do to get out of it.

Still, we might say, there is the ever-present danger of backsliding in ethics; we see the right, but somehow cannot bring ourselves to do it. With principles, we have something capable of stiffening our waning resolve. Without principles, we will fall short all too often. One answer to this is that it is an empirical hypothesis for which there is little real evidence. What is more, the need for moral stiffening only arises once we have already decided what morality requires of us here, and the real question was whether that decision needed to be based on principle. As far as the point about backsliding goes, it does not; the need, if any, for principles comes later.

More to the point might be a worry about special pleading. This is different from backsliding, because the special pleader is the person who makes exceptions in their own favour. It would not be right for most people to do what I propose to do, but I am special; so I am left off the moral hook that others are caught by. This sort of special pleading occurs in the process of making our moral decision; it is not to do with motivation thereafter, as backsliding is. With backsliding I say ‘this is wrong but I am going to do it all the same’; with special pleading I say ‘this would be wrong for others, but not for me’.

The reason there is a genuine worry about special pleading is that one can always find some difference between this act and a plain duty, and there seems to be no way, within the resources available to particularism, to prevent such differences from being appealed to by those who, in bad faith, want to let themselves off the moral hook. A principle, we might say, would, or at least should, stop this sort of thing.

What is really going on here is that we are appealing to principles to rectify a natural distortion in moral judgement. If such judgement focuses only on the reasons present in the case before us, it is all too easy to twist those reasons to suit oneself. So we use principles to stop ourselves from doing that. But really the remedy for poor moral judgement is not a different style of moral judgement, principle-based judgement, but just better moral judgement. There is only one real way to stop oneself distorting things in one's own favour, and that is to look again, as hard as one can, at the reasons present in the case, and see if really one is so different from others that what would be required of them is not required of oneself. This method is not infallible, I know; but then nor was the appeal to principle.

Bibliography

- Aristotle, *Nicomachean Ethics*.
- Audi, R., (1998) ‘Moderate Intuitionism and the Epistemology of Moral Judgement’, *Ethical Theory and Moral Practice* vol.1, 15-44, esp. pp. 36-41.
- Blackburn, S., (1992) ‘Through Thick and Thin’, *Proceedings of The Aristotelian Society* Supp. Vol. 66, pp. 285-299.
- Broad, C. D., (1930) *Five Types of Ethical Theory* (London: Routledge & Kegan Paul).
- Crisp, R., (2000) ‘Particularizing Particularism’, in Hooker and Little (2000), pp. 23-47.
- Dancy, J., (1981) ‘On Moral Properties’ *Mind* 90, pp. 367-85.
- Dancy, J., (1982) ‘Intuitionism in Meta-epistemology’, *Philosophical Studies* 42, pp. 395-408.
- Dancy, J., (1983) ‘Ethical Particularism and Morally Relevant Properties’, *Mind* 92, pp. 530-47.
- Dancy, J., (1985) ‘The Role of Imaginary Cases in Ethics’, *Pacific Philosophical Quarterly* 66, pp. 141-53.
- Dancy, J., (1993) *Moral Reasons* (Oxford: Blackwell).
- Dancy, J., (2000a)‘The Particularist's Progress’, in Hooker and Little (2000), pp. 130-56.
- Dworkin, G., (1995) ‘Unprincipled Ethics’, *Midwest Studies in Philosophy* 20: *Moral Concepts*, pp. 224-39.
- Frazier, R., (1995) ‘Moral Relevance and Ceteris Paribus Principles’, *Ratio* 8, pp. 113-27.
- Garfield, J., (2000) ‘Particularity and Principle: The Structure of Moral Knowledge’ in Hooker

and Little (2000), pp. 178-204.

- Gert B., (1998) *Morality: Its Nature and Justification* (New York: Oxford University Press).
- Hooker, B. W., (2000) ‘Moral Particularism -- Wrong and Bad’, in Hooker and Little (2000), pp. 1-23.
- Hooker, B. W. and Little, M., eds. (2000) *Moral Particularism* (Oxford: Oxford University Press).
- Jackson, F., Pettit, P., and Smith, M., (2000) ‘Ethical Particularism and Patterns’, in Hooker and Little (2000), pp. 79-99.
- Jonsen, A. R. and Toulmin, S., (1988) *The Abuse of Casuistry: A History of Moral Reasoning* (Berkeley: University of California Press).
- Kagan, S., (1988) ‘The Additive Fallacy’, *Ethics* 99, pp. 5-31.
- Little, M., (2000) ‘Moral Generalities Revisited’, in Hooker and Little (2000), pp. 276-304.
- Little, M., ‘Wittgensteinian Lessons on Particularism’, in Carl Elliot (ed.), *Wittgensteinian Bioethics* (Durham, NC: Duke University Press, forthcoming).
- Little, M., (1994) ‘Moral Realism: Non-Naturalism’, *Philosophical Books* 35, pp. 225-32.
- Little, M., (1995) ‘Seeing and Caring: The Role of Affect in Feminist Moral Epistemology’, *Hypatia* 10, 117-37.
- McDowell, J., (1979) ‘Virtue and Reason’ *The Monist* 62, pp. 331-50.
- McDowell, J., (1981) ‘Non-cognitivism and Rule-following’, in *Wittgenstein: To Follow a Rule*, S. Holtzman and C. Leich (eds) (London: Routledge and Kegan Paul), pp. 141-62; reprinted in McDowell (1998), pp. 98-218.
- McDowell, J., (1982) ‘Criteria, Defeasibility, and Knowledge’, *Proceedings of the British Academy*, 68, pp. 455-79.
- McDowell, J., (1998) *Mind, Value, and Reality* (Cambridge, MA: Harvard University Press).
- McKeever, S. and Ridge, M., ‘The Many Moral Particularisms’, forthcoming.
- McNaughton, D. A., (1988) *Moral Vision* (Oxford: Blackwell).
- McNaughton, D. A., (1996) ‘An Unconnected Heap of Duties?’, *Philosophical Quarterly* 46, pp. 433-47.
- McNaughton, D. A. and Rawling, P., (2000) ‘Unprincipled Ethics’, in Hooker and Little (2000), pp. 256-75.
- Moore, G. E., (1903) *Principia Ethica* (Cambridge: Cambridge University Press).
- Murdoch, I., (1970) *The Sovereignty of Good* (Oxford: Blackwell), esp. pp. 32-3, 44.
- Nussbaum, M., (1990) *Love's Knowledge* (New York: Oxford University Press).
- O'Neill, O., (1996) *Towards Justice and Virtue: A Reconstructive Account of Practical Reasoning* (Cambridge: Cambridge University Press).
- Philips, M., (1987) ‘Weighing Moral Reasons’, *Mind* 96, pp. 367-76.
- Rachels, J., (1993) *The Elements of Moral Philosophy* (New York: Harper Collins).
- Rawls, J., (1971) *A Theory of Justice*. Harvard University Press.
- Raz, J., (2000c) ‘The Truth in Particularism’, in Hooker and Little (2000), pp. 48-78.
- Richardson, H. S., (1990) ‘Specifying Norms’, *Philosophy and Public Affairs* vol. 19, pp. 279-310.
- Ross, W. D., (1930) *The Right and the Good* (Oxford: Clarendon Press).
- Ross, W. D., (1939) *Foundations of Ethics* (Oxford: Clarendon Press).

- Scanlon, T. M., (1998) *What We Owe Each Other* (Cambridge, MA: Harvard University Press).
- Scheffler, S., (1987) 'Morality through Thick and Thin', *Philosophical Review* 96, pp. 411-34.
- Shafer-Landau, R., (1997) 'Moral Rules', *Ethics* 107, pp. 584-611.
- Sinnott-Armstrong, W., (1999) 'Some Varieties of Particularism', *Metaphilosophy* 30, pp. 1-12.
- Stratton-Lake, P. J., (2000) *Kant, Duty, and Moral Worth* (London: Routledge).
- Tännsjö, T., (1995) 'In Defence of Theory in Ethics', *Canadian Journal of Philosophy*, 25, pp. 571-93.
- Walker, M. U., (1997) *Moral Understandings* (London: Routledge, 1997).
- Wallace, J., (1988?) *Ethical Norms, Particular Cases* (Ithaca: Cornell University Press, 1996).
- Wiggins, D., (1976a) 'Deliberation and Practical Reason', *Proceedings of the Aristotelian Society* 76, pp. 29-51; reprinted in Wiggins (1987), pp. 215-37.
- Wiggins, D., (1976b) 'Truth, Invention and the Meaning of Life', *Proceedings of the British Academy* 62, pp. 331-78; reprinted in Wiggins's (1987), pp. 87-137.
- Wiggins, D., (1987) *Needs, Values, Truth* (Oxford, Blackwell).
- Wittgenstein, L., (1953) *Philosophical Investigations* (Oxford: Blackwell).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

practical reason | reasons: justification vs. explanation

Jonathan Dancy
j.p.dancy@reading.ac.uk

First published: June 5, 2001
Content last modified: June 5, 2001

Logical Form

Some inferences are impeccable. Consider:

> (1) John danced if Mary sang, and Mary sang; so John danced.
>
> (2) Every politician is deceitful, and every senator is a politician; so every senator is deceitful.
>
> (3) The tallest man is in the garden; so someone is in the garden.

Such reasoning cannot lead from true premises to false conclusions. The premises may be false. But a thinker takes no epistemic risk by endorsing the conditional claim: if the premises are true, then the conclusion is true. Given the premises, the conclusion follows immediately--without any further assumptions that might turn out to be false. By contrast, it would be very risky to infer that John danced, given only the assumption that Mary sang. More interesting examples include:

> (4) John danced if Mary sang, and John danced; so Mary sang.
>
> (5) Every hairless biped is a bird, Tweety is a hairless biped; so Tweety can fly.
>
> (6) Every human born before 1850 has died; so every human will die.

Inference (4) is not secure. Suppose John dances whenever Mary sings, and he sometimes dances when Mary doesn't sing. Similarly, (5) relies on unstated assumptions--e.g., that Tweety is not a penguin. Even (6) falls short of the demonstrative character exhibited by (1-3). While laws of nature may preclude immortality, it is conceivable that someone will escape the grim reaper; and the conclusion of (6) goes beyond its premise, even if it is (in some sense) foolish to resist the inference.

Appeals to logical form arose in the context of attempts to say more about this intuitive distinction between impeccable inferences, which invite metaphors of security and immediacy, and inferences that involve a risk of slipping from truth to falsity. The motivations for saying more are both practical and theoretical. Experience teaches us that an inference can initially seem more secure than it is; and if we knew which inferences are risk-free, we might be more alert to the points at which we risk error. (Such alertness might be valuable, even if the risks are tolerable.) Claims about inference are also connected, in various ways, with claims about thought and its relation to language. So we would like to know *in virtue of what*, if anything, an inference is impeccable. The most common suggestion has been that certain inferences are absolutely secure by virtue of their (logical) form; although conceptions of form have

evolved in tandem with conceptions of logic and language.

---

1. Patterns of Reason

An ancient thought is that impeccable inferences exhibit patterns that can be characterized schematically by abstracting away from the specific contents of particular premises and conclusions, thereby revealing a general form common to many other impeccable inferences. Such forms, along with the inferences that exemplify them, are said to be *valid*. With regard to (1), it seems that the conclusion is part of the first premise, and the second premise is another part of the first. We can express this point by saying that (1) is an inference of the form: **B** if **A**, and **A**; so **B**. And many other inferences, like (7), share this form:

(7) Chris swam if Pat was asleep, and Pat was asleep; so Chris swam.

The Stoics discussed such inferences, all of which are equally secure, using numbers (instead of letters) to reflect the abstract form: if the first then the second, and the first; so the second. In addition to this variant of '**B** if **A**, and **A**; so **B**', the Stoics formulated other valid schemata:

If **the first** then **the second**, but not **the second**; so not **the first**.

Either **the first** or **the second**, but not **the second**; so **the first**.

Not both **the first** and **the second**, but **the first**; so not **the second**.

Schematic formulations like these require variables. And let us introduce 'proposition' as a term of art for whatever the variables above (represented in bold) range over. Propositions are thus the sorts of things that can be true or false; for they are potential premises/conclusions--things that can figure in valid

inferences. (This leaves it open what propositions are: sentences, states of affairs, or whatever.)

We can now draw an important distinction. In speaking of an inference, one might be talking about (i) a mental *process* in which a thinker draws a conclusion from some premises, or (ii) the *propositions* a thinker would accept (perhaps tentatively or hypothetically) if she accepted the premises and conclusion, with one proposition designated as an alleged consequence of the others. The latter notion seems to be primary with respect to what makes an inference risk-free. For a risky thought process is one in which a thinker who accepts certain propositions comes to accept, without further evidence, a proposition that does not follow from the initial assumptions. So let us focus on premises and conclusions, as opposed to episodes of reasoning; and let us assume that propositions themselves have forms. Then the inference

> (1) John danced if Mary sang, and Mary sang; so John danced.

is secure, in part *because* its first premise has the form '**B** if **A**'. If the proposition lacked this form, one could not explain the impeccability of (1) by saying that '**B** if **A**, and **A**; so **B**' is a form of valid inference.

It is not obvious that *all* impeccable inferences are instances of some valid form, and thus inferences whose impeccability is due to the forms of the relevant propositions. But this thought has served as an ideal for the study of inference, at least since Aristotle's treatment of examples like

> (2) Every politician is deceitful, and every senator is a politician; so every senator is deceitful.

The first premise in (2) seems to have several parts, each of which is a part of the second premise or the conclusion; and the inference is presumably valid because these proposition-parts exhibit the right pattern. Aristotle (who predated the Stoics) captured this idea by noting that conditional claims of the following form are sure to be true: if every *P* is *D* and every *S* is a *P*, then every *S* is *D*. Correspondingly, the following inference schema is valid:

> Every *P* is *D*, and every *S* is a *P*; so every *S* is *D*.

Aristotle discussed a range of such inferences, called syllogisms, involving quantificational propositions---indicated by words like 'every' (or 'all') and 'some'. Two other syllogistic forms are expressed below as valid schemata, although Aristotle typically presented his syllogisms as conditional claims:

> Every *P* is *D*, and some *S* is a *P*; so some *S* is *D*

> Some *S* is not *D*, every *S* is a *P*; so some *P* is not *D*.

These (italicized) variables are intended to range over certain *parts* of propositions. There is a sense in

which common nouns like ‘politician’ and adjectives like ‘deceitful’ are *general*, since they can apply to many individuals; and just so, it seems, propositions contain correspondingly general elements. For example, the proposition that every senator is deceitful contains two such elements, both of which are relevant to the validity of inferences involving this proposition.

Propositions thus seem to have structure that bears on the (im)peccability of inferences, even ignoring premises/conclusions with propositional parts. That is, even ‘simple’ propositions have logical form. As Aristotle noted, pairs of such propositions can be related in interesting ways. If every *P* is *D* (and there is at least one *P*), then some *P* is *D*. If no *P* is *D*, then some *P* is not *D*. It is certain that either every *P* is *D* or some *P* is not *D*; and whichever of these propositions is true, the other is false. Similarly, the following propositions cannot both be true: every *P* is *D*; and no *P* is *D*. But it is not certain that either every *P* is *D* or no *P* is *D*. (Perhaps some *P* is *D* and some *P* is not *D*.) This network of logical relations strongly suggests that the propositions in question contain a quantificational element; two general elements; and sometimes, an element of negation. This raises the question of whether other propositions have a similar structure.

2. Propositions and Traditional Grammar

Consider the proposition that Venus is bright, which can figure in inferences like

> (8) Every planet is bright, and Venus is a planet; so Venus is bright.

Aristotle said little about such inferences. But others formulated the schema ‘every *P* is *D*, and *n* is a *P*; so *n* is *D*’, where the new variable is intended to range over proposition-parts of the sort indicated by names. (On some views, discussed below, the proposition that Venus is bright contains a quantifier and two elements of generality; though unsurprisingly, such views are tendentious.) In any case, Aristotle knew about propositions like the conclusion of (8). Indeed, he thought such propositions exemplify a subject-predicate structure shared by all propositions--and the sentences that indicate (or express) them.

The *sentence* ‘every politican is deceitful’ intuitively divides into two major parts, as shown by the slash: ‘every politician / is deceitful’. Similarly for ‘Venus / is bright’, ‘some politician / swam’, and ‘the brightest planet / rises early in the evening’. Until fairly recently, it was held that every declarative sentence can be divided into subject and predicate in this way, *and* that propositions were like sentences in this respect. Aristotle would have said that in ‘Venus is bright’, (the property of being) bright is predicated of Venus; in ‘every *P* is *D*’, *D* is predicated of every *P*. Thus, ‘Venus’ and ‘every politician’ can both indicate subjects; and the word ‘is’ introduces a predicate. Using slightly different terminology, other theorists treated all elements of generality as predicates, and propositions with the following structure were said to have *categorical form*: subject-copula-predicate; where a copula, indicated by words like ‘is’ or ‘was’, links a subject (which can consist of a quantifier and predicate) to a predicate.

In this context, it is relevant that sentences like ‘Every man swam’ can---with some awkwardness---be

paraphrased by sentences like ‘Every man was a swimmer’. So perhaps both sentences indicate the same proposition, while the second better reflects the true categorical form of the proposition. Maybe ‘swim’ is an abbreviation for ‘was a swimmer’, in the way that ‘bachelor’ is arguably short for ‘unmarried man’. (One would violate English grammar by saying ‘Every man was swimmer’; the article ‘a’ is needed. But let us assume that this feature of English, not found in all languages, is not reflected in propositions.)

The proposition that *every man is tall if Venus is bright* seems to be a ‘molecular’ compound of categorical propositions; the same complex proposition is presumably indicated by ‘if Venus is bright then every man is tall’. The proposition that *not only every man is tall* apparently extends a categorical proposition, via the elements indicated by ‘not’ and ‘only’. Such reflections suggest the possibility, explored with great ingenuity by medieval logicians, that all propositions are composed of categorical propositions and a small number of logical (or so-called syncategorematic) elements. This is not to say that all propositions were, or could be, analyzed in this manner. But by formulating various Aristotelian inference schemata, in ways that complemented proposed analyses of complex propositions, many impeccable infererences were revealed as instances of valid syllogistic forms.

Medieval logicians also discussed the relation of logic to grammar. Many viewed their project, in part, as an attempt to uncover principles of a mental language common to all thinkers. (Aristotle had said that spoken sounds symbolize ‘affections of the soul’.) From this perspective, one expects a few differences between the logical forms of propositions, and overt features of sentences. Spoken languages must mask certain aspects of logical structure, if the proposition that *every man swam* has categorical form3and thus contains a copula. Ockham held that a mental language would have no need for Latin's declensions, and that logicians could ignore such aspects of spoken language. The ancient Greeks were aware of sophisms like: that dog is a father, and that dog is yours; so that dog is your father. This bad inference cannot share its form with the superficially parallel (but impeccable) variant: that dog is a mutt, and that mutt is yours; so that dog is your mutt. So the superficial features of sentences are not infallible guides to the logical forms of propositions. But the divergence was held to be relatively minor. Spoken sentences have structure; they are composed, in systematic ways, of words. And the assumption was that sentences reflect the major aspects of logical form, including subject-predicate structure. There is a distinction between logic and the task of describing the sentences used in spoken languages. But the connection between logic and grammar was thought to run deep, making it tempting to say that the logical form of a proposition is the grammatical form of some (perhaps mental) sentence.

3. Motivations for Revision

Towards the end of the eighteenth century, Kant could say (without much exaggeration) that logic had followed a single path since its inception, and that ‘since Aristotle it has not had to retrace a single step’. He also said that syllogistic logic was ‘to all appearance complete and perfect’; but this was exuberance. There were too few schemata, yet there were also too many.

Some valid schemata are reducible to others, in that any inference of the reducible form can be revealed as valid (with a little work) given other schemata. And this turns out to be important. Consider

(9) If Al ran then either Al did not run or Bob did not swim, and Al ran; so Bob did not swim.

And assume that ‘Al did not run’ negates ‘Al ran’, while ‘Bob did not swim’ negates ‘Bob swam’. Then (9) is an instance of ‘if **A** then either not-**A** or not-**B**, and **A**; so not-**B**’. But we can treat this as a derived form, reducible to other valid schemata, like ‘if **the first** then **the second**, and **the first**; so **the second**’. Given the premises of (9), it follows that either not-**A** or not-**B**; so given **A**, it follows that not-**B**, in accordance with the schema ‘either not **the first** or not **the second**, and **the first**; so not **the second**’. Similarly, as Aristotelian logicians recognized, the following schema can be treated as a derived form:

(10) Some *P* is not *D*, and every *S* is *D*; so not every *P* is an *S*.

We have already seen that if some *P* is not *D*, then not every *P* is *D*; and every *P* is *D*, if both every *S* is *D* and every *P* is an *S*. So it cannot be that both of these conditions obtain, given the first premise of (10); and if every *S* is *D*, as stated in the second premise, then it follows that not every *P* is an *S*.

This invites the thought that further reduction is possible, especially given another respect in which there are more syllogistic schemata than one would like. Consider the contribution indicated by ‘every’ to propositions of the form: every *P* is *D*; not every *P* is *D*; every *P* is not *D*, etc. Ideally, one would specify this contribution--the logical role of the universal quantifer--with a single schema that reveals a *common* contribution to all propositions containing a universal quantifier. Similarly, one would like to characterize the common contribution of negation to propositions of diverse forms: not **A**; not every *P* is *D*; some *P* is not *D*; not both **A** and **B**; etc. But syllogistic logic is not ideal in this respect.

Correspondingly, one might suspect that there are relatively few *basic* inferential patterns. Some inferences may reflect inherently compelling transitions in thought. Perhaps ‘**B** if **A**, and **A**; so **B**’ is so obvious that logicians are entitled to take this rule of inference as axiomatic. But how many rules are plausibly regarded as fundamental in this sense? Theoretical elegance and explanatory depth favor theories with fewer irreducible assumptions. Indeed, Euclid's geometry had long provided a model for how to present a body of knowledge as a network of propositions that follow from a few basic axioms; and for several reasons, foundational questions played an important role in nineteenth century logic and mathematics. The work of Boole and others showed that progress in this regard was possible with respect to the logic of inferences involving propositional variables. But the syllogisms remained disunified and incomplete, for reasons related to another failing of traditional logic/grammar.

Propositions involving *relations*--e.g., the proposition that Juliet kissed Romeo--are evidently not categorical. One might suggest ‘Juliet was a kisser of Romeo’ as an overtly categorical paraphrase. But the predicate ‘kisser of Romeo’ differs, in ways that matter to inference, from predicates like ‘woman’. If some kisser of Romeo died, it follows that someone was kissed; whereas the proposition that some woman died has no comparable consequence (stated in the passive voice). Correlatively, if Juliet kissed Romeo, then Juliet kissed *someone*. This last proposition is of interest, even if we express it by saying

'Juliet was a kisser of Romeo'; for such a proposition involves a quantifier (outside the subject) as part of a complex predicate. And traditional logic does not provide the resources needed for capturing the validity of inferences whose validity depends on quantifiers *within* predicates, as in:

(11) Some patient respects some doctor, and every doctor is a sailor; so some patient respects some sailor.

If 'respects some doctor' and 'respects some sailor' are nonrelational, like 'is tall', (11) has the following form:

Some *P* is *T*, and every *D* is an *S*; so some *P* is *O*.

(Replacing '*T*' with 'a respecter of some doctor' and '*O*' with 'a respecter of some sailor'.) But this schema, which fails to reflect any quantificational structure in the *predicates*, is not valid. Its instances include the bad inference: some patient is tall, and every doctor is a sailor; so some patient is omniscient. This dramatizes the point that 'respects some doctor' and 'respects some sailor' are logically related, in ways that 'is tall' and 'is omniscient' are not.

One can introduce a variable '$\underline{R}$', intended to range over relations, and formulate the following schema: some *P* $\underline{R}$ some *D*, and every *D* is an *S*; so some *P* $\underline{R}$ some *S*. But the problem remains. Quantifiers can still appear in the predicates, as in the following (complex, but still impeccable) inference:

(12) Every patient who met every doctor is tall, and some patient who met every doctor respects every senator; so some patient who respects every senator is tall.

If 'patient who met every doctor' and 'patient who respects every senator' are nonrelational, then (12) has the form: every *P* is *T*, and some *P* $\underline{R}$ every *S*; so some *O* is *T*. And this is not a valid form. Consider: every politician is tall, and some politician respects every senator; so some obstetrician is tall. Again, one might abstract a valid schema that covers (12), letting parentheses indicate a relative clause:

Every $P(\underline{R}_1$ every $D)$ is T, and some $P(\underline{R}_1$ every $D)$ $\underline{R}_2$ every S; so some $P(\underline{R}_2$ every $S)$ is T.

But there can be still further quantificational structure in these predicates; and so on. This suggests that it really is important--if impeccability is to be revealed as a matter of form--to find a *general* characterization of how quantifiers contribute to propositions in which quantifiers can appear.

4. Frege and Formal Language

Frege showed how to resolve these difficulties in one fell swoop. His system of logic, published in 1879 (and still in use, with notational modifications), was the single greatest contribution to the subject. So it is

significant that on Frege's view, propositions do not have the subject-predicate form of sentences. Frege's achievement required a substantial distinction between logical form and grammatical form (as traditionally conceived). It is hard to overemphasize the impact of this point on subsequent discussions of thought and its relation to language.

Frege's leading idea was that propositions have 'function-argument' structure. The intended analogy to mathematical functions, developed by Frege in later work, intimates his view. The successor function maps every integer onto its successor; it maps the number 1 onto the number 2, 2 onto 3, etc. We can represent this function, using a variable that ranges over integers, as follows: $S(x) = x + 1$. As this notation makes clear, the function takes integers as *arguments*; and given an argument, the *value* of the function is the successor of that argument. The division function, representable as '$Q(y, z) = y/z$', maps *ordered pairs* of numbers onto quotients; the pair (8, 4) onto 2; (9, 3) onto 3; etc. Mappings can also be conditional, as with the function that maps every even integer onto itself and odd integer onto its successor: $F(x) = x$ if x is even, and $x + 1$ otherwise. By itself, however, no function has a value. Frege says that functions are *saturated* by arguments. The metaphor, encouraged by his claim that we can indicate functions with expressions like 'S( ) = ( ) + 1', is that a function has 'holes' that can be 'filled' by arguments (of the right sort). Variable letters, such as the '*y*' and '*z*' in as '$Q(y, z) = y/z$', are convenient for representing functions that take more than one argument. But we could also express the division function by indexing the argument places as follows:

$$Q([_]_i , [_]_j) = [_]_i / [_]_j$$

Similarly, propositions are said to contain unsaturated elements in combination with the requisite number of arguments. The proposition that Mary sang is said to contain a functional component indicated by 'sang', and an argument indicated by 'Mary': Sang(Mary). Frege thinks of the relevant function as a conditional mapping from individuals to truth values: Sang(x) = *true* if x sang, and *false* otherwise. The proposition that John admired Mary contains an ordered pair of arguments and a function: Admired(John, Mary). One can think of the function as a mapping from ordered pairs of individuals to truth values--or alternatively, as a function from individuals to *functions* from individuals to truth values. In any case, the proposition that Mary was admired by John has the same function-argument structure; although pace Frege, this is not yet reason to deny that all propositions have subject-predicate structure. (A traditional grammarian could say that the passive sentence fails to reflect the true structure of the indicated proposition, in which the subject is John, the first argument.) But Frege's treatment of quantified propositions, which flows from his claims about saturation, does depart from the tradition.

Let ***F*** be the function indicated by 'Sang( )'. Mary sang, if and only if: ***F*** maps the individual Mary onto the value *true*; i.e., ***F*** has the value *true* given Mary as argument. If Mary (or anyone else) sang, then someone sang. Thus, someone sang if and only if ***F*** maps at least one individual onto the value true; or using a modern variant of Frege's notation, $\exists x$[Sang(x)]. The quantifier '$\exists x$' is said to bind the variable '*x*', which appears in 'Sang(x)'; and this variable ranges over objects in a domain of discourse. (For now, assume the domain contains only persons.) If everyone sang, then each individual in the domain sang, so ***F*** maps each individual on the value *true*; or using formal notation, $\forall x$[Sang(x)]. A quantifier binds each

occurrence of its variable, as in '$\exists x[D(x)$ & $C(x)]$', which reflects the logical form of 'someone is deceitful and clever'. Here the quantifier combines with a conjunction of functions:

$$[\text{Quantifier}]_i\ [\ \text{Predicate}([_]_i)\ \&\ \text{Predicate}([_]_i)\ ]$$

Turning now to the syllogistic proposition that some politician is deceitful, traditional grammar suggests the division 'some politician / is deceitful', with the noun 'politician' going with 'some'. Frege suggests, however, that the logically relevant division is between the quantifier and the rest: $\exists x[P(x)$ & $D(x)]$; someone is both a politician and deceitful. (If the logical form is '$\exists x[P(x)$? $D(x)]$', where '?' is a connective, then '&' is the connective we want; some politician is deceitful, if and only if someone is both a politican and deceitful.) With respect to the proposition that every politician is deceitful, Frege thinks the logical form is given by '$\forall x[P(x) \rightarrow D(x)]$': everyone is such that if he is a politician then is deceitful. (Clearly, '$\forall x[P(x)$ & $D(x)]$' would be wrong, since this implies that everyone is a politician. Frege defines '$\rightarrow$' so that '$\forall x[P(x) \rightarrow D(x)]$' is equivalent to '$\forall x[\neg P(x) \vee D(x)]$', everyone either fails to be a politician or is deceitful; and '$\forall$' is defined so that $\forall x[P(x)] \leftrightarrow \neg\exists x[\neg P(x)]$.) This notation suggests that grammar is misleading in at least two respects. First, grammar leads us to think that 'some politician' indicates a major *constituent* of the proposition that some politician is deceitful. But if Frege is right, no constituent of the *proposition* contains both the quantifier and the predicate indicated by 'politician'; quantified propositions divide along lines that keep the predicates together (apart from the quantifier). Second, grammar masks a difference between existential and universal syllogistic propositions: the main predicates are related conjunctively in the former, and conditionally in the latter. Initially, one might object to Frege's departure from intuition here; but on his view, multiply quantified propositions present no special difficulty.

Just as a single quantifier can bind an unsaturated position associated with a function that takes a single argument, two quantifiers can bind two unsaturated positions associated with a function that takes a pair of arguments. The proposition that everyone trusts everyone, for example, has the following (noncategorical) form: $\forall x \forall y[T(x,y)]$. Assuming that 'John' and 'Mary' indicate arguments, it follows that John trusts everyone, and that everyone trusts Mary--i.e., $\forall y[T(j,y)]$ and $\forall x[T(x,m)]$. It follows from all three propositions that John trusts Mary: $T(j,m)$. Frege's rule of inference, which captures the key logical role of the universal quantifier, is that one can replace a variable bound by a universal quantifier with a name for some individual in the domain. Similarly, one can replace a name with a variable bound by an existential quantifier. (Schematically: $\forall x(...x...)$, therefore $...n...$; and $...n...$, therefore $\exists x(....x....)$. Given $T(j,m)$, it follows that someone trusts Mary and that John trusts someone: $\exists x[T(x, m)]$; and $\exists x[T(j,x)]$. It follows from all three propositions that someone trusts someone: $\exists x \exists y[T(x,y)]$. (A single quantifier can bind multiple argument positions, as in '$\exists x[T(x,x)]$'; but this means that someone trusts herself.)

Mixed quantification introduces an interesting wrinkle. The propositions indicated by '$\exists x \forall y[T(x,y)]$' and '$\forall y \exists x[T(x,y)]$' differ. We can paraphrase the first as 'there is someone who trusts everyone' and the second as 'everyone is trusted by someone or other'; the second follows from the first, but not *vice versa*. This raises the possibility that 'someone trusts everyone' is *ambiguous*--that it can indicate either of two propositions. (Although this in turn raises difficult questions about what natural language expressions

are, and what it is for an expression to indicate a proposition. But for Frege, the important point was the distinction between the propositions--or *thoughts*, as he often called them. Similar remarks apply to '$\forall x \exists y[T(x,y)]$' and '$\exists y \forall x[T(x,y)]$'.) A related phenomenon is exhibited by 'John danced if Mary sang and Chris slept'. Is the indicated proposition of the form '(**A** if **B**) and **C**' or '**A** if (**B** and **C**)'? Is someone who says 'The artist drew a club' talking about a sketch or a card game? One can use 'is' to express identity, as in 'Hesperus is the planet Venus'; but in 'Venus is bright', 'is' indicates predication. In 'Venus is a planet', 'a' is logically inert; yet in 'John saw a planet', 'a' indicates existential quantification: $\exists x[P(x) \& S(j,x)]$. (Rendering 'Venus is a planet' as '$\exists x[P(x) \& x = v]$' treats 'is a planet' differently than 'is bright' by appealing, unnecessarily, to quantification and identity.)

According to Frege, such ambiguities provide further evidence that natural language is not ideally suited to the task of representing propositions and inferential relations *perspicuously*, and he wanted a language that was up to the task. (Leibniz and others had envisioned a 'Characteristica Universalis', but without suggestions for how to proceed beyond syllogistic logic in creating one; and given Frege's interest in the foundations of arithmetic, he was especially interested in claims, like 'every number has a successor'.) This is not to say that natural language is ill-suited for other purposes, like communication, or that natural language is useless as a tool for representing propositions. Rather, Frege suggested that natural language is like the eye, whereas a good formal language is like a microscope that reveals structure not otherwise observable.

5. Descriptions and Analysis

Frege did not distinguish--or at least did not emphasize any distinction between--names like 'John' and descriptions like 'the boy' or 'the tall boy in the garden'. Initially, both kinds of expression appear to indicate arguments, as opposed to functions; so one might think that the propositional contribution of 'the...', whatever its syntactic complexity, is just the individual that satisfies the description. On this view, the logical form of 'the boy sang' is 'Sang(*b*)', where '*b*' is a an unstructured symbol that designates the boy in question. But this makes syntactic elements of the description logically irrelevant, and this seems wrong. If the boy sang, it follows that some boy sang; if the tall boy in the garden sang, some tall boy sang. Moreover, 'the' implies *uniqueness* in a way that 'some' does not. One can say 'the boy sang' without denying that universe contains a plurality of boys, but one implies that there is exactly one contextually relevant boy.

In general, if the *P* is *D*, it follows that some *P* is *D*, and that there is only one (relevant) *P*. Or put another way: there is a *P*, and there is at most one *P*, and it is *D*. Russell held that the logical form of 'the boy sang' reflects these implications: $\exists x\{\text{Boy}(x) \& \forall y[\text{Boy}(y) \leftrightarrow x = y] \& \text{Sang}(x)\}$. The middle conjunct was Russell's way of expressing uniqueness; given an object *x*, '$\forall y[\text{Boy}(y) \leftrightarrow x = y)]$' says that everything is such that it is a boy if and only if it is identical with *x*. But however one formulates the middle conjunct, 'the boy' does not correspond to any constituent of Russell's formalism. This reflects his main point: while a speaker may refer to a boy in saying 'the boy sang', the boy is not a constituent of the proposition indicated. The proposition has the form of an existential quantification with a bound variable; it does *not* have the form of a function saturated by an argument--the boy referred to. In this respect, 'the

boy' is like 'some boy'; but not even 'the' indicates a constituent on Russell's view.

Natural language can thus mislead us about the constituency of the propositions we assert. Russell went on to apply this point to a now famous puzzle. Even though France is kingless, 'the king of France is bald' indicates a proposition; the sentence is not, in that sense, meaningless. If the proposition consists of the function indicated by 'Bald( )' and an argument indicated by 'the king of France', there must be an argument so indicated. But what is it? Appeal to nonexistent kings is, to say the least, dubious. Russell concluded that 'the King of France is bald' indicates a quantified proposition: $\exists x\{K(x) \,\&\, \forall y[K(y) \leftrightarrow x = y] \,\&\, B(x)\}$. And one should not be led into thinking otherwise by the following spurious reasoning: every proposition is true or false; so the king of France is bald or not; so there is a present king of France (somewhere), and he is either bald or not. For let '¶' stand for the proposition that the king of France is bald. Russell grants that ¶ is true or false. In fact, it is false, since there is no king of France; given $\neg\exists x[K(x)]$, it follows that $\neg\exists x\{K(x) \,\&\, \forall y[K(y) \leftrightarrow x = y] \,\&\, B(x)\}$. But it hardly follows that there is a king of France who is either bald or not. Russell thinks the confusion lies in a failure to distinguish the negation of ¶ from: $\exists x\{K(x) \,\&\, \forall y[K(y) \leftrightarrow x = y] \,\&\, \neg B(x)\}$; the king of France is not bald. And the natural language expression 'the king of France is bald or not' fosters such a confusion.

The idea that philosophical puzzles might *dissolve*, if only we understood the logical forms of our claims, attracted copious attention. Wittgenstein argued, in his influential *Tractatus Logico-Philosophicus*, that: (i) the very possibility of meaningful sentences, which can be true or false depending on how the world is, requires propositions with structures of the sort Frege and Russell were getting at; (ii) all propositions are logical compounds of--and thus analyzable into--atomic propositions that are inferentially independent of one another, though even simple natural language sentences may indicate complex propositions; and (iii) the right analyses would, given a little reflection, reveal all philosophical puzzles as confusions about how language is related to the world. Russell never endorsed (iii). But for reasons related to epistemological puzzles, Russell held that we are *directly acquainted* with the constituents of those propositions into which every proposition (that we can grasp) can be analyzed; we are not directly acquainted with mind-independent objects that cause our various sensory perceptions; and so the apparent referents of proper names (in natural language) are not constituents of basic propositions.

This led Russell to say that such names are disguised descriptions. On this view, 'Venus' is associated with a (perhaps complex) predicate, and 'Venus is bright' indicates a proposition of the form: $\exists x\{V(x) \,\&\, \forall y[V(y) \leftrightarrow x = y] \,\&\, B(x)\}$. This has the tendentious consequence that for some predicate 'V', it follows that $\exists x[V(x)]$, if Venus is bright; though perhaps, *pace* Russell, 3V' is an unanalyzable predicate true of exactly one mind-indepedent planet. (A related view is that 'Venus is bright' shares its logical form, not with 'The nearest planet is bright', but with '*That* planet is bright'.)

Questions about names are related to psychological reports, like 'Mary thinks Venus is bright', which present puzzles of their own. At least intially, such reports seem to indicate propositions that are neither atomic nor *logical* compounds of simpler propositions; and as Frege noted in a famous paper, it seems that replacing one name with another name for the same object can affect the truth of a psychological report. If Mary fails to know that Hesperus *is* Venus, she might think Venus is a planet without thinking

Hesperus is a planet; yet any function that has the value *true* given Venus as argument has the value *true* given Hesperus as argument. So unsurprisingly, Frege, Russell, and Wittgenstein all held--in varying ways--that psychological reports are misleading with respect to their logical form, since intuitively coreferential names can be associated with distinct propositional contributions. (Wittgenstein later noted that claims like ‘This is red’ and ‘This is yellow’ present difficulties for his view: if the indicated propositions are unanalyzable, and thus logically independent, each should be compatible with the other; yet it is hard to envision any analysis that accounts for the apparent impeccabilty of ‘This is red, so this is not yellow’. This raises questions about whether *all* inferential security is due to logical form.)

6. Restricted Quantifiers

Within the Frege-Russell-Wittgenstein tradition, which flourished in England and America, it became a commonplace that logical form and grammatical form often diverge--even if the most ambitious analytic projects did not succeed. But recent work on quantifiers suggests that the divergence may have been exaggerated, because of how the idea of variable-binding was implemented. Consider again the proposition that some boy sang, and the proposed logical division into the quantifier and the rest:

> $\exists x$[Boy(x) & Sang(x)];
> For some x, x is a boy and x sang.

While this captures the truth conditions of the English sentence, one can also offer a ‘logical paraphrase’ that more closely parallels the grammatical division ‘some boy / sang’: for some x such that x is a boy, x sang. One can formalize this alternative, by using *restricted* quantifiers, which (as the terminology suggests) incorporate a restriction on the domain over which the variable in question ranges. For example, ‘$\exists x$:Boy(x)’ is an existential quantifier that binds a variable ranging over boys in the unrestricted domain. So ‘$\exists x$:Boy(x)[Sang(x)]’ is interpreted as: for some x such that Boy(x), x sang; that is, some (individual who is a) boy sang. This is logically equivalent to ‘$\exists x$[Boy(x) & Sang(x)]’.

Universal quantifiers can also be restricted, as in ‘$\forall x$:Boy(x)[Sang(x)]’, which is interpreted as follows: for every x such that Boy(x), x sang; that is, every (individual who is a) boy sang. This is logically equivalent to ‘$\forall x$[Boy(x) $\rightarrow$ Sang(x)]’; although one might well think the inferential difference between ‘some boy sang’ and ‘every boy sang’ lies entirely with the propositional contributions of ‘some’ and ‘every’--and not (in part) with the contribution of alleged connectives between predicates. Restrictors can be logically complex, as in ‘Some tall boy sang’ or ‘Some boy who respects Mary sang’, which are rendered as ‘$\exists x$:Tall(x) & Boy(x)[Sang(x)]’ and ‘$\exists x$:Boy(x) & Respects(x, mary)[Sang(x)]’.

From this perspective, words like ‘someone’, and the grammatical requirement that ‘every’ be followed by a noun (like ‘boy’), reflect the form of those quantifiers that figure in propositions indicated by natural language sentences. Such quantifiers are composed of a *determiner*, indicated by words like ‘some’ and ‘every’, and a (restricting) predicate that is true of individuals of some sort. One can think of determiners as functions from ordered pairs of predicates to truth values--or equivalently, as functions *from* predicates *to* functions from predicates to truth values. This makes explicit that ‘every’ and ‘some’ indicate relations

between predicates, much as transitive verbs express relations between individuals.

Since ‘*x*’ and ‘*y*’ are variables ranging over individuals, one can say that the function indicated by the transitive verb ‘loves’ has the value *true* given the ordered pair <*x*,*y*> as argument if and only if *x* loves *y*. In this notational scheme, ‘*y*’ corresponds to the direct object (or *internal* argument), which combines with the verb to form a complex predicate (like ‘loves Chris’); ‘*x*’ corresponds to the grammatical subject (or *external* argument) of the verb. If we think about ‘every boy sang’ analogously, ‘boy’ is the internal argument with which ‘every’ combines to form a phrase. We can now introduce ‘*X*’ and ‘*Y*’ as variables ranging over functions from individuals to truth values, stipulating that the *extension* of such a function is the set of things that the function maps onto the value *true*. Then one can say that the function indicated by ‘every’ has the value *true* given the ordered pair <*X*, *Y*> as argument if and only if the extension of *X* includes the extension of *Y*. Similarly, the function indicated by ‘some’ has the value *true* given the ordered pair <*X*, *Y*> as argument if and only if the extension of *X* intersects with the extension of *Y*. At this point, we no longer need the symbols ‘∃’ and ‘∀’. For we can say that ‘Every:Boy(*x*)[Sang(*x*)]’ is true if and only if the singers included the boys; and ‘Some:Boy(*x*)[Sang(*x*)]’ is true if and only if the singers and the boys intersect. (Moreover, the truth condition for ‘Most boys sang’ cannot be captured with ‘∃’, ‘∀’, and the sentential connectives. But we can say that most boys sang if and only if: the boys who are also singers outnumber the boys who are not singers; or put another way, the number of boys who sang exceeds the number of boys who did not sing. So we can say that ‘most’ indicates a function that takes the value *true* given <*X*, *Y*> as argument if and only if: the number of things that *Y* and *X* both map onto *true* exceeds the number of things that *Y* maps onto *true* but *X* does not.)

The Russellian formula

$$\exists x\{\text{Boy}(x) \,\&\, \forall y[\text{Boy}(y) \rightarrow x = y] \,\&\, \text{Sang}(x)\}$$

can be replaced with

$$\text{Some:Boy}(x)[\text{Sang}(x)] \,\&\, |\text{Boy}| = 1$$

interpreted as follows: for some *x*, *x* a boy, *x* sang; and there is exactly one (relevant) boy. On this view, ‘the boy’ still does not correspond to a constituent of the formalism; nor does ‘the’. But one can also treat ‘the’ as a determiner in its own right, as in

‘The:Boy(*x*)[Sang(*x*)]’,

while specifying the propositional contribution of this determiner as follows:

The <*X*, *Y*> = *true* if |*Y*| = 1 & the extensions of *X* and *Y* intersect, and *false* otherwise.

This version of Russell's theory still preserves his central claim. While there may be a boy one refers to in saying ‘the boy sang’, the boy is not a constituent of the indicated proposition, which is quantificational--

involving two predicates and a logical relation between them. But far from showing that the logical form of ‘the boy sang’ diverges dramatically from its grammatical form, the restricted quantifier notation suggests that the logical form closely parallels the grammatical form.

It is worth noting, briefly, an implication of this point for the inference ‘the boy sang, so some boy sang’. If the logical form of ‘the boy sang’ is

Some:Boy(x)[Sang(*x*)] & |Boy|=1,

then the inference is an instance of the simple schema ‘*A* & *B*; so A’. But if the logical form of ‘the boy sang’ is simply ‘The:Boy(*x*)[Sang(*x*)]’, both premise and conclusion of the inference have the abstract form ‘Determiner:Boy(*x*)[Sang(*x*)]’; in which case, the impeccability of the inference depends on the specific propositional contributions of ‘the’ and ‘some’. (If the validity of an inference depends solely on the propositional contributions of its logical elements, perhaps ‘the’ and ‘some’ indicate logical elements that are inferentially related.)

7. Transformational Grammar

Still, the subject/predicate structures of ‘Mary trusts every doctor’ and ‘Some boy trusts every doctor’ diverge from the logical forms of the indicated propositions. Even with restricted quantifiers, and rewriting ‘*T*(*x*, *y*)’ as ‘(*x*)*T*(*y*)’, the formal sentences ‘Every:Doctor(*y*){(mary)Trusts(*y*)}’ and ‘[Some:Boy(*x*)][Every:Doctor(*y*)]{(*x*)Trusts(*y*)]}’ differ from the spoken English sentences. But in thinking about the relation of logic to grammar, one must not assume a naive conception of the latter. For example, the grammatical form of a sentence need not be determined by the linear order of its words. We can distinguish the logical form ‘(**A** if **B**) and **C**’ from ‘**A** if (**B** and **C**)’. And we can distinguish ‘Mary{saw [the boy (with binoculars)]}’ from ‘Mary{[saw (the boy)](with binoculars)}’; with the former indicating that the boy had binoculars, while the latter indicates that Mary used binoculars to see the boy. Diagnosing ambiguity may be just one among many reasons for positing unobvious grammatical structure. And it turns out that the study of *natural* language suggests a rich conception of grammatical form that diverges from the Aristotelian tradition in just the way that modern quantificational logic does.

A leading idea of modern linguistics is that at least some grammatical structures are *transformations* of other structures; and this is related to the idea that words (and phrases) often appear to be *displaced* from the positions typically associated with the relevant grammatical roles. For example, the word ‘who’ in (13) is associated with the second (direct object) argument position of the verb ‘saw’:

(13) Mary wondered who John saw.

Correspondingly, (13) can be glossed as ‘Mary wondered for which x (x a person) John saw x’. This suggests that (13) reflects a transformation of the ‘deep structure’ (13d) into the ‘surface structure’ (13s):

(13d) Mary wondered [John saw who]

(13s) Mary wondered [who_i [John saw $[_]_i$]]

In (13d), the embedded clause has the same basic form as ‘John saw Bill’; and in (13s), ‘who’ has been displaced from the argument position represented by the blank. Evidently, ‘who’ is also displaced in ‘Who did John see’. (Similar remarks apply to ‘why’, ‘what’, ‘when’, ‘how’, etc.)

One might also explain the synonymy of (14) and (15), by positing a common deep structure, (14d):

(14) John seems to like Mary

(14s) John_i seems ($[_]_i$ to like Mary)

(14d) Seems(John likes Mary)

(15) It seems John likes Mary

Since every English sentence needs a grammatical subject, (14d) must be modified: either by displacing ‘John’, as in (14s); or by inserting a dummy subject ‘it’, which does not indicate an argument, as in (15). (Compare ‘There is something in the garden’, which is synonymous with ‘Something is in the garden’.) Appeal to displacement also lets one distinguish the superficially parallel sentences (16) and (17):

(16) John is easy to please.

(17) John is eager to please

If (16) is true, it is easy (for somone) to please John, and so someone can easily please John; but if (17) is true, John is eager that he please someone or other. This asymmetry is effaced by representations like ‘Easy-to-please(John)’ and ‘Eager-to-please(John)’. But the contrast is made manifest by:

(16s) John_i is easy [*e* to please $[_]_i$]

(17s) John_i is eager [$[_]_i$ to please *e*]

where ‘*e*’ indicates an unvoiced (or empty) argument position. On this view of grammar, the ‘surface subject’ may in fact be the *object* of a verb embedded within the main predicate. Of course, such hypotheses about (hidden) grammatical structure require defense. But Chomsky and others have argued that such hypotheses are needed to account for a range of data concerning human linguistic capacities.

As an illustration of the *kind* of data that is relevant, note that (18-20) are perfectly fine expressions of English, while (21) is word salad.

(18) The boy who sang was happy.

(19) Was the boy who sang happy.

(20) The boy who was happy sang.

(21) *Was the boy who happy sang.

This suggests that the auxiliary verb ‘was’ can be displaced from some positions but not others. That is, while (19s) is a permissible transformation of (18d), (21s) is not a permissible transformation of (20d):

(18d) {The [boy (who sang)]} [was happy]

(19s) Was$_i$ <{the [boy (who sang)]} [[_]$_i$ happy]>

(20d) {The [boy (who <was happy>)]} sang}

(21s) *Was$_i$ <{the [boy (who < [_]$_i$ happy>)]} sang}>

The ill-formedness of (21s) is striking, since one can ask if the boy who was happy sang. One can also ask whether (22) is true. But (23) corresponds to ‘The boy who lost was kept crying’, and *not* (22):

(22) The boy who was lost kept crying.

(23) Was the boy who lost kept crying.

This is precisely what one would expect, if ‘was’ cannot be displaced from its position in (22):

*Was$_i$ <{the [boy (who [_]$_i$ lost)]} [kept crying]}>

Such explanations appeal to nonobvious grammatical structure, and constraints on transformations. (For example, no ‘fronting’ an auxiliary verb from an embedded clause.) A sentence was thus said to have a deep structure (DS), which reflects Fregean function-argument stucture, as well as a surface structure (SS); and linguists posited various constraints on DS, SS, and the transformations that relate them. But as the theory was elaborated and refined under empirical pressure, linguists concluded that DS and SS failed to reflect all the grammatically relevant features of sentences. It was argued that another level of grammatical structure, obtained by an operation on SS, was needed to account for certain linguistic facts. The hypothesized transformation, called *quantifier raising* because it targeted the kinds of expressions that indicate (restricted) quantifiers, mapped structures like (24s) onto structures like (24LF).

(24s) {(some boy) [trusts (every doctor)]}.

(24LF) [some boy]$_i$[every doctor]$_j${ [_]$_i$ [trusts [_]$_j$]}

Clearly, (24LF) does not reflect the pronounced word order in English. But the idea is that (24s) can determine the pronounced form of the sentence, while also serving as input to the new transformation. The label ‘LF’ (intimating ‘logical form’) was used for a *grammatical* structure, since the *scope* of a

natural language quantifier is determined by its position at LF--and not by its position at DS or SS. Moreover, the mapping between (24LF) and the following logical formalism is trivial:

[some:boy(*x*)][every:doctor(*y*)]{Trusts(*x*,*y*)}

And (24LF) differs from another structure, interpreted as:

[every:doctor(*y*)][some:boy(*x*)]{Trusts(*x*,*y*)}.

[every doctor]$_i$[some boy]$_j${ [_]$_j$ [trusts [_]$_i$]}

There is a large body of work suggesting that many logical properties of quantifiers, names, and pronouns are similarly reflected in properties of LF. To take just one example, if (25) is true, it follows that some doctor treated some doctor; whereas (26) does not have this consequence:

(25) Every boy saw the doctor who treated himself.
(26) Every boy saw the doctor who treated him.

The truth conditions of (25-26) seem to be (respectively):

[every:boy(*x*)][the:doctor(*y*) & treated(*y*,*y*)]{saw(*x*, *y*)]}

[every:boy(*x*)][the:doctor(*y*) & treated(*y*,*x*)]{saw(*x*, *y*)}

This suggests that ‘himself’ is behaving like a variable bound by ‘the doctor’, while ‘every boy’ can bind ‘him’. And there are independent grammatical reasons for saying that ‘himself’ must be linked to ‘the doctor’, while ‘him’ must not be so linked:

Every boy saw [the doctor]$_i$ who treated [himself]$_i$
*[Every boy]$_i$ saw the doctor who treated [himself]$_i$

[Every boy]$_i$ saw the doctor who treated [him]$_i$
*Every boy saw [the doctor]$_i$ who treated [him]$_i$

While there is a conceptual distinction between LF and the traditional notion of logical form, perhaps the LF of a sentence is at least isomorphic to the logical form of the indicated proposition. (If so, one might avoid certain questions prompted by Frege's view of natural language. How can a sentence indicate a proposition with a *different* structure? And if grammar is deeply misleading, why think our intuitions of (im)peccability provide *reliable* evidence about which propositions follow from which?

8. Semantic Structure and Events

The notion of logical form has also played a significant role in theories of meaning for natural languages. One would like to show how the meaning of a complex expression depends on (i) the meanings of its constituents, and (ii) the ways in which the constituents are arranged. It seems that ‘every tall sailor respects some doctor’ and ‘some short boy likes every politician’ exhibits a common mode of semantic combination; and the meaning of each sentence is presumably fixed by this mode of combination, given the relevant word meanings. This claim is sometimes expressed as follows: the meaning of a sentence is determined by its logical form and the meanings of its parts. This assumes that sentences, like propositions, have logical forms. But in so far as grammatical form mirrors logical form, one might say that the meaning of a sentence S is determined by S's grammatical form and the meanings of its words. (Indeed, some theorists identify the logical form of a sentence with its semantic structure3the way in which its parts are arranged so as to create a complex whose meaning is determined compositionally; and some identify grammatical form with LF, especially in light of Chomsky's recent work, which eschews constraints on DS and SS in favor of constraints on the generation of LFs.)

Davidson and others have urged a version of this idea that draws on Tarski's development of Frege's work. Tarski showed how to provide finitely statable interpretations for certain formal languages, which generate arbitrarily many sentences, by employing recursive rules that assign semantic values to every expression of the language given assignments of semantic values to primitive elements of the language. (This is related to the idea that an inference is valid if and only if: every interpretation that makes the premises true makes the conclusion true, holding fixed the interpretations of logical elements like ‘if’ and ‘every’.) Davidsonians try to construct Tarski-style theories that assign truth conditions to natural language sentences, given the extensions of words and the relevant logical forms. For on their view, a sentence S means that p, if a recursive theory of the right sort assigns the following truth conditions to S: *true* if p, and *false* otherwise. (Montague and others have pursued a similar line of thought.)

For present purposes, the details of this program are less important than Davidson's claim that in constructing any such theory of meaning, we should attend to inferential relations like those exhibited in

(27) Juliet kissed Romeo quickly at midnight.

(28) Juliet kissed Romeo quickly.

(29) Juliet kissed Romeo at midnight.

(30) Juliet kissed Romeo.

If (27) is true, so are (28-30); if (28) or (29) is true, so is (30). The inferences seem impeccable. But the function-argument structure of (27) is not obvious. If we represent ‘kissed quickly at midnight’ as a unstructured predicate that takes two arguments, like ‘kissed’ or ‘kicked’, we will represent the inference from (27) to (30) as having the form: $K^*(x,y)$; so $K(x,y)$. But this form is exemplified by the bad inference: Juliet kicked Romeo; so Juliet kissed Romeo. Put another way, if ‘kissed quickly at midnight’ is a logically unstructured binary predicate, then the following claim is an extra (nonlogical) assumption:

if *x* kissed *y* in a certain manner at a certain time, *x* kissed *y*. But this seems like a tautology, not an assumption that introduces epistemic risk. Davidsonians thus hold that the surface appearances of sentences like (27-30) mask semantic structure; in particular, there is hidden quantification over *events*.

The true form of (30) is said to be manifested by the paraphrase ‘there was a kissing of Romeo by Juliet’. One can formalize this proposal in various ways: ∃*e*[Kissing(*e*) & Of(*e*, Romeo) & By(*e*, Juliet)]; or ∃*e*[Kiss(*e*, Juliet, Romeo)], with the verb ‘kiss’ indicating a function that takes three arguments; or

> (30) ∃*e*[Agent(*e*, Juliet) & Kissing(*e*) & Patient(e, Romeo)]

with Juliet and Romeo being explicitly represented as players of certain *roles* in the event--roughly, the doer (Agent) and the done to (Patient). But whatever the notation, adverbs like ‘quickly’ and ‘at midnight’ are said to indicate further features of the event described, as shown below:

> (27a) ∃*e*[Agent(*e*, Juliet) & Kissing(*e*) & Patient(*e*, Romeo) & Quick(*e*) & At-midnight(*e*)]
>
> (28a) ∃*e*[Agent(*e*, Juliet) & Kissing(*e*) & Patient(*e*, Romeo) & Quick(*e*)]
>
> (29a) ∃*e*[Agent(*e*, Juliet) & Kissing(*e*) & Patient(*e*, Romeo) & At-midnight(*e*)]

If this is correct, then the inference from (27) to (30) is an instance of the following valid form:

> $\exists e[A(e, x) \ \& \ K(e) \ \& \ P(e, y) \ \& \ Q(e) \ \& \ M(e)]$,
> therefore $\exists e[A(e, x) \ \& \ K(e) \ \& \ P(e, y)]$

And the other impeccable inferences involving (27-30) are similarly instances of conjunction reduction. If the grammatical form of (30) is simply ‘{Juliet [kissed (Romeo)]}’, then the mapping from grammatical to logical form is not transparent; and the grammar is misleading--or at least not entirely forthcoming--in that no *word* corresponds to the event quantifier. But a growing body of literature (in philosophy and linguistics) suggests that Davidson's proposal captures an important feature of natural language semantics, and that ‘event analyses’ provide a useful framwork for future discussions of logical form.

Bibliography

The following books provide a useful overview of the history and basic subject matter of logic:

- Kneale, W. & Kneale, M. 1962: *The Development of Logic*. Oxford: OUP, reprinted 1984.
- Sainsbury, M. 1991: *Logical Forms*. Oxford: Blackwell.
- Broadie, A. 1987: *Introduction to Medieval Logic*. Oxford: OUP.

Frege's *Begriffschrift* can be found in:

- *From Frege to Gödel*, J. van Heijenoort, ed. Cambridge, MA: Harvard, 1967.

His later work on functions (‘Function and Concept’) and belief ascriptions (‘Sense and Reference’) can be found in:

- *Translations from the Philosophical Writings of Gottlob Frege*, P. Geach & M. Black (trans.), Oxford: Blackwell, 1952.

For these purposes, Russell's most important books are:

- *Introduction to Mathematical Philosophy* (London: George Allen and Unwin, 1919)
- *Our Knowledge of the External World* (New York: Norton, 1929)
- *The Philosophy of Logical Atomism* (La Salle, Ill: Open Court, 1985).

See also the introduction to the latter, by David Pears; and *Russell*, by Mark Sainsbury (London: Routledge & Kegan Paul, 1979). Stephen Neale's *Descriptions* (Cambridge, MA: MIT Press, 1990) is a recent development of Russell's theory. Wittgenstein's *Tractatus Logico-Philosophicus* (London: Routledge & Kegan Paul, 1961) is importantly related to Russell's logical atomism.

Two key articles on restricted quantifiers, and a third reviewing recent work, are:

- Barwise, J. & Cooper, R. 1981: Generalized Quantifiers and Natural Language. *Linguistics and Philosophy* 4: 159-219.
- Higginbotham, J. & May, R. 1981: Questions, Quantifiers, and Crossing. *Linguistic Review* 1: 47-79.
- Keenan, E. 1996: The Semantics of Determiners. In S. Lappin, ed., *The Handbook of Contemporary Semantic Theory*, Oxford: Blackwell

For introductions to Transformational Grammar and Chomsky's conception of natural language, see:

- Radford, A., 1988: *Transformational Grammar*. Cambridge: CUP.
- Haegeman, L. 1994: *Introduction to Government & Binding Theory*. Cambridge: Blackwell.
- Chomsky, N. 1986: *Knowledge of Language*. New York: Praeger.

And for discussions of work in linguistics bearing directly on issues of logical form:

- Higginbotham, J. 1985: On Semantics. *Linguistic Inquiry*: 16 547-93.
- Hornstein, N. 1995: *Logical Form: From GB to Minimalism*. Oxford: Blackwell.
- Larson, R. and Segal, G. 1995: *Knowledge of Meaning*. Cambridge, MA: MIT Press.
- May, R. 1985: *Logical Form: Its Structure and Derivation*. Cambridge, MA: MIT Press.

- Neale, S. 1993: *Grammatical Form, Logical Form, and Incomplete Symbols*. In A. Irvine & G. Wedeking, eds., *Russell and Analytic Philosophy* Toronto: University of Toronto.

For discussions of the Davidsonian program (briefly described in section 8) and appeal to events:

- Davidson, D. 1984: *Essays on Truth and Interpretation*. Oxford: OUP.
- Davidson, D. 1967: The Logical Form of Action Sentences. In his *Essays on Actions and Events*, Oxford: OUP, 1984.
- Davidson, D. 1985: Adverbs of Action. In B. Vermazen and M. Hintikka, eds., *Essays on Davidson: Actions and Events*, Oxford: Clarendon Press
- Evans, G. & McDowell, J. (eds.) 1976: *Truth and Meaning*. Oxford: OUP.
- Higginbotham, J. 1986: Davidson's Program in Semantics. In E. Lepore, ed., *Truth and Interpretation (*Oxford: Blackwell).
- Lycan, W. 1984: *Logical Form in Natural Language*. Cambridge, MA: MIT Press.
- Parsons, T. 1990: *Events in the Semantics of English* Cambridge, MA: MIT Press.
- Schein, B. 1993: *Plurals*. Cambridge, MA: MIT Press.
- Taylor, B. 1985: *Modes of Occurrence*. Oxford: Blackwell.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle: logic | Davidson, Donald | descriptions | Frege, Gottlob | Frege, Gottlob: logic, theorem, and foundations for arithmetic | logic: classical | logical consequence | propositions: singular | propositions: structured | quantification | Russell, Bertrand

Acknowledgements

The author would like to Christopher Menzel for spotting an error in the formulation of the truth conditions for the generalized quantifier ‘every’, at the end of Section 6. This led to some revision and improvement in that discussion.

Paul M. Pietroski
pietro@wam.umd.edu

First published: October 19, 1999
Content last modified: March 14, 2002

Intrinsic vs. Extrinsic Properties

I have some of my properties purely in virtue of the way I am. (My mass is an example.) I have other properties in virtue of the way I interact with the world. (My weight is an example.) The former are the intrinsic properties, the latter are the extrinsic properties. This seems to be an intuitive enough distinction to grasp, and hence the intuitive distinction has made its way into many discussions in ethics, philosophy of mind, metaphysics and even epistemology. Unfortunately, when we look more closely at the intuitive distinction, we find reason to suspect that it conflates a few related distinctions, and that each of these distinctions is somewhat resistant to analysis.

1. Introduction

The standard way to introduce the distinction between intrinsic and extrinsic properties is by the use of a few platitudes. Stephen Yablo provides perhaps the most succinct version: “You know what an intrinsic property is: it’s a property that a thing has (or lacks) regardless of what may be going on outside of itself.” (1999: 479). David Lewis provides a more comprehensive list of platitudes.

> A sentence or statement or proposition that ascribes intrinsic properties to something is entirely about that thing; whereas an ascription of extrinsic properties to something is not entirely about that thing, though it may well be about some larger whole which includes that thing as part. A thing has its intrinsic properties in virtue of the way that thing itself, and nothing else, is. not so for extrinsic properties, though a thing may well have these in virtue of the way some larger whole is. The intrinsic properties of something depend only on that thing; whereas the extrinsic properties of something may depend, wholly or partly, on something else. If something has an intrinsic property, then so does any perfect

> duplicate of that thing; whereas duplicates situated in different surroundings will differ in their extrinsic properties. (Lewis 1983a: 111-2)

As we shall see, the last claim Lewis makes (that duplicates never differ with respect to intrinsic properties) is somewhat controversial. The other way to introduce the subject matter is by providing examples of paradigmatic intrinsic and extrinsic properties. One half of this task is easy: everyone agrees that being an uncle is extrinsic, as is being six metres from a rhodadendron. The problem with using this method to introduce the distinction is that there is much less agreement about which properties are intrinsic. Lewis has in several places (1983a, 1986a, 1988) insisted that shape properties are intrinsic, but one could hold that an object's shape depends on the curvature of the space in which it is embedded, and this might not even be intrinsic to that space (Nerlich 1979), let alone the object. Lewis also mentions charge and internal structure as being examples of intrinsic properties.

1.1 Philosophical Importance

The distinction between intrinsic and extrinsic properties plays an essential role in stating several interesting philosophical problems. Historically, the most prominent of these has to do with notions of intrinsic value. G. E. Moore (1903: §18) noted that we can make a distinction between things that are good in themselves, or possess intrinsic value, and those that are good as a means to other things. To this day there is still much debate over whether this distinction can be sustained (Feldman 1998, Kagan 1998), and if it can which kinds of things possess intrinsic value (Krebs 1999). In particular, one of the central topics in contemporary environmental ethics is the question of which kinds of things (intelligent beings, conscious beings, living things, species, etc) might have intrinsic value. While this is the oldest (and still most common) use of the intrinsic/extrinsic distinction in philosophy, it has not played much role in the discussions of the distinction in metaphysics, to which we now turn.

As P. T. Geach (1969) noted, the fact that some object *a* is not *F* before an event occurs but is *F* after that event occurs does not mean that the event constitutes, in any deep sense, a change in *a*. To use a well-worn example, at the time of Socrates's death Xanthippe became a widow; that is, she was not a widow before the event of her husband's death, but she was a widow when it ended. Still, though that event constituted (or perhaps was constituted by) a change in Socrates, it did not in itself constitute a change in Xanthippe. Geach noted that we can distinguish between real changes, such as what occurs in Socrates when he dies, from mere changes in which predicates one satisfies, such as occurs in Xanthippe when Socrates dies. The latter he termed 'mere Cambridge' change. There is something of a consensus that an object undergoes real change in an event iff there is some **intrinsic** property they satisfied before the event but not afterwards.

David Lewis (1986a, 1988) built on this point of Geach's to mount an attack on **endurantism**, the theory that objects persist by being wholly located at different times, and that there can be strict identity between an object existing at one time and one existing at another time. Lewis argues that this is inconsistent with the idea that objects undergo real change. If the very same object can be both *F* (at one time) and not *F* (at another), this means that *F*-ness must be a relation to a time, but this means that it is

not an intrinsic property. So any property that an object can change must be extrinsic, so nothing undergoes real change. Lewis says that this argument supports the rival theory of **perdurantism**, which says that objects persist by having different temporal parts at different times. While this argument is controversial (see Haslanger (1989), Johnston (1987) and Lowe (1988) for some responses), it does show how considerations about intrinsicness can resonate within quite different areas of metaphysics.

The other major area where the concept of intrinsicness has been put to work is in stating various supervenience theses. Frank Jackson (1998) defines physicalism in terms of duplication and physical duplication, which are in turn defined in terms of intrinsic properties. This definition builds upon a similar definition offered by Lewis (1983b). Similarly, Jaegwon Kim (1982) defines a mind/body supervenience thesis in terms of intrinsic properties. As Theodore Sider (1993) notes, the simplest way to define the **individualist** theory of mental content that Tyler Burge (1979) attacks is as the claim that the content of a thinker's propositional attitudes supervenes on the intrinsic properties of the thinker. And many **internalist** theories in epistemology are based around the intuition that whether a thinker is justified in believing some proposition supervenes on the intrinsic properties of the thinker.

Though these are the most prominent uses of the intrinsic/extrinsic distinction in philosophy, they by no means exhaust its uses. Many applications of the distinction are cited by I. L. Humberstone (1996), including the following. George Schlesinger (1990) uses the distinction between intrinsic and extrinsic properties to state a non-trivial version of Mill's principle of the uniformity of nature, though Schlesinger gives his distinction a different name. Wlodzimierz Rabinowicz (1979) uses the distinction to formulate principles of universalizability for moral principles and natural laws. And E. J. Khamara (1988) uses a distinction between relational and non-relational properties to state a non-trivial version of the principle of Identity of Indiscernibles.

1.2 Global and Local

Whether a property is intrinsic, and whether some individual that has that property has it intrinsically, are different issues. The property *being square or married* is no doubt an extrinsic property; but it is a property that is had intrinsically by all squares (assuming *being square* is intrinsic). Once we have these two concepts, a 'global' concept of intrinsicness of properties, and a 'local' concept of a particular object being intrinsically such as to possess some property, we might wonder how they are connected. (The names 'global' and 'local' are taken from Humberstone 1996). In particular, we might wonder which of these should be primary in an analysis of intrinsicness.

At first glance, the principles (GTL) and (LTG) seem to connect the two concepts.

> (GTL) If *F* is a (globally) intrinsic property, and *a* is *F*, then *a* is intrinsically *F*

> (LTG) If every *a* that is *F* is intrinsically *F*, then *F* is a (globally) intrinsic property.

(GTL) is undoubtedly true, but (LTG) is more problematic. If the quantifier in it is actualist (i.e. only

ranges over actual objects), then it is clearly false. Let *F* be the property *being square or being inside a golden mountain*. Even if the quantifier is possibilist, it is not clear that (LTG) should be true. For a problematic example, let *F* be the property *being square or being such that the number 21 does not exist*. Every possible object that is *F* is square, and hence intrinsically *F*, but it is not clear that *F* is an intrinsic property. This question (like a few others we will discuss below) turns on the metaphysics of properties. If two properties that are necessarily coextensive are identical (as Lewis believes), or are guaranteed to be alike in whether they are intrinsic or extrinsic (as Sider 1993 argues), then *F* will be intrinsic. If properties can be individuated more finely than this, and if their intrinsicness or otherwise turns on this fine-grained individuation, then maybe *F* is not intrinsic. We will return to this issue in some of the discussions below.

2. Notions of Intrinsicness

Many different distinctions have been called the intrinsic/extrinsic distinction. As J. Michael Dunn (1990) notes, some authors have used ‘intrinsic’ and ‘extrinsic’ to mean ‘essential’ and ‘accidental’. Dunn is surely right in saying that this is a misuse of the terms. A more interesting distinction is noted by Brian Ellis (1991; discussed in Humberstone 1996: 206). Ellis suggests we should distinguish between properties that objects have independently of any outside forces acting on them (what we will call the Ellis-intrinsic properties), and those that they have in virtue of those outside forces (the Ellis-extrinsic properties). For many objects (such as, say, a stretched rubber band) their shape will be dependent on the outside forces acting on them, so their shape will be Ellis-extrinsic. If one is committed to the idea that shapes are intrinsic, one should think this means that the distinction between the Ellis-intrinsic and Ellis-extrinsic properties is not the same as the intrinsic/extrinsic distinction. Such a judgement may seem a little hasty, but in any case we will turn now to distinctions that have received more attention in the philosophical literature.

2.1 Relational vs. Non-Relational Properties

Many writers, especially in the literature on intrinsic value, use ‘relational’ for the opposite of intrinsic. This seems to be a mistake for two reasons. The first reason is that many properties seem to be both be relational and intrinsic. For example, most people have the property *having longer legs than arms*, and indeed seem to have this property intrinsically, even though the property consists in a certain relation being satisfied. Maybe the property is not intrinsic if whether or not something is an arm or a leg is extrinsic, so perhaps this isn’t a conclusive example, but it seems troubling. As Humberstone notes, some might respond by suggesting that a relational property is one such that if an object has it, then it bears some relation to a distinct thing. But this won’t do either. *Not being within a mile of a rhodadendron* is clearly relational, but does not consist in bearing a relation to any distinct individual, as we can see by the fact that a non-rhodadendron all alone in a world can satisfy it.

A larger problem is that it seems *being intrinsic* and *being relational* are properties of two very different kinds of things. Consider again the property *F*, *being square or being such that the number 21 does not*

exist. Assuming (as we do for now) that we can make sense of the relational/non-relational distinction, *F* is a relational property. But *F* is necessarily co-extensive with the property *being square*, which is surely non-relational. So two necessarily co-extensive properties can differ with respect to whether they are relational. We can put this point a few different ways. If any two properties that are necessarily co-extensive are identical, then being relational is not a property of *properties*, but a property of *concepts*, or in any case of some things individuated as finely as Fregean concepts. If we think that intrinsicness is not a property that makes such fine distinctions, then the relational/non-relational and intrinic/extrinsic distinctions are quite different, for they are distinctions between different kinds of things.

2.2 Qualitative vs. Non-Qualitative Properties

As noted above, one of the platitudes Lewis lists when isolating the concept of intrinsicness is that duplicates never differ with respect to their intrinsic properties. Lewis holds a further principle that may not be obvious from the above quote: that any property with respect to which duplicates never differ is intrinsic. Of course, this is only true if the quantifiers in it are interpreted as possibilist. Otherwise the property *having a greater mass than any man that has ever existed* will be intrinsic, since it never differs between actual duplicates. We will assume from now on that all quantifiers are possibilist, unless otherwise noted. And, following Humberstone, we will say that the properties that do not differ between duplicates are the qualitative properties, which is not to say they are not also the intrinsic properties.

Despite this two-way connection between intrinsicness and duplication, we do not yet have an analysis because the relevant concept of duplication can only be (easily) analysed in terms of intrinsic properties. In the next section we will look at the two ways Lewis has attempted to analyse that concept and hence break into the circle. But for now it is worth looking at some results that follow directly from the idea that intrinsic properties are those that do not differ between duplicates.

First, as Humberstone (1996: 227) notes, if this is our definition of intrinsicness, then we can easily analyse local intrinsicness in terms of global intrinsicness. An object *x* is intrinsically *P* iff all its duplicates are *P*, that is, if all objects that have the same (global) intrinsic properties as it does are *P*. And having this concept of local intrinsicness is quite useful, because it lets us explain what is right about an intuitively attractive (though ultimately mistaken) claim about intrinsic properties. Let *P* be an intrinsic property and *Q* a property such that an object's having *Q* is entailed by its having *P*. One might think in those circumstances that *Q* would be intrinsic, since its possession follows from a fact solely about the object in question, namely that it is *P*. This isn't right in general; to see why let *P* be the property of *being square* and *Q* be the property *being square or being inside a golden mountain*. For some objects that are *Q* their *Q*-ness follows from facts solely about the object, but for others it follows from facts quite extrinsic to the object in question. But, Humberstone notes, something similar is true. If *x* possess *P* intrinsically, and being *P* entails being *Q*, then *x* possesses *Q* intrinsically. This local concept of intrinsicness might also do philosophical work; presumably the intrinsic value of an object depends on which properties it intrinsically possesses, not on which intrinsic properties it possesses.

Secondly, it follows from the definition that necessarily co-extensive properties are alike in being

intrinsic or not. In particular, any property that every possible object has, such as *being such that the number 21 exists*, will be an intrinsic property. Robert Francescotti (1999) takes this to be a decisive mark against Lewis's theory, but others (e.g. Sider 1993, Weatherson 2001) have been willing to treat it as a philosophical discovery. It is crucial to the proof that Lewis's theory entails that this property is intrinsic that the quantifiers in the theory are possibilist. If we let the quantifiers range over the right kind of impossibilities (such as the situations of Barwise and Perry 1983, or the impossible worlds of Nolan 1997) then one can have duplicates that, for example, differ with respect to whether the number 21 exists. Since this approach has not been developed in any detail it is impossible to say at this time whether it would have serious untoward consequences.

Thirdly, the duplication relation is transitive, so any duplicate of a duplicate of David Lewis, is a duplicate of David Lewis. That means that *being a duplicate of David Lewis* is an intrinsic property of all those objects. While this might plausibly be a property that Lewis intrinsically possesses, it is somewhat surprising that it is intrinsic to all of his duplicates. Dunn (1990) reports that Lewis in conversation said that this property (being a duplicate of David Lewis) is equivalent to an infinite conjunction of intrinsic properties (the ones Lewis has) so it should turn out to be intrinsic.

Conversely, assuming the metaphysics of counterpart theory, none of these duplicates of David Lewis is David Lewis himself, so the property *being (identical with) David Lewis* turns out to be extrinsic on this account. Even if we drop the counterpart theory, and allow that objects in different worlds might be strictly identical, still not all duplicates of Lewis will be identical with Lewis, so the property *being (identical with) David Lewis* will still not be intrinsic. It might seem rather odd that a property so internal to Lewis should not be intrinsic. Yablo (1999) notes that in some cases, such as this one, we can make an argument for identity properties not being intrinsic. If it is essential to David Lewis that he be descended from a particular zygote Z, then the fact that something is David Lewis entails that something else is a zygote, and any property whose possession entails the existence of other objects is usually held to be extrinsic. Still, Yablo argues, it is very plausible that the identity properties of some things (especially atoms) should be intrinsic.

Finally, we can define relative notions of duplication, and hence relative notions of intrinsicness (Humberstone 1996: 238). To give just one interesting example, say that a property is nomically intrinsic iff it never differs between duplicates in worlds with the same laws. Then many dispositional properties might turn out to be nomically intrinsic, capturing nicely the idea that they are in a sense internal to the objects that possess them, while their manifestation depends both on external facts, and on the laws being a certain way.

2.3 Interior vs. Exterior Properties

J. Michael Dunn (1990) suggests that odd consequences of Lewis's theory are sufficient to look for a new concept of intrinsicness. He suggests that the notion of intrinsicness is governed by platitudes like the following. "Metaphysically, an *intrinsic* property of an object is a property that the object has by virtue of itself, depending on no other thing. Epistemologically, an intrinsic property would be a property

that one could determine by inspection of the object itself - in particular, for a physical object, one would not have to look outside its region of space-time" (1990: 178) As Dunn notes, the metaphysical definition here is the central one, the epistemological platitude is at best a heuristic. On this view, *being identical to X* will be intrinsic (except in cases where *X* has essential properties rooted outside itself), while *being a duplicate of X* will not be. Also, *having X as a part* will be intrinsic, though it is not on Lewis's account.

Dunn argues that the appropriate logic in which to formulate claims of intrinsicness and to investigate what consequences they have is the relevant logic **R**, and he provides some of the details of how this should be done. But until we see a specific formulation of the idea (comparable in specificity to the accounts of Peter Vallentyne and Stephen Yablo, discussed below in section 3.3), we cannot comment on its consequences. Still, there is an intuitive distinction here, and it clearly differs from the distinction Lewis discusses.

2.4 Which is the real distinction?

If we grasp the three distinctions discussed above, we might well ask which of them is the intrinsic/extrinsic distinction? It is possible that this question has no determinate answer. Humberstone suggests that we have three interesting distinctions here, each of which can do some philosophical work, and there is not much interest in the issue of which of them is called the distinction between intrinsic and extrinsic properties. If we do decide to investigate this seriously, we should perhaps be prepared to be disappointed - there is no guarantee that there will be a fact of the matter which distinction the words 'intrinsic' and 'extrinsic' latch onto.

Should we just give up on identifying *the* intrinsic/extrinsic distinction; then, on pain of having some indeterminacy in our philosophical theories, we must reformulate the theories that are framed using this distinction, specifying which distinction should take the role of the intrinsic/extrinsic distinction in each case. Sider, in the course of defending the philosophical interest of the qualitative/non-qualitative distinction, makes a start on doing this. He notes that in the debates about supervenience, the distinction that is usually relevant is the qualitative/non-qualitative one. If we let *being (identical to) X* be an intrinsic property, then most of the supervenience theses discussed will be trivially true, because it will be impossible to have duplicates that are different objects, and hence impossible to have duplicates that differ with respect to the contents of their beliefs, or the justificatory status of their beliefs, or their phenomenal states, or whatever. But these theses are not trivially true; so if we are to formulate the distinctions this way, we had better not let identity properties be intrinsic *in these contexts*.

This, of course, does not show that the qualitative/non-qualitative distinction is the only one that can do philosophical work. Indeed, when trying to grasp what real change amounts to, it seems to be the interior/exterior distinction that is relevant. Say that *a* has *b* as a part, and consider the event whereby *b* is replaced in *a* by *c*, which happens to be a duplicate of *b*. This event seems to constitute a real change in *a*, not merely a Cambridge change, but it does not constitute a change in qualitative properties.

3. Attempts at Analysis

We will first look at two attempts to analyse the qualitative/non-qualitative distinction, and then at two more ambitious projects that aim to capture intrinsicness in all of its facets.

3.1 Combinatorial Theories

As Yablo noted, if an object has a property intrinsically, then it has it independently of the way the rest of the world is. The rest of the world could disappear, and the object might still have that property. Hence a *lonely* object, an object that has no wholly distinct worldmates, could have the property. Note that in the sense relevant here, two objects are only ‘wholly distinct’ if they have no parts in common, not if they are merely non-identical. The idea is that a lonely object could have proper parts. This is good, since *having six proper parts* is presumably an intrinsic property. Many extrinsic properties could not be possessed by lonely objects – no lonely object is six metres from a rhododendron, for example.

This suggests an analysis of intrinsicness: *F* is an intrinsic property iff it is possible for a lonely object to be *F*. This analysis is usually attributed to Kim (1982) (e.g. in Lewis 1983a and Sider 1993), though Humberstone (1996) dissents from this interpretation. Both directions of the biconditional can be challenged.

Some objects change in mass over time: this is presumably an intrinsic property of those objects. If necessitarian theories of laws are true (as endorsed by Ellis 2001 and Shoemaker 1984), then there could not be a world with just that object, as the conservation of matter would be violated. If any kind of combinatorial analysis of intrinsicness can work, we have to assume something like Hume’s dictum that there are no necessary connections between distinct existences. Indeed, all combinatorial theories of intrinsicness do assume this, and further that the range of what is possible can be taken as given in crafting a theory of intrinsicness. This might be thought problematic, since the best way to formally spell out Hume’s dictum itself appeals to the concept of intrinsicness (Lewis 1986a: 87-91).

The analysis is only viable as an analysis of the qualitative/non-qualitative distinction, since it rules that *being a duplicate of David Lewis* is an intrinsic property. This feature, too, is shared by all combinatorial theories of intrinsicness.

The major problem with this analysis is that the ‘if’ direction of the biconditional is clearly false. As Lewis (1983) pointed out, the property *being lonely* is had by some possible lonely objects, but it is not intrinsic.

Rae Langton and David Lewis (1998) designed a theory to meet this objection. Their theory resembles, in crucial respects, the theory sketched in an appendix to Dean Zimmerman’s paper "Immanent Causation" (Zimmerman 1997). The two theories were developed entirely independently. We will focus on Langton and Lewis’s version here, because it is more substantially developed, and more widely

discussed in the literature. On their theory, a property *F* is *independent of accompaniment* iff the following four conditions are met:

a. There exists a lonely *F*
b. There exists a lonely non-*F*
c. There exists an accompanied (i.e. not lonely) *F*
d. There exists an accompanied non-*F*

Langton and Lewis's idea is that if *F* is intrinsic, then whether *or not* an object is *F* should not depend on whether *or not* it is lonely. So all four of these cases should be possible. Still, some extrinsic properties satisfy all four conditions. Consider, for instance, the property *being lonely and round or accompanied and cubical*. A lonely sphere suffices for (a), a lonely cube for (b), an actual cube for (c) and an actual sphere for (d). So they have to rule out this property. They do it by the following five-step process.

First, Langton and Lewis identify a class of privileged *natural* (or non-disjunctive) properties. Lewis (1983b) had argued that we need to recognise a distinction between natural and non-natural properties to make sense of many debates in metaphysics, philosophy of science, philosophy of language and philosophy of mind, and suggested a few ways we might draw the distinction. We might take the natural properties to be those that correspond to real universals, or those that appear in the canonical formulations of best physics or regimented common sense, or even take the distinction to be primitive. Langton and Lewis say that it should not matter how we draw the distinction for present purposes, as long as we have it, and properties like *being lonely and round or accompanied and cubical* are not natural.

Secondly, they say properties are *disjunctive* iff they "can be expressed by a disjunction of (conjunctions of) natural properties; but are not themselves natural properties." Thirdly, they say a property is *basic intrinsic* iff it is non-disjunctive and satisfies (a) through (d). Fourthly, they say two (possible) objects are duplicates iff they share the same basic intrinsic properties. Finally, they say *F* is an intrinsic property iff two duplicates never differ with respect to it.

Three objections have been pressed against this view. Stephen Yablo (1999) objected to the role of natural properties in the analysis, which he argued introduced irrelevant material, and implied that the theory was at best *de facto*, but not *de jure*, correct. Dan Marshall and Josh Parsons (2001) claimed that according to this definition, the property *being such that a cube exists* is non-disjunctive, but it satisfies (a) through (d), so it would be basic intrinsic, despite being extrinsic. Theodore Sider (2001) claimed that the theory could not handle *maximal* properties: properties of objects that are not shared by their large proper parts. Sider claims that *being a rock* is such a property: large parts of rocks are not rocks. So *being a rock* is extrinsic, since a duplicate of a large part of a rock might be a rock if it is separated from the rest of the rock. But, argued Sider, on some interpretations of 'natural property' it is natural, and hence basic intrinsic.

Brian Weatherson's (2001) theory was designed to meet these three objections. In his theory,

combinatorial principles of possibility are not used to derive characteristics of individual intrinsic properties, as Kim and Langton and Lewis do, but characteristics of the whole set of intrinsic properties. He argues that this set, call it *SI*, will have the following properties:

- If $F \in SI$ and $G \in SI$, then *F and G* $\in SI$ and *F or G* $\in SI$ and *not F* $\in SI$
- If $F \in SI$ then *Having* n *parts that are F* $\in SI$ and *Being entirely composed of exactly* n *things that are F* $\in SI$
- If $F \in SI$ and $G \in SI$ and there is a possible world with n+1 pairwise distinct things, and something in some world is *F* and something in some world is *G*, then there is a world with exactly n+1 pairwise distinct things such that one is *F* and the other n are *G*.
- If $F \in SI$ and $G \in SI$ and it is possible that regions with shapes d_1 and d_2 stand in relation *A*, and it is possible that an *F* wholly occupy a region with shape d_1 and a *G* wholly occupy a region with shape d_2, then there is a world where regions with shapes d_1 and d_2 stand in *A*, and an *F* wholly occupies the region with shape d_1 and a *G* wholly occupies the region with shape d_2.

The first two principles are closure principles on the set. The third principle says that any two intrinsic properties that can be instantiated can be instantiated together any number of times. And the fourth says that if objects having two intrinsic properties can be in two regions, and those two regions can be in a particular spatial relation, then the regions can be in that relation while filled by objects having those properties. The third principle suffices to show that *being such that a cube exists* could not be in *SI*, and the fourth to show that *being a rock* could not be.

Weatherson's theory does not entirely avoid appeals to a concept of naturalness, though the counterexamples that prompt the appeal are now much more recherché. Without such an appeal, then if *F* and *G* are intrinsic properties that atoms could have, nothing in his theory rules out the property *being simple, lonely and F or being G* from being intrinsic. There are a few ways for the appeal to go at this point, see Weatherson (2001) and Lewis (2001) for a few suggestions. The following moves, taken directly from Langton and Lewis, will probably work if any will. Say that the basic intrinsic properties are those non-disjunctive properties such that their membership in *SI* is consistent with the above four principles. Two objects are duplicates if they do not differ with respect to basic intrinsic properties. A property is intrinsic if it never differs between duplicates.

Finally, John Hawthorne (2001) has suggested that all these combinatorial theories have a problem with properties of the form *being R-related to something*, where *R* is a perfectly natural relation that is neither reflexive nor irreflexive. Such properties are extrinsic, but Hawthorne suggests they will satisfy all the combinatorial principles, and their close connection to natural relations means that they will be natural enough to cause problems for all these combinatorial approaches.

3.2 Natural Kind Theories

In *On the Plurality of Worlds*, David Lewis presents a quite different analysis of intrinsic properties. As

with the combinatorial theory that he and Rae Langton defend, it heavily exploits the idea that some properties are more natural than others. In fact, it rests even more weight on it. Here is Lewis's statement of the theory:

> [I]t can plausibly be said that all perfectly natural properties are intrinsic. Then we can say that two things are *duplicates* iff (1) they have exactly the same perfectly natural properties, and (2) their parts can be put into correspondence in such a way that corresponding parts have exactly the same perfectly natural properties, and stand in the same perfectly natural relations…Then we can go on to say that an *intrinsic* property is one that can never differ between duplicates. (Lewis 1986a: 61-2)

Like the combinatorial theories, this is an attempt at analysing qualitative intrinsicness. Anyone who thinks this is too modest an aim to be worthwhile will be disappointed. It rests heavily on the 'plausible' claim that all perfectly natural properties are intrinsic, and, implicitly, that the perfectly natural properties are sufficient to characterise the world completely. The last assumption is needed because the theory rules out the possibility that there are two objects that share all their perfectly natural properties, but differ with respect to some intrinsic property or other. One consequence of these assumptions is that a world is fully characterised by the intrinsic properties of its inhabitants and the perfectly natural relations between those inhabitants. Lewis thinks this is true for the actual world, it is just his doctrine of Humean supervenience. (Lewis 1986b: i-xiii). But it might be thought a stretch to think it is true of *all* worlds.

In their paper of 1998, Langton and Lewis claim the only advantage of their theory over Lewis's old theory is that it makes fewer assumptions about the nature of natural properties. They also note that Lewis still believes those assumptions, but they think it is worthwhile to have a theory that gets by without them. It also seems that Lewis's new theory, perhaps as amended, provides more insight into the nature of intrinsicness.

3.3 Contractions

Peter Vallentyne (1997) develops a theory based around the idea that *x*'s intrinsic properties are those properties it would have if it were alone in the world. He defines a *contraction* of a world as "a world 'obtainable' from the original one solely by 'removing' objects from it." (211) As a special case of this, an *x*-*t* contraction, where *x* is an object and *t* a time, is "a world 'obtainable' from the original one by, to the greatest extent possible, 'removing' all objects wholly distinct from *x*, all spatial locations not occupied by *x*, and all times (temporal states of the world) except *t*, from the world." (211) Vallentyne allows that there might be unique *x*-*t* contraction; sometimes we can remove one of two objects, but not both, from the world while leaving *x*, so there will be one *x*-*t* contraction which has one of these in it, and another that has the other.

He then says that *F* is intrinsic iff for all *x*, *t*, all *x*-*t* contractions are such that *Fx* is true in the contraction iff it is true in the actual world. In short, a property is intrinsic to an object iff removing the rest of the world doesn't change whether the object has the property.

Vallentyne notes that this definition will not be very enlightening unless we understand the idea of a contraction. This seems related to the objection Langton and Lewis (1998) urge against Vallentyne. They say that Vallentyne's account reduces to the claim that a property is intrinsic iff possession of it never differs between an object and its lonely duplicates, a claim they think is true but too trivial to count as an analysis. Their position is that we cannot understand contractions without understanding duplication, but if we understand duplication then intrinsicness can be easily defined, so Vallentyne's theory is no advance.

Stephen Yablo (1999) argues that this criticism is too quick. Vallentyne should best be understood as working within a very different metaphysical framework to Lewis. For Lewis, no (ordinary) object exists at more than one world, so Vallentyne's contractions, being separate worlds, must contain separate objects. Hence *x-t* contractions can be nothing other than lonely duplicates, and the theory is trivial. Yablo suggests that the theory becomes substantive relative to a metaphysical background in which the very same object can appear in different worlds. (In chapter 4 of *Plurality* Lewis has a few arguments against this idea, and Yablo has interesting responses to these arguments. A thorough investigation of this debate would take us too far from the topic.) If this is the case then we can get a grip on contractions without thinking about duplications - the *x-t* contraction of a world is the world that contains *x* itself, and as few other things as possible.

3.4 Francescotti

Robert Franscecotti (1999) recently outlined an analysis that takes the concept of intrinsicness as non-relationality to be primary. Francescotti takes a property to be extrinsic iff an object possesses it in virtue of its relations to other objects. So *being a duplicate of Jack* and *being such that the number 17 exists* are extrinsic, while *being identical to Jack* and *being a vertebrate* (i.e. *having a vertebral column*) are intrinsic. As noted above, this means that we must either have a hyper-intensional notion of properties, or we say that intrinsicness is a property of concepts, not of properties. Francescotti takes the former option.

Francescotti notes that not all relational properties are extrinsic. *Having a vertebral column*, for instance, seems to be relational in that it consists of a relation to a vertebral column, but it is also an intrinsic property. So he focuses on relations to distinct objects. The definition of intrinsicness goes as follows. First we define a d-relational property. *F* is d-relational iff:

> (a) there is a relation *R*, and an item *y*, such that (i) *x*'s having *F* consists in *x*'s bearing *R* to *y*, and (ii) *y is distinct from x*; or
>
> (b) there is a relation *R*, and a class of items *C*, such that (i) *x*'s having *F* consists in there being some member of *C* to which *x* bears *R*, and (ii) at least one member of *C* to which *x* bears *R* is distinct from *x*; or
>
> (c) there is a relation *R*, and a class of items *C*, such that (i) *x*'s having *F* consists in *x*'s

bearing *R* to every member of *C*, and (ii) it is possible that there is a member of *C* that is distinct from *x*.

We then define intrinsic properties as being those that are not d-relational.

F is an intrinsic property of $x =_{df} x$ has *F*, and *F* is not a d-relational property of *x*.

Francescotti's theory provides intuitively plausible answers to all the cases he considers, provided of course that we identify the intrinsic/non-intrinsic distinction with the relational/non-relational distinction, rather than one of the other two distinctions considered in section two. Like the other three theories, it has one unexplained (or perhaps underexplained) primitive, in this case the *consists-in* relation. As Francescotti notes, following Khamara (1988), it won't do to say *x*'s having *F* consists in *x*'s bearing *R* to *y* just in case it is necessary that *Fx* iff *xRy*. That makes it too easy for having a property to consist in some necessary being (say God, or the numbers) being a certain way. Rather, he says, "*x*'s having *F*, consists in the event or state, *x*'s having *G*, just in case *x*'s having *F* is the very same event or state as *x*'s having *G*." (599) Whether this can handle all the hard cases seems to depend how the theory of identity conditions for events and states turns out. (See the entries on Donald Davidson and events for some start on this.)

Bibliography

- Barwise, Jon and John Perry (1983), *Situations and Attitudes*. Cambridge, MIT Press.
- Burge, Tyler (1979), "Individualism and the Mental", in P. French, T. Euhling, and H. Wettstein (eds.), *Studies in Epistemology*. Vol 4, *Midwest Studies in Philosophy*. Minneapolis: University of Minnesota Press.
- Dunn, J. Michael (1990), "Relevant Predication 2: Intrinsic Properties and Internal Relations", *Philosophical Studies* 60: 177-206
- Ellis, Brian (1991), "Scientific Essentialism", Paper presented to the 1991 conference of the Australasian Association for the History and Philosophy of Science.
- Ellis, Brian (2001), *Scientific Essentialism*. Cambridge, Cambridge
- Feldman, Fred (1998), "Hyperventilating about Intrinsic Value", *Journal of Ethics* 2: 339-354
- Francescotti, Robert (1999), "How to Define Intrinsic Properties", *Noûs* 33: 590-609
- Geach, P. T. (1969), *God and the Soul*. London, Routledge
- Haslanger, Sally (1989), "Endurance and Temporary Intrinsics", *Analysis* 49:119-125
- Hawthorne, John (2001), "Intrinsic Properties and Natural Relations", *Philosophy and Phenomenological Research* 63: 399-403
- Humberstone, I. L. (1996), "Intrinsic/Extrinsic", *Synthese* 108: 205-67
- Jackson, Frank (1998), *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford, Oxford
- Johnston, Mark (1987), "Is There a Problem About Persistence?", *Proceedings of the Aristotelian Society* (Supp) 61:107-35
- Kagan, Shelly (1998), "Rethinking Intrinsic Value", *Journal of Ethics* 2: 277-297

- Kim, Jaegwon (1982), “Psychophysical Supervenience”, *Philosophical Studies* 41: 51-70
- Khamara, E. J. (1988), “Indiscernables and the Absolute Theory of Space and Time”, *Studia Leibnitiana* 20: 140-59
- Krebs, Angelika (1999), *Ethics of Nature: A Map*. Hawthorne, de Gruyter
- Langton, Rae and David Lewis (1998), “Defining ‘Intrinsic’”, *Philosophy and Phenomenological Research* 58: 333-45.
- Langton, Rae and David Lewis (2001), “Marshall and Parsons on ‘Intrinsic’”, *Philosophy and Phenomenological Research* 63: 353-5
- Lewis, David (1983a), “Extrinsic Properties”, *Philosophical Studies* 44: 197-200
- Lewis, David (1983b), “New Work for a Theory of Universals”, *Australasian Journal of Philosophy* 61: 343-77
- Lewis, David (1986a), *On the Plurality of Worlds*. Oxford, Blackwell
- Lewis, David (1986b), *Philosophical Papers: Volume 2*. Oxford, Oxford
- Lewis, David (1988), “Rearrangement of Particles: Reply to Lowe”, *Analysis* 48: 65-72
- Lowe, E. J. (1988), “The Problems of Intrinsic Change: Rejoinder to Lewis”, *Analysis* 48: 72-77
- Marshall, Dan and Josh Parsons (2001), “Langton and Lewis on “*Intrinsic*””, *Philosophy and Phenomenological Research* 63: 347-51
- Moore, G. E. (1903), *Principia Ethica*. Cambridge, Cambridge
- Nerlich, Graham (1979), “Is Curvature Intrinsic to Physical Space”, *Philosophy of Science* 46: 439-58.
- Nolan, Daniel (1998), “Impossible Worlds: A Modest Approach”, *Notre Dame Journal of Formal Logic* 38: 535-73
- Rabinowicz, Wlodzimierz (1979), *Universalizability*. Dordrecht, Reidel
- Schlesinger, George (1990), “Qualitative Identity and Uniformity”, *Noûs* 24: 529-41
- Shoemaker, Sydney (1984), *Identity, Cause and Mind*. Cambridge, Cambridge
- Sider, Theodore (1993), “Intrinsic Properties”, *Philosophical Studies* 83: 1-27
- Sider, Theodore (2001), “Maximality and Intrinsic Properties”, *Philosophy and Phenomenological Research* 63: 357-64
- Vallentyne, Peter (1997), “Intrinsic Properties Defined”, *Philosophical Studies* 88: 209-19.
- Weatherson, Brian (2001), “Intrinsic Properties and Combinatorial Principles”, *Philosophy and Phenomenological Research*, 63: 365-80
- Yablo, Stephen (1999), “Intrinsicness”, *Philosophical Topics* 26: 479-505
- Zimmerman, Dean (1997), “Immanent Causation”, *Philosophical Perspectives* 11: 433-71

Other Internet Resources

- "Intrinsic Properties", preprint of a paper by Ted Sider (Syracuse University)
- "The Stage View and Temporary Intrinsics", a paper by Ted Sider (Syracuse University)
- "Intrinsicness", draft of a paper by Steven Yablo (MIT)
- "Intrinsic Properties and Combinatorial Principles", paper by Brian Weatherson (Brown University)

[Please contact the author with other suggestions.]

Related Entries

abstract objects | Davidson, Donald | events | identity: of indiscernibles | justification, epistemic: internalist vs. externalist conceptions of | mereology | properties

Brian Weatherson
Brian_Weatherson@Brown.edu

First published: January 4, 2002
Content last modified: January 4, 2002

Thermodynamic Asymmetry in Time

Macroscopic processes appear to be temporally "directed" in some sense. The spontaneous evolution of systems is always to a future but not past equilibrium state. The nature of this directedness concerns many deep questions at the foundations of philosophy and science.

Thermodynamics is the science that describes much of the time-asymmetric behavior found in the world. This entry's first task, consequently, is to show how thermodynamics treats temporally 'directed' behavior. It then concentrates on the following two questions. (1) What is the origin of the thermodynamic asymmetry in time? In a world possibly governed by time-symmetric laws, how should we understand the time-asymmetric laws of thermodynamics? (2) Does the thermodynamic time asymmetry explain the other temporal asymmetries? Does it account, for instance, for the fact that we know more about the past than the future? The discussion thus divides between thermodynamics being an explanandum or explanans. In the former case the answer will concern philosophy of physics; in the latter case it will concern metaphysics, epistemology, and other fields, though in each case there will be blurring between the disciplines.

1. Thermodynamic Time Asymmetry: A Brief Guide

Consider the following.

> Place some chlorine gas in a small closed flask into the corner of a room. Set it up so that an automaton will remove its cover in 1 minute. Now we know what to do: run. Chlorine is poison, and furthermore, we know the gas will spread reasonably quickly through its available volume. The chlorine originally in equilibrium in the flask will, upon being freed,

> 'relax' to a new equilibrium.

Or less dramatically:

> Place an iron bar over a flame for half an hour. Place another one in a freezer for the same duration. Remove them and place them against one another. Within a short time the hot one will 'lose its heat' to the cold one. The new combined two-bar system will settle to a new equilibrium, one intermediate between the cold and hot bar's original temperatures. Eventually the bars will together settle to roughly room temperature.

These are two examples of a tendency of systems to spontaneously evolve to equilibrium; but there are indefinitely more examples in all manner of substance. The physics first used to describe such processes is thermodynamics.

First systematically developed in S. Carnot's *Reflections on the Motive Power of Fire* 1824, the science of classical thermodynamics is intimately associated with the industrial revolution. Most of the results responsible for the science originated from the practice of engineers trying to improve steam engines. Begun in France and England in the late eighteeth and early nineteenth centuries, the science quickly spread throughout Europe. By the mid-nineteenth century, Clausius in Germany and Thompson in England had developed the theory in great detail.

Thermodynamics is a 'phenomenal' science, in the sense that the variables of the science range over macroscopic parameters such as temperature and volume. Whether the microphysics underlying these variables are motive atoms in the void or an imponderable fluid is largely irrelevant to this science. The developers of the theory both prided themselves on this fact and at the same time worried about it. Clausius, for instance, was one of the first to speculate that heat consisted solely of the motion of particles (without an ether), for it made the equivalence of heat with mechanical work less surprising. However, as was common, he kept his ontological beliefs separate from his statement of the principles of thermodynamics because he didn't wish to (in his words) "taint" the latter with the speculative character of the former.[1.]

A treatment of thermodynamics naturally begins with the statements it takes to be laws of nature. These laws are founded upon observations of relationships between particular macroscopic parameters and they are justified by the fact they are empirically adequate. No further justification of these laws is to be found -- at this stage -- from the details of microphysics. Rather, stable, counterfactual-supporting generalizations about macroscopic features are enshrined as law. The typical textbook treatment of thermodynamics describes some basic concepts, states the laws in a more or less rough way and then proceeds to derive the concepts of temperature and entropy and the various thermodynamic equations of state. It is worth remarking, however, that in the last forty years the subject has been presented with a degree of mathematical rigor not previously achieved. Originating from the early axiomatization by Caratheodory in 1909, the development of 'rational thermodynamics' has clarified the concepts and logic of classical thermodynamics to a degree not generally appreciated. There now exist many quite different,

mathematically exact approaches to thermodynamics, each starting with different observational regularities as axioms. (For a popular presentation of a recent axiomatization, see Lieb and Yngvson 2000.)

In the traditional approach classical thermodynamics has two laws, the second of which is our main focus. (Readers may have heard of a ’third law’ as well, but it was added later and is not relevant to the present discussion.) The first law expresses the conservation of energy. The law uses the concept of the internal energy of a system, $U(x)$, which is a function of variables such as volume. For thermally isolated (adiabatic) systems--think of systems such as coffee in a thermos--the law states that this function, $U(x)$, is such that the work W delivered to a system’s surroundings is compensated by a loss of internal energy, i.e., $dW = -dU$. When Joule and others showed that mechanical work and heat were interconvertible, consistency with the principle of energy conservation demanded that heat, Q, considered as a different form of energy, be taken into account. For non-isolated systems we extend the law as $dQ = dU + dW$, where dQ is the differential of the amount of heat added to the system (in a reversible manner).

The conservation of energy tells us nothing about temporally asymmetric behavior. In particular, it doesn’t follow from the first law that interacting systems quickly tend to approach equilibrium (a state where the values of the macroscopic variables remain approximately stable), and once achieved, never leave this state. It is perfectly consistent with the first law that systems in equilibrium leave equilibrium. Since this tendency of systems cannot be derived from the First Law, another law is needed. Although S. Carnot was the first to state it, the formulations of Kelvin and Clausius are standard:

> Kelvin’s Second Law: There is no thermodynamic process whose sole effect is to transform heat extracted from a source at uniform temperature completely into work.
>
> Clausius’ Second Law: There is no thermodynamic process whose sole effect is to extract a quantity of heat from a colder reservoir and deliver it to a hotter reservoir.

Kelvin’s version is essentially the same as the version arrived at by both Carnot and Planck, whereas Clausius’ version differs from these in a few ways.[2.]

Clausius’ version transparently rules out anti-thermodynamic behavior such as a hot iron bar extracting heat from a neighboring cold iron bar. The cool bar cannot give up a quantity of heat to the warmer bar (without something else happening). Kelvin’s statement is perhaps less obvious. It stems from the fact familiar from steam engines that heat energy is a ‘poor’ grade of energy. Consider a gas-filled cylinder with a frictionless piston holding the gas down at one end. If we put a flame under the cylinder, the gas will expand and the piston can perform work, e.g., it might move a ball. However, we can never convert the heat energy straight into work without some other effect occurring. In this case, the gas occupies a larger volume.

In 1865 Clausius introduced the notion of the ‘equivalence value’ of a system, a concept that is the ancestor of the modern day concept of entropy. Later in 1865 Clausius used the term ‘entropy’ from the

Greek word for transformation. The entropy of a state *A*, *S*(*A*), for instance, is defined as the integral $S(A) = \int_O^A dQ/T$ over a reversible transformation, where *O* is some arbitrary fixed state. For *A* to have an entropy, the transformation from *O* to *A* must be quasi-static, i.e., a succession of equilibrium states. Continuity considerations then imply that the initial and final states *O* and *A* must also be equilibrium states.

In terms of entropy, the Second Law states that in a transformation from equilibrium state *A* to equilibrium state *B*, the inequality *S*(*B*) - *S*(*A*) is greater than or equal to the $\int_B^A dQ/T$. Loosely put, for realistic systems, this implies that in the spontaneous evolution of a thermally closed system the entropy can never decrease and that it attains its maximum value for states at equilibrium. We are invited to think of the Second Law as driving the gas to its new, higher entropy equilibrium state. Using this concept of entropy, thermodynamics is able to capture an extraordinary range of phenomena under one simple law. Remarkably, whether they are gases filling their available volumes, two iron bars in contact coming to the same temperature, or milk mixing in your coffee, they all have an observable property in common: their entropy increases. Coupled with the First Law, the Second Law is remarkably powerful. It appears that all classical thermodynamical behavior can be derived from these two simple statements (Penrose 1970).[3.]

There are a number of philosophical questions one might ask about the the laws of thermodynamics. For instance, where exactly is time-asymmetry found in the above statement of the Second Law? Why think the Second Law is universal? (See Uffink 2001 for an interesting discussion of these topics.) How are these laws framed in a relativistic universe? Do Lorentz boosted gases appear hotter or colder in the new frame? Surprisingly, the correct (special) relativistic transformation rules for thermodynamic quantities, and thus the relativistic understanding of thermodynamic time asymmetry, is still controversial. With all the current activity of physicists being focused on the thermodynamics of black holes in general relativity and quantum gravity, it is amusing to note that special relativistic thermodynamics is still a field with many open questions, both physically and philosophically. (See Earman 1981 and Liu 1994.)

Another important question concerns the reduction of thermodynamic concepts such as entropy to their mechanical, or statistical mechanical, basis. As even a cursory glance at statistical mechanics reveals, there are many candidates for the statistical mechanical entropy, each the center of a different program in the foundations of the field. Here, again, surprisingly, there is no consensus as to which entropy is best suited to be the reduction basis of the thermodynamic entropy (see Sklar 1993; Callender 1999). Consequently, there is little agreement about what the Second Law looks like in statistical mechanics. Despite the worthiness of these issues, this article will focus on the particularly important problem of the direction of time (though as we'll see, many issues go by this name.)

2. The Problem of the Direction of Time I

This 'problem of the direction of time' has its source in the debates over the status of the second law of

thermodynamics between L. Boltzmann and some of his contemporaries, notably, J. Loschmidt, E. Zermelo and E. Culverwell. Boltzmann sought the mechanical underpinning of the second law. He came up with a particularly ingenious explanation for why systems tend toward equilibrium. Consider an isolated gas of N particles in a box, where N is large enough to make the system macroscopic ($N \approx 10^{23}+$). For the sake of familiarity we will work with classical mechanics. We can characterize the gas by the coordinates and momenta x_{in}, p_{in} of each of its particles and represent the whole system by a point $X = (q,p)$ in a $6N$-dimensional phase space known as Γ, where $q = (q_1 \ldots q_{3N})$ and $p = (p_1 \ldots p_{3N})$.

Boltzmann's great insight was to see that the thermodynamic entropy arguably "reduced" to the volume in Γ picked out by the macroscopic parameters of the system. The key ingredient is partitioning Γ into compartments, such that all of the microstates X in a compartment are macroscopically (and thus thermodynamically) indistinguishable. To each macrostate M, there corresponds a volume of Γ, $|\Gamma_M|$, whose size will depend on the macrostate in question. For combinatorial reasons, almost all of Γ corresponds to a state of thermal equilibrium. There are simply many more ways to be distributed with uniform temperature and pressure than ways to be distributed with nonuniform temperature and pressure. There is a vast numerical imbalance in Γ between the states in thermal equilibrium and the states in thermal nonequilibrium.

We can now introduce Boltzmann's famous entropy formula (up to an additive constant):

$$S_B(M(X)) = k \log |\Gamma_M|$$

where $|\Gamma_M|$ is the volume in Γ associated with the macrostate M, and k is Boltzmann's constant. S_B provides a relative measure of the amount of Γ corresponding to each M. Given the mentioned asymmetry in Γ, almost all microstates are such that their entropy value is overwhelmingly likely to increase with time. When the constraints are released on systems initially confined to small sections of Γ, typical systems will evolve into larger compartments. Since the new equilibrium distribution occupies *almost all* of the newly available phase space, nearly all of the microstates originating in the smaller volume will tend toward equilibrium. Except for those incredibly rare microstates conspiring to stay in small compartments, microstates will evolve in such a way as to have S_B increase. Though substantial questions can be raised about the details of this approach, and philosophers can rightly worry about the justification of the standard probability measure on Γ, this explanation seems to offer the correct *framework* for understanding why the entropy of systems tends to increase with time. (For further explanation and discussion see Bricmont 1996, Callender 1999, Goldstein 2001, Klein 1973 and Lebowitz 1993.)

Before Boltzmann described entropy increase as described above, he proposed a now notorious "proof" known as the "H-theorem" to the effect that entropy must always increase. Loschmidt and Zermelo launched objections to the H-theorem. But an objection in their spirit can also be advanced against Boltzmann's later view sketched above. Loosely put, because the classical equations of motion are time reversal invariant (TRI), nothing in the original explanation necessarily referred to the direction of time.

(See Hurley 1985.) Though I just stated the Boltzmannian account of entropy increase in terms of entropy increasing into the future, the explanation can be turned around and made for the past temporal direction as well. Given a gas in a box that is in a nonequilibrium state, the vast majority of microstates that are *antecedents* of the dynamical evolution leading to the present macrostate correspond to a macrostate with *higher entropy* than the present one. Therefore, not only is it highly likely that typical microstates corresponding to a nonequilibrium state will evolve to higher entropy states, but it is also highly likely that they *evolved from* higher entropy states.

Concisely put, the problem is that given a nonequilibrium state at time t_2, it is overwhelmingly likely that

> (1) the nonequilibrium state at t_2 will evolve to one closer to equilibrium at t_3

but that due to the reversibility of the dynamics it is also overwhelmingly likely that

> (2) the nonequilibrium state at t_2 has evolved from one closer to equilibrium at t_1

where $t_1 < t_2 < t_3$. However, transitions described by (2) do not seem to occur; or phrased more carefully, not both (1) and (2) occur. However we choose to use the terms ‘earlier’ and ‘later,’ clearly entropy doesn’t increase in both temporal directions. For ease of exposition let us dub (2) the culprit.

The traditional problem is not merely that nomologically possible (anti-thermodynamic) behavior does not occur when it could. That is not straightforwardly a problem: *all sorts* of nomologically allowed processes do not occur. Rather, the problem is that statistical mechanics seems to make a prediction that is falsified, and that is a problem according to anyone’s theory of confirmation.

Many solutions to this problem have been proposed. Generally speaking, there are two ways to solve the problem: eliminate transitions of type (2) either with special boundary conditions or with laws of nature. The former method works if we assume that *earlier* states of the universe are of comparatively low-entropy *and* that (relatively) *later* states are not also low-entropy states. There are no high-to-low-entropy processes simply because earlier entropy was very low. Alternatively, the latter method works if we can somehow restrict the domain of physically possible worlds to those admitting only low-to-high transitions. The laws of nature are the straightjacket on what we deem physically possible. Since we need to eliminate transitions of type (2) while keeping those of type (1) (or vice versa), a necessary condition of the laws doing this job is that they be time reversal noninvariant. Our choice of strategy boils down to either assuming temporally asymmetric boundary conditions or of adding (or changing to) time reversal noninvariant laws of nature. Many approaches to this problem have thought to avoid this dilemma, but a little analysis of any proposed ‘third way’ arguably proves this to be false.

2.1 Past Hypothesis

Without changing the TRI laws of nature, there is no way to eliminate transition (2) in favor of (1).

Nevertheless, appealing to temporally asymmetric boundary conditions, as we've seen, allow us to describe a world wherein (1) but not (2) occur. A cosmological hypothesis claiming that in the very distant past entropy was much lower will work. Boltzmann, as well as many of this century's greatest scientists, e.g., Einstein, Feynman, and Schroedinger, saw that this hypothesis is necessary given our laws. (Boltzmann, however, explained this low-entropy condition by treating the observable universe as a natural statistical fluctuation away from equilibrium in a vastly larger universe.) Earlier states do not have higher entropy than present states because we make the cosmological posit that the universe began in an extremely tiny section of its available phase space. Albert 2000 calls this the "Past Hypothesis" and provides a detailed discussion of its role in statistical mechanics.

Classical mechanics is also compatiable with a "Future Hypothesis": the claim that entropy is very low in the distant future. The restriction to "distant" is needed, for if the near future were of low-entropy, we would not expect the thermodynamic behavior that we see -- see Cocke 1967, Price 1996 and Schulman 1997 for discussion of two-time boundary conditions.

The main dissatisfaction with this solution is that many do not find it sufficiently *explanatory* of thermodynamic behavior. That a gas in the lab last Wednesday filled its available volume due to special initial conditions may be credible. But that gases everywhere for all time should expand through their available volumes due to special initial conditions is, for some, incredible. The common cause of these events is viewed as unlikely. Expressing this feeling, Penrose 1989 estimates that the probability, given the standard measure on phase space, of the universe starting in the requisite state is astronomically small. Callender 1997, however, assimilates the problem to the general one facing the special sciences -- all special science laws require conspiratorial initial conditions for their generalizations to hold. If the problem really is a problem, according to Callender, it is not necessarily one specific to thermodynamics and time's direction.

2.2 Electromagnetism

The physicist E. Ritz and others have claimed that electromagnetism accounts for the thermodynamic arrow. The wave equation for both mechanical and electromagnetic processes is well-known to include both 'advanced' and 'retarded' solutions. The retarded solution

$$\phi_{\text{ret}}(r,t) = \int dr' \rho \frac{(r', t - \frac{|r'-r|}{c})}{|r'-r|}$$

gives the field amplitude φ_{ret} at *r,t* by finding the source density *r* at r' at earlier times. The advanced solution

$$\phi_{\mathrm{adv}}(r,t) = \int dr' \rho \frac{(r', t + \frac{|r'-r|}{c})}{|r'-r|}$$

gives the field amplitude in terms of the source density at r' at later times. Despite this symmetry nature seems to contain only processes obeying the retarded solutions. (This popular way of stating the electromagnetic asymmetry is actually misleading. The advanced solutions describe the radiation sink's receiving waves, and this happens all the time. The asymmetry of radiation instead lay with the form (concentrated or dispersed) the sources take.)

If we place an isolated concentrated gas in the middle of a large volume, we would expect the particles to spread out in an expanding sphere about the center of the gas, much as radiation spreads out. It is therefore tempting to think that there is a relationship between the thermodynamic and electromagnetic arrows of time. In a debate in 1909, A. Einstein and E. Ritz disagreed about the nature of this relationship. Ritz took the position that the asymmetry of radiation had to be judged lawlike and that the thermodynamic asymmetry could be derived from this law. Einstein's position is instead that "irreversibility is exclusively based on reasons of probability" (Einstein and Ritz 1909, quoted from Zeh 1989, 13). It is unclear whether he meant probability plus the right boundary conditions, or simply probability alone. In any case, Ritz believes the radiation arrow causes the thermodynamic one, whereas Einstein seems to hold something closer to the opposite position.

It seems that Einstein must be right, or at least, closer to being correct than Ritz. Ritz' position appears implausible if only because it implies gases composed of neutral particles will not tend to spread out. That aside, we now think that the wave asymmetry must originate in asymmetric boundary conditions, just as the statistical mechanical asymmetry may. Recall the statistical version of the Second Law. It implies that with the right (improbable) initial conditions a system will undergo improbable-to-probable transitions rather than the reverse. The crucial point to see is that the usual retarded radiation is a kind of improbable-to-probable transition. A concentrated source is improbable, but given its existence, a system will evolve toward more probable regions of the phase space, i.e., the waves will spread. Advanced radiation is likewise a species of improbable-to-probable transitions. Given an improbable source in the past, it will spread backwards in time to more probable regions of the phase space too. Using Popper's famous mechanical wave example as an analogy, throwing a rock into a pond so that waves on the surface spread out into the future requires every bit the conspiracy that is needed for waves to converge on a point in order to eject a rock from the bottom. Both are equally likely, pace Popper; whether one or both happen depends upon the boundary conditions. The real asymmetry lies in the fact that in the past there are concentrated sources for waves, whereas in the future there tend not to be. See Price 1996, Arntzenius 1993, and Frisch 2000 for discussion of this controversial point.

These considerations do not mean the radiation arrow *reduces* in any sense to the thermodynamic arrow. Rather, the thing to say is that the radiation arrow just seems to *be* the statistical mechanical one, with the qualification that the media sustaining the improbable-to-probable transition is electromagnetic.

2.3 Cosmology

Cosmology presents us with a number of apparently temporally asymmetric mechanisms. The most obvious one is the inexorable expansion of the universe. In cosmology the spatial scale factor $a(t)$, which gives the distance between co-moving observers, is increasing. The universe seems to be uniformly expanding relative to our local frame. Since this temporal asymmetry occupies a rather unique status it is natural to wonder whether it might be the 'master' arrow. The cosmologist T. Gold 1962 proposed just this. Believing that entropy values covary with the size of the universe, Gold asserts that at the maximum radius the thermodynamic arrow will 'flip' due to the re-contraction. However, as Tolman 1936 has shown in some detail, a universe filled with non-relativistic particles will not suffer entropy increase due to expansion, nor will an expanding universe uniformly filled with blackbody radiation increase its entropy either. Interestingly, Tolman demonstrated that more realistic universes containing both matter and radiation *will* change their entropy contents. Coupled with expansion, various processes will contribute to entropy increase, e.g., energy will flow from the 'hot' radiation to the 'cool' matter. So long as the relaxation time of these processes is larger than the expansion time scale, they should generate entropy. We thus have a purely cosmological method of entropy generation.

Others (e.g., Davies 1994) have thought inflation provides a kind of entropy-increasing behavior -- again, given the sort of matter content we have in our universe. The inflationary model is an alternative of sorts to the standard big bang model, although by now it is so well entrenched in the cosmology community that it really deserves the tag 'standard'. In this scenario, the universe is very early in a quantum state called a 'false vacuum', a state with a very high energy density and negative pressure. Gravity acts like Einstein's cosmological constant, so that it is repulsive rather than attractive. Under this force the universe enters a period of exponential inflation, with geometry resembling de Sitter space. When this period ends any initial in-homogeneities will have been smoothed to insignificance. At this point ordinary stellar evolution begins. Loosely associating gravitational homogeneity with low-entropy and inhomogeneity with higher entropy, inflation is arguably another source of cosmological entropy generation.

There are other proposed sources of cosmological entropy generation, but these should suffice to give the reader a flavor of the idea. We shall not be concerned with evaluating these scenarios in any detail. Rather, our concern is about how these proposals explain time's arrow. In particular, how do they square with our earlier claim that the issue boils down to either assuming temporally asymmetric boundary conditions or of adding time reversal non-invariant laws of nature?

The answer is not always clear, owing in part to the fact that the separation between laws of nature and boundary conditions is especially slippery in the science of cosmology. Advocates of the cosmological explanation of time's arrow typically see themselves as explaining the origin of the needed low-entropy cosmological condition. Some explicitly state that special initial conditions are needed for the thermodynamic arrow, but differ with the conventional 'statistical' school in deducing the origin of these initial conditions. Earlier low-entropy conditions are not viewed as the boundary conditions of the spacetime. They came about, according to the cosmological schools, about a second or more after the big

bang. But when the universe is the size of a small particle, a second or more is enough time for some kind of cosmological mechanism to bring about our low-entropy 'initial' condition. What cosmologists (primarily) differ about is the precise nature of this mechanism. Once the mechanism creates the 'initial' low-entropy we have the same sort of explanation of the thermodynamic asymmetry as discussed in the previous section. Because the proposed mechanisms are supposed to make the special initial conditions inevitable or at least highly probable, this maneuver seems like the alleged 'third way' mentioned above.

The central question about this type of explanation, as far as we're concerned, is this: Is the existence of the low 'initial' state a consequence of the laws of nature alone or the laws plus boundary conditions? In other words, first, does the proposed mechanism produce low-entropy states given *any* initial condition, and second, is it a *consequence* of the laws *alone* or a consequence of the laws *plus initial conditions*? We want to know whether our question has merely been shifted back a step, whether the explanation is a disguised appeal to special initial conditions. Though we cannot here answer the question in general, we can say that the two mechanisms mentioned are not lawlike in nature. Expansion fails on two counts. There are boundary conditions in expanding universes that do not lead to an entropy gradient, i.e., conditions without the right matter-radiation content, and there are boundary conditions that do not lead to expansion, e.g., matter-filled Friedman models that do not expand. Inflation fails at least on the second count. Despite advertising, arbitrary initial conditions will not give rise to an inflationary period (Earman 1995, pp. 152-3). Furthermore, it's not clear that inflationary periods will give rise to thermodynamic asymmetries (Price 1996, ch. 2). The cosmological scenarios do not seem to make the thermodynamic asymmetries a result of nomic necessity. The cosmological hypotheses may be true, and in some sense, they may even explain the low-entropy initial state. But they do not appear to provide an explanation of the thermodynamic asymmetry that makes it nomologically necessary or even likely.

Another way to see the point is to consider the question of whether the thermodynamic arrow would 'flip' if (say) the universe started to contract. Gold, as we said above, asserts that at the maximum radius the thermodynamic arrow must 'flip' due to the re-contraction. Not positing a thermodynamic flip while maintaining that entropy values covary with the radius of the universe is clearly inconsistent -- it is what Price 1996 calls the fallacy of a "temporal double standard". Gold does not committ this fallacy, and so he claims that the entropy must decrease if ever the universe started to re-contract. However, as Albert 2000 writes, "there are plainly locations in the phase space of the world from which ... the world's radius will inexorably head up and the world's entropy will exorably head down". Since that it is the case, it doesn't follow from law that the thermodynamic arrow will flip during re-contraction; therefore, without changing the fundamental laws, the Gold mechanism cannot explain the thermodynamic arrow in the sense we want.

From these considerations we can understand what Price 1996 calls the *basic dilemma*: either we explain the earlier low-entropy condition Gold-style or it is inexplicable by time-symmetric physics (82). Because there is no net asymmetry in a Gold universe, we might paraphrase Price's conclusion in a more disturbing manner as the claim that the (local) thermodynamic arrow is explicable just in case (globally) there isn't one. However, notice that this remark leaves open the idea that the laws governing expansion or inflation are not TRI. (For more on Price's basic dilemma, see Callender 1998 and Price 1995.)

2.4 Quantum Cosmology

Quantum cosmology, it is often said, is the *theory* of the universe's initial conditions. Presumably this entails that its posits are to be regarded as lawlike. Because theories are typically understood as containing a set of laws, quantum cosmologists apparently assume that the distinction between laws and initial conditions is fluid. Particular initial conditions will be said to obtain as a matter of law. Hawking 1987 writes, for example, "we shall not have a complete model of the universe until we can say more about the boundary conditions than that they must be whatever would produce what we observe," (163). Combining such aspirations with the observation that thermodynamics requires special boundary conditions leads quite naturally to the thought that “the second law becomes a selection principle for the boundary conditions of the universe [for quantum cosmology]” (Laflamme 1994, 358). In other words, if one is to have a theory of initial conditions, it would certainly be desirable to deduce initial conditions that will lead to the thermodynamic arrow. This is precisely what many quantum cosmologists have sought.[5.] Since quantum cosmology is currently very speculative, it has been argued that it is premature to start worrying about what it says about time's arrow (Callender 1998). Nevertheless, there has been a substantial amount of debate on this issue (see Haliwell *et al*, 1994).

2.5 Time Itself

Some philosophers have sought an answer to the problem of time's arrow by claiming that time *itself* is directed. They do not mean time is asymmetric in the sense intended by advocates of the tensed theory of time. Their proposals are firmly rooted in the idea that time and space are properly represented on a four-dimensional manifold. The main idea is that the asymmetries in time indicate something about the nature of time itself. Christensen 1993 argues that this is the most economical response to our problem since it posits nothing besides time as the common cause of the asymmetries, and we already believe in time. A proposal similar to Christensen's is Weingard's 1977 ‘time-ordering field’. Weingard's speculative thesis is that spacetime is temporally oriented by a ‘time potential,’ a timelike vector field that at every spacetime point directs a vector into its future light cone. In other words, supposing our spacetime is temporally orientable, Weingard wants to actually orient it. The main virtue of this is that it provides a time sense everywhere, even in spacetimes containing closed timelike curves (so long as they're temporally orientable). As he shows, any explication of the ‘earlier than’ relation in terms of some other physical relation will have trouble providing a consistent description of time direction in such spacetimes. Another virtue of the idea is that it is in principle capable of explaining *all* the temporal asymmetries. If coupled to the various asymmetries in time, it would be the ‘master arrow’ responsible for the arrows of interest. As Sklar 1985 notes, Weingard's proposal makes the past-future asymmetry very much like the up-down asymmetry. As the up-down asymmetry was reduced to the existence of a gravitational potential -- and not an asymmetry of space itself -- so the past-future asymmetry would reduce to the time potential -- and not an asymmetry of time itself. Of course, if one thinks of the gravitional metric field as part of spacetime, there is a sense in which the reduction of the up-down asymmetry really was a reduction to a spacetime asymmetry. And if the metric field is conceived as part of spacetime -- which is itself a huge source of contention in philosophy of physics -- it is natural to think of Weingard's time-ordering field as also part of spacetime. Thus his proposal shares a lot in common

with Christensen's suggestion.

This sort of proposal has been criticized by both Earman and Sklar on methodological grounds. Sklar 1985, for instance, claims that scientists would not accept such an explanation (111-2). One might point out, however, that many scientists did believe in analogues of the time-ordering field as possible causes of the CP violations.[6.] The time-ordering field, if it exists, would be an unseen (except through its effects) common cause of strikingly ubiquitous phenomena. Scientists routinely accept such explanations. To find a problem with the time-ordering field we need not invoke methodological scruples; instead we can simply ask whether it does the job asked of it. Is there a mechanism that will couple the time-ordering field to thermodynamic phenomena? Weingard says the time potential field needs to be suitably coupled (p. 130) to the non-accidental asymmetric processes, but neither he nor Christensen elaborate on how this is to be accomplished. Until this is addressed satisfactorily, this speculative idea must be considered interesting yet embryonic.

2.6 Interventionism

When explaining time's arrow, many philosophers and physicists have focused their attention upon the unimpeachable fact that real systems are open systems that are subjected to interactions of various sorts.[7.] We can not truly isolate thermodynamic systems, and even if we could, it would probably not be for all time. To take the most obvious example, we can not shield a system from the influence of gravity. At best, we can move systems to locations feeling less and less gravitational force, but we can never completely decouple a system from the gravitational field. Not only do we ignore the weak gravitational force when doing classical thermodynamics, but we also ignore less exotic matters, such as the walls in the standard gas in a box scenario. We can do this because the time it takes for a gas to reach equilibrium with itself is vastly shorter than the time it takes the gas plus walls system to reach equilibrium. For this reason we typically discount the effects of the box walls on the gas.

In this approximation many have thought there lies a possible solution to the problem of the direction of time. Indeed, many have thought herein lies a solution that *does not* change the laws of classical mechanics and *does not* allow for the nomological possibility of anti-thermodynamic behavior. In other words, advocates of this view seem to believe it embodies a third way.

The idea is to take advantage of what a random perturbation of the representative phase point would do to the evolution of a system. In phase space there is a tremendous asymmetry between the volume of points leading to equilibrium and points leading away from equilibrium. If the representative point of a system were knocked about randomly, then due to this asymmetry, it would be very probable that the system at any given time be on a trajectory leading toward equilibrium. Thus, if it could be argued that the earlier treatment of the statistical mechanics of ideal systems ignored a random perturber in the environment of the system, then one would seem to have a solution to our problems. Even if the perturbation were weak it would still have the desired effect. The weak 'random' previously ignored knocking of the environment is the sought after cause of the approach to equilibrium. Prima facie, this answer to the problem escapes the appeal to special initial conditions and the appeal to new laws.

But only prima facie. A number of criticisms have been leveled against this maneuver. One that seems on the mark is the observation that if classical mechanics is to be a universal theory, then the environment must be governed by the laws of classical mechanics as well. The environment is not some mechanism outside the governance of physical law, after all, and when we treat it too, the 'deus ex machina' -- the random perturber -- disappears. If we treat the gas-plus-the-container walls as a classical system, it is still governed by time-reversible laws that will cause the same problem as we met with the gas alone. At this point one sometimes sees the response that that combined system of gas plus walls has a neglected environment too, and so on, and so on, until we get to the entire universe. It is then questioned whether we have a right to expect laws to apply universally (Reichenbach 1956, 81ff). Or the point is made that we cannot write down the Hamiltonian for all the interactions a real system suffers, and so there will always be something 'outside' what is governed by the time-reversible Hamiltonian. Both of these points rely, we suspect, on an underlying instrumentalism about the laws of nature. Our problem only arises if we assume or pretend that the world literally is the way the theory says; dropping this assumption naturally 'solves' the problem. Rather than further address these responses, let us turn to the claim that this maneuver need not modify the laws of classical mechanics.

If one does not make the radical proclamation that physical law does not govern the environment, then it is easy to see that whatever law describes the perturber's behavior, it cannot be the laws of classical mechanics *if* the environment is to do the job required of it. A time-reversal noninvariant law, in contrast to the TRI laws of classical mechanics, must govern the external perturber. Otherwise we can in principle subject the whole system, environment plus system of interest, to a Loschmidt reversal. The system's velocities will reverse, as will the velocities of the millions of tiny perturbers. 'Miraculously', as if there were a conspiracy between the reversed system and the millions of 'anti-perturbers', the whole system will return to a time reverse of its original state. What is more, this reversal will be just as likely as the original process if the laws are TRI. A minimal criterion of adequacy, therefore, is that the random perturbers be time reversal *noninvariant*. But the laws of classical mechanics are TRI. Consequently, if this 'solution' is to succeed, it must exercise new laws and modify or supplement classical mechanics. (Since the perturbations need to be genuinely random and not merely unpredictable, and since classical mechanics is deterministic, the same sort of argument could be run with indeterminism instead of irreversibility. See Price 2002 for a diagnosis of why people have made this mistake, and also for an argument objecting to interventionism for offering a 'redundant' physical mechanism responsible for entropy increase.) [8.]

2.7 Quantum Mechanics

To the best of our knowledge, our world is fundamentally quantum mechanical, not classical mechanical. Does this change the situation? 'Maybe' is perhaps the best answer. Not surprisingly, answers to the question are affected by one's interpretation of quantum mechanics. Quantum mechanics suffers from the notorious measurement problem, a problem which demands one or another interpretation of the quantum formalism. These interpretations fall broadly into two types, depending on their view of the unitary evolution of the quantum state (e.g., evolution according to the Schroedinger equation): they either say that there is something more than the quantum state, or that the unitary evolution is not entirely correct.

The former are called ‘no-collapse’ interpretations while the latter are dubbed ‘collapse’ interpretations. This is not the place to go into the details of these interpretations, but we can still sketch the outlines of the picture painted by quantum mechanics (for more see Albert 1992).

Modulo some philosophical concerns about the meaning of time reversal (see Albert 2000, Callender 2000), the equation governing the unitary evolution of the quantum state is time reversal invariant. For interpretations that add something to quantum mechanics, this typically means that the resulting theory is time reversal invariant too (since it would be odd or even inconsistent to have one part of the theory invariant and the other part not). Since the resulting theory is time reversal invariant, it is possible to generate the problem of the direction of time just as we did with classical mechanics. While many details are altered in the change from classical to no-collapse quantum mechanics, the logical geography seems to remain the same.

Collapse interpretations are more interesting with respect to our topic. Collapses interrupt or outright replace the unitary evolution of the quantum state. To date, they have always done so in a time reversal *non*invariant manner. The resulting theory, therefore, is not time reversal invariant. This fact offers a potential escape from our problem: the transitions of type (2) in our above statement of the problem may not be lawful. And this has led many thinkers throughout the century to believe that collapses somehow explain the thermodynamic time asymmetry.

Mostly these postulated methods fail to provide what we want. We think gases relax to equilibrium even when they’re not measured by Bohrian observers or Wignerian conscious beings. This complaint is, admittedly, not independent of more general complaints about the adequacy of these interpretations. But perhaps because of these controversial features they have not been pushed very far in explaining thermodynamics.

More satisfactory collapse theories exist, however. One, commonly known as GRW, can describe collapses in a closed system -- no dubious appeal to observers outside the quantum system is required. Albert (1994; 2001) has extensively investigated the impact GRW would have on statistical mechanics and thermodynamics. GRW would ground a temporally asymmetric probabilistic tendency for systems to evolve toward equilibrium. Anti-thermodynamic behavior is not impossible according to this theory. Instead it is tremendously unlikely. The innovation of the theory lies in the fact that although entropy is overwhelmingly likely to increase toward the future, it is not also overwhelmingly likely to increase toward the past (because there are no dynamic backwards transition probabilities providied by the theory). So the theory does not suffer from a problem of the direction of time as stated above.

This does not mean, however, that it removes the need for something like the Past Hypothesis. GRW is capable of explaining why, given a present nonequilibrium state, later states should have higher entropy; and it can do this without also implying that earlier states have higher entropy too. But it does not explain how the universe ever got into a nonequilibrium state in the first place. As indicated before, some are not sure *what* would explain this fact, if anything, or whether it’s something we should even aspire to explain. The principal virtue GRW would bring to the situation, Albert thinks, is that it would solve or

bypass various troubles involving the nature of probabilities in statistical mechanics.

More detailed discussion of the impact quantum mechanics has on our problem can be found in Albert 2000, North 2002, Price 2002. But if our superficial review is correct, we can say that quantum mechanics will not obviate our need for a Past Hypothesis though it may well solve (on a GRW interpretation) at least one problem related to the direction of time.

2.8 Lawlike Initial Conditions?

Without some new physics that eliminates or explains the Past Hypothesis, or some satisfactory 'third way', it seems we are left with a bald posit of special initial conditions. Again, one can question whether there really is anything unsatisfactory about this (Sklar 1993; Callender 1997). But perhaps we were wrong in the first place to think of the Past Hypothesis as a contingent boundary condition. The question 'why these special initial conditions?' would be answered with 'it's physically impossible for them to be otherwise,' which is always a conversation stopper. Indeed, Feynman (1965, 116) speaks this way when explaining the statistical version of the second law.

Absent a particular understanding of laws of nature, there is perhaps not much to say about the issue. But given particular conceptions of lawhood, it is clear that various judgments about this issue follow naturally -- as we will see momentarily. However, let's acknowledge that this may be to get matters backwards. It might be said that we *first* ought to find out whether the boundary conditions *are* lawlike, and *then* devise a theory of law appropriate to the answer. To decide whether or not the boundary conditions are lawlike based merely on current philosophical theories of law is to prejudge the issue. Perhaps this objection is really evidence of the feeling that settling the issue based on one's conception of lawhood seems particularly unsatisfying. And it is hard to deny this. Even so, it is illuminating to have a brief look at the relationships between some conceptions of lawhood and the topic of special initial conditions.

For instance, if one agrees with Mill that from the laws one should be able to deduce everything and one considers the thermodynamic part of that 'everything,' then the special initial condition will be needed for such a deduction. The modern heir of this conception of lawhood, the one associated with Ramsey and Lewis (see Loewer 1994), sees laws as the axioms of the simplest, most powerful, consistent deductive system possible. It is likely that the specification of a special initial condition would emerge as an axiom in such a system, for such a constraint may well make the laws much simpler than they otherwise would be.

We should not expect the naïve regularity view of laws to follow suit, however. On this sort of account, roughly, if As always follow Bs, then it is a law of nature that A causes B. To avoid finding laws everywhere, however, this account needs to assume that As and Bs are instantiated plenty of times. But the initial conditions occur only once.

For more robust realist conceptions of law, it's difficult to predict whether the special initial conditions

will emerge as lawlike. Necessitarian accounts like Pargetter's 1984 maintain that it is a law that *P* in our world iff *P* obtains at every possible world joined to ours by a nomic accessibility relation. Without more specific information about the nature of the accessibility relations and the worlds to which we're related, one can only guess whether all of the worlds relative to ours have the same special initial conditions. Nevertheless some realist theories offer apparently prohibitive criteria, so they are able to make negative judgments. For instance, 'universalist' theories associated with Armstrong say that laws are relations between universals. Yet a constraint on initial conditions isn't in any natural way put in this form; hence it would seem the universalist theory would not consider this constraint lawlike.

Philosophical opinion is certainly divided. The problem is that a lawlike boundary condition lacks many of the features we ordinarily attribute to laws, e.g., multiple instances, governing temporal evolution, etc., yet different accounts of laws focus on different subsets of these features. When we turn to the issue at hand, what we find is the disagreement we expect.

3. The Problem of the Direction of Time II

A completely different problem going by the name 'problem of the direction of time' is the question of grounding various non-thermodynamic temporal asymmetries (to be described in detail below). In this problem, we take the thermodynamic arrow as given and use it to explain other temporally asymmetric features of the world, e.g., causation, knowledge. Boltzmann famously suggested that many of these asymmetries are given by the direction of entropy increase. And Reichenbach 1956 modified this to some of these temporal asymmetries being given by the direction of dominant entropy increase among all so-called "branch systems."

Sklar 1985 provides a useful discussion of this topic. He points out that conceiving of the reduction of these temporal asymmetries to that of the entropic arrow evades many of its obvious shortcomings if we conceive of it as a potential a posteriori scientific reduction of the kind now very familiar. The question is then whether it is so reduced (as for instance, the up-down plausibly reduces to the local gravitational gradient) or whether there is merely a correlation between the two (as, for example, there is between left-right and parity violations in high-energy particle physics).

The question is not easily answered partly due to vagueness about what is meant by both the concept to be reduced and the reducing concept. What temporal asymmetries are we concerned with, and exactly what kind of entropic relation do we intend?

The temporal asymmetries with which we are concerend are all the phenomena that we associate with the past and future directions being different. In addition to all of the temporal asymmetries from physics (thermodynamic arrow, electromagnetic arrow, Hubble expansion, etc.), there are a number of different asymmetries with which we are all familiar. The 'direction of time' might then be a broad umbrella covering the following:

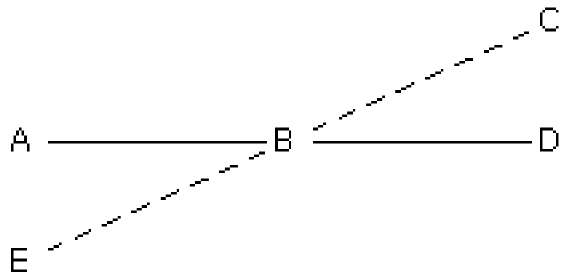
1. The psychological arrow. This controversial arrow is actually many different asymmetries. One, though much disputed, is that we seem to share a psychological sense of passage through time. Allegedly, we sense a moving ‘now’, the motion of the present as events are transformed from future to past. Another is that we have very different attitudes toward the past than toward the future. We dread future but not past headaches and prison sentences.

2. The mutability arrow. We feel the future is ‘open’ or indeterminate in a way the past is not. The past is closed, fixed for all eternity. Related to this, no doubt, is the feeling that our actions are essentially tied to the future and not the past. The future is mutable whereas the past is not.

3. The epistemological arrow. Although we believe that we know some facts about the future, the vast majority of propositions we claim to know about the past. I know that yesterday’s broken egg on the floor had a similar outline to Chile’s boundaries, but I have no idea what country tomorrow’s broken egg will look like. There are many more traces of events in the future than in the past. When I say something embarrassing, information representing that event is encoded on sound and light waves that form a continually increasing spherical shell in my future light-cone. I am potentially further embarrassed throughout my whole future lightcone. Yet in the backward lightcone stretching from the event there is little or no indication of the unfortunate event.

4. The explanation-causation-counterfactual arrow. This arrow is actually three, though it seems plausible that there are connections among them. Backwards causation may be physically possible, but if it is, it seems either to never happen or be exceedingly rare. Causes typically occur before their effects. Related to the causal asymmetry in some fashion or other is the asymmetry of explanation. Usually good explanations appeal to events in the past of the event to be explained, not to events in the future. It may be that this is just a prejudice that we ought to dispense with, but it is an intuition that we frequently have. Finally, and no doubt this is again related to the other two arrows as well as the mutability arrow, we -- at least naively -- believe the future depends counterfactually on the present in a way that we do not believe the past depends counterfactually on the present.

For example, consider a body moving uniformly from point *A* to point *C* in accord with Newton’s first law of motion.[9.] A force is impressed on the body at *B* and the body changes direction and proceeds uniformly towards *C*.



We will assume the body is a molecule travelling in a relative vacuum, and that the only trace left by the force is the altered path of the body. The solid lines in the diagram represent what we take to be the actual path of the body, the broken lines the alternative paths. Now consider two competing subjunctive conditionals:

> If no force had been impressed upon the body at *B*,
>
> (i) it would have moved uniformly in the right line *ABD*.
>
> (ii) it would have moved uniformly in the right line *EBC*.

The problem is to find an objective reason for our preference of (i). It seems that *AB* is co-tenable with the counterfactual antecedent. If the antecedent were true, it seems the body would have continued from *B* to *D*. But *BC* is also a leg of the actual path of the body, and to what do we appeal besides temporal asymmetry to reject *BC* as co-tenable with the counterfactual supposition? Perhaps after our intuitions have been tutored by physics we should say that either (i) or (ii) is correct. Or perhaps the asymmetry relies on thermodynamics, in which case the world described above is too bare to support our asymmetry.

Some authors -- particularly defenders of the tensed theory of time -- dismiss out of hand the idea of grounding the direction of time on the direction of material processes in time. But with so many asymmetric processes in the world, and with homo sapiens being just a part of this world, there are strong reasons to favor a connection between the two in many cases. But what is the connection?

Many authors have explicitly or implicitly proposed various ‘dependency charts’ that are supposed to explain which of the above arrows depend on which for their existence. Horwich 1987, for instance, argues for an explanatory relationship wherein the counterfactual arrow depends for its existence on the causal arrow, which depends on the arrow of explanation, which depends on the epistemological arrow, which in turn depends on the fork asymmetry that he associates with some chaotic conditions in the early universe. One can imagine other ways to plausibly arrange the dependency chart. Lewis 1979 thinks an alleged over-determination of traces grounds the asymmetry of counterfactuals and that this in turn grounds the rest. The chart one judges most appropriate will depend, to a large degree, upon one’s general philosophical stance on realism and Humeanism, etc., and one’s understanding of the above arrows.

Which chart is the correct one is not our concern here. Rather, returning to our main topic, the Boltzmann

entropic reduction of time-direction, we now have a somewhat clearer question: do any or all of the above temporal asymmetries depend for their existence upon the thermodynamic time-asymmetry? At the end of his 1979, for instance, Lewis hints that the asymmetry of traces is linked to the thermodynamic arrow, but he can offer no further explanation. Reichenbach 1956, Gruenbaum 1963, and Smart 1967 have developed entropic accounts of the knowledge asymmetry. Various people, for instance Dowe 1992, have tied the direction of causation to the entropy gradient. And some have also tied the psychological arrow to this gradient (for a discussion see Kroes 1985).

One can think of reasons for being quite pessimistic about any straightforward positive link between these temporal asymmetries and the entropy gradient. We really don't know how to bridge the gap between the thermodynamic arrow and the other arrows. And the gap is huge when you start thinking about the science of thermodynamics. Thermodynamics is a science with very precise and definite restrictions on the applicability of its concepts. A system has an entropy, for instance, only when it is thermally isolated and in equilibrium. Yet it is clear that our experience of the above temporal asymmetries carves up the world much differently than thermodynamics does. System A's doing f at time t might cause system B's doing g at time t* (where t* > t), yet A and B may not, and typically will not, have well-defined entropies.

The objections (see Earman 1974, Horwich 1987) to the entropic account of the knowledge asymmetry are worth recalling. The entropic account claimed that because we know there are many more entropy-increasing rather than entropy-decreasing systems in the world (or our part of it), we can infer when we see a low-entropy system that it was preceded and caused by an interaction with something outside the system. To take the canonical example, upon seeing a footprint in the sand, we can infer, due to its high order, that it was caused by something previously also of high (or higher) order, i.e, someone walking.

Though this brief sketch does not do justice to the entropic account, one can still see that it faces some very severe and basic challenges. First, do footprints on beaches have well-defined thermodynamic entropies? To describe the example we switched from low-entropy to high order, but the association between entropy and our ordinary concept of order is teneuous at best and usually completely misleading. To describe the range of systems about which we have knowledge, the account needs something broader than the thermodynamic entropy. But what? And why expect whatever it is to behave like entropy in some respects but not (in terms of its definability) in others? Second, the entropic account doesn't license the inference to a human being walking on the beach. All it tells you is that the grains of sand in the footprint interacted with its environment previously, which barely scratches the surface of our ability to tell detailed stories about what happened in the past. Third, even if we have a broader understanding of entropy, it still doesn't seem that this broader concept always works. Consider Earman's 1974 example of bomb destroying a city. From the destruction we may infer that a bomb went off; yet the bombed city does not have lower entropy than its surroundings or even any type of intuitively higher order than its surroundings.

Boltzmann's suggestion that the temporal asymmetries discussed above are explained by the direction of increasing entropy, though attractive at an abstract level, is hard to maintain when one looks at the details. Still, the more general idea, that these temporal asymmetries are due to the asymmetric behavior

of physical processes in our world (whatever their origin ,law or Past Hypothesis) as opposed to more metaphysical sources seems very plausible. Much work remains to be done on this problem.

Bibliography

- Albert, D. 2000. *Time and Chance*. Harvard University Press.
- Albert, D. 1992. *Quantum Mechanics and Experience*. Harvard University Press.
- Arntzenius, F. 1993. “The Classical failure to Account for Electromagnetic Arrows of Time” in G. Massey, T. Horowitz and A. Janis (eds) *Scientific Failure*.
- Blatt, J.M. 1959. *Progress in Theoretical Physics*, 22, 745.
- Bricmont, J. 1996: ‘Science of Chaos or Chaos in Science?’ in The Flight from Science and Reason, *New York Academy of Science Annals*, **775**, 131.
- Callender, C. 2000. “Is Time ‘Handed’ in a Quantum World?” *Proceedings of the Aristotelian Society*, June, 247-269.
- Callender, C. 1999. “Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy” *Journal of Philosophy*, XCVI, 348-373.
- Callender, C. 1998. “The View From No-when” *British Journal for the Philosophy of Science* 49, 135-159.
- Callender, C. 1997. "What is ‘The Problem of the Direction of Time’?” *Philosophy of Science (Supplement*), 63, v.2, 223-34.
- Christensen, F. M. 1993. *Space-like Time*. Toronto: University of Toronto Press.
- Cocke, J. 1967. ‘Statistical Time Symmetry and Two-Time Boundary Conditions in Physics and Cosmology’, *Physical Review* **160**, 1165-70.
- Davies, P. C. W. 1994. “Stirring Up Trouble” in Haliwell *et al* 1994, 119-30.
- Dowe, P. 1992. “Process Causality and Asymmetry”, *Erkenntnis* 37, 179-196.
- Earman, J. 1969. "The Anisotropy of Time" *Australasian Journal of Philosophy*, 67, 273-295.
- Earman, J. 1974. "An Attempt to Add a Little Direction to ‘The Problem of the Direction of Time’" *Philosophy of Science* **41**, 15-47.
- Earman, J. 1981. “Combining Statistical-Thermodynamics and Relativity Theory: Methodological and Foundations Problems” *Philosophy of Science Association* 1978, 2, pp. 157-185
- Fermi, E. 1936. *Thermodynamics*. NY: Dover.
- Feynman, R. 1965. *The Character of Physical Law* Cambridge, MA: MIT Press.
- Frisch, M. 2000. “(Dis-)solving the Puzzle of the Arrow of Radiation” *British Journal for the Philosophy of Science*, 51, pp. 381-410.
- Gold, T. 1962. “The Arrow of Time”, *American Journal of Physics*, **30**, 403-10.
- Goldstein, S., 2001, ‘Boltzmann’s Approach to Statistical Mechanics’, in *Chance in Physics: Foundations and Perspectives*, edited by Jean Bricmont, Detlef Dürr, Maria C. Galavotti, Giancarlo Ghirardi, Francesco Petruccione, and Nino Zanghi, (Lecture Notes in Physics 574), Springer-Verlag, 2001 [Available online]
- Grünbaum, A. 1973. *Philosophical Problems of Space and Time* NY: Knopf.
- Haliwell, J., Perez-Mercader, J., and W. Zurek,(eds) 1994. *Physical Origins of Time Asymmetry*. Cambridge: Cambridge University Press.

- Hurley, J. 1986. 'The Time-asymmetry Paradox' *American Journal of Physics* 54 (1), 25-28.
- Hawking, S. 1987. 'The Boundary Conditions of the Universe' in *Quantum Cosmology* ed. Fang and Ruffini. NJ: World Scientific, 162-174.
- Healey, R. 1981. "Statistical Theories, QM and the Directedness of Time" in *Reduction, Time and Reality* Cambridge: Cambridge University Press.
- Horwich, P. 1987. *Asymmetries in Time*. MIT Press.
- Joos, E. and Zeh, D. 1985. "The Emergence of Classical Properties through Interaction with the Environment", Z. Phys. B59, 223.
- Klein, M. 1973. "The Development of Boltzmann's Statistical Ideas" in *The Boltzmann Equation: Theory and Applications*, edited by E.G.D. Cohen and W. Thirring. Wien: Springer, 53-106.
- Kroes, P. 1985. *Time: Its Structure and Role in Physical Theories* Boston: D. Reidel.
- Laflamme, R. 1994. "The Arrow of Time and the No-boundary Proposal" in Haliwell *et al* , *The Physical Origins of Time Asymmetry*, pp. 358-68.
- Lebowitz, J. 1993. "Boltzmann's Entropy and Time's Arrow", *Physics Today* Sept.: 32-38.
- Lewis, D. 1979. "Counterfactual Dependence and Time's Arrow" *Nous* 13, 455-76.
- Lieb, E. H. and Yngvason, J. 2000. "A Fresh Look at Entropy and the Second Law of Thermodynamics" *Physics Today*, April, 32-37.
- Liu, C. 1994. "Is There a Relativistic Thermodynamics? A Case Study of the Meaning of Special Relativity" *Studies in the History and Philosophy of Modern Physics*, 25, 983-1004.
- Loewer, B. 1996. "Humean Supervenience and Laws of Nature" *Philosophical Topics* 24, 101-127.
- North, J. 2002. "What is the Problem about the Time-asymmetry of Thermodynamics? Reply to Price" *British Journal for the Philosophy of Science*, forthcoming.
- Partovi, M.H., 1989. "Irreversibility, Reduction, and Entropy Increase in Quantum Measurements", *Physics Letters* A **137**, 445-450.
- Penrose, O. 1970. *Foundations of Statistical Mechanics*. NY: Pergamon Press.
- Penrose, O. and Percival, I.C. 1962. "The Direction of Time" *Proceedings of the Physical Society*, **79**, 605-615.
- Penrose, R. 1989. *The Emperor's New Mind* Oxford: Oxford University Press.
- Pippard, A.B. 1964. *The Elements of Classical Thermodynamics*. Cambridge: Cambridge University Press.
- Popper, K. 1956. "The Arrow of Time", *Nature* 177, 538.
- Price, H. 2002. "Burbury's Last Case: The Mystery of the Entropic Arrow." Forthcoming in Craig Callender, ed., *Time, Reality and Experience*, Cambridge University Press.
- Price, H. 1996.: *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time*. New York: Oxford University Press. [Table of Contents and Chapter 1 available online]
- Price, H. 1995.: "Cosmology, Time's Arrow, and That Old Double Standard", in S. Savitt (ed.), 1995.
- Psillos, S. 1994. "A Philosophical Study of the Transition from the Caloric Theory of Heat to Thermodynamics". *Studies in the History and Philosophy of Science*, **25**, 159-90.
- Redhead, M. and Ridderbos, K. 1998. "The Spin-Echo Experiments and the Second Law of Thermodynamics" *Foundations of Physics*, 28, 1237-1270.
- Reichenbach, H. 1956. *The Direction of Time*. Berkeley: UCLA Press.

- Sanford, D. "The Direction of Causation and the Direction of Time," in *Midwest Studies in Philosophy* v.IX.
- Savitt, S. (ed.) 1995. *Time's Arrow Today*. Cambridge University Press.
- Savitt, S. 1996. "Survey Article: The Direction of Time" *British Journal for the Philosophy of Science*, 47, 347-370.
- Sklar, L. 1985. *Philosophy and Spacetime Physics* Los Angeles: UCLA Press.
- Sklar, L. 1993. *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge: Cambridge University Press.
- Schulman, L.S. *Time's Arrows and Quantum Measurement*. NY: Cambridge University Press.
- Smart, J. J. C. 1967. "Time" *Encyclopedia of Philosophy*. 8, NY: Macmillan.
- Tolman, R. 1934. *Relativity, Thermodynamics and Cosmology*. Oxford: Oxford University Press.
- Uffink, J. 2001. "Bluff Your Way Through the Second Law of Thermodynamics" *Studies in the History and Philosophy of Modern Physics*, forthcoming.
- Weingard, R. 1977. "Spacetime and the Direction of Time" *Nous* **11**, 119-131.
- Zeh, H.D. 1989: *The Physical Basis of the Direction of Time*. Berlin: Springer-Verlag. [4th edition available online]

Other Internet Resources

- PhilSci Archive (U. Pittsburgh); contains a section containing papers in the foundations of thermodynamics and statistical mechnaics. Many papers relevant to this entry are available.

Related Entries

physics: intertheory relations in | probability calculus: interpretations of | statistical physics: Boltzmann's work in | statistical physics: philosophy of statistical mechanics | time: the experience and perception of

Craig Callender
ccallender@ucsd.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 15, 2001
Content last modified: November 15, 2001

Stanford Encyclopedia of Philosophy
Notes to Thermodynamic Asymmetry in Time

Notes

1. For a nice discussion of this point within the debate about scientific realism, see Psillos 1994.

2. A third version, inequivalent to both of the above, is Caratheodory's law. It states that in the neighborhood of any equilibrium state of a system there are states that are inaccessible by an adiabatic process. This considerably more abstract version of the second law forbids a more general type of process than the other formulations, but it does so at the cost of making the law less intuitive, since its relationship to the familiar types of thermodynamic processes is often quite convoluted.

3. The so-called Third Law is essentially Nernst's Theorem, which states that the entropy of every system at absolute zero can always be taken equal to zero. The third law thus allows one to calculate the absolute value of the entropy; however, this is not necessary for classical thermodynamics to work successfully. In addition, there is the so-called Zero-th Law that expresses the transitivity of equilibrium: if two bodies A and B are separately in equilibrium with a third body C, then A and B are in equilibrium with one another. Interestingly, this law follows from Kelvin and Caratheodory's formulations but not from Clausius' (see Pippard 1964, 95).

4. Quantum field theory and the so-called CPT theorem suggest that neutral kaon decay is not TRI. One might also worry about whether quantum mechanics is really TRI (Albert 2001; Callender 2000).

5. This should be contrasted with the arrows of time discussed in semiclassical quantum gravity, for example, the idea that quantum scattering processes in systems with black holes violate the CPT theorem.

6. That is, a case can be made that both the 'super-weak' and 'milli-weak' fields postulated to account for CP-violations are quite analogous in method to the postulation of the time-ordering field. See Sachs 1986, 236, and references therein.

7. Blatt 1959, Reichbach 1956, Redhead and Ridderbos 1999, and to some extent, Horwich 1974 are a few works charmed by this idea.

8. The confusion surrounding interventionism is especially bad in quantum mechanics. Coupled with one of the primary sources of confusion, the measurement problem, and various mistaken views conflating information and entropy, one encounters many misguided ideas. The mysterious loss of information

essential to the observation process, for instance, is frequently said to be the source of quantum mechanical collapses and also the direction of time. Partovi (1989; 1990) has a project along these lines that exploits much such confusion. The point of his project is to show that interactions with the environment bring about wave function reduction that in turn causes entropy increase. Since environmental noise is often interpreted as environmental decoherence in QM, the decoherence approach to quantum measurement is also said to explain time's arrow (Joos and Zeh 1985). But since decoherence respects the unitarity and reversibility of the Schroedinger evolution, it is hard to see how it could make itself immune to a Loschmidt reversal, and so, this approach would still need to posit special initial conditions.

9. This example is from Sanford 1979; but see also Reichenbach 1956, pp. 43-47.

Craig Callender
ccallender@ucsd.edu

First published: November 15, 2001
Content last modified: November 15, 2001

Intertheory Relations in Physics

Many issues in the philosophy of science concern the nature of theories and certain relations that may obtain between them. Typically, one is interested in the degree to which a successor to a given theory "goes beyond" (both descriptively and explanatorily) the theory it succeeds. Most often these issues are framed in the context of *reductive* relations between theories. When does a theory T' reduce to a theory T? How is one to understand the nature of this reduction relation? Interestingly, there are two distinct, yet, related ways of understanding the reductive relationship between T and T'. Thomas Nickles noted this in a paper entitled "Two Concepts of Intertheoretic Reduction." On the one hand, there is the "philosopher's" sense of reduction on which the supplanted theory is said to reduce to the newer more encompassing theory. On the other hand, the "physicist's" sense of reduction puts things the other way. The newer, typically more refined theory is said to reduce to the older typically less encompassing theory in some sort of limit. These two senses of reduction will be discussed in turn.

---

1. Philosopher's Sense of Reduction

Most contemporary discussions of reductive relations between a pair of theories owe considerable debt to the work by Ernest Nagel. In *The Structure of Science*, Nagel asserts that "[r]eduction ... is the explanation of a theory or a set of experimental laws established in one area of inquiry, by a theory usually though not invariably formulated for some other domain." (Nagel, 1961, p. 338) The general schema here is as follows:

- T reduces T' just in case the laws of T' are derivable from those of T.

Showing how these derivations are possible for "paradigm" examples of intertheoretic reduction turns out

to be rather difficult.

Nagel distinguishes two types of reductions on the basis of whether or not the vocabulary of the reduced theory is a subset of the reducing theory. If it is---that is, if the reduced theory T' contains no descriptive terms not contained in the reducing theory T, and the terms of T' are understood to have approximately the same meanings that they have in T, then Nagel calls the reduction of T' by T "homogeneous." In this case, while the reduction may very well be enlightening in various respects, and is part of the "normal development of a science," most people believe that there is nothing terribly special or interesting from a philosophical point of view going on here. (Nagel, 1961, p. 339.)

Lawrence Sklar (1967, p. 110--111) points out that, from a historical perspective, this attitude is somewhat naive. The number of actual cases in the history of science where a genuine homogeneous reduction takes place are few and far between. Nagel, himself, took as a paradigm example of homogeneous reduction, the reduction of the Galilean laws of falling bodies to Newtonian mechanics. But, as Sklar points out, what actually can be derived from the Newtonian theory are *approximations* to the laws of the reduced Galilean theory. The approximations, of course, are strictly speaking *incompatible* with the actual laws and so, despite the fact that no concepts appear in the Galilean theory that do not also appear in the Newtonian theory, there is no deductive derivation of the *laws* of the one from the *laws* of the other. Hence, strictly speaking, there is no reduction on the deductive Nagelian model.

One way out of this problem for the proponent of Nagel-type reductions is to make a distinction between explaining a theory (or explaining the laws of a given theory) and explaining it away. (Sklar, 1967, pp. 112--113) Thus, we may still speak of reduction if the derivation of the approximations to the reduced theory's laws serves to account for why the reduced theory works as well as it does in its (perhaps more limited) domain of applicability. This is consonant with more sophisticated versions of Nagel-type reductions in which part of the very process of reduction involves revisions to the reduced theory. This process arises as a natural consequence of trying to deal with what Nagel calls "heterogeneous" reductions.

The task of characterizing reduction is more involved when the reduction is heterogenous---that is, when the reduced theory contains terms or concepts that do not appear in the reducing theory. Nagel takes, as a paradigm example of heterogeneous reduction, the (apparent) reduction of thermodynamics, or at least some parts of thermodynamics, to statistical mechanics.[1] For instance, thermodynamics contains the concept of temperature (among others) that is lacking in the reducing theory of statistical mechanics.

Nagel notes that "if the laws of the secondary science [the reduced theory] contain terms that do not occur in the theoretical assumptions of the primary discipline [the reducing theory] ... , the logical derivation of the former from the latter is *prima facie* impossible." (Nagel, 1961, p. 352) As a consequence, Nagel introduces two "necessary formal conditions" required for the reduction to take place:

1. *Connectability*. "Assumptions of some kind must be introduced which postulate suitable relations between whatever is signified by ‘A’ [the term to be reduced, that is, an element of the vocabulary of theory T'] and traits represented by theoretical terms already present in the primary [reducing] science."
2. *Derivability*. "With the help of these additional assumptions, all the laws of the secondary science, including those containing the term ‘A,’ must be logically derivable from the theoretical premises and their associated coordinating definitions in the primary discipline." (Nagel, 1961, pp. 353--354)

The connectability condition brings with it a number of interpretive problems. Exactly what is, or should be, the status of the "suitable relations," often called bridge "laws" or bridge hypotheses? Are they established by linguistic investigation alone? Are they factual discoveries? If the latter, what sort of necessity do they involve? Are they identity relations that are contingently necessary or will some sort of weaker relation, such as nomic coextensivity, suffice? Much of the philosophical literature on reduction addresses these questions about the status of the bridge laws.[2]

The consideration of certain examples lends plausibility to the idea, prevalent in the literature, that the bridge laws should be considered to express some kind of identity relation. For instance, Sklar notes that the reduction of the "theory" of physical optics to the theory of electromagnetic radiation proceeds by *identifying* one class of entities -- light waves -- with (part of) another class -- electromagnetic radiation. He says "... the place of correlatory laws [bridge laws] is taken by empirically established *identifications* of two classes of entities. Light waves are not correlated with electromagnetic waves, for they *are* electromagnetic waves." (Sklar, 1967, p. 120) In fact, if something like Nagelian reduction is going to work, it is generally accepted that the bridge laws should reflect the existence of some kind of synthetic identity.

Kenneth Schaffner calls the bridge laws "reduction functions." He too notes that they must be taken to reflect synthetic identities since, at least initially they require empirical support for their justification. "Genes were not discovered to be DNA via the analysis of *meaning*; important and difficult empirical research was required to make such an identification." (Schaffner, 1976. pp. 614--615)

Now one problem facing this sort of account was forcefully presented by Feyerabend in "Explanation, Reduction, and Empiricism." (Feyerabend, 1962) Consider the term "temperature" as it functions in classical thermodynamics. This term is defined in terms of Carnot cycles and is related to the strict, nonstatistical second law as it appears in that theory. The so-called reduction of classical thermodynamics to statistical mechanics, however, fails to identify or associate *nonstatistical* features in the reducing theory, statistical mechanics, with the nonstatistical concept of temperature as it appears in the reduced theory. How can one have a genuine reduction, if terms with their meanings fixed by the role they play in the reduced theory get identified with terms having entirely different meanings? Classical thermodynamics is not a statistical theory. The very possibility of finding a reduction function or bridge law that captures the concept of temperature and the strict, nonstatistical, role it plays in the thermodynamics seems impossible.

The plausibility of this argument, of course, depends on certain views about how meaning accrues to theoretical terms in a theory. However, just by looking at the historical development of thermodynamics one thing seems fairly clear. Most physicists, now, would accept the idea that our concept of temperature and our conception of other "exact" terms that appear in classical thermodynamics such as "entropy," need to be modified in light of the alleged reduction to statistical mechanics. Textbooks, in fact, typically speak of the theory of "statistical thermodynamics." The very process of "reduction" often leads to a corrected version of the reduced theory.

In fact, Schaffner and others have developed sophisticated Nagelian type schemas for reduction that explicitly try to capture these features of actual theory change. The idea is explicitly to include in the model, the "corrected reduced theory" such as statistical thermodynamics. Thus, Schaffner (1976, p. 618) holds that T reduces T' if and only if there is a corrected version of T', call it T'^* such that

1. The primitive terms of T'^* are associated via reduction functions (or bridge laws) with various terms of T.
2. T'^* is derivable from T when it is supplemented with the reduction functions specified in 1.
3. T'^* corrects T' in that it makes more accurate predictions than does T'.
4. T' is explained by T in that T' and T'^* are *strongly analogous* to one another, and T indicates why T' works as well as it does in its domain of validity.

Much work clearly is being done here by the intuitive conception of "strong analogy" between the reduced theory T' and the corrected reduced theory T'^*. In some cases, as suggested by Nickles and Wimsatt, the conception of strong analogy may find further refinement by appeal to what was referred to as the "physicists" sense of reduction.

2. Physicist's Sense of Reduction

Philosphical theories of reduction would have it that, say, quantum mechanics reduces classical mechanics through the derivation of the laws of classical physics from those of quantum physics. Most physicists would, on the other hand, speak of quantum mechanics reducing to classical mechanics in some kind of correspondence limit (e.g., the limit as Planck's constant ($h/2\pi$) goes to zero). Thus, the second type of intertheoretic reduction noted by Nickles fits the following schema:

Schema R: $\lim_{\varepsilon \to 0} T_f = T_c$

Here T_f is the typically newer, more fine theory, T_c is the typically older, coarser theory, and ε is a fundamental parameter appearing in T_f.

One must take the equality here with a small grain of salt. In those situations where **Schema R** can be said to hold, it is likely not the case that every equation or formula from T_f will yield a corresponding

equation of T_c .

Even given this caveat, the equality in **Schema R** can hold only if the limit is "regular." In such circumstances, it can be argued that it is appropriate to call the limiting relation a "reduction." If the limit in **Schema R** is singular, however, the schema fails and it is best to talk simply about intertheoretic relations.

One should understand the difference between regular and singular limiting relations as follows. If the solutions of the relevant formulae or equations of the theory T_f are such that for small values of ε they *smoothly* approach the solutions of the corresponding formulas in T_c, then **Schema R** will hold. For these cases we can say that the "limiting behavior" as $\varepsilon \to 0$ equals the "behavior in the limit" where $\varepsilon = 0$. On the other hand, if the behavior in the limit is of a *fundamentally different character* than the nearby solutions one obtains as $\varepsilon \to 0$, then the schema will fail.

A nice example illustrating this distinction is the following: Consider the quadratic equation $x^2 + x - 9\varepsilon = 0$. Think of ε as a small expansion or perturbation parameter. The equation has two roots for any value of ε as $\varepsilon \to 0$. In a well-defined sense, the solutions to this quadratic equation as $\varepsilon \to 0$ smoothly approach solutions to the "unperturbed" ($\varepsilon = 0$) equation $x^2 + x = 0$; namely, $x = 0, -1$. On the other hand, the equation $x^2\varepsilon + x - 9 = 0$ has two roots for any value of $\varepsilon > 0$ but has for the its "unperturbed" solution only one root; namely, $x = 9$. The equation suffers a reduction in order when $\varepsilon = 0$. Thus, the character of the behavior in the limit $\varepsilon = 0$ differs fundamentally from the character of its limiting behavior. Not all singular limits result from reductions in order of the equations. Nevertheless, these latter singular cases are much more prevalent than the former.

A paradigm case where a limiting reduction of the form **R** rather straightforwardly holds is that of classical Newtonian particle mechanics (NM) and the special theory of relativity (SR). In the limit where $(v/c)^2 \to 0$, SR reduces to NM. Nickles says "epitomizing [the intertheoretic reduction of SR to NM] is the reduction of the Einsteinian formula for momentum,

$$p = m_0 v / \sqrt{(1 - (v/c)^2)}$$

where m_0 is the rest mass, to the classical formula $p = m_0 v$ in the limit as $v \to 0$."[3] (Nickles, 1973, p. 182)

This is a regular limit---there are no singularities or "blowups" as the asymptotic limit is approached. As noted one way of thinking about this is that the exact solutions for small but nonzero values of $|\varepsilon|$ "smoothly [approach] the unperturbed or zeroth-order solution [ε set identically equal to zero] as $\varepsilon \to 0$." In the case where the limit is *singular* "the exact solution for $\varepsilon = 0$ is *fundamentally different in character* from the 'neighboring' solutions obtained in the limit $\varepsilon \to 0$." (Bender and Orszag, 1978, p. 324)

In the current context, one can express the regular nature of the limiting relation in the following way. The fundamental expression appearing in the Lorentz transformations of SR , can be expanded in a Taylor series as

$$1/\sqrt{(1-(v/c)^2)} = 1 - 1/2\,(v/c)^2 - 1/8\,(v/c)^4 - 1/16\,(v/c)^6 - \ldots$$

and so the limit is analytic. This means that (at least some) quantities or expressions of SR can be written as Newtonian or classical quantities plus an expansion of corrections in powers of $(v/c)^2$. So one may think of this relationship between SR and NM as a *regular* perturbation problem.

Examples like this have led some investigators to think of limiting relations as forming a kind of new rule of inference which would allow one to more closely connect the physicists' sense of reduction with that of the philosophers'. Fritz Rohrlich, for example, has argued that NM reduces (in the philosophers' sense) to SR because the *mathematical framework* of SR reduces (in the physicists' sense) to the *mathematical framework* of NM. The idea is that the mathematical framework of NM is "rigorously derived" from that of SR in a "derivation which involves limiting procedures." (Rohrlich, 1988, p. 303) Roughly speaking, for Rohrlich a "coarser" theory is reducible to a "finer" theory in the philosophers' sense of being rigorously deduced from the latter just in case the mathematical framework of the finer theory reduces in the physicists' sense to the mathematical framework of the coarser theory. In such cases, we will have a systematic explication of the idea of "strong analogy" to which Schaffner appeals in his model of philosophical reduction. The corrected theory T'^* in this context is the perturbed Newtonian theory as expressed in the Taylor expansion given above. The "strong analogy" between Newtonian theory T' and the corrected T'^* is expressed by the existence of the *regular* Taylor series expansion.

As noted the trouble with maintaining that this relationship between the philosophical and "physical" models of reduction holds generally is that far more often than not the limiting relations between the theories are *singular* and not regular. In such situations, **Schema R** fails to hold. Paradigm cases here include the relationships between classical mechanics and quantum mechanics, the ray theory of light and the wave theory, and thermodynamics and statistical mechanics of systems in critical states.

Intertheory Relations

It seems reasonable to expect something like philosophical reductions to be possible in those situations where **Schema R** holds. On the other hand, neither philosophical nor "physical" reduction seems possible when the limiting correspondence relation between the theories is singular. Perhaps in such cases it is best to speak simply of intertheoretic relations rather than reductions. It is here that much of philosophical and physical interest is to be found. This claim and the following discussion should not be taken to be anything like the received view among philosophers of science. Instead, it reflects the views of the author.

Nevertheless, here is a passage from a recent paper by Michael Berry which expresses a similar point of

view.

> Even within physical science, reduction between different levels of explanation is problematic--indeed, it is almost always so. Chemistry is supposed to have been reduced to quantum mechanics, yet people still argue over the basic question of how quantum mechanics can describe the shape of a molecule. The statistical mechanics of a fluid reduces to its thermodynamics in the limit of infinitely many particles, yet that limit breaks down near the critical point, where liquid and vapour merge, and where we never see a continuum no matter how distantly we observe the particles . . . . The geometrical (newtonian) optics of rays should be the limit of wave optics as the wavelength becomes negligibly small, yet . . . the reduction (mathematically similar to that of classical to quantum mechanics) is obstructed by singularities ... .
>
> My contention ... will be that many difficulties associated with reduction arise because they involve *singular limits*. These singularities have both negative and positive aspects: they obstruct the smooth reduction of more general theories to less general ones, but they also point to a great richness of borderland physics between theories. (Berry, forthcoming, p. 3)

When **Schema R** fails this is because the mathematics of the particular limit ($\varepsilon \to 0$) is singular. One can ask what, physically, is responsible for this mathematical singularity. In investigating the answer to this question one will often find that the mathematical blow-up reflects a physical impossibility. For instance, if **Schema R** held when T_f is the wave theory of light and T_c is the ray theory (geometrical optics), then one would expect to recover rays in the shortwave limit $\lambda \to 0$ of the wave theory. On the ray theory, rays are the carriers of energy. But in certain situations families of rays can focus on surfaces or lines called "caustics." These are not strange estoteric situations. In fact, rainbows are, to a first approximation, described by the focusing of sunlight on these surfaces following its refraction and reflection through raindrops. However, according to the ray theory, the intensity of the light on these focusing surfaces would be *infinite*. This is part of the physical reason for the mathematical singularities.

One is led to study the asymptotic domain in which the parameter ε in **Schema R** approaches 0. In the example above, this is the short wavelength limit. Michael Berry (1980,1990, 1994a, 1994b) has done much research on this and other asymptotic domains. He has found that in the asymptotic borderlands between such theories there emerge phenomena whose explanation requires in some sense appeal to a third intermediate theory. The emergent structures (the rainbow itself is one of them) are not fully explainable either in terms of the finer wave theory or in terms of the ray theory alone. Instead, aspects of both theories are required for a full understanding of these emergent phenomena.

This fact calls into question certain received views about the nature of intertheoretic relations. The wave theory, for example, is surely the fundamental theory. Nevertheless, these considerations seem to show that that theory is itself explanatorily deficient. There are phenomena within its scope whose explanations require reference to structures that exist only in the superseded, *false*, ray theory. A similar

situation arises in the asymptotic domain between quantum mechanics and classical mechanics where Planck's constant can be considered asymptotically small.

There is much here worthy of further philosophical study. See (Batterman 1993, 1995, and forthcoming) for the beginnings of some work in this direction.

Bibliography

- Batterman, R.W., 1991, "Chaos, quantization, and the correspondence principle", *Synthese*, 89: 189-227.
- -------, 1993, "Quantum chaos and semiclassical mechanics", in *PSA 1992*, volume 2, pages 50-65. Philosophy of Science Association.
- -------, 1995, "Theories between theories: Asymptotic limiting intertheoretic relations", *Synthese*, 103: 171-201.
- -------, forthcoming, *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press, New York.
- Bender, C.M., and Orszag, S.A., 1978, *Advanced Mathematical Methods for Scientists and Engineers*. McGraw-Hill, New York.
- Berry, M.V., 1990, "Beyond rainbows", *Current Science*, 59/(21-22): 1175-1191.
- -------, 1991, "Asymptotics, singularities and the reduction of theories", in Dag Prawitz, Brian Skyrms, and Dag Westerståhl, editors, *Logic, Methodolog and Philosophy of Science, IX: Proceedings of the Ninth International Congress of Logic, Methodology and Philosophy of Science, Uppsala, Sweden, August 7-14, 1991*, volume 134 of *Studies in Logic and Foundations of Mathematics*, pages 597-607, Amsterdam, 1994. Elsevier Science B. V.
- -------, 1994, "Singularities in waves and rays", in R. Balian, M. Kléman, and J. P. Poirier (eds), *Physics of Defects (Les Houches, Session XXXV, 1980)*, pages 453-543, Amsterdam, 1994. North-Holland.
- -------, forthcoming, "Chaos and the Semiclassical Limit of Quantum Mechanics (Is the Moon There When Somebody Looks?)", in *Proceedings of the CTNS-Vatican Conference on Quantum Physics and Quantum Field Theory*, in press (preprint in PDF format)
- Berry, M.V., and Upstill, C., 1980, "Catastrophe optics: Morphologies of caustics and their diffraction patterns", in E. Wolf (ed), *Progress in Optics*, volume XVIII, pages 257-346, Amsterdam, 1980. North-Holland.
- Feyerabend, P.K., 1962, "Explanation, reduction, and empiricism", in H. Feigl and G. Maxwell, (eds), *Minnesota Studies in the Philosophy of Science*, volume 3, pages 28-97. D. Reidel Publishing Company.
- Nagel, E., 1961, *The Structure of Science*. Routledge and Kegan Paul, London.
- Nickles, T., 1973, "Two concepts of intertheoretic reduction", *The Journal of Philosophy*, 70/7: 181-201.
- Rohrlich, F., 1988, "Pluralistic ontology and theory reduction in the physical sciences", *The British Journal for the Philosophy of Science*, 39: 295-312.
- Schaffner, K. 1976, "Reductionism in biology: Prospects and problems", in R.S. Cohen, *et al.*

(eds), *PSA 1974*, pages 613-632. D. Reidel Publishing Company.
- Sklar, L., 1967, "Types of inter-theoretic reduction", *The British Journal for the Philosophy of Science*, 18: 109-124.
- -------, 1993, *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge University Press, Cambridge.
- Wimsatt, W. C., 1976, "Reductive Explanation: A Functional Account", in A. C. Michalos, C. A. Hooker, G. Pearce, and R. S. Cohen, eds., PSA-1974 (Boston Studies in the Philosophy of Science, volume 30) Dordrecht: Reidel, pp. 671-710.

Other Internet Resources

- Berry, M.V., and Howls, C.J., 1993, "Infinity Interpreted
, *Physics World*(June 1993): 35-39

[Please contact the author with further suggestions.]

Related Entries

reduction and reductionism | scientific explanation

Robert W. Batterman
batterman.1@osu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 2, 2001
Content last modified: January 2, 2001

Stanford Encyclopedia of Philosophy
Notes to Intertheory Relations in Physics

Notes

1. That this is a paradigm example of reduction practically has the status of dogma in the philosophical literature on reduction. Unfortunately, this is entirely misguided. See Sklar's *Physics and Chance* for an extended discussion.

2. For a good introduction to the subtleties involved see Sklar, 1967, pp. 118--121.

3. It is best to think of this limit as $(v/c)^2 \to 0$ since $(v/c)^2$ is a dimensionless quantity and so this limit will not depend on the units used to measure the velocity.

Robert W. Batterman
batterman.1@osu.edu

First published: January 2, 2001
Content last modified: Januaray 2, 2001

Philosophy of Statistical Mechanics

Statistical mechanics was the first foundational physical theory in which probabilistic concepts and probabilistic explanation played a fundamental role. For the philosopher it provides a crucial test case in which to compare the philosophers' ideas about the meaning of probabilistic assertions and the role of probability in explanation with what actually goes on when probability enters a foundational physical theory. The account offered by statistical mechanics of the asymmetry in time of physical processes also plays an important role in the philosopher’s attempt to understand the alleged asymmetries of causation and of time itself.

---

1. Historical Sketch

From the seventeenth century onward it was realized that material systems could often be described by a small number of descriptive parameters that were related to one another in simple lawlike ways. These parameters referred to geometric, dynamical and thermal properties of matter. Typical of the laws was the ideal gas law that related product of pressure and volume of a gas to the temperature of the gas.

It was soon realized that a fundamental concept was that of equilibrium. Left to themselves systems would change the values of their parameters until they reached a state where no further changes were observed, the equilibrium state. Further, it became apparent that this spontaneous approach to equilibrium was a time-asymmetric process. Uneven temperatures, for example, changed until temperatures were

uniform. This same "uniformization" process held for densities.

Profound studies by S. Carnot of the ability to extract mechanical work out of engines that ran by virtue of the temperature difference between boiler and condenser led to the introduction by R. Clausius of one more important parameter describing a material system, its entropy. How was the existence of this simple set of parameters for describing matter and the lawlike regularities connecting them to be explained? What accounted for the approach to equilibrium and its time asymmetry? That the heat content of a body was a form of energy, convertible to and from mechanical work formed one fundamental principle. The inability of an isolated system to spontaneously move to a less orderly state, to lower its entropy, constituted another. But why were these laws true?

One approach, that of P. Duhem and E. Mach and the "energeticists," was to insist that these principles were autonomous phenomenological laws that needed no further grounding in some other physical principles. An alternative approach was to claim that the energy in a body stored as heat content was an energy of motion of some kind of hidden, microscopic constituents of the body, and to insist that the laws noted, the thermodynamic principles, needed to be accounted for out of the constitution of the macroscopic object out of its parts and the fundamental dynamical laws governing the motion of those parts. This is the kinetic theory of heat.

Early work on kinetic theory by W. Herepath and J. Waterston was virtually ignored, but the work of A. Krönig made kinetic theory a lively topic in physics. J. C. Maxwell made a major advance by deriving from some simple postulates a law for the distribution of velocities of the molecules of a gas when it was in equilibrium. Both Maxwell and L. Boltzmann went further, and in different, but related, ways derived an equation for the approach to equilibrium of a gas. The equilibrium distribution earlier found by Maxwell could then be shown to be a stationary solution of this equation.

This early work met with vigorous objections. H. Poincaré had proven a recurrence theorem for bounded dynamical systems that seemed to contradict the monotonic approach to equilibrium demanded by thermodynamics. Poincaré's theorem showed that any appropriately bounded system in which energy was conserved would of necessity, over an infinite time, return an infinite number of times to states arbitrarily close to the initial dynamical state in which the system was started. J. Loschmidt argued that the time irreversibility of thermodynamics was incompatible with the symmetry under time reversal of the classical dynamics assumed to govern the motion of the molecular constituents of the object.

Partly driven by the need to deal with these objections explicitly probabilistic notions began to be introduced into the theory by Maxwell and Boltzmann. Both realized that equilibrium values for quantities could be calculated by imposing a probability distribution over the microscopic dynamical states compatible with the constraints placed on the system, and identifying the observed macroscopic values with averages over quantities definable from the microscopic states using that probability distribution. But what was the physical justification for this procedure?

Both also argued that the evolution toward equilibrium demanded in the non-equilibrium theory could

also be understood probabilistically. Maxwell, introducing the notion of a "demon" who could manipulate the microscopic states of a system, argued that the law of entropic increase was only probabilistically valid. Boltzmann offered a probabilistic version of his equation describing the approach to equilibrium. Without considerable care, however, the Boltzmannian picture can still appear contrary to the objections from recurrence and reversibility interpreted in a probabilistic manner.

Late in his life Boltzmann responded to the objections to the probabilistic theory by offering a time-symmetric interpretation of the theory. Systems were probabilistically almost always close to equilibrium. But transient fluctuations to non-equilibrium states could be expected. Once in a non-equilibrium state it was highly likely that both after and before that state the system was closer to equilibrium. Why then did we live in a universe that was not close to equilibrium? Perhaps the universe was vast in space and time and we lived in a "small" non-equilibrium fluctuational part of it. We could only find ourselves in such an "improbable" part, for only in such a region could sentient beings exist. Why did we find entropy increasing toward the future and not toward the past? Here the answer was that just as the local direction of gravity defined what we meant by the downward direction of space, the local direction in time in which entropy was increasing fixed what we took to be the future direction of time.

In an important work (listed in the bibliography), P. and T. Ehrenfest also offered a reading of the Boltzmann equation of approach to equilibrium that avoided recurrence objections. Here the solution of the equation was taken to describe not "the overwhelmingly probable evolution" of a system, but, instead, the sequence of states that would be found overwhelmingly dominant at different times in a collection of systems all started in the same non-equilibrium condition. Even if each individual system approximately recurred to its initial conditions, this "concentration curve" could still show monotonic change toward equilibrium from an initial non-equilibrium condition.

Many of the philosophical issues in statistical mechanics center around the notion of probability as it appears in the theory. How are these probabilities to be understood? What justified choosing one probability distribution rather than another? How are the probabilities to be used in making predictions within the theory? How are they to be used to provide explanations of the observed phenomena? And how are the probability distributions themselves to receive an explanatory account? That is, what is the nature of the physical world that is responsible for the correct probabilities playing the successful role that they do play in the theory?

2. Philosophers on Probability and Statistical Explanation

Philosophers concerned with the interpretation of probability are usually dealing with the following problem: Probability is characterized by a number of formal rules, the additivity of probabilities for disjoint sets of possibilities being the most central of these. But what ought we to take the formal theory to be a theory of? Some interpretations are "objectivist," taking probabilities to be, possibly, frequencies of outcomes, or idealized limits of such frequencies or perhaps measures of "dispositions" or

"propensities" of outcomes in specified test situations.

Other interpretations are "subjectivist," taking probabilities to be measures of "degrees of belief," perhaps evidenced in behavior in situations of risk by choices of available lotteries over outcomes. Still another interpretation reads probabilities as measures of a kind of "partial logical entailment" among propositions.

Although subjectivist (or, rather, logical) interpretations of probability in statistical mechanics have been proffered (by E. Jaynes, for example), most interpreters of the theory opt for an objectivist interpretation of probability. This still leaves open, however, important questions about just what "objective" feature the posited probabilities of the theory are and how nature contrives to have such probabilities evinced in its behavior.

Philosophers dealing with statistical explanation have generally focussed on everyday uses of probability in explanation, or the use of probabilistic explanations in such disciplines as the social sciences. Sometimes it has been suggested that to probabilistically explain an outcome is to show it likely to have occurred given the background facts of the world. In other cases it is suggested that to explain an outcome probabilistically is to produce facts which raise the probability of that outcome over what it would have been those facts being ignored. Still others suggest that probabilistic explanation is showing an event to have been the causal outcome of some feature of the world characterized by a probabilistic causal disposition.

The explanatory patterns of non-equilibrium statistical mechanics place the evolution of the macroscopic features of matter in a pattern of probabilities over possible microscopic evolutions. Here the types of explanation offered do fit the traditional philosophical models. The main open questions concern the explanatory grounds behind the posited probabilities. In equilibrium theory, as we shall see, the statistical explanatory pattern has a rather different nature.

3. Equilibrium Theory

The standard method for calculating the properties of an energetically isolated system in equilibrium was initiated by Maxwell and Boltzmann and developed by J. Gibbs as the microcanonical ensemble. Here a probability distribution is imposed over the set of microscopic states compatible with the external constraints imposed on the system. Using this probability distribution, average values of specified functions of the microscopic conditions of the gas (phase averages) are calculated. These are identified with the macroscopic conditions. But a number of questions arise: Why this probability distribution? Why average values for macroscopic conditions? How do phase averages related to measured features of the macroscopic system?

Boltzmann thought of the proper average values to identify with macroscopic features as being averages over time of quantities calculable from microscopic states. He wished to identify the phase averages with such time averages. He realized that this could be done if a system started in any microscopic state

eventually went through all the possible microscopic states. That this was so became known as the ergodic hypothesis. But it is provably false on topological and measure theoretic grounds. A weaker claim, that a system started in any state would go arbitrarily close to each other microscopic state is also false, and even if true would not do the job needed.

The mathematical discipline of ergodic theory developed out of these early ideas. When can a phase average be identified with a time average over infinite time? G. Birkhoff (with earlier results by J. von Neumann) showed that this would be so for all but perhaps a set of measure zero of the trajectories (in the standard measure used to define the probability function) if the set of phase points was metrically indecomposable, that is if it could not be divided into more than one piece such that each piece had measure greater than zero and such that a system started in one piece always evolved to a system in that piece.

But did a realistic model of a system ever meet the condition of metric indecomposability? What is needed to derive metric indecomposability is sufficient instability of the trajectories so that the trajectories do not form groups of non-zero measure which fail to wander sufficiently over the entire phase region. The existence of a hidden constant of motion would violate metric indecomposability. After much arduous work, culminating in that of Ya. Sinai, it was shown that some "realistic" models of systems, such as the model of a gas as "hard spheres in a box," conformed to metric indecomposability. On the other hand another result of dynamical theory, the Kolmogorov-Arnold-Moser (KAM) theorem shows that more realistic models (say of molecules interacting by means of "soft" potentials) are likely not to obey ergodicity in a strict sense. In these cases more subtle reasoning (relying on the many degrees of freedom in a system composed of a vast number of constituents) is also needed.

If ergodicity holds what can be shown? It can be shown that for all but a set of measure zero of initial points, the time average of a phase quantity over infinite time will equal its phase average. It can be shown that for any measurable region the average time the system spends in that region will be proportional to the region's size (as measured by the probability measure used in the microcanonical ensemble). A solution to a further problem is also advanced. Boltzmann knew that the standard probability distribution was invariant under time evolution given the dynamics of the systems. But how could we know that it was the only such invariant measure? With ergodicity we can show that the standard probability distribution is the only one that is so invariant, at least if we confine ourselves to probability measures that assign probability zero to every set assigned zero by the standard measure.

We have, then, a kind of "transcendental deduction" of the standard probability assigned over microscopic states in the case of equilibrium. Equilibrium is a time-unchanging state. So we demand that the probability measure by which equilibrium quantities are to be calculated be stationary in time as well. If we assume that probability measures assigning non-zero probability to sets of states assigned zero by the usual measure can be ignored, then we can show that the standard probability is the only such time invariant probability under the dynamics that drives the individual systems from one microscopic state to another.

As a full "rationale" for standard equilibrium statistical mechanics, however, much remains questionable.

There is the problem that strict ergodicity is not true of realistic systems. There are many problems encountered if one tries to use the rationale as Boltzmann hoped to identify phase averages with measured quantities relying on the fact that macroscopic measurements take "long times" on a molecular scale. There are the problems introduced by the fact that all of the mathematically legitimate ergodic results are qualified by exceptions for "sets of measure zero." What is it physically that makes it legitimate to ignore a set of trajectories just because it has measure zero in the standard measure? After all, such neglect leads to catastrophically wrong predictions when there really are hidden, global constants of motion. In proving the standard measure uniquely invariant, why are we entitled to ignore probability measures that assign non-zero probabilities to sets of conditions assigned probability zero in the standard measure? After all, it was just the use of that standard measure that we were trying to justify in the first place.

In any case, equilibrium theory as an autonomous discipline is misleading. What we want, after all, is a treatment of equilibrium in the non-equilibrium context. We would like to understand how and why systems evolve from any initially fixed macroscopic state, taking equilibrium to be just the "end point" of such dynamic evolution. So it is to the general account of non-equilibrium we must turn if we want a fuller understanding of how this probabilistic theory is functioning in physics.

4. Non-Equilibrium Theory

Boltzmann provided an equation for the evolution of the distribution of the velocities of particles from a non-equilibrium initial state for dilute gases, the Boltzmann equation. A number of subsequent equations have been found for other types of systems, although generalizing to, say, dense gases has proven intractable. All of these equations are called kinetic equations.

How may they be justified and explained? In the discussions concerning the problem of irreversibility that ensued after Boltzmann's work, attention was focussed on a fundamental assumption he made: the hypothesis with regard to collision numbers. This time-asymmetrical assumption posited that the motions of the molecules in a gas were statistically uncorrelated prior to the molecules colliding. In deriving any of the other kinetic equations a similar such posit must be made. Some general methods for deriving such equations are the master equation approach and an approach that relies upon coarse-graining the phase space of points representing the micro-states of the system into finite cells and assuming fixed transition probabilities from cell to cell (Markov assumption). But such an assumption was not derived from the underlying dynamics of the system, and, for all they knew so far, might have been inconsistent with that dynamics.

A number of attempts have been made to do without such an assumption and to derive the approach to equilibrium out of the underlying dynamics of the system. Since that dynamics is invariant under time reversal and the kinetic equations are time asymmetric, time asymmetry must be put into the explanatory theory somewhere.

One approach to deriving the kinetic equations relies upon work which generalizes ergodic theory.

Relying upon the instability of trajectories, one tries to show that a region of phase points representing the possible micro-states for a system prepared in a non-equilibrium condition will, if the constraints are changed, eventually evolve into a set of phase points that is "coarsely" spread over the entire region of phase space allowed by the changed constraints. The old region cannot "finely" cover the new region by a fundamental theorem of dynamics (Liouville's theorem). But, in a manner first described by Gibbs, it can cover the region in a coarse-grained sense. To show that a collection of points will spread in such a way (in the infinite time limit at least) one tries to show the system possessed of an appropriate "randomization" property. In order of increasing strength such properties include weak-mixing, mixing, being a K system or being a Bernoulli system. Other, topological as opposed to measure-theoretic, approaches to this problem exist as well.

As usual, many caveats apply. Can the system really be shown to have such a randomizing feature (in the light of the KAM theorem, for example)? Are infinite time limit results relevant to our physical explanations? If the results are finite time, are they relativized in the sense of saying that they only hold for some coarse partitionings of the system rather than to those of experimental interest?

Most importantly, mixing and its ilk cannot be the whole story. All the results of this theory are time symmetric. To get time asymmetric results, and to get results that hold in finite times and which show evolution in the manner described by the kinetic equation over those finite times, requires an assumption as well about how the probability is to be distributed over the region of points allowed as representing the system at the initial moment.

What must that probability assumption look like and how may it be justified? These questions were asked, and partly explored, by N. Krylov. Attempts at rationalizing this initial probability assumption have ranged from Krylov's own suggestion that it is the result of a non-quantum "uncertainty" principle founded physically on the modes by which we prepare systems, to the suggestion that it is the result of an underlying stochastic nature of the world described as in the Ghirardi-Rimini-Weber approach to understanding measurement in quantum mechanics. The status and explanation of the initial probability assumption remains the central puzzle of non-equilibrium statistical mechanics.

There are other approaches to understanding the approach to equilibrium at variance with the approaches that rely on mixing phenomena. O. Lanford, for example, has shown that for an idealized infinitely dilute gas one can show, for very small time intervals, an overwhelmingly likely behavior of the gas according to the Boltzmann equation. Here the interpretation of that equation by the Ehrenfests, the interpretation suitable to the mixing approach, is being dropped in favor of the older idea of the equation describing the overwhelmingly probable evolution of a system. This derivation has the virtue of rigorously generating the Boltzmann equation, but at the cost of applying only to one severely idealized system and then only for a very short time (although the result may be true, if unproven, for longer time scales). Once again an initial probability distribution is still necessary for time asymmetry.

5. Irreversibility

The thermodynamic principles demand a world in which physical processes are asymmetric in time. Entropy of an isolated system may increase spontaneously into the future but not into the past. But the dynamical laws governing the motion of the micro-constituents are, at least on the standard views of those laws as being the usual laws of classical or quantum dynamics, time reversal invariant. Introducing probabilistic elements into the underlying theory still does not by itself explain where time asymmetry gets into the explanatory account. Even if, following Maxwell, we take the Second Law of thermodynamics to be merely probabilistic in its assertions, it remains time asymmetric.

Throughout the history of the discipline suggestions have often been made to the effect that some deep, underlying dynamical law itself introduces time asymmetry into the motion of the micro-constituents.

Other proposals take the entropic change of a system to be mediated by an actually uneliminable "interference" into the system of random causal influences from outside the system. It is impossible, for example, to genuinely screen the system from subtle gravitational influences from the outside. The issue of the role of external interference in the apparently spontaneous behavior of what is idealized as an isolated system has been much discussed. Here the existence of special systems (such as spin echo systems encountered in nuclear magnetic resonance) plays a role in the arguments. For these systems seem to display spontaneous approach to equilibrium when isolated, yet can have their apparent entropic behavior made to "go backward" with an appropriate impulse from out side the system. This seems to show entropic increase without the kind of interference from the outside that genuinely destroys the initial order implicit in the system. In any case, it is hard to see how outside interference would do the job of introducing time asymmetry unless such asymmetry is put in "by hand" in characterizing that interference.

It was Boltzmann who first proposed a kind of "cosmological" solution to the problem. As noted above he suggested a universe overall close to equilibrium with "small" sub-regions in fluctuations away from that state. In such a sub-region we would find a world far from equilibrium. Introducing the familiar time-symmetric probabilistic assumptions, it becomes likely that in such a region one finds states of lower entropy in one time direction and states of higher entropy in the other. Then finish the solution by introducing the other Boltzmann suggestion that what we mean by the future direction of time is fixed as that direction of time in which entropy is increasing.

Current cosmology sees quite a different universe than that posited by Boltzmann. As far as we can tell the universe as a whole is in a highly non-equilibrium state with parallel entropic increase into the future everywhere. But the structure of the cosmos as we know it allows for an alternative solution to the problem of the origin of time asymmetry in thermodynamics. The universe seems to be spatially expanding, with an origin some tens of billions of years ago in an initial singularity, the Big Bang. Expansion, however, by itself does not provide the time asymmetry needed for thermodynamics, for an expanding universe with static or decreasing entropy is allowed by physics. Indeed, in some cosmological models in which the universe contracts after expanding, it is usually, though not always, assumed that even in contraction entropy continues to increase.

The source of entropic asymmetry is sought, rather, in the physical state of the world at the Big Bang.

Matter "just after" the Big Bang is usually posited to be in a state of maximum entropy – to be in thermal equilibrium. But this does not take account of the structure of "space itself," or, if you wish, of the way in which the matter is distributed in space and subject to the universal gravitational attraction of all matter for all other matter. A world in which matter is distributed with uniformity is one of low entropy. A high entropy state is one in which we find a clustering of matter into dense regions with lots of empty space separating these regions. This deviation from the usual expectation – spatial uniformity as the state of highest entropy – is due to the fact that gravity, unlike the forces governing the interaction of molecules in a gas for example, is a purely attractive force.

One can then posit an initial "very low entropy" state for the Big Bang, with the spatial uniformity of matter providing an "entropic resevoir." As the universe expands, matter goes from a uniformly distributed state with temperature also uniform to one in which matter is highly clumped into hot stars in an environment of cold empty space. One then has the universe as we know it, with its thermally highly non-equilibrium condition. "Initial low entropy," then, will be a state in the past not (as far as we know) matched by any singularity of any kind, much less one of low entropy, in the future. If one conditionalizes on that initial low entropy state one then gets, using the time symmetric probabilities of statistical mechanics, a prediction of a universe whose entropy increased in time.

But it is not, of course, the entropy of the whole universe with which the Second Law is concerned, but, rather, that of "small" systems temporarily energetically isolated from their environments. One can argue, in a manner tracing back to H. Reichenbach, that the entropic increase of the universe as a whole will lead, again using the usual time symmetric probabilistic posits, to a high probability that a random "branch system" will show entropic increase parallel to that of the universe and parallel to that of other branch systems. Most of the arguments in the literature that this will be so are flawed, but the inference is reasonable nonetheless.

Positing initial low entropy for the Big Bang gives rise to its own set of "philosophical" questions: Given the standard probabilities in which high entropy is overwhelmingly probable, how could we explain the radically "unexpected" low entropy of the initial state? Indeed, can we apply probabilistic reasoning appropriate for systems in the universe as we know it to an initial state for the universe as a whole? The issues here are reminiscent of the old debates over the teleological argument for the existence of God.

6. The Reduction(?) of Thermodynamics to Statistical Mechanics

It comes as no surprise that the relationship of the older thermodynamic theory to the new statistical mechanics on which it is "grounded" is one of some complexity.

The older theory had no probabilistic qualifications to its laws. But as Maxwell was clearly aware, it could not then be "exactly" true if the new probabilistic theory correctly described the world. One can either keep the thermodynamic theory in its traditional form and carefully explicate the relationship its

principles bear to the newer probabilistic conclusions, or one can, as has been done in deeply interesting ways, generate a new "statistical thermodynamics" that imports into the older theory probabilistic structure.

Conceptually the relationship of older to newer theory is quite complex. Concepts of the older theory (volume, pressure, temperature, entropy) must be related to the concepts of the newer theory (molecular constitution, dynamical concepts governing the motion of the molecular constituents, probabilistic notions characterizing either the states of an individual system or distributions of states over an imagined ensemble of systems subject to some common constraints).

A single term of the thermodynamic theory such as ‘entropy’ will be associated with a wide variety of concepts defined in the newer account. There is, for example, Boltzmann entropy which is the property of a single system defined in terms of the spatial and momentum distribution of its molecules. On the other hand there are the Gibbs'entropies, definable out of the probability distribution over some Gibbsian ensemble of systems. Adding even more complications there is, for example, Gibbs' fine grained entropy which is defined by the ensemble probability alone and is very useful in characterizing equilibrium states and Gibbs' coarse grained entropy whose definition requires some partitioning of the phase space into finite cells as well as the original probability distribution and which is a useful concept in characterizing approach to equilibrium from the ensemble perspective. In addition to these notions which are measure theoretic in nature, there are topological notions which can play the role of a kind of entropy as well.

Nothing in this complexity stands in the way of claiming that statistical mechanics describes the world in a way that explains why thermodynamics works and works as well as it does. But the complexity of the inter-relationship between the theories should make the philosopher cautious in using this relationship as a well understood and simple paradigm of inter-theoretic reduction.

It is of some philosophical interest that the relationship of thermodynamics to statistical mechanics shows some similarity to aspects uncovered in functionalist theories of the mind-body relationship. Consider, for example, the fact that systems of very different physical constitutions (say a gas made up of molecules interacting by means of forces on the one hand and on the other hand radiation whose components are energetically coupled wave lengths of light) can share thermodynamic features. They can, for example, be at the same temperature. Physically this means that the two systems, if initially in equilibrium and then energetically coupled, will retain their original equilibrium conditions. The parallel with the claim that a functionally defined mental state (a belief, say) can be instantiated in a wide variety of physical devices is clear.

7. The Direction of Time

We have noted that it was Boltzmann who first suggested that our very concept of the future direction of time was fixed by the direction in time in which entropy was increasing in our part of the universe. Numerous authors have followed up this suggestion and the "entropic" theory of time asymmetry remains a much debated topic in the philosophy of time.

We must first ask what the theory is really claiming. In a sensible version of the theory there is no claim being made to the effect that we find out the time order of events by checking the entropy of systems and taking the later event as the one in which some system has its higher entropy. The claim is, rather, that it is the facts about the entropic asymmetry of systems in time that "ground" the phenomena that we usually think of as marking out the asymmetrical nature of time itself.

What are some features whose intuitive temporal asymmetry we think of as, perhaps, "constituting" the asymmetrical nature of time? There are asymmetries of knowledge: We have memories and records of the past, but not of the future. There are asymmetries of determination: We think of causation as going from past through present to future, and not of going the other way round. There are asymmetries of concern: We may regret the past, but we anxiously anticipate the future. There are alleged asymmetries of "determinateness" of reality: It is sometimes claimed that past and present have determinate reality, but that the future, being a realm of mere possibilities, has no such determinate being at all.

The entropic theory in its most plausible formulation is a claim to the effect that we can explain the origin of all of these intuitive asymmetries by referring to fact about the entropic asymmetry of the world.

This can be best understood by looking at the very analogy used by Boltzmann: the gravitational account of up and down. What do we mean by the downward direction at a spatial location? All of the phenomena by which we intuitively identify the downward direction (as the direction in which rocks fall, for example) receive an explanation in terms of the spatial direction of the local gravitational force. Even our immediate awareness of which direction is down is explainable in terms of the effect of gravity on the fluid in our semi-circular canals. It comes as no shock to us at all that "down" for Australia is in the opposite direction as "down" for Chicago. Nor are we dismayed to be told that in outer space, far from a large gravitating object such as the Earth, there is no such thing as the up-down distinction and no direction of space which is the downward direction.

Similarly the entropic theorist claims that it is the entropic features that explain the intuitive asymmetries noted above, that in regions of the universe in which the entropic asymmetry was counter-directed in time the past-future directions of time would be opposite, and that in a region of the universe without an entropic asymmetry neither direction of time would count as past or as future.

The great problem remains in trying to show that the entropic asymmetry is explanatorily adequate to account for all the other asymmetries in the way that the gravitational asymmetry can account for the distinction of up and down. Despite many interesting contributions to the literature on this, the problem remains unresolved.

Bibliography

A comprehensive treatment of the issues from a philosophical perspective is Sklar 1993. Of important historical interest is Reichenbach 1956. An accessible and up-to-date discussion of the fundamental

issues is Albert 2000. Additional philosophical discussion is in Guttman 1999. A spirited defense of the initial low entropy approach to time asymmetry is Price 1996. English translations of many original fundamental papers are in Brush 1965. Brush 1976 provides an historical treatment of the development of the theory. Two foundational works that are essential are Gibbs 1960 and Ehrenfest and Ehrenfest 1959.

- Albert, D., 2000, *Time and Chance,* Cambridge MA, Harvard University Press.
- Brush, S., ed., 1965, *Kinetic Theory*, Oxford, Pergamon Press.
- Brush, S., 1976, *The Kind of Motion That We Call Heat*, Amsterdam, North-Holland.
- Ehrenfest, P. and T., 1959, *The Conceptual Foundations of the Statistical Approach in Mechanics*, Ithaca NY, Cornell University Press.
- Gibbs, J., 1960, *Elementary Principles in Statistical Mechanics*, New York, Dover.
- Guttman, Y., 1999, *The Concept of Probability in Statistical Physics*, Cambridge, Cambridge University Press.
- Price, H., 1996, *Time's Arrow and the Archimedean Point*, Oxford, Oxford University Press.
- Reichenbach, H., 1956, *The Direction of Ti*me, Berkeley, University of California Press.
- Sklar, L., 1993, *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*, Cambridge, Cambridge University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

causation: backward | physics: intertheory relations in | probability calculus: interpretations of | statistical physics: Boltzmann’s work in

Lawrence Sklar
University of Michigan
lsklar@umich.edu

First published: April 12, 2001
Content last modified: April 12, 2001

Backward Causation

Sometimes also called retro-causation. A common feature of our world seems to be that in all cases of causation, the cause and the effect are placed in time so that the cause precedes its effect temporally. Our normal understanding of causation assumes this feature to such a degree that we intuitively have great difficulty imagining things differently. The notion of backward causation, however, stands for the idea that the temporal order of cause and effect is a mere contingent feature and that there may be cases where the cause is causally prior to its effect but where the temporal order of the cause and effect is reversed with respect to normal causation, i.e. there may be cases where the effect temporally, but not causally, precedes its cause.

The idea of backward causation should not be confused with that of time travel. These two notions are related to the extent that both agree that it is possible to causally affect the past. The difference, however, is that time travel involves a causal loop whereas backward causation does not. Causal loops for their part can only occur in a universe in which one has closed time-like curves. In contrast, backward causation may take place in a world where there are no such closed time-like curves. In other words, an ordinary system *S* taking part in time travel would preserve the temporal order of its proper time during its travel, it would keep the same time sense during its entire flight (a watch measuring *S*'s proper time would keep moving clockwise); but if the same system *S* were to become involved in a process of backward causation, the order of its proper time would have to reverse in the sense that the time sense of the system would become opposite of what it was before its back-in-time travel (the watch will start to move counter-clockwise). So neither backward causation nor time travel logically entails each other and time travel is distinct from back-in-time travel.

1. History

The philosophical debate about backward causation is relatively new. Only little consideration of the problem can be found in the philosophical literature before Michael Dummett and Anthony Flew initiated their discussion in the mid 1950s. The reason for this is twofold. No empirical phenomena seem to demand a notion of backward causation for our understanding of them. And for a long time it was thought that such a notion involved either a contradiction in terms or a conceptual impossibility. David Hume's definition of the cause as the one of two events that happens before the other thus rules out that the cause can happen after its effect. Moreover, according to Kant's idea of synthetic a priori truth the claim that the cause temporally precedes its effect was considered to state such a truth. In 1954 Michael Dummett and Anthony Flew had a discussion about whether an effect can precede its cause. Dummett defended the idea whereas Flew argued that it involved contradictions in terms.

Two years later, Max Black (1956) presented an argument against backward causation, which became known as the bilking argument, and later attempts to meet the argument seemed to generate all kinds of paradoxes. Imagine *B* to be earlier than *A*, and let *B* be the alleged effect of *A*. Thus we assume that *A* causes *B* even though *A* is later than *B*. The idea behind the bilking argument is that whenever *B* has occurred, it is possible, in principle, to intervene in the course of events and prohibit *A* from occurring. But if this is the case, *A* cannot be the cause of *B*; hence, we cannot have backward causation. Since then philosophers have debated the effectiveness of the bilking argument in particular and, in general, the validity and the soundness of the concept of backward causation.

In the 1960s and 1970s, physicists began to discuss the possibilities of particles travelling with a speed greater than light, the so-called tachyons, and as a consequence a similar debate about paradoxes involving backward causation arose among them. In case superluminal particles, like tachyons, exist and could be used to generate signals, it seemed possible to communicate with the past because tachyons going forward in time with respect to one set of reference frames would always be seen as travelling backwards in time from another set of reference frames.

2. Philosophy

A general notion of backward causation raises two sets of questions: those concerning conceptual problems and those that relate to empirical or physical matters. Among the first sets of questions that require a satisfactory answer are the following:

(i) *Can metaphysics provide a notion of time that allows that the effect precede its cause?* A proper notion of backward causation requires a static account of time in the sense that there is no objective becoming, no coming into being such that future events exist on the par with present and past events. It means that the future is real, the future does not merely consist of unrealised possibilities or even nothing at all. Ordinarily we may think of the past as a nothing that once was a something. But when asked what makes sentences about the past true or false, we would probably also say that it is the facts of the past that make present sentences about the past either true or false. The fact that I went to the cinema yesterday makes it true today when I say that I went to the cinema yesterday. This view is a realist one with respect to the past. If backward causation is to be conceptually possible it forces us to be realists with respect to the

future. The future must contain facts, events with certain properties, and these facts can make sentences about the future true or false. Such a realist account is provided by static and tenseless theories of time. A static theory holds that the participation of time into the past, the present and the future depends on the perspective we human beings put on the world. The attribution of pastness, presentness and futureness to events is determined by what we take to exist at times earlier than and times later than the time of our experience.

(ii) *Does backward causation mean that a future cause is changing something in the past?* Even most protagonists consider it an unwarranted consequence that the notion, if true, involves the idea that the future is able to change the past. Their answer has therefore usually been that if we have the power to bring something about in the past, what came about really already existed when the past was present. We have to make a distinction between changing the past so it becomes different from what it was and influencing the past so it becomes what it was. A coherent notion of backward causation only requires that the future is able to have an influence on what happens in the past.

(iii) *Can the cause be distinguished from its effect so that the distinction does not depend on a temporal ordering of the events?* The adherents have usually tried to give an account of causation in which the cause and the effect are not seen as regularities between types of events. Various alternative proposals refer to counterfactuals, probabilities, manipulation and intervention, common cause or causal forks. It is, apparently, only a Humean notion of causation that needs the temporal identification of the cause and the effect.

(iv) *Can the bilking argument be challenged in such a way that the mere possibility of intervention does not generate any serious paradoxes?* The force of the bilking argument can, it seems, be weakened in various ways. First, one may hold that it is not a problem for our notion of backward causation that we can in principle intervene in the course of the events. If we actually do so and prevent *A* after *B* has occurred, then of course a particular later *A* (which does not exist) cannot be the cause of a particular earlier *B* (which exists). But in all those cases where nobody actually intervenes, events of the same type as *A* may be the cause of events of the same type as *B*. This is not different from what can happen in some cases of forward causation. Assume that *P* causes *Q* in the relevant circumstances. We may still prevent a particular *P* from happening, but at the same time a particular *Q* may nevertheless occur because in the given circumstances it is caused by another event than *P*. Second, if a later event *A* really causes an earlier one *B*, then it would be impossible to intervene into the cause of the event after *B* has happened and therefore impossible to prevent *A* from happening. If someone tries, she will by all means fail. It may intuitively sound strange as long as we think of backward causation as consisting of something we can control directly by our everyday actions. But if backward causation is a notion that is applicable only to processes that human beings are unable to control in any foreseeable way the notion would not provoke our intuitions so much.

3. Physics

The notion of backward causation raises a very different set of questions that need to be answered before

a physically adequate notion has been developed.

(i) *What, if anything, would in physical terms characterize backward causation?* One has to remember that causality as such is an everyday notion that has no natural application in physics. How we could physically identify backward causal processes depends very much on which feature we take our ordinary notion of causation to apply to a physical process. In physics we may be tempted to associate it with different physical notions of processes. Four suggestions have been put forward: (a) the causal link can be identified with the transference of energy; (b) it can be identified with the conservation of physical quantities like charge, linear and angular momentum; (c) it can be identified with interaction of forces; or (d) it can be identified with the microscopic notion of interaction. It appears with respect to all four suggestions, however, that the involved descriptions are invariant under the time reversal operation.

The most fundamental laws of nature are time reversal invariant in the sense that our physical theories allow description of the fundamental reactions and processes in terms of the time reversed order. Such processes are said to be reversible in time. Maxwell's theory of electromagnetism, for instance, admits two kinds of mathematical solutions for the equations describing the radiation of energy in an electromagnetic field. One is called the *retarded* solution where radiation appears as outgoing concentric waves, the other is named the *advanced* solution according to which radiation appears as incoming concentric waves. Apparently the advanced solution describes the temporal inverse phenomena of the retarded solution so that these two solutions are usually regarded as the time reverse solution of the other. Nevertheless, retarded waves, like the increase of entropy in quasi-closed systems, appear to be *de facto* irreversible although they are described in terms of time invariant laws. Nature seems to prefer certain processes rather than their temporally inversed counterparts in spite of the fact that the laws of nature do not show such a preference. Light, radiation and ripples on a pond always spread outwards from their source rather than inwards just like entropy of a quasi-closed system is always moving from lower to higher states.

3.1 The Wheeler-Feynman Absorber Theory

Why do we not see any advanced waves in nature? Wheeler and Feynman (1945) came up with an answer. If we assume, they said, that radiation from an isolated accelerated charged particle is equally retarded and advanced, that is half retarded and half advanced to be exact, we can explain why it appears to be fully retarded in terms of the influence distant absorbers make on the source. The absorber consists of charged material that reacts with the source field by radiating with half retarded and half advanced waves. It is this half advanced field of the charged particles of the absorber which is added to the half retarded field of the source. The advanced waves of the absorber interfere constructively with the retarded waves of the source, whereas the same waves cancel out the advanced waves of the source in a destructive interference. Thus one of the consequences of Wheeler and Feynman Absorber Theory is the idea that emitters are intrinsically symmetric, another is that there is no intrinsic difference between so-called emitters and so-called absorbers. In other words, if this theory is true we have to conclude that radiation from a source is a time symmetric process but the presence of an absorber makes it asymmetric.

The Wheeler-Feynman theory takes for granted that outgoing, expanding waves are identical with

retarded radiation and incoming, contracting waves with advanced radiation. But is such identification without any problems? Not quite. An example with retarded and advanced emitters illustrates clearly why. Think of a stone being thrown directly into the middle of a circular pond. The ripples move outwards from the point where the stone hits the water (the source) in a coherent, organized wave front and eventually reach the edges (the absorber). Moreover, the source acts earlier than the absorber. What will the inverse process look like? It depends on how we understand such a process, whether or not we consider a case that includes a reversed source and a reversed absorber. (A) If they are included, the edges of the pond will now act as the source and the converging waves will eventually reach the middle of the pond. We may create something like this if we dropped a big ring horizontally into the pond. Inside the ring the waves would move inwards in an organized wave front towards the centre. In this case the source (the drop of the ring) would still act earlier than the absorber (the ripples meeting at the middle of the pond from all sides). (B) But if our understanding of the inverse process does not include an exchange of the source with the absorber and *vice versa*, then the ripples reach the edges of the pond (the absorber) earlier than the stone plunges into the water (the source). This is definitely not a state of affairs we could bring about. Furthermore, if we were to observe such a process, the ripples would seem to move inwards as contracting waves. The problem is that both kinds of inverse processes would seem to appear to us as organized incoming waves but one would be a case of retarded radiation and the other of advanced radiation.

This may not be the only problematic assumption of the Wheeler and Feynman theory. Huw Price (1996) has singled out other problems. Among them is the question of how we may experience the difference between retarded and advanced waves. When Wheeler and Feynman attributed to the source a field of half retarded and half advanced waves, they assumed that the field actually consists of retarded as well as an advanced component. Price objects, however, that there is no measurable difference between the two kinds of waves, and we cannot justify such a distinction by an appeal to the nature of the source because both emitters and absorbers can be associated with retarded as well as advanced waves. Instead he believes that these components are fictitious and that Wheeler and Feynman's formalism merely offer two different descriptions of the same wave. The problem of the asymmetry, as he sees it, has nothing to do with the fact that transmitters are associated with outgoing radiation rather than incoming radiation but that transmitters are centered on organized outgoing wave fronts whereas receivers are not centered on similar organized incoming wave fronts.

3.2 Tachyons

When the discussion of tachyons began to appear in physics in the 1960s, it was soon noticed that such particles according to some frames of reference were associated with negative energies going backwards in time. To understand how, consider the trajectory of the same tachyon in relation of three different reference frames, *S*, *S**, and *S*** in the Minkowski-space. Now assume that *A* is, in relation to *S*, the emission of a tachyon at t_1 and *B* is the absorption of the tachyon at t_2. According to an observer in *S*, *A* will be earlier than *B* and the tachyon will carry positive energy forward in time. Nevertheless it is always possible to select a reference frame *S** in relation to which an observer will see *A* happen simultaneously with *B* and yet another reference frame *S*** in relation to which an observer sees *A* happens at t_2**

whereas *B* happens at t_1**. According to the observer in *S***, *A* will take place later than *B* and the tachyon carries negative energy backwards in time (See Figure 1).

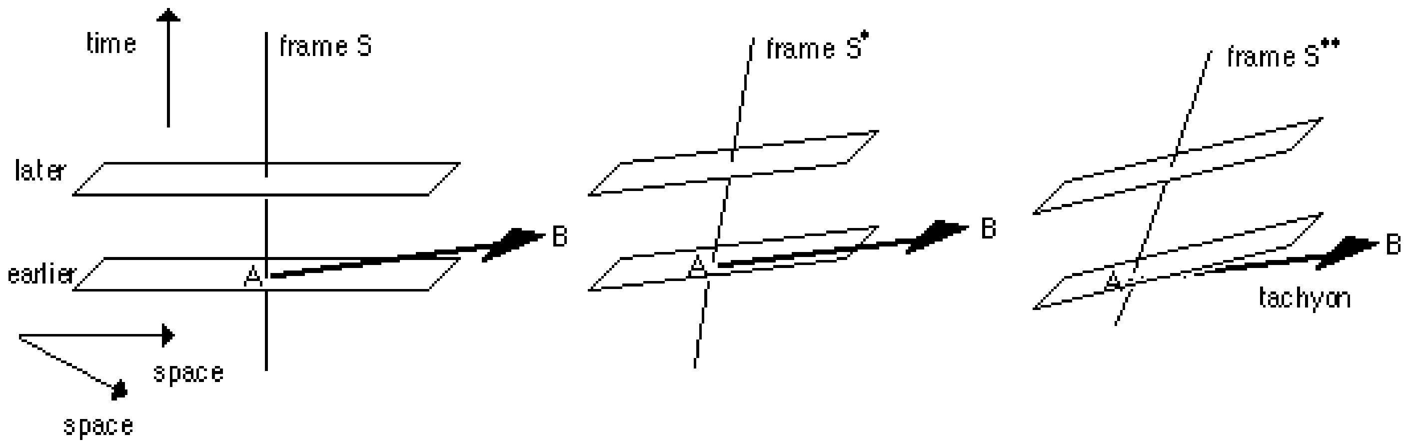


Figure 1

In Figure 1 the planes represent the hypersurfaces of simultaneity. In relation to frame *S* the taychon source is at rest, and a tachyon is emitted at event *A*, with a superluminal but finite velocity. The absorption of the tachyon, event *B*, will accordingly occur later than *A* in relation to the observer in *S*, and the arrow of trajectory is for that reason pointing into the future above the hypersurface passing through *A* and standing perpendicular to the world-line of the source. But neither with respect to the frame *S** nor *S*** is the tachyon source at rest and the hypersurfaces are therefore tilted in relation to the arrow of trajectory. An observer in *S** observes the tachyon to have infinite speed, and therefore the hypersurface is tilted so much that it coincides with the arrow. The observer in *S*** is moving so fast with respect to the tachyon source that the hypersurface becomes titled so much that the arrow points into the past below the hypersurface.

E. Recami (1978) tried to avoid the idea that tachyons could move backwards in time by introducing the so-called reinterpretation principle according to which all negative energy tachyons should be interpreted as if they have positive energy and move forward in time. This would mean that the causal order of tachyons should not be regarded objective since both *A* and *B* sometimes denoted the emission and sometimes the absorption depending on the frame of reference. There are, however, good reasons to believe that this suggestion does not solve the problems it was intended to (Faye, 1981/1989).

3.3 Quantum Mechanics

Other physical candidates for backward causation can be founded in the physics literature. Richard Feynman once came up with the idea that the electron could go backwards in time as a possible interpretation of the positron (Feynman, 1949). In fact he imagined the possibility that perhaps there were only one electron in the world zig-zaging back and forth in time. An electron moving backwards in time would carry negative energy whereas it would with respect to our ordinary time sense have positive charge and positive energy. But few consider this as a viable interpretation today (Earman, 1967, 1976).

More recently the Bell type experiments have by some been interpreted as if quantum events could be connected in such a way that the past light cone might be accessible under non-local interaction; not only in the sense of action at a distance but as backward causation. Costa de Beauregard (1977, 1979), for instance, has suggested that when a system of two photons in a singlet state is measured by two observers in two regions separated by a space-like distance, then it is precisely the act of observation that produces the past of the measuring process in the sense that it influences the source that emitted the two photons. de Beauregard's idea is that the element of reality being revealed in the formulation of the EPR paradox is real only because it was created by actually performed acts of observation that was propagated backwards in time with one of the two correlated quantum objects from the measuring device to the source of the photons. Several other philosophers and physicists have come forward with similar ideas. The basic assumption behind all of them is that in the micro-world we find only causal symmetry and this fact together with proper boundary conditions can be used to give an explanation of outcomes that seem otherwise paradoxical. Such quantum correlation experiments can, however, be interpreted in many other ways.

3.4 Two alternatives

These alleged examples of backward causation have one thing in common. They are all based on the idea that fundamental physical processes are by themselves symmetric in nature. Our ordinary notion of causation does not track any nomological feature of the world. What counts as the cause and the effect depends on the observer's projection of his or her temporal sense onto the world. So it is still an open question how a coherent notion of backward causation can fit into this general understanding of nature. The question we therefore have to answer is the following:

(ii) *How can we distinguish between forward causation and backward causation if all basic physical processes are time symmetric according to our description of nature?* Two very different reactions to this problem seem possible.

3.4.1 Boundary Conditions

One proposal is to say that if we came across reversed cases of *de facto* irreversible processes, such as running a film backwards in which the cream converged in a coffee cup, such cases should be interpreted as examples of backward causation (Price, 1996). The point is here to argue that it is the absence of the right initial or boundary conditions that makes backward causation so rare or nearly empirically impossible. This interpretation is based on three basic assumptions: (i) there is no objective asymmetry in the world, causal processes are intrinsically symmetric in nature, or causation is bidirectional, and therefore the fundamental processes of the micro-world are temporally symmetric; (ii) causal asymmetry is subjective in the sense that any attribution of an asymmetry between cause and effect depends on our use of counterfactuals and our own temporal orientation; (iii) backward causation, or advanced action, is nonetheless possible because sometimes the correlation of certain past events depends on the existence of causally symmetric processes and some future boundary conditions. For instance, advanced actions in electrodynamics require that the existence of transmitters in the future are centered on organized incoming wave fronts; and advanced actions in quantum mechanics require that their present states are in part

determined by the future conditions (measurements) they are to encounter. This feature is then taken to explain Bell's results in quantum mechanics.

A simple consideration seems to support this interpretation. Think of a particle travelling between two boxes. The normal observer and the counter-observer who has an inverse time sense will describe the exchange in conflicting terms. To the normal observer Box 1, say, will be considered as the emitter because it loses energy before anything in Box 2 happens. Therefore, Box 2 will be considered as the receiver since it gains energy at a later time. So in relation to the normal observer, the particle travels from Box 1 to Box 2. The counter-observer, however, sees the situation with opposite eyes. In relation to him, Box 2 loses energy and not until thereafter does Box 1 gain a similar amount of energy. Accordingly, in relation to the counter-observer, the particle moves from Box 2 to Box 1. In other words whether a box is considered to be an emitter or a receiver depends on the observer's time sense.

3.4.2 Nomic conditions

The other proposal denies that basic physical processes are time symmetric and argues, in contrast, that the causal asymmetry is objective and therefore that there exists an intrinsic difference between the cause and the effect of all physical processes. Hence backward causation should not be considered as a notion about boundary conditions but as a notion concerned with processes that nomically distinguish themselves from forward causal processes. Thus, if there are processes in the world that might be seen as a manifestation of backward causation, these are not to be depicted by a description that leaves them to be time reversed cases of ordinary forward causal processes (Faye, 1981/1989, 1997, 2002). This alternative interpretation rests on a basic claim and four assumptions.

The fundamental claim is that for any observer it is possible to identify experimentally the cause and the effect so that these remain the same even in relation to counter-observers, i.e. observers having the opposite time sense of ours. In support of this claim consider the following thought experiment. Two boxes, each having a shutter, are facing each other. Assume, ex hypothesis, that Box 1 is the particle source and Box 2 is the particle receiver. The question is how a normal observer and a counter-observer can come to agreement that particles move from Box 1 to Box 2. The answer can be found through a series of manipulations with the shutters, I would say. There are four possible combinations of the two shutters: open-open, close-close, open-close, close-open. Let us call any change of energy in Box 1, regardless of whether it emits or receives a particle, *A* and, similarly, any change of energy in Box 2 *B*. Whether *A* or *B* stand for a gain or a loss of energy can be determined by weighing the two boxes. (i) In case both boxes are closed, no particle will leave Box 1 and no particle is received by Box 2, thus no gain or loss of energy occurs, and both the normal observer and the counter-observer see a situation of not-*A*, not-*B*. (ii) In case both boxes are open a particle leaves Box 1 and is received by Box 2. Again this can be observed by measuring the change of energy in the two boxes. Thus the observers will see a situation of both *A* and *B*. (iii) In case Box 1 is closed and Box 2 is open, they will observe no change of energy in Box 1 (because it is closed) and, since no particle is leaving Box 1, no particle will reach Box 2 although its shutter is open. Hence the observers measure no energy change in this box. Thus they see not-*A* and not-*B*. (iv) Finally, if Box 1 is open and Box 2 is closed, a particle leaves Box 1, but none is received by Box 2. In other words, there is a loss or a gain of energy in Box 1, but no loss or gain of energy in Box 2.

So the observers see *A* and not-*B*. The upshot of this toy experiment is that the normal observer as well as the counter-observer experience two *A*s but only one *B*, and one not-*A* but two not-*B*s; therefore both will agree that the particles move from Box 1 to Box 2.

This means that what a normal observer identifies as a forward causal process will be regarded as a backward causal process in relation to the counter-observer in the sense that the very same event acting as a past cause for the normal observer will act as a future cause for the counter-observer. This indicates, too, that in relation to a normal observer forward causation and backward causation cannot be regarded as two different manifestations of nomologically reversible (but *de facto* irreversible) processes since both manifestations - the common process and the very improbable reversed process - would develop forward in time. If this claim is true, it implies that the description of physical processes should reflect such an intrinsic asymmetry in a way that the nomic description varies according to whether the process in question goes forward or backwards in time. Moreover, we must also be able to distinguish theoretically (and not only experimentally) between the normal observer's report and the counter-observer's report of the same process by a separate convention in respect to whether the process is forward moving or backward moving. What we want is a characterization of every physical process so that the invariance of cause and effect corresponds to *nomological* irreversibility.

In order to establish a nomic, intrinsic distinction between forward causal processes and backward causal processes one has to take departure in four assumptions. (i) Process *tokens* and process *types* are distinct in the sense that only process types are reversible, process tokens are not. (ii) A normal observer will describe causal processes propagating forward in time *in terms of positive mass and positive energy states pointing into her future* whereas she will describe the same tokens *in terms of negative mass and energy states pointing into her past*. This reflects two possible solutions of the four-momentum vector in the theory of relativity. (iii)Thus, one must distinguish between a *passive* time reversal operation and an *active* time reversal operation. The passive transformation is applied to the same process token by *describing* it in terms of opposite coordinates and opposite energy states. The active transformation, in contrast, brings about another token of the same process type in virtue of some physical translation or rotation of the system itself, both tokens having the same energy sign pointing in the same direction of time. (iv) The description in terms of positive mass and the possitive energy flow corresponds to the intrinsic order of the propagation.

Now, let us try to apply the nomic interpretation to the above consideration concerning the exchange of a particle between two boxes. In relation to the normal observer who describes the particle in terms of its *positive energy component*, it travels from Box 1 to Box 2 because Box 1 looses energy at an earlier time and Box 2 gains energy at a later time. The same situation is by the counter-observer described in terms of the particle's *negative energy component* as a situation where something happens in Box 2 before it happens in Box 1. In relation to the counter-observer, Box 2 would not, as the boundary interpretation suggests, loose energy. On the contrary, Box 2 would seem to gain energy, but the counter-observer would describe the particle as a series of negative energy states reaching into his future supposing the particle to be moving from Box 2 to Box 1 carrying negative energy. But, as we have just argued, the particle really moves from Box 1 to Box 2, from the counter-observer's future into his past carrying positive energy.

Consequently, the nomic interpretation holds that in relation to our normal time sense the causal direction of ordinary processes is identical with that of their reversed processes. In other words, take two tokens of a nomologically reversible process type, say *A* and *B*, and let *B* be the actively time reversed process of *A*, then this interpretation claims that *A* and *B* causally develop in the same direction of time. So, according to this view, neither incoming, contracting electromagnetic waves nor the decrease of entropy would count as examples of backward causation as long as such processes involve ordinary types of matter, i.e., matter that possesses positive mass and/or energy pointing, in relation to our normal time sense, towards the future. The notion of backward causation should instead be applied to matter of a different type, particles that appear to have, according to usual conventions, *negative mass and/or energy pointing, in relation to our normal time sense, towards the future but positive mass and/or energy pointing towards the past*. Such advanced matter, if it exists, should be distinguished from both ordinary retarded matter as well as tachyons by always being described with respect to our time sense in terms of negative mass and energy stretching forward in time. A consequence is that a world in which advanced matter exists together with retarded matter, and where advanced matter is able to interact directly with the same amount of retarded matter, both would, in case they actually did interact, annihilate without leaving any trace of energy.

How and whether the notion of backward causation has a role to play in physics has yet to be seen. But as long as no common agreement exists among philosophers and physicists about what in the physical description of the world corresponds to our everyday notion of causation, it would still be a matter of theoretical dispute what counts as empirical examples of backward causation.

Bibliography

- Arons, M.E. & E.C.G. Sudarshan (1968), "Lorentz Invariance, Local Field Theory and Faster-than-Light Particles", *Physical Review*, 173, 5, pp. 1622-1628.
- Bilaniuk, O.M.P, V.K. Despande and E.C.G. Sudarshan (1962), "'Meta' Relativity", *American Journal of Physics*, 30, 2, pp. 718-723.
- Bilaniuk, O.M.P. et al. (1969), "More About Tachyons", *Physics Today*, 22, 12, pp. 47-51.
- Bilaniuk, O.M.P., and E.C.G. Sudarshan (1969), *Physics Today*, 22, 5, pp. 43-51.
- Black, M. (1956), "Why Cannot an Effect Precede its Cause", *Analysis*, 16, pp. 49-58.
- Csonka, P.L. (1970), "Causality and Faster than Light Particles", *Nuclear Physics*, B21, pp. 436-444.
- de Beauregard, C. (1977), "Time Symmetry and the Einstein Paradox," *Il Nuovo Cimento*, 42B, pp. 41-64.
- de Beauregard, C. (1979), "Time symmetry and the Einstein Paradox- II," *Il Nuovo Cimento*, 51B, pp. 267-279.
- Dorato, M. (1995), *Time and Reality: Space-Time Physics and the Objectivity of Temporal Becoming*, Bologna: CLUEB.
- Dummett, M. (1954), "Can an Effect Precede its Cause", *Proceedings of the Aristotelian Society*, Supp 28, pp. 27-44.
- Dummett, M. (1964), "Bringing about the Past", *Philosophical Review*, 73, pp. 338-359.

- Earman, J. (1967), "On going Backward in Time", *Philosophy of Science*, 34, pp. 211-222.
- Earman, J. (1976), "Causation: A Matter of Life and Death", *Journal of Philosophy*, 73, pp. 5-25.
- Faye, J. (1981/1989), *The reality of the future*, Odense: Odense University Press.
- Faye, J. (1997), "Causation, Reversibility, and the Direction of Time", in J. Faye, U. Scheffler & M. Urchs (eds.), *Perspectives on Time*. In *Boston Studies in the Philosophy of Science*, 189, pp. 237-266. Dordrecht: Kluwer Academic Publisher.
- Faye, J. (2002), "When Time Gets Off Track", in C. Callender (ed.) *Time, Reality, and Experience*, Cambridge: Cambridge University Press.
- Feinberg, G. (1967) "Possibility of Faster-Than-Light Particles", *Physical Review*, 159, 5, pp.1089-1105.
- Feynman, P.R. (1949), "The Theory of Positrons", *Physical Review*, 76, pp. 749-459.
- Flew, A. (1954), "Can an Effect Precede its Cause", *Proceedings of the Aristotelian Society*, Supp 28, pp. 45-62.
- Flew, A. (1956), "Effects before their Causes - Addenda and Corrigenda", *Analysis*, 16, pp. 104-10.
- Flew, A. (1956-7), "Causal Disorder Again", *Analysis*, 17, pp. 81-86.
- Gale, R. (1965), "Why a Cause Cannot be Later than its Effects", *Review of Metaphysics*, 19, pp. 209-234.
- Gorovitz, G. (1964), "Leaving the Past Alone", *Philosophical Review*, 73, pp.360-371.
- Horwich, P. (1987), *Asymmetries in Time*, Cambridge Mass.: MIT Press.
- Mellor, D.H. (1981), *Real Time*, Cambridge: Cambridge University Press.
- Price, H. (1996), *Time's Arrow and Archimedes' Point*, Oxford: Oxford University Press.
- Recami, E. (1978), "How to Recover Causality in Special Relativity for Tachyons", *Foundations of Physics*, 8, pp.329-340.
- Schlesinger, G. (1980), *Aspects of Time*, Indiana: Hackett.
- Tanaka, S. (1960), "Theory of Matter with Super Light Velocity", *Progress of Theoretical Physics*, 24, 1, pp.171-200.
- Wheeler, J.A., and Feynman, R.P. (1945), "Interaction with the Absorber as the Mechanism of Radiation," *Reviews of Modern Physics*, 17, pp. 157-181.

Other Internet Resources

[Please contact author with suggestions.]

Related Entries

causation: causal processes | space and time: being and becoming in modern physics | time | time travel: and modern physics

Acknowledgements

Thanks to John Norton for his editorial suggestions and for his drawing of Figure 1.

Jan Faye
faye@hum.ku.dk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 27, 2001
Content last modified: November 21, 2001

Being and Becoming in Modern Physics

Does time flow or lapse or pass? Are the future or the past as real as the present? These metaphysical questions have been debated for more than two millennia, with no resolution in sight. Modern physics provides us, however, with tools that enable us to sharpen these old questions and generate new arguments. Does the special theory of relativity, for example, show that there is no passage or that the future is as real as the present? The focus of this entry will be these new questions and arguments.

---

1. Introduction

Around 500 B. C. Heraclitus wrote the following:

> Everything flows and nothing abides; everything gives way and nothing stays fixed.
>
> You cannot step twice into the same river, for other waters and yet others, go flowing on.
>
> Time is a child, moving counters in a game; the royal power is a child's.[1]

Transience is basic, and the present is primary. Those things which exist now do not abide. They slip into

the past and non-existence, devoured by time, as all experience attests.

A generation or so later we have a classic statement of the opposing view by Parmenides:

> There remains, then, but one word by which to express the [true] road: Is. And on this road there are many signs that What Is has no beginning and never will be destroyed: it is whole, still, and without end. It neither was nor will be, it simply is—now, altogether, one, continuous…

Permanence is basic. No things come to be or, slipping into the past, cease to be. Past, present, and future are distinctions not marked in the static Is. Time and becoming are at best secondary, at worst illusory, as our understanding of the world confirms.

Turn now to modern times and to a paragraph in Rudolf Carnap's intellectual autobiography (Carnap 1963, pp. 37-38):

> Once Einstein said that the problem of the Now worried him seriously. He explained that the experience of the Now means something special for man, something essentially different from the past and the future, but that this important difference does not and cannot occur within physics. That this experience cannot be grasped by science seemed to him a matter of painful but inevitable resignation. I remarked that all that occurs objectively can be described in science; on the one hand the temporal sequence of events is described in physics; and, on the other hand, the peculiarities of man's experiences with respect to time, including his different attitude towards past, present, and future, can be described and (in principle) explained in psychology. But Einstein thought that these scientific descriptions cannot possibly satisfy our human needs; that there is something essential about the Now which is just outside the realm of science. We both agreed that this was not a question of a defect for which science could be blamed, as Bergson thought. I did not wish to press the point, because I wanted primarily to understand his personal attitude to the problem rather than to clarify the theoretical situation. But I definitely had the impression that Einstein's thinking on this point involved a lack of distinction between experience and knowledge. Since science in principle can say all that can be said, there is no unanswerable question left. But though there is no theoretical question left, there is still the common human emotional experience, which is sometimes disturbing for special psychological reasons.

This difference as expressed here between Einstein and Carnap (that is, between the Heraclitean and Parmenidean attitude towards time and change) is the subject of this article, which will use modern physics--especially modern spacetime theory--as a set of lenses through which it is hoped the riddles of time will come into sharper focus. There are many ways, however, to approach these questions. Early in the twentieth century, Anglo-American philosophy turned to consideration of language as way to clarify philosophical disputes. Philosophers of time debated the relative primacy of tensed language (concerning

the notions of present, past, and future) or tenseless language (concerning the relations of simultaneity and temporal precedence). Our considerations of physics will generally, though not completely, skirt linguistic disputes. The reader interested in following these debates can find a sophisticated review and discussion in Tooley (1999).

Other philosophers have been influenced by analogies between time and modality. The reader interested in this way of thinking about time should consult the article Temporal Logic. The present article will focus on time in physics and the relations between time and space. Other philosophical approaches focus on the primacy of experience in our understanding of time. The reader interested in these approaches may wish to consult The Experience and Perception of Time.

2. Newtonian Spacetime

Modern physical theories are often formulated in a language that permits one to express a variety of different views with respect to time and its relation to space. One can, for example, formulate the basic ideas of classical (that is, Newtonian) physics, the special theory of relativity, and the general theory of relativity in this language. For a brief introduction to the spacetime view, see Section 1 ("Modern Spacetime Theories: A Beginner's Guide") of John Norton's entry on The Hole Argument in this Encyclopedia. For more detail with minimal technical demands the reader should see the first four chapters of Geroch (1978) or (more demanding) chapter 2 of Friedman (1983).

For our purposes, the defining feature of a manifold that is a Newtonian spacetime is that the temporal interval between any two points or events in the spacetime, *p* and *q*, is a well-defined quantity. This quantity is well-defined in that it does not depend upon point of view, reference frame, coordinate system or "observer". This quantity, then, is *absolute* in the sense of being frame- or observer-independent. (In the special theory of relativity the temporal interval between two distinct spacetime points fails to be absolute in this sense.)

If the temporal interval between two events is 0, then we say that the two events are *simultaneous*. This relation of (absolute) simultaneity is an equivalence relation (That is, it is reflexive, symmetric, and transitive.) that slices (partitions or foliates) the spacetime or manifold into mutually exclusive and exhaustive planes of simultaneity. These planes of simultaneity can then be completely ordered by the relation 'is earlier than' or its converse 'is later than'.

2.1 Presentism, Possibilism, Eternalism

The geometrical structure of Newtonian spacetime reflects the way we ordinarily think about time and is the proper backdrop for introducing the three major rival metaphysical views of time, as illustrated below:

Figure 1

Figure 1. Three Metaphysics of Time

The first view, represented on the left, is the ontologically austere view called *presentism*, the view that only the present exists. The past has been but is no longer, while the future will come to be but is not yet. Note that it is the convention of these diagrams that one spatial dimension is suppressed. The present is actually a three dimensional global slice of the spacetime. Moreover, the illustration necessarily represents the spatial extent of the present as finite and may suggest that time also has a beginning and/or end. These views are, however, merely artifacts of the representation and not integral to presentism, possibilism, or eternalism. The diagram illustrating presentism also has four arrows pointing up (conventionally, towards the future) attached to the plane representing the present. These arrows are meant to indicate something that is integral to presentism, the idea that the present (and hence the existent) constantly shifts or changes. These arrows represent, then, the dynamic aspect of time called *temporal becoming* or *passage.* The deepest problem in the metaphysics of time is how to understand passage or becoming and its relation to existence.

In contrast to the radical Heraclitean view of presentism, the Parmenidean eternalist picture on the far right lacks these arrows and indicates that there is no more special about the temporal present (the *now*) than the spatial present (the *here*). Future and past events at a place, on this view, are no more or less real than distant events at a time. The *now* like the *here* is a function of one's perspective, one's position in the spacetime, and these positions are indicated by the line in the spacetime representing the history of spacetime locations of a particular object or person. Such a line is often called a *world line.*

The middle view, possibilism, is indeed an intermediate view. It is a passage view, but it is less ontologically sparse than presentism. While on this view the future is still merely possible rather than actual (hence its name), the past has become and is fully actual. If one thinks of the future as a branching structure of alternative possibilities (as the result, for instance, of free human choices or indeterministic quantum measurements), then one can think of the past and present as the trunk of that tree, growing as possibilities become actual in the present.

Possibilism seems to capture much of the way we think about time and being. While the sparse symmetry of presentism is attractive, there are many deep asymmetries concerning past and future that it fails to reflect. I can easily ascertain, for instance, yesterday's closing number for the Dow Jones Industrial Average, but by no efforts, however great, can I now ascertain tomorrow's close. And it seems as if my actions (or certain sorts of quantum measurements) can actualize some future possibilities as opposed to others, whereas past actions (or the results of past quantum measurements) seem no longer to admit of alternatives. Even if one allows for the possibility of retrocausation, for the possibility of an effect preceding its cause in time, it is generally held that a present cause can not change or alter the past. It would merely make the past what it was. (See the entry Backwards Causation for further consideration of this topic.)

Eternalism too, *prima facie*, would seem to have trouble accounting for the asymmetries built into

possibilism, in addition to its implausible denial of passage. But the first topic to which we shall turn is an argument, prominent in twentieth century philosophy of time, that passage or becoming is a self-contradictory idea. If the argument is correct, then neither presentism nor possibilism can be correct metaphysical views of time and being.

2.2 McTaggart's Argument

At the beginning of the 20th century, J. M. E. McTaggart (1908) presented an argument which purported to prove that time is unreal. According to McTaggart (1927, pp. 9-10):

> Positions in time, as time appears to us prima facie, are distinguished in two ways. Each position is Earlier than some and Later than some of the other positions… . In the second place, each position is either Past, Present, or Future. The distinctions of the former class are permanent, while those of the latter are not. If *M* is ever earlier than *N*, it is always earlier, But an event, which is now present, was future, and will be past.

The first structure of "positions in time," McTaggart called *the B-series.* I will assume that McTaggart intended the B-series to coincide with the classical spacetime structure described above. McTaggart noted that there was something static or "permanent" about the B-series. If, for example, event e_1 is earlier than event e_2 at some time or other, then it is earlier than e_2 at all times.

The dynamic element of time must be represented, in McTaggart's view, by the series of properties of pastness, presentness, and futurity, which (in contrast to the static B-series) are constantly changing. A given event becomes less future, becomes present, and then becomes increasingly past. This latter ever-shifting series McTaggart called *the A-series.*

While there are many obscurities in McTaggart's writing, it seems clear that the argument intended to prove that time is unreal runs along the following lines:

(1) there can be no time unless it has a dynamic element (that is, on his view, unless there is an A-series),

(2) there can be no A-series, because the supposition that there is an A-series leads to contradiction.

The contradiction alleged by McTaggart is that:

(A_1) every event must have many, if not all, the A-properties (or A-determinations, as they are sometimes called) whereas,

(A_2) since the A-properties are mutually exclusive, no event can have more than one of them.

Near the end of career in which he spent much time and effort in thinking about McTaggart's argument, C. D. Broad (1959, p. 765) wrote:

> I felt from the first, and still feel, that the difficulty which arises is (a) embarrassing enough prima facie to demand the serious attention of anyone who philosophises about time, and (b) almost certainly due to some purely linguistic source (common, and perhaps peculiar, to the Indo-European verb-system), which it ought to be possible to indicate and make harmless.

Broad's claim (a) was vindicated by the fact that McTaggart's argument has received serious attention from most subsequent philosophers who pondered the metaphysics of time. Much of this debate concerns the relative relations of the two series. Is the A-series fundamental and the B-series derived from it, or vice versa, or does, perhaps, one series supervene upon the other? In the formal mode, the questions become whether the B-series may somehow be reduced to, may be defined in terms of, the A-series (or vice versa). These debates concern mainly language rather than physics and will not be considered here.[2]

What emerges from the McTaggart literature that is relevant to this discussion is, first of all, a tendency to identify the existence of passage or temporal becoming with the existence of the A-series (that is, to think of becoming as events changing their properties of pastness, presentness or nowness, and futurity) and hence the tendency for debates about the existence of passage to focus on the merits or incoherence of the A-series rather than examining alternative accounts of becoming. (But Cf. Fitzgerald, 1985)

There is a tendency amongst those philosophers who take modern physics seriously to be sceptical of entities like constantly shifting temporal properties of events, since such properties seem to play no role in modern physical theory. One view, defended by Paul Horwich (1987, Chapter 2) and Huw Mellor(1981, 1998), is that even though McTaggart showed that passage (that is, the A-series) is impossible, the B-series (that is, static classical spacetime structure) suffices for time.

Before we expand on this theme, though, first a few words about Broad's (b), his suspicion that there is some peculiarity of our language(s) that creates or at least reinforces the credibility of McTaggart's anti-passage argument. Broad suspected that there was a subtle ambiguity in the copula ‘is’ between tensed and tenseless uses, between the uses in, for instance:

> It is raining

and

> Seven is prime,

the former sentence containing a tensed and the latter sentence a non-tensed or tenseless copula. It has

been further suggested (Sellars 1962) that one might understand a non-tensed copula (indicated by 'be' rather than 'is') after the following fashion

> *S* be *F* at *t* iff (*S* was *F* at *t* or *S* is *F* at *t* or *S* will be *F* at *t*),

where the verbs to the right of the 'iff' (a logician's abbreviation for 'if and only if') are usual tensed verbs.

Alternatively, one might think of a tenseless copula as the usual copula stripped of temporal information (Quine, 1960, p. 170, Mellor 1981, 1998, Chapter 7), just as the usual copula carries no spatial information. If we indicate this tenseless copula by writing 'BE' instead of 'is,' we can say that 'It BE windy in Chicago' carries information about the place but not the time of the wind, just as 'It BE windy at t' tells us about its time but not its place.

These distinctions will prove helpful in the subsequent discussion of being in modern physics. For the moment, one might note that Broad could argue that McTaggart's (A_1) seems plausible if the copula is understood in some tenseless fashion, whereas (A_2) is plausible if the copula is tensed. If, however, the copula is not univocal in (A_1) and (A_2), then there is no contradiction involved in accepting both. (Savitt, 2001)

2.3 How (and How Not) To Think About Passage

If McTaggart's argument that passage is conceptually absurd or self-contradictory fails, philosophers mindful of modern physics are still left with Einstein's concern that passage and the now, while deeply embedded in human experience, seem to find no place in physics. One may agree with Carnap that "all that occurs objectively can be described in science" and then argue that passage reflects something perspectival or subjective and so is implicit in the physics or rightly omitted by it.

The most popular version of this view holds that *now* is a token-reflexive or indexical term, like *here* (Smart 1963, chapter VII; Mellor 1981, 1998). Physics is not felt to be incomplete because it fails to treat *hereness.* Why should its indifference to *nowness* be of any greater concern?

Early proponents of this view often claimed that '*S* is now *F*' meant '*S*'s being *F* is simultaneous with this utterance,' a quite implausible claim. A more sophisticated version of the view is that the truth-conditions of sentences like '*S* is now *F*' can be given solely in terms of the (tenseless) facts that exist or events that occur at the time of the utterance or inscription of the given sentence. It should be obvious how to extend the position to treat *past* and *future* in a similar fashion.

Smart claimed that excessive attention to the tensed notions of *now, past,* and *future* serve to project a "sort of anthropocentric idea on the universe at large."(1963, 132) But even if the tensed temporal locutions are anthropocentric and do locate *us* in the universe, it may still be asked whether these

temporal locations are in a static structure, “a four-dimensional continuum of spacetime entities,” (132) or in an unfolding or dynamic universe. Smart dismisses this latter view because, in his view, it involves the obscure or mistaken idea that events “become” or “come into existence.” Becoming and passage are mistakes, and harmful ones at that. Smart writes: “Our notion of time as flowing, the transitory aspect of time as Broad has called it, is an illusion which prevents us seeing the world as it really is.“ (132)

It will be useful to untangle a couple of ideas that are confounded in these quotes from Smart, with the help of some arguments of (mostly) Broad's (1938, section 1.22 of Chapter 35). First is the idea that time “flows” or, more generally, that passage is somehow to be thought of as like motion. Perhaps time itself somehow moves. Or perhaps, as Broad wrote in a famous sentence, “[t]he characteristic of presentness is … supposed to move along this series of event-particles, in the direction from earlier to later, as the light from a policeman's bullseye [flashlight] might move along a row of palings.”

Motion is one sort of change, change of spatial position with respect to time. The motion of time, then, must be change of time with respect to … What? If the answer, by analogy with motion, is “time”, one might be rightly puzzled as to how time (or anything else, for that matter) can change *with respect to itself.* Furthermore, if it is just time again, then the ratio of these two quantities expressing the rate of change is a pure or dimensionless number if the dimensions of the quantities in this ratio cancel. (See Price 1996, p. 13.) A pure number is not a rate of change, although it may represent various rates of change (for instance, 30 meters/second or 30 miles/hour). As Price remarks, “We might just as well say that the ratio of the circumference of a circle to its diameter flows at π seconds per second!”

If (in order to avoid this absurdity) the time in the denominator of the ratio expressing the rate of time's motion is held to be a different temporal dimension from the one in the numerator, then for it to be a genuine time there will have to be passage in it, requiring yet a third temporal dimension. One can see that we are at the beginning of an infinite regress, unless the third temporal dimension is identified with the first (as in Schlesinger 1980, Chapter II), leaving us in the uncomfortable position of having two temporal dimensions. It seems at best heroic, at worst hopeless, to try to understand passage as a kind of motion.

Broad also thought that trying to explain or represent passage in terms of qualitative change was “doomed to failure.” A thing or substance, *S*, can change in terms of a quality or property if property P_1 and property P_2 are determinates under a given determinable and *S* is P_1 at t_1 but P_2 at t_2. The passage of time, then, is to be thought of as an event's having (say) the property of presentness and then immediately losing that property but gaining (and losing in turn) a long and possibly endless series of properties of the increasing degrees of pastness.

In order for a thing to change it must evidently persist at least from t_1 to t_2, but the events usually supposed in discussions of passage are instantaneous events, which have no duration at all. They can not undergo qualitative change. It is sometimes argued that the properties that make up the A-series (and so change of which represents passage) are special properties, which even instantaneous events can gain and lose, but this is special pleading. As noted above, physics has so far no need of such special properties

and such special change and so is unlikely to be sympathetic to this special pleading.

Finally, Broad notes that (assuming one wishes to think of passage as like qualitative change) the acquisition and loss of (say) presentness by an event would itself be an event, a second-order event, in the history of a first-order event. Since the first-order events are, by hypothesis, durationless, it is tempting to suppose that this history takes place in a second temporal dimension. We find ourselves again launched on what looks to be an infinite regress of temporal dimensions.

These are strong arguments against two perennially tempting ways to construe temporal becoming--as like motion or qualitative change. They are strong arguments against the existence of temporal becoming if there is no other way to understand it. Broad thought, however, that he had a third way. Having pointed out the superficial grammatical similarity between '*E* became louder' and '*E* became present', Broad said that our understanding of these two kinds of assertions need not be dictated by it. He wrote (1938, p. 280-1):

> Again, any subject of which we can significantly say that it "became louder" must be a more or less prolonged noise-process, which divides into an earlier phase of less loudness adjoined to a later phase of greater loudness. But a literally *instantaneous* event-particle can significantly be said to "become present"; and, indeed, in the strict sense of "present" *only* instantaneous event-particles can be said to "become present". To "become present" is, in fact, just to "become", in an absolute sense; i.e., to "come to pass" in the Biblical phraseology, or, most simply, to "happen". Sentences like "This water became hot" or "This noise became louder" record facts of *qualitative change.* Sentences like "This event became present" record facts of *absolute becoming.*

The terminology may be pretentious, but the idea is simple. Absolute becoming is just the happening of events. The *raison d'être*, the very being or existence of events, is in their happening (at some place and time). If one is willing to embrace this category of entity at all, then one has the tools for a minimalist understanding of passage. Given the geometric richness of Newtonian spacetime, we can say that some events occur at the same time and so form a class of simultaneous events. If these classes can be, somehow, ordered, then we can say that some events occur before or after others. The passage of time is just the successive happening of (simultaneity sets of) events. It may be this picture of passage that the great logician Kurt Gödel had in mind when he wrote (1949, p. 558): "The existence of an objective lapse of time … means (or, at least, is equivalent to the fact) that reality consists of an infinity of layers of 'now' which come into existence successively."

There is an ambiguity in this last quote, however, that we must note. Did Gödel think that the layers of now come into existence (as what is to be becomes what is now) and then immediately cease to exist (as what is now becomes what once was), which is the presentist metaphysics of time? Or did he think that the layers of now come into existence and forever stay in existence, as the possibilist picture maintains? If one's basic ontology consists of the sort of events characterized above and often invoked in discussions of time, (idealized) instantaneous happenings, then the presentist picture seems inevitable.

The metaphysics of time is, however, one of the cross-roads of philosphy where issues intersect. If one thinks of a basic ontology consisting not of events but of substances or continuants, then one is apt to wonder what it is that makes sentences marking episodes in the histories of such substances--sentences like ‘*S* is Φ at *t*’--true. One frequent suggestion is that the “truth-makers” of such sentences are facts, the fact that at *t*, *S is* Φ. Then one might further note that in the current year, 2001, we can say:

1. It is a fact that Mount St. Helens erupted in Washington in 1980.
2. It is a fact that Jean Chretien is now Prime Minister of Canada.
3. It is a fact that there will be an eclipse of the sun in the Eastern United States in 2017.

These facts, when compared to evanescent events, seem to have great stability, the first one lasting (since it *is* a fact …) at least from 1980 till the present. The third one is, however, a special sort of fact, clearly not dependent on human will or choice and almost certainly not dependent upon any quantum measurements either. Future facts that do depend upon human choice or quantum measurement, should they be facts now, would seem to constrain human choice or quantum measurement in ways that many philosophers find undesirable. It is easy to convince oneself, then, that future facts of those two sorts can not really be part of the existing. Perhaps, then, facts like fact 3 above can be argued away as well. The result of this (lightly sketched) train of thought is, of course, the possibilist picture of time.

It seems unlikely that a simple argument will decide between these two metaphysical pictures of time, presentism and possiblism. Showing that McTaggart's argument is flawed, because it relies on an ambiguity in the copula ‘is’, and that there is a way to construe passage that side-steps the traditional objections, moreover, does not show that eternalism is false but only that it is optional. In Newtonian spacetime it may appear implausible, but it may fare better when we turn to Minkowski spacetime.

3. The Special Theory of Relativity

The Special Theory of Relativity (Einstein, 1905) was presented as a geometric theory of spacetime in Minkowski (1908).[3] For our purposes, the key change from Newtonian spacetime to Minkowski spacetime is that in the latter it is no longer the case that the temporal interval between any two points or events in the spacetime, *p* and *q*, is a well-defined quantity. In fact, the temporal interval between two points in the spacetime (and hence the simultaneity of two points in the spacetime) is not defined at all until a coordinate system or frame of reference (with some arbitrarily chosen spacetime point as origin of the frame) is chosen. A peculiar feature of special relativity (as opposed Newtonian physics) is that each coordinate system or frame of reference defined by an “observer” passing through the chosen origin and moving with some constant non-zero speed that is less than the speed of light (as measured in the first frame) picks out a *distinct* set of points as simultaneous with the origin. This feature of special relativity is called *the relativity of simultaneity*.

The relativity of simultaneity is a consequence of the even more startling assumption that each of these “observers”, no matter at what speed or in which direction they or the source of the light are moving (as

long as neither the speed nor the directions change), must come to the same result (conventionally indicated as *c*) when they measure the speed of light. We will not attempt to justify the assumption of the constancy of the speed of light here, though many standard texts present the empirical and theoretical background the led to it. Nor is it obvious that this assumption leads to the relativity of simultaneity, though one of the joys of even elementary presentations of the subject is that this *prima facie* astonishing connection can be convincingly demonstrated to persistent non-specialists.

A second assumption typically made in presentations of the special theory is *the Principle of Relativity:* All inertial frames of reference are completely equivalent for the formulation of the laws of physics.[4]

A glance back at Figure 1 reminds us that presentism and possibilism suppose that one plane of simultaneity is uniquely metaphysically important. In the former view, it represents all that exists. In the latter view, it is the locus of becoming, the dividing line between the merely possible future and the actual past-plus-present. The special theory of relativity tells us that there is an infinity of planes of simultaneity passing through any given spacetime point and that no physical test can distinguish one from amongst the lot. What was metaphysically distinguished is now physically indistinguishable. Assuming that we humans are complex physical systems, then we have no way to distinguish *the* present from amongst the multitude of presents.

An enthusiast could make much of this fact. For instance, the mathematician (and science fiction writer) Rudy Rucker wrote (1984, p. 149):

> As it turns out, it is actually *impossible* to find any objective and universally acceptable definition of "all of space, taken at this instant."" This follows … from Einstein's special theory of relativity. The idea of the block universe is, thus, more than an attractive metaphysical theory. It is a well-established scientific fact.

On the other hand, the distinguished philosopher and logician Arthur Prior thought that the above conclusion showed that special relativity is an incomplete view of reality (Prior, 1970):[5]

> One possible reaction to this situation, which to my mind is perfectly respectable though it isn't very fashionable, is to insist that all that physics has shown to be true or likely is that in some cases we can never *know*, we can never *physically find out*, whether something is actually happening or merely has happened or will happen.

We shall look at more nuanced reactions to the relativity of simultaneity below, but first it will be useful to introduce an argument that plays somewhat the same role in Minkowski spacetime as McTaggart's argument did in Newtonian spacetime. Versions of the argument are endorsed in papers by the physicist Cornellis Rietdijk (1966, 1976) and the philosopher Hilary Putnam (1967), but the presentation here will be based on an example found in Roger Penrose's book, *The Emperor's New Mind*.

Imagine that the Andromeda galaxy, which is about two million light years or 2×10^{19} kilometers from

Earth, is at rest with respect to Earth. On Earth two friends walk past each other, Alice walking along the Earth-Andromeda line towards Andromeda, Bob walking along that line but away from Andromeda. Each is walking at a comfortable pace, say 4 km/hour. One can calculate that their planes (or spaces) of simultaneity at the instant at which they pass each other on Earth (Call the event of their meeting **O**) intersect the history or *world line* of Andromeda about 5 ¾ days apart. (Call these two events **A** and **B**, respectively. We are idealizing Andromeda as a point, for the purpose of this example.) Imagine, finally, that during this 5 ¾ day period between **B** and **A** a momentous thing happens. The Andromedeans launch a space fleet aimed at invading Earth.

Figure 2. The Andromedean Invasion

The launch of the invading fleet is prior to **A** and so *in some sense* in Alice's past. But since the launch is after **B**, it is in that same sense in Bob's future. Penrose comments:

> Two people pass each other on the street; and according to one of the two people, an Andromedean space fleet has already set off on its journey, while to the other, the decision as to whether or not the journey will actually take place has not yet been made. How can there still be some uncertainty as to the outcome of that decision? If to *either* person the decision has already been made, then surely there *cannot* be any uncertainty. The launching of the space fleet is an inevitability. (p. 303)

This is an odd situation indeed. An event in Bob's future seems in some way to become fixed or inevitable by being in Alice's past. But that is not the end of the oddness here. Imagine that at point **A** (where Alice's plane of simultaneity intersects the world line of Andromeda) there is an Andromedean, Carol, who is walking directly towards Earth at about 4 km/hour. Then Carol's plane of simultaneity intersects Earth at some point **C** which is about 11 ½ days after **O**, the meeting of Alice and Bob. If all events (like **A**) in Alice's past or present at **O** have happened, are fixed, or are real, then the principle of relativity suggests that we must also extend the same courtesy to Carol; and so simultaneous with the fixed and real event **A** (Carol's walking towards Earth at exactly the point at which Alice's plane of simultaneity intersects the history of Andromeda) is the event **C** (and so fixed and real), the intersection of Carol's plane of simultaneity with Earth, which is in the future of *both* Alice and Bob. It is easy to see that, by adjusting the speeds of Alice and Carol, any event to the future of **O** can be shown to be fixed or real or inevitable. But **O** itself was just an arbitrarily chosen point in the spacetime. “It begins to seem that if anything is definite at all,” we might echo Penrose, “then the entire space-time must indeed be definite! There can be no ‘uncertain’ future.” (p. 304)

Roberto Torretti (1983, p. 249) calls the resulting view of the definiteness or fixity of all events in the spacetime *chronogeometrical determinism.* A slightly better name might be *chronogeometrical fatalism*, as we will see below. In order to see more clearly, however, what has gone wrong in the argument above, it will be useful first to look more closely at the problems attendant upon trying to import our commonsense or classical intuitions about time into the understanding of Minkowski spacetime and then to describe briefly the structures peculiar to that spacetime itself. To begin with the first task, one of the

most notable attempts to bring our time into Minkowski spacetime is to be found in Sellars (1962), a determined attempt by one of the most profound systematic metaphysicians of the latter half of the 20th century.

3.1 Relativizing the Present

Wilfrid Sellars believed that the various invariant or observer-*independent* elements of Minkowski spacetime (like the light cone structure to be described below) that are typically given primary consideration in treatments of relativity from a spacetime perspective are abstractions from and secondary to the ‘perspectival’ pictures, the myriad of coordinate systems or reference frames. When it comes to time, however, he believed there is something even more fundamental than these perspectives:

> …we must distinguish between a moment, t, and the event of the moment's being present with respect to a given perspective and, above all, between the event of the moment's being present with respect to a given perspective and the event of the moment's being *present.* The latter, of course, is the essential feature of a temporal picture of the world. (577)

While there is in Sellars' paper a lengthy and illuminating series of reflections on the relation between events, facts, and substances, there is no guidance offered on the relation between a moment's being present with respect to a given perspective and a moment's simply being present, a concept which is ill-formed from a relativistic point of view. If this latter is indeed an essential feature of a temporal picture of the world, then special relativity does not provide us with a temporal picture of the world. If the world *is* fundamentally temporal in the way that Sellars insists it is, then (at least as far a special relativity as a representation of that world is concerned), Sellars' famous scientific realism is compromised.

Even though Sellars' conservative attempt to import pre-relativistic categories into Minkowski spacetime fails, there are some useful lessons to learn from it. First, Sellars is careful to distinguish between events as things that happen or occur or take place and the ‘events’ (the use of single quotes is Sellars') that are basic in relativity. The latter are just spacetime points. They do not take place or occur, and they are not the relata in causal relations, whereas events are. (But cf. Tooley (1997, Chapter 9)) While it is not clear what precisely Sellars took the distinction to be, he is careful to mark a distinction between events and ‘events’.

Sellars also presents a distinction between what he calls (p. 586) *categorial existence statements* and what, for lack of a better term, I will call *non-categorial existence statements.* The former invoke frameworks, like the framework of substances or the framework of ‘events’, the frameworks that Sellars takes great pains to compare in his essay. He is inclined towards a view he credits (without source) to Carnap that to say that, for instance, ‘Things exist’ is to make the metalinguistic claim that there are thing words in our language L now. This use of ‘exist,’ claims Sellars, has no (future or past) tensed contrast.

Non-categorial existence statements, on the other hand, assert the existence of individuals or less general kinds in a fully tensed fashion. Sellars would construe them in the following way (p. 592):

> *a* be existent { before *now*, *now*, after *now* } $\equiv$
> $\exists x$(*x* be { before *now*, *now*, after *now* } *and* *x* be $\Phi_1,\ldots,\Phi_n$ *and*
> 'Φ_1',…,'Φ_n' be our criteria *now* for [being] '*a*')

Leaving aside Sellars' idiosyncratic way of construing existence statements, if a distinction like the one indicated here can be made, then it would be perfectly coherent to indicate that one is adopting or working in the framework of 'events' by asserting that 'events exist' (in the categorial sense) without being committed to the "tenseless existence" of particular 'events,' which may be past, present, or future (in the non-categorial sense).

It has sometimes been thought that commitment to a spacetime framework, as is often explicit in treatments of special relativity, is tantamount to commitment to eternalism, since to say that spacetime points exist seems inconsistent with saying that some spacetime points are future and so do not exist yet or are past and so exist no longer. *If* some distinction of the type just sketched can be made between categorial and non-categorial existence statements, then eternalism is not a straightforward consequence of adopting the spacetime viewpoint.[6]

Granting Sellars all the distinctions that he wishes, however, does not give him the tools to avoid the central problem sketched above. Since the problem is, in one form or another, the problem that any view that tries to define a notion of becoming in Minkowski spacetime must address, it is worth examining it a bit more closely. Sellars wrote (p. 591):

> … in the case of an 'event' framework, a primary temporal picture is a picture with a *now*. And even if one observer's *now* is another observer's *then*, or one observer's simultaneous cross sections of the world are another observer's sets of differently dated 'events',… each of their *now*-pictures is a primary picture, and the purely topological picture (which includes the measurements performed by *S* and *S'* as topological facts) which is common to them is not *the primary picture* of the world construed as a system of 'events,' but merely a topological abstraction common to the various primary pictures; and the topologically formulated location of individual events in the topological picture is merely the topologically invariant features of the criteria which identify these 'events' in a primary picture.

In this quote Sellars is using the term 'topological' where one would now normally use the term 'geometric', and he is forcefully reiterating his view that the spacetime manifold of 'events' is merely an abstraction from the infinity of distinct *primary* now-pictures of individual observers.

The first question one will surely want to ask about this view is: how can an infinity of distinct "now-pictures" each be primary? No answer is forthcoming. The second, and more troubling question, is: how can this infinity of distinct "now-pictures" be related to the traditional metaphysical views under discussion? What, in short, is the connection (if any) between the temporal notions implicit in each of the

pictures and existence of the past, present, and future? The striking fact about Sellars' schema above for ‘a be existent *now*’ is that it is *not* relativized to a reference frame, coordinate system or “observer” and so is not meaningful relativistically. The definition gives us no guidance as to how to parcel out existence to elements in the infinity of reference frames that are admissible at a spacetime point.

If the definition or schema above were relativized to frames F, F', etc., so as to connect existence to relativistically acceptable "primary now-pictures", its interpretation would be either unhelpful or mysterious. Consider the following modification of Sellars' schema above:

> *a* be existent *now* in *F* ≡
> ∃*x*(*x* be *now* in *F and x* be $\Phi_1, \ldots \Phi_n$ *and*
> ‘Φ_1,’…,‘Φ_n’ be our criteria *now* for ‘*a*’)

Suppose it is not the case that *a* be existent *now* in some other frame F'. It seems as if this difference must result from *a*'s being simultaneous with some spacetime point **O**, say, in F while not being simultaneous with the same point **O** as coordinatized in F'. But on this reading Sellars' schema is just a round-about way to indicate that simultaneity is relative--the point of departure for our metaphysical questions rather than the answer to any.

The schema looks as if it is meant to do something more, to connect temporal notions to existence. But if so, how is existence relative to a frame to be understood? Classical presentism, for instance, wishes to identify existence with present existence or existence *now*. Since the present is relativized to frames in special relativity, may not existence be relativized to frames as well? This is a difficult notion to understand or accept. Kurt Gödel (1949, p. 558) said flatly, “The concept of existence … cannot be relativized without destroying its meaning completely.” Is the concept of existence, then, like the concept of truth, which, when relativized (as true-for-me, true-for-you), comes to something more like belief than truth? Or is it like simultaneity, about which thoughtful persons a century or so ago might have made pronouncements much like Gödel's? This difficult and fundamental question has by no means been resolved.

Were this question resolved in favor of the relativization of existence, what would be the import of a relativized version of presentism? It would have to hold that what existed changed radically with one's state of motion. Certain events (say on Mars or a planet orbiting a distant star) may be existent for you now, sitting at your computer screen or reading a printout, but other events will replace those as existent should you decide to walk one way or another. This seems (once again) less like an interesting metaphysical insight than a restatement of the relativity of simultaneity. Possibilism is not better off in this regard, for it relies on a metaphysically distinguished present to separate the real from the potential. (See the symposium "The Prospects for the Present in Spacetime Theories" in Howard (2000) for further arguments and references.)

To sum up, then, Sellars' attempt to tie existence to temporal notions, when properly relativized, is either a bland re-statement of what special relativity tells us already about simultaneity or an opaque statement

about relativized existence. This dilemma confronts any attempt to import pre-relativistic notions in Minkowski spacetime. Let us, then, turn to efforts to understand Minkowski spacetime in a different way, efforts that will help clarify the puzzling argument about the Andromedean invasion presented above.

3.2 Chronogeometrical Fatalism Again

We have said much about the relativity of simultaneity above but little directly about the invariance of the speed of light. We must now rectify that situation.

Imagine that at some point **O** of the spacetime an idealized point-sized flashbulb flashes for (literally) an instant. It follows from the invariance of the speed of light that Alice, passing through **O** as above, will find herself at the center of a sphere of photons. The radius of the sphere expands with speed *c*. (It follows that Bob, also passing through **O** but moving with some constant velocity with respect to Alice, must find himself also at the center of such a sphere, even though he and Alice are walking away from each other. Such is relativistic life!) If we try to diagram this situation, it is helpful to suppress one spatial dimension, as we have in all the figures above, and so the two-dimensional cut through the expanding sphere looks like an expanding circle, which becomes a cone when that growth is plotted vertically up the diagram (and so is called *the light cone.*) More precisely, this figure is just half the light cone. If two photons (restricting ourselves to two dimensions now) converged on point **O** from opposite directions, the lines indicating their histories would mark the other half, the past lobe, of the light cone.[7]

The light cone exists at each point of the spacetime and is an invariant structure. Since the speed of light is an invariant quantity, all “observers” agree as to which points of the spacetime are illuminated by the popping of the flashbulb at **O**. Furthermore, as special relativity is standardly understood, the speed of light is a limiting speed. No material particle can be accelerated from a speed less that *c* to a speed equal to or greater than *c*. Electromagnetic radiation (including light) always propagates in a vacuum at speed *c*. (To see why *c* is held to be limiting velocity, speed, see Mermin (1968, Chapter 15) and Nahin (1999, pp. 342-353 and Tech Note 7.) Given these suppositions, the light cone structure divides all spacetime into three distinct sorts of regions relative to each spacetime point **O**. (See chapters 5 and 6 of Geroch (1978) for a thorough discussion.)

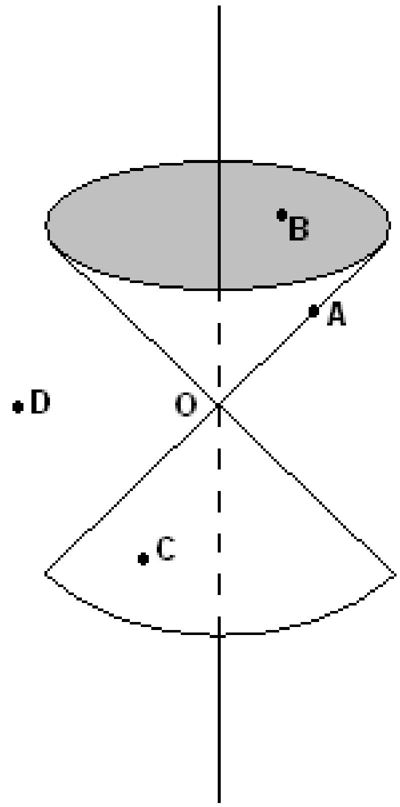


Figure 3. The Light Cone

First, there are the points from which a photon may travel to **O** or which may be reached by a photon from **O**. We say that these points are *lightlike separated* from **O**. If a photon can travel from **O** to **A**, we can indicate this briefly by writing **O** < **A**. In this case, **A** lies on the future light cone of **O**.

Second, there are the points inside (rather than on) the future or past light cone of **O**. We say that these points are *timelike separated* from **O**. If **B** is a point in the spacetime timelike separated from **O** and future to it (that is, inside **O**'s future light cone), then a material particle travelling at some relativistically acceptable speed (that is, less than *c*) can travel from **O** to **B**. Similarly, a material particle at a point inside the past light cone of **O**, can travel at some speed less than *c* from **C** to **O**. In this case we write **C** << **O**; in the former case, **O** << **B**.

Finally, there are the points of the spacetime that are neither in nor on the light cone of **O**. We say that such points are *spacelike separated* from **O**. If **D** is spacelike separated from **O**, then no light signal and no material body can travel from **O** to **D** or *vice versa*, because such travel would require superluminal speed. If one makes the natural assumption that information and causal influence are propagated by electromagnetic signals and material particles, then if **D** is spacelike separated from **O**, events or occurrences at **O** can have no causal influence at all on events at **D**.

We have reached this last conclusion by means of quite straightforward reasoning from the invariance of the speed of light. But consider the following observation of Torretti (1983, p. 247):

> Before Einstein … nobody appears to have seriously disputed that any two events might be

> causally related to each other, regardless of their spatial and temporal distance. The denial of this seemingly modest statement is perhaps the deepest innovation in natural philosophy brought about by Relativity. It has completely upset our traditional views of time, space, and causality …

As one illustration of how our traditional views of time and causality are upset by restricting the propagation of causal influence to the light cone structure, let us revisit the reasoning of the example of the Andromedean invasion that we used to illustrate and motivate chronogeometrical fatalism. We may be able to see now that this reasoning is not so compelling as it first seemed, and we may be able to see why some philosophers have proposed that we look at becoming in Minkowski spacetime in a way quite different from the traditional way.

To make the exposition easier, let us add to the story of the Andromedean invasion a fourth observer, Ted, who is at rest with respect to Earth (and so also Andromeda) at the spot where Alice and Bob meet. Ted too defines a coordinate system or frame of reference, and there is a point at Andromeda (We can call it **D**) that (in Ted's frame) is simultaneous with the meeting of Alice and Bob and Ted. To make our exposition easier still, let us suppose that Alice and Bob and Ted all set their clocks to read 0 at the instant at which they all meet.[8] Let us focus on **D**.

Ted (at the meeting of Alice and Bob) assigns to **D** the time 0, since it is simultaneous (in his frame) with his time 0. Alice assigns **D** (roughly) the time -3 days, whereas Bob assigns it time (roughly) +3 days. **D** is, of course, spacelike separated from **O**, and we have been at pains to explain that from a special relativistic standpoint this spacelike separation precludes the (physical) possibility that there is any causal influence upon **D** of the events at **O**. Once the labelling of spacetime points like **D** with coordinates is complete, what further content is there, what further could be meant, by adding that for Alice and Ted **D** is real or fixed? If there is indeed no further content, then what possible implications with regard to ‘reality’ or ‘fixity’ or ‘determinateness’ can be drawn from the fact that Bob labels this point with a positive number, Alice labels it with a negative number, and Ted labels it with 0?[9]

A good text in special relativity will sooner or later prove that for *any* pair of spacelike separated points (but let us continue to call them **O** and **D**) there is precisely one admissible coordinate system (with **O** as origin) in which **O** and **D** are simultaneous, an infinity of admissible coordinate systems in which **D** is assigned a positive number (that is, in which **O** occurs before **D**), and an infinity of other admissible coordinate systems in which **D** is assigned a negative number (that is, in which **D** occurs before **O**). What metaphysical significance could be gleaned from the fact that some observers (the usual anthropomorphized way to refer to admissible coordinate systems) at **O** must assign positive times, some negative times, and one time 0 to the distant event **D**, which, again, can not be influenced by and can not itself influence the events at **O**, according to special relativity at least?

Inability to provide any positive answer to this question can motivate a different approach to conceptualizing becoming in Minkowski spacetime, an approach presented by the philosopher Howard Stein (1968, 1991). The basic idea of this approach is to begin from or to define concepts in terms of the

geometric structure intrinsic to the spacetime. In the present case, this approach leads one to try to define ‘becoming’ in terms of spacetime points and light cones. Pre-relativistically, ‘has become’ is defined relative to a plane of simultaneity. We have seen the limitations of the notion of a plane of simultaneity in special relativity. Stein begins, then, by proposing that one defines the relation of ‘having become’ or ‘already definite’ with respect to spacetime points. A two-place relation schematically written as *Rxy* will be intended to capture the idea that point *y* has already become or is definite with respect to point *x*.

There are two other formal features that this relation *R* should possess. It should be *transitive* -- that is, if *z* has already become with respect to *y* and *y* has already become with respect to *x*, then it seems reasonable to require that *z* has already become with respect to *x*. It should also be *reflexive* -- that is, is seems reasonable to require that *x* has become with respect to *x* itself.

(We can indicate these conditions briefly as (1) *Rzy* and *Rxz* entail *Rxy*, for all *x*, *y*, *z* and (2) *Rxx*, for all *x*.)

Finally, Stein proposes that the relation *R* not hold between every two points in the spacetime. That is, he proposes that given some choice of spacetime point *x*, there is at least one distinct point *y* that has not become, that is not already definite, with respect to *x*. But is there any such relation, any relation that has all these intuitively desirable characteristics? The answer is *yes*. The relation is that between a point *x* and each point in or on its past light cone.[10] If one can accept that the relation *Rxy* represents in special relativity the notion of becoming (or, having become), then the existence of the relation specified and found by Stein is a formal refutation of the Rietdijk-Putnam-Penrose argument for chronogeometric fatalism.

It is this last issue, of course, that is controversial. Stein, who wishes to tie his definitions of temporal concepts to intrinsic geometric structure, holds that “in Einstein-Minkowski space-time *an event's present is constituted by itself alone.*” (1968, p. 15) If one wishes to include even *one* other event in an event's present--that is, if one specifies that for each point *x* there must be one other distinct point *y* such that not only *Rxy* but also *Ryx*--then the only relation that satisfies this *desideratum* and the other conditions specified by Stein is the universal relation.[11]

Callender (2000, S592) remarks that requiring that an event‘s present must contain at least one event distinct from it, which he calls the *non-uniqueness condition*, “seems the thinnest requirement one might put on becoming.” He would then not accept Stein's relation *R* as representing a genuine relation of becoming since it fails to meet this condition, but then he also must accept the conclusion of the Rietdijk-Putnam-Penrose argument, since the only alternative to *R* is the universal relation. If one wishes to evade chronogeometric fatalism, as far as the special theory of relativity is concerned, then it seems there is no alternative to accepting Stein's relation *R* as representing a genuine relation of becoming and to considering that an event's present is constituted by itself alone. It is a truism that the relativistic revolution in physics has profound implications for our concepts of space and time. This last dilemma shows why that truism is true.

There may seem to be an insuperable obstacle to accepting Stein's relation *R* as representing a genuine relation of becoming. *R* is supposed to represent becoming, but the light cone structure of Minkowski spacetime, in terms of which it is defined, is inert. This reaction was voiced, for instance, by Palle Yourgrau, who wrote that “Stein's mistake is to adduce a *structural* property as what ‘justifies the use of our notion of *“becoming”* in relativistic spacetime.”’ (1999, p. 77) If Yourgrau has put his finger on a “mistake”, then it is a “mistake” at the very heart of Stein's effort. There are, however, a few remarks to be made on this score.

First, there have been attempts to articulate positions like Stein's that try to account for passage in terms of geometric structure and that seem to incorporate more dynamic elements, exploiting the fact that persistent objects or substances (including “observers”) are represented by timelike world lines, rather than by points. The mathematician G. J. Whitrow (1980, p. 348) wrote:

> At a given instant *E* on the world line of an observer *A* (who need not be regarded as anything more than a recording instrument), all the events from which *A* can have received signals lie within the backwards-directed light cone with its vertex at *E*… . Signals from events [outside the light cone at *E*] can only reach *A* after the event *E*, and when they do reach *A* they will then lie within *A*'s backward-directed light cone at that instant. The passage of time corresponds to the continual advance of this light cone.

The physicist-philosopher Abner Shimony, in responding to the claim that special relativity shows that becoming is subjective or “mind-dependent,” wrote (1993, p. 284):

> Something fleeting does indeed traverse the world line, but that something is not subjective; it is the transient *now,* which as a matter of objective fact is momentarily present and thereafter is past.

In the felicitous phrase of Park (1971), we have here two different sorts of *animated Minkowski diagram.* Each seems to involve a kind motion, of the light cone or of the transient *now* advancing along a world line. Our initial restrictions on accounts of transience inspired by Broad's arguments should make us wary of invoking motion to account for passage. Park, moreover, sees no benefit to adding the animation.

> I want now to make the vital point that the animated diagram may be more intuitive, or more picturesque, or make better cinema than the atemporal one, but that it contains no more specific, verifiable information. All of the science of dynamics, that is, all we know about how complex systems (including ourselves) behave and interact, is already represented on the atemporal Minkowski diagram.

The non-animated Minkowski diagram may be “static”, but, as Park points out, the static diagram *represents* the evolution in (proper) time of systems along their world lines. The diagram, if Park is correct, need not itself be animated to represent dynamical phenomena. If Park is correct, then what Yourgrau called a “mistake” is in fact a virtue of Stein's account, that he makes no attempt to animate his

geometric *picture* but leaves whatever transience there may be in what it depicts.

Bibliography

- Born, M. 1962. *Einstein's Theory of Relativity*. New York City: Dover Publications. (This is a revised and translated version of the original German text of 1920.)
- Broad, C. D. 1938. *Examination of McTaggart's Philosophy,* Vol. II, Part I. Cambridge: Cambridge University Press.
- Broad, C. D. 1959. "A Reply to My Critics" in *The Philosophy of C. D. Broad,* P. A. Schilpp (ed.), pp. 711-830. New York City: Tudor Publishing.
- Butterfield, J. (ed.) 1999. *The Arguments of Time.* Oxford: Oxford University Press.
- Callender, C. 2000. "Shedding Light on Time," in Howard (ed.) 2000, pp. S587-S599.
- Carnap, R. 1963. "Carnap's Intellectual Biography" in *The Philosophy of Rudolf Carnap,* P. A. Schilpp (ed.), pp. 3-84. La Salle, IL: Open Court.
- Clifton, R. and Hogarth, M. 1995. "The Definability of Objective Becoming in Minkowski Spacetime," *Synthese* **103**: 355-387.
- Einstein, A. 1905. "On the Electrodynamics of Moving Bodies," as reprinted and translated in *The Principle of Relativity,* pp. 35-65. New York City: Dover Publications, 1952).
- Fitzgerald, P. 1985. "Four Kinds of Temporal Becoming," *Philosophical Topics* **13**: 145-177.
- Friedman, M. 1983. *Foundations of Space-Time Theories: Relativistic Physics and Philosophy of Science.* Princeton: Princeton University Press.
- Gale, R. 1967. *The Philosphy of Time: a Collection of Essays.* Garden City, NY: Doubleday and Company.
- Geroch, R. 1978. *General Relativity from A to B.* Chicago: The University of Chicago Press.
- Gödel, K. 1949. "A Remark about the Relationship Between Relativity and Idealistic Philosophy," in *Albert-Einstein: Philosopher-Scientist,* Schilpp, P. (ed.), pp. 557-62. La Salle, IL: Open Court.
- Grünbaum, A. 1971. "The Meaning of Time," in *Basic Issues in the Philosophy of Time,* Freeman, E. and W. Sellars (eds.), pp 195-228. La Salle, IL: Open Court.
- Grünbaum, A. 1973. *Philosophical Problems of Space and Time,* (second, enlarged edition). Dordrecht, Holland and Boston, MA: D. Reidel Publishing Company.
- Horwich, P. 1987. *Asymmetries in Time: Problems in the Philosophy of Science.* Cambridge, MA: The MIT Press.
- Howard, D. (ed.) 2000. *PSA 1998: Proceedings of the 1998 Biennial Meeting of the Philosophy of Science Association,* Part II: Symposia papers. *Philosophy of Science,* Supplement to Volume 67, Number 3.
- McTaggart, J. M. E. 1908. "The Unreality of Time," *Mind*, New Series **68**: 457-484.
- McTaggart, J. M. E. 1927. *The Nature of Existence,* Vol. II. Cambridge: Cambridge University Press.
- Mellor, D. H. 1981. *Real Time.* Cambridge: Cambridge University Press.
- Mellor, D. H. 1998. *Real Time II* London and New York: Routledge.
- Mermin, N. D. 1968. *Space and Time in Special Relativity.* Prospect Heights, IL: Waveland Press, Inc.

- Minkowski, H. 1908. "Space and Time," as reprinted and translated in *The Principle of Relativity,* pp. 73-91. New York City: Dover Publications, 1952).
- Nahin, P. 1999. *Time Machines: Time Travel in Physics, Metaphysics, and Science Fiction.* 2nd edn. New York, Berlin, and Heidelberg: Springer-Verlag.
- Oaklander, N. and Smith, Q. (eds.) 1994. *The New Theory of Time.* New Haven and London: Yale University Press. Park, D. 1971. "The Myth of the Passage of Time," *Studium Generale* **24**: 19-30. Reprinted in *The Study of Time,* J.T. Fraser, F. C. Haber, and G. H. Müller (eds.) Berlin, Heidelberg, and New York: Springer-Verlag, 1972.
- Penrose, R. 1989. *The Emperor's New Mind: Concerning Computers, Minds, and Laws of Physics.* New York and Oxford: Oxford University Press.
- Price, H. 1996. *Time's Arrow & Archimedes' Point: New Directions for the Physics of Time.* New York and Oxford: Oxford University Press.
- Prior, A. 1970. "The Notion of the Present," *Studium Generale* **23**: 245-48. Reprinted in *The Study of Time*, J.T. Fraser, F. C. Haber, and G. H. Müller (eds.) Berlin, Heidelberg, and New York: Springer-Verlag, 1972.
- Putnam, H. 1967. "Time and Physical Geometry," *Journal of Philosophy* **64**: 240-247. Reprinted in Putnam's *Collected Papers,* Vol. I. Cambridge: Cambridge University Press, 1975.
- Quine, W. V. O. 1960. *Word and Object.* Cambridge, MA: The MIT Press.
- Rietdijk, C. 1966. "A Rigorous Proof of Determinism Derived from the Special Theory of Relativity," *Philosophy of Science,* **33**: 341-4.
- Rietdijk, C. 1976. "Special Relativity and Determinism," *Philosophy of Science,* **43**: 598-609.
- Rucker, R. 1984. *The Fourth Dimension.* Boston: Houghton Mifflin Co.
- Savitt, S. 2001. "A Limited Defense of Passage," *American Philosophical Quarterly,* **38**: 261-270.
- Schlesinger, G. N. 1980. *Aspects of Time.* Indianapolis, IN: Hackett Publishing Company.
- Sellars, W. 1962. "Time and World Order" in *Minnesota Studies in the Philosophy of Science,* Vol. III, Feigl, H. and Maxwell, G. (eds.), pp 527-616. Minneapolis: University of Minnesota Press.
- Shimony, A. 1993. "The Transient *now*," in *Search for a Naturalistic World View* Vol. II. Cambridge: Cambridge University Press.
- Smart, J. J. C. 1963. *Philosophy and Scientific Realism.* New York: The Humanities Press.
- Stein, H. 1968. "On Einstein-Minkowski Space-Time," *The Journal of Philosophy* **65**: 5-23.
- Stein, H. 1991. "On Relativity Theory and Openness of the Future," *Philosophy of Science* **58**: 147-167.
- Taylor, E. F. and J. A. Wheeler. 1963. *Spacetime Physics.* San Francisco and London: W. H. Freeman and Company.
- Tooley, M. 1997. *Time, Tense, and Causation.* Oxford: Oxford University Press.
- Tooley, M. 1999. "The Metaphysics of Time" in *The Arguments of Time.* J. Butterfield (ed.), pp 21-42. Oxford: Oxford University Press.
- Torretti, R. 1983. *Relativity and Geometry*. Oxford, New York, Toronto, Sydney, Paris, Frankfurt: Pergamon Press.
- Wheelwright, Philip. 1960. *The Presocratics.* Indianapolis: Bobbs- Merrill.
- Whitrow, G. 1961. *The Natural Philosophy of Time.* Oxford: Oxford University Press. (2nd edn., 1980.)

- Yourgrau, P. 1999. *Gödel Meets Einstein: Time Travel in the Gödel Universe.* Chicago and La Salle, IL: Open Court. (Revised and expanded edn. of *The Disappearance of Time: Kurt Gödel and the Idealistic Tradition in Philosophy*. Cambridge: Cambridge University Press, 1991.)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Broad, Charles Dunbar | events | facts | McTaggart, John M. E. | Sellars, Wilfrid

Steven Savitt
savitt@interchange.ubc.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 11, 2001
Content last modified: July 11, 2001

Stanford Encyclopedia of Philosophy
Notes to Being and Becoming in Modern Physics

Notes

1. As translated in Wheelwright (1960). The quote by Parmenides in what follows is also from this volume.

2. Many recent papers on these issues are collected in Oaklander and Smith (1994). Gale (1967) has some good older papers and a useful bibliography.

3. There are many excellent non-technical introductions to the special theory. Two fine books that are currently available are Mermin (1968) and Born (1962). A more demanding introduction mathematically is Taylor and Wheeler (1963). An excellent philosophical discussion is Chapter IV of Friedman (1983).

All the concepts needed for the present discussion are outlined briefly in the opening paragraphs of section 4 of Shimony (1993), but there is no substitute for working through in detail at least one presentation of the special theory at whatever level of mathematical sophistication one is equipped to handle.

4. While most popular presentations of special relativity explicitly employ only these two assumptions, Friedman (1983) points out that another assumption of a more technical nature, the flatness of Minkowski spacetime, is needed in order to derive all the characteristic results of the theory. We will ignore this refinement here.

One should note, however, that the two assumptions explicitly made are assumptions concerning *invariance*--the invariance of the speed of light and the laws of physics. That certain other quantities classically thought to be invariant turn out not to be so in special relativity has sometimes obscured the fact that there is a fundamental invariant special relativisic four-dimensional quantity called *the spacetime interval* that will enter our considerations in due course.

5. Hans Reichenbach indicated the same view in 1925. See Grünbaum (1973, p. 318).

6. Whether this suggested distinction overlaps or is independent of the distinction between tensed and tenseless uses of ‘is’ invoked above in the section on Newtonian Spacetime is an open question. Questions about the viability of this distinction are connected to deep questions in ontology and philosophy of language on which Carnap, Quine, and Sellars differed. See the discussion in Jay Rosenberg's entry in this Encyclopedia, Wilfrid Sellars.

7. Minkowski spacetime is a *time orientable* manifold. If one chooses one of the two lobes of the light cone at a point **O** to be, say, future, that choice can be extended smoothly throughout the whole of the spacetime. We say nothing as to how this choice is to be made in this entry, but we assume that it has been, somehow, made.

8. The three are free to choose **O** as the origin of each of their coordinate systems and to assign it spatial coordinate (0,0,0) and temporal coordinate 0. But what position and time values are assigned by each of them to other spacetime points now follows rigorously from the rules, the Lorentz transformations, of special relativity.

9. It is the fact the Rietdijk-Putnam-Penrose argument for the fixity of the future does not rely on features of natural laws or causation that leads me to call the thesis chronogeometric *fatalism* rather than chronogeometric *determinism.* Determinist and fatalist arguments have the same conclusion, that the future is somehow fixed and not within our control, but the former do so from causal or nomological considerations while the later do not.

10. Briefly, Rxy iff ($y < x$ or $y << x$). Clifton and Hogarth (1995) point out the relation betwen x and each point in (but not on) its past light cone also satisfies all the criteria of adequacy specified in the text.

11. This result is implicit in the proofs offered by Stein and by Clifton and Hogarth. It is made explicitly in Callender (2000).

Steven Savitt
href="mailto:savitt@interchange.ubc.ca">*savitt@interchange.ubc.ca*

First published: July 11, 2001
Content last modified: July 11, 2001

The Experience and Perception of Time

We see colours, hear sounds and feel textures. Some aspects of the world, it seems, are perceived through a particular sense. Others, like shape, are perceived through more than one sense. But what sense or senses do we use when perceiving time? It is certainly not associated with one particular sense. In fact, it seems odd to say that we see, hear or touch time passing. And indeed, even if all our senses were prevented from functioning for a while, we could still notice the passing of time through the changing pattern of our thought. Perhaps, then, we have a special faculty, distinct from the five senses, for detecting time. Or perhaps, as seems more likely, we notice time through perception of other things. But how?

Time perception raises a number of intriguing puzzles, including what it means to say we *perceive* time. In this article, we shall explore the various processes through which we are made aware of time, and which influence the way we think time really is. Inevitably, we shall be concerned with the psychology of time perception, but the purpose of the article is to draw out the philosophical issues, and in particular whether and how aspects of our experience can be accommodated within certain metaphysical theories concerning the nature of time and causation.

What is ‘the perception of time’?

The very expression ‘the perception of time’ invites objection. Insofar as time is something different from events, we do not perceive *time* as such, but changes or events *in* time. But, arguably, we do not perceive events only, but also their temporal relations. So, just as it is natural to say that we perceive spatial distances and other relations between objects (I see the dragonfly as hovering above the surface of the

water), it seems natural to talk of perceiving one event following another (the thunderclap as following the flash of lightening), though even here there is a difficulty. For what we perceive, we perceive as *present*-as going on right now. Can we perceive a relation between two events without also perceiving the events themselves? If not, then it seems we perceive both events as present, in which case we must perceive them as simultaneous, and so not as successive after all. There is then a paradox in the notion of perceiving an event as occurring after another, though one that perhaps admits of a straightforward solution. When we perceive B as coming after A, we have, surely, ceased to perceive A. In which case, A is merely an item in our memory. Now if we wanted to construe 'perceive' narrowly, excluding any element of memory, then we would have to say that we do not, after all, perceive B as following A. But in this article, we shall construe 'perceive' more broadly, to include a wide range of experiences of time that essentially involve the senses. In this wide sense, we perceive a variety of temporal aspects of the world. We shall begin by enumerating these, and then consider accounts of how such perception is possible.

Kinds of temporal experience

There are a number of what Ernst Pöppel (1978) calls 'elementary time experiences', or fundamental aspects of our experience of time. Among these we may list the experience of (i) duration; (ii) non-simultaneity; (iii) order; (iv) past and present; (v) change, including the passage of time. It might be thought that experience of non-simultaneity is the same as experience of time order, but it appears that, when two events occur very close together in time, we can be aware that they occur at different times without being able to say which one came first (see Hirsh and Sherrick (1961)). We might also think that perception of order was itself explicable in terms of our experience of the distinction between past and present. There will certainly be links here, but it is a contentious question whether the experience of *tense*-that is, experiencing an event as past or present-is more fundamental than the experience of order, or vice versa, or whether indeed there is such a thing as the experience of tense at all. This issue is taken up below. Finally, we should expect to see links between the perception of time order and the perception of motion if the latter simply involves perception of the order of the different spatial positions of an object. This is another contentious issue that is taken up below.

Duration

One of the earliest, and most famous, discussions of the nature and experience of time occurs in the autobiographical *Confessions* of St Augustine. Augustine was born in Numidia (now Algeria) in 354 AD, held chairs in rhetoric at Carthage and Milan, and become Bishop of Hippo in 395. He died in 430. As a young adult, he had rejected Christianity, but was finally converted at the age of 32. Book XI of the *Confessions* contains a long and fascinating exploration of time, and its relation to God. During the course of it Augustine raises the following conundrum: when we say that an event or interval of time is short or long, what is it that is being described as of short or long duration? It cannot be what is past, since that has ceased to be, and what is non-existent cannot presently have any properties, such as being long. But neither can it be what is present, for the present has no duration. (For the reason why the present must be regarded as durationless, see the section on the specious present, below.) In any case, while an event is

still going on, its duration cannot be assessed.

Augustine's answer to this riddle is that what we are measuring, when we measure the duration of an event or interval of time, is in the memory. From this he derives the radical conclusion that time itself (or, at least, the past and future) is something in the mind. While not following Augustine all the way to his theory of the subjectivity of time, we can concede that the perception of temporal duration is crucially bound up with memory. It is some feature of our memory of the event (and perhaps specifically our memory of the beginning and end of the event) that allows us to form a belief about its duration. This process need not be described, as Augustine describes it, as a matter of measuring something wholly in the mind. Arguably, at least, we are measuring the event or interval itself, a mind-independent item, but doing so by means of some psychological process.

Whatever the process in question is, it seems likely that it is intimately connected with what William Friedman (1990) calls ‘time memory’: that is, memory of when some particular event occurred. That there is a close connection here is entailed by the plausible suggestion that we infer (albeit subconsciously) the duration of an event, once it has ceased, from information about how long ago the beginning of that event occurred. That is, information that is *metrical* in nature (e.g. ‘the burst of sound was very brief’) is derived from *tensed* information, concerning how far in the past something occurred. The question is how we acquire this tensed information. It may be direct or indirect, a contrast we can illustrate by two models of time memory described by Friedman. He calls the first the *strength model* of time memory. If there is such a thing as a memory trace that persists over time, then we could judge the age of a memory (and therefore how long ago the event remembered occurred) from the strength of the trace. The longer ago the event, the weaker the trace. This provides a simple and direct means of assessing the duration of an event. Unfortunately, the trace model comes into conflict with a very familiar feature of our experience: that some memories of recent events may fade more quickly than memories of more distant events, especially when those distant events were very salient ones (visiting a rarely seen and frightening relative when one was a child, for instance.) A contrasting account of time memory is the *inference model*. According to this, the time of an event is not simply read off from some aspect of the memory of it, but is inferred from information about relations between the event in question and other events whose date or time is known.

The inference model may be plausible enough when we are dealing with distant events, but rather less so for much more recent ones. In addition, the model posits a rather complex cognitive operation that is unlikely to occur in non-human animals, such as the rat. Rats, however, are rather good at measuring time over short intervals of up to a minute, as demonstrated by instrumental conditioning experiments involving the ‘free operant procedure’. In this, a given response (such as depressing a lever) will delay the occurrence of an electric shock by a fixed period of time, such as 40 seconds, described as the R-S (response-shock) interval. Eventually, rate of responding tracks the R-S interval, so that the probability of responding increases rapidly as the end of the interval approaches. (See Mackintosh (1983) for a discussion of this and related experiments.) It is hard to avoid the inference here that the mere passage of time itself is acting as a conditioned stimulus: that the rats, to put it in more anthropocentric terms, are successfully estimating intervals of time. In this case, the strength model seems more appropriate than the inference model.

The specious present

The term ‘specious present’ was first introduced by the psychologist E.R. Clay, but the best known characterisation of it was due to William James, widely regarded as one of the founders of modern psychology. He lived from 1842 to 1910, and was professor of philosophy at Harvard. His definition of the specious present goes as follows: ‘the prototype of all conceived times is the specious present, the short duration of which we are immediately and incessantly sensible’ (James (1890)). How long is this specious present? Elsewhere in the same work, James asserts ‘We are constantly aware of a certain duration-the specious present-varying from a few seconds to probably not more than a minute, and this duration (with its content perceived as having one part earlier and another part later) is the original intuition of time.’ This surprising variation in the length of the specious present makes one suspect that more than one definition is hidden in James' rather vague characterisation. One could define it, for example, as the extent of short-term memory, in which case it might well vary from person to person, and also from one sense modality to another. Or it might be the interval in which information is experienced as a single unit (say a sentence, or musical phrase)-a rather ambiguous and unsatisfactory definition. A quite different definition is this: the interval of time such that events occurring within that interval are experienced as present. This is how the specious present tends to be treated in recent discussions, though it is inconsistent with James' remark that we can discern earlier and later parts in the specious present. As we remarked at the beginning of this article, if two events are experienced as present, they are surely experienced as simultaneous.

Taking the specious present as defined by this third characterisation, the *doctrine of the specious present* holds that the group of events we experience at any one time as present contains successive events spanning an interval. The experienced present is ‘specious’ in that, unlike the objective present, it is an interval and not a durationless instant. The ‘real’ present, as we might call it, must be durationless for, as Augustine argued, in an interval of any duration, there are earlier and later parts. So if any part of that interval is present, there will be another part that is past or future. This definition needs to be tightened up a little, to distinguish the tendency of the mind to group together successive events from the familiar fact that light and sound travel at finite speeds, and so events experienced as present will in fact be past (the degree of pastness varying with distance). What matters, as far as the doctrine is concerned, is not when an event occurred, but when information from that event reached our sense organs. Thus, light beams from two events may reach the retina at slightly different times, and yet the two events be perceived as simultaneous.

A number of arguments have been advanced in favour of the doctrine of the specious present (see Mundle (1966)):

(A) We see things as moving, such as the second-hand of a clock, and ‘to see a second-hand moving is quite a different thing from "seeing" that a hour-hand has moved.’ (Broad (1923)) More formally:

> (1) What we see, we see as present.

(2) We see motion.

(3) Motion occurs over an interval.

Therefore: What we see as present occurs over an interval.

(B) If the experienced present were only an durationless instant, then we could not understand a spoken sentence, because what would be presented to the senses at any one point would only be a meaningless phoneme—indeed not even that, since any sound necessarily takes up time (Gombrich (1964)).

(C) If the experienced present were only a durationless instant, then we would not see pictures on the television screen or VDU of a computer, since these are built up from a moving electron beam. More generally, we would not see anything at all, since light itself is a motion (*Ibid.*).

However, the first two of these arguments are questionable (at least as arguments for the doctrine as we have characterised it; they may be more appropriate for other conceptions of the specious present). If events e_1 and e_2 are registered in a single specious present, then we perceive them both as present, and so as simultaneous. But we do not see, e.g., the successive positions of a moving object as simultaneous, for if we did we would see a blurred object and not a moving one. So (in response to A) to see an object as moving is not to see as present something that occurs over an interval. Similarly, we do not hear all the parts of a spoken sentence as simultaneous, for if we did it would be a meaningless jumble. So (in response to B) we do not hear all the parts of the sentence as present.

C, in contrast, appears to be sound, and does not involve the contradictory suggestion that we experience some things as both ordered and as present. When events occur sufficiently fast, such as the movement of the electron beam over the television screen, we simply fail to perceive the temporal order of certain components of our experience. In these cases, we see things as simultaneous when they are not simultaneously presented to our sensory apparatus, and that is the basis of the true doctrine of the specious present.

Past, present and the passage of time

The previous section indicated the importance of distinguishing between perceiving the present and perceiving something *as* present. We may perceive as present items that are past. Indeed, given the finite speed of the transmission of both light and sound (and the finite speed of transmission of information from receptors to brain), it seems that we only ever perceive what is past. However, this does not by itself tell us what it is to perceive something as present, rather than as past. Nor does it explain the most striking feature of our experience as-of the present: that it is constantly changing. The passage (or apparent passage) of time is its most striking feature, and any account of our perception of time must account for this aspect of our experience.

Here is one attempt to do so. The first problem is to explain why our temporal experience is limited in a way in which our spatial experience is not. We can perceive objects that stand in a variety of spatial

relations to us: near, far, to the left or right, up or down, etc. Our experience is not limited to the immediate vicinity (although of course our experience is spatially limited to the extent that sufficiently distant objects are invisible to us). But, although we perceive the past, we do not perceive it as past, but as present. Moreover, our experience does not only appear to be temporally limited, it is so: we do not perceive the future, and we do not continue to perceive transient events long after information from them reached our senses. Now, there is a very simple answer to the question why we do not perceive the future, and it is a causal one. Briefly, causes always precede their effects; perception is a causal process, in that to perceive something is to be causally affected by it; therefore we can only perceive earlier events, never later ones. So one temporal boundary of our experience is explained; what of the other?

There seems no *logical* reason why we should not directly experience the distant past. We could appeal to the principle that there can be no action at a temporal distance, so that something distantly past can only causally affect us via more proximate events. But this is inadequate justification. We can only perceive a spatially distant tree by virtue of its effects on items in our vicinity (light reflected off the tree impinging on our retinas), but this is not seen by those who espouse a direct realist theory of perception as incompatible with their position. We still see the *tree*, they say, not some more immediate object. Perhaps then we should look for a different strategy, such as the following one, which appeals to biological considerations. To be effective agents in the world, we must represent accurately what is currently going on: to be constantly out of date in our beliefs while going about our activities would be to face pretty immediate extinction. Now we are fortunate in that, although we only perceive the past it is, in most cases, the very recent past, since the transmission of light and sound, though finite, is extremely rapid. Moreover, although things change, they do so, again in most cases, at a rate that is vastly slower than the rate at which information from external objects travels to us. So when we form beliefs about what is going on in the world, they are largely accurate ones. (See Butterfield (1984) for a more detailed account along these lines.) But, incoming information having been registered, it needs to move into the memory to make way for more up to date information. For, although things may change slowly relative to the speed of light or of sound, they do change, and we cannot afford to be simultaneously processing conflicting information. So our effectiveness as agents depends on our not continuing to experience a transient state of affairs (rather in the manner of a slow motion film) once information from it has been absorbed. Evolution has ensured that we do not experience anything other than the very recent past (except when we are looking at the heavens).

To perceive something as present is simply to perceive it: we do not need to postulate some extra item in our experience that is ‘the experience of presentness.’ It follows that there can be no ‘perception of pastness’. In addition, if pastness were something we could perceive, then we would perceive *everything* in this way, since every event is past by the time we perceive it. But even if we never perceive anything as past (at the same time as perceiving the event in question) we could intelligibly talk more widely of the experience of pastness: the experience we get when something comes to an end. And it has been suggested that memories-more specifically, *episodic memories*, those of our experiences of past events-are accompanied by a feeling of pastness (see Russell (1921)). The problem that this suggestion is supposed to solve is that an episodic memory is simply a memory of an event: it *represents* the event simpliciter, rather than the fact that the event is past. So we need to postulate something else which alerts us to the fact that the event remembered is past. An alternative account, and one which does not appeal to

any phenomenological aspects of memory, is that memories dispose us to form past-tensed beliefs, and is by virtue of this that they represent an event as past.

We have, then, a candidate explanation for our experience of being located at a particular moment in time, the (specious) present. And as the content of that experience is constantly changing, so that position in time shifts. But there is still a further puzzle. Change in our experience is not the same thing as experience of change. We want to know, not just what it is to perceive one event after another, but also what it is to perceive an event as occurring after another. Only then will we understand our experience of the passage of time. We turn, then, to the perception of time order.

Time order

How do we perceive precedence amongst events? A temptingly simple answer is that the perception of precedence is just a sensation caused by instances of precedence, just as a sensation of red is caused by instances of redness. Hugh Mellor (1998), who considers this line, rejects it for the following reason. If this were the correct explanation, then we could not distinguish between *x* being *earlier* than *y*, and *x* being *later* than *y*, for whenever there is an instance of one relation, there is also an instance of the other. But plainly we are able to distinguish the two cases, so it cannot simply be a matter of perceiving a relation, but something to do with our perception of the relata. But mere perception of the relata cannot be all there is to perceiving precedence. Consider again Broad's point about the second hand and the hour hand. We first perceive the hour hand in one position, say pointing to 3 o'clock, and later we perceive it in a different position, pointing to half-past 3. So I have two perceptions, one later than the other. I may also be aware of the temporal relationship of the two positions of the hand. Nevertheless, I do not perceive that relationship, in that I do not see the hand moving. In contrast, I do see the second hand move from one position to another: I see the successive positions *as* successive.

Mellor's proposal is that I perceive *x* precede *y* by virtue of the fact that my perception of *x* causally affects my perception of *y*. As I see the second hand in one position, I have in my short-term memory an image (or information in some form) of its immediately previous position, and this image affects my current perception. The result is a perception of movement. The perceived order of different positions need not necessarily be the same as the actual temporal order of those positions, but it will be the same as the causal order of the *perceptions* of them. Since causes always precede their effects, the temporal order perceived entails a corresponding temporal order in the perceptions.

In effect, Mellor's idea is that the brain represents time by means of time: that temporally ordered events are represented by similarly temporally ordered experiences. This would make the representation of time unique. (For example, the brain does not represent spatially separated objects by means of spatially separated perceptions, or orange things by orange perceptions.) But why should time be unique in this respect? In other media, time can be represented spatially (as in cartoons, graphs, and analogue clocks) or numerically (as in calendars and digital clocks). So perhaps the brain can represent time by other means. One reason to suppose that it must have other means at its disposal is that time needs to be represented in *memory* (I recall, both that a was earlier than b, and also the experience of seeing a occur before b) and

intention (I intend to F after I G), but there is no obvious way in which Mellor's ‘representation of time by time’ account can be extended to these.

On Mellor's model, the mechanism by which time-order is perceived is sensitive to the *time* at which perceptions occur, but indifferent to their *content* (what the perceptions are of). Daniel Dennett (1991) proposes a different model, on which the process is time-independent, but content-sensitive. For example, the brain may infer the temporal order of events by seeing which sequence makes sense of the causal order of those events. One of the advantages of Dennett's model is that it can account for the rather puzzling cases of ‘backwards time referral’, where perceived order does not follow the order of perceptions. (See Dennett (1991) for a discussion of these cases, and also Roache (1999) for an attempt to reconcile them with Mellor's account.)

The metaphysics of time perception

In giving an account of the various aspects of time perception, we inevitably make use of concepts that we take to have an objective counterpart in the world: the past, temporal order, causation, change, the passage of time and so on. But one of the most important lessons of philosophy, for many writers, is that there may be a gap, perhaps even a gulf, between our representation of the world and the world itself, even on a quite abstract level. (It would be fair to add that, for other writers, this is precisely *not* the lesson philosophy teaches.) Philosophy of time is no exception to this. Indeed, it is interesting to note how many philosophers have taken the view that, despite appearances, time, or some aspect of time, is unreal. In this final section, we will take a look at how two metaphysical debates concerning the nature of the world interact with account of time perception.

The first debate concerns the reality of tense, that is, our division of time into past, present and future. Is time really divided in this way? Does what is present slip further and further into the past? Or does this picture merely reflect our perspective on a reality in which there is no uniquely privileged moment, the present, but simply an ordered series of moments? *Tensed theorists* say that our ordinary picture of the world as tensed reflects the world as it really is: the passage of time is an objective fact. *Tenseless theorists* deny this. For them, the only objective temporal facts concern relations of precedence and simultaneity between events. (I ignore here the complications introduced by the Special Theory of Relativity, since tenseless theory-and perhaps tensed theory also-can be reformulated in terms which are compatible with the Special Theory.) Tenseless theorists do not deny that our tensed beliefs, such that that a cold front is *now* passing, or that Sally's wedding *was two years ago*, may be true, but they assert that what makes such beliefs true are not facts about the pastness, presentness or futurity of events, but tenseless facts concerning precedence and simultaneity (see Mellor (1998), Oaklander and Smith (1994)). On one version of the tenseless theory, for example, my belief that there is a cold front now passing is true because the passing of the front is *simultaneous with* my forming the belief. Now one very serious challenge to the tenseless theorist is to explain why, if time does not pass in reality, it appears to do so. What, in tenseless terms, is the basis for our experience as-of the passage of time?

The accounts we considered above, first of the temporal restrictions on our experience, and secondly of

our experience of time order, did not explicitly appeal to tensed notions. The facts we did appeal to look like purely tenseless ones: that causes are always earlier than their effects, that things typically change slowly in relation to the speed of transmission of light and sound, that our information-processing capacities are limited, and that there can be causal connections between memories and experiences. So it may be that the tenseless theorist can discharge the obligation to explain why time seems to pass. But two doubts remain. First, perhaps the tensed theorist can produce a simpler explanation of our experience. Second, it may turn out that supposedly tenseless facts are dependent upon tensed ones, so that, for example, *a* and *b* are simultaneous by virtue of the fact that both are *present*.

The second metaphysical issue that has a crucial bearing on time perception concerns causal asymmetry. The account of our sense of being located at a time which we considered under Past, present and the passage of time rested on the assumption that causation is asymmetric. Later events, it was suggested, cannot affect earlier ones, as a matter of mind-independent fact, and this is why we do not perceive the future, only the past. But attempts to explain the basis of causal asymmetry, in terms for example of counterfactual dependence, or in probabilistic terms, are notoriously problematic. One moral we might draw from the difficulties of reducing causal asymmetry to other asymmetries is that causal asymmetry is primitive, and so irreducible. Another is that that the search for a mind-independent account is mistaken. Perhaps causation in intrinsically symmetric, but some feature of our psychological constitution and relation to the world makes causation appear asymmetric. This *causal perspectivalism* is the line taken by Huw Price (1996). That causal asymmetry should be explained in part by our psychological constitution, in a way analogous to our understanding of secondary qualities such as colour, is a radical reversal of our ordinary assumptions, but then our ordinary understanding of a number of apparently objective features of the world—tense, absolute simultaneity—have met with similarly radical challenges. Now, if causal asymmetry is mind-dependent in this way, then we cannot appeal to it in accounting for our experience of temporal asymmetry—the difference between past and future.

But the facts of perception may themselves constitute a problem for perspectivalism over causal asymmetry. We will leave the topic of time perception with the following conundrum for proponents of causal perspectivalism. Consider the following causally ordered (but not directed) series:

$\Phi - \beta - \kappa$

Assuming, as perspectivalism holds, that causation is intrinsically symmetric, β stands in exactly the same causal relation to Φ as it does to κ. However, although not directed, the series is ordered in that the relation of causal betweenness holds between items. Thus β is causally between Φ and κ. But then, if this is so, it is not clear how perspectivalism could explain why the following principle holds:

If β is a perceptual experience, then it cannot have both Φ and κ as its object

This principle does not beg the question against perspectivalism by smuggling in an assumption about causal asymmetry. For it is surely a trivial fact about our perception of time that if A is experienced as occuring before B, A and B cannot be experienced as simultaneous. And it is surely an objective

(although non-trivial) fact that our experience of A will be causally between A and our experience of B. Now if perspectivalism cannot answer the challenge to explain the truth of the above principle, it seems that our experience of temporal asymmetry, insofar as it has a causal explanation, requires causation to be objectively asymmetric.

One strategy the causal perspectivalist could adopt (indeed, the only one available) is to explain the asymmetric principle above in terms of some objective non-causal asymmetry. Price, for example, allows an objective thermodynamic asymmetry, in that an ordered series of states of the universe will exhibit what he calls a thermodynamic gradient: entropy will be lower at one end of the series than at the end. We should resist the temptation to say that entropy increases, for that would be like asserting that a road goes uphill rather than downhill without conceding the perspectival nature of descriptions like ‘uphill’. Could such a thermodynamic asymmetry explain why perception points in one direction? That is a thought for the reader to ponder.

Bibliography

- Augustine, St. (1961) *Confessions*, ed. R.S. Pinecoffin, Harmondsworth: Penguin
- Broad, C.D. (1923) *Scientific Thought*, London:
- Butterfield, Jeremy (1984) ‘Seeing the Present’, *Mind* 93, 161-76; reprinted with corrections in R. Le Poidevin (ed.) *Questions of Time and Tense*, Oxford: Clarendon Press, 61-75
- Campbell, John (1994) *Past, Space and Self*, Cambridge, Mass.: MIT Press
- Dennett, Daniel (1991) *Consciousness Explained*, London: Allen Lane
- Fotheringham, Heather (1999) ‘How Long is the Present?’, *Stoa* 1, No. 2, 56-65
- Friedman, William J. (1990) *About Time: Inventing the Fourth Dimension*, Cambridge, Mass.: MIT Press
- Gombrich, Ernst (1964) ‘Moment and Movement in Art’, *Journal of the Warburg and Courtauld Institutes* XXVII, 293-306
- Hestevold, H. Scott (1990) ‘Passage and the Presence of Experience’, *Philosophy and Phenomenological Research* 50, 537-52; reprinted in Oaklander and Smith (1994), 328-43
- Hirsh, I.J. and Sherrick, J.E. (1961) ‘Perceived Order in Different Sense Modalities’, *Journal of Experimental Psychology* 62, 423-32
- Hoerl, Christoph (1998) ‘The Perception of Time and the Notion of a Point of View’, *European Journal of Philosophy* 6, 156-71
- James, William (1890) *The Principles of Psychology*, New York: Henry Holt
- Le Poidevin, Robin (1997) ‘Time and the Static Image’, *Philosophy* 72, 175-88
- Le Poidevin, Robin (1999) ‘Egocentric and Objective Time’, *Proceedings of the Aristotelian Society* XCIX, 19-36
- Mabbott, J.D. (1951) ‘Our Direct Experience of Time’, *Mind* 60, 153-67
- Mackintosh, N.J. (1983) *Conditioning and Associative Learning*, Oxford: Clarendon Press
- Mayo, Bernard (1950) ‘Is There a Sense of Duration?’, *Mind* 59, 71-8
- Mellor, D.H. (1998) *Real Time II*, London: Routledge
- Mundle, C.W.K. (1966) ‘Augustine's Pervasive Error Concerning Time’, *Philosophy*41, 165-8
- Myers, Gerald (1971) ‘James on Time Perception’, *Philosophy of Science* 38, 353-60

- Oaklander, L. Nathan (1993) ‘On the Experience of Tenseless Time’, *Journal of Philosophical Research* 18, 159-66; reprinted in Oaklander and Smith (1994), 344-50
- Oaklander, L. Nathan, and Smith, Quentin (1994), eds., *The New Theory of Time*, New Haven: Yale University Press
- Odegard, D. (1978) ‘Phenomenal Time’, *Ratio* 20, 116-22
- Ornstein, R.E. (1969) *On the Experience of Time*, Harmondsworth: Penguin
- Plumer, Gilbert (1985) ‘The Myth of the Specious Present’, *Mind* 94
- Plumer, Gilbert (1987) ‘Detecting Temporalities’, *Philosophy and Phenomenological Research* 47, 451-60
- Pöppel, Ernst (1978) ‘Time Perception’, in Richard Held et al., eds., *Handbook of Sensory Physiology*, Vol. VIII: Perception, Berlin: Springer-Verlag
- Price, Huw (1996) *Time's Arrow and Archimedes' Point: New Directions in the Physics of Time*, Oxford: Oxford University Press
- Roache, Rebecca (1999) ‘Mellor and Dennett on the Perception of Temporal Order’, *Philosophical Quarterly* 49, 231-38
- Russell, Bertrand (1915) ‘On the Experience of Time’, *Monist* 25, 212-33
- Russell, Bertrand (1921) *The Analysis of Mind*, London: George Allen and Unwin
- Smith, Quentin (1988) ‘The Phenomenology of A-Time’, *Diálogos* 52, 143-53; reprinted in Oaklander and Smith (1994), 351-9
- Walsh, W.H. (1967) ‘Kant on the Perception of Time’, *Monist* 51, 376-96
- Williams, Clifford (1992) ‘The Phenomenology of B-Time’, *Southern Journal of Philosophy* 30, 123-37; reprinted in Oaklander and Smith (1994), 360-72

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Augustine, Saint | memory | perception | space and time: being and becoming in modern physics

Robin Le Poidevin
r.d.lepoidevin@leeds.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 28, 2000
Content last modified: August 28, 2000

Stanford Encyclopedia of Philosophy

The Hole Argument

What is space? What is time? Do they exist independently of the things and processes in them? Or is their existence parasitic on these things and processes? Are they like a canvas onto which an artist paints; they exist whether or not the artist paints on them? Or are they akin to parenthood; there is no parenthood until there are parents and children? That is, is there no space and time until there are things with spatial properties and processes with temporal durations?

These questions have long been debated and continue to be debated. The hole argument arose when these questions were asked in the context of modern spacetime physics. In that context, space and time are fused into a single entity, spacetime, and we inquire into its status. One view is that spacetime is a substance, a thing that exists independently of the processes occurring within spacetime. This is spacetime substantivalism. The hole argument seeks to show that this viewpoint leads to unpalatable conclusions in a large class of spacetime theories. Spacetime substantivalism requires that we ascribe such a surfeit of properties to spacetime that neither observation nor even the laws of the relevant spacetime theory itself can determine which are the correct ones. Such abundance is neither logically contradictory nor refuted by experience. But there must be some bounds on how rich a repertoire of hidden properties can be ascribed to spacetime. The hole argument urges that spacetime substantivalism goes beyond those bounds.

The hole argument was invented for slightly different purposes by Albert Einstein late in 1913 as part of his quest for the general theory of relativity. It was revived and reformulated in the modern context by John^3 = John Earman x John Stachel x John Norton.

1. Modern Spacetime Theories: A Beginner's Guide

Virtually all modern spacetime theories are now built in the same way. The theory posits a manifold of events and then assigns further structures to those events to represent the content of spacetime. A standard example is Einstein's general theory of relativity. As a host for the hole argument, we will pursue one of its best known applications, the expanding universes of modern relativistic cosmology.[1]

Manifold of Events. Consider our universe, which relativistic cosmologies attempt to model. Events in the universe correspond to the dimensionless points of familiar spatial geometry. Just as a geometric point is a particular spot in a geometrical space, an event is a particular point in a cosmological space at a particular time. To be a four-dimensional manifold, the events must be a little bit more organized than they are if they merely form a set. That extra organization comes from the requirement that we can smoothly label the events with four numbers--or at least we can do this for any sufficiently small chunk of the manifold. These labels form coordinate systems. Figure 1 shows how a set of events may be made into a two dimensional manifold.

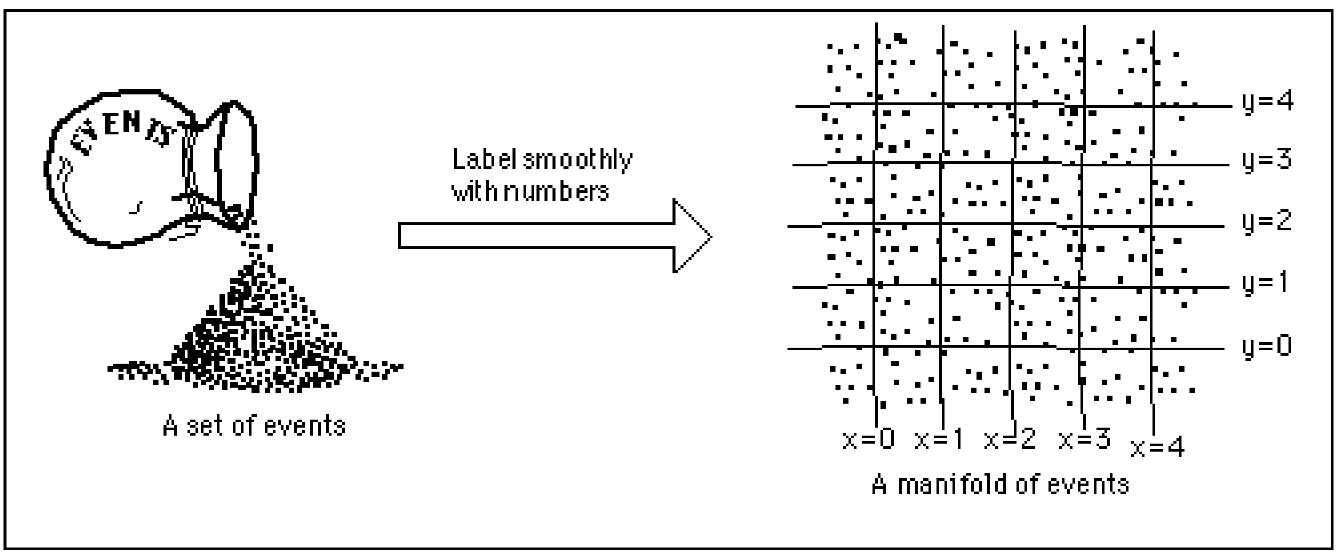


Figure 1. Forming a manifold of events

Metrical Structure and Matter Fields. In specifying that events form a four dimensional manifold, we have still not specified which events lie in the future and past of which other events, how much time elapses between these events, which events are simultaneous with others so that they can form three dimensional spaces, what spatial distances separates these events and many more related properties. These additional properties are introduced by specifying the metric field.

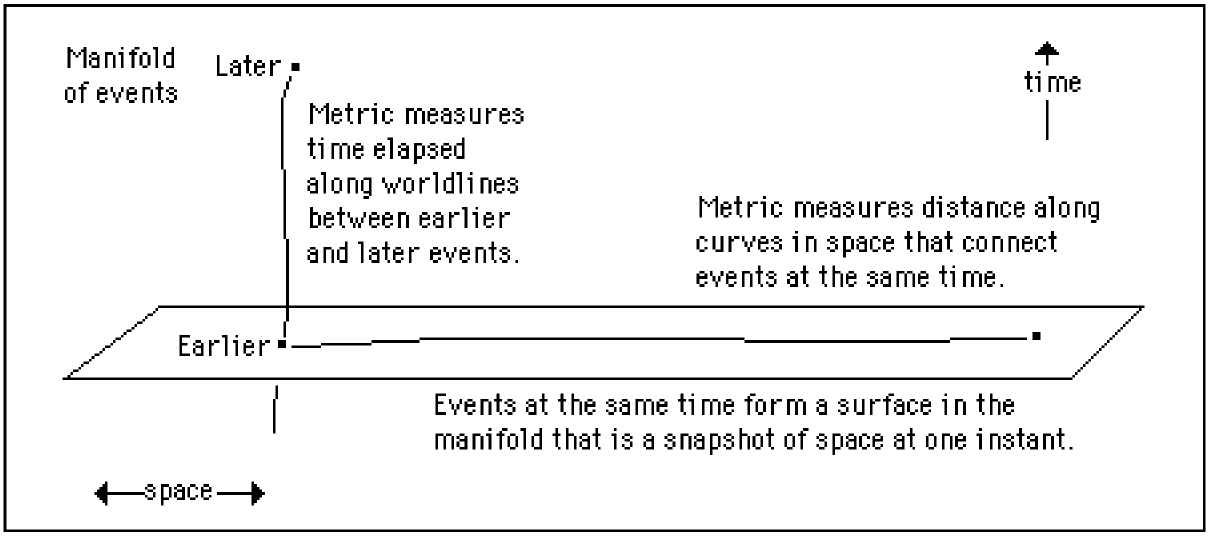


Figure 2. The function of the metric field.

Imagine all the curves in some spacetime and an exhaustive catalog of all the distances in space and duration elapsed in time between any two events on any curve that connects them. The information carried by the metric is just that catalog, reduced to a very much more compact form.

The matter of the universe is represented by matter fields. The simplest form of matter--the big lumps that make galaxies--can be represented by worldlines that trace out the history of each galaxy through time. In standard models, the galaxies recede from one another and this is represented by a spreading apart of the galactic worldlines as we proceed to later times.

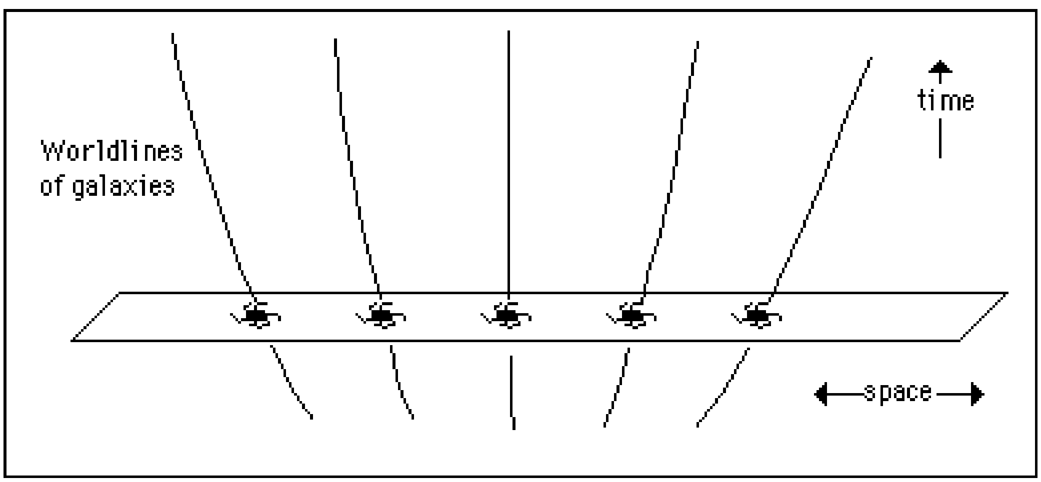


Figure 3. Galaxies in an expanding universe.

2. The Freedom of General Covariance

When Einstein first introduced his general theory of relativity in the 1910s, its novel feature was its general covariance. It was the first spacetime theory in which one was free to use arbitrary spacetime coordinate systems. This feature is now shared by virtually all modern formulations of spacetime

theories, including modern versions of special relativity and Newtonian spacetime theory. In the modern context, the freedom of general covariance is expressed in a more vividly geometric manner.[2] When we assemble a relativistic universe, we spread the metrical structure and matter fields over a manifold:

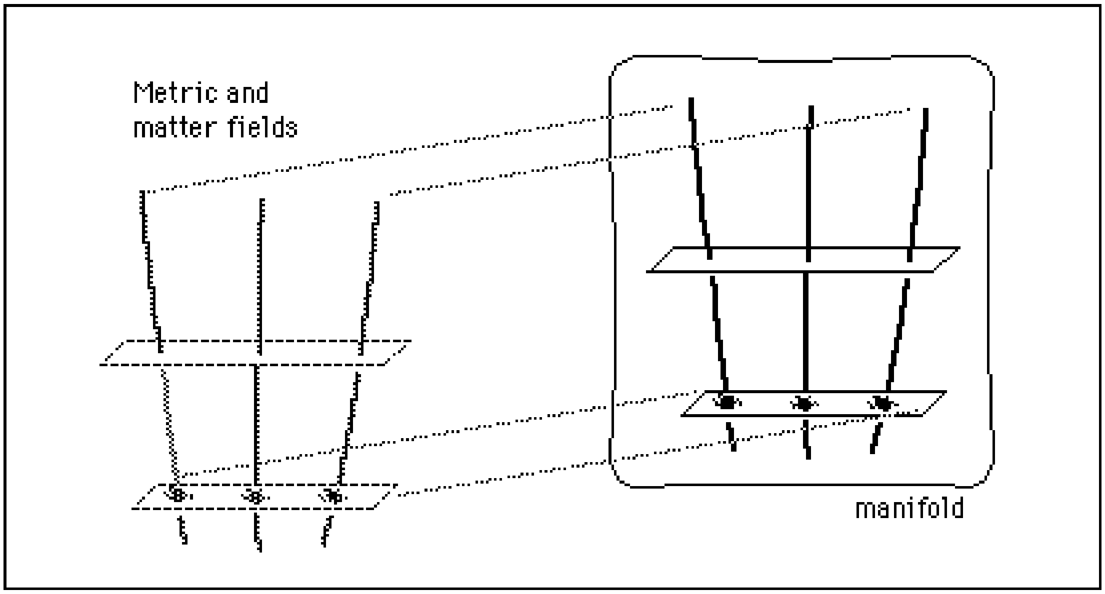


Figure 4. One way to spread metric and matter over the manifold.

What general covariance does is to license different spreadings of metric and matter fields. If the theory allows one spreading, such as that given in Figure 4, then through general covariance it allows any other that represents a smooth deformation of the original, such as shown in Figure 5:

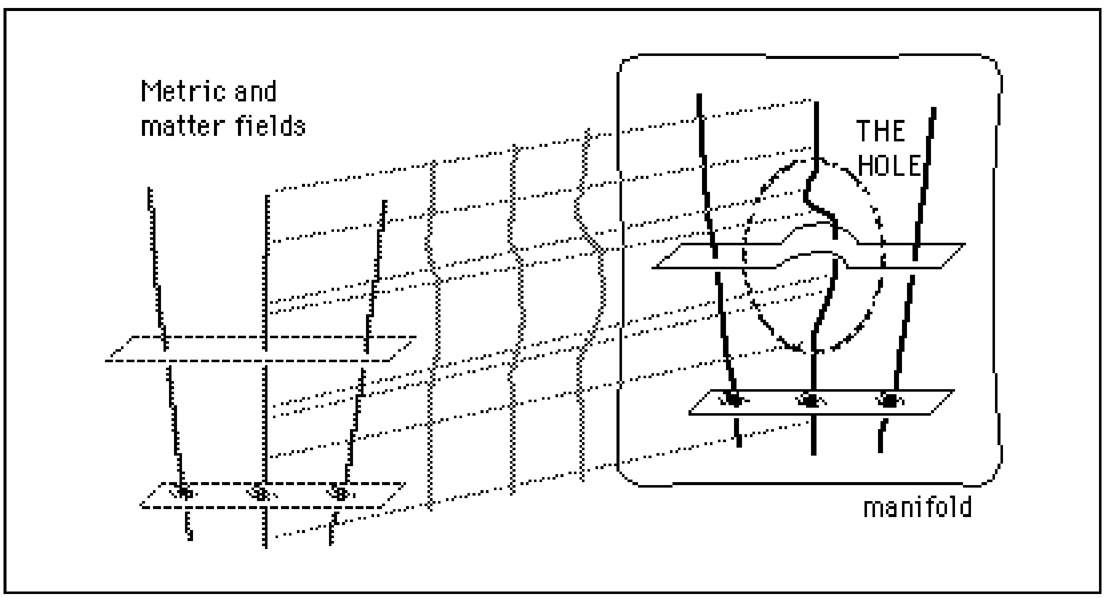


Figure 5. Another way to spread metric and matter over the manifold.

Figures 4 and 5 illustrate a hole transformation on the metric and matter fields. The dotted region is The Hole. The first distribution of metric and matter fields is transformed into the second in a special way. The transformation leaves the fields unchanged outside the hole; within it they have been spread

differently on the manifold; the spreadings inside and outside the hole join smoothly.[3]

3. The Preservation of Invariants

The two different spreadings share one vital characteristic upon which the hole argument depends: the two spreadings agree completely on all invariant properties. These invariant properties are, loosely speaking, the ones that are intrinsic to the geometry and dynamics, such as distance along spatial curves and time along worldlines of galaxies, the rest mass of the galaxy, the number of particles in it, as well as a host of other properties, such as whether the spacetimes are metrically flat or curved. While the fields are spread differently in the two cases, in invariant terms, they are the same.

This last result actually explains the prevalence of general covariance. The laws of a spacetime theory are typically stated as relations between invariant properties. Therefore if they are satisfied by one spacetime, they must also be satisfied by a transform of that spacetime that shares all the original's invariant geometric properties.

All observables can be reduced to invariants. For example, if one makes a journey from one galaxy to another, all observables pertinent to the trip will be invariants. These include the time elapsed along the journey, whether the spaceship is accelerating or not at any time in its journey, the age of the galaxy one leaves at the start of the trip and the age of the destination galaxy at the end and all operations that may involve signaling with particles or light pulses.

Therefore, since the two spreadings or distributions agree on invariants, they also agree on all observables and are observationally indistinguishable.

4. What Represents Spacetime? Manifold Substantivalism

We want to know whether we can conceive of spacetime as a substance, that is, as something that exists independently. To do this, we need to know what in the above structures represents spacetime. One popular answer to that question is that the manifold of events represents spacetime. We shall see shortly that this popular form of the answer is the one that figures in the hole argument. This choice is natural since modern spacetime theories are built up by first positing a manifold of events and then defining further structures on them. So the manifold plays the role of a container just as we expect spacetime does.[4]

The notion that the manifold represents an independently existing thing is quite natural in the realist view of physical theories. In that view one tries to construe physical theories literally. If formulated as above, a spacetime is a manifold of events with certain fields defined on the manifold. The literal reading is that this manifold is an independently existing structure that bears properties.

Appealing as manifold substantivalism once was, after one sees what trouble it causes, it becomes tempting to insist that the manifold of events lacks properties essential to spacetime. For example, there is no notion of past and future, of time elapsed or of spatial distance in the manifold of events. Thus one might be tempted to identify spacetime with the manifold of events plus some further structure that supplies these spatiotemporal notions. In relativistic cosmologies, that further structure would be the metrical structure. Readers who pursue this issue further will find that this escape from the hole argument sometimes succeeds and sometimes fails. In certain important special cases, alternative versions of the hole argument can be mounted against the manifold-plus-further-structure substantivalists. (See Norton 1988.)

In general relativity, there is a further problem with arguing that the metric field properly belongs to spacetime since it carries essential spatiotemporal properties. The metric field of general relativity also carries energy and momentum--the energy and momentum of the gravitational field. This energy and momentum is freely interchanged with other matter fields in spacetimes. It is the source of the huge quantities of energy released as radiation and heat in stellar collapse, for example. To carry energy and momentum is a natural distinguishing characteristic of matter contained within spacetime. So the metric field of general relativity seems to defy easy characterization as exclusively part of spacetime the container or matter the contained.

5. The Price of Spacetime Substantivalism

So far we have characterized the substantivalist doctrine as the view that spacetime has an existence independent of its contents. This formulation conjures up powerful if vague intuitive pictures, but it is not clear enough for interpretation in the context of physical theories. If we represent spacetime by a manifold of events, how do we characterize the independence of its existence? Is it the counterfactual claim that were there no metric or matter fields, there would still be a manifold of events? That counterfactual is automatically denied by the standard formulation which posits that all spacetimes have at least metrical structure. That seems too cheap a refutation of manifold substantivalism. Surely, there must be an improved formulation. Fortunately, we do not need to wrestle with finding it. For present purposes we need only consider a consequence of the substantivalist view and can set aside the task of giving a precise formulation of the substantivalist view.

In their celebrated debate over space and time, Leibniz taunted the substantivalist Newton's representative, Clarke, by asking how the world would change if East and West were switched. For Leibniz there would be no change since all spatial relations between bodies would be preserved by such a switch. But the Newtonian substantivalist had to concede that the bodies of the world were now located in different spatial positions, so the two systems were physically distinct.

Correspondingly, when we spread the metric and matter fields differently over a manifold of events, we are now assigning metrical and material properties in different ways to the events of the manifold. For example, imagine that a galaxy passes through some event E in the hole. After the hole transformation,

this galaxy might not pass through that event. For the manifold substantivalist, this must be a matter of objective physical fact: either the galaxy passes through E or not. The two distributions represent two physically distinct possibilities.

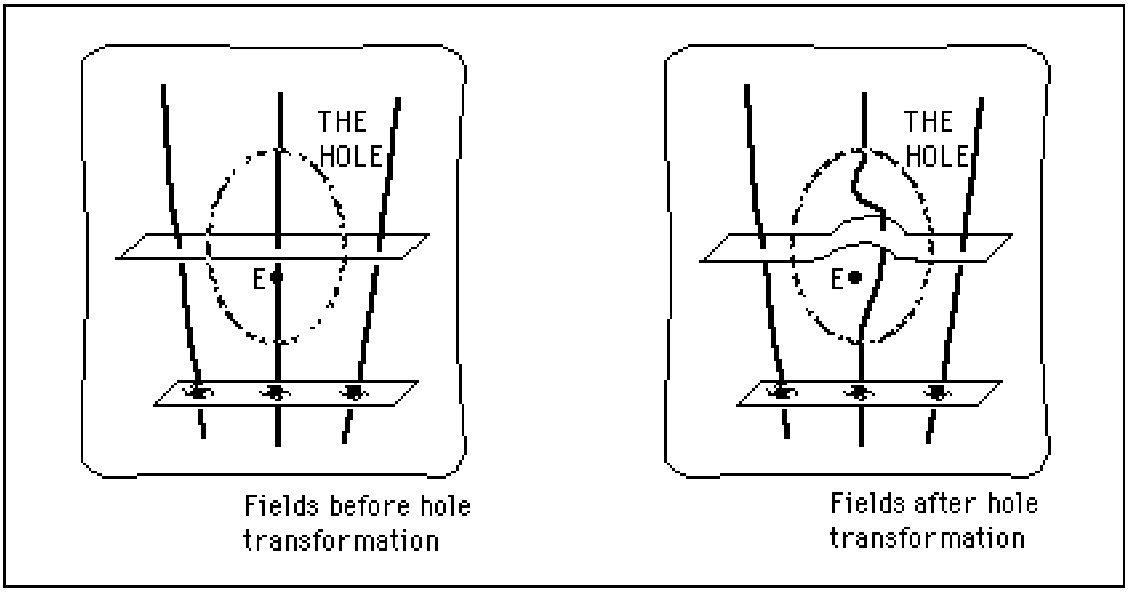


Figure 6. Does the galaxy pass through event E?

That is, manifold substantivalists must deny an equivalence inspired by Leibniz' taunt and is thus named after him:[5]

> **Leibniz Equivalence**. If two distributions of fields are related by a smooth transformation, then they represent the same physical systems.

6. Unhappy Consequences

We can now assemble the pieces above to generate unhappy consequences for the manifold substantivalist. Consider the two distributions of metric and matter fields related by a hole transformation. Since the manifold substantivalist denies Leibniz equivalence, the substantivalist must hold that the two systems represent distinct physical systems. But the properties that distinguish the two are very elusive.

We have already noticed that the two are observationally equivalent. So the substantivalist must insist that it makes a physical difference as to whether the galaxy passes through event E or not. But no observation can tell us if we are in a world in which the galaxy passes through event E or misses event E, for universes with either are observationally equivalent.

Worse, the physical theory of relativistic cosmology is unable to pick between the two cases. This is manifested as an indeterminism of the theory. We can specify the distribution of metric and matter fields throughout the manifold of events, excepting within the region designated as The Hole. Then the theory

is unable to tell us how the fields will develop into The Hole. Both the original and the transformed distribution are legitimate extensions of the metric and matter fields outside The Hole into The Hole, since each satisfies all the laws of the theory of relativistic cosmology. The theory has no resources which allow us to insist that one only is admissible.

It is important to see that the unhappy consequence does not consist merely of a failure of determinism. We are all too familiar with such failures and it is certainly not automatic grounds for dismissal of a physical theory. The best known instance of a widely celebrated, indeterministic theory is quantum theory, where, in the standard interpretation, the measurement of a system can lead to an indeterministic collapse onto one of many possible outcomes. Less well known is that it is possible to devise indeterministic systems in classical physics as well.

The problem with the failure of determinism in the hole argument is not the fact of failure but the way that it fails. If we deny manifold substantivalism and accept Leibniz equivalence, then the indeterminism induced by a hole transformation is eradicated. While there are uncountably many mathematically distinct developments of the fields into the hole, under Leibniz Equivalence, they are all physically the same. That is, there is a unique development of the physical fields into the hole after all. Thus the indeterminism is a direct product of the substantivalist viewpoint. Similarly, if we accept Leibniz equivalence, then we are no longer troubled that the two distributions cannot be distinguished by any possible observation. They are merely different mathematical descriptions of the same physical reality and so should agree on all observables.

We can load up any physical theory with superfluous, phantom properties that cannot be fixed by observation. If their invisibility to observation is not sufficient warning that these properties are illegitimate, finding that they visit indeterminism onto a theory that is otherwise deterministic in this set-up ought to be warning enough. These properties are invisible to both observation and theory; they should be discarded along with any doctrine that requires their retention.

7. The Hole Argument in Brief

In sum the hole argument amounts to this:[6]

> **1.** If one has two distributions of metric and matter fields related by a hole transformation, manifold substantivalists must maintain that the two systems represent two distinct physical systems.
>
> **2.** This physical distinctness transcends both observation and the determining power of the theory since:
> -- The two distributions are observationally identical.
> -- The laws of the theory cannot pick between the two developments of the fields into the hole.

3. Therefore the manifold substantivalist advocates an unwarranted bloating of our physical ontology and the doctrine should be discarded.

8. The History of the Hole Argument

8.1 Einstein Falls into the Hole...

The hole argument was created by Albert Einstein late in 1913 as an act of desperation when his quest for his general theory of relativity had encountered what appeared to be insuperable obstacles. Over the previous year, he had been determined to find a gravitation theory that was generally covariant, that is, whose equations were unchanged by arbitrary transformation of the spacetime coordinates. He had even considered essentially the celebrated, generally covariant equations he would settle upon in November 1915 and which now appear in all the text books.

Unfortunately Einstein had been unable to see that these equations were admissible. Newton's theory of gravitation worked virtually perfectly for weak gravitational fields. So it was essential that Einstein's theory revert to Newton's in that case. But try as he might, Einstein could not see that his equations and many variants of them could properly mesh with Newton's theory. In mid 1913 he published a compromise: a sketch of a relativistic theory of gravitation that was not generally covariant. (For further details of these struggles, see Norton (1984).)

His failure to find an admissible generally covariant theory troubled Einstein greatly. Later in 1913 he sought to transform his failure into a victory of sorts: he thought he could show that no generally covariant theory at all is admissible. Any such theory would violate what he called the Law of Causality--we would now call it determinism. He sought to demonstrate this remarkable claim with the hole argument.

In its original incarnation, Einstein considered a spacetime filled with matter excepting one region, the hole, which was matter free. (So in this original form, the term "hole" makes more sense than in the modern version.) He then asked if a full specification of both metric and matter fields outside the hole would fix the metric field within. Since he had tacitly eschewed Leibniz Equivalence, Einstein thought that the resulting negative answer sufficient to damn all generally covariant theories.

8.2 ...and Climbs out Again

Einstein struggled on for two years with his misshapen theory of limited covariance. Late in 1915, as evidence of his errors mounted inexorably, Einstein was driven to near despair and ultimately capitulation. He returned to the search for generally covariant equations with a new urgency, fueled in part by the knowledge that none other than David Hilbert had thrown himself into analysis of his theory. Einstein's quest came to a happy close in late November 1915 with the completion of his theory in

generally covariant form.

For a long time it was thought that Hilbert had beaten Einstein by 5 days to the final theory. New evidence in the form of the proof pages of Hilbert's paper now suggests he may not have. More important, it shows clearly that Hilbert, like Einstein, at least temporarily believed that the hole argument precluded all generally covariant theories and that the belief survived at least as far as the proof pages of his paper. (See Corry, Renn and Stachel 1997.)

While Einstein had tacitly withdrawn his objections to generally covariant theories, he had not made public where he thought the hole argument failed. This he finally did when he published what John Stachel calls the "point-coincidence argument." This argument, well known from Einstein's (1916, p.117) review of his general theory of relativity, amounts to a defense of Leibniz equivalence. He urges that the physical content of a theory is exhausted by the catalog of the spacetime coincidences it licenses. For example, in a theory that treats particles only, the coincidences are the points of intersection of the particle worldlines. These coincidences are preserved by transformations of the fields. Therefore two systems of fields that can be intertransformed have the same physical content; they represent the same physical system.

Over the years, the hole argument was deemed to be a trivial error by an otherwise insightful Einstein. It was John Stachel (1980) who recognized its highly non-trivial character and brought this realization to the modern community of historians and philosophers of physics. (See also Stachel, 1986.) In Earman and Norton (1987), the argument was recast as one that explicitly targets spacetime substantivalism. For further historical discussion, see Howard and Norton (1993), Janssen (forthcoming), Klein (1995) and Norton (1987).

9. Responses to the Hole Argument

There are at least as many responses to the hole argument as authors who have written on it. One line of thought simply agrees that the hole argument makes acceptance of Leibniz equivalence compelling. It seeks to make more transparent what that acceptance involves by trying to find a single mathematical structure that represents a physical spacetime system rather than the equivalence class of intertransformable structures licensed by Leibniz equivalence. One such attempt involves the notion of a "Leibniz algebra." (See Earman, 1989, Ch. 9, Sect. 9) It has become unclear that such attempts can succeed. Just as intertransformable fields represent the same physical system, there are distinct but intertransformable Leibniz algebras with the same physical import. If the formalisms of manifolds and of Leibniz algebras are intertranslatable, one would expect the hole argument to reappear in the latter formalism as well under this translation. (See Rynasiewicz, 1992.)

Einstein's original hole argument was formulated in the context of general relativity. The hole argument as formulated in Earman and Norton (1987) applies to all local spacetime theories and that includes generally covariant formulations of virtually all known spacetime theories. One view is that this goes too far, that general relativity is distinct from many other spacetime theories in that its spacetime geometry

has become dynamical and it is only in such theories that the hole argument should be mounted. (See Earman, 1989, Ch.9, Section 5; Stachel, 1993)

To critics, the hole argument presents a huge target. It consists of a series of assumptions all of which are needed to make good on its conclusion. The argument can be blocked by denying just one of its presumptions. Different authors have sought to sustain denial of virtually every one of them.

Perhaps the most promising of these attacks is one that requires the least modification of the ideas used to mount the hole argument. It is the proposal that spacetime is better represented not by the manifold of events alone but by some richer structure, such as the manifold of events in conjunction with metrical properties. (See, for example, Hoefer, 1996.) A slight and very popular variant allows that each event of the manifold represents a physical spacetime event, but which physical event that might be depends on the spreading of metric and matter fields on the manifold. Thus the indeterminism of the hole transformation can be eradicated since the metric and matter properties of an event can be carried with the transformation. (See, for example, Brighouse, 1994.)

More generally, we may well wonder whether the problems faced by spacetime substantivalism is an artifact of the particular formalism described above. Bain (1998) has explored the effect of a transition to other formalisms.

The simplest challenge notes that Leibniz equivalence is a standard presumption in the modern mathematical physics literature and suggests that even entertaining its denial (as manifold substantivalists must) is some kind of mathematical blunder unworthy of serious attention. While acceptance of Leibniz equivalence is widespread in the physics literature, it is not a logical truth that can only be denied on pain of contradiction. That it embodies non-trivial assumptions whose import must be accepted with sober reflection is indicated by the early acceptance of the hole argument by David Hilbert. If denial of Leibniz equivalence is a blunder so egregious that no competent mathematician would do it, then our standards for competence have become unattainably high, for they must exclude David Hilbert in 1915 at the height of his powers.

Another challenge seeks principled reasons for denying general covariance. One approach tries to establish that a spacetime can be properly represented by at most one of two intertransformable systems of fields on some manifold. So Maudlin (1990) urges that each spacetime event carries its metrical properties essentially, that is, it would not be that very event if (after redistribution of the fields) we tried to assign different metrical properties to it. Butterfield (1989) portrays intertransformable systems as different possible worlds and uses counterpart theory to argue that at most one can represent an actual spacetime.

These responses are just a few of a large range of responses of increasing ingenuity and technical depth. In the course of the scrutiny of the argument, virtually all its aspects have been weighed and tested. Is the indeterminism of the hole argument merely an artifact of an ill-chosen definition of determinism? Is the problem merely a trivial variant of the philosophical puzzle of inscrutability of reference? Or are there

deep matters of physics at issue? The debate continues over these and further issues. To enter it, the reader is directed to the bibliography below.

Bibliography

- Bain, Jon (1998) *Representations of Spacetime: Formalism and Ontological Commitment* Dissertation, Dept. H.P.S., University of Pittsburgh.
- Belot, Gordon (1995) "Indeterminism and Ontology," *International Studies in the Philosophy of Science*, **9**, pp.85-101.
- Belot, Gordon (1996) *Whatever is Never and Nowhere is Not: Space, Time and Ontology in Classical and Quantum Gravity* Dissertation, Department of Philosophy, University of Pittsburgh.
- Belot, Gordon (1996a) "Why General Relativity *Does* Need an Interpretation," *Philosophy of Science*, **63**(supplement), pp.S80-S88.
- Brighouse, Carolyn (1994) "Spacetime and holes," in D. Hull, M. Forbes and R. M. Burian (eds.) *PSA 1994* Vol.1 pp. 117-125.
- Butterfield, Jeremy "Albert Einstein meets David Lewis," pp. 56-64 in A. Fine and J. Leplin (eds.) *PSA 1988* Vol. 2.
- Butterfield, Jeremy (1989) "The Hole Truth," *British Journal for the Philosophy of Science,* **40**, 1-28.
- Corry, Leo, Renn, Juergen, and Stachel, John (1997) "Belated Decision in the Hilbert-Einstein Priority Dispute," *Science,* **278**, pp. 1270-73.
- Earman. John (1986), "Why Space is not a Substance (At Least Not to First Degree)" *Pacific Philosophical Quarterly*, **67**, pp. 225-244.
- Earman, John (1989) *World Enough and Space-Time: Absolute Versus Relational Theories of Space and Time* Cambridge, MA:MIT Bradford.
- Earman, John and Norton, John D. (1987) "What Price Spacetime Substantivalism," *British Journal for the Philosophy of Science*, **38**, 515-525.
- Einstein, Albert (1916), "The Foundation of the General Theory of Relativity," pp. 111-164 in H.A.Lorentz et al., *The Principle of Relativity.* New York: Dover, 1952.
- Hoefer, Carl and Cartwright, Nancy (1993) "Substantivalism and the Hole Argument," in J. Earman et al. (eds.) *Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Gruenbaum* Pittsburgh: University of Pittsburgh Press/Konstanz: Universitaetsverlag Konstanz.
- Hoefer, Carl (1996) "The Metaphysics of Space-Time Substantivalism," *Journal of Philosophy*, **93**, pp. 5-27.
- Howard, Don and Norton, John D. (1993), "Out of the Labyrinth? Einstein, Hertz and the Goettingen Answer to the Hole Argument," pp. 30-62 in John Earman, Michel Janssen, John D. Norton (eds.) *The Attraction of Gravitation: New Studies in History of General Relativity* Boston: Birkhäuser.
- Janssen, Michel (forthcoming) "Rotation as the Nemesis of Einstein's ‘Entwurf’ Theory," *Einstein Studies.*
- Jammer, Max (1993) *Concepts of Space: The History of Theories of Space in Physics* Third

enlarged edition. New York: Dover. Chapter 6. "Recent Developments."

- Klein, Martin J. et al. (eds.) (1995) *The Collected Papers of Albert Einstein. Volume 4. The Swiss Years: Writing, 1912-1914.* Princeton: Princeton Univ. Press.
- Liu, Chuang (1996) "Realism and Spacetime: Of Arguments Against Metaphysical Realism and Manifold Realism," *Philosophia Naturalis*, **33**, pp. 243-63.
- Liu, Chuang (1996a) "Gauge Invariance, Indeterminism, and Symmetry Breaking," *Philosophy of Science* **63**(S), pp. S71-S80.
- Leeds, S. (1995) "Holes and Determinism: Another Look," *Philosophy of Science* **62**, pp.425-178.
- Maudlin, Tim (1989), "The Essence of Spacetime," pp. 82-91 in A. Fine and J. Leplin (eds.) *PSA 1988* Vol. 2.
- Maudlin, Tim (1990) "Substances and Spacetimes: What Aristotle Would have Said to Einstein," *Studies in the History and Philosophy of Science*, **21**, pp. 531-61.
- Muller, F. A. (1995) "Fixing a Hole," *Foundations of Physics Letters*, **8**, pp.549-562.
- Mundy, Brent (1992) "Spacetime and Isomorphism," in D. Hull, M. Forbes and K. Okruhlik (eds.) *PSA 1992* Vol.1, pp.515-527.
- Norton, John D. (1984) "How Einstein found his Field Equations: 1912-1915," *Historical Studies in the Physical Sciences,* **14**, 253-316; reprinted in Don Howard and John Stachel (eds.) *Einstein and the History of General Relativity: Einstein Studies,* Vol. 1 Boston: Birkhäuser, 1989, pp.101-159.
- Norton, John D.(1987) "Einstein, the Hole Argument and the Reality of Space," pp. 153-188 in John Forge (ed.) *Measurement, Realism and Objectivity.* Reidel.
- Norton, John D. (1988) "The Hole Argument," pp. 56-64 in A. Fine and J. Leplin (eds.) *PSA 1988* Vol. 2.
- Norton, John D. (1989) "Coordinates and Covariance: Einstein's view of spacetime and the modern view," *Foundations of Physics*, **19**, 1215-1263.
- Norton, John D. (1992) "The Physical Content of General Covariance" in J. Eisenstaedt and A. Kox eds., *Studies in the History of General Relativity: Einstein Studies,* Vol.3, Boston: Birkhauser.
- Norton, John D. (1992a) "Philosophy of Space and Time," in M.H.Salmon *et al., Introduction to the Philosophy of Science*, Englewood Cliffs, NJ: Prentice-Hall; reprinted Hackett, pp.179-231.
- Rynasiewicz, Robert (1992) "Rings, Holes and Substantivalism: On the Program of Leibniz Algebras," *Philosophy of Science*, **45**, pp. 572-89.
- Rynasiewicz, Robert (1994) "The Lessons of the Hole Argument," *British Journal for the Philosophy of Science*, **45**, pp.407-436.
- Rynasiewicz, Robert (1996) "Is There a Syntactic Solution to the Hole Problem," *Philosophy of Science*(S), pp. 55-62.
- Stachel, John (1980) "Einstein's Search for General Covariance," paper read at the Ninth International Conference on General Relativity and Gravitation, Jena; printed in Don Howard and John Stachel (eds.) *Einstein and the History of General Relativity: Einstein Studies,* Vol. 1 (Boston: Birkhäuser, 1989) pp.63-100.
- Stachel, John (1986) "What can a Physicist Learn from the Discovery of General Relativity?" *Proceedings of the Fourth Marcel Grossmann Meeting on Recent Developments in General Relativity* ed. R. Ruffini. Amsterdam: North-Holland.

- Stachel, John (1993) "The Meaning of General Covariance," pp. 129-160 in J. Earman et al. (eds.) *Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Gruenbaum* Pittsburgh: University of Pittsburgh Press/Konstanz: Universitaetsverlag Konstanz.
- Teller, Paul (1991), "Substances, Relations and Arguments About the Nature of Spacetime," *The Philosophical Review*, **C**(No.3), pp. 363-97.
- Wilson, Mark (1993) "There's a Hole and a Bucket, Dear Leibniz," pp. 202-241.in P. A. French, T. E. Uehling and H. K. Wettstein (eds.) *Philosophy of Science*, Notre Dame: University of Notre Dame Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

determinism, causal | substance

Acknowledgements

I am grateful to Erik Curiel, Robert Rynasiewicz and Ed Zalta for helpful comments on earlier drafts.

John D. Norton
jdnorton+@pitt.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 1, 1999
Content last modified: February 15, 1999

Stanford Encyclopedia of Philosophy
Notes to The Hole Argument

Notes

1. This one example illustrates the core content of the hole argument. With only a little further effort, the argument can be made more precise and general. This will be done concurrently in these notes, intended for readers with some background in differential geometry and general relativity. The form of the argument follows that originally laid out in Earman and Norton (1987). For a development of the hole argument at a technical level intermediate between the body of this article and these notes, see Norton (1992a, Section 5.12).

In general, the class of theories in which the hole argument is mounted are "local spacetime theories." These are theories whose models consist of n+1 tuples

$$(M, O_1, ..., O_n)$$

where the $O_1, ..., O_n$ are geometric object fields defined on a differential manifold M. The class of models in the theory is delimited by some set of invariant (e.g. tensorial) equations, which are the laws of the theory:

$$L_1=0, ..., L_m=0$$

where the quantities L_i are functions of the geometric objects O_i.The theory is presumed complete in the sense that any model satisfying these laws will be in the model set of the theory. This characterization of local spacetime theories is sufficiently general to include formulations of virtually all common spacetime theories, including general relativity, special relativity and Newtonian theories.

2. The distinction at issue here is between the passive and active reading of general covariance. Passive general covariance allows use of all coordinate charts of the differential manifold and is conferred automatically on theories formulated by modern methods. Active general covariance considers the dual point transformations induced by coordinate transformations. These amount to diffeomorphisms on the manifold M and the transformations of the fields correspond to maps that associate an object field O with its carry along h^*O under diffeomorphism h.

The need to convert Einstein's original analysis from passive to active transformations is awkward. I have argued that the distinction between them was not so clear cut when Einstein originally formulated the hole argument because of the more impoverished mathematical environment in which he worked and that this is responsible for much of the present confusion in interpreting Einstein's pronouncements on coordinate systems. See Norton (1989, 1992).

3. That is, a hole transformation is a diffeomorphism on M that is the identity outside some arbitrarily selected neighborhood but comes smoothly to differ from the identity within that neighborhood. For an explicit construction of such a transformation, see Muller (1995).

4. More generally, manifold substantivalism asserts that the manifold M of local spacetime theories is the mathematical structure that represents spacetime.

5. For any spacetime model $(M, O_1, ..., O_n)$ and any diffeomorphism on M, Leibniz equivalence asserts that the two models

$$(M, O_1, ..., O_n) \text{ and } (hM, h^*O_1, ..., h^*O_n)$$

represent the same physical system.

6. The general form of the argument is essentially identical. The first sentence is generalized to read:

> **1.** If one has two models of a local spacetime theory $(M, O_1, ..., O_n)$ and $(hM, h^*O_1, ..., h^*O_n)$ related by a hole transformation h, manifold substantivalists must maintain that the two systems represent two distinct physical systems.

Note also that statements 1 and 2 are premises. Statement 3 is the conclusion drawn from them. There is a suppressed premise that it is inadmissible to load up a physical theory with hidden properties that outstrip both observation and the determining power of the theory.

John D. Norton
jdnorton+@pitt.edu

First published: February 1, 1999
Content last modified: February 15, 1999

Time Travel and Modern Physics

Time travel has been a staple of science fiction. With the advent of general relativity it has been entertained by serious physicists. But, especially in the philosophy literature, there have been arguments that time travel is inherently paradoxical. The most famous paradox is the grandfather paradox: you travel back in time and kill your grandfather, thereby preventing your own existence. To avoid inconsistency some circumstance will have to occur which makes you fail in this attempt to kill your grandfather. Doesn't this require some implausible constraint on otherwise unrelated circumstances? We examine such worries in the context of modern physics.

A Botched Suicide

You are very depressed. You are suicidally depressed. You have a gun. But you do not quite have the courage to point the gun at yourself and kill yourself in this way. If only someone else would kill you, that would be a good thing. But you can't really ask someone to kill you. That wouldn't be fair. You decide that if you remain this depressed and you find a time machine, you will travel back in time to just about now, and kill your earlier self. That would be good. In that way you even would get rid of the depressing time you will spend between now and when you would get into that time machine. You start to muse about the coherence of this idea, when something amazing happens. Out of nowhere you suddenly see someone coming towards you with a gun pointed at you. In fact he looks very much like you, except that he is bleeding badly from his left eye, and can barely stand up straight. You are at peace. You look straight at him, calmly. He shoots. You feel a searing pain in your left eye. Your mind is in chaos, you stagger around and accidentally enter a strange looking cubicle. You drift off into unconsciousness. After a while, you can not tell how long, you drift back into consciousness and stagger out of the cubicle. You see someone in the distance looking at you calmly and fixedly. You realize that it is your younger self. He looks straight at you. You are in terrible pain. You have to end this, you have to kill him, really kill him once and for all. You shoot him, but your eyesight is so bad that your aim is off. You do not kill him, you merely damage his left eye. He staggers off. You fall to the ground in agony, and decide to study the paradoxes of time travel more seriously.

Why Do Time Travel Suicides Get Botched?

The standard worry about time travel is that it allows one to go back and kill one's younger self and thereby create paradox. More generally it allows for people or objects to travel back in time and to cause events in the past that are inconsistent with what in fact happened. (See e.g. Gödel 1949, Earman 1972, Malament 1985a&b, Horwich 1987.) A stone-walling response to this worry is that by logic indeed inconsistent events can not both happen. Thus in fact all such schemes to create paradox are logically bound to fail. So what's the worry?

Well, one worry is the question as to why such schemes always fail. Doesn't the necessity of such failures put *prima facie* unusual and unexpected constraints on the actions of people, or objects, that have traveled in time? Don't we have good reason to believe that there are no such constraints (in our world) and thus that there is no time travel (in our world)? We will later return to the issue of the palatability of such constraints, but first we want to discuss an argument that no constraints are imposed by time travel.

Topology and Constraints

Wheeler and Feynman (1949) were the first to claim that the fact that nature is continuous could be used to argue that causal influences from later events to earlier events, as are made possible by time travel, will not lead to paradox without the need for any constraints. Maudlin (1990) showed how to make their argument precise and more general, and argued that nonetheless it was not completely general.

Imagine the following set-up. We start off having a camera with a black and white film ready to take a picture of whatever comes out of the time machine. An object, in fact a developed film, comes out of the time machine. We photograph it, and develop the film. The developed film is subsequently put in the time machine, and set to come out of the time machine at the time the picture is taken. This surely will create a paradox: the developed film will have the opposite distribution of black, white, and shades of gray, from the object that comes out of the time machine. For developed black and white films (i.e. negatives) have the opposite shades of gray from the objects they are pictures of. But since the object that comes out of the time machine is the developed film itself it we surely have a paradox.

However, it does not take much thought to realize that there is no paradox here. What will happen is that a uniformly gray picture will emerge, which produces a developed film that has exactly the same uniform shade of gray. No matter what the sensitivity of the film is, as long as the dependence of the brightness of the developed film depends in a continuous manner on the brightness of the object being photographed, there will be a shade of gray that, when photographed, will produce exactly the same shade of gray on the developed film. This is the essence of Wheeler and Feynman's idea. Let us first be a bit more precise and then a bit more general.

For simplicity let us suppose that the film is always a uniform shade of gray (i.e. at any time the shade of gray does not vary by location on the film) The possible shades of gray of the film can then be represented by the (real) numbers from 0, representing pure black, to 1, representing pure white.

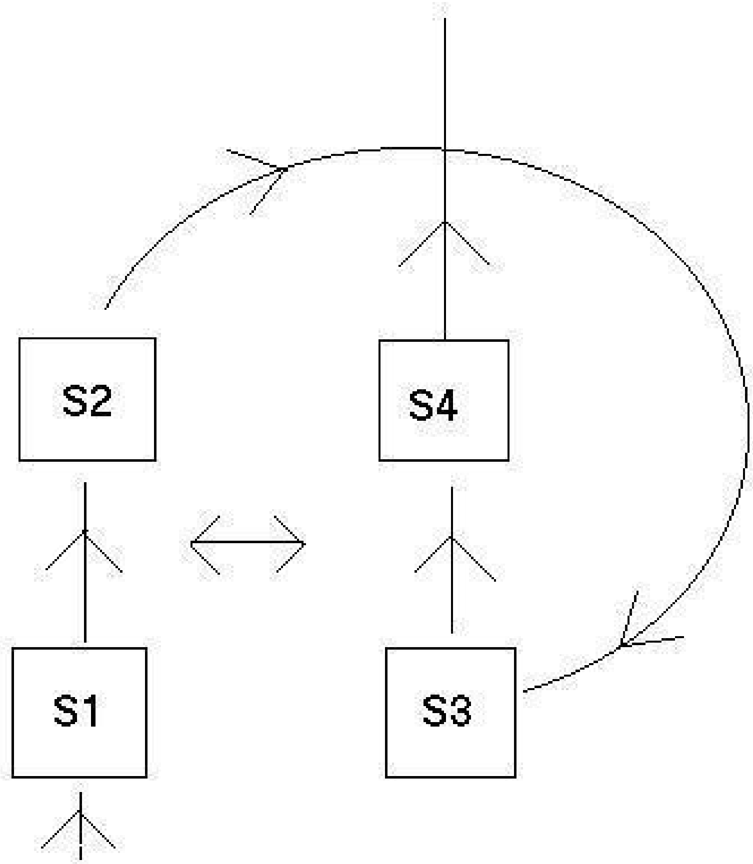


Figure 0

Let us now distinguish various stages in the chronogical order of the life of the film. In stage S_1 the film is young; it has just been placed in the camera and is ready to be exposed. It is then exposed to the object that comes out of the time machine. (That object in fact is a later stage of the film itself). By the time we come to stage S_2 of the life of the film, it has been developed and is about to enter the time machine. Stage S_3 occurs just after it exits the time machine and just before it is photographed. Stage S_4 occurs after it has been photographed and before it starts fading away. Let us assume that the film starts out in stage S_1 in some uniform shade of gray, and that the only significant change in the shade of gray of the film occurs between stages S_1 and S_2. During that period it acquires a shade of gray that depends on the shade of gray of the object that was photographed. I.e. the shade of gray that the film acquires at stage S_2 depends on the shade of gray it has at stage S_3. The influence of the shade of gray of the film at stage S_3, on the shade of gray of the film at stage S_2, can be represented as a mapping, or function, from the real numbers between 0 and 1 (inclusive), to the real numbers between 0 and 1 (inclusive). Let us suppose that the process of photography is such that if one imagines varying the shade of gray of an object in a smooth, continuous manner then the shade of gray of the developed picture of that object will also vary in a smooth, continuous manner. This implies that the function in question will be a continuous function. Now any continuous function from the real numbers between 0 and 1 (inclusive) to the real numbers between 0 and 1 (inclusive) must map at least one number to itself. One can quickly convince oneself of this by graphing such functions. For one will quickly see that any continuous function f from [0,1] to [0,1] must intersect the line x=y somewhere, and thus there must be at least one point

x such that f(x)=x. Such points are called fixed points of the function. Now let us think about what such a fixed point represents. It represents a shade of gray such that, when photographed, it will produce a developed film with exactly that same shade of gray. The existence of such a fixed point implies a solution to the apparent paradox.

Let us now be more general and allow color photography. One can represent each possible color of an object (of uniform color) by the proportions of blue, green and red that make up that color. (This is why television screens can produce all possible colors.) Thus one can represent all possible colors of an object by three points on three orthogonal lines x, y and z, that is to say, by a point in a three-dimensional cube. This cube is also known as the 'Cartesian product' of the three line segments. Now, one can also show that any continuous map from such a cube to itself must have at least one fixed point. So color photography can not be used to create time travel paradoxes either!

Even more generally, consider some system *P* which, as in the above example, has the following life. It starts in some state S_1, it interacts with an object that comes out of a time machine (which happens to be its older self), it travels back in time, it interacts with some object (which happens to be its younger self), and finally it grows old and dies. Let us assume that the set of possible states of *P* can be represented by a Cartesian product of *n* closed intervals of the reals, i.e. let us assume that the topology of the state-space of *P* is isomorphic to a finite Cartesian product of closed intervals of the reals. Let us further assume that the development of *P* in time, and the dependence of that development on the state of objects that it interacts with, is continuous. Then, by a well-known fixed point theorem in topology (see e.g. Hocking and Young 1961, p 273), no matter what the nature of the interaction is, and no matter what the initial state of the object is, there will be at least one state S_3 of the older system (as it emerges from the time travel machine) that will influence the initial state S_1 of the younger system (when it encounters the older system) so that, as the younger system becomes older, it develops exactly into state S_3. Thus without imposing any constraints on the initial state S_1 of the system *P*, we have shown that there will always be perfectly ordinary, non-paradoxical, solutions, in which everything that happens, happens according to the usual laws of development. Of course, there is looped causation, hence presumably also looped explanation, but what do you expect if there is looped time?

Unfortunately, for the fan of time travel, a little reflection suggests that there are systems for which the needed fixed point theorem does not hold. Imagine, for instance, that we have a dial that can only rotate in a plane. We are going to put the dial in the time machine. Indeed we have decided that if we see the later stage of the dial come out of the time machine set at angle *x*, then we will set the dial to *x*+90, and throw it into the time machine. Now it seems we have a paradox, since the mapping that consists of a rotation of all points in a circular state-space by 90 degrees does not have a fixed point. And why wouldn't some state-spaces have the topology of a circle?

However, we have so far not used another continuity assumption which is also a reasonable assumption. So far we have only made the following demand: the state the dial is in at stage S_2 must be a continuous function of the state of the dial at stage S_3. But, the state of the dial at stage S_2 is arrived at by taking the state of the dial at stage S_1, and rotating it over some angle. It is not merely the case that the effect of the interaction, namely the state of the dial at stage S_2, should be a continuous function of the cause, namely the state of the dial at stage S_3. It is additionally the case that path taken to get there, the way the dial is rotated between stages S_1 and S_2 must be a continuous function of the state at stage S_3. And, rather surprisingly, it turns out that this can not be done. Let us illustrate what the problem is before going to a more general demonstration that there must be a fixed point solution in the dial case.

Forget time travel for the moment. Suppose that you and I each have a watch with a single dial neither of which is running. My watch is set at 12. You are going to announce what your watch is set at. My task is going to be to adjust my watch to yours no matter what announcement you make. And my actions should have a continuous (single valued) dependence on the time that you announce. Surprisingly, this is not possible! For instance, suppose that if you announce "12", then I achieve that setting on my watch by doing nothing. Now imagine slowly and continuously increasing the announced times, starting at 12. By continuity, I must achieve each of those settings by rotating my dial to the right. If at some point I switch and achieve the announced goal by a rotation of my dial to the left, I will have introduced a discontinuity in my actions, a

discontinuity in the actions that I take as a function of the announced angle. So I will be forced, by continuity, to achieve every announcement by rotating the dial to the right. But, this rotation to the right will have to be abruptly discontinued as the announcements grow larger and I eventually approach 12 again, since I achieved 12 by not rotating the dial at all. So, there will be a discontinuity at 12 at the latest. In general, continuity of my actions as a function of announced times can not be maintained throughout if I am to be able to replicate all possible settings. Another way to see the problem is that one can similarly reason that, as one starts with 12, and imagines continuously making the announced times earlier, one will be forced, by continuity, to achieve the announced times by rotating the dial to the left. But the conclusions drawn from the assumption of continuous increases and the assumption of continuous decreases are inconsistent. So we have an inconsistency following from the assumption of continuity and the assumption that I always manage to set my watch to your watch. So, a dial developing according to a continuous dynamics from a given initial state, can not be set up so as to react to a second dial, with which it interacts, in such a way that it is guaranteed to always end up set at the same angle as the second dial. Similarly, it can not be set up so that it is guaranteed to always end up set at 90 degrees to the setting of the second dial. All of this has nothing to do with time travel. However, the impossibility of such set ups is what prevents us from enacting the rotation by 90 degrees that would create paradox in the time travel setting.

Let us now give the positive result that with such dials there will always be fixed point solutions, as long as the dynamics is continuous. Let us call the state of the dial before it interacts with its older self the initial state of the dial. And let us call the state of the dial after it emerges from the time machine the final state of the dial. We can represent the possible initial and final states of the dial by the angles *x* and *y* that the dial can point at initially and finally. The set of possible initial plus final states thus forms a torus. (See figure 1.)

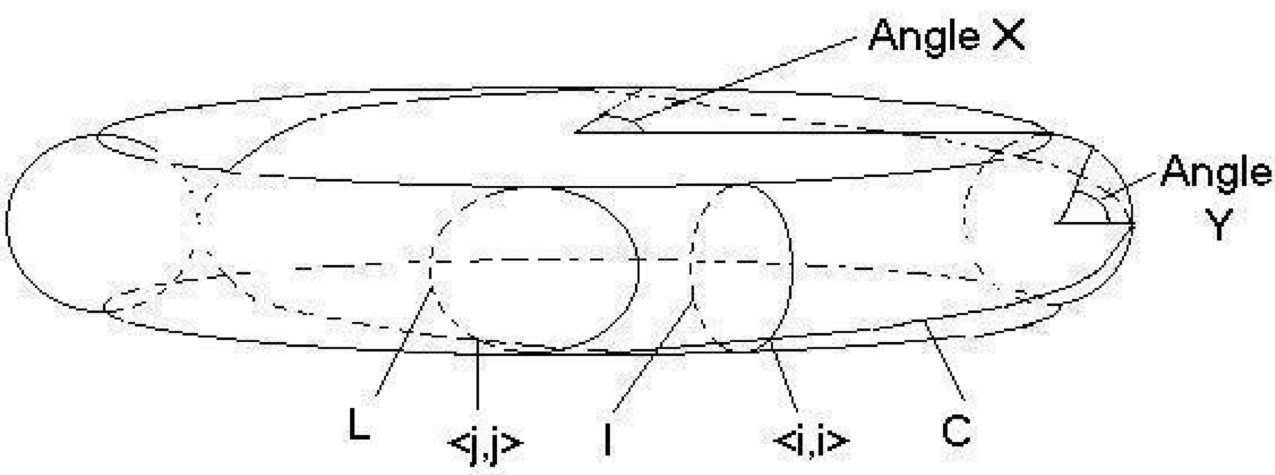


Figure 1

Suppose that the dial starts at angle *I*. The initial angle *I* that the dial is at before it encounters its older self, and the set of all possible final angles that the dial can have when it emerges from the time machine is represented by the circle *I* on the torus

(see figure 1). Given any possible angle of the emerging dial the dial initially at angle *I* will develop to some other angle. One can picture this development by rotating each point on *I* in the horizontal direction by the relevant amount. Since the rotation has to depend continuously on the angle of the emerging dial, ring *I* during this development will deform into some loop *L* on the torus. Loop *L* thus represents the angle *x* that the dial is at when it is thrown into the time machine, given that it started at angle *I* and then encountered a dial (its older self) which was at angle *y* when it emerged from the time machine. We therefore have consistency if *x*=*y* for some *x* and *y* on loop *L*. Now, let loop *C* be the loop which consists of all the points on the torus for which *x*=*y*. Ring *I* intersects *C* at point <*i*,*i*>. Obviously any continuous deformation of *I* must still intersect *C* somewhere. So *L* must intersect *C* somewhere, say at <*j*,*j*>. But that means that no matter how the development of the dial starting at *I* depends on the angle of the emerging dial, there will be some angle for the emerging dial such that the dial will develop exactly into that angle (by the time it enters the time machine) under the influence of that emerging dial. This is so no matter what angle one starts with, and no matter how the development depends on the angle of the emerging dial. Thus even for a circular state-space there are no constraints needed other than continuity.

Unfortunately there are state-spaces that escape even this argument. Consider for instance a pointer that can be set to all values between 0 and 1, where 0 and 1 are not possible values. I.e. suppose that we have a state-space that is isomorphic to an open set of real numbers. Now suppose that we have a machine that sets the pointer to half the value that the pointer is set at when it emerges from the time machine.

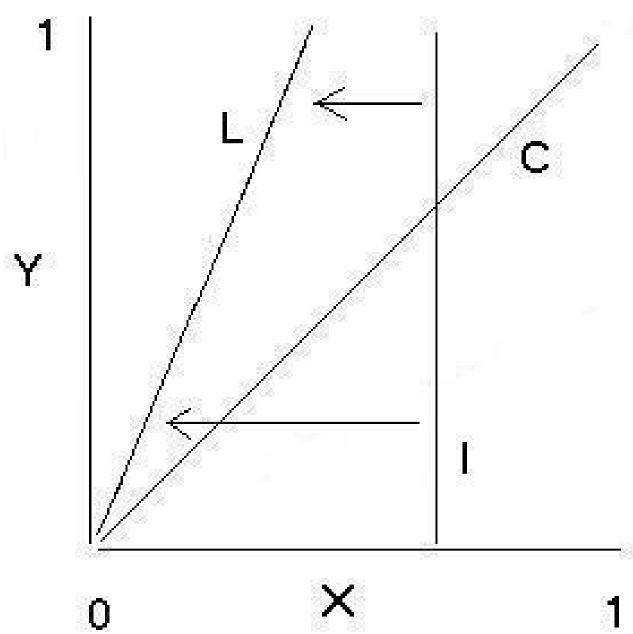


Figure 2

Suppose the pointer starts at value *I*. As before we can represent the combination of this initial position and all possible final positions by the line *I*. Under the influence of the pointer coming out of the time machine the pointer value will develop to a value that equals half the value of the final value that it encountered. We can represent this development as the continuous deformation of line *I* into line *L*, which is indicated by the arrows in Figure 2. This development is fully continuous. Points <*x*,*y*> on line *I* represent the initial position *x*=*I* of the (young) pointer, and the position y of the older pointer as it emerges

from the time machine. Points <*x,y*> on line *L* represent the position x that the younger pointer should develop into, given that it encountered the older pointer emerging from the time machine set at position y. Since the pointer is designed to develop to half the value of the pointer that it encounters, the line *L* corresponds to $x=1/2y$. We have consistency if there is some point such that it develops into that point, if it encounters that point. Thus, we have consistency if there is some point <*x,y*> on line *L* such that $x=y$. However, there is no such point: lines *L* and *C* do not intersect. Thus there is no consistent solution, despite the fact that the dynamics is fully continuous.

Of course if 0 were a possible value *L* and *C* would intersect at 0. This is surprising and strange: adding one point to the set of possible values of a quantity here makes the difference between paradox and peace. One might be tempted to just add the extra point to the state-space in order to avoid problems. After all, one might say, surely no measurements could ever tell us whether the set of possible values includes that exact point or not. Unfortunately there can be good theoretical reasons for supposing that some quantity has a state-space that is open: the set of all possible speeds of massive objects in special relativity surely is an open set, since it includes all speeds up to, but not including, the speed of light. Quantities that have possible values that are not bounded also lead to counter examples to the presented fixed point argument. And it is not obvious to us why one should exclude such possibilities. So the argument that no constraints are needed is not fully general.

An interesting question of course is: exactly for which state-spaces must there be such fixed points. We do not know the general answer.

The General Possibility of Time Travel in General Relativity

Time travel has recently been discussed quite extensively in the context of general relativity. Time travel can occur in general relativistic models in which one has closed time-like curves (CTC's). A time like curve is simply a space-time trajectory such that the speed of light is never equalled or exceeded along this trajectory. Time-like curves thus represent the possible trajectories of ordinary objects. If there were time-like curves which were closed (formed a loop), then travelling along such a curve one would never exceed the speed of light, and yet after a certain amount of (proper) time one would return to a point in space-time that one previously visited. Or, by staying close to such a CTC, one could come arbitrarily close to a point in space-time that one previously visited. General relativity, in a straightforward sense, allows time travel: there appear to be many space-times compatible with the fundamental equations of General Relativity in which there are CTC's. Space-time, for instance, could have a Minkowski metric everywhere, and yet have CTC's everywhere by having the temporal dimension (topologically) rolled up as a circle. Or, one can have wormhole connections between different parts of space-time which allow one to enter ‘mouth *A*’ of such a wormhole connection, travel through the wormhole, exit the wormhole at ‘mouth *B*’ and re-enter ‘mouth *A*’ again. Or, one can have space-times which topologically are R4, and yet have CTC's due to the ‘tilting’ of light cones (Gödel space-times, Taub-NUT space-times, etc.)

General relativity thus appears to provide ample opportunity for time travel. Note that just because there are CTC's in a space-time, this does not mean that one can get from any point in the space-time to any other point by following some future directed timelike curve. In many space-times in which there are CTC's such CTC's do not occur all over space-time. Some parts of space-time can have CTC's while other parts do not. Let us call the part of a space-time that has CTC's the "time travel region" of that space-time, while calling the rest of that space-time the "normal region". More precisely, the "time travel region" consists of all the space-time points *p* such that there exists a (non-zero length) timelike curve that starts at *p* and returns to *p*. Now let us start examining space-times with CTC's a bit more closely for potential problems.

Two Toy Models

In order to get a feeling for the sorts of implications that closed timelike curves can have, it may be useful to consider two simple models. In space-times with closed timelike curves the traditional initial value problem cannot be framed in the usual way. For it presupposes the existence of Cauchy surfaces, and if there are CTCs then no Cauchy surface exists. (A Cauchy surface is a spacelike surface such that every inextendible timelike curve crosses it exactly once. One normally specifies

initial conditions by giving the conditions on such a surface.) Nonetheless, if the topological complexities of the manifold are appropriately localized, we can come quite close. Let us call an edgeless spacelike surface *S* a *quasi-Cauchy* surface if it divides the rest of the manifold into two parts such that a) every point in the manifold can be connected by a timelike curve to *S*, and b) any timelike curve which connects a point in one region to a point in the other region intersects *S* exactly once. It is obvious that a quasi-Cauchy surface must entirely inhabit the normal region of the space-time; if any point *p* of *S* is in the time travel region, then any timelike curve which intersects *p* can be extended to a timelike curve which intersects *S* near *p* again. In extreme cases of time travel, a model may have no normal region at all (e.g. Minkowski space-time rolled up like a cylinder in a time-like direction), in which case our usual notions of temporal precedence will not apply. But temporal anomalies like wormholes (and time machines) can be sufficiently localized to permit the existence of quasi-Cauchy surfaces.

Given a timelike orientation, a quasi-Cauchy surface unproblematically divides the manifold into its *past* (i.e. all points that can be reached by past-directed timelike curves from *S*) and its *future* (ditto *mutatis mutandis*). If the whole past of *S* is in the normal region of the manifold, then *S* is a *partial Cauchy surface*: every inextendible timelike curve which exists to the past of *S* intersects *S* exactly once, but (if there is time travel in the future) not every inextendible timelike curve which exists to the future of *S* intersects *S*. Now we can ask a particularly clear question: consider a manifold which contains a time travel region, but also has a partial Cauchy surface *S*, such that all of the temporal funny business is to the future of *S*. If all you could see were *S* and its past, you would not know that the space-time had any time travel at all. The question is: are there any constraints on the sort of data which can be put on *S* and continued to a global solution of the dynamics which are different from the constraints (if any) on the data which can be put on a Cauchy surface in a simply connected manifold and continued to a global solution? If there is time travel to our future, might we we able to tell this now, because of some implied oddity in the arrangement of present things?

It is not at all surprising that there might be constraints on the data which can be put on a locally space-like surface which passes through the time travel region: after all, we never think we can freely specify what happens on a space-like surface and on another such surface to its future, but in this case the surface at issue lies to its own future. But if there were particular constraints for data on a partial Cauchy surface then we would apparently need to have to rule out some sorts of otherwise acceptable states on *S* if there is to be time travel to the future of *S*. We then might be able to establish that there will be no time travel in the future by simple inspection of the present state of the universe. As we will see, there is reason to suspect that such constraints on the partial Cauchy surface are non-generic. But we are getting ahead of ourselves: first let's consider the effect of time travel on a very simple dynamics.

The simplest possible example is the Newtonian theory of perfectly elastic collisions among equally massive particles in one spatial dimension. The space-time is two-dimensional, so we can represent it initially as the Euclidean plane, and the dynamics is completely specified by two conditions. When particles are traveling freely, their world lines are straight lines in the space-time, and when two particles collide, they exchange momenta, so the collision looks like an 'X' in space-time, with each particle changing its momentum at the impact.[1] The dynamics is purely local, in that one can check that a set of world-lines constitutes a model of the dynamics by checking that the dynamics is obeyed in every arbitrarily small region. It is also trivial to generate solutions from arbitrary initial data if there are no CTCs: given the initial positions and momenta of a set of particles, one simply draws a straight line from each particle in the appropriate direction and continues it indefinitely. Once all the lines are drawn, the worldline of each particle can be traced from collision to collision. The boundary value problem for this dynamics is obviously well-posed: any set of data at an instant yields a unique global solution, constructed by the method sketched above.

What happens if we change the topology of the space-time by hand to produce CTCs? The simplest way to do this is depicted in figure 3: we cut and paste the space-time so it is no longer simply connected by identifying the line $L-$ with the line $L+$. Particles "going in" to $L+$ from below "emerge" from $L-$, and particles "going in" to $L-$ from below "emerge" from $L+$.

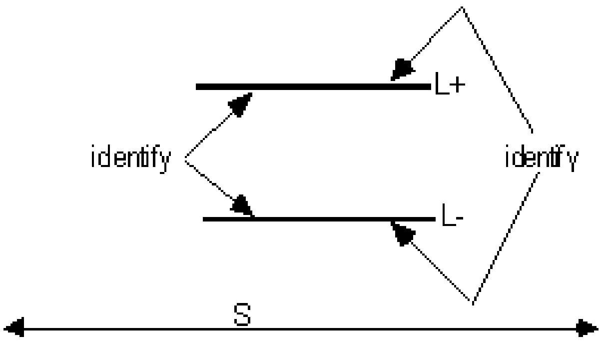


Figure 3: Inserting CTCs by Cut and Paste

How is the boundary-value problem changed by this alteration in the space-time? Before the cut and paste, we can put arbitrary data on the simultaneity slice *S* and continue it to a unique solution. After the change in topology, *S* is no longer a Cauchy surface, since a CTC will never intersect it, but it is a partial Cauchy surface. So we can ask two questions. First, can arbitrary data on *S* always be continued to a global solution? Second, is that solution unique? If the answer to the first question is *no*, then we have a backward-temporal constraint: the existence of the region with CTCs places constraints on what can happen on *S* even though that region lies completely to the future of *S*. If the answer to the second question is *no*, then we have an odd sort of indeterminism: the complete physical state on *S* does not determine the physical state in the future, even though the local dynamics is perfectly deterministic and even though there is no other past edge to the space-time region in *S*'s future (i.e. there is nowhere *else* for boundary values to come from which could influence the state of the region).

In this case the answer to the first question is *yes* and to the second is *no*: there are no constraints on the data which can be put on *S*, but those data are always consistent with an infinitude of different global solutions. The easy way to see that there always is a solution is to construct the minimal solution in the following way. Start drawing straight lines from *S* as required by the initial data. If a line hits *L*— from the bottom, just continue it coming out of the top of *L*+ in the appropriate place, and if a line hits *L*+ from the bottom, continue it emerging from *L*— at the appropriate place. Figure 4 represents the minimal solution for a single particle which enters the time-travel region from the left:

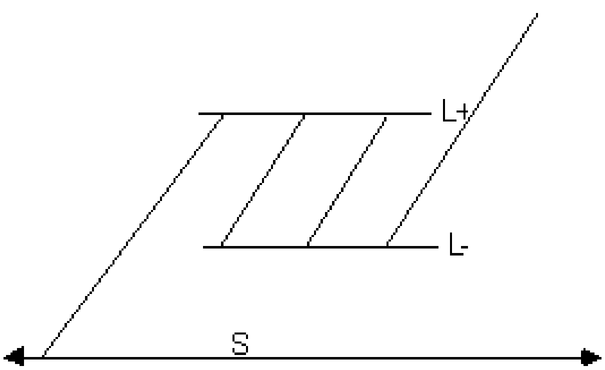


Figure 4: The Minimal Solution

The particle ‘travels back in time’ three times. It is obvious that this minimal solution is a global solution, since the particle always travels inertially.

But the same initial state on *S* is also consistent with other global solutions. The new requirement imposed by the topology is just that the data going into *L*+ from the bottom match the data coming out of *L*— from the top, and the data going into *L*- from the bottom match the data coming out of *L*+ from the top. So we can add any number of vertical lines connecting *L*- and *L*+ to a solution and still have a solution. For example, adding a few such lines to the minimal solution yields:

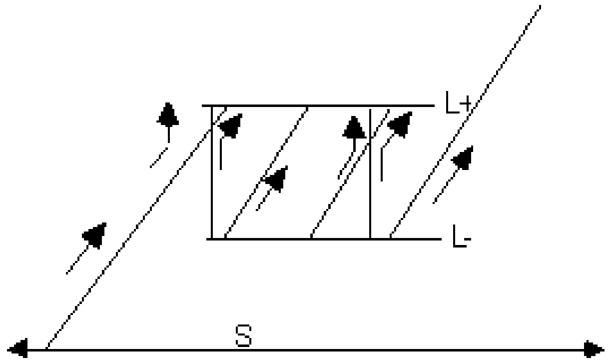


Figure 5: A Non-Minimal Solution

The particle now collides with itself twice: first before it reaches *L*+ for the first time, and again shortly before it exits the CTC region. From the particle's point of view, it is traveling to the right at a constant speed until it hits an older version of itself and comes to rest. It remains at rest until it is hit from the right by a younger version of itself, and then continues moving off, and the same process repeats later. It is clear that this is a global model of the dynamics, and that any number of distinct models could be generating by varying the number and placement of vertical lines.

Knowing the data on *S*, then, gives us only incomplete information about how things will go for the particle. We know that the particle will enter the CTC region, and will reach *L*+, we know that it will be the only particle in the universe, we know exactly where and with what speed it will exit the CTC region. But we cannot determine how many collisions the particle will undergo (if any), nor how long (in proper time) it will stay in the CTC region. If the particle were a clock, we could not predict what time it would indicate when exiting the region. Furthermore, the dynamics gives us no handle on what to think of the various possibilities: there are no probabilities assigned to the various distinct possible outcomes.

Changing the topology has changed the mathematics of the situation in two ways, which tend to pull in opposite directions. On the one hand, *S* is no longer a Cauchy surface, so it is perhaps not surprising that data on *S* do not suffice to fix a unique global solution. But on the other hand, there is an added constraint: data "coming out" of *L*− must exactly match data "going in" to *L*+, even though what comes out of *L*− helps to determine what goes into *L*+. This added consistency constraint tends to cut down on solutions, although in this case the additional constraint is more than outweighed by the freedom to consider various sorts of data on *L*+/*L*-.

The fact that the extra freedom outweighs the extra constraint also points up one unexpected way that the supposed paradoxes of time travel may be overcome. Let's try to set up a paradoxical situation using the little closed time loop above. If we send a single particle into the loop from the left and do nothing else, we know exactly where it will exit the right side of the time travel region. Now suppose we station someone at the other side of the region with the following charge: if the particle should come out on the right side, the person is to do something to *prevent* the particle from going in on the left in the first place. In fact, this is quite easy to do: if we send a particle in from the right, it seems that it can exit on the left and *deflect* the incoming left-hand particle.

Carrying on our reflection in this way, we further realize that if the particle comes out on the right, we might as well send *it* back in order to deflect itself from entering in the first place. So all we really need to do is the following: set up a perfectly reflecting particle mirror on the right-hand side of the time travel region, and launch the particle from the left so that--*if nothing interferes with it*--it will just barely hit *L*+. Our paradox is now apparently complete. If, on the one hand, nothing interferes with the particle it will enter the time-travel region on the left, exit on the right, be reflected from the mirror, re-enter from the right, and come out on the left to prevent itself from ever entering. So if it enters, it gets deflected and never enters. On the other hand, if it never enters then nothing goes in on the left, so nothing comes out on the right, so nothing is reflected back, and there is nothing to deflect it from entering. So if it doesn't enter, then there is nothing to deflect it and it enters. If it enters, then it is deflected and doesn't enter; if it doesn't enter then there is nothing to deflect it and it enters: paradox complete.

But at least one solution to the supposed paradox is easy to construct: just follow the recipe for constructing the minimal solution, continuing the initial trajectory of the particle (reflecting it the mirror in the obvious way) and then read of the number and trajectories of the particles from the resulting diagram. We get the result of figure 6:

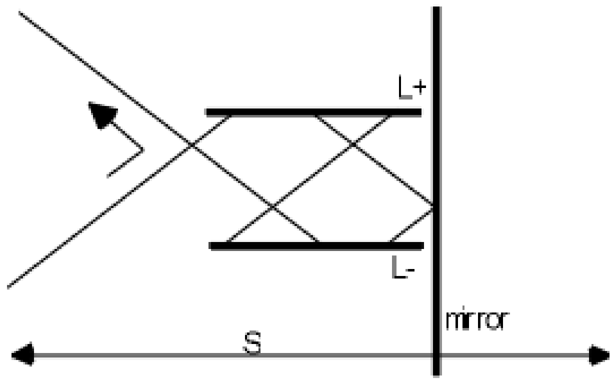


Figure 6: Resolving the "Paradox"

As we can see, the particle approaching from the left never reaches *L*+: it is deflected first by a particle which emerges from *L*-. But it is not deflected *by itself*, as the paradox suggests, it is deflected by another particle. Indeed, there are now *four* particles in the diagram: the original particle and three particles which are confined to closed time-like curves. It is not the leftmost particle which is reflected by the mirror, nor even the particle which deflects the leftmost particle; it is another particle altogether.

The paradox gets it traction from an incorrect presupposition: if there is only one particle in the world at *S* then there is only one particle which could participate in an interaction in the time travel region: the single particle would have to interact with its earlier (or later) self. But there is no telling what might come out of $L-$: the only requirement is that whatever comes out must match what goes in at *L*+. So if you go to the trouble of constructing a working time machine, you should be prepared for a different kind of disappointment when you attempt to go back and kill yourself: you may be prevented from entering the machine in the first place by some completely unpredictable entity which emerges from it. And once again a peculiar sort of indeterminism appears: if there are many self-consistent things which could prevent you from entering, there is no telling which is even likely to materialize.

So when the freedom to put data on $L-$ outweighs the constraint that the same data go into *L*+, instead of paradox we get an embarrassment of riches: many solution consistent with the data on *S*. To see a case where the constraint "outweighs" the freedom, we need to construct a very particular, and frankly artificial, dynamics and topology. Consider the space of all linear dynamics for a scalar field on a lattice. (The lattice can be though of as a simple discrete space-time.) We will depict the space-time lattice as a directed graph. There is to be a scalar field defined at every node of the graph, whose value at a given node depends linearly on the values of the field at nodes which have arrows which lead to it. Each edge of the graph can be assigned a weighting factor which determines how much the field at the input node contributes to the field at the output node. If we name the nodes by the letters *a*, *b*, *c*, etc., and the edges by their endpoints in the obvious way, then we can label the weighting factors by the edges they are associated with in an equally obvious way.

Suppose that the graph of the space-time lattice is *acyclic*, as in figure 7. (A graph is *Acyclic* if one can not travel in the direction of the arrows and go in a loop.)

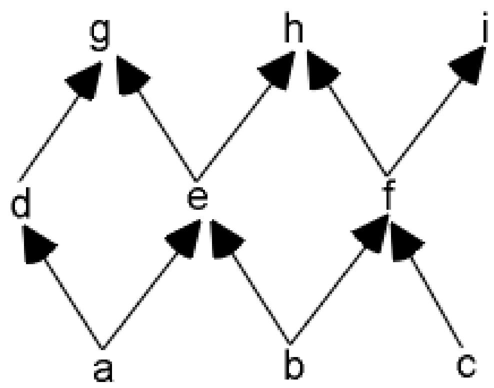


Figure 7: An Acyclic Lattice

It is easy to regard a set of nodes as the analog of a Cauchy surface, e.g. the set {*a*, *b*, *c*}, and it is obvious if arbitrary data are put on those nodes the data will generate a unique solution in the future.[2] If the value of the field at node *a* is 3 and at node *b* is 7, then its value at node *d* will be $3W_{ad}$ and its value at node *e* will be $3W_{ae} + 7W_{be}$. By varying the weighting factors we can adjust the dynamics, but in an acyclic graph the future evolution of the field will always be unique.

Let us now again artificially alter the topology of the lattice to admit CTCs, so that the graph now is cyclic. One of the simplest such graphs is depicted in figure 8: there are now paths which lead from *z* back to itself, e.g. *z* to *y* to *z*.

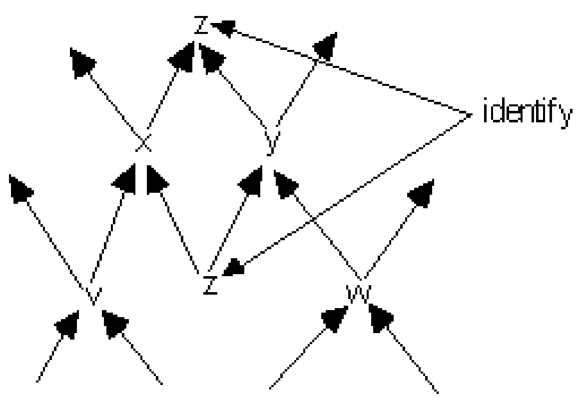


Figure 8: Time Travel on a Lattice

Can we now put arbitrary data on *v* and *w*, and continue that data to a global solution? Will the solution be unique?

In the generic case, there will be a solution and the solution will be unique. The equations for the value of the field at *x*, *y*, and *z* are:

$$x = vW_{vx} + zW_{zx}$$
$$y = wW_{wy} + zW_{zy}$$
$$z = xW_{xz} + yW_{yz}.$$

Solving these equations for *z* yields

$$z = (vW_{vx} + zW_{zx})W_{xz} + (wW_{wy} + zW_{zy})W_{yz},$$

or

$$z = (vW_{vx}W_{xz} + wW_{wy}W_{yz})/\ (1 - W_{zx}W_{xz} - W_{zy}W_{yz}),$$

which gives a unique value for z in the generic case. But looking at the space of all possible dynamics for this lattice (i.e. the space of all possible weighting factors), we find a singularity in the case where $1 - W_{zx}W_{xz} - W_{zy}W_{yz} = 0$. If we choose weighting factors in just this way, then arbitrary data at *v* and *w* cannot be continued to a global solution. Indeed, if the scalar field is everywhere non-negative, then this particular choice of dynamics puts ironclad constraints on the value of the field at *v* and *w*: the field there must be zero (assuming W_{vx} and W_{wy} to be non-zero), and similarly all nodes in their past must have field value zero. If the field can take negative values, then the values at *v* and *w* must be so chosen that $vW_{vx}W_{xz} = -wW_{wy}W_{yz}$. In either case, the field values at *v* and *w* are severely constrained by the existence of the CTC region even though these nodes lie completely to the past of that region. It is this sort of constraint which we find to be unlike anything which appears in standard physics.

Our toy models suggest three things. The first is that it may be impossible to prove in complete generality that arbitrary data on a partial Cauchy surface can *always* be continued to a global solution: our artificial case provides an example where it cannot. The second is that such odd constraints are not likely to be generic: we had to delicately fine-tune the dynamics to get a problem. The third is that the opposite problem, namely data on a partial Cauchy surface being consistent with *many* different global solutions, is likely to be generic: we did not have to do any fine-tuning to get this result. And this leads to a peculiar sort of indeterminism: the entire state on *S* does not determine what will happen in the future even though the local dynamics is deterministic and there are no other "edges" to space-time from which data could influence the result. What happens in the time travel region is constrained but not determined by what happens on *S*, and the dynamics does not even supply any *probabilities* for the various possibilities. The example of the photographic negative discussed in section 3, then, seems likely to be unusual, for in that case there is a *unique* fixed point for the dynamics, and the set-up plus the dynamical laws *determine* the outcome. In the generic case one would rather expect *multiple* fixed points, with no room for anything to influence, even probabilistically, *which* would be realized.

It is ironic that time travel should lead generically not to contradictions or to constraints (in the normal region) but to *underdetermination* of what happens in the time travel region by what happens everywhere else (an underdetermination tied neither to a probabilistic dynamics or to a free edge to space-time). The traditional objection to time travel is that it leads to contradictions: there is no consistent way to complete an arbitrarily constructed story about how the time traveler intends to act. Instead, though, it appears that the problem is underdetermination: the story can be consistently completed in many different ways. Let us now discuss some results regarding some slightly more realistic models that have been discussed in the physics literature.

Slightly More Realistic Models of Time Travel

Echeverria, Klinkhammer and Thorne (1991) considered the case of 3-dimensional single hard spherical ball that can go through a single time travel wormhole so as to collide with its younger self.

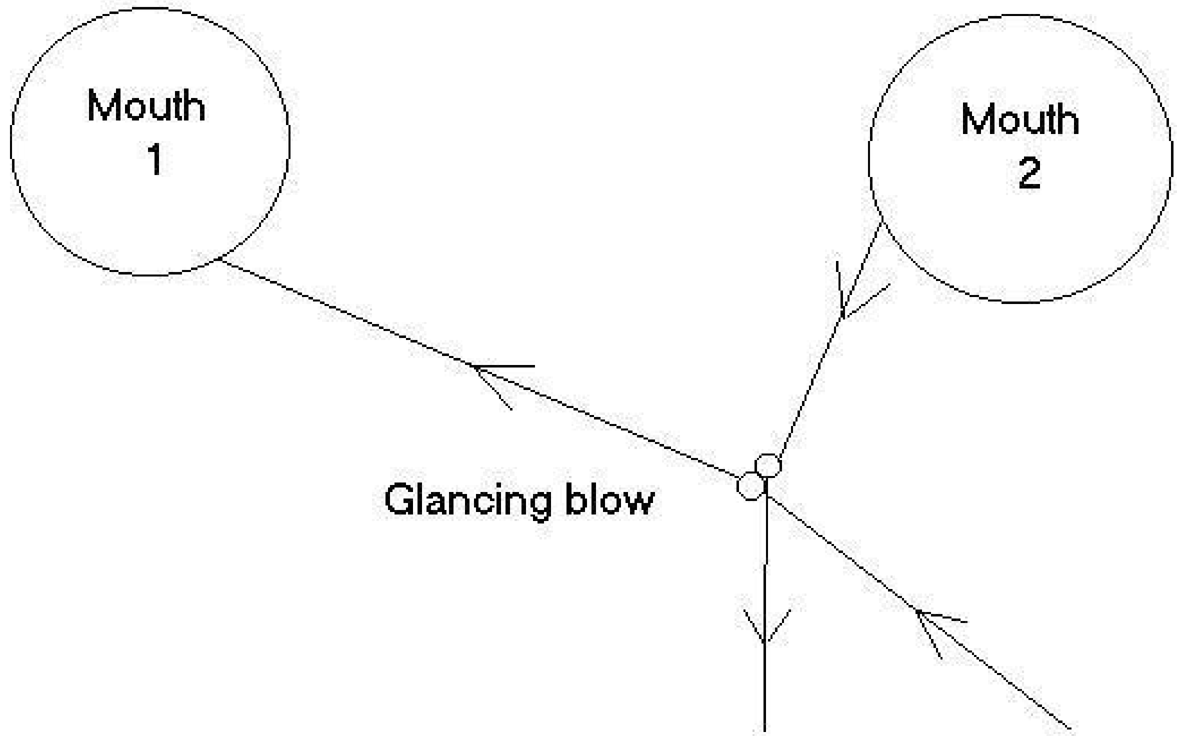


Figure 9

The threat of paradox in this case arises in the following form. There are initial trajectories (starting in the non-time travel region of space-time) for the ball such that if such a trajectory is continued (into the time travel region), assuming that the ball does not undergo a collision prior to entering mouth 1 of the wormhole, it will exit mouth 2 so as to collide with its earlier self prior to its entry into mouth 1 in such a way as to prevent its earlier self from entering mouth 1. Thus it seems that the ball will enter mouth 1 if and only if it does not enter mouth 1. Of course, the Wheeler-Feynman strategy is to look for a 'glancing blow' solution: a collision which will produce exactly the (small) deviation in trajectory of the earlier ball that produces exactly that collision. Are there always such solutions?[3]

Echeverria, Klinkhammer & Thorne found a large class of initial trajectories that have consistent 'glancing blow' continuations, and found none that do not (but their search was not completely general). They did not produce a rigorous proof that every initial trajectory has a consistent continuation, but suggested that it is very plausible that every initial trajectory has a consistent continuation. That is to say, they have made it very plausible that, in the billiard ball wormhole case, the time travel structure of such a wormhole space-time does not result in constraints on states on spacelike surfaces in the non-time travel region.

In fact, as one might expect from our discussion in the previous section, they found the opposite problem from that of inconsistency: they found underdetermination. For a large class of initial trajectories there are multiple different consistent 'glancing blow' continuations of that trajectory (many of which involve multiple wormhole traversals). For example, if one initially has a ball that is traveling on a trajectory aimed straight between the two mouths, then one obvious solution is that the ball passes between the two mouths and never time travels. But another solution is that the younger ball gets knocked into mouth 1 exactly so as to come out of mouth 2 and produce that collision. Echeverria et al. do not note the possibility (which we pointed out in the previous section) of the existence of additional balls in the time travel region. We conjecture (but have no proof) that for every initial trajectory of A there are some, and generically many, multiple ball continuations.

Friedman et al. 1990 examined the case of source free non-self-interacting scalar fields traveling through such a time travel wormhole and found that no constraints on initial conditions in the non-time travel region are imposed by the existence of such time travel wormholes. In general there appear to be no known counter examples to the claim that in ‘somewhat realistic’ time-travel space-times with a partial Cauchy surface there are no constraints imposed on the state on such a partial Cauchy surface by the existence of CTC's. (See e.g. Friedman and Morris 1991, Thorne 1994, Earman 1995, Earman and Smeenk 1999.)

How about the issue of constraints in the time travel region *T*? *Prima facie*, constraints in such a region would not appear to be surprising. But one might still expect that there should be no constraints on states on a spacelike surface, provided one keeps the surface ‘small enough’. In the physics literature the following question has been asked: for any point *p* in *T*, and any space-like surface *S* that includes *p* is there a neighborhood *E* of *p* in *S* such that any solution on *E* can be extended to a solution on the whole space-time? With respect to this question, there are some simple models in which one has this kind of extendibility of local solutions to global ones, and some simple models in which one does not have such extendibility, with no clear general pattern. (See e.g. Yurtsever 1990, Friedman et. al. 1990, Novikov 1992, Earman 1995, Earman and Smeenk 1999). What are we to think of all of this?

Even If There are Constraints, So What?

Since it is not obvious that one can rid oneself of all constraints in realistic models, let us examine the argument that time travel is implausible, and we should think it unlikely to exist in our world, in so far as it implies such constraints. The argument goes something like the following. In order to satisfy such constraints one needs some pre-established divine harmony between the global (time travel) structure of space-time and the distribution of particles and fields on space-like surfaces in it. But it is not plausible that the actual world, or any world even remotely like ours, is constructed with divine harmony as part of the plan. In fact, one might argue, we have empirical evidence that conditions in any spatial region can vary quite arbitrarily. So we have evidence that such constraints, whatever they are, do not in fact exist in our world. So we have evidence that there are no closed time-like lines in our world or one remotely like it. We will now examine this argument in more detail by presenting four possible responses, with counterresponses, to this argument.

Response 1. There is nothing implausible or new about such constraints. For instance, if the universe is spatially closed, there has to be enough matter to produce the needed curvature, and this puts constraints on the matter distribution on a space-like hypersurface. Thus global space-time structure can quite unproblematically constrain matter distributions on space-like hypersurfaces in it. Moreover we have no realistic idea what these constraints look like, so we hardly can be said to have evidence that they do not obtain.

Counterresponse 1. Of course there are constraining relations between the global structure of space-time and the matter in it. The Einstein equations relate curvature of the manifold to the matter distribution in it. But what is so strange and implausible about the constraints imposed by the existence of closed time-like curves is that these constraints in essence have nothing to do with the Einstein equations. When investigating such constraints one typically treats the particles and/or field in question as test particles and/or fields in a given space-time, i.e. they are assumed not to affect the metric of space-time in any way. In typical space-times without closed time-like curves this means that one has, in essence, complete freedom of matter distribution on a space-like hypersurface. (See response 2 for some more discussion of this issue). The constraints imposed by the possibility of time travel have a quite different origin and are implausible. In the ordinary case there is a causal interaction between matter and space-time that results in relations between global structure of space-time and the matter distribution in it. In the time travel case there is no such causal story to be told: there simply has to be some pre-established harmony between the global space-time structure and the matter distribution on some space-like surfaces. This is implausible.

Response 2. Constraints upon matter distributions are nothing new. For instance, Maxwell's equations constrain electric fields **E** on an initial surface to be related to the (simultaneous) charge density distribution ρ by the equation ρ = div(**E**). (If

we assume that the E field is generated solely by the charge distribution, this conditions amounts to requiring that the E field at any point in space simply be the one generated by the charge distribution according to Coulomb's inverse square law of electrostatics.) This is not implausible divine harmony. Such constraints can hold as a matter of physical law. Moreover, if we had inferred from the apparent free variation of conditions on spatial regions that there could be no such constraints we would have mistakenly inferred that ρ = div(**E**) could not be a law of nature.

Counterresponse 2. The constraints imposed by the existence of closed time-like lines are of quite a different character from the constraint imposed by ρ = div(**E**). The constraints imposed by ρ = div(**E**) on the state on a space-like hypersurface are: (i) local constraints (i.e. to check whether the constraint holds in a region you just need to see whether it holds at each point in the region), (ii) quite independent of the global space-time structure, (iii) quite independent of how the space-like surface in question is embedded in a given space-time, and (iv) very simply and generally stateable. On the other hand, the consistency constraints imposed by the existence of closed time-like curves (i) are not local, (ii) are dependent on the global structure of space-time, (iii) depend on the location of the space-like surface in question in a given space-time, and (iv) appear not to be simply stateable other than as the demand that the state on that space-like surface embedded in such and such a way in a given space-time, do not lead to inconsistency. On some views of laws (e.g. David Lewis' view) this plausibly implies that such constraints, even if they hold, could not possibly be laws. But even if one does not accept such a view of laws, one could claim that the bizarre features of such constraints imply that it is implausible that such constraints hold in our world or in any world remotely like ours.

Response 3. It would be strange if there are constraints in the non-time travel region. It is not strange if there are constraints in the time travel region. They should be explained in terms of the strange, self-interactive, character of time travel regions. In this region there are time-like trajectories from points to themselves. Thus the state at such a point, in such a region, will, in a sense, interact with itself. It is a well-known fact that systems that interact with themselves will develop into an equilibrium state, if there is such an equilibrium state, or else will develop towards some singularity. Normally, of course, self-interaction isn't true instantaneous self-interaction, but consists of a feed-back mechanism that takes time. But in time travel regions something like true instantaneous self-interaction occurs. This explains why constraints on states occur in such time travel regions: the states ‘ab initio’ have to be ‘equilibrium states’. Indeed in a way this also provides some picture of why indeterminism occurs in time travel regions: at the onset of self-interaction states can fork into different equi-possible equilibrium states.

Counterresponse 3. This is explanation by woolly analogy. It all goes to show that time travel leads to such bizarre consequences that it is unlikely that it occurs in a world remotely like ours.

Response 4. All of the previous discussion completely misses the point. So far we have been taking the space-time structure as given, and asked the question whether a given time travel space-time structure imposes constraints on states on (parts of) space-like surfaces. However, space-time and matter interact. Suppose that one is in a space-time with closed time-like lines, such that certain counterfactual distributions of matter on some neighborhood of a point *p* are ruled out if one holds that space-time structure fixed. One might then ask "Why does the actual state near *p* in fact satisfy these constraints? By what divine luck or plan is this local state compatible with the global space-time structure? What if conditions near *p* had been slightly different?". And one might take it that the lack of normal answers to these questions indicates that it is very implausible that our world, or any remotely like it, is such a time travel universe. However the proper response to these question is the following. There are no constraints in any significant sense. If they hold they hold as a matter of accidental fact, not of law. There is no more explanation of them possible than there is of any contingent fact. Had conditions in a neighborhood of *p* been otherwise, the global structure of space-time would have been different. So what? The only question relevant to the issue of constraints is whether an arbitrary state on an arbitrary spatial surface *S* can always be embedded into a space-time such that that state on *S* consistently extends to a solution on the entire space-time.

But we know the answer to that question. A well-known theorem in general relativity says the following: any initial data set on a three dimensional manifold *S* with positive definite metric has a unique embedding into a maximal space-time in which *S* is a Cauchy surface (see e.g. Geroch and Horowitz 1979, p. 284 for more detail), i.e. there is a unique largest space-time

which has *S* as a Cauchy surface and contains a consistent evolution of the initial value data on *S*. Now since *S* is a Cauchy surface this space-time does not have closed time like curves. But it may have extensions (in which *S* is not a Cauchy surface) which include closed timelike curves, indeed it may be that any maximal extension of it would include closed timelike curves. (This appears to be the case for extensions of states on certain surfaces of Taub-NUT space-times. See Earman and Smeenk 1999). But these extensions, of course, will be consistent. So properly speaking, there are no constraints on states on space-like surfaces. Nonetheless the space-time in which these are embedded may or may not include closed time-like curves.

Counterresponse 4. This, in essence, is the stonewalling answer which we indicated at the beginning of section 2. However, whether or not you call the constraints imposed by a given space-time on distributions of matter on certain space-like surfaces ‘genuine constraints’, whether or not they can be considered lawlike, and whether or not they need to be explained, the existence of such constraints can still be used to argue that time travel worlds are so bizarre that it is implausible that our world or any world remotely like ours is a time travel world.

Suppose that one is in a time travel world. Suppose that given the global space-time structure of this world, there are constraints imposed upon, say, the state of motion of a ball on some space-like surface when it is treated as a test particle, i.e. when it is assumed that the ball does not affect the metric properties of the space-time it is in. (There is lots of other matter that, via the Einstein equation, corresponds exactly to the curvature that there is everywhere in this time travel worlds.) Now a real ball of course does have some effect on the metric of the space-time it is in. But let us consider a ball that is so small that its effect on the metric is negligible. Presumably it will still be the case that certain states of this ball on that space-like surface are not compatible with the global time travel structure of this universe.

This means that the actual distribution of matter on such a space-like surface can be extended into a space-time with closed time-like lines, but that certain counterfactual distributions of matter on this space-like surface can not be extended into the same space-time. *But note that the changes made in the matter distribution (when going from the actual to the counterfactual distribution) do not in any non-negligible way affect the metric properties of the space-time.* Thus the reason why the global time travel properties of the counterfactual space-time have to be significantly different from the actual space-time is not that there are problems with metric singularities or alterations in the metric that force significant global changes when we go to the counterfactual matter distribution. The reason that the counterfactual space-time has to be different is that in the counterfactual world the ball's initial state of motion starting on the space-like surface, could not ‘meet up’ in a consistent way with its earlier self (could not be consistently extended) if we were to let the global structure of the counterfactual space-time be the same as that of the actual space-time. Now, it is not bizarre or implausible that there is a counterfactual dependence of manifold structure, even of its topology, on matter distributions on spacelike surfaces. For instance, certain matter distributions may lead to singularities, others may not. We may indeed in some sense have causal power over the topology of the space-time we live in. But this power normally comes via the Einstein equations. But it is bizarre to think that there could be a counterfactual dependence of global space-time structure on the arrangement of certain tiny bits of matter on some space-like surface, where changes in that arrangement by assumption do not affect the metric *anywhere in space-time in any significant way*. It is implausible that we live in such a world, or that a world even remotely like ours is like that.

Let us illustrate this argument in a different way by assuming that wormhole time travel imposes constraints upon the states of people prior to such time travel, where the people have so little mass/energy that they have negligible effect, via the Einstein equation, on the local metric properties of space-time. Do you think it more plausible that we live in a world where wormhole time travel occurs but it only occurs when people's states are such that these local states happen to combine with time travel in such a way that nobody ever succeeds in killing their younger self, or do you think it more plausible that we are not in a wormhole time travel world?[4]

Quantum Mechanics to the Rescue?

There has been a particularly clear treatment of time travel in the context of quantum mechanics by David Deutsch (see

Deutsch 1991, and Deutsch and Lockwood 1994) in which it is claimed that quantum mechanical considerations show that time travel never imposes any constraints on the pre-time travel state of systems. The essence of this account is as follows.

A quantum system starts in state *S*1, interacts with its older self, after the interaction is in state S_2, time travels while developing into state S_3, then interacts with its younger self, and ends in state S_4 (see figure 10).

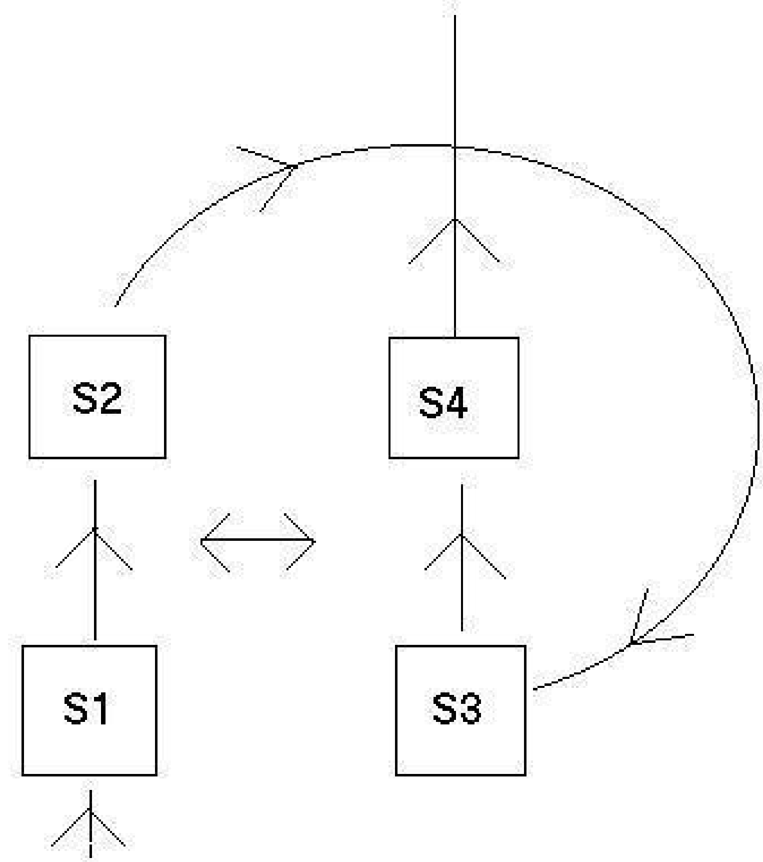


Figure 10

Deutsch assumes that the set of possible states of this system are the mixed states, i.e. are represented by the density matrices over the Hilbert space of that system. Deutsch then shows that for any initial state S_1, any unitary interaction between the older and younger self, and any unitary development during time travel, there is a consistent solution, i.e. there is at least one pair of states S_2 and S_3 such that when S_1 interacts with S_3 it will change to state S_2 and S_2 will then develop into S_3. The states S_2, S_3 and S_4 will typically be not be pure states, i.e. will be non-trivial mixed states, even if S_1 is pure. In order to understand how this leads to interpretational problems let us give an example. Consider a system that has a two dimensional Hilbert space with as a basis the states $|+\rangle$ and $|-\rangle$. Let us suppose that when state $|+\rangle$ of the young system encounters state $|+\rangle$ of the older system, they interact and the young system develops into state $|-\rangle$ and the old system remains in state $|+\rangle$. In obvious notation:

$|+\rangle_1|+\rangle_3$ develops into $|-\rangle_2|+\rangle_4$.

Similarly, suppose that:

$|+\rangle_1|-\rangle_3$ develops into $|+\rangle_2|-\rangle_4$,

$|-\rangle_1|+\rangle_3$ develops into $|-\rangle_2|-\rangle_4$, and

$|-\rangle_1|-\rangle_3$ develops into $|+\rangle_2|+\rangle_4$.

Let us furthermore assume that there is no development of the state of the system during time travel, i.e. that $|+\rangle_2$ develops into $|+\rangle_3$, and that $|-\rangle_2$ develops into $|-\rangle_3$.

Now, if the only possible states of the system were $|+\rangle$ and $|-\rangle$ (i.e. if there were no superpositions or mixtures of these states), then there is a constraint on initial states: initial state $|+\rangle_1$ is impossible. For if $|+\rangle_1$ interacts with $|+\rangle_3$ then it will develop into $|-\rangle_2$, which, during time travel, will develop into $|-\rangle_3$, which inconsistent with the assumed state $|+\rangle_3$. Similarly if $|+\rangle_1$ interacts with $|-\rangle_3$ it will develop into $|+\rangle_2$, which will then develop into $|+\rangle_3$ which is also inconsistent. Thus the system can not start in state $|+\rangle_1$.

But, says Deutsch, in quantum mechanics such a system can also be in any mixture of the states $|+\rangle$ and $|-\rangle$. Suppose that the older system, prior to the interaction, is in a state S_3 which is an equal mixture of 50% $|+\rangle_3$ and 50% $|-\rangle_3$. Then the younger system during the interaction will develop into a mixture of 50% $|+\rangle_2$ and 50% $|-\rangle_2$, which will then develop into a mixture of 50% $|+\rangle_3$ and 50% $|-\rangle_3$, which is consistent! More generally Deutsch uses a fixed point theorem to show that no matter what the unitary development during interaction is, and no matter what the unitary development during time travel is, for any state S_1 there is always a state S_3 (which typically is not a pure state) which causes S_1 to develop into a state S_2 which develops into that state S_3. Thus quantum mechanics comes to the rescue: it shows in all generality that no constraints on initial states are needed!

One might wonder why Deutsch appeals to mixed states: will superpositions of states $|+\rangle$ and $|-\rangle$ not suffice? Unfortunately such an idea does not work. Suppose again that the initial state is $|+\rangle_1$. One might suggest that that if state S_3 is $1/\sqrt{2}\,|+\rangle_3 + 1/\sqrt{2}\,|-\rangle_3$ one will obtain a consistent development. For one might think that when initial state $|+\rangle_1$ encounters the superposition $1/\sqrt{2}\,|+\rangle_3 + 1/\sqrt{2}\,|-\rangle_3$, it will develop into superposition $1/\sqrt{2}\,|+\rangle_2 + 1/\sqrt{2}\,|-\rangle_2$, and that this in turn will develop into $1/\sqrt{2}\,|+\rangle_3 + 1/\sqrt{2}\,|-\rangle_3$, as desired. However this is not correct. For initial state $|+\rangle_1$ when it encounters $1/\sqrt{2}\,|+\rangle_3 + 1/\sqrt{2}\,|-\rangle_3$, will develop into the entangled state $1/\sqrt{2}\,|-\rangle_2|+\rangle_4 + 1/\sqrt{2}\,|+\rangle_2|-\rangle_4$. In so far as one can speak of the state of the young system after this interaction, it is in the mixture of 50% $|+\rangle_2$ and 50% $|-\rangle_2$, not in the superposition $1/\sqrt{2}\,|+\rangle_2 + 1/\sqrt{2}\,|-\rangle_2$. So Deutsch does need his recourse to mixed states.

This clarification of why Deutsch needs his mixtures does however indicate a serious worry about the simplifications that are part of Deutsch's account. After the interaction the old and young system will (typically) be in an entangled state. Although for purposes of a measurement on one of the two systems one can say that this system is in a mixed state, one can not represent the full state of the two systems by specifying the mixed state of each separate part, as there are correlations between observables of the two systems that are not represented by these two mixed states, but are represented in the joint entangled state. But if there really is an entangled state of the old and young systems directly after the interaction, how is one to represent the subsequent development of this entangled state? Will the state of the younger system remain entangled with the state of the older system as the younger system time travels and the older system moves on into the future? On what

space-like surfaces are we to imagine this total entangled state to be? At this point it becomes clear that there is no obvious and simple way to extend elementary non-relativistic quantum mechanics to space-times with closed time-like curves. There have been more sophisticated approaches than Deutsch's to time travel, using technical machinery from quantum field theory and differentiable manifolds (see e.g. Friedman et al 1991, Earman and Smeenk 1999, and references therein). But out of such approaches no results anywhere near as clear and interesting as Deutsch's have been forthcoming.

How does Deutsch avoid these complications? Deutsch assumes a mixed state S_3 of the older system prior to the interaction with the younger system. He lets it interact with an arbitrary pure state S_1 younger system. After this interaction there is an entangled state S' of the two systems. Deutsch computes the mixed state S_2 of the younger system which is implied by this entangled state S'. His demand for consistency then is just that this mixed state S_2 develops into the mixed state S_3. Now it is not at all clear that this is a legitimate way to simplify the problem of time travel in quantum mechanics. But even if we grant him this simplification there is a problem: how are we to understand these mixtures?

If we take an ignorance interpretation of mixtures we run into trouble. For suppose that we assume that in each individual case each older system is either in state $|+\rangle_3$ or in state $|-\rangle_3$ prior to the interaction. Then we regain our paradox. Deutsch instead recommends the following, many worlds, picture of mixtures. Suppose we start with state $|+\rangle_1$ in all worlds. In some of the many worlds the older system will be in the $|+\rangle_3$ state, let us call them *A*-worlds, and in some worlds, *B*-worlds, it will be in the $|-\rangle_3$ state. Thus in *A*-worlds after interaction we will have state $|-\rangle_2$, and in *B*-worlds we will have state $|+\rangle_2$. During time travel the $|-\rangle_2$ state will remain the same, i.e turn into state $|-\rangle_3$, but the systems in question will travel from *A*-worlds to *B*-worlds. Similarly the $|+\rangle_2$ states will travel from the *B*-worlds to the *A*-worlds, thus preserving consistency.

Now whatever one thinks of the merits of many worlds interpretations, and of this understanding of it applied to mixtures, in the end one does not obtain genuine time travel in Deutsch's account. The systems in question travel from one time in one world to another time in another world, but no system travels to an earlier time in the same world. (This is so at least in the normal sense of the word ‘world’, the sense that one means when, for instance, one says "there was, and will be, only one Elvis Presley in this world".) Thus, even if it were a reasonable view, it is not quite as interesting as it may have initially seemed.

Conclusions

What remains of the killing-your-earlier-self paradox in general relativistic time travel worlds is the fact that in some cases the states on edgeless spacelike surfaces are ‘overconstrained’, so that one has less than the usual freedom in specifying conditions on such a surface, given the time-travel structure, and in some cases such states are ‘underconstrained’, so that states on edgeless space-like surfaces do not determine what happens elsewhere in the way that they usually do, given the time travel structure. There can also be mixtures of those two types of cases. The extent to which states are overconstrained and/or underconstrained in realistic models is as yet unclear, though it would be very surprising if neither obtained. The extant literature has primarily focused on the problem of overconstraint, since that, often, either is regarded as a metaphysical obstacle to the possibility time travel, or as an epistemological obstacle to the plausibility of time travel in our world. As we have discussed, using responses and counterresponses, it is not entirely clear that it is indeed an epistemological or a metaphysical obstacle. It is true that our world would be quite different from the way we normally think it is, if states were overconstrained given the time travel structure. If anything, underconstraint seems even more bizarre to us than overconstraint. However, time travel is quite strange to begin with, and it does not appear to be a terribly strong additional argument against time travel that it has strange consequences.

Bibliography

- Deutsch, D. 1991. "Quantum mechanics near closed timelike curves," *Physical Review D* 44, 3197-3217.
- Deutsch, D. and Lockwood, M. 1994. "The quantum physics of time travel," *Scientific American*, March 1994, 68-74.
- Earman, J. 1972. "Implications of causal propagation outsider the null cone," in *Foundations of Space-Time Theory, Minnesota Studies in the Philosophy of Science*, Vol VII, Earman, J., Glymour, C., and Stachel, J. (eds), pp 94- 108. Minneapolis, University of Minnesota Press.
- Earman, J. 1995. *Bangs, Crunches, Whimpers and Shrieks: Singularities and Acausalities in Relativistic Spacetimes*. New York: Oxford University Press.
- Earman, J. and Smeenk, C. 1999. "Take a ride on a time machine," Manuscript.
- Echeverria, F., Klinkhammer, G., and Thorne, K. 1991. "Billiard ball in wormhole spacetimes with closed timelike curves: classical theory," *Physical Review D*, Vol 44 No 4, 1077-1099.
- Friedman, J. et al. 1990. "Cauchy problem in spacetimes with closed timelike lines," *Physical Review D* 42, 1915-1930.
- Friedman, J. and Morris, M. 1991. "The Cauchy problem for the scalar wave equation is well defined on a class of spacetimes with closed timelike curves," *Physical Review letters* 66, 401-404.
- Geroch, R. and Horowitz, G. 1979. "Global structures of spacetimes," in *General Relativity, an Einstein Centenary Survey*, Hawking, S., and Israel, W., eds.
- Gödel, K. 1949. "A remark about the relationship between relativity theory and idealistic philosophy," in *Albert Einstein: Philosopher-Scientist*, edited by P. Schilpp, pp 557-562. Open Court, La Salle.
- Hocking, J., and Young, G. 1961. *Topology*. New York: Dover Publications.
- Horwich, P. 1987. "Time travel," in *Asymmetries in time*. Cambridge: MIT Press.
- Malament, D. 1985a. "'Time travel' in the Gödel universe," *PSA* 1984, Vol 2, 91-100. Asquith, P., and Kitcher, P. editors. Philosophy of Science Association, East Lansing, Michigan.
- Malament, D. 1985b. "Minimal acceleration requirements for 'time travel' in Gödel spacetime," *Journal of Mathematical Physics* 26, 774-777.
- Maudlin, T. 1990. "Time Travel and topology," *PSA* 1990, Vol 1, 303-315. Philosophy of Science Association, East Lansing, Michigan.
- Novikov, I. 1992. "Time machine and self-consistent evolution in problems with self-interaction," *Physical Review D* 45, 1989-1994.
- Thorne, K. 1994. *Black Holes and Time Warps, Einstein's Outrageous Legacy*. W.W. Norton: London and New York.
- Wheeler, J. and Feynman, R. 1949. "Classical electrodynamics in terms of direct interparticle action," *Reviews of Modern Physics* 21, 425-434.
- Yurtsever, U. 1990. "Test fields on compact space-times," *Journal of Mathematical Physics* 31, 3064-3078.

Other Internet Resources

- Time Travel in Flatland (Cal Tech Particle Theory Group)

Related Entries

determinism, causal | Gödel, Kurt: contributions to relativity theory

Acknowledgements

Thanks to Edward N. Zalta, who spotted that we incorrectly stated one of the consequences of Maxwell's equations as **E** = div(ρ) rather than as ρ = div(**E**).

Frank Arntzenius
arntzeni@rci.rutgers.edu
and
Tim Maudlin
maudlin@rci.rutgers.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 17, 2000
Content last modified: March 9, 2000

Stanford Encyclopedia of Philosophy
Notes to Time Travel and Modern Physics

Notes

[1.] Multiple collisions are handled in the obvious way by continuity considerations: just continue straight lines through the collision point and identify which particle is which by their ordering in space.

[2.] The dynamics here is radically non-time-reversible. Indeed, the dynamics is deterministic in the future direction but not in the past direction.

[3.] One might hope that fixed point theorems can be used to prove the existence of solutions in this type of cases too. Consider, for instance, a fixed initial state of motion I of the ball. Then consider all the possible velocities and locations and times $<v,x,t>$ at which such a ball could enter mouth 1 of the wormhole. Each such triple $<v,x,t>$ will determine the trajectory of that ball out of mouth 2. One can then look at the continuation of the trajectory from state I and that from state s, and see whether these trajectories collide. Then one can see for each possible triple $<v,x,t>$ whether the ball that starts in state I will be collided into mouth 1, and if it is, with which speed at what location and at which time this will occur. Thus given state I, each triple $<v,x,t>$ maps onto another triple $<v',x',t'>$. One might then suggest appealing to a fixed point theorem to argue that there must be a solution for each initial state I. However, in the first place the set of possible speeds and times are open sets. And in the second place there can be multiple wormhole traversals. Thus the relevant total state-space of wormhole mouth crossings consists of discretely many completely disconnected state-spaces (with increasing numbers of dimensions). So standard fixed point theorems do not apply directly. It should be noted that the results that have been achieved regarding this case do make use of fixed points theorems quite extensively. But their application is limited to certain sub-problems, and do not yield a fully general proof of the lack of constraints for arbitrary I.

[4.] This argument, especially the second illustration of it, is similar to the one in Horwich 1987, p 124-128. However, we do not share Horwich's view that it only tells against time travel of humans into their local past.

Frank Arntzenius
arntzeni@rci.rutgers.edu
and
Tim Maudlin
maudlin@rci.rutgers.edu

First published: February 17, 2000
Content last modified: February 17, 2000

The Epistemology of Religion

Contemporary epistemology of religion may conveniently be treated as a debate over whether Evidentialism applies to the belief-component of religious faith, or whether we should instead adopt a more permissive epistemology. Here by Evidentialism I mean the initially plausible position that a belief is warranted only if "it is proportioned to the evidence". Evidentialism implies that it is not warranted to have a full religious belief unless there is conclusive evidence for it. It follows that if the known arguments for there being a God, including any arguments from religious experience, are at best probable ones, no one would be warranted in having full belief that there is a God. And the same holds for other religious beliefs, such as the Christian belief that Jesus was God incarnate. Likewise, it would be unwarranted to believe even with less than full confidence if there is not a balance of evidence for belief.

Simplifications

Epistemology is confusing because there are several sorts of items to be evaluated and several sorts of evaluation. Since the topic of this article is the epistemology of religion not general epistemology I shall simplify matters by assuming that what is being evaluated is something related to faith, namely individual religious beliefs, and that the way of evaluating religious beliefs is as warranted or unwarranted. (See Plantinga 1993a: 3-5 for the characterisation of warrant in terms of knowledge.)

I shall also ignore disputes between coherence theorists and foundationalists; disputes between internalists and externalists and disputes over whether belief is voluntary. Although these have some implications for the epistemology of religion they are primarily topics in general epistemology.

The Rejection of Enlightenment Evidentialism

Most contemporary epistemology of religion is *postmodern* in being a reaction to the Enlightenment thesis of the *hegemony* of Evidentialism. I discuss hegemony below, but let us first consider Evidentialism. This is the initially plausible position that a belief is warranted only if "it is proportioned to the evidence". (Beliefs proportioned to the evidence include, as a degenerate case, the evidence itself.) Here several sorts of evidence are allowed. One consists of beliefs in that which is "evident to the senses", that is, beliefs directly due to sense-experience. Another sort of evidence is that which is "self-evident", that is, obvious once you think about it. Evidence may also include the beliefs directly due to memory and introspection. Again moral convictions might count as evidence, even if not treated as "self-evident". But in order to state the sort of Evidentialism characteristic of Enlightenment thought, I stipulate that no beliefs asserting the content of religious or mystical experiences count as evidence. That does not prevent the fact that someone has had a religious experience with a certain content as counting as evidence. Likewise the fact that various people report miracles counts as evidence.

Evidentialism implies that it is not warranted to have a full religious belief (ie a religious belief held with full confidence) unless there is conclusive evidence for it. The content of religious experience has been stipulated not to count as evidence. Religious beliefs do not seem to be self-evident. So the only available evidence would seem to be non-religious premisses, from which the religious beliefs are inferred. Therefore, the only way of deciding whether the religious beliefs are warranted would be to examine various arguments with the non-religious beliefs as premisses and the religious beliefs as conclusions.

According to Evidentialism it follows that if the known arguments for there being a God, including any arguments from religious experience, are at best probable ones, no one would be warranted in having full belief that there is a God. And the same holds for other religious beliefs, such as the Christian belief that Jesus was God incarnate. Likewise, it would be unwarranted to believe even partially (ie with less than full confidence) if there is not a balance of evidence for belief.

In fact it seems that many religious believers combine full belief with "doubts" in the sense of some reasons for doubting, or they combine partial belief with what they take to be weighty reasons for disbelief. According to Evidentialism they are unwarranted. Other believers consider that, on reflection, they have little reason for doubting but that they have almost no positive evidence for their religious beliefs. According to Evidentialism they too are unwarranted. This raises the question, how can we adjudicate between an epistemological thesis which might otherwise be believed and a religious belief which that thesis implies is unwarranted? One component of the Enlightenment (with the notable exception of Hume) was the *hegemony* of epistemology. By that I mean the assumptions that (a) we can discover the correct epistemology in isolation from discovering actual human tendencies to form beliefs, and that (b) we have an overriding reason to use this epistemology to correct those tendencies. If,

according to Evidentialism, full or even partial religious beliefs were unwarranted, then, given the hegemony of epistemology we have an overriding reason to reject those beliefs. Perhaps the clearest exponent of this position is the comparatively recent Clifford whose use of moral vocabulary conveys well the overriding character of the reasons epistemology is meant to provide. His position is summed up in the famous quote: "It is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence" (Clifford 1879: 186).

At the other extreme from Clifford is the position of Fideism, namely that if an epistemological theory such as Evidentialism conflicts with the warrant of religious beliefs then that is so much the worse for the epistemological theory.

The Enlightenment position was the hegemony of *Evidentialism*. Its rejection is quite compatible with holding a hegemony thesis for a fragment of epistemology, weaker than Evidentialism. Such a fragment might, for instance, contain the principle of self-referential consistency, relied upon by Plantinga (1983: 60). This states that it is not warranted to have a belief according to which that belief is itself not warranted.

As I understand it, postmodernism implies more than being postmodern in my sense. Postmodernism is the rejection of the hegemony of even a fragment of epistemology. That might seem agreeable to fideists. Postmodernism tends, however, to trivialise Fideism by obliterating any contrast between faith in divine revelation and trust in human capacities to discover the truth.

Much contemporary epistemology of religion seeks to avoid the extremes both of the Enlightenment Evidentialism and of Fideism. It is thus postmodern without necessarily being postmodernist. Let us call the injunction to avoid these extremes the *problematic* of contemporary epistemology of religion.

Evidentialism Defended

One response to the problematic is to separate Evidentialism from the hegemony of epistemology. Evidentialism may then be argued for by noting how we implicitly rely upon evidentialist principles in many different areas of enquiry, or by noting which principles generalise various particular examples of warranted and unwarranted reasoning. Such a defence of Evidentialism is part of the project of some contemporary philosophers who seek to attack theism in favour of agnosticism and/or atheism. This defence may well be implicit in Flew’s famous "The Presumption of Atheism" (1972). It is more explicit in Scriven’s *Primary Philosophy* (1966, ch 4). Scriven anbd Flew are relying on the Ockhamist principle that, in the absence of evidence for the existence of things of kind X, belief in Xs is unwarranted. This they can defend by means of examples in which non-Ockhamist thinking is judged unwarranted. So even if the whole of Evidentialism is not defended the Ockhamist fragment of it may be.

Not surprisingly the reliance of non-theist philosophers on Evidentialism has been criticised. First there is an ad hominem. Shalkowski (1989) has pointed out that these defenders of Evidentialism tend in fact

to be atheists not agnostics, yet a careful examination, he says, of the examples used to support Ockham's Razor show that either they are ones in which there is independent evidence for denying the existence of Xs or ones in which suspense of judgement seems to be the appropriate response, not denial. Another criticism is Plantinga's claim that Evidentialism is self-referentially inconsistent for there is no evidence for Evidentialism (Plantinga 1983: 60). This might be met in either of two ways. First, it could be said that all that is being defended is the Ockhamist fragment of Evidentialism and that this is not itself vulnerable to Ockham's Razor. Or it could be argued that deriving an epistemology from a wide range of examples is evidence for it. To be sure this is far from conclusive evidence. But even a less than full belief in an epistemological thesis which showed theism to be unwarranted would be damaging. This may be illustrated using an example with artificial numerical precision: 80% confidence in an epistemology which showed that no degree of belief in theism greater than 60% was warranted would demonstrate, I take it, that no degree belief in theism greater than 68% was warranted.(68% = 20% plus 80% of 60%.)

Natural theology

Theistic philosophers may, of course, grant Evidentialism and even grant its hegemony, but defend theism by providing the case which evidentialists demand. Here the details of the arguments are not within the scope of an article on epistemology. What is of interest is the kind of argument put forward. For a start there is the project of *demonstrating* God's existence, and this project is not restricted to neo-Thomists. (See Craig 1979, Braine 1988, Miller 1991.)To show the warrant of full belief that there is a God it is sufficient (a) to have a deductively valid argument from premisses which are themselves warrantably held with full belief unless defeated by an objection and (b) to have considered and defeated all available objections to either the premisses, the conclusion or any intermediate steps. Some of the premisses of these argument are said to be self-evident, that is obvious once you think about it. (Eg the denial of the explanatory power of an infinite causal regress, or the principle that the existence of any composite thing needs to be explained). And that raises a further epistemological problem. Does something's being self-evident to me warrant my full belief in it even if I know of those of equal or greater intellectual ability to myself to whom it is not self-evident?

Many natural theologians have, however, abandoned the search for demonstrative arguments, appealing instead to ones which are probable, either in sense of having weight but being inconclusive or in the sense of having a mathematical probability assigned to them. Notable in this regard are Mitchell's cumulative argument (Mitchell 1973) and Swinburne's Bayesian reliance on probability (Swinburne 1979). In a popular exposition of his argument Swinburne appeals instead to an inference to the best explanation (Swinburne 1995; see also Forrest 1996). While there are differences of approach, the common theme is that there is evidence for theism but evidence of a probable rather than a conclusive kind, warranting belief but not full belief.

The Relevance of Newman

Although pre-dating the current debate, Newman's rejection of Locke's and Paley's Evidentialism is relevant to the problematic of contemporary epistemology of religion. First he quite clearly rejected the hegemony of epistemology. His procedure was to examine how in fact people made up their minds on non-religious issues and argue that by the same standards religious beliefs were warranted. As a result he qualified Evidentialism by insisting that an *implicit* and *cumulative* argument could lead to warranted certainty. (See Mitchell 1990.)

Newman's position has two interpretations. One, which differs little from Swinburne's probabilistic approach to natural theology, asserts that the consilience of a number of independent pieces of probable reasoning can result in a probability so high as to be negligibly different from certainty. If, to use an example Newman would not have liked, Aquinas's five ways were independent and each had probability 75% then taken together their probability is about 99.9%. One difficulty with this interpretation is that even a highly probable argument differs from a demonstration in that the former is vulnerable to probabilistic counter-arguments. Thus a probabilistic version of the Argument from Evil might subsequently reduce the probability from 99.9% down to 75% again.

The other interpretation of Newman's position is to say that Evidentialism falsely presupposes that there are fine gradations on a scale from full belief through partial belief to partial disbelief to full disbelief. Newman claims we are not like that when it comes to those beliefs which form part of religious faith. In such cases the only available states are those of full belief and full disbelief or, perhaps, full belief, and lack of full belief. Of course we can believe that theism has a probability between 90% and 60%, say, but that could be interpreted as believing that relative to the evidence theism has a probability between 90% and 60%, which, in turn, is a comment on the strength of the case for theism not the expression of a merely partial belief.

If Newman is right then Evidentialism is slightly wrong. Instead of requiring belief to be proportioned to the evidence, full belief is warranted if the case for it holds "on the balance of probabilities". In that case a natural theology, such as Swinburne's, consisting of merely probable arguments can still show full religious belief to be warranted.

Wittgensteinian Fideism

Another reaction to the problematic is Wittgensteinian Fideism. I take this to be thesis that there are various different "language games", and that while we can ask questions about warrant within a language game it is a mistake to ask about the warrant of "playing" the game in question. In this way epistemology is relativised to language games, themselves related to forms of life, and the one used for assessing religious claims is less stringent than Evidentialism. Here there seems to be both an autonomy thesis and an incommensurability thesis. The autonomy thesis tells us that religious utterances are only to be judged as warranted or unwarranted by the standards implicit in the religious form of life, and this may be further restricted to Christianity or Hinduism, or any other religion (Malcolm 1992). The incommensurability thesis tells us that religious utterances are unlike scientific or metaphysical claims and so we are confusing different uses of language if we judge religious utterances by the standards of

science or metaphysics (Phillips 1992). Stress on the autonomy thesis brings Wittgensteinian Fideism close to the Fideism of many religious conservatives, but stress on the incommensurability thesis brings it close to the extreme liberal position of Braithwaite (1955), namely that religion is about attitudes not facts, which would, of course, be rejected by religious conservatives.

Perhaps the most obvious criticism of Wittgensteinian Fideism is that even if the underlying theory of forms of life and language games is granted it is an historical fact, warranted by the criteria of the "game" of history, that the tradition to which the majority of Jews, Christians and Muslims belong to is a form of life with heavy metaphysical commitments, and in which such utterances as "There is a God" are intended as much like "There is a star ten times more massive than the Sun" as like "There is hope". So Wittgensteinian Fideism is only appropriate for such religions as Zen Buddhism and for some, relatively recent, liberal strands of Judaism and Christianity which have rejected the traditional metaphysical commitment (as in Cupitt 1984).

We could modify the Wittgensteinian position to allow a metaphysical "language game" with its own criteria for warrant etc, and in which natural theology should be pursued. Then the Judo-Christian-Islamic "language game" would be part of this larger, autonomous metaphysical "language game". That modified account would cohere with the historical fact of the metaphysical commitment of that religious tradition. In that case, though, it would seem that, not just the Judo-Christian-Islamic "language game", but all serious intellectual enquiry should also be treated as parts of the one "game", with one set of rules. Thus Wittgensteinian Fideism would have been qualified out of existence.

Even if we reject Wittgensteinian Fideism we might still take a lesson from it. For it must surely be granted that religious utterances are not made in a purely intellectual way. Their entanglement with commitment to a way of life and their emotional charge might help to explain the fact, if it is one, that those who take religion seriously, whether believers or not, do not in fact have a continuous range of degrees of confidence but operate instead with full belief or full disbelief. For, normally, emotionally charged beliefs are either full on or full off, and in abnormal cases tend to be divided rather than partial. Thus, confronted with conflicting evidence about whether your affection is reciprocated you are far less likely to suspend judgement than to oscillate between full belief and full disbelief. Likewise it seems more normal to oscillate between full belief in God in moments of crisis and full disbelief when things go well than to suspend judgement at all times. This ties in with the Newmanian modification of Evidentialism, mentioned above.

Reformed Epistemology

An influential contemporary rejection of Evidentialism is Reformed Epistemology, due to Wolterstorff (1976) and Plantinga (1983). As Plantinga develops it, beliefs are warranted without Enlightenment-approved evidence provided they are (a) grounded, and (b) defended against known objections. Such beliefs then can themselves be used as evidence for other beliefs. Quite what grounding amounts to could be debated. Recently, however, Plantinga has proposed an account of warrant as proper functioning. This account seems to entail that S’s belief that p is grounded in event E if (a) in the circumstances E caused S

to believe that p, and (b) S’s coming to believe that p was a case of proper functioning (Plantinga 1993b).

While the details of grounding might be controversial I shall assume that reformed epistemologists assert that ordinary religious experiences of awe, gratitude, contrition, etc ground the beliefs implied by believer’s sincere reports of such experiences, provided they can be said to cause those beliefs. Such grounded beliefs are warranted provided they can be defended against known objections. They can then be used as evidence for further religious beliefs. Thus if religious experience grounds the belief that God has forgiven me for doing what is wrong to other humans beings, then that is evidence for a personal God who acts in a morally upright fashion. For, it can be argued, only such a God would find anything to forgive in the wrongs I do to my fellow human beings.

One difference between Reformed Epistemology and Fideism is that the former requires defence against known objections, such as the Argument from Evil, which the latter might dismiss such objections as either irrelevant or, worse, intellectual temptations.

Reformed Epistemology could be correct and yet far less significant than its proponents take it to be. That would occur if in fact rather few religious beliefs are grounded in the sorts of ordinary religious experiences most believers have. For it may well be that the beliefs are part of the cause of the experience rather than the other way round (Katz 1978).

One way of comparing Reformed Epistemology with Wittgensteinian Fideism is to note that the former proposes a universal relaxation of the stringent conditions of Evidentialism while the latter only proposes a relaxation for the case of religious beliefs.

Religious Experience, Revelation and Tradition

Reformed Epistemology may be thought of as a modification of Evidentialism in which the permissible kinds of evidence are expanded. In this context we should note especially Alston’s work arguing that certain kinds of religious experience can be assimilated to perception (Alston 1991).

The difference between Reformed Epistemology and Evidentialism is also shown by a consideration of revelation and inspiration. An evidentialist will consider arguments from the premiss that it is said such and such was revealed or the premiss that so and so claimed to be inspired by God, but a reformed epistemologist might allow as warranted religious beliefs grounded in the event of revelation or inspiration. Thus Mavrodes has argued that any belief due to a genuine revelation is warranted, and has discussed several modes of revelation (Mavrodes 1988). This would have the, to my mind unacceptable, consequence that warrant becomes totally inaccessible either to the person concerned or the community (Zagzebski 1993a:204-205). A similar criticism could be made of beliefs grounded in religious experience. In both cases, the question of whether a belief is genuinely grounded in religious experience or is genuinely grounded in inspiration is one that several religious traditions have paid attention to, with such theories as that of discernment of spirits (Murphy, 1990, ch 5).

Finally in what might be called Counter-Reformed Epistemology it could be allowed that a belief can be warranted if grounded in a religious tradition. Such a belief would have to be caused in the right sort of way by the right sort of tradition. As in the previous cases we might note that such grounding should be partially accessible to the believer. As far as I know rather little work has been done on this extension of Reformed Epistemology, but the social dimension of warrant has been noted (Zagzebski 1993a).

Bibliography

Works Cited

- Alston, William P. (1991), *Perceiving God: The Epistemology of Religious Experience*, Ithaca: Cornell University Press.
- Braine, David (1988), *The Reality of Time and the Existence of God: The Project of Proving God's Existence*, Oxford: Clarendon Press.
- Braithwaite, Richard B. (1955), *An Empricist's View of the Nature of Religious Belief*, Cambridge: Cambridge University Press.
- Clifford, William Kingdon (1879), *Lectures and Essays* (ed. Pollock, F.), London: Macmillan.
- Craig, William Lane (1979), *The Kalam Cosmological Argument*, London: Macmillan
- Cupitt, Don, (1984) *The Sea of Faith*, Cambridge: Cambridge University Press.
- Flew, Antony (1972), "The Presumption of Atheism", *Canadian Journal of Philosophy*, **2**: 29-46
- Forrest, Peter (1996), *God without the Supernatural: A Defense of Scientific Theism*, Ithaca: Cornell University Press
- Katz, Steven (1978), "Language Epistemology and Mysticism" in *Mysticism and Philosophical Analysis* (ed. Steven Katz), Oxford: Oxford University Press.
- Malcolm, Norman (1992), "The Groundlessness of Belief", in Geivett and Sweetman (1992): 92-103. Reprinted from *Reason and Religion*, ed. Stuart C. Brown, (1977), Ithaca: Cornell university Press.
- Mavrodes, George I. (1988), *Revelation in Religious Belief*, Philadelphia: Temple University Press.
- Miller, Barry (1991), *From Existence to God : A Contemporary Philosophical Argument*, London : Routledge.
- Mitchell, Basil (1973), *The Justification of Religious Belief*, London: Macmillan.
- Mitchell, Basil (1990), "Newman as a Philosopher" in *Newman after a Hundred Years*, eds Ian Ker and Alan G. Hill, Oxford: Clarendon Press.
- Murphy, Nancey (1990), *Theology in an Age of Scientific Reasoning*, Ithaca: Cornell University Press
- Phillips, D. Z. (1992), "Faith, Skepticism, and Religious Understanding", in Geivett and Sweetman (1992): 81-91. Reprinted from D. Z. Phillips, *Faith and Philosophical Enquiry* (1970), London: Routledge & Kegan Paul.
- Plantinga, Alvin (1993a), *Warrant: The Current Debate*, Oxford: The Clarendon Press.
- Plantinga, Alvin (1993b), *Warrant and Proper Function*, Oxford: Oxford University Press.

- Plantinga, Alvin and Wolterstorff, Nicholas, eds. (1983) *Faith and Rationality*, Notre Dame: University of Notre Dame Press.
- Scriven, Michael (1966), *Primary Philosophy*, New York: McGraw Hill.
- Swinburne, Richard (1979), *The Existence of God*, Oxford: Clarendon Press.
- Swinburne, Richard (1996), *Is There a God?*, Oxford: Oxford University Press
- Wolterstorff, Nicholas (1976), *Reason within the Bounds of Religion*, Grand Rapids: Eerdmans.
- Zagzebski Linda (1993a), "Religious Knowledge and the Virtues of the Mind", in Zagzebski (1993b).

Other Important Works

- Audi, Robert and Wainwright, William J., eds. (1986), *Rationality, Religious Belief, and Moral Commitment*, Ithaca: Cornell University Press.
- Geivett, Douglas R. and Sweetman, Brendan, eds. (1992), *Contemporary Perspectives on Religious Epistemology*, Oxford: Oxford University Press.
- Plantinga, Alvin (1983), "Reason and Belief in God" in Plantinga and Wolterstorff (1983): 16-93.
- Zagzebski, Linda, ed. (1993b) *Rational Faith: Catholic Responses to Reformed Epistemology*, Notre Dame: University of Notre Dame.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

miracles | Pascal's wager | probability calculus: interpretations of | Wittgenstein, Ludwig

Peter Forrest
pforrest@metz.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 23, 1997
Content last modified: April 24, 1997

Informal Logic

Sometimes informal logic is portrayed as a theoretical alternative to formal logic. While some informal logicians may see the discipline this way, this description places too much emphasis on a rejection of formal methods of analysis -- a rejection which does not characterize all informal logicians. The field can, therefore, be better described as the attempt to develop logical tools that can analyze and assess the "informal" reasoning that occurs in natural language contexts in, for example, political debate, legal proceedings, social commentary, and the opinion pieces featured in the mass media (in newspapers, magazines, television, the Internet, and so on).

Historically, informal logic can be described as a field which has ‘broken away’ from formal logic. In some cases, this break is characterized by a vehement rejection of formal methods of analysis but the relationship of the latter to informal logic is a matter of dispute. Few commentators would maintain that formal methods can be applied with full rigor, but many believe that they can contribute to an understanding of informal reasoning and much of the work in the field assumes a premise/conclusion model of argument which is derived from a formal paradigm. In the future, informal logic may be linked to formal logic by attempts to convert its insights into formal analogues which can be used to construct computer models of natural language reasoning, in research in AI (Artificial Intelligence), and so on.

Three distinct approaches to argument characterize informal logic. The first is founded on fallacy theory, the second is rhetorical, and the third is dialogical. The fallacy approach has been criticized by many theorists and is slowly giving way to a more general account of good reasoning which subsumes it. The account of argument that results establishes the criteria for good causal arguments, arguments by analogy, etc., and treats fallacies as failures to live up to these criteria.

Many informal logicians borrow freely from all three of these approaches, and from a variety of related disciplines and traditions. Literature in the field may, in view of this, apply techniques borrowed from formal logic, philosophy of language, communication theory, game theory, AI, speech act theory and so on. Important topics in recent literature include legal argumentation, the viability of the distinction between linked and convergent premises, deductivism, enthymemes, relevance, the management of argumentative exchange, adversarial and non-adversarial models of argument, visual modes of reasoning, premise acceptability, the logic of humor, arguments by analogy, and antecedents of informal logic in ancient, medieval and early modern times.

History

The origins of informal logic are found in the call for more relevant higher education that characterizes the social and political movements of the nineteen sixties. In logic this precipitated a more concerted attempt to apply its methods to concrete examples of everyday reasoning. The beginning of informal logic is the attempt to replace the artificial examples that characterize earlier logic books (early editions of Copi, e.g.) with instances of reasoning, argument and debate which are directly taken from newspapers, television and the mass media.

In attempting to analyze such reasoning, informal logicians were greatly influenced by a number of earlier works which analyzes ordinary language arguments. Aristotle's treatment of fallacies and his theory of rhetoric remain a frequent basis for discussion and research. The two modern works which most anticipate and influence informal logic are Hamblin's *Fallacies* and Toulmin's *The Uses of Argument*. The latter is especially notable for its emphasis on "the standards and values of practical reasoning," as opposed to "the abstract and formal criteria relied on in mathematical logic and much twentieth century epistemology."

Informal logic proper begins in North America in the nineteen seventies. The most influential figures in its development are Ralph H. Johnson and J. Anthony Blair. Their *Logical Self-Defense* was one of the first introductory texts to emphasize concrete examples of informal reasoning and their *Informal Logic Newsletter* quickly became a focus for discussion, news and research. Now the journal *Informal Logic*, it remains a barometer for developments in the field in its eighteenth year of publication.

Other journals which have played a significant role in the development of informal logic include *Argumentation*, *Philosophy and Rhetoric*, *Argumentation and Advocacy* (formerly the *Journal of the American Forensic Association*) and *Teaching Philosophy*. Philosophy journals like the *American Philosophical Quarterly* and the *Canadian Journal of Philosophy* have also published significant articles in the field.

In keeping with an emphasis on concrete examples of actual reasoning, the development of informal logic has been tied to pedagogical discussions of the ways in which students can best be taught to reason well in the social, political and work related contexts. One prominent feature of the evolution of informal logic is, therefore, the publication of dozens (and probably hundreds) of textbooks designed to teach

students how to reason in such contexts. In many cases, these texts (e.g., those by Govier, Kahane and Ruggiero) are also of theoretical interest, for they implicitly or explicitly advocate and elaborate a particular theoretical approach.

Three Approaches

Early approaches to informal logic emphasized fallacy theory. Its limits have been the subject of a great deal of discussion and debate (witness the articles in Hans Hansen & Robert Pinto's *Fallacies*), but formal and informal fallacies continue to be a vigorous subject of research. Hamblin's treatment serves as a common reference point, as does the work of John Woods and Douglas Walton, who have discussed a great variety of fallacies in a series of articles and books, first as co-authors and then as individuals. It is notable that their early work (and Woods' later work) frequently employs formal methods of analysis.

One significant development in the literature on fallacies is a number of articles which point out that instances of traditional fallacies -- *ad bacculum*, *ad hominem*, two wrongs reasoning, etc. -- often constitute good arguments. Though many introductory textbooks have ignored such developments, those commentators who theoretically defend the fallacy approach have responded by developing a more careful treatment of the fallacies than was previously the case. Groarke, Tindale and Fisher have tried to consolidate concerns about individual fallacies by developing criteria that distinguish between good and bad variants of traditional fallacy forms, in the process exchanging fallacies for corresponding forms of good argument.

A second approach to informal logic borrows from rhetorical perspectives which include classical rhetoric (especially as it is found in Aristotle) and contemporary rhetorical theory. To a much greater extent that formal logic, they have traditionally addressed practical concerns about argument and persuasion. In the attempt to understand informal argument, much can be learned from their willingness to discuss and analyze features of informal reasoning that extend beyond traditional logical concerns. Rhetoric's emphasis on audience does, for example, underscore an essential ingredient of effective argument that extends beyond logical notions of validity and soundness: in a real life context, an argument which is sound (i.e. deductively valid with true premises) is unlikely to achieve its desired end if it is constructed in a way that violates the deeply held sentiments (the *pathos*) of its intended audience.

A third prominent approach to informal logic borrows from communication theory. It sees argumentation as a form of dialogical exchange and dispute resolution which must conform to implicit normative rules. These rules determine what moves and counter-moves are and are not acceptable in a dialogue. An understanding of such rules -- and a recognition that the rules for dialogue differ in different kinds of contexts (scientific, political, etc.) -- is, on this account, the key to understanding argument. Problems with particular arguments -- fallacies, for example -- are explained as violations of the rules of dialogue.

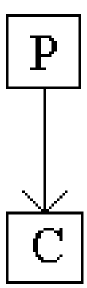
The foremost representative of the dialogical approach to informal logic is Douglas Walton. Like other commentators who develop this approach, he is greatly influenced by the Dutch "pragma-dialectic" theory of argumentation developed by Frans H. Van Eemeren and Rob Grootendorst of the University of

Amsterdam. Taken in conjunction with informal logic, the critical thinking movement and related disciplines, the latter has helped make argumentation theory an area of intensive international research and discussion (for a collection of papers which well represents the state of the art in this multi-disciplinary field, see van Eemeren et. al.).

An Example

The three approaches to informal logic can easily be demonstrated with a simple example which is taken from a Danish television debate over the question whether the Danish church should be separated from the Danish state (see Charlotte Jorgensen, "Hostility in public debate," in Van Eemeren et. al., Vol. IV). At one point in this debate, the debater arguing against the separation of church and state declares that "My opponent wants to sever the Danish church from the state for his own personal sake. His motion is an attempt to take over the church and further his ecumenical theology by his usual mafia methods."

We can plausibly understand this remark as a simple argument with one premise and an implicit (or "missing" or "hidden") conclusion. The premise (P) is the claim that "My opponent wishes to sever the Danish church from the state for the sake of his personal interests (i.e., in order to take it over and further his ecumenical theology by his usual mafia methods)." The implied conclusion (C) is the implicit claim that "We should (therefore) reject his motion to separate the Danish church and state." The argument can be diagrammed as:



Looked at from the point of view of fallacy theory, this is a classic case of *ad hominem*. Kahane, for example, describes it as a fallacy that occurs when an arguer is guilty "of attacking his opponent rather than his opponent’s evidence and arguments" (65). In this case, the debater in question attacks the motivation and the character of the person promoting a separate Danish church instead of showing what is wrong with his evidence for the claim that this is a good idea. On these grounds, the fallacy approach rejects the proposed reasoning as fallacious.

Though dialogical approaches employ a different theoretical structure, they invite a very similar analysis. Van Eemeren and Grootendorst explain ad hominem as a violation of their first rule for "critical discussion," which maintains that "Parties [to a dispute] must not prevent each other from advancing standpoints or casting doubts on arguments." Different kinds of *ad hominem* (i.e., abusive, circumstantial and *tu quoque ad hominem*) are construed as different violations of this rule. In the case in question, it suffices to say that the debater’s attack on his opponent illegitimately denies him his right to make a case

for his position.

Rhetorical approaches are characterized by a more sympathetic attitude to *ad hominem* arguments. This attitude can be understood in terms of Aristotle's suggestion that the *ethos* of a speaker plays a crucial role in determining whether an argument is persuasive or not. Looked at from this point of view, an *ad hominem* is, in principle, an acceptable attack on the *ethos* of a speaker (or writer). This being said, this particular example of *ad hominem* remains problematic, for it is founded on a very heavy handed and (at least in the excerpt we have looked at) unsubstantiated charge against the debater who advocates the separation of the Danish church and state. One might therefore argue that this kind of strongly worded charge undermines the ethos of the *speaker who forwards it* rather than the person he attacks.

It is significant that the rhetorical approach clearly recognizes that *ad hominem* attacks can be entirely appropriate. One may, to take a different example, reasonably cast doubt on an arguer's reasoning by pointing out that they lack the requisite knowledge to make appropriate judgments in the area in question, by pointing out that they have a vested interest, and so on. If someone with no physics credentials wishes to sell one a device to see quarks, then one might very reasonably dismiss their arguments with an *ad hominem*. Such appeals play an essential role in ordinary language reasoning, for time constraints make it impossible to carefully analyze all the arguments presented to us and we must therefore decide which ones we pay attention to by relying on an assessment of the arguer.

It is in this regard important to recognize that one can easily adapt dialogical and fallacy approaches to informal reasoning so that they too recognize the possibility of reasonable *ad hominem*s. If one adopts a fallacy approach, then this can be accomplished by defining *ad hominem* more narrowly, so that the fallacy only encompasses those attacks against the person which are fallacious. Alternatively, one can recognize *ad hominem* as in principle legitimate, but unreliable when it is not properly formulated -- when care is not taken to ensure that the argument in question does not attack an opponent's character in an irrelevant or an unsubstantiated manner).

If one adopts a dialogical approach, then one may recognize the possibility of good *ad hominem* reasoning by arguing that there are forms of dialogue which do not allow the participation of individuals who have vested interests or are in some other way not qualified to speak to an issue in some contexts. Such refinements are in keeping with our common understanding that, for example, a judge who adjudicates over a civil dispute must not have a vested interest in the outcome.

Relationship to Philosophy

Philosophy and philosophers have played a central role in defining informal logic and developing its approach to ordinary reasoning (communication theory and rhetoric have also made significant contributions). In part because of this, work in the field reflects philosophical concerns about the nature of rationality, the standards of good reasoning, the value of formal methods of analysis, the value of logic and rhetoric, and the social and political role of reasoning (especially in democracy). In the future, the analysis of ordinary reasoning which informal logic makes possible may allow new insights in

philosophy of mind, ethics and epistemology.

Within the subdisciplines of philosophy, the emphasis that informal logic places on natural language reasoning and on concrete examples of everyday argumentation might be readily compared to the emphasis that the various kinds of "applied ethics" (biomedical ethics, business ethics, etc.) place on concrete moral problems. In both cases, one finds an approach to philosophy which places great emphasis on its relevance to practical concerns. In many ways, one might compare this attitude to philosophy to the one that characterizes ancient philosophical perspectives which emphasize the ways in which philosophy can contribute to day to day life (e.g., Stoicism, Skepticism and Epicureanism).

The ideal theory of informal logic will encompass both a general theory of argument and a procedure for applying it to concrete instances of reasoning. If the field can avoid the fragmentation that has tended to accompany a sometimes staggering proliferation of approaches derived from a variety of related disciplines, the end result may be a model of ordinary argumentation which makes much better sense from an informal and a formal point of view.

Bibliography

- Copi, Irving. *Introduction to Logic*. New York: Macmillan, 1957
- Eemeren, Frans H. van and Rob Grootendorst. *Argumentation, Communication, and Fallacies: A Pragma-Dialectical Perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992
- Eemeren, Frans H. van, Rob Grootendorst, J. Anthony Blair and Charles Willard, eds. *Perspectives and Approaches*, *Analysis and Evaluation*, *Reconstruction and Application*, *Special Fields and Case Studies*, Vols. I - IV, *Proceedings of the Third ISSA [International Society for the Study of Argument] Conference on Argumentation*. Amsterdam: International Centre for the Study of Argumentation, 1995
- Gilbert, Michael. *Coalescent Argumentation*. Mahwah: Erlbaum, 1997.
- Govier, Trudy. *Problems in Argument Analysis and Evaluation*. Dordrecht: Foris, 1987
- Govier, Trudy. *A Practical Study of Argument*. 3rd ed. Belmont, Calif.: Wadsworth, 1992
- Groarke, Leo, Christopher Tindale and Linda Fisher. *Good Reasoning Matters!*. 2nd ed. Toronto: Oxford University Press, 1997
- Hamblin, Charles Leonard. *Fallacies*. London: Methuen, 1970
- Hansen, Hans V. and Roberts C. Pinto, eds. *Fallacies: Classical and Contemporary Readings*. University Park, PA: Penn State Press, 1995
- Johnson, Ralph J. *The Rise of Informal Logic*. Newport News: Vale Press, 1996.
- Johnson, Ralph H. and J. Anthony Blair, "Informal Logic: Past and Present." Johnson, Ralph H. and J. Anthony Blair, eds. *New Essays in Informal Logic*. Windsor: Informal Logic, 1994: 1-19
- Johnson, Ralph H. and J. Anthony Blair. *Logical Self-Defense*. 3rd ed. Toronto: McGraw Hill-Ryerson, 1995
- Kahane, Howard. *Logic and Contemporary Rhetoric*. 7th ed. Belmont: Wadsworth Publishing Company, 1995
- Ruggiero, Vincent Ryan. *Becoming a Critical Thinker*. 2nd ed. Rapid City, SD: Houghton

Mifflin, 1992

- Toulmin, Stephen. *The Uses of Argument*. Cambridge: Cambridge University Press, 1964
- Walton, Douglas N. *Informal Logic: A Handbook for Critical Argumentation*. New York: Cambridge University Press, 1989
- Woods, John and Douglas Walton. *Fallacies: Selected Papers 1972-1982*. Dordrecht-Holland/Providence-RI: Foris, 1989

Other Internet Resources

- Institute for Critical Thinking
- The Critical Thinking Community
- Journals in Logic, Informal Logic, and Rhetoric
- Mission Critical/Critical Thinking Web Page
- Critical Thinking Accross the Curriculum Project
- Propaganda Techniques: Logical Fallacies
- Logic and Fallacies

Related Entries

logic: classical

Leo Groarke
Wilfrid Laurier University
lgroarke@wlu.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 25, 1996
Content last modified: April 19, 1998

Cosmology and Theology

Reasoning known as the cosmological argument (Burrill 1967; Craig 1979, 1980; Hepburn 1967) tries to justify belief in God by pointing to the existence of the cosmos, its causal orderliness, and alleged evidence of its being in some sense designed to include life and intelligence. [Often the appeal to such evidence is instead called the argument from design, or the teleological argument.] Some cosmologists believe, however, that the existence and order of the cosmos can be accounted for scientifically. Its life-permitting character might itself, they consider, be explained through its being divided into multiple domains worth the name of "universes". These could vary randomly in their features, ours being one of the perhaps very rare ones in which life had any chance of evolving. As the anthropic principle reminds us, only the life-permitting universes could give rise to observers. They should hesitate before concluding that an omnipotent, omniscient, all-creating person had made their surroundings life-permitting.

Philosophers, too, have doubted that so remarkable a person would be needed to explain such affairs, or that this person's own existence could be any less in need of explanation. They may here conceive God in a way not everybody would accept, interactions between cosmology and theology often depending on which picture of God is preferred. Such interactions include discussions of the nature of time and of the human mind, and of whether intelligent life is widespread in the cosmos.

Why is there a Cosmos?

People disagree over whether the sheer existence of the cosmos could call for explanation. Some have held that it always would, no matter how long the cosmos had existed; others that it never would; and still others that this would depend on whether the cosmos had existed eternally, or else on the nature of time.

It is nowadays usually believed that our universe came out of a Big Bang, a violent explosion occurring perhaps fifteen billion years ago. Thinking that the Bang could be explained only by God, the atheistic Hoyle (1950) preferred a universe existing eternally in a Steady State: it expanded, but new hydrogen atoms constantly materialized to keep the average cosmic density the same. Although at first offering no explanation for the new atoms, he thought them much less of a difficulty than the materialization of everything at once. Among theists, Pius XII (1952) then agreed that a Bang could indeed be counted as specially strong evidence of God's creative activity, yet many theologians protested that this activity should not be particularly associated with any first moment of a universe's existence. Tables, for instance, would vanish immediately if God failed to "conserve" their existence through exercise of his creative power.

When Hawking (1987, 1988) suggested that his own cosmology left "no place for a creator" since it made "What happened before the Bang, to cause it?" comparable to "What's Earth like to the north of the North Pole?", the theologians renewed their protests. They accompanied them with quotations from Saint Augustine, who had written that God created time and the world together. It is a theological commonplace, though, that God could be described as "creator" even of an eternally existing world. Both in theology and in philosophy, talk of creation or causation does not necessarily assume the temporal priority of creative or causal agencies. Descartes emphasized that in calling God "the cause of himself" he meant only that God's eternal existence was guaranteed by God's nature.

G.Gamow and W.B.Bonnor (both reprinted in Leslie ed. 1990) had meanwhile joined Hoyle in his eagerness to avoid a cosmic beginning. Gamow's article of 1954 favoured a universe which had contracted for infinitely long before rebounding in a Big Bang, whereas Bonnor's of 1960 proposed infinitely many oscillations, the rebounds always occurring before the cosmic material reached infinite densities, which Bonnor called "signs of error". In contrast, Milne (1952) welcomed the Bang as evidence of God's hand. He maintained, too, that the universe at the time of its creation had to be point-like and therefore infinitely dense. Creation of a spatially extended universe was a logical impossibility.

This last contribution to the debate was an outright blunder, there being no contradiction in the idea of creating something extended. Other contributions are harder to evaluate, for intuitions about what should be viewed as a universe's "natural state" -- where this means something not calling for explanation by a divine person or any other external factor -- can be defended or attacked only very controversially. Grünbaum (1990, and in Leslie ed. 1990) thinks it perverse to imagine that tables or the entire universe would disappear if God did nothing to prevent this; in his firm opinion, to discover what happens naturally we must look to see what actually occurs; while Hume (1739) holds that no cause would be needed even for the abrupt entry of a universe into a time which had previously been flowing. Yet one main message which many find in Hume's writings is that we can never learn anything from experience unless helped by various unprovable basic principles: for instance, that the future is likely to resemble the

past. Now, one such basic principle might be the need to deny that anything -- let alone an entire universe -- could come to exist for no reason whatever. True enough, quantum theory is often thought to tell us that the universe is indeterministic in some absolute way, but not even this would be a clear declaration that its existence could be utterly reasonless. Again, people can accept quantum laws without thinking that their operation is "natural" in the technical sense of not being a product of God's will.

If we saw a problem in a universe's existence, could it be removed by taking that universe to have existed for infinite past time? Some argue that this would allow its presence at any one instant to be explained by its presence at earlier instants; yet would they then say the same about, e.g., the notion that a particular book had existed eternally? Presumably not. Those Moslems who think the Koran eternal typically consider its existence due to an eternal divine decree. "It exists because it always has done" is not what they think. Moreover, there might be some force in the Kalam Cosmological Argument (Craig 1979) which opposes the world's eternity on the grounds that no infinite series of years could have been traversed to reach the present day. [A complicating factor is that whether the Big Bang's earliest, hottest stages were of finite or infinite duration might depend on one's choice of clock. What if infinitely many events occurred during those stages? The clocks most appropriate to timing the events might then show the stages as taking infinitely long to unfold. As measured by ours, such clocks would tick ever more slowly as the universe cooled.]

A better approach could be that of the quantum cosmologists, Vilenkin (1982) and Hawking (1987) for example, who speculate that quantum theory can give probabilities for worlds of certain kinds to exist. Many recent cosmological models have been inspired by E.P.Tryon's idea of 1973 (Leslie ed. 1990) that even very large universes could begin their existence by taking advantage of quantum indeterminism. Each universe would "cost" little or nothing since the energy tied up in all its particles could be cancelled by their gravitational potential energy, which is standardly treated as *negative energy*. They could therefore be akin to the quantum fluctuations in which individual particles such as electrons exist fleetingly by "borrowing" energies too small to upset quantum theory's rather disorderly balance sheets. It is at present unclear whether this could make sense except against the background of an already existing space, or at least a "space-time foam" lacking clear distinctions between space and time (see Atkins 1992; Craig and Smith 1993; Halliwell 1992; Russell, Murphy and Isham eds. 1996). A common verdict, though, is that even if no such background were needed, there could still be a problem of why quantum laws applied.

Causal Orderliness

Some read Kant as arguing that various principles which we apply to the world, for instance when we view it as causally ordered throughout, are valid only as giving insight into how we necessarily see things. Could this mean that the world was not itself ordered causally, the situation instead being that our unconscious minds were super-geniuses regulating everything we fancied we saw, to ensure that it all obeyed the laws of nature (quantum electrodynamics, general relativity, or whatever)? This would be bizarre, which leads many of Kant's admirers to suspect that he intended something more subtle. Admittedly the laws which people tend to regard as inviolable may some day break down; that is a

possibility which cosmologists find easy to defend, familiar as they are with the idea of cosmic "phase transitions" comparable to the change from ice to water and then to steam; but none the less, it remains obvious that the world up to the present moment has had considerable causal orderliness. Now, is this something that calls for explanation?

In the mid twentieth century a widespread opinion was that explaining any one causal law, for instance that heated water changes to steam, could proceed only by appeal to some more basic causal law such as that faster-moving molecules find it easier to break free from one another. The very most basic laws could not possibly be explained, therefore. They would concern mere regularities: as a matter of brute fact, events of one kind would always (or very often) be succeeded by events of some other kind. Discussing the idea that various individuals can know hidden cards by "extra-sensory perception", A.J.Ayer (1970) writes that the only thing which would be remarkable here would be someone's being "consistently rather better at guessing cards than the ordinary run of people"; the fact of doing "better than chance" would prove "nothing at all". If everybody just did know all about playing cards without looking at them, then there would be no mystery in this.

Recently, many philosophers have found such an approach dissatisfying. It is usually ascribed to Hume (1739), yet Hume suggests that causation's patterns would have characterized various events which did not actually occur. As a matter of causal necessity, a window (for example) would have broken had a brick been thrown at it, or would have remained intact had a mere peanut been thrown. This is often thought obviously right, the problem then being how causal patterns could have a *necessity* which was not just a matter of what always in fact occurred.

Might physics and cosmology throw light on any such problem? Quantum theory suggests to many people that we should think in terms of propensities, tendencies, rather than of absolutely firm necessities (a brick might "quantum-tunnel" across a window without smashing it, but this would be exceedingly unlikely) and that the Big Bang grew from a "quantum foam" in which causation as ordinarily conceived could not have acted since there was no firm direction of time. But all this leaves the fundamental issue largely untouched. Why does our world ever obey anything worth the name of a physical law? Or, to express the point differently (Wigner 1960), why "the unreasonable effectiveness of mathematics in the natural sciences"? Why does our world have the kind of elegance which made Jeans (1930) talk of a mathematically minded creator? While no logically possible world could violate mathematical principles, it is easy enough to imagine worlds in which they had little application.

When authors present a divine person or divine creative principle as responsible for the world's existence, they sometimes view causal orderliness as a matter directly attributable to this person or principle (Swinburne 1968 and 1979; Leslie 1979). Whitehead proposed instead (1938) that all things in nature in some sense strive to achieve aims, even atomic particles enjoying some very low level of awareness -- a theme echoed by the physicist D.Bohm (Bohm and Hiley 1993, chapter 15).

Note that the alleged problem of why there are causal laws is rather different from the alleged problem of why anything ever moves or changes, which is what led Aristotle to propose a divine prime mover.

Design and Fine Tuning

Living organisms provide seemingly overwhelming evidence of divine design, their parts forming immensely complicated mechanisms which permit them to survive and to reproduce themselves. To say "That's how they just happen to be" would be silly, although Philo in Hume's *Dialogues*(1779), sometimes thought to speak for Hume himself, may at times be guilty of saying it. Had Darwin said it, then he would never have discovered his famous way of undermining the seemingly overwhelming evidence: his theory of evolution, that is to say. Darwin's theory is now known to be right, so supporters of design seek their evidence elsewhere. They point to how the world's laws combine to make Darwinian evolution possible (Henderson 1913 supplies an early example of this). Why, they ask, is there a friendly environment in which extremely complex living machinery could appear after long ages, through selective inheritance of more and more complicated genes?

They can here direct our attention not just to the general fact of causal orderliness as discussed above but also (Leslie 1989, chapter 3) to the fortunate effects of particular causal principles. Consider, for instance, the principles met with in quantum theory. As well as allowing apparently dissipated wave-energy to concentrate itself so that it can do useful work, quantum laws ensure that atoms come in standardized types, making the genetic code possible. Similar things can next be said of the laws of special relativity, and of various laws controlling elementary particles.

However, much more attention has been directed towards the apparent "fine tuning" of fundamental cosmic parameters: the strengths of physical forces, the masses of elementary particles, the expansion speed and degree of turbulence at early moments in the Big Bang, and so forth (Barrow and Tipler 1986; Davies 1982; Ellis 1993; Leslie 1989, chapter 2; Leslie ed. 1990; Polkinghorne 1986; Rolston 1987). For example, it appears that electromagnetism, gravity, and the two main forces which control the atomic nucleus, had all of them to have strengths which fell inside very narrow limits if there were to be any stars of the long-living, steadily burning sort: the sort which encourage life to evolve. Again, life's complex chemistry appears possible only thanks to very precise adjustment of the masses of the neutron, the proton and the electron.

For there to be life of any readily imaginable kind, anything up to several dozen factors can appear to have needed fine tuning. Because the number of factors to be listed seems so large, this supposed evidence of design can survive many doubts about what exactly should be on the list. Take the case of the early cosmic expansion speed. It is often held that cosmic inflation, a brief burst of tremendously rapid expansion occurring early in the Bang, resulted in a universe whose subsequent more leisurely expansion was in no need of tuning. Yes, the expansion speed after inflation had to fall inside very narrow limits for stars to be able to form but, it is often said, inflation more or less forced the speed to fall inside those limits. To this we might reply that inflation itself stood in need of very delicate tuning, yet we could instead simply drop the expansion speed from the list, pointing out that plenty of other items remained on it. [We could find the list impressive without claiming knowledge of all logically possible universes, most of them presumably with properties very distant from those of our universe, and of what proportion of these possible universes were life-permitting. Imagine that a bullet hits a fly surrounded by a large empty

area. The bullet's trajectory needed fine tuning to achieve this result, which can help to show that a marksman was at work. It can help to show it regardless of whether distant areas are all of them so covered with flies that any bullet striking them would hit one. The crucial point is that *the local area* contained just the one fly.]

It is sometimes held that we could avoid belief in fine tuning by believing instead in various exotic life-forms (Feinberg and Shapiro 1980). Rather than being based on chemistry (which means, in effect, on electromagnetism), intelligent organisms might be based on the strong nuclear force so that they could inhabit neutron stars. Alternatively, they might be plasma beings inside the sun, or complex patterns in frozen hydrogen, or intricately organised interstellar gas clouds. None of these intelligent life-forms would dream of arguing that chemistry, something possible only through fine tuning, was necessary to intelligent life. Yet people willing to believe in such strange life-forms could still say that much tuning was required for there to be neutron stars, suns, planets covered with frozen hydrogen or interstellar clouds. It is often reasoned (see Rozental 1988 in particular) that a universe taken at random from among the apparent physical possibilities would almost certainly lack such objects. It would be likely either to collapse in a Big Crunch after a very brief, intensely hot career, or else to expand so fast that any matter which it contained soon became too rarefied to form clouds, let alone stars and planets. It could well consist almost entirely of light rays or black holes.

Various other doubts about fine tuning and the need to explain it are fairly easily dismissed. For instance, it seems wrong to reason (i) that no possible evidence of design could have any force *because we can see only the one universe*, and therefore cannot know whether its patterns are at all extraordinary, or (ii) that all possible patterns *would be equally probable*, just like all possible hands of cards, or (iii) that *probabilities depend on repetitions being possible* whereas the universe is unrepeatable: a universe can occur only once. Such reasoning delivers the strange conclusion that not even the words "God designed all this", written on every rabbit, tree and snowflake, could be in special need of explanation. It forgets such facts as that a hand of cards which includes four queens, four kings and four aces can, thanks to the possibility of cheating, be considerably more probable than many others. And the claim "that a universe can occur only once" itself runs into trouble: see the next section, "Multiple Universes".

Again, it would surely be wrong to protest that fine tuning needs no explanation "since if the universe hadn't been tuned in appropriate ways, then there'd not have been anybody to consider the affair". What would you think of the man who, untouched by all the bullets of a fifty-marksman firing squad, failed to suspect that the marksmen had wanted to miss him, commenting instead "that he'd otherwise not be alive to discuss anything"?

A divine designer's influence might be limited to creating a universe with life-permitting laws and fine-tuned force strengths, particle masses, etcetera. Some (e.g., Ward 1996b) argue, however, that God could be expected to influence the course taken by Darwinian evolution, ensuring that various crucial events occurred in favourable ways. [Assuming that quantum physics makes the reign of natural law into something only probabilistic, not deterministic, then there is actually some difficulty in deciding whether God, by ensuring that such and such an event occurred in the most favourable of various ways which quantum physics allowed, would be "intervening miraculously".] Also, it is sometimes thought that God

created a universe in which the evolution of intelligent, truly conscious minds, and the workings of those minds during free decision-making, are at least in part inexplicable by physical laws (Swinburne 1986).

Multiple Universes

If by "universe" you mean Absolutely Everything, then there must be just a single universe. However, people often picture the cosmos as containing numerous huge domains, very varied in their characters and largely or entirely isolated from one another. Now, "universes" is what they typically call them nowadays. Understood in this way, universes can be used to explain any observed fine tuning without introducing a divine designer. While most universes could well be hostile to intelligent life, observers would clearly have to find themselves in the life-permitting ones.

Numerous universe-generating mechanisms have been proposed (Atkins 1992; Barrow and Tipler 1986; Barrow 1988; Halliwell 1992; Leslie 1989; Leslie ed. 1990; Linde 1990; Rees 1997; Rozental 1988; Smolin 1997). Universes could be successive cycles of an oscillating cosmos (Big Bang, Big Crunch, Big Bang, etcetera). They could be huge areas of a gigantic, perhaps infinite cosmos. They could be the "worlds" of Many-Worlds Quantum Theory, which says that reality continually branches, every alternative allowed by quantum laws occurring in some branch or other. They could be quantum fluctuations in a pre-existing space or in a space-time foam. Or they might "quantum-tunnel from nothing" (if that makes sense), or bud off from other universes, or form bubbles in which expansion speeds had slowed inside a cosmos which was perpetually inflating. They could even be born in the depths of black holes, then expanding into spaces of their own without disturbing their parent universes.

All this could help an opponent of divine design only if the universes differed widely so that sooner or later, somewhere, one or several of them might be expected to be fine tuned in life-permitting ways. Difference-generating mechanisms are easily invented, however. An early suggestion was that an oscillating cosmos would "forget" its properties in the quantum-fuzzy depths of its Big Crunches. Among many later suggestions, perhaps the most plausible is that tiny domains form in the cosmos like ice crystals on a pond, each domain then being enormously enlarged (to "universe size") by cosmic inflation so that living beings deep inside it cannot see the other domains. Wide differences between the domains can readily be attributed to *scalar fields*. Having no directionality such as makes a magnetic field obvious to a compass needle, scalar fields are hard to detect, yet it is now standardly considered that such fields tore apart the initially unified forces of nature (electromagnetism and the nuclear forces, for instance) and gave them their various strengths, also causing various types of particle to become massive to differing degrees. Appearing early in the Big Bang, the scalar fields could have differed randomly from place to place because different field intensities were more or less equal in their potential energies (which are what physical systems try to minimize, like balls rolling down into valleys).

Belief in multiple domains, alias universes, and in the likelihood that they differ widely, is nowadays extremely common among cosmologists. It is considered quaint to assume that all of reality must be like the region visible to human telescopes. This would have pleased Hume's Philo, who cautioned us against any such assumption (Hume 1779). A reality consisting of infinitely many, very varied universes may

actually be thought simpler than the alternatives. Why, after all, should a universe-generating mechanism operate only a limited number of times? And if it operated again and again, why should it produce identical results on each occasion?

This does not mean that belief in divine design must be abandoned. While the manner in which our universe appears "fine tuned for permitting life to evolve" could encourage a story about many widely differing universes, it could equally well support belief in a designer. The fact remains, though, that the designer does not supply the sole plausible explanation for any fine tuning. This is largely because universes, like ice crystals, could differ widely while remaining identical in the fundamental laws they obeyed.

The Anthropic Principle

In the early 1970s, Brandon Carter stated what he called "the anthropic principle": that what we can expect to observe "must be restricted by the conditions necessary for our presence as observers" (Leslie ed. 1990). Carter's word "anthropic" was intended as applying to *intelligent beings in general*. The "weak" version of his principle covered the spatiotemporal districts in which observers found themselves, while its "strong" version covered their universes, but the distinction between spatiotemporal districts and universes, and hence between the weak principle and the strong, *could not always be made firmly*: one writer's "universe" could sometimes be another's "gigantic district". Moreover, the necessity involved was never -- not even in the case of the "strong anthropic principle" -- a matter of saying that some factor, for instance God, had made our universe *utterly fated* to be intelligent-life-permitting, let alone intelligent-life-*containing*. However, all these points have often been misunderstood and, at least when it comes to stating what words mean, errors regularly repeated can cease to be errors. Has Carter therefore lost all right to determine what "anthropic principle" and "strong anthropic principle" really mean? No, he has not, for his suggestion that observership's prerequisites *might set up observational selection effects* is of such importance. Remember, it could throw light on any observed fine tuning without introducing God. Everything is thrust into confusion when people say that belief in God "is supported by the anthropic principle", meaning simply that they believe in fine tuning and think God can explain it. As enunciated by Carter, the anthropic principle does not so much as mention fine tuning.

Being aware of possible "anthropic" observational selection effects can encourage one set of expectations, and belief in God another set. If suspecting that Carter's anthropic principle has practical importance, you will be readier to believe (i) that there exist multiple universes and (ii) that their characteristics have been settled randomly, some mechanism such as cosmic inflation ensuring that all was settled in the same fashion throughout the region visible to our telescopes. True, the believer in God can accept these things too, yet he or she may feel far less pressure to accept them. Even if there existed only a single universe, God could have fine tuned it in ways that encouraged intelligent life to evolve.

A possible argument for preferring the God hypothesis runs as follows. A physical force strength or elementary particle mass can often seem to have required tuning to such and such a numerical value, plus or minus very little, *for several different reasons*. Random variations from universe to universe might

explain why it took any particular value somewhere or other, yet how could they account for the fact that one and the same value satisfied many different requirements? Why is such consistency possible? Why does electromagnetism, for example, not need to have one strength to allow atoms to be stable, and another strength for stars to burn at a life-encouraging rate, and yet another to permit carbon (quite probably crucial to life) to be produced plentifully? Here a religiously minded physicist could think in terms of many possible fundamental theories, God selecting a theory which permitted life's requirements to be fulfilled without contradictions.

God

It is sometimes protested that God cannot adequately explain the existence and orderliness of the cosmos, for the following reason: that God's own existence, and the orderliness which would have to characterize his mind before he could bring order to anything else, would in turn need explanation. How might theists reply?

It might seem that an infinitely knowledgeable divine mind would be infinitely harder to explain than any finite cosmos or than one which, while infinite in both time and space, was still limited in, for example, the number of its dimensions -- whereas the divine mind would know every possible universe, including those with a million billion dimensions. Again, such a mind might be thought to go infinitely far beyond any evidence we could collect. [If you saw a pound of butter rising on a balance pan, would you conclude that it was being outweighed by an infinitely heavy weight?] It can be replied, however, that an all-knowing mind *would be in a way extremely simple*, as can be seen by how a single word -- "everything" -- can describe what it knows; and such an all-knowing mind, it could next be said, might well be expected to create a complex cosmos for the sake of all the living beings in it. But while the combination of the cosmos and a divine creator might be considered simple for reasons such as these, there remains the difficulty that a complete and utter blank could well be thought simpler still. Note that the "ontological argument" which tries to prove God's existence from his mere notion is generally dismissed today. There would seem to be no contradiction in the idea that a perfect being *was a logical possibility only*, not an actual existent.

On the other hand, logical possibilities can be *real* without being *actual existents*, and once this is appreciated their reality can be seen to be guaranteed. If God or anything else is a logical possibility, then that is an unconditional fact. It is eternally the case, non-fictitious, genuine, real, that this or that is indeed a logical possibility. [How odd it would be to fancy that thinkers, for instance, could become *really logically possible* only after they had come into existence and developed logics! Before thinkers evolved, the sun which helped them to evolve was logically possible, surely. It was not like a round square, and nor were the thinkers.] Furthermore, some logical possibilities are such that their failure to be actualized -- their absence from the realm of actually existing things -- can be thought needed, ethically required, in an eternal and unconditional way. Take the case of a logically possible world consisting solely of people in torment. To declare that the actual existence of such a world would be evil is the same as calling its non-existence needed or ethically required, and it could seem utterly wrong to add "just so long as there exists somebody who is contemplating the affair, somebody with a moral duty to prevent the existence of such a

world". Likewise, to say that a world full of interest and happiness would be a good thing is to say that the existence of such a world is ethically required, and would (presumably) be so regardless of whether anyone ever contemplated that fact. Might we point to these matters when trying to account for the existence of God or of the cosmos?

Certainly, the concept of an ethical requirement is distinct from that of a requirement which is fulfilled; yet cannot a flower be red, despite how the concept of redness differs from that of being a flower? Flowers are "the right sort of reality" for being red. They are "in the right ball park". The idea of a red flower is not conceptually confused. And rather similarly, it can be argued, a requirement for the existence of something, for instance a divine mind, might be the right sort of reality to carry responsibility for this something's actual existence, *even if it were an ethical requirement*. There is no conceptual confusion here. So long as it was recognized that no logical necessity was involved, it could actually be suspected that some ethical requirement (or consistent set of requirements) carried such responsibility *necessarily*. Compare, perhaps, the necessity that red as we experience it (say, in an after-image produced by a bright light) is nearer to purple than to blue. This can be argued to be an absolute necessity, without being a logical necessity. Again, it can be argued that various states of mind are necessarily in themselves worth having, without this being logically demonstrable.

An ethical requirement is, at any rate, what Ewing (1973, chapter 7) proposes as the ground of a divine person's reality. As a matter of necessity -- necessity which is absolute despite not being provable by logicians -- the unconditionally real ethical need for a divine mind to exist is adequate, Ewing suggests, to ensure its eternal existence. And a very similar theme is central to the long neoplatonic tradition in theology. This treats "God" as the name not of any mind, but of the supposed fact that the world owes its existence to its ethical requiredness (Leslie 1979, 1989 chapter 8; Mackie 1982, chapter 13; Levine ed. 1997; Tillich 1953-63).

Ewing's picture may at times be hard to separate from that of the neoplatonists. For suppose we joined the pantheist Spinoza in thinking (to take an interpretation of his writings which can seem to make sense of them) that all the complexities of the cosmos are simply the complex thoughts of a divine mind, so that your consciousness and mine, or the consciousness of a bird or of a bat, is just the divine knowledge of what it is like -- exactly what it feels like -- to be particular living beings with strictly limited power and knowledge. Suppose we also adopted the belief (which can again be the suggested by Spinoza's writings) that the divine mind exists *because this is ethically required*, rather than through any logical necessity which has nothing to do with good or bad. There might then be no real difference, so far as concerned the situation in which we believed, between our declaring (a) that God was a divine mind, as Ewing thought, or (b) that God was the cosmos, as Spinoza thought, or (c) that, as neoplatonists think, God is a creative force or principle: the principle that a supreme ethical requirement (or consistent set of requirements) is responsible for the existence of the cosmos.

Sure enough, this disregards distinctions which have appeared important to many people. For many variants on pantheism and on neoplatonism, consult Forrest (1996), Laird (1940), Levine (1994) and Whitehead (1938). Sometimes Spinoza's eternal and all-knowing divine mind is replaced by one which constantly improves its power and its knowledge. Sometimes God is thought to some degree separate

from the cosmos, instead of just being the cosmos or the creative ethical requiredness of the cosmos. Sometimes the notion that the natural world is alive, and that it strives after value, plays a greater role than the idea that value is actually achieved. Sometimes identification of the cosmos with God is oddly taken to imply the illusoriness of most of the things we think we see.

Time and the Human Mind

Believing (as Spinoza appears to) that we are all parts of a divine mind, we might think we could answer the theological problem of evil: the problem of why the cosmos contains so many items which can seem so very unsatisfactory. Knowing everything, or (if this is different) everything in the least worth knowing, a divine mind could know many things which might be very little worth knowing if they were taken in isolation. It might know, for example, not only all the detailed structure of innumerable universes, but also exactly how it would feel to be each of the intelligent living beings in those universes. Knowing this eternally, it could none the less know what it felt like to be engaged in actual struggles, in constant ignorance of what the next moment would bring: see Williams (1951) to gain further insight into the theory about time's flow which would be involved here, a theory often adopted because it appears to reflect Einstein's theory of relativity. While such items of knowledge might be nowhere near the best that it possessed, the divine consciousness could still be better for not being ignorant of them.

A competing approach to the problem of evil is often preferred, though. Instead of viewing themselves as elements of a divine mind, many religious people think they exist separately from that mind and are given absolutely free choice of whether to join it in a heavenly hereafter.

Absolutely free choice, as they conceive it, depends on time's flowing in a way which Einstein rejected when he wrote of the world as having "a four-dimensional existence". [Suppose that you are about to choose "with absolute freedom" whether to give up smoking. It cannot already be true, they say, that "at points a little further along the fourth dimension" you are lighting a cigarette.] It probably also demands a fairly strong division between the operations of human minds and those of material objects.

Extraterrestrials

Belief in God need not involve believing that our universe is crammed with intelligent life from side to side and from start to finish, or that God will save humans from driving themselves to extinction through polluting their planet or by germ warfare, or perhaps (Rees 1997, chapter 12) by experiments at extremely high energies. The fraction of our universe which we can see contains many hundred million trillion sun-like stars. Even if only a small proportion of these had hospitable planets, intelligent beings could well exist in tremendously many places. What is more, the universe as a whole might be very much larger, perhaps infinitely larger, and there may be up to infinitely many other universes. Wishing for intelligent beings to exist in large numbers, a divine person could have them without ensuring that humans survived long enough to colonize their entire galaxy.

We have failed to detect extraterrestrials, although calculations suggest that an intelligent species could spread across its galaxy in a few million years. This, together with our observed position in the midst of a population explosion, might reinforce whatever other grounds we had for thinking that rapid extinction, not galactic colonization, was the likely fate of an intelligent species. Suppose human extinction occurred during the next century. Roughly ten per cent of all humans who had ever been born would be alive when it occurred. On the other hand, if humans spread right across their galaxy then perhaps well under a thousandth of one per cent would have lived during that period. This may be found disturbing. [The point is that one ought to hesitate before adopting theories whose truth would have made one's own observations highly unlikely, when other theories would have made them fairly likely. It is a point first noted by Brandon Carter; for a discussion drawing very pessimistic conclusions from it, see Gott (1993). Various reasons in favour of guarded optimism are given by Leslie (1996), reasons centred on the fact that our universe is probably indeterministic, so that the number of humans who will ever have lived is something which has not yet been fixed.]

For religious people asking whether humans will soon be extinct, here is a point to bear in mind. If you hope to solve the theological problem of evil, then you have to assume that there are strong reasons *against* divine intervention to prevent calamities.

Bibliography

- Atkins, P.W. 1992. *Creation Revisited*, Oxford: W.H. Freeman.
- Ayer, A.J. 1970. "Chance," chapter 7 of *Metaphysics and Common Sense*, Freeman, Cooper: San Francisco.
- Balashov, Y.V. 1991. "Resource Letter AP-1: The anthropic principle," *American Journal of Physics*, **59**: 1069-1076.
- Barrow, J.D., and Tipler, F.J. 1986. *The Anthropic Cosmological Principle*, Oxford: Clarendon Press.
- Barrow, J.D. 1988. *The World within the World*, Oxford: Clarendon Press.
- Bertola, F., and Curi, U., eds. 1993. *The Anthropic Principle*, Cambridge: Cambridge University Press.
- Bohm, D., and Hiley, B.J. 1993. *The Undivided Universe*, London: Routledge.
- Burrill, D.R. 1967. *The Cosmological Arguments*, New York: Doubleday.
- Carr, B.J., and Rees, M.J. 1979. "The anthropic principle and the structure of the physical world," *Nature*, **278**: 605-612.
- Carter, B. 1989. "The anthropic principle: self-selection as an adjunct to natural selection," 185-206 of S.K.Biswas et al., eds., *Cosmic Perspectives*, Cambridge: Cambridge University Press.
- Craig, W.L. 1979. *The Kalam Cosmological Argument*, London: Macmillan.
- Craig, W.L. 1980. *The Cosmological Argument from Plato to Leibniz*, London: Macmillan.
- Craig, W.L., and Smith, Q. 1993. *Theism, Atheism and Big Bang Cosmology*, Oxford: Clarendon Press.
- Davies, P.C.W. 1982. *The Accidental Universe*, Cambridge: Cambridge University Press.
- Davies, P.C.W. 1992. *The Mind of God*, New York: Simon and Schuster.

- Ellis, G. 1993. *Before the Beginning*, London: Bowerdean Press/ Marion Boyers.
- Ewing, A.C. 1973. *Value and Reality*, London: Allen and Unwin.
- Feinberg, G., and Shapiro, R. 1980. *Life Beyond Earth*, New York: William Morrow.
- Flew, A. 1966. *God and Philosophy*, London: Hutchinson.
- Forrest, P. 1996. *God without the Supernatural*, Ithaca: Cornell University Press.
- Gott, J.R. 1993. "Implications of the Copernican principle for our future prospects," *Nature*, **363**: 315-319.
- Grünbaum, A. 1990. "Pseudo-creation of the big bang," *Nature*, **344**: 821-822.
- Halliwell, J.J. 1992. *Quantum Cosmology*, Cambridge: Cambridge University Press.
- Hassing, R.F., ed. 1997. *Final Causality in Nature and Human Affairs*, Washington: The Catholic University of America Press.
- Hawking, S.W. 1987. "Quantum cosmology," pages 631-651 of S.W.Hawking and W.Israel, eds., *Three Hundred Years of Gravitation*, Cambridge: Cambridge University Press.
- Hawking, S.W. 1988. *A Brief History of Time*, New York: Bantam Books.
- Henderson, L.J. 1913. *The Fitness of the Environment*, New York: Macmillan.
- Hepburn, R.W. 1967. "Cosmological Argument for the Existence of God," 234-237 of P.Edwards, ed., *The Encyclopedia of Philosophy, Vol. 2*, New York: Macmillan.
- Hetherington, N.S., ed. 1993. *Encyclopedia of Cosmology*, New York: Garland.
- Hoyle, F. 1950. *The Nature of the Universe*, Oxford: Blackwell.
- Hume, D. 1739. *A Treatise of Human Nature*, London.
- Hume, D. 1779. *Dialogues concerning Natural Religion*, London.
- Jeans, J. 1930. *The Mysterious Universe*, London: Macmillan.
- Kragh, H. 1996. *Cosmology and Controversy*, Princeton University Press: Princeton.
- Kuiper, B.H., and Brin, G.D. 1989. "Resource Letter ETC-1: Extraterrestrial Civilization," *American Journal of Physics*, **57**: 12-18.
- Laird, J. 1940. *Theism and Cosmology*, London: Allen and Unwin.
- Leslie, J. 1979. *Value and Existence*, Oxford: Blackwell.
- Leslie, J. 1989. *Universes*, London and New York: Routledge.
- Leslie, J., ed. 1990. *Physical Cosmology and Philosophy*, New York: Macmillan.
- Leslie, J. 1996. *The End of the World: the science and ethics of human extinction*, London and New York: Routledge.
- Levine, M. 1994. *Pantheism*, London and New York: Routledge.
- Levine, M., ed. 1997. *Pantheism* (special issue, volume 80, of *The Monist*).
- Linde, A.D. 1990. *Inflation and Quantum Cosmology*, San Diego: Academic Press.
- Mackie, J.L. 1982. *The Miracle of Theism*, Oxford: Oxford University Press.
- Margenau, H., and Varghese, R.A., eds. 1995. *Cosmos, Bios, Theos*, La Salle: Open Court.
- Matthews, C.N., and Varghese, R.A., eds. 1995. *Cosmic Beginnings and Human Ends*, Chicago: Open Court.
- McMullin, E., ed. 1985. *Evolution and Creation*, Notre Dame: University of Notre Dame Press.
- Milne, E.A. 1952. *Modern Cosmology and the Christian Idea of God*, Oxford: Clarendon Press.
- Munitz, M.K., ed. 1957. *Theories of the Universe*, New York: Macmillan.
- Munitz, M.K. 1986. *Cosmic Understanding*, Princeton: Princeton University Press.
- Peacocke, A.R. 1979. *Creation and the World of Science*, Oxford: Clarendon Press.

- Parfit, D. 1992. "The puzzle of reality," *Times Literary Supplement*, July 3: 3-5.
- Pius XII (Pope). 1952. "Science and the catholic church," *Bulletin of Atomic Scientists* **8**: 142-146, 165.
- Polkinghorne, J. 1986. *One World: the Interaction of Science and Theology*, Princeton: Princeton University Press.
- Rees, M. 1997. *Before the Beginning: Our Universe and Others*, Reading, Mass.: Addison-Wesley.
- Rescher, N. 1984. *The Riddle of Existence*, Lanham: University Press of America.
- Rolston, H. 1987. *Science and Religion*, New York: Random House.
- Rozental, I.L. 1988. *Big Bang, Big Bounce*, Berlin: Springer-Verlag.
- Russell, J.R., Murphy, N., and Isham, C.J., eds. 1996. *Quantum Cosmology and the Laws of Nature,* Notre Dame: University of Notre Dame Press.
- Russell, R., Stoeger, W., and Coyne, G., eds. 1988. *Physics, Philosophy and Theology*, Notre Dame: University of Notre Dame Press.
- Smart, J.J.C. 1989. *Our Place in the Universe*, Oxford: Blackwell.
- Smart, J.J.C., and Haldane, J.J. 1996. *Atheism and Theism*, Oxford: Blackwell.
- Smolin, L. 1997. *The Life of the Cosmos,* New York: Oxford University Press.
- Swinburne, R. 1968. "The Argument from Design," *Philosophy*, **43**: 199-211.
- Swinburne, R. 1979. *The Existence of God*, Oxford: Clarendon Press.
- Swinburne, R. 1986. *The Evolution of the Soul*, Oxford: Clarendon Press.
- Tillich, P. 1953-63. *Systematic Theology*, 3 vols., London: Nisbet.
- Vilenkin, A. 1982. "Creation of Universes from Nothing," *Physics Letters*, **117 B**: 25-28.
- Ward, K. 1996a. *Religion and Creation*, Oxford: Clarendon Press.
- Ward, K. 1996b. *God, Chance and Necessity*, Oxford: Oneworld Publications.
- Whitehead, A.N. 1938. *Modes of Thought*, New York: Macmillan.
- Wigner, E.P. 1960. "The unreasonable effectiveness of mathematics in the natural sciences," *Communications in Pure and Applied Mathematics*, **13**: 1-14.
- Williams, D.C. 1951. "The Myth of Passage," *Journal of Philosophy*, **48**: 457-472

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

causation: in science | consciousness | cosmological argument | Descartes, René | determinism, causal | Einstein, Albert: philosophy of science | evil, problem of | existence | free will | God | Hume, David | infinity | Kant, Immanuel | laws of nature | Leibniz, Gottfried Wilhelm | materialism | metaphysics | mind: philosophy of | modality, metaphysics of | monism | natural religion | Newton, Isaac: views on space, time, and motion | panpsychism | pantheism | physicalism | Platonism: in metaphysics | possible worlds | principle of sufficient reason | quantum mechanics | rationalism vs. empiricism | religion: philosophy of |

science, philosophy of | space and time: being and becoming in modern physics | Spinoza, Baruch [Benedict] | theism | time | truth: necessary vs. contingent | Whitehead, Alfred North

John Leslie
johnlesl@uoguelph.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 2, 1998
Content last modified: July 2, 1998

Immutability

The doctrine of divine immutability (DDI) asserts that God cannot undergo real or intrinsic change in any respect. To understand the doctrine, then, we must first understand these kinds of change. Both "intrinsic" and "real" (in the relevant sense) are hard notions to elucidate. I cannot here attempt anything like a full account of them. I instead provide very rough characterizations, which would be acceptable on a wide variety of competing accounts of these notions.

1. Kinds of Change

A change is real if and only if it makes on its own a real difference to the world. Real change is (roughly) change involved in causing something, change an item is caused to undergo or change not "logically parasitic" on change in other things. (These "or"s are inclusive, not exclusive -- one change can satisfy more than one clause of this account.) Smith's kicking me involves changes in Smith. Kicking me makes on its own a real difference in Smith. When Smith kicks me, I undergo various changes which count as real, intuitively -- being kicked makes many real differences in me. On the other hand, when I become shorter than Smith because Smith grows, Smith changes really but I do not. Gaining height makes a real difference in Smith, but Smith's gaining height doesn't make a real difference in me. Rather, my becoming shorter than Smith in this way is logically parasitic on Smith's growing. It is simply a logical consequence of a real change in something else, not a real change itself.

Coming at the notion another way, a change is real if and only if it is an event which logically could be either the only event that occurs in an entire universe or else the sum of all events in a universe, while a change is a "logical parasite" if and only if this is not true of it. Consider Smith's kicking me. This looks like it could be the only occurrence in a universe. It is logically possible that the universe simply sits there wholly static, then Smith kicks me, then the universe winks out of existence before anything else happens. But actually, when we look more carefully, we see that in this case, the kick is not the only

change. For it has parts which are shorter or spatially smaller changes. In the kick, for instance, Smith's leg first moves an inch, then another inch, etc., and each movement consists spatially of discrete movements by various muscles. The kick is not the only change, but in the odd universe I described, it is the sum of all changes in a universe. No change with parts can be the only change in a universe, strictly speaking. On the other hand, it is not logically possible that my becoming shorter than Smith in the way described occur alone or be a sum of all changes occurring in a universe. For it to occur, Smith's growth must also occur. And Smith's growth is not a part of it. So changes are real if and only if they are logically independent events in the world's history. If God cannot change really, then, nothing can so act on him as to change him, his actions do not change him, and no change in God could be the only event in a universe. If God cannot change really, then if had there been nothing other than God, there would have been no change at all, of any sort.

Intrinsic changes are changes like learning or expanding, which (roughly) occur entirely within the changing item -- which could occur if the universe ended at the item's "skin." Putting this more carefully, a change is intrinsic to the changing thing if and only if its occurring does not imply the existence of anything "outside" the boundaries of the changing thing. Changes which are not intrinsic are extrinsic. All changes in relations to other things are extrinsic. For instance, if a dog moves to my left, I become a man with a dog on his left. This does not take place wholly within me. It also involves the dog, who exists outside me; it implies the existence of the dog.

Changes can be intrinsic or extrinsic in spatial or temporal terms. The dog's coming to be on my left is a spatially extrinsic change in me. If Smith's guardian angel gets angry at Smith's kicking me an hour earlier, this change has no spatial aspect -- angels have no spatial location or boundaries. But it is a temporally extrinsic change in the angel, for it implies the occurrence of an event, the kicking, which occurred outside the temporal boundaries of the becoming-angry. Extrinsic changes aren't "real" in the sense above. I change extrinsically when a dog comes to be on my left. The dog does not so act on me at to make me have it on my left: the dog does nothing to me at all. And my becoming a man with a dog on his left could not be the only event in a universe. For it to happen, a dog must also move. But to say that extrinsic changes aren't "real" in the above sense is not to say that they don't occur. It is merely to classify them.

DDI lets God change extrinsically. On DDI, God can, e.g., come to have new relational properties involving other things which are not due to those things' acting on him. Suppose that at *t*, Quine worships God. Then at *t*, God comes to have a new relational property, being-worshipped-by-Quine. This is clearly an extrinsic change, since its occurring in God implies the existence of someone "outside" God, namely Quine. God does not (let's say) do anything to cause this. It's a matter of Quine's free will. Quine doesn't do anything to God by worshipping him. And this change in God is a logical parasite of the real changes in Quine which constitute his worshipping God. It's not possible that becoming worshipped by Quine be the only change in a universe. Quine's worshipping must occur too if this one does. DDI thus lets God become Quine-worshipped. It rules out only real and intrinsic changes.

All of this raises a question: why do only real, intrinsic changes matter? It would be silly to ascribe to friends of DDI some sort of pre-given general antipathy against real, intrinsic change. Disliking change

makes as little sense as disliking qualities or relations. DDI's friends simply followed various specific arguments, discussed below, where they seemed to lead. And they seemed to lead to a denial that God can change really and intrinsically, but not to a denial of all divine change *tout court*. The arguments which led philosophers to exempt God from real, intrinsic change all had to do with ways involvement in change might (they thought) make God less perfect. It isn't plausible that merely extrinsic change could make something more or less perfect. I am exactly as impressive a man with a dog on my left as I am without a dog on my left.

2. Immutability vs. Impassibility

DDI is sometimes conflated with the doctrine of divine *impassibility*, which asserts that nothing external can affect God -- that nothing external can cause God to be in any state, and in particular can cause him to feel negative emotions like grief. Actually, DDI neither implies nor is implied by divine impassibility. Something could be impassible but mutable if it could change itself, but nothing else could change or affect it. God could be immutable but passible. For he could be changelessly aware of events outside himself -- perhaps even caused to be aware of them by the events themselves -- and due to them changelessly feel such responsive emotions as grief. But he would feel them without change, and so always feel them. If temporal, such a God would grieve for us before, while and after we suffer what he grieves for. There is nothing counter-intuitive in this. It's standard theism to hold that God has full foreknowledge of what is to befall us: he sees our pain before we feel it, not just while we feel it, and so grieves it beforehand if He ever grieves it at all. There would be no difference in the quality of God's grief before and while the pain occurs. For were there anything about it He did not know beforehand, the foreknowledge would not be full, and full knowledge beforehand should elicit the same reaction as full knowledge during. Likewise, it's standard theism to hold that God is cognitively perfect. If he is so and exists in time, He has a past to recall and so has perfect memory. If God perfectly remembers your pain, it is as fresh for him years later as it was while it occurred, and if he perfectly loves you, perhaps he never gets over it. So we can make sense of unchanging grief; if God does grieve, we might well expect it from a God with full foreknowledge, cognitive perfection and a perfect affective nature. If He is timeless, an immutable but passible God would just timelessly suffer for us -- responsively, i.e., because of our pain. The case would be just as if God were temporal, save that His knowledge would not be temporally located and so would not literally involve either foreknowledge or memory. So whether God is temporal or timeless, DDI implies nothing odd here. And it need not "depersonalize" God as some feel impassibility would.

Still, it is surprising that Western theists have held DDI. For Western Scriptures seem to conflict with the full DDI. Some Scriptural texts depict human sin as making God sadder than he was (e.g., Gen. 6:6), then bringing God to new decisions, e.g., to flood the world. According to *John*, "the Word became flesh" (1:14), i.e., God took on a human nature he did not always have. So Western theism's Scriptural roots seem to deny DDI. Yet by the first century A.D., DDI was central to the main theory of God's nature, "classical theism." In such "classical theist" writers as Augustine and Aquinas, being immutable makes God eternal, and eternality is God's distinctive mode of being. So DDI is at the roots of such writers' understandings of God's nature. And "classical theism" ruled the theological roost till the 19th century. So

one wonders: what made DDI so attractive for so long?

3. The Case for Immutability

For one thing, the Scriptural witness is not really so thoroughly on the side of divine real intrinsic change. Much that Scripture says of God is clearly metaphor. And it is not hard to show that Old Testament texts which ascribe change to God could be speaking metaphorically. As I note later, one can parse even the Incarnation in ways which avoid divine real or intrinsic change. Standard Western theism clearly excludes many sorts of change in God. Western theists deny that God can begin or cease to be. If God cannot, He is immutable with respect to existence. Nothing can change its essential nature: a thing's essence is by definition a property (or set of them) it cannot fail to have. For Western theists, God is by nature a spirit, without body. If he is, God cannot change physically -- he is physically immutable. So it is not clear that the Western God could undergo other than mental changes -- changes in knowledge, will, or affect. Further, Scripture amply supports the claim that God is in all respects perfect. In conjunction with certain other Scriptural claims, God's perfection seems to rule out many sorts of mental change.

This is a broader topic than can be tackled here; instead, a few examples involving God's knowledge will be examined. If perfect, God is all-knowing. If God learns something new, then before that he was not all-knowing, unless the new item could not have been foreknown. Only free beings' future actions and what depends on these are even *prima facie* beyond God's foreknowledge. But Scripture is full of claims that God foreknows our free actions. So if Scripture calls God's knowledge perfect and asserts that it includes foreknowledge, there is a sort of fact about which God is omniscient and his knowledge does not change. Suppose that today God knows that I will finish this article tomorrow and tomorrow God knows that I *am* finishing the article. There is a fact God knows both days, that on this particular day, referred to one day as "tomorrow" and another as "today," I finish the article. Such facts involve no real tense: they are ‘tenseless’ facts. If God has foreknowledge even of free creaturely actions, he always knows all tenseless facts. His knowledge of these does not change, and if He necessarily foreknows, then his knowledge of these is immutable. Our own knowledge of these constantly changes. I do not know till tomorrow that I tenselessly-finish the article tomorrow, for I do not know this until I know that I *am* finishing it, a truth which is present-tensed rather than tenseless. Similarly, if God always knows all truths of mathematics and logic, his knowledge of these never changes, and here too it is plausible to suppose that His knowledge of these cannot change. For these truths are necessary. Truths of mathematics and logic cannot be or become false. So God could change with respect to his belief in these only if he could at some time hold a false belief. If God is necessarily omniscient, he cannot.

Unlike us, then, God has a constant, unchangeable store of tenseless knowledge about all of history, mathematical knowledge, logical knowledge and indeed knowledge of any other sort of necessary truth -- and this on Scriptural grounds. It is a small step indeed from divine perfection to necessary divine perfection. For it is surely more perfect to be unable not to be perfect than to be perfect but able not to be. Again, it is a small step from foreknowledge to necessary foreknowledge: wouldn't the latter be more perfect? So God's necessary and tenseless historical knowledge -- tenseless knowledge of contingent truths -- looks to be unchangeable, given only a small step beyond Scripture. We can take this a step

further. The only sort of knowledge we've left out so far is knowledge of tensed contingent truths -- such truths as that tomorrow I *will* finish the article or yesterday I *did* finish the article. If God always knows the tenseless correlates of these truths -- e.g., that on March 27, 2002, it is the case that I tenselessly-finish the article on March 28, 2002 -- then his contingent knowledge changes only in ways for which the passage of time accounts. For which tensed truth God knows -- that I will finish, am finishing or did finish -- depends simply on what time it is. So to speak, he never has to learn about whether I finish on March 28; he merely has to learn where in time he is in relation to March 28, and this tells him what tensed propositions are true about my finishing. Thus one can make a case on what are basically Scriptural grounds that God's knowledge changes at most due to the bare passage of time.

So Scriptural considerations suggest a God at least much less changeable than we in some respects. But the roots of the full DDI are also philosophical. In thinking out their views of God's nature, Western philosophers have largely filled out the concept of God by ascribing to him the properties they thought he must have to count as absolutely perfect. God's perfection seems to rule out many sorts of change, as we've just seen. More general arguments from perfection convinced classical theists that God cannot change in any way.

Plato argued for the full DDI. He asserted that a god is "the… best possible" in virtue and beauty. Virtue is a perfection of mind. Beauty is a non-mental perfection. So Plato's examples are probably meant to do duty for all mental and non-mental perfections, i.e. all perfections *simpliciter*. If a god is already the best possible in these respects, Plato reasoned, a god cannot change for the better. But being perfect includes being immune to change for the worse -- too powerful to have it imposed without permission and too good to permit it. Thus a god cannot improve or deteriorate. Plato's argument had great historical influence. But it overlooked the possibility of changes which neither better nor worsen. If one first knows that it is 11:59:59 and then knows that it is midnight, is one the better or the worse for it? If the best possible state of mind includes omniscience, then perhaps it includes constant change in respects which neither better nor worsen God, e.g., in what precise time God knows it is. Perhaps changes to 'keep up with' time are required by a constant perfection, his omniscience. At 11:59:59 it is surely better to know that it is now 11:59:59, and then at midnight it is better to know that it is now midnight. Plato's argument does not rule out such changes.

Aristotle also contributed to acceptance of the full DDI. For many medieval theists accepted Aristotle's case for God's existence. Aristotle's *Physics* reasoned that if change occurs, it has a final source, an eternally unchanged changer. Aristotle's *De Caelo* added that something is eternally unchanged only if unchangeable. Later theists thought the role of first cause of change too lofty not to be God's. Writers who took Aristotle's argument or its descendants to prove God's existence found themselves committed to DDI.

Boethius also played a role in DDI's popularity. He held that being in time involves, necessarily, at least two things which are defects. For being temporal, as Boethius saw it, entailed having past and future parts of one's life. Temporal beings no longer live the past parts of their lives. They do not yet live the future parts of their lives. Both things are defects, according to Boethius. So if God is free of all defects, Boethius reasoned, then God has no past or future. What has no past or future does not change. For what

changes goes from what it was to what it then was going to be, and so has a past and a future. Hence, to Boethius, perfection required changelessness. If it does, necessary perfection -- which is better than contingent perfection, and so by perfect-being reasoning is God's -- requires being immutable. Now Boethius' reasoning requires at least some cleaning up. For some things are temporal but have neither past nor future parts, e.g. instantaneous changes. And if one's past or future is a bad time, it is not obviously a bad thing not to be living them. But at least some current philosophers thing that Boethius' argument has a kernel of truth. For it would not be a perfection of an individual to have a life lasting only a single temporal instant. Surely longer would be better, at least if there was a good chance that the longer existence would be overall a good thing. And God's life, at least, cannot contain parts which it is overall bad to live if God truly is perfect.

Boethius actually followed his reasoning about divine perfection to the conclusion that God exists outside time by his very nature -- that God can't be temporal. For whatever has neither past nor future is not located in time. But change requires existence in time. Suppose that a turnip, aging, goes from fresh to spoiled. It also then goes from fresh to not-fresh. So first "the turnip is fresh" is true, then "it is not the case that the turnip is fresh" is true. The two cannot be true at once. So things change only if they exist at at least two distinct times. Hence, if God is necessarily atemporal -- via *necessary* divine perfection -- God is necessarily changeless, i.e. immutable.

Aquinas (like Augustine) derived DDI from the deeper classical-theist doctrine of divine simplicity. If God is simple, God has no parts of any sort. Now when the turnip aged, it became partly different -- its smell and texture altered. Were this not so, no change would have occurred. But if the turnip had changed in *every* respect, it would not have been a case of change either. For it would have changed with respect to such properties as *being a turnip* and *being identical with this turnip*. And if first we have something identical with this turnip and then we have something not identical with this turnip, the turnip has not changed, but disappeared and been replaced by something else. So whatever changes must stay partly the same (else there was not change in one selfsame surviving thing). So only things with parts can change. If so, a simple God cannot change. DDI's connection with divine simplicity and the classical theist theory of God's perfection which centers on divine simplicity is one of the deepest reasons for DDI's broad historical appeal; one cannot fully explain what moved thinkers to accept DDI without also treating the motivation for the doctrine of divine simplicity. That, however, is too large a topic to broach here.

DDI, then, has a variety of religious and philosophical roots.

4. Arguments Against Immutability

There are many arguments against DDI. One that some find particularly forceful begins from the fact that God is omniscient. If God is omniscient, they say, God knows what time it is now. What he knows, then, is constantly changing, since what time it is now is constantly changing. But knowledge of what time it is is intrinsic to God. And change with respect to something intrinsic is intrinsic change. So it seems that God cannot be intrinsically immutable.

One reply is that knowing the correct time is not an intrinsic state of God's. Intrinsic states are those settled entirely within one's own skin. But then unless *P* is a truth entirely about matters within one's own skin, knowing that *P* is not an intrinsic state. For that one knows that *P* rather than believes falsely that *P* only if *P* is true, and if *P* is not a truth entirely about matters within one's own skin, whether *P* is true is at least partly settled by matters outside one's own skin. But when it is now is not a matter settled within God's own skin. What time it is now is not a fact about God alone. It is a fact about time, which is not God, and also about the entire temporal universe. Further, it arguably is not the case even that what is within God's skin settles what time it is in the sense that were nothing else than God to exist, facts about God would suffice to determine what time it is. For this would be at best a contentious claim, as it implies that God would be in time if he existed alone, without a universe. Many would say that time is an aspect of the physical universe -- no universe, no time either. Moreover, even if God existed alone and were then in time, knowing what time it is would not be a *temporally* intrinsic matter. God knows that it is now *t* just at *t*. So knowing the correct time at *t* is temporally intrinsic only if it does not entail the existence or occurrence of anything that exists at other times. But for any time after the first instant (if there was one), knowing what time it is involves knowing the temporal distance between the present time and some other time: knowing that it is noon, April 16, 2002, implies knowing a relation between that time and date and the date traditionally assigned to the birth of Christ. Thus knowing what time it is entails knowing that there was some time other than the present -- a time outside the period in which God knows that it is now this time. Knowing what time it is at any time after the first instant thus is not a temporally intrinsic state. And if this is so, it is hard to see why even knowledge that it is time's first instant would be. Thus the objection has a false premise. If God's cognitive state with respect to what time it is changed, it would not follow from this that he is not *intrinsically* immutable. This result may seem a bit hard to swallow. One wants to know *how* a change in knowledge could fail to be an intrinsic change. One answer might appeal to externalist theories of mental content. On these, my knowledge that *P* is a complex consisting of my inner mental state plus certain items in the world. Perhaps God's knowledge that it is now *t* is something like a complex consisting of God's inner cognitive state and an external component, *t*. If it is, perhaps the only change involved when God first knows that it is now *t* and then knows that it is now t^* ($> t$) is the change from *t* to t^*.

Other arguments against DDI appeal to Scripture's depiction of God as changing, e.g., in the Flood story. Facing such texts, DDI's friends defuse the appearance of divine change by appeal to doctrines less speculative and theoretical than DDI. Thus Philo argues from God's foreknowledge of the future and constancy of character that God cannot repent or feel regret, as the Flood story suggests. The Incarnation is an especially knotty problem for DDI's Christian friends. In general, these argue that all change it involved occurred in the human nature God the Son assumed rather than in God; God was eternally ready to be incarnate, and eternally had those experiences of the earthly Christ which the Incarnation makes part of his life. Through changes in Mary and the infant she bore, what was eternally in God eventually took place on earth.

Another argument against DDI appeals to God's power. Before Creation, God could assure that no universe ever existed. God has this power now only if he can alter the past. Few think he can. So events seem to change God's power. DDI's defenders reply that any change here is purely extrinsic. God has the power he always has. He has lost a chance to use it, and so we no longer want to *call* his power a power

to prevent a universe. But God is intrinsically as able as ever to do so.

A final objection goes thus: God does change extrinsically. Even full DDI grants this. But whatever changes extrinsically is in time. For different things are true of it at different times, even if the 'real' changes due to which this is so are in other things. And whatever is in time changes intrinsically -- it grows older. So DDI is false. Some would reply by denying that growing older is an intrinsic change. A more controversial response might be to deny even extrinsic change to God. This can be done by holding that God is atemporal. For if God exists at t and at a later t^*, and, at t, P is true of him and, at t^*, P is false of him, he does change extrinsically. He bears different relations to proposition P, at least. If God is atemporal, he exists neither at t nor at t^* -- his existence is not temporally located. If this is so, there are never two times such that different things are true of him at different times. Rather, all that is ever true of him is true of him timelessly. But a thing changes, even extrinsically, only if different things are true of it at different times. Perhaps, then, defending DDI requires commitment to divine timelessness.

Bibliography

- Aristotle, *De Caelo* (*De Cael.*), trans. R. Hardie and R. Gaye, in *The Basic Works of Aristotle*, ed. Richard McKeon, N.Y.: Random House, 1941.
- Aristotle, *Physics* (*Phys.*), tr. J. Stocks, in *The Basic Works of Aristotle*, ed. Richard McKeon, N.Y.: Random House, 1941.
- Boethius (1936), *The Consolation of Philosophy*, trans. H. Stewart, in *Boethius: The Theological Tractates*, trans. H. Stewart and E. K. Rand, Cambridge, Mass.: Loeb Classical Library, Harvard University Press.
- Gale, Richard (1986), "Omniscience-Immutability Arguments," *American Philosophical Quarterly* 23, 319-35.
- Hallman, Joseph (1981), "The Mutability of God: Tertullian to Lactantius," *Theological Studies* 42, pp. 373-93.
- Hartshorne, Charles (1948), *The Divine Relativity*, New Haven, Conn.: Yale University Press.
- Helm, Paul (1988), *Eternal God*, N.Y.: Oxford University Press.
- Kretzmann, Norman (1966), "Omniscience and Immutability," *Journal of Philosophy* 63, 409-421.
- Leftow, Brian (1991), *Time and Eternity*, Ithaca, N.Y.: Cornell University Press.
- Mann, William (1987), "Immutability and Predication," *International Journal for Philosophy of*

Religion 22, pp. 21-39.

- Philo, *On the Unchangeableness of God*, trans. F. Colson and G. Whitaker, in *Philo*, trans. F. Colson and G. Whitaker, Cambridge, Mass.: Loeb Classical Library, Harvard University press, 1960, vol. 3.
- Plato (1977) *Phaedo*, trans. G. Grube, Indianapolis, In.: Hackett.
- Plato (1992), *Republic*, trans. G. Grube and C. Reeve, Indianapolis, In.: Hackett.
- Sorabji, Richard (1983), *Time, Creation and the Continuum*, Ithaca, N.Y.: Cornell University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

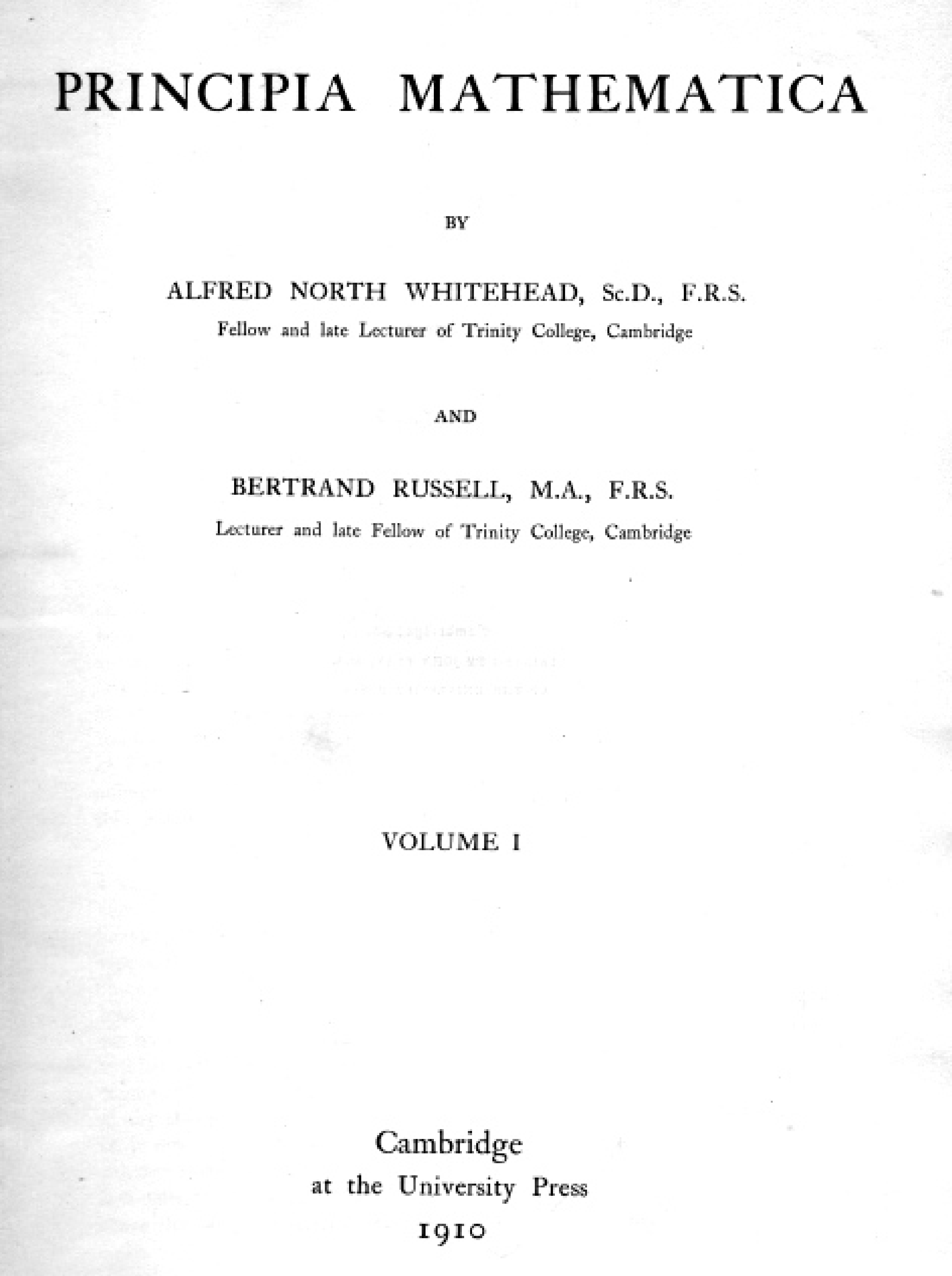
Aquinas, Saint Thomas | Aristotle | Boethius, Anicius Manlius Severinus | change | intrinsic vs. extrinsic properties | perfect being theology

Brian Leftow
leftow@fordham.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 1, 2002
Content last modified: July 1, 2002



PRINCIPIA MATHEMATICA

BY

ALFRED NORTH WHITEHEAD, Sc.D., F.R.S.
Fellow and late Lecturer of Trinity College, Cambridge

AND

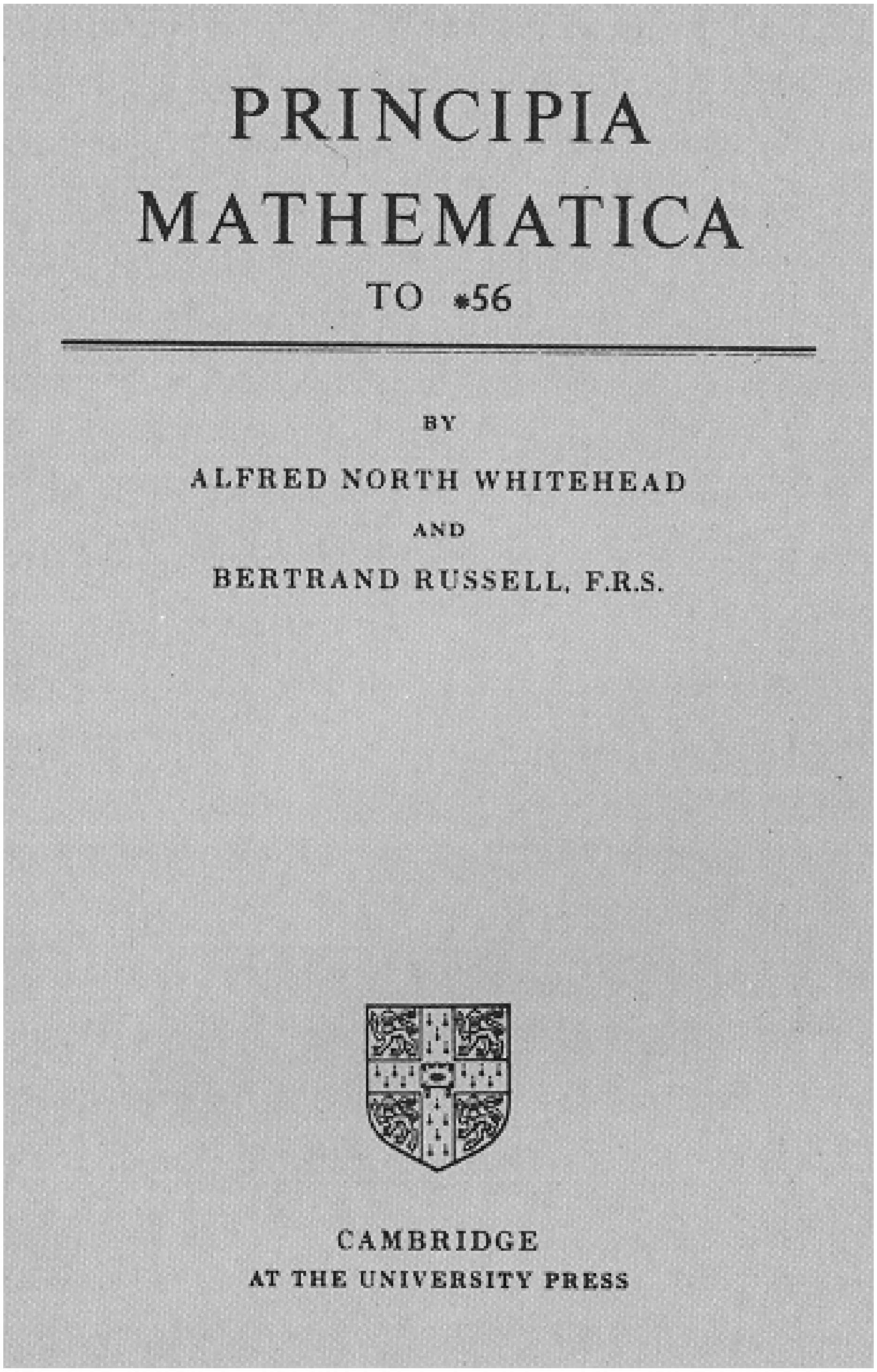
BERTRAND RUSSELL, M.A., F.R.S.
Lecturer and late Fellow of Trinity College, Cambridge

VOLUME I

Cambridge
at the University Press
1910

Title page of the 1st edition of *Principia Mathematica*, Volume 1

(*This image appears courtesy of the Bertrand Russell Archives at McMaster University*.)



PRINCIPIA
MATHEMATICA
TO ∗56

BY
ALFRED NORTH WHITEHEAD
AND
BERTRAND RUSSELL, F.R.S.

CAMBRIDGE
AT THE UNIVERSITY PRESS

Cover of the 1st edition of *Principia Mathematica to 56**

(*This image appears courtesy of the* *Bertrand Russell Archives* *at McMaster University*.)

Difficulties for *N*-relation Accounts of Natural Laws

N-relation theories of natural laws face difficulties in accounting for some of the very things they were introduced to explain. These difficulties may not be insuperable, but they are serious enough to be worth noting here.

Modality It isn't clear that *N*-relation theorists are any better off than regularity theorists when it comes to explaining the modal features of laws. If it is a contingent matter, as most aver, which first-order properties stand in the *N*-relation, then this relation doesn't link those properties because of what those properties are *like*; it just happens to link some properties in the actual world, and it could link completely different properties in other possible circumstances. On this account light could have had the phenomenal properties of molasses, photons the mass of the solar system, and two quite different kinds of elementary particles could retain their identities while swapping all their quantum numbers.

The reason why modality is a problem for such *N*-relation theorists is that you cannot infer a genuine nomological *must* from an *is* (although you may be able to contrive some sort of ersatz *must*). If we want genuine modal force to fall out at the end, we have to build it in at the beginning, and on *N*-relation accounts there is no place to put it except in the *N*-relation itself. We might do this by holding that the *N*-relation links the same properties in every nomologically possible world, but without an independent characterization of nomological possibility this doesn't illuminate the nature of laws. So it can be tempting, as a sort of last resort, to conclude that the necessity involved is a very strong, metaphysical necessity. If we do build such a strong necessity into the *N* relation, so that if it actually relates two properties, it necessarily does so, we can explain the modal force of laws. But in accordance with the fundamental ontological tradeoff, investing the *N*-relation with this strong modal force raises the epistemic cost.

Confirmation If laws involve a metaphysical relation holding among properties, it may be more difficult to determine when an observed regularity is actually backed by a law. *N*-relation theorists may urge that we do this using whatever methods actual scientists employ. But argument is required to show that such methods actually give us reasons to believe that there are laws with the strong features that *N*-relation theorists suppose, rather than merely supplying reasons to draw the weaker sorts of conclusions proposed by regularity theorists.

A Nice Sharp Line It is far from clear that there is a sharp line between laws non-laws. Indeed, we often talk as though some laws (e.g., various conservation laws) are very fundamental and robust, while other laws (e.g., Kepler's laws) are less so.

Explanation We often (indeed very often, outside fundamental physics) explain things by citing the causal mechanisms and processes they involve rather than by subsuming them under general laws. For example, we do not explain why all crows are black by saying (in some more idiomatic way) that the *N*-relation holds between the properties *being a crow* and *being black*. We explain it by finding causal (in this case genetic) mechanisms that link the two properties. In other cases we appeal to a deeper theory, e.g., we explain why Kepler's laws hold (to the extent that they do) by deriving (approximations of) them from Newton's laws.

Chris Swoyer
cswoyer@ou.edu

Return to Properties

First published: December 15, 2000
Content last modified: December 15, 2000

Stanford Encyclopedia of Philosophy
Notes to Properties

Notes

1. Let **I** be an interpretation function with respect to a model, so that **I**(*a*) is the individual **I** assigns to the individual constant '*a*', **I**(F^n) is the *n*-place relation that **I** assigns to *n*-place predicate F^n. Then for all *n*-place predicates 'F^n' and individuals constants 'c_1', ..., 'c_n', the sentence '$F^n c_1 \ldots c_n$' is true in the interpretation just in case < **I**(c_1), ..., **I**(c_n)> ∈ **ext**(**I**(F^n)). We would relativize this clause to variable assignments in the usual way to define satisfaction conditions for atomic sentences.

2. More generally, if φ is a formula with free occurrences of exactly the variables v_1, ..., v_n, then '[λv_1,..., v_n φ]' is an *n*-place complex predicate (normal quotation marks are used here as stand-ins for quasi-quotation). In more complicated applications we could allow φ to contain no free variables (in which case '[λ φ]' denotes the proposition *that*-φ) or more free variables than are bound by the λ-operator (to allow expressions like '[λ*x Fxyz*]', which we could quantify into, as with '∃*y*([λ*x Fxyz*])'.

3. A standard sort of comprehension schema holds that for each open formula φ, $\exists X^n \forall x_1 \ldots \forall x_n (X^n x_1 \ldots x_n$ if and only if φ).

Chris Swoyer
cswoyer@ou.edu

First published: September 23, 1999
Content last modified: September 23, 1999

Vagueness

There is wide agreement that a term is vague to the extent that it has borderline cases. This makes the notion of a borderline case crucial in accounts of vagueness. I shall concentrate on an historical characterization of borderline cases that most commentators would accept. Vagueness will then be contrasted with ambiguity and generality. This will clarify the nature of the philosophical challenge posed by vagueness. I will then discuss some rival theories of vagueness with a special emphasis on supervaluationism. I will conclude with the issue of whether all vagueness is linguistic.

1. Inquiry Resistance

If you cut one head off of a two headed man, have you decapitated him? What is the maximum height of a short man? When does a fertilized egg develop into a person?

These questions are impossible to answer because they involve absolute borderline cases. In the vast majority of cases, the unknowability of a borderline statement is only relative to a given means of settling the issue (Sorensen 2001a, chapter 1). For instance, a boy may count as a borderline case of ‘obese’ because people cannot tell whether he is obese just by looking at him. A curious mother could try to settle the matter by calculating her boy's body mass index. The formula is to divide his weight (in kilograms) by the square of his height (in meters). If the value exceeds 30, this test counts him as obese. The calculation will itself leave some borderline cases. The mother could then use a weight-for-height chart. These charts are not entirely decisive because they do not reflect the ratio of fat to muscle, whether the child has large bones, and so on. The boy will only count as an absolute borderline case of ‘obese’ if no possible method of inquiry could settle whether he is obese. When we reach this stage, we start to suspect that our uncertainty is due to the concept of obesity rather than to our limited means of testing for

obesity.

Absolute borderline cases are targeted by Charles Sander Peirce's entry for ‘vague’ in the 1902 *Dictionary of Philosophy and Psychology*:

> A proposition is vague when there are possible states of things concerning which it is *intrinsically uncertain* whether, had they been contemplated by the speaker, he would have regarded them as excluded or allowed by the proposition. By intrinsically uncertain we mean not uncertain in consequence of any ignorance of the interpreter, but because the speaker's habits of language were indeterminate. (Peirce 1902, 748)

In the case of relative borderline cases, the question is clear but our means for answering it are incomplete. In the case of absolute borderline cases, there is incompleteness in the question itself.

When a term is applied to one of its absolute borderline cases the result is a statement that resists all attempts to settle whether it is true or false. No amount of conceptual analysis or empirical inquiry can settle whether removing one head from a two headed man counts as decapitating him. We could give the appearance of settling the matter by stipulating that ‘decapitate’ means ‘remove a head’ (as opposed to ‘make headless’ or ‘remove the head’ or ‘remove the most important head’). But that would amount to changing the topic to an issue that merely sounds the same as decapitation.

Vagueness is standardly defined as the possession of borderline cases. For example, ‘tall’ is vague because a man who is 1.8 meters in height is neither clearly tall nor clearly non-tall. No amount of conceptual analysis or empirical investigation can settle whether a 1.8 meter man is tall. Borderline cases are inquiry resistant. Indeed, the inquiry resistance typically recurses. For in addition to the unclarity of the borderline case, there is normally unclarity as to where the unclarity begins. In other words ‘borderline case’ has borderline cases. This higher order vagueness shows that ‘vague’ is vague.

2. Comparison with Ambiguity and Generality

‘Tall’ is relative. A 1.8 meter pygmy is tall for a pygmy but a 1.8 meter Masai is not tall for a Masai. Although relativization disambiguates, it does not eliminate borderline cases. There are shorter pygmies who are borderline tall for a pygmy and taller Masai who are borderline tall for a Masai. The direct bearers of vagueness are a word's full disambiguations such as ‘tall for an eighteenth century French man’. Words are only vague indirectly, by virtue of having a sense that is vague. In contrast, an ambiguous word has its ambiguity directly -- simply in virtue of having multiple meanings.

This contrast between vagueness and ambiguity is obscured by the fact that most words are both vague and ambiguous. ‘Child’ is ambiguous between ‘offspring’ and ‘immature offspring’. The latter reading of ‘child’ is vague because there are borderline cases of immature offspring. The contrast is further complicated by the fact that most words are also general. For instance, ‘child’ covers both boys and girls.

Mathematical terms such as ‘prime number’ show that a term can be general without being vague. A term can also be vague without being general. Borderline cases of analytically empty predicates illustrate this possibility.

Generality is obviously useful. Often, lessons about a particular *F* can be projected to other *F*s in virtue of their common *F*-ness. When a girl learns that her *cat* has a nictating membrane that protects its eyes, she rightly expects her neighbor's cat also has a nictating membrane. Generality saves labor. When the girl says that she wants a toy rather than clothes, she narrows the range of acceptable gifts without going through the trouble of specifying a particular gift. The girl also balances values: a gift should be intrinsically desired and yet also be a surprise. If uncertain about which channel is the weather channel, she can hedge by describing the channel as ‘forty-something’. There is an inverse relationship between the contentfulness of a proposition and its probability: the more specific a claim, the less likely it is to be true. By gauging generality, we can make sensible trade-offs between truth and detail.

‘Vague’ has a sense which is synonymous with abnormal generality. This precipitates many equivocal explanations of vagueness. For instance, many commentators say that vagueness exists because broad categories ease the task of classification. If I can describe your sweater as red, then I do not need to figure out whether it is scarlet. This freedom to use wide intervals obviously helps us to learn, teach, communicate, and remember. But so what? The problem is to explain the existence of borderline cases. Are they present because vagueness serves a function? Or are borderline cases side-effects of ordinary conversation -- like echoes?

Unless special preventive measures are taken the attempt to classify objects will unintentionally leave some unanswerable questions in its wake. Predicates are like multi-stage rockets. Rockets which shed components are not designed to spread space age thingamajigs to inaccessible places. But if no precautions are taken, rockets will sow mysteries. At the extreme end of this accidental continuum lie the analogues of borderline cases.

Every natural language is both vague and ambiguous. However, both features seem eliminable. Indeed, both are eliminated in miniature languages such as checkers notation, computer programming languages, and mathematical descriptions. Moreover, it seems that both vagueness and ambiguity ought to be minimized. ‘Vague’ and ‘ambiguous’ are pejorative terms. And they deserve their bad reputations. Think of all the automotive misery that has been prefaced by

> Driver: Do I turn left?
> Passenger: Right.

English can be lethal. Philosophers have long motivated appeals for an ideal language by pointing out how ambiguity creates the menace of equivocation:

> No child should work.
> Every person is a child of someone.

> Therefore, no one should work.

Happily, we know how to criticize and correct all equivocations. Indeed, every natural language is self-disambiguating in the sense that each has all the resources needed to uniquely specify any reading one desires. Ambiguity is often the cause but rarely the object of philosophical rumination.

3. The Philosophical Challenge Posed by Vagueness

Vagueness, in contrast, precipitates a profound problem: the sorites paradox. For instance,

> Base step: A one day year old human being is a child.
>
> Induction step: If an *n* day old human being is a child, then that human being is also a child when it is *n* + 1 days old.
>
> Conclusion: Therefore, a 36,500 day old human being is a child.

The conclusion is false because a 100 year old man is clearly a non-child. Since the base step of the argument is also plainly true and the argument is valid by mathematical induction, we seem to have no choice but to reject the second premise.

George Boolos (1991) observes that we have an autonomous case against the induction step. In addition to implying plausible conditionals such as ‘If a 1 day old human being is a child, then that human being is also a child when it is 2 days old’, the induction step also implies ludicrous conditionals such as ‘If a 1 day old human being is a child, then that human being is also a child when it is 36,500 days old’. For some reason, we tend to overlook these easy counterexamples to the induction step.

With Boolos' helping hand, we have driven *two* stakes into the heart of the sorities paradox. Yet the paradox seems far from dead. The negation of the second premise classically implies a sharp threshold for childhood. For it implies the existential generalization that there is a number *n* such that an n day old human being is a child but is no longer a child *n* + 1 days.

Epistemicists accept this astonishing consequence. They think vagueness is a form of ignorance. Timothy Williamson (1994) traces the ignorance of the threshold for childhood to "margin for error" principles. If one knows that an *n* day old human being is a child, then that human being must also be a child when *n* + 1 days old. Otherwise, one is right by luck.

Most philosophers believe that epistemicism is tantamount to the acceptance of a linguistic miracle. They boggle at the possibility that our rough and ready terms such as ‘child’ could so sensitively classify

objects. Epistemicists counter that this bafflement rests on an over-estimate of the role of stipulation in meaning. They say much meaning is acquired passively by default rather than actively by decision.

Still, most philosophers blame logic rather than our beliefs about language. Surprisingly, H. G. Wells was amongst the first to suggest that we must moderate the *application* of logic:

> Every species is vague, every term goes cloudy at its edges, and so in my way of thinking, relentless logic is only another name for stupidity -- for a sort of intellectual pigheadedness. If you push a philosophical or metaphysical enquiry through a series of valid syllogisms -- never committing any generally recognized fallacy -- you nevertheless leave behind you at each step a certain rubbing and marginal loss of objective truth and you get deflections that are difficult to trace, at each phase in the process. Every species waggles about in its definition, every tool is a little loose in its handle, every scale has its individual. -- *First and Last Things* (1908)

Many more believe that problem is with logic itself rather than the manner in which it is applied. They favor solving the sorites paradox by replacing standard logic with an earthier deviant logic.

There is a desperately wide range of opinions as to how the revision of logic should be executed. Every form of deviant logic has been applied in the hope of resolving the sorites paradox.

An early favorite was many-valued logic. On this approach, borderline statements are assigned truth-values that lie between full truth and full falsehood. New rules are introduced to calculate the truth value of compound statements that contain statements with intermediate truth-values. For instance, the revised rule for conjunctions is to assign the conjunction the same truth-value as the conjunct with the lowest truth-value.

Most theorems of standard logic break down when intermediate truth-values are involved. (An irrelgular minority, such as $P \rightarrow P$, survive.) Even the classical contradiction ‘Bozo is bald and it is not the case that he is bald’ receives a truth-value of .5 when ‘Bozo is bald’ has a truth-value of .5. Many-valued logicians note that the error they are imputing to classical logic is often so small that classical logic can still be fruitfully applied. But they insist that the sorites paradox illustrates how tiny errors can accumulate into a big error.

Critics of the many-valued approach complain that it botches phenomena such as hedging. If I regard you as a borderline case of ‘tall man’, I cannot sincerely assert that you are tall and I cannot sincerely assert that you are of average height. But I can assert the hedged claim ‘You are tall or of average height’. The many-valued rule for disjunction is to assign the whole statement the truth-value of its highest disjunct. Normally, the added disjunct in a hedged claim is not more plausible than the other disjuncts. Thus it cannot increase the degree of truth. Disappointingly, the proponent of many-valued logic cannot trace the increase of assertibility to an increase the degree of truth.

Epistemicists explain the rise in assertibility by the increasing probability of truth. Since the addition of disjuncts can raise probability indefinitely, the epistemicists correctly predict that we can hedge our way to full assertibility. However, epistemicists does not have a monopoly on this prediction.

4. Supervaluationism

According to supervaluationists, borderline statements lack a truth-value. This neatly explains why it is universally impossible to know the truth-value of a borderline statement. Supervaluationism offers details about the nature of absolute borderline cases. Simple sentences about borderline cases lack a truth-value. Compounds of these statements can have a truth-value if they come out true regardless of how the statement is precisified. For instance, ‘Either Mr. Stoop is tall or it is not the case that Mr. Stoop is tall’ is true because it comes out true under all ways of sharpening ‘tall’. Thus the method of supervaluations allows one to retain all the theorems of standard logic while admitting "truth-value gaps".

One may wonder whether this striking result is a genuine convergence with standard logic. Is the supervaluationist characterizing vague statements as propositions? Or is he merely pointing out that certain non-propositions have a structure isomorphic to logical theorems? (Some electrical circuits are isomorphic to tautologies but this does make the circuits tautologies.) Kit Fine (1975, 282), and especially David Lewis, characterize vagueness as hyper-ambiguity. Instead of there being one vague concept, there are many precise concepts that closely resemble each other. ‘Child’ can mean a human being at most one day old or mean a human being at most two days old or mean a human being at most three days old . . . Thus the logic of vagueness is a logic for equivocators. Lewis' idea is that ambiguous statements are true when they come out true under all disambiguations. But logicians normally require that a statement be disambiguated *before* logic is applied. The laws of logic are about *propositions*. The mere fact that an ambiguous statement comes out true under all its disambiguations does not show that the statement itself is true. Sentences which are *actually* disambiguated may have truth-values. But the best that can be said of those that merely *could* be disambiguated is that they *would* have had a truth-value had they been disambiguated (Tye 1989).

Supervaluationism will converge with classical logic only if each word of the supervaluated sentence is uniformly interpreted. For instance, ‘Either a carbon copy of Teddy Roosevelt's signature is an autograph or it is not the case that a carbon copy of Teddy Roosevelt's signature is an autograph’ comes out true only if ‘autograph’ is interpreted the same way in both disjuncts. Vague sentences resist mixed interpretations. However, mixed interpretations are permissible for ambiguous sentences. As Lewis himself notes in a criticism of relevance logic, ‘Scrooge walked along the bank on his way to the bank’ can receive a mixed disambiguation. When exterminators offer ‘non-toxic ant poison’, we charitably relativize: the substance is safe for human beings but deadly for ants.

Even if one agrees that supervaluationism converges with classical logic about theoremhood, they clearly differ in other respects. Supervaluationism requires rejection of inference rules such as contraposition, conditional proof and reductio ad absurdum (Williamson 1994, 151-152). In the eyes of the supervaluationist, a demonstration that a statement is not true does not guarantee that the statement is

false.

The supervaluationist is also under pressure to reject semantic principles which are intimately associated with the application of logical laws. According to Alfred Tarski's Convention T, a statement ‘*S*’ is true if and only if *S*. In other words, truth is disquotational. Supervaluationists say that being supertrue (being true under all precisifications) suffices for being true. But given Convention T, supertruth would then be disquotational. Since the supervaluationists accept the principle of excluded middle, they would be forced to say ‘*P*’ is supertrue or ‘Not *P*’ is supertrue (even if ‘*P*’ applies a predicate to a borderline case). This would imply that either ‘*P*’ is true or ‘Not *P*’ is true. (Williamson 1994, 162-163) And that would be a fatal loss of truth-value gaps for supervaluationism.

There is a final concern about the "ontological honesty" of the supervaluationist's existential quantifier. As part of his solution to the sorites paradox, the supervaluationist will assert ‘There is a human being who was a child when *n* days old but not when *n* + 1 days old’. For this statement comes out true under all admissible precisifications of ‘child’. However, when pressed the supervaluationist will add an unofficial clarification: "Oh, of course I do not mean that there really is a sharp threshold for childhood."

After the clarification, some wonder how supervaluationism differs from drastic metaphysical skepticism. In his nihilist days, Peter Unger (1979) admitted that it is useful to talk *as if* there are children. But he insisted that strictly speaking, vague terms such as ‘child’ cannot apply to anything. Unger was free to use supervaluationism as a theory to explain our ordinary discourse about children. (Unger instead used other resources to explain how we fruitfully apply empty predicates.) But once the dust had cleared and the precise rubble came into focus, Unger had to conclude that there are no children.

Officially, the supervaluationist rejects the induction step of the sorites argument. Unofficially, he seems to instead reject the *base* step of the sorites argument.

5. Is All Vagueness Linguistic?

Supervaluationism encourages the view that all vagueness is a matter of linguistic indecision: the reason why there are borderline cases is that we have not bothered to make up our minds. Many supervaluationists maintain that this indecision is functional. Instead of committing ourselves prematurely, we can fill in meanings as we go along in light of new information and interests. This conjecture is promising for the highly stipulative enterprise of promulgating and enforcing laws (Endicott 2000). Judges frequently seem to exercise and control discretion by means of vague language. However, there is the suspicion that the real work is being done by the generality of the legal propositions rather than their vagueness (Sorensen 2001b).

Objects themselves do not seem to be the sort of thing that can be general, ambiguous, or vague. Thus the Romantics appear to be committing a category mistake when they characterize sea foam as vague. Indeed, there used to be a consensus that believers in vague objects were committing the fallacy of verbalism -- inferring that an object has the property that its representation has.

A minority of philosophers now believe that there are vague objects (clouds, the sky, perhaps even entities of quantum physics). There is a precedent for this revival. Peter van Inwagen recalls that thirty years ago, there was a consensus that all necessity is linguistic. Most philosophers now take the possibility of essential properties seriously.

Indeed, some allege that supervaluationists inadvertently rely on metaphysical vagueness to characterize linguistic vagueness (Merricks 2001). If ‘Bozo is bald’ lacks a truth-value because there is no fact to make the statement true, then the shortage appears to be ontological rather than linguistic.

The view that vagueness is always linguistic has been attacked from other directions. Mental imagery seems vague. When rising suddenly after a prolonged crouch, I "see stars before my eyes". I can tell there are more than ten of these hallucinated lights but I cannot tell how many. Is this indeterminacy in thought to be reduced to indeterminacy in language? Why not vice versa? Language is an outgrowth of human psychology. Thus it seems natural to view language as merely an accessible intermediate bearer of vagueness.

Bibliography

- Boolos, George (1991) ‘Zooming Down the Slippery Slope’. *Nous* 25: 695-706.
- Endicott, Timothy (2000) *Vagueness in the Law* (Oxford University Press).
- Evans, Gareth: ‘Can there be Vague Objects?’ *Analysis*, 38 (1978), 208.
- Fine, Kit (1975). ‘Vagueness, truth and logic’. *Synthese* 54: 235-59. Reprinted in *Vagueness: A Reader* ed. Rosanna Keefe and Peter Smith. Cambridge: MIT Press, 1996): 119-150.
- Lewis, David (1982) "Logic for Equivocators" *Nous* 16: 431-441.
- Lewis, David (1993) "Many, but almost one" in *Ontology, Causality, and Mind: Essays on the Philosophy of D.M. Armstrong* (Cambridge: Cambridge University Press) ed. Keith Campbell, John Bacon, and Lloyd Reinhardt.
- Merricks, Trenton (2001) "Varieties of Vagueness" *Philosophy and Phenomenological Research* vol. LXII, no. 1: 145-157.
- Peirce, C.S. (1902). ‘Vague’ in *Dictionary of Philosophy and Psychology* ed. J.M. Baldwin (New York: MacMillan): 748.
- Sorensen, Roy (2001a) *Vagueness and Contradiction* Oxford: Oxford University Press.
- Sorensen, Roy (2001b) "Vagueness has no function in law" *Legal Theory* 7/4: 385-415.
- Tye, Michael (1989) "Supervaluationism and the Law of Excluded Middle" *Analysis* 49/3: 141-143.
- Unger, Peter (1979) ‘There are no ordinary things’ *Synthese* 4: 117-54.
- Van Inwagen, Peter (1990) *Material Beings* (Ithaca, New York: Cornell University Press).
- Williamson, Timothy: *Vagueness* (London: Routledge).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

ambiguity | Sorites paradox

Roy Sorensen
roy.sorensen@dartmouth.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published:
Content last modified: February 22, 2002

Sorites Paradox

The sorites paradox is the name given to a class of paradoxical arguments, also known as little-by-little arguments, which arise as a result of the indeterminacy surrounding limits of application of the predicates involved. For example the concept of a heap appears to lack sharp boundaries and, as a consequence of the subsequent indeterminacy surrounding the extension of the predicate 'is a heap', no one grain of wheat can be identified as making the difference between being a heap and not being a heap. Given then that one grain of wheat does not make a heap, it would seem to follow that two do not, thus three do not, and so on. In the end it would appear that no amount of wheat can make a heap. We are faced with paradox since from apparently true premises by seemingly uncontroversial reasoning we arrive at an apparently false conclusion.

This phenomenon at the heart of the paradoxes is now recognised as the phenomenon of vagueness. Once identified, vagueness can be seen to be a feature of syntactic categories other than predicates, nonetheless one speaks primarily of the soriticality of predicates. Names, adjectives, adverbs and so on are only susceptible to paradoxical sorites reasoning in a derivative sense.

Sorites arguments of the paradoxical form are to be distinguished from multi-premise syllogisms (polysyllogisms) which are sometimes also referred to as sorites arguments. Whilst both polysyllogisms and sorites paradoxes are chain-arguments, the former need not be paradoxical in nature and the latter need not be syllogistic in form.

The Sorites In History

The name 'sorites' derives from the Greek word *soros* (meaning 'heap') and originally referred, not to a

paradox, but rather to a puzzle known as *The Heap*: Would you describe a single grain of wheat as a heap? No. Would you describe two grains of wheat as a heap? No. ... You must admit the presence of a heap sooner or later, so where do you draw the line?

It was one of a series of puzzles attributed to the Megarian logician Eubulides of Miletus. Also included were:
The Liar: A man says that he is lying. Is what he says true or false?
The Hooded Man: You say that you know your brother. Yet that man who just came in with his head covered is your brother and you did not know him.
The Bald Man: Would you describe a man with one hair on his head as bald? Yes. Would you describe a man with two hairs on his head as bald? Yes. ... You must refrain from describing a man with ten thousand hairs on his head as bald, so where do you draw the line?

This last puzzle, presented as a series of questions about the application of the predicate 'is bald', was originally known as the *falakros* puzzle. It was seen to have the same form as the Heap and all such puzzles became collectively known as sorites puzzles.

It is not known whether Eubulides actually invented the sorites puzzles. Some scholars have attempted to trace its origins back to Zeno of Elea however the evidence seems to point to Eubulides as the first to employ the sorites. Nor is it known just what motives Eubulides may have had for presenting this puzzle. It was, however, employed by later Greek philosophers to attack various positions. Most notably by the Sceptics against the Stoics' claims to knowledge.

These puzzles of antiquity are now more usually described as paradoxes. Though the conundrum can be presented informally as a series of questions whose puzzling nature gives it dialectical force it can be, and was, presented as a formal argument having logical structure. The following argument form of the sorites was common:

> 1 grain of wheat does not make a heap.
> If 1 grain of wheat does not make a heap then 2 grains of wheat do not.
> If 2 grains of wheat do not make a heap then 3 grains do not.
> ...
> If 9,999 grains of wheat do not make a heap then 10,000 do not.
> --
> 10,000 grains of wheat do not make a heap.

The argument certainly seems to be valid, employing only *modus ponens* and *cut* (enabling the chaining together of each sub-argument which results from a single application of *modus ponens*). These rules of inference are endorsed by both Stoic logic and modern classical logic.

Moreover its premises appear true. Some Stoic presentations of the argument recast it in a form which replaced all the conditionals, 'If *A* then *B*', with 'Not(*A* and not-*B*)' to stress that the conditional should

not be thought of as being a strong one, but rather the weak Philonian conditional (the modern material conditional) according to which ‘If *A* then *B*’ was equivalent to ‘Not(*A* and not-*B*)’. Such emphasis was deemed necessary since there was a great deal of debate in Stoic logic regarding the correct analysis for the conditional. In thus judging that a connective as weak as the Philonian conditional underpinned this form of the paradox they were forestalling resolutions of the paradox that denied the truth of the conditionals based on a strong reading of them. This interpretation then presents the argument in its strongest form since the validity of *modus ponens* seems assured whilst the premises are construed so weakly as to be difficult to deny. The difference of one grain would seem to be too small to make any difference in the application of the predicate; it is a difference so negligible as to make no apparent difference in the truth values of the respective antecedents and consequents.

Yet the conclusion seems false. Thus paradox confronted the Stoics just as it does the modern classical logician. Nor are such paradoxes isolated conundrums. Innumerable sorites paradoxes can be expressed in this way. For example, one can present the puzzle of the Bald Man in this manner. Since a man with one hair on his head is bald and if a man with one is then a man with two is, so a man with two hairs on his head is bald. Again, if a man with two is then a man with three is, so a man with three hairs on his head is bald, and so on. So a man with ten thousand hairs on his head is bald yet we rightly feel that such men are hirsute, i.e. not bald. Indeed, it seems that almost any vague predicate admits of such a sorites paradox and vague predicates are ubiquitous.

As presented, the paradox of the Heap and the Bald Man proceed by addition (of grains of wheat and hairs on the head respectively). Alternatively though, one might proceed in reverse, by subtraction. If one is prepared to admit that ten thousand grains of sand do make a heap then would can argue that one grain of sand does since the removal of any one grain of sand cannot make the difference. Similarly, if one is prepared to admit a man with ten thousand hairs on his head is not bald, then one could prove that even with one hair on his head he is not bald since the removal of any one hair from the originally hirsute scalp cannot make the relevant difference. It was thus recognised even in antiquity that sorites arguments come in pairs: non-heap and heap; bald and hirsute; poor and rich; few and many; small and large; and so on. For every argument which proceeds by addition there is another reverse argument which proceeds by subtraction.

The paradox attracted little subsequent interest until the late nineteenth century when formal logic once again assumed a central role in philosophy. With the demise of ideal language doctrines in the latter half of the twentieth century interest in the vagaries of natural language and the sorites paradox in particular has greatly increased.

Its Paradoxical Forms

The common form of the sorites paradox presented for discussion in the literature is the form discussed above. Let ‘*F*’ represent the soritical predicate (e.g. ‘is bald’, or ‘does not make a heap’) and let the expression ‘a_i’ (where *i* is a natural number) represent a subject expression in the series with regard to which ‘*F*’ is soritical (e.g., ‘a man with *i* hair(s) on his head’ or ‘*i* grain(s) of wheat’, depending on *F*).

Then the sorites proceeds by way of a series of conditionals and can be schematically represented as follows:

Conditional Sorites

Fa_1
If Fa_1 then Fa_2
If Fa_2 then Fa_3
...
If Fa_{i-1} then Fa_i

Fa_i (where i can be arbitrarily large)

Whether the argument is taken to proceed by addition or subtraction will depend on how one views the series.

Barnes (1982) states conditions under which any argument of this form is soritical. Initially, the series $\langle a_1, ..., a_i \rangle$ must be ordered; for example, scalps ordered according to number of hairs, heaps ordered according to number of grains of wheat, and so on. Secondly, the predicate ‘*F*’ must satisfy the following three constraints: (i) it must appear true of a_1, the first item in the series; (ii) it must appear false of a_i, the last item in the series; and (iii) each adjacent pair in the series, a_n and a_{n+1}, must be sufficiently similar as to appear indiscriminable in respect of ‘*F*’ -- that is, both a_n and a_{n+1} appear to satisfy ‘*F*’ or neither do. Under these conditions ‘*F*’ will be soritical relative to the series $\langle a_1, ..., a_i \rangle$ and any argument of the above form using ‘*F*’ and $\langle a_1, ..., a_i \rangle$ will be soritical.

In recent times the explanation of the fact that sorites arguments come in pairs has shifted from consideration of the sorites series itself to the predicate involved. It is now common to focus on the presence or absence of a negation in the predicate, noting the existence of both a positive form which bloats the predicate's extension and negative form which shrinks the predicate's extension. With the foregoing analysis of the conditions for sorites susceptibility it is easy to verify that ‘*F*’ will be soritical relative to $\langle a_1, ..., a_i \rangle$ if and only if ‘not-*F*’ is soritical relative to $\langle a_i, ..., a_1 \rangle$. Thus verifying that for every positive sorites there is an analogous negative variant.

The key feature of soritical predicates which drives the paradoxes, constraint (iii), is described in Wright (1975) as "tolerance" and is thought to arise as a result of the vagueness of the predicate involved. Predicates such as ‘is a heap’ or ‘is bald’ appear tolerant of sufficiently small changes in the relevant respects -- namely number of grains or number of hairs. The degree of change between adjacent members of that series relative to which ‘*F*’ is soritical would seem too small to make any difference to the application of the predicate ‘*F*’. Yet large changes in the relevant respects will make a difference, even though large changes are the accumulation of small ones which don't seem to make a difference.

This is the very heart of the conundrum which has delighted and perplexed so many for so long.

Any resolution of the paradoxes is further complicated by the fact that they can be presented in a variety of forms and the problem they present can only be considered solved when all forms have been defused.

One variant replaces the set of conditional premises with a universally quantified premises. Let '*n*' be a variable ranging over the natural numbers and let '∀*n*(...*n*...)' assert that every number *n* satisfies the condition ...*n*... . Further, let us represent the claim of the form "for all *n*, if Fa_n then Fa_{n+1}" as follows:

$$\forall n(Fa_n \rightarrow Fa_{n+1})$$

Then the sorites is now seen as proceeding by the inference pattern known as mathematical induction:

Mathematical Induction Sorites

Fa_1
$\forall n(Fa_n \rightarrow Fa_{n+1})$

$\forall nFa_n$

So, for example, it is argued that since a man with 1 hair on his head is bald and since the addition of one hair cannot make the difference between being bald and not bald (for any number *n*, if a man with *n* hairs is bald then so is a man with *n*+1 hairs), then no matter what number *n* you choose, a man with *n* hairs on his head is bald.

Yet another form is a variant of this inductive form. Assume that it is not the case that for every *n*, a man with *n* hairs on his head is not bald, i.e., that for some number *n*, it is not the case that a man with *n* hairs on his head is bald. Then by the least number principle (equivalent to the principle of mathematical induction) there must be a least such number, say *i*+1, such that it is not the case that a man with *i*+1 hairs on his head is bald. Since a man with 1 hair on his head is bald it follows that *i*+1 must be greater than 1. So, there must be some number *n* (= *i*) such that a man with *n* hairs counts as bald whilst a man with *n*+1 does not. Thus it is argued that though a_1 is bald, not every number *n* is such that a_n is bald, so there must be some point at which baldness ceases. Let '∃*n*(...*n*...)' assert that some number *n* satisfies the condition ...*n*... . Then we can represent the chain of reasoning just described as follows:

Line-drawing Sorites

Fa_1
$\sim\forall nFa_n$

$$\exists n \geq 1(Fa_n \;\&\; {\sim}Fa_{n+1})$$

Now obviously, given that sorites arguments have been presented in these three forms, "the sorites paradox" will not be solved by merely claiming, say, mathematical induction to be invalid for soritical predicates. All forms need to be addressed one way or another. One would hope to solve it, if at all, by revealing some general underlying fault common to all forms of the paradox. No such general solution could depend on the diagnosis of a fault peculiar to any one form. On the other hand, were no general solution available then "the sorites paradox" will only be solved when each of its forms separately have been rendered toothless. This piecemeal approach holds little attraction though. It is less economical than a unified approach, arguably less elegant, and would fail to come to grips with the underlying unifying phenomenon which is considered to give rise to the paradoxes, namely vagueness. A logic of vagueness, be it classical or otherwise, ought to be able to defuse all those paradoxes that have their source in this phenomenon.

Responses

The various responses to soritical reasoning can be most easily catalogued by focussing on that form most commonly discussed in the literature -- the conditional form. As with any paradox, four responses appear to be available. One might:

> (1) deny that logic applies to soritical expressions.

According to this response the problem cannot legitimately be set up in the first place. On the other hand one might accept that the sorites paradox constitutes a legitimate argument to which logic applies and deny its soundness by:

> (2) denying some premise(s),

or

> (3) denying its validity.

Finally, seemingly as a last resort, one might embrace the paradox and

(4) accept it as sound.

Ideal Language Approaches

Committed as Frege and Russell were to ideal language doctrines, it is not surprising to find them pursuing response (1). A key attribute of the ideal language is its precision; the vagueness of natural language is a defect to be eliminated. Since soritical terms are vague, the elimination of vagueness will entail the elimination of soritical terms. They cannot then, as some theorists propose, be marshalled as a challenge to classical logic.

A modern variation on this response, promoted most notably by Quine, sees vagueness as an eliminable feature of natural language. The class of vague terms, including soritical predicates, can as a matter of fact be dispensed with. There is, perhaps, some cost to ordinary ways of talking, but a cost that is nonetheless worth paying for the simplicity it affords -- namely, our thereby being able to defend classical logic with what Quine describes as its "sweet simplicity".

However, with the demise of ideal language doctrines and subsequent restoration of respect for ordinary language, vagueness is increasingly considered less superficial than response (1) suggests. If logic is to have teeth it must be applicable to natural language as it stands. Soritical expressions are unavoidable and the paradox must be squarely faced.

Responses of type (2) do just this. Logic is seen as applicable to natural language, in particular the sorites paradox, but the conditional form of the argument is seen as proceeding from a faulty premise.

The Epistemic Theory

According to Williamson (1994) Stoic logicians pursued just such a route. Given their acceptance of the principle of bivalence and their presentation of the argument as invoking a material conditional, they blocked the sorites by claiming some one conditional to be false (since not true) and that there comes a point in any sorites series where the relevant predicate ceases to apply and its negation does. For example vague terms like ‘heap’ or ‘knowledge’, though soritical relative to an appropriately chosen series, are semantically determinate so, in spite of appearances to the contrary, there is a sharp cut-off point to their application. The inclination to validate all the premises of a sorites argument (along with the inference pattern employed, which the Stoics accepted) was to be explained via ignorance -- more exactly, the unknowable nature of the relevant sharp semantic boundary.

In this way the threat of wholesale scepticism urged by the Sceptics was met by the limited scepticism arising from our inability to know the precise boundaries to knowledge. "Nothing can be known" was rejected in favour of "The precise boundaries to knowledge itself cannot be known". This epistemological response has been elaborated on in Sorensen (1988) and Williamson (1994). Though soritical predicates are admittedly indeterminate in their extension the indeterminacy is not semantic. The conundrum presented by the sorites paradox is an epistemological one which in no way challenges

classical semantics or logic.

Until recently such a solution was ruled out by definition. Vagueness was characterised as a semantic phenomenon whereby the apparent semantic indeterminacy surrounding a soritical term's extension was considered real. In the absence of any apparent barrier to knowledge of a soritical predicate's precise extension it was assumed that there was simply no precise extension to be known. The philosophical landscape has now changed. Williamson (1994) contains an impressive array of arguments defending an epistemological account of vagueness which, if successful, would make possible an epistemological solution to the sorites.

The key concern with the epistemological approach however is its counter intuitive nature. Even if such an analysis is possible, the indeterminacy surrounding the application of soritical terms is generally considered to be a semantic phenomenon. Once seen in this way, classical semantics appears in need of revision, and with it classical logic. In the second half of this century there have been a number of attempts to develop non-classical logics of vagueness, a major constraint being the provision of a solution to the sorites paradox. The extent of the proposed logical innovation varies.

Supervaluationism

In accord with a principle of least mutilation, Dummett (1975) and Fine (1975) adapt Van Fraassen's supervaluation semantics to the sorites paradox, and vagueness more generally, resulting in a non-bivalent logic which, initially at least, retains the classical consequence relation and classical laws whilst admitting truth value gaps. The challenge posed by the sorites paradox can, on this view, to be met by logical revision in the metatheory alone.

According to the semantic conception of vagueness, a vague predicate is characterised by the existence of border cases; that is, cases to which the predicate neither definitely applies nor definitely doesn't apply. If we define the positive extension as those objects to which the predicate definitely applies, the negative extension as those to which the predicate definitely does not apply, and any remaining border cases as the penumbra, then vague predicates are characterised by their having a penumbra.

Given a vague predicate, for example ‘heap’, we can then stipulate a sharpening thereof, ‘heap*’ which resolves any border cases by placing them either in the positive or negative extension of ‘heap*’. Intuitively, for a sharpening to be admissible as a sharpening of the original vague predicate it ought to also be constrained by its only resolving vague semantic aspects of the predicate's meaning. For example if a pile of *i* grains of wheat definitely counts as a heap then it ought to be definitely counted a heap*; positive cases ought remain unchanged, as should negative cases. Additionally it should not draw a line between positive and negative cases in such a way as to alter definite ordering relations. Since it is definitely the case that *i*+1 grains counts as a heap if *i* does, *i*+1 will count as a heap* if *i* does. An admissible sharpening of any soritical predicate will simply extend the positive and negative extensions to form a sharp boundary somewhere in the predicate's penumbra.

It is easy to see that vague predicates in general have no unique admissible sharpening; there will be a number of possibilities depending on where the cut-off point between the positive and negative extension of the now precise predicate is drawn. Nonetheless, to predicate ‘heap’ of something in the positive extension will be true for all admissible sharpenings of the predicate and to predicate ‘heap’ of something in the negative extension will be false for all admissible sharpenings. Defining "truth" *simpliciter* (or, as it is sometimes called, "supertruth") as "truth on all admissible sharpenings", the former predications will be, quite simply, true and the latter false. To predicate ‘heap’ of a border case will be true on some admissible sharpenings and false on others and so is considered neither true nor false but indeterminate in truth value.

Validity can now be defined so that an argument is supervaluationally valid just in case every model in which the premises are true is one in which the conclusion is also true. Validity thus defined coincides with classical validity, reflecting the fact that the basic underlying notion of ‘truth on an admissible sharpening’ is a classical two-valued valuation function. In particular, treating laws as zero-premise arguments, such supervaluationism preserves all classical laws. In spite of its being non-bivalent then it validates the law of excluded middle. For example, it is true in every model that i grains of wheat either does or does not make a heap since in any model the corresponding disjunction with ‘heap*’ in place of ‘heap’ will be true regardless of which admissible sharpening ‘heap*’ is considered.

Supervaluation semantics then is no longer truth-functional. Consider disjunction for example. Some disjunctions with indeterminate disjuncts will count as indeterminate, for example ‘(Border case) a_{i} makes a heap or it makes a heap’; whereas some will count as true, for example ‘(Border case) a_{i} makes a heap or it does not make a heap’. Moreover the semantics countenances instances of true disjunctions neither of whose disjuncts is true. Conjunction and the conditional exhibit analogous non-classical features.

Since all the forms taken by the sorites are classically valid, they are also supervaluationally valid. The conclusion of the conditional form is resisted by noticing that some conditional premise fails to be true. For example, the premise ‘If (border case) a_i makes a heap then (border case) a_{i+1} makes a heap’ will admit of a sharpening which draws the line between a_i and a_{i+1} thus making the conditional false on that sharpening, and one which does not so draw the line, making the conditional true on that sharpening. It is therefore neither true nor false but indeterminate. The conditional sorites is valid but unsound.

More revealing is the diagnosis with regard to the mathematical induction form. It is also deemed unsound due to the failure of one of the premises -- the universal premise. The universally quantified conditional is not true; in fact it's false. While there is no one conditional premise of the conditional form which is false, it is nonetheless true according to supervaluation theory that some conditional is. That is to say, it is false that for all n, if Fa_n then Fa_{n+1} (where ‘F’ is soritical relative to the subjects of the form a_i). For any sharpening of the vague predicate ‘F’ there will always be some a_i which counts as the cut-off point relative to that sharpening and thus falsifies the relevant conditional ‘if Fa_i then Fa_{i+1}’. Every sharpening produces a cut-off point even though no single cut-off point is produced by every sharpening.

Hence "For all n, if Fa_n then Fa_{n+1}" is false on every sharpening and so false *simpliciter*.

Given that supervaluation semantics admits that the falsity of "For all n, if Fa_n then Fa_{n+1}" is logically equivalent to the truth of "For some n, Fa_n & $\sim Fa_{n+1}$", the line-drawing form of the sorites is also solved. The argument is supervaluationally valid since classically valid and its premises are uncontestably true. What supervaluation semantics provides is an account of how it is that such a conclusion could be true; it is true since true no matter how one admissibly sharpens the soritical predicate involved.

In this way then the sorites paradoxes are said to be defused. Classical logic is no longer appropriate to reasoning in vague contexts and supervaluation semantics is proposed in its place. One immediate concern facing this solution however is the fact that it ultimately treats the mathematical induction and line-drawing forms of the sorites in just the same way as the logically conservative epistemic theory. We are forced to accept the avowedly counter intuitive truth of "For some n, Fa_n & $\sim Fa_{n+1}$" which seems to postulate the existence of a sharp boundary yet the existence of just such a boundary is what the semantic theory of vagueness is supposed to deny.

Supervaluationists respond by denying that the conclusion of the line-drawing sorites expresses the existence of a sharp boundary. Though committed to the claim

(a) T '$\exists n\, (Fa_n \,\&\, \sim Fa_{n+1})$',

semantic precision is only properly captured by the claim that

(b) $\exists n$ T '$(Fa_n \,\&\, \sim Fa_{n+1})$'

and this is clearly denied by supervaluation theory. Whilst it is true that there is some cut-off point, there is no particular point of which it is true that it is the cut-off point. Since it is only this latter claim which is taken to commit one to the existence of a sharp boundary there is no commitment to there being such a boundary of which we are ignorant (contra the epistemic theorist).

With this explanation however, doubts arise as to the adequacy of the logic. Not only must (b) be properly taken to represent the semantic precision of '*F*' but we must also be prepared to admit that some existential statements can be true without having any true instance, thus blocking any inference from (a) to (b). Just as the failure of the metatheoretic principle of bivalence in conjunction with the retention of the law of excluded middle leads to the presence of true disjunctions lacking true disjuncts, so too must we countenance analogous non-standard behaviour in the logic's quantification theory. In effect, the counter intuitive aspects of the epistemic theory are avoided only at a cost to other intuitions.

At this point the supervaluationist might seek to explain the non-standard behaviour of the quantifiers, and for that matter the non-truth-functional two-place connectives, by showing how such behaviour follows from a proper understanding of the underlying phenomenon of vagueness. More exactly, the

suggestion is that a view of vagueness as merely semantic and in no way a reflection of any underlying phenomenon of ontological vagueness might underpin a supervaluationist approach to vagueness. Fine (1975) promotes this representational view of vagueness when defending the law of excluded middle for example. The suggestion would not only prescribe the counter intuitive aspects of supervaluation semantics but would also provide a principled justification of the common *de facto* linkage of supervaluation theory and a representational view of vagueness.

If this explanation is to be pursued then the formal machinery of supervaluation semantics dissolves the paradox only in conjunction with the metaphysical assumption of the impossibility of ontological vagueness. The metaphysical debate is ongoing.

The supervaluation approach has also come under fire for its semantic ascent into the metatheory when defusing the sorites. The problem with accepting the major premise of the mathematical induction sorites as false is simply that it runs counter to our conviction that a grain of wheat can make the difference between a heap and a non-heap; but this conviction can be expressed in the object-language, so why should the elaborate metalinguistic theory be relevant here?

As it happens, such assent is not essential to the account. The object language can be extended to include an operator ‘It is determinate (or definite) that ...’ (‘Det’) appropriate to the expression of vagueness in the object language. The vagueness of expressions like ‘heap’ is characterised by their possessing border cases; this can now be expressed as the existence of cases which are neither determinately heaps nor determinately non-heaps.

By means of the extended language, (a) and (b) are now recast as claims within the object language:

(a') $\text{Det}\ \exists n\ (Fa_n \ \&\ {\sim}Fa_{n+1})$

(b') $\exists n\ \text{Det}\ (Fa_n \ \&\ {\sim}Fa_{n+1})$.

The first is again affirmed and the latter denied. Any inference from (a') to (b') is analogous to the modal inference from ‘Nec $\exists x\ Fx$’ to ‘$\exists x$ Nec Fx’ and is seen as fallacious, just as the modal inference is commonly said to be. Vagueness, like modality, is viewed as *de dicto*, and, as in many modal logics, the resulting quantification theory reflects important scope distinctions.

Williamson (1994) points to two further problems which now beset the account. If the definition of validity as necessary truth-preservation is retained then classical inferences like *conditional proof*, *dilemma* and *reductio ad absurdum* are no longer valid. Moreover problems arise with regard to the phenomenon of higher order vagueness.

Many-Valued Logic

As alternatives to the non-truth-functional supervaluation semantics, non-classical logics have been proposed, and in particular, ‘many-valued logics’. Again vagueness is seen as grounds for rejecting the principle of bivalence, however truth-functionality is preserved. The approaches vary as regards the number of non-classical truth values deemed appropriate to model vagueness and defuse the sorites paradox.

An initial proposal, first developed in Halldén (1949) and Körner (1960) and recently revamped in Tye (1994), uses a three-valued logic. The motivation for such a logic is similar to the supervaluationist's. Just as a vague predicate divides objects into the positive extension, negative extension and the penumbra, vague sentences can be divided into the true, the false and the indeterminate. The truth-set is then {1 (true), 1/2 (indeterminate), 0 (false)}. Unlike supervaluation semantics, however, the connectives can then be defined by means of truth tables. There are a range of proposals with Kleene's strong three-valued tables as the preferred choice. (See Haack (1974), Appendix.)

The particular response to the sorites paradox then depends on the definition of validity adopted. A common generalisation of the concept of validity to many-valued logic involves the designation of certain values. A sentence holds (or is assertible) in a many-valued interpretation just if it takes a designated value. Validity is then defined as the necessary preservation of designated value. (In classical logic, of course, only truth is designated and thus the generalised concept reduces to the classical concept of necessary truth preservation.) There are then two non-trivial choices: let the set of designated values be {1} or {1, 1/2}. The former proposal results in a type (2) response, the latter a type (3) response.

If validity is defined as necessary preservation of truth-only then *modus ponens* and *cut* are valid and so, subsequently, is the conditional sorites. Yet not all its premises hold since they are not all true. Some conditional premise is considered to have a true antecedent and indeterminate consequent and so counts as indeterminate. Like supervaluationism such a logic may be said to be paracomplete, admitting non-trivial incomplete theories. (That is, a vague sentence ‘*A*’ and its negation ‘~*A*’ can both fail to be designated in an interpretation which nonetheless designates some sentences.)

If, on the other hand, validity is defined so that the conclusion of an argument is either true or indeterminate whenever the premises are either true or indeterminate then a paraconsistent logic results, admitting non-trivial inconsistent theories. (A vague sentence ‘*A*’ and its negation ‘~*A*’ might both be designated in an interpretation which nonetheless fails to designate every sentence.) Such an approach solves the sorites paradox by declaring it invalid. The premises hold since they are at worst indeterminate in truth-value, yet *modus ponens* is no longer valid. In an interpretation where ‘Fa_i’ is indeterminate and ‘Fa_{i+1}’ is false both ‘Fa_i’ and ‘If Fa_i then Fa_{i+1}’ are designated yet ‘Fa_{i+1}’ is not.

A general concern with three-valued approaches is that their tripartite division of sentences faces similar objections to those which led to the abandonment of the bipartite division effected by two-valued classical logic. Due to the phenomenon of higher order vagueness (in particular second order vagueness) there would seem to be no more grounds for supposing there to exist a sharp boundary between the true sentences and indeterminate ones or the indeterminate sentences and false sentences than there was for

supposing a sharp boundary to exist between the true sentences and the false ones. The phenomenon of vagueness which drives the sorites paradox no more suggests two sharp boundaries than it did one. Vague concepts appear to be concepts without boundaries at all. No finite number of divisions seems adequate. Tye (1994) seeks to avoid such difficulties by employing a vague metalanguage.

Goguen (1969) and Zadeh (1975) suggest replacing classical two-valued logic with an infinite-valued one. Infinite-valued or fuzzy logics have also been promoted for their recognition of degrees of truth. Just as baldness comes in degrees so too it is argued does the truth of sentences predicating baldness of things. The fact that John is more bald than Jo is reflected in the sentence 'John is bald' having a higher degree of truth than 'Jo is bald'. With this logical innovation infinite-valued logics are then offered as a means to solve the sorites paradox.

The classical two-valued truth set {0, 1} is replaced by set of real numbers in the interval [0, 1]. Sentences which are neither definitely true nor definitely false take values other than 0 or 1, but some number in between. As with all many-valued logics, the connectives can be defined in a number of ways, giving rise to a number of distinct logics. A standard proposal proceeds by way of the continuum-valued truth-functional semantics devised by Lukasiewicz. (See Haack (1974), Appendix.) As with the three-valued case, the type of response offered to the paradox depends on the definition of validity. If only the maximum value 1 is designated then the conditional form of the argument is valid, however the conditional premises are not completely true (that is, they do not take the value 1) and do not hold. The diagnosis of the paradox is that we mistake nearly true sentences for completely true ones and the error compounds each time a new conditional premise is invoked in the chain of reasoning. Beginning with a completely true categorical premise we thus proceed to heap neglected difference on neglected difference until finally complete falsity is reached. This type (2) response contrasts with another approach sometimes advocated.

In the foregoing account the laws of excluded middle and non-contradiction fail. Consider some sentence '*A*' with truth-value 1/2. By the definition of negation '~*A*' has value 1/2 and so their disjunction also has value 1/2, which is not designated. Thus the law of excluded middle fails. Similarly for the law of non-contradiction. To reinstate these classical laws one can consider all values in the interval [1, 1/2] as designated. In this case all premises of the paradox hold yet *modus ponens* fails and a type (3) response results. In an interpretation where 'Fa_i' takes the value 1/2 and 'Fa_{i+1}' takes the value 0 both 'Fa_i' and 'If Fa_i then Fa_{i+1}' take the value 1/2 and so are designated yet 'Fa_{i+1}' is not. Such a proposal is paraconsistent, admitting contradictions as sometimes taking a designated value; namely when a sentence and its negation both take the value 1/2. Other type (3) responses can be developed by taking the set of designated values to be the interval [1, n] where $1/2 < n < 1$.

There are, however, a number of problems which beset any infinite-valued approach to vagueness. Firstly, the very idea of a degree of truth needs explanation. Secondly, if numerical truth-values are to used some justification seems required for the particular truth-value assignments. Thirdly, the full implications of abandoning the well-understood classical theory in favour of degree theory need spelling out before a proper evaluation of its worth can be made. (On these points see Sainsbury (1995) Ch 2, sec.

6.) Furthermore, it is far from clear whether such an approach successfully avoids problems of higher order vagueness. And the assumption of a totally ordered truth-set is overly simple. Not all natural language sentences are comparable as regards their truth. Due to the multi-dimensional nature of a concept such as redness we may be unable to say of two reddish patches which differ in hue or brightness or colour-saturation whether one is redder than the other. (On these latter points see Williamson (1994) Ch 4, sec. 12-13.)

Embracing the Paradox

A final option is to simply embrace the paradox. (See Dummett (1975), Wright (1975).) Conditional sorites paradoxes are, contrary to appearances, sound. For example, no amount of grains of wheat makes a heap. This initial claim in favour of a universal type (4) response immediately runs into difficulty however with the realisation that, as noted above, such paradoxes come in pairs. There are negative and positive versions depending on whether the soritical predicate is negated or not. To accept all sorites as sound requires assent to the additional claim that, since one grain of wheat makes a heap, any number do. A radical incoherence follows since there is a commitment to all and any number both making a heap and not making a heap. Similarly, everyone is bald and no-one is; everyone is rich and no-one is, and so on.

The problem is that the soundness of any positive conditional sorites undercuts the truth of the unconditional premise of the corresponding negative version, and vice versa. Unless one is prepared to countenance the almost total pervasiveness of contradictions in natural language, it seems that not all sorites can be sound. Unger (1979) and Wheeler (1979) propose a more restricted embrace. Following dissatisfaction with responses of type (1) and (3) one accepts the applicability and validity of classical norms of reasoning. Nonetheless, dissatisfaction with responses of type (2) considered so far -- rejecting some conditional premise -- leaves open the possibility of either rejecting the unconditional premise or accepting it and, with it, the soundness of the paradox. What is advocated is the soundness of those sorites which deny heapness, baldness, hirsuteness, richness, poverty, etc. of everything -- a type (4) response -- and the corresponding falsity of the unconditional premise of all respective positive variants of the argument -- a type (2) response. Terms like ‘heap’, ‘bald’, ‘hirsute’, ‘rich’ and ‘poor’ apply to nothing. It is admitted that they apply to everything if they apply to anything, but the all-or-nothing choice is resolved in favour of the latter option. (See Williamson (1994) Ch 6.)

Bibliography

General

- Sainsbury, M. 1995 (2nd edn). *Paradoxes*. Cambridge: Cambridge University Press. Chapter 2.
- Williamson, T. 1994. *Vagueness*. London: Routledge. Includes detailed exposition and criticism of all the major approaches to the paradox.

Sorites in History

- Barnes, J. 1982. 'Medicine, experience and logic', in J. Barnes, J. Brunschwig, M.F. Burnyeat and M. Schofield (eds), *Science and Speculation*. Cambridge: Cambridge University Press.
- Burnyeat, M.F. 1982. 'Gods and heaps', in M. Schofield and M.C. Nussbaum (eds), *Language and Logos*. Cambridge: Cambridge University Press.
- Williamson, T. 1994. *Vagueness*. London: Routledge. Chapter 1.

Its Paradoxical Forms

- Sainsbury, M. and Williamson, T. 1995. 'Sorites' in B. Hale and C. Wright (eds), *Blackwell Companion to the Philosophy of Language*. Oxford: Blackwell.

Responses

Ideal Language Approaches

- Quine, W.V. 1981. 'What price bivalence?', *Journal of Philosophy* 77: 90-5.
- Russell, B. 1923. 'Vagueness', *The Australian Journal of Philosophy and Psychology* 1: 84-92.

The Epistemic Response

- Sorensen, R. 1988. *Blindspots*. Oxford: Clarendon Press.
- Williamson, T. 1994. *Vagueness*. London: Routledge. Chapters 7-8.

Supervaluationism

- Dummett, M. 1975. 'Wang's paradox', *Synthese* 30: 301-24; reprinted in his *Truth and Other Enigmas*.
- Fine, K. 1975. 'Vagueness, truth and logic', *Synthese* 30: 265-300.

Many-Valued Approaches

- Goguen, J.A. 1969. 'The logic of inexact concepts', *Synthese* 19: 325-373.
- Haack, S. 1974. *Deviant Logic*. Cambridge: Cambridge University Press.
- Halldén, S. 1949. *The Logic of Nonsense*. Uppsala: Uppsala Universitets Arsskrift.
- Körner, S. 1960. *The Philosophy of Mathematics*. London: Hutchinson.
- Tye, M. 1994. 'Sorites paradoxes and the semantics of vagueness', in J. Tomberlin (ed.), *Philosophical Perspectives: Logic and Language*. Atascadero, California: Ridgeview.
- Zadeh, L. 1975. 'Fuzzy logic and approximate reasoning', *Synthese* 30: 407-428.

Embracing the Paradox

- Dummett, M. 1975. 'Wang's paradox', *Synthese* 30: 301-24; reprinted in his *Truth and Other Enigmas*.
- Unger, P. 1979. 'There are no ordinary things', *Synthese* 41: 117-54.
- Wheeler, S.C. 1979. 'On that which is not', *Synthese* 41: 155-94.
- Wright, C. 1975. 'On the coherence of vague predicates', *Synthese* 30: 325-65.

Other Internet Resources

- Bibliography of literature on vagueness and the Sorites Paradox, maintained by Justin Needle (Ph.D./Philosophy, U. Dundee, 1995)

Related Entries

logic: non-classical | logic: paraconsistent | vagueness

Dominic Hyde
D.Hyde@mailbox.uq.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 17, 1997
Content last modified: July 31, 2002

Stanford Encyclopedia of Philosophy
Notes to Bertrand Russell

Notes

1. Bertrand Russell, in a letter to Lady Ottoline Morrell dated 13 December 1911, quoted in John G. Slater, *Bertrand Russell,* Bristol: Thoemmes, 1994, p. 67.

2. For example, see Nicholas Griffin, *Russell's Idealist Apprenticeship*, Oxford: Clarendon, 1991 and Peter W. Hylton, *Russell, Idealism, and the Emergence of Analytic Philosophy*, Oxford: Clarendon, 1990.

3. For example, see Paul J. Hager, *Continuity and Change in the Development of Russell's Philosophy,* Dordrecht: Nijhoff, 1994 and Morris Weitz, "Analysis and the Unity of Russell's Philosophy", in Paul Arthur Schilpp, *The Philosophy of Bertrand Russell,* 3rd ed., New York: Tudor, 1951, pp. 55-121.

4. For example, see A.D. Irvine, "Epistemic Logicism and Russell's Regressive Method", *Philosophical Studies,* 55 (1989), 303-327.

5. Bertrand Russell, *Sceptical Essays*, New York: Norton, 1928, p. 11.

A. D. Irvine
andrew.irvine@ubc.ca

First published: July 20, 2000
Content last modified: July 20, 2000

Logical Constructions

Bertrand Russell referred to several different definitions and philosophical analyses as providing "logical constructions" of certain entities and expressions. Examples he cited were the Frege/Russell definition of numbers as classes of equinumerous classes, the theory of definite descriptions, the construction of matter from sense data, and several others. Generally expressions for such entities are called "incomplete symbols" and the entities themselves "logical fictions". The notion originates with Russell's logicist program of reducing mathematics to logic, was widely used by Russell, and led to the later Logical Positivist notion of construction and ultimately the widespread use of set theoretic models in philosophy.

Honest Toil

Russell's most specific formulation of logical construction as a method in Philosophy comes from his essay "Logical Atomism":

> One very important heuristic maxim which Dr.Whitehead and I found, by experience, to be applicable in mathematical logic,and have since applied to various other fields, is a form of Occam's Razor. When some set of supposed entities has neat logical properties, it turns out, in a great many instances, that the supposed entities can be replaced by purely logical structures composed of entities which have not such neat properties. In that case, in interpreting a body of propositions hitherto believed to be about the supposed entities, we can substitute the logical structures without altering any of the detail of the body of propositions in question.This is an economy, because entities with neat logical properties are always inferred, and if the propositions in which they occur can be interpreted without making this inference, the ground for the inference fails, and our body of propositions is

> secured against the need of a doubtful step. The principle may be stated in the form: ‘Whenever possible, substitute constructions out of known entities for inferences to unknown entities’ (1924, p.160)

Russell was speaking of logical constructions in this memorable passage from his *Introduction to Mathematical Philosophy*: "The method of ‘postulating’ what we want has many advantages; they are the same as the advantages of theft over honest toil. Let us leave them to others and proceed with our honest toil." (1919, p. 71)

The notion of logical construction appears frequently with the idea that what is defined is a "logical fiction", and an "incomplete symbol". The latter term derives from the use of contextual definitions, providing an analysis of each sentence in which a defined symbol may occur without, however, giving an explicit definition, an equation or universal statement giving necessary and sufficient conditions for the application of the term in isolation. The terms "fiction" and "incomplete symbol" apply with differing aptness to various constructions.

Russell's first use of construction, and the model for later constructions, is the Frege/Russell definition of numbers as classes. This follows the kind of definitions used in the arithmetization of analysis of the preceding century, in particular, Dedekind's earlier construction of real numbers as bounded classes in the rational numbers. Russell's logicist program could not rest content with postulates for the fundamental objects of mathematics such as the Peano Axioms for the natural numbers. Instead numbers were to be defined as classes of equinumerous classes. Russell also refers to this method as "abstraction", now known as the abstraction of an equivalence class. The definition of equinumerosity, or of the existence of a one to one mapping between two classes, also called "similarity", is solely in terms of logical notions of quantifiers and identity. With the numbers defined, for example, two as the class of all two membered sets, or pairs, the properties of numbers could be derived by logical means alone.

Definite Descriptions and Classes

The most influential of Russell's constructions was the *theory of descriptions* from his paper "On Denoting" in 1905. Russell's theory provides an analysis of sentences of the form ‘The F is G’ where ‘The F’ is called a *definite decription*. The analysis proposes that ‘The F is G’ is equivalent to ‘There is one and only one F and it is G’. With this analysis, the logical properties of descriptions can now be deduced using just the logic of quantifiers and identity. Among the theorems in *14 of *Principia Mathematica* are those showing that, (1) if there is just one F then ‘The F is F’ is true, and if there is not, then ‘The F is G’ is always false and, crucially for the logical manipulation of descriptions, (2) if the F = the G, and the F is H, then the G is H. In other words, proper (uniquely referring) descriptions behave like singular terms. Some of these results are contentious---Strawson noted that ‘The present king of france is bald’ should be truth valueless since there is no present king of France, rather than "plainly false", as Russell's theory predicts.

The theory of descriptions introduces Russell's notion of *incomplete symbol*. Definite descriptions ‘The

F' do not show up in the formal analysis of sentences in which they occur, thus 'The F is H' becomes:

$$(\exists x) [(y)(Fy \leftrightarrow y=x) \& Hx]$$

of which no subformula, or continuous segment, can be identified as the analysis of 'The F'. Much as talk about "the average family" as in "The average family has 2.2 children" becomes "The number of children in families divided by the number of families = 2.2", there is no portion of that analysis that corresponds with "the average family". Instead we have a formula for eliminating such expressions from contexts in which they occur, hence the notion of "incomplete symbol" and the related "contextual definition". It is standard to see in this the origins of the distinction between between surface grammatical form and logical form, and thus the origin of linguistic analysis as a method in philosophy which operates by seeing past superficial linguistic form to underlying philosophical analysis. The theory of descriptions has been criticized by some linguists who see descriptions and other noun phrases as full fledged constituents of sentences, and who see the sharp distinction between grammatical and logical form as a mistake.

The theory of descriptions is often described as a model for avoiding ontological commitment to objects such as Meinongian subsistent entities, and so logical constructions in general are often seen as being chiefly aimed at ontological goals. In fact, that goal is at most peripheral to most constructions. Rather the goal is to allow the proof of propositions that would otherwise have to be assumed as axioms or hypotheses. Nor need the ontological goal be always elimination of problematic entities. Other constructions should be seen more as reductions of one class of entity to another, or replacements of one notion by a more precise, mathematical, substitute.

Russell's "No-Class" theory of classes from *20 of *Principia Mathematica* provides a contextual definition like the theory of descriptions. One of Russell's early diagnoses of the paradoxes was that they showed that classes could not be objects. Indeed he seems to have come across his paradox of the class of all classes that are not members of themselves by applying Cantor's argument to show that there are more classes of objects than objects. Hence, he concluded, classes could not be objects. Inspired by the theory of descriptions, Russell proposed that to say something G of the class of Fs, G{x: Fx}, is to say that there is some property H coextensive with (true of the same things as) F such that H is G. Extensionality of sets is thus derivable, rather than postulated. If F and H are coextensive then anything true of {x: Fx} will be true of {x: Hx}. Features of sets then follow from the features of the logic of properties, the "ramified theory of types". Because classes would seem to be individuals of some sort, but on analysis are found not to be, Russell speaks of them as "logical fictions", an expression which echoes Jeremy Bentham's notion of a "legal fiction". Because statements attributing a property to particular classes are analyzed by existential sentences saying that there is some propositional function having that property, this construction should not be seen as avoiding ontological commitment entirely, but rather of reducing classes to propositional functions. The properties of classes are really properties of propositional functions and for every class said to have a property there really is some propositional function having that property.

Other Constructions

For other constructions such as propositions a contextual definition is not provided. In any case, constructions do not appear as the referents of logically proper names, and so by that account are not part of the fundamental "furniture" of the world. (Early critical discussions of constructions, such as Wisdom's, stressed the contrast between logically proper names, which do refer, and constructions, which were thus seen as ontologically innocent.)

Beginning with *The Problems of Philosophy* in 1912, Russell turned repeatedly to the problem of matter. Part of the problem is to find a refution of Berkeleyan idealism, of showing how the existence and real nature of matter can be proved. In *Problems* Russell argues that matter is a well supported hypothesis that explains our experiences. Matter is known only indirectly, "by description", as the cause, whatever it may be, of our sense data, which we know "by acquaintance". This is the notion of hypothesis which Russell contrasts with construction in the passage above. Russell saw an analogy between the case of simply hypothesizing the existence of numbers with certain properties, those described by axioms, and hypothesizing the existence of matter. While we distinguish the certain knowledge we may have of mathematical entities from the contingent knowledge of material objects, Russell says that there are certain "neat" features of matter which are just too tidy to have turned out by accident. Examples include the most general spatiotemporal properties of objects, that no two can occupy the same place at the same time, and so on. Material objects are now to be seen as collections of sense data. Influenced by William James, Russell defended a "neutral monism" by which matter and minds were both to be constructed from sense data, but in different ways. Intuitively, the sense data occuring as they do "in" a mind, are material to construct that mind, the sense data derived from an object from different points of view to constructthat object. Russell saw some support for this in the theory of relativity, and the fundumental importance of frames of reference in the new physics.

These prominent examples are not the only use of the notion of construction in Russell's thought. In *Principia Mathematica* the *multiple relation* theory of propositions is introduced by saying that propositions are "incomplete symbols". Russell's multiple relation theory, that he held from 1910 to 1919 or so, argued that the constitituents of propositions, say ‘Desdemona loves Cassio’, which is false, are unified in a way that does not make it the case that they constitute a fact by themselves. Those constituents occur only in the context of beliefs, say, ‘Othello judges that Desdemona loves Cassio’. The real fact consists of a relation of Belief holding between the constituents Othello, Desdemona and Cassio, thus B(o, d , L, c). Because one might also have believed propositions of other structures, such as B (o, F, a) there need to be many such relations B, thus the "multiple" relation theory. Like the construction of numbers, this construction abstracts out what a number of occurrences of a belief have in common, a believer and various objects in a certain order. The analysis also makes the proposition an incomplete symbol because there is no constituent in the analysis of ‘x believes that p’ that corresponds to ‘p’.

Russell also suggests that propositional functions are logical constructions when he says that they are "nothing", but "nonetheless important for that". (1918, p. 96) Propositional functions are abstracted from their values, propositions. The propositional function ‘x is human’ is abstracted from its values ‘Socrates

is human', 'Plato is human' , etc. Viewing propositional functions as constructions from propositions which are in turn constructions by the multiple relation theory helps to make sense of the theory of types of propositional functions in *Principia Mathematica*. The notion of "incomplete symbol" does not make as much sense as "construction" when applied to propositional functions and propositions. This usage requires a broadening of the notion.

The notion of logical construction had a great impact on the future course of analytic philosophy. One line of influence was via the notion of a contextual definition, or paraphrase, intended to minimize ontological commitment and to be a model of philosophical analysis. The distinction between the surface appearance of definite descriptions, as singular terms, and the fully analyzed sentences from which they seem to disappear was seen as a model for making problematic notions disappear upon analysis. The theory of descriptions has been viewed as a paradigm of philosophical analysis.

A more technical strand in analytic philosophy was influenced by the construction of matter. Rudolf Carnap was attempted to carry out the construction of matter from sense data, and later Nelson Goodman continued the project. More generally, however, the use of set theoretic constructions became widespread among philosophers, and continues in the construction of set theoretic models, both in the sense of logic where they model formal theories, and as objects of interest in their own right.

Bibliography

- Carnap, R., *The Logical Structure of the World & Pseudo Problems in Philosophy*, trans. R.George, Berkeley: University of California Press, 1967.
- Goodman, N., *The Structure of Appearance*, Cambridge Mass: Harvard University Press, 1951.
- Russell, B., 1919, *Introduction to Mathematical Philosophy* , London: Routledge, reprinted 1993.
- Russell, B., 1905, "On Denoting", in Robert Marsh, *Logic and Knowledge: Essays 1901-1950* , London: George Allen and Unwin, 1956, 39-56.
- Russell, B., 1918, "The Philosophy of Logical Atomism" in *The Philosophy of Logical Atomism* , D.F.Pears, ed. Lasalle: Open Court, 1985, 35-155.
- Russell, B., 1924, "Logical Atomism", in *The Philosophy of Logical Atomism* , D.F.Pears, ed., Lasalle: Open Court, 1985, 157-181.
- Russell, B., 1912, *The Problems of Philosophy* , Oxford: Oxford University Press, reprinted 1967.
- Whitehead, A.N., and Russell,B.: 1925, *Principia Mathematica* Vol.I., second ed., Cambridge: Cambridge University Press, 1925.
- Wisdom, J., 1931, "Logical Constructions (I.).", *Mind* , **XL** , April, 188-216.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Carnap, Rudolf | descriptions | Russell, Bertrand | Russell's paradox

Bernard Linsky
bernie@cs.ualberta.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 20, 1996
Content last modified: June 29, 2001

Voluntary Euthanasia

The entry sets out five individually necessary conditions for anyone to be a candidate for legalised voluntary euthanasia (or, in some usages, physician-assisted suicide), outlines the moral case advanced by those in favour of legalising voluntary euthanasia, and discusses six of the more important objections made by those opposed to the legality of voluntary euthanasia.

Introduction

When a person commits an act of euthanasia he brings about the death of another person because he believes the latter's present existence is so bad that she would be better off dead, or believes that unless he intervenes and ends her life, it will become so bad that she would be better off dead. The motive of the person who commits an act of euthanasia is to benefit the one whose death is brought about. (Though what was just said also holds for many instances of physician-assisted suicide, some wish to restrict the use of the latter term to forms of assistance which stop short of the physician 'bringing about the death' of the patient, such as those involving mechanical means which have to be activated by the patient.)

Our concern will be with *voluntary* euthanasia—that is, with those instances of euthanasia in which a clearly competent person makes a voluntary and enduring request to be helped to die. There shall be occasion to mention *non-voluntary euthanasia*—instances of euthanasia where a person is either not competent to, or unable to, express a wish about euthanasia, and there is no one authorised to make a substituted judgment (wherein a proxy chooses as the no longer competent patient would have chosen had she remained competent)— in the context of considering the claim that permitting voluntary euthanasia will lead via a slippery slope to permitting non-voluntary euthanasia. Nothing will be said here about *involuntary euthanasia*, where a competent person's life is brought to an end despite an explicit

expression of opposition to euthanasia, beyond saying that, no matter how honourable the perpetrator's motive, such a death is, and ought to be, unlawful.

Debate about the morality and legality of voluntary euthanasia is, for the most part, a phenomenon of the second half of the twentieth century. Certainly, the ancient Greeks and Romans did not consider life needed to be preserved at any cost and were, in consequence, tolerant of suicide in cases where no relief could be offered to the dying or, in the case of the Stoics and Epicureans, where a person no longer cared for his life. In the sixteenth century, Thomas More, in describing a utopian community, envisaged such a community as one that would facilitate the death of those whose lives had become burdensome as a result of ‘torturing and lingering pain’. But it has only been in the last hundred years that there have been concerted efforts to make legal provision for voluntary euthanasia. Until quite recently there had been no success in obtaining such legal provision (though assisted suicide has been legally tolerated in Switzerland for many years). However, in, the nineteen seventies and eighties a series of court cases in The Netherlands culminated in agreement being reached between the legal and medical authorities to ensure that no physician would be prosecuted for assisting a patient to die as long as certain guidelines were strictly adhered to (see Griffiths, et al. 1998) In brief, the guidelines were established to permit physicians to practise voluntary euthanasia in instances where a competent patient had made a voluntary and informed decision to die, the patient's suffering was unbearable, there was no way of making that suffering bearable which was acceptable to the patient, and the physician's judgements as to diagnosis and prognosis were confirmed after consultation with another physician. In the nineteen nineties the first legislative approval for voluntary euthanasia was achieved with the passage of a bill in the parliament of Australia's Northern Territory to enable physicians to practise voluntary euthanasia. Subsequent to the Act's proclamation in 1996 it faced a series of legal challenges from opponents of voluntary euthanasia. In 1997 the challenges culminated in the Australian National Parliament overturning the legislation when it prohibited Australian Territories (the Australian Capital Territory and the Northern Territory) from enacting legislation to permit euthanasia. In Oregon in the United States legislation was introduced in 1997 to permit physician-assisted suicide when a second referendum clearly endorsed the proposed legislation. Later in 1997 the Supreme Court of the United States ruled that there is no constitutional right to physician-assisted suicide. However, the Court did not preclude individual States from legislating in favour of physician-assisted suicide. The Oregon legislation has, in consequence, remained operative and has been successfully utilised by a number of people. In November 2000 The Netherlands passed legislation to legalise the practice of voluntary euthanasia. The legislation passed through all the parliamentary stages early in 2001 and so became law. The Belgium parliament passed similar legislation in May 2002.

With that brief sketch of the historical background in place, we now proceed to set out the conditions which those who have advocated making voluntary euthanasia legally permissible have wished to insist should be satisfied. The conditions are stated with some care so as to focus the moral debate about legislation. Second, we shall go on to outline the positive moral case underpinning the push to make voluntary euthanasia legally permissible. Third, we shall then consider the more important of the morally grounded objections which have been advanced by those opposed to the legalisation of voluntary euthanasia.

Five Individually Necessary Conditions for Candidacy for Voluntary Euthanasia

Advocates of voluntary euthanasia contend that if a person is

> (a) suffering from a terminal illness;
>
> (b) unlikely to benefit from the discovery of a cure for that illness during what remains of her life expectancy;
>
> (c) as a direct result of the illness, either suffering intolerable pain, or only has available a life that is unacceptably burdensome (because the illness has to be treated in ways which lead to her being unacceptably dependent on others or on technological means of life support);
>
> (d) has an enduring, voluntary and competent wish to die (or has, prior to losing the competence to do so, expressed a wish to die in the event that conditions (a)-(c) are satisfied); and
>
> (e) unable without assistance to commit suicide,

then there should be legal and medical provision to enable her to be allowed to die or assisted to die.

It should be acknowledged that these conditions are quite restrictive, indeed more restrictive than some would think appropriate. In particular, the conditions concern access only to voluntary euthanasia for those who are *terminally ill*. While that expression is not free of all ambiguity, for present purposes it can be agreed that it does not include the bringing about of the death of, say, victims of accidents who are rendered quadriplegic or victims of early Alzheimer's Disease. Those who consider that such cases show the first condition to be too restrictive may nonetheless accept that including them would, at least for the time being, make it far harder to obtain legal protection for helping those terminally ill persons who wish to die. The fifth condition further restricts access to voluntary euthanasia by excluding those capable of ending their own lives, and so will not only be thought unduly restrictive by those who think physician-assisted suicide a better course to follow, but will be considered morally much harder to justify by those who think health care practitioners may never justifiably kill their patients. More on this anon.

The second condition is intended simply to reflect the fact that we normally are able to say that someone's health status is incurable. So-called ‘miracle’ cures may be spoken of by sensationalist journalists, but progress toward medical breakthroughs is typically painstaking. If there are miracles wrought by God that will be quite another matter entirely, but it is at least clear that not everyone's death is thus to be staved off.

The third condition recognises what many who oppose the legalisation of voluntary euthanasia do not, namely that it is not only release from pain that leads people to want to be helped to die. In The Netherlands, for example, it has been found to be a less significant reason for requesting assistance with dying than other forms of suffering and frustration with loss of independence. Sufferers from some terminal conditions may have their pain relieved but have to endure side effects that for them make life unbearable. Others may not have to cope with pain but instead be incapable, as with motor neurone disease, of living without life supports which at the same time rob their lives of quality.

A final preliminary point is that the fourth condition requires that the choice to die not only be voluntary but that it be made in an enduring (not merely a one-off) way and be competent. The choice is one that will require discussion and time for reflection and so should not be settled in a moment. As in other decisions affecting matters of importance, normal adults are presumed to choose voluntarily unless the presence of defeating considerations can be established. The onus of establishing lack of voluntariness or lack of competence is on those who refuse to accept the person's choice. There is no need to deny that it can sometimes be met (e.g. by pointing to the person's being in a state of clinical depression). The claim is only that the onus falls on those who deny that a normal adult's choice is not competent.

A Moral Case for Voluntary Euthanasia

The central ethical argument for voluntary euthanasia—that respect for persons demands respect for their autonomous choices as long as those choices do not result in harm to others—is directly connected with this issue of competence (cp. Brock, 1992) because autonomy presupposes competence. People have an interest in making important decisions about their lives in accordance with their own conception of how they want their lives to go. In exercising autonomy or self-determination people take responsibility for their lives and, since dying is a part of life, choices about the manner of their dying and the timing of their death are, for many people, part of what is involved in taking responsibility for their lives. Most people are concerned about what the last phase of their lives will be like, not merely because of fears that their dying might involve them in great suffering, but also because of the desire to retain their dignity and as much control over their lives as possible during this phase.

The technological interventions of modern medicine have had an effect on how drawn out the dying phase may be. Sometimes this added life is an occasion for rejoicing, sometimes it may serve to stretch out the period of significant physical and intellectual decline in such a way as to impair and burden the end of life so that life comes to be no longer worth living. There is no single, objectively correct answer, which has application to everyone, as to when, if at all, life becomes a burden and unwanted. But that simply points up the importance of individuals being able to decide autonomously for themselves whether their own lives retain sufficient quality and dignity. In making such decisions individuals decide about the mix between their self-determination and their well-being that suits them. Given that a critically ill person is typically in a severely compromised and debilitated state it is, other things being equal, the patient's judgement of whether continued life is a benefit that must carry the greatest weight, provided always that the patient is competent.

Suppose it is agreed that a person's exercise of her autonomy warrants our respect. If medical assistance is to be provided to help a person achieve her autonomously chosen goal of an easeful death (because she cannot end her own life), the autonomy of the assisting professional(s) also has to be respected. The value (or right) of self-determination does not entitle a patient to compel a medical professional to act contrary to her own moral or professional values. If voluntary euthanasia is to be legally permitted it must be against a backdrop of respect for professional autonomy. Thus, if a doctor's view of her moral or professional responsibilities is at odds with the request of her patient for euthanasia, provision must be made for the transfer of the patient's care to another who faces no such conflict.

Opponents of voluntary euthanasia have endeavoured to counter this very straightforward moral case for the practice in a variety of ways. Some of the counter-arguments are concerned only with whether the moral case warrants making the practice of voluntary euthanasia legal, others are concerned with trying to undermine the moral case itself. In what follows, consideration will be given to the six most important of the counter-arguments. (Some less important moral objections to the practice of voluntary euthanasia are considered in Young, 1976, esp. pp. 265-275.)

Six Objections to the Moral Permissibility of Voluntary Euthanasia

Objection 1

It is often said that it is not necessary nowadays for anyone to die while suffering from intolerable or overwhelming pain. We are getting better at providing effective palliative care and hospice care is available. Given these considerations it is urged that voluntary euthanasia is unnecessary.

There are several flaws in this counter-argument. First, while both good palliative care and hospice care make important contributions to the care of the dying neither is a panacea. To get the best palliative care for an individual involves trial and error with some consequent suffering in the process. But, far more importantly, even high quality palliative care commonly exacts a price in the form of side effects such as nausea, incontinence, loss of awareness because of semi-permanent drowsiness, and so on. A rosy picture is often painted as to how palliative care can transform the plight of the dying. Such a picture is misleading according to those who have closely observed the effect of extended courses of treatment with drugs like morphine, a point acknowledged as well by many skilled palliative care specialists. Second, though the sort of care provided through hospices is to be applauded, it is care that is available only to a small proportion of the terminally ill and then usually only in the very last stages of the illness (typically a matter of a few weeks). Third, the point of greatest significance is that not everyone wishes to avail themselves of either palliative care or hospice care. For those who prefer to die in their own way and in their own time neither palliative care nor hospice care may be attractive. For many dying patients it is having their autonomous wishes frustrated that is a source of the deepest distress. Fourth, as indicated earlier when the conditions under which voluntary euthanasia is advocated were outlined, not everyone who is dying is suffering because of the pain occasioned by their illness. For those for whom what is

intolerable is their dependence on others or on machinery, the availability of effective pain control will be quite irrelevant.

Objection 2

A second, related objection to permitting the legalisation of voluntary euthanasia is to the effect that we never have sufficient evidence to be justified in believing that a dying person's request to be helped to die is competent, enduring and genuinely voluntary.

Notice first that a request to die may not reflect an enduring desire to die (cf. some attempts to commit suicide may similarly reflect temporary despair). That is why advocates of voluntary euthanasia have argued that normally a cooling off period should be allowed. But that said, the objection claims we can *never* be justified in believing someone's request to die reflects a settled preference for death. This goes too far. If someone discusses the issue with others on different occasions, or reflects on the issue over an extended period, and does not waver in her conviction, her wish to die is surely an enduring one.

But, it might be said, what if a person is racked with pain, or befuddled because of the measures taken to relieve her pain, and so not able to think clearly and rationally about the alternatives? It has to be agreed that a person in those circumstances who wants to die cannot be assumed to have a competent, enduring and genuinely voluntary desire to die. However, there are at least two important points to make about those in such circumstances. First, they do not account for all of the terminally ill, so even if it is acknowledged that such people are incapable of agreeing to voluntary euthanasia that does not show that no one can ever voluntarily request help to die. Second, it is possible for a person to indicate in advance of losing the capacity to give competent, enduring and voluntary consent, how she would wish to be treated should she become terminally ill and be suffering intolerably from pain or from loss of control over her life. ‘Living wills’ or ‘advance declarations’ are legally useful instruments for giving voice to people's wishes while they are capable of giving competent, enduring and voluntary consent, including to their wanting help to die. As long as they are easily revocable in the event of a change of mind (just as ordinary wills are), they should be respected as evidence of a well thought out conviction. It should be noted, though, that any request for voluntary euthanasia or physician-assisted suicide will not be able lawfully to be implemented (outside of The Netherlands, Belgium and Oregon).

Perhaps, though, what is really at issue in this objection is whether anyone can ever form a competent, enduring and voluntary wish about being better off dead rather than continuing to suffer from an illness *before actually suffering the illness*. If this is what underlies the objection it is surely too paternalistic to be acceptable. Why cannot a person have sufficient inductive evidence (e.g. based on the experience of the deaths of friends or family) to know her own mind and act accordingly?

Objection 3

According to one interpretation of the traditional ‘doctrine of double effect’ it is permissible to act in ways which it is foreseen will have bad consequences provided only that

(a) this occurs as a side effect (or indirectly) to the achievement of the act which is directly aimed at or intended;

(b) the act directly aimed at is itself morally good or, at least, morally neutral;

(c) the good effect is not achieved by way of the bad, that is, the bad must not be a means to the good; and

(d) the bad consequences must not be so serious as to outweigh the good effect.

In line with the doctrine of double effect it is, for example, held to be permissible to alleviate pain by administering drugs like morphine which it is foreseen will shorten life, whereas to give an overdose or injection with the direct intention of terminating a patient's life (whether at her request or not) is considered morally indefensible. This is not the appropriate forum to give full consideration to this doctrine. However, there is one vital criticism to be made of the doctrine concerning its relevance to the issue of voluntary euthanasia. With that point made we will be able to turn to the more general question of the moral permissibility of intentional killing.

The criticism of the relevance of the doctrine of double effect to any critique of voluntary euthanasia, at least on what seems to me to be a defensible reading of that doctrine, is simply this: the doctrine can only be relevant where a person's death is an evil or, to put it another way, a *harm*. Sometimes 'harm' is understood simply as damage to a person's interest whether consented to or not. At other times it is more strictly understood as wrongfully inflicted damage. On either account, if the death of a person who wishes to die is not harmful (because from that person's standpoint it is, in fact, beneficial), the doctrine of double effect can have no relevance to the debate about the permissibility of voluntary euthanasia. (For an extended discussion of the doctrine of double effect and its bearing on the moral permissibility of voluntary euthanasia see McIntyre, 2001.)

Objection 4

There is a widespread belief that passive (voluntary) euthanasia, where life-sustaining or life-prolonging measures are withdrawn or withheld, is morally acceptable because steps are simply not taken which could preserve or prolong life (and so a patient is allowed to die), whereas active (voluntary) euthanasia is not, because it requires an act of killing. The distinction, despite its widespread popularity, is very unclear. Whether behaviour is described in terms of acts or omissions (which underpins the alleged distinction between active and passive (voluntary) euthanasia), is generally a matter of pragmatics not of anything of deeper importance. Consider, for instance, the practice of deliberately proceeding slowly to a ward in response to a request to provide assistance for a patient who is subject to a 'not for resuscitation' code. Or consider 'pulling the plug' on an oxygen machine keeping an otherwise dying patient alive as against not replacing the tank when it runs out. Are these acts or omissions; cases of passive euthanasia or active euthanasia?

More fundamentally, though, those who think some reliance can be placed on the distinction think that, at least in a medical context, killing is morally worse than letting die. Consider the case of a patient suffering from motor neurone disease who is completely respirator dependent, finds her condition intolerable, and competently and persistently requests to be removed from the respirator so that she may die. Even the Catholic Church in recent times has been prepared to agree in cases like this one to the turning off of the respirator. Is this merely a case of letting the patient die?

It is often said that even if *motives* and *consequences* are agreed to be in common, if someone's life is *intentionally* terminated she has been killed, whereas if she is no longer being aggressively treated her life is not ended by the withdrawal of such aggressive treatment but by the underlying disease.One way to show that it is in most cases implausible to think that the withdrawal of life sustaining measures involves no intention to terminate the patient's life is to consider the growing practice of withholding artificial nutrition and hydration in those instances where a decision has been made to cease aggressive treatment, and then to see if we can generalise to cases like that of the motor neurone sufferer (cf. Winkler, 1995). Many physicians would say that their intention in withholding life-sustaining artificial nutrition is simply to respect the patient's wishes, and this is plausible in those instances where the patient is still able competently to ask for such treatment no longer to be given (or the patient's proxy makes the request). However, unless there has been such a request from a competent patient (or the patient's proxy), the best explanation of the physician's behaviour in withdrawing life-sustaining nutrition will be that the physician intends thereby to end the life of the patient. Permanently withdrawing nutrition from someone in, say, an irreversible coma (a persistent vegetative state), thereby starving the patient, is not merely to foresee that death will ensue, but to intend the death. What could be the point of the action, the goal aimed at, the intended outcome, if not the ending of the patient's life? No sense can be made of the action as being intended to serve to palliate the disease, or to keep the patient comfortable, or even, in the case of a person in a permanently vegetative state, as allowing the underlying disease to carry the person off. The loss of brain activity is not going to kill the person. What is going to kill the patient is the act of starving her to death. That is the clear intention, not merely something foreseen as an unfortunate side effect, but in no way aimed at.

Can this claim be extended to other circumstances than those involving the withdrawal of life-sustaining nutrition? The giving of massive doses of morphine, way beyond what is needed to control pain, or the removal of a respirator from a sufferer from motor neurone disease would seem, by parallel reasoning, to amount to the intentional bringing about of the death of the person being cared for.

So that there is no misunderstanding, it should be conceded that there are circumstances where doctors can truthfully say that actions which they perform, or omissions which they make, do lead to the deaths of their patients without them intending that those patients should die. Thus, for instance, if a patient refuses life prolonging medical treatment because she considers it useless, it might reasonably be said that the doctor's intention in complying is simply to respect the patient's wishes. But the point made earlier was of much wider significance and was aimed at showing that it is utterly stilted to claim, as some doctors do, that it can never be the intention in performing certain actions and omissions to intend to bring about death and hence that those actions and omissions cannot count as killings.

Two final points need to be added to round off the discussion of the fourth objection. First, much of the debate surrounding the objection is premised on the belief that killing, at least in medical contexts, cannot morally be justified. For that reason alone the medical profession has long found psychological comfort in the belief that even if killing cannot be justified it is quite another thing to allow a patient to die (where that involves no negligence) because there the cause of death is natural. This underlying assumption is one that is open to challenge (and has been challenged in e.g. Rachels, 1986, chs. 7, 8; Kuhse, 1987). First, there will be cases, namely those where someone who has requested assistance to die and is allowed to die, rather than killed, where it is morally worse to allow to die because all that does is prolong the patient's suffering. The second point to make is that despite the longstanding legal doctrine that no one can justifiably consent to be killed (on which more later), it surely is relevant to the justification of an act of killing someone that she has autonomously decided that that would be best for her.

Objection 5

It is often said that if society allows *voluntary* euthanasia to be legally permitted we will have set foot on a slippery slope that will lead us inevitably to support other forms of euthanasia, especially non-voluntary euthanasia. Whereas it was once the common refrain that that was precisely what happened in Hitler's Germany, nowadays the claim tends to be that the experience of The Netherlands in the last decade or so confirms the reality of the slippery slope. Slippery slope arguments come in at least three different versions: logical, psychological and arbitrary line. What the different forms share is the contention that once the first step is taken on a slippery slope the subsequent steps follow inexorably, whether for logical reasons, psychological reasons or to avoid arbitrariness in ‘drawing a line’ across a person's actions. (For further discussion see e.g. Rachels, 1986, ch. 10; Brock, 1992, pp. 19ff.).

We first consider why, at the theoretical level, none of these forms of argument appears powerful enough to trouble an advocate of the legalisation of voluntary euthanasia. We then comment on the alleged empirical support from the experiences of Hitler's Germany and The Netherlands of today for the existence of a slippery slope beginning from voluntary euthanasia.

There is nothing logically inconsistent in supporting voluntary euthanasia but rejecting non-voluntary euthanasia as morally inappropriate. Since the two issues are logically separate there will be some advocates of voluntary euthanasia who will wish also to lend their support to some acts of non-voluntary euthanasia (e.g. for those in persistent vegetative states who have never indicated their wishes about being helped to die or for some severely disabled infants for whom the outlook is hopeless). Others will think that what may be done with the consent of the patient sets a strict limit on the practice of euthanasia. The difference is not one of logical acumen. It has to be located in the respective values of the different supporters (e.g. whether self-determination alone or the best interests of a person should prevail).

As regards the alleged psychological inevitability of moving from voluntary to non-voluntary euthanasia (where there is no way of knowing the patient's views because the patient is neither competent nor has made any provision for a proxy to make a substituted judgment), again it is hard to see the alleged inevitability. Why should it be supposed that those who value the autonomy of the individual and so

support provision for voluntary euthanasia will, as a result, find it psychologically easier to kill patients who are not able competently to request assistance with dying? What reason is there to believe that they will, as a direct result of their support for voluntary euthanasia, be psychologically driven to practise non-voluntary euthanasia?

Finally, if there is nothing arbitrary about distinguishing voluntary euthanasia from non-voluntary euthanasia (because the line between them is based on clear principles) there can be no substance to the charge that there is a slide from voluntary to non-voluntary euthanasia that can only be prevented by arbitrarily drawing a line between them.

What, though, of Hitler's Germany and The Netherlands of today? The former is easily dismissed as a provider of evidence for an inevitable descent from voluntary euthanasia to non-voluntary. There never was a policy in favor of, or a legal practice of, voluntary euthanasia in Germany in the 1920s to the 1940s (see, for example, Burleigh (1994)). There was, prior to Hitler coming to power, a clear practice of killing some disabled persons. The justification was never suggested to be that their being killed was in *their* best interests, rather it was said to be society that benefited. Hitler's later revival of the practice and its widening to take in other groups such as Jews and gypsies was part of a program of *eugenics*, not euthanasia.

Since the publication of the Remmelink Report in 1991 into the medical practice of euthanasia in The Netherlands it has frequently been said that the Dutch experience shows decisively that legally protecting voluntary euthanasia is impossible without also affording protection to the non-voluntary euthanasia that will come in its train. Unfortunately, many of those who have made this claim have paid insufficient attention to the serious studies carried out by van der Maas, *et al.* (1991), and van der Wal, *et al.* (1992a and 1992b) into what the Report revealed. In a second nation-wide investigation of physician-assisted dying in the Netherlands carried out in 1995 a similar picture emerged as had in the earlier Remmelink Report. Again no evidence was found of any descent down a slippery slope toward ignoring people's voluntary choices to be assisted to die (see van der Maas, *et al.* (1996); van der Wal, *et al.* (1996); Griffiths, *et al.* (1998)). The true picture is that, of those terminally ill persons assisted to die under the agreement between the legal and medical authorities, a little over one half were clearly cases of voluntary euthanasia as it has been characterised in this article. Of the remainder, the vast majority of cases were of patients who at the time of the assisted death were no longer competent. The deaths of some of these were brought about by withdrawal of treatment, that of others by interventions such as the giving of lethal doses of anaesthetics. But there are two critical points to be made: first, in the overwhelming majority of such cases the decision to end life was taken after consultation between the doctor(s) and family members and, second, according to the researchers, most of the cases are to be seen as like the practice common in other countries where voluntary euthanasia is not legally tolerated of giving large doses of opioids to relieve pain knowing all the while that this will also end life. It is true that in a very few cases of this kind there was no consultation with relatives, only with other medical personnel. This is explained by the researchers as having occurred because families in The Netherlands strictly have no final authority to act as surrogate decision-makers for incompetent persons. That there have only been a handful of prosecutions of Dutch doctors for failing to follow agreed procedures (Griffiths, *et al.* (1998)), that none of the doctors prosecuted has had a significant penalty imposed, and that the Dutch public have regularly

reaffirmed their support for those agreed procedures suggests that, contrary to the claims of some critics of The Netherlands' experience of legally protecting voluntary euthanasia, social life has not broken down. Indeed, such studies as have been published about what happens in other countries, like Australia (see Kuhse, *et al.* (1997)), where no legal protection is in place, suggest that the pattern of things in The Netherlands and elsewhere is quite similar. If active euthanasia is widely practised but in ways that are not legally recognized there is apt in fact to be more danger that the distinction between voluntary cases and non-voluntary ones will be blurred or ignored than in a situation where the carrying out of euthanasia is transparent and subject to monitoring.

We can bring this discussion of the fifth objection to a close with two observations. First, nothing that has been said should be taken as suggesting that there is no need to put in place safeguards against potential abuse of any legal protection for voluntary euthanasia. This is particularly important for those who have become incompetent by the time decisions need to be taken about assisting them to die. As was mentioned very early on, there are ways of addressing this issue (such as by way of advance declarations or living wills) which are widely thought to be effective, even if they are not perfect. The main point to be stressed at the present, though, is that there is surely no need for anyone to be frightened into thinking that the legalisation of voluntary euthanasia will inevitably end in her having her life snatched away from her should she become incapable of exercising a competent judgement on her own behalf. Second, it is, of course, possible that the reform of any law may have unintended effects. It is sometimes said in discussions about legalising voluntary euthanasia that experience with abortion law reform should remind us of how quickly and easily practices can become accepted which were never among the reformers' intentions, and that the same thing could occur if voluntary euthanasia were to become legally permitted. No amount of theorising, it is said, can gainsay that possibility. There is no need to deny that it is possible that reform of the laws that presently prohibit voluntary euthanasia could have untoward consequences. However, if the arguments given above are sound (and the Dutch experience is not only the best evidence we have that they are sound, but the *only* relevant evidence), that does not seem very likely.

Objection 6

I turn now to the final objection to be considered here. It is often claimed that whatever the morality of an individual's deciding for herself that her life is no longer of value to her, that provides no basis for the formulation of public policy. The fear of the slippery slope is, no doubt, part of the concern expressed here. But, as well, there are concerns about the role of the law and more particularly, its contribution to the regulation of medicine.

Legal permission for doctors to perform voluntary euthanasia cannot simply be grounded in the right of self-determination of patients. We have already had occasion to note that the law does not presently permit an individual to consent to her own death. Nevertheless, the very same fundamental basis of the right to decide about life-sustaining treatment—respect for a person's autonomy—underpins voluntary euthanasia as well. Extending the right of self-determination to cover cases of voluntary euthanasia would not, therefore, amount to a dramatic shift in legal policy. No novel legal values or principles need to be invoked. Indeed, the fact that suicide and attempted suicide are no longer criminal offences in many jurisdictions indicates that the central importance of individual self-determination in a closely analogous

setting has been accepted. The fact that assisted suicide and voluntary euthanasia have not yet been widely decriminalised is probably best explained along the lines that have frequently been offered for excluding consent of the victim as a justification for an act of killing, namely the difficulties thought to exist in establishing the genuineness of the consent. The establishment of suitable procedures for giving consent to assisted suicide and voluntary euthanasia would seem to be no harder than establishing procedures for competently refusing burdensome or otherwise unwanted medical treatment. The latter has already been accomplished in many jurisdictions, so the former should be capable of establishment as well.

Suppose that the moral case for legalising voluntary euthanasia does come to be judged as stronger than the case against (as the drift of this article would imply), and voluntary euthanasia is made legally permissible. Should doctors take part in the practice? Should only doctors perform voluntary euthanasia? The proper administration of medical care is not at odds with an understanding of medical care that both promotes patients' welfare interests and respects their self-determination. It is these twin values which should guide medical care, not a commitment to preserving life at all costs, or preserving life without regard to whether patients want their lives prolonged when they judge that life is no longer of benefit or value to themselves. Many doctors in The Netherlands and, to judge from available survey evidence, in other Western countries as well, see the practice of (voluntary) euthanasia as not only compatible with their professional commitments but also with their conception of the best medical care for the dying. That being so, they should not be prohibited by law from lending their professional assistance to those competent, terminally ill persons for whom no cure is possible and who wish for an easy death.

Bibliography

- D. Brock, 1993, “Voluntary Active Euthanasia”, *Hastings Center Report* 22, no. 2 (1993) pp. 10-22.
- M. Burleigh, 1994, *Death and Deliverance: Euthanasia in Germany c. 1900-1945* (Cambridge: Cambridge University Press).
- *Commission on the Study of Medical Practice Concerning Euthanasia: Medical Decisions Concerning the End of Life* (The Hague: SdU, 1991)—otherwise known as ‘The Remmelink Report’.
- J. Griffiths, A. Bood, and H. Weyers, 1998, *Euthanasia and Law in The Netherlands* (Amsterdam: Amsterdam University Press).
- H. Kuhse, 1987, *The Sanctity-of-Life Doctrine in Medicine: A Critique* (Oxford: Clarendon Press).
- H. Kuhse, P. Singer, P. Baume, A. Clark, and M. Rickard, 1997, “End-of-Life Decisions in Australian Medical Practice”, *The Medical Journal of Australia* 166, pp. 191-196.
- A. McIntyre, 2001, “Doing Away With Double Effect”, *Ethics* 111, pp. 219-255.
- J. Rachels, 1986, *The End of Life: Euthanasia and Morality* (Oxford: Oxford University Press).
- P.J. van der Maas, J.J.M. van Delden, L. Pijnenborg, C.W.N. Looman, 1991, “Euthanasia and other Medical Decisions Concerning the End of Life”, *The Lancet* 338, pp. 669-674.
- P.J. van der Maas, G. van der Wal, I. Haverkate, C.L.M. de Graaf, J.G.C. Kester, B.D. Onwuteaka-Philipsen, A. van der Heide, J.M. Bosma and D.L. Willems, 1996, “Euthanasia, Physician-

Assisted Suicide, and other Medical Practices Involving the End of Life in the Netherlands, 1990-1995”, *The New England Journal of Medicine* 335, pp. 1699-1705.
- G. van der Wal, J.Th.M. van Eijk, H.J.J. Leenen, C. Spreeuwenberg, 1992a, “Euthanasia and Assisted Suicide, I: How Often is it Practised by Family Doctors in the Netherlands?”, *Family Practice* 9, pp. 130-134.
- G. van der Wal, J.Th.M. van Eijk, H.J.J. Leenen, C. Spreeuwenberg, 1992b, “Euthanasia and Assisted Suicide, II: Do Dutch Family Doctors Act Prudently?”, *Family Practice* 9, pp. 135-140.
- G. van der Wal, P.J. van der Maas, J.M. Bosma, B.D. Onwuteaka-Philipsen, D.L. Willems, I. Haverkate and P.J. Kostense, 1996, “Evaluation of the Notification Procedure for Physician-Assisted Death in the Netherlands”, *The New England Journal of Medicine* 335, pp. 1706-1711.
- E. Winkler, 1995, “Reflections on the State of Current Debate Over Physician-Assisted Suicide and Euthanasia”, *Bioethics* 9, pp. 313-326.
- R. Young, 1976, “Voluntary and Nonvoluntary Euthanasia”, *The Monist* 59, pp. 264-283.

Other Internet Resources

- The Case For Voluntary Euthanasia, maintain by Kevin Williams
- The Case Against Voluntary Euthanasia, maintained by Kevin Williams
- Voluntary Euthanasia, authored by Richard Epstein (University of Chicago), site maintained by the Hayek Society, the London School of Economics and Political Science
- Euthanasia and Assisted Suicide: Seven Reasons Why They Should Not Be Legalized, authored by Luke Gormally (Linacre Centre for Healthcare Ethics)
- Oregon's Death with Dignity Act, (Oregon Department of Human Services)
- Dutch Voluntary Euthanasia Society
- Euthanasia and End-of-Life Decisions, (Ethics Updates, L. Hinman, University of San Diego)

Related Entries

ethics: biomedical

Robert Young
La Trobe University
Robert.Young@latrobe.edu.au

Table of Contents

First published: April 18, 1996
Content last modified: May 20, 2002

Aristotle's Political Theory

Aristotle (b. 384 - d. 322 BC), was a Greek philosopher, logician, and scientist. Along with his teacher Plato, Aristotle is generally regarded as one of the most influential ancient thinkers in a number of philosophical fields, including political theory. Aristotle was born in Stagira in northern Greece, and his father was a court physician to the king of Macedon. As a young man he studied in Plato's Academy in Athens. After Plato's death he left Athens to conduct philosophical and biological research in Asia Minor and Lesbos, and he was then invited by King Philip II of Macedon to tutor his young son, Alexander the Great. Soon after Alexander succeeded his father, consolidated the conquest of the Greek city-states, and launched the invasion of the Persian Empire. Aristotle returned as a resident alien to Athens, and was close friend of Antipater the Macedonian viceroy. At this time (335-323 BC) he wrote or at least completed some of his major treatises, including the *Politics*. When Alexander died suddenly, Aristotle had to flee from Athens because of his Macedonian connections, and he died soon after. Aristotle's life seems to have influenced his political thought in various ways: his interest in biology seems to be expressed in the naturalism of his politics; his interest in comparative politics and his sympathies for democracy as well as monarchy may have been encouraged by his travels and experience of diverse political systems; he criticizes harshly, while borrowing extensively, from Plato's *Republic*, *Statesman*, and *Laws*; and his own *Politics* is intended to guide rulers and statesmen, reflecting the high political circles in which he moved.

1. Political Science in General

The modern word 'political' derives from the Greek *politikos*, 'of, or pertaining to, the polis'. (The Greek term *polis* will be translated here as 'city-state'. It is also translated as 'city' or 'polis', or simply anglicized as 'polis'. City-states like Athens and Sparta were relatively small and cohesive units, in which political, religious, and cultural concerns were intertwined. The extent of their similarity to modern nation-states is controversial.) Aristotle's word for 'politics' is *politikê*, which is short for *politikê epistêmê* or 'political science'. It belongs to one of the three main branches of science, which Aristotle distinguishes by their ends or objects. Contemplative science (including physics and metaphysics) is concerned with truth or knowledge for its own sake; practical science with good action; and productive science with making useful or beautiful objects (*Top*. VI.6.145a14-16, *Met*. VI.1.1025b24, XI.7.1064a16-19, *EN* VI.2.1139a26-8). Politics is a practical science, since it is concerned with the noble action or happiness of the citizens (although it resembles a productive science in that it seeks to create, preserve, and reform political systems.) Aristotle thus understands politics as a normative or prescriptive discipline rather than as a purely empirical or descriptive inquiry.

In *Nicomachean Ethics* I.2 Aristotle characterizes politics as the most authoritative science. It prescribes which sciences are to be studied in the city-state, and the other capacities -- such as military science, household management, and rhetoric -- fall under its authority. Since it governs the other practical sciences, their ends serve as means to its end, which is nothing less than the human good. "Even if the end is the same for an individual and for a city-state, that of the city-state seems at any rate greater and more complete to attain and preserve. For although it is worthy to attain it for only an individual, it is nobler and more divine to do so for a nation or city-state." (*EN* I.2.1094b7-10) Aristotle's political science encompasses the two fields which modern philosophers distinguish as ethics and political philosophy. (See the entry on Aristotle's ethics.) Political philosophy in the narrow sense is roughly speaking the subject of his treatise called the *Politics*. For a further discussion of this topic, see the following supplementary document:

> Supplement: Characteristics and Problems of Aristotle's Politics

2. Aristotle's View of Politics

Political science studies the tasks of the politician or statesman (*politikos*), in much the way that medical science concerns the work of the physician (see *Politics* IV.1). It is, in fact, the body of knowledge that such practitioners, if truly expert, will also wield in pursuing their tasks. The most important task for the politician is, in the role of lawgiver (*nomothetês*), to frame the appropriate constitution for the city-state. This involves enduring laws, customs, and institutions (including a system of moral education) for the citizens. Once the constitution is in place, the politician needs to take the appropriate measures to maintain it, to introduce reforms when he finds them necessary, and to prevent developments which might subvert the political system. This is the province of legislative science, which Aristotle regards as more important than politics as exercised in everyday political activity such as the passing of decrees (see *EN*

VI.8).

Aristotle frequently compares the politician to a craftsman. The analogy is imprecise because politics, in the strict sense of legislative science, is a form of practical wisdom or prudence, but valid to the extent that the politician produces, operates, and maintains a legal system according to universal principles (*EN* VI.8 and X.9). In order to appreciate this analogy it is helpful to observe that Aristotle explains production of an artifact in terms of four causes: the material, formal, efficient, and final causes (*Phys*. II.3 and *Met*. A.2). For example, clay (material cause) is molded into a vase shape (formal cause) by a potter (efficient or moving cause) so that it can contain liquid (final cause). (For discussion of the four causes see the entry on Aristotle's physics.)

One can also explain the existence of the city-state in terms of the four causes. It is a kind of community (*koinônia*), that is, a collection of parts having some functions and interests in common (*Pol*. II.1.1261a18, III.1.1275b20). Hence, it is made up of parts, which Aristotle describes in various ways in different contexts: as households, or economic classes (e.g., the rich and the poor), or demes (i.e., local political units). But, ultimately, the city-state is composed of individual citizens (see III.1.1274a38-41), who, along with natural resources, are the "material" or "equipment" out of which the city-state is fashioned (see VII.14.1325b38-41).

The formal cause of the city-state is its constitution (*politeia*). Aristotle defines the constitution as "a certain ordering of the inhabitants of the city-state" (III.1.1274b32-41). He also speaks of the constitution of a community as "the form of the compound" and argues that whether the community is the same over time depends on whether it has the same constitution (III.3.1276b1-11). The constitution is not a written document, but an immanent organizing principle, analogous to the soul of an organism. Hence, the constitution is also "the way of life" of the citizens (IV.11.1295a40-b1, VII.8.1328b1-2). Here the citizens are that minority of the resident population who are adults with full political rights.

The existence of the city-state also requires an efficient cause, namely, its ruler. On Aristotle's view, a community of any sort can possess order only if it has a ruling element or authority. This ruling principle is defined by the constitution, which sets criteria for political offices, particularly the sovereign office (III.6.1278b8-10; cf. IV.1.1289a15-18). However, on a deeper level, there must be an efficient cause to explain why a city-state acquires its constitution in the first place. Aristotle states that "the person who first established [the city-state] is the cause of very great benefits" (I.2.1253a30-1). This person was evidently the lawgiver (*nomothetês*), someone like Solon of Athens or Lycurgus of Sparta, who founded the constitution. Aristotle compares the lawgiver, or the politician more generally, to a craftsman (*dêmiourgos*) like a weaver or shipbuilder, who fashions material into a finished product (II.12.1273b32-3, VII.4.1325b40-1365a5).

The notion of final cause dominates Aristotle's *Politics* from the opening lines:

> Since we see that every city-state is a sort of community and that every community is established for the sake of some good (for everyone does everything for the sake of what

> they believe to be good), it is clear that every community aims at some good, and the community which has the most authority of all and includes all the others aims highest, that is, at the good with the most authority. This is what is called the city-state or political community. [I.1.1252a1-7]

Soon after, he states that the city-state comes into being for the sake of life but exists for the sake of the good life (2.1252b29-30). The theme that the good life or happiness is the proper end of the city-state recurs throughout the *Politics* (III.6.1278b17-24, 9.1280b39; VII.2.1325a7-10).

To sum up, the city-state is a hylomorphic (i.e., matter-form) compound of a particular population (i.e., citizen-body) in a given territory (material cause) and a constitution (formal cause). The constitution itself is fashioned by the lawgiver and is governed by politicians, who are like craftsmen (efficient cause), and the constitution defines the aim of the city-state (final cause, IV.1.1289a17-18). For a further discussion of this topic, see the following supplementary document:

> Supplement: Presuppositions of Aristotle's Politics

It is in these terms that Aristotle understands the fundamental normative problem of politics: What constitutional form should the lawgiver and politician establish and preserve in what material for the sake of what end?

3. General Theory of Constitutions and Citizenship

Aristotle states that "the politician and lawgiver is wholly occupied with the city-state, and the constitution is a certain way of organizing those who inhabit the city-state" (III.1.1274b36-8). His general theory of constitutions is set forth in *Politics* III. He begins with a definition of the citizen (*politês*), since the city-state is by nature a collective entity, a multitude of citizens. Citizens are distinguished from other inhabitants, such as resident aliens and slaves; and even children and seniors are not unqualified citizens (nor are most ordinary workers). After further analysis he defines the citizen as a person who has the right (*exousia*) to participate in deliberative or judicial office (1275b18-21). In Athens, for example, citizens had the right to attend the assembly, the council, and other bodies, or to sit on juries. The Athenian system differed from a modern representative democracy in that the citizens were more directly involved in governing. Although full citizenship tended to be restricted in the Greek city-states (with women, slaves, foreigners, and some others excluded), the citizens were more deeply enfranchised than in modern representative democracies because they were more directly involved in governing. This is reflected in Aristotle's definition of the citizen (without qualification). Further, he defines the city-state (in the unqualified sense) as a multitude of such citizens which is adequate for a self-sufficient life (1275b20-21).

Aristotle defines the constitution as a way of organizing the offices of the city-state, particularly the sovereign office (III.6.1278b8-10; cf. IV.1.1289a15-18). The constitution thus defines the governing body, which takes different forms: for example, in a democracy it is the people, and in an oligarchy it is a select few (the wealthy or well born). Before attempting to distinguish and evaluate various constitutions

Aristotle considers two questions. First, why does a city-state come into being? He recalls the thesis, defended in *Politics* I.2, that human beings are by nature political animals, who naturally want to live together. For a further discussion of this topic, see the following supplementary document:

> Supplement: Political Naturalism

He then adds that "the common advantage also brings them together insofar as they each attain the noble life. This is above all the end for all both in common and separately." (III.6.1278b19-24) Second, what are the different forms of rule by which one individual or group can rule over another? Aristotle distinguishes several types. He first considers despotic rule, which is exemplified in the master-slave relationship. Aristotle thinks that this form of rule is justified in the case of natural slaves who (he asserts without evidence) lack a deliberative faculty and thus need a natural master to direct them (I.13.1260a12; slavery is defended at length in *Politics* I.4-8). Although a natural slave allegedly benefits from having a master, despotic rule is still primarily for the sake of the master and only incidentally for the slave (III.6.1278b32-7). (Aristotle provides no argument for this: if some persons are congenitally incapable of self-governance, why should they not be ruled primarily for their own sakes?) He next considers paternal and marital rule, which he also views as defensible: "the male is by nature more capable of leadership than the female, unless he is constituted in some way contrary to nature, and the elder and perfect [is by nature more capable of leadership] than the younger and imperfect." (I.12.1259a39-b4) Aristotle is persuasive when he argues that children need adult supervision because their rationality is "imperfect" (*ateles*) or immature. But he also alleges (without substantiation) that, although women have a deliberative faculty, it is "without authority" (*akuron*), so that females require male leadership (I.13.1260a13-14). (Aristotle's arguments about slaves and women appear so weak that some commentators take them to be ironic. However, what is obvious to a modern reader need not have been so to an ancient Greek, so that it is not necessary to suppose that Aristotle's discussion is ironic.) It is noteworthy, however, that paternal and marital rule are properly practiced for the sake of the ruled (for the sake of the child and of the wife respectively), just as arts like medicine or gymnastics are practiced for the sake of the patient (III.6.1278b37-1279a1). In this respect they resemble political rule, which involves equal and similar citizens taking turns in ruling for one another's advantage (1279a8-13). This sets the stage for the fundamental claim of Aristotle's constitutional theory: "constitutions which aim at the common advantage are correct and just without qualification, whereas those which aim only at the advantage of the rulers are deviant and unjust, because they involve despotic rule which is inappropriate for a community of free persons" (1279a17-21).

The distinction between correct and deviant constitutions is combined with the observation that the government may consist of one person, a few, or a multitude. Hence, there are six possible constitutional forms (*Politics* I.7):

	Correct	**Deviant**
One Ruler	Kingship	Tyranny

Few Rulers	Aristocracy	Oligarchy
Many Rulers	Polity	Democracy

This six-fold classification (which is adapted from Plato's *Statesman*) sets the stage for Aristotle's inquiry into the best constitution, although it is modified in various ways throughout the *Politics*. For example, he observes that the dominant class in oligarchy (literally rule of the *oligoi*, i.e., few) is typically the wealthy, whereas in democracy (literally rule of the *dêmos*, i.e., people) it is the poor, so that these economic classes should be included in the definition of these forms (see *Politics* III.8, IV.4, and VI.2 for alternative accounts). Also, polity is later characterized as a kind of "mixed" constitution typified by rule of the "middle" group of citizens, a moderately wealthy class between the rich and poor (*Politics* IV.11).

Aristotle turns to arguments for and against the different constitutions, which he views as different applications of the principle of distributive justice (III.9.1280a7-22). Everyone agrees, he says, that justice involves treating equal persons equally, and treating unequal persons unequally, but they do not agree on the standard by which individuals are deemed to be equally (or unequally) meritorious or deserving. He assumes his own analysis of distributive justice set forth in *Nicomachean Ethics* V.3: Justice requires that benefits be distributed to individuals in proportion to their merit or desert. The oligarchs mistakenly think that those who are superior in wealth should also have superior political rights, whereas the democrats hold that those who are equal in free birth should also have equal political rights. Both of these conceptions of political justice are mistaken in Aristotle's view, because they assume a false conception of the ultimate end of the city-state. The city-state is neither a business association to maximize wealth (as the oligarchs suppose) nor an agency to promote liberty and equality (as the democrats maintain). Instead, Aristotle argues, "the good life is the end of the city-state," that is, a life consisting of noble actions (1280b39-1281a4). Hence, the correct conception of justice is aristocratic, assigning political rights to those who make a full contribution to the political community, that is, to those with virtue as well as property and freedom (1281a4-8). This is what Aristotle understands by an "aristocratic" constitution: literally, the rule of the *aristoi*, i.e., best persons. Aristotle explores the implications of this argument in the remainder of *Politics* III, considering the rival claims of the rule of law and the rule of a supremely virtuous individual. Here absolute kingship is a limiting case of aristocracy. Again, in books VII-VIII, Aristotle describes the ideal constitution in which the citizens are fully virtuous.

4. Study of Specific Constitutions

The purpose of political science is to guide "the good lawgiver and the true politician" (IV.1.1288b27). Like any complete science or craft, it must study a range of issues concerning its subject matter. For example, gymnastics (physical training) studies what sort of training is advantageous for what sort of body, what sort of training is best or adapted to the body that is naturally the best, what sort of training is best for most bodies, and what capacity is appropriate for someone who does not want the condition or knowledge appropriate for athletic contests. Political science studies a comparable range of constitutions

(1288b21-35): first, the constitution which is best without qualification, i.e., "most according to our prayers with no external impediment"; second, the constitution that is best under the circumstances "for it is probably impossible for many persons to attain the best constitution"; third, the constitution which serves the aim a given city-state population happens to have, i.e., the one that is best "based on a hypothesis": "for [the political scientist] ought to be able to study a given constitution, both how it might originally come to be, and, when it has come to be, in what manner it might be preserved for the longest time; I mean, for example, if a particular city happens neither to be governed by the best constitution, nor to be equipped even with necessary things, nor to be the [best] possible under existing circumstances, but to be a baser sort."

Hence, Aristotelian political science is not confined to the ideal system, but also investigates the second-best constitution, the one which is the best that most city-states are capable of supporting. For it is the closest approximation to full political justice which the lawgiver can attain under the circumstances. Although Aristotle’s political views were influenced by his teacher Plato, he is very critical of the ideal city-state set forth in Plato’s *Republic* on the grounds that it overvalues political unity, it embraces a system of communism that is impractical and inimical to human nature, and it neglects the happiness of the individual citizens (*Politics* II.1-5). In contrast, in Aristotle’s own "best constitution" (described in *Politics* VII-VIII) each and every citizen will possess moral virtue and the equipment to carry it out in practice, and thereby attain a life of excellence and complete happiness (see VII.13.1332a32-8). All of the citizens will hold political office and possess private property because "one should call the city-state happy not by looking at a part of it but at all the citizens." (VII.9.1329a22-3). Moreover, there will be a common system of education for all the citizens, because they share the same end (*Pol*. VIII.1). But if (as is the case with most city-states) the population lacks the capacities and resources for complete happiness, the lawgiver must be content with fashioning a suitable constitution (*Politics* IV.11). The second-best system typically takes the form of a polity (in which citizens possess an inferior, more common grade of virtue) or mixed constitution (combining features of democracy, oligarchy, and aristocracy, so that no group of citizens is in a position to abuse its rights).

In addition, the political scientist must understand existing constitutions even when they are bad. Aristotle adds that "to reform a constitution is no less a task [of politics] than it is to establish one from the beginning," and in this way "the politician should also help existing constitutions." (IV.1.1289a1-7) The political scientist should also be cognizant of forces of political change which can undermine an existing regime. Aristotle criticizes his predecessors for excessive utopianism and neglect of the practical duties of a political theorist. However, he is no Machiavellian. The best constitution still serves as a regulative ideal by which to evaluate existing systems.

These topics occupy the remainder of the *Politics*. Books IV-VI are concerned with the existing constitutions: that is, the three deviant constitutions, as well as polity or the mixed constitution, the best attainable under most circumstances (IV.2.1289a26-38). The whole of book V investigates political change and revolution. Books VII-VIII are devoted to the ideal constitution. As might be expected, Aristotle’s attempt to carry out this program involves many difficulties, and scholars disagree about how the two series of books (IV-VI and VII-VIII) are related to each other: for example, which were written first, which were intended to be read first, and whether they are ultimately consistent with each other. For

a further discussion of this topic, see the following supplementary document:

> [Supplement: Characteristics and Problems of Aristotle's Politics]

Aristotle's *Politics* did not have an immediate impact because it defended the Greek city-state, which was already becoming obsolete in his own lifetime. (As mentioned above, the Greek city-states permanently lost their independence due to the conquest by the kings of Macedon.) For similar reasons much of his discussion of particular political institutions is not directly applicable to modern nation-states (apart from his objectionable defenses of slavery, female subservience, and disenfranchisement of the working classes). Even so, Aristotle's *Politics* has had a deep influence on political philosophy until the present day, because it contains deep and thought-provoking discussions of perennial concerns of political philosophy: the role of human nature in politics, the relation of the individual to the state, the place of morality in politics, the theory of political justice, the rule of law, the analysis and evaluation of constitutions, the relevance of ideals to practical politics, the causes and cures of political change and revolution, and the importance of a morally educated citizenry.

Glossary of Aristotelian Terms

- action: *praxis*
- citizen: *politês*
- city-state: *polis*
- community: *koinônia*)
- constitution: *politeia*
- excellence: *aretê* (also 'virtue')
- free: *eleutheros*
- good: *agathos*
- happiness: eudaimonia
- happy: *eudaimôn*
- justice: *dikaiosunê*
- law: *nomos*
- lawgiver: *nomothetês*
- master: *despotês*
- nature: *phusis*
- noble: *kalon* (also 'beautiful')
- political: *politikos* (of, or pertaining to, the *polis*)
- political science: *politikê epistêmê*
- practical: *praktikos*
- practical wisdom: *phronêsis*
- right: *exousia*
- ruler: *archôn*
- self-sufficient: *autarkês*
- sovereign: *kurios*

- without qualification: *haplôs* (also 'absolute')
- without authority: *akuron*

Bibliography

Translations

- Ernest Barker, rev. by Richard Stalley (Oxford, 1995).
- Benjamin Jowett, rev. Jonathan Barnes (in *The Complete Works of Aristotle*, vol. 2, Princeton, 1984).
- Carnes Lord (Chicago, 1984).
- C. D. C. Reeve (Indianapolis, 1998).
- Peter L. P. Simpson (Chapel Hill, 1996).
- T. A. Sinclair, rev. Trevor J. Saunders (Harmondsworth, 1983).
- The Clarendon Aristotle Series (Oxford University Press) will include translation and commentary of the *Politics* in four volumes:
    - Trevor J. Saunders, *Politics* I-II (1995).
    - Richard Robinson with a supplementary essay by David Keyt, *Politics* III-IV (1995).
    - David Keyt, *Politics* V-VI (1999).
    - Richard Kraut, *Politics* VII-VIII (1997).

Scholarly literature

- Jonathan Barnes et al., eds., *Articles on Aristotle*, vol. 2, Ethics and Politics (London, 1977).
- Richard Bodéüs, *The Political Dimensions of Aristotle's Ethics* (Albany, 1993).
- Otfried Höffe, ed., *Aristoteles Politik* (Berlin, 2001).
- David Keyt and Fred D. Miller, Jr., eds., *A Companion to Aristotle's Politics* (Oxford, 1991).
- Richard Kraut, *Aristotle: Political Philosophy* (Oxford, 2002).
- Carnes Lord and David O'Connor, eds., *Essays on the Foundations of Aristotelian Political Science* (Berkeley, 1991).
- Fred D. Miller, Jr., *Nature, Justice, and Rights in Aristotle's Politics* (Oxford, 1995).
- Richard G. *Mulgan, Aristotle's Political Theory* (Oxford, 1977).
- W. L. Newman, *The Politics of Aristotle*, 4 vols. (Oxford, 1887-1902).
- Mary Nichols, *Citizens and Statesmen: A Study of Aristotle's Politics* (Savage, Md., 1992).
- Günther Patzig, ed., *Aristoteles' Politik* (Göttingen, 1990).
- Stephen G. Salkever, *Finding the Mean: Theory and Practice in Aristotelian Political Philosophy* (Princeton, 1990).
- Peter Simpson, *A Philosophical Commentary on the Politics of Aristotle* (Chapel Hill, 1998).
- Judith A. Swanson, *The Public and the Private in Aristotle's Political Philosophy* (Ithaca, 1991).
- Bernard Yack, *The Problems of a Political Animal: Community, Justice, and Conflict in Aristotelian Political Thought* (Berkeley, 1993).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle: biology | Aristotle: ethics | Aristotle: physics

Fred D. Miller, Jr.
fmiller@bgnet.bgsu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 1, 1998
Content last modified: July 19, 2002

Stanford Encyclopedia of Philosophy
Supplement to Aristotle's Politics

Characteristics and Problems of Aristotle's Politics

The work which has come down to us under this name appears to be less an integrated treatise than a collection of essays on various topics in political philosophy, which may have been compiled by a later editor rather than by Aristotle. The following topics are discussed in the eight books:

> I Naturalness of the city-state and of the household
>
> II Critique of ostensibly best constitutions
>
> III General theory of constitutions
>
> IV Inferior constitutions
>
> V Preservation and destruction of constitutions
>
> VI Further discussion of democracy and oligarchy
>
> VII-VIII Blueprint of the best constitution

This ordering of the books reflects, very roughly, the program for the study of constitutions which concludes the Nicomachean Ethics:

First, then, if any particular point has been treated well by those who have gone before us, we must try to review it; then from the constitutions that have been collected we must try to see what it is that preserves and destroys each of the constitutions, and for what reasons some city-states are well governed and others the reverse. For when these things have been examined, we will perhaps better understand what sort of constitution is best, and how each is structured, and which laws and customs it uses. Let us then begin our discussion. [X.9.1181b15-23]

However, scholars have raised problems with the Politics as we have it. The first concerns the intended order of its eight books. Some (including W. L. Newman) have questioned the traditional ordering, arguing that the discussion of the best constitution (books VII-VIII) should follow directly after book III. Indeed, book III concludes with a transition to a discussion of the best constitution (although this may be due to a later editor). However, cross-references between various passages of the Politics indicate that books IV-V-VI form a connected series, as do books VII-VIII, but these series do not refer to each other.

However, both series refer back to book III which in turn refers to book I. With some oversimplification, the Politics is comparable to a tree trunk supporting two separate branches: the root system is I, the trunk is II-III, and the branches are IV-V-VI and VII-VIII. (The summary in Nicomachean Ethics X.9 describes only the visible part of the tree.)

The second problem concerns the order in which the books were actually written. If they were composed at very different dates, they might represent discordant stages in the development of Aristotle's political philosophy. For example, Werner Jaeger argued that books VII-VIII contain a youthful utopianism, motivating Aristotle to emulate his teacher Plato in erecting "an ideal state by logical construction." In contrast, books IV-VI are based on "sober empirical study." Other scholars have seen a more pragmatic, even Machiavellian approach to politics in books IV-VI. A difficulty for this interpretation is that in book IV Aristotle regards the business of constructing ideal constitutions as perfectly compatible with that of addressing actual political problems. Although much ink has been spilled since Jaeger attempted to discern different chronological strata in the Politics, it has not resulted in a clear scholarly consensus. Because there is no explicit evidence of the dates at which the various books of the Politics were written, argument has turned on alleged inconsistencies between different passages.

This leads to the third problem, whether there are major inconsistencies of doctrine or method in the Politics. For example, Aristotle's account of the best constitution assumes his theory of justice, a moral standard which cannot be met by the actual political systems (democracies and oligarchies) of his own day. He does discuss practical political reforms in books IV-VI but more in terms of stability than justice. This raises the question of whether books IV-VI mark a radical departure from the political philosophy of the other books. Resolution of this problem requires careful study of the Politics as a whole.

Fred D. Miller, Jr.
fmiller@bgnet.bgsu.edu

Return to Aristotle's Politics [Section 1]
Return to Aristotle's Politics [Section 4]

First published: July 1, 1998
Content last modified: July 1, 1998

Stanford Encyclopedia of Philosophy
Supplement to Aristotle's Politics

Presuppositions of Aristotle's Politics

Aristotle's political philosophy is distinguished by its underlying philosophical doctrines. Of these the following four principles are especially noteworthy:

(1) The principle of teleology Aristotle begins the *Politics* by invoking the concept of nature (see Political Naturalism). In the Physics Aristotle identifies the nature of a thing above all with its end or final cause (Phys. II.2.194a28-9, 8.199b15-18). The end of a thing is also its function (*EE* II.1.1219a8), which is its defining principle (*Meteor*. IV.12.390a10-11). On Aristotle's view plants and animals are cardinal examples of natural existents, because they have a nature in the sense of an internal causal principle which explains how it comes into being and behaves (*Phys*. II.1.192b32-3). For example, an acorn has an inherent tendency to grow into an oak tree, so that the tree exists by nature rather than by craft or by chance. The thesis that human beings have a natural function has a fundamental place in th*e Eudemian Ethics* II.1, *Nicomachean Ethics* I.7, and *Politics* I.2. The *Politics* further argues that it is part of the nature of human beings that they are political or adapted for life in the city-state. Thus teleology is crucial for the political naturalism which is at the foundation of Aristotle's political philosophy. (For discussion of teleology see the entry on Aristotle, biology.)

(2) The principle of perfection Aristotle understands good and evil in terms of his teleology. The natural end of the organism (and the means to this end) is good for it, and what defeats or impedes this end is bad. For example, he argues that animals sleep in order to preserve themselves, because "nature operates for the sake of an end, and this is a good," and sleeping is necessary and beneficial for entities which cannot move continuously (*De Somno* 2.455b17-22). For human beings the ultimate good or happiness (*eudaimonia*) consists in perfection, the full attainment of their natural function, which Aristotle analyzes as the activity of the soul according to reason (or not without reason), i.e., activity in accordance with the most perfect virtue or excellence (*EN* I.7.1098a7-17). This also provides a norm for the politician: "What is most choiceworthy for each individual is always the highest it is possible for him to attain" (*Pol*. VII.14.1333a29-30; cf. *EN* X.7.1177b33-4). This ideal is to be realized in both the individual and the city-state: "that way of life is best, both separately for each individual and in common for city-states, which is equipped with virtue" (*Pol*. VII.1.1323b40-1324a1). However, Aristotle recognizes that it is generally impossible to fully realize this ideal, in which case he invokes a fall-back principle: it is best to attain perfection, but, failing that, a thing is better in proportion as it is nearer to the end (see *DC* II.12.292b17-19).

Aristotle's perfectionism was opposed to the subjective relativism of Protagoras, according to which good and evil is defined by whatever human beings happened to desire. Like Plato, Aristotle maintained that the good was objective and independent of human wishes. However, he rejected Plato's own theory that

the good was defined in terms of a transcendent form of the good, holding instead that good and evil are in a way relative to the organism, that is, to its natural end.

(3) The principle of community Aristotle maintains that the city-state is the most complete community, because it attains the limit of self-sufficiency, so that it can exist for the sake of the good life (*Pol*. I.2.1252b27-30). Individuals outside of the city-state are not self-sufficient, because they depend on the community not only for material necessities but also for education and moral habituation. "Just as, when perfected, a human is the best of animals, so also when separated from law and justice, he is the worst of all" (1253a31-3). On Aristotle's view, then, human beings must be subject to the authority of the city-state in order to attain the good life. The following principle concerns how authority should be exercised within a community.

(4) Principle of rulership Aristotle believes that the existence and well-being of any system requires the presence of a ruling element: "Whenever a thing is established out of a number of things and becomes a single common thing, there always appears in it a ruler and ruled. . . . This [relation] is present in living things, but it derives from all of nature." (1254a28-32)

Just as an animal or plant can survive and flourish only if its soul rules over its body (Pol. I.5.1254a34-6, DA I.5.410b10-15; compare Plato Phaedo 79e-80a), a human community can possess the necessary order only if it has a ruling element which is in a position of authority, just as an army can possess order only if it has a commander in control. Although Aristotle followed Plato on this principle, he rejected Plato's further claim that one form of rule is appropriate for all. For Aristotle different forms of rule are necessary for different systems: e.g., political rule for citizens and despotic rule for slaves. The imposition of an inappropriate type of rule results in disorder and injustice.

The aforementioned principles account for much of the distinctive flavor of Aristotle's political philosophy, and they also indicate where many modern theorists have turned away from him. Modern philosophers such as Thomas Hobbes have challenged the principles of teleology and perfectionism, arguing against the former that human beings are mechanistic rather than teleological systems, and against the latter that good and bad depend upon subjective preferences of valuing agents rather than on objective states of affairs. Liberal theorists have criticized the principle of community on the grounds that it cedes too much authority to the state. Even the principle of rulership which Aristotle, Plato, and many other theorists thought self-evident has come under fire by modern theorists like Adam Smith and F. A. Hayek who argued that social and economic order may arise spontaneously as if by an "invisible hand." Modern neo-Aristotelian political theorists are committed to defending one or more of these doctrines against such criticisms.

Fred D. Miller, Jr.
fmiller@bgnet.bgsu.edu

Return to Aristotle's Politics

First published: July 1, 1998
Content last modified: July 1, 1998

Stanford Encyclopedia of Philosophy
Supplement to Aristotle's Politics

Political Naturalism

Aristotle lays the foundations for his political theory in Politics book I by arguing that the city-state and political rule are "natural." The argument begins with a schematic, quasi-historical account of the development of the city-state out of simpler communities. First, individual human beings combined in pairs because they could not exist apart. The male and female joined in order to reproduce, and the master and slave came together for self-preservation. The natural master uses his intellect to rule, and the natural slave uses his body to labor. Second, the household arose naturally from these primitive communities in order to serve everyday needs. Third, when several households combined for other needs a village emerged also according to nature. Finally, "the complete community, formed from several villages, is a city-state, which at once attains the limit of self-sufficiency, roughly speaking. It comes to be for the sake of life, and exists for the sake of the good life." (I.2.1252b27-30)

Aristotle defends three claims about nature and the city-state: First, the city-state exists by nature, because it comes to be out of the more primitive natural associations and it serves as their end, because only it attains self-sufficiency (1252b30-1253a1). Second, human beings are by nature political animals, because nature, which does nothing in vain, has equipped them with speech, which enables them to communicate moral concepts such as justice which are formative of the household and city-state (1253a1-18). Third, the city-state is naturally prior to the individuals, because individuals cannot perform their natural functions apart from the city-state, since they are not self-sufficient (1253a18-29). However, these three claims are immediately followed by a fourth: the city-state is a creation of human intelligence. "Therefore, everyone naturally has the impulse for such a [political] community, but the person who first established [it] is the cause of very great benefits." This great benefactor is evidently the lawgiver, for the legal system of the city-state makes human beings just and virtuous and lifts them from the savagery in which they would otherwise languish (1253a29-39).

Aristotle's political naturalism presents the difficulty that he does not explain how he is using the term "nature" (phusis). In the Physics nature is understood as an internal principle of motion or rest (see III.1.192b8-15). (For discussion of nature see Aristotle's Physics.) If the city-state were natural in this sense, it would resemble a plant or an animal which grows naturally to maturity out of a seed. However, this cannot be reconciled with the important role which Aristotle also assigns to the lawgiver as the one who established the city-state. For on Aristotle's theory a thing either exists by nature or by craft; it cannot do both. (This difficulty is posed by David Keyt.) Aristotle can seemingly escape this dilemma only if it is supposed that he speaks of the city-state as "natural" in another sense of the term. For example, he might mean that it is "natural" in the extended sense that it arises from human natural inclinations (to live in communities) for the sake of human natural ends, but that it remains unfinished until a lawgiver provides it with a constitution. (This solution was proposed by Ernest Barker and is

defended more recently by Fred Miller and Trevor Saunders.)

Fred D. Miller, Jr.
fmiller@bgnet.bgsu.edu

Return to Aristotle's Politics

First published: July 1, 1998
Content last modified: July 1, 1998

Stanford Encyclopedia of Philosophy
Unabridged Table of Contents

(Projected and Assigned Entries)

Search Encyclopedia	Editorial Information
What's New	How to Cite This Encyclopedia
Encyclopedia Archives	Abridged Table of Contents

Navigation Panel:

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

A

- abduction
- **Abelard [Abailard], Peter** (Peter King)
- absolute, the
- **abstract objects** (Gideon Rosen)
- **Academy, Plato's** (Wolfgang Mann)
- accidental properties -- see essential vs. accidental properties
- Achillini, Alessandro
- **action** (George Wilson)
- **action at a distance** (Joseph Berkovitz)
- **actualism** (Christopher Menzel)
- **adaptation** (Robert Brandon)
- **Adorno, Theodore** (Lambert Zuidervaart)
- Aegidius Romanus -- see Giles of Rome
- Aenesidemus -- see skepticism: ancient
- aesthetics
    - in African Philosophy -- see African Philosophy: aesthetics
    - in the 18th century
    - **and objectivity** (Nick Zangwill)
- Aetius -- see Doxography of Ancient Philosophy

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

B

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

C

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

D

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

E

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

F

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

G

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

H

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

I

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

J

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

K

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

L

M

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

N

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

O

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

P

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Q

R

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

S

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

T

U

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

V

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

W

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

X

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Y

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Z

Editorial Information

Stanford Encyclopedia of Philosophy What's New

- Updates for the Last Three Months Listed in Reverse Chronological Order
- Entries by First Publication Date Listed in Reverse Chronological Order

Updates for the Last Three Months Listed in Reverse Chronological Order

- **Homosexuality** (Brent Pickett) [NEW: *August 6, 2002*]
- **The Identity Theory of Truth** (Stewart Candlish) [REVISED: *August 6, 2002*]
  Changes to: Main text, Bibliography
- **Pyrrho** (Richard Bett) [NEW: *August 5, 2002*]
- **Francis Herbert Bradley** (Stewart Candlish) [REVISED: *August 2, 2002*]
  Changes to: Main text, Bibliography
- **Events** (Roberto Casati and Achille Varzi) [REVISED: *August 2, 2002*]
  Changes to: Main text
- **Sorites Paradox** (Dominic Hyde) [REVISED: *July 31, 2002*]
  Changes to: Internet resources
- **Max Stirner** (David Leopold) [REVISED: *July 31, 2002*]
  Changes to: Bibliography
  First published: *June 27, 2002*
- **Conventionality of Simultaneity** (Allen Janis) [REVISED: *July 23, 2002*]
  Changes to: Main text, Bibliography
- **Aristotle's Political Theory** (Fred Miller) [REVISED: *July 19, 2002*]
  Changes to: Main text, Bibliography
- **Identity Politics** (Cressida Heyes) [REVISED: *July 18, 2002*]
  Changes to: Bibliography
  First published: *July 15, 2002*
- **Leibniz's Philosophy of Mind** (Mark Kulstad and Laurence Carlin) [REVISED: *July 15, 2002*]
  Changes to: Main text, Bibliography, Internet resources
- **Naturalism in Legal Philosophy** (Brian Leiter) [NEW: *July 15, 2002*]
- **Virtue Epistemology** (John Greco) [REVISED: *July 10, 2002*]
  Changes to: Bibliography
- **Set Theory** (Thomas Jech) [NEW: *July 10, 2002*]
- **Color** (Barry Maund) [REVISED: *July 9, 2002*]
  Changes to: Main text, Bibliography, Internet resources
- **Realism** (Alexander Miller) [NEW: *July 8, 2002*]

- **The Mathematics of Boolean Algebra** (J. Donald Monk) [NEW: *July 5, 2002*]
- **Animal Consciousness** (Colin Allen) [REVISED: *July 3, 2002*]
  Changes to: Main text, Bibliography
- **Species** (Marc Ereshefsky) [NEW: *July 3, 2002*]
- **Confucius** (Jeffrey Riegel) [NEW: *July 3, 2002*]
- **Salomon Maimon** (Peter Thielke and Yitzhak Melamed) [REVISED: *July 2, 2002*]
  Changes to: Main text
- **Mally's Deontic Logic** (Gert-Jan Lokhorst) [REVISED: *July 2, 2002*]
  Changes to: mally.pl.txt
  Changes prior to June 21, 2002 (notes.html) available in Summer 2002 Edition.
- **Theological Voluntarism** (Mark Murphy) [NEW: *July 1, 2002*]
- **Immutability** (Brian Leftow) [NEW: *July 1, 2002*]
- **Russell's Paradox** (A. D. Irvine) [REVISED: *June 29, 2002*]
  Changes to: Main text, Bibliography, notes.html
- **Distributive Justice** (Julian Lamont) [REVISED: *June 29, 2002*]
  Changes to: Main text, Bibliography
- **Johann Georg Hamann** (Gwen Griffith-Dickson) [NEW: *June 29, 2002*]
- **Cosmopolitanism** (Pauline Kleingeld and Eric Brown) [REVISED: *June 28, 2002*]
  Changes to: Bibliography
- **John Buridan** (Jack Zupko) [REVISED: *June 27, 2002*]
  Changes to: notes.html
  First published: *May 13, 2002*
- **Consciousness and Intentionality** (Charles Siewert) [NEW: *June 22, 2002*]
- **Globalization** (William Scheuerman) [NEW: *June 21, 2002*]
- **Diagrams** (Sun-Joo Shin and Oliver Lemon) [REVISED: *June 19, 2002*]
  Changes are prior to June 21, 2002 (Bibliography) and are available in Summer 2002 Edition.
- **Moral Skepticism** (Walter Sinnott-Armstrong) [NEW: *June 14, 2002*]
- **Scientific Realism** (Richard Boyd) [NEW: *June 12, 2002*]
- **Peter John Olivi** (Robert Pasnau) [REVISED: *June 12, 2002*]
  Changes are prior to June 21, 2002 (Bibliography, Internet resources) and are available in Summer 2002 Edition.
- **Common Knowledge** (Peter Vanderschraaf) [REVISED: *June 12, 2002*]
  Changes are prior to June 21, 2002 (Main text, Bibliography, notes.html) and are available in Summer 2002 Edition.
- **Harriet Taylor Mill** (Dale E. Miller) [REVISED: *June 11, 2002*]
  Changes are prior to June 21, 2002 (Main text, notes.html) and are available in Summer 2002 Edition.
- **Environmental Ethics** (Andrew Brennan and Yeuk-Sze Lo) [NEW: *June 3, 2002*]
- **Alan Turing** (Andrew Hodges) [NEW: *June 3, 2002*]
- **Personal Autonomy** (Sarah Buss) [NEW: *May 28, 2002*]
- **The Biological Notion of Self and Non-self** (Alfred Tauber) [REVISED: *May 27, 2002*]
  Changes are prior to June 21, 2002 (Main text, Bibliography) and are available in Summer 2002 Edition.
  First published: *May 21, 2002*

- **Nicolas Malebranche** (Tad Schmaltz) [NEW: *May 24, 2002*]
- **Existence** (Barry Miller) [REVISED: *May 23, 2002*]
  Changes are prior to June 21, 2002 (Main text, Bibliography) and are available in Summer 2002 Edition.
- **Death** (Steven Luper) [NEW: *May 21, 2002*]
- **Omnipotence** (Joshua Hoffman and Gary Rosenkrantz) [NEW: *May 21, 2002*]
- **Voluntary Euthanasia** (Robert Young) [REVISED: *May 20, 2002*]
  Changes are prior to June 21, 2002 (Main text, Bibliography, Internet resources) and are available in Summer 2002 Edition.
- **Georg Wilhelm Friedrich Hegel** (Paul Redding) [REVISED: *May 20, 2002*]
  Changes are prior to June 21, 2002 (Main text, Bibliography, Internet resources) and are available in Summer 2002 Edition.
- **Cosmology: Methodological Debates in the 1930s and 1940s** (George Gale) [NEW: *May 17, 2002*]
- **Substructural Logics** (Greg Restall) [REVISED: *May 16, 2002*]
  Changes are prior to June 21, 2002 (Bibliography, Internet resources) and are available in Summer 2002 Edition.
- **Paul Feyerabend** (John Preston) [REVISED: *May 15, 2002*]
  Changes are prior to June 21, 2002 (Bibliography) and are available in Summer 2002 Edition.
- **The Identity of Indiscernibles** (Peter Forrest) [REVISED: *May 15, 2002*]
  Changes are prior to June 21, 2002 (Main text, Bibliography) and are available in Summer 2002 Edition.
- **Doing vs. Allowing Harm** (Frances Howard-Snyder) [NEW: *May 14, 2002*]
- **Desert** (Owen McLeod) [NEW: *May 14, 2002*]
- **Privacy** (Judith DeCew) [NEW: *May 14, 2002*]
- **Modal Fictionalism** (Daniel Nolan) [NEW: *May 14, 2002*]
- **Singular Propositions** (Greg Fitch) [REVISED: *May 13, 2002*]
  Changes are prior to June 21, 2002 (Bibliography) and are available in Summer 2002 Edition.
- **Philosophy and Christian Theology** (Michael Murray) [NEW: *May 13, 2002*]
- **The Correspondence Theory of Truth** (Marian David) [NEW: *May 10, 2002*]
- **Frege's Logic, Theorem, and Foundations for Arithmetic** (Edward N. Zalta) [REVISED: *May 9, 2002*]
  Changes are prior to June 21, 2002 (Bibliography) and are available in Summer 2002 Edition.

Entries by First Publication Date Listed in Reverse Chronological Order

- **Homosexuality** (Brent Pickett) [*August 6, 2002*]
- **Pyrrho** (Richard Bett) [*August 5, 2002*]
- **Identity Politics** (Cressida Heyes) [*July 15, 2002*]
- **Naturalism in Legal Philosophy** (Brian Leiter) [*July 15, 2002*]
- **Set Theory** (Thomas Jech) [*July 10, 2002*]
- **Realism** (Alexander Miller) [*July 8, 2002*]

- **The Mathematics of Boolean Algebra** (J. Donald Monk) [*July 5, 2002*]
- **Species** (Marc Ereshefsky) [*July 3, 2002*]
- **Confucius** (Jeffrey Riegel) [*July 3, 2002*]
- **Theological Voluntarism** (Mark Murphy) [*July 1, 2002*]
- **Immutability** (Brian Leftow) [*July 1, 2002*]
- **Johann Georg Hamann** (Gwen Griffith-Dickson) [*June 29, 2002*]
- **Max Stirner** (David Leopold) [*June 27, 2002*]
- **Consciousness and Intentionality** (Charles Siewert) [*June 22, 2002*]
- **Globalization** (William Scheuerman) [*June 21, 2002*]
- **Moral Skepticism** (Walter Sinnott-Armstrong) [*June 14, 2002*]
- **Scientific Realism** (Richard Boyd) [*June 12, 2002*]
- **Environmental Ethics** (Andrew Brennan and Yeuk-Sze Lo) [*June 3, 2002*]
- **Alan Turing** (Andrew Hodges) [*June 3, 2002*]
- **Personal Autonomy** (Sarah Buss) [*May 28, 2002*]
- **Nicolas Malebranche** (Tad Schmaltz) [*May 24, 2002*]
- **Death** (Steven Luper) [*May 21, 2002*]
- **The Biological Notion of Self and Non-self** (Alfred Tauber) [*May 21, 2002*]
- **Omnipotence** (Joshua Hoffman and Gary Rosenkrantz) [*May 21, 2002*]
- **Cosmology: Methodological Debates in the 1930s and 1940s** (George Gale) [*May 17, 2002*]
- **Doing vs. Allowing Harm** (Frances Howard-Snyder) [*May 14, 2002*]
- **Desert** (Owen McLeod) [*May 14, 2002*]
- **Privacy** (Judith DeCew) [*May 14, 2002*]
- **Modal Fictionalism** (Daniel Nolan) [*May 14, 2002*]
- **Philosophy and Christian Theology** (Michael Murray) [*May 13, 2002*]
- **John Buridan** (Jack Zupko) [*May 13, 2002*]
- **The Correspondence Theory of Truth** (Marian David) [*May 10, 2002*]
- **Thomas of Erfurt** (Jack Zupko) [*May 5, 2002*]
- **The Epsilon Calculus** (Jeremy Avigad and Richard Zach) [*May 3, 2002*]
- **Copenhagen Interpretation of Quantum Mechanics** (Jan Faye) [*May 3, 2002*]
- **Aristotle's Rhetoric** (Christof Rapp) [*May 2, 2002*]
- **Philosophy for Children** (Michael Pritchard) [*May 2, 2002*]
- **Zeno's Paradoxes** (Nick Huggett) [*April 30, 2002*]
- **Determinates vs. Determinables** (David H. Sanford) [*April 26, 2002*]
- **Events** (Roberto Casati and Achille Varzi) [*April 22, 2002*]
- **Relative Identity** (Harry Deutsch) [*April 22, 2002*]
- **The Definition of Morality** (Bernard Gert) [*April 16, 2002*]
- **Friedrich Daniel Schleiermacher** (Michael Forster) [*April 16, 2002*]
- **Moral Dilemmas** (Terrance McConnell) [*April 15, 2002*]

- **Descartes' Modal Metaphysics** (David Cunning) [*April 15, 2002*]
- **The Social Dimensions of Scientific Knowledge** (Helen Longino) [*April 12, 2002*]
- **Mally's Deontic Logic** (Gert-Jan Lokhorst) [*April 5, 2002*]
- **Finitism in Geometry** (Jean-Paul Van Bendegem) [*April 3, 2002*]
- **Process Philosophy** (Nicholas Rescher) [*April 2, 2002*]
- **Space and Time: Inertial Frames** (Robert DiSalle) [*March 29, 2002*]
- **Impartiality** (Troy Jollimore) [*March 25, 2002*]
- **Many-Worlds Interpretation of Quantum Mechanics** (Lev Vaidman) [*March 24, 2002*]
- **Action** (George Wilson) [*March 17, 2002*]
- **Lord Shaftesbury [Anthony Ashley Cooper, 3rd Earl of Shaftesbury]** (Michael Gill) [*March 12, 2002*]
- **Harriet Taylor Mill** (Dale E. Miller) [*March 11, 2002*]
- **18th Century German Philosophy Prior to Kant** (Brigitte Sassen) [*March 10, 2002*]
- **Justice as a Virtue** (Michael Slote) [*March 7, 2002*]
- **Bruno Bauer** (Douglas Moggach) [*March 7, 2002*]
- **Collapse Theories** (Giancarlo Ghirardi) [*March 7, 2002*]
- **Cosmopolitanism** (Pauline Kleingeld and Eric Brown) [*February 22, 2002*]
- **Hobbes's Moral and Political Philosophy** (Sharon A. Lloyd) [*February 12, 2002*]
- **George Santayana** (Herman Saatkamp) [*February 11, 2002*]
- **Medieval Theories: Properties of Terms** (Stephen Read) [*February 5, 2002*]
- **Relational Quantum Mechanics** (Federico Laudisa and Carlo Rovelli) [*February 4, 2002*]
- **Quantum Logic and Probability Theory** (Alexander Wilce) [*February 4, 2002*]
- **Formal Learning Theory** (Oliver Schulte) [*February 2, 2002*]
- **Scottish Philosophy in the 19th Century** (Gordon Graham) [*January 29, 2002*]
- **Salomon Maimon** (Peter Thielke and Yitzhak Melamed) [*January 28, 2002*]
- **Robert Boyle** (J. J. MacIntosh) [*January 15, 2002*]
- **Jonathan Edwards** (William Wainwright) [*January 15, 2002*]
- **Evolutionary Game Theory** (J. McKenzie Alexander) [*January 14, 2002*]
- **Free Will** (Timothy O'Connor) [*January 7, 2002*]
- **Intrinsic vs. Extrinsic Properties** (Brian Weatherson) [*January 4, 2002*]
- **Ralph Waldo Emerson** (Russell Goodman) [*January 3, 2002*]
- **John Stuart Mill** (Fred Wilson) [*January 3, 2002*]
- **The Epistemic Closure Principle** (Steven Luper) [*December 31, 2001*]
- **Antonio Rosmini** (Denis Cleary) [*December 28, 2001*]
- **Affirmative Action** (Robert Fullinwider) [*December 28, 2001*]
- **Giles of Rome** (Roberto Lambertini) [*December 20, 2001*]
- **Exploitation** (Alan Wertheimer) [*December 20, 2001*]
- **Legal Rights** (Kenneth Campbell) [*December 20, 2001*]
- **Laozi** (Alan Chan) [*December 14, 2001*]

- **Non-monotonic Logic** (Aldo Antonelli) [*December 10, 2001*]
- **Skepticism** (Peter Klein) [*December 8, 2001*]
- **Friedrich Heinrich Jacobi** (George di Giovanni) [*December 6, 2001*]
- **Replication** (David Hull) [*December 4, 2001*]
- **Mental Illness** (Christian Perring) [*November 30, 2001*]
- **Nationalism** (Nenad Miscevic) [*November 29, 2001*]
- **Early Philosophical Interpretations of General Relativity** (Thomas A. Ryckman) [*November 27, 2001*]
- **Legal Philosophy: The Economic Analysis of Law** (Lewis Kornhauser) [*November 26, 2001*]
- **Thermodynamic Asymmetry in Time** (Craig Callender) [*November 15, 2001*]
- **Zhuangzi** (Harold Roth) [*November 9, 2001*]
- **Model Theory** (Wilfrid Hodges) [*November 9, 2001*]
- **First-order Model Theory** (Wilfrid Hodges) [*November 9, 2001*]
- **Tarski's Truth Definitions** (Wilfrid Hodges) [*November 9, 2001*]
- **Causation in the Law** (Antony Honoré) [*November 8, 2001*]
- **Well Being** (Roger Crisp) [*November 6, 2001*]
- **Bohmian Mechanics** (Sheldon Goldstein) [*October 26, 2001*]
- **Johann Gottfried von Herder** (Michael Forster) [*October 23, 2001*]
- **Spinoza's Psychological Theory** (Michael LeBuffe) [*October 22, 2001*]
- **Law and Ideology** (Christine Sypnowich) [*October 22, 2001*]
- **Friedrich Wilhelm Joseph von Schelling** (Andrew Bowie) [*October 22, 2001*]
- **Nicholas of Autrecourt** (Hans Thijssen) [*October 14, 2001*]
- **The Uncertainty Principle** (Jan Hilgevoord and Jos Uffink) [*October 8, 2001*]
- **Communitarianism** (Daniel Bell) [*October 3, 2001*]
- **The Cambridge Platonists** (Sarah Hutton) [*October 3, 2001*]
- ***Sophismata*** (Fabienne Pironet) [*September 29, 2001*]
- **Gregory of Rimini** (Christopher Schabel) [*September 24, 2001*]
- **Johannes Sharpe** (Alessandro Conti) [*September 24, 2001*]
- **John Wyclif** (Alessandro Conti) [*September 18, 2001*]
- **Indexicals** (David Braun) [*September 14, 2001*]
- **Antoine Le Grand** (Patricia Easton) [*September 13, 2001*]
- **The Equivalence of Mass and Energy** (Francisco Flores) [*September 12, 2001*]
- **John Locke** (William Uzgalis) [*September 2, 2001*]
- **Johann Gottlieb Fichte** (Dan Breazeale) [*August 30, 2001*]
- **Diagrams** (Sun-Joo Shin and Oliver Lemon) [*August 28, 2001*]
- **Common Knowledge** (Peter Vanderschraaf) [*August 27, 2001*]
- **Insolubles** (Paul Vincent Spade) [*August 27, 2001*]
- **Backward Causation** (Jan Faye) [*August 27, 2001*]
- **Paul of Venice** (Alessandro Conti) [*August 22, 2001*]

- **Godfrey of Fontaines** (John Wippel) [*August 17, 2001*]
- **Causation and Manipulability** (James Woodward) [*August 17, 2001*]
- **Gersonides** (Tamar Rudavsky) [*August 17, 2001*]
- **Marsilius of Inghen** (Maarten Hoenen) [*August 16, 2001*]
- **Computer Ethics: Basic Concepts and Historical Overview** (Terrell Bynum) [*August 13, 2001*]
- **Quantum Entanglement and Information** (Jeffrey Bub) [*August 13, 2001*]
- **Medieval Theories of Causation** (Graham White) [*August 10, 2001*]
- **Conditionals** (Dorothy Edgington) [*August 8, 2001*]
- **Richard Kilvington** (Elzbieta Jung-Palczewska) [*August 6, 2001*]
- **Divine Providence** (Hugh J. McCann) [*August 1, 2001*]
- **Comparative Philosophy: Chinese and Western** (David Wong) [*July 31, 2001*]
- **Logic and Games** (Wilfrid Hodges) [*July 27, 2001*]
- **William Penbygull** (Alessandro Conti) [*July 25, 2001*]
- **Robert Alyngton** (Alessandro Conti) [*July 25, 2001*]
- **Robert Holkot** (Hester Gelber) [*July 23, 2001*]
- **Charles Hartshorne** (Dan Dombrowski) [*July 23, 2001*]
- **Abstract Objects** (Gideon Rosen) [*July 19, 2001*]
- **Automated Reasoning** (Frederic Portoraro) [*July 18, 2001*]
- **Bayesian Epistemology** (William Talbott) [*July 12, 2001*]
- **Epistemological Problems of Perception** (Laurence BonJour) [*July 12, 2001*]
- **Being and Becoming in Modern Physics** (Steven Savitt) [*July 11, 2001*]
- **Truthlikeness** (Graham Oddie) [*July 10, 2001*]
- **Naturalized Epistemology** (Richard Feldman) [*July 5, 2001*]
- **Baruch Spinoza** (Steven Nadler) [*June 29, 2001*]
- **Scottish Philosophy in the 18th Century** (Alexander Broadie) [*June 27, 2001*]
- **Charles Sanders Peirce** (Robert Burch) [*June 22, 2001*]
- **Descartes' Ontological Argument** (Lawrence Nolan) [*June 18, 2001*]
- **Moral Particularism** (Jonathan Dancy) [*June 5, 2001*]
- **John Duns Scotus** (Thomas Williams) [*May 30, 2001*]
- **Medieval Theories of Relations** (Jeffrey Brower) [*May 29, 2001*]
- **Interpretation and Coherence in Legal Reasoning** (Julie Dickson) [*May 29, 2001*]
- **The Nature of Law** (Andrei Marmor) [*May 27, 2001*]
- **Panpsychism** (William Seager) [*May 23, 2001*]
- **Aristotle's Ethics** (Richard Kraut) [*May 1, 2001*]
- **Peter of Spain** (Joke Spruyt) [*April 12, 2001*]
- **Philosophy of Statistical Mechanics** (Lawrence Sklar) [*April 12, 2001*]
- **Descartes' Life and Works** (Kurt Smith) [*April 9, 2001*]
- **Integrity** (Damian Cox, Marguerite La Caze, and Michael Levine) [*April 9, 2001*]

- **Higher-Order Theories of Consciousness** (Peter Carruthers) [*April 3, 2001*]
- **The Unity of Consciousness** (Andrew Brook) [*March 27, 2001*]
- **Equality** (Stefan Gosepath) [*March 26, 2001*]
- **Francis of Marchia** (Christopher Schabel) [*March 23, 2001*]
- **Robert Desgabets** (Patricia Easton) [*March 22, 2001*]
- **Social Epistemology** (Alvin Goldman) [*February 26, 2001*]
- **David Hume** (William Edward Morris) [*February 26, 2001*]
- **John Austin** (Brian Bix) [*February 23, 2001*]
- **Physicalism** (Daniel Stoljar) [*February 13, 2001*]
- **The Analysis of Knowledge** (Matthias Steup) [*February 5, 2001*]
- **Richard Rorty** (Bjørn Ramberg) [*February 3, 2001*]
- **Benjamin Peirce** (Ivor Grattan-Guinness and Alison Walsh) [*February 3, 2001*]
- **Albert of Saxony** (Joël Biard) [*January 29, 2001*]
- **Dante Alighieri** (Winthrop Wetherbee) [*January 29, 2001*]
- **Semantic Challenges to Realism** (Drew Khlentzos) [*January 11, 2001*]
- **Evolutionary Epistemology** (Michael Bradie and William Harms) [*January 10, 2001*]
- **Constitutionalism** (Wil Waluchow) [*January 10, 2001*]
- **Counterfactual Theories of Causation** (Peter Menzies) [*January 10, 2001*]
- **Curry's Paradox** (JC Beall) [*January 10, 2001*]
- **Moral Responsibility** (Andrew Eshleman) [*January 5, 2001*]
- **Disjunction** (Ray Jennings) [*January 5, 2001*]
- **Legal Punishment** (Antony Duff) [*January 2, 2001*]
- **Intertheory Relations in Physics** (Robert Batterman) [*January 2, 2001*]
- **William Whewell** (Laura J. Snyder) [*December 22, 2000*]
- **The Modern History of Computing** (B. Jack Copeland) [*December 17, 2000*]
- **Quantum Mechanics** (Jenann Ismael) [*November 28, 2000*]
- **Brentano's Theory of Judgement** (Johannes Brandl) [*November 22, 2000*]
- **Feminist History of Philosophy** (Charlotte Witt) [*November 3, 2000*]
- **Aristotle's Metaphysics** (S. Marc Cohen) [*October 8, 2000*]
- **Classical Logic** (Stewart Shapiro) [*September 15, 2000*]
- **The Kochen-Specker Theorem** (Carsten Held) [*September 10, 2000*]
- **The Medieval Problem of Universals** (Gyula Klima) [*September 10, 2000*]
- **William James** (Russell Goodman) [*September 7, 2000*]
- **Thomas Reid** (Gideon Yaffe) [*August 28, 2000*]
- **The Experience and Perception of Time** (Robin Le Poidevin) [*August 28, 2000*]
- **Incompatibilist (Nondeterministic) Theories of Free Will** (Randolph Clarke) [*August 16, 2000*]
- **Feminist Epistemology and Philosophy of Science** (Elizabeth Anderson) [*August 9, 2000*]
- **Substructural Logics** (Greg Restall) [*July 4, 2000*]

- **Contractarianism** (Ann Cudd) [*June 18, 2000*]
- **Actualism** (Christopher Menzel) [*June 16, 2000*]
- **Behaviorism** (George Graham) [*May 26, 2000*]
- **Representational Theories of Consciousness** (William Lycan) [*May 22, 2000*]
- **Saint Anselm** (Thomas Williams) [*May 18, 2000*]
- **Many-Valued Logic** (Siegfried Gottwald) [*April 25, 2000*]
- **Mental Representation** (David Pitt) [*March 30, 2000*]
- **Saint Augustine** (Michael Mendelson) [*March 24, 2000*]
- **Aristotle's Logic** (Robin Smith) [*March 18, 2000*]
- **Modal Logic** (James Garson) [*February 29, 2000*]
- **Foundationalist Theories of Epistemic Justification** (Richard Fumerton) [*February 21, 2000*]
- **Time Travel and Modern Physics** (Frank Arntzenius and Tim Maudlin) [*February 17, 2000*]
- **Propositional Attitude Reports** (Thomas McKay) [*February 16, 2000*]
- **Identity and Individuality in Quantum Theory** (Steven French) [*February 15, 2000*]
- **War** (Brian Orend) [*February 4, 2000*]
- **Infinitary Logic** (John L. Bell) [*January 23, 2000*]
- **William Godwin** (Mark Philp) [*January 16, 2000*]
- **The Identity Theory of Mind** (J. J. C. Smart) [*January 12, 2000*]
- **Aristotle's Psychology** (Christopher Shields) [*January 11, 2000*]
- **Temporal Logic** (Antony Galton) [*November 29, 1999*]
- **Medieval Theories of Analogy** (E. Jennifer Ashworth) [*November 29, 1999*]
- **Divine Illumination** (Robert Pasnau) [*November 2, 1999*]
- **Peter John Olivi** (Robert Pasnau) [*November 2, 1999*]
- **Logical Form** (Paul Pietroski) [*October 19, 1999*]
- **Measurement in Quantum Theory** (Henry Krips) [*October 12, 1999*]
- **Medieval Theories of Practical Reason** (Anthony Celano) [*October 8, 1999*]
- **Properties** (Chris Swoyer) [*September 23, 1999*]
- **Reichenbach's Common Cause Principle** (Frank Arntzenius) [*September 22, 1999*]
- **Intuitionistic Logic** (Joan Moschovakis) [*September 1, 1999*]
- **Richard the Sophister** (Paul Streveler) [*August 4, 1999*]
- **Nineteenth Century Geometry** (Roberto Torretti) [*July 26, 1999*]
- **Holism and Nonseparability in Physics** (Richard Healey) [*July 22, 1999*]
- **Saint Thomas Aquinas** (Ralph McInerny) [*July 12, 1999*]
- **Virtue Epistemology** (John Greco) [*July 9, 1999*]
- **Medieval Theories of Modality** (Simo Knuuttila) [*June 30, 1999*]
- **Supertasks** (Jon Pérez Laraudogoitia) [*June 29, 1999*]
- **Feminist Perspectives on the Self** (Diana Meyers) [*June 28, 1999*]
- **The Philosophy of Neuroscience** (John Bickle and Peter Mandik) [*June 7, 1999*]

- **Philip the Chancellor** (Colleen McCluskey) [*February 23, 1999*]
- **The Hole Argument** (John Norton) [*February 1, 1999*]
- **Epiphenomenalism** (William Robinson) [*January 18, 1999*]
- **Artifact** (Risto Hilpinen) [*January 5, 1999*]
- **Indispensability Arguments in the Philosophy of Mathematics** (Mark Colyvan) [*December 21, 1998*]
- **Dialetheism** (Graham Priest) [*December 4, 1998*]
- **Medieval Theories of Conscience** (Douglas Langston) [*November 23, 1998*]
- **Multiple Realizability** (John Bickle) [*November 23, 1998*]
- **The St. Petersburg Paradox** (Robert Martin) [*November 4, 1998*]
- **Experiment in Physics** (Allan Franklin) [*October 5, 1998*]
- **Conventionality of Simultaneity** (Allen Janis) [*August 31, 1998*]
- **Cosmology and Theology** (John Leslie) [*July 2, 1998*]
- **Aristotle's Political Theory** (Fred Miller) [*July 1, 1998*]
- **Relevance Logic** (Edwin Mares) [*June 17, 1998*]
- **Frege's Logic, Theorem, and Foundations for Arithmetic** (Edward N. Zalta) [*June 10, 1998*]
- **Everett's Relative-State Formulation of Quantum Mechanics** (Jeffrey Barrett) [*June 3, 1998*]
- **The Language of Thought Hypothesis** (Murat Aydede) [*May 28, 1998*]
- **Pascal's Wager** (Alan Hájek) [*May 2, 1998*]
- **Feminist Ethics** (Rosemarie Tong) [*February 16, 1998*]
- **Leibniz on the Problem of Evil** (Michael Murray) [*January 4, 1998*]
- **Jacques Maritain** (William Sweet) [*December 5, 1997*]
- **Descartes' Epistemology** (Lex Newman) [*December 3, 1997*]
- **Color** (Barry Maund) [*December 1, 1997*]
- **Mental Imagery** (Nigel Thomas) [*November 18, 1997*]
- **Constructive Mathematics** (Douglas Bridges) [*November 18, 1997*]
- **Karl Popper** (Stephen Thornton) [*November 13, 1997*]
- **Ancient Skepticism** (Leo Groarke) [*November 4, 1997*]
- **Leibniz's Philosophy of Mind** (Mark Kulstad and Laurence Carlin) [*September 22, 1997*]
- **Structured Propositions** (Jeffrey C. King) [*September 22, 1997*]
- **Folk Psychology as a Theory** (Ian Ravenscroft) [*September 22, 1997*]
- **Prisoner's Dilemma** (Steven Kuhn) [*September 4, 1997*]
- **The Deflationary Theory of Truth** (Daniel Stoljar) [*August 28, 1997*]
- **Paul Feyerabend** (John Preston) [*August 26, 1997*]
- **Qualia** (Michael Tye) [*August 20, 1997*]
- **The Traditional Square of Opposition** (Terence Parsons) [*August 8, 1997*]
- **Singular Propositions** (Greg Fitch) [*July 19, 1997*]
- **Probabilistic Causation** (Christopher Hitchcock) [*July 11, 1997*]
- **Bernard Bosanquet** (William Sweet) [*June 15, 1997*]

- **Friedrich Nietzsche** (Robert Wicks) [*May 30, 1997*]
- **Connectionism** (James Garson) [*May 18, 1997*]
- **The Epistemology of Religion** (Peter Forrest) [*April 23, 1997*]
- **Wilfrid Sellars** (Jay Rosenberg) [*February 22, 1997*]
- **Tropes** (John Bacon) [*February 19, 1997*]
- **Georg Wilhelm Friedrich Hegel** (Paul Redding) [*February 13, 1997*]
- **Vagueness** (Roy Sorensen) [*February 8, 1997*]
- **Game Theory** (Don Ross) [*January 25, 1997*]
- **Sorites Paradox** (Dominic Hyde) [*January 17, 1997*]
- **Folk Psychology as Mental Simulation** (Robert M. Gordon) [*January 9, 1997*]
- **The Church-Turing Thesis** (B. Jack Copeland) [*January 8, 1997*]
- **Thought Experiments** (James R. Brown) [*December 28, 1996*]
- **Causal Processes** (Phil Dowe) [*December 8, 1996*]
- **Category Theory** (Jean-Pierre Marquis) [*December 6, 1996*]
- **Holes** (Roberto Casati and Achille Varzi) [*December 5, 1996*]
- **Søren Kierkegaard** (William McDonald) [*December 3, 1996*]
- **Liberalism** (Gerald Gaus) [*November 28, 1996*]
- **Informal Logic** (Leo Groarke) [*November 25, 1996*]
- **Logical Constructions** (Bernard Linsky) [*November 20, 1996*]
- **Arthur Prior** (B. Jack Copeland) [*October 7, 1996*]
- **Paraconsistent Logic** (Graham Priest and Koji Tanaka) [*September 24, 1996*]
- **Cognitive Science** (Paul Thagard) [*September 23, 1996*]
- **Distributive Justice** (Julian Lamont) [*September 22, 1996*]
- **Miracles** (Michael Levine) [*September 4, 1996*]
- **The Coherence Theory of Truth** (James O. Young) [*September 3, 1996*]
- **Existence** (Barry Miller) [*August 22, 1996*]
- **Historicist Theories of Rationality** (Carl Matheson) [*August 12, 1996*]
- **The Identity of Indiscernibles** (Peter Forrest) [*July 31, 1996*]
- **Private Language** (Stewart Candlish) [*July 26, 1996*]
- **Inconsistent Mathematics** (Chris Mortensen) [*July 2, 1996*]
- **Pantheism** (Michael Levine) [*June 4, 1996*]
- **Donald Davidson** (Jeff Malpas) [*May 29, 1996*]
- **Alfred North Whitehead** (A. D. Irvine) [*May 21, 1996*]
- ***Principia Mathematica*** (A. D. Irvine) [*May 21, 1996*]
- **Francis Herbert Bradley** (Stewart Candlish) [*May 9, 1996*]
- **Voluntary Euthanasia** (Robert Young) [*April 18, 1996*]
- **Stoicism** (Dirk Baltzly) [*April 15, 1996*]
- **The Identity Theory of Truth** (Stewart Candlish) [*March 28, 1996*]

- **Teleological Notions in Biology** (Colin Allen) [*March 20, 1996*]
- **Contemporary Approaches to the Social Contract** (Fred D'Agostino) [*March 3, 1996*]
- **Original Position** (Fred D'Agostino) [*February 27, 1996*]
- **Public Justification** (Fred D'Agostino) [*February 27, 1996*]
- **Ontological Arguments** (Graham Oppy) [*February 8, 1996*]
- **Animal Consciousness** (Colin Allen) [*December 23, 1995*]
- **The Revision Theory of Truth** (Eric Hammer) [*December 15, 1995*]
- **Peirce's Logic** (Eric Hammer) [*December 15, 1995*]
- **Russell's Paradox** (A. D. Irvine) [*December 8, 1995*]
- **Bertrand Russell** (A. D. Irvine) [*December 7, 1995*]
- **Gottlob Frege** (Edward N. Zalta) [*September 14, 1995*]
- **Turing Machine** (Editors at the SEP) [*September 14, 1995*]

Homosexuality

The term ‘homosexuality’ was coined in the late 19th century by a German psychologist, Karoly Maria Benkert. Although the term is new, discussions about sexuality in general, and same-sex attraction in particular, have occasioned philosophical discussion ranging from Plato's *Symposium* to contemporary queer theory. Since the history of cultural understandings of same-sex attraction is relevant to the philosophical issues raised by those understandings, it is necessary to review briefly some of the social history of homosexuality. Arising out of this history, at least in the West, is the idea of natural law and some interpretations of that law as forbidding homosexual sex. References to natural law still play an important role in contemporary debates about homosexuality in religion, politics, and even courtrooms. Finally, perhaps the most significant recent social change involving homosexuality is the emergence of the gay liberation movement in the West. In philosophical circles this movement is, in part, represented through a rather diverse group of thinkers who are grouped under the label of queer theory. A central issue raised by queer theory, which will be discussed below, is whether homosexuality, and hence also heterosexuality and bisexuality, is socially constructed or purely driven by biological forces.

History

As has been frequently noted, the ancient Greeks did not have terms or concepts that correspond to the contemporary dichotomy of ‘heterosexual’ and ‘homosexual’. There is a wealth of material from ancient Greece pertinent to issues of sexuality, ranging from dialogues of Plato, such as the *Symposium*, to plays by Aristophanes, and Greek artwork and vases. What follows is a brief description of ancient Greek attitudes, but it is important to recognize that there was regional variation. For example, in parts of Ionia there were general strictures against same-sex *eros*, while in Elis and Boiotia (e.g., Thebes), it was

approved of and even celebrated (cf. Dover, 1989; Halperin, 1990).

Probably the most frequent assumption of sexual orientation is that persons can respond erotically to beauty in either sex. Diogenes Laeurtius, for example, wrote of Alcibiades, the Athenian general and politician of the 5th century B.C., “in his adolescence he drew away the husbands from their wives, and as a young man the wives from their husbands.” (Quoted in Greenberg, 1988, 144) Some persons were noted for their exclusive interests in persons of one gender. For example, Alexander the Great and the founder of Stoicism, Zeno of Citium, were known for their exclusive interest in boys and other men. Such persons, however, are generally portrayed as the exception. Furthermore, the issue of what gender one is attracted to is seen as an issue of taste or preference, rather than as a moral issue. A character in Plutarch's *Erotikos* (*Dialogue on Love*) argues that “the noble lover of beauty engages in love wherever he sees excellence and splendid natural endowment without regard for any difference in physiological detail.” (*Ibid*., 146) Gender just becomes irrelevant “detail” and instead the excellence in character and beauty is what is most important.

Even though the gender that one was erotically attracted to (at any specific time, given the assumption that persons will likely be attracted to persons of both sexes) was not important, other issues were salient, such as whether one exercised moderation. Status concerns were also of the highest importance. Given that only free men had full status, women and male slaves were not problematic sexual partners. Sex between freemen, however, was problematic for status. The central distinction in ancient Greek sexual relations was between taking an active or insertive role, versus a passive or penetrated one. The passive role was acceptable only for inferiors, such as women, slaves, or male youths who were not yet citizens. Hence the cultural ideal of a same-sex relationship was between an older man, probably in his 20's or 30's, known as the *erastes*, and a boy whose beard had not yet begun to grow, the *eromenos* or *paidika*. In this relationship there was courtship ritual, involving gifts (such as a rooster), and other norms. The *erastes* had to show that he had nobler interests in the boy, rather than a purely sexual concern. The boy was not to submit too easily, and if pursued by more than one man, was to show discretion and pick the more noble one. There is also evidence that penetration was often avoided by having the *erastes* face his beloved and place his penis between the thighs of the *eromenos*, which is known as intercrural sex. The relationship was to be temporary and should end upon the boy reaching adulthood (Dover, 1989). To continue in a submissive role even while one should be an equal citizen was considered troubling, although there certainly were many adult male same-sex relationships that were noted and not strongly stigmatized. While the passive role was thus seen as problematic, to be attracted to men was often taken as a sign of masculinity. Greek gods, such as Zeus, had stories of same-sex exploits attributed to them, as did other key figures in Greek myth and literature, such as Achilles and Hercules. Plato, in the *Symposium*, argues for an army to be comprised of same-sex lovers. Thebes did form such a regiment, the Sacred Band of Thebes, formed of 500 soldiers. They were renowned in the ancient world for their valor in battle.

Ancient Rome had many parallels in its understanding of same-sex attraction, and sexual issues more generally, to ancient Greece. This is especially true under the Republic. Yet under the Empire, Roman society slowly became more negative in its views towards sexuality, probably due to social and economic turmoil, even before Christianity became influential.

Exactly what attitude the New Testament has towards sexuality in general, and same-sex attraction in particular, is a matter of sharp debate. John Boswell argues, in his fascinating *Christianity, Social Tolerance, and Homosexuality*, that many passages taken today as condemnations of homosexuality are more concerned with prostitution, or where same-sex acts are described as “unnatural” the meaning is more akin to ‘out of the ordinary’ rather than as immoral (Boswell, 1980, ch.4; see also Boswell, 1994). Yet others have criticized, sometimes persuasively, Boswell's scholarship (see Greenberg, 1988, ch.5). What is clear, however, is that while condemnation of same-sex attraction is marginal to the Gospels and only an intermittent focus in the rest of the New Testament, early Christian church fathers were much more outspoken. In their writings there is a horror at any sort of sex, but in a few generations these views eased, in part due no doubt to practical concerns of recruiting converts. By the fourth and fifth centuries the mainstream Christian view allowed for procreative sex.

This viewpoint, that procreative sex within marriage is allowed, while every other expression of sexuality is sinful, can be found, for example, in St. Augustine. This understanding leads to a concern with the gender of one's partner that is not found in previous Greek or Roman views, and it clearly forbids homosexual acts. Soon this attitude, especially towards homosexual sex, came to be reflected in Roman Law. In Justinian's Code, promulgated in 529, persons who engaged in homosexual sex were to be executed, although those who were repentant could be spared. Historians agree that the late Roman Empire saw a rise in intolerance towards sexuality, although there were again important regional variations.

With the decline of the Roman Empire, and its replacement by various barbarian kingdoms, a general tolerance (with the sole exception of Visigothic Spain) of homosexual acts prevailed. As one prominent scholar puts it, “European secular law contained few measures against homosexuality until the middle of the thirteenth century.” (Greenberg, 1988, 260) Even while some Christian theologians continued to denounce nonprocreative sexuality, including same-sex acts, a genre of homophilic literature, especially among the clergy, developed in the eleventh and twelfth centuries (Boswell, 1980, chapters 8 and 9).

The latter part of the twelfth through the fourteenth centuries, however, saw a sharp rise in intolerance towards homosexual sex, alongside persecution of Jews, Muslims, heretics, and others. While the causes of this are somewhat unclear, it is likely that increased class conflict alongside the Gregorian reform movement in the Catholic Church were two important factors. The Church itself started to appeal to a conception of “nature” as the standard of morality, and drew it in such a way so as to forbid homosexual sex (as well as extramarital sex, nonprocreative sex within marriage, and often masturbation). For example, the first ecumenical council to condemn homosexual sex, Lateran III of 1179, stated that “Whoever shall be found to have committed that incontinence which is against nature” shall be punished, the severity of which depended upon whether the transgressor was a cleric or layperson (quoted in Boswell, 1980, 277). This appeal to natural law (discussed below) became very influential in the Western tradition. An important point to note, however, is that the key category here is the ‘sodomite,’ which differs from the contemporary idea of ‘homosexual’. A sodomite was understood as act-defined, rather than as a type of person. Someone who had desires to engage in sodomy, yet did not act upon them, was not a sodomite. Also, persons who engaged in heterosexual sodomy were also sodomites. There are

reports of persons being burned to death or beheaded for sodomy with a spouse (Greenberg, 1988, 277). Finally, a person who had engaged in sodomy, yet who had repented of his sin and vowed to never do it again, was no longer a sodomite. The gender of one's partner is again not of decisive importance, although some medieval theologians single out same-sex sodomy as the worst type of sexual crime.

For the next several centuries in Europe, the laws against homosexual sex were severe in their penalties. Enforcement, however, was episodic. In some regions, decades would pass without any prosecutions. Yet the Dutch, in the 1730's, mounted a harsh anti-sodomy campaign (alongside an anti-Gypsy pogrom), even using torture to obtain confessions. As many as one hundred men and boys were executed and denied burial (Greenberg, 1988, 313-4). Also, the degree to which sodomy and same-sex attraction were accepted varied by class, with the middle class taking the narrowest view, while the aristocracy and nobility often accepted public expressions of alternative sexualities. At times, even with the risk of severe punishment, same-sex oriented subcultures would flourish in cities, sometimes only to be suppressed by the authorities. In the 19th century there was a significant reduction in the legal penalties for sodomy. The Napoleonic code decriminalized sodomy, and with Napoleon's conquests that Code spread. Furthermore, in many countries where homosexual sex remained a crime, the general movement at this time away from the death penalty usually meant that sodomy was removed from the list of capital offenses.

In the 18th and 19th centuries an overtly theological framework no longer dominated the discourse about same-sex attraction. Instead, secular arguments and interpretations became increasingly common. Probably the most important secular domain for discussions of homosexuality was in medicine, including psychology. This discourse, in turn, linked up with considerations about the state and its need for a growing population, good soldiers, and intact families marked by clearly defined gender roles. Doctors were called in by courts to examine sex crime defendants (Foucault, 1980; Greenberg, 1988). At the same time, the dramatic increase in school attendance rates and the average length of time spent in school, reduced transgenerational contact, and hence also the frequency of transgenerational sex. Same-sex relations between persons of roughly the same age became the norm.

Clearly the rise in the prestige of medicine resulted in part from the increasing ability of science to account for natural phenomena on the basis of mechanistic causation. The application of this viewpoint to humans led to accounts of sexuality as innate or biologically driven. The voluntarism of the medieval understanding of sodomy, that sodomites chose sin, gave way to the modern notion of homosexuality as a deep, unchosen characteristic of persons, regardless of whether they act upon that orientation. The idea of a ‘latent sodomite’ would not have made sense, yet under this new view it does make sense to speak of a person as a ‘latent homosexual.’ Instead of specific acts defining a person, as in the medieval view, an entire physical and mental makeup, usually portrayed as somehow defective or pathological, is ascribed to the modern category of ‘homosexual.’ Although there are historical precursors to these ideas (e.g., Aristotle gave a physiological explanation of passive homosexuality), medicine gave them greater public exposure and credibility (Greenberg, 1988, ch.15). The effects of these ideas cut in conflicting ways. Since homosexuality is, by this view, not chosen, it makes less sense to criminalize it. Persons are not choosing evil acts. Yet persons may be expressing a diseased or pathological mental state, and hence medical intervention for a cure is appropriate. Hence doctors, especially psychiatrists, campaigned for the

repeal or reduction of criminal penalties for consensual homosexual sodomy, yet intervened to "rehabilitate" homosexuals. They also sought to develop techniques to prevent children from becoming homosexual, for example by arguing that childhood masturbation caused homosexuality, hence it must be closely guarded against.

In the 20th century sexual roles were redefined once again. For a variety of reasons, premarital intercourse slowly became more common and eventually acceptable. With the decline of prohibitions against sex for the sake of pleasure even outside of marriage, it became more difficult to argue against gay sex. These trends were especially strong in the 1960's, and it was in this context that the gay liberation movement took off. Although gay and lesbian rights groups had been around for decades, the low-key approach of the Mattachine Society (named after a medieval secret society) and the Daughters of Bilitis had not gained much ground. This changed in the early morning hours of June 28, 1969, when the patrons of the Stonewall Inn, a gay bar in Greenwich Village, rioted after a police raid. In the aftermath of that event, gay and lesbian groups began to organize around the country. Gay Democratic clubs were created in every major city, and one fourth of all college campuses had gay and lesbian groups (Shilts, 1993, ch.28). Large gay urban communities in cities from coast to coast became the norm. The American Psychiatric Association removed homosexuality from its official listing of mental disorders. The increased visibility of gays and lesbians has become a permanent feature of American life despite the two critical setbacks of the AIDS epidemic and an anti-gay backlash (see Berman, 1993, for a good survey). The post-Stonewall era has also seen marked changes in Western Europe, where the repeal of anti-sodomy laws and legal equality for gays and lesbians has become common.

Natural Law

Today natural law theory offers the most common intellectual defense for differential treatment of gays and lesbians, and as such it merits attention. The development of natural law is a long and very complicated story, but a reasonable place to begin is with the dialogues of Plato, for this is where some of the central ideas are first articulated, and, significantly enough, are immediately applied to the sexual domain. For the Sophists, the human world is a realm of convention and change, rather than of unchanging moral truth. Plato, in contrast, argued that unchanging truths underpin the flux of the material world. Reality, including eternal moral truths, is a matter of *phusis*. Even though there is clearly a great degree of variety in conventions from one city to another (something ancient Greeks became increasingly aware of), there is still an unwritten standard, or law, that humans should live under.

In the *Laws*, Plato applies the idea of a fixed, natural law to sex, and takes a much harsher line than he does in the *Symposium* or the *Phraedrus*. In Book One he writes about how opposite-sex sex acts cause pleasure by nature, while same-sex sexuality is "unnatural" (636c). In Book Eight, the Athenian speaker considers how to have legislation banning homosexual acts, masturbation, and illegitimate procreative sex widely accepted. He then states that this law is according to nature (838-839d). Probably the best way of understanding Plato's discussion here is in the context of his overall concerns with the appetitive part of the soul and how best to control it. Plato clearly sees same-sex passions as especially strong, and hence particularly problematic, although in the *Symposium* that erotic attraction could be the catalyst for

a life of philosophy, rather than base sensuality (Cf. Dover, 1989, 153-170; Nussbaum, 1999, esp. chapter 12).

Other figures played important roles in the development of natural law theory. Aristotle, with his emphasis upon reason as the distinctive human function, and the Stoics, with their emphasis upon human beings as a part of the natural order of the cosmos, both helped to shape the natural law perspective which says that “True law is right reason in agreement with nature,” as Cicero put it. Aristotle, in his approach, did allow for change to occur according to nature, and therefore the way that natural law is embodied could itself change with time, which was an idea Aquinas later incorporated into his own natural law theory. Aristotle did not write extensively about sexual issues, since he was less concerned with the appetites than Plato. Probably the best reconstruction of his views places him in mainstream Greek society as outlined above; the main issue is that of active versus a passive role, with only the latter problematic for those who either are or will become citizens. Zeno, the founder of Stoicism, was, according to his contemporaries, only attracted to men, and his thought had no prohibitions against same-sex sexuality. In contrast, Cicero, a later Stoic, was dismissive about sexuality in general, with some harsher remarks towards same-sex pursuits (Cicero, 1966, 407-415).

The most influential formulation of natural law theory was made by Thomas Aquinas in the thirteenth century. Integrating an Aristotelian approach with Christian theology, Aquinas emphasized the centrality of certain human goods, including marriage and procreation. While Aquinas did not write much about same-sex sexual relations, he did write at length about various sex acts as sins. For Aquinas, sexuality that was within the bounds of marriage *and* which helped to further what he saw as the distinctive goods of marriage, mainly love, companionship, and legitimate offspring, was permissible, and even good. Aquinas did not argue that procreation was a necessary part of moral or just sex; married couples could enjoy sex without the motive of having children, and sex in marriages where one or both partners is sterile (perhaps because the woman is postmenopausal) is also potentially just (given a motive of expressing love). So far Aquinas' view actually need not rule out homosexual sex. For example, a Thomist could embrace same-sex marriage, and then apply the same reasoning, simply seeing the couple as a reproductively sterile, yet still fully loving and companionate union.

Aquinas, in a significant move, adds a requirement that for any given sex act to be moral it must be of a generative kind. The only way that this can be achieved is via vaginal intercourse. That is, since only the emission of semen in a vagina can result in natural reproduction, only sex acts of that type are generative, even if a given sex act does not lead to reproduction, and even if it is impossible due to infertility. The consequence of this addition is to rule out the possibility, of course, that homosexual sex could ever be moral (even if done within a loving marriage), in addition to forbidding any non-vaginal sex for opposite-sex married couples. What is the justification for this important addition? This question is made all the more pressing in that Aquinas does allow that how broad moral rules apply to individuals may vary considerably, since the nature of persons also varies to some extent. That is, since Aquinas allows that individual natures vary, one could simply argue that one is, by nature, emotionally and physically attracted to persons of one's own gender, and hence to pursue same-sex relationships is ‘natural’ (Sullivan, 1995). Unfortunately, Aquinas does not spell out a justification for this generative requirement.

More recent natural law theorists, however, have tried a couple different lines of defense for Aquinas' ‘generative type’ requirement. The first is that sex acts that involve either homosexuality, heterosexual sodomy, or which use contraception, frustrate the purpose of the sex organs, which is reproductive. This argument, often called the ‘perverted faculty argument’, is perhaps implicit in Aquinas. It has, however, come in for sharp attack (see Weitham, 1997), and the best recent defenders of a Thomistic natural law approach are attempting to move beyond it (e.g., George, 1999, dismisses the argument). If their arguments fail, of course, they must allow that some homosexual sex acts are morally permissible (even positively good), although they would still have resources with which to argue against casual gay (and straight) sex.

Although the specifics of the second sort of argument offered by various contemporary natural law theorists vary, the common elements are strong (Finnis, 1994; George, 1999). As Thomists, their argument rests largely upon an account of human goods. The two most important for the argument against homosexual sex (though not against homosexuality as an orientation which is not acted upon, and hence in this they follow official Catholic doctrine; see George, 1999, ch.15) are personal integration and marriage. Personal integration, in this view, is the idea that humans, as agents, need to have integration between their intentions as agents and their embodied selves. Thus, to use one's or another's body as a mere means to one's own pleasure, as they argue happens with masturbation, causes ‘dis-integration’ of the self. That is, one's intention then is just to use a body (one's own or another's) as a mere means to the end of pleasure, and this detracts from personal integration. Yet one could easily reply that two persons of the same sex engaging in sexual union does not necessarily imply any sort of ‘use’ of the other as a mere means to one's own pleasure. Hence, natural law theorists respond that sexual union in the context of the realization of marriage as an important human good is the only permissible expression of sexuality. Yet this argument requires drawing how marriage is an important good in a very particular way, since it puts procreation at the center of marriage as its “natural fulfillment” (George, 1999, 168). Natural law theorists, if they want to support their objection to homosexual sex, have to emphasize procreation. If, for example, they were to place love and mutual support for human flourishing at the center, it is clear that many same-sex couples would meet this standard. Hence their sexual acts would be morally just.

There are, however, several objections that are made against this account of marriage as a central human good. One is that by placing procreation as the ‘natural fulfillment’ of marriage, sterile marriages are thereby denigrated. Sex in an opposite-sex marriage where the partners know that one or both of them are sterile is not done for procreation. Yet surely it is not wrong. Why, then, is homosexual sex in the same context (a long-term companionate union) wrong (Macedo, 1995)? The natural law rejoinder is that while vaginal intercourse is a potentially procreative sex act, considered in itself (though admitting the possibility that it may be impossible for a particular couple), oral and anal sex acts are never potentially procreative, whether heterosexual or homosexual (George, 1999). But is this biological distinction also morally relevant, and in the manner that natural law theorists assume? Natural law theorists, in their discussions of these issues, seem to waver. On the one hand, they want to defend an ideal of marriage as a loving union wherein two persons are committed to their mutual flourishing, and where sex is a complement to that ideal. Yet that opens the possibility of permissible gay sex, or heterosexual sodomy, both of which they want to oppose. So they then defend an account of sexuality which seems crudely

reductive, emphasizing procreation to the point where literally a male orgasm anywhere except in the vagina of one's loving spouse is impermissible. Then, when accused of being reductive, they move back to the broader ideal of marriage.

Natural law theory, at present, has made significant concessions to mainstream liberal thought. In contrast certainly to its medieval formulation, most contemporary natural law theorists argue for limited governmental power, and do not believe that the state has an interest in attempting to prevent all moral wrongdoing. Still, they do argue against homosexuality, and against legal protections for gays and lesbians in terms of employment and housing, even to the point of serving as expert witnesses in court cases or helping in the writing of *amicus curae* briefs. They also argue against same sex marriage (Bradley, 2001; George, 2001).

Queer Theory and the Social Construction of Sexuality

With the rise of the gay liberation movement in the post-Stonewall era, overtly gay and lesbian perspectives began to be put forward in politics, philosophy and literary theory. Initially these often were overtly linked to feminist analyses of patriarchy (e.g., Rich, 1980) or other, earlier approaches to theory. Yet in the late 1980's and early 1990's queer theory was developed, although there are obviously important antecedents which make it difficult to date it precisely. There are a number of ways in which queer theory differed from earlier gay liberation theory, but an important initial difference can be gotten at by examining the reasons for opting for the term ‘queer’ as opposed to ‘gay and lesbian.’ Some versions of, for example, lesbian theory portrayed the essence of lesbian identity and sexuality in very specific terms: non-hierarchical, consensual, and, specifically in terms of sexuality, as not necessarily focused upon genitalia (e.g., Faderman, 1985). Lesbians arguing from this framework, for example, could very well criticize natural law theorists as inscribing into the very “law of nature” an essentially masculine sexuality, focused upon the genitals, penetration, and the status of the male orgasm (natural law theorists never mention female orgasms).

This approach, based upon characterizations of ‘lesbian’ and ‘gay’ identity and sexuality, however, suffered from three difficulties. First, it appeared even though the goal was to critique a heterosexist regime for its exclusion and marginalization of those whose sexuality is different, any specific or “essentialist” account of gay or lesbian sexuality had the same effect. Sticking with the example used above, of a specific conceptualization of lesbian identity, it denigrates women who are sexually and emotionally attracted to other women, yet who do not fit the description. Sado-masochists and butch/fem lesbians arguably do not fit this ideal of ‘equality’ offered. A second problem was that by placing such an emphasis upon the gender of one's sexual partner(s), other possible important sources of identity are marginalized, such as race and ethnicity. What is of utmost importance, for example, for a black lesbian is her lesbianism, rather than her race. Many gays and lesbians of color attacked this approach, accusing it of re-inscribing an essentially white identity into the heart of gay or lesbian identity (Jagose, 1996).

The third and final problem for the gay liberationist approach was that it often took this category of ‘identity’ itself as unproblematic and unhistorical. Such a view, however, largely because of arguments developed within poststructuralism, seemed increasingly untenable. The key figure in the attack upon identity as an ahistorical thing is Michel Foucault. In a series of works he set out to analyze the history of sexuality from ancient Greece to the modern era (1980, 1985, 1986). Although the project was tragically cut short by his death in 1984, from complications arising from AIDS, Foucault articulated how profoundly understandings of sexuality can vary across time and space, and his arguments have proven very influential in gay and lesbian theorizing in general, and queer theory in particular (Spargo, 1999).

One of the reasons for the historical review above is that it helps to give some background for understanding the claim that sexuality is socially constructed, rather than given by nature. Moreover, in order to not prejudge the issue of social constructionism versus essentialism, I avoided applying the term ‘homosexual’ to the ancient or medieval eras. In ancient Greece the gender of one's partner(s) was not important, but instead whether one took the active or passive role. In the medieval view, a ‘sodomite’ was a person who succumbed to temptation and engaged in certain non-procreative sex acts. Although the gender of the partner was more important than in the ancient view, the broader theological framework placed the emphasis upon a sin versus refraining-from-sin dichotomy. With the rise of the notion of ‘homosexuality’ in the modern era, a person is placed into a specific category even if one does not act upon those inclinations. What is the common, natural sexuality expressed across these three very different cultures? The social constructionist answer is that there is no ‘natural’ sexuality; all sexual understandings are constructed within and mediated by cultural understandings. The examples can be pushed much further by incorporating anthropological data outside of the Western tradition (Halperin, 1990; Greenberg, 1988). Yet even within the narrower context offered here, the differences between them are striking. The assumption in ancient Greece is that men (less is known about women) can respond erotically to either sex, and the vast majority of men who engaged in same-sex relationships were also married (or would later become married). Yet the contemporary understanding of homosexuality divides the sexual domain in two, heterosexual and homosexual, and most heterosexuals cannot respond erotically to their own sex.

In saying that sexuality is a social construct, these theorists are not saying that these understandings are not real. Since persons are also constructs of their culture (in this view), we are made into those categories. Hence today persons of course understand themselves as straight or gay (or perhaps bisexual), and it is very difficult to step outside of these categories, even once one comes to seem them as the historical constructs they are.

Gay and lesbian theory was thus faced with three significant problems, all of which involved difficulties with the notion of ‘identity.’ Queer theory thus arose in large part as an attempt to overcome them. How queer theory does so can be seen by looking at the term ‘queer’ itself. In contrast to gay or lesbian, ‘queer,’ it is argued, does not refer to an essence, whether of a sexual nature or not. Instead it is purely relational, standing as an undefined term that gets its meaning precisely by being that which is outside of the norm, however that norm itself may be defined. As one of the most articulate queer theorists puts it: “Queer is … *whatever* is at odds with the normal, the legitimate, the dominant. *There is nothing in particular to which it necessarily refers*. It is an identity without an essence” (Halperin, 1995, 62,

original emphasis). By lacking any essence, queer does not marginalize those whose sexuality is outside of any gay or lesbian norm, such as sado-masochists. Since specific conceptualizations of sexuality are avoided, and hence not put at the center of any definition of queer, it allows more freedom for self-identification for, say, black lesbians to identify as much or more with their race (or any other trait, such as involvement in an S & M subculture) than with lesbianism. Finally, it incorporates the insights of poststructuralism about the difficulties in ascribing any essence or non-historical aspect to identity.

This central move by queer theorists, the claim that the categories through which identity is understood are all social constructs rather than given to us by nature, opens up a number of analytical possibilities. For example, queer theorists examine how fundamental notions of gender and sex which seem so natural and self-evident to persons in the modern West are in fact constructed and reinforced through everyday actions, and that this occurs in ways that privilege heterosexuality (Butler, 1990, 1993). Also examined are medical categories which are themselves socially constructed (Fausto-Sterling, 2000, is an erudite example of this, although she is not ultimately a queer theorist). Others examine how language and especially divisions between what is said and what is not said, corresponding to the dichotomy between ‘closeted’ and ‘out,’ especially in regards to the modern division of heterosexual/homosexual, structure much of modern thought. That is, it is argued that when we look at dichotomies such as natural/artificial, or masculine/feminine, we find in the background an implicit reliance upon a very recent, and arbitrary, understanding of the sexual world as split into two species (Sedgwick, 1990).

Another critical perspective opened up by a queer approach, although certainly implicit in those just referred to, is especially important. Since most anti-gay and lesbian arguments rely upon the alleged naturalness of heterosexuality, queer theorists attempt to show how these categories are themselves deeply social constructs. An example helps to illustrate the approach. In an essay against gay marriage, chosen because it is very representative, James Q. Wilson (1996) contends that gay men have a “great tendency” to be promiscuous. In contrast, he puts forward loving, monogamous marriage as the natural condition of heterosexuality. Heterosexuality, in his argument, is an odd combination of something completely natural yet simultaneously endangered. One is born straight, yet this natural condition can be subverted by such things as the presence of gay couples, gay teachers, or even excessive talk about homosexuality. Wilson's argument requires a radical disjunction between heterosexuality and homosexuality. If gayness is radically different, it is legitimate to suppress it. Wilson has the courage to be forthright about this element of his argument; he comes out against “the political imposition of tolerance” towards gays and lesbians (Wilson, 1996, 35).

It is a common move in queer theory to bracket, at least temporarily, issues of truth and falsity (Halperin, 1995). Instead, the analysis focuses upon the social function of discourse. Questions of who counts as an expert and why, and concerns about the effects of the expert's discourse are given equal status to questions of the verity of what is said. This approach reveals that hidden underneath Wilson's (and other anti-gay) work is an important epistemological move. Since heterosexuality is the natural condition, it is a place that is spoken from but not inquired into. In contrast, homosexuality is the aberration and hence it needs to be studied but it is not an authoritative place from which one can speak. By virtue of this heterosexual privilege, Wilson is allowed the voice of the impartial, fair-minded expert. Yet, as the history section above shows, there are striking discontinuities in understandings of sexuality, and this is

true to the point that, according to queer theorists, we should not think of sexuality as having any particular nature at all. Through undoing our infatuation with any specific conception of sexuality, the queer theorist opens space for marginalized forms.

Queer theory, however, has been criticized in a myriad of ways (Jagose, 1996). One set of criticisms comes from theorists who are sympathetic to gay liberation conceived as a project of radical social change. An initial criticism is that precisely because ‘queer’ does not refer to any specific sexual status or gender object choice, for example Halperin (1995) allows that straight persons may be ‘queer,’ it robs gays and lesbians of the distinctiveness of what makes them marginal. It desexualizes identity, when the issue is precisely about a sexual identity (Jagose, 1996). A related criticism is that queer theory, since it refuses any essence or reference to standard ideas of normality, cannot make crucial distinctions. For example, queer theorists usually argue that one of the advantages of the term ‘queer’ is that it thereby includes transsexuals, sado-masochists, and other marginalized sexualities. How far does this extend? Is transgenerational sex (e.g., pedophilia) permissible? Are there any limits upon the forms of acceptable sado-masochism or fetishism? While some queer theorists specifically disallow pedophilia, it is an open question whether the theory has the resources to support such a distinction. Furthermore, some queer theorists overtly refuse to rule out pedophiles as ‘queer’ (Halperin, 1995, 62) Another criticism is that queer theory, in part because it typically has recourse to a very technical jargon, is written by a narrow elite for that narrow elite. It is therefore class biased and also, in practice, only really referred to at universities and colleges (Malinowitz, 1993).

Queer theory is also criticized by those who reject the desirability of radical social change. For example, centrist and conservative gays and lesbians have criticized a queer approach by arguing that it will be “disastrously counter-productive” (Bawer, 1996, xii). If ‘queer’ keeps its connotation of something perverse and at odds with mainstream society, which is precisely what most queer theorists want, it would seem to only validate the attacks upon gays and lesbians made by conservatives. Sullivan (1996) also criticizes queer theorists for relying upon Foucault's account of power, which he argues does not allow for meaningful resistance. It seems likely, however, that Sullivan's understanding of Foucault's notions of power and resistance are misguided.

Conclusion

The debates about homosexuality, in part because they often involve public policy and legal issues, tend to be sharply polarized. Those most concerned with homosexuality, positively or negatively, are also those most engaged, with natural law theorists arguing for gays and lesbians having a reduced legal status, and queer theorists engaged in critique and deconstruction of what they see as a heterosexist regime. Yet the two do not talk much to one another, but rather ignore or talk past one another. There are some theorists in the middle. For example, Michael Sandel takes an Aristotelian approach from which he argues that gay and lesbian relationships can realize the same goods that heterosexual relationships do (Sandel, 1995). He largely shares the account of important human goods that natural law theorists have, yet in his evaluation of the worth of same-sex relationships, he is clearly sympathetic to gay and lesbian concerns. Similarly, Bruce Bawer (1993) and Andrew Sullivan (1995) have written eloquent defenses of

full legal equality for gays and lesbians, including marriage rights. Yet neither argue for any systematic reform of broader American culture or politics. In this they are essentially conservative. Therefore, rather unsurprisingly, these centrists are attacked from both sides. Sullivan, for example, has been criticized at length both by queer theorists (e.g., Phelan, 2001) and natural law theorists (e.g., George, 1999).

Yet as the foregoing also clearly shows, the policy and legal debates surrounding homosexuality involve fundamental issues of morality and justice. Perhaps most centrally of all, they cut to issues of personal identity and self-definition. Hence there is another, and even deeper, set of reasons for the polarization that marks these debates.

Bibliography

- Bawer, Bruce, 1993, *A Place at the Table: The Gay Individual in American Society*. New York: Poseidon Press.
- -----, 1996. *Beyond Queer: Challenging Gay Left Orthodoxy*. New York: The Free Press.
- Berman, Paul, 1993, “Democracy and Homosexuality” in *The New Republic*. Vol.209, No.25 (December 20): pp.17-35.
- Boswell, John, 1980, *Christianity, Social Tolerance, and Homosexuality: Gay People in Western Europe from the Beginning of the Christian Era to the Fourteenth Century*. Chicago: The University of Chicago Press.
- -----, 1994, *Same-Sex Unions in Premodern Europe*. New York: Vintage Books.
- Bradley, Gerard V., 2001, “The End of Marriage” in *Marriage and the Common Good*. Ed. by Kenneth D. Whitehead. South Bend, IN: St. Augustine's Press.
- Butler, Judith, 1990, *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.
- -----, 1993, *Bodies That Matter: On the Discursive Limits of “Sex”*. New York: Routledge.
- Cicero, 1966, *Tusculan Disputations*. Cambridge, MA: Harvard University Press.
- Dover, K.J., 1978, 1989, *Greek Homosexuality*. Cambridge, MA: Harvard University Press.
- Faderman, Lillian, 1985, *Surpassing the Love of Men: Romantic Friendship and Love Between Women from the Renaissance to the Present*. London: The Women's Press.
- Fausto-Sterling, Anne, 2000, *Sexing the Body: Gender Politics and the Construction of Sexuality*. New York: Basic Books.
- Finnis, John, 1994, “Law, Morality, and ‘Sexual Orientation’” *Notre Dame Law Review* 69: 1049-1076.
- Foucault, Michel, 1980, *The History of Sexuality. Volume One: An Introduction*. Translated by Robert Hurley. New York: Vintage Books.
- -----,1985, *The History of Sexuality. Volume Two: The Use of Pleasure*. New York: Pantheon Books.
- -----, 1986, *The History of Sexuality. Volume Three: The Care of the Self*. New York: Pantheon.
- George, Robert P., 1999, *In Defense of Natural Law*. New York: Oxford University Press.
- -----, 2001, “‘Same-Sex Marriage’ and ‘Moral Neutrality’” in *Marriage and the Common Good*. Ed. by Kenneth D. Whitehead. South Bend, IN: St. Augustine's Press.

- Greenberg, David F., 1988, *The Construction of Homosexuality*. Chicago: The University of Chicago Press.
- Halperin, David M., 1990, *One Hundred Years of Homosexuality: and other essays on Greek love*. New York: Routledge.
- -----, 1995, *Saint Foucault: Towards a Gay Hagiography*. New York: Oxford University Press.
- Jagose, Annamarie, 1996, *Queer Theory: An Introduction*. New York: New York University Press.
- Macedo, Stephen, 1995, “Homosexuality and the Conservative Mind” *Georgetown Law Journal* 84: 261-300.
- Malinowitz, Harriet, 1993, “Queer Theory: Whose Theory?” *Frontiers*, Vol.13: 168-184.
- Nussbaum, Martha, 1999, *Sex and Social Justice*. New York: Oxford University Press.
- Phelan, Shane, 2001, *Sexual Strangers: Gays, Lesbians, and Dilemmas of Citizenship*. Philadelphia: Temple University Press.
- Plato, 1981, *The Symposium*. Translated by Walter Hamilton. New York: Penguin Books.
- Plato, 1970, *The Laws*. Translated by Trevor Saunders. New York: Penguin Books.
- Rich, Adrienne, 1980, “Compulsory Heterosexuality and Lesbian Existence” in *Women, Sex, and Sexuality*. Edited by Catharine Stimpson and Ethel Spector Person. Chicago: University of Chicago Press.
- Sandel, Michael J., 1995, “Moral Argument and Liberal Toleration: Abortion and Homosexuality” in *New Communitarian Thinking: Persons, Virtues, Institutions, and Communities*. Edited by Amitai Etzioni. Charlottesville: University Press of Virginia.
- Sedgwick, Eve Kosofsky, 1990, *Epistemology of the Closet*. Berkeley: University of California Press.
- Shilts, Randy, 1993, *Conduct Unbecoming: Gays and Lesbians in the U.S. Military*. New York: St. Martin's Press.
- Spargo, Tasmin, 1999, *Foucault and Queer Theory*. New York: Totem Books.
- Sullivan, Andrew, 1995, *Virtually Normal: An Argument about Homosexuality*. New York: Knopf.
- Weitham, Paul J., 1997, “Natural Law, Morality, and Sexual Complementarity” in *Sex, Preference, and Family: Essay on Law and Nature*. Edited by David M. Estlund and Martha C. Nussbaum. New York: Oxford University Press.
- Wilson, James Q., 1996, “Against Homosexual Marriage” *Commentary*, Vol.101, No.3 (March): 34-39.

Other Internet Resources

- "Disorder and Diversity,", by Alan Soble (Philosophy, University of New Orleans). This was the keynote lecture to a University of New Orleans Colloquium titled "Towards 2002: Gender Diversity Now and Into the Next Century".
- Lesbian, Gay, Bisexual and Transgendered Philosophy Web Site

[Please contact the author with further suggestions.]

Related Entries

Aquinas, Saint Thomas | Ethics, natural law tradition | Feminism, topics: Feminist Perspectives onSexuality; Feminist Perspectives on the Self | Foucault, Michel

Brent Pickett
bpickett@csc.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 6, 2002
Content last modified: August 6, 2002

Pyrrho

Pyrrho was the starting-point for a philosophical movement known as Pyrrhonism that flourished several centuries after his own time. This later Pyrrhonism was one of the two major traditions of sceptical thought in the Greco-Roman world (the other being located in Plato's Academy during much of the Hellenistic period). Perhaps the central question about Pyrrho is whether or to what extent he himself was a sceptic in the later Pyrrhonist mold. The later Pyrrhonists claimed inspiration from him; and, as we shall see, there is undeniably some basis for this. But it does not follow that Pyrrho's philosophy was identical to that of this later movement, or even that the later Pyrrhonists thought that it was identical; the claims of indebtedness that are expressed by or attributed to members of the later Pyrrhonist tradition are broad and general in character (and in Sextus Empiricus' case notably cautious -- see *Outlines of Pyrrhonism* 1.7), and do not in themselves point to any particular reconstruction of Pyrrho's thought. It is necessary, therefore, to focus on the meager evidence bearing explicitly upon Pyrrho's own ideas and attitudes. How we read this evidence will also, of course, affect our conception of Pyrrho's relations with his philosophical contemporaries and predecessors.

1. Life

Pyrrho appears to have lived from around 365-360 BC until around 275-270 BC (for the evidence see von Fritz (1963), 90). We have several reports of philosophers from whom he learned, the most significant

(and the most reliable) of which concern his association with Anaxarchus of Abdera. Alongside Anaxarchus (and several other philosophers) he accompanied Alexander the Great on his expedition to India. We are told that in the course of this expedition he encountered some “naked wise men” (*gumnosophistai*); Diogenes Laertius (9.61) claims that his philosophy developed as a result of this meeting, but it is not clear what basis, if any, he has for this assertion. In any case, after his return to Greece Pyrrho did espouse a philosophy that attracted numerous followers, of whom the most important was Timon of Phlius. The travel writer Pausanias reports (6.24.5) seeing a statue of him in the marketplace in Pyrrho's home town of Elis; Diogenes also reports (9.64) that he was made high priest, and that in his honor philosophers were made exempt from city taxation. While Pausanias' report is easier to accept at face value than Diogenes', both suggest that (at least locally) he achieved considerable celebrity. However, there is no good reason to believe that a Pyrrhonist ‘movement’ continued beyond his own immediate followers; it was not until the first century BC that Pyrrhonism became the name of an ongoing philosophical tradition.

2. The Nature of the Evidence

With the possible exception of a poem honoring Alexander, Pyrrho wrote nothing; we are therefore obliged to try to reconstruct his philosophy from reports by others. Unfortunately the reports that we have are fragmentary and, in many cases, of doubtful reliability. Pyrrho's follower Timon wrote numerous poems and prose works; but of these only fragments and in some cases second-hand reports survive. It is likely that much of what we hear about Pyrrho in later sources also derives indirectly from Timon, but it is frequently not possible to judge, in individual cases, whether or not this is so. Other difficulties concerning Timon's evidence are, first, that he writes as a devotee rather than as a neutral reporter, and second, that the great majority of Timon's fragments derive from a poem called *Silloi* (*Lampoons*), and consist of satirical thumbnail sketches of other philosophers rather than any kind of direct exposition of Pyrrho's outlook. Nevertheless, Timon is clearly the most important and the most trustworthy source of information about Pyrrho. Of the evidence from Timon, one piece is especially important. This is a summary, by the Peripatetic Aristocles of Messene (late 1st c. BC), of an account by Timon of Pyrrho's most general philosophical attitudes; it is generally agreed that this passage must be the centerpiece of any interpretation of Pyrrho's philosophy.

With the exception of a very few snippets of information, the only other source of evidence on Pyrrho that is close to contemporary is Antigonus of Carystus, a biographer of the mid-third century BC. Diogenes Laertius, whose life of Pyrrho is our major source of biographical anecdotes, frequently cites Antigonus, and is probably indebted to him in many places even when he does not cite him; Antigonus is also mentioned by Aristocles, and is likely to be the origin of much or even most of the surviving biographical material on Pyrrho. But Antigonus needs to be handled even more carefully than Timon. He was not a philosopher, and there is no reason to think that he was capable of or particularly interested in comprehending philosophical nuances. Nor is it clear that accuracy was high on his list of priorities; on the contrary, what we know about him suggests that he was a purveyor of sensationalist gossip rather than a reliable historian. His account probably did contain genuine information about Pyrrho's demeanor and activities; but it would certainly be unwise to take his anecdotes as a group on trust.

Cicero refers to Pyrrho about ten times, and Cicero is, in general, a responsible reporter of other people's views. Unfortunately his information about Pyrrho appears to be very incomplete, and, as we shall see, the impression he conveys of him may very well be inaccurate. With one exception, he never mentions him except in conjunction with the Stoics Aristo and (sometimes) Herillus, and the remarks in question always concern the same one or two points in ethics; it looks as if his knowledge of Pyrrho derives almost entirely from a single source that conveyed little or nothing about Pyrrho individually, and nothing at all about any views of his in metaphysics or epistemology. In addition, a number of authors in later antiquity make isolated comments about Pyrrho. But since these comments postdate the rise of the later Pyrrhonist movement, there is often room for suspicion as to whether they reflect genuine information about Pyrrho himself, as opposed to the later philosophy that took his name.

The safest way to proceed, then, in attempting to interpret Pyrrho's philosophy, is to begin with a detailed analysis of the summary of Timon in Aristocles; the remaining evidence should be interpreted, and assessed for its trustworthiness, as far as possible in light of what the Aristocles passage has to tell us about Pyrrho's philosophy.

3. The Aristocles Passage

The main burden of the previous section is uncontroversial. By contrast, the interpretation of the Aristocles passage itself is fraught with controversy. Given the admitted centrality of this passage in any interpretation of Pyrrho, it follows that the character of Pyrrho's thinking is, at its very core, a matter for sharp disagreement. The present survey will attempt to lay out the two main alternatives and explore the consequences of each.

A word, first, about the provenance and evidential value of the passage. Several chapters from book 8 of Aristocles' *Peri philosophias* (*On Philosophy*) appear in quotation in the *Praeparatio evangelica* of Eusebius, the fourth-century bishop of Caesarea. These chapters expound, and then attack from an Aristotelian perspective, a number of philosophies that impugn the reliability of one or other of our cognitive faculties. Among these Aristocles counts Pyrrhonism, as represented by Pyrrho himself and by the initiator of the later Pyrrhonist tradition, Aenesidemus; Aristocles speaks of Aenesidemus as recent, and is apparently not aware of any subsequent members of the Pyrrhonist tradition. The chapter on Pyrrhonism opens with a brief summary of the outlook of Pyrrho specifically, as reported by Timon (14.18.1-5).

This convoluted transmission might lead one to doubt how much authentic information can be extracted from the passage. But there is no reason, first, to doubt Eusebius' explicit claim to be quoting Aristocles verbatim. Aristocles is not quoting Timon verbatim; some of the vocabulary in this passage is clearly reminiscent of Aristocles' own vocabulary in other chapters. However, other terms in the passage are quite distinct both from Aristocles' own normal usage and from any of the terminology familiar to us from later Pyrrhonism, yet were in use prior to Timon's day; it seems easiest to account for these as authentic reproductions of Timon's own language. In addition, the frequent mention of Timon either by

name or by the word *phêsi*, ‘he says’, suggests that Aristocles is taking pains to reproduce the essentials of his account as faithfully as possible. And finally, Aristocles' other chapters, where we are in a position to check his summaries against other evidence of the views being summarized, suggest that he is in general a reliable reporter of other people's views, even views to which he himself is strongly opposed -- or at least, that he is reliable when he has access to systematic written expositions of those views, as seems to be the case here. As for Timon's reliability as a source for Pyrrho's views, a word was said about this in the previous section; this passage stands out from most of the surviving record from Timon in that it appears to be based on a systematic account of the central points of Pyrrho's outlook. Thus, while we clearly cannot be sure that every detail of the passage accurately reflects something that Pyrrho thought, all the indications are that the passage deserves to be taken seriously as evidence for Pyrrho's philosophy.

Timon is reported as telling us that in order to be happy, one must pay attention to three connected questions: first, what are things like by nature? second, how should we be disposed towards things (given our answer to the first question)? and third, what will be the outcome for those who adopt the disposition recommended in the answer to the second question? And the passage then gives us Pyrrho's and Timon's answers to each of the three questions in order.

The answer to the question “what are things like by nature?” is given by a sequence of three epithets; things are said to be *adiaphora* and *astathmêta* and *anepikrita*. Taken by themselves, these epithets can be understood in two importantly different ways: they may be taken as characterizing how things are (by nature) in themselves, or they may be taken as commenting on human beings' lack (by our nature) of cognitive access to things. *Adiaphora* is normally translated ‘indifferent’. But this might be taken as referring either to an intrinsic characteristic of things -- namely that, in themselves and by nature, they possess no differentiating features -- or to our natural inability to discern any such features. (In the latter case ‘undifferentiable’ might be a more perspicuous translation.) Similarly, *astathmêta* might mean ‘unstable’ or ‘unbalanced’, describing an objective property of things; or it might mean ‘not subject to being placed on a balance’, and hence ‘unmeasurable’, which would again place the focus on our cognitive inabilities. And *anepikrita* might mean ‘indeterminate’, referring to an objective lack of any definite features, or ‘indeterminable’, pointing to an inability on our part to determine the features of things. The statement as a whole, then, is either answering the question “what are things like by nature?” by stating that things are, in their very nature, indefinite or indeterminate in various ways -- the precise nature of the thesis would be a matter for further speculation -- or by stating that we human beings are not in a position to pin down or determine the nature of things. Let us call these the metaphysical and the epistemological interpretations respectively.

It is clear that the metaphysical interpretation gives us a Pyrrho who is not in any recognizable sense a sceptic. Pyrrho, on this interpretation, is issuing a declaration about the nature of things in themselves -- precisely what the later Pyrrhonists who called themselves sceptics were careful to avoid. On the epistemological interpretation, on the other hand, Pyrrho is very much closer to the tradition that took his name. There is still some distance between them. To say that we *cannot* determine the nature of things -- as opposed to saying that we have so far failed to determine the natural features of things -- is already a departure from the sceptical suspension of judgement promoted by Sextus Empiricus. And to put it, as on this reading Pyrrho does put it, by saying that *things* are indeterminable, is a further departure, in that it

does attribute at least one feature to things in themselves -- namely, being such that humans cannot determine them. Nevertheless, the epistemological interpretation clearly portrays Pyrrho as a forerunner -- a naive and unsophisticated forerunner, perhaps -- of later Pyrrhonist scepticism, whereas the metaphysical interpretation puts him in a substantially different light.

The natural way to try to choose between these two interpretations is to see which of them fits best with the logic of the passage as a whole. But here we encounter a further complication. The text of the phrase that follows the words we have just been discussing is subject to dispute. According to the manuscripts this phrase reads “*for this reason* (*dia touto*) neither our sensations nor our opinions tell the truth or lie”. Now, if this reading is correct, the thought expressed seems to favor the metaphysical interpretation; the idea is that, since things are in their real nature indeterminate, our sensations and opinions, which represent things as having certain determinate features, are neither true nor false. They are not true, since reality is not the way they present it as being; but they are not false either, since that too would require that reality have certain determinate features, namely features that are the negations of the ones that our sensations and opinions portray it as having. By contrast, it is quite unclear how the claim that the nature of things cannot be determined by us could be thought to license the inference “for this reason neither our sensations nor our opinions tell the truth or lie”. That claim would seem at most to license the very different inference that *we* cannot tell whether they tell the truth or lie.

But some scholars have suggested that the manuscripts are in error at this point, and that what the text should say is “*on account of the fact that* (*dia to*) neither our sensations nor our opinions tell the truth or lie”. The change is justified on linguistic grounds; it is alleged that the text in the manuscripts as they stand is not acceptable Greek. The considerations for and against this proposal are technical, and debate has yielded no consensus on this question. However, it is clear that if one does make this small alteration to the text, the direction of the inference is reversed; the point about our sensations and opinions now becomes a reason for the point about the nature of things, not an inference from it. And this, it has been argued, points towards the epistemological interpretation. The idea would then be that, since our sensations and opinions fail to be consistent deliverers of true reports (or, for that matter, false reports) about the world around us, there is no prospect of our being able to determine the nature of things.

This is the crux on which the decision between the two main lines of interpretation of Pyrrho's philosophy turns. The remainder of the Aristocles passage, and indeed the remainder of the evidence on Pyrrho in general, can be read so as to fit with either the metaphysical or the epistemological reading of his answer to the question about the nature of things. The Aristocles passage continues with the answer to the second question, namely the question of the attitude we should adopt given the answer to the first question. We are told, first, that we should not trust our sensations and opinions, but should adopt an unopinionated attitude. On the epistemological reading, the significance of this is obvious. But on the metaphysical reading, too, we have already been told that our sensations and opinions are not true, which is presumably reason enough for us not to trust them; and the unopinionated attitude that is here recommended may be understood as one in which one refrains from positing any definite characteristics as inherent in the nature of things -- given that their real nature is wholly indefinite. (To the objection that this thesis of indefiniteness is itself an opinion, it may be replied that *doxa*, ‘opinion’, is regularly used in earlier Greek philosophy, especially in Parmenides and Plato, to refer to those opinions -- misguided opinions, in the

view of these authors -- that take on trust a view of the world as conforming more or less to the way it appears in ordinary experience. In this usage, the claim that reality is indefinite would not be a (mere) opinion, but would be a statement of the truth.)

The passage now introduces a certain form of speech that is supposed to reflect this unopinionated attitude. We are supposed to say "about each single thing that it no more is than is not or both is and is not or neither is nor is not". There are a number of intricate questions about the exact relations between the various parts of this complicated utterance, and especially about the role and significance of the 'both' and 'neither' components. But it is clear that this too is susceptible of being read along the lines of either of the two interpretations introduced above. On the metaphysical interpretation, we are being asked to adopt a form of words that reflects the utter indefiniteness of the way things are; we should not say of anything that it *is* any particular way any more than that it is *not* that way (with 'is' being understood, as commonly in Greek philosophy, as shorthand for 'is *F*', where *F* stands for any arbitrary predicate). On the epistemological interpretation, we are being asked to use a manner of speaking that expresses our suspension of judgement about how things are. Sextus Empiricus specifically tells us that 'no more' (*ou mallon*) is a term used by the sceptics to express suspension of judgement, and its occurrence in the Aristocles passage can be taken as an early example of the same type of usage.

Finally, in answer to the third question, we are told that the result for those who adopt the unopinionated attitude just recommended is first *aphasia* and then *ataraxia*. *Ataraxia*, 'freedom from worry', is familiar to us from later Pyrrhonism; this is said by the later Pyrrhonists to be the result of the suspension of judgement that they claimed to be able to induce. The precise sense of *aphasia* is less clear. It might mean 'non-assertion', as in Sextus -- that is, a refusal to commit oneself to definite alternatives; or it might mean, more literally, 'speechlessness', which could in turn be taken to be an initial reaction of stunned silence to the radical position with which one has been presented (an uncomfortable reaction that is subsequently replaced by *ataraxia* -- the passage does say that *aphasia* comes first and *ataraxia* comes later). But the decision between these two ways of understanding the term is independent of the broader interpretive issues bearing upon the passage as a whole. For some form of 'non-assertion' is clearly licensed by either the metaphysical or the epistemological interpretation; and on either interpretation, the view proposed might indeed render someone (initially) uncomfortable to the point of 'speechlessness'. The important point, though, is that *ataraxia* is the end result; and this links back to the introductory remark to the effect that the train of thought to be summarized has the effect of making one happy.

We have, then, two major possibilities. On the one hand, Pyrrho can be read as advancing a sweeping metaphysical thesis, that things are in their real nature indefinite or indeterminate, and encouraging us to embrace the consequences of that thesis by refusing to attribute any definite features to things (at least, as belonging to their real nature) and by refusing to accept at face value (again, as revelatory of the real nature of things) those myriad aspects of our ordinary experience that represent things as having certain definite features. Or, on the other hand, Pyrrho can be read as declaring that the nature of things is inaccessible to us, and encouraging us to withdraw our trust (and to speak in such a way as to express our withdrawal of trust) in ordinary experience as a guide to the nature of things. As noted earlier, the second, epistemological interpretation makes Pyrrho's outlook a great deal closer to that of the later Pyrrhonists who took him as an inspiration. But that is not in itself any reason for favoring this interpretation over the

other, metaphysical one. For on either interpretation Pyrrho is said to promise *ataraxia*, the later Pyrrhonists' goal, and to promise it as a *result* of a certain kind of withdrawal of trust in the veracity of our everyday impressions of things; the connection between these two points aligns Pyrrho with the later Pyrrhonists, and sets him apart from every other Greek philosophical movement that preceded later Pyrrhonism. The fact that this later sceptical tradition took Pyrrho as an inspiration is therefore readily understandable whichever of the two interpretations is correct (or whichever they thought was correct). It is also true that, on the metaphysical interpretation of the passage, the grounds on which Pyrrho advanced his metaphysical thesis of indeterminacy are never specified; this too, like the precise character of the thesis itself, must be a matter for speculative reconstruction. But Aristocles only purports to be giving the key points of Timon's summary; the lack of detail, though disappointing, would not be surprising.

4. Other Reports on Pyrrho's General Approach

The other evidence bearing upon Pyrrho's central philosophical attitudes does nothing to settle the dispute between these two possible interpretations of the Aristocles passage. A number of texts explicitly or implicitly represent Pyrrho's outlook as essentially identical to the sceptical outlook of the later Pyrrhonists. However, all of these texts postdate the rise of later Pyrrhonism itself, and none of them contains anything like the level of detail of the Aristocles passage; there is no particular reason to suppose that any of them represents a genuine understanding of Pyrrho as distinct from the later Pyrrhonists. There are also a few texts that appear to offer a picture of Pyrrho's thought that is congenial to the metaphysical interpretation. The most significant of these is a passage from near the beginning of Diogenes Laertius' life of Pyrrho (9.61); Pyrrho, we are told, “said that nothing is either fine or ignoble or just or unjust; and similarly in all cases that nothing is the case in reality (*mêden einai têi alêtheiai*), but that human beings do everything by convention and habit; for each thing is no more this than that”. This looks as if it is attributing to Pyrrho a metaphysical thesis, and a thesis according to which things have no definite features; it does not seem to be portraying Pyrrho as an advocate of suspension of judgement. However, the passage is undoubtedly in some way confused. For immediately beforehand we are told that Pyrrho was the one who initiated the type of philosophy consisting of “inapprehensibility and suspension of judgement”; and the previously quoted passage is then cited as confirmation of this. This passage of Diogenes as a whole, then, fails to maintain any clear distinction between epistemological and metaphysical readings of Pyrrho's thought, and cannot safely be used as evidence for either.

5. Reports on Pyrrho's Demeanor and Lifestyle

Most of the remaining evidence about Pyrrho has to do with his practical attitudes and behavior. There are a number of biographical anecdotes, and there are a few fragments of Timon that purport to depict Pyrrho's state of mind. Some of the biographical anecdotes are clearly polemical inventions. For example, Diogenes (9.62) reports Antigonus as saying that Pyrrho's lack of trust in his senses led him to ignore precipices, oncoming wagons and dangerous dogs, and that his friends had to follow him around to protect him from these various everyday hazards. But he then reports the dissenting verdict of Aenesidemus, according to which Pyrrho was perfectly capable of conducting himself in a sensible

manner. This reflects a longstanding ancient dispute as to whether it is possible to live if one radically abandons common-sense attitudes to the world (as, on either of the two major interpretations outlined above, Pyrrho in some sense did). Antigonus has transformed a hostile criticism of Pyrrho -- that, if one really were to adopt the attitudes he recommends, one *would* be unable consistently to live as a sane human being -- into an account of how Pyrrho actually did act; but there is no reason to take this seriously as biography. (It does, however, raise the question of what it means to mistrust the senses -- as the Aristocles passage tells us, on either interpretation, that Pyrrho did -- if this is to be consistent with self-preservation in one's ordinary behavior; we shall return to this point at the end of this section.)

But there are many other anecdotes, preserved in Diogenes and elsewhere, that cannot be dismissed in this fashion. The dominant impression that they convey as a group is of an extraordinary impassivity or imperturbability; with rare exceptions (which he himself is portrayed as regretting), Pyrrho is depicted as maintaining his calm and untroubled attitude no matter what happens to him. This extends even to extreme physical pain -- he is reported not to have flinched when subjected to the horrific techniques of ancient surgery -- but it also encompasses dangers such as being on a ship in a storm. (This is not to say that he did not avoid such troubles if he could, as suggested by the apocryphal stories mentioned in the previous paragraph; it is just to say that he did not lose his composure in the face of life's inevitable hardships.) There is another aspect to this untroubled attitude as well. In numerous anecdotes Pyrrho is shown as unconcerned with adhering to the normal conventions of society; he wanders off for days on end by himself, and he performs tasks that would normally be left to social inferiors, such as housework and even washing a pig. Here, too, the suggestion is that he does not care about things that ordinary people do care about -- in this case, the disapproval of others. (The passage from Diogenes quoted in the previous section, according to which Pyrrho held “that human beings do everything by convention and habit” is not necessarily in conflict with this; by ‘human beings’ Pyrrho might have meant *ordinary* human beings, among whom he would not have included himself.)

The fragments of Timon also emphasize Pyrrho's exceptional tranquillity, and add a further, more philosophical dimension to it. Several fragments suggest that this tranquillity results from his not engaging in theoretical inquiry like other philosophers, and with his not engaging in debate with those philosophers. Other thinkers are perturbed by their need to discover how the universe works, and to prevail in arguments with their rivals; Pyrrho is unconcerned about any of this.

Clearly it would be foolish to accept every detail of this composite account. Timon's picture is no doubt idealized, and the biographical material surely includes a measure of embellishment. Collectively, however, these fragments and anecdotes add up to a highly consistent portrait; it does not seem overconfident to take this at least as reflecting an ideal towards which Pyrrho strived, and which he achieved to a sufficient degree to have attracted notice.

What connections can be drawn between this portrait of Pyrrho's demeanor and the philosophy expounded in the Aristocles passage? What we have here, plainly, is a fuller specification of the *ataraxia* that the Aristocles passage promised as the outcome of the process of responding in the recommended way to the three questions. But *why* should this process be thought to yield *ataraxia*, and why should the outcome take the specific form suggested by the material discussed in this section? Now is a convenient

point to address these questions (as we did not do in initially examining the Aristocles passage). We may distinguish the overall emotional tranquillity depicted in the biographical material and the particular tranquillity, derived from the avoidance of theoretical inquiry, that is emphasized in the fragments of Timon.

The most obvious explanation for the overall emotional tranquillity would seem to be along the following lines. If one adopts the position recommended in the answers to the Aristocles passage's first two questions, one will not hold any definite beliefs, about any object or state of affairs, to the effect that the object or state of affairs really is good, or valuable, or worth seeking -- or, on the other hand, bad or to be avoided. (This will be true on either the metaphysical or the epistemological interpretations of the Aristocles passage.) Hence one will attach far less importance to the attainment or the avoidance of any particular objects, or the occurrence or non-occurrence of any particular state of affairs, than one would if one thought that these things really did have some kind of positive or negative value; nothing will matter to one to anything like the same extent as it matters to most people. If this is on the right lines, then Pyrrho's route to *ataraxia* closely resembles the one described in several places by Sextus Empiricus (*PH* 1.25-30, 3.235-8, *M* 11.110-67); the account I have given is in fact modeled after Sextus' explanation of how suspension of judgement produces *ataraxia*. Indeed, it is difficult to see how else the process is to be reconstructed. One difference, however, between the practical attitudes of Pyrrho and Sextus is that Sextus attributes an important role to convention in the shaping of one's behavior; more on this in a moment.

As for the avoidance of theoretical inquiry and debate, it is fair to assume that Pyrrho will have regarded such inquiry and debate as troublesome because it is necessarily fruitless and interminable. If the real nature of things is indeterminable by us, as the epistemological interpretation would have it, then to attempt to determine the nature of things, and to provide cogent grounds for the superiority of one's own theories, is to attempt the impossible; such matters are simply beyond our grasp. But if, as the metaphysical interpretation would have it, the world is in its real nature indefinite, then one is also attempting the impossible, though for a different reason; one is attempting to determine a fixed character for things that inherently lack any fixed character. Of course, the thesis that reality is indefinite is itself a definite statement. But it follows from this statement that any attempt to establish fixed and definite characteristics for specific items in the universe is doomed to failure. And it seems to be this kind of theoretical inquiry that Timon represents Pyrrho as eschewing; as one fragment puts it (addressing Pyrrho), “you were not concerned to inquire what winds hold sway over Greece, from where everything comes and into what it passes” (Diogenes Laertius 9.65). On this reading, then, Pyrrho's avoidance of inquiry and debate resembles the disdain for physical speculation that is apparent in some writings of Plato (the *Phaedo* for example); physical speculation is fruitless because the physical world is not susceptible to rational inquiry.

Here another difference may be discerned from the Pyrrhonism of Sextus Empiricus. For although Sextus certainly does not claim to have definite answers of his own concerning the nature of things, he has no qualms about engaging with his opponents in debate, or about pitting them against one another. Indeed, the Pyrrhonism of Sextus depends on a constant interplay of competing arguments on as many topics as possible. One's suspension of judgement results from one's experience of the ‘equal strength’

(*isostheneia*) of the competing considerations on all sides of a given issue; but this requires that one be regularly exposed to such competing considerations, and the works of Sextus are themselves bountiful sources for this. Sextus may agree with Pyrrho that such debates are interminable, but they have an important role in later Pyrrhonist practice nonetheless. Pyrrho's own practice is quite different, because his *ataraxia* flows not from an ongoing practice of intellectual juggling, but from a *conclusion* about the nature of things, as reported in the Aristocles passage -- either that it is indeterminable by us or that it is indefinite -- that he has arrived at once and for all.

We have seen that Pyrrho was unconcerned, or aspired to be unconcerned, about things that most of us care about very deeply. But how does one make any decisions at all, if one adopts this kind of attitude? Unless we are to believe the stories of Pyrrho as a madman depending on his friends to rescue him from precipices and dogs, we are entitled to expect an answer to this question. There is some evidence to suggest that the answer proposed -- if not by Pyrrho himself, then at least by Timon -- was that one relies on the appearances. If this is correct, we have another link, at least at a general level, with the Pyrrhonism of Sextus, for whom appearances are what he calls the “criterion of action”. One difference, as mentioned earlier, is that Sextus lists laws and customs as one of the four main varieties of appearances that one may use to guide one's behavior; on Sextus' account, then, certain courses of action will appear to one as desirable or undesirable given that they are approved or disapproved of by the prevailing mores. But Pyrrho seems to have been thoroughly unconventional in some aspects of his behavior. It is impossible to say specifically what kinds of phenomena were included for him and Timon under the heading of ‘the appearances’ -- or whether they had anything very precise in mind here. Nevertheless, it looks as if the early or proto-Pyrrhonist answer to the question of how one acts and makes decisions is that one does so in light of the way things appear to one. On the epistemological interpretation, this will amount to “in light of the characteristics that things seem to have (but, for all we know, may not really have)”; on the metaphysical interpretation, it will amount to “in light of the temporary and contingent characteristics that things manifest on any given occasion (but that are no part of how they *really are*, since how they really are is indefinite)”. In either case, the way in which one mistrusts sensations and common-sense opinions, as the Aristocles passage recommends, is not that one pays no attention to them in one's everyday behavior; one mistrusts them simply in that one does not take them as a guide to the underlying nature of things.

6. “The Nature of the Divine and the Good”

There is one further, highly problematic fragment of Timon that seems to belong in the area of practical philosophy. This is a set of four lines of verse that make reference to “the nature of the divine and the good”. There is no consensus on how to translate these lines, and different translations yield very different consequences for interpretation. The first two lines can be rendered either

> (A) “For I will say, as it appears to me to be,
> A word of truth, having a correct standard:”

or

(B) "For I will say, as it is plain to me that it is,
A word of truth, having a correct standard:"

The second two lines can be rendered either

(C) "That the nature of the divine and the good is eternal,
From which a most even-tempered life for a man is derived."

or

(D) "That the nature of the divine and the good is at any time
That from which life becomes most even-tempered for a man."

There are numerous other, less momentous disputes over the translation, but these alternatives put on display the central options for interpretation.

The lines are quoted by Sextus Empiricus (*M* 11.20); there is no further trace of them in the surviving record. The context in which Sextus introduces them shows that he is inclined to understand the first couplet according to reading (A) rather than reading(B); but he suggests that he is not sure about this. The speaker into whose mouth Timon put these lines is never identified, but it has generally been assumed to be Pyrrho.

If one understands the second couplet according to reading (C), then the speaker is apparently endorsing a position that attributes definite natures to things; and this appears flatly inconsistent with the Aristocles passage, on either the metaphysical or the epistemological reading. The tension is especially bad if one reads the first couplet according to reading (B), in which case the speaker is insisting that he is in possession of the truth. But even on reading(A), where the effect is to weaken the assertion to one about what appears to the speaker to be the case, the statement about "the nature of the divine and the good" seems strikingly out of keeping with the rest of what we hear about Pyrrho's philosophy. It has been suggested that Pyrrho made an exception, in the case of the divine and the good, to his general prohibition on attributing definite natures to things; but it is hard to see the motivation for such a move, and this interpretation has not been generally accepted. It is true that Cicero speaks of Pyrrho as holding that virtue is the sole good, and that no distinctions of value are to be drawn among things other than virtue and vice. However, as was mentioned earlier, Cicero always attributes this view to Pyrrho alongside the unorthodox Stoic Aristo of Chios; he never gives any details about Pyrrho's thinking specifically. We know from other sources that Aristo did hold this position; and, as we have seen, there is good reason to think that Pyrrho did refrain quite generally from attributing positive or negative value to ordinary objects of concern (things other than virtue and vice), which is one part of the position Cicero attributes to him and Aristo. So it looks as if Cicero has been misled (probably by the sketchiness of the information in his source) into thinking that Pyrrho agreed with Aristo in both parts of his position rather than in just one part. At any rate, Cicero cannot be regarded as offering any credible support for an interpretation of Pyrrho that has him believing in a natural good; again, this is just too discordant with the remainder of the evidence.

Reading (D) of Timon's second couplet (which is due, with minor modifications, to Burnyeat (1980)) is intended to eliminate the troublesome reference to an eternal real nature. According to this interpretation, the phrase “the nature of the divine and the good” refers simply to a characteristic that is attributed to Pyrrho, and labeled by poetic hyperbole as ‘divine’, in another fragment of Timon, namely his extraordinary tranquillity; the couplet as a whole, then, is saying that tranquillity is the source of an even-tempered life. And if one combines this with the less dogmatic reading (A) of the first couplet, this yields a set of remarks that are not obviously in conflict with anything else in the record on Pyrrho.

The acceptability of the translation in reading (D) is not beyond question. There is also some question whether the claim that tranquillity is the source of an even-tempered life is anything more than vacuous. However, if one assumes that Pyrrho is the speaker of these lines, then this interpretation or something close to it seems to be the only way to rescue his thought (as reported by Timon) from inconsistency. Another possibility is to drop the assumption that Pyrrho is the speaker; in this case, there is no reason to assume that the thought expressed by the lines must be consistent with what we know of Pyrrho's philosophy. But if one takes this option, one must devise an alternative explanation for why Timon would have written these lines -- including some account of who else the speaker might be. It is fair to say that no resolution of these matters is in sight; the fragment continues to be regarded as one of the most intractable pieces of evidence in the corpus of material on Pyrrho and Timon..

7. Influences on Pyrrho

Many different philosophical antecedents have been claimed for Pyrrho. Since we know very little about which philosophical currents Pyrrho may have been acquainted with, such claims are bound to be in large measure speculative. There are, however, a couple of exceptions to this; as noted at the outset, Pyrrho was associated with Anaxarchus and was reported to have encountered some unnamed Indian thinkers. The little that we know of Anaxarchus seems to suggest that his philosophy had a good deal in common with Pyrrho's. Diogenes Laertius (9.60) ascribes to him an attitude of *apatheia* and *eukolia*, ‘freedom from emotion’ and ‘contentedness’; *apatheia* is used in some sources to describe Pyrrho's attitude as well, and the combination of the two terms seems to describe something close to the state cultivated by Pyrrho. We also hear from Sextus Empiricus that Anaxarchus “likened existing things to stage-painting and took them to be similar to the things which strike us while asleep or insane” (*M* 7.88). This has often been taken as an early expression of a form of epistemological scepticism. But it may also be taken as an ontological comment on the insubstantiality of the world around us; it is *things* (as opposed to our impressions of things) that are assimilated to stage-sets and the contents of dreams and fantasies. Either way, the remark looks like an anticipation of Pyrrho's position in the Aristocles passage; the first reading conforms to the epistemological interpretation of that passage, and the second to the metaphysical interpretation. It appears, then, that Pyrrho may have borrowed to a considerable extent from Anaxarchus. There is, however, no indication that Anaxarchus drew a connection between his view of the nature of things and his attitude of emotional contentment, such as we have seen that Pyrrho did.

We do not know the identity of the “naked wise men” whom Pyrrho met in India, or what they thought. There are reports of other meetings between Indian and Greek thinkers during Alexander's expedition,

and these tend to emphasize the Indians' extraordinary impassivity and insensitivity to pain and hardship. It is not unlikely that Pyrrho, too, was impressed by traits of this kind. Though precedents for his ideal of *ataraxia* exist in earlier Greek philosophy as well, his reported ability to withstand surgery without flinching is exceptional in the Greek context (and quite distinct from anything in later Pyrrhonism); if we believe this story, it is tempting to explain it by way of some form of training from the Indians. Some scholars have sought to establish more detailed links between the thought of the Aristocles passage and various currents in ancient Indian philosophy. But it is not clear how far these similarities really go; and in any case, it is not easy to imagine an interchange of any great philosophical subtlety occurring between the Indians and Pyrrho, given the linguistic barriers.

Beyond these figures with attested connections to Pyrrho, it is plausible to suppose a certain influence on Pyrrho from Democritus. Pyrrho is reported to have had a special admiration for Democritus (Diogenes Laertius 9.67, citing Pyrrho's associate Philo); Democritus is one of the few philosophers besides Pyrrho himself who seems to escape serious criticism in Timon's *Lampoons*; and Anaxarchus belonged in the tradition of thinkers stemming from Democritus. The influence may have been mainly in the ethical area; Democritus, too, had an ethical ideal that is recognizably a forerunner to Pyrrho's *ataraxia*. If one adopts the epistemological interpretation of Pyrrho's philosophy, one may see an additional area of influence in Democritus' sceptical pronouncements about the prospects for knowledge of the world around us.

Alternatively, if one interprets Pyrrho along metaphysical lines, one may be inclined to look to Plato and the Eleatics as possible influences. Timon's verdicts on these figures in the *Lampoons* are at least partially favorable; and, as was hinted at earlier on, the dim view of sensibles that is suggested by a number of Plato's dialogues -- but also anticipated by the Eleatics -- seems to have something in common with Pyrrho's view of reality (on the metaphysical interpretation) as indeterminate. The difference, of course, is that Pyrrho does not suggest any higher level of reality such as Plato's Forms or the Eleatic Being.

8. Pyrrho's Influence

Pyrrho's relation to the later Pyrrhonists has already been discussed. Given the importance of Pyrrhonism in earlier modern philosophy, Pyrrho's indirect influence may be thought of as very considerable. But beyond his being adopted as a figurehead in later Pyrrhonism -- itself never a widespread philosophical movement -- Pyrrho seems to have had very little impact in the ancient world after his own lifetime. Both Cicero and Seneca refer to Pyrrho as a neglected figure without a following, and the surviving testimonia do not contradict this impression. It is possible that he had some influence on the form of scepticism adopted by Arcesilaus and other members of the Academy; the extent to which this is so is disputed and difficult to assess. It is also possible that the Epicureans, whose aim was also *ataraxia*, learned something from Pyrrho; there are indications of an association between Pyrrho and Nausiphanes, the teacher of Epicurus. But if so, the extent of the Epicureans' borrowing was strictly limited. For them, *ataraxia* is to be attained by coming to understand that the universe consists of atoms and void; and the Epicureans' attitude towards the senses was anything but one of mistrust.

Bibliography

Ancient Texts

- Decleva Caizzi, F., 1981, *Pirrone: Testimonianze*, Naples: Bibliopolis. Complete collection of texts referring to Pyrrho, with Italian translation and commentary.
- Long, A. A. and D. N. Sedley, 1987, *The Hellenistic Philosophers*, Cambridge: Cambridge University Press (2 vols), sections 1-3. Vol. 1 contains texts in English translation with philosophical commentary; vol. 2 contains original texts with philological commentary.

Secondary Literature

- Bett, R., 2000, *Pyrrho, his Antecedents and his Legacy*, Oxford: Oxford University Press
- Brennan, T., 1998, “Pyrrho on the Criterion”, *Ancient Philosophy* 18: 417-34
- Brunschwig, J., 1994, “Once again on Eusebius on Aristocles on Timon on Pyrrho”, in J. Brunschwig, *Papers in Hellenistic Philosophy*, Cambridge: Cambridge University Press: 190-211
- Burnyeat, M., 1980, “Tranquillity without a Stop: Timon, Frag. 68”, *Classical Quarterly* NS 30: 86-93
- Hankinson, J., 1995, *The Sceptics*, London: Routledge
- Stopper, M., 1983, “Schizzi Pirroniani”, *Phronesis* 28: 265-97
- von Fritz, K., 1963, “Pyrrhon”, in *Paulys Realencyclopädie der Classischen Altertumswissenschaft* xxiv: 89-106 89-106

Other Internet Resources

- Entry on Ancient Greek Skepticism, by Harald Thorsrud, in *The Internet Encyclopedia of Philosophy*

Related Entries

Sextus Empiricus | skepticism, ancient | Timon of Phlius

Richard Bett
rbett1@jhu.edu

Table of Contents

First published: August 5, 2002
Content last modified: August 5, 2002

Max Stirner

Max Stirner (1806-56) is best known as the author of the idiosyncratic and provocative *Der Einzige und sein Eigenthum* (1844). Familiar in English as *The Ego and Its Own* (a more literal translation might be *The Individual and his Property*), both the form and content of Stirner's work are disconcerting. He challenges expectations about how political and philosophical argument should be conducted, and seeks to shake confidence in the superiority of contemporary civilisation. He provides a sweeping attack on the modern world as dominated by religious modes of thought and oppressive social institutions, together with a brief sketch of a radical ‘egoistic’ alternative in which individual autonomy might flourish. The historical impact of *The Ego and Its Own* is not easy to assess. However, Stirner's book can plausibly be claimed to have had a destructive impact on his left-Hegelian contemporaries, to have played a significant role in the intellectual development of Karl Marx (1818-1883), and to have influenced the tradition of individualist anarchism.

1. Stirner's Life and Work

There is a stark contrast between the often melodramatic tone of Stirner's best-known work and the rather less sensational, indeed occasionally abject, events of his own life.

Stirner was born Johann Caspar Schmidt on 25 October 1806, the only child of lower middle class Lutheran parents living in Bayreuth. (‘Stirner’ was originally a nickname, resulting from his large

forehead, and only later adopted, as ‘Max Stirner’, first as a literary pseudonym, and then as his preferred name.) His father died when Stirner was only six months old, and he was brought up by his mother (who remarried) and then later by an aunt (who, when his mother moved from Bayreuth, looked after Stirner in order that he could continue his schooling at the renowned local Gymnasium). Stirner subsequently pursued his undergraduate studies, with little notable academic distinction, at the universities of Berlin, Erlangen, and Königsberg. At Berlin, he is known to have attended three lecture-series given by G.W.F. Hegel (1770-1831): on the philosophy of religion; on the history of philosophy; and on the philosophy of subjective spirit (whose subject matter concerned the cognitive structures and processes of the individual mind).

Towards the end of his university career, Stirner is said to have devoted much of his time to ‘family affairs’, a euphemism for his mother's deteriorating mental health. In 1832, he returned with his mother to Berlin, and sought, with only partial success, to qualify as a teacher. (Stirner's mother was committed to a mental home in 1837 and she would finally outlive him by some three years.) A period of private study and irregular work followed, including eighteenth months as an unpaid Latin teacher. During this time he married Agnes Butz (1815-1838), the illegitimate daughter of his landlady. Edgar Bauer (1820-1886) would later record that Stirner had confessed that having once caught sight of his wife naked he had been unable to touch her again. In August 1838, Agnes died giving birth to a still-born child.

Between 1839 and 1844 Stirner maintained something of a double life in Berlin. He obtained a position at a well-regarded private girls' school, and spent the next five years teaching history and literature, establishing a reputation as a polite and reliable teacher. Away from his teaching post, however, Stirner began to frequent the more avant-garde of Berlin's intellectual haunts. He used the reading room of the novelist Willibald Alexis (1798-1871), spent afternoons at the Café Stehely, and from 1841 onwards was a regular visitor to Hippel's wine bar on the Friedrichstrasse. The latter was the main meeting place of ‘the free’, an increasingly bohemian group of teachers, students, officers, and journalists, under the loose intellectual leadership of the left-Hegelian Bruno Bauer (1809-1882), who had recently been dismissed from his teaching post at the University of Bonn following an offical inquiry into the orthodoxy of his writings on the New Testament. This group included Marie Dähnhardt (1818-1902) who became Stirner's second wife (and the dedicatee of *The Ego and Its Own*). In this unconventional environment, and despite his calm and unassuming personal appearance, Stirner gained a reputation for his hostility to religion, intolerance of moderation, and ability to provoke fierce argument.

Stirner's earliest published writings date from this time in Berlin. In addition to some short and unremarkable pieces of journalism for the *Rheinische Zeitung* and the *Leipziger Allgemeine Zeitung*, they include a knowing review of Bruno Bauer's anonymous and parodic attack on Hegel in *The Trumpet of the Last Judgement* (1842) and an article on pedagogy entitled “The False Principle of Our Education” (1842), which adumbrated some of the themes of his own later work (for example, contrasting the training of individuals to an alien calling with the cultivation of the predisposition to become ‘sovereign characters’). During this period, Stirner is said to have occasionally alluded to a book that he was working on, but it seems that few of his associates took its existence seriously. The impact of *The Ego and Its Own* on these left-Hegelian circles was to be both considerable and unexpected. Stirner finished the book in April 1844 and *The Ego and Its Own* was subsequently published by the Leipzig bookseller Otto Wigand

(1795-1870) in an edition of a thousand copies (although dated 1845 the book appears to have been widely available by November of the previous year).

Measured by the reaction that it produced, *The Ego and Its Own* could be described as a critical success. The book was widely reviewed, and attracted attention from such leading figures as Bettina von Arnim (1785-1859), the doyenne of the Berlin literati, and Kuno Fischer (1824-1907), the distinguished neo-Kantian historian of philosophy. The book also generated responses from many of its left-Hegelian targets: Bruno Bauer, Ludwig Feuerbach (1804-1872), Moses Hess (1812-1875), and Arnold Ruge (1802-1880), all ventured into print in order to defend their own views against Stirner's polemic. However, *The Ego and Its Own* was neither a popular nor a financial success. Stirner had left his teaching post shortly before the book was published, and, by 1846, having squandered much of his second wife's inheritance, he was reduced to advertising in the *Vossische Zeitung* for a loan. Marie Dähnhardt left him towards the end of the same year. Many years later she was traced by Stirner's loyal biographer, the poet and novelist John Henry Mackay (1864-1933), to a Roman Catholic religious community in England. She refused to meet Mackay in person but wrote to him describing Stirner as a very sly man whom she had neither respected nor loved, and claiming that their relationship together had been more of a cohabitation than a marriage.

From 1847, Stirner led a quiet and miserable existence. He remained curiously detached from contemporary events (he seems, for example, to have largely ignored the revolution of 1848), and his daily life was increasingly dominated by economic hardship. Stirner continued to write intermittently, but commentators have generally found his later work to be of little independent interest (that is, apart from its uncertain potential to illuminate *The Ego and Its Own*). He translated into German some of the economic writings of Jean-Baptiste Say (1767-1832) and Adam Smith (1723-1790), and may have written a series of short journalistic pieces for the *Journal des oesterreichischen Lloyd*. In 1852, he published part of a *History of Reaction*, mainly consisting of excerpts from other authors, including Edmund Burke (1729-1797). Stirner's main strategy for economic survival in this period relied on changing addresses in order to evade his creditors, although he does not appear to have moved quickly enough to avoid two brief periods in a debtors' prison in 1853 and 1854.

In May 1856, still living in reduced circumstances in Berlin, Stirner fell into a fever after being stung in the neck by a winged insect. Following a brief remission, he died on 25 June. His death went largely unnoticed by the outside world.

2. *The Ego and Its Own*

2.1 Form and Structure

Modern readers hoping to understand *The Ego and Its Own* are confronted by several obstacles, not least the form, structure, and argument, of Stirner's book.

Much of Stirner's prose—which is crowded with aphorisms, italicisation, and hyperbole—appears calculated to disconcert. Most striking, perhaps, is the use of word play. Rather than reach a conclusion through the conventional use of argument, Stirner often approaches a claim that he wishes to endorse by exploiting words with related etymologies or formal similarities. For example, he frequently associates words for property (such as ‘*Eigentum*’) with words connoting distinctive individual characteristics (such as ‘*Eigenheit*’) in order to promote the claim that property is expressive of selfhood. (Stirner's account of egoistic property—see below—gives this apparently orthodox Hegelian claim a distinctive twist.)

This rejection of conventional forms of intellectual discussion is linked to Stirner's substantive views about language and rationality. His unusual style reflects a conviction that both language and rationality are human products which have come to constrain and oppress their creators. Stirner maintains that accepted meanings and traditional standards of argumentation are underpinned by a conception of truth as a privileged realm beyond individual control. As a result, individuals who accept this conception are abandoning a potential area of creative self-expression in favour of adopting a subordinate role as servants of truth. In stark contrast, Stirner insists that the only legitimate restriction on the form of our language, or on the structure of our arguments, is that they should serve our individual ends. It is the frequent failure of ordinary meanings and standard forms of argument to satisfy his interpretation of this criterion which underpins Stirner's remorselessly idiosyncratic style.

The Ego and Its Own has an intelligible, but scarcely transparent, structure. It is organised around a tripartite account of human experience, initially introduced in a description of the stages of an individual life. The first stage in this developmental narrative is the *realistic* one of childhood, in which children are constrained by material and natural forces such as their parents. Liberation from these external constraints is achieved with what Stirner calls the self-discovery of mind, as children find the means to outwit those forces in their own determination and cunning. The *idealistic* stage of youth, however, contains new internal sources of constraint, as individuals once more become enslaved, this time to the spiritual forces of conscience and reason. Only with the adulthood of *egoism* do individuals escape both material (external) and spiritual (internal) constraints, learning to value their personal satisfaction above all other considerations.

Stirner portrays this dialectic of individual growth as an analogue of historical development, and it is a tripartite account of the latter which structures the remainder of the book. Human history is reduced to successive epochs of realism (the ancient, or pre-Christian, world), idealism (the modern, or Christian, world), and egoism (the future world). Part One of *The Ego and Its Own* is devoted to the first two of these subjects, whilst Part Two is concerned with the third of them.

2.2 The Ancient and Modern Worlds

Part One of *The Ego and Its Own* is backward-looking, in that it is concerned with the ancient and modern worlds rather than with the future, and negative, in that its primary aim is to demonstrate the failure of modernity to escape from the very religious modes of thought which it claims to have outgrown. The bulk of Stirner's historical account is devoted to the modern epoch, and he discusses the

ancient world only insofar as it contributes to the genesis of modernity. In both cases, however, the majority of his examples are taken from the realm of cultural and intellectual affairs. Cumulatively these examples are meant both to undermine historical narratives which portray the modern development of humankind as the progressive realisation of freedom and to support an account of individuals in the modern world as increasingly oppressed by the spiritual. For Stirner, the subordination of the individual to spirit—in any of its guises—counts as religious servitude.

Stirner's account of the historical development of modernity largely revolves around a single event, the Reformation. He attempts to show that, from the perspective of the individual, the movement from Catholic to Protestant hegemony was not a liberating one, but instead constituted both an extension and intensification of the domination of spirit. The Reformation extended, rather than contracted, the sphere of religious control over the individual because it refused to recognise the distinction between the spiritual and the sensuous. Rather than prevent priests marrying, for example, Protestantism made marriage religious, thereby absorbing the sensuous into the sphere of the spiritual. The Reformation also intensified, rather than relaxed, the bond between individuals and religion. The more inward faith of Protestantism, for example, established a perpetual internal conflict between natural impulses and religious conscience. In a typically vivid and belligerent metaphor, Stirner describes this internal conflict in the individual as analogous to the struggle between the population and the secret police in the contemporary body politic.

Stirner's claim that the modern world reproduced, rather than abolished, religious modes of thought provides the opportunity for a sustained attack on the writings of his left-Hegelian contemporaries, Ludwig Feuerbach in particular, for failing to overcome the subordination of the individual to spirit. Stirner's expansive definition of religion enables him to portray Feuerbach's work as sustaining rather than undermining religious modes of thought. The primary error of Christianity, according to Feuerbach, was that it took human predicates and projected them into another world as if they constituted an independent being. However, Stirner insists that Feuerbach's rejection of God as a transcendental subject leaves the divinity of the Christian predicates untouched. In short, rather than describing human nature as it is, Feuerbach is said to have deified a prescriptive account of what being human involves. As a result, the real kernel of religion, the positing of an ‘essence over me’ (46), had been left intact. (All page references in parenthesis are to the English translation of *The Ego and Its Own* cited in the Bibliography below.) Indeed, Stirner suggests that Feuerbach's achievement was to have effected a ‘change of masters’ (55) which made the tyranny of the divine over the individual even more complete, because human nature (unlike the conventional Christian God) was an immanent divinity which could possess both believers and unbelievers alike.

Stirner extends his critique to the work of all the left-Hegelians, including those with whom he had associated in Berlin. Although they disagreed about the content of human nature—for ‘political liberals’ like Arnold Ruge human nature was identified with citizenship, for ‘social liberals’ like Moses Hess human nature was identified with labour, and for ‘humane liberals’ like Bruno Bauer human nature was identified with critical activity—all the left-Hegelians are said to have reproduced the basic Feuerbachian error: separating the individual from his human essence, and setting that essence above the individual as something to be striven for. In contrast, Stirner maintains that because it has no universal or prescriptive

content, human nature cannot ground any claim about how we ought to live. His own intellectual project—which he describes as an attempt to rehabilitate the prosaic and mortal self, the ‘un-man’ (124) for whom the notion of a calling is alien—is intended as a radical break with the work of these contemporaries.

2.3 The Egoistic Future

Part Two of *The Ego and Its Own* is forward-looking, in that it is concerned with the egoistic future rather than the ancient or modern worlds, and positive, in that it aims to establish the possibility that Stirner's contemporaries could escape the tyranny of religion.

Stirner's account of the developing historical relationship between the individual and society is advanced in a series of parallels which are designed to portray egoism as the embodiment of a more advanced civilisation. At one point, he neatly inverts the terms of a familiar progression (rehearsed by countless early modern political thinkers) from a state of nature to civil society. It is membership of society, and not isolation, Stirner suggests, which is humankind's “state of nature” (271), in that it constitutes an early stage of development whose inadequacies are, in due course, outgrown. Elsewhere, he describes the developing relationship between the individual and society as analogous to that between a mother and her child. As the individual (the child) develops a mature preference for a less suffocating environment, it must throw off the claims of society (the mother) which seeks to maintain it in a subordinate position. In both cases, Stirner draws the lesson that the individual must move from social to egoistic relationships in order to escape subjection.

What is meant by ‘egoism’, however, is not always clear. Stirner is occasionally portrayed as a psychological egoist, that is, as a proponent of the descriptive claim that all (intentional) actions are motivated by a concern for the self-interest of the agent. However, this characterisation of Stirner's position appears mistaken. Not least, *The Ego and Its Own* is structured around the opposition between egoistic and non-egoistic forms of experience. Indeed, he appears to hold that non-egoistic action has predominated historically (in the epochs of realism and idealism). Moreover, at one point, Stirner explicitly considers adopting the explanatory stance of psychological egoism only to reject it. In a discussion of a young woman who sacrifices her love for another in order to respect the wishes of her family, Stirner remarks that an observer might be tempted to maintain that selfishness has still prevailed in this case since the woman clearly preferred the wishes of her family to the attractions of her suitor. However, Stirner rejects this hypothetical explanation, insisting that, provided “the pliable girl were conscious of having left her self-will unsatisfied and humbly subjected herself to a higher power” (197), we should see her actions as governed by piety rather than egoism.

It is also a mistake to think of Stirner as advocating a normative proposition about the value of self-interested action. Stirnerian egoism needs to be distinguished from the individual pursuit of narrow self-interest as it is conventionally understood. In *The Ego and Its Own,* Stirner discusses the important example of an avaricious individual who sacrifices everything in pursuit of material riches. Such an individual is clearly self-interested (he acts only to enrich himself) but it is an egoism which Stirner

rejects as one-sided and narrow. Stirner's reason for rejecting this form of egoism is instructive. He suggests that the avaricious man has become enslaved to a single end, and such enslavement is incompatible with egoism properly understood.

Stirnerian egoism is perhaps best thought of, not in terms of the pursuit of self-interest, but rather as a variety of individual self-government or autonomy. Egoism properly understood is to be identified with what Stirner calls ‘ownness [*Eigenheit*]’, a type of autonomy which is incompatible with any suspension, whether voluntary or forced, of individual judgement. “I am my own”, Stirner writes, “only when I am master of myself, instead of being mastered … by anything else” (153). This Stirnerian ideal of self-mastery has external and internal dimensions, requiring both that we avoid subordinating ourselves to others and that we escape being ‘dragged along’ (56) by our own appetites. In short, Stirner not only rejects the legitimacy of any subordination to the will of another but also recommends that individuals cultivate an ideal of emotional detachment towards their own appetites and ideas.

Judged against this account of egoism, characterisations of Stirner as a ‘nihilist’—as rejecting all normative judgement—would also appear to be mistaken. The popular but inaccurate description of Stirner as a ‘nihilist’ is encouraged by his explicit rejection of morality. Morality, on Stirner's account, involves the positing of obligations to behave in certain fixed ways. As a result, he rejects morality as incompatible with egoism properly understood. However, this rejection of morality is not grounded in the rejection of values as such, but in the affirmation of what might be called non-moral goods. That is, Stirner allows that there are actions and desires which, although not moral in his sense (because they do not involve obligations to others), are nonetheless to be assessed positively. Stirner is clearly committed to the non-nihilistic view that certain kinds of character and modes of behaviour (namely autonomous individuals and actions) are to be valued above all others. His conception of morality is, in this respect, a narrow one, and his rejection of the legitimacy of moral claims is not to be confused with a denial of the propriety of all normative judgement. There is, as a result, no inconsistency in Stirner's frequent use of an explicitly evaluative vocabulary, as when, for example, he praises the egoist for having the ‘courage’ (265) to lie, or condemns the ‘weakness’ (197) of the individual who succumbs to pressure from his family.

Two features of Stirner's position emerge as fundamental. First, he values ‘ownness’ neither as one good amongst many, nor as the most important of several goods, but rather as the only good. Second, he adopts an account of self-mastery which is incompatible with the existence of any legitimate obligations to others, even those which an individual has voluntarily undertaken (thereby rejecting perhaps the most familiar way of reconciling individual autonomy with the existence of binding obligations). In short, Stirner appears to value nothing other than individual self-mastery, and he interprets the latter in a stringent and idiosyncratic manner.

2.4 Some Consequences of Egoism

The consequences of Stirner's position appear extreme and far-reaching. As the example of morality suggests, egoists are likely to find themselves in conflict with some cherished social institutions and

practices. Stirner consistently associates (non-egoistic) society with relationships of ‘belonging’, which he treats as involving the subjugation of individuals. For example, he maintains that ‘the forming of family ties binds a man’ (102). (Stirner never appears to consider seriously the possibility that, in at least some of these social relationships, belonging might have more positive associations; for example, of being at home or of feeling secure.) Confronted with the conflict between egoism and ‘society’, Stirner is not prompted to re-examine his commitment to, or understanding of, self-mastery, but instead confidently denies the legitimacy of those conventional institutions and practices. Two examples of this response may suffice.

On Stirner's account, there is a necessary antipathy between the egoistic individual and the state. This inevitable hostility is based on the conflict between Stirner's conception of autonomy and any obligation to obey the law. “Own will and the state”, he writes, “are powers in deadly hostility, between which no ‘perpetual peace’ is possible” (175). Since self-mastery is incompatible with, and valued more highly than, any obligation to obey the law, Stirner rejects the legitimacy of political obligation. Note that this rejection stands irrespective of the foundation of that political obligation, and whatever the form of the state. “Every state”, Stirner insists, “is a despotism, be the despot one or many.” (175) Even in the hypothetical case of a direct democracy in which a collective decision had been made unanimously, Stirner denies that the egoist would be bound by the result. To be bound today by “my will of yesterday”, he maintains, would be to turn my ‘creature’, that is ‘a particular expression of will’, into my ‘commander’; it would be to freeze my will, and Stirner denies that ‘because I was a fool yesterday I must remain such’ (175).

Promise-keeping is another early victim of this commitment to, and understanding of, self-mastery. Stirner associates the institution of promising with illegitimate constraint, since the requirement that duly made promises be kept is incompatible with his understanding of individual autonomy. Stirner rejects any general obligation to keep promises as just another attempt to bind the individual. The egoist, he suggests, must embrace the heroism of the lie, and be willing to break even his own word “in order to determine himself instead of being determined” (210). Note that Stirner's enthusiasm is reserved not for those who break their word in the service of some larger spiritual goal (as Luther, for example, became unfaithful to his monastic vows for God's sake), but rather for the individual who is willing to break his word for his own sake.

As well as a negative account of the institutions and practices that egoists must reject as incompatible with autonomy, *The Ego and Its Own* also contains some positive suggestions about the possible shape of egoistic relationships which do not conflict with individual self-mastery. In particular, Stirner provides a brief sketch of what he calls the “union of egoists [*Verein von Egoisten*]” (161).

The egoistic future is said to consist not of wholly isolated individuals but rather in relationships of ‘uniting’, that is, in impermanent connections between individuals who themselves remain independent and self-determining. The central feature of the resulting union of egoists is that it does not involve the subordination of the individual. The union is “a son and co-worker” (273) of autonomy, a constantly shifting alliance which enables individuals to unite without loss of sovereignty, without swearing allegiance to anyone else's ‘flag’ (210). This union of egoists constitutes a purely instrumental association

whose good is solely the advantage that the individuals concerned may derive for the pursuit of their individual goals; there are no shared final ends and the association is not valued in itself.

Stirner occasionally appears uncertain as to how best to elaborate this basic account of egoistic social relations. In *The Ego and Its Own*, he appears to be pulled in two divergent directions.

In the first, and least typical, of these moods, Stirner shies away somewhat from the suggestion that his views might have radical consequences. More precisely, he seeks to suggest that certain familiar and worthwhile relationships (such as 'love') might continue into the egoistic future. This suggestion is presumably aimed at making that future appear more attractive (not least to those attached to these familiar and worthwhile relationships). However, it is far from certain that all of the relationships he mentions would emerge intact from their reincarnation in egoistic form.

Consider, for example, Stirner's contrast between two different kinds of love: the 'bad case' where 'ownness' is sacrificed, and egoistic love in which self-mastery is retained. Egoistic love allows the individual to deny himself something in order to enhance the pleasure of another, but only because his own pleasure is enhanced as a result. The object of egoistic love, in other words, remains the individual himself. The egoist will not sacrifice his autonomy and interests to another, but rather loves only as long as "love makes me happy" (258). At one point, Stirner characterises this relationship as one in which the individual 'enjoys' the other (258). The description is a revealing one, since enjoying another person and loving them would appear to be rather different matters. Loving another person in the conventional (and non-egoistic) sense might be thought to include the desire to promote the welfare of that person, even when it is not in our interests, or when it conflicts with our own wants and happiness. In this respect, it stands at some distance from Stirner's account of egoistic love. The point here is not a terminological one—Stirner rightly cares little whether we call egoistic love 'love' and "hence stick to the old sound" (261) or whether we invent a new vocabulary—but rather that a world without this experience would be an unfamiliar and impoverished one. Stirner appears to have failed to establish that this particular familiar and worthwhile relationship would survive this reestablishment on egoistic premises.

In the second, and more representative, of these moods, Stirner acknowledges the radical and unfamiliar consequences of adopting an egoistic order. Indeed, in places, he might be said to revel in the acknowledgement that his views have startling consequences from which few will take any solace. This is one of the sources of the melodramatic tone of parts of *The Ego and Its Own*.

Stirner describes the relation between the egoist and his objects (which include, of course, other persons) as a property relation. The egoist stands in a relation of 'ownership' to the wider world. This notion of 'egoistic property' is not to be confused with more familiar juridical concepts of ownership (such as private property or collective ownership). These more familiar forms of property rest on notions of right, and involve claims to exclusivity or constraints on use, which Stirner rejects. Egoistic property is rather constituted by the 'unlimited dominion' (223) of individuals over the world, by which Stirner appears to mean that there are no moral constraints on how an individual might relate to things and other persons. Stirner sometimes describes the resulting association between people as involving relationships "of utility, of use" (263). The egoist, he suggests, views others as "nothing but—my food, even as I am fed

upon and turned to use by you" (263). Stirner embraces the stark consequences of this rejection of any general obligation towards others, insisting, for example, that the egoist does not renounce "even the power over life and death" (282). Over the course of the book, he variously declines to condemn the officer's widow who strangles her child (281), the man who treats his sister 'as wife also' (45), and the murderer who no longer fears his act as a 'wrong' (169). In a world in which "we owe each other nothing" (263), it seems that acts of infanticide, incest, and murder, might all turn out to be justified.

Stirner acknowledges that few readers of *The Ego and Its Own* will draw any comfort from his vision of an egoistic future, but insists that the welfare of this audience is not of any interest to him. Indeed, Stirner suggests that, if he had been motivated by a concern for others, then he would have had to conceal rather than propagate his ideas. As it is, even if he had believed that these ideas would lead to the "bloodiest wars and the fall of many generations" (263) Stirner maintains that he would still have disseminated them.

3. Stirner's Influence

At the time of his death, Stirner's brief period of notoriety was long over, his book had been out of print for several years, and there was little sign that his work might have any longer term impact. Since then, however, *The Ego and Its Own* has been translated into at least eight languages, and appeared in over one hundred editions.

Many of the shifting claims that have been made for the influence of Stirner's ideas would appear to reflect changing historical enthusiasms as much as they accurately capture central features of his thought. For example, at the beginning of the twentieth century, Stirner was frequently portrayed as a precursor of Friedrich Nietzsche (1844-1900), as having anticipated, if not influenced (it is far from certain that Nietzsche had ever read Stirner's work), both the style and substance of Nietzsche's work. In the 1960s and early 1970s, Stirner was rediscovered as a forerunner of existentialism, whose anti-essentialist concept of the self as a 'creative nothing' had affinities with the notion of human nature employed by Jean-Paul Sartre (1905-1980). More recently, Stirner has been identified as a nascent poststructuralist, employing a genealogical critique of humanist discourses of power and identity. It would be wrong to suggest that these various parallels are wholly implausible. Nevertheless, they may not offer the most accurate account of Stirner's impact on philosophical and political thought.

The influence of Stirner's work is perhaps more plausibly located in two rather different contexts. As far as its contemporary impact on the cultural life of *Vormärz* Germany is concerned, *The Ego and Its Own* had a destructive impact on Stirner's left-Hegelian contemporaries, and played a significant role in the intellectual development of Karl Marx. As far as its longer term historical influence is concerned, Stirner's work has become a founding text in the tradition of individualist anarchism.

Stirner's insistence that his radical contemporaries had failed to break with religious modes of thought prompted most of the leading left-Hegelians to defend their own work in public against this attack. In perhaps the most important of these replies, a defensive and ill-tempered Feuerbach (who suspected

Stirner of trying to make a name for himself at his own expense) was widely seen as struggling to maintain a besieged and outdated position. Stirner responded to three of these left-Hegelian reviews—the defence of Bauer's ‘humane liberalism’ by ‘Szeliga’ (the pseudonym of Franz Zychlinski (1816-1900)); the defence of socialism by Moses Hess; and the defence of Feuerbach by Feuerbach himself—in an article entitled ‘Stirner's Critics’ (1845). In this confident rejoinder, Stirner reiterated some of the central themes of *The Ego and Its Own* and clarified the character of his commitment to egoism.

Stirner's work also had a significant impact on a little known contemporary associate of these left-Hegelians, one Karl Marx. Between 1845 and 1846, Marx collaborated with Friedrich Engels (1820-1895) on *The German Ideology*, a fierce and sustained attack on their erstwhile philosophical contemporaries. They were unsuccessful in finding a publisher for their lengthy polemic and it was 1932 before this critical engagement with the work of Bauer, Feuerbach, and Stirner, appeared in print. The account of Stirner contained in *The German Ideology* takes up over three hundred pages of the published text (unfortunately abridged editions occasionally omit this dense but fascinating part of the book), and, although Marx is remorselessly critical of Stirner's position, it scarcely follows that *The Ego and Its Own* was without influence on the former's own work. Not least, Stirner's book appears to have been decisive in motivating Marx's break with the work of Feuerbach, whose influence on many of Marx's earlier writings is readily apparent, and in forcing Marx to reconsider the role that concepts of human nature should play in social criticism.

Finally, and over a longer period of time, the author of *The Ego and Its Own* has become best-known as a member of, and influence upon, the anarchist tradition. In particular, Stirner's name appears with familiar regularity in historically-orientated surveys of anarchist thought as one of the earliest and best-known exponents of individualist anarchism. The affinity between Stirner and anarchism lies in his rejection of political obligation and in his denial of the legitimacy of the state. However, unlike many anarchists, Stirner does not maintain that individuals have a positive obligation to destroy the state (insofar as this may lie within their power), but rather suggests that individuals should simply cheat and evade the state's demands in order to maintain their autonomy. Anarchists influenced by Stirner's individualism and his suspicion of the state can be found in several European countries. However, his best-known anarchist admirers were in America, in the circle which formed around Benjamin R. Tucker (1854-1939) and the remarkable journal *Liberty* (founded in 1881).

Stirner is unlikely to have regretted these disputes about the nature and influence of *The Ego and Its Own*. In considering the various interpretative accounts of the Bible, he himself declined to chose between the judgement of the child who played with the book, the Inca emperor Atahualpa (c.1502-1533) who threw it away when it failed to speak to him, the priest who praised it as the word of God, and the critic who dissected it as a purely human invention. The plurality of interpretations of his own work might well have both amused Stirner and encouraged him in his view that there could be no legitimate constraints on the meaning of a text. Stirner once described himself as writing to procure for his thoughts an existence in the world, and insisted that what subsequently happened to these ideas ‘is your affair and does not trouble me’ (263).

Bibliography

Works by Stirner

- *Der Einzige und sein Eigentum* (Stuttgart: Philipp Reclam, 1972). (A modern edition of Stirner's best-known work.)
- *Max Stirner's Kleinere Schriften und seine Entgegnungen auf die Kritik seines Werkes "Der Einzige und sein Eigentum". Aus den Jahren 1842-1848*, edited by J.H. Mackay, second revised edition (Berlin: Bernhard Zack, 1914). (An extensive collection of Stirner's lesser writings.)
- *Parerga Kritiken Repliken*, edited by Bernd A. Laska (Nürnberg: LSR, 1986). (A modern selection of Stirner's lesser writings.)

Works by Stirner in Translation

- *The False Principle of Our Education*, edited by James J. Martin (Colorado Springs: Ralph Myles, 1967). (An early article on pedagogy.)
- "Stirner's Critics", *The Philosophical Forum*, volume 8 (1978) pp.66-80. (A partial translation of Stirner's 1845 response to critics, covering his reply to Feuerbach.)
- "Art and Religion", Lawrence S. Stepelevich (edited), *The Young Hegelians. An Anthology* (Cambridge: Cambridge University Press, 1983) pp.327-334. (An article on an eminently Hegelian topic from 1842.)
- *The Ego and Its Own*, edited by David Leopold (Cambridge: Cambridge University Press, 1995). (A well-annotated English edition of Stirner's best-known work.)

Works by Others

- Carroll, John, *Break-Out from the Crystal Palace. The Anarcho-Psychological Critique: Stirner, Nietzsche, Dostoevsky* (London: Routledge and Kegan Paul, 1974).
- Clark, John P, *Max Stirner's Egoism* (London: Freedom Press, 1976).
- Helms, Hans G., *Die Ideologie der anonymen Gesellschaft. Max Stirner 'Einziger' und der Fortschritt des demokratischen Selbstbewußtseins vom Vormärz bis zur Bundesrepublik* (Köln: Du Mont Schauberg, 1966). (Contains an excellent bibliography.)
- Koch, Andrew M., "Max Stirner: The Last Hegelian or the First Poststructuralist", *Anarchist Studies*, volume 5 (1997) pp.95-108.
- Laska, Bernd A., *Ein dauerhafter Dissident. 150 Jahre Stirners 'Einziger'. Eine kurze Wirkungsgeschichte*, Nürnberg: LSR-Verlag, 1996.
- Lobkowicz, Nicholas, "Karl Marx and Max Stirner", Frederick J. Adelmann (edited), *Demythologising Marxism* (The Hague: Martinus Nijhoff, 1969) pp.64-95.
- Leopold, David, "Introduction", Max Stirner, *The Ego and Its Own* (Cambridge: Cambridge University Press, 1995) pp.xi-xxxii.
- Löwith, Karl, *From Hegel to Nietzsche. The Revolution in Nineteenth Century Thought* (London:

Constable, 1965). (First published in German in 1941.)

- Mackay, John Henry, *Max Stirner. Sein Leben und sein Werk*, second edition (Berlin, 1914).
- Martin, James J., *Men Against the State. The Expositors of individualist Anarchism in America, 1827-1908* (DeKalb, Illinois: Adrian Allen, 1953).
- Maruhn, Jürgen, *Die Kritik an der Stirnerschen Ideologie im Werk von Karl Marx und Friedrich Engels* (Frankfurt: R.G. Fischer, 1982).
- Marx, Karl and Friedrich Engels, *The German Ideology*, *Marx Engels Collected Works* (London: Lawrence and Wishart, 1976) volume 5.
- Paterson, R.W.K., *The Nihilistic Egoist: Max Stirner* (Oxford: Oxford University Press for University of Hull Publications, 1971).
- Stepelevich, Lawrence S., “Max Stirner and Ludwig Feuerbach”, *Journal of the History of Ideas*, volume 39 (1978) pp.451-463.
- Stepelevich, Lawrence S., “Max Stirner as Hegelian”, *Journal of the History of Ideas*, volume 46 (1985) pp.597-614.
- Tucker, Benjamin R., *Instead of a Book. By a Man Too Busy to Write One. A Fragmentary Exposition of Philosophical Anarchism, culled from the writings of Benj. R. Tucker* (New York: Haskell House, 1969). (First published in 1897.)

Other Internet Resources

- Max Stirner Page at "Non Serviam" (maintained by Svein Olav Nyberg)
- Max Stirner im LSR Projekte (maintained by Bernd A. Laska)

Related Entries

Bauer, Bruno | Hegel, Georg Wilhelm Friedrich | Marx, Karl

David Leopold
david.leopold@merton.ox.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 27, 2002
Content last modified: July 31, 2002

Bruno Bauer

Bruno Bauer (1809-1882), philosopher, historian, and theologian. His career falls into two main phases, divided by the revolutions of 1848. In the 1840's, the period known as the Vormärz or the prelude to the German revolutionary events of March 1848, Bauer was a leader of the Left-Hegelian movement, developing a republican interpretation of Hegel, which combined ethical and aesthetic motifs. His theory of infinite self-consciousness, derived from Hegel's account of subjective spirit, stressed rational autonomy and historical progress. Investigating the textual sources of Christianity, Bauer described religion as a form of alienation, which, because of the deficiencies of earthly life, projected irrational, transcendent powers over the self, while sanctioning particularistic sectarian and material interests. He criticized the Restoration state, its social and juridical base, and its orthodox religious ideology. Analyzing the emergence of modern mass society, he rejected liberalism for its inconsequent opposition to the existing order, and for its equation of freedom with property, but he accused socialism of an inadequate appreciation of individual autonomy. After the defeats of 1848, Bauer repudiated Hegel. He predicted a general crisis of European civilization, caused by the exhaustion of philosophy and the failure of liberal and revolutionary politics. New prospects of liberation would, he believed, issue from the crisis. His late writings examined the emergence of Russia as a world power, opening an era of global imperialism and war. These writings influenced Nietzsche's thinking on cultural renewal. Friedrich Engels and Karl Kautsky claimed Bauer's religious criticism for the socialist movement, while the anti-traditionalist conservatism and anti-Semitism of his late work link him to the revolutionary right in the twentieth century.

1. Career

Bauer's family moved from Saxony to Berlin in 1815. At the University of Berlin (1828 -1834), he studied under Hegel, Schleiermacher, and the Hegelians Hotho and Marheineke. His 1829 essay on

Kant's aesthetics won the Prussian royal prize in philosophy, on Hegel's recommendation. From 1834 to 1839, he lectured on theology and biblical texts in Berlin. He was transferred to the theology faculty at Bonn after publishing an attack on his colleague and former teacher Hengstenberg. He taught in Bonn from 1839 till spring 1842, when he was dismissed for the unorthodoxy of his writings on the New Testament. The dismissal followed a consultation by the ministry of education with the theology faculties of the six Prussian universities, but no consensus emerged from the academic inquest. The order for Bauer's dismissal came directly from the king of Prussia, Friedrich Wilhelm IV, who had decreed the suspension from state employment of participants in a banquet to honour the South German liberal Karl Welcker, held in Berlin in 1841. On that occasion, Bauer had proposed a toast to Hegel's conception of the state.

From 1842 to 1849, Bauer was active in political journalism and historical research on the Enlightenment and the French Revolution. He argued against the emancipation of Prussian Jews in 1842-43, seeing this proposal as a political legitimation of particular religious interests. He was the object of polemical attacks by Marx and Engels in *The Holy Family* (1844) and *The German Ideology* (written in 1845-46). With his brother Edgar, Bauer founded the Charlottenburg Democratic Society in 1848, and stood unsuccessfully for election to the Prussian National Assembly on a platform of popular sovereignty.

Remaining in Prussia after the defeats of 1848-49, Bauer continued to produce work of biblical criticism and political analysis. He wrote in the mid 1850's for *Die Zeit*, a government-sponsored newspaper, in which his anti-liberalism took a conservative turn. He contributed articles on European affairs to other newspapers, such as *Die Post*, the *Kleines Journal*, and the *New York Daily Tribune*. From 1859-66 he collaborated with F.W.H. Wagener on his conservative *Staats- und Gesellschafts-Lexikon*, editing almost all 23 volumes, and writing numerous articles, several with anti-Semitic themes. In 1865 he acquired a small farm in Rixdorf, on the outskirts of Berlin. He died there in April 1882.

2. Bauer's Writings, 1829-50

Bauer was a prolific writer, publishing a dozen substantial books and over 60 articles between 1838 and 1848 alone, but no critical edition of these works exists. They included analyses of Hegel, the Bible, modern theologies, the Enlightenment, and the French Revolution and its aftermath. The interpretation of Bauer's work is problematic for several reasons. Because of anonymous, pseudonymous, and collaborative publication, some attributions are disputed; and divergences exist between Bauer's published texts and private correspondence. In the anonymous *Trumpet of the Last Judgement* (1841) and *Hegel's Doctrine of Religion and Art* (1842), Bauer spoke not in his own voice, but in the ironic guise of a conservative critic of Hegel, attributing to Hegel his own revolutionary views.

Three lines of interpretation of Bauer can be distinguished. These focus on his early work; his later writings have attracted little critical attention. The first sees Bauer as a radical subjectivist, whose social and religious criticism was closer to Enlightenment rationalism than to Hegel (Sass, 1978; Brudney, 1998). The second, largely influenced by Marx, insists on Bauer's abandonment of the Hegelian left after 1843 (Rossi, 1974; Pepperle, 1978). The third emphasizes the continuity throughout the Vormärz of

Bauer's thought and of his republicanism, based on the Hegelian idea of the unity of thought and being (Moggach, 2002).

Bauer's prize manuscript of 1829, *De pulchri principiis*, presented the unity of concept and objectivity as the central idea of Hegel's idealism. It examined this unity as expressed in art, comparing Hegel's aesthetic theory to Kant's Third Critique. The manuscript supplemented the criticisms of Kant from Hegel's Berlin Aesthetics lectures with the logical analysis of categories provided by the 1827 *Encyclopaedia of the Philosophical Sciences*. Bauer argued that while the Critique of Judgement attempted to bridge thought and being, and thus opened the way to Hegel, it reproduced the antinomies characteristic of the first two critiques. Kant's synthesis failed, since he continued to define the concept as merely subjective, and the object as the unknowable thing-in-itself, transcending the cognitive power. Self-consciousness, or the subject of the transcendental unity of apperception, was likewise impervious to cognition from the Kantian standpoint. In Hegel's syllogisms of the idea, objectivity attained rational form, while the concept acquired an explicit, material existence. Beauty, life, and idea were moments in the process which constituted the actuality of reason. As the immediate unity of thought and objectivity, art illustrated the inexhaustible fecundity of the philosophical Idea. The manuscript underlined the opposition of faith and reason in its critique of the religious conceptions of the unity of thought and being. Faith was taken to be inimical to free inquiry, which is the element of reason.

Throughout the 1830's, however, Bauer sought to reconcile thought and being through the idea of rational faith. In *Zeitschrift für spekulative Theologie*, which he edited between 1836 and 1838, and in *Jahrbücher für wissenschaftliche Kritik*, he offered a speculative account of Christian doctrines as exemplifications of logical categories. In the 1838 *Religion of the Old Testament*, Bauer depicted religious experience as a product of self-consciousness. He proposed both a transcendental account, stressing the conditions of possibility of religious experiences, and a phenomenological sequence of their forms: a legalistic subordination to an authoritarian deity in the early books of the Old Testament expressed a merely external relation between God and man, while the messianic consciousness of later books heralded a higher form, the immanence of the universal in the community; but this consciousness could only point up the inadequacy of the law, not yet propose the effective overcoming of estrangement. The texts of the 1830's located the logical structure which, for Bauer, defined the religious consciousness: the immediate identity between particularity, whether of a subject or a community, and the abstract universal, a unity achieved without self-transformation. By 1839, Bauer deduced the political implications of this view: the religious consciousness asserted this immediate identity as a monopolistic, sectarian claim, excluding other particulars from equivalent status. The essence of religion for Bauer was now a hubristic particularism, which also conferred a transcendent status on the universal, as a realm divorced from concrete social relationships. This position received its fullest exposition in Bauer's *Christianity Revealed* (1843). He already sketched this argument in *Herr Dr. Hengstenberg*, of 1839, publicly breaking with orthodox and conservative versions of Christianity, and stressing the discontinuity between Christianity and Judaism. By 1840-41, Bauer would present the emancipated philosophical self-consciousness as opposed to all forms of religious representation. His political radicalism and republicanism were cemented by his recognition of the structural identity between the private interests fostered by the Restoration order, and the monopolistic religious consciousness.

Bauer's political and theoretical radicalization is evidenced in his biblical studies. The series is comprised of *Critique of the Gospel of John* (1840), and the three-volume *Critique of the Synoptic Gospels* (1840-42). Together with his 1838 study of the Old Testament, these volumes criticized the stages of revealed religion, and forms of self-alienated spirit in history. Bauer's critique of John's gospel demonstrated the opposition between the free self-consciousness and the religious spirit. His stated purpose was to restore the Christian principle to its source in creative self-consciousness; he did not yet openly oppose the principle itself, but sought to differentiate it, as a rational idea, from ecclesiastical dogmatism. The positivity of Christianity derived from the abstract understanding, rather than from speculative reason, which led religious experience back to its subjective roots. The rational core of Christianity was the identity of God and man, but theology had built an untenable doctrinal system on this foundation. Speculation now undermined dogma; it was not confirmed by it, as Bauer's mid-1830's articles had maintained. In his correspondence, though not in this text itself, Bauer indicated that this restoration of the Christian principle was also its overthrow, as the unity of universal and particular could now be grasped in more tangible and earthly forms. Christianity was a necessary but now transcended stage in the development of the human spirit, to be supplanted by new expressions of autonomous self-consciousness.

In his critique of the Synoptics, Bauer's object was more openly to negate dogmatic Christianity, mobilized in defence of the absolutist order. The incidents described in the gospels were products of the religious consciousness, rather than factual reports. Bauer's critique of John convinced him that the gospel narrative was a purely literary product, and he now argued that the Synoptics too contained no historically authentic material. Bauer attempted to establish the historical priority of Mark, and the specific elaborations undertaken sequentially by Luke and Matthew. He depicted miracles as fallaciously displaying the immediate causality of the universal in nature, and criticized the naturalistic explanations favoured by theological rationalism. The third volume of the series denied the historicity of Christ. The Christian idea that God and mankind share the same essence appeared as the religious representation of a single empirical individual who assumed the universal power of spirit. Like his contemporaries D.F. Strauss and Ludwig Feuerbach, Bauer understood this synthesis instead as a project immanent in human history. As Bauer's political writings from this period show, he proposed that the assumption of universality, and the transcendence of particular interest, were historical tasks, undertaken by the state and the republican citizenry. In the Synoptics texts, Bauer equated Christianity and feudalism, and defended the freedom and equality of self-consciousness. Religion and the absolutist state were mutually sustaining, sharing the essential features of alienation and repression. Christianity represented the completion of the religious consciousness in pure abstraction, and the dissolution of all ethical bonds. Bauer contended that Judaism presupposed the subordination of nature to religious interests, but still maintained the natural links of kinship and ethnicity. Christianity eliminated this limited *Sittlichkeit* in favour of the purely abstract self, thus perfecting alienation and requiring its definitive resolution.

The political application of self-consciousness can be traced through two texts on the state, also dating from 1840-41. In *The Evangelical State Church of Prussia and Science*, Bauer described the essence of the state as free development. The state was the dialectical agency of historical progress and of the universality of the will, manifesting the capacity to abstract from any given content and express itself in ever new forms. While signalling empirical tendencies which might limit the state's progressive function (the prominence of religious interests, and hesitancy before the social question), Bauer contended that the

genuine state, as the expression of freedom, was in constant transformation. His surface claim was that the Prussian state is such an institution, though his contemporary correspondence belied this view. He defended the 1817 union of the Lutheran and Reformed churches in Prussia as the political overcoming of religious oppositions, whose basis had been eroded by the Enlightenment. Through its (still abstract) grasp of the universal concept of man, against religious particularity, the Enlightenment had transformed religious consciousness into self-consciousness. (This process formed one of the major themes of Bauer's *Christianity Revealed*, along with a critique of French materialism for its inadequate grasp of freedom). The churches were now impotent to perpetuate their own existence without the support of the state. Countering conservative historians like F.J. Stahl, who championed the independence of the churches, Bauer's "The Christian State and Our Times," of 1841, again identified the state as the focus of ethical life. Stahl's position implied a derogation of the spirituality of the state, presenting it an agency of external constraint, to uphold the orthodox ecclesiastical and political order against the flow of history, and to defend a social order governed by irrational privileges and immunities. Bauer denounced not only the Christian state of Friedrich Wilhelm IV, but also the formal *Rechtsstaat*, or liberal constitutionalism. For Bauer, both these positions defined freedom as private interest, religious or economic; but as particularity, these attitudes had to be purged away in the name of a new political order. Bauer maintained that Hegel's view of freedom as universality was far in advance of liberal views, even if the Philosophy of Right was inconsistent or incomplete. This was Bauer's provocative claim at the Welcker banquet of 1841. The elimination of egoistic atomism by moral self-consciousness was the pre-requisite for the republic, or the free state.

The anonymous *Trumpet of the Last Judgement*, or *Posaune* (November, 1841), and its sequel, *Hegel's Doctrine of Religion and Art* (1842), interpreted Hegel as sounding a call for revolution, to bring this state into being. Bauer claimed that the consequences of Hegel's system were the overthrow of church and state; and that Hegel's conservative critics were right to see him as the most dangerous adversary of the Restoration. Written ironically as pietistic denunciations, Bauer's two texts attributed to Hegel a theory of infinite self-consciousness, in which the concept of substance and a transcendent absolute were necessary but self-annulling illusions. Recapitulating the issue in his own voice in 1845, Bauer identified a tension in Hegel's thought between Spinoza and Fichte, between inert, undifferentiated substance and creative form. The *Posaune*, however, argued that the Spinozist moment, though necessary to Hegel's dialectic, was fully assimilated to infinite self-consciousness. In absolute spirit, properly understood, all religious pretensions dissipated, while the absolute itself dissolved into the critical activities of conscious individual subjects. Nothing transcendent remained. Yet, the *Posaune* recognized, Hegel also stressed the concept of substance. Its role had to be accounted for. In its apparent transcendence, substance disciplined the immediate, particular self. This was necessary because, as Hegel argued, particularity cannot be the criterion of theoretical or practical reason; rather, individuals must first internalize substance as a stage in reaching infinite self-consciousness. The undifferentiated, pure universal of substance subsumed all particularity, including the self. This initial, Spinozist moment created an appearance of pantheism in Hegel, which misled interpreters like D. F. Strauss. In Bauer's depiction, however, Hegel proceeded to dissolve substantiality as a power independent of consciousness. This dialectical resolution was not equivalent to renouncing objectivity, but meant that substance, once it had demonstrated to the particular consciousness the need to transcend itself, might not claim an immediate validity either. Forgoing immediacy to substantiality, individuals could then become the organs through which the universal

attained conscious form. By overcoming the dialectical illusion of substance, the unity of concept and objectivity could first be glimpsed. The subject must appear as potentially universal, and the objective must show itself as a purposive order, responding to the subject's striving for rational freedom. This development entailed transforming substance into the record of the acts of conscious spirit, an inner relation of self-consciousness to itself. Subjectivity thus assimilated the principle of universality, which it now contained as its own character, not as something alien to it. But this relation was not confined to an inward experience, since reason must realize itself in the world. The externalization of reason produced a historical sequence, including the forms of alienated life. The stages in this sequence could be grasped as moments in the unfolding unity of thought and being. Bauer described self-consciousness, conceived as an immanent and subjective universality, as the motive force of history, generating historical content by taking up and transforming the given. As Bauer's 1829 manuscript had declared, at stake was not only the subjective realisation of the concept, but the fate of the idea, the unity of thought and being; and this required that the objectivity of the historical process be equally emphasized. This historical and critical idealism, which the Posaune attributes to Hegel, was politically revolutionary: it affirmed the rights of free self-consciousness against any positive institution which could not justify its existence before rational thinking, against state, religion, and social hierarchy.

Bauer used his central concept of infinite self-consciousness, a term taken from Hegel's theory of subjective spirit, to reconfigure the Hegelian absolute, bringing art and philosophy into close proximity, and excluding religion as a form of alienated reason, while recognising its past historical necessity. Bauer insisted on the immanence of the universal in history, as the record of struggles for liberation, and of alienation, which was necessary to discover the meaning of rational autonomy. Bauer's ethical idealism resembles what Kant calls perfectionism, or *Vollkommenheit*, a form of rational heteronomy, one of whose meanings is that action is validated by its contribution to historical progress. Bauer equated perfectionism and autonomy, as an uncompromising commitment to remodel political and social relations and institutions. Subjects acquired autonomy by freeing themselves from particular interests, and by repudiating transcendent universals, religious and political institutions which claimed to be underivable from self-consciousness, and exempt from history. Bauer denounced the *ancien régime* and its Restoration surrogates as a feudal system of tutelage and irrational privileges. Arrogating universality to itself, the authoritarian state which arose over these exclusive particulars thwarted the self-activity of its people, and concealed the source of its authority behind a veil of religious sanctification. Bauer maintained that the state, and not religion, was the principal adversary. Against this order of alienated spirit, he insisted that the decisive political question was the source of the state's authority, whether in tradition and religious sanction, or in the popular will. This question was to be resolved without compromise. Bauer asserted that his objective was not merely political, but social emancipation. The social question, the polarizations and crises of civil society to which Hegel had been alert, could be resolved not by direct appeals to the particular interests of one class, but by a common republican struggle against multiform privilege. The result of this combat would be the attainment of justice in all spheres of social life.

Two texts, dating from late 1842 and early 1843, *The Jewish Question*, and "The Capacity of Present-Day Jews and Christians to Become Free," elaborate Bauer's critique of the religious consciousness and of political reformism. The consequence of their publication, however, was that Bauer forfeited his leading

position in the opposition movement, as he challenged one of its central demands. The question was whether the explicitly Christian state of Prussia could eliminate restrictions on Jewish participation in civil institutions. While liberals and republicans advocated emancipation, conservative opponents defended the state's exclusive confessional allegiance. Bauer's interventions attacked the state for defending privilege, and claimed that it used religion as a mask for its interests in maintaining relations of subordination; but he also criticized Jews and their supporters for claiming freedom on the basis of a particular identity. Political and social freedom required the renunciation of all particularistic ties with the past; thus, as a precondition of juridical equality, Jews must renounce their religious allegiance, as must Christians. Christianity demonstrated a historically higher degree of consciousness, since it cancelled the externality of the deity. But this was not a unilateral progress upon Judaism, because Christianity, and especially Protestantism, universalized alienation to encompass all aspects of life. The superiority of Christianity consisted in its radical negativity, making requisite a transition to a new and higher form of ethical life. By exacerbating the contradiction between self-determination and self-abasement, the way was cleared for an epochal resolution. These interventions were censured by Marx, and by leading liberal spokesmen. Bauer remained adamant that his position was the correct progressive stance.

In his studies of the French Revolution and its impact on Germany, Bauer traced the emergence of mass society, based on conformity and inchoate particularism. The dissolution of the feudal estates by the Revolution produced a purely atomistic society, characterized by the assertion of individual property right. The attachment to private economic interest made impossible a concerted opposition to privilege and to the existing order, and had caused the ultimate defeat of the revolutions that had spawned it. Jacobinism, which Bauer in many ways endorsed, had been directed against this attitude, but had failed to overcome it; and it now threatened the republican movement of the Vormärz. The masses, encompassing both the proletariat and the bourgeoisie, represented inertia and stagnation, and formed the bulwark of the existing order. Their opposition to it was merely apparent. Liberalism unconsciously expressed this development of mass society, defining freedom as acquisition. Bauer criticized liberal constitutionalism as a vacillating, compromising attitude toward the feudal regime. Even in its most advanced form, that endorsed by Hegel, constitutionalism juxtaposed two diametrically opposed principles of sovereignty, popular and princely, and was unable to resolve the contention between them. Incipient socialism shared the same terrain as liberalism, the defence of private interest, but proposed inconsequent and unacceptable solutions to the conditions which liberalism simply affirmed. For Bauer, socialism was irredeemably heteronomous. The socialist movement, he claimed, sought to organize the workers in their immediate, particular existence, and not to transform them. He saw in the proletariat pure particularity, and, unlike Marx, denied that this particularity could transform itself into a genuine universal unless it first renounced its own sectional interests. Bauer also anticipated the negative effects of a socialist organization of labour. While criticizing capitalism for its irrational competitive forms, he defended the principle of competition itself as a necessary condition for progress, the independence of persons, and the possibility of conscious, free self-determination. Bauer's pre-1848 work revived the classical republican themes of the opposition of commerce and virtue, but gave them a new shape, consistent with his Hegelianism. In 1842-43, Bauer confidently predicted the triumph of republican principles and institutions, though this confidence waned as the political crisis deepened. In his two electoral addresses of 1848-49, he defended popular sovereignty and the right of revolution, demanding that the new constitution be promulgated as an act of revolutionary will, and not received as a concession from the king.

3. Bauer’s Late Work, 1850-1882

While he continued to proclaim the continuity of his thought, Bauer’s late work was characterized by the definitive abandonment of his Vormärz republicanism. The failure of 1848, he argued, demonstrated the bankruptcy of the European philosophical tradition. Instead of the triumph of republics, Bauer now foresaw an age of global imperialism. The decisive political question after 1848 was the rise of Russia. Bauer predicted that Russian pressure would promote a pan-European union, as a stage in a movement toward a global absolutism. The revolutionaries of 1848 still presupposed, uncritically, that states were independent units. The next historical period would initiate a genuine continental crisis. Anticipating Nietzsche, Bauer contended that the impending collapse of European civilisation would make possible a new beginning, a liberation from traditional forms and values, together with their metaphysical and religious sanctions. Bauer’s abiding opposition to liberalism now induced him to collaborate in conservative causes; but his conservatism was unconventional. Like Nietzsche, he continued to repudiate tradition and religion. Because of his anti-Semitism, Bauer was claimed as a precursor by some National-Socialist authors, though Ernst Barnikol, for example, disputes a direct connection (Barnikol 1972, pp. 350-53).

For Bauer, the revolutions of 1848 were so closely connected with the Enlightenment, Kantian, and Hegelian projects that their failure sounded the death-knell of philosophy and its claims to rational individual autonomy. Bauer’s late critique assimilated Hegel with Spinoza and the metaphysics of substance, understood as the negation of form and subjectivity. Unlike his Vormärz position, he asserted in texts of 1852 and 1853 that Hegel had yielded to the influence of Spinoza, effacing individuality, and submerging concrete particulars under illusory, abstract logical categories. Bauer now described the Hegelian idea as a transcendent illusion. Its inability to admit concrete particulars derived from the substantiality of the system itself. The result was that Hegel had discounted individuality in favour of conformity. While prior to 1848 Bauer proclaimed that Hegel had taught "the republic and the revolution," he now decried the absolutist tendencies of the Hegelian system, whose oppressive unity paralleled the historical trend toward an all-encompassing political despotism. Bauer accused philosophy of contributing to an inexorable process of levelling and uniformity in the post-revolutionary state (Bauer, *Russland und das Germanenthum*, I, pp. 40-54). These criticisms anticipated Rudolph Haym’s polemic in *Hegel und seine Zeit* (1857).

In common with many post-1848 intellectuals, Bauer’s abandonment of metaphysics led him to a new conception of critique as a positive science or empirical investigation. Bauer no longer contended that history represents an unfolding dialectic of self-consciousness. Critique was to permit the observer to examine historical phenomena without distortion or partiality, and without an a priori systemic concern. Bauer maintained that scientific research must remain independent of ecclesiastical and political tutelage. Its objective was to determine the relation of nature to rights and freedom of the will (concepts which the late Bauer retained, while rejecting their metaphysical foundations); but critique did not enjoin practical intervention in political affairs. The correct stance was now disinterested contemplation of the inevitable processes of cultural decay and regeneration.

The conclusion of this new critique was that the future belonged not to the republican people, or to separate peoples, but to a transnational imperialism, involving the confrontation of two absolutist programmes. In one of these, the Western European, political absolutism arose over modern mass society as its necessary complement. Bauer had earlier criticized this configuration as an outmoded form of state, to be supplanted by the republic; it now described the result of an incomplete political development, which would issue in a contention for world domination. Within the Western European form, Bauer distinguished two variants: Bismarck's state socialism, imitating eighteenth-century Prussian militarism, attempted to subject economic production to political control, suppressing innovation and personal independence; Disraeli's romantic imperialism sought to level and subordinate English society before a paternalistic monarchy. In opposition to the west, the second major absolutist form was that of Russia, a substantial power with limited internal distinctions. Its cohesiveness derived from the fusion of political and ecclesiastical power, and the absence of the modern idea of subjectivity. Bauer noted that Hegel had mistakenly discounted this zone from world history. Like the anarchist Michael Bakunin, Bauer claimed that Russia owed its original state formation to Germany; but Russia had otherwise been impervious to western philosophical influence, adopting only what served its immediate, concrete ends. Animated by hatred and shame of its past insignificance, Russia too was ambivalent. It did not directly provide the solution to the contemporary political crisis, but elicited the decisive struggle with the west. The vigour of an alien adversary would force Europe to transform itself. This process involved the extension of imperialism across the continent and the globe, and the clash of rivals for dominance within the new empire. World war was inevitable.

Bauer's prognosis anticipated aspects of Karl Kautsky's 1915 theory of ultraimperialism, though without the latter's optimism that this trend heralded a reduction in conflicts among contenders for hegemony. Imperialism, moreover, did not stimulate, but hampered economic growth, since insecurity and permanent military mobilization undermined productive activity. The historic function of the globalizing process was to eliminate national identities, laying the basis for an eventual cosmopolitan rebirth. Bauer saw nationalism as a dissipated force. The emerging world order was framed not by the defence of national interests, but by a struggle for transnational supremacy among elites with no local loyalties. The growing centralization of political power was abetted by the levelling forces of the socialist movement, with its own internationalist pretensions. This trend also underlay what Bauer called political pauperism, a generalised disqualification of individuals from participation in political activity. The conclusion of this process would be to perfect mass society, which Bauer had analyzed since the 1840's. The principle of substance, non-differentiation, and conformity would reach its ultimate extension, and could then be overthrown. World imperialism would issue in an all-embracing catastrophe, the apocalyptic end of the old, Christian-Germanic order. Only then would new cultural possibilities emerge. Though these could not be predicted in detail, they would involve the emergence of an unprecedented creative individuality, freed from religious and metaphysical illusions.

Bauer likened the present crisis to the end of the classical world in Roman imperialism. His studies in the 1850's located the origins of Christianity in the second century A.D., concluding that the first gospel was written under Hadrian (117-138 AD), though slightly predated by some of the Pauline epistles. Bauer traced the evolution of Christian ideas from Hellenism and Stoicism, deriving the logos doctrine of John's gospel from Philo and neo-Platonic sources. As in *Herr Dr. Hengstenberg*, he denied that Christianity

had emerged directly from Judaism. More than in his early work, though, he now stressed the revolutionary power of the early Christian religion, as a source of liberation for the excluded and impoverished elements of the Roman Empire. His final book described Christianity as the socialist culmination of Greek and Roman history. Responding to this argument in his very positive obituary of Bauer, Friedrich Engels acknowledged the importance of Bauer's late work for the socialist critique of religion (*Sozialdemokrat*, 1882). In 1908, Karl Kautsky's book, *The Origins of Christianity*, applied Bauer's thesis.

Bauer's late writings identified sentiment and pietistic feeling-certainty, rather than autonomous reason, as the principal force in shaping modern subjectivity. His studies of the Quakers and of pietism described passive inwardness and feeling as the dominant characteristics of the German Enlightenment. The practical reason of Kant and Fichte merely translated the inner voice of pietist conscience into a rationalist idiom. Bauer also described pietism as the end of Christianity, since it destroyed dogma in favour of inner illumination and personal moral rectitude. Consistent with his *Christianity Revealed*, Bauer continued to define positive or statutory religions by their exclusive dogmas and symbols; and he still saw the general course of history as dissipating these dogmas as mere illusions. He discounted the mobilizing potential of religion in the modern imperial order. In the Posaune, he had denounced Schleiermacher's efforts to restore dogmatic Christianity through an appeal to feelings of dependency. Now he claimed that the force of sentiment, contrary to Schleiermacher's supposition, was to dissolve dogmatic religion into personal conviction. The new world empire would end with the inner erosion of religious belief. Not rational speculation, but sentiment, would effect this transformation.

A stringent anti-nationalism and a marked anti-Semitism characterized Bauer's later thought. He defended German culture against its political appropriation by the Prussian and Austrian regimes, but criticized its insufficiencies, in Goethe, for example, who remained enthralled to the metaphysical tradition. Bauer stressed that Germany was not a racial unit, but a historical and cultural artefact, reinforced by racial mixing, and not by racial purity (Barnikol 1972, p. 393). It is clear, however, that some elements were excluded from the mix: unlike his earlier treatment of the Jewish question as historical, cultural, and religious, he now asserted that a natural distinction of race created an impassable divide between Jews and Europeans (Bauer, "Present Position of the Jews," 1852). His claim that the political significance of the Jews throughout the political spectrum was a testimony to the debility of European culture and to the approaching crisis was greeted by National-Socialist authors.

Bauer's late work contains prescient observations on globalization and world war, and has affinities with a variety of twentieth-century ideological forms, from socialism to imperialism and anti-Semitism. In contrast, his early work bespeaks an original, Hegelian republicanism, and offers cogent analyses of Restoration political thought and the rise of mass society. His intellectual legacy is complex and contentious.

Bibliography

Major Works by Bruno Bauer, 1829-1882

- *De pulchri principiis*. Prussian royal prize manuscript, University of Berlin,1829. First published as *Über die Prinzipien des Schönen. De pulchri principiis. Eine Preisschrift*, hrsg. Douglas Moggach und Winfried Schultze, mit einem Vorwort von Volker Gerhardt (Berlin: Akademie Verlag, 1996)
- "Rezension: *Das Leben Jesu, kritisch bearbeitet* von David Friedrich Strauss," *Jahrbücher für wissenschaftliche Kritik*, Dec. 1835, no. 109, 879-880; no. 111, 891; no. 113, 905-912. May 1836, no. 86, 681-688; no. 88, 697-704.
- *Kritik der Geschichte der Offenbarung. Die Religion des alten Testaments in der geschichtlichen Entwicklung ihrer Prinzipien dargestellt*, 2 vol. (Berlin, 1838).
- *Herr Dr. Hengstenberg* (Berlin, 1839)
- (anon., 1st ed.) *Die evangelische Landeskirche Preußens und die Wissenschaft* (Leipzig, 1840); second edition, with author indicated, 1840.
- *Kritik der evangelischen Geschichte des Johannes* (Bremen, 1840).
- "Der christliche Staat und unsere Zeit," *Hallische Jahrbücher für deutsche Wissenschaft und Kunst*, 7 - 12 June 1841, no 135-140, pp. 537-558.
- *Kritik der evangelischen Geschichte der Synoptiker*, 2 vol. (Leipzig, 1841); *Kritik der evangelischen Geschichte der Synoptiker und des Johannes, Dritter und letzter Band* (Braunschweig, 1842)
- (anon.) *Die Posaune des jüngsten Gerichts über Hegel den Atheisten und Antichristen. Ein Ultimatum* (Leipzig,1841); *The Trumpet of the Last Judgement against Hegel the Atheist and Antichrist. An Ultimatum*, trans. L. Stepelevich (Lewiston, N.Y.: E. Mellen Press, 1989)
- (anon.) *Hegels Lehre von der Religion und Kunst von dem Standpuncte des Glaubens aus beurtheilt* (Leipzig, 1842); new edition Aalen: Scientia Verlag, 1967
- *Die gute Sache der Freiheit und meine eigene Angelegenheit* (Zürich und Winterthur, 1842)
- *Die Judenfrage* (Braunschweig, 1843).
- articles in Arnold Ruge (ed.), *Anekdota zur neuesten deutschen Philosophie und Publizistik*, vol. 2 (Zürich und Winterthur, 1843): "Leiden und Freuden des theologischen Bewußtseins", 89-112; "Rezension: 'Bremisches Magazin für evangelische Wahrheit gegenüber dem modernen Pietismus. Erstes Heft'", 113-134; "Rezension: *Einleitung in die Dogmengeschichte* von Theodor Kliefoth."135-159; "Rezension: *Die Geschichte des Lebens Jesu mit steter Rücksicht auf die vorhandenen Quellen* dargestellt von Dr. von Ammon. Leipzig, 1842" 160-185.
- *Das entdeckte Christenthum. Eine Erinnerung an das 18. Jahrhundert und ein Beitrag zur Krisis des 19*. (Zürich und Winterthur, 1843)
- "Die Fähigkeit der heutigen Juden und Christen, frei zu werden," in Georg Herwegh (ed.), *Einundzwanzig Bogen aus der Schweiz* (Zürich und Winterthur, 1843), 56-71.
- *Geschichte der Politik, Kultur und AufklÄrung des achtzehnten Jahrhunderts*, 4 vols. (Charlottenburg: 1843-45).
- *Denkwürdigkeiten zur Geschichte der neueren Zeit seit der französischen Revolution* (Charlottenburg, 1843-1844).
- *Briefwechsel zwischen Bruno Bauer und Edgar Bauer während der Jahre 1839-1842 aus Bonn und Berlin* (Charlottenburg, 1844).
- "Was ist jetzt der Gegenstand der Kritik?" *Allg. Lit-Ztg*. VIII, July 1844, 18-26.

- "Die Gattung und die Masse", *Allg. Lit.-Ztg.* X, September 1844, 42-48.
- *Acktenstücke zu den Verhandlungen über die Beschlagnahme der "Geschichte der Politik, Kultur und Aufklärung des achtzehnten Jahrhunderts",* von Bruno Bauer. Teil I herausgegeben von Bruno Bauer (Christiania: Verlag von C.C. Werner, 1844).
- "Charakteristik Ludwig Feuerbachs," *Wigands Vierteljahrschrift* III, 1845, 86-146.
- *Geschichte Deutschlands und der französischen Revolution unter der Herrschaft Napoleons*, 2 vols. (Charlottenburg, 1846)
- *Vollständige Geschichte der Parteikämpfe in Deutschland während der Jahre 1842-1846* (Charlottenburg, 1847).
- *Die bürgerliche Revolution in Deutschland seit dem Anfange der deutschkatholischen Bewegung* (Charlottenburg, 1847)
- "Erste Wahlrede von 1848," and "Verteidigungsrede Bruno Bauers vor den Wahlmännern des Vierten Wahlbezirkes am 22.2. 1849," in E. Barnikol, *Bruno Bauer: Studien und Materialien*, 518-531.
- *Untergang des Frankfurter Parlaments* (Charlottenburg, 1849).
- *Kritik der paulinischen Briefe* (Berlin, 1850-1851).
- *Kritik der Evangelien und Geschichte ihres Ursprungs*, 3 vol. (Berlin, 1850-1851); 4th vol. under the title *Die theologische Erklärung der Evangelien* (Berlin, 1852).
- "The Present Position of the Jews," *New York Daily Tribune*, June 7, 1852.
- *Russland und das Germanenthum*, 2 vol. (Charlottenburg, 1853).
- *De la dictature occidentale* (Charlottenburg, 1854).
- *Deutschland und das Russenthum* (Charlottenburg, 1854).
- *Die russische Kirche. Schlussheft* (Charlottenburg, 1855).
- *Das Judenthum in der Fremde. Separat-Abdruck aus dem Wagener'schen Staats- und Gesellschaftslexikon* (Berlin, 1863).
- *Freimaurer, Jesuiten und Illuminaten in ihrem geschichtlichen Zusammenhange* (Berlin, 1863).
- *Philo, Strauss und Renan und das Urchristenthum* (Berlin, 1874).
- *Einfluss des englischen Quäkerthums auf die deutsche Cultur und auf das englisch-russische Project einer Weltkirche* (Berlin, 1878).
- *Christus und die Cäsaren. Der Ursprung des Christenthums aus dem römischen Griechenthum* (Berlin, 1879).
- *Zur Orientierung über die Bismarck'sche Ära* (Chemnitz, 1880).
- *Disraelis romantischer und Bismarcks socialistischer Imperialismus* (Chemnitz,1882).

Journals edited by Bruno Bauer

- *Zeitschrift für spekulative Theologie* (Berlin, three volumes,1836-1838).
- *Allgemeine Literatur-Zeitung* (Charlottenburg, 12 issues December 1843-October 1844). Second edition under the title: *Streit der Kritik mit den modernen Gegensätzen* (Charlottenburg, 1847).
- *Norddeutsche Blätter. Eine Monatschrift für Kritik, Literatur und Unterhaltung* (Charlottenburg. 10 issues July 1844 - April 1845). Second edition under the title: *Beiträge zum Feldzuge der Kritik. Norddeutsche Blätter für 1844 und 1845* (Berlin, 1846).
- collaboration in Friedrich Wilhelm Hermann Wagener, ed., *Neues Conversations-Lexikon. Staats-*

und Gesellschafts-Lexikon, 23 vol. (Berlin, 1859-1867).
- collaboration in *Schmeitzner's Internationale Monatsschrift. Zeitschrift für allgemeine nationale Kultur und deren Literatur* (Chemnitz, 1882).

Secondary Sources

Books and Articles

- Barnikol, Ernst, *Bruno Bauer, Studien und Materialien*, aus dem Nachlass ausgewählt und zusammengestellt von P. Riemer und H.-M. Sass (Assen: van Gorcum, 1972).
- Brazill, W.J., *The Young Hegelians* (New Haven: Yale University Press, 1970).
- Brudney, Daniel, *Marx's Attempt to Leave Philosophy* (Cambridge, MA: Harvard University Press, 1998).
- Cesa, Claudio, *Studi sulla Sinistra hegeliana* (Urbino: Argalia, 1972).
- Engels, Friedrich, "Bruno Bauer und das Urchristentum," *Sozialdemokrat*, May 4 and 11, 1882.
- Eßbach, Wolfgang, *Die Junghegelianer*. *Soziologie einer Intellektuellengruppe* (München: Wilhelm Fink Verlag, 1988).
- Hertz-Eichenrode, Dieter, *Der Junghegelianer Bruno Bauer im Vormärz*. Inauguraldissertation (Berlin: Freie Universität, 1959).
- Kautsky, Karl, *Der Ursprung des Christentums* (Stuttgart: Dietz, 1908).
- Kautsky, Karl, *Nationalstaat, imperialistischer Staat und Staatenbund* (Nürnberg: Fränkische Verlagsanstalt, 1915.
- Leopold, David, "The Hegelian Antisemitism of Bruno Bauer," *History of European Ideas* 25 (1999), 179-206.
- Löwith, Karl, *From Hegel to Nietzsche* (Garden City: Doubleday, 1967).
- Mah, Harold, *The End of Philosophy and the Origin of Ideology*. *Karl Marx and the Crisis of the Young Hegelians* (Berkeley: University of California Press, 1987).
- Marx, Karl, "On the Jewish Question," *Collected Works*, vol. 3 (New York: International Publishers, 1975), 146-74.
- Marx, Karl, Frederick Engels, "The Holy Family, or Critique of Critical Criticism," *Collected Works*, vol. 4 (New York: International Publishers, 1975), 5-211; "The German Ideology," *Collected Works*, vol. 5 (New York: International Publishers, 1976), 19-539.
- Mayer, Gustav, "Die Anfänge des politischen Radikalismus im vormärzlichen Preußen," *Zeitschrift für Politik* (1913), Heft 1, Sonderdruck, 1-113.
- McLellan, David, *The Young Hegelians and Karl Marx* (Toronto: Macmillan, 1969).
- Moggach, Douglas, *The Philosophy and Politics of Bruno Bauer* (Cambridge: Cambridge University Press, forthcoming 2002).
- Peled, Yoav, "From Theology to Sociology: Bruno Bauer and Karl Marx on the Question of Jewish Emancipation," *History of Political Thought* 13/3 (1992), 463-85.
- Pepperle, Ingrid, *Junghegelianische Geschichtsphilosophie und Kunsttheorie* (Berlin: Akademie Verlag, 1978).
- Rambaldi, Enrico, *Le origini della sinistra hegeliana* (Florence: Nuova Italia,1966).
- Rosen, Zvi, *Bruno Bauer and Karl Marx* (the Hague: Nijhoff, 1978).

- Rossi, Mario, *Da Hegel a Marx III: La Scuola Hegeliana. Il giovane Marx*, 2nd edition (Milan: Feltrinelli, 1974).
- Sass, Hans-Martin, "Bruno Bauers Idee der Rheinischen Zeitung", *Zeitschrift für Religions- und Geistesgeschichte* 19 (1967), 221-276.
- Sass, Hans-Martin, "Bruno Bauer's Critical Theory," *Philosophical Forum* 8 (1978), 93-103.
- Schweitzer, Albert, *The Quest of the Historical Jesus. A Critical Study of its Progress from Reimarus to Wrede* (Baltimore: Johns Hopkins University Press, 1998).
- Schläger, Eduard, "Bruno Bauer und seine Werke," *Schmeitzner's Internationale Monatsschrift*, vol. 1 (Chemnitz, 1882), 377-400.
- Stepelevich, L.S., ed., *The Young Hegelians, An Anthology* (Cambridge: Cambridge University Press, 1983).
- Stuke, Horst, *Philosophie der Tat, Studien zur 'Verwirklichung der Philosophie' bei den Junghegelianern und den Wahren Sozialisten* (Stuttgart: Ernst Klett Verlag, 1963).
- Toews, J.E., *Hegelianism. The Path toward Dialectical Humanism* (Cambridge: Cambridge University Press, 1980).
- van den Bergh van Eysinga, G.A., "Aus einer unveröffentlichten Biographie von Bruno Bauer. Bruno Bauer in Bonn 1839-1842," *Annali Feltrinelli* (1963), 329-386.
- Waser, Ruedi, *Autonomie des Selbstbewußtseins. Eine Untersuchung zum Verhältnis von Bruno Bauer und Karl Marx (1835-1843)* (Tübingen: Francke Verlag, 1994).
- Zanardo, Aldo, "Bruno Bauer hegeliano e giovane hegeliano," *Rivista critica di storia della filosofia*, 1965, 1-57.

Unpublished Manuscripts

- Barnikol, Ernst, *Bruno Bauer. Darstellung und Quellen*, ca. 1965 (International Institute for Social History, Amsterdam).
- van der Bergh van Eysinga, Gustaaf Adolf, *Bruno Bauer. Sein Leben und seine theologische Bedeutung* (International Institute for Social History, Amsterdam).

Other Internet Resources

- Hegel Society of America

Related Entries

Hegel, Georg Wilhelm Friedrich | idealism | Kant, Immanuel

Douglas Moggach
dmoggach@ottawa.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 7, 2002
Content last modified: March 7, 2002

Conventionality of Simultaneity

In his first paper on the special theory of relativity, Einstein indicated that the question of whether or not two spatially separated events were simultaneous did not necessarily have a definite answer, but instead depended on the adoption of a convention for its resolution. Some later writers have argued that Einstein's choice of a convention is, in fact, the only possible choice within the framework of special relativistic physics, while others have maintained that alternative choices, although perhaps less convenient, are indeed possible.

The Conventionality Thesis

Prior to Einstein's first paper on relativity (Einstein, 1905), it was generally agreed that simultaneity was absolute; i.e., that there was a unique event at location A that was simultaneous with a given event at location B. Einstein, however, said that it was necessary to make an assumption in order to be able to compare the times of occurrence of events at spatially separated locations (Einstein, 1905, pp. 38-40 of the Dover translation or pp. 125-127 of the Princeton translation; but note Scribner, 1963, for correction of an error in the Dover translation). His assumption, which defined what is usually called standard synchrony, can be described in terms of the following idealized thought experiment, where the spatial locations A and B are fixed locations in some particular, but arbitrary, inertial (i.e., unaccelerated) frame of reference: Let a light ray, traveling in vacuum, leave A at time t_1 (as measured by a clock at rest there), and arrive at B coincident with the event E at B. Let the ray be instantaneously reflected back to A, arriving at time t_2. Then standard synchrony is defined by saying that E is simultaneous with the event at A that occurred at time $(t_1 + t_2)/2$. This definition is equivalent to the requirement that the one-way

speeds of the ray be the same on the two segments of its round-trip journey between A and B.

The thesis that the choice of standard synchrony is a convention, rather than one necessitated by facts about the physical universe (within the framework of the special theory of relativity), has been argued particularly by Reichenbach (see, for example, Reichenbach, 1958, pp. 123-135) and Grünbaum (see, for example, Grünbaum, 1973, pp. 342-368). They argue that the only nonconventional basis for claiming that two distinct events are not simultaneous would be the possibility of a causal influence connecting the events. In the pre-Einsteinian view of the universe, there was no reason to rule out the possibility of arbitrarily fast causal influences, which would then be able to single out a unique event at A that would be simultaneous with E. In an Einsteinian universe, however, no causal influence can travel faster than the speed of light in vacuum, so from the point of view of Reichenbach and Grünbaum, any event at A whose time of occurrence is in the open interval between t_1 and t_2 could be defined to be simultaneous with E. In terms of the ε-notation introduced by Reichenbach, any event at A occurring at a time $t_1 + \varepsilon(t_2 - t_1)$, where $0 < \varepsilon < 1$, could be simultaneous with E. That is, the conventionality thesis asserts that any particular choice of ε within its stated range is a matter of convention, including the choice $\varepsilon=1/2$ (which corresponds to standard synchrony). If ε differs from 1/2, the one-way speeds of a light ray would differ (in an ε-dependent fashion) on the two segments of its round-trip journey between A and B. If, more generally, we consider light traveling on an arbitrary closed path in three-dimensional space, then (as shown by Minguzzi, 2002, pp.155-156) the freedom of choice in the one-way speeds of light amounts to the choice of an arbitrary scalar field (although two scalar fields that differ only by an additive constant would give the same assignment of one-way speeds).

It might be argued that the definition of standard synchrony makes use only of the relation of equality (of the one-way speeds of light in different directions), so that simplicity dictates its choice rather than a choice that requires the specification of a particular value for a parameter. Grünbaum (1973, p. 356) rejects this argument on the grounds that, since the equality of the one-way speeds of light is a convention, this choice does not simplify the postulational basis of the theory but only gives a symbolically simpler representation.

Phenomenological Counterarguments

Many of the arguments against the conventionality thesis make use of particular physical phenomena, together with the laws of physics, to establish simultaneity (or, equivalently, to measure the one-way speed of light). Salmon (1977), for example, discusses a number of such schemes and argues that each makes use of a nontrivial convention. For instance, one such scheme uses the law of conservation of momentum to conclude that two particles of equal mass, initially located halfway between A and B and then separated by an explosion, must arrive at A and B simultaneously. Salmon (1977, p. 273) argues, however, that the standard formulation of the law of conservation of momentum makes use of the concept of one-way velocities, which cannot be measured without the use of (something equivalent to) synchronized clocks at the two ends of the spatial interval that is traversed; thus, it is a circular argument to use conservation of momentum to define simultaneity.

It has been argued (see, for example, Janis, 1983, pp. 103-105, and Norton, 1986, p. 119) that all such schemes for establishing convention-free synchrony must fail. The argument can be summarized as follows: Suppose that clocks are set in standard synchrony, and consider the detailed space-time description of the proposed synchronization procedure that would be obtained with the use of such clocks. Next suppose that the clocks are reset in some nonstandard fashion (consistent with the causal order of events), and consider the description of the same sequence of events that would be obtained with the use of the reset clocks. In such a description, familiar laws may take unfamiliar forms, as in the case of the law of conservation of momentum in the example mentioned above. Indeed, all of special relativity has been reformulated (in an unfamiliar form) in terms of nonstandard synchronies (Winnie, 1970a and 1970b). Since the proposed synchronization procedure can itself be described in terms of a nonstandard synchrony, the scheme cannot describe a sequence of events that is incompatible with nonstandard synchrony. A comparison of the two descriptions makes clear what hidden assumptions in the scheme are equivalent to standard synchrony.

Transport of Clocks

A phenomenological scheme that deserves special mention, because of the amount of attention it has received over the course of many years, is to define synchrony by the use of clocks transported between locations A and B in the limit of zero velocity. Eddington (1924, p. 15) discusses this method of synchrony, and notes that it leads to the same results as those obtained by the use of electromagnetic signals (the method that has been referred to here as standard synchrony). He comments on both of these methods as follows (1924, pp. 15-16): "We can scarcely consider that either of these methods of comparing time at different places is an essential part of our primitive notion of time in the same way that measurement at one place by a cyclic mechanism is; therefore they are best regarded as conventional."

One objection to the use of the slow-transport scheme to synchronize clocks is that, until the clocks are synchronized, there is no way of measuring the one-way velocity of the transported clock. Bridgman (1962, p. 26) uses the "self-measured" velocity, determined by using the transported clock to measure the time interval, to avoid this problem. Using this meaning of velocity, he suggests (1962, pp. 64-67) a modified procedure that is equivalent to Eddington’s, but does not require having started in the infinite past. Bridgman would transport a number of clocks from A to B at various velocities; the readings of these clocks at B would differ. He would then pick one clock, say the one whose velocity was the smallest, and find the differences between its reading and the readings of the other clocks. Finally, he would plot these differences against the velocities of the corresponding clocks, and extrapolate to zero velocity. Like Eddington, Bridgman does not see this scheme as contradicting the conventionality thesis. He says (1962, p.66), "What becomes of Einstein’s insistence that his method for setting distant clocks -- that is, choosing the value 1/2 for ε -- constituted a ‘definition’ of distant simultaneity? It seems to me that Einstein’s remark is by no means invalidated."

Ellis and Bowman (1967) take a different point of view. Their means of synchronizing clocks by slow

transport (1967, pp. 129-130) is again somewhat different from, but equivalent to, those already mentioned. They would place clocks at A and B with arbitrary settings. They would then place a third clock at A and synchronize it with the one already there. Next they would move this third clock to B with a velocity they refer to as the "intervening ‘velocity’", determined by using the clocks in place at A and B to measure the time interval. They would repeat this procedure with decreasing velocities and extrapolate to find the zero-velocity limit of the difference between the readings of the clock at B and the transported clock. Finally, they would set the clock at B back by this limiting amount. On the basis of their analysis of this procedure, they argue that, although consistent nonstandard synchronization appears to be possible, there are good physical reasons (assuming the correctness of empirical predictions of the special theory of relativity) for preferring standard synchrony. Their conclusion (as summarized in the abstract of their 1967, p. 116) is, "The thesis of the conventionality of distant simultaneity espoused particularly by Reichenbach and Grünbaum is thus either trivialized or refuted."

A number of responses to these views of Ellis and Bowman (see, for example, Grünbaum et al., 1969; Winnie, 1970b, pp. 223-228; and Redhead, 1993, pp. 111-113) argue that nontrivial conventions are implicit in the choice to synchronize clocks by the slow-transport method. For example, Grünbaum (Grünbaum et al., 1969, pp. 5-43) argues that it is a nontrivial convention to equate the time interval measured by the infinitely slowly moving clock traveling from A to B with the interval measured by the clock remaining at A and in standard synchrony with that at B, and the conclusion of van Fraassen (Grünbaum et al., 1969, p. 73) is, "Ellis and Bowman have not proved that the standard simultaneity relation is nonconventional, which it is not, but have succeeded in exhibiting some *alternative conventions* which also yield that simultaneity relation." Winnie (1970b), using his reformulation of special relativity in terms of arbitrary synchrony, shows explicitly that synchrony by slow-clock transport agrees with synchrony by the standard light-signal method when both are described in terms of an arbitrary value of ε within the range $0 < \varepsilon < 1$, and argues that Ellis and Bowman err in having assumed the $\varepsilon=1/2$ form of the time-dilation formula in their arguments. He concludes (Winnie, 1970b, p. 228) that "it is not possible that the method of slow-transport, or any other synchrony method, could, within the framework of the *nonconventional* ingredients of the Special Theory, result in fixing *any* particular value of ε to the exclusion of any other particular values." Redhead (1993) also argues that slow transport of clocks fails to give a convention-free definition of simultaneity. He says (1993, p. 112), "There is no absolute factual sense in the term ‘slow.’ If we estimate ‘slow’ relative to a moving frame K’, then slow-clock-transport will pick out standard synchrony in K’, but this ... corresponds to nonstandard synchrony in K."

An alternative clock-transport scheme, which avoids the issue of slowness, is to have the clock move from A to B and back again (along straight paths in each direction) with the same self-measured speed throughout the round trip (Mamone Capria, 2001, pp. 812-813; as Mamone Capria notes, his scheme is similar to those propoosed by Brehme, 1985, pp. 57-58, and 1988, pp. 811-812). If the moving clock leaves A at time t_1 (as measured by a clock at rest there), arrives at B coincident with the event E at B, and arrives back at A at the time t_2, then standard synchrony is obtained by saying that E is simultaneous with the event at A that occurred at the time $(t_1 + t_2)/2$. It would seem that this transport scheme is sufficiently similar to the slow-transport scheme that it could engender much the same debate, apart from

those aspects of the debate that focussed specifically on the issue of slowness.

Malament's Theorem

An entirely different sort of argument against the conventionality thesis has been given by Malament (1977), who argues that standard synchrony is the only simultaneity relation that can be defined, relative to a given inertial frame, from the relation of (symmetric) causal connectibility. Let this relation be represented by κ, let the statement that events p and q are simultaneous be represented by S(p,q), and let the given inertial frame be specified by the world line, O, of some inertial observer. Then Malament's uniqueness theorem shows that if S is definable from κ and O, if it is an equivalence relation, if points p on O and q not on O exist such that S(p,q) holds, and if S is not the universal relation (which holds for all points), then S is the relation of standard synchrony.

Some commentators have taken Malament's theorem to have settled the debate on the side of nonconventionality. For example, Torretti (1983, p. 229) says, "Malament proved that simultaneity by standard synchronism in an inertial frame F is the *only* non-universal equivalence between events at different points of F that is definable ('in any sense of "definable" no matter how weak') in terms of causal connectibility alone, for a given F"; and Norton (Salmon et al., 1992, p. 222) says, "Contrary to most expectations, [Malament] was able to prove that the central claim about simultaneity of the causal theorists of time was false. He showed that the standard simultaneity relation was the only nontrivial simultaneity relation definable in terms of the causal structure of a Minkowski spacetime of special relativity."

Other commentators disagree with such arguments, however. Grünbaum (as reported by Norton in Salmon et al., 1992, p. 226) and Redhead (1993, p.114) cite Malament's need to postulate that S is an equivalence relation as a weakness in the argument. Havas (1987, p. 444) says, "What Malament has shown, in fact, is that in Minkowski space-time ... one can always introduce time-orthogonal coordinates ..., an obvious and well-known result which implies ε=1/2." Janis (1983, pp. 107-109) argues that Malament's theorem leads to a unique (but different) synchrony relative to any inertial observer, that this latitude is the same as that in introducing Reichenbach's ε, and thus Malament's theorem should carry neither more nor less weight against the conventionality thesis than the argument (mentioned above in the last paragraph of the first section of this article) that standard synchrony is the simplest choice. Similarly, Redhead (1993, p. 114) says that "we can use the same argument as we did for slow-clock-transport to demonstrate that we are faced with a conventional choice betweeen standard synchronies defined à la Malament in all possible inertial frames." In a comprehensive review of the problem of the conventionality of simultaneity, Anderson, Vetharaniam, and Stedman (1998, pp. 124-125) claim that Malament's proof is erroneous. Although they appear to be wrong in this claim, the nature of their error highlights the fact that Malament's proof, which uses the time-symmetric relation κ, would not be valid if a temporal orientation were introduced into space-time (see, for example, Spirtes, 1981, Ch. VI, Sec. F; and Stein, 1991, p. 153n).

Sarkar and Stachel (1999) argue that there is no physical warrant for the requirement that a simultaneity

relation be invariant under temporal reflections. Dropping that requirement, they show that Malament's other criteria for a simultaneity relation are then also satisfied if we fix some arbitrary event in space-time and say either that any pair of events on its backward null cone are simultaneous or, alternatively, that any pair of events on its forward null cone are simultaneous. They show further that, among the relations satisfying these requirements, standard synchrony is the unique such relation that is independent of the position of an observer and the half-null-cone relations are the unique such relations that are independent of the motion of an observer. If the backward-cone relation were chosen, then simultaneous events would be those seen simultaneously by an observer at the cone's vertex. As Sarkar and Stachel (1999, p. 209) note, Einstein (1905, p. 39 of the Dover translation or p. 126 of the Princeton translation) considered this possibility and rejected it because of its dependence on the position of the observer. Since the half-null-cone relations define causally connectible events to be simultaneous, it would seem that they would also be rejected by adherents of the views of Reichenbach and Grünbaum.

Giulini (2001, p.653) argues that it is too strong a requirement to ask that a simultaneity relation be invariant under causal transformations (such as scale transformations) that are not physical symmetries, which Malament as well as Sarkar and Stachel do. Using "Aut" to refer to the appropriate invariance group and "nontrivial" to refer to an equivalence relation on spacetime that is neither one in which all points are in the same equivalence class nor one in which each point is in a different equivalence class, Giulini (2001, pp. 657-658) defines two types of simultaneity: Absolute simultaneity is a nontrivial Aut-invariant equivalence relation on spacetime such that each equivalence class intersects any physically realizable timelike trajectory in at most one point, and simultaneity relative to some structure X in spacetime (for Malament, X is the world line of an inertial observer) is a nontrivial Aut_X-invariant equivalence relation on spacetime such that each equivalence class intersects any physically realizable timelike trajectory in at most one point, where Aut_X is the subgroup of Aut that preserves X. First taking Aut to be the inhomogeneous (i.e., including translations) Galilean transformations, Giulini (2001, pp. 660-662) shows that standard Galilean (i.e., pre-relativistic) simultaneity is the unique absolute simultaneity relation. Then taking Aut to be the inhomogeneous Lorentz tansformations (also known as the Poincaré transformations), Giulini (2001, pp. 664-666) shows that there is no absolute simultaneity relation and that standard Einsteinian synchrony is the unique relative simultaneity when X is taken to be a foliation of spacetime by straight lines (thus, like Malament, singling out a specific inertial frame, but in a way that is different from Malament's choice of X).

Other Considerations

Since the conventionality thesis rests upon the existence of a fastest causal signal, the existence of arbitrarily fast causal signals would undermine the thesis. If we leave aside the question of causality, for the moment, the possibility of particles (called tachyons) moving with arbitrarily high velocities is consistent with the mathematical formalism of special relativity (see, for example, Feinberg, 1967). Just as the speed of light in vacuum is an upper limit to the possible speeds of ordinary particles (sometimes called bradyons), it would be a lower limit to the speeds of tachyons. When a transformation is made to a different inertial frame of reference, the speeds of both bradyons and tachyons change (the speed of light in vacuum being the only invariant speed). At any instant, the speed of a bradyon can be transformed to

zero and the speed of a tachyon can be transformed to an infinite value. The statement that a bradyon is moving forward in time remains true in every inertial frame (if it is true in one), but this is not so for tachyons. Feinberg (1967) argues that this does not lead to violations of causality through the exchange of tachyons between two uniformly moving observers because of ambiguities in the interpretation of the behavior of tachyon emitters and absorbers, whose roles can change from one to the other under the transformation between inertial frames. He claims to resolve putative causal anomalies by adopting the convention that each observer describes the motion of each tachyon interacting with that observer's apparatus in such a way as to make the tachyon move forward in time. However, all of Feinberg's examples involve motion in only one spatial dimension. Pirani (1970) has given an explicit two-dimensional example in which Feinberg's convention is satisfied but a tachyon signal is emitted by an observer and returned to that observer at an earlier time, thus leading to possible causal anomalies.

A claim that no value of ε other than 1/2 is mathematically possible has been put forward by Zangari (1994). He argues that spin-1/2 particles (e.g., electrons) must be represented mathematically by what are known as complex spinors, and that the transformation properties of these spinors are not consistent with the introduction of nonstandard coordinates (corresponding to values of ε other than 1/2). Gunn and Vetharaniam (1995), however, present a derivation of the Dirac equation (the fundamental equation describing spin-1/2 particles) using coordinates that are consistent with arbitrary synchrony. They argue that Zangari mistakenly required a particular representation of space-time points as the only one consistent with the spinorial description of spin-1/2 particles.

The debate about conventionality of simultaneity seems far from settled, although some proponents on both sides of the argument might disagree with that statement. The reader wishing to pursue the matter further should consult the sources listed below as well as additional references cited in those sources.

Bibliography

- Anderson, R., Vetharaniam, I., and Stedman, G. 1998. "Conventionality of Synchronisation, Gauge Dependence and Test Theories of Relativity," *Physics Reports* **295**, 93-180.
- Brehme, R. 1985. "Response to 'The Conventionality of Synchronization'," *American Journal of Physics* **53**, 56-59.
- Brehme, R. 1988. "On the Physical Reality of the Isotropic Speed of Light," *American Journal of Physics* **56**, 811-813.
- Bridgman, P. 1962. *A Sophisticate's Primer of Relativity*. Middletown: Wesleyan University Press.
- Eddington, A. 1924. *The Mathematical Theory of Relativity*, 2nd ed. Cambridge: Cambridge University Press.
- Einstein, A. 1905. "Zur Elektrodynamik bewegter Körper," *Annalen der Physik* **17**, 891-921. English translations in *The Principle of Relativity*, pp. 35-65. New York: Dover, 1952; and in J. Stachel, ed., *Einstein's Miraculous Year*, pp. 123-160. Princeton: Princeton University Press, 1998.
- Ellis, B. and Bowman, P. 1967. "Conventionality in Distant Simultaneity," *Philosophy of Science*

34, 116-136.

- Feinberg, G. 1967. "Possibility of Faster-Than-Light Particles," *Physical Review* **159**, 1089-1105.
- Giulini, D. 2001. "Uniqueness of Simultaneity," *British Journal for the Philosophy of Science* **52**, 651-670.
- Grünbaum, A. 1973. *Philosophical Problems of Space and Time*, 2nd, enlarged ed. (*Boston Studies in the Philosophy of Science*, vol. 12). Dordrecht/Boston: D. Reidel.
- Grünbaum, A., Salmon, W., van Fraassen, B., and Janis, A. 1969. "A Panel Discussion of Simultaneity by Slow Clock Transport in the Special and General Theories of Relativity," *Philosophy of Science* **36**, 1-81.
- Gunn, D. and Vetharaniam, I. 1995. "Relativistic Quantum Mechanics and the Conventionality of Simultaneity," *Philosophy of Science* **62**, 599-608.
- Havas, P. 1987. "Simultaneity, Conventionalism, General Covariance, and the Special Theory of Relativity," *General Relativity and Gravitation* **19**, 435-453.
- Janis, A. 1983. "Simultaneity and Conventionality," in R. Cohen and L. Laudan, eds., *Physics, Philosophy and Psychoanalysis* (*Boston Studies in the Philosophy of Science*, vol. 76), pp. 101-110. Dordrecht/Boston: D. Reidel.
- Malament, D. 1977. "Causal Theories of Time and the Conventionality of Simultaniety," *Noûs* **11**, 293-300.
- Mamone Capria, M. 2001. "On the Conventionality of Simultaneity in Special Relativity," *Foundations of Physics* **31**, 775-818.
- Minguzzi, E. 2002. "On the Conventionality of Simultaneity," *Foundations of Physics Letters* **15**, 153-169.
- Norton, J. 1986. "The Quest for the One Way Velocity of Light," *British Journal for the Philosophy of Science* **37**, 118-120.
- Pirani, F. 1970. "Noncausal Behavior of Classical Tachyons," *Physical Review* **D1**, 3224-3225.
- Redhead M. 1993. "The Conventionality of Simultaneity," in J. Earman, A. Janis, G. Massey, and N. Rescher, eds., *Philosophical Problems of the Internal and External Worlds*, pp. 103-128. Pittsburgh: University of Pittsburgh Press; Konstanz: Universitätsverlag Konstanz.
- Reichenbach H. 1958. *The Philosophy of Space & Time*. New York: Dover.
- Salmon, M., Earman, J., Glymour, C., Lennox, J., Machamer, P., McGuire, J., Norton, J., Salmon, W., and Schaffner, K. 1992. *Introduction to the Philosophy of Science*. Englewood Cliffs: Prentice Hall.
- Salmon, W. 1977. "The Philosophical Significance of the One-Way Speed of Light," *Noûs* **11**, 253-292.
- Sarkar, S. and Stachel, J. 1999. "Did Malament Prove the Non-Conventionality of Simultaneity in the Special Theory of Relativity?" *Philosophy of Science* **66**, 208-220.
- Scribner, C. 1963. "Mistranslation of a Passage in Einstein's Original Paper on Relativity," *American Journal of Physics* **31**, 398.
- Spirtes, P. 1981. *Conventionalism and the Philosophy of Henri Poincaré*. Ph. D. Dissertation, University of Pittsburgh.
- Stein, H. 1991. "On Relativity Theory and Openness of the Future," *Philosophy of Science* **58**, 147-167.
- Torretti, R. 1983. *Relativity and Geometry*. Oxford/New York: Pergamon.

- Winnie, J. 1970a. "Special Relativity Without One-Way Velocity Assumptions: Part I," *Philosophy of Science* **37**, 81-99.
- Winnie, J. 1970b. "Special Relativity Without One-Way Velocity Assumptions: Part II," *Philosophy of Science* **37**, 223-238.
- Zangari, M. 1994. "A New Twist in the Conventionality of Simultaneity Debate," *Philosophy of Science* **61**, 267-275.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

causation: causal processes | Einstein, Albert: philosophy of science | Reichenbach, Hans

Allen I. Janis
aij@pitt.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 31, 1998
Content last modified: July 23, 2002

Identity politics

The laden phrase "identity politics" has come to signify a wide range of political activity and theorizing founded in the shared experiences of injustice of members of certain social groups. Rather than organizing solely around ideology or party affiliation, identity politics typically concerns the liberation of a specific constituency marginalized within its larger context. Members of that constituency assert or reclaim ways of understanding their distinctiveness that challenge dominant oppressive characterizations, with the goal of greater self-determination.

1. History and Scope

The second half of the twentieth century saw the emergence of large-scale political movements -- second wave feminism, Black Civil Rights in the U.S., gay and lesbian liberation, and the American Indian movements, for example -- based in claims about the injustices done to particular social groups. These social movements are undergirded by and foster a philosophical body of literature that takes up questions about the nature, origin and futures of the identities being defended. Identity politics as a mode of organizing is intimately connected to the idea that some social groups are oppressed; that is, that one's identity as a woman or as a Native American, for example, makes one peculiarly vulnerable to cultural imperialism (including stereotyping, erasure, or appropriation of one's group identity), violence,

exploitation, marginalization, or powerlessness (Young 1990). Identity politics starts from analyses of oppression to recommend, variously, the reclaiming, redescription, or transformation of previously stigmatized accounts of group membership. Rather than accepting the negative scripts offered by a dominant culture about one's own inferiority, one transforms one's own sense of self and community, often through consciousness-raising. For example, in their germinal statement of Black feminist identity politics, the Combahee River Collective argued that "as children we realized that we were different from boys and that we were treated different -- for example, when we were told in the same breath to be quiet both for the sake of being 'ladylike' and to make us less objectionable in the eyes of white people. In the process of consciousness-raising, actually life-sharing, we began to recognize the commonality of our experiences and, from the sharing and growing consciousness, to build a politics that will change our lives and inevitably end our oppression" (Combahee River Collective 1982, 14-15).

The scope of political movements that may be described as identity politics is broad: the examples used in the philosophical literature are predominantly of struggles within Western capitalist democracies, but indigenous rights movements worldwide, nationalist projects, or demands for regional self-determination use similar arguments. Predictably, there is no straightforward criterion that makes a political struggle into an example of "identity politics;" rather, the term signifies a loose collection of political projects that each articulate a collective with a distinctively different social location that has hitherto been neglected, erased, or suppressed. It is beyond the scope of this essay to offer historical or sociological surveys of the many different social movements that might be described as identity politics, although some references to this literature are provided in the bibliography; instead the focus here is to provide an overview of the philosophical issues in the expansive literature in political theory.

The phrase "identity politics" is also something of a philosophical punching-bag for a variety of critics. Often challenges fail to make sufficiently clear their object of critique, using "identity politics" as a blanket description that invokes a range of tacit political failings (as discussed in Bickford 1997). From a contemporary perspective, some early identity claims by political activists certainly seem naive, totalizing, or unnuanced. However, the public rhetoric of identity politics both served useful and empowering purposes for some, and belied more subtle philosophical understandings of what political liberation requires. Since the twentieth century heyday of the well known political movements that made identity politics so visible, a vast academic literature has sprung up; although "identity politics" can draw on intellectual precursors from Mary Wollstonecraft to Frantz Fanon, writing that actually uses this specific phrase, with all its contemporary baggage, is limited almost exclusively to the last 15 years. Thus it was barely as intellectuals started to systematically outline and defend the philosophical underpinnings of identity politics that we simultaneously began to deconstruct them. At this historical juncture, then, asking whether one is for or against identity politics is to ask an impossible question. Wherever they line up in the debates, thinkers agree that the notion of *identity* has become indispensable to contemporary political discourse, at the same time as they concur that it has troubling implications for models of the self, political inclusiveness, and our possibilities for solidarity and resistance.

2. Philosophy and Identity

From this brief examination of how identity politics fits into the political landscape it is already clear that the use of the controversial term “identity” raises a host of philosophical questions. Logical uses aside, it is likely familiar to philosophers from the literature in metaphysics on personal identity -- one's sense of self and its persistence. Indeed, underlying many of the more overtly pragmatic debates about the merits of identity politics are philosophical questions about the nature of subjectivity and the self (Taylor 1989). Charles Taylor argues that the modern identity is characterized by an emphasis on its inner voice and capacity for *authenticity* -- that is, the ability to find a way of being that is somehow true to oneself (Taylor in Gutmann, ed. 1994). While doctrines of equality press the notion that each human being is capable of deploying his or her reason or moral sense to live an authentic live qua individual, the politics of difference has appropriated the language of authenticity to describe ways of living that are true to the identities of marginalized social groups. As Sonia Kruks puts it:

> What makes identity politics a significant departure from earlier, pre-identarian forms of the politics of recognition is its demand for recognition on the basis of the very grounds on which recognition has previously been denied: it is *qua* women, *qua* blacks, *qua* lesbians that groups demand recognition. The demand is not for inclusion within the fold of “universal humankind” on the basis of shared human attributes; nor is it for respect “in spite of” one's differences. Rather, what is demanded is respect for oneself *as* different (2001, 85).

For many proponents of identity politics this demand for authenticity includes appeals to a time before oppression, or a culture or way of life damaged by colonialism, imperialism, or even genocide. Thus for example Taiaiake Alfred, in his defense of a return to traditional indigenous values, argues that:

> Indigenous governance systems embody distinctive political values, radically different from those of the mainstream. Western notions of domination (human and natural) are noticeably absent; in their place we find harmony, autonomy, and respect. We have a responsibility to recover, understand, and preserve these values, not only because they represent a unique contribution to the history of ideas, but because renewal of respect for traditional values is the only lasting solution to the political, economic, and social problems that beset our people. (Alfred 1999, 5)

What is crucial about the “identity” of identity politics appears to be the experience of the subject, especially his or her experience of oppression and the possibility of a shared and more authentic alternative. Thus identity politics rests on unifying claims about the meaning of politically laden experiences to diverse individuals. Sometimes the meaning attributed to a particular experience will diverge from that of its subject: thus, for example, the woman who struggles desperately to be thin may think that she is simply trying to be a better person, rather than understanding her experience as part of the disciplining of female bodies in a patriarchal culture. Making sense of such disjunctions relies on notions such as false consciousness -- the systematic mystification of the experience of the oppressed by the perspective of the dominant. Thus despite its conflicts with Marxism and other radical political models, identity politics shares with them the anti-liberal view that individuals' perceptions of their own interests may be systematically distorted by ideology and must be somehow freed of their misperceptions by group-

based transformation.

Concern about this aspect of identity politics has crystallized around the transparency of experience to the oppressed, and the univocality of its interpretation. Experience is never, critics argue, epistemically available with a singular meaning (Scott 1992); rather it requires a theoretical framework -- implicit or explicit -- to give it sense. Moreover, if experience is the origin of politics, then some critics worry that what Kruks (2001) calls "an epistemology of provenance" will become the norm: on this view, political perspectives gain legitimacy by virtue of their articulation by subjects of particular experiences. This closes off the possibility of critique of these perspectives by those who don't share the experience, which in turn inhibits political dialogue and coalition-building.

From these understandings of subjectivity, it is easy to see how critics of identity politics, and even some cautious supporters, have feared that it is prone to *essentialism*. This term is another philosophical term of abuse, intended to capture a multitude of sins. In its original contexts in metaphysics, the term implies the belief that an object has a certain quality by virtue of which it is what it is; for Locke, famously, the essence of a triangle is that it is a three-sided shape. In the contemporary humanities the term is used more loosely to imply, most commonly, an illegitimate generalization about identity (Heyes 2000). In the case of identity politics, two claims stand out as plausibly "essentialist:" the first is the understanding of the subject that makes a single axis of identity stand in for the whole, as if being Asian-American, for example, were entirely separable from being a woman. To the extent that identity politics urges mobilization around a single axis, it will put pressure on participants to identify that axis as their defining feature, when in fact they may well understand themselves as heterogeneous selves with multiple identities and political goals (Spelman 1988). The second form of essentialism is closely related to the first: generalizations made about particular social groups in the context of identity politics may come to have a disciplinary function within the group, not just describing but also dictating the self-understanding that its members should have. Thus, the supposedly liberatory new identity may inhibit autonomy, as Anthony Appiah puts it, replacing "one kind of tyranny with another" (Appiah in Gutmann ed. 1994, 163). Just as dominant groups in the culture at large insisted that the marginalized integrate by assimilating to dominant norms, so within some practices of identity politics dominant sub-groups may, in theory and practice, impose their vision of the group's identity onto all its members. For example, in his films *Black Is, Black Ain't* and *Tongues Untied* Marlon Riggs eloquently portrays the exclusion of Black women and gay Black men from heterosexist and masculinist understandings of African-American identity politics.

Philosophical discussion around the identities identity politics defends has thus centered on a familiar metaphysical tension between identity and difference, and the possibilities for solidarity when these opposites are transposed to political contexts. Postmodern critics have suggested that alterity from dominant norms and within and between marginalized group members is a better descriptive and normative social ontology. How can a politics of difference mediate a conventional liberal individualism and more traditional identity politics? This question reflects the tremendous ambivalence with which all interlocutors approach identity politics. Many commentators describe and theorize the experience of hybridity for those whose identities are especially far from norms of univocality: Gloria Anzaldúa, for example, famously writes of her *mestiza* identity as a Chicana, American, raised poor, a lesbian and a

feminist, living in the metaphoric and literal Borderlands of the American Southwest (Anzaldúa 1999 [1987]). Some suggest the deployment of "strategic essentialism:" we should act *as if* an identity were uniform only to achieve interim political goals, without implying any deeper authenticity (Spivak 1990, 1-16). Others argue that a relational social ontology, which makes clear the fluidity and interdependence of social groups, should be developed as an alternative to the reification of other approaches to identity politics (Young 2000; Nelson 2001). These new accounts of subjectivity, new ontologies, and new ways of understanding solidarity and relationships are perhaps the most interesting and important face of contemporary scholarship in identity politics.

3. Liberalism and Identity Politics

A key condition of possibility for contemporary identity politics was institutionalized liberal democracy (Brown 1995). The perceived paucity of rewards offered by liberal capitalism after the extension of formal rights to most adult citizens spurred forms of radical critique that sought to explain the persistence of oppression. At the most basic philosophical level, critics of liberalism suggested that liberal social ontology -- the model of the nature of and relationship between subjects and collectives -- was misguided. The social ontology of most liberal political theories consists of citizens conceptualized as essentially similar individuals, as for example in John Rawls' famous thought experiment using the "original position," in which representatives of the citizenry are conceptually divested of all specific identities or affiliations in order to make rational decisions about social welfare (Rawls 1970). To the extent that group interests are represented in liberal polities, they tend to be understood as associational, forms of interest group pluralism whereby those sharing particular interests voluntarily join together to create a political lobby. Citizens are free to register their individual preferences (through voting, for example), or to aggregate themselves for the opportunity to lobby more systematically (e.g. by forming an association such as a neighborhood community league). These lobbies, however, are not defined by the identity of their members so much as by specific shared interests and goals, and their members are not taken to be peculiarly disadvantaged in pressing their case. Indeed, interest groups continue to include very powerful associations such as the National Rifle Association in the U.S., or tobacco company lobbies. Finally, political parties, the other primary organs of liberal democratic government, critics suggest, have few moments of inclusivity, being organized around party discipline, responsiveness to lobby groups, and broad-based electoral popularity. Ultimately conventional liberalism, diverse radical critics claim, cannot effectively address the ongoing structural marginalization that persists in late capitalist liberal states, and may even be complicit with it (Young 1990; P. Williams 1991; Brown 1995; M. Williams 1998).

On a philosophical level, these understandings of the political subject and its relationship to collectivity came to seem inadequate to ensuring representation for women, gays and lesbians, or racial-ethnic groups (M. Williams 1998). Critics charged that the neutral citizen of liberal theory was in fact the bearer of an identity coded white, male, bourgeois, able-bodied, and heterosexual (Young 1990). This implicit ontology in part explained the persistent historical failure of liberal democracies to achieve anything more than token inclusion in power structures for members of marginalized groups. A richer understanding of political subjects as deeply shaped by their social location was required. In particular, the history and experience of oppression brought with it certain perspectives and needs that could not be assimilated

through existing liberal structures. Individuals are oppressed by virtue of their membership in a particular *social group* -- that is, a collective whose members have relatively little mobility into or out of the collective, who usually experience their membership as involuntary, who are generally identified as members by others, and whose opportunities are deeply shaped by the relation of their group to corollary groups through privilege and oppression. Oppression, then, is the systematic limiting of opportunity or constraints on self-determination because of such membership: for example, Frantz Fanon eloquently describes the experience of being always constrained by the white gaze as a Black man: “I already knew that there were legends, stories, history, and above all *historicity*… I was responsible at the same time for my body, my race, for my ancestors” (Fanon 1968, 112). Conversely, members of dominant groups are *privileged* -- systematically advantaged by the deprivations imposed on the oppressed. For example, in a widely cited article Peggy McIntosh identifies whiteness as a dominant identity, and lists 47 ways in which she is advantaged by being white compared with her colleagues of color. These range from being able to buy “flesh-colored” Band-Aids that will match her skin tone, to knowing that she can be rude without provoking negative judgments of her racial group, to being able to buy a house in a middle-class community without risking neighbors' disapproval (1993).

Critics have also charged that *assimilation* (or, less provocatively, integration) is a guiding principle of liberalism. If the liberal subject is coded in the way Young (1990) suggests, then attempts to apply liberal norms of equality will risk demanding that the marginalized conform to the identities of their oppressors. For example, many gays and lesbians have objected to campaigns to institute “gay marriage” on the grounds that these legal developments assimilate same-sex relationships to a heterosexual model, rather than challenging its terms. If this is *equality*, they claim, then it looks suspiciously like the erasure of socially subordinate identities rather than their genuine incorporation into the polity. This suspicion helps to explain the affiliation of identity politics with *separatism*. This latter is a set of positions that share the view that attempts at integration of dominant and marginalized groups so consistently compromise the identity or potential of the less powerful that a distinct social and political space is the only structure that will adequately protect them. In Canada, for example, Québec separatists claim that the French language and francophone culture are persistently erased within an overwhelmingly dominant Anglo-American continent, despite the efforts of the Canadian state to maintain its official bilingualism and to integrate Québec into the nation. Given their long history of conflict and marginalization, a separate and sovereign Québec, they argue, is the only plausible solution (e.g. Laforest in Beiner and Norman 2001). Analogous arguments have been made on behalf of Native American and other indigenous peoples and African Americans (e.g. Alfred 1999, Asante 2000). Lesbian feminist separatists have claimed that the central mechanism for the oppression of women under patriarchy is heterosexuality. Understanding heterosexuality as a forced contract or compulsory institution, they argue that women's relationships with men are persistently characterized by domination and subordination. Only divorce (literal and figurative) and the creation of new geographic and political communities of woman-identified women will end patriarchal exploitation, and forge a liberatory female identity (Rich 1980; Frye 1983; Radicalesbians 1988; Wittig 1992).

One of the central charges against identity politics by liberals, among others, has been its alleged reliance on notions of sameness to justify political mobilization. Looking for people who are *like* you rather than who share your political values as allies runs the risk of sidelining critical political analysis of complex

social locations and ghettoizing members of social groups as the only persons capable of making or understanding claims to justice. After an initial wave of relatively uncompromising identity politics, proponents have taken these criticisms to heart and moved to more philosophically nuanced accounts that appeal to *coalitions* as better organizing structures. On this view, separatism around a single identity formation must be muted by recognition of the internally heterogeneous and overlapping nature of social group memberships. The idea of a dominant identity from which the oppressed may need to disassociate themselves remains, but the alternative becomes a more fluid and diverse grouping, less intent on guarantees of internal homogeneity and more concerned with identifying "family resemblances" than literal identity (Heyes 2000).

This trajectory -- from formal inclusion in liberal polities, to assertions of difference and new demands under the rubric of identity politics, to internal and external critique of identity political movements -- has taken different forms in relation to different identities. Increasingly it is difficult to see what divides contemporary positions, and some commentators have suggested possible *rapprochements* between liberalism and identity politics (e.g. Laden 2001). A problem in sorting through such claims is the vagueness of philosophical discussions of identity politics, which are often content to list their rubric under the mantra of "gender, race, class, etc." although these three are not obviously analogous, nor is it clear which identities are gestured toward by the predictable "etc." (or why they do not merit naming). Class in particular has a distinctively different political history, and contemporary critics of identity politics, as I'll discuss below, often take themselves to be defending class analysis against identity politics' depoliticizing effects. Of those many forms of identity politics to which large academic literatures attach, however, I'll briefly highlight key issues concerning gender, sexuality, disability, and a complex cluster of race, ethnicity and multiculturalism.

4. Gender and Feminism

Twentieth century feminism has consistently opposed biological determinism: the view that shared biological features among a certain group lead inevitably to certain social roles or functions. For example, one early opponent of women's suffrage suggested that women and men had different metabolic systems -- katabolic (or "energy-expending") in men, and anabolic (or "energy-conserving") in women -- that precluded women's effective or informed participation in politics (see Moi 2000, 3-21 for discussion). Feminist identity politics, then, takes up the task of articulating women's understandings of themselves (and of men) without reducing femininity (or masculine dominance) to biology, and instead situating women as oppressed under patriarchy. Two philosophical questions dog this endeavor: first, what to make of biological difference? In their eagerness to present gender as socially constructed and entirely separable from sex, feminists sidelined women's experiences of childbirth, menopause, or embodiment generally as "essentialist" and irrelevant to feminist politics. This gap is now being filled by so-called "sexual difference" feminism, but the legacy of social constructionism that reads gender identity only as a set of ideas rather than also embodied experiences has proved hard to shake. Second, the very idea of reclaiming women's identities from patriarchy has been criticized as merely an affirmation of a slave morality -- a Nietzschean term describing the *ressentiment* of the oppressed as they rationalize and prescribe their condition. Attempts from various quarters to capture and revalue the distinctively feminine

(by theorizing, for example, “maternal thinking,” [Ruddick 1989], or *écriture féminine* [Irigaray 1985]) risk endorsing existing power relations. Thus the heated debates surrounding the “ethic of care” in moral psychology, for example, line up around two constellations of positions: on the one hand, advocates of the ethic of care as a distinctively feminine contribution to moral reasoning point to its benefits for negotiating a social world characterized by webs of relationship, and to the pathologies of masculine disassociation. Carol Gilligan is the best known proponent of this position (although the details of her complex paradigm are often glossed or misrepresented) (Gilligan 1993 [1982]). Her critics charge that she reifies femininity -- were women not oppressed, they would not speak in the voice of care, thus casting doubt on its usefulness as a liberatory strategy. The current construction of femininity is so deeply imbricated with the oppression of women that such attempts will always end up reinforcing the very discourse they seek to undermine (Butler 1999 [1990]); this critique has strong affiliations with poststructuralism (which are discussed below).

The narrative of feminist interpretation of gender relations most commonly offered points to universalizing claims made on behalf of women during the so-called “second wave” of the feminist movement in the late 1960's and 1970's in Western countries. The most often discussed (and criticized) second wave feminist icons -- women such as Betty Friedan or Gloria Steinem -- are white, middle-class, and heterosexual, although this historical picture too often neglects the contributions of lesbian feminists, feminists of color, and working-class feminists, which were less visible in popular culture, perhaps, but arguably equally influential in the lives of women. For some early radical feminists, women's oppression *as women* was the core of identity politics, and should not be diluted with other identity issues. For example, Shulamith Firestone, in her classic book *The Dialectic of Sex*, argued that “*racism is sexism extended*,” and that the Black Power movement represented only sexist cooptation of Black women into a new kind of subservience to Black men. Thus for Black women to fight racism (especially among white women) was to divide the feminist movement, which properly focused on challenging patriarchy, understood as struggle between men and women, the foundational dynamic of all oppressions (Firestone 1970, esp. 103-120).

Claims about the universality of gender made during the second wave have been extensively criticized in feminist theory for failing to recognize the specificity of their own constituencies. For example, Friedan's famous proposition that women needed to get out of the household and into the professional workplace was, bell hooks pointed out, predicated on the experience of a post-war generation of white, middle-class married women confined to housekeeping and child-rearing by their professional husbands (Friedan 1963; hooks 1981). Many women of color and working-class women had worked outside their homes (sometimes in *other* women's homes) for decades; some lesbians had a history of working in traditionally male occupations or living alternative domestic lives without a man's “family wage.” Similarly, some women from the Southern hemisphere have been critical of Northern feminist theory for globalizing its claims. Such moves construct Southern women, they argue, as less developed or enlightened versions of their Northern counterparts, rather than understanding their distinctively different situation (Mohanty 1988); or, they characterize liberation for Northern women in ways that exacerbate the exploitation of the South: by supporting economic conditions in which increasing numbers of western women can abuse immigrant domestic workers, for example (Anderson 2000).

Thus feminist claims made about the oppression of women founded in a notion of shared experience and identity are now invariably greeted with philosophical suspicion. Some critics have charged that this suspicion itself has become excessive, undercutting the very possibility of generalizations about women that gives feminist theory its force (Martin 1994), or that it marks the distancing of feminist philosophy from its roots in political organizing. Others suggest alternative methods for feminist theory that will minimize the emphasis on shared criteria of membership in a social group and stress instead the possibilities for alliances founded on non-identical connections (Young 1997; Heyes 2000; Cornell 2000). It is commonplace to hear that "identity" is a term in serious crisis in feminist thought, and that feminist praxis must move beyond identity politics (Dean 1996). Nonetheless, sex-gender as a set of analytical categories continues to guide feminist thought, albeit in troubled and troubling ways.

5. From Gay and Lesbian to Queer

Nowhere have conceptual struggles over identity been more pronounced than in the lesbian and gay liberation movement. The notion that sexuality provides a stable and authentic core identity has itself been profoundly challenged by the advent of queer politics. Most early lesbian and gay activists emphasized the authenticity of their identities; they were a distinctively different natural kind of person, with the same rights as heterosexuals (another natural kind) to find fulfillment in marriage, child-rearing, property ownership, and so on. This conformist strand of gay organizing (perhaps associated more closely with white, middle-class gay men, at least until the radicalizing effects of the AIDS pandemic) has a genealogy going back to pre-Stonewall homophilic activism. While early lesbian feminists had a very different politics, oriented around liberation from patriarchy and the creation of separate spaces for woman-identified women, many still appealed to a more authentic, distinctively feminist self. Heterosexual feminine identities were products of oppression, yet the literature imagines a utopian alternative where woman-identification will liberate the lesbian within every woman.

The paradigm shift that the term "queer" signals, then, is a shift to a model in which identities are more self-consciously historicized, seen as contingent products of particular genealogies rather than enduring or essential natural kinds (Phelan 1989 and 1994; Blasius 2001). Michel Foucault's work, especially his *History of Sexuality*, is the most widely cited progenitor of this view: Foucault famously argues that "homosexuality appeared as one of the forms of sexuality when it was transposed from the practice of sodomy onto a kind of interior androgyny, a hermaphrodism of the soul. The sodomite had been a temporary aberration; the homosexual was now a species" (Foucault 1980, 43). Although Foucault is the most often cited as the originator of social constructionist arguments about sexuality, other often neglected writers contributed to the emergence of this new paradigm (e.g. M. McIntosh 1968). In western popular culture such theories co-exist uneasily with biologically essentialist accounts of sexual identity, which look for a particular gene, brain structure, or other biological feature that will explain same-sex sexual desire. At stake are not only epistemic questions about the correct explanation for certain human behaviors, but also a host of moral and political questions. If sexual identity is biological, then no individual is morally responsible for it, any more than it makes sense to say that an individual is responsible for his or her race. Some gay activists thus see biological explanations of sexuality as offering a defense against homophobic commentators who believe that gays can voluntarily change their

“immoral” behaviors. Indeed, much of the intuitive hostility to social constructionist accounts of sexuality within gay and lesbian communities seems to come from the dual sense of many individuals that they could not have been other than gay, and that anything less than a radically essentialist view of sexuality will open the door to further attempts to “cure” them of their homosexuality (through “ex-gay ministries,” for example).

Whatever the truth of these fears, Eve Sedgwick is right to say that no specific form of explanation for the origins of sexual preference will be proof against the infinitely varied strategies of homophobia (Sedgwick 1990, esp. 22-63). Queer politics, then, both stresses the usefulness of social constructionism while eschewing a genetic quest for the origins of homosexuality as always a presentist history. In addition to historicizing and contextualizing sexuality, including the very idea of sexual identity, the shift to queer is also characterized by deconstructive methods. Rather than understanding sexual identities as a set of discrete and independent social types, queer theorists emphasize their mutual implication: for example, the word “homosexuality” first appears in English in 1897, but the term “heterosexuality” is back-formed, first used some years later (Garber 1995, 39-42). Heterosexuality comes into existence as a way of understanding the nature of individuals *after* the homosexual has been diagnosed; homosexuality *requires* heterosexuality as its opposite, despite its self-professed essentialism. Queer theorists point out that the homo/hetero dichotomy, like many others in western intellectual history that it arguably draws on and reinforces, is not only mutually implicated, but also hierarchical (heterosexuality is superior, normal, and inevitable) and masquerades as natural or descriptive. The task of a more radical “identity politics,” on this vision, is to constantly denaturalize and deconstruct the identities in question, with a political goal of their subversion rather than their accommodation.

An exemplary conflict within the identity politics of sexuality focuses on the expansion of gay and lesbian organizing to those with other queer affiliations, especially bisexual and transgendered activists. Skepticism about inclusion of these groups in organizational mandates, community centers, parades, and festivals has origins in more traditional understandings of identity politics that see reclaiming lesbian and/or gay identity from its corruption in a homophobic society as a task compromised by those whose identities are read as diluted, treacherous, ambiguous, or peripheral. Lesbian feminist critiques of transgender, for example, see male-to-female transsexuals in particular as male infiltrators of women's space, individuals so intent on denying their male privilege that they will modify their bodies and attempt to pass as women to do it; bisexual women dabble in lesbian life, but flee to straight privilege when occasion demands (see Heyes 2003 for references and discussion). These arguments have been challenged in turn by writers who see them as attempts to justify purity of identity that merely replace the old exclusions with new dictatorships (Stone 1991, Lugones 1994) and inhibit coalitional organizing against conservative foes.

6. Disability

The trope of social constructionism reappears in disability studies as the argument that disability is not a natural or objective flaw of certain individuals, but rather a set of challenges faced by those whose needs the dominant culture fails to accommodate (Wendell 1996; Davis 1997). In a society in which buildings

or communications systems, for example, were differently designed, many of the struggles of people currently labeled "disabled" would no longer be disadvantaged. This has been the basis for legislation (such as the 1990 Americans with Disabilities Act), which requires of employers that they make reasonable accommodation for disabled employees, to minimize the disadvantages they face. Disability rights language here draws on feminist discourse: if employers are obliged to accommodate childbirth and parental leave (when once they would have fired or not hired women who had or were thought likely to have children), then other embodied differences merit the same treatment in order that employees can perform to their full potential. Thus disability rights advocates use the familiar strategy of shifting moral responsibility for an identity away from those who involuntarily share it, towards those whose actions have made it oppressive.

That disability should be involuntary rather than chosen has become an important issue here (in some ways paralleling debates in sexuality studies). It is a trope of liberal discourse that one should not be held morally responsible for traits (or their consequences) that one cannot control; thus arguments from accommodation have come to depend on disability being understood as analogous to sex or race in its immutability. Yet life-long smokers with chronic lung disease, or the very obese (to take a particularly tricky example), also require changes to work schedules or physical environments to participate in public life. If choosing to act differently *could* ameliorate or dispel the disability in question, why should others accommodate it?

Many philosophers of disability suggest that these questions conceal a deeper issue: like many objectors to identity politics, those skeptical of accommodating the disabled are liable to assume a charity model. The disabled, they suggest, are simply a burden on the larger society, and their demands constitute special pleading. Yet like many marginalized groups, disabled rights advocates point out that experiences of disability are far from entirely negative, and in fact have value as source of knowledge or a standpoint unavailable to others (Wendell 1996). This argument has taken a particularly forceful turn within Deaf studies, where scholars have argued that the Deaf constitute a separate culture and linguistic minority (as users of sign languages) that should be preserved and fostered. New technologies (such as cochlear implants) promise to reduce the numbers of deaf people, and increase the proportion of the deaf and hard-of-hearing who can use spoken language to communicate with hearing culture. For some Deaf activists, this move is less the lifting of an unfortunate affliction than the genocide of an under-valued and stigmatized culture. Again, the identity of being Deaf is affirmed contra attempts to assimilate it to the terms of the hearing.

7. Race, Ethnicity, and Multiculturalism

Similar debates in philosophy of race highlight the socially constructed and historical nature of "race" as a category of identity. Despite a complex history of biological essentialism in the presentation of racial typologies, the notion of a genetic basis to racial difference has been discredited; the criteria different societies (at different times) use to organize and hierarchize "racial formations" are political and contingent (Omi and Winant 1986). While skin color, appearance of facial features, or hair type are in some trivial sense genetically determined, the grouping of different persons into *races* does not pick out

any patterned biological difference. What it does pick out is a set of social meanings (Alcoff 1997). The most notorious example of an attempt to rationalize racial difference as biological is the U.S. "one-drop rule," under which an individual was characterized as Black if they had "one drop" or more of "Black blood." Adrian Piper points out that not only does this belief persist into contemporary readings of racial identity, it also implies that given the prolonged history of racial mixing in the US -- both coerced and voluntary -- very significant numbers of nominally "white" people in the U.S. today should be re-classified as "Black" (Piper 1996). In those countries that have had official racial classifications, individuals' struggles to be re-classified almost always as a member of a more privileged racial group are often invoked to highlight the contingency of race, especially at the borders of its categories. And a number of histories of racial groups that have apparently changed their racial identification -- Jews, Italians, or the Irish, for example -- also illustrate social constructionist theses (Ignatiev 1995).

The claim that race is socially constructed does not in itself mark out a specific identity politics. Indeed, the very contingency of race and its lack of correlation with ethnicity or culture circumscribes its political usefulness: just as feminists have found the limits of appeals to "women's identity," so Asian-Americans may find with ethnicities and cultures as diverse as Chinese, Indian, or Vietnamese that their racial designation itself provides little common ground. That a US citizen of both Norwegian and Ashkenazi Jewish heritage will check that they are "white" on a census form says relatively little (although nonetheless something) about their experience of their identity, or indeed of their very different relationship to anti-Semitism. Tropes of separatism and the search for forms of authentic self-expression are related to race via ethno-cultural understandings of identity: for example, the U.S. Afro-centric movement appeals to the cultural significance of African heritage for Black Americans (Asante 2000).

Where perhaps racial categories are most politically significant is in their contested relation to racism. Racism attempts to reduce members of social groups to their racial features, drawing on a complex history of racial stereotypes to do so. Racism is arguably analogous to other forms of oppression in being both overt and institutionalized, manifested both as deliberate acts by individuals and as unplanned systemic outcomes. The specific direction of US discussion of the categories of race has been around color-blind versus color-conscious public policy (Appiah and Gutmann 1996). Color-blindness -- that is, the view that race *should* be ignored in public policy and everyday exchange -- has hegemony in popular discourse. Drawing attention to race -- whether in a personal description or in university admissions procedures -- is unfair and racist. Advocates of color-consciousness argue that racism will not disappear without proactive efforts, which require the invocation of race. Thus affirmative action, for example, requires statistics about the numbers of members of oppressed racial groups employed in certain contexts, which in turn requires racial identification and categorization. Thus those working against racism face a paradox familiar in identity politics: the very identity they aim to dispel must be invoked to make their case.

The literature on multiculturalism takes up questions of race, ethnicity, and cultural diversity in relation to the liberal state. Some multicultural states -- notably Canada -- allegedly aim to permit the various cultural identities of their residents to be preserved rather than assimilated, despite the concern that the over-arching liberal aims of such states may be at odds with the values of those they claim to protect. For example, Susan Moller Okin argues that multiculturalism is sometimes bad for women, especially when it

works to preserve patriarchal values in minority cultures. If multiculturalism implies a form of cultural relativism that prevents judgment of or interference with the “private” practices of minorities, female genital mutilation, forced marriage, compulsory veiling, or being deprived of education may be the consequence. Okin's critics counter that she falsely portrays culture as static, internally homogeneous, and defined by men's values, allowing liberalism to represent a culturally unmarked medium for the defense of individual rights (Okin et al 1999). For many commentators on multiculturalism this is the nub of the issue: is there an inconsistency between defending the rights of minority cultures, while prohibiting those (allegedly) cultural practices that the state judges illiberal? Can liberalism sustain the cultural and value-neutrality that some commentators still ascribe to it, or to what extent should it embrace its own cultural specificity (Taylor, Habermas in Gutmann, ed. 1994; Lawrence and Herzog, eds. 1994; Kymlicka, ed. 1995)? Defenders of the right to cultural expression of minorities in multicultural states thus practice forms of identity politics that are both made possible by liberalism and sometimes in tension with it.

8. Other Challenges to Identity Politics

Since its 1970s vogue, identity politics as a mode of organizing and set of political philosophical positions has undergone numerous attacks by those motivated to point to its flaws, whether by its pragmatic exclusions or more programmatically as liberals, Marxists, or poststructuralists. For many leftist commentators, identity politics is something of a *bête noire*, representing the capitulation to cultural criticism in place of analysis of the material roots of oppression. Marxists, both orthodox and revisionist, and socialists -- especially those who came of age during the rise of the New Left in western countries -- have often interpreted the perceived ascendancy of identity politics as representing the end of radical materialist critique (see discussion in Farred 2000). Identity politics, for these critics, is both factionalizing and depoliticizing, drawing attention away from the ravages of late capitalism toward superstructural cultural accommodations that leave economic structures unchanged. For example, while allowing that both recognition and redistribution have a place in contemporary politics, Nancy Fraser laments the supremacy of perspectives that take injustice to inhere in “cultural” constructions of identity that the people to whom they are attributed want to reject. Such recognition models, she argues, require remedies that “valorize the group's ‘groupness’ by recognizing its specificity,” thus reifying identities that themselves are products of oppressive structures. By contrast, injustices of distribution require redistributive remedies that aim “to put the group out of business as a group” (Fraser 1997, 19).

The reasons given for this alleged turn away from economic oppression to themes of culture, language, and identity in contemporary politics differ. First, the institutionalization of North American radicalism in the middle-class bastion of academia creates incentives for intellectuals to minimize the political importance of their own class privilege, and focus instead on other identities (in turn divorced from their economic inflections). Second, Wendy Brown suggests that capitalist suffering has been displaced onto other identities, interpreted through the lens of class aspiration (Brown 1995, 59-60). Third, the turn away from economic analysis may be less dramatic than some critics believe: recent activism against global capitalism indicates a resurgence in economic critique that is now arguably more fully imbricated with identity politics (Lott 2000). Finally, the rise of diverse “postmodern” paradigms offers sophisticated theoretical alternatives to Marxism for those on the left.

Poststructuralist challenges to identity politics are perhaps the most philosophically developed and profound. Poststructuralists charge that identity politics rests on a mistaken view of the subject that assumes a *metaphysics of substance* -- that is, that a cohesive, self-identical subject can be identified and reclaimed from oppression (Butler 1999). This subject has certain core essential attributes that define her or his identity, over which are imposed forms of socialization that cause her or him to internalize other nonessential attributes. This position, they suggest, misrepresents both the psychology of identity and its political significance. The alternative view offered by poststructuralists is that the subject is itself always already a product of discourse, that possibilities for subjecthood are set out in advance of any possible expression by an individual. Some critics are uncomfortable with the limitations on agency this seems to assume, but advocates argue that changing discourse itself is a better possible goal than reclaiming authentic identities; individual choices have a role to play in this project.

Also key to poststructuralist positions is the mutually sustaining opposition of identity and difference:

> An identity is established in relation to a series of differences that have become socially recognized. These differences are essential to its being. If they did not coexist as differences, it would not exist in its distinctness and solidity. Entrenched in this indispensable relation is a second set of tendencies, themselves in need of exploration, to conceal established identities into fixed forms, thought and lived as if their structure expressed the true order of things. When these pressures prevail, the maintenance of one identity (or field of identities) involves the conversion of some differences into otherness, into evil, or one of its numerous surrogates. Identity requires differences in order to be, and it converts difference into otherness in order to secure its own self-certainty. (Connolly 2002, 64)

The dangers of identity politics, then, are that it casts as authentic to the self or group an identity that in fact is defined by its opposition to an Other. Reclaiming such an identity as one's own merely reinforces its dependence on this dominant Other, and further internalizes and reinforces an oppressive discourse. These moves cultivate *ressentiment* (the moralizing revenge of the powerless): while the charge that identity politics promotes a victim mentality is often facile, Brown makes a sophisticated version of the critique. She argues that identity politics has its own genealogy in liberal capitalism that relentlessly reinforces the “wounded attachments” it claims to sever: “Politicized identity thus enunciates itself, makes claims for itself, only by entrenching, restating, dramatizing, and inscribing its pain in politics; it can hold out no future -- for itself or others -- that triumphs over this pain” (Brown 1995, 74).

What political alternatives does this model imply? Proponents of identity politics have suggested that poststructuralism is politically impotent, capable only of deconstruction and never of action (Hartsock 1998, 205-226). Yet there are political projects motivated by poststructuralist theses. For example, Judith Butler's famous articulation of *performativity* as a way of understanding subject-development suggests to her and others the possibility of disarticulating seamless performances to subvert the meanings with which they are invested (Butler 1999). Drag can constitute such a disarticulation, although other critics have suggested other examples; Adrian Piper's conceptual art seeks to disrupt the presumed self-identity

of race by showing how it is actively interpreted and reconstituted, never determinate and self-evident.

9. Identity Politics in the 21st Century

The continuing intellectual crisis surrounding identity politics paradoxically marks its importance to contemporary political philosophy and practice. Both flexible and extensible, identity political tropes continue to influence new political claims: can identity politics be extended to children, for example, as the emergent children's rights movement implies? Identity politics has limits, too: is it too person-centered? How can identity politics also be an environmental politics (Sandilands 2000)? Perhaps most important for philosophers, the idea of identity itself appears to be in a period of rapid evolution. Changing technologies are having a profound impact on our philosophical understandings of who we are. Attempts to decode human genetics and possibly shape the genetic make-up of future persons (Wald 2000), to clone human beings, or to xeno-transplant animal organs, and so on, all raise deep philosophical questions about the kind of thing a person is. We are capable of changing our bodies in ways that dramatically change our identities, including through sex change or cosmetic surgeries, with immediate consequences for the kinds of identities I have been discussing in this essay. As more and more people form political alliances using disembodied communications technologies, the kinds of identities that matter seem also to shift (Turkle 1995). Our identities are increasingly pathologized as syndromes and disorders and treated by psychiatrists, and political thinkers should continue to criticize and resist the tendency to dehistoricize and naturalize them (Elliott 2003). At the same time, familiar mechanisms of oppression are further entrenching the very identities that in some western, wealthy contexts look set to fragment. Global capitalism appears to be widening the gap between the North and South, and working to further marginalize women, ethnic or indigenous minorities, and the disabled in the so-called Third and Fourth Worlds.[1] This mass of shifts and contradictions helps explain one move that almost all intellectuals agree on: identity politics must adopt a local focus. Structures of oppression may operate at macro-levels, but their consequences for the lived experience of those whose self-determination they undermine are myriad.

Bibliography

References cited

- Alcoff, Linda Martin. 1997. “Philosophy and Racial Identity.” *Philosophy Today* 41:1-4:67-76.
- Alfred, Taiaiake. 1999. *Peace, Power, and Righteousness: An Indigenous Manifesto*. Oxford: Oxford University Press.
- Anderson, Bridget. 2000. *Doing the Dirty Work? The Global Politics of Domestic Labour*. London: Zed Books.
- Anzaldúa, Gloria. 1999 (1987). *Borderlands/La Frontera: The New Mestiza*. San Francisco: Aunt Lute.
- Appiah, Anthony and Amy Gutmann. 1996. *Color Conscious: The Political Morality of Race*. Princeton: Princeton University Press.

- Asante, Molefi K. 1998. *The Afrocentric Idea*. Philadelphia: Temple University Press.
- -----. 2000. *The Painful Demise of Eurocentrism: An Afrocentric Response to Critics*. Africa World Press.
- Beiner, Ronald and Wayne Norman (eds). 2001. *Canadian Political Philosophy: Contemporary Reflections*. Don Mills, ON: Oxford University Press.
- Bickford, Susan. 1997. “Anti-Anti-Identity Politics: Feminism, Democracy, and the Complexities of Citizenship.” *Hypatia* 12:4: 111-31.
- Blasius, Mark (ed). 2001. *Sexual Identities, Queer Politics*. Princeton: Princeton University Press.
- Brown, Wendy. 1995. *States of Injury: Power and Freedom in Late Modernity*. Princeton: Princeton University Press.
- Butler, Judith. 1999 [1990]. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.
- The Combahee River Collective. 1982. “A Black Feminist Statement,” in *All the Women are White, All the Blacks are Men, But Some of Us Are Brave: Black Women's Studies*, Gloria T. Hull, Patricia Bell Scott, and Barbara Smith (eds). New York: Feminist Press.
- Connolly, William. 2002. *Identity\Difference: Democratic Negotiations of Political Paradox*. Minneapolis: University of Minnesota Press.
- Cornell, Drucilla. 2000. *Just Cause: Freedom, Identity, and Rights*. Lanham, MD: Rowman and Littlefield.
- Davis, Lennard J. (ed.). 1997. *The Disability Studies Reader*. New York: Routledge.
- Dean, Jodi. 1996. *Solidarity of Strangers: Feminism after Identity Politics*. Berkeley: University of California Press.
- Elliott, Carl. 2003. “Does Your Patient Have A Beetle in His Box? Language Games and Psychopathology,” in *The Grammar of Politics: Wittgenstein and Political Philosophy*, Cressida J. Heyes (ed.). Ithaca, NY: Cornell University Press.
- Fanon, Frantz. trans. Charles Markmann. 1968. *Black Skin, White Masks*. New York: Grove Press.
- Farred, Grant. 2000. “Endgame Identity? Mapping the New Left Roots of Identity Politics.” *New Literary History* 31:4: 627-48.
- Firestone, Shulamith. 1970. *The Dialectic of Sex: The Case for Feminist Revolution*. New York: Morrow.
- Foster, Lawrence and Patricia Herzog (eds). 1994. *Defending Diversity: Contemporary Philosophical Perspectives on Pluralism and Multiculturalism*. Amherst, MA: University of Massachusetts Press.
- Foucault, Michel. 1980. *The History of Sexuality*. New York: Vintage.
- Fraser, Nancy. 1997. *Justice Interruptus: Critical Reflections on the “Post-Socialist” Condition*. New York. Routledge.
- Friedan, Betty. 1963. *The Feminine Mystique*. New York: Norton.
- Frye, Marilyn. 1983. *The Politics of Reality: Essays in Feminist Theory*. Trumansberg, NY: The Crossing Press.
- Garber, Marjorie. 1995. *Vice Versa: Bisexuality and the Eroticism of Everyday Life*. New York: Routledge.
- Gilligan, Carol. 1993 (1982). *In a Different Voice: Psychological Theory and Women's Development*. Cambridge: Harvard University Press.

- Gutmann, Amy (ed.). 1994. *Multiculturalism: Examining the Politics of Recognition*. Princeton: Princeton University Press.
- Hartsock, Nancy C. M. 1998. *The Feminist Standpoint Revisited and Other Essays*. Boulder: Westview.
- Heyes, Cressida J. 2000. *Line Drawings: Defining Women through Feminist Practice*. Ithaca, NY: Cornell University Press.
- -----. Forthcoming. “Feminist Solidarity after Queer Theory: The Case of Transgender.” *Signs* 28:2.
- Hooks, Bell. 1981. *Ain't I a Woman? Black Women and Feminism*. Boston: South End Press.
- Ignatiev, Noel. 1995. *How the Irish Became White*. New York: Routledge.
- Irigaray, Luce, trans. Catherine Porter. 1985. *This Sex Which is Not One*. Ithaca, NY: Cornell University Press.
- Kruks, Sonia. 2000. *Retrieving Experience: Subjectivity and Recognition in Feminist Politics*. Ithaca, NY: Cornell University Press.
- Kymlicka, Will (ed.). 1995. *The Rights of Minority Cultures*. Oxford: Oxford University Press.
- Laden, Anthony. 2001. *Reasonably Radical: Deliberative Liberalism and the Politics of Identity*. Ithaca: Cornell University Press.
- Lott, Eric. 2000. “After Identity, Politics: The Return of Universalism.” *New Literary History* 31:4: 665-80.
- Lugones, María. 1994. “Purity, Impurity, and Separation.” *Signs* 19: 458-79.
- McIntosh, Mary. 1968. “The Homosexual Role.” *Social Problems* 16:2.
- McIntosh, Peggy. 1993. “White Privilege and Male Privilege: A Personal Account of Coming to See Correspondences Through Work in Women's Studies,” in *Gender Basics*: *Feminist Perspectives on Women and Men,* Anne Minas (ed.). Belmont, CA: Wadsworth.
- Martin, Jane Roland. 1994. “Methodological Essentialism, False Difference, and Other Dangerous Traps.” *Signs* 19: 630-57.
- Mohanty, Chandra Talpade. 1988. “Under Western Eyes: Feminist Scholarship and Colonial Discourse,” in *Third World Women and the Politics of Feminism*, Chandra Talpade Mohanty, Ann Russo, and Lourdes Torres (eds.). Bloomington: Indiana University Press, 1991.
- Moi, Toril. 2000. *What is a Woman? And Other Essays*. Oxford: Oxford University Press.
- Nelson, Hilde Lindemann. 2001. *Damaged Identities, Narrative Repair*. Ithaca, NY: Cornell University Press.
- Okin, Susan Moller, et al. 1999. *Is Multiculturalism Bad for Women?* Princeton: Princeton University Press.
- Omi, Michael and Howard Winant. 1994 (1986). *Racial Formation in the United States: From the 1960s to the 1990s*. New York: Routledge.
- Phelan, Shane. 1994. *Getting Specific: Postmodern Lesbian Politics*. Minneapolis: University of Minnesota Press.
- -----. 1989. *Identity Politics: Lesbian Feminism and the Limits of Community*. Philadelphia: Temple University Press.
- Piper, Adrian. 1996. “Passing for Black, Passing for White,” in *Passing and the Fictions of Identity*, Elaine K. Ginsberg (ed.). Durham, NC: Duke University Press.
- Radicalesbians. 1988 (first published 1975). “The Woman Identified Woman,” in *For Lesbians*

Only: A Separatist Anthology, Sarah Hoagland and Julia Penelope (eds.). London: Onlywomen Press.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Rich, Adrienne. 1980. “Compulsory Heterosexuality and Lesbian Existence.” *Signs* 5:4.
- Ruddick, Sara. 1989. *Maternal Thinking: Toward a Politics of Peace*. New York: Ballantine.
- Sandilands, Catriona. 1995. “From Natural Identity to Radical Democracy.” *Environmental Ethics* 17: 75-91.
- Scott, Joan. 1992. “Experience,” in *Feminists Theorize the Political*, Judith Butler and Joan W. Scott (eds.). New York: Routledge.
- Sedgwick, Eve Kosofsky. 1990. *Epistemology of the Closet*. Berkeley: University of California Press.
- Spelman, Elizabeth V. 1988. *Inessential Woman: Problems of Exclusion in Feminist Thought*. Boston: Beacon Press.
- Spivak, Gayatri, ed. Sara Harasym. 1990. *The Post-Colonial Critic: Interviews, Strategies, Dialogues*. New York: Routledge.
- Taylor, Charles. 1989. *Sources of the Self: The Making of the Modern Identity*. Cambridge, MA: Harvard University Press.
- Turkle, Sherry. 1995. *Life on the Screen: Identity in the Age of the Internet*. New York: Simon and Schuster.
- Wald, Priscilla. 2000. “Future Perfect: Grammar, Genes, and Geography.” *New Literary History* 31:4: 681-708.
- Wendell, Susan. 1996. *The Rejected Body: Feminist Philosophical Reflections on Disability*. New York: Routledge.
- Williams, Melissa. 1998. *Voice, Trust, and Memory: Marginalized Groups and the Failings of Liberal Representation*. Princeton: Princeton University Press.
- Williams, Patricia. 1991. *The Alchemy of Race and Rights*. Cambridge, MA: Harvard University Press.
- Young, Iris Marion. 1990. *Justice and the Politics of Difference*. Princeton: Princeton University Press.
- -----. 1997. *Intersecting Voices: Dilemmas of Gender, Political Philosophy and Policy*. Princeton: Princeton University Press.
- -----. 2000. *Inclusion and Democracy*. Oxford: Oxford University Press.

Other important works

- Appiah, Anthony. 1992. *In My Father's House: Africa in the Philosophy of Culture*. New York: Oxford University Press.
- Appiah, Kwame Anthony and Henry Louis Gates, Jr. (eds.). 1995. *Identities*. Chicago: University of Chicago Press.
- Benhabib, Seyla. 1999. “The Liberal Imagination and the Four Dogmas of Multiculturalism.” *Yale Journal of Criticism* 12:2: 401-13.
- Benhabib, Seyla, ed. 1996. *Democracy and Difference: Contesting the Boundaries of the Political*. Princeton: Princeton University Press.

- Boxill, Bernard, ed. 2001. *Race and Racism*. Oxford: Oxford University Press.
- Bulkin, Elly, Minnie Bruce Pratt, and Barbara Smith. 1984. *Yours in Struggle: Three Feminist Perspectives on Anti-Semitism and Racism*. Ithaca, NY: Firebrand Books.
- Collins, Patricia Hill. 1991. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. New York: Routledge.
- Cuomo, Chris and Kim Hall (eds.). 2000. *Whiteness: Feminist Philosophical Reflections*. Lanham, MD: Rowman and Littlefield.
- Danielson, Dan and Karen Engle (eds.). 1995. *After Identity: A Reader in Law and Culture*. New York: Routledge.
- Gilroy, Paul. 1993. *The Black Atlantic: Modernity and Double Consciousness*. Cambridge: Harvard University Press.
- Goldberg, David Theo. 1993. *Racist Cultures: Philosophy and the Politics of Meaning*. Oxford: Blackwell.
- Hooks, Bell. 1990. *Yearning: Race, Gender, and Cultural Politics*. Boston: South End Press.
- Honneth, Axel. 1996. *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. Cambridge, MA: MIT Press.
- Johnson, Troy, Alvin M. Josephy and Joane Nagel (eds.). 1998. *Red Power: The American Indians' Fight for Freedom*. Lincoln: University of Nebraska Press.
- Laclau, Ernesto, ed. 1994. *The Making of Political Identities*. London: Verso.
- Lorde, Audre. 1984. *Sister/Outsider: Essays and Speeches*. Trumansburg, NY: The Crossing Press.
- Mills, Charles. 1998. *Blackness Visible: Essays on Philosophy and Race*. Ithaca, NY: Cornell University Press.
- Minow, Martha. 1997. *Not Only for Myself: Identity, Politics, and the Law*. New York: New Press.
- Moraga, Cherríe and Gloria Anzaldúa (eds.). 1981. *This Bridge Called My Back: Writings by Radical Women of Color*. Watertown, MA: Persephone Press.
- Narayan, Uma and Sandra Harding (eds.). 2000. *Decentering the Center: Philosophy for a Multicultural, Postcolonial, and Feminist World*. Bloomington: Indiana University Press.
- Nicholson, Linda and Steven Seidman (eds.). 1995. *Social Postmodernism: Beyond Identity Politics*. Cambridge: Cambridge University Press.
- Ryan, Barbara (ed.). 2001. *Identity Politics in the Women's Movement*. New York: New York University Press.
- Shklar, Judith. 1998. *Redeeming American Political Thought*. Chicago: University of Chicago Press.
- Simpson, David. 2002. *Situatedness: Or, Why We Keep Saying Where We're Coming From*. Durham, NC: Duke University Press.
- Stychin, Carl. 1998. *A Nation by Rights: National Cultures, Sexual Identity Politics, and the Discourse of Rights*. Philadelphia: Temple University Press.
- Touraine, Alain. 2000. *Can We Live Together? Equality and Difference*. Stanford, CA: Stanford University Press.
- Tremain, Shelley. 2001. "On the Government of Disability." *Social Theory and Practice* 27:4.
- Trinh, Minh-ha. 1989. *Woman, Native, Other: Writing Postcoloniality and Feminism*. Bloomington: Indiana University Press.

- Tully, James. 1995. *Strange Multiplicity: Constitutionalism in an Age of Diversity*. Cambridge: Cambridge University Press.
- Wing, Adrien Katherine (ed.). 1997. *Critical Race Feminism: A Reader*. New York: New York University Press.
- Zack, Naomi. 1997. *Thinking About Race*. Belmont, CA: Wadsworth.
- Zack, Naomi, Laurie Shrage and Crispin Sartwell (eds.). 1998. *Race, Class, Gender, and Sexuality: The Big Questions*. Malden, MA: Blackwell.

Other Internet Resources

- "Gender and Race: (What) Are They? (What) Do We Want Them To Be?", by Sally Haslanger (Philosophy/MIT)

[Please contact the author with further suggestions.]

Related Entries

difference | disability | essentialism | ethnicity | feminism | feminism, topics: feminist perspectives on the self | homosexuality | identity | liberalism | Marxism | multiculturalism | race and racism | recognition | self | sex and gender | sexuality | social groups | social ontology | subjectivity

Cressida J. Heyes
cressida.heyes@ualberta.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 15, 2002
Content last modified: July 18, 2002

Stanford Encyclopedia of Philosophy
Notes to Identity Politics

Notes

1. The term "Third World" is clearly problematic in that it indicates a part of the world that has fallen behind, or lost the race to development altogether, rather than highlighting the exploitation and over-development of the “First World.” The term “Fourth World” is used to describe the increasing numbers of very poor who live in the world's wealthy countries.

Cressida J. Heyes
cressida.heyes@ualberta.ca

First published: July 15, 2002
Content last modified: July 15, 2002

Feminist Perspectives on the Self

The topic of the self has long been salient in feminist philosophy, for it is pivotal to questions about personhood, identity, the body, and agency that feminism must address. In some respects, Simone de Beauvoir's trenchant observation, "He is the Subject, he is the Absolute -- she is the Other," sums up why the self is such an important issue for feminism. To be the Other is to be the non-subject, the non-person, the non-agent -- in short, the mere body. In law, in customary practice, and in cultural stereotypes, women's selfhood has been systematically subordinated, diminished, and belittled, when it has not been outright denied. Since women have been cast as lesser forms of the masculine individual, the paradigm of the self that has gained ascendancy in U.S. popular culture and in Western philosophy is derived from the experience of the predominantly white and heterosexual, mostly economically advantaged men who have wielded social, economic, and political power and who have dominated the arts, literature, the media, and scholarship. Responding to this state of affairs, feminist philosophical work on the self has taken three main tacks: (1) critique of established views of the self, (2) reclamation of women's selfhood, and (3) reconceptualization of the self to incorporate women's experience. This entry will survey feminist perspectives on the self from all three of these angles.

---

1. Critique

Two views of the self have been prominent in contemporary Anglo-American moral and political

philosophy -- a Kantian ethical subject and homo economicus. Both of these conceptions see the individual as a free and rational chooser and actor -- an autonomous agent. Nevertheless, they differ in their emphasis. The Kantian ethical subject uses reason to transcend cultural norms and to discover absolute moral truth, whereas homo economicus uses reason to rank desires in a coherent order and to figure out how to maximize desire satisfaction. Whether the self is identified with pure abstract reason or with the instrumental rationality of the marketplace, though, these conceptions of the self isolate the individual from personal relationships and larger social forces. For the Kantian ethical subject, emotional bonds and social conventions imperil objectivity and undermine commitment to duty. For homo economicus, it makes no difference what social forces shape one's desires provided they do not result from coercion or fraud, and one's ties to other people are to be factored into one's calculations and planning along with the rest of one's desires. Some feminist philosophers modify and defend these conceptions of the self. But their decontextualized individualism and their privileging of reason over other capacities trouble many feminist philosophers.

Twentieth century philosophy's regnant conceptions of the self minimize the personal and moral import of unchosen circumstances and interpersonal relationships. They eclipse family, friendship, passionate love, and community, and they downplay the difficulty of resolving conflicts that arise between these commitments and personal values and aspirations. Since dependency is dismissed as a defective form of selfhood, caregiving responsibilities vanish along with children, the disabled, and the frail elderly. Prevailing conceptions of the self ignore the multiple, sometimes fractious sources of social identity constituted by one's gender, sexual orientation, race, class, age, ethnicity, and so forth. Structural domination and subordination do not penetrate the "inner citadel" of selfhood. Likewise, these conceptions deny the complexity of the intrapsychic world of unconscious fantasies, fears, and desires, and they overlook the ways in which such materials intrude upon conscious life. The homogenized -- you might say sterilized -- rational subject is not prey to ambivalence, anxiety, obsession, prejudice, hatred, or violence. A disembodied mind, the body is peripheral -- a source of desires for homo economicus to weigh and a distracting temptation for the Kantian ethical subject. Age, looks, sexuality, and physical competencies are extraneous to the self. As valuable as the capacities for rational analysis and free choice undoubtedly are, it is hard to believe that there is nothing more to the self.

Feminist philosophers have charged that these views are, at best, incomplete and, at worst, fundamentally misleading. Many feminist critiques take the question of who provides the paradigm for these conceptions as their point of departure. Who models this free, rational self? Although represented as genderless, sexless, raceless, ageless, and classless, feminists argue that the Kantian ethical subject and homo economicus mask a white, healthy, youthfully middle-aged, middleclass, heterosexual MAN. He is pictured in two principle roles -- as an impartial judge or legislator reflecting on principles and deliberating about policies and as a self-interested bargainer and contractor wheeling and dealing in the marketplace. It is no accident that politics and commerce are both domains from which women have historically been excluded. It is no accident either that the philosophers who originated these views of the self typically endorsed this exclusion. Deeming women emotional and unprincipled, these thinkers advocated confining women to the domestic sphere where their vices could be neutralized, even transformed into virtues, in the role of submissive wife and nurturant mother.

Feminist critics point out, furthermore, that this misogynist heritage cannot be remedied simply by condemning these traditional constraints and advocating equal rights for women, for these conceptions of the self are themselves gendered. In western culture, the mind and reason are coded masculine, whereas the body and emotion are coded feminine (Lloyd 1992). To identify the self with the rational mind is, then, to masculinize the self. If selfhood is not impossible for women, it is only because they resemble men in certain essential respects -- they are not altogether devoid of rational will. Yet, feminine selves are necessarily deficient, for they only mimic and approximate the masculine ideal.

Problematic, as well, is the way in which these gendered conceptions of the self contribute to the valorization of the masculine and the stigmatization of the feminine. The masculine realm of rational selfhood is a realm of moral decency -- principled respect for others and conscientious fidelity to duty -- and of prudent good sense -- adherence to shrewd, fulfilling, long-range life plans. However, femininity is associated with emotionally rooted concern for family and friends that spawns favoritism and compromises principles. Likewise, femininity is associated with immersion in unpredictable domestic exigencies that forever jeopardize the best-laid plans and often necessitate resorting to hasty retreats or charting new directions. By comparison, the masculinized self appears to be a sturdy fortress of integrity. How flattering! The self is essentially masculine, and the masculine self is essentially good and wise.

Feminists object that this philosophical consolidation of the preeminence of the masculine over the feminine rests on untenable assumptions about the transparency of the self, the immunity of the self to noxious social influences, and the reliability of reason as a corrective to distorted moral judgment. Today people grow up in social environments in which culturally normative prejudice persists, even in communities where overt forms of bigotry are strictly proscribed (Meyers 1994). Although official cultural norms uphold the values of equality and tolerance, cultures continue to transmit camouflaged messages of the inferiority of historically subordinated social groups through stereotypes and other imagery. These deeply ingrained schemas commonly structure attitudes, perception, and judgment despite the individual's conscious good will (Valian 1998). As a result, people often consider themselves objective and fair, and yet they systematically discriminate against "different" others while favoring members of their own social group (Piper 1990; Young 1990). Fortified by culture and ensconced in the unconscious, such prejudice cannot be dispelled through rational reflection alone (Meyers 1994). In effect, then, the Kantian moral subject countenances "innocent" wrongdoing and occluded reinforcement of the social stratification that privileges the minority of men whom this conception takes as paradigmatic.

These oversights necessitate reconceptualizing the self in two respects. To account for the residual potency of this form of prejudice, feminists urge, the self must be understood as socially situated and murkily heterogeneous. To account for the self's ability to discern and resist culturally normative prejudice, the moral subject must not be reduced to the capacity for reason.

Complementing this line of argument, a number of feminists argue that conceptualizing the self as a seamless whole has invidious social consequences. To realize this ideal, it is necessary to repress inner diversity and conflict and to police the boundaries of the purified self. Alien desires and impulses are consigned to the unconscious, but this unconscious material inevitably intrudes upon conscious life and

influences people's attitudes and desires. In particular, the feared and despised Other within is projected onto "other" social groups, and hatred and contempt are redirected at these imagined enemies (Scheman 1993; Kristeva 1991). Misogyny and other forms of bigotry are thus borne of the demand that the self be unitary together with the impossibility of meeting this demand. Worse still, these irrational hatreds cannot be cured unless this demand is repudiated, but to repudiate this demand is to be resigned to a degraded, feminized self. Far from functioning as the guarantor of moral probity, the Kantian moral subject is the condition of the possibility of intractable animosity and injustice.

Another strand of feminist critique targets homo economicus's preoccupation with independence and planning. In an eery suspension of biological reality, selves are conceived as sufficient unto themselves. No one seems to be born and raised, for birth mothers and caregivers are driven offstage (Baier 1987; Code 1987; Held 1987; Benhabib 1987; Kittay 1999). The self appears to materialize on its own, endowed with a starter set of basic desires, ready to select additional desires and construct overarching goals, and skilled in performing instrumental rationality tasks. No one's powers ever seem to deteriorate either, for time is suspended along with biology. Since dependency is denied, no morally significant preconsensual or nonconsensual entanglements at the beginning or the end of life need be acknowledged. All affiliations are to be freely chosen, and all transactions are to be freely negotiated. The repudiation of feminine caregiving underwrites the illusion of independence, and the illusion of independence underwrites homo economicus's voluntarism.

To achieve maximal fulfillment, homo economicus must organize his chosen pursuits into a rational life plan. He must decide which desires are most urgent; he must ensure that his desires are co-satisfiable; and he must ascertain the most efficient way to satisfy this set of desires. Madcap spontaneity and seat-of-the-pants improvisation are registered as defeats for "The Man with the Plan." Not only is this vision of a life governed by a self-chosen plan distinctly middleclass, it is gendered (Addelson 1994; Walker 1999). The mother coping with the vagaries of early childhood and the wife accommodating her man's plan are the antitheses of this conception of the self. Uncertain of where they are ultimately headed and seldom sure how to achieve the goals they embrace as they go along, these women violate norms of selfhood. Ironically, middleclass men who grow old also have difficulty measuring up to homo economicus's standards of control. Unable to count on continued health and vigor, unable to anticipate the onset of serious disease or disabling conditions, unable finally to outwit the grim reaper, affluent elderly men violate norms of selfhood along with women and the poor. The price of denying the relationality of the self and idolizing rational self-regulation is that full selfhood eludes all but a lucky, albeit transitory, male elite.

A further problem with this view from a feminist standpoint is that it fails to furnish an adequate account of internalized oppression and the process of overcoming it. It is common for women to comport themselves in a feminine fashion, to scale down their aspirations, and to embrace gender-compliant goals (Bartky 1990; Babbitt 1993). Feminists account for this phenomenon by explaining that women internalize patriarchal values and norms -- that is, these pernicious values and norms become integrated in the cognitive, emotional, and conative structure of the self. Once embedded in a woman's psychic economy, internalized oppression conditions her desires. To maximize satisfaction of her desires, then, would be to collaborate in her own oppression. Paradoxically, the more completely she fulfills these

desires, the worse off she becomes. Advantaged as he is, homo economicus can safely accept his desires as given and proceed without ado to orchestrate a plan to satisfy them. But women and members of other subordinated groups can ill-afford such complacency, and homo economicus's instrumental reason is too superficial a form of mastery to serve their interests (Babbitt 1993). They need a conception of the self that renders emancipatory transformation of one's values and projects intelligible.

Feminist critique exposes the partiality of the ostensibly universal Kantian ethical subject and homo economicus. These conceptions of the self are: 1) androcentric because they replicate masculine stereotypes and ideals; 2) sexist because they demean anything that smacks of the feminine; and 3) masculinist because they help to perpetuate male dominance. I leave the heterosexist, racist, ethnocentric, ableist, and classist dimensions of these conceptions to other encyclopedia articles.

2. Reclamation

Feminist critiques, we have seen, accuse regnant philosophical accounts of masculinizing the self. One corollary of this masculinized view of selfhood is that women are consigned to selflessness -- that is, to invisibility, subservient passivity, and self-sacrificial altruism.

Feminist critiques, we have seen, accuse regnant philosophical accounts of masculinizing the self. One corollary of this masculinized view of selfhood is that women are consigned to selflessness -- that is, to invisibility, subservient passivity, and self-sacrificial altruism.

This nullification of women's selfhood was once explicitly codified in law. The legal doctrine of coverture held that a woman's personhood was absorbed into that of her husband when she married (McDonagh 1996). The wife's assuming her husband's surname symbolizes this revocation of her separate identity. In addition, coverture deprived the wife of her right to bodily integrity, for rape within marriage was not recognized as a crime, nor was it illegal for a husband to beat his wife. She lost her right to property, as well, for her husband was entitled to control her earnings, and she was barred from making contracts in her own name. Lacking the right to vote or to serve on juries, she was a second-class citizen whose enfranchised husband purportedly represented her politically.

Although coverture has been rescinded, vestiges of this denial of women's selfhood can be discerned in recent legal rulings, and the doctrine remains influential in culture. For example, pregnant women remain vulnerable to legally sanctioned violations of their right to bodily integrity. Courts have forced pregnant women to submit to invasive medical procedures for the sake of the fetuses they were carrying, although no court would compel any other woman or man to undergo comparable procedures for the sake of a living individual, including a family member (Bordo 1993). Selflessness remains the pregnant woman's legal status. Moreover, the stereotype of feminine selflessness still thrives in the popular imagination. Any self-confident, self-assertive woman is out of step with prevalent gender norms, and a mother who is not unstintingly devoted to her children is likely to be perceived as selfish and face severe social censure. Despite the fact that it is no longer legally mandatory for wives to give up their maiden names, many women adhere to this custom and perpetuate this traditional gesture of self-renunciation.

A tension within feminism complicates the project of reclaiming women's selfhood, however. The claim that women are systematically subordinated and that this subordination has a grievous impact on women's lives is central to feminism. Yet, this key insight seems to belie the claim that women's selfhood and agency have been overlooked. To be unjustly subordinated, it would seem, is to be diminished in one's selfhood and to have one's agency curtailed. Otherwise, what's the harm?

Some feminists have endorsed this very position. Arguing that moral virtues have no gender, Mary Wollstonecraft regards "feminine" virtues as perversions of true human virtues and laments women's conscription into a bogus ideal (Wollstonecraft 1792). Similarly but more vividly, Simone de Beauvoir labels women "mutilated" and "immanent" (Beauvoir 1952). Socialized to objectify themselves, women become narcissistic, small-minded, and dependent on others' approval. Excluded from careers, waiting to be chosen by their future husbands, taken over by natural forces during pregnancy, busy with tedious, repetitive housework, women never become transcendent agents. Indeed, they are content not to assume the burden of responsibility for their own freedom. Cast in the role of man's Other and at the mercy of feminine vices, women succumb to bad faith and surrender their agency.

This portrayal of women as abject victims has been challenged and modulated in contemporary feminist philosophy. I shall review four major reclamation strategies: 1) rethinking the activities of mothering, 2) developing an ethic of care, 3) exploring separatist practices, and 4) reconceiving autonomy.

The conventional view of pregnancy and birth classifies them as merely biological processes, and the conventional view of mothering classifies it as a merely instinctual activity. Feminists demonstrate that these assessments are sorely mistaken. Both pregnancy and birthgiving engage women's agentic powers. Not only does pregnancy raise the question of whether to have an abortion, but also a woman's decision to proceed with a pregnancy entails learning to care for herself in previously unnecessary ways (Held 1989). In the last few decades, medical technologies, such as sonography and fetal surgery, have raised new issues for pregnant women and sometimes confront them with wrenching choices that test their agentic resilience. Arguably, routine pregnancy and birthgiving mobilize specific agentic capacities, such as "active waiting" and coping with "chosen and predictable pain" (Ruddick 1994).

A related feminist innovation focuses on analyzing the discipline of mothering to grasp its aims, its forms of thought, its ideal form, and its characteristic values and disvalues (Ruddick 1989). Caring for a child imposes a set of demands -- for preservation (survival), for growth (development into a healthy adult), and for acceptability (enculturation that ensures fitting into a community). Meeting these demands involves a range of activities that are governed by a distinctive set of values: protecting a fragile existence, acknowledging the limits of one's power and the unpredictability of events, cheerful determination to persist despite setbacks, responsive adaptability, sensitivity to the child's subjective viewpoint, and tolerance for inconclusive processes of disclosure. Although the practice of mothering places no premium on independence, self-interest, free choice, power, advance planning, or control, it clearly calls upon a wide range of interpersonal and reflective skills and enlists caregivers' agentic capacities. Dumb instinct hardly suffices for good childcare.

Like feminists who have reclaimed women's agency as mothers, feminists who have developed different versions of care ethics insist on taking women's experience seriously and use this experience as a basis for new approaches to morality and social policy. The aim of the psychological studies that first made the voice of care audible was to recognize and understand the capacities for moral judgment of women whose competency had been underrated. Previous research comparing boys' and girls' moral development had concluded that girls' development was stunted, but Carol Gilligan argues that this assessment misconstrued the data (Gilligan 1982). According to Gilligan, there are two paths of moral development. Many girls and women but almost no men follow the care trajectory (Gilligan 1987). Since earlier investigations first studied U.S. boys and men and used these interviews to generalize about people's moral development, researchers noticed only one path, namely, the justice trajectory. By repudiating the assumption that the masculine is the human norm and by studying girls and women, Gilligan discovered an alternative mode of moral cognition -- the Care Perspective. Constituted by a distinctive set of framing concepts and a distinctive set of reflective skills, the morality of care is not translatable into the morality of justice that Gilligan's predecessors had taken to be the gauge of moral development. The Care Perspective, in Gilligan's view, is a different and equally good way to interpret moral situations and to decide how to act. Moreover, by noticing this alternative, we are able to recognize women's moral agency and defend women against the age-old charge that they are morally inferior to men.

Although some feminist philosophers criticize Gilligan's investigations on empirical or philosophical grounds (Moody-Adams 1991; Friedman 1993; Card 1996; Fraser and Nicholson 1990), her research prompted a number of feminist philosophers to develop moral theories marked by quite different emphases from those of traditional moral theories. The theme of human interconnectedness and the value of intersubjectivity are prominent in contemporary feminist moral philosophy. A climate of trust forms an indispensable background for all sorts of undertakings, but no voluntaristic ethic can account for trust (Baier 1986). The ability to empathize with other individuals and imaginatively reconstruct their unique subjective viewpoints is vital to moral insight and wise moral choice, but ethics that base moral judgment on a universal conception of the person marginalize this skill (Meyers 1994). By developing narratives of one's moral identity, one's relationships, and one's values and sharing those narratives with one's associates, one endows one's life with moral meaning and integrity, but rationalistic ethics overlook this process of self-disclosure and interpersonal mediation (Walker 1998). Taking responsibility for who one is and how one shall respond is a salient feature of informal personal relationships, yet justice oriented ethics focus exclusively on being held responsible for what one has done and the credit or blame one's actions may deserve (Card 1996). Appreciating the inevitability of dependency and the need for care demonstrates the poverty of conceiving justice exclusively in terms of rights not to be interfered with and the urgency of developing a theory of justice that includes provisions for care (Kittay 1999). In each instance, feminist moral theorists revalue that which is traditionally deemed feminine -- feeling, intimacy, nurturance, and so forth. By highlighting these contexts and values, they reclaim the venues traditionally associated with women as morally significant sites, and they reclaim the moral agency of the individuals whose lives are centered in these sites.

A third approach to reclaiming women's agency spotlights several types of separatist practice -- including friendship among women, lesbianism, support groups for rape victims and battered women, and women's consciousness raising and activist groups. Establishing and maintaining such affiliations presupposes self-

willed defiance of norms of heterosexual fidelity and familial commitment. Thus, the very existence of these relationships testifies to women's awareness of their own needs and their capacity to act on them despite a repressive social context. Moreover, noting that unchosen relationships and communities of origin often prove oppressive to women and inimical to their agency, some feminists stress that it is important not to underestimate the role of chosen relationships and communities as sites of women's agency (Friedman 1993; Brison 1997; Hoagland 1988; Ferguson 1987; Frye 1983; MacKinnon 1982; Hartsock 1983). By cordoning off a social sphere of mutually attuned, mutually concerned women, separatism in all its forms turns down the racket of patriarchy. Separatist associations provide forums in which women can exchange personal confidences, secure in the knowledge that other participants will empathize with their dissatisfactions and frustrations as well as their joys and triumphs and that others will be receptive to their worries and complaints. Isolation often confounds women's subjectivity and agency, for in isolation their problems appear to be personal failings if not pathologies. Separatism overcomes isolation. It affords women the opportunity to develop language that makes sense of their anomalous experience and that restores their self-esteem together with the opportunity to reflect on the social meanings of their experience. Within separatist contexts, women find support for their resistance to social norms and their struggles to overcome personal privations or pressures.

Separatist practices relieve women of the burden of Otherness. Each woman is an equal among equals. Each woman's subjectivity and agency are affirmed. Thus, feminist philosophers seek not only to chronicle the ways in which women have created pockets of separatism within patriarchal systems but also to theorize the forms of subjectivity and agency that flourish in these sites.

Autonomy is a key issue for this theoretical project. Although some feminists dismiss autonomy as an androcentric relic of modernism (Jaggar 1983; Addelson 1994; Hekman 1995; Card 1996), others assert women's need for self-determination (de Lauretis 1986; King 1988; Lugones and Spelman 1983; Govier 1993). In light of the history of figuring women as driven by their reproductive biology and in need of rational male guidance and the history of women's enforced economic dependence on men or relegation to poorly paid, often despised forms of labor, feminists can hardly ignore the topic of self-determination. Thus, a number of feminist philosophers take up this challenge and present accounts of autonomy that do not devalue the interpersonal capacities and social contributions that are conventionally coded feminine (Nedelsky 1989; Meyers 1989 and 2000; Benhabib 1995 and 1999; Weir 1995). In feminist accounts, autonomy is not conflated with self-sufficiency and free will, but rather it is seen to be facilitated by supportive relationships and also to be a matter of degree.

Whereas standard philosophical accounts of autonomy confine their social critique to the observation that pluralistic societies that respect basic rights are more conducive to autonomy, feminist accounts point up how subordination constrains autonomy and how egalitarian communities augment it (Meyers 1989; Babbitt 1993; Benhabib 1995). Whereas standard philosophical accounts intellectualize autonomy and stress rational decision making, feminist accounts accent the role of feelings in autonomous lives (Nedelsky 1989; Meyers 1989; Weir 1995). Whereas standard philosophical accounts spotlight the autonomous individual's independence and immunity to others' influence, feminist accounts stress the autonomous individual's need for constructive feedback, advice, and encouragement (Meyers 1989; Brison 1997). Whereas standard philosophical accounts trace autonomy to endorsing and prioritizing a

coherent set of desires and goals and to scheduling fulfillment of these objectives, feminist accounts view autonomy as an ongoing and improvisational process of exercising self-discovery, self-definition, and self-direction skills (Meyers 1989 and 2000). Whereas standard philosophical accounts see autonomy as an all-or-nothing achievement, feminist accounts note how autonomy skills piggyback on seemingly unrelated ancillary skills, how autonomy skills may be exercised in certain contexts yet deactivated in others, and how different degrees of skillfulness yield varying degrees of autonomy (Friedman 1993; Meyers 1989).

Feminist accounts of autonomy strike a balance between recognizing the injury that subordination does to women's sense of self and agency and respecting the measure of autonomy women gain despite this subjugation. Subordination endangers women's autonomy in a number of ways. Not only does internalized oppression mold women's desires and alienate them from themselves, but also those in subordinate positions are offered all sorts of incentives to minimize friction and ease their lot by placating those with power (Card 1996). Likewise, well-meaning friends are all too likely to counsel the course of least resistance, namely, compliance with convention regardless of one's personal values and aspirations. Another effect of systematic subordination is that women's autonomy skills may be poorly developed or poorly coordinated, and exercising these skills is rarely rewarded and generally discouraged (Meyers 1989). Deficient autonomy skills compound the threat internalized oppression poses.

Still, feminist accounts of autonomy enable us to understand why women do not completely lack autonomy and how women's autonomy can be augmented. The self-discovery, self-definition, and self-direction skills that secure autonomy are commonplace (Meyers 1989). Indeed, some of them,such as introspective attunement to feelings and receptiveness to others' feedback, are gender-compatible for and often promoted in women. Although others, such as rational planning and self-assertion, are coded masculine, many women in fact have considerable proficiency in these areas. All too often, however, they exercise these skills only in narrowly restricted, gender-appropriate contexts. For example, a homemaker may demonstrate remarkable instrumental reason skills in running her household, or a mother may exhibit effective self-assertion skills in dealing with a teacher who has mistreated her child. Yet, these women may come off as inept, helpless, and meek in other situations. Thus, augmenting women's autonomy is often a matter of emboldening women to extend the range of application of their existing autonomy skills and fostering the development of weak skills. It is evident, then, why separatist practices of various kinds are conducive to women's autonomy. By inviting women to marshall their autonomy skills and reinforcing women's determination to carry out their decisions, they function as autonomy workshops.

Still, from a feminist perspective, separatist practices are merely transitional and ameliorative. In addition, the patriarchal social structures that relentlessly undermine women's autonomy must be changed, and women's selfhood and agency must be legally and culturally affirmed. Thus, feminist philosophers defend a variety of social policy initiatives that expand the scope of women's choices and that respect women as self-directing individuals. Feminist philosophers have been in the forefront in arguing for egalitarian families, in legitimating economic opportunity for women, in opposing harassment of women in workplaces, in defending women's reproductive rights, and in condemning violence against women in all its forms. In each instance, greater justice for women strikes a blow against the masculinized self of traditional philosophy by securing greater social recognition for the female agentic self.

3. Reconceptualizations

A. The Nature of the Self

The primary task of a philosophy of the self is to clarify what makes something a self. Feminist philosophers are acutely aware that this is not a value-free task. To get an analysis of the nature of the self off the ground, one must decide which entities count as selves (or, at least, which entities are noncontroversially counted as selves within one's linguistic community). Since we regard selves as valuable -- as members of our moral community and as worthy of respect -- these judgments are in part judgments about which entities are valuable. Moreover, values enter into these judgments because we consider selves to be the sorts of things that can achieve (or fail to achieve) ideals of selfhood. Thus, philosophical accounts of the self have implications for conceptions of what it is to lead a good life. As we have seen, many feminist philosophers argue that it is a mistake to hold that rationality alone is essential to the self and that the ideal self is transparent, unified, coherent, and independent, for they discern misogynist subtexts in the atomistic individualism of the Kantian ethical subject and homo economicus (see Section 1). It is incumbent on feminist philosophers, then, to develop more satisfactory accounts of the self -- accounts that are compatible with respect for women. Thus, a number of feminist philosophers propose reconstructions of alternative theories of the nature of the self.

Three traditions have been especially influential in recent feminist thought -- classic psychoanalysis, object relations theory, and poststructuralism. Feminist philosophers gravitate toward these approaches to understanding selfhood because they do not share the drawbacks that prompt feminist critique of the Kantian ethical subject and homo economicus. None of these approaches regards the self as homogeneous or transparent; none supposes that a self should be coherent and speak in a single voice; none removes the self from its cultural or interpersonal setting; n one sidelines the body. In appropriating these views, feminists bring out their implications in regard to gender, incorporate feminist insights into these theories, and modify the theories to address feminist concerns.

Julia Kristeva transposes the classic Freudian conception of the self and the distinction between consciousness and the unconscious into an explicitly gendered discursive framework (Kristeva 1980). For Kristeva, the self is a subject of enunciation -- a speaker who can use the pronoun ‘I’. But speakers are not unitary, nor are they fully in control of what they say because discourse is bifurcated. The symbolic dimension of language, which is characterized by referential signs and linear logic, corresponds to consciousness and control. The clear, dry prose of scientific research reports epitomizes symbolic discourse. The semiotic dimension of language, which is characterized by figurative language, cadences, and intonations, corresponds to the unruly, passion-fueled unconscious. The ambiguities and nonstandard usages of poetry epitomize semiotic discourse. These paradigms notwithstanding, Kristeva maintains that all discourse combines elements of both registers. Every intelligible utterance relies on semantic conventions, and every utterance has a tone, even if it is a dull monotone. This contention connects Kristeva's account to feminist concerns about gender and the self. Since the rational orderliness of the symbolic is culturally coded masculine while the affect-laden allure of the semiotic is culturally coded

feminine, it follows that no discourse is purely masculine or purely feminine. The masculine symbolic and the feminine semiotic are equally indispensable to the speaking subject, whatever this individual's socially assigned gender may be. It is not possible, then, to be an unsulliedly masculine self or an unsulliedly feminine self. Every subject of enunciation -- every self -- amalgamates masculine and feminine discursive modalities.

Like the unconscious in classic psychoanalytic theory, the semiotic decenters the self. One may try to express one's thoughts in definite, straightforward language, yet because of the semiotic aspects of one's utterances, what one says carries no single meaning and is amenable to being interpreted in more than one way. In Kristeva's view, this is all to the good, for accessing the semiotic -- that which is conveyed, often inadvertently, by the style of an utterance -- kindles social critique. The semiotic gives expression to repressed, unconscious material. According to Kristeva, what society systematically represses provides clues to what is oppressive about society and how society needs to be changed. Thus, she discerns a vital ethical potential in the semiotic (Kristeva 1987). Since this ethical potential is explicitly linked to the feminine, moreover, Kristeva's account of the self displaces "masculine" adherence to principle as the prime mode of ethical agency and recognizes the urgent need for a "feminine" ethical approach. Viewing the self as a "questionable-subject-in-process" -- a subject who is responsive to the encroachments of semiotic material into conscious life and who is therefore without a fixed or unitary identity -- and valorizing the dissident potential of this decentered subjectivity, Kristeva seeks to neutralize the fear of the inchoate feminine that, in her view, underwrites misogyny. In one respect, Nancy Chodorow's appropriation of object relations theory parallels Kristeva's project of reclaiming and revaluing femininity, for Chodorow's account of the relational self reclaims and revalues feminine mothering capacities. But whereas Kristeva focuses on challenging the homogeneous self and the bright line between reason, on the one hand, and emotion and desire, on the other, Chodorow focuses on challenging the self-subsisting self with its sharp self-other boundaries. Chodorow's claim that the self is inextricable from interpersonal relationships calls into question the decontextualized individualism of the Kantian ethical subject and homo economicus.

Chodorow sees the self as relational in several respects (Chodorow 1981). Every child is cared for by an adult or adults, and every individual is shaped for better or worse by this emotionally charged interaction. As a result of feelings of need and moments of frustration, the infant becomes differentiated from its primary caregiver and develops a sense of separate identity. Concomitantly, a distinctive personality emerges. By selectively internalizing and recombining elements of their experience with other people, children develop characteristic traits and dispositions. Moreover, Chodorow attributes the development of a key interpersonal capacity to nurturance. A caregiver who is experienced as warmly solicitous is internalized as a "good internal mother" (Chodorow 1980). Children gain a sense of their worthiness by internalizing the nurturance they receive and directing it toward themselves, and they learn to respect and respond to other people by internalizing their experience of nurturance and projecting it toward others. Whereas Kristeva understands the self as a dynamic interplay between the feminine semiotic and the masculine symbolic, Chodorow understands the self as fundamentally relational and thus linked to cultural norms of feminine interpersonal responsiveness. For Chodorow, the rigidly differentiated, compulsively rational, stubbornly independent self is a masculine defensive formation -- a warped form of the relational self -- that develops as a result of fathers' negligible involvement in childcare.

Feminist philosophers have noted strengths and weaknesses in both of these views. For example, Kristeva's questionable-subject-in-process seems to enshrine and endorse the very gender dichotomy that causes women so much grief. Yet, Chodorow's relational self seems to glorify weak individuation and scorn the independence and self-assertiveness that many women desperately need. Still, Kristeva's analyses of the psychic, social, and political potency of gender figurations underscore the need for feminist counter-imagery to offset culturally entrenched, patriarchal images of womanhood. And Chodorow's appreciation of the relational self together with her diagnosis of the damage wrought by hyperindividuation advances feminist demands for equitable parenting practices. These contributions notwithstanding, both of these views have come under attack for heterosexist biases as well as for inattention to other forms of difference among women.

Critical race theorists and poststructuralists have been particularly vocal about this failure to come to grips with the diversity of gender, and they have offered accounts of the self designed to accommodate difference. Poststructuralist Judith Butler maintains that personal identity -- the sense that there are answers to the questions ‘who am I?’ and ‘what am I like?’ -- is an illusion (Butler 1990). The self is merely an unstable discursive node -- a shifting confluence of multiple discursive currents -- and sexed/gendered identity is merely a "corporeal style" -- the imitation and repeated enactment of ubiquitous norms. For Butler, psychodynamic accounts of the self, including Kristeva's and Chodorow's, camouflage the performative nature of the self and collaborate in the cultural conspiracy that maintains the illusion that one has an emotionally anchored, interior identity that is derived from one's biological nature, which is manifest in one's genitalia. Such accounts are pernicious. In concealing the ways in which normalizing regimes deploy power to enforce the performative routines that construct "natural" sexed/gendered bodies together with debased, "unnatural" bodies, they obscure the arbitrariness of the constraints that are being imposed and deflect resistance to these constraints. The solution, in Butler's view, is to question the categories of biological sex, polarized gender, and determinate sexuality that serve as markers of personal identity, to treat the construction of identity as a site of political contestation, and to embrace the subversive potential of unorthodox performances and parodic identities.

African American feminists are less sanguine than poststructuralists about the felicitous social impact of playful deviations from norms and the laughter they may prompt (Williams 1991; Crenshaw 1993). Nevertheless, some of them have adapted poststructuralist theory to the purposes of critical race theory. Noting that gender, race, and class stratification do not operate in isolation from one another but rather interact to produce compound effects, these theorists conceive of the individual as an intersectional subject -- a site where structures of domination and subordination converge (King 1988; Crenshaw 1993). Intersectional theory does not purport to offer a comprehensive theory of the self. Its aim is to capture those aspects of selfhood that are conditioned by membership in subordinated or privileged social groups. Accenting the liabilities of belonging to more than one subordinated group, Kimberle Crenshaw likens the position of such individuals to that of a pedestrian hit by several speeding vehicles simultaneously, and Maria Lugones likens their position to that of a stateless border-dweller who is not at home anywhere (Crenshaw 1991; Lugones 1992). Nevertheless, some theorists of mixed ancestry embrace border-dwelling as a model of positive identity (Anzaldua 1987; Alcoff 1995). Moreover, proponents of the intersectional self credit multiply oppressed people with a certain epistemic advantage. In virtue of their

suffering and alienation, these individuals are well situated not only to discern which values and practices in their heritage deserve allegiance but also to identify shortcomings in the traditions of the groups to which they belong. Thus, African American women are acutely aware of racism within feminism and sexism within the struggle for racial justice. Their intersectional positioning and subjectivity makes such insight virtually unavoidable.

By and large, recent feminist philosophy of the self reflects skepticism about modernist, unitary accounts of the self. In seeking to remedy the androcentric biases of the latter views, feminist philosophers emphasize features of selfhood that other philosophical schools neglect, including intersubjectivity, heterogeneity, and social construction. Still, some contemporary feminist philosophers express concern that the sorts of conceptions I have sketched are detrimental to feminist aims. Influenced by Jurgen Habermas's communicative ethics, Seyla Benhabib refuses to join poststructuralists in declaring the death of the autonomous, self-reflective individual who is capable of taking responsibility and acting on principle (Benhabib 1995). Although Benhabib is committed to viewing people as socially situated, interpersonally bonded, and embodied, she is also committed to the feasibility of rational philosophical justification of universal moral norms. Moreover, she argues that a narrative conception of the self renders the idea of a core self and coherent identity intelligible without suppressing difference and without insulating the self from social relations (Benhabib 1999). Autobiographical stories can include the many voices within us and the many relationships we have experienced, and these stories are constantly under revision, for they are always being contested by our associates' disparate self-narratives with their divergent versions of events. Nevertheless, these narratives do not collapse into incoherence, and they presuppose a core capacity to describe and reflect on one's experience. For Benhabib, this view of selfhood and reason is indispensable to feminist emancipatory objectives.

B. Gender and Identity

Postmodern challenges to the idea of a stable self and to the coherence of the category woman' have sparked a lively debate about the relation between gender and the self. If there is no such thing as a self with persistent attributes, it seems that gender cannot be a feature of every woman's identity. But if there is nothing that all women have in common, it seems that there are no interests that all women share, and there is nothing for feminism to be about. Several feminist philosophers have proposed accounts of the relation between gender and the self that aim to rescue feminism from this reductio. Linda Alcoff rejects both the universalized conceptions of gender that cultural feminists advocate and the deconstructions of the category ‘woman’ that poststructuralist feminists tender. Her alternative is to construe femininity as "positionality." Positionality has two dimensions (Alcoff 1994). First, it is the social context that locates the individual and that deprives her of power and mobility. Second, it is a political point of departure -- the affirmation of women's collective right to take charge of their gendered identity. To be a woman is, then, to be deprived of equality, and to be a feminist is to take responsibility for redressing this wrong and for redefining the meaning of being a woman. Alcoff salvages the category ‘woman’ by defending an interpretation of the social meaning of being assigned to that category.

Iris Young introduces an additional layer of analysis. To explain gender, Young invokes Jean-Paul Sartre's idea of seriality (Young 1994). A social series is "a social collective whose members are unified

passively by the objects around which their actions are oriented or by the objectified results of the material effects of the actions of others." In other words, a series is constituted by a behavior-directing, meaning-defining environment. The lives of series members are affected by being assigned to particular social series, for serial existence is experienced as a "felt necessity." People feel impelled to act in ways that are consonant with their series memberships. Yet, series membership "does not define the person's identity in the sense of forming his/her individual purposes, projects, and sense of self in relation to others." Indeed, a woman "can choose to make none of her serial memberships important for her sense of identity." For Young, then, a gendered self is optional although membership in the series ‘woman’ is not.

Sally Haslanger underscores three conceptions of ‘gender’, and her distinctions clear up much of what seems puzzling in Alcoff's and Young's views (Haslanger 2000). Haslanger argues in favor of a "critical analytical" approach to gender -- that is, focusing on the "work the concepts of gender and race might do for us in a critical -- specifically feminist and antiracist -- social theory" and suggesting "concepts that can accomplish at least some important elements of that work." To pursue this politicized approach to gender, it is necessary to decide which conception of gender best serves the aims of emancipatory political theory and politics. As Haslanger points out, feminist theorists have interpreted gender as the experience of sexed embodiment, a broad psychological orientation to the world, a set of internalized norms, a system of sexual symbolism, a set of traditional roles, and a social position or class. In her view, the principal task for feminist theory is to provide an analysis of gender as a "pattern of social relations that constitute the social classes of men as dominant and women as subordinate." The other forms of gender, including gendered identity, should be explained in terms of this fundamental structure.

Like Alcoff and Young, Haslanger is acutely aware of the political damage essentialist theories of gender have done -- especially, the alienation of women of color and lesbians from feminism -- and also the factual shortcomings of essentialist theories of gender -- that is, many women do not fit the proposed accounts of what it is to be a woman. Although none of these feminist philosophers repudiates the conception of gender as a gendered self, each treats this conception as secondary or derivative. Alcoff politicizes gendered identity and ties it to a self-ascribed commitment to progressive social change. Young stresses the indeterminacy of gendered identity and allows that women's gender-based political commitments can be antifeminist as well as feminist. Although Haslanger insists that the connections between gender as social class and gendered identities are highly variable, she claims that gender "implicates each of us at the heart of our self-understandings," and she advocates conscientious reflection on "who we think we are."

Alcoff remarks that every woman's subjectivity is engendered; Young observes that every woman's identity is marked by gender; and Haslanger notes that every woman is invested in her gender. From the standpoint of a feminist philosophy of the self, it is crucial to account for this engendering, marking, and investment. In my view, the alternative to a common homogeneous feminine identity is gendered and individualized identities (Meyers forthcoming 2000). Individualization does not, however, entail optionality, for gender insinuates itself into identity in ways that we may not be conscious of and in ways that we may not be able to change no matter how much we try. Gender is constitutive of who we are -- our personalities, our capabilities and liabilities, our aspirations, and how we feel about all of these attributes. Yet, there is no feature of identity that all women share. How is this possible? Nancy

Chodorow uses psychoanalytic theory to make sense of individualized, gendered identities (Chodorow 1995). Psychoanalysis explains how individuals' affective dispositions, unconscious fantasies, and interpersonal relationships filter the culturally entrenched conception of gender they encounter. Through various psychic processes -- projection and introjection together with the defense mechanisms -- gender acquires a "personal meaning" that is inspired by, but that does not wholly replicate culturally transmitted strictures and iconography.

It is a mistake to picture gender as a toxic capsule full of norms and interpretive schemas that individuals swallow whole and that lodges intact in their psychic structure. The diversity of individuals' experience of gender belies this view. But it is also a mistake to picture attributes like gender as systems of social and economic opportunities, constraints, rewards, and penalties that need not impinge upon individual identity. The seeming naturalness of enacting gendered characteristics, the passion with which people cling to their sense of their gender, and the intractability of many gendered attributes when people seek to change them testify to the embeddedness of gender in identity. Still, it is important to recognize, as Alcoff, Young, Haslanger, and Chodorow do, that the potency of the impact of gender on the self does not altogether deprive women of control over their gendered attributes. Neither personal resistance to one's own gendered dispositions and evaluative standards nor political resistance to the social structures that gender women's identities is ruled out on this view (See Section 2).

4. Conclusion

As this article attests, there is tremendous foment and variety within the field of feminist work on the self. Yet, in reviewing this literature, I have been struck by a recurrent theme -- the inextricability of metaphysical issues about the self from moral and political theory. Feminist critiques of regnant philosophical theories of the self expose the normative underpinnings of these theories. Feminist analyses of women's agentic capacities both acknowledge traditional feminine social contributions and provide accounts of how women can overcome oppressive norms and practices. Feminist reconstructions of the nature of the self are interwoven with arguments that draw out the emancipatory benefits of conceiving the self one way rather than another. There is nothing surprising, to be sure, about the salience of normative concerns in feminist philosophizing. Still, I mention it because I believe that feminists' attention to political concerns leads to fresh questions and also that asking novel questions enriches philosophical understanding of the self. Moreover, I would urge that this forthrightness about the political viewpoint that informs philosophy is a virtue, for overlooking the political suppositions and implications of esoteric philosophical views has led to considerable mischief.

Bibliography

Comprehensive Bibliography

In the interests of concision and readability, the present essay mentions only some of the representative works on the feminist literature on the self. These cited works are collated in the Bibliography which

appears in the next section of this essay. However, the feminist literature on the self is vast. Lisa Cassidy has put together a comprehensive bibliography of this literature; it attempts to cite all of the books and articles that are relevant to the present entry. This comprehensive bibliography is linked into the present essay as the following supplementary document:

Comprehensive Bibliography of Feminist Perspectives on the Self (by Lisa Cassidy)

Readers are therefore encouraged to pursue additional references by following the above link.

References

The following works are cited in the entry:

- Addelson, Kathryn Pyne. 1994. *Moral Passages:* New York: Routledge.
- Alcoff, Linda. 1994. "Cultural Feminism versus Post-Structuralism: The Identity Crisis in Feminist Theory." In *Culture/Power/History* , eds., Nicholas Dirks et. al. Princeton: Princeton University Press.
- Alcoff, Linda. 1995. "Mestizo Identity." In *American Mixed Race: The Culture of Microdiversity* , ed., Naomi Zack, 257-278. Lanham, MD and London: Rowman and Littlefield, 1995.
- Anzaldua, Gloria. 1987. *Borderlands: The New Mestiza/La Frontera.* San Fransisco: Spinters/Aunt Lute.
- Babbitt, Susan E. 1993. "Feminism and Objective Interests? The Role of Transformation Experiences in Rational Deliberation." In *Feminist Epistemologies,* eds., Linda Alcoff and Elizabeth Potter. New York: Routledge, 1993.
- Baier, Annette. 1997. "Trust and Anti-Trust."*Ethics* 96 (January 1986): 231-260. ALSO IN: *Feminist Social Thought,* ed., Diana Tietjens Meyers. New York: Routledge, 1997.
- Baier, Susan. 1987. "The Need for More than Justice." In *Science, Morality, and Feminist Theory,* eds., Marsha Hanen and Kai Nielsen. Minneapolis: University of Minnesota Press.
- Bartky, Sandra Lee. 1990. *Femininity and Domination.* New York: Routledge.
- Beauvoir, Simone de. 1952. *The Second Sex,* H.M. Parshley (Trans). New York: Vintage Press.
- Benhabib, Seyla. 1999. "Sexual Difference and Collective Identities: The New Global Constellation." *Signs* 24: 335-361.
- Benhabib, Seyla and Drucilla Cornell, eds. 1987. *Feminism as Critique.* Minneapolis: University of Minnesota Press.
- Benhabib, Seyla, et. al. 1995. *Feminist Contentions.* New York: Routledge.
- Bordo, Susan. 1993. *Unbearable Weight.* Berkeley: University of California Press.
- Brison, Susan J. 1997. "Outliving Oneself: Trauma, Memory, and Personal Identity." In *Feminists Rethink the Self,* ed., Diana Tietjens Meyers. Boulder: Westview Press.
- Butler, Judith. 1990. "Gender Trouble, Feminist Theory, and Psychoanalytic Discourse." In *Feminism/Postmodernism,* ed., Linda Nicholson. New York: Routledge.
- Card, Claudia. 1996. *The Unnatural Lottery.* Philadelphia: Temple University Press.
- Chodorow, Nancy. 1980. "Gender, Relation, and Difference in Psychoanalytic Perspective," In

The Future of Difference, ed., Hester Eisenstein and Alice Jardine, Alice. Boston: G.K. Hall.

- Chodorow, Nancy. 1981. "On The Reproduction of Mothering: A Methodological Debate." *Signs* 6: 500-514.
- Chodorow, Nancy. 1995. "Gender as Personal and Cultural Construction.", *Signs* 20 (Spring): 516-544.
- Code, Lorraine. 1987. "Second Persons." In *Science, Morality, and Feminist Theory* (Supplement to *Canadian Journal of Philosophy* 13), ed., Marsha Hanen and Kai Nielsen. Calgary: University of Calgary Press.
- Crenshaw, Kimberlé. 1991. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics." In *Feminist Legal Theory,* eds. Katherine T. Bartlett and Rosanne Kennedy. Boulder: Westview Press.
- Crenshaw, Kimberlé. 1993. "Beyond Race and Misogyny: Black Feminism and 2 Live Crew." In *Words that Wound,* ed. Mari J. Matsuda, et. al.. Boulder: Westview Press.
- de Lauretis, Teresa. 1986. "Feminist Studies/Critical Studies: Issues, Terms, Contexts." In *Feminist Studies/Critical Studies,* ed., Teresa de Lauretis. Bloomington: Indiana University Press.
- Ferguson, Ann. 1987. "A Feminist Aspect Theory of Self." In *Science, Morality and Feminist Theory* (Supplement to Canadian Journal of Philosophy 13), eds., Marsha Hanen and Kai Nielsen, Kai. Calgary: University of Calgary Press.
- Fraser, Nancy and Linda Nicholson. 1990. "Social Criticism without Philosophy." In *Feminism/Postmodernism*, ed. Linda Nicholson. New York: Routledge.
- Friedman, Marilyn A. 1993. *What are Friends For?*. Ithaca: Cornell University Press.
- Frye, Marilyn. 1983. *The Politics of Reality.* Trumansburg: Crossing Press.
- Gilligan, Carol. 1982. *In a Different Voice.* Cambridge: Harvard University Press.
- Gilligan, Carol. 1987. "Moral Orientation and Moral Development." In *Women and Moral Theory,* eds., Eva Feder Kittay and Diana T. Meyers. Totowa: Rowman and Littlefield.
- Govier, Trudy. 1993. "Self-Trust, Autonomy, and Self-Esteem." *Hypatia,* 8:1 (Winter): 99-120.
- Hartsock, Nancy. 1983. *Money, Sex, Power.* New York: Longman.
- Haslanger, Sally. 2000. "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" *Nous* 34:1 (March): 31-55.
- Hekman, Susan J. 1995. *Moral Voices, Moral Selves.* University Park: Pennsylvania State University Press.
- Held, Virginia. 1987. "Feminism and Moral Theory." In *Women and Moral Theory,* eds., Eva Feder Kittay and Diana T. Meyers. Totowa: Rowman and Littlefield. ALSO IN: *Feminist Social Thought* (Meyers 1997)
- Held, Virginia. 1989. "Birth and Death." *Ethics* 99 (January): 362-388.
- Hoagland, Sarah Lucia. 1988. *Lesbian Ethics.* Palo Alto: Institute for Lesbian Studies.
- King, Deborah K. 1988. "Multiple Jeopardy, Multiple Consciousness: The Context of a Black Feminist Ideology." *Signs* 14(1): 42-72. ALSO IN: *Feminist Social Thought* (Meyers 1997).
- Kittay, Eva Feder. 1999. *Love's Labor.* New York: Routledge.
- Kristeva, Julia. 1980. *Desire in Language*, eds., Thomas Gora, Alice Jardine, and Leon Roudiez (Trans). New York: Columbia University Press.
- Kristeva, Julia. 1987. *Tales of Love*, Leon Roudiez (Trans). New York: Columbia University Press.

- Kristeva, Julia. 1991. *Strangers to Ourselves,* Leon S. Roudiez (Trans.). New York: Columbia University Press.
- Lloyd, Genevieve. 1992. "Maleness, Metaphor, and the 'Crisis' of Reason." In *A Mind of One's Own,* eds., Louise Antony and Charlotte Witt. Boulder: Westview Press.
- Lugones, María. 1992. "On 'Borderlands/La Frontera': An Interpretive Essay." *Hypatia* 7:4 (Fall): 31-37.
- Lugones, María and Elizabeth Spelman. 1983. "Have We Got a Theory for You!" *Hypatia* (WSIF) 1: 573-581.
- MacKinnon, Catherine. 1982. "Feminism, Marxism, Method and State: An Agenda for Theory." *Signs* 7(3): 514-544.
- McDonagh, Eileen L. 1996. *Breaking the Abortion Deadlock.* New York: Oxford University Press.
- Meyers, Diana T. 1989. *Self, Society, and Personal Choice.* New York: Columbia University Press.
- Meyers, Diana Tietjens. 1994. *Subjection and Subjectivity.* New York: Routledge.
- Meyers, Diana Tietjens, ed. 1997. *Feminist Social Thought: A Reader*. New York: Routledge.
- Meyers, Diana Tietjens. 2000. "Intersectional Identity and the Authentic Self? Opposites Attract!" In *Relational Autonomy, eds.,* Catriona Mackenzie and Natalie Stoljar. New York: Oxford University Press.
- Meyers, Diana Tietjens. Forthcoming 2000. "Marginalized Identities -- Individuality, Agency, and Theory." In *Smudges in the Margins,* eds., Patricia Smith, et. al. Lawrence KA: Kansas University Press.
- Moody-Adams, Michelle. 1991. "Gender and the Complexity of Moral Voices." In *Feminist Ethics,* ed., Claudia Card. Kansas City: University of Kansas Press.
- Nedelsky, Jennifer. 1989. "Reconceiving Autonomy: Sources, Thoughts, and Possibilities." *Yale Journal of Law and Feminism* 1:1 (Spring): 7-16.
- Piper, Adrian M.S. 1990. "Higher-Order Discrimination." In *Identity, Character and Morality,* eds., Owen Flanagan and Amelie Okensberg Rorty. Cambridge: MIT Press.
- Ruddick, Sara. 1989. *Maternal Thinking.* Boston: Beacon Press.
- Ruddick, Sara. 1994. "Thinking of Mothers/Conceiving Birth." In *Representations of Motherhood,* eds., Donna Bassin, et. al. New Haven: Yale University Press.
- Scheman, Naomi. 1993. *Engenderings.* New York: Routledge.
- Valian, Virginia. 1998. *Why So Slow? The Advancement of Women.* Cambridge: MIT Press.
- Walker, Margaret Urban. 1999. "Gettting Out of Line: Alternatives to Life as a Career." In *Mother Time,* ed., Margaret Urban Walker. Lanham, MD: Rowman and Littlefield.
- Walker, Margaret Urban. 1998. *Moral Understandings.* New York: Routledge.
- Weir, Allison. 1995. "Toward a Model of Self-Identity: Habermas and Kristeva." In *Feminists Read Habermas,* ed., Johanna Meehan. New York: Routledge.
- Williams, Patricia J. 1991. *The Alchemy of Race and Rights.* Cambridge: Harvard University Press.
- Wollstonecraft, Mary. (1792) 1982. *A Vindication of the Rights of Woman,* 2nd Edition. London: 1792. Recently, ed., Miriam Brody Dramnick. New York: Penguin, 1982.
- Young, Iris Marion. 1994. "Gender as Seriality: Thinking About Women as a Social Collective."

Signs 19: 3 (Spring): 713-738, Spring 1994.
- Young, Iris Marion. 1990. *Stretching Out.* Bloomington: Indiana University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

feminism, topics: feminist perspectives on class and work | feminism, topics: feminist perspectives on reproduction and the family

Diana Meyers
dmeyers@uconnvm.uconn.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 28, 1999
Content last modified: March 23, 2000

Stanford Encyclopedia of Philosophy
Supplement to Feminist Perspectives on the Self

Comprehensive Bibliography on Feminist Perspectives on the Self

Introduction

This bibliography serves as a supplement to Diana Tietjens Meyers' entry "Feminist Perspectives on the Self". I hope it will be useful to philosophers as well as a general audience. All the listings are written in English. Individual chapters in anthologies are cited, and all the mentioned anthologies are also listed separately. Although I have tried to be thorough, my apologies are offered in advance for misclassifying

or omitting relevant pieces.

This Bibliography is up to date with respect to publications prior to March 1997. Because the feminist literature on the self is voluminous, it is organized around three central philosophical disciplines: metaphysics, epistemology, and social philosophy.

- Metaphysical treatments of the self address the question ‘What is the self?’. Four sub-categories are included: agency, the body, gender & identities, and personhood.
- Epistemological treatments of the self ask ‘What and how do we gain knowledge of the self?’. Three sub-categories are included: emotional knowledge, self-knowledge, and knowledge of others & women's ways of knowing.
- Social treatments of the self have a dual aspect of ethics and politics. The key question addressed is ‘What are the ethics and/or politics of the self?’. Seven sub-categories are included: care ethics, critiques of individualism, ethical & political theory, families & friendship, self-respect, social groups, and violence & self.

Two additional interdisciplinary categories are included: literary & cultural studies and psychology & psychoanalysis.

Anthologies

Alcoff, Linda, and Potter, Elizabeth (Eds). Feminist Epistemologies. New York: Routledge, 1993.

Ames, Roger T. (Ed). Self and Deception. Albany: State University of New York Press, 1996.

Andolsen, Barbara Hilkert, and Gudorf, Christine E., and Pellauer, Mary D. (Eds). Women's Consciousness, Women's Conscience. Minneapolis: Winston Press, 1985.

Antony, Louise M., and Witt, Charlotte (Eds). A Mind of One's Own. Boulder: Westview Press, 1993.

Bar On, Bar-Ami (Ed). Engendering Origins. Albany: State University of New York Press, 1994

Benhabib, Seyla, et. al. Feminist Contentions. New York: Routledge, 1995.

Bock, Gisela, and James, Susan (Eds). Beyond Equality and Difference. New York: Routledge, 1992.

Broch-Due, Vigdis, and Rudie, Ingrid, and Bleie, Tony (Eds). Carved Flesh/Cast Selves. Providence: Berg, 1992.

Bushnell, Dana (Ed). Nagging Questions. Lanham: Rowman and Littlefield, 1995.

Butler, Judith, and Scott, Joan W. (Eds). Feminists Theorize the Political. New York: Routledge, 1992.

Card, Claudia (Ed). Feminist Ethics. Lawrence: University of Kansas Press, 1991.

Cole, Eve Browning, and Coultrap-McQuin, Susan (Eds). Explorations in Feminist Ethics. Bloomington: Indiana University Press, 1992.

Cornell, Drucilla, and Benhabib, Seyla (Eds). Feminism as Critique. Minneapolis: University of Minnesota Press, 1987.

Crosby, Donald A. (Ed). Religious Experience and Ecological Responsibility. New York: Lang, 1996.

Dallery, Arleen B., and Schott, Charles E., and Roberts, P. Holley (Eds). Ethics and Danger. Albany: State University of New York Press, 1992.

Dallery, Arleen B. (Ed). Continental Philosophy. Albany: State University of New York Press, 1990.

Dickens, David R. (Ed). Postmodernism and Social Inquiry. New York: Guilford, 1994.

Dillon, Robin S. (Ed). Dignity, Character, and Self-Respect. New York: Routledge, 1995.

Eisenstein, Hester, and Jardine, Alice (Eds). The Future of Difference. New York: Barnard College Women's Center, 1980.

Feldstein, Richard, and Roofs, Judith (Eds). Feminism and Psychoanalysis. New York: Routledge, 1989.

Flanagan, Owen, and Rorty, Amelie Okensberg (Eds). Identity, Character, and Morality. Cambridge: MIT Press, 1990.

Fraser, Nancy and Bartky, Sandra Lee (Eds). Revaluing French Feminism. Bloomington: Indiana University Press, 1992.

Garry, Ann, and Pearsall, Marilyn. Women, Knowledge, and Reality. Boston: Unwin Hyman, 1989 (1st Edition); New York: Routledge, 1996 (2nd Edition).

Gergen, Mary M., and Davis, Sara N. (Eds). Towards a New Psychology of Gender. New York: Routledge, 1997.

Goldberger, Nancy Rule, et. al., (Eds). Knowledge, Difference, and Power. New York: Basic Books, 1996.

Gould, Carol C. (Ed). Beyond Domination. Totowa: Rowman and Allenheld, 1984.

Gould, Carol C., and Wartofsky, Marx W. (Eds). Women and Philosophy. New York: Putnam, 1976.

Griffiths, A. Phillips (Ed). Ethics (Royal Institute of Philosophy Supplement 35). New York: Cambridge University Press, 1993.

Griffiths, Morwenna, and Whitford, Margaret (Eds). Feminist Perspectives in Philosophy. Bloomington: Indiana University Press, 1988.

Hanen, Marsha, and Nielsen, Kai (Eds). Science, Morality, and Feminist Theory (Supplement to Canadian Journal of Philosophy 13). Calgary: University of Calgary Press, 1987.

Harding, Sandra, and Hintikka, Merrill B. (Eds). Discovering Reality. Dordrecht: Reidel, 1983.

Held, Virginia (Ed). Justice and Care. Boulder: Westview Press, 1995.

Hirsch, Marianne, and Keller, Evelyn Fox. Conflicts in Feminism (Eds). New York: Routledge, 1990.

Jaggar, Alison M., and Bordo, Susan R. (Eds). Gender/Body/Knowledge. New Brunswick: Rutgers University Press, 1989.

James, Stanlie M., and Busia, Abena P.A. (Eds). Theorizing Black Feminims. New York: Routledge, 1993.

Katz, Michael J. (Ed). Philosophy and Education 1994. Urbana: Philosophy Education Society, 1994.

Kittay, Eva Feder, and Meyers, Diana T. (Eds). Women and Moral Theory. Totowa: Rowman and Littlefield, 1987.

Lauretis, Teresa de (Ed). Feminist Studies/Critical Studies. Bloomington: Indiana University Press, 1986.

MacAlister, Linda Lopez (Ed). Hypatia's Daughters. Bloomington: Indiana University Press, 1996.

Marks, Elaine, and Courtivron de, Isabelle (Eds). New French Feminisms. Amherst, University of Massachusetts Press, 1980.

Meehan, Johanna (Ed). Feminists Read Habermas. New York: Routledge, 1995.

Meyers, Diana Tietjens (Ed). Feminist Social Thought. New York: Routledge, 1997.

Meyers, Diana Tietjens (Ed). Feminists Rethink the Self. Boulder: Westview Press, 1997.

Nelson, Hilde Lindemann (Ed). Feminism and Families. New York: Routledge, 1997.

Nicholson, Linda J. (Ed). Feminism/Postmodernism. New York: Routledge, 1990.

Ogilvy, James (Ed). Revisioning Philosophy. Albany: State University of New York Press, 1992.

Orstein, P. (Ed). The Search for Self. New York: International Universities Press, 1978.

Pearsall, Marilyn (Ed). Women and Values. Belmont: Wadsworth Publishing, 1986.

Ramazanoglu, Caroline (Ed). Up Against Foucault. New York: Routledge, 1993.

Silverman, Hugh J. (Ed). Writing the Politics of Difference. Albany: State University of New York Press, 1991.

Smeyers, Paul (Ed). Identity, Culture, and Education. Leuven: Leuven University Press, 1994.

Sterba, James, and Peden, Creighton (Eds). Freedom, Equality, and Social Change. Lewiston: Mellen, 1990.

Stuhr, John (Ed). Philosophy and the Reconstruction of Culture. Albany: State of New York University Press, 1991.

Trebilcot, Joyce (Ed). Mothering. Totowa: Rowman and Allanheld, 1984.

Tuana, Nancy, and Tong, Rosemarie (Eds). Feminism and Philosophy. Boulder: Westview Press, 1995

Warren, Karen J. (Ed). Ecofeminism. Bloomington: Indiana University Press, 1997.

Warren, Karen J. (Ed). Ecological Feminist Philosophies. Bloomington: Indiana University Press, 1996.

Wasserstrom, Richard (Ed). Today's Moral Problems. New York: Macmillan, 1975.

Zack, Naomi (Ed). Race/Sex. New York: Routledge, 1997

The Metaphysics of Self

Agency

Babbitt, Susan E., "Feminism and Objective Interests? The Role of Transformation Experiences in Rational Deliberation" in Feminist Epistemologies, Alcoff, Linda, and Potter, Elizabeth (Eds). New York: Routledge, 1993.

Benhabib, Seyla, "Feminism and Postmodernism" in Feminist Contentions, Benhabib, Seyla, et. al. New York: Routledge, 1995.

Benhabib, Seyla, "Subjectivity, Historiography, and Politics" in Feminist Contentions, Benhabib, Seyla, et. al. New York: Routledge, 1995.

Boisvert, Raymond D., "Heteronomous Freedom" in Philosophy and the Reconstruction of Culture, Stuhr, John (Ed). Albany: State University of New York Press, 1993.

Butler, Judith, "Contingent Foundations", in Feminist Contentions, Benhabib, Seyla, et. al. New York: Routledge, 1995.

Butler, Judith, "For a Careful Reading", in Feminist Contentions, Benhabib, Seyla, et. al. New York: Routledge, 1995.

Butler, Judith. Gender Trouble. New York: Routledge, 1989.

Caust, Lesley, "Community, Autonomy and Justice: The Gender Politics of Identity and Relationship", History of European Ideas, 17(5), 639-650, Spring 1993

Christman, John, "Feminism and Autonomy" in Nagging Questions, Bushnell Dana (ed). Lanham:Rowman and Littlefield, 1995.

Christman, John, "Autonomy: A Defense of the Split-Level Self", Journal of Social Philosophy, 25, 281-293, Fall 1987.

Cornell, Drucilla, "What is Ethical Feminism?", in Feminist Contentions, Benhabib, Seyla, et. al. New York: Routledge, 1995

Cornell, Drucilla, "Rethinking the Time of Feminism", in Feminist Contentions, Benhabib, Seyla, et. al. New York: Routledge, 1995.

Ferguson, Ann, "Moral Responsibility and Social Change: A New Theory of Self", Hypatia, 12(3), 116-141, Summer 1997.

Flax, Jane. Disputed Subjects. New York: Routledge, 1993.

Fraser, Nancy, "False Antitheses", in Feminist Contentions, Benhabib, Seyla, et. al. New York: Routledge, 1995.

Fraser, Nancy, "Pragmatism, Feminism, and the Linguistic Turn", in Feminist Contentions, Benhabib, Seyla, et. al. New York: Routledge, 1995.

Friedman, Marilyn, "Autonomy and Social Relationships: Rethinking the Feminist Critique" in Feminists Rethink the Self, Meyers, Diana Tietjens (Ed). Boulder: Westview Press, 1997.

Friedman, Marilyn A., "Autonomy and the Split-Level Self", Journal of Social Philosophy, 24, 19-35, Spring 1986. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Friedman, Marilyn, "Women's Autonomy and Feminist Aspirations", Journal of Philosophical Research, 21, 331-340, January 1996.

Friedman, Marilyn A., "Self Rule in Social Context", in Freedom, Equality, and Social Change, Sterba, James, and Peden, Creighton (Eds). Lewiston: Mellen Press, 1990.

Govier, Trudy, "Self-Trust, Autonomy, and Self-Esteem", Hypatia, 8(1), 99-120, Winter 1993.

Gowens, Pat, "Sexual Empowerment: To Empower Women and Topple Male Supremacy", Off Our Backs, 23(9), 10-11, October 1993.

Griffiths, Morwenna, "Autonomy and the Fear of Dependence", Women's Studies International Forum, 15, 351-362, May-June 1992.

Grimshaw, Jean, "Practices of Freedom", in Up Against Foucault, Ramazanoglu, Caroline (Ed). New York: Routledge, 1993.

Grimshaw, Jean, "Autonomy and Identity in Feminist Thinking" in Feminist Perspectives in Philosophy, Griffiths, Morwenna, and Whitford, Margaret (Eds). Bloomington: Indiana University Press, 1988.

Hekman, Susan J., "Reconstituting the Subject: Feminism, Modernism, and Postmodernism", Hypatia, 6(2), Spring 1991.

Held, Virginia, "Birth and Death", Ethics, 99, 362-388, January 1989.

Herman, Barbara, "Agency, Attachment, and Difference", Ethics, 101(4), 775-797, July 1991.

Hill, Sharon Bishop, "Self-Determination and Autonomy" in Today's Moral Problems, Wasserstrom,

Richard (Ed). New York: Macmillan, 1975.

Hill, Thomas E. Jr., "The Importance of Autonomy", in Women and Moral Theory, Kittay, Eve Feder and Meyers, Diana T. (Eds). Totowa: Rowman & Littlefield, 1987.

Hoagland, Sarah, "Lesbian Ethics and Female Agency" in Explorations in Feminist Ethics, Cole, Eve Browning, and Coultrap-McQuin, Susan (Eds). Bloomington: Indiana University Press, 1992.

hooks, bell, "Out of the Academy and Into the Streets", Ms., 3(1), 80-87, July-August 1992.

Huntington, Patricia, "Toward a Dialectical Conception of Autonomy", Philosophy and Social Criticism, 21(1), 37-55.

Kasprisin, Lorraine, "Ideas of Self and Community: Ethical Implications for a Communitarian Concept of Moral Autonomy", Studies in Philosophy and Education, 15(1-2), 41-49, January-April 1996.

Keller, Jean, "Autonomy, Relationality, and Feminist Ethics", Hypatia, 12(2), 152-164, Spring 1997.

Mann, Patricia S. Micro-Politics. Minneapolis: University of Minnesota Press, 1994.

Meehan, Johanna, "Autonomy, Recognition, and Respect: Habermas, Benjamin, and Honneth", in Feminists Read Habermas, Meehan, Johanna (Ed). New York: Routledge, 1995.

Meyers, Diana T., "Personal Autonomy or the Deconstructed Subject? A Reply to Hekman", Hypatia, 7(1), 124-132, Winter 1992.

Meyers, Diana T. Self, Society, and Personal Choice. New York: Columbia University Press, 1991.

Meyers, Diana T., "Personal Autonomy and the Paradox of Feminine Socialization", Journal of Philosophy, 84, 619-629, November 1987.

Meyers, Diana T., "The Socialized Individual and Individual Autonomy" in Women and Moral Theory, Kittay, Eva Feder and Meyers, Diana T. (Eds). Totowa: Rowman and Littlefield, 1987.

Nedelsky, Jennifer, "Reconceiving Autonomy: Sources, Thoughts, and Possibilities", Yale Journal of Law and Feminism, 1(1), 7-16, Spring 1986.

The Body

Bordo, Susan. Unbearable Weight. Berkeley: University of California Press, 1993.

Butler, Judith. Bodies That Matter. New York: Routledge, 1993.

Dallery, Arleen B., "Sexual Embodiment: Beauvoir and French Feminism (Ecriture Feminine)", Women's Studies International Forum, 8(3), 197-202, 1985.

Davis, Kathy. Reshaping the Female Body. New York: Routledge, 1995.

Enns, Diane, "'We Flesh' Re-Membering the Body Beloved", Philosophy Today, 39(3-4), 263-279, Fall 1995.

Fallding, Helen, "Our Bodies, Our Selves", The Optimist, 14(3), 17, September 1988.

Ferguson, Ann, "Motherhood and Sexuality: Some Feminist Questions", Hypatia, 1,33-22, Fall 1986.

Hengehold, Laura, "An Immodest Proposal: Foucault, Hysterization, and the 'Second Rape'", Hypatia, 9(3), 88-107, Summer 1994.

Morgan, Kathryn Pauly, "Women and the Knife: Cosmetic Surgery and the Colonization of Women's Bodies", Hypatia, 6(3), 25-53, Fall 1991.

Probyn, Elspeth. Sexing the Self. New York: Routledge, 1993.

Probyn, Elspeth, "This Body Which Is Not One: Speaking an Embodied Self", Hypatia, 6(3), 111-124, Fall 1991.

Spelman, Elizabeth V., "Women as Body: Ancient and Contemporary Views", Feminist Studies, 8(1), 25-53, Fall 1991.

Stearns, Deborah C., "Gendered Sexuality: The Privileging of Sex and Gender in Sexual Orientation", National Women's Studies Association Journal, 7(1), 8-29, Spring 1995.

Wittig, Monique. The Straight Mind. Boston: Beacon Press, 1992.

Wittig, Monique. The Lesbian Body, David Le Vay (Trans). Boston: Beacon Press, 1986.

Young, Iris Marion. Stretching Out. Bloomington: Indiana University Press, 1990.

Young, Iris Marion. Throwing Like a Girl and Other Essays in Feminist Philosophy and Social Theory. Bloomington: Indiana University Press, 1990.

Young, Iris Marion, "Pregnant Embodiment: Subjectivity and Alienation", Journal of Medical Ethics, 9,

45-62, Fall 1984.

Young, Iris Marion, "Throwing Like A Girl: A Phenomenology of Feminine Body Comportment Motility and Spatiality", Human Studies, 3, 137-156, April 1980. ALSO IN: Feminism and Philosophy, Tuana, Nancy, and Tong, Rosemarie (Eds). Boulder: Westview Press, 1995.

Gender and Identities

Abel, Elizabeth, "Race, Class, and Psychoanalysis? Opening Questions", in Conflicts in Feminism, Hirsch, Marianne, and Keller, Evelyn Fox (Eds). New York: Routledge, 1990. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Allen, Anita L., "Forgetting Yourself" in Feminists Rethink the Self, Meyers, Diana Tietjens (Ed). Boulder: Westview Press, 1997.

Bar On, Bat-Ami, "Reading Bartky: Identity, Identification, and Critical Self-Reflection", Hypatia, 8(1), 159-163, Winter, 1993.

Brennan, Teresa, "Essence Against Identity", Metaphilosophy, 27(1-2), 92-103, January 1996.

Busia, Abena P.A., "Performance, Translation, and the Language of the Self: Interrogating Identity as a ‘Post-Colonial’ Poet" in Theorizing Black Feminisms, James, Stanlie M., and Busia, Abena P.A. (Eds). New York: Routledge, 1993.

Calhoun, Cheshire, "Separating Lesbian Theory from Feminist Theory", Ethics, 104(3), 558-581, April 1994. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens. New York: Routledge, 1997.

Campell, Sue, "Women, ‘False’ Memory, and Personal Identity", Hypatia, 12(20), 51-82, Spring 1997.

Estenberg, Kristin G. Lesbian and Bisexual Identities. Philadelphia: Temple University Press, 1997.

Ferguson, Ann, "Can I Choose Who I Am? And How Would That Empower Me? Gender, Race, Identities, and the Self", in Women, Knowledge, and Reality, Garry, Ann, and Pearsall, Marilyn (Eds). New York: Routledge, 1996 (2nd Edition).

Ferguson, Ann, "Lesbian Identity: Beauvoir and History", Hypatia, (WSIF) 3, 203-208, 1985.

Friedman, Marilyn, "The Unholy Alliance of Sex and Gender", Metaphilosophy, 27(1-2), 78-91, January 1996.

Griffiths, Morwenna. Feminisms and the Self. New York: Routledge, 1995.

Hale, Jacob, "Are Lesbians Women?", Hypatia, 11(2), 94-121, Spring 1996.

Honig, B., "Toward an Agnostic Feminism: Hannah Arendt and the Politics of Identity" in Feminists Theorize the Political, Butler, Judith, and Scott, Joan W. (Eds). New York: Routledge, 1992.

hooks, bell. Talking Back. Boston: South End Press, 1989.

Houston, Barbara, "In Defense of a Politics of Identity" in Philosophy of Education 1994, Katz, Michael S. (Ed). Urbana: Philosophy Education Society, 1994.

Kristeva, Julia. Strangers to Ourselves, Leon Roudiez (Trans). New York: Columbia University Press, 1992.

Lee-Lampshire, Wendy, "Decisions of Identity: Feminist Subjects and Grammars of Sexuality", Hypatia, 10(4), 32-45, Fall 1995.

Lugones, María, "Structure/Antistructure and Agency Under Oppression" Journal of Philosophy, 500-507, October 1990.

McNamee, Michael, "Identity and the Self", Studies in Philosophy and Education, 15(1-2), 107-111, January-April 1996.

Scheman, Naomi, "Queering the Center by Centering the Queer: Reflections on Transsexuals and Secular Jews" in Feminists Rethink the Self, Meyers, Diana Tietjens (Ed). Boulder: Westview Press, 1997.

Smiley, Marion, "Feminist Theory and the Question of Identity", Women & Politics, 13(2), 91-122, 1993.

Spelman, Elizabeth. Inessential Woman. Boston: Beacon Press, 1988.

Vegetti Finzi, Silvia, "Female Identity between Sexuality and Maternity" in Beyond Equality and Difference, Bock, Gisela, and James, Susan (Eds). New York: Routledge, 1992.

Personhood

Antony, Louise M., "Is Psychological Individualism a Piece of Ideology?", Hypatia, 10(3), 157-74, Summer 1995.

Assister, Alison. Enlightened Woman. New York: Routledge, 1996.

Bartky, Sandra Lee. Femininity and Domination. New York: Routledge, 1990.

Beauvoir, Simone de. The Second Sex, H.M. Parschley (Trans). New York: Vintage Press, 1953.

Braidotti, Rosi. Nomadic Subjects. New York: Columbia University Press, 1994.

Braidotti, Rosi, "On the Female Feminist Subject, or: ‘From She-Self’ to ‘She-Other’" in Beyond Equality and Difference, Bock, Gisela, and James, Susan (Eds). New York: Routledge, 1992.

Butler, Judith. Excitable Speech. New York: Routledge, 1997.

Butler, Judith, "Gender Trouble, Feminist Theory, and Psychoanalytic Discourse" in Feminism/Postmodernism, Nicholson, Linda J. (Ed). New York: Routledge, 1990.

Card, Claudia. The Unnatural Lottery. Philadelphia: Temple University Press, 1996.

Code, Lorraine, "Second Persons" in Science, Morality, and Feminist Theory (Supplement to Canadian Journal of Philosophy 13). Hanen, Marsha, and Nielsen, Kai (Eds). Calgary: University of Calgary Press, 1987.

Davion, Victoria, "Competition, Recognition, and Approval Seeking", Hypatia, 3, 165-166, Summer 1988.

Davion, Victoria, "Do Good Feminists Compete?", Hypatia, 2, 55-63, Summer 1987.

Duchamp, Linda Timmel, "Desperately Seeking Approval: The Importance of Distinguishing Between Approval and Recognition", Hypatia, 3, 163-164, Summer 1988.

Ferguson, Ann, "A Feminist Aspect Theory of Self" in Science, Morality and Feminist Theory (Supplement to Canadian Journal of Philosophy 13). Hanen, Marsha, and Nielsen, Kai (Eds). Calgary: University of Calgary Press, 1987.

Ferguson, Kathy. The Man Question. Berkeley: University of California Press, 1993.

Frye, Marilyn. The Politics of Reality. Trumansburg: Crossing Press, 1983.

Jaggar, Alison M. Feminist Politics and Human Nature. Totowa: Rowman and Allanheld, 1983.

Kittay, Eva Feder, "Woman as Metaphor’, Hypatia, 3, 63-86, Summer 1988. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Kruks, Sonia, "Gender and Subjectivity: Simone de Beauvoir and Contemporary Feminism", Signs, 18(1), 89-110, Autumn 1992.

Kukla, Rebecca, "Decentering Women", Metaphilosophy, 27(1-2), 28-52, January 1996.

Kotzin, Rhoda Hadassa, "Bribery and Intimidation: A Discussion of Sandra Lee Bartky's ‘Femininity and Domination’", Hypatia, 8(1), 164-172, Winter 1993.

Lee-Lampshire, Wendy, "Moral ‘I’: The Feminist Subject and the Grammar of Self-Reference", Hypatia, 7(1) 34-51, Winter 1992.

Radden, Jennifer, "Shame and Blame: The Self through Time and Change", Hypatia 11(3), 71-96, Summer 1996.

Scheman, Naomi, "Individualism and the Objects of Psychology" in Discovering Reality, Harding, Sandra, and Hintikka, Merrill B. (Eds). Dordrecht: Reidel, 1983.

Weir, Allison. Sacrificial Logics. New York: Routledge, 1996.

Weir, Allison, "Toward a Model of Self-Identity: Habermas and Kristeva", in Feminists Read Habermas, Meehan, Johanna (Ed). New York: Routledge, 1995.

Whitbeck, Caroline, "A Different Reality: Feminist Ontology" in Beyond Domination, Gould, Carol C., (Ed). Totowa: Rowman and Allenheld, 1984.

The Epistemologies of Self

Emotional Knowledge

Bartky, Sandra Lee, "Sympathy and Solidarity: On a Tightrope with Scheler" in Feminists Rethink the Self, Meyers, Diana Tietjens (Ed). Boulder: Westview Press, 1997.

Calhoun, Cheshire, "Changing One's Heart", Ethics, 103(1), 76-96, October 1992.

Calhoun, Cheshire, "Subjectivity and Emotion", Philosophical Forum, 20, 195-210, Spring 1989.

Irigaray, Luce, "Sorcer Love: A Reading of Plato's Symposium, Diotima's Speech", Eleanor Kuykendal (Trans.), Hypatia, 3(3) Winter 28-44 1989.

Jaggar, Alison M. "Love and Knowledge: Emotion in Feminist Epistemology" in Women, Knowledge, and Reality, Garry, Ann, and Pearsall, Marilyn (Eds). New York: Unwin Hyman, 1989 (1st Edition). ALSO IN: Gender/Body/Knowledge, Jaggar, Alison M., and Bordo, Susan R. (Eds). New Brunswick: Rutgers University Press, 1989.

Meyers, Diana Tietjens, "Emotion and Heterodox Moral Perception: An Essay in Moral Social Psychology" in Feminists Rethink the Self, Meyers, Diana Tietjens (Ed). Boulder: Westview Press, 1997.

Morgan, Kathryn Pauly, "Romantic Love, Altruism, and Self-Respect", Hypatia, 1, 117-148, Spring 1986.

Morrow, Frances. Unleashing Our Unknown Selves. New York: Praeger, 1990.

Nussbaum, Martha C. Love's Knowledge. New York: Oxford University Press, 1990.

Rappaport, Elizabeth, "On the Future of Love: Rousseau and the Radical Feminists", Philosophical Forum (Boston), 5, 186-205, Winter 1973.

Scheman, Naomi, "Property, Authority, and the Emotions", Resources for Feminist Research, 8(1), March, 27-28, 1979.

Spelman, Elizabeth V., "Anger and Insubordination" in Women, Knowledge, and Reality, Garry, Ann, and Pearsall, Marilyn (Eds). New York: Unwin Hyman, 1989 (1st Edition).

Self-Knowledge

Baier, Annette C., "The Vital but Dangerous Art of Ignoring: Selective Attention and Self-deception" in Self and Deception, Ames, Roger T. (Ed). Albany: State University of New York Press, 1996.

Bernstein, Susan David, "Confessing Feminist Theory -- What's ‘I’ Got To Do With It?", Hypatia, 7(2), 120-147, Spring 1992.

Grimshaw, Jean, "Ethics, Fantasy and Self-Transformation" in Ethics (Royal Institute of Philosophy Supplement 35), Griffiths, A. Phillips (Ed). New York: Cambridge University Press, 1993.

Hardwig, John, "Privacy, Self-knowledge, and Pluralistic Communes: An Invitation to the Epistemology of the Family", in Feminism and Families, Nelson, Hilde Lindemann (Ed). New York: Routledge, 1997.

Mullet, Sheila, "Only Connect: The Place of Self-Knowledge in Ethics" in Science, Morality, and Feminist Theory (Supplement to Canadian Journal of Philosophy 13), Hanen, Marsha, and Nielsen, Kai (Eds). Calgary: University of Calgary Press, 1987.

Park, Shelly M., "False Memory Syndrome: A Feminist Philosophical Approach", Hypatia, 12(2), 1-50, Spring 1997.

Radden, Jennifer, "Defining Self-Deception", Dialogue 23, 103-120, March 1984.

Tomm, Winnie, "Ethics and Self-Knowing: The Satisfaction of Desire" in Explorations in Feminist Ethics, Cole, Eve Browning, and Coultrap-McQuin, Susan (Eds). Bloomington: Indiana University Press, 1992.

Women's Knowledge and Knowledge of Others

Addelson, Kathryn Pyne. Impure Thoughts. Philadelphia: Temple University Press, 1991.

Baier, Annette. Postures of the Mind. Minneapolis: University of Minnesota Press, 1985.

Belenky, Mary Field, et. al. Women's Ways of Knowing. New York: Basic Books, 1986.

Code, Lorraine, "Taking Subjectivity into Account" in Feminist Epistemologies, Alcoff, Linda, and Potter, Elizabeth (Eds). New York: Routledge, 1993.

Crawford, Mary, "Agreeing to Differ: Feminist Epistemologies and Women's Ways of Knowing" in Towards a New Psychology of Gender, Gergen, Mary M., and Davis, Sara N. (Eds). New York: Routledge, 1997.

Gaten-Robinson, Eugenie, "Finding Out Feminist Ways in Natural Philosophy and Religious Thought", Hypatia, 9(4), 207-228, Winter 1994.

Gilligan, Carol, "Hearing the Difference: Theorizing Connection", Hypatia, 10(2) 120-127, Spring 1995.

Gilligan, Carol, "In a Different Voice: Women's Conceptions of Self and of Morality", Harvard Educational Review, 47(4), 481-517, 1977. ALSO IN: The Future of Difference, Eisenstein, Hester, and Jardine, Alice (Eds). New York: Barnard College Women's Center, 1980.

Grimshaw, Jean. Philosophy and Feminist Thinking. Minneapolis: University of Minnesota Press, 1986.

Irigaray, Luce. Speculum of the Other Woman, Gillian C. Gill (Trans). Ithaca: Cornell University Press, 1985.

Lugones, María, "Playfulness, ‘World’-Traveling, and Loving Perception", Hypatia, 2(2), 3-19, Fall 1990. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Scheman, Naomi. Engenderings. New York: Routledge, 1993.

Scott, Joan, "Gender as a Useful Category for Historical Analysis", American Historical Review, 91, 1986.

Taylor, Gabrielle. Pride, Shame, and Guilt. Oxford: Clarendon Press, 1985.

The Social Philosophies of Self

Care Ethics

Baier, Annette C., "Whom Can Women Trust?" in Feminist Ethics, Card, Claudia (Ed). Lawrence: University of Kansas Press, 1991.

Bartlett, Elizabeth Ann, "Beyond Either/Or: Justice and Care in the Ethics of Albert Camus" in Explorations in Feminist Ethics, Cole, Eve Browning, and Coultrap-McQuin, Susan (Eds). Bloomington: Indiana University Press, 1992.

Benhabib, Seyla, "The Generalized and the Concrete Other: The Kohlberg-Gilligan Controversy and Feminist Theory", Praxis International, 5(4), 402-424, January 1986. ALSO IN: Women and Moral Theory, Kittay, Eva Feder, and Meyers, Diana T. (Eds). Totowa: Rowman and Littlefield, 1987.

Brabeck, Mary M. Who Cares. Westport: Greenwood Press, 1989.

Dillon, Robin S., "Care and Respect", in Explorations in Feminist Ethics, Cole, Eve Browning, and Coultrap-McQuin, Susan (Eds). Bloomington: Indiana University Press, 1992.

Friedman, Marilyn, "Beyond Caring: The De-Moralization of Gender" in Science, Morality, and Feminist Theory (Supplement to Canadian Journal of Philosophy 13). Calgary: University of Calgary Press, 1987.

Gilligan, Carol, "Moral Orientation and Moral Development" in Women and Moral Theory, Kittay, Eve Feder, and Meyers, Diana T. (Eds). Totowa: Rowman and Littlefield, 1987.

Gilligan, Carol. In a Different Voice. Cambridge: Harvard University Press, 1982.

Golden, Jill, "The Care of the Self: Poststructuralist Questions About Moral Education and Gender", Journal of Moral Education, 25(4), 381-393, December 1996.

Haaken, Janice, "From Al-Anon to ACOA: Codependence and the Reconstruction of Caregiving", Signs, 18(2), 321-345, Winter 1993.

Held, Virginia, "Feminism and Moral Theory" in Women and Moral Theory, Kittay, Eva Feder, and Meyers, Diana T. (Eds). Totowa: Rowman and Littlefield, 1987. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Hoagland, Sarah Lucia, "Some Thoughts About Caring" in Feminist Ethics, Card, Claudia (Ed).

Lawrence: University Press of Kansas, 1991.

Nelson, Hilde Lindemann, "Against Caring", Journal of Clinical Ethics, 3(1), 8-15, Spring 1992.

Noddings, Nel, "In Defense of Caring", Journal of Clinical Ethics, 3(1), 15-18, Spring 1992.

Noddings, Nel. Caring. Berkeley: University of California Press, 1984.

Ruddick, Sara, "From Maternal Thinking to Peace Politics" in Explorations in Feminist Ethics, Cole, Eve Browning, and Coultrap-McQuin, Susan (Eds). Bloomington: Indiana University Press, 1992.

Ruddick, Sara. Maternal Thinking. Boston: Beacon Press, 1989.

Ruddick, Sara, "Maternal Thinking", Feminist Studies, 6(2), 342-67, Summer 1980. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Critiques of Individualism

Baier, Annette, "Trust and Anti-Trust", Ethics, 96, 231-260, January 1986. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Cornell, Drucilla. The Philosophy of the Limit. New York: Routledge, 1992.

Ferguson, Ann. Sexual Democracy. Boulder: Westview Press, 1991.

Friedman, Marilyn, "The Social Self and the Partiality Debates" in Feminist Ethics, Card, Claudia, (Ed). Lawrence: University Press of Kansas, 1991.

Ginzberg, Ruth, "Philosophy Is Not a Luxury" in Feminist Ethics, Card, Claudia (Ed). Lawrence: University Press of Kansas, 1991.

Held, Virginia, "Non-Contractual Society: A Feminist View" in Science, Morality, and Feminist Theory (Supplement to Canadian Journal of Philosophy 13), Hanen, Marsha, and Nielsen, Kai (Eds). Calgary: University of Calgary Press, 1987.

Held, Virginia, "On Rawls and Self-Interest", Midwest Studies in Philosophy, 1, 57-59, 1976.

hooks, bell. Feminist Theory From Margin to Center. Boston: South End Press, 1984.

hooks, bell. Ain't I a Woman?. Boston: South End Press, 1981.

Irigaray, Luce. Je, Tu, Nous, Alison Martin (Trans). New York: Routledge, 1993.

Irigaray, Luce. This Sex Which is Not One, Catherine Porter with Carolyn Burke (Trans). Ithaca: Cornell University Press, 1985.

Irigaray, Luce, "Any Theory of the ‘Subject’ has Always Been Appropirated by the Masculine", Gillian C. Gill (Trans.), Trivia, 6, 38-51, Winter 1985.

Keller, Catherine. From a Broken Web. Boston: Beacon Press, 1986.

Kittay, Eva Feder, "Human Dependency and Rawlsian Equality" in Feminists Rethink the Self, Meyers, Diana Tietjens. Boulder: Westview Press,1997.

Mahowald, Mary B., "Feminism: Individualistic or Communalistic?", Proceedings of Catholic Philosophy Association, 50, 219-228, 1976.

Meyers, Diana Tietjens, "Moral Reflection: Beyond Impartial Reason", Hypatia, 8(3), 21-47, Summer 1993.

May, Larry. The Socially Responsive Self. Chicago: University of Chicago Press, 1996.

Plumwood, Val, "Nature, Self, and Gender: Feminism, Environmental Philosophy, and the Critique of Rationalism", Hypatia, 3-27, Spring 1991. ALSO IN: Ecological Feminist Philosophies, Warren, Karen J. (Ed). Bloomington: Indiana University Press, 1996.

Radden, Jennifer, "Relational Individualism and Feminist Theory", Hypatia, 11(3), 71-96, Summer 1996.

Schmitt, Richard. Beyond Separateness. Boulder: Westview Press, 1995.

Schwartzenbach, Sybil, "Rawls and Ownership: The Forgotten Category of Reproductive Labor", in Science, Morality, and Feminist Theory (Supplement to Canadian Journal of Philosophy 13), Hanen, Marsha, and Nielsen, Kai (Eds). Calgary: University of Calgary Press, 1987.

Spelman, Elizabeth, "Good Grief! It's Plato!" in Feminists Rethink the Self, Meyers, Diana Tietjens (Ed). Boulder: Westview Press, 1997.

Wager, Joseph, "Incommensurable Differences: Cultural Relativism and Antirationalism Concerning Self and Other", Philosophy of the Continental World, 3(2), 18-26, Summer 1996.

Wittig, Monique, "On the Social Contract", Feminist Issues, 9(1), 3-12, Spring 1989.

Young, Iris Marion, "The Ideal of Community and the Politics of Difference" in Feminism/Postmodernism, Nicholson, Linda J. (Ed). New York: Routledge, 1990.

Ethical and Political Theory

Addelson, Kathryn Pyne. Moral Passages. New York: Routledge, 1994.

Allen, Anita, "Privacy and Reproductive Liberty" in Nagging Questions, Bushnell, Dana (Ed). Lanham: Rowman and Littlefield, 1995.

Anderson, Olive, "The Feminism of T.H. Green: A Late-Victorian Success Story?", History of Political Thought, 12(4), 671-693, Winter 1991.

Blum, Larry, et. al., "Altruism and Women's Oppression", Philosophical Forum (Boston), 5, 222-247, Fall-Winter 1973.

Benhabib, Seyla. Situating the Self. New York: Routledge, 1992.

Burks, Valerie C., "Women's Place: An Arendtian Critique of Feminism", Women & Politics, 14(3), 19-56, 1994.

Card, Claudia, "Gender and Moral Luck" in Identity, Character, and Morality, Flanagan, Owen, and Rorty, Amelie Okensberg (Eds). Cambridge: MIT Press, 1990. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Code, Lorraine, "Simple Equality Is Not Enough", Australian Journal of Philosophy (Supplement, 64, 48-65, June 1986.

Cornell, Drucilla. Transformations. New York: Routledge, 1993.

Cuomo, Chris, "Toward Thoughtful Ecofeminist Activism" in Ecological Feminist Philosophies, Warren, Karen J. (Ed).

Cutting-Gray, Joanne, "Hannah Arendt, Feminism, and the Politics of Alterity: ‘What Will We Lose If We Win?’", Hypatia, 8(1), 35-54, Winter 1993.

Davion, Victoria, "When Lives Become Logic Problems: Nuclear Deterrence, an Ecological Feminist Critique" in Ecological Feminist Philosophies, Warren, Karen J. (Ed). Bloomington: Indiana University Press, 1996.

Donner, Wendy, "John Stuart Mill's Liberal Feminism", Philosophical Studies, 69(2-3), 155-166, March

1993.

Duran, Jane, "The Intersection of Pragmatism and Feminism", Hypatia, 8(2), 159-171, Spring 1993.

Ferguson, Kathy. Self, Society, and Womankind. Westport: Greenwood Press, 1980.

Flax, Jane, "Beyond Equality: Gender, Justice and Difference" in Beyond Equality and Difference, Bock, Gisela, and James, Susan (Eds). New York: Routledge, 1992.

Fox, Ellen L., "Seeing Through Women's Eyes: The Role of Vision in Women's Moral Theory" in Explorations in Feminist Ethics, Cole, Eve Browning, and Coultrap-McQuin, Susan (Eds). Bloomington: Indiana University Press, 1992.

Fraser, Nancy, "Toward A Discourse Ethic of Solidarity", Praxis International, 5, 425-429, January 1988.

Hartsock, Nancy. Money, Sex, Power. New York: Longman, 1983.

Hekman, Susan J. Moral Voices, Moral Selves. University Park: Pennsylvania State University Press, 1995.

Hekman, Susan J., "Moral Voices, Moral Selves: About Getting it Right in Moral Theory", Human Studies, 16(1-2), 143-162, April 1993.

Held, Virginia. Feminist Morality. Chicago: University of Chicago Press, 1993.

Held, Virginia, "Feminist Transformations of Moral Theory", Philosophy and Phenomenological Research, 321-344, 1990.

Held, Virginia, "Marx, Sex, and the Transformation of Society", Philosophical Forum (Boston), 5, 168-184, Fall-Winter 1973.

Howe, Leslie A., "Kierkegaard and the Feminine Self", Hypatia, 9(4), 131-157, Fall 1994.

Jaggar, Alison M., "Feminist Ethics: Some Issues for the Nineties", Journal of Social Philosophy, 20, 91-107, Spring-Fall 1989.

Mackenzie, Catriona, "Reason and Sensibility: The Ideal of Women's Self-Governance in the Writings of M. Wollstonecraft", Hypatia, 8(4), 35-55, 1993. ALSO IN: Hypatia's Daughters, MacAlister, Linda Lopez (Ed). Bloomington: Indiana University Press, 1996.

Meyers, Diana Tietjens. Subjection and Subjectivity. New York: Routledge, 1994.

Millett, Kate. Sexual Politics. New York: Doubleday, 1970.

Minow, Martha. Making All the Difference. Ithaca: Cornell University Press, 1990.

Morgan, Kathryn, "Women and Moral Madness", in Science, Morality, and Feminist Theory (Supplement to Canadian Journal of Philosophy 13), Hanen, Marsha, and Nielsen, Kai (Eds). Calgary: University of Calgary Press, 1987.

Mouffe, Chantal, Feminism, "Citizenship, and Radical Democratic Politics" in Feminists Theorize the Political, Butler, Judith, and Scott, Joan W. (Eds). New York: Routledge, 1992. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Nelson, Hilde Lindemann, "Resistance and Insubordination", Hypatia, 10(2)23-40, Spring 1995.

Nelson, Julie A., "Thinking About Gender", Hypatia, 7(3), 138-154, Summer 1992.

Pangerl, Susan, "Biophilic Intuition and Intersubjectivity: The Relational Roots of the Moral Self" in Religious Experience and Ecological Responsibility, Crosby, Donald A. (Ed). New York: Lang, 1996.

Scott, Joan W., "Deconstructing Equality-Versus-Difference: Of the Uses of Poststructuralist Theory for Feminism", in Conflicts in Feminism, Hirsch, Marianne, and Keller, Evelyn Fox (Eds). New York: Routledge, 1990. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Sousa, Ronald B. de, and Morgan, Kathryn Pauly, "Philosophy, Sex, and Feminism", Atlantis, 13, 1-10, Spring 1988.

Spelman, Elizabeth, "Who's Who in the Polis?" in Engendering Origins, Bar on, Bat-Ami (Ed). Albany: State University of New York Press, 1994.

Spelman, Elizabeth, "On Treating Persons as Persons", Ethics, 88, 150-161, January 1978.

Straumanis, Joan, "Duties to Oneself: An Ethical Basis for Self-Liberation", Journal of Social Philosophy, 15, 1-13, Summer 1984.

Thomas, Lawrence, "The Reality of the Moral Self", Monist, 76(1), 3-21, January 1993.

Walker, Margaret Urban. Moral Understandings. New York: Routledge, 1998.

Walker, Margaret Urban, "Picking Up Pieces: Lives, Stories, and Integrity" in Feminists Rethink the Self, Meyers, Diana Tietjens (Ed). Boulder: Westview Press, 1997.

Walker, Margaret Urban, "Feminism, Ethics, and the Question of Theory", Hypatia, 7(3), 23-28, Summer 1992.

Walker, Margaret Urban, "Moral Particularity", Metaphilosophy, 18, 171-185, July-October 1987

Warren, Karen J., and Cheney, Jim, "Ecological Feminism and Ecosystem Ecology", in Ecological Feminist Philosophies, Warren, Karen J. (Ed). Bloomington: Indiana University Press, 1996.

Young, Iris Marion, "Self-Determination as a Principle of Justice", Philosophical Forum (Boston), 11, 30-46, Fall 1979.

Families and Friendship

Calhoun, Cheshire, "Emotional Work" in Explorations in Feminist Ethics, Cole, Eve Browning, and Coultrap-McQuin, Susan (Eds). Bloomington: Indiana University Press, 1992.

Collins, Patricia Hill, "The Meaning of Motherhood in Black Culture" in Towards a New Psychology of Gender, Gergen, Mary M., Davis, Sara N. (Eds). New York: Routledge 1997.

Friedman, Marilyn A. What are Friends For?. Ithaca: Cornell University Press, 1993.

Friedman, Marilyn A., "Feminism and Modern Friendship: Dislocating the Community", Ethics, 99, 275-290, January 1989. ALSO IN: Explorations in Feminist Ethics, Cole, Eve Browning and Coultrap-McQuin, Susan (Eds). Bloomington: Indiana University Press, 1992.

Irigaray, Luce, "And One Doesn't Stir Without the Other", Signs, 7, 60-67, 1981. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens. New York: Routledge, 1997.

Kaplan, Laura Duhan, "Women as Nurturer: An Archetype Which Supports Patriarchal Militarism", Hypatia, 9(2), 123-133, Spring 1994.

Herman, Barbara, "Could It Be Worth Thinking About Kant on Sex and Marriage?", in A Mind of One's Own, Antony, Louise M., and Witt, Charlotte (Eds). Boulder: Westview Press, 1993.

LaFollette, Hugh. Personal Relationships. Cambridge: Blackwell, 1996.

McMahon, Martha. Motherhood. New York: Guilford, 1995.

Meyers, Diana Tietjens, "The Family Romance: A Fin-de-Siècle Tragedy", in Feminism and Families, Nelson, Hilde Lindemann (Ed). New York: Routledge, 1997. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge, 1997.

Nelson, Hilde Lindemann, "Sophie Doesn't: Families and Counterstories of Self-Trust", Hypatia, 11(1), 91-104, Winter 1996.

Raymond, Janice. A Passion for Friends. Boston: Beacon Press, 1986.

Robinson, Elise L.E., Nelson, Hilde Lindemann, and Nelson, James, "Fluid Families: The Role of Children in Custody Arrangements" in Feminism and Families, Nelson, Hilde Lindemann (Ed). New York: Routledge, 1997.

Scheman, Naomi, "Who is that Masked Woman? Reflections on Power, Privilege, and Home-ophobia" in Revisioning Philosophy, Ogilvy, James (Ed). Albany: State University of New York Press, 1992.

Weinzweig, Marjorie, "Should a Feminist Choose A Marriage-Like Relationship?", Hypatia, 3(1), 139-160, 1988.

Young, Iris Marion, "Is Male Gender Identity the Cause of Male Domination?", in Mothering, Trebilcot, Joyce (Ed). Totowa: Rowman and Allanheld, 1984. ALSO IN: Feminist Social Thought, Meyers, Diana Tietjens (Ed). New York: Routledge.

Self-Respect

Calhoun, Cheshire, "Standing for Something", Journal of Philosophy, 9(2), 235-260, May 1995.

Dillon, Robin, "Toward a Feminist Conception of Self-Respect", Hypatia, 7(1), 352-69, Winter 1992. ALSO IN: Dignity, Character, and Self Respect, Dillon, Robin S. (Ed). New York: Routledge, 1995.

French, Marilyn, "Self-Respect: A Female Perspective", Humanist, 46, 18-23, November-December 1986.

Held, Virginia, "Reasonable Progress and Self-Respect", Monist, 57, 12-27, January 1973.

Hill, Thomas E. Jr., "Servility and Self-Respect", Monist, 57, 87-104, 1973.

Meyers, Diana Tietjens, "Self-Respect and Autonomy" in Dignity, Character, and Self-Respect, Dillon, Robin, S. (Ed). New York: Routledge, 1995.

Meyers, Diana T., "The Politics of Self-Respect: A Feminist Perspective", Hypatia, 1, 86-100, Spring 1986.

Postow, B.C., "Economic Dependence and Self-Respect", Philosophical Forum (Boston), 10, 181-205,

January 1979.

Social Groups

Alcoff, Linda, "Feminist Politics and Foucault" in Continental Philosophy, Dallery, Arleen B. (Ed). Albany: State University of New York Press, 1990.

Alcoff, Linda, "Cultural Feminism Versus Post-Structuralism: the Identity Crisis in Feminist Theory", Signs, 13, 405-436, Spring 1988. ALSO IN: Feminism and Philosophy, Tuana, Nancy, and Tong, Rosemarie (Eds). Boulder: Westview Press, 1995.

Benhabib, Seyla, "From Identity Politics to Social Feminism: A Plea for the Nineties", in Philosophy and Education 1994, Katz, Michael S. (Ed). Urbana: Philosophy Education Society, 1994.

Brown-Collins, Alice Ray, and Sussewell, Deborah Ridley, "The Afro-American Women's Emerging Selves", Journal of Black Psychology, 13 (August 1), 1-11, 1986.

Childers, Mary and hooks, bell, "A Conversation About Race and Class", in Conflicts in Feminism, Hirsch, Marianne, and Keller, Evelyn Fox (Eds). New York: Routledge, 1992.

Collins, Patricia Hill. Black Feminist Thought. Boston: Unwin Hyman, 1990.

Daumer, Elizabeth, "Queer Ethics, or the Challenge of Bisexuality to Lesbian Ethics", Hypatia, 7(4), 91-105, Fall 1992.

Frye, Marilyn, "A Response to Lesbian Ethics: Why Ethics?" in Feminist Ethics, Card, Claudia (Ed). Lawrence: University of Kansas Press, 1991.

Hoagland, Sarah Lucia, "Why Lesbian Ethics?", Hypatia, 7(4), 195-206, Fall 1992.

Hoagland, Sarah Lucia. Lesbian Ethics. Palo Alto: Institute for Lesbian Studies, 1988.

hooks, bell. Outlaw Culture. New York: Routledge, 1994.

King, Deborah K., "Multiple Jeopardy, Multiple Consciousness: The Context of a Black Feminist Ideology", Signs, 14(1), 42-72., 1988 ALSO IN: Feminist Social Thought, Meyers, Diana

Tietjens (Ed). New York: Routledge, 1997.

Kondo, Dorinne K. Crafting Selves. Chicago: University of Chicago Press, 1990.

Lâm, Maivan Clêch, "Feeling Foreign in Feminism", Signs, 19(4), 865-893, Summer 1994.

Loomba, Ania, "Tangled Histories: Indian Feminism and Anglo-American Feminist Criticism", Tulsa Studies in Women's Literature, 12(2), 271-278, Fall 1993.

Lorde, Audre. Sister Outsider. Trumansburg: Crossing Press, 1984.

Lugones, María, "Purity, Impurity, and Separation", Signs, 19(2), 458-479, Winter 1994.

Lugones, María, "On the Logic of Pluralist Feminism" in Feminist Ethics, Card, Claudia (Ed). Lawrence: University of Kansas Press, 1991.

Mullin, Amy, "Selves, Diverse and Divided: Can Feminists Have Diversity Without Multiplicity?", Hypatia, 10(4), 1-31, Fall 1995.

Spelman, Elizabeth and Lugones, María, "Have We Got a Theory for You!", Hypatia, (WSIF) 1, 573-581, 1983.

Yeatman, Anna, "A Feminist Theory of Social Differentiation" in Feminism/Postmodernism, Nicholson, Linda J. (Ed). New York: Routledge, 1990.

Young, Iris Marion, "Gender as Seriality: Thinking About Women as a Social Collective", Signs 19(3), 713-738, Spring 1994.

Violence and Self

Alcoff, Linda, "Survivor Discourse: Transgression or Recuperation?", Signs, 18(2), 260-290, Winter 1993.

Brison, Susan J., "Outliving Oneself: Trauma, Memory, and Personal Identity" in Feminists Rethink the Self, Meyers, Diana Tietjens (Ed). Boulder: Westview Press, 1997.

Brison, Susan "Surviving Sexual Violence: A Philosophical Perspective", Journal of Social Philosophy, 24(1), 5-22, 1986.

Culbertson, Roberta, "Embodied Memory, Transcendence, and Telling: Re-counting Trauma, Re-establishing the Self", New Literary History, 26, 169-195.

Tourmey, Judith, "Exploitation, Oppression, and Self Sacrifice" in Women in Philosophy, Gould, Carol C., and Wartofsky, Marx, W. (Eds). New York: Putnam, 1976

Wendell, Susan, "Oppression and Victimization: Choice and Responsibility" in Nagging Questions, Bushnell, Dana (Ed). Lanham: Rowman and Littlefield, 1995.

Interdisciplinary

Literary and Cultural Studies

Ahmed, Sara, "Beyond Humanism and Postmodernism: Theorizing a Feminist Practice", Hypatia, 11(2), 71-93, Spring 1996.

Barnes, Hazel E., "Sartre and Sexism", Philosophy and Literature, 340-347, October 1990.

Benstock, Shari, "The Female Self Engendered - Autobiographical Writing and Theories of Selfhood", Women's Studies, 20(1), 5-14, 1991.

Bowles, Gloria, "Going Back Through My Journals: The Unsettled Self", National Women's Studies Association Journal, 6(2), 255-275, Summer 1994.

Brod, Harry, "The New Men's Studies: From Feminist Theory to Gender Scholarship", Hypatia, 2, 179-196, Winter 1987.

Carroll, Shireen, and Carse, Wendy, and Trefzer, Annette, "Fashioning Professional Selves", Critical Matrix, 7(1), 63-79, 1993.

Cocks, Joan, "Cultural Theory Looks at Identity and Contradiction", Quest, 38-60, December 1990.

Coldwell, C., "Discipline and Control: Butler and Deleuze on Individuality and Dividuality", Philosophy Today, 40(1-4)211-216, Spring 1996.

Deveaux, Monique, "Feminism and Empowerment: A Critical Reading of Foucault", Feminist Studies, 20(2), 223-247, Summer 1994.

Farganis, Sondra, "Postmodernism and Feminism" in Postmodernism and Social Inquiry, Dickens, David R. (Ed). New York: Guiford, 1994.

Flax, Jane, "Postmodernism and Gender Relations in Feminist Theory", Signs, 12(4), 621-643, Summer 1987.

Fraser, Nancy, "Uses and Abuses of French Discourse Theory" in Revaluing French Feminism, Fraser, Nancy and Bartky, Sandra Lee (Eds). Bloomington: Indiana University Press, 1992.

Fraser, Nancy. Unruly Practices. Minneapolis: University of Minnesota Press, 1989.

Henke, Jill Birnie, and Umble, Diane Zimmerman, and Smith, Nancy, J., "Construction of the Female Self: Feminist Readings of the Disney Heroine", Women's Studies in Communication, 19(2), 229-249, Summer 1996.

hooks, bell. Sisters of the Yam. Boston: South End Press, 1993.

Kamuf, Peggy, and Miller, Nancy K., "Parisian Letters: Between Feminism and Deconstruction", in Conflicts in Feminism, Hirsch, Marianne, and Keller, Evelyn Fox (Eds). New York: Routledge, 1990.

Keller, Catherine, "'To Illuminate Your Trace': Self in Late Modern Feminist Theology", Listening, 211-224, Fall 1990.

Kozel, Susan, "The Diabolical Strategy of Mimesis: Luce Irigaray's Reading of Maurice Merleau-Ponty", Hypatia, 11(3) 114-129, Summer 1996.

Kristeva, Julia. Tales of Love, Leon Roudiez (Trans). New York: Columbia University Press, 1987.

Kupfer, Joseph H., "Prostitutes, Musicians, and Self-Respect", Journal of Social Philosophy, 26(3), 75-88, Winter 1995.

Linden, Robin Ruth. Making Stories, Making Selves. Columbus: Ohio State University Press, 1992.

Lugones, María, "On 'Borderlands/La Frontera': An Interpretive Essay", Hypatia 7(4), 31-37, Fall 1992.

McNay, Lois. Foucault and Feminism. Cambridge: Polity Press, 1992.

Perreault, Jeanne. Writing Selves. Minneapolis: University of Minnesota Press, 1995.

Pilardi, Jo-Ann, "Philosophy Becomes Autobiography: The Development of the Self in the Writings of Simone de Beauvoir", in Writing the Politics of Difference, Silverman, Hugh J. (Ed). Albany: State University of New York Press, 1991.

Porritt, Ruth, "Surpassing Derrida's Deconstructed Self - Virginia Woolf's Poetic Disarticulation of the Self", Women's Studies, 21(3), 323-338, 1992.

Robinson, Sally. Engendering the Subject. Albany: State University of New York Press, 1991.

Scholz, Sally, "A Critique of Jean Bethke Elsthain's Reconstruction of the Public and the Private", Continental Philosophy, 19-23, July-August 1991.

Smith, Sidonie, "Self, Subject, and Resistance - Marginalities and Twentieth-Century Autobiographical Practice", Tulsa Studies in Women's Literature, 9, 11-24, Spring 1990.

Torres, Lourdes, "The Construction of the Self in U.S. Latina Autobiographies" in Women, Knowledge, and Reality, Garry, Ann and Pearsall Marilyn (Eds). New York: Routledge, 1996 (2nd Edition).

Weedon, Chris. Feminist Practice and Poststructuralist Theory. Oxford: Blackwell, 1987.

Wetherell, Margaret, "Linguistic Repertoires and Literary Criticism: New Directions for a Social Psychology of Gender" in Towards a New Psychology of Gender, Gergen, Mary M., and Davis, Sara N. (Eds). New York: Routledge, 1997.

Psychology and Psychoanalysis

Benjamin, Jessica. Like Subjects, Love Objects. New Haven: Yale University Press, 1995.

Benjamin, Jessica. The Bonds of Love. New York: Pantheon, 1988.

Benjamin, Jessica, "A Desire of One's Own: Psychoanalytic Feminism and Subjective Space" in Feminist Studies/Critical Studies, Lauretis, Teresa de (Ed). Bloomington: Indiana University Press, 1986.

Chodorow, Nancy, "Gender as Personal and Cultural Construction", Signs, 20, 516-544, Spring 1995.

Chodorow, Nancy. Feminism and Psychoanalytic Theory. New Haven: Yale University Press, 1989.

Chodorow, Nancy. The Reproduction of Motherhood. Berkeley: University of California Press, 1978.

Dinnerstein, Dorothy. The Mermaid and the Minotaur. New York: Harper, 1976.

Donchin, Anne, "Concepts of Woman in Psychoanalytic Theory: The Nature-Nurture Controversy Revisited", in Beyond Domination, Gould, Carol C. (Ed). Totowa: Rowman and Allenheld, 1984.

Flax, Jane, "Forgotten Forms of Close Combat: Mothers and Daughters Revisited" in Towards a New Psychology of Gender, Gergen, Mary M., and Davis, Sara N. (Eds). New York: Routledge, 1997.

Flax, Jane, "Multiples: On the Contemporary Politics of Subjectivity", Human Studies, 16(1-2), 33-49, April 1993.

Flax, Jane. Thinking Fragments. Berkeley: University of California Press, 1989.

Gallop, Jane, "The Daughter's Seduction: Feminism and Psychoanalysis", Women's Studies International

Forum, 7(6), 522-524, 1984.

Gallop, Jane. The Daughter's Seduction. Ithaca: Cornell University Press, 1982.

Gallop, Jane, and Burke, Carolyn G., "Psychoanalysis and Feminism in France" in The Future of Difference, Eisenstein, Hester, and Jardine, Alice (Eds). New York: Barnard College Women's Center, 1980.

Garner, Shirley Nelson, "Feminism, Psychoanalysis, and the Heterosexual Imperative", in Feminism and Psychoanalysis, Feldstein, Richard, and Roofs, Judith (Eds). Ithaca: Cornell University Press, 1989. ALSO IN: Feminism and Philosophy, Tuana, Nancy, and Tong, Rosemarie (Eds). Boulder: Westview Press, 1995.

Gergen, Mary M., "Life Stories: Pieces of a Dream" in Towards a New Psychology of Gender, Gergen, Mary M., and Davis, Sara N. (Eds). New York: Routledge, 1997.

Gotz, Ignacio L., "Education and the Self: Cross-Cultural Perspectives", Educational Theory, 45(4), 479-795, Fall 1995.

Huff, Margaret C., "The Interdependent Self: An Integrated Concept From Feminist Theology and Feminist Psychology", Theology, 2, 160-172, Winter 1987.

Kristeva, Julia. Black Sun, Leon Roudiez (Trans). New York: Columbia University Press, 1989.

Kristeva, Julia. Revolution in Poetic Language, Margaret Waller (Trans). New York: Columbia University Press, 1984.

Kristeva, Julia. Powers of Horror, Leon Roudiez (Trans). New York: Columbia University Press, 1982.

Kristeva, Julia. Desire in Language, Thomas Gora, Alice Jardine, and Leon Roudiez (Trans), Roudiez (Ed). New York: Columbia University Press, 1980.

Leland, Dorothy, &;quot;Lacanian Psychoanalysis and French Feminism: Toward an Adequate Political Psychology", in Revaluing French Feminism, Fraser, Nancy, and Bartky, Sandra Lee (Eds). Bloomington: Indiana University Press, 1992.

Lauretis, Teresa de. Alice Doesn't. Bloomington: Indiana University Press, 1993.

Mahoney, Maureen A., and Yogvesson, Barbara, "The Construction of Subjectivity and the Paradox of Resistance: Reintegrating Feminist Anthropology and Psychology", Signs, 18(1), 44-73, Autumn 1992.

Meyers, Diana T., "The Subversion of Women's Agency in Psychoanalytic Feminism: Chodorow, Flax, Kristeva", in Revaluing French Feminism, Fraser, Nancy, and Bartky, Sandra Lee (Eds). Bloomington: Indiana University Press, 1992.

Miller, Jean Baker. Toward A New Psychology of Women. Boston: Beacon Press, 1972.

Radden, Jennifer. Divided Minds and Successive Selves. Cambridge: MIT Press, 1996.

Radden, Jennifer. Madness and Reason. London: G. Allen and Unwin, 1985.

Simonds, Wendy. Women and Self-Help Culture. New Brunswick: Rutgers University Press, 1992.

Acknowledgements

My special thanks to Diana Tietjens Meyers and Sally Haslanger for both this opportunity and their guidance.

Return to Feminist Perspectives on the Self

Lisa Cassidy
University of Connecticut
lmc96002@uconnvm.uconn.edu

First published: June 28, 1999
Content last modified: June 28, 1999

Leibniz's Philosophy of Mind

In a more popular view, Leibniz's place in the history of the philosophy of mind is best secured by his pre-established harmony, that is, roughly, by the thesis that there is no mind-body interaction strictly speaking, but only a non-causal relationship of harmony, parallelism, or correspondence between mind and body. Certainly, the pre-established harmony is important for a proper understanding of Leibniz's philosophy of mind, but there is much more to be considered as well, and even in connection with the pre-established harmony, the more popular view needs to be refined, particularly insofar as it suggests that Leibniz accepts a roughly Cartesian, albeit non-interactionist dualism, which he does not. In fact, Leibniz is justly famous for his critiques, not only of materialism, but also of such a dualism. (Whether Leibniz accepts, throughout his maturity, the idealistic view that all substances are simple unextended substances or monads is an important interpretive issue that has been discussed widely in recent years. We shall not try to resolve the issue here.) In short, Leibniz made important contributions to a number of classical topics of the philosophy of mind, including materialism, dualism, idealism and mind-body interaction.

But Leibniz has much to say about the philosophy of mind that goes well beyond these traditionally important topics. Perhaps surprisingly, his system sometimes contains ideas of relevance even to contemporary discussions in the cognitive sciences. More generally, he discusses in depth the nature of perception and thought (conscious and unconscious), and of human motivation and striving (or, as he would say, appetition). We will take up such topics in what follows.

1. Matter and Thought

For present purposes, we may think of materialism as the view that everything that exists is material, or

physical, with this view closely allied to another, namely, that mental states and processes are either identical to, or realized by, physical states and processes. Leibniz remained opposed to materialism throughout his career, particularly as it figured in the writings of Epicurus and Hobbes. The realms of the mental and the physical, for Leibniz, form two distinct realms—but not in a way conducive to dualism, or the view that there exists both thinking substance, and extended substance. By opposing both materialism and dualism, Leibniz carved himself an interesting place in the history of views concerning the relationship between thought and matter.

Most of Leibniz's arguments against materialism are directly aimed at the thesis that perception and consciousness can be given mechanical (i.e. physical) explanations. His position is that perception and consciousness cannot *possibly* be explained mechanically, and, hence, could not be physical processes. His most famous argument against the possibility of materialism is found in section 17 of the *Monadology* (1714):

> One is obliged to admit that *perception* and what depends upon it is *inexplicable on mechanical principles*, that is, by figures and motions. In imagining that there is a machine whose construction would enable it to think, to sense, and to have perception, one could conceive it enlarged while retaining the same proportions, so that one could enter into it, just like into a windmill. Supposing this, one should, when visiting within it, find only parts pushing one another, and never anything by which to explain a perception. Thus it is in the simple substance, and not in the composite or in the machine, that one must look for perception.

Leibniz's argument seems to be this: the visitor of the machine, upon entering it, would observe nothing but the properties of the parts, and the relations they bear to one another. But no explanation of perception, or consciousness, can possibly be deduced from this conglomerate. No matter how complex the inner workings of this machine, nothing about them reveals that what is being observed are the inner workings of a conscious being. Hence, materialism must be false, for there is no possible way that the purely mechanical principles of materialism can account for the phenomena of consciousness.

In other writings, Leibniz suggests exactly what characteristic it is of perception and consciousness that the mechanical principles of materialism cannot account for. The following passages, the first from the *New System of Nature* (1695), the second from the *Reply to Bayle* (1702), are revealing in this regard:

> Furthermore, by means of the soul or form, there is a true unity which corresponds to what is called the *I* in us; such a thing could not occur in artificial machines, nor in the simple mass of matter, however organized it may be.

> But in addition to the general principles which establish the monads of which compound things are merely the results, internal experience refutes the Epicurean [i.e. materialist] doctrine. This experience is the consciousness which is in us of this *I* which apperceives things which occur in the body. This perception cannot be explained by figures and movements.

Leibniz's point is that whatever is the subject of perception and consciousness must be truly one, a single "I" properly regarded as *one* conscious being. An aggregate of matter is not truly one and so cannot be regarded as a single *I*, capable of being the subject of a unified mental life. This interpretation fits nicely with Lebniz's oft-repeated definition of perception as "the representation in the simple of the compound, or of that which is outside" (*Principles of Nature and Grace*, sec.2 (1714)). More explicitly, in a letter to Antoine Arnauld of 9 October 1687, Leibniz wrote that "in natural perception and sensation, it is enough for what is divisible and material and dispersed into many entities to be expressed or represented in a single indivisible entity or in a substance which is endowed with genuine unity." If perception (and hence, consciousness) essentially involves a representation of a variety of content in a simple, indivisible "I," then we may construct Leibniz's argument against materialism as follows: Materialism holds that matter can explain (is identical with, can give rise to) perception. A perception is a state whereby a variety of content is represented in a true unity. Thus, whatever is not a true unity cannot give rise to perception. Whatever is divisible is not a true unity. Matter is infinitely divisible. Hence, matter cannot form a true unity. Hence, matter cannot explain (be identical with, give rise to) perception. If matter cannot explain (be identical to, give rise to) perception, then materialism is false. Hence, materialism is false.

Leibniz rejected materialism on the grounds that it could not, in principle, ever capture the "true unity" of perceptual consciousness, that characteristic of the self which can simultaneously unify a manifoldness of perceptual content. If this is Leibniz's argument, it is of some historical interest that it bears striking resemblances to contemporary objections to certain materialist theories of mind. Many contemporary philosophers have objected to some versions of materialism on the basis of thought experiments like Leibniz's: experiments designed to show that qualia and consciousness are bound to elude certain materialist conceptions of the mind (cf. Searle 1980; Nagel 1974; McGinn 1989; Jackson 1982).

Leibniz's rejection of materialist conceptions of the mind was coupled with a strong opposition to dualistic views concerning the relationship between mind and body, particularly the substance dualism that figured in the philosophy of Descartes and his followers. According to this dualism, the world fundamentally consists of two disparate substances: extended material substance (body) and unextended thinking substance (mind). This bifurcation, of course, carries no burden of holding that the operations of the mental are realized by the operations of the physical. But despite his claim that consciousness and perception cannot be realized by, nor reduced to, the mechanical operations of matter, Leibniz found the alternative of postulating two distinct kinds of substance equally implausible.

Leibniz's opposition to Cartesian dualism stems not from a rejection of unextended substance, but from his denial of the existence of genuine extended material substance. To begin with, Leibniz held the Scholastic thesis that "being" and "one" are equivalent. He writes to Arnauld: "To be brief, I hold as axiomatic the identical proposition which varies only in emphasis: that what is not truly *one* being is not truly one *being* either" (30 April 1687). For Leibniz, in order for something to count as a real being—a substance—it must be "truly one," or an entity endowed with genuine unity. And, as we saw above, in order for something to be a genuine unity, it must be a simple, indivisible entity. "Substantial unity," he writes, "requires a complete, indivisible and naturally indestructible entity" (to Arnauld, 28 November 1686). But matter is extended, and thus, Leibniz believes, infinitely divisible. Hence, there is no such

thing, for Leibniz, as material substance.

There is a positive thesis which goes hand-in-hand with Leibniz's negative thesis against material substance, and which helps to explain further his rejection of material substance. It is summarized in the following passage from a letter to Arnauld of 30 April 1687:

> I believe that *where there are only beings through aggregation, there will not even be real beings*. For every being through aggregation presupposes beings endowed with a true unity, because it obtains its reality from nowhere but that of its constituents, so that it will have no reality at all if each constituent being is still an entity through aggregation; or else, one must yet seek another basis to its reality, which in this way, if one must constantly go on searching, can never be found…. If there are aggregates of substances, there must also be genuine substances from which all the aggregates result. One must therefore necessarily arrive either at mathematical points from which certain authors make up extension, or at Epicurus' and M. Cordemoy's atoms (which you, like me, dismiss), or else one must acknowledge that no reality can be found in bodies, or finally one must recognize certain substances in them that possess a true unity.

According to Leibniz, bodies (qua material) are aggregates, and an aggregate, of course, is not a substance on account of its lack of unity. The claim in the above passage is that whatever being, or reality, an aggregate has derives from the being and reality of its constituents. Thus, Leibniz thinks that if a body is to have any reality at all, if it is to be more than a mere "phenomenon, lacking all reality as would a coherent dream," then it must ultimately be composed of things which are real beings. Atoms, he claims, are unfit for this role, because they are themselves extended beings, and for Leibniz, divisibility is of the essence of extension. That is, those who believe in indivisible atoms make matter "divisible in one place, indivisible in another" (*On Nature Itself* (1698)), but "we cannot explain why bodies of a definite smallness [i.e. atoms] should not be further divisible" (*Primary Truths* (1686)). Every extended mass, for Leibniz, is composed of extended parts, and so even if we could conceive of an atom as composed of parts which cannot be physically divided, "an invincible attachment of one part to another would not at all destroy the diversity of these parts" (*New System of Nature*, (1695)), or it would not at all overcome the fact that it is an aggregate composed of parts, and not truly one being. Likewise, mathematical points, "even an infinity of points gathered into one, will not make extension," (to Des Bosses, 30 April 1709) and so cannot be understood as the constituents of extended bodies. Hence, Leibniz opts for the last of the above quoted alternatives: the constituents of bodies are "certain substances … that possess a true unity." These substances are partless, unextended, and indivisible, and therefore real beings in Leibniz's sense. Indeed, in several writings, Leibniz invites us to conceive of these substances on the model of our notion of souls. These simple substances are the only things which suffice for grounding the reality of bodies. To be sure, substances, Leibniz tells us, do not constitute a body as parts of the body, but as the "first elements," or "primitive unities," of the body. As he sometimes puts it, bodies "result from" these constitutive unities. That is, bodies just are aggregates of substances which *appear to us* as extended corporeal phenomena, though they are "well-founded" phenomena; they have their foundation in real beings.

In short, Leibniz stands in a special position with respect to the history of views concerning thought and its relationship to matter. He rejects the materialist position that thought and consciousness can be captured by purely mechanical principles. But he also rejects the dualist position that the universe must therefore be bifurcated into two different kinds of substance, thinking substance, and material substance. Rather, it is his view that the world consists solely of one *type* of substance, though there are infinitely many substances of that type. These substances are partless, unextended entities, some of which are endowed with thought and consciousness, and others of which found the phenomenality of the corporeal world. The sum of these views secures Leibniz a distinctive position in the history of the philosophy of mind.

2. Denial of Mind-Body Interaction, Assertion of Pre-established Harmony

A central philosophical issue of the seventeenth century concerned the apparent causal relations which hold between the mind and the body. In most seventeenth-century settings this issue was discussed within the context of substance dualism, the view that mind and body are different kinds of substance. For Leibniz, this is a particularly interesting issue in that he remained fundamentally opposed to dualism. But although Leibniz held that there is only one type of substance in the world, and thus that mind and body are ultimately composed of the same kind of substance (a version of monism), he also held that mind and body are metaphysically distinct. There are a variety of interpretations of what this metaphysical distinctness consists in for Leibniz, but on any plausible interpretation it is safe to assume (as Leibniz seems to have done) that for any person *P*, *P*'s mind is a distinct substance (a soul) from *P*'s body. With this assumption in hand, we may formulate the central issue in the form of a question: how is it that certain mental states and events are coordinated with certain bodily states and events, and vice-versa? There were various attempts to answer this question in Leibniz's time period. For Descartes, the answer was mind-body interactionism: the mind can causally influence the body, and (most commentators have held) vice-versa. For Malebranche, the answer was that neither created minds nor bodies can enter into causal relations because God is the only causally efficient being in the universe. God causes certain bodily states and events on the occasion of certain mental states and events, and vice-versa. Leibniz found Descartes' answer unintelligible (cf. *Theodicy*, sec. 60), and Malebranche's excessive because miraculous (cf. Letter to Arnauld, 14 July 1686).

Leibniz's account of mind-body causation was in terms of his famous doctrine of the *preestablished harmony*. According to the latter, (1) no state of a created substance has as a real cause some state of another created substance (i.e. a denial of inter-substantial causality); (2) every non-initial, non-miraculous, state of a created substance has as a real cause some previous state of that very substance (i.e. an affirmation of intra-substantial causality); and (3) each created substance is programmed at creation such that all its natural states and actions are carried out in conformity with all the natural states and actions of every other created substance.

Formulating (1) through (3) in the language of minds and bodies, Leibniz held that no mental state has as

a real cause some state of another created mind or body, and no bodily state has as a real cause some state of another created mind or body. Further, every non-initial, non-miraculous, mental state of a substance has as a real cause some previous state of that mind, and every non-initial, non-miraculous, bodily state has as a real cause some previous state of that body. Finally, created minds and bodies are programmed at creation such that all their natural states and actions are carried out in mutual coordination.

According to Leibniz, what *appear* to be real causal relations between mind and body are, in metaphysical reality, the mutual conformity or coordination of mind and body—in accordance with (3)—with no interaction or divine intervention involved. For example, suppose that Smith is pricked with a pin (call this bodily state *Sb*) and pain ensues (call this mental state *Sm*), a case of apparent body to mind causation. Leibniz would say that in such a case some state of Smith's mind (soul) prior to *Sm* was the real cause of *Sm,* and *Sb* was not a causal factor in the obtaining of *Sm*. Suppose now that Smith has a desire to raise his arm (call this mental state *Sm*), and the raising of his arm ensues (call this bodily state *Sb*), a case of apparent mind to body causation. Leibniz would say that in such a case some state of Smith's body prior to *Sb* was the real cause of *Sb* and *Sm* was not a causal factor in the obtaining of *Sb*. So although substances do not causally interact, their states accommodate one another as if there were causal interaction among substances.

It should be noted, however, that Leibniz did think that there was a sense in which one could say that mental events influence bodily events, and vice-versa. He wrote to Antoine Arnauld that although “one particular substance has no physical influence on another … nevertheless, one is quite right to say that my will is the cause of this movement of my arm …; for the one expresses distinctly what the other expresses more confusedly, and one must ascribe the action to the substance whose expression is more distinct” (28 November 1686 (draft)). In this passage, Leibniz sets forth what he takes the metaphysical reality of apparent inter-substantial causation to amount to. We begin with the thesis that every created substance perceives the entire universe, though only a portion of it is perceived distinctly, most of it being perceived unconsciously, and, hence, confusedly. Now consider two created substances, *x* and *y* (*x* not identical to *y*), where some state of *x* is said to be the cause of some state of *y*. Leibniz's analysis is this: when the causal state of affairs occurred, the relevant perceptions of substance *x* became more distinct, while the relevant perceptions of substance *y* became more confused. Insofar as the relevant perceptions of *x* become increasingly distinct, it is “causally” active; insofar as the relevant perceptions of substance *y* become increasingly confused, it is passive. In general, causation is to be understood as an increase in distinctness on the part of the causally active substance, and an increase in confusedness on the part of the passively effected substance. Again, each substance is programmed at creation to be active/passive at the relevant moment, with no occurrence of real substantial interaction.

It is difficult to say exactly why Leibniz denied inter-substantial causation. Some of the things he tells us, in both private and public writings, seem unsatisfactory for one reason or another. For example, in *Primary Truths* (1686?), we are given this:

> Strictly speaking, one can say that *no created substance exerts a metaphysical action or influx on any other thing*. For, not to mention the fact that one cannot explain how something can pass from one thing into the substance of another, we have already shown

> that from the notion of each and every thing follows all of its future states. What we call causes are only concurrent requisites, in metaphysical rigor.

Here Leibniz gives a reason tied to his complete concept theory of substance, according to which “the nature of an individual substance or of a complete being is to have a notion so complete that it is sufficient to contain and to allow us to deduce from it all the predicates of the subject to which this notion is attributed” (*Discourse on Metaphysics*, sec. 8). But there are, it seems, at least two problems with this explanation. First, Leibniz moves rather quickly from a conceptual explanation of substance in terms of the complete concept theory, to the conclusion that this consideration is sufficient to explain the activity of concrete substances. Second, even if conceptual considerations about substances were sufficient to explain their apparent causal activity, it does not seem to follow that substances do not interact—unless one is assuming that causal overdetermination is not a genuine possibility. Leibniz seems to be assuming just that, but without argument.

Sometimes Leibniz gives a more familiar line of reasoning. At *Monadology* 7, we read this:

> There is no way of explaining how a monad can be altered or changed internally by some other creature, since one cannot transpose anything in it, nor can one conceive of any internal motion that can be excited, directed, augmented, or diminished within it, as can be done in composites, where there can be change among the parts. The monads have no windows through which something can enter or leave. Accidents cannot be detached, nor can they go about outside of substances, as the sensible species of the Scholastics once did. Thus, neither substance nor accident can enter a monad from without.

He seems to think that causal interaction between two beings requires the transmission or transposition of the parts of those beings. But substances are simple unextended entities which contain no parts. Thus, there is no way to explain how one substance could influence another. Unfortunately, however, this line of reasoning would seem to also rule out one case of inter-substantial causation which Leibniz allows, viz., God's causal action on finite simple substances.

3. Language and Mind

Some scholars have suggested that Leibniz should be regarded as one of the first thinkers to envision something like the idea of artificial intelligence (cf. Churchland 1984; Pratt 1987). Whether or not he should be regarded as such, it is clear that Leibniz, like contemporary cognitive scientists, saw an intimate connection between the form and content of language, and the operations of the mind. Indeed, according to his own testimony in the *New Essays*, he “really believe[s] that languages are the best mirror of the human mind, and that a precise analysis of the signification of words would tell us more than anything else about the operations of the understanding” (bk.III, ch.7, sec.6 (RB, 333)). This view of Leibniz's led him to formulate a plan for a “universal language,” an artificial language composed of symbols, which would stand for concepts or ideas, and logical rules for their valid manipulation. He believed that such a language would perfectly mirror the processes of intelligible human reasoning. It is this plan that has led

some to believe that Leibniz came close to anticipating artificial intelligence. At any rate, Leibniz's writings about this project (which, it should be noted, he never got the chance to actualize) reveal significant insights into his understanding of the nature of human reasoning. This understanding, it turns out, is not that different from contemporary conceptions of the mind, as many of his discussions bear considerable relevance to discussions in the cognitive sciences.

According to Leibniz, natural language, despite its powerful resources for communication, often makes reasoning obscure since it is an imperfect mirror of intelligible thoughts. As a result, it is often difficult to reason with the apparatus of natural language, “since it is full of innumerable equivocations” (*On the Universal Science: Characteristic* (undated); G VII, 205 (S, 18)). Perhaps this is because of his view that the terms of natural language stand for complex, or derivative, concepts—concepts which are composed of, and reducible to, simpler concepts. With this “combinatorial” view of concepts in hand, Leibniz notices “that all human ideas can be resolved into a few as their primitives” (*On the Universal Science: Characteristic*; G VII, 205 (S, 18)). We could then assign symbols, or “characters,” to these primitive concepts from which we could form characters for derivative concepts by means of combinations of the symbols. As a result, Leibniz tells us, “it would be possible to find correct definitions and values and, hence, also the properties which are demonstrably implied in the definitions” (*On the Universal Science: Characteristic*; G VII, 205 (S, 19)). The totality of these symbols would form a “universal characteristic,” an ideal language in which all human concepts would be perfectly represented, and their constitutive nature perfectly transparent. Now it is true that Leibniz eventually came to doubt “whether any concept of this [primitive] kind appears distinctly to men, namely, in such a way that they know they have it” (*An Introduction to a Secret Encyclopedia* (1679?); C, 513 (MP, 7)). But it is also clear that he did not see this skepticism concerning our ability to reach the primitive concepts as much of a barrier to the project of a universal language. He writes in *The Art of Discovery* (1685) that “there are certain primitive terms which can be posited, if not absolutely, at least relatively to us” (C, 176 (W, 51)). The suggestion seems to be that even if we cannot provide a catalog of absolutely primitive concepts, we can nevertheless construct a characteristic based on concepts which cannot be further resolved by humans.

In addition to the resolution of concepts, and their symbolic assignments, Leibniz envisages the formulation of logical rules for the universal characteristic. He claims that “it is plain that men make use in reasoning of several axioms which are not yet quite certain” (*The Method of Certitude and the Art of Discovery* (undated); G VII, 183 (W, 49)). Yet with the explicit formulation of these rules for the logical manipulation of the symbols—rules which humans use in reasoning—we would be in possession of a universal language which would mirror the relations between the concepts used in human reasoning. Indeed, the universal characteristic was intended by Leibniz as an instrument for the effective calculation of truths. Like formal logic systems, it would be a language capable of representing valid reasoning patterns by means of the use of symbols. Unlike formal logic systems, however, the universal language would also express the content of human reasoning in addition to its formal structure. In Leibniz's mind, “this language will be the greatest instrument of reason,” for “when there are disputes among persons, we can simply say: Let us calculate, without further ado, and see who is right” (*The Art of Discovery* (1685); C, 176 (W, 51)).

Judging from Leibniz's plans for a universal language, it is clear that Leibniz had a specific view about the

nature of human cognitive processes, particularly about the nature of human reasoning. According to this view, cognition is essentially symbolic: it takes place in a system of representations which possesses language-like structure. Indeed, it was Leibniz's view that “all human reasoning uses certain signs or characters,” (*On the Universal Science: Characteristic*; G VII, 204 (S, 17)) and “if there were no characters, we could neither think of anything distinctly nor reason about it” (*Dialogue* (1677); G VII, 191 (A&G, 271)). Add to this conception Leibniz's view that human cognitive processes follow determinable axioms of logic, and the picture that emerges is one according to which the mind operates, at least when it comes to intelligible reasoning, by following implicit algorithmic procedures. Regardless of whether or not Leibniz should be seen as the grandfather of artificial intelligence, he did conceive of human cognition in essentially computational terms. In fact, as early as 1666, remarking favorably on Hobbes' writings, Leibniz wrote: “Thomas Hobbes, everywhere a profound examiner of principles, rightly stated that everything done by our mind is a *computation*” (*On the Art of Combinations* (1666); G IV, 64 (P, 3)).

4. Perception and Appetition

What do we find in the human mind? Representations on the one hand, and tendencies, inclinations, or strivings on the other, according to Leibniz. Or, to put this in Leibniz's more customary terminology, what is found within us is perception and appetition. For human minds count for Leibniz as simple substances, and, as he says in a letter to De Volder, “it may be said that there is nothing in the world except simple substances, and, in them, perception and appetite.” (30 June 1704)

Perception has already been discussed briefly above. But it will be advisable to consider also a definition from a letter to Des Bosses (and echoed in many other passages), in which Leibniz discusses perception as the representation or “expression” of “the many in the one” (letter to Des Bosses, 11 July 1706). We shall return to this definition below. Appetitions are explained as “tendencies from one perception to another” (*Principles of Nature and Grace*, sec.2 (1714)). Thus, we represent the world in our perceptions, and these representations are linked with an internal principle of activity and change (*Monadology*, sec.15 (1714)) which, in its expression in appetitions, urges us ever onward in the constantly changing flow of mental life. More technically explained, the principle of action, that is, the primitive force which is our essence, expresses itself in momentary derivative forces involving two aspects: on the one hand, there is a representative aspect (perception), by which that the many without are expressed within the one, the simple substance; on the other, there is a dynamical aspect, a tendency or striving towards new perceptions, which inclines us to change our representative state, to move towards new perceptions.

It should not be inferred that this appetitive tendency to change is entirely mechanistic, entirely governed by efficient causation only. For in Leibniz's view, value and final causes are not excluded from the action of the mind, the change of mental states. As he says in section 13 of the *Discourse on Metaphysics* (1686), just as “God will always do the best, … a man shall always do … that which appears to him to be the best.” Appearance, of course, has to do with perception; doing, with appetition. So this principle of human action applies directly, as one would expect, to the two key factors of monadic interior life, only with the role of value, or an end in view, now more clearly in focus. This is why Leibniz says that, at the

level of bodies (that is, for Leibniz, at the level of well-founded phenomena), all occurs according to the laws of efficient causes; whereas with respect to perceptions and appetites (or at least with some of these—interpretations differ here) all occurs according to the laws of final causes. But there is no clash here, given the harmony of the kingdom of nature and the kingdom of grace in Leibniz's system, the harmony of final and efficient causes.

To be sure, at an ultimate level, the only actions of substances *are* changes of perceptions. Thus, at the ultimate level, the appetitions are not so much the tendencies impelling a person towards voluntary motions of the human body (although at the level of well-founded phenomena this may indeed be the case) but rather tendencies arising out of present perceptions (present appearances) towards new perceptions. This explains why Leibniz defines appetitions in the initially surprising way noted above, as “tendencies from one perception to *another*”—another *perception,* that is.

The last two paragraphs have helped to clarify appetition. It is time to return to perception. In Leibniz's definition (the expression of the many in the one) the two key terms are ‘expression’ and ‘one’. Both of them bear considerable weight in Leibniz's metaphysics. Representation or expression (Leibniz uses the two terms interchangeably) has its own definition: “One thing expresses another … when there is a constant and regulated relation between what can be said of the one and of the other” (letter to Arnauld, 9 October 1687). Examples, in addition to perception, include a map expressing or representing a geographical region and an algebraic equation representing or expressing a geometric figure, such as a circle or an ellipse.

With respect to oneness, Leibniz famously claims a connection with being. He says, “I hold this identical proposition, differentiated only by the emphasis, to be an axiom, namely, *that what is not truly* one *being is not truly one* being *either*” (letter to Arnauld, 30 April 1687). For Leibniz, what truly is is substance, so it is not surprising that at one point he clarifies his definition of perception by saying that perception is “the expression of many things in one, *or in simple substance*” (*A New Method of Learning and Teaching Jurisprudence*, revision notes of 1697-1700).

Finally, it should be recalled that for Leibniz there are quite distinct levels of perception among created substances. Some of these will be taken up in more detail in the following section, but the basic point for now is that the three major levels, from the lowest to the highest, are bare perception (without special distinctness or memory), sensation (with heightened distinctness and memory), and thought (with distinctness, memory, and reflection). These are distinctive of the three levels of monads, respectively, the bare monads, souls, and spirits. Only the last of these may properly be said to have reason. Only the last of these is strictly a mind in the Leibnizian classification.

5. Apperception, Desire and the Unconscious

One of the better-known terms of Leibniz's philosophy, and of his philosophy of mind, is apperception. A famous definition is presented in section 4 of the *Principles of Nature and of Grace* (1714), where Leibniz says that apperception is “*consciousness,* or the reflective knowledge of this internal state.” He

adds that this is “something not given to all souls, nor at all times to a given soul.”

Despite being well known, Leibniz's concept of apperception is not necessarily well understood. In particular, the place of apperception in the three-fold classifications given just above—of three kinds of perceptions and of simple substances—is not agreed upon, despite the fact that this would seem to be of considerable importance. A common understanding is that for Leibniz apperception is distinctive of spirits, and is not present in even the highest of animals beneath humans. While there is evidence that Leibniz at least sometimes adopts this position, there is also evidence that he sometimes endorses the view that (at least some) beasts also apperceive. Since we may assume that at a minimum apperception involves consciousness (though not necessarily certain higher forms of consciousness, e.g., self-consciousness, or reflective consciousness, in one sense or another), this leads to some uncertainty as to whether Leibniz assigns consciousness to beasts, that is, whether he does or does not agree with the famous Cartesian principle that beasts are not conscious, but only material automata.

There are at least three specific lines of evidence for apperception in beasts. The first is that Leibniz sometimes uses very similar definitions and examples when talking about the contrast between, on the one hand, apperceptions and *petites* perceptions (perceptions which are not apperceived), and, on the other, sensation and bare perceptions. This suggests, though it does not demonstrate, that Leibniz is identifying apperception and sensation, not apperception and rational thought. The second line of evidence is that he often appears to take the side of the common man against Descartes' position on beasts, for example, when he says,

> It will be difficult to rid mankind of this opinion which has been held always and everywhere and which is universal if any opinion deserves that term, namely, that beasts have feelings (letter to Arnauld, 9 October 1687).

Finally, there are passages, notably in the *New Essays concerning Human Understanding* (1704), in which Leibniz quite simply ascribes apperception, directly or indirectly, to beasts, as, for example, when he discusses the case of a wild boar that has only a bare perception of a human until the human shouts at it, at which point the boar apperceives the person (”*s'apperçoit d'une personne*“)and begins a charge (Bk.II ch.21, sec.5; A vi VI 173).

Without trying to proceed further with this issue here, we can see that whichever of these views is ultimately adopted, it remains the case that Leibniz's theory of perception involves something very distinctive in an age dominated by Descartes' theory of ideas, the thesis that there are some perceptions of which we are not conscious, the much-discussed *petites* perceptions. Although Leibniz was not the first to propose such an idea (Aquinas, for example, had a similar view), and although the view in his hands did not have the explosive quality that it did in the hands of Freud, the thesis remains an intriguing and important part of his philosophy of mind. Indeed, the Preface of the *New Essays concerning Human Understanding* contains as strong a statement as one is likely to find about the centrality of this view in a particular metaphysical system. Among other things, Leibniz makes it very clear that it is not just lower simple substances that have such unconscious perceptions but also human minds.

Having raised the issue of unconscious perceptions, we should consider also the question of unconscious appetitions. This is infrequently discussed, but the question should not be overlooked. Since appetitions are tendencies or strivings, ones which profoundly influence human actions, it is of distinct human relevance whether or not an individual human is conscious of all of these strivings. Certainly, some have taken the possibility of urges of which we are not conscious as highly important for the proper understanding of individual humans, and indeed of the human condition generally.

There is evidence, notably from the *New Essays*, that Leibniz did indeed draw a parallel between perceptions and appetitions with respect to consciousness. Although he did not always explain the distinction between conscious and unconscious appetitions with care and uniformity, it seems clear that he committed himself to appetitions of which we are not conscious, or which we do not apperceive, just as he had committed himself to perceptions which are not apperceived. Consider the following two statements in combination: “desires and tendencies which are apperceived are often called volitions” (*New Essays*, Bk.II, ch.21, sec.39); and, “There are also efforts that result from insensible perceptions which one does not apperceive, and these I prefer to call *appetitions* rather than volitions (although there are also apperceptible appetitions)” (*New Essays*, Bk.2, ch.21, sec.5).

In short, and perhaps oversimplifying to a certain extent, we can say that in the Leibnizian realm of mind there are indeed only perceptions and appetitions, but in these there is a fundamental divide between the realm of consciousness and unconsciousness. In the former, there are apperceptions and desires, the perceptions and appetitions of which we are conscious. In the latter, there are perceptions and appetitions of which we are not conscious. That does not mean, however, that this latter realm is unimportant in our mental lives. As Leibniz says, “insensible perceptions are as important to [the science of minds, souls, and soul-like substances] as insensible corpuscles are to natural science, and it is just as unreasonable to reject the one as the other on the pretext that they are beyond the reach of our senses.” ((*New Essays*, Preface) He would have said the same, no doubt, about inapperceptible appetitions.

Bibliography

Works of Leibniz

A *Gottfried Wilhelm Leibniz: Sämtliche Schriften und Briefe*. Edited by the German Academy of Science. Darmstadt and Berlin: Berlin Academy, 1923-. Cited by series, volume, and page.

A&G *Philosophical Essays*. Edited and translated by Roger Ariew and Daniel Garber. Indianapolis: Hackett, 1989.

C *Opuscules et Fragments Inédits de Leibniz*. Edited by Louis Couturat. Paris: Felix Alcan, 1903.

G *Die Philosophischen Schriften von Gottfried Wilhelm Leibniz*. Edited by C.I. Gerhardt. Berlin: Weidman, 1875-1890. Cited by volume and page.

Grua *Textes Inédits*. Edited by Gaston Grua. Paris: Presses Universitaires de France, 1948.

H *Theodicy*. Edited by Austin Farrer and translated by E.M. Huggard. New Haven: Yale UP, 1952.

L *Philosophical Papers and Letters*. Edited by Leroy Loemker, 2nd ed. Dordrecht: Reidel, 1969.

LA *The Leibniz-Arnauld Correspondence*. Translated and edited by H.T. Mason. Manchester: Manchester UP, 1967.

MP *Philosophical Writings*. Translated and edited by Mary Morris and G.H.R. Parkinson. London: Dent, 1973.

P *Leibniz: Logical Papers*. Translated and edited by G.H.R. Parkinson. Oxford: Oxford UP, 1966.

RB *New Essays on Human Understanding*. Translated and edited by Peter Remnant and Jonathon Bennett. Cambridge: Cambridge UP, 1982.

S *Monadology and Other Philosophical Essays*. Translated and edited by Paul Schrecker and Anne Martin Schrecker. New York: Bobbs-Merrill Co., 1965.

W *Leibniz: Selections*. Edited by Philip P. Wiener. New York: Charles Scribner's Sons, 1951.

Secondary Literature

- Churchland, P. *Matter and Consciousness*, Cambridge: MIT Press, 1984.
- Cole, D. “Thought and Thought Experiments,” *Philosophical Studies*, **45** (1984): 431-444.
- Cook, D. “Leibniz and Hegel on the Philosophy of Language,” *Studia Leibnitiana Supplementa*, **15** (1972): 229-238.
- Dascal, M. *Leibniz: Language, Signs, and Thought*. Philadelphia: John Benjamins, 1987.
- Jackson, F. “Epiphenomenal Qualia,” *Philosophical Quarterly*, **32** (1982): 127-136.
- Kulstad, M. “Leibniz's Conception of Expression,” *Studia Leibnitiana*, **9** (1977): 55-76.
- Kulstad, M. “Some Difficulties in Leibniz's Definition of Perception,” *Leibniz: Critical and Interpretive Essays*, M. Hooker (ed.), Minneapolis: University of Minnesota, 1982: 65-78.
- Kulstad, M. *Leibniz on Apperception, Consciousness, and Reflection*, München: Philosophia, 1991.
- Kulstad, M. “Appetition in the Philosophy of Leibniz,” *Mathesis rationis: Festschrift für Heinrich Schepers*, Münster: Philosophia, 1991: 133-52.
- Lodge, P. and Bobro, M. “Stepping back inside Leibniz's Mill,” *Monist*, **81** (1998): 554-73.
- McGinn, C. “Can We Solve the Mind-Body Problem,” *Mind*, **98** (1989): 346-366.
- McRae, R. *Leibniz: Perception, Apperception, and Thought*, Toronto: Toronto UP, 1976.
- Nagel, T. “What is it Like to be a Bat?” *Philosophical Review*, **83** (1974): 435-450.
- Pratt, V. *Thinking Machines: The Evolution of Artificial Intelligence*, Oxford: Basil Blackwell, 1987.
- Rossi, P. “The Twisted Roots of Leibniz's Characteristic,” *The Leibniz Renaissance*, Florence:

Olschki, 1989: 271-289.
- Rutherford, D. “Philosophy and Language in Leibniz,” *The Cambridge Companion to Leibniz*, Cambridge: Cambridge UP, 1995: 224-269.
- Seager, W. “The Worm in the Cheese: Leibniz, Consciousness, and Matter,” *Studia Leibnitiana*, **23** (1991): 79-91.
- Searle, J. “Minds, Brains, and Programs,” *Behavioral and Brain Sciences*, **3** (1990): 417-457.
- Sleigh, R.C. “Leibniz on Malebranche on Causality,” *Central Themes in Early Modern Philosophy*, J.A.Cover and Mark Kulstad, eds. Indianapolis: Hackett, 1990: 161-194.
- Wilson, M. “Leibniz and Materialism,” *Canadian Journal of Philosophy*, **3** (1974): 495-513.

Other Internet Resources

- Markku Roinila's Leibniz webpage (University of Helsinki)
- Leibnitiana maintained by Gregory Brown (Universtity of Houston)
- J. A. Cover's Leibniz webpage (Purdue University)
- Paul Lodge's collection of Leibniz Links (Tulane University)
- Leibniz E-list, maintained by George Gale (University of Missouri-Kansas City)

Related Entries

Leibniz, Gottfried Wilhelm | Leibniz, Gottfried Wilhelm: ethics | Leibniz, Gottfried Wilhelm: modal metaphysics | Leibniz, Gottfried Wilhelm: on the problem of evil

Mark Kulstad
Rice University
kulstad@ruf.rice.edu
and
Laurence Carlin
University of Wisconsin/Oshkosh
carlin@uwosh.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 22, 1997

Content last modified: July 15, 2002

Leibniz on the Problem of Evil

Without question, the problem of evil vexed Leibniz as much as any philosophical problem during his career. This is obvious from the fact that the first and the last book length works that he authored, the PhilosopherÕs Confession (written at age 26 in 1672) and the Theodicy (written in 1709, seven years before his death) were both devoted to this problem. It is, as well, equally striking that this latter work was the only book length treatise Leibniz saw fit to publish during his life. In this entry we will examine to two main species of the problem of evil which Leibniz addresses. The first, "the underachiver problem," is the one raised by the critic who argues that the evil in our world indicates that God cannot be as knowledgeable, powerful, or good, as traditional monotheists have claimed. The second, "the holiness problem," is one raised by a critic who argues that God's intimate causal entanglements with the world make God the cause of evil. God is thereby implicated in the evil at the expense of his holiness.

The Variety of Problems of Evil in Leibniz

Without question, the problem of evil vexed Leibniz as much as any philosophical problem during his career. This is obvious from the fact that the first and the last book length works that he authored, the *Philosopher's Confession* (written at age 26 in 1672) and the *Theodicy* (written in 1709, seven years before his death) were both devoted to this problem. It is, as well, equally striking that this latter work was the only book length treatise Leibniz saw fit to publish during his life.

Before we take a closer look at Leibniz's views on the problem of evil, we will need to do some stage-setting to help us understand just what sort of problem Leibniz thought evil presented. Open any contemporary introductory textbook and philosophy and it becomes clear that the problem of evil in contemporary philosophy is thought of as an argument for atheism. Since, the atheist contends, God and

evil are incompatible, and evil clearly exists, there is no God. Some, thinking that the claimed incompatibility in the above argument is too strong, argue that even if the existence of God and the existence of evil prove compatible, the existence (or duration, or amount, or distribution) of evil provides us with at least strong evidence that God does not exist.

Framed in this way, the "atheistic problem of evil" invites certain sorts of responses. In particular, it invites the theist to explain how a being that is all-knowing, all-powerful, and all-good can allow evil to exist. And thus, contemporary responses to the problem of evil focus largely on presenting "theodicies" that is, reasons why a perfect being does or might allow evils of the sort (or duration, or amount, or distribution) we find in our world.

When we turn back, however, to the works of those medieval philosophers who treat the problem of evil, the "atheistic problem" is not to be found. Since these figures believed that the arguments of natural theology demonstrated overwhelmingly the existence of God, the problem that evil presented was quite different. For them, the problem was how the existence of evil was compatible with divine moral purity or holiness. Since, they argued, God is the author of everything that exists, and evil is one of the things that exists, God is thereby the author of evil. And if someone is an "author of evil," they are thereby implicated in the evil and thus cannot be morally pure or holy. Thus, God cannot be morally pure nor holy. Let's call this problem of evil the "holiness problem."

Because traditional theists held that God is the "author" or cause of everything in the cosmos in at least three different respects, discussions of the holiness problem often branch off into three correspondingly different directions. First, God is regarded as the *creative cause* of everything in the cosmos. Everything which exists contingently is caused to come into being by the creative activity of God. Second, God is the *conserving cause* of everything that exists. This means that God not only brings into existence every contingent thing that exists, but that every contingent thing which remains in existence does so by God's continuously *maintaining* it in existence. Third, every action by a created substance requires direct divine activity as *concurrent cause*. Thus, every whack of the hammer, every hit of my fingertip on the keyboard, every tug of a magnet on a piece of iron requires not merely that a created substance act, but also that the creator act concurrently with the substance to bring about the particular effect. [For a classic exposition of these various modes of divine causal involvement see St. Thomas Aquinas, *Disputationes de Potentia Dei* , Q.3, a.7, resp.]

Of course, since this traditional picture had God so intimately connected with the workings of the cosmos, the holiness problem seemed all the more intractable. In light of these intimate connections the problem is not just that God created a world in which evil happens to occur, but that God seems to be causally (and thus morally) implicated in, for example, every particular act of murder, every earthquake, and every death caused by plague. As a result, responses to the holiness problem sought to explain not only how God could remain holy in light of having *created* a world such as ours, but also how he could remain holy in light of *conserving* it in existence and *cooperating* in all the events that occur in it.

Since Leibniz lived in between these two eras, eras in which evil was taken to present quite different

problems for the monotheistic philosopher, we are immediately led to wonder which sort of problem he sought to address. Without a doubt Leibniz expends a great deal of effort attempting to solve the holiness problem. But he also frequently takes up something much like the atheistic problem. It would be somewhat deceptive, however, to call it the atheistic problem at this stage in the history of ideas however, since, when raised in this way, evil was seen more an argument for an unorthodox form of theism, than an argument for atheism. Thus, for example, a group of thinkers collectively known as "Socinians" held, among other things, that the existence of evil was not incompatible with God's *existence*, but that it was incompatible with the existence of a God who is *all-knowing*. Thus, Socinians held that God must not be all-knowing, lacking at least knowledge of the future. [For Leibniz's view on the Socinians see Theodicy 364 (H343; G VI 318) et passim. More details on Socinianism can be found in Jolley, c.2, and Maclachlan.]

We might then characterize the problem raised by atheists in our own century and by Socinians, to cite just one example, in the seventeenth century, in more broad terms as what I will call the "underachiever problem." According to the underachiever problem, if there were the sort of being that traditional monotheism describes as God, the existence of this world would represent a vast underachievement on his part; thus there is no such being as this. Atheists take this conclusion to show that there is no God, Socinians take it to show that God is not the sort of being the traditionalist supposes.

Without a doubt, Leibniz is concerned about the underachiever problem, though it is the Socinian, and not the atheist, version of the problem that occupies his attention. The winds of atheism simply had not reached the gale force proportions that it would in succeeding centuries. As a result, this stronger conclusion was not yet taken as a serious, or at least the main, threat presented by evil.

It is important to distinguish these various problems of evil since we cannot understand Leibniz's treatment of evil in a given text until we know what problem it is that he is addressing. Having set the stage in this way, we can know turn to look at Leibniz's solutions to these problems of evil, beginning with the underachiever problem, then turning to the holiness problem.

The Underachiever Problem

The core of Leibniz's solution to the underachiever problem is quite straightforward. Leibniz argues that there is no underachieving involved in creating this world since this world is the best of all possible worlds. Many thinkers have supposed that commitment to the claim that this world is the best of all possible worlds follows straightforwardly from monotheism. Since God is all-powerful, all-knowing, and all-good, there is certainly nothing that can prevent God from creating the best world. And God's goodness further obliges God to create the best world. Thus, the actual world is the best world.

Leibniz's reasoning to his conclusion does not, however, follow this straightforward path since, among other things, it is not clearly cogent as it stands. A number of seventeenth century figures recognized that God would not be obliged to create the best if there were no best world. There might be no such best world if the series of possible worlds formed a continuum of increasingly good worlds ad infinitum. And

if there were no such best world, we cannot fault God for failing to create the best since to do so is as impossible as, say, naming the highest number. There is no such number of course, and likewise no such world. Thus, while God may obliged to create a world which has at least some measure of goodness, he could not be obliged, on this view, to create the best. Thus, God simply chose arbitrarily to bring about one among the range of morally acceptable worlds. [This line of argument was common among certain Jesuit scholastics of the period. For discussions of this see, for example, Ruiz de Montoya, *Commentaria ac Disputationes in primam partem Summae Thologicae S. Thomae. De voluntate Dei et propiis actibus eius* , Lyon 1630, disp. 9 and 10, and Diego Granado, Diego Granado, *Comentarii in primam partem Summae Theologicae S. Thomae* , Pont-a-Mousson, 1624, pp.420-433.]

Leibniz was aware of this argument which denied God's obligation to create the best, but he was firmly committed to rejecting it. The reason for this is that a central principle of Leibniz's system, the Principle of Sufficient Reason, forced him to reject it. According to this principle, any state of affairs must have a reason sufficient which explains why it and not some other state of affairs obtains. When it comes to our world, then, there must be some reason which explains why it, and not some other world, obtains. Clearly, however, there can be no such reason on the view that the goodness of worlds increases ad infinitum. Thus, Leibniz held, there must be no such infinite continuum.

One might be tempted to resist Leibniz's argument here by saying that even on the "infinite continuum of good worlds" view there is something which can play the role of sufficient reason for the fact that this world is actual, namely, *God's decree that this world be actual* . But this, as Leibniz notes, would just push the problem back one step further, since the Principle of Sufficient Reason applies to free choices as much as any other event. Thus, we would have to provide a sufficient reason for God choosing this world over some other on the continuum. And it looks like providing such a sufficient reason is the very thing we cannot do on the infinite continuum of good worlds view. Notice that the sufficient reason cannot be provided by some feature or fact about the world actually chosen. For this would raise the obvious question: why did *this* feature provide the sufficient reason for God's choice? The only possible answers, it appears, would be: a) because God arbitrarily selected that feature as the one he would favor in deciding which world to create, or b) because that feature made that world better than the competitors. But notice that neither of these answers are acceptable. The first is inconsistent with the Principle of Sufficient Reason. The second is incompatible with the hypothesis we are trying to defend here: that there is no "best world."

One might think that declaring this world to be the best possible world is hardly a response to the underachiever problem. In fact, one might think it just provides ammunition for a new underachiever argument along the following lines:

> 1) If God were all-powerful, all-knowing, and all-good, then this world would be the best possible world.
>
> 2) But surely this world is not the best possible world.

3) Thus, God is not all-powerful, all-knowing, and all-good.

Leibniz believed that the evidence that the conclusion of this argument was false was simply overwhelming. So, he is committed to thinking that one of the two premises in this argument is false. Since he himself is committed to the first, he ought to reject the second. And this is what he does.

What reason, Leibniz asks, does the critic have for thinking that 2) is true? When Leibniz addresses this issue, he usually has the critic saying something along the following lines:

> Surely this world is not the best possible world since we can easily conceive of possible worlds that are better. Take some token instance of suffering: the tragic bombing of the Oklahoma City federal building. Surely a world without that event would be better than the actual world. And there is no reason why God couldn't have created the world without that event. Thus, this is not the best possible world. [See Theodicy 118-119 (H 188-191; G VI 168-172)

Leibniz's response to this sort of criticism comes in two stages. First, Leibniz says that while we can think of certain token features of the world that might be better than they are taken individually, we don't know whether or not it is possible to create a better world without those features, since we are never sure of what the connections between the token events and other events in the world might be. If we could improve the token event without otherwise changing the world, we may well have a better world. Unfortunately, we have no way of knowing if changing the token would leave the world otherwise unchanged, or might instead make things, on balance, worse. [See Theodicy 211-214 (H260-2; G VI 244-7) and Grua, p.64f., for examples]

Second, examples such as these are deceptive because they presume that God utilizes standards of world goodness that he does not use. For example, it might presume that a world is only good if each part taken in isolation is good (a standard we have seen Leibniz argues against). Or, it might presume that a world is good only if earthly humans enjoy happiness.

Leibniz argued repeatedly that it was surely too parochial to think that the standard by which the goodness of worlds is to be judged is earthly human happiness. A more reasonable standard, says Leibniz, would be the happiness of all sentient beings. But once we admit this, it may turn out that the amount of unhappiness in the created realm is quite small since for all we know, the sentient beings on the earth might represent a very small percentage of the sentient beings God has created. Here he includes not only preternatural beings such as angels, but the possibility of extra-terrestrial rational beings as well [Theodicy 19 (H134-5; G VI 113-4)].

Leibniz scholars differ about which standard Leibniz thought was applicable in judging the goodness of worlds. Various scholars have defended one or more of the following:

1) The best world is the one which maximizes happiness (i.e., virtue) of rational beings.

> 2) The best world is the one which maximizes the "quantity of essence."
>
> 3) The best world is the one which yields the greatest variety of phenomena governed by the simplest set of laws.

Whether or not Leibniz believed that the maximizing the happiness or virtue of rational beings is one of the standards by which God judges world goodness is a disputed question. [For supporters of this claim see Rutherford, c.3; Blumenfeld, Brown; for detractors see Russell, p.199, Gale]. It is unlikely that Leibniz believed that 1) alone was the true standard of world goodness in light of the fact that he says, in commenting on an argument by Bayle:

> the author is still presupposing that false maxim . . . stating that the happiness of rational creatures is the sole aim of God [Theodicy 120 (H192; G VI 172)]

In part, the dispute over this standard hangs on whether or not 1) is compatible with the more metaphysical standards embodied in 2) and 3), since it is these more metaphysical standards that Leibniz seems to endorse most consistently. In some cases, Leibniz writes as if the standard of happiness is fully compatible with the more metaphysical criteria For example, within a single work, the *Discourse on Metaphysics* , Leibniz entitled section 5 "What the rules of the perfection of divine conduct consist in, and that the simplicity of the ways is in balance with the richness of effects," and entitled section 36: "God is the monarch of the most perfect republic, composed of all minds, and the happiness of this city of God is his principle purpose." Here Leibniz seems to advance both standards 1) and 3) in the same work [For another example, see Riley, p.105 (K X pp.9-10)]. In other places however, he writes as if they compete with one another [See Theodicy 124 (H197-8;G VI 178-9)].

Whatever position one comes to hold on this matter, Leibniz often points to the more metaphysical standards as the ones God utilizes in assessing world goodness. But there is further controversy over exactly which metaphysical standard, 2) or 3), Leibniz endorses. In general, Leibniz holds that God creates the world in order to share his goodness with created things in the most perfect manner possible [Grua 355-6]. Since limited created things can only mirror the divine goodness in limited respects, God creates a variety of things, each of which has an essence that reflects different facets of divine perfection in unique ways. Since this is God's goal in creating, it would be reasonable to think that maximizing the mirroring of divine goodness in creation is the goal that God seeks in creating. And this in fact is one of the standards Leibniz seems to endorse. We might call this the "maximization of essence" standard. Leibniz seemed convinced that the actual world met this standard and that we could therefore find creatures which mirrored the divine perfections in all the sorts of ways that creatures could do this. Thus, there are creatures with bodies and creatures without, creatures with freedom and intelligence and creatures without, creatures with sentience and creatures without, etc. [See, for example, MP pp.75-6 and 138 (G VII 303-4 and 310)].

In some texts, however, Leibniz frames the standard of goodness in what some have taken to be a third distinct way. In these places he argues that the goodness of a world is measured by the ratio of the variety

of phenomena a world contains to the simplicity of the laws which govern it. Here Leibniz emphasizes the fact that the perfection of a world which maximizes the variety of phenomena it contains is enhanced by the simplicity of its laws since this displays the intelligence of the creator who created it.

Various scholars have made the case that one or the other of these two more metaphysical standards represents Leibniz's settled view on the true standard of goodness [Gale, for example]. Others have argued that, in the end, the two standards are not exclusive of one another [See Rutherford, cc.2-3 and Rescher, c.1 for two very different ways of harmonizing 2) and 3)].

Whichever might be the case, if these are the standards by which one thinks God judges the world's goodness, it becomes much more difficult to defend the claim that this is not the best possible world. We can use standard 3) to illustrate. If God were to eliminate the Oklahoma City bombing, what would be required to do so? There are presumably a number of ways in which this might be done. The most obvious would involve miraculous intervention somewhere in the chain of events leading up to the explosion. God might miraculously prevent the explosives from detonating, or might make the entire truck and its contents vanish. But any sort of miraculous intervention will involve making the laws governing the phenomena more complex. As a result, Leibniz, and others who share this view of what world goodness consists in, such as Malebranche, think that miraculous intervention is generally repugnant and would require vastly outweighing goods in order for them to be permissible. [See Theodicy 129 (H192-3; G VI 182)].

In any case, Leibniz holds that we are simply unable to know how changing certain events would change the worlds capacity to meet the standards of goodness described in 2) and 3). As a result, we can never, with any confidence, make the claim that this world is not as good, all things considered, as some other world we might try to imagine. According to Leibniz, then, the underachiever problem cannot get off the ground unless the critic is able to defend the claim that this world is not the best possible world. While we might think such a defense would be easy to mount, our inability to know how changing certain events in the world would affect other events, and our inability to know how such changes would affect the true overall goodness of the world makes such a defense impossible for us.

The Holiness Problem

Far less scholarly attention has been devoted to Leibniz's treatment of the holiness problem, if only because this way of seeing the problem has only recently been recognized by Leibniz scholars. As mentioned above, the main problem here is that God's character seems to be stained by evil since God knowingly and causally contributes to the existence of everything in the world, and evil is one of those things. [For two recent treatments see Sleigh (1997) and Murray (1998)]
The standard solution adopted by medieval thinkers was to deny something the above argument affirms, namely, that evil is a "something." Evil, they claimed was not a positive reality, but a "privation" or "lack." As a result, evil has no more reality than the hole in the center of a donut. Making a donut does not require putting together two components, the cake and the hole. Instead, the cake is all that there is to the donut. The hole is just "privation of cake." Thus, it would be silly to say that making the donut

requires something to cause the cake, and then something to cause the hole. Causing the cake causes the hole as a "by-product." Thus, we need not assume any additional cause for the hole beyond that assumed for the causing of the cake.

The upshot of our pastry analogy is simply this: since evil, like the hole, is merely a privation, it needs no cause on its own (or as the medievals, and Leibniz, liked to say, it needs no "cause *per se* "). Thus, God is not a "knowing causal contributor to evil" since evil *per se* has no cause at all. But since God does not contribute to evil, God cannot be implicated in the evil. Thus, the holiness problem evaporates.

Early in his career Leibniz, like many seventeenth century figures, scoffed at this solution. In a short piece entitled "The Author of Sin," Leibniz explains why he thinks the privation response to the holiness problem fails. Since, Leibniz argues, God is the author of all that is real and positive in the world, God is, by extension, "author" of all of its privations, "It is a manifest illusion to hold that God is not the author of sin because there is no such thing as an author of a privation, even though he can be called the author of everything which is real and positive in the sinful act." [A.6.3.150]

The reason, says Leibniz, can be gleaned from an example. Consider a painter who creates two paintings, one a small scale version of the other. The details of the pictures are identical in every respect, only the scale is different. It would be absurd, Leibniz remarks,

> . . . to say that the painter is the author of all that is real in the two paintings, without however being the author of what is lacking or the disproportion between the larger and the smaller painting. . . . In effect, what is lacking is nothing more than a simple result of an infallible consequence of that which is positive, without any need for a distinct author [of that which is lacking] [A.6.3.151]

Thus, even if it is true that evil is a privation, this does not have as a consequence that God is not the author of sin. Since what is positively willed by God is a sufficient condition for the evil state of affairs obtaining, willing what is positive makes God the author of that which is privative as well [A similar early critique is found at A.6.3.544].

Thus, in his early years Leibniz looks to develop a different strategy. In the *Philosopher's Confession* , his most significant treatise on evil aside for the *Theodicy* , Leibniz claims that God wills everything in the world, though his will with respect to goods in the world is *decretory* , while his will respect to evils is merely *permissive* . Further, Leibniz argues, permissive willing of evils is morally permissible as long as the permitting the evil is a necessary condition for meeting one's outweighing obligations.

It is important to note here that Leibniz does not think that the permission of evil is morally permissible because allowing the evil *brings about a greater good not otherwise attainable* . To put the matter this way leaves God, according to Leibniz, in the position of violating the Biblical injunction "not to do evil that good may come" [Causa Dei 36 (S 121; G VI 444)]. Thus, Leibniz casts permission in such a way that the resultant evil is a *necessary consequence of God's performing his duty* (namely, to create the best

world). As a result, Leibniz characterizes (morally permissible) permission as follows:

> P permits E iff:
>
> 1. P fails to will that E
> 2. P fails to will that not-E
> 3. P brings it about that state of affairs S obtains by willing that S obtains
> 4. If S obtains then E obtains
> 5. P knows that 4)
> 6. P believes that it is P's duty to will S and that the good of performing one's duty outweighs the evil entailed by E's obtaining
>
> [This account is distilled from A.6.3.129-131]

This, Leibniz believes, resolves any holiness problem that might arise in so far as God is considered as creator of the cosmos. However, from the time that Leibniz composed the *Philosopher's Confession* until at least the mid-1680's, Leibniz grew increasingly concerned that a tension might arise in his account when applied to the holiness problem considered as a problem for *concurrence* . Recall that traditional theists held that God was not only creator and conserver of all created things, but that in addition God acted as concurrent cause of every act of a created substance.

There were heated debates in the sixteenth and seventeenth centuries concerning the nature of divine concurrence. And much of the dispute focused on the way in which God concurred with the free acts of creatures. This was an especially pressing problem for the obvious reason that positing too close of a connection between God and the creature in cases where moral evils are committed runs the risk of implicating God in the evil, thus raising the holiness problem all over again. This debate often focused on a certain type of proposition and on what made this type of proposition true. The type of proposition in question are called "conditional future contingents" and they are propositions of the form:

> If agent, S, were in circumstances, C, and time, t, S would freely chose to f.

These propositions were particularly important in discussions of philosophical theology in the sixteenth and seventeenth centuries because God's having knowledge of token propositions of this type was regarded as essential for God exercising providential control over the activities of free beings in creation. In order to be able to providentially superintend the activities of free beings in the created world, God must know how each such being will choose to act in each circumstance in which they will be found. If God did not know what Eve would choose when confronted by the serpent, or what I would choose when confronted with the tuna sandwich, God could not know, in advance, how the order of events would unfold in the universe he deigns to create.

But *how* does God know whether or not a token proposition of this type is true? Speaking generally, disputants in this period held that there were only two answers. God either knows that a token

proposition of this type is true because God wills it to be true, or because something independent of God's will makes it true, and God, being omniscient, thereby knows it. In keeping with recent tradition, we will call the first view the "postvolitional view" (since the truth of the proposition is determined only "after" God wills it) and the latter view the "prevolitional" view (since the propositions has truth independently of what God wills). In his early years, Leibniz seemed inclined to adopt the postvolitional answer. So, take the token proposition:

> If Peter were accused of consorting with Christ immediately after the crucifixion, Peter would deny Christ.

The early Leibniz would have held that this proposition, and others of this type, is true because God decrees that it is, that is, he decrees that Peter will deny under these circumstances [See C 26-7 and Grua 312-3]. Furthermore, those who held this view generally held that it was through divine concurrence that God makes the proposition true in the actual world. So, by causally influencing Peter at the moment of decision, God brings it about that Peter denies in these circumstances.

This view obviously faces a number of difficulties. For our purposes, the most pressing one is that it seems to undercut Leibniz's solution to the holiness problem based on permission. For if the above proposition is true because God wills that it is, then it appears that God wills that Peter sin, and if he wills that Peter sin, he cannot merely permit it, in light of condition 1) of the definition of permission. As a result, it appears that Leibniz must surrender his initial answer to the question "what makes conditional future contingents true?" and adopt the alternative answer.

There are troubles here as well, however. What does it mean to say that the truth of the proposition is determined independently of God's will? Defenders of this view usually held that the human will was entirely free from any determining cause whatsoever. In choosing one alternative over another, nothing could be regarded as "determining" or "causing" the choice, else it would not be free. Thus, for those who defended this view, the answer to the question "what makes conditional future contingents true?" must be "nothing." For if something *made* it true, then *that thing* would be determining the choice, and the choice could then not be free.

In light of his commitment to the Principle of Sufficient Reason, however, Leibniz could not support such a view. Does Leibniz, then, have answer to this question that will rescue him from the holiness problem? Scholars disagree about this. Some have held that Leibniz is obliged to hold the postvolitional view in spite of the troubles it raises for him [See Davidson, Sleigh (1995)(1997)]. Others have held that Leibniz tried to forge a third alternative in order to avoid this seemingly intractable dilemma [See Murray (1998)]. I will close with a look at this latter suggestion.

According to Leibniz, free choice in humans is brought about through the activity of the human intellect and the human will working in concert with one another. The intellect deliberates about alternatives and selects the one that it perceives to be the best all things considered. The intellect then represents this alternative to the will as the one that is best to pursue. The will, which for Leibniz is a faculty

characterized by "appetite for the good," then chooses that alternative which is represented to it as containing the most good. [*Theodicy* , 311 (H314; G VI 300-1).]

On this picture, it appears that there are two ways in which I might exercise "control" over my acts of will. First, I might be able to control what appears to me to be the best all things considered. That is, I might control the process of deliberation. Second, I might be able to control the will's choosing that alternative which is presented to it by the intellect as representing the greatest good at that time. There are places where Leibniz seems to accept each of these possibilities. In some passages he argues that by engaging in moral therapy of certain sorts I can control which things appear to me good, and thus control the outcome of my deliberations. In other passages, he seems to say that while the will does "infallibly" choose that which the intellect deems to be best, the will retains the power to resist since the intellect does not "cause" the will to choose as it does. [Concerning the first strategy see, for example, *Reflections on Hobbes* , 5 (H396-7; G VI 391-1). For more on this aspect of Leibniz's view of freedom see Seidler (1985). Concerning the second strategy see, for example *Theodicy* 282 (H298-300; G VI 284-5).]

Difficulties arise in following through on either suggestion. Consider the first. How might I go about engaging in "moral therapy"? First, I would have to choose to do something to begin to bring about a change in how I see things. But of course, I can only make a choice to do this if I first deliberate about it and see that making this change is the best thing to be done. So, did I have control over this process of "coming to see that a change is the best thing to be done"? It looks as if I have control here only if I have control over the actions that led to my coming to see things this way in the first place. Do I have control over these actions? If the answer is yes, it is only because I had control over my prior deliberations, and it looks as if this will lead us back in the chain of explanation to certain very early formative stages of my moral and intellectual life, stages over which it is hard to believe I had any control. Thus, this route seems hard to sustain.

Let us consider the second alternative then, according to which I have control because the will is never "causally determined" to choose that which the intellect deems to be best in those circumstances. Leibniz held that the will was not causally determined in the act of choice but merely "morally necessitated." Scholars disagree about exactly how we should understand this phrase. Some think it just means "causally necessitated." But if this is right, it appears that God, who establishes the laws of nature, determines how creatures act, and this leads us back to the suggestion that Leibniz was a postvolitionalist in these matters. As we noted above, this is a troubling position for Leibniz to adopt since it seems to undermine his response to the holiness problem. [For various positions on the nature of "moral necessity" see Adams, pp.21-2, Sleigh (1998), Murray (1995), pp.95-102, and (1996), esp. Section IV].

Others have held that moral necessity is a philosophical novelty, invented to explain the unique relationship between intellect and will. On this view, the will infallibly follows the outcome of deliberation, without being causally necessitated by it. Leibniz sometimes hints at this reading such as in the following example he borrows from Pierre Nicole:

> It is considered impossible that a wise and serious magistrate, who has not taken leave of

> his senses, should publicly commit some outrageous action, as it would be, for instance, to run about the streets in order to make people laugh [*Theodicy* 282 (H299; G VI 284)]

Here, the wise magistrate is not causally determined to refrain from streaking to make people laugh. Instead, he just considers it so unseemly that "he can't bring himself to do it." Something about his psychological constitution prevents him from seeing this as a live option for him, even though there is surely some sense in which he *could do it* nonetheless.

If we allow Leibniz to locate control over actions in the fact that the will is only morally necessitated by the intellect, is there a way for him to avoid the postvolitional/ prevolitional dilemma discussed earlier? It is not clear. One would have to say that the will infallibly choosing in accordance with the deliverances of deliberation is a fact, the truth of which is *independent of God's will* , while also saying that the deliverances of deliberation provide a *sufficient reason* for the wills choice. If this can be done, Leibniz may have a way of avoiding the difficulty posed by conditional future contingents.

However we might think these questions should be resolved, Leibniz himself seemed to think that the prevolitional route is the one to take. This seems relatively clear in light of the fact that from the mid-1680's onward, Leibniz consistently uses language which implies that God merely "discovers" how free human beings would act if created. [See, for example, Grua 227, 232, A&G 32, and C 23-4]. It is not something that God makes so. In these later works, Leibniz speaks of these truths about how human beings will act as "limitations" which prevent God from making them, and the world that contains them, more perfect. In the end, it is these limitations, Leibniz argues, which prevents there being a better world than the actual one [On the notion of "limitations" see A&G 60-2, 11,*Theodicy* 20 (H86-7; G VI 114-5), *Causa Dei* 69-71 (S 128-30; 457-8)]. If this is right, then we might think that the permission strategy will work as a solution to the holiness problem both when it comes to considering God as creator and as concurrent cause of all effects in the cosmos.

It is interesting to note that, for reasons still not clear, Leibniz comes to favor, in later life, the scholastic "privation" view he rejected in his youth. [See, for example, *Theodicy* 20, 30, 153 (respectively, H86-7, 91-2, 219-20; G VI 114-5, 119-20, 201]. The role that this revived position is supposed to play for Leibniz in his later writings awaits further scholarly investigation.

The issues that arise in thinking about Leibniz's views on the problem of evil are vigorously discussed among contemporary Leibniz scholars and philosophers of religion. It is clear that his thinking about evil impacts, and is impacted by, some of the most centrally important components of his philosophical and theological system. Because of this, scholars have found inquiry into this feature of Leibniz's philosophy a particularly rewarding route for discovering some of the deeper motivations standing behind some of Leibniz's more puzzling philosophical views. Yet because this topic represents such an active area of current Leibniz scholarship, it is clear that any conclusions we might draw on his views are, for now, preliminary and subject to revision.

Bibliography

A	*Sämtliche Schriften und Briefe* . Darmstadt and Berlin: Berlin Academy, 1923-. Cited by series, volume, and page.
A&G	*Philosophical Essays*. Roger Ariew and Daniel Garber (eds. and trans.), Indianapolis: Hackett, 1989.
Adams	Robert Adams, *Leibniz: Determinist, Theist, Idealist* . Oxford: Oxford University Press, 1995.
Blumenfeld	David Blumenfeld, "Perfection and Happiness in the Best Possible World." In *The Cambridge Companion to Leibniz* . Nicholas Jolley (ed.), Cambridge: Cambridge University Press, 1994.
Brown	Gregory Brown, "Leibniz's Theodicy and the Confluence of Worldly Goods." *Journal of the History of Philosophy* 26:571-91.
C	Louis Couturat (ed.), *Opuscules et Fragments Inédits de Leibniz.* Hildesheim: Georg Olms, 1966.
Davidson	Jack Davidson, "Untying the Knot: Leibniz's on God's Knowledge of Future Free Contingents," *History of Philosophy Quarterly* , 13.
Gale	George Gale, "On What God Chose: Perfection and God's Freedom." *Studia Leibnitiana* , 8:69-87.
Grua	Gaston Grua (ed.), *Textes Inédits* . Paris: Presses Universitaires de France, 1948
Jolley	Nicholas Jolley, *Leibniz and Locke: A Study of the New Essays in Human Understanding* . Oxford: Clarendon Press, 1984.
Klopp	Onno Klopp (ed.), *Die Werke von Leibniz* . *Reihe I: Historisch-politische und staatswissenschaftliche Schriften.* Hannover: Klindworth, 1864-84.
MacLachlan	H.J. MacLachlan, *Socinianism in Seventeenth-century England* . Oxford: Oxford University Press, 1951.
MP	Mary Morris and G.H.R. Parkinson (eds. and trans.), *Leibniz-Philosophical Writings* , London: J.M. Dent and Sons, 1973.
Murray	Michael J. Murray: 1995. "Leibniz on Divine Knowledge of Conditional Future Contingents and Human Freedom," *Philosophy and Phenomenological Research* , 55: 75-108.
1996.	"Intellect, Will, and Freedom: Leibniz and His Precursors," *The Leibniz Society Review* , 6: 25-60.
1998.	"Evil, Divine Holiness, and Future Contingents in Leibniz." forthcoming.
Rescher	Nicholas Rescher, *Leibniz's Metaphysics of Nature* . Dordrecht: D. Reidel, 1981.
Riley	G.W. Leibniz, *Political Writings* . Patrick Riley (ed. and trans.), Cambridge: Cambridge University Press, 1988.

Rutherford	Donald Rutherford, *Leibniz and the Rational Order of Nature.* Cambridge: Cambridge University Press, 1995.
Schrecker	Paul Schrecker and Anne Martin Schrecker (eds. and trans.), *Leibniz: Monadology and Other Philosophical Essays* . Indianapolis: Bobbs-Merrill, 1965.
Sleigh	Robert C. Sleigh: 1995. "Leibniz on Divine Foreknowledge." *Faith and Philosophy* , 11.
1997.	"Leibniz's First Theodicy." forthcoming
1998.	"Leibniz on Freedom." forthcoming

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Leibniz, Gottfried Wilhelm

Michael J. Murray
m_murray@acad.fandm.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 4, 1998
Content last modified: January 4, 1998

Naturalism in Legal Philosophy

The “naturalistic turn” that has swept so many areas of philosophy over the past three decades has also had an impact, especially recently, in legal philosophy. Methodological naturalists (M-naturalists) view philosophy as continuous with empirical inquiry in the sciences. Some M-naturalists want to *replace* conceptual and justificatory theories with empirical and descriptive theories; they take their inspiration from more-or-less Quinean arguments against conceptual analysis and foundationalist programs. Other M-naturalists retain the *normative* and regulative ambitions of traditional philosophy, but emphasize that it is an *empirical* question what normative advice is actually useable and effective for creatures like us. Some M-naturalists are also *substantive* naturalists (S-naturalists). Ontological S-naturalism is the view that there exist only *natural* or *physical* things; semantic S-naturalism is the view that a suitable philosophical analysis of any concept must show it to be amenable to empirical inquiry. Each of these varities of naturalism has applications in legal philosophy. *Replacement* forms of M-naturalism hold that: (1) conceptual analysis of the concept of law should be replaced by reliance on the best social scientific explanations of legal phenomena, and (2) normative theories of adjudication should be replaced by empirical theories. These views are associated with American Legal Realism and Brian Leiter's reinterpretation of Realism. Normative M-naturalists, by contrast, inspired and led by Alvin Goldman, seek to bring empirical results to bear on philosophical and foundational questions about adjudication, the legal rules of evidence and discovery, the adversarial process, and so forth. An older form of S-naturalism in legal philosophy, associated with Scandinavian Legal Realism, seeks a reduction of legal concepts to behavioral and psychological categories. More recent forms of S-naturalism, associated with a revival of a kind of natural law theory defended by David Brink and Michael Moore (among others), applies the “new” or “causal” theory of reference to questions of legal interpretation, including the interpretation of moral concepts as they figure in legal rules.

---

1. Varieties of Naturalism: Methodological and Substantive

Different philosophical doctrines travel under the heading of “naturalism.” We can usefully distinguish two broad and important categories: *methodological* (or M-naturalism) and *substantive* (or S-naturalism) (Leiter 1998; cf. Railton 1990 and Goldman 1994). Naturalism in philosophy is most often a *methodological* view to the effect that philosophical theorizing should be continuous with empirical inquiry in the sciences. Such a view need not presuppose a solution to the so-called “demarcation problem”—i.e., the problem of what demarcates genuine science from pseudo-science—as long as there remain clear, paradigmatic cases of successful sciences. Some M-naturalists want “continuity with” only the *hard* or *physical* sciences (Hard M-naturalists); others seek “continuity with” any successful science, natural or social (Soft M-naturalists). Soft M-naturalism is probably the dominant strand in philosophy today.

For M-naturalists, “continuity with” the sciences includes, in the first instance, the Quinean repudiation of a “first philosophy”, a philosophical solution to problems that proceeds entirely *a priori*, that is, without the benefit of empirical evidence. (Most M-naturalists do not go as far as Quine, however, in repudiating *any* role for *a priori* conceptual analysis: see, e.g., Goldman 1986 for a more typical M-naturalist approach). Beyond hostility to methods that are exclusively *a priori*, M-naturalists require continuity with the sciences in two more precise senses, what we may call “Results Continuity” and “Methods Continuity”.

Results Continuity requires that the substantive claims of philosophical theories be supported or justified by the results of the sciences. Epistemologists like Goldman look to the results of psychology and cognitive science to find out how the human cognitive apparatus really works; only with that information in hand can the epistemologist construct norms for how humans *ought* to form beliefs (Goldman 1978, 1986). Moral philosophers like Gibbard and Railton, despite profound substantive disagreements, both think that a satisfactory account of morality's nature and function must be supported by the results of evolutionary biology, our best going theory for how we got to be the way we are (Gibbard 1990, Railton 1986). A philosophical account of morality that explains its nature and function in ways that would be impossible according to evolutionary theory would not, by naturalistic scruples, be an acceptable philosophical theory.

“Methods Continuity,” by contrast, demands only that philosophical theories emulate the “methods” of inquiry of successful sciences. “Methods” should be construed broadly here to encompass not only, say, the experimental method, but also the styles of explanation (e.g., via appeal to causes that determine, *ceteris paribus*, their effects) employed in the sciences. Such a view does not presuppose the methodological *unity* of the various sciences, only that successful sciences have some methodological *uniqueness*, even if this is not exactly the same across all the sciences. Historically, Methods Continuity has constituted the most important type of naturalism in philosophy, found in writers like Spinoza, Hume, and Nietzsche. (However, unlike the contemporary M-naturalists who draw on the actual *results*

of established sciences, many historical M-naturalists drawn to Methods Continuity simply try to *emulate* a scientific way of understanding the world in developing their philosophical theories.)

M-naturalists, then, construct philosophical theories that are continuous with the sciences either in virtue of their dependence upon the actual results of scientific method in different domains or in virtue of their employment and emulation of distinctively scientific ways of looking at and explaining things. We may still distinguish between two different branches of M-naturalism, represented best by Quine, on the hand, and Goldman, on the other. The former we will call Replacement Naturalism, the latter Normative Naturalism. Goldman's paradigm of Normative Naturalism has dominated philosophical research in the area (see Kitcher 1992), though it is Quine's notion of Replacement Naturalism that proves useful for understanding the American Legal Realists as naturalists in legal philosophy (Leiter 1997). Since both Replacement and Normative Naturalists share the *methodological* commitment distinctive of naturalism—to make philosophical theorizing continuous with and dependent upon scientific theorizing—the difference must be located elsewhere: not in methodology, but in goal. According to Replacement Naturalists, the goal of theorizing is description or explanation; to that end, conceptual and justificatory theories are to be *replaced* by empirical and descriptive theories. According to Normative Naturalists, the goal is regulation of practice through the promulgation of norms or standards. Of course, traditional epistemology also shares the regulative goal of Normative Naturalism; what distinguishes the Normative Naturalist is simply the *methods* employed to realize this objective (cf. Goldman 1986, pp. 6-9).

Many naturalists go beyond methodological naturalism, however, and embrace a *substantive* doctrine. S-naturalism in philosophy is either the (ontological) view that the only things that exist are *natural* or *physical* things; or the (semantic) view that a suitable philosophical analysis of any concept must show it to be amenable to empirical inquiry. In the ontological sense, S-naturalism is often taken to entail physicalism, the doctrine that only those properties picked out by the laws of the physical sciences are real. In the semantic sense, S-naturalism is just the view that predicates must be analyzable in terms that admit of empirical inquiry: so, e.g., a semantic S-naturalist might claim that “morally good” can be analyzed in terms of characteristics like “maximizing human well-being” that admit of empirical inquiry by psychology and physiology (assuming that well-being is a complex psycho-physical state).

Many philosophers are drawn to some type of S-naturalism in virtue of their M-naturalism: being a philosophical naturalist in the methodological sense sometimes leads a philosopher to think that the best philosophical account of some concept or domain will be in terms that are substantively naturalistic. (So, e.g., while both Gibbard (1990) and Railton (1986) are M-naturalists—in the sense that both seek to locate morality in a world-view circumscribed by the results of evolutionary biology—only Railton is an S-naturalist since only he thinks that the account of morality that makes it square best with the results of evolutionary biology is one which interprets moral properties as being reducible to “natural” properties amenable to inquiry in the sciences.) It is important to notice that a commitment to Methodological Naturalism does *not* entail any substantive conclusions, however: methodologically, it is an open question whether the best philosophical account of morality or mentality or law must be in substantively naturalistic terms.

The varieties of philosophical naturalism map on to a variety of naturalistic approaches in legal philosophy. The most radical version of M-naturalism, Replacement Naturalism, is articulated and defended in Leiter (2001b, 2001c) and, arguably, in the American Legal Realists (Llewellyn 1930; Moore and Callahan 1943; Leiter 1997). The less radical form of M-naturalism, Normative Naturalism, is exemplified in epistemology by Goldman (1978, 1986), as noted, but its implications for jurisprudence and law have, to date, only been partly developed (Allen & Leiter 2001; Goldman 1999; Leiter 1998, 2001d; Talbott & Goldman 1998). One form of S-naturalism is well-represented by Scandinavian Realism of mid-century (Hägerström 1953, Ross 1958), although their S-naturalism, conjoined as it is with skepticism about normative concepts, ends up requiring an implausible semantics of legal concepts. More recently, S-naturalism without normative skepticism has been defended by contemporary moral realists and natural law theorists like Brink (1988, 1989, 2001) and Moore (1985, 1992b).

2. Replacement Naturalism I: Against Conceptual Analysis

Replacement Naturalism holds that conceptual and justificatory theories—the traditional fare of philosophy—are to be *replaced* by empirical and descriptive theories. There are two kinds of argumentative routes to Replacement Naturalism, both due to Quine: the first arises from doubts about the analytic-synthetic distinction (Quine 1951); the second from doubts about foundationalism (Quine 1969). Here we consider the former.

Philosophers have long thought that some truths were *necessary* while others were *contingent*; in the twentieth-century, under the influence of logical positivism, this was taken to be the distinction between those statements that were “true in virtue of meaning” (hence *necessarily* true) and those that were “true in virtue of fact” (hence only *contingently* true). The former “analytic” truths were the proper domain of philosophy; the latter “synthetic” truths the proper domain of empirical science. Quine argued that the distinction could not be sustained: all statements are, in principle, answerable to experience, and, conversely, all statements can be maintained in the face of experience as long as we adjust other parts of our picture of the world. So there is no real distinction between claims that are “true in virtue of meaning” and “true in virtue of facts,” or between ‘necessary’ and ‘contingent’ truths; there is simply the socio-historical fact that, at any given point in the history of inquiry, there are some statements we are unlikely to give up in the face of recalcitrant empirical evidence, and others that we are quite willing to give up when empirical evidence conflicts.

Without a domain of analytic truths—truths that are *a priori* and hold in virtue of meaning—it becomes unclear what special domain of expertise for philosophical reflection remains. If all claims are, in principle, revisable in light of empirical evidence, why not let all questions fall to empirical science? Philosophy would be out of business, except perhaps as the abstract, reflective branch of empirical science. This Quinean attack has consequences for the traditional philosophical business of conceptual analysis, since on the dominant view from Plato through Carnap “every analysis of a concept is inextricably bound to a collection of purported analyticities” (Laurence & Margolis 1999, p. 18). (Even

the more recent “posession-condition” account of concepts in Peacocke (1992) requires that it be *analytic* that certain inferential transitions are privileged by a particular concept.) The conclusion Replacement Naturalists draw from the preceding is that since any claim of conceptual analysis is vulnerable to the demands of *a posteriori* (i.e., empirical) theory construction, philosophy must proceed in tandem with empirical science, not as the arbiter of its claims, but as a reflective attempt at synoptic clarity about the state of empirical knowledge.

Many resist this conclusion. According to one proponent, a conceptual analysis proceeds “by appeal to what seems to us most obvious and central about [the concept in question]…as revealed by our intuitions about possible cases” (Jackson 1998, p. 31). “[T]he general coincidence in intuitive responses [to possible cases] reveals something about the folk theory of [the concept in question]” (Jackson, 1998, p. 32). The question that plagues conceptual analysis, post-Quine, is what kind of *knowledge* such a procedure actually yields? Why should ordinary intuitions about the extension of a concept be deemed reliable or informative? Why think the “folk” are right?

The track record of *a priori* methods like appeal to intuitions and conceptual analysis is not a promising one (e.g., Harman 1994; Hintikka 1999). Kant, for example, took it to be *a priori* that space necessarily had the structure described by Euclidean geometry; subsequent physics showed his intuitions to be mistaken. The moral naturalists would draw from this track record of *a priori* philosophy is well-expressed by Cummins (1999, pp. 117-18):

> We can give up on intuitions about the nature of space and time and ask instead what sort of beasts space and time must be if current physical theory is to be true and explanatory. We can give up on intuitions about representational content and ask instead what representation must be if current cognitive theory is to be true and explanatory.

For the Replacement Naturalist, in short, the only sound reason to prefer a proposed conceptual analysis is not because it seems intuitively obvious, but because it earns its place by figuring in successful *a posteriori* theories of the world. Philosophy *cum* conceptual analysis and intuition-pumping should be abandoned in favor of empirical science; philosophy is simply the more abstract and reflective part of empirical science and lays claim to no distinctive methods or body of knowledge.

Defenders of conceptual analysis, it is true, commonly proclaim the modesty of their ambitions; indeed, Jackson specifically chastises conceptual analysis in its “immodest role,” namely when “it gives intuitions…too big a place in determining what the world is like” (1998, pp. 43-44): “There is nothing sacrosanct about folk theory. It has served us well but not so well that it would be irrational to make changes to it in the light of reflection on exactly what it involves, and in the light of one or another empirical discovery about us and our world” (Jackson 1998, p. 44). The question is, having conceded this much, what remains? Conceptual analysis, as Jackson conceives it, becomes hard to distinguish from banal descriptive sociology of the Gallup-poll variety. (Jackson even says he advocates, when necessary, “doing serious opinion polls on people's responses to various cases” (1998, p. 36).) Such a procedure might deliver some insight in to what some people, at some time, in some place, think about “mind” or

"law" or "justice," but Replacement Naturalists wonder what philosophical import any of this data could have, since it is bounded not simply by time and place, but also ignorance.

How might Replacement Naturalism in legal philosophy, motivated by these Quinean doubts about conceptual analysis and intuitions, proceed? One possibility is suggested in Leiter (2001b, 2001c), which invokes the following example. Raz (1985) has offered an influential conceptual argument against Soft Positivism's claim that there is no constraint on the content of a rule of recognition beyond the fact that it is a social rule: its existence-conditions are given by the actual practice of officials in deciding disputes, but what criteria of legality officials appeal to (i.e., the *content* of the rule of recognition) is dependent upon whatever the conventional practice of officials in that society happens to be. Raz offers an analysis of the concept of authority to show that Soft Positivism is incompatible, even in principle, with the law's possessing the authority it claims to possess. According to Raz, it is a non-normative prerequisite for a claim to authority that it be possible to identify the authority's directive without reference to the underlying "dependent" reasons for that directive. This is a *prerequisite* for authority because what distinguishes a (practical) authority, on Raz's "service" conception, is that its directives preempt consideration of the underlying reasons for what we ought to do, and in so doing actually makes it more likely that we will do what we really ought to do. Authoritative reasons are claimed to be exclusionary reasons, excluding from consideration those dependent reasons (including, importantly, moral reasons) on which the authoritative directive rests. Soft Positivism, then, undermines the possibility of the rule of recognition claiming authority, since for Soft Positivism a rule of recognition can, in principle, employ dependent reasons as criteria of legal validity: to identify, then, the directives about legal validity of such a rule of recognition would be impossible without recourse to precisely the dependent reasons the rule was supposed to preempt.

One line of response to Raz has appealed to contrary intuitions about the concept of authority. Perry (1987), for example, argues that authoritative reasons need not be exclusionary in Raz's sense; it suffices, Perry says, that they simply be "weightier" than other reasons. Some commentator's intuitions line up with Raz's (Leiter 2001b), others with Perry (Waluchow 1994). Now, of course, the Quinean worries about conceptual analysis hold even in cases where everyone's intuitions about a concept coincide; but when they do not coincide, the inadequacies of the philosophical "methods" at hand seem especially acute. Some proponents of traditional methods of legal philosophy object that "the mere fact that there is disagreement about what the conceptual truths of law are…does not mean that conceptual analysis of law is fruitless. If that were the case, we should have to conclude the same about philosophy generally" (Coleman 2001, p. 211 n. 38). Unfortunately, this *reductio* response depends on a conclusion that the Replacement Naturalist is, in fact, prepared to embrace—and not because the Replacement Naturalist naively believes empirical methods "would…put an end to disputes about the nature of law or anything else" (Coleman 2001, p. 211 n.38). The worry, rather, is that intuitions about concepts enjoy no privileged epistemic status, while claims in empirical science do. Even if empirical science does not resolve these disputes, it at least delineates criteria with epistemic weight for adjudicating them. The crucial question, then, becomes whether our best empirical science requires drawing the conceptual lines one way rather than another.

The leading social scientific accounts of judicial decision-making—both the informal ones (Pritchett

1949, Powe 2000) and the formal ones (Segal & Spaeth 1993)—have two striking features in this regard: first, they all aim to account for the relative causal contribution of “law” and non-law factors (e.g., political ideologies or “attitudes”) to judicial decisions; and second, they demarcate “law” from non-law factors in typical Hard Positivist terms, i.e., they generally treat as “law” only pedigreed norms, like legislative enactments and prior holdings of courts (as well, sometimes, as the interpretive methods applied to these kinds of legal sources: see the treatment of the “legal model” in Segal & Spaeth 1993, pp. 33-53). Supposing that these models are ultimately vindicated empirically—and not just for American courts—this would give the Replacement Naturalist reason to abandon any *a priori*, intuitive confidence we had about the concept of law that conflicted with Hard Positivism—just as the role of non-Euclidean geometry in parts of physics has led *everyone* to repudiate Kant's *a priori* intuitive confidence about the Euclidean structure of space. If social science cuts the causal joints of the legal world in Hard Positivist terms, the Replacement Naturalist argues, that is a compelling reason to work with that concept of law as against its competitors.

Proponents of conceptual analysis, by contrast, are skeptical that the explanatory premises of empirical social scientists give us any reason to prefer one concept of law to another. Notice, of course, that an analogous skepticism is available to the diehard Euclidean: after all, non-Euclidean geometries are famously non-intuitive and hard to grasp. But Kantians recognize that such a response would be unmotivated: if non-Euclidean geometry does explanatory work within successful physical theory, then the right conclusion to draw is that our intuitions about the structure of space need tutoring to keep pace with empirical knowledge. So, too, the analogous question for the natural lawyer or Soft Positivist is: why think your intuitions are epistemically privileged as opposed to simply untutored by the best empirical science?

The skeptic, however, might refine the challenge as follows: “It's not,” she might say, “that I insist on sticking to my intuitions, empirical science be damned. Rather, I do not see why the empirical science at issue *needs* to take sides on a dispute about the concept of law.” Of course, it is clear that the empirical social science at issue *does* draw the line between legal and non-legal norms based on pedigree criteria, but the question is whether it *needs* to: the natural lawyer could agree with the social scientists that, e.g., moral and political considerations determine judicial decisions, but contest the assumption that these considerations are not themselves legally binding.

The difficulty, of course, is that the candidate non-law explanatory factors at issue (e.g., an ideological commitment to the platforms of the Republican Party) are not plausible candidates for being legal norms, on any existing theory of the concept of law. Moreover, there are good reasons why social science treats the explanatory factors at issue as non-legal: for example, the moral and political attitudes invoked to explain decisions do not, for example, appear explicitly in the text of the decisions, or in the explicit rationales for the decisions; they are often *hidden* and *hard to detect*, which make them quite unlike any of the paradigm instances of legal norms, like statutory provisions or precedent. Finally, the legal/non-legal demarcation in empirical social science usually reflects more general explanatory premises about the psycho-social factors that account for behavior, well beyond the realm of the legal. The motivation for demarcating the legal/non-legal in essentially Hard Positivist terms is, for most social scientists, to effect an explanatory unification of legal phenomena with other political and social behavior.

Yet the very talk of “legal phenomena” may invite a different kind of objection to the proposed naturalization of jurisprudential questions. For how is it, one might wonder, that the social scientist knows these are *legal* phenomena he is explaining, and not phenomena of some other kind? Does that not already presuppose an analysis of the concept of law? (Cf. Coleman 2001, pp. 213-214.) It is not obvious, though, why a shared language and dictionaries won't suffice to get empirical science off the ground; it is not that empirical science *needs* conceptual analysis to tell his explanatory story, it's rather that *after the fact* the philosopher may be able to offer some greater reflective clarity about the concepts invoked in the explanatory story. Conceptual philosophers are keen to insist that they are not lexicographers; but the intelligibililty of empirical science can get a long way with lexicography alone. To the extent a conceptual analysis helps, it helps *after* we discover which way of cutting the causal joints of the social world works best, according to the naturalist.

3. Replacement Naturalism II: American Legal Realism

The *locus classicus* of the second kind of Replacement Naturalism—the one deriving from an attack on foundationalism—is Quine (1969). The central enterprise of epistemology on Quine's view is to understand the relation between our theories of the world and the evidence (sensory input) on which they are based. Quine's target is one influential construal of this project: Cartesian foundationalism, particularly in the sophisticated form given to it in the twentieth-century by Rudolf Carnap in *Der Logische Aufbau der Welt* (1928). The foundationalist wants an account of the theory-evidence relation that would vindicate the privileged epistemic status of at least some subset of our theories: our theories (in particular, our best theories of natural science) are to be “grounded” in indubitable evidence (i.e., immediate sense impressions). Quine deems foundationalism to be a failure: the semantic part of the program is rendered unrealizable by meaning holism on the one hand (theoretical terms get their meanings from their place in the whole theoretical framework, not in virtue of some point-by-point contact with sensory input), while the epistemic part of the program is defeated by the Duhem-Quine thesis about the underdetermination of theory by evidence on the other (there is always more than one theory consistent with the evidence, in part, because a theoretical hypothesis can always be preserved in the face of recalcitrant evidence by abandoning the auxiliary hypotheses that informed the test of the hypothesis) (see Kim 1988, pp. 385-386).

What becomes, then, of epistemology? Hilary Kornblith has summed up Quine's view as follows: “Once we see the sterility of the foundationalist program, we see that the only genuine questions there are to ask about the relation between theory and evidence and about the acquisition of belief are psychological questions” (Kornblith 1994, p. 4). This view Kornblith aptly dubs Quine's “replacement thesis”: “the view that epistemological questions may be replaced by psychological questions” (Kornblith 1994, p. 3). Here is how Quine puts it:

> The stimulation of his sensory receptors is all the evidence anybody has had to go on,

> ultimately, in arriving at his picture of the world. Why not just see how this construction really proceeds? Why not settle for psychology? Such a surrender of the epistemological burden to psychology is a move that was disallowed in earlier time as circular reasoning. If the epistemologist's goal is validation of the grounds of empirical science, he defeats his purpose by using psychology or other empirical science in the validation. However, such scruples against circularity have little point once we have stopped dreaming of deducing science from observations. (1969, pp. 75-76)

Several pages later, Quine continues that on his proposal,

> Epistemology, or something like it, simply falls into place as a chapter of psychology and hence of natural science. It studies a natural phenomenon, viz., a physical subject. This human subject is accorded a certain experientially controlled input—certain patterns of irradiation in assorted frequencies, for instance—-and in the fullness of time the subject delivers as output a description of the three-dimensional external world and its history. The relation between the meager input and the torrential output is a relation that we are prompted to study for somewhat the same reasons that always prompted epistemology; namely, in order to see how evidence relates to theory, and in what ways one's theory of nature transcends any available evidence. (1969, pp. 82-83)

Thus Quine: the central concern of epistemology is the theory-evidence relationship; if the foundationalist story about this relationship is a failure, then that leaves only one story worth telling about this relationship: namely, the story told by “a purely descriptive, causal-nomological science of human cognition” (Kim 1988, p. 388). The science of human cognition *replaces* armchair epistemology: we naturalize epistemology by turning over its central question—the relationship between theory and evidence—to the relevant empirical science.

We can now generalize Quine's point as follows (Leiter 1998). Let us say that a Replacement Naturalist in any branch of philosophy holds that:

> For any pair of relata that might stand in a *justificatory* relation—e.g., evidence and theory, reasons and belief, causal history and semantic or intentional content, legal reasons and judicial decision—if no normative account of the relation is possible, then the only theoretically fruitful account is the descriptive/explanatory account given by the relevant science of that domain.

This goes beyond Quine in one important respect: for Quine infers Replacement Naturalism only from the failure of *foundationalism*—which is simply one possible normative account of the evidence-theory relationship, but not the only one. Quine's arguments simply do not show that no other normative account of the evidence-theory relationship is possible.

Quine has been extensively criticized on precisely this score (e.g., Goldman 1986, pp. 2-3; Kim 1988).

The key to a successful defense of Replacement Naturalism lies in an explanation of why normative theory without foundationalism is *sterile*. One worry is that without foundationalism, normative theories are *banal*. Consider: it is now a familiar result of cognitive psychology that human beings regularly make mistakes in logical reasoning (cf. Stich 1994). So a mere descriptive theory of belief-formation, of the sort Quine appears to recommend, would simply record these mistakes. But shouldn't epistemology tell us that beliefs *ought* not to be formed illogically? One can hardly imagine why Quine would disagree: one *ought* not to form beliefs illogically. But the question is whether this piece of banal advice adds up to a fruitful research program? The descriptive project of Replacement Naturalism may record certain irrational cognitive processes in studying the evidence-theory relationship, but given the underdetermination of theory by evidence, even when we correct for logical mistakes, we still won't have an account of which of our theoretical beliefs are warranted and which are not. The Quinean intuition is we'll learn more from the empirical inquiry, than from systematizing our banal normative intuitions about irrationality. More generally, *unless* we have some foundational point outside our epistemic practices from which to assess the epistemic issues, the project of systematizing our mundane normative intuitions will simply collapse into the descriptive sociology of knowledge. If we can't stand outside the epistemological boat, then we can do no more than report what it is we do. But it is precisely the viability of such an external standpoint that Quine denies in his embrace of the metaphor of Neurath's boat. So from within the boat, there is nothing *to do* but description.

Quine's argument for Replacement Naturalism, recall, moved in two steps. Step one was *anti-foundationalism*: no unique theory is justified on the basis of the evidential input. Step two was *replacement*: since no foundational story can be told about the relation between input (evidence) and output (theory), we should replace the normative program with a purely descriptive inquiry, e.g. the psychological study of what input causes what output. We can find analogues of both steps in the approach to the theory of adjudication offered by American Legal Realism.

Theory of adjudication is concerned *not* with the relationship between "evidence" and "scientific theory," but rather with the justificatory relationship between "legal reasons" (the input, as it were) and judicial decision (the output): theory of adjudication tries to tell judges how they *ought* to justify their decisions, i.e. it seeks to "ground" judicial decision-making in reasons that require unique outcomes. The American Legal Realists are "anti-foundationalists" about judicial decisions in the sense that they deny that the legal reasons justify a unique decision: the legal reasons underdetermine the decision (at least in most cases actually litigated). More precisely, the Realists claim that the law is *rationally* indeterminate in the sense that the class of legal reasons—i.e., the class of legitimate reasons a judge may offer for a decision—does not provide a *justification* for a unique outcome. Just as sensory input does not *justify* a unique scientific theory, so legal reasons, according to the Realists, do not *justify* a unique decision.

The Realists also take the second step that Quine takes: replacement. According to the Realist indeterminacy thesis, legal reasons do not justify a unique decision, meaning that the foundationalist enterprise of theory of adjudication is impossible. Why not replace, then, the "sterile" foundational program of justifying some one legal outcome on the basis of the applicable legal reasons, with a descriptive/explanatory account of what input (that is, what combination of facts and reasons) produces what output (i.e., what judicial decision)? As Underhill Moore puts it at the beginning of one of his

articles: “This study lies within the province of jurisprudence. It also lies within the field of behavioristic psychology. It places the province within the field” (Moore & Callahan 1943, p. 1). Notice how closely this echoes Quine's idea that, “Epistemology…simply falls into place as a chapter of psychology…” (1969, p. 82). Jurisprudence—or, more precisely, the theory of adjudication—is “naturalized” because it falls into place, for the Realist, as a chapter of psychology (or economics or sociology, etc.). Moreover, it does so for essentially Quinean reasons: because the foundational account of adjudication is a failure—a consequence of accepting the Realists' famous claim that the law is indeterminate.

Of course, this argument for Replacement Naturalism only seems to work against “formalist” theories of adjudication that are committed to the rational determinacy of law. But, some object, “No contemporary analytic jurisprudent is a formalist” (Coleman 1998, p. 284), and some have even claimed that the “formalists” the Legal Realists opposed were not committed to the rational determinacy of law (Paulson 2001, p. 78). Both objections seem mistaken: Dworkin, for example, is committed to the rational determinacy of law in exactly the sense at issue for the Replacement argument. And it is even conceded that all legal theorists are committed to the rational determinacy of law in “at least some legal disputes” (Coleman 1998, p. 284), thus making them vulnerable, in principle, to the Replacement argument. The targets of the Legal Realist critique were, equally, committed to the rational determinacy of law; indeed, it would be impossible to make sense of what the Realists were doing if that were not so. The Replacement Naturalist might take the view that there is no reason to call for “naturalizing” theory of adjudication in those range of cases where legal reasons *are* satisfactory predictors of legal outcomes (i.e., precisely those cases where the foundationalist program can be carried out). One may worry, again, about whether there is an *interesting* or *fruitful* normative story to be told (rather than a merely banal descriptive sociology), but it suffices for the analogy with Quine that there remains some substantial domain of cases where the foundational program can not be carried out, so that the case for replacement remains intact.

The real difficulty, of course, pertains not to these historical points, but to whether or not the project of a normative theory of adjudication warrants replacement just because rational determinacy does not obtain. As in the Quinean case, the Replacement Naturalist must maintain that without rational determinacy, normative theories of adjudication are banal, mere excercises in descriptive sociology. Critics of Replacement Naturalism contest this conclusion, though more by way of affirmation than argument (Coleman 1998, p. 285 n. 44). However, if the objection under consideration were correct, then a *normative* theory that specifies what the anti-foundationalist concedes—namely, that there is more than one (though not simply any) judicial decision that can be justified on the basis of the class of legal reasons—must, in some measure, be a theory worth having. Arguably, such a theory might be adequate to deflect the challenge to the political *legitimacy* of adjudication based on the indeterminacy of law, but does it provide the *normative* guidance to judges we want from a theory? Does a theory that tells judges they would be justified (on the basis of the class of legal reasons) in deciding for the plaintiff on theory *X or* the defendant on theory *Y* (but not the plaintiff or defendant on theory Z!) really provide normative guidance for judges worth having? The Replacement Naturalist answers in the negative: better to have a descriptive account of inputs and outputs, one that would license prediction of judicial behavior, than an indeterminate normative theory. This response, of course, makes Replacement Naturalism vulnerable to conflicting intuitions about the fruitfulness or sterility of different kinds of theorizing.

There are other limits to the Quinean analogy (Leiter 2001a, pp. 284-285). First, the American Legal Realists end up presupposing a theory of the concept of legality in framing their arguments for law's indeterminacy (Leiter 1995; Leiter 2001a, pp. 292-293); thus, while they may believe the only fruitful account of *adjudication* is descriptive and empirical, not normative and conceptual, they themselves need a concept of *law* that is not—at least on the arguments considered so far—empirical or naturalized. As one critic of Replacement Naturalism notes: “the naturalist is committed as a conceptual matter to the existence of a test of legality…. The naturalist is thus in the same boat with ever other analytic philosopher of law” (Coleman 2001, p. 214). The analogy with naturalized epistemology, in other words, must be localized to the theory of adjudication, and not the whole of jurisprudence. Of course, it remains possible for the Replacement Naturalist to argue for the requisite concept of legality on precisely the empirical grounds noted in the previous section (“Replacement Naturalism I: Against Conceptual Analysis”). But as it stands, the analogy to Quine's attack on foundationalist epistemology warrants no radical abandonment of traditional conceptual analysis across the boards.

A second difference from Quine is also important: for the crux of the Legal Realist position (at least for the majority of Realists) is that non-legal reasons (e.g., judgments of fairness, or consideration of commercial norms) *explain* the decisions. They, of course, explain the decisions by *justifying* them, though not necessarily by justifying a unique outcome (i.e., the non-legal reasons might themselves rationalize other decisions as well). Now clearly the descriptive story about the non-legal reasons is not going to be part of a non-mentalistic naturalization of the theory of adjudication: a causal explanation of decisions in terms of reasons (even non-legal reasons) does require taking the normative force of the reasons *qua reasons* seriously. The behaviorism of Quine or Underhill Moore is not in the offing here, but surely this is to be preferred: behaviorism failed as a foundation for empirical social science, while social-scientific theories employing mentalistic categories have flourished. Moreover, if the non-legal reasons are themselves indeterminate—i.e., if they do not justify a *unique* outcome—then any causal explanation of the decision will have to go beyond reasons to identify the psyco-social facts (e.g., about personality, class, gender, socialization, etc.) that cause the decision. Such a “naturalization” of the theory of adjudication might be insufficiently austere in its ontology for Quinean scruples, but it is still a recognizable attempt to subsume what judges do within a (social) scientific framework.

4. Normative Naturalism

Like the traditional epistemologist, the Normative Naturalist embraces as his goal the promulgation of norms by which to regulate our epistemic practices (to govern how we should acquire and weigh evidence, as well as, ultimately, form beliefs). Unlike the non-naturalist, however, the Normative Naturalist does not think epistemic norms can be adequately formulated from the armchair: normative theorizing must be continuous with scientific theorizing. But if this is not just to collapse into Replacement Naturalism then what does the M-naturalist credo amount to in the normative case? Consider Goldman's proposal: “Epistemics assumes that cognitive operations should be assessed instrumentally: given a choice of cognitive procedures, those which would produce the best set of *consequences* should be selected” (1978, p. 520). The Normative Naturalist maintains that the reason the

philosopher can't do armchair epistemology is because it is an *a posteriori*, empirical matter what norms *in fact* serve our epistemic or cognitive goals (e.g., forming true beliefs). Goldman emphasizes a particularly important instance of this general point:

> [A]dvice in matters intellectual, as in other matters, should take account of the agent's capacities. There is no point in recommending procedures that cognizers cannot follow or recommending results that cognizers cannot attain. As in the ethical sphere, "ought" implies "can." Traditional epistemology has often ignored this precept. Epistemological rules often seem to have been addressed to "ideal" cognizers, not human beings with limited information-processing resources. Epistemics [as a type of Normative Naturalism] wishes to take its regulative role seriously. It does not want to give merely idle advice, which humans are incapable of following. This means it must take account of the powers and limits of the human cognitive system, and this requires attention to descriptive psychology. (1978, p. 510)

So the Normative Naturalist thinks that normative epistemology must be continuous with natural and social science in (at least) two senses: (i) we need to know what epistemic norms in fact lead to our forming true beliefs; and (ii) as a special case of (i), we need to identify epistemic norms actually usable by creatures like us. This rules out certain (non-naturalistic) epistemic norms which require of cognizers belief-formation practices beyond their ken (Goldman 1978, pp. 512-513). The Normative Naturalist, in short, emphasizes the *instrumental* character of normative theorizing in epistemology, and then argues that the only way to assess instrumental claims is empirically—to see what means *really* brings about what ends. And that task can never be pursued *a priori*, from the armchair, simply by analyzing the meaning of the words "knowledge" or "justified" or "true."

Of course, it bears noting that the Normative Naturalist does not dispense *entirely* with conceptual analysis—to the contrary. It is precisely, for example, Goldman's proffered conceptual analyses of "knowledge" and "justification" that require him to turn to empirical psychology to fill in the actual content of epistemic norms. Unlike the Quinean program, naturalization enters for the Normativist only, as it were, in *applied* epistemology. What many philosophers might think of as "pure" epistemology—giving an account of knowledge—continues to be an *a priori* enterprise, even though it is an enterprise that invokes notions (like "reliability" and "causation") that require *a posteriori* investigation to apply.

The Normative Naturalist in jurisprudence, too, views theoretical questions *instrumentally*. The philosophical foundation of evidence law has, to date, received the most attention from this perspective (Allen & Leiter 2001; Leiter 2001d). We want to ask, as Goldman puts, "Which [social] practices have a comparatively favorable impact on knowledge as contrasted with error and ignorance?" (1999, p. 5). Normative naturalism is, in this respect, *veritistic* (to borrow Goldman's term): it is concerned with the production of knowledge, meaning (in part) *true* belief (Goldman 1999, pp. 79-100). So the Normative Naturalist embraces as his goal the promulgation of norms by which to regulate our epistemic practices so that they yield knowledge. In the case of *individual* epistemology, this means the norms governing how individuals should acquire and weigh evidence as well as, ultimately, form beliefs; in the case of

social epistemology, this means the norms governing the social mechanisms and practices that inculcate belief. The legal rules of evidence, in turn, are a prime case of the latter: for these rules structure the epistemic process by which jurors arrive at beliefs about disputed matters of fact at trials. As such, the rules of evidence are a natural candidate for investigation by Normative Naturalists. We may ask of any particular rule: does it increase the likelihood that jurors will reach *true* beliefs about disputed matters of fact? (Of course, it does not make sense to ask that of *every* rule, since some rules of evidence—for example, Federal Rules of Evidence (FRE) 407-411—are not meant to facilitate the discovery of truth, but to carry out various policy objectives like reducing accidents and avoiding litigation.) That means, of course, asking an essentially *empirical* question: does this rule of inclusion or exclusion *in fact* increase the likelihood that fact finders, *given what they are actually like,* will achieve knowledge about disputed matters of fact (i.e., does it maximize veritistic value). Of course, many rules that on their face invite one kind of veritistic analysis require a very different kind in practice. So, for example, FRE 404, *on its face*, excludes character evidence in most contexts, though, in fact, the exception in 404(b) largely swallows the rule. Thus, while it might seem that we should ask whether *excluding* character evidence maximizes veritistic value, the real question is whether *admitting* it does. The same may be said for the hearsay rule. Although on its face, the hearsay doctrine is a rule of exclusion, in reality it is a rule of admission: what the advocate must really know is how to get the proferred hearsay admitted under one of the multitude of exceptions to the nominal rule of exclusion (FRE 802). Thus, the pertinent veritistic question concerns the veritistic credentials of the grounds on which hearsay is admitted, rather than the veritistic reasons for excluding it in most cases. (Such questions, in fact, are already a staple of much evidence scholarship.)

In theory of adjudication, by contrast, the Normative Naturalist wants to identify norms for adjudication that will help judges realize adjudicative goals. Such norms must, once again, satisfy two naturalistic constraints: first, they must, as a matter of empirical fact, be effective means to goals ("the Instrumental Constraint"); second, they must be constrained by relevant empirical facts about the nature and limitations of judges ("the Ought-Implies-Can Constraint") (Leiter 1998).

Dworkin's theory of ajudication (Dworkin 1986) makes a popular target for the Normative Naturalist. Dworkin's theory says, very roughly, that a judge should decide a case in such a way that it coheres with the principle that *explains* some significant portion of the prior institutional history *and* provides the best *justification* for that history as a matter of political morality. Can a Normative Naturalist be a Dworkinian?

(1) **Instrumental Constraint**: The naturalist assesses normative advice relative to its actual effectiveness for realizing relevant goals. What, then, is the relevant goal in *adjudication*? One candidate is surely this: we want to give judges normative advice that will lead them to reach fair or just outcomes. Thus, the naturalist's question becomes: which piece of normative advice is most effective in *really* helping actual judges realize justice and fairness? It is, at least, an open question whether Dworkin's methodology will be effective in leading judges to do fair things. The fact that his normative theory has had almost no impact whatsoever on American judicial practice over the last thirty years is at least defeasible evidence that it does not appear to be an *effective* methodology (let alone one effective for realizing justice!) (Leiter 1998, p. 102). This latter point is related to the naturalist's second, and more important objection.

(2) **Ought-Implies-Can-Constraint:** One thing judges can not do is what Dworkin's Judge Hercules does. This is a familiar complaint about Dworkin's theory, but naturalized jurisprudence gives it a principled foundation. The naturalistic jurisprudent eschews all normative guidance unusable by real judges; like his naturalized counterpart in epistemology, he does “not want to give merely idle advice, which humans [including judges] are incapable of following” (Goldman 1978, p.510). Dworkin may give judges an “aspirational model” (to borrow Jules Coleman's apt phrase), and the naturalistic jurisprudent need not dispute that; but Descartes gave us an aspirational model in epistemology as well, and that does not make his program any more adequate or relevant by the naturalist's lights. (It *would* be attractive if we could take certain “clear and distinct” ideas, and build up all of knowledge from them.) The naturalist wants normative advice effective for *creatures like us;* demanding of judges Herculean philosophical ingenuity violates this constraint. Aspiration, the naturalist concludes, is not a fit aim of normative advice, which must, first and foremost, offer *effective* means to ends.

5. Substantive Naturalism

There have been two prominent groups of S-naturalists in legal philosophy: the Scandinavian Legal Realists (like Axel Hägerström and Alf Ross), whose austere views about the ontology of the natural world, conjoined with moral skepticism, led them to unusual conclusions about the semantics of legal propositions; and contemporary defenders of a kind of natural law theory (like David Brink, Michael Moore, and Nicos Stavropoulos), who invoke the Causal Theory of Reference associated with Kripke and Putnam to offer an interpretation of some legal predicates in substantively naturalistic terms.

The S-naturalism of the Scandinavian Realists is, today, more a museum piece than a live contender in jurisprudential debate. This is unsurprising in light of its dependence, especially in the work of Ross, on the semantic doctrines of logical positivism and cognate developments, like emotivism in ethics. As one well-known commentator has written, the Scandinavian Realists held that “a proposition which does not admit to being reduced to enunciations of facts can have no real and intelligible meaning” (Bjarup 1999, p. 774). Their allegiance to emotivism (and non-cognitivism more generally) in ethics led the Scandinavian Realists to conclude that normative statements (at least in morality and law—like the emotivists, they simply ignored the question about the status of epistemic norms) had to have a naturalistically respectable reduction if they were to have truth values. For Ross (1958, Ch. 2), this reduction base was a combination of behavioral and psychological facts: to say that, “*X* is a valid law” means to predict that (1) judges will *act* in accordance with that law, and (2) in so acting, they will “feel” themselves bound to do so.

In an influential essay (reviewing Ross 1958), H.L.A. Hart famously demolished this analysis. “A valid law,” said Hart, can not be “a verifiable hypothesis about future judicial behavior and its special motivational feeling” since such an account makes no sense of the “meaning” of judgments of legal validity “in the mouth of a judge who is not engaged in predicting his own or others' behavior or feelings”: “‘This is a valid rule’ said by a judge is an act of recognition; in saying it he recognizes the rule in question as one satisfying certain accepted general criteria for admission as a rule of the system and so as a legal standard of behavior” (1959, p. 165). This critique, expanded upon in Chapter 7 of Hart

(1961), did much to consign Scandinavian Realism to the history of ideas, though it, unfairly, had the same impact on American Legal Realism, which was not, in fact, committed to this semantic analysis (on the latter point, see Leiter 2001a, esp. pp. 290-293).

Just as semantic doctrines more-or-less derived from logical positivism were central to the S-naturalism of Scandinavian Realism, the more recent S-naturalism of writers like Brink, Moore, and Stavropoulos is indebted to the revolution in semantics initiated by Hilary Putnam and Saul Kripke known as the "new" or "causal" theory of reference. These latter doctrines are not—yet in any case—museum pieces, so the derived jurisprudential theses are very much live items of debate. Stavropoulos explains the core semantic ideas on which jurisprudential writers draw as follows:

> Both Kripke and Putnam attack what they call the traditional theory of reference. That theory holds that an expression refers to whatever fits the description with which speakers associate the expression. The relevant description…captures necessary properties of the referent which are knowable *a priori*, as in the case of knowing that a bachelor is an unmarried man. This cannot be true, Kripke and Putnam argue, since expressions refer to the same object in the lips of speakers who can only associate the expression with vague or mistaken descriptions. Indeed, not only individual speakers but the community as a whole can be in error about the true properties of the relevant object. … .The important suggestion being made by Kripke and Putnam is that reference is *object-dependent*. *Which* object 'Aristotle' or 'water' refers to is not determined by the associated description, but turns instead on a matter of fact, namely which object the name-using or term-using practice is directed at. (1996, p. 8)

Thus, if on the old view the "meaning" of an expression (the descriptions speakers associated with it) fixed the reference of the expression, on the new theory, the referent fixes the meaning. "Water" picks out whatever stuff we happened to baptize with the name "water" at the beginning of the "term-using practice." As it happens, that stuff has a distinctive micro-constitution: it is H_20. Thus, "water" refers to stuff that is H_20, and that is what the term means: the stuff that is H_20.

Writers like Brink, Moore, and Stavropoulos propose that when the meaning of expressions in legal rules—and, in particular, the meaning of the moral concepts (like "equality") that figure in some legal rules—is understood in the same way, then it follows that all rules have *determinate* applications: either the facts do or do not fall within the extension of the meaning of the key terms in the rule. The meaning of the rule determines its application, but the meaning is fixed by the *real* referents of the terms in the legal rule. Of course, for this to be a version of S-naturalism, the claim must be that the real referents are themselves things cognizable within a naturalistic framework: so, e.g., it would have to be the case that the legal and moral features of situations picked out by our legal concepts (and the moral concepts embedded in them) must be identical with (or supervenient upon) natural facts: just as there are necessary, *a posteriori* statements of property identity about water, so too there are such statements about legal and moral facts. For example, perhaps the property of being "morally right" is just identical with the property of "maximizing human well-being," where the latter may be understood in purely

psychological and physiological terms. In that case, whether an action *X* is morally right is simply a scientific question about whether action *X* will in fact maximize the relevant kinds of psychological and physiological states in the world. (Most naturalistic moral realisms are based on versions of utilitarianism, precisely because it is easy to see what the naturalistic base of moral properties would be in a utilitarian schema. One peculiar feature of the moral realism of Moore (1992b) is that it is conjoined with a deontological moral theory, yet within a purportedly naturalistic moral realist framework.) The crucial claim, plainly, is that moral facts are to be identified with (or treated as supervenient upon) certain kinds of natural facts. Of course, many philosophers are skeptical that this claim can be made out (e.g., Gibbard 1990).

Problems arise at several different levels with this more recent S-naturalism, though all are traceable to the reliance on the new theory of reference. To begin, there are familiar reasons to be skeptical about whether the new theory of reference is correct, reasons that won't be rehearsed here (see, e.g., Evans 1973, Blackburn 1988). Even granting, however, the correctness of the new theory, it is not obvious how it helps in the case of law. After all, the new theory always seemed most plausible for a limited class of expressions: proper names and natural kind terms. The reason has to do with the implicit essentialism required for the new theory: unless referents have *essential* characteristics—just as “water” has a distinctive and essential molecular constitution—they can not fix meanings. But what is the essence of “due process” or “equal protection?” And what is the “essential” nature of the many artifact terms that populate legal rules (terms like “contract” or “vehicle” or “security interest”)? Unsurprisingly, S-naturalists like Brink and Moore are also moral realists, and also try to give accounts of artifact terms as picking out not “natural kinds” but “functional kinds” (Moore 1992a, pp. 207-208).

Of course, even if the new theory of reference gives the correct account of the meaning of some terms (like natural kind terms), that still does not show that it gives us the right account of meaning for purposes of legal interpretation (cf. Munzer 1985). Suppose the legislature prohibits the killing of “fish” within 100 miles of the coast, intending quite clearly (as the legislative history reveals) to protect whales, but not realizing that “fish” is a natural kind term that does not include whales within its extension. The new theory of reference tells us that the statute protects sea bass but not whales, yet surely a court that interpreted the statute as also protecting whales would not be making a mistake. Indeed, one might think the reverse is true: for a court *not* to protect whales would be to contravene the will of the legislature, and thus, indirectly, the will of the people. What the example suggests is that the correct theory of legal interpretation is *not* a mere matter of philosophical semantics: issues about *political legitimacy*—about the conditions under which the exercise of coercive power by courts can be justified—must inform theories of legal interpretation, and such considerations may even trump considerations of semantics.

Bibliography

- Allen, Ronald J. and Brian Leiter (2001). “Naturalized Epistemology and the Law of Evidence,” *Virginia Law Review* 87: 1491-1550.
- Bix, Brian (ed.) (1998). *Analyzing Law: New Essays in Legal Theory* (Oxford: Clarendon Press).
- Bjarup, Jes (1999). “Scandinavian Legal Realism,” in C.B. Gray (ed.), *The Philosophy of Law: An*

Encyclopedia (New York: Garland, 1999).

- Blackburn, Thomas (1988). “The Elusiveness of Reference,” *Midwest Studies in Philosophy* 12: 179-194.
- Brink, David O. (1988). “Legal Theory, Legal Interpretation, and Judicial Review,” *Philosophy & Public Affairs* 17: 105-148.
- ----- (1989). *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press).
- ----- (2001). “Legal Interpretation, Objectivity, and Morality,” in B. Leiter (ed.), *Objectivity in Law and Morals* (New York: Cambridge University Press).
- Coleman, Jules L (1998). “Second Thoughts and Other First Impressions,” in Bix (1998).
- ----- (2001). *The Practice of Principle* (Oxford: Clarendon Press).
- Cummins, Robert (1999). “Reflection on Reflective Equilibrium,” in M. DePaul & W. Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry* (Lanham, Md.: Rownman & Littlefield).
- Dworkin, Ronald (1986). *Law's Empire* (Cambridge, Mass.: Harvard University Press).
- Evans, Gareth (1973). “The Causal Theory of Names,” reprinted in *The Varieties of Reference* (Oxford: Clarendon Press, 1982).
- Gibbard, Allan (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Cambridge, Mass.: Harvard University Press).
- Goldman, Alvin I. (1978). “Epistemics: The Regulative Theory of Cognition,” *Journal of Philosophy* 75: 509-523.
- -----. (1986) *Epistemology and Cognition* (Cambridge, Mass.: Harvard University Press).
- ----- (1994). “Naturalistic Epistemology and Reliabilism,” *Midwest Studies in Philosophy* 19: 301-320.
- ----- (1999). *Knowledge in a Social World* (Oxford: Oxford University Press).
- Hägerström, Axel (1953). *Inquiries into the Nature of Law and Morals*, ed. K. Olivecrona, trans. C.D. Broad (Stockholm: Almquist & Wiksell).
- Harman, Gilbert (1994). “Doubts About Conceptual Analysis,” in M. Michael & J. O'Leary-Hawthorne (eds.), *Philosophy in Mind* (Dordrecht: Kluwer).
- Hart, H.L.A. (1961). *The Concept of Law* (Oxford: Clarendon Press).
- ----- (1959). “Scandinavian Realism,” reprinted in *Essays in Jurisprudence and Philosophy* (Oxford: Clarendon Press, 1983).
- Hintikka, Jaakko (1999). “The Emperor's New Intuitions,” *Journal of Philosophy* 96: 127-147.
- Jackson, Frank (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Oxford: Clarendon Press).
- Kim, Jaegwon (1988). “What is ‘Naturalized’ Epistemology?” *Philosophical Perspectives* 2: 381-405.
- Kitcher, Philip (1992). “The Naturalists Return,” *Philosophical Review* 101: 53-114.
- Kornblith, Hilary (ed.) (1994a) *Naturalizing Epistemology*, 2nd ed. (Cambridge, Mass.: MIT Press).
- ----- (1994b) “Introduction: What is Naturalistic Epistemology?” in Kornblith (1994a).
- Laurence, Stephen and Eric Margolis (1999). “Concepts and Cognitive Science,” in E. Margolis & S. Laurence (eds.), *Concepts: Core Readings* (Cambridge, Mass.: MIT Press).

- Leiter, Brian (1995). "Legal Indeterminacy," *Legal Theory* 1: 481-492.
- -----. (1997). "Rethinking Legal Realism: Toward a Naturalized Jurisprudence," *Texas Law Review* 76: 267-315.
- ----- (1998). "Naturalism and Naturalized Jurisprudence," in Bix (1998).
- ----- (2001a). "Legal Realism and Legal Positivism Reconsidered," *Ethics* 111: 278-301.
- ----- (2001b). "Legal Realism, Hard Positivism, and the Limits of Conceptual Analysis," in J.L. Coleman (ed.), *The Postscript: Essays on Hart's Postscript to The Concept of Law* (Oxford: Clarendon Press).
- ----- (2001c). "The Naturalistic Turn in Legal Philosophy," *APA Newsletter on Law and Philosophy* (Spring): 142-146.
- ----- (2001d). "Prospects and Problems for the Social Epistemology of Evidence Law," forthcoming, *Philosophical Topics*
- Llewellyn, Karl (1930) *The Bramble Bush* (New York: Oceana).
- Moore, Underhill and Charles Callahan (1943). "Law and Learning Theory: A Study in Legal Control," *Yale Law Journal* 53: 1-136.
- Moore, Michael S. (1985). "A Natural Law Theory of Interpretation," *Southern California Law Review* 58: 277-398.
- -----. (1992a). "Law as a Functional Kind," in R. George (ed.), *Natural Law Theory: Contemporary Essays* (Oxford: Clarendon Press).
- ----- (1992b). "Moral Reality Revisited," *Michigan Law Review* 90: 2424-2533.
- Munzer, Stephen R. (1985). "Realistic Limits on Realist Interpretation," *Southern California Law Review* 58: 459-475.
- Paulson, Stanley (2001). Review of Bix (1998), *Philosophical Books* 42: 76-80.
- Peacocke, Christopher (1992). *A Study of Concepts* (Cambridge, Mass.: MIT Press).
- Perry, Stephen R. (1987). "Judicial Obligation, Precedent, and the Common Law," *Oxford Journal of Legal Studies* 7: 215-257.
- Powe, Jr., Lucas A. (2000). *The Warren Court and American Politics* (Cambridge, Mass.: Harvard University Press).
- Pritchett, C. Herman (1948). *The Roosevelt Court: A Study in Judicial Politics and Values, 1937-1947* (New York: Macmillan Co.).
- Quine, W.V.O. (1951). "Two Dogmas of Empiricism," reprinted in *From a Logical Point of View* (Cambridge, Mass.: Harvard University Press, 1953).
- -----. (1969). "Epistemology Naturalized," in *Ontological Relativity and Other Essays* (New York: Columbia University Press).
- Railton, Peter (1986). "Moral Realism," *Philosophical Review* 95: 163-205.
- -----. (1990). "Naturalism and Prescriptivity," in E.F. Paul, *et al*. (eds), *Foundations of Moral and Political Philosophy* (Oxford: Blackwell).
- Raz, Joseph (1985). "Authority, Law and Morality," *The Monist* 68: 295-324.
- Ross, Alf (1958). *On Law and Justice* (Berkeley: University of California Press).
- Segal, Jeffrey A. & Harold J. Spaeth (1993). *The Supreme Court and the Attitudinal Model* (New York: Cambridge University Press).
- Stavropoulos, Nicos (1996). *Objectivity in Law*. (Oxford: Clarendon Press).
- Stich, Stephen P (1994). "Could Man Be an Irrational Animal?" in Kornblith (1994a).

- Talbott, William J. & Alvin I. Goldman (1998). “Games Lawyers Play: Legal Discovery and Social Epistemology,” *Legal Theory* 4 (1998): 93-163.
- Waluchow, W.J. (1994). *Inclusive Legal Positivism* (Oxford: Clarendon Press).

Other Internet Resources

[Please contact the author with suggestions]

Related Entries

adjudication | atheism-agnoticism | conceptual analysis | epistemology: naturalizedlegal realism: American | legal realism: Scandinavian | naturalism | naturalism: methodological | naturalism: substantive | natural law | reference: causal theory of

Brian Leiter
bleiter@mail.law.utexas.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 15, 2002
Content last modified: July 15, 2002

The Mathematics of Boolean algebra

Boolean algebra is the algebra of two-valued logic with only sentential connectives, or equivalently of algebras of sets under union and complementation. The rigorous concept is that of a certain kind of algebra, analogous to the mathematical notion of a group. This concept has roots and applications in logic (Lindenbaum-Tarski algebras and model theory), set theory (fields of sets), topology (totally disconnected compact Hausdorff spaces), foundations of set theory (Boolean-valued models), measure theory (measure algebras), functional analysis (algebras of projections), and ring theory (Boolean rings). The study of Boolean algebras has several aspects: structure theory, model theory of Boolean algebras, decidability and undecidability questions for the class of Boolean algebras, and the indicated applications. In addition, although not explained here, there are connections to other logics, subsumption as a part of special kinds of algebraic logic, finite Boolean algebras and switching circuit theory, and Boolean matrices.

1. Definition and simple properties

A Boolean algebra (BA) is a set *A* together with binary operations + and $\cdot$ and a unary operation $-$, and elements 0, 1 of *A* such that the following laws hold: commutative and associative laws for addition and multiplication, distributive laws both for multiplication over addition and for addition over multiplication, and the following special laws:

$$x + (x \cdot y) = x$$
$$x \cdot (x + y) = x$$
$$x + (-x) = 1$$

$$x \cdot (-x) = 0$$

These laws are better understood in terms of the basic example of a BA, consisting of a collection *A* of subsets of a set *X* closed under the operations of union, intersection, complementation with respect to *X*, with members $\varnothing$ and *X*. One can easily derive many elementary laws from these axioms, keeping in mind this example for motivation. Any BA has a natural partial order $\leq$ defined upon it by saying that $x \leq y$ if and only if $x + y = y$. This corresponds in our main example to $\subseteq$. Of special importance is the two-element BA, formed by taking the set *X* to have just one element. An important elementary result is that an equation holds in all BAs if and only if it holds in the two-element BA. Next, we define $x \oplus y = (x \cdot -y) + (y \cdot -x)$. Then *A* together with $\oplus$ and $\cdot$, along with 0 and 1, forms a ring with identity in which every element is idempotent. Conversely, given such a ring, with addition $\oplus$ and multiplication, define $x + y = x \oplus y \oplus (x \cdot y)$ and $-x = 1 \oplus x$. This makes the ring into a BA. These two processes are inverses of one another, and show that the theory of Boolean algebras and of rings with identity in which every element is idempotent are definitionally equivalent. This puts the theory of BAs into a standard object of research in algebra. An atom in a BA is a nonzero element *a* such that there is no element *b* with $0 < b < a$. A BA is atomic if every nonzero element of the BA is above an atom. Finite BAs are atomic, but so are many infinite BAs. Under the partial order $\leq$ above, $x + y$ is the least upper bound of *x* and *y*, and $x \cdot y$ is the greatest lower bound of *x* and *y*. We can generalize this: $\sum X$ is the least upper bound of a set *X* of elements, and $\prod X$ is the greatest lower bound of a set *X* of elements. These do not exist for all sets in all Boolean algebras; if they do always exist, the Boolean algebra is said to be complete.

2. The elementary algebraic theory

Several algebraic constructions have obvious definitions and simple properties for BAs: subalgebras, homomorphisms, isomorphisms, and direct products (even of infinitely many algebras). Some other standard algebraic constructions are more peculiar to BAs. An ideal in a BA is a subset *I* closed under +, with 0 as a member, and such that if $a \leq b \in I$, then also $a \in I$. Although not immediately obvious, this is the same as the ring-theoretic concept. There is a dual notion of a filter (with no counterpart in rings in general). A filter is a subset *F* closed under $\cdot$, having 1 as a member, and such that if $a \geq b \in F$, then also $a \in F$. An ultrafilter on *A* is a filter *F* with the following properties: $0 \notin F$, and for any $a \in A$, either $a \in F$ or $-a \in F$. For any $a \in A$, let $S(a) = \{F : F$ is an ultrafilter on A and $a \in F\}$. Then *S* is an isomorphism onto a BA of subsets of the set *X* of all ultrafilters on *A*. This establishes the basic Stone representation theorem, and clarifies the origin of BAs as concrete algebras of sets. Moreover, the sets $S(a)$ can be declared to be a base for a topology on *X*, and this turns *X* into a totally disconnected compact Hausdorff space. This establishes a one-one correspondence between the class of BAs and the class of such spaces. As a consequence, used very much in the theory of BAs, many topological theorems and concepts have consequences for BAs. If *x* is an element of a BA, we let $0x = -x$ and $1x = x$. If $(x(0), \ldots x(m-1))$ is a finite sequence of elements of a BA *A*, then every element of the subalgebra of *A* generated by $\{x(0), \ldots, x(m-1)\}$ can be written as a sum of monomials $e(0)x(0) \cdot \ldots \cdot e(m-1)x(m-1)$ for *e* in some set of functions mapping $m = \{0, \ldots, m-1\}$ into $2 = \{0, 1\}$. This is an algebraic expression of the disjunctive normal form theorem of sentential logic. A function *f* from a set *X* of generators of a BA *A* into a BA *B* can be extended to a homomorphism if and only if $e(0)x(0) \cdot \ldots \cdot e(m-1)x(m-1) = 0$

always implies that $e(0)f(x(0)) \cdot \ldots \cdot e(m - 1)f(x(m - 1)) = 0$. This is Sikorski's extension criterion. Every BA A can be embedded in a complete BA B in such a way that every element of B is the least upper bound of a set of elements of A. B is unique up to A-isomorphism, and is called the completion of A. If f is a homomorphism from a BA A into a complete BA B, and if A is a subalgebra of C, then f can be extended to a homomorphism of C into B. This is Sikorski's extension theorem. Another general algebraic notion which applies to Boolean algebras is the notion of a free algebra. This can be concretely constructed for BAs. Namely, the free BA on κ is the BA of closed-open subsets of the two element discrete space raised to the κ power.

3. Special classes of Boolean algebras

There are many special classes of Boolean algebra which are important both for the intrinsic theory of BAs and for applications:

- Atomic BAs, already mentioned above
- Atomless BAs, which are defined to be BAs without any atoms. For example, any infinite free BA is atomless.
- Complete BAs, defined above. These are specially important in the foundations of set theory.
- Interval algebras. These are derived from linearly ordered sets $(L, <)$ with a first element as follows. One takes the smallest algebra of subsets of L containing all of the half-open intervals $[a, b)$ with a in L and b in L or equal to ∞. These BAs are useful in the study of Lindenbaum-Tarski algebras. Every countable BA is isomorphic to an interval algebra, and thus a countable BA can be described by indicating an ordered set such that it is isomorphic to the corresponding interval algebra.
- Tree algebras. A tree is a partially ordered set $(T, <)$ in which the set of predecessors of any element is well-ordered. Given such a tree, one considers the algebra of subsets of T generated by all sets of the form $\{b : a \leq b\}$ for some a in T.
- Superatomic BAs. These are BAs which are not only atomic, but are such that each subalgebra and homomorphic image is atomic.

4. Structure theory and cardinal functions on Boolean algebras

Much of the deeper theory of Boolean algebras, telling about their structure and classification, can be formulated in terms of certain functions defined for all Boolean algebras, with infinite cardinals as values. We define some of the more important of these cardinal functions, and state some of the known structural facts, mostly formulated in terms of them

1. The cellularity $c(A)$ of a BA is the supremum of the cardinalities of sets of pairwise disjoint elements of A.

2. A subset *X* of a BA *A* is independent if *X* is a set of free generators of the subalgebra that it generates. The independence of *A* is the supremum of cardinalities of independent subsets of *A*.
3. A subset *X* of a BA *A* is dense in *A* if every nonzero element of *A* is $\geq$ a nonzero element of *X*. The π-weight of *A* is the smallest cardinality of a dense subset of *A*.
4. Two elements *x*, *y* of *A* are incomparable if neither one is $\leq$ the other. The supremum of cardinalities of subset *X* of *A* consisting of pairwise incomparable elements is the incomparability of *A*.
5. A subset *X* of *A* is irredundant if no element of *X* is in the subalgebra generated by the others.

An important fact concerning cellularity is the Erdos-Tarski theorem: if the cellularity of a BA is a singular cardinal, then there actually is a set of disjoint elements of that size; for cellularity regular limit (inaccessible), there are counterexamples. Every infinite complete BA has an independent subset of the same size as the algebra. Every infinite BA *A* has an irredundant incomparable subset whose size is the π-weight of *A*. Every interval algebra has countable independence. A superatomic algebra does not even have an infinite independent subset. Every tree algebra can be embedded in an interval algebra. A BA with only the identity automorphism is called rigid. There exist rigid complete BAs, also rigid interval algebras and rigid tree algebras.

5. Decidability and undecidability questions

A basic result of Tarski is that the elementary theory of Boolean algebras is decidable. Even the theory of Boolean algebras with a distinguished ideal is decidable. On the other hand, the theory of a Boolean algebra with a distinguished subalgebra is undecidable. Both the decidability results and undecidablity results extend in various ways to Boolean algebras in extensions of first-order logic.

6. Lindenbaum-Tarski algebras

A very important construction, which carries over to many logics and many algebras other than Boolean algebras, is the construction of a Boolean algebra associated with the sentences in a first-order theory. Let *T* be a first-order theory in a first-order language *L*. We call formulas φ and ψ equivalent provided that $T \vdash \varphi \leftrightarrow \psi$. The equivalence class of a sentence φ is denoted by $[\varphi]$. Let *A* be the collection of all equivalence classes under this equivalence relation. We can make *A* into a BA by the following definitions, which are easily justified:

$$[\varphi] + [\psi] = [\varphi \vee \psi]$$

$$[\varphi] \cdot [\psi] = [\varphi \wedge \psi]$$

$$-[\varphi] = [\neg \varphi]$$

$$0 = [F]$$

$$1 = [T]$$

Every BA is isomorphic to a Lindenbaum-Tarski algebra. However, one of the most important uses of these classical Lindenbaum-Tarski algebras is to describe them for important theories (usually decidable theories). For countable languages this can be done by describing their isomorphic interval algebras. Generally this gives a thorough knowledge of the theory. Some examples are:

	Theory	**Isomorphic to interval algebra on**
(1)	essentially undecidable theory	**Q**, the rationals
(2)	BAs	$\mathbb{N} \times \mathbb{N}$, square of the positive integers, ordered lexicographically
(3)	linear orders	**A** $\times$ **Q** ordered antilexicographically, where **A** is $\mathbb{N}$ to the $\mathbb{N}$ power in its usual order
(4)	abelian groups	(**Q** + **A**) $\times$ **Q**

7. Boolean-valued models

In model theory, one can take values in any complete BA rather than the two-element BA. This Boolean-valued model theory was developed around 1950--1970, but has not been worked on much since. But a special case, Boolean-valued models for set theory, is very much at the forefront of current research in set theory. It actually forms an equivalent way of looking at the forcing construction of Cohen, and has some technical advantages and disadvantages. Philosophically it seems more satisfactory than the forcing concept. We describe this set theory case here; it will then become evident why only compete BAs are considered. Let B be a complete BA. First we define the Boolean valued universe $V(B)$. The ordinary set-theoretic universe can be identified with $V(2)$, where 2 is the 2-element BA. The definition is by transfinite recursion, where α, β are ordinals and λ is a limit ordinal:

$V(B, 0) = \varnothing$

$V(B, \alpha + 1) =$ the set of all functions f such that the domain of f is a subset of $V(B, \alpha)$ and the range of f is a subset of B

$V(B, \lambda) =$ the union of all $V(B, \beta)$ for $\beta < \lambda$.

The B-valued universe is the proper class $V(B)$ which is the union of all of these Vs. Next, one defines by a rather complicated transfinite recursion over well-founded sets the value of a set-theoretic formula with elements of the Boolean valued universe assigned to its free variables

$$||x \in y|| = \sum \{(||x = t|| \cdot y(t)) : t \in \text{domain}(y)\}$$

$$||x \subseteq y|| = \prod \{-x(t) + ||t \in y|| : t \in \text{domain}(x)\}$$

$$||x = y|| = ||x \subseteq y|| \cdot ||y \subseteq x||$$

$$\|\neg\varphi\| = -\|\varphi\|$$

$$\|\varphi \vee \psi\| = \|\varphi\| + \|\psi\|$$

$$\|\exists x\varphi(x)\| = \sum\{\|\varphi(a)\| : a \in V(B)\}$$

Bibliography

- Halmos, P., 1963, *Lectures on Boolean Algebras*, Princeton: Van Nostrand
- Heindorf, L., and Shapiro, L., 1994, *Nearly projective Boolean algebras*, Lecture Notes in Mathematics no. 1596, Berlin: Springer-Verlag
- Jech, T., 1997, *Set Theory*, 2nd corrected edition, Berlin, New York: Springer-Verlag
- Monk, J. D., and Bonnet, R., (eds), 1989, *Handbook of Boolean algebras*, 3 volumes, Amsterdam: North-Holland.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

algebra | Boolean algebra

J. Donald Monk
Don.Monk@colorado.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 5, 2002
Content last modified: July 5, 2002

Species

The nature of species is controversial in biology and philosophy. Biologists disagree on the definition of the term "species." Philosophers disagree over the ontological status of species. A proper understanding of species is important for a number of reasons. Species are the fundamental taxonomic units of biological classification. Environmental laws are framed in terms of species. Even our conception of human nature is affected by our understanding of species. In this entry, three philosophical issues concerning species are discussed. The first is the ontological status of species. The second is whether biologists should be species pluralists or species monists. The third is whether the theoretical term "species" refers to a real category in nature.

Overview

What are biological species? At first glance, this seems like an easy question to answer. *Homo sapiens* is a species, and so is *Canis familaris* (dog). Many species can be easily distinguished. When we turn to the technical literature on species, the nature of species becomes much less clear. Biologists offer a dozen definitions of the term "species" (Claridge, Dawah, and Wilson 1997). These definitions are not fringe accounts of species but prominent definitions in the current biological literature. Philosophers also disagree on the nature of species. Here the concern is the ontological status of species. Some

philosophers believe that species are natural kinds. Others maintain that species are particulars or individuals.

The concept of species plays an important role both in and outside of biology. Within biology, species are the fundamental units of biological classification. Species are also units of evolution — groups of organisms that evolve in a unified way. Outside of biology, the concept of species plays a role in debates over environmental law and ecological preservation. Our conception of species even affects our understanding of human nature. From a biological perspective, humans are the species *Homo sapiens*.

This entry discusses three philosophical issues concerning species. The first issue is their ontological status. Are species natural kinds, individuals, or sets? The second issue concerns species pluralism. Monists argue that biologists should attempt to find the correct definition of "species." Pluralists disagree. They argue that there is no single correct definition of "species" but a plurality of equally correct definitions. The third issue concerns the reality of species. Does the term "species" refer to a real category in nature? Or, as some philosophers and biologists argue, is the term "species" a theoretically empty designation?

The Ontological Status of Species

The Death of Essentialism

Since Aristotle, species have been paradigmatic examples of natural kinds with essences. An essentialist approach to species makes perfect sense in a pre Darwinian context. God created species and an eternal essence for each species. After God's initial creation, each species is a static, non evolving group of organisms. Darwinism offers a different view of species. Species are the result of speciation. No qualitative feature — morphological, genetic, or behavioral — is considered essential for membership in a species. Despite this change in biological thinking, many philosophers still believe that species are natural kinds with essences. Let us start with a brief introduction to kind essentialism and then turn to the biological reasons why species fail to have essences.

Kind essentialism has a number of tenets. One tenet is that all and only the members of a kind have a common essence. A second tenet is that the essence of a kind is responsible for the traits typically associated with the members of that kind. For example, gold's atomic structure is responsible for gold's disposition to melt at certain temperatures. Third, knowing a kind's essence helps us explain and predict those properties typically associated with a kind. The application of any of these tenets to species is problematic. But to see the failure of essentialism we need only consider the first tenet.

Biologists have had a hard time finding biological traits that occur in all and only the members of a species. Even such pre Darwinian essentialists as Linnaeus could not locate the essences of species (Ereshefsky 2001). Evolutionary theory explains why. A number of forces conspire against the universality and uniqueness of a trait in a species (Hull 1965). Suppose a genetically based trait were

found in all the members of a species. The forces of mutation, recombination and random drift can cause the disappearance of that trait in a future member of the species. All it takes is the disappearance of a trait in one member of a species to show that it is not essential. The universality of a biological trait in a species is fragile.

Suppose, nevertheless, that a trait occurs in all the members of a species. That trait is the essence of a species only if it is unique to that species. Yet organisms in different species often have common characteristics. Again, biological forces work against the uniqueness of a trait within a single species. Organisms in related species inherit similar genes and developmental programs from their common ancestors. These common stores of developmental resources cause a number of similarities in the organisms of different species. Another source of similar traits in different species is parallel evolution. Species frequently live in similar habitats with comparable selection pressures. Those selection pressures cause the prominence of similar traits in more than one species. The parallel evolution of opposable thumbs in primates and pandas is an example.

The existence of various evolutionary forces does not rule out the possibility of a trait occurring in all and only the members of a species. But consider the conditions such a trait must satisfy. A species" essential trait must occur in all the members of a species for the entire life of that species. Moreover, if that trait is to be unique to that species, it cannot occur in any other species for the entire existence of life on this planet. The temporal parameters that species essentialism must satisfy are quite broad. The occurrence of a biological trait in all and only the members of a species is an empirical possibility. But given current biological theory, that possibility is unlikely.

Other arguments have been mustered against species essentialism. Hull (1965) contends that species have vague boundaries and that such vagueness is incompatible with the existence of species specific essences. According to Hull, essentialist definitions of natural kinds require strict boundaries between those kinds. But the boundaries between species are vague. In all but a few cases, speciation is a long and gradual process such that there is no principled way to draw a precise boundary between one species and the next. As a result, species cannot be given essentialist definitions. (Hull's argument against species essentialism is very similar to one of Locke's (1894[1975], III, vi) arguments against kind essentialism.)

Sober (1980) raises a different objection to species essentialism. He illustrates how essentialist explanations have been replaced by evolutionary ones. Essentialists explain variation within a species as the result of interference in the ontogenetic development of a species" organisms. Organisms have species specific essences, but interference often prevents the manifestations of those essences. Contemporary geneticists offer a different explanation of variation within a species. They cite the gene frequencies of a species as well as the evolutionary forces that affect those frequencies. No species specific essences are posited. Contemporary biology can explain variation within a species without positing a species" essence. So according to Sober, species essentialism has become theoretically superfluous.

In a pre Darwinian age, species essentialism made sense. Such essentialism, however, is out of step with

contemporary evolutionary theory. Evolutionary theory provides its own methods for explaining variation within a species. It tells us that the boundaries between species are vague. And it tells us that a number of forces conspire against the existence of a trait in all and only the members of a species. From a biological perspective, species essentialism is no longer a plausible position.

Species as Individuals

Let us turn to the prevailing view of the ontological status of species. Ghiselin (1974) and Hull (1978) suggest that instead of viewing species as natural kinds we should think of them as individuals. Hull draws the ontological distinction this way. (Instead of talking of "natural kinds," Hull uses the term "classes.") Classes are groups of entities that can function in scientific laws. One requirement of such laws is that they are true at any time and at any place in the universe. If "All water freezes at 0?C" is a law, then that law is true here and now, as well as 100,000 years ago on some distant planet. Water is a class because samples of water are spatiotemporally unrestricted —water can occur anywhere in the universe. Individuals, unlike classes, consist of parts that are spatiotemporally restricted. Think of a paradigmatic individual, a single mammalian organism. The parts of that organism cannot be scattered around the universe at different times if they are parts of a living, functioning organism. Various biological processes, such as digestion and respiration, require that those parts be causally and spatiotemporally connected. The parts of such an organism can only exist in a particular space-time region. In brief, individuals consist of parts that are spatiotemporally restricted. Classes consist of members that are spatiotemporally unrestricted.

Given the class/individual distinction, Ghiselin and Hull argue that species are individuals, not classes. Their argument assumes that the term "species" is a theoretical term in evolutionary theory. So their argument concerning the ontological status of species focuses on the role of "species" in evolutionary biology. Here is Hull's version of the argument, which can be dubbed the "evolutionary unit argument."

The Evolutionary Unit Argument: Since Darwin, species have been considered units of evolution. When Hull and others assert that species are units of evolution, they do not simply mean that the gene frequencies of a species change from one generation to the next. They have a more significant form of evolution in mind, namely a trait going from being rare to being prominent in a species. A classic example of such evolution is the change in coloration of peppered moths in Nineteenth Century England. Prior to the industrial revolution, peppered moths were light gray with black specks. During the industrial revolution, selection caused peppered moths to become coal black.

A number of processes can cause a trait to become prominent in a species. Hull highlights selection. Selection causes a trait to become prominent in a species only if that trait is passed down from one generation to the next. If a trait is not heritable, the frequency of that trait will not increase cumulatively. Hereditary relations, genetic or otherwise, require the generations of a species to be causally connected. Reproduction requires the generations of a species to be causally and hence spatiotemporally connected. So, if species are to evolve in non trivial ways by natural selection, they must be spatiotemporally continuous entities. Given that species are units of evolution, species are individuals and not classes.

The conclusion that species are individuals has a number of interesting implications. For one, the relationship between an organism and its species is not a member/class relation but a part/whole relation. An organism belongs to a particular species only if it is appropriately causally connected to the other organisms in that species. The organisms of a species must be parts of a single evolving lineage. If belonging to a species turns on an organism's insertion in a lineage, then qualitative similarity can be misleading. Two organisms may be very similar morphologically, genetically, and behaviorally, but unless they belong to the same spatiotemporally continuous lineage they cannot belong to the same species. Think of an analogy. Being part of my immediate family turns on my wife, my children and I having certain biological relations to one another, not our having similar features. It does not matter that my son's best friend looks just like him. That friend is not part of our family. Similarly, organisms belong to a particular species because they are appropriately causally connected, not because they look similar (if they indeed do).

Another implication of the species are individuals thesis concerns our conception of human nature (Hull 1978). As we have seen, species are first and foremost genealogical lineages. An organism belongs to a species because it is part of a lineage not because it has a particular qualitative feature. Humans may be a number of things. One of them is being the species *Homo sapiens*. From an evolutionary perspective, there is no biological essence to being a human. There is no essential feature that all and only humans must have to be part of *Homo sapiens*. Humans are not essentially rational beings or social animals or ethical agents. An organism can be born without any of these features and still be a human. From a biological perspective, being part of the lineage *Homo sapiens* is both necessary and sufficient for being a human. (For further implications of the individuality thesis, see Hull 1978.)

Responses to the Individuality Thesis

Some philosophers think that Hull and Ghiselin too quickly dismiss the assumption that species are natural kinds. Kitcher (1984) believes that species are sets of organisms. Thinking of species as sets is an ontologically neutral stance. It allows that some species are spatiotemporally restricted sets of organisms, that is, individuals. And it allows that other species are spatiotemporally unrestricted sets of organisms.

Why does Kitcher believe that some species are individuals and other species are spatiotemporally unrestricted sets? Following the biologist Ernst Mayr, Kitcher suggests that there are two fundamental types of explanation in biology: those that cite *proximate* causes and those that cite *ultimate causes*. Proximate explanations cite the more immediate cause of a trait, for example, the genes or developmental pathways that cause the occurrence of a trait in an organism. Ultimate explanations cite the evolutionary cause of a trait in a species, for example, the selection forces that caused the evolution of thumbs in pandas and their ancestors.

For each type of explanation, Kitcher believes that there are corresponding definitions of the term "species" (what biologists call 'species concepts"). Proximate explanations cite species concepts based on structural similarities, such as genetic, chromosomal and developmental similarities. These species concepts assume that species are spatiotemporally unrestricted sets of organisms. Ultimate explanations

cite species concepts that assign species evolutionary roles. These species concepts assume that species lineages and thus individuals.

Kitcher is correct that biologists attempt to explain the traits of organisms in two ways: sometimes they cite the ultimate, or evolutionary, cause of a trait; other times they cite a structural feature of an organism with that trait. A problem with Kitcher's approach is his characterization of biological practice. Biologists since Darwin have taken species to be evolutionary units. A glance at a biology text book will reveal that the evolutionary approach to species is the going concern in biology. The groups that correspond to Kitcher's structural concepts are not considered species by taxonomists. Groups of organisms that have genetic, developmental, behavioral and ecological similarities, are natural kinds in biology, but they are not considered species. Consider such groups of organisms as males, females, tree nesters and diploid organisms. These groups of organisms cut across species. For instance, some but not all humans are males and many organisms in other species are males. Male is a kind in biology, but it is not a species. Kitcher's motivation for asserting that species are sets is to allow spatiotemporally unrestricted groups of organisms to form species. That motivation, however, is not substantiated by biological theory or practice.

A more recent account of species as natural kinds is found in Boyd (1999), Griffiths (1999), and Wilson (1999). Their approach to species relies on Boyd's theory of natural kinds. According to Boyd, natural kinds are *homeostatic property cluster kinds*. The members of such a kind share a number of co-occurring properties that can be cited in prediction and explanation. The co-occurrence of such properties is due to a kind's causal homeostatic mechanisms. Turning to species, the organisms in *Canis familaris* share a number of similar properties that are sufficiently stable for use in explanation and prediction. The stability of those properties are the result of such causal homeostatic mechanisms as gene flow, stabilizing pressure and developmental homeostasis.

Boyd's approach to natural kinds is distinct from the traditional essentialist approach to kinds in several ways. First, membership in a kind does not require the occurrence of a universal and unique property in all the members of a kind. Thus Boyd's account allows that species can be natural kinds without requiring that the members of a species share a qualitative essence. Second, Boyd's account allows species to vary at a time and over time. Species can be natural kinds even though they evolve. There is some limit to the variability allowed in a species, however. Species must be sufficiently stable so that better than chance predictions can be made about some of the properties of a species.

A third way that Boyd's account of natural kinds differs from the traditional account is that Boyd allows natural kinds to be spatiotemporally restricted entities. Natural kinds, in other words, can be individuals, so long as the members of a natural kind are sufficiently stable to allow prediction and explanation. Boyd's account brings species back into the fold of natural kinds by allowing that species can be *both* natural kinds and individuals. Species are natural kinds because species are homeostatic property cluster kinds. Species are individuals because one of their homeostatic mechanisms is genealogy.

Boyd's approach to species brings unity to the debate over the ontological status of species. Such unity is

appealing if one does not mind deflating the distinction between kinds and individuals. On Boyd's approach, a particular species, even a particular human (Boyd 1999, 163), is both an individual and a natural kind. Bill Clinton's arms are parts of the individual Bill Clinton. And Bill Clinton's arms are members of the natural kind Bill Clinton. If the idea that Bill Clinton is a natural kind seems unproblematic, then Boyd's assertion that species are both individuals and natural kinds should be just the ontological ticket.

Species Pluralism

Biologists offer various definitions of the term "species" (Claridge, Dawah, and Wilson 1997). Biologists call these different definitions 'species concepts." The Biological Species Concept defines a species as a group of organisms that can successfully interbreed and produce fertile offspring. The Phylogenetic Species Concept (which itself has multiple versions) defines a species as a group of organisms bound by a unique ancestry. The Ecological Species Concept defines a species as a group of organisms that share a distinct ecological niche. These species concepts are just three of a dozen prominent species concepts in the biological literature.

What are we to make of this variety of species concepts? Monists believe that an aim of biological taxonomy is to identify the single correct species concept. Perhaps that concept is among the species concepts currently proposed and we need to determine which concept is the right one. Or perhaps we have not yet found the correct species concept and we need to wait for further progress in biology. Pluralists take a different stand. They do not believe that there is a single correct species concept. Biology, they argue, contains a number of legitimate species concepts. Pluralists believe that the monist's goal of a single correct species concept should be abandoned.

Varieties of Pluralism

Species pluralism comes in various forms (for example, Kitcher 1984, Mishler and Brandon 1987, Dupré 1993, and Ereshefsky 2001). Kitcher and Dupré offer forms of species pluralism that recognize the species concepts mentioned above — biological species, phylogenetic species, and ecological species — as well as other species concepts. As we saw in Section 1.2, Kitcher accepts species concepts that require species to be individuals, and he accepts species concepts based on the structural similarities of organisms. The latter type of species are not spatiotemporally continuous entities. Such species merely need to contain organisms that share theoretically significant properties. Dupré's version of species pluralism is more robust. He recognizes all of the species concepts found in Kitcher's version of pluralism. Dupré's pluralism also allows species concepts based on similarities highlighted by non biologists. For example, Dupré accepts species concepts based on gastronomically significant properties.

If one thinks that the term "species" is a theoretical term found within evolutionary biology, then one might find Dupré's version of pluralism too promiscuous. If the question is how the term "species" is defined in biology, then how it is defined outside of biology does not count. Think of a parallel situation

in physics. When we are interested in the scientific meaning of the term "work" we do not attend to its meaning in the sentence "How was work today?" Similarly, the use of the word "species" by culinary experts does not reveal the theoretical meaning of "species."

Kitcher's pluralism is more circumspect: it limits species concepts to those that are legitimized by theoretical biology. Still, one might worry that Kitcher's form of pluralism is too liberal. Kitcher's pluralism allows that some species are spatiotemporally continuous entities (individuals), while other species may be spatiotemporally unrestricted entities (natural kinds). As we saw in Section 2.1, Hull's evolutionary unit argument states that within the purview of evolutionary biology, species must be individuals. Kitcher's pluralism does not satisfy this requirement: some species can be non individuals. If one assumes that "species" is a theoretical term in evolutionary theory and that species are individuals, then Kitcher's pluralism is too inclusive.

Another version of species pluralism is found in Ereshefsky (2001). This version of pluralism adopts Hull's conclusion that species must be spatiotemporally continuous lineages. Nevertheless, this version of pluralism asserts that there are different types of lineages called "species." The Biological Species Concept and related concepts highlight those lineages bound by the process of interbreeding. The Phylogenetic Species Concepts highlight those lineages of organisms that share a common and unique ancestry. Ecological approaches to species highlight lineages of organisms that are exposed to common sets of stabilizing selection. On this form of species pluralism, the tree of life is segmented by different processes into different types of species lineages.

It is worth noting that the motivation behind Dupré's, Kitcher's and Ereshefsky's versions of pluralism is ontological not epistemological. Some authors (for example, Rosenberg 1994) suggest that we adopt pluralism because of our epistemological limitations. The world is exceedingly complex and we have limited cognitive abilities, so we should accept a plurality of simplified and inaccurate classifications of the world. The species pluralism offered by Dupré, Kitcher, and Ereshefsky is not epistemologically driven. Evolutionary theory, a well substantiated theory, tells us that the organic world is multifaceted. According to Dupré, Kitcher, and Ereshefsky, species pluralism is a result of a fecundity of biological forces rather than a paucity of scientific information.

Responses to Pluralism

Not everyone is willing to accept species pluralism. Monists (for example, Sober 1984, Ghiselin 1987, Hull 1987) have launched a number of objections to species pluralism. One objection centers on the type of lineage that should be accepted as species. Some monists allow the existence of different types of base lineages but contend that only one type of lineage should be called "species" (Ghiselin 1987). Supporters of the Biological Species Concept and related concepts believe that lineages of interbreeding sexual organisms are much more important in the evolution of life on this planet (Eldredge 1985). They argue that only the Biological Species Concept, or some interbreeding concept, should be accepted.

Adopting only an interbreeding approach to species has its costs: it would exclude all asexual organisms

from forming species. Interbreeding requires the genetic contributions of two sexual organisms. Asexual organisms reproduce by themselves, either through cloning, vegetative means or self fertilization. Some reptiles and amphibians reproduce asexually. Many insects reproduce asexually. And asexuality is rampant in plants, fungi and bacteria. In fact, asexual reproduction is the prominent form of reproduction on Earth (Hull 1988). If one adopts an interbreeding approach to species, then most organisms do not form species. This seems a high price to pay for species monism.

Another objection to species pluralism is that pluralism is an overly liberal position (Sober 1984, Ghiselin 1987, Hull 1987). Pluralists allow a number of legitimate species concepts, but how do pluralists determine which concepts should be accepted as legitimate? Should any species concept proposed by a biologist be accepted? What about those concepts proposed by non biologists? Without criteria for determining the legitimacy of a proposed species concept, species pluralism boils down to a position of anything goes.

Species pluralists respect this objection and attempt to respond to it (Dupré 1993, Ereshefsky 2001). They have suggested criteria for judging the legitimacy of a proposed species concept. Such criteria can be used to determine which species concepts should be accepted into the plurality of legitimate species concepts. Candidate criteria are the epistemic virtues that scientists typically use for determining the scientific worthiness of a theory. For example, in judging a species concept, one might ask if the theoretical assumptions of a concept are empirically testable. The Biological Species Concept relies on the assumption that interbreeding causes the existence of stable lineages. It also assumes that organisms that cannot interbreed do not form stable lineages. Whether interbreeding and only interbreeding causes the existence of stable lineages is empirically testable. So the Biological Species Concept has the virtue of empirical sensitivity. Other criteria for judging species concepts include intertheoretic coherence and internal consistency.

The point here is not to bring out an array of proposed species concepts and show how such criteria work in action (Dupré 1993 and Ereshefsky 2001 perform that task). The point is to show that pluralists can provide criteria for discerning which species concepts should be accepted as legitimate. If pluralists can successfully provide such criteria, then the anything goes objection to pluralism is answered.

Does the Species Category Exist?

There is one other item concerning species pluralism worth discussing. Suppose one accepts species pluralism. The term "species" then refers to a variety of different types of lineages. Some species are groups of interbreeding organisms, other species are groups of organisms that share a common ecological niche, and still other species are phylogenetic units. Given that there are different types of species, one might wonder what feature causes these different types of species to be species?

Perhaps they share a common property that renders them species. If one adopts the thesis that all species are genealogical lineages, then a common feature of species is their being lineages. However, this feature is also shared by other types of taxa in the Linnaean Hierarchy. From an evolutionary perspective, all

taxa, whether they be species, genera, or tribes, are genealogical lineages. We need to locate a feature that is not only common in species but also distinguishes species from other types of taxa.

Biological taxonomists often talk in terms of the patterns and processes of evolution. Perhaps there is a process or a pattern that occurs in species but not in other types of taxa. Such a process or pattern would unify the types of lineages we call "species." Let us start with process. The Biological Species Concept highlights those species bound by the process of interbreeding. The Ecological Species Concept identifies those species unified by stabilizing selection. The species highlighted by Phylogenetic Species Concepts are unified by such historical processes as genetic and developmental homeostasis. A survey of these different species concepts reveals that species are bound by different types of processes. So no single type of process is common to all species. Arguably, none of these processes are unique to species either (Mishler and Donoghue 1982).

What about pattern? Do species display a pattern that distinguishes them from other types of taxa? If by pattern we mean ontological structure, then species have different patterns. Species are individuals, but they are different types of individuals. Species of asexual organisms and species of sexual organisms have different structures. Both types of species contain organisms that are genealogically connected to a common ancestor. But the organisms in a sexual species are also connected by interbreeding. Thus species of sexual organisms form causally integrated entities: within a given generation, their members exchange genetic material through sexual reproduction. Species of asexual organisms do not form causally integrated entities: their organisms are merely connected to a common ancestor.

There are other suggestions for the common and unique pattern of species. Many observe that the organisms of a species often look the same or that the organisms of a species share a cluster of reoccurring properties. To the extent that this is true, it is also true of genera and some other higher taxa. The members of some genera tend to look the same and have a cluster of stable properties. Another suggestion for the pattern that distinguishes species is their ability to evolve as a unit — species are the units of evolution, other types of taxa are not. But again, many higher taxa have such unity as well (Mishler and Donoghue 1982).

The above survey of candidate unifying features is far from exhaustive. But the result is clear enough. Species vary in their unifying processes and ontological structure. Furthermore, many features that biologists and philosophers highlight as unique to species occur in many higher taxa as well. Given this survey, what position should we adopt concerning the nature of species? There are several options. According to one option we should keep looking for the unifying feature of species. This is the option favored by some monists (for example, Sober 1984). Contemporary biology may not have discovered the unifying feature of species, but that does not mean that biology will not find such a feature in the future. To give up the search for the unifying nature of species would be too hasty.

Another option starts with the assumption that the search for the unifying feature of species has gone on long enough. Biologists have looked long and hard for the correct definition of "species." The result of that search is not that we do not know what species are. The result is that the organic world contains

different types of species. The conclusion drawn by some pluralists (Kitcher 1984, Dupré 1993) is that the term "species" should be given a disjunctive definition. Species are either interbreeding lineages, or ecological lineages, or phylogenetic units, or? .

A third option, like the previous one, assumes that biologists have looked long enough for the unifying feature of species. In that search, biology has learned that there are different types of lineages called "species." But proponents (Ereshefsky 1998) of this option do not opt for a disjunctive definition of "species." According to this option, we should doubt the very existence of the category species. Those lineages we call "species" vary in their patterns and processes. Furthermore, the distinction between species and other types of taxa is riddled with vagueness. Consequently, we should doubt whether the term "species" refers to a real category in nature.

To better understand this third option it is useful to see more precisely what is being doubted. Biologists make a distinction between the *species category* and *species taxa*. Species taxa are the individual lineages we call "species." *Homo sapiens* and *Canis familaris* are species taxa. The species category is a more inclusive entity. The species category is the class of all species taxa. The third option does not call into question the existence of *Homo sapiens* or *Canis familaris* or any other lineage that we call "species." The third option just calls into question the existence of the categorical rank of species.

Darwin may have had this third option in mind when he wrote his friend Joseph Hooker:

> It is really laughable to see what different ideas are prominent in various naturalists' minds, when they speak of "species"; in some, resemblance is everything and descent of little weight — in some, resemblance seems to go for nothing, and Creation the reigning idea — in some, sterility an unfailing test, with others it is not worth a farthing. It all comes, I believe, from trying to define the indefinable (December 24, 1856; in F. Darwin 1887, vol. 2, 88.).

Darwin considers the term "species" indefinable. He could mean a couple of things by this. Perhaps Darwin meant that the term "species" is an indefinable primitive term in evolutionary theory though he still believed in the existence of the species category. Or perhaps Darwin doubted the existence of the species category altogether. According to Beatty (1985), Darwin doubted the existence of the species category. There is also evidence that he doubted the existence of the other Linnaean categories (Ereshefsky 2001). Still, Darwin believed in the existence of those lineages we call "species.' He just doubted whether the species category and the other Linnaean categories — the grid we place on the tree of life — exists in nature.

This encyclopedia entry started with the observation that at an intuitive level the nature of species seems fairly obvious. But a review of the technical literature reveals that our theoretical understanding of species is far from settled. The debate over the nature of species involves a number of issues. One issue is their ontological status: are species natural kinds or individuals or both? A second issue concerns pluralism: should we adopt species monism or species pluralism? A third issue, and perhaps the most

fundamental issue, is whether the term "species" refers to a real category in nature.

Bibliography

- Beatty, J.,1985, "Speaking of Species: Darwin's Strategy", in *The Darwinian Heritage*, D. Kohn (ed.), Princeton: Princeton University Press.
- Boyd, R., 1999, "Homeostasis, Species, and Higher Taxa", in *Species: New Interdisciplinary Studies*, R. Wilson (ed.), Cambridge, Massachusetts: MIT Press.
- Claridge, M., Dawah, H., and Wilson, R, (eds.), 1997, *Species: The Units of Biodiversity*, London: Chapman and Hall.
- Darwin, F., (ed.), 1877, *The Life and Letters of Charles Darwin, including an Autobiographical Chapter*, London: John Murray.
- Dupré, J., 1993, *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*, Cambridge, Massachusetts: Harvard University Press.
- Eldredge, N., 1985 *Unfinished Synthesis*, New York: Oxford University Press.
- Ereshefsky, M.,1998, "Species Pluralism and Anti-Realism", *Philosophy of Science*, 65:103-120.
- Ereshefsky, M., 2001, *The Poverty of the Linnaean Hierarchy: A Philosophical Study of Biological Taxonomy*, Cambridge: Cambridge University Press.
- Ghiselin, M., 1974, "A Radical Solution to the Species Problem", *Systematic Zoology*, 23:536-544
- Ghiselin, M., 1987, "Species Concepts, Individuality, and Objectivity", *Biology and Philosophy*, 2:127-143.
- Griffiths, P., 1999, "Squaring the Circle: Natural Kinds with Historical Essences", in *Species: New Interdisciplinary Studies, R. Wilson* (ed.), Cambridge, Massachusetts: MIT Press.
- Hull, D., 1965, "The Effect of Essentialism on Taxonomy: Two Thousand Years of Stasis", *British Journal for the Philosophy of Science*, 15:314-326, 16:1-18.
- Hull, D., 1978, "A Matter of Individuality", *Philosophy of Science*, 45:335-360.
- Hull, D.,1987, "Genealogical Actors in Ecological Roles", *Biology and Philosophy*, 2:168-183.
- Hull, D., 1988 , *Science as a Process*, Chicago: University of Chicago Press.
- Kitcher, P., 1984, "Species", *Philosophy of Science*, 51: 308-333.
- Locke, J., 1894[1975], *An Essay Concerning Human Understanding*, P. Nidditch (ed.), New York: Oxford University Press.
- Mishler, B. and Brandon, R., 1987, "Individuality, Pluralism, and the Phylogenetic Species Concept", *Biology and Philosophy*, 2: 397-414.
- Mishler, B. and Donoghue, M., 1982, "Species Concepts: A Case for Pluralism", *Systematic Zoology*, 31: 491-503.
- Rosenberg, A.,1994, *Instrumental Biology or the Disunity of Science*, Chicago: Chicago University Press.
- Sober, E., 1980, "Evolution, Population Thinking and Essentialism", *Philosophy of Science*, 47:350-383.
- Sober, E., 1984, "Sets, Species, and Natural Kinds: A Reply to Philip Kitcher's ‘Species’", *Philosophy of Science,51:334-341*.
- Wilson, R., 1999, "Realism, Essence, and Kind: Resuscitating Species Essentialism?", in *Species:*

New Interdisciplinary Studies, R. Wilson (ed.), Cambridge, Massachusetts: MIT Press.

Other Internet Resources

- What is a Species and What is Not?, by Ernst Mayr
- Systematic Lectures, by John Heraty and Richard Whitkus

Related Entries

essentialism | evolution | individual | natural kinds | ontology | pluralism: in biology | scientific realism

Marc Ereshefsky
ereshefs@ucalgary.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 3, 2002
Content last modified: July 3, 2002

Confucius

Confucius (551-479 BCE), according to Chinese tradition, was a thinker, political figure, educator, and founder of the *Ru* School of Chinese thought. His teachings, preserved in the *Analects*, form the foundation of much of subsequent Chinese speculation on the education and comportment of the ideal man, how such an individual should live his live and interact with others, and the forms of society and government in which he should participate. Fung Yu-lan, one of the great 20th century authorities on the history of Chinese thought, compares Confucius' influence in Chinese history with that of Socrates in the West.

1. Confucius' Life

The sources for Confucius' life are later and do not carefully separate fiction and fact. Thus it is wise to regard much of what is known of him as legendary. Many of the legends surrounding Confucius at the end of the 2nd century BCE were included by the Han dynasty court historian, Sima Qian (145-c.85 BCE), in his well-known and often-quoted *Records of the Grand Historian* (*Shiji*). This collection of tales opens by identifying Confucius' ancestors as members of the Royal State of Song. It notes as well that his great grandfather, fleeing the turmoil in his native Song, had moved to Lu, somewhere near the present town of Qufu in southeastern Shandong, where the family became impoverished. Confucius is described, by Sima Qian and other sources, as having endured a poverty-stricken and humiliating youth and been forced, upon reaching manhood, to undertake such petty jobs as accounting and caring for livestock. Sima Qian's account includes the tale of how Confucius was born in answer to his parents' prayers at a sacred hill (*qiu*) called Ni. Confucius' surname *Kong* (which means literally an utterance of thankfulness when prayers have been answered), his tabooed given name *Qiu*, and his social name

Zhongni, all appear connected to the miraculous circumstances of his birth. This casts doubt, then, on Confucius' royal genealogy as found in Sima Qian. Similarly, Confucius' recorded age at death, ‘seventy-two,’ is a ‘magic number’ with far-reaching significance in early Chinese literature. We do not know how Confucius himself was educated, but tradition has it that he studied ritual with the Daoist Master Lao Dan, music with Chang Hong, and the lute with Music-master Xiang. In his middle age Confucius is supposed to have gathered about him a group of disciples whom he taught and also to have devoted himself to political matters in Lu. The number of Confucius' disciples has been greatly exaggerated, with Sima Qian and other sources claiming that there were as many as three thousand of them. Sima Qian goes on to say that, “Those who, in their own person, became conversant with the Six Disciplines [taught by Confucius], numbered seventy-two.” The 4th century BCE *Mencius* and some other early works give their number as seventy. Perhaps seventy or seventy-two were a maximum, though both of these numbers are suspicious given Confucius' supposed age at death.

At the age of fifty, when Duke Ding of Lu was on the throne, Confucius' talents were recognized and he was appointed Minister of Public Works and then Minister of Crime. But Confucius apparently offended members of the Lu nobility who were vying with Duke Ding for power (or was it the duke himself that Confucius had rubbed the wrong way?) and he was subsequently forced to leave office and go into exile. As in other ancient cultures, exile and suffering are common themes in the lives of the heroes of the early Chinese tradition. In the company of his disciples, Confucius left Lu and traveled in the states of Wei, Song, Chen, Cai, and Chu, purportedly looking for a ruler who might employ him but meeting instead with indifference and, occasionally, severe hardship and danger. Several of these episodes, as preserved in the *Records of the Grand Historian*, appear to be little more than prose retellings of songs found in the ancient Chinese *Book of Songs*, Confucius' life is thus rendered a re-enactment of the suffering and alienation of the personas of the poems.

In any case, by most traditional accounts, Confucius returned to Lu in 484 BCE and spent the remainder of his life teaching, putting in order the *Book of Songs*, the *Book of Documents*, and other ancient classics, as well as editing the *Spring and Autumn Annals*, the court chronicle of Lu. Sima Qian's account also provides background on Confucius' connection to the early canonical texts on ritual and on music (the latter of which was lost at an early date). Sima Qian claims, moreover, that, “In his later Years Confucius delighted in the *Yi*”—the famous, some might say infamous, divination manual popular to this day in China and in the West. The *Analects* passage which appears to corroborate Sima Qian's claim seems corrupt and hence unreliable on this point. Confucius' traditional association with these works led them and related texts to be revered as the “Confucian Classics” and made Confucius himself the spiritual ancestor of later teachers, historians, moral philosophers, literary scholars, and countless others whose lives and works figure prominently in Chinese intellectual history.

Book X of the *Analects* consists of personal observations of how Confucius comported himself as a thinker, teacher, and official. Some have argued that these passages were originally more general prescriptions on how a gentleman should dress and behave that were relabeled as descriptions of Confucius. Traditionally, Book X has been regarded as providing an intimate portrait of Confucius and has been read as a biographical sketch. The following passages provide a few examples.

> Confucius, at home in his native village, was simple and unassuming in manner, as though he did not trust himself to speak. But when in the ancestral temple or at Court he speaks readily, though always choosing his words with due caution. (*Lunyu* 10.1)
>
> When at court conversing with the officers of a lower grade, he is friendly, though straightforward; when conversing with officers of a higher grade, he is restrained but precise. When the ruler is present he is wary, but not cramped. (*Lunyu* 10.2)
>
> On entering the Palace Gate he seems to contract his body, as though there were not sufficient room to admit him. If he halts, it must never be in the middle of the gate, nor in going through does he ever tread on the threshold. (*Lunyu* 10.4)
>
> When fasting in preparation for sacrifice he must wear the Bright Robe, and it must be of linen. He must change his food and also the place where he commonly sits. He does not object to his rice being thoroughly cleaned, nor to his meat being finely minced. (*Lunyu* 10.7, 10.8)
>
> When sending a messenger to enquire after someone in another country, he bows himself twice while seeing the messenger off. (*Lunyu* 10.15)
>
> In bed he avoided lying in the posture of a corpse … On meeting anyone in deep mourning he must bow across the bar of his chariot. (*Lunyu* 10.24, 10.25)

Analects passages such as these made Confucius *the* model of courtliness and personal decorum for countless generations of Chinese officials.

By the 4th century BCE, Confucius was recognized as a unique figure, a sage who was ignored but should have been recognized and become a king. At the end of the 4th century, Mencius says of Confucius: “Ever since man came into this world, there has never been one greater than Confucius.” And in two passages Mencius implies that Confucius was one of the great sage kings who, according to his reckoning, arises every five hundred years. Confucius also figures prominently as the subject of anecdotes and the teacher of wisdom in the writing of Xunzi, a third century BCE follower of Confucius' teachings. Indeed chapters twenty-eight to thirty of the *Xunzi*, which some have argued were not the work of Xunzi but compilations by his disciples, look like an alternative, and considerably briefer, version of the *Analects*.

Confucius and his followers also inspired considerable criticism from other thinkers. The authors of the *Zhuangzi* took particular delight in parodying Confucius and the teachings conventionally associated with him. But Confucius' reputation was so great that even the *Zhuangzi* appropriates him to give voice to Daoist teachings.

2. Confucius' Social Philosophy

Confucius' teachings and his conversations and exchanges with his disciples are recorded in the *Lunyu* or *Analects*, a collection that probably achieved something like its present form around the second century BCE. While Confucius believes that people live their lives within parameters firmly established by Heaven—which, often, for him means both a purposeful Supreme Being as well as 'nature' and its fixed cycles and patterns—he argues that men are responsible for their actions and especially for their treatment of others. We can do little or nothing to alter our fated span of existence but we determine what we accomplish and what we are remembered for.

Confucius represented his teachings as lessons transmitted from antiquity. He claimed that he was "a transmitter and not a maker" and that all he did reflected his "reliance on and love for the ancients." (*Lunyu* 7.1) Confucius pointed especially to the precedents established during the height of the royal Zhou (roughly the first half of the first millennium, BCE). Such justifications for one's ideas may have already been conventional in Confucius' day. Certainly his claim that there were antique precedents for his ideology had a tremendous influence on subsequent thinkers many of whom imitated these gestures. But we should not regard the contents of the *Analects* as consisting of old ideas. Much of what Confucius taught appears to have been original to him and to have represented a radical departure from the ideas and practices of his day.

Confucius also claimed that he enjoyed a special and privileged relationship with Heaven and that, by the age of fifty, he had come to understand what Heaven had mandated for him and for mankind. (*Lunyu* 2.4). Confucius was also careful to instruct his followers that they should never neglect the offerings due Heaven. (*Lunyu* 3.13) Some scholars have seen a contradiction between Confucius' reverence for Heaven and what they believe to be his skepticism with regard to the existence of 'the spirits.' But the *Analects* passages that reveal Confucius's attitudes toward spiritual forces (*Lunyu* 3.12, 6.20, and 11.11) do not suggest that he was skeptical. Rather they show that Confucius revered and respected the spirits, thought that they should be worshipped with utmost sincerity, and taught that serving the spirits was a far more difficult and complicated matter than serving mere mortals.

Confucius' social philosophy largely revolves around the concept of *ren*, "compassion" or "loving others." Cultivating or practicing such concern for others involved deprecating oneself. This meant being sure to avoid artful speech or an ingratiating manner that would create a false impression and lead to self-aggrandizement. (*Lunyu* 1.3) Those who have cultivated *ren* are, on the contrary, "simple in manner and slow of speech." (*Lunyu* 13.27). For Confucius, such concern for others is demonstrated through the practice of forms of the Golden Rule: "What you do not wish for yourself, do not do to others;" "Since you yourself desire standing then help others achieve it, since you yourself desire success then help others attain it." (*Lunyu* 12.2, 6.30). He regards devotion to parents and older siblings as the most basic form of promoting the interests of others before one's own and teaches that such altruism can be accomplished only by those who have learned self-discipline.

Learning self-restraint involves studying and mastering *li*, the ritual forms and rules of propriety through

which one expresses respect for superiors and enacts his role in society in such a way that he himself is worthy of respect and admiration. A concern for propriety should inform everything that one says and does:

> Look at nothing in defiance of ritual, listen to nothing in defiance of ritual, speak of nothing in defiance or ritual, never stir hand or foot in defiance of ritual. (*Lunyu* 12.1)

Subjecting oneself to ritual does not, however, mean suppressing one's desires but instead learning how to reconcile one's own desires with the needs of one's family and community. Confucius and many of his followers teach that it is by experiencing desires that we learn the value of social strictures that make an ordered society possible (See *Lunyu* 2.4.). Nor does Confucius' emphasis on ritual mean that he was a punctilious ceremonialist who thought that the rites of worship and of social exchange had to be practiced correctly at all costs. Confucius taught, on the contrary, that if one did not possess a keen sense of the well-being and interests of others his ceremonial manners signified nothing. (*Lunyu* 3.3). Equally important was Confucius' insistence that the rites not be regarded as mere forms, but that they be practiced with complete devotion and sincerity. “He [i.e., Confucius] sacrificed to the dead as if they were present. He sacrificed to the spirits as if the spirits were present. The Master said, ‘I consider my not being present at the sacrifice as though there were no sacrifice.’” (*Lunyu* 3.12)

While ritual forms often have to do with the more narrow relations of family and clan, *ren*, however, is to be practiced broadly and informs one's interactions with all people. Confucius warns those in power that they should not oppress or take for granted even the lowliest of their subjects. “You may rob the Three Armies of their commander, but you cannot deprive the humblest peasant of his opinion.” (*Lunyu* 9.26) Confucius regards loving others as a calling and a mission for which one should be ready to die (*Lunyu* 15.9).

3. Confucius' Political Philosophy

Confucius' political philosophy is also rooted in his belief that a ruler should learn self-discipline, should govern his subjects by his own example, and should treat them with love and concern. “If the people be led by laws, and uniformity among them be sought by punishments, they will try to escape punishment and have no sense of shame. If they are led by virtue, and uniformity sought among them through the practice of ritual propriety, they will possess a sense of shame and come to you of their own accord.” (*Lunyu* 2.3; see also 13.6.) It seems apparent that in his own day, however, advocates of more legalistic methods were winning a large following among the ruling elite. Thus Confucius' warning about the ill consequences of promulgating law codes should not be interpreted as an attempt to prevent their adoption but instead as his lament that his ideas about the moral suasion of the ruler were not proving popular.

Most troubling to Confucius was his perception that the political institutions of his day had completely broken down. He attributed this collapse to the fact that those who wielded power as well as those who occupied subordinate positions did so by making claim to titles for which they were not worthy. When

asked by a ruler of the large state of Qi, Lu's neighbor on the Shandong peninsula, about the principles of good government, Confucius is reported to have replied: “Good government consists in the ruler being a ruler, the minister being a minister, the father being a father, and the son being a son.” (*Lunyu* 12.11) If I claim for myself a title and attempt to participate in the various hierarchical relationships to which I would be entitled by virtue of that title, then I should live up to the meaning of the title that I claim for myself. Confucius' analysis of the lack of connection between actualities and their names and the need to correct such circumstances is usually referred to as Confucius' theory of *zhengming*. Elsewhere in the *Analects*, Confucius says to his disciple Zilu that the first thing he would do in undertaking the administration of a state is *zhengming*. (*Lunyu* 13.3). Xunzi composed an entire essay entitled *Zhengming*. But for Xunzi the term referred to the proper use of language and how one should go about inventing new terms that were suitable to the age. For Confucius, *zhengming* does not seem to refer to the ‘rectification of names’ (this is the way the term is most often translated by scholars of the *Analects*), but instead to rectifying behavior of people so that it exactly corresponds to the language with which they identify and describe themselves. Confucius believed that this sort of rectification had to begin at the very top of the government, because it was at the top that the discrepancy between names and actualities had originated. If the ruler's behavior is rectified then the people beneath him will follow suit. In a conversation with Ji Kangzi (who had usurped power in Lu), Confucius advised: “If your desire is for good, the people will be good. The moral character of the ruler is the wind; the moral character of those beneath him is the grass. When the wind blows, the grass bends.” (*Lunyu* 12.19)

For Confucius, what characterized superior rulership was the possession of *de* or ‘virtue.’ Conceived of as a kind of moral power that allows one to win a following without recourse to physical force, such ‘virtue’ also enabled the ruler to maintain good order in his state without troubling himself and by relying on loyal and effective deputies. Confucius claimed that, “He who governs by means of his virtue is, to use an analogy, like the pole-star: it remains in its place while all the lesser stars do homage to it.” (*Lunyu* 2.1) The way to maintain and cultivate such royal ‘virtue’ was through the practice and enactment of *li* or ‘rituals’—the ceremonies that defined and punctuated the lives of the ancient Chinese aristocracy. These ceremonies encompassed: the sacrificial rites performed at ancestral temples to express humility and thankfulness; the ceremonies of enfeoffment, toasting, and gift exchange that bound together the aristocracy into a complex web of obligation and indebtedness; and the acts of politeness and decorum—such things as bowing and yielding—that identified their performers as gentlemen. In an influential study, Herbert Fingarette argues that the performance of these various ceremonies, when done correctly and sincerely, involves a ‘magical’ quality that underlies the efficacy of royal ‘virtue’ in accomplishing the aims of the ruler.

4. Confucius and Education

A hallmark of Confucius' thought is his emphasis on education and study. He disparages those who have faith in natural understanding or intuition and argues that the only real understanding of a subject comes from long and careful study. Study, for Confucius, means finding a good teacher and imitating his words and deeds. A good teacher is someone older who is familiar with the ways of the past and the practices of the ancients. (See *Lunyu* 7.22) While he sometimes warns against excessive reflection and meditation,

Confucius' position appears to be a middle course between studying and reflecting on what one has learned. “He who learns but does not think is lost. He who thinks but does not learn is in great danger.” (*Lunyu* 2.15) Confucius, himself, is credited by the tradition with having taught altogether three thousand students, though only seventy are said to have truly mastered the arts he cherished. Confucius is willing to teach anyone, whatever their social standing, as long as they are eager and tireless. He taught his students morality, proper speech, government, and the refined arts. While he also emphasizes the “Six Arts” -- ritual, music, archery, chariot-riding, calligraphy, and computation -- it is clear that he regards morality the most important subject. Confucius' pedagogical methods are striking. He never discourses at length on a subject. Instead he poses questions, cites passages from the classics, or uses apt analogies, and waits for his students to arrive at the right answers. “I only instruct the eager and enlighten the fervent. If I hold up one corner and a student cannot come back to me with the other three, I do not go on with the lesson.” (*Lunyu* 7.8).

Confucius' goal is to create gentlemen who carry themselves with grace, speak correctly, and demonstrate integrity in all things. His strong dislike of the sycophantic “petty men,” whose clever talk and pretentious manner win them an audience, is reflected in numerous *Lunyu* passages. Confucius finds himself in an age in which values are out of joint. Actions and behavior no longer correspond to the labels originally attached to them. “Rulers do not rule and subjects do not serve,” he observes. (*Lunyu* 12.11; cf. also 13.3) This means that words and titles no longer mean what they once did. Moral education is important to Confucius because it is the means by which one can rectify this situation and restore meaning to language and values to society. He believes that the most important lessons for obtaining such a moral education are to be found in the canonical *Book of Songs*, because many of its poems are both beautiful and good. Thus Confucius places the text first in his curriculum and frequently quotes and explains its lines of verse. For this reason, the *Lunyu* is also an important source for Confucius' understanding of the role poetry and art more generally play in the moral education of gentlemen as well as in the reformation of society.

Bibliography

- Brooks, E. & A., 1998, *The Original Analects*, New York: Columbia University Press.
- Creel, H., 1949, *Confucius*, Harper.
- Fingarette, H., 1972, *The Secular as Sacred*, Harper.
- Knoblock, J., 1988, 1990, 1994, *Xunzi: A Translation and Study of the Complete Works* (Three Volumes), Stanford University Press.
- Lau, D. C., 1979, *Confucius: The Analects*, Harmondsworth: Penguin.
- Nivison, D., 1996, *The Ways of Confucianism*, Open Court.
- Waley, A., 1938, *The Analects of Confucius*, New York: Vintage Books.
- Yang, Bojun, 1958, *Lunyu yizhu*, Beijing: Zhonghua shuju.

Other Internet Resources

- German Website with Various Links to Sites and Materials on Confucius and Confucianism, maintained by Erling Weinrich
- Online English Translation of the Analects, the Internet Classics Archive (MIT)
- Bibliography of Chinese Philosophy, maintained by Bryan Van Norden (Vassar College)

Related Entries

Analects of Confucius | *Book of Songs* | Confucianism | Daoism [Taoism] | *De*, or virtues | Golden Rule | Heaven | Lao Dan | Laozi | Mencius | *Ren*, compassion | ritual*Ru* School of Thought | Sima Qian | social philosophy | Xunzi | *Yi*, *Book of Changes* | Zhengming or, rectification of names | Zhou dynasty | Zhuangzi

Jeffrey Riegel
jkriegel@socrates.berkeley.edu

First published: July 3, 2002
Content last modified: July 3, 2002

Laozi

Confucianism, Daoism (Taoism), and Buddhism form the three main pillars of Chinese thought, keeping in mind that they are not monolithic but multifaceted traditions with complex internal divisions. Laozi (Lao-tzu, in "Wade-Giles" romanization) flourished during the sixth century B.C.E. and was the "founder" of Daoism, according to Chinese tradition. According to some modern scholars, however, Laozi is entirely legendary; there was never an historical Laozi. Daoism appears as a school of philosophy (*daojia*) as well as a religious tradition (*daojiao*); in the latter, Laozi is revered as a supreme deity. The name "Laozi" is best taken to mean "Old (*lao*) Master (*zi*)," and Laozi the ancient philosopher is said to have written a short book, which has come to be called simply the *Laozi*. When the *Laozi* was recognized as a "classic" (*jing*) -- that is, a work of such profound insight as to merit canonical status -- it acquired a more exalted and hermeneutically instructive title, the *Daodejing* (*Tao-te ching*), commonly translated as the "Classic of the Way and Virtue." Its influence on Chinese culture is pervasive, and it reaches beyond China. Next to the Bible, the *Daodejing* is the most translated work in world literature. It is concerned with the "Way" or Dao and how it finds expression in "virtue" (*de*), especially through what the text calls "naturalness" (*ziran*) and "nonaction" (*wuwei*). These concepts, however, are open to interpretation. While some see them as proof that the *Laozi* is a deeply "mystical" work, others emphasize their contribution to ethics and/or political philosophy. Interpreting the *Laozi* demands careful hermeneutic reconstruction, which requires both analytic rigor and an informed historical imagination.

The Laozi Story

The *Shiji* (Records of the Historian) by the Han dynasty court historian Sima Qian (ca. 145-86 B.C.E.) offers a "biography" of Laozi. Its reliability has been questioned, but it serves as a common point of

departure for scholarly debate. Laozi was a native of Chu, a southern state in the Zhou dynasty (see map and discussion in Loewe and Shaughnessy 1999 (*The Cambridge History of Ancient China*), 594, 597). His surname was Li; his given name was Er, and he was also called Dan. Laozi served as a keeper of archival records at the court of Zhou. Confucius (551-479 B.C.E.) had consulted him on the rites and praised him highly (*Shiji* 63). This establishes the traditional claim that Laozi was a senior contemporary of Confucius. A meeting or meetings between Confucius and Laozi, identified as "Lao Dan," is reported also in the *Zhuangzi* and other early Chinese sources.

"Laozi cultivated Dao and virtue," as Sima Qian goes on to relate, and "his learning was devoted to self-effacement and not having fame. He lived in Zhou for a long time; witnessing the decline of Zhou, he departed." When he reached the northwest border then separating China from the outside world, Yin Xi, the official in charge of the border pass, asked that he put his thoughts to writing. The result was a book consisting of some five thousand Chinese characters, divided into two parts, which discusses "the meaning of Dao and virtue." Thereafter, Laozi left; no one knew where he had gone. This completes the main part of Sima Qian's account. The remainder puts on record attempts to identify the legendary Laozi with certain known historical individuals and concludes with a list of Laozi's purported descendants (see W. T. Chan 1963 and D. C. Lau 1963 for an English translation).

Few scholars today would subscribe fully to the *Shiji* report. Indeed, according to William Boltz, it "contains virtually nothing that is demonstrably factual; we are left no choice but to acknowledge the likely fictional nature of the traditional Lao tzu [Laozi] figure" (1993, 270). Disagreements abound on every front, including the name Laozi itself. Although the majority takes "Laozi" to mean "Old Master," some scholars believe that "Lao" is a surname. The *Zhuangzi* and other early texts refer to "Lao Dan" consistently but not "Li Er." According to Fung Yu-lan, Sima Qian had "confused" the legendary Lao Dan with Li Er, who flourished during the Warring States period (480-221 B.C.E.) and was the "real" founder of the Daoist school (1983, 171). In an influential essay, A. C. Graham (1986) argues that the story of Laozi reflects a conflation of different legends. The earliest strand revolved around the meeting of Confucius with Lao Dan and was current by the fourth century B.C.E. During the first half of the third century, Lao Dan was recognized as a great thinker in his own right and as the founder of a distinct "Laoist" school of thought. It was not until the Han dynasty (206 B.C.E.-220 C.E.), when the teachings of Laozi, Zhuangzi, and others were seen to share certain insights centering on the concept of Dao, that they were classified together under the rubric of philosophical "Daoism" (*daojia*).

It is clear that by 100 B.C.E. if not earlier Laozi was already shrouded in legends, and that Sima Qian could only exercise his judgment as an historian to put together a report that made sense to him, based on the different and sometimes competing sources at his disposal. The fact that Lao Dan appears favorably in both Confucian and Daoist sources seems to argue against the likelihood that the figure was fabricated for polemical purposes. Lao Dan could have gathered around him a group of disciples, who out of respect would address him as "Laozi." Confucius had sought his advice presumably on mourning and funeral rites, given that the Confucian work *Liji* (Record of Rites) has Confucius citing Lao Dan four times specifically on these rites. Indeed, various dates have been proposed for the encounter -- for example, 501 B.C.E., following the account in the *Zhuangzi* (ch. 14) -- about which different versions circulated later among the elite during the Warring States period, when the Confucian and "Laoist" schools had secured

their place in the intellectual arena. Admittedly, this is conjecture; what is certain is that the identity of Laozi will continue to attract and divide scholarly opinion.

The story of Laozi occupies a cherished place in the Daoist tradition. It is important also because it raises certain hermeneutic expectations and affects the way in which the *Laozi* is read. If the work was written by a single author, one might expect, for example, a high degree of consistency in style and content. If the *Laozi* was a work of the sixth or fifth century B.C.E., one might interpret certain sayings in the light of what we know of the period. There is little consensus among scholars, however, on the date or authorship of the *Laozi*, as we shall see below.

With the arrival of the "Way of the Celestial Master" (*tianshidao*), the first organized religious Daoist establishment (*daojiao*) in the second century C.E., the story of Laozi gained an important hagiographic dimension. In the eyes of the faithful, the Dao is a divine reality, and Laozi is seen as the personification of the Dao. Lao Dan is but one manifestation of the divine Laozi, albeit a pivotal one because of the writing of the *Daodejing*, which in religious Daoism commands devotion as a foundational scripture that promises not only wisdom but also immortality and salvation to those who submit to its power. During the Tang dynasty (618-906 C.E.), the imperial Li family traced its ancestry to Laozi. Today, Laozi's "birthday" is celebrated in many parts of Asia on the fifteenth day of the second lunar month.

The influence of the *Laozi* on Chinese culture is both deep and far-reaching. One indication of its enduring appeal and hermeneutical openness is the large number of commentaries devoted to it throughout Chinese history -- some seven hundred, according to one count (W. T. Chan 1963, 77). The *Laozi* has inspired an intellectual movement known as *xuanxue*, "Learning of the Mysterious (Dao)" -- or "Neo-Daoism," as some scholars prefer, emphasizing its roots in classical Daoism -- that dominated the Chinese elite or high culture from the third to the sixth century C.E. Consequently, the *Laozi* played a significant role in informing not only philosophic thought but also the development of literature, calligraphy, painting, music, and other cultural traditions.

Imperial patronage enhanced the prestige of the *Laozi* and enlarged its scope of influence. In 731 C.E., the emperor Xuanzong decreed that all officials should keep a copy of the *Daodejing* at home and placed the classic on the list of texts to be examined for the civil service examinations. In religious Daoism, recitation of the *Daodejing* is a prescribed devotional practice and figures centrally in ritual performance. The *Daodejing* has been set to music from an early time. The term "*Laozi* learning" (*Laoxue*) has come to designate an important field of study; a recent effort that sketches the major landmarks in this development is *Zhongguo Laoxue shi* (A History of *Laozi* Learning in China) (Xiong Tieji, et al. 1995).

The influence of the *Laozi* extends beyond China, as Daoism reaches across Asia and in the modern period, the Western world. In Hong Kong, Taiwan, and among the Chinese in Southeast Asia, Daoism is a living tradition. Daoist beliefs and practices have contributed also to the formation of Korean and Japanese culture, although here the process of cultural transmission, assimilation, and transformation is highly complex, especially given the close interaction among Daoism, Buddhism, and indigenous traditions such as Shintō (see Fukui, et al. 1983, vol. 3). During the seventh century, the *Laozi* was translated into Sanskrit; in the eighteenth century a Latin translation was brought to England, after which

there has been a steady supply of translations into Western languages, yielding a handsome harvest of some 250 to date (LaFargue and Pas 1998, 277).

Laozi is an "axial" philosopher whose insight helps shape the course of human development, according to Karl Jaspers (1974). Memorable phrases from the *Laozi* such as "governing a large country is like cooking a small fish" (ch. 60) have found their way into Western political rhetoric. At the popular level, several comic versions of the *Laozi* reach out to a younger and wider readership (e.g., Tsai Chih Chung, et al. 1995). Some may have come to learn about the *Laozi* through such best-selling works as *The Tao of Physics* (Capra 1975) or *The Tao of Pooh* (Hoff 1982); and there is also *A Taoist Cookbook* (Saso 1994), which comes with "meditations" from the *Daodejing*. From nature lovers to management gurus, a growing audience is discovering that the *Laozi* has something to offer to them. The reception of the *Laozi* in modern Asia and the West falls outside the scope of this article; nevertheless, it is important to note that the *Laozi* should be regarded not only as a work of early Chinese philosophy but also in a larger context as a classic of world literature with keen contemporary relevance.

Date and Authorship of the *Laozi*

The date of composition refers to the time when the *Laozi* reached more or less its final form; it does not rule out later interpolations or corruptions. Generally, three positions can be distinguished. First, some scholars maintain that we should accept on the whole Sima Qian's account that the *Laozi* was written by Lao Dan in the sixth or early fifth century B.C.E. A second and more widely held view traces the *Laozi* to the fourth century, while a third argues for an even later date, not earlier than the mid-third century B.C.E. These are general indicators; the situation is more complex because the *Laozi* may turn out to be a composite work involving a long process of textual formation.

Both external and internal considerations play a role in determining the date of the *Laozi*. Quotations from the *Laozi* in other classical works are often cited as evidence. For example, if the *Mozi* quotes from the *Laozi*, and if the *Mozi* can be dated to the fifth century, then the *Laozi* would have been current by that time. There is in fact one such quotation preserved in the Song dynasty encyclopedic work, the *Taiping yulan* (322.5b), although it is not found in the present *Mozi*. Unless new archaeological evidence comes to light, the available external evidence can only confirm that parts of the current *Laozi* were available around 300 B.C.E. (see further discussion in the next section) and that the work became widely recognized by the middle of the third century, when it was quoted extensively in such works as the *Hanfeizi* and the "outer" and "miscellaneous" chapters of the *Zhuangzi*.

The language of the *Laozi* provides important clues. Much of the text is rhymed. Focusing on rhyme patterns, Liu Xiaogan (1994 and 1997) concludes that the poetic structure of the *Laozi* is closer to that of the *Shijing* (Classic of Poetry) than that of the later *Chuci* ("Songs of the South"; David Hawkes, trans. 1985). For this reason, the traditional view first articulated by Sima Qian should be upheld. Examining a wider range of linguistic evidence, William Baxter agrees that the *Laozi* should be dated earlier than the *Zhuangzi* and the *Chuci*, but he traces "the bulk of the *Lao-tzu* to the mid or early fourth century" (1998, 249). Both Liu and Baxter provide a concise analysis of the different theories of the date of the *Laozi*.

It is possible that the *Laozi* has "preserved" the ideas of Lao Dan. W. T. Chan, for example, believes that the text "embodied" the teachings of Laozi, although it was not written until the fourth century (1963, 74). According to A. C. Graham, the *Laozi* was ascribed to Lao Dan around 250 B.C.E. by the text's author or "publiciser," capitalizing on Lao Dan's reputation (1986, 119; also see Graham 1989). This leaves open the possibility that the book or parts of it existed before the middle of the third century. It also raises the question whether the *Laozi* was the work of a single author.

Conceivably, an editor or compiler, or a group or succession of them, could have brought together diverse sources. D. C. Lau, for example, is of the view that the *Laozi* is an "anthology" (1963, 14). According to Bruce Brooks and Taeko Brooks, the *Laozi* contains different layers of material spanning the period between 340 and 249 B.C.E. -- "its long timespan precludes a single author" (1998, 151). Indeed, Chad Hansen describes the "dominant current textual theory" of the *Daodejing* as one which "treats the text as an edited accumulation of fragments and bits drawn from a wide variety of sources … there was no single author, no Laozi" (1992, 201). In contrast, Rudolf Wagner (1984 and 2000) asserts that the *Laozi* has a consistent "rhetorical structure," characterized by an intricate "interlocking parallel style," which would cast doubt on the "anthology" thesis.

The idea of an oral tradition that preceded the writing of the *Laozi* has gained wide acceptance in recent years. However, it is not always clear what that entails. On the one hand, it could lend support to W. T. Chan's view cited above, that Lao Dan's disciples had kept alive the teachings of the master orally before some later student(s) committed them to writing. On the other hand, it could also mean that the redactor(s) or compiler(s) had access to disparate sayings originated from and circulated in different contexts. As Michael LaFargue emphasizes, oral tradition need not refer to the sayings of one person; it functions rather as a reservoir of "aphorisms," which were circulated among like-minded "Laoist" scholars and formed the basis of the *Daodejing* (1992, 197). This does not prejudge whether the final product contains sayings that were put together at random, or reflects a careful distillation on the part of the redactor(s) who arranged and/or altered the material at their disposal. LaFargue appears to favor the latter view, but other scholars (e.g., Lau 1963 and Mair 1990) see little sign of tight editorial control.

Much remains uncertain. It may be argued that date and authorship are immaterial to and may detract from interpretation. The "truth" of the *Laozi* is "timeless," according to this view, transcending historical and cultural specificities. Issues of provenance are important, however, if context has any role to play in the production of meaning. Polemics among different schools of thought, for example, were far more pronounced during the Warring States period than in the earlier "Spring and Autumn" period (770-481 B.C.E.). The Zhou government had been in decline; warfare among the "feudal" states intensified both in scale and frequency from the fourth century onward. As the political conditions deteriorated, philosophers and strategists vied to convince the rulers of the various states of their program to bring order to the land. At the same time, perhaps with the increased displacement and disillusionment of intellectuals, a stronger eremitic tradition also emerged. If the *Laozi* had originated from the fourth century, it might reflect some of these concerns. From this perspective, the origin of the *Laozi* is as much a hermeneutical issue as it is an historical one.

Textual Traditions

The discovery of two *Laozi* silk manuscripts at Mawangdui, near Changsha, Hunan province in 1973 marks an important milestone in modern *Laozi* research. The manuscripts, identified simply as "A" (*jia*) and "B" (*yi*), were found in a tomb that was sealed in 168 B.C.E. The texts themselves can be dated earlier, the "A" manuscript being the older of the two, copied in all likelihood before 195 B.C.E. (see Lau 1982, Boltz 1984, and Henricks 1989). Before this find, access to the *Laozi* was mainly through the received text of Wang Bi (226-249 C.E.) and Heshanggong, a legendary figure depicted as a teacher to the Han Emperor Wen (r. 179-157 B.C.E.). There are other manuscript versions, but by and large they play a secondary role in the history of the classic. A more recent archaeological find in Guodian, the so-called "Bamboo-slip *Laozi*," which predates the Mawangdui manuscripts, has rekindled debates on the origin and composition of the *Laozi*. But first a note on the title and structure of the *Daodejing*.

The *Laozi* did not acquire its canonical status until the Han dynasty. According to the *Shiji* (49.5b), the Empress Dowager Dou -- wife of Emperor Wen and mother of Emperor Jing (r. 156-141) -- was a dedicated student of the *Laozi*. Later sources add that it was Emperor Jing who established the text officially as a "classic" (*jing*). However, the title *Daodejing* appears not to have been widely used until later, toward the close of the Han era. The *Daodejing* is also referred to as the *Daode zhenjing* (True Classic of the Way and Virtue), the *Taishang xuanyuan Daodejing* (Classic of the Way and Virtue of the Highest Primordial Mystery), and less formally the "five-thousand character" text, on account of its approximate length. Most versions exceed five thousand characters by about five to ten percent, but it is interesting to note that numerological considerations later became an integral part of the history of the work. According to the seventh-century Daoist master Cheng Xuanying, Ge Xuan (fl. 200 C.E.) shortened the text that accompanied the Heshanggong commentary to fit the magical number of five thousand. This claim cannot be verified, but a number of *Laozi* manuscripts discovered at Dunhuang contain 4,999 characters.

The current *Daodejing* is divided into two parts (*pian*) and 81 chapters or sections (*zhang*). Part one, comprising chapters 1-37, has come to be known as the *Daojing*, while chapters 38-81 make up the *Dejing*. This is understood to be a thematic division -- chapter 1 begins with the word Dao, while chapter 38 begins with the phrase "superior virtue" -- although the concepts of Dao and virtue (*de*) feature in both parts. As a heuristic guide, some commentators have suggested that the *Daojing* is more "metaphysical," whereas the *Dejing* focuses more on sociopolitical issues.

In this context, it is easy to appreciate the tremendous interest occasioned by the discovery of the Mawangdui *Laozi* manuscripts. The two manuscripts contain all the chapters that are found in the current *Laozi*, although the chapters follow a different order in a few places. For example, in both manuscripts, the sections that appear as chapters 80 and 81 in the current *Laozi* come immediately after a section that corresponds with chapter 66 of the present text. Both manuscripts are similarly divided into two parts, but in contrast with the current version, in reverse order; i.e., both manuscripts begin with the *Dejing*, corresponding to chapter 38 of the received text. "Part one" of the "B" manuscript ends with the editorial notation, "Virtue, 3,041 [characters]," while the last line of "Part two" reads: "Dao, 2,426." Does this

mean that the classic should be renamed? One scholar, in fact, has adopted the title *Dedaojing* for his translation of the Mawangdui *Laozi* (Henricks 1989). Does this imply that the "original" *Laozi* gives priority to sociopolitical issues? This raises important questions for interpretation.

The division into 81 chapters reflects numerological interest and is associated particularly with the Heshanggong version, which also carries chapter titles. It was not universally accepted until much later, perhaps the Tang period, when the text was standardized under the patronage of Emperor Xuanzong (r. 712-755). Traditional sources report that some versions were divided into 64, 68, or 72 chapters; and some did not have chapter divisions (Henricks 1982). The Mawangdui "A" manuscript contains in some places a dot or "period" that appears to signal the beginning of a chapter. The earlier Guodian text is not divided into two parts, but in many places it employs a black square mark to indicate the end of a section. The sections so marked generally agree with the division in the present *Laozi*. Thus, although the 81-chapter formation may be relatively late, some attempt at chapter division seems evident from an early stage of the textual history of the *Daodejing*.

Until recently, the Mawangdui manuscripts have held the pride of place as the oldest extant manuscripts of the *Laozi*. In late 1993, the excavation of a tomb (identified as M1) in Guodian, Jingmen city, Hubei province, has yielded among other things some 800 bamboo slips, of which 730 are inscribed, containing over 13,000 characters. Some of these, amounting to about 2,000 characters, match the *Laozi* (see Allan and Williams 2000, and Henricks 2000). The tomb is located near the old capital of the state of Chu and is dated around 300 B.C.E. Robbers entered the tomb before it was excavated, although the extent of the damage is uncertain. The bamboo texts, written in a Chu script, have been transcribed into standard Chinese and published under the title *Guodian Chumu zhujian* (Beijing: Wenwu, 1998), which on the basis of the size and shape of the slips, calligraphy, and other factors divides the *Laozi* material into three groups. Group A contains thirty-nine bamboo slips, which correspond in whole or in part to the following chapters of the present text: 19, 66, 46, 30, 15, 64, 37, 63, 2, 32, 25, 5, 16, 64, 56, 57, 55, 44, 40 and 9. Groups B and C are smaller, with eighteen (chs. 59, 48, 20, 13, 41, 52, 45, 54) and fourteen slips (chs. 17, 18, 35, 31, 64), respectively.

On the whole, the Guodian "bamboo-slip *Laozi*" is consistent with the received text, although the placement or sequence of the chapters is different and there are numerous variant and/or archaic characters. Particularly, whereas chapter 19 of the current *Laozi* contains what appears to be a strong attack on Confucian ideals -- "Cut off benevolence (*ren*), discard rightness (*yi*)" -- the Guodian "A" text directs its readers to "cut off artificiality, discard deceit." This has been taken to suggest that in the course of its transmission, the *Laozi* has taken on a more "polemical" outlook. However, the Guodian "C" text indicates that *ren* and *yi* arose only after the "Great Dao" had gone into decline, which agrees with chapter 18 of the current *Laozi*.

It is not clear whether the Guodian bamboo manuscripts were copied from one source and were meant to be read as one text divided into three parts, whether they were "selections" from a longer original, or whether they were three different texts copied from different sources at different times. The "A" and "C" texts give two different versions of what is now part of chapter 64 of the *Laozi*, which may suggest different sources. One scholar at least has suggested a chronology to the making of the Guodian *Laozi*

bamboo slips, with the “A” group being the oldest of the three, copied around 400 B.C.E. (Ding 2000, 7-9). It is possible that the Guodian texts only furnished some of the textual “raw material” or “building blocks” that were used later to create the *Laozi* (Boltz 1999). In other words, they were independent writings and not versions of or excerpts from the *Laozi*, which in this scenario did not yet exist when the Guodian texts were made. Nevertheless, taking into account all the available evidence, it seems likely that a body or bodies of sayings attributed to Laozi gained currency during the fourth century B.C.E. They may have been derived from earlier, oral or written sources. By the mid-third century if not earlier, the *Laozi* probably reached more or less its final form and began to attract commentarial attention.

Two approaches to the making of the *Laozi* warrant consideration, for they bear directly on interpretation. A linear “evolutionary” model of textual formation would suggest that there was an original *Laozi*, by Lao Dan or of unknown authorship, and that the Guodian *Laozi* was close to this original text. Concerned with the decline of Zhou rule, the original *Laozi* addressed above all issues of governance. During the third century B.C.E., the *Laozi* had undergone substantial change and grown into a longer and more complex work, becoming in this process more polemical against the Confucian and other schools of thought, and acquiring new material of stronger metaphysical or cosmological interest. The Mawangdui manuscripts were based on this mature version of the *Laozi*; the original emphasis on politics, however, can still be detected in the placement of the *Dejing* before the *Daojing*. Later versions reversed this order and in so doing subsumed politics under a broader philosophical vision of Dao as the beginning and end of all beings.

The Guodian and Mawangdui manuscripts are certainly older than the received text of the *Laozi*, but this does not necessarily mean that they are therefore closer to the “original,” if there was an original. As opposed to a linear evolutionary model, it is conceivable that there were several overlapping collections of sayings attributed to Laozi from the start, each inhabiting a particular interpretive context, from which different versions of the *Laozi* were derived. Although some key chapters in the current *Laozi* that deal with the nature of Dao (e.g., chs. 1, 14) are not found in the Guodian corpus, the idea that the Dao is “born before heaven and earth,” for example, which is found in chapter 25 of the received text is already present. The critical claim that “being [*you*] is born of nonbeing [*wu*]” in chapter 40 also figures in the Guodian “A” text. This seems to argue against any suggestion that the *Laozi*, and for that matter ancient Chinese philosophical works in general were not interested or lacked the ability to engage in abstract philosophic thought, an assumption that sometimes appears to underlie evolutionary approaches to the development of Chinese philosophy.

The Guodian and Mawangdui finds are extremely valuable. They are syntactically clearer than the received text in some instances, thanks to the larger number of grammatical particles they employ. However, they cannot resolve all the controversies and uncertainties surrounding the *Laozi*. Perhaps the two approaches identified above are not mutually exclusive. Different written collections of Laozi sayings, leaving open the time and the way in which they were first formed, circulated during the fourth century. Overlapping in some cases and with varying emphases in others, they address both the nature of Dao and Daoist government. These were then developed in several ways -- e.g., some collections were combined; new sayings were added; and explanatory comments, illustrations, and elaboration on individual sayings were integrated into the text. The demand for textual uniformity rose when the *Laozi*

gained recognition, and consequently the different textual traditions eventually gave way to the received text of the *Laozi*.

As mentioned, the current *Laozi* on which most reprints, studies and translations are based is the version that comes down to us along with the commentaries by Wang Bi and Heshanggong. Three points need to be made in this regard. First, technically there are multiple versions of the Wang Bi and Heshanggong *Laozi* -- over thirty Heshanggong versions are extant -- but the differences are on the whole minor. Second, the Wang Bi and Heshanggong versions are not the same, but they are sufficiently similar to be classified as belonging to the same line of textual transmission. Third, the Wang Bi and Heshanggong versions that we see today have suffered change. Prior to the invention of printing, when each manuscript had to be copied by hand, editorial changes and scribal errors are to be expected. In particular, the *Laozi* text that now accompanies Wang Bi's commentary bears the imprint of later alteration, mainly under the influence of the Heshanggong version, and cannot be regarded as the *Laozi* that Wang Bi himself had seen and commented on. Boltz (1985) and Wagner (1989) have examined this question in some detail.

The "current" version refers to the "Sibu beiyao" and the "Sibu congkan" editions of the *Daodejing*. The former contains the Wang Bi version and commentary, together with a colophon by the Song scholar Chao Yuezhi (1059-1129), a second note by Xiong Ke (ca. 1111-1184)), and Lu Deming's (556-627) *Laozi yinyi* (Glosses on the Meaning and Pronunciation of the *Laozi*). It is a reproduction of the Qing dynasty "Wuying Palace" edition, which in turn is based on a Ming edition (see especially Hatano 1979). The Heshanggong version preserved in the Sibu congkan series is taken from the library of the famous bibliophile Qu Yong (fl. 1850). According to Qu's own catalogue, this is a Song version, published probably after the reign of the emperor Xiaozong (r. 1163-1189). Older extant versions include two incomplete Tang versions and fragments found in Dunhuang.

Besides the Guodian bamboo texts, the Mawangdui manuscripts, and the received text of Wang Bi and Heshanggong, there is an "ancient version" (*guben*) edited by the early Tang scholar Fu Yi (fl. 600). Reportedly, this version was recovered from a tomb in 574 C.E., whose occupant was a consort of the Chu general Xiang Yu (d. 202 B.C.E.), the rival of Liu Bang before the latter emerged victorious and founded the Han dynasty. A later redaction of the "ancient version" was made by Fan Yingyuan in the Song dynasty. There are some differences, but these two can be regarded as having stemmed from the same textual tradition.

Manuscript fragments discovered in the Dunhuang caves form another important source in *Laozi* research. Among them are several Heshanggong fragments (especially S. 477 and S. 3926 in the Stein collection, and P. 2639 in the Pelliot collection) and the important *Xiang'er Laozi* with commentary. Another Dunhuang manuscript that merits attention is the Suo Tan fragment, now at the University Art Museum, Princeton University, which contains the last thirty-one chapters of the *Daodejing* beginning with chapter 51 of the modern text. It is signed and dated at the end, bearing the name of the third-century scholar and diviner Suo Tan, who is said to have made the copy, written in ink on paper, in 270 C.E. According to Rao Zongyi (1955), the Suo Tan version belongs to the Heshanggong line of the *Laozi* text. A more recent study by William Boltz (1996) questions its third-century date and argues that the fragment in many instances also agrees with the Fu Yi "ancient version."

While manuscript versions inform textual criticism of the *Laozi*, stone inscriptions provide further collaborating support. Over twenty steles, mainly of Tang and Song origins, are available to textual critics, although some are in poor condition (Yan 1957). Students of the *Laozi* today can work with several Chinese and Japanese studies that make use of a large number of manuscript versions and stone inscriptions (notably Ma 1965, Jiang 1980, Zhu 1980, and Shima 1973). Boltz (1993) offers an excellent introduction to the manuscript traditions of the *Laozi*.

Commentaries

Commentaries to the *Laozi* offer an invaluable guide to interpretation and are important also for their own contributions to Chinese philosophy and religion. Two chapters in the *Hanfeizi* (chs. 21-22) are entitled "Explaining the *Laozi*" (*Jie Lao*) and "Illustrating the *Laozi*" (*Yu Lao*), which can be regarded as the earliest extant commentary to the classic. The "bibliographical" section of the *Hanshu* (History of the Former Han Dynasty) lists four commentaries to the *Laozi*, but they have not survived. Nevertheless, *Laozi* learning began to flourish from the Han period. The commentaries by Heshanggong, Yan Zun, Wang Bi, and the *Xiang'er* commentary will be introduced in what follows. Some mention will also be made of later developments in the history of the *Daodejing*. The late Isabelle Robinet has contributed an important pioneering study of the early *Laozi* commentaries (1977; see also Robinet 1998).

Traditionally, the Heshanggong commentary is regarded as a product of the early Han dynasty. The name Heshanggong means an old man who dwells by the side of the river, and some have identified the river in question to be the Yellow River. An expert on the *Laozi*, he caught the attention of Emperor Wen, who went personally to consult him. Heshanggong revealed to the emperor his true identity as a divine emissary sent by the "Supreme Lord of the Dao" -- i.e., the divine Laozi -- to teach him. The emperor proves a humble student, as the legend concludes, worthy of receiving the *Daodejing* with Heshanggong's commentary (Chan 1991a).

Recent Chinese studies generally place the commentary at the end of the Han period, although some Japanese scholars would date it to as late as the sixth century C.E. It is probably a second-century C.E. work and reflects the influence of the "Huang-Lao" (Yellow Emperor and Laozi) school, which flourished during the early Han dynasty (Chan 1991b). Called in early sources the *Laozi zhangju*, it belongs to the genre of *zhangju* literature, prevalent in Han times, which one may paraphrase as commentary by "chapter and sentence." Its language is simple; its imagination, down-to-earth. The Heshanggong commentary shares with other Han works the cosmological belief that the universe is constituted by *qi* or "vital energy." On this basis, interpreting the text in terms of yin-yang theory, the *Laozi* is seen to disclose not only the mystery of the origin of the universe but also the secret to personal well-being and sociopolitical order.

What the *Laozi* calls the "One," according to Heshanggong, refers to the purest and most potent form of *qi*-energy that brings forth and continues to nourish all beings. This is the meaning of *de*, the "virtue" or power with which the "ten thousand things" -- i.e., all beings -- have been endowed and without which

life would cease. The maintenance of “virtue,” which the commentary also describes as “guarding the One,” is thus crucial to self-cultivation. A careful diet, exercise, and some form of meditation are implied, but generally the commentary focuses on the diminishing of selfish desires. The government of the “sage” -- a term common to all schools of Chinese thought but which is given a distinctive Daoist meaning in the commentary -- rests on the same premise. Policies that are harmful to the people such as heavy taxation and severe punishment are to be avoided, but the most fundamental point remains that the ruler himself must cherish what the *Laozi* calls “emptiness” and “nonaction.” Disorder stems from the dominance of desire, which reflects the unruly presence of confused and agitated *qi*-energy. In this way, self-cultivation and government are shown to form an integral whole.

A second major commentary is the *Laozi zhigui* (The Essential Meaning of the *Laozi*) attributed to the Han dynasty scholar Yan Zun (fl. 83 B.C.E.-10 C.E.). Styled Junping, Yan’s surname was originally Zhuang; it was changed in later written records to the semantically similar Yan to comply with the legal restriction not to use the name Zhuang, which was the personal name of Emperor Ming (r. 57-75) of the Later Han dynasty. Yan Zun is well remembered in traditional sources as a recluse of great learning and integrity, a diviner of legendary ability, and an author of exceptional talent. The famous Han poet and philosopher Yang Xiong (53 B.C.E.-18 C.E.) studied under Yan and spoke glowingly of him.

The *Laozi zhigui*, as it now stands, is incomplete; only the commentary to the *Dejing*, chapters 38-81 of the current *Laozi*, remains. The best edition of the *Zhigui* is that contained in the *Daozang* (Daoist Canon 693, fasc. 375-377), which clearly indicates that the work had originally thirteen *juan* or books, the first six of which have been lost. Judging from the available evidence, it can be accepted as a Han product (Chan 1998b). The *Laozi* text that accompanies Yan Zun’s commentary agrees in many instances with the wording of the Mawangdui manuscripts.

Like Heshanggong, Yan Zun also subscribes to the yin-yang cosmological theory characteristic of Han thought. Unlike Heshanggong’s commentary, however, the *Zhigui* does not prescribe a program of nourishing one’s *qi*-energy or actively cultivating “long life.” This does not mean that it rejects the ideal of longevity. On the contrary, it recognizes that the Dao “lives forever and does not die” (8.9b), and that the man of Dao, correspondingly, “enjoys long life” (7.2a). Valuing one’s spirit and vital energy is important, but the *Zhigui* is concerned that self-cultivation must not violate the principle of “nonaction.” Any effort contrary to what the *Laozi* has termed “naturalness” (*ziran*) is counter-productive and doomed to failure.

The concept of *ziran* occupies a pivotal position in Yan Zun’s commentary. It describes the nature of the Dao and its manifestation in the world. It also points to an ethical ideal. The way in which natural phenomena operate reflects the workings of the Dao. The “sage” follows the Dao in that he, too, abides by naturalness. In practice this means attending to one’s heart-mind (*xin*) so that it will not be enslaved by desire. Significantly, the *Zhigui* suggests that just as the sage “responds” to the Dao in being simple and empty of desire, the common people would in turn respond to the sage and entrust the empire to him. In this way, the *Laozi* is seen to offer a comprehensive guide to order and harmony at all levels.

An early commentary that maximizes the religious import of the *Laozi* is the *Xiang’er Commentary*.

Although it is mentioned in catalogues of Daoist works, there was no real knowledge of it until a copy was discovered among the Dunhuang manuscripts (S. 6825 in the Stein collection). The manuscript copy, now housed in the British Library, was probably made around 500 C.E. The original text, disagreement among scholars notwithstanding, is generally traced to around 200 C.E. It is closely linked to the "Way of the Celestial Master" and has been ascribed to Zhang Daoling, the founder of the sect, or his grandson Zhang Lu, who was instrumental in ensuring the group's survival after the collapse of the Han dynasty. A detailed study and translation of the work in English is now available (Bokenkamp 1997).

The *Xiang'er* manuscript is unfortunately incomplete; only the first part has survived, beginning with the middle of chapter 3 and ending with chapter 37 in the current chapter division of the *Laozi*. It is not clear what the title, *Xiang'er*, means. Following Rao Zongyi and Ōfuchi Ninji, Stephen Bokenkamp suggests that it is best understood in the literal sense that the Dao "thinks (*xiang*) of you (*er*)" (1997, 61). This underscores the central thesis of the commentary, that devotion to the Dao in terms of self-cultivation and compliance with its precepts would assure boundless blessing in this life and beyond.

The *Xiang'er* commentary accepts without question the divine status of Laozi. While Yan Zun and Heshanggong direct their commentary primarily to those in a position to effect political change, the *Xiang'er* invites a larger audience to participate in the quest for the Dao, to achieve union with the Dao through spiritual and moral discipline. It is possible to attain the "life-span of an undefiled, godlike being" (*xianshou*). Nourishing one's vital *qi*-energy through meditation and other practices remains the key to attaining "long life" and ultimately to forming a spiritual body devoid of the blemishes of mundane existence (Rao 1991).

Spiritual discipline, however, is not sufficient; equally important is the accumulation of moral merit. Later Daoist sources refer to the "nine precepts" of the *Xiang'er*. There is also a longer set known as the "twenty-seven precepts" of the *Xiang'er*. These include general positive steps such as being tranquil and yielding, as well as specific injunctions against envy, killing, and other morally reprehensible acts. Likening the human body to the walls of a pond, the essential *qi*-energy to the water in it, and good deeds the source of the water, the *Xiang'er* commentary makes clear that deficiency in any one would lead to disastrous consequences (see Bokenkamp 1993).

Compared with the *Xiang'er*, Wang Bi's *Laozi* commentary could not be more different. There is no reference to "immortals"; no deified Laozi. The *Daodejing*, as Wang Bi sees it, is fundamentally not concerned with the art of "long life" but offers profound insights into the radical otherness of Dao as the source of being, and the practical implications that follow from it.

Styled Fusi, Wang Bi (226-249) was one of the acknowledged leaders of the movement of the "Learning of the Mysterious (Dao)" (*xuanxue*), a revival of Daoist philosophy that came into prominence during the Wei period (220-265) and dominated the Chinese intellectual scene well into the sixth century. The word *xuan* denotes literally a shade of dark red and is used in the *Laozi* (esp. ch. 1) to suggest the mystery or profundity of Dao. The movement has been identified, perhaps inappropriately, as "Neo-Daoism" in some Western sources. It signifies a broad philosophical front united in its attempt to discern the "true" meaning of Dao but not a homogeneous, sectarian school. Alarmed by what they saw as the decline of

Dao, influential intellectuals of the day initiated a radical reinterpretation of the classical heritage. They did not neglect the Confucian classics but drew inspiration especially from the *Yijing*, the *Laozi*, and the *Zhuangzi*, which were then referred to as the "Three Treatises on the Mysterious (Dao)" (*sanxuan*). Wang Bi, despite his short life, distinguished himself as a brilliant interpreter of the *Laozi* and the *Yijing* (see Chan 1991b and Wagner 2000).

According to Wang Bi, Dao is indeed the "beginning" of the "ten thousand things." Unlike Heshanggong or the *Xiang'er*, however, he did not pursue a cosmological or religious interpretation of the process of creation. Rather, Wang seems more concerned with what may be called the logic of creation. Dao constitutes the absolute "beginning" in that all beings have causes and conditions that derive logically from a necessary foundation. The ground of being, however, cannot be itself a being; otherwise, infinite regress would render the logic of the *Laozi* suspect. For this reason, the *Laozi* would only speak of Dao as "nonbeing" (*wu*).

The transcendence of Dao must not be compromised. To do justice to the *Laozi*, it is also important to show how the function of Dao translates into basic "principles" (*li*) governing the universe. The regularity of the seasons, the plenitude of nature, and other expressions of "heaven and earth" all attest to the presence of Dao. Human beings also conform to these principles, and so are "modeled" ultimately after Dao.

Wang Bi is often praised in later sources for having given the concept of "principle" its first extended philosophical treatment. In the realm of Dao, principles are characterized by "naturalness" (*ziran*) and "nonaction" (*wuwei*). Wang Bi defines *ziran* as "an expression of the ultimate." In this regard, attention has been drawn to Yan Zun's influence. Nonaction helps explain the practical meaning of naturalness. In ethical terms, Wang Bi takes nonaction to mean freedom from the dictates of desire. This defines not only the goal of self-cultivation but also that of government. Wang Bi's *Laozi* commentary has exerted a strong influence on modern interpretations of the *Laozi* in both Asia and the West. There are three English translations available (Lin 1977, Rump 1979, and Lynn 1999).

Among these four commentaries, Heshanggong's *Laozi zhangju* occupied the position of preeminence in traditional China, at least until the Song dynasty. For a long period, Wang Bi's work was relatively neglected. The authority of the Heshanggong commentary can be traced to its place in the Daoist religion, where it ranks second only to the *Daodejing* itself. Besides Heshanggong's work and the *Xiang'er*, there are two other commentaries entitled the *Laozi jiejie* (Sectional Explanation) and the *Laozi neijie* (Inner Explanation) closely associated with religious Daoism. Both have been ascribed to Yin Xi, the keeper of the pass who "persuaded" Laozi to write the *Daodejing* and who, according to Daoist hagiographic records, later studied under the divine Laozi and became an "immortal." These texts, however, only survive in citations (see Kusuyama 1979).

From the Tang period, one begins to find serious attempts to collect and classify the growing number of *Laozi* commentaries. An early pioneer is the eighth-century Daoist master Zhang Junxiang, who cited some thirty commentaries in his study of the *Daodejing* (Wang 1981). Du Guangting (850-933) provided a larger collection, involving some sixty commentaries (*Daode zhenjing guangshengyi*, *Daozang* 725).

According to Du, there were those who saw the *Laozi* as a political text, while others focused on spiritual self-cultivation. There were Buddhist interpreters (e.g., Kumārajīva and Sengzhao), and there were those who explained the "Twofold Mystery" (*chongxuan*). This latter represents an important development in the history of interpretation of the *Daodejing*.

The term "Twofold Mystery" comes from chapter 1 of the *Laozi*, where Dao is said to be the mystery of all mysteries (*xuan zhi you xuan*). As a school of Daoist learning, "Twofold Mystery" seizes this to be the key to understanding the *Laozi*. Daoist sources relate that the school goes back to the fourth-century master Sun Deng. Through Gu Huan (fifth century) and others, the school reached its height during the Tang period, represented by such thinkers as Cheng Xuanying and Li Rong in the seventh century. The school reflects the growing interaction between Daoist and Buddhist thought, particularly Mādhyamika philosophy. Unlike Wang Bi, it sees "nonbeing" as equally one-sided as being when applied to the transcendence of Dao. Nonbeing may highlight the profundity or mystery of Dao, but it does not yet reach the highest truth, which according to Cheng Xuanying can be called the "Dao of Middle Oneness" (Kohn 1992, 144). Like other polar opposites, the distinction between being and nonbeing must also be "forgotten" before one can achieve union with Dao.

The *Laozi* has been viewed in still other ways. For example, a Tang commentary by Wang Zhen, the *Daodejing lunbing yaoyishu* (*Daozang* 713; fasc. 417), presented to Emperor Xianzong (r. 806-820) in 809, sees the text as a treatise on military strategy (Rand 1979-80; see also Wang Ming 1984 and Mukai 1994). The diversity of interpretation is truly remarkable (see Robinet 1998 for a typological analysis). The *Daodejing* was given considerable imperial attention, with no fewer than eight emperors having composed or at least commissioned a commentary on the work. These include Emperor Wu and Emperor Jianwen of the Liang dynasty, Xuanzong of the Tang, Huizong of the Song, and Taizu of the Ming dynasty (see Liu Cunren 1969 for a discussion of the last three).

By the thirteenth century, students of the *Daodejing* were already blessed, as it were, with an embarrassment of riches, so much so that Du Daojian (1237-1318) could not but observe that the coming of the Dao to the world takes on a different form each time. That is to say, different commentators were shaped by the spirit of their age in their approach to the classic, so that it would be appropriate to speak of a "Han *Laozi*," "Tang *Laozi*," or "Song *Laozi*," each with its own agenda (*Xuanjing yuanzhi fahui*, DZ 703; fasc. 391).

Approaches to the *Laozi*

Is the *Laozi* a manual of self-cultivation and government? Is it a metaphysical treatise, or does it harbor deep mystical insights? Chapter 1 of the current *Laozi* begins with the famous words: "The Way that can be spoken of is not the constant Way." Chapter 10 speaks of nourishing one's "soul" and embracing the "One." Chapter 80 depicts the ideal polity as a small country with few inhabitants. The *Laozi* is a difficult text. Its language is often cryptic; the sense or reference of the many symbols it employs remains unclear, and there seems to be conceptual inconsistencies. For example, whereas chapter 2 refers to the "mutual production of being and nonbeing," chapter 40 declares, "Being originates in nonbeing" (Henricks, trans.

1989). Is it more meaningful to speak of the "worldviews" of the *Daodejing*, instead of a unified vision? If the *Laozi* were an "anthology" put together at random by different compilers over a long period of time, coherence need not be an issue. Traditionally, however, this was never a serious option. Most modern studies are equally concerned to disclose the "deeper" unity and meaning of the classic. While some seek to recover the "original" meaning of the *Laozi*, others celebrate its contemporary relevance. Consider, first of all, some of the main modern approaches to the *Daodejing* (cf. Hardy 1998).

One view is that the *Laozi* reflects a deep mythological consciousness at its core. The myth of "chaos," in particular, helps shape the Daoist understanding of the cosmos and the place of human beings in it (Girardot 1983). Chapter 25, for example, likens the Dao to an undifferentiated oneness. The myth of a great mother earth goddess may also have informed the worldview of the *Laozi* (Erkes 1935; Chen 1969), which explains its emphasis on nature and the feminine (Chen 1989). Chapter 6, for example, refers to the "spirit of the valley," which is also called the "mysterious female."

A second view is that the *Laozi* gives voice to a profound mysticism. According to Victor Mair (1990), it is indebted to Indian mysticism (see also Waley 1958). According to Benjamin Schwartz (1985), the mysticism of the *Daodejing* is *sui generis*, uniquely Chinese and has nothing to do with India. Indeed, as one scholar suggests, it is unlike other mystical writings in that ecstatic vision does not play a role in the ascent of the Daoist sage (Welch 1965, 60). According to another interpretation, however, there is every indication that ecstasy forms a part of the world of the *Laozi*, although it is difficult to gauge the "degree" of its mystical leanings (Kaltenmark 1969, 65). It is possible to combine the mystical and mythological approaches to yield a third view. Although the presence of ancient religious beliefs can still be detected, they have been raised to a "higher" mystical plane in the *Laozi* (e.g., Ching 1997).

A fourth view sees the *Laozi* mainly as a work of philosophy, which gives a metaphysical account of reality and insight into Daoist self-cultivation and government; but fundamentally it is not a work of mysticism (W. T. Chan 1963). The strong practical interest of the *Laozi* distinguishes it from any mystical doctrine that eschews worldly involvement. It is, in Creel's (1977) words, "purposive" and not "contemplative." Fifth, to many readers the *Laozi* offers essentially a philosophy of life. Remnants of an older religious thinking may have found their way into the text, but they have been transformed into a naturalistic philosophy. The emphasis on naturalness translates into a way of life characterized by simplicity, calmness, and freedom from the tyranny of desire (e.g., Liu Xiaogan 1997). Unlike the claim that the *Laozi* espouses a mystical or esoteric teaching directed at a restricted audience, this view tends to highlight its universal appeal and contemporary relevance.

A sixth and influential view is that the *Laozi* is above all concerned with realizing peace and sociopolitical order. It is an ethical and political masterpiece intended for the ruling class, with concrete strategic suggestions aimed at remedying the moral and political turmoil engulfing late Zhou China. Self-cultivation is important, but the ultimate goal extends beyond personal fulfillment (Lau 1963; LaFargue 1992). The *Laozi* criticizes the Confucian school not only for being ineffectual in restoring order but more damagingly as a culprit in worsening the ills of society at that time. The ideal seems to be a kind of "primitive" society, where people would dwell in harmony and contentment, not fettered by ambition or desire (Needham 1956).

This list is far from exhaustive; there are other views of the *Laozi*. Chad Hansen (1992), for example, focuses on the "anti-language" philosophy of the text. Different combinations are also possible. A. C. Graham, for example, emphasizes both the mystical and political elements, arguing that the *Laozi* was probably targeted at the ruler of a small state (1989, 234). The *Laozi* could be seen as encompassing all of the above -- such categories as the metaphysical, ethical, political, mystical, and religious form a unified whole in Daoist thinking and are deemed separate and distinct only in Western thought. Alternatively, coming back to the question of multiple authorship and coherence, it could be argued that the *Laozi* contains "layers" of material put together by different people at different times (Emerson 1995).

Is it fair to say that the *Laozi* is inherently "polysemic" (Robinet 1998), open to diverse interpretations? This concerns not only the difficulty of the *Laozi* but also the interplay between reader and text in any act of interpretation. Polysemy challenges the assertion that the "intended" meaning of the *Laozi* can be recovered fully. But it does not follow that context is unimportant, that parameters do not exist, or that there are no checks against particular interpretations. Questions of provenance, textual variants, as well as the entire tradition of commentaries and modern scholarship are important for this reason. Put differently, while hermeneutic reconstruction should be given full attention, it remains an open process. The following presents some of the main concepts and symbols in the *Laozi* based on the current text, focusing on the key conceptual cluster of Dao, *de* (virtue), *ziran* (naturalness), and *wuwei* (nonaction), in a way that highlights their philosophical significance and suggests a degree of coherence.

To begin with Dao, the etymology of the graph suggests a pathway, or heading in a certain direction along a path. Most commentators agree in translating *dao* as "way." As a verb, perhaps on account of the directionality involved, *dao* also conveys the sense of "speaking." Thus, the opening phrase of chapter 1, *dao ke dao*, literally "Dao that can be dao-ed," is often rendered, "The Way that can be spoken of." In most cases, the capitalized form -- "Way" or "Dao" -- is used, to distinguish it from other usages of the term.

The concept of *dao* figures centrally also in Confucian writings, and as mentioned some parts of the current *Laozi* represent a critique of the Confucian school (especially chs. 18 and 19). In general, whereas *dao* signifies a means to a higher end in other schools of Chinese philosophy, the *Laozi* sees it as an end in itself. This distinction is captured in the *Oxford English Dictionary* (online edition), which defines "Dao" as follows: "In Taoism, an absolute entity which is the source of the universe; the way in which this absolute entity functions." "In Confucianism and in extended uses," however, the term means "the way to be followed, the right conduct; doctrine or method."

The *Laozi* underscores both the ineffability and creative power of Dao. Chapter 1 states that the "constant" (*chang*, also translated as "eternal" -- e.g., W. T. Chan 1963) Dao cannot be described; it is "nameless." Chapter 14 brings out clearly that Dao transcends sensory perception; it has no shape or form. Nameless and formless, Dao can only be described as "dark" (*xuan*) or *wu*, literally "not having" any name, form, or other characteristics of things (see also chs. 21 and 32). Indeed, though suggestive, the term "Dao" itself is no more than a symbol -- as the *Laozi* makes clear, "I do not know its name; I style it Dao" (ch. 25; see also ch. 34). This suggests a sense of radical transcendence, which may explain why the

Laozi has been approached so often as a mystical text.

The concept of *wu* is difficult and has been translated variously as “nothing,” “nothingness,” or “nonbeing.” It marks not only the mystery of Dao but also its limitlessness or inexhaustibility (e.g., ch. 4). Names serve to delimit, to set boundaries; in contrast, Dao is without limits and therefore cannot be captured fully by language. This suggests a positive dimension to transcendence, which brings into view the creative power of Dao: “All things under heaven are born of being (*you*); being is born of *wu*” (ch. 40). What does this mean?

Elsewhere in the *Laozi*, Dao is said to be the “beginning” of all things (chs. 1, 25). Daoist creation involves a process of differentiation from unity to multiplicity: “Dao gives birth to One; One gives birth to Two; Two gives birth to Three; Three gives birth to the ten thousand things” (ch. 42). The text does not indicate tense or spell out what the numbers refer to -- is it saying that something called “the One” produced or produces “the Two”? The “nothingness” of Dao helps impose certain constraints on interpretation. Specifically, the idea of a creator god with attributes, like the “Lord on High” (Shangdi) in ancient Chinese religion, does not seem to fit with the emphasis on transcendence.

The dominant interpretation in traditional China is that Dao represents the source of the original, undifferentiated, essential *qi*-energy, the “One,” which in turn produces the yin and yang cosmic forces. While the yang energy rises to form heaven, yin solidifies to become earth. A further “blending” of the two generates a “harmonious” *qi*-energy that informs human beings. This is essentially the reading of the Heshanggong commentary. Although the *Laozi* may not have entertained a fully developed yin-yang cosmological theory, which took shape during the Han period, it does suggest at one point that natural phenomena are constituted by yin and yang (ch. 42). That which gave rise to the original *qi*-energy is indescribable. The *Laozi* calls it Dao, or perhaps more appropriately in this context, “the Dao,” with the definite article, to signal its presence as the source of the created order. In modern terms, minus the language of yin-yang cosmology, this translates into an understanding of the Dao as “an absolute entity which is the source of the universe.” Not being anything in particular, the Dao may be described as “nothing” (*wu*). However, *wu* does not mean “nothingness” or absence in the nihilistic sense, in view of the creative power of the Dao.

Alternatively, one could argue that Dao signifies a conceptually necessary ontological ground; it does not refer to any indescribable original substance or energy. “Beginning” is not a term of temporal reference but suggests ontological priority in the *Laozi*. The process of creation does proceed from unity to multiplicity, but the *Laozi* is only concerned to show that “two” would be impossible without the idea of “one.” The assertion “One gives birth to Two” affirms that duality presupposes unity; to render it as “The One gave birth to the Two” is to turn what is essentially a logical relation into a cosmological event.

As the source of being, Dao cannot be itself a being, no matter how powerful or perfect; otherwise, the problem of infinite regress cannot be overcome. For this reason, the *Laozi* makes use of the concept of *wu*, “nonbeing,” not to suggest a substance or something of which nothing can be said, but to signify the conceptual “otherness” and radical transcendence of the ground of being. This agrees with Wang Bi’s interpretation. If *wu* points to a necessary ontological foundation, the distinction between “Dao” and

"One" seems redundant. Commenting on chapter 42 of the *Laozi*, Wang Bi writes, "One can be said to be *wu*"; "One is the beginning of numbers and the ultimate of things" (commentary on ch. 39; see also Wang's commentary on the *Yijing*, trans. in Lynn 1994, 60). The concept of "One" and the concept of *wu* thus complement each other in disclosing different aspects of the logic of creation -- both unity and nonbeing are necessary for understanding the generation of beings.

Comparing the two interpretations, whereas the first, "cosmological" reading has to explain the sense in which the Dao can be said to be "nothing," the second emphasizes the centrality of *wu*, for which "Dao" is but one designation. Depending on the interpretation, *wu* may be translated as "nothing" or "nonbeing" accordingly. The metaphor of "Dao" is apt. It shows that all things are derived ultimately from an absolute "beginning," in either sense of the word, like the start of a pathway. It also suggests a direction to be followed, which brings out the ethical interest of the *Laozi*.

The *Daodejing* is concerned with both Dao and *de*. The graph *de* has also made it into the *Oxford*: "In Taoism, the essence of Tao inherent in all beings"; "in Confucianism and in extended use, moral virtue." *De* has been translated variously as virtue, potency, integrity, or power (for an etymological study, see Nivison 1978-79, and Hall and Ames 1987, 216). The Confucian usage is quite clear; virtue is a matter of moral character and presupposes self-cultivation. The *Laozi* seems to be suggesting a "higher" *de* against any moral achievement attained through repeated effort (e.g., ch. 38). The different translations aim at bringing out the uniquely Daoist sense of the term. Though ambiguous, "virtue" reminds us that the *Laozi* is giving new meaning to an established concept, as opposed to introducing a new concept not found in other schools of Chinese philosophy.

The marriage of Dao and *de* effectively bridges the gap between transcendence and immanence. Traditional commentaries beginning with the *Hanfeizi* often play on the homonymic relation between *de* (virtue) and another graph also pronounced *de*, which means to "acquire" or "obtain" something. *De* is thus what one has "obtained" from (the) Dao, a "latent power" by "virtue" of which any being becomes what it is (Waley 1958, 32). In this sense, the *Laozi* speaks of *de* as that which nourishes all beings (e.g., ch. 51).

Within these parameters, interpretations of *de* follow from the understanding of Dao and *wu*. On the one hand, for Heshanggong and other proponents of the cosmological view, what one has obtained from the Dao refers specifically to one's *qi*-endowment, which determines one's physical, intellectual, moral, and spiritual capacity. Read this way, the *Daodejing* should be translated as the "Classic of the Way and Its Virtue," given that *de* is understood to have emanated from the Dao. On the other hand, for Wang Bi and others who do not subscribe to a substantive view of Dao, *de* represents what is "genuine" or "authentic" (*zhen*) in human beings (e.g., see Wang Bi's commentary on *Laozi* chs. 3, 5, 16, 51). Because *wu* does not refer to any substance or cosmological power, what the *Laozi* means by *de*, the "virtue" that one has "obtained" from Dao, can only be understood as what is originally, naturally present in human beings. In either case, the concept of *de* emerges as a Daoist response to the question of human nature, which was one of the most contested issues in early Chinese philosophy. The two readings of the *Laozi*, despite their differences, agree that there is a prescriptive side to *de*. The empowerment enables a person to conform to the way in which Dao operates. When realized, "virtue" signifies the full embodiment of the Dao or the

flourishing of authenticity. As such, Dao points not only to the "beginning" but also through *de* to the "end" of all things.

The *Laozi* makes use of the concept of *ziran*, literally what is "self (*zi*) so (*ran*)," to describe the workings of Dao. As an abstract concept, it gives no specific information, except to say that Dao does not "model" after anything (ch. 25). However, since "heaven and earth" -- interpreted as nature in most modern studies -- are said to be born of Dao and come to be in virtue of their *de*, the *Laozi* is in effect saying that the ways of nature reflect the function of Dao. In a cosmological reading, this suggests an understanding of nature as governed by the operation of *qi*-energies in an ideal yin-yang system characterized by harmony and fecundity. As interpreted by Wang Bi, the *Laozi* means more generally that there are "principles" (*li*) inherent in nature. Human beings are, in turn, born of heaven and earth and so are "modeled" after them, either in terms of their *qi*-constitution or in the sense that they are governed also by the same basic principles. Usually translated as "naturalness" or "spontaneity," *ziran* thus builds on the concept of *de* in suggesting not only that the power of Dao finds expression in nature but also at the practical level a mode of being and way of action in accordance with the ways of nature.

Nature in the Daoist sense, it is important to note, need not exclude the spiritual and the social. The existence of gods and spirits was hardly questioned in early China. The *Laozi* makes clear that they, too, stem from Dao and form a part of the order of *ziran* (e.g., chs. 39, 60). Further, "nature" encompasses not only natural phenomena but also sociopolitical institutions. The king clearly occupies a central place in the realm of Dao (chs. 16, 25); the family also should be regarded as a "natural" institution (chs. 18, 54). As an ethical concept, *ziran* thus extends beyond the personal to the sociopolitical level. It is worth mentioning that *ziran* remains an influential idea today, especially in conceptions of romantic love and beauty in Chinese thinking.

The concept of *wuwei*, "nonaction," serves to explain naturalness in practice. "Nonaction" is awkward, and some translators prefer "non-assertive action" or "non-aggressive action," but it identifies *wuwei* as a technical term. It does not mean total inaction. Later Daoists may see a close connection between *wuwei* and techniques of spiritual cultivation -- the practice of "sitting in forgetfulness" (*zuowang*) discussed in the *Zhuangzi* is often mentioned in this regard. In the *Laozi*, the concept seems to be used more broadly as a contrast against any form of action characterized especially by self-serving desire (e.g., chs. 3, 37).

It is useful to recall the late Zhou context, where disorder marched on every front. The *Laozi*, one assumes, is not indifferent to the forces of disintegration tearing the country asunder, although the remedy it proposes is subject to interpretation. The problems of political decline are traced to excessive desire, a violation of *ziran*. Nonaction entails at the personal level simplicity and quietude, which naturally follow from having few desires. At the political level, the *Laozi* condemns aggressive measures such as war (ch. 30), cruel punishment (ch. 74), and heavy taxation (75), which reflect but the ruler's own desire for wealth and power. If the ruler could rid himself of desire, according to the *Laozi*, the world would be at peace of its own accord (chs. 37, 57).

In this sense, the *Laozi* describes the ideal sage-ruler as someone who understands and follows *ziran* (e.g., chs. 2, 17, 64). In this same sense, it also opposes the Confucian program of benevolent intervention,

which as the *Laozi* understands it, addresses at best the symptoms but not the root cause of the disease. The Confucian project is in fact symptomatic of the decline of the rule of Dao. Conscious efforts at cultivating moral virtues only accentuate the loss of natural goodness, which in its original state would have been entirely commonplace and would not have warranted distinction or special attention (chs. 18, 38). Worse, Confucian ethics assumes that learning and moral self-cultivation can bring about personal and social improvement. From the Daoist perspective, artificial effort to "improve" things or to correct the order of *ziran* only fuels a false sense of self that alienates human beings from their inherent "virtue."

The concept of nonaction is exceedingly rich. It brings into play a cutting discernment that value distinctions are ideological, that human striving and competitive strife spring from the same source. Nonaction entails also a critique of language and conventional knowledge, which to the Daoist sage has become impregnated with ideological contaminants. The use of paradoxes in the *Laozi* especially heightens this point. For example, the person of Dao is depicted as "witless" or "dumb," whereas people driven by desire appear intelligent and can scheme with cunning (ch. 20). The way of learning, as one would normally understand, "increases" the store of knowledge and adds value to goods and services; in contrast, questioning the very meaning of such "knowledge" and "value," the *Laozi* describes the pursuit of Dao as constantly "decreasing" or chipping away at the artifice built by desire (ch. 48). Driving home the same point, to cite but one more example, the *Laozi* states, "The highest virtue is not virtuous; therefore it has virtue" (ch. 38). In other words, those who fully realize "virtue" in the Daoist sense do not act in the way that men and women of conventional morality typically act or are expected to act. Paradoxes of this kind function as a powerful rhetorical device, which forces the reader, so to speak, to move out of his or her "comfort zone" and to take note of the proposed higher truth of Dao (see also, e.g., chs. 41, 45, 56). In this context, one can also understand some of the provocative statements in the *Laozi* telling the ruler, for example, to keep the people in a state of "ignorance" (ch. 65).

Some scholars would object that this interpretation misses the religious import of the *Daodejing*, while others would question whether it is too eager to defend the philosophical coherence of the classic. Perhaps the *Laozi* in chapter 65 of the current text did mean to tell the ruler literally to keep the people ignorant or stupid for better control, which as a piece of political advice is not that extraordinary. The remarks offered here take nonaction as central to the Daoist view of life, taking into account that the concept of *wuwei* does not only initiate a critique of value but also points to a higher mode of knowledge, action, and being.

At the critical level, the *Laozi* emphasizes the relativity of knowledge and value. Things appear big or small only in relation to other things; knowledge and ignorance are meaningful only in relation to each other. Good and bad, being and nonbeing, and other opposites should be understood in the same light (ch. 2). Distinctions as such are not necessarily problematic; for example, an object can be described as rare or difficult to find as compared with other objects. Problems arise, however, when objects that are rare are deemed more valuable than commonplace objects, when "big" is deemed superior to "small," or in general terms when distinctions become a basis for value discrimination. When certain things or features (e.g., precious stones, reputation, being slim, skin color) are regarded as "beautiful" or "worthy" -- i.e., desirable -- other things will inevitably be deemed "ugly" and "unworthy," with serious social, economic, and political consequences (ch. 3).

The recognition of the relativity of value does not end in a kind of ethical paralysis. The *Laozi* also does not appear to be advocating the obliteration of all distinctions, and by extension civilization as a whole, in a state of mystical oneness. For example, while there is some concern that technology may bring a false sense of progress, the antidote does not lie in a deliberate rejection of technology but rather in a life of natural simplicity and contentment that stems from having few desires (ch. 80). The critique of value demonstrates the way in which desire (*yu*), as distinguished from basic needs, perverts the heart-mind (*xin*) and colors our judgment and experience of reality. Nonaction contrasts sharply with the way people stereotypically act, with profit motives, calculated steps, expectations, longings, regrets, and other expressions of desire. In this way, the *Laozi* aims at making us understand why we do what we do. As a philosophical concept, *wuwei* intimates a mode of being that governs existential engagement at all levels, transforming the way in which we think, feel, and experience the world. It does not stipulate what one ought to do or ought not to do in particular cases -- terms such as quietude, emptiness, and simplicity favored by the *Laozi* describe a general ethical orientation rather than specific practices. Although in following *wuwei* there are things that a person of Dao naturally would not do (e.g., wage a war of aggression), philosophically *wuwei* is not about not doing certain things (thus, military engagement is not ruled out entirely -- e.g., see chs. 67, 68, 69) but suggests a reorientation of perception and value that ideally would bring an end to the dominance of desire and a return to the order of *ziran*.

Nonaction need not exclude meditation or other forms of spiritual practice; the point is rather that once realized, the transformative power of nonaction would ensure not only personal fulfillment but also sociopolitical order. This seems to weigh against a mystical reading of the *Laozi*, if mysticism is understood to entail a kind of personal union with the Dao transcending all political interests. The concept of “virtue,” whether interpreted in terms of authenticity or the purity and fullness of *qi*-energy, depicts a pristine natural and sociopolitical order in which naturalness and nonaction are the norm. The ethics of *wuwei* rests on this insight.

In this interpretive framework, a number of symbols which both delight and puzzle readers of the *Laozi* can be highlighted. Suggestive of its creativity and nurturance, Dao is likened to a mother (e.g., chs. 1, 25). This complements the paradigm of the feminine (e.g., chs. 6, 28), whose “virtue” is seen to yield fecundity and to find expression in yieldingness and non-contention. The infant (e.g., chs. 52, 55) serves as a fitting symbol on two counts. First, it brings out the relationship between Dao and world; second, the kind of innocence and wholesome spontaneity represented by the infant exemplifies the pristine fullness of *de* in the ideal Daoist world.

Natural symbols such as water (e.g., chs. 8, 78) further reinforce the sense of yielding and deep strength that characterizes nonaction. The low-lying and fertile valley (e.g., chs. 28, 39) accentuates both the creative fecundity of Dao and the gentle nurturance of its power. Carefully crafted and ornately decorated objects are treasured by the world, and as such can be used as a powerful symbol for it. In contrast, the utterly simple, unaffected, and seemingly valueless *pu*, a plain uncarved block of wood, brings into sharp relief the integrity of Daoist virtue and of the person who embodies it (e.g., chs. 28, 32). Finally, one may mention the notion of reversal (e.g., chs. 40, 65), which suggests not only the need to “return” to Dao, but also that the Daoist way of life would inevitably appear the very opposite of “normal” existence, and that

it involves a complete revaluation of values.

In sum, any interpretation of the current *Laozi* as a whole must take into account the way in which *wuwei* and *ziran* provide a guide to the good life. Specifically, two related issues need to be addressed. First, naturalness and nonaction are seen to reflect the function of the nameless and formless Dao. As such, ethical ideals are anchored in a non-empirical view of nature, which raises the concept of *de* to a higher level than "virtues" in the sense of moral attainments. The understanding of *de*, however, is dependent on that of Dao, which in turn hinges on the interpretation of *wu* as either original substance or nonbeing. Both readings are plausible. A cosmological reading is attractive also because it aligns the *Laozi* more closely with other early Chinese philosophical texts. An ontological reading emphasizing the conceptual nature of *wu* has the advantage of allowing the reader to draw out more readily the philosophical implications of the *Laozi*; it cannot be ruled out, especially if the only appeal is to a perceived evolutionary development in Chinese philosophy.

The second issue concerns the ethics of the *Laozi*. As mentioned, one main approach to the classic stresses its political orientation. In many chapters, the text seems to be addressing the ruler or the ruling elite, explaining to them the ideal government of the Daoist sage. This is not surprising given the Zhou context and given that the production of and access to written documents were generally the preserve of the ruling class in ancient China. But this need not restrict interpretation to politics in the narrow sense of statecraft or political strategies. In the light of the emphasis on *ziran* and *wuwei*, there is sufficient evidence that the *Laozi* views politics in a larger ethical context. The more difficult question is whether *wuwei* suggests a kind of moral relativism that would render any suggestion of a "higher" Daoist truth problematic.

If things are relative, they should be regarded as being of equal value. There are no objective criteria beyond relative utility for specific purposes in favoring, for example, what is "soft" over something "hard." Yet, as D. C. Lau (1963) has pointed out, at the ethical level the *Laozi* seems to favor the "lower" term -- for example, the "weakness" or "yieldingness" of water is singled out as a metaphor for ideal ethical conduct (ch. 78). We have suggested that the *Laozi* makes use of a critical mode to deconstruct conventional knowledge and value. The sense of moral relativity, however, gives way to a "higher" ethical mode based on insight into the order of *ziran* derived from an idealized view of nature. In this sense, the *Laozi* thus also makes use of such expressions as the "highest good" (ch. 8) and "highest virtue" (ch. 38). *Wuwei* ultimately derives its meaning from *wu*, which as an ethical orientation privileges "not having" over the constant strivings of the mundane world. This constitutes a powerful critique of a world given to the pursuit of wealth and power. More important, in being "empty," the person of Dao is said to be "full"; without desire, he or she is able to rediscover the riches of *ziran* and finds fulfillment. This invites reflection and continuing dialogue with the *Laozi*.

Bibliography

- Allan, Sarah, and Crispin Williams. 2000. *The Guodian Laozi*. Berkeley: Society for the Study of Early China and the Institute of East Asian Studies, University of California.

- Barrett, T. H. 1996. *Taoism under the T'ang*. London: Wellsweep Press.
- Baxter, William H. 1998. "Situating the Language of the *Lao-tzu*: The Probable Date of the *Tao-te-ching*." In *Lao-tzu and the Tao-te-ching*, edited by Livia Kohn and Michael LaFargue. Albany: State University of New York Press. pp. 231-53.
- Bokenkamp, Stephen. 1989. "Death and Ascent in Ling-pao Taoism." *Taoist Resources* 1.2: 1-20.
- ------. 1993. "Traces of Early Celestial Master Physiological Practice in the *Xiang'er* Commentary." *Taoist Resources* 4.2: 37-52.
- ------. 1997. *Early Taoist Scriptures*. Berkeley: University of California Press.
- Boltz, William G. 1984. "Textual Criticism and the Ma Wang Tui *Lao-tzu*." *Harvard Journal of Asiatic Studies* 44: 185-224.
- ------. 1985. "The *Laozi* Text that Wang Pi and Ho-shang Kung Never Saw." *Bulletin of the School of Oriental and African Studies* 48: 493-501.
- ------. 1993. "Lao tzu Tao te ching." In *Early Chinese Texts: A Bibliographical Guide*, edited by Michael Loewe. Berkeley: University of California, Institute of East Asian Studies. pp. 269-92.
- ------. 1996. "Notes on the Authenticity of the So Tan Manuscript of the *Lao-tzu*." *Bulletin of the School of Oriental and African Studies* 59: 508-15.
- ------. 1999. "The Fourth-Century B.C. Guodiann Manuscripts From Chuu and the Composition of the *Laotzyy*." *Journal of the American Oriental Society* 119.4: 590-608.
- Brooks, E. Bruce, and A. Taeko Brooks. 1998. *The Original Analects*. New York: Columbia University Press.
- Capra, Fritjof. 1975. *The Tao of Physics*. London: Wildwood House.
- Chan, Alan K. L. 1991a. "The Formation of the Heshang gong Legend." In *Sages and Filial Sons: Mythology and Archaeology in Ancient China*, edited by Julia Ching and R. Guisso. Hong Kong: Chinese University Press. pp. 101-34.
- ------. 1991b. *Two Visions of the Way: A Study of the Wang Pi and Ho-shang Kung Commentaries on the Lao-tzu*. Albany: State University of New York Press.
- ------. 1998a. "A Tale of Two Commentaries: Ho-shang-kung and Wang Pi on the *Lao-tzu*." In *Lao-tzu and the Tao-te-ching*, edited by Livia Kohn and Michael LaFargue. Albany: State University of New York Press. pp. 89-117.
- ------. 1998b. "The Essential Meaning of the Way and Virtue: Yan Zun and '*Laozi* Learning' in Early Han China." *Monumenta Serica* 46: 105-127.
- ------. 1999. "The *Daodejing* and Its Tradition." In *Daoism Handbook*, edited by Livia Kohn. Leiden: E. J. Brill. pp. 1-29.
- Chan, Wing-tsit. 1963. *The Way of Lao Tzu*. Indianapolis: Bobbs-Merrill.
- Chen, Ellen M. 1969. "Nothingness and the Mother Principle in Early Chinese Taoism." *International Philosophical Quarterly* 9.3: 391-405.
- ------. 1974. "Tao as the Great Mother and the Influence of Motherly Love in the Shaping of Chinese Philosophy." *History of Religions* 14.1: 51-64.
- ------. 1989. *The Tao Te Ching: A New Translation with Commentary*. New York: Paragon House.
- Ching, Julia. 1997. *Mysticism and Kingship in China*. Cambridge: Cambridge University Press.
- Creel, Herlee G. 1970. *What is Taoism? And Other Studies in Chinese Cultural History*. Chicago: University of Chicago Press.
- Ding Sixin. 2000. *Guodian Chumu zhujian sixiang yanjiu*. Beijing: Dongfang chubanshe.

- Emerson, John. 1995. “A Stratification of *Lao Tzu*.” *Journal of Chinese Religions* 23: 1-28.
- Erkes, Eduard. 1935. “Arthur Waley’s Laotse-Übersetzung.” *Artibus Asiae* 5: 285-307.
- ------, trans. 1958. *Ho-shang Kung’s Commentary on Lao-tse*. Ascona: Artibus Asiae.
- Fukui Kōjun, et al. 1983. *Dōkyō*. 3 vols. Tokyo: Hirakawa shuppansha.
- Fung Yu-lan. 1983. *A History of Chinese Philosophy*. Translated by Derk Bodde. 2 vols. Princeton: Princeton University Press.
- Gao Heng. 1981. *Chongding Laozi zhenggu*. Taipei: Xinwenfeng. Orig. pub. 1940.
- Graham, A. C. 1986. *The Origins of the Legend of Lao Tan*. Singapore: Institute of East Asian Philosophies. Reprinted in A. C. Graham, *Studies in Chinese Philosophy and Philosophical Literature*. Albany: State University of New York Press, 1990. pp. 111-24.
- ------. 1989. *Disputers of the Tao: Philosophical Argument in Ancient China*. La Salle, IL: Open Court.
- Girardot, Norman J. 1983. *Myth and Meaning in Early Taoism*. Berkeley: University of California Press.
- Hall, David, and Roger Ames. 1989. *Thinking Through Confucius*. Albany: State University of New York Press.
- Hansen, Chad. 1992. *A Daoist Theory of Chinese Thought*. New York: Oxford University Press.
- Hardy, Julia. 1998. “Influential Western Interpretations of the *Tao-te-ching*.” In *Lao-tzu and the Tao-te-ching*, edited by Livia Kohn and Michael LaFargue. Albany: State University of New York Press. pp. 165-88.
- Hatano Tarō. 1979. *Rōshi dōtokukyō kenkyū*. Tokyo: Kokusho kankōkai.
- Hawkes, David, trans. 1985. *The Songs of the South: An Anthology of Ancient Chinese Poems by Qu Yuan and Other Poets*. Harmondsworth: Penguin Books.
- Henricks, Robert. 1982. “On the Chapter Divisions in the *Lao-tzu*.” *Bulletin of the School of Oriental and African Studies* 45: 501-24.
- ------. 1989. *Lao-Tzu Te-Tao Ching: A New Translation Based on the Recently Discovered Ma-wang-tui Texts*. New York: Ballantine Books.
- ------. 2000. *Lao Tzu’s Tao Te Ching: A Translation of the Startling New Documents Found at Guodian*. New York: Columbia University Press.
- Hoff, Benjamin. 1982. *The Tao of Pooh*. New York: E.P. Dutton.
- Hurvitz, Leon. 1961. “A Recent Japanese Study of *Lao-tzu*: Kimura Eiichi’s *Rōshi no shin kenkyū*.” *Monumenta Serica* 20: 311-67.
- Jaspers, Karl. 1974. *Anaximander, Heraclitus, Parmenides, Plotinus, Lao-tzu, Nagarjuna*. From *The Great Philosophers*, Vol. 2, *The Original Thinkers*. Ed. Hannah Arendt; trans. Ralph Manheim. A Harvest Book. New York and London: Harcourt Brace Jovanovich, 1974.
- Jiang Xichang. 1937. *Laozi jiaogu*. Taipei: Dongsheng.
- Kaltenmark, Max. 1969. *Lao Tzu and Taoism*. Translated by Roger Greaves. Stanford: Stanford University Press.
- Kimura Eiichi. 1959. *Rōshi no shin kenkyū*. Tokyo: Sōbunsha.
- Kohn, Livia. 1992. *Early Chinese Mysticism: Philosophy and Soteriology in the Taoist Tradition*. Princeton: Princeton University Press.
- ------. 1998a. “The Lao-tzu Myth.” In *Lao-tzu and the Tao-te-ching*, edited by Livia Kohn and Michael LaFargue. Albany: State University of New York Press. pp. 41-62.

- Kusuyama Haruki. 1979. *Rōshi densetsu no kenkyū*. Tokyo: Sōbunsha.
- LaFargue, Michael. 1992. *The Tao of the Tao Te Ching*. Albany: State University of New York Press.
- ------. 1994. *Tao and Method: A Reasoned Approach to the Tao Te Ching*. Albany: State University of New York Press.
- ------, and Julian Pas. 1998. “On Translating the *Tao-te-ching*.” In *Lao-tzu and the Tao-te-ching*, edited by Livia Kohn and Michael LaFargue. Albany: State University of New York Press. pp. 277-301.
- Lau, D. C. 1963. *Lao Tzu Tao Te Ching*. Harmondsworth: Penguin Books.
- ------. 1982. *Chinese Classics: Tao Te Ching*. Hong Kong: Chinese University Press.
- Legge, James. 1962. *The Texts of Taoism*, Part 1. The Sacred Books of the East, vol. 39. New York: Dover. Orig. pub. 1891.
- Lin, Paul J. 1977. *A Translation of Lao Tzu’s Tao Te Ching and Wang Pi’s Commentary*. Ann Arbor: Center for Chinese Studies, University of Michigan.
- Liu Cunren. 1969. “Daozangben sansheng zhu Daode jing zhi deshi.” *Chongji xuebao* 9.1: 1-9.
- ------. 1971-73. “Daozangben sansheng zhu Daode jing huijian.” Pts. 1-3. *Zhongguo wenhua yanjiusuo xuebao* 4: 287-343; 5: 9-75; 6: 1-43.
- Liu, Xiaogan. 1991. “*Wuwei* (Non-Action): From *Laozi* to Huainanzi.” *Taoist Resources* 3.1: 41-56.
- ------. 1994. *Classifying the Zhuangzi Chapters*. Ann Arbor: University of Michigan, Center for Chinese Studies.
- ------. 1997. *Laozi*. Taipei: Dongda.
- ------. 1998. “Naturalness (*Tzu-jan*), the Core Value in Taoism: Its Ancient Meaning and Its Significance Today.” In *Lao-tzu and the Tao-te-ching*, edited by Livia Kohn and Michael LaFargue. Albany: State University of New York Press. pp. 211-28.
- Loewe, M. and Shaughnessy, E., eds., 1999, *Cambridge History of Ancient China*. Cambridge: Cambridge University Press.
- Lynn, Richard John, trans. 1994. *The Classic of Changes: A New Translation of the I Ching as Interpreted by Wang Bi*. New York: Columbia University Press.
- ------. 1999. *The Classic of the Way and Virtue: A New Translation of the Tao-te ching of Laozi as Interpreted by Wang Bi*. New York: Columbia University Press.
- Ma Xulun. 1965. *Laozi jiaogu*. Hong Kong: Taiping shuju.
- Mair, Victor. 1990. *Tao Te Ching: The Classic Book of Integrity and the Way*. New York: Bantam Books.
- Mukai Tetsuo. 1994. “Rokuto no kisoteki kenkyū.” *Tōhō shūkyō* 83: 32-51.
- Needham, Joseph. 1956. *Science and Civilisation in China*. Vol. 2. *History of Scientific Thought*. Cambridge: Cambridge University Press.
- Nivison, David. 1978-79. “Royal ‘Virtue’ in Shang Oracle Inscriptions.” *Early China* 4: 52-55.
- Pelliot, Paul. 1912. “Autour d’une traduction sanscrite du Tao To King.” *T’oung-pao* 13: 351-430.
- Rand, Christopher. 1979-80. “Chinese Military Thought and Philosophical Taoism.” *Monumenta Serica* 34: 171-218.
- Rao Zongyi. 1955. “Wu Jianheng er’nian suotan xieben Daode jing canjuan kaozheng.” *Journal of Oriental Studies* 2: 1-71.

- ------. 1991. *Laozi Xiang'er zhu jiaozheng*. Shanghai: Guji.
- Robinet, Isabelle. 1977. *Les commentaires du Tao to king jusqu'au VIIe siècle*. Paris: Presses Universitaires de France.
- ------. 1998. "Later Commentaries: Textual Polysemy and Syncretistic Interpretations." In *Lao-tzu and the Tao-te-ching*, edited by Livia Kohn and Michael LaFargue. Albany: State University of New York Press. pp. 119-42.
- Rump, Ariane. 1979. *Commentary on the Lao-tzu by Wang Pi*. In Collaboration with Wing-tsit Chan. Honolulu: University of Hawaii Press.
- Saso, Michael. 1994. *A Taoist Cookbook: With Meditations Taken from the Laozi Daode Jing*. Boston: Tuttle Press.
- Schwartz, Benjamin. 1985. *The World of Thought in Ancient China*. Cambridge, MA: Harvard University Press.
- Seidel, Anna. 1969. *La divinisation de Lao-tseu dans le taoïsme des Han*. Paris: Ecole Française d'Extrême-Orient.
- Shima Kunio. 1973. *Rōshi kōsei*. Tokyo: Kyūkoshoin.
- Tsai Chih Chung, et al. 1995. *The Silence of the Wise: The Sayings of Laozi*. Asiapac Comic Series. Singapore: Asiapac Books. Orig. pub. 1989.
- Wagner, Rudolf. 1980. "Interlocking Parallel Style: Laozi and Wang Bi." *Études Asiatiques* 34.1: 18-58.
- ------. 1989. "The Wang Bi Recension of the *Laozi*." *Early China* 14: 27-54.
- ------. 2000. *The Craft of a Chinese Commentator: Wang Bi on the Laozi*. Albany: State University of New York Press.
- Waley, Arthur. 1958. *The Way and Its Power: A Study of the Tao Te Ching and Its Place in Chinese Thought*. New York: Grove Press.
- Wang Ming. 1984. "Lun Laozi bingshu." In Wang Ming, *Daojia he daojiao sixiang yanjiu*. Chongqing: Zhongguo shehui kexue chubanshe. pp. 27-36.
- Wang Zhongmin. 1981. *Laozi kao*. Taibei: Dongsheng. Orig. pub. 1927.
- Welch, Holmes. 1965. *Taoism: The Parting of the Way*. Boston: Beacon Press.
- Xiong Tieji, et al. 1995. *Zhongguo Laoxueshi*. Fuzhou: Fujian renmin chubanshe.
- Xu Kangsheng. 1985. *Boshu Laozi zhuyi yu yanjiu*. Zhejiang renmin chubanshe.
- Yan Lingfeng. 1957. *Zhongwai Laozi zhushu mulu*. Taipei: Zhonghua congshu weiyuanhui.
- ------. 1965a. *Wuqiubeizhai Laozi jicheng, chubian*. Taipei: Yiwen yinshuguan.
- ------. 1965b. *Wuqiubeizhai Laozi jicheng, xubian*. Taipei: Yiwen yinshuguan.

Other Internet Resources

- Taoism Information Page (maintained by Gene Thursby, Department of Religion, University of Florida)
  This forms a part of the World Wide Web Virtual Library and contains useful links to internet resources on Daoism and Chinese philosophy in general. The section on the *Daodejing* currently only lists a handful of English translations; but I expect the site to grow.
- Taoist Culture and Information Center

This site forms a part of the Daoist Culture Database and provides a simple introduction to Religious Daoism. It is the work of the Fung Ying Seen Koon, a major Daoist temple and a non-profit charitable organization active in community work in Hong Kong. Stephen Bokenkamp, a specialist in religious Daoism and professor at Indiana University is listed as a translator. Several well-known scholars from China are listed as advisors. The entries I have read are reliable, albeit very short.

- Taoist Studies in the World Wide Web, maintained by Fabrizio Pregadio.
  This is an excellent site, which also has links to other websites on Daoism.
- The China WWW Virtual Library. Internet Guide for China Studies, maintained at the Institute of Chinese Studies, Heidelberg University
  This is one of the most comprehensive websites on China Studies available today. Highly recommended.
- The Institute for Research in the Humanities, Kyoto University
  The "Jinbunken" has been building a large textual database for sinological research. The above link will take you to the English version page, but see also the following page for specialists.

Related Entries

Confucius | Mencius | Mohism | Taoism | Zhuangzi

Acknowledgments

Transliteration of Chinese terms in this article follows the *hanyu pinyin* romanization system, except for a few proper names and quotations. Some of the material presented above first appeared in “The *Daodejing* and Its Tradition,” *Daoism Handbook*, edited by Livia Kohn [Leiden: E. J. Brill, 1999], pp. 1-29; permission by the publisher to rework them here is gratefully acknowledged.

Alan K. L. Chan
National University of Singapore
alanchan@nus.edu.sg

First published: December 14, 2001
Content last modified: December 14, 2001

Stanford Encyclopedia of Philosophy

Zhuangzi

'Zhuangzi' is the name of the second foundational text of the Daoist philosophical and religious tradition and the name of the putative author of this text, who early historical sources say flourished between about 350 and 300 B.C.E. As one of the two most popular Daoist texts in the Chinese tradition, the *Zhuangzi* has been the subject of more than sixty major East Asian commentaries since the third century C.E., some of which contain philosophically significant interpretations of the text. The most important of these are the commentary by Guo Xiang, which focuses on his understanding of Zhuangzi's philosophy of spontaneity, the commentary by Cheng Xuanying (ca. 620-670), a religious Daoist master with strong interests in emptiness theory, and commentaries by the following Sung and Ming dynasty literati scholars: Wang Pang (1042-76), Lin Xiyi (ca. 1200-73), Lo Miandao (ca. 1240-1300), and Jiao Hong (1541-1620). None of these has been fully translated into English and modern studies of them in any language are few, thus yielding a fertile field for future research. The existence of these commentaries demonstrates the great popularity of the *Zhuangzi* among Chinese literati who saw within it support for a withdrawal from a life of social and political service into a private life of reclusion and self-cultivation. If Confucianism came to stand for the foundational philosophy of this ethos of self-sacrifice, the Daoism of the *Zhuangzi* came to stand for its opposite, the escape from societal pressure into an individual path of freedom. While thus important to literati scholars, the work was also significant for Daoist religious practitioners who often took ideas and themes from it for their meditation practice, for example Sima Chengzhen's 'Treatise on Sitting and Forgetting' (ca. 660 C.E.) (Kohn 1987).

1. Provisos Concerning the Text

What we know of the philosophy of Zhuangzi comes primarily from this work but readers of translations of the received recension (Watson, Graham 1981, Mair 1998) should be aware of the following provisos.

First, the received recension contains thirty-three chapters and is not the original recension of the text. Guo Xiang (d. 312 C.E.) revised a fifty-two chapter original recension first listed in Imperial bibliographies circa 110 C.E. by removing material he thought was superstitious and generally not of philosophical interest to his literati sensibilities. He appended a philosophical commentary to the text that became famous and within four centuries his shorter and snappier expurgated recension became the only one known. This recension is traditionally divided into three sections: ‘Inner Chapters’ (1-7), ‘Outer Chapters’ (8-22), ‘Miscellaneous Chapters’ (23-33). This division is quite old and is likely to have been part of the original recension.

Second, the *Zhuangzi* text is clearly not the work of a single author and it is difficult to affix definitive authorship to any one person. At the very least there are five authorial voices best summarized by A.C. Graham: the historical Zhuangzi, later followers of Zhuangzi, followers influenced by the individualist thinker Yang Zhu, a ‘Primitivist’ Daoist author whose ideas are akin to those of the *Daode jing*, and the ‘Syncretist’ Daoist authors who Graham thinks compiled the first recension of the text (Graham 1979).

While it is true that many of the philosophical insights for which this work has become renowned in China and more recently in the West are found in the ‘Inner Chapters’ that have traditionally been ascribed to the historical Zhuangzi, we cannot fully understand these ideas and their significance without grasping how they relate to the entire thirty-three chapter text and the variety of ideas it contains. In this entry we shall accept the convention that a historical Zhuangzi authored most of the seven ‘Inner chapters’ while noting that there have been questions about this attribution that are not sufficient to overturn this traditional belief.

2. The Philosophy of the Inner Chapters

The *Zhuangzi* has become renowned for a series of original insights into human nature and the nature of the cosmos and many of these are found in the ‘Inner chapters.’ These insights are communicated in a variety of literary styles: didactic narratives, poetry, and very short prose essays. Like its famous companion, the *Daode jing*, the *Zhuangzi* is grounded in the complementary ideas of Dao and De. Dao, the ‘Way,’ is an ineffable monistic principle that infuses and guides the spontaneous processes of all phenomena; De, ‘Inner Power,’ is the realized manifestation of this Way within all phenomena. Despite sharing these foundational ideas, these two Daoist works discuss them very differently. The *Daode jing* often presents the characteristics and features of the Way in a direct discursive analysis (e.g., DDJ 1: "The Way that can be told of is not the Constant Way"). On the other hand, the *Zhuangzi* often approaches the Way indirectly through narratives and poetry. Witness the following rhetorical pointing to the Way:

> Heaven turns circles, yes!
> Earth sits firm, yes!
> Sun and Moon vie for a place, yes!
> Whose is the bow that shoots them?
> Whose is the net that holds them?
> Who is it sits with nothing to do and gives them the push that sends them away?
>
> (Graham 1981, p. 49; all following translations based on this work)

Or consider this rumination on epistemological relativity that ends with a vivid pointing to the Way:

> What is It is also Other, what is Other is also It. There they say, "this is true and that is false" from one point of view; here we say, "this is true and that is false" from another point of view. Are there really It and Other? Or really no It and Other. Where neither It nor Other finds its opposite is called the axis of the Way. When once the axis is found at the center of the circle there is no limit to responding with either, on the one hand no limit to what is it, on the other no limit to what is not.... (Graham, p. 53, modified)

This questioning of the certainty of knowledge from any normal human viewpoint is another hallmark of the 'Inner chapters,' as is the considerable degree of humor and irony with which the most profound insights into the cosmos are presented. This is true as well for Zhuangzi's presentation of Inner Power, which is done through narratives in which the most highly evolved human beings in terms of realizing this power are skilled butchers (see chapter 3) and criminals punished by mutilation (see chapter 5). This flaunting of societal prejudices is another way in which Zhuangzi challenges entrenched beliefs and demonstrates the breathtaking freedom from fixed conventions that has delighted readers for two millennia.

The Zhuangzi of the 'Inner Chapters' is also known for a thorough questioning of the canons, methods, and value of discursive reason and logic as practiced by contemporary thinkers in the traditions of the Mohists, Confucians, and Terminologists (*ming jia*). He is particularly critical of assumptions such as the one to one correspondence between words and the objects to which they refer that is an offshoot of the Confucian doctrine of the rectification of names. He demonstrates that naming is purely arbitrary and conventional and cannot be used to give any objective certainty about the world. Furthermore no matter how sophisticated the logic involved, no argument can establish objective truths because all knowing remains confined to the standpoint of the knower:

> Gaptooth put a question to Wang Ni: "Would you know something which all things agreed is true?" "How would I know that?" he replied. "Would you know what you did not know?" "How would I know that?" he replied again. "Then does no thing know anything?" "How would I know that?" he replied again. He then continued, "however, let me try to put this in words: how do I know that what I call knowing is not ignorance? How do I know

> that what I call ignorance is not knowing? ... Gibbons are sought by baboons as mates, elaphures like the company of deer, loaches play with fish. Maoqiang and Lady Li were beautiful in the eyes of men but when the fish saw them they plunged into the deep and when the birds saw them they flew away. Which of these four knows what is truly beautiful in the world? In my judgment, the principles of Humaneness and Rightness, the paths of True and False are inextricably confused: how could I know how to discriminate between them?" (Graham, p. 58, mod.)

Inspired by such ideas -- which are principally located in the second chapter of the *Zhuangzi*, entitled 'Discourse on Seeing all things as Equal' (*Qiwu lun*) -- Western comparative philosophers during the past two decades have taken a considerable interest in Zhuangzi and have engaged in spirited debate about whether he can be classified as a skeptic, a relativist, or a perspectivist. For example, Hansen argues that Zhuangzi espouses a 'perspectival relativism,' that shows that all discrimination and classification are relative to some changeable context of judgment (Hansen 1983). Relative judgments yield relative knowledge and so there are no standpoints from which anything can be known to be objectively true. Raphals finds that in this chapter Zhuangzi employs skeptical methods but not skeptical doctrines (Raphals 1996). This is because while denying the objectivity of common forms of knowledge ('small knowledge' in the text) Zhuangzi does acknowledge the existence of a greater form of knowledge ('illumination') and hence does not advance a 'true skepticism.' Ivanhoe argues that Zhuangzi was neither a sense skeptic nor an ethical skeptic, but that he definitely was both an epistemological skeptic about intellectual (in contrast to intuitive) knowledge and a language skeptic who doubted distinctions between right and wrong and the ability of words to express the Way (Ivanhoe 1993).

Yearley and Roth maintain that in addition to demonstrating elements of skepticism and relativism, that Zhuangzi was also a mystic. Yearley argues for an 'intraworldly mysticism' in which the goal is not union with some unchanging monistic principle but instead full participation in the natural world (Yearley 1983). Roth sees a 'bimodal' mystical experience in Zhuangzi that shows evidence for Stace's categories of introvertive and extrovertive mysticm (Roth 2000). He argues that the greater form of knowledge seen by Raphals and the acceptance of intuitive knowledge seen by Ivanhoe derive from a firm grounding in a meditative practice attested to in both the *Zhuangzi* and in many other sources of early Daoism. It is this practice that Roth calls 'inner cultivation' (Roth 1999). Simply put, it involves sitting quietly and systematically circulating the breath until mind and body become tranquil and the contents of consciousness gradually empty. Taken to its ultimate levels, this apophatic practice gives the adept a direct experience of the Way. In the 'Inner chapters' Zhuangzi talks of meditation in two famous narratives, one in which, ironically, his spokesman is the uncontemplative Confucius who teaches his favorite disciple Yan Hui about a technique he calls 'the fasting of the mind,' and the other in which the tables are turned and Yan teaches Confucius about 'sitting and forgetting.' In the former Confucius gives the following instructions:

> ...Unify your attention. Do not listen with the ears, listen with the mind. Do not listen with the mind but listen with the vital breath (*qi*). The ears only listen to sounds. The mind is only aware of its objects. But to focus on the vital breath is to be empty and await the arising of objects. It is only the Way that settles in emptiness. Emptiness is the fasting of

> the mind. (compare Graham, 68)

In the latter, Yan teaches Confucius:

> I let organs and members drop away, dismiss hearing and eyesight, part from the body and expel knowledge, and merge with the Great Pervader. This is what I mean by 'just sit and forget.' (Graham, 92 mod.)

Both passages attest to this apophatic inner cultivation practice in which all normal perceptions and thoughts are removed from consciousness and one eventually achieves union with the Way (the 'Great Pervader'). This yields what Stace has called an 'introvertive mystical experience' (Stace, 1960) Yet for Zhuangzi this experience, although profound, is not the ultimate goal. Speaking metaphorically, through Confucius, he says: 'to stop making footprints is easy but it is difficult to walk without touching the ground.'(comp. Graham 69). This type of ungrounded 'walking' has a significant epistemological dimension: a distinctive mode of cognition that Zhuangzi refers to as 'flowing' (*yin-shi*: literally 'to affirm by following along') in opposition to the 'fixed' mode of cognition (*wei-shi*: literally 'to affirm by forcing') that is bound to one individual perspective. Zhuangzi playfully contrasts these modes in the following famous story:

> A monkey keeper handing out nuts said, "Three every morning and four every evening." The monkeys were all in a rage. "All right," he said, "four every morning and three every evening." The monkeys were all delighted. Without anything being missed out either in name or in substance, their pleasure and anger were put to use: his too was a flowing cognition (*yin-shi*). This is why the sage smoothes things out with his flowing categories and stays at the point of rest on the Potter's Wheel of Heaven... (Graham, 54 mod.)

The monkeys exemplify the fixed mode of cognition in their rigid attachment to one and only one way of seeing the seven nuts. In this they symbolize all the contemporary intellectual traditions -- Confucians, Mohists, Terminologists -- that were arguing the truth of their individual positions against all others. The keeper is able to shift his conceptual categories to harmonize with those of the monkeys because he is free of attachment to any one particular way of seeing the nuts. His is a flowing cognition that adapts spontaneously to the situation, an 'illumined'(*ming*) awareness that exhibits an intuitive knowledge of how to act without even knowing that it is acting. Zhuangzi also calls this 'illuminating things with the light of heaven.' For him, 'heaven' stands for the spontaneous and intuitive aspect of our being that emerge when someone is grounded in the empty Way.

Zhuangzi further makes clear that abandonment of fixed cognition is concomitant with abandonment of attachment to the self: "Without an Other there is no Self; without Self, no choosing one thing rather than another." (Graham, 51) That is, if you lose the distinction between self and other then you lose the self and with it, any bias towards choosing one thing rather than another: "No thing is not 'Other;' no thing is not 'It.' If you treat yourself too as 'Other,' they do not appear. If you know of yourself you know of them" (Graham, 52). That is, in this situation It and Other do not appear because 'treating yourself as

Other' involves abandoning attachment to your self -- in other words having the same degree of attachment to Self as you do to Other. This lack of self-attachment is an essential characteristic of the free and spontaneously functioning consciousness of Zhuangzi's 'flowing cognition.' That this flowing mode, which is similar to Yearley's 'intraworldly mysticism' and which Roth asserts is a type of 'extrovertive mysticism,' develops from a union with the Way is articulated in the following passage from the 'Discourse on Seeing all things as Equal':

> If being 'so' is inherent in a thing ... then from no perspective would it not be 'so' ... . Therefore when a fixed cognition picks out a stalk from a pillar, a hag from beautiful Xishi, things however peculiar and incongruous, the Way pervades and unifies them (*tong wei yi*). As they divide they develop, as they develop they dissolve. All things whether developing or dissolving revert to being pervaded and unified. Only those who penetrate this know how to pervade and unify things. Fixed cognition they do not use, but find lodging places in daily life. It is in daily life that they make use of this perspective. It is in making use of this perspective that they pervade things. It is in pervading things that they attain it. And when they attain it they are almost there. 'Flowing cognition' comes to an end. It ends and when it does, that of which we do not know what is 'so' of it, we call 'the Way.' (Graham, 53-4 mod.)

Here Zhuangzi argues that there are no perspectives from which a thing is always 'so,' is always true. Common fixed cognition clearly differentiates things such as a stalk and a pillar and so on and it simultaneously makes preferences based on these perceptual distinctions. However it is only the Way that can pervade these things and unify them. That is, it is the one and only perspective from which all things are seen just as they are, without bias and without preference, from which 'all things are seen to be equal.' It is just this kind of seeing that is the essential defining characteristic of the 'great knowledge' or 'illumination' of the 'flowing cognition' that is developed by those rare sages who can penetrate through 'fixed cognition' and who can, as the Way does, 'pervade and unify' all things.

Using this ability, sages find temporary 'lodging-places,' i.e. viewpoints to which they are completely unattached within daily life. Yet even this 'flowing cognition' does eventually rest on something more profound: it comes to an end in the experience of union with the Way, described here as 'that of which we do not know what is so.' Knowing what is 'so' of something implies knowing its essential truth as an object of cognition and hence from the dualistic perspective of Self and Other. Yet the Way can never be known as an object: it can only be realized when the distinction between Self and Other dissolves in the introvertive mystical experience of uniting with the Way. Thus the extrovertive mystical experience of 'pervading and unifying' depends on the introvertive mystical experience of union referred to elsewhere in chapter six as 'merging with the Great Pervader' (*tong yu datong*). Once one temporarily loses the Self in this union and then returns to the everyday dualistic world one is no longer attached to Self. 'Flowing cognition' arises from this detachment.

Hence Zhuangzi's renowned questioning of logic and his skepticism and relativism are based upon this shift from fixed cognition to flowing cognition, from self-centered perspective to 'Way-centered' perspective. His epistemological critique is thus applied to knowledge derived from 'fixed cognition.'

‘Flowing cognition’ is exempted from this critique because it is this continually changing ‘Way-centered’ position from which the critique is made.

These complementary mystical experiences (‘merging with the Great Pervader’ and pervading and unifying using flowing cognition) are critical for the understanding of the other important philosophical themes for which the *Zhuangzi* is renowned. Naturalness and spontaneity arise directly from the ‘flowing cognition’ that is free of attachment to any one perspective. When sages act from this mode they can respond without self-consciousness to whatever situation in which they find themselves. Yet their spontaneity is grounded not in their individual separate selves but rather in the Way; we might say they have moved from a ‘self-centered’ perspective to a ‘Way-centered’ perspective. Thus living at the ‘axis of the Way’ they are able to blend the perspectives of Heaven and of human beings:

> The Realized Ones (*zhenren*) ... were one with what they liked and one with what they disliked, one when they were one and one when they were not one. When one, they were of Heaven’s faction and when not one they were of the human faction. Someone in whom neither Heaven nor human is victor over the other, this is what is meant by ‘Realized Ones.’ (Graham p. 85 mod.)

This freedom from attachment to any individual perspective that characterizes Zhuangzi’s ‘flowing cognition’ also leads to the freedom from fear of death and acceptance of it as part of the natural processes of life that is another of the famous themes of this work. In this narrative from chapter six a dying Daoist master addresses his friends:

> A child that has father and mother goes east and west, north and south, and has only their commands to obey. For humans the Yin and Yang are more than father and mother. As something other than me approaches, I am dying; if I were to refuse to listen it would be defiance on my part so how can I blame it? That Vast Clod of Soil (the Way) loaded me with a body, had me toiling through a life, eased me with old age, rests me with death. Therefore that I found it good to live is the very reason that I find it good to die. If today a master sword smith were smelting metal and the metal should jump up and say "I insist on being made into an Excalibur," the sword smith would surely think it metal with a curse on it. If now having once happened on the shape of a human being I were to say, "I’ll be a human, nothing but a human," that which fashions and transforms us would surely think me a baleful sort of person. Now if once and for all I think of heaven and earth as a vast foundry and the fashioner and transformer as a master smith, wherever I am going why should I object? ... (Graham 88-9 mod.)

3. The Philosophy of the Outer and Miscellaneous Chapters

3.1 The Disciples of Zhuangzi

With writings as profound and vibrant as these the historical Zhuangzi must have had quite a devoted group of followers and it is to them -- in all likelihood -- that we owe both the transmission of his ideas beyond his lifetime and at least six chapters of new material, much of it consisting of narratives written in the style of the ‘Inner chapters’ but generally not demonstrating the same creativity and rhetorical skill. Zhuangzi is a figure in about one quarter of these narratives, which were probably based on stories told by his immediate disciples and written down after his death. The chapters in this section, 17-22, are almost completely devoid of the philosophical essays, jottings, or even the diatribes we find in the first third of the book. Yet they contain some of the most famous narratives in the entire text.

The ‘autumn floods’ passage that dominates chapter 17 continues the theme of the relativity of different perspectives and the wholeness of the Way-centered perspective. This epistemological relativity is also the theme of the well-known dialogue between Zhuangzi and his Terminologist friend and debating rival Huishi while strolling over the Hao River Bridge found in this chapter. Chapter 18, ‘Complete Happiness,’ centers around the theme of the acceptance of death as part of the natural processes of Heaven and Earth and contains the famous narrative about Huishi’s visit to Zhuangzi after the death of the latter’s wife. Chapter 19 is perhaps the most famous of this grouping as it contains a series of ‘skill’ or ‘knack’ passages that feature heroes who can be seen as masters of the flowing mode of cognition emphasized in the ‘Inner chapters.’ These include the cicada-catching hunchback, the swimmer at Spinebridge Falls, and the bellstand carver who fasts for seven days before undertaking his task, thus recalling the mind-fasting advice Confucius gave to Yan Hui in chapter 4. Chapter 20 contains a group of narratives loosely organized around the theme of uselessness first presented in chapters 1 and 4. Only things that are not of use to anyone else are able to flourish and attain their full potential. Chapter 21 is filled with stories featuring exemplars of self-cultivation who have achieved the utmost inner power. The famous ‘knowledge wanders north’ narrative that begins chapter 22 contains insights on the limitations of the fixed mode of cognition to comprehend the Way. Filled with ideas from *Daode jing* and with references to breath meditation, it also contains the famous dialogue in which Zhuangzi details where the Way can be found.

Unlike the ‘Inner chapters’ that contain no references to Lao Tzu the man and to the text of the *Daode jing*, many of these chapters show an awareness of the *Daode jing* by their use of ideas and quotations from this text. This indicates that they were most likely written after this work began widely circulating in China after in about 260 B.C. E. To the extent that they recast material from the ‘Inner chapters’ in new narrative frameworks and frequently see it in light of ideas from the *Daode jing*, these chapters represent a unique blending of the two intellectually foundational sources of early Daoism.

The first group of the ‘Miscellaneous chapters,’ 23-27, and chapter 32 are much more heterogeneous in their content. They appear to contain more writings of the followers of Zhuangzi into which are interspersed passages from the other major authorial voices in the complete work, mostly the *Zhuangzi* of the ‘Inner chapters,’ the Primitivist, and, on occasion, the Syncretist. Given this lack of coherence, these ‘Miscellaneous chapters’ could contain material from some of the nineteen chapters that Guo Xiang deleted from the original recension of the text. In these chapters Zhuangzi’s followers continue their

engagement with their master's teachings from the 'Inner chapters' and attempt to integrate it with the teachings of the *Daode jing* now often attributed in narratives to Lao Dan, the shadowy fifth-century B.C.E. figure to whom this text began to be attributed after about 250 B.C.E. Perhaps the most interesting narrative in this grouping is the one that constitutes almost the entirety of chapter 23. In it the character Nanguo Chu goes on a quest for mystical knowledge and ends up being instructed by Lao Dan in a meditative practice that blends together ideas from the 'Inner chapters,' the *Daode jing*, and other sources of 'inner cultivation' such as *Guanzi*'s 'Inward Training' (*Neiye*) text (Roth 1999). This narrative, as well as several others in this group of chapters from the disciples of Zhuangzi, indicates that such meditation practices continued to be as central to the followers of Zhuangzi as they were to their teacher himself.

3.2 The Yangist Chapters

Chapters 28-31 of the received recension of the *Zhuangzi* were the first to be perceived as so different from the philosophy of the renowned 'Inner Chapters' that they were thought to be the work of an entirely different intellectual lineage. Indeed, these chapters are now seen to be similar in thought to five essays from the first two chapters of the compendium *Lüshi chunqiu* (240 B.C.E.) that consitute the only surviving works of the long-lost tradition of the philosopher Yang Zhu. Graham regards these *Zhuangzi* chapters themselves as Yangist while Liu Xiaogan links them to the 'Primitivist' material. Close examination reveals many common philosophical themes between these two groups of chapters but also reveals some key differences as well, as we shall see.

Yang Zhu was a fourth century B.C.E. contemporary of Mencius who engendered great antipathy in this Confucian thinker for suggesting that the basic tendencies of human nature were not what we might call 'other-regarding.' Mencius condemned Yang for being so egotistical as to be unwilling to sacrifice even a single hair in order to benefit the state (Mencius 7A26; Lau, p. 275). However if we are to base our understanding of Yang's doctrines on these two surviving sources, a much more complex and interesting picture of his philosophy emerges.

Yang Zhu may have been the first Chinese philosopher to speak of the concept of human nature (*xing*), and the parameters for all early Chinese discussions of this concept seem to have been established by Yang and Mencius. In brief, human nature is given to us by Heaven, the power responsible for everything in life beyond human control. The early Chinese conceived of two major aspects of our lives that fall into this category: *ming* (fate, destiny), the various things that occur as the result of agencies other than ourselves and *xing* (nature), the sum total of our genetic inheritance both as a species and as unique individual members of it. According to Graham and Ames, human nature in early China is conceived as totally dynamic, in contrast to the implicit static basis of human nature we find in the West (Graham 1967, Ames 1991). The Chinese concept of human nature can be best understood as referring to the spontaneous tendencies that an individual has from birth that govern its development as a particular individual within a species and which also act as forces in its daily life. Thus this concept implies both the potential to develop in a certain way and the spontaneous tendencies for this development and for certain characteristic types of activities. We might call the former tendencies 'genetic' and the latter

tendencies ‘instinctive.’ Mencius argued that the essential goodness of human nature rested in the spontaneous tendencies to act selflessly and respectfully, tendencies that persist throughout the lifetime of an individual even if left undeveloped. In other words, it is a basic human instinct to act selflessly. For him the purpose of self-cultivation was to nurture these spontaneous instinctive tendencies until they blossomed into complete ethical virtues. The Yangist challenge to the social emphasis of the Confucians consisted in the primacy they placed on the maintenance of the individual life and the fact that they supported this mode of living with the theory that to act in this fashion was to nourish the nature that we receive from Heaven. Since Confucians placed a high value on the sanctions and approvals of Heaven, it was incumbent upon them to argue for a different vision of human nature.

The single most basic of the spontaneous tendencies of human nature for the Yangists is longevity. They argued that human beings tend to live long if they keep themselves from being disturbed by the ‘external things’ of this world such as fame and profit. The second important aspect of human nature is the desire of the five sense faculties (eye, ear, nose, tongue, skin) for sense-objects. It is the senses’ desire for their objects that in a fundamental way helps to maintain the health and the development of the organism, thus enabling it to realize its inherent tendency for longevity. However the senses themselves need to be regulated and limited to only the ‘suitable’ amount of stimulation. Over-stimulation causes the senses to be impaired and eventually damaged. Thus there is a suitable amount of stimulation that is conducive to the health and development of the human organism and that suitable amount must be determined by Sages; the senses on their own do not have the ability to do this. Self-cultivation for the Yangists therefore consists of nourishing one’s inherent nature by strictly limiting sense stimulation to the appropriate degree needed to maintain health and vitality. One of their principal practices was to prevent the loss of one’s finite supply of *jing* (essential vital energy), which is lost due to over-stimulation of the senses. The Yangists shared an understanding of how the human organism functioned with the thinkers of the ‘inner cultivation’ tradition and with early Chinese medical philosophers and practitioners who envisioned a body-mind complex made up of various systems of *qi* (vital energy). It is also worth noting that, in keeping with their dynamic concept of human nature, the Yangists included the senses’ desire for sense objects in it and not the senses themselves, which would imply a static basis.

Implicit in the Yangist authors’ inclusion of longevity within human nature is the understanding that the various systems of vital energy that constitute a living organism tend to function harmoniously if left unimpaired. Nurturing the nature by limiting the senses to their appropriate degree of stimulation and avoiding activities that would damage the organism involve assisting in this inherent tendency for harmony. Human beings, as well as all things in the world, cohere and function if undamaged; they do not fall apart. The individual microcosm, just as the universal macrocosm, is not random and chaotic. It functions according to certain basic laws and patterns. To understand them, and to live according to them, to understand the spontaneous tendencies of human nature and to nurture them by conscious choice, is the basis of the Yangist method of self-cultivation.

Given their concept of human nature and their resultant ideas about self-cultivation, and their emphasis on avoiding placing oneself in jeopardy for fame and profit, it is not surprising that the Yangists do not proffer an elaborate social and political philosophy. "The most genuine in the Way is for supporting one’s own person ..." reads *Zhuangzi* 28, (‘Yielding the Throne’), "its left-overs are for running a state,

its discards are for ruling the empire. Seen from this viewpoint, the achievements of emperors and kings are the left-over deeds of the sage, they are not the means by which he keeps his person whole and nurtures life." (Graham, p. 227) In chapter 29 of *Zhuangzi* ('Robber Zhi') we find the dictum: 'If you can't look after yourself, you can't look after others.' (Graham, p. 238) In the 'Valuing Life' essay of the *Lüshi chunqiu*, we read, "only those who would not impair their natures are able to be entrusted with ruling the empire" (Riegel/Knoblock p. 80, mod.). A very similar idea is expressed in chapter 13 of the *Daode jing*: "Hence he who values his body more than dominion over the empire can be entrusted with the empire." (Lau 1982, p. 19) This seems to imply that such a ruler would treat others as carefully as he treats himself, although the ethical implications of putting oneself first for government are never worked out in the surviving Yangist documents.

What we do find in the Yangist-oriented chapters of the *Zhuangzi* are many stories in which the concern for not impairing one's nature leads people to either resign the throne, or to never accept public office. The life of the recluse is commended, but the authors are sharply critical of those moralists who would rather kill themselves than participate in government. The Yangist political philosophy is clear: do not seek after fame, wealth, and power, all of which are far beyond your essential needs. Never do anything to impair your inherent tendency to live a long and fulfilling life. To know this is to differentiate the important from the unimportant, to understand, in the words of the 'Giving Weight to the Self' essay in Book 1 of the *Lüshi chunqiu,* 'the essentials of our nature and destiny' (*xingming zhi qing*) Riegel/Knoblock, pp. 67-68 mod.) Only those who can do this are truly fit to govern.

Structurally, the four Yangist-oriented chapters of the *Zhuangzi* are each unique. Chapter 28 is a collection of eighteen narratives, ten of which are found in slightly varying forms dispersed throughout the *Lüshi chunqiu*. Chapter 29 contains only three narratives, including the famous long one in which Robber Zhi berates Confucius, accusing him of practicing a way that 'is a crazy obsession, a thing of deception, trickery, vanity, and falsehood [that] will not serve to keep the genuine in us intact...'(Graham, 239). Chapter 30 consists of one short narrative in which Zhuangzi demonstrates the inferiority of sword fighting. Chapter 31 contains a dialogue between Confucius and a hermit known as the 'Old Fisherman,' who teaches him that the art of self-cultivation lies in 'guarding the genuine within you.' He defines 'the genuine' as the spontaneous expression of emotions. This concept of 'guarding the genuine' resonates most closely with the central theme of the rest of the text, that of cultivating the flowing mode of cognition. However, there are virtually no mystical elements in these Yangist chapters and this both distinguishes them from the remainder of the *Zhuangzi* and prevents them from being classified as 'Daoist.' Yet in their discussion of human nature, their attacks on the Confucians and their praise of ancient primitive Utopias, they so much echo ideas from the Primitivist chapters that scholars such as Liu Xiaogan conclude they belong to the same intellectual tradition.

3.3 The Primitivist Chapters

There are three chapters (8-10) and half of a fourth (11) that present such a consistent literary structure, technical terminology, and written style that Graham maintains them to be the work of a single author, one who espouses a viewpoint similar to that found in the *Daode jing* differing principally in that it is not

addressed to the ruler. To these Roth would add chapter 16, ‘Menders of Nature,’ when modified by the removal of a short Syncretist gloss that has previously prevented scholars from linking it with the others (Roth 2002). Because of their advocacy of a return to a government and social organization similar to that found in primitive tribal Utopias, Graham has labeled these chapters as ‘Primitivist.’ While similar in their Utopian vision, critique of the Confucians, and their focus on a theory of human nature to the Yangist-oriented chapters, they depart from them in containing key elements of mystical philosophy such as a cosmology of the Way and Inner power and reference to methods of ‘inner cultivation.’ In addition, rather than totally eschewing political philosophy, they advocate a government by Non-action similar to that found in the *Daode jing*. These chapters are important because they contain the first discussion of human nature in the entire Daoist tradition and they seem to define it by building on the Yangist conceptions of it.

Like the Yangists, the Primitivists commend, and, in ‘Menders of Nature,’ even defend the life of the recluse who withdraws from political involvement. Both label their Utopias ages ‘when Inner Power was at its utmost,’ and Robber Zhi, the Yangist critic of Confucius in chapter 29 is a somewhat sympathetic, if ironic, figure in the chapter 10, ‘Rifling Trunks.’ Like the Yangists, the Primitivists recognize the two complementary aspects of human nature, both the ‘genetic’ tendencies that determine the characteristic course of development of members of a given species and the spontaneous ‘instinctive’ tendencies that arise within the daily activity of individual members of said species. Both agree that this nature is not the actual physical attributes (such as horses hooves for the Primitivist, or the sense organs for both), but rather the tendencies for them to develop in a species-characteristic way and to function spontaneously without additional help. Both believe that these tendencies can be nurtured or can be interfered with, and both maintain that to avoid interfering with them is to understand the ‘essentials of our nature and destiny,’ which is mentioned once in the Yangist-oriented chapters and eight times in the Primitivist chapters. If we take the word ‘destiny’ to refer to one’s allotted years, we might argue that the Primitivists as well, at least tacitly accept longevity as an attribute of human nature. To this extent there is a Yangist basis for the Taoist Primitivists. However there are also substantial differences.

As we have seen, the Yangist inclusion of longevity in human nature contains an implicit understanding that the physiological systems that constitute the individual organism tend to function spontaneously and harmoniously. This understanding is made explicit by the Primitivists; indeed, the inherent spontaneous activity of the individual becomes one of the cornerstones of their unique concept of human nature. It is so important that it replaces the Yangist longevity as the overall goal of self-cultivation and the raison d’etre for human existence. In contrast to the Yangists, the Primitivists maintain that the natural tendency of the senses is not merely to desire sense-objects, but, rather, to perceive them clearly. When the senses are allowed to operate unimpeded by the culturally established normative categories (the Five Tastes, the Five Colors, the Five Tones, and so on) and they hence operate spontaneously, they will clearly grasp their objects. These cultural norms insert an element of bias and inclination into the perceptual process thereby leading perception astray. In an analogous manner, Confucian ethical norms of humaneness and rightness, inject ‘inclinations and aversions’ into the spontaneous functioning of the mind and disrupt its inherent tendency to operate in an unbiased and harmonious fashion. As the author of ‘Menders of Nature’ succinctly states: "If someone else lays down the direction for you you blinker your Inner Power." (Graham, p. 171, mod.). For the Primitivists, Inner Power is simply allowing the spontaneous

tendencies of one’s nature to operate without interference. It is our nature for them to do so; culture, with the self-consciousness it forces on human beings, disrupts and damages it. As we read in chapter 8, ‘Webbed Toes,’:

> To depend on the carpenter’s curve and line, compasses, L-square, to straighten you out, is to pare away your nature; to depend on cords, knots, glue, lacquer, to hold you together, is to violate your Inner Power; and to bow and crouch for Rituals and Music, and smirk and simper over Humaneness and Rightness, in order to soothe the hearts of the world, is to lose the constant in you. There is such as thing as constancy in the world.... Thus everything in the world springs into life without knowing how it is born, attains unthinkingly without knowing how it attains. Hence the present is no different from the past, and nothing can be missing from its place. (Graham, p. 201, mod.)

That which is the constant in the world for the Primitivists, is its tendency for spontaneous and harmonious activity. In human beings this is inherent in their very natures. The author of ‘Menders of Nature,’ in harkening back to an earlier time when people "lived in the midst of the merged and featureless, and found tranquility and mildness with those of their own time..." says, "At this era no one took any (deliberate) action and things were constantly so of themselves." (Graham, p. 171, mod.) This reiterates two important ideas in the *Daode jing*, Non-action (*wu-wei*) and ‘so-ness’ or spontaneity (*ziran*). The former refers to acting without egotism and self-consciousness. The latter refers to allowing things to be just as they naturally are. This theme of the constancy in human nature is central to the Primitivist vision of an ideal society. It is fully described in chapter 9, ‘Horses Hooves:’

> The people have a nature that is constant:
> By their weaving clothed, by their ploughing fed-
> This is called ‘sharing Inner Power’.
> In oneness and without faction:
> The name for it is ‘free as the air’.
>
> Therefore, in the age when Inner Power was at its utmost....
> people lived in sameness with birds and beasts, side by side as fellow clansmen with the myriad creatures; how would they know a gentleman from a knave?
>
> In sameness, knowing nothing!
> Not parted from their Inner Power.
> In sameness, desiring nothing!-
> Call it "the simple and unhewn."
>
> In the simple and unhewn the nature of the people is found. (Graham pp. 204-5, mod.)

The ‘simple’ (*so*) and the ‘unhewn’ (*pu*) are important ideas in the *Daode jing* wherein to be ‘simple’ mean to be unselfish (DDJ 19) and to be ‘unhewn’ means to be without desires (DDJ 19, 38) The

definitions in 'Horses Hooves' are similar, the only difference being that to be simple means 'knowing nothing.' Both words suggest a state of mind totally devoid of self-consciousness, a state of mind in which people act spontaneously and without self-reflection. It is a state of mind that is reminiscent of the 'flowing cognition' of the 'Inner chapters.' For the Primitivists, it is human nature to attain this state of mind when people are left on their own, when the institutions of culture do not interfere with their spontaneous tendencies. To attain this state is to realize your Inner Power. Throughout their writings the Primitivists harken back to an earlier Utopian age when people lived in selfless harmony with one another and with all things in the world and when the Way and Inner Power were fully realized. The Confucian sage-rulers, who established cultural norms and thereby forced people to think about how to attain them, destroyed this harmony and made it much more difficult for people to attain the 'simple and unhewn' state of mind. However by doing away with the sages and their cultural norms we can return to a primitive Utopia: 'Utterly demolish the laws of the sages throughout the world, and for the first time it will be possible to sort out and discuss things with people...cast away Humaneness and Rightness, and at last Inner Power throughout the world will be the same from its profoundest depths.' (Graham, 208-9, mod.) Then society can be governed by a ruler who learns how to practice Non-action. In chapter 11, 'To Locate and Circumscribe,' we read:

> So if the gentleman is left with no choice but to preside over the world, his best policy is Non-action. Only by Non-action will he find security in the essentials of his nature and destiny. So if you value regard for you own person more than governing the world, you are fit to be entrusted with the world; if you love the care of your own person more than governing the world, you deserve to have the world delivered to you. If then a gentleman does prove able not to dislocate his Five Spheres [of vital energy] and not to stretch his eyesight and hearing, then sitting as still as a corpse he will look majestic as a dragon, from the silence of the abyss he will speak with a voice of thunder, he will move like a spirit and veer like Heaven, he will be relaxed and take no action.... (Graham, p. 212, mod.)

So if forced to rule the Primitivist author here follows the dictum of the *Daode jing* to take no action, that is, to act effortlessly, not interfering with your spontaneous response. By so acting you can 'locate and circumscribe' your nature by identifying and nurturing its spontaneous tendencies and preventing it from being led astray by Confucian cultural norms. The Primitivist here echoes the Yangists by arguing that to do this rulers must place their own self-cultivation first. However unlike the Yangists who maintain that personal longevity is the goal and purpose of human existence, the Primitivists maintains that returning to the 'simple and unhewn' unselfconscious mode of experience is this goal. In order to accomplish this rulers must practice inner cultivation and through this attain the profound tranquility and silence needed to find 'security in the essentials of nature and destiny.'

Thus the Primitivist authors in the *Zhuangzi* extend the definition of human nature past that of the Yangists by arguing that it consists not simply of the desire of the senses for sense-objects and the tendency for the biological systems these senses support to function harmoniously but of the tendency of these senses to spontaneously function clearly and for people to attain the 'simple and unhewn' unself-conscious state of mind. In other words, it is our most basic tendency to be able to experience the flowing

mode of cognition so often advocated in the 'Inner chapters.' It is this that links these chapters most strongly with the remainder of the *Zhuangzi*.

3.4 The Syncretist Chapters

The final stratum of the *Zhuangzi* contains a distinctive and largely consistent viewpoint that connects with the rest of the text and with a larger philosophical context. It is contained in three complete essays: 1. the first two-thirds of chapter 13, 'The Way of Heaven;' 2. chapter 15 'Inveterate Ideas,' and 3.the final chapter, 33, 'Below in the Empire,' as well as in narratives that play key roles in chapters 12 and 14. This material shares a common cosmology of the Way and interest in inner cultivation that we have seen in most of the rest of the text but veers in a different direction in its political thought, advocating a hierarchical social and political structure that incorporates the best ideas of other earlier intellectual lineages within a Daoist cosmological framework. However it agrees with the Primitivist idea that government should be led by a sage enlightened through inner cultivation techniques. In its general intellectual viewpoint it exemplifies many of the characteristics of the Daoist tradition that were first enunciated by the Han dynasty historian, Sima Tan (d. 110 B.C.E.), the man who coined the very term 'Daoism' (*dao-jia*):

> The Daoists enables the essential vital energy and spirit of human beings to be concentrated and unified. They move in unison with the Formless and provide adequately for all living things. In deriving their techniques, they follow the grand compliances [between humans and the cosmos] of the Naturalists, select the best of the Confucians and Mohists, and extract the essentials of the Terminologists and Legalists...[Daoists] take no action but also say that nothing is left undone. Their substance is easy to practice but their words are difficult to understand. Their techniques take emptiness and nothingness as the foundation, adaptation and compliance [with cosmic patterns] as the application....[They show how] the ruler [can] unite with the Great Dao, obscure and mysterious, and after illuminating the whole world revert to the Nameless. (Queen and Roth, pp. 279-82, mod.)

Thus, according to Sima Tan, Daoists advocate these essential ideas: 1. Humans can cultivate themselves to attain harmony of body and mind and to realize their essential connection to the Way and to the entire cosmos; 2. When rulers become adept at such 'inner cultivation' they can govern dispassionately and humanely according to the greater patterns of heaven and earth, upon which they model their social and political institutions; 3. While remaining faithful to this general Daoist orientation rulers should make use of the best ideas of other early intellectual lineages; 4. With these institutions and practices established, rulers can govern by taking no action while leaving nothing undone.

All these ideas are found in the Syncretist *Zhuangzi*. Inner cultivation practice is advocated to attain a deep and tranquil state of mind to enable both sages and rulers to act efficaciously in the world. In 'Inveterate Ideas' we read:

> Hence it is said that calm repose, silent stillness, empty nothingness, and Non-action: these

> are the even level of heaven and earth, the substance of the Way and its Inner Power; therefore sages find rest in them. At rest they are unperturbed and relaxed; being unperturbed and relaxed, they are calm and reposed. If they are unperturbed and relaxed, calm and reposed,
>
> Cares and misfortunes cannot enter
> Deviant vital energy cannot make inroads.
>
> Therefore their inner power is intact and their spirit is unimpaired. Hence it is said that sages:
>
> In life proceed with heaven,
> In death transform with other things.
> In stillness share the inner power of the Yin,
> In motion share the surge of the Yang.
> They won't make the first move to gain advantages,
> They won't take the first step to avoid trouble.
> Only when stirred will they respond,
> Only when pressed will they move.
> Only when it is inevitable will they arise,
> Rejecting knowledge and precedent
> They take their course from the patterns of heaven.
>
> (Graham, 265, mod.)

So sages attain profound levels of tranquility and through them they get in touch with the Way and preserve its inner power within them. They then make use of this experience to act spontaneously and harmoniously while being guided by the greater patterns of the cosmos. In other words, this is the Syncretists' version of attaining the flowing mode of cognition advocated in the other parts of the text. As if embracing the other viewpoints in the *Zhuangzi*, the Syncretist author of chapter 13, 'The Way of Heaven,' argues that this flowing mode can be applied to a variety of life circumstances:

> Empty tranquility, calm repose, silent stillness, and Non-action are the foundation of all living things. To be clear about these when you sit facing south is to be the kind of ruler [the legendary sage-king] Yao was. To be clear about these when you sit facing south is to be the kind of minister that [the legendary sage-minister] Shun was. To have these as your resources in a high position is the inner power which is in emperor, king, Son of Heaven. To have these as your resources in a low position is the Way of the obscure sage, the untitled king. Use these to settle in retirement or wander at leisure, and the hermits of river, sea, mountain, and forest will submit to you. Use these to come forward and act in order to bring comfort to the age and your achievement is great and name illustrious, and the empire is united. In stillness a sage, in motion a king, you do nothing yet are exalted, you

> are simple and unhewn yet no one in the empire is able to rival your glory. (Graham 259-60, mod.)

Hence, ruler and minister, politician and hermit can utilize flowing cognition; but it is perhaps when applied to rulership that it attains its greatest achievement. Government by enlightened sages who attain flowing cognition is the pinnacle of Daoist political thought for the Syncretist: it is symbolized above by the phrase ‘in stillness a sage, in motion a king,’ which is elsewhere in the final chapter 33 referred to as ‘to be inwardly a sage and outwardly a king.’ It is with this cultivated mind that the sage ruler establishes human society in parallel to the greater patterns of the cosmos. For example, human social and political hierarchies are based upon normative natural patterns:

> The ruler comes first, the minister follows; the father comes first, the son follows ... the senior comes first, the junior follows ... . Being exalted or lowly, first or last, belongs to the progressions of heaven and earth; therefore sages take their model from them ... spring and summer first, autumn and winter last, is the sequence of the four seasons ... . Heaven and earth are extremely numinous yet have sequences of the exalted and lowly, the first and the last, and how much more should this be true of the Way of human beings! If you expound a Way without their sequences, it is not their Way; and if you expound a Way which is not their Way, from what will you derive a Way? (Graham, 261, mod.)

This coordination of human society with cosmic patterns is a characteristic tenet of the early Daoist syncretic lineage that historians first called the ‘Way of the Yellow Emperor and Laozi’ (*Huang-Lao zhi dao*). This concern with cosmic patterns and their human parallels is incorporated from the Naturalist thinkers who first systematized the concepts of Yin and Yang (complementary negative and positive aspects of all living things). The Syncretist author of *Zhuangzi* 13 demonstrates further evidence of this characteristic adaptation of ideas from other early philosophical lineages:

> ...Therefore, the ancients who made clear the great Way first made heaven clear and the Way and its Inner Power were next. When the Way and its Inner Power were clear, [Confucian concepts of] humaneness and rightness were next. When humaneness and rightness were clear [Legalist concepts of] portions and responsibilities were next. When portions and responsibilities were clear, [Terminologist concepts of] title and performance were next ... when [Terminologist concepts of] judging right and wrong were clear, [Legalists concepts of] reward and punishment were next. ... This is how one served the ruler above or was pastor to the people below, put other things in order or cultivated one’s own person. Cleverness and strategy were not used and they invariably referred back to what was from heaven in them. It is this that is meant by supreme tranquillity, the utmost in government. (Graham 261-2, mod.)

This is an excellent example of the kind of Daoist syncretism Sima Tan describes and that survives in other early texts such as the *Huang-Lao Silk Manuscripts* found in 1973 at Ma-wang-tui (ca. 250 B.C.), parts of the philosophical compendia *Guanzi* (ca. 320-150 B.C.) and *Lüshi chunqiu* (240 B.C.), and the

Huai-nan Tzu (139 B.C.). All these works are the products of a loosely organized intellectual tradition of masters and disciples that carried Daoist ideas into the unified empire of second century B.C. China and that transmitted what we know of this philosophy into the modern world.

The final chapter of the *Zhuangzi*, 'Below in the Empire,' exemplifies this kind of syncretism in its analysis of early intellectual traditions. After establishing its own position, the comprehensive 'Way of Heaven and Earth' that embraces all the 'techniques of the Way,' it analyzes how each of these earlier traditions understood one part of this comprehensive Way but ultimately failed to grasp the whole. Zhuangzi himself is included in this analysis:

> ...Above he roamed with the maker of things; below, he made friends with those for whom life and death are externals and there is neither end nor beginning. As for the Foundation, he opened it up in all its comprehensiveness, ran riot in the vastness of its depths. As for the Ancestor, it may be said that by being in tune he withdrew all the way back to it. However, when one assents to transformation and is released from things, the body has not exhausted its pattern, having come it will not be shaken off. Abstruse! Obscure! A man who did not succeed in getting it all. (Graham 283, mod.)

The Syncretist author thus praises Zhuangzi for his depth of mystical cultivation but chides him for failing to realize that there are practical affairs in the world that must be attended to. It is an interesting yet telling comment. The Syncretist criticizes the very impractibility for which Zhuangzi became renowned to later literati. The irony is that without the more prosaic and responsible Syncretists, who transmitted earlier parts of the text that contained the authentic voice of Zhuangzi, the iconoclastic ideas of this philosophical giant may have never been known. This is because of the extensive support and patronage of early philosophy by the kings of the various 'warring states' starting in the middle of the fourth century B.C. and continuing into imperial times. These rulers were interested in learning how to govern more efficaciously. Philosophical works were valued for their teachings about rulership and virtually all the texts that were widely circulated during this period had some political relevance, including, of course, the *Daode jing*. The 'Inner Chapters' of *Zhuangzi* had none, but were preserved, as we have seen, in a larger collection that did contain important ideas for rulership.

So, in the last analysis, what is it that holds together this disparate collection of materials that were written over a span of more than a century? All the various authorial voices in the text (except possibly the Yangist) share a common interest in cultivating the spontaneous 'flowing mode' of cognition that is symbolized throughout the work by the 'heavenly' side of human beings. They vary somewhat in how they conceive of this mode and they apply it in different ways but nonetheless they concur in emphasizing its importance. It is possible that all the authors were part of an intellectual tradition that descended from the historical founder, Zhuangzi, by direct master-disciple lines and each authorial voice constitutes a generation of such descent. It is also possible that the *Zhuangzi* collection was compiled by later Daoists, perhaps those at the court of Liu An, the second king of Huai-nan, who were attempting to reconstruct a lost Zhuangzi lineage (Roth 1991). We cannot know for certain. The precise way this text formed has, in all likelihood, disappeared into the mists of time, but we can rejoice in the fact that because of its politically relevant final stratum, the entire text has not met the same fate.

Bibliography

- Ames, Roger T. (1991), ‘The Mencian Concept of *Ren Xing*: Does it Mean Human Nature?’ in *Chinese Texts and Philosophical Contexts*, ed. Henry Rosemont, Jr. LaSalle, Ill.: Open Court Press.
- Ames, Roger T. (1998) ed. *Wandering at Ease in the Zhuangzi*. Albany: State University of New York Press.
- Graham, Angus Charles, (1967), ‘The Background of the Mencian Theory of Human Nature,’ *Tsing Hua Journal of Chinese Studies* 6:1,2. Repr. *Studies in Chinese Philosophy and Philosophical Literature,* ed. Graham. Albany: State University of New York Press, 1990.
- Graham, Angus Charles (1979), ‘How Much of *Chuang Tzu* Did Chuang Tzu Write?’ in *Studies in Early Chinese Thought* ed. Henry Rosemont Jr. And Benjamin Schwartz. Repr. in *Studies in Chinese Philosophy and Philosophical Literature*, ed. Graham. Albany: State University of New York Press, 1990.
- Graham, Angus Charles (1981), *Chuang Tzu: The Inner Chapters*. London: Allen and Unwin. Repr. Boston: Hackett, 2000.
- Graham, Angus Charles (1989), *The Disputers of the Tao*. LaSalle, Ill: Open Court.
- Hansen, Chad (1983), ‘A Tao of Tao in *Chuang Tzu,*’ in *Experimental Essays on Chuang Tzu*. Ed Victor Mair. Honolulu: University of Hawaii Press.
- Ivanhoe, Philip J. (1993) ‘*Zhuangzi* on Skepticism, Skill, and the Ineffable Tao,’ *Journal of the American Academy of Religion* 64.4: 639-54.
- Kjellberg, Paul and Philip J. Ivanhoe eds. (1996), *Essays on Skepticism, Relativism, and Ethics in the Zhuangzi*. Albany: State University of New York Press.
- Knoblock, John and Jeffrey Riegel (2000), *The Annals of Lü Buwei: A Complete Study and Translation*. Stanford: Stanford University Press.
- Kohn, Livia (1987), *Seven Steps to the Dao: Sima Chengzhen's Zuowang lun*. Monumenta Serica Monograph XX. Nettetal: Steyler Verlag.
- Lau, D.C. (1982), *Chinese Classics Tao Tao Ching.* Hong Kong: Chinese University Press.
- Liu Xiaogan (1994), *Classifying the Zhuangzi Chapters*. Ann Arbor: University of Michigan Center for Chinese Studies.
- Mair, Victor (1983), *Experimental Essays on the Chuang Tzu*. Honolulu: University of Hawaii Press.
- Mair, Victor (1998), *Wandering on the Way: Early Taoist Tales and Parables of Chuang Tzu*. Honolulu: University of Hawaii Press.
- Peerenboom, R.P. (1993), *Law and Morality in Ancient China: The Silk Manuscripts of Huang-Lao*. Albany: State University of New York Press.
- Queen, Sarah, Nathan Sivin, and Harold Roth (1999), ‘Syncretic Visions of State, Society, and Cosmos,’ in *Sources of Chinese Tradition*. 2nd ed. Compiled by Wm. Theodore deBary and Irene Bloom. New York: Columbia University Press.
- Raphals, Lisa (1996), ‘Skeptical Strategies in *Zhuangzi* and Theaetetus,’ in)*Essays on Skepticism, Relativism, and Ethics in the Zhuangzi*. Albany: State University of New York Press.

- Rickett, W. Allyn (1985, 1998), *Guanzi: Political, Economic, and Philosophical Essays from Early China*. 2 vols. Princeton: Princeton University Press.
- Roth, Harold D. (1991)¸’Who Compiled The *Chuang Tzu*?’ in *Chinese Texts and Philosophical Contexts*, ed. Henry Rosemont, Jr. LaSalle, Ill.: Open Court Press.
- Roth, Harold D. (1999), *Original Tao: Inward Training (Nei-yeh) and the Foundations of Taoist Mysticism*. New York, Columbia University Press.
- Roth, Harold D. (2000), ‘Bimodal Mystical Experience in the ‘Qiwu lun’ Chapter of *Zhuangzi*,’ *Journal of Chinese Religions*, 28: 31-50.
- Roth, Harold D. (2002), ‘Graham’s Scholarship on the *Chuang Tzu*: A Reassessment,’ in *A Companion to Angus C. Graham’s Chuang Tzu: the Inner Chapters*, ed. Harold D. Roth. Society for Asian and Comparative Philosophy Monograph. Honolulu: University of Hawaii Press.
- Stace, Walter (1960), *Mysticism and Philosophy*. London: MacMillan Press.
- Watson, Burton (1968), *The Complete Works of Chuang Tzu*. New York: Columbia University Press.
- Yates, Robin D.S. (1997), *Five Lost Classics: Tao, Huang-Lao, and Yin-Yang in Han China*. New York: Ballantine Books.
- Yearley, Lee (1983), ‘The Perfected Person in the Radical *Chuang Tzu*,’ in *Experimental Essays on the Chuang Tzu*, ed. Victor Mair. Honolulu: University of Hawaii Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Laozi

Harold Roth
Brown University
Harold_Roth@brown.edu

First published: November 9, 2001
Content last modified: November 9, 2001

Mally's Deontic Logic

In 1926, Mally presented the first formal system of deontic logic. His system had several consequences which Mally regarded as surprising but defensible. It also had a consequence ("A is obligatory if and only if A is the case") which Menger (1939) and almost all later deontic logicians have regarded as unacceptable. We will not only describe Mally's system but also discuss how it may be repaired.

1. Introduction

In 1926, the Austrian philosopher Ernst Mally (1879-1944) proposed the first formal system of deontic logic. In the book in which he presented this system, *The Basic Laws of Ought: Elements of the Logic of Willing*, Mally gave the following motivation for his enterprise:

> In 1919, everybody was using the word self-determination. I wanted to obtain a clear understanding of this word. But then, of course, I immediately stumbled on the difficulties and obscurities surrounding the concept of Ought, and the problem changed. The concept of Ought is the basic concept of the whole of ethics. It can only serve as a usable foundation for ethics when it is captured in a system of axioms. In the following I will

present such an axiomatic system.[1]

As Mally's words indicate, he was not primarily interested in deontic logic for its own sake: he mainly wanted to lay the foundation of "an exact system of pure ethics" (*eine exakte reine Ethik*). More than half of his book is devoted to the development of this exact system of ethics. In the following, we will, however, concentrate on the formal part of his book, both because it is its "hard core" and because it is the part that has attracted the most interest.

2. Mally's Formal Language

Mally based his formal system on the classical propositional calculus as formulated in Whitehead's and Russell's *Principia Mathematica* (vol. 1, 1910).

The non-deontic part of Mally's system had the following vocabulary: the sentential letters A, B, C, P and Q (these symbols refer to states of affairs), the sentential variables M and N, the sentential constants V (the *Verum*, Truth) and Λ (the *Falsum*, Falsity), the propositional quantifiers $\exists$ and $\forall$, and the connectives $\neg$, & , $\vee$, $\rightarrow$ and $\leftrightarrow$. Λ is defined by $\Lambda = \neg V$.

The deontic part of Mally's vocabulary consisted of the unary connective !, the binary connectives f and ∞, and the sentential constants U and $\cap$.

- Mally read !A as "A ought to be the case" (*A soll sein*) or as "let A be the case" (*es sei A*).
- He read A f B as "A requires B" (*A fordert B*).
- He read A ∞ B as "A and B require each other."
- He read U as "the unconditionally obligatory" (*das unbedingt Geforderte*).
- He read $\cap$ as "the unconditionally forbidden" (*das unbedingt Verbotene*).

f, ∞ and $\cap$ are defined by:

Def. f . A f B = A $\rightarrow$!B (Mally 1926, p. 12)

Def. ∞ . A ∞ B = (A f B) & (B f A)

Def. $\cap$. $\cap = \neg U$

Mally did not only read !A as "it ought to be the case that A." Because a person's willing that a given state of affairs A be the case is often expressed by sentences of the form "A ought to be the case" (for example, someone might say "it ought to be the case that I am rich and famous" to indicate that she wants to be rich and famous), he also read !A as "A is desirable" or "I want it to be the case that A." As a result, his formal system was as much a theory about *Wollen* (willing) as a theory about *Sollen* (ought to be the case). This explains the subtitle of his book. In modern deontic logic, the basic deontic connective O is seldom read in this way.

We have just described one respect in which Mally's deontic logic was different from more modern proposals. There are two other conspicuous differences:

- Mally was only interested in the deontic status of states of affairs; he paid no special attention to the deontic status of actions. Thus, his *Deontik* was a theory about *Seinsollen* (what ought to be the case) rather than *Tunsollen* (what ought to be done). Modern authors often regard the concept of *Tunsollen* as fundamental.
- In modern deontic logic, the notions of prohibition F, permission P and waiver W are usually defined in terms of obligation O: $FA = O\neg A$, $PA = \neg FA$, $WA = \neg OA$. Such definitions are not to be found in Mally's book.

3. Mally's Axioms

Mally adopted the following informal deontic principles (Mally 1926, pp. 15-19):

(i) If A requires B and if B then C, then A requires C.

(ii) If A requires B and if A requires C, then A requires B and C.

(iii) A requires B if and only if it is obligatory that if A then B.

(iv) The unconditionally obligatory is obligatory.

(v) The unconditionally obligatory does not require its own negation.

Mally did not offer much support for these principles. They simply seemed intuitively plausible to him.

Mally formalized his principles as follows (Mally 1926, pp. 15-19):

I. $((A \mathrm{f} B) \& (B \rightarrow C)) \rightarrow (A \mathrm{f} C)$

II. $((A \mathrm{f} B) \& (A \mathrm{f} C)) \rightarrow (A \mathrm{f} (B \& C))$

III. $(A \mathrm{f} B) \leftrightarrow !(A \rightarrow B)$

IV. $\exists U\, !U$

V. $\neg(U \mathrm{f} \cap)$

Axiom IV is strange:

- $!U$ is a more natural formalization of (iv).
- Axiom IV seems redundant: $!A \rightarrow !A$ is a tautology, so we have $!A \mathrm{f} A$ by Def. f, whence !(!A

$\rightarrow$ A) by Axiom III($\rightarrow$), whence $\exists$M !M by existential generalization. Axiom IV seems to add nothing to this.

- Axiom IV is the only axiom or theorem mentioned by Mally in which U occurs as a bound variable: in Axiom V and in theorems (15)-(17), (20)-(21), (23), (23′) and (27)-(35) (to be displayed below), U is either a constant or a free variable. One should treat it in the same way in the formalization of (iv).

For these reasons, we replace Axiom IV by the following axiom:[2]

IV. !U

Mally could hardly have objected to this version of Axiom IV because it is equivalent with his theorem (23′), i.e., V f U, in virtue of Def. f. In the following "Axiom IV" will always refer to our version of Axiom IV rather than Mally's.

Using Def. f, Axioms I-V may also be written as follows (Mally 1926, pp. 15-19 and p. 24):

I′. $((A \rightarrow\ !B)\ \&\ (B \rightarrow C)) \rightarrow (A \rightarrow\ !C)$

II′. $((A \rightarrow\ !B)\ \&\ (A \rightarrow\ !C)) \rightarrow (A \rightarrow\ !(B\ \&\ C))$

III′. $(A \rightarrow\ !B) \leftrightarrow\ !(A \rightarrow B)$

IV′. V f U

V′. $\neg(U \rightarrow\ !\cap)$

4. Mally's Theorems

Mally derived the following theorems from his axioms (Mally 1926, pp. 20-34).[3]

(1) $(A \text{ f } B) \rightarrow (A \text{ f } V)$

(2) $(A \text{ f } \wedge) \leftrightarrow \forall M\ (A \text{ f } M)$

(3) $((M \text{ f } A) \vee (M \text{ f } B)) \rightarrow (M \text{ f } (A \vee B))$

(4) $((M \text{ f } A)\ \&\ (N \text{ f } B)) \rightarrow ((M\ \&\ N) \text{ f } (A\ \&\ B))$

(5) $!P \leftrightarrow \forall M\ (M \text{ f } P)$

(6) $(!P\ \&\ (P \rightarrow Q)) \rightarrow\ !Q$

(7) $!P \rightarrow\ !V$

(8) $((A \text{ f } B) \,\&\, (B \text{ f } C)) \rightarrow (A \text{ f } C)$

(9) $(!P \,\&\, (P \text{ f } Q)) \rightarrow\ !Q$

(10) $(!A \,\&\, !B) \leftrightarrow\ !(A \,\&\, B)$

(11) $(A \infty B) \leftrightarrow\ !(A \leftrightarrow B)$

(12) $(A \text{ f } B) \leftrightarrow (A \rightarrow\ !B) \leftrightarrow\ !(A \rightarrow B) \leftrightarrow\ !\neg(A \,\&\, \neg B) \leftrightarrow\ !(\neg A \vee B)$

(13) $(A \rightarrow\ !B) \leftrightarrow \neg(A \,\&\, \neg!B) \leftrightarrow (\neg A \vee\ !B)$

(14) $(A \text{ f } B) \leftrightarrow (\neg B \text{ f } \neg A)$

(15) $\forall M\, (M \text{ f } U)$

(16) $(U \rightarrow A) \rightarrow\ !A$

(17) $(U \text{ f } A) \rightarrow\ !A$

(18) $!!A \rightarrow\ !A$

(19) $!!A \leftrightarrow\ !A$

(20) $(U \text{ f } A) \leftrightarrow (A \infty U)$

(21) $!A \leftrightarrow (A \infty U)$

(22) $!V$

(23) $V \infty U$

(23′) $V \text{ f } U$

(24) $A \text{ f } A$

(25) $(A \rightarrow B) \rightarrow (A \text{ f } B)$

(26) $(A \leftrightarrow B) \rightarrow (A \infty B)$

(27) $\forall M\, (\bigcap \text{ f } \neg M)$

(28) $\bigcap \text{ f } \bigcap$

(29) $\bigcap \text{ f } U$

(30) $\bigcap \text{ f } \Lambda$

(31) $\bigcap \infty \Lambda$

(32) $\neg(U \text{ f } \Lambda)$

(33) $\neg(U \rightarrow \Lambda)$

(34) $U \leftrightarrow V$

(35) $\bigcap \leftrightarrow \Lambda$

5. Surprising Consequences

Mally called theorems (1), (2), (7), (22) and (27)-(35) "surprising" (*befremdlich*) or even "paradoxical" (*paradox*). He viewed (34) and (35) as the most surprising of his surprising theorems. But Mally's reasons for calling these theorems surprising are puzzling if not confused.

Consider, for example, theorem (1). Mally interpreted this theorem as follows: "if A requires B, then A requires everything that is the case" (Mally 1926, p. 20). He regarded this as a surprising claim, and we agree. However, Mally's interpretation of (1) is not warranted. (1) only says that if A requires B, then A requires the *Verum*. The expression "if A requires B, then A requires everything that is the case" is to be formalized as

$$(1') \quad (A \text{ f } B) \rightarrow (C \rightarrow (A \text{ f } C))$$

This formula is an immediate consequence of (1) in virtue of Axiom I. In other words, Mally should have reasoned as follows: (1′) is surprising; but (1′) is an immediate consequence of (1) in virtue of Axiom I; Axiom 1 is uncontroversial; so (1) is to be regarded as surprising.

A similar pattern is to be seen in many of Mally's other remarks about theorems which surprised him. He generally read too much into them and confused them with some of the consequences they had in his system:

- Mally was surprised by (2) because he thought that it says that if A requires B and B is not the case, then A requires every state of affairs whatsoever (Mally 1926, p. 21). But (2) says no such thing. Mally's paraphrase is a paraphrase of $(A \text{ f } B) \rightarrow (\neg B \rightarrow (A \text{ f } C))$ (a consequence of (2) in virtue of Axiom I) rather than (2).
- Mally paraphrased (7) as "if anything is required, then everything that is the case is required" (Mally 1926, p. 28), which is indeed surprising. However, Mally's paraphrase corresponds with $!A \rightarrow (B \rightarrow !B)$ (a consequence of (7) in virtue of Axiom I) rather than (7).
- Mally paraphrased (22) as "the facts ought to be the case" (Mally 1926, p. 24). We grant that this is a surprising claim. But the corresponding formula in Mally's language is $A \rightarrow !A$ (a consequence of (22) in virtue of Axiom I), not (22).
- Mally read (27) as "if something which ought not to be the case is the case, then anything whatsoever ought to be the case" (Mally 1926, pp. 24, 33), but this is a paraphrase of $!\neg A \rightarrow (A \rightarrow !B)$ (a theorem of Mally's system) rather than (27).
- Mally paraphrased (33) as "what is not the case is not obligatory" (Mally 1926, p. 25) and as "everything that is obligatory is the case" (Mally 1926, p. 34). These assertions are indeed surprising, but Mally's readings of (33) are not warranted. They are paraphrases of $!A \rightarrow A$ rather than (33).
- Mally made the following remark about (34) and (35):

> The latter sentences, which seem to identify being obligatory with being the case, are surely the most surprising of our "surprising consequences."[4]

> However, (34) and (35) do not assert that being obligatory is equivalent with being the case, for the latter statement should be formalized as $A \leftrightarrow\, !A$. The latter formula is a theorem of Mally's system, as will be shown in a moment, but it is not to be found in Mally's book.

Mally regarded theorems (28)-(32) as surprising because of their relationships with certain other surprising theorems:

- (28)-(30) are instantiations of (27). But this is not sufficient to call these theorems surprising. Mally actually viewed (28) as less surprising than (27): one might use it to justify retaliation and revenge (Mally 1926, p. 24).
- (31) implies (28)-(30) and is therefore at least as surprising as those theorems.
- Mally viewed (32) as surprising because the surprising theorem (33) is an immediate consequence of (32) and the apparently non-surprising theorem (25).

Mally's list of surprising theorems seems too short: for example, (24) is equivalent with $A \rightarrow\, !A$ in virtue of Def. f. But $A \rightarrow\, !A$ may be paraphrased as "the facts ought to be the case," an assertion which Mally regarded as surprising (Mally 1926, p. 24). So then why didn't he call (24) surprising? Did it not surprise him after (22)?

Even though Mally regarded many of his theorems as surprising, he thought that he had discovered an interesting concept of "correct willing" (*richtiges Wollen*) or "willing in accordance with the facts" which should not be confused with the notions of obligation and willing used in ordinary discourse. Mally's "exact system of pure ethics" was mainly concerned with this concept, but we will not describe this system because it belongs to the field of ethics rather than deontic logic.

6. Menger's Criticism

In 1939, Karl Menger published a devastating critique of Mally's formal system. He first pointed out that $A \leftrightarrow\, !A$ is a theorem of this system. In other words, if A is the case, then A is obligatory, and if A ought to be the case then A is indeed the case. As we have already noted in connection with theorems (34) and (35), Mally made the same claim in informal terms, but the formula $A \leftrightarrow\, !A$ does not occur in his book.

Menger's theorem $A \leftrightarrow\, !A$ may be proven as follows (Menger's proof was different; **PC** denotes the propositional calculus).

First, $A \rightarrow\, !A$ is a theorem:

1.	$A \rightarrow ((!B \rightarrow !B) \,\&\, (B \rightarrow A))$	[**PC**][5]
2.	$((!B \rightarrow !B) \,\&\, (B \rightarrow A)) \rightarrow (!B \rightarrow !A)$	[I′]
3.	$A \rightarrow (!B \rightarrow !A)$	[1, 2, **PC**]
4.	$!B \rightarrow (A \rightarrow !A)$	[3, **PC**]
5a.	$!U$	[Ax. IV]
5b.	$!(!A \rightarrow A)$	[III′($\rightarrow$), **PC**]
6.	$A \rightarrow !A$	[4, either 5a or 5b, **PC**]

Second, $!A \rightarrow A$ is a theorem:

1.	$((U \rightarrow !A) \,\&\, (A \rightarrow \cap)) \rightarrow (U \rightarrow !\cap)$	[I′]
2.	$\neg((U \rightarrow !A) \,\&\, (A \rightarrow \cap))$	[1, V′, **PC**]
3.	$\neg((U \rightarrow !A) \,\&\, (A \rightarrow \cap)) \rightarrow (!A \rightarrow A)$	[**PC**][6]
4.	$!A \rightarrow A$	[2, 3, **PC**]

Because $A \rightarrow !A$ and $!A \rightarrow A$ are theorems, $A \leftrightarrow !A$ is a theorem as well.

Menger gave the following comment:

> This result seems to me to be detrimental for Mally's theory, however. It indicates that the introduction of the sign ! is superfluous in the sense that it may be cancelled or inserted in any formula at any place we please. But this result (in spite of Mally's philosophical justification) clearly contradicts not only our use of the word "ought" but also some of Mally's own correct remarks about this concept, e.g. the one at the beginning of his development to the effect that $p \rightarrow (!q \text{ or } !r)$ and $p \rightarrow !(q \text{ or } r)$ are not equivalent. Mally is quite right that these two propositions are not equivalent according to the ordinary use of the word "ought." But they are equivalent according to his theory by virtue of the equivalence of p and !p (Menger 1939, p. 58).

Almost all deontic logicians have accepted Menger's verdict. After 1939, Mally's deontic system has seldom been taken seriously.

7. Where Did Mally Go Wrong?

Where did Mally go wrong? How could one construct a system of deontic logic which does more justice to the notion of obligation used in ordinary discourse? Three types of answers are possible:

- Mally should not have added his deontic axioms to classical propositional logic;
- Some of his deontic principles should be modified; and
- Both of the above. Menger advocated the latter view: "One of the reasons for the failure of Mally's interesting attempt is that it was founded on the 2-valued calculus of propositions" (Menger 1939, p. 59).

We will only explore the first two suggestions. Each of them will turn out to be sufficient, so the third proposal is overkill.

We will first show that if Mally's deontic principles are added to a system in which the so-called paradoxes of material and strict implication are avoided, many of the "surprising" theorems (such as (34) and (35)) are no longer derivable and A ↔ !A is no longer derivable either. But most of the theorems which Mally regarded as "plausible" are still derivable. The resulting system is closely related to Anderson's relevant deontic logic (1967).

After this, we will show that one might also modify some of the specifically deontic principles: Def. f and Axiom I may be modified in such a way that the resulting system is almost identical with the system nowadays known as standard deontic logic. The resulting system agrees less well with Mally's deontic expectations (as revealed by his expressions of surprise) than the first one does, but it is adequate in the sense that (34), (35) and A ↔ !A are no longer derivable.

8. Alternative Non-Deontic Bases

Mally's informal postulates (i)-(iii) and (v) are conditionals or negations of conditionals, i.e., of the form "if ... then --" or "not: if ... then --." Føllesdal and Hilpinen (1981, pp. 5-6) have suggested that such conditionals should not be formalized in terms of material implication and that some sort of strict implication would be more appropriate. But this suggestion is not altogether satisfactory, for if I′ and III′ are added to the rather weak system of strict implication **S3** (with → as the symbol of strict implication), then the undesirable theorem A → !A is again derivable.[7]

In systems of strict implication the so-called paradoxes of material implication (such as A → (B → A)) are avoided, but the so-called paradoxes of strict implication (such as (A & ¬A) → B) remain. What would happen to Mally's system if *both* kinds of paradoxes were avoided? This question can be answered by adding Mally's axioms to a system in which none of the so-called "fallacies of relevance" can be derived (see the entry on relevance logic).

In the following, we will add Mally's axioms to the most popular system of this type, relevant system **R**. The result is better than in the case of **S3**: most of the theorems which Mally regarded as surprising are no longer derivable, and Menger's theorem A ↔ !A is not derivable either. But many "plausible" theorems can still be derived.

Relevant system **R** with the propositional constant t ("the conjunction of all truths") has the following axioms and rules (Anderson & Belnap 1975, ch. V; $\leftrightarrow$ is defined by $A \leftrightarrow B = (A \to B) \& (B \to A)$):

Self-implication.	$A \to A$
Prefixing.	$(A \to B) \to ((C \to A) \to (C \to B))$
Contraction.	$(A \to (A \to B)) \to (A \to B)$
Permutation.	$(A \to (B \to C)) \to (B \to (A \to C))$
& Elimination.	$(A \& B) \to A$, $(A \& B) \to B$
& Introduction.	$((A \to B) \& (A \to C)) \to (A \to (B \& C))$
$\vee$ Introduction.	$A \to (A \vee B)$, $B \to (A \vee B)$
$\vee$ Elimination.	$((A \to C) \& (B \to C)) \to ((A \vee B) \to C)$
Distribution.	$(A \& (B \vee C)) \to ((A \& B) \vee C)$
Double Negation.	$\neg\neg A \to A$
Contraposition.	$(A \to \neg B) \to (B \to \neg A)$
Ax. t.	$A \leftrightarrow (t \to A)$
Modus Ponens.	$A, A \to B / B$
Adjunction.	$A, B / A \& B$

A relevant version **RD** of Mally's deontic system may be defined as follows:

- The language is the same as the language of **R**, except that we write V instead of t, add the propositional constant U and the unary connective !, and define $\wedge$, $\cap$, f and ∞ as in Mally's system.
- Axiomatization: add Mally's Axioms I-V to the axioms and rules of **R**.

RD has the following properties.

- Axioms I, II and III may be replaced by the following three simpler axioms:[8]

I*. $(A \to B) \to (!A \to !B)$

II*. $(!A \& !B) \to !(A \& B)$

III*. $!(!A \to A)$

- Formulas I′-V′, (3), (4), (6), (8)-(11), (14), (16)-(18), (23′) and (30) are theorems of **RD**.[9]
- Formulas (1), (2), (5), (7), (12), (13), (15), (19)-(23), (24)-(29), (31)-(35), A → !A and !A → A are not derivable.[10]
- There are 12 mismatches between **RD** and Mally's expectations: (5), (12)-(13), (15), (19)-(21), (23) and (24)-(26) are not derivable even though Mally did not regard these formulas as surprising, and (30) is a theorem even though Mally viewed it as surprising.
- Formulas (34) and (35) (the formulas which Mally viewed as the most surprising theorems of his system) are in a sense stranger than Menger's theorem A ↔ !A because the latter theorem is derivable in **RD** supplemented with (34) or (35) while neither (34) nor (35) is derivable in **RD** supplemented with A ↔ !A.[11]

Although most of Mally's surprising theorems are not derivable in **RD**, this has nothing to do with Mally's own reasons for regarding these theorems as surprising. They are not derivable in **RD** because they depend on fallacies of relevance. Mally never referred to such fallacies to explain his surprise. His considerations were quite different, as we have already described.

RD is closely related to Anderson's relevant deontic logic **ARD**, which is defined as **R** supplemented with the following two axioms (Anderson 1967, 1968, McArthur 1981; Anderson used the unary connective O instead of !):

ARD1. $!A \leftrightarrow (\neg A \rightarrow \cap)$

ARD2. $!A \rightarrow \neg!\neg A$

- All theorems of **RD** are theorems of **ARD**.[12]
- ARD1(→) is not a theorem of **RD**+ARD2.[13] This formula does not occur in Mally's book. According to Anderson, Bohnert (1945) was the first one to propose it.[14]
- ARD2 is not a theorem of **RD**+ARD1.[15] This formula does not occur in Mally's book, but Mally endorsed the corresponding informal principle: "a person who wills correctly does not will (not even implicitly) the negation of what he wills; correct willing is free of contradictions."[16]
- **RD** supplemented with ARD1(→) and ARD2 has the same theorems as **ARD**.[17]

Anderson's system has several problematical features (McArthur 1981, Goble 1999, 2001) and **RD** shares most of these features. But we will not go into this issue here. It is at any rate clear that **RD** is better than Mally's original system.

9. Alternative Deontic Principles

Instead of changing the non-deontic propositional basis of Mally's system, one might also modify the specifically deontic axioms and rules. This might of course be done in various ways, but the following approach works well without departing too much from Mally's original assumptions.[18]

First, regard f as primitive and replace Mally's definition of f in terms of → and ! (Def. f, the very first specifically deontic postulate in Mally's book) by the following definition of ! in terms of V and f :

Def. !. !A = V f A

Second, replace Axiom I, which may also be written as (B → C) → ((A f B) → (A f C)), with the following *rule of inference*:

R f . B → C / (A f B) → (A f C)

We may then derive:

1.	B → C / !B → !C	[Def. !, R f]
2.	(!A & !B) → !(A & B)	[Def. !, Ax. II]
3.	!V	[1, Ax. IV, **PC**]
4.	¬!⋀	[1, Ax. III(←), Ax. V, **PC** (ex falso)]

The so-called standard system of deontic logic **KD** is defined as **PC** supplemented with 1-4 (except that ! is usually written as O: see the entry on deontic logic), so the new system is at least as strong as **KD**. It is not difficult to see that it is in fact identical with **KD** supplemented with OU (Mally's Axiom IV) and the following definition of f : A f B = O(A → B). In modern deontic logic, the notion of *commitment* is sometimes defined in this way.

In the new system, Mally's theorems have the following status.

- II′, IV′, (1)-(5), (7)-(11), (13)-(15), (17), (20)-(24) and (27)-(32) are derivable.
- I′, III′, V′, (6), (12), (16), (18)-(19), (25)-(26), (33)-(35), A → !A and !A → A are not derivable.
- There are 20 mismatches with Mally's deontic expectations: 10 "plausible" formulas are no longer derivable, namely I′, V′, (6), (12), (16), (18)-(19) and (25)-(26), and 10 "surprising" theorems are still derivable, namely (1), (2), (7), (22) and (27)-(32).
- Although (34) and (35) are not derivable, adding them would by no means lead to the theoremhood of A ↔ !A.

There were only 12 mismatches in the case of **RD**, so the new system does less justice to Mally's deontic expectations than **RD** did. But it agrees better with his general outlook because it is still based on classical propositional logic, a system to which Mally did not object (not that he had much choice in the 1920s).

Many of Mally's surprising theorems are derivable in **KD**, but they have, as it were, lost their sting: those theorems lead to surprising consequences when combined with Mally's Axiom I and his definition of f , but they are completely harmless without these postulates.

The standard system of deontic logic has several problematical features. The fact that each provable statement is obligatory is often regarded as counterintuitive, and there are many other well-known "paradoxes." The revised version of Mally's system shares these problematical features. But we will not discuss these issues here. The standard system is at any rate better than Mally's original proposal.

10. Conclusion

Mally's deontic logic is unacceptable for the reasons stated by Menger (1939). But it is not as bad as it may seem at first sight. Only relatively minor modifications are needed to turn it into a more acceptable system. One may either change the non-deontic basis to get a system that is similar to Anderson's system, or apply two patches to the deontic postulates to obtain a system similar to standard deontic logic.

Some authors have refused to view Mally's deontic logic as a "real" deontic system and say that they "mention it only by way of curiosity" (Meyer and Wieringa 1993, p. 4). The above shows that this judgment is too harsh. It is only a small step, not a giant leap, from Mally's system to modern systems of deontic logic, so Mally's pioneering effort deserves rehabilitation rather than contempt.

Bibliography

- Anderson, Alan Ross, 1967, "Some nasty problems in the formal logic of ethics," *Noûs*, vol. 1, pp. 345-360.
- Anderson, Alan Ross, 1968, "A new square of opposition: Eubouliatic logic," in *Akten des XIV. Internationalen Kongresses für Philosophie*, vol. 2, pp. 271-284, Vienna: Herder.
- Anderson, Alan Ross, and Nuel D. Belnap, Jr., 1975, *Entailment: The Logic of Relevance and Necessity*, vol. 1, Princeton, N. J.: Princeton University Press.
- Anderson, Alan Ross, Nuel D. Belnap, Jr., and J. Michael Dunn, 1992, *Entailment: The Logic of Relevance and Necessity*, vol. 2, Princeton, N. J.: Princeton University Press.
- Bohnert, Herbert G., 1945, "The semiotic status of commands," *Philosophy of Science*, vol. 12, pp. 302-315.
- Føllesdal, Dagfinn, and Risto Hilpinen, 1981, "Deontic logic: An introduction," in Risto Hilpinen, ed., *Deontic Logic: Introductory and Systematic Readings*, pp. 1-35, Dordrecht: D. Reidel, 2nd edn.
- Goble, Lou, 1999, "Deontic logic with relevance," in Paul McNamara and Henry Prakken, eds., *Norms, Logics and Information Systems: New Studies on Deontic Logic and Computer Science*, pp. 331-345, Amsterdam, etc.: IOS Press.
- Goble, Lou, 2001, "The Andersonian reduction and relevant deontic logic," in Brown Byson and

John Woods, eds., *New Studies in Exact Philosophy: Logic, Mathematics and Science--Proceedings of the 1999 Conference of the Society of Exact Philosophy*, pp. 213-246, Paris: Hermes Science Publications.

- Hacking, Ian, 1963, "What is strict implication?," *Journal of Symbolic Logic*, vol. 28, pp. 51-71.
- Lokhorst, Gert-Jan C., 1999, "Ernst Mally's *Deontik* (1926)," *Notre Dame Journal of Formal Logic*, vol. 40, pp. 273-282.
- Mally, Ernst, 1926, *Grundgesetze des Sollens: Elemente der Logik des Willens*. Graz: Leuschner und Lubensky, Universitäts-Buchhandlung, viii+85 pp. Reprinted in Ernst Mally, *Logische Schriften: Großes Logikfragment, Grundgesetze des Sollens*, edited by Karl Wolf and Paul Weingartner, pp. 227-324, Dordrecht: D. Reidel, 1971.
- McArthur, Robert P., 1981, "Anderson's deontic logic and relevant implication," *Notre Dame Journal of Formal Logic*, vol. 22, pp. 145-154.
- Menger, Karl, 1939, "A logic of the doubtful: On optative and imperative logic," in *Reports of a Mathematical Colloquium*, 2nd series, 2nd issue, pp. 53-64, Notre Dame, Indiana: Indiana University Press.
- Meyer, John-Jules Ch., and Roel J. Wieringa, 1993, "Deontic logic: A concise overview," in John-Jules Ch. Meyer and Roel J. Wieringa, editors, *Deontic Logic in Computer Science*, pp. 3-16, Chichester: John Wiley and Sons.
- Morscher, Edgar, 1998, "Mallys Axiomsystem für die deontische Logik: Rekonstruktion und kritische Würdigung," *ProPhil -- Projekte zur Philosophie*, vol. 2 (*Ernst Mally: Versuch einer Neubewertung*, A. Hieke, ed.), Sankt Augustin: Academia Verlag, pp. 81-165.
- Weinberger, Ota, to appear, "Ernst Mallys Deontik: Ein kritischer Rückblick und ein konstruktiver Ausblick nach einem dreiviertel Jahrhundert," in Thomas Binder, Reinhard Fabian, Uld Höfer and Jutta Valent, eds., *Bausteine zu einer Geschichte der Philosophie an der Universität Graz*, Amsterdam/Atlanta (*Studien zur Oesterreichischen Philosophie*, vol. 32). [We have not yet seen this paper.]
- Whitehead, Alfred North, and Bertrand Russell, 1910, *Principia Mathematica*, vol. 1, Cambridge: Cambridge University Press.

Other Internet Resources

- Basic information page on Ernst Mally (Edward Zalta, Stanford University).
- MaGIC 2.1.4. MaGIC (Matrix Generator for Implication Connectives) is a program which finds matrices for implication connectives for a wide range of propositional logics. MaGIC was written by John Slaney of the Automated Reasoning Group, The Research School of Information Sciences and Engineering, The Australian National University.

Related Entries

logic: deontic | logic: relevance

Acknowledgments

The author is very grateful to Lou Goble, whose extensive comments on two earlier drafts led to many substantial improvements. The author would also like to thank Lou Goble and Edgar Morscher for making their papers on Anderson's and Mally's deontic logic available to him, and Robert K. Meyer for helping him find the matrices used in note 10.

Gert-Jan C. Lokhorst
lokhorst@fwb.eur.nl

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 5, 2002
Content last modified: April 5, 2002

Stanford Encyclopedia of Philosophy
Notes to Mally's Deontic Logic

Notes

1. *Im Jahre 1919 wurde mir das Wort Selbstbestimmung, das in aller Leute Munde war, Anlaß eines Versuches, mir einen klaren Begriff zu dem Wort zu bilden. Natürlich stieß ich dabei alsbald auf die Schwierigkeiten und Dunkelheiten des Sollensbegriffes: das Problem wandelte sich. Grundbegriff aller Ethik, kann der Begriff des Sollens ein brauchbares Fundament ihres Aufbaus nur geben, wenn er in einem System von Axiomen festgelegt ist. Ein solches Axiomensystem führe ich hier vor* (Mally 1926, Preface, p. I).

2. Menger (1939, p. 57) and Føllesdal and Hilpinen (1981, pp. 2-3) made the same decision.

3. Menger (1939, p. 58) said that Mally derived "fifty" theorems, while Føllesdal and Hilpinen (1981, p. 3) said that Mally derived "about fifty" theorems. They probably included some of Mally's unnumbered theorems, which Mally mentioned only in passing. Our list consists of Mally's thirty-six explicitly numbered theorems.

4. *Diese letzten Sätze, die Seinsollen und Tatsächlichsein zu identifizieren scheinen, sind unter unseren "befremdlichen Folgerungen" wohl die befremdlichsten* (Mally 1926, p. 25).

5. The formula on this line is a theorem of the classical propositional calculus because A → (B → B) and A → (B → A) are theorems of this calculus.

6. The formula on this line is a theorem of the classical propositional calculus because

$$
\begin{aligned}
\neg((U \to !A) \,\&\, (A \to \cap)) &\leftrightarrow \neg((U \to !A) \,\&\, (U \to \neg A)) \\
&\leftrightarrow \neg(U \to (!A \,\&\, \neg A)) \\
&\leftrightarrow U \,\&\, \neg(!A \,\&\, \neg A) \\
&\to !A \to A
\end{aligned}
$$

7. On **S3**, see Hacking 1963. The proof that A → !A is a theorem of **S3** enriched with ! and supplemented with I′ and III′ is straightforward and left to the reader.

8. We only prove that I* is a theorem of **RD**. The other five cases (II* and III* are theorems of **RD** and I-

III follow from I*-III*) are left to the reader.

1. $(!A \rightarrow !A) \,\&\, (A \rightarrow ((A \rightarrow B) \rightarrow B))$ [**R**]
2. $!A \rightarrow !((A \rightarrow B) \rightarrow B)$ [1, Ax. I, Def. f]
3. $!A \rightarrow ((A \rightarrow B) \rightarrow !B)$ [2, Ax. III($\leftarrow$), Def. f]
4. $(A \rightarrow B) \rightarrow (!A \rightarrow !B)$ [3, Permutation]

9. Most cases are obvious but some hints for (16)-(18) and (30) may be helpful. (16) is a consequence of I* and IV. (17) follows from (16) and (18). (18) may be proven as follows:

1. $!(!A \rightarrow A) \,\&\, !(!!A \rightarrow !A)$ [III* (twice), Adjunction]
2. $!((!A \rightarrow A) \,\&\, (!!A \rightarrow !A))$ [1, II*]
3. $((!A \rightarrow A) \,\&\, (!!A \rightarrow !A)) \rightarrow (!!A \rightarrow A)$ [**R**]
4. $!((!A \rightarrow A) \,\&\, (!!A \rightarrow !A)) \rightarrow !(!!A \rightarrow A)$ [3, I*]
5. $!(!!A \rightarrow A)$ [2, 4]
6. $!!A \rightarrow !A$ [5, Ax. III($\leftarrow$), Def. f]

(30) is proven as follows: we have $\cap \rightarrow (V \rightarrow \cap)$ by Ax. t, whence $\cap \rightarrow (U \rightarrow \wedge)$ by Contraposition and Double Negation, whence $\cap \rightarrow !\wedge$ by (16).

10. In order to prove this claim, we first have to remove an obstacle: (2), (5), (15) and (27) do not belong to the language of **RD** because they contain propositional quantifiers. If such quantifiers were added in the usual way (Anderson, Belnap & Dunn 1992, sec. 32), then we could easily prove that (2), (5), (15) and (27) are deductively equivalent with (2′), (5′), (15′) and (27′), respectively:

(2′) $(A \text{ f } \wedge) \rightarrow (A \text{ f } B)$

(5′) $(!A \rightarrow (B \text{ f } A)) \,\&\, ((V \text{ f } A) \rightarrow !A)$

(15′) $A \text{ f } U$

(27′) $\cap \text{ f } A$

To avoid needless complications, we will not equip **RD** with quantifiers, but confine our attention to the unquantified versions of (2), (5), (15) and (27). We may now proceed with the actual proof. We use the following matrices:

⥽	0	1	2	3	4	5	¬
0	5	5	5	5	5	5	5
1	0	1	2	3	4	5	4
2	0	0	1	1	3	5	3
3	0	0	0	1	2	5	2
4	0	0	0	0	1	5	1
5	0	0	0	0	0	5	0

&	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	0	1	0	1
2	0	0	2	2	2	2
3	0	1	2	3	2	3
4	0	0	2	2	4	4
5	0	1	2	3	4	5

Each theorem M of **RD** has the following property: v(M)∈{1, 3, 5} for every assignment v of values to the variables in M such that

- v(A) ∈ {0, 1, 2, 3, 4, 5},
- v(V) = 1,
- v(U) = 2,
- v(¬A), v(A ⥽ B) and v(A & B) are as indicated in the tables,
- v(A ∨ B) = v(¬(¬A & ¬B)), and
- v(!A) = v(U ⥽ A).

(1), (2), (5), (7), (12), (13), (15), (19)-(23), (24)-(29), (31)-(35), A ⥽ !A and !A ⥽ A (unquantified versions) lack this property, so these formulas are not theorems of **RD**. (These matrices were discovered by the computer program MaGIC. See the section *Other Internet Resources* for information on this program. The claims we have made were verified by a simple Perl-script of our own making.)

11. Menger's theorem A ↔ !A is a theorem of both **RD**+(34) and **RD**+(35):

1. A ↔ (V ⥽ A)	[Ax. t]
2. A ↔ (U ⥽ A)	[1, either (34) or (35)]
3. A ⥽ !A	[2, (16)]
4. (A ⥽ ∩) ⥽ ((U ⥽ !A) ⥽ (U ⥽ !∩))	[I*, Prefixing]
5. (A ⥽ ∩) ⥽ (¬(U ⥽ !∩) ⥽ ¬(U ⥽ !A))	[4, Contraposition]
6. ¬(U ⥽ !∩) ⥽ ((A ⥽ ∩) ⥽ ¬(U ⥽ !A))	[5, Permutation]
7. (A ⥽ ∩) ⥽ ¬(U ⥽ !A)	[6, Ax. V, Def. f]
8. !A ⥽ A	[2, 7, **R**]

9. A ↔ !A [3, 8, Adjunction]

However, neither (34) nor (35) is derivable in **RD** supplemented with A ↔ !A. To prove this, we use the following table from Anderson and Belnap 1975, p. 148:

→	0 1 2	¬
0	2 2 2	2
1	0 1 2	1
2	0 0 2	0

Each theorem M of **RD**+{A ↔ !A} has the following property: v(M) ∈ {1, 2} for every assignment v of values to the variables in M such that

- v(A)∈ {0, 1, 2},
- v(V) = 1,
- v(U) = 2,
- v(¬A) and v(A → B) are as indicated in the table,
- v(A & B) = min{v(A), v(B)},
- v(A ∨ B) = max{v(A), v(B)}, and
- v(!A) = v(A).

But v((34)) = v(U → V) = 0 and v((35)) = v(⋀ → ⋂) = 0, so (34) and (35) are not theorems of **RD**+{A ↔ !A}.

12. See Lokhorst 1999, pp. 277-278.

13. One may use the three-valued matrices of note 11 to prove this.

14. Anderson 1967, p. 348. However, Menger (1939, p. 59) had already defined "I command p" as "unless p, something unpleasant will happen," or in symbols: Cp ↔ (¬p → A), where Cp stands for "I command p" and A denotes the statement that the unpleasant thing will happen.

15. One may use the six-valued matrices of note 10 to prove this.

16. *Wer richtig will, will nicht (auch nicht impliziterweise) das Negat des Gewollten; richtiges Wollen ist widerspruchsfrei* (Mally 1926, p. 49). Mally regarded this as a paraphrase of Axiom V, but this is not entirely correct. Morscher (1998, p. 106) has suggested that ARD2 expresses Mally's intentions more

adequately than Axiom V does.

17. ARD1 is a theorem of **RD**+ARD1($\rightarrow$) by virtue of theorem (16), i.e., ARD1($\leftarrow$). This completes the proof. Notice that Axiom V is redundant because we have !U by Ax. IV, whence $\neg!\cap$ by ARD2, whence $\neg(U \rightarrow \cap)$ by ARD1, whence $\neg(U \rightarrow (U \rightarrow \cap))$ by **R**, whence $\neg(U \rightarrow !\cap)$ by ARD1, whence Axiom V by Def. f.

18. Morscher (1998) has made a somewhat similar proposal, but the details are different. The present approach was inspired by a number of suggestions made by Lou Goble.

Gert-Jan C. Lokhorst
lokhorst@fwb.eur.nl

First published: June 12, 2002
Content last modified: June 12, 2002

```
#!/usr/bin/perl
#
# $Id: mally.pl,v 1.19 2002/07/02 10:24:48 lokhorst Exp $
#
# Supplement to G.J.C. Lokhorst, "Mally's Deontic Logic",
# in Edward N. Zalta, ed., Stanford Encyclopedia of Philosophy.
# The Metaphysics Research Lab at the Center for the Study of
# Language and Information, Stanford University, Stanford, CA,
# 2002.
#
# Perl script that can also be run as a cgi program.
#

use strict;

$count::main = 0;

sub definitionA {

    # These matrices were discovered by MaGIC version 2.1.4:
    # see the output of MaGIC at the end of this program.
    @values::main    = ( 0 .. 5 );
    @matrixNeg::main = reverse @values::main;
    @matrixCon::main = (
        [ 0, 0, 0, 0, 0, 0 ], [ 0, 1, 0, 1, 0, 1 ],
        [ 0, 0, 2, 2, 2, 2 ], [ 0, 1, 2, 3, 2, 3 ],
        [ 0, 0, 2, 2, 4, 4 ], [ 0, 1, 2, 3, 4, 5 ],
    );
    @matrixArrow::main = (
        [ 5, 5, 5, 5, 5, 5 ], [ 0, 1, 2, 3, 4, 5 ],
        [ 0, 0, 1, 1, 3, 5 ], [ 0, 0, 0, 1, 2, 5 ],
        [ 0, 0, 0, 0, 1, 5 ], [ 0, 0, 0, 0, 0, 5 ],
    );
    @matrixObl::main = ( 0, 0, 1, 1, 3, 5 );
    @matrixDes::main = ( 0, 1, 0, 1, 0, 1 );
}

sub definitionB {
    @values::main      = ( 0 .. 2 );
    @matrixNeg::main   = reverse @values::main;
    @matrixCon::main   = ( [ 0, 0, 0 ], [ 0, 1, 1 ], [ 0, 1, 2 ] );
    @matrixArrow::main = ( [ 2, 2, 2 ], [ 0, 1, 2 ], [ 0, 0, 2 ] );
    @matrixObl::main   = @values::main;
    @matrixDes::main = ( 0, 1, 1 );
}

# connectives
# Internal notation: fully parenthesized prefix notation
$V::main = 1;
$U::main = 2;

# conjunction
sub K {
    return @matrixCon::main[ $_[0] ]->[ $_[1] ];
}

# relevant implication
sub C {
    return @matrixArrow::main[ $_[0] ]->[ $_[1] ];
}

# negation
```

```
sub N {
    return @matrixNeg::main[ $_[0] ];
}

# obligation
sub O {
    return @matrixObl::main[ $_[0] ];
}

# disjunction
sub A {
    return N( K( N( $_[0] ), N( $_[1] ) ) );
}

# equivalence
sub E {
    return K( C( $_[0], $_[1] ), C( $_[1], $_[0] ) );
}

# requirement
sub R {
    return C( $_[0], O( $_[1] ) );
}

# mutual requirement
sub M {
    return K( R( $_[0], $_[1] ), R( $_[1], $_[0] ) );
}

# subroutines for display

# conjunction
sub K_p {
    return "($_[0]" . "&" . "$_[1])";
}

# relevant implication
sub C_p {
    return "($_[0]" . "->" . "$_[1])";
}

# negation
sub N_p {
    return "~" . "$_[0]";
}

# obligation
sub O_p {
    return "!" . "$_[0]";
}

# disjunction
sub A_p {
    return "($_[0]" . "v" . "$_[1])";
}

# equivalence
sub E_p {
    return "($_[0]" . "<->" . "$_[1])";
}

# requirement
sub R_p {
```

```
    return "($_[0]" . "f" . "$_[1])";
}

# mutual requirement
sub M_p {
    return "($_[0]" . "<=>" . "$_[1])";
}

sub commify {
    ( @_ == 0 ) ? ".\n" :
      ( @_ == 1 ) ? "$_[0].\n" :
      ( @_ == 2 ) ? join ( " and ", @_ ) . ".\n" :
      join ( ", ", @_[ 0 .. ( $#_ - 1 ) ], "and $_[-1].\n" );
}

sub hdr {
    my ( $name, $conn, $sep ) = @_;
    printf "%s:\n%${sep}s%${sep}s", $name, $conn, "|";

    foreach (@values::main) { printf "%${sep}d", $_ }
    print "\n";
    printf "-" x ( ( 2 * $sep ) - 1 );
    print "+";
    printf "-" x ( scalar(@values::main) * $sep );
    print "\n";
}

sub conn1 {
    my ( $name, $conn, $symbol, $sep ) = @_;
    hdr $name, "", $sep;
    printf "%${sep}s%${sep}s", $symbol, "|";
    foreach my $x (@values::main) { printf "%${sep}d", eval("$conn($x)") }
    print "\n";
}

sub conn2 {
    my ( $name, $conn, $symbol, $sep ) = @_;
    hdr $name, $symbol, $sep;
    foreach my $y (@values::main) {
        printf "%${sep}d%${sep}s", $y, "|";
        foreach my $x (@values::main) {
            printf "%${sep}d", eval("$conn( $y, $x )");
        }
        print "\n";
    }
}

sub showmatrices {
    my @des;
    my $sep = 3;
    print "Admissible values: " . commify(@values::main);
    foreach (@values::main) {
        if ( $matrixDes::main[$_] ) { push ( @des, $_ ) }
    }
    print "Designated values: " . commify(@des);
    printf "Constants: v(V)=%d and v(U)=%d.\n", $V::main, $U::main;
    conn1 'Negation', 'N', '~', $sep;
    conn2 'Relevant Implication', 'C', '->', $sep;
    conn2 'Conjunction', 'K', '&', $sep;
    conn1 'Obligation',  'O', '!', $sep;
    conn2 'Disjunction', 'A', 'v', $sep;
    conn2 'Relevant Equivalence', 'E', '<->', $sep;
```

```
    conn2 'Requirement',        'R', 'f',   $sep;
    conn2 'Mutual Requirement', 'M', '<=>', $sep;
}

sub showformula {
    my ( $name, $wff, $result ) = @_;
    printf "%-12s%-32s%s", $name, $wff, $result;
}

# subroutines for checking
# print and store
sub process {
    my $s;
    my @list;
    my ( $name, $wff, $result ) = @_;

    if ( @$result == 0 ) {
        showformula( $name, $wff, "valid\n" );
        push ( @validnrs::main, $name );
    }
    else {

        $s = "v($name)=@$result[0]";
        if ( @$result > 1 ) {
            $s = $s . " if ";
            push ( @list, "v(A)=@$result[1]" );
            if ( @$result > 2 ) {
                push ( @list, "v(B)=@$result[2]" );

                if ( @$result > 3 ) {
                    push ( @list, "v(C)=@$result[3]" );

                    if ( @$result > 4 ) {
                        push ( @list, "v(D)=@$result[4]" );
                    }
                }
            }
        }
        showformula( $name, $wff, $s . commify(@list) );
        push ( @invalidnrs::main, $name );
    }
}

# returns string in infix notation
sub wffstr {
    my ($wff) = @_;
    $wff =~ s/([A-Z])/${1}_p/g;
    my ( $a, $b, $c, $d, $u, $v ) = ( "A", "B", "C", "D", "U", "V" );
    return eval($wff);
}

# main loops
sub check {
    my $u = $U::main;
    my $v = $V::main;
    my ( $name, $wff ) = @_;
    my @msg;

    if ( $wff =~ /\$d/ ) {
        CHECK4:
        foreach my $a (@values::main) {
            foreach my $b (@values::main) {
```

```
                foreach my $c (@values::main) {

                    foreach my $d (@values::main) {
                        my $z = eval($wff);
                        $count::main++;
                        if ( !$matrixDes::main[$z] ) {
                            @msg = ( $z, $a, $b, $c, $d );
                            last CHECK4;
                        }
                    }
                }
            }
        }
    }
    elsif ( $wff =~ /\$c/ ) {
        CHECK3:

        foreach my $a (@values::main) {
            foreach my $b (@values::main) {
                foreach my $c (@values::main) {
                    my $z = eval($wff);
                    $count::main++;

                    if ( !$matrixDes::main[$z] ) {
                        @msg = ( $z, $a, $b, $c );
                        last CHECK3;
                    }
                }
            }
        }
    }

    elsif ( $wff =~ /\$b/ ) {
        CHECK2:
        foreach my $a (@values::main) {
            foreach my $b (@values::main) {
                my $z = eval($wff);
                $count::main++;

                if ( !$matrixDes::main[$z] ) {
                    @msg = ( $z, $a, $b );
                    last CHECK2;
                }
            }
        }
    }
    elsif ( $wff =~ /\$a/ ) {
        CHECK1:

        foreach my $a (@values::main) {
            my $z = eval($wff);
            $count::main++;

            if ( !$matrixDes::main[$z] ) {
                @msg = ( $z, $a );
                last CHECK1;
            }
        }
    }
    else {
        my $z = eval($wff);
        $count::main++;
```

```
        if ( !$matrixDes::main[$z] ) {
            @msg = ($z);
        }
    }

    process( $name, wffstr($wff), \@msg );
}

sub checkrule {
    my $u = $U::main;
    my $v = $V::main;
    my ( $name, $wff0, $wff1, $wff2 ) = @_;
    my @msg;
    CheckRule2:

    foreach my $a (@values::main) {
        if ( $matrixDes::main[ eval($wff0) ] ) {
            foreach my $b (@values::main) {
                if ( $matrixDes::main[ eval($wff1) ] ) {
                    my $z = eval($wff2);
                    $count::main++;

                    if ( !$matrixDes::main[$z] ) {
                        @msg = ( $z, $a, $b );
                        last CheckRule2;
                    }
                }
            }
        }
    }

    my $s = wffstr($wff0) . "," . wffstr($wff1) . "/" . wffstr($wff2);
    process( $name, $s, \@msg );
}

# tests
sub testformulas {
    my $prime = "'";
    @validnrs::main   = ();
    @invalidnrs::main = ();

    printf "## %d-valued matrices\n", scalar(@values::main);
    showmatrices;

    print "## Alethic Axioms\n";

    check( "Self-impl",  'C($a,$a)' );
    check( "Pref",       'C(C($a,$b),C(C($c,$a),C($c,$b)))' );
    check( "Contract",   'C(C($a,C($a,$b)),C($a,$b))' );
    check( "Perm",       'C(C($a,C($b,$c)),C($b,C($a,$c)))' );
    check( "A<->(V->A)", 'E($a,C($v,$a))' );
    check( "DblNeg",     'C(N(N($a)),$a)' );
    check( "Contrapos",  'C(C($a,N($b)),C($b,N($a)))' );
    check( "&-Elim1",    'C(K($a,$b),$a)' );
    check( "&-Elim2",    'C(K($a,$b),$b)' );
    check( "&-Int",      'C(K(C($a,$b),C($a,$c)),C($a,K($b,$c)))' );
    check( "v-Int1",     'C($a,A($a,$b))' );
    check( "v-Int2",     'C($b,A($a,$b))' );
    check( "v-Elim",     'C(K(C($a,$c),C($b,$c)),C(A($a,$b),$c))' );
    check( "Distr",      'C(K($a,(A($b,$c))),A(K($a,$b),$c))' );
```

```
print "## Rules of Inference\n";

checkrule( "MP",  '$a', 'C($a,$b)', '$b' );
checkrule( "Adj", '$a', '$b',       'K($a,$b)' );

print "## Deontic Axioms\n";

check( "I",   'C(K(R($a,$b),C($b,$c)),R($a,$c))' );
check( "II",  'C(K(R($a,$b),R($a,$c)),R($a,K($b,$c)))' );
check( "III", 'E(R($a,$b),O(C($a,$b)))' );
check( "IV",  'O($u)' );
check( "V",   'N(R($u,N($u)))' );

print "## Formulas (Mally/Menger/Anderson)\n";

check( "01",       'C(R($a,$b),R($a,$v))' );
check( "02",       'C(R($a,N($v)),R($a,$b))' );
check( "03",       'C(A(R($a,$b),R($a,$c)),R($a,A($b,$c)))' );
check( "04",       'C(K(R($a,$b),R($c,$d)),R(K($a,$c),K($b,$d)))' );
check( "05",       'K(C(O($a),R($b,$a)),C(R($v,$a),O($a)))' );
check( "06",       'R(K(O($a),C($a,$b)),$b)' );
check( "07",       'C(O($a),O($v))' );
check( "08",       'C(K(R($a,$b),R($b,$c)),R($a,$c))' );
check( "09",       'R(K(O($a),R($a,$b)),$b)' );
check( "10",       'E(K(O($a),O($b)),O(K($a,$b)))' );
check( "11",       'E(M($a,$b),O(E($a,$b)))' );
check( "12",       'E(O(C($a,$b)),O(N(K($a,N($b)))))' );
check( "13",       'E(R($a,$b),N(K($a,N(O($b)))))' );
check( "14",       'E(R($a,$b),R(N($b),N($a)))' );
check( "15",       'R($a,$u)' );
check( "16",       'R(C($u,$a),$a)' );
check( "17",       'R(R($u,$a),$a)' );
check( "18",       'C(O(O($a)),O($a))' );
check( "19",       'E(O(O($a)),O($a))' );
check( "20",       'E(R($u,$a),M($a,$u))' );
check( "21",       'E(O($a),M($a,$u))' );
check( "22",       'O($v)' );
check( "23",       'M($v,$u)' );
check( "23$prime", 'R($v,$u)' );
check( "24",       'R($a,$a)' );
check( "25",       'C(C($a,$b),R($a,$b))' );
check( "26",       'C(E($a,$b),M($a,$b))' );
check( "27",       'R(N($u),$a)' );
check( "28",       'R(N($u),N($u))' );
check( "29",       'R(N($u),$u)' );
check( "30",       'R(N($u),N($v))' );
check( "31",       'M(N($u),N($v))' );
check( "32",       'N(R($u,N($v)))' );
check( "33",       'N(C($u,N($v)))' );
check( "34",       'E($u,$v)' );
check( "35",       'E(N($u),N($v))' );
check( "A<->!A",   'E($a,O($a))' );
check( "ARD1->",   'C(O($a),C(N($a),N($u)))' );
check( "ARD1<-",   'C(C(N($a),N($u)),O($a))' );
check( "ARD2",     'C(O($a),N(O(N($a))))' );

printf "## Summary (%d-valued matrices)\n", scalar(@values::main);
print "Valid formulas: " . commify(@validnrs::main);
print "Invalid formulas: " . commify(@invalidnrs::main);
if ( scalar(@values::main) == 6 ) { @invalidnrs6::main = @invalidnrs::main }
elsif ( scalar(@values::main) == 3 ) {
    @invalidnrs3::main = @invalidnrs::main;
```

```
    }
}

sub summary {
    print "## Global summary\n";
    print "Formulas that are invalid according to either ";
    print "the 3-valued or the 6-valued matrices: ";
    my %union;

    foreach my $e (@invalidnrs6::main) { $union{$e} = 1 }
    foreach my $e (@invalidnrs3::main) { $union{$e} = 1 }
    print commify( sort { $a cmp $b } keys %union );
    printf "## Done. %d values tested.\n", $count::main;
}

# Main program
# next line turns this program into a cgi program
print "Content-type: text/plain\n\n";

definitionA;
testformulas;
definitionB;
testformulas;
summary;
exit;

# Output from MaGIC. Definition: u o u = ~(u -> ~u).
#
# Logic: R
# Extra: (u o u) o u
# Fragment: ->, &, v, ~, o, t, f, T, F
# Definition: u Primitive (cut)
# Fail: u o u
# Search concludes when 1 matrix found
# or when size 14 finished.
#
# Size: 6
#
# Negation table 6.1
#  a| 0 1 2 3 4 5
# --+------------
# ~a| 5 4 3 2 1 0
#
# Order 6.1.1
# < | 0 1 2 3 4 5
# --+------------
# 0 | + + + + + +
# 1 | - + - + - +
# 2 | - - + + + +
# 3 | - - - + - +
# 4 | - - - - + +
# 5 | - - - - - +
#
# Choice 6.1.1.1 of t: 1
#
# Implication matrix 6.1.1.1.1
# ->| 0 1 2 3 4 5
# --+------------
# 0 | 5 5 5 5 5 5
# 1 | 0 1 2 3 4 5
# 2 | 0 0 1 1 3 5
```

```
# 3 | 0 0 0 1 2 5
# 4 | 0 0 0 0 1 5
# 5 | 0 0 0 0 0 5
# Choice 6.1.1.1.1.1 of u: 2
# Failure: u o u
#
# That's all, for now.
```

Theological Voluntarism

There is a class of metaethical and normative views that commonly goes by the name ‘divine command theory.’ What all members of this class have in common is that they hold that what God wills is relevant to determining the moral status of some set of entities (acts, states of affairs, character traits, etc., or some combination of these). But the name ‘divine command theory’ is a bit misleading: what these views have in common is their appeal to the divine will; while many of these views hold that the relevant act of divine will is that of commanding, some deny it. So we would do well to have a label for this class of views that does not prejudge the issue of the relevant act of divine will. The label that I will use, following Quinn 1990, is ‘theological voluntarism.’

I have three aims in this article. I want first to distinguish metaethical versions of theological voluntarism from normative versions of that view, putting to the side normative versions. Second, I will say something about the main lines of defense of theological voluntarism, the various theoretical options that confront defenders of theological voluntarism, and some of the reasons for affirming or rejecting these different possible formulations. And finally I will say a bit about the sort of difficulties that seem to confront any such views. (I do not, however, give an account of the history of theological voluntarism in moral philosophy; for an anthology of readings covering a broad swath of this history, see Idziak 1979.)

1. Metaethical and normative theological voluntarism

To be a theological voluntarist is to hold that entities of some kind have at least some of their moral statuses in virtue of certain acts of divine will. But some instances of this view are metaethical theses; some instances of it are normative theses.

Consider, for example, theological voluntarism about the status of acts as obligatory or non-obligatory. One might hold that there is a single supreme obligation, the obligation to obey God. Every particular type of act that one might perform thus has its moral status as obligatory or non-obligatory in virtue of God's having commanded the performance of acts of that type or God's not having commanded acts of that type. This is a common version of divine command theory, according to which all of the more workaday obligations that we are under (not to steal from each other, not to murder each other, to help each other out when it would not be inconvenient, etc.) bind us as a result of the exercise of God's supreme practical authority.

The view just described is a version of *normative* theological voluntarism. It is a normative view because it asserts that some normative state of affairs obtains -- namely, the normative state of affairs *its being obligatory to obey God*. And it is a version of theological voluntarism because it holds that all other normative states of affairs, at least those involving obligation, obtain in virtue of God's commanding activity.

Metaethical theological voluntarist views, by contrast, do not assert the obtaining of any normative state of affairs. It is possible for one to be a metaethical theological voluntarist and to hold that no normative states of affairs obtain. Rather, metaethical theological voluntarists aim to say something interesting and informative about moral concepts, properties, or states of affairs; and they want to say something interesting and informative about them by connecting them to acts of the divine will. Metaethical theological voluntarists might claim that (e.g.) obligation is a theological concept, or that the property of being obligatory is a theological property, or that obligations are caused immediately by the divine will. But note that none of these views asserts that there are any obligations.

1.1 Theological voluntarism and theism

One does not have to be a theist in order to be a theological voluntarist. One can affirm normative theological voluntarism or metaethical theological voluntarism while failing to affirm theism; atheists and agnostics can be theological voluntarists of either stripe. With respect to normative theological voluntarism: one might claim that while it is true that any being that merits the title of ‘God’ merits obedience, we should not believe that there is such a being. (Compare: if, through some glitch in

promotions, there happened to be no lieutenants in the army at some time, it would not cease to be true that privates ought to obey lieutenants. One could believe that that there are no lieutenants while believing that privates ought to obey their lieutenants.) With respect to metaethical theological voluntarism: one might claim that, for example, the concept of obligation is ineliminably theistic, though there is no God; that God does not exist counts not against metaethical theological voluntarism but rather against the claim that the concept of obligation has application. (Compare: one might believe that 'sin' is properly defined as 'offense against God.' One can clearly affirm this definition while rejecting God's existence; all that one is committed to thereby is that there really are no sins.)

1.2 Theological voluntarism and moral skepticism

Call a 'moral skeptic' one who disbelieves or withholds judgment on the claim that any normative state of affairs obtains. One can affirm metaethical theological voluntarism while being a moral skeptic; one cannot affirm normative theological voluntarism while being a moral skeptic. A metaethical theological voluntarist might claim that no normative state of affairs could be made to obtain without certain acts of divine will, but because there is no God, or because there is a God that has not performed the requisite acts of will, no normative states of affairs obtain. A normative theological voluntarist cannot, however, be a moral skeptic. Because the normative theological voluntarist is committed to the obtaining of at least one normative state of affairs -- for example, *its being obligatory to obey God* -- the conjunction of moral skepticism and normative theological voluntarism is not a coherent combination of views.

My concern in the rest of this article will be with the metaethical version of theological voluntarism; any further references to theological voluntarism are, unless otherwise noted, to the metaethical version of the position. Theological voluntarism thus understood is consistent either with the affirmation or with the denial of theism and moral skepticism. Taking a negative stand on theism or a positive stand on moral skepticism should not prevent one from taking seriously theological voluntarism as a philosophical position. This is an important point, because it is often thought that theological voluntarism is only for theists, or only for moral nonskeptics. While it is true that some of the arguments for theological voluntarism take theism, or the existence of moral obligations, as premises, not all of them do.

2. Metaethical theological voluntarism

Metaethics is concerned with the formulation of interesting and informative accounts of normative concepts, properties, and states of affairs; and a metaethics that is a version of theological voluntarism will formulate such accounts in terms of some acts of divine will. This statement of the position is highly abstract, but it cannot be made less abstract without making difficult choices among rival formulations of the view.

2.1 Considerations in Favor

The considerations to be offered in favor of theological voluntarism are, at this level, similarly abstract. I

will discuss three types of consideration: *historical*, *theological*, and *metaethical*.

Historical considerations in favor of theological voluntarism

Some of the considerations in favor of metaethical theological voluntarism are historical. Both theists and nontheists have been impressed by the extent to which at least some moral concepts developed in tandem with theological concepts, and it may therefore be the case that there could be no adequate explication of some moral concepts without appeal to theological ones. On this view, it is not merely historical accident that at least some moral concepts had their origin in contexts of theistic belief and practice; rather, these concepts have their origin essentially in such contexts, and become distorted and unintelligible when exported from those contexts (see, for example, Anscombe 1958).

Theological considerations in favor of theological voluntarism

Some of the considerations in favor of theological voluntarism have their source in matters regarding the divine nature. Several such arguments are summarized in Idziak 1979 (pp. 8-10). Some appeal to *omnipotence*: since God is both omnipotent and impeccable, theological voluntarism must be true: for if God cannot act in a way that is morally wrong, then God's power would be limited by other normative states of affairs were theological voluntarism not the case. Some appeal to God's *freedom*: since God is free and impeccable, theological voluntarism must be true: for if moral requirements existed prior to God's willing them, requirements that an impeccable God could not violate, God's liberty would be compromised. Some appeal to God's status as supremely *lovable* and *deserving of allegiance*: if theism is true, then the world of value must be a theocentric one, and so any moral view that does not place God at its center is bound to be inadequate. Even if individually insufficient as justifications for adopting theological voluntarism, collectively they may suggest some desiderata for a moral view: that God must be at the center of a moral theory, and, in particular, that the realm of the moral must be dependent on God's free choices. It seems that any moral theory that met these desiderata would count as a version of theological voluntarism.

Metaethical considerations in favor of theological voluntarism

A third set of considerations in favor of theological voluntarism has its source in metaethics proper, in the attempt to provide adequate philosophical accounts of the various formal features exhibited by moral concepts, properties, and states of affairs. One might claim, that is, that theological voluntarism makes the best sense of the formal features of morality that both theists and nontheists acknowledge.

Consider first the *normativity* of morals. Both theists and nontheists have been impressed by the weirdness of normativity, with its very otherness, and have thought that whatever we say about normativity, it will have to be a story not about natural properties but nonnatural ones (cf. Moore 1903, section 13). John Mackie, an atheist, and George Mavrodes, a theist, have both drawn from this the same moral: if there is a God, then the normativity of morality can be understood in theistic terms; otherwise, the normativity of morality is unintelligible (Mavrodes 1986; Mackie 1977, p. 48). As Robert Adams has

suggested, given the serious difficulties present in understanding morality as a natural property, it is worthwhile taking seriously the hypothesis that morality is not just a nonnatural matter but a supernatural one (Adams 1973, p. 105). For the standard objections against understanding normativity as a nonnatural property concern our inability to say anything further about that nonnatural property itself and about our ability to grasp that property (see, e.g., Smith 1994, pp. 21-25). But if morality is to be understood in terms of God's commands, we can give an informative account of what these unusual properties are; and if it is understood in terms of God's commands, then we can give an informative account of how God, being the creator and sustainer of us rational beings, can ensure that we can have an adequate epistemic grasp of the moral domain (Adams 1979a, pp. 137-138).

Consider next the *impartiality* of morals. The domain of the moral, unlike the domain of value generally, is governed by the requirements of impartiality. To use Sidgwick's phrase, the point of view of morality is not one's personal point of view but rather "the point of view … of the Universe" (Sidgwick 1907, p. 382). But, to remark on the perfectly obvious, the Universe does not have a point of view. Various writers have employed fictions to try to provide some sense to this idea: Smith's impartial and benevolent spectator, Firth's ideal observer, and Rawls' contractors who see the world *sub specie aeternitatis* come to mind most immediately (Smith 1759, Pt III, Ch 8; Firth 1958; and Rawls 1971, p. 587). But theological voluntarism can provide a straightforward understanding of the impartiality of morals by appealing to the claim that the demands of morality arise from the demands of someone who in fact has an impartial and supremely deep love for all of the beings that are morality's proper objects.

Consider next the *overridingness* of morals. The domain of the moral, it is commonly thought, consists in a range of values that can demand absolute allegiance, in the sense that it is never reasonable to act contrary to what those values finally require. One deep difficulty with this view, formulated in a number of ways but perhaps most memorably by Sidgwick (1907, pp. 497-509), is that it is hard to see how moral value automatically trumps other kinds of value (e.g. prudential value) when they conflict. But if the domain of the moral is to be understood in terms of the will of a being who can make it possible that, or even ensure that, the balance of reasons is always in favor of acting in accordance with the moral demand, then the overridingness of morals becomes far easier to explain.

Consider next the *content* of morals. There is a strong case to be made that moral judgments cannot have just any content: they must be concerned, somehow, with what exhibits respect for certain beings, or with what promotes their interests (cf. Foot 1958, pp. 510-512; Smith 1994, p. 40). Theological voluntarism has a ready explanation for the content of morals being what it is: it is that moral demands arise from a being that loves that being's creation.

So there are some general reasons to think theological voluntarism promising. The reasons are stronger yet when one is proceeding from theistic starting points. (This is not trivial, since a number of theistic philosophers reject theological voluntarism.) But these reasons, while suggestive, are rather generic: they point to the promise possessed by theological voluntarism, though they do not fix for us on a particular formulation of the view. The general schema for a particular theological voluntarist position is 'evaluative status *M* stands in dependence relation *D* to divine act *A*' (cf. Quinn 1999, p. 53, which I follow here except to substitute the more general 'evaluative' for Quinn's more specific 'moral'). So

there are at least three choices that have to be made. We need to say something about what sorts of evaluative statuses depend on God's will. We need to say something about what are the relevant acts of divine will. And we need to say something about what the dependence relation is supposed to be. (These are not independent questions, of course.)

2.2 Which evaluative statuses?

A metaethical view can be more or less comprehensive, aiming to cover more or fewer evaluative statuses. A metaethical view might claim to provide an account of all evaluative notions, or of all normative notions, or of all moral notions, or or some set of moral notions. (Roughly, and taking the notion of an evaluative property as fundamental: for a notion to be normative is for it to be a certain sort of evaluative notion, one that is essentially action-guiding; for a notion to be moral is for it to be a certain sort of normative notion, one that exhibits impartiality.) No one claims that theological voluntarism provides an account of all evaluative notions. The real contenders are the latter three.

There are good reasons to reject the claim that all normative notions are to be understood in relation to God's will. The main reason is that, as we will see below, it is important that there be items with normative statuses independent of God's will in order to explain how God's will, even if free, is not arbitrary. And it is not as if the view that some normative statuses are not to be explained in terms of God's will must be repugnant to a theocentric metaethics: for, after all, one might understand such statuses in theological, even if not voluntaristic, terms. Adams, for example, understands some notions of goodness in terms of likeness to God, an understanding that is unquestionably theocentric though not voluntaristic (Adams 1999, pp. 28-38).

Most of the current debate over the evaluative statuses to be explained by theological voluntarism, then, concerns whether the entire set or only some proper subset of moral statuses is to be understood in both theological and voluntaristic terms. Quinn (1978) offers a theological voluntarist view on which all moral statuses are to be understood in terms of God's will. But Adams rejects this view, and Quinn, following Adams and Alston (1990), now rejects it as well. These writers hold that only moral properties in the "obligation family," properties like those of *being obligatory*, *being permissible*, *being required*, and *being right* (where *being right* involves a constraint on conduct, rather than being merely fitting), are to be understood in theological voluntarist terms. Call their view the *restricted* moral view; call the view that all moral statuses are to be understood in voluntarist terms the *unrestricted* moral view.

The restricted moral view has been defended with more and less impressive arguments. The less impressive arguments are those that appeal to the idea that there must be moral properties that are not explained in terms of God's will in order to deal with some of the classic objections to theological voluntarism. To preserve the notion that God is good, one might say, we need to restrict the aspirations of theological voluntarism to those of explaining a proper subset of moral notions, leaving the remainder for an account of God's goodness; or, to make intelligible the commands that God chooses to give, we need to set aside some group of moral notions to be explained in other than theological voluntarist terms and that can therefore enter into our account of the intelligibility of God's choices to give certain

commands rather than others. But these considerations are not, after all, entirely persuasive. For one might well explain the notion that God is good and account for the intelligibility of divine commands by appeal to normative notions that are nonmoral. (See Section 3 below for further discussion of these arguments.)

More plausible are arguments that suggest that there is something in particular about obligation that makes it fit for a theological voluntarist explanation, some feature that is not shared with notions like moral virtue and moral good. Adams suggests, with some plausibility, that the notion of obligation is ineliminably social, that it must involve a relationship between persons, a relationship in which a demand is made (Adams 1987b, p. 264; also Adams 1999, pp. 245-246). This feature of obligation makes it different from notions of goodness and virtue, which do not seem to have this essentially social element. That obligation is special in this way does not, of course, show that notions of moral virtue and moral goodness do not also need to be treated in a theological voluntarist way. It could be that even if obligation most obviously requires this treatment, the points made earlier about the promise of theological voluntarism also extend to other moral notions, even if in a less pressing way. (See the Section 3.3 below for further discussion, and evaluation, of this point.)

There are at present no decisive reasons for the theological voluntarist to adhere to either the restricted or the unrestricted moral view. But theological voluntarists want to say that at least that the properties in the obligation family are to be accounted for in terms of this view. In the remainder of this article, it will be assumed that theological voluntarism is about properties in the obligation family, though we will occasionally consider how the view could be extended to other moral properties as well.

2.3 Which act of divine will?

Assume, then, that theological voluntarism is an account of obligation-type properties. A second issue concerning the proper formulation of the view concerns the relevant act of divine will. Is the requisite act of divine will is to be understood as an act of commanding, or instead as some mental act like choosing, intending, preferring, or wishing? And if one holds that the act of the divine will is a mental act, should the mental act to which the theological voluntarist appeals in order to account for obligation be one whose object is the action that is made obligatory, or one whose object is the state of affairs that the action is obligatory? We have, to simplify matters, three options:

> (1) That it is obligatory for *A* to ϕ depends on God's commanding *A* to ϕ.
>
> (2) That it is obligatory for *A* to ϕ depends on God's willing that *A* ϕ.
>
> (3) That it is obligatory for *A* to ϕ depends on God's willing that it be obligatory for *A* to ϕ.

One might think that the central issue here would be to decide between the speech-act view (1) and the mental acts views (2) and (3); it might be thought to be less important, an issue of intramural interest

only, to decide between (2) and (3). But this is not right. The important debate is between (1) and (2). For (3) is, understood in one way, no competitor with either (1) or (2); and understood differently, it has little argumentative support.

The dispute between (1) and (2).

There is an ongoing debate whether (1) or (2) is the better formulation of theological voluntarism about obligation. There are initially plausible points on both sides of the issue. In favor of (1), one might appeal to the centrality of the image of God as commander in the Abramic faiths. In favor of (2), one might appeal to the centrality to theistic belief and practice of the idea that doing God's will is the standard for the moral life.

With only these initial points, there can be no resolution, and so defenders of these two formulations of theological voluntarism have sought other argumentative routes. One might try to reduce (2) to absurdity. One who is rational does not intend what one knows will not happen; and, on the orthodox conception of God, God is both rational and omniscient. This entails that God will never intend something that will not happen. If obligation arises from divine intentions, though, then no obligations will ever be violated. Since this is absurd, one should prefer (1) over (2). But defenders of (2) have a plausible response. First, defenders of (1) are in no better a position than defenders of (2). For it is a sincerity condition on the giving of commands that the commander intend that the commanded perform the action; and so if this objection reduces (2) to absurdity, the only way that the defender of (1) can avoid having his or her position reduced to absurdity is by holding that God is not necessarily sincere. Second, the notion of intention admits of various readings, and there is a reading of intention suitable for theological voluntarism that does not have the untoward result that no created rational beings could ever act contrary to a divine intention. It is standard to distinguish between God's *antecedent* and God's *consequent* will: God's consequent will is God's will absolutely considered, as bearing on all actual circumstances; God's antecedent will is God's will considered with respect to some proper subset of actual circumstances. (To use an example of Aquinas's, one drawn from the discussion in which the distinction between antecedent and consequent will is made [*Summa Theologiae*, Ia, Q. 19, A. 6]: while in one way God wills that all persons be saved, in another way God does not will that all be saved; indeed, God wills that some persons be damned. What makes this coherent is that the sense of willing in which God wills that all be saved is antecedent: prior to a consideration of all of the particulars of persons' situations, God wills their salvation; but in light of all of the particulars of persons' situations -- including the circumstance that some persons have willingly rejected friendship with God -- God wills their damnation.) So while there is a sense in which it is true that all that God intends must come to pass, this sense is that of consequent intending rather than antecedent intending. On (2), then, one can say that the theological voluntarist holds that obligation depends on certain of God's *antecedent* intentions. (See also Murphy 1998, pp. 17-21.)

We might ask, in order to bring the differences between these views better out into the light, what our considered opinion is in cases in which a divine antecedent intention that $A \wp$ and a divine command that $A \wp$ pull apart. It is far from clear that it is a real option for God to command that $A \wp$ while not intending that $A \wp$. Though this possibility is endorsed by Wierenga 1983 (p. 390) and is at least

entertained in Murphy 1998 (p. 9), for God to issue such a command would be for God to command insincerely -- something that many would be loath to allow. (See also Adams 1999, p. 260 and Murphy 2002, 2.5) But the other option appears unproblematic enough. God might intend for humans to act a certain way while not commanding them to do so. In such a scenario, one might ask, is obligation engendered? If yes, then it seems to count in favor of the divine will view (2); if no, then it seems to count in favor of the divine command view (1).

Adams claims that obligations are not engendered in such cases; actually made demands are necessary. He offers three reasons for preferring the command conception in these cases. The first is that holding that obligation is a matter of divine command rather than divine will makes possible a distinction between the obligatory and the supererogatory: we can say that while God's commands somehow makes certain acts obligatory, if God does will that we perform some act but does not command it, performing that act is supererogatory. The second is an appeal to the idea that theological voluntarism is a social conception of obligation: obligation arises in the context of the social relationship between God and created rational beings. (Not all versions of theological voluntarism affirm this; see the discussion of (3) below.) But, Adams says, in social relationships obligations arise only when demands are actually made. And the third reason is that there is something unsavory about obligations allegedly resulting from an act of divine will that is not expressed as a command: “Games in which one party incurs guilt for failing to guess the unexpressed wishes of the other party are not nice games. They are no nicer if God is thought of as a party to them” (Adams 1999, p. 261).

But the defender of the divine will view (2) has some responses available. The defender of this view can say, with respect to the first point, that a divine will view can capture the difference between the obligatory and the supererogatory not by appeal to the difference between acts of divine will that are expressed as commands and those that are not but rather as a difference between distinct types of act of divine will: the difference, say, between what God intends that we do and what God merely prefers that we do (cf. Quinn 1999, p. 56). With respect to the second and third points, the defender of the divine will view can directly challenge Adams' view that obligations generated within social relationships must always be expressed as demands. Spouses, for example, often take themselves to be obligated by what their spouses intend with respect to their behavior; indeed, it would be unseemly to hold oneself to be bound by one's spouse's will only if the spouse has actually made a demand on one. (“How can you blame me for not helping you empty the dishwasher? You didn't tell me to!” does not often go over well.) One often wants another to perform some action without being told to; many actions have their value only through being performed without being prompted by a command. But, on (1), no act of the form ‘φ*-ing*, though God has not told me to φ’ could ever be obligatory.

Here is a thought experiment that may help to decide the dispute between these two camps. For it to be possible for one to give another a command to φ, there must be a linguistic practice available to the addressee in terms of which the speaker can formulate a command. This is not just for the sake of having the means to communicate a command: rather, commands are essentially linguistic items, and cannot be defined except in such terms. Imagine, though, that a certain created rational being, Mary, inhabits a linguistic community in which there is no practice of commanding. One can successfully make assertions to Mary, and among these assertions can be assertions about one's own psychological states, but one

cannot successfully command Mary to do anything. Here is the question: so long as Mary's linguistic resources are confined to those afforded by this practice, can God impose obligations on her? The defender of (1) will have to say No: for Mary cannot be commanded to do anything. The defender of (2) will have to say Yes: God could have an antecedent intention that Mary perform some action and (to sidestep worries about being under an obligation that one cannot know about) could inform Mary that God has that intention with respect to her conduct.

The debate between defenders of (1) and defenders of (2) is ongoing, and at present far from conclusive.

Option (3).

The other formulation of theological voluntarism that we noted is that in which the act of divine will is that of willing that the state of affairs that it is obligatory for *A* to $\wp$ obtain. Unlike the formulations (1) and (2), which admit of various sorts of dependence relationship between the act of divine will and the obligation, (3) is limited to something like a causal picture. (It obviously could not be that *its being obligatory for A to* $\wp$ is identical to *God's willing that it be obligatory for A to* $\wp$, on pain of a vicious regress.) The idea expressed here is that ultimately all obligations are present because of efficacious acts of the divine will, in particular, acts of willing that those obligations be in force.

This account is compatible with (1) and (2), because it could be that the way that God makes it the case that an act is obligatory is necessarily through the giving of commands (as in (1)) or through antecedently intending (as in (2)) the performance of that action. So it is not really, in its most general form, a *competitor* to (1) and (2). It could be made a competitor by adding claims about the way that the divine will brings it about that these obligations obtain. A defender of (3) might add that on his or her view the divine intention that *A* be obligated to $\wp$ is the immediate, total, and exclusive cause of its being obligatory for *A* to $\wp$ (cf. Quinn 1999, p. 55). If so, then a divine command that *A* $\wp$ or a divine antecedent intention that *A* $\wp$ could not be partial or mediate causes of its being obligatory for *A* to $\wp$. But note that even thus strengthened (3) is compatible with (1) or (2) understood as an identity claim: if the claim is that obligations just are divine commands or divine intentions, then the compatibility with (1) and (2) is reestablished.

Even if it turns out that (3) is not an obvious competitor with (1) and (2), it is still worth asking whether it is true. Quinn no longer is concerned to defend it, but he once argued for it in terms of an argument from divine sovereignty: because every state of affairs that obtains that does not involve God's existing depends on God's will, a fortiori every normative state of affairs that does not involve God's existing depends on God's will. (Quinn thought that *its being obligatory to obey God* is a state of affairs that involves God [Quinn 1990, pp. 298-299], but for a reason to reject that claim see Murphy 1998, pp. 12-13.) It does not seem, though, that this argument would support (3) in the strengthened version that holds that the dependence must be immediate, total, and exclusive. After all, very few folks want to say that every state of affairs that is brought about by God's will is brought about by God's will exclusively, totally, and immediately. While it is plausibly part of theism that every state of affairs that obtains, apart from those that involve God, is somehow dependent on God's will, this does not show that deontic states

of affairs are more interestingly connected to the divine will than states of affairs involving mathematics, or physics, or accounting (Murphy 1998, pp. 14-16).

We may put (3) to the side, then. While some formulations of it may very well be true, those formulations for which there is argumentative support do not establish much in the way of interesting metaethical conclusions. The debate concerning whether (1) and (2) is the more adequate formulation of theological voluntarism is ongoing, and we should thus proceed in a way that is as far as possible neutral between the two (though admittedly unwieldy) by saying that *A*s moral obligations to φ depend on God's commands/intentions that *A* φ. (By 'intentions' I will mean antecedent intentions.) Allowing for both of these possibilities, what can be said about the relationship of dependence holding between divine commands/intentions and moral obligations?

2.4 What sort of dependence?

The third issue that must be dealt with in providing a formulation of theological voluntarism is that of the specification of the dependence relationship that holds between divine commands/intentions that A φ and the moral obligation of *A* to φ. There have been several options considered whose nature and merits are worth discussing here. On an *analysis* view, it is part of the meaning of 'it is morally obligatory for *A* to φ' that God commands/intends that *A* φ. On a *reduction* view, the state of affairs *its being obligatory for A to* φ is the state of affairs *God's commanding/intending that A* φ. On a *supervenience* view, *its being obligatory for A* to φ supervenes on *God's commanding/intending that A* φ On a causal view, necessarily, *its being obligatory for A* to φ is caused by *God's commanding/intending that A* φ, and necessarily, *God's commanding/intending that A* φ causes it to be obligatory for *A* to φ.

Causation

The causal view is defended by Quinn (1979, 1990, 1999), and in a particularly strong form: on Quinn's view, the causal connection between *God's antecedently intending that A* φ and *its being obligatory for A to* φ exhibits totality, exclusivity, activity, immediacy, and necessity.

There are at least three serious difficulties for the causation formulation. The first we may call the 'Humean worry'. Once we allow that *its being morally obligatory to* φ is distinct from *God's commanding/intending* φ*-ing*, there is the question of what reason we would have for thinking that *its being morally obligatory to* φ necessarily obtains if *God's commanding/intending* φ*-ing* obtains. And whatever answer the defender of this view offers, it must be consistent with the causation formulation of theological voluntarism. But it is unclear what would do the trick. One way to try to make this necessary connection is by holding that there is a prior moral obligation to obey God; and so, whenever God gives a command/has an intention that one perform some action, it follows that one is morally obligated to perform the action commanded/intended. But we cannot take this route, because if the causation formulation is correct, then *all* moral obligations are caused entirely by God's commanding/intending activity; there cannot be, then, this *prior* moral obligation to obey God that would serve as part of the *explanans* for the necessary connection between divine commands/intentions and moral obligations.

The causal view that is a version of (2) (that is, that *its being obligatory for A* to φ depends causally on *God's willing that A* φ) should be distinguished carefully from the causal view that is a version of (3) (that is, that *its being obligatory for A to* φ depends causally on *God's willing that it be obligatory for A to* φ). The causal formulation of (3) has at least some plausibility as a result of God's sovereignty and omnipotence -- though it is in the end unclear why we should move from the claim that God is the ultimate source of all being to the claim that, for all deontic states of affairs, God's willing that that deontic state of affairs obtain is the *immediate*, *total*, and *exclusive* cause of its obtaining. Most of us would not, after all, find intuitively compelling a move from the claim that God is ultimate source of all being to the claim that, for all *physical* states of affairs, God's willing that that physical state of affairs obtain is the immediate, total, and exclusive cause of its obtaining. The causal view as an instance of (2), though, seems to have even less in the way of argumentative support. Why would one think that God's intending that *A* φ is an immediate, total, and exclusive cause of a deontic state of affairs' obtaining? In the absence of some evidence for such a connection, it is hard to see why one would be attracted to this formulation of theological voluntarism.

The second worry about the causal formulation I will call the ‘lack of precedent worry’. Moral properties and states of affairs supervene on nonmoral properties and states of affairs. The intuitive idea is that there can be no differences in moral status without some difference in nonmoral status. The causal formulation satisfies the supervenience constraint -- the differences in nonmoral status concern God's commands/intentions -- but it does so in a way that is unprecedented and mysterious. When we look at the specific ways in which changes in nonmoral facts can make a difference to the moral facts that hold, there is a pretty limited number of intelligible relationships that can hold between these nonmoral facts and the moral facts that supervene on them. A nonmoral fact can be part of what constitutes a reason to perform an action. (That you promised to φ can be cited in explaining why you have a reason to φ; your promising to φ constitutes, at least in part, the reason that you have to φ.) It can be part of an enabling condition for that reason. (The existence of a social practice of promising can be cited in explaining why you have a reason to φ; the existence of that practice might explain why your promise has the reason-giving force that it has.) It can be cited as a defeater-defeater for a reason. (While the fact the promisee told you that you need not fulfill your promise to φ typically releases you from your promise to φ, the fact that you threatened to beat up the promisee if he or she did not tell you that you need not fulfill your promise invalidates that release, and can be cited in explaining why you have a reason to φ.) But while theological voluntarism holds that a fact -- the fact that God commands/intends that one φ -- explains why one has a reason (in this case, an obligation) to φ, the causal view holds that this fact falls into none of the familiar explanatory categories: it is not constitutive of the reason, it is not an enabling condition for the reason, it is not a defeater-defeater for the reason. The way that the fact is supposed to explain the reason is *merely causal*: it just brings the reason about, exclusively, totally, immediately. This is an entirely unfamiliar phenomenon: nowhere else do we encounter a merely causal connection between a nonmoral fact and a moral one. (The appeal to the very strangeness of divine causation itself is not sufficient to answer the objection. For there is an extra strangeness here: that the relationship between nonmoral and moral facts is in every case with which we are familiar a rational relationship, whereas on the causal formulation of theological voluntarism the relationship is merely causal. Creation *ex nihilo* does not constitute carte blanche to multiply strangenesses.)

The third worry is the ‘no authority worry’. Theological voluntarism can be defended on the basis of considerations proper to metaethics -- that, for example, theological voluntarism provides the best explanation for the impartiality of morals, or for its overridingness, or for its normativity, or for its content. But theological voluntarists have tended to argue that theological voluntarism has something specific to offer to theists. One of these benefits on offer is that theological voluntarism fits well with the centrality of the virtue of obedience in theistic thought and practice (Quinn 1992, p. 510; Adams 1973, pp. 99-103). God is a being who is *to be obeyed*, is someone who is a *practical authority* over us.

For one to be a practical authority over another is, at least, for one to have some sort of control over others' reasons for action. Whatever else practical authority is, it is the ability to make a difference with respect to someone's reasons to act. The control involved in practical authority is, however, of a specific sort: it is *constitutive* control. When a party is an authority over another, his or her dictates constitute, at least in part, reasons for action for that other. (One piece of evidence for this is that when we take *A* to be an authority over us, we will cite ‘*A* told us to’ as a reason for action.) But if God's commands to φ have merely causal power to bring about obligations to φ, then the resultant state of affairs that is the reason for action is *its being obligatory to* φ -- a state of affairs that need not be in any way constituted by God's issuing any commands. No version of theological voluntarism that is built simply around God's causal role in actualizing moral obligations implies that God is a practical authority. (See Murphy 2002, 4.3.)

Supervenience

Suppose that we continue to interpret supervenience intuitively as the no-difference-in-moral-properties-without-some-difference-in-nonmoral-properties thesis. We can see very quickly that the theological voluntarist has to say something more about the sort of supervenience he or she has in mind in order to present what is genuinely a theological voluntarist account of moral obligation. For suppose that one puts forward a view on which both of the following claims are true: the moral law does not depend on, nor is it identical with, God's commands; but God necessarily commands us to follow the moral law. While it is obvious is that this not a version of theological voluntarism at all -- moral obligation in no way depends on divine command -- it satisfies the intuitive description of what is involved in the supervenience of the moral on the nonmoral: for there could be, on this view, no differences in moral status without some difference in divine commands. So if one is to put forward a supervenience formulation of theological voluntarism, then one will have to either be a little bit more doctrinaire about the supervenience relationship, so that it will exclude the nonvoluntarist view just described, or one will have to say more than that moral obligations supervene on divine commands. For our purposes here, they come to the same thing: that there is something more to the supervenience formulation of theological voluntarism than the claim that there are no differences in agents' moral obligations without some differences in the divine commands that have been imposed on that agent.

What is called for here is, pretty obviously, just some particular relationship of ontological dependence. It will not be that of causation, for reasons we have already examined. But neither does the defender of the supervenience view want it to be the extreme dependence of moral obligations on divine commands affirmed by the reduction formulation, on which moral obligations *just are* divine commands. To avoid

collapse into the reduction formulation, it has to hold that moral obligations are distinct from divine commands. It can make this distinction in one of two ways. It could say that moral obligation is wholly distinct from divine command -- that is, that the state of affairs *its being morally obligatory* to φ is not constituted even in part by *God's commanding φ-ing*. Or it could say that moral obligation is only partially constituted by divine command -- that is, that the state of affairs *its being obligatory to φ*, while not identical with *God's commanding φ-ing*, includes the state of affairs *God's commanding φ-ing* (and some other state of affairs besides). Let us consider each of these possibilities in turn.

Suppose first that the defender of the supervenience view affirms that moral obligation is wholly distinct from divine command. If so, then all of the arguments that were raised against the causation formulation can be leveled against the supervenience view. The no authority issue will arise. Because the states of affairs *its being obligatory to φ* and *God's commanding φ-ing* will be distinct, the supervenience account lacks the resources to underwrite divine authority. For God is authoritative only if God's commands are themselves reasons for action, but if the states of affairs *its being obligatory to φ* and *God's commanding φ-ing* are distinct, then God's commands will not be themselves reasons for action on the adequately strengthened supervenience view. And if these commands are not themselves reasons for action, then God does not constitutively actualize reasons for action by His commands; and if God does not constitutively actualize reasons for action by His commands, then God is not authoritative. The lack of precedent issue will arise. For the adequately strengthened supervenience view cannot view obligations as constituted by divine commands, and no theological voluntarist worthy of the name will see God's commands as merely enablers or defeater-defeaters for obligations; and so the relationship between divine commands and moral obligations is bound to be unprecedented and mysterious. And the Humean issue will arise. For the causation view is, after all, just the adequately strengthened supervenience view plus the claim that the dependence relationship involved in a particular sort of causal dependence. So, understood as affirming a dependence relationship between wholly distinct moral obligations and divine commands, the supervenience view has all of the problems of the causation view.

So the only hope for the supervenience formulation is to hold that God's commands are proper parts of moral obligations: for if those commands are identical with moral obligations, then the supervenience view collapses into the reduction view, and if moral obligations are wholly distinct from divine commands, then the supervenience view fails for the reasons that the causation view fails. There are, however, serious difficulties for this partial constitution version: in particular, if one is committed to saying that *God's commanding φ-ing* partly constitutes *its being morally obligatory to φ*, it is hard to see what state of affairs the theological voluntarist would be tempted to say is also necessary for moral obligation to be fully constituted. Obviously this other state of affairs cannot be one that involves moral obligation, on pain of circularity. (So, the theological voluntarist cannot say that *its being morally obligatory to φ* just is the complex state of affairs consisting both of *God's commanding φ-ing* and *its being morally obligatory to do what God commands*.) Further, in order to remain faithful to the basic idea of the supervenience version, we would have to say that any state of affairs that is held to constitute *its being morally obligatory to φ* along with *God's commanding φ-ing* must be a state of affairs that is certain to obtain if *God's commanding φ-ing* obtains. Otherwise, it might be the case that *its being morally obligatory to φ* does not supervene on *God's commanding φ-ing*, for there would be two possible worlds, in both of which *God's commanding φ-ing* obtains, but in only one of which does *its*

being morally obligatory to φ obtain. This runs contrary to even the basic idea of the supervenience view, on which there are no differences in moral obligations without a difference in divine commands.

These limitations make it hard to imagine what a motivated version of this form of the supervenience view would look like. We have to imagine a view of the following form. It is nonnegotiable that the state of affairs *its being morally obligatory to* φ is partially constituted by *God's commanding* φ*-ing*. It is nonnegotiable that there is, apart from *God's commanding* φ*-ing*, at least one state of affairs S that partially constituted *its being morally obligatory to* φ. It is nonnegotiable that S either obtains necessarily or at the very least necessarily obtains if *God's commanding* φ*-ing* obtains (otherwise moral obligation would not supervene on divine command). And it is nonnegotiable that S not involve moral obligation. The only remotely plausible candidates for S that come to mind are normative states of affairs that fall short of the obligatory, for example, φ*-ing's being good* (or *virtuous*, or *praiseworthy*). One might say, for example, that *its being morally obligatory to* φ is constituted jointly by *God's commanding* φ*-ing* and φ*-ing's being virtuous*. But it is unclear what motivation one would have for affirming such a position. It cannot be for the sake of making sure that God cannot impose a moral obligation to do something that is not virtuous: for, ex hypothesi, we know already that φ*-ing's being virtuous* obtains whenever *God's commanding* φ*-ing* obtains, for otherwise *its being morally obligatory to* φ would not supervene on *God's commanding* φ*-ing*.

The difficulty that faces the defender of the supervenience view can be framed as a dilemma. If the defender of that view holds that the state of affairs *its being morally obligatory to* φ is wholly distinct from *God's commanding* φ*-ing*, then he or she is refuted by the considerations that refute the causation view. If, on the other hand, the defender of the supervenience view holds that the state of affairs *its being morally obligatory to* φ is partially but not wholly constituted by *God's commanding* φ*-ing*, then there is pressure to explain why he or she does not simply affirm the reduction view, on which *its being morally obligatory to* φ just is *God's commanding* φ*-ing*. Unless the defender of the supervenience view identifies the state of affairs that, in addition to *God's commanding* φ*-ing*, makes for a moral obligation to φ, then his or her unwillingness to adopt the reduction view will look unmotivated and arbitrary.

Analysis

According to the analysis view, defended in Adams 1973, the concept of the morally obligatory is to be analyzed as that of being commanded by a loving God. Adams did not put this view forward as an account of the meaning of 'obligation' generally, but only of its meaning as employed in Judeo-Christian moral discourse. As evidence for this analysis, Adams appealed to the freedom with which users of that discourse moved between claims of the form '*x* is obligatory' and '*x* is God's will' or '*x* is God's command.'

There are a couple of central difficulties for this position. The first is that it seems to imply that those inside and those outside the Judeo-Christian practice of moral discourse have never disagreed when one has affirmed a claim of the form ' φ-ing is obligatory' and the other denied a claim of that form. For they do not, on Adams' account, mean the same thing when they use these terms. In atheistic moral

discourse, a masterful user of the language can say ‘it is not true that God has commanded φ-ing, but φ-ing is nonetheless obligatory’; in Judeo-Christian moral discourse, on Adams' view, one shows oneself to be either unintelligible or not a masterful user of moral language if one were to speak thus. Adams was aware of this difficulty, and attempted to mitigate it: he argued that the agreement over which items the term ‘obligatory’ applied to, and the appropriate attitudinal and volitional responses to those things correctly described as ‘obligatory,’ made possible substantive moral discourse (Adams 1973, pp. 116-120). But all this seems to do is to explain how a simulacrum of genuine moral discourse is preserved; it does not show that what we get is the real thing.

The second difficulty is that of dealing with those within the Judeo-Christian tradition of moral discourse who employed or continue to employ moral language in a way that is out of step with Adams' analysis. Now, it is not sufficient to refute a suggested analysis of some term that users of that term have questioned or even rejected that analysis. But if we take the task of analyzing terms to be that of making explicit and systematizing the platitudes employing that term affirmed by masterful users of that term (Smith 1994, pp. 29-32), and we note that many thoughtful Jews and Christians who otherwise appear to be masterful users of the language of moral obligation have rejected, either explicitly or implicitly, the notion that an act is obligatory if and only if it has been commanded by God, then we would have some reason to doubt whether the analysis formulation of theological voluntarism is defensible.

Reduction

Adams' maneuver in the face of these difficulties was to move from the analysis to the reduction version of theological voluntarism. He decided that the meaning of the term ‘morally obligatory’ was common to theists and nontheists. There is a common concept of the morally obligatory, a common concept that makes possible substantive agreement and disagreement between theists and nontheists. This common concept is neutral between theism and nontheism. But, following the now standard Kripke-Putnam line, Adams affirms that there are necessary a posteriori truths, among which are included property identifications. He argued that the property *being wrong* is identical to the property *being contrary to the commands of (a loving) God* because the property *being contrary to the commands of (a loving) God* best fills the role assigned by the concept of wrongness (Adams 1979a, pp. 133-142; see also Adams 1999, pp. 252-258). By conceptual analysis alone we can know only that wrongness is a property of actions (and perhaps intentions and attitudes); that people are generally opposed to what they regard as wrong; that wrongness is a reason, perhaps a conclusive reason, for opposing an act; and that there are certain acts (e.g. torture for fun) that are wrong. But given traditional theistic beliefs, the best candidate property to fill the role set by the concept of wrongness is that of being contrary to (a loving) God's commands. For that property is an objective property of actions. Further, given Christian views about the content of God's commands, this identification fits well with widespread pre-theoretical intuitions about wrongness; and given Christian views about human receptivity to divine communication and God's willingness to communicate both naturally and supernaturally, God's commands have a causal role in our acquisition of moral knowledge (Adams 1979, p. 139; see also Adams 1999, pp. 257).

The reduction formulation avoids the most troublesome implications of the analysis formulation, for it allows that there is a common concept of obligation, so that those within the Judeo-Christian tradition

and those outside it can engage in moral debate and can have substantive agreements and disagreements with each other, and so that those within the Judeo-Christian tradition can raise substantive questions about the relationship between God and obligation without ipso facto excluding themselves from the class of masterful users of the moral concepts of that community. The reduction formulation allows that the concept of obligation may be nontheistic while the property that best fills the role assigned to it by that concept is a theistic one.

Analysis vs. reduction.

Nevertheless, it remains an open question whether the reduction view is superior to the analysis view. One might argue that Adams' analogy to ‘H_2O is water’ is inappropriate, as the identification with water with H_2O is clearly a posteriori, whereas the identification of the morally obligatory with the commanded by God is a priori. For if Adams is right in his characterization of the concept of obligation, it is not as if those who do not have the ability to infer from ‘this is morally obligatory’ to ‘this is commanded by a loving God’ (and vice versa) are just missing out on an interesting extra fact, the way that those without rudimentary chemistry are missing out on an interesting extra fact if they do not know that water is H_2O. The term ‘water’ can play its role in our practical lives perfectly well without our knowing that it is H_2O. The term ‘morally obligatory’ cannot play its role in our practical lives without our knowing that the morally obligatory is the commanded by God. No unintelligibility creeps into the life of agents that do not grasp that water is H_2O; unintelligibility creeps into the life of agents that do not grasp that the morally obligatory is what is commanded by God.

Why might one think that the masterful use of ‘morally obligatory’ requires recognition that the morally obligatory is what is commanded by God? If Adams is right, it is part of the meaning of obligation that obligations are social in character (Adams 1999, p. 233) and involve actually made demands by one party in the social relationship on another (Adams 1999, pp. 245-246). It is the fact that a demand is actually made that gives sense to the notion that one has to perform an action, rather than merely that it would be good, even the best, to do it (Adams 1999, p. 246). But if it is part of the meaning of ‘morally obligatory’ that one is part of a certain social relationship in which demands are actually made, then it is no longer just an interesting further fact that the property that best answers to the concept ‘morally obligatory’ is the property *being commanded by God*. Rather, one who denies that there is a God or that God actually makes demands on human beings must fail to use the term ‘morally obligatory’ masterfully. For think of the other marks of the moral, especially those of impartiality and overridingness. For one to think of an act as obligatory is for one to think of it as being actually imposed on one as a demand; for every obligation, on Adams' view, there is someone who imposes that obligation by commanding. It is clear a priori that the only being that could impose the sort of obligation that could plausibly be classified as moral would be God. How, then, could one be a masterful user of ‘moral obligation’ without grasping that moral obligations are demands imposed by God?

This analysis view would not, unlike Adams' earlier formulation, require the subdivision of linguistic communities. One could say that the meaning of ‘morally obligatory’ includes ‘being commanded by God,’ for both theists *and* nontheists. Those who do not grasp that it is of the essence of obligations to be

divinely commanded -- whether theists or nontheists -- fail to be masterful users of the language of moral obligation. To embrace this view is to return to the position of Anscombe 1958, according to which we should hold that the concept of obligation is inherently theological. On this view, we should not allow that Judeo-Christian moral practice has a different concept of obligation. Rather, the theological understanding of obligation is the authentic one, and nontheological concepts of obligation are unintelligible truncations.

3. Perennial difficulties for metaethical theological voluntarism

Apart from the difficulties that must be handled by particular formulations of theological voluntarism, there are a number of objections that have been levelled against theological voluntarist views as such. A wide variety of these objections are helpfully discussed in Quinn 1999 (pp. 65-71). Here I will consider only two objections, but they are the two that are characteristically taken to be the most powerful perennial objections to theological voluntarism: first, that theological voluntarism is incompatible with any substantive sense in which God is good; second, that theological voluntarism entails the arbitrariness of morality. While these objections have been answered plausibly in recent formulations of theological voluntarism, the way that these objections have been answered leave theological voluntarists open to a different objection: that theological voluntarism is not adequately motivated as a philosophical position. I conclude with a brief discussion of this worry.

3.1 Theological voluntarism and God's goodness

God is, by definition, good. This is both a fixed point concerning God's nature and a plausibility-making feature of theological voluntarism. If one were to deny that God is good (understood *de dicto* -- that is, ‘if there is a being that qualifies as God, then that being is good’), one would call one's own competence in use of the term ‘God’ into question. And even if it were allowed that one can employ the term ‘God’ masterfully while denying that God is good, if one were to deny that God is good, then one would undercut one's capacity to defend theological voluntarism. For theological voluntarism is plausible only if God is an exalted being; but a being that is not good is not an exalted being.

That God is good is a fixed point for theistic discourse in general and for theological voluntarism in particular provides the basis for a common objection to theological voluntarism: that theological voluntarism makes it impossible to say that, in any substantive sense, God is good. The most straightforward formulation of the objection is as follows. For God to be good is for God to be morally good. But if moral goodness is to be understood in theological voluntarist terms, then God's goodness consists only in God's measuring up to a standard that God has set for Himself. While this is perhaps an admirable resoluteness -- it is, other things being equal, a good thing to live up to your own standards -- it is hardly the sort of thing that provokes in us the admiration that God's goodness is supposed to provoke.

Now, one might dispute the claim that if God's goodness consists simply in God's living up to a standard

that God has set for Himself, then that goodness is far less admirable than we would have supposed. (See, for a nice discussion of this issue, Clark 1982, esp. pp. 341-343.) Suppose, though, that we grant this part of the argument. How powerful is the objection from God's goodness against theological voluntarism?

As we noted earlier, theological voluntarism comes in a variety of strengths. One dimension along which a theological voluntarist view might be assessed as stronger or weaker is in terms of the range of normative properties that it attempts to account for in theological voluntarist terms. The strength of the objection from God's goodness is directly proportional to the size of the range of normative properties that one wishes to explain in theological voluntarist terms (see also Alston 1990). If one wishes only to account for a proper subset of moral notions, such as obligation, with one's theological voluntarism, then the objection from God's goodness is very weak; if one wishes to provide a sweeping account of normativity in theological voluntarist terms, then the objection is much stronger.

Suppose, for example, that one defends a version of theological voluntarism that accounts only for obligation. If moral obligation only is dependent on acts of the divine will, one can appeal to moral notions other than deontic ones in order to provide a substantive sense in which God is good. Granting to some extent the force of the objection, we can say, on this view, that God's moral goodness cannot consist in God's adhering to what is morally obligatory. But there are other ways to assess God morally other than in terms of the morally obligatory. Adams, for example, holds that God should be understood as benevolent and as just, and indeed concedes that his theological voluntarist account of obligation as the divinely commanded is implausible unless God is thus understood (Adams 1999, pp. 253-255). The ascription to God of these moral virtues is entirely consistent with his theological voluntarism, for his theological voluntarism is not meant to provide any account of the moral virtues. One can hold that God's moral goodness involves supereminent possession of the virtues, at least insofar as those virtues do not presuppose weakness and vulnerability. God is good because God is supremely just, loyal, faithful, benevolent, and so forth. It seems that ascribing to God supereminent possession of these virtues would be enough to account for God's supreme moral goodness: it is, after all, in such terms that God is praised in the Psalms.

Matters become more difficult for theological voluntarist views that aim to provide accounts of all moral notions in terms of God's will. If one held to such an ambitious version of theological voluntarism -- if one were to hold, say, that a state of affairs is morally good because it is a state of affairs that God wishes to obtain for its own sake, and that a character trait is a moral virtue because it is a property that God wants one to have for its own sake, and that an action is morally obligatory because it is antecedently intended by God, and so forth -- then obviously the gambit employed by the less ambitious theological voluntarist is unavailable. The more ambitious theological voluntarist should hold, instead, that God's goodness is not to be understood in moral terms. God's being good might be understood in terms of God's being *good to us*, where us includes all created rational beings, or all created sentient beings, or whatever class of created beings to which one thinks that God has a special relationship. What it is for God to be good to us would be for God to be loving -- to will each of our goods, and to do so in a way that plays no favorites. This understanding of ‘loving’ does not run afoul of theological voluntarism construed as an account of all moral goods, because ‘our goods’ is to be interpreted in terms of prudential goodness, what

makes each of us well-off (but see Chandler 1985).

Suppose, though, that one were to go all the way, holding that theological voluntarism is the correct account of all normative notions: on this extremely ambitious view, anything that is intrinsically action-guiding depends on God's will. I think that the ‘God is good to us’ understanding of God's goodness is ruled out on this approach: for the notion of ‘good to us’ is a normative notion. Perhaps one could hold, on this view, that ‘God is good’ affirms of God some sort of metaphysical goodness, fullness of being. This surely makes God exalted, but it is not clear whether the will of such a being is plausibly understood as the source of all normative statuses. It is also less than clear that ‘God is good,’ on this reading, is the claim that God possesses a particular perfection, rather than is merely a reminder that God has a variety of perfections.

To sum up, then: for each of the various formulations of theological voluntarism, there seems to be some way of answering the charge that the view undercuts the notion that God is good. But the strain needed to answer the charge becomes greater the wider the range of normative properties that the formulation of theological voluntarism aims to explain.

3.2 Theological voluntarism and arbitrariness

It is also an extraordinarily popular charge against theological voluntarism that it entails, objectionably, that morality is arbitrary. There is, however, more than one objection here, and the different objections need to be distinguished and answered individually. One claim is that theological voluntarism implies that God's commands/intentions, on which moral statuses depend, must be arbitrary. A distinct claim is that theological voluntarism implies that the content of morality is itself arbitrary, that it is of the essence of morality to exhibit a certain rational structure, and that theological voluntarism precludes its having that structure. I will consider each of these objections in turn.

One arbitrariness objection against theological voluntarism is that if theological voluntarism is true, then God's commands/intentions must be arbitrary; and it cannot be that morality could wholly depend on something arbitrary; and so theological voluntarism must be false. In favor of the claim that if theological voluntarism were true, then morality would be arbitrary: morality would be arbitrary, on theological voluntarism, if God lacks reasons for the commands/intentions that God gives/has; but because theological voluntarism holds that reasons depend on God's commands/intentions, it is clear that there could ultimately be no reason for God's commanding/intending one thing rather than another. In favor of the claim that morality could not wholly depend on something arbitrary: when we say that some moral state of affairs obtains, we take it that there is a reason for that moral state of affairs obtaining rather than another. Moral states of affairs do not just happen to obtain.

Just as in the case of the objection from God's goodness, the strength of this version of the objection from arbitrariness depends on the formulation of theological voluntarism that is being attacked. The arbitrariness objection becomes more difficult to answer the stronger the relationship between God's intentions/commands and moral properties is held to be; and it becomes more difficult to answer the

more normative properties one attempts to account for by appeal to God's intentions/commands.

The arbitrariness objection has less force if one holds that, say, only moral obligations are to be accounted for by theological voluntarism. The claim made by the objector is that morality is arbitrary on theological voluntarism, because God has no reason for having one set of commands/intentions rather than another. But this is so only if one appeals to the very strong form of theological voluntarism on which all normative states of affairs depend on God's will. If one holds that only moral obligations are determined by God's will, then God might have moral reasons for selecting one set of commands/intentions rather than another: that, for example, one set of commands/intentions is more benevolent, or just, or loyal, than another. If one holds that all moral properties are determined by God's will, then God might have nonmoral reasons for selecting one set of commands/intentions rather than another: that one set of commands/intentions is more loving than another. The fewer normative properties that a version of theological voluntarism attempts to account for, the less susceptible it is to the claim that theological voluntarism implies the arbitrariness of God's commands/intentions.

Now, one might respond on behalf of this version of the arbitrariness objection that even if it is true that there can be reasons for God to choose the commands/intentions that God chooses, it is unlikely that these reasons would wholly determine God's choice of commands/intentions, and so there would be some latitude for arbitrariness in God's choices/intentions. But of itself this is not much of a worry. The initial claim pressed against theological voluntarism was that it made all of God's commands/intentions ultimately arbitrary, and morality could not depend on something so thoroughly arbitrary. But the chastened claim -- that there is some arbitrariness in God's commands -- is far less troubling on its own. We are already familiar with morality depending to some extent on arbitrary facts about the world: if one thinks about the particular requirements that he or she is under, one will note straightaway the extent to which these requirements have resulted from contingent and indeed fluky facts about oneself, one's relationships, and one's circumstances. It does not seem that allowing that God has some choices to make concerning what to command/intend with respect to the conduct of created rational beings that are undetermined by reasons must introduce an intolerable arbitrariness into the total set of divine commands/intentions.

Allowing for such pockets of divine discretion does not provide backing for this version of the objection from arbitariness, but rather offers a premise for the other version of the objection from arbitrariness. This other version of the objection from arbitrariness holds that moral states of affairs exhibit a certain rational structure that they would not have if theological voluntarism were true. Here is the idea, roughly formulated. Suppose that some moral state of affairs obtains -- that it is the case that murder is wrong, or that lying is objectionable, or that courage is a virtue, or that Sharon's snubbing me in that way was unforgivable. The idea is that for any such moral state of affairs, the following is true: either we can provide a *justification* for the obtaining of that moral state of affairs, or that moral state of affairs is *necessary*. A justification of an obtaining moral state of affairs *A* is some obtaining moral state of affairs *B* (where *A* is not identical with *B*), which in conjunction with the other non-moral facts entails that *A* obtains. So, for example: it may be the justification for *murder's being prima facie wrong* that murder is an intentional harm (non-moral fact) and intentionally harming is prima facie wrong (obtaining moral state of affairs). Now, presumably not all moral states of affairs can be justified: eventually there will be

basic moral states of affairs, for which no justification can be given. But it would be very unsatisfactory to say that these basic moral states of affairs just happen to obtain. So any basic moral states of affairs must obtain necessarily. Perhaps *unrelieved suffering's being bad* is a state of affairs of this sort, or perhaps *rational beings'being worthy of respect*.

The claim that the structure of morality is not arbitrary is, put positively, the claim that every obtaining moral state of affairs either has a justification or is necessary. And, thus, what those who claim that theological voluntarism entails that morality is objectionably arbitrary mean is that if theological voluntarism is true, then there are some moral states of affairs that both lack a justification and are not necessary. The view that God's commands/intentions are not wholly determined by reasons offer our basis for holding that there are some moral states of affairs that both lack a justification and are not necessary. For consider some act of φ-ing that is not subsumed under any other issued divine command and which is such that God lacks decisive reasons to command or not to command its performance. In the possible world in which God issues a command to φ, there is a moral state of affairs -- *its being obligatory to* φ -- which lacks a justification (for the action is subsumed under no other divine command) and is not necessary (for God might have failed to command the action).

Some theological voluntarists have responded to this sort of worry by claiming that what God commands/intends with respect to human action, God does so necessarily. But this seems either to understate the divine freedom or to overstate the determination of God's commands by reasons. More plausible is the denial of the claim that morality must exhibit the particular structure presupposed in the objection. While I think that in general the subsumption model of justification is innocuous enough -- even particularists can affirm it, if they affirm even the most minimal doctrine of moral supervenience -- its appeal to necessary moral states of affairs as the only proper starting point is dubious. It is not clear why the starting points for justification have to be necessary moral states of affairs. Surely if these moral states of affairs are basic, their moral status must not be explained by appeal to other moral states of affairs, but that does not mean they must be necessary; they might be contingent, and have their moral status explained in some way other than an appeal to another moral state of affairs. It could be, for example, that the explanation of them appeals to a contingent nonmoral state of affairs plus some necessary state of affairs concerning a connection between that contingent nonmoral state of affairs and the moral state of affairs. Theological voluntarism would be an instance of this latter model.

3.3 Is theological voluntarism adequately motivated?

Both with respect to the objection from God's goodness and with respect to the objection from arbitrariness, the now-standard theological voluntarist response is not to bite the bullet but rather to restrict the range of normative properties of which theological voluntarism is supposed to provide an account. So Adams, Quinn, and Alston all recommend theological voluntarism only as a theory about properties like *being morally obligatory*, and not about any other normative properties. The worry is that allowing that there is adequate motivation to refuse to understand these other normative properties in theological voluntarist terms might commit one to holding that there is adequate motivation to refuse to understand obligation-type properties in theological voluntarist terms. Look, one might say: if you are

willing to hold that all moral properties other than those in the obligation family are to be understood in non-theological voluntarist terms, what is to stop us from holding that obligation is to be understood in non-theological voluntarist terms as well? If we are willing to give up theological voluntarism in some moral domains, why not in all of them?

The most well-developed account of why we should treat obligation as special is Adams', on which obligation is apt for theological voluntarist treatment because of its intrinsic link to demands made within social relationships (Adams 1987b, and Adams 1999, pp. 231-258). But it is also unclear whether this is persuasive. We may grant that obligations result from demands, but only if we emphasize (as Adams does) that it is demands from *authorities* that result in obligations. But what makes someone an authority is that by his or her dictates he or she can give reasons for action of a certain kind. There is some dispute over what kind of reasons they must be, but for our purposes we can just follow Joseph Raz, who holds that genuine authorities give “protected reasons” by their dictates, where a protected reason to φ is a reason to φ and a reason to disregard some reasons against φ-ing. If one is an authority over another with respect to φ-ing, then one's dictate that the other φ is a protected reason for the other to φ (Raz 1979, p. 18).

But now here is the question. We agree that obligations arise from authoritative dictates. And we agree that for a dictate to be authoritative is for it to constitute a certain sort of reason, let us say a protected one. But why, then, would we identify obligations with protected reasons that result from demands, rather than (as Raz does) with protected reasons themselves, whatever their source? If, after all, there were other ways of producing protected reasons other than through the giving of commands, what would be the point of saying ‘oh, but though there is a protected reason to φ, it isn't really *obligatory* to φ.’ Surely if there were any point to this remark it would be purely verbal, and of no philosophical / normative interest.

It turns out, then, that whether Adams' move is enough to motivate theological voluntarism about obligation is dependent on whether there are in fact any protected reasons (or reasons of whatever structure that one thinks that authoritative dictates must give) that are not dependent on demands being made. If there are such reasons -- which natural law theorists, for example, would hold -- then Adams' gambit will not work, and the theological voluntarist will have to look elsewhere for motivation to understand obligation in those terms.

Bibliography

- Adams, Robert M. 1973. “A Modified Divine Command Conception of Ethical Wrongness.” Reprinted in Adams 1987a, pp. 97-122.
- Adams, Robert M. 1979a. “Divine Command Metaethics Modified Again.” Reprinted in Adams 1987a, pp. 128-143.
- Adams, Robert M. 1979b. “Moral Arguments for Theistic Belief.” Reprinted in Adams 1987a, pp. 144-163.
- Adams, Robert M. 1987a. *The Virtue of Faith and Other Essays in Philosophical Theology*.

Oxford.
- Adams, Robert M. 1987b. “Divine Commands and the Social Nature of Obligation.” *Faith and Philosophy* 4, pp. 262-275.
- Adams, Robert M. 1999. *Finite and Infinite Goods: A Framework for Ethics*. Oxford.
- Alston, William P. 1990. “Some Suggestions for Divine Command Theorists.” In Beaty 1990, pp. 303-326.
- Anscombe, G. E. M. 1958. “Modern Moral Philosophy.” *Philosophy* 33, pp. 1-19.
- Audi, Robert and William Wainwright. 1986. *Rationality, Religious Belief, and Moral Commitment.* Cornell.
- Beaty, Michael, ed. 1990. *Christian Theism and the Problems of Philosophy.* Notre Dame.
- Chandler, John. 1985. “Divine Command Theories and the Appeal to Love.” *American Philosophical Quarterly* 22, pp. 231-239.
- Clark, Stephen R. L. 1982. “God's Law and Morality.” *Philosophical Quarterly* 32, pp. 339-347.
- Firth, Roderick. 1952. “Ethical Absolutism and the Ideal Observer.” *Philosophy and Phenomenological Research* 12, pp. 317-345.
- Foot, Philippa. 1958. “Moral Arguments.” *Mind* 67, pp. 502-513.
- Idziak, Janine Marie. 1979. *Divine Command Morality*. Edwin Mellen.
- LaFollette, Hugh, ed. 1999. *Guide to Ethical Theory*. Blackwell.
- Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong.* Penguin.
- Mavrodes, George. 1986. “Religion and the Queerness of Morality.” In Audi and Wainwright 1986, pp. 213-226.
- Moore, G. E. 1903. *Principia Ethica.* Cambridge.
- Murphy, Mark C. 1998. “Divine Command, Divine Will, and Moral Obligation.” *Faith and Philosophy* 15, pp. 3-27.
- Murphy, Mark C. 2002. *An Essay on Divine Authority*. Cornell.
- Quinn, Philip. 1978. *Divine Commands and Moral Requirements.* Oxford.
- Quinn, Philip. 1979. “Divine Command Ethics: A Causal Theory.” In Idziak 1979, pp. 305-325.
- Quinn, Philip. 1990. “An Argument for Divine Command Ethics.” In Beaty 1990, pp. 289-302.
- Quinn, Philip. 1992. “The Primacy of God's Will in Christian Ethics.” *Philosophical Perspectives* 6, pp. 493-513.
- Quinn, Philip. 1999. “Divine Command Theory.” In LaFollette 1999, pp. 53-73.
- Rawls, John. 1971. *A Theory of Justice*. Harvard.
- Raz, Joseph. 1979. *The Authority of Law*. Oxford.
- Sidgwick, Henry. 1907 [1981]. *The Methods of Ethics*, 7th edition. Hackett.
- Smith, Adam. 1759. *Theory of Moral Sentiments.*
- Smith, Michael. 1994. *The Moral Problem*. Blackwell.
- Wierenga, Edward. 1983. “A Defensible Divine Command Theory.” *Nous* 17, pp. 387-407.

Other Internet Resources

- Religion and Ethics page, at the *Ethics Updates* site, maintained by Lawrence Hinman (University of San Diego)

Related Entries

divine command | metaethics | obligation | religion: and morality | theism

Mark C. Murphy
Georgetown University
murphym@georgetown.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 1, 2002
Content last modified: July 1, 2002

Stanford Encyclopedia of Philosophy

Distributive Justice

Principles of distributive justice are normative principles designed to allocate goods in limited supply relative to demand. The principles vary in numerous dimensions. They vary in what goods are subject to distribution (income, wealth, opportunities, etc.); on the nature of the subjects of the distribution (natural persons, groups of persons, reference classes, etc.); and on what basis the goods should be distributed (equality, according to individual characteristics, according to free market transactions, etc.).

This entry will focus on principles of distributive justice designed to cover the distribution of material goods and services to individuals. Principles of this kind have been the dominant source of Anglo-American debate on distributive justice over the last three decades.

---

1. Strict Egalitarianism

One of the simplest principles of distributive justice is that of strict or radical equality. The principle says that every person should have the same level of material goods and services. The principle is most commonly justified on the grounds that people are owed equal respect and that equality in material goods and services is the best way to give effect to this ideal.

Even with this ostensibly simple principle some of the difficult specification problems of distributive principles can be seen. The two main problems are the construction of appropriate indices for

measurement (the index problem), and the specification of time frames. Because there are numerous proposed solutions to these problems, the ‘principle of strict equality’ is not a single principle but a name for a group of closely related principles. This range of possible specifications occurs with all the common principles of distributive justice.

The index problem arises primarily because the goods to be distributed need to be measured if they are going to be distributed according to some pattern (such as equality). The strict equality principle stated above says that there should be ‘the same *level* of material goods and services’. The problem is how to specify and measure levels. One way of solving the index problem in the strict equality case is to specify that everyone should have the same *bundle* of material goods and services rather than the same *level* (so everyone would have 4 oranges, 6 apples, 1 bike, etc.). The main problem with this solution is that there will be many other allocations of material goods and services which will make some people better off without making anybody else worse off. For instance, a person preferring apples to oranges will be better off if she swaps some of the oranges from her bundle for some of the apples belonging to a person preferring oranges to apples. Indeed, it is likely that everybody will have something they would wish to trade in order to make themselves better off. As a consequence, requiring identical bundles will make virtually everybody materially worse off than they would be under an alternative allocation. So specifying that everybody must have the same *bundle* of goods does not seem to be a satisfactory way of solving the index problem. Some index for measuring the value of goods and services is required.

Money is an index for the value of material goods and services. It is an imperfect index and its pitfalls are well-documented in most economics textbooks. Moreover, once the goods to be allocated are extended beyond material ones to include opportunities, etc. it needs to be combined with other indices. (For instance, John Rawls' index of primary goods - see Rawls 1971.) Nevertheless, using money as index for the value of material goods and services is the most practical response so far suggested to the index problem and is widely used in the specification and implementation of distributive principles.

The second main specification problem involves time frames. Many distributive principles identify and require that a particular pattern of distribution be achieved. But they also need to specify *when* the pattern is required. One version of the principle of strict equality requires that all people should have the same wealth at some initial point, after which people are free to use their wealth in whatever way they choose. Principles specifying initial distributions after which the pattern need not be preserved are commonly called ‘starting-gate’ principles. (See Ackerman 1980, 53-59,168-170,180-186)

Because ‘starting-gate’ forms of the strict equality principle may lead in time to very inegalitarian wealth distributions they are not common. The most common form of strict equality principle specifies that *income* (measured in terms of money) should be equal in *each* time-frame, though even this may lead to significant disparities in wealth if variations in savings are permitted. Hence, strict equality principles are commonly conjoined with some society-wide specification of just saving behavior.

There are a number of direct moral criticisms made of strict equality principles: that they unduly restrict freedom, that they do not give best effect to equal respect for persons, that they conflict with what people

deserve, etc. (see Desert-Based Principles) But the most common criticism is a welfare-based one: that everyone can be materially better off if incomes are not strictly equal. (see Carens) It is this fact which partly inspired the Difference Principle.

2. The Difference Principle

The wealth of an economy is not a fixed amount from one period to the next. More wealth can be produced and indeed this has been the experience of industrialized countries over the last few centuries. The most common way of producing more wealth is to have a system where those who are more productive earn greater incomes. This partly inspired the formulation of the Difference Principle.

The most widely discussed theory of distributive justice in the past three decades has been that proposed by John Rawls in *A Theory of Justice,* (Rawls 1971), and *Political Liberalism,* (Rawls 1993). Rawls proposes the following two principles of justice:

> 1. Each person has an equal claim to a fully adequate scheme of equal basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value.
>
> 2. Social and economic inequalities are to satisfy two conditions: (a) They are to be attached to positions and offices open to all under conditions of fair equality of opportunity; and (b), they are to be to the greatest benefit of the least advantaged members of society. (Rawls 1993, pp. 5-6. The principles are numbered as they were in Rawls' original *A Theory of Justice.*)

Under Rawls' proposed system Principle (1) has priority over Principle (2). In addition to (2b) it is possible to think of Principles (1) and (2a) as principles of distributive justice: (1) to govern the distribution of liberties, and (2a) the distribution of opportunities. Looking at the principles of justice in this way makes all principles of justice, principles of distributive justice (even principles of retributive justice will be included on the basis that they distribute negative goods). Keeping in line with the primary focus of this entry though, let us concentrate on (2b), known as the Difference Principle.

The main moral motivation for the Difference Principle is similar to that for strict equality: equal respect for persons. Indeed the Difference Principle materially collapses to a form of strict equality under empirical conditions where differences in income have no effect on the work incentive of people. The overwhelming opinion though is that in the foreseeable future the possibility of earning greater income will bring forth greater productive effort. This will increase the total wealth of the economy and, under the Difference Principle, the wealth of the least advantaged. Opinion divides on the size of the inequalities which would, as a matter of empirical fact, be allowed by the Difference Principle, and on how much better off the least advantaged would be under the Difference Principle than under a strict equality principle. Rawls' principle however gives fairly clear guidance on what type of arguments will

count as justifications for inequality. Rawls is not opposed to the principle of strict equality *per se*, his concern is about the *absolute* position of the least advantaged group rather than their *relative* position. If a system of strict equality maximizes the absolute position of the least advantaged in society, then the Difference Principle advocates strict equality. If it is possible to raise the absolute position of the least advantaged further by having some inequalities of income and wealth, then the Difference Principle prescribes inequality up to that point where the absolute position of the least advantaged can no longer be raised.

Because there has been such extensive discussion of the Difference Principle in the last 30 years, there have been numerous criticisms of it from the perspective of all the other theories of distributive justice outlined here. Briefly, the main criticisms are as follows.

Advocates of strict equality argue that inequalities permitted by the Difference Principle are unacceptable even if they do benefit the least advantaged. The problem for these advocates is to explain in a satisfactory way why the relative position of the least advantaged is more important than their absolute position, and hence why society should be prevented from materially benefiting the least advantaged when this is possible. The most common explanation appeals to solidarity (Crocker): that being materially equal is an important expression of the equality of persons. Another common explanation appeals to the power some may have over others, if they are better off materially. Rawls' response to this latter criticism appeals to the priority of his first principle: The inequalities consistent with the Difference Principle are only permitted so long as they do not compromise the fair value of the political liberties. So, for instance, very large wealth differentials may make it practically impossible for poor people to be elected to political office or to have their political views represented. These inequalities of wealth, even if they increase the material position of the least advantaged group, may need to be reduced in order for the first principle to be implemented.

The Utilitarian objection to the Difference Principle is that it does not maximize utility. In *A Theory of Justice*, Rawls uses Utilitarianism as the main theory for comparison with his own, and hence he responds at length to this Utilitarian objection and argues for his own theory in preference to Utilitarianism (some of these arguments are outlined in the section on Welfare-Based Principles).

Libertarians object that the Difference Principle involves unacceptable infringements on liberty. For instance, the Difference Principle may require redistributive taxation to the poor, and Libertarians commonly object that such taxation involves the immoral taking of just holdings. (see Libertarian Principles)

The Difference Principle is also criticized as a primary distributive principle on the grounds that it mostly ignores claims that people *deserve* certain economic benefits in light of their actions. Advocates of Desert-Based Principles argue that some may deserve a higher level of material goods because of their hard work or contributions even if their unequal rewards do not also function to improve the position of the least advantaged. They also argue that the explanations of *how* people come to be in more or less advantaged positions is relevant to their fairness, yet the Difference Principle wrongly ignores these

explanations.

Like Desert theorists, advocates of Resource-Based Principles criticize the Difference Principle on the grounds that it is not ‘ambition-sensitive’ enough, i.e. it is not sensitive to the consequences of people's choices. They also argue that it is not adequately ‘endowment-sensitive’: it does not compensate people for natural inequalities (like handicaps or ill-health) over which people have no control.

3. Resource-Based Principles

Resource-based principles (also called Resource Egalitarianism) prescribe equality of resources. Interestingly, resource-based principles do not normally prescribe a patterned outcome - the idea being that the outcomes are determined by people's free use of their resources. Resource-theorists claim that the Difference Principle is insufficiently ‘ambition-sensitive’ and that provided people have equal resources they should live with the consequences of their choices. They argue, for instance, that people who choose to work hard to earn more income should not be required to subsidize those choosing more leisure and hence less income.

Resource-theorists also make a related complaint that the Difference Principle is not sufficiently ‘endowment-sensitive.’ While part of Rawls' motivation for the Difference principle is that people have unequal endowments, resource-theorists explicitly emphasize this feature of their theory though they differ on which endowments are relevant to questions of distributive justice. They agree that, ideally, social circumstances over which people have no control should not adversely affect life prospects or earning capacities. Some resource-theorists further argue that, for the same sorts of reasons, unequal natural endowments should attract compensation. For instance, people born with handicaps, ill-health, or low levels of natural talents have not brought these circumstances upon themselves and hence, should not be disadvantaged in their life prospects.

The most prominent Resource-based theory, developed by Ronald Dworkin, (Dworkin 1981a, 1981b), proposes that people begin with equal resources but end up with unequal economic benefits as a result of their own choices. What constitutes a just material distribution is to be determined by the result of a thought experiment designed to model fair distribution. Suppose that everyone is given the same purchasing power and each use that purchasing power to bid, in a fair auction, for resources best suited to their life plans. They are then permitted to use those resources as they see fit. Although people may end up with different economic benefits, none of them is given less consideration than another in the sense that if they wanted somebody else's resource bundle they could have bid for it instead.

As mentioned above, many resource-theorists, including Dworkin, add to this system of equal resources and ambition-sensitivity, a sensitivity to inequalities in natural endowments. They note that natural inequalities are not distributed according to people's choices, nor are they justified by reference to some other morally relevant fact about people. Dworkin proposes a hypothetical compensation scheme in which he supposes that, before the hypothetical auction described above, people do not know their own natural endowments. However, they are able to buy insurance against being disadvantaged in the natural

distribution of talents and they know that their payments will provide an insurance pool to compensate those people who are unlucky in the ‘natural lottery’.

Because the Resource-based theory has a similar motivation to the Difference Principle the moral criticisms of it tend to be variations on those leveled against the Difference Principle. However, unlike the Difference Principle, it is not at all clear what would constitute an implementation of Resource-based theories and their variants in a real economy. It seems impossible to measure differences in people's natural talents - unfortunately, people's talents do not neatly divide into the natural and developed categories. A system of special assistance to the physically and mentally handicapped and to the ill would be a partial implementation of the compensation system, but most natural inequalities would be left untouched by such assistance while the theory requires that such inequalities be compensated for. It is simply not clear how to implement equality of resources in a complex economy and hence despite its theoretical advantages, it is difficult to see it as a practical improvement on the Difference Principle.

4. Welfare-Based Principles

Welfare-based principles are motivated by the idea that what is of primary moral importance is the level of welfare of people. Advocates of Welfare-based principles view the concerns of other theories - equality, the least advantaged, resources, desert-claims, or liberty as derivative concerns. Resources, equality, desert-claims, or liberty are only valuable in so far as they increase welfare, so that all distributive questions should be settled according to which distribution maximizes welfare. However, ‘maximizes welfare’ is imprecise, so welfare theorists propose particular welfare functions to maximize. The welfare functions proposed vary enormously both on what will count as welfare and the weighting system for that welfare. For almost any distribution of material benefits there is a welfare function whose maximization will yield that distribution (at least in a one sector period).

Distribution according to some welfare function is most commonly advocated by economists, who normally state the explicit functional form. Philosophers tend to avoid this. Philosophers also tend to restrict themselves to a small subset of the available welfare functions. Although there are a number of advocates of alternative welfare functions (like ‘equality of well-being’), most philosophical activity has concentrated on a variant known as Utilitarianism. This theory can be used to illustrate most of the main characteristics of Welfare-based principles.

Historically, Utilitarians have used the term ‘utility’ rather than ‘welfare’ and utility has been defined variously as pleasure, happiness, or preference-satisfaction. So, for instance, the principle for distributing economic benefits for Preference Utilitarians is to distribute them so as to maximize preference-satisfaction. The welfare function for such a principle has a simple theoretical form: it involves choosing that distribution maximizing the arithmetic sum of all satisfied preferences (unsatisfied preferences being negative), weighted for the intensity of those preferences.

The basic theory of Utilitarianism is one of the simplest to state and understand. Much of the work on the theory therefore has been directed towards defending it against moral criticisms, particularly from the

point of view of ‘commonsense’ morality. The criticisms and responses have been widely discussed in the literature on Utilitarianism as a general moral theory. Only two of the criticisms will be mentioned here.

The first is that Utilitarianism fails to take the distinctness of persons seriously. Maximization of preference-satisfaction is often taken as prudent in the case of individuals - people may take on greater burdens, suffering or sacrifice at certain periods of their lives so that their lives are overall better. The complaint against Utilitarianism is that it takes this principle, commonly described as prudent for individuals, and uses it on an entity, society, unlike individuals in important ways. While it may be acceptable for a person to choose to suffer at some period in her life (be it a day, or a number of years) so that her overall life is better, it is often argued against Utilitarianism that it is immoral to make some people suffer so that there is a net gain for other people. In the individual case, there is a single entity experiencing both the sacrifice and the gain. Also, the individuals, who suffer or make the sacrifices, choose to do so in order to gain some benefit they deem worth their sacrifice. In the case of society as a whole, there is no single experiential entity - some people suffer or are sacrificed so that others may gain. Furthermore, under Utilitarianism, there is no requirement for people to consent to the suffering or sacrifice.

A related criticism of Utilitarianism involves the way it treats individual preferences or interests referring to the holdings of others. For instance, some people may have a preference that some minority racial group should have less material benefits. Under Utilitarian theories, in their classical form, this preference or interest counts like any other in determining the best distribution. Hence, if racial preferences are widespread and are not outweighed by the minorities' contrary preferences, Utilitarianism will recommend an inegalitarian distribution based on race.

Utilitarians have responded to these criticisms in a number of ways. Utilitarians may apply their theory to the preferences themselves, arguing that utility is best promoted in the long run when people's preferences are shaped (if this is possible) in ways that are harmonious with one another, suggesting racist preferences should be discouraged. However, the utilitarian then must supply an account of why racist or sexist preferences should be discouraged when the same level of total long term utility could be achieved by encouraging the less powerful to be contented with a lower position. It is difficult for utilitarians to explain why the position of the oppressed is aptly desribed as one of oppression. Utilitarians have also argued that the empirical conditions are such that utility maximizing will rarely require women or racial minorities to sacrifice or suffer for the benefit of others, or to satisfy the prejudices of others. But if their theory on rare occasions does require people sacrifice or suffer in these ways, Utilitarians have defended this unintuitive consequence on the grounds that our judgements about what is wrong provide us with ‘rules of thumb’ which are useful at the level of commonsense morality but ultimately mistaken at the level of ‘critical theory’.

Utilitarian distribution principles, like the other principles described here, have problems with specification and implementation. Most formulations of Utilitarianism require interpersonal comparisons of utility. This means, for instance, that we must be able to compare the utility one person gains from eating an apple with that another gains from eating an apple. Furthermore, Utilitarianism requires that

differences in utility be measured and summed for widely disparate goods (so, for instance, the amount of utility a particular person gains from playing football is measured and compared with the amount of utility another gains from eating a gourmet meal). Critics have argued that such interpersonal utility comparisons are impossible, even in theory, due to one or both of the following: (1) It is not possible to combine all the diverse goods into a single index of ‘utility’ which can measured for an individual; (2) Even if you could do the necessary weighing and combining of the goods to construct such an index for an individual, there is no conceptually adequate way of calibrating such a measure between individuals. (see Elster 1991)

Utilitarians face a greater problem than this theoretical one in determining what material distribution is prescribed by their theory. Those who share similar Utilitarian theoretical principles frequently recommend very different material distributions to implement the principle. This problem occurs for other theories, but appears worse for Utilitarian and Welfare-based distribution principles. Recommendations for distributions or economic structures to implement other distributive principles commonly vary among advocates with similar theoretical principles, but the advocates tend to cluster around particular recommendations. This is not the case for Utilitarianism, with adherents dispersed in their recommendations across the full range of possible distributions and economic structures. For instance, many Preference Utilitarians believe their principle prescribes strongly egalitarian structures with lots of state invention while many other Preference Utilitarians believe it prescribes a *laissez faire* style of capitalism.

There is an explanation for why Utilitarians are faced with greater difficulties in implementation. Other distributive principles can rule out, relatively quickly, some practical policies on the grounds that they clearly violate the guiding principle, but Utilitarians must examine, in great detail, all the policies on offer. For each policy, they must determine the distribution of goods and services yielded by the policy and at least three other factors: the identity of each person in the distribution (if individuals' utility functions differ); the utility of each person from the goods and services distributed to them; the utility of each person from the policy itself. The size of the information requirements make this task impossible. Hence, broad assumptions must be made and each different set of assumptions will yield a different answer, and so the answers range across the full set of policies on offer. Moreover, there is no obvious way to arbitrate between the different sets of assumptions. For instance, suppose three Utilitarians agree on the same Utilitarian distributive principle. Utilitarian 1 however, asserts that the population's utility function conforms to function A (e.g. people's marginal utility is linear in the goods and services they consume) and is maximized by Policy 1; while Utilitarian 2 asserts that half the population's utility function conforms to function A and half to function B (e.g. people's marginal utility is diminishing) and is maximized by Policy 2; Utilitarian 3 asserts Utilitarian 2 is correct about the utility functions of the population but claims that Policy 3 will maximize utility. What seems impossible for advocates of Utilitarian-distribution principles to answer is how we would arbitrate these claims. If Utilitarian principles are to play a role in debates about distributive justice then this is the most important question to answer.

5. Desert-Based Principles

Another complaint against welfarism is that it ignores, and in fact cannot even make sense of, claims that people *deserve* certain economic benefits in light of their actions. The complaint is often motivated by the concern that various forms of welfarism treat people as mere containers for well-being, rather than purposeful beings, responsible for their actions and creative in their environments.

The different desert-based principles of distribution differ primarily according to what they identify as the basis for deserving. Most contemporary proposals for desert-bases fit into one of three broad categories:

1. Contribution: People should be rewarded for their work activity according to the value of their contribution to the social product. (Miller 1976, Miller 1989, Riley)
2. Effort: People should be rewarded according to the effort they expend in their work activity.
3. Compensation: People should be rewarded according to the costs they incur in their work activity. (Sadurski, Lamont 1997)

Aristotle argued that virtue should be a basis for distributing rewards, but most contemporary principles owe a larger debt to John Locke. Locke argued people deserve to have those items produced by their toil and industry, the products (or the value thereof) being a fitting reward for their effort. His underlying idea was to guarantee to individuals the fruits of their own labor and abstinence. According to the contemporary desert theorist, people freely apply their abilities and talents, in varying degrees, to socially productive work. People come to deserve varying levels of income by providing goods and services desired by others. (Feinberg) Distributive systems are just insofar as they distribute incomes according to the different levels earned or deserved by the individuals in the society for their productive labors, efforts, or contributions.

Contemporary desert-principles all share the value of raising the standard of living - collectively, ‘the social product’. Under each principle, only activity directed at raising the social product will serve as a basis for deserving income. The concept of desert itself does not yield this value of raising the social product; it is a value societies hold independently. Hence, desert principles identifying desert-bases tied to socially productive activity (productivity, compensation, and effort all being examples of such bases) do not do so because the concept of desert requires this. They do so because societies value higher standards of living, and therefore choose the raising of living standards as the primary value relevant to desert-based distribution. This means that the full development of desert-based principles requires specification (and defense) of those activities which will or will not count as socially productive, and hence as deserving of remuneration. (Lamont 1994)

It is important to distinguish desert-payments from entitlements. For desert theorists a well-designed institutional structure will make it so that many of the entitlements people have are deserved. But, of course, entitlements and just deserts can come apart – a person can be entitled to a payment without it being deserved, just as a person can be entitled to assume the presidential office without deserving it. (Feinberg 1970, 86) Similarly, a person may deserve a payment but not being entitled to it (such

instances are potential areas for institutional reform for a desert theorist). Payments designed to give people incentives are a form of entitlement particularly worth distinguishing from desert-payments as they are commonly confused. Incentive-payments are ‘forward-looking’ (Barry 1965, 111-112) in that they are set up to create a situation in the future, while desert-payments are ‘backwards-looking’ that they are justified with reference to work in the present or past. Even though it is possible for the same payment to be both deserved and an incentive, incentives and desert provide distinct rationales for income and should not be conflated. (Lamont 1997)

While some have sought to justify current capitalist distributions via desert-based distributive principles, John Stuart Mill and many since have forcefully argued the contrary claim - that the implementation of a productivity principle would involve dramatic changes in modern market economies and would greatly reduce the inequalities characteristic of them. It is important to note, though, that contemporary Desert-based principles are rarely complete distributive principles. They usually are only designed to cover distribution among working adults, leaving basic welfare needs to be met by other principles.

The specification and implementation problems for desert-based distribution principles revolve mainly around the desert-bases: it is difficult to identify what is to count as a contribution, an effort or a cost, and it is even more difficult to measure these in a complex modern economy.

The main moral objection to desert-based principles is that they make economic benefits depend on factors over which people have little control. John Rawls has made one of the most widely discussed arguments to this effect (Rawls 1971), and while the strong form of this argument has been clearly refuted (Zaitchik, Sher), it remains a problem for desert-based principles. The problem is most pronounced in the case of productivity-based principles - a person's productivity seems clearly to be influenced by many factors over which the person has little control.

It is interesting to note that under most welfare-based principles, it is also the case that people's level of economic benefits depend on factors beyond their control. But welfarists view this as a virtue of their theory, since they think the only morally relevant characteristic of any distribution is the welfare resulting from it. Whether the distribution ties economic benefits to matters beyond our control is morally irrelevant from the welfarist point of view. (As it happens, welfarists often hold the empirical claim that people have little control over their contributions to society anyway.) However, for people's benefits to depend on factors beyond their control is a more awkward result for desert theorists who emphasize the responsibility of people in choosing to engage in more or less productive activities.

6. Libertarian Principles

Most contemporary versions of the principles discussed so far allow some role for the market as a means of achieving the desired distributive pattern - the Difference Principle uses it as a means of helping the least advantaged; utilitarian principles commonly use it as a means of achieving the distributive pattern maximizing utility; desert-based principles rely on it to distribute goods according to desert, etc. In contrast, advocates of Libertarian distributive principles rarely see the market as a means to some desired

pattern, since the principle(s) they advocate do not ostensibly propose a ‘pattern’ at all, but instead describe the sorts of acquisitions or exchanges which are themselves just. The market will be just, not as a means to some pattern, but insofar as the exchanges permitted in the market satisfy the conditions of just exchange described by the principles. For Libertarians, just outcomes are those arrived at by the separate just actions of individuals; a particular distributive pattern is not required for justice. Robert Nozick has advanced this version of Libertarianism (Nozick 1974), and is its most well-known contemporary advocate.

Nozick proposes a 3-part "Entitlement Theory".

> If the world were wholly just, the following inductive definition would exhaustively cover the subject of justice in holdings:
>
> a. A person who acquires a holding in accordance with the principle of justice in acquisition is entitled to that holding.
> b. A person who acquires a holding in accordance with the principle of justice in transfer, from someone else entitled to the holding, is entitled to the holding.
> c. No one is entitled to a holding except by (repeated) applications of (a) and (b).
>
> The complete principle of distributive justice would say simply that a distribution is just if everyone is entitled to the holdings they possess under the distribution. (Nozick, p.151)

The statement of the Entitlement Theory includes reference to the principles of justice in acquisition and transfer. (For details of these principles see Nozick, pp.149-182.) The principle of justice in transfer is the least controversial and is designed to specify fair contracts while ruling out stealing, fraud, etc. The principle of justice in acquisition is more complicated and more controversial. The principle is meant to govern the gaining of exclusive property rights over the material world. For the justification of these rights, Nozick takes his inspiration from John Locke's idea that everyone ‘owns’ themselves and, by mixing one's labors with the world, self-ownership can generate ownership of some part of the material world. However, of Locke's mixing metaphor, Nozick legitimately asks: ‘...why isn't mixing what I own with what I don't own a way of losing what I own rather than a way of gaining what I don't? If I own a can of tomato juice and spill it in the sea so its molecules... mingle evenly throughout the sea, do I thereby come to own the sea, or have I foolishly dissipated my tomato juice?’ (Nozick 1974, p.174) Nozick concludes that what is significant about mixing our labor with the material world is that in doing so, we tend to increase the value of it, so that self-ownership can lead to ownership of the external world in such cases (Nozick 1974, pp. 149-182).

The obvious objection to this claim is that it is not clear why the first people to acquire some part of the material world should be able to exclude others from it (and, for instance, be the land owners while the later ones become the wage laborers). In response to this objection, Nozick puts a qualification on just acquisition, called the *Lockean Proviso,* whereby an exclusive acquisition of the external world is just, if, after the acquisition, there is ‘enough and as good left in common for others’. One of the main challenges

for Libertarians has been to formulate a morally plausible interpretation of this proviso. According to Nozick's interpretation, an acquisition is just if and only if the position of others after the acquisition is no worse than their position was when the acquisition was unowned or ‘held in common’. For Nozick's critics, his proviso is unacceptably weak. This is because it fails to consider the position others may have achieved under alternative distributions and thereby instantiates the morally dubious criterion of whoever is first gets the exclusive spoils. For example, one can satisfy Nozick's proviso by ‘acquiring’ a beach and charging $1 admission to those who previously were able to use the beach for free, so long as one compensates them with a benefit they deem equally valuable, such as a clean up or life-guarding service on the beach. However, the beach-goers would have been even better off had the more efficient organizer among them acquired the beach, charging only 50 cents for the same service, but this alternative is never considered under Nozick's proviso. (Cohen, 1995)

Will Kymlicka has given a summary of the steps in Nozick's self-ownership argument:

1. People own themselves.
2. The world is initially unowned.
3. You can acquire absolute rights over a disproportionate share of the world, if you do not worsen the condition of others.
4. It is relatively easy to acquire absolute rights over a disproportionate share of the world. Therefore:
5. Once private property has been appropriated, a free market in capital and labor is morally required. (Kymlicka, p.112)

The assessment of this argument is quite complex, but the difficulties mentioned above with the proviso call into question the step from (3) to (4).

The challenge for Libertarians then is to find a plausible reading of (3) which will yield (4). Moreover, at one point, Nozick claims the proviso must apply to both acquisitions and transfers, compounding the problem.

Of course, many existing holdings are the result of acquisitions or transfers which at some point did not satisfy principles (a) and (b) above. Hence, Nozick must supplement those principles with a principle of rectification for past injustice. Although he does not specify this principle he does describe its purpose:

> This principle uses historical information about previous situations and injustices done in them... and information about the actual course of events that flowed from these injustices, until the present, and it yields a description (or descriptions) of holdings in the society. The principle of rectification presumably will make use of its best estimate of subjunctive information about what would have occurred... if the injustice had not taken place. If the actual description of holdings turns out not to be one of the descriptions yielded by the principle, then one of the descriptions yielded must be realized. (Nozick 1974, pp. 152-153)

Nozick does not make an attempt to provide a principle of rectification. The absence of such a principle is much worse for a historical theory than for a patterned theory. Past injustices systematically undermine the justice of every subsequent distribution in historical theories. Nozick is clear that his historical theory is of no use in evaluating the justice of actual societies until such a theory of rectification is given:

> In the absence of [a full treatment of the principle of rectification] applied to a particular society, one *cannot* use the analysis and theory presented here to condemn any particular scheme of transfer payments, unless it is clear that no considerations of rectification of injustice could apply to justify it. (Nozick 1974, p.231)

Unfortunately for the theory, no such treatment will ever be forthcoming because the task is, for all practical purposes, impossible. The numbers of injustices perpetrated throughout history, both within nations and between them, are enormous and the necessary details of the vast majority of injustices are unavailable. Even if the details of the injustices were available, the counterfactual causal chains could not be reliably determined and, as Derek Parfit has pointed out, in a different context, even the people who would have been born would have been different. (Parfit 1986) As a consequence, Nozick's entitlement theory will never provide any guidance as to what the current distribution of material holdings should be nor what distributions or redistributions are legitimate or illegitimate. (Indeed Nozick suggests, for instance, the Difference Principle may be the best implementation of the principle of rectification.) Although Nozick is fairly candid about this consequence, many of his supporters and critics have ignored it and have carried on a vigorous debate as though his theory is an attempt to tell us something about the justice of current economic distributions.

Libertarians inspired by Nozick usually advocate a system in which there are exclusive property rights, with the role of the government restricted to the protection of these property rights. The property rights commonly rule out taxation for purposes other than raising the funds necessary to protect property rights. The strongest critique of any attempt to institute such a system of legally protected property rights comes, as we have seen, from Nozick's theory itself - there seems no obvious reason to give strong legal protection to property rights which have arisen through violations of the just principles of acquisition and transfer. But putting this critique to one side for a moment, what other arguments are made in favor of exclusionary property rights?

As already noted, Nozick argues that because people own themselves and hence their talents, they own whatever they can produce with these talents. Moreover, it is possible in a free market to sell the products of exercising one's talents. Any taxation of the income from such selling, according to Nozick, 'institute[s] (partial) ownership by others of people and their actions and labor'. (Nozick, p.172) People, according to this argument, have these exclusive rights of ownership. Taxation then, simply involves violating these rights and allowing some people to own (partially) other people. Moreover, it is argued, any system not legally recognizing these rights violates Kant's maxim to treat people always as ends in themselves and never merely as a means. The two main difficulties with this argument have been: (1) to show that self-ownership is only compatible with having such strong exclusive property rights; and (2) that a system of exclusive property rights is the best system for treating people with respect, as ends in

themselves.

Nozick candidly accepts that he does not himself give a systematic moral justification of the exclusionary property rights he advocates: ‘This book does not present a precise theory of the moral basis of individual rights.’ (Nozick, p.xiv) But others have tried to provide more systematic justifications of similar rights (Lomasky, Steiner) or to develop, more fully, justifications to which Nozick alludes.

In addition to the arguments from self-ownership, and the requirement to treat people as ends in themselves, the most common other route for trying to justify exclusive property rights has been to argue that they are required for the maximization of freedom and/or liberty or the minimization of violations of these. (Hayek) As an empirical claim though, this appears to be false. If we compare countries with less exclusionary property rights (e.g. more taxation) with countries with more exclusionary property regimes, we see no systematic advantage in freedoms/liberties enjoyed by people in the latter countries. (Of course, we do see a *difference* in distribution of such freedoms/liberties in the latter countries, the richer have more and the poorer less, while in the former they are more evenly distributed.) Now if Libertarians restrict what counts as a valuable freedom/liberty (and discount other freedoms/liberties people value), it will follow that exclusionary property rights are required to maximize freedom/liberty or to minimize violations of these. But the challenge for these Libertarians is to show why only their favored liberties and freedoms are valuable, and not those which are weakened by a system of exclusive property rights.

7. Feminist Principles

There is no one feminist conception of distributive justice; theorists who name themselves feminists defend positions across the political spectrum. Hence, feminists offer distinctive versions of all the theories considered so far as well as others. One way of thinking about what unifies many feminist theorists is an interest in what difference, if any, the practical experience of gender makes to the subject matter or study of justice; how different feminists answer this question distinguishes them from each other and from those alternative distributive principles which most inspire their thinking.

The distributive principles so far outlined, with the exception of strict egalitarianism, could be classified as liberal theories - they both inform, and are the product of, the liberal democracies which have emerged over the last two centuries. Lumping them together this way, though clumsy, makes the task of understanding the emergence of feminist critiques (and the subsequent positive theories) much easier.

John Stuart Mill in *The Subjection of Women* (1869) gives one of the clearest early feminist critiques of the political and distributive structures of the emerging liberal democracies. His writings provide the starting point for many contemporary liberal feminists. Mill argued that the principles associated with the developing liberalism of his time required equal political status for women. The principles Mill explicitly mentions include a rejection of the aristocracy of birth, equal opportunity in education and in the marketplace, equal rights to hold property, a rejection of the man as the legal head of the household, and equal rights to political participation. Feminists who follow Mill believe that a proper recognition of the

position of women in society requires that women be given equal and the same rights as men have, and that these primarily protect their liberty and their status as equal persons under the law. Thus, government regulation should not prevent women from competing on equal terms with men in educational, professional, marketplace and political institutions. From the point of view of other feminisms, the liberal feminist position is a conservative one, in the sense that it requires the proper inclusion for women of the rights, protections, and opportunities previously secured for men, rather than a fundamental change to the traditional liberal position. The problem for women, on this view, is not liberalism but the failure of society and the State to properly instantiate liberal principles.

One phrase or motto around which a whole range of feminists have rallied, however, marks a significant break with Mill's liberalism: ‘the personal is political.’ Feminists have offered a variety of interpretations of this motto, many of which take the form of a critique of liberal theories. Mill was crucial in developing the liberal doctrine of limiting the state's intervention in the private lives of citizens. Many contemporary feminists have argued that the resulting liberal theories of justice have fundamentally been unable to accommodate the injustices that have their origins in this ‘protected’ private sphere. This particular feminist critique has also been a primary source of inspiration for the broader multicultural critique of liberalism. The liberal commitments to government neutrality and to a protected personal sphere of liberty, where the government must not interfere, have been primary critical targets.

While issues about neutrality and personal liberty go beyond debates about distributive justice they also have application within these debates. The feminist critics recognize that liberalism correctly identifies the government as one potential source of oppression against individuals, and therefore recommends powerful political protections of individual liberty. They argue, however, that liberal theories of distributive justice are unable to address the oppression which surfaces in the so-called private sphere of government non-interference. Susan Moller Okin, for example, documents the effects of the institution of the nuclear family, arguing that the consequence of this institution is a position of systematic material and political inequality for women. Standard liberal theories, committed to neutrality in the private sphere, seem powerless to address (or sometimes even recognize) striking and lasting inequalities for women, minorities, or historically oppressed racial groups, when these are merely the cumulative effect of individuals' free behavior. Okin and others demonstrate, for example, that women have substantial disadvantages in competing in the market because of childrearing responsibilities which are not equally shared with men. As a consequence, any theory relying on market mechanisms, including most liberal theories, will yield systems which result in women systematically having less income and wealth than men. Thus, feminists have challenged contemporary political theorists to rethink the boundaries of political authority in the name of securing a just outcome for women and other historically oppressed groups.

While the political effects of personal freedom pose a serious challenge to contemporary liberal theories of distributive justice, the feminist critiques are somewhat puzzling because, as Jean Hampton puts it, many feminists appear to complain in the name of liberal values. In other words, their claims about the fundamental flaws of liberalism at the same time leave in tact the various ideals of liberty and equality which inspire the liberal theories of justice. Moreover, the task of defining feasible pathways for modifying the structure of liberal democracies without undermining their virtues and protections has

proved more difficult than the setting out the criticisms of liberalism. Indeed, despite a legitimate feminist worry about the effects of so-called government neutrality on women's material status, the relative neutrality of liberal democracies compared to non-liberal societies has been one of the significant contributing factors both to the flourishing of feminist theory and to the many significant practical gains women in liberal democracies have made relative to women in other parts of the world. The challenge, being taken up by many, is to navigate both a coherent theoretical and practical path in response to the best feminist critiques available (see the entry on feminist ethics).

Bibliography

Strict Egalitarianism

- Carens, Joseph, *Equality, Moral Incentives and the Market* (Chicago: Chicago University Press, 1981)
- Rawls, John, *A Theory of Justice* (Harvard, MA: Harvard University Press, 1971)
- Nielsen, Kai, ‘Radical Egalitarian Justice: Justice as Equality’ *Social Theory and Practice*, 1979, 209-226

The Difference Principle

- Crocker, Lawrence, ‘Equality, Solidarity, and Rawls' Maximin’ *Philosophy and Public Affairs*, 1977, 262-266
- Rawls, John, *A Theory of Justice* (Harvard, MA: Harvard University Press, 1971)
- Rawls, John, *Political Liberalism*, (New York: Columbia University Press, 1993)
- Wellbank, J. H., *John Rawls and his critics: an annotated bibliography* (New York: Garland Pub., 1982)

Resource-Based Principles

- Arneson, Richard, ‘Liberalism, Distributive Subjectivism, and Equal Opportunity for Welfare’ *Philosophy and Public Affairs* 19, 1990, 158-194
- Daniels, Norman, ‘Equality of What: Welfare, Resources, or Capabilities?’ *Philosophy and Phenomenological Research*, 1990
- Dworkin, Ronald, ‘What is Equality? Part 1: Equality of Resources’, *Philosophy and Public Affairs*, 10, 1981, 185-246
- Dworkin, Ronald, ‘What is Equality? Part 2: Equality of Welfare’, *Philosophy and Public Affairs*, 10, 1981, 283-345
- Dworkin, Ronald, *Sovereign Virtue* (Cambridge, MA: Harvard University Press, 2000)
- Kronman, Anthony T., ‘Talent Pooling’ in J. Roland Pennock and John W.Chapman (eds.), *Human Rights: Nomos 23*, (New York: New York University Press, 1981), 58-79
- Sen, Amartya, ‘Equality of What?’ in: Sen, Amartya, *Choice, Welfare and Measurement*

(Cambridge: Cambridge University Press, 1982)

Welfare-Based Principles

- Elster, Jon, and John E. Roemer (eds.), *Interpersonal Comparisons of Well-Being*, (Cambridge: Cambridge University Press, 1991)
- Glover, Jonathan (ed.), *Utilitarianism and Its Critics*, (New York: Macmillan Publishing Company, 1990)
- Goodin, Robert E., *Utilitarianism as a Public Philosophy* (New York: Cambridge University Press, 1995)
- Hardin, Russell, *Morality within the Limits of Reason* (Chicago: University of Chicago, 1988)
- Rescher, Nicholas, *Distributive Justice: A Constructive Critique of the Utilitarian Theory of Distribution* (Indianapolis: Bobbs-Merrill Co., 1966)
- Sen, Amartya, and Bernard Williams (eds.), *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982)

Desert-Based Principles

- Feinberg, Joel, 'Justice and Personal Desert', *Doing and Deserving* (Princeton, NJ, Princeton University Press, 1970), 55-94
- Lamont, Julian, 'Incentive Income, Deserved Income, and Economic Rents', *Journal of Political Philosophy*, 5, 1997, 26-46
- Lamont, Julian, 'Problems For Effort-Based Distribution Principles', *Journal of Applied Philosophy*, 12, 1995, 215-229
- Lamont, Julian, 'The Concept of Desert in Distributive Justice', *The Philosophical Quarterly*, 44, 1994, 45-64
- Mill, John Stuart, Principles of Political Economy
- Miller, David, *Market, State, and Community* (Oxford: Clarendon Press, 1989)
- Miller, David, *Social Justice* (Oxford: Clarendon Press, 1976)
- Riley, Jonathan, 'Justice Under Capitalism', *Markets and Justice*, ed. John W. Chapman (New York: New York University Press, 1989), 122-162
- Sadurski, Wojciech, *Giving Desert Its Due* (Dordrecht, Holland: D. Reidel, 1985)
- Sher, George, *Desert* (Princeton, NJ: Princeton University Press, 1987)
- Zaitchik, Alan, 'On Deserving to Deserve', *Philosophy and Public Affairs* 6, 1977, 370-388

Libertarian Principles

- Bogart, J. H., 'Lockean Provisos and State of Nature Theories' *Ethics*, 1985, 824-836
- Christman, John, 'Self-Ownership, Equality, and the Structure of Property Rights' *Political Theory*, 1991, 28-46
- Cohen, G. A., *Self-Ownership, Freedom, and Equality* (New York: Cambridge University Press,

1995)

- Hayek, Friedrich A., *The Constitution of Liberty* (London, Routledge and Kegan Paul, 1960)
- Kymlicka, Will, *Contemporary Political Philosophy* (Oxford: Clarendon Press, 1990)
- Lomasky, Loren E., *Persons, Rights, and the Moral Community* (New York: Oxford University Press, 1987)
- Nozick, Robert, *Anarchy, State and Utopia* (New York: Basic Books, 1974)
- Steiner, H., Liberty and Equality, *Political Studies*, 1981, 555-569

Feminist Principles

- Okin, Susan Moller, *Justice, Gender and the Family* (New York: Basic Books, 1991)
- Hampton, Jean, *Political Philosophy* (Boulder, Colorado: Westview Press, 1997)
- Held, Virginia *Rights and Goods: justifying social action* (New York: Free Press, 1994)
- Gatens, Moira, *Feminism and Philosophy: Perspectives on Difference and Equality* (Indianapolis: Indianan University Press, 1991)
- MacKinnon, Catherine A., *Sex Equality* (Foundation Press, 2001)
- MacKinnon, Catherine A., *Feminism Unmodified: Discourses of Life and Law* (Cambridge, MA: Harvard Univ Press, 1987)
- Pateman, Carol, *The Sexual Contract* (Stanford: Stanford University Press, 1988)
- Tong, Rosemary, *Feminine and Feminist Ethics* (Belmont, California: Wadsworth Publishing Company, 1993)

Further Theories and General Reference

- Ackerman, Bruce A., *Social Justice and the Liberal State* (New Haven: Yale University Press, 1980)
- Alstott, Anne and Bruce A. Ackerman, *The Stakeholder Society* (New Haven: Yale University Press, 1999)
- Arthur, John and William Shaw (eds.), *Justice and Economic Distribution 2nd Ed.*, (Englewood Cliffs, NJ: Prentice-Hall, 1991)
- Barry, Brian, *Political Argument* (London: Routledge and Keagan Paul, 1965).
- Barry, Brian, *Theories of Justice, Vol. 1* (Berkeley: University of California Press, 1989)
- Cohen, G.A., ‘Where the Action Is: On the Site of Distributive Justice’, in *Philosophy and Public Affairs*, 26, 1997, 3-30
- Cohen, G.A., *If You’re an Egalitarian, How Come You’re so Rich?* (Cambridge, MA: Harvard University Press, 2000)
- Gauthier, David, *Morals by Agreement* (Cambridge: Cambridge University Press, 1987)
- Kymlicka, Will, *Contemporary Political Philosophy* (Oxford: Clarendon Press, 1990)
- Parfit, Derek, *Reasons and Persons* (Oxford University Press, 1986)
- Roemer, John E., *Theories of Distributive Justice* (Harvard University Press, 1996)
- Scheffler, Samuel, *Boundaries and Allegiances* (Oxford: Oxford University Press, 2001)
- Walzer Michael, *Spheres of Justice* (New York: Basic Books, 1984)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

consequentialism | desert | egalitarianism | feminism, interventions: feminist ethics | liberalism | liberty | Locke, John | Mill, John Stuart

Julian Lamont
University of Queensland
j.lamont@uq.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 22, 1996
Content last modified: June 29, 2002

Desert

The concept of desert is deeply entrenched in everyday morality. We say that effort deserves success, wrongdoing deserves punishment, innocent suffering deserves sympathy or compensation, virtue deserves happiness, and so on. We think that the getting of what's deserved is *just*, and that failure to receive what's deserved is *unjust*. We also believe it's *good* that a person gets what she deserves, and *bad* that she doesn't — even if she deserves something bad, like punishment. We assume, too, that it's *wrong* to treat people better or worse than they deserve, and *right* to treat them according to their deserts. In these and other ways, the notion of desert pervades our ethical lives.

In spite of its ubiquity, or perhaps because of it, the notion of desert is not especially well understood. This isn't surprising, since there are many difficult questions surrounding desert. For instance, what are the ingredients (as it were) of desert? What sorts of thing can be deserving? What are the grounds or bases for desert? How do bases for desert manage to make a thing that has them deserving? What connections does desert have to other moral-normative concepts, such as justice and goodness? This article sketches some possible answers to these and subsidiary questions about desert.

1. The Ingredients of Desert

Consider some ordinary desert claims: "Hans deserves praise in virtue of his efforts," "Because of her outstanding scholarly contributions, Nkechi deserves promotion to full professor," "Financial compensation is what the innocent victims of September 11 deserve." These desert claims have several things in common: each involves a deserving *subject* (Hans, Nkechi, innocent victims), a deserved *object*

(praise, promotion, compensation) and a desert *basis* (effort, contribution, innocent suffering). This suggests that desert itself is a three-place relation that holds among a subject, an object, and a basis. Of course, sometimes the desert claims we utter do not explicitly refer to all three of these ingredients. For example, one might say that Hans deserves praise (without specifying the basis of his desert), or that Nkechi is deserving (without specifying what she deserves). But unless one can fill these claims out further -- say, by explaining why one thinks that Hans deserves praise, or what it is one thinks that Nkechi deserves -- then the concept of desert is being misused.

It might be thought that desert involves more than three ingredients -- more, that is, than a basis, object, and subject. This thought might be due to reflection on desert claims like these: “In virtue of his efforts, Hans deserves praise *from his teacher*,” “Because of her scholarly contributions, Nkechi deserves *from the University* the position of full professor,” “Financial compensation *from groups that sponsor terrorism* is what the victims of September 11 deserve.” These desert claims specify a "source" from which the subject deserves the object. Hans deserves praise not from just anyone, but from his teacher. Nkechi deserves promotion not just by any institution, but by her university. The victims of September 11 deserve compensation not from just any group, but from those who harbor terrorists. Thus, it might be supposed that desert has four ingredients: a basis, a subject, an object, and its source (Kleinig 1971).

There are two reasons to resist adding this fourth ingredient to desert. One is that some legitimate cases of desert involve either no object-source, or an object-source so general that specifying it would be otiose. For example, you might deserve some good fortune after a long streak of undeserved bad luck, even though the good luck you now deserve is not deserved from any particular source. Or, in virtue of being a person, you might deserve respect -- but from whom? Well, everyone! A second reason for not allowing an independent fourth ingredient (the object-source) into desert is that in fact it is already contained in one of the undisputed ingredients -- namely, the desert object itself. For example, the object of Hans’s desert is not simply praise, but *praise from his teacher*; the object of Nkechi’s desert is not simply promotion, but *promotion by her university*; and so on. In these cases, the desert object includes a specific source. In other cases, the desert object will not involve a specific source -- as when you deserve good luck from no particular source, or respect from everyone.

2. Subjects of Desert

The most uncontroversial bearers of desert are human beings. Humans are thought to deserve, or be capable of deserving, many things: punishment, reward, apologies, compensation, admiration, contempt, wages, grades, prizes, and so on. But is desert limited to human beings? Can non-human animals, for example, be deserving too?

A conservative view is that *only* humans can be deserving, and that any attribution of desert to non-humans is either incoherent, false, or translatable into some claim of "human" desert, or perhaps into a claim that does not mention desert at all. So, for example, suppose someone says “The dog deserves a treat.” On the conservative view, this claim could be incoherent, since dogs (it is claimed) neither deserve nor fail to deserve anything. The concept of desert simply does not apply to them. Or it could be

coherent but false: the dog does not deserve a treat, since dogs cannot literally deserve anything. Or it could be coherent and even true, provided that it's translatable without change of meaning into some other claim, e.g., "You deserve the satisfaction of giving your dog a treat," or "Giving the dog a treat will reinforce its good behavior."

This view seems a bit *too* conservative. First, it flies in the face of the way we ordinarily apply the concept of desert. Indeed, if ordinary language is any guide, then just about anything can be literally deserving. For in addition to making desert claims about humans, we also say, for example, that the pet deserves our love, the Olympic Peninsula our respect, the proposal our support, the nation our loyalty, the painting our admiration. (It's possible that some of these claims are translatable in ways that would be consistent with the conservative view, but it's not obvious that all of them are. The burden of proof -- and of translation! -- is on the conservatives.) Second, what makes a human being capable of desert probably is not the biological fact that she is a human being -- that is, a member of the species *Homo sapiens*. It might instead be the fact that she is capable of reason or self-reflection, or that she is capable of experiencing pain and pleasure. If so, then since there are non-human creatures (e.g., dogs) who can either reason or experience pain and pleasure, then at least some non-human creatures are subjects of desert too.

Unfortunately, this isn't saying very much. For one thing, it's not clear *how* having the ability to reason or experience pleasure or pain would make a creature capable of being deserving. Furthermore, it's not obvious that possession of one or more of those abilities is even necessary for being a subject of desert. Dead people, who don't have those abilities, are sometimes thought to deserve various things, such as a decent burial. Likewise, inanimate objects -- such as the Grand Canyon or Big Ben -- are also said to deserve certain things, such as preservation or protection.

Clearly, the question of what is necessary or sufficient for being a subject of desert is a difficult one. Is there any way to answer it? One possibility suggests itself: first, draw up a list of the possible bases of desert; then look around and see what sorts of things can have them. The result would be at least a partial catalogue of the sorts of things that can be deserving. In order to do this, of course, we would need a list of possible desert bases. What would that list look like? What are the bases for desert?

3. Bases for Desert

Suppose we are asked to draw up a list of the bases for desert -- that is, a catalogue of the sorts of things such that having any of them would make a subject deserving. What would go on the list, and how would each entry be justified? We might agree that effort, for example, is a basis for deserving reward or success, but we might be less sure about whether *need*, for instance, is a basis for deserving medical care, or whether *moral worth* is a basis for deserving happiness. In order to settle such doubts and arrive at a justifiable catalogue, some general and defensible principles for identifying desert bases seem necessary.

There is agreement that the basis of a thing's desert must in some important sense be a fact "about" that thing (Feinberg 1970). So, for example, the fact that 3 is a prime number, or that a particular star is two

billion years old, is not a basis for your deserving anything. In order for a fact to be a basis of your desert, it must be a fact about you -- for example, that you have worked hard, innocently suffered, are a person, and so on. So the basis of a thing's desert must be a fact about that thing. Is there anything more that can be said?

Many writers have claimed that another necessary condition on something's being a desert base is that the deserving subject is *responsible* for it (see Cupit 1996a and Feldman 1995b, 1996 for references and discussion). A standard version of this condition can be stated as follows:

> (DR) *A* deserves *x* in virtue of *y* only if *A* is responsible for *y*.

DR is plausible at first glance, and there are many cases that seem to support it. If, for instance, a professor discovers that his student isn't responsible for the high-quality paper she submitted (because she stole it from the Internet), then she can't deserve a good grade for it. Or consider the fact that criminal action deserves punishment, but not if the agent was insane at the time of the action. Presumably, the justification for this is that an insane agent is not responsible for his or her actions, and no one deserves punishment for actions for which he or she isn't responsible.

DR plays a central role in contemporary discussions about desert. It occurs, for instance, in a much-discussed argument for the conclusion that no one deserves anything. (This argument is suggested by some passages in Rawls 1971, though it's not clear that Rawls himself means to endorse the following versions of it. See Miller 1976 for discussion.) The argument goes something like this: all the actions we perform and attributes we possess -- in short, all would-be bases for desert -- are determined by factors for which we are not responsible, such as our genetic makeup, early training, and environment. But if no one is responsible for his or her possession of any would-be basis for desert, then these are not bases for desert at all (*DR*). Therefore, no one deserves anything. On another version of this argument, it involves an extra premise: if *A* deserves *x* in virtue of *y*, then *A* must also deserve *y* (Nozick 1974 and Zaitchik 1977 add this premise to the argument, and both reject it). So, for example, Rostropovich deserves praise for his musical talent only if he deserves to have that talent. But does he? This is where *DR* comes in. For if *DR* is true, then Rostropovich can deserve his musical talent only if he's responsible for it -- which he isn't, since he was born with it. Thus, he can't deserve praise for it. Likewise, it's argued, for any alleged desert base: no one could deserve to have any so-called desert base, since no one is responsible for having such things. Thus, once again, no one deserves anything.

DR appears elsewhere in theorizing about desert. A good example is found in the literature on desert of wages. Much of this literature centers on the question of whether the wage a worker deserves depends on the effort exerted (regardless of actual productivity), or on productivity (regardless of effort). Sometimes the following argument is proposed: wages are deserved either for effort or productivity; people are responsible only for their efforts, not for the success or productivity of those efforts; DR is true; thus, it's the worker's effort, not productivity, that determines the deserved wage. (For more discussion of this argument, see Lamont 1994, McLeod 1996, and Sadurski 1985.)

In spite of *DR*s intuitive appeal and wide acceptance, it has not gone unchallenged. To some, it seems obvious that we can be deserving in virtue of things for which we are not responsible (Cupit 1996a, Feldman 1995b, 1996). Consider a person who innocently suffers a brutal mugging. The victim is not responsible for the attack, but deserves compensation anyway. Or consider a person who innocently suffers an excruciating disease that is very expensive to treat. She is not responsible for contracting this disease, yet she might deserve medical care and sympathy. Or consider the fact that you are a person. No one is responsible for this, yet in virtue of being a person you deserve a modicum of respect. These examples suggest that if there is a connection between desert and responsibility, it's more complicated than the one posited by *DR*. (For more discussion of the connection, see Secton 5.4 below, as well as Cupit 1996a,b and Smilansky 1996a, b.)

Another attempt to place conditions on desert bases does so by attempting to link them with certain emotions or attitudes (Miller 1976). A version of this view begins by calling attention to the class of attitudes that we take up toward people in virtue of various qualities they possess or actions they perform. Those attitudes include admiration, gratitude, disgust, resentment, and so on. These have been called the "appraising attitudes". The idea here is that the bases for appraising attitudes are, or at least coincide with, the bases for desert. Put another way:

> (DAA) *x* is a desert base if and only if *x* is the basis of an appraising attitude.

If correct, *DAA* would provide a useful principle for determining what the desert bases are. For example, it's sometimes held that need is a basis for desert (say, of medical care). Yet need doesn't seem to be the basis of any appraising attitude. We don't respect or resent people, admire or detest them, because of their needs. Thus, if *DAA* is true, then need isn't a basis for desert. On the other hand, it seems that exerting effort (for example) is a basis for admiring a person. If *DAA* is true, it follows that effort is a basis for desert.

However, there may be a problem with *DAA*. To see it, consider the distinction between something's actually being admired or resented, on the one hand, and that thing's being *appropriately* admired or resented on the other. For example, many admired Hitler even though it would have been appropriate to detest him. And many resented Martin Luther King, Jr., even though it would have been appropriate to admire him. This allows us to ask: is *DAA* the thesis that *x* is a basis for desert if and only if *x* happens to be the object of an appraising attitude? Or is it instead the view that *x* is a basis for desert if and only if *x* is an appropriate basis for an appraising attitude? If the former, then *DAA* reduces to an implausible form of relativism about desert bases. (Desert bases would be properties whose presence we just happened to admire or detest, even if we admired nasty properties and detested noble ones.) If the latter, then *DAA* is in danger of being vacuous. For what else could make *x* an "appropriate" basis for an appraising attitude, except for the fact that those who have *x* *deserve* to be the object of that attitude?

Still another way of attempting to determine the bases for desert proceeds in two stages (Feinberg 1970). First, draw up a list of all the sorts of treatment that can be deserved: prizes, rewards, punishments, grades, compensation, and so on. Second, for each form of treatment, attempt to specify the basis or

bases for which it is deserved. There are two potential problems with this approach. First, just as it is difficult to draw up a catalogue of bases for desert, it is hard to draw up a list of forms of deservable treatment. Second, this way of proceeding might reinforce the assumption that for every sort of deservable treatment, there is a desert base or set of desert bases unique to it. This assumption may be correct, but another possibility is that there is a single set of desert bases, and possession of any or all of them can influence the extent to which one deserves any given form of deservable treatment. If this latter view of the relationship between desert bases and deserved treatment is correct, then attempting to match deserved treatments to their bases might not be the best way to determine the bases of desert (McLeod 1996).

Yet another method for determining desert bases follows directly from an "institutional" theory of desert (Cummisky 1987, Arnold 1987). On that sort of theory, the bases for desert are determined by the rules or purposes of social institutions. For example, if the purpose of the ‘institution’ of Olympic gymnastics is to award the gold medal to the gymnast who receives the highest number of points from the judges, then the gymnast who achieves that score deserves the gold medal. If a theory of this sort is correct, then discovering the bases for desert will be as easy (or as difficult) as discovering the rules or purposes of social institutions. A possible source of trouble for this view, however, is that desert bases do not seem to be entirely determined by the rules or purposes of social institutions. Otherwise morally repugnant rules such as racist restrictions on voting eligibility, sexist restrictions on employee benefits, and so on, could not be fairly criticized on the grounds that race, gender, and so on are not legitimate grounds for deserving the loss of such benefits. But such criticism certainly seems warranted. This has persuaded some writers that there are "pre-institutional" facts about the bases for desert, and that social rules and purposes can be evaluated in terms of them (Feinberg 1970, McLeod 1999).

4. How desert bases work

How do desert bases work? In particular, how does possession of a basis for desert manage to make its possessor deserving? The difficulty of this question can be brought out by imagining a person who has, for example, exerted quite a lot of *effort* toward achieving some end. Effort is widely thought to be a basis for desert, so let’s suppose that it is. Would this mean that the person who exerted effort now deserves to achieve his end? One reason to doubt this is that his end might have been an *evil* one -- for example, the bombing of a building full of innocent people. Surely he can’t deserve that this horrible end be realized, no matter how hard he works for it. So maybe we should add that effort is a basis for desert only if it’s directed toward a morally unobjectionable end. If correct, this would explain why the terrorist can’t deserve that his efforts succeed. After all, those efforts are directed toward the morally objectionable end of murdering innocent people.

However, it’s probably a mistake to conclude that exerting effort toward a morally unobjectionable end is *sufficient* for deserving the end itself. This is because there might be other factors at work that could weigh against one’s deserving it, even if the end itself is morally okay. For example, suppose you’ve worked hard toward getting an A on your term paper. You spent hours in the library doing research, you composed several drafts, you went hours without sleep, and so on. Your end is morally okay, since

there's nothing wrong with trying (in these ways!) to get an A. Even so, it's still possible that, in spite of your tremendous efforts, your paper is terrible. In that case, you probably don't deserve an A.

The general phenomenon emerging here is that, at least in many cases, possession of any particular basis for desert is not going to be sufficient for being deserving. You might work hard toward an end, but what if the end itself is wrong? Or the end is okay, and your effort intense, but what if the product of your effort is of very low quality? Even if the product of your effort is of high quality, you still might not be deserving, since other factors could count against you. Until all these of factors are taken into consideration and weighed against each other, it's impossible to render a verdict of desert.

This sort of phenomenon is not unprecedented. In fact it comes up often in philosophical ethics, most notably in theorizing about moral rightness and wrongness. There, it's standard to distinguish *prima facie* moral rightness from *all-things-considered* moral rightness (Ross 1930). To illustrate this distinction, imagine that you've made a solemn promise to help your friend. In virtue of this, it would be morally right, in some sense, for you to keep your promise. However, the rightness here is merely *prima facie*, for it might not be all-things-considered morally right for you to help your friend. This is because your promise might, for example, have been to help your friend by murdering his enemy (who, let's suppose, is an innocent person). In that case, you almost certainly have an all-things-considered obligation to break your promise. Still, the fact that you made a solemn promise to perform a certain action does seem to count toward the action's being right. For in other circumstances, a solemn promise to do something could generate an all-things-considered obligation to do it. It just so happens that, in this case, the *prima facie* rightness that the act would inherit in virtue of being a promise-keeping is outweighed by the *prima facie* wrongness of committing murder. Thus, on this way of thinking, the all-things-considered rightness or wrongness of any action is determined by all the respects in which the action would be *prima facie* right, when weighed against all the respects in which it would be *prima facie* wrong.

It might be that desert bases work in a similar way. That is, there might be a distinction between *prima facie* desert and *all-things-considered* desert. Possession of any given desert base makes one *prima facie* deserving, but it might not make one all-things-considered deserving. For example, you might have exerted intense effort toward achieving some end. If, as is usually held, effort is a basis for deserving success, then you *prima facie* deserve success in virtue of your effort. However, you might not be all-things-considered deserving of success. Perhaps this is because your effort was directed toward an evil end. In such a case, the *prima facie* desert of success that you gained in virtue of exerting effort is outweighed by the *prima facie* desert of failure and punishment that you acquired in virtue of plotting an evil end. If these are all the desert bases in play, then you all-things-considered deserve failure and punishment, not success.

The obvious drawback to thinking of desert bases as functioning in this way is that the method, if any, for determining the weight or importance of desert bases in particular cases is pretty mysterious. How, for instance, are we to weigh effort against productivity in the context of deserving a wage, or moral worth against need in the context of deserving medical care, or a person's humanity against his criminal behavior in the context of deserving punishment? Until we have some principled way of weighing *prima facie* desert bases against each other, the distinction between *prima facie* and all-things-considered desert

might serve only to describe, rather than solve, the problem of how desert bases work.

An alternative approach would involve abandoning the distinction between *prima facie* and all-things-considered desert, and working toward a catalogue of more “finely-grained” desert bases. The idea here is that if the desert bases are specified in enough detail, then the relationship between having them and being deserving would be simple: possession of a desert base would be sufficient for being deserving. On this sort of view, effort itself, for example, is not a basis for desert, nor is effort-directed-toward-a-morally-unobjectionable-end, since neither one seems sufficient for being deserving (say, of an A on a paper). Instead, possession of a much more complicated desert base -- viz., effort-directed-toward-a-morally-unobjectionable-end-that-is-not-also-a-grade-on-a-paper-etc. -- is sufficient for desert.

This kind of view has the advantage of positing a simple relationship between possessing a desert base and being deserving. That relationship is this: possession of the desert base *entails* being deserving. However, this simplicity is purchased at the cost of making the desert bases themselves hopelessly complicated. The prior view, by contrast, makes desert bases very simple (e.g., ‘effort’, ‘productivity’, ‘moral worth’), but the price for this simplicity is a complicated and admittedly mysterious relationship between possessing desert bases and being deserving.

5. Desert’s Relationship to Some Other Concepts

Fully grasping a concept involves understanding its relationships to other concepts. Thus, in order to fully grasp the concept of desert, it’s important to see what connections it has to other concepts. Desert probably bears interesting connections to a large number of concepts, but this concluding section focuses only on four: justice, intrinsic value, entitlement, and responsibility.

5.1 Justice

There are many theories of justice (and, some would say, many sorts of justice -- distributive, retributive, social, etc. -- for there to be theories about). Some of these theories are egalitarian, since they state that some sort of equality is most central to justice. Other theories of justice are libertarian, because of the supreme importance that they place on liberty or freedom. But an ancient idea is that justice involves the getting of what’s deserved -- even if this results in inequalities, and even if distribution according to desert involves or requires some loss of liberty. On an old-fashioned version of this view, for instance, justice obtains entirely to the extent that the morally virtuous are happy, and the morally wicked suffer. If happiness were somehow to be distributed according to moral goodness in this way, the result would be inequality with respect to happiness, since the more virtuous would be happier than the less virtuous. There would also be a loss of liberty or freedom for the morally wicked, since they would be punished or otherwise made to suffer. But these inequalities and losses of freedom wouldn’t detract from the justice of the world; instead, they would be required by justice itself.

However, contemporary theorists don’t agree about the relationship between justice and desert. Some

seem willing to accept that justice is entirely a matter of getting what's deserved (Feldman 1992, 1995a, 1995c). A more moderate position is that getting what's deserved is part, but only a part, of justice (Feinberg 1974, Lucas 1980, Slote 1973). Another possible component of justice is *fairness*, which has to do with the way one's treatment compares to the treatment received by others. Suppose, for example, that every student in the class deserves a C, but one of them is arbitrarily given an A and the rest are given Cs. Some will say, quite plausibly, that although these other students are given the grades they deserve, they are treated unjustly because they are treated unfairly. If this is correct, then justice cannot consist simply in getting what's deserved. Other factors, such as fairness, might be relevant to justice as well (Feinberg 1974). Another such factor might be *consent*. For suppose a person has worked hard to earn money, for example, and thereby comes to deserve it. Even so, there is no injustice if the person freely consents to giving this money away (Slote 1973).

Some theorists have gone so far as to argue that desert has nothing at all to do with justice. This view contradicts commonsense morality, but the arguments for holding it have been influential. One such argument is that the concept of desert, rather than providing a basis for the explanation of justice, is in fact conceptually parasitic on the notion of justice. On this way of seeing things, to deserve something is to be entitled to it according to rules that are just. The justice of these rules is then explained not in terms of how well or regularly they result in deserved distributions, which would be circular, but rather by some criterion (such as agreement on those rules by rational parties) that has nothing to do with desert (Rawls 1971, Scanlon 1984). The motivation behind a view like this might be that a more robust notion of desert would involve metaphysical mysteries, such as freedom or responsibility (Scheffler 1995). Or perhaps the motivation is simply the pragmatic one that any system designed to distribute goods and evils according to individual deserts would be hopelessly impractical (Rawls 1971). For how could we determine each individual's moral worth, level of effort, productivity, and so on, and then distribute benefits and burdens accordingly?

There is another argument for thinking that desert is irrelevant to justice -- or, more precisely, that distribution according to 'desert' would actually involve *injustice*. This argument relies on the intuition that just distributions cannot be based on factors over which the recipients have no control. The distribution of economic and political benefits according to race or gender, for instance, seems unjust since neither one's race nor gender is within one's control. But what if the alleged bases for desert -- effort, moral worth, productivity, being a person, and so on -- are also beyond or largely outside of one's control, and due instead to factors such as such as genetic endowment and early training? It would seem to follow that distributions based on these factors, these "desert bases", are unjust (Rawls 1971).

5.2 Intrinsic value

The *intrinsic* value of a thing is the value it has simply in virtue of what it is, rather than the value it has in virtue of what it leads to, signifies, entails, purchases, and so on. (For reasons that can't be discussed here, the concept of intrinsic value is of central importance in philosophical ethics.) The branch of ethics concerned with intrinsic value is known as *axiology*. A helpful assumption often made in axiology is that intrinsic value is had not just by anything at all, but rather by states of affairs or propositions. And one of

the central questions in axiology is this: what elements can contribute to the intrinsic value of a state of affairs?

Some have argued that one such element involves desert (Feldman 1992, 1995c, Kagan 1999; Hurka 2001). To see why, consider two states of affairs: (i) that Smith is happy, and (ii) that Jones is unhappy. What is the intrinsic value of these states of affairs? One might suppose that (i) is intrinsically good, since happiness seems a likely constituent of intrinsically good states of affairs, and that (ii) is intrinsically bad, since unhappiness seems a likely ingredient of intrinsically bad states of affairs. So far, so good. However, it might be that happiness and unhappiness are not the only contributors to the intrinsic values of states of affairs. Another such factor might be the "fit" between the happiness or unhappiness *received*, on the one hand, and the happiness or unhappiness *deserved*, on the other. For suppose that Smith and Jones are morally despicable people, and that therefore their happiness is undeserved. In that case, it seems wrong to say that (i) is intrinsically good and (ii) intrinsically bad. On the contrary, a more plausible view is that (i) is intrinsically bad and (ii) intrinsically good! In any case, an intuitively appealing position is that a state of affairs in which a person who deserves unhappiness but receives happiness is intrinsically bad, and a state of affairs in which a person who deserves unhappiness and gets it is not intrinsically bad, and quite possibly intrinsically good. If this is correct, then the fit between desert and receipt within a state of affairs is at least one determinant of its intrinsic value.

That said, the quality of fit between desert and receipt is almost surely not the only determinant of intrinsic value (Persson 1996). Compare, for instance, two possible worlds (here thought of as extremely large states of affairs). One world contains a million sinners, all of whom suffer the horrible punishment that they deserve. Another world contains a million saints, all of whom enjoy the bliss that they deserve. These are rather different worlds, but in one respect they are the same: the quality of fit between desert and receipt in the sinners' world is equal to that of the saints' world. This is because the fit in each world is perfect: each person in each world receives precisely what he or she deserves, and the worlds have the same population. However, it seems absurd to say that these worlds have the same intrinsic value. On the contrary, the sinners' world seems intrinsically much worse than the saints'. If so, then the intrinsic value of a state of affairs is not exhausted by the quality of fit between desert and receipt.

5.3 Entitlement

We often speak of being *entitled* to something, like a vacation, a grade, an inheritance, or an apology. Notice, however, that in many of these cases we could just as easily have said that we *deserve* these things. Thus, the terms 'entitled to' and 'deserve' are often used interchangeably. This raises the question of desert's relationship to entitlement. Are they one and the same, or are they different? And if they are distinct notions, then what connection, if any, does desert bear to entitlement?

First, though, what exactly is entitlement? Reflecting on a few cases can bring out the sense of 'entitlement' of interest here. Suppose, then, that the rules of a certain corporation state that its employees shall receive two weeks of paid vacation after one year of full-time employment. Now suppose that an employee of the company has worked full-time for one year. This employee is now entitled to two weeks

of paid vacation. Or, to take another example, suppose the rules of a certain board game specify that a player who rolls "snake eyes" must lose a turn. If a player now rolls snake eyes, then that player is "entitled" to lose a turn. (Admittedly, we don't usually think of penalties as objects of entitlement, but, in the sense of 'entitlement' that concerns us, they can be.) Suppose, as a final example, that the accepted rules of etiquette state that the hostess of an upcoming dinner party shall receive RSVPs from her invited guests, even if she fails to request the favor of a response. In that case, the hostess is entitled to a response from all her invitees. These cases have several things in common. First, there is a conventional rule that specifies that a person shall receive a certain treatment in virtue of possessing an "entitlement base" (that is, a particular attribute or performance of a certain action). Second, there is a person who falls under the rule and who has the entitlement base. Third, this person thereby comes to be entitled to the treatment in question. These are paradigmatic cases of entitlement.

Clearly, entitlement thus understood is structurally similar to desert. For entitlement, like desert, is a three-place relation among an entitled subject, a basis of entitlement, and an object of entitlement. Also, as noted above, many objects of entitlement -- vacations, punishment, replies to invitations -- are also objects of desert. Furthermore, failure to treat in accordance with entitlement, like failure to treat in accordance with desert, can be an injustice. These considerations might lead some to conclude that there is a profound relationship between entitlement and desert.

Some might want to say that the relationship between desert and entitlement is extremely intimate. Indeed, the "institutional" theories of desert mentioned in Section 3 are precisely those that *identify* desert with some sort of entitlement. However, this proposed connection (identity) is a bit too intimate, for there are cases in which a person is entitled to something but doesn't deserve it, and also cases in which what's deserved isn't something to which the person is entitled. For instance, the rules that govern the state lottery might entitle the winning ticket holder to one hundred million dollars, even if the lucky winner doesn't deserve so much money. Or, it might be that everyone in the United States deserves free or affordable access to basic health care, even though there are no rules that entitle us to it. These cases suggest that if there is an interesting relationship between desert and entitlement, it isn't identity.

At least some authors have claimed that although entitlement is not identical to desert, it is nevertheless a basis for desert (Feldman 1992, 1995a; McLeod 1999). On their view, being entitled to something is a basis for deserving it. Perhaps the main motivation for this position springs from combining the conviction that justice is simply the getting of what's deserved with the apparent fact that failing to treat in accordance with entitlement can involve injustice. If this combination of views is correct, it would seem to follow that being entitled to something must be a basis for deserving it. But this view also suffers from serious problems. One is its implication that evil or morally repugnant rules are capable of generating desert. Suppose, for example, that Nazi laws entitled officers of the SS to property confiscated from Jews. Even so, it hardly seems correct to say that, in virtue of this, the Nazi officers came to deserve (even prima facie) the stolen property. Another problem with the view that entitlement is a basis for desert is that it seems to violate the requirement that a desert basis must be a fact about the deserving subject. Unlike the property of being morally virtuous or being lazy, the property of being entitled by the rules to x is a highly extrinsic property of an individual, and thus cannot count as a basis for desert.

A third proposed connection between desert and entitlement is found in Cupit (1996b). Suppose, for purposes of illustration, that the orchestra has just pulled off an electrifying performance of Dvorak's Ninth Symphony. The musicians deserve the audience's admiration, and there are rules for how it can be expressed. If, as in this case, their admiration is immense, then the rules call for a standing ovation. Put another way, there is a rule that entitles the musicians to a standing ovation from the audience. Thus, if the audience fails to give the orchestra a standing ovation, the message is conveyed that the musicians don't deserve it. In other words, if the audience fails to give the players that to which they are entitled, then it fails to give them what they deserve. In this way, a link seems to be forged between desert and entitlement: the rules that generate entitlement also help to shape the meanings of the actions that fall under those rules; those actions are meant to express what the recipient deserves; therefore, failure to treat in accordance with entitlement can result in failure to treat in accordance with desert.

5.4 Responsibility

Much of the relevant literature presupposes or explicitly considers some conception of the relationship, if any, that desert bears to the concept of *responsibility*. As noted in Section 3, some authors (e.g., Rawls 1971, Rachels 1978, Pojman 1997) claim that desert always presupposes responsibility. On their view, one cannot deserve anything in virtue of an action or attribute for which one isn't responsible. Feldman (1995a), Cupit (1996a, b), and others challenge this claim by pointing to properties that seem to be bases for desert but that do not presuppose responsibility. Innocent suffering, being a person, being beautiful, being a member of an endangered species - all of these properties are plausibly regarded as bases for desert of various forms of treatment, such as compensation, respect, admiration, protection, and so on, in spite of the fact that one need not be responsible for possessing them. Thus, it seems that if there is a desert-responsibility connection, it's not as simple as some have thought.

There are ways in which proponents of the view that desert always involves responsibility might deal these apparent counterexamples. One is to distinguish "moral" from "nonmoral" desert -- moral desert presupposes responsibility, while nonmoral desert does not -- and to relegate nonmoral desert to the philosophical scrap heap. But this maneuver seems both *ad hoc* and implausible: *ad hoc* because what motivates the distinction is simply a desire to save the alleged desert-responsibility connection; implausible because treatment (or failure to treat) in accordance with innocent suffering, being a person, being beautiful, etc., doesn't really seem to be nonmoral. On the contrary, failure to treat in accordance with a thing's possession of one or more of those properties would be unfitting or wrong, while treatment in accordance with one or more of them would be fitting or right. A similar objection applies to treating desert as a species of "merit". On this view, merit is based on possession of any quality that is an appropriate basis for treatment, whether the meriting subject is responsible for the quality or not, while desert is a species of merit that requires responsibility for the merit basis (Pojman 1997). This too is *ad hoc*, since ordinary language doesn't support a sharp merit-desert distinction. It also contradicts the actual practice of competent speakers who readily speak of desert in cases where the subject isn't responsible for the basis.

This isn't to say that desert and responsibility are not connected in any way. In fact, there are cases where

desert does require responsibility (perhaps the most obvious is punishment). Thus, rather than focus on simple versions of a desert-responsibility connection, theorists of desert should distinguish those cases in which desert involves responsibility from those in which it does not, and articulate general principles that explain the difference.

Bibliography

- Adkins, A. W. H., 1960, *Merit and Responsibility: A Study of Greek Values*, Chicago: University of Chicago Press
- Adler, Jonathan, 1987, 'Luckless Desert is Different Desert,' *Mind* 96: 247-49
- Annis, David B., and Bohanon, Cecil E., 1992, 'Desert and Property Rights,' *Journal of Value Inquiry* 26: 537-546
- Arnold, N. Scott, 1987, 'Why Profits are Deserved,' *Ethics* 97: 387-402
- Barry, Brian, 1965, *Political Argument*, London: Routledge and Kegan Paul
- -----, 1991, *Liberty and Justice: Essays in Political Theory* 2, Oxford: Clarendon Press
- Becker, Lawrence C., 1977, *Property Rights: Philosophic Foundations*, London: Routledge and Kegan Paul
- Benn, S. I., and Peters, R. S., 1959, *The Principles of Political Thought*, New York, NY: The Free Press
- Campbell, Tom, 1988, *Justice*, Atlantic Highlands, NJ: Humanities Press International
- Card, Claudia, 1972, 'On Mercy,' *Philosophical Review* 81: 182-207
- Cummiskey, David, 1987, 'Desert and Entitlement: A Rawlsian Consequentialist Account,' *Analysis* 47: 15-19
- Cupit, Geoffrey, 1996a, 'Desert and Responsibility,' *Canadian Journal of Philosophy* 26: 83-100
- -----, 1996b, *Justice as Fittingness*, Oxford: Clarendon Press, 1996
- Daniels, Norman, 1978, 'Merit and Meritocracy,' *Philosophy and Public Affairs* 7: 206-223
- Davis, Michael, 1985, 'How to Make the Punishment Fit the Crime,' in *Criminal Justice: Nomos XXVII*, edited by Pennock, Roland J. and Chapman, John W. New York, NY: New York University Press, 1985: 119-55
- Dennett, Daniel, 1984, *Elbow Room*, Cambridge, MA: Massachusetts Institute of Technology Press
- Dick, James, 1975, 'How to Justify a Distribution of Earnings,' *Philosophy and Public Affairs* 4: 248-72
- Feinberg, Joel, 1970, 'Justice and Personal Desert,' in Feinberg, *Doing and Deserving*, Princeton University Press
- -----, 1973, *Social Philosophy*, Englewood Cliffs, NJ: Prentice-Hall
- -----, 1974, 'Noncomparative Justice,' *Philosophical Review* 83: 297-338
- Feldman, Fred, 1992. *Confrontations With the Reaper*, Oxford: Oxford University Press
- ------, 1995a, 'Adjusting Utility for Justice: A Consequentialist Reply to the Objection from Justice,' *Philosophy and Phenomenological Research* 55: 567-585
- ------, 1995b, 'Desert: Reconsideration of Some Received Wisdom,' *Mind* 104: 63-77
- ------, 1995c, 'Justice, Desert, and the Repugnant Conclusion,' *Utilitas* 7: 189-206

- ------, 1996, ‘Responsibility as a Condition for Desert,’ *Mind* 105: 165-168
- ------, 1997. *Utilitarianism, Hedonism, and Desert*, Cambridge: Cambridge University Press
- Fields, Lloyd, 1987, ‘Parfit on Personal Identity and Desert,’ *Philosophical Quarterly* 37: 432-440
- Galston, William, 1992, *Justice and the Human Good*, Chicago, IL: University of Chicago Press
- Garcia, J. L. A., 1988, ‘A Problem About the Basis of Desert,’ *Journal of Social Philosophy*: 11-19
- Gaus, Gerald F., 1990, *Value and Justification: The Foundations of Liberal Theory*, Cambridge: Cambridge University Press
- Goodin, Robert, 1985, ‘Negating Positive Desert Claims,’ *Political Theory* 13: 575-98
- Griffin, James, 1986, *Well-Being: Its Meaning, Measurement, and Moral Importance*, Oxford: Clarendon Press
- Hestevold, H. Scott, 1983, ‘Disjunctive Desert,’ *American Philosophical Quarterly* 20: 357-63
- Hill, Christopher, 1985, ‘Desert and the Moral Arbitrariness of the Natural Lottery,’ *Philosophical Forum* 16: 207-22
- Holborow, Les, 1975, ‘Desert, Equality, and Injustice,’ *Philosophy* 50: 157-68
- Holmgren, Margaret, 1986, ‘Justifying Desert Claims: Desert and Opportunity,’ *Journal of Value Inquiry* 20: 265-78
- Hurka, Thomas, 2001, ‘The Common Structure of Virtue and Desert,’*Ethics* 112: 6-31
- Husak, Douglas, 1992, ‘Why Punish the Deserving?’ *Nous* 26: 447-464
- Jackson, Michael W., 1986, *Matters of Justice*, Wolfeboro, NH: Croom Helm
- Kagan, Shelly, 1999, ‘Equality and Desert,’ in Pojman and McLeod, eds., *What Do We Deserve?* New York: Oxford University Press, 1999: 298-314
- Kernohan, Andrew, 1993, ‘Desert and Self-Ownership,’ *Journal of Value Inquiry* 27: 197-202
- Kleing, John, 1971, ‘The Concept of Desert,’ *American Philosophical Quarterly* 8: 71-78
- Kleinig, John, 1973, *Punishment and Desert*, The Hague: Martinus Nijhoff
- Lamont, Julian, 1994, ‘The Concept of Desert in Distributive Justice,’ *The Philosophical Quarterly* 44: 45-64
- Lucas, J. R., 1980, *On Justice*, Oxford: Clarendon Press
- McLeod, Owen, 1996, ‘Desert and Wages,’ *Utilitas* 8: 205-221
- -----, 1999, ‘Desert and Institutions,’ in Pojman and McLeod, eds., *What Do We Deserve?* New York: Oxford University Press: 186-95
- Miller, David, 1976, *Social Justice*, Oxford: Oxford University Press
- -----, 1989, *Market, State, and Community*, Oxford: Clarendon Press
- -----, 1992a, ‘Deserving Jobs,’ *Philosophical Quarterly* 42: 161-81
- -----, 1992b, ‘Distributive Justice: What the People Think,’ *Ethics* 102: 555-93
- -----, 1996, ‘Two Cheers for Meritocracy,’ *Journal of Political Philosophy* 4: 277-301
- New, Christopher, 1992, ‘Time and Punishment,’ *Analysis* 52: 35-40
- Nozick, Robert, 1974, *Anarchy, State, and Utopia*, New York: Basic Books
- Persson, Ingmar, 1996, ‘Feldman’s Justicized Act Utilitarianism,’ *Ratio* 9: 39-46
- Pojman, Louis, 1997, ‘Equality and Desert,’ *Philosophy* 72: 549-70
- ------, 1999, ‘Merit: Why Do We Value It?’ *Journal of Social Philosophy* 30: 83-102
- ------ and McLeod, Owen, eds., 1999, *What Do We Deserve?* New York: Oxford University Press

- Primoratz, Igor, 1989, *Justifying Legal Punishment*, Atlantic Highlands, NJ: Humanities Press International
- Rachels, James, 1978, 'What People Deserve,' in *Justice and Economic Distribution*, Arthur, John and Shaw, William, eds., Englewood Cliffs, NJ: Prentice-Hall: 167-196
- Rawls, John, 1971, *A Theory of Justice*, Cambridge, MA: Harvard University Press
- Rescher, Nicholas, 1966, *Distributive Justice*, Indianapolis: Bobbs-Merrill
- Richards, Norvin, 1986, 'Luck and Desert,' *Mind* 95: 198-209
- Ross, W. D., 1930, *The Right and the Good*, Oxford: Oxford University Press
- Sadurski, Wojceich, 1985, *Giving Desert its Due*, Dordrecht: D. Reidel Publishing
- Sandel, Michael, 1982, *Liberalism and the Limits of Justice*, Cambridge: Cambridge University Press
- Scanlon, Thomas, 1988, 'The Significance of Choice,' in *The Tanner Lectures on Human Values* VIII, McMurrin, Sterling, ed., Salt Lake City, UT: University of Utah Press: 149-216
- Scheffler, Samuel, 1992, 'Responsibility, Reactive Attitudes, and Liberalism in Philosophy and Politics,' *Philosophy and Public Affairs* 21: 299-323
- Sher, George, 1987, *Desert*, Princeton, NJ: Princeton University Press
- Slote, Michael, 1973, 'Desert, Consent, and Justice,' *Philosophy and Public Affairs* 2: 323-347
- Smilansky, Saul, 1996a, 'Responsibility and Desert: Defending The Connection,' *Mind* 105: 157-163
- -----, 1996b, 'The Connection between Responsibility and Desert: The Crucial Distinction,' *Mind* 105: 485-486
- Sterba, James, 1974, 'Justice as Desert,' *Social Theory and Practice* 3: 101-116
- -----, 1976, 'Justice and the Concept of Desert,' *The Personalist* 57: 188-97
- Sverdlik, Steven, 1983a, 'The Logic of Desert,' *Journal of Value Inquiry* 17: 317-324
- -----, 1983b, 'The Nature of Desert,' *Southern Journal of Philosophy* 21: 585-594
- Vallentyne, Peter, 1995, 'Taking Justice Too Seriously,' *Utilitas* 7: 207-216
- Waller, Bruce, 1987, 'Just and Nonjust Deserts,' *The Southern Journal of Philosophy* 25: 229-238
- -----, 1989, 'Uneven Starts and Just Deserts,' *Analysis* 49: 209-213
- Wicclair, Mark, 1986, 'Preferential Treatment and Desert,' *Social Theory and Practice* 12: 287-308
- Young, Robert, 1992, 'Egalitarianism and Personal Desert,' *Ethics* 102: 319-341
- Zaitchik, Alan, 1977, 'On Deserving to Deserve,' *Philosophy and Public Affairs* 6: 370-88

Other Internet Resources

[To be supplied by the author before publication.]

Related Entries

justice | moral responsibility | value: intrinsic vs. extrinsic

Owen McLeod
Lafayette College
mcleodo@lafayette.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 14, 2002
Content last modified: May 14, 2002

Globalization

Covering a wide range of distinct political, economic, and cultural trends, the term “globalization” has quickly become one of the most fashionable buzzwords of contemporary political and academic debate. In popular discourse, globalization often functions as little more than a synonym for one or more of the following phenomena: the pursuit of classical liberal (or “free market”) policies in the world economy (“economic liberalization”), the growing dominance of western (or even American) forms of political, economic, and cultural life (“westernization” or “Americanization”), the proliferation of new information technologies (the “Internet Revolution”), as well as the notion that humanity stands at the threshold of realizing one single unified community in which major sources of social conflict have vanished (“global integration”). Fortunately, recent social theory has formulated a more precise concept of globalization than those typically offered by pundits. Although sharp differences continue to separate participants in the ongoing debate, most contemporary social theorists endorse the view that globalization refers to fundamental changes in the spatial and temporal contours of social existence, according to which the significance of space or territory undergoes shifts in the face of a no less dramatic acceleration in the temporal structure of crucial forms of human activity. Geographical distance is typically measured in time. As the time necessary to connect distinct geographical locations is reduced, distance or space undergoes compression or “annihilation.” The human experience of space is intimately connected to the temporal structure of those activities by means of which we experience space. Changes in the temporality of human activity inevitably generate altered experiences of space or territory. Theorists of globalization disagree about the precise sources of recent shifts in the spatial and temporal contours of human life. Nonetheless, they generally agree that alterations in humanity's experiences of space and time are working to undermine the importance of local and even national boundaries in many arenas of human endeavor. Since globalization contains far-reaching implications for virtually every facet of human life, it necessarily suggests the need to rethink key questions of normative political theory.

1. Globalization in the History of Ideas

The term globalization has only become commonplace in the last two decades, and academic commentators who employed the term as late as the 1970s accurately recognized the novelty of doing so (Modelski, 1972). At least since the advent of industrial capitalism, however, intellectual discourse has been replete with allusions to phenomena strikingly akin to those that have garnered the attention of recent theorists of globalization. Nineteenth and twentieth-century philosophy, literature, and social commentary include numerous references to an inchoate yet widely shared awareness that experiences of distance and space are inevitably transformed by the emergence of high-speed forms of transportation (for example, rail and air travel) and communication (the telegraph or telephone) that dramatically heighten possibilities for human interaction across existing geographical and political divides (Harvey, 1989; Kern, 1983). Long before the introduction of the term globalization into recent popular and scholarly debate, the appearance of novel high-speed forms of social activity generated extensive commentary about the compression of space.

Writing in 1839, an English journalist commented on the implications of rail travel by anxiously postulating that as distance was “annihilated, the surface of our country would, as it were, shrivel in size until it became not much bigger than one immense city” (Harvey, 1996: 242). A few years later, Heinrich Heine, the émigré German-Jewish poet, captured this same experience when he noted: “space is killed by the railways. I feel as if the mountains and forests of all countries were advancing on Paris. Even now, I can smell the German linden trees; the North Sea's breakers are rolling against my door” (Schivelbusch, 1978: 34). Another German émigré, the socialist theorist Karl Marx, in 1848 formulated the first theoretical explanation of the sense of territorial compression that so fascinated his contemporaries. In Marx's account, the imperatives of capitalist production inevitably drove the bourgeoisie to “nestle everywhere, settle everywhere, and establish connections everywhere.” The juggernaut of industrial capitalism constituted the most basic source of technologies resulting in the annihilation of space, helping to pave the way for “intercourse in every direction, universal interdependence of nations,” in contrast to a narrow-minded provincialism that had plagued humanity for untold eons (Marx, 1979 [1848]: 476). Despite their ills as instruments of capitalist exploitation, new technologies that increased possibilities for human interaction across borders ultimately represented a progressive force in history. They provided the necessary infrastructure for a cosmopolitan future socialist civilization, while simultaneously functioning in the present as indispensable organizational tools for a working class destined to undertake a revolution no less oblivious to traditional territorial divisions than the system of capitalist exploitation it hoped to dismantle.

European intellectuals have hardly been alone in their fascination with the experience of territorial compression, as evinced by the key role played by the same theme in early twentieth-century American thought. In 1904, the literary figure Henry Adams diagnosed the existence of a “law of acceleration,” fundamental to the workings of social development, in order to make sense of the rapidly changing spatial and temporal contours of human activity. Modern society could only be properly understood if the seemingly irrepressible acceleration of basic technological and social processes was given a central place in social and historical analysis (Adams, 1931 [1904]). John Dewey argued in 1927 that recent economic

and technological trends implied the emergence of a “new world” no less noteworthy than the opening up of America to European exploration and conquest in 1492. For Dewey, the invention of steam, electricity, and the telephone offered formidable challenges to relatively static and homogeneous forms of local community life that had long represented the main theatre for most human activity. Economic activity increasingly exploded the confines of local communities to a degree that would have stunned our historical predecessors, for example, while the steamship, railroad, automobile, and air travel considerably intensified rates of geographical mobility. Dewey went beyond previous discussions of the changing temporal and spatial contours of human activity, however, by suggesting that the compression of space posed fundamental questions for democracy. Dewey observed that small-scale political communities (for example, the New England township), a crucial site for the exercise of effective democratic participation, seemed ever more peripheral to the great issues of an interconnected world. Increasingly dense networks of social ties across borders rendered local forms of self-government ineffective. Dewey wondered, “How can a public be organized, we may ask, when literally it does not stay in place?” (Dewey, 1954 [1927]: 140). To the extent that democratic citizenship minimally presupposes the possibility of action in concert with others, how might citizenship be sustained in a social world subject to ever more astonishing possibilities for movement and mobility? New high-speed technologies attributed a shifting and unstable character to social life, as demonstrated by increased rates of change and turnover in many arenas of activity (most important perhaps, the economy) directly affected by them, and the relative fluidity and inconstancy of social relations there. If citizenship requires some modicum of constancy and stability in social life, however, did not recent changes in the temporal and spatial conditions of human activity bode poorly for political participation? How might citizens come together and act in concert when contemporary society's “mania for motion and speed” made it difficult for them even to get acquainted with one another, let alone identify objects of common concern? (Dewey, 1954 [1927]: 140).

The unabated proliferation of high-speed technologies is probably the main source of the numerous references in intellectual life since 1950 to the annihilation of distance. The Canadian cultural critic Marshall McLuhan made the theme of a technologically based “global village,” generated by social “acceleration at all levels of human organization,” the centerpiece of an anxiety-ridden analysis of new media technologies in the 1960s (McLuhan, 1964: 103). Arguing in the 1970s and '80s that recent shifts in the spatial and temporal contours of social life exacerbated authoritarian political trends, the French social critic Paul Virilio seemed to confirm many of Dewey's darkest worries about the decay of democracy. According to his analysis, the high-speed imperatives of modern warfare and weapons systems strengthened the executive and debilitated representative legislatures. The compression of territory thereby paved the way for executive-centered emergency government (Virilio, 1986 [1977]). But it was probably the German philosopher Martin Heidegger who most clearly anticipated contemporary debates about globalization. Heidegger not only described the “abolition of distance” as a constitutive feature of our contemporary condition, but he linked recent shifts in spatial experience to no less fundamental alterations in the temporality of human activity: “All distances in time and space are shrinking. Man now reaches overnight, by places, places which formerly took weeks and months of travel” (Heidegger, 1971 [1950]: 165). Heidegger also accurately prophesied that new communication and information technologies would soon spawn novel possibilities for dramatically extending the scope of *virtual reality*: “Distant sites of the most ancient cultures are shown on film as if they stood this very

moment amidst today's street traffic…The peak of this abolition of every possibility of remoteness is reached by television, which will soon pervade and dominate the whole machinery of communication" (Heidegger, 1971 [1950]: 165). Heidegger's description of growing possibilities for simultaneity and instantaneousness in human experience ultimately proved no less apprehensive than the views of many of his predecessors. In his analysis, the compression of space increasingly meant that from the perspective of human experience "everything is equally far and equally near." Instead of opening up new possibilities for rich and multi-faceted interaction with events once distant from the purview of most individuals, the abolition of distance tended to generate a "uniform distanceless" in which fundamentally distinct objects became part of a bland homogeneous experiential mass (Heidegger, 1971 [1950]: 166). The loss of any meaningful distinction between "nearness" and "distance" contributed to a leveling down of human experience, which in turn spawned an indifference that rendered human experience monotonous and one-dimensional.

2. Globalization in Contemporary Social Theory

Since the mid-1980s, social theorists have moved beyond the relatively underdeveloped character of previous reflections on the compression or annihilation of space to offer a rigorous conception of globalization. To be sure, major disagreements remain about the precise nature of the causal forces behind globalization, with David Harvey (1989, 1996) building directly on Marx's pioneering explanation of globalization, while others (Giddens, 19990; Held, McGrew, Goldblatt, Perraton, 1999) question the exclusive focus on economic factors characteristic of the Marxist approach. Nonetheless, a consensus about the basic rudiments of the concept of globalization appears to be emerging.

First, contemporary analysts associate globalization with *deterritorialization*, according to which a growing variety of social activities takes place irrespective of the geographical location of participants. As Jan Aart Scholte observes, "global events can -- via telecommunication, digital computers, audiovisual media, rocketry and the like -- occur almost simultaneously anywhere and everywhere in the world" (Scholte, 1996: 45). Globalization refers to increased possibilities for action between and among people in situations where latitudinal and longitudinal location seems immaterial to the social activity at hand. Even though geographical location remains crucial for many undertakings (for example, farming to satisfy the needs of a local market), deterritorialization manifests itself in many social spheres. Business people on different continents now engage in electronic commerce; television allows people situated anywhere to observe the impact of terrible wars being waged far from the comfort of their living rooms; academics make use of the latest video conferencing equipment to organize seminars in which participants are located at disparate geographical locations; the Internet allows people to communicate instantaneously with each other notwithstanding vast geographical distances separating them. Territory in the sense of a traditional sense of a geographically identifiable location no longer constitutes the whole of "social space" in which human activity takes places. In this initial sense of the term, globalization refers to the spread of new forms of non-territorial social activity (Ruggie, 1993; Scholte, 2000).

Second, recent theorists conceive of globalization as linked to the growth of social *interconnectedness* across existing geographical and political boundaries. In this view, deterritorialization is a crucial facet of

globalization. Yet an exclusive focus on it would be misleading. Since the vast majority of human activities is still tied to a concrete geographical location, the more decisive facet of globalization concerns the manner in which distant events and forces impact on local and regional endeavors (Tomlinson, 1999: 9). For example, this encyclopedia might be seen as an example of a deterritorialized social space since it allows for the exchange of ideas in cyberspace. The only prerequisite for its use is access to the Internet. Although substantial inequalities in Internet access still exist, use of the encyclopedia is in principle unrelated to any specific geographical location. However, the reader may very well be making use of the encyclopedia as a supplement to course work undertaken at a school or university. That institution is not only located at a specific geographical juncture, but its location is probably essential for understanding many of its key attributes: the level of funding may vary according to the state or region where the university is located, or the same academic major might require different courses and readings at a university in China, for example, than in Argentina or Norway. Globalization refers to those processes whereby geographically distant events and decisions impact to a growing degree on "local" university life. For example, the insistence by powerful political leaders in the First World that the International Monetary Fund (IMF) should require that Latin and South American countries commit themselves to a particular set of economic policies might result in poorly paid teachers and researchers as well as large, understaffed lecture classes in San Paolo or Lima; the latest innovations in information technology from a computer research laboratory in India could quickly change the classroom experience of students in British Columbia or Tokyo. Globalization refers "to processes of change which underpin a transformation in the organization of human affairs by linking together and expanding human activity across regions and continents" (Held, McGrew, Goldblatt, Perraton, 1999: 15). Globalization in this sense is a matter of degree since any given social activity might influence events more or less faraway: even though a growing number of activities seems intermeshed with events in distant continents, certain human activities remain primarily local or regional in scope. Also, the magnitude and impact of the activity might vary: geographically removed events could have a relatively minimal or a far more extensive influence on events at a particular locality. Finally, we might consider the degree to which interconnectedness across frontiers is no longer merely haphazard but instead predictable and regularized (Held, McGrew, Goldblatt, Perraton, 1999).

Third, globalization must also include reference to the *speed* or *velocity* of social activity. Deterritorialization and interconnectedness initially seem chiefly spatial in nature. Yet it is easy to see how these spatial shifts are directly tied to the acceleration of crucial forms of social activity. As we observed above in our discussion of the conceptual forerunners to the present-day debate on globalization, the proliferation of high-speed transportation, communication, and information technologies constitutes the most immediate source for the blurring of geographical and territorial boundaries that prescient observers have diagnosed at least since the mid-nineteenth century. The compression of space presupposes rapid-fire forms of technology; shifts in our experiences of territory depend on concomitant changes in the temporality of human action. High-speed technology only represents the tip of the iceberg, however. The linking together and expanding of social activities across borders is predicated on the possibility of relatively fast flows and movements of people, information, capital, and goods. Without these fast flows, it is difficult to see how distant events could possibly posses the influence they now enjoy. High-speed technology plays a pivotal role in the velocity of human affairs. But many other factors contribute to the overall pace and speed of social activity. The

organizational structure of the modern capitalist factory offers one example; certain contemporary habits and inclinations, including the "mania for motion and speed" described by Dewey, represent another. Deterritorialization and the expansion of interconnectedness are intimately tied to the acceleration of social life, while social acceleration itself takes many different forms (Eriksen, 2001). Here as well, we can easily see why globalization is always a matter of degree. The velocity or speed of flows, movements, and interchanges across borders can vary no less than their magnitude, impact, or regularity.

Fourth, even though analysts disagree about the causal forces that generate globalization, most agree that globalization should be conceived as a relatively *long-term process*. The triad of deterritorialization, interconnectedness, and social acceleration hardly represents a sudden or recent event in contemporary social life. Globalization is a constitutive feature of the modern world, and modern history includes many examples of globalization (Giddens, 1990). As we saw above, nineteenth-century thinkers captured at least some of its core features; the compression of territoriality composed an important element of their lived experience. Nonetheless, some contemporary theorists believe that globalization has taken a particularly intense form in recent decades, as innovations in communication, transportation, and information technologies (for example, computerization) have generated stunning new possibilities for simultaneity and instantaneousness (Harvey, 1989). In this view, present-day intellectual interest in the problem of globalization can be linked directly to the emergence of new high-speed technologies that tend to minimize the significance of distance and heighten possibilities for deterritorialization and social interconnectedness. Although the intense sense of territorial compression experienced by so many of our contemporaries is surely reminiscent of the experiences of earlier generations, some contemporary writers nonetheless argue that it would be mistaken to obscure the countless ways in which ongoing transformations of the spatial and temporal contours of human experience are especially far-reaching. While our nineteenth-century predecessors understandably marveled at the railroad or the telegraph, a comparatively vast array of social activities is now being transformed by innovations that accelerate social activity and considerably deepen longstanding trends towards deterritorialization and social interconnectedness. To be sure, the impact of deterritorialization, social interconnectedness, and social acceleration are by no means universal or uniform: migrant workers engaging in traditional forms of low-wage agricultural labor in the fields of southern California, for example, probably operate in a different spatial and temporal context than the Internet entrepreneurs of San Francisco or Seattle. Distinct assumptions about space and time often coexist uneasily during a specific historical juncture (Gurvitch, 1964). Nonetheless, the impact of recent technological innovations is profound, and even those who do not have a job directly affected by the new technology are shaped by it in innumerable ways as citizens and consumers (Eriksen, 2001: 16).

Fifth, globalization should be understood as a *multi-pronged* process, since deterritorialization, social interconnectedness, and acceleration manifest themselves in many different (economic, political, and cultural) arenas of social activity. Although each facet of globalization is linked to the core components of globalization described above, each consists of a complex and relatively autonomous series of empirical developments, requiring careful examination in order to disclose the causal mechanisms specific to it (Held, McGrew, Goldblatt, Perraton, 1999). Each manifestation of globalization also generates distinct conflicts and dislocations. For example, there is substantial empirical evidence that cross-border flows and exchanges, as well as the emergence of directly transnational forms of production

by means of which a single commodity is manufactured simultaneously in distant corners of the globe, are gaining in prominence (Castells, 1996). High-speed technologies and organizational approaches are employed by transnationally operating firms, the so-called “global players,” with great effectiveness. The emergence of “around-the-world, around-the-clock” financial markets, where major cross-border financial transactions are made in cyberspace at the blink of an eye, represents a familiar example of the economic face of globalization. Global financial markets also challenge traditional attempts by liberal democratic nation-states to rein in the activities of bankers, spawning understandable anxieties about the growing power and influence of financial markets over democratically elected representative institutions. In political life, globalization takes a distinct form, though the general trends towards deterritorialization, interconnectedness across borders, and the acceleration of social activity are fundamental here as well. Transnational movements, in which activists employ rapid-fire communication technologies to join forces across borders in combating ills that seem correspondingly transnational in scope (for example, the depletion of the ozone layer), offer an example of political globalization. Another would be the tendency towards ambitious supranational forms of social and economic lawmaking and regulation, where individual nation-states cooperate to pursue regulation whose jurisdiction transcends national borders no less than the cross-border economic processes that may undermine traditional modes of nation state-based regulation. Political scientists typically describe the trend towards ambitious forms of supranational organization (the European Union, for example, or North America Free Trade Association) as important recent manifestations of political globalization. The proliferation of supranational organizations has been no less conflict-laden than economic globalization, however. Critics insist that local, regional, and national forms of self-government are being rapidly supplanted by insufficiently democratic forms of global governance remote from the needs of ordinary citizens, whereas their defenders describe new forms of supranational legal and political decision as indispensable forerunners to more inclusive and advanced forms of self-government.

3. The Normative Challenges of Globalization

The wide-ranging impact of globalization on human existence means that it necessarily touches on many basic philosophical questions. At a minimum, globalization suggests that academic philosophers in the rich countries of the West should pay closer attention to the neglected voices and intellectual traditions of peoples with whom our fate is intertwined in ever more intimate ways (Dallmayr, 1998). In this section, however, we focus exclusively on the immediate challenges posed by globalization to normative political theory.

Western political theory has traditionally presupposed the existence of territorially bound communities, whose borders can be more or less neatly delineated from those of other communities. The contemporary liberal political philosopher John Rawls continues to speak of bounded communities whose fundamental structure consists of “self-sufficient schemes of cooperation for all the essential purposes of human life” (Rawls, 1993: 301). Although political and legal thinkers historically have exerted substantial energy in formulating defensible normative models of relations between states (Nardin and Mapel, 1992), they typically have relied on a clear delineation of “domestic” from “foreign” affairs. In addition, they have often argued that the domestic arena represents a normatively privileged site, since fundamental

normative ideals and principles (for example, liberty or justice) are more likely to be successfully realized in the domestic arena than in relations among states. According to one influential strand within international relations theory, relations between states are fundamentally lawless. Since the achievement of justice or democracy, for example, presupposes an effective political sovereign, the lacuna of sovereignty at the global level means that justice and democracy are necessarily incomplete and probably unattainable there. In this “Realist” view of international politics, core features of the modern system of sovereign states relegate the pursuit of western political thought's most noble normative goals primarily to the domestic arena (Morgenthau, 1954).

Globalization poses a fundamental challenge to each of these traditional assumptions. It is no longer self-evident that nation-states can be described as “self-sufficient schemes of cooperation for all the essential purposes of human life” in the context of intense deterritorialization and the spread and intensification of social relations across borders. The idea of a bounded community seems suspect given recent shifts in the spatio-temporal contours of human life. Even the most powerful and privileged political units are now subject to increasingly deterritorialized activities (for example, global financial markets) over which they have limited control, and they find themselves nested in webs of social relations whose scope explodes the confines of national borders. Of course, in much of human history social relations have transcended existing political divides. However, globalization implies a profound quantitative increase in and intensification of social relations of this type. While attempts to offer a clear delineation of the “domestic” from the “foreign” probably made sense at an earlier juncture in history, this distinction no longer accords with core developmental trends in many arenas of social activity. As the possibility of a clear division between domestic and foreign affairs dissipates, the traditional tendency to picture the domestic arena as a privileged site for the realization of normative ideals and principles becomes problematic as well. As an empirical matter, the decay of the domestic-foreign frontier seems highly ambivalent, since it might easily pave the way for the decay of the more attractive attributes of domestic political life: as “foreign” affairs collapse inward onto “domestic” political life, the relative lawlessness of the former potentially makes disturbing inroads onto the latter (Scheuerman, 1999). As a normative matter, however, the disintegration of the domestic-foreign divide probably calls for us to consider, to a greater extent than ever before, how our fundamental normative commitments about political life can be effectively achieved on a global scale. If we take the principles of justice or democracy seriously, for example, it is no longer self-evident that the domestic arena is the main site for their pursuit, since domestic and foreign affairs are now deeply and irrevocably intermeshed. In a globalizing world, the lack of democracy or justice in the global setting necessarily impacts deeply on the pursuit of justice or democracy at home. Indeed, it may no longer be possible to achieve our normative ideals at home without undertaking to do so transnationally as well.

To claim, for example, that questions of distributive justice have no standing in the making of foreign affairs represents at best empirical naivete about economic globalization. At worst, it constitutes a disingenuous refusal to grapple with the fact that the material existence of those fortunate enough to live in the rich countries is inextricably tied to the material status of the vast majority of humanity residing in poor and underdeveloped regions. Growing material inequality spawned by economic globalization is linked to growing domestic material inequality in the rich democracies (Falk, 1999). Similarly, in the context of the ongoing destruction of the ozone layer by privileged countries like Australia, Japan, and

the United States, a dogmatic insistence on the sanctity of national sovereignty risks constituting a cynical fig leaf for irresponsible activities whose impact extends well beyond the borders of the polluting countries. Ozone-depletion cries out for ambitious forms of transnational cooperation and regulation, and the refusal by the rich democracies to accept this necessity implies a failure to take the process of globalization seriously when doing so conflicts with their material interests. Although it might initially seem to be illustrative of clever *Realpolitik* on the part of the polluting nations to ward off strict cross-border environmental regulation, their stubbornness is probably short-sighted: ozone depletion will affect the children of Americans who drive gas-guzzling SUVs or use environmentally unsound air-conditioning as well as the future generations of South Africa or Afghanistan. If we keep in mind that environmental degradation probably impacts negatively on democratic politics (for example, by undermining its legitimacy and stability), the failure to pursue effective transnational environmental regulation potentially undermines democracy at home as well as abroad.

In recent years, philosophers and political theorists have been busy addressing the normative implications of our globalizing world. A lively debate about the possibility of achieving justice at the global level now pits representatives of cosmopolitanism against communitarianism. Cosmopolitans underscore our universal moral obligations to those who reside faraway and with whom we share little in the way of language, custom, or culture, arguing that claims to “justice at home” can and should be applied elsewhere as well. In this way, cosmopolitanism builds directly on the universalistic impulses of modern moral and political thought. In contrast, communitarians dispute the view that our moral obligations to foreigners possess the same status as those to members of particular communities (for example, the nation-state) of which we remain very much a part. Communitarians by no means deny the need to redress global inequality, for example, but they often express skepticism in the face of cosmopolitanism's tendency to defend significant legal and political reforms as necessary to address the inequities of a planet where eighteen million people a year die of starvation (Jones, 1999; Pogge, 2001: 9). Nor do communitarians necessarily deny that the process of globalization is real, though some of them believe its impact has been grossly exaggerated (Kymlicka, 1999). Nonetheless, they doubt that humanity has achieved a rich or sufficiently articulated sense of a common fate such that far-reaching attempts to achieve greater global justice (for example, substantial redistribution from the rich to poor) could prove successful. Cosmopolitans not only typically counter with a flurry of universalist and egalitarian moral arguments, but they also accuse communitarians of obscuring the threat posed by globalization to the particular forms of community whose ethical primacy the communitarians endorse. From the cosmopolitan perspective, the communitarian tendency to favor moral obligations to fellow members of the nation-state represents a misguided and increasingly reactionary nostalgia for a rapidly decaying constellation of political practices and institutions.

A similar intellectual divide characterizes the ongoing debate about the prospects of democratic institutions at the global level. In a cosmopolitan mode, David Held (1995) argues that globalization requires the extension of liberal democratic institutions (including the rule of law and elected representative institutions) to the transnational level. Nation state-based liberal democracy is poorly equipped to deal with deleterious side effects of present-day globalization such as ozone depletion or burgeoning material inequality. In addition, a growing array of genuinely transnational forms of activity calls out for no less intrinsically transnational modes of liberal democratic decision-making. According

to this model, “local” or “national” matters should remain under the auspices of existing liberal democratic institutions. But in those areas where deterritorialization and social interconnectedness across national borders are especially striking, new transnational institutions (for example, cross-border referenda), along with a dramatic strengthening and further democratization of existing forms of supranational authority (in particular, the United Nations), are necessary if we are to assure that popular sovereignty remains an effective principle. In the same spirit, Jürgen Habermas has tried to formulate a defense of the European Union that conceives of it as a key steppingstone towards supranational democracy. If the EU is to help succeed in salvaging the principle of popular sovereignty in a world where the decay of nation state-based democracy makes democracy vulnerable, the EU will need to strengthen its elected representative organs and better guarantee the civil, political, and social and economic rights of all Europeans (Habermas, 2001: 58-113).

In opposition to Held, Habermas, and other defenders of global democracy, communitarian-minded skeptics underscore the purportedly utopian character of such proposals, arguing that democratic politics presupposes deep feelings of trust, commitment, and belonging that remain uncommon at the transnational level. Largely non-voluntary commonalities of belief, history, and custom compose necessary preconditions of any viable democracy, and since these commonalities are missing beyond the sphere of the nation-state, global or cosmopolitan democracy is doomed to fail (Archibugi, Held, and Koehler, 1998). In an analogous vein, critics inspired by Realist theory argue that cosmopolitanism obscures the fundamentally pluralistic, dynamic, and conflictual nature of political life on our divided planet. Notwithstanding its pacific self-understanding, cosmopolitan democracy inadvertently opens the door to new and even more horrible forms of political violence. Cosmpolitanism's universalistic moral discourse not only ignores the harsh and unavoidably agonistic character of political life, but it also tends to serve as a convenient ideological cloak for terrible wars waged by political blocs no less self-interested than the traditional nation state. For these critics, the fact that the recent Allied war against Iraq, conducted as a so-called “humanitarian intervention” with the blessings of the United Nations, probably resulted in at least 220,000 civilian deaths, vividly underscores the profound dangers intrinsic to the quest for novel forms of global democracy (Zolo 1997, 24).

Bibliography

- Adams, Henry (1931), *The Education of Henry Adams* (New York: Modern Library).
- Archibugi, Daniele, Held, David, and Koehler, Martin (ed.) (1998), *Re-imagining Political Community: Studies in Cosmopolitan Democracy* (Stanford: Stanford University Press).
- Castells, Manuel (1996), *The Rise of Network Society* (Oxford, UK: Blackwell).
- Dallmayr, Fred (1998), *Alternative Visions: Paths in the Global Village* (Lanham, Md.: Rowman & Littlefield).
- Dewey, John (1927), *The Public and Its Problems* (Athens, Ohio: Swallow Press).
- Giddens, Anthony (1990), *The Consequences of Modernity* (Stanford: Stanford University Press).
- Eriksen, Thomas Hylland (2001), *Tyranny of the Moment: Fast and Slow Time in the Information Age* (London: Pluto Press).
- Falk, Richard (1999), *Predatory Globalization* (Cambridge, UK: Polity Press).

- Gurvitch, Georges (1965), *The Spectrum of Social Time* (Dordrecht, Holland: Reidel).
- Habermas, Jürgen (2001), *The Postnational Constellation: Political Essays* (Cambridge, USA: MIT Press, 2001)
- Harvey, David (1989), *The Condition of Postmodernity* (Oxford, UK: Blackwell).
- Harvey, David (1996), *Justice, Nature, & the Geography of Difference* (Oxford, UK: Blackwell).
- Heidegger, Martin (1971), “The Thing,” in *Poetry, Language, Thought* (New York: Harper & Row).
- Held, David (1995), *Democracy and the Global Order: From the Modern State to Cosmopolitan Governance* (Stanford: Stanford University Press).
- Held, David, McGrew, Anthony, Goldblatt, David, and Perraton, Jonathan (1999), *Global Transformations: Politics, Economics and Culture* (Stanford: Stanford University Press).
- Jones, Charles (1999), *Global Justice: Defending Cosmopolitanism* (Oxford: Oxford University Press).
- Kymlicka, Will (1999), “Citizenship in an Era of Globalization: A Response to Held,” in Ian Shapiro and Casiano Hacker-Cordon (eds.), *Democracy's Edges* (Cambridge, UK: Cambridge University Press).
- Kern, Stephen (1983), *The Culture of Time and Space, 1880-1918* (Cambridge, USA: Harvard University Press).
- Marx, Karl (1979), “Communist Manifesto,” in Robert Tucker (ed.), *The Marx-Engels Reader* (New York: Norton).
- McLuhan, Marshall (1964), *Understanding Media: The Extensions of Man* (New York: McGraw Hill).
- Modelski, George (1972), *Principles of World Politics* (New York: Free Press, 1972).
- Morgenthau, Hans (1954), *Politics Among Nations: The Struggle for Power and Peace* (New York: Knopf).
- Nardin, Terry and Mapel, David (ed.) (1992), *Traditions of International Ethics* (Cambridge, UK: Cambridge University Press).
- Pogge, Thomas (2001), “Priorities of Global Justice,” *Metaphilosophy* 32; 6-24.
- Rawls, John (1993), *Political Liberalism* (New York: Columbia University Press).
- Ruggie, John Gerard (1993), “Territoriality and Beyond: Problematizing Modernity in International Relations,” *International Organization* 47; 139-74.
- Scheuerman, William E. (1999), “Globalization, Exceptional Powers, and the Erosion of Liberal Democracy,” *Radical Philosophy* 93; 14-23.
- Schivelbusch, Wolfgang (1978), “Railroad Space and Railroad Time,” *New German Critique* 14; 31-40.
- Scholte, Jan Aart (1996), “Beyond the Buzzword: Towards a Critical Theory of Globalization,” in Eleonore Kofman and Gillians Young (ed.), *Globalization: Theory and Practice* (London: Pinter).
- Scholte, Jan Aart (2000), *Globalization: A Critical Introduction* (New York: St. Martin's).
- Tomlinson, John (1999), *Globalization and Culture* (Cambridge, UK: Polity Press).
- Virilio, Paul (1986), *Speed and Politics* (New York: Semiotext[e]).
- Zolo, Danilo (1997), *Cosmopolis: Prospects for World Government* (Cambridge, UK: Polity Press).

Other Internet Resources

- Global Transformations website (maintained by David Held, Political Science, London School of Economics, and Anthony McGrew, International Relations, Southampton University)

Related Entries

communitarianism | cosmopolitanism | democracy

William E. Scheuerman
scheuerm@polisci.umn.edu

First published: June 21, 2002
Content last modified: June 21, 2002

Diagrams

All of us engage in and make use of valid reasoning, but the reasoning we actually perform differs in various ways from the inferences studied by most (formal) logicians. Actual reasoning as performed by human beings typically involves information obtained through more than one medium. Formal logic, by contrast, has thus far been primarily concerned with valid reasoning which is based on information in one form only, i.e. in the form of sentences. Recently, many philosophers, psychologists, logicians, mathematicians, and computer scientists have been increasingly aware of the importance of multi-modal reasoning and, moreover, much research has been undertaken in the area of non-symbolic, especially diagrammatic, representation systems. This entry outlines the overall directions of this new research area and focuses on the logical status of diagrams in proofs, their representational function and adequacy, different kinds of diagrammatic systems, and the role of diagrams in human cognition.

1. Introduction

Diagrams or pictures probably rank among the oldest forms of human communication. They are not only used for representation but can also be used to carry out certain types of reasoning, and hence play a particular role in logic. However, sentential representation systems (e.g., first-order logic) have been dominant in the modern history of logic, while diagrams have largely been seen as only of marginal interest. Diagrams are usually adopted as a heuristic tool in exploring a proof, but not as part of a proof. It is a quite recent movement among philosophers, logicians, cognitive scientists and computer scientists to focus on different types of representation systems, and much research has been focussed on diagrammatic representation systems.

Challenging a long-standing prejudice against diagrammatic representation, those working on multi-modal

reasoning have taken different kinds of approaches which we may categorize into three distinct groups. One branch of research can be found in philosophy of mind and cognitive science. Since the limits of linguistic forms are clear to those who have been working on mental representation and reasoning, some philosophers and cognitive scientists have embraced this new direction of multi-modal reasoning with enthusiasm and have explored human reasoning and mental representation involving non-linguistic forms [Cummins 1996, Chandrasekaran *et al*. 1995]. Another strand of work on diagrammatic reasoning shows that there is no intrinsic difference between symbolic and diagrammatic systems as far as their logical status goes. Some logicians have presented case studies to prove that diagrammatic systems can be sound and complete in the same sense as a symbolic system. This type of result directly refuted a widely-held assumption that diagrams are inherently misleading, and abolished theoretical objections to diagrams being used in proofs [Shin 1994, Hammer 1995]. A third direction in multi-modal reasoning has been taken by computer scientists, whose interest is much more practical than those of the other groups. Not so surprisingly, those working in many areas in computer science - for example, knowledge representation, systems design, visual programming, GUI design, and so on - found new and exciting opportunities in this new concept of 'heterogeneous system' and have implemented diagrammatic representations in their research areas.

We have the following goals for this entry. First of all, we would like to acquaint the reader with some details of specific diagrammatic systems. At the same time, the entry will address theoretical issues involved in the topic, by exploring the nature of diagrammatic representation and reasoning in terms of expressive power and correctness. The case study presented below will not only satisfy our first goal but also provide us with solid material for the more theoretical and general discussion in the third section. As mentioned above, the topic of diagrams has attracted much attention with important results from many different research areas. Hence, our fourth section aims to introduce various approaches to diagrammatic research taken in different areas.

For further discussions, we need to clarify two related but distinct uses of the word 'diagram': diagram as internal mental representations and diagram as external representation. The following quotation from Chandrasekaran *et al*. [1995, p. xvii] succinctly sums up the distinction between internal versus external diagrammatic representations:

- *External diagrammatic representations*: These are constructed by the agent in a medium in the external world (paper, etc), but are meant as representations by the agent.
- *Internal diagrams or images*: These comprise the (controversial) internal representations that are posited to have some pictorial properties.

As we will see below, logicians focus on external diagrammatic systems, the imagery debate among philosophers of mind and cognitive scientists is mainly about internal diagrams, and research on the cognitive role of diagrams touches on both forms.

2. Diagrams as Representational Systems

The dominance of sentential representation systems in the history of modern logic has obscured several important facts about diagrammatic systems. One of them is that several well-known diagrammatic systems

were available as a heuristic tool before the era of modern logic. Euler circles, Venn diagrams, and Lewis Carroll's squares have been widely used for certain types of syllogistic reasoning [Euler 1768, Venn 1881, Carroll 1896]. Another interesting, but neglected, story is that a founder of modern symbolic logic, Charles Peirce, not only revised Venn diagrams but also invented a graphical system, Existential Graphs, which has been proven to be equivalent to a predicate language [Peirce 1933, Roberts 1973, Zeman 1964].

These existing diagrams have inspired those researchers who have recently drawn our attention to multi-modal representation. Logicians who participate in the project have explored the subject in two distinct ways. First, their interest has focused exclusively on externally-drawn representation systems, as opposed to internal mental representations. Second, their aim has been to establish the logical status of a system, rather than to explain its heuristic power, by testing the correctness and the expressive power of selective representation systems. If a system fails to justify its soundness or if its expressive power is too limited, a logician's interest in that language will fade.

In this section, we examine the historical development of Euler and Venn diagrams as a case study to illustrate the following aspects: First, this process will show us how one mathematician's simple intuition about diagramming syllogistic reasoning has gradually been developed into a formal representation system. Second, we will observe different emphases given to different stages of extension and modification of a diagrammatic system. Thirdly and relatedly, this historical development illustrates an interesting tension and trade-off between the expressive power and visual clarity of diagrammatic systems. Most importantly, the reader will witness logicians tackle the issue of whether there is any intrinsic reason that sentential systems, but not diagrammatic systems, could provide us with rigorous proofs, and their success in answering this question in the negative.

Hence, the reader will not be surprised by the following conclusion drawn by Barwise and Etchemendy, the first logicians to launch an inquiry into diagrammatic proofs in logic,

> there is no principled distinction between inference formalisms that use text and those that use diagrams. One can have rigorous, logically sound (and complete) formal systems based on diagrams. (Barwise & Etchemendy [1995], p. 214.)

This conviction was necessary for the birth of their innovative computer program *Hyperproof*, which adopts both first-order languages and diagrams (in a *multi-modal* system) to teach elementary logic courses [Barwise & Etchemendy 1994].

2.1 Euler Diagrams

Leonhard Euler, an 18th century mathematician, adopted closed curves to illustrate syllogistic reasoning [Euler 1768]. The four kinds of categorical sentences are represented by him as shown in Figure 1.

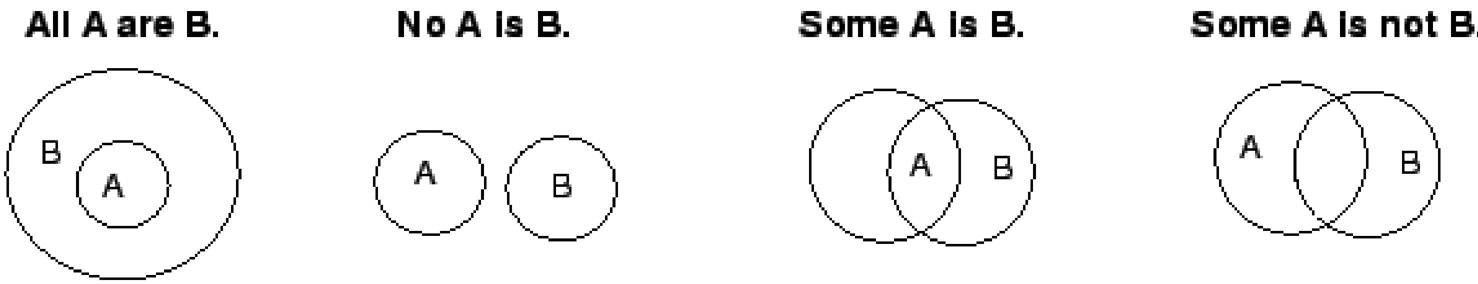


Figure 1: Euler Diagrams

For the two universal statements, the system adopts spatial relations among circles in an intuitive way: If the circle labelled 'A' is *included* in the circle labelled 'B,' then the diagram represents the information that all A is B. If there is *no overlapping* part between two circles, then the diagram conveys the information that no A is B.

This representation is governed by the following convention:[1]

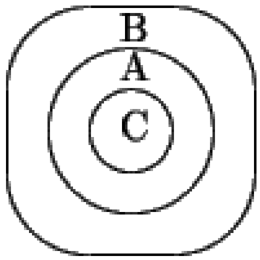
> Every object *x* in the domain is assigned a unique location, say *l*(*x*), in the plane such that *l*(*x*) is *in* region *R* if and only if *x* is a member of the set that the region *R* represents.

The power of this representation lies in the fact that an object being a member of a set is easily conceptualized as the object falling inside the set, just as locations on the page are thought of as falling inside or outside drawn circles. The system's power also lies in the fact that no additional conventions are needed to establish the meanings of diagrams involving more than one circle: relationships holding among sets are asserted by means of the same relationships holding among the circles representing them. The representations of the two universal statements, "All A are B" and "No A is B," illustrate this strength of the system.

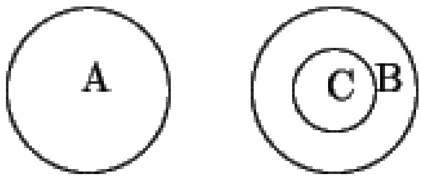
Moving on to two existential statements, this clarity is not preserved. Euler justifies the diagram of "Some A is B" saying that we can infer *visually* that something in A is also contained in B since part of area A is contained in area B.[2] Obviously, Euler himself believed that the same kind of visual containment relation among areas can be used in this case as well as in the case of universal statements. However, Euler's belief is not correct and this representation raises a damaging ambiguity. In this diagram, not only is part of circle A contained in area B (as Euler describes), but the following are true: (i) part of circle B is contained in area A (ii) part of circle A is not contained in circle B (iii) part of circle B is not contained in circle A. That is, the third diagram could be read off as "Some B is A," "Some A is not B," and "Some B is not A" as well as "Some A is B." In order to avoid this ambiguity, we need to set up several more conventions.[3]

Euler's own examples nicely illustrate the strengths and weaknesses of his diagrammatic system.

> **Example 1**. All A are B. All C are A. Therefore, all C are B.



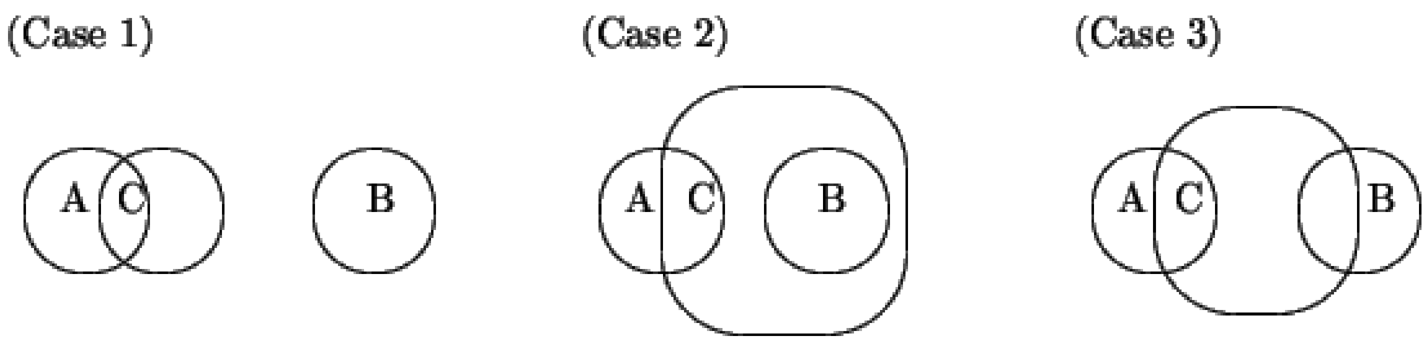
Example 2. No A is B. All C are B. Therefore, no C is A.



In both examples, the reader can easily infer the conclusion, and this illustrates visually powerful features of Euler diagrams. However, when existential statements are represented, things become more complicated, as explained above. For instance:

Example 3. No A is B. Some C is A. Therefore, Some C is not B.

No single diagram can represent the two premises, since the relationship between sets B and C cannot be fully specified in one single diagram. Instead, Euler suggests the following three possible cases:



Euler claims that the proposition ‘Some C is not B’ can be read off from all these diagrams. However, it is far from being visually clear how the first two cases lead a user to reading off this proposition, since a user might read off “No C is B” from case 1 and “All B is C” from case 2.

Hence, the representation of existential statements not only obscures the visual clarity of Euler Circles but also raises serious interpretational problems for the system. Euler himself seemed to recognize this potential problem and introduced a new syntactic device, ‘*’ (representing non-emptiness) as an attempt to repair this flaw. (Letter 105)

However, a more serious drawback is found when this system fails to represent certain compatible (that is, consistent) pieces of information in a single diagram. For example, Euler's system prevents us from drawing a single diagram representing the following pairs of statements: (i) “All A are B” and “No A is B” (which are

consistent if A is an empty set). (ii) “All A are B” and “All B are A” (which are consistent when A = B). (iii) “Some A is B” and “All A are B”. (Suppose we drew an Euler diagram for the former proposition and try to add a new compatible piece of information, i.e., the latter, to this existing diagram.) This shortcoming is closely related to Venn's motivation for his own diagrammatic system (see Section 3.1 for other shortcomings of Euler's system).

2.2 Venn Diagrams

Venn's criticism of Euler Circles is summarized in the following passage:

> The weak point in this [Euler diagrams], and in all similar schemes, consists in the fact that they only illustrate in strictness the actual relation of classes to each other, rather than the imperfect knowledge of these relations which we may possess, or may wish to convey by means of the proposition. [Venn 1881, p. 510.]

Because of its strictness, Euler's system sometimes fails in representing consistent pieces of information in a single diagram, as shown above. In addition to this expressive limitation, Euler's system also suffers other kinds of expressive limitations with respect to non-empty sets, due to topological restrictions on plane figures (see Section 3.1).

Venn's new system [1881] was to overcome these expressive limitations so that partial information can be represented. The solution was his idea of ‘primary diagrams’. A primary diagram represents all the possible set-theoretic relations between a number of sets, without making any existential commitments about them. For example, Figure 2 shows the primary diagram about sets A and B.

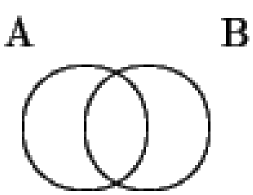


Figure 2: Venn's Primary diagrams

According to Venn's system, this diagram does not convey any specific information about the relation between these two sets. This is the major difference between Euler and Venn diagrams.

For the representation of universal statements, unlike the visually clear spatial containment relations in the case of Euler diagrams, Venn's solution is ‘to shade them [the appropriate areas] out’ ([Venn 1881], p. 122). By using this syntactic device, we obtain diagrams for universal statements as shown in Figure 3.

Figure 3: Venn's shading

Venn's choice of shading might not be absolutely arbitrary in that a shading could be interpreted as a

visualization of set emptiness. However, it should be noted that a shading is a new syntactic device which Euler did not use. This revision gave flexibility to the system so that certain compatible pieces of information may be represented in a single diagram. In the following, the diagram on the left combines two pieces of information, “All A are B” and “No A is B,” to visually convey the information “Nothing is A.” The diagram below, representing both “All A are B” and “All B are A,” clearly shows that A is the same as B:

In fact, using primary diagrams also avoids some other expressivity problems (to do with spatial properties of diagram objects) discussed below, in Section 3. Surprisingly, Venn was silent about the representation of existential statements, which was another difficulty of Euler diagrams. We can only imagine that Venn might have introduced another kind of a syntactic object representing existential commitment. This is what Charles Peirce did about twenty years later.

2.3 Peirce's extension

Peirce points out that Venn's system has no way of representing the following kinds of information: existential statements, disjunctive information, probabilities, and relations. Peirce aimed to extend Venn's system in expressive power with respect to the first two kinds of propositions, i.e., existential and disjunctive statements. This extension was completed by means of the following three devices. (i) Replace Venn's shading representing emptiness with a new symbol, ‘o’. (ii) Introduce a symbol ‘x’ for existential import. (iii) For disjunctive information, introduce a linear symbol ‘-’ which connects ‘o’ and ‘x’ symbols.

For example, Figure 4 represents the statement, ‘All A are B or some A is B’, which neither Euler's nor Venn's system can represent in a single diagram.

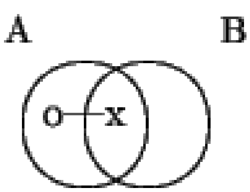
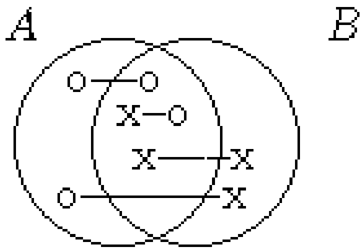


Figure 4: A Peirce diagram

The reason that Peirce replaced Venn's shading for emptiness with the symbol ‘o’ seems to be obvious: It would not be easy to connect shadings or shadings and x's in order to represent disjunctive information. In this way, Peirce increased the expressive power of the system, but this change was not without its costs.

For example, the following diagram represents the proposition ‘ *Either* all A are B and some A is B, *or* no A is B and some B is not A’:



Reading off this diagram requires more than reading off visual containment among circles (as in Euler diagrams) or shadings (as in Venn diagrams), but also requires extra conventions for reading combinations of the symbols ‘o,’ ‘x,’ and lines. Peirce's new conventions increased the expressive power of single diagrams, but the arbitrariness of its conventions and more confusing representations (for example, the above diagram) sacrificed the visual clarity which Euler's original system enjoys. At this point, Peirce himself confesses that ‘there is a great complexity in the expression that is essential to the meaning’ ([Peirce 1933], 4.365). Thus, when Peirce's revision was completed, most of Euler's original ideas about visualization were lost, except that a geometrical object (the circle) is used to represent (possibly empty) sets.

Another important contribution Peirce made to the study of diagrams starts with the following remark:

> ‘Rule’ is here used in the sense in which we speak of the ‘rules’ of algebra; that is, as a permission under strictly defined conditions. ([Peirce 1933], 4.361.)

Peirce was probably the first person to discuss rules of transformation in a non-sentential representation system. In the same way that the rules of algebra tell us which transformations of symbols are permitted and which are not, so should the rules of diagram manipulation. Some of Pierce's six rules needed more clarification and turn out to be incomplete - a problem which Peirce himself anticipated. However, more importantly, Peirce did not have any theoretical tool - a clear distinction between syntax and semantics - to convince the reader that each rule is correct or to determine whether more rules are needed. That is, his important intuition (that there could be transformation rules for diagrams) remained to be justified.

2.4 Diagrams as formal system

In [1994], Shin follows up Peirce's work in two directions. One is to improve Peirce's version of Venn diagrams, and the other is to prove the soundness and the completeness of this revised system.

Shin's work alters Peirce's modifications of Venn diagrams to achieve an increase in expressive power without such a severe loss of visual clarity. This revision is made in two stages: (i) Venn-I: retains Venn's shadings (for emptiness), Peirce's ‘*x*’ (for existential import) and Peirce's connecting line between *x*'s (for disjunctive information). (ii) Venn-II: This system, which is proven to be logically equivalent to monadic predicate logic, is the same as Venn-I except that a connecting line between diagrams is newly introduced to display disjunctive information.

Returning to one of Euler's examples we will see the contrast among these different versions clearly:

Example 3. No A is B. Some C is A. Therefore, Some C is not B.

Euler admits that no single Euler diagram can be drawn to represent the premises, but that three possible cases must be drawn. Venn's system is silent about existential statements. Now, Peirce's and Shin's systems represent these two premises in a single diagram as follows:

In the case of Shin's diagram, Venn's shading convention for emptiness, as opposed to Peirce's ‘o’, much more naturally leads the reader to the inference “Some C is not B” than in the case of Peirce's diagram.

However, Venn-I cannot express disjunctive information between universal statements or between universal and existential statements. Retaining Venn-I's expressive power, Venn-II allows diagrams to be connected by a line. Peirce's confusing looking diagram above is equivalent to the following Venn-II diagram:

In addition to this revision, Shin [1994] presented each of these two systems as a standard formal representation system equipped with its own syntax and semantics. The syntax tells us which diagrams are acceptable, that is, which are well-formed, and which manipulations are permissible in each system. The semantics defines logical consequences among diagrams. Using these tools, it is proven that the systems are sound and complete, in the same sense that some symbolic logics are.

This approach has posed a fundamental challenge to some of the assumptions held about representation systems. Since the development of modern logic, important concepts, e.g., syntax, semantics, inference, logical consequence, validity, and completeness, have been applied to sentential representation systems only. However, none of these turned out to be intrinsic to these traditional symbolic logics only. For *any* representation system, whether it is sentential or diagrammatic, we can discuss two levels, a syntactic and a semantic level. What inference rules tell us is how to manipulate a given unit, whether symbolic or diagrammatic, to another. The definition of logical consequence is also free from any specific form of a representation system. The same argument goes for the soundness and the completeness proofs. When a system is proven to be sound, we should be able to adopt it in proofs. In fact, much current research explores the use of diagrams in automated theorem proving (see Section 4 and [Barker-Plummer & Bailin 1997, Jamnik *et al*. 1999]).

2.5 Euler Circles revisited

It is interesting and important to notice that the gradual changes made from Euler Circles through to Shin's systems share one common theme: to increase both the expressive and logical power of the system so that it is sound, complete, and logically equivalent to monadic predicate logic. The main revision from Euler to Venn diagrams, introducing primary diagrams, allows us to represent partial knowledge about relations between sets. The extension from Venn to Peirce diagrams is made so that existential and disjunctive

information may be represented more effectively.

Both Venn and Peirce adopted the same kind of solution in order to achieve these improvements: to introduce new syntactic objects, that is, shadings by Venn, and x's, o's, and lines by Peirce. However, on the negative side, these revised systems suffer from a loss of visual clarity, as seen above, mainly because of the introduction of more arbitrary conventions. The modifications from Peirce to Shin diagrams concentrate on restoring visual clarity, but without loss of expressive power.

Hammer and Shin take a different path from these revisions: To revive Euler's homomorphic relation between circles and sets -- containment among circles represents the subset relation among sets, and non-overlapping of regions represents the disjoint relation -- and at the same time, to adopt Venn's primary diagrams by default. On the other hand, this revised Euler system is not a self-sufficient tool for syllogistic reasoning, since it cannot represent existential statements. For more details of this revised system, refer to [Hammer & Shin 1998].

This case study raises an interesting question for further research on diagrammatic reasoning. Throughout the different developments of Euler diagrams, increasing its expressive power and enhancing its visual clarity seem to be complementary to each other. Depending on purposes, we need to give priority to one over the other. Hammer and Shin's alternative system provides a simple model for the development of other efficient non-sentential representational systems, a topic that has been receiving increasing attention in computer science and cognitive science.

3. Consequences of Spatial Properties of Diagrams

While it is often possible to afford diagrams the same logical status as formulae (as argued above), there are still important differences (which can have ramifications for correctness of the system) between diagrams and traditional linear proof calculi. An important point to note about diagrams (as Russell did [1923]) is that spatial relations between objects in a diagram can be used to represent relations between objects in some other domain. Sequential languages (e.g., symbolic logics, natural languages), however, use only the relation of concatenation to represent relations between objects. The peculiar representational use of spatial relations in the case of diagrams is direct and intuitive, as seen in the development of Euler Diagrams above, but also has its perils - as we shall discuss. Spatial constraints, being peculiar to diagrammatic systems, can be expected to be an important source of both their strengths and weaknesses. Psychological considerations concerning human capacities for visual processing of information, and skill at qualitative spatial reasoning, also have ramifications for the effectiveness of reasoning with diagrams, but we shall not survey them here.

A particular distinguishing feature of diagrams is that they obey certain “nomic” or “intrinsic” constraints due to their use of plane surfaces as a medium of representation. The idea is that sentential languages are based on acoustic signals which are sequential in nature, and so must have a compensatingly complex syntax in order to express certain relationships - whereas diagrams, being two-dimensional, are able to display some relationships without the intervention of a complex syntax [Stenning & Lemon 2001]. Diagrams exploit this possibility - the use of spatial relations to represent other relations. The question is; how well can spatial relations and objects represent other (possibly more abstract) objects and relations?

Logical reasoning with diagrams is often carried out in virtue of their depiction of all possible models of a situation, up to topological equivalence of the diagrams (this, of course, depends on the particular diagrammatic system in use). A single diagram is often an abstraction over a class of situations, and once a suitable diagram has been constructed, inferences can simply be read-off the representation without any further manipulation. In some diagrammatic systems (e.g., Euler Circles) inference is carried out by constructing diagrams correctly and reading information off them. The complexity of using inference rules in a symbolic logic is, in these cases, replaced by the problem of drawing particular diagrams correctly.[4] For instance, an Euler Circles diagram ventures to capture relationships between sets using topological relationships between plane regions in such a way that that it depicts all the possible ways that a certain collection of set-theoretic statements could be true. This has two important consequences: (1) if a certain diagram cannot be drawn then the described situation must be impossible (termed "self-consistency"), and (2) if a certain relationship between diagram objects must be drawn, then the corresponding relation can be inferred as logically vaild. (See the numerous examples in Section 2.) This phenomenon is often termed a "free-ride" [Barwise & Shimojima 1995]. This style of diagrammatic reasoning is thus dependent on a particular representational use of diagrams - that they represent classes of models. If a particular class of models cannot be represented by a diagrammatic system, then those cases will not be taken into account in inferences using the system, and incorrect inferences might be drawn. This fact makes the representational adequacy of diagrammatic systems, restricted by their spatial nature, of paramount importance, as we shall now explore.

3.1 Limitations on diagrammatic representation and reasoning

The representational use of the spatial relations in the plane constrains diagrammatic representation, and therefore reasoning with diagrams, in certain important ways. In particular, there are topological and geometrical (let us lump them together as "spatial") properties of diagrammatic objects and relations which limit the expressive power of diagrammatic systems. For instance, in graph theory it is known that some simple structures cannot be drawn in the plane. For example, the graph G_5 is the graph consisting of 5 nodes, each joined to the other by an arc. This graph is non-planar, meaning that it cannot be drawn without at least two of the arcs crossing. This is just the sort of constraint on possible diagrams that limits the expressive power of diagrammatic systems. Now, since diagrammatic reasoning can occur by enumeration of all possible models of a situation, this representational inadequacy (a type of incompleteness) renders many diagrammatic systems incorrect if they are used for logical reasoning (e.g., see the critque of [Englebretsen 1992] in [Lemon & Pratt 1998]).

Perhaps the most simple example of this is due to Lemon and Pratt[5] (see e.g., [1997]). Consider Euler Circles -- where convex regions of the plane represent sets, and overlap of the regions represents non-empty intersection of the corresponding sets. A result of convex topology known as Helly's Theorem states (for the 2 dimensional case) that if every triple of 4 convex regions has a non-empty intersection then all four regions must have a non-empty intersection.

To understand the ramifications of this, consider the following problem:

Example 4. Using Euler Circles, represent the following premises:

$$A \cap B \cap C \neq \varnothing$$

$$B \cap C \cap D \neq \varnothing$$

$$C \cap D \cap A \neq \varnothing$$

Note that, in terms of set-theory, only trivial consequences follow from these premises. However, an Euler diagram of the premises, such as Figure 5, leads to the incorrect conclusion that $A \cap B \cap C \cap D \neq \varnothing$ (due to the quadruple overlap region in the centre of the diagram):

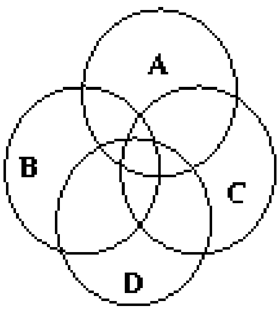


Figure 5: An Euler's Circles representation exhibiting Helly's Theorem

In other words, a user of Euler Circles is forced[6] to represent a relationship between the sets which is not logically necessary. This means both that there are logically possible situations which the system cannot represent, and that a user would make incorrect inferences if they relied on the system for reasoning. More generally, this type of result can be generated for many different types of diagrammatic system, depending on the particular spatial relations and objects which they use in representation - a research programme which is ongoing.

For example, using non-convex regions (e.g., “blobs” instead of circles) leads to a similar problem, only that non-planar graphs are involved instead of Helly's Theorem. A similar result concerns linear diagrams for syllogisms [Englebretsen 1992], where lines are used to represent sets, points represent individuals, point-line intersection represents set-membership, and intersection of lines represents set-intersection. Again, planarity constraints restrict the expressive power of the system and lead to incorrect inferences.

Atsushi Shimojima's “constraint hypothesis” perhaps best sums all this up:

> Representations are objects in the world, and as such they obey certain structural constraints that govern their possible formation. The variance in inferential potential of different modes of representation is largely attributable to different ways in which these structural constraints on representations match with the constraints on targets of representation ([Shimojima 1996a, 1999]).

3.2 Efficacy of diagrams

As discussed above, much of the interest in diagrams has been generated by the claim that they are somehow more “effective" than traditional logical representations for certain types of task. Certainly, for example, a map is a greater aid to navigation than a verbal description of a landscape. However, while there are certainly psychological advantages to be gained through the use of diagrams, they are (as in the case of Euler Circles) often ineffective as representations of abstract objects and relationships. Once a purely intuitive notion, non-psychological claims about “efficacy” of diagrammatic systems can be examined in terms of standard formal properties of languages [Lemon *et al*. 1999]. In particular, many diagrammatic systems are self-consistent, incorrect, and incomplete, and complexity of inference with the diagrams is NP-hard. By way of contrast, most sentential logics, while able to express inconsistencies, are complete and correct.

On the other hand, not being able to represent contradictions could provide us with interesting insights about the nature of diagrammatic representation. If a central goal of a language is to represent the world or a state of affairs, then representing contradictions or tautologies is called into question. Neither contradictions nor tautologies are part of the world. How can we draw a picture, or take a picture, of the contradiction that it is raining and it is not raining? How about the picture of it is either raining or not raining? Now, we seem to be much closer to Wittgenstein's classic picture theory of language [Wittgenstein 1921].

4. Diagrams and Cognition, Applications

Despite the formal limitations of some diagrammatic systems noted above, many different systems are currently used in a wide variety of contexts; logic teaching, automated reasoning, specifying computer programs, reasoning about situations in physics, graphical user interfaces to computer programs, and so on. In general, it is not yet known how effective (in the above sense) many of these diagrammatic systems are. We now give a brief survey of other diagrammatic systems and their uses, as well as the more philosophical issues raised by the debate over the status of diagrammatic reasoning.

4.1 Some other Diagrammatic Systems

It is worth noting that many mathematicians and philosophers proposed diagrammatic systems, often with a didactic motivation. Some systems, like Lewis Carroll's in “The Game of Logic" [1896] are variants on the proposals of Euler and Venn. Others, like Frege's [1879], used lines rather than plane regions. (For a description of Frege's notation, see the notation table in the entry Frege's logic, theorem and foundations for arithmetic. See also [Englebretsen 1992].) Carroll's system supercedes Venn's in that the complements of sets are explicitly represented as regions of the diagram, rather than being left as the background region against which the circles appear. This means that Carroll's system is able to draw inferences about relations between complements of properties, at the expense of representing some properties as disjoint (i.e., non-connected) regions. This shift closely mirrors the shift in logic from subject-predicate argumentation to a function-argument representation [Stenning 1999].

As a more practical side of the project, AI researchers, one of whose main concerns is the heuristic power of

a representation system in addition to its expressive power, have been debating for decades about different forms of representation [Sloman 1971, 1985, 1995]. Hence, they have welcomed discussions of the distinct role of visual reasoning and have recently hosted interdisciplinary symposiums on diagrammatic reasoning at AI conferences.[7] At the same time, realizing that human beings adopt different representation forms depending on the kinds of problems they face, some AI researchers and design theorists have practiced domain-specific approaches to bringing in problem-tailored representation forms.[8]

For instance, Harel [1988] invented highgraphs to represent system specifications in computer science. This idea has been taken up in industrial applications (e.g., UML [Booch *et al*. 1998]). Several authors have also worked on the problem of automating reasoning with diagrams in mathematical contexts. For instance, that the sum of the first n odd natural numbers is n squared is easily seen by decomposing an n x n grid into “ells” [Jamnik *et al*. 1999]. Getting computers to carry out this kind of analogical reasoning is also the task of [Barker-Plummer & Bailin 1997] amongst others.

It should also be mentioned that scientists such as chemists and physicists also use diagrams in order to perform certain computations. Feynman diagrams, for example, are used to perform calculations in sub-atomic physics. In Knot Theory (which has applications in physics [Kauffman 1991]) the three Reidemeister Moves are diagrammatic operations which make up a complete calculus for proving knots equivalent.

4.2 Diagrams as Mental Representations

Do our mental representations have diagram-like or picture-like entities as components? This question has a long history both in philosophy and in psychology, independently of each other. More recently, however, some philosophers have participated in this “imagery debate”, one of the most time-honored controversies in psychology, and some cognitive psychologists find certain epistemological theories in philosophy useful to support their views on the issue.

The nature of mental representation has been one of the perennial topics in philosophy, and we can easily trace back philosophical discussions on images and mental representation to ancient times.[9] The writings of Hobbes, Locke, Berkeley, and Hume concern themselves in large part with mental discourse, the meaning of words, mental images, particular ideas, abstract ideas, impressions, and so on. Descartes' well-known distinction between imagining and conceiving something has generated much discussion about the unique role of visual images in mental representations. The development of cognitive science in the 20th century naturally has brought certain group of philosophers and psychologists closer and we find a number of authors whose works easily belong to both disciplines [Block 1983, Dennett 1981, Fodor 1981].

Imagery based on mental inspection was the main focus in the early development of psychology until the behavioristic approach became predominant in the discipline. During the era of behaviorism, anything related to mental inspection, including images, was excluded from any serious research agenda. Finally when the topic of mental images made a comeback in psychology in 1960s, researchers adopted a more humble agenda for mental imagery than before: Not all mental representations involve imagery, and imagery is one of many ways of manipulating information in the mind. Also, thanks to the influence of behaviorism, it is

acknowledged that introspection is not enough to explore imagery, but a claim about mental imagery needs to be confirmable by experiments in order to show that we successfully externalize mental events. That is, if what a certain mental introspection tells us is genuine, then there would be observable external consequences of that mental state.

Thus the contemporary imagery debate among cognitive scientists is about the claim that picture-like images exist as mental representations and about how we interpret certain experiments.[10]

Kosslyn [1980, 1994] and other pictorialists [Shepard & Metzler 1971] present experimental data to support their position that some of our mental images are more like pictures than a linear form of language (for example, natural languages or artificial symbolic languages) in some important aspects, even though not all visual mental images and pictures are of exactly the same kind. By contrast, Pylyshyn [1981] and other descriptionalists [Dennett 1981] raise questions about the picture-like status of mental images and argue that mental images are formed out of structured descriptions. To them, mental images represent in the manner of language rather than pictures and, hence, there are no picture-like visual mental images.

Both sides of the debate sometimes used a philosophical theory as a supporting factor. For example, pictorialists in the imagery debate found the modern sense-datum theory in philosophy quite close to their point of view. By the same token, the critics of the sense-datum theory argued that the mistaken pictorial view of mental images arises mainly from our confusion about ordinary language and claimed that mental images are epiphenomena.

4.3 The Cognitive Role of Diagrams

Without being heavily involved in the imagery debate, some researchers have focused on a distinct role that diagrams or pictures - as opposed to traditional sentential forms - play in our cognitive activities. Based on the conjectures that humans adopt diagrammatic or spatial internal mental representations in their reasoning about concrete or abstract situations (see [Howell 1976, Sober 1976]), some cognitive scientists have concentrated on the functions of images or diagrams in our various cognitive activities, for example, memory, imagination, perception, navigation, inference, problem-solving, and so on. Here, the distinct nature of "visual information," which is obtained either through internal mental images or through externally drawn diagrams, has become a major topic of research. Even though most of these works assume that there are mental images (that is, they accept the pictorialists' claim), strictly speaking they do not have to commit themselves to the view that these images exist as basic units in our cognition. Descriptionalists do not have to discard discussions of the functions of images, but only need to add that these images are not primitive units stored in our memory, but formed out of structured descriptions more like the sentences of a language. (See [Pylyshyn 1981].)

A search for the distinct role of diagrams has led researchers to explore the differences among different forms of external or internal representations, and mainly between diagrammatic and sentential representations. Many important results have been produced in cognitive science. Starting from Larkin and Simon's classic case study [1987] to illustrate a difference between informational and computational equivalence among representation systems, Lindsay's work locates where this computational difference lies,

which he calls a 'non-deductive' method. As briefly pointed out above, this inference process is called 'free ride' by Barwise and Shimojiima [1995], i.e., the kind of an inference in which the conclusion seems to be read off automatically from the representation of premises. In Gurr, Lee, and Stenning [1998] and Stenning and Lemon [2001], there is an explanation of the uniqueness of diagrammatic inference in terms of a degree of 'directness' of interpretation, and it is argued that this property is relative, and hence that "some rides are cheaper than others". Having the role of graphs in mind, Wang and Lee ([1993]) present a formal framework as a guideline for correct visual languages. At this point, we are very close to applied aspects of research in multi-modal reasoning - design theory and AI research - by providing these disciplines with with computational support for visual reasoning.

Related to the issue of imagistic mental representation is the examination of the semantics of various diagrammatic systems and what they can teach us about the nature of languages in general (e.g., Goodman [1968]). For instance, Robert Cummins [1996], amongst others, argues that too much attention has been given to sentential representations and that focus on a notion of "structural representation" more akin to diagrammatic representation can help to explicate the nature of representation itself. We believe that the considerations presented above give us some empirical handle on this type of claim at least - depending on the imagistic objects and relations used, patterns of incorrect inference should be predictable and detectable. An important article, if little-known, article on this theme is [Malinas 1991]. Here Malinas explores the concepts of pictorial representation and "truth in" a picture via the notion of resemblance, and considers various semantic puzzles about pictorial representation. He develops Peacocke's "Central Thesis" of depiction ([Peacocke 1987]), where experienced similarities between properties of pictorial objects and their referents in the visual field give rise to the relation of depiction. He goes on to provide a formal semantics for pictures which is "analogous to a semantics for an ideal language".

Summary

We began by motivating the philosophical interest of diagrams, by way of their role in human reasoning and their relation to the study of language in general, and multi-modal information processing. We then explained the trade-off between expressive power and visual clarity of diagrammatic systems, by examining the historical development of diagram systems from Euler and Venn, through Peirce's work, to recent work by Shin and Hammer. It was argued that diagrammatic systems can be afforded the same logical status as traditional linear proof calculi. We then explained some of the potential pitfalls of diagrammatic representation and reasoning, by examining spatial constraints on diagrammatic systems and how they can affect correctness and expressive power. We closed by surveying other diagrams systems, the interest in diagrams generated in computer science and cognitive science, and gave an introduction to the imagery debate in the philosophy of mind.

Bibliography

References

- Allwein, G., and Barwise, J., (eds.), 1996, *Logical Reasoning with Diagrams*. Oxford: Oxford University Press.
- Barker-Plummer, D., and Bailin, S., 1997, "The Role of Diagrams in Mathematical Proofs", *Machine GRAPHICS and VISION*, 6(1): 25-56. (Special Issue on Diagrammatic Representation and Reasoning).
- Barwise, J., 1993, "Heterogeneous reasoning", in G. Mineau, B. Moulin, and J. Sowa, (eds), *ICCS 1993: Conceptual Graphs for Knowledge Representation*, volume 699 of *Lecture Notes in Artificial Intelligence*, pages 64-74, Berlin: Springer Verlag.
- Barwise, J., and Etchemendy, J., 1989, "Information, Infons, and Inference", in Cooper, Mukai, and Perry, (eds), *Situation Theory and its Applications*, volume 1, Stanford: CSLI Publications.
- -----, 1991, "Visual Information and Valid Reasoning", in Zimmerman and Cunningham, (eds), *Vizualization in Teaching and Learning Mathematics*, pages 9-24. Washington: Mathematical Association of America.
- -----, 1993, *The Language of First-Order Logic*. Stanford: CSLI Publications.
- -----, 1994, *Hyperproof*. Stanford: CSLI Publications.
- -----, 1995, "Heterogeneous Logic", in J. Glasgow, N. Hari Narayanan, and B. Chandrasekaran, (eds), *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, pages 209-232. Cambridge, MA: AAAI Press/The MIT Press.
- Barwise, J., and Shimojima, A., 1995, "Surrogate Reasoning", *Cognitive Studies: Bulletin of Japanese Cognitive Science Society*, 4(2): 7-27.
- Block, N., (ed.), 1981, *Imagery*. Cambridge, MA: MIT Press.
- -----, 1983, "Mental pictures and cognitive science", *The Philosophical Review*, 92: 499-541
- Booch, G., Rumbaugh, J., and Jacobson, I., 1998, *The Unified Modeling Language User Guide*. Addison-Wesley.
- Carroll, L., 1896, *Symbolic Logic*. Dover.
- Chandrasekaran, B., Glasgow, J., and Narayanan, N. Hari, (eds.), 1995, *Diagrammatic Reasoning: Cognitive and Computational Perspectives*. Cambridge, MA: AAAI Press/The MIT Press.
- Cummins, R., 1996, *Representations, Targets, and Attitudes*. Cambridge, MA: MIT Press.
- Dennett, D., 1981, "The nature of images and the introspective trap", in Block 1981, pages 87-107.
- Englebretsen, G.,, 1992, "Linear Diagrams for Syllogisms (with Relationals)", *Notre Dame Journal of Formal Logic*, 33(1): 37-69.
- Euler, L., 1768, *Lettres à une Princesse d'Allemagne*. St. Petersburg; l'Academie Imperiale des Sciences.
- Fodor, J., 1981, "Imagistic representation", in Block 1981, pages 63-86.
- Frege, G., 1879, *Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*, Halle am See: Louis Nebert
- Gardner, M., 1958, *Logic Machines and Diagrams*. Sussex: Harvester Press.
- Goodman, N., 1968, *Languages of Art: an approach to a theory of symbols*. London: Oxford University Press.
- Grigni, M., Papadias, D., and Papadimitriou, C., 1995, "Topological Inference", in *International Joint Conference on Artificial Intelligence (IJCAI '95)*, pages 901-907, Cambridge, MA: AAAI Press.
- Gurr, C., Lee, J., and Stenning, K., 1998, "Theories of diagrammatic reasoning: Distinguishing component problems", *Minds and Machines*, 8: 533-557.
- Hammer, E., 1995, "Reasoning with Sentences and Diagrams", *Notre Dame Journal of Formal Logic*,

35(1): 73-87

- Hammer, E., and Shin, S., 1998, "Euler's Visual Logic", *History and Philosophy of Logic*, 19: 1-29.
- Harel, D., 1988, "On Visual Formalisms", *Communications of the ACM*, 31(5): 514-530.
- Howell, R., 1976, "Ordinary Pictures, Mental Representations, and Logical Forms", *Synthese*, 33: 149-174.
- Jamnik, M., Bundy, A., and Green, I., 1999, "On Automating Diagrammatic Proofs of Arithmentic Arguments", *Journal of Logic, Language, and Information*, 8(3): 297-321.
- Kauffman, L. 1991, *Knots and Physics*. World Scientific, Singapore.
- Kosslyn, S., 1980, *Image and Mind*. Cambridge, MA: Harvard University Press.
- -----, 1994, *Image and Brain: the resolution of the imagery debate*. Cambridge, MA: MIT Press.
- Lambert, J. H., 1764, *Neues Organon*, Berlin: Akademie Verlag, 1990.
- Larkin, J., and Simon, H., 1987, "Why a Diagram is (Sometimes) Worth 10,000 Words", *Cognitive Science*, 11: 65-99.
- Lemon, O., de Rijke, M., and Shimojima, A., 1999, "Efficacy of Diagrammatic Reasoning" (Editorial), *Journal of Logic, Language, and Information*, 8(3): 265-271.
- -----, 1997b, Spatial Logic and the Complexity of Diagrammatic Reasoning. *Machine GRAPHICS and VISION*, 6(1): 89-108, 1997. (Special Issue on Diagrammatic Representation and Reasoning).
- -----, 1998, "On the insufficiency of linear diagrams for syllogisms", *Notre Dame Journal of Formal Logic*, 39(4): 573-580.
- Malinas, G., 1991, "A Semantics for Pictures", *Canadian Journal of Philosophy*, 21:3, pages 275-298.
- Narayanan, N., 1993, "Taking issue/forum: The imagery debate revisited", *Computational Intelligence*, 9(4): 303-435.
- Peacocke, C., 1987, "Depiction", *The Philosophical Review*, 96: 383-410
- Peirce, C.S., 1933, *Collected Papers*. Cambridge, MA: Harvard University Press.
- Pylyshyn, Z., 1981, "Imagery and Artificial Intelligence", in N. Block, (ed.), *Readings in Philosophy of Psychology*, volume 2, pages 170 -196. Cambridge, MA: Harvard University Press.
- Roberts, D., 1973, *The Existential Graphs of Charles S. Peirce*. The Hague: Mouton.
- Russell, B., 1923, "Vagueness", in J. Slater, (ed.), *Essays on Language, Mind, and Matter: 1919-26*, The Collected Papers of Bertrand Russell, pages 145-154. London: Unwin Hyman.
- Shepard, R., and Metzler, J., 1971, "Mental rotation of three-dimensional objects", *Science*, (171): 701-3.
- Shimojima, A., 1996a, *On the Efficacy of Representation*. Ph.D. thesis, Indiana University.
- -----, 1999, "Constraint-Preserving Representations", in L. Moss, J. Ginzburg, and M. de Rijke, (eds), *Logic, Language and Computation: Volume 2*, CSLI Lecture Notes #96, pages 296-317. Stanford: CSLI Publications.
- Shin, S., 1994, *The Logical Status of Diagrams*. Cambridge: Cambridge University Press.
- Sloman, A., 1971, "Interaction between philosophy and ai: The role of intuition and non-logical reasoning in intelligence", in *Proceedings Second International Joint Conference on Artificial Intelligence*, Morgan Kaufmann.
- -----, 1985, "Why we need many knowledge representation formalisms", in M. Bramer, (ed.), *Research and Development in Expert Systems*, pages 163-183.
- -----, 1995, "Musings on the roles of logical and nonlogical representations in intelligence", in [Chandrasekaran et al., 1995], pages 7-32.
- Sober, E., 1976, "Mental Representations", *Synthese*, 33: 101-148

- Stenning, K., 1999, "Review of *Das Spiel der Logik*, by Lewis Carrol", *Journal of Symbolic Logic*, 64: 1368-1370.
- Stenning, K., and Lemon, O., 2001, "Aligning Logical and Psychological Perspectives on Diagrammatic Reasoning", *Artificial Intelligence Review*, 15(1-2): 29-62. (Reprinted in *Thinking with Diagrams*, Kluwer, 2001.)
- Tye, M., 1991, *The Imagery Debate*, Cambridge, MA: MIT Press.
- Venn, J., 1881, *Symbolic Logic*, London: Macmillan.
- Wang, D., and Lee, J., 1993, "Visual Reasoning: its Formal Semantics and Applications", *Journal of Visual Languages and Computing*, 4: 327-356.
- Wittgenstein, L., 1921, *Tractatus Logico-Philosophicus*, B. Pears and B. McGuinness (trans), London: Routledge & Kegan Paul, 1961
- Zeman, J., 1964, *The Graphical Logic of C. S. Peirce*. Ph.D. thesis, University of Chicago.

Relevant Literature

- Barwise, J., and Hammer, E., 1994, "Diagrams and the Concept of a Logical System", in D. Gabbay, (ed), *What is a Logical System?* New York: Oxford University Press.
- Greaves, M., 2002, *The Philosophical Status of Diagrams*, Stanford: CSLI Publications.
- Hammer, E., 1998, "Semantics for Existential Graphs", *Journal of Philosophical Logic*, 27: 489-503
- -----, 1995, *Logic and Visual Information*. Studies in Logic, Language, and Computation. Stanford: CSLI Publications and FoLLI.
- Hammer, E., and Shin, S., 1996, "Euler and the Role of Visualization in Logic", in J. Seligman and D. Westerståhl, (eds), *Logic, Language and Computation: Volume 1*, CSLI Lecture Notes #58, pages 271-286. Stanford: CSLI Publications.
- Kneale, W., and Kneale, M., 1962, *The Development of Logic*. Oxford: Clarendon Press
- Lemon, O., 1997, "Review of *Logic and Visual Information*, by E. M. Hammer", *Journal of Logic, Language, and Information*, 6(2): 213-216.
- Lemon, O., 2001, "Comparing the Efficacy of Visual Languages", in D. Barker-Plummer, D. Beaver, P. Scotto di Luzio, and J. van Benthem, (eds), *Logic, Language, and Diagrams*. Stanford: CSLI Publications (in press).
- Roberts, D., 1992, "The Existential Graphs of Charles S. Peirce", *Computer and Math. Applic.*, (23): 639-663.
- Shimojima, A., 1996b, "Operational constraints in diagrammatic reasoning", in J. Barwise and G. Allwein, (eds), *Logical Reasoning with Diagrams*, New York: Oxford University Press, pages 27-48
- -----, 1996c, "Reasoning with Diagrams and Geometrical Constraints", in J. Seligman and D. Westerståhl, (eds), *Logic, Language and Computation: Volume 1*, CSLI Lecture Notes #58, pages 527-540. Stanford, CSLI Publications.
- Shin, S., 1991, "A Situation-Theoretic Account of Valid Reasoning with Venn Diagrams", in J. Barwise, J. Gawron, G. Plotkin, and S. Tutiya, (eds), *Situation Theory and its Applications: Volume 2*, CSLI Lecture Notes #26, pages 581-605. Stanford: CSLI Publications.
- -----, 1999, "Reconstituting Beta Graphs into an Efficacious System", *Journal of Logic, Language, and Information*, 8: 273-295.
- -----, 2000, "Reviving the Iconicity of Beta Graphs", in Anderson, Cheng, and Haarslev, (eds), *Theory and Application of Diagrams*, pages 58-73. Springer-Verlag

- -----, 2001a, "Multiple Readings of Peirce's Alpha Graphs", in M. Anderson and B. Meyer, (eds), *Thinking with Diagrams*. Dordrecht: Kluwer. (in press).
- -----, 2002, *The Iconic Logic of Peirce's Graphs*. Cambridge, MA: MIT Press.
- Sowa, J., 1984, *Conceptual Structures: Information Processing on Mind and Machine*. London: Addison Wesley.
- -----, 2000, *Knowledge Representation: Logical, Philosophical, Computational Foundations*. Belmont, CA: Brooks/Cole.
- Stenning, K., and Oberlander, J., 1995, "A Cognitive Theory of Graphical and Linguistic Reasoning: Logic and Implementation", *Cognitive Science*, 19(1): 97-140.
- Tufte, E., 1983, *The Visual Display of Quantitative Information*, Connecticut, Graphics Press.
- -----, 1990, *Envisioning Information*, Connecticut, Graphics Press.

Other Internet Resources

- Diagrammatic Reasoning site (Michael Anderson, Computer and Information Sciences, Fordham University)
- Visual Inference Lab (Indiana University)
- Venn Site (University of Victoria, Frank Ruskey)
- Carroll Site
- Feynman diagrams (Stanford University)
- Diagrammatics
- Knot Theory (York University)

Related Entries

cognitive science | Frege, Gottlob: logic, theorem, and foundations for arithmetic | mental imagery | Peirce, Charles Sanders: logic

Sun-Joo Shin
Sun-Joo.Shin@yale.edu
and
Oliver Lemon
lemon@csli.stanford.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

A to Z Table of Contents

First published: August 28, 2001
Content last modified: June 19, 2002

Stanford Encyclopedia of Philosophy
Notes to Diagrams

Notes

1. Note that, however natural this convention may sound, this is still an arbitrary convention. For example, Lambert and Englebretsen's systems visualize individuals as points and sets as lines [Lambert 1764, Englebretsen 1992].

2. Euler [1768], p. 233.

3. For more details, see Hammer and Shin [1998].

4. Such problems have been studied under the banner of "Topological Inference" and are nearly all NP hard [Grigni *et al*. 1995, Lemon & Pratt 1997b].

5. Now Ian Pratt-Hartmann.

6. As a practical instance of Helly's Theorem.

7. For example: Reasoning with Diagrammatic Representations: 1992 AAAI Spring Symposium; Cognitive and Computational Models of Spatial Representation: 1996 AAAI Spring Symposium; Reasoning with Diagrammatic Representations II: 1997 AAAI Fall Symposium; and Formalizing Reasoning with Visual and Diagrammatic Representations: 1998 AAAI Fall Symposium. See also Narajanan [1993].

8. The following conferences are good evidence for this effort: VISUAL '98: Visualization Issues in Formal Methods (Lisbon); International Roundtable Conference on Visual and Spatial Reasoning in Design (MIT, 1999); and Theories of Visual Languages -- Track of VL '99 : 1999 IEEE Symposium on Visual Languages.

9. See Aristotle *On the Soul* and *On the Memory and Recollection*.

10. Block [1981] is one of the best collections of important papers on this debate, and Block [1983] presents a succinct summary of this controversy and raises insightful philosophical questions about the debate. Chapters 1-4 of Tye [1991] are a good overview of both cognitive scientists' and philosophers' various positions on this issue.

Sun-Joo Shin
sjshin@nd.edu
and
Oliver Lemon
lemon@csli.stanford.edu

First published: August 28, 2001
Content last modified: August 28, 2001

Peirce's Logic

Charles Peirce's contributions to logical theory are numerous and profound. His work on relations building on ideas of De Morgan influenced Schroder, and through Schroder Peano, Russell, Lowenheim and much of contemporary logical theory. Although Frege anticipated much of Peirce's work on relations and quantification theory, and to some extent developed it to a greater extent, Frege's work remained out of the mainstream until the twentieth century. Thus it is plausible that Peirce's influence on the development of logic has been of the same order as Frege's. Further discussion of Peirce's influence can be found in Dipert (1995).

In contrast to Frege's highly systematic and thoroughly developed work in logic, Peirce's work remains fragmentary and extensive, rich with profound ideas but most of them left in a rough and incomplete form. Three of the Peirce's contributions to logic that are not as well-known as others are described below:

Among Peirce's other contributions to logic are: (i) quantification theory (see Peirce (1885) and Berry (1952)), (ii) propositional logic (see Berry (1952)), (iii) Boolean algebra (see Lewis (1918)), and (iv) "Peirce's Remarkable Theorem" (see Herzberger (1981) and Berry (1952)).

Three-Valued Logic

In three unnumbered pages from his unpublished notes written before 1910, Peirce developed what amounts to a semantics for three-valued logic. This is at least ten years before Emil Post's dissertation, which is usually cited as the origin of three-valued logic. A good source of information about these three pages is Fisch and Turquette (1966), which also includes reproductions of the three pages from Peirce's

notes.

In his notes, Peirce experiments with three symbols representing truth values: V, L, and F. He associates V with "1" and "T", indicating truth. He associates F with "0" and "F", indicating falsehood. He associates L with "1/2" and "N", indicating perhaps an intermediate or unknown value.

Peirce defines a large number of unary and binary operators on these three truth values. The semantics for the operators is indicated by truth tables. Two examples are given here. First, the bar operator (indicated here by a minus sign) is defined as follows:

x	V	L	F
-x	F	L	V

Applied to truth the bar operator yields falsehood, applied to unknown it yields unknown and applied to falsehood it yields truth.

The Z operator is a binary operator which Peirce defines as follows:

	V	L	F
V	V	L	F
L	L	L	F
F	F	F	F

Thus, the Z operator applied to a falsehood and anything else yields a falsehood. The Z operator applied to an unknown and anything but a falsehood yields an unknown. And the Z operator applied to a truth and some other value yields the other value.

The bar operator and the Z operator provide the essentials of a truth-functionally complete strong Kleene semantics for three-valued logic. In addition to these two strong Kleene operators, Peirce defines several other forms of negation, conjunction, and disjunction. The notes also provide some basic properties of some of the operators, such as being symmetric and being associative.

Calculus of Relations

Building on ideas of De Morgan, Peirce fruitfully applied the concepts of Boolean algebra to relations. Boolean algebra is concerned with operations on general or class terms. Peirce applied the same idea to

what he called "relatives" or "relative terms." While his ideas evolved continually over time on this subject, fairly definitive presentations are found in Peirce (1870) and Peirce (1883). The calculus of relatives is developed further in Tarski (1941). A history of work on the subject is Maddux (1990).

Given relative terms such as "friend of" and "enemy of" (more briefly "f" and "e"), Peirce studied various operations on these terms such as the following:

```
(union)                   friend of or enemy of
```

A pair <a,b> stands in this relation if and only if if stands in one or both of the relations. In symbols "f + e".

```
(intersection)            friend of and enemy of
```

A pair <a,b> stands in this relation if and only if if stands in both of the relations. In symbols "f . e".

```
(relative product)        friend of an enemy of
```

A pair <a,b> stands in this relation if and only if there is a c such a is a friend of c and c is an enemy of b. In symbols "f ; e".

```
(relative sum)            friend of every enemy of
```

A pair <a,b> stands in this relation if and only if a is the friend of every object c that is the enemy of b. In symbols "f , e" (Peirce uses a dagger rather than a comma)

```
(complement)              is not a friend of
```

A pair <a,b> stands in this relation if and only if <a,b> does not stand in the friend-of relation. In symbols "-f" (Peirce places a bar over the relative term).

```
(converse)                is one to whom the other is friend
```

A pair <a,b> stands in this relation if and only if b is a friend of a. In symbols "~f" (Peirce places an upwards facing semi-circle over the relative term).

Peirce presented numerous theorems involving his operations on relative terms. Examples of the numerous such laws identified by Peirce are:

```
~(r + s)     =   ~r + ~s
```

```
-(r ; s)     =   -r , -s

(r . s) , t  =   (r , s) . (r , t)
```

Peirce's calculus of relations has been criticized for remaining unnecessarily tied to previous work on Boolean algebra and the equational paradigm in mathematics. It has been frequently claimed that real progress in logic was only realized in the work of Frege and later work of Peirce in which the equational paradigm was dropped and the powerful expressive ability of quantification theory was realized.

Nevertheless, Peirce's calculus of relations has remained a topic of interest to this day as an alternative, algebraic approach to the logic of relations. It has been studied by Lowenheim, Tarski and others. Lowenheim's famous theorem was originally a result about the calculus of relations rather than quantification theory, as it is usually presented today. Some of the subsequent work on the calculus of relations is outlined in Maddux (1990).

Existential Graphs

Following his development of quantification theory, Peirce developed a graphical system for analyzing logical reasoning that he felt was superior in analytical power to his algebraic and quantificational notations. A large portion of this material is reprinted as volume 4, book 2 of Peirce (1933) and is discussed, for example, in Roberts (1964), Roberts (1973), Zeman (1964) and Hammer (1996). This system of "existential graphs" encompassed propositional logic, first-order logic with identity, higher-order logic, and modal logic.

The "alpha" portion of the system of existential graphs is concerned with propositional logic. Conjunction is indicated by juxtaposing graphs next to one another. Negation is indicated by enclosing a graph within an enclosed circle or other closed figure, which Peirce calls a "cut". Here (and occasionally in Peirce's writings) cuts will be indicated by matching parentheses. So

```
(P)
```

is equivalent to "not P", and

```
(P(Q))
```

is equivalent to "if P then Q". Observe that this is the same graph as

```
((Q)P)
```

because order is irrelevant. Juxtaposition and enclosure are the only relevant logical operations. Peirce provides five elegant rules of inference that form a complete set. The rules are Insertion in Odd, Erasure

in Even, Iteration, Deiteration, and Double Cut.

Insertion in Odd Any graph can be added to an area enclosed within an odd number of cuts.

The following table gives some examples of this rule:

```
((B))          (A (B (C)))      (A)
((B)A)         (A (B (C D)))    ((B)A)
```

In the first case from "not not B", "If A then B" is obtained. In the second case from "If A, then if B then C", "If A, then if B then both C and D" is obtained. In the third case from "not A", "If A then B" is obtained.

Erasure in Even Any graph can be erased that occurs within an even number of cuts.

The following table gives some examples of this rule:

```
(A(B))         (A (B (C)))      B(A)
(A())          (A (  (C)))      B
```

In the first case from "if A then B", "if A then true" is obtained. In the second case from "If A, then if B then C", "if A, then not C" is obtained. In the third case from "not A and B", "B" is obtained.

Iteration Any graph can be recopied to any other area that occurs within all the cuts the original occurs within.

Here are some examples of Iteration:

```
(A(B))         ( (A) (B) )      B(A)
(A(AB))        ( (A) (B(A)) )   B(A)B(A)
```

In the first case from "if A then B", "if A then both A and B" is obtained. In the second case from "If not A then B", "if not A then both B and not A" is obtained. In the third case from "B and not A", "B and not A and B and not A" is obtained.

Deiteration Any graph that could have been obtained by iteration can be erased.

Here are some examples of Deiteration:

```
(A(AB))        ( (A) (B(A)) )   B(A)B(A)
(A(B))         ( (A) (B) )      B(A)
```

These are just the exact converses of the examples of Iteration.

Double cut Two cuts can be put immediately around any graph, and two cuts immediately around any graph can be erased.

Here are some examples of Deiteration:

```
(A(B))           (A)              ((B))(A)
(((A))(B))       (((A)))          B(A)
```

From "if A then B", "if not not A, then B" is obtained. From "not A", "not not not A" is obtained. From "not not B and not A", "B and not A" is obtained.

A proof of modus ponens:

```
P  (P(Q))       premises: (i) if P then Q.  (ii) P
P  ( (Q))       Deiteration
P     Q         Double Cut
   Q            Erasure in Even
```

A proof of "if A, then if B then A":

```
(())            Double Cut
(A())           Insertion in Odd
(A( A ))        Iteration
(A( ((A)) ))    Double Cut
(A( (B(A)) ))   Insertion in Odd
```

A proof of "if not B then not A" from "if A then B":

```
(A(B))          premise
( ((A)) (B))    Double Cut
```

Finally, a proof of "if A then C" from "if A then B" and "if B then C":

```
(A(B))  (B(C))          premises
(A( B (B(C)) )) (B(C))  Iteration
(A( B (B(C)) ))         Erasure in Even
(A( B ( (C)) ))         Deiteration
(A( B    C   ))         Double Cut
(A(C))                  Erasure in Even
```

The "beta" portion of Peirce's system of existential graphs is equivalent to first-order logic with identity. It does not use variables to fill the argument places of predicates. Instead, the argument places are filled by drawn lines. Two or more such argument places can be identified (analogous to filling them with the same variable) by connecting them with a drawn line. These "lines of identity" play the role of quantifiers as well as variables. The order of interpretation of the various lines of identity and cuts of a beta graph is determined by the portions of lines of identity that are enclosed within the fewest cuts. Elements enclosed by fewest cuts are interpreted before more deeply embedded elements. Rules of inference for the beta portion are generalizations of the rules for the alpha portion. They allow lines of identity to be manipulated in various ways, such as erasing portions of lines connecting loose ends, extending loose ends, and retracting loose ends. More information about the beta portion of the system of existential graphs can be found in Roberts (1973).

Bibliography

- Berry, George D. W. (1952) "Peirce's Contributions to the Logic of Statements and Quantifiers." In P. Wiener and F. Young (Eds.) *Studies in the Philosophy of Charles S. Peirce* Cambridge: Harvard University Press.
- Burch, Robert W. (1991) *A Peircean Reduction Thesis.* Texas Tech University Press.
- Dipert, Randall (1995) "Peirce's Underestimated Role in the History of Logic." In Kenneth Ketner (Ed.) *Peirce and Contemporary Thought.* New York: Fordham University Press.
- Fisch, Max and Atwell Turquette (1966) "Peirce's Triadic Logic." *Transactions of the Charles S. Peirce Society* 11: 71 - 85.
- Hammer, Eric M. (1996) "Semantics for Existential Graphs." *Journal of Philosophical Logic* (to appear).
- Herzberger, Hans (1981) "Peirce's Remarkable Theorem." In L. W. Sumner, J. G. Slater, and F. Wilson (Eds.) *Pragmatism and Purpose: Essays Presented to Thomas A. Goudge* Toronto: University of Toronto Press.
- Lewis, C. I. (1918) *A Survey of Symbolic Logic.* Berkeley: University of California Press.
- Maddux, Roger D. (1990) "The Origin of Relation Algebras in the Development and Axiomatization of the Calculus of Relations." *Studia Logica* 3: 421 - 55.
- Peirce, Charles S. (1870) "Description of a Notation for the Logic of Relatives, Resulting from an Amplification of the Conceptions of Boole's Calculus of Logic." *Memoirs of the American Academy of Sciences* 9: 317 - 78. Reprinted in Peirce (1933).
- Peirce, Charles S. (1883) "Note B: The Logic of Relatives." In *Studies in Logic by Members of the Johns Hopkins University* Boston: Little Brown and Co. Reprinted in Peirce (1933).
- Peirce, Charles S. (1885) "On the Algebra of Logic; A Contribution to the Philosophy of Notation." *American Journal of Mathematics* 7: 180 - 202. Reprinted in Peirce (1933).
- Peirce, Charles S. (1933) "Collected Papers." Edited by Charles Hartshorne and Paul Weiss. Cambridge: Harvard University Press. In Edward Moore and Richard Robin (Eds.) *Studies in the Philosophy of Charles S. Peirce.*, Amherst, University of Massachusetts Press.
- Roberts, Don (1964) "The Existential Graphs and Natural Deduction." In Edward Moore and

Richard Robin (Eds.) *Studies in the Philosophy of Charles S. Peirce*, Amherst, University of Massachusetts Press.

- Roberts, Don (1973) *The Existential Graphs of Charles S. Peirce*. Mouton and Co.
- Tarski, Alfred (1941) "On the Calculus of Relations." *Journal of Symbolic Logic* 6: 73 - 89.
- Zeman, J. Jay (1964) *The Graphical Logic of C. S. Peirce*. Ph.D. Diss., University of Chicago.

Other Internet Resources

- The MacTutor History of Mathematics Archive

Related Entries

Peirce, Charles Sanders

Eric M. Hammer
v-erhamm@microsoft.com

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 15, 1995
Content last modified: January 2, 1996

Moral Skepticism

"Moral Skepticism" names a diverse collection of views that deny or raise doubts about various roles of reason in morality. Different versions of moral skepticism deny or doubt moral knowledge, justified moral belief, moral truth, moral facts or properties, and reasons to be moral.

Despite this diversity among the views that get called "moral skepticism", many people have very strong feelings about moral skepticism in general. One large group finds moral skepticism obvious, because they do not see how anyone could have real knowledge of anything's moral status. Others see moral skepticism as so absurd that any moral theory can be refuted merely by showing that it leads to moral skepticism. Don't you know, they ask, that slavery is morally wrong? Or terrorism? Or child abuse? Skeptics who deny that we have reason to believe or obey these moral judgments are seen as misguided and dangerous. The stridency and ease of these charges suggests misunderstanding, so we need to be more charitable and more precise.

1. Varieties of Moral Skepticism

Moral skeptics differ in many ways, but they share a common core that makes them all moral skeptics. What makes moral skepticism *moral* is that it concerns morality rather than other topics. Moral skeptics might go on to be skeptics about the external world or about other minds or about induction or about all beliefs, but

these other skepticisms are not entailed by moral skepticism alone.

What makes moral skeptics *skeptics* is that they raise doubts about common beliefs. Moral skeptics then differ in the kinds of doubts that they raise. Since general skepticism is an epistemological view about the limits of knowledge, the most central version of moral skepticism is the one that raises doubts about moral knowledge.

There are two main traditions in epistemological skepticism. Cartesian skepticism (which gets its name from Descartes, although he argued against it) is the claim that nobody ever knows anything. This claim is not made by Pyrrhonian skeptics. They also don't deny it. They have so much doubt that they refrain from taking any position on whether anything can be known.

Moral skepticism comes in two corresponding kinds. *Pyrrhonian moral skeptics* refuse to admit that some people sometimes know that some substantive moral belief is true. They doubt that moral knowledge is possible. Still, they do not go on to claim that moral knowledge is impossible. They doubt that, too. Their doubts are so extreme that they do not make any claim one way or the other about the possibility of moral knowledge. Similar views can be adopted regarding justified moral belief.

In contrast, Cartesian-style moral skeptics make definite claims about the epistemic status of moral beliefs:

> *Skepticism about moral knowledge* is the claim that nobody ever knows that any substantive moral belief is true. (Cf. Butchvarov 1989, 2.)

Some moral skeptics also make this stronger claim:

> *Skepticism about justified moral belief* is the claim that nobody is ever justified in holding any substantive moral belief.

(The relevant way of being justified is specified in Sinnott-Armstrong 1996, 17-24.) These two claims and Pyrrhonian moral skepticism all fall under the general heading of *epistemological moral skepticism*.

The relation between these two claims depends on the nature of knowledge. If knowledge implies justified belief, as is traditionally supposed, then skepticism about justified moral belief implies skepticism about moral knowledge. However, even if knowledge does require justified belief, it does not require *only* justified belief, so skepticism about moral knowledge does not imply skepticism about justified moral belief.

One reason is that knowledge implies truth, but justified belief does not. Thus, if moral beliefs cannot be true, they can never be known to be true, but they still might be justified in some way that is independent of truth. As a result, skepticism about moral knowledge is implied, but skepticism about justified moral belief is not implied, by yet another form of moral skepticism:

> *Skepticism about moral truth* is the claim that no substantive moral belief is true.

This claim is usually based on one of two more specific claims:

> *Skepticism about moral truth-aptness* is the claim that no substantive moral belief is either true or false.
>
> *Skepticism with moral error* is the claim that every substantive moral belief is false.

These last three kinds of moral skepticism are not epistemological, for they are not directly about knowledge or justification. Instead, they are usually based on views of moral language or metaphysics.

Some philosophers of language argue that sentences like "Cheating is morally wrong" are neither true nor false, because they resemble pure expressions of emotion (such as "Boo Knicks") or prescriptions for action (such as "Go Celtics"). Such expressions and prescriptions are neither true nor false. Thus, if these analogies hold, then substantive moral beliefs are also neither true nor false. They are not apt for evaluation in terms of truth. For this reason, such linguistic theories are often taken to imply skepticism about moral truth-aptness. Views of this general sort are defended by Ayer (1952), Stevenson (1944), Hare (1981), Gibbard (1990), and Blackburn (1993), although recent versions often allow some minimal kind of moral truth while denying that moral beliefs can be true or false in the same robust way as factual beliefs.

Such views are often described as *non-cognitivism*. That label is misleading, since etymology suggests that cognitivism is about cognition, which is knowledge. Since knowledge implies truth, skepticism about moral truth-aptness has implications for moral knowledge, but it is directly about truth-aptness and not about moral knowledge.

Whatever you call it, skepticism about moral truth-aptness runs into several problems. If moral assertions have no truth-value, it is hard to see how they can fit into truth-functional contexts, such as negation, disjunction, and conditionals. Such contexts are also unassertive, so they do not express the same emotions or prescriptions as when moral claims are asserted. Indeed, no particular emotion or prescription seems to be expressed when someone says, "Eating meat is not morally wrong." Expressivists and prescriptivists respond to such objections, but their responses remain controversial. (Cf. Sinnott-Armstrong 2000.)

Many moral theorists conclude that moral assertions express not only emotions or prescriptions but also *beliefs*. In particular, they express beliefs that certain acts, institutions, or people have certain moral properties (such as moral rightness or wrongness) or beliefs in moral facts (such as the fact that a certain act is morally right or wrong). This non-skeptical linguistic analysis still does not show that such moral claims can be true, since assertions can express beliefs that are false. Indeed, the falsity of all substantive moral assertions and beliefs follows if one combines the linguistic view that moral assertions express beliefs, a view of truth on which a belief cannot be true unless it corresponds to a fact, and one more thesis:

> *Skepticism about moral reality* is the claim that no moral facts or properties exist.

This metaphysical claim is, thus, a reason for skepticism with moral error, as developed by Mackie (1977). Opponents of such "error theories" often object that not all moral beliefs can be false because some moral beliefs deny other moral beliefs. However, error theorists can allow a negative moral belief (such as that

eating meat is *not* morally wrong) to be true, but only if it merely denies the truth of the corresponding positive moral belief (that eating meat *is* morally wrong). If such denials of moral beliefs are not substantive moral beliefs (as denials of astrological beliefs are not astrology), then *error theorists* can maintain that all substantive moral beliefs are false.

Error theorists and skeptics about moral truth-aptness disagree about the content of moral assertions, but they still agree that no substantive moral claim or belief is true, so they are both skeptics about moral truth. None of these skeptical theses is implied by either skepticism about moral knowledge or skepticism about justified moral belief. Some moral claims might be true, even if we cannot know or have justified beliefs about which ones are true. However, a converse implication seems to hold: If knowledge implies truth, and if moral claims are never true, then there is no knowledge of what is moral or immoral (assuming that skeptics deny the same kind of truth that knowledge requires). Nonetheless, since the implication holds in only one direction, skepticism about moral truth is still distinct from all kinds of epistemological moral skepticism.

Yet another non-epistemological form of moral skepticism answers the question "Why be moral?" This question is used to raise many different issues. Almost everyone admits that there is sometimes some kind of reason to be moral. However, many philosophers deny various universal claims, including the claims that there is always *some* reason to be moral, that there is always a distinctively *moral* (as opposed to self-interested) reason to be moral, and/or that there is always an *adequate* reason to make it irrational not to be moral or at least not irrational to be moral. These distinct denials can be seen as separate forms of *practical moral skepticism*, which are discussed in more detail in the following supplementary document:

> Supplement on Practical Moral Skepticism

Practical moral skepticism resembles epistemological moral skepticism in that both kinds of skepticism deny a role to reasons in morality. However, epistemological moral skepticism is about reasons for *belief*, whereas practical moral skepticism is about reasons for *action*. Moreover, practical moral skeptics usually deny that there is *always* enough reason for moral action, whereas epistemological moral skeptics usually deny that there is *ever* an adequate reason for moral belief. Consequently, practical moral skepticism does not imply epistemological moral skepticism. Some moral theorists do assume that a reason to believe that an act is immoral cannot be adequate unless it also provides a reason not to do that act. However, even if the two kinds of reasons are related in this way, they are still distinct, so practical moral skepticism must not be confused with epistemological moral skepticism.

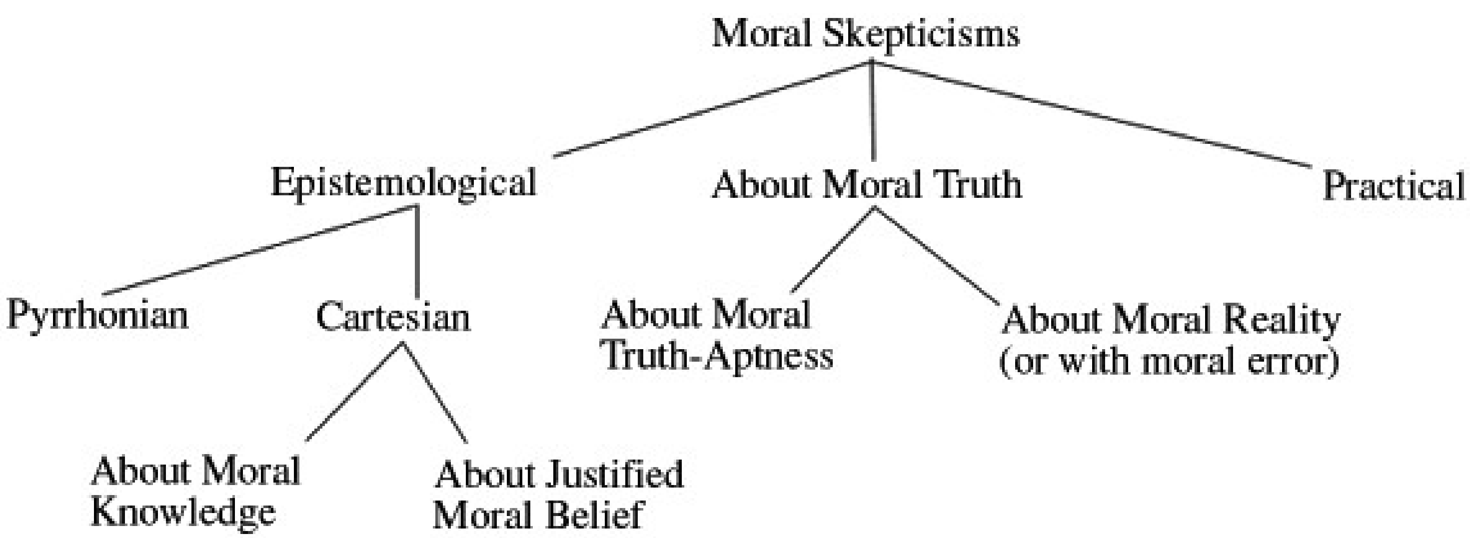
Overall, then, we need to distinguish the following kinds of epistemological moral skepticism:

> *Skepticism about moral knowledge* = nobody ever knows that any substantive moral belief is true.
>
> *Skepticism about justified moral belief* = nobody is ever justified in holding any substantive moral belief.
>
> *Pyrrhonian moral skepticism* withholds assent from skepticism about moral knowledge, skepticism about justified moral belief, and their denials.

We also need to distinguish these from several non-epistemological kinds of moral skepticism:

> *Skepticism about moral truth* = no substantive moral belief is true.
>
> *Skepticism about moral truth-aptness* = no substantive moral belief is either true or false.
>
> *Skepticism with moral error* = every substantive moral belief is false.
>
> *Skepticism about moral reality* = no moral properties or facts exist.
>
> *Practical moral skepticism* = there is not always any or adequate or distinctively moral reason to be moral.

These kinds of moral skepticism can be diagrammed as follows:



Skepticism about justified moral belief will be the primary topic for the rest of this entry, and I will refer to it henceforth simply as moral skepticism.

2. A Presumption against Moral Skepticism?

Opponents often accuse moral skepticism of leading to immorality. However, skeptics about justified moral belief can act well and be nice people. They need not be any less motivated to be moral, nor need they have (or believe in) any less reason to be moral than non-skeptics have (or believe in). Moral skeptics can hold substantive moral beliefs just as strongly as non-skeptics. Their substantive moral beliefs can be common and plausible ones. Moral skeptics can even believe that their moral beliefs are true by virtue of corresponding to an independent moral reality. All that moral skeptics deny is that their (or anyone's) moral beliefs are justified. This meta-ethical position about the epistemic status of moral beliefs need not trickle down and

infect anyone's substantive moral beliefs or actions.

Critics still argue that moral skepticism conflicts with commonsense. Most people think that they are justified in holding many moral beliefs, such as that it is morally wrong to beat your opponent senseless with a baseball bat just because she beat you in a baseball game. People also claim moral knowledge, such as when a neighbor says, "I know that it is wrong for him to spank his daughter so hard, but I don't know what I should do about it." Moral skepticism conflicts with these common ways of talking and thinking, so moral skeptics seem to owe us some argument for their claim.

Moral skepticism is, moreover, a universal and abstruse claim. It is the claim that all moral beliefs have a certain epistemic status. Normally one should not make such a strong claim without some reason. One should not, for example, claim that all astronomical beliefs are unjustified unless one has some reason for this claim. Analogously, it seems that one should not claim that all moral beliefs are unjustified unless one has some positive argument. Thus, its form, like its conflict with commonsense, seems to create a presumption against moral skepticism.

Moral skeptics, in response, sometimes try to shift the burden of proof to their opponents. Anyone who makes the positive moral claim that homosexual sodomy is morally wrong seems to need some reason for that claim, just as someone who claims that there is life on Mars seems to need evidence for that claim. If the presumption is always against those who make positive moral claims, then it is opponents of moral skepticism who must carry the burden of proof. Or, at least, moral skeptics can deny that the burden of proof is on moral skeptics. Then moral skeptics may criticize any moral belief or theory without needing to offer any positive argument for moral skepticism, and their opponents need to take moral skepticism seriously enough to argue against it. (Cf. Copp 1991.)

This controversy about burden of proof might be resolved by distinguishing Cartesian-style skepticism from Pyrrhonian moral skepticism. Cartesian-style skeptics about justified moral belief make a universal claim that conflicts with commonsense, so they seem to have the burden of arguing for their claim. In contrast, Pyrrhonian moral skeptics neither make nor deny any claim about the epistemic status of any moral belief. They simply raise doubts about whether moral beliefs are ever justified. This difference suggests that Pyrrhonian moral skeptics do not take on as much burden of proof as do Cartesian-style skeptics about justified moral belief.

3. Arguments for Moral Skepticism

Whether or not they need to, moral skeptics do offer a variety of arguments for their position. Here I will focus on arguments for Cartesian-style skepticism about justified moral belief, but essentially the same arguments could be formulated to support skepticism about moral knowledge. I will return later to Pyrrhonian moral skepticism. Also, although here I will sometimes formulate these arguments in terms of moral truth, they could be restated in ways more congenial to skeptics about moral truth-aptness.

3.1 Moral Disagreements

The simplest and most common argument for moral skepticism is based on observed facts: Smart and well-meaning people disagree about the moral permissibility of abortion, affirmative action, capital punishment, active euthanasia, nuclear deterrence, welfare reform, civil rights, and so on. Many observers generalize to the conclusion that no moral claim will be accepted by everyone.

However, all of these disagreements together still do not exclude the possibility of agreement on other moral beliefs. Maybe nobody denies that it is morally wrong to torture babies just to get sexual pleasure. Moreover, even if no moral belief is immune to disagreement, the fact that some people disagree with me does not prove that I am unjustified in holding my moral belief. I might be able to show them that I am right, or they might agree with me under ideal circumstances, where they are better informed, more thoughtful, less partial, and so on. Moral disagreements that are resolvable do not support moral skepticism, so any argument for moral skepticism from moral disagreement must show that moral disagreements are unresolvable on every issue. That will require a separate argument.

3.2 Moral Explanations

Another way to argue for moral skepticism is to cite a requirement on justified belief. On one view, we cannot be justified in believing any claim unless the truth of that claim is necessary for the best explanation of some independent fact. Some philosophers then argue that moral truths are never necessary for the best explanation of any non-moral fact. (Cf. Harman 1977.) It follows that we cannot be justified in believing any moral claim.

This argument can be countered in two ways. First, one could deny that justified belief must always involve inference to the best explanation. It is not clear, for example, that beliefs about mathematics or colors are or must be grounded in this way, although such beliefs still seem justified. (Compare Harman 1977 on mathematics and color.)

Another common response is that sometimes a moral truth is necessary for the best explanation of a non-moral fact. (Cf. Sturgeon 1985.) Hitler's vices are sometimes cited to explain his actions. Slavery's injustice has been said to explain its demise. And the fact that everyone agrees that it is morally wrong to torture babies just to get sexual pleasure might be best explained by the fact that this common belief is true.

Moral skeptics usually reply that such explanations can be replaced by non-moral descriptions of Hitler, slavery, and torture. If such replacements are always available, then moral truths are not necessary for the best explanation of anything. However, it is not clear whether or not non-moral explanations really do work as well as moral explanations in all cases. Nor is it clear whether inference to the best explanation must lie behind all justified belief.

3.3 A Regress

The next argument develops a skeptical regress. This form of argument, which derives from Sextus Empiricus (2000), is sometimes used to support the more general skeptical claim that no belief about any topic is justified. Nonetheless, it might seem to have special force within morality.

The argument's goal is to rule out all of the ways in which a person might be justified in believing something. It starts with a definition:

> A person *S* is *inferentially* justified in believing a claim that *p* if and only if what makes *S* justified is (at least in part) *S*'s ability to infer *p* from some belief of *S*.

There are, then, only two ways to be justified:

> (1) If any person *S* is justified in believing any moral claim that *p*, then *S* must be justified either inferentially or non-inferentially.

The moral skeptic denies both possibilities in turn. First:

> (2) No person *S* is ever non-inferentially justified in believing any moral claim that *p*.

Moral intuitionists and some moral contextualists deny premise (2), but moral skeptics argue that too many beliefs would be justified if people did not need any reason or inference to support their moral beliefs. If Thelma could be non-inferentially justified in believing that eating meat *is* wrong, then Louise could also be non-inferentially justified in believing that eating meat is *not* wrong, and Nick could be non-inferentially justified in believing that it is morally wrong to eat vegetables. Conflicting beliefs can sometimes both be justified, but it seems less plausible to hold that such conflicting moral beliefs are all justified without any inference when each believer knows that other people disagree. If such conflicting beliefs are not justified in the absence of a reason, and if such conflicts are pervasive enough to undermine all non-inferential justification, then premise (2) is true.

Premises (1) and (2) together imply an intermediate conclusion:

> (3) If any person *S* is justified in believing any moral claim that *p*, then *S* must be justified inferentially.

This means that, to be justified, *S* must be able to infer *p* from some other beliefs held by *S*. But which other beliefs? There are three main possibilities:

> (4) If any person *S* is inferentially justified in believing any moral claim that *p*, then S must be justified by an inference with either (a) no normative premises or (b) some normative premises but no moral premises or (c) some moral premises.

To the first possibility, moral skeptics respond with a variation on the maxim that you can't get "ought" from "is":

> (5) No person *S* is ever justified in believing any moral claim that *p* by an inference with no normative premises.

Naturalists in moral epistemology deny (5) when they try to derive a conclusion that an act is morally wrong from purely non-normative features of the act. However, moral skeptics retort that such derivations always depend on a suppressed premise that all acts with those features are morally wrong. Such a suppressed premise seems moral and, hence, normative. If so, the naturalist's inference does not really work without any normative premises. Naturalists still might invoke inferences to the best moral explanation, but then moral skeptics can deny that any moral hypothesis provides the best explanation independently of prior moral assumptions.

The next possibility is to justify a moral conclusion with an inference whose premises are not moral but are normative in another way. This approach, which is adopted by contractarians among others, can be called *normativism*. Normativists usually start with premises about rationality and impartiality that are each supposed to be normative but morally neutral. If rational impartial people under relevant circumstances would agree to certain moral standards, this is supposed to show that the corresponding moral beliefs are true or justified.

One problem for this general approach is that different theories of rationality, impartiality, and relevant circumstances are all questionable and lead to contrary moral beliefs. This suggests that such theories are not morally neutral, so these derivations do not avoid moral premises. Other arguments from non-moral norms to moral conclusions run into similar problems. Moral skeptics conclude that:

> (6) No person *S* is ever justified in believing any moral claim that *p* by an inference with some normative premises but no moral premises.

Premises (4)-(6) imply another intermediate conclusion:

> (7) If any person *S* is justified in believing any moral claim that *p*, then *S* must be justified by an inference with some moral premise.

In short, moral beliefs must be justified by moral beliefs.

This creates a problem. Although the justifying beliefs must include some moral beliefs, not just any moral beliefs will do:

> (8) No person *S* is ever justified in believing a moral claim that *p* by an inference with a moral premise unless *S* is also justified in believing that moral premise itself.

Premise (8) is denied by some contextualists, who claim that, even if a moral belief is not justified, if it is shared within a certain social context, then it may be used to justify other moral beliefs. However, moral skeptics reply that social contexts are often corrupt, and no social context by itself can show that a moral belief is true, reliable, or, hence, justified in the relevant way.

But then how can moral premises be justified? Given (7)-(8), the moral premises must be justified by inferring them from still other moral beliefs which must also be justified by inferring them from still other moral beliefs, and so on. To justify a moral belief thus requires a branching tree or chain of justifying beliefs

or premises, which must have one of two forms:

> (9) If any person *S* is justified in believing any moral claim that *p*, then *S* must be justified by a chain of inferences that either goes on infinitely or circles back to include *p* itself as an essential premise.

The first of these two alternatives is almost never defended, since most accept:

> (10) No person *S* is ever justified in believing any moral claim that *p* by a chain of inferences that goes on infinitely.

Moral skeptics also deny the other possibility:

> (11) No person *S* is ever justified in believing any moral claim that *p* by a chain of inferences that includes *p* as an essential premise.

Any argument that includes its conclusion as a premise will be valid. However, anyone who doubts the conclusion will have just as much reason to doubt the premise. So, according to skeptics, nothing is gained when a premise just restates the belief to be justified.

Premise (11) is opposed by moral coherentists. Recent coherentists emphasize that they do not infer a belief from itself in a linear way. Instead, a moral belief is supposed to be justified because it coheres in some way with a body of beliefs that is coherent in some way. Still, moral skeptics deny that coherence is enough to make a moral belief justified. One reason is that the internal coherence of a set of beliefs is not evidence of any relation to anything outside the beliefs. Another reason is that every belief — no matter how ridiculous — can cohere with some body of beliefs that is internally coherent. Because so many incompatible systems seem coherent, moral skeptics deny that coherence alone is sufficient to make beliefs justified.

Now the moral skeptic can draw a final conclusion. (9)-(11) imply:

> (12) No person is ever justified in believing any moral claim.

This is skepticism about justified moral beliefs.

Many opponents find this conclusion implausible, but the regress argument is valid. Hence, its conclusion cannot be avoided without denying one of its premises. Different opponents of moral skepticism deny different premises, as I indicated. However, it remains to be seen whether any of these responses to the regress argument is defensible in the end.

3.4 Skeptical Hypotheses

The final kind of argument derives from René Descartes (1979). I do not seem justified in believing that what I see is a lake if I cannot rule out the possibility that it is a bay or a bayou. Generalizing, if there is any

contrary hypothesis that I cannot rule out, then I am not justified in believing that what I see is a lake. This is supposed to be a common standard for justified belief. When this principle is applied thoroughly, it leads to skepticism. All a skeptic needs to show is that, for each belief, there is some contrary hypothesis that cannot be ruled out. It need not be the same hypothesis for every belief, but skeptics usually buy wholesale instead of retail, so they seek a single hypothesis that is contrary to all (or many common) beliefs and which cannot be ruled out in any way.

The famous Cartesian hypothesis is of a demon who deceives me in all of my beliefs about the external world, while also ensuring that my beliefs are completely coherent. This possibility cannot be ruled out by any experiences or beliefs, because of how the deceiving demon is defined. This hypothesis is also contrary to my beliefs about the lake. So my beliefs about the lake are not justified, according to the above principle. And there is nothing special about my beliefs about the lake. Everything I believe about the external world is incompatible with the deceiving demon hypothesis. Skeptics conclude that no such belief is justified.

This argument is often dismissed on the grounds that there is no reason to believe in a deceiving demon or that nobody really doubts whether there is an external world. In contrast, some people do seem to adopt a parallel skeptical hypothesis in morality:

> Moral Nihilism = Nothing is morally wrong.

Moral nihilism here is not about what is semantically or metaphysically possible. It is just a substantive, negative, existential claim that there does not exist anything that is morally wrong. This thesis has been supported by various reasons, including the pervasiveness of moral disagreement and our supposed ability (with the help of sociobiology and other sciences) to explain moral beliefs without reference to moral facts. Since people do take moral nihilism seriously and even argue for it (Joyce 2001), moral nihilism cannot be dismissed as readily as Descartes's deceiving demon.

Moral skeptics can then argue that the definition of moral nihilism forestalls any refutation. Since moral nihilists question all of our beliefs in moral wrongness, they leave us with no starting points on which to base arguments against them without begging the question at issue. If this trick works, then it fits right into a skeptical hypothesis argument.

This argument is clearest when applied to an example. If nothing is morally wrong, as moral nihilists claim, then it is not morally wrong to torture babies just for fun. So, according to the general principle above, one must be able to rule out moral nihilism in order to be justified in believing that torturing babies just for fun is morally wrong. Moral skeptics conclude this moral belief is not justified. More precisely:

> (1) I am not justified in believing the denial of moral nihilism.
>
> (2) I am justified in believing that [(*p*)"It is morally wrong to torture babies just for fun" entails (*q*) the denial of moral nihilism].
>
> (3) If I am justified in believing that *p*, and I am justified in believing that *p* entails *q*, then I am justified in believing that *q*.

(4) Therefore, I am not justified in believing that it is morally wrong to torture babies just for fun.

This moral belief is not especially problematic in any way. It seems as obvious as any moral belief. So the argument can be generalized to cover any moral belief. Moral skeptics conclude that no moral belief is justified.

There are two main responses to such skeptical hypothesis arguments. First, some anti-skeptics deny (1) and claim that skeptical hypotheses can be ruled out somehow. They might argue that moral nihilism is internally inconsistent or meaningless. If so, it can be ruled out by logic and semantics alone. However, moral nihilism does seem consistent and meaningful, according to all plausible theories of moral language, including expressivism, realism, and constructivism (Sinnott-Armstrong 1995). Moral nihilism is also not subject to the kind of argument that Putnam (1981) deploys against more general skeptical scenarios. Anti-skeptics still might argue that moral nihilism is incompatible with some non-moral facts or observations or their best explanations. If so, it can be ruled out by arguments with only non-moral premises. However, all such attempts to cross the dreaded is-ought gap are questionable. A third way to rule out moral nihilism would be based on common moral beliefs that are incompatible with moral nihilism. However, just as it would beg the question to use common beliefs about the external world to rule out a deceiving demon hypothesis, so it would also beg the question to argue against moral nihilism on the basis of common moral beliefs — no matter how obvious those beliefs might seem to us, and no matter how well these common beliefs cohere together. Moral skeptics conclude that there is no way to rule out moral nihilism, just as premise (1) claims.

Another recent response is to deny premise (3). This is a principle of *closure*. Since a belief entails the denial of every contrary hypothesis, this closure principle in effect says that I cannot be justified in believing *p* unless I am justified in denying every hypothesis contrary to *p* — that is, unless I can rule out *all* contrary hypotheses. This principle has been denied by relevant alternative theorists, who claim instead that only relevant hypotheses need to be ruled out. On this theory, if skeptical hypotheses are not relevant, then a belief that it is morally wrong to torture babies just for fun can be justified, even if the believer cannot rule out moral nihilism.

For this response to have force, however, opponents of moral skepticism need to say why moral nihilism is irrelevant. It seems relevant, for the simple reason that it is directly contrary to the moral belief that is supposed to be justified. Moreover, there can be reasons to believe in moral nihilism. Some people are led to moral nihilism by the absence of any defensible theory of morality. If consequentialism is absurd or incoherent, as some critics argue, and if deontological restrictions and permissions are mysterious and unfounded, as their opponents argue, then some people might believe moral nihilism for reasons similar to those that led scientists to reject phlogiston. Another basis for moral nihilism cites science. If all of our moral beliefs can be explained by sociobiology and/or other social sciences without assuming that any moral belief is true, then some might accept moral nihilism for reasons similar to those that lead many people to reject witches or elves. The point is not that such reasons for moral nihilism are adequate. The point here is only that there is enough prima facie reason to believe moral nihilism that it cannot be dismissed as irrelevant on this basis. If moral nihilism is relevant, and if closure holds for all or at least relevant alternatives, and if moral nihilism cannot be ruled out in any way, then moral skepticism seems to follow.

3.5 Relations Among the Arguments

These arguments for moral skepticism differ in many ways, but they seem mutually supportive. One crucial premise in the skeptical hypothesis argument claims that nothing can rule out moral nihilism. The best way to support that premise is to criticize each method for ruling out moral nihilism. That is just one instance of what the regress argument does more generally. The argument from moral explanations excludes yet another way to rule out moral nihilism. So, if these other arguments work, they support a crucial premise in the skeptical hypothesis argument.

Conversely, one crucial premise in the regress argument claims that no moral belief can be justified non-inferentially. Another crucial premise, (8), claims that an inference cannot justify its conclusion unless its premises are justified. These premises claim, in effect, that a moral belief needs a certain kind of justification. One way to establish this need is to point to a contrary possibility that is not yet ruled out. That is what the skeptical hypothesis argument does. Another way to confirm this need is to show that the moral belief is controversial. That is what the argument from moral disagreement does. Thus, if these other arguments work, they support a crucial premise in the regress argument.

To skeptics, this mutual support might seem desirable. Anti-skeptics, however, might object that this mutual support makes the arguments jointly circular. In the end, the force of the arguments depends on the defensibility of non-skeptical views in moral epistemology. If moral intuitionism, coherentism, naturalism, or normativism works to justify some moral beliefs and/or to rule out moral nihilism, then this will undermine the crucial premises in the arguments for moral skepticism. But that remains to be seen.

4. Pyrrhonian Moral Skepticism

Although the arguments for moral skepticism are hard to refute, most people reject their conclusion. This makes it natural to seek some compromise. Various compromises have been proposed, but here I will focus on one in the Pyrrhonian tradition.

This Pyrrhonian position can be explained in terms of contrast classes, which should be familiar from shopping: Are jumbo shrimp large? An answer of "Yes" or "No" would be too simple. Jumbo shrimp are large for shrimp, but they are not large for seafood. Analogously, someone can be justified in believing a claim out of one contrast class, even if the same person is not justified in believing the same claim out of a different contrast class. For example, suppose a father sees an animal in a zoo and believes it to be a zebra. If the father has adequate evidence that the animal is not a lion or a horse, then the father can be justified in believing that it is a zebra out of the contrast class {lion, horse, zebra}. Nonetheless, the father still might not have any evidence that the animal is not a mule painted to look like a zebra. Then the father is not justified in believing that the animal is a zebra out of the contrast class {lion, horse, zebra, painted mule}.

The same situation arises with moral beliefs. The father might be justified in believing that he should tell his children the truth rather than lying to them, even if the father is not justified in believing that he should tell his children the truth as opposed to keeping quiet. Or someone might be justified in favoring Kantian moral

theory over act-utilitarianism, because of counterexamples to act-utilitarianism, without being justified on that basis in favoring Kantian moral theory over rule-utilitarianism, if that alternative is not subject to the same counterexamples.

More generally, we can distinguish two contrast classes:

> The *extreme contrast class* for a moral belief that *p* includes every moral claim that is contrary to *p*, including moral nihilism.
>
> The *modest contrast class* includes all and only those contrary moral beliefs that most people would take seriously in an ordinary discussion.

Since most people do not take moral nihilism seriously in ordinary discussions, the modest contrast class does not include moral nihilism. Thus, anyone who can rule out all other members of the modest contrast class but cannot rule out moral nihilism is justified in believing the moral claim out of the modest contrast class but not out of the extreme contrast class.

These classes enable us to distinguish two versions of moral skepticism:

> *Skepticism about modestly justified moral belief* is the claim that nobody is ever justified out of the modest contrast class in holding any substantive moral belief.
>
> *Skepticism about extremely justified moral belief* is the claim that nobody is ever justified out of the extreme contrast class in holding any substantive moral belief.

The latter but not the former follows if nobody can ever rule out moral nihilism, but some believers sometimes can rule out all other members of the modest contrast class.

Critics will ask, "If someone is justified out of the modest contrast class but not out of the extreme contrast class, is this believer just plain justified (period or without qualification)?" That, of course, depends on what it means to say that a believer is justified (without qualification). On one plausible account, to say that a believer is justified (without qualification) is to say that the believer is justified out of the *relevant* contrast class. But which contrast class is relevant when?

Contextualists say that the modest contrast class is relevant in everyday contexts, such as hospital ethics committees, where it would be seen as a distraction to discuss moral nihilism. Nonetheless, the extreme contrast class is said to be relevant in philosophical contexts, such as philosophy classes where moral nihilism is taken seriously. This allows contextualists to hold that a doctor in a hospital ethics committee is justified in believing a moral claim that a philosophy student with the same evidence would not be justified in believing.

Problems arise when contexts cross. Consider a philosophy student who says that the doctor on the ethics committee is not justified in believing the moral claim. Is the student's contrast class (with moral nihilism) or

the doctor's contrast class (without moral nihilism) really relevant to the student's judgment about the doctor's belief? And what if the doctor says that the student really is justified while in the philosophy class? When epistemic assessments cross contexts in such ways, sometimes the believer's context seems relevant, but sometimes the assessor's context seems relevant, so it is hard to see any basis for claiming that either context or either contrast class really is the relevant one for assessing whether the believer really is justified (without qualification).

Such paradoxes lead Pyrrhonian moral skeptics to renounce all claims about which contrast class is really relevant. They can still talk about whether someone is justified in believing a moral claim out of a specified contrast class, but they refuse to take any position on whether the believer is justified (without qualification). Pyrrhonian moral skeptics can then (i) accept skepticism about extremely justified moral belief but (ii) deny skepticism about modestly justified moral belief and (iii) refuse to either assert or deny (Cartesian-style) skepticism about any moral belief being justified (without qualification).

Whether or not this view is finally defensible, the point here is just that such a Pyrrhonian compromise is available and attractive to those who want to avoid Cartesian-style moral skepticism but see no way to refute it. There are also other possible compromises that combine different strands in moral skepticism. That is what makes it so fascinating to study this important view.

Bibliography

- Ayer, A. J., 1952, *Language, Truth, and Logic*, New York: Dover. (First edition originally published in 1935.)
- Bambrough, Renford, 1979, *Moral Skepticism and Moral Knowledge*, London: Routledge.
- Blackburn, Simon, 1993, *Essays in Quasi-Realism*, New York: Oxford University Press.
- Brink, David, 1989, *Moral Realism and the Foundations of Ethics*, Cambridge: Cambridge University Press.
- Butchvarov, Panayot, 1989, *Skepticism in Ethics*, Bloomington and Indianapolis; Indiana University Press.
- Copp, David, 1991, "Moral Skepticism", *Philosophical Studies*, 62: 203-233.
- Descartes, Rene, 1979, *Meditations on First Philosophy*, translated by D. C. Cress, Indianapolis: Hackett. (Originally published in 1641.)
- Gibbard Allan, 1990, *Wise Choices, Apt Feeling*, Cambridge: Harvard University Press.
- Hare, R. M., 1981, *Moral Thinking*, Oxford: Clarendon Press.
- Harman, Gilbert, 1977, *The Nature of Morality*, New York: Oxford University Press.
- Joyce, Richard, 2001, *The Myth of Morality*, Cambridge: Cambridge University Press.
- Mackie, J. L., 1977, *Ethics: Inventing Right and Wrong*, New York: Penguin.
- Putnam, Hilary, 1981, *Reason, Truth, and History*, Cambridge: Cambridge University Press.
- Russell, Bruce, 1988, "Two Forms of Ethical Skepticism", in L. Pojman, ed., *Ethical Theory*, Belmont, Cal.: Wadsworth.
- Sextus Empiricus, 2000, *Outlines of Scepticism*, translated by Julia Annas and Jonathan Barnes, Cambridge: Cambridge University Press.
- Sidgwick, Henry, 1966, *Methods of Ethics*, New York: Dover. (First edition originally published in 1874.)

- Singer, Marcus, 1973, "Moral Skepticism", in *Skepticism and Moral Principles*, C. Carter (ed.), Evanston, Ill.: New University Press.
- Sinnott-Armstrong, Walter, 1995, "Nihilism and Skepticism about Moral Obligations", *Utilitas* 7, 217-236.
- -----, 1996, "Moral Skepticism and Justification", in *Moral Knowledge? New Readings in Moral Epistemology*, W Sinnott-Armstrong and M. Timmons (eds.), New York: Oxford University Press.
- -----, 2000, "Expressivism and Embedding", *Philosophy and Phenomenological Research* 61, 677-693.
- Stevenson, Charles, 1944, *Ethics and Language*, New Haven: Yale University Press.
- Sturgeon, Nicholas, 1985, "Moral Explanations", in *Morality, Reason, and Truth*, D. Copp and D. Zimmerman, (eds.), Totowa, N.J.: Rowman and Allanheld.
- Williams, Bernard, 1985, *Ethics and the Limits of Philosophy*, Cambridge: Harvard University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

cognitivism vs. non-cognitivism, moral | Descartes, René: epistemology | knowledge: analysis of | moral epistemology | moral realism | moral relativism | non-naturalism, moral | skepticism: ancient

Walter Sinnott-Armstrong
wsa@dartmouth.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 14, 2002
Content last modified: June 14, 2002

Stanford Encyclopedia of Philosophy
Supplement to Moral Skepticism

Practical Moral Skepticism

Practical moral skepticism answers the common question, "Why be moral?" This question, like many philosophical questions, is too short to be clear. It can be expanded and explained in several different ways.

The first word that needs to be clarified is "Why". This interrogative asks for a reason, but reasons are understood in different ways. Some philosophers suggest that all reasons are self-interested. Then the question "Why be moral?" asks "Why is it in my interest to be moral?" Others, in contrast, argue that some reasons concern effects on others (rather than oneself) or do not concern any effects on anyone. Then the question "Why be moral?" is not just asking why it is in my interest to be moral. Instead, it asks what, if anything, makes it irrational to be immoral or at least keeps it from being irrational to be moral.

The next clarification concerns the phrase "be moral". The question "Why be moral?" might seem to ask, "Why should I be a moral person?" This should be distinguished from the question, "Why should I do moral acts?", if there can be isolated cases where I have no reason not to do a particular immoral act (such as hurting this friend or cheating on this exam) while I still have reason not to want a wider tendency to do immoral acts (such as hurting and cheating regularly). Only such wider tendencies or character traits make someone immoral as a person, but it seems harder to imagine how there could be no reason to avoid such widespread tendencies. If immoral tendencies always directly or indirectly hurt other people, and if some reasons are facts about interests of other people, then there will always be reason not to be an immoral person. Even if reasons are restricted to self-interest, widespread tendencies to immorality might always be dangerous to the agent's self-interest, and then, again, there will always be reason not to be an immoral person. Practical moral skeptics might try to describe cases where a widespread tendency to immorality is in one's self-interest, but critics will respond by calling such cases unrealistic. Whether or not a realistic example can be found, the focus on immoral people or character traits at least makes it easier to argue that real agents always have some reason to be moral people.

The other question, "Why should I do moral acts?", can still be interpreted in different ways, including "Why should I do acts that are morally good?" or "Why should I do acts that are morally required?" These questions are distinct if some acts, such as giving to a particular charity, are morally good but not morally required. If I have no reason to do such an act, then I do not always have a reason to do what is morally good, but I still might always have a reason to do what is morally required. For simplicity, the rest of this entry will focus on practical moral skepticism about what is morally required.

What such practical moral skeptics deny is that I always have reason to do what is morally required. In other words, they deny that I always have reason not to do what is morally wrong. These claims are

equivalent because it is morally wrong not to do what is morally required. Practical moral skeptics do not deny that there is *sometimes* reason not do what is morally wrong. After all, some wrongdoers are caught and punished. However, practical moral skeptics can still deny that there is *always* reason to do what is morally required or to avoid what is morally wrong.

How much reason? Some practical moral skeptics claim that sometimes there is no reason at all to do what is morally required. If all reasons are self-interested, this means that sometimes doing what is morally required does not serve the agent's interest in any way. That extreme position would be refuted if doing what is morally wrong always creates even a slight risk of some negative repercussion. A more plausible and common version of practical moral skepticism denies, instead, that there is always an adequate (or non-overridden) reason to do what is morally required. To establish this position, practical moral skeptics need only one case where there is overriding reason not to do what is morally required.

It is not hard to imagine such a case if reasons are restricted to self-interest. Just consider an agent who would receive great satisfaction from killing another person whom he hates and whom he can kill without cost because he will die soon anyway. Killing then serves that agent's self-interest, even if it is still morally wrong. Other cases would work as well. If overall self-interest ever conflicts with moral requirements, then there will be overriding reason not to do what is morally required, on the assumption that all reasons are self-interested.

Without that assumption, such practical moral skepticism becomes much less plausible. If harms to others give agents reasons for and against actions, then this agent has some reason not to kill the victim. If that reason overrides the agent's reason to kill, then the agent will not have an adequate reason to kill. Of course, the agent's reason not to kill might also not be overriding. The reasons might be equal or incomparable in some way, in which case each is adequate, but neither is overriding.

This position is closely related to the claim that, when self-interest conflicts with moral requirements, neither alternative is irrational. If so, it is not always irrational to do what is morally wrong, but it still might never be irrational to do what is morally required. (Cf. Sidgwick 1966.) It does not matter much whether this position is classified as a version of practical moral skepticism. It is skeptical insofar as it denies that immoral actions are always irrational. It is anti-skeptical insofar as it claims that moral actions are never irrational.

Walter Sinnott-Armstrong
wsa@dartmouth.edu

Return to Moral Skepticism

First published: June 14, 2002

Content last modified: June 14, 2002

Common Knowledge

A proposition *A* is *mutual knowledge* among a set of agents if each agent knows that *A*. Mutual knowledge by itself implies nothing about what, if any, knowledge anyone attributes to anyone else. Suppose each student arrives for a class meeting knowing that the instructor will be late. That the instructor will be late is mutual knowledge, but each student might think only she knows the instructor will be late. However, if one of the students says openly "Peter told me he will be late again," then each student knows that each student knows that the instructor will be late, each student knows that each student knows that each student knows that the instructor will be late, and so on, *ad infinitum.* The announcement made the mutually known fact *common knowledge* among the students.

Common knowledge is a phenomenon which underwrites much of social life. In order to communicate or otherwise coordinate their behavior successfully, individuals typically require mutual or common understandings or background knowledge. Indeed, if a particular interaction results in "failure", the usual explanation for this is that the agents involved did not have the common knowledge that would have resulted in success. If a married couple are separated in a department store, they stand a good chance of finding one another because their common knowledge of each others' tastes and experiences leads them each to look for the other in a part of the store both know that both would tend to frequent. Since the spouses both love cappuccino, each expects the other to go to the coffee bar, and they find one another. But in a less happy case, if a pedestrian causes a minor traffic jam by crossing against a red light, she explains her mistake as the result of her not noticing, and therefore not knowing, the status of the traffic signal that all the motorists knew. The spouses coordinate successfully given their common knowledge, while the pedestrian and the motorists miscoordinate as the result of a breakdown in common knowledge.

Given the importance of common knowledge in social interactions, it is remarkable that only quite recently have philosophers and social scientists attempted to analyze the concept. David Hume (1740) was perhaps the first to make explicit reference to the role of mutual knowledge in coordination. In his account of convention in *A Treatise of Human Nature,* Hume argued that a necessary condition for coordinated activity was that agents all know what behavior to expect from one another. Without the requisite mutual knowledge, Hume maintained, mutually beneficial social conventions would disappear. Much later, J. E. Littlewood (1953) presented some examples of common-knowledge-type reasoning, and Thomas Schelling (1960) and John Harsanyi (1967-1968) argued that something like common knowledge is needed to explain certain inferences people make about each other. Yet it was David Lewis (1969) who first gave an explicit analysis of common knowledge in the monograph *Convention.* Stephen Schiffer (1972), Robert Aumann (1976), and Gilbert Harman (1977) independently gave alternate definitions of common knowledge. Jon Barwise (1988, 1989) gave a precise formulation of Harman's intuitive account.

Schiffer's analysis of common knowledge as a *hierarchy* of epistemic claims has become standard in the philosophical and social science literature. Lewis', Aumann's, and Barwise's accounts all imply the hierarchical account. In some contexts, Schiffer's, Aumann's, and Barwise's definitions of common knowledge are more convenient to use than Lewis' original definition. More recently, Margaret Gilbert (1989) proposed a somewhat different account of common knowledge which she argues is preferable to the standard account. Others have developed accounts of mutual knowledge, *approximate common knowledge*, and *common belief* which require less stringent assumptions than the standard account, and which serve as more plausible models of what agents know in cases where strict common knowledge seems impossible (Brandenburger and Dekel 1987, Stinchcombe 1988, Monderer and Samet 1989, Rubinstein 1992). The analysis and applications of common knowledge and related multi-agent knowledge concepts has become a lively field of research.

The purpose of this essay is to overview of some of the most important results stemming from this contemporary research. The topics reviewed in each section of this essay are as follows: Section 1 gives motivating examples which illustrate a variety of ways in which the actions of agents depend crucially upon their having, or lacking, certain common knowledge. Section 2 discusses alternative analyses of common knowledge. Section 3 reviews applications of multi-agent knowledge concepts, particularly to *game theory* (von Neumann and Morgenstern 1944), in which common knowledge assumptions have been found to have great importance in justifying *solution concepts* for mathematical games. Section 4 discusses skeptical doubts about the attainability of common knowledge. Finally, Section 5 discusses the *common belief* concept which result from weakening the assumptions of Lewis' account of common knowledge.

---

1. Motivating Examples

Most of the examples in this section are familiar in the common knowledge literature, although some of the details and interpretations presented here are new. Readers may want to ask themselves what, if any, distinctive aspects of mutual and common knowledge reasoning each example illustrates.

1.1. The Clumsy Waiter[1]

A waiter serving dinner slips, and spills gravy on a guest's white silk evening gown. The guest glares at the waiter, and the waiter declares "I'm sorry. It was my fault." Why did the waiter say that he was at fault? He knew that he was at fault, and he knew from the guest's angry expression that she knew he was at fault. However, the sorry waiter wanted assurance that the guest *knew that he knew* he was at fault. By saying openly that he was at fault, the waiter knew that the guest knew what he wanted her to know, namely, that he knew he was at fault. Note that the waiter's declaration established at least three levels of nested knowledge.

Certain assumptions are implicit in the preceding story. In particular, the waiter must know that the guest knows he has spoken the truth, and that she can draw the desired conclusion from what he says in this context. More fundamentally, the waiter must know that if he announces "It was my fault" to the guest, she will interpret his intended meaning correctly and will infer what his making this announcement ordinarily implies in this context. This in turn implies that the guest must know that if the waiter announces "It was my fault" in this context, then the waiter indeed knows he is at fault. Then on account of his announcement, the waiter knows that the guest knows that he knows he was at fault. The waiter's announcement was meant to generate *higher-order* levels of knowledge of a fact each already knew.

Just a slight strengthening of the stated assumptions results in even higher levels of nested knowledge. Suppose the waiter and the guest each know that the other can infer what he infers from the waiter's announcement. Can the guest now believe that the waiter does not know that she knows that he knows he is at fault? If the guest considers this question, she reasons that if the waiter falsely believes it is possible that she does not know that he knows he is at fault, then the waiter must believe it to be possible that she

cannot infer that he knows he is at fault from his own declaration. Since she knows she *can* infer that the waiter knows he is at fault from his declaration, she knows that the waiter knows she can infer this, as well. Hence the waiter's announcement establishes the fourth-order knowledge claim: The guest knows that the waiter knows that she knows that he knows he is at fault. By similar, albeit lengthier, arguments, the agents can verify that corresponding knowledge claims of even higher-order must also obtain under these assumptions.

1.2 The Barbecue Problem

This is a variation of an example first published by Littlewood (1953), although he notes that his version of the example was already well-known at the time.[2] N individuals enjoy a picnic supper together which includes barbecued spareribs. At the end of the meal, $k \geq 1$ of these diners have barbecue sauce on their faces. No one wants to continue the evening with a messy face. No one wants to wipe her face if it's not messy, for this would make her appear neurotic. And no one wants to take the risk of being thought rude by telling anyone else that he has barbecue sauce on his face. Since no one can see her own face, none of the messy diners makes a move to clean her face. Then the cook who served the spareribs returns with a carton of ice cream. Amused by what he sees, the cook rings the dinner bell and makes the following announcement: "At least one of you has barbecue sauce on her face. I will ring the dinner bell over and over, until anyone who is messy has wiped her face. Then I will serve dessert." For the first $k - 1$ rings, no one does anything. Then, at the k^{th} ring, each of the messy individuals suddenly reaches for a napkin, and soon afterwards, the diners are all enjoying their ice cream.

How did the messy diners finally realize that their faces needed cleaning? The $k = 1$ case is easy, since in this case, the lone messy individual will realize he is messy immediately, since he sees that everyone else is clean. Consider the $k = 2$ case next. At the first ring, messy individual i_1 knows that one other person, i_2, is messy, but does not yet know about himself. At the second ring, i_1 realizes that he must be messy, since had i_2 been the only messy one, i_2 would have known this after the first ring when the cook made his announcement, and would have cleaned her face then. By a symmetric argument, messy diner i_2 also concludes that she is messy at the second ring, and both pick up a napkin at that time.

Let's next consider $k = 3$. Again at the first ring, each of the messy diners i_1, i_2, and i_3 knows the status of the other diners, but not her own. The situation is apparently unchanged after the second ring. But on the third ring, i_1 realizes that she is messy. For if i_2 and i_3 were the only messy ones, then they would have discovered this after the second ring by the argument of the previous paragraph. Since i_1 can see that all of the diners other than i_2 and i_3 are clean, she concludes that she must be messy. i_2 and i_3 draw similar conclusions at the third ring, and all clean their faces at that time.

The general case follows by induction. Suppose that if $k = j$, then each of the j messy diners can determine that he is messy after j rings. Then if $k = j + 1$, then at the $j + 1^{\text{st}}$ ring, each of the $j + 1$ individuals will realize that he is messy. For if he were not messy, then the other j messy ones would have all realized their messiness at the j^{th} ring and cleaned themselves then. Since no one cleaned herself after

the j^{th} ring, at the $j + 1^{st}$ ring each messy person will conclude that someone besides the other *j* messy people must also be messy, namely, himself.

The "paradox" of this argument is that for $k > 1$, like the case of the clumsy waiter of Example 1.1, the cook's announcement told the diners something that each already knew. Yet apparently the cook's announcement also gave the diners useful information. How could this be? By announcing a fact already known to every diner, the cook made this fact *common knowledge* among them, enabling each of them to eventually deduce the condition of his own face after sufficiently many rings of the bell. Note that the inductive argument the agents run through depends upon the conclusions they each draw from several *counterfactual conditionals.* In general, the consequences of agents' common knowledge are intimately related to how they evaluate subjunctive and counterfactual conditionals.[3]

1.3 The Farmer's Dilemma

Does meeting one's obligations to others serve one's self-interest? Plato and his successors recognized that in certain cases, the answer seems to be "No." Hobbes (1651, pp. 101-102) considers the challenge of a "Foole", who claims that it is irrational to honor an agreement made with another who has already fulfilled his part of the agreement. Noting that in this situation one has gained all the benefit of the other's compliance, the Foole contends that it would now be best for him to break the agreement, thereby saving himself the costs of compliance. Of course, if the Foole's analysis of the situation is correct, then would the other party to the agreement not anticipate the Foole's response to agreements honored, and act accordingly?

Hume (1740, pp. 520-521) takes up this question, using an example: Two neighboring farmers each expect a bumper crop of corn. Each will require his neighbor's help in harvesting his corn when it ripens, or else a substantial portion will rot in the field. Since their corn will ripen at different times, the two farmers can ensure full harvests for themselves by helping each other when their crops ripen, and both know this. Yet the farmers do not help each other. For the farmer whose corn ripens later reasons that if she were to help the other farmer, then when her corn ripens he would be in the position of Hobbes' Foole, having already benefited from her help. He would no longer have anything to gain from her, so he would not help her, sparing himself the hard labor of a second harvest. Since she cannot expect the other farmer to return her aid when the time comes, she will not help when his corn ripens first, and of course the other farmer does not help her when her corn ripens later.

The structure of Hume's *Farmers' Dilemma* problem can be summarized using the following tree diagram:

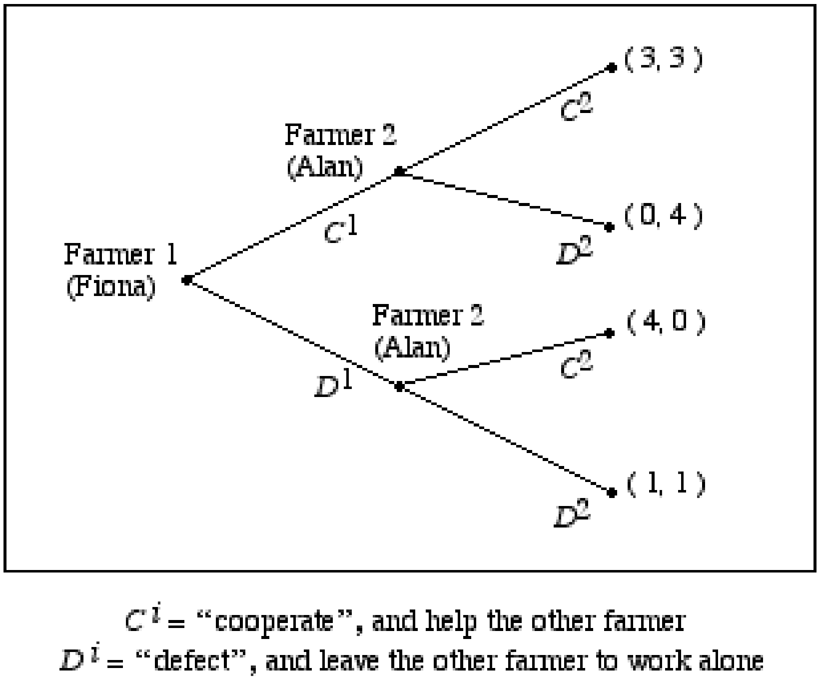


Figure 1.1a

This tree is an example of a *game in extensive form.* At each stage *i*, the agent who moves can either choose C^i, which corresponds to helping or *cooperating*, or D^i, which corresponds to not helping or *defecting*. The relative preferences of the two agents over the various outcomes are reflected by the ordered pairs of *payoffs* each receives at any particular outcome. If, for instance, Fiona chooses C^i and Alan chooses D^i, then Fiona's payoff is 0, her worst payoff, and Alan's is 4, his best payoff. In a game such as the Figure 1.1.a game, agents are *(Bayesian) rational* if each chooses an act that maximizes her expected payoff, given what she knows.

In the Farmers' Dilemma game, following the C^1,C^2-path is strictly better for both farmers than following the D^1,D^2-path. However, Fiona chooses D^1, as the result of the following simple argument: "If I were to choose C^1, then Alan, who is rational and who knows the payoff structure of the game, would choose D^2. I am also rational and know the payoff structure of the game. So I should choose D^1." Since Fiona knows that Alan is rational and knows the game's payoffs, she concludes that she need only analyze the *reduced* game in the following figure:

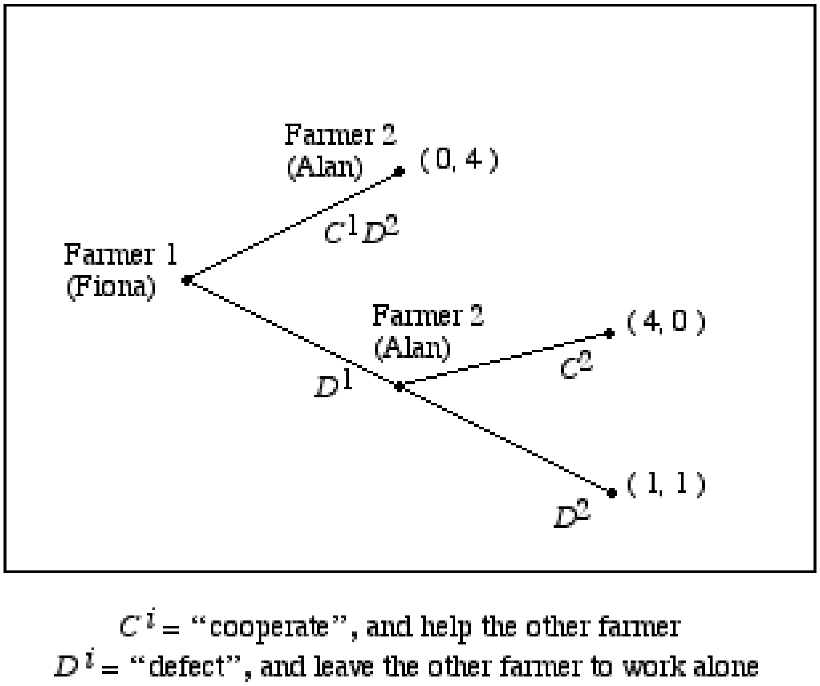


Figure 1.1b

In this reduced game, Fiona is certain to gain a strictly higher payoff by choosing D^1 than if she chooses C^1, so D^1 is her unique best choice. Of course, when Fiona chooses D^1, Alan, being rational, responds by choosing D^2. If Fiona and Alan know: (i) that they are both rational, (ii) that they both know the payoff structure of the game, and (iii) that they both know (i) and (ii), then they both can predict what the other will do at every node of the Figure 1.1.a game, and conclude that they can rule out the D^1,C^2-branch of the Figure 1.1.b game and analyze just the reduced game of the following figure:

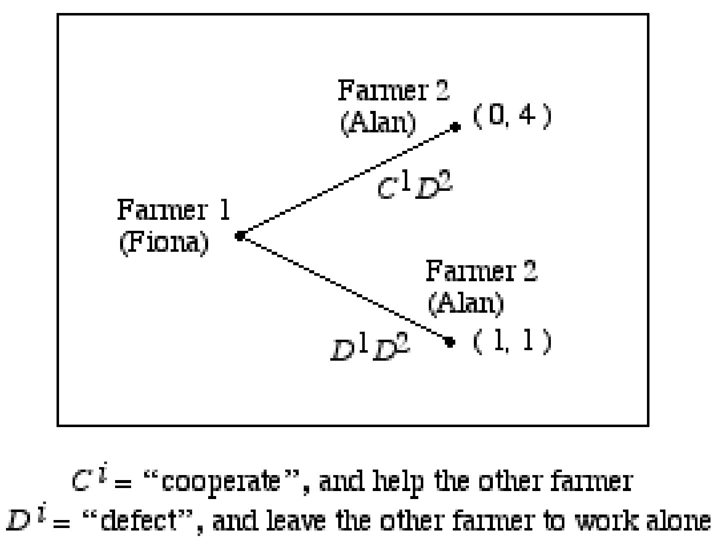


Figure 1.1c

On account of this *mutual knowledge*, both know that Fiona will choose D^1, and that Alan will respond with D^2. Hence, the D^1,D^2-outcome results if the Farmers' Dilemma game is played by agents having this

mutual knowledge, though it is suboptimal since both agents would fare better at the C^1,C^2-branch.[4] This argument, which in its essentials is Hume's argument, is an example of a standard technique for solving sequential games known as *backwards induction*.[5] The basic idea behind backwards induction is that the agents engaged in a sequential game deduce how each will act throughout the entire game by ruling out the acts that are not payoff-maximizing for the agents who would move last, then ruling out the acts that are not payoff-maximizing for the agents who would move next-to-last, and so on. Clearly, backwards induction arguments rely crucially upon what, if any, mutual knowledge the agents have regarding their situation, and they typically require the agents to evaluate the truth values of certain subjunctive conditionals, such as "If I (Fiona) were to choose C^1, then Alan would choose D^2".

1.4 The Centipede

The mutual knowledge assumptions required to construct a backwards induction solution to a game become more complex as the number of stages in the game increases. To see this, consider the sequential *Centipede* game depicted in the following figure:

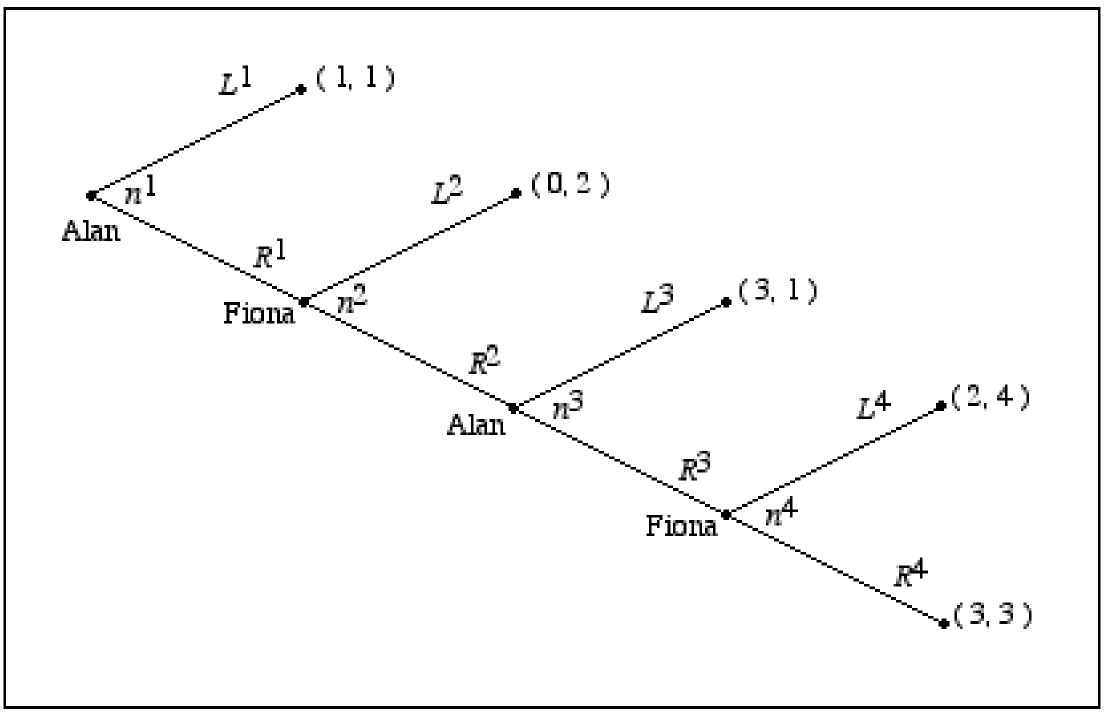


Figure 1.2

At each stage i, the agent who moves can either choose R^i, which in the first three stages gives the other agent an opportunity to move, or L^i, which ends the game.

Like the Farmers' Dilemma, this game is a commitment problem for the agents. If each agent could trust the other to choose R^i at each stage, then they would each expect to receive a payoff of 3. However, Alan chooses L^1, leaving each with a payoff of only 1, as the result of the following backwards induction argument: "If node n_4 were to be reached, then Fiona, (being rational) would choose L^4. I, knowing this,

would (being rational) choose L^3 if node n_3 were to be reached. Fiona, knowing *this*, would (being rational) choose L^2 if node n_2 were to be reached. Hence, I (being rational) should choose L^1." To carry out this backwards induction argument, Alan implicitly assumes that: (i) he knows that Fiona knows he is rational, and (ii) he knows that Fiona knows that he knows she is rational. Put another way, for Alan to carry out the backwards induction argument, at node n_1 he must know what Fiona must know at node n_2 to make L^2 her best response should n_2 be reached. While in the Farmer's Dilemma Fiona needed only *first-order* knowledge of Alan's rationality and *second-order* knowledge of Alan's knowledge of the game to derive the backwards induction solution, in the Figure 1.2 game, for Alan to be able to derive the backwards induction solution, the agents must have *third-order mutual knowledge* of the game and *second-order mutual knowledge* of rationality, and Alan must have *fourth-order* knowledge of this mutual knowledge of the game and *third-order* knowledge of their mutual knowledge of rationality. This argument also involves several counterfactuals, since to construct it the agents must be able to evaluate conditionals of the form, "If node n_i were to be reached, Alan (Fiona) would choose L^i (R^i)", which for $i > 1$ are counterfactual, since third-order mutual knowledge of rationality implies that nodes n_2, n_3, and n_4 are never reached.

The method of backwards induction can be applied to any sequential game of *perfect information*, in which the agents can observe each others' moves in turn and can recall the entire history of play. However, as the number of potential stages of play increases, the backwards induction argument evidently becomes harder to construct. This raises certain questions: (1) What precisely are the mutual or common knowledge assumptions that are required to justify the backwards induction argument for a particular sequential game? (2) As a sequential game increases in complexity, would we expect the mutual knowledge that is required for backwards induction to start to fail?

1.5 The Department Store

> When a man loses his wife in a department store without any prior understanding on where to meet if they get separated, the chances are good that they will find each other. It is likely that each will think of some obvious place to meet, so obvious that each will be sure that it is "obvious" to both of them. One does not simply predict where the other will go, which is wherever the first predicts the second to predict the first to go, and so *ad infinitum.* Not "What would I do if I were she?" but "What would I do if I were she wondering what she would do if she were wondering what I would do if I were she . . . ?"
>
> Thomas Schelling, *The Strategy of Conflict*

Schelling's department store problem is an example of a *pure coordination problem*, that is, an interaction problem in which the interests of the agents coincide perfectly. Schelling (1960) and Lewis (1969), who were the first to make explicit the role common knowledge plays in social coordination, were also among the first to argue that coordination problems can be modeled using the analytic vocabulary of game theory. A very simple example of such a coordination problem is given in the next figure:

Robert

Liz	s_1	s_2	s_3	s_4
s_1	(1, 1)	(0, 0)	(0, 0)	(0, 0)
s_2	(0, 0)	(1, 1)	(0, 0)	(0, 0)
s_3	(0, 0)	(0, 0)	(1, 1)	(0, 0)
s_4	(0, 0)	(0, 0)	(0, 0)	(1, 1)

s_i = search on floor i, $1 \leq i \leq 4$

Figure 1.3

The matrix of Figure 1.3 is an example of a *game in strategic form.* At each outcome of the game, which corresponds to a cell in the matrix, the row (column) agent receives as payoff the first (second) element of the ordered pair in the corresponding cell. However, in strategic form games, each agent chooses without first being able to observe the choices of any other agent, so that all must choose as if they were choosing simultaneously. The Figure 1.3 game is a game of *pure coordination* (Lewis 1969), that is, a game in which at each outcome, each agent receives exactly the same payoff. One interpretation of this game is that Schelling's spouses, Liz and Robert, are searching for each other in the department store with four floors, and they find each other if they go to the same floor. Four outcomes at which the spouses coordinate correspond to the strategy profiles (s_j, s_j), $1 \leq j \leq 4$, of the Figure 1.3 game. These four profiles are strict *Nash equilibria* (Nash 1950, 1951) of the game, that is, each agent has a decisive reason to follow her end of one of these strategy profiles provided that the other also follows this profile.[6]

The difficulty the agents face is trying to select an equilibrium to follow. For suppose that Robert hopes to coordinate with Liz on a particular equilibrium of the game, say (s_2, s_2). Robert reasons as follows: "Since there are several strict equilibria we might follow, I should follow my end of (s_2, s_2) if, and only if, I have sufficiently high expectations that Liz will follow her end of (s_2, s_2). But I can only have sufficiently high expectations that Liz will follow (s_2, s_2) if she has sufficiently high expectations that I will follow (s_2, s_2). For her to have such expectations, Liz must have sufficiently high (second-order) expectations that I have sufficiently high expectations that she will follow (s_2, s_2), for if Liz doesn't have these (second-order) expectations, then she will believe I don't have sufficient reason to follow (s_2, s_2) and may therefore deviate from (s_2, s_2) herself. So I need to have sufficiently high (third-order) expectations that Liz has sufficiently high (second-order) expectations that I have sufficiently high expectations that she will follow (s_2, s_2). But this implies that Liz must have sufficiently high (fourth-order) expectations that I have sufficiently high (third-order) expectations that Liz has sufficiently high

(second-order) expectations that I have sufficiently high expectations that she will follow (s_2, s_2), for if she doesn't, then she will believe I don't have sufficient reason to follow (s_2, s_2), and then she won't, either. Which involves me in fifth-order expectations regarding Liz, which involves her in sixth-order expectations regarding me, and so on." What would suffice for Robert, and Liz, to have decisive reason to follow (s_2, s_2) is that they each *know* that the other *knows* that . . . that the other will follow (s_2, s_2) for any number of levels of knowledge, which is to say that between Liz and Robert it is common knowledge that they will follow (s_2, s_2). If agents follow a strict equilibrium in a pure coordination game as a consequence of their having common knowledge of the game, their rationality and their intentions to follow this equilibrium, and no other, then the agents are said to be following a *Lewis-convention* (Lewis 1969).

Lewis' theory of convention applies to a more general class of games than pure coordination games, but pure coordination games already model a variety of important social interactions. In particular, Lewis models conventions of language as equilibrium points of a pure coordination game. The role common knowledge plays in games of pure coordination sketched above of course raises further questions: (1) Can people ever attain the common knowledge which characterizes a Lewis-convention? (2) Would less stringent epistemic assumptions suffice to justify Nash equilibrium behavior in a coordination problem?

2. Alternative Accounts of Common Knowledge

Informally, a proposition *A* is *mutually known* among a set of agents if each agent knows that *A*. Mutual knowledge by itself implies nothing about what, if any, knowledge anyone attributes to anyone else. Suppose each student arrives for a class meeting knowing that the instructor will be late. That the instructor will be late is mutual knowledge, but each student might think only she knows the instructor will be late. However, if one of the students says openly "Peter told me he will be late again," then the mutally known fact is now *commonly known.* Each student now knows that the instructor will be late, and so on, *ad infinitum.* The agents have common knowledge in the sense articulated informally by Schelling (1960), and more precisely by Lewis (1969) and Schiffer (1972). Schiffer uses the formal vocabulary of *epistemic logic* (Hintikka 1962) to state his definition of common knowledge. Schiffer's general approach was to augment a system of sentential logic with a set of knowledge operators corresponding to a set of agents, and then to define common knowledge as a hierarchy of propositions in the augmented system. Bacharach (1992) and Bicchieri (1993) adopt this approach, and develop logical theories of common knowledge which include soundness and completeness theorems. One can also develop alternate formal accounts of common knowledge in set-theoretic terms, which is the approach taken in this article.[7]

2.1 The Hierarchical Account

Monderer and Samet (1988) and Binmore and Brandenburger (1989) give a particularly elegant set-theoretic definition of common knowledge. I will review this definition here, and then show that it is logically equivalent to the '*i* knows that *j* knows that … *k* knows that A' hierarchy that Lewis (1969) and Schiffer (1972) argue characterizes common knowledge.[8]

Some preliminary notions must be stated first. Following C. I. Lewis (1943-1944) and Carnap (1947), propositions are formally subsets of a set Ω of *state descriptions* or *possible worlds.* One can think of the elements of Ω as representing Leibniz's possible worlds or Wittgenstein's possible states of affairs. Some results in the common knowledge literature presuppose that Ω is of finite cardinality. If this admittedly unrealistic assumption is needed in any context, this will be explicitly stated in this essay, and otherwise one may assume that Ω may be either a finite or an infinite set. A distinguished actual world ω_α is an element of Ω. A proposition $A \subseteq \Omega$ obtains (or is true) if the actual world $\omega_\alpha \in A$. In general, we say that *A obtains at* a world $\omega \in \Omega$ if $\omega \in A$. What an agent *i* knows about the possible worlds is stated formally in terms of a *knowledge operator* $\mathbf{K}_i$. Given a proposition $A \subseteq \Omega$, $\mathbf{K}_i(A)$ denotes a new proposition, corresponding to the set of possible worlds at which agent *i* knows that A obtains. $\mathbf{K}_i(A)$ is read as '*i* knows (that) *A* (is the case)'. The knowledge operator $\mathbf{K}_i$ satisfies certain axioms, including:

K1: $\mathbf{K}_i(A) \subseteq A$

K2: $\Omega \subseteq \mathbf{K}_i(\Omega)$

K3: $\mathbf{K}_i(\bigcap_k A_k) = \bigcap_k \mathbf{K}_i(A_k)$

K4: $\mathbf{K}_i(A) \subseteq \mathbf{K}_i\mathbf{K}_i(A)$[9]

In words, K1 says that if *i* knows *A*, then *A* must be the case. K2 says that *i* knows that some possible world in Ω occurs no matter which possible world ω occurs. K3 says that *i* knows a conjunction if, and only if, *i* knows each conjunct. K4 is a *reflection axiom*, which says that if *i* knows *A*, then *i* knows that she knows *A*. Note that by K3, if $A \subseteq B$ then $\mathbf{K}_i(A) \subseteq \mathbf{K}_i(B)$, by K1 and K2, $\mathbf{K}_i(\Omega) = \Omega$, and by K1 and K4, $\mathbf{K}_i(A) = \mathbf{K}_i\mathbf{K}_i(A)$. Any system of knowledge satisfying K1 - K4 corresponds to the modal system S4 (Kripke 1963). If one drops the K1 axiom and retains the others, the resulting system would give a formal account of what an agent *believes*, but does not necessarily *know.*

A useful notion in the formal analysis of knowledge is that of a *possibility set.* An agent i's possibility set at a state of the world Ω is the smallest set of possible worlds that *i* thinks could be the case if ω is the actual world. More precisely,

Definition 2.1
Agent *i*'s *possibility set* $\mathcal{H}_i(\omega)$ at $\omega \in \Omega$ is defined as

$$\mathcal{H}_i(\omega) \equiv \bigcap \{ E \mid \omega \in \mathbf{K}_i(E) \}$$

The collection of sets

$$\mathcal{H}_i = \bigcup_{\omega \in \Omega} \mathcal{H}_i(\omega)$$

is *i*'s *private information system.*

Since in words, $\mathcal{H}_i(\omega)$ is the intersection of all propositions which *i* knows at ω, $\mathcal{H}_i(\omega)$ is the smallest proposition in Ω that *i* knows at ω. Put another way, $\mathcal{H}_i(\omega)$ is the most specific information that *i* has about the possible world ω. The intuition behind assigning agents private information systems is that while an agent *i* may not be able to perceive or comprehend every last detail of the world in which *i* lives, *i* does know certain facts about that world. The elements of *i*'s information system represent what *i* knows immediately at a possible world. We also have the following:

Proposition 2.2
$\mathbf{K}_i(A) = \{ \omega \mid \mathcal{H}_i(\omega) \subseteq A \}$

In many formal analyses of knowledge in the literature, possibility sets are taken as primitive and Proposition 2.2 is given as the definition of knowledge. If one adopts this viewpoint, then the axioms K1 - K4 follow as consequences of the definition of knowledge. In many applications, the agents' possibility sets are assumed to *partition*[10] the set, in which case $\mathcal{H}_i$ is called i's *private information partition.*

To illustrate the idea of possibility sets, let us return to the Barbecue Problem described in Example 1.2. Suppose there are three diners: Cathy, Jennifer and Mark. Then there are 8 relevant states of the world, summarized by Table 2.1:

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
Cathy	clean	messy	clean	clean	messy	messy	clean	messy
Jennifer	clean	clean	messy	clean	messy	clean	messy	messy
Mark	clean	clean	clean	messy	clean	messy	messy	messy

Table 2.1

Each diner knows the condition of the other diners' faces, but not her own. Suppose the cook makes no

announcement, after all. Then none of the diners knows the true state of the world whatever $\omega \in \Omega$ the actual world turns out to be, but they do know *a priori* that certain propositions are true at various states of the world. For instance, Cathy's information system before any announcement is made is depicted in Figure 2.1a:

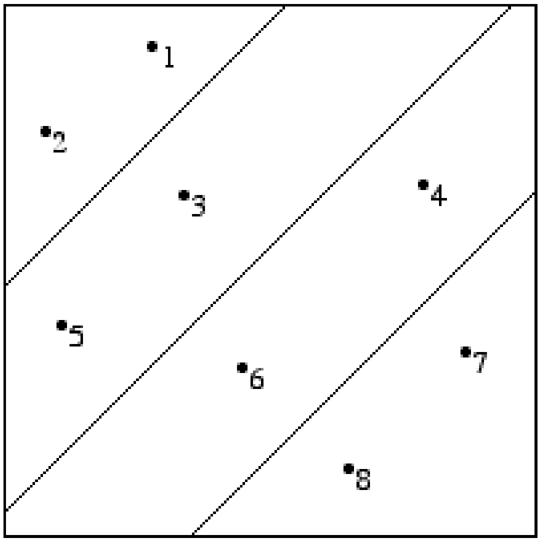


Figure 2.1a

In this case, Cathy's information system is a partition $\mathcal{H}_1$ of Ω defined by

$$\mathcal{H}_1 = \{H_{CC}, H_{CM}, H_{MC}, H_{MM}\}$$

where

$H_{CC} = \{\omega_1, \omega_2\}$ (i.e., Jennifer and Mark are both clean)

$H_{CM} = \{\omega_4, \omega_6\}$ (i.e., Jennifer is clean and Mark is messy)

$H_{MC} = \{\omega_3, \omega_5\}$ (i.e., Jennifer is messy and Mark is clean)

$H_{MM} = \{\omega_7, \omega_8\}$ (i.e., Jennifer and Mark are both messy)

Cathy knows immediately which cell $\mathcal{H}_1(\omega)$ in her partition is the case at any state of the world, but does not know which is the true state at any $\omega \in \Omega$.

If we add in the assumption stated in Example 1.2 that if there is at least one messy diner, then the cook announces the fact, then Cathy's information partition is depicted by Figure 2.1b:

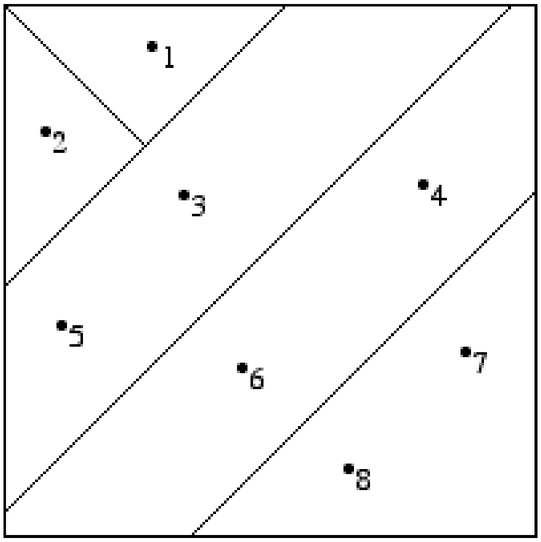


Figure 2.1b

In this case, Cathy's information system is a partition $\mathcal{H}_1$ of Ω defined by

$$\mathcal{H}_1 = \{H_{CCC}, H_{MCC}, H_{CM}, H_{MC}, H_{MM}\}$$

where

$H_{CCC} = \{\omega_1\}$ (i.e., Jennifer, Mark, and I are all clean)
$H_{MCC} = \{\omega_2\}$ (i.e., Jennifer and Mark are clean and I am messy)
$H_{CM} = \{\omega_4, \omega_6\}$ (i.e., Jennifer is clean and Mark is messy)
$H_{MC} = \{\omega_3, \omega_5\}$ (i.e., Jennifer is messy and Mark is clean)
$H_{MM} = \{\omega_7, \omega_8\}$ (i.e., Jennifer and Mark are both messy)

In this case, Cathy's information partition is a *refinement* of the partition she has when there is no announcement, for in this case, then Cathy knows *a priori* that if ω_1 is the case there will be no announcement and will know immediately that she is clean, and Cathy knows *a priori* that if ω_2 is the case, then she will know immediately from the cook's announcement that she is messy.

A slightly more complex case occurs if we alter the Barbecue problem so that the cook makes an announcement only if he sees at least two messy diners. Cathy's possibility set is now depicted by the diagram in Figure 2.1c:

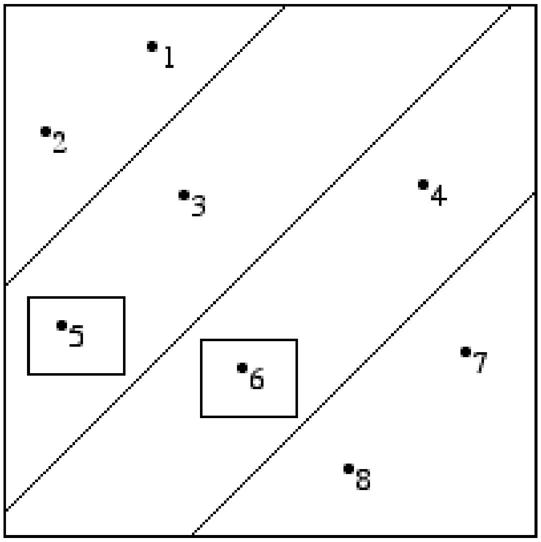


Figure 2.1c

This time, Cathy's information system does not partition Ω. For Cathy knows *a priori* that at ω_5, the cook will make his announcement, and since at ω_5 Jennifer is messy and Mark is clean, Cathy will realize immediately that she is messy. However, Cathy also knows *a priori* that at ω_3, either ω_3 or ω_5 could be the case, since at ω_3 she does not know in advance whether or not the cook will make an announcement. Hence $\mathcal{H}_1(\omega_5) = \{\omega_5\}$, but $\mathcal{H}_1(\omega_3) = \{\omega_3, \omega_5\}$. Similarly, $\mathcal{H}_1(\omega_6) = \{\omega_6\}$, but $\mathcal{H}_1(\omega_4) = \{\omega_4, \omega_6\}$. Jennifer's and Mark's information systems given any of the above three scenarios are derived similarly to Cathy's information system, and the details of this are left as an exercise for the reader.

We can now define mutual and common knowledge as follows:

Definition 2.3
Let a set Ω of possible worlds together with a set of agents N be given.

1. The proposition that A is *(first level* or *first order) mutual knowledge for the agents of N*, $\mathbf{K}^1_N(A)$, is the set defined by

$$\mathbf{K}^1_N(A) \equiv \bigcap_{i \in N} \mathbf{K}_i(A).$$

2. The proposition that A is m^{th} *level* (or m^{th} *order*) *mutual knowledge among the agents of N*, $\mathbf{K}^m_N(A)$, is defined recursively as the set

$$\mathbf{K}^m_N(A) \equiv \bigcap_{i \in N} \mathbf{K}_i(\mathbf{K}^{m-1}_N(A)).$$

3. The proposition that A is *common knowledge* among the agents of N, $\mathbf{K}^*_N(A)$, is defined as the set

$$\mathbf{K}^*_N(A) \equiv \bigcap_{m=1}^{\infty} \mathbf{K}^m_N(A).$$

As a consequence of Proposition 2.2, the agents' private information systems determine an *a priori* structure of propositions over the space of possible worlds regarding what they can know, including what mutual and common knowledge they potentially have. The world $\omega \in \Omega$ which obtains determines *a posteriori* what individual, mutual and common knowledge agents in fact have. Hence, one can read $\omega \in \mathbf{K}_i(A)$ as '*i* knows *A* at (possible world) ω', $\omega \in \mathbf{K}^m_N(A)$ as '*A* is m^{th} level mutual knowledge for the agents of *N* at ω', and so on. If ω obtains, then one can conclude that *i* does know *A*, that *A* is m^{th} level mutual knowledge, and so on. Common knowledge of a proposition *E* implies common knowledge of all that *E* implies, as is shown in the following:

> **Proposition 2.4**
> If $\omega \in \mathbf{K}^*_N(E)$ and $E \subseteq F$, then $\omega \in \mathbf{K}^*_N(F)$.
>
> Proof.

Note that $(\mathbf{K}^m_N(E))_{m \geq 1}$ is a decreasing sequence of events, in the sense that $\mathbf{K}^{m+1}_N(E) \subseteq \mathbf{K}^m_N(E)$, for all $m \geq 1$. It is also easy to check that if everyone knows *E*, then *E* must be true, that is, $\mathbf{K}^1_N(E) \subseteq E$. If Ω is assumed to be finite, then if *E* is common knowledge at ω, this implies that there must be a finite *m* such that

$$\mathbf{K}^m_N(E) = \bigcap_{n=1}^{\infty} \mathbf{K}^n_N(E).$$

The following result relates the set-theoretic definition of common knowledge to the hierarchy of '*i* knows that *j* knows that … knows *A*' statements.

> **Proposition 2.5**
> $\omega \in \mathbf{K}^m_N(A)$ iff
>
> (1) For all agents $i_1, i_2, \ldots, i_m \in N$, $\omega \in \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_m}(A)$
>
> Hence, $\omega \in \mathbf{K}^*_N(A)$ iff (1) is the case for each $m \geq 1$.
>
> Proof.

The condition that $\omega \in \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_m}(A)$ for all $m \geq 1$ and all $i_1, i_2, \ldots, i_m \in N$ is Schiffer's definition of common knowledge, and is often used as the definition of common knowledge in the literature.

2.2 Lewis' Account

Lewis is credited with the idea of characterizing common knowledge as a hierarchy of '*i* knows that *j* knows that … knows that *A*' propositions. However, it is far less well recognized that in *Convention*, Lewis also gives an algorithm which generates such a hierarchy from a finite set of assumptions regarding the agents' knowledge. These assumptions taken together constitute Lewis' official definition of common knowledge. Lewis' presentation of this definition is informal, and occasionally lacking in detail. It is probably for this reason that Aumann is often credited with presenting the first finitary method of generating the common knowledge hierarchy (Aumann 1976). A mathematically precise account of Lewis' analysis of common knowledge is given here, and it is shown that Lewis' analysis does result in the common knowledge hierarchy following from a finite set of axioms.

Lewis presents his account of common knowledge on pp. 52-57 of *Convention*. Lewis does not specify what account of knowledge is needed for common knowledge. As it turns out, Lewis' account is satisfactory for any formal account of knowledge in which the knowledge operators $\mathbf{K}_i$, $i \in N$, satisfy K1, K2, and K3. A crucial assumption in Lewis' analysis of common knowledge is that agents know they share the same "rationality, inductive standards and background information" (Lewis 1969, p. 53) with respect to a state of affairs A', that is, if an agent can draw any conclusion from A', she knows that all can do likewise. This idea is made precise in the following:

> **Definition 2.6**
> Given a set of agents N and a proposition $A' \subseteq \Omega$, the agents of N are *symmetric reasoners with respect to* A' (*or* A'*-symmetric reasoners*) iff, for each $i, j \in N$ and for any proposition $E \subseteq \Omega$, if $\mathbf{K}_i(A') \subseteq \mathbf{K}_i(E)$ and $\mathbf{K}_i(A') \subseteq \mathbf{K}_i\mathbf{K}_j(A')$, then $\mathbf{K}_i(A') \subseteq \mathbf{K}_i\mathbf{K}_j(E)$.[11]

The definiens says that for each agent *i*, if *i* can infer from A' that E is the case and that everyone knows that A' is the case, then *i* can also infer that everyone knows that E is the case.

> **Definition 2.7**
> A proposition E is *Lewis-common knowledge at* $\omega \in \Omega$ among the agents of a set $N = \{1, \ldots, n\}$ iff there is a proposition A^* such that $\omega \in A^*$, the agents of N are A^*-symmetric reasoners, and for every $i \in N$,
>
> L1: $\omega \in \mathbf{K}_i(A^*)$
>
> L2: $\mathbf{K}_i(A^*) \subseteq \mathbf{K}_i(\bigcap_{j \in N} \mathbf{K}_j(A^*))$

L3: $\mathbf{K}_i(A^*) \subseteq \mathbf{K}_i(E)$

> A^* is a *basis* for the agents' common knowledge. $\mathbf{L}^*_N(E)$ denotes the proposition defined by L1 - L3 for a set N of A^*-symmetric reasoners, so we can say that E is Lewis-common knowledge for the agents of N iff $\omega \in \mathbf{L}^*_N(E)$.

In words, L1 says that i knows A^* at ω. L2 says that if i knows that A^* obtains, then i knows that everyone knows that A^* obtains. This axiom is meant to capture the idea that common knowledge is based upon a proposition A^* that is *publicly known*, as is the case when agents hear a public announcement. If the agents' knowledge is represented by partitions, then a typical basis for the agents' common knowledge would be an element $\mathcal{M}(\omega)$ in the meet[12] of their partitions. L3 says that i can infer from A^* that E.

A human agent obviously cannot work her way mentally through an infinite mutual knowledge hierarchy. Lewis argues that this is not a problem for his analysis of common knowledge, since the mutual knowledge claims of a common knowledge hierarchy for a chain of logical consequences, not a series of steps in anyone's actual reasoning. Lewis uses an example to show how his definition of common knowledge generates the first few levels of mutual knowledge. In fact, Lewis' definition implies the entire common knowledge hierarchy, as is shown in the following result.

> **Proposition 2.8**
> $\mathbf{L}^*_N(E) \subseteq \mathbf{K}^*_N(E)$, that is, Lewis-common knowledge of E implies common knowledge of E.
>
> Proof.

2.3 Aumann's Account

Aumann (1976) gives a different characterization of common knowledge which gives another simple algorithm for determining what information is commonly known. Aumann's original account assumes that the each agent's possibility set forms a private information partition of the space Ω of possible worlds. Aumann shows that a proposition C is common knowledge if, and only if, C contains a cell of the meet of the agents' partitions. One way to compute the meet $\mathcal{M}$ of the partitions $\mathcal{H}_i$, $i \in N$ is to use the idea of "reachability".

> **Definition 2.9**
> A state $\omega' \in \Omega$ is *reachable* from $\omega \in \Omega$ iff there exists a sequence $\omega=\omega_0, \omega_1, \omega_2, \dots,$ $\omega_m=\omega'$ such that for each $k \in \{0,1, \dots, m-1\}$, there exists an agent $i_k \in N$ such that $\mathcal{H}_{i_k}(\omega_k) = \mathcal{H}_{i_k}(\omega_{k+1})$.

In words, ω' is reachable from ω if there exists a sequence or "chain" of states from ω to ω' such that two consecutive states are in the same cell of some agent's information partition. To illustrate the idea of reachability, let us return to the modified Barbecue Problem in which Cathy, Jennifer and Mark receive no announcement. Their information partitions are all depicted in Figure 2.1d:

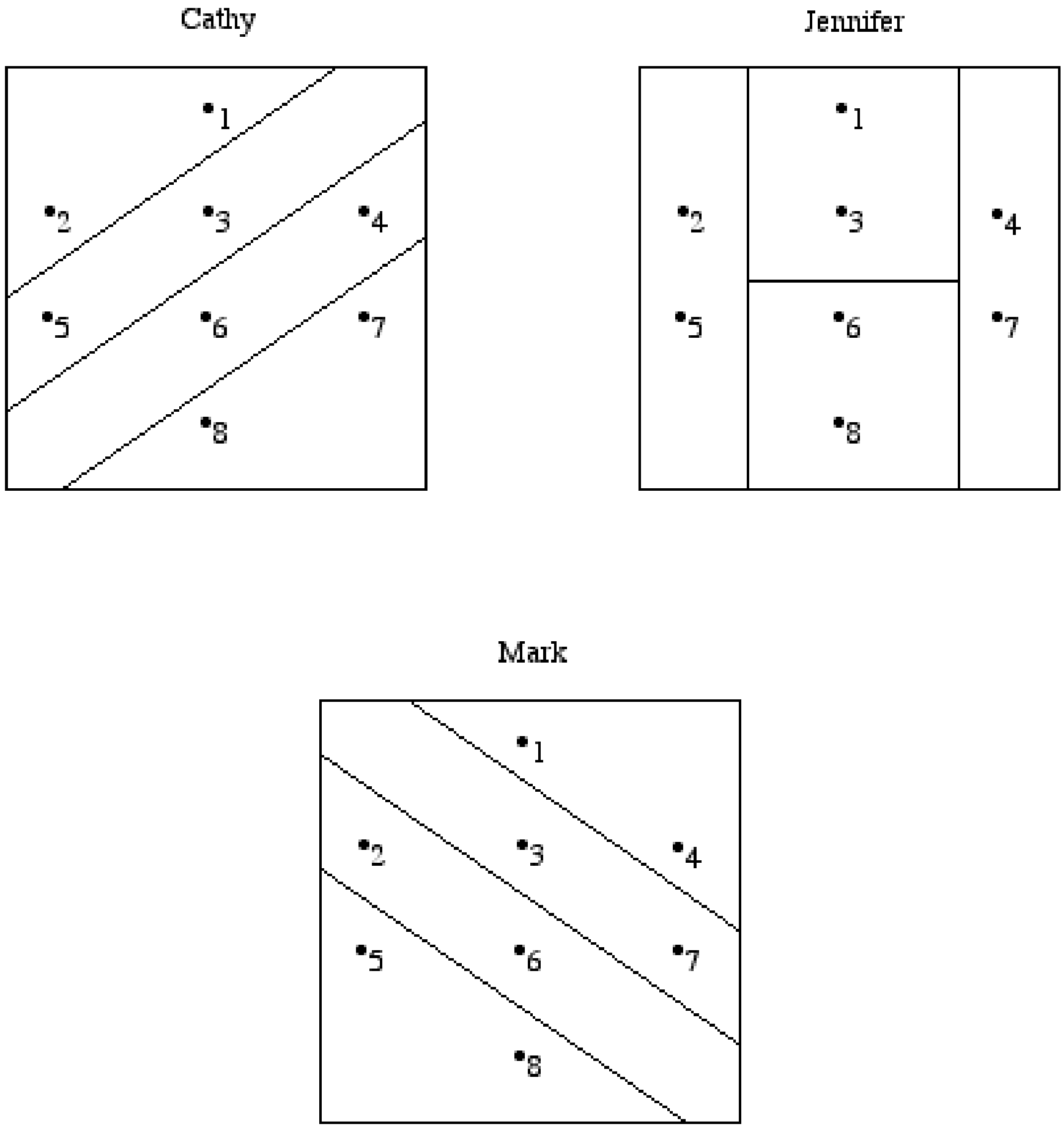


Figure 2.1d

One can understand the importance of the notion of reachability in the following way: If ω' is reachable from ω, then if ω obtains then some agent can reason that some other agent thinks that ω' is possible. Looking at Figure 2.1d, if $\omega = \omega_1$ occurs, then Cathy (who knows only that $\{\omega_1, \omega_2\}$ has occurred) knows that Jennifer thinks that ω_5 might have occurred (even though Cathy knows that ω_5 did not occur). So Cathy cannot rule out the possibility that Jennifer thinks that Mark thinks that that ω_8 might have occurred. And Cathy cannot rule out the possibility that Jennifer thinks that Mark thinks that Cathy believes that ω_7 is possible. In this sense, ω_7 is reachable from ω_1. The chain of states which establishes this is ω_1, ω_2, ω_5, ω_8, ω_7, since $\mathcal{H}_1(\omega_1) = \mathcal{H}_1(\omega_2)$, $\mathcal{H}_2(\omega_2) = \mathcal{H}_2(\omega_5)$, $\mathcal{H}_3(\omega_5) = \mathcal{H}_3(\omega_8)$, and $\mathcal{H}_1($

$\omega_8) = \mathcal{H}_1(\omega_7)$. Note that one can show similarly that in this example any state is reachable from any other state. This example also illustrates the following immediate result:

Proposition 2.10
ω' is reachable from ω iff there is a sequence $i_1, i_2, \ldots, i_m \in N$ such that

$$(1)\ \omega' \in \mathcal{H}_{i_m}(\ \ldots (\mathcal{H}_{i_2}(\mathcal{H}_{i_1}(\omega))))$$

One can read (1) as: 'At ω, i_1 thinks that i_2 thinks that … , i_m thinks that ω' is possible.'

We now have:

Lemma 2.11
$\omega' \in \mathcal{M}(\omega)$ iff ω' is reachable from ω.

Proof.

and

Lemma 2.12.
$\mathcal{M}(\omega)$ is common knowledge for the agents of N at ω.

Proof.

and

Proposition 2.13 (Aumann 1976)
Let $\mathcal{M}$ be the meet of the agents' partitions $\mathcal{H}_i$ for each $i \in N$. A proposition $E \subseteq \Omega$ is common knowledge for the agents of N at ω iff $\mathcal{M}(\omega) \subseteq E$. (In Aumann (1976), E is *defined* to be common knowledge at ω iff $\mathcal{M}(\omega) \subseteq E$.)

Proof.

If $E = \mathbf{K}^1_N(E)$, then E is a *public event* (Milgrom 1981) or a *common truism* (Binmore and Brandenburger 1989). Clearly, a common truism is common knowledge whenever it occurs, since in this case $E = \mathbf{K}^1_N(E) = \mathbf{K}^2_N(E) = \ldots$, so $E = \mathbf{K}^*_N(E)$. The proof of Proposition 2.13 shows that the common truisms are precisely the elements of $\mathcal{M}$ and unions of elements of $\mathcal{M}$, so any commonly known event is the consequence of a common truism.

2.4 Barwise's Account

Barwise (1988) proposes another definition of common knowledge that avoids explicit reference to the hierarchy of '*i* knows that *j* knows that … knows that *A*' propositions. Barwise's analysis builds upon an informal proposal by Harman (1977). Consider the situation of the guest and clumsy waiter in Example 1 when he announces that he was at fault. They are now in a setting where they have heard the waiter's announcement and know that they are in the setting. Harman adopts the circularity in this characterization of the setting as fundamental, and propses a definition of common knowledge in terms of this circularity. Barwise's formal analysis gives a precise formulation of Harman's intuitive analysis of common knowledge as a *fixed point*. Given a function *f*, *A* is a fixed point of *f* if *f(A)=A*. Now note that

$$\begin{aligned}
\mathbf{K}^1_N\left(E \cap \bigcap_{m=1}^{\infty} \mathbf{K}^m_N(E)\right) &= \mathbf{K}^1_N(E) \cap \mathbf{K}^1_N\left(\bigcap_{m=1}^{\infty} \mathbf{K}^m_N(E)\right) \\
&= \mathbf{K}^1_N(E) \cap \left(\bigcap_{m=1}^{\infty} \mathbf{K}^1_N(\mathbf{K}^m_N(E))\right) \\
&= \mathbf{K}^1_N(E) \cap \left(\bigcap_{m=2}^{\infty} \mathbf{K}^m_N(E)\right) \\
&= \bigcap_{m=1}^{\infty} \mathbf{K}^m_N(E)
\end{aligned}$$

So we have established that $\mathbf{K}^*_N(E)$ is a fixed point of the function f_E defined by $f_E(X) = \mathbf{K}^1_N(E \cap X)$. f_E has other fixed points. For instance, any contradiction $B \cap B^c = \varnothing$ is a fixed point of f_E.[13] Note also that if $A \subseteq B$, then $E \cap A \subseteq E \cap B$ and so

$$f_E(A) = \mathbf{K}^1_N(E \cap A) \subseteq \mathbf{K}^1_N(E \cap B) = f_E(B)$$

that is, f_E is *monotone*. (We saw that $\mathbf{K}^1_N$ is also monotone in the proof of Proposition 2.4.) Barwise's analysis of common knowledge can be developed using the following result from set theory:

> **Proposition**
> A monotone function *f* has a unique fixed point *C* such that if *B* is a fixed point of *f*, then $B \subseteq C$. *C* is the *greatest fixed point of f.*

This proposition establishes that f_E has a greatest fixed point, which characterizes common knowledge in Barwise's account. As Barwise himself observes, the fixed point analysis of common knowledge is closely related to Aumann's partition account. This is easy to see when one compares the fixed point

analysis to the notion of common truisms that Aumann's account generates. Some authors regard the fixed point analysis as an alternate formulation of Aumann's analysis. Barwise's fixed point analysis of common knowledge is favored by those who are especially interested in the applications of common knowledge to problems in logic, while the hierarchical and the partition accounts are favored by those who wish to apply common knowledge in social philosophy and social science. When knowledge operators satisfy the axioms (K1)-(K4), the Barwise account of common knowledge is equivalent to the hierarchical account.

Proposition 2.14
Let C^*_N be the greatest fixed point of f_E. Then $C^*_N(E) = K^*_N(E)$. (In Barwise (1988, 1989), E is *defined* to be common knowledge at ω iff $\omega \in C^*_N(E)$.)

Proof.

Barwise argues that in fact the fixed point analysis is more flexible and consequently more general than the hierachical account. This may surprise readers in light of Proposition 2.14, which shows that Barwise's fixed point definition is *equivalent* to the hierarchical account. Indeed, while Barwise (1988, 1989) proves a result showing that the fixed point account implies the hierarchical account and gives examples that satisfy the common knowledge hierarchy but fail to be fixed points, a number of authors who have written after Barwise have given various proofs of the equivalence of the two definitions, as was shown in Proposition 2.14. In fact, there is not a true controversy, at least with respect to the analytical results. Barwise's fixed point account is indeed equivalent to the hierarchical and the partition accounts given the account of knowledge characterized by (K1)-(K4) that most practitioners accept. Barwise does not make explicit which axioms of (K1)-(K4) he accepts, but he wishes to analyze a weaker notion of knowledge that is not closed under logical implication, and so he is committed to rejecting (K3). By doing so, Barwise is able to prove the nonequivalence between the fixed point and the hierarchical account he claims. But Barwise's result comes at a price most analysts are not willing to pay. To formulate his results given his very weak conception of knowledge, Barwise must use *non-well-founded set theory* (Aczel 1988) in order to allow him to make the necessary circular definitions. As we have seen in this section, when one adopts the conventional analysis of knowledge that satisfies (K1)-(K4), the equivalence of the hierarchical and the fixed point accounts follows without the need to introduce non-well-founded set-theoretic concepts.

2.5 Gilbert's Account

Gilbert (1989, Chapter 3) presents an alternative account of common knowledge, which is meant to be more intuitively plausible than Lewis' and Aumann's accounts. Gilbert gives a highly detailed description of the circumstances under which agents have common knowledge.

Definition 2.15
A set of agents N are in a *common knowledge situation* $\mathcal{S}(A)$ with respect to a proposition A if, and only if, $\omega \in A$ and for each $i \in N$,

G_1: *i* is *epistemically normal*, in the sense that *i* has normal perceptual organs which are functioning normally and has normal reasoning capacity.[14]

G_2: *i* has the concepts needed to fulfill the other conditions.

G_3: *i* perceives the other agents of *N*.

G_4: *i* perceives that G_1 and G_2 are the case.

G_5: *i* perceives that the state of affairs described by *A* is the case.

G_6: *i* perceives that all the agents of *N* perceive that *A* is the case.

Gilbert's definition appears to contain some redundancy, since presumably an agent would not perceive A unless A is the case. Gilbert is evidently trying to give a more explicit account of single agent knowledge than Lewis and Aumann give. For Gilbert, agent *i* knows that a proposition *E* is the case if, and only if, $\omega \in E$, that is, *E* is true, and either *i* perceives that the state of affairs *E* describes obtains or *i* can infer *E* as a consequence of other propositions *i* knows, given sufficient inferential capacity.

Like Lewis, Gilbert recognizes that human agents do not in fact have unlimited inferential capacity. To generate the infinite hierarchy of mutual knowledge, Gilbert introduces the device of an agent's *smooth-reasoner counterpart.* The smooth-reasoner counterpart i' of an agent *i* is an agent that draws every logical conclusion from every fact that *i* knows. Gilbert stipulates that i' does not have any of the constrains on time, memory, or reasoning ability that *i* might have, so i' can literally think through the infinitely many levels of a common knowledge hierarchy.

Definition 2.16
If a set of agents *N* are in a common knowledge situation $\mathcal{S}_N(A)$ with respect to *A*, then the corresponding set N' of their smooth-reasoner counterparts is in a *parallel situation* $\mathcal{S}'_{N'}(A)$ if, and only if, for each $i' \in N$,

G_1': i' can perceive anything that the counterpart *i* can perceive.

G_2': G_2 - G_6 obtain for i' with respect to *A* and N', same as for the counterpart *i* with respect to *A* and *N*.

G_3': i' perceives that all the agents of N' are smooth-reasoners.

From this definition we get the following immediate consequence:

Proposition 2.17
If a set of smooth-reasoner counterparts to a set *N* of agents are in a situation $\mathcal{S}'_{N'}(A)$ parallel to a common knowledge situation $\mathcal{S}_N(A)$ of *N*, then

$$\text{for all } m \in \mathbb{N} \text{ and for any } i_1', \ldots, i_m',\ \mathbf{K}_{i_1'}\mathbf{K}_{i_2'}\ldots\mathbf{K}_{i_m'}(A).$$

Consequently, $\mathbf{K}^m_N{}'(A)$ for any $m \in \mathbb{N}$.

Gilbert argues that, given $\mathcal{S}'_N{}'(A)$, the smooth-reasoner counterparts of the agents of N actually satisfy a much stronger condition, namely mutual knowledge $\mathbf{K}^{\alpha}_N{}'(A)$ to the level of any ordinal number α, finite or infinite. When this stronger condition is satisfied, the proposition A is said to be *open* to the agents of N*. With the concept of open*-ness, Gilbert gives her definition of common knowledge.

Definition 2.18
A proposition $E \subseteq \Omega$ is *Gilbert-common knowledge* among the agents of a set $N = \{1, \ldots, n\}$, if and only if,

G_1*: E is open* to the agents of N.

G_2*: For every $i \in N$, $\mathbf{K}_i(G_1^*)$.

$\mathbf{G}_N^*(E)$ denotes the proposition defined by G_1* and G_2* for a set N of A*-symmetric reasoners, so we can say that E is Lewis-common knowledge for the agents of N iff $\omega \in \mathbf{G}_N^*(E)$.

One might think that an immediate corollary to Gilbert's definition is that Gilbert-common knowledge implies the hierarchical common knowledge of Proposition 2.5. However, this claim follows only on the assumption that an agent knows all of the propositions that her smooth-reasoner counterpart reasons through. Gilbert does not explicitly endorse this position, although she correctly observes that Lewis and Aumann are committed to something like it.[15] Gilbert maintains that her account of common knowledge expresses our intuitions with respect to common knowledge better than Lewis' and Aumann's accounts, since the notion of open*-ness presumably makes explicit that when a proposition is common knowledge, it is "out in the open", so to speak.

3. Applications of Mutual and Common Knowledge

Readers primarily interested in philosophical applications of common knowledge may want to focus on the No Disagreement Theorem and Convention subsections. Readers interested in applications of common knowledge in game theory may continue with the Strategic Form Games, and Games of Perfect Information subsections.

3.1 The "No Disagreement" Theorem

Aumann (1976) originally used his definition of common knowledge to prove a celebrated result that says that in a certain sense, agents cannot "agree to disagree" about their beliefs, formalized as probability distributions, if they start with common prior beliefs. Since agents in a community often hold different opinions and know they do so, one might attribute such differences to the agents' having different private information. Aumann's surprising result is that even if agents condition their beliefs on private information, mere common knowledge of their conditioned beliefs and a common prior probability distribution implies that their beliefs cannot be different, after all!

Proposition 3.1
Let Ω be a finite set of states of the world. Suppose that

- Agents i and j have a common prior probability distribution $\mu(\cdot)$ over the events of Ω such that $\mu(\omega) > 0$, for each $\omega \in \Omega$, and
- It is common knowledge at ω that i's posterior probability of event E is $q_i(E)$ and that j's posterior probability of E is $q_j(E)$.

Then $q_i(E) = q_j(E)$.

Proof.
[Note that in the proof of this proposition, and in the sequel, $\mu(\cdot|B)$ denotes conditional probability; that is, given $\mu(B)>0$, $\mu(A|B) = \mu(A\cap B)/\mu(B)$.]

In a later article, Aumann (1987) argues that the assumptions that Ω is finite and that $\mu(\omega) > 0$ for each $\omega \in \Omega$ reflect the idea that agents only regard as "really" possible a finite collection of salient worlds to which they assign positive probability, so that one can drop the states with probability 0 from the description of the state space. Aumann also notes that this result implicitly assumes that the agents have common knowledge of their partitions, since a description of each possible world includes a description of the agents' possibility sets. And of course, this result depends crucially upon (i), which is known as the *common prior assumption* (CPA).

Aumann's "no disagreement" theorem has been generalized in a number of ways in the literature (McKelvey and Page 1986, Monderer and Samet 1989, Geanakoplos 1994). However, all of these "no disagreement" results raise the same philosophical puzzle raised by Aumann's original result: How are we to explain differences in belief? Aumann's result leaves us with two options: (1) admit that at some level, common knowledge of the agents' beliefs or how they form their beliefs fails, or (2) deny the CPA. For instance, agents in the real world often do not express their opinions probabilistically. If one agent announces'I believe that E is the case' while another announces'I doubt that E is the case', then they might attribute their divergent opinions to a lack of common knowledge of each other's true posteriors for E. Even if agents do assign precise posterior probabilities to an event, Aumann shows that if they have

merely first-order mutual knowledge of the posteriors, they can "agree to disagree". Suppose the following all hold:

$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$,
$\mathcal{H}_1 = \{\{\omega_1,\omega_2\}, \{\omega_3,\omega_4\}\}$
$\mathcal{H}_2 = \{\{\omega_1,\omega_2,\omega_3\}, \{\omega_4\}\}$
$\mu(\omega_i) = 1/4$

Then if $E = \{\omega_1,\omega_4\}$, then at ω_1, we have:

$q_1(E) = \mu(E \mid \{\omega_1,\omega_2\}) = 1/2$, and

$q_2(E) = \mu(E \mid \{\omega_1,\omega_2,\omega_3\}) = 1/3$

Moreover, at $\omega = \omega_1$, Agent 1 knows that $\mathcal{H}_2(\omega) = \{\omega_1,\omega_2,\omega_3\}$, so she knows that $q_2(E) = 1/3$. Agent 2 knows at ω_1 that either $\mathcal{H}_1(\omega) = \{\omega_1,\omega_2\}$ or $\mathcal{H}_1(\omega) = \{\omega_3,\omega_4\}$, so either way he knows that $q_1(E) = 1/2$. Hence the agents' posteriors are mutually known, and yet they are unequal. The reason for this is that the posteriors are not common knowledge. For Agent 2 does not know what Agent 1 thinks $q_2(E)$ is, since if $\omega = \omega_3$, which is consistent with what Agent 2 knows, then Agent 1 will believe that $q_2(E) = 1/3$ with probability 1/2 (if $\omega = \omega_3$) and $q_2(E) = 1$ with probability 1/2 (if $\omega = \omega_4$).

Aumann's result could fail if the agents' partitions are not common knowledge. For suppose in the example just given, the agents do not know each other's partitions. Then at $\omega = \omega_1$, if their posteriors are common knowledge, then Agent 1, who knows that $\omega \in \{\omega_1,\omega_2\}$, can explain Agent 2's posterior as the result of Agent 2 having observed either $\{\omega_1,\omega_2,\omega_3\}$, $\{\omega_1,\omega_2,\omega_4\}$, $\{\omega_1,\omega_3,\omega_4\}$ or $\{\omega_2,\omega_3,\omega_4\}$. Still another way Aumann's result might fail is if agents do not have common knowledge that they update their beliefs by Bayesian conditionalization. Then clearly, agents can explain divergent opinions as the result of others having modified their beliefs in the "wrong" way. However, there are cases in which none of these explanations will seem convincing. For instance, odds makers sometimes publicly announce different probabilities for an event, such as a particular winner of a prize at a forthcoming Academy Awards presentation, and they will know that none of them have *any* private information regarding the event. In cases such as this, the agents have common knowledge that they all have the same information structure and common knowledge of their posteriors. And knowing that they are all competent odds makers, they have common knowledge that they update by Bayesian conditionalization. Still, the odds makers' beliefs violate the conclusion of Aumann's result. More generally, denying the requisite common knowledge seems a rather *ad hoc* move. For instance, to deny that agents have common knowledge of information structures is simply to deny that agents can all infer the same conclusions regarding possible worlds as Aumann defines them. To deny that agents have common knowledge that they update their beliefs by Bayesian conditionalization is to assert that some believe that some might be updating their beliefs *incoherently*, in the sense that their belief updating leaves them open to a *Dutch book* (Skyrms

1984). As just noted, these failures of agents' beliefs in each others' competence do not fail in all cases. Why should one think that such failures of common knowledge provide a general explanation for divergent beliefs?

What of the second option, that is, denying the CPA?[16] The main argument put forward in favor of the CPA is that any differences in agents' probabilities should be the result of their having different information only, that is, there is no reason to think that the different beliefs that agents have regarding the same event are the result of anything other than their having different information. However, one can reply that this argument amounts simply to a restatement of the Harsanyi Doctrine.[17] And while defenders of the Harsanyi Doctrine may be right in thinking that there is apparently no compelling reason to think that agents' priors can be different, neither is there compelling reason to think they must be the same! In any event, while the controversy over the Harsanyi Doctrine remains unresolved, we can conclude that the "no disagreement" results have interesting implications for the viability of common knowledge and the very nature of probability. Defenders of the CPA take an *objectivist* view of probability, and by virtue of the "no disagreement" results are evidently committed to the view that common knowledge of agents beliefs and how they are formed is a rare phenomenon in the world. Those who are prepared to deny the CPA allow for a genuinely *subjectivist* conception of probability. They take the view that common knowledge of agents' beliefs and how they come by them can be a commonplace phenomenon, and that differences in opinion can stem from differences in (subjective) prior probabilities.

3.2 Convention

Schelling's Department Store problem of Example 1.5 is a very simple example in which the agents "solve" their coordination problem appropriately by establishing a *convention.* Using the vocabulary of game theory, Lewis (1969) defines a convention as a *strict coordination equilibrium* of a game which agents follow on account of their common knowledge that they all prefer to follow this coordination equilibrium. A coordination equilibrium of a game is a strategy combination such that no agent is better off if any agent unilaterally deviates from this combination. As with equilibria in general, a coordination equilibrium is *strict* if any agent who deviates unilaterally from the equilibrium is strictly worse off. The strategic form game of Figure 1.3 summarizes Liz's and Robert's situation. The Department Store game has four Nash equilibrium outcomes in pure strategies: (s_1, s_1), (s_2, s_2), (s_3, s_3), and (s_4, s_4).[18] These four equilibria are all strict coordination equilibria. If the agents follow either of these equilibria, then they coordinate successfully. For agents to be following a Lewis-convention in this situation, they must follow one of the game's coordination equilibria. However, for Lewis to follow a coordination equilibrium is not a sufficient condition for agents to be following a convention. For suppose that Liz and Robert fail to analyze their predicament properly at all, but Liz chooses s_2 and Robert chooses s_2, so that they coordinate at (s_2, s_2) by sheer luck. Lewis does not count accidental coordination of this sort as a convention.

Suppose next that both agents are Bayesian rational, and that part of what each agent knows is the payoff structure of the Intersection game. If the agents expect each other to follow (s_2, s_2) and they consequently coordinate successfully, are they then following a convention? Not necessarily, contends Lewis, in a

subtle argument on p. 59 of *Convention*. For while each knows the game and that she is rational, she might not attribute like knowledge to the other agent. If each agent believes that the other agent will follow her end of the (s_2, s_2) equilibrium mindlessly, then her best response is to follow her end of (s_2, s_2). But in this case the agents coordinated as the result of their each falsely believing that the other acts like an automaton, and Lewis thinks that any proper account of convention must require that agents have *correct* beliefs about one another. In particular, Lewis requires that each agent involved in a convention must have mutual expectations that each is acting with the aim of coordinating with the other, that is, that each knows that:

> A_1: Both are rational,
> A_2: Both know the payoff structure of the game, and
> A_3: Both intend to follow (s_2, s_2), and not some other strategy combination.

Suppose that the agents' beliefs are appropriately augmented so that each agent knows that A_1, A_2, and A_3 are the case. Again they coordinate on (s_2, s_2). Are they following a convention this time? Still not necessarily, says Lewis. For what if it turns out that Liz thinks that Robert does not know that they are both rational? Then Liz has a false belief about Robert. Beyond this, there are two other points which Lewis does not himself raise in this argument, but which clearly support his view. First, it would be counterintuitive, at the very least, to suppose that any agent following a convention believes that he has reasoning abilities that the other agents lack. If Liz has determined that A_1, A_2, and A_3 are the case, then if they are following a convention she should expect that Robert has arrived at the same conclusion. Second, what could explain Liz's knowledge of A_3? The most natural explanation for Liz's expectation that Robert will follow his end of (s_2, s_2) is that Liz knows that Robert knows that A_1, A_2, and A_3 are the case. So convention evidently involves agents having at least *second-order* mutual knowledge of A_1, A_2, and A_3, that is, Robert (Liz) must know that Liz (Robert) knows that A_1, A_2, and A_3 are the case. But this raises the question: Can *third-order* mutual knowledge that A_1, A_2, and A_3 obtain fail? No, argues Lewis. For if Robert thought that Liz did not know that Robert knew that A_1, A_2, and A_3 were the case, then Robert would have a false belief about Liz. The additional supporting points also kick in again: If Robert has second-order mutual knowledge that A_1, A_2, and A_3 obtain, then he should conclude that Liz also has this second-order mutual knowledge. To conclude otherwise would require Robert to assume, counterintuitively, that he has analyzed their deliberations in this situation in a way that Liz cannot. And how did Robert get his second-order mutual knowledge of A_3? The most obvious way to account for Robert's second-order mutual knowledge would be to attribute to Robert the knowledge that Liz has second-order mutual knowledge that A_1, A_2, and A_3 are the case. So convention requires third-order mutual knowledge that A_1, A_2, and A_3 are the case. And the argument can be continued for any higher level of mutual knowledge.

Lewis concludes that a necessary condition for agents to be following a convention is that their preferences to follow the corresponding coordination equilibrium be common knowledge. So on Lewis' account, a convention for a set of agents is a coordination equilibrium which the agents follow on account

of their common knowledge of their rationality, the payoff structure of the relevant game and that each agent follows her part of the equilibrium.

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in any instance of S among the members of P,
>
> 1. everyone conforms to R;
> 2. everyone expects everyone else to conform to R;
> 3. everyone has approximately the same preferences regarding all possible combinations of actions;
> 4. everyone prefers that everyone conform to R, on condition that at least all but one conform to R;
> 5. everyone would prefer that everyone conform to R', on condition that at least all but one conform to R',
>
> where R' is some possible regularity in the behavior of members of P in S, such that no one in any instance of S among members of P could conform both to R' and to R.
> (Lewis 1969, p. 76)[19]

Lewis includes the requirement that there be an alternate coordination equilibrium R' besides the equilibrium R that all follow in order to capture the fundamental intuition that how the agents who follow a convention behave depends crucially upon how they expect the others to behave. In the Department Store game, the (s_2, s_2) equilibrium is a Lewis-convention when Liz and Robert have common knowledge of A_1, A_2, and A_3. Had their expectations been different, so either had believed that the other would not follow (s_2, s_2), then the outcome might have been very different.

Sugden (1986) and Vanderschraaf (1997) argue that it is not crucial to the notion of convention that the corresponding equilibrium be a coordination equilibrium. Lewis' key insight is that a convention is a pattern of mutually beneficial behavior which depends on the agents' common knowledge that all follow *this* pattern, and no other. Vanderschraaf gives a more general definition of convention as a *strict* equilibrium together with common knowledge that all follow this equilibrium and that all would have followed a different equilibrium had their beliefs about each other been different. An example of this more general kind of convention is given below in the discussion of the Figure 3.1 example.

3.3 Strategic Form Games

Lewis formulated the notion of common knowledge as part of his general account of conventions. In the years following the publication of *Convention*, game theorists have recognized that any explanation of a particular pattern of play in a game depends crucially on mutual and common knowledge assumptions. More specifically, *solution concepts* in game theory are both motivated and justified in large part by the mutual or common knowledge the agents in the game have regarding their situation.

To establish the notation that will be used in the discussion that follows, the usual definitions of a game in strategic form, expected utility and agents' distributions over their opponents' strategies, are given here:

Definition 3.2
A *game* Γ is an ordered triple $(N, S, \boldsymbol{u})$ consisting of the following elements:

- A finite set $N = \{1,2, \ldots , n\}$, called the *set of agents* or *players.*
- For each agent $k \in N$, there is a finite set $S_k = \{s_{k1},s_{k2}, \ldots , s_{kn_k}\}$, called the *alternative pure strategies* for agent k. The Cartesian product $S = S_1 \times \ldots \times S_n$ is called the *pure strategy set* for the game Γ.
- A map $\boldsymbol{u} : S \rightarrow \mathbb{R}^n$, called the *utility* or *payoff function* on the pure strategy set. At each strategy combination $\boldsymbol{s} = (s_{1\,j_1}, \ldots , s_{n\,j_n}) \in S$, agent k's particular payoff or utility is given by the k^{th} component of the value of $\boldsymbol{u}$, that is, agent k's utility u_k at $\boldsymbol{s}$ is determined by

$$u_k(\boldsymbol{s}) = I_k(\boldsymbol{u}(s_{1\,j_1}, \ldots , s_{n\,j_n}))$$

where $I_k(\boldsymbol{x})$ projects $\boldsymbol{x} \in \mathbb{R}^n$ onto its k^{th} component.

The subscript '$-k$' indicates the result of removing the k^{th} component of an n-tuple or an n-fold Cartesian product. For instance,

$$S_{-k} = S_1 \times \ldots \times S_{k-1} \times S_{k+1} \times \ldots \times S_n$$

denotes the pure strategy combinations that agent k's opponents may play.

Now let us formally introduce a system of the agents' beliefs into this framework. $\Delta_k(S_{-k})$ denotes the set of probability distributions over the measurable space $(S_{-k}, \mathfrak{F}_k)$, where $\mathfrak{F}_k$ denotes the Boolean algebra generated by the strategy combinations S_{-k}. Each agent k has a probability distribution $\mu_k \in \Delta_k(S_{-k})$, and this distribution determines the *(Savage) expected utilities* for each of k's possible acts:

$$E(u_k(s_{k\,j})) = \sum_{A_{-k} \in S_{-k}} u_k(s_{k\,j}, \boldsymbol{s}_{-k})\, \mu_k(\boldsymbol{s}_{-k}), \quad j = 1, 2, \ldots , n_k$$

If i is an opponent of k, then i's individual strategy $s_{i\,j}$ may be characterized as a union of strategy combinations $\bigcup\{ \boldsymbol{s}_{-k} \mid s_{i\,j} \in \boldsymbol{s}_{-k} \} \in \mathfrak{F}_k$, and so k's marginal probability for i's strategy $s_{i\,j}$ may be calculated as follows:

$$\mu_k(s_{ij}) = \sum_{\{ s_{-k} \mid s_{ij} \in s_{-k} \}} \mu_k(s_{-k})$$

$\mu_k(\cdot \mid A)$ denotes k's conditional probability distribution given a set A, and $E(\cdot \mid A)$ denotes k's conditional expectation given $\mu_k(\cdot \mid A)$.

Suppose first that the agents have common knowledge of the full payoff structure of the game they are engaged in and that they are all rational, and that no other information is common knowledge. In other words, each agent knows that her opponents are expected utility maximizers, but does not in general know exactly which strategies they will choose or what their probabilities for her acts are. These common knowledge assumptions are the motivational basis for the solution concept for noncooperative games known as *rationalizability*, introduced independently by Bernheim (1984) and Pearce (1984). Roughly speaking, a *rationalizable strategy* is any strategy an agent may choose without violating common knowledge of Bayesian rationality. Bernheim and Pearce argue that when only the structure of the game and the agents' Bayesian rationality are common knowledge, the game should be considered "solved" if every agent plays a rationalizable strategy. For instance, in the "Chicken" game with payoff structure defined by Figure 3.1,

		Joanna s_1	Joanna s_2
Lizzi	s_1	(3, 3)	(2, 4)
	s_2	(4, 2)	(0, 0)

s_1 = cooperate, s_2 = defect

Figure 3.1

if Joanna and Lizzi have common knowledge of all of the payoffs at every strategy combination, and they have common knowledge that both are Bayesian rational, then any of the four pure strategy profiles is rationalizable. For if their beliefs about each other are defined by the probabilities

$\alpha_1 = \mu_1$ (Joanna plays s_1), and
$\alpha_2 = \mu_2$ (Lizzi plays s_1)

then

$$E(u_i(s_1)) = 3\alpha_i + 2(1 - \alpha_i) = \alpha_i + 2$$

and

$$E(u_i(s_2)) = 4\alpha_i + 0(1 - \alpha_i) = 4\alpha_i, \quad i = 1, 2$$

so each agent maximizes her expected utility by playing s_1 if $\alpha_i + 2 \geq 4\alpha_i$ or $\alpha_i \leq 2/3$ and maximizes her expected utility by playing s_2 if $\alpha_i \geq 2/3$. If it so happens that $\alpha_i > 2/3$ for both agents, then both conform with Bayesian rationality by playing their respective ends of the strategy combination (s_2,s_2) *given their beliefs*, even though each would want to defect from this strategy combination were she to discover that the other is in fact going to play s_2. Note that the game's pure strategy Nash equilibria, (s_1, s_2) and (s_2, s_1), are rationalizable, since it is rational for Lizzi and Joanna to conform with either equilibrium given appropriate distributions. In general, the set of a game's rationalizable strategy combinations contains the set of the game's pure strategy Nash equilibria, and this example shows that the containment can be proper.

To show that rationalizability is a nontrivial notion, consider the 2-agent game with payoff structure defined by Figure 3.2a:

		Joanna		
		s_1	s_2	s_3
	s_1	(4, 3)	(1, 2)	(3, 4)
Lizzi	s_2	(1, 1)	(0, 5)	(1, 1)
	s_3	(3, 4)	(1, 3)	(4, 3)

Figure 3.2a

In this game, s_1 and s_3 strictly dominate s_2 for Lizzi, so Lizzi cannot play s_2 on pain of violating Bayesian rationality. Joanna knows this, so Joanna knows that the only pure strategy profiles which are possible outcomes of the game will be among the six profiles in which Lizzi does not choose s_2. In effect, the 3×3 game is reduced to the 2×3 game defined by Figure 3.2b:

Lizzi \ Joanna	s_1	s_2	s_3
s_1	(4, 3)	(1, 2)	(3, 4)
s_3	(3, 4)	(1, 3)	(4, 3)

Figure 3.2b

In this reduced game, s_2 is strictly dominated for Joanna by s_1, and so Joanna will rule out playing s_2. Lizzi knows this, and so she rules out strategy combinations in which Joanna plays s_2. The rationalizable strategy profiles are the four profiles that remain after deleting all of the strategy combinations in which either Lizzi or Joanna play s_2. In effect, common knowledge of Bayesian rationality reduces the 3 × 3 game of Figure 3.2a to the 2 × 2 game defined by Figure 3.2c:

Lizzi \ Joanna	s_1	s_3
s_1	(4, 3)	(3, 4)
s_3	(3, 4)	(4, 3)

Figure 3.2c

since Lizzi and Joanna both know that the only possible outcomes of the game are (s_1, s_1), (s_1, s_3), (s_3, s_1), and (s_3, s_3).

Rationalizability can be defined formally in several ways. A variation of Bernheim's original (1984) definition is given here.

Definition 3.3
Given that each agent $k \in N$ has a probability distribution $\mu_k \in \Delta_k(s_{-k})$, the system of beliefs

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) \in \Delta_1(S_{-1}) \times \ldots \times \Delta_n(S_{-n})$$

is *Bayes concordant* if and only if,

(3.i) For $i \neq k$, $\mu_i(s_{k\,j}) > 0 \Rightarrow s_{k\,j}$ maximizes k's expected utility for some $\sigma_k \in \Delta_k(s_{-k})$,

and (3.i) is common knowledge. A pure strategy combination $\boldsymbol{s} = (s_{1\,j_1}, \ldots , s_{n\,j_n}) \in S$ is *rationalizable* if and only if the agents have a Bayes concordant system $\boldsymbol{\mu}$ of beliefs and, for each agent $k \in N$,

(3.ii) $E(u_k(s_{k\,j_k})) \geq E(u_k(s_{k\,i_k}))$, for $i_k \neq j_k$.[20]

The following result shows that the common knowledge restriction on the distributions in Definition 3.1 formalizes the assumption that the agents have common knowledge of Bayesian rationality.

Proposition 3.4
In a game Γ, common knowledge of Bayesian rationality is satisfied if, and only if, (3.i) is common knowledge.

Proof.

When agents have common knowledge of the game and their Bayesian rationality only, one can predict that they will follow a rationalizable strategy profile. However, rationalizability becomes an unstable solution concept if the agents come to know more about one another. For instance, in the Chicken example above with $\alpha_i > 2/3$, $i = 1, 2$, if either agent were to discover the other agent's beliefs about her, she would have good reason not to follow the (s_2, s_2) profile and to revise her own beliefs regarding the other agent. If, in the other hand, it so happens that $\alpha_1 = 1$ and $\alpha_2 = 0$, so that the agents maximize expected payoff by following the (s_2, s_1) profile, then should the agents discover their beliefs about each other, they will still follow (s_2, s_1). Indeed, if their beliefs are common knowledge, then one can predict with certainty that they will follow (s_2, s_1). The Nash equilibrium (s_2, s_1) is characterized by the belief distributions defined by $\alpha_1 = 1$ and $\alpha_2 = 0$.

The Nash equilibrium is a special case of *correlated equilibrium concepts*, which are defined in terms of the belief distributions of the agents in a game. In general, a correlated equilibrium-in-beliefs is a system of agents' probability distributions which remains stable given common knowledge of the game, rationality and the *beliefs themselves*. We will review two alternative correlated equilibrium concepts (Aumann 1974, 1987; Vanderschraaf 1995), and show how each generalizes the Nash equilibrium concept.

Definition 3.5
Given that each agent $k \in N$ has a probability distribution $\mu_k \in \Delta_k(s_{-k})$, the system of beliefs

$$\mu^* = (\mu_1^*, \ldots, \mu_n^*) \in \Delta_1(s_{-1}) \times \ldots \times \Delta_n(s_{-n})$$

is an *endogenous correlated equilibrium* if, and only if,

(3.iii) For $i \neq k$, $\mu_i^*(s_{k\,j}) > 0 \Rightarrow s_{k\,j}$ maximizes k's expected utility given μ_k^*.

If μ^* is an endogenous correlated equilibrium a pure strategy combination $\boldsymbol{s}^* = (s_1^*, \ldots, s_n^*) \in$ S is an *endogenous correlated equilibrium strategy combination given* μ^* if, and only if, for each agent $k \in N$,

(3.iv) $E(u_k(s_k^*)) \geq E(u_k(s_{k\,i}))$ for $s_{k\,i} \neq s_k^*$.

Hence, the endogenous correlated equilibrium μ^* restricts the set of strategies that the agents might follow, as do the Bayes concordant beliefs of rationalizability. However, the endogenous correlated equilibrium concept is a proper refinement of rationalizability, because the latter does not presuppose that condition (3.iii) holds with respect to the beliefs one's opponents actually have. If exactly one pure strategy combination $\boldsymbol{s}^*$ satisfies (3.iv) given μ^*, then μ^* is a *strict equilibrium,* and in this case one can predict with certainty what the agents will do given common knowledge of the game, rationality and their beliefs. Note that Definition 3.5 says nothing about whether or not the agents regard their opponents' strategy combinations as probabilistically independent. Also, this definition does not require that the agents' probabilities are *consistent*, in the sense that agents' probabilities for a mutual opponent's acts agree. A simple refinement of the endogenous correlated equilibrium concept characterizes the Nash equilibrium concept.

Definition 3.6

A system of agents' beliefs μ^* is a *Nash equilibrium* if, and only if,

- condition (3.iii) is satisfied,
- For each $k \in N$, μ_k^* satisfies probabilistic independance, and
- For each $s_{k\,j} \in s_k$, if $i, l \neq k$ then $\mu_i^*(s_{k\,j}) = \mu_l^*(s_{k\,j})$.

In other words, an endogenous correlated equilibrium is a Nash equilibrium-in-beliefs when each agent regards the moves of his opponents as probabilistically independent and the agents' probabilities are consistent. Note that in the 2-agent case, conditions (b) and (c) of the Definition 3.6 are always satisfied, so for 2-agent games the endogenous correlated equilibrium concept reduces to the Nash equilibrium concept. Conditions (b) and (c) are traditionally assumed in game theory, but Skyrms (1991) and Vanderschraaf (1995) argue that there may be good reasons to relax these assumptions in games with 3 or more agents.

Brandenburger and Dekel (1988) show that in 2-agent games, if the beliefs of the agents are common knowledge, condition (3.iii) characterizes a Nash equilibrium-in-beliefs. As they note, condition (3.iii) characterizes a Nash equilibrium in beliefs for the *n*-agent case if the probability distributions are consistent and satisfy probabilistic independence. Proposition 3.7 extends Brandenburger and Dekel's result to the endogenous correlated equilibrium concept by relaxing the consistency and probabilistic independence assumptions.

Proposition 3.7
Assume that the probabilities

$$\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n) \in \Delta_1(s_{-1}) \times \ldots \times \Delta_n(s_{-n})$$

are common knowledge. Then common knowledge of Bayesian rationality is satisfied if, and only if, $\boldsymbol{\mu}$ is an endogenous correlated equilibrium.

Proof.

In addition, we have:

Corollary 3.8 (Brandenburger and Dekel, 1988)
Assume in a 2-agent game that the probabilities

$$\boldsymbol{\mu} = (\mu_1,\mu_2) \in \Delta_1(s_{-1}) \times \Delta_2(s_{-2})$$

are common knowledge. Then common knowledge of Bayesian rationality is satisfied if, and only if, $\boldsymbol{\mu}$ is a Nash equilibrium.

Proof.
The endogenous correlated equilibrium concept reduces to the Nash equilibrium concept in the 2-agent case, so the corollary follows by Proposition 3.7.

If $\boldsymbol{\mu}*$ is a strict equilibrium, then one can predict which pure strategy profile the agents in a game will follow given common knowledge of the game, rationality and $\boldsymbol{\mu}*$. But if $\boldsymbol{\mu}*$ is such that several distinct pure strategy profiles satisfy (3.iv) with respect to $\boldsymbol{\mu}*$, then one can no longer predict with certainty what the agents will do. For instance, in the Chicken game of Figure 3.1, the belief distributions defined by $\alpha_1 = \alpha_2 = 2/3$ together are a Nash equilibrium-in-beliefs. Given common knowledge of this equilibrium, either pure strategy is a best reply for each agent, in the sense that either pure strategy maximizes expected utility. Indeed, if agents can also adopt randomized or *mixed* strategies at which they follow one of several pure strategies according to the outcome of a chance experiment, then any of the infinitely mixed strategies an agent might adopt in Chicken is a best reply given $\boldsymbol{\mu}*$.[21] So the endogenous

correlated equilibrium concept does not determine the exact outcome of a game in all cases, even if one assumes probabilistic consistency and independence so that the equilibrium is a Nash equilibrium.

Another correlated equilibrium concept formalized by Aumann (1974, 1987) does give a determinate prediction of what agents will do in a game given appropriate common knowledge. To illustrate Aumann's correlated equilibrium concept, let us consider the Figure 3.1 game once more. If Joanna and Lizzi can tie their strategies to their knowledge of the possible worlds in a certain way, they can follow a system of correlated strategies which will yield a payoff vector they both prefer to that of the mixed Nash equilibrium and which is itself an equilibrium. One way they can achieve this is to have their friend Ron play a variation of the familiar shell game by hiding a pea under one of three walnut shells, numbered 1, 2 and 3. Joanna and Lizzi both think that each of the three relevant possible worlds corresponding to ω_k = {the pea lies under shell k} is equally likely. Ron then gives Lizzi and Joanna each a private recommendation, based upon the outcome of the game, which defines a system of strategy combinations f as follows

$$(\bigstar)\ f(\omega) = \begin{cases} (s_1,s_1) \text{ if } \omega_k = \omega_1 \\ (s_1,s_2) \text{ if } \omega_k = \omega_2 \\ (s_2,s_1) \text{ if } \omega_k = \omega_3 \end{cases}$$

f is a *correlated* strategy system because the agents tie their strategies, by following their recommendations, to the same set of states of the world Ω. *f* is also a strict *Aumann correlated equilibrium*, for if each agent knows how Ron makes his recommendations, but knows only the recommendation he gives her, either would do strictly worse were she to deviate from her recommendation.[22] Since there are several strict equilibria of Chicken, *f* corresponds to a convention as defined in Vanderschraaf (1997). The overall expected payoff vector of *f* is (3,3), which lies outside the convex hull of the payoffs for the game's Nash equilibria and which Pareto-dominates the expected payoff vector (4/3, 4/3), of the mixed Nash equilibrium defined by α_1 = 2/3, *i* = 1, 2.[23] The correlated equilibrium f is characterized by the probability distribution of the agents' play over the strategy profiles, given in Figure 3.3:

Lizzi \ Joanna	s_1	s_2
s_1	$\frac{1}{3}$	$\frac{1}{3}$
s_2	$\frac{1}{3}$	0

Figure 3.3

Aumann (1987) proves a result relating his correlated equilibrium concept to common knowledge. To review this result, we must give the formal definition of Aumann correlated equilibrium.

Definition 3.9
Given a game $\Gamma = (N, S, \boldsymbol{u})$ together with a finite set of possible worlds Ω, the vector valued function $f\colon \Omega \rightarrow S$ is a *correlated n-tuple*. If $f(\omega) = (f_1(\omega), \ldots, f_n(\omega))$ denotes the components of f for the agents of N, then agent k's *recommended strategy* at ω is $f_k(\omega)$. f is an *Aumann correlated equilibrium* iff

$$E(u_k \circ f) \geq E(u_k(f_{-k}, g_k)),$$

for each $k \in N$ and for any function g_k that is a function of f_i.

The agents are at Aumann correlated equilibrium if at each possible world $\omega \in \Omega$, no agent will want to deviate from his recommended strategy, given that the others follow their recommended strategies. Hence, Aumann correlated equilibrium uniquely specifies the strategy of each agent, by explicitly introducing a space of possible worlds to which agents can correlate their acts. The deviations g_i are required to be functions of f_i, that is, compositions of some other function with f_i, because i is informed of $f_i(\omega)$ only, and so can only distinguish between the possible worlds of Ω that are distinguished by f_i. As noted already, the primary difference between Aumann's notion of correlated equilibrium and the endogenous correlated equilibrium is that in Aumann's correlated equilibrium, the agents correlate their strategies to some event $\omega \in \Omega$ that is external to the game. One way to view this difference is that agents who correlate their strategies exogenously can calculate their expected utilities conditional on their own strategies.

In Aumann's model, a description of each possible world ω includes descriptions of the following: the game Γ, the agent's private information partitions, and the actions chosen by each agent at ω, and each agent's prior probability distribution $\mu_k(\cdot)$ over Ω. The basic idea is that conditional on ω, everyone knows everything that can be the object of uncertainty on the part of any agent, but in general, no agent necessarily knows which world ω is the actual world. The agents can use their priors to calculate the probabilities that the various act combinations $\boldsymbol{s} \in S$ are played. If the agents' priors are such that for all i, $j \in N$, $\mu_i(\omega) = 0$ iff $\mu_j(\omega) = 0$, then the agents' priors are *mutually absolutely continuous*. If the agents' priors all agree, that is, $\mu_1(\omega) = \ldots = \mu_n(\omega) = \mu(\omega)$ for each $\omega \in \Omega$, then it is said that the *common prior assumption*, or CPA, is satisfied. If agents are following an Aumann correlated equilibrium f and the CPA is satisfied, then f is an *objective* Aumann correlated equilibrium. An Aumann correlated equilibrium is a Nash equilibrium if the CPA is satisfied and the agents' distributions satisfy probabilistic independence.[24]

Let $s_i(\omega)$ denote the strategy chosen by agent i at possible world ω. Then $s\colon \Omega \rightarrow S$ defined by $s(\omega) = (s_1(\omega),\ldots,s_n(\omega))$ is a correlated n-tuple. Given that $\mathcal{H}_i$ is a partition of Ω,[25] the function $s_i\colon \Omega \rightarrow s_i$

defined by s is $\mathcal{H}_i$-*measurable* if for each $\mathcal{H}_{ij} \in \mathcal{H}_i$, $s_i(\omega')$ is constant for each $\omega' \in \mathcal{H}_{ij}$. $\mathcal{H}_i$-measurability is a formal way of saying that i knows what she will do at each possible world, given her information.

Definition 3.10
Agent i is *Bayes rational* with respect to $\omega \in \Omega$ (alternatively, ω-*Bayes rational*) iff s_i is $\mathcal{H}_i$-measurable and

$$E(u_i \circ s \mid \mathcal{H}_i)(\omega) \geq E(u_i(v_i, s_{-\mathrm{i}}) \mid \mathcal{H}_i)(\omega)$$

for any $\mathcal{H}_i$-measurable function $v_i : \Omega \rightarrow s_i$.

Note that Aumann's definition of ω-Bayesian rationality implies that $\mu_i(\mathcal{H}_i(\omega)) > 0$, so that the conditional expectations are defined. Aumann's main result, given next, implicitly assumes that $\mu_i(\mathcal{H}_i(\omega)) > 0$ for every agent $i \in N$ and every possible world $\omega \in \Omega$. This poses no technical difficulties if the CPA is satisfied, or even if the priors are only mutually absolutely continuous, since if this is the case then one can simply drop any ω with zero prior from consideration.

Proposition 3.11 (Aumann 1987)
If each agent $i \in N$ is ω-Bayes rational at each possible world $\omega \in \Omega$, then the agents are following an Aumann correlated equilibrium. If the CPA is satisfied, then the correlated equilibrium is objective.

Proof.

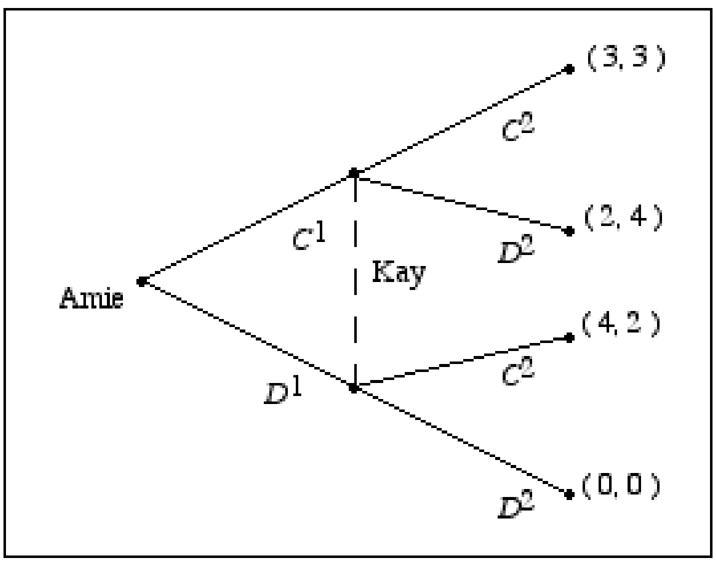
Part of the uncertainty the agents might have about their situation is whether or not all agents are rational. But if it is assumed that all agents are ω-Bayesian rational at each $\omega \in \Omega$, then a description of this fact forms part of the description of each possible ω and thus lies in the meet of the agents' partitions. As noted already, descriptions of the agents' priors, their partitions and the game also form part of the description of each possible world, so propositions corresponding to these facts also lie in the meet of the agents' partitions. So another way of stating Aumann's main result is as follows: *Common knowledge of ω-Bayesian rationality at each possible world implies that the agents follow an Aumann correlated equilibrium.*

Propositions 3.7 and 3.11 are powerful results. They say that common knowledge of rationality and of agents beliefs about each other, quantified as their probability distributions over the strategy profiles they might follow, implies that the agents' beliefs characterize an equilibrium of the game. Then if the agents' beliefs are unconditional, Proposition 3.7 says that the agents are rational to follow a strategy profile consistent with the corresponding endogenous correlated equilibrium. If their beliefs are conditional on their private information partitions, then Proposition 3.11 says they are rational to follow the strategies the corresponding Aumann correlated equilibrium recommends. However, we must not overestimate the

importance of these results, for they say nothing about the *origins* of the common knowledge of rationality and beliefs. For instance, in the Chicken game of Figure 3.1, we considered an example of a correlated equilibrium in which it was *assumed* that Lizzi and Joanna had common knowledge of the system of recommended strategies defined by (★). Given this common knowledge, Joanna and Lizzi indeed have decisive reason to follow the Aumann correlated equilibrium f. But where did this common knowledge come from? How, in general, do agents come to have the common knowledge which justifies their conforming to an equilibrium? Philosophers and social scientists have made only limited progress in addressing this question.

3.4 Games of Perfect Information

In extensive form games, the agents move in sequence. At each stage, the agent who is to move must base her decisions upon what she knows about the preceding moves. This part of the agent's knowledge is characterized by an *information set*, which is the set of alternative moves that an agent knows her predecessor might have chosen. For instance, consider the extensive form game of Figure 3.4:



C^i = "cooperate", D^i = "defect"

Figure 3.4

When Joanna moves she is at her information set $I^{22} = \{C^1, D^1\}$, that is, she moves knowing that Lizzi might have chosen either C^1 or D^1, so this game is an extensive form representation of the Chicken game of Figure 3.1.

In a game of perfect information, each information set consists of a single node in the game tree, since by definition at each state the agent who is to move knows exactly how her predecessors have moved. In Example 1.4 it was noted that the method of backwards induction can be applied to any game of perfect

information.[26] The backwards induction solution is the unique Nash equilibrium of a game of perfect information. The following result gives sufficient conditions to justify backwards induction play in a game of perfect information:

> **Proposition 3.12** (Bicchieri 1993)
> In an extensive form game of perfect information, the agents follow the backwards induction solution if the following conditions are satisfied for each agent i at each information set I^{ik}:
>
> - i is rational, i knows this and i knows the game, and
> - At any information set I^{jk+1} that immediately follows I^{ik}, i knows at I^{ik} what j knows at I^{jk+1}.
>
> Proof.

Proposition 3.12 says that far less than common knowledge of the game and of rationality suffices for the backwards induction solution to obtain in a game of perfect information. All that is needed is for each agent at each of her information sets to be rational, to know the game and to know what the next agent to move knows! For instance, in the Figure 1.2 game, if R_1 (R_2) stands for "Alan (Fiona) is rational" and $\mathbf{K}_i(\Gamma)$ stands for "i knows the game Γ", then the backwards induction solution is implied by the following:

- At I^{24}, R_2 and $\mathbf{K}_2(\Gamma)$.
- At I^{13}, R_1, $\mathbf{K}_1(\Gamma)$, $\mathbf{K}_1(R_2)$, and $\mathbf{K}_1\mathbf{K}_2(\Gamma)$.
- At I^{22}, $\mathbf{K}_2(R_1)$, $\mathbf{K}_2\mathbf{K}_1(R_2)$, and $\mathbf{K}_2\mathbf{K}_1\mathbf{K}_2(\Gamma)$.
- At I^{11}, $\mathbf{K}_1\mathbf{K}_2(R_1)$, $\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1(R_2)$, and $\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1\mathbf{K}_2(\Gamma)$.[27]

One might think that a corollary to Proposition 3.11 is that in a game of perfect information, common knowledge of the game and of rationality implies the backwards induction solution. This is the *classical argument* for the backwards induction solution. Many game theorists continue to accept the classical argument, but in recent years, the argument has come under strong challenge, led by the work of Reny (1987, 1992), Binmore (1987) and Bicchieri (1989, 1993). The basic idea underlying their criticisms of backwards induction can be illustrated with the Figure 1.2 game. According to the classical argument, if Alan and Fiona have common knowledge of rationality and the game, then each will predict that the other will follow her end of the backwards induction solution, to which his end of the backwards induction solution is his unique best response. However, what if Fiona reconsiders what to do if she finds herself at the information set I^{22}? If the information set I^{22} is reached, then Alan has of course not followed the backwards induction solution. If we assume that at I^{22}, Fiona knows only what is stated in (iii), then she can explain her being at I^{22} as a failure of either $\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1(R_2)$ or $\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1\mathbf{K}_2(\Gamma)$ at I^{11}. In this case, Fiona's thinking that either $\sim\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1(R_2)$ or $\sim\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1\mathbf{K}_2(\Gamma)$ at I^{11} is compatible with what Alan in fact does know at I^{11}, so Fiona should not necessarily be surprised to find herself at I^{22}, and given that what

she knows there is characterized by (iii), following the backwards induction solution is her best strategy. But if rationality and the game are common knowledge, or even if Fiona and Alan both have just have mutual knowledge of the statements characterized by (iii) and (iv), then at I^{22}, Fiona knows that $\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1(R_2)$ or $\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1\mathbf{K}_2(\Gamma)$ at I^{11}. Hence given this much mutual knowledge, Fiona no longer can explain why Alan has deviated from the backwards induction solution, since this deviation contradicts part of what is their mutual knowledge. So if she finds herself at I^{22}, Fiona does not necessarily have good reason to think that Alan will follow the backwards induction solution of the subgame beginning at I^{22}, and hence she might not have good reason to follow the backwards induction solution, either. Bicchieri (1993), who along with Binmore (1987) and Reny (1987, 1992) extends this argument to games of perfect information with arbitrary length, draws a startling conclusion: If agents have strictly too few or *strictly too many* levels of mutual knowledge of rationality and the game relative to the number of potential moves, one cannot predict that they will follow the backwards induction solution. This would undermine the central role backwards induction has played in the analysis of extensive form games. For why should the number of levels of mutual knowledge the agents have depend upon the length of the game?

The classical argument for backwards induction implicitly assumes that at each stage of the game, the agents discount the preceding moves as strategically irrelevant. Defenders of the classical argument can argue that this assumption makes sense, since by definition at any agents' decision node, the previous moves that led to this node are now fixed. Critics of the classical argument question this assumption, contending that when reasoning about how to move at any of his information sets, *including those not on the backwards induction equilibrium path*, part of what an agent must consider is what conditions might have led to his being at that information set. In other words, agents should incorporate reasoning about the reasoning of the previous movers, or *forward induction* reasoning, into their deliberations over how to move at a given information set. Binmore (1987) and Bicchieri (1993) contend that a backwards induction solution to a game should be consistent with the solution a corresponding forward induction argument recommends. As we have seen, given common knowledge of the game and of rationality, forward induction reasoning can lead the agents to an apparent contradiction: The classical argument for backwards induction is predicated on what agents predict they would do at nodes in the tree that are never reached. They make these predictions based upon their common knowledge of the game and of rationality. But forward induction reasoning seems to imply that if any off-equilibrium node had been reached, common knowledge of rationality and the game must have failed, so how could the agents have predicted what would happen at these nodes?

This section has barely scratched the surface of this controversy over common knowledge and backwards induction. The key unresolved issue is of course explaining what happens at the off-equilibrium information sets. To date, there is not a generally accepted theory of what agents having certain mutual or common knowledge will do at off-equilibrium nodes. However, we can at least repeat one generally accepted conclusion: In a game of perfect information, mutual knowledge of rationality and the game which falls far short of common knowledge can suffice to explain why agents follow the game's Nash equilibrium, the backwards induction solution. On the other hand, unlike other examples we have considered in which agents have mutual and even common knowledge without having to reason through levels of knowledge, backwards induction arguments in games of perfect information require that at each

information set, the agent who would move were the information set to be reached must reason her way through at least as many levels of knowledge as there are remaining potential moves in the game.

4. Is Common Knowledge Attainable?

Lewis formulated an account of common knowledge which generates the hierarchy of'*i* knows that *j* knows that … *k* knows that *A*' propositions in order to ensure that in his account of convention, agents have correct beliefs about each other. But since human agents obviously cannot reason their way through such an infinite hierarchy, it is natural to wonder whether any group of people can have full common knowledge of any proposition. More broadly, the analyses of common knowledge reviewed in §3 would be of little worth to social scientists and philosophers if this common knowledge lies beyond the reach of human agents.

Fortunately for Lewis' program, there are strong arguments that common knowledge is indeed attainable. Lewis (1969) and Schiffer (1972) argue that the common knowledge hierarchy should be viewed as a chain of implications, and not as steps in anyone's actual reasoning. They give informal arguments that the common knowledge hierarchy is generated from a finite set of axioms. We saw in §2 that it is possible to formulate Lewis' axioms precisely and to derive the common knowledge hierarchy from these axioms. Again, the basic idea behind Lewis' argument is that for a set of agents, if a proposition A is publicly known among them and each agent knows that everyone can draw the same conclusions from A that she can, then A is common knowledge. These conditions are obviously context dependent, just as an individual's knowing or not knowing a proposition is context dependent. Yet there are many cases where it is natural to assume that Lewis' conditions are satisfied. If, for instance, a group of English speaking persons in an automobile are listening to the radio, and the following special news announcement, "The Pope has abdicated", is audibly broadcast, then one may safely conclude that it is common knowledge for this group that the Pope has abdicated. If one has skeptical doubts about the agents' common knowledge in this situation, then one would have to explain the failure of common knowledge as the result of some circumstance that would be quite surprising in this context. Common knowledge could fail if some of the people failed to hear the announcement, or if some of them believed that some of the others could not understand the announcement, but circumstances such as these would be quite peculiar given the stated assumptions in this story. In this context, skeptical doubt about common knowledge is certainly possible, but such doubt relies upon *ad hoc* assumptions similar to those that are needed to explain failure of *individual* knowledge, not with the attainability of common knowledge in principle.

Aumann (1976) gives an alternate finitary procedure for generating the common knowledge hierarchy in the special case in which the relevant number of possible worlds in Ω is finite and each agent's information system partitions Ω. To be sure, knowledge does not always come so neatly packaged, but in many applications a finite state space together with partitions is a good model of the actual situation agents face. Aumann shows that a proposition *A* is common knowledge for a set *N* of agents at ω, if and only if, $\mathcal{M}(\omega) \subseteq A$ where $\mathcal{M}(\omega)$ is the element in the meet of the agents' private information partitions containing ω. In words, anything implied by the agents' common information partition is common

knowledge. If the set Ω is finite, then the meet $\mathcal{M}$ of the agents' partitions $\mathcal{H}_i$, $i \in N$, can be computed in finitely many steps. In a certain sense, the issue of skepticism regarding common knowledge never arises in Aumann's model. Common knowledge is built into Aumann's model, as a result of the agents' having private knowledge which is defined by *partitions* over the possible worlds. Put another way, common knowledge could fail in Aumann's model only if at some $\omega \in \Omega$, some individual i's knowledge of $\mathcal{H}_i(\omega)$ in i's private information partition could fail, which reinforces the point made in the previous paragraph. To reiterate, if one accepts Lewis' and Aumann's analysis of common knowledge, then common knowledge is in principle no more problematic than individual knowledge.

Nevertheless, care must be taken in ascribing common knowledge to a group of human agents. Common knowledge is a phenomenon highly sensitive to the agents' circumstances. The following section gives an example that shows that in order for *A* to be a common truism for a set of agents, they ordinarily must perceive an event which implies *A* *simultaneously* and *publicly*.

5. Coordination and Common *p*-Belief

In certain contexts, agents might not be able to achieve common knowledge. Might they achieve something "close"? One weakening of common knowledge is of course m^{th} level mutual knowledge. For a high value of *m*, $\mathbf{K}^m{}_N(A)$ might seem a good approximation of $\mathbf{K}^*{}_N(A)$. However, the following example, due to Rubinstein (1989, 1992), shows that simply truncating the common knowledge hierarchy at any finite level can lead agents to behave as if they had no mutual knowledge at all.[28]

5.1 The E-mail Coordination Example

Lizzi and Joanna are faced with the coordination problem summarized in the following figure:

ω_1, $\mu(\omega_1) = .51$

Lizzi \ Joanna	A	B
A	(2, 2)	(0, −4)
B	(−4, 0)	(0, 0)

ω_2, $\mu(\omega_2) = .49$

Lizzi \ Joanna	A	B
A	(0, 0)	(0, −4)
B	(−4, 0)	(2, 2)

Figure 5.1

In Figure 5.1, the payoffs are dependent upon a pair of possible worlds. World ω_1 occurs with probability $\mu(\omega_1) = .51$, while ω_2 occurs with probability $\mu(\omega_2) = .49$. Hence, they coordinate with complete success by both choosing *A* (*B*) only if the state of the world is ω_1 (ω_2).

Suppose that Lizzi can observe the state of the world, but Joanna cannot. We can interpret this game as follows: Joanna and Lizzi would like to have a dinner together prepared by Aldo, their favorite chef. Aldo alternates between *A* and *B*, the two branches of Sorriso, their favorite restaurant. State ω_i is Aldo's location that day. At state ω_1 (ω_2), Aldo is at *A* (*B*). Lizzi, who is on Sorriso's special mailing list, receives notice of ω_i. Lizzi's and Joanna's best outcome occurs when they meet where Aldo is working, so they can have their planned dinner. If they meet but miss Aldo, they are disappointed and do not have dinner after all. If either goes to *A* and finds herself alone, then she is again disappointed and does not have dinner. But what each really wants to avoid is going to *B* if the other goes to *A*. If either of them arrives at *B* alone, she not only misses dinner but must pay the exorbitant parking fee of the hotel which houses *B*, since the headwaiter of *B* refuses to validate the parking ticket of anyone who asks for a table for two and then sits alone. This is what Harsanyi (1967) terms a game of *incomplete information*, since the game's payoffs depend upon states which not all the agents know.

A is a "play-it-safe" strategy for both Joanna and Lizzi.[29] By choosing *A* whatever the state of the world happens to be, the agents run the risk that they will fail to get the positive payoff of meeting where Aldo is, but each is also sure to avoid the really bad consequence of choosing *B* if the other chooses *A*. And since only Lizzi knows the state of the world, neither can use information regarding the state of the world to improve their prospects for coordination. For Joanna has no such information, and since Lizzi knows this, she knows that Joanna has to choose accordingly, so Lizzi must choose her best response to the move she anticipates Joanna to make regardless of the state of the world Lizzi observes. Apparently Lizzi and Joanna cannot achieve expected payoffs greater than 1.02 for each, their expected payoffs if they choose (*A*, *A*) at either state of the world.

If the state ω were common knowledge, then the conditional strategy profile (*A*, *A*) if $\omega = \omega_1$ and (*B*, *B*), if $\omega = \omega_2$ would be a strict Nash equilibrium at which each would achieve a payoff of 2. So the obvious remedy to their predicament would be for Lizzi to tell Joanna Aldo's location in a face-to-face or telephone conversation and for them to agree to go where Aldo is, which would make the state ω and their intentions to coordinate on the best outcome given ω common knowledge between them. Suppose for some reason they cannot talk to each other, but they prearrange that Lizzi will send Joanna an e-mail message if, and only if, ω_2 occurs. Suppose further that Joanna's and Lizzi's e-mail systems are set up to send a reply message automatically to the sender of any message received and viewed, and that due to technical problems there is a small probability, $\varepsilon > 0$, that any message can fail to arrive at its destination. Then if Lizzi sends Joanna a message, and receives an automatic confirmation, then Lizzi knows that Joanna knows that ω_2 has occurred. If Joanna receives an automatic confirmation of Lizzi's automatic confirmation, then Joanna knows that Lizzi knows that Joanna knows that ω_2 occurred, and so on. That ω_2 has occurred would become common knowledge if each agent received infinitely many automatic confirmations, assuming that all the confirmations could be sent and received in a finite amount of time.[30] However, because of the probability ε of transmission failure at every stage of communication, the sequence of confirmations stops after finitely many stages with probability one. With probability one, therefore, the agents fail to achieve full common knowledge. But they do at least achieve something

"close" to common knowledge. Does this imply that they have good prospects of settling upon (*B*, *B*)?

Rubinstein shows by induction that if the number of automatically exchanged confirmation messages is finite, then *A* is the only choice that maximizes expected utility for each agent, given what she knows about what they both know.

Rubinstein's Proof

So even if agents have "almost" common knowledge, in the sense that the number of levels of knowledge in "Joanna knows that Lizzi knows that . . . that Joanna knows that ω_2 occurred" is very large, their behavior is quite different from their behavior given common knowledge that ω_2 has occurred. Indeed, as Rubinstein points out, given merely "almost" common knowledge, the agents choose as if no communication had occurred at all! Rubinstein also notes that this result violates our intuitions about what we would expect the agents to do in this case. (See Rubinstein 1992, p. 324.) If $T_i = 17$, wouldn't we expect agent *i* to choose *B*? Indeed, in many actual situations we might think it plausible that the agents would each expect the other to choose *B* even if $T_1 = T_2 = 2$, which is all that is needed for Lizzi to know that Joanna has received her original message and for Joanna to know that Lizzi knows this!

5.2 Common *p*-Belief

The example in Section 5.1 hints that mutual knowledge is not the only weakening of common knowledge that is relevant to coordination. Brandenburger and Dekel (1987), Stinchcombe (1988) and Monderer and Samet (1989) explore another option, which is to weaken the properties of the $\mathbf{K}^*_N$ operator. Monderer and Samet motivate this approach by noting that even if a mutual knowledge hierarchy stops at a certain level, agents might still have higher level mutual *beliefs* about the proposition in question. So they replace the knowledge operator $\mathbf{K}_i$ with a *belief operator* $\mathbf{B}^p_i$:

Definition 5.1
If $\mu_i(\cdot)$ is agent *i*'s probability distribution over Ω, then

$$\mathbf{B}^p_i(A) = \{ \omega \mid \mu_i(A \mid \mathcal{H}_i(\omega)) \geq p \}$$

$\mathbf{B}^p_i(A)$ is to be read '*i* believes *A* (given *i*'s private information) with probability at least *p* at ω', or '*i* *p*-believes *A*'. The belief operator $\mathbf{B}^p_i$ satisfies axioms K2, K3, and K4 of the knowledge operator. $\mathbf{B}^p_i$ does not satisfy K1, but does satisfy the weaker property

$$\mu_i(A \mid \mathbf{B}^p_i(A)) \geq p$$

that is, if one believes *A* with probability at least *p*, then the probability of *A* is indeed at least *p*.

One can define *mutual* and *common p-beliefs* recursively in a manner similar to the definition of mutual and common knowledge:

Definition 5.2
Let a set Ω of possible worlds together with a set of agents N be given.

(1) The proposition that A is *(first level or first order) mutual p-belief for the agents of N*, $\mathbf{B}^p{}_{\mathrm{N}}{}^1(A)$, is the set defined by

$$\mathbf{B}^p{}_{\mathrm{N}}{}^1(A) \equiv \bigcap_{i \in N} \mathbf{B}^p{}_i(A).$$

(2) The proposition that A is m^{th} *level* (or m^{th} *order*) *mutual p-belief among the agents of N*, $\mathbf{B}^p{}_{\mathrm{N}}{}^m(A)$, is defined recursively as the set

$$\mathbf{B}^p{}_{\mathrm{N}}{}^m(A) \equiv \bigcap_{i \in N} \mathbf{B}^p{}_i(\mathbf{B}^p{}_{\mathrm{N}}{}^{m-1}(A))$$

(3) The proposition that A is *common p-belief* among the agents of N, $\mathbf{B}^p{}_N{}^*(A)$, is defined as the set

$$\mathbf{B}^p{}_{\mathrm{N}}{}^*(A) \equiv \bigcap_{m=1}^{\infty} \mathbf{B}^p{}_N{}^m(A).$$

If A is common (or m^{th} level mutual) knowledge at world ω, then A is common (m^{th} level) p-belief at ω for every value of p. So mutual and common p-beliefs formally generalize the mutual and common knowledge concepts. However, note that $\mathbf{B}^1{}_N{}^*(A)$ is not necessarily the same proposition as $\mathbf{K}^*{}_N(A)$, that is, even if A is common 1-belief, A can fail to be common knowledge.

Common p-belief forms a hierarchy similar to a common knowledge hierarchy:

Proposition 5.3
$\omega \in \mathbf{B}^p{}_N{}^m(A)$ iff

(1) For all agents $i_1, i_2, \ldots, i_m \in N$, $\omega \in \mathbf{B}^p{}_{i_1}\mathbf{B}^p{}_{i_2} \ldots \mathbf{B}^p{}_{i_m}(A)$

Hence, $\omega \in \mathbf{B}^p{}_N{}^*(A)$ iff (1) is the case for each $m \geq 1$.

Proof. Similar to the Proof of Proposition 2.5.

One can draw several morals from the e-mail game of Example 5.1. Rubinstein (1987) argues that his conclusion seems paradoxical for the same reason the backwards induction solution of Alan's and Fiona's perfect information game might seem paradoxical: Mathematical induction does not appear to be part of our "everyday" reasoning. This game also shows that in order for A to be a common truism for a set of agents, they ordinarily must perceive an event which implies A *simultaneously* in each others' presence. A third moral is that in some cases, it may make sense for the agents to employ some solution concept weaker than Nash or correlated equilibrium. In their analysis of the e-mail game, Monderer and Samet (1989) introduce the notions of *ex ante* and *ex post* ε-*equilibrium.* An *ex ante* equilibrium *h* is a system of strategy profiles such that no agent *i* expects to gain more than ε-utiles if *i* deviates from *h*. An *ex post* equilibrium h' is a system of strategy profiles such that no agent *i* expects to gain more than ε-utiles by deviating from h' given *i*'s private information. When $\varepsilon = 0$, these concepts coincide, and *h* is a Nash equilibrium. Monderer and Samet show that, while the agents in the e-mail game can never achieve common knowledge of the world ω, if they have common *p*-belief of ω for sufficiently high *p*, then there is an *ex ante* equilibrium at which they follow (*A*,*A*) if $\omega = \omega_1$ and (*B*,*B*), if $\omega = \omega_2$. This equilibrium turns out not to be *ex post*. However, if the situation is changed so that there are no replies, then Lizzi and Joanna could have at most first order mutual knowledge that $\omega = \omega_2$. Monderer and Samet show that in this situation, given sufficiently high common *p*-belief that $\omega = \omega_2$, there is an *ex post* equilibrium at which Joanna and Lizzi choose (*B*,*B*) if $\omega = \omega_2$! So another way one might view this third moral of the e-mail game is that agents' prospects for coordination can sometimes improve dramatically if they rely on their common beliefs as well as their mutual knowledge.

Bibliography

Annotations

Lewis (1969) is the classic pioneering study of common knowledge and its potential applications to conventions and game theory. As Lewis acknowledges, parts of his work are foreshadowed in Hume (1740) and Schelling (1960).

Aumann (1976) gives the first mathematically rigorous formulation of common knowledge using set theory. Schiffer (1972) uses the formal vocabulary of *epistemic logic* (Hintikka 1962) to state his definition of common knowledge. Schiffer's general approach is to augment a system of sentential logic with a set of knowledge operators corresponding to a set of agents, and then to define common knowledge as a hierarchy of propositions in the augmented system. Bacharach (1992), Bicchieri (1993) and Fagin, et. al. (1995) adopt this approach, and develop logical theories of common knowledge which include soundness and completeness theorems. Fagin, et. al. show that the syntactic and set-theoretic approaches to developing common knowledge are logically equivalent.

Aumann (1995) gives a recent defense of the classical view of backwards induction in games of imperfect information. For criticisms of the classical view, see Binmore (1987), Reny (1992), Bicchieri (1989) and

especially Bicchieri (1993). Brandenburger (1992) surveys the known results connecting mutual and common knowledge to solution concepts in game theory. For more in-depth survey articles on common knowledge and its applications to game theory, see Binmore and Brandenburger (1989), Geanakoplos (1994) and Dekel and Gul (1996). For her alternate account of common knowledge along with an account of conventions which opposes Lewis' account, see Gilbert (1989).

Monderer and Samet (1989) remains one of the best resources for the study of common p-belief.

References

- Aumann, Robert. 1974, "Subjectivity and Correlation in Randomized Strategies", *Journal of Mathematical Economics* 1, 67-96.
- Aumann, Robert. 1976, "Agreeing to Disagree", *Annals of Statistics* 4, 1236-9.
- Aumann, Robert. 1987, "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica* 55, 1-18.
- Aumann, R. 1995. "Backward Induction and Common Knowledge of Rationality", *Games and Economic Behavior* 8: 6-19.
- Bacharach, Michael. 1989. "Mutual Knowledge and Human Reason", mimeo.
- Barwise, Jon. 1988. "Three Views of Common Knowledge", in *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, ed. M.Y. Yardi. San Francisco: Morgan Kaufman, pp. 365-379.
- Barwise, Jon. 1989.*The Situation in Logic.* Stanford: Center for the Study of Language and Information.
- Bernheim, B. Douglas. 1984. "Rationalizable Strategic Behavior", *Econometrica*, 52: 1007-1028.
- Bicchieri, Cristina. 1989. "Self Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge", *Erkenntnis*, 30: 69-85.
- Bicchieri, Cristina. 1993. Rationality and Coordination. Cambridge: Cambridge University Press.
- Binmore, Ken. 1987. "Modelling Rational Players I", *Economics and Philosophy*, 3: 179-241.
- Binmore, Ken. 1992. *Fun and Games.* Lexington, Massachusetts: D. C. Heath.
- Binmore, Ken and Brandenburger, Adam. 1988, "Common knowledge and Game theory" ST/ICERD Discussion Paper 88/167, London School of Economics.
- Brandenburger, Adam. 1992. "Knowledge and Equilibrium in Games", *Journal of Economic Perspectives* 6: 83-101.
- Brandenburger, Adam, and Dekel, Eddie. 1987. "Common knowledge with Probability 1", *Journal of Mathematical Economics* 16, 237-245.
- Brandenburger Adam and Dekel, Eddie. 1988. "The Role of Common Knowledge Assumptions in Game Theory", in *The Economics of Missing Markets, Information and Games*, ed. Frank Hahn. Oxford: The Clarendon Press: 46-61.
- Carnap, Rudolf. 1947, *Meaning and Necessity: A Study in Semantics and Modal Logic*, Chicago, University of Chicago Press.
- Dekel, Eddie and Gul, Faruk. 1996. "Rationality and Knowledge in Game Theory", working paper, Northwestern and Princeton Universities.
- Fagin, Ronald, Halpern, Joseph Y., Moses, Yoram and Vardi, Moshe Y. 1995. *Reasoning About*

Knowledge. Cambridge, Massachusetts: MIT Press.

- Geanakoplos, John. 1994. "Common Knowledge", in *Handbook of Game Theory,* Volume 2, ed. Robert Aumann and Sergiu Hart. Elsevier Science B.V.: 1438-1496.
- Gilbert, Margaret. 1989, *On Social Facts,* Princeton University Press, Princeton.
- Harsanyi, J. 1967. "Games with incomplete information played by "Bayesian" players, I: The basic model." *Management Science* 14: 159-82.
- Harsanyi, J. 1968a. "Games with incomplete information played by "Bayesian" players, II: Bayesian equilibrium points." *Management Science* 14: 320-324.
- Harsanyi, J. 1968b. "Games with incomplete information played by "Bayesian" players, III: The basic probability distribution of the game." *Management Science* 14: 486-502.
- Hintikka, Jaako. 1962. *Knowledge and Belief.* Ithaca, New York: Cornell University Press.
- Hume, David. (1740, 1888) 1976, *A Treatise of Human Nature,* ed. L. A. Selby-Bigge. rev. 2nd. ed., ed. P. H. Nidditch. Clarendon Press, Oxford.
- Lewis, C. I. 1943, "The Modes of Meaning", *Philosophy and Phenomenological Research,* 4, 236-250.
- Lewis, David. 1969, *Convention: A Philosophical Study,* Harvard University Press, Cambridge, Massachusetts.
- Littlewood, J. E. 1953. *Mathematical Miscellany,* ed. B. Bollobas.
- Mckelvey, Richard and Page, Talbot, "Common knowledge, consensus and aggregate information", *Econometrica* 54: 109-127.
- Milgrom, Paul. 1981. "An axiomatic characterization of common knowledge", *Econometrica* 49: 219-222.
- Monderer, Dov and Samet, Dov. 1989, "Approximating Common Knowledge with Common Beliefs", *Games and Economic Behavior* 1, 170-190.
- Nash, John. 1950, "Equilibrium points in n-person games". *Proceedings of the National Academy of Sciences of the United States* 36, 48-49.
- Nash, John. 1951, "Non-Cooperative Games". *Annals of Mathematics* 54, 286-295.
- Pearce, David. 1984. "Rationalizable Strategic Behavior and the Problem of Perfection". *Econometrica,* 52: 1029-1050.
- Reny, Philip. 1987. "Rationality, Common Knowledge, and the Theory of Games", working paper, Department of Economics, University of Western Ontario.
- Reny, Philip. 1992. "Rationality in Extensive Form Games", *Journal of Economic Perspectives,* 6: 103-118.
- Rubinstein, Ariel. 1987. "A Game with "Almost Common Knowledge": An Example", in *Theoretical Economics,* D. P. 87/165. London School of Economics.
- Schelling, Thomas. 1960, *The Strategy of Conflict,* Harvard University Press, Cambridge, Massachusetts.
- Schiffer, Stephen. 1972, *Meaning,* Oxford University Press, Oxford.
- Skyrms, Brian. 1984, *Pragmatics and Empiricism*, Yale University Press, New Haven.
- Stinchcombe, Max. 1988. "Approximate Common Knowledge", mimeo, University of California, San Diego.
- Sugden, Robert. 1986, *The Economics of Rights, Cooperation and Welfare,* Basil Blackwell, New York.

- Vanderschraaf, Peter. 1995, "Endogenous Correlated Equilibria in Noncooperative Games", *Theory and Decision* 38: 61-84.
- Vanderschraaf, Peter. 1995. *A Study in Inductive Deliberation*, Ph.D. thesis, Department of Philosophy, University of California at Irvine.
- Vanderschraaf, Peter. 1997, "Knowledge, Equilibrium and Convention", mimeo.
- von Neumann, John and Morgenstern, Oskar. 1944. *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.

Other Internet Resources

- Applications of Circumscription to Formalizing Common Sense Knowledge
- Reasoning About Common Knowledge with Infinitely Many Agents

Related Entries

game theory | prisoner’s dilemma

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 27, 2001
Content last modified: June 12, 2002

Stanford Encyclopedia of Philosophy
Notes to Common Knowledge

Notes

1. Thanks to Alan Hajek for this example, the only example in this section which does not appear elsewhere in the literature.

2. The version of the story Littlewood analyzes involves a group of cannibals, some of whom are marrried to unfaithful wives, and a missionary who visits the group and makes a public announcement of the fact.

3. Robert Vanderschraaf reminded me in conversation that a crucial assumption in this problem is that the cook is telling the diners the truth, that is, the cook's announcement generates common knowledge and not merely *common belief* that there is at least one messy individual. For if the agents believe the cook's announcements even if the cook does not reliably tell the truth, then should the cook mischeviously announce that there is at least one messy individual when in fact all are clean, all will wipe their faces at once.

4. The mutual knowledge characterized by (i), (ii), and (iii) is sufficient both to account for the agents' following the D^1,D^2-outcome, and for their being able to predict each others' moves. However, weaker knowledge assumptions imply that the agents will play D^1,D^2, even if they might not both be able to predict this outcome before the start of play. As Fiona's quoted argument implies, if both are rational, both know the game, and Fiona knows that Alan is rational and knows the game, then the D^1,D^2-outcome is the result, even if Alan does not know that Fiona is rational or knows the game

5. Hume's analysis of the Farmer's Dilemma is perhaps the earliest example of a backwards induction argument applied to a sequential decision problem. See Skyrms (1996) and Vanderschraaf (1996) for more extended discussions of this argument.

6. See §3 for a formal definition of the Nash equilibrium concept.

7. Aumann (1976) himself gives a set-theoretic account of common knowledge, which has been generalized in several articles in the literature, including Monderer and Samet (1988) and Binmore and Brandenburger (1989). Vanderschraaf (1997) gives the set-theoretic formulation of Lewis' account of common knowledge reviewed in this paper.

8. This result appears in several articles in the literature, including Monderer and Samet's and Binmore

and Brandenburger's articles on common knowledge.

9. I abuse notation slightly, writing '$\mathbf{K}_i\mathbf{K}_j(A)$' for '$\mathbf{K}_i(\mathbf{K}_j(A))$'.

10. A partition of a set Ω is a collection of sets $\mathcal{H} = \{H_1, H_2, \ldots\}$ such that $H_i \cap H_j = \varnothing$ for $i \neq j$, and $\cup_i H_i = \Omega$.

11. Thanks to Chris Miller and Jarah Evslin for suggesting the term 'symmetric reasoner' to decribe the parity of reasoning powers that Lewis relies upon in his treatment of common knowledge. Lewis does not explicitly include the notion of A'-symmetric reasoning into his definition of common knowledge, but he makes use of the notion implicitly in his argument for how his definition of common knowledge generates the mutual knowledge hierarchy.

12. The *meet* $\mathcal{M}$ of a collection $\mathcal{H}_i$, $i \in N$ of partitions is the finest common coarsening of the partitions. More specifically, for any $\omega \in \Omega$, if $\mathcal{M}(\omega)$ is the element of $\mathcal{M}$ containing ω, then

- $\mathcal{H}_i(\omega) \subseteq \mathcal{M}(\omega)$ for all $i \in N$, and
- For any other $\mathcal{M}'$ satisfying (i), $\mathcal{M}(\omega) \subseteq \mathcal{M}'(\omega)$.

13. B^c denotes the complement of B, that is $B^c = \omega - B = \{\omega \in \Omega : \omega \notin \Omega\}$. B^c can be read "not-B".

14. Gilbert does not elaborate further on what counts as epistemic normality.

15. Gilbert (1989, p. 193) also maintains that her account of common knowledge has the advantage of not requiring that the agents reason through an infinite hierarchy of propositions. On her account, the agents' smooth-reasoner counterparts do all the necessary reasoning for them. However, Gilbert fails to note that Aumann's and Lewis' accounts of common knowledge also have this advantage.

16. Harsanyi (1968) is the most famous defender of the CPA. Indeed, Aumann (1974, 1987) calls the CPA the *Harsanyi Doctrine* in Harsanyi's honor.

17. Alan Hajek first pointed this out to me in conversation.

18. An agent's *pure strategies* in a noncooperative game are simply the alternative acts the agent might choose as defined by the game. A mixed strategy $\sigma_k(\cdot)$ is a probability distribution defined over k's pure strategies by some random experiment such as the toss of a coin or the spin of a roulette wheel. k plays each pure strategy s_{kj} with probability $\sigma_k(s_{kj})$ according to the outcome of the experiment, which is assumed to be probabilistically independant of the others' experiments. A strategy is *completely mixed* when each pure strategy has a positive probability of being the one selected by the mixing device.

19. Lewis, (1969), p. 76. Lewis gives a further definition of agents following a convention to a *certain degree* if only a certain percentage of the agents actually conform to the coordination equilibrium corresponding to the convention. See Lewis (1969, pp. 78-89).

20. In their original papers, Bernheim (1984) and Pearce (1984) included in their definitions of rationalizability the requirement that the agents' probability distributions over their opponents satisy *probabilistic independance*, that is, for each agent *k* and for each

$$\mathbf{s}_{-k} = (s_{1_{j_1}}, \ldots, s_{k\text{-}1_{j_{k-1}}}, s_{k+1_{j_{k+1}}}, \ldots, s_{n_{j_n}}) \in S_{-k}$$

k's joint probability must equal the product of k's marginal probabilities, that is,

$$\mu_k(\mathbf{s}_{-k}) = \mu_k(s_{1_{j_1}}) \cdot \ldots \cdot \mu_k(s_{k\text{-}1_{j_{k-1}}}) \cdot \mu_k(s_{k+1_{j_{k+1}}}) \cdot \ldots \cdot \mu_k(s_{n_{j_n}})$$

Brandenburger and Dekel (1987), Skyrms (1990), and Vanderschraaf (1995) all argue that the probabilistic independence requirement is not well-motivated, and do not include this requirement in their presentations of rationalizability. Bernheim (1984) calls a Bayes concordant system of beliefs a "consistent" system of beliefs. Since the term "consistent beliefs" is used in this paper to describe probability distributions that agree with respect to a mutual opponent's strategies, I use the term "Bayes concordant system of beliefs" rather than Bernheim's "consistent system of beliefs".

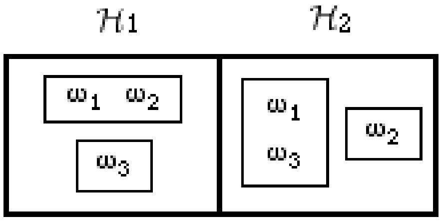
21. A mixed strategy is a propbability distribution $\sigma_k(\cdot)$ defined over *k*'s pure strategies by some random experiment such as the toss of a coin or the spin of a roulette wheel. *k* plays each pure strategy s_{kj} with probability $\sigma_k(s_{kj})$ according to the outcome of the experimentm which is assumed to be probabilistically independent of the others' experiments. A strategy is *completely mixed* when each pure strategy has a positive probability of being the one selected by the mixing device.

Nash (1950, 1951) originally developed the Nash equilibrium concept in terms of mixed strategies. In subsequent years, game theorists have realized that the Nash and more general correlated equilibrium concepts can be defined entirely in terms of agents' beliefs, without recourse to mixed strategies. See Aumann (1987), Brandenburger and Dekel (1988), and Skyrms (1991) for an extended discussion of equilibrium-in-beliefs.

22. Ron's private recommendations in effect partition Ω as follows:

- $\mathcal{H}_1 = \{ \{\omega_1,\omega_2\}, \{\omega_3\} \}$, and
- $\mathcal{H}_2 = \{ \{\omega_1,\omega_3\}, \{\omega_2\} \}$.

These partitions are diagrammed below:



Given their private information, at each possible world ω to which an agent *i* assigns positive probability, following *f* maximizes *i*'s expected utility. For instance, at $\omega = \omega_2$,

$$
\begin{aligned}
E(u_1(A_1) \mid \mathcal{H}_1)(\omega_2) &= ½ \cdot 3 + ½ \cdot 2 \\
&= 5/2 \\
&> 2 = ½ \cdot 4 + ½ \cdot 0 \\
&= E(u_1(A_2) \mid \mathcal{H}_1)(\omega_2)
\end{aligned}
$$

and

$$
\begin{aligned}
E(u_2(A_2) \mid \mathcal{H}_2)(\omega_2) &= 4 \\
&> 3 = E(u_2(A_1) \mid \mathcal{H}_1)(\omega_2)
\end{aligned}
$$

23. An outcome $\mathbf{s}_1$ of a game Pareto-dominates an outcome $\mathbf{s}_2$ if, and only if,

- $E(u_k(\mathbf{s}_1)) \geq E(u_k(\mathbf{s}_2))$ for all $k \in N$.
- $\mathbf{s}_1$ strictly dominates $\mathbf{s}_2$ if the inequalities of (i) are all srict.

24. While both the endogenous and the Aumann correlated equilibrium concepts generalize the Nash equilibrium, neither correlated equilibrium concept contains the other. See Chapter 2 of Vanderschraaf (1995) for examples which show this.

25. Aumann (1987) notes that it is possible to extend the definitions of Aumann correlated equilibrium and $\mathcal{H}_i$-measurability to allow for cases in which Ω is infinite and the $\mathcal{H}_i$'s are not necessarily partitions. However, he argues that there is nothing to be gained conceptually by doing so.

26. In general, the method of backwards induction is undefined for games of imperfect information, although backwards induction reasoning can be applied to a limited extent in such games.

27. By the elementary properties of the knowledge operator, $\mathbf{K}_2\mathbf{K}_1\mathbf{K}_2(\Gamma) \subseteq \mathbf{K}_2\mathbf{K}_1(\Gamma)$ and $\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1\mathbf{K}_2(\Gamma) \subseteq \mathbf{K}_1\mathbf{K}_2\mathbf{K}_1(\Gamma)$, so we needn't explicitly state that at I^{22}, $\mathbf{K}_2\mathbf{K}_1(\Gamma)$ and at I^{11}, $\mathbf{K}_1\mathbf{K}_2\mathbf{K}_1(\Gamma)$. By the same elementary properties, the knowledge assumptions at the latter two information sets imply that Fiona and Alan have third-order mutual knowledge of the game and second-order mutual knowledge of rationality. For instance, since $\mathbf{K}_2\mathbf{K}_1(\Gamma)$ is given at I^{22}, we have $\mathbf{K}_2\mathbf{K}_1\mathbf{K}_1(\Gamma)$ because $\mathbf{K}_1(\Gamma) \subseteq \mathbf{K}_1\mathbf{K}_1(\Gamma)$ and so $\mathbf{K}_2\mathbf{K}_1(\Gamma) \subseteq \mathbf{K}_2\mathbf{K}_1\mathrm{K}_1(\Gamma)$. The other statements which characterize third order mutual knowledge of the game and second order mutual knowledge of rationality are similarly derived.

28. The version of the example Rubinstein presents is more general than the version presented here. Rubinstein notes that this game is closely related to the *coordinated attack problem* analyzed in Halpern (1986).

29. In the terminology of decision theory, *A* is each agents' *maximin* strategy.

30. This could be achieved if the e-mail systems were constructed so that each n^{th} confirmation is sent 2^{-n} seconds after receipt of the n^{th} message.

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

First published: June 12, 2002
Content last modified: June 12, 2002

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Proposition 2.4

Proposition 2.4.
If $\omega \in \mathbf{K}^*_N(E)$ and $E \subseteq F$, then $\omega \in \mathbf{K}^*_N(F)$.

Proof.
If $E \subseteq F$, then as we observed earlier, $\mathbf{K}_i(E) \subseteq \mathbf{K}_i(F)$, so

$$\mathbf{K}^1_N(E) = \bigcap_{i \in N} \mathbf{K}_i(E) = \bigcap_{i \in N} \mathbf{K}_i(F) = \mathbf{K}^1_N(F)$$

If we now set $E' = \mathbf{K}^n_N(E)$ and $F' = \mathbf{K}^n_N(F)$, then by the argument just given we have

$$\mathbf{K}^{n+1}_N(E) = \mathbf{K}^1_N(E') \subseteq \mathbf{K}^1_N(F') = \mathbf{K}^{n+1}_N(F)$$

so we have m^{th} level mutual knowledge for every $n \geq 1$.

Hence if $\omega \in \bigcap_{n=1}^{\infty} \mathbf{K}^n_N(E)$ then $\omega \in \bigcap_{n=1}^{\infty} \mathbf{K}^n_N(F)$. □

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Proof of Proposition 2.5

Proposition 2.5.
$\omega \in \mathbf{K}^m_N(A)$ iff

(1) For all agents $i_1, i_2, \ldots, i_m \in N$, $\omega \in \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_m}(A)$

Hence, $\omega \in \mathbf{K}^*_N(A)$ iff (1) is the case for each $m \geq 1$.

Proof.
Note first that

$$(2) \quad \bigcap_{i_1 \in N} \mathbf{K}_{i_1}\Big(\bigcap_{i_2 \in N} \mathbf{K}_{i_2}\Big(\ldots\Big(\bigcap_{i_{m-1} \in N} \mathbf{K}_{i_{m-1}}\Big(\bigcap_{i_m \in N} \mathbf{K}_{i_m}(A)\Big)\Big)\Big)\Big)$$

$$= \bigcap_{i_1 \in N} \mathbf{K}_{i_1}\Big(\bigcap_{i_2 \in N} \mathbf{K}_{i_2}\Big(\ldots\Big(\bigcap_{i_{m-1} \in N} \mathbf{K}_{i_{m-1}}(\mathbf{K}^1_N(A))\Big)\Big)\Big)$$

$$= \bigcap_{i_1 \in N} \mathbf{K}_{i_1}\Big(\bigcap_{i_2 \in N} \mathbf{K}_{i_2} \ldots \Big(\bigcap_{i_{m-2} \in N} \mathbf{K}_{i_{m-2}}(\mathbf{K}^2_N(A))\Big)\Big)$$

$$= \ldots$$

$$= \bigcap_{i_1 \in N} \mathbf{K}_{i_1}(\mathbf{K}^{m-1}_N(A))$$

$$= \mathbf{K}^m_N(A)$$

By (2), $\mathbf{K}^m_N(A) \subseteq \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_m}(A)$ for $i_1, i_2, \ldots, i_m \in N$ so if $\omega \in \mathbf{K}^m_N(A)$ then condition (1) is satisfied. Condition (1) is equivalent to

$$\omega \in \bigcap_{i_1 \in N} \mathbf{K}_{i_1}\Big(\bigcap_{i_2 \in N} \mathbf{K}_{i_2}\Big(\ldots\Big(\bigcap_{i_{m-1} \in N} \mathbf{K}_{i_{m-1}}\Big(\bigcap_{i_m \in N} \mathbf{K}_{i_m}(A)\Big)\Big)\Big)\Big)$$

so by (2), if (1) is satisfied then $\omega \in \mathbf{K}^{m}_{N}(A)$. □

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Proposition 2.8

Proposition 2.8.
$\mathbf{L}^*_N(E) \subseteq \mathbf{K}^*_N(E)$, that is, Lewis-common knowledge of *E* implies common knowledge of *E*.

Proof.
Suppose that $\omega \in \mathbf{L}^*_N(E)$. By definition, there is a basis proposition A* such that $\omega \in A^*$. It suffices to show that for each $m \geq 1$ and for all agents $i_1, i_2, \ldots, i_m \in N$,

$$\omega \in \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_m}(E)$$

We prove the result by induction on m. The m = 1 case follows at once from (L1) and (L3). Now if we assume that for m = k, $\omega \in \mathbf{L}^*_N(E)$ implies $\omega \in \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_k}(E)$, then $\mathbf{L}^*_N(E) \subseteq \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_k}(E)$ because ω is an arbitrary possible world, so $\mathbf{K}_{i_1}(A^*) \subseteq \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_k}(E)$ by (L3). Since (L2) is the case and the agents of N are A*-symmetric reasoners,

$$\mathbf{K}_{i_1}(A^*) \subseteq \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_k}(E)$$

for any $i_{k+1} \in N$, so $\omega \in \mathbf{K}_{i_1}\mathbf{K}_{i_2} \ldots \mathbf{K}_{i_k}(E)$ by (L1), which completes the induction since $i_1, i_{k+1}, i_2, \ldots, i_k$ are k + 1 arbitrary agents of N. □

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Lemma 2.11

Lemma 2.11.
$\omega' \in \mathcal{M}(\omega)$ iff ω' is reachable from ω.

Proof.
Pick an arbitrary world $\omega \in \Omega$, and let

$$\mathcal{R}(\omega) = \bigcup_{n=1}^{\infty} \bigcup_{i_1, i_2, \ldots, i_n \in N} \mathcal{H}_{i_n}(\ldots(\mathcal{H}_{i_2}(\mathcal{H}_{i_1}(\omega)))$$

that is, $\mathcal{R}(\omega)$ is the set of all worlds that are reachable from ω. Clearly, for each $i \in N$, $\mathcal{H}_i(\omega) \subseteq \mathcal{R}(\omega)$, which shows that $\mathcal{R}$ is a coarsening of the partitions $\mathcal{H}_i$, $i \in N$. Hence $\mathcal{M}(\omega) \subseteq \mathcal{R}(\omega)$, as $\mathcal{M}$ is the finest common coarsening of the $\mathcal{H}_i$'s.

We need to show that $\mathcal{R}(\omega) \subseteq \mathcal{M}(\omega)$ to complete the proof. To do this, it suffices to show that for any sequence $i_1, i_2, \ldots, i_n \in N$

$$(1)\ \mathcal{H}_{i_n}(\ldots(\mathcal{H}_{i_2}(\mathcal{H}_{i_1}(\omega)))$$

We will prove (1) by induction on n. By definition, $\mathcal{H}_i(\omega) \subseteq \mathcal{M}(\omega)$ for each $i \in N$, proving (1) for $n = 1$. Suppose now that (1) obtains for $n = k$, and for a given $i \in N$, let $\omega^* \in \mathcal{H}_i(A)$ where $A = \mathcal{H}_{i_k}(\ldots(\mathcal{H}_{i_2}(\mathcal{H}_{i_1}(\omega)))$. By induction hypothesis, $A \subseteq \mathcal{M}(\omega)$. Since $\mathcal{H}_i(A)$ states that i_1 thinks that i_2 thinks that ... i_k thinks that i thinks that ω^* is possible, A and $\mathcal{H}_i(\omega^*)$ must overlap, that is, $\mathcal{H}_i(\omega^*) \cap A \neq \varnothing$. If $\omega^* \notin \mathcal{M}(\omega)$, then $\mathcal{H}_i(\omega^*) \nsubseteq \mathcal{M}(\omega)$, which implies that $\mathcal{M}$ is not a common coarsening of the $\mathcal{H}_i$'s, a contradiction. Hence $\omega^* \in \mathcal{M}(\omega)$, and since i was chosen arbitrarily from N, this shows that (1) obtains for $n = k + 1$. ☐

peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Lemma 2.12

Lemma 2.12.
$\mathcal{M}(\omega)$ is common knowledge for the agents of N at ω.

Proof.
Since $\mathcal{M}$ is a coarsening of $\mathcal{H}_i$ for each $i \in N$, $\mathbf{K}_i(\mathcal{M}(\omega))$. Hence, $\mathbf{K}^1_N(\mathcal{M}(\omega))$, and since by definition $\mathbf{K}_i(\mathcal{M}(\omega)) = \{\omega \mid \mathcal{H}_i(\omega) \subseteq \mathcal{M}(\omega)\} = \mathcal{M}(\omega)$,

$$\mathbf{K}^1_N(\mathcal{M}(\omega)) = \bigcap_{i \in N} \mathbf{K}_i(\mathcal{M}(\omega)) = \mathcal{M}(\omega)$$

Applying the recursive definition of mutual knowledge, for any $m \geq 1$,

$$\mathbf{K}^m_N(\mathcal{M}(\omega)) = \bigcap_{i \in N} \mathbf{K}_i(\mathbf{K}^{m-1}_N(\mathcal{M}(\omega))) = \bigcap_{i \in N} \mathbf{K}_i(\mathcal{M}(\omega)) = \mathcal{M}(\omega)$$

so, since $\omega \in \mathcal{M}(\omega)$, by definition we have $\omega \in \mathbf{K}^*_N(\mathcal{M}(\omega))$. □

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Proposition 2.13

Proposition 2.13 (Aumann 1976)
Let $\mathcal{M}$ be the meet of the agents' partitions $\mathcal{H}_i$ for each $i \in N$. A proposition $E \subseteq \Omega$ is common knowledge for the agents of N at ω iff $\mathcal{M}(\omega) \subseteq E$. In Aumann (1976), E is *defined* to be common knowledge at ω iff $\mathcal{M}(\omega) \subseteq E$.

Proof.
($\Leftarrow$) By Lemma 2.12, $\mathcal{M}(\omega)$ is common knowledge at ω, so E is common knowledge at ω by Proposition 2.4.

($\Rightarrow$) We must show that $\mathbf{K}^*_N(E)$ implies that $\mathcal{M}(\omega) \subseteq E$. Suppose that there exists $\omega' \in \mathcal{M}(\omega)$ such that $\omega' \notin E$. Since $\omega' \in \mathcal{M}(\omega)$, ω' is reachable from ω, so there exists a sequence 0, 1, . . . , m - 1 with associated states $\omega_1, \omega_2, \ldots, \omega_m$ and information sets $\mathcal{H}_{i_k}(\omega_k)$ such that $\omega_0 = \omega$, $\omega_m = \omega'$, and $\omega_k \in \mathcal{H}_{i_k}(\omega_{k+1})$. But at information set $\mathcal{H}_{i_k}(\omega_m)$, agent i_k does not know event E. Working backwards on k, we see that event E cannot be common knowledge, that is, agent i_1 cannot rule out the possibility that agent i_2 thinks that . . . that agent i_{m-1} thinks that agent i_m does not know E. □

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Proposition 2.14

Proposition 2.14.
Let C^*_N be the greatest fixed point of f_E. Then $C^*_N(E) = K^*_N(E)$.

Proof.

We have shown that $\mathbf{K}^*_N(E)$ is a fixed point of f_E, so we only need to show that $\mathbf{K}^*_N(E)$ is the greatest fixed point. Let B be a fixed point of f_B. We want to show that $B \subseteq \mathbf{K}^k_N(E)$ for each value $k \geq 1$. We will proceed by induction on k. By hypothesis,

$$B = f_E(B) = \mathbf{K}^1_N(E \cap B) \subseteq \mathbf{K}^1_N(E)$$

by monotonicity, so we have the k=1 case. Now suppose that for k=m, $B \subseteq \mathbf{K}^m_N(E)$. Then by monotonicity,

(i) $\mathbf{K}^1_N(B) \subseteq \mathbf{K}^1_N\mathbf{K}^{\mathrm{m}}_N(E) = \mathbf{K}^{\mathrm{m+1}}_N(E)$

We alo have:

(ii) $B = \mathbf{K}^1_N(E \cap B) \subseteq \mathbf{K}^1_N(B)$

by monotonicity, so combining (i) and (ii) we have:

$B \subseteq \mathbf{K}^1_N(B) \subseteq \mathbf{K}^{\mathrm{m+1}}_N(E)$

completing the induction. □

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: June 12, 2002
Content last modified: June 12, 2002

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Proposition 3.1

Proposition 3.1.
Let Ω be a finite set of states of the world. Suppose that

> (i) Agents i and j have a common prior probability distribution $\mu(\cdot)$ over the events of Ω such that $\mu(\omega) > 0$ for each $\omega \in \Omega$, and
>
> (ii) It is common knowledge at ω that i's posterior probability of event E is $q_i(E)$ and that j's posterior probability of E is $q_j(E)$.

Then $q_i(E) = q_j(E)$.

Proof.
Let $\mathcal{M}$ be the meet of all the agents' partitions, and let $\mathcal{M}(\omega)$ be the element of $\mathcal{M}$ containing ω. Since $\mathcal{M}(\omega)$ consists of cells common to every agents information partition, we can write

$$\mathcal{M}(\omega) = \bigcup_k H_{ik},$$

where each $H_{ik} \in \mathcal{H}_i$. Since i's posterior probability of event E is common knowledge, it is constant on $\mathcal{M}(\omega)$, and so

$$q_i(E) = \mu(E \mid H_{ik}) \text{ for all } k$$

Hence,

$$\mu(E \cap H_{ik}) = q_i(E)\,\mu(H_{ik})$$

and so

$$\mu(E \cap \mathcal{M}(\omega)) = \mu(E \cap \bigcup_k H_{ik}) = \mu(\bigcup_k E \cap H_{ik})$$

$$= \sum_k \mu(E \cap H_{ik}) = \sum_k q_i(E)\,\mu(H_{ik})$$

$$= q_i(E) \sum_k \mu(H_{ik}) = q_i(E)\,\mu(\bigcup_k H_{ik})$$

$$= q_i(E)\,\mu(\mathcal{M}(\omega))$$

Applying the same argument to j, we have

$$\mu(E \cap \mathcal{M}(\omega)) = q_j(E)\,\mu(\mathcal{M}(\omega))$$

so we must have $q_i(E) = q_j(E)$. □

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Proof of Proposition 3.4

Proposition 3.4.
In a game Γ, common knowledge of Bayesian rationality is satisfied if, and only if, (3.i) is common knowledge.

Proof.
Suppose first that common knowledge of Bayesian rationality is satisfied. Since it is common knowledge that agent i knows that agent k is Bayesian rational, it is also common knowledge that if $\mu_i(s_{k\,j}) > 0$, then $s_{k\,j}$ must be optimal for k given some belief over S_{-k} , so (3.i) is common knowledge.

Suppose now that (3.i) is common knowledge. Then, by (3.i), agent i knows that agent k is Bayesian rational. Since (3.i) is common knowledge, all statements of the form `For i, j, . . . , k ∈ N, i knows that j knows that . . . is Bayesian rational' follow by induction. □

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Proposition 3.7

Proposition 3.7
Assume that the probabilities

$$\boldsymbol{\mu} = (\mu_1, \ldots , \mu_n) \in \Delta_1 (S_{-1}) \times \ldots \times \Delta_n (S_{-n})$$

are common knowledge. Then common knowledge of Bayesian rationality is satisfied if, and only if, $\boldsymbol{\mu}$ is an endogenous correlated equilibrium.

Proof.
Suppose first that common knowledge of Bayesian rationality is satisfied. Then, by Proposition 3.4, for a given agent $k \in N$, if $\mu_i(s_{kj}) > 0$ for each agent $i \neq k$, then s_{kj} must be optimal for k given some distribution $\sigma_k \in \Delta_k (S_{-k})$. Since the agents' distributions are common knowledge, this distribution is precisely μ_k, so (3.iii) is satisfied for k. (3.iii) is similarly established for each other agent $i \neq k$, so $\boldsymbol{\mu}$ is an endogenous correlated equilibrium.

Now suppose that $\boldsymbol{\mu}$ is an endogenous correlated equilibrium. Then, since the distributions are common knowledge, (3.i) is common knowledge, so common knowledge of Bayesian rationality is satisfied by Proposition 3.4.

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Proposition 3.11

Proposition 3.11 (Aumann 1987)
If each agent $i \in N$ is ω-Bayes rational at each possible world $\omega \in \Omega$, then the agents are following an Aumann correlated equilibrium. If the CPA is satisfied, then the correlated equilibrium is objective.

Proof.
We must show that $s : \Omega \rightarrow S$ as defined by the $\mathcal{H}_i$-measurable s_i's of the Bayesian rational agents is an objective Aumann correlated equilibrium. Let $i \in n$ and $\omega \in \Omega$ be given, and let $g_i : \Omega \rightarrow S_i$ be any function that is a function of s_i. Since s_i is constant over each cell of $\mathcal{H}_i$, g_i must be as well, that is, g_i is $\mathcal{H}_i$-measurable. By Bayesian rationality,

$$E(u_i \circ s \mid \mathcal{H}_i)(\omega) \geq E(u_i(g_i, s_{-i}) \mid \mathcal{H}_i)(\omega)$$

Since ω was chosen arbitrarily, we can take iterated expectations to get

$$E(E(u_i \circ s \mid \mathcal{H}_i)(\omega)) \geq E(E(u_i(g_i, s_{-i}) \mid \mathcal{H}_i)(\omega))$$

which implies that

$$E(u_i \circ s) \geq E(u_i(g_i, s_{-i}))$$

so s is an Aumann correlated equilibrium.

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Proof of Proposition 3.12

Proposition 3.12 (Bicchieri 1993)
In an extensive form game of perfect information, the agents follow the backwards induction solution if the following conditions are satisfied for each agent i at each information set I^{ik}:

- i is rational, i knows this and i knows the game, and
- At any information set I^{jk+1} that immediately follows I^{ik}, i knows at I^{ik} what j knows at I^{jk+1}.

Proof.
The proof is by induction on m, the number of potential moves in the game. If m = 1, then at I^{i1}, by (a) agent i chooses a strategy which yields i her maximum payoff, and this is the backwards induction solution for a game with one move.

Now suppose the proposition holds for games having at most m = r potential moves. Let Γ be a game of perfect information with r + 1 potential moves, and suppose that (α) and (β) are satisfied at every node of Γ. Let I^{i1} be the information set corresponding to the root of the tree for Γ. At I^{i1}, i knows that (a) and (b) obtain for each of the subgames that start at the information sets which immediately follow I^{i1}. Then i knows that the outcome of play for each of these subgames is the backwards induction solution for that subgame. Hence, at I^{i1} i's payoff maximizing strategy is a branch of the tree starting from I^{i1} which leads to a subgame whose backwards induction solution is best for i, and since i is rational, i chooses such a branch at I^{i1}. But this is the backwards induction solution for the entire game Γ, so the proposition is proved for m = r + 1.

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Common Knowledge

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Supplement to Common Knowledge

Rubinstein's Proof

[Note: See Definition 3.2 for the notation used in this proof.]

Let T_2 denote the number of messages that Joanna's e-mail system sends, and T_1 denote the number of messages that Lizzi's e-mail system sends. We might suppose that T_i appears on each agent's computer screen. If $T_1 = 0$, then Lizzi sends no message, that is, ω_1 has occurred, in which case Lizzi's unique best response is to choose *A*. If $T_2 = 0$, then Joanna did not receive a message. She knows that in this case, either ω_1 has occurred and Lizzi did not send her a message, which occurs with probability .51, or ω_2 has occurred and Lizzi sent her a message which did not arrive, which occurs with probability $.49\varepsilon$. If ω_1 has occurred, then Lizzi is sure to choose *A*, so Joanna knows that whatever Lizzi might do at ω_2,

$$\begin{aligned} E(u_2(A) \mid T_2{=}0) &\geq \frac{2(.51) + 0(.49)\varepsilon}{.51 + .49\varepsilon} \\ &> \frac{-4(.51) + 2(.49)\varepsilon}{.51 + .49\varepsilon} \\ &\geq E(u_2(B) \mid T_2{=}0) \end{aligned}$$

so Joanna is strictly better off choosing A no matter what Lizzi does at either state of the world.

Suppose next that for all $T_i < t$, each agents' unique best response given her expectations regarding the other agent is *A*, so that the unique Nash equilibrium of the game is (*A*,*A*). Assume that $T_1 = t$. Lizzi is uncertain whether $T_2 = t$, which is the case if Joanna received Lizzi's t^{th} automatic confirmation and Joanna's t^{th} confirmation was lost, or if $T_2 = t - 1$, which is the case if Lizzi's t^{th} confirmation was lost. Then

$$\begin{aligned} \mu_1(T_2 = t-1 \mid T_1 = t) &= z \\ &= \frac{\varepsilon}{\varepsilon + (1-\varepsilon)\varepsilon} \\ &> \tfrac{1}{2}. \end{aligned}$$

[1]

Thus it is more likely that Lizzi's last confirmation did not arrive than that Joanna did receive this message. By the inductive assumption, Lizzi assesses that Joanna will choose A if $T_2 = t-1$. So

$$\begin{aligned} E(u_1(B) \mid T_1 = t) &\leq -4z + 2(1-z) \\ &= -6z + 2 \\ &< -3 + 2 \\ &= -1, \end{aligned}$$

and

$$E(u_1(A) \mid T_1 = t) = 0$$

since Lizzi knows that ω_2 is the case. So Lizzi's unique best action is A. Similarly, one can show that A is Joanna's best reply if $T_2 = t$. So by induction, (A,A) is the unique Nash equilibrium of the game for every $t \geq 0$.

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

Return to Coordination and Common *p*-Belief

First published: August 27, 2001
Content last modified: August 27, 2001

Stanford Encyclopedia of Philosophy
Notes to Supplement: Rubinstein's Proof

Notes

1. If this does not look immediately obvious, consider that either

> $E = [T_2 = t]$ = my (Lizzi's) t^{th} confirmation was lost,

or

> <em>F = $[T_2 = t]$ = my t^{th} confirmation was received and Joanna's t^{th} confirmation was lost

must occur, and that $\mu_1(T_1 = t \mid E) = \mu_1(T_1 = t \mid F) = 1$ because Lizzi can see her own computer screen, so we can apply Bayes' Theorem as follows:

$$\mu_1(E \mid T_1 = t) = \frac{\mu_1(T_1 = t \mid E)\ \ \mu_1(E)}{\mu_1(T_1 = t \mid E)\ \ \mu_1(E) + \mu_1(T_1 = t \mid F)\ \ \mu_1(F)}$$

$$= \frac{\mu_1(E)}{\mu_1(E) + \mu_1(F)}$$

$$= \frac{\varepsilon}{\varepsilon + (1 - \varepsilon)\varepsilon}$$

Peter Vanderschraaf
peterv@cyrus.andrew.cmu.edu

First published: August 27, 2001
Content last modified: August 27, 2001

Harriet Taylor Mill

> *Were I but capable of interpreting to the world one half the great thoughts and noble feelings which are buried in her grave, I should be the medium of a greater benefit to it, than is ever likely to arise from anything that I can write, unprompted and unassisted by her all but unrivalled wisdom.*
>
> - John Stuart Mill

Harriet Taylor Mill (1807-1858) wrote primarily in the area of social-political philosophy, and had a particular interest in women's rights, but—her essay, "The Enfranchisement of Women" notwithstanding—the body of work that she penned is probably not substantial enough for her to be judged a major figure in the history of philosophy on the basis of it alone. However, John Stuart Mill, her second husband, not only lavished praise on her intellect, emotional depth, and moral character, but also credited her with exerting a tremendous influence on his thought, with making major intellectual contributions to many of the works published in his name, and even with having been intimately involved in the composition of some of his most important works. Today scholars debate how much of a difference she really made to 'his' corpus, whether whatever effect she had on it was an improving one, and whether she even came close to meriting the lavish praise that he heaped on her in passages such as the one quoted above from the dedication to *On Liberty* (J. S. Mill 1977, p. 216).

---

1. Life and Character

The woman who is today most commonly known as Harriet Taylor Mill (hereafter HTM) was born Harriet Hardy in London in October 1807. She married the pharmaceuticals wholesaler John Taylor on 14 March 1826; she was eighteen, and he was twenty nine. The couple had three children together: Herbert, born in 1827; Algernon ("Haji"), born in 1830; and Helen ("Lily"), born in 1831. John Taylor died of cancer in 1849, and in the Spring of 1851 Harriet Taylor married again, this time to John Stuart Mill (hereafter JSM). A tuberculosis sufferer, as was JSM (it is possible that she caught the disease, then called "consumption," from him), she died of a respiratory failure on 3 November 1858.

HTM and JSM met for the first time in 1830. Their meeting was arranged by the leader of HTM's Unitarian congregation, the Reverend W. J. Fox, to whom she had complained about John Taylor's lack of interest in philosophy and the arts. Of course there is no way to know if Fox anticipated that passionate feelings would spring up between JSM and HTM, although Josephine Kamm speculates that Fox, who was already married, might have sought to eliminate JSM as a competitor for the attentions of his soon-to-be mistress Eliza Flower (1977, 29f.). Whether this was Fox's intention or not, the two young people did very quickly fall in love.

Their conduct during the long period in which HTM was Mrs. John Taylor was quite scandalous by Victorian standards. Early on JSM would frequently, indeed almost nightly, visit the Taylors' home; John Taylor would usually facilitate these visits by going to his club. On the whole, John Taylor was remarkably tolerant of the fact that his wife, to whom he was utterly devoted, was on the most intimate of terms with another man, but his tolerance did have some limits. In 1833, at his insistence, HTM established a separate residence, and she lived apart from John Taylor for most of the rest of his life, seeing Mill at her convenience (Helen lived with her, Herbert and Haji with their father). In 1848 Taylor refused to allow JSM to dedicate *The Principles of Political Economy* to HTM, although the dedication was inserted into copies of the book that they distributed to friends. In 1849 John Taylor began to suffer from the cancer that would eventually take his life, and he asked HTM to return home to care for him. She declined, on the grounds that her first duty was to JSM, who at the time was suffering himself from an injured hip and temporary near-blindness. While JSM eventually mended John Taylor's condition only worsened, and at the end HTM did dedicate herself to caring for her husband. In fact, she rebuked JSM very sharply for having failed, while paying her a short visit, to ask about John Taylor's condition. She upbraided him even more severely for having suggested that she might write to him during an "odd time" when she might find a "change of subject of thought a relief": "Good God, sh[oul]d you think it a relief to think of something else some acquaintance or what not while *I* was dying?" (H. T. Mill 1998, 360).

After John Taylor's death in 1849, JSM and HTM waited nearly two years before marrying. They had already largely withdrawn from society, due perhaps to the gossip that their relationship generated. After their marriage they spent much of their time in their Blackheath Park home with just Haji and Helen Taylor for company (although JSM did still go to work each day at the India House). JSM was so excessively sensitive to any perceived slight of his wife that he never forgave his mother and sisters for their failure to call on her immediately after being told of his impending nuptials. (When one bears in mind that he had not allowed his family to make any mention of HTM to that point one can understand why, out of their eagerness to please him, they had waited for instructions before doing anything.) The Mills did sometimes interrupt their reclusion to travel, separately or together, to the south of England or to

the Continent in pursuit of a more healthful climate. In the Fall of 1858 they set out for Montpellier, but HTM's fragile health gave out in Avignon. JSM bought a small house there, next to the cemetery in which she was buried, where he spent a considerable portion of the remainder of his life. He died and was buried in Avignon in May 1873.

There have always been questions about the sexual dimension of the Mills' relationship, or rather about whether it *had* such a dimension. The most interesting question to many of their contemporaries was no doubt that of whether they were sexually involved prior to the death of John Taylor. They adamantly denied this. In his *Autobiography* JSM wrote that:

> …our relation to each other at that time was one of strong affection and confidential intimacy only. For though we did not consider the ordinances of society binding on a subject so entirely personal, we did feel bound that our conduct should be such as in no degree to bring discredit on her husband, nor therefore on herself. (J. S. Mill 1981, 237.)

Their private correspondence seems to bear this out (see for example H. T. Mill 1998, 375).

Today there is probably more interest in whether the Mills had a sexual relationship *after* their marriage, a question that—however sordid and trivial it may appear—could bear on the serious matter of their views about sexuality. Some of their comments could be taken to suggest that the sex act is inherently degrading and that sexual gratification has no value (see for example J. S. Mill 1984, 285—Mill is writing specifically about rape in this passage, but he seems to imply that in every instance the "animal function" of coitus is abjective—and H. T. Mill 1998, 18, 226). Of course, they could have had non-philosophical reasons for failing to consummate their marriage. Sheer lack of interest is one obvious explanation; additionally, commentators have speculated both that JSM might have been impotent (Kamm 1977, 41) and that HTM might have contracted syphilis from her first husband (Jacobs 1998, xxx-xxxii). This is, in any case, a question to which we are unlikely to find a definitive answer.

HTM's contemporaries offer radically different impressions of her as a person. JSM's view is already fairly clear from the lines from the dedication to *On Liberty* that were quoted above, but in a venue free from the normal restrictions on length his description of her in his *Autobiography* really is worth quoting in full:

> Although it was years after my introduction to Mrs. Taylor before my acquaintance with her became at all intimate or confidential, I very soon felt her to be the most admirable person I had ever known. It is not to be supposed that she was, or that any one, at the age at which I first saw her, could be, all that she afterwards became. Least of all could this be true of her, with whom self-improvement, progress in the highest and in all senses, was a law of her nature; a necessity equally from the ardour with which she sought it, and from the spontaneous tendency of faculties which could not receive an impression or an experience without making it the source or the occasion of an accession of wisdom. Up to the time when I first saw her, her rich and powerful nature had chiefly unfolded itself according to the received type of feminine genius. To her outer circle she was a beauty and

a wit, with an air of natural distinction, felt by all who approached her: to the inner, a woman of deep and strong feeling, of penetrating and intuitive intelligence, and of an eminently meditative and poetic nature. Married at a very early age, to a most upright, brave, and honourable man, of liberal opinions and good education, but without the intellectual or artistic tastes which would have made him a companion for her, though a steady and affectionate friend, for whom she had true esteem and the strongest affection through life, and whom she most deeply lamented when dead; shut out by the social disabilities of women from any adequate exercise of her highest faculties in action on the world without; her life was one of inward meditation, varied by familiar intercourse with a small circle of friends, of whom one only (long since deceased) was a person of genius,[1] or of capacities of feeling or intellect kindred with her own, but all had more or less of alliance with her in sentiments and opinions. Into this circle I had the good fortune to be admitted, and I soon perceived that she possessed in combination, the qualities which in all other persons whom I had known I had been only too happy to find singly. In her, complete emancipation from every kind of superstition (including that which attributes a pretended perfection to the order of nature and the universe), and an earnest protest against many things which are still part of the established constitution of society, resulted not from the hard intellect, but from strength of noble and elevated feeling, and co-existed with a highly reverential nature. In general spiritual characteristics, as well as in temperament and organisation, I have often compared her, as she was at this time, to Shelley: but in thought and intellect, Shelley, so far as his powers were developed in his short life, was but a child compared with what she ultimately became. Alike in the highest regions of speculation and in the smaller practical concerns of daily life, her mind was the same perfect instrument, piercing to the very heart and marrow of the matter; always seizing the essential idea or principle. The same exactness and rapidity of operation, pervading as it did her sensitive as well as her mental faculties, would, with her gifts of feeling and imagination, have fitted her to be a consummate artist, as her fiery and tender soul and her vigorous eloquence would certainly have made her a great orator, and her profound knowledge of human nature and discernment and sagacity in practical life, would, in the times when such a *carriére* was open to women, have made her eminent among the rulers of mankind. Her intellectual gifts did but minister to a moral character at once the noblest and the best balanced which I have ever met with in life. Her unselfishness was not that of a taught system of duties, but of a heart which thoroughly identified itself with the feelings of others, and often went to excess in consideration for them by imaginatively investing their feelings with the intensity of its own. The passion of justice might have been thought to be her strongest feeling, but for her boundless generosity, and a lovingness ever ready to pour itself forth upon any or all human beings who were capable of giving the smallest feeling in return. The rest of her moral characteristics were such as naturally accompany these qualities of mind and heart: the most genuine modesty combined with the loftiest pride; a simplicity and sincerity which were absolute, towards all who were fit to receive them; the utmost scorn of whatever was mean and cowardly, and a burning indignation at everything brutal or tyrannical, faithless or dishonourable in conduct and character, while making the broadest distinction between *mala in se* and mere *mala prohibita*—between acts giving evidence of intrinsic badness in

> feeling and character, and those which are only violations of conventions either good or bad, violations which whether in themselves right or wrong, are capable of being committed by persons in every other respect lovable or admirable (J. S. Mill 1981, 193-7).

No one else who knew HTM personally spoke of her in anything like these terms, as far as we know, and indeed several of her contemporaries held her in low estimation. The Carlyles were admirers initially, but soon had changes of heart. Jane said that HTM was "a peculiarly affected body" who "was not easy unless she startled you with unexpected sayings," and was in fact "somewhat of a humbug" (quoted in Packe 1954, 325f). Thomas said that "She was full of unwise intellect, asking and re-asking stupid questions" (quoted in Packe 1954, 315). Harold Laski related that "Morley told me that Louis Blanc told him that he once sat for an hour with her and that she repeated to him what afterwards turned out to be an article that Mill had just finished for the *Edinburgh*. … If she was what he thought, someone at least should have given us indications" (quoted in Stillinger 1961, 24f).

There is, of course, a vast middle territory between these extremes, and it seems highly plausible that the truth about HTM lies somewhere within it. We certainly do not want to make the mistake of uncritically accepting the reports of the notoriously catty Carlyles. But at the same time we have clear evidence that JSM was unable to be absolutely objective where HTM was concerned. His description of her character, for example, leaves absolutely no room for vanity, yet HTM clearly was vain, if not, as Diana Trilling adds, also "prideful" and "mean-spirited" (1952, 120). Although it is important not to make this fault seem any more significant than it would in reference to a male philosopher, only a vain person would have allowed her husband to write a passage about her like the one that appeared in JSM's *Autobiography* (this passage was drafted in HTM's lifetime and she did help to edit the manuscript—on this see JSM 1981, xix).[2] Of course, because JSM's praise for HTM is so hyperbolic, to say that she was a little less gifted or marvelous than he took her to be is not necessarily to denigrate her; it is perfectly consistent with her being a singular person in every respect. She was not often described in judicious or balanced terms, but JSM's brother George, who knew her reasonably well, did relate to JSM's disciple and biographer Alexander Bain that "Mrs. Taylor was a clever and remarkable woman, but nothing like what John took her to be" (Bain 1882, 166).

2. Philosophical Contributions

2.1 Harriet Taylor Mill's Writings

Determining what philosophical contributions HTM made and assessing their value is extremely challenging. The problem is that HTM's close relationship with JSM makes it difficult to say how much responsibility each bears for the work that they collectively produced. It is not always clear how much each of them actually contributed to the writing of some works, and even where this does seem clear it is often unclear which of the ideas each of them contributed. We do know that HTM composed relatively few works herself (relative, at least, to JSM), and not all of her writings were really philosophical in nature; her longest piece, for example, was an essay on William Caxton and the history of printing, which was part of a collection published in 1833 by the Society for the Diffusion of Useful Knowledge (H. T.

Mill 1998, 238-91). Some poems, book reviews, and an essay on the aesthetic appreciation of the seasons were published in the *Monthly Repository* in the early 1830s, when Fox was its editor. She is usually credited with having written “The Enfranchisement of Women,” published in *The Westminster Review* in 1851 (the only reason for any doubt is that JSM did on one occasion describe it to the *Westminster*’s editor as his own, although he later attributed it to her). In it HTM maintains that to remove restrictions on women’s political participation and choice of occupations would not only promote their interests and improve their characters but would do the same for men; this essay contains many of the same lines of argument as *The Subjection of Women*, written by JSM and published in 1869 (although some interpreters take it to express a slightly more radical view of gender roles than the later essay; see Rossi 1970, 41-3). The recently published volume *The Complete Works of Harriet Taylor Mill* also includes some unpublished essays and fragments of essays written by HTM; the topics include women’s rights, the obligation to obey laws that one thinks are unjust, the role of proverbs in moral education, and toleration. These pieces are generally very short, however, and many are just fragments.

2.2 John Stuart Mill’s Assessment of Harriet Taylor Mill

HTM’s collaboration with JSM, however, makes it impossible to reach any judgment about her importance in the history of philosophy simply on the basis of what she herself wrote. As he observed:

> When two persons have their thoughts and speculations completely in common; when all subjects of intellectual or moral interest are discussed between them in daily life, and probed to much greater depths than are usually or conveniently sounded in writings intended for general readers; when they set out from the same principles, and arrive at their conclusions by processes pursued jointly, it is of little consequence in respect to the question of originality, which of them holds the pen; the one who contributes least to the composition may contribute most to the thought; the writings which result are the joint product of both, and it must often be impossible to disentangle their respective parts, and affirm that this belongs to one and that to the other (J. S. Mill 1981, 251).

So while JSM implicitly acknowledged that his hand held the pen most often, when he and HTM collaborated, he also suggested that she contributed ideas to works that he was solely or primarily responsible for composing. “In this wide sense,” he continues, “not only during the years of our married life, but during many of the years of confidential friendship which preceded it, all my published writings were as much my wife's work as mine; her share in them constantly increasing as years advanced” (1981, 251). (Of course, this point cuts both ways. It also entails that we cannot uncritically assume that HTM was the source of all of the ideas in anything *she* penned, including “The Enfranchisement of Women.”)

JSM related that most of the ideas that HTM contributed to their joint work were at either the highest or the lowest levels of abstraction, where she particularly excelled as a thinker:

> With those who, like all the best and wisest of mankind, are dissatisfied with human life as it is, and whose feelings are wholly identified with its radical amendment, there are two

> main regions of thought. One is the region of ultimate aims; the constituent elements of the highest realizable ideal of human life. The other is that of the immediately useful and practically attainable.[3] In both these departments, I have acquired more from her teaching, than from all other sources taken together (1981, 197).

While HTM's strengths were in the most abstract and practical realms of thought, JSM took his own strength to lie in "the uncertain and slippery intermediate region, that of theory, or moral and political science," including "political economy, analytic psychology, logic, philosophy of history," etc. He acknowledged that HTM had very little to do with his first major work, *A System of Logic* (first published in 1843), or with his discussions of the more technical aspects of political economy. But much of his philosophical output, he suggested, conveyed her thoughts:

> During the greater part of my literary life I have performed the office in relation to her, which from a rather early period I had considered as the most useful part that I was qualified to take in the domain of thought, that of an interpreter of original thinkers, and mediator between them and the public; for I had always a humble opinion of my own powers as an original thinker, except in abstract science … but thought myself much superior to most of my contemporaries in willingness and ability to learn from everybody…. I had, in consequence, marked out this as a sphere of usefulness in which I was under a special obligation to make myself active: the more so, as the acquaintance I had formed with the ideas of the Coleridgians, of the German thinkers, and of Carlyle, all of them fiercely opposed to the mode of thought in which I had been brought up, had convinced me that along with much error they possessed much truth…. Thus prepared, it will easily be believed that when I came into close intellectual communion with a person of the most eminent faculties, whose genius, as it grew and unfolded itself in thought, continually struck out truths far in advance of me, but in which I could not, as I had done in those others, detect any mixture of error, the greatest part of my mental growth consisted in the assimilation of those truths, and the most valuable part of my intellectual work was in building the bridges and clearing the paths which connected them with my general system of thought (1981, pp. 251ff).

Moreover, there are two major works published in JSM's name of which he said that HTM was a co-author in the narrower sense of having been heavily involved in their actual composition. One is the *Principles of Political Economy*. The full title of this book is *The Principles of Political Economy, With Some of Their Applications to Social Philosophy*, and it was the portion of the book dealing with social philosophy that HTM helped to shape. This is especially true with respect to one chapter in particular, the one titled "On the Probable Futurity of the Working Classes." This chapter argues that when they have made enough moral and intellectual progress the working classes can be expected to refuse to settle for mere wages; they will instead insist on arrangements in which they cooperatively own the firms for which they labor, and will at least experiment with Socialist and Communist communities of the sorts depicted by Saint-Simon, Fourier, Blanc, and Owen. JSM wrote that:

> In the first draft of the book, that chapter did not exist. She pointed out the need of such a

> chapter, and the extreme imperfection of the book without it: she was the cause of my writing it; and the more general part of the chapter, the statement and discussion of the two opposite theories respecting the proper condition of the labouring classes, was wholly an exposition of her thoughts, often in words taken from her own lips. The purely scientific part of the Political Economy I did not learn from her; but it was chiefly her influence that gave to the book that general tone by which it is distinguished from all previous expositions of Political Economy that had any pretension to being scientific, and which has made it so useful in conciliating minds which those previous expositions had repelled…. The economic generalizations which depend, not on necessities of nature but on those combined with the existing arrangements of society, it deals with only as provisional, and as liable to be much altered by the progress of social improvement. I had indeed partially learnt this view of things from the thoughts awakened in me by the speculations of the St. Simonians; but it was made a living principle pervading and animating the book by my wife's promptings (1981, 255ff).

HTM also played an active role in the process of revising later editions of the *Principles*. For example, the second (1849) edition was considerably more favorable to Socialism and even Communism, and this largely seems to reflect a change in HTM’s thinking that JSM came to share as a result of her persuasion. In a letter written to HTM early in 1849, JSM pointed out that she now “had marked dissent” from a passage in the first edition raising an objection to Communism that “was inserted on your proposition & very nearly in your own words.” He continued:

> This is probably only the progress we have always been making, & by thinking sufficiently I should probably come to think the same—as is almost always the case, I believe always when we think long enough (Hayek 1951, pp. 134f).

The other major work of which JSM says that HTM ought to be considered a co-author in the narrow sense is the essay *On Liberty*, which was published in the year after her death. The dedication of this essay, a portion of which has already been quoted, says that “Like all that I have written for many years, it belongs as much to her as to me,” and Mill later elaborated on HTM’s role in the essay’s production:

> The “Liberty” was more directly and literally our joint production than anything else which bears my name, for there was not a sentence of it that was not several times gone through by us together, turned over in many ways, and carefully weeded of any faults, either in thought or expression, that we detected in it…. With regard to the thoughts, it is difficult to identify any particular part or element as being more hers than all the rest. The whole mode of thinking of which the book was the expression, was emphatically hers…. The “Liberty” is likely to survive longer than anything else that I have written (with the possible exception of the “Logic”), because the conjunction of her mind with mine has rendered it a kind of philosophic text-book of a single truth… (1981, 257ff).[4]

2.3 The Countervailing Assessment

But just as there is doubt about whether HTM was really the person that JSM made her out to be, so too is there doubt about how valuable she really was to him as an intellectual partner source of ideas. H. O. Pappe, for example, concludes his monograph *John Stuart Mill and the Harriet Taylor Myth* by questioning whether HTM introduced any substantial alterations into the pattern of JSM's thought:

> [Harriet's] early writings evince her dependence on Mill. For the later period of their partnership we have no valid evidence to show that Harriet turned Mill's mind toward new horizons or gave an unexpected significance to his thought…. Mill without Harriet would still have been Mill. Mill married to George Eliot (or to Mary Wollstonecraft—permitting the anachronism) might have been transformed. Mary Ann Evans might have given him something new by way of independent thought and deeper feeling. Yet, considering her equality of stature, there would have been no need for him in masochistic guilt to magnify her contribution (1960, 47f).

Similarly, Francis Mineka says that "Neither he [JSM] nor his recent biographers have convinced us that she was the originating mind behind his work …." (1963, 306). Other commentators are willing to concede that HTM did exert a considerable influence over the direction of JSM's thought, but allege that her impact was decidedly for the worse, i.e., that she persuaded him to adopt positions that are rather obviously weaker than contrary ones that would otherwise have held—or, indeed, worse than positions he did hold during the periods at the beginning and/or end of his intellectual career when she was not in a position of 'ascendancy' over him. Gertrude Himmelfarb makes this claim explicitly in her characterization of HTM's influence on JSM's views on liberty (1974), and it may be implicit in the celebrated free market economist Friedrich Hayek's account of her influence on his views on Socialism (1951).

Those who argue that HTM was something less than what JSM took her to be, and that he was even mistaken about her role in their intellectual partnership, must explain how he was so misled. There is a tradition of stating that, in essence, he fell victim to her "feminine wiles." It was commonly held among their contemporaries that, as Bain observes, "she imbibed all his views, and gave them back in her own form, by which he was flattered and pleased" (1882, 173). Ruth Borchard says that "Accustomed by training and experience to the acceptance of ascetic, masculine values, he was completely overpowered by her intensely feminine atmosphere" (1957, 46). And Laski speculates: "I should guess that she was a comfortable and sympathetic person and that Mill, brought up to fight Austin, Praed, Macaulay and Grote, had never met a really soft cushion before." (*op. cit.*). Some writers have even advanced the idea that after the death of his domineering father James Mill, JSM felt a need to invent another parental authority in order that he might submit to it (e.g., Trilling 1952, 118; Mazlish 1975, 286ff). Recently, Jo Ellen Jacobs has been outspoken in denying both that HTM lacked influence over Mill and that this influence had any other basis than his appreciation of her superb intellect. She writes that JSM "was a big boy and could evaluate her reasoning" (1998, xv), and claims that the failure of HTM's critics to recognize that she possessed a first-rate mind and that she engaged in genuine collaboration with JSM on roughly equal terms is largely due to sexism.

2.4 A More Balanced Perspective

The available evidence somewhat underdetermines judgments about the value to JSM of his collaboration with HTM. Too much of their collaborative activity took the form of conversations behind closed doors, of which we have no record. When one of them was traveling then they did correspond, but few of HTM's letters survive (and Jack Stillinger remarks on "how seldom *ideas* are touched on" in the correspondence between them that we do have (1961, 26)). As with the descriptions of HTM's character and ability discussed above, the proposition that the truth about the nature of the Mills' intellectual partnership lies somewhere between some of the more extreme views that have been advanced is intuitively very plausible, and it is at least consistent with the evidence. One such intermediate view of their collaboratory activity is that of Bain, who suggests that just as JSM's friend John Sterling "overflowed in suggestive talk, which Mill took up and improved in his own way," so HTM might have done as well (1882, 173ff). If this was HTM's role, then her contribution might largely have taken the form of articulating and passionately advocating certain progressive social and political views—Socialism, women's rights, an absolutist stance about individual liberty—for which JSM then applied himself to developing sophisticated utilitarian arguments. JSM's statement that his task was that of "building the bridges and clearing the paths which connected" HTM's "truths" with his "general system of thought" can be read as implying that this was the case. Of course, it is the strength of those arguments—which is in part a function of the attractiveness of that general system of thought—on which JSM's claim to be regarded as a major figure in the field of social-political philosophy rests. This is not necessarily to say that he held these positions only because she did, which he explicitly told us was not the case with respect to women's rights (1881, 253n), but only that her strong attachment to them was an important part of his motivation for writing about them.

But even if this was her role, she must have been able to talk about those positions in an intelligent manner, to make them seem reasonable; otherwise JSM would never have been moved to try and make a case for them. At the very least, HTM surely did more than simply declare herself a Socialist, a radical (for the time) feminist, etc. Bain knew JSM extremely well, and even though he says that his friend was under "an extraordinary hallucination as to the personal qualities of his wife," and "outraged all reasonable credibility in describing her matchless genius," he is also adamant not only that JSM "was not such an egoist as to be captivated by the echo of his own opinions" but also that he would only have been stimulated by someone with "independent resources" who had a "good mutual understanding as to the proper conditions of the problem at issue" (1882, 173f). If HTM's primary contribution was that of stimulating JSM by "suggestive talk," then her suggestive talk very likely included the idea that the social and political reforms she favored could help to foster the improvement of mankind. John Robson, the editor of JSM's *Collected Works* and author of a book on JSM titled *The Improvement of Mankind*, comments that "[I]n what we have of her writings, Harriet constantly has her eye on the future, even when criticizing the present; she was a woman of dreams and aspirations, and she must constantly have breathed into Mill a hopeful and expansive view of human possibilities" (1966, 178). JSM's defenses of Socialism, women's rights, and liberty were progressive not only in the sense that were in advance of Victorian opinion, but also in the sense of being grounded on a conception of "man as a progressive being" (1977, 224). Each was presented as a way of securing the necessary conditions for the moral and intellectual development of humanity, and JSM continually asserted that in the absence of this

development the species would enjoy only a small fraction of the happiness of which it is capable. HTM's encouragement may have been largely responsible for the fact that the proposition that human improvement is desirable, and that in the right social and political context it is possible, loomed so large in his thought. If this is correct, then even if HTM bears little responsibility for the specifics of JSM's arguments for the social and political institutions and practices that she advocated—arguments on which much of his philosophical reputation rests—she would still be responsible not only for inspiring him to make those arguments but also for the fact that they have the general character they do.

Bibliography

- Bain, A., 1882, *John Stuart Mill: A Criticism with Personal Reflections*, London: Longmans, Green, and Co.
- Hayek, F. A., 1951, *John Stuart Mill and Harriet Taylor: Their Correspondence and Subsequent Marriage*, Chicago: University of Chicago Press.
- Himmelfarb, G., 1974, *On Liberty & Liberalism: The Case of John Stuart Mill*, New York: Knopf.
- Jacobs, J. E., "'The Lot of Gifted Ladies is Hard': A Study of Harriet Taylor Mill Criticism," *Hypatia's Daughters: Fifteen Hundred Year of Women Philosophers*, ed. L. L. McAlister, Hypatia, 215-47.
- -----, 1998, "Introduction," *The Complete Works of Harriet Taylor Mill*, ed. Jo Ellen Jacobs, Bloomington: Indiana University Press, xi-xxxv.
- -----, 2000, "Harriet Taylor Mill's Collaboration with John Stuart Mill," *Presenting Women Philosophers*, Philadelphia: Temple University Press, 155-66.
- Kamm, J., 1977, *John Stuart Mill in Love*, London, Gordon and Cremonesi.
- MacMinn, N., Hainds J. R., and McCrimmon J., 1945, *Bibliography of the Published Writings of John Stuart Mill*, Bloomington: Pantagraph Press.
- Mazlish, B., 1975, *James and John Stuart Mill: Father and Son in the Nineteenth Century*, New York, Basic Books.
- Mill, H. T., 1998, *The Complete Works of Harriet Taylor Mill*, ed. Jo Ellen Jacobs, Bloomington: Indiana University Press.
- Mill, J. S., 1965, *Principles of Political Economy*, *Collected Works of John Stuart Mill* vol. II and III, ed. J. Robson, Toronto: Toronto University Press.
- -----, 1973, *A System of Logic: Ratiocinative and Inductive*, *Collected Works of John Stuart Mill* vol. VII and VIII, ed. J. Robson, Toronto: Toronto University Press.
- -----, 1977, *On Liberty*, *Essays on Politics and Society*, *Collected Works of John Stuart Mill* vol. XVIII, ed. J. Robson, Toronto: Toronto University Press, 215-310.
- -----, 1981, *Autobiography*, *Autobiography and Literary Essays*, *Collected Works of John Stuart Mill* vol. I, ed. J. Robson and J. Stillinger, Toronto, Toronto University Press, 1-290.
- -----, 1984, *The Subjection of Women*, *Essays on Equality, Law, and Education*, *Collected Works of John Stuart Mill* vol. XXI, ed. J. Robson, Toronto: Toronto University Press, 259-348.
- Mineka, F., 1963. "The *Autobiography* and the Lady," *University of Toronto Quarterly*, 32: 301-6.
- Packe, M., 1954, *The Life of John Stuart Mill*, New York, MacMillan.
- Pappe, H. O., 1960, *John Stuart Mill and the Harriet Taylor Myth*, Melbourne: Melbourne

University Press.
- Robson, J., 1966, “Harriet Taylor and John Stuart Mill: Artist and Scientist,” *Queens Quarterly*, 73: 167-86.
- ----- and Stillinger, J., 1981, “Introduction,” *Autobiography and Literary Essays, Collected Works of John Stuart Mill* vol. I, ed. J. Robson and J. Stillinger, Toronto, Toronto University Press, vii-liv.
- Rossi, A., 1970, “Sentiment and Intellect: The Story of John Stuart Mill and Harriet Taylor Mill,” *Essays on Sex Equality*, Chicago: University of Chicago Press, 3-63.
- Stillinger, J., 1961, “Introduction,” *The Early Draft of John Stuart Mill’s* Autobiography, Urbana: University of Illinois Press, 1-33.
- Trilling, D. “Mill’s Intellectual Beacon,” *Partisan Review*, 19: 115-20.

Other Internet Resources

- Spartacus Educational page on Harrriet Taylor
- Excerpts from "The Enfranchisement of Women"
- "On the Probable Futurity of the Working Class" *Principles of Political Economy*, Bk. 4, Ch. 7
- Helen Taylor's essay "The Claim of Englishwomen to the Suffrage Constitutionally Considered" (1867)

Related Entries

liberalism | Mill, John Stuart | socialism

Dale E. Miller
demiller@odu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 11, 2002
Content last modified: June 11, 2002

Stanford Encyclopedia of Philosophy
Notes to Harriet Taylor Mill

Notes

1. Eliza Flower.

2. For purposes of comparison, note that Alexander Bain convinced Helen Taylor, who prepared the manuscript of Mill's *Autobiography* for publication after his death, to strike some passages about her that were almost as laudatory. When the book was first published, the lines in question were replaced by strings of asterisks; they have been replaced in most contemporary editions (see Robson and Stillinger 1981, xxix n55). Bain also urged Helen to delete some parts of JSM's description of HTM, but she decided against this.

3. In *The Subjection of Women* JSM writes:

> Hardly anything can be of greater value to a man of theory and speculation who employs himself not in collecting materials of knowledge by observation, but in working them up by processes of thought into comprehensive truths of science and laws of conduct, than to carry on his speculations in the companionship, and under the criticism, of a really superior woman. There is nothing comparable to it for keeping his thoughts within the limits of real things, and the actual facts of nature. A woman seldom runs wild after an abstraction. The habitual direction of her mind to dealing with things as individuals rather than in groups, and (what is closely connected with it) her more lively interest in the present feelings of persons, which makes her consider first of all, in anything which claims to be applied to practice, in what manner persons will be affected by it—these two things make her extremely unlikely to put faith in any speculation which loses sight of individuals, and deals with things as if they existed for the benefit of some imaginary entity, some mere creation of the mind, not resolvable into the feelings of living beings. Women's thoughts are thus as useful in giving reality to those of thinking men, as men's thoughts in giving width and largeness to those of women. In depth, as distinguished from breadth, I greatly doubt if even now, women, compared with men, are at any disadvantage (1984, 306).

It is not exactly clear what of this nature JSM thinks that HTM added to his social and political thought, however. John Robson writes of the passage from the *Autobiography* that:

> Perhaps he means nothing more unusual than that one can quickly see whether one's actions fulfill one's intentions in the daily concerns of life. He may also be merely referring to his pronounced inability to manage such practical matters as ordering groceries

> and dealing with difficult neighbors, for here his reliance on her was unusual, and will become notorious…. (1966, 179).

4. JSM prepared a comprehensive bibliography of his publications, in which a number of works are described as "joint productions" with HTM—of some of these he even notes that "very little" in them was his, or that he "acted chiefly as amanuensis to my wife"). Many of these are newspaper articles, most of which concern domestic violence. *The Principles of Political Economy* is also described as a joint production. Interestingly, Mill's bibliography does *not* list *On Liberty* as a joint production.

Dale E. Miller
demiller@odu.edu

First published: June 11, 2002
Content last modified: June 11, 2002

Stanford Encyclopedia of Philosophy

Personal Autonomy

To be autonomous is to be a law to oneself; autonomous agents are self-governing agents. Most of us want to be autonomous because we want to be accountable for what we do, and because it seems that if we are not the ones calling the shots, then we cannot be accountable. More importantly, perhaps, the value of autonomy is tied to the value of self-integration. We don't want to be alien to, or at war with, ourselves; and it seems that when our intentions are not under our own control, we suffer from self-alienation. What conditions must be satisfied in order to ensure that we govern ourselves when we act? Philosophers have offered a wide range of competing answers to this question.

Introduction

When people living in some region of the world declare that their group has the right to live autonomously, they are saying that they ought to be allowed to govern themselves. In making this claim, they are, in essence, rejecting the political and legal authority of those not in their group. They are insisting that whatever power these outsiders may have over them, this power is illegitimate; they, and they alone, have the authority to determine and enforce the rules and policies that govern their lives.

When an individual makes a similar declaration about some sphere of her own life, she, too, is denying that anyone else has the authority to control her activity within this sphere; she is saying that any exercise of power over this activity is illegitimate unless she authorizes it herself. Most of the reasons that can be offered in support of this claim have correlates in the case of demands for group autonomy. But there is one very important exception: a reason that takes us beyond politics, to the metaphysics of agency.

An agent is one who acts. In order to act, one must initiate one's action. And one cannot initiate one's action without exercising one's power to do so. Since nothing and no one has the power to act except the agent herself, she alone is entitled to exercise this power, if she is entitled to act. This means that insofar as someone is an agent—i.e., insofar as she is one who acts—she is correct to regard her own commitments to acting, her own judgments and decisions about how she should act, as authoritative. Indeed, if she were to challenge the authority that is an essential feature of her judgments and decisions, then they would cease to be her own practical conclusions. Their power to move her would cease to be a manifestation of her power to move herself; it would not be the power of her own agency.

In short, every agent has an authority over herself that is grounded, not in her political or social role, nor in any law or custom, but in the simple fact that she alone can initiate her actions. To be sure, it might be unwise for someone to follow the commands she gives to herself when she "makes up her mind." The point, however, is that she has no conceivable option. In order to form an intention to do one thing rather than another, an agent must regard her own judgment about how to act as authoritative—even if it is only the judgment that she should follow the command or advice of someone else. This tight connection between being an agent and having authority has no correlate in cases where the authority at issue is political. Anyone can coherently (and often plausibly) challenge the political authority of some individual or group. Even a political leader herself can with good reason believe that her political power is illegitimate, and that exercising this power is unjustified.

Despite the special inalienable nature of our authority over ourselves, it is possible for us to fail to govern ourselves, just as it is possible for a political leader to fail to govern those who fall within her domain. Indeed, precisely because our authority over our own actions is a formal feature of agency, our deference to this authority is but the form of self-government. It does not imply that whenever we act, the forces that move us are manifestations of our power as agents; the power of our motives is not necessarily an expression of the decision-making power that is constitutive of our agency, and the basis of our authority to decide how to act. Just as a political leader's official status is compatible with her having no real power to call the shots, so too, a person can have an authoritative status with respect to her motives without having any real power over what she does. Though it is an agent's job to determine how she will act, she can do this job without really being in control. Of course, no one can govern herself without being subject to influences whose power does not derive from her own authority: everything we do is a response to past and present circumstances over which we have no control. But some of these influences are different. Some of the forces that move us to act do not merely affect which actions we choose to perform, nor how we govern ourselves in making these choices. They influence us in a way that makes a mockery of our authority over our actions. They undermine our autonomy.

What distinguishes autonomy-undermining influences on a person's will from those motivating forces that merely play a role in the self-governing process? This is the question that all accounts of autonomy try to answer. As the number and variety of these accounts indicate, the distinction is extremely elusive. There is certainly widespread agreement about the paradigm threats to personal autonomy: brainwashing and addiction are the favorite examples in the philosophical literature. But philosophers seem unable to reach a consensus about the precise nature of these threats. They cannot agree about how it is that certain

influences on our behavior prevent us from governing ourselves.

This disagreement about the defining characteristics of autonomous agency reflects the fact that even as concrete examples call attention to a very real difference between those who govern themselves and those who do not, there are significant conceptual obstacles to making sense of this distinction. These obstacles are tied to the very feature of agency mentioned above—the feature that appears to support the demand that individuals be granted considerable political and legal power. If an agent fails to govern herself, this must be because, at the time, she lacks the power to do so. But in what can this powerlessness consist? If she necessarily has the authority to determine how she will act, and if this essential feature of agency is inseparable from the fact that she necessarily defers to herself whenever she initiates her action, then how can her behavior possibly escape her control? Intuitively, agents can fall under the sway of desires, or urges, or compulsions whose power is at odds with their power as agents; they can be moved by such impulses "in spite of themselves." But in what sense, exactly, are such motives "external" to the agent herself? How can their power to cause her to act fail to be a manifestation of her power to act?

The puzzle at the heart of these questions is a puzzle about the relationship between the agent's power and the power of the forces that move her. And it is a puzzle about the relationship between the agent's authority and the status of these motivating forces. What distinguishes motives whose power is attributable to the agent herself from motives whose power is external to the agent's? What distinguishes motives on which the agent has conferred her authority from motives whose power has reduced her authorization to a mere formality? When the governing agent and the agent she governs are the very same self, we cannot answer either of these questions without answering the other. This is why it is so difficult to produce a satisfactory account of personal autonomy.

(Again, the perplexity to which these questions give voice does not have a correlate in the political case. We can easily grasp the idea of a country's army (or legislative body, or cabinet ministers) dictating to the president what legislation he must approve; for in this case there are (at least) two independently identifiable decision-makers—each with its own point of view, each with its own power. The difficulty in the case where the relevant powers are all within the psyche of the individual agent is that there is no such independently identifiable pair of standpoints in terms of which we can distinguish the powers that bully the agent from the powers that can be attributed to the agent herself. An account of the conditions under which an individual agent is bullied by her motives is, at the same time, an account of what makes a motive external to the agent's own standpoint.)

Four More or Less Overlapping Approaches to Personal Autonomy

Philosophers have proposed many different accounts of the autonomous agent's special relation to her own motives. According to one prominent approach, which one might call coherentist, the governing self in this relation is represented by reflexive attitudes—"higher-order" attitudes toward the mental states

that move her to act.[1] Someone is an autonomous agent, the coherentists argue, if and only if she accepts her motives, or identifies with them, or approves of them, or believes that they make sense in terms of her long term commitments or plans. Intuitively, if someone repudiates the causal efficacy of her own motives, then their power is independent of her authority. If, on the other hand, she endorses these motives, then her actions occur with her permission, if not necessarily at her command. In the second case, the motives are truly her own; in the first case, they are external forces in conflict with her own causal powers.

On a strict coherentist approach to autonomy, autonomous agents can be moved by desires they are helpless to resist: though an addict fails to govern herself if she would rather resist her irresistible urge to take drugs, she is an autonomous agent if she has no objection to her addiction and its motivational effects. According to the coherentist, moreover, both the origin and the content of a person's higher-order attitudes are irrelevant to whether she is an autonomous agent. She need have done nothing to bring it about that she has these attitudes; and the attitudes need not be especially rational or well-informed. Coherentist accounts are thus doubly internalist. They express the intuition that whether we govern ourselves depends on neither how we came to be who we are (a fact that is external to the action itself) nor how our beliefs and attitudes relate to reality (a fact that is external to the beliefs and attitudes themselves). There need be no special relation between our autonomy-constituting attitudes and either the past circumstances that caused these attitudes or the present circumstances in response to which they move us to act.

Other accounts of autonomy introduce conditions that are externalist in one or both of these ways. According to those who advocate a reasons-responsive conception of autonomous agency, an agent does not really govern herself unless her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does.[2] On accounts of this type, an agent who is unresponsive to the reasons for “standing behind,” or “backing up,” certain motives and not others is not in the proper position to authorize her own actions. Whether the relevant reasons are grounded in facts about her own desires and interests, or whether they have some independent source, the idea is that someone is not qualified to govern herself if she cannot understand what she has reason to do, or (if this is a distinct handicap) is incapable of being moved by these reasons. In effect, her exercise of authority is so ill-conceived that it is powerless to confer legitimacy on her motives.

The feature of this account that most distinguishes it from the coherentist account is the importance it attributes to an agent's ability to appreciate the reasons she has. (Once she appreciates these reasons, her inability to act accordingly is, essentially, the inability to conform her act to her (higher-order) desire to be moved accordingly.) What, exactly, is the connection supposed to be between being out of touch with (evaluative and nonevaluative) reality and failing to govern oneself? Clearly, a person who fails to appreciate a wide range of reasons for action is unlikely to govern herself well: she is likely to do things that will, in the long run, thwart her own purposes and interests. The reasons-responsiveness conception of autonomy thus appears to reflect the intuition that when we do something very poorly, we do not really do it at all. There is, however, another possible underlying rationale for regarding ignorance as a threat to self-government. If doing *Y* is constitutive of doing *Z*, then if I authorize myself to be moved by the desire to do *Y* because I mistakenly believe that doing *Y* is a way of not doing *Z*, then there is an

obvious sense in which I have not authorized myself to do what I am now doing when I am moved by the desire to do *Y*. So, if I have a general desire to do what is right and prudent, or, even more generally, a desire to do what I can justify to others, then insofar as I am moved to act in ways that are incompatible with satisfying these desires, there is a sense in which I—who am committed to doing only what I have good (enough) reason to do—have not really authorized my action. Alternatively, we could say that, under these circumstances, something external to my power to guide myself by reasons has prevented me from exercising this power, and so has prevented me from governing myself.

An additional source of support for the reasons-responsive conception of autonomy comes from the thought that someone who cannot respond to the reasons there are must have a limited ability to reason. This brings us to a third popular approach to autonomous agency—an approach that stresses the importance of the reasoning process itself.[3] According to responsiveness to reasoning accounts, the essence of self-government is the capacity to evaluate one's motives on the basis of whatever else one believes and desires, and to adjust these motives in response to one's evaluations. It is the capacity to discern what "follows from" one's beliefs and desires, and to act accordingly. One can exercise this capacity despite holding false beliefs of all kinds about what one has reason to do. Being autonomous is not the same thing as being guided by correct evaluative and normative judgments.

The emphasis on an autonomous agent's responsiveness to her own reasoning reflects the intuition that someone whose education consisted of a method of indoctrination that deprived her of the ability to call her own attitudes into question would, in effect, be governed by her "programmers," not by herself. So, too, someone whose practical reasoning was directly manipulated by others would not govern herself by means of this reasoning. And so, it seems, she would have no power over the motives that this reasoning produced.

Like the coherentists, advocates of responsiveness to reasoning accounts believe that the key to autonomous agency is the ability to distance oneself from one's attitudes and beliefs—to occupy a standpoint that is not constituted by whatever mental states are moving one to act. They agree that motives authorized from this reflective standpoint are internal to the agent herself in a way that her other motives are not. Unlike the coherentists, however, the reasoning-responsive theorists believe that there is more to the capacity for self-reflection than the capacity to hold higher-order attitudes. The authority of our higher-order attitudes is grounded, they claim, in the authority of the practical reasoning that supports these attitudes. So a self-governing agent does not merely endorse her motives: her endorsements are implicit claims about which motives have the support of her reason.

This fact is closely tied to another. Like many accounts that stress an autonomous agent's responsiveness to reasons, responsiveness to reasoning accounts often suggest that self-government requires the capacity for self-transformation. On this assumption, an autonomous agent is capable of changing her mind when she discovers good reason to do so.[4] In contrast, strict coherentists insist that it is possible to act autonomously while being moved by desires that are not only irresistible when they produce their effects, but so integral to one's identity that one could not possibly will to resist them, no matter how convincing one found the arguments in favor of doing so.

The conception of autonomous agency as responsiveness to reasoning clearly has a more internalist character than the conception of autonomous agency as responsiveness to reasons: according to those who stress the autonomous agent's ability to evaluate her own motives, what counts is not the relation between the agent's attitudes and external reality, but her ability to draw inferences from what she wants and believes, and by so doing, to reconsider—to rationally reflect upon—her other desires and beliefs. Insofar, however, as a responsiveness to reasoning account presupposes a particular conception of practical reasoning, it appeals to standards, or principles, that the agent herself might misapply, or fail to recognize altogether. Moreover, even if advocates of autonomy as responsiveness to reasoning have nothing in particular in mind when they speak of the process of "reflection," "rational evaluation," etc., reasoning is a norm-governed process that an agent might reject for reasons of her own. Responsiveness to reasoning accounts thus contain an externalist element that is absent from strict coherentist accounts. They imply that an agent can be mistaken about whether she is really reasoning—and so can be mistaken about whether the power of her motives reflects her authority over her own actions.

This weak externalism naturally expands into more robust varieties. In particular, it supports the idea that whether an agent's reasoning is really her way of authorizing her actions depends on which forces exert a nonrational influence on this reasoning. Even when indoctrination and other more or less imaginary forms of "mind control" do not prevent a person from reaching evaluative conclusions about her own motives, they can prevent her from thinking for herself. So, too, it seems, someone in the grip of compulsion or addiction can be so bullied by this condition that whatever facts she considers, and whatever conclusions she draws, cannot legitimately be attributed to her. One way to interpret these cases is to say that the person's reasoning falls so far short of the norms of "rational reflection" that she is not really reasoning at all. Alternatively, one can say that her reasoning does not guarantee her autonomy because it is under the control of external forces.

Insofar as accounts of autonomy simply stipulate that certain influences on an agent's intention-forming process "interfere with," or "pervert" this process—insofar as they do not explain what distinguishes "internal" from "external" forces—these accounts are incomplete. For they leave it mysterious why certain influences, and not others, are a threat to self-government. One response to the mystery is offered by the reasons-responsive account: the autonomy-undermining influences are the ones that prevent the reasoning process from being sufficiently sensitive to the reasons there are. A fourth approach to autonomy, very different from the other three mentioned so far, rejects the mystery as a symptom of confusion. Thus some philosophers argue that cases of mind-control simply call our attention to the fact that whenever our motives are causally determined by events over which we have no control, their power does not depend on our authority. According to this incompatibilist conception of autonomy, autonomy is incompatible with determinism. If our actions can be fully explained as the effects of causal powers that are independent of us, then even if our beliefs and attitudes are among these effects, we do not govern them, and so we do not govern ourselves.[5]

The approaches just sketched have been developed in many subtly different ways. Some of these differences reflect disagreements over the extent to which the relevant conditions—coherence among higher- and lower-order attitudes, responsiveness to reasons, responsiveness to reasoning, freedom from

determination by external causes—must actually be manifest when an agent determines her will, or whether it is enough that under certain specified circumstances the agent would relate to her motives in the stipulated manner. There is also a difference of opinion about the scope of the relevant capacities: Must an autonomous agent be able to respond to a wide range of reasons for and against her action? or is it enough that her motives are responsive to the “strongest,” “most compelling” reasons? and can these reasons include the sort of credible threats that figure in cases of coercion? What range of attitudes must an autonomous agent be capable of calling into question? How well must she be capable of reasoning? Does it matter whether she is guided by certain principles of rationality? Must it be possible for her to draw different conclusions on the basis of the reasons she considers? Is it essential that she could have considered a different set of reasons instead? Clearly, the many possible answers to these questions can be combined in many different ways. And, more generally, the basic approaches themselves often figure together as necessary or sufficient conditions in a single complex account.

Challenges to Identifying the Minimal Conditions of Personal Autonomy

Whatever specific form the competing proposals take, they all contribute to our understanding of the various ways in which agents can play a governing role in their own actions. They articulate an ideal that agents can realize to various degrees. And in so doing, they shed light on how, with the proper training, a very young child, whose deference to the authority of her own judgments is little more than the form of self-government, can develop into a fully autonomous agent.

This is a very important contribution. Nonetheless, it falls short of giving us everything we have reason to expect from an account of autonomy. In particular, challenges to the different approaches sketched above suggest that no account built from these materials can succeed in distinguishing autonomous from nonautonomous agency. In other words, none of these accounts seems to identify the minimal conditions under which a person can be said to act on her own authority—the conditions that must be satisfied if someone's exercise of authority over her action is to be, not a mere formality, but a way of relating to herself that renders her accountable for the forces that move her to act.

Consider, first, the alleged requirement of responsiveness to reasons. To many critics, there is an obvious problem with this requirement. A person, they argue, can govern herself even if she does not understand the significance of what she is doing. To govern oneself is to maintain a certain self-relation; and, many insist, the elements in this relation include one's own beliefs, however unreliable these may be. In killing Desdemona, Othello fails to accomplish his aim of doing what he has good reason to do. But this does not prevent him from being the author of his own misguided actions. So, too, an envious, vengeful, and very stubborn person does not fail to govern herself just because she is unresponsive to the wide range of reasons against trying to sabotage her colleague's career. An agent's failure to respond to certain reasons may be good evidence of the fact that she is not really the author of her actions. But being out of touch with (evaluative or nonevaluative) reality is not the same thing as lacking autonomy.

Nor does autonomous agency require that one's actions be compatible with one's long-term plans. Such plans often enable a person to exercise some measure of control over her life as a whole; they are her way of governing her more local exercises of self-government. But a person can govern herself at a particular time even while defying her earlier attempts to place constraints on how she will govern herself at this time. She can take it upon herself to abandon her plans, or to modify them in ways she did not anticipate when she first made them. She can even reject the counsel of her long-term values.

Even under such circumstances, an autonomous agent “identifies with” the mental states that move her to act. Many philosophers have thus embraced the coherentist position that the attitude of identification is the key to personal autonomy. But this approach, too, has problems that cannot be solved by combining it with other approaches, problems that become evident when we try to spell out what is involved in identifying with one's motives.

Without an account of identification, we have not advanced beyond our initial intuition that the actions of self-governing agents are caused by motives that are, in some sense, internal to the agents themselves. According to the most popular account, someone identifies with her motives if and only if she endorses, or approves of, them. This, allegedly, is what makes them internal; it is what ensures that their power is really her own. One obvious problem with such accounts is that a person could be brainwashed, or otherwise compelled, to endorse a given motive. Indeed, her brain could be manipulated in such a way that her endorsement is highly responsive to reasons. This has led many to supplement coherentist accounts of autonomy with additional conditions that place constraints on the causal history of an agent's endorsements. But there is reason to doubt that such endorsements are even a necessary condition of autonomous agency. In particular, if to endorse one's motive is, essentially, to judge that this motive—or acting from this motive—is good, then the endorsement account of autonomy does not appear to accommodate cases of weakness of will.

By definition, a weak-willed action is an action that someone performs against her best judgment, even while “acting of her own free will” in whatever sense suffices to render an agent accountable for her behavior. A weak-willed agent authorizes herself to act as she does, despite her belief that she has good reason to act otherwise. When someone asserts her authority in this way, she is criticizably irrational; and it is notoriously difficult to make sense of this form of irrationality.[6] For our purposes, however, it is enough to note that if weakness of will is a genuine phenomenon, then human agents have the capacity to govern themselves in a way that they themselves take to be unjustified. They can claim for themselves an authority that challenges the authority of their own reason.

Again, someone whose action

is caused in this way does not govern herself as thoroughly as someone whose will is “strong.” For she acts for a reason that she herself deems inadequate; and so she is not (adequately) governed by the norms of her own thought. The point, however, is that even under these conditions, she is an autonomous agent. Self-conflicted though she may be, she is still accountable for what she does. This is not simply because what she does is the result of an earlier autonomous action. Rather, her accountability is intrinsic to her

weak-willed action itself. She is a weak-willed self-governor.

The possibility of weakness of will implies that the authorization an agent must give to her motives if she is to count as (minimally) governing their effects need not take the form of the judgment that no alternative action would be better. Some philosophers have lent support to this conclusion by arguing that there are other cases in which a person's authorizations can come apart from her evaluations. Sometimes, they claim, a person cannot follow the recommendations of her own reason without betraying herself. A woman, for example, may conclude that even though she has very good reason to give up her child for adoption, she cannot recognize herself in this action, and so cannot identify with the desire to perform it.[7] Reasonable people will surely disagree about how best to interpret any particular example. But human experience does seem to support the general point: the human capacity for self-reflection enables human agents to distance themselves in thought from every aspect of their own psyches—even their rational reflections. Given this possibility, a person's identification with her motives cannot be cashed out in terms of higher-order attitudes of approval and disapproval, or in terms of the rational reflections that typically ground these attitudes.

What is it, then, to "identify with" certain features of one's mental life? It would seem to be nothing more than to confer one's authority upon them, i.e., to authorize their influence. But if this is right, then the concept of identification cannot help us to distinguish autonomous agents from the rest. It does not provide us with the looked-for explanation of what distinguishes an autonomous agent from someone who exercises her authority at the bidding of external powers, and whose authorization of her motives is thus a mere formality.

Most philosophers fail to recognize the extent of this problem. They acknowledge the possibility of autonomy-undermining influences, like brainwashing, that exert their power behind the scenes. But they also believe that there are straightforward cases in which a person lacks autonomy because she performs an action without authorizing herself to perform it. They point to the case of someone who takes drugs when she would rather resist the motivating force of her addiction. They note that even if many people who fit this description are merely weak-willed, not all of them are. Some addicts, compulsives, and others suffering from emotional and psychic distress are the helpless victims of their own psychological states; and these unfortunate agents lack autonomy precisely because they repudiate their own motives.

Compelling as this diagnosis may be, however, it proves too much. It assimilates addicts to people like those with Tourette's Syndrome, whose behavior is not even voluntary. It reduces the distinction between autonomous and nonautonomous agents to the distinction between agents and nonagents. If someone's motives directly defy her authority (rather than ensuring that she exercises this authority on their terms), then her behavior does not reflect her deference to this authority, and so it fails to satisfy an essential feature of agency. Under such circumstances, a person's motives have a power that is not only unauthorized by the agent herself, and hence, distinct from—external to—her power as an agent; they produce their effects in a way that bypasses her (her agency) altogether. Even if she can acknowledge that there is something to be said for behaving as she does, she is a passive bystander to this behavior, as alienated from the causal efficacy of her own motives as she is from the causal efficacy of the physiological states that produce her reflex movements. She is not an autonomous agent because she is

not an agent at all.

Agents as Causes and the Practical Point of View

If agents cannot initiate their own actions “on purpose” without authorizing themselves (their motives) to do so, then the distinguishing feature of autonomous agents is not that they identify with their motives but that the authority they assert in doing so is more than a mere formality. What does this difference come to? The incompatibilist, we saw, has a ready answer: in the case of autonomous agency, and only in the case of autonomous agency, there is more to the agent's assertion of authority than the expression of external power. The familiar problem with this answer is that there seems to be no way for an agent to gain an extra measure of control over her motives simply by acquiring attitudes or judgments or other mental states that are not determined by anything else. If someone's attitude toward her motives is not determined by any earlier state of affairs, then how can it be determined by her?

This question pushes those with incompatibilist intuitions to attribute a special causal power to agents—a power that is not reducible to the power one event transmits to another. A person, some incompatibilists argue, can agent-cause a certain response to earlier events in a way that is not itself the effect of these earlier events. Simply by virtue of being the particular person she is, she can bring it about that she is motivated in a certain way. And the explanatory fact that she is this particular person cannot be reduced to any more basic facts about her dispositions to respond in certain ways to certain inputs; her power over her actions cannot be reduced to the power of external motivating forces.[8]

The obscurities of agent-causation are enough to prevent most philosophers from embracing this conception of autonomous agency. To mention just a few familiar challenges: If agent-causing an event does not involve doing anything, then how does the agent exert her causal power? and why does this power produce its effects at one time rather than another? If, on the other hand, the agent must do something in order to agent-cause an action, then doesn't this require that she undergo some change? and isn't this change of state itself an event? On the basis of these and other difficulties, many conclude that the appeal to agent-causation provides no more insight into autonomy than the simple assertion that we can sometimes govern the effects of our own motives. Yet the agent-causation theorists call our attention to something important. The strong conviction that we are often autonomous agents is grounded in the basic experience of determining our own wills. We believe that we have the capacity for self-government because we believe that, whatever forces may be pressing us to act, it is ultimately “up to us” to determine what to “make of” the pushes and pulls that constitute our mental life.

The seeming incoherence of agent-causation, and more generally, the seeming impossibility of articulating a conception of agency according to which the capacity for self-government depends on the agent's freedom from determining causes, leads some to conclude that autonomous agency is an illusion: our deep conviction notwithstanding, we do not really know that we can govern ourselves. Others, however, see things differently. They argue that this pessimistic conclusion reflects a misunderstanding of the very nature of rational agency. In making their case, they take their lead from the philosopher who has contributed more than any other to our understanding of autonomy. Kant, they note, stresses the deep

differences between the two points of view from which we can think about ourselves and our world.[9] We take up the theoretical point of view in order to gain knowledge about the nature of reality, and on this basis make predictions about which effects will follow from which causes. When we want to make up our minds about what to do, however, we take up the practical point of view. From this point of view, too, we survey the facts that are relevant to our decisions. But none of these facts, taken singly or together, is intrinsically action-guiding; none can free us from the task of drawing our own conclusions about what we have reason to do. This is true even of facts about reasons. Whatever the basis of these facts may be, the normative relations among them are far from determinate. We have to make the necessary determinations ourselves. Given everything we know about what is and what ought to be, we have to determine how we are going to act.

Necessarily, theoretical reasoners are passive bystanders to the events on the basis of which they predict future events. But practical reasoners are not mere observers of the passing scene. As practical reasoners, we have no choice but to determine our responses to what we observe—even if everything we do—and so, everything we decide to do—is determined by events in the past. To make up our minds, we need not be sophisticated reasoners. We need not even be capable of doubting the legitimacy of our most powerful motives. We must, however, find a reason to do one thing rather than another. And since no fact can play the role of a reason unless someone takes it to be a reason, practical reasoners necessarily have the ultimate authority over the powers that move them.

Conclusion

We are back where we started. The demand to be permitted to govern ourselves reflects the conviction that we are, in essence, self-governors. In essence, but not always in fact. Sometimes our authority over our actions is nothing but the form of self-government. Sometimes we are not autonomous agents. If, then, the structure of rational agency justifies our conviction that we are capable of governing our own actions, it does not hold the key to the distinction between those cases in which we fail to exercise this capacity and those in which we succeed. The conviction that there is such a distinction is grounded in the obvious fact that victims of brainwashing, compulsion, addiction, depression, anxiety, and many other conditions are prevented from governing themselves. If their lack of autonomy is not simply a function of the fact that their actions are causally determined by states of affairs over which they have no control, and if it is not equivalent to any fact about the considerations they are disposed to recognize and be moved by, then it would seem to be a more intrinsic feature of their agency. No particular attitude seems to be essential to autonomous agency, however—except, of course, the attitude of authorization that is essential to all action for a reason. Nor is it necessary that any particular principles of reasoning serve the autonomous agent as guides—except, again, whatever principles must guide the action of even nonautonomous agents. The content of our desire to govern ourselves when we act thus remains obscure to us, even as the legitimacy of this desire is clear.

Bibliography

- Albritton, R., 1985, "Freedom of Will and Freedom of Action," *Proceedings and Addresses of the American Philosophical Association* 59, no. 2.
- Allison, H., 1990, *Kant's Theory of Freedom*, Cambridge: Cambridge University Press.
- Aristotle, 1985, *Nicomachean Ethics*, I and VII, T. Irwin (trans.), Indianapolis: Hackett Publishing Co.
- Audi, R., 1993, *Action, Intention, and Reason*, Ithaca: Cornell University Press.
- Benn, S., , 1988, *A Theory of Freedom*, New York: Cambridge University Press.
- Bennett, J, 1980, "Accountability," in *Philosophical Subjects: Essays Presented to P.F. Strawson*, Z van Straaten (ed.), Oxford: Clarendon Press, 14-47.
- Benson, J., 1983, "Who is the Autonomous Man?" *Philosophy* 58: 5-17.
- Berlin, I., *Four Essays on Liberty*, Oxford: Oxford University Press.
- Berofsky, B., (ed.), 1966, *Free Will and Determinism*, New York: Harper and Row.
- -----, 1987, *Freedom from Necessity: The Metaphysical Basis of Responsibility*, New York: Routledge and Kegan Paul.
- -----, 1983, "Autonomy," in *How Many Questions? Essays in Honor of Sidney Morgenbesser*, L. S. Cauman, I. Levi, C.D. Parsons, and R. Schwartz (eds.), Indianapolis: Hackett Publishing Co., 301-19.
- -----, 1992, "On the Absolute Freedom of the Will," *American Philosophical Quarterly* 29: 279-89
- -----, 1995, *Liberation from Self: A Theory of Personal Autonomy*, New York: Routledge and Kegan Paul.
- Bishop, J., 1989, *Natural Agency: An Essay on the Causal Theory of Action*, Cambridge: Cambridge University Press.
- Blumenfeld, D, "The Principle of Alternate Possibilities".
- Bok, H., 1998, *Freedom and Responsibility*, Princeton: Princeton University Press.
- Bratman, M., 1979, "Practical reasoning and Weakness of the Will," *Nous* 13: 131-51.
- -----, 1987, Intention, Plans, and Practical Reason, Cambridge, Mass.: Harvard University Press.
- -----, 1996, "Identification, Decision, and Treating as a Reason," *Philosophical Topics* 24: 1-18.
- -----, 1999, Faces of Intention: Selected Essays on Intention and Agency, Cambridge: Cambridge University Press.
- -----, 2002, "Hierarchy, Circularity, and Double Reduction," in Contours of Agency, Buss and Overton (eds.), 65-85.
- Buss, S., 1994, "Autonomy Reconsidered," Midwest Studies 9:95-121.
- -----, 1997, "Weakness of Will," *Pacific Philosophical Quarterly* 78, no. 1: 13-44.
- Buss, S. and H. Overton, (eds.), 2002, *Contours of Agency: Essays on Themes from Harry Frankfurt*, Cambridge, Mass.: MIT Press.
- Chisholm, R., 1966, "Freedom and Action," in *Freedom and Determinism*, K. Lehrer (ed.), New York: Random House, 11-44.
- -----, 1971, "Reflections on Human Agency," Idealistic Studies 1: 36-46
- -----, 1976a, *Person and Object*, La Salle, Ill.: Open Court Press.
- -----, 1976b, "The Agent as Cause," in *Action Theory*, M. Brand and D. Walton (eds.), Dordrecht: D. Reidel Publishing Co.
- -----, 1982a, "Replies," in *Roderick M. Chisholm*, R.J. Bogdan (ed.), 1-16.

- -----, 1982b, “Human Freedom and the Self,” in *Free Will*, G. Watson (ed.), 24-45.
- -----, 1995, “Agents, Causes, and Events: The Problem of Free Will,” in *Agents, Causes, and Events*, T. O'Connor (ed.), New York: Oxford University Press, 95-100.
- Christman, J., 1988, “Constructing the Inner Citadel: Recent Work on the Concept of Autonomy,” *Ethics* 99: 109-124.
- -----, (ed.), 1989, *The Inner Citadel: Essays on Individual Autonomy*, New York: Oxford University Press.
- -----, 1991, “Autonomy and Personal History,” *Canadian Journal of Philosophy* 21: 1-24.
- -----, 1993, “Defending Historical Autonomy: A Reply to Professor Mele,” *Canadian Journal of Philosophy* 23: 281-90.
- Clarke, R., 1993, “Toward a Credible Agent-Causal Account of Free Will,” *Nous* 27: 191-203, reprinted in *Agents, Causes and Events*, T. O'Connor (ed.), 201-15.
- -----, 1996, “Agent-Causation and Event-Causation in the Production of Free Action,” *Philosophical Topics* 24: 19-48.
- Davidson, D., 1980, *Essays on Actions and Events*, Oxford: Clarendon Press.
- Dennett, D., 1984, *Elbow Room: The Varieties of Free Will Worth Wanting*, Cambridge, Mass.: MIT Press.
- -----, “On Giving Libertarians What They Say They Want,” in *Agents, Causes, and Events*, T. O'Connor (ed.), 43-56.
- Double, R., 1991, *The Non-Reality of Free Will*, Oxford: Oxford University Press.
- Dworkin, G., 1970, “Acting Freely,” *Nous* 4: 367-83.
- -----, 1976, “Autonomy and Behavior Control,” *Hastings Center Report* 6: 23-28.
- -----, 1988, *The Theory and Practice of Autonomy*, New York: Cambridge University Press.
- Ekstrom, L., 1993, “A Coherence Theory of Autonomy,” *Philosophy and Phenomenological Research* 53: 599-616.
- Feinberg, J., 1970a, “Causing Voluntary Actions,” in *Doing and Deserving*, Feinberg, Princeton: Princeton University Press, 152-86.
- -----, 1970b, “What Is So Special about Mental Illness?” in *Doing and Deserving*, Feinberg, Princeton: Princeton University Press, 272-92.
- Fischer, J.M., 1982, “Responsibility and Control,” *Journal of Philosophy* 79: 24-40, reprinted in *Moral Responsibility*, Fischer (ed.), 174-90.
- -----, (ed.), 1986, *Moral Responsibility*, Ithaca: Cornell University Press.
- -----, “Responsiveness and Moral Responsibility,” in *Responsibility, Character and the Emotions*, F. Schoeman (ed.), 88-106.
- -----, 1994, *The Metaphysics of Free Will: An Essay on Control*, Cambridge: Cambridge University Press.
- -----, 1999, “Recent Work on Moral Responsibility,” *Ethics* 110: 93-139
- -----, 2002, “Frankfurt-Style Compatibilism,” in *Contours of Agency*, Buss and Overton (eds.), 1-26.
- Fischer, J.M. and M. Ravizza, 1993, *Perspectives on Moral Responsibility*, Ithaca: Cornell University Press.
- -----, 1998, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

- Foot, Philippa, 1957, “Free Will as Involving Determinism,” *The Philosophical Review*, 66: 439-450, reprinted in *Free Will and Determinism*, B. Berofsky (ed.), 95-108.
- Frankfurt, H., 1988a, *The Importance of What We Care About*, Cambridge: Cambridge University Press.
- -----, 1988b, “Alternate Possibilities and Moral Responsibility,” in *The Importance of What We Care About*, Frankfurt, 1-10.
- -----, 1988c, “Freedom of the Will and the Concept of a Person,” in The Importance of What We Care About, Frankfurt, 11-25
- -----, 1988d, “Three Concepts of Free Action: II,” in *The Importance of What We Care About*, Frankfurt, 47-57.
- -----, 1988e, “Identification and Externality,” in *The Importance of What We Care About*, Frankfurt, 30-94.
- -----, 1988f, “Identification and Wholeheartedness,” in *The Importance of What We Care About*, Frankfurt, 159-76.
- -----, 1988g, “Rationality and the Unthinkable,” in *The Importance of What We Care About*, Frankfurt, 177-90.
- -----, 1999a, *Necessity, Volition and Love*, Cambridge: Cambridge University Press.
- -----, 1999b, “The Faintest Passion,” in *Necessity, Volition, and Love*, Frankfurt, 95-107.
- -----, 1999c, “Autonomy, Necessity, and Love,” in *Necessity, Volition, and Love*, Frankfurt, 129-41.
- -----, 1999d, “On Caring,” in *Necessity, Volition, and Love*, Frankfurt, 155-80.
- -----, 2002a, “Reply to J. David Velleman,” in *Contours of Agency*, Buss and Overton (eds.), 124-28.
- -----, 2002b, “Reply to Gary Watson,” in *Contours of Agency*, Buss and Overton (eds.), 160-64.
- -----, 2002c, “Reply to T.M. Scanlon,” in *Contours of Agency*, Buss and Overton (eds.), 184-88.
- -----, 2002d, “Reply to Richard Moran,” in *Contours of Agency*, Buss and Overton (eds.), 218-25.
- -----, 2002e, “Reply to Susan Wolf,” in Contours of Agency, Buss and Overton (eds.), 245-53.
- -----, 2002f, “Reply to Barbara Herman,” in *Contours of Agency*, Buss and Overton (eds.), 275-78.
- -----, 2002g, “Reply to Jonathan Lear,” in *Contours of Agency*, Buss and Overton (eds.), 293-97.
- -----, 2002h, “Reply to Susan Wolf,” in *Contours of Agency*, Buss and Overton (eds.), 245-52.
- Friedman, M., 1986, “Autonomy and the Split-Level Self,” *Southern Journal of Philosophy* 24: 19-35.
- Gert, B. and T. Duggan, 1979, “Free Will as the Ability to Will,” Nous 13: 197-217, reprinted in *Moral Responsibility*, Fischer (ed.), 205-224.
- Ginet, C., 1996, “Might We Have No Choice?” in *Freedom and Determinism*, K. Lehrer (ed.), 87-104.
- -----, 1990, *On Action*, Cambridge: Cambridge University Press.
- -----, “In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument convincing”.
- Greenspan, P., 1978, “Behavior Control and Freedom of Action,” *Philosophical Review*: 87: 225-40, reprinted in *Moral Responsibility*, Fischer (ed.), 191-204.
- -----, 1999, “Impulse and Self-reflection: Frankfurtian Responsibility versus Free Will”, in *Philosophical Topics* 3, no. 4: 325-40.

- Hampshire, S, 1983, *Thought and Action*, Notre Dame, Ind.: University of Notre Dame Press.
- Hobart, R.E., 1934, "Free Will as Involving Determinism and Inconceivable without It," *Mind* 43: 1-27, reprinted in *Free Will and Determinism*, B. Berofksy (ed.), 63-94.
- Honderich, T., (ed.), 1973, *Essays on Freedom of Action*, London: Rutledge & Kegan Paul.
- -----, 1988, *The Consequences of Determinism*, Oxford: Clarendon Press.
- Hume, D., 1955, *An Inquiry Concerning Human Understanding*, Indianapolis: Bobbs-Merrill.
- Kane, R., 1985, *Free Will and Values*, Albany: State University of New York Press.
- -----, 1996, *The Significance of Free Will*, New York: Oxford University Press.
- Kant, I., 1964, *The Groundwork of the Metaphysics of Morals*, H.J. Paton (trans.), New York: Harper & Row Publishers.
- -----, 1956, *Critique of Practical Reason*, trans. by L.W. Beck, Indianapolis: Bobbs-Merrill.
- -----, 1960, *Religion Within the Limits of Reason Alone*, T. M. Greene and H. H. Hudson (trans.), New York: Harper & Row, Publishers.
- Korsgaard, C., 1996, "Morality as Freedom," in *Creating the Kingdom of Ends*, Korsgaard, Cambridge: Cambridge University Press.
- -----, 1996, *The Sources of Normativity*, Cambridge: Cambridge University Press.
- Lehrer, K., (ed.), 1966, *Freedom and Determinism*, New York: Random House.
- -----, 1997, *Self-Trust*, Oxford; Oxford University Press.
- Lewis, D., 1981, "Are We Free to Break the Laws?" Theoria 47: 113-21.
- Locke, D., 1975, "Three Concepts of Free Action: I," *Proceedings of the Aristotelian Society*, suppl. vol. 49: 95-112.
- MacKay, D.M., 1960, "On the Logical Indeterminiacy of a Free Choice," *Mind* 69: 31-46.
- -----, 1971, "Choice in a Mechanistic Universe: A Reply to Some Critics," *British Journal for the Philosophy of Science* 22: 275-85.
- -----, 1973, "The Logical Indeterminateness of Human Choice," *British Journal for the Philosophy of Science* 24: 405-408.
- Malcolm, N., 1968, "The Conceivability of Mechanism," *The Philosophical Review* 77: 45-72, reprinted in *Free Will*, G. Watson (ed.), 127-49.
- Mele, A., 1993, "History and Personal Autonomy," *Canadian Journal of Philosophy* 23: 271-80
- -----, 1995, *Autonomous Agents: From Self-Control to Autonomy*, New York: Oxford University Press.
- -----, "Soft Libertarianism and Frankfurt-style Scenarios" .
- Meyers, D., 1987, "Personal Autonomy and the Paradox of Feminine Socialization," *Journal of Philosophy* 84: 619-28.
- Moran, R., 2002, "Frankfurt on Identification: Ambiguities of Activity in Mental Life," in *Contours of Agency*, Buss and Overton (eds.), 189--217.
- Morgenbesser, S. and J. Walsh, (eds.), 1962, *Free Will*, Englewood Cliffs, N.J.: Prentice-Hall.
- Nagel, T., 1995, "The Problem of Autonomy," reprinted from *The View from Nowhere*, Nagel, New York: Oxford University Press, 1986, in Agents, Causes, and Events, T. O'Connor (ed.), 33-42.
- Neely, W., 1974, "Freedom and Desire," *Philosophical Review* 83: 32-54.
- Nozick, R., 1981, *Philosophical Explanations*, Cambridge, Mass.: Harvard University Press.
- O'Connor, T., 1993, "Indeterminism and Free Agency: Three Recent Views," *Philosophy and*

Phenomenlogical Research 53: 499-525.

- -----, (ed.), 1995a, *Agents, Causes and Events: Essays on Indeterminism and Free Will*, New York: Oxford University Press.
- -----, 1995b, “Agent Causation,” in *Agents, Causes, and Events*, T. O'Connor (ed.), 173-200.
- -----, “Why Agent Causation?” *Philosphical Topics* 24: 143-51.
- Plato, *Protagoras*, W.K. Guthrie (trans.), in *Collected Dialogues*, E. Hamilton and H. Cairns (eds.), 308-52.
- Reid, T., 1969, *Essays on the Active Powers of the Human Mind*, Cambridge, Mass.: MIT Press.
- Richardson, H., 2001, “Autonomy”s Many Normative Presuppositions,” *American Philosophical Quarterly* 38: 287-303.
- Rowe, W., 1987, “Two Concepts of Freedom,” *The Proceedings and Addresses of the American Philosophical Association* 61: 43-64, reprinted in *Agents, Causes, and Events*, T. O'Connor (ed.), 151-72.
- -----, 1991, “Responsibility, Agent-Causation, and Freedom: An Eighteenth-Century View,” *Ethics* 101: 237-57, reprinted in *Perspectives on Moral Responsibility*, Fischer and Ravizza, (eds.), 263.
- Sartre, J.-P., 1956, *Being and Nothingness*, H. Barnes, (trans.), New York: Simon & Schuster.
- Schatz, D., 1985, “Free Will and the Structure of Motivation,” in *Midwest Studies in Philosophy 10: Studies in the Philosophy of Mind*, P. French, T. Uehling, Jr., and H. Weinstein (eds.), Minneapolis: University of Minnesota Press, 451-82.
- Schoeman, F., (ed.), 1987, *Responsibility, Character and the Emotions: New Essays in Moral Philosophy*, F. Schoeman (ed.), Cambridge: Cambridge University Press.
- Slote, M., 1980, “Understanding Free Will,” *Journal of Philosophy* 77: 136-51, reprinted in *Moral Responsibility*, Fischer (ed.), 124-39.
- -----, 1982, “Selective Necessity and the Free-Will Problem,” *Journal of Philosophy* 79: 5-24.
- Spinoza, B. de, 1985, *Ethics*, E. Curley, (trans.), Princeton: Princeton University Press.
- Strawson, G., 1986, *Freedom and Belief*, Oxford: Clarendon Press.
- Stump, E. “Sanctification, Hardening of the Heart, and Frankfurt's Concept of Free Will,” in *Perspecitves on Moral Responsibility*, Fischer and Ravizza, (eds.), 211-236.
- Taylor, C, 1982, “Responsibility for Self,” in *Free Will*, Watson (ed.), 111-26.
- -----, “What”s Wrong with Negative Liberty,” in *The Idea of Freedom*, Alan Ryan (ed.), 175-93.
- Taylor, R., 1966, *Action and Purpose*, Englewood Cliffs, N.J.: Prentice-Hall.
- -----, 1982, “Agent and Patient,” *Erkenntnis* 18: 111-26.
- Thalberg, I., 1978, “Hierarchical Analyses of Unfree Action,” *Canadian Journal of Philosophy* 8: 211-26.
- Van Inwagen, P., 1982, “The Incompatibility of Free Will and Determinism,” in *Free Will*, G. Watson (ed.), 46-58.
- -----, 1983, *An Essay on Free Will*, Oxford: Clarendon Press.
- Velleman, J. D., 1989a, “Epistemic Freedom,” *Pacific Philosophical Quarterly* 70: 73-97.
- -----, 1989b, *Practical Reflection*, Princeton: Princeton University Press.
- -----, 2000, *The Possibility of Practical Reason*, Oxford: Clarendon Press.
- -----, “What Happens When Someone Acts?” in *The Possibility of Practical Reason*, 123-43.
- -----, “Identification and Identity,” in *Contours of Agency*, Buss and Overton (eds.), 91-123.

- Wallace, R.J., 1994, *Responsibility and the Moral Sentiments*, Cambridge, Mass.: Harvard University Press.
- Watson, G., 1975, “Free Agency,” *Journal of Philosophy* 72: 205-20, reprinted in *Free Will*, Watson (ed.), 96-110.
- -----, 1977, “Skepticism about Weakness of Will,” *Philosophical Review* 85: 316-39.
- -----, (ed.), 1982, *Free Will*, Oxford: Oxford University Press.
- -----, 1987, “Free Action and Free Will,” *Mind* 96: 147-72.
- -----, 1993, “Responsibility and the Limits of Evil: Variations on a Strawsonian Theme,” in *Perspectives on Moral Responsibility*, Fischer and Ravizza (eds.), 119-50.
- -----, “Volitional Necessity,” in *Contours of Agency*, Buss and Overton (eds.), 129-59.
- Wiggins, D., 1973, “Towards a Reasonable Libertarianism,” in *Essays on Freedom of Action*, T. Honderich (ed.), 31-62.
- Williams, B., 1993, *Shame and Necessity*, Berkeley, California: University of California Press.
- Wilson, G., 1989, *The Intentionality of Human Action*, Stanford: Stanford University Press.
- Wolf, S., 1980, “Asymmetrical Freedom.” *Journal of Philosophy* 77: 151-66, reprinted in *Moral Responsibility*, Fischer (ed.), 225-40.
- -----, “The Importance of Free Will,” *Mind* 90: 386-405, reprinted in *Perspectives on Moral Responsibility*, Fischer and Ravizza (eds.), 101-118.
- -----, “Sanity and the Metaphysics of Responsibility,” in *Responsibility, Character, and the Emotions*, 46-62.
- -----, 1990, *Freedom within Reason*, New York: Oxford University Press.
- Yaffe, G., “Manipulation”.
- Young, Robert, 1986, *Personal Autonomy: Beyond Negative and Positive Liberty*, London: Croom Helm Ltd.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

action | addiction | agent causation | Aristotle | authority | coercion | compatibilism | Compulsion | determinism, causal | free action/freedom of action | free will/freedom of the will | incompatibilism | Kant, Immanuel | manipulation | moral responsibility | motivation | Plato | practical rationality/practical reason | reason | Reid, Thomas | Schopenhauer, Arthur | self | values | volition/will | weakness of will

Sarah Buss
sarah-buss@uiowa.edu

Table of Contents

First published: May 28, 2002
Content last modified: May 28, 2002

Stanford Encyclopedia of Philosophy
Notes to Personal Autonomy

Notes

1. For three different versions of this approach, see Frankfurt 1988c, Watson 1975, and Bratman 1979. In Frankfurt's early work, the relevant attitude appears to be a higher-order desire. Later, however, he speaks of a distinct, irreducible attitude of "satisfaction." In contrast, Watson's account singles out the evaluative judgment that one's action makes sense in terms of one's values. And Bratman stresses the extent to which the motives of self-governors conform to their policies. (Note that, on Watson's account, the "attitude" that distinguishes autonomous, morally responsible agents is only by implication an attitude they take toward their motives.) Whereas Watson appeals to Plato's philosophy of mind, Frankfurt has deep affinities with Hume. In stressing the importance of the autonomous agent's attitude toward her motives, he also takes his inspiration from Spinoza.

2. Bernard Berofsky calls attention to the fact that there is support "across a broad spectrum" of theorizing in psychology and psychiatry, for the idea that certain inner states undermine autonomy by creating "barriers to objectivity." They do this, he explains, by "persisting in a way that removes them from experiential review and the influence of newly acquired information." (Berofsky 1995, p. 199) For two more examples of recent philosophical accounts that stress the importance of the autonomous agent's responsiveness to reasons, see Wolf, 1990 and Fischer and Ravizza,1993. These philosophers argue that responsiveness to reasons is a necessary condition of moral responsibility. But in presenting and defending their views, they suggest that unless an agent satisfies this condition, she does not really govern herself. As Wolf puts it, they are interested in the "relation to one's will which is necessary in order for one's actions … to be 'up to oneself' in the way that is necessary for responsibility."(4) (Note that Wolf reserves the term 'autonomy' for one particular conception of this self-relation: the view that an agent's control over her behavior must be "ultimate" — that "her will must be determined by her self, and her self must not, in turn, be determined by anything external to itself"(10).) According to Wolf, responsiveness to reasons is a necessary supplement to a coherentist condition: "a person's status as a responsible agent rests not only on her ability to make her behavior conform to her deepest values but also on her ability to form, assess, and revise those values on the basis of a recognition and appreciation of … the True and the Good." In stressing that an autonomous agent "must be in a position that allows her reasons to be governed by what reasons there are… [i.e.,] by what is valuable and worthless" (117-118), Wolf evokes a tradition that goes back to Plato. But she rejects the Platonic conception of values as "things" that can be "apprehended by some special faculty"(123). Fischer and Ravizza likewise try to steer clear of controversial metaethical assumptions. "Regular reasons-receptivity" is, they argue, essential to having "guidance control" over one's action. "It involves a pattern of actual and hypothetical recognition of reasons (some of which are moral reasons) that is understandable by some appropriate external observer. And the pattern must be at least minimally grounded in reality."(90) (Unlike Wolf, Fischer and Ravizza do not think that reasons responsiveness suffices for moral responsibility. They

believe that in order for an agent to "own" the reasons-responsive mechanism that produces his action, he must "take responsibility" for it. To do this, he must "see himself as an agent" and "accept that he is a fair target of the reactive attitudes as a result of how he exercises this agency in certain contexts"(210-211).

3. The importance of acquiring motives in a way that is responsive to one's own reasoning is taken for granted by most writers on autonomy—though it is widely acknowledged that agents need not actively deliberate prior to every autonomous action. On this view autonomy is "achieved when the individual subjects the norms with which he or she is confronted to critical evaluation and then proceeds to reach practical decisions by way of independent and rational reflection"(Young 1986). In other words, one's conduct is autonomous only if one exercises "assorted introspective, imaginative, reasoning, and volitional skills."(Meyers 1987) This procedural requirement of this kind is often grafted onto coherentist accounts. Thus, in his early work, Gerald Dworkin argues that someone who identifies with his motives lacks autonomy if this identification reflects the fact that he has "been influenced in decisive ways by others in such a fashion that we are not prepared to think of it as his own choice"(Dworkin 1976, 25,) (In his more recent work, Dworkin suggests that an autonomous agent need not actually identify with her motives as long as she is capable of altering her preferences in light of her uncompelled reflection.) (Dworkin 1988) For some examples of recent attempts to work out the details of such a condition, see Mele 1995 and Christman 1991. Mele argues that an autonomous agent must be capable of reflecting critically upon her desires, and of altering them in light of this reflection. Her beliefs must be "conducive to informed deliberation"; and she must be a "reliable deliberator"(**chap. 10). Similarly, Christman stresses the importance of the autonomous agent's ability to reflect (in a "minimally rational" way) on the process whereby she acquired a given desire. He argues that someone acts autonomously when she is moved by a given desire, only if she would not reject the desire if she reflected on its genesis. [**others: Haworth, D. Meyers**]

4. As Dworkin puts it, what is necessary for autonomy is "some ability both to alter one's preferences and to make them effective in one's actions" (Dworkin 1988).

5. For a powerful defense of the incompatibilist position, see van Inwagen 1983. Much of the debate over the relationship between causal determinism, on the one hand, and autonomy, free will, and moral responsibility, on the other, consists of attempts to challenge and defend the modal argument that is at the heart of this defense. This argument spells out the widespread intuition that someone cannot govern her own action if she could not have done anything to refrain from performing it. Frankfurt's alleged counterexamples to this "Principle of Alternative Possibilities" have also been the focus of intense philosophical debate. (See Frankfurt 1988b) For a thorough discussion of the modal argument for incompatibilism, see Fischer 1994. For a recent attempt to work out a rigorous incompatibilist conception of autonomy, see Kane1996. According to Kane, the desire to be an autonomous agent is the desire to have "the power to be the ultimate producers of [one's] own ends… the power to make choices which can only and finally be explained in terms of [one's] own [will] (i.e., character, motives, and efforts of will)." "No one," he argues, "can have this power in a determined world."(254)

6. Attempts to make sense of weakness of will go back to Plato and Aristotle (see *). For some more recent discussions, see Donald Davidson 1980, Gary Watson 1977, Michael Bratman 1979, Alfred Mele 1995 and Sarah Buss 1997.

7. The example is Frankfurt's (Frankfurt 2002b). The mother, he says, may be “glad to be putting her need for the relationship above what is best by a measure that she now refuses to regard as decisive”(163). More generally, Frankfurt argues that “the fact that something is important to someone is a circumstance that naturally has its causes, but it may neither originate in, nor be at all supported by, reasons. It may simply be a brute fact, which is not derived from any assessment or appreciation whatever”(161). “Suppose,” Frankfurt elsewhere writes, “I were to conclude for some reason that it is not desirable for me to seek the well-being of my children. I suspect that I would continue to love them and to care about their well-being anyhow. This discrepancy between my judgment and my desire would not show that I had become alienated from the desire.”(Frankfurt 2002, 223) For a thorough discussion of Frankfurt's position, see Watson 2002.

8. Thomas Reid is an early champion of this approach. More recently, agent-causation theories have been defended by Richard Taylor and Roderick Chisholm; and more recently still, by Randolph Clarke and Timothy O'Connor.

9. See, for example, Korsgaard 1996 and Bok 1998. As Bok explains,

> When I act for reasons, the events that cause me to act as I do might be external to me, but the reasons that I regard as determining what I do cannot be. For while, *qua* event, my acceptance of some reason for action might or might not ultimately be caused by something outside myself, in regarding it as a reason for action I must regard it as having a justification that is independent of those causes. This justification might at various points appeal to theoretical claims. But I cannot regard it, *qua* justification, as having been produced or foisted on me by any natural event. When I consider it as a justification, I consider not its causal origins but its rational grounds; and I accept or reject it on that basis. When I explain what leads me to accept it, I will adduce not the causes that led me to do so, but the reasons that convinced me that it was sound. Because any reasons I adduce must themselves be reasons I accept, this type of explanation will not ultimately lead me to adduce determining factors that I do not regard as my own.(206)

For essentially the same point in a less Kantian context, see D.M. McKay 1960 & 1973, and Hampshire 1983. Like others, they point out that the question “How will *A* act?” has no determinate answer for *A* until she decides how to act. What is a simple fact from the perspective of a third-person observer is not a fact from the perspective of the agent herself. Similarly, J. David Velleman argues that the freedom that counts where autonomous agency is concerned is epistemic freedom with respect to one's alternatives (Velleman 2000). Intentions, he claims, are a special sort of belief that the agent has the power to make true.

> Our expectation of doing something embodies an invention rather than a discovery. For we can simply adopt the expectation that we're going to do any one of the things for which we have some antecedent motives, and this expectation will modify the balance of forces so as to make itself true. We are thus in a position to make up our forthcoming behavior. Making up what we will do is, in fact, our way of making up our minds to do it.(24)

This essential formal feature of the practical point of view is at the center of most existentialist conceptions of human agency. Thus, for example, Jean-Paul Sartre claims that “[M]otives are only for consciousness. And due to the very fact that the motive can arise only as appearance, it constitutes itself as ineffective. … [C]onsciousness is not subject to it because of the very fact that consiciousness posits it; for consciousness has now the task of conferring on the motive its meaning and its importance.”(Sartre 1956, 71)

Sarah Buss
sarah-buss@uiowa.edu

First published: May 28, 2002
Content last modified: May 28, 2002

Stanford Encyclopedia of Philosophy

Portrait courtesy of the National Trust for Scotland

Thomas Reid

Thomas Reid (1710-1796) is a Scottish philosopher and one of the founders of the "common sense" school of philosophy. (Also included in the common sense school are such less well-known 18th century philosophers as James Beattie, George Campbell, and Dugald Stewart.) Reid is best known for his epistemology of sensation -- he believes that sensations serve to make us directly aware of real objects without the aid of any intervening medium -- and for his view of free will -- he holds that the only free actions are those that come about through a causal process originated by the agent. In the explication of both he offers perceptive and important criticisms of the philosophy of Locke, Berkeley and especially Hume. He is also well known for his criticisms of Locke's view of personal identity and Hume's view of causation. However, Reid also wrote on a wide variety of other philosophical topics including ethics, aesthetics and various topics in the philosophy of mind.

After studying at the University of Aberdeen, Reid entered the ministry in New Machar in 1737. In 1748 he published a short essay entitled "An Essay on Quantity" which concerned Hutchison's *Inquiry into the Origin of Our Ideas of Beauty and Virtue*. Although this was his only published work, he was given a professorship at King's College Aberdeen in 1752. There he wrote *An Inquiry Into the Human Mind on the Principles of Common Sense* (published in 1764). Shortly afterward he was given a much more prestigious professorship at the University of Glasgow. He resigned from this position in 1781 in order to give himself greater time to write, and published *Essays on the Intellectual Powers of Man* in 1785 and *Essays on the Active Powers of Man* in 1788.

In the *Inquiry*, which is primarily a work in epistemology, Reid examines each of the five senses and discusses the ways in which we achieve knowledge of the world through employing them. Much of the

view developed in the *Inquiry* reappears in the *Essays on the Intellectual Powers*, which expands his epistemological picture beyond the apprehension of the world through the senses to consideration of memory, imagination, knowledge concerning kinds of things, the nature of judgment, reasoning and taste. The *Essays on the Active Powers* examines a collection of topics concerning ethics, the nature of agency generally, and the distinctive features of human agency.

Common Sense and Ordinary Language

Reid is a staunch defender of "common sense", or, as he sometimes puts it, the opinions of "the vulgar". In fact, in almost every arena of philosophical inquiry, Reid's positions are in various ways tied up with his overall project of defending common sense. Common sense, for Reid, are those tenets that we cannot help but believe, given that we are constructed the way we are constructed. This is not to say that nobody fails to believe the dictates of common sense. People often have beliefs that are in manifest conflict with common sense, but to have such beliefs, Reid thinks, is to be in deep conflict with one's nature as a human being. What this suggests is that Reid takes on a burden of showing, each time that he claims some particular view to be a dictate of common sense, that belief in it is dictated by human nature. In addition, much of Reid's philosophy proceeds from the assumption that the dictates of common sense could not possibly conflict with one another. Human nature could not be such as to lead us to contradictory beliefs. The collection of tenets that make up common sense are consistent with one another, and non-optional to those who are believing as a human being does.

He sees a close tie between the dictates of common sense and distinctions and positions that could be found buried in the structure of ordinary language. Ordinary language, for Reid, is the mirror of our ordinary, everyday thought. (The connection between ordinary language and common sense that Reid espouses was of great influence on much later philosophers such as G. E. Moore and J. L. Austin.) Reid

does not believe, however, that every feature of ordinary language is indicative of some important tenet of common sense. Reid often suggests that the relevant features are those that can be found in "the structure of all languages", suggesting that the linguistic features of relevance are features of syntactic structure shared among languages. For example, Reid repeatedly says that it is a dictate of common sense that there is some important difference between the active and the passive, since "all languages" have a passive and active voice. The mere fact, then, that we say something in ordinary language does not imply that, for Reid, it is to be taken as a tenet of common sense. Rather, syntactic structures shared among languages often indicate, Reid thinks, features of our common sense conception of the world. When a certain tenet is implied by some feature shared by "all languages", Reid thinks it very likely that the best explanation for its being so shared is that it is a dictate of common sense.

Reid's attachment to common sense as found in ordinary language is easy to caricature unfairly. Reid did not hold that every position that can be deduced from a linguistic feature shared by all languages is a dictate of common sense, something that we cannot help but believe if we are to be true to our natures. He says, for instance,

> A philosopher is, no doubt, entitled to examine even those distinctions that are to be found in the structure of all languages; and, if he is able to shew that there is no foundation for them in the nature of the things distinguished; if he can point out some predjudice common to mankind which has led them to distinguish things which are not really different; in that case, such a distinction may be imputed to a vulgar error, which ought to be corrected in philosophy. But when, in the first setting out, he takes it for granted without proof, that distinctions found in the structure of all languages, have no foundation in nature; this surely is too fastidious a way of treating the common sense of mankind. (*Essays on the Intellectual Powers of Man*, p.14)

However, he does hold that the burden of proof is squarely on the shoulders of those who deny something suggested by a syntactic feature shared across languages. If it is possible to find a claim in another philosopher's argument which sits uneasily with the facts about ordinary language, and the philosopher has failed to show that the claim suggested by ordinary language rests on a deep error, then, on Reid's view, we have found a sufficient reason to reject the philosopher's argument.

What this means is that Reid is not concerned to answer certain questions of justification that can seem enormously pressing to us in certain philosophical moods. He is not, for instance, interested in providing a justification for our belief in the external world by appeal to first principles of some sort. For instance, Reid feels he can refute skeptical hypotheses -- such as Descartes's hypothesis of an evil demon who makes us believe that the world is the way we take it to be when it is really vastly different -- simply by showing that such a hypothesis is no more likely to be true than the common-sensical belief that the world is much the way we perceive it to be. Since the belief in the external world is a dictate of common sense, it is, Reid thinks, as justified as it needs to be when it is shown to be on the same footing as any alternative. Justification, therefore, does not necessarily require providing positive reasons in favor of common-sensical beliefs; common sense beliefs could be adequately justified simply by undermining the force of the reasons in favor of alternatives to common sense. Common sense, as found in the structure of

ordinary language, then, constrains, rather than dictates, acceptable philosophical positions.

Epistemology

Attacking the Way of Ideas

One of Reid's most important critical contributions is his attack on the model of the mental offered by Locke (although Locke derived it in large part from Descartes) and accepted in broad outline by Berkeley and Hume. As Reid reads Locke, all thought-like mental attitudes -- as opposed to emotions, desires, choices and the like -- whether they be thoughts concerning real objects, imaginative reveries, or even sensations, can be analyzed as "perceptions" of ideas. "Perception" here is used in a somewhat technical sense not to describe any of the the exercises of the five senses -- nor even to describe all that is going on when we are aware of external objects -- but, instead, to describe a special mental operation, a kind of mental "vision", so to speak. On the Lockean model, for instance, to think about an apple is to perceive an idea of an apple; the idea is a distinctive mental object grasped or perceived through the exercise of one's mental "eyes". It is through the perceiving of an idea -- in this technical sense of "perceive" -- that we perceive -- in a more ordinary sense of the term -- objects that correspond to those ideas; we are aware of the apple by perceiving an idea of it. Berkeley, and especially Hume, revise and expand this picture in various ways, but they maintain the overall schema: mental phenomena are perceptions of mental objects.

From within the Lockean model, there are at least three things that might be said regarding the sense in which we are aware of external objects. The Lockean might say, (1) that we are aware of external objects directly *by* perceiving representations of them, ideas; (2) that we *infer* the existence and nature of external objects by perceiving ideas of them, the nature of which we grasp directly; or, (3) that there is no distinction between external objects and ideas, and, thus, when we perceive ideas we are perceiving external objects. On the first approach, the Lockean would need to offer an explanation of what it is that is so special about ideas that makes it the case that whenever we are perceiving one we are directly aware of that which it represents. If this burden can be discharged, then it is possible to say that we are aware of external objects and remain consistent with the Lockean conception of the mental. Reid, however, thinks that the history of philosophy from the Ancients through his own time has been marked by a series of failed efforts to explain how it is that perception of something in the mind could amount, automatically, to perception of some external object. All such attempts make the mistake, he thinks, of giving unexplained and unexplainable powers to the internal mental objects: how could they possibly, all by themselves, attach our minds to objects whenever they are perceived? Since this question cannot be answered, he thinks, we are left with the second and third alternatives: either we infer the nature of external objects from the features of directly perceived ideas, or else Berkeleyan Idealism is true: external objects are ideas.

Reid thinks that nobody who has absorbed Hume's lessons regarding causation would think that we can avoid skepticism about the external world while insisting that we infer its nature from the features of directly perceived ideas. After all, on the Lockean model, external objects are the causes of our ideas. But if Hume is right about causation, then we can only infer the nature of a particular unobserved cause of a

particular observed effect when we have had repeated experience of conjunction of similar causes with similar effects. Therefore, we can infer nothing about external objects by examination of the ideas which they cause in us: we have never had any experience of the relevant causes, but only experience of the relevant effects. Thus, Reid thinks, Lockeans about the mental are committed either to outright skepticism or Berkeleyan Idealism.

While his desire to overthrow skepticism and idealism -- both of which he takes to be in violation of common sense -- is part of Reid's motivation for attacking the Lockean model of the mental, he is careful not to attack the model merely on the grounds that it is in violation of common sense. In fact, Reid offers a range of criticisms that could be accepted quite readily even by those who deny that philosophical theses ought to be constrained by common sense. He says, for instance,

> When, therefore, in common language, we speak of having an idea of anything, we mean no more by that expression, but thinking of it. The vulgar allow that this expression implies a mind that thinks, an act of that mind which we call thinking, and an object about which we think. But, besides these three, the philosopher conceives that there is a fourth--to wit, the *idea*, which is the immediate object. The idea is in the mind itself, and can have no existence but in a mind that thinks; but the remote or mediate object may be something external, as the sun or moon; it may be something past or future; it may be something which never existed. This is the philosophical meaning of the word *idea*; and we may observe that this meaning of that word is built upon a philosophical opinion: for, if philosophers had not believed that there are such immediate objects of all our thoughts in the mind, they would never have used the word idea to express them. (*Essays on the Intellectual Powers of Man*, p. 20)

Reid can be taken here to be imagining a kind of conversation between "the vulgar" on the one hand and " the philosopher" on the other. The vulgar say, "When I think about an apple in front of me, for instance, the immediate object of my perception is the real apple." The philosopher responds, "No, the immediate object of your perception is a mental object, an idea of the apple." But, Reid points out, the philosopher's response, which seems to be instructing the vulgar on their mistake, is actually predicated on a prior rejection of the vulgar's conception of everyday thoughts about objects. The Lockean claim that there is a "fourth" element in every thought -- an idea -- is "built upon a philosophical opinion", that is, the model arises from rejection of the common sense view and thus can't be given as a reason to reject the "vulgar" position. What Reid is doing here is shifting the burden of proof on to those who hold the Lockean model; there is nothing inherently superior about the Lockean picture, and hence the common sense picture -- under which we are aware of real objects directly and without mediation -- is, as yet, no less well defended.

In one of Reid's more powerful arguments against the Lockean model, he points out that the Lockean model is supposed to explain the fact that our mental states manage to connect to real objects, manage to be *about* real objects. However, this fact about beliefs is only explained by the model if the model is less obscure than the phenomena to be explained by it. But, Reid points out, if we start by noticing that we don't understand how it is that we manage to connect our minds to objects in the world, it can't help to say

that we do it by first connecting our minds to mental objects (ideas) unless we understand how it is that we manage to connect our minds to those mental objects. That is, why is mental perception, or awareness, of ideas thought to be any more intelligible than awareness of objects? If it is no more intelligible, then the Lockean model is not serving to explain what it was intended to explain. (One place in which this objection appears is at *Essays on the Intellectual Powers of Man*, p. 229.)

Reid also uses the distinction between real and apparent magnitude developed in Berkeley's *New Theory of Vision* to respond to one of Hume's arguments for the claim that the immediate objects of awareness are mental objects rather than real objects. (See *Essays on the Intellectual Powers of Man*, pp. 224-225.) As Reid reads him, Hume argues that since, for instance, objects get smaller in our visual field as we move away from them, and real objects don't change size merely as a result of the fact that we move away from them, we must not be directly aware of the real sizes of objects. Reid claims that Hume is equivocating on two different notions of magnitude. The *apparent* magnitude of the object is the size that the object appears to have when looked at from a certain place. Apparent magnitude is a relational property of objects: the apparent magnitude of an object is a function not just of intrinsic features of the object, but also of the location of a particular observer. Real magnitude, on the other hand, is an intrinsic property of objects, not dependent on the position of any observer. So, when we move away from an object, we are perceiving properties of that object: its apparent magnitude relative to the locations through which we pass. Thus, Reid concludes, the fact that the real magnitude of the object doesn't change as we move away from the object is irrelevant to the question of whether or not the immediate objects of awareness are mental items, like ideas, or qualities of objects. Just because we aren't perceiving the real magnitude of the object as we move away from it doesn't mean that we aren't perceiving a propery of the object itself. We *are* perceiving a property of the object itself, namely, it's apparent magnitude.

None of Reid's arguments against the "way of ideas", as the Lockean model of the mental was sometimes called, are definitive. But they all aim to shift the burden of proof back to those who subscribe to the Lockean model. Since Reid saw the Lockean model as manifestly in violation of common sense, and since he took it to be a violation of correct philosophical methodology to accept any view contrary to common sense which was no better defended than the common-sensical view, he felt that to shift the burden of proof back to the subscribers to the Lockean model was to refute the model.

Sensation, Conception and Perception

Reid supplants the Lockean model with an act-based conception of the mental. That is, states of the mind cannot be analyzed into mental act and mental object, as on the Lockean model. The mind's states are all acts, loosely conceived; the only objects involved in thought are in the world, not the mind. The three most important mental acts are labelled by Reid "sensation", "conception", and "perception".

While none of these operations is defined or analyzed by Reid with exact philosophical rigor -- in fact, Reid suggests that such analyses aren't really possible in this domain -- it is possible to make some rough remarks about what Reid has in mind. Sensations are the feelings that are the immediate mental causal consequences of the influence of objects on us. Sensations are always associated with a particular organ

of sense; they are always distinctly *of* , for instance, touch or vision. Conceptions, on the other hand, are ways of being aware of objects. To conceive of an object is to be aware of that object *as* the bearer of some particular property. One might conceive of an apple *as* red or *as* hard or *as* both red and hard. Further, one could conceive of an object -- and thus be directly aware of that object -- as possessing a particular relational property: to conceive of an object as having a particular apparent magnitude, for instance, is to be aware of the object itself as possessing the property of appearing a certain way from a certain location. Perceptions are a species of conception. To perceive an object is to be aware of it in a particular way, as the possessor of a particular quality, *and*, at the same time, to be convinced that the object exists and is as you conceive it to be. Objects, then, act on our bodies and cause us to have sensations -- a feeling of coldness, a visual image of color. These sensations, in turn, lead us -- Reid sometimes says "suggest to us" -- conceptions of their causes; we become aware of the causes of our sensations as possessing various qualities. Sometimes, although not always, these conceptions are accompanied by a conviction in their accuracy, and when they are, they are called "perceptions".

To defend common sense, Reid thinks, he needs to show that we are directly aware of real objects and are, most of the time, roughly right about the nature of the objects of which we are aware. Reid does not think that anything comparable needs to be shown about sensation. That is, sensation -- the direct effect of objects on our minds -- needn't, in itself, accurately represent the world. Rather, our sensations must help us -- and much of the rest of this section is aimed at explaining how -- to place our minds in direct contact with the world. Sensations are merely tools for obtaining what Reid thinks we all believe, common-sensically, to be the case: that in having conceptions, we are aware of real objects that are roughly the way we conceive them to be.

Reid, then, is offering a form of "direct realism": the view that our minds connect to the world directly, rather than through some sort of medium (such as ideas) to which our minds connect and which itself, somehow, connects to the world. While there is debate over the precise sense in which, for Reid, we are directly aware of objects, this much seems clear: whatever the sense of "direct" is in which the subscribers to the Lockean model take us to be directly aware of ideas, it is in that sense that Reid takes us to be directly aware of real objects.

Primary and Secondary Qualities

Intertwined with Reid's view that we can be immediately aware of objects through their relational properties -- such as their apparent magnitudes -- are Reid's views regarding the distinction between primary and secondary qualities. Locke thought that our ideas of "primary qualities" -- shapes, sizes and motions -- and configurations of the primary qualities of the stuff out of which bodies are built -- texture and construction -- resemble the qualities which cause them. However, according to Locke, our ideas of a variety of other qualities -- particularly, colors, sounds, tastes and smells -- do not. Ideas of colors, sounds, tastes and smells are caused by certain complex configurations of primary qualities -- when such configurations are characterized as powers to produce ideas in us, then they are called "secondary qualities" -- that bear no resemblance whatsoever to the ideas which they cause. Reid is deeply struck by Berkeley's attack on this distinction (at, for instance, *Principles of Human Knowledge*, Part 1, Sections 9-15), and agrees with Berkeley that no mental state or object could possibly resemble anything that was

not, itself, a mental state or object; as Berkeley puts it "An idea can be like nothing but an idea" (*Principles of Human Knowledge*, Part 1, Section 8). Mental states and objects have only mental properties, but only something that is, itself, a mental state or object can have a mental property; hence, nothing can resemble -- that is, share a property in common with -- a mental state or object other than another mental state or object. Berkeley took this point to show that *no* non-mental cause of an idea could resemble it, whether the relevant idea were an idea of a shape, say, on the one hand, or a color on the other, and, thus, there could be no distinction between primary and secondary qualities of the sort drawn by Locke.

Reid accepts, for roughly Berkeley's reasons, that sensations cannot possibly resemble their causes (with one qualification regarding visual sensations discussed below). Further, he accepts Berkeley's objections to Locke, and takes them to show that no mental events or states, whether sensations or the conceptions of objects that follow them, could possibly resemble any non-mental object. In addition, there is another reason that Reid cannot draw the distinction between primary and secondary qualities in the way that Locke did: he does not accept that the conceptions of objects which we have following the sensations which they cause in us are to be analyzed as perceptions of ideas; they are, rather, awareness of the qualities in the object that caused the sensations. For Reid, there is no immediately perceived mental object that could succeed or fail to resemble its cause; it is the qualities of objects themselves of which we are directly aware when we have conceptions of those objects. All of this would make it seem that Reid would simply side with Berkeley and deny that there is any important difference between primary qualities and qualities like colors, sounds, tastes and smells. However, he does not take Berkeley's side, but, instead, defends the distinction between primary and secondary qualities on grounds quite different from Locke's, grounds that he takes to be immune to Berkeley's criticisms of the distinction.

Reid accepts that the qualities which we ordinarily conceive objects to have -- whether shapes, sizes and motions, on the one hand, or colors, sounds, tastes and smells, on the other -- are genuinely possessed by those objects (barring illusions and disorders of various sorts, which are, incidentally, difficult for Reid to explain). However, he thinks that shapes, sizes and motions are intrinsic properties of objects while colors, sounds, tastes and smells are relational properties of objects; and, it is to be emphasized, neither (with rare exceptions discussed below) resemble the mental states which they immediately cause in us, namely sensations. Colors, sounds, tastes and smells are powers to produce certain characteristic sensations in us in normal conditions; to ascribe such a quality to an object is not to perceive any intrinsic qualities of the object, but is, rather, to perceive that the object bears a certain relation to something else: namely, ourselves. So, for instance, say that the skin of the apple in front of me has a certain molecular structure that results in its reflecting light at a certain wavelength which in turn causes in me a certain characteristic visual sensation of red. If I am speaking correctly when I say, "That apple is red", I am reporting the fact that I conceive of the apple as possessing a particular relational property: I am aware that the apple has the property of being-such-as-to-cause-in-me-sensations-of-red-in-normal-conditions. Ultimately, the apple possesses this relational property because of facts about its molecular structure that account for its reflecting light in a certain way, and facts about me that account for the fact that such wavelengths of light cause certain sensations in me. But when I am aware of the redness of the apple, I am aware of none of that; I am aware only of the fact that there is something about the apple that makes it cause in me certain sensations in normal conditions.

Our conceptions of qualities such as colors are to be contrasted with our conceptions of primary qualities or configurations of primary qualities, such as hardness. Say I'm holding the apple in my hand while I'm looking at it. I'm having, then, two importantly different sensations: a visual sensation of red, and a tactile sensation of hardness. For Reid, neither sensation resembles anything in the object; both give rise to conceptions of the object as possessing certain properties. The visual sensation gives rise to a conception of the object as possessing a particular relational property: its power to produce certain sensations in me in normal conditions; the tactile sensation gives rise to a conception of the apple as possessing a particular intrinsic property: the complex configuration of primary qualities that is hardness.

So, there is a difference between primary and secondary qualities for Reid, although he draws the distinction in an importantly different way from the way in which it was drawn by Locke. For both Locke and Reid, we are aware of objects as they are intrinsically only when our awareness is caused by the primary qualities of objects. But for Reid, and not for Locke, we are genuinely aware of objects as they are when our awareness results from the secondary qualities of objects; but we are aware of those objects only as they are *relative* to us, and not as they are in themselves.

Sensation, Natural Signs and Suggestion

The immediate effect that objects have on us is to cause sensations. Further, we become aware of the qualities of objects following the sensations that those objects cause. However, for Reid, the conceptions of objects that follow from our sensations are not *derived* from our sensations; our sensations, after all, generally do not bear any kind of resemblance to the qualities which cause them. That is, according to Reid, we don't conceive of objects as possessing particular qualities by, say, representing our sensations, or drawing conclusions about the world through some sort of scrutiny of our sensations. Rather, our sensations give rise to our conceptions of objects by a process that he calls "suggestion": the qualities of objects are "suggested" by our sensations and so when we have sensations we come to be aware of those objects as possessing those qualities. But what is suggestion supposed to be?

Suggestion is a pseudo-linguistic notion, for Reid. Signs suggest conceptions of that which they signify. The word "pigs", for instance, leads those who are familiar with the word to think of certain pinkish barnyard animals. However, we don't come to think of such creatures on encountering the word "pigs" by somehow scrutinizing the word and thereby locating, in the world, some object that has some peculiar fitness to the word. After all, the word "pigs" is utterly arbitrary. While there are probably reasons why it came to signify what it signifies, there is no similarity between pigs and the word "pigs": pigs have four legs, for instance, not four letters.

Reid draws a distinction between natural and artificial signs. Artificial signs signify what they signify as a result of some sort of compact or tacit agreement between people: "pigs" signify pigs, according to Reid, only because people have agreed to use a particular sound and a particular configuration of letters to signify pigs. Natural signs, on the other hand, signify what they signify for other reasons entirely. For instance, blushing signifies embarassment only because of the fact that blushing and embarassment are customarily found together. The connection between them, just like the connection between the word

"pigs" and pigs, is utterly arbitrary: were it the case that, when embarassed, people stamped their left feet, then the stamping of the left foot would signify embarassment. (In fact, the distinction between natural and artificial signs has a long history. It can be found in, at least, Locke, Hobbes, Gassendi and the Port Royal Logic.)

We discover that blushing is a sign of embarassment through experience in ourselves and others of the co-occurrence of the two states: it is because of our acquaintance with human nature and with the conjunction of blushing and embarassment that we think of embarassment when we encounter blushing. However, some natural signs lead us to think of what they signify without any experience whatsoever. Reid describes this category of natural signs in unhesitatingly mystical terms. He says that there is a category of natural signs

> which, though we never before had any notion or conception of the things signified, do suggest it, or conjure it up, as it were, by a natural kind of magic, and at once give us a conception, and create a belief of it. (*Inquiry*, ch. 5, section 3, p. 60)

Reid has an example in mind of a natural sign which works "magically" in this way: the sensation of hardness. This tactile sensation leads us immediately to conceive of that which caused it as being hard, as having a certain resistive construction. (Reid assumes, perhaps wrongly, that the quality of hardness is a non-relational quality.) But we are aware of this quality, in the object which caused the sensation, automatically, "as it were, by a natural kind of magic". We cannot hope to understand why it is that we think of this special kind of construction after having the right kind of tactile sensation, but must "conclude, that this connection is the effect of our constitution, and ought to be considered as an original principle of human nature" (*Inquiry*, ch. 5, section 3, p. 61). In fact, this is the defining feature of the kind of natural signs of which the tactile sensation of hardness is an instance: these are signs that lead us to conceive of what they signify simply because we are built in such a way as to have such conceptions on encountering such signs; such a tendency is an inescapable feature of our constitution.

When does one's conception of an object as having a particular quality amount to a perception, a conception accompanied by a conviction of its accuracy? The answer is: when one comes to have that conception because one has encountered a natural sign which leads one to the conception, and that natural sign leads one to the conception merely because of one's constitution. Conviction in the accuracy of a conception is bestowed on the conception, Reid thinks, just when the relevant conception comes about because of our nature or constitution. When we conceive of an object in a particular way merely because it is in our nature to conceive of the object that way, then the conception is non-optional, unavoidable, and is thus one that we cannot help but trust; the conception of an object which one has when one encounters a natural sign which signifies merely because of one's constition is given with one's consititution; we can no more reject it than we can give up our own humanity. Perceptions, then, are dictates of common sense: to be aware of an object in perception is to have a belief which you cannot give up given your constitution.

So, for Reid, in the standard case, a quality of an object impinges on our bodies causing a sensation that has no resemblance whatsoever to the quality. This sensation, in turn, leads us to conceive of the object as

having the quality -- and thus to be directly aware of the object as possessing the quality -- merely because we are wired in such a way as to have such a conception after having the sensation; that is what it is for the sensation to "suggest" the conception of the quality. And, further, since the conception comes to be had as a result of our natures as human beings, we cannot help but trust it, and thus we are convinced of its accuracy and can be said to be perceiving the object. However, there are non-standard cases, for there are sensations that do bear a resemblance to -- or, at least, are the rational result of-the qualities which cause them. In particular, certain aspects of visual sensation are like this.

Reid is very impressed by results in geometric optics and holds that while there is no similarity between sensations of color and the quality in objects which cause those sensations, there is a non-arbitrary connection between visual sensations of shape and size and the qualities of objects which cause those sensations. One could rationally determine the shape and size of an object given nothing but a particular visual sensation -- even understood just as a particular impression on the retina -- and the laws of geometric optics, together with a few other pieces of information (such as the distance from oneself to the object). Or, conversely, even a blind person could construct the features of the retinal impression of the shape and size of an object when equipped with information about the actual size and shape of the object, the location of the observer, and the laws of geometrical optics. (This point is controversial. Keith Lehrer presents an alternative interpretation under which Reid holds that not even visual sensations resemble their causes. See Lehrer, *Thomas Reid*, 1989.) But, visual sensations of shape and size are natural signs not of the actual, but rather of the apparent shape and size of objects: when we encounter a visual sensation, we are immediately aware (that is, aware because of our natures) of the object as having a particular *apparent* shape and size (and a particular color). Since the actual shape and size are deducible from the apparent shape and size, together with some further pieces of information, we often conceive of the actual shape and size immediately after conceiving of the apparent shape and size. In fact, this movement of the mind from the conception -- that is, *perception* -- of apparent shape and size to the conception of actual shape and size is so quick, and so hard to notice, Reid thinks, that it is much like, although not in fact, perception. It is not perception, strictly speaking, since it doesn't come to pass merely because of our natures -- and therefore the route through which we come to have the conception doesn't necessarily deliver to us a conviction in the conception's accuracy -- but, rather, in some cases, the conception of actual size and shape comes about through a process of reasoning. He calls such conceptions, in keeping with Berkeley's terminology, "acquired perception".

Our acquired perceptions needn't be, although they sometimes are, acquired through *reasoning* from those rare sensations -- namely, visual sensations -- that bear a rational relationship to certain qualities of objects -- namely the real size and shape of objects. We sometimes acquire a perception of an object as having a certain quality simply by correlating data from more than one sense. So, for instance, the real size and shape of objects are immediately suggested by our tactile sensations; such sensations are natural signs of those sizes and shapes and signify only because of our constitutions. However, since we often encounter certain visual sensations -- which are natural signs only of the apparent size and shape of objects -- together with tactile sensations, we come to infer that a body that is merely seen would give rise to a certain tactile sensation were it touched, a tactile sensation that is itself a natural sign of the real size and shape of the object. Thus we come to conceive of the object as possessing a particular actual size or shape merely on seeing it despite the fact that our visual sensations are not natural signs of actual size or

shape. Since such deliverances of conceptions do not come only from our nature, they do not, in general, bring with them a conviction of their accuracy. Even if we are convinced of their accuracy, such a conviction is not guaranteed by the history of their occurrence, as is the case in direct perception, but comes from some other source.

In summary, then, on Reid's view our minds come to be connected to the world in something like the way that we come to grasp objects through a language designed for the purpose. However, when we come to be aware of objects through our senses, we do so by utilizing something like a language embedded in our constitutions: our sensations function like a language that nature has constructed, and that nature has constructed us to understand, for the purpose of signifying real objects. So while it is in a sense only a metaphor to say that, for Reid, we know about the world because the world speaks to us, it is a metaphor that illuminates the facts as he sees them.

Responding to External World Skepticism

We are now in a position to understand the force of Reid's most important response to any argument purporting to show that the external world either might not exist, or might not be anything like the way we take it to be. In one canonical statement of his position, Reid says,

> The sceptic asks me, Why do you believe the existence of the external object which you perceive? This belief, sir, is none of my manufacture; it came from the mint of Nature; it bears her image and superscription; and, if it is not right, the fault is not mine: I even took it upon trust, and without suspicion. Reason, says the sceptic, is the only judge of truth, and you ought to throw off every opinion and every belief that is not grounded on reason. Why, sir, should I believe the faculty of reason more than that of perception? -- they came both out of the same shop, and were made by the same artist; and if he puts one piece of false ware into my hands, what should hinder him from putting another? (*Inquiry*, ch. 6, section 20, pp. 168-169)

The mistake that the skeptic makes, according to Reid, is to deny the truth of something that is demanded by our constitutions. To perceive an object as possessing a particular property is to have a conception of the object which was, itself, delivered by one's nature. What makes us convinced of the accuracy of the conceptions of objects involved in perception is that they arise from our constitutions. But, asks Reid here, why do we find skeptical arguments so compelling? Why do we accept that skeptical conclusions follows from, for instance, Descartes's hypothesis of the evil demon? Ultimately, we think that such arguments lead to their conclusions because we accept certain logical principles -- such as the law of non-contradiction, or modus ponens -- which appear to us to be self-evident. But to say that such principles are self-evident is just to say that we cannot help but accept them; it is to say only that we are compelled to believe them. But the irresistibility of a belief is a very good indicator, Reid thinks, that we hold that belief merely because of the way we are built, merely because of our constitution. But then the skeptic has merely placed the skeptical conclusion on the same footing as the common sense belief about the external world: both rest on something that we are compelled to believe by our constitutions. However, in

order to overthrow common sense, the skeptic must place the skeptical conclusion, rather, on a *firmer* footing than the common sense conclusion. Thus, the skeptic gives us no reason whatsoever to reject common sense beliefs about the external world.

Reid, therefore, takes himself to have defended common sense through construction of an epistemological model that serves as an alternative to the Lockean view and, he thinks, thereby manages to avoid denial of our conception of ourselves as creatures capable of knowing about a non-mental world directly through our senses, a denial which he takes the work of Berkeley and Hume to show to be the inevitable result of acceptance of the Lockean picture.

Bibliography

Major Works

- 1748: *An Essay on Quantity; Occasioned by Reading a Treatise in which Simple and Compound Ratios are Applied to Virtue and Merit*
- 1764: *An Inquiry into the Human Mind on the Principles of Common Sense*
- 1774: *A Brief Account of Aristotle's Logic*
- 1785: *Essays on the Intellectual Powers of Man*
- 1788: *Essays on the Active Powers of Man*
- 1799: *A Statistical Account of the University of Glasgow*

All are included in *The Works of Thomas Reid, D.D.*.

Editions

- *Essays on the Active Powers of the Human Mind*, Baruch A. Brody (ed.), MIT Press, Cambridge 1969
- *Essays on the Intellectual Powers of Man*, Abridged, A.D. Woozley (ed.), Macmillan, London 1941
- *Essays on the Intellectual Powers of Man*, Baruch A. Brody (ed.), MIT Press, Cambridge 1969.
- *Inquiry and Essays*, Ronald E. Beanblossom and Keith Lehrer (eds.), Hackett, Indianapolis 1983.
- *An Inquiry into the Human Mind on the Principles of Common Sense*, Derek R. Brookes (ed.), Pennsylvania State University Press, University Park 1997
- *Lectures on the Fine Arts*, Peter Kivy (ed.), Martinus Nijhoff, The Hague 1973.
- *Lectures on Natural Theology (1780)* , Elmer H. Duncan (ed.), University Press of America 1981.
- *Philosophical Orations of Thomas Reid, delivered at Graduation Ceremonies in King's College, Aberdeen, 1753, 1756, 1759, 1762*, D.D. Todd, Southern Illinois University Press, Carbondale 1989.
- *Practical Ethics*, Knud Haakonssen (ed.), Princeton University Press, Princeton 1990.
- *Thomas Reid on the Animate Creation: Papers Relating to the Life Sciences*, Paul Wood (ed.),

Pennsylvania State University Press, University Park 1996.
- *The Works of Thomas Reid, D.D.*, Sir William Hamilton (ed.), G. Olms Verlagsbuchhandlung, Hildesheim 1983. (first edition, 1846)

Secondary Sources

Books and Collections of Articles:

- Dalgarno, Melvin and Eric Matthews (eds.). *The Philosophy of Thomas Reid*, Kluwer, Dordrecht 1989.
- Daniels, Norman. *Thomas Reid's Inquiry: The Geometry of Visibles and the Case for Realism*, Burt Franklin, New York 1974.
- Fraser, A. Campbell. *Thomas Reid*, Oliphant, Anderson and Ferrier, Edinburgh and London 1898.
- Gallie, Roger. *Thomas Reid and 'The Way of Ideas'*, Kluwer, Dordrecht 1989.
- Lehrer, Keith. *Thomas Reid*, Routledge, London 1989.
- Schneewind, J. B. *The Invention of Autonomy: A History of Modern Moral Philosophy*, Cambridge University Press, Cambridge 1998. (esp. pp. 395-403)
- Yolton, John W. *Perceptual Acquaintance from Descartes to Reid*, University of Minnesota Press, Minneapolis, 1984.
- Two issues of *The Monist* have been dedicated to Reid's philosophy: v. 70, 1987 and v. 61, 1978.

Selected Articles:

- Alston, William P. "Thomas Reid On Epistemic Principles" in *History of Philosophy Quarterly*. 1985; 2, 435-452.
- Anstey, Peter. "Thomas Reid and the Justification of Induction" in *History of Philosophy Quarterly*. 1995; 12, 1, 77-93.
- Bourdillon, Philip. "Thomas Reid's Account of Sensation as a Natural Principle of Belief" in *Philosophical Studies*. 1975; 27, 19-36.
- Chisholm, Roderick M. "Keith Lehrer and Thomas Reid" in *Philosophical Studies*. 1990; 33-38
- Cummins, Phillip D. "Reid's Realism" in *Journal of the History of Philosophy*. 1974; 12, 317-340
- Daniels, Norman. "Thomas Reid's Discovery of a Non-Euclidean Geometry" in *Philosophy of Science*. 1972; 3, 219-234.
- De Rose, Keith. "Reid's Anti-Sensationalism and His Realism" in *Philosophical Review*. 1989; 98, 313-348.
- Duggan, Timothy J. "Thomas Reid's Theory Of Sensation" in *Philosophical Review*. 1960; 69, 90-100.
- Gallie, Roger. "Reid: Conception, Representation and Innate Ideas" in *Hume Studies*. 1997; 23, 2, 315-335.
- Lehrer, Keith. "Conception Without Representation, Justification Without Inference: Reid's Theory" in *Nous*. 1989; 23, 145-154.
- McDermid, Douglas. "Thomas Reid on Moral Liberty and Common Sense" in *British Journal for*

the History of Philosophy. 1999; 7, 2, 275-303.
- Madden, Edward H. "Commonsense and Agency Theory" in *Review of Metaphysics*. 1982; 36, 319-342.
- ---------. "Was Reid a Natural Realist?" in *Philosophy and Phenomenological Research*. 1986; 47, 255-276.
- Manns, James. "Beauty and Objectivity in Thomas Reid" in *British Journal of Aesthetics*. 1988; 28, 119-131.
- Nadler, Steven M. "Reid, Arnauld, and the Objects of Perception" in *History of Philosophy Quarterly*. 1986; 3, 165-174.
- Nauckhoff, Josefine C. "Objectivity and Expression in Thomas Reid's Aesthetics" in *Journal of Aesthetics and Art Criticism*. 1994; 52, 2, 183-191.
- O'Connor, Timothy. "Thomas Reid on Free Agency" in *Journal of the History of Philosophy*. 1994; 32, 4, 605-622.
- Pappas, George S. "Sensation and Perception in Reid" in *Nous*. 1989; 23, 155-167.
- ----------. "Causation and Perception in Reid" in *Philosophy and Phenomenological Research*. 1990; 50, 4, 763-766.
- Robbins, David O. "The Aesthetics of Thomas Reid" in *Journal of Aesthetics and Art Criticism*. 1942; 2, 30-41.
- Schumann, Karl. "Elements of Speech Act Theory in the Work of Thomas Reid" in *History of Philosophy Quarterly*. 1990; 47-66.
- Van Woudenberg, Rene. "Thomas Reid on Memory" in *Journal of the History of Philosophy*. 1999; 37, 1, 117-133.

Other Internet Resources

- The Reid Project at the University of Aberdeen
- Reid Studies
- Martino Squillante's Reid Bibliography
- James Van Cleve's NEH seminar on Reid (Brown University)

Related Entries

Berkeley, George | causation: the metaphysics of | Hume, David | Locke, John | ordinary language | primary and secondary qualities | realism | Scottish Philosophy: in the 18th Century | skepticism

Gideon Yaffe
University of Southern California
yaffe@usc.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 28, 2000
Content last modified: August 28, 2000

Scottish Philosophy in the Eighteenth Century

Philosophy was at the core of the eighteenth century movement known as the Scottish Enlightenment. The movement included major figures, such as Francis Hutcheson, David Hume, Adam Smith, Thomas Reid and Adam Ferguson, and also many others who produced notable works, such as George Turnbull, George Campbell, James Beattie, Alexander Gerard, Henry Home (Lord Kames) and Dugald Stewart. I discuss some of the leading ideas of these thinkers, though paying less attention than I otherwise would to Hume, Smith and Reid, who have separate Encyclopedia entries. Amongst the topics covered in this entry are aesthetics (particularly Hutcheson's), Moral philosophy (particularly Hutcheson's and Smith's), Turnbull's providential naturalism, Kames's doctrines on divine goodness and human freedom, Campbell's criticism of the Humean account of miracles, the philosophy of rhetoric, Ferguson's criticism of the idea of a state of nature, and finally the concept of conjectural history, a concept especially associated with Dugald Stewart.

Major figures

The major figures in Scottish eighteenth century philosophy were Francis Hutcheson, David Hume, Adam Smith, Thomas Reid and Adam Ferguson. Others who produced notable works included George Turnbull, George Campbell, James Beattie, Alexander Gerard, Henry Home (Lord Kames) and Dugald Stewart.

Hutcheson on aesthetics

The first of the major philosophers was Francis Hutcheson (1694-1746). His reputation rests chiefly on his earlier writings, especially *An Inquiry into the Original of our Ideas of Beauty and Virtue* (London 1725), *Reflections upon Laughter and Remarks on the Fable of the Bees* (both in the Dublin Journal 1725-6), and *Essay on the Nature and Conduct of the Passions with Illustrations on the Moral Sense* (London 1728). His magnum opus, *A System of Moral Philosophy*, was published posthumously in Glasgow in 1755; a modern critical edition is awaited. During his period as a student at Glasgow University (c.1711-17) Hutcheson studied moral philosophy and jurisprudence under Gershom Carmichael, and in 1730 he took up the moral philosophy chair left vacant on Carmichael's death. Hutcheson is known principally for his ideas on moral philosophy and aesthetics. First moral philosophy.

Hutcheson reacted against both the psychological egoism of Thomas Hobbes and the rationalism of Samuel Clarke and William Wollaston. As regards Hobbes, Hutcheson thought his doctrine was both wrong and dangerous; wrong because by the frame of our nature we have compassionate, generous and benevolent affections which owe nothing at all to calculations of self-interest, and dangerous because people may be discouraged from the morally worthy exercise of cultivating generous affections in themselves on the grounds that the exercise of such affections is really an exercise in dissimulation or pretence. As against Hobbes Hutcheson held that a morally good act is one motivated by benevolence, a desire for the happiness of others. Indeed the wider the scope of the act the better, morally speaking, the act is; Hutcheson was the first to speak of ‘the greatest happiness for the greatest numbers’.

He believed that moral knowledge is gained via our moral sense. A sense, as the term is deployed by Hutcheson, is every determination of our minds to receive ideas independently of our will, and to have perceptions of pleasure and pain. In accordance with this definition, the five external senses determine us to receive ideas which please or pain us, and the will does not intervene -- we open our eyes and by natural necessity see whatever it is that we see. But Hutcheson thought that there were far more senses than the five external ones. Three in particular play a role in our moral life. The public sense is that by which we are pleased with the happiness of others, and are uneasy at their misery. The moral sense is that by which we perceive virtue or vice in ourselves or others, and derive pleasure, or pain, from the perception. And the sense of honour is that which makes the approbation, or gratitude of others, for any good actions we have done, the necessary occasion of pleasure. In each of these cases the will is not involved. We see a person acting with the intention of bringing happiness to someone else, and by the frame of our nature pleasure wells up in us.

Hutcheson emphasises both the complexity of the relations between our natural affections and also the need, in the name of morality, to exercise careful management of the relations between the affections. We

must especially be careful not to let any of our affections get too ‘passionate’, for a passionate affection might become an effective obstacle to other affections that should be given priority. Above all the selfish affections must not be allowed to over-rule ‘calm universal benevolence’.

Hutcheson's opposition to Hobbesian egoism is matched by his opposition to ethical rationalism, an opposition which emerges in the *Illustrations upon the Moral Sense,* where he demonstrates that his account of the affections and the moral sense makes sense of the moral facts whereas the doctrines of Clarke and Wollaston totally fail to do so. Hutcheson's main thesis against ethical rationalism is that all exciting reasons presuppose instincts and affections, while justifying reasons presuppose a moral sense. An exciting reason is a motive which actually prompts a person to act; a justifying reason is one which grounds moral approval of the act. Hutcheson demonstrates that reason, unlike affection, cannot furnish an exciting motive, and that there can be no exciting reason previous to affection. Reason does of course play a role in our moral life, but only as helping to guide us to an end antecedently determined by affection, in particular the affection of universal benevolence. On this basis, an act can be called ‘reasonable’, but this is not a point on the side of the rationalists, since they hold that reason by itself can motivate, and in this case it is affection, not reason that motivates, that is, that gets us doing something rather than nothing.

If we add to this the fact, as Hutcheson sees it, that it has never been demonstrated that reason is a fit faculty to determine what the ends are that we are obliged to seek, we shall see that Hutcheson's criticism of rationalism is that it can account for neither moral motivation nor moral judgment. On the other hand our natural affections, in particular benevolence, account fully for our moral motivation and our faculty of moral sense accounts fully for our ability to make an assessment of actions whether our own or others'.

Certain features of Hutcheson's moral philosophy appear in his aesthetic theory also. Indeed the two fields are inextricably related, as witness Hutcheson's reference to the ‘moral sense of beauty’. Two features especially work hard. He contends that we sense the beauty, sublimity or grandeur of a sight or of a sound. The sense of the thing's beauty, so to say, wells up unbidden. And associated with that sense, and perhaps even part of it -- Hutcheson does not give us a clear account of the matter -- is a pleasure that we take in the thing. We *enjoy* beautiful things and that enjoyment is not merely incidental to our sensing their beauty.

A question arises here regarding the features of a thing that cause us to see it as beautiful and to take pleasure in it. Hutcheson suggests that a beautiful thing displays unity (or uniformity) amidst variety. If a work has too much uniformity it is simply boring. If it has too much variety it is a jumble. An object, whether visual or audible, requires therefore to occupy the intermediate position if it is to give rise to a sense of beauty in the object. But if Hutcheson is right about the basis of aesthetic judgment how does disagreement arise? Hutcheson's reply is that our aesthetic response is affected in part by the associations that the thing arouses in our mind. If an object that we had found beautiful comes to be associated in our mind with something disagreeable this will affect our aesthetic response; we might even find the thing ugly. Hutcheson gives an example of wines to which men acquire an aversion after they have taken them in an emetic preparation. On this matter his position may seem extreme, for he holds that if two people have the same visual experience and if the thing experienced carries the same identical associations for

the two people, then they will have the same aesthetic response to the visible object. The position is however difficult to disprove, since if two people do in fact disagree about the aesthetic merit of an object, Hutcheson can say that the object produces different associations in the two spectators.

Hutcheson, Hume and Turnbull

Hutcheson influenced most of the Scottish philosophers who succeeded him, perhaps all of them, whether because he helped to set their agenda or because they appropriated, in a form suitable to their needs, certain of his doctrines. In the field of aesthetics for example, where Hutcheson led, many, including Hume, Reid, and Archibald Alison, followed. But influences can be hard to pin down and there is much dispute in particular concerning his influence on David Hume (1711-76). It is widely held that Hume's moral philosophy is essentially Hutchesonian, and that Hume took a stage further Hutcheson's projects of internalisation and of grounding our experience of the world on sentiment or feeling. For Hume agreed with Hutcheson that moral and aesthetic qualities are really sentiments existing in our minds, but he also argued that the necessary connection between any pair of events E1 and E2 which are related as cause to effect is also in our minds, for it is nothing more than a determination of the mind, due to custom or habit, to have a belief (a kind of feeling) that an event of kind E2 will occur next when we experience an event of kind E1. Furthermore Hume argues that what we think of as the ‘external’ world is almost entirely a product of our own imaginative activity. As against these reasons for thinking Hume indebted to Hutcheson there are the awkward facts that Hutcheson greatly disapproved of the draft of *Treatise* Book III that he saw in 1739 and that Hutcheson did his best to prevent Hume being appointed to the moral philosophy chair at Edinburgh University in 1744-5. In addition many of their contemporaries, such as Adam Smith and Thomas Reid, held that Hume's moral philosophy was significantly different from Hutcheson's. (See James Moore, ‘Hutcheson and Hume’.)

But if there is ample room for doubt over whether major areas of Hume's philosophy are developments of, or are even compatible with, Hutcheson's thought, the same cannot be said of George Turnbull (1689-1749), regent at Marischal College, Aberdeen (1721-7), and teacher of Thomas Reid. In view of the fact that the subtitle of Hume's *Treatise of Human Nature* is ‘An attempt to introduce the experimental method of reasoning into moral subjects’, it should be noted that Turnbull's *Principles of Moral Philosophy*, published in 1740 (the year of publication of Bk. III of the *Treatise*) but based on lectures given in Aberdeen in the mid-1720s, contains a defence of the claim that natural and moral philosophy are very similar types of enquiry. When Turnbull tells us that all enquiries into fact, reality, or any part of nature must be set about, and carried on in the same way, he is bearing in mind the fact , as he sees it to be one, that there are moral facts and a moral reality, and that our moral nature is part of nature and therefore to be investigated by the methods appropriate to the investigation of the natural world. As the natural philosopher relies on experience of the external world, so the moral philosopher relies on his experience of the internal world. It is, in Turnbull's judgment, the failure to do this that led to the moral scepticism (as Turnbull thought it to be) of Hobbes and Mandeville, whose reduction of morality to self-love flies in the face of experience and is a shock to common sense.

The experience in question is of the reality of the public affection in our nature, the immediate object of

which is the good of others, and the reality of the moral sense by which we are determined to approve such affections. This moral sense, of whose workings we are all aware, is the faculty by which, without the mediation of rational activity, we approve of virtuous acts and disapprove of vicious ones; and the approval and disapproval rise up in us without any regard for self-love or self-interest. In a very Hutchesonian way Turnbull invites us to consider the difference we feel when faced with two acts which are the same except for the fact that one of them is performed from love of another and the other is performed from self-interest. These facts about our nature have to be accommodated within moral philosophy just as the fact that heavy bodies tend to fall has to be accommodated within natural philosophy.

Turnbull is committed to a form of reliabilism according to which the faculties that we have by the frame or constitution of our nature are trustworthy. It is not simply that we are so constructed that we cannot but accept their deliverances; it is that we are also entitled to accept them. Turnbull, a deeply committed Christian, believed that the author of our nature would not have so constituted us as to accept the deliverances of our nature if our nature could not be relied upon to deliver up truth. We are in the hands of providence, and live directed towards the truth for that reason. This doctrine has been termed ‘providential naturalism’, and bears a marked resemblance to the language and also to the substance of Reid's position.

Kames on aesthetics and religion

Henry Home, Lord Kames, likewise taught a version of providential naturalism. In his *Essays on the Principles of Morality and Natural Religion* he has a good deal to say about the senses external and internal, treating them as enabling us, by the original frame of our nature, to gain access, without the use of reasoning processes, to the realities in the corresponding domains, including the moral domain. Kames's moral sense has as much to do with aesthetics as with morality; or rather, for Kames, no less than for Hutcheson, virtue is a kind of beauty, moral beauty, as vice is moral deformity. Beauty itself is ascribed to anything that gives pleasure. And as there are degrees of pleasure and pain, so also there are degrees of beauty and ugliness. In the lowest rank are things considered without regard to an end or a designing agent. The possibility of greater pleasure, and of the ascription of greater beauty, arises when an object is considered with respect to the object's end. A house, considered in itself, might be beautiful, but how much more beautiful is it judged to be if it seen to be well designed for human occupancy.

Approbation, as applied to works of art, is our pleasure at them when we consider them to be well fitted or suited to an end. The approbation is greater if the end for which the object is well suited also gives pleasure. A ship may give pleasure because it is so shapely, and also give pleasure because it is well suited to trade, and also give pleasure because trade also is a fine thing. If these further thing are taken into account the beauty of the ship is enhanced. Kames argues that these kinds of pleasure can also be taken in human actions, and that human acts can cause pleasure additionally by the special fact about them that they proceed from intention, deliberation and choice. In the case of, for example, an act of generosity towards a worthy person, the act is intentionally well suited, or fitted, to an end whose beauty is recognised by the agent. The fact that observation of acts displaying generosity, and other virtues, gives

us pleasure is due to the original constitution of our nature. The pleasure arises unbidden, and no exercise of will or reason is required, any more than we require to use our reason to see the beauty of a landscape or a work of art.

Kames wrote extensively on revealed and natural theology. As regards the latter, he often has Hume in his sights, particularly Hume's *Dialogues Concerning Natural Religion* (1779), with whose contents Kames was familiar decades before the work's publication. Hume held that in an inference from effect to cause no more should be assigned to the cause than is sufficient to explain the effect. In particular, if we argue from the existence of the natural world to the existence of God we should ascribe to God only such attributes as are requisite for the explanation of the world. And since the world is imperfect, why not say that we are not constrained by the facts in the natural order to ascribe perfection to God? Kames, on the other hand, holds that there are principles implanted in our nature that permit us to draw conclusions that reason alone does not sanction. If something is a tendency of our nature then we have to rely on it as a source of truth. Invoking just such a tendency Kames affirms that though we see both good and evil around us we do not conclude that the cause of the world must also be a mixture of good and evil: ‘it is a tendency of our nature to reject a mixed character of benevolence and malevolence, unless where it is necessarily pressed home upon us by an equality of opposite effects; and in every subject that cannot be reached by the reasoning faculty, we justly rely on the tendency of our nature.’ (*Essays* p.353) In any case Kames sees a world which is predominantly good even though it has ‘a few cross instances’. But the few cross instances might not look so cross, or even at all cross, if we had a fuller perspective, and Kames anticipates the time when that perspective will be granted us.

This latter position did not raise the hackles of the zealots among the Presbyterian clergy in Scotland, but Kames's position on free will caused a furore and he had to defend himself from attempts to expel him from the Kirk. Kames, accepting the concept of history, natural and human, as the gradual realisation of a divine plan, believed in universal necessity. The laws ordained by God ‘produce a regular train of causes and effects in the moral as well as material world, bringing about those events which are comprehended in the original plan, and admitting the possibility of none other.’ (*Essays* p.192) On the other hand, if we are to fulfill our role in the grand scheme we must see ourselves as able to initiate things, that is, to be the free cause of their occurrence. God has therefore, according to Kames, concealed from us the necessity of our acts and he is therefore a deceitful God. Kames sought to explain how this divine deceit enables us to live as morally accountable beings, but this latter part of his philosophy did nothing to placate those in the Kirk for whom the affirmation of a deceitful God was a sacrilege. Kames, however, could not see any difference between the deception by which we believe ourselves to be free when in fact we are necessitated and the deception by which we believe secondary qualities, such as colours and sounds, to be in the external world and able to get along without us, when in fact they depend for their existence upon the exercise of our own sensory powers.

Campbell on miracles

Kames did not dedicate an entire book to an attack on Hume on religion, but George Campbell (1719-96) did. This interesting man, a student at Marischal College, Aberdeen, of which in 1759 he became

Principal, was a founder-member of the Aberdeen Philosophical Society, the ‘Wise Cub’, which also included Thomas Reid, John Gregory, David Skene, Alexander Gerard, James Beattie and James Dunbar. It is probable that many of Campbell's writings began life as papers to the Club. In 1763 he published *A Dissertation on Miracles* which was intended as a demolition of Hume's essay ‘On miracles’, Chapter Ten in *An Enquiry Concerning Human Understanding*. Miracles were commonly discussed in eighteenth century Scotland. On the one hand the Kirk required people to accept miracle claims on the basis either of eyewitness reports or of reports of such reports, and on the other hand the spirit of Enlightenment required that claims based on the authority of others be put before the tribunal of reason. Hume focuses especially on the credibility of testimony, and argues that the credence we place in testimony is based entirely on experience, the experience of the occasions when testimony has turned out to be true as against those experiences where it has not. Likewise it is on the basis of experience that we judge whether a reported event occurred. If the reported event is improbable we ask how probable it is that the eyewitness is speaking truly. We have to balance the probability that the eyewitness is speaking truly against the improbability of the occurrence of the event. Hume held that the improbability of a miracle is always so great that no testimony could tell effectively in its favour. The wise man, proportioning his belief to the evidence, would believe that the testimony in favour of the miracle is false.

Campbell's opening move against this argument is to reject Hume's premiss that we believe testimony solely on the basis of experience. For, according to Campbell, there is in all of us a natural tendency to believe other people. This is not a learned response based on repeated experience but an innate disposition. In practice this principle of credulity is gradually finessed in the light of experience. Once testimony is placed before us it becomes the default position, something that is true unless or until proved false, not false unless or until proved true. The credence we give to testimony is much like the credence we give to memory. It is the default position as regards beliefs about the past, even though in the light of experience we might withhold belief from some of its deliverances

Because our tendency to accept testimony is innate, it is harder to overturn than Hume believes it to be. Campbell considers the case of a ferry that has safely made a crossing two thousand times. I, who have seen these safe crossings, meet a stranger who tells me solemnly that he has just seen the boat sink taking with it all on board. The likelihood of my believing this testimony is greater than would be implied by Hume's formula for determining the balance of probabilities. Reid, a close friend of Campbell's, likewise gave massive emphasis to the role of testimony, stressing both the innate nature of the credence we give to testimony and also the very great proportion of our knowledge of the world that we gain, not through perception or reason, but through the testimony of others. Reid's comparison of the credence we naturally give to the testimony of others and the credence we naturally give to the deliverances of our senses, is one of the central features of his *Inquiry into the Human Mind* (1764).

Campbell and the rhetorical tradition

A number of eighteenth century Scots, including James Burnett (Lord Monboddo), Adam Smith, Thomas Reid, Hugh Blair and James Dunbar, made significant contributions in the field of language and rhetoric. George Campbell's *The Philosophy of Rhetoric* (London 1776) is a large-scale essay in which he takes a

roughly Aristotelian position on the relation between logic and rhetoric, since he holds that convincing an audience, which is the province of rhetoric or eloquence, is a particular application of the logician's art. The central insight from which Campbell is working is that the orator seeks to persuade people, and in general the best way to persuade is to produce perspicuous arguments. Good orators have to be good logicians. Their grammar also must be sound. This double requirement of orators leads Campbell to make a sharp distinction between logic and grammar, on the grounds that though both have rules, the rules of logic are universal and those of grammar particular. Though there are many natural languages there is but one set of rules of logic, and on the other hand different languages have different rules of grammar. It is against a background of discussion by prominent writers on language such as Locke and James (‘Hermes’) Harris that Campbell takes his stand with the claim that there cannot be such a thing as a universal grammar. His argument is that there cannot be a universal grammar unless there is a universal language, and there is no such thing as a universal language, just many particular languages. There are, he grants, collections of rules that some have presented under the heading ‘universal grammar’. But, protests, Campbell, ‘such collections convey the knowledge of no tongue whatever’. His position stands in interesting relation to Reid's frequent appeals to universals of language in support of the claim that given beliefs are held by all humankind.

Common sense

Campbell was a leading member of the school of common sense philosophy. For him common sense is an original source of knowledge common to humankind, by which we are assured of a number of truths that cannot be evinced by reason and ‘it is equally impossible, without a full conviction of them, to advance a single step in the acquisition of knowledge’. (*Philosophy of Rhetoric*, vol.1, p.114) His account is much in line with that of his colleague James Beattie: ‘that power of the mind which perceives truth, or commands belief, not by progressive argumentation, but by an instantaneous, instinctive, and irresistible impulse; derived neither from education nor from habit, but from nature; acting independently on our will, whenever its object is presented, according to an established law, and therefore properly called Sense; and acting in a similar manner upon all, or at least upon a great majority of mankind, and therefore properly called *Common Sense*.’ (*An Essay on the Nature and Immutability of Truth*, 40) We are plainly in the same territory as Reid's account: ‘there are principles common to [philosophers and the vulgar] which need no proof, and which do not admit of direct proof’, and these common principles ‘are the foundation of all reasoning and science’. (*Essays on the Intellectual Powers*, ed. Hamilton, vol.1, 230A-B)

These philosophers do however disagree about substantive matters. In particular, Reid lists as the first principle of common sense: ‘The operations of our mind are attended with consciousness; and this consciousness is the evidence, the only evidence, which we have or can have of their existence.’ (*Essays on the Intellectual Powers*, 231B) Campbell on the other hand lists three sorts of intuitive evidence. The first concerns our unmediated insight into the truth of mathematical axioms and the third concerns common sense principles. The second concerns the deliverances of consciousness, consciousness being the faculty through which we learn directly of the occurrence of mental acts -- thinking, remembering, being in pain, and so on. What is listed as a principle of common sense by Reid is, therefore, according to Campbell, to be contrasted with such principles. Aside from this, however, it is clear that Campbell is

philosophically very close to Reid, even if Reid is unquestionably the greater philosopher.

Smith on moral sentiments

Reid and Hume both owed an immense debt to Hutcheson. So also did Adam Smith (1723-90) who, unlike the others, studied under Hutcheson at Glasgow University. In 1751 Smith was appointed to the chair of logic and rhetoric at Glasgow and the following year transferred to the chair of moral philosophy that Hutcheson had occupied. Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations* appeared in 1776. *Essays on Philosophical Subjects* appeared postumously in 1795. He also published an essay on the first formation of languages, and student notes of his lectures on rhetoric and belles lettres, and on jurisprudence have survived. But much his most important work in philosophy is the *Theory of Moral Sentiments* which appeared in 1759 and of which six authorised editions appeared during Smith's lifetime.

The concepts of sympathy and spectatorship, central to the doctrine of *TMS*, had already been put to work by Hutcheson and Hume, but Smith's account is distinct. As spectator of an agent's suffering we form in our imagination a copy of such ‘impression of our own senses’ as we have experienced when we have been in a situation of the kind the agent is in: ‘we enter as it were into his body, and become in some measure the same person with the agent.’ (*Theory of Moral Sentiments*, p.9) Smith gives two spectacular examples of cases where the spectator has a sympathetic feeling that does not correspond to the agent's. The first concerns the agent who has lost his reason. He is happy, unaware of his tragic situation. The spectator imagines how he himself would feel if reduced to this same situation. In this imaginative experiment, in which the spectator is operating on the edge of a contradiction, the spectator's idea of the agent's situation plays a large role while his idea of the agent's actual feelings has a role only in that the agent's happiness is itself evidence of his tragedy. The second of Smith's examples is the spectator's sympathy for the dead, deprived of sunshine, conversation and company. Again Smith emphasises the agent's situation, and asks how the spectator would feel if in the agent's situation, deprived of everything that matters to people.

Smith relates sympathy to approval. For a spectator to approve of an agent's feelings is for him to observe that he sympathises with the agent. This account is used as the basis of the analysis of propriety. For a spectator to judge that an agent's act is proper or appropriate is for him to approve of the agent's act. The agent's act lacks propriety, in the judgment of the spectator, if the spectator does not sympathise with the agent's performance.

Propriety and impropriety are based on a bilateral relation, between spectator and agent. Smith attends also to a trilateral relation, between a spectator, an agent who acts on someone, and the person who is acted on, the ‘recipient’ of the act. There are several kinds of response that the recipient may make to the agent's act, and Smith focuses on two, gratitude and resentment. If the spectator judges the recipient's gratitude proper or appropriate then he approves of the agent's act and judges it meritorious, or worthy of reward. If he judges the recipient's resentment proper or appropriate then he disapproves of the agent's act and judges it demeritorious, or worthy of punishment. Judgments of merit or demerit concerning a

person's act are therefore made on the basis of an antecedent judgment concerning the propriety or impropriety of another person's reaction to that act. Sympathy underlies all these judgments, for in the cases just mentioned the spectator sympathises with the recipient's gratitude and with his resentment. He has direct sympathy with the affections and motives of the agent and indirect sympathy with the recipient's gratitude; or in judging the agent's behaviour improper the spectator has indirect sympathy with the agent's resentment. (*Theory of Moral Sentiments*, p.74)

We have supposed, in each of these cases, that the recipient really does have the feeling in question, whether of gratitude or resentment. However, in Smith's account the spectator's belief about what the recipient actually feels about the agent is not important for the spectator's judgment concerning the merit and demerit of the agent. The recipient may, for whatever reason, resent an act that was kindly intentioned and in all other ways admirable, and the spectator, knowing the situation better than the recipient does, puts himself imaginatively in the shoes of the recipient while taking with him into this spectatorial role information about the agent's behaviour that the recipient lacks. The spectator judges that were he himself in the recipient's situation he would be grateful for the agent's act; and on that basis, and independently of the recipient's actual reaction, he approves of the agent's act and judges it meritorious. Here the spectator considers himself as a better (because better informed) spectator of the agent's act than the recipient is.

As regards judgments of merit and demerit, Smith sets up a model of three people, but the three differ in respect of the weight that has to be given to their work, for the recipient does almost nothing. He is acted on by the agent, but apart from that he is no more than a place holder for the spectator who will imaginatively occupy his shoes and make a judgment concerning merit or demerit on the basis solely of his conception of how he would respond to the agent if he were in the place of the recipient. He does not judge on the basis of the actual reaction of the recipient, who might approve of the agent's act or disapprove or have no feelings about it one way or the other.

Up to this point we have attended to the spectator's moral judgment of the acts of others. What of his judgment of his own acts? In judging the other the spectator has the advantage of disinterest, but he may lack requisite information and much of the work of creative imagination goes into his rectifying the lack. In judging himself he has, or may be presumed to have, the requisite information but he has the problem of overcoming the tendency to a distorted judgment caused by self-love or self-interest. He must therefore factor out of his judgment those features that are due to self-love. He does this by setting up, by an act of creative imagination, a spectator, an *other* who, *qua* spectator, is at a distance from him.The point about the distance is that it creates the possibility of disinterest or impartiality, but it is still necessary to ask how disinterest or impartiality is achieved if it is the agent himself who imagines the spectator into existence.

Let us move to an answer by wondering who or what it is that is imagined into existence? Is it the voice of society, representing established social attitudes? At times in the first edition of *The Theory of Moral Sentiments* Smith comes close to saying that it is. In the second edition Smith is clear that this is not the role of the impartial spectator for the latter can, and occasionally does, speak against established social attitudes. Nor can the judgment of the impartial spectator be reduced to the judgment of society, even where those two judgments coincide. Nevertheless the impartial spectator exists because of real live

spectators. Were it not for our discovery that while we are judging other people, those same people are judging us, we would not form the idea of a spectator judging us impartially.

The impartial spectator is a product of the imagination, and its mode of existence is therefore intentional -- it has what medieval philosophers termed *esse intentionale* as against *esse naturale*. In one sense therefore it should be thought of not as a real spectator who has the merit of being impartial, but as an ideal spectator in the sense of one that exists as an idea. In another sense the impartial spectator is real, for it is no other than the agent who is imagining it into existence.

Smith's account of justice is built upon his account of the spectator's sympathetic response to the recipient of an agent's act. If a spectator sympathises with a recipient's resentment at the agent's act then he judges the act demeritorious and the agent worthy of punishment. In the latter case the moral quality attributed to the act is injustice. An act of injustice ‘does a real and positive hurt to some particular persons, from motives which are naturally disapproved of’. (*Theory of Moral Sentiments* p.79) Since a failure to act justly has a tendency to result in injury, while a failure to act charitably or generously does not, a distinction is drawn by Smith, in line with Humean thinking, between justice and the other social virtues, on the basis that it is so much more important to prevent injury than promote positive good that the proper response to injustice is punishment, whereas we do not feel it appropriate to punish someone who does not act charitably or gratefully. In a word, we have a stricter obligation to justice than to the other virtues.

Though there are important points of contact between Smith's account of justice and Hume's, the differences are considerable, chief of them being the fact that Hume grounds our approval of justice on our recognition of its utility, and Smith does not. We do sometimes take it into account in coming to a judgment, but more often than not it is something of a quite different nature that wells up in us: ‘All men, even the most stupid and unthinking, abhor fraud, perfidy, and injustice, and delight to see them punished. But few men have reflected upon the necessity of justice to the existence of society, how obvious soever that necessity may appear to be.’ (*Theory of Moral Sentiments*, p.89) There are a few cases where utility is plainly involved in our judgment, but they are few, and they are in a distinct psychological class. Smith instances the sentinel who fell asleep while on watch and was executed because such carelessness might endanger the whole army. Smith's comment is: ‘When the preservation of an individual is inconsistent with the safety of a multitude, nothing can be more just than that the many should be preferred to the one. Yet this punishment, how necessary soever, always appears to be excessively severe. The natural atrocity of the crime seems to be so little, and the punishment so great, that it is with great difficulty that our heart can reconcile itself to it.’ (*Theory of Moral Sentiments*, p.90) And our reaction in this kind of case is to be contrasted with our reaction to the punishment of ‘an ungrateful murderer or parricide’, where we applaud the punishment with ardour and would be enraged and disappointed if the murderer escaped punishment. These very different reactions demonstrate that our approval of punishment in the one case and in the other are founded on very different principles.

Blair's Christian stoicism

Smith devotes considerable space to the Stoic virtue of self-command. Another eighteenth century

Scottish thinker who devotes considerable space to it is Hugh Blair (1718-1800), minister of the High Kirk of St Giles in Edinburgh and first professor of rhetoric and belles lettres at Edinburgh University. Blair's sermons bear ample witness to his interest in Stoic virtue. For example, in the sermon 'On our imperfect knowledge of a future state' he wonders why we have been left in the dark about our future state. Blair replies that to see clearly into our future would have disastrous consequences. We would be so spellbound by the sight that we would neglect the arts and labours which support social order and promote the happiness of society. We are, believes Blair, in 'the childhood of existence', being educated for immortality. The education is of such a nature as to enable us to develop virtues such as self-control and self-denial. These are Stoic virtues, and Blair's sermons are full of the need to be Stoical. In his sermon 'Of the proper estimate of human life' he says: 'if we cannot control fortune, [let us] study at least to control ourselves.' Only through exercise of self- control is a virtuous life possible, and only through virtue can we attain happiness. He adds that the search for worldly pleasure is bound to end in disappointment and that that is just as well. For it is through the failure of the search that we come to a realisation both of the essential vanity of the life we have been living and also of the need to turn to God and to virtue. For many, the fact of suffering is the strongest argument there is against the existence of God. Blair on the contrary holds that our suffering provides us with a context within which we can discover that our true nature is best realised by the adoption of a life-plan whose overarching principle is religious.

Ferguson and the social state

One of Blair's colleagues at Edinburgh University was Adam Ferguson (1723-1816). He succeeded David Hume as librarian of the Advocates' Library in Edinburgh and then held in succession two chairs at Edinburgh University, that of natural philosophy (1759-64) and of pneumatics and moral philosophy (1764-85). His most influential work is *An Essay on the History of Civil Society* (1767). Ferguson attended to one of the main concepts of the Enlightenment, that of human progress, and expressed doubts about whether over the centuries the proportion of human happiness to unhappiness had increased. He believed that each person accommodates himself to the conditions in his own society and the fact that we cannot imagine that we would be contented if we lived in an earlier society does not imply that people in earlier societies were not, more or less, as happy in their own society as we are in ours. As against our unscientific conjectures about how we would have felt in a society profoundly unlike the only one we have ever lived in, Ferguson commends the use of historical records. He talks disparagingly about boundless regions of ignorance in our conjectures about other societies, and among those he has in mind who speak ignorantly about earlier conditions of humanity are Hobbes, Rousseau and Hume in their discussions of the state of nature and the origins of society.

Hobbes and Rousseau in particular had a good deal to say about the pre-social condition of humankind. Ferguson argues, against their theories, that there are no records whatever of a pre-social human condition; and since on the available evidence humankind has always lived in society he concludes that living in society comes naturally to us. Hence the state of nature is a social state and is not antecedent to it.

Dugald Stewart on history and philosophy

One colleague of Blair and Ferguson at Edinburgh University was Dugald Stewart (1753-1828), who was a student first at Edinburgh, and then at Glasgow where his moral philosophy professor was Thomas Reid. Stewart succeeded his father in the chair of mathematics at Edinburgh, and then in 1785 became professor of pneumatic and moral philosophy at Edinburgh when Ferguson resigned the chair. Stewart shared with Ferguson an interest in the kind of historical (or pseudo-historical) writings to be found in Hobbes' *Leviathan* and Rousseau's *Contrat Social*. In his *Account of the Life and Writings of Adam Smith LL.D.* Dugald Stewart says of one of Smith's works, the *Dissertation on the Origin of Languages*, that ‘it deserves our attention less, on account of the opinions it contains, than as a specimen of a particular sort of inquiry, which, so far as I know, is entirely of modern origin’. (Smith, *Essays on Philosophical Subjects*, p.292) Stewart then spells out the ‘particular sort of inquiry’ that he has in mind. He notes the lack of direct evidence for the origin of language, of the arts and the sciences, of political union, and so on, and affirms: ‘In this want of direct evidence, we are under a necessity of supplying the place of fact by conjecture; and when we are unable to ascertain how men have actually conducted themselves upon particular occasions, of considering in what manner they are likely to have proceeded, from the principles of their nature, and the circumstances of their external situation.’ (*Essays on Philosophical Subjects* p.293)

For Stewart such enquiries are of practical importance, for by them ‘a check is given to that indolent philosophy, which refers to a miracle, whatever appearances, both in the natural and moral worlds, it is unable to explain’. (*Essays on Philosophical Subjects*, p.293) Stewart uses the term ‘conjectural history’ for the sort of history exemplified by Smith's account of the origin of language. Conjectural history works against the illegitimate encroachment of religion into the lives of people who are too quick to reach for God as the solution to a problem when extrapolation from scientifically established principles of human nature would provide a solution satisfying to the intellect. Knowing what we do about human nature, about our intellect and will, our emotions and fundamental beliefs, we ask how people would have behaved in given circumstances. Love and hate, anger and jealousy, joy and fear, do not change much through the generations. Much the same things, speaking generally, have much the same effect first on the emotions and then on behaviour. Dugald Stewart formulates the principle underlying conjectural history: it has ‘long been received as an incontrovertible logical maxim that the capacities of the human mind have been in all ages the same, and that the diversity of phenomena exhibited by our species is the result merely of the different circumstances in which men are placed’. (Stewart, *Collected Works*, ed. William Hamilton vol.1, p.69)

As regards the credentials of Stewart's ‘incontrovertible logical maxim’, if the claim that human nature is invariant is an empirical claim, it must be based on observation of our contemporaries and on evidence of people's lives in other places and at other times. Such evidence needs however to be handled with care. The further back we go the more meagre it is, and so the more we need to conjecture to supplement the few general facts available to us. Indeed we can go back so far that we have no facts beyond the generalities that we have worked out in the light of our experience. But to rely on conjecture in order to support the very principle that forms the first premiss in any exercise in conjectural history is to come

suspiciously close to arguing in a circle. The incontrovertible logical maxim of Dugald Stewart should probably be accorded at most the status of a well-supported empirical generalisation.

Conjectural history is certainly not pure guesswork. We argue on the basis of observed uniformities, and the more experience we have of given uniformities the greater credence we will give to reports that speak of the occurrence of the uniformities, whether they concern dead matter or living people and their institutions. In a famous passage Hume writes: ‘Whether we consider mankind according to the difference of sexes, ages, governments, conditions, or methods of education; the same uniformity and regular operation of natural principles are discernible. Like causes still produce like effects; in the same manner as in the mutual action of the elements and powers of nature.’ (*Treatise* p.401)

For Hume the chief point about the similarity between ourselves and our ancestors is that histories greatly contribute to the scientific account of human nature by massively extending our otherwise very limited observational data base. Hume writes: ‘Mankind are so much the same, in all times and places, that history informs us of nothing new or strange in this particular. Its chief use is only to discover the constant and universal principles of human nature, by showing men in all varieties of circumstances and situations, and furnishing us with materials from which we may form our observations and become acquainted with the regular springs of human action and behaviour. These records of wars, intrigues, factions, and revolutions, are so many collections of experiments, by which the politician or moral philosopher fixes the principles of his science, in the same manner as the physician or natural philosopher becomes acquainted with the nature of plants, minerals, and other external objects, by the experiments which he forms concerning them.’ (*Enquiry Concerning Human Understanding* 83-4) On this account of history, it is perhaps the single most important resource for the philosopher seeking to construct a scientific account of human nature. Among the historians produced by eighteenth century Scotland were Turnbull, Hume, Smith and Ferguson. In light of Hume's observation it is not surprising that so much history was written by men prominent for their philosophical writings on human nature.

Bibliography

- Alison, Archibald. *Essays on the Nature and Principles of Taste*. 2 vols. Edinburgh, 1790.
- Beattie, James. *An Essay on the Nature and Immutability of Truth, in Opposition to Sophistry and Scepticism*. Edinburgh, 1770.
- Blair, Hugh. *Sermons*. 2 vols. Edinburgh: Anderson, 1824-5.
- Brown, Thomas. *Lectures on the Philosophy of the Human Mind*. 4 vols. Edinburgh, 1820.
- Campbell, George. *A Dissertation on Miracles*. Edinburgh, 1762.
- Campbell, George. *The Philosophy of Rhetoric*. 2 vols. London, 1776.
- Campbell, T. D. *Adam Smith's Science of Morals*. London: Allen and Unwin, 1971.
- Dunbar, James. *Essays on the History of Mankind in Rude and Cultivated Ages*. 2nd ed. London, 1781. Reprint, with an introduction by C. J. Berry, Bristol: Thoemmes Press, 1995.
- Ferguson, Adam. *An Essay on the History of Civil Society*. Edited by F. Oz-Salzberger. Cambridge: Cambridge University Press, 1995.
- Gaskin, J. C. A. *Hume's Philosophy of Religion*. 2nd ed. Basingstoke: Macmillan, 1988.

- Gerard, Alexander. *An Essay on Taste*. London, 1759. Reprint, Menston: Scolar Press, 1971.
- Grave, S. A. *The Scottish Philosophy of Common Sense*. Oxford: Clarendon Press, 1960.
- Haakonssen, Knud. *The Science of a Legislator: The Natural Jurisprudence of David Hume and Adam Smith*. Cambridge: Cambridge University Press, 1981.
- Haakonssen, Knud. *Natural Law and Moral Philosophy, from Grotius to the Scottish Enlightenment*. Cambridge: Cambridge University Press, 1995.
- Kames, Henry Hume, Lord. *Essays on the Principles of Morality and Natural Religion*. 3rd ed. Edinburgh, 1779.
- Hope, Vincent, ed. *Philosophers of the Scottish Enlightenment*. Edinburgh: Edinburgh University Press, 1984.
- Hume, David. *A Treatise of Human Nature*. Edited by L. A. Selby-Bigge. 2nd ed., revised by P. H. Nidditch. Oxford: Clarendon Press, 1978.
- Hume, David. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Edited by L. A. Selby-Bigge. 3rd ed., revised by P. H. Nidditch. Oxford: Clarendon Press, 1975.
- Hutcheson, Francis. *An Inquiry into the Original of our Ideas of Beauty and Virtue; in Two Treatises*. London, 1725.
- Hutcheson, Francis. *An Essay on the Nature and Conduct of the Passions and Affections. With Illustrations on the Moral Sense*. Dublin, 1728.
- Hutcheson, Francis. *A System of Moral Philosophy*. 2 vols. Glasgow, 1755.
- Lehrer, Keith. *Thomas Reid*. London: Routledge, 1989.
- McCosh, James. *The Scottish Philosophy: Biographical, Expository, Critical, from Hutcheson to Hamilton*. London: Macmillan,1875. Reprint, Bristol: Thoemmes Press, 1990.
- Monboddo, James Burnett, Lord. *Antient Metaphysics: or, the Science of Universals*. 6 vols. Edinburgh, 1779-99.
- Moore, James. ‘Hutcheson and Hume’, in *Hume and Hume's Connexions*, edited by M. A. Stewart and John P. Wright, 23-57. Edinburgh: Edinburgh University Press, 1990.
- Reid, Thomas. *The Works of Thomas Reid*. 2 vols. 6th ed., edited by Sir William Hamilton. Edinburgh: Maclachlan and Stewart, 1863. Reprint, Bristol: Thoemmes Press, 1994.
- Reid, Thomas. *An Inquiry into the Human Mind*. Edited by Derek R. Brookes. Edinburgh: Edinburgh University Press, 1997.
- Reid, Thomas. *Practical Ethics*. Edited by Knud Haakonssen. Princeton, N.J.: Princeton University Press, 1990.
- Scott, William R. *Francis Hutcheson: his Life, Teaching and Position in the History of Philosophy*. Cambridge: Cambridge University Press, 1990.
- Smith, Adam. *Essays on Philosophical Subjects*. London, 1795. Reprint, edited by W. P. D. Wightman and J. C. Bryce, Oxford: Clarendon Press, 1980; Indianapolis: Liberty Classics, 1982.
- Smith, Adam. *The Theory of Moral Sentiments*. London, 1790. Reprint, edited by D. D. Raphael and A. L. Macfie, Oxford: Clarendon Press, 1976; Indianapolis: Liberty Classics, 1982.
- Stewart, Dugald. *Collected Works*. Edited by Sir William Hamilton.11 vols. Edinburgh: Constable, 1854-58.
- Stewart, M. A., ed. *Studies in the Philosophy of the Scottish Enlightenment*. Oxford: Clarendon Press, 1990.

- Stewart, M. A. and John P. Wright, eds. *Hume and Hume's Connexions*. Edinburgh: Edinburgh University Press, 1995.
- Turnbull, George. *Principles of Moral and Christian Philosophy*. London, 1740.
- Turnbull, George. *Observations upon Liberal Education*. London, 1742.
- Wolterstorff, Nicholas. *Thomas Reid and the Story of Epistemology*. Cambridge: Cambridge University Press, 2001.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Hobbes, Thomas | Hume, David | Reid, Thomas

Alexander Broadie
A.Broadie@philosophy.arts.gla.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 27, 2001
Content last modified: June 27, 2001

The Biological Notion of Self and Non-self

Fundamental to biology are (1) defining the characteristics of identity which distinguish individual organisms from those of similar kind, and (2) describing the mechanisms that defend organisms from their predators. Immunology is the science devoted to these problems. A progeny of late 19th-century pathology and microbiology, immunology did not attain a formal theoretical construction until after World War II, when “the self” was introduced to provide a ready and convenient metaphor for deciphering immune reactivity. In the original formulation, normally, host constituents are ignored by the immune system, while “the other”—pathogens, foreign substances, altered host elements—are processed and destroyed. By the late 1970s, “the self” became the foundation of immune theorizing, and immunology dubbed itself the science of “self/non-self discrimination.” But this dominant model has recently been challenged, for the self is polymorphous and ill-defined. Contemporary transplantation biology and autoimmunity have demonstrated phenomena that fail to allow strict adherence to such a dichotomy of self/non-self, and as new models are emerging, “the self” has been left exposed as a metaphor, whose grounding—philosophically and scientifically—is unsteady and thus increasingly elusive as the putative nexus of immunology's doctrines.

Introduction

At an important nexus of pathology, clinical medicine, and basic biology, immunology has served

several research agendas and thus defies a single, unifying experimental framework. Rather it is (and has been) characterized by multiple, even competing thought-styles (Crist and Tauber 1997), each requiring a different methodological apparatus to order its experimental program. But underlying each branch of immunology, the concept of an identified and protected "self," a theoretical construction and fecund metaphor, has served as the central theme which integrates this diverse discipline. Indeed, the fate of "the self" in immunology offers a historical understanding of how the science has evolved.

Immunology during the first half of the twentieth century was pre-occupied with the more focused chemical questions of immune specificity, and the broader biological questions concerning immune identity remained unformulated. But after World War II, transplantation and autoimmunity became increasingly relevant both to basic immunologists and clinicians. These later concerns required a theoretical apparatus that explicitly addressed the question of biological identity and individuality. It was at this juncture that Sir Frank Macfarlane Burnet introduced the "self" into the immunological lexicon (Burnet and Fenner 1949), and upon that metaphor erected a theory of immunological tolerance that still dominates the field.

"Tolerance" refers to the immune system's "silence" to potential targets of destruction, thus allowing host constituents and some foreign elements an adopted co-equal status within the organism. Tolerance and autoimmunity are two sides of the proverbial coin: In one instance, the immune system is seen to ignore the host, and even foreign components, while in the other modality, the immune system attacks what is regarded by the outside observer as "self." These findings challenged the notion of a "one-directional" schema of immune reactivity, for tolerance was shown to be more than a passive silence of immune function, but required a more complex balance of responses. By the 1990s, immunologists increasingly appreciated that an immune self, representing a fortress from which attacking lymphocytes might sally forth to destroy invaders, offered a naive depiction of what was, in fact, a dynamic equilibrium in which "attacked" and "tolerated" were not easily predicated.

The simple model of immunity as committed to discerning those mechanisms by which the "self" discriminates host elements from the foreign requires revision. No longer is the identity of the host organism given or assumed, and, indeed, immune selfhood embraces diverse definitions. The designation of "self" and the "other" ignores that such neat divisions or boundaries were adopted, or at best, were drawn with a certainty that remained problematic. In fact, early discrepancies accompanying the full embrace of a self/non-self discriminatory mode to explain immune function remain vexing. So, while in Burnet's original formulation, the host organism, perceiving an invasion by microbial pathogens, mounts a defensive response, contemporary immunology has broadened this agenda to include surveillance of the body for malignant, effete, damaged, or dead host constituents (altered "normal" cells), as well as auto-immune processes directed against undamaged elements—some of which may be part of ordinary physiological economy, while others are pathological. The challenges to define a basis for immune identity, within the coupled ambiguities of autoimmunity and tolerance, has consequently generated debate about selfhood as an organizing concept for the discipline.

So when immunology is summarily defined as the science of self/non-self discrimination, and Burnet's theory by which selfhood is currently understood, "with only slight modification…has passed from the

status of theory to that of paradigm" (Golub and Green 1991, p. 15) and "no longer a theory but a fact" (Klein 1990, p. 335), a vast body of experimental data and explanation is ignored. Despite such dogma, the immune self, an implicit entity in the late 19th century (Tauber and Chernyak 1991; Mazumdar 1995), has became a hotly contested one today (Langman 2000), and offers a rich philosophical topic, both in terms of its epistemological standing, as well as its metaphysical foundations (Tauber 1994; 1999).

This article will outline, in a historical analysis, the two principle theories governing immunology's research program—the theory of immune identity, and a more recent one that challenges the very notion of selfhood. (Critics of this reading [Cohn 1998a; 1998b; Howes 1998] have been answered elsewhere [Tauber 1998b; 1999; 2000]). In those theoretical constructions are reflected the prevailing attempts to define the concept of organism. Note, while the immune self is rooted historically in the problematics of biological individuality (Loeb 1945; Buss 1987), its philosophical attention is distinct from those concerns (Wilson 1999) and subsumed to the broader questions of reductionism.

Historical Antecedents

The first medical use of the term "immunity" (originally a legal designation conferring exemption and distinction) appears in 1775, when Van Sweiten, a Dutch physician, used "immunitas" to describe the effects induced by an early attempt at variolization (Moulin 1991, p. 24). But the concept was not developed until the mid-19th century, when Claude Bernard set the theoretical stage for the autonomous organism (E. Cohen 2001). In contradistinction to an animal in humoral balance with a pervasive environment, Bernard postulated the primacy of the organism's essential independence. Physiology became the mode of inquiry for medical experimentation, one that instantiated a reductive strategy based on positivist principles. Later, biochemistry and genetics pursued this methodological and theoretical approach, thereby providing medicine with its modern experimental basis.

Bernard furnished biology with a new concept of the organism, one which would have wider ramifications than the establishment of a scientific method. Obviously, interchange with the environment was a necessary requirement for life, but Bernard emphasized how boundaries provided the crucial metabolic limits required for normal physiological function. With his concept of the *milieu interieur*, the body was envisioned as a demarcated, inter-dependent yet autonomous entity ("corporeal atomism" [E. Cohen 2001, p. 190]), thereby establishing the theoretical grounding that became the *sine qua non* for the development of the models for infectious diseases, genetics, neurosciences, and immunology in all of its various guises. He thus introduced a revolutionary approach to the study of the organism, and immunology became one of its defining sciences, indeed, immunity was alien to the older humoral view.

> Given the inclusive and fluid metaphoric system underlying pre-modern medicine, to speak of "immunity" with respect to embodied states would not only be improper but nonsensical. If disease signified a relation among elemental qualities and humors that were materially constitutive of both the living organism and its life context, then "exemption from" on the model of juridico-political immunity would be a non sequitur at best. (E. Cohen 2001, p. 183)

By radically changing the inside/outside topology so that the organism's interior becomes the determining context of function, Bernard effectively isolated the organism from its environment, and joined a complex cultural movement of redefining the body more generally.

Bernard's notion of the body as independent of the environment complemented Malthusian economics, liberal political philosophy, and Comtian sociology. From these and other disparate sources, the autonomous body as a political, social, economic and medical entity was redefined in the 19th century (Foucault 1973; Agamben 1998), and Bernard played a central role in providing a theoretical biological foundation for its critical use in various discourses. Notwithstanding that “independence” is a political term, and neither fairly represents the dialectical relationships of the organism and its environment (Levins and Lewontin 1985), nor the evolutionary peculiarities of individuality itself (Buss 1987), the formulation has served as the touchstone for various cultural constructions of identity. Indeed, culture critics have seized on immunology as paradigmatic for the modern notions of identity, where boundaries are contested and the body becomes the localized site of battle between self and other (Haraway 1989, Martin 1994). The warfare metaphors—“attack,” “defense,” “invaders”—so prevalent in immunology's lexicon, dramatically illustrate this construction, both in terms of the self/other dichotomy, as well as the privileged regard of individuality over community.

Origins of the Immune Self

Immunology's history is generally regarded as intimately tied to those discoveries leading to the elucidation of the bacterial etiology of infectious diseases, which draws together twin disciplines—microbiology, the study of the offenders, and immunology, the examination of host defense. Thus, in this pathological context, immunology began as the study of how a host animal reacts to pathogenic injury and defends itself against the deleterious effects from such microbial insult. This is the typical historical account of immunology as a clinical science, a tool of medicine, and as such it focused almost exclusively on the role of immunity as a defender of the infected. The paradigmatic host is the patient, an infected “self,” which is the critical element for the power of this view. The clinical orientation, which *assumes* a given entity—the self—is obviously a dominant organizing perspective, but another perspective turns this assumption into a question or a problem: Rather than the science that seeks to discern the basis of self/non-self discrimination, immunology may also be regarded as more fundamentally concerned with the *establishment* of organismal identity.

This latter point of view was offered by Elie Metchnikoff, who came to the nascent field of immunology from an unexpected theoretical and methodological perspective—an embryologist, who sought to discover genealogical relationships in the context of Darwinism (Tauber and Chernyak 1991; Gourko, Williamson, and Tauber 2000). Intrigued with the problem of how divergent cell lineages were integrated into a coherent, functioning organism, Metchnikoff was thus preoccupied with the problems of development as process, which he regarded as analogous to Darwinian inter-species struggle: Cell lineages were inherently in conflict to establish their own hegemony, but unlike nature writ large he hypothesized that a regulatory system was required to impose order, or what he called “harmony” on the

disharmonious elements of the animal. He found such an agent in the phagocyte, which retained its ancient phylogenetic eating function, to devour effete, dead, or injured cells that violated the phagocyte's sense of organismal identity. When pathogenic microbes were discovered in the 1870s, Metchnikoff soon applied the phagocyte the new role of defending the organism against invaders. Indeed, on this view, the phagocyte became an exemplar combatant of Darwinian struggle, now occurring within the organism.

In Metchnikoff's theory, immunity was a particular case of physiological inflammation, a normal process of animal economy. But there was a more subtle message: 1) immunity was an active process with the phagocyte's response seemingly mounted with a sense of independent arbitration, and 2) organismal identity was a problem bequeathed from a Darwinian perspective that placed all life in an evolutionary context. In short, Metchnikoff combined a Darwinian sensibility to a Bernardian conceptualization of autonomy.

Metchnikoff's overall representation constituted the phagocyte as an *agent* (Crist and Tauber 2001), an actor that is the cause of its own action—as a matter of endogenously generated and directed behaviors. The portrayal of the phagocyte as autonomous is largely derivative from the linked features of its capacity to sense its environment and move freely within it, and the various degrees of unpredictability and meaningfulness that characterize this behavior. The play of these features assemble an entity that is irreducibly the center of its own actions, one seen as analogous to the more complex organism with multiple functions: sensibility, locomotion, engulfment, ingestion, digestion, and excretion. Indeed, the phagocyte, as an agent, becomes a metaphorical “self,” a primordial microcosmic expression of what later immunologists would extend into an epistemology of biological identity. But while placing the identity function at the core of immunology's concerns, Metchnikoff failed to provide the necessary pre-conditions for those who would seek to demonstrate those reactions that conferred protection of such an *entity*. Much of the subsequent history of immunology may be traced to the attempts of establishing a definition of organismal identity and providing an experimental basis that would describe identity-making functions. These were scientific aspirations 19th-century biology could not fulfill.

Twentieth Century Constructions of the Immune Self

The first half of 20th-century immunology was devoted to establishing the chemical basis of immunity, leaving the parameters of selfhood tacit and assumed (Silverstein 1989). This chemical perspective dominated immunology until shortly after World War II, when transplantation and autoimmunity became increasingly relevant both to basic immunologists and clinicians. It was at this juncture that Sir Frank Macfarlane Burnet formally introduced the “self” into the immunological lexicon, and upon that metaphor erected a theory of immunological tolerance that was to henceforth dominate the field (Burnet and Fenner 1949; Tauber 1994). From this perspective, the foreign is destroyed by immune cells and their products, whereas the normal constituents of the animal are ignored. In other words, the host organism was a given identity within the Bernardian construct, one with implicit boundaries as defined by immune reactivity. What was “attacked” was “other;” that which was regarded by immune silence

became “the self.”

As currently understood, "self" and "nonself" may be discerned by either arm of the immune system. The more phylogenetically ancient phagocyte and its associated recognition proteins (together comprising the components of 'innate immunity') rely on at least three strategies to distinguish the host from "other" (Medzhitov and Janeway, 2002): 1) Recognizing "microbial nonself" depends on the ability of the host to identify conserved products of microbial metabolism that are unique to pathogens and are not produced by the host animal. These invariant structures are referred to as pathogen-associated molecular patterns, and their recognition plays a crucial role in host defense against bacteria. 2) Identifying "missing self" relies on the detection of "markers of normal self" which are dedicated gene products and metabolic products unique to the host. Missing such markers may initiate immune destruction, whereas recognizing such normal self-markers requires that immune reactivity is coupled to inhibitory pathways of immune activation. (Interestingly, certain microbes have assumed such markers by horizontal gene transfer to encode self-markers and thus avoid detection.) 3) Markers of "altered self" may be induced by infection and cellular transformation, so when such "neo-antigens" arise they become targets for immune destruction. This mechanism is the principle "house cleaning" function of phagocytes that ingest apoptotic (dying) and necrotic (dead) cells.

Historically, the innate system has not been the focus of immunology, and, correspondingly, the mechanisms of how the innate immune system distinguishes self and nonself have only recently been discerned. Immunologists have, instead, been preoccupied with lymphocyte biology, whose so-called mechanisms of "acquired immunity" are characterized by the ability to 1) mount an increasingly robust immune reaction once appropriate lymphocytes "learn" of pathogen insult, and 2) "remember" such insult so that upon repeated invasion, the lymphocyte-antibody response is both quickened and augmented. This was the immunology that intrigued Burnet, who was intent on explaining how immune reactivity develops in three stages - recognition, amplification, and memory - and more fundamentally, how self and nonself were discerned by this system. He invoked "tolerance" to explain how auto-destructive immune reactivity was controlled, or more specifically, he proposed a hypothesis that might explain how the immune system ignores host constituents. He thereby provided immunology with a theory of the self: Tolerance, the negative image of the self (or that which is absent in the space of immune recognition), became the central motif of understanding immune reactivity.

Unlike Metchnikoff, Burnet sought a firm definition of the immune self. Burnet's theory proposed that the animal, during prenatal development, exercised a purging function of self-reactive lymphocytes (the cells responsible for synthesizing reactive antibodies and mediating so-called cellular reactions) so that all antigens (substances that initiate immune responses) encountered during this period would attain a neutrality status. Thus lymphocytes with reactivity against host constituents are putatively destroyed during development, and only those “tolerant” lymphocytes that are non-reactive are left to engage the antigens of the foreign universe. Accordingly, potentially deleterious substances would select lymphocytes with high affinity for them, and through clonal amplification a population of lymphocytes differentiates and expands to combat the offending agents. The hypothesis (first presented in 1949 and later developed into the “clonal selection theory” (CST [Burnet 1959]) contained two key challenges which dominated contemporary immunology: 1) How was tolerance induced and auto-immunity

controlled? and 2) What was the mechanism that accounted for antibody and lymphocyte diversity? The latter issue was solved by molecular biologists by the mid-1980s (Podolsky and Tauber 1997); the former question, involving systems analysis, apparently requires a comprehensive model of the immune system as a whole and while theories of immune tolerance abound, the issue remains unresolved.

Aside from incomplete accounts of tolerance, there were early discrepancies arising from a continuum of auto-immune reactions, ranging from normal physiological and inflammatory processes to uncontrolled disease initiated by an immune reaction gone awry.

> During this century, the evolution of concepts on autoimmunity could be summarized by "never, sometimes, always." Thus from the early "horror autotoxicus" [Ehrlich] to the 1960s, immune autoreactivity was simply not considered…. With the first identification of autoreactive antibodies in patients and the subsequent conceptual association with autoaggressive immune behaviors, the "sometimes" phase was entered, necessarily equated with disease. By this time, immunology had laid its foundation on the clonal selection theory, which forbids autoreactive clones in normal individuals. Immunologists thereafter devoted 30 years discovering ways by which autoreactive lymphocyte clones can be deleted and why they fail to be deleted in autoimmune patients…. In the 1970s at least three sets of observations and ideas began to alter this course of events and to herald the "always" period. (Coutinho and Kazatchkine 1994, pp. 1-2)

Bountiful evidence in recent years has shown that autoimmunity is also a normal finding, and in these newer views, such functions are regarded as integrated within a more complex normal physiology (Schwartz and Cohen 2000; Horn et al 2001). Thus, immune reactivity, rather than functioning only in an "other-directed" mode is in fact bidirectional. This position contrasts with the "one-way" definition of selfhood, where there is a genetic self, whose constitutive agents see the foreign, and immune reactivity arises from this polarization with attack directed only against non-self (Tauber 1998a). Not unexpectedly, in this turn inwards, the immune self becomes increasingly difficult to define, unable to accommodate these new appraisals easily. There are at least half a dozen different conceptions of what constitutes the immune self (Matzinger 1994, p.993): 1) everything encoded by the genome; 2) everything under the skin including/excluding immune "privileged" sites; 3) the set of peptides complexed with T-lymphocyte antigen-presenting complexes of which various sub-sets vie for inclusion; 4) cell surface and soluble molecules of B-lymphocytes; 5) a set of bodily proteins that exist above a certain concentration; 6) the immune network itself, variously conceived (detailed below). While these versions may be situated along a continuum between a severe genetic reductionism and complex organismal constructions (Tauber 1998; 1999), each shares an unsettled relationship to Burnet's original dichotomous model of self and other (Langman 2000).

The Deconstruction of the Immune Self

Well before the current debate about the immune self, Niels Jerne attempted to dispel the many ad hoc caveats and paradoxes encumbering it by deconstructing the self concept altogether. He went beyond the

current notion of the immune network composed of lymphocyte subsets, secreting immuno-stimulatory and inhibitory substances (essentially a simple mechanical model with interlaced, first order feedback loops) to propose a novel conception of immune regulation (Jerne 1974). His network theory was, from its very inception, a complex amalgam of fitting the pieces of the regulation puzzle in place, with an overriding desire of understanding the immune system as a cognitive enterprise, one that spawned different formulations (e.g., Varela et al 1988; Atlan and Cohen 1989; Stewart 1994a). In introducing this metaphoric construction of the immune system as analogous to the nervous system as early as 1960, Jerne set the stage for understanding newer immune metaphors—recognition, memory, learning—which built on that parallel with human cognition.

Jerne's idiotypic network theory hypothesis proposed that antibodies formed a highly complex interwoven system, where the various specificities “referred” to each other (Jerne 1974). Under the general rubric of “cognition,” he conceived of the immune system as self-regulating, where antibody not only recognizes foreign antigen, but is capable of recognizing self constituents as antigens (the so-called idiotopes). There was no essential difference between the “recognized” and the “recognizer,” since any given antibody might serve either, or both, functions. In other words, immune regulation was based on the reactivity of antibody (and later lymphocytes) with its own repertoire forming a set of self-reactive, self-reflective, self-defining immune activities. Strong experimental support notwithstanding (Horn et al 2001), the relative importance of Jerne's network compared to other systems models remains contested, not the least for it radical reformulation of immune identity.

According to Jerne's model, the “self” and “other” dichotomy collapses, for the system is complete unto itself. Consisting of interlocking recognizing units, each component reacts with certain other constituents to form a complex network. When the system is perturbed by the introduction of a substance that is “recognized” (i.e., it reacts with a members of the system), this disturbance initiates the immune response. Thus foreignness per se does not exist in this formulation. In short, the system “knows” only itself. In Burnet's simplified world of self/non-self discrimination, the immune system learned host/foreign distinctions, generated an army of reactive antibody and lymphocytes, and acted accordingly when “antigen” was encountered. But Jerne coupled the simple antibody-antigen interactions to the far more complex and non-discriminatory functions of the immune system that built upon self-recognition. On his view, “autoimmunity,” instead of an aberrancy, became the organizational rule to explain immune function. Strikingly, there is no explicit mechanism for self/non-self discrimination, and this apparent lacuna served as the nexus of critiques (reviewed in Podolsky and Tauber 1997; Tauber 1999; 2000). But for Jerne, the need to define the “self” as distinct from the “other” receded from his primary theoretical concerns, and this posture was to have important repercussions.

When the immune system is regarded as essentially self-reactive and interconnected, the “meaning” of immunogenicity, that is reactivity, must be sought in some larger framework. Antigenicity then is only a question of degree, where “self” evokes one kind of response, and the “foreign” another, based not on its intrinsic foreignness, but rather because the immune system sees that foreign antigen in the context of invasion or degeneracy. There is no foreignness per se, because if a substance was truly foreign, it would not be recognized, i.e., there would be no image by which the immune system might engage it. So in the Jernian network, “foreign” is defined as perturbation of the system above a certain threshold. Only as

observers do we designate “self” and “non-self.” From the immune system's perspective, it only “knows” itself (Varela et al 1988). In this scheme, the immune system both disqualifies and abdicates any responsibility for discriminating “self” and “other.” Indeed, for Jerne, if one “needed” a self, it was the immune system itself. Most importantly, the singular defensive purpose of immunity was widened to include an array of physiological functions, each of them now regarded as fully integrated within the immune system itself (Matzinger 1994; Anderson and Matzinger 2000a; 2000b). If eventually successful, this move heralds a decisive shift in immunology's theoretical foundations, one more attuned to the diversity of immune functions which contribute to evolutionary fitness (I. Cohen 1992; 1994; Stewart 1994a). While host defense is a critical function, it is hardly the only one of interest. Indeed, the immune system might be regarded as primarily fulfilling an altogether different role if its phylogeny is carefully examined. On this basis, John Stewart has provocatively suggested that the immune system became defensive only after its primordial neuroendocrine communicative capabilities (Rabin, 1999; Ader, Felton, and Cohen 2001) were usurped for ‘immunity’ (Stewart 1994b).

In this spirit, Irun Cohen, and other contemporary theorists, refer to an “immune dialogue,” where the immune system continuously exchanges molecular signals with its interlocutor, the body (I. Cohen 1992; 1994). This “contextualist sensibility” highlights attention to complex systems that function in a self-organizing, dialectical interchange within itself (however its boundaries are drawn) *and* with its “outside” world (Levins and Lewontin 1985). No longer content with only defining the elements of the system and the local interactions of those components, biologists have increasingly come to appreciate that such systems are highly integrated within larger wholes and require analysis of how adjustments are made in relation to these other systems. This means, simply, that immune reactivity is determined by context (I. Cohen 1994; Podolsky and Tauber 1997; Grossman and Paul 2000), where agent and object play upon each other. In other words, as applied to the problem of self/non-self discrimination, from this ecological perspective, there can be no circumscribed, self-defined entity that is designated the Self, but rather there is an organism that is under constant challenge to respond along a continuum of behaviors, and it adapts and changes accordingly. In the case of the immune system, reactivity may vary from a full fledged immune response to mild irritation to quiescence.

Powerful molecular support for this contextual (or in another sense, ecological) orientation has been gathered. Consider the dominant model concerning lymphocyte activation, where it is generally appreciated that specific recognition of antigen by a lymphocyte receptor is not sufficient for activation, and that additional signals determine whether a cellular response or cell inactivation follows. In short, an antigen is neither self nor non-self except as it attains its “meaning” within a broader construct. Orthodox immune theory encompasses this idea in the so-called “two-signal model,” which does not require any of Jerne's hypotheses to fulfill its agenda. But there are more radical readings of the “contextualist” setting by which antigens are sensed, and debate concerning what constitutes the milieu of “meaning” of antigenicity and ensuing reaction have spawned certain provocative, and potentially important models of immune regulation (reviewed in Podolsky and Tauber 1997; Tauber 2000).

But Why Does the Immune Self Linger?

If we look at the “big picture,” as a chapter of biology, immunology is, on the one hand, adjusting to the twin demands of increasing molecular elucidation, and, on the other hand, an “ecological” sensibility. In both contexts, the “self” has slipped into an archaic formulation: From the molecularists' perspective, atomic delineations have outstripped explanations of immune regulation so that no molecular “signature” of selfhood suffices to explain the complex interactions of immunocytes, their regulatory products, and the targets of their actions. Reactivity has become the functional definition of immune identity. But when non-reactivity occurs, this may be because of active or implicit tolerance, which in turn is determined by many factors beyond Burnet's original formulation. Indeed, a new metaphor, “danger,” has been introduced to account for the integration of the immune system into the body as a whole, so that immune reactivity is regarded as determined not by a police function arbitrating self and non-self, but rather as a response to repair damage and defend against further deleterious agents of any kind—microbial, chemical, mechanical, etc. (Matzinger 1994).

When perceived as an attack on the centrality of self/non-self discrimination, much controversy has ensued (e.g., Langman 2000). While some detractors have generously called for a pluralistic approach (Vance 2000), and others have regarded the crisis over the self as overblown (e.g., Silverstein and Rose 2000), most would agree, at the very least, that immune selfhood is increasingly a polymorphous and ill-defined construct. Contemporary transplantation biology and autoimmunity have demonstrated phenomena that fail to allow faithful adherence to a strict dichotomy of self/non-self discrimination (Horn et al, 2001), and as new models are emerging, the immune self has been left exposed as a metaphor, whose grounding is unsteady and thus increasingly elusive as the putative nexus of immunology's doctrines. Quite simply, the immune system, now regarded as fully integrated with all the systems of the organism, no longer is seen as exclusively serving a separate “policing” function, one that protects *a* Self. Instead of searching for criteria of “self” and “other,” immune responses are increasingly studied as arising within a complex context which determines reactivity or dormancy. Self/non-self discrimination recedes as a governing principle when immunity is appreciated as both “outer-directed” against the deleterious, and “inner-directed” in an on-going communicative system of internal homeostasis. From this dual perspective, immune function falls on a continuum of reactivity, where the character of the immune “object” is determined by the context in which it appears, not its character as “foreign” per se.

Central to the context question, immunology must successfully integrate two, hitherto conceptually separate systems of inflammation. Virtually all attention regarding immune identity has been paid to the lymphocyte and its product, antibody. But there is an older phylogenetic system of immunity, the so-called “innate system” (as opposed to the “acquired” immunity of lymphocyte/antibody reactivity), which employs an ancient protein attack complex—complement (named originally as complementary to antibody)—and lectin proteins that together serve as opsonizing (coating) recognition proteins for phagocytes. This innate system is, in fact, the first line of defense against invading pathogens, and while it lacks the exquisite specificity of the lymphocyte system, the phagocyte with its attendant co-factors readily recognizes bacteria, some viruses, fungi, and protozoans, and also responds to non-specific damage to body tissues. The interesting issue for this discussion is that the innate system of immunity lacks the ability to distinguish self from non-self in the terms defined for lymphocyte biology, yet a second (or perhaps third) signal is required from these non-distinguishing antigen-presenting phagocytes

to activate lymphocytes. And there lies the rub. How are the two systems integrated to account for identity discrimination? The “danger” theorists maintain that these second signals must arise from such non-specifying sources so that the “self” concept essentially deconstructs, while the protectors of the self concept argue that the immune system simply needs to broaden its scope to integrate such “non-discriminatory” signals into a more discriminating system. In either case, a new theoretical consensus seems to be emerging as the “self” is either bypassed or expanded beyond its original formulation. (For review of contending theories see Anderson and Matzinger 2000b)

These developments continue a trajectory of two major theoretical developments: Originally, Metchnikoff regarded immunology as effecting dual functions: first, establishing organismal identity, and then protecting its integrity. His immunochemical contemporaries and their direct heirs followed the second agenda to the exclusion of the first. The primacy of the identity issue was re-introduced by Burnet, and his program defined immunology for the latter half of the 20th century. The second theoretical advance was made by Jerne, who moved past the identity issue altogether. No longer in service to a “self,” on his view the immune system functioned within a greater whole as a cognitive faculty, perceiving only what it might know—*itself*. Jerne thus introduced, perhaps ironically, a revision of the self metaphor, not its final elimination. For him, patterns, context, and interlocution become organizing principles, so that the self, assuming a Jernian perspective, is eclipsed by another catch-all metaphor, *cognition*, a direct descendent of the self concept, which itself readily lends to the scientific dictionary a host of meanings borrowed from other human experience (Tauber 1994; 1997; 1999). Without pursuing the ramifications of the cognitive approach to immunity, it is still evident that this turn in the language—“perception,” “memory,” “learning”—are in service to a more elusive “knowing entity.” Thus hidden within new formulations, the self still resides, reflecting a deep struggle over the character of biology, one that has its roots in Bernard's original understanding of autonomy, and now linked to our own more complex ecological views of agency and determinism.

On balance, notwithstanding the weakness of the metaphor, “self” maintains important uses despite its indistinct borders. Arguably, its elusive character is crucial to its utility and evocative power. But there is a deeper concern with the self's philosophical underpinnings, so that beyond the scientific rationale, the history of the concept of self suggests that the metaphor has specific uses not readily applied to the investigation of immune phenomena. Charles Taylor's description of “the punctual self” clarifies this issue (1989):

> John Locke attempted to construct “the self” as an autonomous legal unit to fulfill certain seventeenth-century liberal political goals. To do so he extrapolated from a philosophical invention. The “punctual” self was part of the early modern scientific conceit that regarded the knowing subject as totally divorced from the world in order to attain objectivity. The self itself also became a subject of scrutiny, so that ordinary experience might be seen from afar, objectified, and thereby controlled. He thereby reified the mind to an extraordinary degree, adopting a kind of atomism. This construction of the thinking subject supported a radical disengagement of oneself with a view toward remaking the world: “the real self is ‘extensionless;’ it is nowhere but in the power to fix things as objects” (Taylor 1989, p. 172). This power reposes in consciousness, a theme traced to Hume and later to William

James (Tauber 1994, pp. 207-15) and the phenomenologists (ibid., pp. 215-29).

Taylor argues that the self, as an object (or entity) must fulfill certain criteria: it must have objective status, standing independent of any description or interpretation, and be capturable in explicit description. Most importantly, an object can and must be understood without reference to its surroundings or contingent circumstances (Taylor 1989, pp. 34-5). The self of Locke and David Hume cannot fulfill these criteria. Taylor makes the critical point that “the self” is the answer to the question, “Who am I?”, which requires answers totally dependent on cultural or moral contexts, frameworks, or orientation—human categories of personal and social action, of value. Epistemological criteria are obviously operative, but the question of personal identity is a *moral* issue, not an epistemological one. *The self, then, is a moral description or category, one that fulfills criteria of human identification.*

Thus the “punctual” or “detached” self arose from two tributaries: the agent of a moral philosophy in post-Reformation England and the scientific actor who would scrutinize the world, and himself, apart from any constitutive concerns: The self's “only constitutive property is self-awareness. This is the self Hume set out to find and, predictably, failed to find” (Taylor 1989, p. 50). This identity over time is a construction that simply fails to fulfill the essential criteria of an entity given above. Most importantly, defining personal identity, the “Who am I?” question, can only be answered in the context of social and moral concerns. Context-determined, the self simply cannot be reduced to some single psychological construction of continuity, not only because self-consciousness cannot be captured or defined, but more fundamentally, as Hume himself realized, this was only a psychological conceit, a vague awareness of an identity function held together by psychological reflection. It was a useful invention, but it had no basis that could be defined by any analytical criteria.

The problematic status of “the self” in ascribing human agency perhaps should have alerted immunologists to the limits of its application to their science. Just as humans are loosely understood as selves, so too might scientists refer to immune identity in a similarly vulgar sense. There has been, to be sure, metaphorical utility in the term, but scientists seek to move from metaphor to more concrete and precise definitions. If theory is based on phenomena that indeed may be described more objectively, then those terms of discussion are adopted, because the scientific lexicon seeks more precise reference to natural entities and their relationships. And if metaphors become distorting to the evidence, they are abandoned. On the view presented here, immunology's cooption of the “self” was, from the very beginning, restricted by the metaphor's vague meanings, and immunology must—and, indeed, is—now moving beyond its original models to embrace new theories and novel metaphors to build its evolving theoretical edifice. In its wake, the science leaves us to ponder the significance of its failure to define the self and the cultural implications of the attempt (Haraway 1989; Martin 1994).

Bibliography

- Ader, Robert, Felten, David L., and Cohen, Nicholas. (2001) *Psychoneuroimmunology, 3rd edition.* San Diego: Academic Press
- Agamben, Giorgio. (1998) *Homo Sacer: Sovereign Power and Bare Life*. Stanford: Stanford

University Press
- Anderson, C. C. and Matzinger, P. (2000a) “Danger: the view from the bottom of the Cliff,” *Seminars in Immunology* 12: 231-38
- -----. (2000b) “Anderson and Matzinger round 2,” *Seminars in Immunology* 12: 277-91
- Atlan, Henri and Cohen, Irun R. (eds.) (1989) *Theories of Immune Networks*. Berlin: Springer-Verlag
- Burnet, Frank Macfarlane. (1959) *The Clonal Selection Theory of Acquired Immunity*. Nashville: Vanderbilt University Press
- Burnet, Frank Macfarlane and Fenner, Frank. (1949) *The Production of Antibodies, 2nd edition.* Melbourne: Macmillan and Co.
- Buss, Leo. (1987) *The Evolution of Individuality*. Princeton: Princeton University Press.
- Cohen, Edward. (2001) “Figuring immunity: Towards the genealogy of a metaphor,” in *Singular Selves: Historical Issues and Contemporary Debates in Immunology*, edited by A-M Moulin and A. Cambrosio. Amsterdam: Elsevier, pp. 179-201
- Cohen, Irun R. (1992) “The cognitive paradigm and the immunological homunculus,” *Immunology Today* 13: 490-4
- -----. (1994) “Kadishman's tree, Escher's angels, and the immunological homunculus,” in *Autoimmunity: Physiology and Disease*, edited by Antonio Coutinho and Michel D. Kazatchkine. New York: Wiley-Liss, pp. 7-18
- Cohn, M. (1998a) “The self-nonself discrimination in the context of function,” *Theoretical Medicine and Bioethics* 19: 475-84
- -----. (1998b) “A reply to Tauber,” *Theoretical Medicine and Bioethics* 19: 495-504
- Coutinho, A., and Kazatchkine, M. (1994) “Autoimmunity today,” in *Autoimmunity: Physiology and Disease*, edited by A.Coutinho and M. Kazatchkine. New York: Willey Liss., pp.3-6
- Crist, Eileen and Tauber, Alfred I. (1997) “Debating humoral immunity and epistemology: The rivalry of the immunochemists Jules Bordet and Paul Ehrlich,” *Journal of the History of Biology*, 30: 321-356
- -----. (2001) “The phagocyte, the antibody, and agency: Contending turn-of-the-century approaches to immunity,” in *Singular Selves: Historical Issues and Contemporary Debates in Immunology*, edited by Anne Marie Moulin and Alberto Cambrosio. Amsterdam: Elsevier, pp. 115-39
- Foucault, Michel (1973) *The Birth of the Clinic: An Archaeology of Medical Perception*. New York: Vintage
- Golub, E.S. and Green, D.R. (1991) *Immunology, a Synthesis*, 2nd ed. Sunderland, MA: Sinauer.
- Gourko, H., Williamson, D.I., and Tauber, A.I. (eds.) (2000) *The Evolutionary Biology Papers of Elie Metchnikoff*. Dordrecht: Kluwer Academic Publishers
- Grossman, Z. and Paul, W.E. (2000) “Self-tolerance: context dependent tuning of T cell antigen recognition,” *Seminars in Immunology* 12: 197-203
- Haraway, Donna. (1989) “The biopolitics of postmodern bodies: Determinations of self in immune system discourse,” *Differences* 1: 3-43
- Horn, M. P., Lacroix-Desmazes, S., Stahl, D, et al. (2001) “Natural antibodies—benefits of recognizing ‘self,’” *Modern Aspects of Immunobiology* 1: 267-270
- Howes, M. (1998) “The Self of philosophy and the self of immunology,” *Perspectives in Biology*

and Medicine 42: 118-130

- Jerne, Niels K. (1974) "Towards a network theory of the immune system," *Annals of Institute Pasteur/Immunology* (Paris) 125C: 373-389
- Klein, J. (1990) *Immunology*. Boston and Oxford: Blackwell Scientific Publications
- Langman, Rodney. (ed.) (2000) *Self-Nonself Discrimination Revisited*. *Seminars in Immunology* 12: Issue no. 3
- Levins, Richard and Lewontin, Richard. (1985) *The Dialectical Biologist*. Cambridge: Harvard University Press
- Loeb, Leo. (1945) *The Biological Basis of Individuality*. Springfield: Thomas
- Martin, Emily. (1994) *Flexible Bodies. The Role of Immunity in American Culture from the Days of Polio to the Age of AIDS*. Boston: Beacon Press
- Matzinger, Polly. (1994) "Tolerance, danger, and the extended family," *Annual Review of Immunology* 12:991-1045
- Mazumdar, Pauline M. H. (1995) *Species and Specificity. An Interpretation of the History of Immunology*. Cambridge: Cambridge University Press
- Medzhitov, Rusian and Janeway, Charles A., Jr. (2002) "Decoding the pattern of self and nonself by the innate immune system." *Science* 296:298-300.
- Moulin, Anne Marie. (1991) *Le Dernier Langage de la Medicine: Histoire de l'Immunologie de Pasteur au Sida*. Paris: Presses Universitaires de France
- Podolsky, Scott H. and Tauber, Alfred I. (1997) *The Generation of Diversity: Clonal Selection Theory and the Rise of Molecular Immunology*. Cambridge: Harvard University Press
- Rabin, B.S. (1999) *The Connection: Stress, Immune Function, and Health*. Wiley-Liss
- Schwartz, M. and Cohen, I. R. (2000) "Autoimmunity can benefit self-maintenance" Immunology Today 21: 265-68
- Silverstein, Arthur. (1989) *A History of Immunology*. San Diego: Academic Press
- Silverstein A. and Rose, N. R. (2000) "There is only one immune system! The view from immunopathology," *Seminars in Immunology* 12: 173-78
- Stewart, John. (1994a) "Cognition without neurons: Adaptation, learning and memory in the immune system," CC AI 11: 7-30
- -----. (1994b) *The Primordial VRM System and the Evolution of Vertebrate Immunity*. Austin, TX: R. G. Landes.
- Tauber, Alfred I. (1994) *The Immune Self: Theory or Metaphor?* New York and Cambridge: Cambridge University Press
- -----. (1997) "Historical and philosophical perspectives on immune cognition," *Journal of the History of Biology* 30: 419-440
- -----. (1998a) "Conceptual shifts in immunology: Comments on the 'two-way paradigm,'" *Theoretical Medicine and Bioethics* 19: 457-473
- -----. (1998b) "Conceptual shifts in immunology: Response to Melvin Cohn: How Cohn's two-signal model was turned." *Theoretical Medicine and Bioethics* 19:485-94
- -----. (1999) "The elusive self: A case of category errors," *Perspectives in Biology and Medicine* 42: 459-74
- -----. (2000) "Moving beyond the immune self?" *Seminars in Immunology*. 12: 241-48
- Tauber, Alfred I. and Chernyak, Leon. (1991) *Metchnikoff and the Origins of Immunology: From*

Metaphor to Theory. New York and Oxford: Oxford University Press

- Taylor, C. (1989) *Sources of the Self. The Making of the Modern Identity*. Cambridge: Harvard University Press
- Vance, R. E. (2000) “Cutting edge commentary: A Copernican Revolution? Doubts about the danger theory,” *The Journal of Immunology* 165: 1725-8
- Varela, Francisco J., Coutinho, Antonio, Dupire, B., and Vaz, N.N. (1988) “Cognitive networks: Immune, neural, and otherwise,” in *Theoretical Immunology, Part Two*, edited by Alan S. Perelson, Redwood City: Addison-Wesley Publishing Co., pp. 359-75
- Wilson, Jack. (1999) *Biological Individuality. The Identity and Persistence of Living Entities*. Cambridge: Cambridge University Press

Other Internet Resources

- Understanding the Immune System, Developed by Lydia Schindler, Donna Kerrigan M.S., Jeanne Kelly (National Cancer Institute)
- An Overview of the Immune System, maintained by Stephanie Forrest (University of New Mexico)
- General Information about the immune system

Related Entries

biology, philosophy of | biology: notion of individual | biology: notion of self | feminism, topics: feminist perspectives on the self

Alfred I. Tauber
ait@bu.edu

First published: May 21, 2002
Content last modified: May 27, 2002

Death

Attempts to understand death and its ramifications have generated much controversy. In what follows we will examine six areas of debate. *First*, what constitutes a person's death? It is clear enough that people die when their lives end, but less clear what constitutes the ending of a person's life.

A *second* controversy is whether one or more of several arguments defeats the *harm thesis*, the claim that death harms the individual who dies. (Often, discussions of the harm thesis ask whether post-mortem events, as well as death, can harm the individual who dies.) These arguments are as follows: (1) The *no-self theory* rejects the very idea of a self; this is a challenge to the harm thesis, since, unless there are selves, there is nothing that death can harm. (2) According to the *comparative good objection*, saying that death is harmful would make sense only if saying that the dead are worse off than the living made sense, which it does not, because an individual who has died is nonexistent and hence does not have a level of well-being. (3) The *symmetry argument* claims that it is irrational to think death is bad for us, because we do not think the nonexistence that preceded our births is bad for us, and when we compare this period of nonexistence to death, we see the two are mirror images, alike in all respects. (4) What might be called the *timing puzzle* is an attempt to impale the harm thesis on the horns of a dilemma. If death, or some post-mortem event, harms us, it does so before we die, or afterwards. The first option seems absurd, so the harm must occur while we are dead. But a person can be harmed only if caused to be in some sort of objectionable condition, and since the dead do not exist, they are in no condition at all. Either way, the harm thesis is false.

A *third* controversy concerns attempts to show that even if the dead cannot be harmed, the harm thesis is correct, since death, and some post-mortem events, harm the living. That is, there is something in the way death affects an individual who is not yet dead that constitutes harm. Admittedly, the idea that death or a post-mortem event harms the living may appear mysterious, but only if we operate with two overly narrow conceptions. First, we may think, mistakenly, that unless one thing *A*, has a causal impact on another *B*, *A* cannot affect *B* at all. This overlooks the possibility that *A* might affect *B* by influencing what is true of *B*. This is the way people who care about how they are thought of can be affected for the worse by posthumous events that destroy their reputations: these events have no causal impact on them, but they make it true of them that their desire always to be thought well of will be thwarted. Death, also, can affect us, by making it true of us that many of our desires will be thwarted. A second overly narrow conception we might have concerns harm. Everyone acknowledges that an event is harmful if it causes the presence of a bad condition of some sort, such as a wound. But there is another kind of harm: an event is sometimes harmful because it causes the absence of a good condition of some sort. When our teacher is killed, we lose knowledge we would otherwise have gained. Some theorists argue that all harms reduce to the first sort (so that, if the loss of a teacher is harmful, that is because it leaves us in a bad -- in effect

‘wounded’ -- condition), and reject the harm thesis on the grounds that death and posthumous events cannot cause wound-type harms. Other theorists counter that both kinds of harm exist, and death can harm us in the second way, by precluding our achieving all sorts of goods, yet leaving no ‘wounds’.

Assuming that the harm thesis is correct, a *fourth* controversy arises, concerning the specific nature of the harm death and various post-mortem events do, and whether such harms constitute misfortune. Presumably, these events harm us by putting certain goods out of our reach, but we are not always harmed by states of affairs that block our access to goods. My not having a magic lamp blocks me from getting three wishes, but it would be silly to say that I am harmed by my lack of a lamp. As an approximation, we might say that an event or state of affairs harms me if it ensures that I will lack some good that, otherwise, I would have had, but this criterion is open to objections.

A *fifth* controversy concerns whether all deaths are misfortunes or only some. Of particular interest here is a dispute between Thomas Nagel, who says that death is always an evil, since continued life always makes good things accessible, and Bernard Williams, who argues that, while premature death is a misfortune, it is a good thing that we are not immortal, since we cannot continue to be who we are now and remain meaningfully attached to life forever.

A *final* controversy concerns whether or not the harmfulness of death can be reduced. It may be that, by adjusting our conception of the self, or the good life, and by altering our attitudes, we can reduce or eliminate the harm death can do to us. Indeed, the adaptation of our views and attitudes might be a way to reduce or eliminate the harmfulness of anything that might happen to us, as certain ancient theorists, such as Gautama and perhaps Epicurus suggested. But there is a case to be made that such efforts will backfire if taken to extremes.

---

1. What is Death?

The term ‘death’ is ambiguous. Undergoing a disease that undermines the ability to live is one thing, the ending of life is a second, and the condition of having life over is a third. The first of these is a process,

the second an event, while the third is a state or condition. ‘Death’ can refer to any or all of the three, and it is particularly easy to run the last two together. To avoid confusion, it helps to use the term *death* (or *dying*) for the event of life's ending, the term *being dead* for the state in which life has ended, and neither for the process of physiological decline that (unless halted) undermines life and leads to death.

‘Death’ is also unclear in at least two ways. First, the concept of life is not entirely clear, and to the extent that we are puzzled about what life entails, we will be puzzled about what is entailed by the ending of life, that is, death (Feldman 1992). Second, it seems somewhat indeterminate whether a temporary absence of life suffices for death, or whether death entails a permanent loss of life. For practical purposes, whenever a creature loses life, or becomes nonexistent, the condition is permanent; so ‘death’, as commonly used, need not be sensitive to the distinction between temporary nonexistence and permanent nonexistence. But in thought experiments we can imagine the temporary loss of life and existence. Suppose, for example, that I were frozen and later revived, as is sometimes done to simple organisms: it is tempting to say that I cease to be alive, and cease to exist, while frozen in a state of suspended animation. Or imagine a futuristic device that reduces me to disconnected atoms which it stores and later reassembles just as they were before. Many of us will say that I would survive—my life would continue—after the reassembly, but it is quite clear that I would not exist during intervals when my atoms are stacked in storage. In these cases, our linguistic intuitions give no definitive verdict concerning the applicability of ‘death’. On the one hand, it seems appropriate to say that I die when my body is completely frozen or my atoms are disconnected, since the term ‘death’ seems applicable when a creature's life ceases. On the other hand it seems correct to deny that I die, since the cessation of my existence is only temporary given that my life is eventually restored, and ‘death’ seems applicable only when a creature is made permanently nonexistent. Nonetheless, once we allow our competing intuitions to work themselves out, we are likely to conclude that permanent nonexistence more fully captures what we mean by ‘death’; hence in what follows we may as well assume that death entails permanent nonexistence.

According to some religious traditions, people need not permanently cease to exist when their bodies break down. There are perhaps two main competing ideas about how an afterlife is possible. First, our physical demise could be merely temporary, since God might resurrect our bodies (restoring our mental life in doing so). Second, we might avoid even temporary nonexistence, assuming we are immaterial (nonphysical) souls who survive the demise of the body. Proponents of the first idea of the afterlife sometimes apply ‘death’ to the breakdown of bodies, and proponents of the second sometimes apply it to the soul’s departure from the body, but both groups presumably will also acknowledge that ‘death’ would apply to our permanent nonexistence (even though they would deny that such death is inevitable).

1.1 Transition v. Ending

Does our existence come to a final end when our bodies break down, or is there some sort of transition to an afterlife? As evidence for the latter, one might cite anecdotes by gravely ill people, who sometimes report out-of-body experiences, whereby they seem to be souls traveling outside of their bodies. But these data can be accommodated by the hypothesis that death is our complete annihilation, together with the claim that out-of-body experiences are misleading. In support of this alternative, we might cite the fact that these experiences can be produced pharmacologically (Blackmore 1993), in people who are perfectly

healthy. Unless one supposes that souls are knocked loose by drugs, only to wander back once the drugs wear off, such experiments suggest that out-of-body experiences are illusory, even when triggered in the brains of dying people. Moreover, scientists have never been able to detect souls, and many of the people who claim this ability have been proven frauds. Aside from such empirical considerations, there are further arguments for an afterlife in the philosophical and religious literature. However, most philosophers are skeptical about these arguments, and we must leave these aside. Hereinafter we will adopt the conservative assumption that death is annihilation, or permanent nonexistence.

1.2 Death and Identity

Even if immaterial souls do not exist, there is good reason not to identify the deaths of people with the deaths of their bodies. For you can survive the demise of parts of your body, and your body can survive while you do not. *You* die if and only if your identity is destroyed. Hence we can clarify what it is for a person to die only if we clarify what is essential to a person's identity (Green and Winkler 1980). This is a complicated matter, which we must leave largely unexamined. But a few points are in order.

First, theorists such as Derek Parfit (1984), building on the work of John Locke (1689), have made a strong case for the view that psychological attributes such as memories and character traits, which change gradually over time, are central to our identities (see the essays in Perry 1975). Two separate but related ideas of identity vie for our acceptance: *identity as connectedness* requires that one's psychological profile not change significantly over time if one is to remain the same person, while *identity as continuity* allows changes in one's profile so long as these are gradual. According to the first idea, we can gradually lose our identities; identity is a matter of degree, since we retain our psychological attributes in varying degrees. By the second idea, identity is all or nothing; we either remain the same person or we do not; either there is not more than a gradual change in our psychological profiles or there is. Hence if we think of identity as connectedness, we will conclude that death, too, can come in degrees, and becomes complete when our psychological profiles are greatly altered or destroyed. Thinking of identity as continuity will lead us to say that death is all or nothing -- that people live through gradual, but not sudden and drastic, psychological changes.

Second, it is important to distinguish between the concept of death and a criterion for death. The concept of death says what death *is*: the cessation of personal survival. A criterion for death, by contrast, lays out a condition by which a person's death may be determined. The traditional criterion for death says that you will be dead when your heart and lungs cease to function (not that death *is* cessation of respiration and cardiac functioning). A more recent criterion is brain death -- meaning the death of the entire brain -- since the brain is the seat of our psychological features. The brain death criterion is more accurate since, with modern technology, respiration and blood circulation can be maintained artificially even when the brain is dead. As things stand, authorities in the legal and medical context frequently rely on the brain death criterion (President's Commission, 1981). For example, tissues are not to be harvested from organ donors unless the entire brain is dead. But there is good reason to consider a person dead even if certain parts of the brain are still alive. Personality is most closely associated with the higher brain (the cerebral cortex). Unsurprisingly, then, there is increasing support for a *higher* brain criterion for death, according to which death occurs when the higher brain is no longer alive.

2. How is Death Not Harmful?

Typically, those who value life accept a view that might be called *the harm thesis:* annihilation is, at least sometimes, bad for those who die, and in this sense something that 'harms' them. It is important to know what to make of this thesis, since our response itself can be harmful. This might happen as follows: suppose that we love life, and reason that since it is good, more would be better. Our thoughts then turn to death, and we decide it is bad: the better life is, we think, the better more life would be, and the worse death is. At this point, we are in danger of condemning the human condition, which embraces life and death, on the grounds that it has a tragic side, namely death. It will help some if we remind ourselves that our situation also has a good side. Indeed, our condemnation of death is here based on the assumption that more life would be good. But such consolations are not for everyone. (They are unavailable if we crave immortality on the basis of demanding standards by which the only worthwhile projects are endless in duration, for then we will condemn the condition of mere mortals as tragic through and through, and may, as Unamuno (1913) points out, end up suicidal, fearing that the only life available is not worth having.) And a favorable assessment of life may be a limited consolation, since it leaves open the possibility that, viewing the human condition as a whole, the bad cancels much of the good. In any case it is grim enough to conclude that, given the harm thesis, the human condition has a tragic side. It is no wonder that theorists over the millennia have sought to defeat the harm thesis. Let us examine their efforts.

2.1 The No-self Challenge

The first challenge to the harm thesis confronts us if we object to death on the grounds that it takes away our existence. According to Gautama (563-483 B.C.), it is a mistake to say that death ends our existence. If he is correct, our objection is moot. But why does Gautama deny that death causes our nonexistence? It is not because there is an afterlife. Instead, his thought is that there is no self and never was. The notion of a self is defective, so it makes no sense to ask whether selves are annihilated, and those who fear annihilation are confused. The no-self view, in turn, Gautama rests on skepticism about the notion of souls or substances, thought of as changeless substrata underlying changing attributes.

This first challenge to the harm thesis is far from conclusive, however, since the notion of personal identity is not tied inextricably to any particular conception of substance. As John Locke pointed out, a criterion of identity based on psychological continuity floats free of any particular conception of the underlying material basis for our psychological attributes. Presumably, our psychological attributes, and hence our identities, depend on the brain, yet survive its material transformation.

2.2 The Comparative Good Objection

Consider a second challenge to the harm thesis. The claim that we are harmed by death seems to imply that we are worse off dead than alive. But being worse off dead seems to require that we have some level of well-being *while* dead which compares badly to our level of well-being while alive. However, the dead

do not have a level of well-being since they do not exist. Apparently we are in a muddle when we claim to be harmed by death.

This objection fails, since being worse off dead than alive need not imply that we have some level of well-being while dead. Perhaps when I say I am worse off dead than alive, I compare alternative ways my life might go, noting that some alternatives would be better than others, and judge that an endless life is the best fate of all for me. Perhaps I assess events in terms of their bearing on which alternative I end up with, as follows: when an event determines that the way my life shall go is inferior to the alternative ways it might have gone, it harms me. Relying on this approach, I conclude that any event, such as death, that ends my life is bad for me in the sense that it brings about one of the inferior alternatives. What I am saying, then, is that endless life is better for me than the briefer alternatives, not that my state while living endlessly is better than my state while nonexistent. Death affects me, not by placing me in some condition or another during some mysterious form of existence, but rather by limiting the duration of my life.

2.3 The Symmetry Argument

A third challenge to the harm thesis is an attempt to show that the state death puts us in, nonexistence, is not bad. According to the *symmetry argument*, posed by Lucretius, a follower of Epicurus, we can prove this to ourselves by thinking about our state before we were born:

> Look back at time … before our birth. In this way Nature holds before our eyes the mirror of our future after death. Is this so grim, so gloomy? (Lucretius 1951)

The idea is clear to a point: it is irrational to object to death, since we do not object to pre-natal nonexistence (the state of nonexistence that preceded our births), and the two are alike in all relevant respects, so that any objection to the one would apply to the other. However, Lucretius' argument admits of more than one interpretation, depending on whether it is supposed to address the event of death or the state of death (or both).

On the first interpretation, the death *event* is not bad, since the only thing we could hold against it is the fact that it is followed by our nonexistence, yet the latter is not objectionable, even to us, as is shown by the fact that we do not object to our nonexistence before birth. So understood, the symmetry argument is weak. Our complaint about death need not be that the state of nonexistence is ghastly. Instead, our complaint might be that death brings life, which is a good thing, to an end, and, all things being equal, what ends good things is bad. Notice that the mirror image of death is birth (or, more precisely, becoming existent), and the two affect us in very different ways: birth makes life possible; it starts a good thing going. Death makes life impossible; it brings a good thing to a close.

Perhaps Lucretius only meant to argue that the death *state* is not bad, since the only thing we could hold against the death state is that it is nonexistence, which is not really objectionable, as witness our attitude about pre-natal nonexistence. So interpreted, there is a kernel of truth in Lucretius' argument. Truly, our pre-natal nonexistence does not concern us much. But that is because pre-natal nonexistence is followed

by existence. Nor would we worry overly about post-natal nonexistence if it, too, were followed by existence. If we could move in and out of existence, say with the help of futuristic machines that could dismantle us, then rebuild us, molecule by molecule, after a period of nonexistence, we would not be overly upset about the intervening gaps, and, rather like hibernating bears, we might enjoy taking occasional breaks from life while the world gets more interesting. But undergoing temporary nonexistence is not the same as undergoing permanent nonexistence. Unlike the former, the latter entails *death* in the fullest sense. What is upsetting is the death that precedes post-natal nonexistence -- or, what comes to the same thing, the permanence of post-natal nonexistence -- not nonexistence per se.

There is another way to use considerations of symmetry to argue against the harm thesis: we want to die later, or not at all, because it is a way of extending life, but this attitude is irrational, Lucretius might say, since we do not want to be born earlier (we do not want to have always existed), which is also a way to extend life. As this argument suggests, we are more concerned about the indefinite *continuation* of our lives than about their indefinite *extension*. (Be careful when you rub the magic lamp: if you wish that your life be extended, the genie might make you older!) A life can be extended by adding to its future *or* to its past. Some of us might welcome the prospect of having lived a life stretching indefinitely into the past, given fortuitous circumstances. But we would prefer a life stretching indefinitely into the future.

Is it irrational to want future life more than past life? No; it is not surprising to find ourselves with no desire to extend life into the past, since the structure of the world permits life extension only into the future, and that is good enough. But what if life extension were possible in either direction? Would we still be indifferent about a lengthier past? And should our attitude about future life match our attitude about past life?

There should be a match if our interests and attitudes are limited in certain ways. If quantity of life is the only concern, a preference for future life is irrational. Similarly, the preference is irrational if our only concern is to maximize how much pleasure we experience over the course of our lives without regard to its temporal distribution. But our attitude is not that of the life- or pleasure-gourmand.

According to Parfit, we have a far-reaching bias extending to goods in general: we prefer that any good things, not just pleasures, be in our future, and that bad things, if they happen at all, be in our past. He argues that if we take this extensive bias for granted, we can explain why it is rational to deplore death more than we do our not having always existed: the former, not the latter, deprives us of good things in the future (he need not say that it is because it is in the past that we worry about the life-limiting event at the beginning of our lives less than the life-limiting event at the end). This preference for future goods is unfortunate, however, according to Parfit. If cultivated, the temporal insensitivity of the life- or pleasure-gourmand could lower our sensitivity to death: towards the end of life, we would find it unsettling that our supply of pleasures cannot be increased in the future, but we would be comforted by the pleasures we have accumulated.

Whether or not we have the extensive bias described by Parfit, it is true that the accumulation of life and pleasure, and the passive contemplation thereof, are not our only interests. We also have active, forward-looking goals and concerns. Engaging in such pursuits has its own value; for many of us, these pursuits,

and not passive interests, are central to our identities. However, we cannot make and pursue plans for our past. We must project our plans (our self-realization) into the future, which explains our forward bias. It is not irrational to prefer that our lives be extended into the future rather than the past, if for no other reason than this: only the former makes forward-looking pursuits possible. It is not irrational to prefer not to be at the end of our lives, unable to shape them further, and limited to reminiscing about days gone by.

Nevertheless, it does not follow that we should be *indifferent* about the extent of our pasts. Being in the grip of forward-looking pursuits is important, but these take time, and their historical development, which underlies our self-realization, is important, too. We have passive interests as well, which would prompt us to extend our pasts if the opportunity presented itself. If fated to die tomorrow, most of us would prefer to have a thousand years of glory behind us rather than fifty years of glory.

2.4 The Timing Puzzle: Death Cannot Affect Us

Another challenge arises when we look for the specific time during which we undergo the harm that death supposedly brings upon us. As Epicurus (341-270) says in his *Letter to Menoeceus*, there does not appear to be such a time:

> Death …, the most awful of evils, is nothing to us, seeing that, when we are, death is not come, and, when death is come, we are not.

His thought is that if death is bad for us, there must be a time when we are made worse off. Given that death follows immediately upon life, the harm must be incurred either while we are alive, or afterwards. Presumably it is not incurred while we are alive, since that implies that we undergo the harm before the death responsible for it occurs. But the alternative is to say that the harm is incurred after we are gone. Yet it is odd to say that nonexistent people can be harmed, for wouldn't that mean people have some sort of ghostly existence after their lives end, and that the condition of these ghosts can be worsened?

Epicurus focuses on death, but if his argument is good, it applies more generally, to include all events that follow death. Let us call something a *mortem event* if it takes place when we die or afterwards, so that death and every event that follows is a mortem event. Epicurus's position is that no mortem event can harm us.

Epicurus's argument can be interpreted in more than one way. The intent might be to show that no mortem event can *affect* us *at all*. This claim, together with the following *impact thesis*, implies that mortem events are harmless:

> An event harms us only if it somehow affects us at some time (the event may affect us well after it occurs).

Let us see if it is possible to show that mortem events do not affect us. Then we can try out a weaker thesis: that no mortem event can affect us *in a way that matters*. This weaker claim is easier to defend; in

all likelihood, it is what Epicurus had in mind, but the stronger claim is worth exploring.

To defend the view that mortem events do not affect us, we need to make some assumptions about when an event can affect us. To this end, let us adopt the *causal account of responsibility*:

a. An event (or state of affairs) can affect some subject (person or thing) *S* only by causally affecting *S* (the *causal impact only* thesis).
b. A subject *S* cannot be causally affected by an event while *S* is nonexistent (the *exist while affected* thesis).
c. A subject cannot be causally affected by an event before the event occurs (the *ban on backwards causation*).

From this account, it follows that a *post*-mortem event, such as the burning of one's corpse, cannot affect us after we are dead, since, by (a), to be affected is to be affected causally, but, by (b), nonexistent people cannot be causally affected by any event. It also follows that a post-mortem event cannot affect us while we are alive, given the ban on backwards causation. We might call this the *inertness of posthumous events* argument:

1. An event can affect us only by causally affecting us (the causal impact only thesis).
2. We cannot be causally affected by an event while we are nonexistent (the exist while affected thesis).
3. The dead do not exist.
4. So no posthumous event can affect us while we are dead.
5. We cannot be causally affected by an event before the event occurs (the ban on backwards causation).
6. So no posthumous event can affect us while we are alive.
7. So no posthumous event can affect us.

So far so good: no post-mortem event can affect us. However, there may still be a *mortem* event that can affect us: death. Of course, the thesis that we must exist to be affected rules out the possibility that death affects us *after* it occurs (after we are nonexistent). But it does not rule out the possibility that death affects us exactly *when* it occurs. Or does it?

Well, it does if death occurs only after we are nonexistent, as do post-mortem events. And some theorists have maintained that the event of death occurs on the nonexistence side of the boundary between our existence and nonexistence. For example, Feinberg (1984, p. 172) adopts this view in the following passage:

> Death is defined simply as the first moment of the subject's nonexistence, so it is not something that ever coexists with the dying person for the time required for it to have a directly harmful effect on him.

However, the idea that we die only after we are nonexistent is unacceptable, for the event of death is a transition from a state of life to a state of death, and it is absurd to say that the transition takes place only after we are gone. It is also absurd to say the transition is completed while we are still alive. Hence defining death as the first moment of our nonexistence, as Feinberg does, is no better than defining it as the last moment of our existence.

But is it reasonable to say that we are alive at least part of the time during which we undergo the transition from life to death? Yes, since death takes time, and we are fully alive when the transition of death begins, partially alive as it progresses, and not at all alive when it ends. We exist while it is under way, and are affected by it in a straightforward way: it makes us less and less alive, until finally we cease to be.

However, conceivably a death might be instantaneous, in the following way: we simply move from being wholly alive to being wholly dead, and no time passes between the two. Can *that* kind of transition affect us? Actually, it is hard to say, since this picture is puzzling in certain ways. For example, if we suppose that no time passes in between our existing and our not existing, it seems to follow that everything that happens occurs either while we exist or while we do not exist (or during a period of time combining the two). We are never in between, never in any condition in between (whether existence, nonexistence, or some mysterious state that is neither), and no events happen in between. So if death is an event, when does it occur? If death is both an event and a transition across the boundary between being wholly alive and wholly dead, don't we have to imagine it overlapping with the sequences of events on both sides of this boundary? But suppose that, in spite of such puzzles, we can make it clear that the causal effects of an instantaneous death occur entirely on the far side of the boundary between existence and nonexistence. Then according to the causal account of responsibility, instant death does not causally affect us when it occurs or at any time thereafter.

Let's review. Granting them some leeway, especially the assumption that only what has a causal impact on us affects us, Epicurians can show that no mortem event other than death can affect us, and that if death can affect us, it can do so only precisely at the time it occurs. But they lack a convincing argument against the possibility that death and some of its effects overlap in time; and hence they cannot prove that mortem events are harmless.

2.5 The Timing Puzzle Again: Death's Impact Is Harmless

Instead of trying to show that no mortem event can affect us at all, Epicureans can try to show that no mortem event can affect us *in a way that should matter to us.* To that end, they can assume that something can affect us in a way that is positive, negative or even neutral from the standpoint of value, only if it causes (or can cause) in us the presence of some salient condition. The new strategy calls for replacing the impact thesis with the *bad impact thesis*:

> An event harms us only if it is responsible for a bad condition's coming to be present in us at some time (the event and the condition's presence in us need not be simultaneous).

Some terminology will be helpful. If an event *E* is responsible for our being in a bad state, let us say that *E* is the *indirect* harm, while the bad state that *E* precipitates is the *direct* harm. Thus *E* indirectly harms us when it occurs, but directly harms us only when the bad state is brought about.

Proponents of the new thesis are committed to the old, but not vice versa. Something cannot affect us in a bad way unless it somehow affects us, but not all ways of affecting us involve the presence in us of some salient condition. Death affects us when it annihilates us, but it is not responsible for any condition's presence in us. For no condition can be present *in* us if there is no *us.* A condition cannot be present in us unless we exist. We can be directly harmed only if we exist--this claim is often called the *existence condition.* Yet we need not exist in order to be indirectly harmed: an event may indirectly harm us long before it has any direct impact on us; indirect harm may come even before we exist, as when someone times a bomb to go off 150 years later, killing everyone around.

Given the bad impact thesis, Epicureans can show that no mortem event can harm us by showing that no mortem event is responsible for any condition's presence in us. We might call this the *absent conditions argument*:

1. Death is responsible for our nonexistence and for the absence of conditions in us at the time it occurs and thereafter, but not for the presence of any conditions in us at any time.
2. Direct harms reduce to the presence of salient conditions in us, not in their absence (the bad impact thesis)
3. So death cannot harm us.

The point can be extended to post-mortem events: given the ban on reverse causation, and the thesis that we are unaffected while not existing, nothing that happens after we die can be responsible for any condition's being present in us at any time, so posthumous events are harmless.

Different Epicureans offer different accounts of the salient condition genuine harm reduces to. However, they seem to agree that direct harm is a kind of experience, and they offer versions of the *experience requirement*, which says that an event can harm us only if we experience it (or only if we can experience it, or only if we (can) experience it as bad). According to Epicurus's own view, the only things that are bad for an individual are things that cause that individual to suffer (this claim is the *painfulness criterion for harm*). Armed with this criterion, he offers the *argument from painlessness*:

1. The immediate result of death is the nonexistence of the self.
2. So dying is painless.
3. Nothing harms me unless it makes me suffer (the painfulness criterion for harm).
4. So dying cannot harm me.

The same goes for post-mortem events: I cannot suffer posthumously, so nothing that happens after I die harms me.

Epicurus goes on to warn against three confusions that might blind us to the force of his view. First, it is easy to confuse death with the dying process leading up to it, such as progressive cancer, and to hold our complaints about the one against the other. But Epicurus admits that the latter can be painful and hence bad for us. Is he then trying to remove our concern about death only to leave in place our concern about the dying process? That would be odd, just as it would be odd to remove objections to the state of death, while leaving intact our objections to the event of death (but see Rosenbaum 1986). The stated goal of Epicureanism is *ataraxia*, or tranquility of mind; this goal is thought to be attainable because, for the enlightened, nothing in life is harmful. From this perspective, it would not be useful to show that being dead is of no concern, while leaving us terrified at the prospect of death, or the dying process. Of the three -- death, being dead, and the dying process -- Epicurus admits only that the dying process can be bad, and even *it* is not *especially* bad. But this last Epicurus rests on the dubious claim that serious afflictions are not very painful:

> Continuous pain does not last long in the flesh; on the contrary, pain, if extreme, is present a very short time…. Illnesses of long duration even permit of an excess of pleasure over pain in the flesh (*Principal Doctrines*, Doctrine 4)

A second confused response results if we fail to distinguish pain caused by anticipating death with pain caused by death itself, and hold the former against the latter. *Anticipating* death upsets us and is, to that extent, a bad thing. However, our (present) anticipatory fear is not caused by our (future) death, since future events are powerless to affect the past. Moreover, fear is irrational unless its object is genuinely evil in some way, which death is not.

The third confusion arises if we do not distinguish what is bad for us from what is bad for others. At most, the fact that your family grieves your death supports the claim that your demise harms *them*, not that it harms you. (Too, your distress at anticipating your family's grief over your death is not grounds for you to regard your death as a bad thing: the suffering your death brings them cannot affect you, and your anticipatory grief is irrational.) Furthermore, their grief should be mitigated by the fact that your death is not bad for *you*. Their grief is entirely self-centered, exactly like the self-pity a gardener might feel at the loss of a familiar plot of land or pleasant flower.

(Would it be morally wrong to kill you, given Epicurus's painfulness criterion? Perhaps, but the moral case against killing is weak, given the fact that killing you harms you in no way, and, as a true Epicurean, you do not mind. But won't killing you displease others? Perhaps, but this reservation will not block the killing of pariahs -- or the complete annihilation of humanity.)

In sum, the Epicurean position comes to this: While one mortem event alone -- death -- can causally affect us (by annihilating us), no mortem event can harm us, since it can never cause in us the presence of any condition, and hence it can never cause in us a condition, such as the experience of pain, that qualifies as direct harm.

3. How Is Death Harmful?

Epicureans must be granted this much: all other things being equal, painless things, in being less frightening, are not as bad as painful things. If the immediate result of death is our nonexistence, and death happens too quickly to be experienced (which is far from obvious), then at least we can say that death and post-mortem events are painless, which makes them less frightening. However, even if we fear mortem events less, or not at all, since they are not experienced, should we conclude that they are not harmful? Let us see if the Epicureans' case holds up. As we shall see, even if we cannot be directly harmed while we are dead, it is possible to defend the harm thesis. We can start by looking for counterexamples to the bad impact thesis.

3.1 Harms that Wound Versus Harms that Deprive

One set of examples centers on the fact that most of us regret the severing of our interpersonal relationships and the thwarting of our aspirations. However, it is possible for these to be destroyed without our noticing. Suppose (Nozick 1971) that by means of an elaborate lie, an enemy convinces someone you love to hate you but to feign love so as to keep tabs on you for the rest of your life. Then you lose the love of your partner yet forever retain the appearance of love. Your loss produces no troubling experiences in you, but it is bad for you nonetheless. Or suppose (Nagel 1979) you are struck by an illness that instantly destroys your faculties and reduces you to the state of a contented infant. Here again is a tragic loss that is not accompanied by troubling experiences. A related case: it is bad to be raped after secretly being drugged into sleep, even though we cannot experience being raped while asleep. A second set of examples exploits the fact that we can experience, and suffer from, nothing that happens after we die, yet many post-mortem events are regrettable. For example, it is terrible to have your will set aside. And there are other examples. Suppose you found out that, starting in two weeks, your family and friends (or everyone in the world, for that matter) were to suffer horribly. But then you learn you are to die in one week, so their fate can have no causal impact on you. If you adopt the Epicurean's bad impact thesis, the fate of your loved ones, under such circumstances, is not cause for concern, and if you could do something now to prevent their suffering, it would be unimportant for you to do so. Most of us would be appalled by the Epicurean's indifference about all of these matters.

Epicureans are committed to denying that any of these examples involve genuine harm, given their view that (direct) harm consists in the *presence* of some condition that is bad for us, such as wounds or pain. Their view, which we might call the *present bad view of harm* or the *wound model of harm*, is that events are harmless -- unable to make us worse off -- unless they leave us in a condition analogous to being wounded. Our examples suggest that this is an overly narrow conception of harm. They suggest that the direct harmfulness of some events may consist in the *absence* of a salient *good*, such as the love of your partner, or the completion of your life's work, or the flourishing of your children. In fact, the salient good might be pleasure itself, so even hedonists can agree that the direct harmfulness of an event might consist in the absence of a good. Epicurus's own hedonist account is negative, in the sense that it restricts harm to the presence of pain; but nothing stops a hedonist from adopting a positive account, according to which an event may harm us by depriving us of pleasure. The idea that direct harms can consist in the absence of some salient good (but might also consist in the presence of a relevant bad condition) we might term the *absent good view of harm*, or the *deprivation account of harm*.

But if the absence of a salient good can constitute a harm, then death can be harmful, for there are two ways in which death can be responsible for harmfully absent goods. First, it destroys goods, such as good conditions present in us when we die (these are *destruction harms*). Second, it precludes our retaining goods or acquiring goods (*preclusion harms*). So the harm thesis is true after all.

However, proponents of the harm thesis still have work to do, for so far nothing they have said tells us *when* we incur the harms associated with absent goods. Now, it seems possible to pin down the time we incur the harms associated with death. Perhaps destruction harms and the death responsible for them occur simultaneously. And perhaps all preclusion harms associated with death can be treated similarly. However, the harm done by post-mortem events resists this treatment: it seems unfortunate if the executor of your will ignores your directives, but when are you made worse off? If we say you are harmed at the time your will is ignored, rather than before, we have to say that you can be made worse off after you are gone. It is as if, before and after you die, you persist in a harm-free condition until your evil relatives toss out your will, at which point you are placed in a state of being harmed, made all the more mysterious by the fact that you are long gone. That seems mistaken. But what are the alternatives? Perhaps, as Thomas Nagel suggests (1979), it is an indeterminate matter when some things, such as the violations of wills, or death, make us worse off, although they do so all the same.

Apparently, the absent good view of harm allows us to explain the harmfulness of death, but leaves us with questions about when posthumous events harm us. Should the air of mystery surrounding post-mortem harms prompt us to deny that they exist? That is one alternative open to the proponents of the harm thesis: They can give up on post-mortem harms and defend the harm thesis on the grounds that death is responsible for temporally locatable deprivation harms. They will have rejected only one of the three main pillars supporting the Epicureans' case against the harm thesis, namely the present bad view of harm. The remaining two pillars are the causal account of responsibility, and the assumption that the dead are in no sense real.

There are other alternatives. Each involves rejecting not just the present bad view of harm, but also one additional pillar of the Epicureans' argument. An option we shall call the *harmed dead view* rejects the assumption that the dead are not real. A further option, which we shall call the *noncausal harm view*, rejects the causal account of responsibility.

3.2 The Harmed Dead View

Suppose that in some sense people are real even after death annihilates them. This assumption may allow us to revive the possibility that post-mortem events can harm us *while* we are dead. But are annihilated people real in any sense? Perhaps; Silverstein (1980) argues that we can say that they exist in a timeless sense of existence, and Palle Yourgrau suggests that we speak of the dead, as well as the unborn, as objects, where an object has a kind of reality even if it does not exist. Doing so, Yourgrau thinks, allows us to say that “the deprivation of nonexistence endured by the unborn is as great as that suffered by the dead… ” (Yourgrau 1987, p. 149).

However, even if we can show that the dead are real in some sense, we will not be in a position to claim that the dead can be harmed unless we abandon the present bad view of harm. Absences are the only candidates for harm to people while they are nonexistent, for even if the dead are in some sense real, it is difficult to imagine a condition whose presence in them constitutes a state of harm. Being deprived of existence, as the unborn are, is itself at worst a harmful absence, and hence, by the bad impact thesis, no harm at all. We must also reject the experience requirement, which says that an event can harm us only if we experience it, for obviously the dead feel no pain. And if well-being is marked out in terms of the presence of some salient condition, it is difficult to see that the mode of reality possessed by the dead permit them to have a level of well-being. Hence if being harmed requires having a level of well-being, the dead cannot be harmed.

But if we do admit that goods deprivations can constitute harm, there is no further need to decide whether the dead can be harmed. As we shall see, we can defend the harm thesis on the basis of the deprivation account of harm together with other assumptions that hold even if the dead cannot be harmed.

3.3 The Noncausal Harm View

Undeniably, an air of mystery surrounds the idea that people who are annihilated come to be in a state of harm while dead. It would be desirable to pin down when post-mortem events harm us without assuming anything about the ontological status of the dead. Fortunately, there is a way. But we will have to reject one tenet of the causal account of responsibility, namely, the causal impact only thesis. Given the ban on backwards causation, the causal impact only thesis forces us to dismiss the idea that harm can occur *before* the event that precipitates it takes place. Yet, as George Pitcher (1984) says, this is precisely the idea we need in order to understand the harmfulness of a post-mortem event such as our will's being ignored. It is while you have a will, and the desires it expresses, that you are harmed. Pitcher's idea is that causation is not the only route via which things affect and thus harm others. The proposition, ‘You have a will which will be ignored’ is true now, even though it is true, in part, because of events that will occur after you and your desires cease to exist. These distant events harm you only *indirectly*. What *directly* harms you is *your having desires or interests that will be thwarted*, or *your having the potential to attain a certain good that will go unrealized*, which becomes true of you when you develop the desires or potential, and ceases to be true of you when they are gone. Perhaps (as the ban on backwards causation and the exist while affected thesis imply) we cannot be causally affected by events that will happen well after we are gone, yet we can be affected by these events in a straightforward non-causal way, because they help determine what is true of us now. It is not the dead who are harmed, but rather the living.

The idea of *non-causal* harm can be applied to death as well as post-mortem events. The verdict about death is that death can harm us *indirectly*, by being partly responsible for our having desires that will be thwarted, or potential that will go unrealized, in which case we are harmed *directly*, during such time as we have desires or potential that death prevents our attaining. Does this verdict force us to reassess our earlier suggestion that death directly harms us when it occurs? Not necessarily. For death (unlike posthumous events) might directly harm us twice: when it occurs and obliterates us, and when, because of death, it is true of us that we will not realize our potential.

There is another way to extend Pitcher's idea. We might object to the *state* of death because of its non-causal impact on us, since coming to be dead makes it true of us that we have desires that will be ignored. But instead of saying that being dead is objectionable, it seems better to say something else, once we notice that the state of death is simply the state of nonexistence initiated by the event of death. Perhaps being dead is powerless to harm us since any harm that might be associated with it is entailed in, and brought about by, death itself, which is responsible for limiting the duration of our lives, and all that that entails.

In sum, if we reject the present bad view of harm in favor of the absent good view, we can say that death is responsible for destruction harms and indirect preclusion harms at the moment it occurs, and for direct preclusion harms while we have the potential to acquire salient goods. We can also say that post-mortem events are responsible for direct and indirect preclusion harms.

4. What Is a Misfortune?

As we have seen, proponents of the harm thesis are committed to condemning a thing as bad when it deprives us of goods. Stated in this rough way, the *good-deprivation criterion* has considerable plausibility. Nonetheless, it requires development, and those who wish to refine it further will face three issues: First, should we adopt a subjectivist or an objectivist account of the good? Second, which of those goods an event or state of affairs precludes contributes to the harmfulness of that event? Third, how is harm related to misfortune?

4.1 Objective Versus Subjective

In clarifying what is good for a person, should our account be objectivist or subjectivist or some sort of mixture of the two? Nagel's well-known version of the good-deprivation criterion is objectivist. By contrast, subjectivists might say that things are good for us to the extent that they satisfy relevant desires (which desires these are must then be specified). Such theorists are likely to accept the *thwarted desire criterion* for misfortune, according to which something is bad for us insofar as it prevents us from satisfying relevant desires. To *satisfy* a desire here means to attain its object, not to gain pleasure from attaining its object, and to *thwart* a desire here means to block the attainment of its object, not to produce a feeling of frustration by such blocking. Bernard Williams is among those who defend this criterion. It was prefigured in the views of ancient Indian theorists -- for example, in Gautama's view that the cause of suffering is thwarted desire.

4.2 Harmfully Precluded Goods

The second question facing proponents of the good-deprivation view is: Which of those goods an event or state of affairs precludes (which of those desires an event thwarts) contributes to the harmfulness of that event? Not all of the goods an event puts out of reach would be attained or even accessible if the event did not occur. Losing my arms precludes my becoming a baseball star; yet, it would be odd to hold my failure

to attain stardom against my injury since I would not have attained stardom uninjured. So apparently the relevant goods are limited to those we would have enjoyed had the event not occurred. And this is Nagel's strategy (refined by McMahan 1988 and Feldman 1991): he wants to measure the harmfulness of mortality in terms of goods immortality brings us. The position is that an event or state of affairs *E* is harmful to me if and only if I would be worse off if *E* held than I would be had *E* not held, and that the degree of harmfulness of *E* is measured in terms of how much worse off I would be if *E* held than I would be had *E* not held. Given this counterfactual criterion for harm, a good *G* is relevant to whether an event or state of affairs *E* is harmful to me if and only if:

1. If *E* held I would lack *G*, and
2. If *E* had not held, I would have had *G*.

Accordingly, the loss of my arms is harmful, since I am worse off without them, which means that there is a good, such as my capacity to use tools, that meets (1) and (2). But my becoming a baseball star is not relevant to the harmfulness of the loss of arms, since it is disqualified by (2): even if I kept my arms, I would not become a baseball star.

Unfortunately, there is a problem with the counterfactual criterion. It works well when we evaluate *losses*, such as the loss of my arms. But it often fails when we evaluate *lacks*. Consider, for example, *my lack of genius*: does it harm me? It does preclude my enjoying goods great intelligence would make possible, such as the ability to discover profound truths about the universe. So it meets (1). It meets (2) as well: if I failed to lack genius -- that is, if I were a genius -- I would enjoy the goods genius brings. However, it is peculiar to say that I am harmed by my lack of genius. Why is this?

4.3 Misfortune Versus Harm

The explanation we need involves clarifying the relationship between harm and misfortune. Let us begin with some observations: it is no misfortune for me not to enjoy the goods genius would bring me, and it is no *misfortune* to be deprived of goods when their absence is not a misfortune for me. Also, lacking genius is not in itself a misfortune, and yet genius is a great good. Similar points can be made about extraordinary beauty or God-like powers of various sorts: while these are great gifts, lacking them is no misfortune. (This is not to deny that beauty could come to be important to a person who makes it the focus of life, so that losing it would be a misfortune, even if never having it would not have been.) So it need not be a misfortune to lack great goods. And it is false that, the greater the good, the greater the misfortune we suffer in being denied it.

Nagel may be making a similar point when he writes, “the question is whether we can regard as a misfortune any limitation, like mortality, that is normal to the species” (Fischer 1993, p. 68). It is not clear what Nagel is saying, because limitations that are typical to a species might be ruled out as misfortunes on the grounds that lacking them is not really humanly good. Following Aristotle, we could define what is humanly good in terms of what enables an exemplary yet actual human being to live as well as possible. Lacking the limitations of the exemplary human being, we might add, is not humanly good, and having

them is not a misfortune. Great beauty, Aristotle would say, is humanly good, but superhuman strength, of which even the best of us is incapable, is not. However, this is not the view we have defended. Our point is that a feature could be a genuine good for a human being, yet lacking it might be no misfortune.

How can lacking a great good fail to be a misfortune? Because some goods are less important for us than others, and it is a misfortune to be deprived of a good if and only if it is important for us to have it. But when is it important for us to have a good? The answer lies in the fact that it is one thing for a life to be (merely) good, and quite another for it to be the best (physically? conceptually?) possible life; some qualities are requisite for a merely good life, or a life that meets the minimal conditions for happiness, while others are essential to the optimal life, or one that provides for a degree of happiness that cannot be exceeded. Failing to have (something essential to) a good life (or minimal happiness) is a misfortune, yet failing to have (what makes for) the best possible life (or maximal happiness) surely is not. So it is plausible to say that the goods it is important to have, and whose absence constitutes a misfortune, are *essential goods*: items essential to a (merely) good life, or a life of (mere) happiness. (Of course, given the flexibility of the term 'misfortune,' some hedging is in order. Perhaps things need not go so far as to deprive us of an essential good to be a misfortune; perhaps it is enough that they significantly impair our chances of attaining the essentials.)

The explanation of why it is awkward to speak of harm when certain good possibilities, such as enjoying God-like powers, are not actualized, is that we tend to use the term 'harm' to refer to misfortune, and often it is *not* a misfortune for us when good possibilities fail to be actualized (since the failure does not bear on our having essential goods). The awkwardness is exacerbated, however, because we also want to use the term 'harm' to refer to things that are bad for us, and 'bad' covers a lot of territory: When, on the whole, something makes us worse off in any way or to any degree, no matter how trivial, it is common to call it a bad thing; we also say, of a good state of affairs that does not actually hold, that its failure to hold is a bad thing, since we would be better off if it did. Our use of the terms 'bad', 'harm,' and 'misfortune' thus makes it difficult for us to express the fact that the nonactualization of a good possibility might be no *misfortune*, even though it is *harmful* since *bad*, and *bad* only in the sense that we would be better off if the possibility were actual.

We have a choice to make. We may apply the terms 'harm' and 'bad' to any overall worsening of our condition, or to any failure to make our overall condition better. If we do, however, we cannot say that all harms are misfortunes.

5. When Is Death a Misfortune?

What is the upshot for death? Clearly some deaths deprive people of essential goods, and are therefore misfortunes. But are all deaths misfortunes?

5.1 Premature Death Is a Misfortune

By applying the thwarted desire criterion, we can reinforce the conclusion that death is not always a

misfortune. Perhaps it is not bad to die at an advanced enough age, for people who live long enough may be ground down by life until they give up many of their goals. Also, they will have attained many of their aspirations. If already satisfied, or given up, a desire cannot be thwarted, even by death, so as we lose our motivation for living, death ceases to be objectionable to us. Perhaps death is bad for us only if premature in the sense that it comes when we are still in the grip of desires that propel us forward in life, and only if satisfying our desires is a real prospect.

We are left to wonder whether death would ever cease to be objectionable were we *not* ravaged by bad health and other setbacks. Williams argues that it would be bad to live forever, even under the best of circumstances. His view is based on an assumption about the relationship between our identities and the desires that motivate us to live.

Consider a woman who wants to die. She might still take the view that if she is to live on, then she should be well fed and clothed. She wants food and clothing on condition she remain alive. In this sense her desires are *conditional*, and do not give her reason to live. Contrast a father who is committed to rearing a beloved daughter: he desires unconditionally that the child do well, and his desire gives him reason to live, because he can rear his child only if he survives. In this sense, his desire is categorical, or unconditional. Williams thinks that categorical desires are essential to identity, and give meaning to life. Through categorical desires, we are attached to projects or relationships that are definitive of the self; faced with their destruction, we would feel our lives are meaningless, and that in an important sense we cannot survive as the persons we once were.

The bearing on death, according to Williams, is, first, that people have good reason to condemn a death that is premature in the sense that it thwarts their categorical desires. Second, mortality is good, since people who live long enough eventually will lose the categorical desires with which they identify. Life will lose its novelty, and oppressive boredom will set in. To avoid ennui, superseniors would have to replace their fundamental desires, again and again. But this is to abandon their identities; it is tantamount to death.

As Williams says, lives of unimaginative routine will eventually grow stale if extended long enough. Of course, this is not supposed to comfort ordinary mortals, most of whom will die long before routine undermines the joy in living. However, as several theorists, including Nagel (1986, p. 224, n. 3) Glover (1977, p. 57), and Fischer (1993, p. 11) have suggested, it is not obvious that life must become dull. Williams may have overlooked how rich and complex life can be, especially for superseniors who pursue multiple open-ended projects in the company of other superseniors. His response to this kind of criticism is that even rich and open-ended projects eventually will become routine (say after a few billion years), so our pursuits must be replaced periodically if we are to remain interested in life. But to phase in wholly new projects is to lose our identity.

Williams's response faces objections. First, we might avoid boredom by adding to our pursuits, and varying the way we approach them, without abandoning certain core interests that define us. Second, Williams is working with a view of identity that may be too narrow. Many of us would welcome a possibility that he downplays: gradually transforming our interests and projects over time. Transformation

is not death. It is distinct from, and preferable to, annihilation. Transformation would be death only if identity were wholly a matter of connectedness. However, we also think of identity as continuity: If we could live endlessly, the stages of our lives would display reduced connectedness, yet they would be continuous, which is a property that is important in the kind of survival most of us prize. Even after drinking at the fountain of eternal youth, we would tend to focus on relatively short stretches of our indefinitely extensive lives, and over these periods we would prize connectedness, since we are animated by specific projects and relationships that can be developed only if there are strong interconnections among the temporal stages of our lives. However, sometimes we would turn our attention to relatively long stretches of life, and then, prizing continuity, we would phase in new and worthwhile undertakings that build upon, and do not wholly replace, the old.

6. Can Death's Harmfulness be Reduced?

We have been asking after the objectively correct answer to the question, Is it bad to die? Instead of treating the value of death as a fact to be discovered, some argue that death *need* not be a misfortune, if we prepare ourselves suitably. Assuming that identity is malleable, we might identify with something durable, such as a family line or the community or the natural order, so that ‘we’ survive, and need not be harmed by, the demise of the individual (of course, the demise of the family or community or natural order would still harm us). Or, like Gautama, we might altogether abandon the notion of the self, in an attempt to convince ourselves that there is no one to die. Another approach does not specifically involve adjusting the boundaries of the self. Ancient philosophers in both the East and West noticed that if we adopt the right conception of the good life, and the right desires, we prevent death from harming us. We can even become invulnerable in an important sense: nothing that happens to us will be able to diminish the goodness of our lives; nothing will be a misfortune for us, including death. Let us see how this idea can be developed.

For invulnerability, what is needed is a view given which the things that affect the goodness of our lives are entirely in our control. To arrive at such a view, we equate goodness with our happiness (or well-being), then construe happiness negatively and in such a way that it demands nothing that is out of our control.

6.1 Two Negative Accounts of Happiness

Epicurus's brand of hedonism is quite suitable in this regard, and it is likely that its proponents consider negative, or pain avoiding, hedonism more attractive than the alternatives, including positive, or pleasure seeking hedonism, due to the superior way it facilitates the goal of invulnerability. Epicurus characterizes happiness as (a) subjective, (b) agent-relative, and (c) largely negative, in that it is understood in terms of what is absent (pain), rather than what is present. These three features help shield our happiness from aspects of the world that are not in our control and that can deprive us of any positive form of happiness, such as death and the suffering of others.

Rather than reducing it to the absence of suffering, we might instead equate happiness with the (again

negative) condition of lacking thwarted desires (contrast the positive condition of having satisfied desires). These two ways of construing happiness are distinct yet closely related: Distinct since not all pain is due to thwarted desire (mashing one's thumb hurts no matter what we want), and not all desire thwartings are painful or even noticed (recall the example of the ignored will). But they are closely related since, as Gautama (and much later Epicurus) noted, the chief cause of suffering is thwarted desire: usually, we suffer from an event only when we wanted it not to happen.

6.2 Becoming Invulnerable Through Desire Adaptation

To convince ourselves that death cannot harm us, either of two strategies will serve (Luper 1987, 1996). First, we could adopt Epicurus's negative hedonism, and adjust our attitudes accordingly. Second, we could adopt the no-thwarted desires view of happiness, then *thanatize* our desires, in this sense: abandon all desires that death might thwart . Epicurus does not distinguish between the two approaches, yet they differ. Negative hedonism does encourage us to pare back our desires, but only so as to avoid the painful experiences associated with their being frustrated or to avoid other painful consequences of having them. In particular, it allows us to retain a desire so long as we cannot experience the events by which it might be thwarted, and so long as retaining it is not painful for other reasons. It does not prompt us to drop a desire simply because death might thwart it, since death thwarts desires in ways we cannot experience. Suffering is caused neither by death nor by the thwarting of any desires for which death is responsible. Thanatizing, by contrast, eliminates desires death might thwart. However, it allows us to retain some desires that are discouraged by negative hedonism. Consider, for example, my desire that the moon orbit the earth after I am dead: whether this desire is satisfied is not affected by my being alive or by anything I might do while alive. In this sense it is *independent*. Thanatizing leaves independent desires in place, yet some of them are discouraged by negative hedonism. The moon's (not) orbiting the earth after I am dead can have no causal impact on me, and hence must be a matter of indifference to me if I am a negative hedonist.

Negative hedonism and thanatizing insulate us from the view that death is harmful to us. However, both strategies leave us vulnerable to harm of other sorts. Both leave us free to make all sorts of plans, only to suffer our plans' failure. To properly assess whether it is advisable to adjust our attitudes so as to fend off the harm of death, it is useful to examine the more general project of developing invulnerability to all harms whatsoever. How might this more extensive adjustment go?

It is accomplished in two steps. First, we equate goodness with happiness, and adopt the no-thwarted desires view of happiness. Second, we eschew all except those desires we can surely satisfy. We might pursue the (autarchist) strategy of allowing ourselves only desires that we can satisfy by our own power. We might limit ourselves to the (conformist) attitude that whatever happens (or whatever happens by necessity) is what we want to happen. More extreme yet, we might adopt the (nihilist) approach of wanting nothing at all. Autarchy, conformism and nihilism stop any event from impairing our happiness. Along the way, they insulate us from the threat of death, by eliminating the desire not to die, as well as any desire whose satisfaction requires our being alive.

Unfortunately, all the strategies we have canvassed have a drawback: they leave us with an impoverished conception of happiness, of what matters. After applying these strategies, our happiness is little more than the absence of unhappiness. For example, Epicurean negative hedonists must say that nothing that happens after they die matters; in particular, they will be indifferent to the suffering of their children -- so long as it will occur after they (the parents) are dead. They are incapable of true love and friendship since these commit us to the judgment that the well-being of another matters for its own sake, while the negative hedonist thinks only one thing matters for its own sake: pain avoidance. (Admittedly, Epicurus claims that friendship is good, but he cannot have it both ways: negative hedonism rules out loving relationships of all sorts.) Arguably, similar remarks apply to people who thanatize their desires (but see Rosenbaum 1989). A thanatized parent cannot sustain real concern for the well-being of her children through independent desires, since such desires leave her indifferent to her children while there is anything that she can do to help her children. At best they can take the attitude, *let my children do well so long as I cannot possibly do anything about them*, which is so bizarre as to be psychologically impossible, and which falls far short of genuine concern, not to say love, since the latter guarantees the attitude, *let my children do well even if I cannot possibly help them*. Autarchics, conformists and nihilists are even more callous: they shrug off the suffering of their children (and everyone else) no matter when it occurs.

Moreover, in avoiding all desires that would leave them vulnerable to death, thanatics, autarchics, conformists and nihilists must give up the view that life is worth living, as well as the projects and concerns that constitute grounds for thinking that life is good, assuming that, in rational beings, the judgment that life is worthwhile prompts the desire to live, which could be thwarted at any time. Any reason to (want to) live is an excellent reason to want not to die; to avoid the latter, we must avoid the former.

However, the core idea of adapting our desires and requirements for happiness is useful, if not taken to an extreme. For what deprives us of happiness is a misfortune. Hence it is imprudent to let our happiness hinge on demands we cannot possibly meet, and better to reshape our most fundamental ideals so that they are manageable. In particular, it is prudent to take the view that we can be happy, or content, with a normal lifespan, which falls far short of immortality. $$This is not to say we should be indifferent about goals we shall never achieve, however. Most of us would be glad to live endlessly under favorable circumstances. But our attitude here should be that attaining immortality moves us well beyond mere happiness, and that being limited to a normal lifespan is not a misfortune.

Bibliography

- Blackmore, S., 1993. *Dying to Live: Near-Death Experiences*. Buffalo, NY: Prometheus Books.
- Braddock, G., 2000. "Epicureanism, Death, and the Good Life," *Philosophical Inquiry* 22, no. 1-2.
- Epicurus, 1966a. *Principal Doctrines*, in Saunders, J., ed., *Greek and Roman Philosophy after Aristotle*. New York: Free Press.
- -----, 1966b. *Letter to Menoeceus*, in Saunders, J., Ed., *Greek and Roman Philosophy after Aristotle*. New York: Free Press.
- Feldman, F., 1991. "Some Puzzles About the Evil of Death," *The Philosophical Review* 100, no.

205-27; reprinted in Fischer 1993, 307-326.
- -----, 1992. *Confrontations with the Reaper*. New York: Oxford University Press.
- Fischer, J.M., ed., 1993. *The Metaphysics of Death*. Stanford University Press.
- Glover, J., 1977. *Causing Death and Saving Lives*. Harmondsworth: Penguin Books.
- Green, M. and Winkler, D., 1980. “Brain Death and personal Identity,” *Philosophy and Public Affairs* 9, 105-133.
- Locke, J., 1689. *An Essay Concerning Human Understanding*.
- Lucretius, 1951. *On the Nature of the Universe*. Latham, reg. trans., Penguin Classics.
- Luper(-Foy), S., 1987. “Annihilation,” *The Philosophical Quarterly* 37, no. 148, 233-52. Reprinted in Fischer 1993.
- -----, 1996. *Invulnerability: On Securing Happiness*. Chicago: Open Court.
- McMahan, J., 1988. “Death and the Value of Life,” *Ethics* 99, no. 1, 32-61; reprinted in Fischer 1993, 233-266.
- Nagel, T., 1979. “Death,” in Nagel, T., *Mortal Questions*. Cambridge: Cambridge University Press.
- -----, 1986. *The View From Nowhere*. Oxford: Oxford University Press.
- Nozick, R., 1971. “On the Randian Argument,” *The Personalist*. Reprinted in Paul, J., ed., *Reading Nozick*. Totowa, NJ: Rowman & Littlefield, 1981.
- Parfit, D., 1984. *Reasons and Persons*. Oxford: Clarendon Press.
- Perry, J., ed., 1975. *Personal Identity*. Berkeley: University of California Press.
- President's Commission, 1981. *Defining Death: Medical, Legal, and Ethical Issues in the Determination of Death*. Washington, D.C.
- Pitcher, G., 1984. “The Misfortunes of the Dead,” in *American Philosophical Quarterly* 21, no. 2, 217-225; reprinted in Fischer 1993, 119-134.
- -----, 1989. “Epicurus and Annihilation,” *Philosophical Quarterly* 39, no. 154, 81-90; reprinted in Fischer 1993, 293-304.
- Rosenberg, J., 1983. *Thinking Clearly About Death*. Englewood Cliffs, NJ: Prentice-Hall.
- Silverstein, H., 1980. “The Evil of Death,” *Journal of Philosophy* 77, no. 7, 401-424; reprinted in Fischer 1993, 95-116.
- Unamuno, M., 1913. Kerrigan, A., Trans., *The Tragic Sense of Life in Men and Nations*. Princeton: Princeton University Press, 1972.
- Williams, B., 1973. “The Makropulos Case: Reflections on the Tedium of Immortality,” in Williams, B., *Problems of the Self*. Cambridge: Cambridge University Press.
- Yourgrau, P., 1987. “The Dead,” *Journal of Philosophy* 86, no. 2, 84-101; reprinted in Fischer 1993, 137-156.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Epicurus | identity

Steven Luper
sluper@trinity.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 21, 2002
Content last modified: May 21, 2002

Omnipotence

Omnipotence is maximal power. Maximal greatness (or perfection) includes omnipotence. According to traditional Western theism, God is maximally great (or perfect), and therefore is omnipotent. Omnipotence seems puzzling, even paradoxical, to many philosophers. They wonder, for example, whether God can create a spherical cube, or make a stone so massive that he cannot move it. Is there a consistent analysis of omnipotence? What are the implications of such an analysis for the nature of God?

1. Introductory Preliminaries

Philosophical reflection upon the notion of omnipotence raises many puzzling questions about whether or not a consistent notion of omnipotence places limitations on the power of an omnipotent agent. Could an omnipotent agent create a stone so massive that that agent could not move it? Paradoxically, it appears that however this question is answered, an omnipotent agent turns out not to be all-powerful. Could such an agent have the power to create or overturn necessary truths of logic and mathematics? Could an agent of this kind bring about or alter the past? Is the notion of an omnipotent agent other than God an intelligible one? Could *two* omnipotent agents coexist? If there are states of affairs that an omnipotent agent is powerless to bring about, then how is the notion of omnipotence intelligibly to be defined? Moreover, an obstacle to traditional Western theism arises if it is impossible for God to be morally perfect and omnipotent. If an omnipotent God is powerless to do evil, then how can he be omnipotent? Rational theology seeks an analysis of the concept of omnipotence that resolves the puzzles and apparent paradoxes that surround this concept. If the notion of omnipotence were found to be unintelligible, or incompatible with moral perfection, then traditional Western theism would be false.

According to some philosophers, omnipotence should be understood in terms of the power to *perform certain tasks*, for instance, to kill oneself, to make 2+2=4, or to make oneself non-omniscient. However, in recent philosophical discussion, the approach has more often been to analyze omnipotence in terms of the power to *bring about certain possible states of affairs* (understood as propositional entities which either obtain or fail to obtain).[1] This approach appears to have been more successful.

One sense of 'omnipotence' is, literally, that of having the power to bring about *any* state of affairs whatsoever, including necessary and impossible states of affairs. Descartes seems to have had such a notion.[2] Yet, Aquinas and Maimonides held the view that this sense of 'omnipotence' is incoherent. Their view can be defended as follows. It is *not* possible for an agent to bring about an *impossible* state of affairs (e.g., *that there is a shapeless cube*), since if it were, it would be possible for an impossible state of affairs to obtain, which is a contradiction.[3] Nor is it possible for an agent to bring about a *necessary* state of affairs (e.g., *that all cubes are shaped*). It is possible for an agent, *a*, to bring about a necessary state of affairs, *s*, only if possibly, (1) *a* brings about *s*, and (2) if *a* had not acted, then *s* would have failed to obtain. Because a necessary state of affairs obtains whether or not anyone acts, (2) is false. As a consequence, it is impossible for an agent to bring about either a necessary or an impossible state of affairs. Obviously, an agent's having the *power* to bring about a state of affairs entails that, *possibly*, the agent brings about that state of affairs. Thus, the first sense of 'omnipotence' is incoherent. Henceforth, it will be assumed that it is not possible for an agent to have the power to bring about *any* state of affairs whatsoever.

A second sense of 'omnipotence' is that of *maximal power*, meaning just that no being could exceed the overall power of an omnipotent being. It does not follow that a maximally powerful being can bring about *any* state of affairs, since, as we have seen, bringing about some such states of affairs is impossible. Nor does it follow that a being with maximal power can bring about whatever any *other* agent can bring about. If *a* can bring about *s*, and *b* cannot, it does not follow that *b* is *not overall* more powerful than *a*, since it could be that *b* can bring about more states of affairs than *a* can, rather than the other way around. This *comparative* sense of 'omnipotence' as maximal power appears to be the only sense that has a chance to be intelligible.

Power should be distinguished from *ability*. Power is ability plus opportunity: a being having maximal ability who is prevented by circumstances from exercising those abilities would not be omnipotent. Nothing could prevent an omnipotent agent from exercising its powers, if it were to endeavor to do so.

In light of the foregoing, could there be two coexistent omnipotent agents, Dick and Jane? If this were even possible, then *possibly*, at some time, *t*, Dick, while retaining his omnipotence, attempts to move a feather, and at *t*, Jane, while retaining her omnipotence, attempts to keep that feather motionless. Intuitively, in this case, neither Dick nor Jane would affect the feather as to its motion or rest. Thus, in this case, at *t*, Dick would be powerless to move the feather, and at *t*, Jane would be powerless to keep the feather motionless! But it is absurd to suppose that an omnipotent agent could lack the power to move a feather or the power to keep it motionless. Therefore, neither Dick nor Jane is omnipotent. Since the idea that there could be two omnipotent beings who are *necessarily* always in perfect agreement is highly

questionable, it seems impossible that there be two coexistent omnipotent agents.

Could an agent be accidentally omnipotent? At first glance, this appears possible, but there is the following argument for the opposite view. On the assumption that God exists, he has necessary existence, is not temporally limited, and is essentially omnipotent. But there could not be two coexistent omnipotent agents. Thus, on the assumption that God exists, an accidentally omnipotent being is impossible.

This argument against the possibility of accidental omnipotence presupposes traditional Western theism. However, traditional Western theism is highly controversial, and *neutrality* about whether God exists has some advantages. If one is neutral about whether God exists, then omnipotence should *not* be assumed to be attributable *only* to the God of traditional Western theism or *only* to an essentially omnipotent being.

2. The Scope of Omnipotence

The intelligibility of the notion of omnipotence has been challenged by the so-called paradox or riddle of the stone. Can an omnipotent agent, Jane, bring it about that there is a stone of some mass, *m*, which Jane cannot move? If the answer is ‘yes’, then there is a state of affairs that Jane cannot bring about, namely, (S1) *that a stone of mass m moves*. On the other hand, if the answer is ‘no’, then there is another state of affairs that Jane cannot bring about, namely, (S2) *that there is a stone of mass m which Jane cannot move*. Thus, it seems that whether or not Jane can make the stone in question, there is some possible state of affairs that an omnipotent agent cannot bring about. And this appears to be paradoxical.

A first resolution of the paradox comes into play when Jane is an *essentially* omnipotent agent. In that case, the state of affairs of Jane’s being non-omnipotent is impossible. Therefore, Jane cannot bring it about that she is not omnipotent. Since, necessarily, an omnipotent agent can move any stone, no matter how massive, (S2) is impossible. But, as we have seen, an omnipotent agent is not required to be able to bring about an impossible state of affairs.

If, on the other hand, Jane is an *accidentally* omnipotent agent, both (S1) and (S2) *are* possible, and it *is* possible for some omnipotent agent to bring it about that (S1) obtains at one time, *and* that (S2) obtains at a different time. Thus, there is a second solution to the paradox. In this case, Jane’s being non-omnipotent is a possible state of affairs; thus, we may assume that it *is* possible for Jane to bring it about that she is non-omnipotent. So, Jane can create and move a stone, *s*, of mass, *m*, while omnipotent, and *subsequently* bring it about that she is not omnipotent and powerless to move *s*. As a consequence, Jane can bring about both (S1) and (S2), but only if they obtain at different times.[4]

It might now be conjectured that omnipotence can be analyzed simply as the power to bring it about that any *contingent* state of affairs obtains. However, the following list of contingent states of affairs shows that there can be contingent states of affairs that an omnipotent agent is powerless to bring about, and hence that this simple analysis is inadequate:

a. that a raindrop fell;
b. that a raindrop falls at *t* (where *t* is a past time);
c. that Parmenides lectures for the first time;
d. that the Amazon River floods an odd number of times less than four;
e. that a snowflake falls and no omnipotent agent ever exists; and
f. that Plato freely decides to write a dialogue.

Note that (a) is a past state of affairs. Presumably, it is not possible for an efficient cause to occur *later* than its effect. However, an agent's bringing about a state of affairs is a kind of efficient causation. Therefore, it is not possible for an agent to bring about anything that is in the *past*. In other words, it is impossible for *any* agent to have power over what is past. Hence, no agent, not even an omnipotent one, can bring it about that (a) obtains. Likewise, despite the fact that (b) can be brought about *prior* to *t*, the impossibility of an agent's having power over what is past implies that *after t* even an omnipotent agent cannot bring it about that (b) obtains. In the case of (c), prior to Parmenides's first lecture, an omnipotent agent can bring about (c). But once Parmenides has lectured, even an omnipotent agent cannot bring it about that (c) obtains. As for (d), prior to the Amazon's third flooding, an omnipotent agent can bring it about that (d) obtains, while after the Amazon's third flooding, even an omnipotent agent cannot bring it about that (d) obtains. (e) introduces a special difficulty. Although it is obvious that (e) could *not* be brought about by an omnipotent agent, it can be argued plausibly that it *is* possible for a non-omnipotent agent to bring about (e) by causing a snowflake to fall, *provided that* no omnipotent agent ever exists.[5] But, as we argued earlier, a maximally powerful being need not have the power to bring about every state of affairs that any other being could. Lastly, while if the libertarian theory of free will is correct, an omnipotent agent (who is, of course, *other than Plato*) cannot bring about (f), apparently a non-omnipotent agent, namely, Plato, can bring it about that (f) obtains.

Consequently, a satisfactory analysis of omnipotence ought not to require that an omnipotent agent have the power to bring about (a), (b), (c), (d), (e), or (f), if it is assumed, *arguendo*, in the case of (f), that libertarianism is true.

Because of the wide disparity among contingent states of affairs, (a)-(f), one might despair of finding an analysis of omnipotence that both deals satisfactorily with all of these states of affairs and implies that an omnipotent being has, intuitively speaking, sufficient power. Is such pessimism warranted, or is omnipotence analyzable?

There are at least two approaches to analyzing omnipotence that hold out some hope of success. The first utilizes the notion of an *unrestrictedly repeatable state of affairs*, and the second utilizes the notion of *two worlds sharing their histories up to a time*. Although these approaches to analyzing omnipotence differ in important ways, they are in broad agreement on the leading idea that maximal power has logical and temporal limitiations, including the limitation that an omnipotent agent cannot bring about, i.e., cause, another agent's free decision in the libertarian sense. In the following two sections, some recent instances of these approaches are set forth and compared.

3. Omnipotence and Unrestricted Repeatability

One attempt to analyze omnipotence in terms of unrestricted repeatability is the account of Hoffman and Rosenkrantz. According to their approach, by identifying certain features of (a)-(f), we can find a feature that none of them possesses, and in terms of which an analysis of omnipotence can be stated.[6] To begin, unless it is possible for *some* agent to bring about a given state of affairs, an omnipotent agent ought not to be required to be able to bring about that state of affairs. But (a) is not possibly brought about by any agent.

Next, while (b) and (c) are possibly brought about by some agent, they are not *repeatable*: it is not possible for either one of them to obtain, subsequently fail to obtain, and then obtain again. Note that if, because (a) is not possibly brought about by someone, an omnipotent agent is not required to be able to bring about (a), then for the same reason, that agent is also not required to be able to bring about impossible or necessary states of affairs. Moreover, if, because (b) and (c) are not repeatable, an omnipotent agent is not required to bring about (b) or (c), then for the same reason, that agent is also not required to be able to bring about impossible or necessary states of affairs. These reasons for not requiring an omnipotent agent to have the power to bring about impossible or necessary states of affairs cohere with our earlier independent arguments for these restrictions.

Third, while (d) *is* repeatable, it is *not unrestrictedly* repeatable, that is, it cannot obtain, then fail to obtain, then obtain again, and so on, eternally.

Fourth, while (e) *is* unrestrictedly repeatable, it is a *complex* state of affairs, namely, a *conjunctive* state of affairs whose second conjunct is *not* repeatable. These examples suggest a hypothesis about repeatability and its relation to power, namely, that an omnipotent agent should *not* be required to have the power to bring about *either* a state of affairs that is not unrestrictedly repeatable, *or* a conjunctive state of affairs one of whose conjuncts is not unrestrictedly repeatable.

Lastly, although (f) *is* unrestrictedly repeatable, (f) is another type of complex state of affairs. In particular, it is identifiable with or analyzable as a conjunctive state of affairs. This state of affairs has three conjuncts, the second of which is not possibly brought about by anyone. The conjunctive state of affairs in question can be informally expressed as follows: Plato decides to write a dialogue; and there is no *antecedent* sufficient causal condition of Plato’s deciding to write a dialogue; and there is no concurrent sufficient causal condition of Plato’s deciding to write a dialogue. Because it is impossible for an agent to have power over what is *past*, the second conjunct of this state of affairs is not possibly brought about by anyone. Thus, an omnipotent agent ought not to be required to have the power to bring about a state of affairs that is identifiable with or analyzable as a conjunctive state of affairs one of whose conjuncts is not possibly brought about by anyone.

According to the account of Hoffman and Rosenkrantz, incoporating these ideas, omnipotence can be analyzed in terms of the following three definitions.

($D1$) The period of time t is a sufficient interval for $s =_{df} s$ is a state of affairs such that: it is possible that s obtains at a time-period which has the duration of t.

For example, any period of time with a duration of 7 seconds is a sufficient interval for the state of affairs *that a ball rolls for 7 seconds*.

($D2$) A state of affairs, s, is unrestrictedly repeatable $=_{df} s$ is possibly such that: $\forall n \exists t_1 \exists t_2 \exists t_3 \ldots \exists t_n [(t_1 < t_2 < t_3 < \ldots < t_n$ are periods of time which are sufficient intervals for s & s obtains at t_1 & s doesn't obtain at t_2 & s obtains at t_3 & … & s obtains at t_n) if and only if n is odd].[7]

For instance, the state of affairs *that a ball rolls for 7 seconds* is unrestrictedly repeatable.

($D3$) x is omnipotent at $t =_{df} \forall s$(if it is possible for some agent to bring about s then at t x has it within his power to bring about s).

In *D3*, x ranges over agents, and s over states of affairs that satisfy the following condition:

(C) (i) s is unrestrictedly repeatable, and of the form 'in n minutes, p', & (p is a complex state of affairs $\rightarrow$ (each of the parts of p is unrestrictedly repeatable & possibly brought about by someone)), or (ii) s is of the form 'q forever after', where q is a state of affairs which satisfies (i).[8]

In applying *D3* to states of affairs like (e) and (f) it should be observed that a *conjunct* of a conjunctive state of affairs is a *part* of such a complex state of affairs.[9]

As intended, *D3* does not require an omnipotent agent to have the power to bring about either impossible or necessary states of affairs, or states of affairs such as (a)-(f). Furthermore, *D3* does not unduly limit the power of an omnipotent agent, since an agent's bringing about a state of affairs can always be "cashed out" in terms of that agent's bringing about an unrestrictedly repeatable state of affairs that it is possible for some agent to bring about. That is, necessarily, for any state of affairs, s, if an agent, a, brings about s, then either s is an unrestrictedly repeatable state of affairs which it is possible for some agent to bring about, or else a brings about s *by* bringing about q, where q is an unrestrictedly repeatable state of affairs which it is possible for some agent to bring about. For instance, an omnipotent agent can bring about the state of affairs, *that in one hour*, *Parmenides lectures for the first time*, by bringing about the state of affairs, *that in one hour*, *Parmenides lectures*, when this lecture is Parmenides's first. And although the former state of affairs is a nonrepeatable one that *D3* does *not* require an omnipotent agent to be able to bring about, the latter state of affairs *is* an unrestrictedly repeatable state of affairs that *D3* *does* require an omnipotent agent to be able to bring about.

4. Omnipotence and the Shared Histories Approach

The alternative approach to analyzing omnipotence in terms of two worlds sharing their histories up to a time is exemplified by the accounts of Flint and Freddoso, and Wierenga. As we shall see, although these two accounts are similar, they differ in certain significant respects.

Flint and Freddoso's account of what it is for an agent *S* at a time *t* to be omnipotent in a possible world *W* is formulated as follows.

> *S* is omnipotent at *t* in *W* if and only if for any state of affairs *p* and world-type-for-*S Ls* such that *p* is not a member of *Ls*, if there is a world *W** such that
>
> 1. *Ls* is true in both *W* and *W**, and
> 2. *W** shares the same history with *W* at *t*, and
> 3. at *t* in *W** someone actualizes *p*, then *S* has the power at *t* in *W* to actualize *p*.[10]

The notion of a world-type-for-*S Ls* is to be understood in the following way. A *world-type* is "a set which is such that for any *counterfactual of freedom*, i.e., any proposition which can be expressed by a sentence of the form 'If individual essence *P* were instantiated in circumstances *C* at time *t* and its instantiation were left free with respect to action *A*, the instantiation of *P* would freely do *A*' -- either that counterfactual or its negation is a member of the set."[11] It may also be stipulated "that for any two members of the set, the conjunction of those two members is a member of the set as well."[12] Moreover, "a world-type is *true* just in case every proposition which is a member of it is true."[13] In addition, it is presupposed that "for any free agent *x*, there will be a set of all and only those true counterfactuals of freedom (or true negations of such counterfactuals) over whose truth-value *x* has no control."[14] A set of this kind is referred to as *the world-type-for-x*. Finally, '*Lx*' designates the true-world-type-for-*x*.

The notion of actualization employed in this account of omnipotence calls for some explanation. If an agent, *S*, brings about a state of affairs, *p*, then *S* actualizes *p*. However, this account presupposes that an agent may [*weakly*] actualize another agent's making a *free* decision without *bringing about* or *causing* that decision. In particular, it is assumed that an agent may weakly actualize a decision that is free in the libertarian sense by bringing about the antecedent of a true "counterfactual of freedom."

The basic idea of this account of omnipotence is that an agent is omnipotent just when he can actualize any state of affairs that it is possible for someone to actualize, except for certain "counterfactuals of freedom", their consequents, and certain states of affairs that are "accidentally impossible" because of the past.

With respect to so-called counterfactuals of freedom, this account presupposes that some of them, for

example,

> If Jessica were offered the grant, then she would freely decide to accept it,

are *true*. Some philosophers hold the contrary view that a subjunctive conditional of this kind is *necessarily false*. Why do these philosophers reject the claim that some "counterfactuals of freedom" are true? Presumably, what distinguishes a subjunctive conditional from a corresponding material conditional is that only the former expresses a strong or necessary connection of some sort between the conditions specified by the antecedent and the consequent. Seemingly, the only kinds of strong or necessary connections available in this case are relations [broadly speaking] of either *causation* or *entailment*. Consequently, it appears that the subjunctive conditional under discussion is necessarily false, since if Jessica *freely* decides to accept the grant [in the relevant libertarian sense], then her making that decision is neither *caused* nor *entailed* by her being offered the grant. If the foregoing line of reasoning is correct, then the notion of a true "counterfactual of freedom" is incoherent. Since Flint and Freddoso's account of omnipotence presupposes that there are such "counterfactuals of freedom," it can be argued that this account is incoherent.

Moreover, it can be argued that a state of affairs discussed earlier provides a counter-example to Flint and Freddoso's account of omnipotence, namely:

> (e) A snowflake falls and no omnipotent agent ever exists.

A non-omnipotent agent can bring about or actualize (e) by bringing it about that a snowflake falls when in fact no omnipotent agent ever exists. But, it is clear that an omnipotent agent cannot bring about or actualize (e). For although an omnipotent agent can bring it about that a snowflake falls, surely, an omnipotent agent cannot bring it about that *no omnipotent ever exists*, nor would this conjunct of (e) obtain if there were an omnipotent agent. Moreover, we may assume that there are possible worlds, *W* and *W**, such that *W* and *W** share the same history up to a time *t*, no omnipotent agent ever exists in *W**, and a contingently omnipotent agent, Oscar, is omnipotent for the first time at *t* in *W*. We may also assume that *W** is a world in which at *t* some non-omnipotent agent actualizes (e). On the other hand, evidently, if in *W*, Oscar is omnipotent at *t*, then at *t* Oscar cannot actualize (e). Note that since the second conjunct of (e) is not unrestrictedly repeatable, this is consistent with Hoffman and Rosenkrantz's account of omnipotence; their account does not require an omnipotent agent to be able to bring about a conjunctive state of affairs one of whose conjuncts is not unrestrictedly repeatable. On the other hand, Flint and Freddoso's account of omnipotence implies that in *W*, at *t* Oscar has the power to actualize (e). This implication holds for the following reasons. First, (e) is not a member of a world-type-for-Oscar, inasmuch as (e) is neither a "counterfactual of freedom," the negation of one, nor a conjunction of such "counterfactuals of freedom." Second, we may assume that a world-type-for-Oscar is true in both *W** and *W*, since the assumption that an agent is not omnipotent in one possible world, and is omnipotent for a time in another possible world, does not necessitate any difference in the world-type for that agent which is true in those worlds. Third, it is possible for someone at *t* to actualize (e) in a world, *W**, that has the same history up to *t* as *W*.[15] Thus, arguably, Flint and Freddoso's account of omnipotence requires that

in *W* an omnipotent agent, Oscar, at *t* has the power to actualize (e), when Oscar lacks this power. If this is right, then their account does not provide a logically necessary condition of omnipotence.[16]

A counter-example of this kind assumes that an analysis of omnipotence should allow for the possibility of an omnipotent agent other than God. Given this assumption, (e) seems to provide a counter-example to Flint and Freddoso's account of omnipotence, but not to the account of Hoffman and Rosenkrantz.

Let us now turn to Wierenga's account of omnipotence.. The basic idea of Wierenga's account of omnipotence is that an agent is omnipotent if and only if he can do anything that it is possible for him to do, given the past. According to this account, we can analyze what it is for an agent, *A*, to be omnipotent at *t* in a world *W* in terms of what it is *possible for A* to strongly actualize at *t* in worlds having the same history as *W* up to *t*.

Wierenga's account of omnipotence, like Flint and Freddoso's, relies on the intuitive idea that two possible worlds can share the same past or history up to a certain point in time, and then diverge. According to Wierenga's account, two worlds of this kind share an *initial segment*, where *S* (*W*, *t*) is an initial segment of a possible world *W* up to a time *t*.[17] Unlike Flint and Freddoso's account, Wierenga's account is not stated in terms of what an agent can actualize, but rather in terms of the narrower notion of what an agent can *strongly actualize*. An agent, *A*, strongly actualizes just those states of affairs that *A* brings about *directly* or those actions that *A* does not do *by* doing something else.[18] Of course, an agent may actualize a state of affairs *indirectly* by strongly actualizing *another* state of affairs. Wierenga's account of omnipotence is formulated as follows.

A being *x* is omnipotent in a world *W* at a time *t* =df. In *W* it is true both that (i) for every state of affairs *A*, if it is possible that both *S* (*W*, *t*) obtains and that *x* strongly actualizes *A* at *t*, then at *t x* can strongly actualize *A*, and (ii) there is some state of affairs which *x* can strongly actualize at *t*.[19]

This account of omnipotence may be vulnerable to a counter-example of the following kind. Arguably, there could be an agent, *x*, such that: *x* has a wide range of powers, *x* is essentially limited to these powers, and *x* essentially lacks a power, *P*, which an omnipotent agent ought to possess. Of course, *x* would not be omnipotent. Yet, Wierenga's account of omnipotence paradoxically implies that *x* would be omnipotent. Hence, it can be argued that Wierenga's account does not provide a logically sufficient condition for omnipotence. The assumption that there could be a non-omnipotent agent that is essentially limited in its powers can be defended as follows. An omnipotent agent has the power to overrule (or supersede) any law of nature (a mere *physical necessity*). For example, God has the power to overrule the law of gravity by bringing it about that a mountain floats in midair without any physical cause. Yet, arguably, it is possible for there to be a non-omnipotent agent who essentially lacks the power to overrule any law of nature. For example, it can be argued that there could be a physical or material agent who is *essentially* subject to certain laws of nature. Surely, such an agent would lack the power to overrule any law of nature, and so would not be omnipotent.

A similar, though weaker, sort of objection concerns McEar, a hypothetical agent who essentially has the

power to do only *one* thing, namely, scratch his ear. It may be objected that Wierenga's analysis of omnipotence falsely implies that McEar would be omnipotent. But it might be replied that an agent such as McEar is impossible. It can be cogently argued that, necessarily, if McEar has the power to scratch his ear, then he *also* has the power to move a part of his body to scratch his ear, for instance, his arm.[20] So, it appears that there could not be an agent that has the power to do only one thing. In reply to the stronger sort of objection discussed earlier, it may be suggested that, necessarily, for *any* power, if an agent lacks that power, then an omnipotent being could give that agent that power.[21] The difficulty with such a reply is that there could be a non-omnipotent agent who *essentially* lacks the power to overrule any law of nature, and hence that not even an omnipotent agent could give this non-omnipotent agent that power.

5. Omnipotence and Divine Moral Perfection

It has been argued that the traditional God has incompatible attributes, namely, necessary existence, essential omnipotence, essential omniscience, and essential moral perfection.[22] The contention has been that it is impossible for God to have the power to bring about evil, while non-omnipotent (and morally imperfect) beings may have this power. The precise form of such an argument varies depending on what precisely the relation between God and evil is assumed to be. However, generally speaking, it is argued that divine moral perfection and omnipotence are incompatible because divine omnipotence entails that God has the power to bring about evil, whereas divine moral perfection entails that God is powerless to bring about evil.

One can respond to arguments of this kind as follows. Assume that if God exists, then this is a best possible world.[23] In that case, if God exists, there could not be an evil unless it were necessary for some greater good, in which case any state of affairs containing evil incompatible with there being a maximally good world is *impossible*. But it be may be assumed that it is not possible for *any* agent to bring about an impossible state of affairs. Thus, if God exists, any moral evil, that is, any evil brought about by anyone, and any natural evil, or any evil which has an impersonal, natural cause, must be necessary for some greater good.

Suppose that God exists and that some other person, for example, Cain, brings it about that an evil, *E*, exists. There are two possibilities that need to be considered here. The first is that Cain's decisions and actions are causally determined, as are all occurrences in the created universe. Then, given our assumptions, since Cain's bringing it about that *E* exists is necessary for some good which more than compensates for *E*'s existence, it is consistent with God's moral perfection that God [remotely] brings it about that Cain brings it about that *E* exists.

The second possibility is that Cain's decision to do evil is uncaused by anything other than Cain and free in the libertarian sense. In that case, God did *not* [remotely cause Cain freely to] bring it about that *E* exists, while [let us assume] Cain *did* freely bring it about that *E* exists. If so, then it must be the case that God's creating Cain and permitting Cain freely to do what he chooses to do [in the context of the entire creation] brings about more good than his *not* creating Cain and thus *not* permitting him freely to do what

he chooses to do. It might be objected that if Cain can bring about a state of affairs that God cannot, namely, *that E exists*, then God is not omnipotent. But, as we have seen, an agent's being omnipotent does not require of that agent that it be able to bring about *every* state of affairs which *any* other agent can bring about. It *does*, of course, require that an omnipotent agent have more power than any other agent. And God, of course, *would* have more power than Cain, even though Cain could bring about something that God could not. For there are many more states of affairs that God could bring about and that Cain could not, than *vice versa*. At this point, it might further be objected that an omnipotent agent, one that was morally imperfect, who *could* bring it about that *E* exists, as well as all the other states of affairs that God could bring about, would be more powerful than God. But recall that if God exists, then he exists eternally in every possible world. Recall, too, that there cannot be more than one omnipotent agent. Thus, if God exists, then an omnipotent agent who is morally imperfect is *impossible*. Thus, this second objection is based on an assumption that is impossible, namely, that if God exists there could exist another omnipotent agent who is morally imperfect and who is therefore more powerful than God.

Of course, if God exists, then any evil state of affairs, *s*, which *is* incompatible with a maximally good world is *impossible*. And if *s* is impossible, then neither God nor any other agent has the power to bring it about that *s* obtains. God would lack the power to bring it about that *s* obtains because of his moral perfection, and any created agent would lack the power to bring it about that *s* obtains either because (i) God would not create an agent who had the power to bring it about that *s* obtains, or (ii) God would not permit any created agent to bring it about that *s* obtains. Thus, to the extent indicated, if God's attributes impose moral restrictions on the nature of the universe and on what he can bring about, then they impose parallel restrictions on what any other agents can bring about.

The foregoing line of reasoning implies that God's moral perfection and omnipotence are not incompatible.[24]

This argument about God and the possibility of evil has been disputed by theists such as Alvin Plantinga, who do not hold that God's existence implies the existence of a maximally good world, but do hold that God seeks to create as good a world as he can.[25] Theists such as Plantinga allow for there to be evil that is *unnecessary* for any greater good that outweighs it. An evil of this kind involves free decisions of non-divine agents, which God does not prevent, but which these other agents can prevent. Plantinga contends that God is not wrong to permit an evil of this kind, since God cannot bring about a vital good, the existence of free human agents, without there being such an evil. Alternatively, it might be argued that God does no wrong in this sort of case, because he does not know how to do better (knowledge of the future free actions of created agents being impossible). However, as an omnipotent God is *not* required to have power over the free decisions of non-divine agents, it follows that on these views, his omnipotence and moral perfection are compatible, roughly to the extent indicated earlier in our discussion of the view that God's existence implies a maximally good world. Of course, nothing that has been said here answers the question of how much, if any, evil is compatible with the existence of the traditional God. This question is central to the problem of evil for theism.

Bibliography

- Aquinas, T., *Summa Theologiae* (New York: Benziger Brothers, 1948).
- Curley, E. M., 1984, "Descartes On the Creation of the Eternal Truths," *Philosophical Review* 93, pp. 569-97.
- Descartes, R., *Meditations on First Philosophy*, trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch, in *The Philosophical Writings of Descartes*, *Volume 2* (Cambridge: Cambridge University Press, 1984).
- Flint, T. and Freddoso, A., 1983, "Maximal Power," in *The Existence and Nature of God*, ed. A. Freddoso (Notre Dame, Ind.: University of Notre Dame Press), pp. 81-113.
- Frankfurt, H., 1977, "Descartes on the Creation of the Eternal Truths," *Philosophical Review* 86, pp. 36-57.
- Hoffman, J., 1979, "Can God Do Evil?," *Southern Journal of Philosophy* 17, pp. 213-20.
- Hoffman, J., and Rosenkrantz, G. S., 1988, "Omnipotence Redux," *Philosophy and Phenomenological Research* 49, pp. 283-301.
- -----, 2002, *The Divine Attributes* (Oxford: Blackwell).
- Maimonides, *Guide for the Perplexed*, trans. M. Friedlander (London: George Routledge and Sons, 1904).
- Pike, N., 1969, "Omnipotence and God's Ability to Sin," *American Philosophical Quarterly* 6, pp. 208-16.
- Plantinga, A., 1974, *God, Freedom, and Evil* (New York: Harper and Row).
- Rosenkrantz, G. S. and Hoffman, J., 1980, "The Omnipotence Paradox, Modality, and Time," *Southern Journal of Philosophy* 18, pp. 473-9.
- -----, 1980, "What An Omnipotent Agent Can Do," *International Journal for Philosophy of Religion* 11, pp. 1-19.
- Wierenga, E., 1989, *The Nature of God: An Inquiry into Divine Attributes* (Ithaca, NY: Cornell University Press).

Other Internet Resources

- http://www.courses.rochester.edu/wierenga/REL111/omnipch.html
- http://www.nd.edu/~afreddos/papers/mp.htm
- [Please contact the authors with additional suggestions.]

Related Entries

evil, problem of | perfectionism: in moral philosophy

Joshua Hoffman
University of North Carolina/Greensboro
j_hoffma@uncg.edu

and
Gary Rosenkrantz
University of North Carolina/Greensboro
g_rosenk@uncg.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 21, 2002
Content last modified: May 21, 2002

Stanford Encyclopedia of Philosophy
Notes to Omnipotence

Notes

1. This approach is taken in Rosenkrantz & Hoffman 1980; Flint & Freddoso 1983; and Wierenga 1989, pp. 12-35.

2. René Descartes *Meditations on First Philosophy*, Meditation 1.

3. St. Thomas Aquinas, *Summa Theologiae*, Ia, 25, 3; and Maimonides, *Guide for the Perplexed*, part I, chap. 15.

4. For a discussion of the paradox of the stone and a useful bibliography concerning this paradox, see Rosenkrantz & Hoffman 1980.

5. Such an argument is based on two premises. The first premise is a plausible version of the principle of *the diffusiveness of power* which implies that for any agent, *A*, and for any states of affairs *p* & *q*, if *A* brings about *p*, *q* obtains, and *q* is not within the power of any agent other than *A*, then *A* brings about (*p* & *q*). The second premise is that possibly [(a non-omnipotent agent brings it about that a snowflake falls) & (no omnipotent agent ever exists) & (it is not within the power of any agent to bring it about that an omnipotent agent exists at some time)]. It is plausible that this conjunction is possible provided that an accidentally omnipotent agent is possible. See Hoffman & Rosenkrantz 1988.

6. Hoffman & Rosenkrantz 2002, chap. 8.

7. In *D2*, ‘*n*’ ranges over all natural numbers, and $t_1 \ldots t_n$ are nonoverlapping. In addition, it is assumed for the purposes of *D2* that either it is possible for time to have no beginning, or it is possible for time to have no end (or both).

8. It should be noted that in (C), ‘*n*’ ranges over real numbers, and *p* is not itself equivalent to a state of affairs of the form ‘in *n* minutes, *r*’, where *n* is not equal to zero.

9. A complex state of affairs is one which is either constructible out of other states of affairs by use of the logical apparatus of first-order logic enriched with whatever modalities one chooses to employ, or else analyzable (in the sense of a philosophical analysis) into a state of affairs which is so constructible. Therefore, a *part* of a complex state of affairs, *s*, is one of those states of affairs out of which *s*, or an analysis of *s*, is constructed. The relevant notion of a *part* in this context is that of a logical part, as

opposed to a spatial part or a temporal part.

10. Flint & Freddoso 1983, p. 99.

11. Flint & Freddoso 1983, p. 96.

12. Flint & Freddoso 1983, pp. 96-97.

13. Flint & Freddoso 1983, p. 97.

14.Flint & Freddoso 1983, p. 97.

15. Note that if *W* and *W** sharing the same history up to *t* implies that *W* and *W** share the same natural laws up to *t*, then in *W* there is no sufficient causal condition for Oscar becoming omnipotent at *t*. However, the libertarianism of Flint and Freddoso presupposes that some events, i.e., all free decisions, lack a sufficient causal condition, and there seems to be no good reason to deny the possibility of events that have no sufficient causal condition, especially in the light of current understandings of quantum mechanics.

16. There are variations on (e) that may provide additional counter-examples to Flint and Freddoso's account of omnipotence. For instance, consider the following possible state of affairs:

> (e*) A snowflake falls & ~(Oscar brings about something at some time during his life).

Suppose that Oscar is a contingently existing omnipotent agent. It can be argued that although it is impossible for Oscar to bring about (e*), it is possible that a non-omnipotent agent other than Oscar brings about (e*) by bringing about *both* of its conjuncts, even when Oscar is omnipotent. Arguably, a non-omnipotent agent of this sort could accomplish this by causing a snowflake to fall and destroying Oscar before he brings about something. It can be argued that such a case provides a counter-example to Flint and Freddoso's account of omnipotence similar to the one based on (*e*).

17. Wierenga 1989, pp. 18-20.

18. Wierenga 1989, pp. 20-23.

19. Wierenga 1989, p. 25.

20. Wierenga 1989, pp. 28-29.

21. Wierenga 1989, p. 29.

22. Pike 1969. Pike argues that divine omnipotence and perfect goodness are incompatible. For a discussion of the compatibility of divine omnipotence and perfect goodness, see Hoffman 1979.

23. Whether divine moral perfection should be understood as perfect goodness, perfect virtue, or an optimal combination of goodness and virtue, depends upon whether the correct theory of morality is consequentialist, deontological, or *mixed* (that is, a mixture of core elements of consequentialist and deontological moral theories). To preserve our neutrality on this controversial question in this context, in the main text we use expressions such as ‘best possible world’ and ‘maximally good possible world’ to refer to *either* a possible world of unsurpassable goodness, a possible world governed by a being of unsurpassable virtue, or a possible world with an optimal balance of goodness and virtuous governance. For a discussion of these alternative conceptions of divine moral perfection see Hoffman & Rosenkrantz 2002, chap. 7.

24. Although from what we have said about the restrictions that any coherent account of God’s power must place on this power, a better term for God’s power than ‘omnipotence’ would be ‘maxipotence’.

25. Plantinga 1974. This work is an influential free will defense of theism against the problem of evil. A number of philosophers have argued against some of the presuppositions of Plantinga’s view, and in particular, against the acceptance of so-called counterfactuals of freedom.

Joshua Hoffman
University of North Carolina/Greensboro
j_hoffma@uncg.edu
and
Gary Rosenkrantz
University of North Carolina/Greensboro
g_rosenk@uncg.edu

First published: May 21, 2002
Content last modified: May 21, 2002

Cosmology: Methodological Debates in the 1930s and 1940s

Sometimes, philosophy drives science. Cosmology between 1932-48 provides an excellent example how explicitly philosophical considerations directed the evolution of a modern science during a crucial period of its development. The following article exhibits these philosophical aspects of cosmological thinking in detail, beginning with a brief sketch of the historical development of general relativity cosmology until 1932. Following this, the historical participants in the philosophical debate are introduced, along with the basic ideas of their competing positions. Then the critical stages of the debate -- 1935-37 -- are closely explored by focussing directly upon the arguments of the participating scientists and philosophers. Finally, the concluding stage of the philosophical debate, namely, the emergence of the steady-state theory of the Universe, is presented in the context of its development from Popper's philosophy of science.

---

1. Introduction

One of the most vigorous philosophical debates of the century broke out among cosmologists during the 1930s and 1940s. At the peak of the debate, 1936-37, many of the most prominent scientists in Britain, as well as several leading philosophers of science, had gotten themselves publically involved. Their

arguments, attacks and rebuttals were chronicled in many of the leading scientific journals, including a special edition of the foremost general scientific journal, *Nature*, devoted entirely to philosophical arguments and counter-arguments.

Methodology was the central issue of the debate, although metaphysical questions also arose, particularly those concerning the actual reality of certain structures and forces imputed to the Universe by the new cosmological theories and observations. But in the end, methodology was the real goad spurring on most of the participants.

At bottom, there were just two opposing positions in the debate, each of which comprised a two-point stance. On one side were those scientists who had their roots mostly in the experimental side of natural science. To them, there was one and only one legitimate method for science. Theory construction, they believed, involved two closely-linked steps. First, one began from the *empirical observations*, that is, from measurements, manipulations, experiments, whose results were evident to the human senses; this is classic *empiricist* epistemology. Observational results would then suggest possible hypotheses to examine via further empirical testing. When enough data concerning the hypothesis had been gathered, *logical generalization* could be carried out, thereby producing a theory; this is classic *inductivist* logic.

Opposing these inductive-empiricist scientists were those whose roots were mostly in the theoretical side of natural science, most especially mathematical physics. To them, there was another, more logically sound, method to construct theories. First, *hypotheses* could be generated in any fashion, although most believed that imagining hypotheses which were based upon very general, very reasonable concepts—that the Universe's physical processes had simple mathematical descriptions, for example—was the best place to begin; this is classic *rationalist* epistemology. Once the hypothesis had been generated, strict analytical reasoning could be used to make predictions about observations; this is classic *deductivist* logic. Scientists who held this view came to be called *hypothetico-deductivists*; their views about both hypothesis generation and deductive predictions were each strongly opposed by the inductive-empiricists.

Part of the controversy may be laid to the fact that cosmology was a new science, and disputes about methodology in new sciences are not rare in the history of the sciences. What is rare about this case, however, is the vigor, sometimes even bitterness, with which the philosophical controversy was waged. Another reason for the controversy lies in the fact that cosmology is a data-poor science: observations are hard-won and rare, and they frequently must be run through elaborate theoretical manipulations and corrections in order to make sense at all. With a paucity of data results, scientists must rely upon philosophical argument to undergird their views about how the scientific work should be done.

One final feature of the debate must be noted. The participants are almost universally scientists, and not philosophers. Yet this does not much affect the level of philosophical thinking going on; these scientists knew their philosophy well, and they wielded philosophy's weapons and defenses with great skill. In the end, their debate shaped cosmology into the science we know today.

It will be useful to look briefly at the history of cosmology leading up to the debate.

2. The Lead-up to the Debate

Since about 1700 theories about the nature and structure of the Universe were derived from Newtonian theory, most especially his theory of gravitation, which was used to account for the behavior of heavenly bodies and their systems. Newton's theory hypothesized a force—gravitation—acting upon material bodies, free to move over time within the passive, inert ‘container’ of three-dimensional space. Bodies, paths, and space itself exemplified the classical geometry of Euclid. All these features changed with the publication of Einstein's General Theory of Relativity, 1915-17.

2.1 Einstein's General Theory of Relativity

Einstein's intended his theory to replace Newton's theory of gravitation completely. In Einstein's view, gravity was not a force existing independently of the spatial ‘container’; rather, gravitation arises as a curvature of the space (and time, which is necessarily connected to space in the new theory), which means that geometry and gravity and astronomical behavior are all intimately connected. For example, near the sun the geometrical structure, the curvature, of spacetime changes radically, which expresses itself as an increasing velocity of incoming orbiting bodies such as comets or satellites. One immediate, and to some, puzzling, consequence of Einstein's theory is that the geometry of the Universe is no longer taken to be Euclidean. Although there are several different candidates for the actual geometry of space, it was not known which is correct.

It was not recognized at first that the General Theory of Relativity could be applied to the Universe as a single, whole, individual object, thereby producing a new cosmological theory, one completely different from its predecessors. Although the mathematics involved are extremely difficult, two solutions, one by Einstein himself, the other by the Dutch astronomer Willem de Sitter, were produced in just a short time in 1917. Unfortunately, the universes predicted by the two solutions were extreme: Einstein's universe would be densely packed with matter, whereas de Sitter's would be essentially empty.

Obviously, the universe as observed by astronomers did not conform at all to the description provided by either solution, a fact many found troubling. Moreover, no additional solutions were forthcoming (even though both Friedmann and LeMaître had developed alternatives, they remained unknown and unnoticed). For nearly twelve years, the new cosmology appeared to be going nowhere. Then Hubble at California's Mt. Palomar made public his astonishing observations of a cosmic Doppler shift, a shift toward the red in the color of light coming from the most distant star systems.

2.2 Hubble's Expanding Universe

Most cosmologists—with the interesting exception of Hubble himself—came to the immediate conclusion that the red shift could only mean that the universe was expanding. Immediately the relativity theorists were able to interpret the expansion as a continuous change in the geometry of spacetime, which was thoroughly accounted for by the General Theory of Relativity. After over a decade of stagnation in

face of the meager choice between just two models of the cosmos, Hubble's observations spurred theorists on to the construction of a melange of new models, each vying in competition with the other.

In the end, it was the Belgian astronomer Georges LeMaître's theory of an expanding universe that came to be accepted. LeMaître's model was publically proclaimed as appropriate and generally correct during a special session of the British Association for the Advancement of Science, 31 Oct 1931. Modern scientific cosmology had been officially born; because of its birth within the context of Einstein's theory of relativity, the new cosmology became quickly and broadly known as *Relativistic Cosmology*. The model of this cosmology is most famously that of the blowing up of a balloon painted with dots to represent galaxies. Over time, the radius of the model's spherical space (the balloon) increases, thereby decreasing the curvature of the space (the balloon's skin), and increasing the distance between the dots.

Although cosmologists came from Europe and America as well as Britain, most of the work in theoretical cosmology took place in London, Cambridge and Oxford. Americans Hubble, Tolman and Robertson did their work at CalTech in Pasadena, but were frequently in England; and most of the European and British workers cycled through Pasadena at one time or another. De Sitter, from Holland, and LeMaître, from Belgium, spent important periods in England, as did various of the German workers. Thus, even though cosmology was done throughout the Western world, its major concentrating point was England; our focus in what follows will be the same.

3. Cosmology and its philosophy

Because it represented such a radical departure from previous scientific thinking about the Universe, relativistic cosmology needed to work out its philosophical underpinnings, most especially in regard to its methodology. ‘How is this new science to be conducted?’ was thus a compelling question. But method, of course, is always linked to metaphysics and epistemology. A full-blown philosophical discussion was evidently required. It came soon enough: within a span of less than a year, a vigorous debate between two philosophically opposed camps developed. The debate required nearly two decades to reach its full resolution. But, with the resolution, cosmology had philosophically certified its methodology within the context of a concensus metaphysics and epistemology. Before going into the details of the debate, however, several general points should be noted. Let us turn to the details of the origin of the debate.

3.1 Relativistic Cosmology: the majority philosophy

To begin with, the cosmological thinking of the majority of scientists—including Eddington, de Sitter, Robertson, Tolman, and their colleagues—had betrayed a relatively unexamined, and apparently uncontroversial cluster of philosophical perspectives: on the metaphysical side they held a modest sort of *explanatory realism*—if an accepted theory referred to entity *x*, then *x* was acceptable as a genuinely, physically real object; to this metaphysical stance was coupled a methodology exhibiting a classic sort of *inductive empiricism*—scientific knowledge consisted of generalizations built up from individual empirical observations. But this cluster of views did not go long unchallenged. In early July 1932, just nine months after relativistic cosmology became the concensus during the British Association meeting,

Oxford astrophysicist E. A. Milne published a short article in *Nature* which directly attacked the current philosophical tenets, proposing their replacement by a new cluster of views, one as radical as the new science it purported to undergird.

3.2 Milne's Philosophical Challenge

Milne's metaphysical views were based in positivism, most especially in *operationalism*: only those objects whose properties could be directly revealed by some observational procedure, or operation, were to be counted among the real. Thus, for Milne, reality did not contain the usual relativistic cosmology referents “curved space”, “expanding space” or even “four-dimensional spacetime”, simply because none of these entities were operationalizable. Milne held that ‘what you saw was what you got’; since “curved space” of various geometries couldn't be observed, space was just as it looked—Euclidean. In the same way, the expanding of the Universe was a genuine expanding—each galaxy was retreating from each other—not just a change in the geometry of the curvature, as it was for the relativists. Moreover, in place of relativistic cosmology's inductive empiricism, Milne opted for a *hypothetico-deductive rationalism*. Cosmologists, Milne believed, were bound to dream up any and all possible models of a universe, and then deduce what, if any, observable consequences followed from their hypotheses. Hypothesizing could be based upon just about anything, although Milne believed that certain very general rational principles derived from aesthetic beliefs about universal order and regularity would be the most fruitful. His ‘*cosmological principle*’, namely, that every cosmologist in the universe should look out upon the same cosmos, was the prime example of such thinking. Later, Bondi would found Steady-state Cosmology on an even more general version of Milne's principle, which he called the ‘*perfect* cosmological principle.’

3.3 Kinematic Relativity—an alternative cosmology

Milne's philosophy of cosmology came with a closely associated cosmological model, *kinematic relativity*, so-called because of its tight links to the kinematics of Einstein's Special Theory of Relativity. Kinematics, in Einstein's Special Theory, is especially concerned with observers trying to make measurements of the behavior of objects moving in systems relative to themselves. According to Milne's understanding, these measurements could only be made by the observers signalling among themselves using light or other electromagnetic radiation, using clocks to time the signals. In the end, all measurement—even space and velocity—reduced to timing the signals as they went from one observer to another. This idea of measurement represented a revolutionary simplification of previous ideas, and held philosophical interest for that reason alone. There were practical consequences as well. As Bondi later remarked, “Milne's idea led straightaway to the radar speed gun!”

Milne's kinematic relativity and its underlying philosophy both immediately brought careful scrutiny, not to mention controversy.

4. The Great Cosmological Debate Begins: 1933-1934

Eddington was the first to attack Milne's views. In a series of lectures given at Harvard during late '32 and published early in the new year as *The Expanding Universe*, Eddington denied the efficacy of both operationalism and hypothetico-deductivism, and not only defended explanatory realism, but strengthened his ontological position heroically: the theoretical entities of relativistic cosmology were not just plausible, they were so necessary to understanding the universe, that cosmological knowledge was essentially impossible without them. (Eddington 1932, p. 19)

4.1 Dingle's First Attacks

Few others would ever go so far as Eddington's ontological heroism; yet epistemological and methodological heroism in the fight against Milne was not rare. Chief hero in this aspect of the attack on Milne was Herbert Dingle, a highly respected astrophysicist, and then-secretary of the Royal Astronomical Society. Dingle's initial foray appeared as a response to Milne's first detailed presentation of kinematic relativity. (Milne 1933) After claiming that while kinematic relativity did not differ appreciably from relativistic cosmology in its mathematical formalism or observable consequences, Dingle asserted that, on the other hand, it invited special criticism because it renounced “the fundamental principles of scientific method,” namely, “Newton's principle of induction from phenomena.”(Dingle 1933 p178) Dingle was never to relent in his attacks upon Milne's hypothetico-deductivism, at all times rejecting the methodology as even a possible candidate for acceptance by genuine science.

Dingle's article appeared back-to-back with an appraisal of kinematic relativity by the important American cosmologist H. P. Robertson. (Robertson 1933a) Robertson focussed upon Milne's hypothetico-deducivism as well, noting that the cosmological principle in kinematic relativity functions as an *a priori* rule, rather than as an empirical generalization, its status in relativistic cosmology. Robertson otherwise is not especially taken with Milne's theory, limiting himself to remarks suggesting that, where they can be compared, kinematic relativity and relativistic cosmology apparently are similar in physical content. As we shall shortly see, Robertson's later work strongly belied this earlier apparent indifference to Milne's theory.

4.2 Two Ways to Disagree with Milne

Right from the start, cosmologists differed in their opinions about both the physical content of Milne's theory, and its underlying philosophy. Dingle disliked Milne's methods, but found the theory itself of not much interest. Robertson initially agreed with Dingle, but within a short time had significant changes of heart. Eddington disliked both the theory and its philosophy, finding them far too deviant from what was typical, i.e., relativistic cosmology and its mainstream philosophy. Younger cosmologists such as McVittie and McCrea, students, respectively, of Eddington and Whittaker, soon joined the fray. McVittie initially found the physics of kinematic relativity quite interesting, and quite different from the physics of relativistic cosmology. (McVittie 1933b) But Milne's philosophy was something else; the strict empiricism McVittie had earlier revealed (McVittie 1933a) made him equally unhappy with both Milne's rationalism and his hypothetico-deductivism.

McCrea's remarks were among the most perceptive and favorable. (Kermack and McCrea 1933) While he thought that Milne's operationalist criticisms of curved and expanding space were of little import, McCrea was the first to notice the parsimony and elegance of Milne's strictly kinematic solution to the problem of the origin of the universe's expansion. Since a search for such a solution had vexed relativistic cosmology for several years (McVittie 1931), McCrea suggested that Milne's mechanism should be immdiately shown to be part of relativistic cosmology. (Kermack and McCrea 1933 p. 529)

Clearly, the so-recently won consensus about relativistic cosmology and its philosophy had dissolved into confusion and controversy over Milne and his methods.

4.3 Milne Makes Philosophical Improvements

Over the next year or so, Milne made strenuous efforts to elucidate both kinematic relativity's physics and its philosophy, beginning, as would be his wont, with the philosophy. In October, Milne addressed the Philosophical Association, giving an explicit and detailed analysis not only of his philosophical views, but also of their history, which, he claimed, extended back to Locke and Hume (Milne 1934a) His description of the two opposed methods—inductive empiricism vs. hypothetico-deductivism—is quite clear and useful:

> Strictly speaking, physics has no philosophy. It has method…Now the methods of theoretical physics seem to be reducible to two species, the method of starting with concepts and the method of starting with things observed. …When a subject is developed from concepts the concepts play the part of the terms occurring in the axioms of geometry…. The concepts are undefined save as being governed by propositions of which they are subjects.

Milne's commitment to axiomatization is notable here. It was based in his earlier admiration for the work of Whitehead and Russell (Crowther 1970); commitment to axiomatization in cosmology, once having been initiated by Milne, would be an enduring hallmark of the work of many, including both Robertson and Walker.

Three weeks after this address, Milne spoke on related topics at the monthly meeting of the Royal Astronomical Society. His main point was that theories differed only insofar as their concepts could be cashed out in observations deduced from them. Eddington took strong exception to Milne's arguments, claiming in response that kinematic relativity and relativistic cosmology differed more importantly in their ontology than in their consequences. Eddington was especially concerned with differences in the spacetime-geometry each theory ascribed to the world. For Eddington and his colleagues, the equivalence of gravitation and spacetime geometry was a genuine reality, a feature of the physical world, just as real as suns and moons and stars.

4.4 A Major Philosophical Issue: What makes a scientific theory

‘good’?

Milne was having none of it. In the next month's *Observatory*—the informal monthly publication of the Royal Astronomical Society—he took Eddington and his theoretical-realist colleagues to task, concluding that “theories differ simply and solely when their predictions as to phenomena differ”; most importantly, “this method of comparison avoids all reference to distance-assignments, world-geometry, schemes of projection or the like.” (Milne 1934b) In other words, metaphysics was to be avoided in cosmology; space, spacetime, geometry and the like were to be rejected as scientific realities, replaced by reference simply and solely to observations. The only realities, according to Milne, were what could be reported among observers about light signals and clocks.

During 1934 Milne worked together with his new student A.G. Walker. Walker never evinced much interest in the philosophical aspects of kinematic relativity, choosing instead to focus tightly upon working out the physical details of the theory itself. He had immediate success. (Walker 1934) One of his important conclusions was that other authors, specifically McVittie and Robertson, were wrong to conclude that the physics of Milne's theory ultimately corresponded to relativistic cosmology: “*Milne's system is fundamentally different from that of general relativity*.” (Walker 1934 p. 489; emphasis in original)

4.5 How to Choose Among Theories and Philosophies?

In an important review of the entire confused situation between kinematic relativity and relativistic cosmology, McVittie confessed that “experimentally it seems hopeless to discriminate between them…at present the choice is almost entirely a matter of personal taste.” (McVittie 1934 p. 29) At almost the same time, de Sitter took on Milne in serious fashion. (deSitter 1934) Responding to Milne's methodological challenge, he showed that, indeed, it is possible to formulate relativistic cosmology in axiomatic fashion, just as Milne had formulated kinematic relativity “from concepts.” But de Sitter explicitly rejected Milne's philosophical use of the cosmological principle, “which asserts that statistically the world pictures of two different observers must be the same.” His objection is founded on the matter-of-fact that “we have, however, no means of communicating with other observers, situated on faraway stars, or moving with excessive velocities.” (deSitter 1934 p. 598) So much for rational principles as hypotheses!

The year in cosmology ended almost as confused as it had begun, with one exception: Milne had gotten much clearer about his philosophical views, and was applying them to an exhaustive presentation of his cosmology, theory and philosophy. His book *Relativity, Gravitation and World Structure* (Milne 1935) would be published in just a few months.

5. The Triumph of Milne's Methods 1935-36.

The new year marked a sudden change. In short order, McCrea, Walker and Robertson succumbed to Milne's methodological recommendations: first, to carry out an operationalist paring of non-observational

concepts, then, secondly, to embed the resulting minimalist concept set in an axiomatic hypothetical-deductive structure. Thus was the famous Robertson-Walker spacetime metric born.

5.1 McCrea, Walker and Robertson Adopt Milne's Methods

McCrea's effort operationalized the concept of “distance”, principally and originally by comparison of certain elements of Newtonian cosmology and de Sitter's axiomatized version of relativistic cosmology. (McCrea 1935) Walker's paper specifically eschewed use of “any indefinable concepts”, in particular, he did not assume that the “associated metric [of relativistic cosmology] has any *a priori* physical significance.” (Walker 1935) Robertson's article, the first of three, is the most important, both in its content, and in the signal it sends, namely, that one of the original mainstream relativistic cosmology proponents has adopted a major element of Milne's new philosophy for cosmology. (Robertson 1935) Robertson's conclusion exhibits this point clearly and explicitly:

> We have examined, from the operational standpoint, the problem of determining the most general kinematical background suitable for an idealized universe in which the cosmological principle holds. Allowing the fundamental observers the use only of clocks and theodolites, and granting them the possibility of sending and receiving we have shown that for each given mode of motion *x(t)* there necessarily exists a quadratic line element which is invariant, in form as well as in fact, under transformation fqrom one fundamental observer to another. (Robertson 1935 p. 300)

Unlike de Sitter, Robertson accepts the cosmological principle, replete with its observers on far-separated particles. Moreover, as this statement shows, Robertson is intimately familiar with Milne's latest operationalist reduction: space is to be reduced to time measurements given by clock readings on signals exchanged between observers. Robertson gets this idea from Milne's book, which he had earlier reviewed (but which was only subsequently published) for *Astrophysical Journal*. (Robertson 1936)

5.2 But Eddington Scoffd…

Eddington disparages these same methods. In his scathing *Nature* review of Milne's book, he rejects Milne's hypothetico-deductivism, his cosmological principle, and, above all, his operationalism: “When I visit the Cavendish Laboratory, I do not find its occupants engaged in flashing light-signals at each other, but I find practically everyone employing rigid scales or their equivalent.” (Eddington 1935, p. 636) Whittaker's review was not so negative as Eddington's. (Whittaker 1935) While the senior physicist rejects Milne's operationalism and attacks upon the geometrical commitments of relativistic proponents, he is considerably more forgiving about Milne's hypothetico-deductivism, and even goes so far as to remark Milne's “brilliant record in astrophysical discovery.” Nonetheless, Milne's break with a tradition including at least “Einstein, de Sitter, Friedmann, LeMaître, Weyl, Eddington, H.P. Robertson and others” is to be regretted. (Whittaker 1935 p. 179) Perhaps, along with Eddington, Whittaker hopes that soon “Professor Milne will return to orthodoxy.” (Eddington 1935, p. 636)

But Whittaker's view on Milne must be put into the perspective he held on the whole ongoing debate. As he saw it

> …a lively debate is in progress at the present moment between Sir Arthur Eddington and Dr Harold Jeffreys of Cambridge, Professor Milne of Oxford, Sir James Jeans, and Professor Dingle of the Imperial College, the subject being the respective shares of reason and observation in the discovery of the laws of nature. (Whittaker 1941, p. 160)

But lively debate is far too gentlemanly a description for what now occurred. After holding his ire somewhat in check for—as he saw it—already far too long, Dingle finally erupted.

6. Dingle's Denoument

Controversy over Milne and his philosophy reached a crescendo in mid-1937. Dingle, his stew having finally boiled over, wrote privately to the editor of *Nature*, first castigating the rampant cosmological ‘mysticism’ passing itself off for science, and then offering to produce an article taking the sword to the mystics themselves. His offer was immediately accepted. The result was Dingle's notorious “Modern Aristotelianism”, a polemical diatribe chiefly against Milne, but aimed as well at Eddington and Dirac on account of their “betrayal” of the scientific method of Newton and his fellow members of the Royal Society. (Dingle 1937)

6.1 Modern Aristotles?

The article is remarkable both for its style and for its content. Dingle's style in the article is vituperative. Thus, emotionally-loaded terms such as “paralysis of reason,” “intoxication of the fancy,” “‘Universe’ mania”, and the like frequently appear, these to be topped only by references to “delusions,” “traitors,” and, of course, “treachery,” each associated with one or more of the guilty parties. (Dingle 1937, p. 786)

Above and beyond his extreme language, Dingle makes certain substantive claims bearing directly upon central philosophical questions. The issue, as he sees it, is nothing more than the question “Whether the *foundation* of science shall be observation or invention” (Dingle 1937, p. 786). As always, talk about ‘foundations’ is philosophical talk. The two opposing positions Dingle here calls “foundational” involve views on both method and epistemology, suitably tangled together. Dingle delineates the opposed alternatives as follows. The way of true science, he claims, shows that “the first step in the study of Nature should be sense observation, no general principles being admitted which are not derived by induction therefrom.” (Dingle 1937, p. 784) Stated more explicitly, Dingle here argues that authentic science is **empiricist** in epistemology (scientific knowledge is founded in sensory observation), and **inductivist** in method (general principles are reached via inductive logic). Opposed to this view, he argues, is “the doctrine that Nature is the visible working-out of general principles known to the human mind apart from sense perception.” (Dingle 1937, p.787) As representative of this latter view Dingle cites Milne, and refers in particular to Milne's claim that “it is, in fact, possible to derive the laws of dynamics

rationally…without recourse to experience." (Milne 1937, p. 329) Obviously, Dingle is here arguing against the view, Milne's view, that authentic science may be **rationalist** in epistemology (scientific knowledge is founded in pure theoretical reasoning apart from sense perception), and **hypothetico-deductive** in method (general principles are justified by their deductively implying correct observations).

Along with Milne, Dingle indicts Eddington, and, by implication, Dirac, all three of whom, Dingle believes, are guilty of inventing scientific hypotheses by free mental imaginings rather than by strict immersion in observations and observational data.

6.2 Dingle as 'True Believer'

What is going on here? Put bluntly, Dingle is an old-fashioned empiricist and inductivist. He believes that the only way to do true science is to first collect data, then, and only then, to hypothesize on the basis of that data. Observation, then hypothesis. As he sees it, Eddington, Milne and Dirac have got it exactly backwards. They first (as he terms it) assume an hypothesis, then, and only then, go about collecting data. Except, according to his lights, the data isn't ever collected: "to [the Aristotelians'] modern representatives it seems as though a fancy is no sooner in the head than it is on paper and sent for publication." (Dingle 1937, p. 785) Obviously Dingle is simply wrong; it never occurred to his opponents that hypotheses would **not** be followed immediately by attempts at deductive prediction of observational consequences. But it was enough, in Dingle's mind, that they didn't use induction, for them to come under blame.

6.3 Wrong from the Very Start

But there is something else at work here as well. Dingle doesn't object solely to his opponents' lack of inductive logic. Of equal importance is the fact that they find the source of their hypotheses in fairly general principles, wide-ranging rational proposals about the structure of the universe at large. These principles Dingle takes to be *a priori*, in the most pejorative sense of that term. They are phantasms, "chimeras" he calls them, which seduce the imaginations of his opponents, and lead them and their dumb-struck admirers away from the genuine, authentic method of science. **This** is what really sticks in Dingle's craw. In turn, Eddington, Milne and Dirac are chastised, each for something slightly different, but at bottom the same, namely, they one and all "appear as a victim of the great 'Universe' mania." (Dingle 1937, p. 786) In the end, Dingle believes, the danger of this new 'methodology' is real, and serious. As he notes in conclusion:

> Nor are we dealing with a mere skin disease which time itself will heal. Such ailments are familiar enough; every age has its delusions and every cause its traitors. But the danger here is radical. Our leaders themselves are bemused, so that treachery can pass unnoticed and even think itself fidelity. It is the noblest minds that are o'erthrown…the very council of the elect can violate its charter and think it is doing science service. (Dingle 1937, p. 786)

Here Dingle obviously goes over the top. Yet overblown as it is, there is no doubting his sincerity: Milne

and the other cosmologists have betrayed the true science bequeathed them by their ancestors in the Royal Society.

How could Dingle be answered?

6.4 The Debate Goes Very Public

The response arrived three months later, on 12 June. On this particular Saturday in June, *Nature* published a fifteen-page special supplement as No. 3528. Contained within were contributions from sixteen “representative investigators”, as the editor referred to them, each responding to “Modern Aristotelianism” *Nature*'s Editor, R.A. Gregory, introduces the occasion by noting that “in *Nature* of May 8, we published an article by Dr. Herbert Dingle entitled ‘Modern Aristotelianism’”. Because the article, as Gregory goes on to say, “created considerable interest”, *Nature* “decided to invite further contributions on the subject from a number of representative investigators.”

“Created considerable interest” is, to understate the issue, an understatement. Some of the contributors were quite obviously livid with rage and other volatile emotions. Others, such as Milne himself, who had come in for particularly scathing criticism in Dingle's article, were patient and careful in rebuttal. Each of the sixteen contributors to the special article chose a side in the controversy, either pro Dingle's inductive empiricism, and con Milne et al.'s rationalist hypothetico-deductivism, or vice versa. Remarks made by the participants exhibit the full diversity of philosophy of science in their contemporary community. Dingle's views, in particular, were not without favor.

Harold Jeffreys, F.R.S., noted geologist and astronomer, and author of a well-regarded philosophy of science book *Scientific Inference* (1957), led off with a nice ad hominem: “Without using induction, Milne and Eddington could not order their lives for a day, and what they are really asserting is that they are entitled to use special axioms in physics, for which no need has been shown.” Jeffreys' criticism here of course ignores the role of deductive observations in justifying the “special” axioms. The problem, as Jeffreys sees it, originates in the perpetrators' “belief that there is some special virtue in mathematics.” L.N.G. Filon, F.R.S., vice-chancellor of the University of London agrees on this point, noting that “some men of science appear to think that they can solve the whole problem of Nature by some all-inclusive mathematical intuition.” R.A. Sampson, the Astronomer Royal, focusses upon the rationalistic aspects of the ‘modern Aristotelians’, to wit, for their “framing a theory independent of experience, such as is denounced in Dr. Dingle's article”, which produces work not unlike that “of a poet or other humanist, who gives us at most a number of illustrative cases.”

6.5 The Counterattack

But Milne, Eddington, and Dirac had their supporters as well. N.R. Campbell, whose theory of science was already well-known, makes an uncontroversial interpretation of the affair. “Science” he begins, “(or at least physics) has long consisted of two distinct but complementary activities”, one of which is experimental and empirical; “its procedure is induction.” The other activity attempts to provide

explanations of scientific laws, which explanations have the “pecularity” that “they often (not always) predict new laws in addition to explaining old ones.” Campbell cannot resist ending on an ad hominem of his own: “If he [Dingle] does not deem it important to observe the distinction between what is and what is not demonstrable experiment, surely he should welcome a movement to amalgamate the Royal with the Aristotelian Society.” Indeed.

G.J. Whitrow, then a young lecturer at Christ Church, Oxford, returns to the mathematical theme. Dingle, he argues, “not only attacks the particular methods adopted by contemporary mathematical investigators in relativistic cosmology, but even refuses to admit that this subject is worthy of scientific investigation as it is based not only on experience but also on reason.” Hypotheses, by this light, may originate rationalistically as well as any other way, certainly there is no problem with this.

6.6 The Coolest Voice

The clearest, most temperate discription of the issues at hand is given by young cosmologist William McCrea, then professor of mathematics at Queen's, Belfast, and editor of the R.A.S.'s *Observatory*. Not to be outdone by Jeffreys, McCrea begins with an ad hominem of his own: “Dr. H. Dingle's objection to ‘modern Aristotelianism’ seems to be itself what he would call Aristotelian rather than Galilean.” In other words, Dingle raises a non-empirical objection about Milne et al.'s non-empiricism! But McCrea soon gets to the heart of the matter, the role of hypothetico-deductivism in mathematical physics:

> What Dr. Dingle has done is to reopen the question of the relation of mathematical physics to experimental physics, since he claims to detect a new and perverted point of view in the former. Now a system of mathematical physics, apart from the alleged perversion, is the working out of the mathematical consequences of certain hypotheses. The worth of the theory is judged…by the closeness of the agreement of its predictions with the results of observation, and also the number of phenomena which it can so predict from the one set of hypotheses. The scientific attitude is, not to cavil at the attempt, but to see if it is successful.

This is an absolutely standard interpretation of how the H-D method works. Throughout his own writings, beginning right from his inaugural lecture in Oxford (Milne 1929), Milne had subscribed to precisely the same interpretation of the Hypothetico-Deductive (H-D) method. Whatever the controversy is about, the issue is not how to interpret hypothetico-deductivism. That much is evident.

Moreover, it is quite clear that Dingle et al. are not mounting opposition to something we, today, would consider philosophically radical; rather, they are objecting to what, today, would be considered completely unobjectionable. Given today's acceptance of the H-D method, yet its rejection by otherwise well-regarded scientists at that time, it seems to follow that it was, at least in part, this debate and its followup which settled the issue. In any case, Dingle and his supporters generally went silent, restricting their activities for the most part to books, or relatively positive statements of their own positions. (A.D.R. 1938, Dingle 1938) Things settled down, just in time for the War.

7. The Calm Between the Storms

During the next several years, it became evident that Milne's methods, and kinematic relativity as well, had reached respectability. One important sign of this progress was exhibited at an early 1939 joint meeting between the Royal Astronomical Society and the Physical Society of London, The meeting had as its goal a thorough review of the situation in cosmology. McVittie was chosen to present the observational situation; his report was soon published. (McVittie, 1939) Reviewing the theoretical situation was George Temple, one of the most highly respected mathematicians of the time. Temple's report saw print almost immediately. (Temple 1939) Within a short time, Temple's paper took on the role of successor to Robertson's definitive 1933 “Relativistic Cosmology.” (Robertson 1933b)

7.1 Two Equal Competitors

Both McVittie and Temple presented kinematic relativity and relativistic cosmology as equal competitors in accounting for the cosmological observations. Unfortunately, as McVittie noted, observations could not, at that time, discriminate between the two theories. Temple's analysis of Milne's work praises its simplicity and elegance, and refers in particular to its operationalism and axiomatization, which “start from a completely novel discussion of the correlation of measurements made by different observers in terms of light signals only.” (Temple 1939, p. 468) Throughout the rest of his discussion of the two theories, Temple utilizes Milne's light-signal correlation method, explicitly rejecting rigid-rod transport for distance measurement. Milne's methods have triumphed.

Later that year McCrea publishes an important paper in *Philosophy of Science*. (McCrea 1939) Put most simply, the paper starts out to defend Milne's methods, but ends up by presenting a full-blown and interesting, although quite unhistorical, account of the evolution and structure of physical theories.

7.2 The Origin and Evolution of Theories

McCrea's overall view is that theories are set up to be hypothetico-deductive in structure. His account is based on his view of the evolution of theories of space-time and mechanics, beginning with Newton, through the General Theory of Relativity and ending in kinematic relativity. His argument reduces to the claim that, insofar as Milne and e.g., Newton, can be shown to follow the same procedure, any attack upon Milne is also an attack upon Newton. First, he states his goals in the paper.

The first goal is to emphasise how each theory leaves us in a position in which the succeeding one appears as a perfectly natural next step in the development of ideas. (McCrea 1939, p. 137) McCrea embeds this argument in an account of how analogue models (à la Campbell) are used to set up new theories—this is essentially an account of how discovery might proceed in linking an older theory with its successor.The second goal is to show how, in spite of superficial differences in character, the theories in question all necessarily possess the same general structure constituted by the presence of hypotheses, from which certain general mathematical relations are deduced, which in their turn are used to predict

relations between observable quantities. As McCrea notes, “this study may claim an interest of its own, but it is presented also for a further reason” namely, that

> it has been contended [by Dingle, most especially] that theories like Milne's represent a fundamentally new outlook on the part of some theorists, in that such theories are purely mental constructs divorced from experince of the physical world. We shall see that on the contrary Milne's theory is easily brought into line with the others in such a way that this criticism is neither more nor less true of it than of the rest. (McCrea 1939, p. 138)

In McCrea's discussion of the theories he asserts that “the consituents [of the theories themselves] which are of physical significance are sets of mathematical relations, coupled with sets of rules of interpretation” which yield, “after observational test, descriptions rather than explanations of physical phenomena.” According to McCrea, one real advantage of his view is that it “leads to simple criteria for comparing the merits of different theories.” Finally, on the metaphysics of the original hypotheses, McCrea claims that “the initial hypotheses from which the mathematical relations are deduced do not ultimately have any direct physical significance.”

McCrea's paper, published in the leading philosophy of science journal of the time, is the final imprimateur on Milne's views.

7.3 Milne's Ultimate Success

Three years later, Milne was awarded the James Scott prize, the most prestigious award for ‘natural philosophy’ in the Anglophone world. Milne's lecture title is telling: “Fundamental Concepts of Natural Philosophy.” (Milne 1943) Although Milne does concede to Dingle that he no longer believes that it is possible to deduce physics completely in the absence of reference to phenomena, for the most part his award lecture is a long reiteration of his previous twelve years' work in cosmology.

One year later, Milne received his ultimate accolade, election as president of the Royal Astronomical Society. In his inaugural lecture, Milne again reviews his work, but adds two remarks of interest. First, he modifies his earlier view that theories are acceptable solely on the basis of their successful predictive power; to this, he now adds that a theory cannot be accepted as satisfactory unless it is philosophically satisfying. (Milne 1943, p. 120) Secondly, on a personal note, he admits that he is still amazed at the outcry that his theory and its philosophy caused. Milne here is being a bit disingenuous. In many places in his letters he not only recognizes the outcry, he delights in it, and seeks to provoke it even more. (Milne 1932-37, 12 May 35; 28 Jul 36)

From this point onward, cosmology's philosophy is no longer directly influenced by Milne himself. Moreover, kinematic relativity began to stagnate as a research programme; except for Whitrow, Milne had no new students, and failed to attract any new converts to the theory. His work was done. But his philosophical influence didn't end, in fact it wasn't to crest until the end of the 40s in the work of another man, Hermann Bondi. Again, however, a storm was generated by Milne's methods, even though they

were now in the hands of another.

8. Steady-state Cosmology

In 1948, a young mathematician, Herman Bondi, in concert with two close friends Thomas Gold and Fred Hoyle, proposed a radical new cosmological theory, the *Steady State* theory. This theory differs from the basic picture shared by both kinematic relativity and relativistic cosmology, namely, that of a universe with a definite origin in a small, dense knot, followed by evolution into the universe we have today. According to Bondi's theory, the universe as far back into the past as we might look would always look the same; there was no evolution, there could be no “fossils”, as Bondi called putative evidence of a universe different in the past from our present one. What we observe today is the same state of a universe that has been and always will be steady. Bondi came to his notion of the steady state primarily from his commitment to the philosophical components of Milne's work, most especially the methodology of rationalism plus hypothetico-deductivism; additionally, Bondi coupled to these Milnean notions some ideas taken directly from the philosophy of Karl Popper.

8.1 Bondi's Philosophical Origins

Bondi reveals his philosophical commitments in several ways. First, he argues against induction and extrapolation from small-scale experiment, that is, against the inductive empiricism of Dingle et al. Secondly, he argues in favor of hypothesis and deduction, that is, in favor of Milne et al. Finally, he specifically remarks the excellence of Milne's Methods, and the theory—kinematic relativity—created therefrom, and remarks the significance of these elements in the creation of his version of the new steady-state cosmology.

From the very beginning Bondi admits the validity of both positions in the methodological debate:

> In particular, there are two important approaches to the subject [cosmology] so different from each other that it is hardly surprising that they lead to different answers...The contrast between the 'extrapolating' and the 'deductive' attitudes to cosmology is very great indeed. (Bondi 1960), p. 3-5)

The *extrapolating* approach, which Bondi sometimes calls the *empirical school*, is represented by Dingle, McVittie, and their colleagues. Opposed to the extrapolative approach is the *deductive* approach, which “is reached from investigations in the borderland between physics and philosophy.” Milne is obviously the major proponent of this view. Although Bondi finds good points in both approaches, he also finds problems in both approaches. In the end, cosmology is the worse for excesses from **either** end of the spectrum:

> Just as some adherents of the ‘empirical’ school tend to regard cosmology as a testing ground for their extrapolations and as a legitimate playground for the geometers, so some

> adherents of the deductive approach appear to regard cosmology as a purely logical subject. (Bondi 1960, p. 7)

In this latter case, the deductive extremists, in their mathematical zeal, seem to forget that cosmology, after all, **should** have some relation to observation: “To them all that is of interest in a theory is its logical character, not its relevance to the interpretation of observational data.” Obviously, this danger must be avoided: according to Bondi, deductivism can be a ***scientific*** approach in cosmology only if its postulates (or axioms) are candidates for disproof.

8.2 Enter Popper

Clearly, with this reference to the connection between science and disproof, Bondi has added a distinctly Poppererian element to the deductivist methodology, one which had not previously appeared in the works of any of the earlier members of the hypothetico-deductive school. According to Popper's philosophy of science, a theory can legitimately be called “scientific” only if that theory makes a prediction that, in principle, can be shown to be false, or *falsified*, to use Popper's own term. Thus astrology, for example, fails to be a scientific theory because it cannot be falsified: although astrology seems to make predictions, these statements about the future are so vague, so general and abstract, that they cannot be tied down to definite claims about observations to be made at a definite time and place. Hence there is no explicit observation to be made in falsification. Astronomy, in comparison, makes explicit, specific predictions about what will occur in the sky on such-and-such a date, in such-and-such a place. If the prediction fails, then we know that the element of astronomical theory which made the prediction is deficient, maybe even false.

Cosmology is a borderline case: since observations of cosmological significance are so rare and hard-won—Hubble's observation of the red shift was one of the first solid ones—it is very difficult, not to mention brave, to tie one's cosmological theory to Popper's falsificationist principle as a guarantee of scientific acceptability. But this is exactly what Bondi did.

Much later Bondi was to make explicit his debt to Popper:

> I think the person from whom we had most help on the philosophical side was Popper. His analysis of science encouraged one to be imaginative, and encouraged one to go for something that was very rigid and therefore empirically disprovable. (Bondi 1990, p. 194)

8.3 But It's Milne In the End

Yet Bondi's major philosophical debt was to Milne. According to Bondi, Milne's theory was through and through deductive, which was reason enough for some of his colleagues to condemn it:

> The aim of this discipline [= kinematic relativity] is to deduce as much as possible merely from the cosmological principle and the basic properties of space, time and the propagation

> of light. The beauty of this, as indeed of any deductive theory, rests on the rigour of the arguments and the small number of the axioms required...When the theory was first developed it met with great hostility and was criticized very severely, often unjustly, and sometimes frivolously. (Bondi 1960), p. 123)

In addition to his admiration of Milne's H-D methodology, Bondi has high praise as well for Milne's operationalism, particularly its use in defining distance:

> Imperfect as Milne's definition of distance may be, it is very much better than the ‘rigid ruler’ one used in most other theories…Milne's definition of distance, by no means perfect as it is, is probably the best yet devised. (Bondi 1960, p. 126-9)

In the end, Bondi sums up Milne's contributions with no uncertain praise:

> The foregoing brief description will have indicated the remarkable success of kinematic relativity in attempting to use the cosmological principle not only for the construction of the substratum but as chief guide in formulating ordinary physics. In this respect it differs greatly from all other cosmologies which either rely on a conventionally obtained body of physics or have not yet succeeded in drawing conclusions of local interest from the cosmological principle. (Bondi 1960, p. 136)

8.4 Return of the Cosmological Principle

Here Bondi speaks of Milne's cosmological principle. According to Milne's principle, every observer in the universe should get the same world picture, that is, should make precisely the same observations of the universe at the same moment as any other observer. (Milne 1934b) Uniformity over spatial slices is guaranted by Milne's invoking of the principle. Yet Milne's universe evolves, it changes its form over time. Hence it has no temporal uniformity. Bondi felt that this raised the possibility that physics itself might change over time. Because of this risk, Bondi generalized Milne's cosmological principle into what he called the *perfect cosmological principle* [=PCP]. According to this principle, all observers at all places and **at all times** will look out upon the same unchanging, unevolving, universe. Such a universe is a universe in a steady-state—hence the name.

Clearly, PCP is a daring, indeed heroic, interpretation of a methodological necessity. Forty years after the fact, Bondi described the “philosophical attitude” which underlay his “implausible” PCP:

> But the essential point of the philosophy was and is that if the universe was evolving and changing, then there is no reason to trust what we call the laws of physics, established by experiments performed here and now, to have permanent validity. (Bondi 1990, p. 192)

Hence, or so Bondi's argument goes, since there is reason **not** to trust the laws of physics if the universe is evolving, let us presume that the universe is **not** evolving and changing; that is, let us presume PCP.

Although the principle (and the theory which results from it—steady state cosmology) is, as McVittie remarked, "much more restrictive than general relativity"; (McVittie 1990, p. 45) it is this very restrictiveness which satisfies Bondi's Popperian wishes:

> For the correct argument has always been that the steady state model was the one that could be disproved most easily by observation, Therefore, it should take precedence over other less disprovable ones until it has been disproved. (Bondi and Kilmister 1959, p. 55-6)

In another place, Bondi makes a similar point: "Comparison with observation becomes then possible and renders the PCP liable to observational disproof. This possibility of a clear-cut disproof establishes the scientific status of PCP." (Bondi 1957, p. 198) Comments such as this make clear Bondi's committment to a Popperian addition to the basic deductive methodology he inherited from Milne.

8.5 A Popperian Conclusion

In the end, the philosophical purity of Bondi's steady state theory served him, and cosmology, well. Of course, the usual suspect, Dingle, and others of his ilk, such as McVittie in particular, were outraged, and loudly, at Bondi's extension of Milne's methods. A passage from Dingle's R.A.S. Presidential Address suffices to show the tenor of the debate's declining days:

> Even idle speculation may not be quite valueless if it is recognized for what it is. If the new cosmologists would observe this proviso, calling a spade a spade and not a perfect agricultural principle, one's only cause for regret would be that such great talents were spent for so little profit. (Dingle 1953, p. 404)

But PCP and the theory which it engendered were exactly as described: eminently falsifiable. No matter the extent of Dingle *et al*'s disdain, Steady State theory stayed right out in front, ready for whatever empirical observations might be slung at it. As Bondi said "Show me some fossils from an evolving universe, and I'll give it up." In 1965, the fossils arrived, courtesy of the observations of the 3° K remnant microwave radiation.

And Bondi, true to his philosophy, gave it up.

Bibliography

- A.D.R. [initials of the semi-anonymous reviewer], 1938, "A Philosophy of Science," *Nature (London)* 141: 95-96.
- Bondi, H., 1957, "Some Philosophical Problems in Cosmology", In British *Philosophy in the Mid-Century*. Edited by C. A. Mace. London: George Allen and Unwin.
- Bondi, H., 1960, *Cosmology*, 2 ed. Cambridge: Cambridge University Press.
- Bondi, H., 1990, "The Cosmological Scene 1945-1952", In *Modern Cosmology in Retrospect*.

Edited by B. Bertotti, *et al*. Cambridge: Cambridge University Press.

- Bondi, H. and C. W. Kilmister, 1959, “The Impact of *Logik der Forschung*,” Brit.J.Phil.Sci. 10: 55-57.
- Crowther, J. G. 1970, *Fifty Years with Science*, London: Barrie & Jenkins.
- deSitter, W., 1934, “On the Foundations of the theory of Relativity, with Special reference to the Theory of the Expanding Universe,” *Proceedings,Royal Academy Amsterdam* 37: 597-601.
- Dingle, H. 1931, *The Evolution of the Universe*, London: Nature.
- Dingle, H., 1933, “On E.A. Milne's theory of world structure and the expansion of the Universe,” *Z.Astrophysik* 6: 173-179.
- Dingle, H., 1937, “Modern Aristotelianism,” *Nature (London)* 139: 784-786.
- Dingle, H., 1938, “Science and the Unobservable,” *Nature (London)* 141: 21-28.
- Dingle, H., 1953, “The President's Address,” *Mon.Not.R .astron.Soc*.: 113: 393-407.
- Eddington, A. S., 1932, *The Expanding Universe*, Ann Arbor: Ann Arbor Paperbacks-U.Mich Press.
- Eddington, A. S., 1935, “*Review of Relativity Gravitation and World-Structure*,” *Nature (London)* 135: 635-636.
- Eddington, A. S. 1939, *The Philosophy of Physical Science*, New York: Macmillan ..
- Kermack, W. O. and W. H. McCrea, 1933, “On Milne's Theory of World Structure,” *Mon.Not.R.astron.Soc*. 93: 519-529.
- McCrea, W. H., 1935, “Observable Relations in relativistic Cosmology,” *Z.Astrophysik* 9: 290-314.
- McCrea, W. H., 1939, “The Evolution of Theories of Space-Time and Mechanics,” *Phil.Sci*. 6: 137-162.
- McVittie, G. C., 1931, “The Problem of *n* Bodies and the Expansion of the Universe,” *Mon.Not.R.astron.Soc*. 91: 274-283.
- McVittie, G. C., 1933a, “The Mass-particle in an Expanding Universe,” *Mon.Not.R.astron.Soc*. 93: 325-339.
- McVittie, G. C., 1933b, “Milne's Theory of the Expansion of the Universe,” *Nature (London)* 131: 533-534.
- McVittie, G. C., 1934, “The Spiral Nebulae and the Expansion of the Universe,” *Phys.Soc.(Lond.)Reports* 1: 24-29.
- McVittie, G.C., 1939, “Observation and Theory in Cosmology,” *Proc.Phys.Soc.Lond* 51: 529-537.
- McVittie, G. C. 1990, Interview, 21 Mar 78, In *Interviews with Astrophysicists*, Edited by American Institute of Physics, New York: American Institute of Physics.
- Milne, E. A., 1929, *The Aims of Mathematical Physics*, Oxford: Oxford University Press.
- Milne, E. A., 1933, “World-structure and the Expansion of the Universe,” *Z.Astrophysik* 6: 1-35.
- Milne, E. A., 1934a, “Some Points in the Philosophy of Physics: Time, Evolution and Creation,” *Philos*. 9: 19-38.
- Milne, E. A., 1934b, “World-models and the World-picture,” *Observatory*, 57: 24-27.
- Milne, E. A. 1935, *Relativity Gravitation and World-Structure*. Oxford: Clarendon Press.
- Milne, E. A., 1937, “Kinematics, Dynamics, and the Scale of Time,” *Proc.Roy.Soc.(Lond.)* A158: 324-329.
- Milne, E. A., 1943, “The Fundamental Concepts of Natural Philosophy,” *ProcRoySoc(Edin)* 63:

10-24.
- Milne, E. A., 1932-37, Correspondence with Geoffrey Milne. This correspondence is in Modern Manuscripts, Bodelian Library, Oxford, and is used with permission. .
- Robertson, H. P., 1933a, "On E.A. Milne's Theory of World Structure," *Z.Astrophysik* 7: 152-162.
- Robertson, H. P., 1933b, "Relativistic Cosmology," *Rev.Mod.Phys*. 5: 62-90.
- Robertson, H. P., 1935, "Kinematics and World-structure," *Ap.J*. 82: 284-301.
- Robertson, H. P., 1936, "Review of Milne's *Relativity Gravitation and World-Structure*," *Ap.J*. 83: 61-66.
- Temple, G., 1939, "Relativistic Cosmology," *PhysSoc.(London)*, 51: 465-478.
- Walker, A. G., 1934, "The Principle of Least Action in Milne's Kinematical Relativity," *Proc.Roy.Soc.(Lond.)* 147A: 478-490.
- Walker, A. G., 1935, "On the formal comparison of Milne's kinematical system with the systems of general relativity," *Mon.Not.R.Astr.Soc*. 95: 263-269.
- Whittaker, E. T., 1935, "Review of *Relativity Gravitation and World-Structure*," *Observ*. 58: 179-188.
- Whittaker, E. T., 1941, "Some Disputed Questions in the Philosophy of the Physical Sciences," *ProcRoySoc(Edin)* 61: 160-175.

Other Internet Resources

At this time there are few if any internet sites devoted to the history and philosophy of modern cosmology. What is typically available is information relating to cosmology within the history of astronomy, or presentations about contemporary cosmology. The URLs given below are the best sources for history of cosmology; there are no known sites relating to the philosophy of modern cosmology.

- The Shapley-Curtis Debate in 1920 (maintained by Robert J. Nemiroff, Michigan Technological University and NASA Goddard)
- History of Astronomy (maintained by Prof. Dr. Wolfgang Dick, Astronomische Institute, Universität Bonn)
- Cosmology Books and Links (compiled by Joseph S. Tenn, Physics and Astronomy, Sonoma State University)
- Cosmology Since 1900 (by Joseph S. Tenn, Physics and Astronomy, Sonoma State University)

Related Entries

epistemology | general relativity: early philosophical interpretations of | induction: new problem of | induction: problem of | metaphysics | Popper, Karl | rationalism vs. empiricism | science, philosophy of | scientific method | space and time: inertial frames

George D. Gale
University of Missouri/Kansas City
galeg@umkc.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 17, 2002
Content last modified: May 17, 2002

Early Philosophical Interpretations of General Relativity

Each of the following philosophical interpretations of general relativity selected certain aspects of that theory for favored recognition. While followers of Mach lauded Einstein's attempt to implement a "relativization of inertia" in the general theory, they were much more comfortable with Einstein's operationalist treatment of concepts in the special theory. Kantians and neo-Kantians, if freed from strict fealty to the doctrine of the Transcendental Aesthetic, pointed to the surpassing importance of certain synthetic "intellectual forms" in the general theory, such as the principle of general covariance. For logical empiricism, the philosophical significance of relativity theory was largely methodological, that conventions must first be laid down in order to express the empirical content of a physical theory. Finally, within a few years of its completion in 1915, attempts were made to extend general relativity's "geometrization" of gravitation to non-gravitational fields. The first of these, by Weyl, and shortly thereafter by Eddington, may be distinguished from others, in particular the many attempts of Einstein, in that they aimed not at a unified field theory, in the sense of a completely geometrical field theory of all fundamental interactions, but at reconstructing general relativity from the epistemological perspectives of transcendental idealism.

1. The Search for Philosophical Novelty

Extraordinary public clamor greeted an announcement of the joint meeting of the Royal Society of London and the Royal Astronomical Society on the 6th of November, 1919. To within acceptable margin of error, astronomical observations during the solar eclipse the previous May 29th revealed the displacement of starlight passing near the surface of the sun predicted by Einstein's gravitational theory

of curved spacetime. By dint of having "overthrown" such a permanent fixture of the cognitive landscape as Newtonian gravitational theory, the general theory of relativity at once became a principal focus of philosophical interest and inquiry. Although some physicists and philosophers initially opposed it, mostly on non-physical grounds, surveyed here are the principal philosophical interpretations of the theory accepting it as a definite advance in physical knowledge. Even so, these include positions ill-informed as to the mathematics and physics of the theory. Further lack of clarity stemmed from the scientific *literati* who provided differing, and at times, conflicting mathematical or physical accounts of the theory's fundamental principles. These are: the principles of equivalence, of general relativity, of general covariance, and finally what Einstein termed "Mach's Principle" of the complete relativization of inertia. In one or another form, all of these controversies have continued into the present literature of physics and philosophy of physics. (See e.g., Ohanian (1977); Norton (1993); Friedman (1983); Barbour and Pfister (1995).) This is not unusual: physical theories, if sufficiently robust, are rarely, if ever, without unproblematic aspects, often taken to say different things at different stages of development. But the very fluidity of physical and mathematical meaning lent interpretative latitude for inherently antagonistic philosophical viewpoints seeking vindication, confirmation or illumination by the revolutionary new theory. Perhaps only semi-facetiously, Russell (1926, 331) observed that

> There has been a tendency, not uncommon in the case of a new scientific theory, for every philosopher to interpret the work of Einstein in accordance with his own metaphysical system, and to suggest that the outcome is a great accession of strength to the views which the philosopher in question previously held. This cannot be true in all cases; and it may be hoped that it is true in none. It would be disappointing if so fundamental a change as Einstein has introduced involved no philosophical novelty.

It cannot be denied that general relativity proved a considerable stimulus to "philosophical novelty". But then the question as to whether it particularly supported any one line of philosophical interpretation over another also must take into account the fact that schools of interpretation in turn "evolved" to accommodate what were regarded as its philosophically salient features. A classic instance of this is the assertion, to become a cornerstone of logical empiricism, that relativity theory had shown the untenability of any "philosophy of the synthetic *a priori"*, despite the fact that early works on relativity theory by both Reichenbach and Carnap were written from within that broad perspective. It will be seen that, however ideologically useful, this claim by no means "follows" from relativity theory although, as physicist Max von Laue noted in his early text on general relativity (1921, 42), "not every sentence of *The Critique of Pure Reason"* might still be held intact. What does "follow" from scrutiny of the various philosophical appropriations of general relativity is rather a consummate illustration that, due to the evolution and mutual interplay of physical, mathematical and philosophical understandings of a revolutionary physical theory, significant "philosophical interpretations" often are works in progress, extending over many years.

2. Machian Positivism

2.1 In the Early Einstein

In 1912, Einstein's name, together with those of the Göttingen mathematicians David Hilbert and Felix Klein, was prominently displayed (in the *Naturwissenschaftliche Rundschau* 27, 336) among those joining Mach's in a call for the formation of a "Society for Positivist Philosophy". Citing the pressing need of science "but also of our age in general" for a "comprehensive world view based on the material facts accumulated in the individual sciences", the appeal appears above all to have been an orchestrated attempt to buttress Mach's positivist conception of science in the face of recent realist criticisms of Mach by Max Planck, then Germany's leading theoretical physicist. More a declaration of allegiance than an act of scholarly neutrality, it provides but further evidence of Einstein's youthful enthusiasm for Mach's writings. Late in life (1949a, 21), Einstein wrote of the "profound influence" that Mach's *Science of Mechanics* (1883) exercised upon him as a student as well as of the "very great influence" in his "younger years" of "Mach's epistemological position". Indeed, in first decade or so of relativity theory, these influences are highly visible. Already in the special theory of relativity (1905), Einstein's operational definition of the "simultaneity" of distantly separated events, whereby clocks are synchronized by sending and receiving light signals, is closely modeled on the operational definition of "mass" in Mach's *Mechanics.* Moreover, occasional epistemological and methodological pronouncements indicated a broad consensus with core parts of Mach's epistemology of science, e.g., "The concept does not exist for the physicist until he has the possibility of discovering whether or not it is fulfilled in an actual case" (1917a/1955, 22). Thus relativity theory was widely viewed as fully compliant with Mach's characterization of theoretical concepts as merely economical shorthand for concrete observations or operations.

2.2 A "Relativization of Inertia"?

Machian influences specific to the general theory of relativity appeared even more extensive. In papers leading up to the definitive presentation of the general theory of relativity in 1916, Einstein made no secret of the fact that Mach had been the inspiration for his epistemologically mandated generalization of the principle of relativity. Holding, with Mach, that no observable facts could be associated with the notions of "absolute acceleration" or "absolute inertia" (i.e., resistance to acceleration), the generalization mandated that the laws of nature be completely independent of the state of motion of any chosen reference system. On Mach's death, Einstein wrote, in a warm obituary, of how close Mach himself had been, years before, to demanding a general theory of relativity, quoting extensively from the famous passages in the latter parts of the *Mechanics* critical of Newton's "absolute" concepts of space, time and motion (1916b, 102-3). With this reference in mind, the physicist Phillip Frank, later to be associated with the Vienna Circle, observed (1917/1949, 68) that "it is universally known today that Einstein's general theory of relativity grew immediately out of the positivistic doctrine of space and motion". In fact, there are both genuine and spurious Machian motivations connected with Einstein's principle of general relativity, a mixture complicated by Einstein's own puzzling remarks regarding the principle of general covariance.

2.3 Positivism and the "Hole Argument"

A passage from §3 of Einstein's first complete exposition of the general theory of relativity (1916a)

appeared to provide further grist for the mill of Machian positivism. There Einstein grandly declared that his requirement of general covariance for the gravitational field equations (i.e., that they remain unchanged under arbitrary, but suitably continuous, transformation of the spacetime coordinates), "takes away from space and time the last remnant of physical objectivity". An accompanying heuristic "reflection" on the reasoning behind this claim seemed nothing less than an endorsement of Mach's phenomenalism. "All our space-time verifications", Einstein wrote, "invariably amount to a determination of space-time coincidences....". This is because, Einstein presumed, all results of physical measurement ultimately amount to verifications of such coincidences, such as the observation of the coincidence of the second hand of a clock with a mark on its dial. Observing that such (topological) relations alone are preserved under arbitrary coordinate transformation, Einstein concluded that "all our physical experience can ultimately be reduced to such coincidences". To Mach's followers, Einstein's illustrative reflection was nothing less than an explicit avowal of the centerpiece of Mach's phenomenalist epistemology, that sensations *(Empfindungen)*, directly experienced sensory perceptions, alone are real and knowable. Thus Josef Petzoldt, a Machian philosopher and editor of the 8th edition of Mach's *Mechanics* , the first to appear after the general theory of relativity, noted that Einstein's remarks meant that the theory "rests, in the end, on the perception of the coincidence of sensations" and so "is fully in accord with Mach's world-view, which is best characterized as relativistic positivism" (1921, 516).

However, contemporary scholarship has shown that Einstein's remarks here were but elliptical references to an argument (the so-called "Hole Argument") that has only fully been reconstructed from his private correspondence. Its conclusion is that, if a theory is generally covariant, the points of the spacetime manifold can have no inherent primitive individuality (inherited say, from the underlying topology), and so no reality independent of physical fields (Stachel (1980); Norton (1984), (1993)). Thus for a generally covariant theory, no physical reality accrues to "empty space" (or "spacetime") in the absence of physical fields. This means that the spacetime coordinates are nothing more than arbitrary labels for the identification of physical events, or, with rhetorical embellishment, that space and time have lost "the last remnant of physical objectivity". Hence this passage was not an endorsement of positivist phenomenalism.

2.4 "Mach's Principle"

To be sure, for a number of years Einstein expressed the ambition of the general theory of relativity to fully implement Mach's program for the relativization of all inertial effects, even appending the so-called "cosmological constant" to his field equations (1917b) for this purpose. This real point of contact of Mach's influence was clearly identified only in 1918, when Einstein distinguished what he baptized as "Mach's Principle" (roughly, that inertial effects stem from an interaction of bodies) from the principle of general relativity which he now interpreted as the principle of general covariance. Taken together with the principle of the equivalence, Einstein asserted that the three principles, were three "points of view" on which his theory rested, even if they could not be thought completely independent of one another. Despite Einstein's intent, there is considerable disagreement about the extent to which, if at all, general relativity conforms to "Mach's Principle". In part this is due to vagaries regarding what the Principle actually asserts and then again, to difficulties in comprehending what physical mechanism might implement the Principle, however interpreted. How, for instance, could a body's inertial mass be accounted due to the

influence of all other bodies in the universe? (See the discussions in Barbour and Pfister (1995)).

2.5 An Emerging Anti-Positivism

As Einstein's principal research activity turned, after 1919, to the pursuit of a geometrical "unified theory of fields", his philosophical pronouncements increasingly took on a realist or at least anti-positivist coloration. Already in (1922, 28) lecturing at the Sorbonne, Einstein pronounced Mach "*un bon mécanicien*" (no doubt in reference to Mach's views of the relativity of inertia) but "*un déplorable philosophe*". Increasingly, Einstein's retrospective portrayals of the genesis of general relativity centered almost entirely on considerations of mathematical aesthetics (see Norton (2000) and §5). On the other hand, positivists and operationalists alike adopted the Einstein analysis of simultaneity as relativity theory's fundamental methodological feature. One, ruefully noting the difficulty of giving an operationalist analysis of the general theory, even suggested that the requirement of general covariance "conceals the possibility of disaster" (Bridgman (1949), 354). Finally there was, for Einstein, an understandable awkwardness in learning of Mach's surprising disavowal of any role as forerunner to relativity theory in the "Preface", dated 1913, to his posthumous book (1921) on physical optics, published by Mach's son Ludwig. Though Einstein died without knowing differently, a recent investigation has built a strong case that this statement was forged after Mach's death by his son Ludwig, under the influence of a rival guardian of Mach's legacy and opponent of relativity theory, the philosopher Hugo Dingler (Wolters, 1987).

3. Kantian and Neo-Kantian Interpretations

3.1 Neo-Kantians on Special Relativity

In the universities of Imperial and early Weimar Germany, the philosophy of Kant, particularly the various neo-Kantian schools, held pride of place. Of these, the "Marburg School" of Hermann Cohen and Paul Natorp, later Ernst Cassirer, exhibited a special interest in the philosophy of the physical sciences and of mathematics. But prior to the general theory of relativity (1915-1916), Kantian philosophers accorded relativity theory only cursory attention. This may be seen in two leading Marburg works appearing in 1910, Cassirer's *Substanzbegriff und Funktionsbegriff*. and Natorp's *Die Logischen Grundlagen der Exakten Wissenschaften*. Both conform to the characteristic Marburg modification of Kant that greatly extended the scope of "transcendental logic", bringing under "pure thought" or "intellectual forms" what Kant had sharply distinguished as "pure intuition" and a conceptual faculty of understanding. Of course, this revisionist tendency greatly transformed the meaning of Kant's Transcendental Aesthetic and with it Kant's conviction that space and time were "forms of sensibility" or "pure intuitions *a priori"* and so as well, his accounts of arithmetic and geometry. As will be seen, it enabled Cassirer, some ten years later, to view even the general theory of relativity as a striking confirmation of the fundamental tenets of transcendental idealism. In 1910, however, Cassirer's brief but diffuse discussion of "the problem of relativity" mentions neither the principle of relativity nor the light postulate nor the names of Einstein, Lorentz or Minkowski. Rather it centers on the question of whether

space and time are aggregates of sense impressions or "independent intellectual *(gedankliche)* forms". Having decided in favor of the latter, Cassirer goes on to argue how and why these ideal mathematical presuppositions are necessarily related to measurable, empirical notions of space, time, and motion (1910, 228-9; 1923, 172-3).

Natorp's treatment, though scarcely six pages is far more detailed (1910, 399-404). In Marburg revisionist fashion, the "Minkowski *(sic)* principle of relativity" was welcomed as a more consistent (as avoiding "Newtonian absolutism") carrying through of the distinction between transcendentally ideal and purely mathematical *concepts* of space and time and the relative physical measures of space and time. The relativization of time measurements, in particular, showed that Kant, shorn of the psychologistic error of "pure intuition", had correctly maintained that time is not an object of perception. Natorp further alleged that from this relativization it followed that events are ordered, not in relation to an absolute time, but as lawfully determined phenomena in mutual temporal relation to one another. This is close to a Leibnizian relationism about time. Similarly, the light postulate had a two-fold significance within the Marburg conception of natural science. On the one hand, the uniformity of the velocity of light, deemed an *empirical* presupposition of all space- and time-measurements, reminded that absolute determinations of these measures, unattainable in empirical natural science, would require a correspondingly absolute bound. Then again, as an upper limiting velocity for physical processes, including gravitational force, the light postulate eliminated the "mysterious absolutism" of Newtonian action-at-a-distance. Natorp regarded the requirement of invariance of laws of nature with respect to the Lorentz transformations as "perhaps the most important result of Minkowski's investigation". However, little is said about this, and in fact there is some confusion regarding these transformations and the Galilean ones they supercede (the former are seen as a "broadening *(Erweiterung)* of the old supposition of the invariance of Newtonian mechanics for a translatory or *circular* (*zirkuläre*, emphasis added) motion of the world coordinates"(403)). He concluded with an observation that the appearance of non-Euclidean and multi-dimensional geometries in physics and mathematics are to be understood only as "valuable tools in the treatment of special problems". In themselves, they furnish no new insight into the (transcendental) logical meaning and ground of the purely mathematically determined concepts of space and time; still less do they require the abandonment of these concepts.

3.2 Immunizing Strategies

Following the experimental confirmation of the general theory in 1919, few Kantians attempted to retain, unadulterated, all of the components of Kant's epistemological views. Several examples will suffice to indicate characteristic "immunizing" strategies (see Hentschel (1990). The *Habilitationsschrift* of E. Sellien (1919), read by Einstein in view of his criticism expressed in an October, 1919 letter to Moritz Schlick (Howard (1984),625), declared that Kant's views on space and time pertained solely to "intuitive space" and so were not touched by the measurable spaces and times of Einstein's empirical theory. The work of another young Kantian philosopher, Ilse Schneider, personally known to Einstein, affirmed that Kant merely had held that the space of three-dimensional Euclidean geometry is the space in which Newton's gravitational law is valid, whereas an analogous situation obtains in general relativity. Furthermore, Einstein's cosmology (1917b) of a finite but unbounded universe could be seen as in complete accord with the "transcendental solution" to the First Antinomy in the Second Book of the

Transcendental Dialectic. Her verdict was that the apparent contradictions between relativity theory and Kantian philosophy disappear on closer examination of both doctrines (1921, 71-75).

3.3 Rejecting or Refurbishing the Transcendental Aesthetic

But most Kantian philosophers did not attempt to immunize Kant from an apparent empirical refutation by the general theory. Rather, their concern was to establish how far-reaching the necessary modifications of Kant must be and whether, on implementation, anything distinctively Kantian remained. Certainly, most at risk appeared to be the claim, in the Transcendental Aesthetic, that all objects of "outer" intuition, and so all physical objects, conform to the space of Euclidean geometry. Since the general theory of relativity employed non-Euclidean (Riemannian) geometry for the characterization of physical phenomena, the conclusion seemed inevitable that any assertion of the necessarily Euclidean character of physical space in finite, if not "infinitesimal", regions, is simply false.

Winternitz (1924) is an example of this tendency that may be singled out on the grounds that it was deemed significant enough to be the subject of a rare book review by Einstein (1924) . Winternitz argued that the Transcendental Aesthetic is inextricably connected to the claim of the necessarily Euclidean character of physical space and so stood in direct conflict with Einstein's theory. It must accordingly be totally jettisoned as a confusing and unnecessary appendage of the fundamental transcendental project of establishing the *a priori* logical presuppositions of physical knowledge. Indeed, these presuppositions have been confirmed by the general theory: They are spatiality and temporality as "unintuitive schema of order" in general (as distinct from any particular chronometrical relations), the law of causality and presupposition of continuity, the principle of sufficient reason, and the conservation laws. Remarkably, the *necessity* of each of these principles was, rightly or wrongly, soon to be challenged by the new quantum mechanics. (For a challenge to the law of conservation of energy, see Bohr, Kramers, and Slater (1924)). According to Winternitz, the *ne plus ultra* of transcendental idealism lay in the claim that the world "is not given but posed *(nicht gegeben, sondern aufgegeben)* (as a problem)" out of the given material of sensation. Interestingly, Einstein, late in life, returns to this formulation as comprising the fundamental Kantian insight into the character of physical knowledge (1949b, 680).

However, a number of neo-Kantian positions, of which that of Marburg was only the best known, did not take the core doctrine of the Transcendental Aesthetic, that space and time are *a priori* intuitions, *à la lettre*. Rather, resources broadly within it were sought for preserving an updated "critical idealism". In this regard, Bollert (1921) merits mention for its technically adroit presentation of both the special and the general theory. Bollert argued that relativity theory had "clarified" the Kantian position in the Transcendental Aesthetic by demonstrating that not space and time, but spatiality (determinateness in positional ordering) and temporality (in order of succession) are *a priori* conditions of physical knowledge. In so doing, general relativity theory with its variably curved spacetime, brought a further advance in the steps or levels of "objectivation" lying at the basis of physics. In this process, corresponding with the growth of physical knowledge since Galileo, each higher level is obtained from the previous through elimination of subjective elements from the concept of physical object. This ever-augmented and revised advance of conditions of objectivity is alone central to critical idealism. For this reason, it is "an error" to believe that "a contradiction exists between Kantian *a priorism* and relativity

theory" (1921,64). As will be seen, these conclusions are quite close to those of the much more widely known monograph of Cassirer (1921). It is worth noting that Bollert's interpretation of critical idealism was cited favorably by Gödel (1946/9-B2, 240, n.24) much later during the course of research which led to his famous discovery of rotating universe solutions to Einstein's gravitational field equations (1949). This investigation had been prompted by Gödel's curiosity concerning the similar denials, in relativity theory and in Kant, of an absolute time.

3.4 General Covariance: A Synthetic Principle of "Unity of Determination"

The most influential of all neo-Kantian interpretations of general relativity was Ernst Cassirer's *Zur Einsteinschen Relativitätstheorie* (1921). Cassirer regarded the theory as a crucial test for *Erkenntniskritik*, the preferred term for the epistemology of Marburg's transcendental idealism. The question, posed right at the beginning, is whether the Transcendental Aesthetic offered a foundation "broad enough and strong enough" to bear the general theory of relativity. Recognizing the theory's principal epistemological significance to lie in the requirement of general covariance ("that the general laws of nature are not changed in form by arbitrary changes of the space-time variables"), Cassirer directed his attention to Einstein's remarks, cited in §2 above, that general covariance "takes away from space and time the last remnant of physical objectivity". Cassirer correctly construed the gist of this passage to mean that in the general theory of relativity, space and time coordinates have no further importance than to be mere labels of events ("coincidences"), independent variables of the mathematical (field) functions characterizing physical state magnitudes. Furthermore,in accord with central tenets of the Marburg Kant interpretation noted above, Cassirer maintained that the requirement of generally covariant laws was a vindication of the transcendental ideality of space and time, not, indeed, as "forms of intuition" but as "objectifying conditions" that further "de-anthropomorphized" the concept of object in physics, rendering it "purely symbolic". In this regard, the requirement of general covariance had significantly improved upon Kant in bringing out far more clearly the exclusively methodological role of these conditions in empirical cognition, a role Kant misleadingly assigned to "pure intuition". Not only has it has shown that space and time are not "things", it has also clarified that they are "ideal principles of order" applying to the objects of the physical world as a necessary condition of their possible experience. According to Cassirer, Kant's *intention* with regard to "pure intuition" was simply to express the methodological presupposition that certain "intellectual forms" *(Denkformen),* among which are the purely ideal *concepts* of coexistence and succession, enter into all physical knowledge. According to the development of physics since the 17th century chronicled in *Substanzbegriff und Functionsbegrif*, these forms have progressively lost their "fortuitous" *(zufälligen)* anthropomorphic features, and more and more take on the character of "systematic forms of unity". From this vantage point, general covariance is but the most recent refinement of the methodological principle of "unity of determination" governing the constitution of objects of physical knowledge, completing the transposition in physics from concepts of substance into functional and relational concepts. In its wake, the fundamental concept of object in physics no longer pertains to particular entities or processes in space and time but rather to "the invariance of relations among (physical state) magnitudes". For this reason, Cassirer concluded, the general theory of relativity exhibits "the most determinate application and carrying through within

empirical science of the standpoint of critical idealism" (1921/1957, 71; 1923, 412).

4. Logical Empiricism

4.1 Lessons of Methodology?

Logical empiricism's philosophy of science was conceived under the guiding star of Einstein's two theories of relativity, as may be seen from the early writings of its founders, for purposes here, Moritz Schlick, Rudolf Carnap, and Hans Reichenbach. The small monograph of Schlick, *Space and Time in Contemporary Physics*, appearing in 1917, initially in successive issues of the scientific weekly *Die Naturwissenschaften*, served as a prototype. Among the first of a host of "philosophical examinations" of the general theory of relativity, it was distinguished both by the comprehensibility of its largely non-technical physical exposition and by Einstein's enthusiastic praise of its philosophical appraisal, favoring Poincaré's conventionalism over both neo-Kantianism and Machian positivism. The transformation of the concept of space by the general theory of relativity was the subject of Rudolf Carnap's dissertation at Jena in 1921. Appearing as a monograph in 1922, it also evinced a broadly conventionalist methodology combined with elements of Husserlian transcendental phenomenology. Distinguishing clearly between intuitive, physical and purely formal conceptions of space, Carnap argues that, subject to the necessary constraints of certain *a priori* phenomenological conditions of the topology of intuitive space, the purely formal and the physical aspects of theories of space, can be adjusted to one another so as to preserve any conventionally chosen aspect. In turn, Hans Reichenbach was one of five intrepid attendees of Einstein's first seminar on general relativity given at Berlin University in the tumultuous winter of 1918-1919; his detailed notebooks survive. The general theory of relativity was the particular subject of Reichenbach's neo-Kantian first book (1920), which is dedicated to Albert Einstein, as well as of his next two books (1924), (1928), and of numerous papers in the 1920s.

But Einstein's theories of relativity provided far more than the subject matter for these philosophical examinations; rather logical empiricist philosophy of science was itself fashioned by lessons allegedly drawn from relativity theory in correcting or rebutting neo-Kantian and Machian perspectives on general methodological and epistemological questions of science. Several of the most characteristic doctrines of logical empiricist philosophy of science — the interpretation of *a priori* elements in physical theories as conventions, the treatment of the role of conventions in linking theory to observation and in theory choice, the insistence on verificationist definitions of theoretical terms — were taken to have been conclusively demonstrated by Einstein in fashioning his two theories of relativity. In particular, Einstein's 1905 analysis of the conventionality of simultaneity in the special theory of relativity became something of a methodological paradigm, prompting Reichenbach's own method of "logical analysis" of physical theories into "subjective" (definitional, conventional) and "objective" (empirical) components. The overriding concern in the logical empiricist treatment of relativity theory was to draw broad lessons from relativity theory for scientific methodology and philosophy of science generally, although issues more specific to the philosophy of physics were also addressed. Only the former are considered here; for a discussion of the latter, we may refer to Ryckman (forthcoming b).

4.2 From the "Relativized A priori to the "Relativity of Geometry"

A cornerstone of Reichenbach's "logical analysis" of the theory of general relativity is the thesis of "the relativity of geometry", that an arbitrary geometry may be ascribed to spacetime (holding constant the underlying topology) if the laws of physics are correspondingly modified through the introduction of "universal forces". This particular argument for metric conventionalism has generated substantial controversy on its own, but is better understood through an account of its genesis in Reichenbach's early neo-Kantianism. Independently of that genesis, the thesis becomes the paradigmatic illustration of Reichenabch's broad methodological claim that conventional or definitional elements, in the form of "coordinative definitions" associating mathematical concepts with "elements of physical reality", are a necessary condition of empirical cognition in science. At the same time, however, Reichenbach's thesis of metrical conventionalism is part and parcel of an audacious program of epistemological reductionism regarding spacetime structures. This was first attempted in his "constructive axiomatization" (1924) of the theory of relativity on the basis of "elementary matters of fact"*(Elementartatbestande)* regarding the observable behavior of lights rays, and rods and clocks. Here, and in the more widely read treatment(1928), metrical properties of spacetime are deemed less fundamental than topological ones, while the latter are derived from the concept of time order. But time order in turn is reduced to that of causal order and so the whole edifice of structures of spacetime is considered epistemologically derivative, resting upon ultimately basic empirical facts about causal order and a prohibition against action-at-a-distance. The end point of Reichenbach's epistemological analysis of the foundations of spacetime theory is then "the causal theory of time", a type of relational theory of time that assumes the validity of the causal principle of action-by-contact *(Nahwirkungsprinzip).*

However, Reichenbach's first monograph on relativity (1920) was written from within a neo-Kantian perspective. As Friedman (1999) and others have discussed in detail (Ryckman, forthcoming a), Reichenbach's innovation, a modification of the Kantian conception of *synthetic a priori* principles, rejecting the sense of "valid for all time" while retaining that of "constitutive of the object (of knowledge)", led to the conception of a theory-specific "relativised *a priori*". According to Reichenbach, any physical theory presupposes the validity of systems of certain, usually quite general, principles, which may vary from theory to theory. Such "coordinating principles", as they are then termed, are indispensable for the ordering of perceptual data; they define "the objects of knowledge" within the theory. The epistemological significance of relativity theory, according to the young Reichenbach, is to have shown, contrary to Kant, that these systems may contain mutually inconsistent principles, and so require emendation to remove contradictions. Thus a "relativization" of the Kantian conception of *synthetic a priori* principles is the direct epistemological result of the theory of relativity. But this finding is also taken to signal a transformation in the method of epistemological investigation of science. In place of Kant's "analysis of Reason", "the method of analysis of science"*(der wissenschaftsanalytische Methode)* is proposed as "the only way that affords us an understanding of the contribution of our reason to knowledge" (1920, 71; 1965, 74). The method's *raison d'être* is to sharply distinguish between the "subjective" role of (coordinating) principles — "the contribution of Reason" — and the "contribution of objective reality", represented by theory-specific empirical laws and regularities ("axioms of connection") which in some sense have been "constituted" by the former. Relativity theory itself is a shining exemplar of this method for it has shown that the metric of spacetime describes an "objective property" of the

world, once the subjective freedom to make coordinate transformations (the coordinating principle of general covariance) is recognized (1920, 86-7; 1965, 90). The thesis of metric conventionalism had yet to appear.

But soon it did. Still in 1920, Schlick objected, both publicly and in private correspondence with Reichenbach, that "principles of coordination" were precisely statements of the kind that Poincaré had termed "conventions" (see Coffa, 1991, 201ff.). Moreover, Einstein, in lecture of January, 1921, entitled "Geometry and Experience", appeared to lend support to this view. Einstein argued that the question concerning the nature of spacetime geometry becomes an empirical question only on certain *pro tem* stipulations regarding the "practically rigid body" of measurement (*pro tem* in view of the inadmissibility in relativity theory of the concept "actually rigid body"). In any case, by 1922, the essential pieces of Reichenbach's "mature" conventionalist view had emerged. The argument is canonically presented in §8 (entitled "The Relativity of Geometry") of *Der Philosophie der Raum-Zeit-Lehre* (completed in 1926, published in 1928). In a move superficially similar to the argument of Einstein's "Geometry and Experience", Reichenbach maintained that questions concerning the empirical determination of the metric of spacetime must first confront the fact that only the whole theoretical edifice comprising geometry and physics admits of observational test. Einstein's gravitational theory is such a totality. However, unlike Einstein, Reichenbach's "method of analysis of science", later re-named "logical analysis of science", is directed to the epistemological problem of factoring this totality into its conventional or definitional and its empirical components.

This is done as follows. The empirical determination of the spacetime metric by measurement requires choice of some "metrical indicators": this can only be done by laying down a "coordinative definition" stipulating, e.g., that the metrical notion of a "length" is coordinated to some physical object or process. A standard choice coordinates "lengths" with "infinitesimal measuring rods" supposed rigid (e.g., Einstein's "practically rigid body"). This however is only a convention, and other physical objects or processes might be chosen. (In Schlick's fanciful example, the Dali Lama's heartbeat could be chosen as the physical process establishing units of time.) Of course, the chosen metrical indicators must be corrected for certain distorting effects (temperature, magnetism, etc.) due to the presence of physical forces. Such forces are termed "differential forces" to indicate that they affect various materials differently. However, Reichenbach argued, the choice of a rigid rod as standard of length is tantamount to the claim that there are no non-differential — "universal" — distorting forces that affect all bodies in the same way and cannot be screened off. In the absence of "universal forces" the coordinative definition regarding rigid rods can be implemented and the nature of the spacetime metric empirically determined, for example, finding that paths of light rays through solar gravitational field are not Euclidean straight lines. Thus, the theory of general relativity, on adoption of the coordinative definition of rigid rods ("universal forces = 0"), affirms that the geometry of spacetime in this region is of a non-euclidean kind. The point, however, is that this conclusion rests on the convention governing measuring rods. One could, alternately, maintain that the geometry of spacetime was Euclidean by adopting a different coordinative definition, for example, holding that measuring rods expanded or contracted depending on their position in spacetime, a choice tantamount to the supposition of "universal forces". Then, consistent with all empirical phenomena, it could be maintained that Euclidean geometry was compatible with Einstein's theory if only one allowed the existence of such forces. Thus whether general relativity affirms a Euclidean or a

non-euclidean metric in the solar gravitational field rests upon a conventional choice regarding the existence of "universal forces". Either hypothesis may be adopted since they are empirically equivalent descriptions; their joint possibility is referred to as "the relativity of geometry". Just as with the choice of "standard synchrony" in Reichenbach's analysis of the conventionality of simultaneity, also held to be "logically arbitrary", Reichenbach recommends the "descriptively simpler" alternative in which "universal forces" do not exist. To be sure, "descriptive simplicity has nothing to do with truth", i.e., has no bearing on the question of whether spacetime has a non-Euclidean structure (1928, 47; 1958, 35).

4.3 Critique of Reichenbachian Metric Conventionalism

In retrospect, it is rather difficult to understand the significance that has been accorded this argument. Carnap, for example, in his "Introductory Remarks" (1958, vii) to the posthumous English translation of this work, singled it out on account of its "great interest for the methodology of physics". Reichenbach himself deemed "the philosophical achievement of the theory of relativity" to lie in this methodological distinction between conventional and factual claims regarding spacetime geometry (1928, 24; 1958, 15), and he boasted of his "philosophical theory of relativity" as an incontrovertible "philosophical result":

> *the philosophical theory of relativity*, i.e., the discovery of the definitional character of the metric in all its details, holds independently of experience….a philosophical result not subject to the criticism of the individual sciences." (1928, 223; 1958, 177)

Yet this result is neither incontrovertible nor an untrammeled consequence of Einstein's theory of gravitation. There is, first of all, the shadowy status accorded to "universal forces". A sympathetic reading (e.g., Dieks (1987)) suggests that the notion serves usefully in mediating between a traditional *a priori* commitment to Euclidean geometry and the view of modern geometrodynamics, where gravitational force is "geometrised away" (see §5). For, as Reichenbach explicitly acknowledged, gravitation is itself a "universal force", coupling to all bodies and affecting them in the same manner (1928, 294-6; 1958, 256-8). Hence the choice recommended by "descriptive simplicity" is merely a stipulation that metrical appliances, regarded as "infinitesimal", be considered as "differentially at rest" in an inertial system (1924, 115; 1969, 147). This is a stipulation that spacetime measurements always take place in regions that are to be considered small Minkowski spacetimes (arenas of gravitation-free physics). By the same token, however, consistency required an admission that "the transition from the special theory to the general one represents merely a renunciation of metrical characteristics" (1924, 115; 1969, 147), or, even more pointedly, that "all the metrical properties of the spacetime continuum are destroyed by gravitational fields" where only topological properties remain (1928, 308; 1958, 268-9). To be sure, these conclusions are supposed to be rendered more palatable in connection with the epistemological reduction of spacetime structures in the causal theory of time.

Despite the influence of this argument on the subsequent generation of philosophers of science, Reichenbach's analysis of spacetime measurement treatment is plainly inappropriate, manifesting a fallacious tendency to view the generically curved spacetimes of general relativity as stiched together from little bits of flat Minskowski spacetimes. Besides being mathematically inconsistent, this procedure

offers no way of providing a non-metaphorical physical meaning for the fundamental metrical tensor $g_{\mu\nu}$, the central theoretical concept of general relativity, nor to the series of curvature tensors derivable from it and its associated affine connection. Since these sectional curvatures at a point of spacetime are empirically manifested and the curvature components can be measured, e.g., as the tidal forces of gravity, they can hardly be accounted as due to conventionally adopted "universal forces". Furthermore, the concept of an "infinitesimal rigid rod" in general relativity cannot really be other than the interim stopgap Einstein recognized it to be. For it cannot actually be "rigid" due to these tidal forces; in fact, the concept of a "rigid body" is already forbidden in special relativity as allowing instantaneous causal actions. Secondly, such a rod must indeed be "infinitesimal", i.e., a freely falling body of negligible thickness and of sufficiently short extension, so as to not be stressed by gravitational field inhomogeneities; just how short depending on strength of local curvatures and on measurement error (Torretti (1983), 239). But then, as Reichenbach appeared to have recognized in his comments about the "destruction" of the metric by gravitational fields, it cannot serve as a coordinately defined general standard for metrical relations. In fact, as Weyl was the first to point out, precisely which physical objects or structures are most suitable as measuring instruments should be decided on the basis of gravitational theory itself. From this enlightened perspective, measuring rods and clocks are objects that are far too complicated. Rather, the metric in the region around any observer O can be empirically determined from freely falling ideally small neutral test masses together with the paths of light rays. More precisely stated, the spacetime metric results from the affine-projective structure of the behavior of neutral test particles of negligible mass and from the conformal structure of light rays received and issued by the observer. (Weyl, 1921) Any purely conventional stipulation regarding the behavior of "measuring rods" as physically constitutive of metrical relations in general relativity is then otiose (Weyl, 1923a; Ehlers, Pirani and Schild (1973)). Alas, since Reichenbach reckoned the affine structure of the gravitational-inertial field to be just as conventional as, on his view, its metrical structure, he was not able to recognize this method as other than an equivalent, but by no means necessarily preferable, account of the empirical determination of the metric through the use of rods and clocks (Coffa, 1979; Ryckman (1994), (1996)).

5. "Geometrization of Physics": Realism and Transcendental Idealism

5.1 Differing Motivations

In the decade or so following the appearance of the general theory of relativity, there was much talk of a "geometrization" of physics (Weyl (1918b), (1919); Haas (1920); Lodge (1921)). While these discussions were largely, and understandably, confined to scientific circles, they nonetheless brought distinctly philosophical issues — of methodology, but also of epistemology and metaphysics — together with technical matters. General relativity revived a geometrizing tendency essentially dormant within physics since the 17th century. In so doing, it opened up the prospect of a "geometrization" of physics, the possibility of finding a unifying representation of all of known physics within a single geometrical theory of the spacetime continuum. Einstein himself, however, was not the first to embark on this audacious quest. Rather he initially followed in the mathematical footsteps of Hermann Weyl, Arthur Stanley

Eddington, and Theodore Kaluza, only gradually (1925) devising the first of his own "homegrown" geometrical "unified field theories". Still, by 1923, Einstein had become the recognized leader of the unification program. (Vizgin (1994), 265)

The first phase of the geometrical unification program essentially ended with Einstein's "distant parallelism" theory of 1928-1931 (1929), perhaps Einstein's final public sensation (Fölsing (1997, 605)). Needless to say, none of these efforts met with success. In a lecture at the University of Vienna on October 14, 1931, Einstein forlornly referred to these failed attempts, each conceived on a different differential geometrical basis, as a "graveyard of dead hopes" (Einstein, 1932). By this time, certainly, the prospects for the geometrical unification program had considerably waned. A consensus emerged among nearly all leading theoretical physicists that while the geometrical unification of the gravitation and electromagnetic fields might be attained in formally different ways, the problem of matter, treated with undeniable empirical success by the new quantum theory, was not to be resolved within the confines of spacetime geometry. In any event, from the early 1930s, any unification program appeared greatly premature, in view of the wealth of data produced by the new physics of the nucleus.

As many will know, the unsuccessful pursuit of the goal of geometrical unification absorbed Einstein, and his various research assistants, for more than three decades, up to Einstein's death in 1955. In the course of it, Einstein's methodology of research diametrically changed. In place of physical or heuristic principles to guide theoretical construction, such as the principle of equivalence, which put him on the path to general relativity, he increasingly relied on considerations of mathematical aesthetics, such as "logical simplicity" and the inevitability of certain mathematical structures under variously adopted constraints. In a talk entitled "On the Method of Theoretical Physics" at Oxford in 1933, the transformation was stated dramatically:

> Experience remains, of course, the sole criterion of the physical utility of a mathematical construction. But the creative principle resides in mathematics. In a certain sense, therefore, I hold it true that pure thought can grasp reality, as the ancients dreamed. (274)

Moreover, the advent and accumulating empirical successes of the new quantum theory did not dislodge Einstein's core metaphysical belief in a physical reality conceived as a continuous "total field" whose components are functions of the spacetime variables, a geometrical conception of physical reality implied, to be sure, by general relativity (e.g., (1950), 348). Yet, whatever may have been Kaluza's philosophical motivations in putting forward his proposal for geometrical unification, neither Einstein's mathematical realism nor his metaphysics guided either Weyl or Eddington, a fact that has often been obscured or ignored in historical treatments. The geometrical unifications of Weyl (1918a,b) and Eddington (1921) were above all explicit attempts to comprehend the nature of physical theory, in the light of general relativity, from systematic epistemological standpoints that were neither positivist nor realist. As such they comprise "early philosophical interpretations" of that theory, although they intertwine philosophy, geometry and physics in a manner unprecedented since Descartes. Before turning attention to their "interpretations", it will be helpful to see how the geometrizing tendency arises within general relativity itself and to note a few details of the geometrical unification program that followed in its wake.

5.2 "Geometrizing" Gravity: the Initial Step

Einstein's so-called "geometrization" of gravitational force in 1915 gave the geometrization program its first, partial, realization as well as its subsequent impetus. In Einstein's theory, the fundamental or "metric" tensor *g* of Riemannian geometry appears in a dual role which thoroughly fuses its geometrical and its physical meanings. As is apparent from the expression for the "interval" between neighboring spacetime events, $ds^2 = g_{\mu\nu}\, dx^\mu\, dx^\nu$ (here, and below there is an implicit summation over repeated upper and lower indices), the metric tensor is at once the geometrical quantity underlying measurable metrical relations of lengths and times. In this role it ties a mathematical theory of events in four-dimensional "curved" spacetime to observations and measurements in space and time. But it is also the "potential" of the gravitational (or "metrical") field whose value, at any point of spacetime, depends, via the Einstein Field Equations (see below), on the presence of physical quantities of mass-momentum-stress in the immediate region. In the new view, the idea of strength of gravitational "force" is replaced by that of degree of "curvature" of spacetime. Such a curvature is manifested, for example, by the "tidal force" of the Earth's gravitational field that occasions two freely falling bodies, released at a certain height and at fixed separation, to approach one another. A freely falling body is no longer to be regarded as moving through space according to the "pull" of an attractive gravitational "force", but simply as tracing out the "laziest" track along the bumps and hollows of spacetime itself. The Earth's mass (or equivalently, energy) determines a certain spacetime curvature and so becomes a source of gravitational action. At the same time, the gross mechanical properties of bodies, comprising all gravitational-inertial phenomena, can be derived as the solution of a single system of generally covariant partial differential equations, the Einstein equations of the gravitational field. According to these equations, spacetime and matter stand in dynamical interaction. One abbreviate way of characterizing the dual role of the $g_{\mu\nu}$ is to say that in the general theory of relativity, gravitation, which includes mechanics, has become "geometrized", i.e., incorporated into the geometry of spacetime.

5.3 Extending "Geometrization"

In making spacetime curvature dependent on distributions of mass and energy, general relativity is indeed capable of encompassing all (non-quantum) physical fields. However, in classical general relativity there remains a fundamental asymmetry between gravitational and non-gravitational fields, in particular, electromagnetism, the only other fundamental physical interaction definitely known at the time. This shows up visibly in one form of the Einstein field equations in which, on the left-hand side, a geometrical object ($G_{\mu\nu}$, the Einstein tensor) built up from the uniquely compatible linear symmetric ("Levi-Civita") connection associated with the metric tensor $g_{\mu\nu}$, and representing the curvature of spacetime, is set identical to a tensorial but non-geometrical phenomenological representation of matter on the right-hand side.

$$G_{\mu\nu} = k\, T_{\mu\nu}, \quad \text{where } G_{\mu\nu} = R_{\mu\nu} - 1/2\, g_{\mu\nu} R$$

The expression on the right side, introduced by a coupling constant, mathematically represents the non-gravitational sources of the gravitational field in a region of spacetime in the form of a stress-energy-

momentum tensor (an "*omnium gatherum*" in Eddington's pithy phrase (1919, 63)). As the geometry of spacetime principally resides on the left-hand side, this situation seems unsatisfactory. Late in life, Einstein likened his famous equation to a building, one wing of which (the left) was built of "fine marble", the other (the right) of "low grade wood" (1936, 311). In its classical form, general relativity accords only the gravitational field a direct geometrical significance; the other physical fields reside *in* spacetime; they are not *of* spacetime.

Einstein's dissatisfaction with this asymmetrical state of affairs was palpable at an early stage and was expressed with increasing frequency beginning in the early 1920s. A particularly vivid declaration of the need for geometrical unification was made in his "Nobel lecture" of July, 1923:

> The mind striving after unification of the theory cannot be satisfied that two fields should exist which, by their nature, are quite independent. A mathematically unified field theory is sought in which the gravitational field and the electromagnetic field are interpreted as only different components or manifestations of the same uniform field,… The gravitational theory, considered in terms of mathematical formalism, i.e. Riemannian geometry, should be generalized so that it includes the laws of the electromagnetic field."(489)

It might be noted that the tacit assumption, evident here, that incorporation of electromagnetism into spacetime geometry requires a generalization of the Riemannian geometry of general relativity, though widely held at the time, is not quite correct (Rainich (1925); Misner and Wheeler (1962); Geroch (1966)).

5.4 A "Pure Infinitesimal Geometry"

Still, it wasn't Einstein, but the mathematician Hermann Weyl who first addressed the asymmetry in 1918 in the course of refashioning Einstein's theory on the preferred epistemological basis of a "pure infinitesimal geometry" *(Reine Infinitesimalgeometrie).* Holding that direct — *evident*, in the sense of Husserlian phenomenology --comparisons of length or duration could be made at neighboring points of spacetime, but not, as the Riemannian geometry of Einstein's theory permitted, "at a distance", Weyl discovered additional terms in his geometry that he identified with the potentials of the electromagnetic field. From these, the electromagnetic field strengths can be immediately derived and so electromagnetism as well as gravitation could be expressed solely within the terms of spacetime geometry. As no other interactions were definitely known to occur, Weyl proudly declared that the concepts of geometry and physics were the same. Hence, everything in the physical world was a manifestation of spacetime geometry.

> (The) distinction between geometry and physics is an error, physics extends not at all beyond geometry: the world is a (3+1) dimensional metrical manifold, and all physical phenomena transpiring in it are only modes of expression of the metric field, …. (M)atter itself is dissolved in "metric" and is not something substantial that in addition exists "in" metric space (1919, 115-16).

By the winter of 1919-1920, for both physical and philosophical reasons (the latter having to do with his conversion to Brouwer's "intuitionist" views about the mathematical continuum, in particular, the continuum of spacetime), Weyl (1920) surrendered the belief, expressed here, that matter, with its corpuscular structure, might be derived within spacetime geometry. Thus he gave up the Holy Grail of the nascent unified field theory program almost before it had begun. Nonetheless, he actively defended his theory well into the 1920s, essentially on the grounds of Husserlian transcendental phenomenology, that his geometry and its central principle, "the epistemological principle of relativity of magnitude" comprised a superior epistemological framework for general relativity. Weyl's postulate of a "pure infinitesimal" non-Riemannian metric for spacetime, according to which it must be possible to independently choose a "gauge" (scale of length or duration) at each spacetime point, met with intense criticism. No observation spoke in favor of it; to the contrary, Einstein pointed out that according to Weyl's theory, the atomic spectra of the chemical elements should not be constant, as indeed they are observed to be. Although Weyl responded to this objection forcefully, and with some subtlety (Weyl, 1923a), he was able to persuade neither Einstein, nor any other leading relativity physicist, with the exception of Eddington. However, the idea of requiring "gauge invariance" of fundamental physical laws was revived and vindicated by Weyl himself in a different form later on (Weyl (1929);see also O'Raifeartaigh (1997)).

5.5 Eddington's "World Geometry"

Despite Weyl's failure to win many friends for his theory, his guiding example of unification launched the geometrical program of "unified field theory", initiating a variety of efforts, all aimed at finding a suitable generalization of the Riemannian geometry of Einstein's theory to encompass as well non-gravitational physics (Vizgin (1994), ch.4). In December, 1921, the Berlin Academy published Theodore Kaluza's novel proposal for unification of gravitation and electromagnetism upon the basis of a five-dimensional Riemannian geometry. But earlier that year, in February, came Arthur Stanley Eddington's further generalization of Weyl's four-dimensional geometry, wherein the sole primitive geometrical notion is the non-metrical comparison of direction or orientation at the same or neighboring points. In Weyl's geometry the magnitude of vectors at the same point, but pointing in different directions, might be directly compared to one another; in Eddington's, comparison was immediate only for vectors pointing in the same direction. His "theory of the affine field" included both Weyl's geometry and the semi-Riemannian geometry of Einstein's general relativity as special cases. Little attention was paid however, to Eddington's claim, prefacing his paper, that his objective had not been to "seek (the) unknown laws (of matter)" as befits a unified field theory. Rather it lay "in consolidating the known (field) laws" wherein "the whole scheme seems simplified, and new light is thrown on the origin of the fundamental laws of physics" (1921, 105).

Eddington was persuaded that Weyl's "principle of relativity of length" was "an essential part of the relativistic conception", a view he retained to the end of his life (e.g., (1939, 28)). But he was also convinced that the largely antagonistic reception accorded Weyl's theory was due to its confusing formulation. The flaw lay in Weyl's failure to make transparently obvious that his locally scale invariant ("pure infinitesimal") "world geometry" was not the physical geometry of actual spacetime, but an entirely mathematical geometry inherently serving to specify the ideal of an observer-independent

external world. To remedy this, Eddington devised a general method of deductive presentation of field physics in which "world geometry" is developed mathematically as conceptually separate from physics. A "world geometry" is a purely mathematical geometry the derived objects of which possess only the structural properties requisite to the ideal of a completely impersonal world; these are objects, as he wrote in *Space, Time and Gravitation* (1920), a semi-popular best-seller, represented "from the point of view of no one in particular". Naturally, this ideal had changed with the progress of physical theory. In the light of relativity theory, such a world is indifferent to specification of reference frame and, after Weyl, of gauge of magnitude (scale). A "world geometry" is not the physical theory of such a world but a framework or "graphical representation" in whose terms existing physical theory might be displayed, essentially by the mathematical identification of known tensors of the existing physical laws of gravitation and electromagnetism, with tensors of the world geometry. Such a geometrical representation of physics cannot really be said to be "right" or "wrong", for it only implements, if it can, current ideas governing the conception of objects and properties of an impersonal objective external world. But when existing physics, in particular, Einstein's theory of gravitation, is set in the context of Eddington's world geometry, it yields a surprising consequence: The Einstein law of gravitation appears as a definition! In the form $R^{\mu\nu} = 0$ it defines what in the "world geometry" appears to the mind as "vacuum" while in the form of the Einstein field equation noted above, it defines what is there encountered by the mind as "matter". This result is what was meant by his stated claim of throwing "new light on the origin of the fundamental laws of physics". Eddington's notoriously difficult and opaque later works (1936), (1946), took their inspiration from this argumentation in attempting to carry out a similar, but algebraic, program of deriving fundamental physical laws, and the constants occurring in them, from epistemological principles.

5.6 Meyerson on "Pangeometrism"

Within physics the idealist currents lying behind the "world geometries" of Weyl and Eddington were largely ignored, whereas within philosophy, with the notable exception of Émile Meyerson's *La Déduction Relativiste* (1925),most philosophers lacked the tools to connect these readily discernible currents with their geometrical theories. Meyerson, who had no doubt of the basic realist impetus of science, carefully distinguished Einstein's "rational deduction of the physical world" from the geometrical unifications of gravitation and electromagnetism of Weyl and Eddington. These theories, as affirmations of a complete *panmathematicism*, or rather of a *pangeometrism* (§§ 157-58), were compared to the rational deductions of Hegel's *Logic*. That general relativity succeeded in partly realizing Descartes' program of reducing the physical to the spatial through geometric deduction, is due to the fact that Einstein "followed in the footsteps" of Descartes, not Hegel (§133). But *pan-geometrism* is also capable of overreaching itself and this is the transgression committed by Weyl and Eddington. Weyl in particular is singled out for criticism for seemingly to have reverted to Hegel's monistic idealism, and so to be subject to its fatal flaw. In regarding nature as completely intelligible, Weyl had abolished the thing-in-itself and so promoted the identity of self and non-self, the great error of the *Naturphilosophien*.

Though he had "all due respect to the writings of such distinguished scientists" as Weyl and Eddington, Meyerson took their overt affirmations of idealism to be misguided attempts "to associate themselves with a philosophical point of view that is in fact quite foreign to the relativistic doctrine" (§150). That

"point of view" is in fact two distinct species of transcendental idealism. It is above all "foreign" to relativity theory because Meyerson cannot see how it is possible to "reintegrate the four-dimensional world of relativity theory into the self". After all, Kant's own argument for Transcendental Idealism proceeded "in a single step", in establishing the subjectivity of the space and time of "our naïve intuition". But this still leaves "the four dimensional universe of relativity independent of the self". Any attempt to "reintegrate" four-dimensional spacetime into the self would have to proceed at a "second stage" where, additionally, there would be no "solid foundation" such as spatial and temporal intuition furnished Kant at the first stage. Perhaps, Meyerson allowed, there is indeed "another intuition, purely mathematical in nature", lying behind spatial and temporal intuition, and capable of "imagining the four-dimensional universe, to which, in turn, it makes reality conform". This would make intuition a "two-stage mechanism". While all of this is not "inconceivable", it does appear, nonetheless, "rather complex and difficult if one reflects upon it". In any case, this is likely to be unnecessary, for considering the matter "with an open mind",

> one would seem to be led to the position of those who believe that relativity theory tends to destroy the concept of Kantian intuition (§§ 151-2).

Meyerson had come right up to the threshold of grasping the Weyl-Eddington geometric unification schemes in something like the sense in which they were intended. The stumbling block for him, and for others, is the conviction that transcendental idealism can be supported only from an argument about the nature of intuition, and intuitive representation. To be sure, the geometric framework for Weyl's construction of the objective four-dimensional world of relativity is based upon the *Evidenz* available in "essential insight", which is limited to the simple linear relations and mappings in what is basically the tangent vector space to a point in a manifold. Thus in Weyl's differential geometry there is a fundamental divide between integrable and non-integrable relations of comparison. The latter are primitive and epistemologically privileged, but nonetheless not justified until it is shown how the infinitesimal homogenous spaces, corresponding to the "essence of space as a form of intution", are compatible with the large-scale inhomogenous spaces (spacetimes) of general relativity. And this required not a philosophical argument about the nature of intution, but one formulated in group-theoretic *conceptual* form. (Weyl, 1923a,b). Eddington, on the other hand, without the cultural context of Husserlian phenomenology or indeed of philosophy generally, jettisoned the intuitional basis of transcendental idealism altogether, as if unaware of its prominence. Thus he sought a superior and completely general *conceptual* basis for the objective four-dimensional world of relativity theory by constituting that world within a geometry (its "world structure" (1923)) based upon a non-metrical affine (i.e., linear and symmetric) connection. He was then free to find his own way to the empirically confirmed integrable metric relations of Einstein's theory without being hampered by the conflict of a "pure infinitesimal" metric with the observed facts about rods and clocks.

5.7 "Structural Realism"?

It has been routinely assumed that all the attempts at a "geometrization of physics" in the early unified field theory program shared something of Einstein's hubris concerning the ability of mathematics to "grasp" the fundamental structure of the external world. The geometrical unified field theory program

thus appears to be inseparably stitched to a form of scientific realism, recently termed "structural realism", with perhaps even an inspired turn toward Platonism. According to "structural realism", whatever the "nature" of the fundamental entities comprising the physical world, only their "structure" can be known as that structure is represented in the equations of the theory. The sole ontological continuity across changes in fundamental physical theory is a continuity of structure, as the equations of the earlier theory can be derived,say as limit cases, from those of the later. Geometrical unification theories seems tailored for this kind of realism. For if a geometrical theory is taken to give a true or approximately true representation of the physical world, what is geometrically represented has the definite structure of the fundamental geometrical relations. But for Weyl and Eddington, geometrical unification was not, nor could be, such a representation, for essentially the reasons articulated two decades before by Poincaré (1906,14):

> Does the harmony the human intelligence thinks it discovers in nature exist outside of this intelligence? No, beyond doubt, a reality completely independent of the mind which conceives it, sees or feels it, is an impossibility. A world as exterior as that, even if it existed, would for us be forever inaccessible. But what we call objective reality is, in the last analysis, what is common to many thinking beings, and could be common to all; this common part,...,can only be the harmony expressed by mathematical laws. It is this harmony then which is the sole objective reality....

In Weyl and Eddington, geometrical unification was an attempt to cast the "harmony" of the Einstein theory of gravitation in a new epistemological and explanatory light, by displaying the great field laws of gravitation and electromagnetism within the common frame of a geometrically represented objective reality. Their unorthodox manner of philosophical argument, cloaked, perhaps necessarily, in the language of differential geometry, has tended to conceal or obscure conclusions about the significance of a "geometrized physics" that push in considerably different directions from either instrumentalism or scientific realism.

Bibliography

- Barbour, J., and Pfister, H. (eds.) (1995). *Mach's Principle: From Newton's Bucket to Quantum Gravity*. (Boston, Basel, Berlin: Birkhäuser).
- Bohr, N., Kramers, H.A., and Slater, J.C. (1924). "The Quantum Theory of Radiation", *Philosophical Magazine* 47, 785-802.
- Bollert, K. (1921). *Einstein's Relativitätstheorie und ihre Stellung im System der Gesamterfahrung.* (Leipzig: Theodor Steinkopff).
- Bridgman, P. (1949). "Einstein's Theories and the Operational Point of View", in P.A. Schilpp (ed), *Albert Einstein: Philosopher-Scientist*. (Evanston: Northwestern University Press). 335-354.
- Carnap, R. (1922). *Der Raum. Ein Beitrag zur Wissenschaftslehre*. (Berlin: Reuther & Reichard). (*Kant-Studien* Ergänzungshefte, no.56).
- ----- (1956). "Introductory Remarks to the English Edition", dated July, 1956, in H. Reichenbach (1958), v- vii.

- Cassirer, E. (1910). *Substanzbegriff und Funktionsbegriff*. (Berlin: Bruno Cassirer). translated by W.C. Swabey and M.C. Swabey in *Substance and Function and Einstein's Theory of Relativity*. (Chicago: Open Court, 1923); reprint (New York: Dover, 1953), 1-346.
- ----- (1921). *Zur Einstein'schen Relativitätstheorie*. (Berlin: Bruno Cassirer). Pagination as reprinted in *Zur Modernen Physik*. (Darmstadt: Wissenschaftliche Buchgesellschaft, 1957), 1-125. Translated by W.C. Swabey and M.C. Swabey in *Substance and Function and Einstein's Theory of Relativity*. (Chicago: Open Court, 1923); reprint (New York: Dover, 1953), 345-460.
- Coffa, J.A. (1979). "Elective Affinities: Weyl and Reichenbach", in W. Salmon (ed.), *Hans Reichenbach: Logical Empiricist* (Dordrecht: D. Reidel), 267-304.
- ----- (1991).*The Semantic Tradition from Kant to Carnap: To the Vienna Station* (Cambridge: Cambridge University Press).
- Dieks, D. (1987). "Gravitation as a Universal Force", *Synthese* 73, 381-97.
- Eddington, A.S. (1919), *Report on the Relativity Theory of Gravitation*. (London: The Physical Society of London, Fleetway Press).
- ----- (1921) "A Generalization of Weyl's Theory of the Electromagnetic and Gravitational Fields, *Proceedings of the Royal Society of London, Series A* 99, 104-122.
- ----- (1923). *The Mathematical Theory of Relativity* (Cambridge: Cambridge University Press).
- ----- (1936). *The Relativity Theory of Protons and Electrons*. (Cambridge: Cambridge University Press).
- ----- (1939). *The Philosophy of Physical Science.(The Tarner Lectures, 1938)*. (Cambridge: Cambridge University Press)
- ----- (1948). *Fundamental Theory*. (Cambridge: Cambridge University Press). Posthumously published.
- Ehlers, J., Pirani, F., and Schild, A. (1973). "The Geometry of Freefall and Light Propagation", in L. O' Raifeartaigh (ed.), *General Relativity: Papers in Honour of J.L. Synge*. (Oxford: Clarendon Press), 63-84.
- Einstein, A. (1905). "Zur Elektrodynamik bewegter Körper", *Annalen der Physik* 17; 891-921; reprinted in Einstein (1989), 275-310. Translation by A. Beck in Einstein (1989), supplement, 140-171.
- ----- (1916a). "Die Grundlage der allgemeine Relativitätstheorie", *Annalen der Physik*, 49: 769-822; separately printed (Berlin: J. Springer). Reprinted in Einstein (1996), 284-339. Translation by W. Parret and G.B. Jeffrey in A. Einstein *et al.*, *The Principle of Relativity*.(London: Methuen, 1923); reprint (New York: Dover, n.d.), 109-173.
- ----- (1916b) "Ernst Mach", *Physikalische Zeitschrift*, 17, 101-04; reprinted in Einstein (1996), 278-281.
- ----- (1917a). *Über die spezielle und die allgemeine Relativitätstheorie*. (Braunschweig: Vieweg). Translation by R.W. Lawson as *Relativity: The Special and the General Theory*. (New York: Crown Publishers, 1955).
- ----- (1917b). "Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie", *Sitzungsberichte der preuß. Akad. Berlin, Math. Kl.*, 142-157. Translation by W. Parret and G.B. Jeffrey in A. Einstein *et al.*, *The Principle of Relativity*. (London: Methuen, 1923); reprint (New York: Dover, n.d.), 175-188.
- ----- (1918). "Prinzipielles zur allgemeinen Relativitätstheorie", *Annalen der Physik* 55, 241-244.

- ----- (1921). "Geometrie und Erfahrung", *Sitzungsberichte der preuß. Akad. Berlin, Math. Kl.*, 123-130. Expanded and issued separately, (Berlin: J.Springer, 1921). Translation in *Sidelights on Relativity*. (London: Methuen,1922); reprint (New York: Dover, 1983), 27-56.
- ----- (1922). Remarks in *Bulletin de la Société Française de Philosophie*, 17; as reprinted in *la pensée*, 210 (1980), 12-29.
- ----- (1923). "Fundamental Ideas and Problems of the Theory of Relativity" (Lecture delivered to the Nordic Assembly of Naturalists at Gothenburg, July 11, 1923), as translated in *Nobel Lectures Physics, 1901-1921* (Elsevier, Amsterdam-London-New York, 1967), 482-90.
- ----- (1924). Review of *Relativitätstheorie und Erkenntnislehre* by J. Winternitz. *Deutsche Literaturzeitung*, paragraphs 20-22.
- ----- (1925). "Einheitliche Feldtheorie von Gravitation und Elektrizität", *Sitzungsberichte der preuß. Akad. Berlin, Math. Kl.*, 414-419.
- ----- (1929). "Zur einheitliche Feldtheorie", *Sitzungsberichte der preuß. Akad. Berlin, Math. Kl.*, 2-7.
- ----- (1932). "Der gegenwärtige Stand der Relativitätstheorie," *Die Quelle (Pedagogischer Führer)*, 82, 440-42
- ----- (1933). "On the Method of Theoretical Physics", as translated by S. Bargmann in *Albert Einstein: Ideas and Opinions*. (New York: Bonanza Books,1954), 270-276.
- ----- (1936). "Physik und Realität", *The Journal of the Franklin Institute*, v.221; 313-337; as translated by S. Bargmann in *Albert Einstein: Ideas and Opinions* (New York: Bonanza Books,1954), 290-323.
- ----- (1949a). "Autobiographical Notes", in P. A. Schilpp (ed.), *Albert Einstein: Philosopher-Scientist* (Evanston: Northwestern University Press), 2-95.
- ----- (1949b). "Replies to Criticisms" in P.A. Schilpp (ed), *Albert Einstein: Philosopher-Scientist*. (Evanston: Northwestern University Press), 665-88.
- ----- (1989). *The Collected Papers of Albert Einstein*, v.2, J. Stachel (ed.) (Princeton: Princeton University Press); English translation supplement published separately.
- ----- (1996). *The Collected Papers of Albert Einstein*, v.6, A.J. Kox, M.J. Klein, and R. Schulmann (eds.) (Princeton: Princeton University Press).
- Fölsing, A. (1997). *Albert Einstein; A Biography*. (New York: Viking).
- Frank, P. (1917). "Die Bedeutung der physikalischen Erkenntnistheorie Machs für die Geistesleben der Gegenwart", *Die Naturwissenschaften* 5, 65-71; translated as "The Importance for Our Times of Ernst Mach's Philosophy of Science", in *Modern Science and its Philosophy*. (Cambridge, MA: Harvard University Press, 1950), 61-78.
- Friedman, M. (1983). *Foundations of Space-Time Theories*. (Princeton: Princeton University Press).
- ----- (1999). "Geometry, Convention, and the Relativized A Priori: Reichenbach, Schlick, and Carnap", reprinted in *Reconsidering Logical Positivism* (Cambridge: Cambridge University Press, 1999), 59-70.
- Geroch, R. (1966). "Electromagnetism as an Aspect of Geometry? Already Unified Field Theory — The Null Field Case", *Annals of Physics* 36, R. (1966), 147-187.
- Gödel, K. (1946/9-B2). "Some Observations about the Relationship between Theory of Relativity and Kantian Philosophy", in *Collected Works*, v.II. (New York and Oxford: Oxford University

Press, 1995), 230-59.

- ----- (1949). "An Example of a New Type of Cosmological Solutions of Einstein’s Field Equations of Gravitation", *Reviews of Modern Physics* 21, 447-450; reprinted in *Kurt Gödel Collected Works* v.II. (New York and Oxford: Oxford University Press, 1990), 190-198.
- Hentschel, K. (1990). *Interpretationen und Fehlinterpretationen der speziellen und der allgemeinen Relativitätstheorie durch Zeitgenossen Albert Einsteins*. (Basel, Boston, Berlin: Birkhäuser).
- Haas, A. (1920). "Die Physik als geometrische Notwendigkeit", *Die Naturwissenschaften* 8, 121-140.
- Howard, D. (1984). "Realism and Conventionalism in Einstein’s Philosophy of Science: The Einstein-Schlick Correspondence", *Philosophia Naturalis* 21, 616-629.
- Howard, D., and Stachel, J. (eds.) (1989). *Einstein and the History of General Relativity.* (Boston, Basel, Berlin: Birkhäuser).
- Kaluza, T. (1921). "Zum Unitätsproblem der Physik", *Sitzungsberichte der. preuß. Akad. Berlin, Math. Kl.*, 966-972. Translation in O’Raifeartaigh (1997), 53-58. Another translation, together with translations of the relevant letters from Einstein to Kaluza, all by C. Hoenselaers, appears in V. de Sabbata and E. Schmutzer (eds.), *Unified Field Theories of More than 4 Dimensions* (Singapore: World Scientific, 1983), 427-33; 447-57.
- Lodge, O. (1921). "The Geometization of Physics, and its Supposed Basis on the Michelson-Morley Experiment", *Nature* 106, 795-802.
- Mach, E. (1883). *Die Mechanik in ihrer Entwicklung: Historisch-Kritisch Dargestellt.* (Leipzig: Brockhaus). Translation of the 7th German edition (1912), by T.J. McCormack, with revisions through the 9th German edition (1933), as *The Science of Mechanics: A Critical and Historical Account of its Development.* (LaSalle, IL: Open Court Publishing Co., 1960).
- ----- (1921). *Die Prinzipien der physikalischen Optik. Historisch und erkenntnispsychologisch entwickelt.* (Leipzig: J.A. Barth). Translation by J. Anderson and A.Young as *The Principles of Physical Optics: An Historical and Philosophical Treatment;* reprint (New York: Dover, 1953).
- Meyerson, É. (1925). *La Déduction Relativiste* (Paris: Payot). Translation with supplementary material by D.A. and M.A. Sipfle as *The Relativistic Deduction*. (Dordrecht: D. Reidel, 1985).
- Misner, C.W., and Wheeler, J.A. (1962) "Classical Physics as Geometry" in J.A. Wheeler *Geometrodynamics* (New York and London: Academic Press), 225-307.
- Natorp, P. (1910). *Die Logischen Grundlagen der Exakten Wissenschaften.* (Leipzig: B.G. Teubner).
- Norton, J. (1984). "How Einstein Found his Field Equations", *Historical Studies in the Physical Sciences 14*, 253-316; reprinted in Howard and Stachel (1989), 101-159.
- ----- (1993). "General Covariance and the Foundations of General Relativity: Eight Decades of Dispute", *Reports on Progress in Physics* 56, 791-858.
- Norton, J. (2000). "‘Nature is the Realisation of the Simplest Conceivable Mathematical Ideas’: Einstein and the Canon of Mathematical Simplicity", *Studies in the History and Philosophy of Modern Physics* 31, 135-170.
- Ohanian, H. (1977). "What is the Principle of Equivalence?", *American Journal of Physics* 49, 903-909.
- O’Raifeartaigh, L. (1997). *The Dawning of Gauge Theory.* (Princeton: Princeton University Press).

- Petzoldt, J. (1921). "Der Verhältnis der Machschen Gedankenwelt zur Relativitätstheorie", an appendix to E. Mach, *Die Mechanik in ihrer Entwicklung: Historisch-Kritisch Dargestellt. Achte Auflage*. (Leipzig: Brockhaus), 490-517.
- Poincaré, H. (1906).*The Value of Science* . 1913 translation from the original French edition by G.B. Halsted; reprint edition (New York: Dover Publishing, 1958)
- Rainich, G. (1925) "Electrodynamics in the General Relativity Theory", *Transactions of the American Mathematical Society* 27, 106-136.
- Reichenbach, H. (1920). *Relativitätstheorie und Erkenntnis Apriori.* (Berlin: J. Springer). Translation by M. Reichenbach, *The Theory of Relativity and A Priori Knowledge.* (Berkeley and Los Angeles: University of California Press, 1965).
- ----- (1922). "La signification philosophique de la théorie de la relativité", *Revue Philosophique de la France et de Étranger* 94 (1922), 5-61, translated into French by L. Bloch. English translation, with omissions, by M. Reichenbach in M. Reichenbach and R.S. Cohen (eds.) *Hans Reichenbach Selected Works,* v.2 (Dordrecht: D. Reidel, 1978), 3-47.
- ----- (1924) *Axiomatik der relativistischen Raum-Zeit-Lehre.* (Braunschweig: Vieweg). Translation by M. Reichenbach, *Axiomatization of the Theory of Relativity* (Berkeley and Los Angeles: University of California Press, 1969).
- ----- (1928). *Philosophie der Raum-Zeit Lehre.* (Berlin: Walter de Gruyter). Translation, with omissions, by M.Reichenbach and J. Freund, *The Philosophy of Space and Time.* (New York: Dover, 1958).
- Russell,, B. (1926). "Relativity: Philosophical Consequences." *Encyclopedia Britannica,* 13th ed., v.31,.331-332 (London and New York).
- Ryckman, T. (1994). "Weyl, Reichenbach and the Epistemology of Geometry", *Studies in the History and Philosophy of Modern Physics,* 25,831-870.
- ----- (1996). Einstein *Agonists:* Weyl and Reichenbach on Geometry and the General Theory of Relativity", in R. Giere and A. Richardson (eds.) *The Origins of Logical Empiricism.(Minnesota Studies in the Philosophy of Science,* v.16) (Minneapolis: University of Minnesota Press), 165-209.
- ----- (forthcoming a). "Two Roads from Kant: Cassirer. Reichenbach and General Relativity", in W. Salmon and P. Parrini (eds.), *Analytical and Continental Aspects of Logical Empiricism* (Pittsburgh: University of Pittsburgh Press).
- ----- (forthcoming b). "Logical Empiricism and the Philosophy of Physics", in A. Richardson and T.E. Uebel (eds.), *The Cambridge Companion to Logical Empiricism.* (Cambridge: Cambridge University Press).
- Schlick, M. (1917). "Raum und Zeit in der gegenwärtigen Physik", *Die Naturwissenschaften* 5, 161-167 (16 März), 177-186 (23 März); also appearing as a monograph (Berlin: J. Springer). Augmented *Vierte Auflage, 1922.* (Berlin: J. Springer). Translation by H.L. Brose and P. Heath in H. Mulder, and B. van de Velde-Schlick (eds.) *Moritz Schlick: Philosophical Papers,* v.1 (Dordrecht: D. Reidel, 1979), 207-269.
- Schneider, I. (1921). *Das Raum-Zeit Problem bei Kant und Einstein*. (Berlin: J. Springer).
- Sellien 1919). *Die erkenntnistheoretische Bedeutung der Relativitätstheorie.*, E. (Berlin: Reuther & Reichard) *("Kant-Studien" Ergänzungshefte, 48).*
- Stachel,, J. (1980). "Einstein's Search for General Covariance", as reprinted in Howard and

Stachel (1989), 63-100.
- Torretti, R. (1983). *Relativity and Geometry* (Oxford: Pergamon Press).
- von Laue, M. (1921). *Die Relativitätstheorie. Zweiter Band: Die Allgemeine Relativitätstheorie und Einsteins Lehre von der Schwerkraft.* (Braunschweig: Vieweg).
- Vizgin, V. (1994) *Unified Field Theories in the first third of the 20th Century*. Translated from the Russian by J. B. Barbour (Basel, Boston, Berlin: Birkäuser).
- Weyl, H. (1918a). "Gravitation und Elektrizität", *Sitzungsberichte der preuß. Akad. Berlin, Math. Kl.*,465-80; reprinted in H. Weyl, *Gesammelte Abhandlungen*,Bd..II (Berlin, Heidelberg, New York: Springer Verlag, 1968), 29-42. Translation in O'Raifeartaigh (1997), 24-37.
- ----- (1918b). "Reine Infinitesimalgeometrie", *Mathematische Zeitschrift*, Bd. II, 384-411; reprinted in H. Weyl, *Gesammelte Abhandlungen*, Bd. II (Berlin, Heidelberg, New York: Springer Verlag, 1968), 1-28.
- ----- (1919). *Raum-Zeit-Materie. Dritte Auflage.* (Berlin: J. Springer).
- ----- (1920). "Das Verhältnis der kausalen zur statistischen Betrachtungsweise in derPhysik", *Schweizerische Medizinische Wochenschrift*, Nr. 34, 737-741; reprinted in H. Weyl, *Gesammelte Abhandlungen*, Bd. II (Berlin, Heidelberg, New York: Springer Verlag, 1968), 113-122.
- ----- (1921). "Zur infinitesimalgeometrie: Einordnung der projektiven und der konformen Auffassung", *Nachrichten: Königlichen Gesellschaft der Wissenschaften zu Göttingen, Math.-Phys. Kl.*, 99-112; reprinted in H. Weyl, *Gesammelte Abhandlungen*,Bd..II (Berlin, Heidelberg, New York: Springer Verlag, 1968), 195-207.
- ----- (1923a) *Raum-Zeit-Materie. Fünfte Auflage.* (Berlin: J. Springer).
- ----- (1923b). *Mathematische Analyse des Raumproblems.* (Berlin: J. Springer). Reprint in *Das Kontinuum und andere Monographien.* (New York: Chelsea Publishing Co., n.d.),
- ----- (1929). "Elektron und Gravitation", *Zeitschrift für Physik* 56, 330-352; reprinted in H. Weyl, *Gesammelte Abhandlungen*,Bd..III (Berlin, Heidelberg, New York: Springer Verlag, 1968), 245-267. Translation in O'Raifeartaigh (1997), 121-144.
- Winternitz, J. (1924). *Relativitätstheorie und Erkenntnislehre.* (Leipzig und Berlin: B.G. Teubner).
- Wolters, G. (1987). *Mach I, Mach II, Einstein und die Relativitätstheorie: Eine Fälschung und ihre Folgen*. (Berlin and New York: Walter de Gruyter).

Other Internet Resources

- The History of Philosophy Working Group (U. Missouri/Kansas City)

[Please contact the author with other suggestions.]

Related Entries

a priori justification and knowledge | Einstein, Albert: philosophy of science | equivalence of mass and energy | geometry: in the 19th century | Kant, Immanuel | Reichenbach, Hans | space and time: conventionality of simultaneity | space and time: the hole argument

Thomas A. Ryckman
tryckman@hotmail.com

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 27, 2001
Content last modified: November 27, 2001

Equivalence of Mass and Energy

Einstein correctly described the equivalence of mass and energy as "the most important upshot of the special theory of relativity" (Einstein, 1919), for it is more than a mere curiosity of physics. According to Einstein's famous equation $E = mc^2$, the energy (E) of a body is numerically equal to the product of its mass (m) and the speed of light (c) squared. It is customary to refer to this result as "the equivalence of mass and energy," or simply "mass-energy equivalence," because one can choose units in which $c = 1$, and hence $E = m$. An important consequence of $E = mc^2$ is that a change in the rest-energy of a body is accompanied by a corresponding change to its inertial mass. (This is discussed further in Section 1.) This has led many philosophers to argue that mass-energy equivalence has profound consequences for ontology, the philosophical study of what there is. There are two main philosophical interpretations of $E = mc^2$. The first is that mass-energy equivalence teaches us that "mass" and "energy" designate the same *property* of physical systems. This is the weaker of the two interpretations because no further ontological claims are made. The second interpretation is that $E = mc^2$ entails that there is only one sort of fundamental *stuff* in the world. (This is discussed further in Section 2.) Recently, the history of $E = mc^2$ has also attracted the attention of some philosophers. This is primarily, though not exclusively, because this history shows that $E = mc^2$ is a direct consequence of changes to the structure of spacetime brought about by Special Relativity. (This is discussed further in Section 3.)

1. Mass-Energy Equivalence: The Result

The equation $E=mc^2$ has two distinct physical consequences. To see this, one needs first to distinguish

between a body's rest-mass and its relativistic mass. In Newtonian physics, the inertial mass of a body is a measure of that body's resistance to acceleration. This is the notion of mass one uses in everyday life when one talks about, for example, 1 kg. of salt. Furthermore, in Newtonian physics, the inertial mass of a body is independent of its relative state of motion. Because this is no longer the case in relativistic physics, one can identify two notions of mass in Special Relativity (SR). The *rest-mass* of a body is the inertial mass of that body when it is at rest relative to an inertial frame. The term m in the equation $E=mc^2$ does *not* represent rest-mass; it represents *relativistic mass*, which is the inertial mass of a body when it is in a state of motion relative to an inertial frame. If we use m_o to designate the rest-mass of a body, then we can re-write Einstein's equation in the following way:

$$E = mc^2 = m_o \gamma(v) c^2, \text{ where } \gamma(v) = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} \text{ is the "Lorentz Factor".}$$

These equations entail that:

> **(I)** In the frame of reference in which a body is at rest, its energy (in this case called the *rest-energy*) is equal to the product of its rest-mass m_o and the speed of light squared. This is because in this case $v = 0$, so the Lorentz factor is one.
>
> **(II)** In a frame of reference in which a body moves with velocity v, the energy of the body is equal to the product of its rest-mass, the speed of light squared, and the Lorentz factor.

From **(I)** it follows that if there is a change in the rest-energy of a body, there must be a corresponding change in its rest-mass. For example, if a body is heated, and thereby absorbs a small amount of energy ΔE (as measured in the frame of reference in which the body is at rest), its rest-mass will increase by a very small amount equal to $\Delta E/c^2$. This increase is tiny because of the high numerical value of the speed of light. Indeed, for mid-sized objects, such an increase in mass would be too small to measure with even the most accurate balance. For example, if a 1 kg block of gold is heated so that its temperature increases by 10 °C, then its mass could increase by as much as 1.4×10^{-14} kg; a cube of gold of this additional tiny mass would have sides smaller than one one-thousandths of a millimeter. Similarly, if a body emits an amount of energy ΔE, say in the form of light or heat, its rest-mass will decrease by a tiny amount $\Delta E/c^2$. In both cases, the important and novel claim made by SR is that the inertial mass of a body can change depending on whether it absorbs or emits energy.

(I) also entails that there are physical interactions in which masses no longer combine by simple addition, as they do in pre-relativistic physics. For example, suppose two bodies *A* and *B* collide to produce a single, more massive body *AB*. Suppose further that a net amount of energy *E* is emitted in this inelastic collision, say in the form of heat. **(I)** entails that the rest-mass of *AB* will be *less* than the rest-mass of *A* plus the rest-mass of *B* by an amount equal to E/c^2. This stands in sharp contrast to the pre-relativistic

prediction that the rest-mass of *AB* will be *equal* to the rest-mass of *A* plus the rest-mass of *B*. So, for example, suppose a meteor (*A*) struck the earth (*B*). After the crash, the earth (*AB*) would have a mass that is a tiny bit less than the mass of the meteor plus the mass of the earth prior to the crash. This is because during the collision the meteor loses part of its kinetic energy as heat radiation. This energy loss corresponds to a loss of mass. It is worth emphasizing that, according to SR, it is the *inertial* mass of bodies that is no longer simply added in collisions such as these. In other words, SR predicts that the resulting body *AB* will resist acceleration a tiny bit less than one would have predicted according to pre-relativistic physics. There is an analogous result for cases where a single body disintegrates into two or more bodies.

These consequences of **(I)** also illustrate how the classical conservation principles are modified by SR. According to Newtonian physics, all physical interactions are separately governed by the principles of conservation of mass and conservation of energy. So, for example, according to pre-relativistic physics the mass of the block of gold discussed above must remain the same as it is heated. However, as we have seen, this is not the case in relativistic physics, because the energy absorbed by the block of gold contributes to an increase in its rest-mass. Similarly, Newtonian physics predicts that mass is conserved when the meteor crashes into the earth in the above example. However, according to relativistic physics, some of the mass is radiated away as energy in the form of heat. In both of these examples, it is the total mass *and* energy of the entire system that is conserved in these interactions. In general, in SR physical interactions no longer satisfy the two classical conservation principles separately. Instead, these two principles are fused into a single principle: the principle of conservation of mass-energy. It is these consequences of **(I)**, and indirectly the fusing of the two classical conservation principles, that have motivated different philosophical interpretations of $E=mc^2$ (see Section 2, Philosophical Interpretations of Mass-Energy Equivalence).

From **(II)** it follows that no bounded amount of energy is sufficient to accelerate a body to the speed of light. This is because as the speed of a body approaches the speed of light its relativistic mass increases without bound. But this means that the body's resistance to acceleration, as measured in the inertial frame relative to which it is moving, also increases without bound. In practice, this means that it takes more and more energy to achieve proportionally smaller increases in the speed of a body. For example, suppose an electron requires an amount of energy E to reach 50% of the speed of light. The electron requires twice that amount of energy to reach 90% of the speed of light, roughly six times E to reach 99% of the speed of light, and nearly two hundred times E to reach 99.999% of the speed of light! This consequence of $E=mc^2$ is thus crucial in the design and operation of particle accelerators, and it is often emphasized in the popular media (e.g., in popular science books and films). However, its philosophical import is relatively minor because the increase in relativistic mass does not result in a change to the body. In the frame of reference in which the body is at rest, its inertial mass continues to be m_o.

A common misconception surrounding $E = mc^2$ is that it entails that the entire rest-mass of a body can become energy. Strictly speaking, mass-energy equivalence only entails that a *change* in the rest-energy of a body is invariably accompanied by a corresponding *change* in the rest-mass of the body. For example, a body may lose a bit of its mass because it radiates a bit of energy. The stronger claim that a

body may lose *all* of its rest-mass as it radiates energy is *not* a consequence of SR. However, this stronger claim is very well confirmed by experiments in atomic physics. Many particle-antiparticle collisions have been observed, such as collisions between electrons and positrons, where the entire mass of the particles is radiated away as energy in the form of light. Nevertheless, SR leaves open the possibility that a form of matter exists whose mass cannot become energy. This is significant because it emphasizes that mass-energy equivalence is not a consequence of a theory of matter; it is instead a direct consequence of changes to the structure of spacetime imposed by SR (see Section 3, Derivations of Mass-Energy Equivalence: History).

2. Philosophical Interpretations of Mass-Energy Equivalence

Philosophical interpretations of mass-energy equivalence can be classified into two main groups. Interpretations in the first group, such as Eddington's (1929), and more recently Torretti's (1983), regard the terms "mass" and "energy" as designating *properties* of physical systems. According to these interpretations, $E = mc^2$ teaches us that properties hitherto regarded as distinct are actually the same. For example, Eddington states that "it seems very probable that mass and energy are two ways of measuring what is essentially the same thing, in the same sense that the parallax and distance of a star are two ways of expressing the same property of location" (1929, p. 146). According to Eddington, the distinction between mass and energy is artificial. We treat mass and energy as different properties of physical systems because we routinely measure them using different units. However, one can measure mass and energy using the *same* units by choosing units in which c =1, i.e., units in which distances are measured in units of time (e.g., light-years). Once we do this, Eddington claims, the distinction between mass and energy disappears.

Torretti (1983) argues along similar lines when he responds to the opposing view, which is held by a minority (e.g., Bunge, 1967; Sachs, 1981). This minority holds that the numerical equivalence of mass and energy is not sufficient to conclude that the two properties are the same. However, according to Torretti, "If a kitchen refrigerator can extract mass from a given jug of water and transfer it by heat radiation or convection to the kitchen wall behind it, a trenchant metaphysical distinction between the mass and the energy of matter does seem far fetched" (1983, p. 307, fn. 13). Like Eddington, Torretti points out mass and energy seem to be different properties because they are measured in different units. But the units of mass and energy are different only if one uses different units for space and time, which one need not do. For Torretti, the apparent difference between mass and energy is thus an illusion that arises from "the convenient but deceitful act of the mind by which we abstract time and space from nature" (1983, p. 307, fn. 13). Interpretations such as Torretti's and Eddington's draw no further ontological conclusions from mass-energy equivalence beyond the claim that the two properties are the same. For example, neither Eddington nor Torretti make any explicit claim concerning whether properties are best understood as universals, or whether one ought to be a realist about such properties.

Interpretations in the second group, such as Zahar's (1984), Einstein and Infeld's (1938), and Russell's

(1915, 1948), regard "mass" as a measure of the quantity of matter and "energy" as a measure of the quantity of energy. Thus, these interpretations regard "mass" and "energy" as the representatives, within physical theory, of stuff in the world. According to these interpretations, the philosophical lesson of $E = mc^2$ is that we should no longer regard the world as consisting of two types of stuff: matter and energy. Instead, the world is composed of only one type of fundamental stuff. For example, According to Zahar, energy in pre-relativistic physics occupies a distinct "ontological level" from matter primarily because the former is regarded as dependent on the latter, but not *vice versa*. In relativistic physics, however, Einstein's famous equation shows that these two ontological levels are in fact identical. According to Zahar, Einstein showed "that ‘energy’ and ‘mass’ *could* be treated as two names for the same basic entity. The stuff which appears to the senses as hard extended substance and the quantity of energy which characterises a process are in fact one and the same thing" (1989, p. 262). For Zahar, the apparent difference between mass and energy arises from the contingent fact that our senses perceive mass and energy differently. On this reading, mass-energy equivalence has the metaphysical implication that what is real, "is no longer the familiar hard substance but a new entity which can be interchangeably called matter or energy" (1989, p. 263). Thus, Zahar holds that the fundamental stuff of physics is a sort of "I-know-not-what" that we can call either "mass" or "energy."

Einstein and Infeld (1938) hold a slightly different version of this interpretation. They claim that mass-energy equivalence implies that we can no longer distinguish between "matter" and "the field". Einstein and Infeld (1938) argue that in pre-relativistic physics there are physical criteria for distinguishing matter and field since matter has mass but fields do not. Hence, there is a qualitative difference between matter and fields, and so it is reasonable to adopt an ontology containing both. However, the equivalence of mass and energy entails that "matter represents vast stores of energy and that energy represents matter" (1938, p. 242). Consequently, Einstein and Infeld argue, the distinction between matter and field is no longer a qualitative one in relativistic physics. Instead, it is merely a quantitative difference, since "matter is where the concentration of energy is great, field where the concentration of energy is small"(1938, p. 242). Thus, mass-energy equivalence entails that we should adopt an ontology consisting only of fields.

Among philosophers, Russell interprets mass-energy equivalence in a similar fashion to Einstein and Infeld. According to Russell, "a unit of matter tends more and more to be something like an electromagnetic field filling all space, though having its intensity in a small region" (1915, p. 121). In his later work, Russell continues to hold this view. For example, in *Human Knowledge, Its Scope and Limits*, he points out that "atoms" are merely small regions in which there is a great deal of energy. Furthermore, these regions are precisely the regions where one would have said, in pre-relativistic physics, that there was matter. For Russell, these considerations suggest that "mass is only a form of energy, and there is no reason why matter should not be dissolved into other forms of energy. It is energy, not matter, that is fundamental in physics" (1948, p. 291). Russell is not claiming, as Zahar does, that there is one unknown type of stuff that we can call either "mass" or "energy". Instead, Russell is proposing that mass is *reducible* to energy in the sense that the world consists only of energy. Thus, for Russell, "mass" and "matter" are otiose in modern physics. Several physicists have held a similar position, though this view is less common now. For example, after a discussion particle-antiparticle annihilation experiments in 1951, Wolfgang Pauli states: "Taking the existence of all these transmutations into account, what remains of the old idea of matter and of substance? The answer is energy. This is the true substance, that which is

conserved; only the form in which it appears is changing" (1951, p. 31).

Despite the marked difference in the ontological claims made by the two groups of interpretations, there is one significant similarity. All interpretations implicitly claim that mass-energy equivalence changes our knowledge concerning the extensions of the concepts "mass" and "energy". In pre-relativistic physics, the terms "mass" and "energy" had different extensions and intensions. Relativistic physics teaches us that the extension of the two terms is actually the same. This is analogous to the discovery that the referents of "the morning star" and "the evening star" are the same. We can push the analogy a bit further. Just as it is possible to verify empirically that the planet Venus is the referent of both "the morning star" and "the evening star," it is possible to verify empirically that the extensions of "mass" and "energy" are the same. (See Section 4, Experimental Verification of Mass-Energy Equivalence.) From this perspective, the various interpretations of mass-energy equivalence disagree only about what kinds of things the terms "mass" and "energy" designate.

3. Derivations of Mass-Energy Equivalence: History

Einstein first derived mass-energy equivalence from the principles of SR in a small article titled "Does the Inertia of a Body Depend Upon Its Energy Content?" (1905b). This derivation, along with others that followed soon after (e.g., Plank (1906), Von Lau (1911)), uses Maxwell's theory of electromagnetism. (See Subsection 3.1, Derivations of $E = mc^2$ that Use Maxwell's Theory.) However, as Einstein later observed (1935), mass-energy equivalence is a result that should be independent of any theory that describes a specific physical interaction. This is the main reason that led physicists to search for "purely dynamical" derivations, i.e., derivations that invoke only mechanical concepts such as "energy" and "momentum", and the principles that govern them. (See Subsection 3.2, Purely Dynamical Derivations of $E = mc^2$.)

3.1 Derivations of $E = mc^2$ that Use Maxwell's Theory

Einstein's original derivation of mass-energy equivalence is the best known in this group. Einstein begins with the following thought-experiment: a body at rest (in some inertial frame) emits two pulses of light of equal energy in opposite directions. Einstein then analyzes this "act of emission" from another inertial frame, which is in a state of uniform motion relative to the first. In this analysis, Einstein uses Maxwell's theory of electromagnetism to calculate the physical properties of the light pulses (such as their intensity) in the second inertial frame. By comparing the two descriptions of the "act of emission", Einstein arrives at his celebrated result: "the mass of a body is a measure of its energy-content; if the energy changes by L, the mass changes in the same sense by $L/9 \times 10^{20}$, the energy being measured in ergs, and the mass in grammes" (1905b, p. 71). A similar derivation using the same thought experiment but appealing to the Doppler effect was given by Langevin (1913) (see the discussion of $E = mc^2$ in Fox (1965)).

Some philosophers and historians of science claim that Einstein's first derivation is fallacious. For example, in *The Concept of Mass*, Jammer says: "It is a curious incident in the history of scientific

thought that Einstein's own derivation of the formula $E = mc^2$, as published in his article in *Annalen der Physik*, was basically fallacious. . . the result of a *petitio principii*, the conclusion begging the question" (Jammer, 1961, p. 177). According to Jammer, Einstein implicitly assumes what he is trying to prove, viz., that if a body emits an amount of energy L, its inertial mass will decrease by an amount $\Delta m = L/c^2$. Jammer also accuses Einstein of assuming the expression for the relativistic kinetic energy of a body. If Einstein made these assumptions, he would be guilty of begging the question. Recently, however, Stachel and Torretti (1982) have shown convincingly that Einstein's (1905b) argument is sound. They note that Einstein indeed derives the expression for the kinetic energy of an "electron" (i.e., a structureless particle with a net charge) in his earlier (1905a) paper. However, Einstein nowhere uses this expression in the (1905b) derivation of mass-energy equivalence. Stachel and Torretti also show that Einstein's critics overlook two key moves that are sufficient to make Einstein's derivation sound, since one need not assume that $\Delta m = L/c^2$.

Einstein's further conclusion that "the mass of a body is a measure of its energy content" (1905b, p. 71) does not, strictly speaking, follow from his argument. As Torretti (1983) and other philosophers and physicists have observed, Einstein's (1905b) argument allows for the possibility that once a body's energy store has been entirely used up (and subtracted from the mass using the mass-energy equivalence relation) the remainder is not zero. In other words, it is only an hypothesis in Einstein's (1905b) argument, and indeed in all derivations of $E = mc^2$ in SR, that no "exotic matter" exists that is *not* convertible into energy (see Ehlers, Rindler, Penrose, (1965) for a discussion of this point). However, particle-antiparticle anihilation experiments in atomic physics, which were first observed decades after 1905, strongly support "Einstein's dauntless extrapolation" (Torretti, 1983, p. 112).

3.2 Purely Dynamical Derivations of $E = mc^2$

Purely dynamical derivations of $E = mc^2$ typically proceed by analyzing an inelastic collision from the point of view of two inertial frames in a state of relative motion (the centre-of-mass frame, and an inertial frame moving with a relative velocity v). One of the first papers to appear following this approach is Perrin's (1932). According to Rindler and Penrose (1965), Perrin's derivation was based largely on Langevin's "elegant" lectures, which were delivered at the College de France in Zurich around 1922. Einstein himself gave a purely dynamical derivation (Einstein, 1935), though he nowhere mentions either Langevin or Perrin. The most comprehensive derivation of this sort was given by Ehlers, Rindler and Penrose (1965). More recently, a purely dynamical version of Einstein's original (1905b) thought experiment, where the particles that are emitted are not photons, has been given by Mermin and Feigenbaum (1990).

Derivations in this group are distinctive because they demonstrate that mass-energy equivalence is a consequence of the changes to the structure of spacetime brought about by SR. The relationship between mass and energy is independent of Maxwell's theory or any other theory that describes a specific physical interaction.We can get a glimpse of this by noting that to derive $E = mc^2$ by analyzing a collision, one must first define relativistic momentum ($\mathbf{p}_{rel}$) and relativistic kinetic energy (T_{rel}), since one cannot use

the old Newtonian notions of momentum and kinetic energy. In Einstein's own purely dynamical derivation (1935), more than half of the paper is devoted to finding the mathematical expressions that define $\mathbf{p}_{rel}$ and T_{rel}. This much work is required to arrive at these expressions for two reasons. First, the changes to the structure of spacetime must be incorporated into the definitions of the relativistic quantities. Second, $\mathbf{p}_{rel}$ and T_{rel} must be defined so that they reduce to their Newtonian counterparts in the appropriate limit. This last requirement ensures, in effect, that SR will inherit the empirical success of Newtonian physics. Once the definitions of $\mathbf{p}_{rel}$ and T_{rel} are obtained, the derivation of mass-energy equivalence is straight-forward. (For a more detailed discussion of Einstein's (1935), see Flores, (1998).)

4. Experimental Verification of Mass-Energy Equivalence

Cockcroft and Walton (1932) are routinely credited with the first experimental verification of mass-energy equivalence. Einstein (1905b) had conjectured that the equivalence of mass and energy could be tested by "weighing" an atom before and after it undergoes radioactive decay. But there was no way of performing this experiment or another experiment that would directly confirm mass-energy equivalence at the time. Technological developments allowed Cockcroft and Walton to take a different approach. They studied the bombardment of a lithium atom (Li) by a proton (p), which produces two alpha particles (α). This reaction is symbolized by the following equation:

$$p + Li = 2\alpha$$

In this reaction, there is a *decrease* in the total rest-mass as the reaction proceeds from left to right: the total rest-mass of proton and the Lithium atom is greater than the total rest-mass of the two alpha particles. Furthermore, there is also an *increase* in the total kinetic energy: the kinetic energy of the proton is less than the total kinetic energy of the two alpha particles. (One only considers the kinetic energy of the proton because the Lithium atom is considered at rest, and hence has zero kinetic energy.) Cockcroft and Walton were able to measure the kinetic energies of the incident proton and the out-going alpha particles very precisely. They found that the decrease in rest-mass corresponds to the increase in kinetic energy according to Einstein's famous equation $E = mc^2$ (to an accuracy of better than 1%). Hence, the total mass *and* energy of the entire system is conserved.

Bibliography

- Anderson, J. L. (1967) *Principles of Relativity Physics*, New York: Academic Press.
- Bunge, M. (1967) *Foundations of Physics*, New York: Springer-Verlag.
- Cockcroft J. D. and E. T. Walton (1932) *Proc. Roy. Soc. London*, **A137**:229.
- Eddington, A. (1929) *Space, Time, and Gravitation*, London: Cambridge University Press,

originally published in 1920.

- Ehlers, J., W. Rindler, and R. Penrose (1965) "Energy Conservation as the Basis of Relativistic Mechanics II," *Am. J. Phys.* **35**:995-997.
- Einstein, A. (1905a) "On the Electrodynamics of Moving Bodies" in A. Einstein *et al.* (1952):35-65.
- Einstein, A. (1905b) "Does the Inertia of a Body Depend Upon Its Energy Content?" in A. Einstein *et al.* (1952):69-71.
- Einstein, A. (1919) "What is the Theory of Relativity?" in A. Einstein (1982):227-232.
- Einstein, A. (1935) "Elementary Derivation of the Equivalence of Mass and Energy," *Am. Math. Soc. Bul.* **41**:223-230.
- Einstein, A., H.A. Lorentz, H. Minkowski and H. Weyl (1952) *The Principle of Relativity*, W. Perrett and G.B. Jeffery (trans.), New York: Dover.
- Einstein, A. (1982) *Ideas and Opinions*, New York: Crown Publishers Inc.
- Einstein, A. and L. Infeld (1938) *The Evolution of Physics*, New York: Simon and Schuster.
- Feynman, R. P., R.B. Leighton, and M.L. Sands (1989) *The Feynman Lectures in Physics*, Redwood City, CA: Addison-Wesley Pub. Co.
- Flores, F. (1998) "Einstein's 1935 Derivation of $E = mc^2$," *Stud. Hist. Phil. Mod. Phys.* **29**(2):223-243.
- Fox, J. G. (1965) "Evidence Against Emission Theories," *Am. J. Phys.* **33**(1):1-17.
- Jammer, M. (1961) *Concepts of Mass in Classical and Modern Physics*, Cambridge, MA: Harvard University Press.
- Mermin D. and J. Feigenbaum (1990) "$E = mc^2$" in *Boojums All the Way Through*, New York: Cambridge University Press.
- Pauli, W. (1951) "Matter" in *Wolfgang Pauli: Writings on Physics and Philosophy*, C.P. Enz and K. von Meyenn (eds.), New York: Springer-Verlag (1994).
- Perrin, F. (1932) "La Dynamique Relativiste et l'Inertie de l'Energie," *Actualities Scientifiques et Industrielles*, Serie 1932, Vol. XLI.
- Rindler, W. and R. Penrose (1965) "Energy Conservation as the Basis of Relativistic Mechanics," *Am. J. Phys.* **35**:55-59.
- Russell, B. (1915) "The Ultimate Constituents of Matter" in *Mysticism and Logic*, Montreal: Pelican Press, (1953).
- Russell, B. (1948) *Human Knowledge, Its Scope and Limits*, New York: Simon and Schuster.
- Sachs, M. (1981) *Ideas of Matter*, Washington D.C.: University Press of America.
- Stachel, J. and R. Torretti (1982) "Einstein's First Derivation of Mass-Energy Equivalence," *Am. J. Phys.* **50**(8):760-761.
- Taylor, E.F. and J.A. Wheeler (1966) *Spacetime Physics*, San Francisco, CA: W. H. Freeman.
- Torretti, R. (1983) *Relativity and Geometry*, New York: Dover (1996).
- Zahar, E. (1989) *Einstein's Revolution: A Study in Heuristic*, La Salle, Il: Open Court.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

ontology | space and time: inertial frames

Francisco Flores
fflores@calpoly.edu

A| B| C| D| E| F| G| H| I| J| K| L| M| N| O| P| Q| R| S| T| U| V| W| X| Y| Z

Table of Contents

First published: September 12, 2001
Content last modified: September 12, 2001

Space and Time: Inertial Frames

A “frame of reference” is a standard relative to which motion and rest may be measured; any set of points or objects that are at rest relative to one another enables us, in principle, to describe the relative motions of bodies. A frame of reference is therefore a purely kinematical device, for the geometrical description of motion without regard to the masses or forces involved. A dynamical account of motion leads to the idea of an “inertial frame,” or a reference frame relative to which motions have distinguished dynamical properties. For that reason an inertial frame has to be understood as a spatial reference frame together with some means of measuring time, so that uniform motions can be distinguished from accelerated motions. The laws of Newtonian dynamics provide a simple definition: an inertial frame is a reference-frame with a time-scale, relative to which the motion of a body not subject to forces is always rectilinear and uniform, accelerations are always proportional to and in the direction of applied forces, and applied forces are always met with equal and opposite reactions. It follows that, in an inertial frame, the center of mass of a system of bodies is always at rest or in uniform motion. It also follows that any other frame of reference moving uniformly relative to an inertial frame is also an inertial frame. For example, in Newtonian celestial mechanics, taking the “fixed stars” as a frame of reference, we can determine an (approximately) inertial frame whose center is the center of mass of the solar system; relative to this frame, every acceleration of every planet can be accounted for (approximately) as a gravitational interaction with some other planet in accord with Newton’s laws of motion.

This appears to be a simple and straightforward concept. By inquiring more narrowly into its origins and meaning, however, we begin to understand why it has been an ongoing subject of philosophical concern. It originated in a profound philosophical consideration of the principles of relativity and invariance in the context of Newtonian mechanics. Further reflections on it, in different theoretical contexts, had extraordinary consequences for 20th-century theories of space and time.

---

1. Relativity and Reference Frames in Classical Mechanics

1.1 The Origins of Galilean Relativity

The term “reference frame” was coined in the 19th century, but it has a long prehistory, beginning, perhaps, with the emergence of the Copernican theory. The significant point was not the replacement of the earth by the sun as the center of all motion in the universe, but the recognition of both the earth and the sun as merely possible points of view from which the motions of the celestial bodies may be described. This implied that the basic task of Ptolemaic astronomy -- to represent the planetary motions by combinations of circular motions -- could take any point to be fixed, and that, as Copernicus suggested in the opening arguments of “On the revolutions of the heavenly spheres,” the choice of any particular point required some justification on other than astronomical grounds. As the basic programme of Ptolemy and Copernicus gave way to that of early classical mechanics, this equivalence of points of view was made more precise and explicit. Galileo demonstrated that the Copernican view does not contradict our experience of a seemingly stable earth, through a principle that, in the precise form that it takes in Newtonian mechanics, has become known as the “principle of Galilean relativity”: mechanical experiments will have the same results in a system in uniform motion that they have in a system at rest. Therefore the experiments claimed as evidence against Copernicus -- e.g., that a stone dropped from a tower falls to the base of the tower, instead of being left behind -- would happen just as they do whether the earth were moving or not, provided that the motion is sufficiently uniform. See Figure 1.

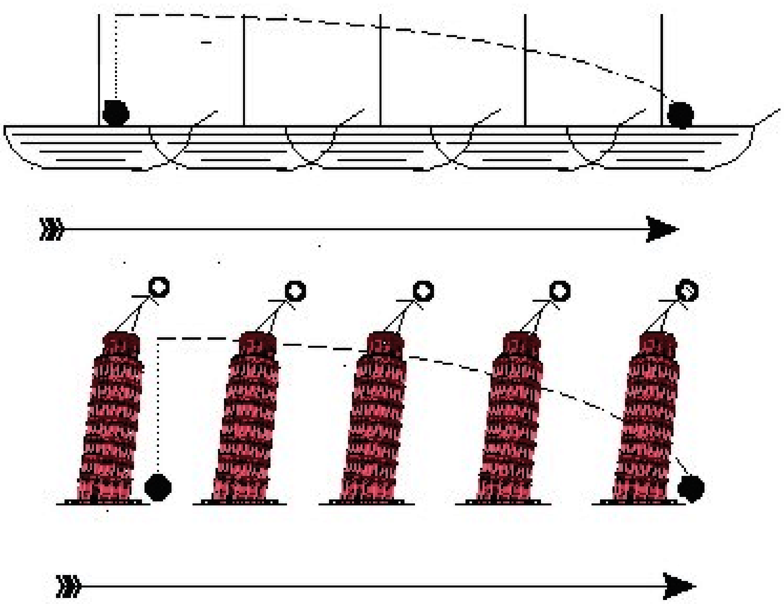


Figure 1: Galileo’s Argument
If the earth is rotating sufficiently uniformly, a stone dropped from the tower will fall straight to the base, just as a stone dropped from the mast of a uniformly moving ship will fall to the foot of the mast. In both cases the stone’s vertical motion will be smoothly composed with its horizontal motion. Hence a sufficiently uniform motion will be indistinguishable from rest.

1.2 Philosophical Controversy Over Absolute and Relative Motion

Leibniz, later, articulated a more general “equipollence of hypotheses”: in any system of interacting bodies, any hypothesis that any particular body is at rest is equivalent to any other. Therefore neither Copernicus’ nor Ptolemy’s view can be true -- though one may be judged simpler than the other -- because both are merely possible hypothetical interpretations of the same relative motions. This principle clearly defines (what we would call) a set of reference frames, differing in their arbitrary choices of a resting point or origin, but agreeing on the relative positions of bodies at any moment and their changing relative distances through time.

For Leibniz and many others, this general equivalence was a matter of philosophical principle, founded in the metaphysical conviction that space itself is nothing more than an abstraction from the geometrical relations among bodies. In some form or other it was a widely shared tenet of the 17th-century “mechanical philosophy”. Yet it was flatly incompatible with physics as Leibniz himself, and the other “mechanists,” actually conceived it. For the basic program of mechanical explanation depended essentially on the concept of

a privileged state of motion, as expressed in the assumption that bodies maintain a state of rectilinear motion until acted upon by an external cause. Thus their fundamental conception of force, as the power of a body to change the state of another, likewise depended on this notion of a privileged state. This dependence was clearly exhibited in the vortex theory of planetary motion, in which every orbit was explained by the balance between the planet's inherent centrifugal tendency (its tendency to follow the tangent to the orbit) and the pressure of the surrounding medium.

For this reason, the notion of a dispute between "relativists" or "relationists" and "absolutists" or "substantivalists", in the 17th century, is a drastic oversimplification. Newton, in his controversial Scholium on space, time, and motion, was not merely asserting that motion is absolute in the face of the mechanists' relativist view; he was arguing that a conception of absolute motion was already implicit in the views of his opponents -- that it was implicit in their conception, which he largely shared, of physical cause and effect. The general equivalence of reference-frames was implicitly denied by a physics that understood forces as powers to change the states of motion of bodies.

Newton therefore held that physics required the conception of absolute space, a distinguished frame of reference relative to which bodies could be said to be truly moving or truly at rest. Assuming, as both Newton and Leibniz did, that states of motion could be distinguished by their causes and effects, the distinguished status of this frame of reference is physically well founded -- and metaphysically well-founded for a metaphysics that, like Newton's or Leibniz's, takes force to be a well-founded notion. On Leibniz's conception of force, in particular, a given force is required to generate *or to maintain* a given velocity -- for objects "passively" resist motion, but maintain their states of motion only by "active" force -- so that, on dynamical grounds, "every body truly does have a certain amount of motion, or, if you will, force." This implies that there is in principle a distinguished frame of reference in which the velocities of bodies correspond to their true velocities, i.e. to the amounts of moving force that they truly possess, and it implies that in any frame that is in motion relative to this one, bodies will not have their true velocities. In short, such a conception of force, if it could be applied physically, would give a precise physical application of Newton's conception of absolute space.

1.3 Galilean Relativity in Newtonian Physics

The difficulty with Newton's view of absolute space comes from the Newtonian conception of force. If force is defined and measured solely by the power to accelerate a body, then obviously the effects of forces -- in short, the causal interactions within a system of bodies -- will be independent of the velocity of the system in which they are measured. So the existence of a set of equivalent "inertial frames" is imposed from the start by Newton's laws. Suppose that we determine for the bodies in a given frame of reference -- say, the rest frame of the fixed stars -- that all observable accelerations are proportional to forces impressed by bodies within the system, by equal and opposite actions and reactions among those bodies. Then we know that these physical interactions will be the same in any frame of reference that is in uniform rectilinear motion relative to the first one. Therefore no Newtonian experiment will be able to determine the velocity of a body, or system of bodies, relative to absolute space. In other words, there is no way to distinguish absolute space itself from any frame of reference that is in uniform motion relative to it. Newton thought that a coherent account of force and motion requires a background space consisting of "places" that "all keep given positions in relation to one another from infinity to infinity" (1726, p. 412). But the laws of motion enable us to determine an infinity of such spaces, all in uniform rectilinear motion relative to each other, and furnish no way of singling out any one as "immovable space."

Oddly enough, no one in the 17th century, or even before the late 19th century, expressed this equivalence of reference-frames more clearly than Newton himself. Newton explicitly derived it from the laws of motion as Corollary V:

> **Corollary V**:
> When bodies are enclosed in a given space, their motions in relation to one another are the same whether the space is at rest or whether it is moving uniformly straight forward without circular motion. (1726, p. 423.)

This is the first clear statement of the Galilean relativity principle. It implied that the dispute between the heliocentric and geocentric views of the universe was mistakenly framed: the proper question about "the system of the world" was not "which body is at rest in the center?" but "where is the center of gravity of the system, and which body is closest to it?" For in a system of orbiting bodies, only their common center of gravity will be unaccelerated, and by Corollary V, the motions of the bodies in the system will be the same, whether its center of gravity is at rest or in uniform rectilinear motion. The system is indeed approximately Keplerian, since the sun has by far the greatest mass and is therefore little disturbed from the center of gravity, which is therefore very close to the common focus of the approximately Keplerian ellipses in which the planets orbit the sun. But by Corollary V, the nearly-Keplerian structure of the system is completely independent of the system's state of motion in absolute space.

The Galilean relativity principle thus expressed the insight that different states of uniform motion, or different uniformly-moving frames of reference, determine only different points of view on the same physically objective quantities, namely force, mass, and acceleration. We can see this insight expressed more explicitly in Newton's understanding of inertia. For Leibniz (among others) , as we saw, moving force, the power of a body to change the motion of another, was determined by velocity. It was therefore seen as an active power, fundamentally different from the passive power of a resting body to resist any change of position. Newton, in contrast, understood the "force of inertia" as a Galilei-invariant quantity:

> [A] body exerts this force only during a change of its state, caused by another force impressed upon it, and the exercise of this force is, depending on viewpoint, both resistance and impetus: resistance in so far as the body, in order to maintain its state, strives against the impressed force, and impetus in so far as the same body, yielding only with difficulty to the force of a resisting obstacle, endeavors to change the state of that obstacle. Resistance is commonly attributed to resting bodies and impetus to moving bodies; but motion and rest, in the popular sense of the term, are distinguished from each other only by point of view, and bodies commonly regarded as being at rest are not always truly at rest. (1726, p. 404-05.)

Newton thus recognized the powers distinguished by Leibniz as the same thing seen from different points of view.

1.4 The Lingering Problem of Absolute Space

Newton understood the Galilean principle of relativity with a degree of depth and clarity that eluded most of his "relativist" contemporaries. It may seem bizarre, therefore, that the notion of inertial frame did not emerge until more than a century and a half after his death. He had identified a distinguished class of dynamically equivalent "relative spaces," in any of which true forces and masses, accelerations and rotations, would have the same objectively measured values. Yet these spaces, though empirically indistinguishable, were not equivalent in principle; evidently Newton conceived them as moving with various velocities in absolute space, though those velocities could not be known. Why should not he, or someone, have recognized the equivalence of these spaces immediately?

This is not the place for an adequate answer to this question, if indeed one is possible. For much of the 20th century, the accepted answer was that of Ernst Mach: Newton lived in an age "deficient in epistemological critique," and so was unable to draw the conclusion that these *empirically* indistinguishable spaces must be equivalent in *every* meaningful sense, so that no one of them deserves even in principle to be designated as "absolute space." Yet even those whom the 20th century credited with more sophisticated epistemological views, such as Leibniz, evidently had difficulties understanding force and inertia in a Galilei-invariant way, despite a philosophical commitment to relativity. Perhaps it suffices to say that to abandon the intuitive association of force or motion with velocity in space, and to accept an equivalence-class structure as the fundamental spatiotemporal framework, requires a level of abstraction that became possible only with the extraordinary development of mathematics, especially of a more abstract view of geometry, that took place in the 19th century. (See geometry: in the 19th century.) In the 17th century only Christiaan Huygens came close to expressing such a view; he held that not velocity, but velocity-difference, was the fundamental dynamical quantity. He therefore understood, for example, that the "absoluteness" of rotation had nothing to do with velocity relative to absolute space, but arose from the difference of velocity among different parts of a rotating body -- a difference which would, evidently, be the same irrespective of the velocity of the body as a whole in absolute space. But of this Huygens gave only the merest suggestion, in manuscripts that remained unpublished for two centuries. (See Stein 1977.) The concept of inertial frame therefore emerged only in the late 19th century, when, as we shall see, it did not seem to be of any great immediate importance.

1.5 19th-Century Analyses of the Law of Inertia

The development of this concept began with a renewed critical analysis of the notion of absolute space, for reasons not anticipated by Newton's contemporary critics. Its starting point was a critical questions about the law of inertia: relative to what is the motion of a free particle uniform and rectilinear? If the answer is "absolute space," then the law would appear to be something other than an empirical claim, for no one can observe the trajectory of a particle relative to absolute space. Two quite different answers to the question were offered in 1870, in the form of revised statements of the law of inertia. Carl Neumann proposed that when we state the law, we must suppose that there is a body somewhere in the universe -- the "body Alpha" -- with respect to which the motion of a free particle is rectilinear, and that there is a time-scale somewhere relative to which it is uniform (Neumann 1870). Ernst Mach (1883) claimed that the law of inertia, and Newton's laws generally, implicitly appeal to the fixed stars as a spatial reference-frame, and to the

rotation of the earth as a time-scale; at least, he held, such is the basis for any genuine empirical content that the laws have. The notion of absolute space, it followed, was only an unwarranted abstraction from the practice of measuring motions relative to the fixed stars.

Mach's proposal had the advantage of a clear empirical motivation; Neumann's "body Alpha" seemed no less mysterious than absolute space, and almost sounds comical to the modern reader. But Neumann's discussion of a time-scale was somewhat more fruitful. He noted that the law of inertia defines a time-scale: equal intervals of time are those in which a free particle travels equal distances. He also noted, however, that this definition is quite arbitrary. For, in the absence of a prior definition of equal times, any motion whatever can be stipulated to be uniform. It is no help to appeal to the requirement of freedom from external forces, since the free particles presumably are known to us only by their uniform motion. We have a genuine empirical claim only when we state of *at least two* free particles that their motions are *mutually* proportional; equal intervals of time can then be defined as those in which two free particles travel mutually proportional distances.

1.6. The Emergence of the Concept of Inertial Frame

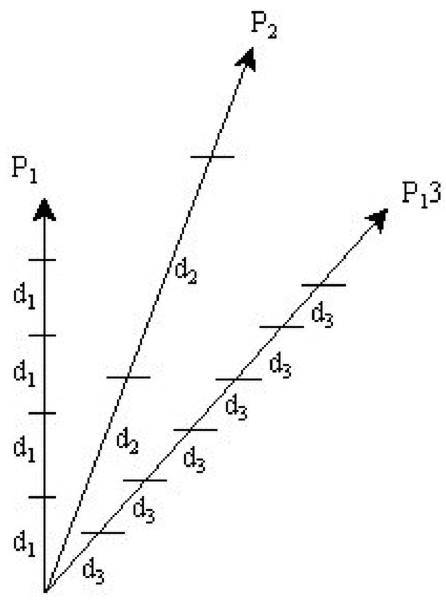
Neumann's definition of a time-scale directly inspired Ludwig Lange's conception of "inertial system," introduced in 1885 . An inertial coordinate system ought to be one in which free particles move in straight lines. But any trajectory may be stipulated to be rectilinear, and a coordinate system can always be constructed in which it is rectilinear. And so, as in the case of the time-scale, we cannot adequately define an inertial system by the motion of one particle. Indeed, for any two particles moving anyhow, a coordinate system may be found in which both their trajectories are rectilinear. So far the claim that either particle, or some third particle, is moving in a straight line may be said to be a matter of convention. We must define an inertial system as one in which at least three non-collinear free particles move in noncoplanar straight lines; then we can state the law of inertia as the claim that, relative to an inertial system so defined, the motion of any fourth particle, or arbitrarily many particles, will be rectilinear. The notions of inertial system and Neumann's time-scale, which Lange called an "inertial time-scale," may be combined as follows: relative to a coordinate system in which three free particles move in straight lines *and* travel mutually-proportional distances, the motion of any fourth free particle will be rectilinear and uniform. The questionable Newtonian concepts of absolute rotation and acceleration, Lange proposed, could now be replaced by the concepts of "inertial rotation" and "inertial acceleration," i.e. rotation and acceleration relative to an inertial system and inertial time-scale. See Figures 2 and 3.

Figure 2: Neumann's Time-Scale:
By Newton's first law, a particle not subject to forces travels equal distances in equal times. But which particles are free of forces? This might appear to be a matter of convention.

Either P_1 or P_2 can be arbitrarily stipulated to be at the origin of a system of coordinates, and to serve as the measure of equal times

But I can say of *two* particles with different velocities: in intervals of time in which one moves a given distance d_1, the other moves a proportional distance $d_2 = kd_1$ (where k is a constant; i.e., $d_1/d_2 = k$). Or I can compare a particle to a freely rotating planet: in intervals of time through which the planet rotates through equal angles, the particle moves equal distances.

Figure 3: Lange's Definition of 'inertial system' (1885):
An inertial system is a coordinate system with respect to which three free particles, projected from a single point and moving in non-coplanar directions, move in straight lines and travel mutually-proportional distances. The law of inertia then states that relative to any inertial system, any fourth free particle will move uniformly.



At about the same time, apparently unaware of the work of Mach, Neumann, and Lange, James Thomson expressed the content of the law of inertia, and the appropriate frame of reference and time-scale (“dial-traveller”), somewhat differently:

> For any set of bodies acted on each by any force, a REFERENCE FRAME and a REFERENCE DIAL-TRAVELLER are kinematically possible, such that relatively to them conjointly, the motion of the mass-centre of each body, undergoes change simultaneously with any infinitely short element of the dial-traveller progress, or with any element during which the force on the body does not alter in direction nor in magnitude, which change is proportional to the intensity of the force acting on that body, and to the simultaneous progress of the dial-traveller, and is made in the direction of the force. (Thomson 1884, p. 387)

More simply, an inertial reference-frame is one in which Newton’s second law is satisfied, so that every acceleration corresponds to an impressed force. Thomson did not reject the term “absolute rotation,” holding instead that it has to be understood as rotation relative to a reference frame that satisfies his definition. The definition does not express, as Lange’s does, the degree of arbitrariness involved in the construction of an inertial system by means of free particles. Moreover, like Lange’s, it leaves out a crucial condition for an inertial system as we understand it: all forces must belong to action-reaction pairs. Otherwise we could have, as on a rotating sphere, merely apparent (centrifugal) forces that are, by definition, proportional to mass and acceleration, and so the rotating sphere would satisfy Thomson’s definition. Therefore the definition needs to be completed by the stipulation that to every action there is an equal and opposite reaction. (This completion was actually proposed by R.F. Muirhead in 1887.)

But, so completed, Thomson’s definition has two advantages over Lange’s. First, by appealing to Newton’s second law instead of his first, it shows that we can apply the notion of inertial frame without having to consider the question whether there really are any free particles in nature. Second, it exhibits more clearly an essential point about the relation between the laws of motion and the inertial frames: that the laws assert the existence of at least one inertial frame. The original question, “relative to what frame of reference do the laws of motion hold?” is revealed to be wrongly posed. For the laws of motion essentially *determine* a class of reference frames, and (in principle) a procedure for constructing them. For the same reason, a skeptical question that is still commonly asked about the laws of motion -- why is it that the laws are true only relative to a certain choice of reference frame? -- is also wrongly posed. If Newton’s laws are true, then we can construct an inertial frame; their truth doesn’t depend on our ability to construct such a frame in advance.

2. Inertial Frames in the 20th Century: Special and General Relativity

2.1 Inertial Frames in Newtonian Spacetime

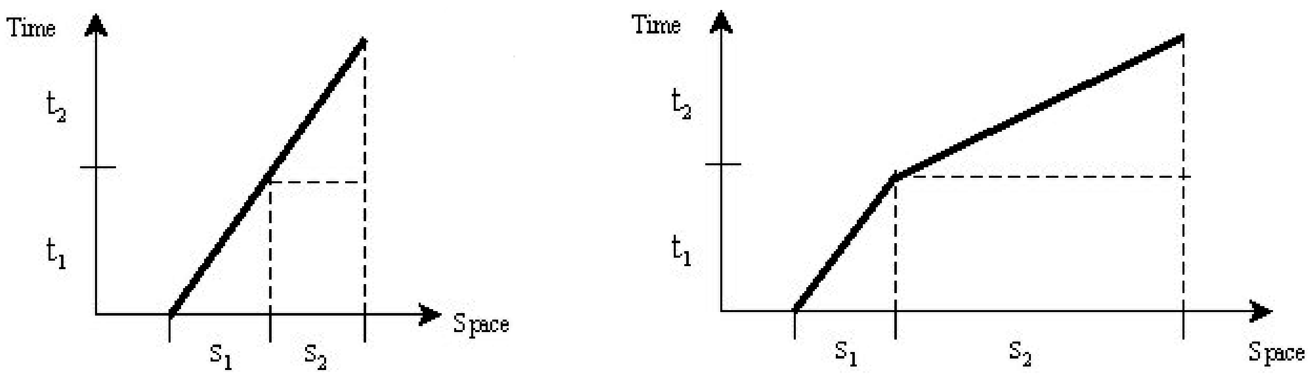
By the early years of the 20th century, this notion of inertial system seems to have been widely accepted, even if the specific works of Lange and Thomson were largely forgotten; in writing "On the electrodynamics of moving bodies" in 1905, Einstein took it to be obvious to his readers that classical mechanics does not require a single privileged frame of reference, but an equivalence-class of frames, all in uniform motion relative to each other, and any of which "the equations of mechanics hold good." Two inertial frames with coordinates (x, y, z, t) and ($x', y' z' t'$) are related by the *Galilean transformations*,

$$x' = x - vt$$
$$y' = y$$
$$z' = z$$
$$t' = t$$

(assuming that the x axis is defined to be the direction of their relative motion). These transformations clearly preserve the invariant quantities of Newtonian mechanics, i.e. acceleration, force, and mass (and therefore time, length, and simultaneity). As far as Newtonian mechanics was concerned, then, the problem of absolute motion was completely solved; all that remained was to express the equivalence of inertial frames in a simpler geometrical structure.

The lack of a privileged spatial frame, combined with the obvious existence of privileged states of motion -- paths defined as rectilinear in space and uniform with respect to time -- suggests that the geometrical situation ought to be regarded from a four-dimensional *spatiotemporal* point of view. The structure defined by the class of inertial frames can be captured in the statement that *spacetime is a four-dimensional affine space, whose straight lines* (geodesics) *are the trajectories of particles in uniform rectilinear motion*. See Figure 4.

Figure 4: Inertial Trajectories as Straight Lines of Spacetime

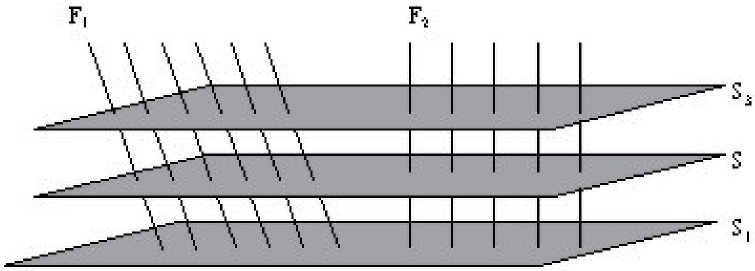


The uniformly moving particle will travel the same distance in the same intervals.

A particle that accelerates after t_1 will move a greater distance during t_2 and therefore its path *in spacetime* changes direction.

That is, spacetime is a structure whose automorphisms -- the Galilean transformations that relate one inertial frame to another -- are equivalent to affine transformations: they take straight lines into straight lines (i.e. an inertial motion in one inertial frame will be an inertial motion in any other inertial frame, and likewise for an accelerating or rotational motion), and parallel lines into parallel lines (i.e. uniformly-moving particles or observers who are relatively at rest in one frame will also be relatively at rest in another). (See Stein 1967, Ehlers 1973, and Friedman 1983 for further explanation.) An inertial frame can be characterized as a family of parallel straight lines "filling" spacetime, representing the possible trajectories of a family of free particles that are relatively at rest. See Figure 5:

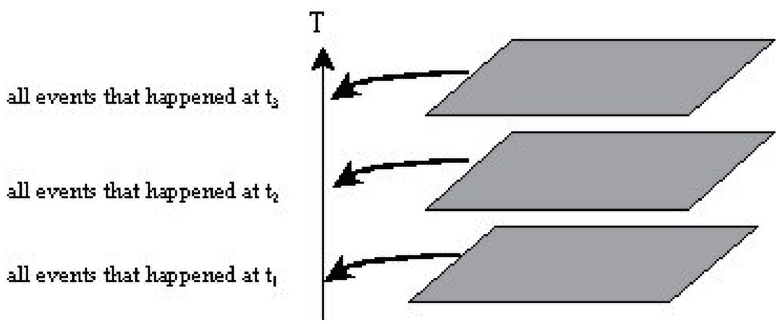
Figure 5:
Each of these families of straight lines, F_1 and F_2, represents the trajectories of a family of free particles that are relatively at rest, and therefore each defines an inertial frame. Relative to each other, the frames defined by F_1 and F_2 are in uniform motion.



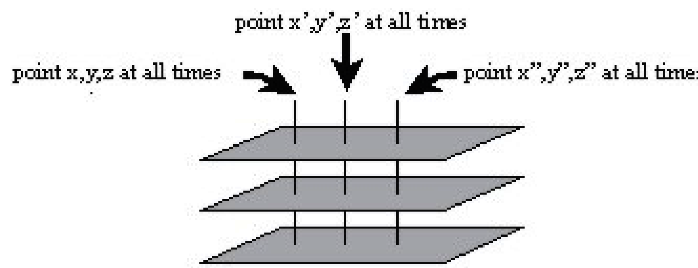
Each of the surfaces S is a "hypersurface of absolute simultaneity" representing all of space at a given moment; evidently (given the Galilean transformations) two inertial frames will agree on which events in spacetime are simultaneous.

From this we can see that the assertion that an inertial frame exists imposes a global structure on spacetime; it is equivalent to the assertion that spacetime is flat. As we can see from the Galilean transformations, distinct inertial frames will agree on time and simultaneity. Therefore, in the four-dimensional picture, the decomposition of spacetime into hypersurfaces of absolute simultaneity is independent of the choice of inertial frame. Another way of putting this is that Newtonian spacetime is endowed with a *projection* of spacetime onto time, i.e. a function that identifies spacetime points that have the same time-coordinate. Similarly, absolute space arises from a projection of spacetime onto space, i.e. a function that identifies spacetime points that have the same spatial coordinates. See Figure 6.

Figure 6:



The relation of simultaneity "decomposes" spacetime into 3-dimensional pieces, each representing "all of space at a given time," by projecting spacetime onto time, i.e., by identifying spacetime points that have the same time coordinates.



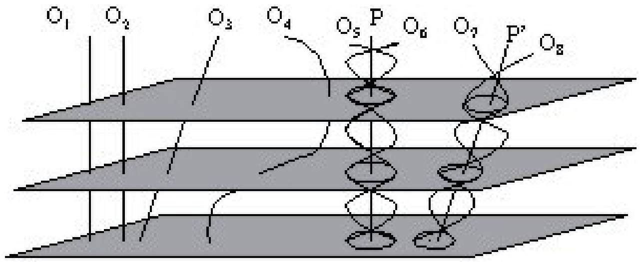
Similarly, one can think of the notion of "same place" as projecting spacetime onto space, i.e., by identifying spacetime points that have the same spatial coordinates; each of the trajectories thus singled out represents "a given place at all times."

But this latter projection is arbitrary: while it assumes that we can identify the same time at different spatial locations, Newtonian mechanics provides no physical way of identifying the same spatial point at different times. Thus the equivalence of inertial frames can be thought of as the arbitrariness of the projection of spacetime onto space, any such projection being, essentially, the arbitrary choice of some particular inertial frame as a rest-frame.

Figure 7: :

Here is a spacetime diagram of motions relative to the inertial frame in which O_1, O_2, and P are at rest. This can be seen as arising from the projection of each of their inertial trajectories onto a single point of space.

Here is the same situation viewed from an inertial frame in which O_3 and P' are at rest. Now O_1, O_2, and P are in uniform motion.



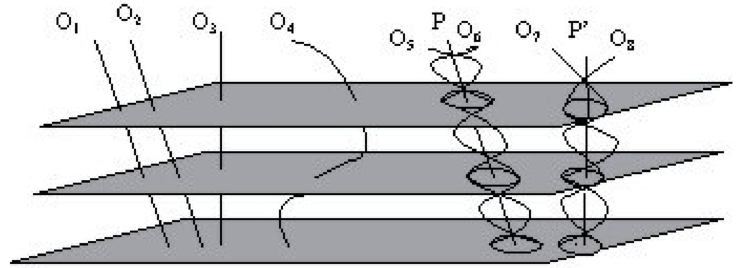
O_1 and O_2 are at rest

O_3 is in uniform motion

O_4 is accelerating any old way

O_5 and O_6 are revolving around their common centre of gravity P, which is at rest

O_7 and O_8 are revolving around their centre of gravity P′, which is in uniform motion.



O_3 is at rest

O_1 and O_2 are in uniform motion

O_4 is accelerating any old way

O_5 and O_6 are revolving around their common centre of gravity P, which is in uniform motion

O_7 and O_8 are revolving around their centre of gravity P′, which is at rest

2.2 The Conflict Between Galilean Relativity and Modern Electrodynamics

By the time that this representation of the Newtonian spacetime structure was developed, however, the Newtonian conception of inertial frame had been essentially overthrown. First, 19th-century electrodynamics raised again the question of a privileged frame of reference: the conception of light as an electromagnetic wave in the ether implied that the rest-frame of the ether itself should play a distinguished role in electrodynamical phenomena. On the one hand, physicists such as Maxwell and Lorentz were careful to point out that velocity relative to the ether was not equivalent to absolute velocity, and that the state of motion of the ether itself was necessarily unknown -- in other words, that this conception of light did not violate the classical principle of relativity. On the other hand, the existence of such a preferred frame made the equivalence of inertial frames correspondingly less interesting, even if it was true in principle. This is why the appearance of the idea of inertial frame in the 1880's, as I suggested earlier, was not of pressing physical interest to the majority of physicists, and seemed to be a mere philosophical sidelight. The attempts to measure the effects of motion relative to the ether commanded considerably more attention.

Second, the abandonment of the ether -- following the failure of attempts to measure velocity relative to the ether and, more generally, the apparent independence of all electrodynamical phenomena of motion relative to the ether -- did not vindicate the Newtonian inertial frame, but required a dramatically revised conception. Special relativity might be said to have applied the relativity principle of Newtonian mechanics to Maxwell's electrodynamics, by eliminating the privileged status of the rest-frame of the ether and admitting that the velocity of light is independent of the motion of the source. As Einstein expressed it, "the same laws of electrodynamics and optics will be valid for all frames of reference for which the equations of mechanics hold good." (1905, p. 38.) But as Einstein also pointed out, the invariance of the velocity of light and the principle of relativity, at least in its Galilean form, are incompatible. It simply makes no sense, according to Galilean relativity, that any velocity should appear to be the same in inertial frames that are in relative motion.

2.3 Special Relativity and Lorentz Invariance

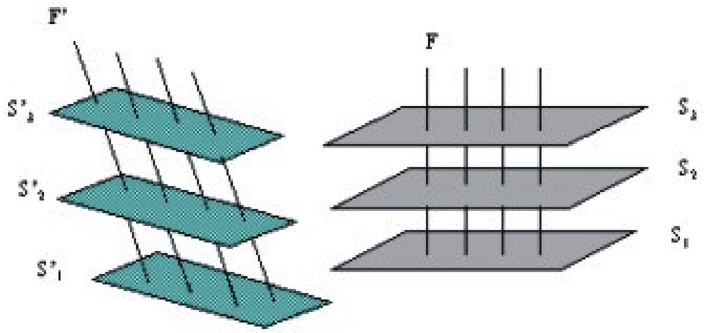
2.3. Special relativity and Lorentz invariance

Einstein solved this difficulty through his analysis of simultaneity: frames in relative motion can agree on the velocity of light only if they disagree on simultaneity; only the relativity of simultaneity makes possible the invariance of the velocity of light. This means that the transformations between inertial frames that preserve the velocity of light will not preserve simultaneity. These are the *Lorentz transformations*:

$$x' = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad y' = y \qquad z' = z \qquad t' = \frac{t - \frac{vx}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Evidently these transformations do not preserve length and time, and so the invariant quantities of Newtonian mechanics, which presuppose invariant measures of length and time, must now depend on the choice of inertial frame. By the same token, the notions of force, mass, and acceleration can no longer be appealed to in the definition of an inertial frame. The definition must instead appeal to the invariant quantities of electrodynamics: an inertial frame is one in which light travels equal distances in equal times in arbitrary directions. What seems impossible, from the point of view of Galilean relativity, is that a frame that moves uniformly relative to such a frame should also satisfy the definition. But that, again, rests on the assumption that two inertial frames will have a common measure of simultaneity. If, as Einstein asserts, the only reasonable definition of simultaneity is one provided by light signals, then there is no determination of simultaneity that will give the same results in different inertial frames. The spacetime structure that is implied by special relativity is thus an affine space, like Newtonian spacetime, but it is not objectively divided into hypersurfaces of absolute simultaneity; the sets of simultaneous events for any inertial frame are the hyperplanes orthogonal to the trajectories that determine that frame. In other words, the choice between two inertial frames determines a choice between two distinct divisions of spacetime into space and time. See Figure 8:

Figure 8:
The inertial frames F and F′ are in relative motion, and therefore, as the Lorentz transformations indicate, they disagree on simultaneity. F and F′ thus determine distinct decompositions of spacetime into instantaneous spaces, S and S′, respectively



2.4 Simultaneity and Reference-Frames

The details of Einstein's argument and the structure of Minkowski spacetime can be found elsewhere (see, e.g., Einstein 1951 and Geroch 1978). Here only one more point is worth making. It could be argued that Einstein's and Lorentz's view are completely equivalent. That is, we could assume that there is indeed a privileged frame of reference, and that the apparent invariance of the velocity of light is explained by the effects on bodies of their motion through the ether (the Lorentz contraction and time dilation). This purported distinction between empirically indistinguishable frames has often been criticized on straightforward methodological grounds, but it could be (and surely has been) argued that it is more intuitively plausible than the relativity of simultaneity. After all, knowing that (as Einstein showed) the Lorentz contraction can be derived from the invariance of the velocity of light does not, by itself, entitle us to say which of the two is the more convincing starting-point.

This is why it is so important that Einstein's 1905 paper begins with a critical analysis of the entire notion of a frame of reference. It is tacitly assumed by Lorentz's theory, and classical electrodynamics generally, that we have a reference-frame in which we can measure the velocity of light. But how is such a reference-frame determined? The distances between points in space can only be determined if it is possible to determine which events are simultaneous. In practice this is always done by light-signalling, if only in the informal sense that we identify simultaneous events when we see them at the same time. But if the spatial frame of reference is determined by light-signals, and is then to be used to measure the speed of light, we would appear to be going in a circle; the underlying assumption must be that, while light-signalling is useful and practical, it is not essential to the definition of simultaneity, and that there is a fact of the matter about which events are simultaneous that is independent of this method of signalling. This assumption was actually made explicit by James Thomson. He recognized -- alone, apparently, before Einstein -- that the measurement of distance involves

> the difficulty as to imperfection of our means of ascertaining or specifying, or clearly idealizing, simultaneity at distant places. For this we do commonly use signals by sound, by light, by electricity, by connecting wires or bars, and by various other means. The time required in the transmission of the signal involves an imperfection in human powers of ascertaining simultaneity of occurrences at distant places. It seems, however, probably not to involve any difficulty of idealizing or imagining the existence of simultaneity. Probably it may not be felt to involve any difficulty comparable to that of attempting to form a distinct notion of identity of place at successive times in unmarked space. (1884, p. 380).

In other words, Thomson assumed that it was not a difficulty in principle, like the difficulty of determining rest in absolute space. But Einstein showed that it was precisely the same kind of difficulty, and that determinations of simultaneity involve reference to an arbitrary choice of reference-frame, just as much as determinations of velocity. Einstein's conclusion is, of course, entirely contingent on the empirical facts of electrodynamics; it could have been avoided if there were in nature a useful signal of some kind whose transmission would provide a criterion of absolute simultaneity, so that the same events would be determined to be simultaneous in all inertial frames. Or, experiments might have been able to reveal the dependence of the velocity of light on the state of motion of the source. Then synchronization by light-signals could still have been regarded as a mere practical substitute for a notion of absolute simultaneity that stood on independent grounds, empirically as well as conceptually. But as Einstein saw, because of the apparent independence of the velocity of light of the motion of the source, even "idealizing or imagining the existence of simultaneity" involves light-signaling more essentially than anyone could have realized. Unless some other criterion of simultaneity is provided, therefore, the establishment of a spatial frame of reference involves light-signaling in an essential way. In the absence of such a criterion the speed of light cannot be, as Lorentz supposed, empirically measured against the background of an inertial frame; in that case the only empirically sound definition of an inertial frame is the one that appeals to the speed of light.

2.5 From Special Relativity and Lorentz Invariance to General Relativity and General Covariance

It may seem surprising that, after this insightful analysis of the concept of inertial frame and its role in electrodynamics, Einstein should have turned almost immediately to call that concept into question. But he had a compelling combination of physical and philosophical motives to do so. On the physical side, he realized (along with many others) that special relativity would require some fundamental revision of the Newtonian theory of gravity. On the philosophical side, he became convinced, largely by his reading of Mach (1883), that the central role of inertial frames was an "epistemological defect" that special relativity shared with Newtonian mechanics. (Einstein 1916, pp. 112-113.) Only relative motions are observable, yet both of these theories purport to identify a privileged state of motion and use it to explain observable effects (such as centrifugal forces). Coordinate systems are not observable, yet both of these theories assign a fundamental physical role to certain kinds of coordinate system, namely, the inertial systems. In either theory, inertial coordinates are distinguished from all others, and the laws of physics are said to hold only relative to inertial coordinate systems. In an epistemologically sophisticated theory, both of these problems would be solved at once: the new theory would only refer to what is observable, which is relative motion; it would admit arbitrary coordinate systems, instead of confining itself to a special class of system. Why, after all, should any genuine physical phenomenon depend on the choice of coordinate system?

Another way of putting the same point is to say that, in Newtonian mechanics and special relativity, rotation is "absolute" because the transformations between inertial frames (Galilean or Lorentzian) preserve rotational states. Thus the "absoluteness" of rotation arises precisely from singling out one type of frame, by one type of transformation, instead of allowing arbitrary transformations and arbitrary frames. Einstein held that this epistemological insight had a natural mathematical representation in the principle of *general covariance*, or the principle that the laws of nature are to be invariant under *arbitrary* coordinate transformations. More precisely, what this means is that coordinate transformations are no longer required (as in the affine spaces of Newtonian mechanics and special relativity) to take straight lines to straight lines, but only to preserve the smoothness of curves (i.e. their differentiability). The general theory of relativity was intended to be a generally covariant account of spacetime, and its general covariance was intended to express the general relativity of motion. And the theory came into being because Einstein perceived a deep connection between this project and that of finding a relativistic theory of gravitation.

2.6 The Equivalence of Inertia and Gravity

The philosophical motivations and implications of Einstein's view are dealt with elsewhere. (See, for example, the entries on Einstein's philosophy of science; the hole argument; and early philosophical interpretations of general relativity.) We will consider here only the bearing of general relativity on the notion of an inertial frame. It is questionable whether Einstein succeeded in

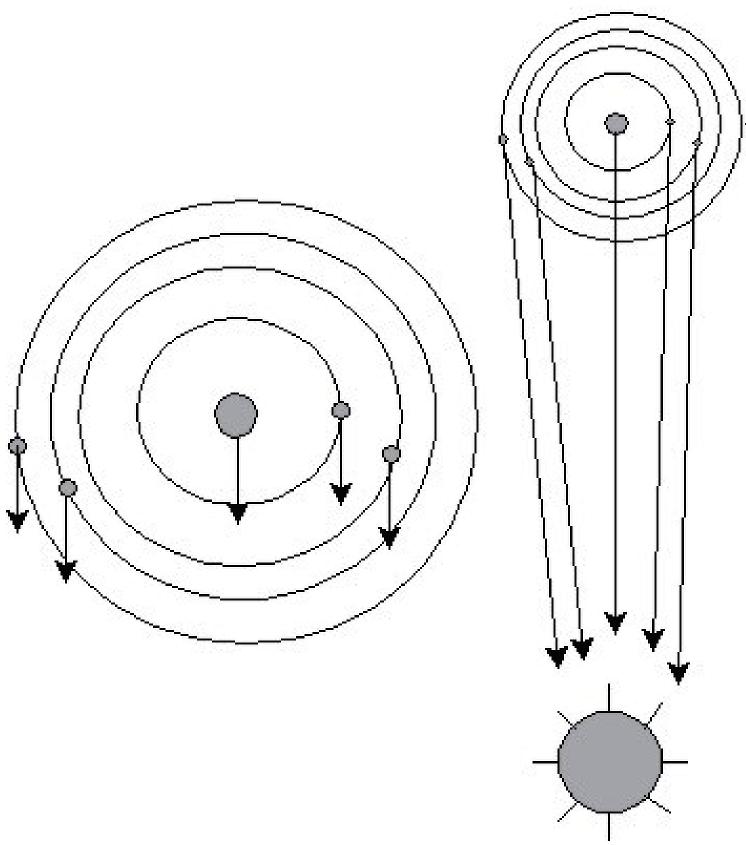
establishing the general relativity of motion, but it is clear that general relativity undermines the concept of inertial frame in important respects. This arises from the *equivalence principle*: that inertial mass -- the quantity that enters into Newton's second law, and that is a measure of a body's resistance to acceleration -- is equivalent to gravitational mass, the quantity that enters into Newton's law of universal gravitation. A more empirical way of expressing it is that all bodies fall with the same acceleration in the same gravitational field, or, the trajectory of a body in a given gravitational field will be independent of its mass and composition. This is the principle that Newton tested by constructing pendulums with wooden boxes as their bobs, which he would fill with different materials in order to see whether those differences made a difference to the speed of falling; they didn't. Eötvös made more precise tests in the late 19th century, and established the principle to much greater accuracy; these are the results on which Einstein would have relied. Newton also tested the principle for bodies whose masses differ greatly, by observing that Jupiter and its four moons all received precisely the same acceleration from the sun's gravitational field.

The equivalence principle suggests, however, that a freely-falling frame of reference is physically indistinguishable from an inertial frame. Newton had already noticed this, and indeed he stated it, more or less, in Corollary VI to the laws of motion:

> If bodies are moving in any way whatsoever with respect to one another and are urged by equal accelerative forces along parallel lines, they will all continue to move with respect to one another in the same way as they would if they were not acted on by those forces. (1726, p. 423.)

For example, he was able to treat the system of Jupiter and its moons as if it were (nearly) at rest or moving uniformly in a straight line, because the attractive force of the sun acts (almost) equally on every part of the system. See Figure 9:

Figure 9: Newton's Corollary VI
What seem, within a given system, like equal and parallel accelerations may be, on a larger scale, unequal and converging on some distant massive object; e.g., the system of Jupiter and its moons is falling toward the sun, but "locally" the accelerations are very nearly equal and parallel, and may therefore be neglected.



He even applied this reasoning to the entire solar system, in order to justify treating it as an isolated system: if there were any outside force acting on it, it must have been acting more or less equally and in parallel directions on all parts of the system.

> It may be alleged that the sun and planets are impelled by some other force equally and in the direction of parallel lines;

> but by such a force (by Cor. VI of the Laws of Motion) no change would happen in the situation of the planets to one another, nor any sensible effect follow; but our business is with the causes of sensible effects. Let us, therefore, neglect every such force as imaginary and precarious, and of no use in the phenomena of the heavens....(1729, volume 2 p. 558)

Now, it is a familiar fact that in an orbiting spacecraft, bodies behave as if no forces were acting on any of them (as if they were “weightless”), because the attraction of the earth acts equally on all of them. But these phenomena are not, by themselves, evidence that *no* phenomena are capable of distinguishing an inertial frame from a falling frame. Einstein was willing to generalize the equivalence principle, and to conclude that the classical idea of a distinguished class of frames of reference has no physical basis. Any frame that we might regard as inertial might be, for all we can tell by experiment, in free fall. By the same token, any frame that is uniformly accelerating is indistinguishable from one that is at rest in a uniform gravitational field. Suppose that you are in a box at rest on the earth; you and everything in the box, by the equivalence principle, will be accelerated downward with the acceleration g (= 9.8 meters/second/second). Now suppose that the box itself is in empty, gravity-free space, but accelerating upward (i.e. in the direction of its roof) with the acceleration $-g$. Obviously, because of their inertia, bodies in the box, including your own, will exert the same force -- have the same “weight” -- on the floor as if the box were at rest and sitting on the earth.

2.7 The Equivalence Principle and General Covariance

To get a clearer idea of the physical significance of the equivalence principle, and its connection with general covariance, consider the Newtonian procedure for analyzing motion in the solar system, here sketched very roughly:

1. Determine the accelerations of all the planets relative to the fixed stars.
2. Using the laws of motion, their corollaries, and all the propositions proved from these in Book I of Principia, derive from the accelerations the forces needed to produce them; in particular, derive from the orbits the centers of those orbits, and the masses of the bodies needed to produce those forces. This crucially involves the law of action and reaction, for otherwise it would be impossible to break down the total acceleration of any planet into the components contributed by particular other planets; the earth’s acceleration, for example, is the sum of its accelerations toward all the other planets, and each individual acceleration is part of an action-reaction pair involving some other planet.
3. When we understand the mutual interactions among the planets, we are in a position to estimate their relative masses. In Newton’s case, this was necessarily restricted to the planets with satellites, because only in those cases could he compare the accelerations they determine at given distances and so deduce the differences in mass. By this reasoning he estimated the ratios of the Sun’s mass to those of Jupiter (1067 to 1), Saturn (3021 to 1), and the earth, and was able to calculate that the center of mass of the entire solar system would never be more than one solar diameter from the center of the Sun.
4. Having found the center of mass, we have in principle determined an inertial frame: by Corollary IV to the laws of motion, the center of mass will be at rest or moving uniformly in a straight line. That is, the mutual actions of the bodies in the system will not change the state of motion of the center of mass. And having determined an inertial frame, we are in a position to say that the accelerations relative to the center of mass frame are the true accelerations.

One might think that the problem of relativity arises right from the start: the reliance on the fixed stars already seems to introduce an arbitrary assumption that threatens to vitiate Newton’s procedure as an account of the true motions. But the framework of the fixed stars, initially just taken for granted, turns out to be justified in the course of the analysis. If it turns out that all the accelerations relative to the fixed stars can be analyzed into action-reaction pairs involving bodies within the system, leaving no “leftover” accelerations that need to be traced to some yet-unknown influence, then we can conclude that the stars are a suitable (sufficiently inertial) frame of reference after all. (By the later 19th century, observations became sufficiently precise to reveal that there is in fact a leftover acceleration, namely the famous extra precession of Mercury. But that could not affect Newton’s analysis in 1687.) In contrast, had we chosen the earth as a frame of reference, we would find that there are accelerations relative to this frame -- e.g. Coriolis and centrifugal accelerations -- that don’t satisfy the law of action and reaction.

2.8 The Extension of the Relativity Principle

The relativistic aspect of this situation arises from the equivalence principle. Newton’s Corollary VI said that the inertial frame we construct by this procedure is effectively indistinguishable from one in which all the bodies are undergoing equal and parallel accelerations caused by some force that acts equally on all of them; the equivalence principle asserts that gravity is such a force. In following the Newtonian procedure for constructing an inertial frame, we have constructed a frame which might be, for all we can determine empirically, falling in the gravitational field of some other system. This means that the accelerations relative to this frame

cannot be known to be the “true accelerations”; they may be accelerations relative to a freely-falling trajectory just in case the center of mass is itself freely falling, in which case they have to be added to the gravitational acceleration of the center of mass before we can arrive at the true accelerations. But the acceleration of the center of mass may have to be added to some larger acceleration--and so on. This means that we can’t know the true strength of the gravitational field by observing the motions in this frame. The only hope of doing so would be to include all the mass in the universe in one dynamical system; if we knew the center of mass of the entire universe, we could rule out the possibility that something else is exerting an accelerative force, since by hypothesis there would be nothing else.

We can see the significance of this more clearly by looking at the equations of motion (in a very simplified form). Newton’s equation of motion for a particle subject to no force asserts that it moves uniformly, with zero acceleration. Obviously, in a gravitational field, the particle’s acceleration will depend on the field. In effect, we are accounting for the trajectory of the falling particle by “decomposing” it into two parts, the part determined by its natural tendency to move uniformly in a straight line, and the part contributed by the gravitational field. But by the analysis of the equivalence principle, determining the inertial part--and therefore determining the gravitational part--depends on our assumption that the center of mass frame is inertial rather than freely falling. And this assumption is arbitrary; that is, it amounts to an arbitrary choice of the coordinate system in which we define the equation of inertial motion. This implies that the gravitational field depends on the coordinate system in precisely the same way.

The principle of general covariance, then, acquires its physical significance in conjunction with the equivalence principle. By itself, it says that the geometrical structures of spacetime don’t depend on the coordinates in which we express them, or on the set of points that we may think comprises spacetime. This is an important principle, but it doesn’t recommend general relativity over other theories, since special relativity and Newtonian mechanics also involve spacetime structures that can be defined in a generally-covariant way, through the same kinds of coordinate-independent mathematical objects that we use in general relativity. Combined with the equivalence principle, however, it implies that a central Newtonian idea--that gravity is a force causing deviations from uniform rectilinear motion-- is based on an arbitrary choice of coordinates. For a trajectory that satisfies all empirical criteria for being inertial in a particular frame of reference--e.g. the trajectory of the center of mass in our example--may be freely falling relative to some other trajectory that satisfies the same criteria. By contrast, a freely-falling trajectory is a freely falling trajectory in any coordinate system; it is only the decomposition of it into its inertial and gravitational parts that will be different in different coordinate systems.

2.9 From Inertial Frames to Curved Spacetime

General covariance is thus not an argument against privileged states of motion, as Einstein had hoped it would be. It is an argument that the privileged states of motion should not be mere artifacts of our choice of coordinates, i.e. that they should be coordinate-independent. Precisely what this means depends, then, on what physical means we have at our disposal to identify states of motion other than by simply setting down coordinates. Combined with the equivalence principle, it is an argument for regarding gravitational free-fall as the privileged state of motion, rather than as a forced deviation from the privileged state of motion. And in this way it provides an argument for spacetime curvature. As we saw, in Newtonian and Minkowski spacetime the inertial trajectories are, by definition, the straight lines or geodesics of spacetime. And the flatness of spacetime consists in the fact that these geodesics behave like straight lines in a flat space or surface: parallel geodesics remain parallel, and non-parallel geodesics do not accelerate relative to one another. (In any inertial frame, the motion of any other inertial frame appears uniform.) By the equivalence principle, however, free-fall trajectories satisfy all empirical criteria for being inertial trajectories, and so the distinction between the two types of trajectory depends on the mere choice of coordinates. General covariance suggests, then, that the free-fall trajectories ought to be identified as the inertial trajectories -- and therefore, as the geodesics of spacetime. But if free-fall trajectories are the geodesics of spacetime, then spacetime is curved. For the free-fall trajectories exhibit relative accelerations, and the relative acceleration of geodesics is a defining characteristic of curved geometry. The curvature of the earth’s surface, for example, is revealed in the fact that geodesics that begin in parallel directions can begin to approach one another -- for example, two lines of longitude can both be perpendicular to the equator, but converge on one another as they approach the poles. And since the relative accelerations of falling bodies depend on the distribution of mass, as we already knew from Newton’s theory, we now conclude not only that spacetime is curved, but that its curvature is determined by the distribution of mass. (For further explanation see Geroch 1978.)

The curvature of spacetime, finally, determines the status of inertial frames in general relativity. The statement that all reference-frames, rather than just inertial frames, are equivalent is a misleading way of describing the situation; rather, the variable curvature of spacetime makes the imposition of a global inertial frame impossible. So the status of the latter is like the status of a plane rectangular coordinate system on the surface of the earth. Over a sufficiently small area, the coordinate plane may be a good approximation to the surface, but over increasingly large areas it diverges increasingly from the contours of the earth. And if two such coordinate systems,

with their origins at different points on the earth, are extended until they meet, they will be seen to be “disoriented” relative to one another. In contrast, a flat plane can be so coordinatized, and coordinate systems originating at different points can be smoothly combined into one system. Similarly, in the affine spaces of Newtonian and special-relativistic physics, any inertial coordinate system can be extended over the whole of spacetime. And in any system so extended, the trajectory of every other inertial observer will be a uniform rectilinear motion. But if spacetime is variably curved, according to the distribution of mass and energy, local inertial systems will be “disoriented” relative to one another; indeed, the degree of this “disorientation” is one of the measures of curvature. And an inertially-moving -- i.e. freely falling -- particle will in that case be accelerating in the local inertial system of another freely-falling particle. Thus there are inertial trajectories, but no extended inertial systems. See Figures 10 - 13:

Figure 10:
This Cartesian coordinate system can evidently be simply “set down” over the plane below. Any coordinate system defined at any point of the plane can be smoothly extended over the entire plane.

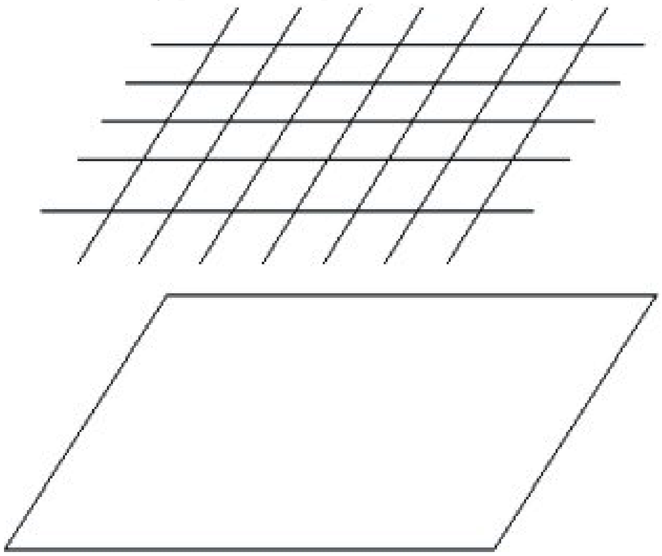
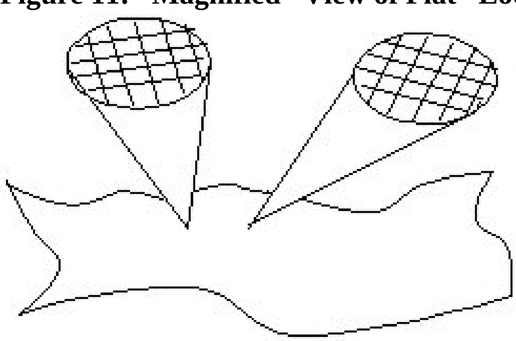


Figure 11: “Magnified” View of Flat “Local” Coordinate Systems on a Curved Surface



This arbitrary curved surface won’t allow for the global laying down of a coordinate system, but must be coordinatized in small overlapping pieces, which generally won’t be parallel to one another.

Figure 12:
In a flat spacetime, the rest-frame of any inertial observer an be “extended” over all of spacetime in such a way that, in this global inertial frame, the trajectory of every other inertial observer will be an inertial trajectory.

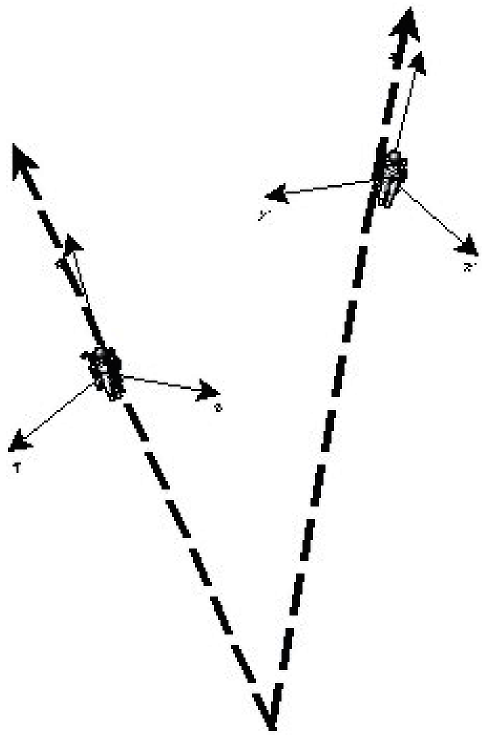
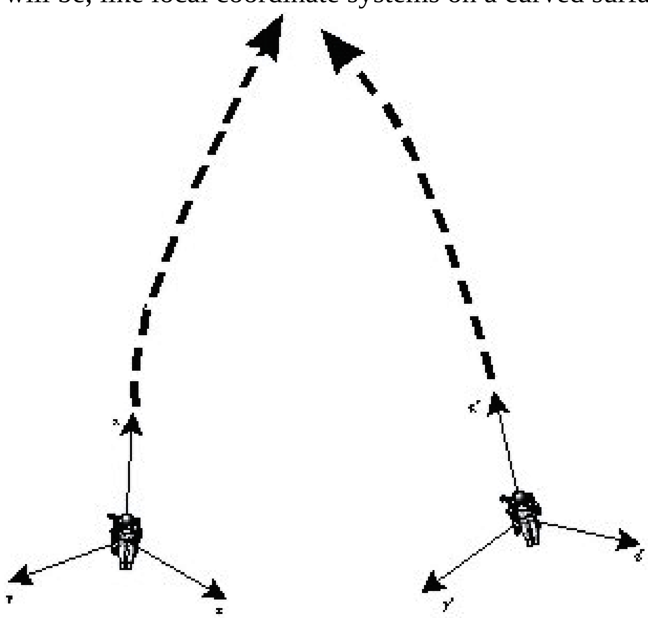


Figure 13:
In a curved spacetime, inertial trajectories will be relatively accelerated; indeed the relative acceleration of geodesics is a measure of curvature. Therefore the local inertial frame of any freely-falling observer cannot be extended into a global frame in which all other inertial observers are moving uniformly. The inertial frames of different freely-falling observers will be, like local coordinate systems on a curved surface, “disoriented” relative to one another.



One could try to express this idea with Einstein’s remark about the need to “free oneself from the idea that coordinates must have an immediate metrical meaning.” (Einstein 1949, p. 67.)But even this might be misleading. Einstein evidently was thinking that, in general relativity, coordinates, and coordinate transformations, no longer represent the possible displacements of rigid bodies or the transport of ideal clocks. The insight underlying this is that the notion of rigid displacement -- therefore of rigid coordinate system, and inertial frame -- imposes *a priori* a degree of uniformity, or symmetry, on spacetime; the displacement of bodies without change of dimension, and the transport of an ideal clock without distortion of time-intervals, requires a homogeneous space. And so rigid displacement cannot be a basic principle in a theory in which spacetime curvature varies according to the distribution of mass and energy. The possibility of a rigid displacement, and therefore the existence of an inertial frame, can only arise *a posteriori*, as the result of a peculiar distribution of mass-energy (for example, in a universe empty of mass and energy, or with a highly symmetrical distribution). The serious defect in the notion of inertial frame is not that it makes an arbitrary distinction among coordinate systems --

for the distinction is quite as genuine as the distinction between flat and curved spacetime -- but that it extends indefinitely over spacetime a structure that, in our universe, only corresponds approximately to very small regions.

Bibliography

- DiSalle, R. (1988). *Space, Time and Inertia in the Foundations of Newtonian Physics*. Unpublished Ph.D. Dissertation, University of Chicago.
- ----- (1991). "Conventionalism and the origins of the inertial frame concept." In *PSA 1990.* East Lansing: The Philosophy of Science Association.
- ----- (2002). “Newton’s philosophical analysis of space and time.” Forthcoming in *The Cambridge Companion to Newton*. Cambridge: Cambridge University Press.
- Earman, J. (1989). *World Enough and Spacetime: Absolute and Relational Theories of Motion*. Boston: M.I.T. Press.
- Ehlers, J. (1973) "The nature and structure of space-time." In *The Physicist’s Conception of Nature*. Edited by Jagdish Mehra, 71-95. Dordrecht: Reidel.
- Einstein, A. (1905). “On the electrodynamics of moving bodies.” In Einstein, *et al*. (1952), pp. 35- 65.
- Einstein, A. (1916). “The foundation of the general theory of relativity.” In Einstein,, *et al*. (1952), pp. 109-164.
- Einstein, A. (1949), “Autobiographical notes.” In P.A. Schilpp, ed., *Albert Einstein, Philosopher-Scientist*. Chicago: Open Court.
- Einstein, A. (1951). *Relativity: The Special and the General Theory*. R. Lawson, tr. New York: Crown Publishers Inc..
- Einstein, A., H. A. Lorentz, H. Minkowski, and H. Weyl (1952). *The Principle of Relativity*. W. Perrett and G.B. Jeffery, trs. New York: Dover Books.
- Friedman, M. (1983). *Foundations of Space-Time Theories*. Princeton: Princeton University Press.
- Lange, L. (1885). "Ueber das Beharrungsgesetz." *Berichte der Königlichen Sachsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-physische Classe* 37 (1885): 333-51.
- Leibniz, G. (1970). *Philosophical Papers and Letters*. Edited by Leroy Loemker. Dordrecht: Reidel.
- Mach, E. (1872). *Die Geschichte und die Wurzel des Satzes von der Erhaltung der Arbeit*. Prague: J.G. Calve’sche K.-K. Univ-Buchhandlung.
- Mach, E. (1883). *Die Mechanik in ihrer Entwickelung, historisch-kritisch dargestellt*. 2nd edition. Leipzig: Brockhaus.
- Minkowski, H. (1908). “Space and time.” In Einstein, *et al*. (1952), pp. 75-91.
- Misner, C., K. Thorne, and J.A. Wheeler (1973). *Gravitation*. San Francisco: Freeman
- Muirhead, R.F. (1887). "The laws of motion." *Philosophical Magazine*, 5th series, 23: 473-89.
- Neumann, C. (1870). *Ueber die Principien der Galilei-Newton’schen Theorie*. Leipzig: B. G. Teubner, 1870.
- Newton, I. (1726). *The Principia: Mathematical Principles of Natural Philosophy*, tr. I. Bernard Cohen and Anne Whitman. Berkeley and Los Angeles: University of California Press, 1999.
- Newton, I. (1729). *Sir Isaac Newton’s Mathematical Principles of Natural Philosophy and his System of the World*. 2 vols. Edited by Florian Cajori. Translated by Andrew Motte. Berkeley: University of California Press, 1962.
- Seeliger, H. (1906). "Über die sogenannte absolute Bewegung." *Sitzungs-Berichte der Bayerische Akademie der Wissenschaft*: 85-137.
- Stein, H. (1967). “Newtonian space-time.” *Texas Quarterly* 10: 174-200.
- Stein, H. (1977). "Some philosophical prehistory of general relativity." In *Foundations of Space-Time Theories*. Edited by John Earman, Clark Glymour, and John Stachel, 3-49. *Minnesota Studies in the Philosophy of Science*, Vol. 8. Minneapolis: University of Minnesota Press.
- Thomson, J. (1884). "On the law of inertia; the principle of chronometry; and the principle of absolute clinural rest, and of absolute rotation." *Proceedings of the Royal Society of Edinburgh* 12: 568-78.
- Torretti, R. (1983). *Relativity and Geometry*. Oxford: Pergamon Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Einstein, Albert: philosophy of science | general relativity: early philosophical interpretations of | geometry: in the 19th century | space and time: conventionality of simultaneity | space and time: the hole argument

Robert DiSalle
rdisalle@uwo.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 29, 2002
Content last modified: March 29, 2002

Nineteenth Century Geometry

In the nineteenth century, geometry, like most academic disciplines, went through a period of growth that was near cataclysmic in proportion. During the course of this century, the content of geometry and its internal diversity increased almost beyond recognition; the axiomatic method, highly touted since antiquity by the admirers of geometry, attained true logical sufficiency, and the ground was laid for replacing the standard geometry of Euclid by Riemann's more pliable system in the description of physical phenomena. Modern philosophers of all tendencies --- Descartes and Hobbes, Spinoza and Locke, Hume and Kant --- had regarded Euclidean geometry as a paradigm of epistemic certainty. The downgrading of Euclidean geometry to a subspecies of the vast family of mathematical theories of space shattered philosophical illusions and prompted important changes in our understanding of human knowledge. After these nineteenth-century developments, those who thirst for an absolute knowledge of right and wrong can no longer propose Euclidean geometry as the one instance in which a like goal has proved attainable. The present article reviews the aspects of nineteenth century geometry that are of major interest for philosophy and hints, in passing, at their philosophical significance.

---

1. Lobachevskian geometry

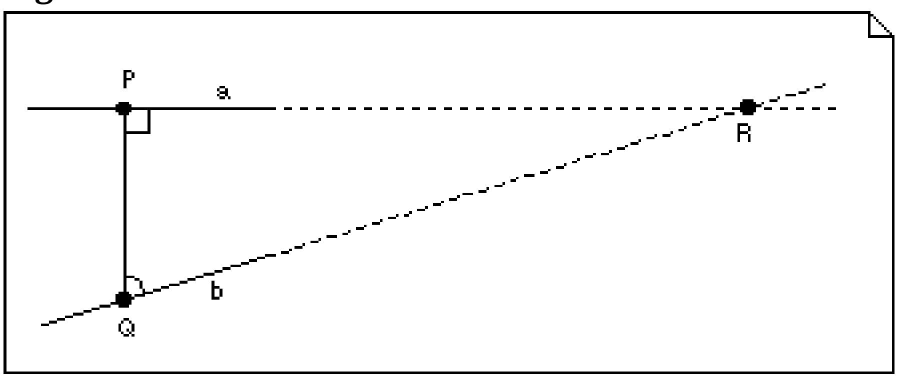
Euclid (fl. 300 b.c.) placed at the head of his *Elements* a series of 'definitions' (e.g., "A point is that which has no part") and 'common notions' (e.g., "If equals be added to equals, the sums are equal"), and five 'requests'. Supposedly these items conveyed all of the information needed for inferring the theorems and solving the problems of geometry, but as a matter of fact they do not. However, the requests

(*aitemata*)---usually called ‘postulates’ in English---must at any rate be granted or Euclid’s proofs will not go through. Some of them are plainly practical:

> 1. To draw a straight line from any point to any point.
> ...
> 3. To draw a circle with any center and any radius.

However, the fifth one sounds more like a statement of fact. Still, it can readily be paraphrased as a recipe for constructing triangles: Given any segment PQ, draw a straight line a through P and a straight line b through Q, so that a and b lie on the same plane; verify that the angles that a and b make with PQ on one of the two sides of PQ add up to less than two right angles; if this condition is satisfied, it should be granted that a and b meet at a point R on that same side of PQ, thus forming the triangle PQR. (See Figure 1.) This request is known as "Euclid’s Postulate".

Figure 1



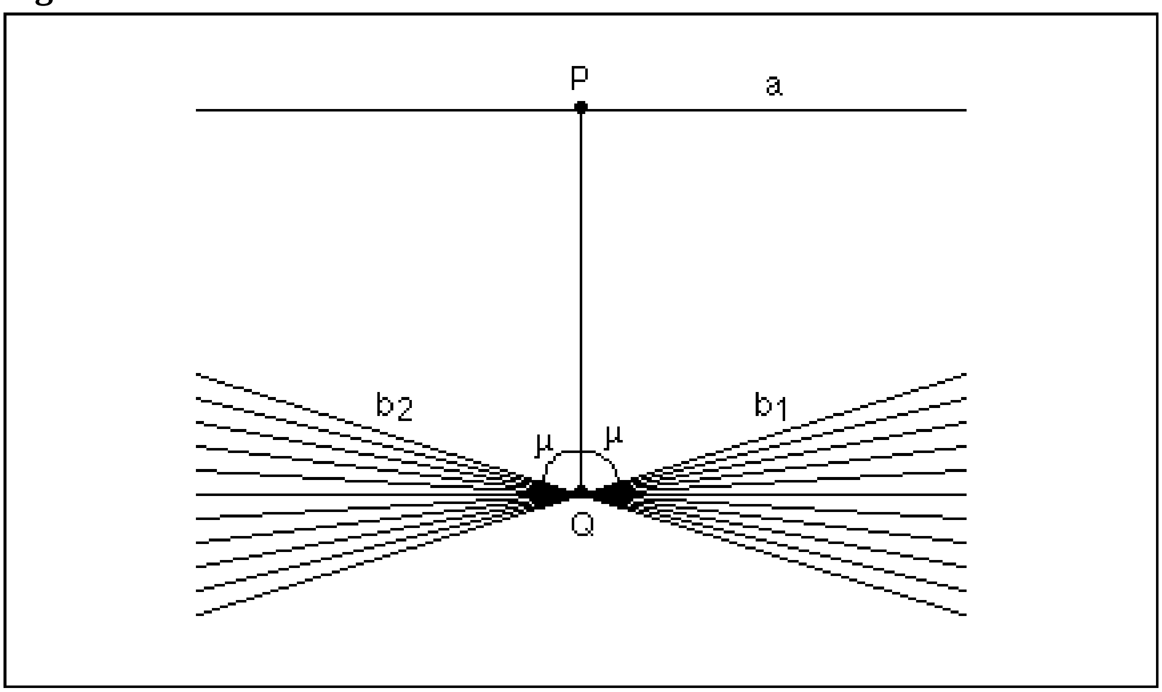
In the darker ages that followed, Euclid’s sense of mathematical freedom was lost and philosophers and mathematicians expected geometry to rest on self-evident grounds. Now, if a is perpendicular and b is almost perpendicular to PQ, a and b approach each other very slowly on one side of PQ and it is not self-evident that they must eventually meet somewhere on that side. After all, the hyperbole indefinitely approaches its asymptotes and yet, demonstrably, never meets them. Through the centuries, several authors demanded---and attempted---a proof of Euclid’s Postulate. John Wallis (b. 1616, d. 1703) derived it from the assumption that there are polygons of different sizes that have the same shape. But then this assumption needs proof in turn. Girolamo Saccheri (b. 1667, d. 1733) tried *reductio*. He inferred a long series of propositions from the negation of Euclid’s Postulate, until he reached one which he pronounced "repugnant to the nature of the straight line". But Saccheri’s understanding of this "nature" was nourished by Euclidean geometry and his conclusion begged the question.

In the 1820’s, Nikolai I. Lobachevsky (b. 1793, d. 1856) and Janos Bolyai (b. 1802, d. 1860) independently tackled this question in a radically new way. Lobachevsky built on the negation of Euclid’s Postulate an alternative system of geometry, which he dubbed "imaginary" and checked---inconclusively---for validity at the astronomical scale. Bolyai excised the postulate from Euclid’s system; the remaining rump is the "absolute geometry", which can be further specified by adding to it either Euclid’s Postulate or its negation. From the 1790’s Carl Friedrich Gauss (b. 1777, d. 1855) had been

working on the subject in the same direction, but he refrained from publishing for fear of scandal. Since Lobachevsky was the first to publish, the system of geometry based on the said "absolute geometry" plus the negation of Euclid's Postulate is properly called *Lobachevskian geometry*.

The construction introduced above to explain Euclid's Postulate can also be used for elucidating its negation. Draw the straight line a through point P at right angles with the segment PQ. If Euclid's Postulate is denied, there are countless straight lines through Q, coplanar with a, that make acute angles with PQ but never meet a. Consider the set of real numbers which are the magnitudes of these acute angles. Let the greatest lower bound of this set be μ. Evidently, μ > 0. There are exactly two straight lines through Q, coplanar with a, that make an angle of size μ with PQ. (See Figure 2.) Call them b_1 and b_2. Neither b_1 nor b_2 meets a, but a meets every line through Q that is coplanar with a and makes with PQ an angle less than μ. Gauss, Lobachevsky and Bolyai---unbeknownst to each other---coincided in calling b_1 and b_2 *the parallels* to a through Q. μ is called the angle of parallellism for segment PQ. Its size depends on the length of PQ, and decreases as the latter increases.

Figure 2



Suppose that the angle of parallellism for PQ is one half a right angle. In this case, b_1 and b_2 make a right angle at Q and we thus have two mutually perpendicular straight lines on the same plane as a, which fail to meet a.

Lobachevsky's geometry abounds in surprising theorems (many of which had already been found by Saccheri). Here are a few: The three interior angles of a triangle add up to *less* than two right angles. The difference or "defect" is proportional to the triangle's area. Hence, in Lobachevskian geometry, similar triangles are congruent. Moreover, if a triangle is divided into smaller triangles, the defect of the whole equals the sum of the defects of the parts. Since the defect cannot be greater than two right angles, the

area of triangles has a finite maximum. If a quadrilateral, by construction, has three right angles, the fourth angle is necessarily acute. Thus, in Lobachevskian geometry there are no rectangles.

There is a simple formal correspondence between the equations of Lobachevskian trigonometry and those of standard spherical trigonometry. Based on it, Lobachevsky argued that any contradiction arising in his geometry would inevitably be matched by a contradiction in Euclidean geometry. This is apparently the earliest instance of a proof of relative consistency, by which a theory is shown to be at least as consistent as another one (whose consistency is taken for granted).

Lobachevskian geometry received little attention before the late 1860’s. When philosophers finally took notice of it, their opinions were divided. Some regarded it as a formal exercise in logical deduction, with no physical or philosophical significance, which employed ordinary words---such as ‘straight’ and ‘plane’---with a covertly changed meaning. Others welcomed it as sufficient proof that, contrary to the influential thesis of Kant, Euclidean geometry does not convey any prerequisites of human experience and that the geometrical structure of physical space is open to experimental inquiry. Still others agreed that Non-Euclidean geometries were legitimate alternatives, but pointed out that the design and interpretation of physical experiments generally presupposes a definite geometry and that this role has been preempted by Euclid’s system.

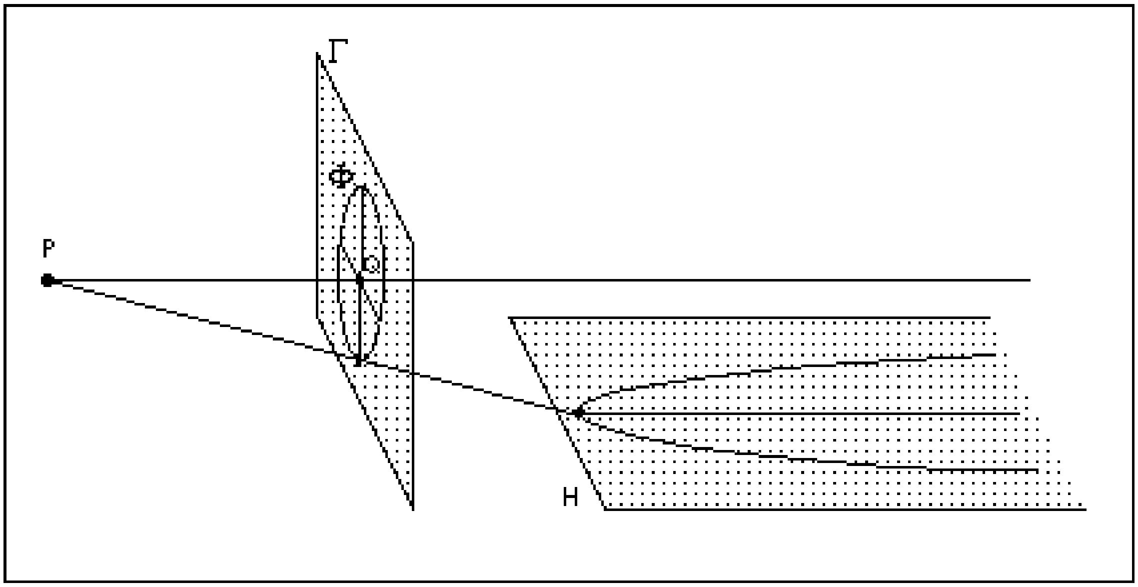
No matter what philosophers might say, for mathematicians Lobachevskian geometry would probably have been no more than an odd curiosity, if a niche had not been found for it within both projective and differential geometry, the two main currents of nineteenth-century geometrical research (§§ 2 and 5).

2. Projective geometry

Today projective geometry does not play a big role in mathematics, but in the late nineteenth century it came to be synonymous with modern geometry. Projective methods had been employed by Desargues (b. 1591, d. 1661) and Pascal (b. 1623, d. 1662), but were later eclipsed by Descartes’s method of coordinates. They prospered, however, after Jean-Victor Poncelet (b. 1788, d. 1867) showed that the projective properties of figures furnished grounds of proof that were at least as powerful as, and certainly more geometric than the Cartesian procedure of setting up and solving equations between numbers representing points.

Projective properties are those preserved by projections. Take, for example, two planes Γ and H and a point P outside them. Let Φ be any figure on Γ. Draw straight lines from P through each point of Φ. The figure formed by the points where these lines meet H is the projection of Φ on H from P. Generally this figure will differ from Φ in size and shape. But the projection of any number of straight lines on Γ meeting each other at certain points generally consists of an equal number of straight lines on H meeting respectively at the projection of those points. What happens, however, if the straight line joining P with a some point Q of Γ never meets H, because PQ happens to lie on a plane parallel to H? (See Figure 3.)

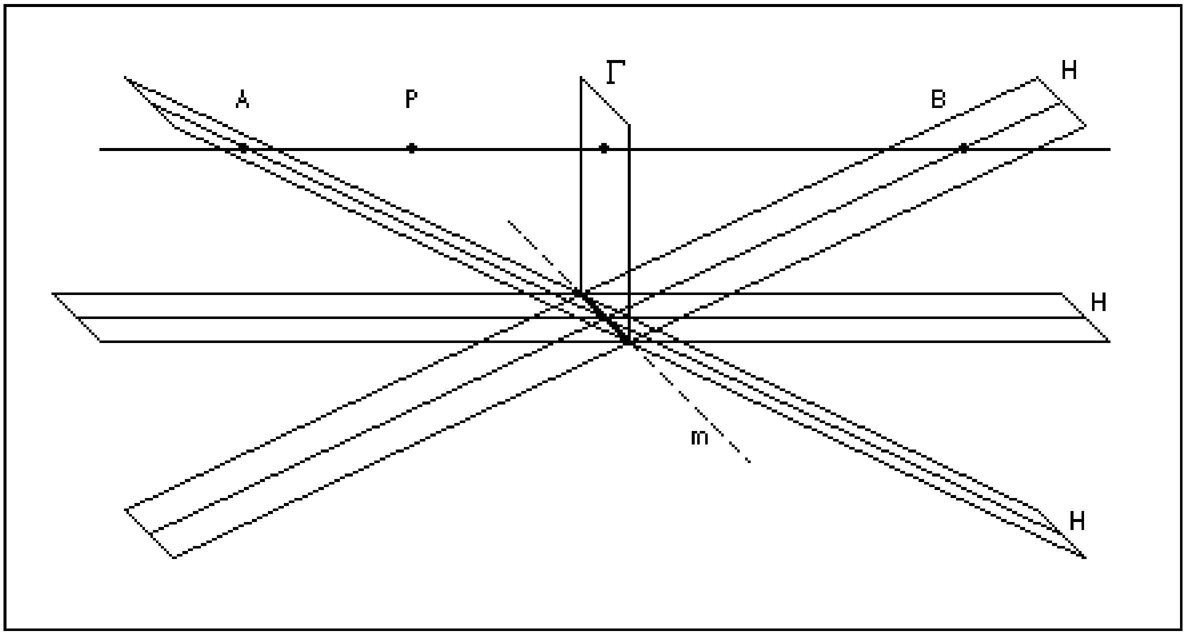
Figure 3



To obviate such irksome exceptions, projective geometry added to each straight line in space an ideal point, shared by every line parallel to it. Continuity requires then that all ideal points lie on a single ideal plane, which meets each family of parallel planes along a different ideal line. Fundamentalists may shudder at this seemingly wanton multiplication of entities. However, it had been practised in arithmetic for centuries, as the initial stock of natural numbers 1, 2, 3, ... , was supplemented with zero, the negative integers, the non-integral rationals, the irrationals, and the so-called imaginary numbers.

The points of a straight line stand in mutual relations of neighborhood and order. To see how the ideal point fits into these relations let H rotate continually about the straight line *m* where it intersects Γ. (See Figure 4.) When H is parallel to PQ---say, at time t---the projection of Q on H from P is the ideal point of the straight line through P and Q. Right before t the said projection is an ordinary point of H, very far from *m*. Right after t the projection is again an ordinary point of H, very far from *m*, but at the opposite end of the plane. Studying the continuous displacement of the projection during a short time interval surrounding t, one concludes that if A and B are any two points of H that stand, respectively, on either side of *m*, the ideal point of the straight line through A and B must be placed between A and B. Thus, in projective geometry, the points of a straight line are ordered cyclically, i.e., like the points of a circle. As a result of this, the neighborhood relations among points in projective space and on projective planes differ drastically from those familiar from standard geometry, and are highly counterintuitive. It is fair to say that projective geometry signified a much deeper and far-reaching revolution in human thought than did the mere denial of Euclid’s Postulate.

Figure 4



In the new setting, the projective properties of figures can be defined unexceptionably. A one-one mapping f of projective space onto itself is a *collineation* if it sends any three collinear points A, B, and C, to three points (A), (B), and (C), which are collinear too. Projective properties (and relations) are those which are preserved by collineations. Here are a few examples of projective properties. Of three or more points: to lie on the same straight line; to lie on the same plane. Of three or more straight lines: to meet at the same point; to lie on the same plane. Of three or more planes: to intersect along the same straight line; to share the same point. Of curves: to be a conic. Of surfaces: to be a quadric.

3. Klein's Erlangen program

In a booklet issued when he joined the faculty at Erlangen (1872), Felix Klein (b. 1849, d. 1925) took stock of the enormous growth and diversification of geometry and proposed a standpoint from which its many branches could be organized into a system. From this standpoint, the task of a branch of geometry can be stated thus:

> Given a manifold and a group of transformations of the manifold, to study the manifold configurations with respect to those features which are not altered by the transformations of the group. (Klein 1893, p. 67)

In nineteenth-century mathematics, 'manifold' often designated what we now call a set, but Klein apparently had something more specific in mind:

> If *n* variables $x_1, \ldots, x_n$ are given, the ... value systems we obtain if we let the variables x independently take the real values from $-\infty$ to $+\infty$ constitute what we shall call ... *a manifold of n dimensions*. Each particular value system $(x_1, \ldots, x_n)$ is called an *element* of the manifold. (Klein 1873, p. 116)

If *S* is a manifold in either sense, by a *transformation of S* we mean a one-one mapping of *S* onto itself. It is clear that

> (i) If T_1 and T_2 are transformations of *S*, the composite mapping $T_2 \circ T_1$, which consists of T_1 followed by T_2, is also a transformation of *S*;
>
> (ii) the composition of transformations is associative, so that, if T_1, T_2 and T_3 are transformations of *S*, $(T_3 \circ T_2) \circ T_1 = T_3 \circ (T_2 \circ T_1)$;
>
> (iii) the identity mapping I that sends each point of *S* to itself is a transformation of *S* such that, for any transformation T, $T \circ I = I \circ T = T$;
>
> (iv) for every transformation T there is a transformation T^{-1}, the *inverse* of T, such that $T^{-1} \circ T = I$ (T^{-1} sends each point of *S* back to where it was brought from by T).

By virtue of conditions (i)-(iv), the transformations of *S* form a *group* G_S in the precise sense that this term has in algebra. G_S includes subgroups, i.e., subsets which contain I and satisfy conditions (i) and (iv). If H is a subgroup of G_S and Φ is a feature of *S*, or of its elements or parts, which is not affected by the transformations of Φ, we say that Φ is H-invariant. The only G_S-invariant is the cardinality of *S* (i.e., the number of elements in the manifold). On the other hand, the group {I}, consisting of the identity alone, trivially preserves every conceivable feature. Between these two extremes there can be many different subgroups with all sorts of interesting invariants, depending on the respective group structure. If *S* is not an arbitrary (structureless) set, but a numerical manifold as described by Klein, it inherits structure from the real number field, which contributes to characterize the different subgroups of G_S and their invariants. Thus, the group of continuous transformations preserves the topological properties (neighborhood relations), and the group of linear transformations preserves the projective properties.

Can metric properties be fixed in this way? Traditionally one defines the distance between two points $(x_1, ... ,x_n)$ and $(y_1, ... ,y_n)$ of a numerical manifold as the positive square root of $(x_1 - y_1)^2 + ... + (x_n - y_n)^2$. The group of isometries consists of the transformations that preserve this function. However, this is no more than a convention, adopted to ensure that the geometry is Euclidean. Using projective geometry, Klein thought of something better. No real-valued function of point pairs, defined on all projective space, can be an invariant of the projective group, but there is a function of collinear point quadruples, called the cross-ratio, which is such an invariant. Drawing on work by Arthur Cayley (b. 1821, d. 1895), Klein (1871, 1873) employed the cross-ratio for defining projectively invariant distance functions on specific regions of projective space. Surprisingly, two of the three distance functions defined by Klein agree respectively, on their respective regions, with the distance functions of Euclid and Lobachevsky, while the third one determines another variety of non-Euclidean geometry which Klein dubbed ‘elliptic’. (In elliptic geometry every straight line meets every other, and the three internal angles of a triangle always

add up to more than two right angles. Klein’s names for the geometries of Euclid and Lobachevsky were ‘parabolic’ and ‘hyperbolic’, respectively.)

This is how Klein’s approach works for Lobachevskian geometry on the plane. Let κ be a real conic---a conic comprising only real points---on the projective plane. Let $G\kappa$ be the set of all collineations that map κ onto itself. $G\kappa$ is a subgroup of the projective group. Consider now the cross-ratio of point quadruples $<P_1,P_2,P_3,P_4>$ such that P_3 and P_4 belong to κ, while P_1 and P_2 range over the interior Int(κ) of κ. ($P \in$ Int(κ) if and only if P is a real point and no real tangent to κ passes through P.) P_3 and P_4 are the points where the straight line through P_1 and P_2 meets κ, so the said cross-ratio may be regarded as a function of the point pair $<P_1,P_2>$, say, $f\kappa(P_1,P_2)$. The function $f\kappa$ is clearly $G\kappa$-invariant. Put $d\kappa(P_1,P_2) = c \log f\kappa(P_1,P_2)$, where *c* is an arbitrary real-valued constant, different from 0, and log x denotes the principal value of the natural logarithm of x. Klein was able to show that $d\kappa$ behaves precisely like a Lobachevskian distance function on Int(κ). In other words, every theorem of Lobachevskian geometry holds for suitable figures formed from points of Int(κ), if the distance between any two of these points is given by the function $d\kappa$. Consider, for instance, four points P_1,P_2,P_3, and P_4 in Int(κ), such that $d\kappa(P_1,P_2) = d\kappa(P_2,P_3) = d\kappa(P_3,P_4) = d\kappa(P_4,P_1)$. They are the vertices of a Lobachevskian equilateral quadrilateral Q, which can have at most three right angles, in which case the fourth interior angle of Q must be acute. (Where ‘right angle’ means, as usual, an angle equal to its adjacent angle, and two angles in Int(κ) are said to be equal if one is the image of the other by a transformation of group $G\kappa$).

If κ stands for a different sort of conic, not an ordinary real one, the function $d\kappa$ obtained by the above procedure behaves on suitably defined regions of the projective plane like a Euclidian distance function or like the distance function of elliptic geometry (this depends on the nature of the conic κ). Thus, depending on whether κ belongs to one or the other of three kinds of conic, the group of collineations that map κ onto itself is structurally identical with one of the three groups of Lobachevskian, Euclidean, or elliptic isometries. Similar results hold for the three-dimensional case, with κ a quadric surface.

Klein’s result led Bertrand Russell (b. 1873, d. 1970) to assert, in his neo-Kantian book on the foundations of geometry (1897), that the general "form of externality" is disclosed to us a priori in projective geometry, but its metric structure---which can *only* be Lobachevskian, Euclidean or elliptic---must be determined a posteriori by experiment. Henri Poincaré (b. 1854, d. 1912) took a more radical stance: If geometry is nothing but the study of a group,

> one may say that the truth of the geometry of Euclid is not incompatible with the truth of the geometry of Lobachevsky, for the existence of a group is not incompatible with that of another group. (Poincaré 1887, p. 290)

The application to physics is immediate: "Among all possible groups we have chosen one in particular, in order to refer to it all physical phenomena, just as we choose three coordinate axes in order to refer to them a geometrical figure" (ibid., p. 291). The choice of this particular group is motivated by its mathematical simplicity, but also by the fact that "there exist in nature some remarkable bodies which are

called *solids*, and experience tells us that the different possible movements of these bodies are related to one another much in the same way as the different operations of the chosen group" (ibid.).

Klein's group-theoretical view of geometry enjoyed much favor among mathematicians and philosophers. It achieved a major success when Minkowski (1909) showed that the gist of Einstein's special theory of relativity was the (spacetime) geometry of the Lorentz group. But Klein's Erlangen program failed to cover the differential geometry of Riemann (§5), which Einstein (1915, 1916) placed at the core of his general theory of relativity.

4. Axiomatics perfected

According to Aristotle, scientific knowledge (*episteme*) must be expressed in statements that follow deductively from a finite list of self-evident statements (axioms) and only employ terms defined from a finite list of self-understood terms (primitives). For over two millennia it was taken for granted that Aristotle's ideal is actually realized in Euclid's *Elements*. In fact, there is a logical gap already in Euclid I.1 (the solution of this problem rests on an unstated assumption of continuity) and it is not clear that Euclid regarded his postulates as self-evident (by calling them "requests" he suggested he did not). The idea of securing knowledge by logical deduction from unquestionable principles had a powerful fascination for modern scientists such as Galileo and Newton, both of whom fondly practised axiomatics, at any rate as a literary form, like Spinoza in his *Ethics*. Still, a truly satisfactory and, if one may say so, serious instance of axiomatization of a branch of knowledge was not available in print until 1882, when Moritz Pasch (b. 1843, d. 1930) published his *Lectures on Modern Geometry*.

Pasch viewed geometry as a natural science, whose successful utilization by other sciences and in practical life rests "exclusively on the fact that geometrical concepts originally agreed exactly with empirical objects" (Pasch 1882, p. iii). Geometry distinguishes itself from other natural sciences because it obtains only very few concepts and laws directly from experience, and aims at obtaining from them the laws of more complex phenomena by purely deductive means. The empirical foundation of geometry was encapsulated by Pasch in a core of basic concepts and basic statements or axioms. The basic concepts refer to the shape, size and reciprocal position of bodies. They are not defined, for no definition could replace the "exhibition of appropriate natural objects," which is the only road to understanding such simple, irreducible notions (ibid., p. 16). All other geometric concepts must be ultimately defined in terms of the basic ones. The basic concepts are connected to one another by the axioms, which "state what has been observed in certain very simple diagrams" (p. 43). All other geometric statements must be proved from the axioms by the strictest deductive methods. Everything that is needed to prove them must be recorded, without exception, in the axioms. These must therefore embody the whole empirical material elaborated by geometry, so that "after they are established it is no longer necessary to resort to sense perceptions" (p. 17). "Every conclusion which occurs in a proof must find its confirmation in the diagram, but it is not justified by the diagram, but by a definite earlier statement (or definition)" (p. 43). Pasch understood clearly the implications of his method. He writes (p. 98):

> If geometry is to be truly deductive, the process of inference must be independent in all its

> parts from the *meaning* of the geometric concepts, just as it must be independent from the diagrams. All that need be considered are the *relations* between the geometric concepts, recorded in the statements and definitions. In the course of deduction it is both permitted and useful to bear in mind the meaning of the geometric concepts that occur in it, but *it is not at all necessary*. Indeed, when it actually beomes necessary, this shows that there is a gap in the proof, and---if the gap cannot be eliminated by modifying the argument---that the premises are too weak to support it.

Pasch's *Lectures on Modern Geometry* dealt with projective geometry. The first rigorous axiomatization of Euclidean geometry---*Foundations of Geometry* by David Hilbert (b. 1862, d. 1943)---appeared in 1899 and exercised enormous influence on twentieth century mathematics and philosophy. Hilbert invites the reader to consider three arbitrary collections of objects, which he calls 'points', 'straight lines' and 'planes', and five undefined relations between (i) a point and a straight line, (ii) a straight line and a plane, (iii) three points, (iv) two pairs of points ('segments') and (v) two equivalence classes of point triples ('angles'). The conditions prescribed in Hilbert's 20 axioms---including the Axiom of Completeness added in the second edition---are sufficient to characterize the said objects and relations up to isomorphism. Isomorphism---i.e., structural equivalence---can hold, however, between different, intuitively disparate, systems of objects. Hilbert availed himself of this feature of axiomatic theories for studying the independence of some axioms from the rest. To prove it he proposed actual instances (models) of the structure determined by all axioms but one, plus the negation of the omitted one. Frege complained that the geometric axioms retained in these exercises could be applied to Hilbert's far-fetched models only by tampering with the natural meaning of words (cf. Alice's conversation with Humpty Dumpty). Hilbert replied on 29 December 1899:

> Every theory is only a scaffolding or schema of concepts together with their necessary mutual relations, and the basic elements can be conceived in any way you wish. If I take for my points any system of things, for example, the system love, law, chimney-sweep, ... and I just assume all my axioms as relations between these things, my theorems---for example, the theorem of Pythagoras---also hold of these things. ... This feature of theories can never be a shortcoming and is in any case inevitable.

All this follows, of course, from the very nature of axiomatics, as explained in the passage quoted from Pasch. Indeed, such truth-preserving semantic permutations were no news in geometry after Gergonne (1771-1859) drew attention in 1825 to the following *principle of duality:* Any true statement of projective plane geometry gives rise to another, equally true, dual statement obtained by substituting 'point' for 'line', 'collinear' for 'concurrent', 'meet' for 'join', and vice versa, wherever these words occur in the former. (In projective space geometry, duality holds for points and planes.) The same result is secured, of course, by exchanging not the words, but their meanings.

5. The differential geometry of Riemann

In a lecture "On the hypotheses that lie at the foundation of geometry", delivered to the Faculty of

Philosophy at Göttingen in 1854 and posthumously published in 1867, Bernhard Riemann (b. 1826, d. 1866) presented some radically innovative views on this matter. He noted that the measurable properties of a discrete manifold can be readily determined by counting. (Think of the population of a country, and such properties as the median, average and aggregate income, or the proportion of drug-addicts and born-again Christians.) But continuous manifolds do not admit this approach. In particular, the metric properties of physical space, which are the subject of geometry, must depend on the binding forces that act on it. The distance between two points in space can be ascertained with a rod, or a tape, or by optical means, and the result depends essentially on the physical behavior of the instruments used. Up to now, the metric properties of space have been successfully described in accordance with Euclidean geometry. However, "the empirical concepts on which the metric determinations of space are based --- the concepts of a rigid body and a light ray--- lose their validity in the infinitely small; it is therefore quite likely that the metric relations of space in the infinitely small do not agree with the assumptions of geometry, and in fact one would have to accept this as soon as the phenomena can thereby be explained in a simpler way" (Riemann 1854, p. 149). To prepare physicists for this eventuality, Riemann proposed a more general conception of geometry. Riemann's basic scheme makes allowance for much greater generality than he actually reaches for; but, in his judgment, it should be enough for the time being to characterize the geometry of continuous manifolds in such a way that it agrees optimally with Euclidean geometry on a small neighborhood of each point.

Riemann extends to *n* dimensions the methods employed by Gauss (1828) in his study of the intrinsic geometry of curved surfaces embedded in Euclidean space (called "intrinsic" because it describes the metric properties that the surfaces display by themselves, independently of the way they lie in space). Looking back at Gauss's work one gets a better intuitive feel for Riemann's concepts (see Torretti 1978, pp. 68-82). However, for the sake of conciseness and perspicuity, it is advisable to look forward and to avail oneself of certain concepts introduced by later mathematicians as they tried to make sense of Riemann's proposal. Consider the following modern formulation of Riemann's theory, in this Supplement.

In his study of curved surfaces, Gauss introduced a real-valued function, the *Gaussian curvature*, which measures a surface's local deviation from flatness in terms of the surface's intrinsic geometry. Riemann extended this concept of curvature to Riemannian n-manifolds. By using his extended concept of curvature, he was able to characterize with great elegance the metric manifolds in which all figures can freely move around without changing their size and shape. They are the Riemannian manifolds of *constant curvature*. This idea can be nicely combined with Klein's classification of metric geometries. Regarded as Riemannian 3-manifolds, Euclidean space has constant zero curvature, Lobachevskian space has constant negative curvature, and elliptic space has constant positive curvature. Pursuant to the Erlangen Program, each of these geometries of constant curvature is characterized by its own group of isometries. But Klein's conception is too narrow to embrace all Riemannian geometries, which include spaces of variable curvature. Indeed, in the general case, the group of isometries of a Riemannian n-manifold is the trivial group consisting of the identity alone, whose structure conveys no information at all about the respective geometry.

Bibliography

Primary sources

- Bolyai, J., 1832. *Scientia absoluta spatii*. Appendix to Bolyai, F., *Tentamen juventutem studiosam in elementa matheseos purae elementis ac sublimioris, methodo intuitiva, evidentiaque huic propria, introducendi*, Tomus Primus. Maros Vasarhely: J. et S. Kali. (English translation by G. B. Halsted printed as a supplement to Bonola 1955.)
- Cayley, Arthur, 1859. "A sixth memoir upon quantics," *Philosophical Transactions of the Royal Society of London*. **149**: 61-90.
- Einstein, A., 1915. "Die Feldgleichungen der Gravitation," *K. Preussische Akademie der Wissenschaften. Sitzungsberichte*, pp. 844-847.
- Einstein, A., 1916. "Die Grundlagen der allgemeinen Relativitätstheorie," *Annalen der Physik*. **49**: 769-822.
- Euclides, 1883-88. *Elementa*. Edidit I. L. Heiberg. Leipzig: B. G. Teubner. 5 vols. (For English translation, see below under Heath).
- Gauss, C. F., 1828. *Disquisitiones generales circa superficies curvas*. Göttingen: Dieterich. (English translation by A. Hiltebietel and J. Morehead: Hewlett, NY, Raven Press, 1965.)
- Hilbert, D., 1899. "Die Grundlagen der Geometrie," in *Festschrift zur Feier der Enthüllung des Gauss-Weber Denkmals*. Leipzig: B.G. Teubner. Pp. 3-92.
- Hilbert, D., 1968. *Grundlagen der Geometrie*, mit Supplementen von P. Bernays. Zehnte Auflage. Stuttgart: Teubner. (Tenth, revised edition of Hilbert 1899.)
- Klein, F., 1871. "Über die sogenannte Nicht-Euklidische Geometrie," *Mathematische Annalen*. **4**: 573-625.
- Klein, F., 1872. *Vergleichende Betrachtungen über neuere geometrische Forschungen*. Erlangen: A. Duchert.
- Klein, F., 1873. "Über die sogenannte Nicht-Euklidische Geometrie (Zweiter Aufsatz)," *Mathematische Annalen*. **6**: 112-145.
- Klein, F., 1893. "Vergleichende Betrachtungen über neuere geometrische Forschungen," *Mathematische Annalen*. **43**: 63-100. (Revised version of Klein 1872).
- Klein, F., 1911. "Über die geometrischen Grundlagen der Lorentz-Gruppe," *Physikalische Zeitschrift*. **12**: 17-27.
- Lobachevsky, N. I., 1837. "Géométrie imaginaire," *Journal für die reine und angewandte Mathematik*. **17**: 295-320.
- Lobachevsky, N. I., 1840. *Geometrische Untersuchungen zur Theorie der Parallellinien*. Berlin: F. Fincke. (English translation by G. B. Halsted printed as a supplement to Bonola 1955.)
- Lobachevsky, N. I., 1856. *Pangéométrie ou précis de géométrie fondée sur une théorie générale et rigoureuse des parallèles*. Kazan: Universitet.
- Locke, J., 1690. *An Essay concerning Humane Understanding*. In four books. London: Printed for Thomas Basset and sold by Edward Mory. (Published anonymously; the author's name was added in the second edition).
- Minkowski, H., 1909. "Raum und Zeit," *Physikalische Zeitschrift*. **10**: 104-111.

- Pasch, M., 1882. *Vorlesungen über neueren Geometrie*. Leipzig: Teubner.
- Poincaré, H., 1887. "Sur les hipothèses fondamentales de la géométrie," *Bulletin de la Société mathématique de France*. **15**: 203-216.
- Poncelet, J. V., 1822. *Traité des propriétés projectives des figures*. Paris: Bachelier.
- Ricci, G. and T. Levi-Cività, 1901. "Méthodes de calcul différentiel absolu et leurs applications," *Mathematische Annalen*. **54**: 125-201.
- Riemann, B., 1854. "Über die Hypothesen, welche der Geometrie zugrunde liegen," *Abhandlungen der Kgl. Gesellschaft der Wissenschaften zu Göttingen*. **13** (1867): 133-152. (For English translation, see below under Spivak.)
- Riemann, B., 1861. "Commentatio mathematica, qua respondere tentatur quaestioni ab illustrissima Acad. Parisiensi propositae," in *Bernhard Riemanns gesammelte mathematische Werke und wissenschaftlicher Nachlass*, Leipzig: Teubner, 1876, pp. 391-404.
- Russell, B., 1897. *An Essay on the Foundations of Geometry*. Cambridge: Cambridge University Press. (Unaltered reprint: New York, Dover, 1956.)
- Saccheri, G. 1733. *Euclides ab omni nævo vindicatus sive conatus geometricus quo stabiliuntur prima ipsa universæ geometriæ principia*. Mediolani: Ex Typographia Pauli Antonii Montani. (Reprint, with facing English translation by G. B. Halsted: New York, Chelsea, 1986.)

Suggested readings

- Blumenthal, L. M., 1961. *A Modern View of Geometry*. San Francisco: Freeman.
- Boi, Luciano, 1995. *Le problème mathématique de l'espace: Une quête de l'intelligible*. Berlin: Springer.
- Bonola, R., 1955. *Non-Euclidean Geometry: A critical and historical study of its development*. English translation with additional appendices by H.S. Carslaw. New York: Dover.
- Freudenthal, H., 1957. "Zur Geschichte der Grundlagen der Geometrie," *Nieuw Archief vor Wiskunde*. **5**: 105-142.
- Freudenthal, H., 1960. "Die Grundlagen der Geometrie um die Wende des 19. Jahrhunderts," *Mathematisch-physikalische Semesterbericht*. **7**: 2-25.
- Heath, T. L., 1956. *The Thirteen Books of Euclid's Elements*. Translated from the text of Heiberg with introduction and commentary. Second edition, revised with additions. New York: Dover. 3 vols.
- Nagel, E., 1939. "The formation of modern conceptions of formal logic in the development of geometry," *Osiris*. **7**: 142-224.
- O'Neill, B., 1983. *Semi-Riemannian Geometry with Applications to Relativity*. New York: Academic Press.
- Rosenfeld, B. A., 1988. *A History of Non-Euclidean Geometry: Evolution of the Concept of a Geometric Space*. Translated by Abe Shenitzer. New York: Springer.
- Spivak, M., 1979. *A Comprehensive Introduction to Differential Geometry*. Second edition. Berkeley: Publish or Perish. 5 vols. (Contains an excellent English translation, with mathematical commentary, of Riemann's lecture "On the hypotheses that lie at the foundation of geometry"; see Vol. 2, pp. 135ff.)
- Torretti, R., 1978. *Philosophy of Geometry from Riemann to Poincaré*. Dordrecht: Reidel.

(Corrected reprint: Dordrecht, Reidel, 1984).
- Trudeau, R. J., 1987. *The Non-Euclidean Revolution*. Boston: Birkhäuser.
- Winnie, J. W., 1986. "Invariants and objectivity: A theory with applications to relativity and geometry". In R. G. Colodny, ed., *From Quarks to Quasars*. Pittsburgh: Pittsburgh University Press. Pp. 71-180.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

a priori justification and knowledge | Hilbert, David | Kant, Immanuel | space and time: the hole argument

Acknowledgements

I thank John Norton for the illustrations and for ideas leading to a better presentation of some mathematical concepts. I am also very grateful to Edward Zalta for his painstaking editorial work and for having identified and firmly rejected a murky passage in the first version of this article.

Roberto Torretti
Universidad de Chile
cordua@terra.cl

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 26, 1999
Content last modified: July 26, 1999

Stanford Encyclopedia of Philosophy
Supplement to Nineteenth Century Geometry

A Modern Formulation of Riemann’s Theory

The multidimensional continua that Riemann was concerned with are essentially instances of what is now known as a real *n-dimensional smooth manifold*. For brevity, call them ‘n-*manifolds*’. (For example, a sphere, a Möbius strip and the surface of a Henry Moore sculpture may be regarded as 2-manifolds; the spacetimes models of current cosmology are 4-manifolds.) Conditions (a) -- (c) provide a full characterization of n-manifolds.

> (a) An n-manifold **M** is a set of points that can be pieced together from partially overlapping patches, such that every point of **M** lies in at least one patch.
>
> (b) **M** is endowed with a neighborhood structure (a topology) such that, if U is a patch of **M**, there is a continuous one-one mapping *f* of U onto some region of $\mathbf{R}^n$, with continuous inverse f^{-1}. ($\mathbf{R}^n$ denotes here the collection of all real number n-tuples, with the standard topology generated by the open balls.) *f* is a coordinate system or *chart* of **M**; the k-th number in the n-tuple assigned by a chart *f* to a point P in *f*’s patch is called the k-th coordinate of P by *f*; the k-th coordinate function of chart *f* is the real-valued function that assigns to each point of the patch its k-th coordinate by *f*.
>
> (c) There is a collection **A** of charts of **M**, which contains at least one chart defined on each patch of **M** and is such that, if g and h belong to **A**, the composite mappings $h \circ g^{-1}$ and $g \circ h^{-1}$ --- known as *coordinate transformations* --- are differentiable to every order wherever they are well defined. (Denote the real number n-tuple $<a_1, \ldots , a_n>$ by **a**. The mapping $h \circ g^{-1}$ is well defined at **a** if **a** is the valued assigned by g to some point P of **M** to which h also assigns a value. Suppose that the latter value $h(P) = <b_1, \ldots ,b_n> = \mathbf{b}$; then, $\mathbf{b} = h \circ g^{-1}(\mathbf{a})$. Since $h \circ g^{-1}$ maps a region of $\mathbf{R}^n$ into $\mathbf{R}^n$, it makes sense to say that $h \circ g^{-1}$ is differentiable.) Such a collection **A** is called an *atlas*.[*]

It is the pair $<\mathbf{M}, \mathbf{A}>$ that, strictly speaking, is an n-manifold, in the sense defined above. If $<\mathbf{M}_1,\mathbf{A}_1>$ and $<\mathbf{M}_2,\mathbf{A}_2>$ are an n-manifold and an m-manifold, respectively, it makes good sense to say that a mapping *f* of $\mathbf{M}_1$ into $\mathbf{M}_2$ is differentiable at a point P of $\mathbf{M}_1$ if, for a chart h defined at P and a chart g defined at *f*(P), the composite mapping $g \circ f \circ h^{-1}$ is differentiable at h(P). (Condition (c) implies that the fulfillment of this requirement does not depend on the choice of h and g.) *f* is differentiable if it is differentiable at every point of $\mathbf{M}_1$.

Let <**M**,**A**> be an n-manifold. To each point P of **M** one associates a vector space, which is known as the *tangent space at P* and is denoted by T_P**M**. The idea is based on the intuitive notion of a plane tangent to a surface at a given point. It can be constructed as follows. Let γ be a one-one differentiable mapping of a real open interval **I** into **M**. We can think of the successive values of γ as forming the path of a point that moves through **M** during a time interval represented by **I**. We call γ a *curve* in **M** (parametrized by u $\in$ **I**). Put $\gamma(t_0)$ = P for a fixed number t_0 in **I**. Consider the collection **F**(P) of all differentiable real-valued functions defined on some neighborhood of P. With the ordinary operations of function addition and multiplication by a constant, **F**(P) has the structure of a vector space. Each function *f* in **F**(P) varies smoothly with u, along the path of γ, in some neighborhood of P. Its rate of variation at P = $\gamma(t_0)$ is properly expressed by the derivative $d(f\circ\gamma)/du$ at u = t_0. As *f* ranges over **F**(P), the value of $d(f\circ\gamma)/du$ at u = t_0 is apt to vary in **R**. So we have here a mapping of **F**(P) into **R**, which we denote by $\dot{\gamma}(u)$. It is in fact a linear function and therefore a vector in the dual space **F***(P) of real-valued linear functions on **F**(P). Call it the *tangent* to γ at P. The tangents at P to all the curves whose paths go through P span an n-dimensional subspace of **F***(P). This subspace is, by definition, the tangent space T_P**M**. The tangent spaces at all points of an n-manifold **M** can be bundled together in a natural way into a 2n-manifold T**M**. The projection mapping π of T**M** onto **M** assigns to each tangent vector **v** in T_P**M** the point π(**v**) at which **v** is tangent to **M**. The structure <T**M**,**M**,π> is the tangent bundle over **M**. A vector field on **M** is a section of T**M**, i.e., a differentiable mapping ***f*** of **M** into T**M** such that $\pi \circ$ ***f*** sends each point P of **M** to itself; such a mapping obviously assigns to P a vector in T_P**M**.

Any vector space **V** is automatically associated with other vector spaces, such as the dual space **V*** of linear functions on **V**, and the diverse spaces of multilinear functions on **V**, on **V***, and on any possible combination of **V** and **V***. This holds, of course, for each tangent space of an n-manifold **M**. The dual of T_P**M** is known as the cotangent space at P. There is a natural way of bundling together the cotangent spaces of **M** into a 2n-manifold, the cotangent bundle. Generally speaking, all the vector spaces of a definite type associated with the tangent and cotangent spaces of **M** can be naturally bundled together into a k-manifolds (for suitable integers k, depending on the nature of the bundled items). A section of any of these bundles is a *tensor field* on **M** (of rank r, if the bundled objects are r-linear functions).

A *Riemannian metric* **g** on the n-manifold <**M**,**A**> is a tensor field of rank 2 on **M**. Thus, **g** assigns to each P in **M** a bilinear function $\mathbf{g}_P$ on T_P**M**. For any P in **M** and any vectors **v**, **w**, in T_P**M**, $\mathbf{g}_P$ must meet these requirements:

(i) $\mathbf{g}_P(\mathbf{v},\mathbf{w}) = \mathbf{g}_P(\mathbf{w},\mathbf{v})$ (symmetry)

(ii) $\mathbf{g}_P(\mathbf{v},\mathbf{w}) = 0$ for all vectors **w** in T_P**M** if and only if **v** is the 0-vector (non-degeneracy)

(iii) $\mathbf{g}_P(\mathbf{v},\mathbf{v}) > 0$ unless **v** is the 0-vector (positive definiteness).

It is worth noting that the so-called Lorentzian metrics defined by relativity theory on its spacetime models meet requirements (i) and (ii), but not (iii), and are therefore usually said to be *semi-Riemannian*.

The length $\lambda(\mathbf{v})$ of a vector **v** in $T_P\mathbf{M}$ is defined by $|\lambda(\mathbf{v})|^2 = \mathbf{g}_P(\mathbf{v},\mathbf{v})$. Let γ be a curve in **M**. Let $\dot{\gamma}(u)$ be the tangent to γ at the point $\gamma(u)$. The length of γ's path from $\gamma(a)$ to $\gamma(b)$ is measured by the integral

$$\int_a^b \lambda(\dot{\gamma}(u))du$$

Thus, in Riemannian geometry, the length of the tangent vector $\dot{\gamma}(u)$ bears witness to the advance of curve g as it passes through the point $\gamma(u)$. The definition of the length of a curve leads at once to the notion of a geodesic (or straightest) curve, which is characterized by the fact that its length is extremal; in other words, a geodesic is either the greatest or the shortest among all the curves that trace out neighboring paths between the same two points.

In his study of curved surfaces, Gauss introduced a real-valued function, the *Gaussian curvature*, which measures a surface's local deviation from flatness in terms of the surface's intrinsic geometry. Riemann extended this concept of curvature to Riemannian n-manifolds. He observed that each geodesic through a point in such a manifold is fully determined by its tangent vector at that point. Consider a point P in a Riemannian n-manifold <**M**,**A**,**g**> and two linearly independent vectors **v** and **w** in $T_P\mathbf{M}$. The geodesics determined by all linear combinations of **v** and **w** form a 2-manifold about P, with a definite Gaussian curvature $K_P(\mathbf{v},\mathbf{w})$ at P. The real number $K_P(\mathbf{v},\mathbf{w})$ measures the curvature of **M** at P in the 'surface direction' (Riemann 1854, p. 145) fixed by **v** and **w**. Riemann (1861) thought up a global mapping, depending on the metric **g**, that yields the said values $K_P(\mathbf{v},\mathbf{w})$ on appropriate arguments P, **v** and **w**. Nowadays this object is conceived as a tensor field of rank 4, which assigns to each point P in a Riemannian n-manifold <**M**,**A**,**g**> a 4-linear function on the tangent space $T_P\mathbf{M}$. It is therefore known as the *Riemann tensor*. Given the above definition of $K_P(\mathbf{v},\mathbf{w})$ it is clear that, if n = 2, the Riemann tensor reduces to the Gaussian curvature function.

Roberto Torretti
Universidad de Chile
cordua@rdc.cl

Return to Nineteenth Century Geometry

First published: July 26, 1999
Content last modified: July 26, 1999

Stanford Encyclopedia of Philosophy
Notes to Supplement: A Modern Formulation of Riemann's Theory

---

Notes

* The informal characterization of n-manifolds in the supplement is not altogether accurate and may cover some far-fetched monsters that we do not wish to cover with this concept. For readers who have a smattering of topology the following characterization is preferable.

> (a) Let **M** be a set of points. Pick a collection of partially overlapping subsets of **M**, or *patches*, such that every point of **M** lies in at least one patch.
>
> (b) If U is a patch of M, there is a one-one mapping *f* of U onto an open space of $\mathbf{R}^n$, whose inverse we denote by f^{-1}. ($\mathbf{R}^n$ denotes here the collection of all real number n-tuples, with the standard topology generated by the open balls.) f is a coordinate system or *chart* of **M**; the k-th number in the n-tuple assigned by a chart *f* to a point P in *f*'s domain is called the k-th coordinate of P by *f*; the k-th coordinate function of chart *f* is the real-valued function that assigns to each point of the patch its k-th coordinate by *f*.
>
> (c) There is a collection **A** of charts of **M** which contains at least one chart defined on each patch of **M** and is such that, if g and h belong to **A**, the composite mappings $h \circ g^{-1}$ and $g \circ h^{-1}$ --- known as *coordinate transformations* --- are differentiable to every order wherever they are well defined. (Denote the real number n-tuple $(a_1, \ldots, a_n)$ by **a**. $h \circ g^{-1}$ is well defined at **a** if **a** is the valued assigned by g to some point P of **M** to which h also assigns a value. Suppose that the latter value $h(P) = (b_1, \ldots, b_n) = \mathbf{b}$; then, $\mathbf{b} = h \circ g^{-1}(\mathbf{a})$. Since $h \circ g^{-1}$ maps a region of $\mathbf{R}^n$ into $\mathbf{R}^n$, it makes sense to say that $h \circ g^{-1}$ is differentiable.) Such a collection **A** is called an *atlas* for **M**.
>
> (d) A given atlas **A** for **M** can be extended in one and only one way to a maximal atlas $\mathbf{A}_{max}$ as follows: add to **A** every one-one mapping g of a subset of **M** onto an open set of $\mathbf{R}^n$ which, combined with any chart h of **A**, satisfies the condition of differentiability stated under (c).
>
> (e) **M** is given the weakest Hausdorff topology that makes every chart g in $\mathbf{A}_{max}$ into a homeomorphism. (A topological space is said to be Hausdorff if any two points in it have open neighborhoods that do not overlap.)

The pair (**M**,**A**) is an n-manifold.

Roberto Torretti
Universidad de Chile
cordua@rdc.cl

First published: July 26, 1999
Content last modified: July 26, 1999

Roberto Torretti
Curriculum Vitæ

Postal Address: Casilla 20017, Correo 20, Santiago, CHILE
Phone: 562-212-5692
Birthdate: January 16, 1930
Electronic Mail: cordua@rdc.cl

Education

Ph.D., Philosophy, University of Freiburg, Germany, 1954
(Thesis Director: Wilhelm Szilasi)

Areas of Specialization

Philosophy of physics, especially relativity theory

Philosophy of mathematics, especially 19th century geometry

Professional Experience

- Professor of Philosophy,
  Universidad de Chile (1999 -- present)
- Professor of Philosophy,
  University of Puerto Rico (1970 -- 1995)
- Editor, *Dialogos* (1971 -- 1995)
- Professor of Philosophy,
  Universidad de Chile (1964 -- 1970)
- Director, Centro de Estudios Humanisticos,
  Faculty of Physics and Mathematics,
  Universidad de Chile (1964 -- 1970)
- Professor of Philosophy,
  Universidad de Concepcion, Chile (1961-- 1964)
- Chairman, Department of Philosophy,
  Universidad de Concepcion, Chile (1961 -- 1964)
- Lecturer, Faculty of General Studies,
  University of Puerto Rico (1958 -- 1961)
- Translator, Secretariat,
  United Nations, New York City(1955 -- 1958)

Academic Awards:

- Fellow, Pittsburgh Center for the Philosophy of Science, 1983-1984
- John Simon Guggenheim Memorial Fellow, 1980-1981
- John Simon Guggenheim Memorial Fellow, 1975-1976
- Alexander-von-Humboldt Dozentenstipendiat, Kant-Archiv, Bonn, 1964-1965

Learned Societies

- Institut International de Philosophie
- Academie Internationale de Philosophie des Sciences
- British Society for the Philosophy of Science
- Philosophy of Science Association

Publications

Books:

- *Manuel Kant. Estudio sobre los fundamentos de la filosofia critica*. Santiago: Ediciones de la Universidad de Chile, 1967. 603 pp. Seco fia de la Naturaleza. Textos Antiguos y Modernos. Santiago: Editorial Universitaria, 1971. 178 pp. Second, enlarged edition, Santiago: Editorial Universitaria, 1998. 170 pp.
- (With Luis O. Gomez) *Problemas de la Filosofia. Textos filosoficos clasicos y contemporaneos.* Rio Piedras: Editorial Universitaria, Universidad de Puerto Rico, 1975. 768 pp. Reprinted several times with minor corrections.
- *Philosophy of Geometry from Riemann to Poincare*. Dordrecht: D. Reidel Publishing Co., 1978. xiii + 458 pp. Corrected reprint. Dordrecht: D. Reidel Publishing Co., 1984. xiii+458 pp.
- *Relativity and Geometry*. Oxford: Pergamon Press, 1983. xi + 395 pp. Corrected reprint. New York: Dover, 1996. xiv + 395 pp.
- *Creative Understanding: Philosophical Reflections on Physics*. Chicago: The University of Chicago Press, 1990. xvi + 369 pp.
- (With Carla Cordua) *Variedad en la Razon: Ensayos sobre Kant*. Rio Piedras: Editorial de la Universidad de Puerto Rico, 1992. x + 248 pp.
- *La geometria del universo y otros ensayos de filosofia natural*. Merida: Consejo de Publicaciones de la Universidad de los Andes, 1994. x + 296 pp.
- *Sophocles' Philoctetes*. Text and Commentary. Bryn Mawr, PA: Thomas Library, Bryn Mawr College, 1997. 94 pp. (Bryn Mawr Greek Commentaries).
- *El Paraiso de Cantor: La tradicion conjuntista en la filosofia matematica*. Santiago de Chile: Editorial Universitaria, 1998. xiv + 589 pp.
- *The Philosophy of Physics*. New York: Cambridge University Press, 1999. xv + 512 pp.

EncyclopediaArticles:

- "Geometry". In Hans Burkhardt and Barry Smith, eds., *Handbook of Metaphysics and Ontology*. Munich: Philosophia Verlag, 1991.
- "El metodo axiomatico". In *Enciclopedia Iberoamericana de Filosofia*, Vol. 4. Madrid: Trotta, 1993, pp. 89-110.
- "Spazio". In *Dizionario delle Scienze Fisiche*. Roma: Istituto della Enciclopedia Italiana, 1995. Vol. 5, pp. 427-433.
- "Space". In *The Routledge Encyclopedia* London: Routledge. 1998.
- "Spacetime". In *The Routledge Encyclopedia of Philosophy*. London: Routledge. 1998.
- "Geometry in the 19th Century", In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*, URL = <http://plato.stanford.edu/entries/geometry-19th/>

Other Articles:

- "Aspectos de la doctrina de Kelsen". *Anales de la Universidad de Chile*, 100: 85-108 (1955).
- "Medio geografico y medio de la conducta en la psicologia moderna". *Atenea*, 361-362: 105-119 (1955).
- "Ha habido progreso de la filosofia en su historia?" *Revista de Filosofia* (Chile), iv.1: 49-56 (1957).
- "Causalidad y evolucion. Consideraciones sobre el presunto dilema que estas ideas plantearian a la ciencia". *Revista de Filosofia* (Chile), iv.2-3: 38-51 (1957).
- "Socialidad del individuo". *Revista de Ciencias Sociales*(Puerto Rico), v.1: 21-29 (1961).
- "Kant, filosofo del mas aca". *La Torre*, 34: 161-179 (1961).
- "Reflexiones sugeridas por el *Tractatus* de Wittgenstein". *Revista de Ciencias Sociales* (Puerto Rico), V.4: 479-490 (1961).
- *Hume y la religion*. Concepcion: Ediciones Atenea, n.d. (1962), 36 pp.
- "Poder politico y opresion". *Revista de Filosofia* (Chile), IX.1-2: 35-48 (1962).
- "Lecciones sobre el empirismo ingles". *Revista de Filosofia* (Chile), IX.3: 113-155 (1962).
- "Para introducir a Heidegger". *La Torre*, 39: 87-102 (1962).
- "Sobre el significado del imperativo categorico". *Revista de Filosofia* (Chile), X.1: 45-66 (1963).
- "Finitud del hombre y limites del conocimiento en Descartes y Leibniz". *Anales de la Universidad de Chile*, 128: 33-58 (1963).
- "Contrato social y economia dirigida en el pensamiento politico de Fichte". *Revista de Ciencias Sociales* (Puerto Rico), VIII.4: 357-375 (1964).
- "Unamuno, pensador cristiano". In: *Unamuno*, Santiago: Ediciones de la Universidad de Chile, 1964, pp. 95-112.
- "Introduccion a un estudio de la deduccion trascendental de las categorias en la primera *Critica* de Kant". *Revista de Filosofia* (Chile), xii: 19-61 (1965).
- "Las contrapartidas incongruentes en la gestacion de la f Kant". *Dialogos*, 3: 7-24 (1965).
- "Las *Investigaciones* de Wittgenstein y la posibilidad de la filosofia". *Dialogos*, 10: 35-59 (1968). Reprinted in Jorge J.E. Gracia, Eduardo Rabossi, Enrique Villanueva and Marcelo Dascal, eds. *El analisis filosofico en America Latina*, Mexico: Fondo de Cultura Economica, 1985, pp. 536-556.
- "Tercer Congreso Internacional de Kant". *Dialogos*, 17: 113-117 (1969).

- "Die Frage nach der Einheit der Welt bei Kant". *Kantstudien*, 62: 77-97 (1972).
- "On the subjectivity of objective space". In: L. W. Beck, ed., *Proceedings of the Third International Kant Congress*, Dordrecht: D. Reidel Publishing Co., 1972, pp. 535-540. Reprinted in L. W. Beck, ed., *Kant's Theory of Knowledge*, Dordrecht: D. Reidel Publishing Co., 1974, pp. 111-116.
- "Remarks on Salmon's Paradox of Primes". *Philosophy of Science*, 39: 260-262 (1972).
- "La filosofia de la aritmetica de Husserl". *Studi internazionali di Filosofia*,4: 183-206 (1972).
- "Logica formal y forma logica". In: *Ueberlieferung und Auftrag. Festschrift fuer Michael de Ferdinandy*, Wiesbaden: Pressler, 1972, pp. 624-633.
- "Juicios sinteticos a priori". *Cuadernos de Filosofia* (Buenos Aires), xi,20: 297-320 (1973).
- "On Mr. Kielkopf's not so sober understanding of standard elementary logic". *Mind*, 83: 575-577 (1974).
- "La geometria en el pensamiento de Kant". *Anales del Seminario de Metafisica* (Madrid), 9: 9-60 (1974).
- "El debate sobre el individualismo metodologico". *Dialogos*, 26: 95-117 (1974).
- "Problemas filosoficos del espacio y el tiempo. A proposito de la nueva edicion de la obra de Adolf Gruenbaum". *Dialogos*, 27: 89-117 (1974).
- "Espacio y tiempo: algunos libros recientes". *Dialogos*, 29-30: 255-294 (1977).
- "Bedingtes und Unbedingtes in der Mathematik". In: *Transzendenz und Immanenz, Philosophie und Theologie in der veraenderten Welt*, hrsg. von D. Papenfuss und J. Soering, Stuttgart: Kohl 1977, pp. 303-308.
- "Tres filosofos de la geometria". *Revista Latinoamericana de Filosofia*, 3: 3-21 (1977).
- "Hugo Dingler's philosophy of geometry". *Dialogos*,32: 85-128 (1978).
- "Presencia e idea del mundo". *Cuadernos de la Facultad de Humanidades*, 1: 3-21 (1978).
- "Jackson and Pargetter's criterion of distant simultaneity". *Philosophy of Science*, 46: 302-306 (1979).
- "Indole y funcion de los principios de la Teoria General de la Relatividad". *Revista Latinoamericana de Filosofia*, 5: 209-233 (1979).
- "Mathematical theories and philosophical insights in cosmology". In: H. Nelkowski *et al.*, eds., *Einstein Symposion Berlin*, Berlin: Springer, 1979, pp. 320-335 (*Springer lecture notes in physics*, 100).
- "Three kinds of mathematical fictionalism". In: J. Agassi and R. S. Cohen, eds., *Scientific Philosophy Today*, Dordrecht: D. Reidel Publishing Co., 1981, pp. 399-414.
- (With John Stachel) "Einstein's first derivation of mass-energy equivalence", *American Journal of Physics*, 50: 760-763 (1982).
- "'Lo que hay'", *Dialogos*, 41: 89-93 (1983).
- "Causality and spacetime structure in Relativity". In: R. S. Cohen and L. Laudan, eds., *Physics, Philosophy and Psychoanalysis*, Dordrecht: D. Reidel Publishing Co., 1983, pp. 273-293.
- "Kosmologie als ein Zweig der Physik". In: Bernulf Kanitscheider, ed., *Moderne Naturphilosophie*, Wuerzburg: Koenigshausen & Neumann, 1984, pp. 183-200.
- "La critica de conceptos en las revoluciones de la fisica basica". *Revista Latinoamericana de Filosofia*, 10: 25-41 (1984).
- "Spacetime physics and the philosophy of science". *British Journal for the Philosophy of Science*,

35: 280-292 (1984).

- "Observation".*British Journal for the Philosophy of Science*, 37: 1-23 (1986).
- "Conceptual reform in scientific revolutions". In: Ruth Barcan Marcus et al., eds., *Logic, Methodology and Philosophy of Science VIII*, Amsterdam: North-Holland, 1986, pp. 413 183-212 (1986); Part II, *Dialogos*, 49: 147-188 (1987).
- "La determinacion omnimoda de las cosas y el fenomenismo de Kant". *Revista Latinoamericana de Filosofia*, 13: 132-141 (1987).
- "Do conjunctive forks always point to a common cause?" *British Journal for the Philosophy of Science*, 38: 384-387 (1987).
- "Probabilidad y determinismo". *Congreso Internacional Extraordinario de Filosofia del 20 al 26 de Setiembre de 1987*. Universidad Nacional de Cordoba, Republica Argentina. Tomo III, pp. 1201-1207.
- "Extractos de una correspondencia" (with C. Ulises Moulines). *Dialogos*, 53: 123-137 (1989).
- "The geometric structure of the universe". In: Evandro Agazzi and Alberto Cordero, eds., *Philosophy and the Origin and Evolution of the Universe*, Dordrecht: Kluwer, 1991, pp. 53-73.
- "'Y se hizo la luz...': Newton y la Ilustracion". *La Torre* n.e., 5 [num. ext.]: 169-178 (1991).
- "Mathematical structures and physical necessity". In: Javier Echeverria, Andoni Ibarra y Thomas Mormann, eds., *The Space of Mathematics: Philosophical, Epistemological, and Historical Explorations*. Berlin: Walter de Gruyter, 1992, pp. 132-140.
- "La tradicion semantica". *Revista Latinoamericana de Filosofia*, 18, 2: 333-340 (1993).
- "Una idea feliz". *Revista Latinoamericana de Filosofia*, 19: 289-301 (1993).
- "Kitcher on the advancement of science". *Dialogos*, 64: 201-215 (1994).
- "El 'observador' en la fisica del siglo XX". In: Francisco Jose Ramos, ed., *Hacer: Pensar. Coleccion de escritos filosoficos*, Rio Piedras: Editorial de la Universidad de Puerto Rico, 1994, pp. 581-610.
- "Einstein's luckiest thought". In: Jarrett Leplin, ed. *The Creation of Ideas in Physics*. Dordrecht: Kluwer, 1995, pp. 89-96.
- "Realismo cientifico y ciencia real". *Theoria*, XI.26: 29-43 (1996).
- "Las analogias de la experiencia de Kant y la filosofia de la fisica". *Anales de la Universidad de Chile*, Sexta Serie, 4: 77-96 (1996).
- " continuidad en la historia de la fisica". *Revista de Filosofia* (Chile), 49/50: 29-44 (1997).
- "From Physics to Metaphysics". *Studies in the History and Philosophy of Modern Physics*. 28: 291-298 (1997).

Book Reviews:

[Note: All reviews are in Spanish unless otherwise indicated.]

- Bogumil Jasinowski, *Saber y dialectica*, Santiago 1957 (*La Torre*, 21: 197-201 (1958)).
- Georg Lukacz, *El asalto a la razon*, trad. por Wenceslao Roces, Mexico 1959 (*Revista de Ciencias Sociales*, iv.2: 390-395 (1960)).
- Hans Freyer, *Teoria de la epoca actual*, trad. por Luis Villoro, Mexico 1959 (*Revista de Ciencias Sociales*, v.1: 85-87 (1961)).

- Willard van Orman Quine, *From a logical point of view*, 2nd ed., Cambridge MA 1961 (RF, ix.1-2: 151-154 (1962)).
- Alfred J. Ayer, ed., *Logical positivism*, Glencoe IL 1959 (RF, ix.3: 157-164 (1962)).
- *Jornadas de Filosofia: Posibilidad de la metafisica*, Tucuman 1961 (IRB, xiii.1: 82-83 (1963)).
- Maine de Biran, *De l'apperception immediate*, Paris 1963 (*Anales de la Universidad de Chile*, 128: 202-207 (1963)).
- Richard B. Braithwaite, *La explicacion cientifica*, trad. por V. Sanchez de Zavala, Madrid 1965 (*Anales de la Universidad de Chile*, 137: 228-236 (1966)).
- Mario Bunge, *The myth of simplicity*, Englewood Cliffs 1963 (*Anales de la Universidad de Chile*, 138: 250-256 (1966)).
- Thomas S. Kuhn, *The structure of scientific revolutions*, Chicago 1962 (*Anales de la Universidad de Chile*, 139: 257-261 (1966)).
- Mario Bunge, *Scientific research*, 2 vols., Berlin 1967 (*Anales de la Universidad de Chile*, 141-144: 346-350 (1967)).
- *The collected papers of Gerhard Gentzen*, Amsterdam 1969 (*Dialogos*, 21: 104-108 (1970)).
- Jaakko Hintikka, ed., *The philosophy of mathematics*, New York 1969 (*Dialogos*, 21: 109-114 (1970)).
- Robert G. Colodny, ed., *The Na contemporary science and philosophy*, Pittsburgh 1970 (*Dialogos*, 21: 115-121 (1970)).
- *Historisches W•örterbuch der Philosophie*, Band I, hrsg. von Joachim Ritter, Basel 1971 (*Dialogos*, 22: 171-175 (1972)).
- Giorgio Tonelli, *A Short-title list of subject dictionaries of the 16th, 17th and 18th centuries as aids to the history of ideas*, London 1971 (*Dialogos*, 22: 175-176 (1972)).
- Henry E. Kyburg Jr., *Probability and inductive logic*, New York 1970 (*Dialogos*, 22: 180-184 (1972)).
- Nicholas Rescher, *Scientific Explanation*, New York 1970 (*Dialogos*, 22: 184-188 (1972)).
- Wesley Salmon, *Statistical Explanation and Statistical Relevance*, Pittsburgh 1971 (*Dialogos*, 22: 188-192 (1972)).
- John R. Searle, ed., *The Philosophy of Language*, New York 1971; Leonard Linsky, ed., *Reference and Modality*, New York 1971 (*Dialogos*, 22: 213-214 (1972)).
- Patrick Suppes, *A Probabilistic Theory of Causality*, Amsterdam 1970 (*Dialogos*, 22: 215-216 (1972)).
- S. W. P. Steen, *Mathematical Logic with Special Reference to the Natural Numbers*, Cambridge 1972 (*Dialogos*, 23: 219-221 (1972)).
- Karel Lambert and Bas C. van Fraassen, *Derivation and Counterexample, an introduction to philosophical logic*, Encino 1972 (*Dialogos*, 23: 221-224 (1972)).
- Robert Rogers, *Mathematical Logic and Formalized Theories, a survey of basic concepts and results*, Amsterdam 1971 (*Dialogos*, 23: 224-227 (1972)).
- Peter Achinstein, *Law and Explanation, an essay in the philosophy of science*, Oxford 1971 (*Dialogos*, 23: 227-229 (1972)).
- *Proceedings of the Third International Kant Congress*, ed. by L.W. Beck, Dordrecht 1972 (*Dialogos*, 23: 229-236 (1972)).
- Hilary Putnam, *Philosophy of Logic*, New York 1971; Norwood Russell Hanson, *Observation and*

Interpretation, a guide to the philosophy of science, New York 1971; Norman Malcolm, *Problems to Wittgenstein*, New York 1971 (*Dialogos*, 23: 236-239 (1972)).

- Bas C. van Fraassen, *Formal Semantics and Logic*, New York 1971 (*Dialogos*, 23: 240 (1972)).
- M. Bunge, F. Halbwachs, T.S. Kuhn, J. Piaget et L. Rosenfed, *Les Theories de la Causalite*, Paris 1971 (*Dialogos*, 23: 241-242 (1972)).
- Carlo Borromeo Giannoni, *Conventionalism in Logic, a study in the linguistic foundation of logical reasoning*, The Hague 1971 (*Dialogos*, 24: 167-174 (1973)).
- Georg Henrik von Wright, *Explanation and Understanding*, Ithaca 1971 (*Dialogos*, 24: 174-179 (1973)).
- Robert G. Colodny, ed., *Paradigms and Paradoxes: The Philosophical Challenge of the Quantum Domain*, Pittsburgh 1972 (*Dialogos*, 24: 196-199, (1973)).
- Robert Borger and Frank Cioffi, eds., *Explanation in the Behavioural Sciences*, Cambridge 1970 (*Dialogos*, 24: 200-203 (1973)).
- Herbert Meschkowski, ed., *Grundlagen der modernen Mathematik*, Darmstadt 1972; Karl Strubecker, ed., *Geometrie*, Darmstadt 1972 (*Dialogos*, 24: 205-207 (1973)).
- Mario Bunge, *Method, Model and Matter*, Dordrecht 1973; *Teoria y realidad*, Barcelona 1972 (*Dialogos*, 25: 156-167 (1973)).
- Karl R. Popper, *Objective Knowledge, an evolutionary approach*, Oxford 1972 (*Dialogos*, 25: 168-174 (1973)).
- J.M. Jauch, *Are Quanta Real? A Galilean dialogue*, Bloomington 1973 (*Dialogos*, 25: 179-181 (1973)).
- Mario Bunge, *Philosophy of Physics*, Dordrecht 1973 (*Dialogos*, 26: 176-181 (1974)).
- Ulises Moulines, *La estructura del mundo sensible (sistemas fenomenalistas)*, Barcelona 1973 (*Dialogos*, 26: 181-183 (1974)).
- Herbert Feigl, Wilfrid Sellars and Keith Lehrer, eds., *New Readings in Philosophical Analysis*, New York 1972 (*Dialogos*, 26: 183-186 (1974)).
- Juergen Kluever, *Operationalismus: Kritik und Geschichte einer Philosophie der exakten Wissenschaften*, Stuttgart-Bad Canns (1974)).
- William A. Wallace, *Causality and Scientific Explanation*, 2 vols., Ann Arbor 1972-1974 (*Dialogos*, 27: 179-181 (1973)).
- Fernando Montero Moliner, *El empirismo kantiano*, Valencia 1973 (*Dialogos*, 27: 177-178 (1974)).
- Jose Ferrater Mora, *Cambio de marcha en filosofia*, Madrid 1974 (*Dialogos*, 28: 153-157 (1975)).
- Imre Lakatos, *Proofs and Refutations, the logic of mathematical discovery*, Cambridge 1976 (*Dialogos*, 31: 182-186 (1978)).
- Jane Bridge, *Beginning Model Theory, the Completeness Theorem and some consequences*, Oxford 1977 (*Dialogos*, 31: 188-192 (1978)).
- Farhang Zabeeh, E.D. Klemke and Arthur Jacobson, eds., *Readings in Semantics*, Urbana 1974 (*Dialogos*, 31: 192-193 (1978)).
- Bernulf Kanitscheider, *Vom absoluten Raum zur dynamischen Geometrie*, Mannheim 1976 (*Dialogos*, 31: 193-195 (1978)).
- Roberto Escobar, *La filosofia en Chile*, Santiago 1976 (*International Review of Bibliography*, 28: 93-94 (1978)).

- Karl R. Popper and John C. Eccles, *The Self and its Brain*, Berlin 1977 (*Dialogos*, 32: 202-206 (1978)).
- John Winnie, ed., *Symposium on Space and Time*[special issue of *Nous*, vol. XI, n•º3, September 1977], (*Dialogos*, 32: 207-211 (1978)).
- Stephen P. Schwartz, ed., *Naming, Necessity and Natural Kinds*, Ithaca 1977 (*Dialogos*, 32: 211-212 (1978)).
- Nelson Goodman, *Ways of Worldmaking*, Indianapolis 1978 (*Dialogos*, 33: 173-174 (1979)).
- Arthur W. Burks, *Chance, Cause and Reason, an Inquiry into the Nature of Scientific Evidence*, Chicago 1977 (*Dialogos*, 33: 174-175 (1979)).
- George S. Pappas and Marshall Swain, eds., *Essays on Knowledge and Justification*, Ithaca 1978 (*Dialogos*, 33: 175-176 (1979)).
- Mario Bunge, *Treatise on Basic Philosophy, Volume 3, Ontology I: The Furniture of the Universe*, Dordrecht 1977 (*Dialogos*, 33: 151-156 (1979)).
- John S. Ear and John J. Stachel, eds., *Foundations of Space-Time Theories*, Minneapolis 1977 [Minnesota Studies in the Philosophy of Science, VIII] (*Dialogos*, 33: 165-171 (1979)).
- Robert Geroch, *General Relativity from A to B*, Chicago 1978 (*Dialogos*, 33: 171-173 (1979)).
- Wladyslaw Krajewski, *Correspondence Principle and Growth of Science*, Dordrecht 1977 (*Dialogos*, 34: 182-186 (1979)).
- George Boolos, *The Unprovability of Consistency, an Essay in Modal Logic*, Cambridge 1979 (*Dialogos*, 35: 183-184 (1980)).
- Angel Jorge Casares, *Sobre la esencia del hombre*, Rio Piedras 1979 (*Dialogos*, 35: 208-209 (1980)).
- Jose Ferrater Mora, *Diccionario de filosofia*, Sexta Edicion, Madrid 1979, 4 vols. (*Dialogos*, 38: 151-152 (1981)).
- Ricardo J. Gomez, *Las teorias cientificas - desarrollo - estructura - fundamentacion*, Tomo I, Buenos Aires 1977 (*Nous*, 15: 244-246 (1981)). In English.
- Nicholas Rescher, *Leibniz's Metaphysics of Nature, a group of essays*, Dordrecht 1981 (*Dialogos*, 39: 160-163 (1982)).
- Jill Vance Buroker, *Space and Incongruence: The Origin of Kant's Idealism*, Dordrecht 1981 (*Dialogos*, 39: 163-167 (1982)).
- Hartry Field, *Science without Numbers, a Defence of Nominalism*, Princeton 1980 (*Dialogos*, 40: 159-162 (1982)).
- Jerzy Giedymin, *Science and Convention, essays on Henri Poincare's philosophy of science and the conventionalist tradition*, Oxford 1981 (*Dialogos*, 40: 162-167 (1982)).
- Richard Swinburne, ed., *Space, Time and Causality*, Dordrecht 1983; D. Mayr and G. Suessmann, eds., *Space, Time and Mechanics: Basic Structures of a Physical Theory*, Dordrecht 1983 (*Dialogos*, 43: 131-136 (1984)).
- John R. Searle, *Intentionality, an essay in the philosophy of mind*, Cambridge 1983 (*Dialogos*, 44: 194-200 (1984)).
- Dudley Shapere, *Reason and the Search for Knowledge, Investigations in the Philosophy of Science* Dordrecht 1984 (*Dialogos*, 45: 166-171 (1985)).
- Andrew Pickering, *Constructing Quarks, a Sociological History of Particle Physics*, Chicago 1984 (*Dialogos*, 47: 177-183 (1986)).

- Isaac Levi, *Decisions and Revisions: Philosophical Essays on Knowledge and Value*, Cambridge 1984 (*Dialogos*, 49: 189-196 (1987)).
- Wolfgang Balzer, David A. Pearce and Heinz-Juergen Schmidt, eds., *Reduction in Science: Structure, Examples, Philosophical Problems*, Dordrecht 1984 (*Dialogos*, 49: 207-213 (1987)).
- Arthur Fine, *The Shaky Game. Einstein Realism and the Quantum Theory*, Chicago 1986 (*Dialogos*, 50: 155-160 (1987)).
- Joseph Agassi, *Science and Society. Studies in the Sociology of Science*, Dordrecht 1981 (*Philosophia*, 17: 223-223 (1987)). Reviewed in English.
- Asim O. Barut, Alwyn van der Merwe and Jean-Pierre Vigier, eds., *Quantum, Space and Time-The Quest Continues. Studies and Essays in Honour of Louis de Broglie, Paul Dirac and Eugene Wigner*, Cambridge 1984. (*Nous*, 21: 442-444 (1987)). Reviewed in English.
- Nicanor Ursua Lezaun, *Filosofia de la ciencia y metodologia critica*, Bilbao 1981. (*Nous*, 22: 327-329 (1988)). Reviewed in English.
- *Los filosofos presocraticos*. I. Introduccion general por Conrado Eggers Lan. Introducciones, traducciones y notas por Conrado Eggers Lan y Victoria E. Julia. II. Introducciones, traducciones y notas por Nestor Luis Cordero, Francisco Jose Olivieri, Ernesto La Croce y Conrado Eggers Lan. III. Introducciones, traducciones y notas por Armando Poratti, Conrado Eggers Lan, Maria Isabel Santa Cruz de Prunes y Nestor Luis Cordero. Madrid 1978-1980. (*Dialogos*, 51: 221-225 (1988)).
- Giovanni Reale, *A History of Ancient Philosophy. I. From the Origins to Socrates. III. The Systems of the Hellenistic Age.* Edited and translated [by] John R. Catan. Albany 1985-1987. (*Dialogos*, 51: 226-229 (1988)).
- G. S. Kirk, J. E. Raven, M. Schofield, *The Pres Selection of Texts.* Second edition. Cambridge 1983. (*Dialogos*, 51: 230-232 (1988)).
- Peter Galison, *How Experiments End*. Chicago 1987. (*Dialogos*, 52: 155-163 (1988)).
- J. V. Field, *Kepler's Geometrical Cosmology*. Chicago 1988. (*Dialogos*, 53: 155-163 (1988))
- B. A. Rosenfeld, *A History of Non-Euclidean Geometry: Evolution of the Concept of a Geometric Space.* Translated [from the Russian] by Abe Shenitzer. New York 1988. (*Dialogos*, 54: 250-254 (1989)).
- Igal Kvart, *A Theory of Counterfactuals.* Indianapolis 1986. (*Dialogos*, 54: 254-260 (1989)).
- Pietro Redondi, *Galileo Heretic (Galileo Eretico)*. English translation by Raymond Rosenthal. Princeton 1987. (*Revista Latinoamericana de Filosofia*, 15: 359-361 (1989)).
- Ronald Giere, *Explaining Science: A Cognitive Approach.* Chicago 1988. (*Dialogos*, 55: 199-203 (1990)).
- Raul Fornet Betancourt, *Introduccion a Sartre.* Mexico 1989. (*Dialogos*, 55: 211 (1990)).
- Aristoteles, *Las Categorias.* Edicion, introduccion, traduccion y notas de Humberto Giannini y Maria Isabel Flisfisch. Santiago de Chile 1988. (*Dialogos*, 55: 212-213 (1990)).
- Gottfried Wilhelm Leibniz, *Investigaciones Generales sobre el analisis de las nociones y las verdades.* Introduccion, traduccion y notas de Mauricio Beuchot y Alejandro Herrera-Ibanez. Mexico 1986. (*Dialogos*, 55: 213 (1990)).
- Immanuel Kant, *Pensamientos sobre la verdadera estimacion de las fuerzas vivas.* Traduccion y comentario de Juan Arana Canedo-Argueelles. Berna 1988. (*Dialogos*, 55: 213-214 (1990)).
- F. W. J. Schelling, *Sistema del idealismo trascendental.* Traduccion, prologo y notas de Jacinto Rivera Rosales y Virginia Lopez Dominguez. Barcelona 1988. (*Dialogos*, 55: 214 (1990)).

- Kurt Goedel, *Obras completas.* Introduccion y traduccion de Jesus Mosterin. Segunda edicion. Madrid 1989. (*Dialogos*, 55: 214-216 (1990)).
- Ian Hacking 1990. (*Dialogos*, 58: 185-187 (1991)).
- Lawrence Sklar, *Philosophy and Spacetime Physics.* Berkeley, 1985. (*Nous*, 25: 574-578 (1991)). Reviewed in English.
- L. T. F. Gamut, *Logic, Language, and Meaning*. Chicago 1991. (*Dialogos*, 59: 191-193 (1992)).
- Hans Joachim Kr•ämer, *Plato and the Foundations of Metaphysics: A Work on the Theory of the Principles and Unwritten Doctrines of Plato with a Collection of the Fundamental Documents.* Albany 1990. (*Dialogos*, 59: 194-200 (1992)).
- John Bigelow y Robert Pargetter, *Science and Necessity*. Cambridge 1990. (*Dialogos*, 60: 219-227 (1992)). Reviewed in English.
- John Earman, *World Enough and Space-Time: Absolute versus Relational Theories of Space and Time.* Cambridge, MA, 1989, (*Philosophical Review*, 101: 723-725 (1992)). Reviewed in English.
- Jon Barwise y John Etchemendy. *The Language of First Order Logic.* Chicago 1991. (*Dialogos*, 61: 195-200 (1993)).
- Ana Maria Vicuna N., *Filosofia, poesia y mito a la luz de Eros en el Symposio de Platon.* [Santiago de Chile]: Pontificia Universidad Catolica de Chile, 1993. (*Dialogos*, 65: 221-224 (1995)).
- Lawrence Sklar. *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics.* Cambridge 1993. (*Dialogos*, 65: 243-249 (1995)).
- Kurt Goedel, *Ensayos ineditos.* Edicion a cargo de Francisco Rodriguez Consuegra. Prologo de W. V. Quine. Barcelona: Mondadori, 1994. (*Dialogos*, 66; 185-190 (1995)).

Translations

- *Cinco escritos de Leibniz.* Introduccion, version castellana y notas de Roberto Torretti. *Revista de Filosofia* (Chile), IX.3: 51-81 (1962).
  (Spanish translation of five Latin texts published by Couturat in *Opuscules et fragments inedits de Leibniz*, Paris 1903, pp. 533-535, 11-16, 518-523, 401-403, 16-24.)
- Immanuel Kant, *La falsa sutileza de las cuatro figuras del silogismo. Dialogos*, 19: 7-22 (1970).
  (Spanish translation of *Die falsche Spitzfindigkeit der vier syllogistischen Figuren erwiesen von M. Immanuel Kant*, Koenigsberg 1762.)
- Immanuel Kant, *Sobre el fundamento primero de la diferencia entre las regiones del espacio.* Traduccion de Roberto Torretti. *Dialogos*, 22: 139-146 (1972).
  (Spanish translation of "Von dem ersten Grunde des Unterschiedes der Gegenden im Raume", published in *Wochentliche Königsbergsche Frag- und Anzeigungsnachrichten,* on February 6th, 13th and 20th, 1768.)
- Gottfried Wilhelm Leibniz, *Ensayos filosoficos alemanes.* Version castellana de Roberto Torretti. *Dialogos*, 23: 139-159 (1972).
  (Spanish translation of four German texts published by Guhrauer in *Deutsche Schriften*, Berlin 1838-40, vol. II, pp. 48-55, vol. I, pp. 410-413, 420-426, 414-419.)
- Gottfried Wilhelm Leibniz, *Principios metafisicos de las matematicas.* Traduccion de Roberto Torretti. *Dialogos*, 24: 131-149 (1973).

(Spanish translation of *Initia rerum mathematicarum metaphysica*, published by Gerhardt in G. W. Leibniz, *Mathematische Schriften*, Berlin and Halle 1849-63, t. VII, pp. 17-29.)

- Immanuel Kant, *Sobre la nitidez de los principios de la teologia natural y de la moral.* Traduccion y notas de Roberto Torretti. *Dialogos*, 27: 57-87 (1974).
  (Spanish translation of *Untersuchung ueber die Deutlichkeit der Grunds•ätze der natuerlichen Theologie und der Moral.* Berlin 1764.)
- Bernhard Riemann, *Sobre las hipotesis que estan en la base de la geometria.* Traduccion de Roberto Torretti. *Dialogos*, 31: 151-168 (1978).
  (Spanish translation of *Ueber die Hypothesen, welche der Geometrie zugrunde liegen,* posthumously published in 1867 by Dedekind in *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen vol.13.)*
- *Immanuel Kant, Monadologia fisica. Traduccion del latin por Roberto Torretti. Dialogos, 32: 173-190 (1978).*
  (Spanish translation of Metaphysicae cum geometria junctae usus in philosophia naturalis cuius Specimen I continet Monadologiam Physicam, Koenigsberg 1756.)
- George Berkeley, *Sobre el movimiento o sobre el principio y la naturaleza del movimiento y sobre al causa de la trasmision de los movimientos.* Traduccion del latin por Roberto Torretti. *Dialogos*, 34: 119-141 (1979).
  (Spanish translation of *De Motu*, London 1721.)
- Immanuel Kant, *Nuevo concepto del movimiento y el reposo.* Traduccion del aleman por Roberto Torretti. *Dialogos*, 34: 143-152 (1979).
  (Spanish translation of *Neuer Lehrbegriff der Bewegung und Ruhe und der damit verknuepften Folgerungen in den ersten Gruenden der Naturwissenschaft*, Koenigsberg 1758.)
- Ludwig Lange, *Sobre la ley de la inercia.* Traduccion del aleman por Roberto Torretti. *Dialogos*, 34: 153-170 (1979).
  (Spanish translation of "Ueber das Beharrungsgesetz", *Leipziger Berichte*, 37: 333-351 (1885).)
- Gottfried Wilhelm Leibniz, *Seis escritos de logica.* Traduccion y notas de Roberto Torretti. *Dialogos*, 51 (1988), 163-215.
  (Spanish translation of six Latin texts published by Couturat in *Opuscules et fragments inedits de Leibniz*, Paris 1903, pp. 49-57, 77-84, 232-237, 421-423, 193-202, 410-416.)
- Hilary Putnam, "Los modelos y la realidad". Traducido del ingles por Francisco Rodriguez Consuegra y Roberto Torretti. *Dialogos,* 63: 7-45 (1994).
  (Annotated translation of "Models and Reality", originally published in *Journal of Symbolic Logic*, 45: 464-482 (1980). The translation includes "Afterthoughts on ‘Models and Reality’", written by Putnam for this issue of *Dialogos*.)

Substructural Logics

Substructural logics are non-classical logics *weaker* than classical logic, notable for the absence of *structural rules* present in classical logic. These logics are motivated by considerations from philosophy (relevant logics), linguistics (the Lambek calculus) and computing (linear logic). In addition, techniques from substructural logics are useful in the study of traditional logics such as classical and intuitionistic logic. This article provides a brief overview of the field of substructural logic. For a more detailed introduction, complete with theorems, proofs and examples, the reader can consult the books and articles in the Bibliography.

Residuation

Logic is about *logical consequence*. As a result, the *conditional* is a central notion in logic because of its intimate connection with logical consequence. This connection is neatly expressed in *residuation condition*:

$$p, q \vdash r \quad \text{if and only if} \quad p \vdash q \rightarrow r$$

It says that r follows from p together with q just when $q \rightarrow r$ follows from p alone. The validity of the transition from q to r (given p) is recorded by the conditional $q \rightarrow r$. (It is called residuation by analogy with residuation in mathematics. Consider the connection between addition and substraction. $a + b = c$ if and only if $a = c - b$. The resulting a (which is $c - b$) is the *residual*, what is left of c when b is taken away.)

However, there is one extra factor in the equation. Not only is there the turnstile, for logical consequence, and the conditional, encoding consequence inside the language of propositions, there is also the *comma*, indicating the *combination* of premises. The behaviour of premise combination is also important in determining the behaviour of the conditional. As the comma's behaviour varies, so does the conditional. In this introduction we will see how this comes about.

Weakening

It is one thing for p to be true. It is another for the conditional $q \rightarrow p$ to be true. Yet, if '$\rightarrow$' is a material conditional, $q \rightarrow p$ follows from p. It seems worthwhile to understand how a conditional might work in the *absence* of this inference. This is tied to the behaviour of premise combination, as can be shown by this demonstration.

$$\frac{\dfrac{p \vdash p}{p, q \vdash p}}{p \vdash q \rightarrow p}$$

From the axiomatic $p \vdash p$ (anything follows from *itself*) we deduce that p follows from p together with q, and then by the residuation condition, $p \vdash q \rightarrow p$. Given that we accept the residuation condition, and the identity axiom at the start of the proof, we must reject the first step in the proof if we are to deny that $q \rightarrow p$ follows from p. This rule, which has the general form:

$$\frac{X \vdash A}{X, Y \vdash A}$$

is called the rule of *weakening*. We step from a stronger statement, that A follows from X to a possibly weaker one, that A follows from X together with Y.

This rule may fail, given different notions of premise combination (the notion encoded by the comma in X,Y). If the conditional $\rightarrow$ is *relevant* (if to say that $p \rightarrow q$ is true is to say, at least, that q truly *depends* on p) then the comma will not satisfy weakening. We may indeed have A following from X, without A following from X,Y, for it need not be the case that A depends on X and Y together.

In relevant logics the rule of weakening fails on the *other* side too, in that we wish *this* argument to be invalid too:

$$q \vdash q$$

$$p, q \vdash q$$

$$p \vdash q \rightarrow q$$

Again, q may follow from q, but this doesn't mean that it follows from *p together with q*, provided that "together with" is meant in an appropriately strong sense. So, in relevant logics, the inference from an arbitrary premise to a logical truth such as $q \rightarrow q$ may well fail.

Commutativity

If the mode of premise combination is commutative (if anything which follows from X, Y also follows from Y, X) then we can reason as follows, using just the identity axiom and residuation:

$$p \rightarrow q \vdash p \rightarrow q$$

$$p \rightarrow q, p \vdash q$$

$$p, p \rightarrow q \vdash q$$

$$p \vdash (p \rightarrow q) \rightarrow q$$

In the absence of commutativity of premise combination, this proof is not available. This is another simple example of the connection between the behaviour of premise combination and that of the conditional.

There are many kinds of conditional for which this inference fails. If $\rightarrow$ has *modal* force (read it as *entails*) then we may have p without also having $(p \rightarrow q) \rightarrow q$. It may be true that Greg is a logician (p) and it is true that Greg's being a logician entails Greg's being a philosopher ($p \rightarrow q$) but this does not *entail* that Greg is a philosopher. (There are many possibilities in which the entailment ($p \rightarrow q$) is true but q is not.) So we have p true but $(p \rightarrow q) \rightarrow q$ is not true.

This makes sense when we consider premise combination. Here when we say $X,A \vdash B$ is true, we are not just saying that B follows when we put X and A together. If we are after a genuine *entailment* $A \rightarrow B$, then we want B to be true in any (related) possibility in which A is true. So, $X,A \vdash B$ says that in *any* possibility in which A is true, so is B. These possibilities mightn't satisfy all of X. (In classical theories of entailment, the possibilities are those in which all that is taken as *necessary* in X are true.)

If premise combination is not commutative, then residuation can go in *two* ways. In addition to the

residuation condition giving the behaviour of $\rightarrow$, we may wish to define a new arrow $\leftarrow$ as follows:

$p, q \vdash r$ if and only if $q \vdash r \leftarrow p$

For the left-to-right arrow we have *modus ponens* in this direction:

$p \rightarrow q, p \vdash q$

For the right-to-left arrow, *modus ponens* is provable with the premises in the opposite order:

$p, q \leftarrow p \vdash q$

This is a characteristic of substructural logics. When we pay attention to what happens when we don't have the full complement of structural rules, then new possibilities open up. We uncover *two* conditionals underneath what was previously one (in intuitionistic or classical logic).

Associativity

Here is another way that structural rules influence proof. The associativity of premise combination provides the following proof:

$$\dfrac{\dfrac{\dfrac{\dfrac{p \rightarrow q, p \vdash q \qquad r \rightarrow p, r \vdash p}{p \rightarrow q, (r \rightarrow p, r) \vdash q}}{(p \rightarrow q, r \rightarrow p), r \vdash q}}{p \rightarrow q, r \rightarrow p \vdash r \rightarrow q}}{p \rightarrow q \vdash (r \rightarrow p) \rightarrow (r \rightarrow q)}$$

This proof uses the *cut* rule at the topmost step. The idea is that inferences can be combined. If $X \vdash A$ and $Y(A) \vdash B$ (where $Y(A)$ is a structure of premises possibly including A one or more times) then $Y(X) \vdash B$ too (where $Y(X)$ is that structure of premises with those instances of A replaced by X). In this proof, we replace the p in $p \rightarrow q, p \vdash q$ by $r \rightarrow p, r$ on the basis of the validity of $r \rightarrow p, r \vdash p$.

Contraction

A final important example is the rule of *contraction* which dictates how premises may be reused.

Contraction is crucial in the inference of $p \rightarrow q$ from $p \rightarrow (p \rightarrow q)$

$$
\dfrac{\dfrac{\dfrac{p \rightarrow (p \rightarrow q) \vdash p \rightarrow (p \rightarrow q)}{p \rightarrow (p \rightarrow q), p \vdash p \rightarrow q} \quad \dfrac{p \rightarrow q \vdash p \rightarrow q}{p \rightarrow q, p \vdash q}}{(p \rightarrow (p \rightarrow q), p), p \vdash q}}{\dfrac{p \rightarrow (p \rightarrow q), p \vdash q}{p \rightarrow (p \rightarrow q) \vdash p \rightarrow q}}
$$

These different examples give you a taste of what can be done by structural rules. Not only do structural rules influence the conditional, but they also have their effects on other connectives, such as conjunction and disjunction (as we shall see below) and negation (Dunn 1993; Restall 2000).

Logics in the Family

There are many different formal systems in the family of substructural logics. These logics can be motivated in different ways.

Relevant Logics

Many people have wanted to give an account of logical validity which pays some attention to conditions of *relevance*. If $X,A \vdash B$ holds, then X must somehow be *relevant* to A. Premise combination is restricted in the following way. We may have $X \vdash A$ without also having $X,Y \vdash A$. The new material Y might not be relevant to the deduction. In the 1950s, Moh (1950), Church (1951) and Ackermann (1956) all gave accounts of what a "relevant" logic could be. The ideas have been developed by a stream of workers centred around Anderson and Belnap, their students Dunn and Meyer, and many others. The canonical references for the area are Anderson, Belnap and Dunn's two-volume *Entailment* (1975 and 1992). Other introductions can be found in Read's *Relevant Logic* and Dunn's "Relevance Logic and Entailment" (1986). A more polemical introduction and defence of relevant logics can be found in Routley, Plumwood, Meyer and Brady's *Relevant Logics and Their Rivals*.

Resource Consciousness

This is not the only way to restrict premise combination. Girard (1987) introduced *linear logic* as a model for processes and resource use. The idea in this account of deduction is that resources must be used (so premise combination satisfies the relevance criterion) and they do not extend *indefinitely*.

Premises cannot be *re*-used. So, I might have $X,X \vdash A$, which says that I can use X twice to get A. I might not have $X \vdash A$, which says that I can use X once alone to get A. A helpful introduction to linear logic is given in Troelstra's *Lectures on Linear Logic* (1992). There are other formal logics in which the *contraction rule* (from $X,X \vdash A$ to $X \vdash A$) is absent. Most famous among these are Lukasiewicz's many-valued logics. There has been a sustained interest in logics without this rule because of Curry's paradox (Curry 1977, Geach 1995; see also Restall 1994 in Other Internet Resources).

Order

Independently of either of these traditions, Joachim Lambek considered mathematical models of language and syntax (Lambek 1958, 1961). The idea here is that premise combination corresponds to composition of strings or other linguistic units. Here X,X differs in content from X, but in addition, X,Y differs from Y,X. Not only does the *number* of premises used count but so does their *order*. Good introductions to the Lambek calculus (also called *categorial grammar*) can be found in books by Moortgat (1988) and Morrill (1994).

Proof Systems

We have already seen a fragment of one way to present substructural logics, in terms of proofs. We have used the residuation condition, which can be understood as including two rules for the conditional, one to *introduce* a conditional

$$\frac{X, A \vdash B}{X \vdash A \to B}$$

and another to *eliminate* it.

$$\frac{X \vdash A \to B \qquad Y \vdash A}{X, Y \vdash B}$$

Rules like these form the cornerstone of a natural deduction system, and these systems are available for the wide sweep of substructural logics. But proof theory can be done in other ways. *Gentzen* systems operate not introducing and eliminating connectives, but by introducing them both on the left and the right of the turnstile of logical consequence. We keep the introduction rule above, and replace the elimination rule by one introducing the conditional on the left:

$$\frac{X \vdash A \qquad Y(B) \vdash C}{Y(A \rightarrow B, X) \vdash C}$$

This rule is more complex, but it has the same effect as the arrow elimination rule: It says that if *X* suffices for *A*, and if you use *B* (in some context *Y*) to prove *C* then you could just as well have used *A* → *B* together with *X* (in that same context *Y*) to prove *C*, since *A* → *B* together with *X* gives you *B*.

Gentzen systems, with their introduction rules on the left and the right, have very special properties which are useful in studying logics. Since connectives are always *introduced* in a proof (read from top to bottom) proofs never *lose* structure. If a connective does not appear in the conclusion of a proof, it will not appear in the proof at all, since connectives cannot be eliminated.

In certain substructural logics, such as linear logic and the Lambek calculus, and in the fragment of the relevant logic **R** without disjunction, a Gentzen system can be used to show that the logic is *decidable*, in that an algorithm can be found to determine whether or not an argument $X \vdash A$ is valid. This is done by searching for proofs of $X \vdash A$ in a Gentzen system. Since premises of this conclusion must feature no language not in this conclusion, and they have no greater complexity (in these systems), there are only a finite number of possible premises. An algorithm can check if these satisfy the rules of the system, and proceed to look for premises for these, or to quit if we hit an axiom. In this way, decidability of some substructural logics is assured.

However, not all substructural logics are decidable in this sense. Most famously, the relevant logic **R** is not decidable. This is partly because its proof theory is more complex than that of other substructural logics. **R** differs from linear logic and the Lambek calculus in having a straightforward treatment of conjunction and disjunction. In particular, conjunction and disjunction satisfy the rule of *distribution*:

$$p \mathbin{\&} (q \vee r) \vdash (p \mathbin{\&} q) \vee (p \mathbin{\&} r)$$

The natural proof of distribution in any proof system uses both weakening and contraction, so it is not available in the relevant logic **R**, which does not contain weakening. As a result, proof theories for **R** either contain distribution as a primitive rule, or contain a second form of premise combination (so called *extensional* combination, as opposed to the *intensional* premise combination we have seen) which satisfies weakening and contraction.

Semantics

While the relevant logic **R** has a proof system more complex than the substructural logics such as linear logic, which lack distribution of (extensional) conjunction over disjunction, its *semantics* is altogether more simple. A Routley-Meyer *model* for the relevant logic **R** is comprised of a set of *points P* with a

three-place relation R on P. A conditional $A \rightarrow B$ is evaluated at a world as follows:

> $A \rightarrow B$ is true at x if and only if for each y and z where $Rxyz$, if A is true at y, B is true at z.

An argument is *valid* in a model just when in any point at which the premises are true, so is the conclusion. The argument $A \vdash B \rightarrow B$ is invalid because we may have a point x at which A is true, but at which $B \rightarrow B$ is not. We can have $B \rightarrow B$ fail to be true at x simply by having $Rxyz$ where B is true at y but not at z.

The three place relation R follows closely the behaviour of the mode of premise combination in the proof theory for a substructural logic. For different logics, different conditions can be placed on R. For example, if premise combination is commutative, we place a *symmetry* condition on R like this: $Rxyz$ if and only if $Ryxz$. Ternary relational semantics gives us great facility to *model* the behaviour of substructural logics. (The extent of the correspondence between the proof theory and algebra of substructural logics and the semantics is charted in Dunn's work on *Gaggle Theory* (1991) and is summarised in Restall's *Introduction to Substructural Logics* (2000).) Furthermore, if conjunction and disjunction satisfy the distribution axiom mentioned in the previous section, they can be modelled straightforwardly too: a conjunction is true at a point just when both conjuncts are true at that point, and a disjunction is true at a point just when at least one disjunct is true there. For logics, such as linear logic, *without* the distribution axiom, the semantics must be more complex, with a different clause for disjunction required to invalidate the inference of distribution.

It is one thing to use a semantics as a formal device to model a logic. It is another to use a semantics as an *interpretive* device to *apply* a logic. For logics like as the Lambek calculus, the interpretation of the semantics is straightforward. We can take the points to be linguistic units, and the ternary relation to be the relation of composition ($Rxyz$ if and only if x concatenated with y results in z). For the relevant logic **R** and its interpretation of natural language conditionals, more work must be done in identifying what features of reality the formal semantics models. Some of this work is reported in the article on relevant logic in this Encyclopedia.

Bibliography

A comprehensive bibliography on relevant logic was put together by Robert Wolff and can be found in Anderson, Belnap and Dunn 1992. The bibliography in Restall 2000 (see Other Internet Resources) is not as comprehensive as Wolff's, but it does include material up to the last part of the 1990s.

Books on Substructural Logic and Introductions to the Field

- Anderson, A.R., and Belnap, N.D., 1975, *Entailment: The Logic of Relevance and Necessity*, Princeton, Princeton University Press, Volume I.
- Anderson, A.R., Belnap, N.D. Jr., and Dunn, J.M., 1992, *Entailment*, Volume II, Princeton,

Princeton University Press
[This book and the previous one summarise the work in relevant logic in the Anderson--Belnap tradition. Some chapters in these books have other authors, such as Robert K. Meyer and Alasdair Urquhart.]

- Dunn, J.M., 1986, "Relevance Logic and Entailment" in F. Guenthner and D. Gabbay (eds.), *Handbook of Philosophical Logic*, Volume 3, Dordrecht: Reidel pp 117--224.
[A summary of work in relevant logic in the Anderson--Belnap tradition. An updated version of this essay, co-authored with Restall, will appear in the new edition of the *Handbook of Philosophical Logic*.]
- Moortgat, Michiel, 1988, *Categorial Investigations: Logical Aspects of the Lambek Calculus* Foris, Dordrecht.
[Another introduction to the Lambek calculus.]
- Morrill, Glyn, 1994, *Type Logical Grammar: Categorial Logic of Signs* Kluwer, Dordrecht
[An introduction to the Lambek calculus.]
- Read, S., 1988, *Relevant Logic*, Oxford: Blackwell.
[An introduction to relevant logic from a distinct philosophical perspective.]
- Restall, Greg, 2000, *An Introduction to Substructural Logics*, Routledge. (online précis)
[A comprehensive introduction to the field of substructural logics.]
- Routley, R., Meyer, R.K., Plumwood, V., and Brady, R., 1983, *Relevant Logics and its Rivals*, Volume I, Atascardero, CA: Ridgeview.
[Another distinctive account of relevant logic, this time from an Australian philosophical perspective.]
- Schroeder-Heister, Peter, and Dosen, Kosta, (eds), 1993, *Substructural Logics*, Oxford University Press.
[An edited collection of essays on different topics in substructural logics, from different traditions in the field.]
- Troestra, Anne, 1992, *Lectures on Linear Logic*, CSLI Publications
[A quick, easy-to-read introduction to Girard's linear logic.]

Other Works Cited

- Ackermann, Wilhelm, 1956, "Begründung Einer Strengen Implikation", *Journal of Symbolic Logic* **21** 113-128.
- Church, Alonzo, 1951, "The Weak Theory of Implication", in *Kontroliertes Denken: Untersuchungen zum Logikkalkül und zur Logik der Einzelwissenschaften*, Kommissions-Verlag Karl Alber, edited by A. Menne, A. Wilhelmy and H. Angsil, 22-37.
- Curry, Haskell B., 1977, *Foundations of Mathematical Logic*, Dover (originally published in 1963).
- Dunn, J.M., 1991, "Gaggle Theory: An Abstraction of Galois Connections and Residuation with Applications to Negation and Various Logical Operations", in *Logics in AI, Proceedings European Workshop JELIA 1990*, Lecture notes in Computer Science, volume **476** Springer-Verlag.

- Dunn, J.M., 1993, "Star and Perp," *Philosophical Perspectives* **7** 331-357.
- Geach, P. T., 1955, "On Insolubilia," *Analysis* **15** 71-72.
- Girard, Jean-Yves, 1987, "Linear Logic," *Theoretical Computer Science* **50** 1-101.
- Lambek, Joachim, 1958, "The Mathematics of Sentence Structure", *American Mathematical Monthly* **65** 154-170.
- Lambek, Joachim, 1961, "On the Calculus of Syntactic Types", in *Structure of Language and its Mathematical Aspects*, edited by R. Jakobson, (Proceedings of Symposia in Applied Mathematics, XII).
- Moh Shaw-Kwei, 1950, "The Deduction Theorems and Two New Logical Systems," *Methodos* **2** 56-75.
- Slaney, John, 1994, "Finite Models for some Substructural Logics," Automated Reasoning Project Technical Report TR-ARP-04-94. Available as either a DVI or Postscript file.

Other Internet Resources

- Restall, Greg, BibTeX Bibliography on Substructural Logic (sourcefile via ftp), from Restall 2000.
- Restall, Greg, 1994, *On Logics Without Contraction*, Ph. D. Thesis, The University of Queensland.
- Slaney, John, 1995, MaGIC: Matrix Generator for Implication Connectives, a software package for generating finite models for substructural logics.

Related Entries

logic: modal | logic: paraconsistent | logic: relevance

Greg Restall
Greg.Restall@mq.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 4, 2000
Content last modified: May 16, 2002

Doing vs. Allowing Harm

Is doing harm worse than allowing harm? If not, there should be no moral objection to active euthanasia in circumstances where passive euthanasia is permissible; and there should be no objection to bombing innocent civilians where doing so will minimize the overall number of deaths in war. There should, however, be an objection—indeed, an outcry—at our failure to prevent the deaths of millions of children in the third world from malnutrition, dehydration, and measles.[1] But is doing harm worse than allowing harm? We might divide approaches to this question into two broad kinds: those that attempt to answer it using examples without saying anything about the nature of the distinction. (Following Shelly Kagan, I'll call this approach 'the contrast strategy.') And those that analyze the distinction in depth and try to show that its underlying nature dictates an answer to the moral question.

---

1. The Contrast Strategy

James Rachels provides a classic example of the first approach.[2] He offers us a pair of cases—in one, Smith drowns his young cousin in the bathtub; in the other, Jones plans to drown his young cousin, but finds the boy already unconscious under water and refrains from saving him. The two cases are exactly alike except that the first is a killing and the second a letting die. Rachels invites us to agree that Smith's behavior is no worse than Jones's. He then concludes that killing per se is no worse than letting die per

se, and that if typical killings are worse than typical lettings die that must be because of other factors.

Although Rachels seems correct about Smith and Jones, the inference from these cases to the moral equivalence of killing and letting die in general (where other things are equal) has been challenged. Shelly Kagan argues that it assumes that "if a factor has genuine moral relevance, then for any pair of cases, where the given factor varies while others are held constant, the cases in that pair will differ in moral status."[3] He claims, moreover, that this assumption assumes the Additive Assumption, the view that "the status of the act is the net balance or sum which is the result of adding up the separate positive and negative effects of the individual factors."[4] He raises several objections to the Additive Assumption. Firstly, one might describe a pair of cases that are exactly alike except that one is a killing and the other a letting die, where the first intuitively seems far worse than the second. If this pair of cases is as good as Rachels' pair, then either the inference is valid in both cases—to prove the contradiction that killing is both worse and not worse than letting die—or it is invalid in both cases. Secondly, one might raise the rhetorical question: why addition—rather than, say, multiplication or some other function?

Instead of using the contrast strategy, let's try to figure out the nature of the distinction between killing and letting die and, more generally, between doing and allowing harm. In both doing and allowing, an agent is responsible for or relevant to a bad upshot—such as a death or injury—in the sense that she could have prevented it. The contrast is most naturally picked out by the terms 'doing' and 'allowing', or 'making' and 'allowing', but since these have vagaries and awkwardnesses in practice, I shall use the terms "positively relevant to an upshot" and "negatively relevant to an upshot" for cases of "doing" and "allowing", respectively.[5]

2. Distinguishing Distinctions

Suppose some upshot occurs and would not have occurred if the agent had behaved in some different way. The question of whether the agent is positively or negatively relevant to the upshot is often conflated with or distorted by questions that should be kept distinct from it, like the questions of (i) whether the agent intended the upshot, (ii) whether she could easily have prevented the upshot, (iii) whether she guaranteed the upshot or merely made it probable, and even (iv) whether the agent's behavior was morally objectionable. It can easily be seen that these do not coincide with the distinction between doing and allowing.

> (i) Consider the distinction between cases where an agent intends the upshot and cases where she does not. If you drive your car into someone's body and she dies as a result, you undoubtedly killed her, even if you did not intend her death. Conversely, someone may intentionally allow a child to drown in order to inherit his fortune.
>
> (ii) It tends to be easier to avoid killing than to avoid letting die, but this is only a tendency. Sometimes saving is easier than not killing. It is easy to throw a life preserver,

and it may be difficult to refrain from killing someone who is threatening one or who has treated one appallingly. There are even cases where it is physically difficult to avoid killing&3151;as for example, where one has to hold tight to a tree to prevent one's (light) vehicle whose brakes have failed from running into a pedestrian.

(iii) Sometimes the terms ‘making’ and ‘allowing’ are used to suggest the difference between making certain and making possible or probable. For example, in discussions of the problem of evil, people sometimes say, “Well, God didn't actually make the murder occur. He just allowed it to occur.” This is best understood, I believe, as a distinction between raising the probability of murder to 1 from something less than 1, on the one hand, and raising the probability of murder from 0 to something higher but still less than 1. This is a morally significant distinction but it is not the distinction between doing and allowing. An agent can kill without guaranteeing death. For example, by adding small quantities of poison to her victim's meals she may bring about the death, even though there was a 20% chance that the poison would not kill her. On the other hand, an agent might guarantee the demise of a plant by failing to water it in a situation where she is the only one who can do so.

(iv) Finally, the distinction between killing and letting die is sometimes thought to have, as part of its conceptual content, a moral element. This thought is rarely made explicit, but the way people are inclined to classify cases suggests that they are guided by it. There are two main difficulties with this way of drawing the line. Firstly, if it is true by definition that killing is worse than letting die, then the question of whether killing is worse than letting die is settled in a trivial, circular, uninteresting way. Secondly, there are obvious counterexamples to this crude account—morally appalling cases of letting die—failing to feed one's children—and morally acceptable cases of killing—We have no hesitation talking of killing in self-defense. Let's turn to some more plausible candidates. In what follows I discuss a series of accounts of the distinction, and where appropriate, the moral significance or insignificance of each account.

3. Causing and Not Causing Not to Occur

One natural suggestion is that the agent who is positively relevant to the upshot causes it to occur; whereas the agent who is negatively relevant to the upshot doesn't cause it, but simply fails to prevent it where she could have done so.[6] This suggestion has immediate moral implications. It seems true by definition (almost) that you can be causally responsible only for upshots that you cause. And it is arguably true that you can be morally responsible only for what you are causally responsible for. So, if you cause a bad state of affairs, you've probably done wrong; whereas if you don't cause a bad state of affairs, you haven't. In choosing between killing and letting die, you are choosing between doing wrong and not doing wrong. The question of what you ought to do is then tautologously easy.

This argument begins to get into trouble when we reflect on the fact that we are often responsible for

upshots we allow: the death of the houseplants or the child's illiteracy. When we notice that, in these cases, the plants die or the child remains uneducated because of some failure on the agent's part, it becomes clear that the agent does, in some sense, cause the upshots. Moreover, most widely accepted contemporary accounts of causation imply that some event or fact involving these agents causes the deaths or illiteracy. For example, the counterfactual account of causation—according to which (very roughly) event *E* causes *F* if and only if had *E* not occurred *F* would not have occurred either—implies that it was the agent's failure to water the plants that caused the deaths.[7] John Mackie's INUS condition—according to which *E* causes *F* if and only if *E* is a(n insufficient but) necessary part of a(n unnecessary but)sufficient condition for *F*—implies that the fact that the agent failed to water the plants causes the plants to die.[8]

4. Counterfactual Accounts

We are concerned then with a contrast between two ways the behavior of agents causes upshots. One suggestion is to say that when the agent is positively relevant to the upshot, the upshot would not have occurred if she had been absent from the scene.[9] Suppose, for example, the victim dies because I push his head under water. He wouldn't have died if I had been absent. On the other hand, suppose he is in deep water and cannot swim and I don't save him. He would have drowned anyway if I had been absent. In these two cases, the counterfactual account draws the line in the intuitively correct way.

This account is sometimes used to support the claim that doing harm is worse than allowing harm, on the grounds that, on this account, allowing harm is simply a matter of letting nature take its course, which, other things being equal, is good, or at least permissible. There are two or three quick objections to this argument. Firstly, it assumes that acting (such as killing or saving lives) is a matter of interfering with the course of nature—in other words, that human action is somehow outside of the course of nature. This is extremely controversial. Secondly, even if human action is outside the course of nature, if the agent is faced with a choice between killing one and allowing two to die at the hand of some other agent, this argument would favor neither option since neither involves letting nature take its course. But, as traditionally understood and used, the Doctrine of Doing and Allowing is supposed to favor letting die in this case just as much as in others. Thirdly, interfering in the course of nature is sometimes obviously the better course of action—to stop the bleeding, restart the heart, and so on.

A different way the counterfactual account can be used to support the claim that doing harm is worse than allowing harm is this: if something bad happens when you are not present (or, especially, if you had never existed) then you aren't responsible for it. If we turn our attention to another world where you are present, but which is otherwise exactly like the first, it seems that your presence makes no difference empirically and, hence, should make no difference morally. This superficially compelling argument seems to prove too much—politicians' careers have hung on the question of whether they were in the room at the time the conspiracy was being hatched. Moreover, suppose an SS officer, Franz, tortures someone to death. But this is standard practice in the Gestapo. If Franz had stayed home with a sore throat, or if Franz had never existed, his pal Hans would have done the torturing, in the same way, at the same time Franz did. If the counterfactual account is correct, then Franz is negatively relevant to the

victim's death by torture. That is, Franz merely allowed the death to occur. This case also creates problems for the idea that killing is worse than letting die, since the fact that Hans was waiting in the wings in no way diminishes Franz's wrongdoing in this case.[10] So, this way of drawing the distinction is problematic, and this argument for the moral significance of the distinction is flawed.

Alan Donagan suggests a similar account of the distinction.[11] To determine whether the agent is positively or negatively relevant to an upshot, we should consider what would have happened if the agent had not acted at the relevant moment, or what would have happened if the agent had 'abstained from intervening in the course of nature'. It isn't entirely clear what we are supposed to imagine when we imagine this but perhaps it's that the agent is asleep or in a trance or in some other way not exercising her agency. Now—with respect to some behavior that led to some upshot, we might ask: would that upshot have occurred if the agent had abstained from intervening in the course of nature? If it would have, the agent allowed the upshot. If it would not have, then she did it (her relevance to the upshot is positive).

The Hans/Franz example can be revised to work against Donagan's version, but here's an additional counterexample. Suppose a man is lying asleep on the ground. He is awoken by a crash and notices a large rock rolling down the hill towards him. He can easily move out of its way, but realizes that if he does so the rock will gain momentum and kill a group of small deaf children further down the hill. He tenses his muscles, fights his desire to run away and stands his ground. The rock hits and seriously injures him. But he stops it. And he saves the children. Donagan's account, however, seems to imply that he merely allows the rock to stop, since, had he remained asleep, the rock would have struck and been stopped by his body.[12]

5. Action, Inaction and Positive and Negative Rights

Warren Quinn offers an account of the distinction—guided he admits by the conviction that doing harm is worse than allowing harm[13] —according to which an agent is positively relevant to a harmful upshot when his most direct contribution to the harm is an action, whether his own or that of some object.[14] His relevance is negative when his most direct contribution is an inaction, a failure to prevent the harm. An agent's most direct contribution to a harmful upshot of his agency is the contribution that most directly explains the harm. And one contribution explains harm more directly than another if the explanatory value of the second is exhausted in the way it explains the first.

The key difference here is between cases where the agent produces the result by an action and cases where she produces it by an inaction—pushing the head under water or refraining from throwing a life preserver. There's an extra complication here, however. Sometimes, Quinn says, your relevance to a death can be positive, you can kill, in other words, even though you don't act. This happens, for example, when you are on a train headed towards some drowning victims you wish to save when you notice someone tied to the tracks ahead of you. You can stop the train but you choose not to in order to reach your destination. Quinn believes that you kill in this case, because the train acts as your agent, taking you

where you want to go, and crushing the person tied to the tracks in the process. On the other hand, if you had chosen not to stop the train for some other reason but you would have not minded had someone else stopped the train, then your failure to stop the train would not have constituted a killing.

What are the moral implications of this way of drawing the line? Following Philippa Foot, Quinn believes that the key here is the distinction between negative and positive rights.[15] Positive relevance to harm involves the violation of negative rights; negative relevance to harm involves the violation of positive rights. Since negative rights are more stringent than positive rights, it is worse to be positively relevant to harm than to be negatively relevant to harm (*ceteris paribus*). But why should we think negative rights more stringent. Here's Quinn:

> [i]n such a morality [neutral vis á vis killing and letting die] the person trapped on the road has a moral say about whether his body may be destroyed only if what he stands to lose is greater than what others stand to gain. But then surely he has no real say at all. For, in cases where his loss would be greater than the gain to others, the fact that he could not be killed would be sufficiently explained not by his authority in the matter but simply by the balance of overall costs. And if this is how it is in general—if we may rightly injure or kill him whenever others stand to gain more than he stands to lose—then surely his body (one might say his person) is not in any interesting moral sense his. It seems rather to belong to the human community, to be dealt with according to its best overall interest…. Whether we are speaking of ownership or more fundamental forms of possession, something is, morally speaking, his only if his say over what may be done to it (and thereby to him) can override the greater needs of others.[16]

To say that one has a negative right against being harmed is to say that it is (at least, prima facie) wrong to harm one unless one wishes to be harmed. It is crucial that we add the phrase "unless one wishes to be harmed", since without it, the precedence of negative rights wouldn't give the victim any special say about his own body, because it would be just as wrong to harm him even if he asked to be harmed, and it would be wrong for him to harm himself. So, the crucial thing is that the victim has some sort of a say about what happens to himself(i.e., others are morally bound to respect his wishes with respect to his body to a certain extent). Quinn's claim is that if there is no extent to which someone's wishes with respect to his body, etc. are to be respected, then we've completely done away with the idea of ownership of one's body, etc.

One person's wishes about what happens to her body do often clash with someone else's wishes about what happen to his. For example, Susie wishes to marry Paul, but Paul doesn't wish to marry Susie. Morality obviously cannot give all such wishes precedence. So, the suggestion goes, let's give some subset of them precedence. For example, let's give negative wishes precedence. Why, what's so special about negative wishes? "Well, nothing," the answer seems to go. "But unless we give some wishes precedence, there will be no domain in which the victim's personal preference takes precedence; and in that case, there will be no sense in which the agent is lord in that domain. In other words, without some such precedence there will be no ownership of one's body or mind, etc." Quinn argues that we cannot

give positive rights precedence over negative rights without incoherence.[17] And, hence, he concludes that we must give negative rights precedence over positive rights.

But there are other ways to divide up rights than the division into positive and negative rights. We might divide them into the rights of children and the rights of adults, rights concerning the upper half of the body and rights concerning the lower half, etc. and then give precedence to one set over the other whenever they come into conflict. These seem arbitrary and wasteful, but their rationale seems no worse than Quinn's.

Quinn's is a funny sort of defense of negative rights. Unless I'm missing something, it doesn't pick out any special feature of negative rights that makes them specially worth respecting.

6. The ‘Most of the Things He Could have Done’ Account

Like Quinn, Jonathan Bennett thinks that the fundamental distinction between doing and allowing is between cases where the upshot occurs because of one's action and cases where the upshot occurs because of one's inaction—although he prefers to replace “action/inaction” talk with “positive/negative fact” talk.[18] When Bennett discusses the contrast between positive and negative relevance to harm, he is attempting to capture a deep, philosophically interesting distinction that underlies our talk of ‘doing and allowing’ ‘making and letting’,‘killing and letting die’. He acknowledges that the correspondence between his distinction and the distinctions we make in everyday life and language may be inexact. He says that my behavior is negatively relevant to an upshot if a negative fact about my behavior is the least informative fact that suffices to complete a causal explanation of it; whereas my behavior is positively relevant to that upshot if a positive fact about my behavior is the least informative fact about my conduct that suffices to complete a causal explanation of it. For example, if I jog while you drown, your drowning could be explained by the fact that I jogged, but it could also be explained by the less informative fact that I did not pull you out.

In a nutshell, on Bennett's view, an agent's relevance to an upshot is positive if most of the ways she could have behaved at the time would not have led to the upshot; otherwise, it is negative.[19] For example, suppose I douse a slug with salt and it dies as a result. My relevance to its demise is positive, since most of the ways I could have behaved would not have led to the death. On the other hand, if it dies because I fail to move it from the path of a car, then most of the ways I could have behaved at the time would have led to its death, so my relevance to the death is negative.

On this account, doing harm is no worse than allowing harm. If some upshot obtains because of the way you behaved, then the fact that there were many ways (rather than only a few) you could have behaved which would also have had that result is morally insignificant. This conclusion is surprising, even shocking. Bennett's account, however, can explain why we tend to think of killing as worse than letting die. He claims, quite plausibly, that it is morally worse to be causally relevant to a bad upshot one could

easily have avoided than a similarly bad upshot one could have avoided only with great difficulty. If most of the ways one could have behaved would have led to an upshot, then it was probably somewhat difficult or onerous to avoid the upshot; whereas if most of the ways one could have behaved would not have led to an upshot, then it was probably fairly easy to avoid the upshot. This correlation is not inevitable, however. It is easy to call 911, and can be difficult to refrain from killing a child at the bottom of a well if your only alternative is to continue hanging from a rope above her.

In spite of its virtues, Bennett's account faces some formidable difficulties. It has often been attacked with counterexamples like the following:

> Raccoon: Returning to the campsite after fetching water, I notice a raccoon eating my food. Hiding behind some trees downwind of him, I know that if I make a noise he will run away. I notice a large bell near me, but decide against using it, allowing the raccoon to eat my food.

Bennett's account, of course, says just that. But let's change the story slightly. Suppose now that I am closely surrounded by bells. If I move at all, I will make a sound loud enough to scare off the raccoon. Again, I don't move and he eats on undisturbed. It seems that I still allow the food to be eaten. The addition of more bells doesn't change this.

Some philosophers argue that such examples refute Bennett's account.[20] Their diagnosis of the difficulty goes like this: "What's important here is that the agent is immobile throughout. Immobility is incompatible with positive relevance to an upshot. Nifty slogan: 'You cannot do anything by doing nothing.' Bennett's account wrongly implies that immobility is compatible with positive relevance to an upshot."

But immobility is not necessarily incompatible with positive relevance to an upshot. Consider the case I described earlier of the agent who was positively relevant to the rock's stopping and the children's being unscathed by standing his ground, precisely by not moving. Here's another case. After a mild earthquake the agent finds herself sitting on a child's chest. Unless she moves soon, the child's lungs will be crushed. She stays put. The child dies. It is not implausible to say that the agent killed the child, although she was completely immobile throughout the relevant time period.[21] These cases make it clear that immobility doesn't rule out positive relevance. But if it is not my immobility that explains the fact that I am negatively relevant to the raccoon's consuming my food, what is it?

Before answering this question, let's consider another difficult case for Bennett's account, Sassan: An assassin, A. Sassan, is preparing to assassinate Victor by shooting him. A second assassin, Baxter, is waiting across the street watching Sassan to ensure his success. If Sassan shows any signs of hesitation, Baxter will shoot Victor himself. Suppose Sassan knows about Baxter and his intentions and also knows that he can turn his gun on Baxter instead of on Victor if he so chooses. Although this thought crosses his mind, he quickly suppresses it, since he is committed to Victor's annihilation. He shoots Victor and Victor dies instantly. Most of the ways he could have behaved would have led to the shooting and death

of Victor (either by himself or by Baxter). By Bennett's account, Sassan's relevance to the fact that Victor is shot and killed is negative. This means that Sassan doesn't kill Victor, but merely lets him die. Hold on! I just said that Sassan shot Victor. He pulled the trigger. The gun fired. A bullet flew out of the barrel and entered Victor's body. Victor died from the bullet wound. A clearer case of killing is impossible to find. Bennett might repeat the point that positive relevance to a death is not exactly the same as killing. Nevertheless, insofar as we have any pre-theoretical grip on (and interest in) the concept of positive relevance to a death, Sassan's relevance to Victor's death must strike us as positive rather than negative.[22] And yet Bennett's account implies that it is negative.

Someone sympathetic with Bennett's account may attempt to demonstrate that that account does accommodate our intuitions on this score by claiming that it implies that Sassan's behavior is positively relevant to the actual death that Victor died, since if Baxter had killed him, he would have died a different death. This response is not consistent with Bennett's approach, however, since it the phrases ‘the actual death’ and ‘a different death’ seem to be referring to events rather than facts. And Bennett is clearly concerned with relevance to facts not events.

Bennett himself has pointed out that the upshot that concerns us is not the fact that Victor died (no-one could prevent that) but the fact that Victor died at T (or perhaps, the fact that Victor died no later than T).[23] He suggested that perhaps Sassan is positively relevant to that, since most of the ways he could have behaved would have resulted in Victor's dying later than T. But we could, with minimal artifice, ensure that Baxter is disposed to kill Victor at exactly T if Sassan does not. For example, we could imagine that there is only a fraction of a second when Victor is vulnerable to a bullet and that Baxter is located closer to him (or has a faster acting gun) so that there is a moment T_2 such that if Sassan does not shoot at T_2, he will not succeed in killing Victor, but such that Baxter still has a chance to get a shot off at T_1 with the result that Victor will die at T.

Perhaps someone may try to argue that Sassan is positively relevant to something—the fact that Victor is killed with this bullet rather than that, or more simply the fact that Victor is killed by him rather than by Baxter. The latter suggestion will not do, since it begs the question. Bennett cannot assume that his account implies that Sassan kills Victor, since that is the very claim at issue. As to the suggestion that Sassan is positively relevant to Victor's being killed with this bullet rather than that, surely we could modify the story in such a way that if Sassan does not pull the trigger, Baxter can push a switch that will guarantee that the gun fires.

So it seems that Bennett is committed to the claim that Sassan is negatively relevant to (the various salient facts concerning) Victor's death, and hence, that Sassan let Victor die. This seems wrong.

7. The Transfer of Energy Account

Why do we think of Sassan's relevance to Victor's death as positive? Surely because Sassan ‘acts on’ him in a way that one does not ‘act on’ a drowning victim if one simply stands by and watches him drown.

Similarly, the woman who crushed the child's chest by pressing on it (while remaining immobile) acted on the child to cause his death. By contrast, in Raccoon, my role in the event of the raccoon's consuming my food was precisely not to act on the raccoon. In cases of ‘acting on’, it seems, physical forces run from the agent to the affected object or patient. This seems to distinguish typical cases of doing from typical cases of allowing harm.

Let's clarify the account. Obviously the agent need not act on the patient directly—it is enough that physical forces run from one to the other, however indirectly. Moreover, it cannot be sufficient for positive relevance to an upshot that physical forces run from the agent to the victim or patient at the appropriate time. The agent may have acted on the patient but done so to produce some other effect. For example, instead of drowning him or pulling him out, she throws him a rose. A passenger on a runaway trolley doesn't seem to kill the victim of that trolley in spite of the fact that his weight adds to the momentum of the trolley and, hence, there is a transfer of energy between the two. We should add that the way in which she acts on him must explain the upshot. In the case just described, because her throwing him a rose doesn't explain his death by drowning, she doesn't count as positively relevant to his death.[24]

A puzzle remains. What about cases where the agent removes a safety net from beneath a falling victim, unplugs a respirator, kicks a rock out of the path of the runaway vehicle, and other similar cases? No physical forces run from the agent to the victim. So, by the account under discussion, they are cases of negative relevance, and yet many of them, at least by many people, are confidently judged to be cases of positive relevance. Of course, such cases are by their nature troublesome and controversial. Unlike Sassan and Raccoon, these may be cases of “spoils to the victor”—cases we should classify in whatever way the otherwise best theory does.[25] The “acting on” account could fairly easily be adapted so that it treated such “safety net” cases as cases of positive relevance. Instead of insisting on one or another version of that account, let me outline three versions of it.

A. According to the first, safety net cases are cases of negative relevance because, although the agent acts on the net, the net does not act on the victim.
B. According to the second, they are cases of positive relevance because the agent acts on the net (physical forces run from the agent to the net) and the position of the net is a necessary part of the causal explanation of the victim's death. More generally: The agent is positively relevant to an upshot *U* if the agent acts on *X* to produce feature *F* in *X* and *X*s having *F* is a necessary part of the causal story leading to *U*.[26]
C. According to the third version, it is not a clear case of positive relevance because although she acts on the net, the net does not act on the victim, but nor is it a clear case of negative relevance, because the relevance of her behavior to the upshot is not in terms of her failure to act on something. It is a borderline case.

I submit that this distinction between cases where the agent acts on the victim and cases where she does not is at least one strand in the complex tangle underlying our commonsense distinction between doing and allowing. If that is right, what are the moral implications? Not clear, I think, but this distinction does

help explain our tendency to be more upset when we kill than when we let die. When physical forces run from the agent to the victim (even where the agent is uncontroversially innocent) there tends to be something like a jolt—as one experiences the death of which one is a causal factor. Consider a case where the victim jumps or is thrown in front of the agent's car. Although there is no question of the agent's guilt in this case, the event must feel more distressing to her than would a case where she is aware of a death that she cannot prevent.[27]

8. More on 'Safety Net' Cases

Jeff McMahan puts "safety net cases" front and center of his own account of the distinction, arguing that some of them qualify as killings and some as lettings die.[28] A number of factors distinguish the killings from the lettings die. "Among these are whether the person who terminates the aid or protection is the person who has provided it, whether the aid or protection is self-sustaining or requires more of the agent, and whether the aid or protection is operative or as yet inoperative."[29] Here are three cases he discusses:

- (Burning Building 1): A person trapped atop a high burning building leaps off. Seeing this, a firefighter quickly stations a self-standing net underneath. But he then immediately notices that two other persons have jumped from a window several yards away. He therefore repositions the net so that it catches the two. The first jumper then hits the ground and dies.
- (Burning Building 2): Just like (Burning Building 1) except that it is a second firefighter who repositions the net.
- (Burning Building 3): Just like (Burning Building 2) except that the second firefighter moves the safety net out of a malicious desire to kill.

Convinced that (Burning Building 2) is importantly like (Burning Building 1), McMahan points out that sometimes one agent can act on behalf of another, or they can act as a team (e.g., of firefighters) so that whether it is the first or the second firefighter doesn't matter, since they are acting in a capacity that is "role-based". In (Burning Building 3), however, where the second firefighter is motivated by malice against the victim, he is acting on his own and kills the victim.

The messy, somewhat ad hoc nature of McMahan's way of drawing the line is clearly a strike against it, as is the fact that it is clearly motivated by a desire to accommodate the moral intuition that killing is worse than letting die. If, however, some or all of the factors that McMahan lists as affecting the question of whether a case is one of killing or letting die can be seen to follow from some simpler, deeper account of the distinction, so much the better for that account and for McMahan's judgments about cases. Here's a way this might be done. It is somewhat arbitrary how we count actions—whether my typing the word 'word' involves four actions or just one, for example. Similarly, it is arbitrary whether the behavior of first writing a (potentially life-saving) check and then tearing it up count as a single action or as two. The simple action of tearing it up might be classified as a killing, whereas the complex 'act' of writing it and tearing it up might seem equivalent to the non-act of never writing it, and, hence, count as letting die. In asking whether the agent killed or let die in such a case, we may sometimes focus on the second (simple)

act and sometimes on the complex act. Whether the two are performed by the same person, the time between the two, whether the first was self-sustaining, etc. all affect our choice here. It seems that Bennett's or the counterfactual account may, with minimal artifice, be modified so that most of the cases that McMahan wants to classify as lettings die are so classified. For example, we might consider what would have happened if the agent had not been present (during the entire period—including the time of writing the check and the time of tearing it up.) or we might consider whether most of the ways the agent could have behaved throughout the entire period would have had the result that upshot obtained.

McMahan's suggestion that the firefighter kills if he acts with a malicious intention, whereas he lets die if he acts with a good intention seems wrong, however. Whatever else we think about doing and allowing, we should be able to distinguish them without reference to internal mental states of the agent.

9. Conclusion

This discussion suggests, I think, that “the distinction between doing and allowing" does not refer uniquely. More likely, it refers indefinitely to a tissue of largely overlapping distinctions—such as Bennett's, the counterfactual account, and the transfer of energy account, in addition to, (if we like) complex, conjunctive distinctions like Quinn's or McMahan's. The fact that each account faces counterexamples may not show that each is incorrect, but simply that none is the unique distinction. In that case, it seems that the sensible approach is to acknowledge this variety of distinctions, and to ask with respect to each, whether it is morally significant.[30] I believe my discussion has shown that there is no decisive reason to say that any of these distinctions is morally significant, as long, that is, as we remember that intention plays no part in the distinction between doing and allowing harm. I have no doubt that the intention with which an agent acts can make a difference to the moral status of her act. (Exactly how is a big question—for another paper.) The claim that doing harm is no worse than allowing harm flies in the face of powerful intuitions to the contrary. I believe that these intuitions can be partially explained away by pointing to other morally significant distinctions (distinctions concerning intentions, difficulty or ease of avoiding the harm, etc) that often coincide with the distinction between doing and allowing harm. A residue remains, however, and we seem faced with a conflict between theory and intuitions about cases.

Bibliography

Cited Works

- Bennett, Jonathan, “Acting and Refraining,” *Analysis* 28 (1967)
- -----, “Morality and Consequences,#148; *The Tanner Lectures on Human Values. II.* S. McMurrin (ed) Salt Lake City: University of Utah Press, 1981.
- -----, “Negation and Abstention: Two Theories of Allowing,#148; *Ethics*, 104 (October 1993), pp. 75-96
- -----, *The Act Itself*. Oxford: Clarendon Press, 1995.

- Callahan, Daniel, "Killing and Allowing to Die," *The Hastings Center Report*, vol. 19 (January/February 1989).
- Casey, John, "Killing and Letting Die: A Reply to Bennett," *Killing and Letting Die*, 1st. ed. Ed. B. Steinbock. Englewood Cliffs, N.J.: Prentice-Hall, 1980.
- Chandler, John, "Killing and Letting Die—Putting the Debate in Context," *Australasian Journal of Philosophy*, 68 (1990).
- Dinello, Daniel, "On Killing and Letting Die,#148; *Analysis*, Vol. 31 (1971), pp. 84-86.
- Donagan, Alan, *The Theory of Morality*, Chicago: The University of Chicago Press, 1977.
- Fischer, John Martin and Mark Ravizza, eds. *Ethics Problems and Principles*, Fort Worth: Harcourt, Brace, Jovanovich, 1992.
- Foot, Philippa, "Morality as a System of Hypothetical Imperatives," *Philosophical Review*, 81 (1972).
- -----, "The Problem of Abortion and the Doctrine of Double Effect," *Virtues and Vices and Other Essays*. Berkeley, CA: University of California Press, 1978.
- -----, "Morality, Action and Outcome," Ted Honderich, ed., *Morality and Objectivity* (London, England: Routledge and Kegan Paul, 1985).
- Frankfurt, Harry, "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 65 (1969).
- Glover, Jonathan, *Causing Death and Saving Lives*. London: Penguin, 1977.—"It Makes No Difference Whether or Not I Do It," *Proceedings of the Aristotelian Society*, 49 (1975).
- Hanser, Matthew, "Killing, Letting Die and Preventing People from Being Saved," *Utilitas* Vol. 11, No. 3, November 1999.
- Kagan, Shelly, The Limits of Morality Oxford: Oxford University Press, 1989.—"The Additive Fallacy," *Ethics*, 99 1988.
- Lewis, David, "Causation," in his *Philosophical Papers*, vol. 2. New York: Oxford University Press, 1986.
- Mackie, John, *The Cement of the Universe*. Oxford: Oxford University Press, 1974.
- Malm, Heidi, "In Defense of the Contrast Strategy," in Fischer and Ravizza, 1992.
- McMahan, Jeff, "Killing, Letting Die and Withdrawing Aid," *Killing and Letting Die*, 2nd ed. Norcross, A. and Steinbock, B. New York: Fordham University Press, 1994. (Originally published in *Ethics*, 103 (January 1993.)
- Munthe, Christian, "The Morality of Interference" in *Theoria*, vol. 65, issue 1, 1999.
- Quinn, Warren S., "Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing," *Killing and Letting Die*, 2nd. Ed. (ed. Norcross, A. and Steinbock, B.) New York: Fordham University Press, 1994. (Originally published in *Philosophical Review* 98, no. 3 (July 1989)).
- Rachels, James, "Active and Passive Euthanasia," *New England Journal of Medicine* 292 (1975). "Killing and Letting People Die of Starvation," *Philosophy*, 54, no. 208 (April 1979).
- Scheffler, Samuel, *The Rejection of Consequentialism*. Oxford: Clarendon Press, 1982.
- Singer, Peter, "Famine, Affluence and Morality," *Philosophy and Public Affairs*, 1. no. 3 (1965).
- Smart, J.J.C, "An Outline of a System of Utilitarian Ethics," in J.J.C. Smart and Bernard Williams, Utilitarianism For and Against. Cambridge: Cambridge University Press, 1973. *Utilitarianism For and Against*. Cambridge: Cambridge University Press, 1973.

- Strudler, Alan and David Wasserman, "The First Dogma of Deontology: the Doctrine of Doing and Allowing and the Notion of a Say," *Philosophical Studies*.
- Thomson, Judith Jarvis, "Killing, Letting Die and the Trolley Problem," *Rights, Restitution, and Risk: Essays in Moral Theory*. Ed. W. Parent. Cambridge: Harvard University Press, 1986.
- -----, *The Realm of Rights*. Cambridge: Harvard University Press, 1990.
- Tooley, Michael "Abortion and Infanticide," *Philosophy and Public Affairs*, 2, no. 1 (1972).
- Trammell, Richard, "Saving and Taking Life," *The Journal of Philosophy* 72 (1975).
- -----, "The Presumption Against Taking Life," *The Journal of Medicine and Philosophy*, 3 (1978).
- -----, "The Nonequivalency of Saving Life and Not Taking Life." *The Journal of Medicine and Philosophy*, 4, no. 3 (September 1979).
- Unger, Peter, *Living High and Letting Die*. Oxford: Oxford University Press, 1996.
- Williams, Bernard, "A Critique of Utilitarianism," in J.J.C. Smart and Bernard Williams, Utilitarianism For and Against. Cambridge: Cambridge University Press, 1973.

Further Reading

- Anscombe, G.E.M., *Intention*. Oxford: Basil Blackwell, 1958.
- -----, "Modern Moral Philosophy," *Ethics, Religion, and Politics. Collected Philosophical Papers. III*. Minneapolis: University of Minnesota Press, 1981.
- Aronsen, J., "On the Grammar of Cause," *Synthese*, 22, 1971.
- Davidson, Donald, "Agency," *Essays on Actions and Events*. Oxford: Oxford University Press, 1980.
- Feinberg, Joel, *Doing and Deserving*. Princeton: Princeton University Press, 1970.
- Haslett, D.W., "Moral Taxonomy and Rachels' Thesis," *Public Affairs Quarterly*, vol. 10, Number 4, October 1996.
- Horowitz, Tamara, "Philosophical Intuition and Psychological Theory," *Ethics*, vol. 108. Issue 2 (1998).
- Howard-Snyder, Frances, "The Heart of Consequentialism," *Philosophical Studies* 1994.
- Isaacs, Tracy, "Moral Theory and Action Theory, Killing and Letting Die," *American Philosophical Quarterly*, Vol. 32. Issue. 4, 1995.
- Kamm, Frances Myrna, "Killing and Letting Die: Methodological and Substantive Issues," *Pacific Philosophical Quarterly*, 64 (1983).
- -----, "Harming Some to Save Others," *Philosophical Studies*, 57 (1989).
- -----, *Morality, Mortality*. 2 vols. Oxford: Oxford University Press, 1993 and 1996.
- McMahon, Christopher, "The Paradox of Deontology," Philosophy & *Public Affairs* 20 (1991).
- Nagel, Thomas, *The View from Nowhere*. Oxford: Oxford University Press, 1986. Norcross, Alastair and Steinbock, Bonnie, ed., Killing and Letting Die, 2nd Ed. New York: Fordham University Press, 1994.
- Oddie, Graham, "Killing and Letting Die: Bare Differences and Clear Differences," *Philosophical Studies*, vol. 88. Issue 3 (1997).
- Rickless, Samuel, "The Doctrine of Doing and Allowing," *Philosophical Review* 1997 vol. 106 (4).
- Russell, Bruce, "On the Relative Strictness of Negative and Positive Duties," *Killing and Letting*

Die, 1st ed. Ed. B. Steinbock. Englewood, Cliffs, N.J.: Prentice-Hall, 1980.

-

Other Internet Resources

- Euthanasia and End-of-Life Decisions (maintained by Lawrence Hinman, University San Diego)
- Entry on Euthanasia, in the *Internet Encyclopedia of Philosophy* (J. Fieser, (ed.), U. Tennessee/Martin)
- Entry on Abortion, in the *Internet Encyclopedia of Philosophy* (J. Fieser, (ed.), U. Tennessee/Martin)

Related Entries

causation: counterfactual theories of | causation: the metaphysics of | consequentialism | euthanasia: voluntary

Acknowledgement

I am grateful to Jonathan Bennett, Tom Downing, Dan Howard-Snyder, Hud Hudson, Phillip Montague, Alastair Norcross, John Hawthorne, Stuart Rachels and Kadri Vihvelin for comments on earlier drafts of this paper. I am also grateful to the Bureau for Faculty Research at Western Washington University for support while writing it.

Frances Howard-Snyder
franhs@cc.wwu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 14, 2002
Content last modified: May 14, 2002

Stanford Encyclopedia of Philosophy
Notes to Doing vs. Allowing Harm

Notes

1. This issue is relevant also to the debate over abortion, the debate over capital punishment, and the more theoretical debate over consequentialism. See Scheffler, Smart and Williams.

2. James Rachels, “Active and Passive Euthanasia,”. See also Michael Tooley, “Abortion and Infanticide,” *Philosophy and Public Affairs* 2 1972, Peter Singer, *Practical Ethics*,1979 and Peter Unger, *Living High and Letting Die*.

3. “The Additive Fallacy”.

4. *Ibid*., 259. See Heidi Malm (1992) for a response to Kagan.

5. We talk about allowing the tree to fall, but we do not talk about “doing the tree fall”. We do of course talk about “making the tree fall” and “causing the tree to fall,” but I am reluctant to use these terms, because they suggest that allowing the tree to fall is not a way of causing it to fall, and, as I shall argue below, this is wrong.

6. See, for example, Daniel Callahan, 1989.

7. This idea, which has its roots in Hume's writings, is developed in Lewis, 1986. Note that the definition I mention in the text represents the gut idea of the counterfactual account of causation. Subsequent versions are much more complicated, but they all imply that letting die is a kind of causing.

8. See Mackie, 1974. According to “transference” accounts of causation, causation consists in the transfer of energy or momentum from one object to another at the point of contact between the objects. Such accounts would indeed imply that allowing was not causing. See J. Aronsen, 1971 . But it seems then that a different premise in the argument for the moral insignificance of the distinction would be in trouble. The thought that one cannot be morally responsible for what one is not causally responsible for now seems dubious. It seemed correct, before, I conjecture, because, on the counterfactual accounts, it is very close to the principle that ‘ought’ implies ‘can’. If the transference account of causation is correct, then lack of causal responsibility does not imply inability to make a difference.

9. Or “ if the agent had never existed.” See Shelly Kagan, 1989, p. 94 ff for this version.

10. It might be objected that, had Franz been absent, the death would not have occurred, because the victim would have died a different death, since he would have been killed by Hans rather than by Franz. It seems to me that this response fails, since it is simply asserting the very point at issue: that the counterfactual account implies that Franz killed.

11. See Donagan, 1977.

12. See Christian Munthe, 1999, for an interesting attack on the claim that the counterfactual distinction is morally significant. Munthe argues that if true, this claim would violate the rule that ‘ought’ implies ‘can’.

13. See Warren Quinn, “Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing.”.

14. Quinn uses the expression, “harmful positive agency” for doing harm and “harmful negative agency” for allowing harm. This idea can be traced back at least to Anscombe and Davidson (see Anscombe, 1958 and Davidson, 1980.)

15. See Foot, 1985. Note that Quinn's argument for the moral significance of the distinction doesn't seem to hang on his account of the distinction.

16. *Ibid*., p. 156.

17. See Strudler and Wasserman for an interesting critique of this point.

18. Bennett has developed and defended these ideas in a series of articles and a book. See Bennett, 1967, 1981, 1993, 1995.

19. This nutshell leaves out much important detail. See the original.

20. For example, Daniel Dinello (1971). Quinn makes a similar point in “Actions, Intentions, and Consequences.”

21. If you feel that whether this counts as a killing depends on other factors such as what the agent knows or desires, add whatever features will make it seem most favorable to me. My point is simply that sometimes an agent can kill while remaining completely immobile. Robin Dillon helped me to see this point. Bennett offers several other reasons for thinking that the immobility objection is not decisive against him. See p. 98 ff of *The Act Itself*.

22. This objection was suggested to me by remarks of Paul Wagoner. Note that Sassan bears obvious resemblances to Harry Frankfurt's famous counterexample to the Principle of Alternative Possibilities.

See Frankfurt, 1969. It is importantly different from that example, however.

23. In correspondence.

24. I am indebted to John Hawthorne for with the ideas of this paragraph. Does every case of this sort have a victim? It is arguable that every upshot consists in a state or event occurring in some object or objects. Let that object or objects be the 'victim' with which, directly or indirectly, the agent interacts.

25. I was surprised to learn—from Jonathan Schaffer—that firing a gun is a safety net case—that most triggers work by removing a barrier and "allowing" pent-up energy to release the bullet. This example suggests that we had better put safety net cases on the positive side of the line.

26. I need to insist that the agent is not identical with *X*. Without this qualification, some very standard cases of allowing would count as positive relevance, since the agent might be said to act on herself to prevent herself from saving the victim. I'm not sure that it even makes sense to talk of an agent acting on herself, or of forces running from her to herself, but if it does, let us exclude such cases.

27. Bernard Williams discusses a case like this. See his 1973, pp. 98-99.

28. In "Killing, Letting Die and Withdrawing Aid". See Tracy Isaacs (1995) for a powerful critique of McMahan.

29. *Ibid*., (in Norcross and Steinbock) p. 396.

30. Matthew Hanser argues that we should recognize a third category here: preventing someone from being saved. I think this treatment of "safety net" cases can be seen as confirming my point about the tissue of overlapping distinctions. Clear cases of killing fall on the positive side of all of the candidate distinctions, clear cases of allowing death fall on the negative side of all the distinctions; indeterminate cases fall on one side of one distinction and on the other side of another distinction. Treating these as a third category is just another way of making the same point. See Hanser, 1999.

Frances Howard-Snyder
franhs@cc.wwu.edu

First published: May 14, 2002
Content last modified: May 14, 2002

Privacy

The term “privacy” is used frequently in ordinary language as well as in philosophical, political and legal discussions, yet there is no single definition or analysis or meaning of the term. The concept of privacy has broad historical roots in sociological and anthropological discussions about how extensively it is valued and preserved in various cultures. Moreover, the concept has historical origins in well known philosophical discussions, most notably Aristotle’s distinction between the public sphere of political activity and the private sphere associated with family and domestic life. Yet historical use of the term is not uniform, and there remains confusion over the meaning, value and scope of the concept of privacy.

Early treatises on privacy appeared with the development of privacy protection in American law from the 1890’s onward, and privacy protection was justified largely on moral grounds. This literature helps distinguish *descriptive* accounts of privacy, describing what is in fact protected as private, from *normative* accounts of privacy defending its value and the extent to which it should be protected. In these discussions some treat privacy as an *interest* with moral value, while others refer to it as a moral or legal *right* that ought to be protected by society or the law. Clearly one can be insensitive to another’s privacy interests without violating any right to privacy, if there is one.

There are several skeptical and critical accounts of privacy. According to one well known argument there is no right to privacy and there is nothing special about privacy, because any interest protected as private can be equally well explained and protected by other interests or rights, most notably rights to property and bodily security (Thomson, 1975). Other critiques argue that privacy interests are not distinctive because the personal interests they protect are economically inefficient (Posner, 1981) or that they are not grounded in any adequate legal doctrine (Bork, 1990). Finally, there is the feminist critique of privacy, that granting special status to privacy is detrimental to women and others because it is used as a shield to dominate and control them, silence them, and cover up abuse (MacKinnon, 1989).

Nevertheless, most theorists take the view that privacy is a meaningful and valuable concept. Philosophical debates concerning definitions of privacy became prominent in the second half of the twentieth century, and are deeply affected by the development of privacy protection in the law. Some defend privacy as focusing on control over information about oneself (Parent, 1983), while others defend it as a broader concept required for human dignity (Bloustein, 1964), or crucial for intimacy (Gerstein, 1978; Inness, 1992). Other commentators defend privacy as necessary for the development of varied and meaningful interpersonal relationships (Fried, 1970, Rachels, 1975), or as the value that accords us the ability to control the access others have to us (Gavison, 1980; Allen, 1988), or as a set of norms necessary not only to control access but also to enhance personal expression and choice (Schoeman,

1992), or some combination of these (DeCew, 1997). Discussion of the concept is complicated by the fact that privacy appears to be something we value to provide a sphere within which we can be free from interference by others, and yet it also appears to function negatively, as the cloak under which one can hide domination, degradation, or physical harm to women and others.

This essay will discuss all of these topics, namely, (1) the historical roots of the concept of privacy, including the development of privacy protection in tort and constitutional law, and the philosophical responses that privacy is merely reducible to other interests or is a coherent concept with fundamental value, (2) the critiques of privacy as a right, (3) the wide array of philosophical definitions or defenses of privacy as a concept, providing alternative views on the meaning and value of privacy (and whether or not it is culturally relative), as well as (4) the challenges to privacy posed in an age of technological advance. Overall, most writers defend the value of privacy protection despite the difficulties inherent in its definition and its potential use to shield abuse. A contemporary collection of essays on privacy provides strong evidence to support this point (Paul *et al*., 2000). The contributing authors examine various aspects of the right to privacy and its role in moral philosophy, legal theory, and public policy. They also address justifications and foundational arguments for privacy rights.

---

1. History

Aristotle's distinction between the public sphere of politics and political activity, the *polis*, and the private or domestic sphere of the family, the *oikos*, as two distinct spheres of life, is a classic reference to a private domain. The public/private distinction is also sometimes taken to refer to the appropriate realm of governmental authority as opposed to the realm reserved for self-regulation, along the lines described by John Stuart Mill in his essay, *On Liberty*. Furthermore, the distinction arises again in Locke's discussion of property in his *Second Treatise on Government*. In the state of nature all the world's bounty is held in common and is in that sense public. But one possesses oneself and one's own body, and one can also acquire property by mixing one's labor with it, and in these cases it is one's private property. Margaret Mead and other anthropologists have demonstrated the ways various cultures protect privacy through concealment, seclusion or by restricting access to secret ceremonies (Mead, 1949). Alan Westin (1967) has surveyed studies of animals demonstrating that a desire for privacy is not restricted to humans. However, what is termed private in these multiple contexts varies. Privacy can refer to a sphere separate from government, a domain inappropriate for governmental interference, forbidden views and knowledge, solitude, or restricted access, to list just a few.

1.1 Informational Privacy

More systematic written discussion of the concept of privacy is often said to begin with the famous essay by Samuel Warren and Louis Brandeis titled "The Right to Privacy" (Warren and Brandeis, 1890). Citing "political, social, and economic changes" and a recognition of "the right to be let alone" they argued that existing law afforded a way to protect the privacy of the individual, and they sought to explain the nature and extent of that protection. Focusing in large part on the press and publicity allowed by recent inventions such as photography and newspapers, but referring as well to violations in other contexts, they emphasized the invasion of privacy brought about by public dissemination of details relating to a person's private life. Warren and Brandeis felt a variety of existing cases could be protected under a more general right to privacy which would protect the extent to which one's thoughts, sentiments, and emotions could be shared with others. Urging that they were not attempting to protect the items produced, or intellectual property, but rather the peace of mind attained with such protection, they said the right to privacy was based on a principle of "inviolate personality" which was part of a general right of immunity of the person, "the right to one's personality" (Warren and Brandeis 1890, 195, 215). The privacy principle, they believed, was already part of common law and the protection of one's home as one's castle, but new technology made it important to explicitly and separately recognize this protection under the name of privacy. They suggested that limitations of the right could be determined by analogy with the law of slander and libel, and would not prevent publication of information about public officials running for office, for example. Warren and Brandeis thus laid the foundation for a concept of privacy that has come to be known as control over information about oneself.

Although the first cases after the publication of their paper did not recognize a privacy right, soon the public and both state and federal courts were endorsing and expanding the right to privacy. In an attempt

to systematize and more clearly describe and define the new right of privacy being upheld in tort law, William Prosser wrote in 1969 that what had emerged were four different interests in privacy. Not claiming to be providing an exact definition, and admitting that there had been confusion and inconsistencies in the development of privacy protection in the law, Prosser nevertheless described the four “rather definite” privacy rights as follows:

1. Intrusion upon a person’s seclusion or solitude, or into his private affairs.
2. Public disclosure of embarrassing private facts about an individual.
3. Publicity placing one in a false light in the public eye.
4. Appropriation of one’s likeness for the advantage of another (Prosser 1969, 389).

Prosser noted that the intrusion in the first privacy right had expanded beyond physical intrusion, and pointed out that Warren and Brandeis had been concerned primarily with the second privacy right. Nevertheless, Prosser felt that both real abuses and public demand had led to general acceptance of these four types of privacy invasions. On his view, answers to three main questions were at the time as yet unclear: i) whether appearance in public implied forfeiture of privacy, ii) whether facts part of a “public record” could still be private, and iii) whether a significant lapse of time affected the privacy of revelations. Note that Warren and Brandeis were writing their normative views about what they felt should be protected under the rubric of privacy, whereas Prosser was describing what courts had in fact protected in the 70 years following publication of the Warren and Brandeis paper. Thus it is not surprising that their descriptions of privacy differ. Because the Supreme Court has been explicit in ruling that privacy is a central reason for Fourth Amendment protection, privacy as control over information about oneself has come to be viewed by many as also including protection against unwarranted searches, eavesdropping, surveillance, and appropriation and misuses of one’s communications.

1.2 The Constitutional Right to Privacy

In 1965 a quite different right to privacy, independent of informational privacy and the Fourth Amendment, was recognized explicitly by the Supreme Court. It is now commonly called the constitutional right to privacy. The right was first announced in the *Griswold v. Connecticut* (381 U.S. 479) case, which overturned convictions of the Director of Planned Parenthood and a doctor at Yale Medical School for dispersing contraceptive related information, instruction, and medical advice to married persons. The constitutional right to privacy was described by Justice William O. Douglas as protecting a zone of privacy covering the social institution of marriage and the sexual relations of married persons. Despite controversy over Douglas’ opinion, the constitutional privacy right was soon cited to overturn a ban against interracial marriage, to allow individuals to possess obscene matter in their own homes, and to allow distribution of contraceptive devices to individuals, both married and single. The most famous application of this right to privacy was as one justification of abortion rights defended in 1973 in *Roe v. Wade* (410 U.S. 113) and subsequent decisions on abortion. While Douglas vaguely called it a “penumbral” right “emanating” from the Constitution, and the Court has been unable to clearly define the right, it has generally been viewed as a right protecting one’s individual interest in independence in making certain important and personal decisions about one’s family, life and lifestyle.

Which personal decisions have been protected by this privacy right has varied depending on the makeup of the Court. In 1986 in *Bowers v. Hardwick* (478 U.S. 186) privacy was not held to cover a ban on anti-sodomy laws in Georgia, despite the intimate sexual relations involved.

1.3 Reductionism vs. Coherentism

One way of understanding the growing literature on privacy is to view it as divided into two main categories, which we may call *reductionism* and *coherentism*. Reductionists are generally critical of privacy, while coherentists defend the coherent fundamental value of privacy interests. Ferdinand Schoeman (1984) introduced somewhat different terminology which makes it easier to understand this distinction. According to Schoeman, a number of authors have believed

> ...there is something fundamental, integrated, and distinctive about the concerns traditionally grouped together under the rubric of "privacy issues." In opposing this position, some have argued that the cases labeled "privacy issues" are diverse and disparate, and hence are only nominally or superficially connected. Others have argued that when privacy claims are to be defended morally, the justifications must allude ultimately to principles which can be characterized quite independently of any concern with privacy. Consequently, the argument continues, there is nothing morally distinctive about privacy. I shall refer to the position that there is something common to most of the privacy claims as the "coherence thesis." The position that privacy claims are to be defended morally by principles that are distinctive to privacy I shall label the "distinctiveness thesis."
>
> Theorists who deny both the coherence thesis and the distinctiveness thesis argue that in each category of privacy claims there are diverse values at stake of the sort common to many other social issues and that these values exhaust privacy claims. The thrust of this complex position is that we could do quite well if we eliminated all talk of privacy and simply defended our concerns in terms of standard moral and legal categories (Schoeman 1984, 5).

These latter theorists, who reject both Schoeman's coherence thesis and distinctiveness thesis, may be referred to as *reductionists*, for they view what are called privacy concerns as analyzable or reducible to claims of other sorts, such as infliction of emotional distress or property interests. They deny that there is anything useful in considering privacy as a separate concept. They conclude, then, that there is nothing coherent, distinctive or illuminating about privacy interests.

On the other side, more theorists have argued that there is something fundamental and distinctive and coherent about the various claims that have been called privacy interests. On this view, privacy has value as a coherent and fundamental concept, and most individuals recognize it as a useful concept as well. Those who endorse this view may be called *coherentists*. Nevertheless, it is important to recognize that coherentists have quite diverse, and sometimes overlapping, views on what it is that is distinctive about

privacy and what links diverse privacy claims.

2. Critiques of Privacy

2.1 Thomson's Reductionism

Probably the most famous reductionist view of privacy is one from Judith Jarvis Thomson (1975). Noting that there is little agreement on what privacy is, Thomson examines a number of cases that have been thought to be violations of the right to privacy. On closer inspection, however, Thomson believes all those cases can be adequately and equally well explained in terms of violations of property rights or rights over the person, such as a right not to be listened to. Ultimately the right to privacy, on Thomson's view, is merely a cluster of rights. Those rights in the cluster are always overlapped by, and can be fully explained by, property rights or rights to bodily security. The right to privacy, on her view, is "derivative" in the sense that there is no need to find what is common in the cluster of privacy rights. Privacy is derivative in its importance and justification, according to Thomson, as any privacy violation is better understood as the violation of a more basic right. Numerous commentators provide strong arguments against Thomson's critique (Scanlon, 1975; Inness, 1992).

2.2 Posner's Economic Critique

Richard Posner (1981) also presents a critical account of privacy, arguing that the kinds of interests protected under privacy are not distinctive. Moreover, his account is unique because he argues that privacy is protected in ways that are economically inefficient. With respect to information, on Posner's view privacy should only be protected when access to the information would reduce its value (e.g. allowing students access to their letters of recommendation make those letters less reliable and thus less valuable, and hence they should remain confidential or private). Focusing on privacy as control over information about oneself, Posner argues that concealment or selective disclosure of information is usually to mislead or manipulate others, or for private economic gain, and thus protection of individual privacy is less defensible than others have thought because it does not maximize wealth. In sum, Posner defends organizational or corporate privacy as more important than personal privacy, because the former is likely to enhance the economy.

2.3 Bork's View

Another strong critic of privacy is Robert Bork (1990), whose criticism is aimed at the constitutional right to privacy established by the Supreme Court in 1965. Bork views the *Griswold v. Connecticut* decision as an attempt by the Supreme Court to take a side on a social and cultural issue, and as an example of bad constitutional law. Bork's attack is focused on Justice William O. Douglas and his majority opinion in *Griswold*. Bork's major point is that Douglas did not derive the right to privacy from some pre-existing right or from natural law, but merely created a new right to privacy with no foundation in the Constitution or Bill of Rights. Bork is correct that the word "privacy" never appears in those

documents. Douglas had argued, however, that the right to privacy could be seen to be based on guarantees from the First, Third, Fourth, Fifth, and Ninth Amendments. Taken together, the protections afforded by these Amendments showed that a basic zone of privacy was protected for citizens, and that it covered their ability to make personal decisions about their home and family life. In contrast, Bork argues i) that none of the Amendments cited covered the case before the Court, ii) that the Supreme Court never articulated or clarified what the right to privacy was or how far it extended, and he charges iii) that the privacy right merely protected what a majority of justices personally wanted it to cover. In sum, he accuses Douglas and the Court majority of inventing a new right, and thus overstepping their bounds as judges by making new law, not interpreting the law.

Theorists including William Parent (1983) and Judith Thomson (1975) argue that the constitutional right to privacy is not really a privacy right, but is more aptly described as a right to liberty. Other commentators believe, to the contrary, that even if Douglas' opinion is flawed in its defense, using vague language about a penumbral privacy right emanating from the Constitution and its Amendments, there is nevertheless a historically and conceptually coherent notion of privacy, distinct from liberty, carved out by the constitutional privacy cases (Inness, 1992; Schoeman, 1992; Johnson, 1994; DeCew, 1997).

2.4 The Feminist Critique of Privacy

There is no single version of the feminist critique of privacy, yet it can be said in general that many feminists worry about the darker side of privacy, and the use of privacy as a shield to cover up domination, degradation and abuse of women and others. If distinguishing public and private realms leaves the private domain free from any scrutiny, then these feminists such as Catharine MacKinnon (1989) are correct that privacy can be dangerous for women when it is used to cover up repression and physical harm to them by perpetuating the subjection of women in the domestic sphere and encouraging nonintervention by the state. Jean Bethke Elshtain (1981, 1995) and others suggest that it appears feminists such as MacKinnon are for this reason rejecting the public/private split, and are, moreover, recommending that feminists and others jettison or abandon privacy altogether. But, Elshtain points out, this alternative seems too extreme.

A more reasonable view, according to Anita Allen (1988), is to recognize that while privacy can be a shield for abuse, it is unacceptable to reject privacy completely based on harm done in private. A total rejection of privacy makes everything public, and leaves the domestic sphere open to complete scrutiny and intrusion by the state. Yet women surely have an interest in privacy that can protect them from state imposed sterilization programs or government imposed drug tests for pregnant women mandating results sent to police, for instance, and that can provide reasonable regulations such as granting rights against marital rape. Thus collapsing the public/private dichotomy into a single public realm is inadequate. What puzzles feminists is how to make sense of an important and valuable notion of privacy that provides them a realm free from scrutiny and intervention by the state, without reverting to the traditional public/private dichotomy that has in the past relegated women to the private and domestic sphere where they are victims of abuse and subjection. The challenge is to find a way for the state to take very seriously the domestic abuse that used to be allowed in the name of privacy, while also preventing the state from

insinuating itself into all the most intimate parts of women's lives. This means drawing new boundaries for justified state intervention and thus understanding the public/private distinction in new ways.

3. Views on the Meaning and Value of Privacy

3.1 Privacy and Control over Information

Narrow views of privacy focusing on control over information about oneself that were defended by Warren and Brandeis and by William Prosser are also endorsed by more recent commentators including Fried (1970) and Parent (1983). In addition, Alan Westin describes privacy as the ability to determine for ourselves when, how, and to what extent information about us is communicated to others (Westin, 1967). Perhaps the best example of a contemporary defense of this view is put forth by William Parent. Parent explains that he proposes to defend a view of privacy that is consistent with ordinary language and does not overlap or confuse the basic meanings of other fundamental terms. He defines privacy as the condition of not having undocumented personal information known or possessed by others. Parent stresses that he is defining the condition of privacy, as a moral value for people who prize individuality and freedom, and not a moral or legal right to privacy. Personal information is characterized by Parent as factual (otherwise it would be covered by libel, slander or defamation), and these are facts that most persons choose not to reveal about themselves, such as facts about health, salary, weight, sexual orientation, etc. Personal information is documented, on Parent's view, only when it belongs to the public record, that is, in newspapers, court records, or other public documents. Thus, once information becomes part of a public record, there is no privacy invasion in future releases of the information, even years later or to a wide audience, nor does snooping or surveillance intrude on privacy if no undocumented information is gained. In cases where no new information is acquired, Parent views the intrusion as irrelevant to privacy, and better understood as an abridgment of anonymity, trespass, or harassment. Furthermore, what has been described above as the constitutional right to privacy, is viewed by Parent as better understood as an interest in liberty, not privacy. In sum, there is a loss of privacy on Parent's view, only when others acquire undocumented personal information about an individual. DeCew (1997) gives a detailed critique of Parent's position.

3.2 Privacy and Human Dignity

In an article written mainly as a defense of Warren and Brandeis' paper and as a response to William Prosser, Edward J. Bloustein (1964) argues that there is a common thread in the diverse legal cases protecting privacy. According to Bloustein, Warren and Brandeis failed to give a positive description of privacy, however they were correct that there was a single value connecting the privacy interests, a value they called "inviolate personality." On Bloustein's view it is possible to give a general theory of individual privacy that reconciles its divergent strands, and "inviolate personality" is the social value protected by privacy. It defines one's essence as a human being and it includes individual dignity and integrity, personal autonomy and independence. Respect for these values is what grounds and unifies the concept of privacy. Discussing each of Prosser's four types of privacy rights in turn, Bloustein defends

the view that each of these privacy rights is important because it protects against intrusions demeaning to personality and against affronts to human dignity. Using this analysis, Bloustein explicitly links the privacy rights in tort law described by Prosser with privacy protection under the Fourth Amendment. He urges that both leave an individual open to scrutiny in a way that leaves one's autonomy and sense of oneself as a person vulnerable, violating one's human dignity and moral personality. The common conceptual thread linking diverse privacy cases prohibiting dissemination of confidential information, eavesdropping, surveillance, and wiretapping, to name a few, is the value of protection against injury to individual freedom and human dignity. Invasion of privacy is best understood, in sum, as affront to human dignity. Although Bloustein admits the terms are somewhat vague, he defends this analysis as conceptually coherent and illuminating.

3.3 Privacy and Intimacy

A more common view has been to argue that privacy and intimacy are deeply related. On one account, privacy is valuable because intimacy would be impossible without it (Fried, 1970; Gerety 1977; Gerstein, 1978). Fried, for example, defines privacy narrowly as control over information about oneself. He extends this definition, however, arguing that privacy has intrinsic value, and is necessarily related to and fundamental for one's development as an individual with a moral and social personality able to form intimate relationships involving respect, love, friendship and trust. Privacy is valuable because it allows one control over information about oneself, which allows one to maintain varying degrees of intimacy. Indeed, love, friendship and trust are only possible if persons enjoy privacy and accord it to each other. Privacy is essential for such relationships on Fried's view, and this helps explain why a threat to privacy is a threat to our very integrity as persons. By characterizing privacy as a necessary context for love, friendship and trust, Fried is basing his account on a moral conception of persons and their personalities, on a Kantian notion of the person with basic rights and the need to define and pursue one's own values free from the impingement of others. Privacy allows one the freedom to define one's relations with others and to define oneself. In this way, privacy is also closely connected with respect and self respect.

Gerstein (1978) argues as well that privacy is necessary for intimacy, and intimacy in communication and interpersonal relationships is required for us to fully experience our lives. Intimacy without intrusion or observation is required for us to have experiences with spontaneity and without shame. Shoeman (1984) endorses these views and stresses that privacy provides a way to control intimate information about oneself and that has many other benefits, not only for relationships with others, but also for the development of one's personality and inner self. Julie Inness (1992) has identified intimacy as the defining feature of intrusions properly called privacy invasions. Inness argues that intimacy is based not on behavior, but on motivation. Inness believes that intimate information or activity is that which draws its meaning from love, liking, or care. It is privacy that protects one's ability to retain intimate information and activity so that one can fulfill one's needs of loving and caring.

3.4 Privacy and Social Relationships

A number of commentators defend views of privacy that link closely with accounts stressing privacy as

required for intimacy, emphasizing not just intimacy but also more generally the importance of developing diverse interpersonal relationships with others. Rachels (1975) acknowledges there is no single answer to the question why privacy is important to us, because it can be necessary to protect one's assets or interests, or to protect one from embarrassment, or to protect one against the deleterious consequences of information leaks, to name just a few. Nevertheless, he explicitly criticizes Thomson's reductionist view, and urges that privacy is a distinctive right. He basically defends the view that privacy is necessary to maintain a variety of social relationships, not just intimate ones. Privacy accords us the ability to control who knows what about us and who has access to us, and thereby allows us to vary our behavior with different people so that we may maintain and control our various social relationships, many of which will not be intimate. An intriguing part of Rachels' analysis of privacy is that it emphasizes ways in which privacy is not merely limited to control over information. Our ability to control both information and access to us allows us to control our relationships with others. Hence privacy is also connected to our behavior and activities.

3.5 Privacy and Restricted Access

Another group of theorists characterize privacy in terms of access. Some commentators describe privacy as exclusive access of a person to a realm of his or her own, and Sissela Bok (1982) argues that privacy protects us from unwanted access by others - either physical access or personal information or attention. Ruth Gavison (1980) defends this more expansive view of privacy in greater detail, arguing that interests in privacy are related to concerns over accessibility to others, that is, what others know about us, the extent to which they have physical access to us, and the extent to which we are the subject of the attention of others. Thus the concept of privacy is best understood as a concern for limited accessibility and one has perfect privacy when one is completely inaccessible to others. Privacy can be gained in three independent but interrelated ways: through secrecy, when no one has information about one, through anonymity, when no one pays attention to one, and through solitude, when no one has physical access to one. Gavison's view is that the concept of privacy is this complex of concepts all part of the notion of accessibility. Furthermore, the concept is also coherent because of the related functions privacy has, namely "the promotion of liberty, autonomy, selfhood, human relations, and furthering the existence of a free society" (Gavison 1980, 347).

Carefully reviewing these various views, Anita Allen (1988) also characterizes privacy as denoting a degree of inaccessibility of persons, their mental states, and information about them to the senses and surveillance of others. She views seclusion, solitude, secrecy, confidentiality, and anonymity as forms of privacy. She also urges that privacy is required by the liberal ideals of personhood, and the participation of citizens as equals. While her view appears to be similar to Gavison's, Allen suggests her restricted access view is broader than Gavison's. This is in part because Allen emphasizes that in public and private women experience privacy losses that are unique to their gender. Noting that privacy is neither a presumptive moral evil nor an unquestionable moral good, Allen nevertheless defends more extensive privacy protection for women in morality and the law. Using examples such as sexual harassment, victim anonymity in rape cases, and reproductive freedom, Allen emphasizes the moral significance of extending privacy protection for women. In some ways her account can be viewed as one reply to the

feminist critique of privacy, allowing that privacy can be a shield for abuse, but can also be so valuable for women that privacy protection should be enhanced, not diminished.

3.6 The Scope of Privacy

There is a further issue that has generated disagreement, even among those theorists who believe privacy is a coherent concept. The question is whether or not the constitutional right to privacy, and the constitutional privacy cases described involving personal decisions about lifestyle and family including birth control, interracial marriage, viewing pornography at home, abortion, and so on, delineate a genuine category of privacy issues, or merely raise questions about liberty of some sort. Parent (1983) explicitly excludes concerns about one's ability to make certain important personal decisions about one's family and lifestyle as genuine privacy issues, saying the constitutional right to privacy cases focus solely on liberty. Among the others who take this view are Henkin (1974), Thomson (1975), Gavison (1980), and Bork (1990). Allen (1988) defines privacy in terms of access and excludes from her definition protection of individual autonomous choice from governmental interference, which she terms a form of liberty. Yet she refers to this latter protection as "decisional privacy" and says determining its category is purely a definitional point and one of labels. Ultimately she believes interference with decisions involving procreation and sexuality raise the same moral concerns as other privacy intrusions, offending the values of personhood. The Supreme Court now claims (*Whalen v. Roe*, 429 U.S. 589, 1977) that there are two different dimensions to privacy: both control over information about oneself and control over one's ability to make certain important types of decisions.

Following this sort of reasoning, a number of theorists defend the view that privacy has broad scope, inclusive of the multiple types of privacy issues described by the Court, even though there is no simple definition of privacy. Most of these theorists explore the links between the types of privacy interests and the similarity of reasons for valuing each. Some stress that privacy is necessary for one to develop a concept of self as a purposeful, self determining agent. Privacy enables control over personal information as well as control over our bodies and personal choices for our concept of self (Kupfer, 1987). Some emphasize the importance of intimacy for all privacy issues, noting the need for privacy to protect intimate information about oneself, access to oneself, as well as intimate relationships and decisions about one's actions (Inness, 1992). Some focus on the importance of privacy norms that allow one to restrict others' access to them as well as privacy norms that enable and enhance personal expression and the development of relationships. Privacy provides protection against overreaching social control by others through their access to information or their control over decision making (Schoeman, 1992). Others suggest that privacy is best understood as a cluster concept covering interests in i) control over information about oneself, ii) control over access to oneself, both physical and mental, and iii) control over one's ability to make important decisions about family and lifestyle in order to be self expressive and to develop varied relationships (DeCew, 1997). These three interests are related because in each of the three contexts threats of information leaks, threats of control over our bodies, and threats to our power to make our own choices about our lifestyles and activities all make us vulnerable and fearful that we are being scrutinized, pressured or taken advantage of by others. Privacy has moral value because it shields us in all three contexts by providing certain freedom and independence - freedom from scrutiny,

prejudice, pressure to conform, exploitation, and the judgment of others.

3.7 Is Privacy Relative?

Schoeman (1984) points out that the question of whether or not privacy is culturally relative can be interpreted in two ways. One question is whether privacy is deemed valuable to all peoples or whether its value is relative to cultural differences. A second question is whether or not there are any aspects of life that are inherently private and not just conventionally so. Most writers have come to agree that while almost all cultures appear to value privacy, cultures differ in their ways of seeking and obtaining privacy, and probably do differ in the level they value privacy (Westin, 1967; Rachels, 1975). Allen (1988) is especially sensitive to the ways obligations from different cultures affect perceptions of privacy. There has been far less agreement on the second question. Some argue that matters relating to one's innermost self are inherently private, but characterizing this realm more succinctly and less vaguely has remained an elusive task. Thus it may well be that one of the difficulties in defining the realm of the private is that privacy is a notion that is strongly culturally relative, contingent on such factors as economics as well as technology available in a given cultural domain.

4. Privacy and Technology

The earliest arguments by Warren and Brandeis for explicit recognition of privacy protection in law were in large part motivated by expanding communication technology such as the development of widely distributed newspapers and multiply printed reproductions of photographs. Similarly Fourth Amendment protection against search and seizure was extended later in the twentieth century to cover telephone wiretaps and electronic surveillance. It is clear that many people still view privacy is a valuable interest and realize it is now threatened more than ever by technological advances. There are massive databases and Internet records of information about individual financial and credit history, medical records, purchases and telephone calls, for example, and most people do not know what information is stored about them or who has access to it. The ability for others to access and link the databases, with few controls on how they use, share, or exploit the information, makes individual control over information about oneself more difficult than ever before.

There are numerous other cases of the clash between privacy and technology. Consider the following new technologies. Caller ID, originally designed to protect people from unwanted calls from harassers, telemarketers, etc., involves privacy concerns for both the caller and the called. There is widespread mandatory and random drug testing of employees and others, although the Supreme Court has disallowed mandatory drug tests on pregnant women for use by police. Officials can now use heat sensors aimed through walls to detect such things as growing marijuana. Surveillance photos are commonly taken of those using Fast Lane, resulting in tickets mailed to speeding offenders, and similar photos are now taken at red lights in San Diego and elsewhere, leading to surprise tickets. Face scanning in Tampa, at casinos, and at large sporting events such as the Super Bowl, matches those photos with database records of felons, resulting in the capture of multiple offenders on the loose. Some rental car drivers are now

tracked by Global Positioning System (GPS) satellites, enabling car rental companies, not police, to levy stiff fines for speeding. Immigration officials in Australia are considering proposals to tag asylum seekers with electronic trackers before sending them into the community to await hearings. The media has recently uncovered an FBI Web surveillance system called Carnivore, that appears to sample the communications of as many Internet users as it chooses, not just suspects. Echelon, a covert global satellite network said to have the ability to intercept all phone, fax, and e-mail messages in the world, may have up to 20 international listening posts. Airline passengers will soon be able to go through customs with a two second biometric scan that confirms identity by mapping the iris of the eye, and U.S. airlines are considering using “smart cards” which will identify passengers by their fingerprints. There is a proliferation of biometric identification using faces, eyes, fingerprints, and other body parts for identifying specific individuals, and the technology for matching the information with other databases is advancing quickly. For many of these cases, it is possible to make a compelling argument for overriding the privacy intrusions. Drug and alcohol tests for airline pilots on the job seem completely justifiable in the name of public safety, for example. With the development of new and more sophisticated technology, however, recent work on privacy is examining the ways in which respect for privacy can be balanced with justifiable uses of new technology (Agre and Rotenberg, 1997; Brin, 1998; Etzioni, 2000). Moreover, in the wake of the terrorist attacks on September 11, 2001, it is likely that the literature on privacy will increasingly focus on how to balance privacy concerns with the need for public safety in an age of terrorism.

Bibliography

- Agre, P. and Rotenberg, M., (eds.), 1997, *Technology and Privacy: The New Landscape*, Cambridge: MIT Press
- Allen, A., 1988, *Uneasy Access: Privacy for Women in a Free Society*, Totowa, N.J.: Rowman and Littlefield
- Bloustein, E., 1964, ‘Privacy as an Aspect of Human Dignity: An Answer to Dean Prosser’, *New York University Law Review* 39:962-1007
- Bok, S., 1982, Secrets: On the Ethics of Concealment and Revelation, New York: Pantheon
- Bork, R., 1990, *The Tempting of America: The Political Seduction of the Law,* New York: Simon and Schuster
- Brin, David, 1998, *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?*, Reading, MA: Addison-Wesley
- DeCew, J., 1997, *In Pursuit of Privacy: Law, Ethics, and the Rise of Technology*, Ithaca: Cornell University Press
- Elshtain, J., 1981, *Public Man, Private Woman: Women in Social and Political Thought*, Princeton: Princeton University Press
- -----, 1995, *Democracy on Trial*, New York, Basic Books
- Etzioni, A., 2000, *The Limits of Privacy*, New York: Basic Books
- Fried, C., 1970, *An Anatomy of Values*, Cambridge: Harvard University Press
- Gavison, R., 1980, ‘Privacy and the Limits of Law’, *Yale Law Journal* 89: 421-71
- Gerety, T., 1977, ‘Redefining Privacy’, *Harvard Civil Rights-Civil Liberties Law Review* 12: 233-

96
- Gerstein, R., 1978, 'Intimacy and Privacy', *Ethics* 89: 76-81
- Henkin, L., 1974, 'Privacy and Autonomy', *Columbia Law Review* 74:1410-33
- Inness, J., 1992, *Privacy, Intimacy and Isolation*, Oxford: Oxford University Press
- Johnson, J., 1994, 'Constitutional Privacy', *Law and Philosophy* 13: 161-193
- Kupfer, J., 1987, 'Privacy, Autonomy and Self-Concept', *American Philosophical Quarterly* 24: 81-89
- MacKinnon, C., 1989, *Toward a Feminist Theory of the State*, Cambridge: Harvard University Press
- Mead, M., 1949, *Coming of Age in Samoa*, New York: New American Library
- Parent, W., 1983, 'Privacy, Morality and the Law', *Philosophy and Public Affairs* 12: 269-88
- Paul, J., Miller, F., and Paul, E., (eds.), 2000, *The Right of Privacy*, Cambridge: Cambridge University Press
- Pennock, J. and Chapman, J., (eds.), 1971, *Privacy*, NOMOS XIII, New York: Atherton Press
- Posner, R., 1981, *The Economics of Justice*, Cambridge: Harvard University Press
- Prosser, W., 1955, *Handbook of the Law of Torts*, 2nd ed., St. Paul: West
- Rachels, J., 1975, 'Why Privacy is Important', *Philosophy and Public Affairs* 4: 323-33
- Scanlon, T., 1975, 'Thomson on Privacy', *Philosophy and Public Affairs* 4: 315-322
- Schoeman, F., (ed.), 1984, *Philosophical Dimensions of Privacy: An Anthology*, Cambridge: Cambridge University Press
- -----, 1992, *Privacy and Social Freedom*, Cambridge: Cambridge University Press
- Thomson, J., 1975, 'The Right to Privacy', *Philosophy and Public Affairs* 4: 295-314
- Turkington, R., Trubow, G., and Allen, A., (eds.), 1992, *Privacy: Cases and Materials*, Texas: John Marshall
- Westin, A., 1967, *Privacy and Freedom*, New York: Atheneum
- Warren, S. and Brandeis, L., 1890, 'The Right to Privacy,' *Harvard Law Review* 4: 193-220.

Other Internet Resources

- Machan, Tibor, 'The Right to Private Property', in *The Internet Encyclopedia of Philosophy*, J. Fieser (University of Tennessee/Martin), editor.
- Online Guide to Privacy Resources, (Electronic Privacy Information Center, Marc Rotenberg, ed.)

Related Entries

autonomy: in moral and political philosophy | feminism, interventions: feminist philosophy of law | legal philosophy | legal rights | liberty (positive and negative) | rights | rights: human | tort law

Judith DeCew
jdecew@clarku.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 14, 2002
Content last modified: May 14, 2002

Legal Rights

Legal rights are, clearly, rights which exist under the rules of legal systems. They raise a number of different philosophical issues. (1) Whether legal rights are conceptually related to other types of rights, principally moral rights; (2) What the analysis of the concept of a legal right is; (3) What kinds of entities can be legal right-holders; (4) Whether there any kinds of rights which are exclusive to, or at least have much greater importance in, legal systems, as opposed to morality; (5) What rights legal systems ought to create or recognise. Issue (5) is primarily one of moral and political philosophy, and is not different in general principle from the issue of what duties, permissions, powers, etc, legal systems ought to create or recognise. It will not, therefore, be addressed here.

A preliminary point should be mentioned. Do all legal systems have a concept of rights? Their use is pervasive in modern legal systems. We talk of legislatures having the legal right to pass laws, of judges to decide cases, of private individuals to make wills and contracts; as well as of constitutions providing legal rights to the citizens against fellow citizens and against the state itself. Yet it has been suggested that even some sophisticated earlier systems, such as Roman law, had no terminology which clearly separated rights from duties (see *Maine* (1861), 269-70). The question is primarily one for legal historians and will not be pursued here, but it may be remarked that it may still be legitimate when describing those systems to talk of rights in the modern sense, since Roman law, for example, clearly achieved many of the same results as contemporary systems. Presumably, it did so by deploying some of the more basic concepts into which rights can, arguably, be analysed.

---

1. Are Legal Rights Conceptually Related to Other Types of Rights?

The position of many important writers on legal rights is difficult to ascertain on this point, because it is not one they addressed directly. Hohfeld (1919), for example, confined his discussion entirely to legal rights and never mentioned moral ones. Hart did write about moral rights (1955, 1979) as well legal ones (1973, 1994), but not in a way that allows for much direct comparison. Bentham (1970 [1782]) wrote extensively about the analysis of legal rights, but, notoriously, thought that the idea of natural moral rights was conceptual nonsense.

Mill (1969 [1861]), whilst endorsing Bentham's overall Utilitarian position, did not share his scepticism about moral rights, and seems to have thought that moral and and legal rights were, analytically, closely connected - "When we call anything a person's right, we mean that he has a valid claim on society to protect him in the possession of it, either by the force of law, or by that of education and opinion." Those things which ought to be so protected were, in his view, those which concerned the fundamentals of human well-being, and were therefore a sub-set of those things which a person ought to have on grounds of utility.

Whilst not necessarily sharing Mill's view about all rights being related to fundamentals of well-being, many contemporary writers (e.g., Raz (1984a, 1984b), Wellman (1985, 1995)) agree that the core concept of a right is something common to law and morality, though some have argued that jurisprudential writers, particularly Hohfeld, provide a better and clearer starting-point for general analysis than previous writers in moral philosophy. The view that the core concept is common to both would appear to be consistent with maintaining that, nevertheless, in terms of justification in practical reasoning, legal rights should be based on moral ones.

2. The Conceptual Analysis of Legal Rights

Not all philosophers have agreed that rights can be fully analysed. White (1984), for example, argued that the task is impossible because the concept of a right is as basic as any of the others, such as duty, liberty, power, etc (or any set of them) into which it is usually analysed. He agrees, however, that rights can in part be explained by reference to such concepts. White's approach, based largely on close linguistic analysis, has remained something of a minority one.

The remaining approaches can be categorised in different ways, but a main division is between those who think that rights are singled out by their great weight as practical reasons, and those who think that rights are not special in this regard, but instead are to be analysed into duties, permissions, powers, etc, or some combination of these, perhaps with the addition of other conditions.

Dworkin (1973, 1975, 1981, 1986) has been the principal proponent of the first view. According to him rights enjoy a categorial priority in weight over any other consideration which is not itself right-based. Clearly, it is true of many legal systems that constitutional rights, or some of them, should outweigh any other consideration which is not itself derived from a constitutional right. But that seems to be primarily because of the constitutional status of the right. Both in law and in morality many rights are of a rather

trivial nature. In morality such rights can, arguably, sometimes be justifiably outweighed even by considerations of personal convenience (cf. Raz (1978)). Similarly in law it seems that many *prima facie* rights can be defeated by what the court regards as considerations of the general interest. Dworkin's (1977) response to the latter type of criticism has been to argue that, on closer inspection, the consideration opposing the right can be seen as itself an instantiation of another general right. But this depends on the contentious claim that the only considerations that courts can justifiably rely upon are pre-existing rights. The objection has also been raised that, as a general theory of the nature of rights, it risks being self-defeating, since any consideration whatsoever can then be argued to be right-based, which leaves rights with no special role in practical reasoning.

Most writers have, instead, favoured the view that rights are to be analysed into other, more basic, notions, principally those of duty, permission and power, with perhaps the addition of other criteria. This means that not all rights will be of great importance. Their importance will vary with the strength of the grounds for the duty, permission or power. Before looking more closely at these accounts, another point should be mentioned. Theorists are divided between those who think that rights are, as it were, the 'reflex' of the duty, permission or power, and those who think that the right has a priority over them. The question is whether the duty, etc, grounds the right, or the right the duty. Most older writers (e.g., Bentham, Austin, Hohfeld, Kelsen) appear to have adhered to the first view, whilst more recent writers (e.g., MacCormick, Raz, Wellman) take the second. The second view has the implication that the force of a right is not necessarily exhausted by any existing set of duties etc, that follow from it, but may be a ground for creating new duties as circumstances change. This latter view seems to accord better at least with the way that constitutional legal rights work.

Amongst those who think that rights can be analysed, at least in part, into duties, permissions and powers, there is a further main division. Some think that the essence of a right is to have choice or control over the corresponding duty etc. Others think that the main thing is that one's interests are protected by the duty etc. Hart and Wellman are amongst the proponents of the first view, Bentham, Austin, MacCormick and Raz are amongst those maintaining some version of the second.

An outline of Hart's (1973) theory may be given as an illustration of the first view. According to Hart, someone (call him '*X*') may be a legal right-holder primarily in one of two ways. First of all, *X* may have a bilateral permission to perform some action, i.e., *X* is permitted both to *A* and to not-*A* (together with there being some prohibitions on others interfering). Secondly, someone else may have a duty (e.g., to pay *X* £10) over which *X* has control, primarily by waiving or enforcing it. Since *X* has a choice in each case that explains why he is referred to as being a right-holder. One difficulty about this kind of theory is to explain our apparent reference to rights when there is no choice, eg when one is not only entitled to vote in elections, but also obliged by law to do so.

Two different versions of the interest theory can be seen, corresponding to the question about the priority of rights mentioned above.

According to older versions, such as those of Bentham and Austin, *X* is a right-holder because he is the beneficiary, or intended beneficiary, of another's duty, or perhaps of the absence of a duty on him which

the law might otherwise have imposed. For example, if *X* has a right to be paid £10 by *Y*, then this is explained by saying that *Y* has a duty, the performance of which (handing over the £10) is intended to benefit X. One problem about this theory is to explain why the criminal law, although it may in part exist to protect moral rights, is not generally regarded as directly conferring legal rights on individual citizens, despite the fact that they are intended beneficiaries of the corresponding duties. (There may, of course, in many systems be parallel civil law rights, but that is a contingent matter. See more on this point below.)

A more modern version of this theory was proposed by MacCormick (1977), who argued that a right-holder was the intended beneficiary of a specific share of benefit, rather than just being a generalised beneficiary of the rules. However, even with this amendment, it remains difficult to explain third party rights under contracts. Suppose *X* and *Y* enter into a contract which imposes duties on each of them with the intention that performance of these will benefit *Z*. According to the theory, *Z* must (conceptually) be a legal right-holder. But it is in fact an entirely contingent matter as to whether *Z* is or not. Some legal systems recognise *Z* as having rights in such a situation and others do not. In Britain, for example, Scots Law recognised such rights under certain conditions, but English Law did not until the English position was recently changed by statute.

More recent versions, such as those of Raz (1984a, 1984b), take a different tack altogether. According to them, to say that *X* is a right-holder is to say that his interests, or an aspect of them, are sufficient reason for imposing duties on others either not to interfere with *X* in the performance of some action, or to secure him in something. This, *inter alia*, gets round the third-party rights' problem, because the explanation is simply that it is all a question of whether the system recognises the interests of *Z* as part of the reason for *X* and Y's duties, or whether it is only the interests of *X* and *Y*. Raz (1997) has emphasised that this does mean that only the right-holder's interests are relevant to the question of whether something should be recognised as a right. Considerations of the general or common interest may be relevant too.

A number of subsidiary questions can be raised.

Firstly, should rights be analysed solely in terms of duties on others (together with some other condition), or do we need to bring in also other concepts, such as permission, power and immunity? Hohfeld thought that, strictly speaking, something was a legal right only if it corresponded to a duty on another, but he argued that legal usage was often confusing because the reference was really to one of the other concepts. Thus, in his view, the law sometimes also said that *X* had a right if (1) he had a permission to *A*, (2) he had a legal power to *A*, (3) *Y* had no legal power to affect him.

Waldron (1981) and Raz (1984a, 1984b) have been exponents of the view that rights should be seen as giving rise only to duties. Hart (1973), following Bentham, had argued that a liberty-right should be seen as a bilateral permission to A together with duties on others not to interfere with *X*'s A-ing. Waldron and Raz argue that it is an important feature of rights that they entitle the right-holder to do not only that which is right, but also (within bounds) that which is wrong. This they regard as best explained by seeing rights as imposing only duties of non-interference on others, not as granting the right-holder a permission. An alternative view (Campbell 1997) is to see some rights as indeed granting permissions,

but to point out that in granting a legal permission the law is not saying that there may not be reasons against performing the action, only that (within the bounds of the permission) the law will act as if there were not.

Powers raise a different issue. Many writers (e.g., Hohfeld, Hart (1973)) have considered them as being a type of right. By a legal power we mean the ability to bring about changes in legal rules or their application (plus some further conditions). Usually, of course, the lawmaker in granting a power also grants a right to exercise it, but occasionally this is not so, for example where the exercising of the right would itself be a crime or a civil wrong. In English Law, for example, until the position was recently changed by statute, a thief had, in certain special circumstances, the legal power to pass good title in the goods he had stolen to a third party, even though by doing so he committed a civil, and possibly also a criminal, wrong. This seems to indicate that powers should not be thought of as being rights themselves.

Powers also illustrate a general problem about the analysis of legal rights, and arguably of rights in general. Namely that of whether an element should be seen as part of the very essence of the concept of a right, or whether it is merely an element in that which is (contingently) its content, i.e., that which there is a right to do or have.

Relatedly, of the four fundamental types of rights which Hohfeld claimed to identify, immunities raise problems, though somewhat different ones. An immunity arises when *Y* has no power to change X's legal position. But is an immunity itself a right or is it simply a means of protecting a right, i.e., by making it immune from removal or alteration? As with powers, views have differed about this.

3. What Kinds of Entities Can be Legal Right-holders?

There has been much dispute amongst philosophers as what to kinds of entities can be right-holders. Corresponding pretty much to the general dispute about the very nature of rights, some have argued that any entity which would benefit from the performance by others of legal duties can be a right-holder; others that it has to be an entity which has interests; others that it has to be a entity capable of exercising some kind of control over the relevant legal machinery. And there are variants of all these positions.

There has to be a sense in which legal systems can confer rights on such entities as they please. This is because it has long been recognised that legal systems can regard as legal persons such entities as they please. In England, for example, 'the Crown' has, for centuries, been regarded as a legal entity, although what this means in terms of office-holders, far less the actual human beings who occupied those offices, has changed greatly over that time. Likewise, all modern societies recognise the legal existence as persons of companies or corporations and frequently of such entities as trade unions, government departments, universities, certain types of partnerships and clubs, etc.

One of the most contentious areas in recent years has been whether young children, the severely mentally

ill, non-human animals, areas of endangered countryside, etc, can properly be regarded as being legal right-holders. Clearly anyone who has *locus standi* before a court must be a holder of some rights within the system. But it does not seem to follow automatically that an entity which does not, or which is physically or mentally incapable of bringing a legal action, is not thereby a right-holder. For it may be the intention of the system that the interests of that entity should be represented by another person. Given then, that all these entities may be protected by law, and that someone can bring some kind of legal action to ensure that those duties are enforced, when would we say that the entity itself is a right-holder and when not?

The answer will often turn upon whether one embraces an interest- or a choice-theory of rights. MacCormick (1976), for example, argued that any theory of rights which could not accommodate childrens' rights must be deficient, and this was a reason, in his view, for adopting an interest theory. Wellman (1995), on the other hand, claims that to assert that very young children or the severely mentally ill can have legal rights is to distort the concept of a right, since they lack the relevant control of the legal machinery. Instead, he argues, the relevant rights should be seen as belonging only to those who can bring the relevant actions on their behalf. For example, in his view a very young child would not have a right not to be negligently injured by the conduct of another. Rather, it would be the case that the child's parent had a right that their child not be negligently injured. One difficulty about this position appears to be that it does not easily square with the relevent remedial rights (e.g., to damages) that the law would recognise. In this example the law would clearly compensate the child's loss in being injured, not the parent's loss in their child being injured (though the latter might be a separate ground of action in some systems).

4. Exclusivity of Rights

The issue here is: whether there are any fundamental aspects of rights which are exclusive to, or at least more important in, legal systems, as opposed to morality.

Five particular sub-issues may be raised here.

4.1 Primary and Remedial Rights

Remedial rights are those which arise because of a breach of a primary one. Clearly they arise also outside the law, for example by the duty to apologise or make amends even if there is no legal obligation to do so. But legal remedial duties are generally more precise, and, just by the nature of law, institutionalised.

It is one of the main functions of legal systems to provide remedies for breach (or sometimes anticipated breach) of the primary rights which they confer. So if someone is injured by the negligence of another there will usually arise a remedial right to damages. If he is killed there may arise in members of his family an independent right to compensation, and so on. Other types of remedial right can include those

for court orders requiring the party at fault to execute, or refrain from, some particular course of action, very often that which they had a duty to do, or to refrain from, under the primary right. Such rights are often very complex in the detail. For example the measure of damages may be different if the wrongful act is a tort/delict, as opposed to a breach of contract. Likewise, in many systems, some remedies must be granted as a matter of right whilst others are at the discretion of the court. By way of illustration of the remedies in the two British legal systems, reference may be made to Lawson (1980) and Walker (1974).

Usually remedial rights will themselves have further remedial rights attached, for example, to have the court impose a more coercive order, perhaps with the threat of a criminal or quasi-criminal sanction, or to have a person's assets frozen or confiscated, in the event, for example, that someone has failed to pay damages previously awarded by the court. The details of these further remedial rights vary from system to system.

A related, more controversial, point is as to whether criminal, as opposed to civil, law confers any legal rights on the citizens protected by it. The orthodox view is that it does not, although there may well be a parallel civil right. Take the case of someone who is wrongfully assaulted. In most legal systems this will be both a crime and a tort/delict. The civil law clearly gives a remedial right, eg. to sue for damages. But since, in most jurisdictions, it is mainly (and sometimes exclusively) the state which decides whether to prosecute for the criminal aspect, the more usual view is that the citizen has no legal right corresponding to the criminal aspect.

The issue is often complicated, legally, by the absence of clear indication from the legislature as whether it intended, by a particular statute, to create only a crime or also to confer civil law rights on citizens. A further complication can be that criminal courts sometimes exercise a quasi-civil function (e.g., to make a restoration or compensation order after a conviction for theft), and *vice versa* (e.g., the power of a civil court to award punitive or exemplary damages).

This issue is different from that of whether criminal law can act to recognise and protect moral rights. It seems possible to suggest that it can, since moral rights can be protected not only by legal rights, but also by legal duties on others (without corresponding legal rights). For example, a legal system could create a criminal offence of harrassment in order to protect a moral right to privacy, without thereby necessarily recognising a legal right to privacy, i.e., something which would act as a positive reason in favour of privacy in interpreting unclear rules, or in developing the law.

4.2 Conditional Rights

In the case of many legal rights a condition has to be satisfied for their possession or exercise. This, in itself, does not make legal rights different from many moral ones. Just as one has a legal right to damages for assault only if one has been assaulted, one has a moral one to an apology for being insulted only if one has been insulted. But legal rights can give rise to more complicated situations, which rarely arise in morality.

In the above examples we can say that the right-token, as opposed to the right-type, comes into existence only when the condition for its instantiation is triggered. But legal systems sometimes say that that the right-token exists before one of the conditions for the exercise of the right exists. Essentially, it is the difference between saying "if *p*, *X* has a right to *A*" and "*X* has a right, if *p*, to *A*." In the latter case the implication is that the right-token exists now, not just that it will exist. Why should we say this? One proposed answer is that legal systems, unlike morality, have devised sets of rules for transmission of rights even before the triggering condition for the exercise of the right has arrived.

Suppose, for example, that *X*, under his will, left a sum of money to *Y*, on condition that *Y* had attained the age of 21. It may be that the correct way of understanding the provision, under the rules of the legal system, is that only if *Y* had attained 21 when *X* died does he have a right to the money. But it may be that the correct way of understanding it as saying that *Y*, even if he has not attained 21 when *X* dies, acquires a right to the money, but it is to be paid only when he is 21. One practical difference is that in the latter case the right can pass to Y's successor in title if *Y*, having survived *X*, nevertheless dies before he is 21. In the latter case, lawyers describe the right as 'vested.' There can be many complex legal rules relating to this type of situation, and they vary greatly from jurisdiction to jurisdiction. Reference should be made to textbooks, primarily on testamentary succession, in the jurisdiction.

4.3 Property Rights

A further particular kind of legal rights, or group of rights, which has received an increasing amount of attention from theorists is that of property rights. Discussion of this belongs more properly to that of property itself -- see the entry on property. Only some very brief points will be made here.

The first is as to whether property rights, and hence the concept of property, are essentially legal in their nature, or whether they are more general social phenomena which are simply recognised and protected by law in all modern societies. According to Bentham (1843) " … there is no natural property … property is entirely the creature of the law." Bentham's argument is essentially that what we mean by property is security of expectation in being able to keep, sell, use, etc, objects, and only the law can guarantee such security.

On the other hand, it is certainly possible to talk coherently about property in a way that does not necessarily correspond to the legal position. A parent may for example say to a young child that a certain toy is theirs, though in law it is the parent's. Likewise it may be plausible to claim that concepts of ownership and possession, though they may be less securely protected, can exist in societies which do not have anything that we normally recognise as a fully-fledged legal system. Some people will perhaps regard these kind of examples as indications that the concept of property is not essentially legal, whilst others may incline to the view that these are simply metaphorical extensions of a concept which is legal *au fond*.

Secondly, it should be noted that, in law, property rights can be of many different types. Although ownership is obviously one of the most important, another major class is that of possession, whether

temporary or relatively permanent. For example, the right to use a car which one has hired for a week or to live in a certain house for the rest of one's life. Yet other types, falling short of either ownership or possession, could be, for example, to walk across the local farmer's field or have one's next-door neighbour maintain his side of the joint garden wall.

The details of property rights vary from jurisdiction to jurisdiction perhaps more than those of almost any other types of right. Further, many jurisdictions have different rules relating to property rights in land (and its fixtures) as opposed to all other types of entity. For these details reference should be made to specialist books in the jursdiction.

Even when considering just ownership, there is debate amongst theorists as to how this should be analysed. Some see it as essentially a cluster of other property rights of particular content, such as those to possession, income, etc, whilst others see it as being basically a structural relation between rights, content being comparatively irrelevant. For example as being the person to whom possession or use, even though those may presently belong to others, would ultimately return if a certain series of contingent events were to occur.

For further discussion of property in a philosophical context see Honore (1960, 1961), Becker (1977), Waldron (1988), Munzer (1990), Campbell (1992), Harris (1996), Penner (1997). (Some of these are concerned more with the moral justification of ownership.)

4.4 Subjective Rights

The above account of rights has been written largely from the point of view of Anglo-American law and philosophy. It should, however, be mentioned that there is one aspect of legal rights which is to be found amongst the European Continental writers, but of which there is no trace in the Anglo-American tradition. That is the description of rights as being 'subjective' (*droits subjectifs*; *subjektives Rechten*).

In French and in German the same word (*droit*, *Recht*) serves as the noun which refers both to rules of law and the rights which are created by them, and therefore disambiguation is required.

In French law the distinction is drawn by distinguishing between *le Droit objectif* (the noun spelt with a capital according to some, but not all, writers) and *les droits subjectifs*. (For general discussion see, for example, Cornu 1996). However, French law seems at the same time to confine the term '*droits subjectifs*' to a sub-class of legal rights, namely rights which are primarily those of private citizens, eg to make a will or contract. The term appears not to extend to such rights as those of a government agency owning property or a government minister making a legal order under delegated powers.

German law seems to draw a basically similar distinction between '*das Recht*' and '*subjectives Rechten*' (see, for example, Dietel 1983).

4.5 Means of Conferring Legal Rights

Many of the issues relating to this are not confined to rights, but are shared with duties and powers, so only a brief outline will be given.

In most modern legal systems certain fundamental rights are conferred by the constitution. This usually gives them a certain degree of priority over competing legal considerations, but this can vary from system to system. Sometimes constitutional rights will have an absolute priority over any other consideration not itself based on a constitutional right. Sometimes they will merely favour one legal outcome rather than another, without dictating it.

Constitutions will vary, too, as to whether certain rights are ‘entrenched’ or not. Entrenchment can be absolute, in which case the rights cannot be removed or altered by any constitutional means (as is the case with some of the ‘basic rights’ in the German Constitution), or it can be relative, requiring only a more onerous procedure than that for normal legislation (as with the Constitution of the USA.).

Constitutions will also vary on the extent to which human rights recognised under international law or treaty are recognised in national law. For example, in some countries in Europe, the European Convention on Human Rights, and decisions of the European Court of Human Rights thereon, are incorporated into national law and override any national law inconsistent with them. In others, such as the United Kingdom, the courts have, so far as possible, to interpret legislation to be consistent with the Convention, but have no power to strike it down if they find it to be clearly inconsistent.

Other rights can be conferred by normal legislation or by common law (ie. the tradition of judge-made law). One interesting point is that, arguably, many legal rights are conferred by no positive law, but arise simply from the absence of any law to the contrary. That is, it is probably a practical necessity that every legal system has an unwritten ‘closure rule’ to the effect that whatever is not prohibited is permitted. If some types of rights are essentially permissions, then many such rights arise in this way. In most legal systems my right to cross the street, for example, is of this nature. Probably no positive law will say that I can do so, and possibly no more general enacted right will imply it.

Bibliography

- Austin, John, (1885). *Lectures on Jurisprudence, or the Philosophy of Positive Law*, 5th edn, ed R Campbell, 2 vols, London: John Murray.
- Becker, Lawrence C (1977). *Property Rights: Philosophic Foundations*, London: Routledge & Kegan Paul.
- Bentham, Jeremy (1970 [1782]). *Of Laws in General*, ed Hart, HLA, London: Athlone Press. (Many of Bentham’s other numerous, but scattered, discussions of rights are referred to in Hart (1973).)
- Bentham, Jeremy (1843). *Principles of the Civil Code*, in Bowring, John, ed, The Works of Jeremy Bentham, Vol 1, Edinburgh: William Tait

- Campbell, Kenneth (1992). "On the General Nature of Property Rights", (1992) 3 King's College Law Journal 79.
- Campbell, Kenneth (1997). "The Variety of Rights", in Martin, R & Sprenger, G, eds, *Challenges to Law at the End of the 20th Century*: *Rights*, Stuttgart: Franz Steiner Verlag, 22.
- Cornu, Gerard, ed, (1996) *Vocabulaire Juridique*, 6th edn, Paris: Presses Universitaires de France.
- Dietel, Clara-Erika, (1983) *Dictionary of Legal, Commercial and Political Terms* [English-German/German-English], New York, Matthew Bender & Co Inc.
- Dworkin, Ronald M (1973). "Taking Rights Seriously", in Simpson, AWB, ed, *Oxford Essays in Jurisprudence, Second Series*, Oxford: Clarendon Press, 202; reprinted in his *Taking Rights Seriously*, revd edn, London: Duckworth, 1978, 184.
- Dworkin, Ronald M (1975). "Hard Cases" (1975) 88 Harvard Law Review 1057; reprinted in his *Taking Rights Seriously*, *supra*, 81.
- Dworkin, Ronald M (1977). "Seven Critics", (1977) 11 Georgia Law Review 1201.
- Dworkin, Ronald M (1981). "Is there a Right to Pornography?", (1981) 1 Oxford Journal of Legal Studies 177; reprinted in his *A Matter of Principle*, Oxford: Clarendon Press, 1985, 336.
- Dworkin, Ronald M (1986). *Law's Empire*, London: Fontana.
- Halpin, Andrew (1997). *Rights and Law: Analysis and Theory*, Oxford: Hart Publishing.
- Harris, JW (1996), *Property and Justice*, Oxford: Clarendon Press.
- Hart, HLA (1955). "Are There any Natural Rights?", (1955) 64 Philosophical Review 175.
- Hart, HLA (1973). "Bentham on Legal Rights", in Simpson, AWB, ed, *Oxford Essays in Jurisprudence, Second Series*, Oxford: Clarendon Press, 171; reprinted in his *Essays on Bentham: Jurisprudence and Political Theory*, Oxford: Clarendon Press, 1982, 162.
- Hart, HLA (1979). "Between Utility and Rights", in Ryan, A, ed, *The Idea of Freedom: Essays in Honour of Isaiah Berlin*, Oxford: Clarendon Press, 77; reprinted in his *Essays in Jurisprudence and Philosophy*, Oxford, Clarendon Press, 1983, 198
- Hart, HLA (1994). *The Concept of Law*, 2nd edn, with posthumous postscript, ed Bulloch, P & Raz, J, Oxford: Clarendon Press.
- Hohfeld, Wesley Newcombe (1919). *Fundamental Legal Conceptions as Applied in Judicial Reasoning*, ed Cooke, WW, New Haven: Yale University Press.
- Honore, Anthony M (1960). "Rights of Exclusion and Immunities Against Divesting", (1960) 34 Tulane Law Review 453
- Honore, Anthony M (1961). "Ownership", in Guest, AG, ed, *Oxford Essays in Jurisprudence: First Series*, Oxford: Clarendon Press, 107
- Kelsen, Hans (1946). *General Theory of Law and State*, trs Wedberg, A, Cambridge, Mass: Harvard University Press.
- Kramer, Matthew H, Simmonds, NE & Steiner, Hillel (1998). *A Debate Over Rights: Philosophical Enquiries*, New York: Oxford University Press, 1998.
- Lawson, FH (1980). *Remedies of English Law*, 2nd edn, London: Butterworths.
- MacCormick, Neil (1976). "Children's Rights: A Test-Case for Theories of Rights",
- (1976) 32 Archiv fur Rechts- und Sozialphilosophie, 305; reprinted in his *Legal Right and Social Democracy: Essays in Legal and Political Philosophy*, Oxford:
- Clarendon Press, 1982, 154.
- MacCormick, Neil (1977). "Rights in Legislation", in Hacker, PMS & Raz, J, eds, *Law, Morality*

and Society: Essays in Honour of HLA Hart, Oxford: Clarendon Press, 189.

- Maine, Henry Sumner (1861). *Ancient Law: Its Connection with the Early History of Society and Its Relation to Modern Ideas*, London: John Murray.
- Martin, Rex (1993). *A System of Rights*, New York: Oxford University Press.
- Mill, John Stuart (1969 [1861]). *Utilitarianism*, in Robson, J, ed, *The Collected Works of John Stuart Mill*, Vol 10, Toronto: Toronto University Press; London: Routledge & Kegan Paul, 203.
- Munzer, Stephen R (1990). *A Theory of Property*, Cambridge: Cambridge University Press.
- Nickel, James W (1987). *Making Sense of Human Rights: Philosophical Reflections on the Universal Declaration of Human Rights*, Berkeley & Los Angeles: University of California Press.
- Penner, JE (1997). *The Idea of Property in Law*, Oxford: Oxford University Press.
- Raz, Joseph (1978). "Professor Dworkin's Theory of Rights", (1978) 26 Political Studies 123.
- Raz, Joseph (1984a). "The Nature of Rights", (1984) 93 Mind 194; reprinted in his *The Morality of Freedom*, Oxford: Clarendon Press, 1986, 165
- Raz, Joseph (1984b). "Legal Rights", (1984) 4 Oxford Journal of Legal Studies 1; reprinted in his *Ethics in the Public Domain: Essays in the Morality of Law and Politics*, Oxford: Clarendon Press, 1994, 238.
- Raz, Joseph (1997). "Rights and Politics", in Tasioulas, J, ed, *Law, Values and Social Practices*, Aldershot: Dartmouth, 75.
- Steiner, Hillel (1994). *An Essay on Rights*, Oxford: Blackwell Publishers.
- Sumner, LW (1987). *The Moral Foundation of Rights*, Oxford: Clarendon Press.
- Thomson, Judith Jarvis (1986). *Rights, Restitution and Risk: Essays in Moral Theory*, ed Parent, W, Cambridge, Mass.:Harvard University Press.
- Waldron, Jeremy (1981). "A Right to do Wrong", (1981) 92 Ethics 21; reprinted in his *Liberal Rights: Collected Papers 1981-1991*, Cambridge: Cambridge University Press, 63.
- Waldron, Jeremy (1988). *The Right to Private Property*, Oxford: Clarendon Press.
- Walker, David M (1974), *The Law of Civil Remedies in Scotland*, Edinburgh: WH Green.
- Wellman, Carl (1985). *A Theory of Rights*, Totowa, NJ: Rowman & Allanheld.
- Wellman, Carl (1995). *Real Rights*, New York: Oxford University Press.
- Wellman, Carl (1999) *The Proliferation of Rights: Moral Progress or Empty Rhetoric?*, Boulder, Colorado: Westview Press.
- White, Alan R (1984). *Rights*, Oxford: Basil Blackwell.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

rights

Kenneth Campbell
kenneth.campbell@kcl.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 20, 2001
Content last modified: December 20, 2001

Philosophy and Christian Theology

Many of the doctrines and concepts central to Christianity have important philosophical implications or presuppositions. In this article we will take a closer look at some of the central doctrines and concepts, and their philosophical relevance.

Of course, many philosophically laden doctrines and concepts are relevant to Christianity, and we cannot discuss them all here. Rather, our focus will be on those concepts and doctrines that are distinctively Christian, and which have been the focus of a good deal of recent discussion in the philosophical literature. Thus, although theism is a central Christian concept, it is not *distinctively* Christian and so will not be covered here. Further, although views about the Eucharist, a central Christian concept, have held a significant place in the philosophical dialogue in former times, it will not be discussed here since it has not been a significant focus of recent discussions. As a result, we will concentrate on three distinctive and central Christian concepts which have received significant attention in the recent literature: the doctrines of the Trinity and the Incarnation, and views on the nature of atonement.

1. Philosophy and Christian Theology

Before we begin, it is worthwhile to consider in brief the general relationship between philosophy and Christian religious dogma. In the history of Christian theology, philosophy has sometimes been seen as a natural complement to theological reflection, while at other times the advocates for the two disciplines have regarded each other as mortal enemies. Some early Christian thinkers such as Tertullian were of the view that any intrusion of secular philosophical reason into theological reflection was out of order. Thus, even if certain theological claims seemed to fly in the face of the standards of reasoning defended by philosophers, the religious believer should not flinch. Other early Christian thinkers, such as St.

Augustine of Hippo, argued that philosophical reflection complemented theology, but only when these philosophical reflections were firmly grounded in a prior intellectual commitment to the underlying truth of the Christian faith. Thus, the legitimacy of philosophy was derived from the legitimacy of the underlying faith commitments.

Into the High Middle Ages, Augustine's views were widely defended. It was during this time however that St. Thomas Aquinas described another model for the relationship between philosophy and theology. According to the Thomistic model, philosophy and theology are distinct enterprises. The primary difference between the two is their intellectual starting points. Philosophy takes as its data the deliverances of our natural mental faculties: what we see, hear, taste, touch, and smell. These data can be accepted on the basis of the reliability of our natural faculties with respect to the natural world. Theology, on the other hand takes as its starting point the divine revelations contained in the Bible. These data can be accepted on the basis of divine authority, in a way analogous to the way in which we accept, for example, the claims made by a physics professor about the basic facts of physics.

On this way of seeing the two disciplines, if at least one of the premises of an argument is derived from revelation, the argument falls in the domain of theology; otherwise it falls into philosophy's domain. Since this way of thinking about philosophy and theology sharply demarcates the disciplines, it is possible in principle that the conclusions reached by one might be contradicted by the other. According to advocates of this model, however, any such conflict must be merely apparent. Since God both created the world which is accessible to philosophy and revealed the texts accessible to theologians, the claims yielded by one cannot conflict with the claims yielded by another unless the philosopher or theologian has made some prior error.

Since the deliverances of the two disciplines must then coincide, philosophy can be put to the service of theology (and perhaps vice-versa). How might philosophy play this complementary role? First, philosophical reasoning might persuade some who do not accept the authority of purported divine revelation of the claims contained in religious texts. Thus, an atheist who is unwilling to accept the authority of religious texts might come to believe that God exists on the basis of purely philosophical arguments. Second, distinctively philosophical techniques might be brought to bear in helping the theologian clear up imprecise or ambiguous theological claims. Thus, theology might provide us with information sufficient to conclude that Jesus Christ was a single person with two natures, one human and one divine, but leave us in the dark about exactly how this relationship between divine and human natures is to be understood. The philosopher can provide some assistance here, since, among other things, he or she can help the theologian discern which models are, for example, logically inconsistent and thus not even candidates for understanding the relationship of divine and human natures in Christ.

For most of the twentieth century, the vast majority of English language philosophy went on without much interaction with theology at all. While there are a number of complex reasons for this divorce, three are especially important. The first is that atheism was the predominant opinion among English language philosophers throughout much of that century.

A second, quite related reason is that, philosophers in the twentieth century regarded theological language as either meaningless, or, at best, subject to scrutiny only insofar as that language had a bearing on religious practice. The former belief (i.e., that theological language was meaningless) was inspired by a tenet of logical positivism, according to which any statement that lacks empirical content is meaningless. Since much theological language, for example, language describing the doctrine of the Trinity, lacks empirical content, such language must be meaningless. The latter belief, inspired by Wittgenstein, holds that language itself only has meaning in specific practical contexts, and thus that religious language was not aiming to express truths about the world which could be subjected to objective philosophical scrutiny.

The third reason is that a great deal of academic theology moved away from defending the claims of orthodox Christian theism in traditional ways, often seeking devices for re-interpreting these claims in ways congenial to contemporary modes of thought which often ran contrary to the methods employed in analytic philosophy.

In the last twenty years, however, philosophers have returned to many of the traditional claims of orthodox Christianity and have begun to apply the tools of contemporary philosophy in ways that are somewhat more eclectic than those described in the Augustinian or Thomistic models described above. In keeping with the recent academic trend, contemporary philosophers of religion have been unwilling to maintain hard and fast distinctions between the two disciplines. As a result, it is often difficult in reading recent work to distinguish what the philosophers are doing from what the theologians of past centuries regarded as strictly within the theological domain. However, like theologians of the medieval period, much recent work in philosophy of religion seems to fall into one of two categories. The first category includes attempts to demonstrate the truth of religious claims by appeal to evidence available apart from purported divine revelations. The second category includes attempts to demonstrate the consistency and plausibility of theological claims using philosophical techniques. In what follows, we will be considering work that falls into this second category.

2. Trinity

From the beginning, Christians have affirmed the claim that there is one God and that three persons are God: God the Father, God the Son, and God the Holy Spirit. In AD 675, the Council of Toledo framed this pair of claims as follows:

> Although we profess three persons we do not profess three substances but *one substance* and three persons … If we are asked about the individual Person, we must answer that he is God. Therefore, we may say God the Father, God the Son, and God the Holy Spirit; but they are not three Gods, he is one God … Each single Person is wholly God in himself and … all three persons together are one God.

Such formulations set forth the Christian doctrine of the Trinity. Cornelius Plantinga, Jr., reflecting on the Council of Toledo's profession, remarks that it “possesses great puzzling power” (Plantinga 1989, 22). No doubt this is an understatement. The Christian doctrine is puzzling, and this has led some of Christianity's

critics to advance the claim that it is, in fact, incoherent.

Perhaps the initial puzzling power of the doctrine of the Trinity is not immediately obvious. After all, someone might think that one thing, Fred, can be “many things” all at the same time, for example, a butcher, a baker, and a candlestick maker. So why can't God be Father, Son, and Holy Spirit all at the same time? Likewise, multiple distinct things can all be “one thing” at the same time. Thus, each member of the Baltimore Orioles baseball team can be Orioles taken individually, as well as “the Orioles” taken collectively. One might then think that defenders of the Trinity might be able to construct models out of such examples that would preserve the logical coherence of the doctrine. But things will not be quite that easy. To see why, we can take a brief detour and then come back to the two examples above.

Traditional Christian theologians have held that however the doctrine of the Trinity is understood, there are two extreme positions that are to be ruled out. These positions are *modalism* and *tritheism*. According to modalism, God is one single entity, object, or substance, and each person of the Trinty is simply a mode or a “way in which the one divine substance manifests itself.” This view has been rejected because it seems to sacrifice the distinctness of the divine persons in order to maintain the notion of divine unity. According to tritheism, on the other hand, the divine persons are each distinct individual persons which are so closely related that they together count as a single thing in some fashion. Nonetheless, despite this oneness, the three persons are still three gods. This view has been rejected for the opposite reason, namely, it preserves the distinctness of persons without maintaining any robust sense of the “oneness” of God.

One can now see why the “butcher, baker, candlestick maker” and the “Orioles” examples will not help us in providing a model for the Trinity. The first, like modalism, leans too heavily towards oneness at the expense of the distinctness of the three persons. It holds, that is, that there is really only one Fred, but that Fred can manifest himself in different ways by carrying out three different tasks. The second, like tritheism, leans too far in the opposite direction. On this example, the individual Orioles only form the “single team” because of certain agreements they have made to act cooperatively on the baseball team. There is no genuine, organic unity here.

Nonetheless, most models of the Trinity that have been proposed and defended have leaned in modalist or tritheistic directions. In order to help sort out which models can be regarded as plausible, one needs first to get clear about just what the Christian means to affirm in confessing the existence of three persons and one God. What is “a person” according to the doctrine, and what is “a God”? One can easily see some initial difficulties in even these questions. Even if we can come up with a single coherent description of God, we are still left with the ambiguous notion of person. Sometimes we use the word “person” in a metaphysical sense, to refer to an individual, rational substance. Other times we use it in a psychological sense to refer to a “center of consciousness or rational awareness.” In other cases we might have in mind a functional notion of person, according to which a person is whatever sort of thing is capable entering into certain sorts of relationships, such as love, friendship, and so forth. Or we might use “person” in a moral or forensic sense, according to which a person is a subject or moral accountability, praise, or blame. And there are others.

Since Christians claim that the doctrine of Trinity is discovered through divine revelation, perhaps the relevant conception of person should be drawn from revealed texts. Unfortunately, the Bible itself does not seem to narrow down the alternatives to a single candidate. As a result, there is a good deal of remaining latitude in constructing a model for the Trinity.

Recent defenses of orthodox conceptions of the Trinity understand the notion in a way that highlights the centrality of persons as distinct centers of rational, conscious, and morally significant volitional activity. Most have concluded that this conception of personhood is incompatible with regarding the three divine persons as somehow mere aspects or modes of presentation of an underlying singular entity. As a result, these recent defenses have leaned in the direction of regarding the divine persons as distinct entities whose unity arises in virtue of certain necessary relations that exist among them. In this way, these models lean more in the tritheistic direction. Still, the necessary relations that these models attribute to the divine persons unify them in special and unique ways.

Richard Swinburne, for example, defends a view according to which each of the three divine persons has all of the essential characteristics of divinity: omniscience, omnipotence, omnipresence, moral perfection, and so forth (Swinburne 1994, ch. 8). He further claims that the persons have necessarily harmonious wills, so that their volitions never come into conflict, and that there is a perfectly loving relation that also necessarily obtains among them. Further, this view is compatible with traditional claims of dependence relations among members of the Trinity. Traditional formulations of the doctrine hold that the Father *generates* the Son and that Father and Son jointly give rise to (or *spirate*, literally “breathe forth”) the Holy Spirit. Such relations are possible as long as one causes the other in such a way that the causing relation has always obtained, and it is impossible for the relation not to obtain.

On this sort of view, there is one God because the community of divine persons is so closely inter-connected that, though they are three distinct persons, they nonetheless count as a single entity in another respect. For if we were to consider a set of three human persons, for example, who exhibited these characteristics of necessary unity, volitional harmony, and love, it is hard to regard them as distinct in the way we do ordinary persons. And that is, of course, just what the doctrine aims to put forth.

Perhaps this view seems to lean too strongly in the tritheistic direction. How could the social Trinitarian respond to this worry? One way would be to focus attention on exactly what is required in order for many “things” to jointly constitute another single “thing.” My (one) body is composed of (many) atoms. My (one) car is composed of (many) parts. In order to assess whether or not social Trinitarianism is viciously tritheistic, one needs to ask what principles govern the relationship between parts and wholes generally. We know many atoms can make a single body and many ingredients can make a single cake. Can many persons constitute a single divine entity? One thing is sure: the answer is not an obvious “no.” And this, perhaps, leaves the door open for the social Trinitarian to make the case that divine unity is not lost on his view after all. Saving such unity, however, will require more metaphysical work.

3. Incarnation

The doctrine of the Incarnation concerns Jesus Christ, the second person of the Trinty. Specifically, the doctrine holds that, at a time roughly two thousand years in the past, the divine person took on himself a second, fully human nature. As a result, he was a single person in full possession of two distinct natures, one human and one divine. The Council of Chalcedon in 451 put forth the canonical statement of the doctrine as follows:

> We confess one and the same our Lord Jesus Christ … the same perfect in Godhead, the same in perfect manhood, truly God and truly man … acknowledged in two natures without confusion, without change, without division, without separation–the difference of natures being by no means taken away because of the union, but rather the distinctive character of each nature being preserved, and combining into one person and hypostasis—not divided or separated into two persons, but one and the same Son and only begotten God, Word, Lord Jesus Christ.

Critics have held this doctrine to be “impossible, self-contradictory, incoherent, absurd, and unintelligible.” The central difficulty for the doctrine is that it seems to attribute to one person characteristics that are not logically compatible. For example, it seems on the one hand that human beings are necessarily created, finite, not-omnipresent, not-omniscient, not-omnipotent, and so forth. On the other hand, divine beings are essentially the opposite of all those things. Thus, one person could bear both natures, human and divine, only if such a person could be both finite and not-finite, created and uncreated, and so forth. And this is surely impossible.

Two main strategies have been pursued in an attempt to resolve this apparent paradox. The first is the *kenotic* strategy. The kenotic view (from the Greek *kenosis* meaning ’to empty’) finds its motivation in a New Testament passage which claims that Jesus “who, though he was in the form of God, did not count equality with God a thing to be grasped, but emptied himself, taking the form of a servant, being born in the likeness of men. And being found in human form he humbled himself and became obedient unto death…” (Phillipians 2:6-8). According to this view, in becoming incarnate, God the Son voluntarily and temporarily laid aside some of his divine attributes in order to take on a human nature and thus his earthly mission.

The main difficulty with this basic version of the kenotic view is that it entails that a thing can lay aside properties essential for it's being a member of a certain kind and still remain a member of that kind. In other words, it allows that God the Son could (temporarily) be non-omnipotent, non-omniscient, and so forth, and still be God. But if those attributes are *essential* to divinity, that is, essential for something's being counted as God, then this solution is simply mistaken. Some have offered more refined versions of the kenotic theory, arguing that the basic view mischaracterizes the divine attributes. Rather, God's properties should be characterized as: omniscient-unless-incarnate, omnipotent-unless-incarnate, and so forth. Thus, when the powers of omnipotence are laid aside at the incarnation, Jesus can be fully human while retaining *these* divine attributes without contradiction. (Feenstra 1989, 128-152)

The other main strategy, defended recently by Thomas V. Morris, is the “two minds view” (Morris 1986, 63-73, 102-7). This view unfolds in two steps, one defensive, the other constructive. First, Morris claims

that the incoherence charge against the incarnation rests on a mistake. The critic assumes that, for example, humans are essentially non-omniscient. But what are the grounds for this assertion? Unless we think that we have some special direct insight into the essential properties of human nature, our grounds are that all of the human beings we have encountered have that property. But this merely suffices to show that the property is *common* to humans, not that it is essential. As Morris points out, it may be universally true that all human beings, for example, were born within ten miles of the surface of the earth, but this does not mean that this is an *essential* property of human beings. An offspring of human parents born on the international space station would still be human. If this is right, the defender of the incarnation can reject the critic's characterization of human nature, and thereby eliminate the conflict between divine attributes and human nature so characterized.

This merely provides a way to fend off the critic, however, without supplying any positive model for how the incarnation should be understood. In the second step, then, Morris proposes that we think about the incarnation as the realization of one person with two minds: a human mind and a divine mind. If possession of a human mind and body is sufficient for something's being a human, then “merging” the divine mind with a human mind and conjoining both to a human body will yield one person with two natures. During his earthly life, Morris proposes, Jesus Christ had two minds, with consciousness centered in the human mind. This human mind had partial access to the contents of the divine mind, while God the Son's divine mind had full access to the corresponding human mind.

The chief difficulty the view faces is the coherence of holding that a single person can possess two distinct minds. Does this view propose an Incarnate Christ with multiple personality disorder? Morris claims that this objection lacks merit. In fact, contemporary psychology seems to provide resources which support the viability of such a model. As Morris points out elsewhere, the human mind is typically characterized as a system of somewhat autonomous subsystems. The normal human mind, for example, includes the workings of the conscious mind, the seat of awareness, and the unconscious mind. Morris proposes that similar sorts of relations can be supposed to obtain between the divine and human mind of Christ.

4. Atonement

Traditional Christianity holds that sin separates human creatures from God, and that reconciliation can occur in virtue of something that happens through the incarnation, life, death, and resurrection of Jesus Christ. But how are these claims of separation and reconciliation to be understood? The answer to these questions makes up the doctrine of atonement. Throughout the history of Christian theology, a variety of models have been proposed. Most of these models fall into one of four types. *Ransom* theories contend that sin has rendered humans enslaved to the Devil. In order to free his beloved creatures from this enslavement God was required to pay a ransom, and the price was the death of his sinless incarnate Son. *Penal Substitution* models contend that through sin humans have incurred a moral debt which needs to be paid. These views hold that the price to be paid is spiritual death and separation from God. No one man can pay the debt of any other since all men have sinned equally. Thus, God chose to send his incarnate Son, free from original or committed sin, to die on behalf of others, and so satisfy their debt.

Sacrifice models are similar to substitution models, but differ in that they do not think that any moral debt of human creatures can be transferred and satisfied by another. Sacrifice theories acknowledge that wrongdoers incur an obligation to “make things right” with the person wronged. Sometimes this means making restitution. Other times it means undertaking acts of penance which demonstrate the wrongdoer's genuine remorse. Thus, if I, in a fit of anger, throw a brick through the window of your house, I might come to seek forgiveness. In doing so I agree to fix the broken window (restitution) but might also do something more, such as bring you a gift as way of demonstrating my genuine remorse. This latter is the act of penance. However, sometimes restitution and suitable penance cannot be carried out by the wrongdoer himself because restitution or suitable penance is beyond his means. In the case of human sinfulness towards God, this is exactly the case. As a result, God sent Christ to earth, where Christ willingly offered his life as a restitution and penance for the sin of the world. Thus, although human sinful creatures cannot make restitution or penance for their wrongdoing on their own, they can, in their repentance, offer up to God the sacrifice of Christ which was made on their behalf.

Finally, *Moral Exemplar* theories hold that the atonement is secured by moral reform of the sinner. But such moral reform was not fully possible without someone to set the moral example for fallen creatures. Christ became incarnate, on these theories, in order to set this example and thus provide a necessary condition for moral reform and thus restoration of the relationship between creature and Creator.

Ransom theories have no defenders in the recent literature. While each of the remaining theories has defenders, each faces certain key difficulties as well. Substitution theories, for example, require a few central controversial claims. For one, these theories seem to entail that a person can incur an infinite moral debt for a finite amount of earthly wrongdoing. Second, they entail that the moral debt in question cannot simply be forgiven by God, but that it must be settled by full payment. Some have argued that this entails that God does not forgive sin at all. (Stump 1988, 61-5) Forgiveness involves remitting some of the payment owed. On these theories however, the debt is paid in full. Most controversial, however, is the claim that moral debts of the sort in question here are transferable. That is, on this view it seems that the punishment of one can be fairly borne by another. While this might be acceptable in certain cases where monetary fines are involved, many think that it cannot apply to specifically moral debts.

Sacrifice theories do not encounter these difficulties. Instead they, like moral exemplar theories, face difficulties of two main sorts. First, both views seem unable to account for the Biblical emphasis on the necessity of Christ's passion to remedy the problems brought forth by sin. It is hard to see why Christ's passion plays any essential role in establishing him as moral exemplar. Further, it is hard to see why Christ's death would provide a suitable sacrifice. Why would it not suffice for Christ to dwell among us and live a perfect human life, resisting all earthly temptation? Second, both views seem unable to account for the necessity of the *horrible nature* of Christ's death on the cross. The reason for this is that both hold that God either could or does forgive the sin of creatures without such grave sacrifices being offered. As a result, one is left to wonder why a solution which does not involve such horrific suffering is preferred to simple forgiveness. This is especially problematic for the moral exemplar theories, which lay almost exclusive emphasis on the importance of Christ's moral example during his life and on the centrality of creaturely moral reform for reconciliation with God.

Defenses of substitution models seem to be on the wane in recent literature, with sacrifice and exemplar theories becoming more widely defended. Can the substitution models overcome the difficulties posed for it above? Some have defended substitution models according to which punishment is a fitting response to human sin, and yet also such that it might nonetheless be fairly borne by a surrogate, in this case, the perfect Christ. Stephen Porter, for example, argues that our moral intuitions generally incline us to view punishment of a surrogate as a bad thing, and that some case needs to be made for its permissibility in this instance (Porter 2001). In run of the mill cases of punishment, the good reasons for punishment (such as reform of the wrongdoer, making reparation, deterrence, and so forth) usually weigh in favor of *not* transferring the punishment to a surrogate. But here, Porter argues, the good reasons for punishing human sinners are not undercut, and that, in fact, there are outweighing reasons for allowing Christ to bear the punishment due human sinners.

Specifically, Porter claims that the goods that come from God's punishment of sin (namely, reparation, manifesting an objective correction to distorted human values, and moral education/reform) justify the punishment. What is more, Porter claims, these ends are more fittingly served through the suffering of Christ on our behalf. The reasons for this are two-fold. First, were we to bear the punishment directly, it might further serve to alienate us from God. Second, the gravity of human sin against an infinite God cannot be suitably expressed by punishment of merely finite humans. Punishment of an infinite God-man better expresses the seriousness of sin.

In Porter's account we have an attempt to respond to the three objections raised earlier against substitution views. First, the (infinite) severity of the punishment is required in order to adequately express the gravity of human sin against an infinite and perfect God. Concerning the second objection (namely, that paying the full price of sin means that there is no forgiveness on God's part), Porter can reply that the objection is simply misguided. God can forgive without any punishment being exacted. However, certain goods arise as a result of punishment being meted out, and God thus metes out punishment suitable for securing those goods. The third difficulty (i.e., the non-transferrability of moral debts) initially seemed to be the most formidable of the three. Porter argues, however, that as long as (a) offender, offended, and surrogate are willing participants, and (b) the goods of punishing can be secured through the punishment of the surrogate, then substitution is permissible, perhaps even preferable. The reason it is permissible, however, is not because the moral debt is “transferred” from sinner to Christ (as the objection assumes) but simply because punishing wrong is a good and punishing a surrogate can equally or better serve the aims of punishing.

Bibliography

Trinity

- Brown, D., 1989, “Trinitarian Personhood and Individuality.” in Feenstra and Plantinga 1989, pp. 48-78.
- Plantinga Jr., C., 1989, “Social Trinty and Tritheism”, in Feenstra, R., and Plantinga, Jr., C.,

(eds.), 1989, *Trinity, Incarnation, and Atonement.* Notre Dame: University of Notre Dame Press, pp. 21-47.
- Swinburne, R., 1994, *The Christian God*. Oxford: Clarendon Press
- van Inwagen, P., 1995, "And yet they are not three Gods but one God", in *God, Knowledge, and Mystery*. Ithaca: Cornell University Press, pp. 222-59.

Incarnation

- Feenstra, R., 1989, "Reconsidering Kenotic Christology." in Feenstra and Plantinga 1989, pp. 128-152.
- Morris, T. V., 1986, *The Logic of God Incarnate*. Ithaca: Cornell University Press.
- Relton, H. M., 1929, *A Study in Christology*. London: MacMillan.
- Senor, T., 1991, "God, Supernatural Kinds, and the Incarnation." *Religious Studies*, 353-370.
- Swinburne, R., 1989, "Could God Become Man?" in *The Philosophy in Christianity*, G. Vesey, (ed.), Cambridge: Cambridge University Press.

Atonement

- Jensen, P., "Forgiveness and Atonement", *Scottish Journal of Theology*, 46, pp. 141-159.
- Porter, S., 2001, "Substitution Reconsidered," in *Philosophy of Religion: A Contemporary Reader*, W.L. Craig (ed.), Edinburgh: Edinburgh University Press, 2002.
- Quinn, P., 1989, "Aquinas on Atonement," in Feenstra and Plantinga 1989, pp. 153-177.
- Stump, E., 1988, "Atonement According to Aquinas" in *Philosophy and the Christian Faith*, T.V. Morris (ed.), Notre Dame: University of Notre Dame Press, pp. 61-91.
- Swinburne, R., 1989, *Responsibility and Atonement*. Oxford: Oxford University Press.

Other Internet Resources

Links on the Trinity

- Key texts from the Church Fathers concerning the Trinity (compiled by Joseph Gallegos, Corunum Catholic Apologetic Web Site
- The Athanasian Creed (Christian Classics Ethereal Library)
- Boethius, *The Trinity is One God* (Christian Classics Ethereal Library)
- St. Augustine of Hippo *On the Holy Trinity* (Christian Classics Ethereal Library)
- Jonathan Edwards 'Unpublished Essay on the Trinity' (Christian Classics Ethereal Library)

Links on the Incarnation:

- Aquinas, *Treatise on the Incarnation* (Christian Classics Ethereal Library)

- St. Anselm, *Cur Deus Homo* (The Internet Medieval Sourcebook, Paul Halsall (ed.), Fordham University)

Links on Atonement

- 'Doctrine of the Atonement' (W.H. Kent, Catholic Encyclopedia)
- Understanding Atonement: A New and Orthodox Theory, (Robin Collins, Philosophy, Messiah College)

Related Entries

Aquinas, Saint Thomas | atonement | Augustine, Saint | faith and reason | incarnation | theology | Trinity

Michael Murray
m_murray@acad.fandm.edu

First published: May 13, 2002
Content last modified: May 13, 2002

Thomas of Erfurt

Thomas of Erfurt was the most influential member of a group of later medieval philosophers known as the speculative grammarians or *Modistae* (Modists), after the central place they assigned to the *modi significandi* (modes of signification) of a word in the analysis of human discourse. The notion that a word, once it has been imposed to signify, carries with it all of its syntactical modes, or possible combinations with other words, had been around since the 12th century. What the *Modistae* did was to explain the origins of the *modi significandi* in terms of parallel theories of *modi intelligendi* (modes of understanding) and *modi essendi* (modes of being). The result was a curious amalgam of philosophy, grammar, and linguistics. Thomas of Erfurt's *De modis significandi* became the standard Modist textbook in the 14th century, though it enjoyed even greater fame later thanks to its misidentification as a work of Duns Scotus. The text was eventually printed as part of Scotus's *Opera Omnia*, where it was read and commented upon by later figures such as Charles S. Peirce and Martin Heidegger, whose 1916 doctoral thesis, *Die Kategorien- und Bedeutungslehre des Duns Scotus*, should have been entitled, *Die Kategorienlehre des Duns Scotus und die Bedeutungslehre des Thomas von Erfurt*.

1. Life

Almost nothing is known about the life of Thomas of Erfurt except that he was active as a teacher and philosopher in the first quarter of the 14th century.[1] Presumably, he came from the city of Erfurt in present-day Germany. His work shows the influence of the Parisian Arts Masters Radulphus Brito (ca. 1270-1320) and Siger of Courtrai (ca. 1280-1341), which suggests that he was educated and perhaps also taught at the University of Paris.[2] Later documents associate him with the school of St. Severus and the *Schottenkloster* of St. Jacob at Erfurt.[3] His most famous grammatical text, *De modi significandi* (On the

Modes of Signifying) was known by 1310 and was already being commented upon by 1324. It is possible that he returned to Paris a number of times over the course of his academic career, although there are no records to attest to this.

Some copies of *De modi significandi* attribute the work to an early 14th-century English cleric named Thomas of Occam, but scholars have been skeptical about this because it occurs in just a handful 15th-century manuscripts.[4] The vast majority of the manuscript evidence, and all of the earliest witnesses, refer to its author as Thomas of Erfurt.

2. Writings

Six works have been attributed to Thomas of Erfurt. In addition to the aforementioned grammatical treatise, whose full title is *Tractatus de modis significandi seu Grammatica speculativa* (ed. Bursill-Hall 1972),[5] there are four short *expositiones*, or literal commentaries: on Porphyry's *Isagoge*, Aristotle's *Categories*, Aristotle's *De interpretatione*, and the anonymous *Liber sex principiorum* (Book of Six Principles). Finally, there is a very brief work of mnemonic verses used to teach grammar to schoolboys, *Commentarius in carmen 'Fundamentum puerorum'* (ed. Gansiniec 1960), although its editor believes that it is actually an anonymous abridgement of *De modis significandi*. In any case, Thomas's entire reputation derives from *De modis significandi*, which is the only one of his works to have been studied in any detail.

De modis significandi proved to be so popular that it became the standard (and later, the representative) text for the Modist tradition (see next section). It exists in over forty 14th- and 15th-century manuscripts. A printed edition appeared in the late 15th century that was reprinted an incredible 11 times before its 'definitive' reprinting in Luke Wadding's 1639 edition of the complete works of Duns Scotus.[6]

3. Modism

Thomas of Erfurt belonged to an interesting though somewhat obscure group of late 13th- and early 14th-century philosophers known as the speculative grammarians or *Modistae*. The term 'speculative grammarian' is ambiguous because it is also used by historians of medieval philosophy to refer to 12th-century Parisian grammar masters such as William of Conches, Peter Helias, and Ralph of Beauvais, who systematically revised the ancient grammars of Donatus and Priscian -- textbooks which had been used to teach Latin to schoolboys -- in order to produce a universal semantics.[7] The two groups are related, as it turns out, since the latter-day grammarians adopted many of the theories as well as the universalizing tendencies of their 12th-century predecessors. Foremost among them was the theory of the *modi significandi*, or modes of signifying. The term '*Modistae*' or 'Modist' properly refers to the later group.

There was already a grammatical concept of the *modi significandi* by the time the first *Modistae*, Martin of Dacia and Boethius of Dacia, appeared in Paris sometime before 1270.[8] Originally, the idea referred

to the different ways in which a word or expression (*dictio*) is able to signify something. Words themselves are the product of a primary act of imposition by which a particular utterance is connected with some thing or property of a thing, the utterance providing the matter, which in turn is said to be 'informed' by the act of imposition. The word acquires its *modi significandi* through a second act of imposition encoding all of the general syntactic roles it can play in connection with other words and expressions, i.e., the various parts of speech it can fulfill (e.g., noun, verb, adverb) and the grammatical forms of these parts (e.g., the gender, number, and case of nouns; the tense and mood of verbs). These modes are said to cause the various lexical forms exhibited by the word in spoken and written languages. Thus, Latin uses the word 'canis' to signify what English speakers mean by 'dog', but the same modes of signifying determine their function as singular and as nouns.[9] We might think of *modi significandi* metaphorically as hooks or fasteners on a word because they reflect its potential for combination with other words in propositions and other grammatically well-formed constructions.

It is easy to see how such an account of meaning could lead to a *bona fide* theory of grammar, what Jan Pinborg has called "the first systematic syntax developed in western linguistics".[10] Indeed, speculative grammarians of both periods have attracted the attention of linguists and philosophers alike.[11] But the stakes were higher for the *Modistae* because the influx of Aristotle's writings on metaphysics and natural philosophy in the late 12th and early 13th centuries forced everyone to think in terms of a new paradigm of knowledge. Could grammar be conceived as the science of language (*scientia sermocinalis*)? The question here was whether grammar could count as an Aristotelian speculative science, i.e., whether it is demonstrative in the sense of being a knowledge-producing activity ordained by a single subject whose principles are universal and necessary.[12] Concerns of this sort clearly underlie Boethius of Dacia's appraisal of traditional grammar:

> subjects in which a demonstrative mode of knowing is possible are seldom taught in a demonstrative way, but descriptively [*sed modo narrativo*] … That is why Priscian states many conclusions for which he offers no reasons, but merely the authority of ancient grammarians. Accordingly, he does not teach, for only those who offer reasons for what they say are teachers.[13]

Thomas of Erfurt indicates in the opening lines of *De modis significativis* that he aims to be one of these teachers:

> *The rationale of the method*. Since in all science, to understand and to know come about from the knowledge of principles, as is written in I *Physics*, text comment 1, therefore, wishing to have knowledge of the science of grammar, it is primarily necessary for us to dwell upon all of its primary *per se* principles of which modes are the modes of signifying. But before knowledge is sought after in the particulars, there are certain things to be set out in advance in general, without which it is not possible to have the fullest understanding of them. (*DMS*, Preamble; Bursill-Hall: 1)

For *Modistae* such as Boethius of Dacia and Thomas of Erfurt, the proper subject of grammar is well-

formed, significant speech (*sermo congrue significativus*), the principles of which are expressed in the *modi significandi*.

But this is where things get interesting, for the *Modistae* also understood the theory of the *modi significandi* to involve claims about the nature of thought and reality. The problem with Priscian, as 12th-century grammarians had pointed out, is that he said nothing about what causes the different parts of speech. Hence the need for a theory of the *modi significandi*. But the *Modistae* saw that this could only be part of the answer, for grammar is a linguistic phenomenon, and linguistic phenomena must have a cause within the natural order of things. Therefore, to complete the explanation, they argued that the formal structure of the *modi significandi* owes its existence to *modi intelligendi*, or modes of understanding, which in turn are caused by *modi essendi*, or the modes of being a thing can exhibit outside the mind.

The heart of the Modist project is the assumption that there is a triadic or parallel relationship between word, concept, and thing. Meaning is based proximately on understanding but ultimately on being. According to Thomas of Erfurt:

> *Every mode of signifying is from some property of the thing*. Concerning the second thing to be noted, that since such notions or modes of signifying are not fictions, it must be that every mode of signifying radically originates from some property of the thing. This is plain thus: since the intellect, in order to signify, imposes the voice under some mode of signifying, it considers the property itself of the thing from which it originally drew the mode of signifying; this is because the intellect, since it is a passive power indeterminate of itself, does not advance to a determinate act unless it is determined from another source. Whence since it imposes the voice in order to signify under the determinate mode of signifying, it is necessarily moved by a determinate property of the thing; therefore some property of a thing, or mode of being of a thing, corresponds to any mode of signifying. (*DMS*, 2.4; Bursill-Hall: 3)

Note the phrase, “since such notions or modes of signifying are not fictions.” The *modi significandi* could not play any causal role in determining the parts of speech, nor the corresponding *modi intelligendi* in determining the *modi significandi*, nor the *modi essendi* in determining the *modi intelligendi*, unless they all actually exist. Here the *Modistae* drew on the Aristotelian idea that although spoken sounds and written marks differ from language to language, “what these are in the first place signs of - affections of the soul - are the same for all; and what these affections are likenesses of - actual things - are also the same” (*De interpretatione* 1.16a3-9).[14] Likewise, the ordering of the modes was intended to replicate the ordering of disciplines in Aristotelian speculative science: just as psychology, the study of moving things *qua* animate, is subordinate to physics, the study of things *qua* moving, and physics is subordinate to metaphysics, the study of things *qua* existing, so the principles of grammar (*modi significandi*) are derived from mental acts of signifying (*modi intelligendi*), which reflect the way things actually are (*modi essendi*).

Tidy conceptual schemes tend to become less so when confronted with the facts. The *Modistae* spent a lot of time trying to explain recalcitrant linguistic data. After sketching the origins of the *modi significandi* in the first chapters of his treatise, Thomas immediately considers a few objections: how can ‘goddess [*dea*]’ be signified with a feminine noun, which connotes passivity? Answer: a mode of signifying need not always be drawn from the thing signified, for sometimes it “can be taken from the property of the thing of another utterance;” thus, when we say ‘in God’, we do not mean to attribute a passive property to God, who suffers not, but only to “imagine him as if being affected by our prayers.” When we use a passive or feminine mode of signifying in relation to God, what we are really doing is signifying our own passive or feminine conception of some *other* thing that *is* a genuine recipient, i.e., something that does correspond to a determinate mode of being *qua* recipient, and then imposing the same word to signify God. In the same way, we impose names on things we cannot sense via the properties of sensible things, thereby attributing “active modes of signifying to their names” (*DMS* 2.5; Bursill-Hall: 4).

But what about words signifying fictions, such as ‘chimera’, or privations, such as ‘blindness’? These do not correspond to any mode of being, active or passive, since they signify nothing (i.e., no thing). According to Thomas, the active modes of signifying chimeras “are taken from the parts from which we imagine a chimera to be composed, which [fiction] we imagine from the head of a lion, the tail of a dragon, etc.” (*DMS* 2.5; Bursill-Hall: 4) Names of privations, on the other hand, “designate the modes of understanding of privations, which are their modes of being, through their own active modes of signifying about privations”. The idea here is that ‘blindness’ corresponds to our positive concept of sight *qua* absent, which enjoys a positive *modus essendi* in our intellects. Thus, “although privations are not positive beings outside the soul, they are nevertheless positive beings in the soul … and since the understanding of them is their being, therefore, their mode of understanding would be their mode of being” (*DMS* 2.6; Bursill-Hall: 4). Blindness outside the soul cannot cause any conception of itself *per se*, since blindness *per se* does not exist, and what does not exist cannot be the cause of anything. Accordingly, when we say, ‘Homer was a blind man,’ the word ‘blind’ actively signifies the passive mode of understanding something as being without sight, and owes its semantic function to the way its corresponding concept is understood. In the case of concepts that are neither fictions nor privations, these modes of understanding are further determined by their corresponding modes of being, i.e., by actual substances and properties outside the soul.

For the *Modistae*, then, the words ‘chimera’ and ‘horse’ differ, but only in terms of the complexity of their underlying modes. Thomas follows Siger of Courtrai in distinguishing between active and passive *modi significandi*/*modi intelligendi* to explain the difference between the act of signifying/understanding (materially construed as a property of the utterance/concept) and the object signified/understood (materially construed as a property of the thing signified/understood). The fact that nothing answers to the name ‘chimera’ simply does not matter. Modism was a theory about meaning (*significatio*) of a word as opposed to its reference (*suppositio*), and reference was regarded as something determined by the logicians. Besides, if grammatical truths are universal and necessary -- i.e., if there really is a science of grammar -- then they cannot be altered by the fact that there are no chimeras. It is the assumption that some palpable phenomenon must underlie every grammatical truth, causing it to be the way it is, which guides Thomas’s discussion in the remainder of the treatise, which soon moves on to more practical

matters such as the different parts of speech (‘*Etymologia*’) and their syntax (‘*Diasynthetica*’). Included in the latter are the concepts of *constructio* (the syntactic joining of one word to another), *congruitas* (the proper formation of such constructions), and *perfectio* (the proper formation of complete expressions).

Modist explanations of semantic phenomena were extremely complicated. Indeed, if we consider the manifold ways in which a word can be a signifier together with its potential for combination with other words in expressions, there could be infinitely many *modi significandi* -- a proliferation of modes repeated all over again at the level of *modi intelligendi*. Philosophers today think of Meinong’s theory of objects as a paradigm case of ontological incontinence. These philosophers have never met the *Modistae*, who would make even Meinong seem parsimonious. But in a way this is unfair, of course, since the *Modistae* were not ontologists and had no interest in the metaphysical consequences of their theories. They were first and foremost concerned with the phenomenon of linguistic meaning, and invoked whatever entities they felt could best explain what they observed.

Thomas’s *De modis significandi* was the last work to develop Modist doctrine to any significant degree.[15] Its clarity and relative brevity led to its adoption as the standard Modist text in medieval universities, replacing the earlier *Modi significandi* of Martin of Dacia. But by 1330 and for reasons that remain unclear, Modism had completely disappeared from Paris, pushed aside by the more powerful and comprehensive approach of the *Summulae de dialectica* of John Buridan (ca. 1300-1361).[16] Modism was never able to overcome certain difficulties, such as its refusal to recognize extra-linguistic context, as a result of which it could not explain how meaning can be communicated via incongruous or imperfect expressions.[17] But the most likely reason for its demise was that it no longer provided satisfying explanations of the phenomena it was supposed to explain. The one thing it could do, provide an account for the syntax of Latin grammar, could be achieved more economically by other means. In addition, the theory became absurdly complicated in order to save the phenomena of the *modi significandi*, suggesting that the *modi* finally collapsed under their own weight, like so many Ptolemaic epicycles.

4. Influence

Thomas of Erfurt’s *De modis significandi* has enjoyed more attention than it might otherwise have received thanks to its early misidentification as a work of Duns Scotus.[18] As a result, it was printed along with authentic works on logic in volume 1 of Luke Wadding’s 17th-century edition of the *Opera Omnia* of Duns Scotus (Lyons 1639), and again in the 19th-century reprint of Wadding by Juan-Luis Vivès (Paris 1891).[19] Until very recently, the Wadding-Vivès edition was the definitive source for the works of Duns Scotus, so that anyone consulting it would have associated *De modis significandi* with him.[20] Complicating the story somewhat is the fact that Duns Scotus was influenced by Modism early in his career, though this was probably due to his exposure to Modist authors such as Simon of Faversham and Andrew of Cornwall.[21] It is unlikely that he was influenced by Thomas of Erfurt, because *De modis significandi* did not circulate widely until after Scotus's death.

One of the many later figures to have unwittingly admired Thomas was the American philosopher,

Charles Sanders Peirce (1839-1914), who regarded Duns Scotus as a fellow traveler in metaphysics and whose own semiotic theory resembles the Modist program in certain respects.[22] Peirce quotes *verbatim* the first six chapters of *De modis significandis* in an 1869 lecture comparing the views of (what he took to be) Duns Scotus and William of Ockham on names and signification. But the lecture is introductory in character, leaving it uncertain whether Peirce fully appreciated what was at stake in Modistic grammar, rather than viewing it, say, as some sort of linguistic addendum to Scotus's metaphysics.[23]

Another figure so influenced was Martin Heidegger, whose *Habilitationsschrift* was published in 1916 under the title, *Die Kategorien- und Bedeutungslehre des Duns Scotus*. This book has not been much studied by Heidegger scholars, which is a pity because it is more about Heidegger's own project of advancing the Husserlian notion of *a priori* grammar than a work of philosophical exposition and interpretation. In what now seems a classic understatement, the historian of medieval philosophy Martin Grabmann wrote of this book in 1926, "Martin Heidegger has demonstrated the continuity of the *Grammatica speculativa* hitherto attributed to Duns Scotus with the terminology and overall intellectual outlook of Husserl, so that the structure and distinctiveness of the medieval original is somewhat obscured".[24] In his defense, Heidegger never pretends to be doing history of philosophy in *Die Kategorien- und Bedeutungslehre des Duns Scotus*. On the contrary, he states at the beginning of Part II that he is mostly interested in exploring the implications of *De modis significandi* for the broader theory of meaning, and that he is following scholarly consensus in attributing it to Duns Scotus.[25]

In 1922, however, Grabmann conclusively demonstrated that it was Thomas of Erfurt who wrote *De modis significandi*, not Duns Scotus.[26] So the claim that Heidegger wrote his *Habilitationsschrift* on Duns Scotus is only partly true. The first half, *die Kategorienlehre*, is correctly addressed to Duns Scotus's theory of the categories as developed in his authentic commentaries on Porphyry's *Isagoge*, and Aristotle's *Categories* and *De sophisticis elenchis*. But the second half, *die Bedeutungslehre*, is based almost entirely on Thomas of Erfurt.[27]

For more recent influence, we need look no further than Jacques Derrida, who mentions Thomas (this time dressed as himself) in connection with the old Peircean idea that logic is a branch of semiotics: "As in Husserl (but the analogy, although it is most thought-provoking, would stop there and one must apply it carefully), the lowest level, the foundation of the possibility of logic (or semiotics) corresponds to the project of the *Grammatica speculativa* of Thomas d'Erfurt, falsely attributed to Duns Scotus".[28] What impresses Derrida is not so much the reduction of logic to grammar as the deconstructive potential of Husserl's phenomenology of signs. The idea of deconstructing the science of language would have struck Thomas as absurd, of course, though he probably would have felt some affinity with the highly variegated notion of the sign in modern semiotics.

Bibliography

Primary Sources

- Bursill-Hall, G. L. (ed. and tr.): 1972, *Thomas of Erfurt: Grammatica Speculativa*, The Classics of Linguistics, 1, London, Longmans.[*DMS*]
- Gansiniec, R. (ed.): 1960, [Thomas of Erfurt,] "*Fundamentum Puerorum*," in *Metrificale marka z opatowca i traktaty gramatyczne XIV i XV wieku [= The Chronicles of Mark of Opatowec and Grammatical Treatises of the XIV*th *and XV*th *Centuries]*, Studia staropolskie, VI, Wroclaw, 105-6.
- Garcia, M. Fernandez (ed.): 1902, *B. Ioannis Duns Scoti Doct. Subtilis O.F.M. Grammaticae speculativae nova editio*, Ad Claras Aquas, Quaracchi.
- Grotz, Stephan (tr.): 1998, *Thomas von Erfurt, Abhandlung über die bedeutsamen Verhaltensweisen der Sprache.* Tractatus De modis significandi, John Benjamins, Amsterdam-Philadelphia. [German translation of *De Modis Significandis*]
- Wadding, Luke (ed.): 1639, *Ioannis Duns Scoti Opera Omnia*, vol. 1, Durand, Lyons, 45-76.

Secondary Sources

- Ashworth, E. J.: 1977, *The Tradition of Medieval Logic and Speculative Grammar*, Pontifical Institute of Mediaeval Studies, Toronto.
- Biard, Joël: 1989, *Logique et théorie du signe au XIV*^{e} *siècle*, Vrin, Paris.
- Bursill-Hall, G. L.: 1971, *Speculative Grammars of the Middle Ages: The Doctrine of the* partes orationis *of the* Modistae, Approaches to Semantics, 11, Mouton, The Hague.
- Derrida, Jacques: 1976, *Of Grammatology*, tr. Gayatri Chakravorty Spivak, The Johns Hopkins University Press, Baltimore-London.
- Ebbesen, Sten: 1998, "The Paris Arts Faculty: Siger of Brabant, Boethius of Dacia, Radulphus Brito," in *Medieval Philosophy*, ed. John Marenbon, The Routledge History of Philosophy, III, Routledge, London-New York, 269-90.
- Fredborg, Karin Margareta: 1980, "Universal Grammar According to Some 12th-Century Grammarians," in *Studies in Medieval Linguistic Thought*, ed. Konrad Koerner et al., Historiagraphia Linguistica, VII.1/2, John Benjamins, Amsterdam, 69-84.
- Fredborg, Karin Margareta: 1988, "Speculative Grammar," in *A History of Twelfth-Century Philosophy*, ed. Peter Dronke, Cambridge University Press, Cambridge-New York, 177-95.
- Grabmann, Martin: 1922, "De Thoma Erfordiensi auctore Grammaticae quae Ioanni Duns Scoto adscribitur speculativae," *Archivum Franciscanum Historicum* 15, 273-7.
- Grabmann, Martin: 1926, *Mittelalterliches Geistesleben. Abhandlung zur Geschichte der Scholastik und Mystic*, Bd. 1, Max Heuber, München, 116-25.
- Grabmann, Martin: 1943, *Thomas von Erfurt und die Sprachlogik des mittelalterlichen Aristotelismus*, Sitzungsberichte der Bayerischen Akademie der Wissenschaft, Heft 2, Heuber, München.
- Heidegger, Martin: 1916, *Die Kategorien- und Bedeutungslehre des Duns Scotus*, J. C. B. Mohr, Tübingen. Rpr. in Martin Heidegger: 1972, *Frühe Schriften*, Vittorio Klostermann, Frankfurt am Main, 130-375.
- Heidegger, Martin: 1970, *Traité des categories et de la signification chez Duns Scot*, tr. Florent Gaboriau, Éditions Gallimard, Paris.

- Kaczmarek, Ludger (ed.): 1994, *Destructiones modorum significandi*, B. R. Grüner, Amsterdam-Philadelphia.
- Kelly, L. G.: 1971, "*De modis generandi*: Points of Contact between Noam Chomsky and Thomas of Erfurt," *Folia Linguistica* 5, 225-52.
- Kneepkens, C. H.: 1995, "The Priscianic Tradition," in *Sprachtheorien in Spätantike und Mittelalter*, ed. Sten Ebbesen, Gunter Narr Verlag, Tübingen, 239-64.
- Lorenz, Sönke: 1989, *Studium Generale Erfordense. Zum Erfurter Schulleben im 13. und 14. Jahrhundert*, Monographien zur Geschichte des Mittelalters, 34, Stuttgart, 312-25.
- Marmo, Costantino: 1994, *Semiotica e linguaggio nella scolastica: Parigi, Bologna, Erfurt, 1270-1330. La semiotica dei Modisti*, Instituto Storico Italiano per il Medioevo, Roma.
- Marmo, Costantino: 1995, "A Pragmatic Approach to Language in Modism," in *Sprachtheorien in Spätantike und Mittelalter*, ed. Sten Ebbesen, Gunter Narr Verlag, Tübingen, 169-83.
- Marmo, Costantino: 1999, "The Semantics of the *Modistae*," in *Medieval Analyses in Language and Cognition*, Acts of the Symposium, 'The Copenhagen School of Medieval Philosophy', January 10-13, 1996, ed. Sten Ebbesen and Russell L. Friedman, Royal Danish Academy of Sciences and Letters - C. A. Reitzels Forlag, Copenhagen, 83-104.
- Peirce, Charles S.: 1984, "Ockam. Lecture 3 [1869]" in *Writings of Charles S. Peirce: A Chronological Edition*, Volume 2: 1867-1871, ed. Edward C. Moore et al., Indiana University Press, Bloomington, 317-36.
- Pinborg, Jan: 1967, *Die Entwicklung der Sprachtheorie im Mittelalter*, BGPM, XLII.2, Aschendorff, Münster.
- Pinborg, Jan, Heinrich Roos and P. J. Jensen (eds.): 1969, *Boethii Dacii Modi Significandi sive Quaestiones super Priscianum Maiorem*, Corpus Philosophorum Danicorum Medii Aevi, 4, G. E. C. Gad, Copenhagen.
- Pinborg, Jan: 1974, Review of Bursill-Hall 1972 in *Lingua* 34 (1974): 369-73.
- Pinborg, Jan: 1982, "Speculative Grammar," in *The Cambridge History of Later Medieval Philosophy*, ed. Norman Kretzmann, Anthony Kenny, and Jan Pinborg, Cambridge University Press, Cambridge-New York, 254-69.
- Pironet, Fabienne: 1997, *The Tradition of Medieval Logic and Speculative Grammar: A Bibliography (1977-1994)*, Brepols, Turnhout.
- Trentman, John: 1975, "Speculative Grammar and Transformational Grammar: A Comparison of Philosophical Presuppositions," in Hermann Parret (ed.), *History of Linguistic Thought and Contemporary Linguistics*, Walter De Gruyter, Berlin.
- Robins, R. H.: 1997, *A Short History of Linguistics*, 4th edition, Longman, London-New York, 79-109.
- Rosier, Irène: 1983, *La grammaire spéculative des Modistes*, PUF, Paris.
- Rosier, Irène: 1995, "*Res significata et modus significandi*: Les implications d'une distinction médiévale," in *Sprachtheorien in Spätantike und Mittelalter*, ed. Sten Ebbesen, Gunter Narr Verlag, Tübingen, 135-68.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Duns Scotus, John | Heidegger, Martin | Peirce, Charles Sanders | Simon of Faversham

Jack Zupko
jzupko@emory.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 5, 2002
Content last modified: May 5, 2002

Stanford Encyclopedia of Philosophy
Notes to Erfurt

Notes

1. The factual information in this section and the next has been gathered from Lorenz 1989, the definitive source for Thomas of Erfurt's life and writings.

2. Stephan Grotz comments on the remarkable similarities between the commentaries of Siger and Thomas on Porphyry's *Isagoge* and Aristotle's *Categories* (1998: vii, n. 1). Likewise, Jan Pinborg notes that "long passages of the *Grammatica speculativa* [i.e. *De modis significandi*] are simply taken over from Radulphus Brito" (Pinborg 1974: 372; see also Pinborg 1967).

3. *Schottenklosteren* (lit. 'Scottish cloisters') were monasteries founded by Scottish and Irish Benedictine monks in continental Europe beginning in the 8th century.

4. See Pinborg 1967: 132 and Lorenz 1989: 312-13.

5. Bursill-Hall's edition unfortunately carries over a number of corrupt readings from the text of this work in Wadding 1639. For a list of corrected readings, see Pinborg 1974.

6. See Lorenz 1989: 317-24.

7. See Pinborg 1982 and especially Fredborg 1988. This earlier movement culminated in the *Summa super Priscianum* of Peter Helias (ca. 1100-1166), which is also the first example of the genre later made famous by Thomas Aquinas, in his theological *summae*.

8. Modism originated in Paris and was for the most part a Parisian phenomenon, although Modist philosophers later came to teach in England and (of course) Erfurt. For the former, see Robert Andrews: 1999, "Andrew of Cornwall and the Reception of Modism in England," in *Medieval Analyses in Language and Cognition*, Acts of the Symposium, 'The Copenhagen School of Medieval Philosophy', January 10-13, 1996, ed. Sten Ebbesen and Russell L. Friedman, Royal Danish Academy of Sciences and Letters - C. A. Reitzels Forlag, Copenhagen.

9. Of course, not all of the possible *modi significandi* of a word will be realized in all languages. In Latin, 'canis' is masculine, but English omits the modus of gender as a means of distinguishing nouns.

10. Pinborg 1982: 260. Karin Margareta Fredborg (1980) has shown that virtually all 12th-century

authors accepted the idea of a universal grammar.

11.See, e.g., Kelly 1971. For a linguistic account of Modism, see Bursill-Hall 1971 and Robins 1997: 79-109. A more details study can be found in Michael A. Covington: 1984, *Syntactic Theory in the High Middle Ages: Modistic Models of Sentence Structure*, Cambridge University Press, London-New York.

12. More than anything else, it was the search for properly Aristotelian foundations for grammar that led to the eclipse of the 12th-century notion of *grammatica universalis* and its gradual replacement by the full-blown *grammatica speculativa*. See Kneepkens 1995: 247-8.

13. Boethius of Dacia, *Modi Significandi*, Q. 9; Pinborg, Roos, and Jensen 1969: 39, ll. 24-33.

14. Aristotle, *De interpretatione* 1.16a3-9, tr. J. L. Ackrill in Jonathan Barnes (ed.): 1984, *The Complete Works of Aristotle*, vol. 1, Bollingen Series LXXI.2, Princeton University Press, Princeton NJ, 25.

15. For a discussion of developments among "second generation" *Modistae* such as Thomas of Erfurt, see Marmo 1995.

16. 1330 is regarded as the *terminus ad quem* of Modism because in that year, the Averroist master John Aurifaber delivered a scathing attack at the University of Erfurt on the doctrine of the *modi significandi* from which it never fully recovered. But by then the doctrine was already in retreat at Paris, where it had suffered a similar fate at the hands of another Averroist, John of Jandun (see Pinborg 1967: 195-209; 1982: 267-8). The most famous critique of modism -- called, appropriately enough, the *Destructiones modorum significandi* -- was composed nearly half a century later by an unknown Parisian author, once believed to have been Peter d'Ailly (Kaczmarek 1994). Also appearing in the third quarter of the 14th century was the *Quaestiones super secundam partem Doctrinalis* of a certain Master Marsilius (= Marsilius of Inghen?), who seems to have been acquainted with the *Destructiones*. Modism found a few isolated advocates in later centuries, but never again emerged into the semantic mainstream. For discussion of the Erfurt and Paris responses to Modism, see Biard 1989: 242-88.

17. On these points, see Pinborg 1982 and Marmo 1995.

18. Grabmann has found evidence that the misidentification occurred as early as the first half of the 15th century (1926: 116-25). And Duns Scotus is identified as its author in one of the earliest (1491) printed editions (Lorenz 1989: 321ff.).

19. A separate edition of the Wadding version of *De modis significandi* was printed in Garcia 1902.

20. The Wadding-Vivès edition has been made obsolete for many of Duns Scotus's works by the new critical edition (in progress) of his *Opera Omnia*, which, unlike its predecessors, is being prepared according to modern editorial principles. His authentic commentaries on Porphyry's *Isagoge* and

Aristotle's *Categories* have now appeared as Volume I of his *Opera Philosophica* (St. Bonaventure, NY: The Franciscan Institute, 1999), with his commentaries on Aristotle's *Sophistical Refutations* and *On Interpretation* (both versions) slated for Volume II (in press).

21. For discussion, see the editors' introduction to *B. Ioannis Duns Scoti, Quaestiones in Librum Porphyrii Isagoge et Quaestiones super Praedicamenta Aristotelis*, ed. R. Andrews et al., Opera Philosophica I, (St. Bonaventure, NY: The Franciscan Institute, 1999): xxxi-xxxiv.

22. Both, for example, understood logic as belonging to semiotics or the general theory of signs. For discussion, see Maurizio Ferriani, "Peirce's Analysis of the Proposition: Grammatical and Logical Aspects," in D. Buzzetti and M. Ferriani (eds.): 1987, *Speculative Grammar, Universal Grammar and Philosophical Analysis of Language*, John Benjamins, Amsterdam-Philadelphia.

23. Peirce viewed logic as a species of semiotics, but Thomas of Erfurt draws no conclusions at all about logic, which he would have regarded as a separate field of inquiry.

24. Grabmann 1926: "Martin Heidegger hat den Gedankengang der bisher dem Duns Scotus zugeteilten *Grammatica speculativa* mit dem Terminologie und der ganzen geistigen Einstellung Husserls wiedergegeben, so daß das mittelalterliche Original in seiner Eigenart und Struktur etwas zurücktritt" (118).

25. Heidegger 1972: 245-6.

26. See Grabmann 1922. It is noteworthy that there were concerns about the authenticity of Duns Scotus's logic within two centuries of his death. The early 16th-century logician Jacob Naveros found the inconsistencies between these texts and his commentary on the *Sentences* sufficient to doubt whether he had written any logical works at all (see E. J. Ashworth, "Jacobus Naveros (fl. ca. 1533) on the Question: 'Do Spoken Words Signify Concepts or Things?'," in *Logos and Pragma. Essays on the Philosophy of Language in Honour of Professor Gabriel Nuchelmans*, ed. L. M. de Rijk and H. A. G. Braakhuis (Nijmegen: Ingenium, 1987): 204). Modern editors have identified four works as authentic: the commentaries on Aristotle's *Categories*, *On Interpretation* (two different versions), *Sophistical Refutations*, and Porphyry's *Isagoge*.

27. Heidegger does not seem ever to have acknowledged the mistake in print, despite having known Grabmann personally, and despite having a splendid opportunity to set the record straight in his introduction to the 1972 reprint of *Die Kategorien- und Bedeutungslehre des Duns Scotus*. This seems to have precipitated needless confusion about his exact relation to Duns Scotus. For example, in a French translation of this work (Heidegger 1970), there is nary a word to suggest that he might have been commenting on anyone other than Duns Scotus. The translator remarks in the introduction that he was able to benefit on several points from correspondence with Heidegger himself. Evidently, the topic of the authenticity of *De modis significandi* was never raised.

28. Derrida 1976: 48.

Jack Zupko
jzupko@emory.edu

First published: May 5, 2002
Content last modified: May 5, 2002

The Epsilon Calculus

The epsilon calculus is a logical formalism developed by David Hilbert in the service of his program in the foundations of mathematics. The epsilon operator is a term-forming operator which replaces quantifiers in ordinary predicate logic. Specifically, in the calculus, a term $\varepsilon x\, A$ denotes *some x* satisfying $A(x)$, if there is one. In Hilbert's Program, the epsilon terms play the role of ideal elements; the aim of Hilbert's finitistic consistency proofs is to give a procedure which removes such terms from a formal proof. The procedures by which this is to be carried out are based on Hilbert's epsilon substitution method. The epsilon calculus, however, has applications in other contexts as well. The first general application of the epsilon calculus was in Hilbert's epsilon theorems, which in turn provide the basis for the first correct proof of Herbrand's theorem. More recently, variants of the epsilon operator have been applied in linguistics and linguistic philosophy to deal with anaphoric pronouns.

Overview

By the turn of the century David Hilbert and Henri Poincaré were recognized as the two most important mathematicians of their generation. Hilbert's range of mathematical interests was broad, and included an interest in the foundations of mathematics: his *Foundations of Geometry* was published in 1899, and of the list of questions posed to the International Congress of Mathematicians in 1900, three addressed distinctly foundational issues.

Following the publication of Russell's paradox, Hilbert presented an address to the Third International Congress of Mathematicians in 1904, where, for the first time, he sketched his plan to provide a rigorous foundation for mathematics via syntactic consistency proofs. But he did not return to the subject in earnest until 1917, when he began a series of lectures on the foundations of mathematics with the assistance of Paul Bernays. Although Hilbert was impressed by the work of Russell and Whitehead in their *Principia Mathematica*, he became convinced that the logicist attempt to reduce mathematics to logic could not succeed, due in particular to the non-logical character of their axiom of reducibility. At the same time, he judged the intuitionistic rejection of the law of the excluded middle as unacceptable to mathematics. Therefore, in order to counter concerns raised by the discovery of the logical and set-theoretic paradoxes, a new approach was needed to justify modern mathematical methods.

By the summer of 1920, Hilbert had formulated such an approach. First, modern mathematical methods were to be represented in formal deductive systems. Second, these formal systems were to be proved syntactically consistent, not by exhibiting a model or reducing their consistency to another system, but by a direct metamathematical argument of an explicit, "finitary" character. The epsilon calculus was to provide the first component of this program, while his epsilon substitution method was to provide the second.

The epsilon calculus is, in its most basic form, an extension of first-order predicate logic with an "epsilon operation" that picks out, for any true existential formula, a witness to the existential quantifier. The extension is conservative in the sense that it does not add any new first-order consequences. But, conversely, quantifiers can be defined in terms of the epsilons, so first-order logic can be understood in terms of quantifier-free reasoning involving the epsilon operation. It is this latter feature that makes the calculus convenient for the purpose of proving consistency. Suitable extensions of the epsilon calculus make it possible to embed stronger, quantificational theories of numbers and sets in quantifier-free calculi. Hilbert expected that it would be possible to demonstrate the consistency of such extensions.

The Epsilon Calculus

In his Hamburg lecture in 1921 (1922), Hilbert first presented the idea of using choice functions to deal with the principle of the excluded middle in a formal system for arithmetic. These ideas were developed into the epsilon calculus and the epsilon substitution method in a series of lecture courses between 1921 and 1923, and in Hilbert's (1923). The final presentation of the epsilon-calculus can be found in Wilhelm Ackermann's dissertation (1924).

This section will describe a version of the calculus corresponding to first-order logic, while extensions to first- and second-order arithmetic will be described below.

Let L be a first-order language, which is to say, a list of constant, function, and relation symbols with specified arities. The set of epsilon terms and the set of formulae of L are defined inductively, simultaneously, as follows:

- Each constant of L is a term.
- Each variable is a term.
- If s and t are terms, then $s = t$ is a formula.
- If $s_1, \ldots, s_k$ are terms and F is a k-ary function symbol of L, $F(s_1, \ldots, s_k)$ is a term.
- If $s_1, \ldots, s_k$ are terms and R is a k-ary relation symbol of L, $R(s_1, \ldots, s_k)$ is a formula.
- If A and B are formulae, so are $A \wedge B$, $A \vee B$, $A \rightarrow B$, and $\neg A$.
- If A is a formula and x is a variable, $\varepsilon x\, A$ is a term.

Substitution and the notions of free and bound variable, are defined in the usual way; in particular, the variable x becomes bound in the term $\varepsilon x\, A$. The intended interpretation is that $\varepsilon x\, A$ denotes *some* x satisfying A, if there is one. Thus, the epsilon terms are governed by the following axiom (Hilbert's "transfinite axiom"):

$$A(x) \rightarrow A(\varepsilon x\, A)$$

In addition, the epsilon calculus includes a complete set of axioms governing the classical propositional connectives, and axioms governing the equality symbol. The only rules of the calculus are the following:

- Modus ponens
- Substitution: from $A(x)$, conclude $A(t)$, for any term t.

Earlier forms of the epsilon calculus (such as that presented in (Hilbert 1923)) use a dual form of the epsilon operator, in which $\varepsilon x\, A$ returns a value *falsifying* $A(x)$. The version above was used in Ackermann's dissertation, and has become standard.

Note that the calculus just described is quantifier-free. Quantifiers can be *defined* as follows:

$$\exists x\, A(x) \equiv A(\varepsilon x\, A)$$

$$\forall x\, A(x) \equiv \mathrm{A}(\varepsilon x\, (\neg A))$$

The usual quantifier axioms and rules can be derived from these, so the definitions above serve to embed first-order logic in the epsilon calculus. The converse is, however, not true: not every formula in the epsilon calculus is the image of an ordinary quantified formula under this embedding. Hence, the epsilon calculus is more expressive than the predicate calculus, simply because epsilon term can be combined in more complex ways than quantifiers.

It is worth noting that epsilon terms are nondeterministic, thereby representing a form of the axiom of choice. For example, in a language with constant symbols a and b, $\varepsilon x\, (x = a \vee x = b)$ is either a or b, but the calculus leaves it entirely open as to which is the case. One can add to the calculus a schema of *extensionality*,

$$\forall x\,(A(x) \leftrightarrow B(x)) \rightarrow \varepsilon x\, A = \varepsilon x\, B$$

which asserts that the epsilon operator assigns the same witness to equivalent formulae *A* and *B*. For many applications, however, this additional schema is not necessary.

The Epsilon Theorems

The second volume of Hilbert and Bernays' *Grundlagen der Mathematik* (1939) provides an account of results on the epsilon-calculus that had been proved by that time. This includes a discussion of the first and second epsilon theorems with applications to first-order logic, the epsilon substitution method for arithmetic with open induction, and a development of analysis (that is, second-order arithmetic) with the epsilon calculus.

The first and second epsilon theorems are as follows:

First epsilon theorem: Suppose $\Gamma \cup \{A\}$ is a set of quantifier-free formulae not involving the epsilon symbol. If *A* is derivable from Γ in the epsilon calculus, then *A* is derivable from Γ in quantifier-free predicate logic.

Second epsilon theorem: Suppose $\Gamma \cup \{A\}$ is a set of formulae not involving the epsilon symbol. If *A* is derivable from Γ in the epsilon calculus, then *A* is derivable from Γ in predicate logic.

In the first epsilon theorem, "quantifier-free predicate logic" is intended to include the substitution rule above, so quantifier-free axioms behave like their universal closures. Since the epsilon calculus includes first-order logic, the first epsilon theorem implies that any detour through first-order predicate logic used to derive a quantifier-free theorem from quantifier-free axioms can ultimately be avoided. The second epsilon theorem shows that any detour through the epsilon calculus used to derive a theorem in the language of the predicate calculus from axioms in the language of the predicate calculus can also be avoided.

More generally, the first epsilon theorem establishes that quantifiers and epsilons can always be eliminated from a proof of a quantifier-free formula from other quantifier-free formulae. This is of particular importance for Hilbert's program, since the epsilons play the role of ideal elements in mathematics. If quantifier-free formulae correspond to the "real" part of the mathematical theory, the first epsilon-theorem shows that ideal elements can be eliminated from proofs of real statements, provided the axioms are also real statements.

This idea is made precise in a certain general consistency theorem which Hilbert and Bernays derive from the first epsilon-theorem, which says the following: Let *F* be any formal system which results from the predicate calculus by addition of constant, function, and predicate symbols plus true axioms which are quantifier- and epsilon-free, and suppose the truth of atomic formulae in the new language is

decidable. Then F is consistent in the strong sense that every derivable quantifier- and epsilon-free formula is true. Hilbert and Bernays use this theorem to give a finitistic consistency proof of elementary geometry (1939, Sec 1.4).

The difficulty for giving consistency proofs for arithmetic and analysis consists in extending this result to cases where the axioms also contain ideal elements, i.e., epsilons.

Herbrand's Theorem

Hilbert and Bernays used the methods of the epsilon calculus to establish theorems about first order logic that make no reference to the epsilon calculus itself. One such example is *Herbrand's theorem* . This is often formulated as the statement that if an existential formula

$$\exists x_1 \ldots \exists x_k\, A(x_1,\ldots, x_k)$$

is derivable in first-order predicate logic (without equality), where A is quantifier-free, then there are sequences of terms $t_1^1, \ldots, t_k^1, \ldots, t_1^n, \ldots, t_k^n$, such that

$$\mathrm{A}(t_1^1,\ldots,t_k^1) \vee \ldots \vee A(t_1^n, \ldots, t_k^n)$$

is a tautology. If one is dealing with first-order logic *with* equality, one has to replace "tautology" by "tautological consequence of substitution instances of the equality axioms"; following Shoenfield we will use the term "quasi-tautology" to describe such a formula.

The version of Herbrand's theorem just described follows immediately from the *Extended First Epsilon Theorem* of Hilbert and Bernays. Using methods associated with the proof of the second epsilon theorem, however, Hilbert and Bernays derived a stronger result that, like Herbrand's original formulation, provides more information. To understand the two parts of the theorem below, it helps to consider a particular example. Let A be the formula

$$\exists x_1\, \forall x_2\, \exists x_3\, \forall x_4\, B(x_1, x_2, x_3, x_4)$$

where B is quantifier-free. The negation of A is equivalent to

$$\forall x_1\, \exists x_2\, \forall x_3\, \exists x_4\, \neg B(x_1, x_2, x_3, x_4).$$

By Skolemizing, i.e., using function symbols to witness the existential quantifiers, we obtain

$$\exists f_2, f_4\, \forall x_1, x_3\, \neg B(x_1, f_2(x_1), x_3, f_4(x_1, x_3)).$$

Taking the negation of this, we see that the original formula is "equivalent" to

$$\forall f_2, f_4\ \exists x_1, x_3\ B(x_1, f_2(x_1), x_3, f_4(x_1, x_3)).$$

The first clause of the theorem below, in this particular instance, says that the formula *A* above is derivable in first-order logic if and only if there is a sequence of terms $t_1^1, t_3^1, \ldots, t_1^n, t_3^n$ in the expanded language with f_2 and f_4 such that

$$B(t_1^1, f_2(t_1^1), t_3^1, f_4(t_1^1, t_3^1)) \vee \ldots \vee B(t_1^n, f_2(t_1^n), t_3^n, f_2(t_1^n, t_3^n))$$

is a quasi-tautology.

The second clause of the theorem below, in this particular instance, says that the formula *A* above is derivable in first-order logic if and only if there are sequences of variables $x_2^1, x_4^1, \ldots, x_2^n, x_4^n$ and terms $s_1^1, s_3^1, \ldots, s_1^n, s_3^n$ in the *original language* such that

$$B(s_1^1, x_2^1, s_3^1, x_4^1) \vee \ldots \vee B(s_1^n, x_2^n, s_3^n, x_4^n)$$

is a quasi-tautology, and such that *A* is derivable from this formula using only the quantifier and idempotency rules described below.

More generally, suppose *A* is any prenex formula, of the form

$$\mathbf{Q}_1 x_1 \ldots \mathbf{Q}_n x_n\ \mathrm{B}(x_1, \ldots, x_n),$$

where *B* is quantifier-free. Then *B* is said to be the *matrix* of *A*, and an *instance of B* is obtained by substituting terms in the language of *B* for some of its variables. The *Herbrand normal form* A^{H} of *A* is obtained by

- deleting each universal quantifier, and
- replacing each universally quantified variable x_i by $f_i(x_i^1, \ldots, x_i^{k(i)})$, where $x_i^1, \ldots, x_i^{k(i)}$ are the variables corresponding to the existential quantifiers preceding $\mathbf{Q}_i$ in *A* (in order), and f_i is a new function symbol designated for this role.

When we refer to an *instance* of the matrix of A^{H}, we mean a formula that is obtained by substituting terms in the expanded language in the matrix of A^{H}. We can now state Hilbert and Bernays’s formulation of

Herbrand's theorem. (1) A prenex formula A is derivable in the predicate calculus if and only if there is a disjunction of instances of the matrix of A^{H} which is a quasi-tautology.

(2) A prenex formula A is derivable in the predicate calculus if and only if there is a disjunction $\bigvee_j B_j$ of instances of the matrix of A, such that $\bigvee_j B_j$ is a quasi-tautology, and A is derivable from $\bigvee_j B_j$ using the following rules:

- from $C_1 \vee \ldots \vee C_i(t) \vee \ldots \vee C_m$
  conclude $C_1 \vee \ldots \vee \exists x\, C_i(x) \vee \ldots \vee C_m$ and

- from $C_1 \vee \ldots \vee C_i(x) \vee \ldots \vee C_m$
  conclude $C_1 \vee \ldots \vee \forall x C_i(x) \vee \ldots \vee C_m$ (if x not in C_j for $j \neq i$),

as well as the idempotence of $\vee$ (from $C \vee C \vee D$ conclude $C \vee D$).

Herbrand's theorem can also be obtained by using cut elimination, via Gentzen's "midsequent theorem." However, the proof using the second epsilon theorem has the distinction of being the first complete and correct proof of Herbrand's theorem. Moreover, and this is seldom recognized, whereas the proof based on cut-elimination provides a bound on the length of the Herbrand disjunction only as a function of the cut rank and complexity of the cut formulas in the proof, the length obtained from the proof based on the epsilon calculus provides a bound as a function of the number of applications of the transfinite axiom, and the rank and degree of the epsilon-terms occurring therein. In other words, the length of the Herbrand disjunction depends only on the quantificational complexity of the substitutions involved, and, e.g., not at all on the propositional structure or the length of the proof.

The version of Herbrand's theorem stated at the beginning of this section is essentially the special case of (2) in which the formula A is existential. In light of this special case, (1) is equivalent to the assertion that a formula A is derivable in first-order predicate logic if and only if A^{H} is. The forward direction of this equivalence is much easier to prove; in fact, for any formula A, $A \rightarrow A^{\mathrm{H}}$ is derivable in predicate logic. Proving the reverse direction involves eliminating the additional function symbols in A^{H}, and is much more difficult, especially in the presence of equality. It is here that epsilon methods play a central role.

Given a prenex formula, the *Skolem normal form* A^{S} is defined dually to A^{H}, i.e., by replacing existentially quantified variables by witnessing functions. If Γ is a set of prenex sentences, let Γ^{S} denote the set of their Skolem normal forms. Using the deduction theorem and Herbrand's theorem, it is not hard to show that the following are pairwise equivalent:

- Γ proves A
- Γ proves A^{H}

- Γ^S proves A
- Γ^S proves A^H

The Epsilon Substitution Method and Arithmetic

As noted above, historically, the primary interest in the epsilon calculus was as a means to obtaining consistency proofs. Hilbert's lectures from 1917-1918 already note that one can easily prove the consistency of propositional logic, by taking propositional variables and formulae to range over truth values 0 and 1, and interpreting the logical connectives as the corresponding arithmetic operations. Similarly, one can prove the consistency of predicate logic (or the pure epsilon calculus), by specializing to interpretations where the universe of discourse has a single element. These considerations suggest the following more general program for proving consistency:

- Extend the epsilon calculus in such a way as to represent larger portions of mathematics.
- Show, using finitary methods, that each proof in the extended system has a consistent interpretation.

For example, consider the language of arithmetic, with symbols for 0, 1, +, $\times$, <. Along with quantifier-free axioms defining the basic symbols, one can specify that the epsilon terms $\varepsilon x A(x)$ picks out the least value satisfying A , if there is one, with the following axiom:

$$(*)\ A(x) \rightarrow A(\varepsilon x A(x)) \wedge \varepsilon x A(x) \leq x$$

The result is a system that is strong enough to interpret first-order (Peano) arithmetic. Alternatively, one can take the epsilon symbol to satisfy the following axiom:

$$A(y) \rightarrow A(\varepsilon x A(x)) \wedge \varepsilon x A(x) \neq y + 1.$$

In other words, if there is any witness y satisfying $A(y)$, the epsilon term returns a value whose predecessor does not have the same property. Clearly the epsilon term described by (*) satisfies the alternative axiom; conversely, one can check that given A, a value of $\varepsilon x (\exists z \leq x A(x))$ satisfying the alternative axiom can be used to interpret $\varepsilon x A(x)$ in (*). One can further fix the meaning of the epsilon term with the axiom

$$\varepsilon x A(x) \neq 0 \rightarrow A(\varepsilon x A(x))$$

which requires that if there is no witness to A, the epsilon term return 0. For the discussion below, however, it is most convenient to focus on (*) alone.

Suppose we wish to show that the system above is consistent; in other words, we wish to show that there is no proof of the formula 0 = 1. By pushing all substitutions to the axioms and replacing free variables

by the constant 0, it suffices to show that there is no propositional proof of 0 = 1 from a finite set of closed instances of the axioms. For that, it suffices to show that, given any finite set of closed instances of axioms, one can assign numerical values to terms in such a way that all the axioms are true under the interpretation. Since the arithmetical operations + and $\times$ can be interpreted in the usual way, the only difficulty lies in finding appropriate values to assign to the epsilon terms.

Hilbert's *epsilon substitution method* can be described, roughly, as follows:

- Given a finite set of axioms, start by interpreting all epsilon terms as 0.
- Find an instance of the axiom (*) above that is false under the interpretation. This can only happen if one has a term t such that $A(t)$ is true in the interpretation, but either $A(\varepsilon x\, A(x))$ is false or the value of t is smaller than the value of $\varepsilon x\, A(x)$.
- "Repair" the assignment by assigning to $\varepsilon x\, A(x)$ the value of t, and repeat the process.

A finitistic consistency proof is obtained once it is shown in a finitistically acceptable manner that this process of successive "repairs" terminates. If it does, all critical formulas are true formulas without epsilon-terms.

This basic idea (the "Hilbertsche Ansatz") was set out first by Hilbert in his 1922 talk (1923), and elaborated in lectures in 1922-23. The examples given there, however, only deal with proofs in which all instances of the transfinite axiom correspond to a single epsilon term $\varepsilon x\, A(x)$. The challenge was to extend the approach to more than one epsilon term, to nested epsilon terms, and ultimately to second-order epsilons (in order to obtain a consistency proof not just of arithmetic, but of analysis).

The difficulty in dealing with nested epsilon terms can be described as follows. Suppose one of the axioms in the proof is the transfinite axiom

$$B(y) \to B(\varepsilon y\, B(y))$$

$\varepsilon y\, B(y)$ may, of course, occur in other formulae in the proof, in particular in other transfinite axioms, e.g.,

$$A(x, \varepsilon y\, B(y)) \to A(\varepsilon x\, A(x, \varepsilon y\, B(y)), \varepsilon y\, B(y))$$

So first, it seems necessary to find a correct interpretation for $\varepsilon y\, B(y)$ before we attempt to find one for $\varepsilon x\, A(x, \varepsilon y\, B(y))$. However, there are more complicated patterns in which epsilon terms may occur in a proof. An instance of the axiom, which plays a role in determining the correct interpretation for $\varepsilon y\, B(y)$ might be

$$B(\varepsilon x\, A(x, \varepsilon y\, B(y))) \to B(\varepsilon y\, B(y))$$

If $B(0)$ is false, then in the first round of the procedure $\varepsilon y\, B(y)$ will be interpreted by 0. A subsequent

change of the interpretation of $\varepsilon x\, A(x, 0)$ from 0 to, say, n, will result in an interpretation of this instance as $B(n) \rightarrow B(0)$ which will be false if $B(n)$ is true. So the interpretation of $\varepsilon y\, B(y)$ will have to be corrected to n, which, in turn, might result in the interpretation of $\varepsilon x\, A(x, \varepsilon y\, B(y))$ to no longer be a true formula.

This is just a sketch of the difficulties involved in extending Hilbert's idea to the general case. Ackermann (1924) provided such a generalization using a procedure which "backtracks" whenever a new interpretation at a given stage results in the need to correct an interpretation already found at a previous stage.

Ackermann's procedure applied to a system of second-order arithmetic, in which, however, second order terms were restricted so as to exclude cross-binding of second-order epsilons. This amounts, roughly, to a restriction to arithmetic comprehension as the set-forming principle available (see the discussion at the end of this section). Further difficulties with second-order epsilon terms surfaced, and it quickly became apparent that the proof as it stood was fallacious. However, no one in Hilbert's school realized the extent of the difficulty until 1930, when Gödel announced his incompleteness results. Until then, it was believed that the proof (at least with some modifications introduced by Ackermann, some of which involved ideas from von Neumann's (1927) version of the epsilon substitution method) would go through at least for the first-order part. Hilbert and Bernays (1939) suggest that the methods used only provides a consistency proof for first-order arithmetic with open induction. In 1936, Gerhard Gentzen succeeded in giving a proof of the consistency of first-order arithmetic in a formulation based on predicate logic without the epsilon symbol. This proof uses transfinite induction up to ε_0. Ackermann (1940) was later able to adapt Gentzen's ideas to give a correct consistency proof of first-order arithmetic using the epsilon-substitution method.

Even though Ackermann's attempts at a consistency proof for second-order arithmetic were unsuccessful, they provided a clearer understanding of the use of second-order epsilon terms in the formalization of mathematics. Ackermann used second-order epsilon terms $\varepsilon f\, A(f)$, where f is a function variable. In analogy with the first-order case, $\varepsilon f\, A(f)$ is a function for which $A(f)$ is true, e.g., $\varepsilon f\, (x + f(x) = 2x)$ is the identity function $f(x) = x$. Again in analogy with the first-order case, one can use second-order epsilons to interpret second-order quantifiers. In particular, for any second-order formula $A(x)$ one can find a term $t(x)$ such that

$$A(x) \leftrightarrow t(x) = 1$$

is derivable in the calculus (the formula A may have other free variables, in which case these appear in the term t as well). One can then use this fact to interpret *comprehension* principles. In a language with function symbols, these take the form

$$\exists f\, \forall x\, (A(x) \leftrightarrow f(x) = 1)$$

for an arbitrary formula $A(x)$. Comprehension is more commonly expressed in terms of set variables, in

which case it takes the form

$$\exists Y\, \forall x\, (A(x) \leftrightarrow x \in Y),$$

asserting that every second order formula, with parameters, defines a set.

Analysis, or *second-order arithmetic,* is the extension of first-order arithmetic with the comprehension schema for arbitrary second-order formulae. The theory is *impredicative* in that it allows one to define sets of natural numbers using quantifiers that range over the entire universe of sets, including, implicitly, the set being defined. One can obtain *predicative* fragments of this theory by restricting the type of formulae allowed in the comprehension axiom. For example, the restriction discussed in connection with Ackermann above corresponds to the *arithmetic comprehension schema,* in which formulae do not involve second-order quantifiers. There are various ways of obtaining stronger fragments of analysis that are nonetheless predicatively justified. For example, one obtains *ramified analysis* by associating an ordinal rank to set variables; roughly, in the definition of a set of a given rank, quantifiers range only over sets of lower rank, i.e., those whose definitions are logically prior.

More Recent Developments

In this section we discuss the development of the epsilon-substitution method for obtaining consistency results for strong systems; these results are of a mathematical nature. We cannot, unfortunately, discuss the details of the proofs here but would like to indicate that the epsilon-substitution method did not die with Hilbert's program, and that a significant amount of current research is carried out in epsilon-formalisms.

Gentzen’s consistency proofs for arithmetic launched a field of research known as *ordinal analysis,* and the program of measuring the strength of mathematical theories using ordinal notations is still pursued today. This is particularly relevant to the *extended Hilbert’s program,* where the goal is to justify classical mathematics relative to constructive, or quasi-constructive, systems. Gentzen’s methods of cut-elimination (and extensions to infinitary logic developed by Paul Lorentzen, Petr Novikov, and Kurt Schütte) have, in large part, supplanted epsilon substitution methods in these pursuits. But epsilon calculus methods provide an alternative approach, and there is still active research on ways to extend Hilbert-Ackermann methods to stronger theories. The general pattern remains the same:

1. Embed the theory under investigation in an appropriate epsilon calculus.
2. Describe a process for updating assignments to the epsilon terms.
3. Show that the procedure is normalizing, i.e., given any set of terms, there is a sequence of updates that results in an assignment that satisfies the axioms.

Since the last step guarantees the consistency of the original theory, from a foundational point of view one is interested in the methods used to prove normalization. For example, one obtains an *ordinal*

analysis by assigning ordinal notations to steps in the procedure, in such a way that the value of a notation decreases with each step.

In the 1960's, William Tait extended Ackermann's methods to obtain an ordinal analysis of extensions of arithmetic with principles of transfinite induction. More recently, Grigori Mints, Sergei Tupailo, and Wilfried Buchholz have considered stronger, yet still predicatively justifiable, fragments of analysis, including theories of arithmetic comprehension and a Δ^1_1-comprehension rule. Toshiyasu Arai has extended the epsilon substitution method to theories that allow one to iterate arithmetic comprehension along primitive recursive well orderings. In particular, his work yields ordinal analyses for predicative fragments of analysis involving transfinite hierarchies and transfinite induction.

As this article is being written, some first steps have been taken in using epsilon substitution method in the analysis of *impredicative* theories. See the annotated bibliography below.

A variation on step 3 above involves showing that the normalization procedure is not sensitive to the choice of updates, which is to say, *any* sequence of updates terminates. This is called *strong normalization.* Mints has shown that many of the procedures considered have this stronger property.

In addition to the traditional, foundational branch of proof theory, today there is a good deal of interest in *structural proof theory*, a branch of the subject that focuses on logical deductive calculi and their properties. This research is closely linked with issues relevant to computer science, having to do with automated deduction, functional programming, and computer aided verification. Here, too, Gentzen-style methods tend to dominate (see, e.g., Troelstra-Schwichtenberg (2000)). But the epsilon calculus can also provide valuable insights; cf. for example Mints (2002) or the discussion of Herbrand's theorem above.

Aside from the investigations of the epsilon calculus in proof theory, two applications should be mentioned. One is the use of epsilon notation in Bourbaki's *Theorie des ensembles* (1958). The second, of perhaps greater current interest, is the use of the epsilon-operator in the theorem-proving systems HOL and Isabelle, where the expressive power of epsilon-terms yields significant practical advantages.

Epsilon Operators in Linguistics, Philosophy, and Non-classical Logics

Reading the epsilon operator as an indefinite choice operator ("an x such that A(x)") suggests that it might be a useful tool in the analysis of indefinite and definite noun phrases in formal semantics. The epsilon notation has in fact been so used, and this application has proved useful in particular in dealing with anaphoric reference.

Consider the familiar example

1. Every farmer who owns a donkey beats it.

The generally accepted analysis of this sentence is given by the universal sentence

2. $\forall x\ \forall y\ (Farmer\ (x) \wedge Donkey(y) \wedge Owns(x, y)) \rightarrow Beats(x, y))$

The drawback is that "a donkey" suggest an existential quantifier, and thus the analysis should, somehow, parallel in form the analysis of sentence 3 given by 4:

3. Every farmer who owns a donkey is happy,
4. $\forall x\ (Farmer(x) \wedge \exists y\ (Donkey(y) \wedge Owns(x, y)) \rightarrow Happy(x))$,

but the closest possible formalization,

5. $\forall\ x\ ((Farmer(x) \wedge \exists y\ (Donkey(y) \wedge Owns(x, y)) \rightarrow Beats(x, y))$

contains a free occurrence of y. Evans suggests that since pronouns are referring expressions, they should be analyzed as definite descriptions; and if the pronoun occurs in the consequent of a conditional, the descriptive conditions are determined by the antecedent. This leads to the following E-type analysis of (1):

$$\forall x\ ((Farmer(x) \wedge \exists y\ (Donkey(y) \wedge Owns(x, y)) \rightarrow Beats(x, \iota y\ (\ Donkey(y) \wedge Owns(x, y)))$$

(ιx is the definite description operator). The trouble with this is that on the standard analysis, the definite description carries a uniqueness condition, and so (5) will be false if there is a farmer who owns more than one donkey. A way out of this is to introduce a new operator, whe (whoever, whatever) which works as a generalizing quantor (Neale, 1991):

$$\forall x\ ((Farmer(x) \wedge \exists y\ (Donkey(y) \wedge Owns(x, y)) \rightarrow Beats(x, \text{whe}\ y\ (Donkey(y) \wedge Owns(x, y)))$$

As pointed out by von Heusinger (1994), this suggests that Neale is committed to pronouns being ambiguous between definite descriptions (ι-expressions) and whe-expressions. Heusinger suggests instead to use epsilon operators indexed by choice functions (which depend on the context). According to this approach, the analysis of (1) is

For every choice function i:
$\forall x\ (Farmer(x) \wedge Owns(x, \varepsilon_i y\ Donkey\ (y)) \rightarrow Beats\ (x, \varepsilon_{a*}\ y\ Donkey\ (\text{y}))$

$a*$ is a choice function which depends on i and the antecedent of the conditional: If i is a choice function which selects $\varepsilon_i y\ Donkey\ (y)$ from the set of all donkeys, then $\varepsilon_{a*}\ y\ Donkey\ (y)$ selects from the set of

donkeys owned by x.

This approach to dealing with pronouns using epsilon operators indexed by choice functions enable von Heusinger to deal with a wide variety of circumstances (see Egli and von Heusinger, 1995; von Heusinger, 2000).

Applications of the epsilon-operator in formal semantics, and choice functions in general, have received significant interest in recent years. Von Heusinger and Egli (2000a) list, among others, the following: representations of questions (Reinhart, 1992), specific indefinites (Reinhart 1992; 1997; Winter 1997), E-type pronouns (Hintikka and Kulas 1985; Slater 1986; Chierchia 1992, Egli and von Heusinger 1995) and definite noun phrases (von Heusinger, 1997).

For discussion of the issues and applications of the epsilon operator in linguistics and philosophy of language, see B. H. Slater's article on epsilon calculi (cited in the Other Internet Resources section below), and the collections (available online) edited by von Heusinger *et al.*, listed in the Bibliography.

Another application of epsilon calculus is as a general logic for reasoning about arbitrary objects. Meyer Viol (1995) provides a comparison of the epsilon calculus with Fine's (1985) theory of arbitrary objects. Indeed, the connection is not hard to see. Given the equivalence $\forall x\, A(x) \equiv \mathrm{A}(\varepsilon x\, (\neg A))$, the term $\varepsilon x\, (\neg A)$ is an arbitrary object in the sense that it is an object of which A is true iff A is true generally.

Meyer Viol (1995, 1995a) contain further proof- and model-theoretic studies of the epsilon calculus; specifically intuitionistic epsilon calculi. Here, the epsilon theorems no longer hold, i.e., introduction of epsilon terms produces non-conservative extensions of intuitionistic logic. In fact, as was shown by Bell (1993a, 1993b; Meyer Viol, 1995), addition of the epsilon operator to intuitionistic predicate logic allows us to prove the intuitionistically non-valid principles $\neg A \vee \neg\neg A$ and $(\mathrm{A} \to \mathrm{B}) \vee (\mathrm{A} \to \mathrm{B})$. The full principle of the excluded middle can only be proved if we also add epsilon extensionality.

$$\forall x\, (A(x) \leftrightarrow B(x)) \to \varepsilon x\, A = \varepsilon x\, B$$

This result provides a rigorous justification of Hilbert's original conjecture that the principle of the excluded middle is, in a sense, a special case of the axiom of choice, and that only with epsilon extensionality do we get the full strength of the choice principle.

Other model-theoretic investigations of epsilon operators in intuitionistic logic can be found in DeVidi (1995). For epsilon-operators in many-valued logics, see Mostowski (1963), for modal epsilon calculus, Fitting (1975).

Bibliography

The following list of references provides a starting point for exploring the literature, but is by no means

comprehensive.

Hilbert's Program

The following source books have many of the original papers:

- Bennacerraf, P., Putnam, H. (eds.), 1983, *Philosophy of Mathematics* , 2nd ed., Cambridge: Cambridge University Press
- Ewald, W. B. (ed.), 1996, *From Kant to Hilbert. A Source Book in the Foundations of Mathematics*, Vol. 2, Oxford: Oxford University Press
- Mancosu, P. (ed.), 1998, *From Brouwer to Hilbert. The Debate on the Foundations of Mathematics in the 1920s*, Oxford: Oxford University Press
- van Heijenoort, J. (ed.), 1967, *From Frege to Gödel. A Source Book in Mathematical Logic*. Cambridge, Mass.: Harvard University Press

Overviews of the historical development of logic and proof theory in the Hilbert school can be found in

- Avigad J. and Reck, E., 2001, 'Clarifying the nature of the infinite": the development of metamathematics and proof theory', Carnegie Mellon University Technical Report CMU-PHIL-120 [Available online in PDF]
- Hallett, M., 1995, 'Hilbert and logic', M. Marion and R. S. Cohen, Quebec Studies in the Philosophy of Science, Vol. 1, Dordrecht: Kluwer, 135-187
- Mancosu, P., 1998a, 'Hilbert and Bernays on metamathematics', in Mancosu, 1998, 149-188
- Moore, G. H., 1997, 'Hilbert and the emergence of modern mathematical logic', *Theoria* (Segunda Epoca), 12:65-90
- Peckhaus, V., 1990, *Hilbertprogramm und Kritische Philosophie*, Göttingen: Vandenhoek & Ruprecht
- Sieg, W., 1988, 'Hilbert's program sixty years later', *Journal of Symbolic Logic*, 53: 338-348
- Sieg, W., 1990, 'Reflections on Hilbert's program', W. Sieg (ed.), *Acting and Reflecting*, Dordrecht: Kluwer
- Sieg, W., 1999, 'Hilbert's Programs: 1917-1922', *Bulletin of Symbolic Logic*, 5: 1-44 [Available online in Postscript]
- Zach, R., 1999, 'Completeness before Post: Bernays, Hilbert, and the development of propositional logic', *Bulletin of Symbolic Logic*, 5: 331--366 [Available online in Postscript]
- Zach, R., 2002, 'The practice of finitism. Epsilon calculus and consistency proofs in Hilbert's Program', *Synthese* (forthcoming). [Preprint available online]

The Early History of the Epsilon Calculus and Epsilon Substitution Method

The original work:

- Ackermann, W., 1924, ‘Begründung des ’’tertium non datur’’ mittels der Hilbertschen Theorie der Widerspruchsfreiheit’, *Mathematische Annalen*, 93:1-36
- Ackermann, W., 1937-38, ‘Mengentheoretische Begründung der Logik’, *Mathematische Annalen,* 115:1-22
- Ackermann, W., 1940, ‘Zur Widerspruchsfreiheit der Zahlentheorie’, *Mathematische Annalen*, 117:162-194
- Hilbert, D., 1922, ‘Neubegründung der Mathematik: Erste Mitteilung’, *Abhandlungen aus dem Seminar der Hamburgischen Universität* , 1:157-177, English translation in Mancosu, 1998, 198-214 and Ewald, 1996, 1115-1134
- Hilbert, D., ‘Die logischen Grundlagen der Mathematik’, *Mathematische Annalen*, 88:151-165, English translation in Ewald, 1996, 1134--1148
- Hilbert, D., Bernays, P., 1934, *Grundlagen der Mathematik* , Vol. 1, Berlin: Springer
- Hilbert, D., Bernays, P., 1939, *Grundlagen der Mathematik* , Vol. 2, Berlin: Springer
- von Neumann, J., 1927, ‘Zur Hilbertschen Beweistheorie’, *Mathematische Zeitschrift*, 26:1-46

Ackermann’s 1940 proof is discussed in

- Hilbert, D., Bernays, P., 1970, ‘Grundlagen der Mathematik’, Vol. 2, 2nd, edition, Berlin: Springer, Supplement V
- Wang, H., 1963, *A Survey of Mathematical Logic,* Peking: Science Press

Maehara showed how to prove the second epsilon theorem using cut elimination, and then strengthened the theorem to include the schema of extensionality, in

- Maehara, S., 1955, ‘The predicate calculus with ε-symbol’, *Journal of the Mathematical Society of Japan*, 7:323-344
- Maehara, S., 1957, ‘Equality axiom on Hilbert’s ε-symbol’, *Journal of the Faculty of Science,* University of Tokyo, Section 1, 7:419-435

An early application of epsilon substitution is Georg Kreisel’s *no-counterexample interpretation*.

- Kreisel, G, 1951, ‘On the interpretation of non-finitist proofs - part I’, *Journal of Symbolic Logic*, 16:241-267

The following provide modern presentations of Hilbert’s epsilon calculus, not just from an introductory standpoint:

- Leisenring, A. C., 1969, *Mathematical Logic and Hilbert’s Epsilon-Symbol*, London: Macdonald
- Mints, G., 1996, ‘Thoralf Skolem and the epsilon substitution method for predicate logic’, *Nordic Journal of Philosophical Logic*, 1:133-146 [Available online]
- Moser, G., 2000, *The Epsilon Substitution Method*, Master’s Thesis, University of Leeds

Corrections to errors in the literature (including Leisenring's book) can be found in

- Flannagan, T. B., 1975, 'On an extension of Hilbert's second ε-theorem', *Journal of Symbolic Logic*, 40:393-397
- Ferrari, P. L., 1987, 'A note on a proof of Hilbert's second ε-theorem', *Journal of Symbolic Logic*, 52:214-215
- Yashahura, M., 1982, 'Cut elimination in ε-calculi', *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 28:311-316

A variation of the epsilon calculus based on Skolem functions, and therefore compatible with first-order logic, is discussed in

- Davis, M., and R. Fechter, 1991, 'A free variable version of the first-order predicate calculus', *Journal of Logic and Computation*, 1:431-451

General References for Proof Theory

- Buss, S. (ed.), 1998. *The Handbook of Proof Theory,* Amsterdam: North-Holland
- Takeuti, G., 1987, *Proof Theory*, second edition. Amsterdam: North-Holland, Amsterdam
- Troelstra, A. S., Schwichtenberg, H., 2000, *Basic Proof Theory*, second edition. Cambridge: Cambridge University Press

The following contains a number of proof-theoretic results that are proved using methods similar to the ones used by Hilbert, Bernays, and Ackermann, though using Skolem functions instead of epsilon terms:

- Shoenfield, J., 1967, *Mathematical Logic*, Reading, Mass.:Addison-Wesley, republished by the Association for Symbolic Logic, 2001

For more on ordinal analysis, see, for example:

- Takeuti, G., *Proof Theory* (see above)
- Pohlers, Wolfram, 1998, 'Subsystems of set theory and second-order number theory', in S. Buss (ed.), *The Handbook of Proof Theory* (see above), 209-335

Herbrand's Theorem

Herbrand's theorem originally appeared in

- Herbrand, J., 1930, *Recherches sur la thèorie de la dèmonstration*, *Dissertation*, University of Paris

English translations can be found in van Heijenoort (see above), and

- Herbrand, J., 1971, *Collected Works*. W. Goldfarb (ed.), Cambridge, Mass.: Harvard University Press

Further historical information can be found in

- Dreben, B., Andrews, P., Aanderaa, S., 1963, 'False lemmas in Herbrand', *Bulletin of the American Mathematical Society*, 69:699-706

The literature on Herbrand's theorem is vast. For some general overviews, in addition to the general proof-theoretic references above, see

- Buss, S., 1995, 'On Herbrand's theorem', *Logic and Computational Complexity* (Indianapolis, IN, 1994), Lecture Notes in Computer Science 960, Berlin: Springer, 195-209 [Available online in PDF]
- Girard, J.-Y., 1982, 'Herbrand's theorem and proof theory', *Proceedings of the Herbrand Symposium*, Amsterdam: North-Holland, 29-38
- Statman, R., 1979, 'Lower bounds on Herbrand's theorem', *Proceedings of the American Mathematical Society*, 75:104-107
- Voronkov, A., 1999, 'Simultaneous rigid E-unification and other decision problems related to the Herbrand theorem', *Theoretical Computer Science*, 224:319-352

A striking application of Herbrand's theorem and related methods is found in Luckhardt's analysis of Roth's theorem:

- Luckhardt, H., 1989, 'Herbrand-Analysen zweier Beweise des Satzes von Roth: Polynomiale Anzahlschranken', *Journal of Symbolic Logic*, 54:234-263

For a discussion of useful extensions of Herbrand's methods, see

- Sieg, W., 1991, 'Herbrand analyses', *Archive for Mathematical Logic*, 30:409-441

A model-theoretic version of this is discussed in

- Avigad, J., 2002, 'Saturated models of universal theories', to appear in the *Annals of Pure and Applied Logic*

More Recent Developments in the Epsilon Substitution Method

In the following two papers, William Tait analyzed the epsilon substitution method in terms of continuity considerations:

- Tait, W. W., 1960, ‘The substitution method.’ *Journal of Symbolic Logic* , 30:175-192.
- Tait, W. W., 1965, ‘Functionals defined by transfinite recursion,’ *Journal of Symbolic Logic* 30:155-174.

More streamlined and modern versions of this approach can be found in:

- Avigad, J., 2002, ‘Update procedures and the 1-consistency of arithmetic’, *Mathematical Logic Quarterly*, 48:3-13.
- Mints, G., 2001, ‘The epsilon substitution method and continuity’, in W. Sieg *et al*. (eds.), *Reflections on the Foundations of Mathematics: Essays in Honor of Solomon Feferman,* Lecture Notes in Logic 15, Association for Symbolic Logic

The following paper shows that the epsilon substitution method for first-order arithmetic is, in fact, strongly normalizing:

- Mints, G., 1996, ‘Strong termination for the epsilon substitution method’, *Journal of Symbolic Logic*, 61:1193-1205

A connection between cut elimination and epsilon substitution method is explored in

- Mints, G., 1994, ‘Gentzen-type systems and Hilbert’s epsilon substitution method. I’, *Logic, Methodology and Philosophy of Science, IX* (Uppsala, 1991), Amsterdam: North-Holland, 91-122

The epsilon substitution method has been extended to predicative fragments of second-order arithmetic in:

- Mints, G., Tupailo, S., Buchholz, W., 1996, ‘Epsilon substitution method for elementary analysis’, *Archive for Mathematical Logic*, 35:103-130
- Mints, G., Tupailo, S., 1999, ‘Epsilon-substitution method for the ramified language and Δ^1_1-comprehension rule’, in A. Cantini *et al*. (eds.), *Logic and Foundations of Mathematics* (Florence, 1995), Dordrecht: Kluwer, 107-130
- Arai, T., 2002, ‘Epsilon substitution method for theories of jump hierarchies’, *Archive for Mathematical Logic*, 2:123-153

The following papers address impredicative theories:

- Arai, T., 2001, ‘Epsilon substitution method for ID1($\Pi^0_1 \vee \Sigma^0_1$)’, preprint
- Mints, G., 2001, ‘An approach to an epsilon-substitution method for ID1’, preprint, Institute

Mittag Leffler, 45, MLI, Stockholm

A development of set theory based on the epsilon-calculus is given by

- Bourbaki, N., 1958, *Theorie des ensembles* , Paris: Hermann

Epsilon Operators in Linguistics, Philosophy, and Non-classical Logics

The following is a list of some publications in the area of language and linguistics of relevance to the epsilon calculus and its applications. The reader is directed in particular to the collections von Heusinger and Egli (2000) and von Heusinger *et al*. (2002) for further discussion and references.

- Bell, J. L., 1993a. 'Hilbert's epsilon-operator and classical logic', *Journal of Philosophical Logic*, 22:1-18
- Bell, J. L., 1993b. 'Hilbert's epsilon operator in intuitionistic type theories', *Mathematical Logic Quarterly* 39:323-337
- Chierchia, G., 1992. 'Anaphora and dynamic logic'. *Linguistics and Philosophy*, 15:111-183
- DeVidi, D., 1995. 'Intuitionistic epsilon- and tau-calculi', *Mathematical Logic Quarterly* 41:523--546
- Evans, G., 1980, 'Pronouns', *Linguistic Inquiry* , 11:337-362
- Egli, U., von Heusinger, K., 1995, 'The epsilon operator and E-type pronouns', in U. Egli *et al*. (eds.), *Lexical Knowledge in the Organization of Language,* Amsterdam: Benjamins, 121-141 (Current Issues in Linguistic Theory 114) [Preprint available online]
- Fine, K., 1985. *Reasoning with Arbitrary Objects*. Oxford: Blackwell.
- Fitting, M., 1975. 'A modal logic epsilon-calculus', *Notre Dame Journal of Formal Logic*, 16:1--16
- Hintikka, J., Kulas, J., 1985. *Anaphora and Definite Descriptions: Two Applications of Game-Theoretical Semantics*. Dordrecht: Reidel
- Kempson, R., Meyer Viol, W., and Gabbay, D., 2001. *Dynamic Syntax: The Flow of Language Understanding*. Oxford: Blackwell
- Meyer Viol, W. P. M., 1995, *Instantial Logic*. *An Investigation into Reasoning with Instances* . Ph.D. thesis, University of Utrecht. ILLC Dissertation Series 1995-11
- Meyer Viol, W., 1995a. 'A proof-theoretic treatment of assignments', *Bulletin of the IGPL*, 3:223-243 [Available online]
- Mostowski, A., 1963. 'The Hilbert epsilon function in many-valued logics', *Acta Philosophica Fennica*, 16:169-188
- Reinhart, T., 1992. 'Wh-in-situ: An apparent paradox'. In: P. Dekker and M. Stokhof (eds.). *Proceedings of the Eighth Amsterdam Colloquium* December 17-20, 1991. ILLC. University of Amsterdam, 483-491
- Reinhart, T., 1997. 'Quantifier scope: How labor is divided between QR and choice functions'. *Linguistics and Philosophy*, 20:335-397

- Slater, B. H. 1986, ‘E-type pronouns and ε-terms’, *Canadian Journal of Philosophy*, 16:27-38
- Slater, B. H. 1988, ‘Hilbertian reference’, *Noûs*, 22:283-97
- Slater, B. H. 1994, ‘The epsilon calculus’ problematic’, *Philosophical Papers*, 23:217-42
- Slater, B.H. 2000, ‘Quantifier/variable-binding’, *Linguistics and Philosophy*, 23:309-21
- von Heusinger, K., 1997. ‘Definite descriptions and choice functions’. In: S. Akama (ed.). *Logic, Language and Computation*. Dordrecht: Kluwer, 61-91 [Preprint available online]
- von Heusinger, K., 2000, ‘The Reference of Indefinites’, in von Heusinger and Egli, (2000), 265-284 [Preprint available online]
- von Heusinger, K., Egli, U., (eds.), 2000. *Reference and Anaphoric Relations*. Dordrecht: Kluwer, 265-284 [Preprints available online]
- von Heusinger, K., Egli, U., 2000a. ‘ Introduction: Reference and the Semantics of Anaphora’, in von Heusinger and Egli (2000), 1-13
- von Heusinger, K., Kempson, R., Meyer-Viol, W., (eds.), 2002. *Proceedings of the Workshop Choice Function and Natural Language Semantics*, Arbeitspapier 110. Fachbereich Sprachwissenschaft, Universität Konstanz [Preprints available online]
- Winter, Y., 1997. ‘Choice functions and the scopal semantics of indefinites’. *Linguistics and Philosophy*, 20:399-467

Other Internet Resources

- Epsilon Calculi by B. Hartley Slater (Internet Encyclopedia of Philosophy)

Please contact the authors with further suggestions.

Related Entries

finitism | Hilbert, David | Hilbert’s Program | logic: classical | mathematics, philosophy of | proof theory | quantification

Jeremy Avigad
avigad@cmu.edu
and
Richard Zach
rzach@ucalgary.ca

Table of Contents

First published: May 3, 2002
Content last modified: May 3, 2002

Philosophy for Children

Philosophy typically makes its formal entry into the curriculum at the college level. A growing number of high schools offer some introduction to philosophy, often in special literature courses for college bound students. This suggests that serious philosophical thinking is not for pre-adolescents. Two reasons might be offered for accepting this view. First, philosophical thinking requires a level of cognitive development that, one may believe, is beyond the reach of pre-adolescents. Second, the school curriculum is already crowded; and introducing a subject like philosophy will not only distract students from what they need to learn, it may encourage them to become skeptics rather than learners. However, there are grounds for challenging both of these reasons for resisting philosophy for children. They will be addressed in turn.

---

1. Are Children Capable of Philosophical Thinking?

Jean Piaget's (1933) well-known theory of cognitive development suggests that prior to age 11 or 12, most children are not capable of philosophical thinking. This is because, prior to this time, children are not capable of “thinking about thinking,” the sort of meta-level thinking that characterizes philosophical thinking. This “formal operational” level of cognitive development includes analogical reasoning about relationships, such as: “Bicycle is to handlebars as ship is to rudder, with “steering mechanism” being the similar relationship.” (Goswami, p. xxi). However, there is a growing body of psychological research suggesting that Piaget's account seriously underestimates children's cognitive abilities. (Astington, 1993; Gopnik, et.al.)

Philosopher Gareth Matthews goes further and argues at length that Piaget failed to see the philosophical thinking manifest in the very children he studied. Matthews (1980) provides a number of delightful examples of very young children's philosophical puzzlement. For example:

- TIM (about six years), while busily engaged in licking a pot, asked, “Papa, how can we be sure that everything is not a dream?” (P. 1)
- JORDAN (five years), going to bed at eight one evening, asked, “If I go to bed at eight and get up at seven in the morning, how do I really know that the little hand of the clock has gone around only once? Do I have to stay up all night to watch it? If I look away even for a short time, maybe the small hand will go around twice.” (P. 3)
- ONE DAY John Edgar (four years), who had seen airplanes take off, rise, and gradually disappear into the distance, took his first plane ride. When the plane stopped ascending and the seat-belt sign went out, John Edgar turned to his father and said in a rather relieved but still puzzled, tone of voice, “Things don't really get smaller up here.” (P. 4)

Matthews acquired many of his anecdotes from friends who knew of his interest in the philosophical thinking of children. It is not uncommon for attentive adults to encounter such examples.

However, it might be objected that more than such anecdotes are needed to show that children are capable of serious philosophical thinking. What is needed is evidence that children are capable of *sustained* philosophical discussion. Matthews (1984) provides illustrations of this, too. Meeting with a group of 8-11 year olds, he used the following example to develop a story for discussion:

> Ian (six year old) found to his chagrin that the three children of his parents' friends monopolized the television; they kept him from watching his favorite program. “Mother,” he asked in frustration, “why is it better for three children to be selfish than one?” (Pp. 92-3)

This generated a lively discussion in which children commented on the inconsiderateness of the three visiting children, the desirability of working out a solution that would satisfy all four children, the importance of respecting people's rights, and how one might feel if he or she were in Ian's place. Matthews then posed a possible utilitarian approach: “What about this argument, that if we let the three visitors have their way, three people will be made happy instead of just one?” One reply was that it would not be fair for three people to get what they want at the expense of a fourth. This triggered a discussion of fairness that addressed more specific concerns about the relative ages of the children, whether they are friends, siblings, or strangers—and what types of television programs are involved.

No doubt, part of the explanation of the children’s ability and willingness to carry on an extended discussion of Ian's circumstance is that they have faced similar challenges. Still, the children exhibited a rather sophisticated conceptual grasp of the issues at hand, which is what one might expect from children once they are invited to reflect on their own experiences.

Stories about those roughly their own age can provide opportunities for children to discuss ideas that are most important to them. Consider this example from Matthew Lipman's novel *Lisa* (1983). Harry and his friend Timmy go to a stamp club to trade stamps. Afterward they stop for ice cream cones, but Timmy discovers he has no money. Harry offers to buy him one, and Timmy says he will buy Harry a cone next time. As they are leaving the store, one of their classmates trips Timmy. Timmy then knocks the tripper's books off the table. After running away from the scene, Timmy and Harry talk about what has happened:

> “I couldn't let them get away with it,” Timmy remarked when they saw that they weren't being pursued and could slow down to a walk. “He didn't have to stick his foot out.” Then he added, “Of course, I didn't have to do what I did either. But, like I said before, turnabout is fair play.”
>
> “Somehow," Harry thought, “it isn't quite the same thing.” But he couldn't figure out why. “I don't know,” he said finally to Timmy. “The purpose of your stamp club is to exchange stamps. So when you give someone stamps, you're supposed to give something back. Just like if someone lends me money, I'm supposed to give it back. But if someone pulls a dirty trick on you, should you do the same thing to him? I'm not so sure.”
>
> “But I had to get even,” Timmy protested. “I couldn't let him get away with it, tripping me like that for no reason.”
>
> A bit later they met Lisa and Laura. Harry told the girls what had happened and why he was puzzled. “It reminds me,” remarked Lisa, “of last year when we were learning about how some sentences could be turned around and would stay true, while others, when you turned them around, would become false.”
>
> “Yeah,” Harry agreed, “but there we found a rule. What's the rule here?” Lisa tossed her long hair so that it hung over her right shoulder. “It looks like there are times when it is right to give back what we got and other times when it is wrong. But how do we tell which is which?” (*Lisa*, pp. 23-4)

This passage is an invitation to explore the moral nuances of reciprocity, or “returning in kind.” What might a group of 10-11 year-olds have to say about this? Here is a sampling from a 30 minute discussion of the *Lisa* passage by a group of fifth graders. (Pritchard 1996) Although the group had been discussing philosophical ideas once a week after school for the past several months, this was an impromptu discussion of this passage. With little prompting from the teacher, the students raised and vigorously pursued the following questions:

- What is likely to happen when we retaliate? Will this simply start a long chain of trying to “get things even” that no one (other than perhaps the initiator) wants?
- Can retaliating really “get things even”? Can we even make sense of “getting things even”?

- Is it really right to respond to a wrong by returning in kind? Is this trying to make a right out of two wrongs?
- What alternatives to trying to "get things even" might there be? What would happen if you just ignore someone who is trying to get a rise out of you?
- When is self-defense the best strategy?
- Is there a difference between trying to "get things even" and trying to "teach someone a lesson"?
- How is hitting back different from: a) exchanging goods; b) paying back a debt; c) returning a favor; d) offering a favor; e) not doing someone a favor unless they do one for you?
- Can the Golden Rule help us here? What does the Golden Rule *mean*? Is it a *good* rule?

Thoughtful and insightful discussions like this are not unusual for children who are given the opportunity to have them. This discussion was prompted by a children's novel. However, students' regular classroom materials, works of art, thought experiments, or even the daily newspaper can be used to trigger philosophical discussions of moral concerns.

Even if one concedes that children are quite capable of engaging in extended discussions of *moral* concepts related to their own experiences, what about philosophical ideas less related to their practical affairs? Here is an illustration that begins with logic and ends up in metaphysics.

The true sentence, "All oaks are trees" becomes false when reversed. So does, "All carrots are vegetables." Can we say that *every* true sentence beginning with 'all' becomes false when reversed?

At least as early as the 3rd grade, children easily find exceptions. What about "All tigers are tigers", many will ask. Others may respond that this is a "boring" sentence, offering something like "All rabbits are hares," or "All mothers have children" as alternatives. With relatively little encouragement, they can come up with good definitions of geometric figures and differentiate them from proposed definitions that cannot be reversed. For example, "All squares are rectangles" is true, but it becomes false when reversed. "All squares are rectangles with equal sides," however, can be reversed.

Although the study of logic is traditionally regarded as a part of philosophy, skeptics might not find the reflections of children on rules of logic terribly interesting philosophically. (Of course, some might say this about elementary logic in the college classroom, as well.) It is not that different from basic math and grammar, they might object. Whether or not this is a fair assessment, for many children it is but a short step from logic to metaphysics. Here is an example from a 4th grade class that had just been asked whether true sentences beginning with 'all' always become false when reversed. (Pritchard 1996) reversed. After the usual "All tigers are tigers" and "All rabbits are hares" were suggested, a student asked, "How about 'All answers have questions' and 'All questions have answers'?" Fortunately, the teacher paused to explore this with the class. "Do all answers have questions?" he asked. Of course, replied the students, otherwise we would not say that we have an *answer*.

The teacher continued, "How about the other sentence? Do you think that all questions have answers?" What followed was a flood of responses:

- Student #1: “Is there life in the center of the sun?”
- Student #2: “Even though we can’t go there to find out, the question still has an answer.”
- Student #3: “How many grains of sand are there on earth?”
- Student #4: “There’s definite number even though we don’t know what it is.”
- Student #3: “The wind will blow them all around, and we’ll count some more than once.”
- Student #5: “There are too many to count.”
- Student #6: “How many grains of sand are there on all the planets?”
- Student #7: “How many trees are there on earth?”
- Student #4: “That’s easier than grains of sand. We could count them.”
- Student #7: “By the time you finish counting them, some would have fallen down and others would have started to grow.”
- Student #8: “Did God make time begin?”
- Student #9: “You mean, *if* there is a God, did he make time begin?”
- Student #7: “Does space have limits?”
- Student #5: “Yeah, what would happen if you got to the end of space and tried to put your hand out? If you couldn’t, what would be holding it back on the outside?”
- Student #6: &#!47;Maybe what would hold your hand back is on the inside. There wouldn’t be any outside.”

During the course of discussing these questions, the students seemed to be struggling to move from questions that are difficult, if not impossible, to answer because of our *practical* limitations (e.g., not being certain that a particular grain of sand has not already been counted) to questions that *in principle* are unanswerable. Finally, with a mischievous grin on his face, one of the students asked, “Will time end?” The problem, he explained, is that if time did end, no one would be able later to confirm that this was so.

Here is another illustration of how quickly a discussion of logic can move to a discussion of deep philosophical issues. (Pritchard, 1985) This is a group of 5th graders considering the sentence, “All people are animals.” One of the students offered this as another example of a true sentence that becomes false when reversed. Jeff objected that “All people are animals” is not true. Chip proceeded to develop a taxonomy that relegated people, along with elephants and tigers, under the heading of mammals, mammals under animals, and animals under living things. Jeff continued to object.

> Chip: “Jeff, what are people? Just tell me, what are people? You can’t answer that, can you?”
>
> Jeff: “Yes, I can.”
>
> Chip: “What are you?”
>
> Jeff: “A person.”

Chip: “What’s a person?”

Jeff: “A living somebody.”

Chip: “A living somebody could be a whale.”

Jeff: “I said, *somebody*, not an animal….”

Chip: “You can check every single book out there in the library—well, every one that’s about us….”

Larry: “I want to know why everyone’s getting so huffy about a little subject.”

Rich: “We’re *thinking*! That’s what we’re here for.”

Amy: “Does anyone have an encyclopedia in here so we can look up either animals, mammals, or persons?”

Jeff: “We’re all humans. So, if this Mars guy saw us, he would say, ‘Hey, look, there’s some human beings.’ He wouldn’t say, “Hey, look, there’s some animals down there.’”

Mike: “Martians, if there are any, would say, ‘Hey, look at those weird looking creatures,’ or something like that. They wouldn’t know *what* we are. They don’t know anything about us. [Returning now to Jeff’s original distinction, Mike continues.] If it’s a person, you say, ‘some*body*.’ If it’s an animal you say, ‘some*thing*.’ Somebody is a human body.”

Chip: “There’s living life, okay? Then you branch off from there. You have animals plants, and whatever the other stuff is—you know, molecules and things like that. Now you go to the animals and you branch off—mammals, amphibians, reptiles, and whatever there is. Then you branch off and you have all these special humans. Is that right so far, Jeff?”

Jeff: “Just go on.”

Chip: “Well, I just want to know if you agree so far.”

Jeff: “Just go on. Go on. I’m not going to change my mind. That’s all…. I’m not an animal. I’m a person, and I’m going to stay that way.”

Chip: “You’re a *type* of animal.”

Jeff: “I’m not going to walk up to Dr. Jekyl and say, ‘Hey, change me into an animal’….”

> Amy: “People are a type of animal, like a bird is. That’s different than like an elephant is. A bird’s different than an elephant. And we’re different than a bird. Mike says we don’t call our dog a person or somebody. But someone might be real close to their pet and consider it part of the family.”

The discussion continued for several more minutes. As the group dispersed, one student remarked to another, “If we want to, we could argue for hours!” “For days,” replied the second. Meeting weekly after school in the local public library, this group of children returned the next week with an encyclopedia to settle the matter. After several minutes of discussion, the teacher asked the students if they thought everything in the encyclopedia is true.

> Emily: “Some things we’re not sure of; and the encyclopedia could put down every word about how the solar system was formed, and it would probably say there was big dust that spun around like a top. But we’re not sure about that. And, so, that could be wrong.”

The teacher asked whether, in such cases, the encyclopedia says, “We’re not sure?”

> Mike: “It’ll say ‘hypothesis’—which is a guess.”

> Kurt: “It’ll say we’re not sure yet.”

So, the discussion retained its philosophical vitality. This particular group continued to meet for the entire school year, discussing a wide range of philosophical topics, including: the relationship between the mind and the brain, differences (and similarities) between dreams and reality, knowledge of other minds, self-knowledge, and relationships between evidence and knowledge.

2. Philosophy in a Crowded Curriculum

Given an already crowded curriculum and growing pressure to provide quantifiable evidence of student mastery of the standard subjects of history, literature, math, and science, teachers may question the suitability of adding philosophy to the curriculum. Where is time to be found for the give-and-take of philosophical discussions? Adding philosophy to the mix, they might object, only makes matters worse. Not only is it yet another subject, it is one that is unfamiliar to most teachers, and they may fear that bringing in philosophy, with its continual questioning, will actually interfere with students’ mastery of the subjects already in the curriculum. Given the unsettling nature of much philosophical inquiry, they may feel vulnerable as teachers because they are not confident of their own answers to the questions posed.

Adding to this problem is increasing pressure on teachers to demonstrate that their students are performing at satisfactory levels in the standard subjects. Standardized tests are commonly used as the

measure of student achievement. Marked by definitive, unambiguous questions and answers, these tests do not place a premium on philosophical reflection. Since student performance is typically linked to school funding, this is not something teachers can take lightly, however skeptical they might be about the educational value of preparing their students to perform well on standardized tests.

In response, Matthew Lipman (1991) and others who advocate bring philosophy into the schools emphasize ways in which philosophy can enhance the entire educational experience of students. The aim is more than simply the introduction of one more subject in the schools. By inviting students to reflect on relationships among different areas of inquiry and to make sense of their educational experiences as a whole, philosophy can add to the meaningfulness of students' education as a whole. In addition, philosophy can make important contributions to another area of concern that cuts across the curriculum, critical thinking.

As the Vietnam War escalated in the mid-1960's, so did heated arguments about the wisdom and morality of the war and society's ills in general. Matthew Lipman became dismayed at the quality of argumentation employed by presumably well-educated citizens. Convinced that the teaching of logic should begin long before college, he tried to figure out a way to do this that would stimulate the interest of 10-11 year olds. Leaving Columbia University for Montclair State College, he launched his efforts with his first children's novel, *Harry Stottlemeier's Discovery*(1974). Lipman's concerns about the level of critical thinking in society in general, and the schools in particular, were not his alone. By the 1970's the hue and cry for teaching critical thinking in the schools was, if not clear, at least loud; and it has continued largely unabated to the present.

What is meant by 'critical thinking'? Characterizations range in complexity from Robert Ennis's admirably brief, "reasonable reflective thinking that is focused on deciding what to believe or do" (Ennis) to a complex statement by group of 46 panelists convened by the American Philosophical Association's Committee on Pre-College Philosophy to employ the Delphi Method of striving for consensus:

> We understand critical thinking to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based…. The ideal critical thinking is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgments, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as the subject and the circumstances of inquiry permit. (Facione 1989)

Lipman was a member of this panel, and it is clear that his novels and teacher's manuals all strive to meet these objectives. His briefer depiction of critical thinking is that it involves judgments based on criteria, or reasons. Criteria, he says, can be appraised in terms of "megacriteria" such as reliability,

relevance, strength, coherence, and consistency. (Lipman 1991, p. 119) Critical thinking, he adds, is characterized as “thinking that (1) facilitates judgment because it (2) relies on criteria, (3) is self-correcting, and (4) is sensitive to context.” (Lipman 1991, p. 116)

Picking up on the idea that critical thinking is sensitive to context, critics challenge the notion that critical thinking can be usefully taught independently of specific disciplinary areas. (McPeck) While conceding that there are some generic features of critical thinking that cut across disciplines, they maintain that even these features acquire their meanings only in specific contexts that vary across disciplines (such as history, sociology, biology, chemistry). However, unless the different disciplines ask questions about their own basic assumptions and their relationships to one another, critical thinking within those disciplines will overlook important questions that need attention. Philosophy does ask such questions about other disciplines, as well as about itself.

Lipman’s hope is that philosophy will acquire a central place in the K-12 curriculum, thus enabling students to develop their critical thinking skills through philosophical questioning. At the same time, he claims, philosophy can help students make better sense of their educational experiences. By seeking to develop comprehensive perspectives, philosophy attempts to understand connections. A curriculum that divides students’s education into discrete, self-contained disciplines without encouraging philosophical questions about the nature of those disciplines and their relationships to one another invites a fragmented view of education.

Short of the ambitious program Lipman has in mind for the schools, there are more modest, but worthwhile, ways of bringing philosophical inquiry into the already existing disciplinary structure in the schools. Teachers can invite their students to reflect on philosophical aspects of their subjects of study. At the same time they study history, students can take some time to ask questions about the extent to which historical accounts can be objective—and questions about what ‘objectivity’ might mean, and why it is or is not important to seek it. Similar questions can be asked about the natural and social sciences, including questions about the extent to which science is, or ought to be, “value-free.” (see Goldfarb and Pritchard, in the Other Internet Resources section below). In fact, if room for such questions is not encouraged, one might well ask to what extent critical thinking itself is encouraged.

3. The Institute for the Advancement of Philosophy for Children (IAPC)

The educational movement known as Philosophy for Children got its start in the early 1970s with the publication of Matthew Lipman’s philosophical novel for children, *Harry Stottlemeier’s Discovery*. In 1970 *Harry* made its entry into the Montclair Public Schools in New Jersey. By the mid-70s the Institute for the Advancement of Philosophy for Children (IAPC) was formally in place at Montclair State College (now Montclair State University). The media quickly picked up on reports of significant improvements in the reading and critical thinking skills of middle school children who were involved in IAPC programs. Subsequently, IAPC has produced materials consisting of children’s novels with

accompanying teachers' workbooks for the entire K-12 curriculum. Thousands of children in New Jersey, across the United States, and even around the world have been introduced to IAPC educational programs.

An unassuming 96 page novel for middle-school children, *Harry Stottlemeier's Discovery* features Harry and his 5th grade classmates. Adults occasionally enter in, but the primary philosophical work is the children's. Harry and his friends discover several basic concepts and rules of Aristotelean logic; and they puzzle over questions about the nature of thought, mind, causality, reality, knowledge and belief, right and wrong, and fairness and unfairness. The story does not introduce any of the special vocabulary of philosophy (not even the word 'philosophy' itself makes an appearance). Philosophical inquiry is initiated by the children in the story rather than adults.

"What *is* Harry Stottlemeier's discovery?" *Harry*'s readers might ask. This question is not directly answered. However, one candidate stands out among the many things that Harry discovers in the course of exploring questions about logic, knowledge, reality, and the mind. Harry and his classmates are asked to write a paper on the topic, "The Most Interesting Thing in the World." Entitled *Thinking*, Harry's essay begins:

> To me, the most interesting thing in the whole world is thinking. I know that lots of other things are also very important and wonderful, like electricity, and magnetism and gravitation. But although we understand them, they can't understand us. So thinking must be something very special.

After writing several more paragraphs, Harry puts his paper aside. Later he thinks, "In school, we think about math, and we think about spelling, and we think about grammar. But who every heard of thinking about thinking?" So, he adds one more sentence to his paper: "If we think about electricity, we can understand it better, but when we think about thinking, we seem to understand ourselves better."

Without using the word 'philosophy,' either here or anywhere else in *Harry*, Lipman shows Harry engaged in serious philosophical thought, "thinking about thinking." This, we might say, reveals Harry's discovery of the joys of philosophical thinking. But there is more. Harry also notices that, as interesting and important as thinking about thinking is, it seems to have no special place in school. Finally, although his paper begins in the first person, it quickly moves to 'we' and focuses on what might be accomplished *with others* in the classroom.

One of the more attractive features of Philosophy for Children for many teachers is that it promotes the idea of the classroom as a "community of inquiry" in which students openly and respectfully exchange ideas. Each student is regarded as having the potential to make valuable contributions to the topics under consideration. Students are encouraged to develop good listening skills, responsiveness to what others say, willingness to try to supports one's own ideas with good reasons, and openness to the possibility that one should modify one's beliefs in light of new considerations. In short, the classroom is designed to reinforce the student's potential for *reasonableness*. This involves more than being able to engage in

skillful reasoning. As Laurance J. Splitter and Ann M. Sharp put it (Splitter and Sharp, p. 6):

> Reasonableness is primarily a social disposition: the reasonable person respects others and is prepared to take into account their views and their feelings, to the extent of changing her own mind about issues of significance, and consciously allowing her own perspective to be changed by others. She is, in other words, willing to be reasoned with.

Teachers who look favorably on the idea that the classroom should be a "community of inquiry" nevertheless may resist the idea that *philosophy* should be the centerpiece for discussion. Other subjects, they might contend, lend themselves well to forming the sort of collaborative learning environment that can fairly be called a "community of inquiry." Philosophy for Children advocates need not deny this; however, they can point to the success IAPC has had in showing how well suited philosophy is for promoting the reasonableness of children in a "community of inquiry."

One reason for resistance is that it may be thought that philosophy is, at best, a suitable subject for relatively few students at the pre-college level. Since philosophy traditionally has been taught only at the college level in the United States, it might be thought that it can be suitable for only a small segment of students at pre-college levels—the two percent of students who are classified as "gifted and talented." However, Philosophy for Children programs have shown themselves to be remarkably successful in drawing virtually all students in the classroom together in inquiry. Teachers are often surprised, and pleased, to see many of their most reticent, "underachieving" students actively join in the discussion of philosophical ideas.

Nevertheless, because they lack background in the formal study of philosophy, many teachers are reluctant to encourage the philosophical thinking of their students. Their fears, however, are exaggerated. Familiarity with some of the standard philosophical literature might be desirable, but it is not necessary for bringing Philosophy for Children into the classroom. What is required is the ability to facilitate philosophical discussion. For this, it is much more important that teachers have some philosophical curiosity themselves than a familiarity with academic philosophical literature. Like their students, teachers unfamiliar with the discipline of philosophy may nevertheless have an aptitude for philosophical thinking—or at least a knack for recognizing when others are engaged in philosophical thought.

Facilitating a Philosophy for Children discussion does not mean dominating it; it is important for teachers to allow their students to develop their own ideas. Teachers are not expected to provide, or even have, answers to all the questions. They can share puzzlement with their students, be open to unexpected but suggestive responses to the questions they and their students pose, and take pleasure in observing the exchanges students have with each other. This means shedding the traditional role of teacher as lecturer and answer-giver. Especially for teachers who are uncertain about what this entails, workshops like those offered by IAPC provide a good introduction to the pedagogy of Philosophy for Children.

IAPC's approach has been to prepare a set of novels with accompanying teacher's workbooks, and to prepare teachers to use these materials by conducting intensive workshops that themselves illustrate the

proposed pedagogy. It is emphasized that the teacher's role in the classroom is to facilitate discussion rather than to present philosophical ideas didactically. The novels provide a stimulus for children to come up with their own questions and ideas. Typically, students first read aloud a few paragraphs of a novel. Then they suggest ideas prompted by their reading that they would find it interesting to explore together. The teacher's workbook contains hundreds of thinking exercises and activities that can help the students advance their inquiry. The teacher's aim should be to foster a "community of inquiry" in which students, insofar as possible, themselves initiate discussion and exchange ideas with each other rather than simply respond to teacher prompts. A robust discussion will find students not only stating their own ideas, but also supporting them with reasons and responding to the similarities and differences between their ideas and those of their classmates. This "community of inquiry" is intended to foster a respect for the ideas of others, as well as a respect for one's own.

4. Philosophizing With Others?

Philosophy for Children encourages children to think *for* themselves at the same time that it encourages them to think *with* others. However, philosophy is often viewed as more a matter of solitary reflection, perhaps involving exchanges between a few other solitary thinkers—something to which the "masses" are neither privy nor attracted. Perhaps many would claim that this is philosophy at its best; like physics or mathematics, "philosophy for everyone" is watered down. There is no need for Philosophy for Children to challenge this analogy. In fact, it can turn it in its favor. However esoteric physics and mathematics at their best may be, the schools nevertheless recognize the importance of making these subjects available to all students. Similarly, Philosophy for Children advocates can counter that there should be a place for the entire classroom—including "gifted and talented," "underachieving," and "ordinary" students— pursuing philosophical questions together.

For this to work, it must be possible for children in the classroom to engage in sustained philosophical discussion with others. As already noted, Gareth Matthew's writings provide ample evidence that many children of capable of having interesting, if not profound, philosophical thoughts. Less obvious, however, is children's ability to sustain and develop this with others. Anecdotes of young children spontaneously sharing a philosophical thought with an observant adult are not sufficient. Matthews' *Dialogues With Children* provides good evidence that children can go well beyond this. Transcripts of lengthy philosophical conversations of children found in Pritchard (1985, 1996) and issues of *Thinking* and *Analytic Teaching* should leave little doubt that children have this ability.

Admittedly, this is quite different from the approach of Jostein Gaarder's popular novel, *Sophie's World*, which introduces young readers to philosophy in a dialectical, but nevertheless didactic manner. (New York: Farrar, Straus and Giroux, 1994) Although Sophie is certainly an apt philosophy student, her older mentor clearly leads the way by introducing Sophie to the *history* of philosophy, which in turn is used to illuminate philosophical questions that confront her. As its subtitle *A Novel About the History of Philosophy* suggests, *Sophie's World* aims as much at acquainting its young (and older) readers with a working familiarity with the *history* of philosophy as at encouraging philosophical reflection itself. Philosophy for Children aims primarily at the latter, however much IAPC materials themselves might

indirectly be informed by the history of philosophy. In any case, IAPC supporters might argue, the mentor/apprentice approach of *Sophie's World* tends to reinforce the idea that philosophy is primarily something passed down from adults rather than, as Gareth Matthews suggests, an important aspect of children's natural curiosity.

5. Philosophy For Children Around The World

In 1985, as a reflection of the rapidly expanding international impact of philosophy for children, educators from around the world established the International Council for Philosophical Inquiry with Children (ICPIC). ICPIC sponsors an international conference every other year, with hosts including Australia, Austria, Brazil, England, Mexico, Spain, and Taiwan. Although retaining strong ties with IAPC, ICPIC members have established their own institutional structures and they have developed centers, associations, and programs independently of IAPC. In North America, there is the North American Association of Community of Inquiry, which meets every other year (years when ICPIC is not meeting). Australia and New Zealand are organized under the Federation of Australasian Philosophy for Children Associations (FAPCA), which meets annually. Philosophy for children endeavors can be found in colleges, universities, and associations in more than 20 countries around the world. (Sasseville)

In addition to IAPC materials, there is a great deal of children's literature with rich enough content to be used to facilitate interesting philosophical discussions. (See Matthews, 1980, 1984, 1994 and his regular contributions to IAPC's journal, *Thinking*.) There is also a growing body of philosophy for children educational materials being developed outside of IAPC. (See, e.g., Cam; DeHaan, MacColl, and McCutcheon; Fisher; Keen; Murris; Sprod, and White.) *Thinking*, *Analytical Teaching*, and *Critical & Creative Thinking* are three longstanding journals that are specifically devoted to the philosophical thinking of children. A new periodical, sponsored by the American Philosophical Association's Pre-College Philosophy Committee, is *Questions: Philosophy for Young People*, which features young people's writings on special philosophical topics. The first issue (Spring 2001) focused on children's rights.

Fortunately, the internet makes it possible to keep up with the latest developments around the world and communicate quickly with other educators interested in Philosophy for Children. See the Other Internet Resources section below for a useful list.

Bibliography

Books and Articles

- Astington, Janet Wilde, *The Child's Discovery of the Mind* (Cambridge, Mass.: Harvard University Press, 1993).
- Cam, Philip, *Thinking Together: Philosophical Inquiry for the Classroom* (Sydney: Primary

English Teaching Association and Hale & Iremonger, 1995).

- Cam, Philip, *Thinking Stories 1, 2, and 3: Philosophical Inquiry for Children* (Sydney: Hare & Iremonger, 1993, 1994, and 1997).
- DeHaan, Chris; MacColl, San; and McCutcheon, Lucy, *Philosophy With Kids*, Books 1-4 (Melbourne: Longman, 1995).
- Dewey, John, *Reconstruction in Philosophy in John Dewey, the Middle Works, 1899-1924*, vol. 12, ed. Jo Ann Boydston (Carbondale: Southern Illinois Press, 1991).
- Ennis, Robert, “A Conception of Critical Thinking--With Some Curriculum Suggestions,” *Newsletter on Teaching Philosophy,* American Philosophical Association, University of Delaware, Summer 1987, pp. 1-5.
- Facione, Peter, ed., “Report on Critical Thinking,” American Philosophical Association Subcommittee on Pre-College Philosophy, University of Delaware, 1989.
- Figueroa, Robert and Goering, Sara, “The Summer Philosophy Institute of Colorado: Building Bridges,” *Teaching Philosophy*, Vol. 20, No. 2, pp. 155-168.
- Fisher, Robert, *Teaching Thinking: Philosophical Inquiry in the Classroom* (Cassell, 1998).
- Gaarder, Jostein, *Sophie’s World*: *A Novel About the History of Philosophy* (New York: Harper, Straus and Giroux, 1994).
- Gilligan, Carol, *In a Different Voice: Psychological Theory and Women’s Development* (Cambridge, Mass.: Harvard University Press, 1982).
- Gopnik, A., Kuhl, and Meltzoff, A., *The Scientist in the Crib: What Early Learnig Tells us About the Mind* (New York: Perennial Books, 1999).
- Goswami, Usha, *Cognition in Children* (East Sussex, UK: Psychology Press, 1998).
- Gregory, Maughn, “Care as a Goal of Democratic Education,” *Journal of Moral Education*, Vol. 29, No. 4, 2000, pp. 445-461.
- Kohlberg, Lawrence, *The Philosophy of Moral Development: Essays on Moral Development*, Vol. 1 (San Francisco: Harper & Row, 1981).
- Keen, Judy, *Brain Strain 1 & 2*. (Melbourne: MacMillan Education, 1997).
- Lipman, Matthew, *Harry Stottlemeier’s Discovery* (Upper Montclair, NJ: Institute for the Advancement of Philosophy for Children, 1974). (Also the author of *Kio and Gus, Pixie, Lisa*, and other K-12 novels and accompanying teachers’ manuals, all available through IAPC.)
- -----, *Philosophy Goes to School* (Philadelphia: Temple University Press, 1988).
- -----, *Thinking in Education* (New York: Cambridge University Press, 1991).
- Lipman, Matthew, ed. *Thinking Children and Education* (Dubuque, Iowa: Kendall/Hunt, 1993).
- Lipman, Matthew; Sharp, Ann M.; and Oscanyan, Frederick, eds., *Growing Up With Philosophy* (Philadelphia: Temple University Press, 1978).
- Matthews, Gareth, *Philosophy and the Young Child* (Cambridge, Mass.: Harvard University Press, 1980).
- -----, *Dialogues With Children* (Cambridge, Mass.: Harvard University Press, 1984).
- -----, *The Philosophy of Childhood* (Cambridge, Mass.: Harvard University Press, 1994)
- -----, “The Ring of Gyges: Plato in Grade School,” *International Journal of Applied Philosophy*, Vol. 14:1, Spring 2000, pp. 3-11.
- McPeck, John, “Critical thinking and the ‘Trivial Pursuit’ Theory of Knowledge,” *Teaching Philosophy*, Vol. 8, No. 4, 1985, pp. 295-308.

- Murris, K., *Teaching Philosophy With Picture Books* (London: Infonet Publications, 1992).
- Partridge, F.; Dubuc, F.; Splitter, L.; and Sprod, T., *Places for Thinking* (Melbourne: Australian Council for Educational Research, 1999).
- Phillips, Christopher, *The Socrates Cafe* (New York: W.W. Norton, 2001).
- Piaget, Jean, "Children's Philosophies," in *A Handbook of Child Psychology*, ed. Carl Murchison, 2nd ed. rev. (Worcester, Mass: Clark University Press, 1933).
- Pritchard, Michael S., *On Becoming Responsible* (Lawrence, KS: University Press of Kansas, 1991).
- -----, *Philosophical Adventures With Children* (Lanham, MD: University Press of America, 1985).
- -----, *Reasonable Children* (Lawrence, KS: University Press of Kansas, 1996)
- -----, "Moral Philosophy for Children and Character Education," *International Journal of Applied Philosophy*, Vol. 14:1, Spring 2000, pp. 13-26.
- Reed, Ronald, *Talking With Children* (Denver: Arden Press, 1983).
- Reed, Ronald, and Sharp, Ann M., eds., *Studies in Philosophy for Children*: *Harry Stottlemeier's Discovery* (Philadelphia: Temple University Press, 1992).
- Reed, Ronald, and Sharp, Ann M., *Studies in Philosophy for Children: Pixie* (Madrid: Ediciones De La Torre, 1996).
- Sasseville, Michel, "The State of International Cooperation in Philosophy for Children," UNESCO Meeting, Paris, March 1998, in *Critical and Creative Thinking*: *The Australasian Journal of Philosophy for Children*, Vol. 7, No. 1, March 1999, pp. 57-79.
- Sharp, Ann M., "The Community of Inquiry: Education for Democracy," *Thinking*, 9(2), 1991, pp. 31-37.
- Sharp, Ann.M., ed., "Women, Feminism, and Philosophy for Children," Special Issue of *Thinking*, Vol. 11, Nos. 3 & 4, 1994.
- Shipman, Virginia, *New Jersey Reasoning Skills Test* (Upper Montclair, NJ: Institute for the Advancement of Philosophy for Children, 1983).
- Splitter, Laurance and Sharp, Ann M., *Teaching for Better thinking: The Classroom Community of Inquiry* (Hawthorn, Vic.: Australian Council for Educational Research, 1995).
- Sprod, T., *Books Into Ideas* (Cheltenham, Vic.: Hawker Brownlow Education, 1993).
- Turner, Susan M. and Matthews, Gareth, eds., *The Philosopher's Child* (Rochester, NY: University of Rochester Press, 1998).
- Weinstein, Mark, "Critical Thinking and Moral Education," *Thinking*, Vol. 7, No. 3, 1989, pp. 42-49.
- White, David A., *Philosophy for Kids* (Prufrock, 2000).
- Wilks, S., *Critical and Creative Thinking*: *Strategies for Classroom Inquiry* (Armadale, Vic.: Eleanor Curtain, 1995).

Periodicals

- *Analytic Teaching: The Community of Inquiry Journal*, Viterbo College, La Crosse, WI.
- *Critical & Creative Thinking: The Australasian Journal of Philosophy for Children*, The Federation of Austrailian Philosophy for Children Associations (FAPCA).

- *Questions: Philosophy for Young People*, published by the Philosophy Documentation Center, Bowling Green State University. Inaugurated in Spring 2001 with the support of the American Philosophical Association and the Northwest Center for Philosophy for Children.
- *Thinking: The Journal of Philosophy for Children*, Institute for the Advancement of Philosophy for Children, Montclair State University, NJ.

Other Internet Resources

- Goldfarb, Theodore and Pritchard, Michael S., *Ethics in the Science Classroom*, NSF Grant No. SBR-9601284
- FAPCA (Federation of Australasian Philosophy for Children)
- IAPC (The Institute for the Advancement of Philosophy for Children)
- ICPIC (International Council for Philosophical Inquiry With Children)
- Institute for Critical Thinking
- The Owl Project
- Philosophy for Children on the World Wide Web

Related Entries

childhood, the philosophy of

Michael Pritchard
pritchard@wmich.edu

First published: May 2, 2002
Content last modified: May 2, 2002

Zeno's Paradoxes

Almost everything that we know about Zeno of Elea is to be found in the opening pages of Plato's *Parmenides*. There we learn that Zeno was nearly 40 years old when Socrates was a young man, say 20. Since Socrates was born in 469 BC we can estimate a birth date for Zeno around 490 BC. Beyond this, really all we know is that he was close to Parmenides (Plato reports the gossip that they were lovers when Zeno was young), and that he wrote a book of paradoxes defending Parmenides' philosophy. Sadly this book has not survived, and what we know of his arguments is second-hand, principally through Aristotle and his commentators (here I have drawn particularly on Simplicius, who, though writing a thousand years after Zeno, apparently possessed at least some of his book). There were apparently 40 'paradoxes of plurality', attempting to show that ontological pluralism -- a belief in the existence of many things rather than only one -- leads to absurd conclusions; of these paradoxes only two definitely survive, though a third argument can probably be attributed to Zeno. Aristotle speaks of a further four arguments against motion (and by extension change generally), all of which he gives and attempts to refute. In addition Aristotle attributes two other paradoxes to Zeno. Sadly again, almost none of these paradoxes are quoted in Zeno's original words by their various commentators, but in paraphrase.

1. Background

Before we look at the paradoxes themselves it will be useful to sketch some of their historical and logical significance. First, Zeno sought to defend Parmenides by attacking his critics. Parmenides rejected pluralism and the reality of any kind of change: for him all was one indivisible, unchanging reality, and any appearances to the contrary were illusions, to be dispelled by reason and revelation. Not surprisingly, this philosophy found many critics, who ridiculed the suggestion; after all it flies in the face of some our most basic beliefs about the world. (Interestingly, general relativity -- particularly quantum general relativity -- arguably provides a novel -- if novelty *is* possible -- argument for the Parmenidean denial of change: Belot and Earman, 2001.) In response to this criticism Zeno did something that may sound obvious, but which had a profound impact on Greek philosophy that is felt to this day: he attempted to show that equal absurdities followed logically from the denial of Parmenides’ views. You think that there are many things? Then you must conclude that everything is both infinitely small and infinitely big! You think that motion is infinitely divisible? Then it follows that nothing moves! (This is what a ‘paradox’ is: a demonstration that a contradiction or absurd consequence follows from apparently reasonable assumptions.)

‘Dialectic’, the technique of arguing for or against a position by careful logical reasoning -- and in particular the technique of arguing against a view by showing that it entails unacceptable consequences -- was a crucial innovation, which has governed philosophical method ever since. In the absence of such a method one can only defend a position by mystical revelation say, or by rhetorical rather than rational appeal, or by force perhaps. And according to Aristotle, Zeno was the inventor of the method (in philosophy of least, for such an approach has been a part of mathematics for even longer). Later philosophers however -- especially Plato and Aristotle -- were far finer exponents of the approach.

As we read the arguments it is crucial to keep this method in mind. They are always directed towards a more-or-less specific target: the views of some person or school. We must bear in mind that the arguments are *ad hominem*, not in the ‘bad sense’ that they attack a person rather than his views but in the ‘good sense’ that they are formulated against a particular philosopher’s assertions. They work by temporarily supposing ‘for argument’s sake’ that those assertions are true, and then arguing that if they are then absurd consequences follow -- that nothing moves for example: they are ‘*reductio ad absurdum*’ arguments. Then, if the argument is logically valid, and the conclusion genuinely unacceptable, the assertions must be false after all. Thus when we look at Zeno’s arguments we must ask two related questions: whom or what position is Zeno attacking, and what exactly is assumed for argument's sake? If we find that Zeno makes hidden assumptions beyond what the position under attack commits one to, then the absurd conclusion can be avoided by denying one of the hidden assumptions, while maintaining the position. Indeed commentators at least since Aristotle have responded to Zeno in this way.

So whom do Zeno's arguments attack? There is a huge literature debating Zeno’s exact historical target.

As we shall discuss briefly below, some say that the target was a technical doctrine of the Pythagoreans, but most today see Zeno as opposing common-sense notions of plurality and motion. I will approach the paradoxes in this spirit, and refer the reader to the literature concerning the interpretive debate.

That said, it is also the majority opinion that -- with certain qualifications -- Zeno's paradoxes reveal some problems that cannot be resolved without the full resources of mathematics as worked out in the Nineteenth century (and perhaps beyond). This is not (necessarily) to say that modern mathematics is required to answer any of the problems that Zeno explicitly wanted to raise; arguably Aristotle and other ancients had replies that would -- or should -- have satisfied Zeno. (Nor yet should we conclude that Zeno's work had any direct influence on the history of mathematics, though surely the kind of worries that he raised did.) However, as mathematics developed, and more thought was given to the paradoxes, new aspects and new difficulties arose from them; these difficulties require modern mathematics for their resolution. These new difficulties arise partly in response to the evolution in our understanding of what mathematical rigor demands: solutions that would satisfy Aristotle's standards of rigor would not satisfy ours. Thus we shall push several of the paradoxes from their common sense formulations to their resolution in modern mathematics. (Another qualification: I will offer resolutions in terms of 'standard' mathematics, but other modern formulations are also capable of dealing with Zeno.)

2. The Paradoxes of Plurality

2.1 The Argument from Denseness

> If there are many, they must be as many as they are and neither more nor less than that. But if they are as many as they are, they would be limited. If there are many, things that are are unlimited. For there are always others between the things that are, and again others between those, and so the things that are are unlimited. (Simplicius(a) *On Aristotle's Physics*, 140.29)

This first argument, given in Zeno's words according to Simplicius, attempts to show that there could not be many things, on pain of contradiction. Assume then that there are many things. First, he says that any collection must contain some definite number of things, neither more nor fewer. But if you have a definite number of things, he further concludes, you must have a finite -- 'limited' -- number of them; he implicitly assumes that to have infinitely many things is not to have any particular number of them. Second, imagine any collection of things arranged in space -- imagine them lined up in one dimension for definiteness. Between and two of them, he claims, is a third; and in between these three elements another two; and another four between these five; and so on without end. Therefore the limited collection is also 'unlimited', which is a contradiction, and hence our original assumption must be false: there are not many things after all. At least, so Zeno's reasoning runs.

But why are there 'always others between the things that are'? (In modern terminology, why must objects always be 'densely' ordered?) Suppose that I had imagined a collection of ten apples lined up;

then there is indeed another apple between the sixth and eighth, but there is none between the seventh and eighth! On the assumption that Zeno is not simply confused, what does he have in mind? There are two possibilities: first, one might hold that for any pair of physical objects (two apples say) to actually be two distinct objects and not just one (a 'double-apple') there must be a third between them, physically separating them, even if it is just air. And one might think that for these three to be distinct, there must be two more objects separating them, and so on (this view presupposes that their being made of different substances is not sufficient to render them distinct). Second, one might hold that any body has parts that can be densely ordered. Of course 1/2s, 1/4s, 1/8s and so on of apples are not dense -- some such parts are adjacent -- but there may be sufficiently small parts -- call them 'point-parts' -- that are. Indeed, if between any two point-parts there lies a finite distance, and if point-parts can be arbitrarily close, then they are dense; a third lies at the half-way point of any two. In particular, familiar geometric points are like this, and hence are dense.

And thus we should read the argument as follows: if you suppose that the world contains many things, then you are faced with a contradiction, for the collection must be both finite and infinite -- finite because it contains a definite number of things, and infinite because they are dense. The assumption that any definite number is finite seems intuitive, but we now know, thanks to the work of Cantor in the Nineteenth century, how to understand infinite numbers in a way that makes them just as definite as finite numbers. One central element of the this theory of the 'transfinite numbers' is a precise definition of when two infinite collections are the same size, and when one is bigger than the other -- with such a definition in hand it is then possible to order the infinite numbers just as the finite numbers are ordered. For example, both the fractions and geometric points in a line are dense, but there are different, definite infinite numbers of them. (See Further Reading below for references to introductions to these mathematical ideas.) Of course, settling the mathematical question of whether infinite numbers can be definite doesn't show that real physical objects actually have geometric point parts, all it shows is that it is a logical possibility.

2.2 The Argument from Finite Size

> … if it should be added to something else that exists, it would not make it any bigger. For if it were of no size and was added, it cannot increase in size. And so it follows immediately that what is added is nothing. But if when it is subtracted, the other thing is no smaller, nor is it increased when it is added, clearly the thing being added or subtracted is nothing. (Simplicius(a) *On Aristotle's Physics*,139.9)
>
> But if it exists, each thing must have some size and thickness, and part of it must be apart from the rest. And the same reasoning holds concerning the part that is in front. For that too will have size and part of it will be in front. Now it is the same thing to say this once and to keep saying it forever. For no such part of it will be last, nor will there be one part not related to another. Therefore, if there are many things, they must be both small and large; so small as not to have size, but so large as to be unlimited. (Simplicius(a) *On Aristotle's Physics*, 141.2)

Once again we have Zeno's own words. According to his conclusion, there are three parts to this argument, but only two survive. The first -- missing -- argument purports to show that if many things exist then they must have no size at all. Second, from this Zeno argues that it follows that they do not exist at all; since the result of joining (or removing) a sizeless object to anything is no change at all, he concludes that the thing added (or removed) is literally nothing. The argument to this point is a self-contained refutation of pluralism, but Zeno goes on to generate a further problem for someone who continues to urge the existence of a plurality. This third part of the argument is rather badly put but it seems to run something like this: suppose there is a plurality, so some spatially extended object exists (after all, he's just argued that inextended things do not exist). Since it is extended, it has two spatially distinct parts (one 'in front' of the other). And the parts exist, so they have extension, and so they also each have two spatially distinct parts; and so on without end. And hence, the final line of argument seems to conclude, the object, if it is extended at all, is infinite in extent.

But what could justify this final step? It doesn't seem that because an object has two parts it must be infinitely big! And neither does it follow from any other of the divisions that Zeno describes here; four, eight, sixteen, or whatever finite parts make a finite whole. Again, surely Zeno is aware of these facts, and so must have something else in mind, presumably the following: he assumes that if the infinite series of divisions he describes were repeated infinitely many times then a definite collection of parts would result. And notice that he doesn't have to assume that anyone could actually carry out the divisions -- there's not enough time and knives aren't sharp enough -- just that an object can be geometrically decomposed into such parts (neither does he assume that these parts are what we would naturally categorize as distinct physical objects like apples, cells, molecules, electrons or so on, but only that they are geometric parts of these objects). Now, if -- as a pluralist might well accept -- such parts exist, it follows from the second part of his argument that they are extended, and, he apparently assumes, an infinite sum of finite parts is infinite.

Here we should note that there are two ways he may be envisioning the result of the infinite division.

First, one could read him as first dividing the object into 1/2s, then one of the 1/2s -- say the second -- into two 1/4s, then one of the 1/4s -- say the second again -- into two 1/8s and so on. In this case the result of the infinite division results in an endless sequence of pieces of size 1/2 the total length, 1/4 the length, 1/8 the length … . And then so the total length is (1/2 + 1/4 + 1/8 + …) of the length, which Zeno concludes is an infinite distance, so that the pluralist is committed to the absurdity that finite bodies are 'so large as to be unlimited'.

What is often pointed out in response is that Zeno gives us no reason to think that the sum is infinite rather than finite. He might have had the intuition that any infinite sum of finite quantities, since it grows endlessly with each new term must be finite, but one might also take this kind of example as showing that some infinite sums are after all finite. Thus, contrary to what he thought, Zeno has not proven that the absurd conclusion follows. However, what is not always appreciated is that the pluralist is not off the hook so easily, for it is not enough just to say that the sum *might* be finite, she must also show that it *is* finite -- otherwise we remain uncertain about the tenability of her position. As an illustration of the

difficulty faced here consider the following: many commentators speak as if it is simply obvious that the infinite sum of the fractions is 1, that there is nothing to infinite summation. But what about the following sum: 1 - 1 + 1 - 1 + 1 - … . Obviously, it seems, the sum can be rewritten (1 -1) + (1 - 1) + … = 0 + 0 + … = 0. Surely this answer seems as intuitive as the sum of fractions. But this sum can also be rewritten 1 - (1 - 1 + 1 - 1 + …) = 1 - 0 -- since we've just shown that the term in parentheses vanishes -- = 1. Relying on intuitions about how to perform infinite sums leads to the conclusion that 1 = 0. Until one can give a theory of infinite sums that can give a satisfactory answer to any problem, one cannot say that Zeno's infinite sum is obviously finite. Such a theory was not fully worked out until the Nineteenth century by Cauchy. (In Cauchy's system 1/2 + 1/4 + … = 1 but 1 - 1 + 1 - … is undefined.)

Second, it could be that Zeno means that the object is divided in half, then both the 1/2s are both divided in half, then the 1/4s are all divided in half and so on. In this case the pieces at any particular stage are all the same finite size, and so one concludes that the result of carrying on the procedure infinitely would be pieces the same size, which if they exist -- according to Zeno -- is greater than zero; but an infinity of equal extended parts is indeed infinitely big.

Actually a little care is needed in drawing this conclusion. The procedure just described involves doubling the number of pieces after every division and so after N divisions there are 2^N pieces. But it turns out that for any natural or infinite number, N, $2^N > N$, and so the number of pieces obtained by the infinity of divisions described is an even larger infinity. This is no problem as we mentioned above, since infinities come in different sizes. The number of times everything is divided in two is said to be 'countably infinite': there is a countable infinity of things in a collection if they can be labeled by the numbers 1, 2, 3, … without remainder on either side. But the number of pieces is 'uncountably infinite', which means that there is no way to label them 1, 2, 3, … without missing some of them -- in fact infinitely many of them. However, Cauchy's definition of an infinite sum only applies to countably infinite series of numbers, and so does not apply to the pieces we are considering. However, we could consider just countably many of them, whose lengths -- since they are all equal and non-zero -- will sum to an infinite length; clearly the length of *all* of the pieces cannot be less than this.

There is however an escape for the pluralist who believes that objects have parts of this kind (which they do if they have parts with the properties of geometric points): she must claim that the parts in fact have no extension, even though they exist. That would block the conclusion that finite objects are infinite, but it seems to push her back to the other horn of Zeno's argument, for how can all these zero length pieces make up a non-zero sized whole? (Note that according to Cauchy 0 + 0 + 0 + … = 0 but this result shows nothing here, for as we saw there are uncountably many pieces to add up -- more than are added in this sum.) We shall postpone this question for the discussion of the next paradox, where it comes up explicitly.

2.3 The Argument from Complete Divisibility

> … whenever a body is by nature divisible through and through, whether by bisection, or generally by any method whatever nothing impossible will have resulted if it has actually

> been divided … . though perhaps nobody in fact could so divide it.
>
> What then will remain? A magnitude? No: that is impossible, since then there will be something not divided, whereas *ex hypothesi* the body was divisible *through and through*. But if it be admitted that neither a body nor a magnitude will remain … the body will *either* consist of points (and its constituents will be without magnitude) *or* it will be absolutely nothing. If the latter, then it might both come-to-be out of nothing and exist as a composite of nothing; and thus presumably the whole body will be nothing but an appearance. But if it consists of points, it will not possess any magnitude. (Aristotle *On Generation and Corruption*, 316a19)

These words are Aristotle's not Zeno's, and indeed the argument is not even attributed to Zeno by Aristotle. However we have Simplicius' opinion ((a) *On Aristotle's Physics*, 139.24) that it originates with Zeno, which is why it is included here. Aristotle begins by hypothesizing that some body is completely divisible, 'through and through'; the second step of the argument makes clear that he means by this that it is divisible into parts that themselves have no size -- parts with any magnitude remain incompletely divided. (Once again what matters is that the body is genuinely composed of such parts, not that anyone has the time and tools to make the division.) So suppose the body is divided into its dimensionless parts. These parts could either be nothing at all -- as Zeno argued above -- or 'point-parts'. If the parts are nothing then so is the body: it's just an illusion. And, the argument concludes, even if they are points, since these are unextended the body itself will be unextended: surely any sum -- even an infinite one -- of zeroes is zero.

One could of course point out that it is only assumed that an infinity of zeroes is itself zero, and deny that assumption. However it has a strong intuitive pull, and once again one should show how any dimensionless points actually do make an extended whole. Fortunately Grünbaum (1967) showed how this is possible according to the modern mathematical treatment of a line. Consider a line segment of unit length. At its most basic level the segment is just a set of points -- if you take any spatial part of it, all you have is a point or set of points. Now Cantor gave a beautiful, astounding and extremely influential 'diagonal' proof that the number of points in the segment is uncountably infinite: there is no way to label *all* the points in the line with the infinity of numbers 1, 2, 3, … . As we noted above, it follows that we cannot apply the Cauchy definition of infinite sums to the points of the line, and so happily we cannot immediately conclude that because they all have zero length so does the whole line. But that still leaves open the question of how the line gets extension from its inextended points.

So suppose that you are just given the number of points in a line and that their lengths are all zero; how would you determine the length? Do we need a new definition, one that extends Cauchy's to uncountably infinite sums? It turns out that that would not help, because Cauchy further showed that any segment, of any length whatsoever (and indeed an entire infinite line) *have exactly the same number of points as our unit segment*. So knowing the number of points won't determine the length of the line, and so nothing like familiar addition -- in which the whole is determined by the parts -- is possible. Instead we must think of the distance properties of a line as logically posterior to its point composition: *first* we have a set of points (ordered in a certain way, so that there is some fact, for example, about which of any two is

before the other) *then* we define a function of two points which specifies how far apart they are (and which satisfies such conditions as that the distance between *A* and *B* plus the distance between *B* and *C* equals the distance between *A* and *C* -- assuming that *C* is not between *A* and *B*). By analogy, the maiden names of a married couple do not determine their surname: they could take either maiden name or hyphenate, or take a wholly new name if they choose. Thus we answer Zeno as follows: the argument assumed that the size of the body was a sum of the sizes of the point parts, but that is not the case; according to modern mathematics, a line is an uncountable infinity of points plus a distance function. (Note that Grünbaum used the fact that the point composition fails to determine a length to support his 'conventionalist' view that a line has no determinate length at all, independent of a standard of measurement.)

3. The Paradoxes of Motion

3.1 The Dichotomy

> The first asserts the non-existence of motion on the ground that that which is in locomotion must arrive at the half-way stage before it arrives at the goal. (Aristotle *Physics*, 239b11)

This paradox is known as the 'dichotomy' because it involves repeated division into two (like the second paradox of plurality). Like the other paradoxes of motion we have it from Aristotle, who sought to refute it.

Suppose a very fast runner -- such as mythical Atalanta -- needs to run for the bus. Clearly before she reaches the bus stop she must run half-way, as Aristotle says. There's no problem there; supposing a constant motion it will take her 1/2 the time to run half-way there and 1/2 the time to run the rest of the way. Now she must also run half-way to the half-way point -- i.e., a 1/4 of the total distance -- before she reaches the half-way point, but again she is left with a finite number of finite lengths to run, and plenty of time to do it. And before she reaches 1/4 of the way she must reach 1/2 of 1/4 = 1/8 of the way; and before that a 1/16; and so on. There is no problem at any finite point in this series, but what if the halving is carried out infinitely many times? The resulting series contains no first distance to run, for any possible first distance could be divided in half, and hence would not be first after all. However it does contain a final distance, namely 1/2 of the way; and a penultimate distance, 1/4 of the way; and a third to last distance, 1/8 of the way; and so on. Thus the series of distances that Atalanta is required to run is: …, then 1/16 of the way, then 1/8 of the way, then 1/4 of the way, and finally 1/2 of the way (of course we are not suggesting that she *stops* at the end of each segment and then starts running at the beginning of the next -- we are thinking of her continuous run being composed of such parts). And now there is a problem, for this description of her run has her travelling an *infinite* number of *finite* distances, which, Zeno would have us conclude, must take an infinite time, which is to say it is never completed. And since the argument does not depend on the distance or who or what the mover is, it follows that no finite distance can ever be traveled, which is to say that all motion is impossible. (Note that the paradox could easily be generated in the other direction so that Atalanta must first run half way, then half the remaining

way, then half of that and so on, so that she must run the following endless sequence of fractions of the total distance: 1/2, 1/4, 1/8 ….)

A couple of common responses are not adequate. One might -- as Simplicius ((a) *On Aristotle's Physics*, 1012.22) tells us Diogenes the Cynic did by silently standing and walking -- point out that it is a matter of the most common experience that things in fact do move, and that we know very well that Atalanta would have no trouble reaching her bus stop. But this would not impress Zeno, who as a paid up Parmenidean held that many things are not as they appear: it may appear that Diogenes is walking or that Atalanta is running, but appearances can be deceptive and surely we have a logical proof that they are in fact not moving at all. And if one doesn't accept that Zeno has given a proof that motion is illusory -- as we hopefully do not -- then one then owes an account of what is wrong with his argument: he has given reasons why motion is impossible, and so an adequate response must show why those reasons are not sufficient. And it won't do simply to point out that there are some ways of cutting up Atalanta's run -- into just two halves, say -- in which there is no problem. For if you accept all of the steps in Zeno's argument then you must accept his conclusion (assuming that he has reasoned in a logically deductive way): it's not enough to show an unproblematic division, you must also show why the given division is unproblematic.

Another response -- given by Aristotle himself -- is to point out that as we divide the distances run we should also divide the total time taken: there is 1/2 the time for the final 1/2, a 1/4 of the time for the previous 1/4, an 1/8 of the time for the 1/8 of the run and so on. Thus each fractional distance has just the right fraction of the finite total time for Atalanta to complete it, and thus the distance can be completed in a finite time. Aristotle felt that this reply should satisfy Zeno, however he also realized (*Physics*, 263a15) that this could not be the end of the matter (and surely Zeno would have made the same point if presented with Aristotle's response). For now we are saying that the *time* Atalanta takes to reach the bus stop is composed of an infinite number of finite pieces -- …, 1/8, 1/4, and 1/2 (of the total time) -- and isn't that an infinite time?

Of course, one could again claim that some infinite sums in fact have finite totals, and in particular that the sum of these pieces is 1 × the total time, which is of course finite (and again a complete solution would demand a rigorous account of infinite summation, like Cauchy's). However, Aristotle did not make such a move. What he said is worth noting because it had a considerable influence on later thinking about Zeno. In his response Aristotle drew a sharp distinction between what he termed a 'continuous' line and a line divided into parts. Consider a simple division of a line into two: on the one hand there is the undivided line, and on the other the line with a mid-point selected as the boundary of the two halves. Aristotle claims that these are two distinct things: and that the later is only 'potentially' derivable from the former. Next, Aristotle takes the common-sense view that time is like a geometric line, and considers the time it takes to complete the run. We can again distinguish the two cases: on the one hand there is the continuous run from start to finish, and on the other there is the run divided into Zeno's infinity of half-runs. The former is 'potentially infinite' in the sense that it could be divided into latter 'actual infinity'. Here's the crucial step: Aristotle thinks that since these times are *geometrically* distinct they must be *physically* distinct. But how could that be? He claims that the runner must do something at the end of each half-run to make it distinct from the next: she must stop. (Why stop rather than cough or something?

Because if the time is discontinuous then so is the motion.) And so Aristotle's full answer to the paradox is that Zeno's question -- whether the infinite series of runs is possible or not -- is ambiguous. One the one hand, the answer is 'yes' if one means the potentially infinite series that form the continuous run. On the other the answer is 'no' if one means the actual infinity of pieces that form the discontinuous run.

It is hard -- from our modern perspective perhaps -- to see how this answer could be completely satisfactory. In the first place it assumes that a clear distinction can be drawn between potential and actual infinities, something that was never fully achieved. Second, suppose that Zeno's problem turns on the claim that infinite sums of finite quantities are invariably infinite. Then Aristotle's distinction will only help if he can explain why potentially infinite sums are in fact finite (and couldn't I potentially add 1 + 1 + 1 + …, which does not have a finite total); or if he can give a reason why potentially infinite sums just don't exist. Or perhaps Aristotle did not see infinite sums as the problem, but rather whether completing an infinity of finite actions is metaphysically and conceptually and physically possible, an idea discussed at length in recent years: see 'Supertasks' below. In this case we need an account of actions that makes precise the sense in which the continuous run is indeed a single action (using rest to individuate motions seems problematic, for humans are probably never completely still, and yet we perform distinct motions -- breathing, eating, skipping and so on.) Finally, the distinction between potential and actual infinities has played no role in mathematics since Cantor tamed the transfinite numbers -- certainly the potential infinite has played no role in the modern mathematical solutions discussed here.

One last point: Zeno's argument seeks most obviously to establish the impossibility of motion, but he also intended it (and the following arguments) as further refutations of plurality -- certainly, Plato interprets Zeno's intentions in this way. How might the argument seek to establish this conclusion? Presumably Zeno has in mind the view that spatial (and perhaps temporal) distances have a plurality of parts; parts which are infinitely divisible into two. Given that assumption, supposedly finite distances (or times) can be decomposed into an infinity of finite parts with no first (or alternatively, last) one. And how can such distances be finite after all? And if the pluralist also believes in motion, how can such a distance be traversed? It seems it could not be.

3.2 Achilles and the Tortoise

> The [second] argument was called "Achilles," accordingly, from the fact that Achilles was taken [as a character] in it, and the argument says that it is impossible for him to overtake the tortoise when pursuing it. For in fact it is necessary that what is to overtake [something], before overtaking [it], first reach the limit from which what is fleeing set forth. In [the time in] which what is pursuing arrives at this, what is fleeing will advance a certain interval, even if it is less than that which what is pursuing advanced … . And in the time again in which what is pursuing will traverse this [interval] which what is fleeing advanced, in this time again what is fleeing will traverse some amount … . And thus in every time in which what is pursuing will traverse the [interval] which what is fleeing, being slower, has already advanced, what is fleeing will also advance some amount.

(Simplicius(b) *On Aristotle's Physics*, 1014.10)

This paradox turns on much the same considerations as the last. Imagine Achilles chasing a tortoise, and suppose that Achilles is running at 1 *m/s*, that the tortoise is crawling at 0.1 *m/s* and that the tortoise starts out 0.9 *m* ahead of Achilles. On the face of it Achilles should catch the tortoise after 1*s*, at a distance of 1*m* from where he starts (and so 0.1*m* from where the Tortoise starts). We could break Achilles' motion up as we did Atalanta's, into halves, or we could do it as follows: before Achilles can catch the tortoise he must reach the point where the tortoise started. But in the time he takes to do this the tortoise crawls a little further forward. So next Achilles must reach this new point. But in the time it takes Achilles to achieve this the tortoise crawls forward a tiny bit further. And so on to infinity: every time that Achilles reaches the place where the tortoise was the tortoise has had enough time to get a little bit further, and so Achilles has another run to make, and so Achilles has in infinite number of finite catch-ups to do before he can catch the tortoise, and so, Zeno concludes, he never catches the tortoise.

One aspect of the paradox is thus that Achilles must traverse the following infinite series of distances before he catches the tortoise: first 0.9*m*, then an additional 0.09*m*, then 0.009*m*, … . These are the series of distances ahead that the tortoise reaches at the start of each of Achilles' catch-ups. Looked at this way the puzzle is identical to the Dichotomy, for it is just to say that 'that which is in locomotion must arrive [nine tenths of the way] before it arrives at the goal'. And so everything we said above applies here too.

But what the paradox in this form brings out most vividly is the problem of completing a series of actions that has no final member -- in this case the infinite series of catch-ups before Achilles reaches the tortoise. But just what is the problem? Perhaps the following: Achilles' run to the point at which he should reach the tortoise can, it seems, be completely decomposed into the series of catch-ups, none of which take him to the tortoise. Therefore, nowhere in his run does he reach the tortoise after all. But if this is what Zeno had in mind it won't do. Of course Achilles doesn't reach the tortoise at any point of the sequence, for every run in the sequence occurs *before* we expect Achilles to reach it! Thinking in terms of the points that Achilles must reach in his run, 1m does not occur in the sequence 0.9*m*, 0.99*m*, 0.999*m*, … , so of course he never catches the tortoise during that sequence of runs! The series of catch-ups does not after all completely decompose the run: the final point -- at which Achilles does catch the tortoise -- must be added to it. So is there any puzzle? Arguably yes.

Achilles run passes through the sequence of points 0.9*m*, 0.99*m*, 0.999*m*, … , 1*m*. But does such a strange sequence -- comprised of an infinity of members followed by one more -- make sense mathematically? If not then our mathematical description of the run cannot be correct, but then what is? Fortunately the theory of transfinites pioneered by Cantor assures us that such a series is perfectly respectable. It was realized that the order properties of infinite series are much more elaborate than those of finite series. Any way of arranging the numbers 1, 2 and 3 gives a series in the same pattern, for instance, but there are many distinct ways to order the natural numbers: 1, 2, 3, … for instance. Or … , 3, 2, 1. Or … , 4, 2, 1, 3, 5, … . Or 2, 3, 4, … , 1, which is just the same kind of series as the positions Achilles must run through. Thus the theory of the transfinites treats not just 'cardinal' numbers -- which depend only on how many things there are -- but also 'ordinal' numbers which depend further on how the things are arranged. Since the ordinals are standardly taken to be mathematically legitimate numbers, and

since the series of points Achilles must pass has an ordinal number, we shall take it that the series is mathematically legitimate. (Again, see ‘Supertasks’ below for another kind of problem that might arise for Achilles’.)

3.3 The Arrow

> The third is … that the flying arrow is at rest, which result follows from the assumption that time is composed of moments … . he says that if everything when it occupies an equal space is at rest, and if that which is in locomotion is always in a now, the flying arrow is therefore motionless. (Aristotle *Physics*, 239b.30)
>
> Zeno abolishes motion, saying "What is in motion moves neither in the place it is nor in one in which it is not". (Diogenes Laertius *Lives of Famous Philosophers*, ix.72)

This argument against motion explicitly turns on a particular kind of assumption of plurality: that time is composed of moments (or ‘nows’) *and nothing else*. Consider an arrow, apparently in motion, at any instant. First, Zeno assumes that it travels no distance during that moment -- ‘it occupies an equal space’ for the whole instant. But the entire period of its motion contains only instants, all of which contain an arrow at rest, and so, Zeno concludes, the arrow cannot be moving.

An immediate concern is why Zeno is justified in assuming that the arrow is at rest during any instant. It follows immediately if one assumes that an instant lasts 0s: whatever speed the arrow has, it will get nowhere if it has no time at all. But what if one held that the smallest parts of time are finite -- if tiny -- so that a moving arrow might actually move some distance during an instant? One way of supporting the assumption -- which requires reading quite a lot into the text we have -- is to assume that instants are indivisible. Then suppose that an arrow actually moved during an instant. It would be at different locations at the start and end of the instant, which implies that the instant has a ‘start’ and an ‘end’, which in turn implies that it has at least two parts, and so is divisible, and so is not an indivisible moment at all. (Note that this argument only establishes that nothing can move during an instant, not that instants cannot be finite.)

So then, nothing moves during any instant, but time is entirely composed of instants, so nothing ever moves. A first response is to point out that determining the velocity of the arrow means dividing the distance traveled in some time by the length of that time. But -- assuming from now on that instants have zero duration -- this formula makes no sense in the case of an instant: the arrow travels 0*m* in the 0*s* the instant lasts, but 0/0 *m/s* is not any number at all. Thus it is fallacious to conclude from the fact that the arrow doesn’t travel any distance in an instant that it is at rest; whether it is in motion at an instant or not depends on whether it travels any distance in a *finite* interval that includes the instant in question.

The answer is correct, but it carries the counter-intuitive implication that motion is not something that happens at any instant, but rather only over finite periods of time. Think about it this way: time, as we said, is composed only of instants. No distance is traveled during any instant. So when does the arrow

actually move? How does it get from one place to another at a later moment? There's only one answer: the arrow gets from point *X* at time 1 to point *Y* at time 2 simply in virtue of being at successive intermediate points at successive intermediate times -- the arrow never changes its position during an instant but only over intervals composed of instants, by the occupation of different positions at different times. In Bergson's memorable words -- which he thought expressed an absurdity -- 'movement is composed of immobilities' (1911, 308): getting from *X* to *Y* is a matter of occupying exactly one place in between at each instant (in the right order of course).

3.4 The Stadium

> The fourth argument is that concerning equal bodies [*AA*] which move alongside equal bodies in the stadium from opposite directions -- the ones from the end of the stadium [*CC*], the others from the middle [*BB*] -- at equal speeds, in which he thinks it follows that half the time is equal to its double…. And it follows that the *C* has passed all the *A*s and the *B* half; so that the time is half … . And at the same time it follows that the first *B* has passed all the *C*s. (Aristotle *Physics*, 239b33)

The final paradox of motion runs as follows: picture three sets of three touching cubes -- all nine exactly the same, with side L m -- in relative motion. One set -- the *A*s --are at rest, and the others -- the *B*s and *C*s -- move from the left and right respectively, at a constant equal speed, *S m/s*. And suppose that at some moment the rightmost *B* is perfectly aligned with the middle *A*, and the leftmost *C* with the rightmost *A*: thus the edges of the rightmost *B* and leftmost *C* are exactly lined up. That is they are arranged as shown.

```
   A A A
 B B B
       C C C
```

Now consider the later time at which the rightmost *A* and *B* are aligned; since the speeds of the *B*s and *C*s are equal, at this moment the middle *A* will be aligned with the leftmost *C*. That is consider the moment when the blocks are configured thus.

```
 A A A
 B B B
   C C C
```

This motion requires the rightmost *B* to move one block -- a distance *L m* -- to the right, at a speed of *S m/s*, so it takes *L/S s*. And the same motion also requires the leftmost *C* to move from just to the right of the rightmost *B* into alignment with the middle *B*, a distance a distance of 2*L m*. So far so good, but now Zeno concludes that since the *C*s are moving at *S m/s*, the motion must also take 2*L/S s*. And hence 'half

the time [L/S] is equal to its double [$2L/S$]', since one and the same motion seems to take both times.

The unanimous verdict on Zeno is that he was hopelessly confused about relative velocity in this paradox. If the *B*s are moving with speed S m/s to the right with respect to the *A*s and if the *C*s are moving with speed *S m/s* to the left with respect to the *A*s then the *C*s are moving with speed $S+S = 2S$ *m/s* to the left with respect to the *B*s. And so, as expected it takes the *C*s $2L/2S = L/S$ *s* to complete the motion after all.

This resolution notwithstanding, recent philosophers have attempted to put a new spin on Zeno's argument (it's arguable whether Zeno himself had anything like what follows in mind). This argument opposed the view that space and time are 'quantized', composed of smallest finite parts. Suppose they are and that *L m* is the 'quantum' of length and that the two moments considered are separated by a single quantum of time. Now something strange has happened, for the rightmost *B* and the leftmost *C* have clearly passed each other during the motion, and yet there is no moment at which they are level: since the two moments are separated by the smallest possible time, there can be no moment between them -- it would be a time smaller than the smallest time from the two moments we considered. Conversely, if one insisted that if they pass then there must be a moment when they are level, then it shows that cannot be a shortest finite interval -- whatever it is, just run this argument against it. However, why should one insist on this assumption? The problem is that one naturally imagines quantized space as being like a chess board, on which the chess pieces are frozen during each quantum of time. Then one wonders when the red queen, say, gets from one square to the next, or how she gets past the white queen without being level with her. But the analogy is misleading. It is better to think of quantized space as a giant matrix of lights that holds some pattern of illuminated lights for each quantum of time. In this analogy a lit bulb represents the presence of an object: for instance a series of bulbs in a line lighting up in sequence represent a body moving in a straight line. In this case there is no temptation to ask when the light 'gets' from one bulb to the next -- or in analogy how the body moves from one location to the next.

4. Two More Paradoxes

4.1 The Paradox of Place

> Zeno's difficulty demands an explanation; for if everything that exists has a place, place too will have a place, and so on *ad infinitum*. (Aristotle *Physics*, 209a23)

When he sets up his theory of place -- the crucial spatial notion in his theory of motion -- Aristotle lists various theories and problems that his predecessors, including Zeno, have formulated on the subject. One can again see here a problem for pluralism, for the second step of the argument concludes that there are many places. It is perhaps a little hard to feel the full force of the conclusion, for why should there not be an infinite series of places of places of places of …? Presumably the worry would be greater for someone who (like Aristotle) believed that there could not be an actual infinity of things, for the argument seems to show that there are. But certainly today we need have no such qualms; there seems nothing

problematic with an actual infinity of places; indeed, it seems very natural to think that every point of space is a distinct place, even if there are an infinity of points.

The only other way one might find the regress troubling is if one had reason to suppose that objects must have ‘absolute’ places, in the sense that there is always a unique answer to the question ‘where is it’? For example, where am I as I write? If the paradox is right then I’m in my place, and I’m also in my place’s place, and my place’s place’s place, and my … . Since I’m in all these places any might seem an appropriate answer to the question. But why think that there must be a unique answer to the question? Why shouldn’t I have many locations? At my desk, in my apartment, in Chicago, Illinois, USA, North America, the Earth, Solar System … . (In fact there is a reason that Aristotle might have had this concern about Zeno’s argument, for in his theory of motion, the natural motion of a body is determined by the relation of its place to the center of the universe: an account that only makes sense if bodies can be attributed a unique place. Interestingly, Newton, in the Scholium to the principal Definitions in Book I of his *Principia*, gives an argument along similar lines: he assumes that every body has a unique, absolute velocity, and argues that only a fixed matter-independent, ‘absolute’ space will provide such uniqueness. That said, there is no evidence either that Zeno had this kind of argument in mind, or that Newton was influenced by Aristotle in this regard.)

4.2 The Grain of Millet

> … Zeno’s reasoning is false when he argues that there is no part of the millet that does not make a sound; for there is no reason why any part should not in any length of time fail to move the air that the whole bushel moves in falling. (Aristotle *Physics*, 250a19)

This argument is a Parmenidean argument against the reliability of the senses. It goes like this: if you drop a sack of millet on the floor then you hear a loud thud; but this noise is the result of the noise made by every grain of millet in the sack; and the result of the noise made by every part of every grain; therefore every part of every grain makes a noise as it hits the ground. But now consider dropping a tiny part of a grain; we know that we won’t hear it. Therefore our sense of hearing is deceptive -- there are noises it cannot hear -- and so we should not trust it. Aristotle’s response seems to be that the part would not move as much air as the sack, but the paradox is not that the part should make as much noise as the sack, but that it should make some noise. A better reply is surely that not every disturbance in the air is audible by us: that a measuring instrument is unreliable over some range is no argument that it is unreliable over every range.

5. Zeno’s Influence on Philosophy

In this final section we should consider briefly the impact that Zeno has had on various philosophers; a search of the literature will reveal that these debates continue.

- The Pythagoreans: For the first half of the Twentieth century the majority reading -- following

Tannery (1885) -- of Zeno held that his arguments were directed against a technical doctrine of the Pythagoreans. According to this reading they held that all things were composed of elements that had the properties of a unit number, a geometric point and a physical atom: this kind of position would fit with their doctrine that reality is fundamentally mathematical. However, in the middle of the century a series of commentators (Vlastos, 1967, summarizes the argument and contains references) forcefully argued that Zeno’s target was instead a common sense understanding of plurality and motion -- one grounded in familiar geometrical notions -- and indeed that the doctrine was not a major part of Pythagorean thought. We have implicitly assumed that these arguments are correct in our readings of the paradoxes. That said, Tannery’s interpretation still has its defenders (see e.g., Matson 2001).

- The Atomists: Aristotle (*On Generation and Corruption* 316b34) claims that our third argument -- the one concerning complete divisibility -- was what convinced the atomists that there must be smallest, indivisible parts of matter. See Abraham (1972) for a further discussion of Zeno’s connection to the atomists.

- Temporal Becoming: In the early part of the Twentieth century several influential philosophers attempted to put Zeno’s arguments to work in the service of a metaphysics of ‘temporal becoming’, the (supposed) process by which the present comes into being. Such thinkers as Bergson (1911), James (1911, Ch 10 -11) and Whitehead (1929) argued that Zeno’s paradoxes show that space and time are not structured as a mathematical continuum: they argued that the way to preserve the reality of motion was to deny that space and time are composed of points and instants. However, we have clearly seen that the tools of standard modern mathematics are up to the job of resolving the paradoxes, so no such conclusion seems warranted: if the present indeed ‘becomes’, there is no reason to think that the process is not captured by the continuum.

- Applying the Mathematical Continuum to Physical Space and Time: Following a lead given by Russell (1929, 182-198), a number of philosophers -- most notably Grünbaum (1967) -- took up the task of showing how modern mathematics could solve all of Zeno’s paradoxes; their work has thoroughly influenced our discussion of the arguments. What they realized was that a purely mathematical solution was not sufficient: the paradoxes not only question abstract mathematics, but also the nature of physical reality. So what they sought was an argument not only that Zeno posed no threat to the mathematics of infinity but also that that mathematics correctly describes objects, time and space. The idea that a mathematical law -- say Newton’s law of universal gravity -- may or may not correctly describe things is familiar, but some aspects of the mathematics of infinity -- the nature of the continuum, definition of infinite sums and so on -- seem so basic that it may be hard to see at first that they too apply contingently. But surely they do: nothing guarantees *a priori* that space has the structure of the continuum, or even that parts of space add up according to Cauchy’s definition. (Salmon offers a nice example to help make the point: since alcohol dissolves in water, if you mix the two you end up with less than the sum of their volumes, showing that even ordinary addition is not applicable to every kind of system.) Our belief that the mathematical theory of infinity describes space and time is justified to the extent that the laws of physics assume that it does, and to the extent that those laws are themselves

confirmed by experience. While it is true that almost all physical theories assume that space and time do indeed have the structure of the continuum, it is also the case that quantum theories of gravity likely imply that they do not. While no one really knows were this research will ultimately lead, it is quite possible that space and time will turn out, at the most fundamental level, to be quite unlike the mathematical continuum that we have assumed here.

One should also note that Grünbaum took the job of showing that modern mathematics describes space and time to involve something rather different from arguing that it is confirmed by experience. The dominant view at the time (though not at present) was that scientific terms had meaning insofar as they referred directly to objects of experience -- such as ‘1 m ruler’ -- or, if they referred to ‘theoretical’ rather than ‘observable’ entities -- such as ‘a point of space’ or ‘1/2 of 1/2 of … 1/2 a racetrack’ -- then they obtained meaning by their logical relations -- via definitions and theoretical laws -- to such observation terms. Thus Grünbaum undertook an impressive program to give meaning to all terms involved in the modern theory of infinity, interpreted as an account of space and time.

- Supertasks: A further strand of thought concerns what Black (1950-51) dubbed ‘infinity machines’. Black and his followers wished to show that although Zeno’s paradoxes offered no problem to mathematics, they showed that after all mathematics was not applicable to space, time and motion. Most starkly, our resolution to the Dichotomy and Achilles assumed that the complete run could be broken down into an infinite series of half runs, which could be summed. But is it really possible to complete any infinite series of actions: to complete was is known as a ‘supertask’? If not, and assuming that Atalanta and Achilles can complete their tasks, their complete runs cannot be correctly described as an infinite series of half-runs, although modern mathematics would so describe them. What infinity machines are supposed to establish is that an infinite series of tasks cannot be completed -- so any completable task cannot be broken down into an infinity of smaller tasks, whatever mathematics suggests.

- Non-standard analysis: Finally, we have seen how to tackle the paradoxes using the resources of mathematics as developed in the Nineteenth century. For a long time it was considered one the great virtues of this system that it finally showed how to do without infinitesimal quantities, smaller than any finite number but larger than zero. (Newton’s calculus for instance effectively made use of such numbers, treating them sometimes as zero and sometimes as finite; the problem with such an approach is that how to treat the numbers is a matter of intuition not rigor.) However, in the Twentieth century Robinson showed how to introduce infinitesimal numbers into mathematics: this is the system of ‘non-standard analysis’ (the familiar system of real numbers, given a rigorous foundation by Dedekind, is by contrast just ‘analysis’). And it has been shown by McLaughlin (1992, 1994) that Zeno’s paradoxes can also be resolved in non-standard analysis; they are no more argument against non-standard analysis than the standard mathematics we have assumed here. It should be emphasized however that -- contrary to McLaughlin’s suggestions -- there is no need for non-standard analysis to solve the paradoxes: either system is equally successful. (The construction of non-standard analysis does however raise a further question about

the applicability of analysis to physical space and time: it seems plausible that all physical theories can be formulated in either terms, and so as far as our experience extends both seem equally confirmed. But they cannot both be true of space and time: either space has infinitesimal parts or it doesn't.)

Further Readings

After the relevant entries in this encyclopedia, the place to begin any further investigation is Salmon (2001), which contains some of the most important articles on Zeno up to 1970, and an impressively comprehensive bibliography of works in English in Twentieth Century.

One might also take a look at Huggett (1999, Ch. 3) for further source passages and discussion. For a discussion of the influence of the paradoxes through the Nineteenth century there is Cajori. For introductions to the mathematical ideas behind the modern resolutions, the Appendix to Salmon (2001) is a good start; Russell (1919) and Courant *et al*. (1996, Chs. 2 and 9) are also both wonderful sources. Finally, three collections of original sources for Zeno's paradoxes: Lee (1936) contains everything known, Kirk *et al* (1983, Ch. 9) contains a great deal of material (in English and Greek) with useful commentaries, and Cohen *et al*. (1995) also has the main passages.

Bibliography

- Abraham, W. E., 1972, 'The Nature of Zeno's Argument Against Plurality in DK 29 B I', *Phronesis* 17: 40-52
- Aristotle, 1984, ' On Generation and Corruption', A. A. Joachim (trans), in *The Complete Works of Aristotle*, J. Barnes (ed.), Princeton: Princeton University Press
- Aristotle, 1984, 'Physics', W. D. Ross(trans), in *The Complete Works of Aristotle*, J. Barnes (ed.), Princeton: Princeton University Press
- Belot, G. and Earman, J., 2001, 'Pre-Socratic Quantum Gravity', in *Physics Meets Philosophy at the Planck Scale: Contemporary Theories in Quantum Gravity*, C. Callender and N. Huggett (eds), Cambridge: Cambridge University Press
- Bergson, H., 1911, *Creative Evolution*, A. Mitchell (trans.), New York: Holt, Reinhart and Winston
- Black, M., 1950, 'Achilles and the Tortoise', *Analysis*, 11: 91-101
- Cohen, S. M., Curd, P. and Reeve, C. D. C. (eds), 1995, *Readings in Ancient Greek Philosophy From Thales to Aristotle*, Indianapolis/Cambridge: Hackett Publishing Co. Inc.
- Courant, R., Robbins, H., and Stewart, I., 1996, *What is Mathematics? An Elementary Approach to Ideas and Methods*, 2nd Edition, New York/Oxford: Oxford University Press
- Diogenes Laertius, 1983, 'Lives of Famous Philosophers', p.273 of *The Presocratic Philosophers: A Critical History with a Selection of Texts*, 2nd Edition, G. S. Kirk, J. E. Raven and M. Schofield (eds), Cambridge, UK: Cambridge University Press
- Grünbaum, A., 1967, *Modern Science and Zeno's Paradoxes*, Middletown: Connecticut Wesleyan

University Press
- Huggett, N. (ed.), 1999, *Space from Zeno to Einstein: Classic Readings with a Contemporary Commentary*, Cambridge, MA: MIT Press
- James, W., 1911, *Some Problems of Philosophy*, New York: Longmans, Green & Co.
- Lee, H. D. P. (ed.), 1967, *Zeno of Elea*, Amsterdam: Adof Hakkert
- Kirk, G. S., Raven J. E. and Schofield M. (eds), 1983, *The Presocratic Philosophers: A critical History with a Selection of Texts*, 2nd Edition, Cambridge, UK: Cambridge University Press
- McLaughlin, W. I., and Miller, S. L., 1992, 'An Epistemological Use of Nonstandard Analysis to Answer Zeno's Objections against Motion', *Synthese*, 92: 371-384
- McLaughlin, W. I., 1994, 'Resolving Zeno's Paradoxes', *Scientific American*, November: 84-89
- Matson, W. I., 2001, 'Zeno Moves!', in *Essays in Ancient Greek Philosophy VI: Before Plato*, A. Preus (ed.), Albany: SUNY Press
- Newton, I., 1999, *The Principia: Mathematical Principles of Natural Philosophy*, I. B. Cohen and A. M. Whitman (trans.), Berkeley: University of California Press
- Plato, 1997, 'Parmenides', M. L. Gill and P. Ryan (trans), in *Plato: Complete Works*, J. M. Cooper (ed.), Indianapolis/Cambridge: Hackett Publishing Co. Inc.
- Russell, B., 1929, *Our Knowledge of the External World*, New York: W. W. Norton & Co. Inc.
- Russell, B., 1919, *Introduction to Mathematical Philosophy*, London: George Allen and Unwin Ltd
- Salmon, W. C., 2001, *Zeno's Paradoxes*, 2nd Edition, Indianapolis: Hackett Publishing Co. Inc.
- Simplicius(a), 1995, 'On Aristotle's Physics', in *Readings in Ancient Greek Philosophy From Thales to Aristotle*, S. M. Chohen, P. Curd and C. D. C. Reeve (eds), Indianapolis/Cambridge: Hackett Publishing Co. Inc. 58-59
- Simplicius(b), 1989, *On Aristotle's Physics 6*, D. Konstan (trans.), London: Gerald Duckworth & Co. Ltd
- Tannery, P., 1885, 'Le Concept Scientifique du continu: Zenon d'Elee et Georg Cantor', *Revue Philosophique de la France et de l'Etranger*, 20: 385
- Vlastos, G., 1967, 'Zeno of Elea', in *The Encyclopedia of Philosophy*, P. Edwards (ed.), New York: The Macmillan Co. and The Free Press
- Whitehead, A. N., 1929, *Process and Reality*, New York: The Macmillan Co.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle | atomism: ancient | Cantor, Georg | Dedekind, Richard | infinity | Parmenides | Plato | Pythagoras | quantum mechanics | Simplicius | space and time: supertasks | time

Acknowledgement

This entry is dedicated to the late Wesley Salmon, who did so much to educate philosophers about the significance of Zeno’s paradoxes. Those familar with his work will see that this discussion owes a great deal to him; I hope that he would find it satisfactory. This material is based upon work supported by National Science Foundation Grant SES-0004375.

Nick Huggett
huggett@uic.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 30, 2002
Content last modified: April 30, 2002

Supertasks

Supertasks have posed problems for philosophy since the time of Zeno of Elea. The term ‘supertask’ is new but it designates an idea already present in the formulation of the old motion paradoxes of Zeno, namely the idea of an infinite number of actions performed in a finite amount of time. The main problem lies in deciding what follows from the performance of a supertask. Some philosophers have claimed that what follows is a contradiction and that supertasks are, therefore, logically impossible. Others have denied this conclusion, and hold that the study of supertasks can help us improve our understanding of the physical world, or even our theories about it.

Section 1: What is a Supertask

1.1: Definitions

A supertask may be defined as an infinite sequence of actions or operations carried out in a finite interval of time. The terms ‘action’ and ‘operation’ must not be understood in their usual sense, which involves a human agent. Human agency may be involved but it is not necessary. To show this, let us see how actions can be characterised precisely without any references to man. We will assume that at each instant of time the state of the world relevant to a specific action can be described by a set S of sentences. Now an action or operation applied to a state of the world results in a change in that state, that is, in the set S corresponding to it. Consequently, an arbitrary action a will be defined (Allis and Koetsier [1995]) as a change in the state of the world by which the latter changes from state S before the change to state a(S) after it. This means that an action has a beginning and an end, but does not entail that there is a finite lapse of time between them. For instance, take the case of a lamp that is on at t = 0 and remains so until t = 1, an instant at which it suddenly goes off. Before t =1 the state of the lamp (which is the only relevant portion of the world here) can be described by the sentence ‘lamp on’, and after t =1 by the sentence ‘lamp off’, without there being a finite lapse of time between the beginning and the end of the action. Some authors have objected to this consequence of the definition of action, and they might be right if we were dealing with the general philosophical problem of change. But we need not be concerned with those objections at this stage, since in the greatest majority of the relevant supertasks instantaneous actions (i.e. actions without any duration) can be replaced by actions lasting a finite amount of time without affecting the analysis at any fundamental point.

There is a particular type of supertask called *hypertasks*. A hypertask is a non-numerable infinite sequence of actions or operations carried out in a finite interval of time. Therefore, a supertask which is not a hypertask will be a numerable infinite sequence of actions or operations carried out in a finite interval of time. Finally, a task can be defined as a finite sequence of actions or operations carried out in a finite interval of time.

1.2: The Philosophical Problem of Supertasks

To gain a better insight into the fundamental nature of the philosophical problem posed by supertasks, consider the distinction between tasks in general (finite sequences of actions of the type (a_1, a_2, a_3, . . . , a_n)) and one particular type of supertasks, namely those consisting of an infinite sequence of actions of the type (a_1, a_2, a_3, . . . , a_n, . . .) and thus having the same type of order as the natural order of positive integers: 1, 2, 3, . . . , n, . . . (it is customary to denote this type of order with letter ‘*w*’ and so the related supertasks can be called supertasks of type *w*).

In the case of a task T = (a_1, a_2, a_3, . . . , a_n) it is natural to say that T is applicable in state S if:

> a_1 is applicable to S,
> a_2 is applicable to a_1(S),
> a_3 is applicable to a_2(a_1(S)),
> . . . , and,
> a_n is applicable to $a_{n-1}(a_{n-2}(. . . (a_2(a_1(S))). . .))$.

The successive states of the world relevant to task T can be defined by means of the finite sequence of sets of sentences:

$$S, \quad a_1(S), \quad a_2(a_1(S)), \quad a_3(a_2(a_1(S))), \quad \ldots \quad , a_n(a_{n-1}(a_{n-2} (\ldots (a_2(a_1(S)))\ldots))),$$

whose last term will therefore describe the relevant state of the world after the performance of T. Or, equivalently, the state resulting from applying T to S will be T(S) =

$$a_n(a_{n-1}(a_{n-2} (\ldots (a_2(a_1(S)))\ldots))).$$

Now take the case of a supertask T = (a_1, a_2, a_3, . . . , a_n, . . .). Let us give the name T_n to the task which consists in performing the first n actions of T. That is, T_n = (a_1, a_2, a_3, . . . , a_n). Now it is natural to say that T is applicable in state S if T_n is applicable in S for each natural number n, and, obviously,

$$T_n(S) \quad = \quad a_n(a_{n-1}(a_{n-2} (\ldots (a_2(a_1(S)))\ldots))).$$

The successive states of the world relevant to supertask T can be described by means of the infinite sequence of sets of sentences:

$$S, \quad T_1(S), \quad T_2(S), \quad \ldots \quad , T_n(S), \quad \ldots$$

A difficulty arises, however, when we want to specify the set of sentences which describe the relevant state of the world after the performance of supertask T, because the infinite sequence above lacks a final term. Put equivalently, it is difficult to specify the relevant state of the world resulting from the application of supertask T to S because there seems to be no final state resulting from such an application. This inherent difficulty is increased by the fact that, by definition, supertask T is performed in a finite time, and so there must exist one first instant of time t* at which it can be said that the performance happened. Now notice that the world must naturally be in a certain specific state at t*, which is the state resulting from the application of T, but that, nevertheless, we have serious trouble to specify this state, as we have just seen.

1.3: Supertask: A Fuzzy Concept

Since we have defined supertasks in terms of actions and actions in terms of changes in the state of the world, there is a basic indeterminacy regarding what type of processes taking place in time should be considered supertasks, which is linked to the basic indeterminacy that there is regarding which type of sets of sentences are to be allowed in descriptions of the state of the world and which are not. For this reason, there are some processes that would be regarded as supertasks by virtually every philosopher and some about which opinions differ. For an instance of the first sort of process, take the one known as ‘Thomson's lamp’. Thomson's lamp is basically a device consisting of a lamp and a switch set on an electrical circuit. The switch can be in one of just two positions and the lamp has got to be lit -when the switch is in position ‘on’- or else dim -when the switch is in position ‘off’. Assume that initially (at t = 12 A.M., say) the lamp is dim and that it is thenceforth subject to the following infinite sequence of actions: when half of the time remaining until t* = 1 P.M. has gone by, we carry out the action a_1 of turning the switch into position ‘on’ and, as a result, the lamp is lit (a_1 is thus performed at t = 1/2 P.M.); when half the time between the performance of a_1 and t* = 1 P.M. has gone by, we carry out action a_2 of turning to the switch into position ‘off’ and, as a result, the lamp is dim (a_2 is thus performed at t = 1/2 + 1/4 P.M.); when half the time between the performance of a_2 and t* = 1 P.M. has gone by, we carry out the action of turning the switch into position ‘on’ and, as a result, the lamp is lit (a_3 is thus performed at t = 1/2 + 1/4 + 1/8), and so on. When we get to instant t* = 1 P.M. we will have carried out an infinite sequence of actions, that is, a supertask T = (a_1, a_2, a_3, . . . , a_n, . . .). If, for the sake of simplicity, we are only concerned about the evolution of the lamp (not the switch) the state of the world relevant to the description of our supertask admits of only two descriptions, one through the unitary set of sentences {lamp lit} and the other through the set {lamp dim}.

As an instance of the second sort of processes we referred to above, those about which no consensus has been reached as to whether they are supertasks, we can take the process which is described in one of the forms of Zeno's dichotomy paradox. Suppose that initially (at t = 12 A.M., say) Achilles is at point A (x = 0) and moving in a straight line, with a constant velocity v = 1 km/h, towards point B (x = 1), which is 1 km. away from A. Assume, in addition, that Achilles does not modify his velocity at any point. In that case, we can view Achilles's run as the performance of a supertask, in the following way: when half the time until t* = 1 P.M. has gone by, Achilles will have carried out the action a_1 of going from point x = 0 to point x = 1/2 (a_1 is thus performed in the interval of time between t =12 A.M. and t = 1/2 P.M.), when half the time from the end of the performance of a_1 until t* = 1 P.M. will have elapsed, Achilles will have carried out the action a_2 of going from point x = 1/2 to point x = 1/2 + 1/4 (a_2 is thus performed in the interval of time between t = 1/2 P.M. and t = 1/2 + 1/4 P.M.), when half the time from the end of the performance of a_2 until t* = 1 P.M. will have elapsed, Achilles will have carried out the action a_3 of going from point x = 1/2 + 1/4 to point x = 1/2 + 1/4 + 1/8 (a_3 is thus performed in the interval of time between t = 1/2 + 1/4 P.M. and t = 1/2 + 1/4 + 1/8 P.M.), and so on. When we get to instant t* = 1 P.M., Achilles will have carried out an infinite sequence of actions, that is, a supertask T = (a_1, a_2, a_3, . . . , a_n, . . .), provided we allow the state of the world relevant for the description of T to be specified, at any arbitrary instant, by a single sentence: the one which specifies Achilles's position at that instant. Several

philosophers have objected to this conclusion, arguing that, in contrast to Thomson's lamp, Achilles's run does not involve an infinity of actions (acts) but of pseudo-acts. In their view, the analysis presented above for Achilles's run is nothing but the breakdown of one process into a numerable infinity of subprocesses, which does not make it into a supertask. In Allis and Koetsier's words, such philosophers believe that a set of position sentences is not always to be admitted as a description of the state of the world relevant to a certain action. In their opinion, a relevant description of a state of the world should normally include a different type of sentences (as is the case with Thomson's lamp) or, in any case, more than simply position sentences.

Section 2: On The Conceptual Possibility Of Supertasks

In section 1.2 I have pointed out and illustrated the fundamental philosophical problem posed by supertasks. Obviously, one will only consider it a problem if one deems the concepts employed in its formulation acceptable. In fact, some philosophers reject them, because they regard the very notion of supertask as problematic, as leading to contradictions or at least to insurmountable conceptual difficulties. Among these philosophers the first well-known one is Zeno of Elea.

2.1: Zeno's Dichotomy Paradox

Consider the dichotomy paradox in the formulation of it given in section 1.3. According to Zeno, Achilles would never get to point B ($x = 1$) because he would first have to reach the mid point of his run ($x = 1/2$), after that he would have to get to the mid point of the span which he has left ($x = 1/2 + 1/4$), and then to the mid point of the span which is left ($x = 1/2 + 1/4 + 1/8$), and so on. Whatever the mid point Achilles may reach in his journey, there will always exist another one (the mid point of the stretch that is left for him to cover) that he has not reached yet. In other words, Achilles will never be able to reach point B and finish his run. According to Owen (Owen [1957-58]), in this as well as in his other paradoxes, Zeno was concerned to show that the Universe is a simple, global entity which is not comprised of different parts. He tried to demonstrate that if we take to making divisions and subdivisions we will obtain absurd results (as in the dichotomy case) and that we must not yield to the temptation of breaking up the world. Now the notion of supertask entails precisely that, division into parts, as it involves breaking up a time interval into successive intervals. Therefore, supertasks are not feasible in the Zenonian world and, since they lead to absurd results, the notion of supertask itself is conceptually objectionable.

In stark contrast to Zeno, the dichotomy paradox is standardly solved by saying that the successive distances covered by Achilles as he progressively reaches the mid points of the spans he has left to go through --- 1/2, 1/4, 1/8, 1/16, . . . --- form an infinite series $1/2 + 1/4 + 1/8 + 1/16 + \ldots$ whose sum is 1. Consequently, Achilles will indeed reach point B ($x = 1$) at $t^* = 1$ P.M. (which is to be expected if he travels with velocity $v = 1$ km/h, as has been assumed). Then there is no problem whatsoever in splitting up his run into smaller sub-runs and, so, no inherent problem about the notion of supertask. An objection can be made, however, to this standard solution to the paradox: it tells us where Achilles is at each instant

but it does not explain where Zeno's argument breaks down. Importantly, there is another objection to the standard solution, which hinges on the fact that, when it is claimed that the infinite series 1/2 + 1/4 + 1/8 + 1/16 + . . . adds up to 1, this is substantiated by the assertion that the sequence of partial sums 1/2, 1/2 + 1/4, 1/2 + 1/4 + 1/8, . . . has limit 1, that is, that the difference between the successive terms of the sequence and number 1 becomes progressively smaller than any positive integer, no matter how small. But it might be countered that this is just a patch up: the infinite series 1/2 + 1/4 + 1/8 + . . . seems to involve infinite sums and thus the performance of a supertask, and the proponent of the standard solution is in fact presupposing that supertasks are feasible just in order to justifiy that they are. To this the latter might reply that the assertion that the sum of the series is 1 presupposes no infinite sum, since, by definition, the sum of a series is the limit to which its partial (and so finite) sums approach. His opponent can now express his disagreement with the response that the one who supports the standard solution is deducing a matter of fact (that Achilles is at $x = 1$ at t* = 1 P.M.) from a definition pertaining to the arithmetic of infinite series, and that it is blatantly unacceptable to deduce empirical propositions from mere definitions.

2.2: The Inverse Form Of The Dichotomy Argument

Before concluding our discussion of the arguments connected with Zeno's dichotomy paradox which have been put forward against the conceptual feasibility of supertasks, we should deal with the so-called inverse dichotomy of Zeno, which can also be formulated as a supertask, but whose status as a logical possibility seems to some philosophers to be even more doubtful than that of the direct version expounded in section 2.1. The process involved in the paradox of inverse dichotomy admits of a supertask kind of description, as follows. Suppose that at t = 12 A.M. Achilles is at point A ($x = 0$) and wishes to do the action of reaching point B ($x = 1$). In order to do this action he must first of all go from point A to the mid point b_1 ($x = 1/2$) of the span AB that he wishes to cover. In order to do this, he must in turn first do the action of going from point A to the mid point b_2 ($x = 1/4$) of the span Ab_1 that he wishes to cover, and so on . In order to reach B, Achilles will have to accomplish an infinite sequence of actions, that is, a supertask T* = (. . . , a_n, . . . , a_3, a_2, a_1), provided we allow the state of the world relevant to the description of T* to be specified, at a given arbitrary instant, by a single sentence, the one specifying Achilles's position at that instant. Notice in the first place that T* has the same type of order as the natural order of negative integers: . . . , -n, . . . , -3, -2, -1 (such order type is usually denoted with the expression '*w**' and the related supertasks can therefore be called supertasks of type *w**). The philosophical problem connected with supertasks of type *w*, already discussed in section 1.2 above, does not arise now because the set of sentences which describes the relevant state of the world after the performance of supertask T* is obviously a_1(S), with S the set of sentences describing the initial relevant state of the world. But as the successive states of the world after S in relation to T* can be described by means of the infinite sequence of sets of sentences . . . , a_n(S), . . . , a_3(S), a_2(S), a_1(S), some philosophers think it puzzling and unacceptable that the initial set of sentences in that sequence cannot be specified. This really means that we cannot specify which is the action in supertask T* that should be carried out first and that we consequently ignore how to begin. Isn't that proof enough that supertasks of type *w** are impossible? Chihara (1965), for example, says that Zeno's inverse dichotomy is even more problematic than the direct one, since Achilles is supposed to be capable of doing something akin to counting the

natural numbers in reverse order. In his opinion, it is just as impossible for Achilles to start his run -- if viewed as a supertask of type *w** -- as it is to start this reverse counting process.

2.3: On Thomson's Impossibility Arguments

Thomson (1954-55) was convinced that he could show supertasks to be logically impossible. To this end, he made up the lamp example analysed in section 1.3, since known as ‘Thomson's lamp’. Thomson argued that the analysis of the workings of his lamp leads to contradiction, and therefore the supertask involved is logically impossible. But then, to the extent that this supertask is representative of ‘genuine’ supertasks, all genuine supertasks are impossible. Thomson's argument is simple. Let us ask ourselves what the state of the lamp is at t* = 1 P.M. At that instant the lamp cannot be lit, the reason being the way we manipulate it: we never light the lamp without dimming it some time later. Nor can the lamp be dim, because even if it is dim initially, we light it and subsequently never dim it without lighting it back again some time later. Therefore,at t* = 1 P.M. the lamp can be neither dim nor lit. However, one of its functioning conditions is that it must be either dim or lit. Thus, a contradiction arises. Conclusion: Thomson's lamp or, better, the supertask consisting in its functioning is logically impossible. Now is Thomson's argument correct? Benacerraf (1962) detected a serious flaw in it. Let us in principle distinguish between the series of instants of time in which the actions a_i of the supertask are performed (which will be called the t-series) and the instant t* = 1 P.M., the first instant after the supertask has been accomplished. Thomson's argument hinges on the way we act on the lamp, but we only act on it at instants in the t-series, and so what can be deduced logically from this way of acting will apply only to instants in the t-series. As t* = 1 P.M. does not belong to the t-series, it follows that Thomson's supposed conclusion that the lamp cannot be lit at t* is fallacious, and so is his conclusion that it cannot be dim at t*. The conditions obtaining in the lamp problem only enable us to conclude that the lamp will be either dim or else lit but not both at t* = 1 P.M., and this follows from the fact that this exclusive disjunction was presupposed from the start to be true at each and every instant of time, independently of the way in which we could act on the lamp in the t-series of instants of time. What cannot be safely inferred is which one of these two states -dim or lit- the lamp will be in at t* = 1 P.M. or, alternatively, the state of the lamp at t* = 1 P.M. is not logically determined by what has happened before that instant. This consequence tallies with what was observed in section 2.1 about the fallacy committed by adherents to the standard solution against Zeno: they seek to deduce that at instant t* = 1 P.M. Achilles will be at point B from an analysis of the sub-runs performed by him before that instant, that is, they assume that Achilles's state at t* follows logically from his states at instants previous to t*, and in so assuming they make the same mistake as Thomson.

Thomson (1954-55) put forward one more argument against the logical possibility of his lamp. Let us assign to the lamp the value 0 when it is dim and the value 1 when it is lit. Then lighting the lamp means adding one unity (going from 0 to 1) and dimming it means subtracting one unity (going from 1 to 0). It thus seems that the final state of the lamp at t* = 1 P.M., after an infinite, and alternating, sequence of lightings (additions of one unity) and dimmings (subtractions of one unity), should be described by the infinite series 1-1+1-1+1. . . If we accept the conventional mathematical definition of the sum of a series, this series has no sum, because the partial sums 1, 1-1, 1-1+1, 1-1+1-1, . . . , etc. take on the values 1 and

0 alternatively, without ever approaching a definite limit that could be taken to be the proper sum of the series. But in that case it seems that the final state of the lamp can neither be dim (0) nor lit (1), which contradicts our assumption that the lamp is at all times either dim or lit. Benacerraf's (1962) reply was that even though the first, second, third, . . . , n-th partial sum of the series 1-1+1-1+1. . . does yield the state of the lamp after one, two, three, . . . , n actions a_i (of lighting or dimming), it does not follow from this that the final state of the lamp after the infinite sequence of actions a_i must of necessity be given by the sum of the series, that is, by the limit to which its partial sums progressively approach. The reason is that a property shared by the partial sums of a series does not have to be shared by the limit to which those partial sums tend. For instance, the partial sums of the series 0.3 + 0.03 + 0.003 + 0.0003 + . . . are 0.3, 0.3 + 0.03 = 0.33, 0.3 + 0.03 + 0.003 = 0.333,. . . , all of them, clearly, numbers less than 1/3; however, the limit to which those partial sums tend (that is, the sum of the original series) is 0.3333... , which is precisely the number 1/3.

2.4: On Black's Impossibility Argument

Another one of the classical arguments against the logical possibility of supertasks comes from Black (1950-51) and is constructed around the functioning of an infinity machine of his own invention. An infinity machine is a machine that can carry out an infinite number of actions in a finite time. Black's aim is to show that an infinity machine is a logical impossibility. Consider the case of one such machine whose sole task is to carry a ball from point A ($x = 0$) to point B ($x = 1$) and viceversa. Assume, in addition, that initially (at t = 12 A.M., say) the ball is at A and that the machine carries out the following infinite sequence of operations: when half the time until t* = 1 P.M. has gone by, it does the action a_1 of taking the ball from position A to position B (a_1 is thus carried out at t = 1/2 P.M.); when half the time between the performance of a_1 and t* = 1 P.M. has gone by, it does the action a_2 of taking the ball from position B to position A (a_2 is thus carried out at t = 1/2 + 1/4 P.M.); when half the time between the performance of a_2 and t* = 1 P.M. has gone by, the machine does the action a_3 of taking the ball from position A to position B (a_3 is thus performed in t = 1/2 + 1/4 + 1/8 P.M.), and so on. When we get to instant t* = 1 P.M. the machine will have carried out an infinite sequence of actions, that is, a supertask T = (a_1, a_2, a_3, . . . , a_n, . . .). The parallelism with Thomson's lamp is clear when it is realised that the ball in position A corresponds to the dim lamp and the ball in position B corresponds to the lit lamp. Nevertheless, Black believes that by assuming that at each instant the ball is either in A or else in B (and note that assuming this means that the machine transports the ball from A to B and viceversa instantaneously, but we need not be worried by this, since all that we are concerned with now is logical or conceptual possibility, not physical possibility), he can deduce, by a totally different route from Thomson's based on the symmetrical functioning of his machine, a contradiction regarding its state at t* = 1 P.M.. However, Benacerraf's criticisms also applies to Black's argument. In effect, the latter hinges on how the machine works, and as this has only been specified for instants of time previous to t* = 1 P.M., it follows that what can be logically inferred from the functioning of the machine is only applicable to those instants previous to t* = 1 P.M.. Black seeks to deduce a contradiction at t* = 1 P.M. but he fails at the same point as Thomson: whatever happens to the ball at t* = 1 P.M. cannot be a logical consequence of what has happened to it before. Of course, one can always specify the functioning of the machine for

instants t greater than or equal to 1 P.M. by saying that at all those instants the machine will not perform any actions at all, but that is not going to help Black. His argument is fallacious because he seeks to reach a logical conclusion regarding instant t* = 1 P.M. from information relative to times previous to that instant.. In the standard argument against Zeno's dichotomy one could similarly specify Achilles's position at t* = 1 P.M. saying, for instance, that he is at B ($x = 1$), but there is no way that this is going to get us a valid argument out of a fallacious one, which seeks to deduce logically where Achilles will be at t* = 1 P.M. from information previous to that instant of time.

2.5: Benacerraf's Critique and the Dichotomy Arguments

The cases dealt with above are examples of how Benacerraf's strategy can be used against supposed demonstrations of the logical impossibility of supertasks. We have seen that the strategy is based on the idea that

> (I) the state of a system at an instant t* is not a logical consequence of which states he has been in before t* (where by ‘state’ I mean ‘relevant state of the world’, see section 1.1)

and occasionally on the idea that

> (II) the properties shared by the partial sums of a series do not have to be shared by the limit to which those partial sums tend.

Since the partial sums of a series make up a succession (of partial sums), (II) may be rewritten as follows:

> (III) the properties shared by the terms of a succession do not have to be shared by the limit to which that succession tends.

If we keep (I), (II) and (III) well in mind, it is easy not to yield to the perplexing implications of certain supertasks dealt with in the literature. And if we do not yield to the perplexing results, we will also not fall into the trap of considering supertasks conceptually impossible. (III), for instance, may be used to show that it is not impossible for Achilles to perform the supertasks of the inverse and the direct dichotomy of Zeno. Take the case of the direct dichotomy: the limit of the corresponding succession of instants of time t_1, t_2, t_3, . . . at which each one of Achilles's successive sub-runs is finished can be the instant at which Achilles's supertask has been accomplished, even if such a supertask is not achieved at any one of the instants in the infinite succession t_1, t_2, t_3, . . . (all of this in perfect agreement with (III)).

2.6: Conclusion

As a corollary it may be said that supertasks do not seem to be intrinsically impossible. The contradictions that they supposedly give rise to may be avoided if one rejects certain unwarranted assumptions that are usually made. The main such assumption, responsible for the apparent conceptual

impossibility of supertasks, is that properties which are preserved after a finite number of actions or operations will likewise be preserved after an infinite number of them. But that is not true in general. For example, we saw in section 1.2 above that the relevant state of the world after the performance of a task T = $(a_1, a_2, \ldots, a_n)$ on the relevant state S was logically determined by T and by S (and was $a_n(a_{n-1}(a_{n-2}(\ldots(a_2(a_1(S)))\ldots))))$, but we have now learned that after the performance of a supertask T = $(a_1, a_2, a_3, \ldots, a_n, \ldots)$ it is not (that is the core of Benacerraf's critique). The same sort of uncritical assumptions seem to be in the origin of infinity paradoxes in general, in which certain properties are extrapolated from the finite to the infinite that are only valid for the finite, as when it is assumed that there must be more numbers greater than zero than numbers greater than 1000 because all numbers greater than 1000 are also greater than zero but not viceversa (Galileo's paradox). In conclusion, if some supertasks are paradoxical, it is not because of any inherent inconsistency of the notion of supertask. This opinion is adhered to by authors such as Earman and Norton (1996).

Section 3: On The Physical Possibility Of Supertasks

We have gone through several arguments for the conceptual impossibility of supertasks and counterarguments to these. Those who hold that supertasks are conceptually possible may however not agree as to whether they are also physically possible. In general, when this issue is discussed in the literature, by physical possibility is meant possibility in relation with certain broad physical principles, laws or ‘circumstances’ which seems to operate in the real world, at least as far as we know. But it is a well-known fact that authors do not always agree about which those principles, laws or circumstances are.

3.1: Kinematical Impossibility

In our model of Thomson's lamp we are assuming that at each moment the switch can be in just one of two set positions (‘off’, ‘on’). If there is a fixed distance d between them, then clearly, since the switch swings an infinite number of times from the one to the other from t = 12 A.M. until t* = 1 P.M., it will have covered an infinite distance in one hour. For this to happen it is thus necessary for the speed with which the switch moves to increase unboundedly during this time span. Grünbaum has taken this requirement to be physically impossible to fulfil. Grünbaum (1970) believes that there is a sort of physical impossibility of a purely kinematical nature (kinematical impossibility) and describes it in more precise terms by saying that a supertask is kinematically impossible if:

> a) At least one of the moving bodies travels at an unboundedly increasing speed,
>
> b) For some instant of time t*, the position of at least one of the moving bodies does not approach any defined limit as we get arbitrarily closer in time to t*.

It is clear then that the Thomson's lamp supertask, in the version presented so far, is kinematically (and eo ipso physically) impossible, since not only does the moving switch have to travel at a speed that increases unboundedly but also -because it oscillates between two set positions which are a constant distance d apart- its position does not approach any definite limit as we get closer to instant t* = 1 P.M., at which the supertask is accomplished. Nevertheless, Grünbaum has also shown models of Thomson's lamp which are kinematically possible. Take a look at Figure 1, in which the switch (in position 'on' there) is simply a segment AB of the circuit connecting generator G with lamp L. The circuit segment AB can shift any distance upwards so as to open the circuit in order for L to be dimmed. Imagine we push the switch successively upwards and downwards in the way illustrated in Figure 2, so that it always has the same velocity $v = 1$.

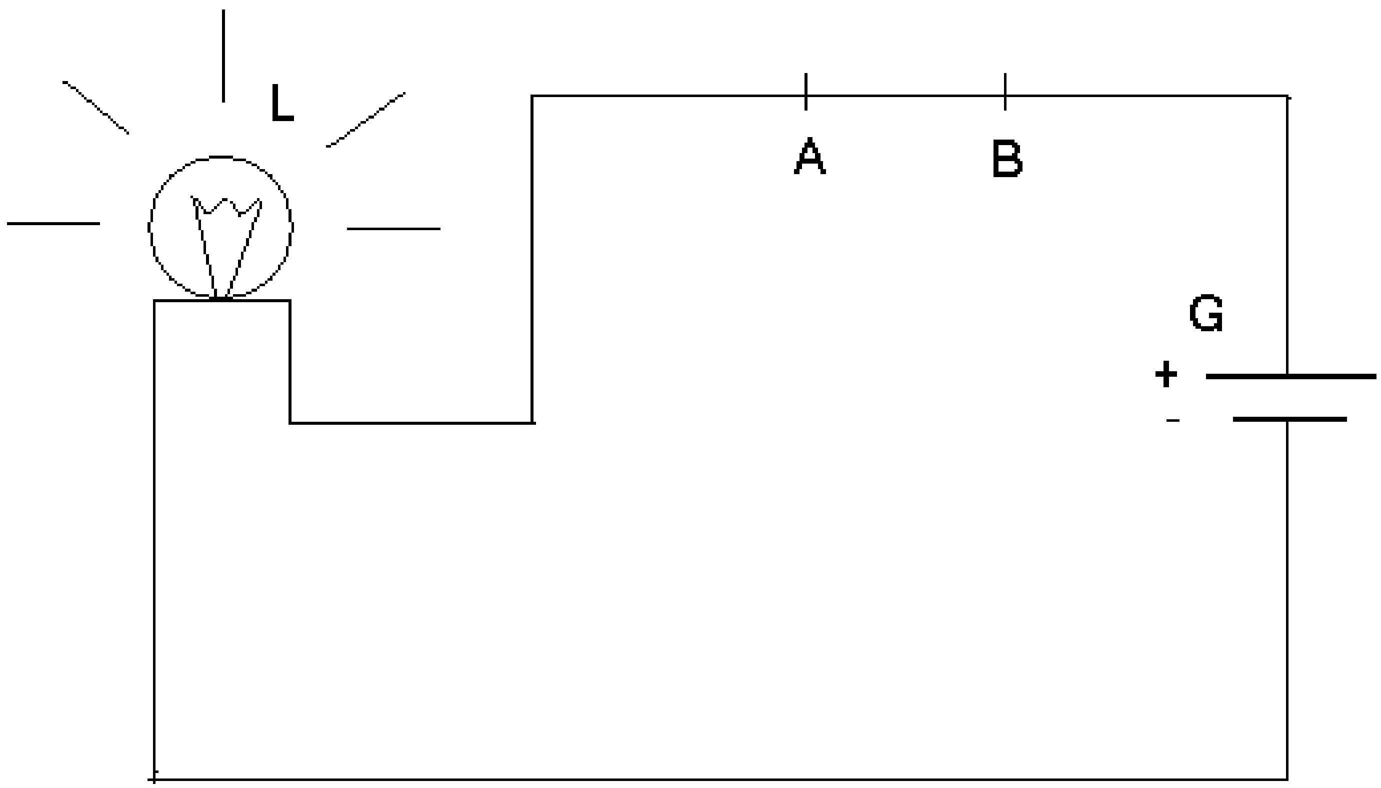


Figure 1

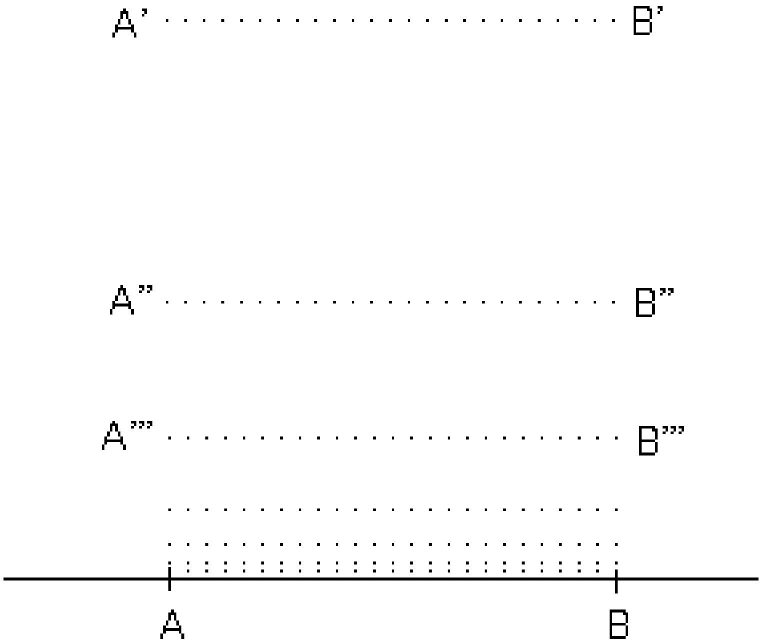


Figure 2

The procedure is the following. Initially (t = 0) the switch is in position $A'B'$ (lamp dim) a height of 0.2 above the circuit and moving downwards (at *v* = 1). At t = 0.2 it will be in position AB (lamp lit) and will begin moving upwards (*v* = 1). At t = 0.2 + 0.01 it will be in position $A''B''$ (lamp dim) and will begin moving downwards (*v* = 1). At t = 0.2 + 0.01 + 0.01 = 0.2 + 0.02 it will be in position AB (lamp lit) and will begin moving upwards (*v* = 1). At t = 0.2 + 0.02 + 0.001 it will be in position $A'''B'''$ (lamp dim), and so on. Obviously, between t = 0 and t* = 0.2 + 0.02 + 0.002 + . . . = 0.2222... = 2/9, the lamp is in the states ‘dim’ and ‘lit’ an infinite number of times, and so a supertask is accomplished. But this supertask is not kinematically impossible, because it has been so designed that the switch always moves with velocity *v* = 1 -- and, therefore, condition (a) for kinematical impossibility is not fulfilled -- and that, additionally, as we get closer to the limit time t* = 2/9 (the only one which could cause us any trouble) the switch approaches more and more a well-defined limit position, position AB (lamp lit) -and, therefore, condition b) for kinematical impossibility is not fulfilled either. Once the kinematical possibility has been established, what is the state of the lamp at t* = 2/9? What has been said so far does not enable us to give a determinate answer to this question (just as the obvious kinematical possibility of Achilles's supertask in the dichotomy paradox does not suffice to determine where Achilles will be at t* = 1 P.M.), but there exists a ‘natural’ result. It seems intuitively acceptable that the position the switch will occupy at t* = 2/9 will be position AB, and so the lamp will be lit at that instant. There is no mysterious asymmetry about this result. Figure 3 shows a model of Thomson's lamp with a switch that works according to exactly the same principles as before, but which will yield the ‘natural’ result that the lamp will in the end be dim at t* = 2/9. In effect, the switch will now finally end up in the ‘natural’ position AB at t* = 2/9 and will thereby bring about an electrical short-circuit that will make all the current in the generator pass through the cable on which the switch is set, leaving nothing for the more resistant path where the lamp is.

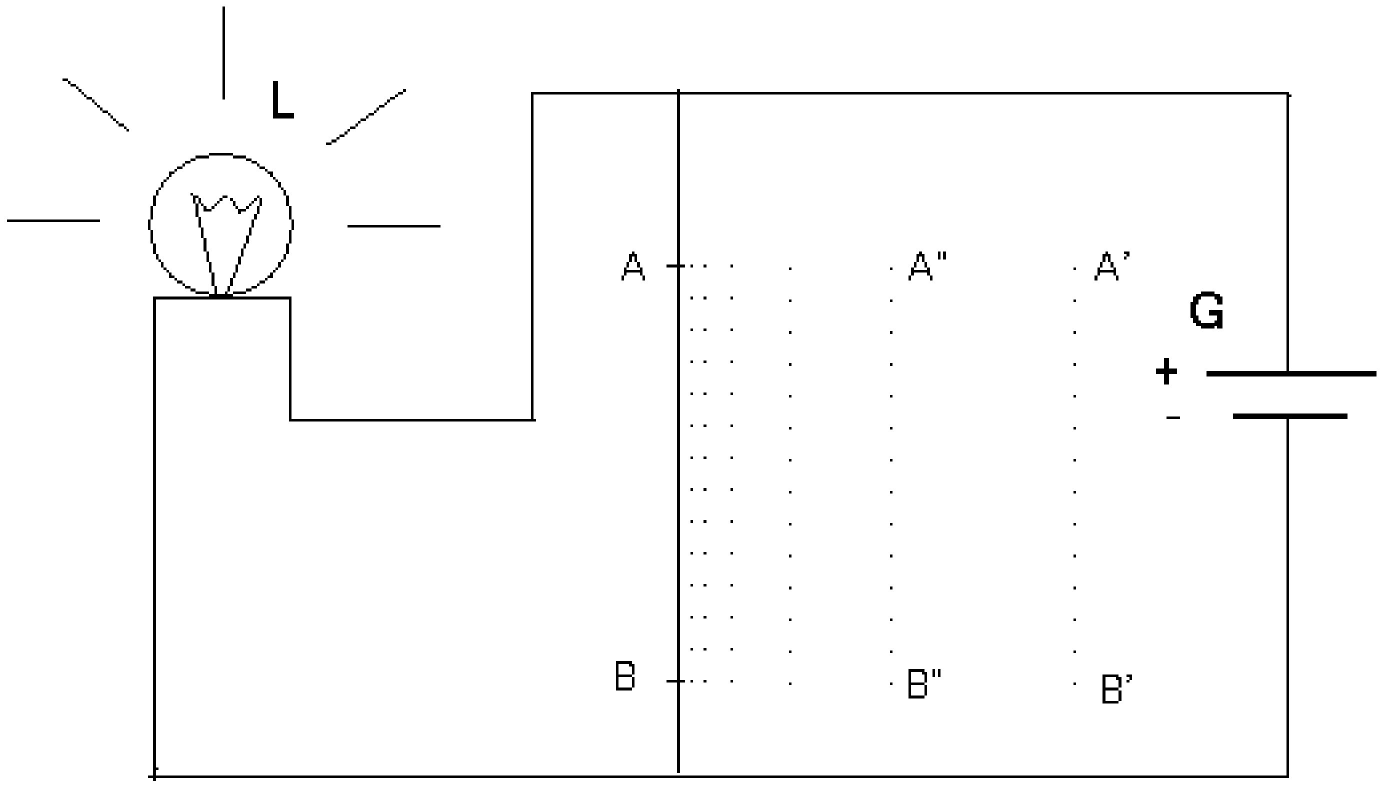


Figure 3

There are some who believe that the very fact that there exist Thomson's lamps yielding an intuitive result of ‘lamp lit’ when the supertask is accomplished but also other lamps whose intuitive result is ‘lamp dim’ brings up back to the contradiction which Thomson thought to have found originally. But we have nothing of that sort. What we do have is different physical models with different end-results. This does not contradict but rather corroborates the results obtained by Benacerraf: the final state is not logically determined by the previous sequence of states and operations. This logical indeterminacy can indeed become physical determinacy, at least sometimes, depending on what model of Thomson's lamp is employed.

A conspicuous instance of a supertask which is kinematically impossible is the one performed by Black's infinity machine, whose task it is to transport a ball from position A ($x = 0$) to position B ($x = 1$) and from B to A an infinite number of times in one hour. As with the switch in our first model of Thomson's lamp, it is obvious that the speed of the ball increases unboundedly (and so condition a) for impossibility is met), while at the same time, as we approach t* = 1 P.M., its position does not tend to any defined limit, due to the fact that it must oscillate continuously between two set positions A and B one unity distance apart from each other (and so also condition b) for impossibility is met).

3.2: The Principle Of Continuity and the Solution to the Philosophical Problem of Supertasks

Up to this point we have seen examples of supertasks which are conceptually possible and, among these, we have discovered some which are also physically possible. For the latter to happen we had to make

sure that at least some requirements were complied with which, plausibly, characterise the processes that can actually take place in our world. But some definitive statement remains to be made about the philosophical problem posed by supertasks: what the state of the world is after they have been accomplished. The principles of physical nature which have so far been appealed to do not enable us to pronounce on this matter. The question thus arises whether some new principle of a physical nature can be discovered which holds in the real world and is instrumental in answering the question what the state of the world will be after a supertask. That discovery would allow us to resolve a radical indeterminacy which still persists -the reader will remember that even in the case of Achilles's dichotomy supertask we were quite unable to prove that it would conclude with Achilles in point B ($x = 1$). In Section 2.1 we saw that such proof cannot be obtained by recourse to the mathematical theory of infinite series exclusively; why should it be assumed that this abstract theory is literally applicable to the physical universe? After all, amounts of money are added up applying ordinary arithmetic but, as Black reminded us, velocities cannot be added up according to ordinary arithmetic.

Since Benacerraf's critique, we know that there is no logical connection between the position of Achilles at t* = 1 P.M. and his positions at instants previous to t* = 1 P.M. Sainsbury [1988] has tried to bridge the gap opened by Benacerraf. He claims that this can be achieved by drawing a distinction between abstract space of a mathematical kind and physical space. No distinction between mathematical and physical space has to be made, however, to attain that goal; one need only appeal to a single principle of physical nature, which is, moreover, simple and general, namely, that the trajectories of material bodies are continuous lines. To put it more graphically, what this means is that we can draw those trajectories without lifting our pen off the paper. More precisely, that the trajectory of a material body is a continuous line means that, whatever the instant t, the limit to which the position occupied by the body tends as time approaches t coincides precisely with the position of the body at t. Moreover, the principle of continuity is highly plausible as a physical hypothesis: the trajectories of all physical bodies in the real world are in fact continuous. What matters is that we realise that, aided by this principle, we can now finally demonstrate that after the accomplishment of the dichotomy supertask, that is, at t* = 1 P.M., Achilles will be in point B ($x = 1$). We know, in fact, that as the time Achilles has spent running gets closer and closer to t* = 1 P.M., his position will approach point $x = 1$ more and more, or, equivalently, we know that the limit to which the position occupied by Achilles tends as time get progressively closer to t* = 1 P.M. is point B ($x = 1$). As Achilles's trajectory must be continuous, by the definition of continuity (applied to instant t = t* = 1 P.M.) we obtain that the limit to which the position occupied by Achilles tends as time approaches t* = 1 P.M. coincides with Achilles's position at t* = 1 P.M. Since we also know that this limit is point B ($x = 1$), it finally follows that Achilles's position at t* = 1 P.M. is point B ($x = 1$). Now is when we can spot the flaw in the standard argument against Zeno mentioned in section 2.1, which was grounded on the observation that the sequence of distances covered by Achilles (1/2, 1/2 + 1/4, 1/2 + 1/4 + 1/8, . . .) has 1 as its limit. This alone does not suffice to conclude that Achilles will reach point $x = 1$, unless it is assumed that if the distances run by Achilles have 1 as their limit, then Achilles will as a matter of fact reach $x = 1$, but assuming this entails using the principle of continuity. This principle affords us a rigorous demonstration of what, in any event, was already plausible and intuitively ‘natural’: that after having performed the infinite sequence of actions ($a_1, a_2, a_3, \ldots, a_n, \ldots$) Achilles will have reached point B ($x = 1$). In addition, now it is easy to show how, with a switch like the one in Figure 2, Thomson's lamp in Figure 1 will reach t* = 2/9 with its switch in position AB and will therefore be lit.

We have in fact already pointed out (3.1) that in this case, as we get closer to the limit time t* = 2/9, the switch indefinitely approaches a well-defined limit position -position AB. Due to the fact that the principle of continuity applies to the switch, because it is a physical body, this well-defined limit position must coincide precisely with the position of the switch at t* = 2/9. Therefore, at t* = 2/9 the latter will be in positon AB and, consequently, the lamp will be lit. By the same token, it can also be shown that the lamp in Figure 3 will be dim at time t* = 2/9.

3.3: The Postulate of Permanence

In Section 3.2, the principle of continuity helped us find the final state resulting from the accomplishment of a supertask in cases in which there exists a ‘natural’ limit for the state of the physical system involved as time progressively approaches the instant at which the supertask is achieved. Now it is considerably more problematic to apply this principle to supertasks for which there is no ‘natural’ limit. For an example, let us consider Black's infinity machine, introduced in Section 2.4, and let us ask ourselves where the ball will be at instant t* = 1 P.M. at which the supertask is achieved. We can set up a reductio ad absurdum type of argument, as follows. Assume that at t* = 1 P.M. the ball were to occupy position P, that it was in point P. According to the principle of continuity, it follows that the limit to which the position occupied by the ball tends as time approaches t* = 1 P.M. is precisely position P. We know, though, that Black's infinity machine makes the ball oscillate more and more quickly between the fixed points A ($x = 0$) and B ($x = 1$) as we get closer to t* = 1 P.M., so the position of the ball does not approach any definite limit as we get closer to t* = 1 P.M. This conclusion patently contradicts what follows from the principle of continuity. Therefore, the assumption that, after Black's supertask is achieved (t* = 1 P.M.), the ball is at point P leads to contradiction with the principle of continuity. Thus, the ball cannot be at point P at t* = 1 P.M., and as the point can be any, given that it has been chosen arbitrarily, the ball cannot be at any single one of the points, which means that at t* = 1 P.M. the ball has ceased to exist. This funny conclusion is consistent with the principle of continuity, as we have just seen, but it enters into contradiction with what could be termed the ‘postulate of permanence’: no material body (and by that we mean a given quantity of matter) can go out of existence all of a sudden, without leaving any traces. The postulate of permanence seems to characterise our world at least as evidently as the principle of continuity. Notice in particular that certain physical bodies (particles) may dematerialise, but that is not inconsistent with the postulate of permanence since such a dematerialisation leaves an energy trace (which is not true of Black's ball). Consequently, we can see that the case Black's infinity machine is one in which the principles of continuity and permanence turn out to be mutually inconsistent. As long as we do not give up any of them, we are forced to accept that such an infinity machine is physically impossible.

Section 4: The Physics of Supertasks

As we do not know exactly what laws of nature there are, it goes without saying that the question whether a particular supertask is physically possible (that is, compatible with those laws) cannot be given a definitive answer in general. What we have done in 3 above is rather to set out necessary conditions for physical possibility which are plausible (such as the principle of continuity) and sufficient conditions for

physical impossibility which are likewise plausible (such as Grünbaum's criterion of kinematical impossibility). In this section we shall look into a problem related to the one just dealt with, but one to which a definitive answer can be given: the problem of deciding whether a certain supertask is possible within the framework of a given physical theory, that is, whether it is compatible with the principles of that theory. These are two distinct problems. At this stage our object are theories whereas in 3 above we were concerned with the real world. What we are after is supertasks formulated within the defined framework of a given physical theory which can tell us something exciting and/or new about that theory. We will discover that this search will lead us right into the heart of basic theoretical problems.

4.1: A New Form of Indeterminism: Spontaneous Self-Excitation

Classical dynamics is a theory that studies the motion of physical bodies which interact among themselves in various ways. The vast majority of interesting examples of supertasks within this theory have been elaborated under the assumption that the particles involved only interact with one another by means of elastic collisions, that is, collisions in which no energy is dissipated. We shall see here that supertasks of type *w** give rise to a new form of indeterministic behaviour of dynamical systems. The most simple type of case (Pérez Laraudogoitia [1996]) is illustrated by the particle system represented in Figure 4 at three distinct moments. It consists of an infinite set of identical point particles P_1, P_2, P_3, . . . , P_n, . . . arranged in a straight line. Take the situation depicted in Figure 4A first. In it P_1 is at one unity distance from the coordinate origin O, P_2 at a distance 1/2 of O, P_3 at a distance 1/3 of O and so on. In addition, let it be that all the particles are at rest, except for P_1, which is approaching O with velocity v = 1. Suppose that all this takes place at t = 0. Now we will employ

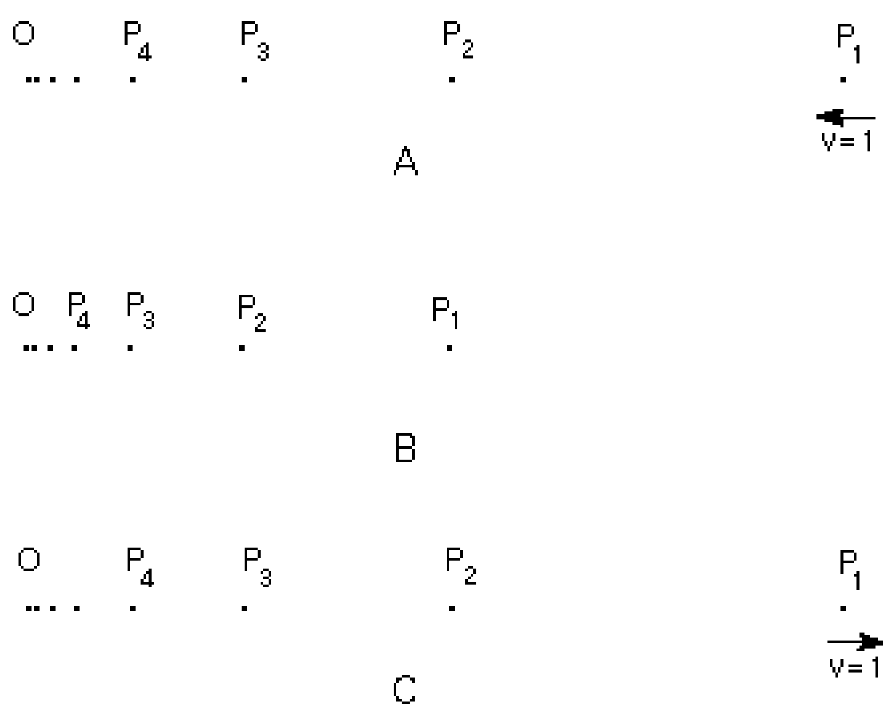


Figure 4

the well-known dynamic theorem by which if two identical particles undergo an elastic collision then

they will exchange their velocities after colliding. If our particles P_1, P_2, P_3, . . . collide elastically, it is easy to predict what will happen after t = 0 with the help of this theorem. In the event that P_1 were on its own, it would reach O at t = 1, but in fact it will collide with P_2 and lie at rest there, while P_2 will acquire velocity $v = 1$. If P_1 and P_2 were on their own, then it would be P_2 that would reach O at t = 1, but P_2 will in fact collide with P_3, and lie at rest there, while P_3 will acquire velocity $v = 1$. Again, it can be said that if P_1, P_2 and P_3 were on their own, then it is P_3 that would reach O at t = 1, but in actual fact it will collide with P_4 and lie at rest there, while P_4 will acquire velocity $v = 1$, and so on. From the foregoing it follows that no particle will get to O at t = 1, because it will be impeded by a collision with another particle. Therefore, at t = 1 all the particles will already lie at rest, which yields the configuration in Figure 4B. Since P_1 stopped when it collided with P_2, it will occupy the position P_2 had initially (at t = 0). Similarly, P_2 stopped after colliding with P_3 and so it will occupy the position P_3 had initially (at t = 0), . . . , etc. If we view each collision as an action (which is plausible, since it involves a sudden change of velocities), it turns out that between t = 0 and t = 1 our evolving particle system has performed a supertask of type *w*. The second dynamic theorem we will make use of says that if a dynamic process is possible, then the process resulting from inverting the direction in which all the bodies involved in it move is also possible. Applying this to our case, if the process leading from the system in the situation depicted in Figure 4A to the situation depicted in Figure 4B is possible (and we have just seen it is), then the process obtained by simply inverting the direction in which the particles involved move will also be possible. This new possible process does not bring the system from configuration 4B back to configuration 4A but rather changes it into configuration 4C (as the direction in which P_1 moves must be inverted). As the direct process lasts one time unity (from t = 0 to t = 1), so will the inverse process, and as in the direct process the system performs a supertask of type *w*, in the inverse process it will perform one of type *w**. What is interesting about this new supertask of type *w**? What's interesting is that it takes the system from a situation (4B) in which all its component particles are at rest to another situation (4C) in which not all of them are. This means that the system has self-excited, because no external influence has been exerted on it, and, what is more, it has done so spontaneously and unpredictably, because the supertask can set off at any instant and there is no way of predicting when it will happen. We have found a supertask of type *w** to be the source of a new form of dynamical indeterminism. The reason we speak of indeterminism is because there is no initial movement to the performance of the supertask. The system self-excites in such a way that each particle is set off by a collision with another one, and it is the ordinal type *w** of the sequence of collisions accomplished in a finite time that guarantees movement, without the need for a ‘prime mover’. Now movement without a ‘prime mover’ is precisely what characterises the dynamical indeterminism linked to supertasks of type *w**.

4.2: Bifurcated Supertasks

Within relativity theory, supertasks have been approached from a radically different perspective from the one adopted here so far. This new perspective is inherently interesting, since it links the problem of supertasks up with the relativistic analysis of the structure of space-time. To get an insight into the nature of that connection, let us first notice that, according to the theory of relativity, the duration of a process will not be the same in different reference systems but will rather vary according to the reference system

within which it is measured. This leaves open the possibility that a process which lasts an infinite amount of time when measured within reference system O may last a finite time when measured within a different reference system O'.

The supertask literature has needed to exploit space-times with sufficiently complicated structure that global reference systems cannot be defined in them. In these and other cases, the time of a process can be represented by its ‘proper time’. If we represent a process by its world-line in space-time, the proper time of the process is the time read by a good clock that moves with the process along its world-line. A familiar example of its use is the problem of the twins in special relativity. One twin stays home on earth and grows old. Forty years of proper time, for example, elapses along his world-line. The travelling twin accelerates off into space and returns to find his sibling forty years older. But much less time -- say only a year of proper time -- will have elapsed along the travelling twin world-line if he has accelerated to sufficiently great speeds.

If we take this into account it is easily seen that the definition of supertask that we have been using is ambiguous. In section 1 above we defined a supertask as an infinite sequence of actions or operations carried out in a finite interval of time. But we have not specified in whose proper time we measure the finite interval of time. Do we take the proper time of the process under consideration? Or do we take the proper time of some observer who watches the process? It turns out that relativity theory allows the former to be infinite while the latter is finite. This fact opens new possibilities for supertasks. Relativity theory thus forces us to disambiguate our definition of supertask, and there is actually one natural way to do it. We can use Black's idea -- presented in 2.4 -- of an infinity machine, a device capable of performing a supertask, to redefine a supertask as an infinite sequence of actions or operations carried out by an infinity machine in a finite interval of the machine's own proper time measured within the reference system associated to the machine. This redefinition of the notion of supertask does not change anything that has been said until now; our whole discussion remains unaffected so long as ‘finite interval of time’ is read as ‘finite interval of the machine's proper time’. This notion of supertask, disambiguated so as to accord with relativity theory, will be denoted by the expression ‘supertask-1’. Thus:

> Supertask-1: an infinite sequence of actions or operations carried out by an infinity machine in a finite interval of the machine's proper time.

However we might also imagine a machine that carries out an infinite sequence of actions or operations in an infinite machine proper time, but that the entire process can be seen by an observer in a finite amount of the observer's proper time.

It is convenient at this stage to introduce a contrasting notion:

> Supertask-2: an infinite sequence of actions or operations carried out by a machine in a finite interval of an observer's proper time.

While we did not take relativity theory into account, the notions of supertask-1 and supertask-2

coincided. The duration of an interval of time between two given events is the same for all observers. However in relativistic spacetimes this is no longer so and the two notions of supertasks become distinct. Even though all supertasks-1 are also supertasks-2, there may in principle be supertasks-2 which are not supertasks-1. For instance, it could just so happen that there is a machine (not necessarily an infinity machine) which carries out an infinite number of actions in an interval of its own proper time of infinite duration, but in an interval of some observer's proper time of finite duration. Such a machine would have performed a supertask-2 but not a supertask-1.

The distinction between supertasks-1 and supertasks-2 is certainly no relativistic hair-splitting. Why? Because those who hold that, while conceptually possible, supertasks are physically impossible (this seems to be the position adopted by Benacerraf and Putnam [1964], for instance) usually mean that supertasks-1 are physically impossible. But from this, it does not follow that supertasks-2 must also be physically impossible. Relativity theory thus adds a brand-new, exciting extra dimension to the challenge presented by supertasks. Earman and Norton (1996), who have studied this issue carefully, use the name ‘bifurcated supertasks’ to refer specifically to supertasks-2 which are not supertasks-1, and I will adopt this term.

4.3: Bifurcated Supertasks and the Solution to the Philosophical Problem of Supertasks

What shape does the philosophical problem posed by supertasks -- introduced in Section 1.2 -- take on now? Remember that the problem lay in specifying the set of sentences which describe the state of the world after the supertask has been performed. The problem will now be to specify the set of sentences which describe the relevant state of the world after the bifurcated supertask has been performed. Before this can done, of course, the question needs to be answered whether a bifurcated supertask is physically possible. Given that we agree that compatibility with relativity theory is a necessary and sufficient condition of physical possibility, we can reply in the affirmative.

Pitowsky (1990) first showed how this compatibility might arise. He considered a Minkowski spacetime, the spacetime of special relativity. He showed that an observer O* who can maintain a sufficient increase in his acceleration will find that only a finite amount of proper time elapses along his world-line in the course of the complete history of the universe, while other unaccelerated observers would find an infinite proper time elapsing on theirs.

Let us suppose that some machine M accomplishes a bifurcated supertask in such a way that the infinite sequence of actions involved happens in a finite interval of an observer O's proper time. If we imagine such an observer at some event on his world-line, all those events from which he can retrieve information are in the ‘past light cone’ of the observer. That is, the observer can receive signals travelling at or less than the speed of light from any event in his past light cone. The philosophical problem posed by the bifurcated supertask accomplished by M has a particularly simple solution when the infinite sequence of actions carried out by M is fully contained within the past light cone of an event on observer O's world-line. In such a case the relevant state of the world after the bifurcated supertask has been performed is M's

state, and this, in principle, can be specified, since O has causal access to it. Unfortunately, a situation of this type does not arise in the simple bifurcated supertask devised by Pitowsky (1990). In his supertask, while the accelerated observer O* will have a finite upper bound on the proper time elapsed on his world-line, there will be no event on his world-line from which he can look back and see an infinity of time elapsed along the world-line of some unaccelerated observer.

To find a spacetime in which the philosophical problem posed by bifurcated supertasks admits of the simple solution that has just been mentioned, we will move from the flat spacetime of special relativity to the curved spacetimes of general relativity. One type of spacetime in the latter class that admits of this simple solution has been dubbed Malament-Hogarth spacetime, from the names of the first scholars to use them (Hogarth [1992]). An example of such a spacetime is an electrically charged black hole (the Reissner-Nordstroem spacetime). A well known property of black holes is that, in the view of those who remain outside, unfortunates who fall in appear to freeze in time as they approach the event horizon of the black hole. Indeed those who remain outside could spend an infinite lifetime with the unfortunate who fell in frozen near the event horizon. If we just redescribe this process from the point of view of the observer who does fall in to the black hole, we discover that we have a bifurcated supertask. The observer falling in perceives no slowing down of time in his own processes. He sees himself reaching the event horizon quite quickly. But if he looks back at those who remain behind, he sees their processes sped up indefinitely. By the time he reaches the event horizon, those who remain outside will have completed infinite proper time on their world-lines. Of course, the cost is high. The observer who flings himself into a black hole will be torn apart by tidal forces and whatever remains after this would be unable to return to the world in which he started.

Section 5: What Supertasks Entail for the Philosophy of Mathematics

5.1: A Critique of Intuitionism

The possibility of supertasks has interesting consequences for the philosophy of mathematics. To start with, take a well-known unsolved mathematical problem, for example that of knowing whether Goldbach's conjecture is or is not correct. Goldbach's conjecture asserts that any even number greater than 2 is the sum of two prime numbers. Nobody has been capable of showing whether this is true yet, but if supertasks are possible, that question can be resolved. Let us, to that effect, perform the supertask of type *w* consisting in the following sequence of actions: action a_1 involves checking whether the first pair greater than 2 (number 4) is the sum of two prime numbers or not; let this action be accomplished at t = 0.3 P.M.; action a_2 involves checking whether the second pair greater than 2 (number 6) is the sum of two prime numbers or not; let this action be accomplished at t = 0.33 P.M.; action a_3 involves checking whether the third pair greater than 2 (number 8) is the sum of two prime numbers or not; let this action be accomplished at t = 0.333 P.M., and so on. It is clear that at t = 0.33333... = 1/3 P.M., the instant at which the supertask has already been performed, we will have checked all the pairs greater than 2, and, therefore, will have found some which is not the sum of two prime numbers or else will have found all of

them to be the sum of two prime numbers. In the first case, we will know at t = 1/3 P.M. that Goldbach's conjecture is false; in the seocnd case we will know at t = 1/3 P.M. that it is true. Weyl (1949) seems to have been the first to point to this intriguing method -the use of supertaks- for settling mathematical questions about natural numbers. He, however, rejected it on the basis of his finitist conception of mathematics; since the performance of a supertask involves the successive carrying out of an actual infinity of actions or operations, and the infinity is impossible to accomplish, in his view. For Weyl, taking the infinite as an actual entity makes no sense. Nevertheless, there are more problems here than Weyl imagines, at least for those who ground their finitist philosophy of mathematics on intuitionism à la Brouwer. That is because Brouwer's rejection of actual infinity stems from the fact that we, as beings, are immersed in time. But this in itself does not mean that all infinities are impossible to accomplish, since an infinity machine is also ‘a being immerse in time’ and that in itself does not prevent the carrying out of the infinity of successive actions a supertask is comprised of. It goes without saying that one can adhere to a constructivist philosophy of mathematics (and the consequent rejection of actual infinity) for diferent reasons from Brouwer's; supertasks will still not be the right kind of objet to study either.

As Benacerraf and Putnam (1964) have observed, the acknowledgement that supertasks are possible has a profound influence on the philosophy of mathematics: the notion of truth (in arithmetic, say) would no longer be doubtful, in the sense of dependent on the particular axiomatisation used. The example mentioned earlier in connection with Goldbach's conjecture can indeed be reproduced and generalised to all other mathematical statements involving numbers (although, depending on the complexity of the statement, we might need to use several infinity machines instead of just one), and so, consequently, supertasks will enable us to decide on the truth or falsity of any arithmetical statement; our conclusion will no longer depend on provability in some formal system or constructibility in a more or less strict intuitionistic sense. This conclusion seems to lead to a Platonist philosophy of mathematics.

5.2: The Importance of the Malament-Hogarth Spacetime

A similar conclusion follows regarding the implications of supertasks for the philosophy of mathematics if one only accepts the possibility of bifurcated supertasks. Of course, a bifurcated supertask performed in a non-Malament-Hogarth space-time would not be so interesting in this sense. The obvious reason is that we would not even have a sound procedure to determine the truth or falsity of Goldbach's conjecture seen in 5.1 by means of the performance of an infinite sequence of actions of order type w. To really have a safe decision procedure in this simple case (as in other, more complex ones) there must necessarily exist an instant of time at which it can be said that the supertask has been accomplished. Otherwise, in the event that the machine finds a counterexample to Goldbach's conjecture we will know it to be false, but in the event of the machine finding none we will not be able to tell that it is true, because for this there must exist an instant of time by which the supertask has been accomplished and at which we can say something like: "the supertask has been performed and the machine has found no counterexamples to Goldbach's conjecture; therefore, the conjecture is true". It follows that, in the case of a bifurcated supertask, possessing a sound decision procedure on Goldbach's conjecture requires the existence of an observer O such that the infinite sequence of actions (of order type w) carried out by the machine lies within the past light cone of an event on observer O's world-line. But this is equivalent to saying that the

relativistic space-time in which the bifurcated supertask is performed is a Malament-Hogarth space-time, and this realisation is one of the main reasons why this sort of relativistic space-times have been studied in the literature.

Note, finally, the intuitionistic criticism of the possibility of supertasks is even less effective in the case of bifurcated supertasks, because in this latter case it is not required that there is any sort of device capable of carrying out an infinite number of actions or operations in a finite time (measured in the reference system associated to the device in question, which is the natural reference system to consider). In contrast, from the possibility of bifurcated supertasks in Malament-Hogarth space-times strong arguments follow against an intuitionistic philosophy of mathematics. As Earman and Norton remind us, it is noteworthy that certain facts relative to the non-Euclidean structure of space-time can have relevant consequences for the nature of mathematical truth.

Bibliography

- Allis, V. and Koetsier, T., 1991, ‘On Some Paradoxes of the Infinite’, *British Journal for the Philosophy of Science*, **42**, pp. 187-194
- Allis, V. and Koetsier, T., 1995, ‘On Some Paradoxes of the Infinite II’, *British Journal for the Philosophy of Science*, **46**, pp. 235-247
- Alper, J.S. and Bridger, M., 1998, ‘Newtonian Supertasks: A Critical Analysis’ *Synthese*, **114**, pp. 355-369
- Alper, J.S., Bridger, M., Earman , J. and Norton , J.D., 2000, ‘What is a Newtonian System? The Failure of Energy Conservation and Determinism in Supertasks’ *Synthese*, **124**, pp.281-293
- Aristotle, *Physics*, (W. Charlton, trans.), Oxford: Oxford University Press, 1970
- Benacerraf, P., 1962, ‘Tasks, Super-Tasks, and Modern Eleatics’, *Journal of Philosophy*, **LIX**, pp. 765-784; reprinted in Salmon [1970]
- Benacerraf, P. and Putnam, H., 1964, Introduction, *Philosophy of Mathematics: Selected Readings*, P. Benacerraf and H. Putnam (eds.), 2nd edition, Cambridge: Cambridge University Press, pp. 1-27
- Benardete, J.A., 1964, *Infinity: An Essay in Metaphysics*, Oxford: Clarendon Press
- Berresford, G. C., 1981, ‘A Note on Thomson's Lamp "Paradox"’, *Analysis*, **41**, pp. 1-7
- Black, M., 1950-1, ‘Achilles and the Tortoise’, *Analysis*, **XI**, pp. 91-101; reprinted in Salmon [1970]
- Black, M., 1954, *Problems of Analysis*, Ithaca: Cornell University Press
- Bostock, D., 1972-73, ‘Aristotle, Zeno and the Potential Infinite’, *Proceedings of the Aristotelian Society*, **73**, pp. 37-51
- Burke, M.B., 1984, ‘The Infinitistic Thesis’, *The Southern Journal of Philosophy*, **22**, pp. 295-305
- Burke, M.B., 2000, ‘The Impossibility of Superfeats’, *The Southern Journal of Philosophy*, **XXXVIII**, pp.207-220
- Burke, M.B., 2000, ‘The Staccato Run: a Contemporary Issue in the Zenonian Tradition’ ,*The Modern Schoolman*, **LXXVIII**, pp.1-8
- Bridger, M., and Alper, J. S., 1999, ‘On the Dynamics of Perez-Laraudogoitia's Supertask’,

Synthese, **119**, pp. 325-337
- Chihara, C., 1965, ‘On the Possibility of Completing an Infinite Task’, *Philosophical Review*, **LXXIV**, pp. 74-87
- Clark, P., and Read, S., 1984, ‘Hypertasks’, *Synthese*, **61**, pp. 387-390
- Earman, J., 1994, *Bangs, Crunches, Shrieks, and Whimpers: Singularities and Acausalities in Relativistic Spacetimes*, New York: Oxford University Press
- Earman, J., and Norton, J.D., 1993, ‘Forever Is a Day: Supertasks in Pitowsky and Malament-Hogarth Spacetimes’, *Philosophy of Science*, **60**, pp. 22-42
- Earman, J., and Norton, J. D., 1996, ‘Infinite Pains: The Trouble with Supertasks’, in *Benacerraf and His Critics*, A. Morton and S. Stich (eds.), Oxford: Blackwell, pp. 231-261.
- Earman, J., and Norton, J. D., 1998, ‘Comments on Laraudogoitia's "Classical Particle Dynamics, Indeterminism and a Supertask"’, *British Journal for the Philosophy of Science*, **49**, pp. 123-133
- Faris, J. A., 1996, *The Paradoxes of Zeno*, Aldershot: Avebury
- Gale, R. M. (ed.), 1968, *The Philosophy of Time*, London: MacMillan
- Glazebrook, T., 2001, ‘Zeno Against Mathematical Physics’, *Journal of the History of Ideas*, **62**, pp.193-210
- Groarke, L., 1982, ‘Zeno's Dichotomy: Undermining the Modern Response’, *Auslegung*, **9**, pp. 67-75
- Grünbaum, A. 1950-52 ‘Messrs. Black and Taylor on Temporal Paradoxes’, *Analysis*, **11-12**, pp. 144-148
- Grünbaum, A., 1967, *Modern Science and Zeno's Paradoxes*, Middletown, CT: Wesleyan University Press
- Grünbaum, A., 1968, ‘Are "Infinity Machines" Paradoxical?’, *Science*, **CLIX**, pp. 396-406
- Grünbaum, A., 1969, ‘Can an Infinitude of Operations be Performed in a Finite Time?’, *British Journal for the Philosophy of Science*, **20**, pp. 203-218
- Grünbaum, A., 1970, ‘Modern Science and Zeno's Paradoxes of Motion’, in Salmon [1970], pp. 200-250
- Hogarth, M. L., 1992, ‘Does General Relativity Allow an Observer to View an Eternity in a Finite Time?’, *Foundations of Physics Letters*, **5**, pp. 173-181
- Hogarth, M. L., 1994, ‘Non-Turing Computers and Non-Turing Computability’, in *PSA 1994*, D. Hull, M. Forbes and R.M. Burian (eds.), **1**, East Lansing: Philosophy of Science Association, pp. 126-138
- Holgate, P., 1994, ‘Discussion: Mathematical Notes on Ross's Paradox’, *British Journal for the Philosophy of Science*, **45**, pp. 302-304
- Koetsier, T. and Allis, V., 1997, ‘Assaying Supertasks’, *Logique & Analyse*, **159**, pp. 291-313
- McLaughlin, W. I., 1998, ‘Thomson's Lamp is Dysfunctional’, *Synthese*, **116**, pp. 281-301
- Moore, A. W., 1989-90, ‘A Problem for Intuitionism: The Apparent Possibility of Performing Infinitely Many Tasks in a Finite Time’, *Proceedings of the Aristotelian Society*, **90**, pp. 17-34
- Moore, A. W., 1990, *The Infinite*, London: Routledge
- Norton, J. D., 1999, ‘A Quantum Mechanical Supertask’, *Foundations of Physics*, **29**, pp. 1265-1302
- Owen, G. E. L., 1957-58, ‘Zeno and the Mathematicians’, *Proceedings of the Aristotelian Society*, **LVIII**, pp. 199-222, reprinted in Salmon [1970]

- Pérez Laraudogoitia, J., 1996, 'A Beautiful Supertask', *Mind*, **105**, pp. 81-83
- Pérez Laraudogoitia, J., 1997, 'Classical Particle Dynamics, Indeterminism and a Supertask', *British Journal for the Philosophy of Science*, **48**, pp. 49-54
- Pérez Laraudogoitia, J., 1998, 'Infinity Machines and Creation *Ex Nihilo*', *Synthese*, **115**, pp. 259-265
- Pérez Laraudogoitia, J., 1998, 'Some Relativistic and Higher Order Supertasks', *Philosophy of Science*, **65**, pp. 502-517
- Pérez Laraudogoitia, J., 1999, 'Earman and Norton on Supertasks that Generate Indeterminism', *British Journal for the Philosphy of Science*, **50**, pp.137-141
- Pérez Laraudogoitia, J., 1999, 'Why Dynamical Self-excitation is Possible', *Synthese*, **119**, pp. 313-323
- Pitowsky, I., 1990, 'The Physical Church Thesis and Physical Computational Complexity', *Iyyun*, **39**, pp. 81-99
- Priest, G., 1982, 'To Be and Not to Be: Dialectical Tense Logic', *Studia Logica*, **XLI**, pp. 157-176
- Ray, C., 1990, 'Paradoxical Tasks', *Analysis*, **50**, pp. 71-74
- Ray, C., 1991, *Time, Space and Philosophy*, London: Routledge
- Sainsbury, R. M., 1988, *Paradoxes*, Cambridge: Cambridge University Press
- Salmon, W., (ed.), 1970, *Zeno's Paradoxes*, Indianapolis: Bobbs-Merril
- Salmon, W., 1980, *Space, Time and Motion: A Philosophical Introduction*, Minneapolis: University of Minnesota Press
- Smith, J. W., 1986, *Reason, Science and Paradox*, London: Croom Helm
- Sorabji, R., 1983, *Time, Creation and the Continuum*, London: Gerald Duckworth and Co. Ltd.
- Svozil, K., 1993, *Randomness and Undecidability in Physics*, Singapore: World Scientific
- Taylor, R., 1951-52, 'Mr. Black on Temporal Paradoxes', *Analysis*, **XII**, pp. 38-44
- Taylor, R., 1952-53, 'Mr. Wisdom on Temporal Paradoxes', *Analysis*, **XIII**, pp. 15-17
- Thomson, J., 1954-55, 'Tasks and Super-Tasks', *Analysis*, **XV**, pp. 1-13; reprinted in Salmon [1970]
- Thomson, J., 1967, 'Infinity in Mathematics and Logic', in *Encyclopedia of Philosophy*, P. Edwards (ed.), **4**, New York: MacMillan, pp. 183-90
- Thomson, J., 1970, 'Comments on Professor Benacerraf's Paper', in Salmon [1970], pp. 130-38.
- Van Bendegem, J. P., 1994, 'Ross' Paradox is an Impossible Super Task', *British Journal for the Philosophy of Science*, **45**, pp. 743-48
- Van Bendegem, J. P., 1995-1997, 'In Defence of Discrete Space and Time', *Logique & Analyse*, **150-151-152**, pp. 127-150
- Vlastos, G., 1966, 'Zeno's Race Course', *Journal of the History of Philosophy*, **IV**, pp. 95-108
- Vlastos, G., 1967, 'Zeno of Elea', in *Encyclopedia of Philosophy*, P. Edwards (ed.), **8**, New York: MacMillan, pp.369-79
- Watling, J., 1953, 'The Sum of an Infinite Series', *Analysis*, **XIII**, pp. 39-46
- Wedeking, G. A., 1968, 'On a Finitist "Solution" to Some Zenonian Paradoxes', *Mind*, **77**, pp. 420-26
- Weyl, H., 1949, *Philosophy of Mathematics and Natural Science*, Princeton: Princeton University Press
- Whitrow, G. J., 1961, *The Natural Philosophy of Time*, Edinburgh: Thomas Nelson & Sons

- Wisdom, J. O., 1951-52, 'Achilles on a Physical Racecourse', *Analysis*, **XII**, pp. 67-72, reprinted in Salmon [1970]

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

infinity | space and time: Malament-Hogarth spacetimes and the new computability | Zeno's paradoxes

Jon Pérez Laraudogoitia
The University of the Basque Country
ylppelaj@vc.ehu.es

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: June 29, 1999
Content last modified: November 26, 2001

Determinates vs. Determinables

Everything red is colored, and all squares are polygons. A square is distinguished from other polygons by being four-sided, equilateral, and equiangular. What distinguishes red things from other colored things? This has been understood as a conceptual rather than scientific question. Theories of wavelengths and reflectance and sensory processing are not considered. Given just our ordinary understanding of color, it seems that what differentiates red from other colors is only redness itself. The Cambridge logician W. E. Johnson introduced the terms determinate and determinable to apply to examples such as red and colored. Chapter XI, of Johnson's Logic, Part I (1921), “The Determinate and the Determinable,” is the main text for discussion of this distinction.

This entry consists of the following sections. Section 1 attends closely to Chapter XI, of W. E. Johnson's Logic, Part I. Section 2 briefly discusses Johnson's use of the determinate-determinable relation elsewhere than Chapter XI of Logic, Part I, and connects this with A. N. Prior, “Determinables, Determinates, and Determinants” (1949). Section 3 describes a 1959 symposium between Stephan Körner and John Searle entitled “On Determinables and Resemblance” and examines both contributions critically. Section 4 describes the Munsell Color Solid so that color examples can be more exact. Section 5 describes and criticizes attempts to define the determinate-determinable relation by means of predicate entailment in the style of Searle. Section 6 pays more attention to this distinction and directs attention to a certain understanding of “disjunctive predicate.” Contrived disjunctive and conjunctive predicates are the typical cause of difficulties in attempts to define the determinate-determinable relation. Section 7 distinguishes independent predicates from non-independent predicates, and thus distinguishes disjunctive and conjunctive predicates from non-disjunctive and non-conjunctive predicates, in a way that assumes no prior classifications of determinates under a determinable. Section 8 explores a view advanced by Johnson and endorsed by many others that the things in *the* world, as distinguished from our descriptions and conceptions of them, are absolutely determinate. This section entertains the contrary view that nothing is absolutely determinate.

- Related Entries

---

1. W. E. Johnson's Chapter on The Determinable

1.1 Substantive and Adjective

Johnson invents phrases and also attaches new meanings to familiar words. His terms determinate, determinable, occurrent, continuant and ostensive definition have entered the philosophical lexicon. Some of his innovations are largely forgotten. Throughout Logic and especially throughout Part I, Chapter XI, Johnson uses a distinction between substantive and adjective. Although he draws and observes many distinctions meticulously, the distinction between the mention of a linguistic expression and its use is not among them. Adjective and substantive sometimes appear in his writings to be linguistic items. More often they are definitely non-linguistic. They are logical categories, “the ultimate comprising classes” (1922, p. 60). “My distinction between substantive and adjective is roughly equivalent to the more popular philosophical antithesis between particular and universal; the notions, however, do not exactly coincide” (1922, p. xiii). The term ‘adjective’ remains in all forthcoming quotations from Johnson. In discussion of Johnson, the term ‘property’ often replaces ‘adjective’ although it is not an exact equivalent.

Johnson is interested in the logical differences between ‘Red is a colour’ and ‘Plato is a man’. He says that the second sentence involves adjectival predication. ‘Human’ is an adjective (property) predicated of Plato. ‘Colour,’ on the other hand, is not an adjective (property) predicated of red (1921, p. 176).

1.2 Similarity and Difference

Color is one of Johnson's central examples. Red and blue are determinate with respect to the determinable color, and Carolina Blue and Duke Blue are determinate with respect to blue.

> The relation of a determinate to its determinable resembles that of an individual to a class, but differs in some important respects. For instance, taking any given determinate, there is only one determinable to which it can belong. Moreover, any one determinable is a literal summum genus not subsumable under any higher genus; and the absolute determinate is a literal infima species under which no other determinate is subsumable. (Part I, Introduction, 1921, p. xxxv)

What makes red and blue and Carolina Blue all colors? Johnson denies that there is some property [some “secondary” adjective] that red and blue share that makes them both colors. The view that color is itself a property that red and blue share requires an explanation of what distinguishes the color red from the color blue. Explanations such as “Red is the color of fire trucks” and “Blue is the color of my true love’s eyes” miss the point. They refer to mere contingent facts. An appropriate explanation should provide a necessary truth such as ‘Triangles are three-sided polygons.’

Rather than resemblance, the sharing of some property, that make red and blue colors, says Johnson, it is

differences between colors.

> In fact, the several colours are put into the same group and given the same name colour, not on the ground of any partial agreement, but on the ground of the special kind of difference which distinguishes one colour from another; whereas no such difference exists between a colour and a shape. (1921, p. 176)

The absence of a difference or exclusion of this kind explains why red and square is not a determinate of a single determinable. “Taking any given determinate, there is only one determinable to which it can belong” (1921, p. xxxv). The nature of the exclusion also explains why no two determinates of the same determinable can qualify exactly the same entire spatio-temporal part of any object (1921, p. 181).

Arguments about the incompatibility of colors in the 1950s and 1960s were concerned with theories of necessity, analyticity, the *a priori*, and meaning. (Edwards and Pap, 1973, has a bibliography of such works, pp. 745-746.) These discussions rarely address specifically the relation between determinables and determinates. Nor do assumptions about the incompatibility of determinates often connect with the problem of understanding incompatibility in general. When the problem of incompatibility reopens there may be less emphasis on philosophy of language and logic and more emphasis on the branch of metaphysics that studies the nature of properties and the recently formed hybrid of metaphysics and science that studies the nature of color and other sensory qualities.

1.3 Determinables and Genera

Treatments of the determinate-determinable relation often contrast it with the species-genus relation. Features such as three-sided differentiate the species triangle under the genus polygon, while the only feature that distinguishes the determinate red under the determinable color is the very determinate itself. A genus-species relation obtains when a proper definition of the form $X = YZ$ is possible. When no such definition is possible and certain other formal requirements are satisfied, the determinable-determinate relation obtains. For this neat, sharp contrast, Johnson provides at best only equivocal support. His remarks are sometimes incompatible with this contrast.

Consider Johnson's examples. The introduction of the term ‘determinable’ in Chapter XI reads: “I propose to call such terms as colour and shape determinables in relation to such terms as red and circular which will be called determinates” (1921, p. 174). So circular is a determinate of the determinable shape despite the existence of a proper definition that distinguishes circles from other shapes. Different shapes are incompatible and are therefore under the same determinable. Being related by incompatibility (in the right way) appears to be necessary and sufficient for items to be determinates under a single determinable. Some determinates such as red cannot be differentiated by a traditional, conjunctive genus-species definition. Others such as square and circular can be so differentiated. Johnson's example of shape shows that a determinable-determinate relation does not require the impossibility of a conjunctive definition.

Rather than insist on a sharp contrast, Johnson attempts to subsume traditional species-genus relations under determinate-determinable relations:

> We have now to point out that the increased determination of adjectival predication which leads to a narrowing of extension may consist—not in a process of conjunction of separate adjectives—but in the process of passing from a comparatively indeterminate adjective to a comparatively more determinate adjective under the same determinable. Thus there is a genuine difference between that process of increased determination which conjunctivally introduces foreign adjectives, and that other process by which without increasing, so to speak, the number of adjectives, we define them more determinately.
>
> In fact, the foreign adjective which appears to be added on in the conjunctive process, is really not introduced from the outside, but is itself a determinate under another determinable, present from the start, though suppressed in the explicit connotation of the genus. (Johnson, 1921, pp. 178-9)

Johnson goes on to provide a symbolic representation of botanical classification in which there are five determinables. One of these determinables is number of cotyledons under which fall the determinates acotyledon, monocotyledon, and dicotyledon. The other determinables concern stamens, the corolla, forms of attachment, and divisibility. The determinables in this botanical example represent “the summum genus ‘plants’ as describable under these five heads” (1921, p. 180). A determinable can have several dimensions.

Johnson's discussion of color variations illustrates such dimensions. Colors vary according to hue, saturation and brightness, and these variations are independent of one another. If hue, saturation, and brightness are determinables, they are not separate, since they depend on each other. There cannot be saturation without hue, for example, even though no determination of saturation requires any particular determination of hue. Johnson says that the determinable color is “single, though complex, in the sense that the several constituent characters upon whose variations its variability depends are inseparable” (1921, p. 183).

There is a difficulty here because the dependence patterns between the three variables are not entirely uniform. Hue and saturation cannot exist without each other, or without brightness, but degrees of brightness do not require either hue or saturation. Black and white movies and photographs and many other achromatic examples come to mind. Dimensions of quality space can vary in their dependence relations on each other.

1.4 Quality Order

There are differences between determinates under the same determinable. Johnson says these differences are comparable. The difference between red and yellow, for example, is greater than that between red and orange.

> In this case the several determinates are to be conceived as necessarily assuming a certain serial order, which develops from the idea of what may be called ‘adjectival betweenness.’ The term ‘between’ is used here in a familiar metaphorical sense imagined most naturally in spatial form. (1921, pp. 181-182)

The three-place relation (*Dabc*) *the difference between a and c is greater than that between a and b,* however, does not by itself provide an adequate definition of ‘between.’ A diagram helps to illustrate this point. Assume for the purpose of diagramming that *the difference between a and c* specifies a certain distance in quality

space. A circle with center *a* and radius *ac* represents points in the space at distance *ac* from point *a*. Any distance between point *a* and any point *b* within this circle is less than the distance between *a* and *c*. A point *b* within this circle represents *Dabc*. As Figure 1 illustrates, *Dabc* is consistent with the distance between *b* and *c* being greater than the distance between *a* and *c*.

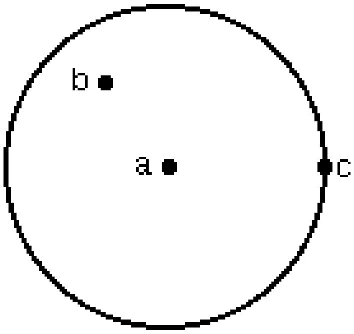


Figure 1

In this situation, *b* is not between *a* and *c* in any sense. For example, the difference (or distance) between red and yellow is greater than the distance between red and purplish red, but purplish red is not between red and yellow.

The word ‘betwixt’, which appears in the next paragraph, comes from Goodman (1951), pp. 244-253, but only the word, not Goodman's definition or intended sense. This is an occasion to remark that Goodman and Carnap (1928) develop constructions of quality order much more elaborate than Johnson's. They do not use a primitive equivalent to *Dabc* in their constructions. Johnson returns to questions of quality order in Part II of Logic, Chapter VII, “The Different Kinds of Magnitude.”

The following conjunctive definition, which overcomes this particular difficulty, is not a revision of a formulation by Johnson. It is the beginning, rather, of a brief attempt to define betweenness by means of Johnson's primitive *Dabc*: Let us say that *b* is betwixt *a* and *c* if and only if *Dabc* and *Dcba*. Orange, but not reddish purple, is betwixt red and yellow. Figure 2 adds to the circle in figure 1 another circle with the same radius with point *c* in the middle. A point *b* betwixt *a* and *c* is within the intersection of these two circles.

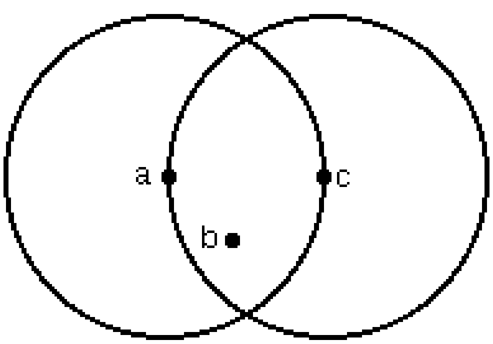


Figure 2

Betwixtness is too wide a notion to explicate betweenness. Suppose that the points in Figure 3, a specification of Figure 2, stand for the following colors:

R: a fully saturated, bright sample of red

Y: a fully saturated, bright sample of yellow

O_1: a fully saturated, bright sample of orange

O_2: a less saturated, less bright sample of orange

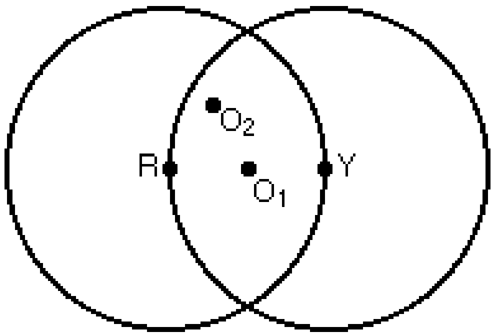


Figure 3

O_1 and O_2 are both betwixt R and Y. But it is natural to represent O_1 as 'right between' R and Y. O_2 is somewhat off to the side. A better definition of 'between' will count O_1 but not O_2 as between R and Y.

Relying again on the notion of *distance*, one can distinguish two senses of *between*. (1) *B* is *somewhere between A* and *C* if and only if the distance between *A* and *B* plus the distance between *B* and *C* is equal to the distance between *A* and *C*. That is, *B* is located somewhere on the straight line (in Euclidean space) between *A* and *B*. (2) *B* is *exactly* or *halfway between A* and *B* if and only if *B* is somewhere between *A* and *C* and also the distance between *A* and *B* is equal to the distance between *B* and *C*. The following definition, built on Johnson's primitive, attempts to define *somewhere between* in sense (1):

> *b* is somewhere between *a* and *c* if and only if *b* is betwixt *a* and *c*, and nothing is both betwixt *a* and *b* and betwixt *b* and *c*.

When two circles with no interior points in common are tangent, the point in common is on the straight line segment between the two centers. Any point on a straight line segment between points *x* and *y* is the point in common between two circles with centers *x* and *y* that have no interior points in common. As Figure 4 illustrates, *b* is on *a* straight line segment between *a* and *c* if and only if the circle *ab* is tangent to the circle *cb*. This is the case if and only if nothing is betwixt *a* and *b* and also betwixt *b* and *c*; for any such thing has to be both an interior point of circle *ab* and an interior point of circle *cb*, and these circles have no interior points in common.

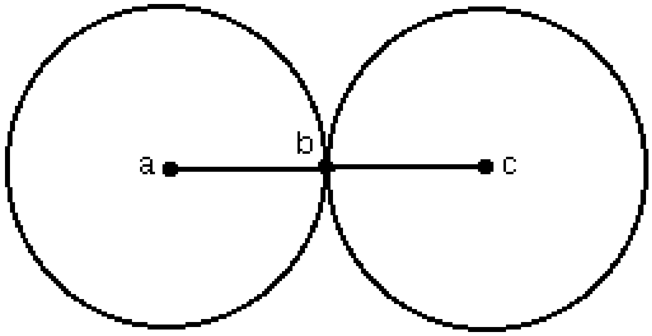


Figure 4

Figure 5, illustrates a point *b* that is betwixt *a* and *c* but is not situated like point *b* in Figure 4. In this case, point *b* is on the intersection of two circle that have interior points in common. Figure 5 shows this region as a shaded area. Anything within this shaded area is both betwixt *a* and *b* and betwixt *b* and *c*.

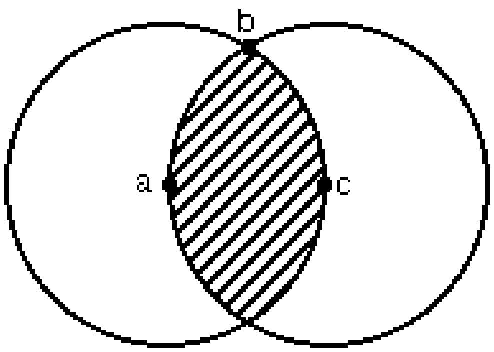


Figure 5

Although a distance between points in these diagrams can be equal, or double, or half, another distance between points, that is due to the conventions of drawing these diagrams. There has been no explication of these distance notions by means of the primitive *Dabc*. A definition of *right between* would provide a sufficient condition for the equality of the distance between *a* and *b* and the distance between *b* and *c*, but there is no attempt here to provide such a definition using only the primitive *Dabc*.

Johnson uses his notion of betweenness to draw two independent distinctions between quality order, interminable series in contrast with cyclic, and continuous series in contrast with discrete (1921, pp. 182-183). His use of the three-place relation *the difference between a and c is greater than that between a and b* undermines a point he insists on earlier that resemblance between determinates does not group them under a determinable. Johnson's three-place relation can also be expressed *the similarity between a and b is greater than that between a and c*. Comparisons between differences are also comparisons between likenesses or similarities. Perhaps his point can be expressed as follows: no two-place relation of resemblance groups determinates under a determinable, although a three-place relation can be useful for this purpose. Johnson's chapter ends with the pronouncement that "The practical impossibility of literally determinate characterization must be contrasted with the universally adopted postulate that the characters of things which we can only characterize more or less indeterminately, are, in actual fact, absolutely determinate" (1921, p. 185). Section 8 examines this claim.

2. W. E. Johnson and A. N. Prior

Johnson discuses determinates and determinables in Parts II and III of Logic (1922, 1924) and also elsewhere in Part I. In Part I, in a chapter entitled “Laws of Thought,” Johnson formulates four principles of adjectival determination that correspond to four more familiar principles of propositional determinations such as ‘Not both *P* and not-*P*’ and ‘Either *P* or not-*P*.’

In Part III, Johnson is concerned primarily with induction and causation. Throughout Part III, he distinguishes the ‘occurrent’ from the ‘continuant’ and often discusses change, cause, and continuants with reference to determinates of determinables.

In Part II, Johnson refers to determinables in several different contexts. One discussion is especially important for understanding Prior's later treatment of the topic. Johnson introduces the notion of a structural proposition which he compares to “what Kant meant by ‘analytic’” (1922, pp. 14-15). In a structural proposition, “it is impossible to realise the meaning of the subject-term without implicitly conceiving it under that category” (1922, p. 15).

Arthur N. Prior takes up the question of structural propositions that relate determinates to determinables in the two-part article “Determinables, Determinates, and Determinants” (Prior, 1949).

> Since a subject's being in a certain universe or category, i.e., its being determinable in certain ways, is presupposed in every genuine characterization of it, an assertion that it is in this category, and is thus determinable, would have for its predicate something which cannot really be separated from the subject in order to be predicated of it. (Prior, 1949, p. 18)

Prior's article reveals his very wide-ranging knowledge of the history of logic. The article together with Prior (1955), reflects a detailed knowledge of Johnson's entire logical system, not only the three-part Logic (1921, 1922, 1924), but also Johnson (1892). Although some of the topics in Prior (1949) have not prompted much subsequent discussion in connection with determinates and determinables, Prior puts his finger on one theme that is now central.

> The problem of fitting the relation between determinates and determinables into a purely “conjunctional” logic might be summarily described as the problem of justifying the inference from “This is red” to “This is coloured” on the assumption that all formal inference consists in the passage from a conjunction to one of its conjuncts. (Prior, 1949, pp. 191-192)

As mentioned earlier, there seems to be no conjunction of the proper kind of the form ‘*x* is *F* and *x* is colored’ that is equivalent to ‘*x* is red.’ Examples of improper conjunctions are:

> *x* is red and *x* is colored,
> *x* is either red or not colored, and *x* is colored.

3. The Körner-Searle Symposium

The 1959 Joint Session of the Aristotelian Society and the Mind Association included a symposium "On Determinables and Resemblance" in which Stephan Körner spoke first and John Searle spoke second.

Körner presents a logic of inexact concepts which reappears in his 1966 book Experience and Theory. This logic recognizes, in addition to traditional set members and non-members, intermediate or neutral set members. Two overlapping sets are related by exclusion-overlap if, by stipulating of each neutral candidate that it is either positive or negative, it is possible to end up with two overlapping sets and it is possible to end up with two sets related by exclusion (Körner, 1966, pp. 45-46. This clarifies or revises Körner, 1959, pp. 127-128). He gives blue and green as examples to illustrate exclusion overlap. Blue and green (strictly, the set of blue things and the set of green things) are not absolutely exact; they have neutral candidates. Since nothing is a member of the green set and also a member of the blue set, these sets exclude each other. When each neutral candidate is converted by stipulation either to a member or to a non-member, the two adjusted sets may still exclude each other, because no neutral candidate has been designated both green and blue, or if there is at least one formerly neutral candidate that is now both green and blue, the adjusted sets overlap.

Körner claims that determinates under the same determinable are linked, directly or indirectly, by exclusion-overlap. Red and green are not directly related by exclusion overlap, but they are presumably related indirectly by direct links between red and orange, orange and yellow, yellow and yellowish green, yellowish green and green. The concepts of red and green are therefore linked. Concepts are linked if and only if they are related by exclusion overlap or the ancestral of exclusion overlap (1959. pp. 130-131). Concepts *P* and *Q* are fully linked if and only if every species of *P* is linked with every species of *Q* (1959, p. 131). Full linkage is crucial to Körner's treatment of determinates and determinables.

Körner attempts to explain the determinate-determinable relation by means of full linkage. He asserts that full linkage is a stronger relation than mere linkage. It is difficult, however, to be convinced that this is true. If at least one species of *P* is linked to at least one species of *Q*, and all the species of *P* are linked to each other, as are all the species of *Q*, then every species of *P* is linked with every species of *Q*, and *P* and *Q* are fully linked. Körner claims that 'angry,' a species of 'yellow or angry,' is not linked with 'green.' This would be an interesting and important result, if true. Körner here identifies a problem that occupies subsequent discussions. How are we to distinguish ordinary predicates such as 'green' from disjunctive predicates such as 'yellow or angry'?

It appears that on Körner's own definitions, 'green' and 'yellow' are linked, 'yellow' is linked to 'yellow or angry,' and 'yellow or angry' is linked to 'angry,' so 'green' is linked to 'angry.' Until explanations are forthcoming how one or more of these alleged linkages violate Körner's requirements, the main influence of his contribution is to direct attention to the problem of disjunctive predicates.

John Searle makes a fresh start. In his attempt to explicate the determinate-determinable relation, he uses the notion of predicate entailment. In the standard sense, entailment is a relation between items, such as propositions, that have truth-values. Searle extends this notion to a relation between predicates. 'Red' entails 'colored' because it is impossible for something to be red and not colored. This is a natural extension, and others have adopted it. Indeed, talk of predicate entailment leads easily to talk of property entailment: the property red entails the property colored.

Searle and others who follow him draw a sharp distinction between the determinate-determinable relation and

the genus-species relation. (He repeats this distinction in Searle, 1967.) The definition of a species is by means of genus and differentia, which are logically independent. (Predicates *F* and *G* are independent when none of the following entailments hold: *F* entails *G*, *G* entails *F*, *F* entails not-*G*, and not-*F* entails *G*.) A determinate of a determinable cannot be defined in this way, by a conjunction of independent predicates. A traditional (although inadequate) definition ‘Man is a rational animal’ passes the genus/differentia test. ‘Rational’ and ‘animal’ are independent terms. The attempted definition ‘Red is a color that is red’ does not pass because ‘red’ entails ‘colored.’

There are both historical and logical difficulties with this view.

The genus-species relation is an ancient philosophical topic. No crisp, clear definition can be consistent with everything that has been said before. Searle's confident exposition, however, contradicts some standard views. A logic text in wide use for many decades gives the following as a rule of definition:

> The better the definition, the more completely will the differentia be something that can only be conceived as a modification of the genus: and the less appropriately therefore will it be called a mere attribute of the subject defined. (Joseph, 1916, p. 112)

Aristotle mentions differentia that entail the genus, as in ‘Walking animal’ (Topics, IV. 6) and ‘Footed animal’ (Metaphysics, Z. 12). In his Commentary on Z. 12, Bostock says that the first differentia should entail the genus (Aristotle, 1994, pp. 176-184). Other philosophers have adopted Searle's proposal, so it is evidently attractive. It does not represent a consensus of earlier writers.

The nature of conjunction poses a logical problem for Searle's account of species. If two conjunctions are logically equivalent, it does not follow that the conjuncts of one are logically equivalent to the conjuncts of the other. The forthcoming example concerns conjunctive propositional functions about pure numbers. It is easy to construct parallel examples about mass, length, temperature, years of service, taxable income, and so on.

> *Ax*: x is greater than 4 but less than 7
> *Bx*: *x* is greater than 4 but less than 6.
> *Cx*: *x* is greater than 5 but less than 7.
> *Dx*: *x* is less than 6.
> *Ex*: *x* is greater than 5.
> *Fx*: *Bx* & *Cx*.
> *Gx*: *Dx* & *Ex*.
> *Hx*: *Bx* & *Ex*.
> *Ix*: *Cx* & *D*x.

The last four predicates, *Fx*, *Gx*, *Hx*, and *Ix*, are equivalent, so they entail the same predicates and are entailed by the same predicates. They all entail *Ax*, and *Ax* entails none of them. Their conjuncts, by design, have various entailment relations. Both conjuncts of *Fx* entail *Ax*. Neither conjunct of *Gx* entails *Ax*. One conjunct of *Hx* and of *Ix* entails *Ax* and the other conjunct in each case does not.

Searle says that “a species is a conjunction of two logically independent properties—the genus and the

differentia" (1959, p. 143). Does he mean (a) that every conjunction equivalent to the species satisfies this requirement or (b) that at least one conjunction satisfies the requirement? Stipulation (a) is too difficult to satisfy, for any species is equivalent to the conjunction of itself and the genus. Stipulation (b) is too easy to satisfy, as will be shown next. The following predicates continue the numerical example:

> *Jx*: *x* is greater than 5 but less than 6.
> *Kx*: *Jx* or (*x* is greater than 2 but less than 3).
> *Lx*: *Ax* & *Kx*.

Jx and *Lx* are equivalent to each other and also to *Fx*, *Gx*, *Hx*, and *Ix*. If we consider *Ax* to be the genus, then *Lx* is a conjunction of the genus and a term *Kx* logically independent of the genus. One can perform a trick of the same kind with the standard example of color. Consider the following contrived 'genus and differentia' definition of *red* as a species of the genus *colored*:

> x is red $=_{df}$ (x is colored) & (x is red or x is not colored).

Searle's distinction between genus-species and determinate-determinable requires some principled way of excluded disjunctive predicates such as '*Kx*' and '*x* is red or *x* is not colored.' Explaining the determinate/determinable relation requires this anyway, whether or not accepts Searle's views about the relation of species to genus.

Searle attempts to define the determinate-determinable relation and to eliminate hybrid, cross-type conjunctive and disjunctive predicates by using only the relation of predicate entailment. When a predicate A entails a predicate *B*, but *B* does not entail *A*, Searle says that *A specifies B* (1959, p. 145). *A* is a non-conjunctive specifier of *B* if and only if *A* specifies *B* and there is no pair of terms *C* and *D* such that *A* is equivalent to (entails and is entailed by) the conjunction of *C* and *D*, *C* specifies *B*, *D* does not entail *B*, and not-*D* does not entail *B*. Take the letters *A*, *B*, *C*, and *D* as abbreviations for some new predicates:

> *A*: red
> *B*: colored
> *C*: colored but not green
> *D*: red or (not colored and not prime)

A specifies *B*. There are terms *C* and *D* such that *A* is equivalent to (*C* & *D*), *C* specifies *B*, *D* does not specify *B*, and not-*D* does not specify *B*. So according to this definition, red is not a non-conjunctive specifier of colored, a result that is opposite to what Searle intends. He says that a necessary condition of *A*s being a determinate of *B* is that *A* is a non-conjunctive specifier of *B*. Like everyone who addresses this topic, Searle takes the relation between red and colored to be a paradigm of the determinate-determinable relation.

Searle adds another condition with the intention of excluding disjunctive predicates such as 'yellow or angry.' A determinate of a determinable must not only be a non-conjunctive specifier of the determinable, it must be logically related to all other non-conjunctive specifiers of the determinable. Suppose that the definition of 'non-conjunctive specifier' can be emended somehow to allow red as a non-conjunctive specifier of colored. The new requirement of logical independence wrecks the project once again because it eliminates colored as a

determinable. Consider darkish red to darkish orange. Darkish red to darkish orange and red are not logically related. Some things are both; some are neither; some are the first but not the second; some are the second but not the first. So Searle's requirements again disqualify his paradigm. (A similar objection occurs in Sanford, 1970, pp. 162-163.)

4. The Munsell Color Solid

In order to construct more exact color examples, this entry will begin to specify colors by reference to the Munsell Color Solid and will use the color designations in the Inter-Society Color Council-National Bureau of Standards (ISCC-NBS) system (see Kelly and Judd, 1976, and the entry “Color” in Webster's Third Unabridged Dictionary). These standard color designation names such as **deep yellow** and **dark grayish yellow** are hereafter printed in boldface. There are 267 names in all, and a number from 1 to 267 is associated with each name. Although the context usually makes it clear when we are talking about (mentioning) a word and when we are using the word to talk about what the word is about, the difference between bold face and the ordinary font will also observe the customary use/mention distinction. The predicate **deep yellow** refers to the color deep yellow.

Any pair of surface color that the human eye can distinguish with respect to any color dimension, hue, brightness, or saturation (chroma), corresponds to a pair of points in The Munsell Color Solid. Estimates of the number of distiguishable colors range from two million to five million. For all practical purposes, one can regard the Munsell Color Solid as a continuum of colors. Pictures of the Munsell Color Solid, however, often depict the whole solid as contstructed out of 267 blocks of uniform determinate color (as in the color plates in Webster's Third or the Adobe Technical Guide to the Munsell Color System). These standard color names are ambiguous between determinate and determinable. For each of the 267 regions in the color solid, there is a determinate representative color, the ‘center of gravity’ or centroid color. These are the colors of the Centroid Color Chips that science and commerce use to standardize color descriptions of minerals, paint, dye, ink, plastic, and so forth. **Pale yellow** is both the name of a determinate centroid color and a determinable color. The scientists who construct the color solid regard the determinable use as primary.

> There are many pairs of easily distinguishable colors which receive in this system the same designation, while there are also many pairs that can scarcely be distinguished which receive different designations. This property is, of course, an unavoidable result of dividing the color solid into an arbitary number of blocks, one for each of the 267 designations. Analogous disadvantages result for identifying the time of events according to date; two events occurring on the same date may be separated by many hours, but on the other hand two scarcely separable midnight events may have to be assigned different dates (Kelly and Judd, 1976. p. 4).

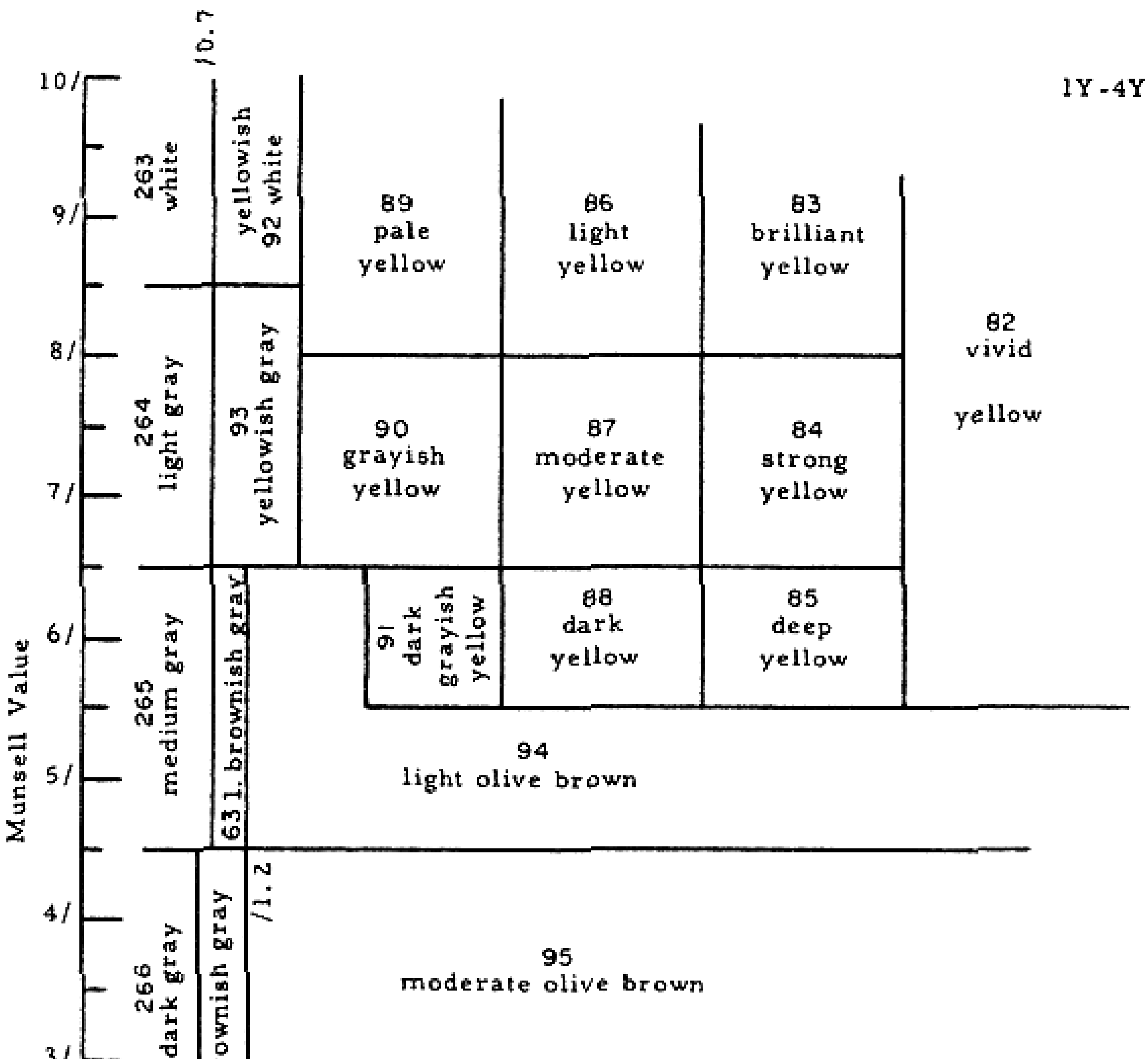
In all the forthcoming uses of the Munsell color names, they should be understood as names of determinables rather than determinates. There are distinguishable instances of pale yellow and some of these pairs are more easily distinguished than some pairs in which one color is pale yellow and the other is grayish yellow.

The bold face of the standard names contrasts in this entry with the upper case italics of invented names defined by means of the standard names. Some forthcoming definitions have the following pattern:

> *WEAK YELLOW*: **pale yellow** or **grayish yellow** (89 or 90)
>
> *ROBUST YELLOW*: **strong yellow** or **moderate yellow** or **grayish yellow** (84 or 87 or 90).

In the color solid, the regions corresponding to *WEAK YELLOW* and *ROBUST YELLOW* are as compact as any of the standard regions. These color names are as comprehensible as the standard names although, of course, they are more determinable. Weak yellow and robust yellow overlap because anything that is grayish yellow is both weak yellow and robust yellow. Neither entails the other.

Kelly and Judd contains thirty-one color charts (not themselves printed in color) that represent cross sections of the Color Solid. The vertical axis represents lightness (Munsell Value), and the horizontal axis represents saturation (Munsell Chroma). All colors that are not on the black-gray-white lightness axis have the same hue. Figure 6 is a reproduction of a chart for yellow (Kelly and Judd, 1976, p. 22) that represents the relation between pale yellow, grayish yellow, and other colors.



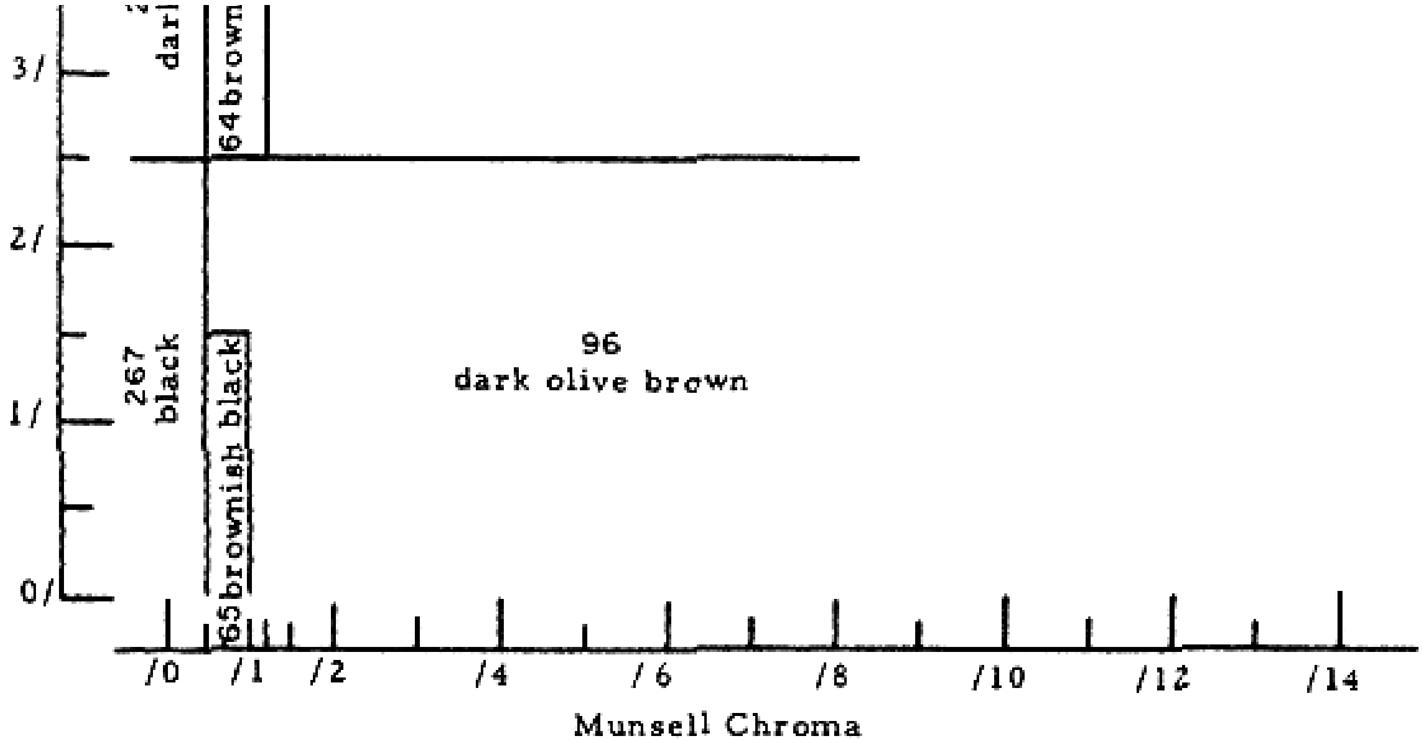


Figure 6

5. After Searle

John Woods attempts to improve Searle's definitions in “On Species and Determinates” (Woods, 1967). As Richmond Thomason demonstrates with a remorseless barrage of difficulties, Woods's efforts only make things worse, if this is possible (Thomason, 1969). Woods's requirements for determinables have, for example, the unwelcome consequence that if *Gx* is a determinable, *Gx* is a theorem of predicate logic (Thomason, 1969, pp. 95-96).

Without offering a rigorous proof, Thomason offers the opinion that the overall project of defining species-genus and determinate-determinable in terms only of entailment and negation is doomed. He proceeds “to search for an abstract, structural characterization” (p. 97) and finds an appropriate structure in the algebraic theory of semi-lattices. The resulting elegant theory is probably useful to theorists who develop taxonomic schemes. It does not, however, appear to help with the problems that Körner and Searle confront. What disqualifies colored or rectangular as a determinable of red? What disqualifies red and square as a determinate of colored?

Besides proposing lattice-theoretic requirements for natural kinds. Thomason recommends a principle of disjointness (D) (p. 98) for taxonomic systems that can be stated as follows:

> (D) If two natural kinds *a* and *b* of a taxonomic system share at least one member, then every member of *a* is a member of *b*, or every member of *b* is a member of *a*.

The two kinds *red* and *darkish red or darkish orange* fail to satisfy (D). They cannot both be natural kinds of the same taxonomic system. The algebraic theory of semi-lattices by itself provides no reason for favoring one

or the other or neither as a natural kind. Many systems of classification appear not to accord with (D). In systems of biological, physical, and chemical taxonomy, its enforcement appears arbitrary. These considerations also undermine Johnson's claim that “Taking any given determinate, there is only one determinable to which it can belong” (Johnson, 1921, p. xxxv, quoted above in Section 1.A).

Although one may nevertheless attempt to respect principle (D), it is important to realize that not every system that respects (D) divides the color solid into natural kinds. Here is another definition in terms of Munsell classification:

> *BELLOW*: **pale yellow** or **deep blue** (89 or 179).

Something pale yellow can change to grayish yellow by continuously becoming a little bit darker and without changing at all in hue or saturation and without occupying regions in the color solid other than 89 or 90. Nothing pale yellow can change to deep blue in the same way. Either the change is discontinuous or the thing occupies many regions other than 89 and 179. The predicate **pale yellow or dark blue** (*BELLOW*) is a disjunction of two determinate predicates but does not itself correspond to a determinate. **Pale yellow or grayish yellow** (*WEAK YELLOW*) is determinable with respect to its disjuncts and is a suitable determinate of ‘colored.’ Disjunctive or conjunctive syntactic forms by themselves are unreliable guides to naturalness or being a proper determinate or determinable

Dean Zimmerman has also suggested improvements to Searle's treatment of determinables. He uses, in addition to the notion of predicate or property entailment, the notion of the Boolean part of a property.

> *F* is a determinate falling under determinable $G =_{df}$ (1) *F* implies *G*, but *G* does not imply *F*; (2) there is no property *H* such that: (a) *G*&*H* implies *F* but (b) neither *H* nor not-*H* implies *G*; (3) every Boolean part of *F* implies *G*; and (4) for every property *I* such that *I* and *G* satisfy the preceding three clauses, *F* and *I* stand in some logical relation. (Zimmerman, 1997, p. 464)

How does this definition fare for our paradigm, ‘*F* (red) is a determinate falling under determinable *G* (colored)’?

Clause (3) requires that every Boolean part of the property red implies the property of being colored. From Zimmerman's discussion (1997, pp. 462-463), it is not obvious whether red has a proper Boolean part or whether its only Boolean part is red itself. Whatever the answer, clause (3) seems to be satisfied for this example. Questions remain, however, about parts of color properties. If the predicate ‘robust yellow’ stands for a property, are strong yellow, moderate yellow, and grayish yellow its Boolean parts? Or is the property of being robust yellow somehow distinct from the properties of being strong or moderate or grayish yellow? On the understanding of ‘disjunctive predicate’ to be recommended below, ‘robust yellow’ is disjunctive if and only if ‘strong or moderate or grayish yellow’ is also disjunctive. An account of disjunctive predicates, in this sense, could be useful in identifying Boolean parts of properties. An appeal to the existence or non-existence of such parts to explain disjunctiveness, on the other hand, appears to assume what it purports to establish.

Clause (1) of the definition above is satisfied and (2), unfortunately, is unsatisfied. Consider the following property:

> *H*: *red or* (*not-colored and square*)

G&*H* implies *F*, that is, *Colored* and (*red or* (*not-colored and square*)) implies *red*. *H* does not imply *G*, that is, *red or* (*not-colored and square*) does not imply *colored*. Not-*H* does not imply *G*, that is, *not-red and* (*colored or not-square*) does not imply *colored*. Zimmerman's formulation does not repair a difficulty in Searle's, namely, that the red-colored paradigm fails to satisfy the definition of the determinate-determinable relation. Predicate *H*, to be sure, is a hideous monstrosity that stands for something stapled together from Boolean parts of unrelated properties. A goal of the definitional enterprise is to distinguish ordinary, healthy predicates from such monstrosities. So we cannot assume the distinction in order to draw the distinction.

Clause (4) had difficulties of its own. Assign a new meaning to the letter '*F*':

> *F*: weak yellow

F satisfies clause (1). Pretend *F* also satisfies some revised and improved version of clause (2). Clause (3) does not stand in the way. But as one would expect, something is lurking in the wings:

> *I*: robust yellow

I satisfies clauses (1), (2), and (3) as well as *F*. But *I* and *F* are logically unrelated. All the following are possible: not-*F* and not-*I*, *F* and not-*I*, *I* and not-*F*, and *F* and *I*. An example of the kind that thwarts Searle also impedes Zimmerman.

6. Predicates and Properties

Philosophical discussion often slides back and forth between talk of predicates and talk of properties. Some philosophers suggest that there is an important logical difference between disjunctive predicates, on the one hand, and disjunctive properties or universals, on the other.

In a brisk Introduction to his early book *Problems of Mind and Matter*, in a section entitled "Generic and Specific," John Wisdom says:

> The fact is *red and hard* and *red or hard* are not universals; for strictly there are no conjunctive or disjunctive universals but only conjunctive and disjunctive facts. "This is red or hard" means "Either this is red or this is hard." (Wisdom, 1963, p. 31)

D. M. Armstrong says something very similar about disjunctive properties (1978, p. 19-23). (Armstrong is willing to admit conjunctive properties.) A disjunction of property predicates such as 'red' and 'hard' is not itself a disjunctive property predicate. There is no property of being red or hard although the disjuncts of the meaningful predicate 'red or hard' do (or might) correspond to the properties red and hard.

At this point the following distinction is relevant:

1. For some property predicates *F* and *G*, the compound predicate *F or G* is not itself a property predicate.

2. For all property predicates *F* and *G (*that do not necessarily have the same extension*)*, the compound predicate *F or G* is not itself a property predicate.

A metaphysician of properties can accept (1) or not. Accepting (2) is not an option. (2) is unacceptable. Armstrong writes:

> Disjunctive properties offend against the principle that a genuine property is identical in its different particulars. Suppose *a* has a property *P* but lacks *Q*, while *b* has *Q* but lacks *P*. It seems laughable to conclude from these premisses that *a* and *b* are identical to some respect. Yet both have the “property”, *P or Q*. (1978, p. 20)

Something that satisfies the first disjunct of the following predicate need not be identical to or resemble in any relevant respect something that satisfies the second disjunct: ‘More than ten million miles from Memphis or sings “All of Me” off key.’ This example is intended to be a disjunction of laughably unrelated components. Not every predicate of the form *P or Q* is good for a laugh in this way.

Perhaps there is no current consensus about the accepted meanings to the technical phrases ‘disjunctive predicate’ and ‘conjunctive predicate.’ If that is so, then the following principles are useful suggestions for fixing their meanings, 'a' stands here for 'acceptable' and 'u' stands for 'unacceptable':

> (Conj-a) If predicates *F* and *G* are equivalent (necessarily apply to the same things), then *F* is conjunctive if and only if *G* is also conjunctive.
>
> (Disj-a) If predicates *F* and *G* are equivalent, then *F* is disjunctive if and only if *G* is also disjunctive.

On the other hand, it is useful to reject both the following principles:

> (Conj-u) If *F* is equivalent to a predicate of the form *K and L*, then *F* is conjunctive.
>
> (Disj-u) If *F* is equivalent to a predicate of the form *K or L* then *F* is disjunctive.

According to (Conj-u) and (Disj-u), all predicates are both conjunctive and disjunctive. Even redundant predicates such as *F and F* and *F or F* demonstrate this result. Additional qualifications can of course eliminate these particular examples. Then there will be more examples with undesirable consequences, and more qualifications to eliminate them. So long as the qualifications must be expressed in the terms of standard logical dependence and independence, the project recapitulates the efforts of Searle and Woods.

According to (Conj-a) and (Disj-a), Wisdom and Armstrong can agree that disjunctive predicates do not stand for properties or universals and they can disagree about whether conjunctive predicates stand for properties or universals. Wisdom says they don't, Armstrong thinks that the arguments against disjunctive universals do not

apply to conjunctive universals. Everyone should agree that if a predicate *F* is equivalent to a disjunction of two different property predicates, *F* may be disjunctive or may not. That depends on how the disjuncts are related. Similar remarks apply to conjunctive predicates.

Some earlier examples of predicates, or similar predicates, appear in the following list:

A. **pale yellow** or **grayish yellow** (*WEAK YELLOW*),
B. **pale yellow** or **deep blue** *(BELLOW)*,
C. (greater than 5 and less than 7) or (greater than 4 and less than 6),
D. (greater than 5 and less than 7) and (greater than 4 and less than 6),
E. yellow or angry,
F. yellow and angry.

So far as one can discriminate just by means of predicate entailment or the presence or absence of logical relations, (A) and (B) are similar. Each is a disjunction of predicates that exclude each other.

Pale yellow and grayish yellow do not differ with respect to hue or saturation. They differ in brightness only to the extent necessary to have distinct locations in the Munsell color solid. Pale yellow and grayish yellow are determinate with respect the determinable weak yellow, and weak yellow in turn is determinate with respect to yellow.

Bellow is not a determinate color with respect to the determinable color. One wants to deny that it is a color at all, even though *BELLOW* is equivalent to a disjunction of color predicates.

Predicates (A)and (B) contrast sharply. **Pale yellow** and **grayish yellow** are as similar as they can be while still excluding each other. **Pale yellow** and **deep blue** are about as dissimilar as two colors can be. Direct ungrounded appeals to resemblance or being in the same dimension or having to do with one another will not solve our problem. A theoretically satisfactory treatment of the determinate-determinable relation should explain these resemblances rather than be explained by them. A new technical term *disjoint* marks the apparent difference between (A) and (B). **Pale yellow** and **deep blue** are disjoint predicate. **Pale yellow** and **grayish yellow** are not disjoint. The next section provides a definition of disjointness.

So far as one can discriminate just by means of predicate entailment or the presence or absence of logical relations, the components of (C) and (D) are logically unrelated. But (C) is not a conjunctive predicate, and (D) is not a disjunctive predicate. Something that satisfies both disjuncts of (C), ‘greater than 5 and less than 7’ and ‘greater than 4 or less than 6,’ can change continuously so as to satisfy the first but not the second, or can change continuously in the other direction along the same dimension so as to satisfy the second but not the first. ‘Greater than 5 and less than 7’ and ‘greater than 4 and less than 6’ obviously indicate overlapping intervals along the same dimension. (C) is a long-winded way of expressing ‘greater than 4 and less than 7’ which has no appearance of being disjunctive. In the same way, (D) is a long-winded way of expressing greater than 5 and less than 6’ which has no appearance of being conjunctive.

So far as one can discriminate just by means of predicate entailment or the presence or absence of logical relations, the components of (E) and (F), these components are related to each other in the same way as those

of (C) and (D). ‘Yellow’ and ‘angry’ are logically independent. A puzzle that the Korner-Searle Symposium poses is still unsolved. (E) is a disjunctive predicate. ‘Yellow’ is not a determinate of the determinable ‘yellow or angry’. (F) is a conjunctive predicate. ‘Yellow and angry’ is not a determinate of the determinable ‘yellow.’

Logically independent predicates can be determinates under the same determinable. ‘Yellow’ and ‘angry’ have an independence of a kind, indicated here by the new technical term *B-independence*, that has conditions in addition to those for logical independence. The next section specifies some conditions of *B-independence*.

7. Boundaries and Borderlines

‘*B-independence*’ stands for ‘boundary independence.’ The Munsell Color Solid is composed of non-overlapping regions that all share boundaries with other regions. The preliminary discussion in this section departs for a while from colors to consider a familiar, two-dimensional array of non-overlapping regions, the states in the United States. In case a map of the United States is not close at hand, the reader can look at Figure 7 , a simple map of the some of the states that figure in the following examples.

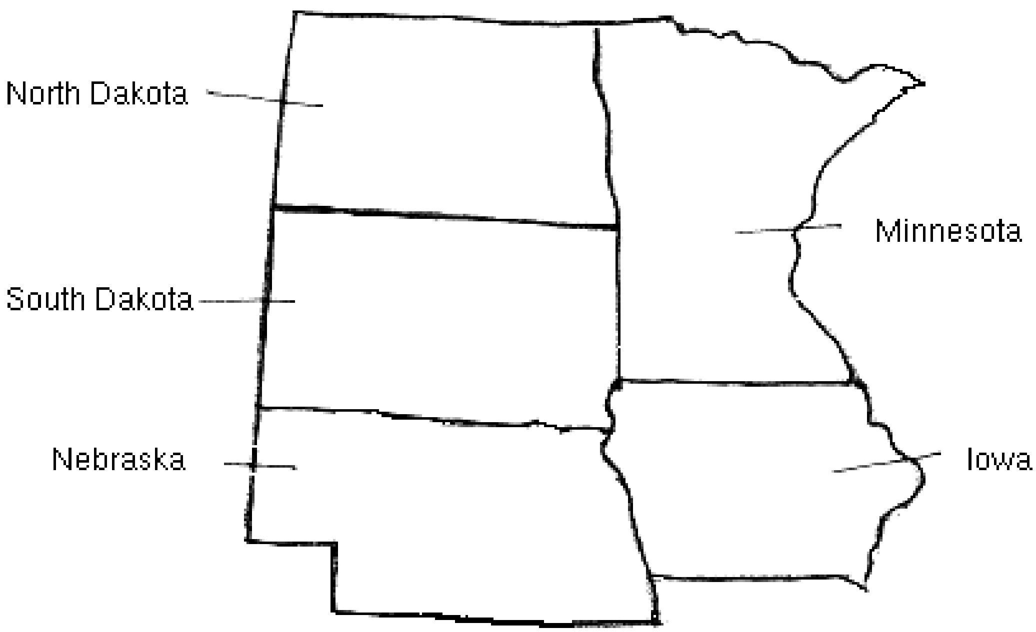


Figure 7

Some spatial analogies about the boundaries of states and other regions composed of the states will motivate the forthcoming discussion of *disjoint* and *B-independent* predicates. Here are two disjunctive definitions:

> x is in the Dakotas $=_{df}$ x is in North Dakota or x is in South Dakota.

> x is in the North States $=_{df}$ x is in North Dakota or x is in North Carolina.

In each case, the disjuncts exclude each other. North Dakota and South Dakota have no points in common. Neither do North Dakota and North Carolina. Predicate entailment fails to capture a topological difference between the Dakotas and the North States. The Dakotas are a coherent, continuous region. The North States

are a discontinuous region. Something can be both on the boundary of North Dakota and on the boundary of South Dakota. Nothing can be both on the boundary of North Dakota and on the boundary of North Carolina, for their boundaries are many miles apart. North Dakota and North Carolina are disjoint. North Dakota and South Dakota are not disjoint.

There is a close analogy here with some the color examples in the last section. Weak yellow is a coherent, continuous region in the quality space of color. Bellow is a discontinuous region. Something can be both on the boundary of pale yellow and on the boundary of grayish yellow. Nothing can be both on the boundary of pale yellow and deep blue. This is a topological difference that predicate entailment does not represent.

Here are two more disjunctive geographical definitions. In the first definition, the disjuncts exclude each other:

> *x* is in Dabraksa $=_{df}$ *x* is in South Dakota or *x* is in Nebraska.

In the second definition, the disjuncts overlap; they are logically independent; neither includes the other; they do not jointly exhaust the total space:

> *x* is in Longkota $=_{df}$ *x* is in the Dakotas or *x* is in Dabraska.

Longkota is a coherent, continuous region. There is nothing inherently disjunctive about it. Boundary relations again indicate topological relations between the Dakotas and Dabraska that logical entailment and non-entailment do not capture. Consider something *A* on the boundary of South Dakota and Minnesota and Iowa. It is on the boundary of the Dakotas and on the boundary of Debraska and on the boundary of Longkota. But it is not on the boundary of the following two regions:

> *x* is in the Dakotas but *x* is not in Debraska (that is, *x* is in North Dakota)

> *x* is in Debraska but *x* is not in the Dakotas (that is, *x* is in Nebraska).

A is not close to any boundary of North Dakota or Nebraska.

One of our main puzzle predicates, 'yellow or angry,' presents a topological contrast. Anything *A* on the boundary of 'yellow' and on the boundary of 'angry' is also on the boundary of the following four predicates:

> 'yellow and angry'
> 'yellow but not angry'
> 'angry but not yellow'
> 'neither angry nor yellow'

This reflects the fact that being yellow and being angry are conceptually independent besides being logically independent. Slight changes in *a* that would move it from the boundary of 'yellow' to being definitely yellow or definitely not yellow are independent of slight changes in *a* that would move it from the boundary of 'angry' to being definitely angry or definitely not angry.

It is possible to define a color predicate analogous to Longkota. This definition uses two earlier definitions that are repeated here:

> *ROBUST YELLOW*: **strong yellow** or **moderate yellow** or **grayish yellow**.
>
> *WEAK YELLOW*: **pale yellow** or **grayish yellow**
>
> *SWELL YELLOW*: *WEAK YELLOW* or *ROBUST YELLOW*.

The disjuncts of this last definition are logically independent. Nevertheless, *SWELL YELLOW* corresponds to a coherent continuous region in color space. Weak yellow and robust yellow have a topological relation that yellowness and anger do not have. There are locations in the Munsell Color Solid (for example, on the borderline of **grayish yellow** and on the borderline of **light olive brown**), where something *x* at this location is on the boundary of weak yellow and on the boundary of robust yellow and on the boundary of swell yellow but is not on the boundary of the following two regions:

> Weak yellow but not robust yellow.
> Robust yellow but not weak yellow.

Although the predicates *WEAK YELLOW* and *ROBUST YELLOW* are logically independent, relations between the boundaries of their regions indicate a significant, objective connection.

Although **pale yellow** is a highly determinate color predicate relative to ‘yellow’, it is far from being maximally determinate. A sentence from Kelly and Judd, quoted above in Section 4, is repeated here:

> There are many pairs of easily distinguishable colors which receive in this system the same designation, while there are also many pairs that can scarcely be distinguished which receive different designations(Kelly and Judd, 1976. p. 4).

Despite its relative specificity, **pale yellow** applies to samples that are visibly different with respect to color. The same goes for **grayish yellow**. So something can change gradually from pale yellow to grayish yellow. Is there some point along the way that is the precise boundary between these two color regions? This is a specific form of a question that divides philosophers who develop theories of vagueness. Without needing to adopt some view about the basic nature of borderline cases, one can admit the possibility of borderline cases between pale yellow and grayish yellow. It is possible that something can be on the borderline of each region. Something is a borderline case of pale yellow if it is neither definitely pale yellow nor definitely not pale yellow.

Körner also uses a logic of inexact concepts to treat ‘yellow or angry.’ One of Searle's complaints about Körner is that "His definition excludes any exact concept as a possible candidate for a determinate" (Searle, 1959, p. 156). Does this objection apply to the suggestions in this section? The next section returns to the question whether there are perfectly exact concepts. Borderline cases are used in this section to locate boundaries. Suppose that ‘more than five feet tall but less than six feet tall’ is perfectly exact. Anything just slightly taller than five feet or just slightly shorter than six feet is on the boundary. For the purposes of

exploring relations to other predicates, we can replace an exact predicate with one that it slightly inexact. For example, amend the definition of the allegedly exact predicate by adding ‘so far as one can tell by using a wall, a pencil, a carpenter’s level, and a yardstick.’ The definitions as amended definitely apply to some things, definitely do not apply to others, and also have some borderline cases left over. (Following a harmless practice, this article refers both to borderline cases of predicates and to borderline cases of properties or regions.)

This project treats borderlines and boundaries as interchangeable. The following is an attempt to generalize and formalize the suggestions made above:

> The ‘**B**’ operator is used to talk about boundaries and borderline cases. ‘**B***Fx*’ means ‘*x* is a borderline case of *F*.’

The definition of *disjoint predicates* promised at the end of the section follows:

> *Fx* and *Gx* are disjoint predicates if and only if *Fx* and *Gx* are exclusive predicates and For any *x*, if **B***Fx*, then not-**B***Gx*.

Disjoint predicates do not, or in a modal version, cannot, share borderline cases. A predicate is *exclusively disjunctive* if and only it is equivalent to a disjunction of disjoint predicates.

A specification of a condition of **B**-*independence* was also promised at the end of Section 5. Let us say that two predicates *Fx* and *Gx* intersect if and only if there is something *x* such that:

> **B**(*F* & *G*)*x* & **B**(*F* & not-*G*)*x* & **B**(not-*F* & *G*)*x* & (not-*F* & not-*G*)*x*.

The boundaries of **B**-*independent* predicates not only intersect; they intersect wherever they have a point in common. *Fx* and *Gx* are **B**-*independent* only if:

> For any *x*, if **B***Fx* and **B***Gx*, then **B**(*F* & *G*)*x* & **B**(*F* & not-*G*)*x* & **B**(not-*F* & *G*)) & **B**(not-*F* & not-*G*)*x*.

A predicate is *inclusively disjunctive* only if it is equivalent to a disjunction of **B**-*independent* predicates. A predicate is *conjunctive* only if it is equivalent to a conjunction of **B**-*independent* predicates. ‘Yellow or angry’ is inclusively disjunctive. *WEAK YELLOW* or *ROBUST YELLOW* is not inclusively disjunctive. ‘Yellow and angry’ is conjunctive. ‘*WEAK YELLOW* and *ROBUST YELLOW*’ is not conjunctive.

This approach appears to solve the puzzle that Körner formulated and that Searle and others attempted to solve without success. Conjunctive predicates do not correspond to determinates. Disjunctive predicates do not correspond to determinables.

The discussion above provides only a necessary condition of B-independence. Attempting to deal with Nelson Goodman's puzzle about ‘grue’ and other perverse artificial predicates requires reference to more complicated relations between boundary conditions. These further conditions which are represented as both necessary and

sufficient for **B**-*independence* are not spelled out here. They appear, in successive versions, in three articles by Sanford 1970, 1981 (in which the later parts are nearly incomprehensible because of over-compression and lack of diagrams), and 1994 (which has some diagrams). All three articles attempt to clarify the determinate-determinable relation by explaining the nature of disjunctive and conjunctive predicates.

8. Absolute Determinacy

“The practical impossibility of literally determinate characterization must be contrasted with the universally adopted postulate that the characters of things which we can only characterise more or less indeterminately, are, in actual fact, absolutely determinate.” This is the final sentence of W. E. Johnson's chapter “The Determinable.” Johnson is not the only philosopher who holds that things are absolutely determinate. One of D. M. Armstrong's six numbered refutations of phenomenalism in Armstrong (1961) maintains that “physical objects, which are determinate, cannot be constructions out of indeterminate sense-impressions” (p. 58).

> A physical object is determinate in all respects, it has a perfectly precise colour, temperature, size, etc. It makes no sense to say that a physical object is light-blue in colour, but is no definite shade of light blue. (p. 59)

Understanding what it is for color, temperature, size, etc., predicates to be perfectly precise helps in understanding what it is for color, temperature, size, etc., properties to be perfectly precise. A precise predicate is not vague; it is exact rather than inexact; it has no borderline cases. Precision contrasts with vagueness.

Specificity, on the other hand, contrasts with generality. **Light blue** is more specific than ‘blue’ which is more specific than ‘colored.’ The more specific a predicate, the narrower the range it covers. Is **light blue** more specific than ‘smooth’? The absence of an inclusion relation in either direction makes it difficult to answer this question. Attempts actually to compare the numerical results after counting all the light blue things in the world and all the smooth things can lead only to frustration and failure. This entry does not address the problem of comparing degrees of specificity of determinates under different determinables.

Specificity and exactness are independent in several ways. There can be predicates *F* and *G* such that:

> *F* and *G* are not identical and are both unspecific and inexact. For example: *F*: about the size of a cat, *G*: about the size of a dog.
>
> *F* is more specific and more exact than *G*. Example: *F*: **pale yellow** and *G*: yellow (in the ordinary rather inclusive sense).
>
> *F* is more specific than *G*, and *G* is more exact than *F*. Example: *F*: about the size of a cat. *G*: has a volume not less than 50.3 cubic inches and not more than 2000.8 cubic inches. Anything *F* is *G* and not everything *G* is *F*, so *F* is more specific than *G*. But *G* is more exact. It requires, at the boundaries, determination to the nearest tenth of a cubic inch.
>
> *F* and *G* are both specific and exact. Examples: *F*: **pale yellow**, *G*: **deep blue**. Neither of these

> color predicates is absolutely precise, but each is quite precise compared to ordinary color terms. The Munsell Color Solid is constructed with the intention that each of the 267 regions has approximately the same degree of specificity.

Although specificity and precision are independent in these ways, they are also significantly connected with respect to absolute determinacy. Any absolutely specific predicate is also absolutely precise. Suppose, for example, that ‘Armstrong blue’ is a predicate for an absolutely specific shade of blue. Two things that are Armstrong blue do not differ at all with respect to hue or brightness or saturation. Given a predicate ‘*F*’ and two objects *a* and *b* such that *a* is a borderline case of ‘*F*’ and *b* is a definite positive case of ‘*F*’, *a* and *b* differ along some relevant dimension. But anything that differs along any relevant dimension from something that is definitely *F*, when ‘*F*’ is an absolutely specific predicate, is not *F*. Absolutely specific predicates cannot have borderline cases.

No predicate that can have borderline cases is absolutely specific. So if there are no absolutely precise or exact predicates, neither are there absolutely specific predicates. Johnson presumably would not question this conclusion since he says that literally determinate characterization is practically impossible. He and Armstrong claim things in the world are absolutely determinate, not the predicates we do or could use to apply to things in the world.

If things in the world are absolutely determinate, this presumably does not require that any absolute determination persists through the passage of time or space. If a cumulus cloud changes continuously in shape and size, this does not by itself preclude its having, at any one time, an absolutely determinate shape and size. The impression that clouds do not have exact boundaries, however, is probably not based entirely on their changeability.

Objects with absolutely determinate sizes have absolutely determinate boundaries. If a thing has an absolutely determinate length along some axis at a given time, then there is exactly one real number *n* such that its length in (say) meters is *n*. Things that we come across in ordinary life such as plants, animals, buildings, furniture, electronic equipment, clothing, and kitchenware do not have exact boundaries, nor do larger items such as mountains, lakes, continents, and stars.

The view that things in the world are absolutely determinate is implausible if it requires clouds, brains, and dinner plates to be absolutely determinate. But this requirement can be put aside. If the microstructure of the world is absolutely determinate, that is absolute determinacy enough. If all the atoms within you and in your vicinity have absolutely determinate properties, then the indeterminate mass and shape and volume of you, your brain, and your teeth somehow supervene on the determinate microstructure. Here is a well-known passage from David Lewis:

> The reason it's vague where the outback begins is not that there's this thing, the outback, with imprecise borders; rather there are many things, with different borders, and nobody has been fool enough to try to enforce a choice of one of them as the official referent of the word ‘outback.’ (Lewis, 1986, p. 212)

The view that there are many things with precise borders does not by itself refute the view that there are things

with imprecise borders. Here is a kind of parody of the passage just quoted. It ends with the same point but begins with a contrary contention:

> The outback is a big thing, and it is vague where it begins. The reason it has imprecise borders is that there are many things, with different precise borders, and nobody has been fool enough to enforce a choice of one of them as the official reference of the word ‘outback’.

For the outback (a cloud, your brain), there are many more-or-less precise aggregates of particles such that each is about as good a candidate as there is to be identified with the outback (a cloud, your brain). This entry does not address the problem of the many, how to understand the relation between a single macro-object and many overlapping aggregates of micro-objects that more or less coincide with it.

However one resolves the problem of the many, the question of absolute determinacy becomes the question of absolute determinacy of the physical basis, the microstructure. After quoting the passage from Lewis above, Roberto Casati and Achille Varzi write:

> There are plenty of objects out there—plenty of slightly distinct and yet precisely determinate aggregates of land molecules. And when we say ‘Mount Everest’ or ‘the outback’, each one of a large variety of such aggregates—each with its own perfectly crisp mereotopological structure—has an equal claim to being a referent of that term. (Casati and Varzi, 1999, p. 95)

And what evidence is there for this precisely determinate perfect crispness? Logic and metaphysics cannot answer this question from its own resources. Science textbooks represent particles and atoms as clouds. Textbook writers fifty years ago knew that the picture of perfect little spheres, the electrons, in elliptical orbits around a nucleus was misleading. Now the picture is simply obsolete.

An attempt to measure the precise dimensions of a polished copper cube might begin using an ordinary school supply ruler, then using a machinist's steel rule, then using a micrometer, then, starting with a low power optical microscope, using a series of increasingly powerful microscopes. At the microscopic level one can discriminate increments of length too small for a mechanical micrometer to detect. This does not produce a more precise determination of the length of the cube if nothing at this level coincides with the boundary of the cube. So the search for the exact measurements of the cube is abandoned and replaced by a hope to find absolute determinacy somewhere at the foundations, the fundamental basis, or the limit. There is a (weak) kind of non-deductive argument here.

> Given a greatest degree of precision determined by the best instruments, sooner or later a more advanced technology produces instruments that are still more precise. This process of making measurement increasingly precise never ends; it asymptotically approaches absolute precision at dimensionless points of matter or spacetime or something.

This vision of absolute determinacy at the limit is apparently attractive. It appears to be internally consistent. It also appears, however, to be inconsistent with physics.

Middle-sized objects do not have perfectly precise boundaries because there are microscopic objects that are

neither definitely included nor definitely excluded from the object. Some larger microscopic objects lack perfectly precise boundaries for the same reason. There is no reason to believe that this process continues infinitely downward. Electron diameters are imprecise, but not because there are swarms of micro-electron-dust, each particle of which is also a swarm of something even smaller. Nor does the process stop with some basic items that really are absolutely determinate.

Instruments can measure the velocity of a tennis ball. They do not, of course, measure velocity with absolutely determinacy. They do not discriminate, say, 114.0 from 114.1 miles per hour. Given some understanding of margins of error, it is meaningful for one to say that a tennis ball was going 114 miles per hour at some temporal instant *t*. The notions of a limit and of convergence provide this meaning. They provide no support for believing in the possibility of a momentary tennis ball that exists neither before instant *t* nor after instant *t* but does exist precisely at instant *t* and travels 114 miles per hour during its instantaneous existence. If it is possible for something to have a property for an instant, it does not follow that an instantaneous thing can have that property.

The same goes for spatial points. If it is possible for something to have a property at a point, it does not follow that it is possible that something punctiform should have this property. When a region is pale yellow, we can say that any point in the region is pale yellow. But no point by itself can be pale yellow.

Johnson said that it is a "universally adopted postulate that the characters of things which we can only characterise more or less indeterminately, are, in actual fact, absolutely determinate." In saying it is a postulate, Johnson does not mean we merely assume it in order to deduce its consequences. He means rather that it is both obviously true and cannot be inferred from truths that are even more obvious. But the so-called postulate seems not to be true.

Bibliography

- Aristotle (1994), *Metaphysics*, Books *Z* and *H*, translated with a commentary by David Bostock, Oxford: Oxford University Press.
- Armstrong, D. M. (1961), *Perception and the Physical World*, London: Routledge and Kegan Paul.
- ------ (1978), *A Theory of Universals* (Volume II of *Universals and Scientific Realism*), Cambridge: Cambridge University Press.
- Carnap, Rudolf (1928), *Der Logishche Aufbau der Welt*, Berlin: Benary. Translated by Rolf A. George as *The Logical Structure of the World* (1967), Berkeley and Los Angeles: University of California Press.
- Casati, Roberto and Varzi, Achille C. (1999), *Parts and Places*, Cambridge: MIT Press.
- Chisholm, Roderick M. (1987), "Brentano and One-Sided Detachability," *Conceptus*, 53-54, pp. 153-159.
- Edwards, Paul, and Pap, Arthur (1973), *A Modern Introduction to Philosophy*, Third Edition, New York: The Free Press.
- Goodman, Nelson (1951), *The Structure of Appearance*, Cambridge: Harvard University Press.
- Johnson, W. E. (1892), "The Logical Calculus", *Mind*, I, New Series, Part I, pp. 3-30, Part II, pp. 235-250. Part III, pp. 340-347.
- ------ (1921), *Logic*, Part I, Cambridge: Cambridge U. P.

- ------ (1922), *Logic*, Part II, Cambridge: Cambridge U. P.
- ------ (1924), *Logic*, Part III, Cambridge: Cambridge U. P.
- Joseph, H. W. B (1925), *An Introduction to Logic*. 2nd edition, Oxford: Oxford University Press. The first edition of this book was printed in 1906.
- Kelly, Kenneth L. and Judd, Deane B., (1976), *Color: Universal Language and Dictionary of Names*, Washington: National Bureau of Standards.
- Körner, Stephan (1959), "On Determinables and Resemblance, I," *The Aristotelian Society Supplementary Volume*, XXXIII, London: Harrison and Sons, pp. 125-140.
- ------ (1966), *Experience and Theory*, London: Routledge and Kegan Paul.
- Prior, Arthur N. (1949), 'Determinables, Determinates, and Determinants,' *Mind*, LVIII, Part I, pp. 1-20, Part II, pp. 178-194.
- ------ (1962), *Formal Logic*, Second Edition, Oxford: Oxford U. P. The first edition of this book was published in 1955.
- Sanford, David H. (1970), "Disjunctive Predicates," *American Philosophical Quarterly*, 7, pp. 162-170.
- ------ (1981), "Independent Predicates," *American Philosophical Quarterly*, 18, pp. 171-174.
- ------ (1994), "A Grue Thought in a Bleen Shade: 'Grue' as a Disjunctive Predicate," *Grue! The New Riddle of Induction*, edited by Douglas Stalker, Chicago and La Salle: Open Court, pp. 173-192.
- ------ (1999), "Determinable," *The Cambridge Dictionary of Philosophy*, Second edition.
- Searle, John (1959), "On Determinables and Resemblance, II," *The Aristotelian Society Supplementary Volume*, XXXIII, London: Harrison and Sons, pp. 141-158.
- ------ (1967), Determinables and Determinates," *The Encyclopedia of Philosophy*, edited by Paul Edwards, New York: Macmillan, Volume II, pp. 357-359.
- Thomason, Richmond (1969), "Species, Determinables and Natural Kinds, *Noûs*, III, pp. 95-101.
- *Webster's Third New International Dictionary of the English Language Unabridged* (1961).
- Wisdom, John (1963), *Problems of Mind and Matter*, Cambridge: Cambridge University Press. This book was first printed in 1934.
- Woods, John (1967), "Species and Determinables," *Noûs* , I. pp. 243-254.
- Zimmerman, Dean W. (1997), "Immanent Causation," *Philosophical Perspectives*, 11, *Mind Causation, and World*, 11, pp. 433-471.

Other Internet Resources

- Munsell Color System
- The Argument from Classification: Determinates vs. Determinables (Mark Armstrong, English, Virginia Tech)
- Berkeley's Likeness Principle (excerpt from article by Phillip D. Cummins, in *Journal of the History of Philosophy*, Volume 4, 1966)

Related Entries

induction: new problem of | induction: problem of | Prior, Arthur | properties | vagueness

David Sanford
Duke University
dhs@duke.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 26, 2002
Content last modified: April 26, 2002

Relative Identity

Identity is often said to be a relation each thing bears to itself and to no other thing (e.g., Zalabardo, 2000). This characterization is clearly circular ("no *other* thing") and paradoxical too, unless the notion of "each thing" is qualified. More satisfactory (though partial) characterizations are available and the idea that such a relation of absolute identity exists is commonplace. Some, however, deny that a relation of absolute identity exists. Identity, they say, is relative: It is possible for objects *x* and *y* to be the same *F* and yet *not* the same *G*, (where *F* and *G* are predicates representing kinds of things (apples, ships, passengers) rather than merely properties of things (colors, shapes)). In such a case 'same' cannot mean absolute identity. For example, the same person might be two different passengers, since one person may be counted twice as a passenger. If to say that *x* and *y* are the same person is to say that *x* and *y* are persons and are (absolutely) identical, and to say that *x* and *y* are different passengers is to say that *x* and *y* are passengers and are (absolutely) distinct, we have a contradiction. Others maintain that while there are such cases of "relative identity," there is also such a thing as absolute identity. According to this view, identity comes in two forms: trivial or absolute and nontrivial or relative (Gupta, 1980). These maverick views present a serious challenge to the received, absolutist doctrine of identity. In the first place, cases such as the passenger/person case are more difficult to dismiss than might be supposed (but see below, §3). Secondly, the standard view of identity is troubled by many persistent puzzles and problems, some of recent and some of ancient origin. The relative identity alternative sheds considerable light on these problems even if it does not promise a resolution of them all.

A word about notation. In what follows, lower case italic letters '*x*', '*y*', etc., are used informally either as variables (bound or free) or as (place-holders for) individual constants. The context should make clear which usage is in play. Occasionally, for emphasis or in deference to logical tradition, other expressions for individual constants are employed. Also, the use/mention distinction is not strictly observed; but again the context should resolve any ambiguity.

- Related Entries

1. The Standard Account of Identity

[Note: The following material is somewhat technical. The reader may wish to casually review it now and return to it as needed, especially in connection with §5. The propositions **Ref**, **LL**, **Ref′**, **LL′**, **NI**, and **ND** are identified in the present section and are referred to as such in the rest of the entry.]

Identity may be formalized in the language *L* of classical first-order logic (FOL) by selecting a two place predicate of *L*, rewriting it as '=', and adopting the universal closures of the following two postulates:

(**Ref**): $x = x$
(**LL**): $x = y \rightarrow [\wp(x) \rightarrow \wp(y)]$,

where the formula $\wp(x)$ is like the formula $\wp(y)$ except for having occurrences of x at some or all of the places $\wp(y)$ has occurrences of y (see Enderton, 2000, for a precise definition). Ref is the principle of the *reflexivity of identity* and LL (*Leibniz' Law*) is the principle of the *indiscernibility of identicals*. It says in effect that identical objects cannot differ in any respect. The other characteristic properties of identity, *symmetry* ($x = y \rightarrow y = x$), and *transitivity* ($x = y$ & $y = z \rightarrow x = z$), may be deduced from Ref and LL. Any relation that is reflexive, transitive, and symmetric is called an 'equivalence relation'. Thus, identity is an equivalence relation satisfying LL. But not all equivalence relations satisfy LL. For example, the relation *x and y are the same size* is an equivalence relation that does not satisfy LL (with respect to a rich language such as English).

Let *E* be an equivalence relation defined on a set *A*. For *x* in *A*, [*x*] is the set of all *y* in *A* such that *E*(*x*, *y*); this is *the equivalence class of x determined by E*. The equivalence relation *E* divides the set *A* into mutually exclusive equivalence classes whose union is *A*. The family of such equivalence classes is called 'the partition of *A* induced by *E*'.

Now let *A* be a set and define the relation *I*(*A*,*x*,*y*) as follows: For *x* and *y* in *A*, *I*(*A*,*x*,*y*) if and only if for each subset *X* of *A*, either *x* and *y* are both elements of *X* or neither is an element of *X*. This definition is equivalent to the more usual one identifying the identity relation on a set *A* with the set of ordered pairs of the form <*x*,*x*> for *x* in *A*. The present definition proves more helpful in what follows.

Suppose for the moment that we do not assign any special interpretation to the identity symbol. We treat it like any other two place predicate. Let *M* be a structure for *L* and assume that Ref and LL are true in *M*. Call the relation defined in *M* by the conjunction of Ref and LL 'indiscernibility' (see Enderton, 2000, for the definition of definability in a structure). There are three important points to note about the relationship between indiscernibility, and the relation *I*(*A*,*x*,*y*). First, indiscernibility need not be the relation *I*(*A*,*x*,*y*)

(where *A* is the domain of the structure). It might be an equivalence relation *E* having the property that for some elements *u*,*v*, of the domain, *E*(*u*,*v*) holds, although *I*(*A*,*u*,*v*) fails. Secondly, there is no way to "correct for" this possibility. There is no sentence or set of sentences that could be added to the list beginning with Ref and LL that would guarantee that indiscernibility coincides with *I*(*A*,*x*,*y*). This fact is usually expressed by saying that identity is not a first-order or "elementary" relation. (For a proof, see Hodges, 1983.) However, in a language such as set theory (as usually interpreted) or second-order logic, in which there is a quantifier 'all *X*' permitting quantification over all subsets of a given set, *I*(*A*,*x*,*y*) is definable.

Third, given any structure *M* for *L* in which Ref and LL are true, there is a corresponding structure *QM*, the 'quotient structure' determined by *M*, in which indiscernibility *does* coincide with *I(A*,*x*,*y*). *QM* is obtained in roughly the following way: Let the elements of *QM* be the equivalence classes [*x*], for elements *x* of *M* determined by indiscernibility in *M*. If *F* is a one-place predicate true in *M* of some object *x* in *M*, then define *F* to be true of [*x*] in *QM*, and similarly for many-place predicates and constants. It can then be shown that any sentence true in *M* is true in *QM*, and vice versa. The existence of quotient structures makes it possible to treat the identity symbol as a logical constant interpreted in terms of *I(A*,*x*,*y*). There is in fact in general no other way to *guarantee* that Ref and LL will hold in every structure. (As Quine (1970) points out, however, a *finite* language will always contain a predicate satisfying Ref and LL in any structure; cf. Hodges, 1983.) The alternative, however, is to view FOL with Ref and LL ($FOL^=$) as a proper theory in whose models (structures in which Ref and LL hold) there will be an equivalence relation *E* such that if *E*(*x*,*y*) holds, then *x* and *y* will be indiscernible with respect to the *defined* subsets of the domain. But we cannot in general assume that every subset of the domain is definable. If the domain is infinite, *L* runs out of defining formulas long before the domain runs out of subsets. Nonetheless, a strong metatheorem asserts that any set of formulas that has a model, has a countable (finite or denumerable) model. This means that the difference between indiscernibility and *I(A*,*x*,*y*) is minimized at least to the extent that, for a sufficiently rich language such as *L*, the valid formulas concerning indiscernibility (i.e., the formulas true in every model of what is termed below 'the pure *L*-theory with identity') coincide with the valid formulas concerning *I*(*A*,*x*,*y*). (See Epstein, 2001 for a sketch of a proof of this fact.) This is not to say, however, that there isn't a significant difference between identity *qua* indiscernibility and identity *qua* *I*(*A*,*x*,*y*) (see below). Both points of view -- that $FOL^=$ is a proper theory and that it is a logic -- may be found in th literature (Quine, 1970). The latter is the more usual view and it will count here as part of the standard account of identity.

Assume that L' is some fragment of *L* containing a subset of the predicate symbols of *L* and the identity symbol. Let *M* be a structure for L' and suppose that some identity statement $a = b$ (where *a* and *b* are individual constants) is true in *M*, and that Ref and LL are true in *M*. Now expand *M* to a structure M' for a richer language -- perhaps *L* itself. That is, assume we add some predicates to L' and interpret them as usual in *M* to obtain an expansion M' of *M*. Assume that Ref and LL are true in M' and that the interpretation of the terms *a* and *b* remains the same. Is $a = b$ true in M'? That depends. If the identity symbol is treated as a logical constant, the answer is "yes." But if it is treated as a non-logical symbol, then it can happen that $a = b$ is false in M'. The indiscernibility relation defined by the identity symbol in *M* may differ from the one it defines in M'; and in particular, the latter may be more "fine-grained" than the former. In this sense, if identity is treated as a logical constant, identity is *not* "language relative;"

whereas if identity is treated as a non-logical notion, it *is* language relative. For this reason we can say that, treated as a logical constant, identity is 'unrestricted'. For example, let L' be a fragment of L containing only the identity symbol and a single one-place predicate symbol; and suppose that the identity symbol is treated as non-logical. The formula

$$\forall x \forall y \forall z(x = y \lor x = z \lor y = z)$$

is then true in any structure for L' in which Ref and LL are true. The reason is that the unique one-place predicate of L' divides the domain of a structure into those objects it satisfies and those it does not. Hence, at least two of any group of three objects will be indiscernible. On the other hand, if the identity symbol is interpreted as $I(A,x,y)$, this formula is false in any structure for L' with three or more elements.

If we do wish to view identity as a non-logical notion, then the phenomenon of language relativity suggests that it is best not to formalize identity using a single identity predicate '='. Instead, we have the following picture: We begin with a language L and define an *L-theory with identity* to be a theory whose logical axioms are those of FOL and which is such that L contains a two-place predicate E_L satisfying the non-logical axiom Ref' and the non-logical axiom schema LL$'$:

(**Ref$'$**): $E_L(x,x)$

(**LL$'$**): $E_L(x,y) \rightarrow (\varphi(x) \rightarrow \varphi(y))$.

The *pure L*-theory with identity is the L-theory whose sole non-logical axiom is Ref' and whose sole non-logical axiom schema is LL$'$.

Now the phenomenon of language relativity can be described more accurately as follows. Let L_1 be a sublanguage of L_2 and assume that T_1 and T_2 are, respectively, the pure L_1-theory with identity and the pure L_2-theory with identity. Let M_1 and M_2 be models of T_1 and T_2, respectively, having the same domain. Assume that a and b are individual constants having the same interpretation in M_1 and M_2. Let E_1 and E_2 be the identity symbols of L_1 and L_2. It can happen that $E_1(a,b)$ is true in M_1 but $E_2(a,b)$ is false in M_2. We can then say, with Geach (1967; see §4) and others, that the self-same objects indiscernible according to one theory may be discernible according to another.

There are two further philosophically significant features of the standard account of identity. First, identity is a *necessary* relation: If a and b are rigid terms (terms whose reference does not vary with respect to parameters such as time or possible world) then

(**NI**): If $a = b$ is true, then it is necessarily true.

Assuming certain modal principles, the necessity of distinctness (ND) follows from NI.

(**ND**): If $a \neq b$ is true, then it is necessarily true.

Note that the necessary truth of *a* = *b* does not imply the necessary existence of objects *a* or *b*. We may assume that what a rigid term *a* denotes at a possible world (or moment of time) *w* need not exist in *w*. Secondly, we do not ordinarily say things of the form "*x* is the same as *y*". Instead, we say "*x* and *y* are the same person" or "*x* and *y* are the same book". The standard view is that the identity component of such statements is just ‘*x* is the same as *y*’. For example, according to the standard view, ‘*x* and *y* are the same person’ reduces to ‘*x* and *y* are persons and *x* is the same as *y*’, where the second conjunct may be formalized as in $FOL^=$.

2. Paradoxes of Identity

The concept of identity, simple and settled though it may seem (as characterized by the standard account), gives rise to a great deal of philosophical perplexity. A few (by no means all) of the salient problems are outlined below. These are presented in the form of paradoxes -- arguments from apparently undeniable premises to obviously unacceptable conclusions. The aim here is to make clear just what options are available to one who would stick close to the standard account. Often (but not always) little or no defense or critique of any particular option is offered. In the next section, we shall see what the relative identity alternative offers by comparison.

2.1 The Paradox of Change

The most fundamental puzzle about identity is the problem of change. Suppose we have two photographs of a dog, Oscar. In one, *A*, Oscar is a puppy, in the other, *B*, he is old and gray muzzled. Yet we hold that he is the same dog, in, it appears, direct violation of LL. More explicitly, *B* is a photograph of an old dog with a gray muzzle; *A* is a photograph of a young dog without a gray muzzle. *A* and *B* are photographs of the same dog. But according to LL, if the dog in *B* has a property (e.g., having a gray muzzle) that the dog in *A* lacks, then *A* and *B* are *not* photographs of the same dog. Contradiction.

Various solutions have been proposed. The most popular are the following two: (1) Simple properties such as having or lacking a gray muzzle are actually relations to times. Oscar has the property of lacking a gray muzzle *at time t* and the property of having a gray muzzle at (a later) t'; but there is no incompatibility, since being thus and so related to time *t* and not being thus and so related to time t' are compatible conditions, and hence change involves no violation of LL. (2) Oscar is an object that is extended in time as well as space. The puppy Oscar and old gray muzzled Oscar are distinct temporal parts or stages of the whole temporally extended Oscar. The photograph of Oscar as a puppy is therefore not a photograph of Oscar at all. There cannot be still photographs of Oscar.

These proposals may seem plausible, and indeed most philosophers subscribe to one or other of them. The most common objections -- that on the temporal parts account, objects are not "wholly present" at any given time, and that on the relations-to-times account, seemingly simply properties of objects, such as

Oscar having a gray muzzle, are complicated relations -- do little more than affirm what their targets deny. Yet the objections are an attempt to give voice to a strong intuition concerning our experience as creatures existing in time. Both (1) and (2) treat time and change from a "God's eye" point of view. (1) presupposes time laid out "all at once", so to speak, and similarly for (2). But we experience no such thing. Instead, while we are prepared to wait to see the whole of a baseball game we are watching, we are not prepared to wait to see the whole of painting we are viewing..

2.2 Chrysippus' Paradox

The following paradox -- a variation of the paradox of change -- raises some new questions. It is due to the Stoic philosopher Chrysippus (c.280 B.C.-c.206 B.C.) and has recently been resurrected by Michael Burke (Burke, 1994). Suppose that at some point t' in the future poor Oscar loses his tail. Consider the proper part of Oscar, as he is now (at t), consisting of the whole of Oscar minus his tail. Call this object ‘ Oscar-minus ‘. Chrysippus wished to know which of these objects -- Oscar or Oscar-minus -- survives at t'. According to the standard account of identity, Oscar and Oscar-minus are distinct at t. and hence, by ND, they are distinct at t'. (Intuitively, Oscar and Oscar-minus are distinct at t' since Oscar has a property at t' that Oscar-minus lacks, namely, the property of having had a tail at t. Notice that this argument involves a tacit appeal to ND -- or NI, depending on how you look at it). Hence, if both survive, we have a case of two distinct physical objects occupying exactly the same space at the same time. Assuming that is impossible, and assuming, as commonsense demands, that Oscar survives the loss of his tail, it follows that Oscar-minus does not survive. This conclusion is paradoxical because it appears that *nothing happens* to Oscar-minus in the interval between t and t' that would cause it to perish.

One extreme option is to deny that there are such things as Oscar-minus. Undetached proper parts of objects don't exist (van Inwagen, 1981). Another is to claim that the parts of an object are essential to it (Chisholm, 1973). A third, less extreme, option is to insist that objects of *different* kinds, e.g., a clay statue and the piece of clay it is composed of, *can* occupy the same space at the same time, but objects of the same kind, e.g., two statues, cannot (Wiggins, 1968; for refinements, see Oderberg, 1996). A fourth option is to claim that Oscar and Oscar-minus are two distinct, temporally extended objects -- a dog part, Oscar-minus, and a dog, Oscar -- that overlap at t'. Temporal parts of distinct objects can occupy the same space at the same time.

2.3 The Paradox of 101 Dalmatians

(This paradox is also known as the paradox of 1001 cats; Geach, 1980, Lewis, 1993): Focus on Oscar and Oscar-minus at t -- before Oscar loses his tail. Is Oscar-minus a dog? When Oscar loses his tail the resulting creature is certainly a dog. Why then should we deny that Oscar-minus is a dog? We saw above that one possible response to Chrysippus' paradox was to claim that Oscar-minus does not exist at t'. But even if we adopt this view, how does it follow that Oscar-minus, existing as it does at t, is not a dog? Yet if Oscar-minus is a dog, then, given the standard account of identity, there are two dogs where we would normally count only one. In fact, for each of Oscar's hairs, of which there are at least 101, there is a proper part of Oscar -- Oscar minus a hair -- which is just as much a dog as Oscar-minus. There are then

at least 101 dogs (and in fact many more) where we would count only one. Some claim that things such as dogs are "maximal." No proper part of a dog is a dog (Burke, 1993). One might conclude as much simply to avoid multiplying the number of dogs populating the space reserved for Oscar alone. But the maximality principle may seem to be independently justified as well. When Oscar barks, do all these different dogs bark in unison? If a thing is a dog, shouldn't it be capable of independent action? Yet Oscar-minus cannot act independently of Oscar. Nevertheless, David Lewis (1993) has suggested a reason for counting Oscar-minus and all the 101 dog parts that differ (in various different ways) from one another and Oscar by a hair, as dogs, and in fact as Dalmatians (Oscar is a Dalmatian). Lewis invokes Unger's (1980) "problem of the many." Oscar sheds continuously but gradually. His hairs loosen and then dislodge, some such remaining still in place. Hence, within Oscar's compass at any given time there are congeries of Dalmatian parts sooner or later to become definitely Dalmatians; some in a day, some in a second, or a split second. It seems arbitrary to proclaim a Dalmatian part that is a split second away from becoming definitely a Dalmatian, a Dalmatian, while denying that one a day away is a Dalmatian. As Lewis puts it, we must either deny that the "many" are Dalmatians, or we must deny that the Dalmatians are many. Lewis endorses proposals of both types but seems to favor one of the latter type according to which the Dalmatians are not many but rather "almost one" In any case, the standard account of identity seems unable on its own to handle the paradox of 101 Dalmatians. It requires that we either deny that Oscar minus a hair is a dog -- and a Dalmatian -- or else that we must affirm that there is a multiplicity of Dalmatians, all but one of which is incapable of independent action and all of which bark in unison no more loudly than Oscar barks alone.

2.4 The Paradox of Constitution

Suppose that on day 1 Jones purchases a piece of clay c and fashions it into a statue s_1. On day 2, Jones destroys s_1, but not c, by squeezing s_1 into a ball and fashions a new statue s_2 out of c. On day 3, Jones removes a part of s_2, discards it, and replaces it using a new piece of clay, thereby destroying c and replacing it by a new piece of clay, c'. Presumably, s_2 survives this change. Now what is the relationship between the pieces of clay and the statues they "constitute?" A natural answer is: identity. On day 1, c is identical to s_1 and on day 2, c is identical to s_2. On day 3, s_2 is identical to c'. But this conclusion directly contradicts NI. If, on day 1, c is (identical to) s_1, then it follows, given NI, that on day 2, s_1 is s_2 (since c is identical to s_2 on day 2) and hence that s_1 exists on day 2, which it does not. By a similar argument, on day 3, c is c' (since s_2 is identical to both) and so c exists on day 3, which it does not. We might conclude, then, that either constitution is not identity or that NI is false. Neither conclusion is wholly welcome. Once we adopt the standard account less NI, the latter principle follows directly from the assumption that individual variables and constants in quantified modal logic are to be handled exactly as they are in first-order logic. And if constitution is not identity, and yet statues, as well as pieces of clay, are physical objects (and what else would they be?), then we are again forced to affirm that distinct physical objects may occupy (exactly) the same space at the same time. The statue s_1 and the piece of clay c occupy the same space on day 1. Even if this is deemed possible (Wiggins, 1980), it is unparsimonious. The standard account is thus *prima facie* incompatible with the natural idea that constitution is identity.

Philosophers have not argued by direct appeal to NI or ND. Typically, (e.g., Gibbard, 1975, Noonan, 1993, Johnston, 1992), arguments that *c* and s_1 are not identical run as follows: *c* exists prior to the existence of s_1 and hence the two are not identical. Again, s_1 possesses the property of being such that it will be destroyed by being squeezed into a ball, but *c* does not possess this property (*c* will be squeezed into a ball but it will not thereby be destroyed). So again the two are not identical. Further, whatever the future in fact brings, *c might* have been squeezed into a ball and not destroyed. Since that is not true of s_1, the two are not identical. On a careful analysis, however, each of these arguments can be seen to rely on NI or ND, *provided one adopts the standard account of modal/temporal predicates.* This last proviso suggests an interesting way out for one who adheres to the standard account of identity but who also holds that constitution is identity (see below).

Some philosophers find it important or at least expedient to frame the issue in terms of the case of a statue *s* and piece of clay *c* that coincide throughout their entire existence. We bring both *c* and *s* into existence by joining two other pieces of clay together, or we do something else that guarantees total coincidence. It seems that total coincidence is supposed to lend plausibility to the claim that, in such a case at least, constitution is identity (and hence NI is false -- Gibbard, 1975). It may do so, psychologically, but not logically. The same sorts of arguments against the thesis that constitution is identity apply in such a case. For example, *s* may be admired for its aesthetic traits, even long after it ceases to exist, but this need not be true of *c*. And *s* has the property, which *c* lacks, of being destroyed if squeezed into a ball. Those who defend the thesis that constitution is identity need to defend it in the general case of partial coincidence; and those who attack the thesis do so with arguments that work equal well against both total and partial coincidence. The assumption that *s* and *c* are totally coincident is therefore inessential.

The doctrine of temporal parts offers only limited help. The statement that *c* is identical to $s_1$on day 1 but identical to s_2 on day 2 can be construed to mean that *c* is a temporally extended object whose day 1 stage is identical to s_1 and whose day 2 stage is identical to s_2. Since the two stages are not identical, NI does not apply. Similarly, we can regard s_2 as a temporally extended object that overlaps *c* on day 2 and c' on day 3. But unless temporal parts theorists are prepared to defend a doctrine of modally extended objects -- objects extended through possible worlds analogous to objects extended in time, there remains a problem. s_2 *might* have been made of a different piece of clay, as is in fact the case on day 3. That is, it is logically possible for s_2 to fail to coincide with the day 2 stage of *c*. But it is not logically possible for the day 2 stage of *c* to fail to coincide with itself.

Lewis recognizes this difficulty and proposes to deal with it by appealing to his counterpart theory (Lewis, 1971, 1986, and 1993). Different concepts, e.g., *statue* and *piece of clay* are associated with different counterpart relations and hence with different criteria of trans-world identity. This has the effect of rendering modal predicates "Abelardian" (Noonan, 1991, 1993). The property determined by a modal predicate may be affected by the subject term of a sentence containing the predicate. The subject term denotes an object belonging to this or that kind or sort. But different kinds or sorts may determine different properties (or different counterpart relations). In particular, the properties determined by the predicate 'might not have coincided with c_2' (where c_2 names the day 2 stage of *c*) in the following

sentences,

(a) s_2 might not have coincided with c_2,

(b) c_2 might not have coincided with c_2,

are *different*, and hence (a) and (b) are compatible, even assuming that s_2 and c_2 are identical. (It should be emphasized that counterpart theory is not the only means of obtaining Abelardian predicates. See Noonan, 1991.)

The upshot seems to be that that the advocate of the standard account of identity must maintain either that constitution is not identity or that modal predicates are Abelardian. The latter option may be the fruitful one, since for one thing it seems to have applications that go beyond the issue of constitution.

2.5 The Ship of Theseus Paradox

(See Plutarch, *Life of Theseus*.) Imagine a wooden ship restored by replacing all its planks and beams (and other parts) by new ones. Plutarch reports that such a ship was "... a model for the philosophers with respect to the disputed arguments ... some of them saying it remained the same, some of them saying it did not remain the same" (cf. Rea, 1995). Hobbes added the catch that the old parts are reassembled to create another ship exactly like the original. Both the restored ship and the reassembled one appear to qualify equally to be the original. In the one case, the original is "remodeled", in the other, it is reassembled. Yet the two resulting ships are clearly not the same ship.

Some have proposed that in a case like this our ordinary "criteria of identity" fail us. The process of dismantling and reassembling usually preserves identity, as does the process of part replacement (otherwise no soldier could be issued just one rifle and body shops would function as manufacturers). But in this case the two processes produce conflicting results: We get two ships, one of which is the same ship as the original, by one set of criteria, and the other is the original ship by another set of criteria. There is a similar conflict of criteria in the case of personal identity: Brain duplication scenarios (Wiggins, 1967, Parfit, 1984) suggest that it is logically possible for one person to split into two competitors, each with equal claim to be the original person. We take it for granted that brain duplication will preserve the psychological properties normally relevant to reidentifying persons and we also take it for granted that the original brain continues to embody these properties even after it is duplicated. In this sense there is a conflict of criteria. Such a case of "fission" gives us two distinct embodiments of these properties.

Perhaps we should conclude that identity is not what matters. Instead, what matters is some *other* relation, but one that accounts as readily as identity for such facts as that the owner of the original ship would be entitled to both the restored version and the reassembled one. For the case of personal identity, Parfit (1984) develops such a response in detail. A related reaction would be to claim that if both competitors have equal claim to be the original, then neither *is* the original. If, however, one competitor is inferior,

then the other wins the day and counts as the original. It seems that on this view certain contingencies can establish or falsify identity claims. That conflicts with NI. Suppose that *w* is a possible world in which no ship is assembled from the discarded parts of the remodeled ship. In this world, then, the remodeled ship is the original. By NI, the restored ship and the original are identical in the actual world, contrary to the claim of the "best candidate" doctrine (which says that neither the remodeled nor the reassembled ship is the original). There are, however, more sophisticated "best candidate" theories that are not vulnerable to this objection (Nozick, 1982).

Some are convinced that the remodeled ship has the best claim to be the original, since it exhibits a greater degree of spatio-temporal continuity with the original (Wiggins, 1967). But it is unclear why the intuition that identity is preserved by spatio-temporal continuity should take precedence over the intuition that identity is preserved in the process of dismantlement and reassembly. Furthermore, certain versions of the ship of Theseus problem do not involve the feature that one of the ships competing to be the original possesses a greater degree of spatio-temporal continuity with the original than does the other (see below). Others are equally convinced that identity is *not* preserved by total part replacement. This view is often suggested blindly, as a stab in the dark, but there is in fact an interesting argument in its favor. Kripke (1980) argues that a table made out of a particular hunk of wood *could not* have been made out of a (totally) different hunk of wood. His reasoning is this: Suppose that in the actual world a table *T* is made out of a hunk of wood *H*; and suppose that there is a possible world *w* in which this very table, *T*, is made out of a different hunk of wood, H'. Then assuming that *H* and H' are completely unrelated (for example, they do not overlap), so that making a table out of the one is not somehow dependent upon making a table out of the other, there is another possible world w' in which *T*, as in the actual world, is made out of *H*, and another table T', exactly similar to *T*, is made out of H'. Since *T* and T' are not identical in w', it follows by ND that the table made out of H' in *w* is not *T*. Note, however, that the argument assumes that the table made out of H' in w' is the same table as the table made out of H' in *w*.

Kripke's reasoning can be applied to the present case (Kripke and others might dispute this claim; see below). Let *w* be a possible world just like the actual world in that *O*, the original ship, is manufactured exactly as it is in the actual world. In *w*, however, another ship, S', exactly similar to *O*, is simultaneously built out of precisely the same parts that *S*, the remodeled ship, is built out of in the actual world. Since S' and *O* are clearly different ships in *w*, it follows by ND that *O* and *S* are not the same ship in the actual world. Note again that the argument assumes that *S* and S' are the same ship, but it seems quite a stretch to deny that. Nevertheless, some have done so. Carter,1987 claims (in effect) that *S* and S' are not identical, but his argument simply *assumes* that *O* and *S* are the same ship. Alternatively, one might view the (Kripkean) argument as showing only that while *S* is the same ship as *O* in the actual world, *S* (that is, S') is not the same ship as *O* in *w*. But this is not an option for one who adheres to the standard account and hence adheres to ND. In defending this view, however, Gallois, (1986, 1988) suggests a weakened notion of rigid designation and a corresponding weakened formulation of ND. (See Carter, (1987) for criticism of Gallois' proposals. See also Chandler, 1975 for a precursor of Gallois' argument.)

If we grant that *O* and *S* cannot be the same ship, we seem to have a solution to the ship of Theseus paradox. By the Kripkean argument, only the reassembled ship has any claim to being the original ship, *O*. But this success is short lived. For we are left with the following additional paradox: Suppose that *S*

eventuates from *O* by replacing one part of *O* one day at a time. There seems to be widespread agreement that replacing just one part of a thing by a new exactly similar part preserves the identity of the thing. If so, then, by the transitivity of identity, *O* and *S* must be the same ship. It follows that either the Kripkean argument is incorrect, or replacement of even a single part (or small portion) does not preserve identity (a view known as "mereological essentialism;" Chisholm, 1973).

As indicated, Kripke denies that his argument (for the necessity of origin) applies to the case of change over time: "The question whether the table could have *changed* into ice is irrelevant here" (1972, 1980). So the question whether *O* could change into *S* is supposedly "irrelevant." But Kripke does not give a reason for this claim, and if cases of trans-temporal identity and trans-world identity differ markedly in relevant respects -- repsects relevant to Kripke's argument for the necessity of origin, it is not obvious what they are. (But see Forbes, 1985, and Lewis, 1986, for discussion.) The argument above was simply that *O* and *S* cannot be the same ship since there is a possible world in which they differ. If this argument is incorrect it is no doubt because there are conclusive reasons showing that *S* and S' differ. Even so, such reasons are clearly not "irrelevant." One may suspect that, if applied to the trans-temporal case, Kripke's reasoning will yield an argument for mereological essentialism. Indeed, a trans-world counterpart of such an argument has been tried (Chandler, 1976, though Chandler views his argument somewhat differently). In its effect, this argument does not differ essentially from the "paradox" sketched in the previous paragraph (which may well be viewed as an argument for mereological essentialism). Subsequent commentators, e.g., Salmon, (1979) and Chandler (1975, 1976), do not seem to take Kripke's admonition of irrelevance seriously.

In any case, there *is* a close connection between the two issues (the ship of Theseus problem and the question of the necessity of origin). This can be seen (though it may already be clear) by considering a modified version of the ship of Theseus problem. Suppose that when *O* is built, another ship O', exactly like *O*, is also built. Suppose that O' never sets sail, but instead is used as a kind of graphic repair manual and parts repository for *O*. Over time, planks are removed from O' and used to replace corresponding planks of *O*. The result is a ship *S* made wholly of planks from O' and standing (in the end), we may suppose, in exactly the place O' has always stood. Now do *O* and O' have equal claim to be *S*? And can we then declare that neither is *S*? Not according to the Kripkean line of thought. It looks for all the world as though the process of "remodeling" *O* is really just an elaborate means of dismantling and reassembling O'. And if O' and *S* are the same ship, then since *O* and O' are distinct, *O* and *S* cannot be the same ship.

This argument is vulnerable to the following two important criticisms: First, it conflicts with the common sense principle that (1) the material of an object can be totally replenished or replaced without affecting its identity (Salmon, 1979); and secondly, as mentioned, it conflicts with the additional common sense principle that (2) replacement by a single part or small portion preserves identity. These objections may seem to provide sufficient grounds for rejecting the Kripkean argument and perhaps restricting the application of Kripke's original argument for the necessity of origin (Noonan, 1983). There is, however, a rather striking problem with (2), and it is uncleear whether the conflict between (1) and the Kripkean argument should be resolved in favor of the former.

The problem with (2) is this. Pick a simple sort of objects, say, shoes, or better, sandals. Suppose *A* and *B* are two exactly similar sandals, one of which (*A*) is brand new and the other (*B*) is worn out. Each consists of a top strap and a sole, nothing more. If *B*'s worn strap is replaced by *A*'s new one, (2) dictates that the resulting sandal is *B* "refurbished." In fact, if the parts of *A* and *B* are simply exchanged, (2) dictates that that the sandal with the new parts, A', is *B* and the sandal with the old parts, B', is *A*. It follows by ND that *A* and A' and *B* and B' are distinct. This is surely the wrong result. The intuition that *A* and A' are the same sandal is very strong; and the process of exchanging the parts of *A* and *B* seems to amount to nothing more than the dismantling and reassembling of each. This example is no different in principle than the more elaborate trans-world cases discussed by Chisholm, (1967), Chandler, (1976), Salmon, (1979), or Gupta, (1980). (One who claims that *A* and A' differ in that A' comes into existence after *A*, does not have much to go on. *A* cannot be supposed to persist after A' comes into existence. We do not end up with two *new* sandals and one old one. Why then couldn't it be *A* itself that reappears at the later time?)

2.6 Church's Paradox

The following paradox -- perhaps the ultimate paradox of identity -- derives from an argument of Church (1982). Suppose Pierre thinks that London and Londres are different cities, but of course doesn't think that London is different from London, or that Londres is different from Londres. Assuming that proper names lack Fregean senses, we can apply LL to get the result that London and Londres *are* distinct. We have here an argument that, given the standard account of identity, merely *thinking* that *x* and *y* are distinct is enough to make them so. There are, of course, a number of ways around this conclusion without abandoning the standard account of identity. Church himself saw the argument (his version of it) as demonstrating the inadequacy of Russellian intensional logic -- in which variables and constants operate as they do in extensional logic, i.e., unequipped with senses. (For another reaction, see Salmon, 1986.) But there are strong arguments against the view that names (or variables) have senses (Kripke, 1980). In light of these arguments, Church's argument may be viewed as posing yet another paradox of identity.

The general form of Church's argument has been exploited by others to reach further puzzling conclusions. For example, it has been used to show that there can be no such thing as vague or "indeterminate" identity (Evans, 1978; and for discussion, Parsons 2000). For *x* is not vaguely identical to *x*; hence, if *x* is assumed to be vaguely identical to *y*, then by LL, *x* and *y* are (absolutely) distinct. As it stands, Evans' argument shows at best that vaguely identical objects must be absolutely distinct, not that there is no such thing as vague identity. But some have tried to amend the argument to get Evans' conclusion (Parsons, 2000; and see the entry on vagueness). In any case, it is useful to see the connection between Evans' argument and Church's. If, for example, 'vaguely identical' is taken to mean 'thought to be identical', then the two arguments collapse into one another. Church's line of argument would seem to lead ultimately to the extreme antirealist position that any perceived difference among objects is a real difference. If one resolves not to attempt to escape the clutches of LL by some clever dodge -- by disallowing straightforward quantifying-in, for example, as with the doctrine of Abelardian predicates -- one comes quickly to the absurd conclusion that no statement of the form $x = y$, where the terms are different, or are just different tokens of the same type, can be true. Yet it might just be that the fault lies

not in ourselves, but in LL.

3. Relative Identity

The fundamental claim of relative identity-the claim the various versions of the idea have in common-is that, as it seems in the passenger/person case, it can and does happen that *x* and *y* are the same *F* and (yet) *x* and *y* are not the same G. Now it is usually supposed that if *x* and *y* are the same *F* (G etc.), then that implies that *x* and *y* are *F*s (*G*s, etc.) If so, then the above schema is trivially satisfied by the case in which *x* and *y* are the same person but *x* (*y*) is not a passenger at all. But let us resolve to use the phrase '*x* and *y* are different *G*s' to mean '*x* and *y* are *G*s and *x* and *y* are not the same *G*'. Then the nontrivial core claim about relative identity is that the following may well be true:

> **(RI)**: *x* and *y* are the same *F* but *x* and *y* are different *G*s.

RI is a very interesting thesis. It seems to yield dramatically simple solutions to (at least some of) the puzzles about identity. We appear to be in a position to assert that young Oscar and old Oscar are the same dog but nonetheless distinct "temporary" objects; that Oscar and Oscar-minus are the same dog but different dog parts; that the same piece of clay can be now (identical to) one statue and now another; that London and Londres are the same city but different "objects of thought," and so forth. Doubts develop quickly, however. Either the *same dog* relation satisfies LL or it does not. If it does not, it is unclear why it should be taken to be a relation of *identity*. But if it satisfies LL, then it follows, given that Oscar and Oscar-minus are different dog parts, that Oscar-minus is not the same dog part as Oscar-minus. Furthermore, assuming that the *same dog part* relation *is* reflexive, it follows from the assumption that Oscar-minus and Oscar-minus are the same dog (and that LL is in force), that Oscar and Oscar-minus are indeed the same dog part, which in fact they are not.

It may seem, then, that RI is simply incoherent. These arguments, however, are a bit too quick. On analysis, they show only that the following three conditions form an inconsistent triad: (1) RI is true (for some fixed predicates *F* and *G*). (2) Identity relations are equivalence relations. (3) The relation *x and y are the same F* figuring in (1) satisfies LL. For suppose that the relation *x and y are the same G*, figuring in (1), is reflexive and that *x* is a *G*. Then *x* is the same *G* as *x*. But according to (1), *x* and *y* are not the same *G*s; hence, according to (3), it is not the case that *x* and *y* are the same *F*; yet (1) asserts otherwise. Now, most relative identity theorists maintain that while identity relations are equivalence relations, they do not in general satisfy LL. However, according to at least one analysis of the passenger/person case (and others), the *same person* relation satisfies LL but the *same passenger* relation is not straightforwardly an equivalence relation (Gupta, 1980). It should be clear though that this view is incompatible with the principle of the *identity of indiscernibles*: If *x* and *y* are different passengers, there must be, by the latter principle, some property *x* possesses that *y* does not. Hence if the *same person* relation satisfies LL, it follows that *x* and *y* are *not* the same person. For the remainder we will assume that identity relations are equivalence relations. Given this assumption, (and assuming that the underlying propositional logic is classical -- cf. Parsons, 2000) RI and LL are incompatible in the sense that within the framework of a single fixed language for which LL is defined, RI and LL are incompatible.

Yet the advocate of relative identity cannot simply reject any form of LL. There are true and indispensable instances of LL: If *x* and *y* are the same dog, then, surely, if *x* is a Dalmatian, so is *y*. The problem is that of formulating and motivating *restricted* forms of LL that are nonetheless strong enough to bear the burden of identity claims. There has been little systematic work done in this direction, crucial though it is to the relative identity project. (See Deutsch, 1997 for discussion of this issue.) There are, however, equivalence relations that do satisfy restricted forms of LL. These are sometimes called 'congruence relations' and they turn up frequently in mathematics. For example, say that integers *n* and *m* are congruent if their difference $n - m$ is a multiple of 3. This relation preserves multiplication and addition, but not every property. The numbers 2 and 11 are thus congruent but 2 is even and 11 is not. There are also non-mathematical congruencies. For example, the relation *x and y are traveling at the same speed* preserves certain properties and not others. If objects *x* and *y* are traveling at the same speed and *x* is traveling faster than *z*, the same is true of *y*. Such similarity relations satisfy restricted forms of LL. In fact, any equivalence relation satisfies a certain minimal form of LL (see below).

There are strong and weak versions of RI. The weak version says that RI has some (in fact, many) true instances but also that there are predicates *F* such that if *x* and *y* are the same *F*, then, for any equivalence relation, *E*, whatsoever (whether or not an identity relation), $E(x,y)$. This last condition implies that the relation *x and y are the same F* satisfies LL. The relation *P* defined so that $P(x,y)$ if and only if $H(x)$ and $H(y)$, where *H* is some predicate, is an equivalence relation. Hence, if *H* holds of *x* but not of *y*, there is an equivalence relation (namely, $P(x,y)$) that fails to hold of *x* and *y*. If we add that in this instance '*x* and *y* are the same *F*' is to be interpreted in terms of the relation $I(A,x,y)$, then the weak version of RI says that there is such a thing as relative identity and such a thing as absolute identity as well. The strong version, by contrast, says that there are (many) true instances of RI but there is no such thing as absolute identity. It is difficult to know what to make of the latter claim. Taken literally, it is false. The notion of unrestricted identity (in the sense of 'unrestricted' explained in §1) is demonstrably coherent. We return to this matter in §5 .

The puzzles about identity outlined in §2 (and there are many others, as well as many variants of these) put considerable pressure on the standard account. A theory of identity that allows for instances of RI is an attractive alternative (see below §4). But there is a certain kind of example of RI, frequently discussed in the literature, that has given relative identity something of a bad name. The passenger/person example is a case in point. The noun 'passenger' is derived from the corresponding relational expression 'passenger in (on) ...'. A passenger is someone who is a passenger in some vehicle (on some flight, etc.). Similarly, a father is man who fathers someone or who is the father of someone. This way of defining a kind of things from a relation between things is perfectly legitimate and altogether open-ended. Given any relation *R*, we can define 'an R' to apply to anything *x* that stands in *R* to something *y*. For example, we can define a 'schmapple' to be an apple in a barrel. All this is fine. But we can't *infer* from such a definition that the same apple might be two different schmapples. From the fact that someone is the father of two different children, we don't judge that he is two different fathers. The fact that airlines choose to count passengers as they do, rather than track persons, is their business, not logic's.

However, when *R* is an equivalence relation, we are entitled to such an inference. Consider the notorious

case of "surmen" (Geach, 1967). A pair of men are the "same surman" if they have the same surname; and a surman is a man who bears this relation to someone. So now it appears that that two different men can be the same surman, since two different men can have the same surname. As Geach (1967) insists (also Geach, 1973), surmen are *defined* to be men, so they are not merely classes of men. Hence we seem to have an instance of RI, and obviously any similarity relation (e.g., *x* and *y* have the same shape) will give rise to a similar case. Yet such instances of RI are not very interesting. It is granted all around that when ‘*F*’ is adjectival, different *G*s may be the same *F*. Different men may have the same surname, different objects, the same color, etc. Turning an adjectival similarity relation into a substantival one having the form of an identity statement yields an identity statement in name only.

A word about the point of view of those who subscribe to the weak version of RI. The view (call it the ‘weak view’) is that ordinary identity relations concerning (largely) the world of contingency and change are equivalence relations answering to restricted forms of LL. The exact nature of the restriction depends on the equivalence relation itself, though there is an element of generality. The kinds of properties preserved by the *same dog* relation are intuitively the same *kinds* of properties as are preserved by the *same cat* relation. From a logical point of view the best that can be said is that any identity relation, like any equivalence relation, preserves a certain minimal set of properties. For suppose *E* is some equivalence relation. Let *S* be the set containing all formulas of the form $E(x,y)$, and closed under the formation of negations, conjunctions, and quantification. Then *E* preserves any property expressed by a formula in *S*. Furthermore, on this view, although absolutely distinct objects may be the same *F*, absolutely identical objects cannot differ at all. Any instance of RI implies that
x and *y* are absolutely distinct.

4. The Paradoxes Reconsidered

Let us look back at the paradoxes of identity outlined in §2 from the perspective of the weak view regarding relative identity. That view allows that absolutely distinct objects may be the same *F*, but denies that absolutely identical objects can be different *G*’s. This implies that if *x* and *y* are relatively different objects, then *x* and *y* are absolutely distinct, and hence only pairs of absolutely distinct objects can satisfy RI. If *x* and *y* are absolutely distinct, we shall say that *x* and *y* are distinct ‘logical objects’; and similarly, if *x* and *y* are absolutely identical objects, then *x* and *y* are identical logical objects. The term ‘logical object’ does not stand for some new and special kind of thing. Absolutely distinct apples, for example, are distinct logical objects.

The following is the barest sketch of relativist solutions to the paradoxes of identity discussed in §2. No attempt is made to fully justify any proposed solution, though a modicum of justification emerges in the course of §6. It should be kept in mind that some of the strength of the relativist solutions derives from the weaknesses of the absolutist alternatives, some of which are discussed in §2.

4.1 The Paradox of Change

Young Oscar and old Oscar are the same dog but absolutely different things, i.e. different logical objects.

The material conditions rendering young Oscar and old Oscar the same dog (and the same Dalmatian) are precisely the same as the material conditions under which young Oscar and old Oscar would qualify as temporal parts of the same dog. The only difference is *logical*. The identity relation between young Oscar and old Oscar can be formalized in an extensional logic (Deutsch, 1997), but a theory of temporal parts requires a modal/temporal apparatus. Young Oscar is wholly present during his youth and possesses the simple, non-relational, property of not having a gray muzzle.

4.2 Chrysippus' Paradox

Oscar and Oscar-minus both survive Oscar's loss of a tail. At both t and t' Oscar and Oscar-minus are the same dog, but at t, Oscar and Oscar-minus are distinct logical objects. This implies (by ND) that Oscar and Oscar-minus are distinct logical objects even at t' Hence, we must allow that distinct logical objects may occupy the same space at the same time. This is not a problem, however. For although Oscar and Oscar-minus are distinct logical objects at t', they are physically coincident.

4.3 The Paradox of 101 Dalmatians

The relativist denies that dogs are "maximal." It is not true that no proper part of a dog is dog. All the 101 (and more) proper parts of Oscar differing from him and from one another by a hair are dogs. In fact, many (though of course not all) identity preserving changes Oscar might undergo correspond directly to proper parts of (an unchanged) Oscar. But there is no problem about barking in unison, and no problem about acting independently. All 101 are the same dog, despite their differences, just as young Oscar and old Oscar are the same dog, . The relativist denies that the dogs are many rather than deny that the many are dogs (Lewis, 1993).

4.4 The Paradox of Constitution

Constitution is identity, *absolute* identity. The relation between the piece of clay c and the statue s_1 on day 1 is one of absolute identity. So we have that $c = s_1$ on day 1, and for the same reason, $c = s_2$ on day 2. Furthermore, since s_1 and s_2 are different statues, it follows (on the weak view) that $s_1 \neq s_2$. In addition, the piece of clay c constituting s_1 on day 1 is (relatively) the same piece of clay as the piece of clay constituting s_2 on day 2. (The identity is relative because we have distinct objects -- the two statues -- that are the same piece of clay.) It follows that *no name of the piece of clay c can be a rigid designator in the standard sense*. That is, no name of c denotes absolutely the same thing on day 1 and on day 2. For on day 1, a name of the piece of clay c would denote s_1 and on day 2, it would denote s_2, and s_1 and s_2 are absolutely distinct. Nevertheless, a name of the piece of clay may be *relatively rigid*: it may denote at each time the *same piece of clay*. Although no name of the piece of clay c is absolutely rigid, that does not prevent the introduction of a name of c that denotes c at any time (or possible world). (Kraut, 1980 discusses a related notion of relative rigidity.)

There is, however, a certain ambiguity in the notion of a name of the piece of clay, inasmuch as the piece of clay may be any number of absolutely distinct objects. The notion of relative rigidity presupposes that a name for the piece of clay refers, with respect to some parameter *p*, to whatever object counts as the piece of clay relative to that parameter. This may be sufficient in the case of the piece of clay, but in other cases it is not. With respect to a fixed parameter *p* there may be no unique object to serve as the referent of the name. For example, if any number of dog parts count, at a fixed time, as the same dog, then which of these objects serves as the referent of 'Oscar'? We shall leave this question open for the time being but suggest that it may be worthwhile to view names such as 'Oscar' as *instantial* terms -- terms introduced into discourse by means of existential instantiation. The name 'Oscar' might be taken as denoting a representative member of the equivalence class of distinct objects qualifiying as the same dog as Oscar. It would follow, then, that most ordinary names are instantial terms. (An alternative is that of Geach, 1980, who draws a distinction between a *name of* and a *name for an* object; see Noonan, 1997 for discussion of Geach's distinction.)

4.5 The Ship of Theseus Paradox

In this case, the relativist, as so far understood, may seem to enjoy no advantage over the absolutist. The problem is not clearly one of reconciling LL with ordinary judgments of identity, and the advantage afforded by RI does not seem applicable. Griffin (1977), for example, relying on RI, claims that the original and remodeled ship are the same ship but not the same collection of planks, whereas the reassembled ship is the same collection of planks as the original but not the same ship. This simply doesn't resolve the problem. The problem is that the reassembled and remodeled ships have, *prima facie*, equal claim to be the original and so the bald claims that the reassembled ship is not--and the remodeled ship is--the original are unsupported. The problem is that of reconciling the intuition that certain small changes (replacement of a single part or small portion) preserve identity, with the problem illustrated by the sandals example of §2.5. It turns out, nevertheless, that the problem *is* one of dealing with the excesses of LL. To resolve the problem, we need an additional level of relativity. To motivate this development, consider the following abstract counterpart of the sandals example:

On the left there is an object P composed of three parts, P_1, P_2, and P_3. On the right is an exactly similar but non-identical object, Q, composed of exactly similar parts, Q_1, Q_2, and Q_3, in exactly the same arrangement. For the sake of illustration, we adopt the rule that only replacement of (at most) a *single* part by an exactly similar part preserves identity. Suppose we now interchange the parts of P and Q. We begin by replacing P_1 by Q_1 in P and replacing Q_1 by P_1 in Q, to obtain objects P^1 and Q^1. So P^1 is composed of parts Q_1, P_2, and P_3, and Q^1 is composed of parts P_1, Q_2, and Q_3. We then replace P_2 in P^1 by Q_2, to obtain P^2, and so on. Given our sample criterion of identity, and assuming the transitivity of identity, P and P^3 are counted the same, as are Q and Q^3. But this appears to be entirely the wrong result. Intuitively, P and Q^3 are the same, as are Q and P^3. For P and Q^3 are composed of exactly the same parts put together in exactly the same way, and similarly for Q and P^3. Futhermore, Q_3 (P_3) can be viewed as simply the result of taking P (Q) apart and putting *it* back together in a slightly different location. And this last difference can be eliminated by switching the locations of P^3 and Q^3 as a last step in the process.

Suppose, however, that we replace our criterion of identity by the following more complicated rule: *x* and *y* are the same *relative to z*, if both *x* and *y* differ from *z* at most by a single part. (This relation is transitive, and is in fact an equivalence relation.) For example, *relative to P*, *P*, P^1, Q^2, and Q^3 are the same, but *Q*, Q^1, P^2 and P^3, are not. Of course, replacement by a single part is an artificial criterion of identity. In actual cases, it will be a matter of the degree or kind of deviation *from the original* (represented by the third parameter, *z*). The basic idea is that identity through change is not a matter of identity through successive, accumulated changes -- that notion conflicts with both intuition (e.g., the sandals example) and the Kripkean argument: Through successive changes objects can evolve into *other* objects. The three-place relation of idenitity does not satisfy LL and is consistent with the outlook of the relativist. Gupta, (1980) develops a somewhat similar idea in detail. Williamson (1990) suggests a rather different approach, but one that, like the above, treats identity through change as an equivalence relation that does not satisfy LL.

4.6 Church’s Paradox

Church’s argument implies that if Pierre’s doxastic position is as described (in §2.6), then London and Londres are distinct objects. Assuming the standard account of identity, the result is that either Pierre’s doxastic position *cannot* be as described or else London and Londres are different *cities* (or else we must punt). Since London and Londres are not different cities, the standard account entails that Pierre’s doxastic position cannot be as described (or else we must punt). This was Church’s own position as regards certain puzzles about synonymy, such as Mate’s puzzle (Mates, 1952). Church held that one who believes that lawyers are lawyers, must indeed believe that lawyers are attorneys, despite any refusal to assent to (or desire to dissent from) ‘Lawyers are attorneys’ (Church, 1954). Kripke later argued (Kripke, 1979) that assent and failure to assent must be taken at face value (at least in the case of Pierre) and Pierre’s doxastic position is as described. Kripke chose to punt -- concluding that the problem is a problem for any "logic" of belief. The relativist concludes instead that (a) Pierre’s doxastic position is as described, (b) if so, London and Londres are distinct objects, and (c) London and Londres are nonetheless the same city. Whether this resolution of Church’s paradox can be exploited to yield solutions to Frege’s puzzle (Salmon, 1986) or Kripke’s puzzle (1979) remains to be seen. Crimmins (1998) has recently suggested that the analysis of propositional attitudes requires a notion of "semantic pretense." In reporting Pierre’s doxastic position we engage in a pretense to the effect that London and Londres are different cities associated with different Fregean senses. Crimmins’ goal is to reconcile (a), (c) and the following, (d): that the pure semantics of proper names (’London’, ‘Londres’) is Millian or directly referential (Kripke, 1979). The relativist proposes just such a reconciliation but suggests that the pretense can be dropped.

5. Absolute Identity

The philosopher P.T. Geach first broached the subject of relative identity and introduced the phrase ‘relative identity’. Over the years, Geach has suggested specific instances of RI (a variant of the case of Oscar and his tail is due to Geach (Geach, 1980)) and in this way he has contributed to the development

of the weak view concerning relative identity, i.e. the view that while ordinary identity relations are often relative, some are not. But Geach maintains that absolute identity does not exist. What is his argument?

That is hard to say. Geach sets up two strawman candidates for absolute identity, one at the beginning of his discussion and one at the end, and he easily disposes of both. In between he develops an interesting and influential argument to the effect that identity, *even as formalized in the system* $FOL^=$, is relative identity. However, Geach takes himself to have shown, by this argument, that absolute identity does not exist. At the end of his initial presentation of the argument in his 1967 paper, Geach remarks:

> We thought we had a criterion for a predicable's expressing strict identity [i.e., as Geach says, "strict, absolute, unqualified identity"] but the thing has come apart in our hands; and no alternative rigorous criterion that could replace this one has thus far been suggested. (Geach, 1972, p. 241)

It turns out, as we'll see, that all that comes apart is the false notion that in $FOL^=$ the identity symbol *defines* the relation $I(A,x,y)$. Let us examine Geach's line of reasoning in detail, focusing on the presentation in his 1967 article, the *locus classicus* of the notion of relative identity.

Geach begins by urging that a plain identity statement '*x* and *y* are the same' is in need of a completing predicate: '*x* and *y* are the same *F*'. Frege had argued that statements of number such as 'this is one' require a completing predicate: 'this is one *F*', and so it is, Geach claims, with identity statements. This is a natural view for one who subscribes to RI. The latter cannot even be stated without the completing predicates. Nevertheless, both the claim itself and the analogy with Frege have been questioned. Some argue that the analogy with Frege is incorrect, others that while the analogy is correct, both Frege and Geach are wrong (Perry, 1978 and Bennett and Alston, 1984). These matters will be discussed in greater detail in an updated version of this article. They do not bear directly on the question of the coherence and truth of RI or the question of absolute identity. One who adopts the weak view would not want to follow Geach on this score. And one could maintain the "completing thesis" without being committed to RI. Furthermore, the completing thesis occupies a puzzling role in Geach's dialectic. Immediately following his statement of the thesis, Geach formalizes $FOL^=$ on the basis of the single formula:

> **(W)**: $\wp(a) \leftrightarrow \exists x(\wp(x) \wedge x = a)$

(The 'W' is for Hao Wang, who first suggested it. The reader is invited to prove Ref and LL from W.) But we hear no complaint about the syntax of W despite its involving a seemingly unrelativized identity symbol. It turns out, however, that Geach apparently thinks of the completing predicate as being given by the whole descriptive apparatus of *L* or a fragment thereof.

Geach now observes

> ... if we consider a moment, we see that an *I*-predicable in given theory *T need not* express strict, absolute, unqualified identity; it need mean no more than that two objects are

> indiscernible by the predicables that form the descriptive resources of the theory -- the ideology of the theory ... . (p. 240)

Here an ‘I-predicable’ is a binary relation symbol ‘=’ satisfying (W). Geach’s focus at this point is on the need to relativize an *I*-predicable to a theory *T*. Geach then immediately saddles the friend of absolute identity with the view that for "real identity" we need not bring in the ideology of a definite theory. This is Geach’s first strawman. When logicians, in discussing FOL$^=$, speak of "real identity" -- and they often do (see Enderton 2000 or Silver 1994, for example) -- they do not mean a relation of *universal identity*, since the universal set does not exist. Nor do they intend, in formulating LL, to use ‘true of’ in a completely unrestrained way which gives rise to semantic paradox. It is no argument against those who wish to distinguish mere indiscernibility from real identity to say that they "will soon fall into contradictions," e.g., Grelling’s or Russe LL's. The relation *I*(*A*,*x*,*y*) is sufficiently relativized. (It is relativized to a *set A*.)

We come next to the main point:

> Objects that are indiscernible when we are confined to the ideology of T may perfectly well be discernible in ideology of a theory T^1 of which T is a fragment. (p. 240)

The warrant for this claim can only be the language relativity of identity when treated as a non-logical notion (see §1). That this is what Geach has in mind is clear from some approving remarks he makes in his 1973 article about Quine’s (1970) proposal to treat identity as a non-logical notion. But how does it follow that absolute identity does not exist? Geach seems to think that the defender of absolute identity will look to Ref and LL (or W) -- and not beyond -- for a full account of "strict, absolute, unqualified" identity. That is not so. The fact that these formulas in themselves define only indiscernibility relations is a logical commonplace. So this is Geach’s second strawman.

Is Geach’s argument at least an argument that identity is relative? Does language relativity support the conclusion that RI is true even of identity as formalized in FOL$^=$? The general idea appears to be that language relativity suggests that we take identity *to be* indiscernibility, and conclude that objects identical relative to one ideology *F* may be different relative to another ideology *G*, and that this confirms RI. Notice first of all that this argument relies on the identity of indiscernibles: that indiscernibility implies identity. This principle is not valid in FOL$^=$ even when the latter is treated as a proper theory. Language relativity does not imply that the distinctness of distinct objects cannot go unnoticed.

Secondly, the interesting cases of RI do not involve a shift from an impoverished point of view to an improved one -- whether this is seen in epistemic terms (which Geach disputes -- Geach, (1973)) or in purely logical terms. We do not affirm that old Oscar and young Oscar, for example, are the same dog on the grounds that there is an ideology with respect to which old Oscar and young Oscar are indistinguishable. Such an ideology would be incapable of describing any change in Oscar. It is true that the *same dog* relation determines a set of predicates that do not discriminate between the members of certain pairs of dogs -- the dogs in the photographs mentioned in earlier, for example. And it is true that

these predicates determine a sublanguage in which the *same dog* relation is a congruence, i.e. no predicate of the sublanguage distinguishes *x* from *y*, if *x* and *y* are the same dog. But the very *sense* of such statements as that old Oscar and young Oscar are the same dog requires a language in which a change in Oscar is expressible. We are talking, after all, about *old* Oscar and *young* Oscar. If we take seriously the idea that change involves the application of incompatible predicates, then the sublanguage cannot express the contrast between old Oscar and young Oscar.

Third, the phenomenon of language relativity (in the technical sense discussed in §1) has led many philosophers, including Geach, to the view that ideology creates ontology. There is no antecedently given domain of objects, already individuated, and waiting to be described. Instead, theories carve up the world in various ways, rendering some things noticeably distinct and others indiscernible, depending on a theories' descriptive resources. The very notion of *object* is theory-bound (Kraut, 1980). This sort of anti-realism may seem to go hand in hand with relative identity. Model theory, however, is realist to its core and language relativity is a model-theoretic phenomenon. It is a matter of definability (in a structure). Referring back to §1, in order to make sense of language relativity we have to start with a pair of *distinct* objects, *a* and *b*, (distinct from the standpoint of the metalanguage), and hence a pair of objects we assume are already individuated. These objects, however, are indistinguishable in *M*, since no formula of L' defines a subset of *M* containing the one object and not the other. When we move to M', we find that there is a formula of the enriched language that defines such a subset in M'. Thus, language relativity is not really any sort relativity of *identity* at all. We must assume that the objects *a* and *b* are distinct in order to *describe* the phenomenon. If we are living in *M*, and suspect that Martians living in M' can distinguish *a* from *b*, our suspicion is not merely to the effect that Martians carve things up differently than we do. Our own model theory tells us that there is more to it than that. Our suspicion must be to the effect that *a* and *b are* absolutely distinct. If we are blind to the difference between *a* and *b*, but the Martians are not, then there must be a difference; and even if we are living in *M*, we know there's a difference, or at least we can suspect there is, since model theory tells us that such suspicion is well founded.

Let us go back to Geach's remark that we "*need not*" interpret identity absolutely. While this is true, we need not interpret it as indiscernibility either. There are always the quotient structures (Quine, 1963) . Instead of taking our "reality" to be *M*, and our "identity" to be indiscernibility in *M*, we can move to the quotient structure, *QM*, whose elements are the equivalence classes, [*x*], for *x* in *M*. If *x* and *y* are indiscernible in *M*, then in *QM*, [*x*] and [*y*] are absolutely identical. We can do this even if we wish to treat $FOL^=$ as a proper theory. For example, suppose L' is a language in which people having the same income are indiscernible. The domain of *M* now consists of people. *QM*, however, consists of income groups, equivalence classes of people having the same income, and identity in *QM* is absolute. Geach objects to such reinterpretation in terms of the quotient structures on the grounds that it increases the ontological commitments not only of L' but of any language of which L' is a sublanguage.

Let's focus on L' first. From a purely model-theoretic point of view the question is moot. We cannot deny that *QM* is a structure for L'. Thus, L' is committed to people *vis à vis* one structure and to income groups *vis à vis* the corresponding quotient structure. But let's pretend that the structures are "representations of reality," and so the question now becomes: Which representation is preferable? Is there then any reason to prefer the ontology of *M* to that of *QM*? *M* contains people but no sets of people, whereas *QM* contains

sets of people but no people. By Quine's criterion of ontological commitment -- that to be is to be the value of a variable -- commitment to a set of objects does not carry a commitment to its elements. That is one of the odd consequences of Quine's criterion. Unless there is some ontological reason to prefer people to sets of people (perhaps because sets are never to be preferred), the ontologies of *M* and *QM* seem pretty much on a par. Both commit L' to one kind of thing.

Geach makes the additional claim that the ontological commitments of a sublanguage L' of a language *L* are inherited by *L* (Geach, 1973). Suppose then that *L* is a language containing expressions for several equivalence relations defined on people: say, *same income*, *same surname*, and *same job*. Geach argues that *L* need only be committed to the existence of people. Things such as income groups, job groups (equivalence classes of people with the same job), and surmen can all be counted using the equivalencies, without bringing surmen, job groups, and income groups into the picture. Consider any sublanguage for which any one of these equivalence relations is a congruence, i.e. for which LL$'$ holds. Pick the language, L_1, for example, in which people having the same job are indiscernible. More precisely, we assume that T_1 is the pure theory with identity whose ideology is confined to the language L_1. Let M_1 be a model of T_1. We may imagine the domain of M_1 to consist of people, and we can interpret indiscernibility in M_1 to be the relation *x and y have the same job*. Geach would argue that if L_1 is committed to the elements of QM_1 -- the job groups -- then so is *L*. But that is not true. If *T* is a theory of the three distinct equivalence relations formulated in *L*, the most *T* (or *L*) would be committed to are the partitions determined by the equivalence relations; and in any case, it would be perfectly consistent to insist that, whatever the ontological commitments of L_1, reality, as described by *L*, consists of people.

The foregoing considerations are rather abstract. To see more clearly what is at stake, let us focus on a specific example. Geach (1967) mentions that rational numbers are defined set-theoretically to be equivalence classes of integers determined by a certain equivalence relation defined on "fractions," i.e. ordered pairs of integers (1/2 is <1,2>, 2/4 is <2,4>, etc.). He suggests that we can instead construe our theory of rational numbers to be about the fractions themselves, taking the *I*-predicable of our theory to be the following equivalence relation, *E*:

(R): $E(\langle x,y\rangle, \langle u,v\rangle)$ iff $xv = yu$.

This approach, Geach says, would have "the advantage of lightening a theory's set-theoretical burdens. (In our present example, we need not bring in *infinite sets* of ordered pairs of integers into the theory of rationals.)."

The first thing to notice about this example is that *E cannot* be the I-predicable of such a theory, since *E* is *defined* in terms of identity (look at the right side of R). It is '=' that must serve as the *I*-predicable, and it renders distinct ordered pairs of integers discernible. The moral is that not all equivalence relations can be drafted to do the job of identity, even given a limited ideology. There is, indeed, a plausible argument that *any* equivalence relation presupposes identity -- not necessarily in the direct way illustrated by (R), but indirectly, nonetheless (see §6). Moreover, from the standpoint of general mathematics, once we have (R), we have the (infinite) equivalence classes it determines and the partition it induces. These are

inescapable. Even from a more limited viewpoint, it seems that once we have enough set theory to give us ordered pairs of integers and the ability to define (R), we get the partition it induces as well.

Geach perceives an ontological advantage in relative identity; but his argument is unconvincing. Shifting to the quotient structures, as Quine suggested, does not induce a "baroque, Meinongian ontology" (Geach, 1967). In particular, the "home language" (L) does not inherit the commitment of the fragment (L_1), and the ontology of an arbitrary model of the pure theory of identity based on the latter language is at least no more various than that of the corresponding quotient model. There are, however, a number of ways in which relative identity does succeed in avoiding commitment to certain entities required by its absolute rival. These are discussed in the replies to objections 4 and 5 in the next section.

6. Objections and Replies

The following constitute a "start up" set of objections and replies concerning relative identity and/or aspects of the foregoing account of relative identity and its rival. Time and space constraints prevent a more extended initial discussion. In addition, there is no presumption that the objections discussed below are the most important or that the initial replies to them are without fault. It is hoped that the present discussion will evolve into a more full blown one, involving contributions by the author and readers alike. Should the discussion become lengthy, old or unchallenged objections and/or replies can be placed in the archives.

Objection 1: "Relativist theories of identity, all of which are inconsistent with Leibniz's principle [LL], currently enjoy little support. The doubts about them are (a) whether they really are theories of numerical identity, (b) whether they can be made internally consistent, and (c) whether they are sufficiently motivated." (Burke, 1994.)

Reply: In reverse order: (c) The issues discussed in §2 and §4 surely provide sufficient motivation. (b) No proof of inconsistency has ever been forthcoming from opponents of relative identity, and in fact the weak view is consistent inasmuch as it has a model in the theory of similarity relations. The arguments outlined in the second paragraph of §3 are frequently cited as showing that relative identity is incoherent; but they show only that RI is incompatible with (unrestricted) LL. (a) See the replies to objections 2 and 3 below.

Objection 2: If an identity relation obeys only a restricted form of LL -- if it preserves only *some* properties and not *all* -- then how do we *tell* which properties serve to individuate a pair of distinct objects?

Reply: Similarity relations satisfy only restricted forms of LL. How then do we *tell* which properties are preserved by the *same shape* relation and which are not? It is no objection to the thesis that identity relations in general preserve some properties and not others to demand to know which are which. At best the objection points to a problem we must face anyway (for the case of similarity). In general, a property is preserved by an equivalence relation if it "spreads" in an equivalence class determined by the relation:

If one member of the class has the property, then every member does. Every property spreads in a singleton, as absolute identity demands.

Objection 3: If identity statements are mere equivalencies, what distinguishes identity from mere similarity?

Reply: The distinction between identity and similarity *statements* (or sentences) is usually drawn in terms of the distinction between substantival and adjectival common nouns. If *F* is a common noun standing for a kind of *things* e.g., 'horse', then '*x* and *y* are the same *F*' is a statement of identity, whereas if *F* is an a common noun standing for a kind of *properties* of things, then '*x* and *y* are the same *F*' is a statement of similarity. (It's interesting to note that when the noun is proper, i.e. a proper name, the result is a statement of similarity, not identity -- as in 'He's not the same Bill we knew before'.) This distinction rests ultimately on the metaphysical distinction between substance and attribute, object and property. While the distinction no doubt presupposes the concept of individuation (the bundle theory, for example, presupposes that we have the means to individuate properties), there is no obvious reason to suppose that it entails the denial of RI, i.e. the claim that no instance of RI is true. For a beginner's review -- from an historical perspective -- of the issues concerning substance and attribute, see O'Connor, (1967); and for more recent and advanced discussion and bibliography, see the entry on properties.

Objection 4: Consider the following alleged instance of RI:

1. *A* is the same word type as *B*, but *A* and *B* are different word tokens.

"If '*A*' and '*B*' refer to the same objects throughout (1), the first conjunct of (1) is not an identity statement, and the counterexample (to the thesis that no instance of RI is true) fails. If both conjuncts are identity statements in the required sense, '*A*' and '*B*' must refer to word types in the first conjunct and word tokens in the second, and the counterexample fails" (Perry, 1970).

Reply: First, if "in the required sense" means "satisfies LL," then the objection buys correctness only at the price of begging the question. Advocates of relative identity will maintain that the relation *A is the same word type as B* is an identity relation, defined on tokens, that does not satisfy LL.

Secondly, even if one insists that in this case intuition dictates that if *A* and *B* refer to tokens in both conjuncts of (1), then '*A* is the same word type as *B*' expresses only the similarity relation: *A and B are tokens of the same type*, there are other cases where, intuitively, both conjuncts of RI involve identity relations and yet the relevant terms all refer to the same kind of things; for example,

2. *A* and *B* are the same dog but *A* and *B* are different physical objects,

as said of young Oscar and old Oscar. Here there is no temptation to suppose that the relation *A and B are the same dog* is *not* an identity relation. One may invoke a theory -- a theory of temporal parts, for example -- that construes the relation as a certain kind of similarity, but that is theory, not pretheoretical

intuition. It is no objection to the relativist's *theory*, which holds in part that '*A* and *B* are the same dog' expresses a relation of primitive identity, that there is an alternative theory according to which it expresses a similarity relation obtaining between two temporal parts of the same object. Furthermore, in the case of (2), *A* and *B* refer, again intuitively, to the same things in both conjuncts.

Third, there are cases in which the relative identity view does possess an ontological advantage. Consider

3. *A* and *B* are the same piece of clay but *A* and *B* are different statues.

Suppose *A* and *B* are understood to refer to one sort of thing -- pieces of clay -- in the first conjunct and another -- statues -- in the second conjunct. Assume that the piece of clay *c* denoted by *A* in the first conjunct constitutes, at time *t*, the statue *s*. Then assuming that statues are physical objects, there are two distinct physical objects belonging to different kinds occupying the same space at *t*. Some, notably Wiggins (1980), hold that this is entirely possible: Distinct physical objects may occupy the same space at the same time, provided they belong to different kinds. The temporal parts doctrine supports and encourages this view. A statue may be a temporal part of a temporally extended piece of clay. But one statue, it seems, cannot be a temporal part of another. Even so, however, the duality of constituter and thing constituted is unparsimonious (cf. Lewis, 1993), and the relativist is not committed to it.

Again, consider

4. *A* and *B* are the same book but *A* and *B* are different copies (of the book).

One can say that in the first conjunct, *A* and *B* refer to books (absolutely the same book), whereas in the second conjunct, *A* and *B* refer to (absolutely distinct) copies. But the alleged duality of books and copies of books is unparsimonious and the relativist is not committed to it. There is no reason to concede to the philosopher that we do not actually purchase or read *books*; instead we purchase and read only *copies* of books. Any copy of a book is just as much the "book itself" as is any other copy. Any copy of a book is *the same book* as any other copy. Nelson Goodman once remarked that "Any accurate copy of a poem is as much the original work as any other" (Goodman, 1968). Goodman was not suggesting that the distinction between poem and copy collapses. If it does collapse, however, we have an explanation of why any accurate copy is as much the original work as any other: any such copy is the same work as any other.

Objection 5: Geach remarks that "As for our recognizing relative identity predicables: any equivalence relation...can be used to specify a criterion of relative identity." But §3 above contains a counterexample. Some equivalence relations are defined in terms of the I-predicable of a theory and hence cannot serve as such. (Any pair of I-predicables for a fixed theory are equivalent.) In fact it seems that any equivalence relation presupposes identity (cf. McGinn, 2000). For example, the relation *x and y are the same color* presupposes identity of colors, since it means that there are colors C and C' such that x has C and y has C', and $C = C'$. Identity, therefore, is logically prior to equivalence.

Reply: This is a good objection. It does seem to show, as the objector says, that identity is logically prior to ordinary similarity relations. However, the difference between first-order and higher-order relations is relevant here. Traditionally, similarity relations such as *x and y are the same color* have been represented, in the way indicated in the objection, as higher-order relations involving identities between higher order objects (properties). Yet this treatment may not be inevitable. In Deutsch (1997), an attempt is made to treat similarity relations of the form '*x* and *y* are the same *F*' (where *F* is adjectival) as primitive, first-order, purely logical relations (see also Williamson, 1988). If successful, a first-order treatment of similarity would show that the impression that identity is prior to equivalence is merely a misimpression -- due to the assumption that the usual higher-order account of similarity relations is the only option.

Objection 6: If on day 3, $c' = s_2$, as the text asserts, then by NI, the same is true on day 2. But the text also asserts that on day 2, $c = s_2$; yet $c \neq c'$. This is incoherent.

Reply: The term s_2 is not an absolutely rigid designator and so NI does not apply.

Objection 7: The notion of relative identity is incoherent: "If a cat and one of its proper parts are one and the same cat, what is the mass of that one cat?" (Burke, 1994)

Reply: Young Oscar and Old Oscar are the same dog, but it makes no sense to ask: "What is the mass of that one dog." Given the possibility of change, identical objects may differ in mass. On the relative identity account, that means that distinct logical objects that are the same *F* may differ in mass -- and may differ with respect to a host of other properties as well. Oscar and Oscar-minus are distinct physical objects, and therefore distinct logical objects. Distinct physical objects may differ in mass.

Objection 8: We can solve the paradox of 101 Dalmatians by appeal to a notion of "almost identity" (Lewis, 1993). We can admit, in light of the "problem of the many" (Unger, 1980), that the 101 dog parts are dogs, but we can also affirm that the 101 dogs are not many; for they are "almost one." Almost-identity is not a relation of indiscernibility, since it is not transitive, and so it differs from relative identity. It is a matter of negligible difference. A series of negligible differences can add up to one that is not negligible.

Reply: The difference between Oscar and Oscar-minus is not negligible and the two are not almost-identical. Lewis concedes this point but proposes to combine almost-identity with supervaluations to give a mixed solution to the paradox. The supervaluation solution starts from the assumption that one and only one of the dog parts is a dog (and a Dalmatian, and Oscar), but it doesn't matter which. It doesn't matter which because we haven't decided as much, and we aren't going to. Since it is true that any such decision renders one and only one dog part a dog, it is plain-true, i.e. supertrue, that there is one and only one dog in the picture. But it is not clear that this approach enjoys any advantage over that of relative identity; in fact, it seems to produce instances of RI. Compare: Fred's bicycle has a basket attached to it. Ordinarily, our discourse slides over the difference between Fred's bicycle with its basket attached and Fred's bicycle minus the basket. (In this respect, the case of Fred's bicycle differs somewhat from that of Oscar and Oscar-minus. We tend not to ignore *that* difference.) In particular, we don't say that Fred has two bicycles

even if we allow that Fred's bicycle-minus is a bicycle. Both relative identity and supervaluations validate this intuition. However, both relative identity and supervaluations also affirm that Fred's bicycle and Fred's bicycle-minus are absolutely distinct objects. That is, the statement that Fred's bicycle and Fred's bicycle-minus are distinct is supertrue. So the supervaluation technique affirms both that Fred's bicycle and Fred's bicycle-minus are distinct objects *and* that there is one and only one (relevant) bicycle. That is RI, or close enough. The supervaluation approach is not so much an alternative to relative identity as a form of it.

Bibliography

- Baker, L. R., 1997: "Why Constitution is not Identity," *Journal of Philosophy*, 94, pp. 599-621.
- Bennett, J., and Alston, W., 1984: "Identity and Cardinality: Geach an Frege," *Philosophical Review* 93, pp. 553-568.
- Burke, M., 1992: "Copper Statues and Pieces of Copper: A Challenge to the Standard Account," *Analysis*, 52, pp. 12-17.
- -----, 1995: "Dion and Theon: An Essentialist Solution to an Ancient Problem," *The Journal of Philosophy*, 91, pp.129-139.
- Carter, W. R., 1982: "On Contingent Identity and Temporal Worms," *Philosophical Studies*, 41, pp. 213-230.
- -----, 1987: "Contingent Identity and Rigid Designation," *Mind*, 96, pp. 250-255.
- -----, 1990: *Elements of Metaphysics.* New York: McGraw-Hill.
- Cartwright, R., 1987: "On the Logical Problem of the Trinity." In Cartwright, R., *Philosophical Essays.* Cambridge, Mass.: MIT Press.
- Chandler, H., 1971: "constitutivity and Identity," *Noûs*, 5, pp. 513-519.
- -----, 1975: "Rigid Designation," *The Journal of Philosophy,* 72, pp. 363-369.
- -----, 1976: "Plantinga on the Contingently Possible," *Analysis*, 36, pp. 106-109.
- Chisholm, R. M., 1967: "Identity through Possible Worlds: Some Questions," *Noûs*, 1, pp. 1-8
- -----, 1969: "The Loose and Popular and Strict and Philosophical Senses of Identity." In Grim, R. H. and Case, R. (eds.), *Perception and Personal Identity*. Cleveland: Ohio University Press.
- -----, 1973: "Parts as Essential to their Wholes," *Review of Metaphysics*, 26, pp. 581-603.
- Church, A., 1954: "Intensional Isomorphism and Identity of Belief," *Philosophical Studies*, 5, pp. 65-73; reprinted in N. Salmon and S. Soames, 1988.
- -----, 1982: "A Remark Concerning Quine's Paradox About Modality," *Analisis Filosophico* 2, (Spanish version). Reprinted in N. Salmon and S. Soames, 1988.
- Crimmins, M., 1998: "Hesperus and Phosphorus: Sense, Pretense, and Reference", *Philosophical Review*, 107, pp. 1-48.
- Deutsch, H., 1997: "Identity and General Similarity," *Philosophical Perspectives* 12, pp.177-200.
- Dummett, M., 1991: "Does Quantification involve Identity?", in H.A. Lewis (ed.), *Peter Geach*: *Philosophical Encounters*. Dordrecht: Kluwer Academic Publishers.
- Eglueta, R. and Jansana, R., 1999: "Definability of Leibniz Equality," *Studia Logica* 63, pp. 223-243.
- Enderton, H., 2000: *A Mathematical Introduction to Logic*. New York: Academic Press.

- Epstein, R., 2001: *Predicate Logic*. Belmont, CA: Wadsworth.
- Evans, G., 1978: "Can There be Vague Objects?", *Analysis*, 38, p. 308.
- Forbes, G., 1985: *The Metaphysics of Modality*. Oxford: Oxford University Press.
- -----, 1994: "A New Riddle of Existence," *Philosophical Perspectives*, 8, pp. 415-430.
- Gallois, A., 1986: "Rigid Designation and the Contingency of Identity," *Mind*, 95, pp. 57-76.
- -----, 1988: "Carter on Contingent Identity and Rigid Designation," *Mind*, 97, pp. 273-278.
- -----, 1990: "Occasional Identity," *Philosophical Studies*, 58, pp. 203-224.
- -----, 1998: *Occasions of Identity*: *A Study in the Metaphysics of Persistence, Change and Sameness*. Oxford: Clarendon Press.
- Geach, P.T., 1967: "Identity," *Review of Metaphysics*, 21, pp.3-12. Reprinted in Geach 1972, pp. 238-247.
- -----, 1972: *Logic Matters*. Oxford: Blackwell.
- -----, 1973: "Ontological Relativity and Relative Identity." In Munitz, M. (ed), *Logic and Ontology*. New York: New York University Press.
- -----, 1980: *Reference and Generality* (third edition). Ithaca: Cornell University Press.
- Gibbard, A., 1975: "Contingent Identity," *Journal of Philosophical Logic* 4, pp.187-221.
- Goodman, N., 1968: *Languages of Art*. Indianapolis and New York: Bobbs-Merrill Company.
- Griffin, N., 1977: *Relative Identity*. New York: Oxford University Press.
- Gupta, A., 1980: *The Logic of Common Nouns*. New Haven and London: Yale University Press.
- Heller, M., 1990: *The Ontology of Physical Objects*. New York: Cambridge University Press.
- Hinchliff, M., 1996: "The Puzzle of Change," *Philosophical Perspectives*,10, pp. 119-133.
- Hodges, W., 1983: "Elementary Predicate Logic," in D. Gabbay and F. Guenthner, (eds), *Handbook of Philosophical Logic*, v.1, Dordrecht: Reidel.
- Johnston, M., 1992: "Constitution is not Identity," *Mind*, 101, pp. 89-105.
- Koslicki, K., 2000: "Constitution and Supervenience," unpublished ms.
- Kraut, R., 1980: "Indiscernibility and Ontology," *Synthese* 44, pp.113-135.
- Kripke, S., 1971: "Identity and Necessity," in M. Munitz, (ed.), *Identity and Individuation*. New York: NYU Press.
- -----, 1972: "Naming and Necessity," in D. Davidson and G. Harmon, (eds.), *Semantics of Natural Language*. Boston: Reidel. Revised and reprinted as Kripke, 1980.
- -----, 1979: "A Puzzle about Belief," in A. Margalit, (ed.), *Meaning and Use:* Dordrecht: Reidel. Reprinted in Salnon and Soames, 1988.
- -----, 1980: *Naming and Necessity*. Oxford: Blackwell.
- Lewis, D. K., 1986: *On the Plurality of Worlds*. Oxford: Blackwell.
- -----, 1993: "Many But Almost One," in K. Campbell, J. Bacon, and L. Reinhardt (eds.), *Ontology, Causality and Mind: Essays on the Philosophy of D.M Armstrong*. Cambridge, England: Cambridge University Press. Reprinted in Lewis, 1999.
- -----, 1999: *Papers in Metaphysics and Epistemology*. Cambridge, England: Cambridge University Press.
- Lowe, E. J., 1982: "The Paradox of the 1,001 Cats," *Analysis*, 42, pp. 128-130.
- -----, 1989: *Kinds of Being*. Oxford: Basil Blackwell.
- -----, 1995: "Coinciding Objects: In Defense of the ‘Standard Account," *Analysis*, 55, pp.171-178.
- Mates, B., 1952: "Synonymity," in L. Linsky, (ed.), *Semantics and the Philosophy of Language,*

Champaign-Urbana: University of Illinois Press.

- McGinn, C., 2000: *Logical Properties*. Oxford: Blackwell.
- Myro, G., 1985: "Identity and Time," in *Philosophical Grounds of Rationality*: *Intentions, Categories, Ends*, R. Grandy and R. Warner, (eds.). New York and Oxford: Oxford University Press.
- Noonan, H., 1980: *Objects and Identity*. The Hague:
- -----, 1983: "The Necessity of Origin," *Mind*, 92, pp. 1-20.
- -----, 1991: "Indeterminate Identity, Contingent Identity and Abelardian Predicates," *The Philosophical Quarterly*, 41, pp. 183-193.
- -----, 1993: "Constitution is Identity," *Mind*, 102, pp. 133-146
- -----, 1997: "Relative Identity." In B. Hale and C. Wright (eds), *A Companion to the Philosophy of Language*. Oxford: Blackwell.
- Nozick. R., 1982: *Philosophical Explanations.* Cambridge, Mass.: Harvard University Press.
- O,Connor, D.J., 1967: "Substance and Attribute." In P. Edwards, (ed.), *The Encyclopedia of Philosophy*, New York: Macmillan.
- Oderberg, D., 1996: "Coincidence under a Sortal," *The Philosophical Review*, 105, pp. 145-171.
- Parsons, T., 2000: *Indeterminate Identity*. Oxford: Oxford University Press.
- Parfit, D., 1984: *Reasons and Persons.* Oxford: Oxford University Press.
- Perry, J., 1970: "The Same F," *The Philosophical Review*, 64, pp.181-200.
- -----, 1978: "Relative Identity and Number," *Canadian Journal of Philosophy* 8, pp. 1-15.
- Quine, W. V. O., 1960: *Word and Object.* Cambridge, Mass.: MIT Press.
- -----, 1963: *From a Logical Point of View.* New York: Harper and Row.
- -----, 1970: *Philosophy of Logic*. Englewood Cliffs, N.J.: Prentice Hall.
- Rea, M., 1995: "The Problem of Material Constitution," *The Philosophical Review*, 104, pp. 525-552.
- Rea, M. (ed.), 1997: *Material Constitution*: *A Reader*. Lanham, MD: Rowman and Littlefield.
- Read, S., 1995: *Thinking About Logic.* Oxford: Oxford University Press.
- Salmon, N., 1979: "How *Not* to Derive Essentialism from the Theory of Reference," *The Journal of Philosophy*, 76, pp. 703-725.
- -----, 1986: "Reflexivity," *Notre Dame Journal of Formal Logic*, 27, pp. 401-429. Reprinted in Salmon and Soames, 1988.
- -----, and Soames, S. (eds.), 1988 : *Propositions and Attitudes.* Oxford: Oxford University Press.
- -----, 1989: "The Logic of What Might Have Been," *Philosophical Review*, 98, pp.
- Sedley, D., 1982: "The Stoic Criterion of Identity," *Phronesis*, 27, pp. 255-275.
- Silver, C., 1994: *From Symbolic Logic...to Mathematical Logic.* Melbourne: Wm. C. Brown, Publishers.
- Simons, P., 1987: *Parts: A Study in Ontology*. Oxford: Clarendon Press.
- Swindler, J. K., 1991: *Weaving: An Analysis of the Constitution of Objects.* Savage, MD.: Rowman & Littlefield.
- Thompson, J., 1998: "The Statue and the Clay," *Noûs*, 32, pp. 149-173.
- Unger, P., 1980: "The Problem of the Many," *Midwest Studies in Philosophy*, 5, pp. 411-467.
- Van Inwagen, P., 1981: "The Doctrine of Arbitrary Undetached Parts," *Pacific Philosophical Quarterly* 62, pp. 123-137.

- -----, 1990: *Material Beings*. Ithaca, NY: Cornell University Press.
- Wiggins, D., 1967: *Identity and Spatio-Temporal Continuity*. Oxford: Blackwell.
- -----, 1968: "On Being in the Same Place at the Same Time," *Philosophical Review*, 77, pp. 90-95.
- -----, 1980: *Sameness and Substance*. Oxford: Blackwell.
- Williams, C.J.F., 1990: *What is Identity*. Oxford: Oxford University Press.
- Williamson, T., 1988: "First-order Logics for Comparative Similarity," *Notre Dame Journal of Formal Logic*, 29, pp.
- -----, 1990: *Identity and Discrimination*. Oxford: Blackwell.
- Yablo, S., 1987: "Identity, Essence and Indiscernibility," *Journal of Philosophy*, 84, pp. 293-314.
- Zalabardo, J., 2000: *Introduction to the Theory of Logic*. Boulder, Colorado: Westview Press.
- Zemach, E., 1974: "In Defense of Relative Identity," *Philosophical Studies*, 26, pp. 207-218.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

identity: of indiscernibles | logic: ancient | logic: classical | personal identity | properties | vagueness

Harry Deutsch
hdeutsch@ilstu.edu

First published: April 22, 2002
Content last modified: April 22, 2002

The Definition of Morality

The term “morality” can be used either

1. descriptively to refer to a code of conduct put forward by a society or,
    a. some other group, such as a religion, or
    b. accepted by an individual for her own behavior or
2. normatively to refer to a code of conduct that, given specified conditions, would be put forward by all rational persons.

How morality is defined plays a crucial, although often unacknowledged, role in formulating ethical theories. To take “morality” to refer to an actually existing code of conduct is quite likely to lead to some form of relativism. Among those who use “morality” normatively, different specifications of the conditions under which all rational persons would put forward a code of conduct result in different kinds of moral theories. To claim that “morality” in the normative sense does not have any referent, that is, to claim that there is no code of conduct that, under any plausible specified conditions, would be put forward by all rational persons, results in moral skepticism. Thus, although, not widely discussed, the definition of morality has great significance for moral theory.

1. Descriptive Definitions of “morality”

“Morality” is an unusual word. It is not used very much, at least not without some qualification. People do sometimes talk about “Christian morality,” “Nazi morality,” or about “the morality of the Greeks,” but they seldom talk simply about morality all by itself. Anthropologists used to claim that morality, like law, applied only within a society. They claimed that “morality” referred to that code of conduct that is put forward by a society. This account seems to fit best those societies that have no written language,

where often no distinctions are made among morality, etiquette, law, and religion. But even for anthropologists “morality” does not often mean simply “code of conduct put forward by a society.” Often, morality is distinguished from etiquette, law, and religion, all of which provide codes of conduct put forward by a society.

Etiquette is sometimes included as a part of morality, but it applies to behavior that is considered less serious than the kinds of actions to which morality usually applies. Law is distinguished from morality by having explicit rules, penalties, and officials who interpret the laws and apply the penalties, but there is often considerable overlap in the conduct governed by morality and that governed by law. Religion differs from morality in that it includes stories, usually about supernatural beings, that are used to explain or justify the behavior that it requires. Although there is often a considerable overlap in the conduct required by religion and that required by morality, morality provides only a guide to conduct, whereas religion always contains more than this. When “morality” is used simply to refer to a code of conduct put forward by a society, whether or not it is distinguished from etiquette, law, and religion, then it is being used in a completely descriptive sense.

When “morality” is used in this descriptive way, moralities can differ from each other in their content and in the foundation that members of the society claim their morality to have. A society might have a morality that is primarily concerned with practices not related to other persons, e.g., which days must be devoted to certain rituals, and might claim that their morality, which is concerned primarily with ritual, is based on the commands of God. Or a society might have a morality that is concerned primarily with sexual practices, and claim that their morality, which has this concern, is based on human nature. Or a society might regard morality as being concerned primarily with practices that minimize the harms that people suffer and claim that their morality, which has this concern, is based on reason. Many societies have moralities that are concerned with all of the above and that are claimed to have all three of the above foundations. But, in this sense of “morality,” regardless of its content, or the justification that those who accept the morality claim for it, the only universal features that all moralities have is that they are put forward by a society and they provide a guide for the behavior of the people in that society. In this sense of “morality,” morality might allow slavery or might allow some people with one skin color to behave in ways that those with a different skin color are not allowed to behave. In this sense of “morality,” it is not even essential that morality incorporate impartiality with regard to all moral agents, those people whose behavior is subject to moral judgments, or that it be universalizable in any significant way.

Although most philosophers do not use “morality” in this purely descriptive sense, some philosophers do. Ethical relativists are interested in these different moralities and claim that they are only kind of morality there is. [Edward Westermarck, *Ethical Relativity*] They deny that there is any universal normative morality. Ethical relativists are not merely, or even primarily, making a linguistic claim about the proper use of the English word, “morality.” However, they hold that only when the term “morality” is used in this descriptive sense is there something that it actually refers to, namely, a code of conduct put forward by a society. They claim that if “morality” is taken to refer to a universal code of conduct that would be endorsed by all rational persons, then there is no referent for the term “morality.” Although ethical relativists admit that many speakers of English use “morality” to refer to such a universal code of

conduct, they claim such persons are mistaken in thinking that there is anything that is the referent of that sense of the word. Ethical relativists are primarily concerned with denying that there is any universal morality that should be used by people in all societies to guide their own conduct and to make judgments about the conduct of others.

When “morality” refers to the codes of conduct of different societies, the features that are essential are that morality is a code of conduct that is put forward by a society and that it is used as a guide to behavior by the members of that society. In this descriptive sense, “morality” can refer to codes of conduct of different societies with widely differing content, and still be used unambiguously. However, there are now other descriptive senses of “morality.” In the sense most closely related to the original descriptive sense, “morality” refers to a guide to behavior put forward by some group other than a society, for example, a religious group. When the guide to conduct put forward by a religious group conflicts with the guide to conduct put forward by a society, it is not clear whether to say that there are conflicting moralities, or that the code of the religious group conflicts with morality. People who are members of that society and also members of the religious group, might differ with regard to the guide that they accept. They are likely to regard the guide they accept as the true morality.

In small homogeneous societies people do not belong to groups which put forward guides to behavior that conflict with the guide put forward by their society. There is only one guide to behavior that is accepted by all members of the society and that is the code of conduct that is put forward by the society. For such societies there is no ambiguity about which guide “morality” refers to. However, in those large societies where people often belong to groups that put forward guides to behavior that conflict with the guide put forward by their society, they do not always accept the guide put forward by their society. If they accept the conflicting guide of some other group to which they belong, often a religious group, rather than the guide put forward by their society, they will not regard the guide put forward by their society as a true or genuine morality.

This reveals an ambiguity in the original descriptive sense of morality that has two essential features: that morality is a code of conduct that is put forward by a society and that it is used as a guide to behavior by the members of that society. This ambiguity was not recognized because of the concentration on small homogeneous societies. Does “morality” refer only to those guides to conduct put forward by a society, or does it refer to guides to conduct put forward by other groups as well? There is another related ambiguity if the “code of conduct put forward by a society” is not “used as a guide to behavior by the members of that society.” Which of these essential features is most essential? The recognition that people in a society do not always accept the code of conduct that is put forward by their society presents problems for the descriptive sense of “morality” as the code of conduct put forward by a society and which used as a guide to behavior by the members of that society.

However, it is not useful to adopt a definition of “morality” as meaning the code of conduct accepted by the members of a society because in many large societies, not all members of the society accept the same code of conduct. Nor is it useful to adopt a somewhat more general definition of “morality” as the code of conduct accepted by the members of a group because it is not only always possible, it is often the case, that not all members of any group accept the same code. A natural outcome of these problems is to

switch attention from groups to individuals. If what is important is what code of conduct people accept, and members of a group do not always accept the same code of conduct, then why be concerned with groups at all?

This consideration leads to a new descriptive sense of “morality.” “morality” is taken to mean that guide to behavior that is regarded by an individual as overriding and that he wants to be universally adopted. [See R. M. Hare, *Moral Thinking*] In this sense of “morality,” it refers to a guide to behavior accepted by an individual rather than that put forward by a society or any other group. But “morality” does not refer to just any guide to behavior accepted by an individual, it is that guide to behavior that the individual adopts as his overriding guide, and wants everyone else to adopt as their overriding guide as well. This sense of “morality” is a descriptive sense, because a person can refer to an individual’s morality without endorsing it. In this sense, like the original descriptive sense, morality has no limitations on content. Whatever guide to behavior an individual regards as overriding and wants to be universally adopted is that individual’s morality.

When people explicitly talk about the morality of a group other than their own or of a person other than themselves, it is usually clear that they are using “morality” in a descriptive sense. However, when a person simply claims that morality prohibits or requires a given action, then the term “morality” is genuinely ambiguous. It is not clear whether it refers to (1) a guide to behavior that is put forward by a society, either his own or some other society; (1a) a guide that is put forward by a group, either one to which he belongs or another; or (1b) a guide that a person, perhaps himself, regards as overriding and wants adopted by everyone else, or (2) is a universal guide that all rational persons would put forward for governing the behavior of all moral agents. When a person uses “morality” to refer to a guide to conduct put forward by a group, unless it is his own group, it is usually only being used in its descriptive sense. No one referring to morality in that sense of “morality” need be endorsing it. When “morality” refers to a guide to conduct accepted by an individual, unless that individual is himself, it is usually being used in its descriptive sense. However, if the individual is referring to his own morality, he is endorsing it. Only (2) is always the normative sense of “morality,” but a person might hold that the morality referred to in (1), (1a), or (1b) is also the morality referred to in (2).

Some philosophers have put forward a sense of morality that seems to be a simple variation of (1b). In this sense, morality is a guide that a person, perhaps himself, regards as overriding, but need not want adopted by everyone else. In this technical philosophical sense of “morality,” “ethical egoism,” the view that one ought to take as one’s own self-interest as the overriding guide to behavior, is a morality. Sidgwick regarded egoism as one of the methods of ethics and, following Plato and Aristotle, “ethics” is sometimes taken as referring to a guide to behavior that an individual adopts as his own guide to life. However, in any normal sense of “morality,” morality cannot be a guide to behavior that a person wants others not to adopt. There is a sense of “morality” such that it does refer to a code of conduct adopted by an individual for his own use, but which he does not claim should be adopted by anyone else. However, this is correctly referred to as “morality” only when the individual would be willing for everyone else to adopt that code of conduct, but does not require that they do so, nor does he judge them to be immoral if they do not adopt it.

2. Normative Definitions of “morality”

When “morality” is used in its universal normative sense, it need not have either of the two features that are essential to moralities referred to by the original descriptive sense: that it be a code of conduct that is put forward by a society and that it be used as a guide to behavior by the members of that society. Indeed, it is possible that “morality” in the normative sense has never been put forward by any particular society, by any group at all, or even by any individual who regards it as overriding. “morality” is thus an ambiguous term, the features that account for what it refers to in any of the descriptive senses are not the features that account for what it refers to in its normative sense. The only feature that the descriptive and normative senses of “morality” have in common is that they refer to guides to behavior.

Those people who claim that there is a universal morality claim that it is a code of conduct that all rational persons would put forward for governing the behavior of all moral agents. They need not hold that every society has a code of conduct that has those features that they claim morality must have. They can admit that the guides to behavior of some societies lack so many of the essential features of morality that these societies do not even have a morality. They can also admit that many, perhaps most, societies have defective moralities, that although their guides to behavior have enough of the features of morality to be classified as moralities, they also lacks some essential features. Although those who hold that morality is universal do not claim that any actual society has or has ever had a guide to conduct that has all of the essential features of morality, they do not deny it either. They do claim that it is possible for any normal adult in any society to know what kinds of actions morality prohibits, requires, discourages, encourages, and allows. They also claim that morality applies to all of these persons, not only those now living, but also those who lived in the past. These are not empirical claims about morality, they are claims about what is essential to morality, or about what is meant by “morality” when it is used normatively.

On all accounts of morality, it is a code of conduct. However, on ethical or group relativist accounts or on individualistic accounts, morality has no special content or features that distinguishes it from nonmoral codes of conduct, such as law or religion. Just as a legal code of conduct can have almost any content, as long as it is capable of guiding behavior, and a religious code of conduct has no limits on content, all of the relativist and individualist accounts of morality, have no limit on the content of a moral code. However, for those, such as Hobbes, Kant, and Mill, who hold that morality is a code of conduct that all rational persons would put forward for governing the behavior of all moral agents, it has a fairly definite content. Although Kant, in accordance with the German word used to translate the English word “morality,” regards morality as applying to behavior that affects no one but the agent, he recognizes that it is commonly related to behavior that affects other people. Hobbes, Bentham, Mill, and most other philosophers writing in English limit morality to behavior that, directly or indirectly, affect others.

Although there are other significant differences between those philosophers who use “morality” to refer to a universal guide that all rational persons would put forward for governing the behavior of all moral agents, there are significant similarities. For all of these philosophers, such as Kurt Baier, Phillipa Foot, and Geoffrey Warnock, morality prohibits actions such as killing, causing pain, deceiving, and breaking

promises. For some, morality also requires charitable actions, but it does not require a justification for not being charitable on every possible occasion in the same way that it requires a justification for any act of killing, causing pain, deceiving, and breaking promises. For others, morality only encourages charitable actions, and no justification is ever needed for not being charitable. Rather, being charitable is supererogatory: it is always morally good to be charitable, but it is not morally required to be charitable.

On all of the accounts of morality as a universal guide that all rational persons would put forward for governing the behavior of all moral agents, it is concerned with promoting people living together in peace and harmony, not causing harm to others, and helping them. For most philosophers, the prohibitions against causing harm, directly or indirectly, are not taken as absolute. However, unlike most kinds of actions, a justification is needed for violating the prohibitions in order to avoid acting immorally. Some philosophers who hold a strict deontology, such as Kant, hold that it is never justified to do some of these kinds of actions. Those who hold that the principle of utility provides the foundation of morality, such as Mill, hold that it is justified to violate moral rules only when the overall direct and indirect consequences would be better. However, all those who use morality in its normative sense agree that the kinds of actions that directly or indirectly harm other people are the kinds of action with which morality is concerned.

The Natural Law tradition, from the Greeks to the present day, explicitly holds that all rational persons know what kinds of actions morality prohibits, requires, discourages, encourages, and allows. They also hold that reason endorses acting morally. Some hold that it is irrational to act immorally, but all hold that it is never irrational to act morally. Even religious thinkers in this tradition, such as Aquinas, hold that morality is known to all those whose behavior is subject to moral judgment, whether or not they know of the revelations of Christianity. Hobbes, who is in the natural law tradition, accepts all of the standard moral virtues, but complains that “the writers of moral philosophy, though they acknowledge the same virtues and vices, yet not seeing wherein consisted their goodness, nor that they come to be praised as the means of peaceable, sociable, comfortable living, place them in the mediocrity of the passions.” (*Leviathan*, Chapter 15, paragraph 40) The differences between those philosophers who hold that there is a universal morality is primarily about the foundation of morality, not about its content.

Neither Kant nor Mill regarded themselves as inventing or creating a new morality. Rather both of them, like Hobbes, regarded themselves as providing a justification for the morality that is accepted by all. Mill explicitly says:

> The intuitive, no less than what may be termed the inductive school of ethics, insists on the necessity of general laws. They both accept that the morality of an individual action is not a question of direct perception, but of the application of a law to an individual case. They recognize also, to a great extent, the same moral laws; but differ as to their evidence, and the source from which they derive their authority. (*Utilitarianism*, Chapter 1, paragraph 3)

According to Mill, Utilitarianism provides the foundation for morality. It explains and justifies the moral rules that are accepted by all. Kant also regards himself as performing the same task, explaining and

justifying a universal moral consciousness.

Some contemporary consequentialists claim that morality requires doing that act that would result in the best overall consequences. Others claim that morality requires following that rule that would result in the best overall consequences if everyone followed or accepted it. Since different consequentialists differ in their views about what consequences count as best, consequentialism does not provide a guide to conduct such that everyone knows what kinds of actions morality prohibits, requires, discourages, encourages, and allows. Of course, consequentialists think that there is a correct answer to the question about what counts as the best consequences, but they may not realize the importance of the fact that until that correct answer is found no one knows the kinds of actions morality prohibits, requires, discourages, encourages, and allows. Most consequentialists also hold that morality is universal, that it applies to all normal adult human beings. However, since not all normal adult human beings agree on what counts as the best consequences, morality no longer has an essential feature, namely, that all those who are subject to moral judgment know what kinds of actions morality prohibits, requires, discourages, encourages, and allows. Some consequentialists are not concerned with this normative sense of “morality,” but only with that guide to conduct that results in the best overall consequences. Others claim that morality does not have as an essential feature that all those subject to it know all of the kinds of actions it prohibits, etc. For them it is sufficient for morality to be that guide to behavior that leads to the best overall consequences.

In trying to provide a definition of the traditional normative sense of “morality,” I find it useful to regard morality as a public system. I use the phrase, “public system” to refer to a guide to conduct such that (1) all persons to whom it applies, all those whose behavior is to be guided and judged by that system, know what behavior the system prohibits, requires, discourages, encourages, and allows; and (2) it is not irrational for any of these persons to accept being guided and judged by that system. The paradigm examples of public systems are card games such as bridge or poker, or athletic games such as baseball, football, and basketball. Although a game is a public system, it applies only to those playing the game. Although, occasionally, someone may participate in a game without knowing its point or all of the rules that apply to those playing the game, the standard case is that all do know the point of the game as well as all of the relevant rules. If a person does not care enough about the game to abide by the rules, she can usually quit. Morality is the one public system that no rational person can quit. This is the point that Kant, without completely realizing it, captured by saying that morality is categorical. Morality applies to people simply by virtue of their being rational persons.

Since the normative sense of “morality” refers to a universal guide to behavior that all rational persons would put forward for governing the behavior of all moral agents, it is important to provide at least a brief account of what is meant by “rational person.” In this context, “rational person” is synonymous with “moral agent” and refers to those persons to whom morality applies. This includes all normal adults with sufficient knowledge and intelligence to understand what kinds of actions morality prohibits, requires, discourages, encourages, and allows, and with sufficient volitional ability to use morality as a guide for their behavior. Such persons must also seek to avoid any harm to themselves unless they believe that their action will result in someone, themselves or others, avoiding a comparable harm, or gaining a compensating good. People lacking these characteristics are not subject to moral judgment. If they lack them only temporarily, they might be excused from moral judgments in those cases.

The following definition of morality incorporates all of the essential features of morality as a guide to behavior that all rational persons would put forward for governing the behavior of all moral agents. *Morality is an informal public system applying to all rational persons, governing behavior thataffects others, and has the lessening of evil or harm as its goal*. In order to show that this definition incorporates all of the essential features of morality, I shall explain how the various parts of the definition incorporate these features.

To say that morality is a public system incorporates the essential feature that everyone who is subject to moral judgment knows what kinds of actions it prohibits, requires, discourages, encourages, and allows. It also guarantees that it is never irrational to act morally. That morality applies to all rational persons makes clear that the sense of "morality" being defined is that guide to conduct that applies to all rational persons. It would take considerably more space than is appropriate here to show that defining morality as a public system that applies to all rational persons also results in morality being a universal guide to behavior that all rational persons would put forward for governing the behavior of all moral agents. I should make clear that the claim that all rational persons would put forward this system only follows if limitations are put on the beliefs that rational persons can use and if they are attempting to reach agreement with similarly limited rational persons.

To say that morality is an informal system means that it has no authoritative judges and no decision procedure that provides unique answers to all moral questions. When it is important that disagreements be settled, societies use political and legal systems to supplement morality. These formal systems have the means to provide unique answers, but they do not provide a moral answer to the question. Rather, the question, being regarded as morally unresolvable, is transferred to the political or legal system. An important example of such a moral question is whether, and if so under what conditions, to allow abortion. There is continuing disagreement about this moral question, even though the legal and political system in the United States has provided fairly clear guidelines about the conditions under which abortion is allowed. Despite this important and controversial issue, morality, like all informal public systems, presupposes overwhelming agreement on most moral questions. No one thinks it is morally justified to cheat, deceive, injure, or kill simply in order to gain sufficient money to take a fantastic vacation. In the vast majority of moral situations, given agreement on the facts, no one disagrees, but for this very reason, these situations are never discussed. Thus, the overwhelming agreement on most moral matters is often overlooked.

The claim that morality governs behavior that affects others is somewhat controversial. Some have claimed that morality governs behavior that affects only the agent herself. Examples of behavior that supposedly affects only oneself, often include taking recreational drugs, masturbation, and developing one's talents. The German word for morality does include behavior that affects only the agent herself, and Kant may provide an accurate account of the German concept of morality. This concept of morality is more closely tied to its religious origin. However, the English concept of morality is more completely secular and almost all who distinguish morality from religion regard morality as governing only that behavior that directly or indirectly affects others. It is likely that regarding self-affecting behavior as governed by morality is a holdover from the time when morality was not clearly distinguished from

religion. This religious holdover might also affect the claim that some sexual practices such as homosexuality are immoral, but those who distinguish morality from religion do not regard homosexuality, per se, as a moral matter. Almost all American colleges and universities prohibit discrimination against homosexuals, which strongly suggests that they agree that only behavior that adversely affects others counts as immoral.

The final characteristic of morality -- that it has the lessening of evil or harm as its goal -- is also somewhat controversial. The Utilitarians talk about producing the greatest good as the goal of morality. However they include the lessening of harm as essential to producing the greatest good and almost all of their examples involve the avoiding or preventing of harm. The paradigm cases of moral precepts involve rules which prohibit causing harm directly or indirectly, such as rules prohibiting killing, causing pain, deceiving, and breaking promises. Even those precepts that require or encourage positive action, such as helping the needy, are almost always related to preventing or relieving harms. An examination of the paradigm examples of moral precepts make it clear that all of them involve the lessening of harms. It would be possible to include these paradigm examples of moral precepts in the definition of morality. This addition would make explicit that the normative sense of “morality” refers to that guide to behavior that we all regard as morality, but it not necessary to do so, because the proposed definition is sufficient to guarantee that these paradigm moral precepts will be part of the moral system.

Dictionary definitions of referring terms are usually just descriptions of the essential features of the referents of those terms. Insofar as all the referents of a term share the features that account for why that term refers to those referents, the term is not regarded as ambiguous. “morality” is an ambiguous term. Unlike the descriptive definitions of morality discussed earlier, which do not have any implications about how a person should behave, this normative definition of morality does have such implications. Hence it is not surprising that it is controversial. Agreeing to this definition commits a person to regarding some behavior as immoral, perhaps even behavior that he is tempted to perform. Although this definition allows as meaningful the question, “Why should I be moral?”, it guarantees that there is an answer that shows that it is not irrational to be moral, even though it may not show that it is irrational to be immoral. This definition also explains why we want others to act morally and why others want us to act morally. It thus does what definitions of referring terms are supposed to do: it clarifies this term's relationship to other terms with which it is related, and explains why we use the word in the way that we do.

Bibliography

- Baier, Kurt, *The Moral Point of View*, Ithaca, New York: Cornell University Press, 1958.
- Brandt, Richard, *A Theory of the Good and the Right*, New York: Oxford University Press, 1979.
- Foot, Phillippa, *Virtues and Vices, and Other Essays in Moral Philosophy*, Berkeley: University of California Press, 1978.
- Frankena, William, *Ethics*, Englewood Cliffs, N.J.: Prentice-Hall 1973.
- Frankena, William, *Thinking about Morality*, Ann Arbor: University of Michigan Press, 1980.
- Gert, Bernard, *Morality: Its Nature and Justification*, New York: Oxford University Press, 1998.
- Griffiths, A. Phillips, editor, *Ethics*, New York: Cambridge University Press, 1993.

- Hare, R. M., *Moral Thinking*, New York: Oxford University Press, 1981.
- Hobbes, Thomas, *Leviathan*, edited by Richard Tuck, New York: Cambridge University Press, 1991.
- Kant, Immanuel, *Groundwork of the Metaphysics of Morals*, New York: Barnes & Noble 1967.
- Mill, John Stuart, *Utilitarianism*, edited by Roger Crisp, New York: Oxford University Press, 1998.
- Moore, G. E., *Ethics*, New York: H. Holt and company, 1912.
- Moore, G. E., *Principia Ethica*, New York: Cambridge University Press, 1993.
- Sidgwick, Henry, *Outlines of the History of Ethics*, Boston: Beacon Press, 1960.
- Sidgwick, Henry, *Methods of Ethics*, Indianapolis: Hackett Pub. Co., 1981.
- Thomson, J. J. and Dworkin, G., editors, *Ethics*, New York: Harper & Row, 1968.
- Toulmin, Stephen, *An Examination of the Place of Reason in Ethics*Cambridge [Eng.]: University Press, 1960.
- Wallace, G. and Walker, A. D. M., editors, *The Definition of Morality*, London: Methuen 1970.
- Warnock, Geoffrey, *The Object of Morality*, London: Methuen, 1971.
- Westermarck, Edward, *Ethical Relativity*, Paterson, N.J.: Littlefield, Adams, 1960.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

consequentialism | ethics: natural law tradition | Hobbes, Thomas: moral and political philosophy | Kant, Immanuel | Mill, John Stuart | moral relativism

Bernard Gert
Bernard.Gert@Dartmouth.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 16, 2002
Content last modified: April 16, 2002

Hobbes’s Moral and Political Philosophy

The 17th Century English philosopher Thomas Hobbes is now widely regarded as one of a handful of truly great political philosophers, whose masterwork *Leviathan* rivals in significance the political writings of Plato, Aristotle, Locke, Rousseau, Kant, and Rawls. Hobbes is famous for his early and elaborate development of what has come to be known as “social contract theory”, the method of justifying political principles or arrangements by appeal to the agreement that would be made among suitably situated rational, free, and equal persons. He is infamous for having used the social contract method to arrive at the astonishing conclusion that we ought to submit to the authority of an absolute -- undivided and unlimited -- sovereign power. While his methodological innovation had a profound constructive impact on subsequent work in political philosophy, his substantive conclusions have served mostly as a foil for the development of more palatable philosophical positions. Hobbes’s moral philosophy has been less influential than his political philosophy, in part because that theory is too ambiguous to have garnered any general consensus as to its content. Most scholars have taken Hobbes to have affirmed some sort of personal relativism or subjectivism; but views that Hobbes espoused divine command theory, virtue ethics, rule egoism, or a form of projectivism also find support in Hobbes’s texts and among scholars. Because Hobbes held that “the true doctrine of the Lawes of Nature is the true Morall philosophie”, differences in interpretation of Hobbes’s moral philosophy can be traced to differing understandings of the status and operation of Hobbes’s “laws of nature”, which laws will be discussed below. The formerly dominant view that Hobbes espoused psychological egoism as the foundation of his moral theory is currently widely rejected, and there has been to date no fully systematic study of Hobbes’s moral psychology.

---

1. Major Political Writings

Hobbes wrote several versions of his political philosophy, including *The Elements of Law, Natural and Politic* (also under the titles *Human Nature* and *De Corpore Politico)* published in 1650, *De Cive* (1642) published in English as *Philosophical Rudiments Concerning Government and Society* in 1651, the English *Leviathan* published in 1651, and its Latin revision in 1668. Others of his works are also important in understanding his political philosophy, especially his history of the English Civil War, *Behemoth* (published 1679), *De Corpore* (1655), *De Homine* (1658), *Dialogue Between a Philosopher and a Student of the Common Laws of England* (1681), and *The Questions Concerning Liberty, Necessity, and Chance* (1656). All of Hobbes's major writings are collected in *The English Works of Thomas Hobbes*, edited by Sir William Molesworth (11 volumes, London 1839-45), and *Thomae Hobbesƒ Opera Philosophica Quae Latina Scripsit Omnia*, also edited by Molesworth (5 volumes; London, 1839-45). Readers new to Hobbes should begin with *Leviathan*, being sure to read Parts Three and Four, as well as the more familiar and often excerpted Parts One and Two. There are many fine overviews of Hobbes's normative philosophy, some of which are listed in the following selected bibliography of secondary works.

2. The Philosophical Project

Hobbes sought to discover rational principles for the construction of a civil polity that would not be subject to destruction from within. Having lived through the period of political disintegration culminating in the English Civil War, he came to the view that the burdens of even the most oppressive government are "scarce sensible, in respect of the miseries, and horrible calamities, that accompany a Civill Warre". Because virtually any government would be better than a civil war, and, according to Hobbes's analysis, all but absolute governments are systematically prone to dissolution into civil war, people ought to submit themselves to an absolute political authority. Continued stability will require that they also refrain from the sorts of actions that might undermine such a regime. In particular, Hobbes aimed to demonstrate the reciprocal relationship between political obedience and peace.

3. The State of Nature

To establish these conclusions, Hobbes invites us to consider what life would be like in a state of nature, that is, a condition without government. Perhaps we would imagine that people might fare best in such a state, where each decides for himself how to act, and is judge, jury and executioner in his own case whenever disputes arise-- and that at any rate, this state is the appropriate baseline against which to judge the justifiability of political arrangements. Hobbes terms this situation "the condition of mere nature", a

state of perfectly private judgment, in which there is no agency with recognized authority to arbitrate disputes and effective power to enforce its decisions.

Hobbes’s near descendant, John Locke, insisted in his *Second Treatise of Government* that the state of nature was indeed to be preferred to subjection to the arbitrary power of an absolute sovereign. But Hobbes famously argued that such a “dissolute condition of masterlesse men, without subjection to Lawes, and a coercive Power to tye their hands from rapine, and revenge” would make impossible all of the basic security upon which comfortable, sociable, civilized life depends. There would be “no place for industry, because the fruit thereof is uncertain; and consequently no culture of the earth; no navigation, nor use of the commodities that may be imported by Sea; no commodious Building; no Instruments of moving and removing such things as require much force; no Knowledge of the face of the Earth; no account of Time; no Arts; no Letters; and which is worst of all, continuall feare, and danger of violent death; And the life of man, solitary, poore, nasty, brutish, and short.” If this is the state of nature, men have strong reasons to avoid it, which can be done only by submitting to some mutually recognized public authority, for “so long a man is in the condition of mere nature, (which is a condition of war,) as private appetite is the measure of good and evill.”

Although many readers have criticized Hobbes’s state of nature as unduly pessimistic, he constructs it from a number of individually plausible empirical and normative assumptions. He assumes that people are sufficiently similar in their mental and physical attributes that no one is invulnerable nor can expect to be able to dominate the others. Hobbes assumes that people generally “shun death”, and that the desire to preserve their own lives is very strong in most people. While people have local affections, their benevolence is limited, and they have a tendency to partiality. Concerned that others should agree with their own high opinions of themselves, people are sensitive to slights. They make evaluative judgments, but often use seemingly impersonal terms like ‘good’ and ‘bad’ to stand for their own personal preferences. They are curious about the causes of events, and anxious about their futures; according to Hobbes, these characteristics incline people to adopt religious beliefs, although the content of those beliefs will differ depending upon the sort of religious education one has happened to receive.

With respect to normative assumptions, Hobbes ascribes to each person in the state of nature a liberty right to preserve herself, which he terms “the right of nature”. This is the right to do whatever one sincerely judges needful for one’s preservation; yet because it is at least possible that virtually anything might be judged necessary for one’s preservation, this theoretically limited right of nature becomes in practice an unlimited right to potentially anything, or, as Hobbes puts it, a right “to all things”. Hobbes further assumes as a principle of practical rationality, that people should adopt what they see to be the necessary means to their most important ends.

4. The State of Nature is a State of War

Taken together, these plausible descriptive and normative assumptions yield a state of nature potentially fraught with divisive struggle. The right of each to all things invites serious conflict, especially if there is competition for resources, as there will surely be over at least scarce goods such as the most desirable

lands, spouses, etc. People will quite naturally fear that others may (citing the right of nature) invade them, and may rationally plan to strike first as an anticipatory defense. Moreover, that minority of prideful or “vain-glorious” persons who take pleasure in exercising power over others will naturally illicit preemptive defensive responses from others. Conflict will be further fueled by disagreement in religious views, in moral judgments, and over matters as mundane as what goods one actually needs, and what respect one properly merits. Hobbes imagines a state of nature in which each person is free to decide for himself what he needs, what he’s owed, what’s respectful, right, pious, prudent, and also free to decide all of these questions for the behavior of everyone else as well, and to act on his judgments as he thinks best, enforcing his views where he can. In this situation where there is no common authority to resolve these many and serious disputes, we can easily imagine with Hobbes that the state of nature would become a “state of war”, even worse, a war of “all against all”.

5. Further Questions about the State of Nature

In response to the natural question whether humanity ever was generally in any such state of nature, Hobbes notes that all sovereigns are in this state with respect to one another. He opined that many now civilized peoples were formerly in that state, and some few peoples -- the savages of 17th C. America, for instance -- were still to his day in the state of nature. Most significantly, Hobbes asserts that the state of nature will be easily recognized by those whose formerly peaceful states have collapsed into civil war. While the state of nature’s condition of perfectly private judgment is an abstraction, something resembling it too closely for comfort remains a perpetually present possibility, to be feared, and avoided.

Do the other assumptions of Hobbes’s philosophy license the existence of this imagined state of isolated individuals pursuing their private judgments? Probably not, since, as feminist critics among others have noted, children are by Hobbes’s theory assumed to have undertaken an obligation of obedience to their parents in exchange for nurturing, and so the primitive units in the state of nature will include families ordered by internal obligations, as well as individuals. The bonds of affection, sexual affinity, and friendship -- as well as of clan membership and shared religious belief -- may further decrease the accuracy of any purely individualistic model of the state of nature. This concession need not impugn Hobbes’s analysis of conflict in the state of nature, since it may turn out that competition, diffidence and glory-seeking are disastrous sources of conflicts among small groups just as much as they are among individuals. Still, commentators seeking to answer the question how precisely we should understand Hobbes’s state of nature are investigating the degree to which Hobbes imagines that to be a condition of interaction among isolated individuals.

Another important open question is that of what, exactly, it is about human beings that makes it the case (supposing Hobbes is right) that our communal life is prone to disaster when we are left to interact according only to our own individual judgments. Perhaps, while people do wish to act for their own best long-term interest, they are shortsighted, and so indulge their current interests without properly considering the effects of their current behavior on their long-term interest. This would be a type of failure of rationality. Alternative, it may be that people in the state of nature are fully rational, but are trapped in a situation that makes it individually rational for each to act in a way that is sub-optimal for

all, perhaps finding themselves in the familiar ‘prisoner’s dilemma’ of game theory. Or again, it may be that Hobbes’s state of nature would be peaceful but for the presence of persons (just a few, or perhaps all, to some degree) whose passions overrule their calmer judgments; who are prideful, spiteful, partial, envious, jealous, and in other ways prone to behave in ways that lead to war. Such an account would understand irrational human passions to be the source of conflict. Which, if any, of these accounts adequately answers to Hobbes’s text is a matter of continuing debate among Hobbes scholars.

6. The Laws of Nature

Hobbes argues that the state of nature is a miserable state of war in which none of our important human ends are reliably realizable. Happily, human nature also provides resources to escape this miserable condition. Hobbes argues that each of us, as a rational being, can see that a war of all against all is inimical to the satisfaction of her interests, and so can agree that “peace is good, and therefore also the way or means of peace ƒ are good”. Humans will recognize as imperatives the injunction to seek peace, and to do those things necessary to secure it, when they can do so safely. Hobbes calls these practical imperatives “Lawes of Nature”, the sum of which is not to treat others in ways we would not have them treat us. These “precepts”, “conclusions” or “theorems” of reason are “eternal and immutable”, always commanding our assent even when they may not safely be acted upon. They forbid many familiar vices such as iniquity, cruelty, and ingratitude. Although commentators do not agree on whether these laws should be regarded as mere precepts of prudence, or rather as divine commands, or moral imperatives of some other sort, all agree that Hobbes understands them to direct people to submit to political authority. They tell us to seek peace with willing others by laying down part of our “right to all things”, by mutually covenanting to submit to the authority of a sovereign, and further direct us to keep that covenant establishing sovereignty.

7. Establishing Sovereign Authority

When people mutually covenant each to the others to obey a common authority, they have established what Hobbes calls “sovereignty by institution”. When, threatened by a conqueror, they covenant for protection by promising obedience, they have established “sovereignty by acquisition”. These are equally legitimate ways of establishing sovereignty, according to Hobbes, and their underlying motivation is the same-- namely fear-- whether of one’s fellows or of a conqueror. Political legitimacy depends not on how a government came to power, but only on whether it can effectively protect those who have consented to obey it; political obligation ends when protection ceases.

8. Absolutism

Although Hobbes offered some mild pragmatic grounds for preferring monarchy to other forms of government, his main concern was to argue that effective government -- whatever its form -- must have absolute authority. Its powers must be neither divided nor limited. The powers of legislation,

adjudication, enforcement, taxation, war-making (and the less familiar right of control of normative doctrine) are connected in such a way that a loss of one may thwart effective exercise of the rest; for example, legislation without interpretation and enforcement will not serve to regulate conduct. Only a government that possesses all of what Hobbes terms the “essential rights of sovereignty” can be reliably effective, since where partial sets of these rights are held by different bodies that disagree in their judgments as to what is to be done, paralysis of effective government, or degeneration into a civil war to settle their dispute, may occur.

Similarly, to impose limitation on the authority of the government is to invite irresoluble disputes over whether it has overstepped those limits. If each person is to decide for herself whether the government should be obeyed, factional disagreement -- and war to settle the issue, or at least paralysis of effective government -- are quite possible. To refer resolution of the question to some further authority, itself also limited and so open to challenge for overstepping its bounds, would be to initiate an infinite regress of non-authoritative ‘authorities’ (where the buck never stops). To refer it to a further authority itself unlimited, would be just to relocate the seat of absolute sovereignty, a position entirely consistent with Hobbes’s insistence on absolutism. To avoid the horrible prospect of governmental collapse and return to the state of nature, people should treat their sovereign as having absolute authority.

9. The Limits of Political Obligation

While Hobbes insists that we should regard our governments as having absolute authority, he reserves to subjects the liberty of disobeying those of their government’s commands that would require them to sacrifice their lives or honor, at least when the commonwealth’s survival does not depend on their doing so. This exception has understandably intrigued those who study Hobbes. His ascription of apparently inalienable rights -- what he calls the “true liberties of subjects” -- seems incompatible with his defense of absolute sovereignty. Moreover, if the sovereign’s failure to provide adequate protection to subjects extinguishes their obligation to obey, and if it is left to each subject to judge for herself the adequacy of that protection, it seems that people have never really exited the fearsome state of nature.

10. Religion and Social Instability

The last crucial aspect of Hobbes’s political philosophy is his treatment of religion. Hobbes progressively expands his discussion of Christian religion in each revision of his political philosophy, until it comes in *Leviathan* to comprise roughly half the book. There is no settled consensus on how Hobbes understands the significance of religion within his political theory. Some commentators have argued that Hobbes is trying to demonstrate to his readers the compatibility of his political theory with core Christian commitments, since it may seem that Christians’ religious duties forbid their affording the sort of absolute obedience to their governors which Hobbes’s theory requires of them. Others have doubted the sincerity of his professed Christianity, arguing that by the use of irony or other subtle rhetorical devices, Hobbes sought to undermine his readers’ religious beliefs. Howsoever his intentions are properly understood, Hobbes’s obvious concern with the power of religious belief is a fact that interpreters of his

political philosophy must seek to explain.

Selected Bibliography

The secondary literature on Hobbes's moral and political philosophy (not to speak of his entire body of work) is vast, appearing across many disciplines and in many languages. The following is a narrow selection of fairly recent works by philosophers, political theorists, and intellectual historians, available in English, on main areas of inquiry in Hobbes's moral and political thought. Very helpful for further reference is the critical bibliography of Hobbes scholarship to 1990 contained in Perez Zagorin, 'Hobbes on Our Mind', *Journal of the History of Ideas*, vol. 51, no. 2 (1990).

Journals

- *Hobbes Studies* is an annually published journal devoted to scholarly research on all aspects of Hobbes's work.

Collections

- K.C.Brown, ed., *Hobbes Studies* (1965), Cambridge, Mass., contains important papers by A.E. Taylor, J.W. N. Watkins, Howard Warrender, and John Plamenatz, among others.
- G.A.J. Rogers and Alan Ryan, eds., *Perspectives on Thomas Hobbes* (Oxford 1988); Mary G.Dietz, ed., *Thomas Hobbes and Political Theory* (Lawrence, Kansas, 1990).
- Tom Sorell, ed., *The Cambridge Companion to Hobbes* (Cambridge 1996).
- S.A. Lloyd, ed., 'Special Issue on Recent Work on the Moral and Political Philosophy of Thomas Hobbes', *Pacific Philosophical Quarterly*, vol. 82, nos. 3&4 (2001).

Books and Articles

- Ashcraft, Richard (1971). 'Hobbes's Natural Man: A Study in Ideology Formation', *Journal of Politics*, 33, pp. 1076-171.
- Baumgold, Deborah (1988). *Hobbes's Political Thought*, Cambridge.
- Boonin-Vail, David (1994). *Thomas Hobbes and the Science of Moral Virtue*, Cambridge.
- Curley, Edwin (1988). 'I durst not write so boldly: or how to read Hobbes' theological-political treatise', E. Giancotti, ed., *Proceedings of the Conference on Hobbes and Spinoza*, Urbino.
- ----- (1994). 'Introduction to Hobbes's *Leviathan*', *Leviathan* with selected variants from the Latin edition of 1668, Edwin Curley ed., Indianapolis, Indiana.
- Darwall, Stephen (1995). *The British Moralists and the Internal "Ought", 1640-1740*, Cambridge/New York
- ----- (2000). 'Normativity and Projection in Hobbes's *Leviathan*', *The Philosophical Review* vol. 109 no.3, pp.313-347.
- Ewin, R.E. (1991). *Virtues and Rights: The Moral Philosophy of Thomas Hobbes*,

Boulder/Oxford.

- Gauthier, David P. (1969). *The Logic of 'Leviathan': the Moral and political Theory of Thomas Hobbes*, Oxford.
- Gert, Bernard (1967). 'Hobbes and psychological egoism', *Journal of the History of Ideas*, 28, pp. 503-520.
- ----- (1978) 'Introduction to Man and Citizen', *Man and Citizen*, Bernard Gert, ed., Humanities Press.
- ----- (1988). 'The law of nature and the moral law', *Hobbes Studies*, I, pp.26-44.
- Goldsmith, M. M. (1966). *Hobbes's Science of Politics*, New York
- Hampton, Jean (1986). *Hobbes and the Social Contract Tradition*, Cambridge.
- Hood, E.C. (1964). *The Divine Politics of Thomas Hobbes*, Oxford.
- Johnston, David (1986). *The Rhetoric of 'Leviathan': Thomas Hobbes and the Politics of Cultural Transformation*, Princeton, N.J.
- Kavka, Gregory S. (1986). *Hobbesian Moral and Political Theory*, Princeton, N.J.
- Lloyd, S.A. (1992). *Ideals as Interests in Hobbes's 'Leviathan': the Power of Mind over Matter*, Cambridge.
- Macpherson, C. B. (1968). 'Introduction', *Leviathan*, C.B. Macpherson, ed., London.
- Martinich, A.P. (1992). *The Two Gods of Leviathan: Thomas Hobbes on Religion and Politics*, Cambridge.
- ----- (1999). *Hobbes: A Biography*, Cambridge.
- Nagel, Thomas (1959). 'Hobbes's Concept of Obligation', *Philosophical Review*, 68.
- Oakeshott, Michael (1975). *Hobbes on Civil Association*, Oxford.
- Raphael, D. D. (1977). *Hobbes: Morals and Politics*, London.
- Ryan, Alan (1986). 'A More Tolerant Hobbes?', Susan Mendus, ed., *Justifying Toleration*, Cambridge.
- Schneewind, J.B. (1997). *The Invention of Autonomy: History of Modern Moral Philosophy*, Cambridge/New York.
- Skinner, Quentin (1996). ,em>Reason and Rhetoric in the Philosophy of Hobbes, Cambridge.
- Sorell, Tom (1986). *Hobbes*, London.
- Strauss, Leo (1936). *The Political Philosophy of Hobbes: its Basis and Genesis*, Oxford.
- Tuck, Richard (1989). *Hobbes*, Oxford.
- ----- (1991). 'Introduction', *Leviathan*, Richard Tuck, ed., Cambridge.
- Warrender, Howard (1957). *The Political Philosophy of Hobbes: his Theory of Obligation*, Oxford.
- Watkins, J.W.N. (1965). *Hobbes's System of Ideas*, London.

Other Internet Resources

- Entry on Thomas Hobbes, Internet Encyclopedia of Philosophy

Related Entries

contractarianism | egoism | ethics: natural law tradition | game theory | Hobbes, Thomas: speculative philosophy | legal rights | liberalism | Locke, John: political philosophy | nature of law: natural law theories | prisoner's dilemma | rights

Sharon A. Lloyd
lloyd@usc.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 12, 2002
Content last modified: February 12, 2002

Contractarianism

"Contractarianism" names both a political theory of the legitimacy of political authority and a moral theory about the origin and/or legitimate content of moral norms. The political theory of authority claims that legitimate authority of government must derive from the consent of the governed, where the form and content of this consent derives from the idea of contract or mutual agreement. The moral theory of contractarianism claims that moral norms derive their normative force from the idea of contract or mutual agreement. Contractarians are thus skeptical of the possibility of grounding morality or political authority in either divine will or some perfectionist ideal of the nature of humanity. Social contract theorists from the history of political thought include Hobbes, Locke, Kant, and Rousseau. The most important contemporary political social contract theorist is John Rawls, who effectively resurrected social contract theory in the second half of the 20th century, along with David Gauthier, who is primarily a moral contractarian. There is no necessity for a contractarian about political theory to be a contractarian about moral theory, although most contemporary contractarians are both. It has been more recently recognized that there are two distinct strains of social contract thought, which now typically go by the names "contractarianism"and "contractualism."

Contractarianism, which stems from the Hobbesian line of social contract thought, holds that persons are primarily self-interested, and that a rational assessment of the best strategy for attaining the maximization of their self-interest will lead them to act morally (where the moral norms are determined by the maximization of joint interest) and to consent to governmental authority. Contractualism, which stems from the Kantian line of social contract thought, holds that rationality requires that we respect persons, which in turn requires that moral principles be such that they can be justified to each person. Thus, individuals are not taken to be motivated by self-interest but rather by a commitment to publicly justify the standards of morality to which each will be held. Where Gauthier or economist James Buchanan are the paradigm Hobbesian contractarians, Rawls or Thomas Scanlon would be the paradigm Kantian contractualists. The rest of this entry will specifically pertain to the contractarian strain wherever the two diverge.

Fundamental Elements of Contractarianism

The social contract has two fundamental elements: a characterization of the initial situation, called variously the "state of nature" by the modern political philosophers, the "original position" by Rawls, or the "initial bargaining position" by Gauthier, and a characterization of the parties to the contract, particularly in terms of their rationality and motivation to come to agreement. The initial situation posits what in bargaining theory is called the "no agreement position," the situation to which the individuals return in case of failure to make an agreement or contract. This situation may be more or less hostile, and more or less social, depending on what the theorist sees as human nature in the absence of rules of justice. But crucial to all contractarian theories, there is some scarcity or motivation for competition in the initial situation and there is some potential for gains from social interaction and cooperation.

In contemporary normative contractarian theories, that is, theories that attempt to ground the legitimacy of government or theories that claim to derive a moral ought, the initial position represents the starting point for a fair, impartial agreement. While contractualists justify the requirement of a fair, impartial agreement by reasons external to the contract itself, contractarians hold that the success of the contract in securing cooperative interaction itself requires that the starting point and procedures be fair and impartial.

Some points of controversy among contractarians concern the role of the initial situation in the theory: is it to be considered an actual historical situation, a possible historical moment, or is the contract situation completely hypothetical? Hume ("Of the Original Contract," pp.470-1) raised the decisive objection to any normative moral or political theory based on a historical contract: the consent of one's ancestors do not bind oneself. But Ronald Dworkin has raised similar concerns about a hypothetical contract: a hypothetical agreement, he objects, is no agreement at all. Hypothetical contract contractarians such as Gauthier counter that the point of the contract device is not to directly bind the contractors, but rather to provide a kind of thought experiment by which to discover the requirements of practical rationality. That is, they argue that if one is rational, and among rational others in circumstances in which agreement is both possible and beneficial, then rationality requires that one abide by the terms of the contract. While mainstream contractarian theories are hypothetical contract theories, a recent interesting and powerfully subversive use of contractarianism (Mills,1997; Pateman, 1989 -- see section on Subversion of Contractarianism below) reads the contract situation as historical agreements to erect and maintain white supremacy and patriarchy or male dominance. These latter contractarian theories are not justifications of the status quo, of course, but rather condemnations, and therefore do not face Hume's objection. Other questions that divide contemporary contractarians include: What are the ideal conditions and who are the ideal contractors that will make obligatory the outcomes of the contract for actual persons? What is the content of the hypothetical agreement?

The second element of a contractarian theory is the rationality of the contractors. First, contractarian (as opposed to contractualist) theories usually take persons to be self-interested in order to justify rules of morality or justice. This is because persons are assumed to have given preferences and interests, that do not necessarily include the well being of others, which is taken to be a moral preference and as such not prior to morality. Such preferences are called (by Gauthier, following the economist Wicksteed) "non-tuistic" preferences. Secondly, persons are presumed to want the benefits of social interaction if they can be had without sacrifice of individual self-interest. (See Feminist Perspectives on the Self (Section 1. Critique) for a critique of this conception of the rational person.) These two aspects of the contractarian individual in part imply what Rawls called the "circumstances of justice": the conditions under which rules for justice could be both possible and necessary. Justice, and so a social contract, is only possible where there is some possibility of benefit to each individual from cooperation. Social contract theories take individuals to be the best judges of their interests and the means to satisfy their desires. For this reason, there is a close connection between liberalism and contractarianism. However, that is not to say that all contractarian thought is liberal. Hobbes, for example, argued in favor of what Jean Hampton has called the "alienation contract," that is, a contract on the part of a people to alienate their rights to adjudicate their own disputes and self-defense to a sovereign, on the grounds that that was the only way to keep the peace given the nature of the alternative, which he famously characterized as life that would be "solitary, poore, nasty, brutish, and short." Thus, given a bad enough initial situation, contractarianism may lead to the legitimation of totalitarianism. Another point of criticism that arises from the characterization of the parties to the contract is that they must be able to contribute to the social product of interaction, or at least to threaten to destabilize it. This is because each individual has to be able to benefit from the inclusion of all those included. But this obviously leaves many, such as the severely disabled(though not the more able disabled -- see Silvers, 1998), outside the realm of justice, an implication that some find completely unacceptable. (Kittay, 1999)

Social contract theories also require some rules to guide the formation of agreement. Since they are prior to the contract, there must be some source of prior moral norms, whether natural, rational, or conventional. The first rule that is normally prescribed is that there must be no force or fraud in the making of the agreement. No one is to be "coerced" into agreement by the threat of physical violence. The reasoning for this is quite straightforwardly prudential: if one is allowed to use violence, then there is no real difference between the "contract" arrived at and the state of nature for the threatened party, and hence no security in the agreement. However, there is a fine line between being coerced by the threat of violence to giving up one's rights and being convinced by the threat of penury to make an unfavorable agreement. For this reason contractarians like Gauthier are able to argue for a fair and impartial starting point for bargaining that will lead to secure and stable agreements. The second rule of contract is that each individual who is a legitimate party to the contract must agree to the rules of justice, which is the outcome of the contract.

The Metaphor of Contract

The metaphor of the social contract requires some interpretation in order to apply it to the situation of

morality or politics. The interpretation can be specified by determining answers to three questions. First, what is the agreement on? Possible answers include the principles of justice (Rousseau, Rawls), the design of the basic social institutions (Rawls), the commitment to give up to a sovereign government (some or all of) one's rights (Hobbes, Locke), the adoption of a disposition to be (conventionally) moral (Gauthier, Hampton). The second question is how the agreement is to be thought of: as a hypothetical agreement? An actual historical agreement? An implicit historical situation? The third question is whether the contract device is to be used as justification or explanation. As discussed above, normative contractarianism uses the contract device primarily as justification, but it may be that Hobbes and Locke thought that there was an explanatory element to the contract device. As will be discussed below (Subversions of Contractarianism), an important contemporary contractarianism uses an implicit contract to explain the origin of oppression.

Morals by Agreement -- David Gauthier's Contractarianism

A brief sketch of an influential contractarian theory, David Gauthier's, is in order. Gauthier's project in *Morals By Agreement* is to employ a contractarian approach to grounding morality in rationality in order to defeat the moral skeptic. Gauthier assumes that humans can have no natural harmony of interests, and that there is much for each individual to be gained through cooperation. According to Gauthier, moral constraint on the pursuit of individual self-interest is required because cooperative activities almost inevitably involve a prisoner's dilemma: a situation in which the best individual outcomes can be had by those who cheat on the agreement while the others keep their part of the bargain. This leads to the socially and individually sub-optimal outcome wherein each can expect to be cheated by the other. But by disposing themselves to act according to the requirements of morality whenever others are also so disposed, they can gain each others' trust and cooperate successfully.

The contractarian element of the theory comes in the derivation of the moral norms. According to Gauthier, the compliance problem -- the problem of justifying rational compliance with the norms that have been accepted -- must always drive the justification of the initial situation and the conduct of the contracting situation. Gauthier likens the contract situation to a bargain, in which each party is trying to negotiate the moral rules that will allow them to realize optimal utility, and then he argues in favor of a bargaining solution that he calls "maximin relative concession." The idea of maximin relative concession is that each bargainer will be most concerned with the concessions that she makes from her ideal outcome relative to the concessions that others make. If she sees her concessions as reasonable relative to the others, considering that she wants to ensure as much for herself as she can while securing agreement (and thereby avoiding the zero-point: no share of the cooperative surplus) and subsequent compliance from the others, then she will agree to it. What would then be the reasonable outcome? Gauthier argues that it is the outcome that minimizes the maximum relative concessions of each party to the bargain.

Equally important to the solution as the procedure is the starting point from which the parties begin. For Gauthier there is no veil of ignorance -- each party to the contract is fully informed of their personal

attributes and holdings. But Gauthier argues that the initial position must have been arrived at non-coercively if compliance to the agreement is to be secured. He thus adopts what he calls the "Lockean proviso" (modeled after Locke's description of the initial situation of his social contract): that one cannot have bettered himself by worsening others. In sum, the moral norms that rational contractors will adopt (and comply with), according to Gauthier, are those norms that would be reached by the contractors beginning from a position each has attained through her own actions which have not worsened anyone else, and adopting as their principle for agreement the rule of maximin relative concession.

Critiques of Normative Contractarianism

Many critiques have been leveled against particular contractarian theories and against contractarianism as a framework for normative thought about justice or morality. (See Contemporary Approaches to the Social Contract.) Jean Hampton criticized Hobbes in her book *Hobbes and the Social Contract Tradition*, in a way that has direct relevance to contemporary contractarianism. Hampton argues that the characterization of individuals in the state of nature leads to a dilemma. Hobbes' state of nature as a potential war of all against all can be generated either as a result of passions (greed and fear, in particular) or rationality (prisoner's dilemma reasoning, in which the rational players each choose to renege on agreements made with each other). But if the passions account is correct, then Hampton argues, the contractors will still be motivated by these passions after the social contract is drawn up, and so will fail to comply with it. And if the rationality account is correct, then rational actors will not comply with the social contract any more than they will cooperate with each other before it is made.

This critique has an analog for Gauthier's theory, in that Gauthier must also claim that without the contract individuals will be stuck in some socially sub-optimal situation that is bad enough to motivate them to make concessions to each other for some agreement, yet the reason for their inability to cooperate without the contract cannot continue to operate after the contract is made. Gauthier's proposed solution to this problem is to argue that individuals will choose to dispose themselves to be constrained (self-interest) maximizers rather than straightforward (self-interest) maximizers, that is, to retrain themselves not to think first of their self-interest, but rather to dispose themselves to keep their agreements, provided that they find themselves in an environment of like-minded individuals. But this solution has been found dubitable by many commentators. (See Vallentyne, 1991)

Hampton also objects to the contemporary contractarian assumption that interaction is merely instrumentally valuable. She argues that if interaction were only valuable for the fruits of cooperation that it bears for self-interested cooperators, then it would be unlikely that those cooperators could successfully solve the compliance problem. In short, they are likely not to be able to motivate morality in themselves without some natural inclination to morality. Interestingly, Hampton agrees with Gauthier that contractarianism is right to require any moral or political norms to appeal to individuals self-interest as a limitation on self-sacrifice or exploitation of any individual.

In an important article, "On Being the Object of Property," African-American law professor Patricia Williams offers a critique of the contract metaphor itself. Contracts require independent agents who are

able to make and carry out promises without the aid of others. Historically, while white men have been treated as these pure wills of contract theory, Blacks and women have been treated as anti-will: dependent and irrational. Both ideals are false; whole people, she says, are dependent on other whole people. But by defining some as contractors and others as incapable of contract, whole classes of people can be excluded from the realm of justice. This point has been taken up by other critics of contractarianism, such as Eva Kittay (1999) who points out that not only are dependents such as children and disabled people left out of consideration by contractarian theories, but their caretakers' needs and interests will tend to be underestimated in the contract, as well.

Subversive Contractarianism

A descriptive use to which contractarianism has recently been put is to exploit the in-group/out-group nature of the contractarian project to illuminate the phenomenon of oppression. Carole Pateman's The Sexual Contract (1989) uses contractarian theory to argue that there has been an implicit contract among men to enforce patriarchy. She calls her approach a "conjectural history," which she uses both to illuminate the actual history of patriarchal oppression of women and the ideology of social contract theory. Similarly, Charles Mills argues in *The Racial Contract* (1997) that whites have had an actual, historical, sometimes explicit, though often only implicit, contract to enforce white supremacy. The arguments are similar in their contractarian outlines, though they differ in the historical and factual details. According to both theories, there are moral, political, and epistemological terms of the contract, and its effect has been to allow one group of persons effectively to dominate, subordinate, and exploit another group. The moral terms require the dominant group to evaluate the lives of their group more highly than those of the subordinated, the political to deprive the subordinate group of effective political power, and the epistemological terms require the members of the dominant group to see themselves as intellectually superior to the dominated. The social contract then can be seen as a justification by the parties to the contract of their interaction, and of their exploitation of those who are not parties to the contract, but only if the fundamental division of in-group and out-group is accepted. If the racial and sexual contracts were to be shown to be rational, they would constitute prima facie critiques of normative contractarianism, since they would then seem to justify racism and sexism.

Several of the critiques surveyed above, then, center on the questions: who is allowed to be a party to the contract, and how are those who are excluded from the contract to be treated? On the normative contractarian view, it is only rational to include all of those who can both benefit and reciprocate benefits to others. Normative contractarianism, then, on the assumption that non-whites and women can both benefit and reciprocate benefits to others, shows the sexual and racial contracts to be fundamentally irrational. Disability rights activists, however, would still have a serious complaint to lodge against normative contractarianism, since it is surely the case that there are persons who cannot reciprocate benefits to others. Such persons would be, on the normative contractarian view, beyond the scope of the rules of justice.

Bibliography

- Binmore, Ken, *Game Theory and the Social Contract*, vol. 1: Playing Fair and vol. 2: Just Playing, MIT, 1994 and 1998.
- Boucher, David and Paul Kelly, eds., *The Social Contract from Hobbes to Rawls*, Routledge, 1994.
- Gauthier, David, *Morals By Agreement*, Oxford, 1986.
- Gauthier, David, *Moral Dealing: Contract, Ethics, and Reason*, Cornell, 1990.
- Hampton, Jean, *Political Philosophy*, Westview, 1998.
- Hume, David, "Of the Original Contract," in *Essays, Moral, Political, and Literary*, Liberty Classics,1987 (first published 1777).
- Kittay, Eva Feder, *Love's Labor*, Routledge, 1999.
- Mills, Charles, *The Racial Contract*, Cornell, 1997.
- Pateman, Carole, *The Sexual Contract*, Stanford, 1989.
- Silvers, Anita, "Formal Justice," in *Disability, Difference, Discrimination*, Silvers, David Wasserman, and Mary B. Mahowald, eds., Rowman & Littlefield, 1998.
- Vallentyne, Peter, ed., *Contractarianism and Rational Choice*, Cambridge, 1991.
- Williams, Patricia, "On Being the Object of Property," in *The Alchemy of Race and Rights*, Harvard, 1991.

Other Internet Resources

- Political Philosophy
- Political Philosophy Reading Room
- Injustice Studies co-edited by Judith Little (Philosophy, SUNY/Potsdam), Thomas Simon (Philosophy, Illinois State), and Bruce Hawkins (English, Illinois State).

Related Entries

justice | liberalism | original position | social contract: contemporary approaches to

Ann E. Cudd
acudd@ukans.edu

First published: June 18, 2000
Content last modified: June 18, 2000

Moral Dilemmas

Moral dilemmas, at the very least, involve conflicts between moral requirements. Consider the cases given below.

1. Examples

In Book I of Plato's *Republic*, Cephalus defines 'justice' as speaking the truth and paying one's debts. Socrates quickly refutes this account by suggesting that it would be wrong to repay certain debts -- for example, to return a borrowed weapon to a friend who is not in his right mind. Socrates' point is not that repaying debts is without moral import; rather, he wants to show that it is not always right to repay one's debts, at least not exactly when the one to whom the debt is owed demands repayment. What we have here is a conflict between two moral norms: repaying one's debts and protecting others from harm. And in this case, Socrates maintains that protecting others from harm is the norm that takes priority.

Nearly twenty-four centuries later, Jean-Paul Sartre described a moral conflict the resolution of which was, to many, less obvious than the resolution to the Platonic conflict. Sartre [1957] tells of a student whose brother had been killed in the German offensive of 1940. The student wanted to avenge his brother and to fight forces that he regarded as evil. But the student's mother was living with him, and he was her one consolation in life. The student believed that he had conflicting obligations. Sartre describes him as

being torn between two kinds of morality: one of limited scope but certain efficacy, personal devotion to his mother; the other of much wider scope but uncertain efficacy, attempting to contribute to the defeat of an unjust aggressor.

While the examples from Plato and Sartre are the ones most commonly cited, it should be clear that there are many others. If a person makes conflicting promises, she faces a moral conflict. Physicians and families who believe that human life should not be shortened and that unpreventable pain should not be tolerated face a conflict in deciding whether to withdraw life support from a dying patient.

2. The Concept of Moral Dilemmas

What is common to the two well-known cases is conflict. In each case, an agent regards herself as having moral reasons to do each of two actions, but doing both actions is not possible. Ethicists have called situations like these *moral dilemmas*. The crucial features of a moral dilemma are these: the agent is required to do each of two (or more) actions; the agent can do each of the actions; but the agent cannot do both (or all) of the actions. The agent thus seems condemned to moral failure; no matter what she does, she will do something wrong (or fail to do something that she ought to do).

The Platonic case strikes many as too easy to be characterized as a genuine moral dilemma. For the agent's solution in that case is clear; it is more important to protect people from harm than to return a borrowed weapon. And in any case, the borrowed item can be returned later, when the owner no longer poses a threat to others. Thus in this case we can say that the requirement to protect others from serious harm *overrides* the requirement to repay one's debts by returning a borrowed item when its owner so demands. When one of the conflicting requirements overrides the other, we do not have a genuine moral dilemma. So in addition to the features mentioned above, in order to have a *genuine* moral dilemma it must also be true that neither of the conflicting requirements is overridden [Sinnott-Armstrong (1988), Chapter 1].

3. Problems

It is less obvious in Sartre's case that one of the requirements overrides the other. Why this is so, however, may not be so obvious. Some will say that our uncertainty about what to do in this case is simply the result of uncertainty about the consequences. If we were certain that the student could make a difference in defeating the Germans, the obligation to join the military would prevail. But if the student made little difference whatsoever in that cause, then his obligation to tend to his mother's needs would take precedence, since there he is virtually certain to be helpful. Others, though, will say that these obligations are equally weighty, and that uncertainty about the consequences is not at issue here.

Ethicists as diverse as Kant [1971/1797], Mill [1979/1861], and Ross [1930 and 1939] have assumed that an adequate moral theory should not allow for the possibility of genuine moral dilemmas. Only recently -- in the last forty years or so -- have philosophers begun to challenge that assumption. And the challenge

can take at least two different forms. Some will argue that it is *not possible* to preclude genuine moral dilemmas. Others will argue that even if it were possible, it is *not desirable* to do so.

To illustrate some of the debate that occurs regarding whether it is possible for any theory to eliminate genuine moral dilemmas, consider the following. The conflicts in Plato's case and in Sartre's case arose because there is more than one moral precept (using 'precept' to designate rules and principles), more than one precept sometimes applies to the same situation, and in some of these cases the precepts demand conflicting actions. One obvious solution here would be to arrange the precepts, however many there might be, hierarchically. By this scheme, the highest ordered precept always prevails, the second prevails unless it conflicts with the first, and so on. There are at least two glaring problems with this obvious solution, however. First, it just does not seem credible to hold that moral rules and principles should be hierarchically ordered. While the requirements to keep one's promises and to prevent harm to others clearly can conflict, it is far from clear that one of these requirements should *always* prevail over the other. In the Platonic case, the obligation to prevent harm is clearly stronger. But there can easily be cases where the harm that can be prevented is relatively mild and the promise that is to be kept is very important. And most other pairs of precepts are like this. This was a point made by Ross in *The Right and the Good* [1930, Chapter 2].

The second problem with this easy solution is deeper. Even if it were plausible to arrange moral precepts hierarchically, situations can arise in which the same precept gives rise to conflicting obligations. Perhaps the most widely discussed case of this sort is taken from William Styron's *Sophie's Choice* [1980] [Greenspan (1983)]. Sophie and her two children are at a Nazi concentration camp. A guard confronts Sophie and tells her that one of her children will be allowed to live and one will be killed. But it is Sophie who must decide which child will be killed. Sophie can prevent the death of either of her children, but only by condemning the other to be killed. The guard makes the situation even more excruciating by informing Sophie that if she chooses neither, then both will be killed. With this added factor, Sophie has a morally compelling reason to choose one of her children. But for each child, Sophie has an apparently equally strong reason to save him or her. Thus the same moral precept gives rise to conflicting obligations. Some have called such cases *symmetrical* [Sinnott-Armstrong (1988), Chapter 2].

4. Dilemmas and Consistency

We shall return to the issue of whether it is possible to preclude genuine moral dilemmas. But what about the desirability of doing so? Why have ethicists thought that their theories should preclude the possibility of dilemmas? At the intuitive level, the existence of moral dilemmas suggests some sort of inconsistency. An agent caught in a genuine dilemma is required to do each of two acts but cannot do both. And since he cannot do both, not doing one is a condition of doing the other. Thus, it seems that the same act is both required and forbidden. But exposing a logical inconsistency takes some work; for initial inspection reveals that the inconsistency intuitively felt is not present. Allowing *OA* to designate that the agent in question ought to do *A* (or is morally obligated to do *A*, or is morally required to do *A*), that *OA* and *OB* are both true is not itself inconsistent, even if one adds that it is not possible for the agent to do both *A* and *B*. And even if the situation is appropriately described as OA and O¬A, that is not a contradiction; the

contradictory of OA is ¬OA. [See Marcus (1980).]

Similarly rules that generate moral dilemmas are not inconsistent, at least on the usual understanding of that term. Ruth Marcus suggests plausibly that we “define a set of rules as consistent if there is some possible world in which they are all obeyable in all circumstances in *that* world.” Thus, “rules are consistent if there are possible circumstances in which no conflict will emerge,” and “a set of rules is inconsistent if there are *no* circumstances, no possible world, in which all the rules are satisfiable” [Marcus (1980), p. 128 and p. 129]. I suspect that Kant, Mill, and Ross were aware that a dilemma-generating theory need not be inconsistent. Even so, they would be disturbed if their own theories allowed for such predicaments. If I am correct in this speculation, it suggests that Kant, Mill, Ross, and others thought that there is an important theoretical feature that dilemma-generating theories lack. And this is understandable. It is certainly no comfort to an agent facing a reputed moral dilemma to be told that at least the rules which generate this predicament are consistent. For a good practical example, consider the situation of the criminal defense attorney. She is said to have an obligation to hold in confidence the disclosures made by a client and to be required to conduct herself with candor before the court (where the latter requires that the attorney inform the court when her client commits perjury) [Freedman (1975), Chapter 3]. It is clear that in this world these two obligations often conflict. It is equally clear that in some possible world -- for example, one in which clients do not commit perjury -- that both obligations can be satisfied. Knowing this is of no assistance to defense attorneys who face a conflict between these two requirements in this world.

Ethicists who are concerned that their theories not allow for moral dilemmas have more than consistency in mind, I think. What is troubling is that theories that allow for dilemmas fail to be *uniquely action-guiding*. A theory can fail to be uniquely action-guiding in either of two ways: by not recommending any action in a situation that is moral or by recommending incompatible actions. Theories that generate genuine moral dilemmas fail to be uniquely action-guiding in the latter way. Since at least one of the main points of moral theories is to provide agents with guidance, that suggests that it is desirable for theories to eliminate dilemmas, at least if doing so is possible.

But failing to be uniquely action-guiding is not the only reason that the existence of moral dilemmas is thought to be troublesome. Just as important, the existence of dilemmas does lead to inconsistencies if one endorses certain widely held theses. Here we shall consider two different arguments, each of which shows that one cannot consistently acknowledge the reality of moral dilemmas while holding selected principles.

The first argument shows that two standard principles of deontic logic are, when conjoined, incompatible with the existence of moral dilemmas. The first of these is the principle of deontic consistency

> **Principle of Deontic Consistency** (**PC**):
> $OA \rightarrow \neg O \neg A$.

Intuitively this principle just says that the same action cannot be both obligatory and forbidden. Note that as initially described, the existence of dilemmas does not conflict with PC. For as described, dilemmas

involve a situation in which an agent ought to do *A*, ought to do *B*, but cannot do both *A* and *B*. But if we add a principle of deontic logic, then we obtain a conflict with PC:

Principle of Deontic Logic (**PD**):
$\Box (A \rightarrow B) \rightarrow (OA \rightarrow OB)$.

Intuitively, PD just says that if doing *A* brings about *B*, and if *A* is obligatory (morally required), then *B* is obligatory (morally required). The *first argument* that generates inconsistency can now be stated.

(1)	OA	
(2)	OB	
(3)	¬ C (A & B)	[where '¬C' means 'cannot']
(4)	$\Box$ (A → B) → (OA → OB)	[where '$\Box$' means physical necessity]
(5)	$\Box$ ¬ (B & A)	(from 3)
(6)	$\Box$ (B → ¬ A)	(from 5)
(7)	$\Box$ (B → ¬ A) → (OB → O¬ A)	(an instantiation of 4)
(8)	OB → O¬ A	(from 6 and 7)
(9)	O¬ A	(from 2 and 8)
(10)	OA and O¬ A	(from 1 and 9)

Line (10) directly conflicts with PC. And from PC and (1), we can conclude

(11) ¬ O¬ A

And, of course, (9) and (11) are contradictory. So if we assume PC and PD, then the existence of dilemmas generates an inconsistency of the old-fashioned logical sort. [Note: In standard deontic logic, the '$\Box$' in PD typically designates logical necessity. Here I take it to indicate physical necessity so that the appropriate connection with premise (3) can be made. And I take it that logical necessity is stronger than physical necessity.]

Two other principles accepted in most systems of deontic logic entail PC. So if PD holds, then one of these additional two principles must be jettisoned too. The first says that if an action is obligatory, it is also permissible. The second says that an action is permissible if and only if it is not forbidden. These principles may be stated as:

(OP): OA → PA;

and

(D): $PA \leftrightarrow \neg O \neg A$.

The *second argument* that generates inconsistency, like the first, has as its first three premises a symbolic representation of a moral dilemma.

(1) OA
(2) OB
(3) $\neg C (A \& B)$

And like the first, this second argument shows that the existence of dilemmas leads to a contradiction if we assume two other commonly accepted principles. The first of these principles is that ‘ought’ implies ‘can’. Intuitively this says that if an agent is morally required to do an action, it must be possible for the agent to do it. We may represent this as

(4) $OA \rightarrow CA$ (for all A)

The other principle, endorsed by most systems of deontic logic, says that if an agent is required to do each of two actions, she is required to do both. We may represent this as

(5) $(OA \& OB) \rightarrow O(A \& B)$

The argument then proceeds:

(6) $O(A \& B) \rightarrow C(A \& B)$	(an instance of 4)
(7) $OA \& OB$	(from 1 and 2)
(8) $O(A \& B)$	(from 5 and 7)
(9) $\neg O(A \& B)$	(from 3 and 6)

So if one assumes that ‘ought’ implies ‘can’ and if one assumes the principle represented in (5) -- dubbed by some the agglomeration principle [Williams (1965)] -- then again a contradiction can be derived.

5. Responses to the Arguments

Now obviously the inconsistency in the first argument can be avoided if one denies either PC or PD. And the inconsistency in the second argument can be averted if one gives up either the principle that ‘ought’ implies ‘can’ or the agglomeration principle. There is, of course, another way to avoid these inconsistencies: deny the possibility of genuine moral dilemmas. It is fair to say that much of the debate concerning moral dilemmas in the last forty years has been about how to avoid the inconsistencies generated by the two arguments above.

Opponents of moral dilemmas have generally held that the crucial principles in the two arguments above are conceptually true, and therefore we must deny the possibility of genuine dilemmas. [See, for example, Conee (1982) and Zimmerman (1996).] Most of the debate, from all sides, has focused on the second argument. There is an oddity about this, however. When one examines the pertinent principles in each argument which, in combination with dilemmas, generates an inconsistency, there is little doubt that those in the first argument have a greater claim to being conceptually true than those in the second. Perhaps the focus on the second argument is due to the impact of Bernard Williams's influential essay [Williams (1965)]. But notice that the first argument shows that if there are genuine dilemmas, then either PC or PD must be relinquished. Even most supporters of dilemmas acknowledge that PC is quite basic. E.J. Lemmon, for example, notes that if PC does not hold in a system of deontic logic, then all that remains are truisms and paradoxes [Lemmon (1965), p. 51]. And giving up PC also requires denying either OP or D, each of which also seems basic. There has been much debate about PD -- in particular, questions generated by the Good Samaritan paradox -- but still it seems basic. So those who want to argue against dilemmas purely on conceptual grounds are better off focusing on the first of the two arguments above.

Some opponents of dilemmas also hold that the pertinent principles in the second argument -- the principle that 'ought' implies 'can' and the agglomeration principle -- are conceptually true. But foes of dilemmas need not say this. Even if they believe that a conceptual argument against dilemmas can be made by appealing to PC and PD, they have several options regarding the second argument. They may defend 'ought' implies 'can', but hold that it is a substantive normative principle, not a conceptual truth. Or they may even deny the truth of 'ought' implies 'can' or the agglomeration principle, though not because of moral dilemmas, of course.

Defenders of dilemmas need not deny all of the pertinent principles, of course. If one thinks that each of the principles at least has some initial plausibility, then one will be inclined to retain as many as possible. Among the earlier contributors to this debate, some took the existence of dilemmas as a counterexample to 'ought' implies 'can' [for example, Lemmon (1962) and Trigg (1971)]; others, as a refutation of the agglomeration principle [for example, Williams (1965) and van Fraassen (1973)]. The most common response to the first argument was to deny PD.

Friends and foes of dilemmas have a burden to bear in responding to the two arguments above. For there is at a *prima facie* plausibility to the claim that there are moral dilemmas and to the claim that the relevant principles in the two arguments are true. Thus each side must at least give reasons for denying the pertinent claims in question. Opponents of dilemmas must say something in response to the positive arguments that are given for the reality of such conflicts. One reason, as noted above, is simply pointing to examples. The case of Sartre's student and that from *Sophie's Choice* are good ones; and clearly these can be multiplied indefinitely. It will tempting for supporters of dilemmas to say to opponents, "If this is not a real dilemma, then tell me *what* the agent ought to do and *why*?" It is obvious, however, that attempting to answer such questions is fruitless, and for at least two reasons. First, any answer given to the question is likely to be controversial, certainly not always convincing. And second, this is a game that will never end; example after example can be produced. The more appropriate response on the part of foes of dilemmas is to deny that they need to answer the question. Examples as such cannot establish the

reality of dilemmas. Surely most will acknowledge that there are situations in which an agent does not know what he ought to do. This may be because of factual uncertainty, uncertainty about the consequences, uncertainty about what principles apply, or a host of other things. So for any given case, the mere fact that one does not know which of two (or more) conflicting obligations prevails does not show that none does.

Another reason in support of dilemmas to which opponents must respond is the point about symmetry. As the cases from Plato and Sartre show, moral rules can conflict. But opponents of dilemmas can argue that in such cases one rule overrides the other. Most will grant this in the Platonic case, and opponents of dilemmas will try to extend this point to all cases. But the hardest case for opponents is the symmetrical one, where the same precept generates the conflicting requirements. The case from *Sophie's Choice* is of this sort. It makes no sense to say that a rule or principle overrides itself. So what do opponents of dilemmas say here? They are apt to argue that the pertinent, all-things-considered requirement in such a case is disjunctive: Sophie should act to save one or the other of her children, since that is the best that she can do [for example, Zimmerman (1996), Chapter 7]. Such a move need not be *ad hoc*, since in many cases it is quite natural. If an agent can afford to make a meaningful contribution to only one charity, the fact that there are several worthwhile candidates does not prompt many to say that the agent will fail morally no matter what he does. Nearly all of us think that he should give to one or the other of the worthy candidates. Similarly, if two people are drowning and an agent is situated so that she can save either of the two but only one, few say that she is doing wrong no matter which she saves. Positing a disjunctive requirement in these cases seems perfectly natural, and so such a move is available to opponents of dilemmas as a response to symmetrical cases.

Supporters of dilemmas have a burden to bear too. They need to cast doubt on the adequacy of the pertinent principles in the two arguments that generate inconsistencies. And most importantly, they need to provide independent reasons for doubting whichever of the principles they reject. If they have no reason other than cases of putative dilemmas for denying the principles in question, then we have a mere standoff. Of the principles in question, the most commonly questioned on independent grounds are the principle that 'ought' implies 'can' and PD. Among supporters of dilemmas, Walter Sinnott-Armstrong [Sinnott-Armstrong (1988), Chapters 4 and 5] has gone to the greatest lengths to provide independent reasons for questioning some of the relevant principles.

6. Moral Residue and Dilemmas

One well-known argument for the reality of moral dilemmas has not been discussed yet. This argument might be called "phenomenological." It appeals to the emotions that agents facing conflicts experience and our assessment of those emotions.

Return to the case of Sartre's student. Suppose that he joins the Free French forces. It is likely that he will experience remorse or guilt for having abandoned his mother. And not only will he experience these emotions, this moral residue, but it is appropriate that he does. Yet, had he stayed with his mother and not joined the Free French forces, he also would have appropriately experienced remorse or guilt. But

remorse or guilt are appropriate only if the agent properly believes that he has done something wrong (or failed to do something that he was all-things-considered required to do). Since no matter what the agent does he will appropriately experience remorse or guilt, then no matter what he does he will have done something wrong. Thus, the agent faces a genuine moral dilemma. [The best known proponents of arguments for dilemmas that appeal to moral residue are Williams (1965) and Marcus (1980).]

Many cases of moral conflict are similar to this example. Certainly the case from *Sophie's Choice* fits here. No matter which of her children Sophie saves, she will experience enormous guilt for the consequences of that choice. Indeed, if Sophie did not experience such guilt, we would think that there was something morally wrong with her. In these cases, proponents of the argument (for dilemmas) from moral residue must claim that four things are true: (1) when the agents acts, she experiences remorse or guilt; (2) that she experiences these emotions is appropriate and called for; (3) had the agent acted on the other of the conflicting requirements, she would also have experienced remorse or guilt; and (4) in the latter case these emotions would have been equally appropriate and called for [McConnell (1996), pp. 37-38]. In these situations, then, remorse or guilt will be appropriate no matter what the agent does and these emotions are appropriate only when the agent has done something wrong. Therefore, these situations are genuinely dilemmatic.

There is much to say about the moral emotions and situations of moral conflict; the positions are varied and intricate. Without pretending to resolve all of the issues here, it will be pointed out that opponents of dilemmas have raised two different objections to the argument from moral residue. The first objection, in effect, suggests that the argument is question-begging [McConnell (1978) and Conee (1982)]; the second objection challenges the assumption that remorse and guilt are appropriate only when the agent has done wrong.

To explain the first objection, note that it is uncontroversial that some bad feeling or other is called for when an agent is in a situation like that of Sartre's student or Sophie. But the negative moral emotions are not limited to remorse and guilt. Among these other emotions, consider regret. An agent can appropriately experience regret even when she does not believe that she has done something wrong. For example, a parent may appropriately regret that she must punish her child even though she correctly believes that the punishment is deserved. Her regret is appropriate because a bad state of affairs is brought into existence (say, the child's discomfort), even when bringing this state of affairs into existence is morally required. Regret can even be appropriate when one has no causal connection at all with the bad state of affairs. It is appropriate for me to regret the damage that a recent fire has caused to my neighbor's house, the pain that severe birth defects cause in infants, and the suffering that a starving animal experiences in the wilderness. Not only is it appropriate that I experience regret in these cases, but I would probably be regarded as morally lacking if I did not.

With remorse or guilt, at least two components are present: the *experiential* component, namely, the negative feeling that the agent has; and the *cognitive* component, namely, the belief that the agent has done something wrong and takes responsibility for it. Although this same cognitive component is not part of regret, the negative feeling is. And the experiential component alone cannot serve as a gauge to distinguish regret from remorse, for regret can range from mild to intense, and so can remorse. In part,

what distinguishes the two is the cognitive component. But now when we examine the case of an alleged dilemma, such as that of Sartre's student, it is question-begging to assert that it is appropriate for him to experience remorse no matter what he does. No doubt, it is appropriate for him to experience *some* negative feeling. To say, however, that it is remorse that is called for is to assume that the agent appropriately believes that he has done something wrong. Since regret is warranted even in the absence of such a belief, to assume that remorse is appropriate is to *assume*, not argue, that the agent's situation is genuinely dilemmatic. Opponents of dilemmas can say that one of the requirements overrides the other, or that the agent faces a disjunctive requirement, and that regret is appropriate because even when he does what he ought to do, some bad will ensue. Either side, then, can account for the appropriateness of some negative moral emotion. To get more specific, however, requires more than is warranted by the present argument. This appeal to moral residue, then, does not establish the reality of moral dilemmas.

Matters are even more complicated, though, as the second objection to the argument from moral residue shows. The argument assumes that remorse or guilt is appropriate only if the agent believes that he has done something wrong. But this is questionable. Consider the case of a middle-aged man, Bill, and a seven-year-old boy, Johnny. It is set in a midwestern village on a snowy December day. Johnny and several of his friends are riding their sleds down a narrow, seldom used street, one that intersects with a busier, although still not heavily traveled, street. Johnny, in his enthusiasm for sledding, is not being very careful. During his final ride he skidded under an automobile passing through the intersection and was killed instantly. The car was driven by Bill. Bill was driving safely, had the right of way, and was not exceeding the speed limit. Moreover, given the physical arrangement, it would have been impossible for Bill to have seen Johnny coming. Bill was not at fault, legally or morally, for Johnny's death. Yet Bill experienced what can only be described as remorse or guilt about his role in this horrible event.

At one level, Bill's feelings of remorse or guilt are not warranted. Bill did nothing wrong. A friend might even recommend that Bill seek therapy. But this is not all there is to say. Most of us understand Bill's response. From Bill's point of view, the response is not inappropriate, not irrational, not uncalled-for. To see this, imagine that Bill had had a very different response. Suppose that Bill had said, "I regret Johnny's death. It is a terrible thing. But it certainly was not my fault. I have nothing to feel guilty about and I don't owe his parents any apologies." Even if Bill is correct intellectually, it is hard to imagine someone being able to achieve this sort of objectivity about his own behavior. When human beings have caused great harm, it is natural for them to wonder if they are at fault, even if to outsiders it is obvious that they bear no moral responsibility for the damage. Human beings are not so finely tuned emotionally that when they have been *causally* responsible for harm, they can easily turn guilt on or off depending on their degree of *moral* responsibility. [See Zimmerman (1988), pp. 134-135.] And this is not a bad thing; for it likely makes agents more cautious about their actions, more sensitive about their responsibilities, and more empathetic regarding the plight of others.

All of this suggests that there are situations in which an agent's remorse or guilt is not inappropriate even though the agent has done nothing wrong. Because of this and because in any given situation the appropriate response may be regret and not remorse, opponents of dilemmas have a way to respond to the argument that appeals to the appropriateness of remorse.

It should be noted, however, that there is a complex array of issues concerning the relationship between ethical conflicts and the moral emotions, and the discussion here has been quite brief. [See Greenspan (1995).]

7. Types of Moral Dilemmas

In the literature on moral dilemmas, it is common to draw distinctions among various types of dilemmas. Only some of these distinctions will be mentioned here. It is worth noting that both supporters and opponents of dilemmas tend to draw some, if not all, of these distinctions. And in most cases the motivation for doing so is clear. Supporters of dilemmas may draw a distinction between dilemmas of type V and W. The upshot is typically a message to opponents of dilemmas: "You think that all moral conflicts are resolvable. And that is understandable, because conflicts of type V are resolvable. But conflicts of type W are not resolvable. Thus, contrary to your view, there are some genuine moral dilemmas." By the same token, opponents of dilemmas may draw a distinction between dilemmas of type X and Y. And their message to supporters of dilemmas is this: "You think that there are genuine moral dilemmas, and given certain facts, it is understandable why this appears to be the case. But if you draw a distinction between conflicts of types X and Y, you can see that appearances can be explained by the existence of type X alone, and type X conflicts are not genuine dilemmas." With this in mind, let us note a few of the distinctions.

One distinction is between *epistemic* conflicts and *ontological* conflicts. The former involve conflicts between two (or more) moral requirements and the agent does not know which of the conflicting requirements takes precedence in her situation. Everyone concedes that there can be situations where one requirement does take priority over the other with which it conflicts, though at the time action is called for it is difficult for the agent to tell which requirement prevails. The latter are conflicts between two (or more) moral requirements, and neither is overridden. This is not simply because the agent does not *know* which requirement is stronger; neither is. Genuine moral dilemmas, if there are any, are ontological. Both opponents and supporters of dilemmas acknowledge that there are epistemic conflicts.

Another distinction is between *self-imposed* moral dilemmas and dilemmas imposed on an agent *by the world*, as it were. Conflicts of the former sort arise because of the agent's own wrongdoing [Aquinas; Donagan (1977 and 1984); and McConnell (1978)]. If an agent made two promises that he knew conflicted, then through his own actions he created a situation in which it is not possible for him to discharge both of his requirements. Dilemmas imposed on the agent by the world, by contrast, do not arise because of the agent's wrongdoing. The case of Sartre's student is an example, as is the case from *Sophie's Choice*. For supporters of dilemmas, this distinction is not all that important. But among opponents of dilemmas, there is a disagreement about whether the distinction is important. Some of these opponents hold that self-imposed dilemmas are possible, but that their existence does not point to any deep flaws in moral theory. Moral theory tells agents how they ought to behave; but if agents violate moral norms, of course things can go askew. Other opponents deny that even self-imposed dilemmas are possible. They argue that an adequate moral theory should tell agents what they ought to do in their current circumstances, regardless of how those circumstances arose. And given the prevalence of

wrongdoing, if a moral theory did not issue uniquely action-guiding "contrary-to-duty imperatives," it would be severely lacking.

Yet another distinction is between *obligation dilemmas* and *prohibition dilemmas*. The former are situations in which more than one feasible action is obligatory. The latter involve cases in which all feasible actions are forbidden. Some [especially, Valentyne (1987 and 1989)] argue that plausible principles of deontic logic may well render obligation dilemmas impossible; but they do not preclude the possibility of prohibition dilemmas. The case of Sartre's student, if genuinely dilemmatic, is an obligation dilemma; Sophie's case is a prohibition dilemma. There is another reason that friends of dilemmas emphasize this distinction. Some think that the "disjunctive solution" used by opponents of dilemmas -- when equally strong precepts conflict, the agent is required to act on one or the other -- is much more plausible when applied to obligation dilemmas than when applied to prohibition dilemmas.

As moral dilemmas are typically described, they involve a *single agent*. The agent ought, all things considered, to do A, ought, all things considered, to do B, and she cannot do both A and B. But we can distinguish *multi-person* dilemmas from single agent ones. The two-person case is representative of multi-person dilemmas. The situation is such that one agent, P1, ought to do A, a second agent, P2, ought to do B, and though each agent can do what he ought to do, it is not possible both for P1 to do A and P2 to do B. [See Marcus (1980), p. 122 and McConnell (1988).] Multi-person dilemmas have been called "interpersonal moral conflicts." Such conflicts are most theoretically worrisome if the same moral system (or theory) generates the conflicting obligations for P1 and P2. A theory that precludes single-agent moral dilemmas remains uniquely action-guiding for each agent. But if that same theory does not preclude the possibility of interpersonal moral conflicts, not all agents will be able to succeed in discharging their obligations, no matter how well-motivated or how hard they try. For supporters of moral dilemmas, this distinction is not all that important. They no doubt welcome (theoretically) more types of dilemmas, since that may make their case more persuasive. But if they establish the reality of single-agent dilemmas, in one sense their work is done. For opponents of dilemmas, however, the distinction may be important. This is because at least some opponents believe that the conceptual argument against dilemmas applies principally to single-agent cases. It does so because the ought-to-do operator of deontic logic and the accompanying principles are properly understood to apply to entities about which decisions can be made. And while an individual act involving one agent can be the object of choice, a compound act involving multiple agents is difficult so to conceive. [See Smith (1986) and Thomason (1981).] To the extent that the possibility of interpersonal moral conflicts raises an intramural dispute among opponents of dilemmas, that dispute concerns how to understand the principles of deontic logic and what can reasonably be demanded of moral theories.

8. Conclusion

Debates about moral dilemmas have been extensive during the last four decades. These debates go to the heart of moral theory. Both supporters and opponents of moral dilemmas have major burdens to bear. Opponents of dilemmas must show why appearances are deceiving. Why are examples of apparent dilemmas misleading? Why are certain moral emotions appropriate if the agent has done no wrong?

Supporters must show why several of many apparently plausible principles should be given up -- principles such as PC, PD, OP, D, ‘ought’ implies ‘can’, and the agglomeration principle. Much progress has been made, but the debate is apt to continue.

Bibliography

Cited Works

- Aquinas, St. Thomas, 1964-1975, *Summa Theologiae*, Trans, Thomas Gilby *et al*, New York: McGraw-Hill.
- Donagan, Alan, 1977, *The Theory of Morality*, Chicago: University of Chicago Press.
- -----, 1984, “Consistency in Rationalist Moral Systems,” *The Journal of Philosophy* **81** : 291-309, [Reprinted in Gowans (1987): 271-290,]
- Freedman, Monroe, 1975, *Lawyers' Ethics in an Adversary System*, Indianapolis: Bobbs-Merrill.
- Gowans, Christopher W. (editor), 1987, *Moral Dilemmas*, New York: Oxford University Press.
- Greenspan, Patricia S., 1983, “Moral Dilemmas and Guilt,” *Philosophical Studies* **43** : 117-125,
- -----, 1995, *Practical Guilt: Moral Dilemmas, Emotions, and Social Norms*, New York: Oxford University Press.
- Kant, Immanuel, 1971/1797, *The Doctrine of Virtue: Part II of the Metaphysics of Morals*, Trans, Mary J. Gregor, Philadelphia: University of Pennsylvania Press.
- Lemmon, E.J., 1962, “Moral Dilemmas,” *The Philosophical Review* **70** : 139-158, [Reprinted in Gowans (1987): 101-114.
- -----, 1965, “Deontic Logic and the Logic of Imperatives,” *Logique et Analyse* **8** : 39-71.
- Marcus, Ruth Barcan, 1980, “Moral Dilemmas and Consistency,” *The Journal of Philosophy* **77** : 121-136, [Reprinted in Gowans (1987): 188-204,]
- Mason, H.E., (editor), 1996, *Moral Dilemmas and Moral Theory*, New York: Oxford University Press.
- McConnell, Terrance, 1978, “Moral Dilemmas and Consistency in Ethics,” *Canadian Journal of Philosophy* **8** : 269-287, [Reprinted in Gowans (1987): 154-173,]
- -----, 1988, “Interpersonal Moral Conflicts,” *American Philosophical Quarterly* **25** : 25-35.
- -----, 1996, “Moral Residue and Dilemmas,” in Mason (1996): 36-47.
- Mill, John Stuart, 1979/1861, *Utilitarianism*, Indianapolis: Hackett Publishing.
- Plato, *The Republic*, trans, Paul Shorey, in *The Collected Dialogues of Plato*, E. Hamilton and H. Cairns (eds.), Princeton: Princeton University Press.
- Ross, W.D., 1930, *The Right and the Good*, Oxford: Oxford University Press.
- -----, 1939, *The Foundations of Ethics*, Oxford: Oxford University Press.
- Sartre, Jean-Paul, 1957/1946, “Existentialism is a Humanism,” Trans, Philip Mairet, in Walter Kaufmann (ed.), *Existentialism from Dostoevsky to Sartre*, New York: Meridian, 287-311,
- Sinnott-Armstrong, Walter, 1988, *Moral Dilemmas*, Oxford: Basil Blackwell.
- Smith, Holly M., 1986, “Moral Realism, Moral Conflict, and Compound Acts,” *The Journal of Philosophy* **83** : 341-345.
- Styron, William, 1980, *Sophie’s Choice*, New York: Bantam Books.

- Thomason, Richmond, 1981, “Deontic Logic as Founded on Tense Logic,” in Risto Hilpinen (ed.), *New Studies in Deontic Logic*, Dordrecht: Reidel, 165-176.
- Trigg, Roger, 1971, “Moral Conflict,” *Mind* **80** : 41-55.
- Vallentyne, Peter, 1987, “Prohibition Dilemmas and Deontic Logic,” *Logique et Analyse* **30** : 113-122.
- -----, 1989, “Two Types of Moral Dilemmas,” *Erkenntnis* **30** : 301-318.
- Van Fraassen, Bas, 1973, “Values and the Heart’s Command,” *The Journal of Philosophy* **70** : 5-19, [Reprinted in Gowans (1987): 138-153,]
- Williams, Bernard, 1965, “Ethical Consistency,” *Proceedings of the Aristotelian Society*, supp, **vol. 39** : 103-124, [Reprinted in Gowans (1987): 115-137,]
- Zimmerman, Michael J., 1988, *An Essay on Moral Responsibility*, Totowa, NJ: Rowman and Littlefield.
- -----, 1996, *The Concept of Moral Obligation*, New York: Cambridge University Press.

Other Worthwhile Readings

- Anderson, Lyle V., 1985, “Moral Dilemmas, Deliberation, and Choice,” *The Journal of Philosophy* **82** : 139-162,
- Atkinson, R.F., 1965, “Consistency in Ethics,” *Proceedings of the Aristotelian Society* supp, **vol. 39** : 125-138.
- Baumrin, Bernard H., and Peter Lupu, 1984, “A Common Occurrence: Conflicting Duties,” *Metaphilosophy* **15** : 77-90.
- Blackburn, Simon, 1996, “Dilemmas: Dithering, Plumping, and Grief,” in Mason (1996): 127-139.
- Bradley, F. H., 1927, *Ethical Studies*, 2nd edition, Oxford: Oxford University Press.
- Brink, David, 1989, *Moral Realism and the Foundations of Ethics*, New York: Cambridge University Press.
- -----, 1994, “Moral Conflict and Its Structure,” *The Philosophical Review* **103** : 215-247, [Reprinted in Mason (1996): 102-126,]
- Bronaugh, Richard, 1975, “Utilitarian Alternatives,” *Ethics* **85** : 175-178.
- Carey, Toni Vogel, 1985, “What Conflict of Duty is Not,” *Pacific Philosophical Quarterly* **66** : 204-215.
- Castañeda, Hector-Neri, 1974, *The Structure of Morality*, Springfield, IL: Charles C. Thomas.
- -----, 1978, “Conflicts of Duties and Morality,” *Philosophy and Phenomenological Research* **38** : 564-574.
- Chisholm, Roderick M., 1963, “Contrary-to-Duty Imperatives and Deontic Logic,” *Analysis* **24** : 33-36.
- Conee, Earl, 1982, “Against Moral Dilemmas,” *The Philosophical Review* **91** : 87-97, [Reprinted in Gowans (1987): 239-249,]
- -----, 1989, “Why Moral Dilemmas are Impossible,” *American Philosophical Quarterly* **26** : 133-141.
- Dahl, Norman O., 1974, “"’Ought’ Implies ‘Can’” and Deontic Logic,“ *Philosophia* **4** : 485-511.
- -----, 1996, “Morality, Moral Dilemmas, and Moral Requirements,” in Mason(1996): 86-101.
- DeCew, Judith Wagner, 1990, “Moral Conflicts and Ethical Relativism,” *Ethics* **101** : 27-41.

- Donagan, Alan, 1996, "Moral Dilemmas, Genuine and Spurious: A Comparative Anatomy," in Mason (1996): 11-22.
- Feldman, Fred, 1986, *Doing the Best We Can*, Dordrecht: D. Reidel Publishing Co.
- Foot, Philippa, 1983, "Moral Realism and Moral Dilemma," *The Journal of Philosophy* **80** : 379-398, [Reprinted in Gowans (1987): 271-290,]
- Gewirth, Alan, 1978, *Reason and Morality*, Chicago: University of Chicago Press.
- Goldman, Holly Smith, 1976, "Dated Rightness and Moral Imperfection," *The Philosophical Review* **85** : 449-487, [See also, *Holly Smith*,]
- Gowans, Christopher W., 1989, "Moral Dilemmas and Prescriptivism," *American Philosophical Quarterly* **26** : 187-197.
- -----, 1994, *Innocence Lost: An Examination of Inescapable Wrongdoing*, New York: Oxford University Press.
- -----, 1996, "Moral Theory, Moral Dilemmas, and Moral Responsibility," in Mason (1996): 199-215.
- Griffin, James, 1977, "Are There Incommensurable Values?" *Philosophy and Public Affairs* **7** : 39-59.
- Guttenplan, Samuel, 1979-80, "Moral Realism and Moral Dilemma," *Proceedings of the Aristotelian Society* **80** : 61-80.
- Hansson, Sven O., 1997, "Should We Avoid Moral Dilemmas?" *Uppsula Philosophical Studies* **46** : 78-89.
- Hare, R.M., 1952, *The Language of Morals*, Oxford: Oxford University Press.
- -----, 1963, *Freedom and Reason*, Oxford: Oxford University Press.
- -----, 1981, *Moral Thinking: Its Levels, Methods, and Point*, Oxford: Oxford University Press.
- Hill, Thomas E., Jr, 1983, "Moral Purity and the Lesser Evil," *The Monist* **66** : 213-232.
- -----, Jr, 1992, "A Kantian Perspective on Moral Rules," *Philosophical Perspectives* **6** : 285-304.
- -----, Jr, " Moral Dilemmas, Gaps, and Residues: A Kantian Perspective," in Mason (1996): 167-198.
- Hoag, Robert W., 1983, "Mill on Conflicting Moral Obligations," *Analysis* **43** : 49-54.
- Howard, Kenneth W., 1977, "Must Public Hands Be Dirty?" *The Journal of Value Inquiry* **11** : 29-40.
- Kant, Immanuel, 1965/1797, *The Metaphysical Elements of Justice: Part I of the Metaphysics of Morals*, Trans, John Ladd, Indianapolis: Bobbs-Merrill.
- Kolenda, Konstantin, 1975, "Moral Conflict and Universalizability," *Philosophy* **50** : 460-465.
- Ladd, John, 1958, "Remarks on Conflict of Obligations," *The Journal of Philosophy* **55** : 811-819.
- Lebus, Bruce, 1990, "Moral Dilemmas: Why They Are Hard to Solve," *Philosophical Investigations* **13** : 110-125.
- MacIntyre, Alasdair, 1990, "Moral Dilemmas," *Philosophical and Phenomenological Research* **50** : 367-382.
- Mallock, David, 1967, "Moral Dilemmas and Moral Failure," *Australasian Journal of Philosophy* **45** : 159-178,
- Mann, William E., 1991, "Jephthah's Plight: Moral Dilemmas and Theism," *Philosophical Perspectives* **5** : 617-647.
- Marcus, Ruth Barcan, 1996, "More about Moral Dilemmas," in Mason (1996): 23-35.

- Marino, Patricia, 2001, “Moral Dilemmas, Collective Responsibility, and Moral Progress,” *Philosophical Studies* **104** : 203-225.
- Mason, H.E., 1996, “Responsibilities and Principles: Reflections on the Sources of Moral Dilemmas,” in Mason (1996): 216-235.
- McConnell, Terrance, 1976, “Moral Dilemmas and Requiring the Impossible,” *Philosophical Studies* **29** : 409-413.
- -----, 1981, “Moral Absolutism and the Problem of Hard Cases,” *Journal of Religious Ethics* **9** : 286-297.
- -----, 1981, “Moral Blackmail,” *Ethics* **91** : 544-567.
- -----, 1981, “Utilitarianism and Conflict Resolution,” *Logique et Analyse* **24** : 245-257.
- -----, 1986, “More on Moral Dilemmas,” *The Journal of Philosophy* **82** : 345-351.
- -----, 1993, “Dilemmas and Incommensurateness,” *The Journal of Value Inquiry* **27** : 247-252.
- McDonald, Julie M., 1995, “The Presumption in Favor of Requirement Conflicts,” *Journal of Social Philosophy* **26** : 49-58.
- Mothersill, Mary, 1996, “The Moral Dilemmas Debate,” in Mason (1996): 66-85.
- Nagel, Thomas, “War and Massacre,” *Philosophy and Public Affairs* **1** : 123-144.
- -----, 1979, “The Fragmentation of Value,” in *Mortal Questions*, New York: Cambridge University Press, [Reprinted in Gowans (1987): 174-187,]
- Nozick, Robert, 1968, “Moral Complications and Moral Structures,” *Natural Law Forum* **13** : 1-50.
- Paske, Gerald H., 1990, “Genuine Moral Dilemmas and the Containment of Incoherence,” *The Journal of Value Inquiry* **24** : 315-323.
- Pietroski, Paul, 1993, “Prima Facie Obligations, Ceteris Paribus Laws in Moral Theory,” *Ethics* **103** : 489-515.
- Price, Richard, 1974/1787, *A Review of the Principal Questions of Morals*, Oxford: Oxford University Press.
- Prior, A.N., 1954, “The Paradoxes of Derived Obligation,” *Mind* **63** : 64-65.
- Quinn, Philip, 1978, *Divine Commands and Moral Requirements*, New York: Oxford University Press.
- -----, 1986, “Moral Obligation, Religious Demand, and Practical Conflict,” in Robert Audi and William Wainwright (eds.), *Rationality, Religious Belief, and Moral Commitment*, Ithaca, NY: Cornell University Press, 195-212.
- Rabinowicz, Wlodzimierz, 1978, “Utilitarianism and Conflicting Obligations,” *Theoria* **44** : 1924.
- Rawls, John, 1971, *A Theory of Justice*, Cambridge: Harvard University Press.
- Railton, Peter, 1992, “Pluralism, Determinacy, and Dilemma,” *Ethics* **102** : 720-742.
- -----, 1996, “The Diversity of Moral Dilemma,” in Mason (1996): 140-166.
- Santurri, Edmund N., 1987, *Perplexity in the Moral Life: Philosophical and Theological Considerations*, Charlottesville, VA: University of Virginia Press.
- Sartorius, Rolf, 1975, *Individual Conduct and Social Norms: A Utilitarian Account of Social Union and the Rule of Law*, Encino, CA: Dickenson Publishing.
- Sayre-McCord, Geoffrey, 1986, “Deontic Logic and the Priority of Moral Theory,” *Nous* **20** : 179-197.
- Sinnott-Armstrong, Walter, 1984, “’Ought’ Conversationally Implies ‘Can’,” *The Philosophical*

Review **93** : 249-261.
- -----, 1985, “Moral Dilemmas and Incomparability,” *American Philosophical Quarterly* **22** : 321-329.
- -----, 1987, “Moral Dilemmas and ‘Ought and Ought Not’,” *Canadian Journal of Philosophy* **17** : 127-139.
- -----, 1987, “Moral Realisms and Moral Dilemmas,” *The Journal of Philosophy* **84** : 263-276.
- -----, 1996, “Moral Dilemmas and Rights,” in Mason (1996): 48-65.
- Slote, Michael, 1985, “Utilitarianism, Moral Dilemmas, and Moral Cost,” *American Philosophical Quarterly* **22** : 161-168.
- Statman, Daniel, 1996, “Hard Cases and Moral Dilemmas,” *Law and Philosophy* **15** : 117-148.
- Steiner, Hillel, 1973, “Moral Conflict and Prescriptivism,” *Mind* **82** : 586-591.
- Stocker, Michael, 1971, “’Ought’ and ‘Can’,” *Australasian Journal of Philosophy* **49** : 303-316.
- -----, 1986, “Dirty Hands and Conflicts of Values and of Desires in Aristotle’s Ethics,” *Pacific Philosophical Quarterly* **67** : 36-61.
- -----, 1987, “Moral Conflicts: What They Are and What They Show,” *Pacific Philosophical Quarterly* **68** : 104-123.
- -----, 1990, *Plural and Conflicting Values*, New York: Oxford University Press.
- Strasser, Mark, 1987, “Guilt, Regret, and Prima Facie Duties,” *The Southern Journal of Philosophy* **25** : 133-146.
- Swank, Casey, 1985, “Reasons, Dilemmas, and the Logic of ‘Ought’,” *Analysis* **45** : 111-116.
- Tannsjo, Torbjorn, 1985, “Moral Conflict and Moral Realism,” *The Journal of Philosophy* **82** : 113-117.
- Thomason, Richmond, 1981, “Deontic Logic and the Role of Freedom in Moral Deliberation,” in Risto Hilpinen (ed.), *New Studies in Deontic Logic*, Dordrecht: Reidel, 177-186.
- -----Vallentyne, Peter, 1992, “Moral Dilemmas and Comparative Conceptions of Morality,” *The Southern Journal of Philosophy* **30** : 117-124.
- Walzer, Michael, 1972, “Political Action: The Problem of Dirty Hands,” *Philosophy and Public Affairs* **2** : 160-180.
- Williams, Bernard, 1966, “Consistency and Realism,” *Proceedings of the Aristotelian Society*, supp, **vol. 40** : 1-22.
- -----, 1972, *Morality: An Introduction to Ethics*, New York: Harper & Row.
- Zimmerman, Michael J., 1987, “Remote Obligation,” *American Philosophical Quarterly* **24** : 199-205.
- -----, 1988, “Lapses and Dilemmas,” *Philosophical Papers* **17** : 103-112.
- -----, 1990, “Where Did I Go Wrong?” *Philosophical Studies* **58** : 83-106.
- -----, 1992, “Cooperation and Doing the Best One Can,” *Philosophical Studies* **65** : 283-304.
- -----, 1995, “Prima Facie Obligation and Doing the Best One Can,” *Philosophical Studies* **78** : 87-123.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Kant, Immanuel | logic: deontic | Mill, John Stuart | Plato | Sartre, Jean-Paul

Terrance McConnell
tcmcconn@uncg.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 15, 2002
Content last modified: April 15, 2002

Finitism in Geometry

In our representations of the world, especially in physics, infinities play a crucial role. The continuum of the real numbers as a representation of time or one-dimensional space is the best known example. However, these same infinities also cause problems. One just has to think about Zeno's paradoxes or the present-day continuation of that discussion, namely, the discussion about supertasks, to see the difficulties. Hence, it is a very tempting idea to investigate whether it is possible to eliminate these infinities and still be able to do physics. This problem reduces first of all to the question of the possibility of a discrete geometry that can approximate classical infinite geometry as closely as possible. If a positive answer can be given to this question, the second question is what could be the possible physical relevance (if any).

---

1. What is Finitism in Geometry?

1.1 Finitism in mathematics

Finitism is one of the foundational views of mathematics that is listed under the broader heading of constructivism. It shares with the many forms of constructivism the fundamental view that mathematical

objects and concepts have to be accessible to the mathematician in terms of constructions that can be executed or performed. The various forms are distinguished from one another as to how ‘execution’ or ‘performance’ is to be understood. Usually outside finitism the potentially infinite is allowed, i.e., if a procedure or algorithm will (provably) terminate at some moment in the future, then the outcome is accepted as constructable. See Bridges & Richman [1997] for an overview. However, finitism goes one step further and argues that an indefinite outcome is not be accepted as an outcome, since, as all computational resources are finite, it could very well be that these resources have been used up before the outcome has been reached.

Usually the label *strict* finitism is used to describe the view sketched above. The additional qualification serves to make the distinction with Hilbert’s finitism which, roughly speaking, can be seen as a form of finitism on the meta-level (e.g., although mathematical theories can talk about infinite structures, still the proofs in such theories must have a finite length). Here the discussion will be limited to strict finitism.

As might be expected, strict finitism is not a popular view in the philosophy of mathematics. Nevertheless a number of proposals have been put forward. A history and an account of the actual (though now somewhat dated) state of affairs can be found in Welti [1987].

1.2 The special case of geometry

It is rather striking that the various proposals that are discussed in Welti’s book focus mainly on numbers and numerals. Less attention has been paid to the problem of geometry. Nevertheless the basic premise seems easy enough to formulate, namely that a geometrical space is finitely extended and that it is composed of a finite number of geometrical basic units, that themselves are not decomposable. These basic units are sometimes called ‘(geometrical) atoms’ or ‘hodons’.

Formulated thus, the problem might seem trivial. After all, there is a separate branch of mathematics that studies finite geometries. There is however an important difference. Although such geometries can be very inspiring for a strict finitist proposal (as will be shown in section 2), their aim is not to provide an *alternative* for continuous infinite geometries. In addition for the strict finitist geometer, success or failure is not solely determined by *internal* mathematical arguments or considerations. Rather the hope is that such alternatives will prove to be relevant for applications to the sciences, the case *par excellence* being physics. If one could successfully replace the classical geometrical continuum structure of space-time by a strict finitist analogue, then, from a philosophical point of view, the consequences would be of great importance. Two obvious topics present themselves as test cases: Zeno’s paradoxes and supertasks. If space and time are discrete, then the runner, the tortoise, Achilles and all other moving objects simply go through a finite number of space locations in a finite number of time elements and all the problems with supertasks vanish as a one-minute time interval is no longer divisible in a denumerable series of intervals. These topics will not be treated here and the reader is referred to the related entries.

In the very same sense, there is a clear difference with all the theories and proposals that have been put forward in the computer sciences. The problem one faces there is precisely to set up a translation from a

classical geometrical model to a model whereof the domain (usually) consists of the finite set of pixels or cells that make up the computer screen. Although equally inspiring, the drawback here is that nearly all these models assume the classical (infinite) model in the background and, hence, do not have a proper foundation of their own (a situation quite analogous to numerical analysis that relies on classical analysis for proving the correctness of the procedures). Note also that these theories should not be confused with computer programs that have the ability to reason *about* geometrical objects. This is part of the research area of automated reasoning and its basic objects are proofs, not necessarily the mathematical objects the proofs are about.

Finally, although there have been numerous proposals for discrete worldviews throughout the history of western culture, see, e.g., White [1992], the focus here is on the twentieth-century developments.

2. A Classic Proposal for a Discrete Geometry

As most of the work that has been done has been limited to the plane, this presentation will also be restricted to that particular case (in most proposals the extension to higher dimensional geometries is considered completely straightforward). There are different routes to follow. One possibility is to take any axiomatisation of the (Euclidean) plane—say, Hilbert's formulation of 1899 in his *Grundlagen der Geometrie*—and show what changes are required to have (a) finite models of the axiomatic theory, and (b) finite models that approximate the classical (infinite, Euclidean) models as closely as possible. One of the very first attempts dates back to the late 40s, early 50s and will therefore be presented here as an exemplar (in the sense that it has both all the positive qualities required as well as the oddities that seem to go together with such attempts). More specifically, it concerns the work of Paul Kustaanheimo in partial collaboration with G. Järnefelt in the period between 1949 and 1957.

2.1 A standard axiomatisation for Euclidean plane geometry

What does a Hilbert-type axiomatisation look like? The first thing one has to do is to fix a (formal) language. Usually one chooses first-order predicate logic with identity, i.e., a language containing names for variables (and, possibly, for constants), names for functions (if necessary), names for predicates including the identity predicate, logical connectives and quantifiers, and a set of grammatical rules to form sentences. The restriction to first-order logic means that only variables can be quantified over. Without going into details, it should be remarked that a more expressive language can be chosen, e.g., whereby quantification over predicates is allowed as well.

If a language has been chosen, the next problem is to determine the primitive terms of the language. For plane Euclidean geometry, these are points and lines, although sometimes lines are defined as particular sets of points. Next the basic predicates have to be selected. There exist a number of different axiomatisations at the present moment. The most frequently used predicates are: the incidence relation ("a point *a* lies on a line *A*"), the betweenness relation ("point *a* lies between points *b* and *c*"), the equidistance relation ("the distance from point *a* to *b* is the same as the distance from point *c* to *d*"), the

congruence relation ("a part of a line, determined by two point *a* and *b* is congruent to a part of a line, determined by two points *c* and *d*"). Note that it is not necessary that all of them occur in an axiomatisation. As an example, if lines are not introduced as primitive terms, then usually there is no incidence relation.

The next step is the introduction of a set of axioms to determine certain properties of the above mentioned relations. As an example, if the axiomatisation uses the incidence relation, then the typical axioms for that relation are:

a. Through two points exactly one straight line can be drawn.
b. There are at least two points on every straight line.
c. There are at least three points that are not on the same straight line.

Finally, one looks for an interpretation or a model of the axiomatisation. This means that we search for a meaning of the primitive terms, such as points and lines, of the functions (if any) and of the predicates in such a way that the axioms become true statements relative to the interpretation. Although we often have a particular interpretation in mind when we develop an axiomatisation, it does not exclude the possibility of the existence of rather unexpected models. In a sense finitist models rely on this very possibility as the next paragraph shows.

2.2 Kustaanheimo's finitist approach

Kustaanheimo's proposal—I reproduce here in rough outline the excellent presentation of his proposal in Welti [1987], pp. 487-521, which is far more accessible than the original work—is based on the following line of reasoning. A standard model for the classical axiomatic theory of Euclidean geometry consists of the cartesian product of the real numbers with itself. Or, as it is usually formulated, a point in the plane is mapped onto a couple of real numbers, its coordinates. The real numbers have the mathematical structure of an infinite field. But finite fields exist as well. So why not replace the infinite real number field with a finite field, a so-called Galois field?

The best result one could obtain would be that every finite Galois field satisfies most of the axioms of Euclidean geometry. That however is not the case. The outcome of Kustaanheimo's research is slightly more complicated:

> (a) Not all finite fields will do. If we call *p* the number of elements in the domain of the finite field, then *p* has to satisfy some conditions. This means that only finite fields of a particular size, i.e., a specific value for *p*, are potential candidates.
>
> (b) For the "good" values of *p*, the full model will not do. As an example, take straight lines. According to their definition in a finite field, it turns out that there are two sorts of straight lines: open and closed ones. The latter ones violate some of the axioms, hence you restrict the model to the open ones. This restriction of the model is called the Euclidean

"kernel" of the model.

In short, one cannot claim that any finite field will do, but only some and for that matter only part of it.

This approach raises some important philosophical questions:

(a) It is clear that the size of the model is an important feature. Does this have any meaning? Or, negatively, what does it mean that fields of a different size are not suitable as models? Suppose as a thought experiment that Euclidean geometry is a good model for the geometrical structure of the universe. Does it make sense to claim that the universe must contain exactly p points (not p-1, not p+1)? A new sort of Pythagoreanism seems to be lurking around the corner here.

(b) The example of the straight lines shows that there are "nice" geometrical objects (those that satisfy most of the axioms) and "bad" geometrical objects. *Ignoring* the bad ones is perhaps a mathematically interesting strategy, but it does not *eliminate* them from the full model. In other words, although they do not play any relevant part in the "kernel" of the model, they are there. What is the meaning of that? To continue the above thought experiment, the question is what corresponds to the "bad" objects in the universe? If they do not correspond to anything, why do we need them in the first place to find the "good" objects?

In defense of Kustaanheimo's approach it must be said that the connections between infinite and finite models are usually far more complex than one expects. A finite model is not merely a scaled-down version of an infinite model. Very often a different structure appears. As an analogy take the (infinite set of) natural numbers. Take a finite part, say the numbers 1 up to L. In the finite case it makes sense to talk about small and large numbers compared to L. This is classically not possible. So one finds additional structure. Metaphorically speaking, by making things finite, a more detailed or "fine-grained" structure appears, that is wiped out in the presence of infinities. Perhaps the distinction between "good" and "bad" geometrical objects is such an additional feature that disappears in the classical Euclidean model. More details about Kustaanheimo's approach are to be found in the following supplementary document

Finite Fields as Models for Euclidean Plane Geometry.

The counterargument could be that one of the major disadvantages of infinite models and theories is that 'fine' details or the 'fine-grained' structure is wiped out by the infinities present, much the same as infinite numbers wipe out the distinction between small and large numbers or numerals. Thus perhaps the prime numbers do have a significance. But still the question remains: is this a new sort of Pythagoreanism?

3. Recent Proposals

In recent years some authors have focused rather on particular problems related to strict finitist geometry instead of attempting to present a full alternative. Usually the problems have been raised by those who doubt the possibility of a strict finitist position. Three problems of this sort will be discussed: the distance function problem, the dimension problem and the isotropy problem.

3.1 Three problems to deal with

The distance function problem. The argument dates back to 1949 and was first formulated by Hermann Weyl: "If a square is built up of miniature tiles, then there are as many tiles along the diagonal as there are along the sides; thus the diagonal should be equal in length to the side" (Weyl [1949], p. 43). At least two solutions to this problem have been formulated.

Van Bendegem [1987] argued that in a finite geometry it should be a basic fact that lines and points have extensions. In particular, lines are supposed to have a constant width (independent of the orientation of the line) N_D. Thus N_D represents a large (finite) number, corresponding to the number of squares that form N_D. Given a line, the width is always defined as perpendicular to that line. Now suppose that the line has an orientation corresponding to an angle α between the line and the *x*-axis. Then the width N_D of that line, when projected on the *x*-axis will be $[N_D/\sin\alpha]$ where the expression $[x]$ indicates the greatest integer less than or equal to *x*.

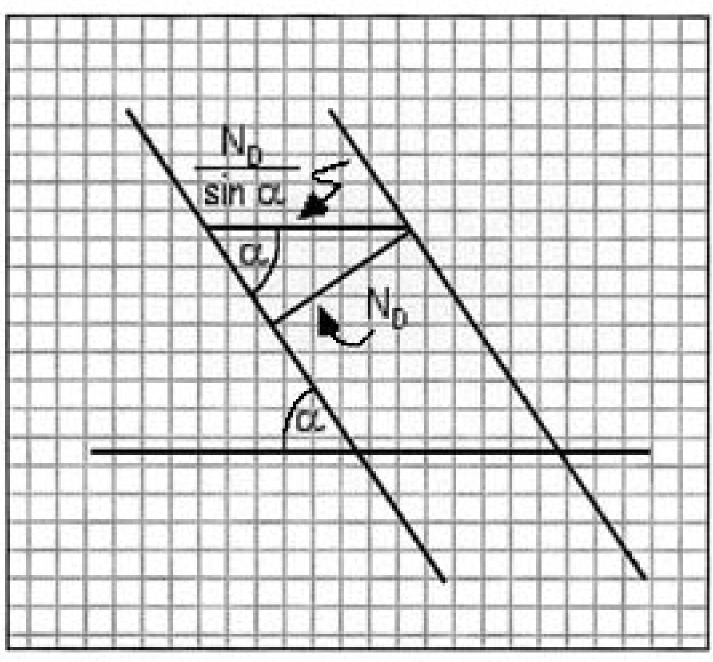


Figure 1

The distance *d* between two points *p* and *q* is then defined as the number of squares in the rectangle formed by the line from *p* to *q* and the width N_D, divided by N_D. The idea is that, although in a discrete geometry, lines must necessarily have a width, this is not an essential feature, so it can be divided out. Hence:

$d(p,q) = N_L\ .[\ N_D/\sin\ \alpha]\ (\text{div}\ N_D).$

N_L here corresponds to the number of layers parallel to the *x*-axis between *p* and *q* and *n* (div *m*) is the quotient of the division of *n* by *m*.

As an illustration, consider the Weyl problem.

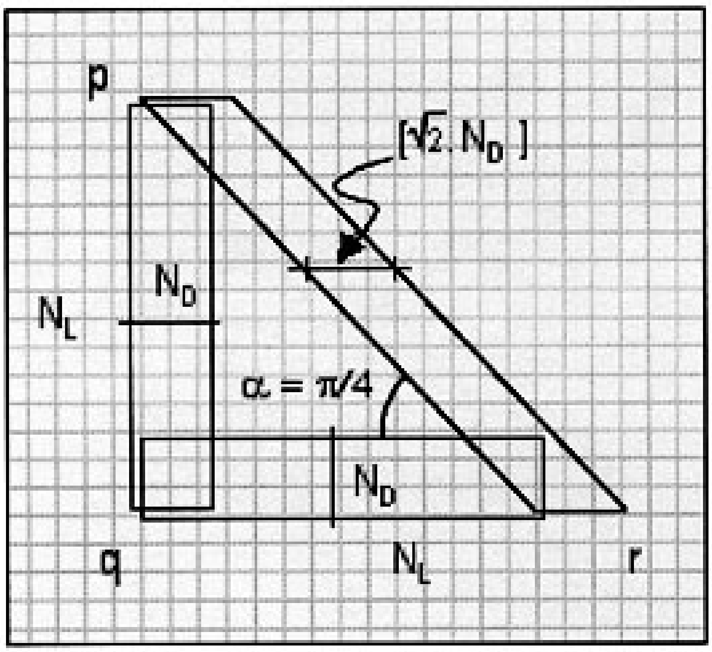


Figure 2

We have a right-angled triangle *pqr* such that for simplicity the right sides *pq* and *qr* are equal to one another and are aligned with the axes of the grid. Suppose that the number of squares in the right sides is N_L. Then

$d(p,q)= d(q,r) = N_L\ .[\ N_D]\ (\text{div}\ N_D) = N_L,$
since, of course, $[N_D] = N_D$.

However, the hypotenuse has an angle of $\alpha = \sqrt{2}/2$. Thus

$d(p,r) = N_L\ .[\ N_D/\sin\ \alpha]\ (\text{div}\ N_D)$

$= N_L.[\sqrt{2}\ .\ N_D]\ (\text{div}\ N_D)$

$= N_L.\ [\sqrt{2}]_n,$

where $[r]_n$ means the number *r* up to *n* decimals

No calculations are needed to show that the Pythagorean theorem holds, i.e., $d^2(p,q) + d^2(q,r) = d^2(p,r)$. Finally, there is an easy explanation why the Weyl problem occurs: it corresponds to the limiting case $N_D = 1$. When $N_D = 1$, then $[\sqrt{2 \cdot N_D}] = [\sqrt{2}] = 1$, hence $d(p,r) = N_L.1 = N_L$ and Pythagoras' theorem fails.

Although the introduction of a width N_D apparently solves the problem, it is equally clear what the drawbacks are. Without classical Euclidean geometry in the background, there is really no way to get the construction going. There is no definition of a line in terms of the discrete geometry, and, above all, the projected width on the *x*-axis of a line *L* is calculated according to a Euclidean distance function that is not explicitly mentioned. In short, there is a mixture of two distance functions.

Peter Forrest [1995] presents another solution. He starts by introducing a family of discrete spaces $E_{n,m}$, where *n* corresponds to the "classical" dimension of space and *m* is a scale factor, to be understood as follows: *m* is a parameter to decide when two points are or are not adjacent, which is the basic (and sole) concept of his geometry. Thus in the case $n = 2$, points are labeled by couples of integers (i, j) and two points (i, j) and (i', j') are adjacent if

a. they are distinct, and
b. $(i - i')^2 + (j - j')^2 \leq m^2$.

Once adjacency has been stipulated, a distance function can be easily derived: the distance between *p* and *q*, $d(p,q)$, is the smallest number of "links" in a chain of points connecting *p* and *q* such that each one is adjacent to the previous one. Next there is no problem to show that a straight line passing through two points is that chain of points that has the shortest distance.

If the parameter m has a small value, then the resulting distance function is not Euclidean. More specifically, if $m = 1$, then we have, once again, the situation presented by Weyl. But if, say, $m = 10^{30}$ (the figure proposed by Forrest himself), then the situation changes. Then it is possible to show that the distance function on the discrete space will approximate the Euclidean distance function as close as one wants. Without presenting all the details, one can show that a Euclidean distance function d_E and the discrete distance function *d* are related by a scale factor, i.e.,

$$d_E(p,q)/d(p, q) = \text{constant}(m),$$

where the constant is determined by the value of *m*. No calculations are needed once again, to show that the original distance function *d* satisfies the Pythagorean theorem.

If one is looking for a weak point in this approach, then inevitably one must end up with the basic notion of adjacency. What is the reason for defining adjacency in Euclidean terms? For, after all, a condition such as $(i - i')^2 + (j - j')^2 \leq m^2$ looks as Euclidean as can be.

A possible way out is suggested in Van Bendegem [1995]. One of the advantages of a discrete approach—and, as a matter of fact, this seems to hold in general for strict finitist proposals—is that definitions that are classically equivalent turn out to be distinct in a strict finitist framework. Thus, more specifically, a circle can be defined in (at least) two ways:

a. as the set of points *p* that have a fixed distance to a fixed point,
b. as the set of points *p* such that, given a fixed line segment *ab*, the angle formed by *apb* is a right angle.

Classically speaking, these two definitions are equivalent. However, they are not in a discrete geometry. If, e.g., the distance function is defined as the lowest number of hodons that connect two given points, then the two definitions are not equivalent. Using definition (a), the circle will have the shape of a square (a well-known fact in so-called taxicab geometry) and thus useless to define adjacency as done above. Definition (b) on the other hand produces a figure that can approximate a Euclidean circle as close as one likes. In that way Forrest's definition for adjacency is acceptable within a discrete framework, as no reference is made to a Euclidean distance function.

The dimension problem. Not much attention has been paid to this problem although it is fundamental. If the plane consists of a discrete set of elements, hodons or atoms, then this set must have dimension zero. For, in order to determine the dimension, this set must be equipped with a topology and the only possible candidate is the discrete topology. This entails that the dimension is zero. Either one takes the option to simply drop the notion of dimension on the basis of the argument that the concept of dimension presupposes a concept of continuity and topology and hence has no finitist meaning. Or one searches for an analog, but it is not clear at all what that could be. Something one should not try to do is to derive a notion of dimension from an ordering relation. Suppose that the hodons are labeled by integers (i, j) in some appropriate coordinate system, such that $-L \leq i, j \leq L$, where L is some upper bound. Then quite different ordering relations are possible. One possibility is to define $(i, j) < (k, l)$ if and only if $i + j < k + l$. But another possibility is to define $(i, j) < (k, l)$ if and only if either $i < k$ or, if $i = k$, then $j < l$. It therefore requires additional arguments to claim that among all possible order relations on a given set, one and only one has a special status. So far no such arguments have been produced.

The isotropy problem. If the plane is built up from square hodons, as in the paragraph above, then the hodons are arranged in such a way that every hodon touches four other hodons, i.e., the plane can be modeled as a square grid, then it is obvious that there are preferred directions, in this case, there will be two preferred directions. However if instead of squares, hexagons are taken as hodons, then there are four preferred directions. Thus no matter what the shape is of the hodon, there will be preferred directions and this implies that the space is anisotropic. However, for physical applications one would like to have isotropy (or at least as close as possible).

Two approaches are possible. Either the hodons have a definite shape or they do not. In the first case it has been suggested that instead of a regular periodic tiling of the plane, one should look for an irregular aperiodic tiling, such as the Penrose tiling.

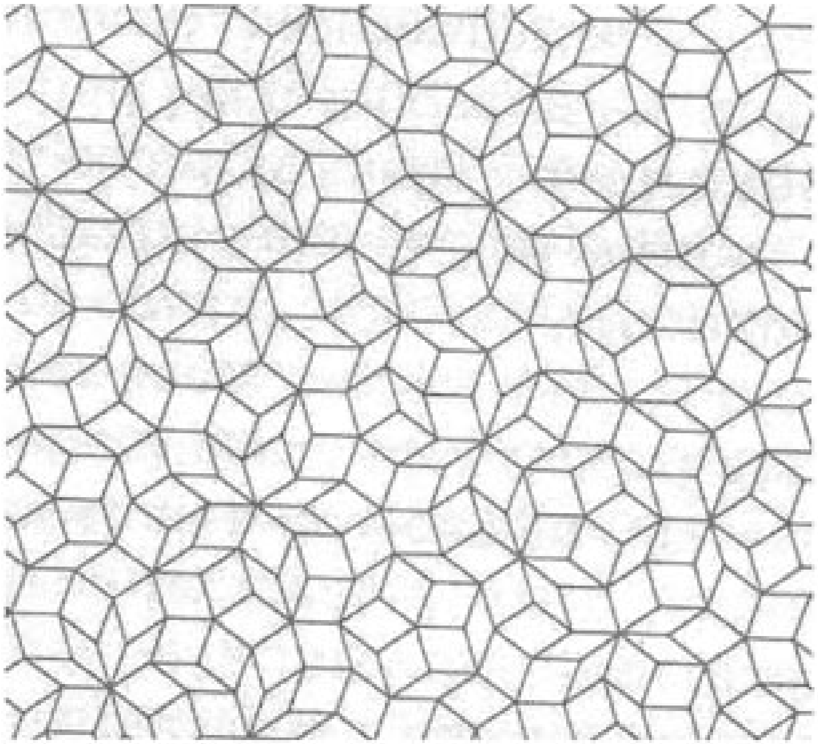


Figure 3

Although no worked-out examples are available, it seems a promising line of attack (see, e.g., Penrose Tilings and Wange Tilings, in the Other Internet Resources section), as a source of inspiration). In the case of the Penrose tiling it is interesting to see that there are no classically straight lines anymore, precisely because of the aperiodicity. In the second case vagueness is a possible way out. As Peter Forrest defends in his [1995], the whole idea of a specific representation of a discrete space, e.g., as built up from tiny squares is fundamentally mistaken. If a hodon has a specific form, then one cannot avoid asking questions about parts of a hodon, such as its border, but that does not make sense if hodons are the smallest spatial entities possible. An intermediate position defended in Van Bendegem [1995] is to consider a series of discrete geometries G_i, each with a hodon of a particular size, h_i, such that $h_i \neq h_j$, for $i \neq j$ and, in addition, there are M and N such that $M < h_i < N$, for all i. One can then apply a supervaluation technique to the series. This means that a statement will be True (False) if it is true (false) in every geometry G_i. In all other cases it is Undecided, i.e. true in some and false in some others. Now if A is the statement ‘hodons have size α’ (where α is a specific number), this will be Undecided, if a corresponds to at least one of the h_i. Such an approach however introduces all problems connected with vagueness into the discussion, which is not necessarily an encouraging situation.

3.2 The empirical question

As said in the beginning, the final test of a discrete geometry will be its application first and foremost to physics. There are basically two ways to think about such a test. Either one can think of a ‘strong’ version, viz., that is possible to design an experiment that would allow us to decide between the

continuous and the discrete case. Or one can think of a ‘weak’ version, viz., that the discrete version does at least the same things as the classical version.

In the first case the aim is to think of a crucial experiment. It is quite interesting to see that already in 1961 Paul Feyerabend suggested such a possibility. However not much more is said, as “the difficulty of the present situation seems to lie in the fact that discrete alternatives for the mathematics, which at the present moment is being used in physics, are missing” (p. 160). Equally interesting is the fact that Feyerabend too mentions the standard argument that the lack of a Pythagorean theorem is a genuine problem. His proposal is that “we need only assume that measurements in different directions do not commute; and then perhaps we can retain the theorem as an operator equation” (p. 161). Here too, unfortunately, nothing more is said. Peter Forrest [1995] maintains that such an experiment is possible. The fundamental reason is that classical mathematics uses continuous variables whereas strict finitist mathematics uses discrete variables. Thus for differentation and integration finite analogs must be found and they will approximate the classical case, but never coincide with it. Hence, there will always be small differences and it cannot be excluded that these could be detectable. Although not much progress has been made in this connection, it is worth mentioning the curious fact that the differential equation, $df/dx = ax(1 - x)$, produces a very neat continuous solution, whereas the analogue difference equation, $\Delta f/\Delta x = ax(1 - x)$, depending on the value of the parameter a, produces chaotic effects. See Van Bendegem [2000] and Welti [1987], pp. 516-518.

In the second case there is a strong tendency to focus on the problem whether elements from the mathematical model can be either directly or indirectly linked to the physical constants. All things considered it seems not very interesting to look for direct connections. Suppose that one would be tempted to identify the hodon with the Planck length, $l_p = 10^{-35}\ m$ and the chronon—the smallest time unit—with Planck time, $t_p = 10^{-43}\ s$, then, if it is accepted that an object can only move through neighbouring hodons, then there is a upper limit to velocities, namely $c = 3.10^8 m/s$. But if an object is not moving at the top speed of one hodon per chronon, then the next lower value must be one hodon per two chronons, but that means a velocity of $c/2$. We seem to have missed out the whole range between $c/2$ and c. There is a way out, but it supposes that ‘jerky’ motion is considered possible, an aesthetically quite ugly idea. An object moves two hodons in two chronons and then waits for one chronon and then repeats the same motion. The average velocity is then $2c/3$.

Nevertheless some philosophers and physicists seem especially interested in this kind of approach, see, e.g., Hahn [1934], Biser [1941], Coish [1959], Finkelstein & Rodriguez [1986], Meessen [1989], Buot [1989], to name but a very few. For the period 1925-1936, Kragh and Carazza [1994] is an excellent overview showing that many physicists were playing around with finitist ideas. Two types of arguments and considerations are often expressed:

The quantum mechanical connection. A strict finitist geometry makes sense because it is clear in quantum mechanics that discreteness is an essential feature of the world and hence should be reflected in our descriptions of that world. Although at first sight the argument seems plausible, it seriously underestimates the difficulty of the task. It is therefore not very surprising that in nearly all cases only

rough suggestions are made and it is rather unclear what mathematical formalism could replace the existing mathematics used in quantum mechanics involving lots of infinities (such as infinite dimensional Hilbert spaces).

The numbers game. In this approach the focus is on the physical constants themselves including such constants as the velocity of light *c*, the Planck constant *h*, the mass of the electron m_e, and so on. As these values are necessarily finite, it seems worthwhile to investigate whether a finitist approach can explain why these constants have the values they happen to have. A very fine example of such an approach is the so-called *Combinatorial Physics Program*. Starting from a very simple model, combinatorial considerations on the size of certain mathematical objects lead to an estimation of some physical constants. The success of this program is rather modest as these models do not connect easily to existing physical theories. A self-contained presentation of this program is to be found in Bastin & Kilmister [1995]. There is a very strong similarity with the work of A. S. Eddington. Not very surprisingly, a presentation of the work of Eddington on his fundamental theory has been written by Kilmister [1994].

All things considered, both on the mathematical as on the physical level an enormous amount of work remains to be done. Probably the best way to characterize the present-day situation is that some 'famous' objections to a finitist approach in geometry have been (partially) answered so that it is legitimate to continue this research program.

Bibliography

- Bastin, T. & Kilmister, C.W., 1995, *Combinatorial Physics*, Singapore: World Scientific.
- Biser, E., 1941, 'Discrete Real Space', *Journal of Philosophy*, 38 (Summer), pp. 518-524
- Bridges, D. & Richman, F., 1987, *Varieties of Constructive Mathematics*. Cambridge: Cambridge University Press (LMS Lecture Notes Series 97).
- Buot, F.A., 1989, 'Discrete Phase-Space Model for Quantum Mechanics', in M. Kafatos (ed.), *Bell's Theorem, Quantum Theory and Conceptions of the Universe*. Dordrecht: Kluwer, pp. 159-162.
- Coish, H.R., 1959, 'Elementary particles in a finite world geometry', *Physical Review*, vol. 114, pp. 383-388.
- Engeler, E., 1983, *Foundations of Mathematics. Questions of Analysis, Geometry, and Algorithmics*. Heidelberg: Springer-Verlag.
- Feyerabend, P., 1961, 'Comments on Grünbaum's "Law and Convention in Physical Theory"', in H. Feigl & G. Maxwell (eds.), *Current Issues in the Philosophy of Science*. New York: Holt, Rinehart and Winston, pp. 155-161.
- Finkelstein, D. & Rodriguez, E., 1986, 'Quantum time-space and gravity', in R. Penrose & C.J. Isham (eds.), *Quantum Concepts in Space and Time*. Oxford: Oxford University Press, pp. 247-254.
- Forrest, P., 1995, 'Is Space-Time Discrete or Continuous? -- An Empirical Question', *Synthese*, 103, pp. 327-354.
- Hahn, H., 1980, 'Does the infinite exist?', in B. Mcguinness (ed.), *Hans Hahn: Empiricism,*

Logic, and Mathematics. Dordrecht: Reidel, pp. 103-131 (originally published in 1934).

- Järnefelt, G., 1951, 'Reflections on a Finite Approximation to Euclidean Geometry: Physical and Astronomical Prospects', *Annales Academiae Scientiarum Fennicae, Series A, I. Mathematica-Physica*, 96, pp. 1-43.
- Kilmister, C.W., 1995, *Eddington's Search for a Fundamental Theory: A Key to the Universe*. Cambridge: Cambridge University Press.
- Kragh, H. & Carazza, B., 1994, 'From Time Atoms to Space-Time Quantization: the Idea of Discrete Time, ca 1925-1936', *Studies in the History and the Philosophy of Science*, vol. 25/3, pp. 437-462.
- Kustaanheimo, P., 1951, 'A Note on a Finite Approximation of the Euclidean Plane Geometry', *Societas Scientiarum Fennica. Commentationes Physico-Mathematicae*, 15/19, pp. 1-11.
- Meessen, A., 1989, 'Is it logically possible to generalize physics through space-time quantization?' in P. Weingartner & G. Schurz (eds.), *Philosophie der Naturwissenschaften. Akten des 13. Internationalen Wittgensteins Symposium*. Vienna: Hölder-Pichler-Tempsky, pp. 19-47.
- Van Bendegem, J.-P., 1987, 'Zeno's Paradoxes and the Weyl Tile Argument', *Philosophy of Science*, 54/2, pp. 295-302.
- -----, 1997, 'In defence of discrete space and time', *Logique et Analyse*, 38/150-152, pp. 127-150.
- -----, 2000, 'How to tell the continuous from the discrete', in François Beets & Eric Gillet (eds.), *Logique en Perspective. Mélanges offerts à Paul Gochet.* Brussels: Ousia, 2000, pp. 501-511.
- Welti, E., 1987, *Die Philosophie des strikten Finitismus. Entwicklungstheoretische und mathematische Untersuchungen über Unendlichkeitsbegriffe in Ideengeschichte und heutiger Mathematik*, Bern: Peter Lang.
- Weyl, H., 1949, *Philosophy of Mathematics and Natural Sciences*, Princeton: Princeton University Press.
- White, M.J., 1992, *The Continuous and the Discrete. Ancient Physical Theories from a Contemporary Perspective*. Oxford: Clarendon Press.

Other Internet Resources

- Penrose Tilings and Wang Tilings (maintained by Frank D. (Tony) Smith)

Related Entries

geometry: in the 19th century | space and time: supertasks | vagueness | Zeno's paradoxes

Jean Paul Van Bendegem
jpvbende@vub.ac.be

Table of Contents

First published: April 3, 2002
Content last modified: April 3, 2002

Stanford Encyclopedia of Philosophy
Supplement to Finitism in Geometry

Finite Fields as Models for Euclidean Plane Geometry

Finite Fields

A field is a set F equipped with two operations +, · such that

a. the pair <F,+> is a group, i.e.,

+ is closed:	$\forall x,y \exists z(x + y = z)$
+ is associative:	$\forall x,y,z(x + (y + z) = (x + y) + z)$
there is a neutral element:	$\forall x(x + 0 = x)$
every element has an inverse:	$\forall x \exists y(x + y = y + x = 0)$

b. likewise <F,·> is a group, the neutral element being 1
c. the following distributivity laws hold:

$\forall x,y,z(x \cdot (y + z) = (x \cdot y) + (y \cdot z))$
$\forall x,y,z((x + y) \cdot z) = (x \cdot z) + (y \cdot z))$

It is straightforward to see that the real numbers $\mathbb{R}$ with the usual addition and multiplication is a field. In this case the set F is infinite, but F can be finite as well. Then we have a finite field or a Galois field. There is however one very important distinction between a field such as $\mathbb{R}$ and a Galois field. In the latter, given the multiplicative neutral element 1, there is a prime number p such that $p \cdot 1 = 0$. p is called the characteristic of the field. It can be shown that if p is the characteristic of a field, then it must have p^n elements, for some natural number n. In addition Galois fields are the only finite fields.

> *Example*: the Galois field with characteristic 3 and number of elements 3, *GF*(3) for short.

The tables for addition and multiplication tell the whole story:

+	0	1	2

0	0	1	2
1	1	2	0
2	2	0	1

·	**0**	**1**	**2**
0	0	0	0
1	0	1	2
2	0	2	1

In the first case, 1 + 1 + 1 = 0.

It is easy to see that the above tables correspond to addition and multiplication modulo p. In other words, in the field $a = b$ if and only if $a \equiv b$ (mod p).

Euclidean Axioms in a Finite Field

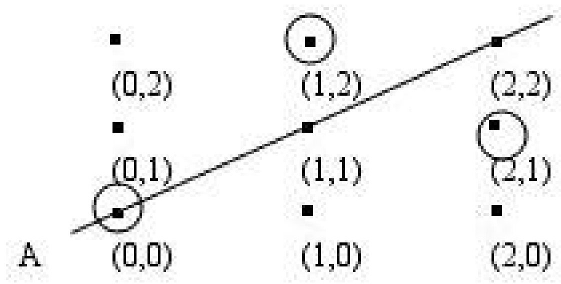
To see how a finite field can be a model, I will take the incidence relation. The axioms for the incidence relation are:

a. Through two points exactly one straight line can be drawn.
b. There are at least two points on every straight line.
c. There are at least three points that are not on the same straight line.

In case we take $GF(3)$ as a possible model, then to a point p corresponds to a couple (x,y), such that $0 \leq x,y \leq 2$. A line corresponds to a triple (a,b,c), such that $0 \leq a,b,c \leq 2$ and at most one of $a = 0$ or $b = 0$, or, equivalently, at most one of $a \equiv 0$ or $b \equiv 0$ (mod 3). The incidence relation is translated as follows: a point p lies on a line A iff if (x,y) corresponds to p and (a,b,c) corresponds to A, then the equation $ax+by+c = 0$ is satisfied, or, equivalently, $ax+by+c \equiv 0$ (mod 3) is satisfied. Some facts are now straightforward to check:

i. There are exactly p^2 points.
ii. There are exactly $p(p+1)$ straight lines.
iii. On every straight line there are exactly p points.
iv. Through every point pass exactly $p+1$ straight lines.
v. The axioms (a), (b) and (c) are true in this model.

A graphical representation could look like this:



The line indicated by *A* corresponds to $x = y$. This seems to correspond nicely to the classical Euclidean case. However something strange happens with the linear equation $x + y = 0$. In the drawing the 3 points that are on this line have been circled. As must be clear these lines are ÒbadÓ and should be "ignored" in the model.

In order to satisfy the remaining axioms further restrictions are required on the size of the domain. These will just be mentioned without details:

a. p must have the specific form $4n+3$ (and not $4n+1$).
b. p must in addition have the specific form $8mq_1q_2...q_k - 1$, where q_i is the i-th odd prime (so $q_1 = 3$) and m a positive integer.

The second condition that is needed to guarantee the existence of the Euclidean "kernel" is a non-trivial statement. It actually requires some essential parts of number theory to prove that there are prime numbers of that form. There are, classically speaking, an infinite number of them. By choosing p large enough, one can make the ‘kernel’ as large as one desires to have the Euclidean approximation as close as one wants.

Note: As Ernst Welti points out, it would be a rather annoying situation for a finitist if the proof that shows that there are an infinite number of primes of the right form were not finitistically acceptable. Although the original proof of Dirichlet was in fact unacceptable, fortunately there does now exist a finitistically acceptable proof of the theorem.

Jean Paul Van Bendegem
jpvbende@vub.ac.be

Return to Finitism in Geometry

First published: April 3, 2002

Content last modified: April 3, 2002

Impartiality

Impartiality is sometimes treated by philosophers as if it were equivalent to *moral* impartiality. Or, at the very least, the former word is often used, without the qualifying adjective 'moral', even when it is the particularly moral concept that is intended. This is misleading, since impartiality in its broadest sense is best understood as a formal notion, while moral impartiality in particular is a substantive concept -- and one concerning which there is considerable dispute. This entry will be predominantly concerned with moral impartiality -- the sort of impartiality, that is, that properly features in normative moral and political theories. We begin, however, by addressing the broader, formal concept.

1. The concept of impartiality

The most common mistake made by philosophers in their use of the words ‘impartial’ and ‘impartiality’ is that of assuming that such words denote a positive, unitary concept -- presumably a concept closely linked with, if not identical to, morality. This, however, is simply not the case. Rather, there are various sorts of behavior that may be described as ‘impartial,’ and some of these obviously have little or nothing to do with morality. A person who chooses an accountant on the basis of her friends’ recommendations may be entirely impartial between the various candidates (members of the pool of local accountants) with respect to their gender, their age, or where they went to school. Yet if her choice is motivated solely by rational self-interested considerations then it is clear that the impartiality she manifests is in no way a form of *moral* impartiality. To take a more extreme case, consider an insane serial killer who chooses his victims on the basis of their resemblance to that some celebrity. The killer may be impartial with respect to his victims’ occupations, religious beliefs, and so forth, but it would be absurd to regard this as a form of *moral* impartiality (despite the fact that, in certain contexts, morality does require impartiality with respect to such considerations.)

It is also worth noting that some types of impartiality may in themselves be immoral or morally questionable. Suppose that I decide to pass along a treasured family heirloom to one of my two sons, Bill and Phil. Flipping a coin would constitute one type of impartial procedure for choosing between the two. But suppose that I have already promised the heirloom to Phil, on several occasions. In this case it would be quite wrong to allow a coin toss to determine whether he gets it. Deciding by means of a coin toss would be an impartial procedure, but it would be the wrong sort of impartiality here, for it would ignore the moral obligation created by my previous promises.

It is important, then, to recognize that *moral* impartiality is a particular species of impartiality, and that the latter, unqualified, denotes a much broader concept. In this broad sense, impartiality is probably best characterized in a negative rather than positive manner: an impartial choice is simply one in which a certain sort of consideration (i.e. some property of the individuals being chosen between) has no influence. An analysis along these lines has been proposed by Bernard Gert. Gert’s analysis holds that "A is impartial in respect R with regard to group G if and only if A’s actions in respect R are not influenced at all by which member(s) of G benefit or are harmed by these actions" (Gert 1995, p.104). Thus, for Gert, impartiality is a property of a set of decisions made by a particular agent, directed toward a particular group.

Gert’s analysis captures the important fact that one cannot simply ask of a given agent whether or not she is impartial. Rather, we must also specify with regard to whom she is impartial, and in what respect. Gert’s analysis, then, permits and indeed requires that we make fairly fine-grained distinctions between various sorts of impartiality. This is necessary, since one and the same agent might manifest various sorts of partiality and impartiality towards various groups of persons. Consider, for instance, a university professor who is also a mother of five children, and who is currently acting as a member of a hiring committee. Such an agent might be impartial between her children with respect to the care they receive (while preferring her own children over others in this respect), and also impartial between the various job candidates; but it is clear that these two uses of the word ‘impartial’ denote very different practices. In particular, the idea of merit applies in one case but not the other: to be impartial between job candidates is presumably to select between them on the basis of merit, whereas to be impartial between one’s children

is *not* to think of merit at all, but rather to provide equal protection and care to all.

Many attempts to characterize impartiality fail to respect the distinction between the broadest, most formalistic sense of the notion, and a more specifically *moral* impartiality. To say, for instance, that an impartial choice is one that is free of bias or prejudice is to presuppose that we are dealing with a certain sort of impartiality, that which is required or recommended by morality, or at least worthy of moral approbation. ‘Bias’ and ‘prejudice’ are loaded terms, suggesting not only that some consideration is being excluded, but also that the exclusion is appropriate and warranted. Similarly, the idea that impartiality requires that we give equal and/or adequate consideration to the interests of all concerned parties goes well beyond the requirements of the merely formal notion. (In the coin toss case, it is quite clear that Phil’s claims to the heirloom are *not* being given equal or adequate consideration.) As a characterization of *moral* impartiality, however, this suggestion is considerably more promising.

2. Morality and impartiality

2.1 The impartial point of view

It is generally agreed that some sort of close connection obtains between morality and impartiality. Indeed, the phrases ‘moral point of view’ and ‘impartial (or ‘impersonal’) point of view’ are sometimes used interchangeably to refer to the imagined impersonal perspective from which, it is supposed, moral judgments are to be made (Baier 1958, chapter 8; Harsanyi 1982; Scheffler 1982, 1985; Smith 1976 [1759]; Wolf 1992; see also Blum 1980, Chapter 3). As noted above, however, the word ‘impartial’ is a general term with many particular species; it follows from this that the phrase ‘impartial point of view’ is itself ambiguous. At most, it might be that the moral point of view constitutes *one sort of* impartial point of view.

Even this claim, however, faces difficulties. While impartiality may play a significant role in moral behavior, it is not clear that it is sufficient to exhaust the demands of morality. Treating a person appropriately and respectfully may well require certain sorts of emotional and/or cognitive responses: sensitivity to her needs and values, empathy for her suffering, and the like. But if these responses are pictured as the results of *positive* traits or attributes (and not simply as, say, the result of a lack of bias or prejudice), then it is not clear that merely being impartial between persons is sufficient to guarantee that one will possess and display the necessary sensitivities. Indeed, characterizations of impartial agents which proceed in negative terms (that is, by defining various preferences, emotions or bits of information that she does not possess or that do not move her) often risk picturing the impartial agent as impersonal and even indifferent (Henberg 1978; Brandt 1954).

A second problem for the claim that the moral point of view is identical with (some version of) the impartial point of view -- or indeed, for any view which identifies morality and impartiality -- is that it seems plausible to regard some forms of moral partiality as morally admirable, and perhaps even morally required (Blum 1980; Cottingham 1983, 1986, 1996; Jeske & Fumerton 1997; Jollimore 2001; Kapur 1991; Kekes 1981; MacIntyre 1984; Oldenquist 1982). Loyalty to one’s family, community or country,

for instance, is commonly regarded as a virtue. Yet such an attitude is a clear and indeed paradigmatic example of partiality, requiring that an agent feel and act differently toward one set of persons than she does toward humanity in general. Similarly, certain specific moral duties arising from certain particular relationships seem to involve partiality in an irreducible manner. Parents, for example, are thought to be morally obliged to take special care of their own children; to regard one's child as merely one among millions would be regarded as highly eccentric if not monstrous. Of course, some moral duties do require that an agent be impartial in performing them. But on common sense moral views at least, impartiality seems mostly to be required in the context of specific roles -- such as when a person is acting as a judge, an umpire, or a representative of some public institution (Baron 1991; Blum 1980; Cottingham 1983). The idea that impartiality is a pervasive and universal moral requirement seems to contradict our ordinary moral intuitions.

Whether there exists such a thing as morally admirable partiality is the main issue that separates the so-called *partialists* from the *impartialists*. Partialists, in general, tend to claim that morally admirable partiality does exist, that it cannot be reduced to any form of impartiality at a more fundamental level, and that these facts pose a serious problem for those who claim that morality and (some form of) impartiality are identical, or even closely related. Impartialists, by contrast, either deny the existence of morally admirable partiality altogether, or hold that any apparent cases are in fact ultimately reducible to impartial standards (see section 6). Thus impartialists hold that -- contrary, perhaps, to appearances -- impartiality is, indeed, a pervasive and universal requirement of morality.

2.2 The ideal observer theory

Rather than being put in terms of an impartial *point of view*, the relation between morality and impartiality is sometimes made out in terms of an impartial *agent* or *observer* -- a person who makes moral judgments without being influenced by the sort of contaminating biases or prejudices that tend to arise from the occupation of some particular point of view. (Smith 1976 [1759]; Hume 1978 [1740]; Firth 1952; Brandt 1954; Hare 1989.) (We should note that this idea is not always clearly distinguished from the conception based on the impartial point of view; Smith 1976 [1759], for instance, seems to advance them both at once, in claiming that the ideal observer simply is the observer who occupies the ideal point of view.)

The observer may also be defined as 'ideal' in various other ways. It is generally stipulated that she is in possession of all the nonmoral facts that are relevant to the judgments she has to make (Firth 1952). It is also fairly common to assume that she is an ideal reasoner, and thus be immune to logical fallacy or mistaken inference, etc. (Indeed, Hare goes so far as to state that his "archangel" possesses "superhuman powers of thought, superhuman knowledge and no human weaknesses" (Hare 1989, p. 44).) The 'ideal observer theory' of morality, in its most straightforward form, states that moral judgments simply are the judgments an ideal observer of this sort will make.

What must be pointed out about such a conception of morality, for our purposes, is that any advantage it has over the conception of morality as an impartial point of view presumably arises from the fact that the ideal observer is not completely defined in terms of impartiality. (If she were, the two conceptions would

simply coincide.) Yet many ideal observer theorists seem to accept a characterization of the ideal observer which concentrates on her impartiality and impersonality. Firth, for example, suggests that the ideal observer is both "distinterested," in the strong sense of being "entirely lacking in particular interests," and "dispassionate," in that she is "incapable of experiencing any emotions at all." (Firth, 1952) Defined in this way, however, the ideal observer sounds not only impersonal but deeply indifferent; and the idea that the moral judgments of a person who had neither emotional responses nor particular interests could be trusted, let alone that they might be considered definitive of morality, strikes some critics as highly implausible (Brandt 1979).

Suppose, then, that the ideal observer theorist decides that the definition of the ideal observer *must* include more than the bare idea of impartiality -- that in addition the observer must be, say, compassionate (and thus not indifferent); and that she must possess a considerable facility for proper moral judgments -- practical wisdom, in the Aristotelian sense. Such a theorist will now face a different problem: the more we build into the definition of our ideal observer, the less useful it becomes as a heuristic device. Stipulating that ideal observer is very wise, for example, is not very helpful if we ourselves are not wise, and so have no idea what an ideally wise observer would choose. Indeed, ideal observer analyses that go too far in this direction seem to become circular -- the ‘ideal’ observer is ideal because she always makes proper judgments, those being defined as just those judgments the ideal observer would make (Broad 1959, p. 263). John Stuart Mill seems to commit an error of this sort when he writes:

> Impartiality, in short, as an obligation of justice, may be said to mean, being exclusively influenced by the considerations which it is supposed ought to influence the particular case in hand; and resisting the solicitation of any motives which prompt to conduct different from what those considerations would dictate. (Mill 1861/1992, p. 154; see also Firth 1952, p. 336)

The ideal observer, then, to be useful, must be given some independent definition, and not simply defined as ‘an agent who always gets it right.’ The challenge, of course, is to find such a definition. Here, as with the conception of morality as defined by an impartial point of view, the phenomenon of morally admirable partiality proves a particularly difficult issue. Should we define the ideal observer as being loyal to her country, or as being above loyalty? If the former, can she serve as an adequate moral example to people who do not share her allegiances? If the latter, how can she serve as an adequate example to *anyone?* The persistent problem that faces the ideal observer approach to moral impartiality seems to be that any process of idealization of the sort required to make such a conception work seems likely to result in an individual so removed from the concrete lives and concerns of actual human moral agents, that her moral judgments will turn out to be in large part irrelevant to the question of how such agents ought to live (see Walker 1991).

2.3 Moral impartiality and equality

The moral importance of the impartial point of view is that from it, every moral agent counts equally: no one, including the person occupying that point of view, is considered to be intrinsically more significant

than anyone else, or to have more powerful claims to attention simply by virtue of who they are. Similarly, one of the primary virtues of the ideal observer seems to be that she also regards persons in this way. Whatever these conceptions may get wrong, then, one thing they seem to get right is the idea that there is a close and important connection between moral impartiality and equality (see especially Nagel 1991, Chapter 7).

Some clarification, however, is required. To say that from the impartial point of view, no one is seen as *intrinsically* more significant than anyone else, is not to say that there is no reason *whatsoever* for which a person might count as more significant than others, or for which a person might legitimately demand disproportionate moral attention in some circumstances. Many moral theorists, after all, will suppose that from the impartial point of view, properly conceived, some individuals *will* count as more significant, at least in certain ways. William Godwin (Godwin 1793) provides an influential, and rather infamous, example. Fenelon, the archbishop of Cambrai, Godwin writes, may be supposed to be more significant than a mere chambermaid, and it follows -- at least according to Godwin -- that in the case of a fire, the archbishop ought to be rescued first. The reason, however, is not simply that the archbishop happens to be *himself* -- that is, his greater significance is not supposed to be intrinsic; rather, the claim is grounded on the fact that the archbishop makes greater contributions to society:

> We are not connected with one or two percipient beings, but with a society, a nation, and in some sense with the whole family of mankind. Of consequence that life ought to be preferred which will be most conducive to the general good. In saving the life of Fenelon, suppose at the moment when he was conceiving the project of his immortal *Telemachus*, I should be promoting the benefit of thousands who have been cured by the perusal of it of some error, vice and consequent unhappiness. Nay, my benefit would extend further than this, for every individual thus cured has become a better member of society and has contributed in his turn to the happiness, the information and improvement of mankind. (Godwin 1793, 41-42)

The claim that the archbishop is more significant, then, is grounded on what are essentially universal properties: if the chambermaid also had it in her to write *Telemachus*, her claim to be rescued would be just as great. Fenelon's claim to priority is not based on his *intrinsic* significance. In Godwin's mind, the fact that one individual has a greater claim to rescue than the other is not only not in conflict with impartiality, but indeed is implied by impartiality, for it is based on the (alleged) fact that an impartial judgment of their worth attributes more to one than to the other.

Thus, viewing persons from an impartial point of view need not imply that we view them equally, in every sense of the word; and it certainly does not imply that everyone must receive *equal treatment.* (In Godwin's Archbishop Fenelon, if we assume that only one person can be saved, the only way to give the archbishop and the chambermaid equal treatment would be to let them both perish in the flames.) What impartiality seems to require is not that everyone receive equal treatment, but rather that everyone be *treated as an equal* (Dworkin 1977, p. 227). While the distinction between equal treatment and treatment as equals is difficult to make out with precision, the main idea is fairly clear: treatment as equals requires that persons are *not* treated equally, but rather treated in accordance with what rights they possess, what

legitimate claims they put forward, and, in general, with what they deserve. Thus, to inflict a one year jail sentence on all accused persons, regardless of whether they are guilty or innocent, is to provide equal treatment to members of that group; but it is not to treat them as equals.

Whether or not moral impartiality obliges us to see the archbishop as having a greater claim to be rescued is, of course, controversial. Moreover, Godwin asserts two even more controversial claims in connection with the Archbishop Fenelon example. The first is that the chambermaid *herself* ought to have perceived the greater significance of the archbishop, and so should have sacrificed her own life for his, were that necessary. The second is that any other person -- even the child or husband of the chambermaid -- ought to have been willing to make the same sacrifice:

> Supposing the chambermaid had been my wife, my mother or my benefactor. That would not alter the truth of the proposition. The life of Fenelon would still be more valuable than that of the chambermaid; and justice -- pure, unadulterated justice -- would still have preferred that which was most valuable. Justice would have taught me to save the life of Fenelon at the expense of the other. What magic is there, in the pronoun "my" to overturn the decisions of everlasting truth? My wife or my mother may be a fool or a prostitute, malicious, lying or dishonest. If they be, what consequence is it that they are mine? (Godwin 1793, pp. 41-42)

Whether this extreme position really is required, either by moral impartiality or by the demand that we treat people as equals, is a matter of great dispute, not only between partialists and impartialists but within the impartialist camp itself. If nothing else, Godwin's position is quite clearly incompatible with the apparent existence of morally admirable partiality. (Williams (1981) holds that even to consider sacrificing one's wife for the sake of impersonal justice constitutes a kind of moral error in its own right.) Moreover, despite the fact that the ultimate evaluation is made on the grounds of perfectly general properties, it is not entirely clear that the objects of the evaluation really are being treated as equals, in the relevant sense -- the fact that the chambermaid's life is to be sacrificed for the greater good at least suggests that her standing as a moral being is not really being taken into account, and that the suggested understanding of moral impartiality is therefore deficient. These issues, which we will return to in section 3.2 and in section 6, are at the heart of the contemporary debate about moral impartiality.

3. Moral impartiality (1): Consequentialist moral theories

3.1 The nature of consequentialist impartiality

Consequentialist moral theories hold that moral evaluations and justifications must ultimately be grounded in the value of the consequences of the actions, rules, policies, strategies, character traits, etc. that are being evaluated (Hooker 1994). That is, the ultimate question to be asked of any action, rule, or character trait under evaluation is, "Does it [the action, rule, or trait in question] promote the good?" For

the purposes of this entry, three important assumptions will be made regarding consequentialist theories. First, consequentialist theories will be assumed to hold that the overall values of sets of consequences can be determined, and thus ranked, independently of the identity of any particular agent (thus, we can speak of the "best" consequences without having to ask, "best for whom?") Second, such theories will be assumed to hold that the impersonal good (i.e. the overall value of some particular state of affairs) is largely if not entirely composed of the interests of individual persons, and that the interests of each person count for just as much as those of every other person. Finally, it will be assumed that we are dealing with *act* consequentialist theories -- theories, that is, which hold that the consequentialist standard is to be applied directly to the actions of agents, and that what is required is that every action (or overall pattern of action) *maximize* the impersonal good. Such a theory, then, requires that every agent always choose an action that will bring about consequences at least as good as those that would be brought about by any other available action. (It should be noted that rule consequentialism, which holds that the consequentialist standard is to be applied to rules of action rather than directly to actions themselves, is excluded by this set of assumptions. Rule consequentialism does incorporate a significant form of moral impartiality, but it is not the sort that is incorporated by more traditional act consequentialist theories (Howard-Snyder 1993).)

Together, these three assumptions regarding the nature of consequentialism result in a theory which, on the surface at least, seems to place each agent under a pervasive obligation to be strictly impartial between all persons, by requiring her always to exclude from her practical deliberations (almost) all considerations that do not bear directly on the ways in which people's interests might be advanced or injured by her actions.

The consequentialist standard, then, is strictly impartial in a very rigorous sense. A consequentialist agent is not permitted to prefer herself, nor any of her loved ones, in choosing a distribution of benefits and burdens. She may not accept a pleasure for herself if doing so involves passing up the opportunity to bring about a slightly larger pleasure for a stranger. Nor is she permitted to feed her own children if she could do more good by feeding hungrier strangers instead. She must sacrifice the life of a spouse, parent or child if, by doing so, she would save more lives, or even save the life of one other person whose contribution to the overall good would be greater than that of the person sacrificed. (Recall Godwin's Archbishop Fenelon case, discussed in section 2.3.) It is for reasons such as this that consequentialist impartiality is accused of being too demanding. By refusing to allow the agent's personal concerns to play a special role in her practical deliberations, it is claimed, consequentialism threatens her integrity and alienates her from herself and others (Kapur 1991, Scheffler 1982, Stocker 1976, Williams 1973, 1981). As Brian Barry has written, the effect of consequentialist impartiality "is, in effect, to extend to the whole of conduct the requirements of impartiality that on the common-sense view are restricted to judges and bureaucrats acting in their official capacities." (Barry 1995, p. 23) The kind of impartiality that features in consequentialist theories, then, seems to be much more pervasive, and much more severe, than that recommended by common sense morality.

3.2 Is consequentialist impartiality too demanding?

The fact that consequentialist impartiality turns out to have such strict and demanding implications is, for

the consequentialist, a double-edged sword. On the one hand, there is no doubt that consequentialism is a deeply impartial moral theory; on the plausible and popular assumption that a moral theory *must* be deeply impartial, consequentialism meets this criterion with flying colors. And consequentialists have typically been adept at exploiting this fact with powerful rhetoric (Godwin's famous query, 'what magic is there in the pronoun 'my'?' being a noteworthy example.) On the other hand, the impartial demands of consequentialism are so strict and so extreme that some critics have found them unacceptable: consequentialism, they claim, simply demands too much and must therefore be rejected (Scheffler 1982, Slote 1985, Williams 1981).

Essentially, this worry is a version of what we referred to above as the puzzle of morally admirable partiality. The common-sense view is that it is permissible for an agent to be partial toward herself; that is, to treat her own projects and concerns as if they had special significance (Scheffler 1982). (From her point of view of course, they *do* have special significance.) This sort of self-concern, then, constitutes a form of partiality which seems, from the vantage point of common sense, to be morally endorsed. Similarly, certain sorts of partiality directed toward *other* people -- friends, family members, and the like -- are also forbidden by consequentialist impartiality, but regarded as justifiable, and in many cases admirable, from the standpoint of common sense (Blum 1980, Cottingham 1983, Kekes 1981, Slote 1985).

Defenders of consequentialism generally respond in one of three ways. First, a consequentialist might argue that *any* genuinely impartial moral theory will make extreme demands of agents. Second, they might argue that even though it is not necessary for a genuinely impartial theory to make extreme demands, the fact that it does is no grounds for an objection against it. Third, they might argue that in fact, the demands of consequentialism are not as extreme as have been supposed.

The first strategy faces a severe difficulty: namely, it at least seems to be the case that certain non-consequentialist moral theories -- in particular, deontological theories -- also incorporate impartial elements in a fundamental manner, and yet make demands on the moral agent which are considerably less extreme than those of consequentialism. Thus, while some consequentialists (e.g. Brink 1989) have argued that the truth of consequentialism can be logically derived more or less directly from the requirement that morality be impartial, this seems to be a mistake (Scheffler 1992, pp. 105-109). Of course, it is open to the consequentialist either to deny that deontological moral theories are *genuinely* impartial (Kagan 1989; Scheffler 1982, 1985), or to argue that, properly understood, any plausible ethical theory will be seen to make demands comparable to those made by consequentialism (Brink 1989, Ashford 2000). Both of these strategies, however, face difficulties; as we will see in section 4, there is in fact a very strong case in favor of viewing at least some deontological theories as genuinely and fundamentally impartial -- a case which nevertheless does not prohibit us from viewing such theories as less demanding than their consequentialist rivals.

The second strategy is to argue that those who object to consequentialism on the grounds that it is too demanding are placing too much importance on the role of morality in practical reasoning (Brink 1989; Wolf 1982, 1992). If moral considerations dominated practical reasoning -- if, that is, they were the only or at any rate by far the most significant considerations in determining our actions -- then

consequentialism would be untenable, on account of its demanding too much. A proper perspective on practical reasoning, however, reveals that moral demands constitute only one set of demands among many. When put in their proper place then in the larger scheme of practical reasons and requirements, the extreme demands of consequentialist morality will no longer seem threatening. To borrow a pair of phrases from David Brink, what appear to be "moral worries" about the tendency of consequentialism to make excessive moral demands, might really be "worries about morality" -- worries, that is, about whether or not we have reason to act as morality requires. Whether the view of morality presupposed by this strategy is true, however, is questionable; at the very least, it does not seem to be the case that the majority of those who have defended consequentialism as a normative theory of ethics have intended it to be viewed as a theory that could be frequently or easily overridden (Jollimore 2001, Chapter 3; Peter Railton (1984) and Richard Miller (1992, Chapter 10) also express doubts about this approach).

The third strategy is perhaps the best known and most frequently employed. It is argued that, given a reasonable and accurate view of human nature and the abilities of agents, it will be seen that what consequentialism requires is *not* a radically different sort of life from the one most of us currently live; rather, consequentialism will require (in most cases, at least) only reasonable, and relatively minor, adjustments in our current lifestyles. In particular, it is argued that consequentialism permits the agent both to give preference to her own projects and concerns, and to favor particular other individuals (friends, family members, etc.), and that all this is consistent with the agent's having as her overriding project the maximizing of the good. The *locus classicus* of this argument is found in Mill's *Utilitarianism:*

> The objectors to utilitarianism cannot always be charged with representing it in a discreditable light. On the contrary, those among them who entertain anything like a just idea of its disinterested character, sometimes find fault with its standard as being too high for humanity. They say it is exacting too much to require that people shall always act from the inducement of promoting the general interests of society. But this is to mistake the very meaning of a standard of morals, and to confound the rule of action with the motive of it […] The great majority of good actions are intended, not for the benefit of the world, but for that of individuals, of which the good of the world is made up; and the thoughts of the most virtuous man need not on these occasions travel beyond the particular persons concerned, except so far as is necessary to assure himself that in benefiting them he is not violating the rights -- that is, the legitimate and authorized expectations -- of any one else. The multiplication of happiness is, according to the utilitarian ethics, the object of virtue: the occasions on which any person (except one in a thousand) has it in his power to do this on an extended scale, in other words, to be a public benefactor, are but exceptional; and on these occasions alone is he called on to consider public utility; in every other case, private utility, the interest or happiness of some few persons, is all he has to attend to. (Mill 1992 [1861], pp. 64-66.)

Similarly, Godwin (1968 [1801]) argues that

> True wisdom will recommend to us individual attachments […] since it is the object of

> virtue to produce happiness; and since the man who lives in the midst of domestic relations will have many opportunities of conferring pleasure, minute in the detail, yet not trivial in the amount. (Quoted in Cannold, *et al*., 1995)

(This position, of course, appears to be in some amount of tension with the more extreme consequentialist position attributed to Godwin in section 2.3).

More recent versions of this argument follow Mill's basic strategy. Peter Railton (1984) argues that a "sophisticated" consequentialist will develop patterns of decision-making that do not, except on rare occasions, refer explicitly to consequentialist aims and goals, and that both the psychology and the outward behavior of such an individual will be similar to that of the typical non-consequentialist. Similarly, Frank Jackson (1991) argues that the most efficient strategy for a dedicated consequentialist is to concentrate on small groups of particular persons, rather than trying to promote the well being of humanity at large, and that this will involve the formation of close personal relationships with other individuals. Others who have deployed versions of this argument include Bales (1971), Brink (1989), and Pettit & Brennan (1986).

The evaluation of this consequentialist strategy is a difficult issue. Consequentialists are surely correct to point out that obsessive consequentialist strategizing is likely, at a certain point, to turn counter-productive, and that a consequentialist agent is therefore well-advised to develop more moderate approaches. On the other hand, Mill and many other consequentialists seem to underestimate the amount of good that a dedicated consequentialist agent might be able to contribute, and thus, to underestimate the amount of good that she will be *required* to contribute. Moreover, our powers to influence the lives of strangers have increased considerably since Mill's day. As Susan Wolf writes, "[T]his argument is simply unconvincing in light of the empirical circumstances of our world. The gain in happiness that would accrue to oneself and one's neighbors by a more well-rounded, richer life than that of the moral saint would be pathetically small in comparison to the amount by which one could increase the general happiness if one devoted oneself explicitly to the care of the sick, the downtrodden, the starving, and the homeless" (Wolf 1982, p. 428; see also Singer 1972). It is not clear, then, that an appeal to the limits of human powers can succeed in converting what is, after all, a fundamentally radical moral theory, into a comfortably conservative one.

3.3 Consequentialist impartiality and justice

In addition to claiming that consequentialist impartiality is too demanding, many critics have also argued that it is too permissive. Since consequentialism makes the permissibility of an action entirely dependent on the value of that action's consequences, it follows that there is no *type* of action that can be prohibited on consequentialist grounds (except, of course, for that "type" which is defined explicitly in terms of sub-optimal consequences.) Thus instances of torture, premeditated murder, rape, and other violations of fundamental human rights are at least potentially justifiable on a consequentialist basis; no such action can be ruled out, morally speaking, until the comparative value of the state of affairs it will bring about has been determined.

The effect of this complaint, like the previous one, is not to deny the claim that consequentialism is a deeply impartial moral theory, but rather to suggest that it incorporates the wrong sort of impartiality. Suppose, to take an example common in the literature, that consequentialism recommends that an man be convicted of, and punished for, a crime he did not commit, in order to prevent the public from rioting (McCloskey 1963). Such an action would, according to common intuitions, constitute a gross violation of justice; and it seems a weak reply to point out that the recommendation was arrived at through an impartial calculation -- a calculation that took the interests of every individual (including the framed man) into equal account. For while the claim is, strictly speaking, true, there is nevertheless a clear and compelling case in favor of concluding that the framed man was *not* treated impartially, in the sense that ought to matter here. We expect a judicial system to allocate punishments in accordance with degree of guilt, not in accordance with the expected value to society of the consequences in each case; and the fact that both methods constitute forms of impartial decision-making does not imply that they are equally morally acceptable.

Again, the classic response to this objection dates back to Mill's *Utilitarianism* (1992 [1861]). If institutions of justice are to be given a general justification, Mill argues, this justification must find its ultimate grounding in utility to society; for what else could explain why justice is valued at all, other than the fact that it serves and protects our interests? But since a justice system will only succeed in this role if it is governed by common principles of justice -- principles including, for instance, that only the guilty should be punished, and that the punishment ought to be proportional to the crime -- it follows that such principles are not opposed to consequentialism at all. Rather, at the deepest justificatory level, consequentialism and the demands of justice coincide.

The claim that such a coincidence generally obtains is probably easy to establish. The challenge for Mill, and for other consequentialists, arises in those particular cases in which the coincidence fails. Assuming that the possibility of such cases does not move one to simply abandon consequentialism in favor of some more justice-friendly conception (such as the rule consequentialism Mill himself sometimes seems to find attractive), there are two general defense strategies for consequentialists to employ. The first strategy argues that there are good consequentialist reasons for being the sort of agent who respects the dictates of justice even in cases in which the coincidence between the demands of justice and those of consequentialism fails (Pettit 1997; cf. Railton 1986). The second strategy admits that there are cases in which unjust actions can be given a consequentialist justification, but holds that when so much as it stake, justice must give way to consequentialism's demands (Smart 1973; Kagan 1989; Pettit 1997). Whether either approach is sufficient, given the apparent depth and force of our common intuitions about the requirements of justice, is a matter of ongoing debate.

4. Moral impartiality (2): Deontological moral theories

4.1 Deontological impartiality and the personal point of view

In section 3.2 we noted that while consequentialist impartiality is one possible interpretation of the demand that morality be impartial, it is not by any means the only available interpretation; nor is it clearly the most plausible. The considerations related to justice discussed in section 3.3 may help us to appreciate this. For consider once more the position of the framed innocent, whose fundamental interests have been sacrificed for the sake of the greater good. Such a person may well complain that he has *not* been treated impartially, in the appropriate sense; for, while it is true that his interests were counted in determining the nature of the overall good, it is nevertheless also true that ultimately, he became the victim of a form of abuse that was both harsh and undeserved. The framed innocent might also back up his complaint by making the plausible claim that, had he been in a position to choose, he would never have consented to a moral system that allowed *anyone* to be accorded such treatment. Thus, while there is a sense in which his interests were counted equally, there is another and very important sense in which his interests -- and perhaps more importantly, his claims and rights -- do not seem to have received full or adequate consideration at all.

Deontologists insist that consequentialism errs by failing to accord proper significance to the moral agent as an individual; in John Rawls' words, consequentialism "does not take seriously the distinction between persons" (Rawls 1971, section 5). (Rawls has utilitarianism in particular as his target, but the point applies more widely.) The fact that consequentialist impartiality makes extraordinary and, to many, unreasonable demands on the individual (section 3.2) indicates that consequentialism fails to take individuals seriously as *agents*. And the fact that consequentialist impartiality permits the individual to be used as a mere means when doing so promotes the greater good (section 3.3) indicates that consequentialism fails to take individuals seriously as *patients*. The conception of impartiality that tends to be favored by deontologists avoids such implications by refusing to view impartial action simply as a matter of maximizing interests (or some other version of the impersonally determined good.) Indeed, deontologists take *the right* rather than *the good* to be fundamental to ethics, and tend to see moral action in terms of acting in accordance with principles that are rationally acceptable to all.

Exactly what these principles are, and exactly what method should be used to determine them, are matters of some disagreement among deontological theorists. But there does seem to be a general consensus among deontologists that moral impartiality does *not* require that an agent be strictly neutral between her own good and the good of other people in ordinary decision-making contexts. Rather, an agent is permitted on deontological views to give special attention to her own projects and interests. An important distinction can be drawn here between first-order and second-order impartiality (Barry 1995). First-order impartiality is that displayed by an agent in ordinary choice situations -- choosing how to spend one's day, who to spend time with, and so forth. Second-order impartiality, by contrast, operates only in a certain, special sort of context: contexts in which the rules, principles and institutions which govern first-order behavior are evaluated and selected. Thus a moral rule granting individuals complete freedom of association, and thus allowing them to display first-order partiality by spending time with whomever they please regardless of whether doing so promotes the greater social good in any particular case, might be given a second-order impartialist justification by demonstrating that such a rule would promote the impersonal good, or that it would be selected by a group of impartial persons who were choosing the moral rules that were to govern society.

The fact that deontological theories generally permit (some degree of) first-order partiality -- that is, that agents are permitted to pay special attention to their own interests, projects, and loved ones -- should not, then, be taken to imply either that the agent's interests are *objectively* more valuable than those of other persons, or that the agent is justified in viewing them as such. Rather, the deontologist will claim, it reflects the fact that it is morally legitimate (perhaps, again, because justifiable in second-order impartialist terms) for an agent to regard her own goals and interests as especially important *to her*. Thus, deontological moral systems tend to incorporate an irreducible element of agent-relativity of a sort that consequentialist theories cannot embody (Nagel 1986; McNaughton & Rawling 1992, 1993, 1998; Jollimore 2001).

The incorporation of agent-relativity of this sort into deontological theories allows such theories to escape the most straightforward versions of the claim that they demand too much of moral agents (Jollimore 2001). Nevertheless, various versions of that objection have been leveled against deontological theories. It has been claimed, for instance, that Kantianism, by insisting that only actions performed out of the motive of duty have moral worth, deligitimizes or even forbids the type of motives which typically (and perhaps necessarily) operate in the context of close personal relationships (Stocker 1976; Williams; 1981). Typically, Kantians have responded by distancing themselves from the view that *only* actions motivated by duty have value, and acknowledging instead that a commitment to duty need only function as a limiting condition, rather than as the primary source of motivation in all cases (Baron 1995). The Kantian account of moral value, of course, is not essential to deontological theories; and those theories which eschew it may well be able to avoid the demandingness objection altogether.

4.2 Impartiality and universalizability

On many deontological views, particularly Kantian ones, the significance of moral impartiality is seen as arising from the fact that a core role is given to the concept of universalizability (Gert 1998; Hare 1981; Kant 1964 [1785]; Kohlberg 1979). The requirement that moral judgments be universalizable is, roughly, the requirement that such judgments be independent of any particular point of view. Thus, an agent who judges that *A* ought morally to do *X* in situation *S* ought to be willing to endorse the same judgment whether she herself happens to be *A*, or some other individual involved in the situation (someone who, perhaps, will be directly affected by *A*'s actions), or an entirely neutral observer. Her particular identity is completely irrelevant in the determination of the correctness or appropriateness of the judgment.

Universalizability, thus formulated, does imply at least one sort of impartiality: an agent whose judgments are universalizable will be morally consistent, in the sense that she will judge her own actions by the same standards she applies to others. Such an agent will not make an exception of herself by allowing herself to break a rule she regards as binding for others, or to perform any other action which she would not accept if performed by another agent. Impartiality of this sort, however, does not necessarily imply any sort of impartiality with respect to other individuals' interests, rights, or claims. On a minimally demanding interpretation of the universalizability requirement, the judgments made by a person whose conception of the good was intrinsically racist -- that is, a person who held that the well-being of members of some one particular race mattered more (or less), objectively speaking, than the well-being of members of other races -- could very well turn out to be universalizable, so long as the racist held that his judgments were

objectively correct, and so ought to be assented to by all individuals -- including those individuals who would be disadvantaged by the general adoption of those views (cf. Gewirth 1978, p. 164; Gert 1998, Chapter 6; Wiggins 1978; Williams 1985, p. 115).

However, the conclusion that the racist's judgments are universalizable presupposes a very minimal account of what universalizability requires. On this account, it requires only that an agent be sincerely committed to the objectivity of his judgments, in the sense that he views them (from his current perspective) as correct from all perspectives, and thus as calling for everyone's assent (whether or not that assent is actually given.) There are two ways of making the universalizability requirement more demanding. The first is to appeal to certain counterfactual claims about what the agent *would* endorse if he actually *did* occupy various perspectives. On this view, a particular judgment by *A* is universalizable if and only if *A* endorses that judgment from his current perspective, *and would endorse the same judgment from any other perspective*. Given this understanding of universalizability, it is much less likely -- indeed, extraordinarily unlikely -- that racist views will turn out to be universalizable; for it is not generally true of individuals that they would endorse the view "The well-being of members of race *R* matters less than the well-being of members of other races" if they themselves were members of race *R*. However, such a view may well require too much, for there are few if any moral judgments or principles that would be endorsed from *every* perspective any given agent might occupy.

A different approach to universalizability eschews the appeal to psychological facts altogether, and holds that whether or not a particular judgment is universalizable is a logical fact rather than a psychological one. Kant's categorical imperative test, for example, holds that universalizability is the distinguishing feature of correct moral judgments, and that a judgment is universalizable if and only if it can, without contradiction, be willed as a universal practical law (Kant 1964 [1785]). Since the test hinges on whether the willing of a judgment as a universal law results in a *contradiction*, it follows that whether or not a judgment *is* universalizable in this way is a matter of practical reason, and does not depend on which particular individual's will happens to be involved.

The types of impartiality implied by both of these more demanding versions of the universalizability requirement are likely to be considerably more substantial than the formal consistency required by the minimal version. Kant, for instance, seems to hold that universalizability implies a certain level of altruism or charity, in the form of the imperfect duties we owe towards other individuals. There are problems, however, with Kant's argument for this. In particular, it is not clear just how the universal willing of a maxim such as "When others are in need of help, I always ignore their needs" give rise to any sort of *contradiction*. It is true, of course, that, were we actually in a position to choose the universal maxims on which all rational persons would act, this would be a poor choice, for we might someday be in need of assistance from another. But to say that the willing of this maxim as a universal law would be *imprudent* is not to say that doing so is *contradictory*. Moreover, as David Wiggins (1978) points out, certain other actions that seem as if they ought to be morally permissible -- the act, for instance, of releasing a debtor from his debt out of generosity -- have maxims that seem to fail the universalizability test so conceived. These examples may point to a general problem with the attempt to derive impartiality from universalizability: whereas the latter, at least on a Kantian interpretation, is a formal property of moral judgments, moral impartiality, as we have seen, is a substantive rather than a formal concept. (See

Herman 1993 and Korsgaard 1996 for attempts to respond to these problems.)

It should be mentioned that some moral theorists have attempted to derive various versions of consequentialist impartiality more or less directly from the universalizability requirement (Hare 1981, Pettit 2000; see also Harsanyi 1982). However, the claim that a conception of impartiality that is not only substantive but also extraordinarily demanding can be derived from a requirement which, as just pointed out, is essentially a formal one, continues to strike a majority of moral philosophers as dubious.

4.3 Contractualist models of deontological impartiality

The requirement that moral judgements be universalizable seems to reflect two fundamental moral insights: first, that morality is objective, and not simply a matter of personal opinion or expression of interest and desire; and second, that from the standpoint of morality, each person matters just as much as, and no more than, any other person. The *contractualis*t approach to moral theorizing provides one method of giving expression to these fundamental ideas about morality.

Contractualism borrows from the social contract tradition the idea that morality may be viewed as the result of an agreement between those who are to be bound by its dictates. Two variants of this approach can be distinguished. The former, sometimes referred to as *contractarianism*, identifies the participants in the bargaining process with actual individuals, and thus is broadly historical. The latter approach, by contrast, appeals to what agents *would* choose under various, quite possibly unrealizable conditions, and is thus *hypothetical* rather than historical. It is the latter approach that will concern us here.

The hypothetical contractualist model, then, regards moral principles as the result of a bargaining process among a group of agents, subject to certain restrictions that are specified so as to guarantee that the chosen principles will meet the demands of second-order impartiality. The most famous example of this approach is John Rawls' "veil of ignorance", as described in (Rawls 1971). According to Rawls, the principles of a just society are those that would be chosen by self-interested rational agents in the "original position" -- a position in which agents possess broad knowledge about human history and the nature of the world they live in, but are denied specific information regarding their own particular identities or prospects in the society in question, the nature of that society, and, crucially, the nature of their own particular conception of the good. Since nobody knows who they will be or what social position they will occupy, there is no opportunity for anyone in an advantaged position to take advantage of that position in order to force a less privileged party to concede to an otherwise unacceptable outcome. It is this fact that allows Rawls to claim that principles chosen under the veil of ignorance are guaranteed to be impartially acceptable to all -- and thus, guaranteed not to be unjust.

It should be noted that Rawls does not intend that morality in its entirely be derived from the original position. Rather, the function of the original position is limited to the choice of the most general principles of social justice in a well-ordered society (Rawls 1971, section 2; 2001, section 12). Nevertheless, Rawls' mechanism is intended to draw the broad outlines of what many see as the most important part of morality: its public or political aspect. By viewing political morality as the result of an agreement between

contractors limited by the strictures of the veil of ignorance, Rawls intends to develop a political philosophy that reflects his commitment to the idea of liberal neutrality: the idea, that is, that each person has a private right to her own conception of the good, and that particular conceptions of the good therefore ought not to be legislatively instituted, nor legislated against

An especially difficult task attending a project of this sort is that of determining what shape this political morality will take -- that is, determining which principles would be chosen by agents in the original position. On Rawls' account, the contractors settle on principles that guaranteed as much liberty as possible for all and, within the limits set by this guarantee, a roughly egalitarian distribution of goods in which inequalities are allowed only if they are to the benefit of the worst off (Rawls 1971, section 11; 2001, Part II). The claim that such principles would recognize all persons as equals -- and thus, their claim to reflect the demands of moral impartiality -- is supported by several considerations, of which three are perhaps most significant: first, that all persons are guaranteed equal (and substantial) civil liberties; second, that the resulting allocation of resources is broadly egalitarian, and in particular, ensures, so far as is possible, that the fundamental needs of all persons are met; and third, that since the only inequalities that are permitted are those that would benefit the least advantaged, it can presumably be assumed that the least advantaged would give their assent to the existence of such inequalities (they would not, even if they could, veto the system.)

In Rawls' scheme, the function of the veil of ignorance is necessary to prevent rational self-interested persons from using their knowledge of their own positions to win unfair advantages over others. An alternative approach abandons both the veil of ignorance and the assumption that the bargaining parties are primarily self-interested. This is the strategy favored by T.M. Scanlon, whose contractors are motivated not by self-interest but by "the desire for reasonable agreement" (Scanlon 1982, p. 115 n. 10; see also Scanlon 1978, 1998; Barry 1995). On the resulting account of moral permissibility, "an act is wrong if its performance under the circumstances would be disallowed by any system of rules for the general regulation of behavior which no one could reasonably reject as a basis for informed, unforced general agreement." (Scanlon, 1982, p. 110) The requirement of impartiality is captured here by the basic fact that the question is whether *everyone* who is to live under the selected rules can reasonably accept them. As in Rawls' theory, however, the principles of second-order impartiality accepted at the contract level allow for considerable first-order partiality at the level of agent-choice.

Harsanyi (1977) argues that a version of utilitarianism can be defended on the basis of an "equiprobability model," according to which an agent ought to choose between social systems "under the assumption that, in either system, he would have the same probability of occupying any one of the available social positions" (Harsanyi 1982, p. 45; cf. Hare 1981). Gert (1998) argues for a list of moral rules which "all impartial rational persons would favor including […] as part of the moral system" (p. 158). Gauthier (1986) defends a contemporary version of contractarianism.

5. Objections to traditional conceptions of moral impartiality

Traditional conceptions of impartiality such as those we have been discussing face a variety of objections. Many of these objections focus on the claim that such conceptions take insufficient account of the nature of the moral agent and of the pragmatics of the situations in which impartial decisions are actually made. Along these lines, some objectors claim that traditional conceptions set the bar too high, demanding superhuman abilities or cognitive efforts from moral agents. Others, meanwhile, concentrate on the fact that the traditional conceptions tend to identify impartiality with an unemotional, dispassionate disposition, or with impersonal bureaucratic institutions; or, more generally, to abstract away from the particular natures of the contesting parties in developing an account of impartial decision-making procedures.

The first charge -- that impartiality, as conceived by traditional ethical theories, makes extraordinary and unreasonable cognitive demands on moral agents -- must be distinguished from the objection to consequentialist impartiality considered earlier, which claimed that the *sacrifices* demanded by consequentialist impartiality were unreasonable and excessive. The objection now being considered is not that impartiality asks the agent to give up too much, but rather that the cognitive feats demanded by these moral theories will exceed the capacity of the typical moral agent. Indeed, one popular version of this objection alleges that an agent will require an unreasonable amount of knowledge or cognitive ability simply to be able to identify what the demands of impartiality *are* (Friedman 1989; Walker 1991). Given the conception of the impartial point of view as a "God's eye" point of view, for example (Baier 1958), it seems questionable whether it is ever reasonable to expect a human moral agent to be able to occupy such a perspective. God, quite obviously, possesses far more knowledge than does any human being; moreover, God's point of view is both objective and impersonal in ways that an individual human's perspective cannot be. . (As Margaret Urban Walker points out, it is often said that human beings have to live with their decisions, but it sounds very odd to say that of God (Walker 1991, p. 765).) Similar remarks apply to the conception of the impartial point of view as "the point of view of the universe" (Sidgwick 1907), to Hare's conception of the ideal moral agent as a so-called "archangel" (Hare, 1981), and, Walker claims, to Firth's conception of the ideal impartial observer (Firth, 1952.) Similarly, Marilyn Friedman points out that even if a person did manage to occupy such a point of view for a period of time-- supposing such a thing to be possible -- there would be no way to confirm that she had successfully done so: standard conceptions of impartiality, she claims, prescribe "methods of normative thinking [which] represent psychological and epistemic feats, the achievement of which we have no independent way to confirm" (Friedman 1991, p. 645).

The second objection finds fault with the traditional tendency to define impartiality in negative or abstract terms -- in terms, that is, of which elements must be *absent* from the psychology of the agent, or which we must *pretend* are absent in the process of idealization. M.C. Henberg, for instance, claims that most if not all procedural accounts of impartiality confuse it with distinterest or impersonality, and thus, ultimately, with indifference. (It should be noted that many impartialists are quite explicit about the link between morality, impartiality, and the lack of emotion; Baier (1958), for instance, writes that "the moral point of view [is] that of an independent, unbiased, impartial, objective, dispassionate, disinterested observer" (p. 201; see also Firth 1952).) Similarly, Richard Brandt argues that it is a mistake to define moral impartiality with reference to an ideal observer who is defined as (among other things) distinterested; for after all, "it is not clear that a purely disinterested being would support a moral system at all" (Brandt

1979, p. 227). While Brandt's complaint is particularly directed at the ideal observer theory of (Firth 1952), this objection seems to apply much more broadly; it is obvious, for instance, that Rawls's veil of ignorance is designed precisely to prevent the contractors from acting in an interested manner.

The problem is not only that impersonal persons of this sort are likely to suffer from massive indifference, but also that there is alleged to be a conceptual difficulty with the very idea of conceiving impartiality in such terms. An abstract or impersonal evaluator, it is argued, could not possibly make reliable judgments about substantive moral matters (whether or not he was motivated to), since he would be unable to appreciate the particular concerns of the contesting parties. Both of these difficulties -- the motivational and the cognitive -- are well expressed by Iris Marion Young, who rejects altogether the idea that morality is primarily a matter of impartiality:

> The ideal of impartiality is an idealist fiction. It is impossible to adopt an unsituated point of view, and if a point of view is situated, it cannot stand apart from and understand all points of view. It is impossible to reason about substantive moral issues without understanding their substance, which always presupposes some particular social and historical context; and one has no motive for making moral judgments and resolving moral dilemmas unless the outcome matters, unless one has a particular and passionate interest in the outcome […] when class, race, ethnicity, gender, sexuality, and age define different social locations, one subject cannot fully empathize with another in a different social location, adopt her point of view; if that were possible then the social locations would not be different. (Young 1990, pp. 104-5; cf. Benhabib 1987, p. 90)

A similar complaint against Rawls is lodged by Shane O'Neill, who writes that Rawls's isolation of political from personal morality results in a theory in which certain contentious moral issues are resolved essentially by fiat: since the moral views of certain contesting parties are defined as personal rather than political matters, they have no chance of effectively defending them in the public sphere (O'Neill 1997, Chapter 1).

If genuine impartiality is an illusion, as Young alleges, then impartialists may be suspected in smuggling in their own substantive moral positions and biases under the guise of neutrality. Rawls' use of the veil of ignorance, for example, has been criticized by Thomas Nagel and others on the basis that, by requiring that agents lack knowledge of their conceptions of the good (a necessary stipulation of the bargainers are to achieve a consensus), the veil of ignorance excludes from the original position information that is morally relevant, and indeed may put some of the bargainers at a disadvantage. "The original position," Nagel writes, "seems to presuppose not just a neutral theory of the good, but a liberal, individualistic conception according to which the best that can be wished for someone is the unimpeded pursuit of his own path, provided it does not interfere with the rights of others" (Nagel 1973; see also. Teitelman 1972; Schwartz 1973; Sandel 1982; Benhabib 1987). Such a conception, it is held, clearly does favor some conceptions of the good over others: in particular, atomic, individualistic conceptions focusing on personal fulfillment (constituted, perhaps, through the acquisition of consumer goods) are privileged over more communal or social ideals that focus on solidarity and mutual interaction between persons (Sandel 1982; cf. O'Neill 1997, Chapter 1).

Feminist critics have paid particular attention to the ways in which liberal conceptions of neutrality and impartiality presuppose and reinforce traditional male-dominated, individualistic approaches to moral theory, and in doing so reinforce the social status quo (Gilligan 1982; Noddings 1984; Benhabib 1987; Young 1990). As Benhabib has pointed out, "Universalistic moral theories in the Western tradition from Hobbes to Rawls are *substitutionalist*, in the sense that the universalism they defend is defined surreptitiously by identifying the experiences of a specific group of subjects as the paradigmatic case of the human as such. These subjects are invariably white, male adults who are propertied or at least professional." (Benhabib 1987, p. 81) As a result, the dominant social positions of such parties tend to be protected and even enhanced in the social and political theories resulting from such allegedly neutral liberal theories.

Such criticisms need not lead us to reject the idea of a moral role for impartiality altogether. Rather, many of the critics considered have thought that there were better models available for understanding impartiality. Many of these models are guided by the thought that the impartial observers, actors or paradigm agents need to be humanized in various ways: by allowing them to retain their conceptions of the good (to a certain extent at least) behind the veil of ignorance; by abandoning totalizing urge that leads us to think of all impartial decisions as being made by essentially one, perfectly abstract agent; and by reducing the cognitive requirements for such agents. Thus, rather than picturing impartial judgments as being made by a single person who is both omniscient and entirely disinterested, we might think of them as the result of a dialogical process between a group of quite distinct participants -- a process that is likely to involve considerably more conflict than the smooth procedure envisioned by Rawls. The general challenge facing such accounts, of course, is obvious: such a picture carried to its logical extreme would simply be identical with the actual world, in which various groups contend with one another in defense of their interests and conceptions of the good; and it is, of course, a salient feature of the actual world that agreement is extremely hard to reach. Thus, it may be worried that while humanizing models of the sort proposed will be able to avoid the difficulties that tend to plague liberal theorists who conceive of impartiality as abstraction, they will only do so by abandoning the primary attraction of such theories: their ability to derive order out of social chaos. Whether a compromise can be found that avoids both horns of this dilemma remains to be seen.

6. The partialist-impartialist debate

Although many people continue to speak of a ‘partialist vs. impartialist debate,’ it should by now be clear that neither ‘partialism’ nor ‘impartialism’ unambiguously denote any single moral position; at best, they designate two poles of a continuum, one of which attributes no moral significance to the demands of (any sort of) impartiality, the other of which sees morality as exhausted by (some version of) impartiality. While a somewhat general distinction can be usefully maintained, it is misleading to think of the partialist-impartialist debate as a dispute between two clearly defined, and clearly opposed, camps (Deigh 1991; Barry 1995, pp. 191-5).

Thus, any general claim beginning with the words ‘partialists (or impartialists) think that …’ is bound to

be both misleading and contentious. In particular, there is good reason to be wary of objections to impartialism which claim that all impartialists endorse extreme moral demands, or that they require that practical reasoning be completely expunged of every vestige of the partial. It is true, of course, that at least some impartialists, such as Godwin, have endorsed such claims. But many do not. Deontologists, as we have seen, hold impartiality to be a deep and significant element of morality, but they also tend to allow for a considerable degree of first-order partiality. And even many consequentialists are prepared to admit the legitimacy of partial reasoning in some contexts, if only on an instrumental basis. It is useful, then, to draw a distinction between two sorts of impartialist moral theory. Impartialist theories which require all agents to display first-order impartiality at all times (Godwin's, for example) might be referred to as *strict impartialist theories.* Impartialist theories which allow for some first-order partiality, but which nevertheless insist that all such behavior be justified in second-order impartialist terms, might be referred to as *fundamentally impartialist* moral theories. The class of fundamentally impartial theories will include not only contractualist, Kantian, and rule consequentialist theories, but also certain act consequentialist theories (e.g. Railton 1986) which allow the practice of first-order partiality as a means of promoting the impersonal good. Such theories allow for partiality that is permissible, justifiable, and perhaps even admirable in moral terms. At the same time, however, they insist that all such partiality is ultimately reducible -- that is, justifiable in impartialist terms at some deeper level.

Within the partialist camp, a *strict partialist* might be defined as holding that no sort of impartiality plays any moral role whatsoever -- a logically possible, but uncommon, position. A *moderate partialist*, by contrast, would admit that impartiality of some sort plays a moral role, but deny that this role encompasses, or grounds, all of morality; in particular, such a figure would be committed to the existence, in some contexts at least, of irreducible morally admirable partiality. A virtue theorist, for instance, might make a significant place for impartiality by selecting it as one of the virtues; but a virtue of this sort would presumably have to compete with other deeply partialist virtues such as loyalty, which would override impartiality in at least some contexts.

To the extent that a deep issue between partialists and impartialists can be identified, it is presumably the question of whether (irreducible) morally admirable partiality does indeed exist; and it is along this line of dispute that the debate seems likeliest to continue. (Whether this debate is identical to the so-called 'justice-care' debate, as contended in Cannold, *et al* (1995), is questionable, though it is undeniable that there are important parallels.) However, this way of classifying the disputants, and of characterizing the issue itself, is meant to be suggestive rather than definitive. The fact remains that there are many types of partialist theories, and many types of impartialist ones, and that continuing to speak of the 'partialist-impartialist debate' in loose and imprecise terms is more likely to obscure than to illuminate.

Bibliography

- Archard, David. 1995. "Moral Partiality." *Midwest Studies in Philosophy* XX: 129-141.
- Ashford, Elizabeth. 2000. "Utilitarianism, Integrity, and Partiality." *The Journal of Philosophy* XCVII(8): 421-39.
- Baier, Kurt. 1958. *The Moral Point of View: A Rational Basis of Ethics.* Cornell University Press.

- Bales, R. Eugene. 1971. "Act-Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedure?" *American Philosophical Quarterly* 8(3): 257-65.
- Baron, Marcia. 1991. "Impartiality and Friendship." *Ethics* 101: 836-57.
- Baron, Marcia. 1995. *Kantian Ethics Almost Without Apology*. Cornell University Press.
- Baron, Marcia, Philip Pettit, and Michael Slote. 1997. *Three Methods of Ethics*. Blackwell.
- Barry, Brian. 1989. *Theories of Justice.* University of California Press.
- Barry, Brian. 1995. *Justice as Impartiality*. Oxford University Press.
- Benhabib, Seyla. 1987. "The Generalized and the Concrete Other: The Kohlberg-Gilligan Controversy and Feminist Theory." In Benhabib and Cornell 1987.
- Benhabib, Seyla, and Drucilla Cornell (eds.) 1987. *Feminism as Critique.* The Polity Press (Cambridge).
- Blum, Lawrence. 1980. *Friendship, Altruism, and Morality*. London: Routledge and Kegan Paul.
- Brandt, Richard. 1954. "The Definition of an 'Ideal Observer' in Ethics." *Philosophy and Phenomenological Research* 15: 407-13.
- Brandt, Richard. 1979. *A Theory of the Good and the Right.* Oxford University Press.
- Brink, David O. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge University Press.
- Broad, C.D. 1959. *Five Types of Ethical Theory*. Paterson, N.J.: Littlefield, Adams & Co.
- Cannold, Leslie, Peter Singer, Helga Kuhse, and Lori Gruen. 1995. "What is the Justice-Care Debate *Really* About?" *Midwest Studies in Philosophy XX*: 357-75.
- Cottingham, John. 1983. "Ethics and Impartiality." *Philosophical Studies* 43: 83-99.
- Cottingham, John. 1986. "Partiality, Favoritism, and Morality," *Philosophical Quarterly* 36: 357-73.
- Cottingham, John. 1996. "Partiality and the Virtues." In Roger Crisp, ed., *How Should One Live? Essays on the Virtues*. Oxford: Clarendon Press, 1996.
- Darwall, Stephen L. 1983. *Impartial Reason.* Cornell University Press.
- Deigh, John. 1991. "Impartiality: A Closing Note." *Ethics* 101: 858-864.
- Double, Richard. "Morality, Impartiality, and What We Can Ask of Persons." *American Philosophical Quarterly* 36(2), April 1999, pp. 149-158.
- Dworkin, Ronald. 1977. *Taking Rights Seriously*. Harvard University Press.
- Firth, Roderick. 1952. "Ethical Absolutism and the Ideal Observer." *Philosophy and Phenomenological Research* 12(3): 317-345.
- Flanagan, Owen, and Jonathan Alder. 1983. "Impartiality and Particularity." *Social Research* L, 3: 576-596.
- Friedman, Marilyn. 1989."The Impracticality of Impartiality," *Journal of Philosophy* 86: 645-56.
- Gauthier, David. 1986. *Morals by Agreement.* Oxford University Press.
- Gert, Bernard, 1998. *Morality: Its Nature and Justification*. Oxford University Press.
- Gert, Bernard. 1995. "Moral Impartiality." *Midwest Studies in Philosophy* XX: 102-127.
- Gewirth, Alan. 1978. *Reason and Morality*. University of Chicago Press.
- Gilligan, Carol. 1982. *In a Different Voice: Psychological Theory and Women's Development.* Harvard University Press.
- Godwin, William. 1926 [1793]. *Enquiry Concerning Political Justice and its Influence on General Virtue and Happiness*, ed. Raymond Preston. New York.
- Godwin, William. 1968 [1801]. *Thoughts Occasioned by the Perusal of Dr. Parr's Spital Sermon*.

In *Uncollected Writings (1785-1822) by William Godwin,* ed. J. Marken and B. Pollin. Scholars' Facsimiles & Reprints (Gainesville, Florida).

- Hare, R.M. 1981. *Moral Thinking*. Oxford University Press.
- Harsanyi, John C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge University Press.
- Harsanyi, John C. 1982. "Morality and the Theory of Rational Behavior." In Sen and Williams, 1982.
- Henberg, M.C. 1978. "Impartiality." *Canadian Journal of Philosophy* 8(4): 715-724.
- Herman, Barbara. 1993. *The Practice of Moral Judgment*. Harvard University Press.
- Hooker, Brad. 1994. "Is Rule-Consequentialism a Rubber Duck?" *Analysis* 54.2: pp. 92-97.
- Howard-Snyder, Frances. 1993. "Rule Consequentialism Is a Rubber Duck." *American Philosophical Quarterly* 30 (1993): 271-78.
- Hume, David. 1978 [1740]. *A Treatise of Human Nature*. Second edition, ed. L.A. Selby-Bigge & P.H. Nidditch. Oxford University Press.
- Jeske, Diane, and Richard Fumerton. 1997. "Relatives and Relativism." *Philosophical Studies* 87: 143-57.
- Jollimore, Troy A. 2000. "Friendship Without Partiality?" *Ratio* 13(1): 69-82.
- Jollimore, Troy A. 2001. *Friendship and Agent-Relative Morality*. Garland Publishing.
- Kagan, Shelley. 1989. *The Limits of Morality.* Oxford: Clarendon Press.
- Kant, Immanuel. 1964 [1785]. *Groundwork of the Metaphysics of Morals.* Translated by H.J. Paton. Harper and Row.
- Kapur, Neera Badwhar. 1991. "Why It Is Wrong to be Always Guided by the Best: Consequentialism and Friendship." *Ethics* 101: 483-504.
- Kekes, John. 1981. "Morality and Impartiality." *American Philosophical Quarterly* 18: 295-303.
- Kohlberg, Lawrence. 1979. "Justice as Reversibility." In P. Laslett and J. Fishkin, ed., *Philosophy, Politics and Society,* Firth Series. Blackwell.
- Korsgaard, Christine. 1996. *Creating the Kingdom of Ends.* Cambridge University Press.
- Locke, Don. 1981. "The Principle of Equal Interests." *Philosophical Review* 90: 531-59.
- MacIntyre, Alasdair. 1984. "Is Patriotism a Virtue?" University of Kansas: The Lindley Lecture Series.
- McCloskey, H.J. 1963. "A Note on Utilitarian Punishment." *Mind* 72.
- McNaughton, David, and Piers Rawling. 1992. "Honoring and Promoting Values." *Ethics* 102: 835-43.
- McNaughton, David, and Piers Rawling. 1993. "Deontology and Agency." *The Monist* 76: 81-100.
- McNaughton, David, and Piers Rawling. 1998. "On Defending Deontology." *Ratio,* 11(1): 37-54.
- Mill, J.S. 1992 [1861]. *Utilitarianism.* In *On Liberty and Utilitarianism.* Knopf: Everyman's Library, Vol. 81.
- Miller, Richard W. 1992. *Moral Differences.* Princeton University Press.
- Monro, D.H. 1950. "Archbishop Fenelon versus My Mother." *Australasian Journal of Philosophy,* xxvii.
- Nagel, Thomas. 1973. "Rawls on Justice." *Philosophical Review* 82. Reprinted in Norman Daniels, ed., *Reading Rawls*. Blackwell, 1975.
- Nagel, Thomas. 1986. *The View from Nowhere.* New York: Oxford University Press.

- Nagel, Thomas. 1991. *Equality and Partiality*. New York: Oxford University Press.
- Noddings, Nel. 1984. *Caring: A Feminine Approach to Ethics and Moral Education*. University of California Press.
- Oldenquist, Andrew. 1982. "Loyalties." *Journal of Philosophy* 79(4): 173-193.
- Okin, Susan Moller. 1989a. *Justice, Gender and the Family*. Basic Books.
- Okin, Susan Moller. 1989b. "Reason and Feeling in Thinking About Justice." *Ethics* 99: 229-49.
- O'Neill, Shane. 1997. *Impartiality in Context.* State University of New York Press.
- Pettit, Philip. 1997. "The Consequentialist Perspective." In Baron, Pettit, and Slote (1997).
- Pettit, Philip. 2000. "Non-consequentialism and Universalizability." *The Philosophical Quarterly* Vol. 50, No. 199: 175-90.
- Pettit, Philip, and Geoffrey Brennan. 1986. "Restrictive Consequentialism." *Australasian Journal of Philosophy* 64(4): 438-55.
- Piper, Adrian. 1990. "Higher-Order Discrimination." In Flanagan and Rorty, ed., *Identity, Character and Morality: Essays in Moral Psychology*. Cambridge: The MIT Press, 1990.
- Piper, Adrian. 1991. "Impartiality, Compassion, and Modal Imagination." *Ethics* 101 (1991).
- Powers, Madison. 1993. "Contractualist Impartiality and Personal Commitments." *American Philosophical Quarterly* 30(1): 63-71.
- Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13: 134-71.
- Rawls, John. 1971. *A Theory of Justice*. Harvard University Press.
- Rawls, John. 1993. *Political Liberalism.* Columbia University Press.
- Rawls, John. 2001. *Justice as Fairness: A Restatement.* Belknap (Harvard).
- Sandel, Michael. 1982. *Liberalism and the Limits of Justice*. Cambridge University Press.
- Scanlon, T.M. 1982. "Contractualism and Utilitarianism." In Sen and Williams, 1982.
- Scanlon, T.M. 1978. "Rights, Goals, and Fairness." In Stuart Hampshire, ed., *Public and Private Morality*. Cambridge University Press.
- Scanlon, T.M. 1998. *What We Owe to Each Other*. Belknap (Harvard).
- Scheffler, Samuel. 1982. *The Rejection of Consequentialism.* Oxford University Press.
- Scheffler, Samuel. 1985. "Agent-Centred Restrictions, Rationality, and the Virtues." *Mind* 94: 409-19.
- Scheffler, Samuel. 1986. "Morality's Demands and Their Limits." *The Journal of Philosophy* 83: 531-37.
- Scheffler, Samuel. 1992. *Human Morality.* Oxford University Press.
- Schwartz, Adina. 1973. "Moral Neutrality and Primary Goods." *Ethics* 83: 294-307.
- Sen, Amartya, and Bernard Williams (eds.) 1982. *Utilitarianism and Beyond*. Cambridge University Press.
- Sidgwick, Henry. 1907. *The Methods of Ethics*. Seventh Edition. Macmillan.
- Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1: 229-43.
- Singer, Peter, Leslie Cannold, and Helga Kuhse. 1995. "William Godwin and the Defense of Impartialist Ethics." *Utilitas* 7 (1): 67-86.
- Slote, Michael. 1985. *Common Sense Morality and Consequentialism.* Boston: Routledge and Kegan Paul.
- Smart, J.J.C. 1973. "An Outline of a System of Utilitarian Ethics." In Smart and Williams 1973.

- Smart, J.J.C, and Bernard Williams. 1973. *Utilitarianism: For and Against*. Cambridge University Press.
- Smith, Adam. 1976 [1759]. *Theory of the Moral Sentiments.* Oxford University Press.
- Stocker, Michael. 1976. "The Schizophrenia of Modern Ethical Theories." *The Journal of Philosophy* 73: 453-66.
- Teitelman, Michael. 1972. "The Limits of Individualism." *Journal of Philosophy* 69: 545-56.
- Walker, Margaret Urban. 1991. "Partial Consideration." *Ethics* 101: 758-74.
- Wiggins, David. 1978. "Universalizability, Impartiality, Truth." In *Needs Values, Truth*. Oxford University Press.
- Williams, Bernard. 1973. "A Critique of Utilitarianism." In Smart and Williams 1973.
- Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. Harvard University Press.
- Wolf, Susan. 1982. "Moral Saints." *Journal of Philosophy* 89: 419-39.
- Wolf, Susan. 1992. "Morality and Partiality." *Philosophical Perspectives* 6: 243-259.
- Young, Iris Marion. 1987. "Impartiality and the Civic Public: Some Implications of Feminist Critiques of Moral and Political Theory." In Benhabib and Cornell, 1987.
- Young, Iris Marion. 1990. *Justice and the Politics of Difference.* Princeton University Press.

Other Internet Resources

- William Godwin: Enquiry Concerning Political Justice
- Kant on the Web
- Utilitarianism.com: John Stuart Mill

Related Entries

consequentialism | equality | ethics: deontological | Godwin, William | Kant, Immanuel: moral philosophy | Mill, John Stuart | original position

Troy Jollimore
tjollimore@csuchico.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 25, 2002
Content last modified: March 25, 2002

Equality

This article is concerned with social and political equality. In its prescriptive usage, ‘equality’ is a loaded and ‘highly contested’ concept. On account of its normally positive connotation, it has a rhetorical power rendering it suitable as a political slogan (Westen 1990). At least since the French Revolution, equality has served as one of the leading ideals of the body politic; in this respect, it is at present probably the most controversial of the great social ideals. There is controversy concerning the precise notion of equality, the relation of justice and equality (the principles of equality), the material requirements and measure of the ideal of equality (equality of what?), the extension of equality (equality among whom?), and its status within a comprehensive (liberal) theory of justice (the value of equality). Each of these five issues will be discussed by turn in the present article.

1. Defining the Concept

‘Equality’ is a contested concept: "People who praise it or disparage it disagree about what they are praising or disparaging" (Dworkin 2000, p. 2). Our first task is therefore to provide a clear definition of equality in the face of widespread misconceptions about its meaning as a political idea.

The terms "equality" (Gr. isotes, Lat. aequitas, aequalitas, Fr. égalité, Ger. Gleichheit), "equal," and "equally" signify a qualitative relationship. ‘Equality’ (or ‘equal’) signifies correspondence between a group of different objects, persons, processes or circumstances that have the same qualities in at least one respect, but not all respects, i.e., regarding one specific feature, with differences in other features. ‘Equality’ needs to thus be distinguished from ‘identity’ -- this concept signifying that one and the same

object corresponds to itself in all its features: an object that can be referred to through various individual terms, proper names, or descriptions. For the same reason, it needs to be distinguished from ‘similarity’ -- the concept of merely approximate correspondence (Dann 1975, p. 997; Menne 1962, p. 44 ff.; Westen 1990, pp. 39, 120). Thus, to say e.g. that men are equal is not to say that they are identical. Equality rather implies similarity but not ‘sameness.’

In distinction to numerical identity, a judgment of equality presumes a difference between the things being compared. According to this definition, the notion of ‘complete’ or ‘absolute’ equality is self-contradictory. Two non-identical objects are never completely equal; they are different at least in their spatiotemporal location. If things do not differ they should not be called ‘equal,’ but rather, more precisely, ‘identical,’ as e.g., the morning and evening star. Here usage might vary. Some authors do consider absolute qualitative equality admissible as a borderline concept (Tugendhat & Wolf 1983, p. 170).

‘Equality’ can be used in the very same sense both to describe and prescribe, as with "thin": "you are thin" and "you are too thin." The approach taken to defining the standard of comparison for both descriptive and prescriptive assertions of the concept of equality is very important (Oppenheim 1970). In the case of *descriptive* use of equality, the common standard is itself descriptive, e.g. two people weigh the same. A *prescriptive* use of equality is present when a prescriptive standard is applied, i.e., a norm or rule, e.g. people ought to be equal before the law. The standards grounding prescriptive assertions of equality contain at least two components. On the one hand, there is a descriptive component, since the assertions need to contain descriptive criteria, in order to identify those people to which the rule or norm applies. The question of this identification -- who belongs to which category? -- may itself be normative, e.g. to whom do the U.S. laws apply? On the other hand, the comparative standards contain something normative -- a moral or legal rule, in the example, the U.S. laws -- specifying how those falling under the norm are to be treated. Such a rule constitutes the prescriptive component (Westen 1990, chap. 3). Sociological and economic analyses of (in-)equality mainly pose the questions of how inequalities can be determined and measured and what their causes and effects are. In contrast, social and political philosophy is in general concerned mainly with the following questions: *what kind* of equality, if any, should be offered, and to *whom* and *when*? Such is the case in this article as well.

‘Equality’ and ‘equal’ are incomplete predicates that necessarily generate one question: equal in what respect? (Rae 1980, p. 132 f.) Equality essentially consists of a tripartite relation between two (or several) objects or persons and one (or several) qualities. Two objects a and b are equal in a certain respect if, in that respect, they fall under the same general terminus. ‘Equality’ denotes the relation between the objects that are compared. Every comparison presumes a *tertium comparationis*, a concrete attribute defining the respect in which the equality applies -- equality thus referring to a common sharing of this comparison-determining attribute. This relevant comparative standard represents a ‘variable’ (or ‘index’) of the concept of equality that needs to be specified in each particular case (Westen 1990, p. 10); differing conceptions of equality here emerge from one or another descriptive or normative moral standard. There is another source of diversity as well: As Temkin (1986, 1993) argues, various different standards might be used to measure inequality, with the respect in which people are compared remaining constant. The difference between a general concept and different specific conceptions (Rawls 1971, p. 21 f.) of equality

may explain why according to various authors producing ‘equality’ has no unified meaning -- or even is devoid of meaning. (Rae 1981, p. 127 f., 132 f.)

For this reason, it helps to think of the idea of equality or for that matter inequality, understood as an issue of social justice, not as a single principle, but as a complex group of principles forming the basic core of today's egalitarianism. Depending on which procedural principle one adopts, contrary answers are forthcoming. Both equality and inequality are complex and multifaceted concepts (Temkin 1993, chap. 2). In any real historical context, it is clear that no single notion of equality can sweep the field. (Rae 1981, p. 132) Many egalitarians concede that much of our discussion of the concept is vague and theoretical. But they believe that there is also a common underlying strain of important moral concerns implicit in it (Williams 1973). Above all it serves to remind us of our common humanity, despite various differences (cf. 2.3. below). In this sense, egalitarians tend to think of egalitarianism as a single coherent normative doctrine -- but one in any case embracing a variety of principles. Following the introduction of different principles and theories of equality, I will return in the last section of this article to the question how best to define egalitarianism and the value of equality.

2. Principles of Equality and Justice

Equality in its prescriptive usage has, of course, a close connection with morality and justice in general and distributive justice in particular. From antiquity onward, equality has been considered a constitutive feature of justice. (On the history of the concept, cf. Albernethy 1959, Benn 1967, Brown 1988, Dann 1975, Thomson 1949.) Throughout history, people and emancipatory movements use the language of justice to pillory certain inequalities. But what exactly is the connection between equality and justice, i.e., what kind of role does equality play in a theory of justice? The role and correct account of equality, understood as an issue of social justice, is itself a difficult philosophical issue. To clarify this, philosophers have defended a variety of principles and conceptions of equality, many of which are mentioned in the following discussion. This section introduces four well known principles of equality, ranging from highly general and uncontroversial to more specific and controversial. The next section reviews various conceptions of the ‘currency’ of equality. Different interpretations of the role of equality in a theory of justice emerge according to which of the four following principles and which measure has been adopted.

Through its connection with justice, equality, like justice itself, has different *justitianda*, i.e., objects the term ‘just’ or ‘equal’ or their opposites can be applied to. These are mainly actions, persons, social institutions, and circumstances (e.g. distributions). These objects of justice stand in an internal connection and order that can here only be hinted at. The predicates "just" or "unjust" are only applicable when *voluntary* actions implying responsibility are in question. Justice is hence primarily related to individual actions. Individual persons are the primary bearer of responsibilities (ethical individualism). Persons have to take responsibility for their individual actions and for circumstances they could change through such actions or omissions. Although people have responsibility for both their actions and circumstances, there is a moral difference between the two *justitianda*, i.e., an injustice due to unjust treatment through an individual or collective action and an injustice due to a failure to correct unjust circumstances (cf. 3.1.v.

below). The responsibility people have to treat individuals and groups they affect in a morally appropriate and, in particular, even-handed way has hence a certain priority over their moral duty to turn circumstances into just ones through some kind of equalization. Establishing justice of circumstances (ubiquitously and simultaneously) is beyond any given individual's capacities. Hence one has to rely on collective actions. In order to meet this moral duty, a basic order guaranteeing just circumstances must be justly created. This is an essential argument of justice in favor of establishing social institutions and fundamental state structures for political communities; with the help of such institutions and structures, individuals can collectively fulfill their responsibility in the best possible manner. If circumstances can be rightly judged to be unjust, all persons have the responsibility and moral duty, both individually and collectively, to change the pertinent circumstances or distributive schemes into just ones. In the following sections, the objects of equality may vary from topic to topic. However, as indicated, there is a close relationship between the objects. The next three principles of equality hold generally and primarily for all actions and treatment of others and for resulting circumstances. From the fourth principle onward, i.e., starting with the presumption of equality, this article is mainly concerned with distributive justice and the evaluation of distribution.

2.1 Formal Equality

When two persons have equal status in at least one normatively relevant respect, they must be treated equally with regard to this respect.This is the generally accepted *formal* equality principle that Aristotle formulated in reference to Plato: "treat like cases as like" (Aristotle, *Nicomachean Ethics*, V.3. 1131a10-b15; *Politics*, III.9.1280 a8-15, III. 12. 1282b18-23). Of course the crucial question is which respects are normatively relevant and which are not. Some authors see this formal principle of equality as a specific application of a rule of rationality: it is irrational, because inconsistent, to treat equal cases unequally without sufficient reasons (Berlin 1955-56). But most authors instead stress that what is here at stake is a moral principle of justice, basically corresponding with acknowledgment of the impartial and universalizable nature of moral judgments. Namely, the postulate of formal equality demands more than consistency with one's subjective preferences. What is more important is possible justification vis-à-vis others of the equal or unequal treatment in question -- and this on the sole basis of a situation's objective features.

2.2 Proportional Equality

According to Aristotle, there are two kinds of equality, numerical and proportional (Aristotle, *Nicomachean Ethics*, 1130b-1132b; cf. Plato, *Laws*, VI.757b-c). A form of treatment of others or as a result of it a distribution is equal *numerically* when it treats all persons as indistinguishable, thus treating them identically or granting them the same quantity of a good per capita. That is not always just. In contrast, a form of treatment of others or distribution is *proportional* or relatively equal when it treats all relevant persons in relation to their due. Just *numerical* equality is a special case of proportional equality. Numerical equality is only just under special circumstances, viz. when persons are equal in the relevant respects so that the relevant proportions are equal. Proportional equality further specifies formal equality; it is the more precise and detailed, hence actually the more comprehensive formulation of formal equality.

It indicates what produces an adequate equality.

Proportional equality in the treatment and distribution of goods to persons involves at least the following concepts or variables: Two or more persons (P_1, P_2) and two or more allocations of goods to persons (*G*) and *X* and *Y* as the quantity in which individuals have the relevant normative quality *E*. This can be represented as an equation with fractions or as a ratio. If P_1 has *E* in the amount of *X* and if P_2 has *E* in the amount *Y*, then *P*1 is due *G* in the amount of X' and P_2 is due *G* in the amount of Y', so that the ratio $X/Y = X'/Y'$ is valid. (N.B. For the formula to be usable, the potentially great variety of factors involved have to be both quantifiable in principle and commensurable, i.e., capable of synthesis into an aggregate value.)

When factors speak for unequal treatment or distribution, because the persons are unequal in relevant respects, the treatment or distribution proportional to these factors is just. Unequal claims to treatment or distribution must be considered proportionally: that is the prerequisite for persons being considered equally.

This principle can also be incorporated into hierarchical, inegalitarian theories. It indicates that equal output is demanded with equal input. Aristocrats, perfectionists, and meritocrats all believe that persons should be assessed according to their differing deserts, understood by them in the broad sense of fulfillment of some relevant criterion. And they believe that reward and punishment, benefits and burdens, should be proportional to such deserts. Since this definition leaves open who is due what, there can be great inequality when it comes to presumed fundamental (natural) rights, deserts, and worth -- and such inequality is apparent in both Plato and Aristotle.

Aristotle's idea of justice as proportional equality contains a fundamental insight. The idea offers a framework for a rational argument between egalitarian and non-egalitarian ideas of justice, its focal point being the question of the basis for an adequate equality (Hinsch, forthcoming 2002). Both sides accept justice as proportional equality. Aristotle's analysis makes clear that the argument involves the features deciding whether two persons are to be considered equal or unequal in a distributive context.

On the formal level of pure conceptual explication, justice and equality are linked through these principles of formal and proportional justice. Justice cannot be explained without these equality principles; the equality principles only receive their normative significance in their role as principles of justice.

Formal and proportional equality is simply a conceptual schema. It needs to be made precise -- i.e., its open variables need to be filled out. The formal postulate remains quite empty as long as it remains unclear when or through what features two or more persons or cases should be considered equal. All debates over the proper conception of justice, i.e., over who is due what, can be understood as controversies over the question of which cases are equal and which unequal (Aristotle, *Politics*, 1282b 22). For this reason equality theorists are correct in stressing that the claim that persons are owed equality becomes informative only when one is told -- what *kind* of equality they are owed (Nagel 1979; Rae

1981; Sen 1992, p. 13). Actually, every normative theory implies a certain notion of equality. In order to outline their position, egalitarians must thus take account of a specific (egalitarian) conception of equality. To do so, they need to identify substantive principles of equality, discussed below.

2.3 Moral Equality

Until the eighteenth century, it was assumed that human beings are unequal by nature -- i.e., that there was a natural human hierarchy. This postulate collapsed with the advent of the idea of natural right and its assumption of an equality of natural order among all human beings. Against Plato and Aristotle, the classical formula for justice according to which an action is just when it offers each individual his or her due took on a substantively egalitarian meaning in the course of time, viz. everyone deserved the same dignity and the same respect. This is now the widely held conception of substantive, universal, moral equality. It developed among the Stoics, who emphasized the natural equality of all rational beings, and in early New Testament Christianity, which elevated the equality of human beings before God to a principle: one to be sure not always adhered to later by the Christian church. This important idea was also taken up both in the Talmud and in Islam, where it was grounded in both Greek and Hebraic elements in both systems. In the modern period, starting in the seventeenth century, the dominant idea was of natural equality in the tradition of natural law and social contract theory. Hobbes (1651) postulated that in their natural condition, individuals possess equal rights, because over time they have the same capacity to do each other harm. Locke (1690) argued that all human beings have the same natural right to both (self-)ownership and freedom. Rousseau (1755) declared social inequality to be a virtually primeval decline of the human race from natural equality in a harmonious state of nature: a decline catalyzed by the human urge for perfection, property and possessions (Dahrendorf 1962). For Rousseau (1755, 1762), the resulting inequality and rule of violence can only be overcome by tying unfettered subjectivity to a common civil existence and popular sovereignty. In Kant's moral philosophy (1785), the categorical imperative formulates the equality postulate of universal human worth. His transcendental and philosophical reflections on autonomy and self-legislation lead to a recognition of the same freedom for all rational beings as the sole principle of human rights (Kant 1797, p. 230). Such Enlightenment ideas stimulated the great modern social movements and revolutions, and were taken up in modern constitutions and declarations of human rights. During the French Revolution, equality -- along with freedom and fraternity -- became a basis of the *Déclaration des droits de l´homme et du citoyen* of 1789.

The principle of equal dignity and respect is now accepted as a minimum standard throughout mainstream Western culture. Some misunderstandings regarding moral equality need to be clarified. To say that men are equal is not to say they are identical. The postulate of equality implies that underneath apparent differences, certain recognizable entities or units exist that, by dint of being units, can be said to be ‘equal.’ (Thomson 1949, p. 4). Fundamental equality means that persons are alike in important relevant and specified respects alone, and not that they are all generally the same or can be treated in the same way (Nagel 1991). In a now commonly posed distinction, stemming from Dworkin (1977, p. 370), moral equality can be understood as prescribing treatment of persons as equals, i.e., with equal concern and respect, and not the often implausible principle of treating persons equally. This fundamental idea of equal respect for all persons and of the equal worth or equal dignity of all human beings (Vlastos 1962) is accepted as a minimal standard by all leading schools of modern Western political and moral culture. Any

political theory abandoning this notion of equality will not be found plausible today. In a period in which metaphysical, religious and traditional views have lost their general plausibility (Habermas 1983, p. 53, 1992, pp. 39-44), it appears impossible to peacefully reach a general agreement on common political aims without accepting that persons must be treated as equals. As a result, moral equality constitutes the 'egalitarian plateau' for all contemporary political theories (Kymlicka 1990, p.5). To recognize that human beings are all equally individual does not mean having to treat them uniformly in any respects other than those in which they clearly have a moral claim to be treated alike. Disputes arise, of course, concerning what these claims amount to and how they should be resolved. That is the crux of the problem to which I now turn.

Since "treatment as an equal" is a shared moral standard in contemporary theory, present-day philosophical debates are concerned with the kind of equal treatment normatively required when we mutually consider ourselves persons with equal dignity. The principle of moral equality is too abstract and needs to be made concrete if we are to arrive at a clear moral standard. Nevertheless, no conception of just equality can be deduced from the notion of moral equality. Rather, we find competing philosophical conceptions of equal treatment serving as interpretations of moral equality. These need to be assessed according to their degree of fidelity to the deeper ideal of moral equality (Kymlicka 1990, p. 44). With this we finally switch the object of equality from treatment to the fair distribution of goods and ills or bads.

2.4 Presumption of Equality

Many conceptions of equality operate along procedural lines involving a *presumption of equality*. While more materially concrete, ethical approaches, as described in the next section below, are concerned with distributive criteria; the presumption of equality, in contrast, is a formal, procedural principle of construction located on a higher formal and argumentative level. What is here at stake is the question of the principle with which a material conception of justice should be constructed -- particularly once the above-described approaches turn out inadequate. The presumption of equality is a prima facie principle of equal distribution for all goods politically suited for the process of public distribution. In the domain of political justice, all members of a given community, taken together as a collective body, have to decide centrally on the fair distribution of social goods, as well as on the distribution's fair realization. Any claim to a particular distribution, including any existing distributive scheme, has to be impartially justified, i.e., no ownership will be recognized without justification. Applied to this political domain, the presumption of equality requires that everyone, regardless of differences, should get an equal share in the distribution unless certain types of differences are relevant and justify, through universally acceptable reasons, unequal distribution. (With different terms and arguments, this principle is conceived as a presumption by Benn & Peters (1959, 111) and by Bedau (1967, 19); as a relevant reasons approach by Williams (1973); as a conception of symmetry by. Tugendhat (1993, 374; 1997, chap. 3); as default option by Hinsch (2001, chap. 5); for criticism of the presumption of equality, cf. Westen (1990, chap. 10).) This presumption results in a principle of prima facie equal distribution for all distributable goods. A strict principle of equal distribution is not required, but it is morally necessary to justify impartially any unequal distribution. The burden of proof lies on the side of those who favor any form of unequal distribution.

The presumption in favor of equality can be justified by the principle of equal respect together with the requirement of universal and reciprocal justification; that requirement is linked to the morality of equal respect granting each individual equal consideration in every justification and distribution. Every sort of public, political distribution is, in this view, to be justified to all relevantly concerned persons, such that they could in principle agree. Since it is immoral to force someone to do something of which he or she does not approve, only reasons acceptable to the other person can give one the moral right to treat the person in accordance with these reasons. The impartial justification of norms rests on the reciprocity and universality of the reasons. Universal norms and rights enforced through inner or external sanctions are morally justified only if, on the one hand, they can be reciprocally justified, i.e., if one person asks no more of the other than what he or she is willing to give (reciprocity), and if, on the other hand, they are justified with respect to the interests of all concerned parties, i.e., if everyone has good reasons for accepting them and no one has a good reason for rejecting them (universality) (Forst 1994, p. 68, Scanlon 1998). In the end, only the concerned parties can themselves formulate and advocate their (true) interests. Equal respect, which we reciprocally owe to one another, thus requires respect for the autonomous decisions of each non-interchangeable individual (Wingert 1993, p. 90-96). This procedural approach to moral legitimation sees the autonomy of the individual as the standard of justification for universal rules, norms, rights etc. Only those rules can be considered legitimate to which all concerned parties can freely agree on the basis of universal, discursively applicable, commonly shared reasons. Equal consideration is thus accorded to all persons and their interests. In a public distribution anyone who claims more owes all others an adequate universal and reciprocal justification. If this cannot be provided, i.e., if there is no reason for unequal distribution that can be universally and reciprocally recognized by all (since, lets assume, all are by and large equally productive and needy), then equal distribution is the only legitimate distribution. How could it be otherwise? Any unequal distribution would mean that someone receives less, and another more. Whoever receives less can justifiably demand a reason for he or she being disadvantaged. Yet there is *ex hyphothesi* no such justification. Hence, any unequal distribution is illegitimate in this case. If no convincing reasons for unequal distribution can be brought forward, there remains only the option of equal distribution. Equal distribution is therefore not merely one among many alternatives, but rather the inevitable starting point that must be assumed insofar as one takes the justificatory claims of all to be of equal weight.

The presumption of equality provides an elegant procedure for constructing a theory of distributive justice. The following questions would have to be answered in order to arrive at a substantial and full principle of justice.

- What goods and burdens are to be justly distributed (or should be distributed)? Which social goods comprise the object of distributive justice?
- What are the spheres (of justice) into which these resources have to be grouped?
- Who are the recipients of distribution? Who has a prima facie claim to a fair share?
- What are the commonly cited yet in reality unjustified exceptions to equal distribution?
- Which inequalities are justified?
- Which approach, conception or theory of egalitarian distributive justice is therefore the best?

What goods and burdens are to be justly distributed (or should be distributed)? There are various opinions as to which social goods comprise the object of distributive justice. Does distributive justice apply only to those goods commonly produced, i.e., through social and economic fair cooperation, or to other goods as well, e.g. natural resources, that are not the result of common cooperation? (At present, the former approach is most apparent in Rawls (1971) and many of his adherents and critics follow Rawls in this respect.)

In the domain of public political distribution, the goods and burdens to be distributed may be divided into various categories. Such a division is essential because reasons that speak for unequal treatment in one area do not justify unequal treatment in another. What are the spheres (of justice) into which these resources have to be grouped? In order to reconstruct our understanding of contemporary liberal, democratic welfare states, four categories seem essential: 1. civil liberties, 2. opportunities for political participation, 3. social positions and opportunities, 4. economic rewards. Despite views to the contrary, liberties and opportunities are seen in this view as objects of distribution. For all four categories, the presumption of equality is the guiding principle. The results of applying the presumption to each category can then be codified as rights.

After dividing social goods into categories, we must next ask what can justify unequal treatment or unequal distribution in each category. Today the following postulates of equality are generally considered morally required.

Strict equality is called for in the legal sphere of civil freedoms, since -- putting aside limitation on freedom as punishment -- there is no justification for any exceptions. As follows from the principle of formal equality, all citizens of a society must have equal general rights and duties. These rights and duties have to be grounded in general laws applying to everyone. This is the postulate of legal equality. In addition, the postulate of equal freedom is equally valid: every person should have the same freedom to structure his or her life, and this in the most far-reaching manner possible in a peaceful and appropriate social order.

In the political sphere, the possibilities for political participation should be equally distributed. All citizens have the same claim to participation in forming public opinion, and in the distribution, control, and exercise of political power. This is the postulate -- requiring equal opportunity -- of equal political power sharing. To ensure equal opportunity, social institutions have to be designed in such a way that persons who are disadvantaged, e.g. have a stutter or a low income, have an equal chance to make their views known and to participate fully in the democratic process.

In the social sphere, social positions, equally gifted and motivated citizens must have approximately the same chances at offices and positions, independent of their economic or social class and native endowments. This is the postulate of fair equality of social opportunity. An unequal outcome has to result from equality of chances at a position, i.e., qualifications alone counting, not social background or influences of milieu.

Since the nineteenth century, the political debate has increasingly centered on the question of economic

and social inequality (this running alongside the question of -- gradually achieved -- equal rights to freedom and political participation) (Marshall 1950). The main controversy here is whether, and if so to what extent, the state should establish far-reaching equality of social conditions for all through political measures such as redistribution of income and property, tax reform, a more equal educational system, social insurance, and positive discrimination.

The equality required in the economic sphere is complex, taking account of several positions that -- each according to the presumption of equality -- justify a turn away from equality. A salient problem here is what constitutes justified exceptions to equal distribution of goods -- the main subfield in the debate over adequate conceptions of distributive equality and its currency. The following sorts of factors are usually considered eligible for justified unequal treatment: (a) need or differing natural disadvantages (e.g. disabilities); (b) existing rights or claims (e.g. private property); (c) differences in the performance of special services (e.g. desert, efforts, or sacrifices); (d) efficiency; and (e) compensation for direct and indirect or structural discrimination (e.g. affirmative action).

These factors play an essential, albeit varied, role in the following alternative egalitarian theories of distributive justice. The following theories offer different accounts of what should be equalized in the economic sphere. Most can be understood as applications of the presumption of equality (whether they explicitly acknowledge it or not); only a few (like strict equality, libertarianism, and sufficiency) are alternatives to the presumption.

3. Conceptions of Distributive Equality: Equality of What?

Every effort to interpret the concept of equality and to apply the principles of equality mentioned above demands a precise measure of the parameters of equality. We need to know the dimensions within which the striving for equality is morally relevant. What follows is a brief review of the seven most prominent conceptions of distributive equality, each offering a different answer to one question: in the field of distributive justice, what should be equalized, or what should be the parameter or "currency" of equality?

3.1 Simple Equality and Objections to Equality in General

Simple equality, meaning everyone being furnished with the same material level of goods and services, represents a strict position as far as distributive justice is concerned. It is generally rejected as untenable.

Hence with the possible exception of Barbeuf (1796), no prominent author or movement has demanded strict equality. Since egalitarianism has come to be widely associated with the demand for economic equality, and this in turn with communistic or socialistic ideas, it is important to stress that neither communism nor socialism -- despite their protest against poverty and exploitation and their demand for social security for all citizens -- calls for absolute economic equality. The orthodox Marxist view of economic equality was expounded in the *Critique of the Gotha Program* (1875). Marx here rejects the

idea of legal equality, on three grounds. In the first place, he indicates, equality draws on a merely limited number of morally relevant vantages and neglects others, thus having unequal effects; right can never be higher than the economic structure and cultural development of the society it conditions. In the second place, theories of justice have concentrated excessively on distribution instead of the basic questions of production. In the third place, a future communist society needs no law and no justice, since social conflicts will have vanished.

As an idea, *simple* equality fails because of problems that are raised in regards to equality in general. It is useful to review these problems, as they require resolution in any plausible approach to equality.

> (i) We need adequate indices for the measurement of the equality of the goods to be distributed. Through what concepts should equality and inequality be understood? It is thus clear that equality of material goods can lead to unequal satisfaction. Money constitutes a usual-index -- although an inadequate one; at the very least, equal opportunity has to be conceived in other terms.
>
> (ii) The time span needs to be indicated for realizing the desired model of equal distribution (McKerlie 1989, Sikora 1989). Should we seek to equalize the goods in question over complete individual lifetimes, or should we seek to ensure that various life segments are as equally well off as possible?
>
> (iii) Equality distorts incentives promoting achievement in the economic field, producing an inefficiency grounded in a waste of assets arising from the administrative costs of redistribution (Okun 1975). Equality and efficiency need to be placed in a balanced relation. Often, pareto-optimality is demanded in this respect -- for the most part by economists. A social condition is pareto-optimal or pareto-efficient when it is not possible to shift to another condition judged better by at least one person and worse by none (Sen 1970, chap. 2, 2*). A widely discussed alternative to the Pareto principle is the Kaldor-Hicks welfare criterion. This stipulates that a rise in social welfare is always present when the benefits accruing through the distribution of value in a society exceed the corresponding costs. A change thus becomes desirable when the winners in such a change could compensate the losers for their losses and still retain a substantial profit. In contrast to the Pareto-criterion, the Kaldor-Hicks criterion contains a compensation rule (Kaldor 1939). For purposes of economic analysis, such theoretical models of optimal efficiency make a great deal of sense. However, the analysis is always made relative to starting situation that can be unjust and unequal. A society can thus be (close to) pareto-optimality -- i.e., no one can increase his or her material goods or freedoms without diminishing those of someone else -- while also displaying enormous inequalities in the distribution of the same goods and freedoms. For this reason, egalitarians claim that it may be necessary to reduce pareto-optimality for the sake of justice if there is no more egalitarian distribution that is also pareto-optimal. In the eyes of their critics, equality of whatever kind should not lead to some people having to do with less even though this equalizing down does not benefit any of those who are in a worse position.

(iv) *Moral objections*: A strict and mechanical equal distribution between all individuals does not sufficiently take into account the differences among individuals and their situations. In essence, since individuals desire different things, why should everyone receive the same? Intuitively, for example, we can recognize that a sick person has other claims than a healthy person, and furnishing each with the same things would be mistaken. With simple equality, personal freedoms are unacceptably limited and distinctive individual qualities insufficiently regarded; in this manner they are in fact unequally regarded. Furthermore, persons not only have a moral right to their own needs being considered, but a right and a duty to take responsibility for their own decisions and their consequences.

Working against the identification of distributive justice with simple equality, a basic postulate of virtually all present-day egalitarians is as follows: human beings are themselves responsible for certain inequalities resulting from their free decisions; aside from minimum aid in emergencies, they deserve no recompense for such inequalities. On the other hand, they are due compensation for inequalities that are not the result of self-chosen options. For egalitarians, the world is morally better when *equality of life conditions* prevail. This is an amorphous ideal demanding further clarification. Why is such equality an ideal, and equality of what, precisely?

By the same token, most egalitarians presently do not advocate an equality of outcome, but different kinds of equality of opportunity, due to their emphasis on a pair of morally central points: firstly, that individuals have responsibility for their decisions; and secondly, that the only things to be considered objects of equality are things serving the real interests of individuals. The opportunities to be equalized between people can be opportunities for well-being (i.e. objective welfare), or for preference satisfaction (i.e., subjective welfare), or for resources. It is not equality of objective or subjective well-being or resources themselves that should be equalized, but an equal opportunity to gain the well-being or resources one aspires to. Such equality of opportunity (to well-being or resources) depends on the presence of a realm of options for each individual equal to the options enjoyed by all other persons, in the sense of the same prospects for fulfillment of preferences or the possession of resources. The opportunity must consist of possibilities one can really take advantage of. Equal opportunity prevails when human beings effectively enjoy equal realms of possibility.

(v) Simple equality is very often associated with equality of results (although these are two distinct concepts). However, to strive only for equality of results is problematic. To illustrate the point, let us briefly limit the discussion to a single action and the event or state of affairs resulting from it. Arguably, actions should not be judged solely by the moral quality of their results as important as this may be. One also has to take into consideration the way in which the events or circumstances to be evaluated have come about. Generally speaking, a moral judgement requires not only the assessment of the results of the action in question (the consequentialist aspect) but, first and foremost, the assessment of the

intention of the actor (the deontological aspect). The source and its moral quality influence the moral judgement of the results (Pogge 1999, sect. V). For example, if you strike me, your blow will hurt me; the pain I feel may be considered bad in itself, but the moral status of your blow will also depend on whether you were (morally) allowed such a gesture (perhaps through parental status, although that is controversial) or even obliged to execute it (e.g. as a police officer preventing me from doing harm to others), or whether it was in fact prohibited but not prevented. What is true of individual actions (or their omission) has to be true *mutatis mutandis* of social institutions and circumstances like distributions resulting from collective social actions (or their omission). Hence social institutions are to be assessed not solely on the basis of information about how they affect individual quality of life. A society in which people starve on the streets is certainly marked by inequality; nevertheless, its moral quality, i.e., whether the society is just or unjust with regard to this problem, also depends on the suffering's causes. Does the society allow starvation as an unintended but tolerable side effect of what its members see as a just distributive scheme? Indeed, does it even defend the suffering as a necessary means, e.g. as a sort of Social Darwinism? Or has the society taken measures against starvation which have turned out insufficient? In the latter case, whether the society has taken such steps for reasons of political morality or efficiency again makes a moral difference. Hence even for egalitarians, equality of results is too narrow and one-sided a focus.

(vi) Finally, there is a danger of (strict) equality leading to uniformity, rather than to a respect for pluralism and democracy (Cohen 1989; Arneson 1993). In the contemporary debate, this complaint has been mainly articulated in feminist and multiculturalist theory. A central tenet of feminist theory is that gender has been and remains a historical variable and internally differentiated relation of domination. The same holds for so called racial and ethnic differences. These differences are often still conceived of as marking different values. The different groups involved here rightly object to their discrimination, marginalization, and domination, and an appeal to equality of status thus seems a solution. However as feminists and multiculturalists have pointed out, equality, as usually understood and practiced, is constituted in part by a denial and ranking of differences; as a result it seems less useful as an antidote to relations of domination. "Equality" can often mean the assimilation to a pre-existing and problematic ‘male’ or ‘white’ or ‘middle class’ norm. In short, domination and a fortiori inequality often arises out of an inability to appreciate and nurture differences -- not out of a failure to see everyone as the same. To recognize these differences should however not lead to an essentialism grounded in sexual or cultural characteristics. In contemporary multiculturalism and feminism, there is a crucial debate between those who insist that sexual, racial, and ethnic differences should become irrelevant, on the one hand, and those believing that such differences, even though culturally relevant, should not furnish a basis for inequality: that rather one should find mechanisms for securing equality, despite valued differences. Neither of these strategies involves rejecting equality. Rather, the dispute is about how equality is to be attained (McKinnon 1989, Taylor 1992).

> Proposing a connection between equality and pluralism, Michael Walzer's theory (1983) aims at what he calls "complex equality". According to Walzer, relevant reasons can only speak in favor of distribution of specific types of goods in specific spheres -- not in several or all spheres. Against a theory of simple equality promoting equal distribution of dominant goods, hence underestimating the complexity of the criteria at work in each given sphere the dominance of particular goods needs to be ended. For instance, purchasing power in the political sphere through means derived from the economic sphere (i.e., money) needs to be prevented. Actually, Walzer's theory of complex equality is not aimed at equality but at the separation of spheres of justice, the theory's designation thus being misleading. Any theory of equality should however follow Walzer's advice not to be monistic but recognize the complexity of life and the plurality of criteria for justice.

We thus arrive at the following desideratum: instead of *simple* equality, we need a concept of more *complex* equality: a concept managing to resolve the above problems through a distinction of various classes of goods, a separation of spheres, and a differentiation of relevant criteria.

3.2 Libertarianism

Libertarianism and *economic liberalism* represent minimalist positions in relation to distributive justice. Citing Locke, they both postulate an original right to freedom and property, thus arguing against redistribution and social rights and for the free market (Nozick 1974; Hayek 1960). They assert an opposition between equality and freedom: the individual (natural) right to freedom can be limited only for the sake of foreign and domestic peace. For this reason, libertarians consider maintaining public order the state's only legitimate duty. They assert a natural right to self-ownership (the philosophical term for "ownership of oneself" -- i.e., one's will, body, work, etc.) that entitles everybody to thus far unowned bits of the external world by means of mixing their labor with it. All individuals can thus claim property if "enough and as good" is left over for others (Locke's proviso). Correspondingly, they defend market freedoms and oppose the use of redistributive taxation schemes for the sake of social justice as equality. A principal objection to libertarian theory is that its interpretation of the Lockean proviso -- nobody's situation should be worsened through an initial acquisition of property -- leads to an excessively weak requirement and is thus unacceptable (Kymlicka 1990, pp.108-117). With a broader and more adequate interpretation of what it means for one a situation to be worse than another, however, justifying private appropriation and, *a fortiori*, all further ownership rights, becomes much more difficult. If the proviso recognizes the full range of interests and alternatives that self-owners have, then it will not generate unrestricted rights over unequal amounts of resources. Another objection is that precisely if one's own free accomplishment is what is meant to count, as the libertarians argue, success should not depend strictly on luck, extraordinary natural gifts, inherited property, and status. In other words, *equal opportunity* also needs to at least be present as a counterbalance, ensuring that the fate of human beings is determined by their decisions and not by unavoidable social circumstances. Equal opportunity thus seems to be the frequently vague minimal formula at work in every egalitarian conception of distributive justice. Many egalitarians, however, wish for more -- namely, an *equality of (at least basic) life conditions*.

In any event, with a shift away from a strictly negative idea of freedom, economic liberalism can indeed

itself point the way to more social and economic equality. For with such a shift, what is at stake is not only assuring an equal right to self-defense, but also furnishing everyone more or less the same chance to actually make use of the right to freedom (e.g. Van Parijs 1995, Steiner 1994). In other words, certain basic goods need to be furnished to assure the equitable or 'fair value of the basic liberties' (Rawls 1993, pp. 356-63).

3.3 Utilitarianism

It is possible to interpret utilitarianism as concretizing moral equality -- and this in a way meant to offer the same consideration to the interests of all human beings (Kymlicka 1990, pp. 31f., Hare 1981, p. 26, Sen 1992, pp. 13f.). From the utilitarian perspective, since everyone counts as one and no one as more than one (Bentham), the interests of all should be treated equally without consideration of contents of interest or an individual's material situation. For utilitarianism this means that all enlightened personal interests have to be fairly aggregated. The morally proper action is the one that maximizes utility (Hare 1984). But this utilitarian conception of equal treatment has been criticized as inadequate by many opponents of utilitarianism. At least in utilitarianism's classical form -- so the critique reads -- the hoped for moral equality is flawed: this because all desires are taken up by the utilitarian calculation -- including "selfish" and "external" preferences (Dworkin 1977, p. 234), all having equal weight, even when they diminish the 'rights' and intentions of others. And this, of course, conflicts with our everyday understanding of equal treatment. What is here at play is an argument involving "offensive" and "expensive" taste: a person cannot expect others to sustain his or her desires at the expense of their own (Kymlicka 1990, p. 40 f.). Rather, according to generally shared conviction, equal treatment consistently requires a basis of equal rights and resources that cannot be taken away from one person, whatever the desire of others. In line with Rawls (1971, pp. 31, 564, cf. 450), many hold that justice entails according no value to interests insofar as they conflict with justice. According to this view, unjustified preferences will not distort mutual claims people have on each other. Equal treatment has to consist of everyone being able to claim a fair portion, and not in all interests having the same weight in disposal over my portion. Utilitarians cannot admit any restrictions on interests based on morals or justice. As long as utilitarian theory lacks a concept of justice and fair allotment, it must fail in its goal of treating all as equals. As Rawls (1971, pp. 27) also famously argues, utilitarianism that involves neglecting the separateness of persons does not contain a proper interpretation of moral equality as equal respect for each individual.

3.4 Equality of Welfare

The concept of welfare equality is motivated by an intuition that when it comes to political ethics, what is at stake is the individual's well-being. The central criterion for justice must consequently be equalizing the level of welfare. But taking welfare as what is to be equalized leads into major difficulties, which resemble those of utilitarianism. If one contentiously identifies subjective welfare with preference satisfaction, it seems implausible to count all individual preferences as equal, some -- such as the desire to do others wrong -- being inadmissible on grounds of justice (the offensive taste argument). Any welfare-centered concept of equality grants people with refined and expensive taste more resources -- something

distinctly at odds with our moral intuitions (the expensive taste argument) (Dworkin 1981a). However, satisfaction in the fulfillment of desires cannot serve as a standard, since we wish for more than a simple feeling of happiness. A more viable standard for welfare comparisons would seem to be success in the fulfillment of preferences. A fair evaluation of such success cannot be purely subjective, rather requiring a standard of what should or could have been achieved. And this itself involves an assumption regarding just distribution; it is thus no independent criterion for justice. An additional serious problem with any welfare-centered concept of equality is that it cannot take account of either desert (Feinberg 1970) or personal responsibility for one's own well-being, to the extent this is possible and reasonable.

3.5 Equality of Resources

Represented above all by both Rawls and Dworkin, resource equality avoids such problems (Rawls 1971; Dworkin 1981b). It holds individuals responsible for their decisions and actions, not, however, for circumstances beyond their control -- race, sex, and skin-color, but also intelligence and social position -- which thus are excluded as distributive criteria. Equal opportunity is insufficient because it does not compensate for unequal innate gifts. What applies for social circumstances should also apply for such gifts, both these factors being purely arbitrary from a moral point of view and requiring adjustment.

According to Rawls, human beings should have the same initial expectations of "basic goods," i.e., all-purpose goods; this in no way precludes ending up with different quantities of such goods or resources, as a result of personal economic decisions and actions. When prime importance is accorded an assurance of equal basic freedoms and rights, inequalities are just when they fulfill two provisos: on the one hand, they have to be linked to offices and positions open to everyone under conditions of fair equality of opportunity; on the other hand, they have to reflect the famous ‘difference principle’ in offering the greatest possible advantage to the least advantaged members of society (Rawls 1993, p. 5 f.; 1971, § 13). Otherwise, the economic order requires revision. Due to the argument of the moral arbitrariness of talents, the commonly accepted criteria for merit (like productivity, working hours, effort) are clearly relativized. The difference principle only allows the talented to earn more to the extent this raises the lowest incomes. According to Rawls, with regard to the basic structure of society, the difference principle should be opted for under a self-chosen "veil of ignorance" regarding personal and historical circumstances and similar factors: the principle offers a general assurance of not totally succumbing to the hazards of a free market situation; and everyone does better than with inevitably inefficient total equal distribution, whose level of well-being is below that of those worst off under the difference principle.

Since Rawls' *Theory of Justice* is the classical focal point of present-day political philosophy, it is worth noting the different ways his theory claims to be egalitarian: First, Rawls upholds a natural basis for equal human worth: a minimal capacity for having a conception of the good and a sense of justice. Second, through the device of the "veil of ignorance," people are conceived as equals in the "original position." Third, the idea of sharing this "original position" presupposes the parties having political equality, as equal participants in the process of choosing the principles by which they would be governed. Fourth, Rawls proposes fair equality of opportunity. Fifth, Rawls maintains that all desert must be institutionally defined, depending on the goals of the society. No one deserves his or her talents or circumstances -- all products of the natural lottery. Finally, the difference principle tends toward equalizing holdings.

Dworkin's equality of resources (1981b) stakes a claim to being even more ‘ambition-sensitive’ and endowment-insensitive’ than Rawls' theory. Unequal distribution of resources is considered fair only when it results from the decisions and intentional actions of those concerned. Dworkin proposes a hypothetical auction in which everyone can accumulate bundles of resources through equal means of payment, so that in the end no one is jealous of another's bundle (the envy test). The auction-procedure also offers a way to precisely measure equality of resources: the measure of resources devoted to a person's life is defined by the importance of the resources to others (Dworkin 1981b, p. 290). In the free market, how the distribution then develops depends on an individual's ambitions. The inequalities that thus emerge are justified, since one has to take responsibility for one's "option luck" in the realm of personal responsibility. In contrast, unjustified inequalities based on different innate provisions and gifts as well as brute luck should be compensated for through a fictive differentiated insurance system: its premiums are established behind Dworkin's own ‘veil of ignorance,’ in order to then be distributed in real life to everyone and collected in taxes. For Dworkin, this is the key to the natural lottery being balanced fairly, preventing an "slavery of the talented" through excessive redistribution.

Objections to all versions of "brute-luck egalitarianism" come from two sides. Some authors criticize its in their view unjustified or excessively radical rejection of merit: The egalitarian thesis of desert only being justifiably acknowledged if it involves desert "all the way down" (Nozick 1974, p. 225) not only destroys the classical, everyday principle of desert, since everything has a basis that we ourselves have not created. In the eyes of such critics, along with the merit-principle this argument also destroys our personal identity, since we can no longer accredit ourselves with our own capacities and accomplishments. (Cf. the texts in Pojman & McLeod 1998.) Other authors consider the criterion for responsibility to be too strong, indeed inhuman in its consequences, since human beings responsible for their own misery would be left alone with their misery (Anderson 1999).

3.6 Equality of Opportunity for Welfare or Advantage

Approaches based on equality of opportunity can be read as revisions of both welfarism and resourcism. Ranged against welfarism and designed to avoid its pitfalls, they incorporate the powerful ideas of choice and responsibility into various, improved forms of egalitarianism. Such approaches are meant to equalize outcomes, insofar as they are the consequences of causes beyond a person's control (i.e., beyond circumstances or endowment), but to allow differential outcomes in so far as they result from autonomous choice or ambition. But the approaches are also aimed at maintaining the insight that individual preferences have to count, as the sole basis for a necessary linkage back to the individual perspective: otherwise, there is an overlooking of the person's value. In Arneson's (1989, 1990) concept of *equal opportunity for welfare*, the preferences determining the measure of individual well-being are meant to be conceived hypothetically -- i.e., a person would decide on them after a process of ideal reflection. In order to correspond to the morally central vantage of personal responsibility, what should be equalized are not enlightened preferences themselves, but rather real opportunities to achieve or receive a good, to the extent that it is aspired to. G.A. Cohen's (1989, p. 916 f.) broader conception of *equality of access to advantage* attempts to link and integrate the perspectives of welfare equality and resource equality

through the overriding concept of advantage. For Cohen, there are two grounds for egalitarian compensation. Egalitarians will be moved to furnish a paralyzed person with a compensatory wheelchair independently of the person's welfare level. This egalitarian response to disability overrides equality of (opportunity to) welfare. Egalitarians also favor compensation for phenomena such as pain, independent of any loss of capacity -- for instance by paying for expensive medicine. But, Cohen claims, any justification for such compensation has to invoke the idea of equality of opportunity to welfare. He thus views both aspects, resources and welfare, as necessary and irreducible. Much of Roemer's (1998) more technical argument is devoted to constructing the scale to calibrate the extent to which something is the result of circumstances. An incurred adverse consequence is the result of circumstances, not choice, precisely to the extent that it is a consequence that persons of one or another specific type can be expected to incur.

3.7 Capabilities Approaches

Theories that limit themselves to the equal distribution of basic means -- this in the hope of doing justice to the different goals of all human beings -- are often criticized as fetishistic, in that they focus on means, rather than on what individuals gain with these means (Sen 1980). For the value goods have for someone depends on objective possibilities, the natural environment, and individual capacities. Hence in contrast to the resourcist approach, Amartya Sen proposes orientating distribution around "capabilities to achieve functionings," i.e., the various things that a person manages to do or be in leading a life (Sen 1992). In other words, evaluating individual well-being has to be tied to a capability for achieving and maintaining various precious conditions and "functionings" constitutive of a person's being, such as adequate nourishment, good health, the ability to move about freely or to appear in public without shame, and so forth. Also important here is the real freedom to acquire well-being -- a freedom represented in the capability to oneself choose forms of achievement and the combination of "functionings." For Sen, capabilities are thus the measure of an *equality of capabilities* human beings enjoy to lead their lives. A problem consistently raised with capability approaches is the ability to weigh capabilities in order to arrive at a metric for equality. The problem is intensified by the fact that various moral perspectives are comprised in the concept of capability (Cohen 1993, p. 17-26, Williams 1987). Martha Nussbaum (1992, 2000) has linked the capability approach to an Aristotelian, essentialistic, "thick" theory of the good -- a theory meant to be, as she puts it, "vague," incomplete, and open-ended enough to leave place for individuality and cultural variations. On the basis of such a "thick" conception of necessary and universal elements of a good life, certain capabilities and functionings can be designated as foundational. In this manner, Nussbaum can endow the capability approach with a precision that furnishes an index of interpersonal comparison, but at some risk: that of not being neutral enough regarding the plurality of personal conceptions of the good ? a neutrality normally required by most liberals (most importantly Rawls 1993).

4. Equality Among Whom?

Justice is primarily related to individual actions. Individual persons are the primary bearers of responsibility (the key principle of ethical individualism). This raises two controversial issues in the

contemporary debate.

One could regard the norms of distributive equality as applying to groups rather than individuals. It is often groups that rightfully raise the issue of an inequality between themselves and the rest of society -- e.g. women; so-called racial and ethnic groups. The question arises of whether inequality among such groups should be considered morally objectionable in itself, or whether even in the case of groups, the underlying concern should be how individuals (as members of such groups) fare in comparative terms.

Do the norms of distributive equality (whatever they are) apply to all individuals, regardless of where (and when) they live? Or rather, do they only hold for members of communities within states and nations? Most theories of equality deal exclusively with distributive equality among people in a single society. But there does not seem to be any rationale for that limitation. Can the group of the entitled be restricted prior to the examination of concrete claims? Many theories seem to imply this when they connect distributive justice or the goods to be distributed with social cooperation or production. For those who contribute nothing to cooperation, such as the disabled, children, or future generations, would have to be denied a claim to a fair share. The circle of persons who are to be the recipients of distribution would thus be restricted from the outset. Other theories are less restrictive, insofar as they do not link distribution to actual social collaboration, yet nonetheless do restrict it, insofar as they bind it to the status of citizenship. In this view, distributive justice is limited to the individuals within a society. Those outside the community have no entitlement to social justice. Unequal distribution among states and the social situations of people outside the particular society could not, in this view, be a problem of social distributive justice. Yet here too, the universal morality of equal respect and the principle of equal distribution demand that we consider each person as prima facie equally entitled to the goods, unless reasons for an unequal distribution can be put forth. It may be that in the process of justification, reasons will emerge for privileging those who were particularly involved in the production of a good. But prima facie, there is no reason to exclude from the outset other persons, e.g. those from other countries, from the process of distribution and justification. That may seem most intuitively plausible in the case of natural resources (e.g. oil) that someone discovers by chance on or beneath the surface of his or her property. Why should such resources belong to the person who discovers them, or on whose property they are located? Nevertheless, in the eyes of many if not most people, global justice, i.e., extending distributive justice globally, demands too much from individuals and their states (Miller 1998). The charge, open, of course, to challenge, is one of excessive demands being made.

5. The Value of Equality: Why Equality?

Does equality play a major role in a theory of justice, and if so, what is this role?

A conception of justice is *egalitarian* when it views equality as a fundamental goal of justice. L. Temkin has put it as follows: "an egalitarian is any person who attaches *some* value to equality *itself* (that is, any person that cares *at all* about equality, over and above the extent it promotes other ideals). So, equality needn't be the only value, or even the ideal she values most....Egalitarians have the deep and (for them) compelling view that it is a bad thing -- unjust and unfair -- for some to be worse off than others through

no fault of their own." (Temkin 1986, p. 100, cf. 1993, p. 7). In general, the focus of the modern egalitarian effort to realize equality is on the possibility of a good life, i.e., on an equality of life prospects and life circumstances -- interpreted in various ways according to various positions in the "equality of what" debate (see above).

I maintain the presence of three sorts of egalitarianism: intrinsic, instrumental and constitutive. (For a two fold distinction cf. Parfit 1997, Temkin 1993, p. 11, McKerlie, 1996, p. 275.) Intrinsic egalitarians view equality as an intrinsic good in itself. As pure egalitarians, they are concerned solely with equality, most of them with equality of social circumstances, according to which it is intrinsically bad if some people are worse off than others through no fault of their own. But it is in fact the case that we do not always consider inequality a moral evil. Intrinsic egalitarians regard equality as desirable even when the equalization would be of no use to any of the affected parties -- e.g. when equality can only be produced through depressing the level of everyone's life. But something can only have an intrinsic value when it is good for at least one person, i.e., makes one life better in some way or another. The following "leveling-down" objection indicates that doing away with inequality in fact ought to produce better circumstances -- it otherwise being unclear why equality should be desired. (For such an objection, cf. Nozick 1974, p. 229, Raz 1986, chap. 9, p. 227, 235, Temkin 1993, pp. 247-8.) Sometimes inequality can only be ended by depriving those who are better off of their resources, rendering them as poorly off as everyone else. (For anyone looking for a drastic literary example, Kurt Vonnegut's science-fiction story *Harrison Bergeron* (1950) is recommended.) This would have to be an acceptable approach according to the intrinsic concept. But would it be morally good if, in a group consisting of both blind and seeing persons, those with sight were rendered blind because the blind could not be offered sight? That would in fact be morally perverse. Doing away with inequality by bringing everyone down contains -- so the objection -- nothing good. Such leveling-down objections would of course only be valid if there were indeed no better and equally egalitarian alternatives available; and nearly always there are such: e.g. those who can see should have to help the blind, financially or otherwise. In case there are no alternatives, in order to avoid such objections, intrinsic egalitarianism cannot be strict, but needs to be *pluralistic*. Then intrinsic egalitarians could say there is something good about the change, namely greater equality -- although they would concede that much is bad about it. Pluralistic egalitarians do not have equality as their only goal; they also admit other values and principles -- above all the principle of welfare, according to which it is better when people are doing better. In addition, pluralistic egalitarianism should be *moderate* enough to not always grant equality victory in the case of conflict between equality and welfare. Instead, it needs to be able to accept reductions in equality for the sake of a higher quality of life for all (as e.g. with Rawls' difference principle).

At present, many egalitarians are ready to concede that equality in the sense of equality of life circumstances has no compelling value in itself; but that, in a framework of liberal concepts of justice, its meaning emerges in pursuit of other ideals: universal freedom, full development of human capacities and the human personality, the mitigation of suffering and defeat of domination and stigmatization, the stable coherence of modern, freely constituted societies, and so forth (Scanlon 1996). For those who are worse off, unequal circumstances often mean considerable (relative) disadvantages and many (absolute) evils; and as a rule these (relative) disadvantages and (absolute) evils are the source for our moral condemnation of unequal circumstances. But this does not mean that inequality as such is an evil. Hence,

the argument goes, fundamental moral ideals other than equality stand behind our aspiring for equality. When we are against inequality on such grounds, we are for equality either as a byproduct or as a means and not as a goal or intrinsic value. In its treatment of equality as a derived virtue, the sort of egalitarianism -- if the term is actually suitable -- here at play is *instrumental*.

As indicated, there is also a third, more suitable approach to the equality ideal: a *constitutive* egalitarianism. According to this approach, we aspire to equality on other moral grounds -- namely, because certain inequalities are unjust. Equality has value, but this is an extrinsic value, since it derives from another, higher moral principle of equal dignity and respect. But it is not instrumental for this reason, i.e., it is not only valued on account of moral equality, but also on its own account. (For the distinction between the origin of a value and the kind of value it is, cf. Korsgaard 1996.) Equality stands in relation to justice as does a part to a whole. The requirement of justification is based on moral equality; and in certain contexts, successful justification leads to the above-named principles of equality, i.e., formal, proportional equality and the presumption of equality. Thus according to constitutive egalitarianism, these principles and the resulting equality are justified and required by justice, and by the same token constitute social justice.

We should further distinguish two levels of egalitarianism and non-egalitarianism, respectively. On a first level, a constitutive egalitarian presumes that every explication of the moral standpoint is incomplete without terms such as ‘equal,’ ‘similarly,’ etc. In contrast, a non-egalitarianism operating on the same level considers such terms misplaced or redundant. On a second level, when it comes to concretizing and specifying conceptions of justice, a constitutive egalitarian gives equality substantive weight. On this level, we can find more and less egalitarian positions according to the chosen currency of equality (the criteria by which just equality is measured) and according to the reasons for unequal distributions (exemptions of the presumption of equality) the respective theories regard as well grounded. Egalitarianism on the second level thus relates to the kind, quality and quantity of things to be equalized. Because of such variables, a clear-cut definition of second level egalitarianism cannot be formulated. In contrast, non-egalitarians on this second level advocate a non-relational entitlement theory of justice.

Alongside the often-raised objections against equality mentioned in the section on "simple equality" there is a different and more fundamental critique formulated by first level non-egalitarians: that equality does not have a foundational role in the grounding of claims to justice. While the older version of a critique of egalitarianism comes mainly from a the right side of the political spectrum, thus arguing in general against "patterned principles of justice" (Nozick 1974, esp. pp. 156-157), the critique's newer version also often can be heard in liberal circles (Walzer 1983, Raz 1986, chap. 9, Frankfurt 1987. 1997, Parfit 1997, Anderson 1999). This first-level critique of equality poses the basic question of why justice should in fact be conceived relationally and (what is here the same) comparatively. Referring back to Joel Feinberg's (1974) distinction between comparative and non-comparative justice, non-egalitarians object to the moral requirement to treat people as equals and many demands for justice emerging from it. They argue that neither the postulate nor these demands involve comparative principles -- let alone any equality principles. They reproach first-level egalitarians for a confusion between "equality" and "universals." As the non-egalitarians see things, within many principles of justice -- at least the especially important ones -- the equality-terminology is redundant. Equality is thus merely a byproduct of the general fulfillment of

actually non-comparative standards of justice: something obscured through the unnecessary inserting of an expression of equality (Raz 1986, p. 227f.). At least the central standards of dignified human life are not relational but "absolute." As Harry Frankfurt puts it: "It is whether people have good lives, and not how their lives compare with the lives of others." (Frankfurt 1997, p. 6) And again: "The fundamental error of egalitarianism lies in supposing that it is morally important whether one person has less than another regardless of how much either of them has." (Frankfurt 1987, p. 34)

From the non-egalitarian vantage, what is really at stake in helping those worse off and improving their lot is *humanitarian concern*, a desire to alleviate suffering. Such concern is understood as not egalitarian. It is not centered on the difference between those better off and those worse off as such (whatever the applied standard), but on improving the situation of persons in bad circumstances. Their distress constitutes the actual moral foundation. The wealth of those better off only furnishes a means that has to be transferred for the sake of mitigating the distress, as long as other, morally negative consequences do not emerge in the process. The strength of the impetus for more equality lies in the urgency of the claims of those worse off, not in the extent of the inequality. For this reason, instead of equality the non-egalitarian critics favor one or another *entitlement theory of justice*, such as Nozick's (1974) libertarianism (cf. 3.2. above) and Frankfurt's (1987) *doctrine of sufficiency*, according to which "What is important from the moral point of view is not that everyone should have *the same* but that each should have *enough*. If everyone had enough, it would be of no moral consequence whether some had more than others." (Frankfurt 1987, p. 21) Parfit's (1997) *priority view* accordingly calls for focus on improving the situation of society's weaker and poorer members and indeed all the more urgently the worse off they are, even if they can be less helped than others in the process. In any case, entitlement-based non-egalitarian arguments can result in praxis in an equality of outcome as far-reaching as egalitarian theories. Hence fulfilling an absolute or non-comparative standard for everyone (e.g. to the effect that nobody should starve) frequently results in a certain equality of outcome, such a standard comprising not only a decent living but a good life. Consequently, the debate here centers on the basis -- is it equality or something else? -- and not so much on the outcome -- are persons or groups more or less equal, according to a chosen metric? Possibly, the difference is even deeper, lying in the conception of morality in general, rather than in equality at all.

Egalitarians can respond to the anti-egalitarian critique by conceding that it is the nature of some (if certainly far from all) essential norms of morality and justice to be concerned primarily with the adequate fulfillment of the separate claims of individuals. However whether a claim can itself be considered suitable can be ascertained only by asking whether it can be agreed on by all those affected in hypothetical conditions of freedom and equality. This justificatory procedure is all the more needed the less evident -- indeed the more unclear or controversial -- it is if what is at stake is actually suffering, distress, an objective need. In the view of the constitutive egalitarians, all the judgments of distributive justice should be approached relationally by asking which distributive scheme all concerned parties can universally and reciprocally agree to. As described at some length in the pertinent section above, many egalitarians argue that a presumption in favor of equality follows from this justification requirement. In the eyes of such egalitarians, this is all one needs for the justification and determination of the constitutive value of equality.

Secondly, even if -- for the sake of argument -- the question is left open of whether demands for distribution according to objective needs (e.g. alleviating hunger) involve non-comparative entitlement-claims, it is nonetheless always necessary to resolve the question of what we do owe needy individuals. And this is tied in a basic way to the question of what we owe persons in comparable or worse situations, and how we need to invest our scarce resources (money, goods, time, energy) in light of the sum total of our obligations. While the claim on our help may well appear non-relational, determining the kind and extent of the help must always be relational -- at least in circumstances of scarcity (and resources are always scarce). Claims are either "satiable" (Raz 1986, p. 235), i.e., an upper limit or sufficiency level can be indicated after which each person's claim to X has been fulfilled, or they are not so. For insatiable claims, to stipulate any level at which one is or ought to be sufficiently satisfied is arbitrary. If the standards of sufficiency are defined as a bare minimum, why should persons be content with that minimum? Why should the manner in which welfare and resources are distributed above the poverty level not also be a question of justice? If, by contrast, we are concerned solely with claims that are in principle "satiable," such claims having a reasonable definitions of sufficiency, then these standards of sufficiency will most likely be very high. In Frankfurt's definition, for example, sufficiency is reached only when persons are satisfied and no longer actively strive for more. Since we find ourselves operating, in practice, in circumstances far beneath such a high sufficiency level, we (of course) live in (moderate) scarcity. Then the above mentioned argument holds as well -- namely, that in order to determine to what extent it is to be fulfilled, each claim has to be judged in relation to the claims of all others and all available resources. In addition, the moral urgency of lifting people above dire poverty cannot be invoked to demonstrate the moral urgency of everyone having enough. In both forms of scarcity, i.e., with satiable and insatiable claims, the social right or claim to goods cannot be conceived as something absolute or non-comparative. Egalitarians may thus conclude that distributive justice is always comparative. This would suggest that distributive equality, especially equality of life-conditions, is due a fundamental role in an adequate theory of justice in particular and of morality in general.

Bibliography

- Albernethy, Georg L. (ed), 1959, *The Idea of Equality*, Richmond: John Knox.
- Anderson, Elizabeth, 1999, "What Is the Point of Equality?" *Ethics* 109, pp. 287-337.
- Aristotle, *Nicomachean Ethics*, in: *The complete works of Aristotle*, ed. Jonathan Barnes, Princeton: Princeton University Press.
- Aristotle, *Politics*, in: *The Complete Works of Aristotle*, ed. Jonathan Barnes, Princeton: Princeton University Press.
- Arneson, Richard, 1989, "Equality and Equal Opportunity for Welfare," *Philosophical Studies* 56, pp. 77-93, reprinted in L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp. 229-241.
- Arneson, Richard, 1990, "Liberalism, Distributive Subjectivism, and Equal Opportunity for Welfare," *Philosophy and Public Affairs* 19, pp. 158-94.
- Arneson, Richard, 1993, "Equality," in: R. Goodin & P. Pettit (eds.), *A Companion to Contemporary Political Philosophy*, Oxford: Blackwell, pp. 489-507.

- Barbeuf, G., 1796, "Manifeste de Égaux," in: *Histoire de G. Barbeuf et du Babouvisme*, Paris 1884, engl. trans. in L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp.49-52.
- Bedau, Hugo Adam, 1967, "Egalitarianism and the Idea of Equality," in: J. R. Pennock, J. Chapman (eds.), *Equality*, New York: Atherton, pp. 3-27.
- Benn, Stanley I. & Richard S. Peters, 1959, *Social Principles and the Democratic State*, London: Allen & Unwin 1959.
- Benn, Stanley, 1967, "Equality, Moral and Social," in: *Encyclopedia of Philosophy*, ed. by Paul Edwards, New York: Macmillan, 1967, Vol. 3. pp. 38-42.
- Berlin, Isaiah, 1955-56, "Equality", *Proceedings of the Aristotelian Society* LVI, pp. 301-326.
- Brown, Henry Phelps, 1988, *Egalitarianism and the Generation of Inequality*, Oxford: Clarendon.
- Cohen, Gerry A., 1989, "On the Currency of Egalitarian Justice," *Ethics* 99, pp. 906-944.
- Cohen, Gerry A., 1993, "Equality of What? On Welfare, Goods, and Capabilities," in: M. Nussbaum & A. Sen (eds.), *The Quality of Life*, Oxford: Oxford University Press, pp. 9?29.
- Dahrendorf, Ralf, 1962, "On the Origin of Social Inequality," in: *Philosophy, Politics, and Society*, 2nd Series, ed. by P. Laslett & W. G. Runciman, Oxford: Blackwell.
- Dann, Otto, 1975, "Gleichheit", in: *Geschichtliche Grundbegriffe*, ed. by V. O. Brunner, W. Conze, R. Koselleck, Stuttgart: Klett-Cotta 1975, pp. 995-1046.
- Dworkin, Ronald, 1977, *Taking Rights Seriously*, Cambridge: Harvard University Press.
- Dworkin, Ronald, 1981a, "What is Equality? Part 1: Equality of Welfare," *Philosophy and Public Affairs* 10, pp. 185-246, reprinted in: R. Dworkin, *Sovereign Virtue. The Theory and Practice of Equality*, Cambridge: Harvard University Press 2000, pp.11-64.
- Dworkin, Ronald, 1981b, "What is Equality? Part 2: Equality of Resources," *Philosophy and Public Affairs* 10, pp. 283-345, reprinted in: R. Dworkin, *Sovereign Virtue. The Theory and Practice of Equality*, Cambridge: Harvard University Press 2000, pp.65-119.
- Dworkin, Ronald, 2000, *Sovereign Virtue. The Theory and Practice of Equality*, Cambridge: Harvard University Press.
- Feinberg, Joel, 1970, "Justice and Personal Desert," in: J. Feinberg, *Doing and Deserving*, Princeton, reprinted in: Louis P. Pojman & Owen McLeod (eds.), *What Do We Deserve? A Reader on Justice and Desert*, Oxford (Oxford University Press) 1998. pp. 70-83.
- Feinberg, Joel, 1974, "Non-Comparative Justice," *Philosophical Review* 83, pp. 297-358.
- Forst, Rainer, 1994, *Kontexte der Gerechtigkeit*, Frankfurt: Suhrkamp; Engl. Trans: *Contexts of Justice*, tr. J. Farrell, Berkeley, Los Angeles : University of California Press, forthcoming 2001.
- Frankfurt, Harry, 1987, "Equality as a Moral Ideal," *Ethics* 98 (1987) 21-42, reprinted in: H. Frankfurt, *The Importance of What We Care About*, Cambridge University Press 1988; reprinted in: L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp. 261-273.
- Frankfurt, Harry, 1997, "Equality and Respect," *Social Research* 64, pp. 3-15.
- Habermas, Jürgen. 1983, "Diskursethik -- Notizen zu einem Begründungsprogramm," in: J. Habermas, *Moralbewußtsein und kommunikatives Handeln*, Frankfurt: Suhrkamp, pp. 53-126; Engl. Trans: "Discourse Ethics: Notes on a Program of Philosophical Justification," in: J. Habermas, *Moral Consciousness and Communicative Action*, tr. C. Lenhardt and S. Weber Nicholsen, Cambridge: MIT Press, 1990, pp. 43-115.

- Habermas, Jürgen. 1992, *Faktizität und Geltung. Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*, Frankfurt: Suhrkamp. 1992; Engl. Trans: *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, Cambridge: MIT Press 1996.
- Hare, Richard M., 1984, "Rights, Utility and Universalization: Reply to J.L. Mackie." in: Raymond G. Frey (ed.), *Utilities and Rights*, Minneapolis, Minn 1984, Oxford: Blackwell 1985.
- Hare, Richard M., *Moral Thinking. Its levels, Method and Point*, Oxford: Oxford University Press 1981.
- Hayek, Friedrich A., 1960, *The Constitution of Liberty*, London: Routledge.
- Hinsch, Wilfried, forthcoming 2001, *Gerechtfertigte Ungleichheiten*, Berlin, New York: de Gruyter.
- Hinsch, Wilfried, forthcoming 2002, "Angemessene Gleichheit" in: *Subjektivität und Anerkennung*, ed. by Barbara Merker, Georg Mohr, Michael Quante.
- Hobbes, Thomas, 1651, *Leviathan*, With Selected Variants from the Latin Edition of 1668, ed. by Edwin Curley, Indianapolis: Hackett 1994
- Kaldor, N., 1939, "Welfare Propositions of Economics and Inter-Personal Comparison of Utility," *The Economic Journal* 49, pp. 549-552.
- Kant, Imanuel, 1785, *Grundlegung zur Metaphysik der Sitten*, in: *Kants Gesammelte Schriften*, ed. by Preußischen Akademie der Wissenschaften, Berlin 1902ff., vol. IV.
- Kant, Imanuel, 1797, *Metaphysik der Sitten*, In: *Kants Gesammelte Schriften*, ed. by Preußischen Akademie der Wissenschaften, Berlin 1902 ff, vol. VI.
- Korsgaard, Christine, 1996, "Two Distinctions in Goodness," in: *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press, pp. 249-53.
- Kymlicka, Will, 1990, *Contemporary Political Philosophy*, Oxford: Clarendon Press.
- Lakoff, Sandford A., 1964, *Equality in Political Philosophy*, Cambridge: Harvard University Press.
- Locke, John, 1690, *The Second Treatise of Government*, ed. C.B. MacPerson, Indianapolis: Hackett 1980.
- Marshall, Thomas Humphrey, 1950, "Citizenship and Social Class," in: T. Marshall, *Citizenship and Social Class and Other Essays*, Cambridge: Cambridge University Press 1950, reprinted London (Pluto) 1981, 1992
- Marx, Karl, 1875, *Critique of the Gotha Program*, reprinted in: *Marx-Engels-Werke* (MEW) vol. 19, Berlin 1978, and in: *Marx-Engels-Gesamtausgabe* (MEGA-B), Berlin 1975 ff., vol. I 25.
- McKerlie, Dennis, 1989, "Equality and Time," *Ethics* 99 (1989) 274-296, reprinted in L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp. 65-75.
- McKerlie, Dennis, 1996, "Equality," *Ethics* 106, pp. 274-296.
- MacKinnon, Catherine, 1989, *Towards a Feminist Theory of the State*, Cambridge: Harvard University Press.
- Menne, Alfred, 1962, "Identität, Gleichheit, Ähnlichkeit," *Ratio* 4, p. 44 ff.
- Miller, David. (1998), "The Limits of Cosmopolitan Justice," in: D. R. Mapel & T. Nardin (eds.), *International Society. Diverse Ethical Perspectives*, Princeton: Princeton University Press, pp. 164-181.
- Nagel, Thomas, 1979, "Equality," in T. Nagel, *Mortal Questions*, Cambridge University Press, pp.

106-127.

- Nagel, Thomas, 1991, *Equality and Partiality*, Oxford University Press.
- Nozick, Robert, 1974, *Anarchy, State, and Utopia*, New York: Basic Books.
- Nussbaum, Martha, 1992, "Human Functioning and Social Justice. In Defense of Aristotelian Essentialism," *Political Theory* 20, pp. 202-246.
- Nussbaum, Martha, 2000, *Women and Human Development: The Capabilities Approach*, Cambridge: Cambridge University Press.
- Okun, Arthur M., 1975, *Equality and efficiency: The Big Tradeoff*, Washington: The Brookings Institution.
- Oppenheim, Felix, 1970, "Egalitarianism as a Descriptive Concept," *American Philosophical Quarterly* 7 (1970) pp. 143-152, reprinted in L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp. 55-65.
- Parfit, Derek, 1997, "Equality and Priority," *Ratio* 10, pp. 202-221.
- Plato, *Republic*, in: Plato, *Complete Works*, ed. John M. Cooper, D.S. Hutchinson, Indianapolis: Hackett 1997.
- Platon, *Laws*, in: Plato, *Complete Works*, ed. John M. Cooper, D.S. Hutchinson, Indianapolis: Hackett 1997.
- Pogge, Thomas W., 1999, "Human Flourishing and Universal Justice," *Social Philosophy and Policy* 16/1 and in Ellen Frankel Paul, et al. (eds.), *Human Flourishing*, Cambridge: Cambridge University Press 1999, pp. 333-361.
- Pojman, Louis P. & R. Westmoreland, (eds.), 1996, *Equality. Selected Readings*, Oxford: Oxford University Press.
- Pojman, Louis P.& Owen McLeod (eds.), 1998, *What Do We Deserve? A Reader on Justice and Desert*, Oxford: Oxford University Press.
- Rae, Douglas et.al., 1981, *Equalities*, Cambridge: Harvard University Press.
- Rawls, John, 1971, *A Theory of Justice*, Cambridge: Harvard University Press, rev. ed. 1999.
- Rawls, John, 1993, *Political Liberalism*, New York: Columbia University Press.
- Raz, Joseph, 1986, *The Morality of Freedom*, Oxford.
- Roemer, John E., 1998, *Equality of Opportunity*, Cambridge: Harvard University Press.
- Rousseau, Jean-Jacques, 1755, *A Discourse on Inequality*, London: Penguin 1984, partly reprinted in L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp. 36-45.
- Rousseau, Jean-Jacques. 1762, *The Social Contract*, Engl. trans. by Maurice Cranston. Harmondsworth: Penguin 1987.
- Scanlon, Thomas, 1996, "The Diversity of Objections to Inequality," in: *The Lindley Lecture*, Lawrence, KA: The University of Kansas.
- Scanlon, Thomas, 1998, *What We Owe to Each Other*, Cambridge: Harvard University Press.
- Sen, Amartya, 1970, *Collective Choice and Social Welfare*, San Fransisco: Holden-Day; reprinted Amsterdam 1979.
- Sen, Amartya, 1992, *Inequality Reexamined*, Oxford: Clarendon Press, Cambridge: Harvard University Press.
- Sikora, R.I., 1989, "Six Viewpoints for Assessing Egalitarian Distribution Schemes," *Ethics* 99, pp. 492-502.

- Steiner, Hillel, 1994, *An Essay on Rights*, Oxford: Blackwell.
- Taylor, Charles, 1992, *Multiculturalism and "The Politics of Recognition"*, Princeton: Princeton University Press.
- Temkin, Larry, 1986, "Inequality," *Philosophy and Public Affairs* 15, reprinted in L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp. 75-88.
- Temkin, Larry, 1993, *Inequality*, Oxford: Oxford University Press.
- Thomson, David, 1949, *Equality*, Cambridge: Cambridge University Press.
- Tugendhat, Ernst & Ursula Wolf, 1983, *Logisch-Semantische Propädeutik*, Stuttgart: Reclam.
- Tugendhat, Ernst, 1993, *Vorlesungen über Ethik*, Frankfurt a.M.: Suhrkamp.
- Tugendhat, Ernst, 1997, *Dialog in Letitia*, Frankfurt a.M: Suhrkamp.
- Van Parijs, Philippe, 1995, *Real Freedom For All. What (If Anything) Can Justify Capitalism?* Oxford: Oxford University Press.
- Vlastos, Gregory, 1962, "Justice and Equality", in: R. Brandt (ed.), *Social Justice*, Englewood Cliffs: Prentice-Hall; rep in: J. Waldron (ed), *Theories of Rights*, Oxford: Oxford University Press 1984, pp. 41-76; reprinted in L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp. 120-133.
- Vonnegut, Kurt, 1950, "Harrison Bergeron," in: K. Vonnegut, *Welcome to the Monkey House*, Delacort Press 1950, pp. 7-13, reprinted in L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp. 315-311.
- Walzer, Michael, 1983, *Spheres of Justice. A Defence of Pluralism and Equality*, New York, London: Basic Books.
- Westen, Peter, 1990, *Speaking Equality*, Princeton: Princeton University Press.
- Williams, Bernard, 1973, "The Idea of Equality," in: B. Williams, *Problems of the Self*, Cambridge: Cambridge University Press, pp. 230-249, reprinted in L. Pojman & R. Westmoreland (eds.), *Equality. Selected Readings*, Oxford: Oxford University Press 1997, pp. 91-102.
- Williams, Bernard, 1987, "The Standard of Living: Interests and Capabilities," in: A. Sen, *The Standard of Living*, Cambridge: Cambridge University Press.
- Wingert, Lutz, 1993, *Gemeinsinn und Moral*, Frankfurt: Suhrkamp.
- Young, Iris Marion, 1990, *Justice and the Politics of Difference*, Princeton: Princeton University Press.

Other Internet Resources

- The Equality Exchange (University of Passau)
- The Equality Studies Centre (University College/Dublin)

[Please contact the author with further suggestions.]

Related Entries

consequentialism | egalitarianism | equality: of opportunity | impartiality | justice: distributive | justification, political: public | libertarianism | luck: justice and bad luck

Stefan Gosepath
Hochschule der Künste Berlin
gosepath@hdk-berlin.de

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 26, 2001
Content last modified: October 8, 2001

William Godwin

William Godwin (1756-1836) was the founder of philosophical anarchism. In his *An Enquiry Concerning Political Justice* (1793) he argued that government is a corrupting force in society, perpetuating dependence and ignorance, but that it will be rendered increasingly unnecessary and powerless by the gradual spread of knowledge. Politics will be displaced by an enlarged personal morality as truth conquers error and mind subordinates matter. In this development the rigorous exercise of private judgment, and its candid expression in public discussion, plays a central role, motivating his rejection of a wide range of co- operative and rule-governed practices which he regards as tending to mental enslavement, such as law, private property, marriage and concerts. Epitomising the optimism of events in France at the time he began writing, Godwin looked forward to a period in which the dominance of mind over matter would be so complete that mental perfectibility would take a physical form, allowing us to control illness and ageing and become immortal.

Godwin's moral theory is often described as utilitarian. He clearly does play an important part in the history of utilitarianism, not least for his invocation of both British and French writers in the tradition, such as Joseph Priestley and d'Holbach and Helvetius, and for the way that his ethical theory is underpinned by a distinctive rationalist necessarianism on the basis of which he insisted on a strong form of first order impartiality. One of Godwin's lasting contributions to moral philosophy, ‘the famous fire cause’, in which we are asked to consider whom I should save from a burning room if I can only save one person and if the choice is between Archbishop Fénelon and a common chambermaid. Fénelon is about to compose his immortal *Télémaque* and the chambermaid turns out to be my mother. Godwin's conclusion that we must save the former relies on consequentialist grounds. However, since his account of the content of utility is inseparable from the development of truth and wisdom, and since we can best promote this through the full and free exercise of private judgment and public discussion, the resulting position looks more like a form of perfectionism than utilitarianism.

Godwin's philosophical importance rests principally on his *Political Justice*. He wrote other philosophical works, *The Enquirer* (1798) and *Thoughts on Man* (1831), but he has become perhaps better known for his novels, the most famous of which is *Caleb Williams* (1794), and for the part he played in literary London from 1783-1836 - from his heyday in the 1790s as the radical philosopher who married Mary Wollstonecraft, through the next forty years in which he was variously the butt of attacks by Thomas Malthus, Samuel Parr and a host of anti-jacobin scribes, friend of the romantic poets, publisher and author of children's books, father-in-law and sponger off Percy Bysshe Shelley, and historian of the Civil War, to his final anomalous position as a government pensioner supported by a Tory administration. His papers and his diary, which sparsely records what he read and wrote and whom he met, provide an

immense resource for scholars of the romantic period.

Life

Godwin was born on 3 March 1756 at Wisbeach, Cambridgeshire, the seventh of thirteen children of John Godwin (1723-1772) a dissenting minister, and his wife Anne (c1723-1809), the daughter of Richard Hull, a ship-owner engaged in the Baltic trade. As a minister Godwin's father was involved in a number of conflicts with his congregations and the family moved first from Wisbeach to Debenham, Suffolk and, in 1760 to Guestwick, near Norwich, Norfolk, where they lived until his father's death. The village was small and the revenue poor; to supplement their income they took in pupils to whom John Godwin taught the classics. The family's financial circumstances improved on the death of Edward Godwin (1695-1764), Godwin's paternal grandfather -- also a dissenting minister and friend of Philip Doddridge, in whose Academy Godwin's father and his uncle Edward had been educated. Godwin's upbringing was rather gloomy. He was not a robust child and his aunt "instructed me to compose myself in sleep, with a temper as if I were never again to wake in this sublunary world." (‘Autobiography’ in *Collected Novels and Memoirs* (*CNM*), 1992, I, 12). At five he was reading *The Pilgrim's Progress* with her, together with James Janeway's *Account of the Conversion, holy and exemplary lives and joyful deaths of several young children* (1671-2), and hymns, catechisms and prayers written by Dr. Isaac Watts. One of Godwin's earliest memories was of composing a poem entitled ‘I wish to be a minister’ (*CNM* I, 15), and a favourite childhood entertainment was to preach sermons in the kitchen on Sunday afternoons.

He was first educated by a Mrs. Gedge, an elderly woman, ‘much occupied in the concerns of religion’, with whom he read the Old and New Testaments. After her death in 1764, he and his brother went to Mr. Akers' school in Hilderston (now Hindolverston). Godwin remained a religious enthusiast and dissenter -- preaching to his fellow schoolchildren, identifying some as ‘children of the devil’, and refusing to answer questions on the Collect of the week, taken from the *Book of Common Prayer* (*CNM* I, 24). His success at Akers' reinforced his commitment to intellectual activity and his aversion to physical toil, and compounded his pride, for which he was frequently admonished by his father. Despite his father's opposition, his resolution to become a minister never wavered, and in 1767 he went to board with a Mr. Samuel Newton, the minister of an independent congregation in Norwich.

Newton was deeply influenced by the writings of Robert Sandeman (1718-1771), a hyper-Calvinist who, scorned faith and presents God as saving or damning a person solely "according to the right or wrong judgment of the understanding" (*CNM* I, 30). Godwin compared Newton in his Autobiography to Caligula or Nero for his spiteful and violent treatment, and he left him in the early summer of 1770, having abandoned his calling and decided to become a bookseller. Six months at Hindolverston persuaded him to resume his pupillage for a further, final year, after which he was pronounced fit for entry into the Dissenting College at Homerton and discharged. Homerton turned him down "on suspicion of Sandemanianism" (*CNM* I, 41). The more tolerant Hoxton Academy, principally run by Andrew Kippis and Abraham Rees admitted him. Hoxton was noted for its Arminianism and Arianism (that is, for the belief that Divine sovereignty was compatible with free will in man and for the rejection of the divinity of Christ), but Godwin's Sandemanianism remained stubbornly untouched, although he supplemented it with "a creed upon materialism and immaterialism, liberty and necessity, in which no subsequent improvement of my understanding has been able to produce any variation." (*CNM* I, 42). In June 1778 he set out to practice his vocation. He had a brief appointment in Ware, followed by a period in London, apparently without income, before obtaining a post in 1780 at Stowmarket, Suffolk. He held the post for two years, during which time his religious beliefs underwent a revolution, moving towards deism after following the suggestion of one of his parishioners and reading Holbach, Helvetius and Rousseau. Not surprisingly, he fell into dispute with his congregation and moved to London in 1782 where friends encouraged him to write for his living.

Later that year he completed his first work, *The history of the Life of William Pitt, Earl of Chatham* (1783), and by the following year was contributing to the English Review, ‘at two guineas a sheet’. At the end of 1782 he returned briefly to his original profession, being employed at Beaconsfield in Buckinghamshire for seven months, during which he produced a volume of sermons, *Sketches of History* (1783). When this appointment broke down he returned to London and resumed his career as an author.

Godwin's output between 1782 and 1784 included, in addition to his *Life of Chatham* and his sermons, three novels, two political pamphlets, a work on education, and a spoof of the critical reviews. None made Godwin much money and it was only when his former tutor, Andrew Kippis, invited him to write the British and Foreign History section for the *New Annual Register*, in July 1784, that he was assured of an adequate income. He probably also made some money from the pieces he wrote in 1785 for the *Political Herald*, a Whig journal, edited by Dr. Gilbert Stuart. The pamphlets, and his pieces for the *Political Herald*, reveal him to be an extremely well informed commentator on contemporary affairs. Between 1785 and 1793 Godwin published little save his work for the *New Annual Register*. Nonetheless, in the summer of 1791, at the height of the debate on the French Revolution, sparked by Edmund Burke's *Reflections on the Revolution in France* (1790), he persuaded his publisher, George Robinson, to support him while he wrote a work summarising recent developments in political philosophy. The work grew from its original conception and was eventually published in two volumes in February 1793 as *An Enquiry Concerning Political Justice*. It is was an immediate success and remains the founding work of philosophical anarchism. Although Godwin drew on principles canvassed in the debate, and on the work of the *philosophes*, *Political Justice* was also powerfully influenced by Godwin's Dissenting education and his involvement in Dissenting circles around Kippis and Timothy and Thomas Brand Hollis. His

success soon made him a central figure in radical political and literary circles of London; he become friends with John Thelwall, Holcroft, and John Horne Tooke (all of whom were indicted for Treason in 1794), he associated with a wide range of other established writers such as Elizabeth Inchbald, James Mackintosh, and Joseph Ritson, and he was sought out by a younger generation of enthusiasts, including William Wordsworth, Samuel Taylor Coleridge and William Hazlitt. In May 1794 Godwin's most successful novel, *Things as they are, or The adventures of Caleb Williams* was published, adding further to his literary reputation, and in the October of that year his shrewd political pamphlet, *Cursory Strictures on the Charge delivered by Lord Chief Justice Eyre to the Grand Jury*, attacked the case for Treason constructed by Eyre against the leaders of the London Corresponding Society and the Society for Constitutional Information, several of whom were his close associates.

A second edition of Godwin's *Political Justice*, in which some of the more rationalist and utopian statements of the first edition were modified, was published at the end of 1795. Shortly thereafter he became reacquainted with Mary Wollstonecraft, whom he had first met briefly in 1791 at a dinner in honour of Paine at which neither was much impressed by the other. Wollstonecraft had subsequently lived in revolutionary France and had had a child in a fraught relationship with Captain Gilbert Imlay, an American merchant. Their second introduction was more successful. As a young man Godwin had been very much the philosopher -- austere in dress, with an angular figure, an intense manner and piercing glance. While approachable he was not socially adept: he both took offence easily and gave it by his over-commitment to the virtue of candour among friends. Only with his increasing success had he come to meet a wide range of clever women with political, literary and philosophical interests -- such as Helen Maria Williams, Inchbald, Amelia Alderson, Maria Reveley, Mary Hays and Mary Robinson. This contact had its effect. He cut his hair short in 1791 and adopted a less ministerial style of dress, he also enjoyed an increasingly extensive social life (albeit without any indication of self-indulgence) and he even experimented in 1796 with holding a dinner party (which included Parr's daughters, Wollstonecraft, and Inchbald). He also developed a basic competence in flirtation. In the last months of 1795 and first half of 1796, Reveley, Samuel Parr's daughter Sarah, Alderson and Inchbald were all candidates for his attention. Following their re-acquaintance Wollstonecraft called on him, unconventionally, in April 1796, and even though Godwin subsequently met and corresponded with her regularly it was only after being turned down by Alderson in July 1796 that they became closer, becoming lovers in August 1796. Their letters and notes provide a touching record of a philosophical relationship gradually subverted by feelings which Godwin found hard to accommodate and Wollstonecraft hard to trust. Wollstonecraft became pregnant in December and after much deliberation to reconcile their actions to their principles, they married in March 1797. Wollstonecraft's death following childbirth in September 1797 left Godwin distraught and burdened with the care of the baby Mary (later Mary Shelley), Imlay's child Fanny, and a succession of debts. He threw himself into work: he revised *Political Justice* for a third, and final time, wrote a hurried memoir of Wollstonecraft, prepared a collection of her works, and embarked on his second major novel, *St Leon* (1799). Wollstonecraft's influence on Godwin's thinking has been detected by critics in his volume of essays, *The Enquirer* (1798), and in the revisions made for the third edition of *Political Justice*, published at the end of 1797. A rather different sense of their relationship was recorded by him in his *Memoirs of the Author of the Vindication of the Rights of Women* (1798), and in his depiction of marriage in *St. Leon* (1799). The *Memoirs* provoked a storm of controversy by their revelations of Wollstonecraft's unconventional sexual mores. Several of Godwin's past acquaintances

spurned him, he found himself increasingly the subject of attack by loyalist newspapers, and his philosophical opinions were parodied and ridiculed in novels, reviews and pamphlets. Godwin reacted with dignity. His *Thoughts Occasioned by the Perusal of Dr. Parr's Spital Sermon* (1801), sought dispassionately to answer his critics and to confess errors which he now recognised -- and which had already been acknowledged both in the revisions to the later editions of his Enquiry, and in his comments in *St. Leon*. But the reply did little to rescue him from the now overwhelming tide of reaction, and incautious remarks in his discussion of *Malthus' Essay on the Principle of Population* (1798) about exposing children and abortion, were seized upon with glee by the reviewers. Godwin's *Political Justice* was a product of the enthusiasm connected with the French Revolution and by the end of the decade the author and his works were exuberantly denounced by loyalism and the forces of order which increasingly dominated the British political and literary scene. From this point on, for much of the rest of his life, Godwinism became a term of opprobrium. In the new, intolerant political climate Godwin turned to literature and history. He tried his hand at drama with a play, *Antonio* (1800), but with little success; in 1803 he wrote a two volume *Life of Chaucer*; and two years later he produced a further novel, *Fleetwood: or The New Man of Feeling* (1805). To cope with his domestic responsibilities he looked for a new wife, approaching Maria Reveley too soon after the death of her husband, and Harriet Lee who found him too pressing. When a widow with two children, Mary Jane Clairmont, leaned over her balcony in 1801 and asked ‘Is it possible that I behold the immortal Godwin’, his fate was sealed.

In 1805, in an effort to establish his finances on a more secure footing, his friends helped establish him as the proprietor of a children's bookshop. Over the next ten years, writing mainly under the pseudonym Edward Baldwin, Godwin produced a variety of books for children: including collections of fables, myths, and bible stories, histories of England, Rome and Greece, and various dictionaries and grammars, but he wrote little of any real political or philosophical significance for ten years.

In 1814 Godwin's domestic life was thrown into turmoil when Percy Bysshe Shelley eloped to France with Godwin's seventeen-year-old daughter Mary, accompanied by Mary's sixteen year old stepsister, Clare Clairmont. The following decade was marked by repeated family and financial crises, by the suicides of Shelley's first wife, Godwin's stepdaughter Fanny, and of his young protégé Patrickson, and by the deaths of three of Mary Shelley's children, followed hard by the death of Shelley himself in 1822. Yet it was also a productive period for Godwin. His *Lives of Edward and John Philips, nephews of Milton* (1815), his chilling tale of madness, *Mandeville* (1817), and his four volume *History of the Commonwealth* (1824-8) each represent his fascination with the republicanism of the civil war period. He also returned to the subject of education in his *Letters of Advice to a Young American* (1818) and in 1820 he produced a critique of Malthus' *Essay*, which won him some respect in some previously hostile quarters, alongside the undisguised enmity of the *Edinburgh Review*. In the last five years of his life he wrote two further novels, and he returned to the philosophical and terrain of his earlier career in his *Thoughts on Man* (1831), his most sustained piece of philosophy since his *Enquirer* (1798). His final work, unpublished in his lifetime, was a series of essays on Christianity, in which he fulfilled an ambition, first noted in 1798, to "sweep away the whole fiction of an intelligent former world and a future state; to call men off from those incoherent and contradictory dreams, that so often occupy their thoughts, and vainly agitate their fears; and to lead them to apply their whole energy to practical objects and genuine realities." (*Political and Philosophical Writings* (*PPW*) IV, 417). In 1833 Godwin finally

received some recognition when he was given a sinecure post by the then Whig government. Peel's subsequent administration agreed to extend the post until Godwin died in April 1836.

Reputation

Hazlitt famously described Godwin's reputation in the 1790s in an essay in his *Spirit of the Age*: No work gave such a blow to the philosophical mind of the country as the celebrated Enquiry ... Tom Paine was considered for a time as Tom Fool to him, Paley and old woman, Edmund Burke a flashy sophist. Truth, moral truth, it was supposed had here taken up its abode; and these were the oracles of thought.’ Godwin himself confirms the view. When travelling in the Midlands in 1794 he found that ‘I was nowhere a stranger. The doctrines of that work, (his Enquiry Concerning Political Justice) coincided in a great degree with the sentiments then prevailing in English Society, and I was everywhere received with curiosity and kindness.’ (Marshall, *William Godwin*, 121). Six years later, reflecting on his reputation, he wrote,

> I have fallen (if I have fallen) in one common grave with the cause and love of liberty; and in this sense I have been more honoured and illustrated in my decline, than ever I was in the highest tide of my success. (*PPW* II, 165)

Philosophically Godwin's greatest supporters were his contemporaries, such as Thomas Holcroft and John Thelwall, and a younger generation of men (and some literary women) who were attracted to Godwin's intellectual rigour and his radical critique of the social and political order. Many later abandoned him, Coleridge, Wordsworth and Southey as part of a rising tide of loyalist reaction, Shelley and Byron, for more personal and domestic reasons. However, his philosophical keyword>anarchismhad a profound influence on Robert Owen, William Thompson and other utopians in the nineteenth century, and there is also evidence of influence on the Chartist movement and on popular labour movements for political reform in the 1840s (see Marshall, 390). His impact in literary circles was long lasting, both through his political writings, and through his novels. *Political Justice* was read and translated by Benjamin Constant in France, and an abridged edition was translated into German in 1803, along with the first three of Godwin's mature novels. Marx and Engels knew of his work and cited him as having contributed to a theory of exploitation. Later in the nineteenth century Anton Menger and Paul Eltzbacher introduced Godwin's work to German audiences, leading to further translation. Caleb Williams appeared in Russian in 1838, and Chernyshevski, Kroptkin and Tolstoy all read and referred to him. In the late nineteenth century the last Book of *Political Justice*, formally titled ‘Of Property’, but dealing with the prospects for progress in the human race and including his attacks on marriage and co-operation, was reprinted as a socialist tract, and the whole work was reprinted again in the 1920s. A critical edition of the third edition with variants appeared in 1946, and an edition of the 1793 text with both later variants and material from the original manuscripts appeared in 1993. Biographies of Godwin have also appeared regularly since the first by C. Kegan-Paul in 1876, which drew heavily on the extensive manuscript sources. Philosophical interest has been less pronounced, although since the 1940s a slow trickle of books has emerged which have sought to do justice to Godwin's essentially liberal political principles and to his moral philosophy. That work has recognised the importance of thinkers of the French Enlightenment, and more recently the

Dissenting inheritance which his education and early career provided. As a result, the traditional view of Godwin as a strict utilitarian has been increasingly challenged. Recent work in political philosophy on the appropriate form and scope of impartiality has looked to Godwin, most commonly to define a position to resist, but not exclusively so.

Political Philosophy

Godwin's major philosophical treatise is his *Enquiry Concerning Political Justice*. The work went through three editions within 5 years, each with substantial changes. No further edition was published in Godwin's lifetime. Although Godwin's other works shed light on changes in Godwin's position after 1798, *Political Justice* is the most coherent expression of Godwin's political philosophy.

The work began as an attempt to review recent developments in political and moral philosophy, but it quickly became more ambitious in scope:

> In the first fervour of enthusiasm I entertained the vain imagination of ‘hewing a stone from rock’ which by its inherent energy and weight should overbear and annihilate all opposition, and place the principles of politics on an unmoveable basis. (*CNM* I, 49)

The discarded first draft centres on the work of Montesquieu and Raynal, while the published work abandons the expository mode and develops its own independent line of argument. Godwin begins by defending the importance of political inquiry and refuting claims that moral and political phenomena are a function of climate, national character or luxury. He argues that character is a function of experience and that the type of government under which people live has an overwhelming impact upon their experience -- bad government produces wretched men and women. Although he is initially prepared to endorse the *philosophe* and republican view that government can have a positive impact on the development of virtue, this view is soon set aside in favour of the argument that moral and political improvement flows from progress in our understanding of moral and political truth -- a process to which there is no limit.

Book Two examines the basic principles of human society, equality, rights, justice, and private judgment. Godwin follows Paine's view in *Common Sense*, that "society is in every state a blessing...government even in its best state is but a necessary evil" (*PPW* III, 48), by seeing society as antecedent to government with its principles setting the bounds of its legitimacy. The basic moral principle is that of justice:

> If justice have any meaning, it is just that I should contribute everything in my power to the benefit of the whole. (*PPW* III, 49)

This principle is filled out by two further principles. The first, equality, is used to establish that we are beings of the same nature, susceptible of the same pleasures and pains, and equally endowed with the capacity for reason. This is to endorse the philosophe principle that birth and rank must not affect the way people are treated -

> the thing really to be desired is the removing as much as possible arbitrary distinctions, and leaving to talents and virtue the field of exertion unimpaired (*PPW* III, 65).

But he also believes (as in the Fénelon case) that some have a higher moral value than others. This second judgment seems rigorously consequentialist, in that we value them more if and only if they contribute more to the general good (a position in line with Godwin's rejection in Book Seven of all desert-based accounts of punishment). Tensions are introduced into his account, however, by the emphasis he places on intention in assessing a person's action -

> It is in the disposition and view of the mind, and not in the good which may accidentally and inintentionally result, that virtue consists. (*PPW* III, 193)

and by his characterisation of the ideal agent as someone devoted to a life of benevolence and virtue. In both instances he appeals to a agent-centred account of virtue, more than to a consequentialist account, and in doing so acknowledges a form of moral worth which is not wholly reducible to consequentialist considerations. The second principle to which he appeals, the doctrine of private judgment, is advanced as the logical complement to the principle of justice:

> to a rational being there is but one rule of conduct, justice, and one mode of ascertaining that rule, the exercise of his understanding. (*PPW*, III, 72).

Here again, although Godwin appeals in part to consequentialist considerations to ground a duty to private judgment, it also plays an integral part of his conception of what it is to be a fully rational agent. When combined with the principle of equality, the principle of private judgment issues in a basic constraint on certain types of consequentialist intervention - each person acts morally only in so far as each acts wholly on the dictates of his or her private judgment. To effect real improvement we must work by appealing to the rational capacities of each of our fellow citizens.

Book Three and the first part of Book Four develop Godwin's case against existing theories of government, in each case making his case by drawing on his opening argument that there is no intrinsic limit on the development of human understanding and enlightenment. The philosophical underpinning for this argument is given in the second half of Book Four where Godwin examines the character of truth and its relationship to virtue and goes on to discuss arguments relating to freedom of the will, the doctrine of philosophical necessity, and the character of moral motivation. He shows that men are capable of recognising truth, and that, because mind acts as a real cause, they will act on it when they perceive it clearly. Nothing beyond the perception of truth is required to motivate our compliance with moral principles. It is this which justifies the description of Godwin's position as ‘rationalist’, and it is on ths point - the motivating power of reason - that later editions show a degree of retraction. One possible source for the position is Richard Price's *Reiview of the Principal Questions of Morals* (1756, but it is noteworthy that Godwin himself later identified this ‘error’ as a function of his Sandemanianism. In *Political Justice*, however, Godwin builds his argument on foundations laid by David Hartley and Joseph

Priestley, albeit he develops their position by insisting that mind is the medium within which sensations, desires, passions and beliefs contend -- so that we should understand the conflict between passion and reason as one of contending opinions. Such contention can be assessed impartially by the mind which will assess the true value of each claim and act on the judgment.

Books Five to Eight apply the principles of justice, equality and private judgment in a critical examination of the institutions of government, issues of toleration and freedom of speech, theories of law and punishment, and, finally, the institution of property. In each case, government and its institutions are shown to constrain the development of our capacity to live wholly in accordance with the full and free exercise of private judgment. In the final book Godwin sketches his positive vision of the egalitarian society of the future, one which, having dispensed with all forms of organised co-operation, including orchestras and marriage, so as to ensure the fullest independence to each persons' judgment, will gradually witness the development of the powers of mind to the point that they gain ascendancy over physiological process allowing life to be prolonged indefinitely.

In 1800 Godwin wrote:

> The Enquiry concerning Political Justice I apprehend to be blemished principally by three errors. 1. Stoicism, or an inattention to the principle, that pleasure and pain are the only bases upon which morality can exist. 2. Sandemanianism, or an inattention to the principle that feeling, and not judgment, is the source of human actions. 3. The unqualified condemnation of the private affections. It will easily be seen how strongly these errors are connected with the Calvinist system, which had been so deeply wrought into my mind in early life, as to enable these errors long to survive the general system of religious opinions of which they formed a part...The first of these errors...has been corrected with some care in the subsequent edition of Political Justice. The second and third owe their destruction to a perusal of Hume's *Treatise of Human Nature* in the following edition. (*CNM*, I, 54)

This account is a fair characterisation of the changes which Godwin made in the second and third editions. Sentiment and feeling are given a much more powerful role, no longer to be expunged by the power of truth; the private affections are allowed to play a part in moral reasoning; and a more consistently utilitarian language is deployed throughout the work. As a consequence, the rationalism which marked the first edition becomes muted and, while the belief in progress is maintained, the more utopian flights of the first edition are omitted.

Moral Philosophy

One of the most powerful attacks on Godwin was that made in Dr. Samuel Parr's ‘Spital Sermon’ of 1800. It was Godwin's advocacy of universal benevolence against which Parr directed his energies, centring his attack on Godwin's early dismissal of family feeling, gratitude and various natural sentiments. For Godwin, these are passions unconstrained by judgment, and so should not play a role in determining how we should act. He exemplifies his case in what has come to be known as the ‘Famous

Fire Cause', in which the reader is asked to imagine being able to save only one of two people in a fire, one of whom is the Archbishop Fénelon, a benefactor to the whole human race, the other of whom is the reader's parent (mother in the first edition, father thereafter!). Godwin's view is that justice demands that we act impartially for the greater good, which means saving Fénelon. He never abandoned this case, nor the view that it is our duty to act to bring about the greatest good. Just as a judge should not be influenced by familial or private concerns in his judgment, so too is the moral agent bound to judge impartially. In replying to Parr, Godwin expresses regret that he had not appealed to the still more persuasive case of Brutus executing his two sons -- a striking example, and a republican commonplace about justice trumping paternal duties. As Godwin says, saving someone just because they are a relation seems bizarre without some additional judgment about their moral worth: a parent who is foolish or evil cannot have an over-riding claim on us against the moral deserts of all other members of the human race. That position, Godwin retains. Moreover, in his reply to Parr, he insists that these extraordinary cases are unlikely to shake the domestic affections in the ordinary intercourse of life. However, from the later editions and other works, it becomes clear that he will admit, in the more normal course of events, a much more substantial role to be played by our natural affections and attachments. They provide us both with information about how best we might benefit others, and a basic moral motivation which can be relied on in normal cases and which can be generalised beyond the narrow domestic sphere (a position much indebted to Adam Smith's *Theory of Moral Sentiments*). These changes are significant: it leaves us a less rationalist, more philosophically robust, account of moral motivation and its relationship to the principle of utility, and it does much to moderate the utopianism of the first edition.

The impact of these changes on Godwin's over-all position is more difficult to assess. What we see in the changes is a consistent shift away from the rationalist account of moral motivation which marked the first edition to a position which is much more sceptical about the power of reason. This scepticism inevitably moderates Godwin's belief in perfectibility, since it becomes more difficult to argue for convergence on principles of morality and the progressive development of knowledge. It also inevitably undermines Godwin's faith in the triumph of mind over physiological processes. That said, neither the doctrine of private judgment nor the principle of utility depend on his earlier rationalism. The former is defended by Godwin on the grounds that only free action has moral value, and that the fullest possible exercise of private judgment is required for one's actions to be free - further evidence of Godwin's attempt to provide an agent-centred account of virtue alongside his consequentialism. With this commitment private judgment remains defensible even if there is a low probability that its exercise will produce true beliefs, so long as no other better method of tracking truth is available (which also becomes proportionately less likely as one's scepticism increases). The defence might require that cognitive status be attributed to moral judgments, but it might also be possible to sustain the argument for private judgment independent of the issue of ethical objectivity. The utility principle might seem to call for an ability to make sound ethical judgments in complex situations but, again, if we are sceptical about people's ability to judge well, this does not entail (and seems to deny) that there is a better way of judging. On both counts then, Godwin's central principles remain intact despite the changes he makes to the account of moral motivation and judgment. Moreover, Godwin's view of man's progressive character might be defended by placing greater weight on the baleful effects of the social and political institutions of the European aristocracies than on the epistemological dimensions of the account.

However, Godwin's endorsement of both the principle of utility as the sole guide to moral duty and the principle of private judgment as a block on the interference of others, is not without tensions. His consistent doctrine is a combination of these two principles: that it is each individual's duty to produce as much happiness in the world as he is able, and that each person must be guided in acting by the exercise of his private judgment, albeit informed by public discussion. If the resulting doctrine is utilitarian it is a highly distinctive form: it is act-utilitarian in that it discounts reliance on rules (although see Barry's suggestion that his act- utilitarianism gives way to motive utilitarianism, *Justice as Impartiality* 224; and see Godwin's invocation of sincerity as a partial rule constraint in the first edition, PPW III, 135-42); it is ideal, in that it acknowledges major qualitative differences in the pleasures; and it is indirect, in that we can only promote over-all utility by improving the understanding of our fellow human beings. More troubling to the view that this none the less amounts to utilitarianism is Godwin's insistence on private judgment as a basic constraint, and his associated characterisation of the fully moral agent in terms of the full development of the individual's intellectual powers and potential. Indeed, Godwin's account of pleasure, in terms of the development of intellect and the exercise of its powers, means that the position looks as much like perfectionism as it does a form of hedonistic utilitarianism (what is valued is the ideal as much as the pleasures which are integral to it). Furthermore, it suggests that no distincton can be drawn between the means which we adopt to promote the general good and the character of the general good itself. That is, what promotes the general good is the development of human intellect, but the general good just is the development of the human intellect. If that is true, Godwin's account cannot be utilitarian because it cannot be consequentialist (because it cannot separate the means to the end from the end which is sought).

Such issues of interpretation remain very much in dispute in studies of Godwin (compare Clarke (1977) with Philp (1986)), complicated by issues concerning the weight to be given to the different editions of *Political Justice* and Godwin's later writings. However, even if a utilitarian reading of Godwin is accepted, it remains the case that the doctrine is strictly a precept of individual moral judgment. Because of his broader views as to the corrupting influence of government, there can be no extension of the principle to politics. Each of us must judge as best we can how to advance the good of all, but every person is owed a respect for their private judgment which precludes us from exercising authority over them. By this constraint, Godwin delivers utilitarianism from the more statist approaches of Bentham and later utilitarians. It also ensures that the doctrine retains a fundamentally egalitarian form. The constraint also supports the view that Godwin reached his philosophical position less through the *philosophes*, than by secularising Dissenting arguments for the sanctity of private judgment and generalising their application to every mode of human activity. This commitment also provides support for a reading of Godwin's position which sees it as concerned with individual moral perfectibility, couched in the language of utility, rather than as strictly utilitarian.

A Philosophy of History

Political Justice condemns all government interference with individual judgment. He claims that over time history has seen gradual progress as knowledge has developed and has spread and as men and women have liberated themselves from their political chains and their subordination to the fraud and

imposture of monarchical and aristocratic government and established religion. His optimistic belief in the impotence of government against advancing opinion (which partly glosses and extends Hume's comment that all government is founded on opinion) is balanced by some sociologically perceptive comments on the baleful influence that certain types of political power have on those who exercise it or are subject to it. These insights are also explored in The *Enquirer*, but it is in Godwin's later novels, from *Caleb Williams* (1794) onward, that it is given its fullest rein. As Godwin indicates in his unpublished essay, ‘On History and Literature’ (1798) (*PPW* V, 290-301), literature can be used to show how the cultures and institutions into which we are born come inexorable to shape our lives, leading us to act in ways which destroy our chances of happiness. The six mature novels effectively follow through the critical enterprise launched in *Political Justice* by their narrative histories of men who are brought to grief by the aristocratic and inegalitarian principles of their societies.

Bibliography

Primary Sources

A complete bibliography of Godwin's work published in his lifetime is given in volume 1 of *The Collected Novels and Memoirs of William Godwin* 8 Volumes ed., Mark Philp (London, Pickering and Chatto, 1992) and in volume 1 of *The Political and Philosophical works of William Godwin*, 7 Volumes ed., Mark Philp, (London, Pickering and Chatto, 1993). These are referred to below as, respectively *CNM* and *PPW*.

- *History of the Life of William Pitt, Earl of Chatham* printed for the Author and sold by G. Kearsley, published anonymously, (London, 1783). (See *PPW* I)
- *An Account of the Seminary that will be opened on Monday the Fourtb Day of August at Epsom in Surrey*, published anonymously, (London, T. Cadell, 1783). (See *PPW* V)
- *Sketches of History in Six Sermons*, (London, T. Cadell, 1784) 190. (Some copies anonymous, others with Godwin's name) (See *PPW* VII).
- *The Herald of Literature, as a Review of the most considerable publications that will be made in the course of the ensuing Winter*, published anonymously, (London, J. Murray, 1784). (See *PPW* V)
- *Instructions to a Statesman. Humbly inscribed to the Right Honourable George Earl Temple*, published anonymously, (London, Murray, J. Debrett & J. Sewell, 1784). (See *PPW* I)
- *An Enquiry concerning Political Justice, and its Influence on General Virtue and Happiness*, 2 volumes, (London, G. G. & J. Robinson, 1793) 4o, xiii, 378, 379-895. A pirated first edition also published 2 volumes 8o in Dublin by Luke White, 1793, xiii, 411, 424. Copies of the octavo first edition with a Robinson flyleaf also exist. 2nd edition, 2 volumes 8o, (London, Robinson, 1796), xviii, 464, v, 545; 3rd edition, 2 volumes 8o, (London, Robinson, 1798), lvi, 463, ix, 554. See also *Enquiry Concerning Political Justice*, 3 volumes, ed., F. E. L. Priestley, (Toronto, University of Toronto Press, 1946, 1969). Facsimile reprint of the third edition with variants from the first and second editions in volume 3; *Enquiry Concerning Political Justice*, ed., Isaac Kramnick (Harmondsworth, Penguin Press, 1976), 825 (third edition); and *PPW* III and IV (first edition text

plus variants from manuscript and from the subsequent editions).

- *Things As They Are; or The Adventures of Caleb Williams*, 3 volumes, (London, B. Crosby, 1794). Critical edition of the fifth edition edited by D. McCracken, Oxford University Press, 1970. Critical edition of the first edition in *CNM* III.
- *Cursory Strictures on the Charge delivered by Lord Chief Justice Eyre to the Grand Jury ... October 2, 1794, first published in the Morning Chronicle October 21*, Published anonymously, (London, D. I. Eaton, 1794) and *A Reply to an Answer to Cursory Strictures, supposed to be wrote by Judge Buller. By the Author of Cursory Strictures*, published anonymously, (London, D. I. Eaton, 1794), 7. (See *PPW* II)
- *Considerations on Lord Grenville's and Mr. Pitt's Bills, concerning Treasonable and Seditious Practices, and Unlawful Assemblies*. By a Lover of Order, published anonymously, (London, J. Johnson, 1795), 86. (See *PPW* II).
- *The Enquirer, Reflections on Education, Manners and Literature*, (London, G. G. & J. Robinson, 1797. (See *PPW* V).
- *Memoirs of the Author of a Vindication of the Rights of Woman*, (London, J. Johnson and G. G. & .1- Robinson, 1798. (See *CNM* I).
- *St. Leon, A Tale of the Sixteenth Century*, 4 volumes, (London, G. G. & J. Robinson, 1799). (See *CNM* IV)
- *Thoughts occasioned by the Perusal of Dr. Parr's Spital Sermon, preached at Christ Church, April I5, 1800: being a Reply to the Attacks of Dr. Parr, Mr. Mackintosh, the Author of an Essay on Population, and Others*, (London, G. G. & J. Robinson, 1801). (See, *PPW* II).
- *Fleetwood. or The New Man of Feeling*, 3 volumes, (London, R. Phillips, 1805). (See *CNM* V)
- *Lives of Edward and John Philips. Nephews and Pupils of Milton. Including Various Particulars of the Literary and Political History of their times*, (London, Longman, Hurst, Rees, Orme & Brown, 1815), xv, 410.
- *Mandeville, a Tale of the Seventeenth Century in England*, 3 volumes, (Edinburgh, A. Constable; London, Longman, Hurst, Rees, Orme & Brown, 1817). (See *CNM* VI)
- *Letter of Advice to a Young American on the course of studies it might be most advantageous for him to pursue*, (London, M. J. Godwin, 1818. First published in Scots Magazine, 1818) and *Further Letters of Advice to Joseph Beavan*, *Analectic Magazine*, Philadelphia, 1818. (See *PPW* V)
- *Of Population. An Enquiry concerning the Power of Increase in the Numbers of Mankind, being an Answer to Mr. Malthus's Essay on that Subject*, (London, Longman, Hurst, Rees, Ornie & Brown, 1820) (See selection in *PPW* II)
- *History of the Commonwealth of England from its commencement to its restoration*, 4 volumes, (London, H. Colburn, 1824-8).
- *Thoughts on Man, his Nature, Productions, and Discoveries. Interspersed with some particulars respecting the author*, (London, Effingham Wilson, 1831) (See *PPW* VI)
- *Essays, Never before published, by the late William Godwin*, ed. C. Kegan Paul, (London, H.S. King, 1873) (See *PPW* VII, re-edited from the manuscript entitled *The Genius of Christianity Unveiled*).

Collected Works

- *Collected Novels and Memoirs of William Godwin*, 8 volumes, ed., Mark Philp, London, Pickering and Chatto Publishers Ltd., 1992. A complete scholarly edition of all Godwin's published novels, his biography of Mary Wollstonecraft, and a range of previously unpublished autobiographical writings. Caleb Williams and Memoirs of the Author of the Vindication of the Rights of Woman are set in the first edition with variants from later editions (and, for Caleb Williams, the manuscript) given in an appendix. All other novels are set from the last edition published within the author's lifetime.
- *Political and Philosophical Writings of William Godwin*, 7 volumes, ed., Mark Philp, London, Pickering and Chatto Publishers Ltd.,1993. A scholarly edition of Godwin's principal writings in politics, philosophy, education and theology, including previously unpublished manuscript material. The edition is made up of two volumes of Godwin's principal political essays, including a substantial unpublished essay; two volumes of his Enquiry Concerning Political Justice, which is set in the first edition (volume III) with variants from the manuscript and the second and third editions given in volume IV (volume IV also includes a previously undiscovered first draft of the Enquiry together with manuscript material relating to the publication of and revisions to the Enquiry); the three later volumes collect Godwin's main educational and literary writings, with previously unpublished material (Volume V), his later essays (volume VI) and his religious writings, including unpublished material and a re-edited edition of Godwin's last,unfinished work, The Genius of Christianity Unveiled (volume VII).
- *Four Early Pamphlets (1783-1784)*, ed., B. R. Pollin, (Gainsville, Florida, Scholars' Facsimiles and Reprints, 1966).
- *Uncollected Writings (1785-1822)*, ed. J. E. Marken and B. R. Pollin, (Gainsville, Florida, Scholars Facsimiles and Reprints, 1968)

Manuscript Collections

- Bodleian Library, Oxford. The Abinger collection, owned by Lord Abinger is a very extensive holding of Godwin's manuscript material, correspondence and diaries. Some of the earlier deposits were microfilmed by Duke University but there have been several deposits made subsequently.
- National Art Library, Victoria and Albert Museum, London. The Forster/Dyce Collection includes the manuscripts to Godwin's Political Justice, Caleb Williams, Life of Chaucer and History of the Commonwealth, and a limited amount of correspondence.
- Pforzheimer Library, New York. Contains the manuscript of Fleetwood and miscellaneous correspondence and material relating to St Leon. (It has been edited by K. N. Cameron, Shelley and his Circle ,volumes I-IV, Cambridge, Mass., Harvard University Press, 1961-70, and D. H. Reiman, volumes V-VI, Ibid., 1973.

Bibliography

- Pollin, Burton R., *Godwin Criticism: A synoptic bibliography*, (Toronto, University of Toronto Press, 1967), 659. A bibliography of all critical work on Godwin to that date.

Biographical Works (in chronological order): :

- Paul, Charles Kegan (1876) William Godwin: his Friends and Contemporaries, 2 volumes (London: H.S King). A substantial biography which remains essential, containing manuscript material no longer available.
- Roussin, Henri (1913) *William Godwin* (Paris: Plon-Nourrit).
- Brown, Ford K. (1926) *The Life of William Godwin* (London: Dent).
- Woodcock, George (1946) *William Godwin. A Biographical Study* (London: Porcupine Press).
- Locke, Don (1980) *A Fantasy of Reason: The Life and Thought of William Godwin* (London: Routledge and Kegan Paul).
- Marshall, Peter, H. (1984) William Godwin (New Haven: Yale University Press
- St. Clair, William (1989) *The Godwins and the Shelleys: The Biography of a Family* (London: Faber and Faber).

Philosophical Commentaries

In chronological order:

- Fleischer, David (1951) *William Godwin, a study in Liberalism* (London: George Allen and Unwin).
- Monro, D. H. (1953) *Godwin's Moral Philosophy: An Interpretation of William Godwin* (Oxford, Oxford University Press).
- Pollin, Bruton R (1962) *Education and Enlightenment in the works of William Godwin* (New York: Las Americas).
- Clark, John P. (1977) *The Philosophical Anarchism of William Godwin* (Princeton, New Jersey, Princeton University Press).
- Tysdahl, B. J.(1981) *William Godwin as Novelist* (London: Athlone Press).
- Philp, Mark (1986) *Godwin's Political Justice* (London: Duckworth).
- Crowder, George (1991) *Classical Anarchism* (Oxford: Oxford University Press).
- Morrow, John (1991) ‘Republicanism and Public Virtue: William Godwin's History of the Commonwealth of England’ *The Historical Journal* **34**, 3 (Cambridge) 645-664
- Clemit, Pamela (1993) *The Godwinian Novel* (Clarendon Press: Oxford University Press)
- Singer, Peter, Cannold, Leslie, and Kuhse, Helga (1995) ‘William Godwin and the Defence of Impartialist Ethics’, *Utilitas*, **7**(1) (Edinburgh University Press) 67-86
- Barry, Brian (1995) *Justice as Impartiality* (Oxford: Oxford University Press).

Other Internet Resources

- Godwin Home Page
- Portraits of Godwin at the National Portrait Gallery (London)

Related Entries

anarchism | consequentialism | Wollstonecraft, Mary

Mark Philp
Oriel College, Oxford University
mark.philp@oriel.ox.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 16, 2000
Content last modified: January 16, 2000

Lord Shaftesbury [Anthony Ashley Cooper, Third Earl of Shaftesbury]

Anthony Ashley Cooper, the third Earl of Shaftesbury, lived from 1671 to 1713. He was one of the most important philosophers of his day, and exerted an enormous influence throughout the eighteenth and nineteenth centuries on British and European discussions of morality, aesthetics and religion.

Shaftesbury's philosophy combined a powerfully teleological approach, according to which all things were part of a harmonious cosmic order, with sharp observations of human nature (see section 2 below). Shaftesbury is often credited with originating the moral sense theory, although his own views of virtue are not as thoroughly anti-rationalist as those of later sentimentalists such as Hutcheson and Hume (section 3). While he argued that virtue leads to happiness (section 4), Shaftesbury was also a fierce opponent of psychological and ethical egoism (section 5) and of the egoistic social contract theory of Hobbes (section 6). Shaftesbury's view of aesthetic judgment was both sentimentalist and objectivist, in that he thought that correct moral judgment was based in human sentiments that reflected accurately the harmonious cosmic order (section 7). Shaftesbury's belief in an harmonious cosmic order also dominated his view of religion, which was based on the idea that the universe clearly exhibited signs of divine design (section 8); according to Shaftesbury, the ultimate end of religion, as well of virtue, beauty and philosophical understanding (all of which are turn out to be one and the same thing), is to identify completely with the universal system of which one is a part.

1. Shaftesbury's Life and Works

Shaftesbury lived from 1671 to 1713. His grandfather, the first Earl of Shaftesbury, oversaw Shaftesbury's early upbringing, and put John Locke in charge of his education. Shaftesbury would eventually come to disagree with many aspects of Locke's philosophy (such as the latter's empiricism, his social contract theory, and what Shaftesbury perceived to be his psychological and ethical egoism), but Locke was clearly a crucially important influence on Shaftesbury's philosophical development, and the two men remained friends until Locke's death.

Shaftesbury served in Parliament and the House of Lords, but ill health curtailed his political career when he was 30 years old. From then on, he concentrated most of his energies on his philosophical and literary writings.

The first work Shaftesbury published was an edited collection of sermons by Benjamin Whichcote, which came out in 1698. Shaftesbury wrote an unsigned preface to the sermons in which he praised Whichcote's belief in the goodness of human beings and urged his readers to use Whichcote's "good nature" as an antidote to the poisonous egoism of Hobbes.

In 1699, John Toland published an early version of Shaftesbury's *Inquiry concerning Virtue*. But Shaftesbury renounced this version of the Inquiry, claiming (probably truthfully) that it was produced without his authorization.

Most of the works for which Shaftesbury is famous were written between 1705-1710. It was during this period that he rewrote the *Inquiry concerning Virtue* and completed versions of *A Letter concerning Enthusiasm*, *Sensus Communis: An Essay on the Freedom of Wit and Humour, The Moralists and Soliloquy, or Advice to an Author*.

In 1711, he collected his mature works into a single volume and added to them extensive notes and commentary, naming the book *Characteristics of Men, Manners, Opinions, Times*. He revised the *Characteristics* over the course of the next two years, up until his death in 1713. A revised edition came out in 1714.

The *Characteristics* is a remarkable volume. It covers a great many topics, ranging freely over morality, art, politics, religion, aesthetics, culture and politeness, and it is written in many different styles, including epistles, soliloquys, dialogues and treatises. The overarching goal of the book, as Klein has put it in his very helpful introduction, is to make its readers "effective participants in the world" (*Characteristics* viii). Shaftesbury saw the *Characteristics* as an exercise in practical (and not merely speculative) philosophy -- as a work that would enable people to live better lives.

The *Characteristics* was extremely popular in Britain and Europe throughout the eighteenth and nineteenth centuries. It was a book that was closely studied by numerous philosophers and artists, as well as widely read by educated people in general.

In addition to the Characteristics, there are two other major posthumous collections of Shaftesbury's writings: the *Second Characteristics,* which is concerned chiefly with the visual arts, and his philosophical notebooks, which Rand has collected in The Life, Unpublished Letters and Philosophical Regimen of Anthony, Earl of Shaftesbury. The notebooks are particularly interesting, as they offer a view of Shaftesbury's private ruminations and of his profound commitment to elements of stoicism.

2. Shaftesbury's View of Human Nature: Teleology and Observation

Shaftesbury's view of human nature is both teleological and observation-based. Indeed, he believes that teleology and observation must go hand-in-hand -- that accurate observation of human psychology requires a teleological conception of humanity, and that one needs to observe human beings to learn about the human telos. He is very critical of philosophers who examine human beings without placing their findings within a teleological context, comparing them to someone who examines the individual parts of a watch without taking into account the purpose for which the watch was designed: just as the latter person will never really come to a proper understanding of the watch, Shaftesbury argues, so too the former will never come to a proper understanding of human nature. Shaftesbury thought that Descartes and Locke were guilty of this narrow non-teleological type of philosophizing.

3. Moral Sentimentalism and Moral Rationalism

Shaftesbury, like most teleologically-minded philosophers, contends that the end or telos of human nature is virtue, and much of his writing is devoted to an explication of his conception of virtue. The account of virtue Shaftesbury proposes has often been taken to be the origin of moral sentimentalism or the moral sense theory, which Hutcheson and Hume would later develop. But while there are parts of Shaftesbury's account of virtue that are undeniably sentimentalist, there are also rationalist elements that defy the sentimentalist or moral sense label. Let us first note the most conspicuously sentimentalist aspects of Shaftesbury's view, and then note the more rationalist aspects.

The place in Shaftesbury in which the moral sense theory most clearly begins to take shape is his *Inquiry concerning Virtue or Merit*, Book I, Part II, Section 3, where he explains his view of "virtue or merit." When other beings "offer themselves" to our senses, Shaftesbury explains in that section, we perform actions and feel affections. So much we have in common with other animals. But we humans also consider our own actions and affections, and form affections about them. As Shaftesbury puts it, "[T]he very actions themselves and the affections of pity, kindness, gratitude and their contraries, being brought into the mind by reflection, become objects. So that, by means of this reflected sense, there arises another

kind of affections towards those very affections themselves, which have been already felt and have no become the subject of a new liking or dislike" (*Characteristics of Men, Manners, Opinions, Times*, edited by Lawrence Klein [Cambridge: Cambridge University Press, 1999], 172).

This reflected sense, Shaftesbury contends, is the origin of virtue, by which he means both that a being can form a judgment about what is virtuous and vicious only if it has this reflected sense, and that a being can be judged virtuous and vicious only if its actions are influenced by its own reflected sense.

> So that if a creature be generous, kind, constant, compassionate, yet if he cannot reflect on what he himself does or sees others do so as to take notice of what is worthy or honest and make that notice or conception of worth and honesty to be an object of his affection, he has not the character of being virtuous. For, thus and no otherwise, he is capable of having a sense of right or wrong, a sentiment or judgment of what is done through just, equal and good affection or the contrary. (*Characteristics* 173)

These passages from Book I, Part II, Section 3 of the Inquiry seem to give us all we need to attribute to Shaftesbury a sentimentalist or moral sense view or virtue. The fact that Shaftesbury attributes to humans a "sense of right or wrong" is not on its own enough to make the attribution, since that phrase is general enough to refer to a rational capacity as well as a sentimentalist one. But earlier in the section Shaftesbury has called the response from which virtue originates an "affection" and a "liking or dislike." He has also maintained that the response that is essential to virtue is similar to (or perhaps the same as) the "heart"-based pleasure we feel when we see or hear something beautiful. Later in the *Inquiry*, moreover, Shaftesbury says that "this sense of right and wrong ... must consist in a real antipathy or aversion to injustice or wrong and in a real affection or love towards equity and right" (*Characteristics* 178). And in the passage quoted above from *Characteristics* 173, he seems to equate moral judgment with "a sentiment." All of this looks to be conclusive evidence that at the core of Shaftesbury's view of morality is something emotional, sentimental or passionate -- not rational.

Further evidence of a sentimentalist moral theory in Shaftesbury occurs in his discussion of how a person can come to lose his sense of right and wrong. He argues (in a manner that anticipates Hume) that because our sense of morality is a sentiment, it can be opposed only by another sentiment, and not by reason or belief. "Sense of right and wrong," he writes, "therefore being as natural to us as natural affection itself, and being a first principle in our constitution and make, there is no speculative opinion, persuasion or belief which is capable immediately or directly to exclude or destroy it... [T]his affection being an original one of earliest rise in the soul or affectionate part, nothing beside contrary affection, by frequent check and control, can operate upon it so as either to diminish it in part or destroy it in the whole" (*Characteristics* 179).

There are other passages, however, in which Shaftesbury sounds a more rationalist note. He speaks, for instance, of the "eternal measures and immutable independent nature of worth and virtue" (*Characteristics* 175) and of "a fitness and decency in actions" (*Characteristics* 327) -- phrases that are touchstones for the rationalists of the period and which Hutcheson and Hume would later attack. And he

contends that "partial affection, or social love in part, without regard to a complete society or whole, is in itself an inconsistency, and implies an absolute Contradiction," as well as maintaining that an affection that is "applied only to some one part of society, or of a species, not to the species of society itself ... has no foundation or establishment in reason" (*Characteristics* 205). Moreover, Shaftesbury often links worth and virtue to a "universal system" (*Characteristics* 167-170) yet another idea that Hutcheson and Hume would eschew.

How are the rationalist-sounding passages in Shaftesbury to be reconciled with his sentimentalist-sounding ones? Shaftesbury seems to think that value exists independently of human affections but that our affections are what enable us to make value judgments (which in turn is what makes us capable of virtue or merit). Shaftesbury's "sense of right and wrong" is, then, truly a sentiment, but it is a sentiment that accurately represents an objective reality -- i.e., a reality that is independent of human sentiments. Shaftesbury tells the same story about aesthetics: there is, he maintains, an eternal and immutable standard of beauty, one that is independent of human affections, but our affections are what enable us to learn about beauty and make aesthetic judgments.

Does this combination of rationalism and sentimentalism constitute a coherent view? Butler might have thought so, and it's just possible that Hutcheson did as well (although Hutcheson's views of this matter are very difficult to pin down). But Balguy, Hume and most twentieth century meta-ethicists would contend that all of Shaftesbury's statements about morality cannot be consistently combined with each other -- that something has to go. However that may be, it is worth keeping in mind that Shaftesbury himself was probably unconcerned about the problem of reconciling aspects of moral rationalism with aspects of moral sentimentalism. This is partly due to the fact that the sharp distinction between rationalism and sentimentalism had not yet been drawn at the time at which Shaftesbury was writing. It is also partly due to the fact that Shaftesbury did not have as a goal for all of his writings the construction of a single systematic theory. His underlying purpose was to improve the character of his readers, and toward this end he was quite willing to write in different styles and different voices, and to use different arguments from one writing to the next. That is not to say that Shaftesbury was happy to contradict himself. But he did seem to think that one could advance the cause of virtue in different ways and that a strict adherence to systematic rigor could frustrate the higher ends to which philosophy should be put.

4. Virtue and Happiness

One point on which Shaftesbury never wavers is that virtue promotes the good of all humankind. As he says, "To love the public, to study universal good, and to promote the interest of the whole world, as far as lies within our power, is surely the height of goodness" (*Characteristics* 20). Or as he puts it elsewhere, the virtuous person is the one who strives to develop an "equal, just and universal friendship" with all humankind. This view of the content of virtue -- that to be virtuous is to promote the good of all humankind -- fits well with Shaftesbury's teleological approach. For he believes that every thing is designed to promote the good of the system of which it is a part. And he also believes that every human being is a part of the system that is the human species as a whole. It is natural for him to think, therefore, that every human being is designed to promote the good of the human species as a whole. (It is important

to note, however, that this view of a system and its parts explains only Shaftesbury's view of the content of goodness, which is something that non-human can also attain. Virtue or merit, which humans alone can attain, involves not merely acting for the good of the system but performing such actions in a self-aware or reflective manner.)

Shaftesbury also consistently maintains that in addition to promoting the good of humanity, virtue promotes the happiness of the virtuous person him or herself, just as vice harms humanity as a whole as well as making the vicious person unhappy. On Shaftesbury's view, in other words, "virtue and interest may be found at last to agree" (*Characteristics* 167). Or as he puts it in the conclusion of the *Inquiry*, "And thus virtue is the good and vice the ill of everyone" (*Characteristics* 229-330). The coincidence of virtue and happiness is just what Shaftesbury's teleological approach should lead us to expect. For teleological thinking generally includes the idea that the best life for a being is one that fulfills the being's natural end or purpose, and being virtuous is the end or purpose for which humans were designed.

Shaftesbury corroborates this teleological connection between virtue and happiness by investigating the pleasures and pains of which human happiness and unhappiness consist. He begins this investigation by drawing a broad distinction between pleasures of the body and pleasures of the mind. He next contends that a person's happiness depends more on mental pleasures than on bodily pleasures. And he then seeks to show that living virtuously is by far the best way to gain the crucially important mental pleasures. Shaftesbury bases much of his "mental pleasures" argument for the connection between virtue and happiness on the idea that the mental pleasures are within one's own control, insulated from the vicissitudes of "fortune, age, circumstances and humour." As one of Shaftesbury's characters rhetorically asks, "How can we better praise the goodness of Providence than in this, 'That it has placed our happiness and good in things we can bestow upon ourselves'?" The importance Shaftesbury places on our control over our mental pleasures grows directly out of his appreciation for the Stoics. And indeed, it can be plausibly maintained that Stoicism is one of the strongest and most fundamental commitments of Shaftesbury's thought overall.

5. Attacks on Egoism

But although Shaftesbury believes that being virtuous makes a person happy, it would be wrong to label him an egoist. In fact, he launches many attacks on both psychological egoism and ethical egoism, attacks that have as their main target Hobbes and which clearly anticipate the influential anti-egoist arguments in Butler, Hutcheson and Hume.

Shaftesbury argues, first of all, that psychological egoism does a simply terrible job of explaining the wide spectrum of observable activities humans engage in. He ridicules, for instance, egoistic interpretations of things as "civility, hospitality, humanity towards strangers or people in distress," maintaining that it is much easier to explain such phenomena simply by positing real sociability and benevolence. He also points out that humans are often motivated by "caprice, zeal, faction and a thousand other springs, which are counter to self-interest" and suggests that the only way psychological

egoism can be plausibly maintained is at the expense of becoming tautologous.

Against ethical egoism, Shaftesbury argues that virtue can exist only if it's possible for people to be motivated by something other than self-interest. For a person's virtue, according to Shaftesbury, consists not of the actions he performs but of the motives he has for performing them. And the motive with which we identify virtue is benevolence, not self-interest. Shaftesbury emphasizes this point by drawing attention to the difference between a knave and a saint. We judge the saint virtuous, he explains, because we think he is motivated by something other than the selfishness of the knave. And if we came to believe that the saint's motive were mere selfishness, we would no longer judge him to be virtuous. As he puts it, "If the love of doing good be not of itself a good and right inclination, I know not how there can possibly be such a thing as goodness or virtue"(*Characteristics* 46).

Shaftesbury's belief that true virtue must flow from non-egoistic motives leads him to criticize sharply the emphasis many religious moralists place on reward and punishment in the afterlife. As one of his characters explains when summarizing the goal of the Inquiry, "[The author of the *Inquiry*] endeavors chiefly to establish virtue on principles by which he is able to argue with those who are not as yet induced to own a god or future state. If he cannot do thus much, he reckons he does nothing" (*Characteristics* 266). Shaftesbury eschews considerations of the afterlife in his case for virtue because he believes that persons who perform virtuous actions only because they desire reward and fear punishment have no real virtue in them at all. And persons who are constantly made to dwell on reward and punishment are likely to become overly concerned with their own "self-good and private interest," which must "insensibly diminish the affections towards public good or the interest of society and introduce a certain narrowness of spirit" (*Characteristics* 184). So an emphasis on reward and punishment cannot make people more virtuous, and it may very well make them less so (*Characteristics* 45-46).

Shaftesbury's anti-egoistic view also leads him to an interesting consideration of what we should say to someone who asks for a reason to be virtuous when he knows he will not be punished for vice, or, as Shaftesbury puts the question, "Why should a man be honest in the dark?" (*Characteristics* 58). At times Shaftesbury suggests that a person who asks this question is already lost to virtue -- that someone who cares about virtue for its own sake won't need another reason to act virtuously, and that someone who needs another reason doesn't have what it takes to be truly virtuous in the first place. At other times, Shaftesbury suggests that we should be honest even in the dark (i.e., virtuous even when we will not be punished for vice) because such conduct is a necessary condition for having a unified self at all (*Characteristics* 127). These suggestions of how to deal with the question "Why be moral?" are almost certainly antecedents of Hume's response to the sensible knave at the end of his *Enquiry concerning Morals*.

6. Attacks on Social Contract Theory

Another point on which Hume seems to be indebted to Shaftesbury is criticism of social contract theories. Shaftesbury argues the selfish beings Hobbes describes in his state of nature bear no

resemblance to humans as they actually are. For naturally, Shaftesbury contends, humans are sociable. And society is thus humankind's natural condition. "In short, if generation be natural, if natural affection and the care and nurture of the offspring be natural, things standing as they do with man and the creature being of that form and constitution he now is, it follows that society must be also natural to him and that out of society and community he never did, nor ever can, subsist" (*Characteristics* 287). Shaftesbury also argues that if Hobbes's description of an amoral state of nature were correct, then it would be impossible for Hobbes ever to establish a duty to obey the laws of society. For if there is no duty to keep one's promises in the state of nature, then the original contract cannot create a duty. And if the original contract does give rise to a duty, then there must have been a duty to keep one's promises even in the state of nature (*Characteristics* 51). Shaftesbury was not the first to criticize social contract theories in this way, but his version of this criticism is stated very clearly and was probably among the most influential.

7. Aesthetics

Shaftesbury's views on aesthetics were also very influential in the eighteenth and nineteenth centuries. The conception of beauty Shaftesbury proposes is very similar to his conception of virtue. Indeed, Shaftesbury contends that proper taste in morals and proper taste in art turn out to be much the same thing, and that this is because the beautiful and the good ("natural beauty" and "moral beauty" [*Characteristics* 65]) are themselves "one and the same" (*Characteristics* 330). Or as he puts it elsewhere, "Thus are the arts and virtues mutually friends and thus the science of virtuosos and that of virtue itself become, in a manner, one and the same" (*Characteristics* 150). Shaftesbury believes that we judge beauty through our affections, but that the beautiful itself is not dependent upon our affections. There is, he contends, a real aesthetic standard that is affection-independent. And what we must do is try to develop and improve our affections in order to acquire taste that is in line with that real aesthetic standard -- just as we must try to develop and improve our affections in order to acquire a character in line with the reality of virtue. Shaftesbury's teleology is crucial to his aesthetics as well, in that he equates beauty with the harmony of the universe. "For all beauty is truth" (*Characteristics* 65). In addition to these general aesthetic claims, Shaftesbury makes a great many specific stylistic points about literature, criticism, music and the visual arts.

8. Religion

Although he resisted complete identification with them, Shaftesbury's religious views share much with the English Deists, and he, like them, was a strong proponent of natural religion. Shaftesbury repeatedly advances versions of the argument from Design for the existence of God, and his general teleological approach is deeply theistic (it could perhaps be said that his teleology and his religion were one and the same thing). At the same time, Shaftesbury places little stock in the institutions and rituals of organized religion, and he maintains that the Scriptures are not self-verifying but must always submit to the judgment of our natural understanding. Shaftesbury argues that a truly religious frame of mine -- a frame of mind attuned to the harmonious workings of the universe as a whole -- bolsters one's commitment to virtue. But he also maintains that an atheist can be virtuous, and that bad religion is more destructive to

virtue than no religion at all. Shaftesbury criticized certain kinds of fanatical religious enthusiasm, but he also believed that the pinnacle of religious-moral-aesthetic experience consisted of an enthusiastic embrace of the spirit of the entire world.

Bibliography

Shaftesbury's Works

- *Characteristics of Men, Manners, Opinions, Times*, edited by Lawrence E. Klein, Cambridge: Cambridge University Press, 1999.
- *Second Characters or the Language of Forms by the Right Honourable Anthony, Early of Shaftesbury*, edited by Benjamin Rand, Cambridge: Cambridge University Press, 1914; reprinted, New York: Greendwood Press, 1969.
- *The Life, Unpublished Letters and Philosophical Regimen of Anthony, Earl of Shaftesbury*, edited by Benjamin Rand, London: Swan Sonnenschein, 1900.

Secondary Literature

Biography of Shaftesbury with extensive discussion of his thought as a whole:

- Voitle, Robert, *The Third Earl of Shaftesbury 1671-1713*, Baton Rouge: Louisiana University Press, 1984.

Book length treatment of Shaftesbury's thought as a whole:

- Grean, Stanley, *Shaftesbury's Philosophy of Religion and Ethics*, Athens: Ohio University Press, 1967.

Detailed discussions of many aspects of Shaftesbury's philosophy and its historical context:

- Darwall, Stephen, *The British Moralists and the Internal Ought: 1640-1740*, Cambridge: Cambridge University Press, 1995.
- Schneewind, J. B., *The Invention of Autonomy: A History of Modern Moral Philosophy*, New York: Cambridge University Press, 1998.

Specifically on Shaftesbury's views of morality:

- Grean, Stanley, "Self-Interest and Public Interest in Shaftesbury's Philosophy," *Journal of the History of Philosophy* 1964; 2: 37-46.
- Gill, Michael B., "Shaftesbury's Two Accounts of the Reason to Be Virtuous," *Journal of the*

History of Philosophy, 2000; 38(4): 529-548.
- Trianosky, Gregory W., “On the Obligation to be Virtuous: Shaftesbury and the Question, Why be Moral?” *Journal of the History of Philosophy* 1978; 16: 289-300.

Specifically on Shaftesbury’s views of religion:

- Bernstein, John A., “Shaftesbury’s Reformation of the Reformation: Reflections on the Relation between Deism and Pauline Christianity,” *Journal of Religious Ethics* 1978, 6: 257-278.
- Toole, Robert, “Shaftesbury on God and His Relationships to the World,” *International Studies in Philosophy* 1976; 8: 81-100.

Specifically on Shaftesbury’s views of aesthetics:

- McAllister, James W., “Scientists’ Aesthetic Judgments,” *British Journal of Aesthetics* 1991, 332-341.
- Townsend, Dabney, “Shaftesbury’s Aesthetic Theory,” *Journal of Aesthetics and Art Criticism* 1982, 205-213.

Specifically on Shaftesbury’s views of personal identity:

- Winkler, Kenneth P., “‘All Is Revolution in Us’: Personal Identity in Shaftesbury and Hume,” *Hume Studies* 2000 April; 26(1): 3-40.

Other Internet Resources

- Entry on Earl of Shaftesbury (Internet Encyclopedia of Philosophy)
  Good general account of Shaftesbury's thought.
- Anthony Ashley Cooper, 3rd Earl of Shaftesbury (Thoemmes Press Encyclopedia)
  Another good general account of Shaftesbury's thought.
- Deism (Francis Aveling, Catholic Encylcopedia)
  On Shaftesbury's relationship to Deism
- Selected Bibliography on Shaftesbury (Laurent Jaffro, Université Paris 1 Panth&#eacute;on-Sorbonne)
  Excellent bibliography of works by and about Shaftesbury.
- Stanford Encyclopedia of Philosophy: Francis Hutcheson in Scottish Philosophy in the Eighteenth Century by Alexander Broadie (University of Glasgow)
- Stanford Encyclopedia of Philosophy: Benjamin Whichcote in The Cambridge Platonists by Sarah Hutton (Middlesex University)

Related Entries

aesthetics | beauty | contractarianism | creationism | deism | egoism | Hobbes, Thomas | Hume, David | Locke, John | moral rationalism | moral sense theory or moral sentimentalism | Stoicism | teleology

Michael Gill
College of Charleston
gillm@cofc.edu

First published: March 12, 2002
Content last modified: March 12, 2002

The Cambridge Platonists

The Cambridge Platonists were a group of English seventeenth-century thinkers associated with the University of Cambridge. The most important philosophers among them were Henry More (1614-1687) and Ralph Cudworth (1617-1689), both fellows of Christ's College, Cambridge. The group also included Benjamin Whichcote (1609-1683), Peter Sterry (1613-1672), John Smith (1618-1652), Nathaniel Culverwell (1619-1651), John Worthington (1618-1671), all one-time fellows of Emmanuel College, Cambridge. Their younger followers included George Rust (d. 1670), Anne Conway (1630-1679) and John Norris (1657-1711). In so far as they all held the philosophy of Plato and Plotinus in high regard, the designation 'Platonist' is apt. However, they drew on a wide range of philosophical sources besides Platonism. Among ancient philosophers, they were well acquainted with Aristotle and with Stoicism. But they were also very much abreast of new developments in philosophy and science - with Descartes, Hobbes and Spinoza as well as Bacon, Boyle and the Royal Society. (Smith, Culverwell, Cudworth and More were among the first Englishmen to read Descartes). The framework within which they read and understood ancient and modern philosophy was that of the 'perennial philosophy' (*philosophia perenni*s) proposed originally by Italian Renaissance philosophers such as Marsilio Ficino, and Agostino Steucho, but also employed by Gottfried Wilhelm Leibniz. Not only did they share the Renaissance Humanist regard for the achievements of ancient philosophy, but like the Humanists of the Renaissance, their interest was dictated by their sense of the relevance of classical philosophy to contemporary life. They also emphatically repudiated the scholasticism that prevailed in academic philosophy and took a lively interest in the developments that brought about the scientific revolution. They therefore form part of the philosophical revolution of the seventeenth century, especially since they sought an alternative philosophical foundation to Aristotelianism which was waning fast in the face of challenges from scepticism and competing alternative philosophies, notably those of Hobbes and Descartes. They were the first philosophers to write primarily and consistently in the English language.

One difference between the Cambridge Platonists and their more famous philosophical contemporaries is that they all had a theological background. Nevertheless, convinced of the compatibility of reason and faith, they regarded philosophy as the legitimate concern of theologians and are distinguished by the high value they accorded human reason. They and devoted their considerable philosophical learning to religious and moral issues, to defending the existence of God and the immortality of the soul, and to formulating a practical ethics for Christian conduct. They held the eternal existence of moral principles and of truth and that the human mind is equipped with the principles of reason and morality. Their optimistic view of human nature is underscored by their emphasis on the freedom of the will. Their anti-determinism lead them to propose arguments for human autonomy. They were all dualists for whom mind is ontologically prior to matter, and for whom the truths of the mind are superior to sense-

knowledge. They were nevertheless moderns in natural philosophy who accepted post-Galilean science, and propounded an atomistic theory of matter. But they repudiated mechanistic natural philosophy in favour of the view that spirit is the fundamental causal principle in the operations of nature.

Benjamin Whichcote

Benjamin Whichcote is usually considered to be the founding father of Cambridge Platonism, by virtue of the fact that so many of them studied at Emmanuel College when he taught there. During the Civil War period, Whichcote was appointed Provost of King’s College, Cambridge, and he served as Vice Chancellor of the University in 1650. However, he was removed from his post at King’s College at the Restoration in 1660. and was obliged to seek employment elsewhere, as a clergyman in London. The interruption to his academic career may explain why he never published any philosophical treatises as such. The main source for his philosophical views are his posthumously-published sermons and aphorisms. Whichcote’s tolerant, optimistic and rational outlook set the intellectual tone for Cambridge Platonism. Whichcote’s philosophical views are grounded in his liberal theology. He held that God being supremely perfect is necessarily good, wise and loving. Whichcote regarded human nature as rational and perfectible, and he believed that it is through reason as much as revelation that God communicates with man. ‘God is the most knowable of any thing in the world’ (Patrides, 1969, p.58). Without reason we would have no means of demonstrating the existence of God, and no assurance that revelation is from God. By reason Whichcote did not mean the disputatious logic of the schools but discursive, demonstrative and practical reason enlightened by contemplation of the divine. He held that moral principles are immutable absolutes exist independently of human minds and institutions, and that virtuous conduct is grounded in reason. Whichcote’s *Aphorisms* amount to a manual of practical ethics which amply illustrates his conviction that the fruit reason is not ‘bare knowledge’ but action, or knowledge which ‘doth go forth into act’. It is through reason that we gain knowledge of the natural world, and recognise natural phenomena as ‘the EFFECTS OF GOD’. Although Whichcote’s published writings do not discuss natural philosophy as such, his recognition of the demonstrative value of natural philosophy for the argument from design anticipates the use of natural philosophy in the apologetics of Cudworth and More.

Culverwell, Smith and Sterry

Whichcote's optimism about human reason and his conviction that philosophy properly belonged within the domain of religion, is an outlook shared by the other Cambridge Platonists, all of whom affirmed the compatibility of reason and faith. The fullest statement of this position is Henry More's *The Apology of Henry More* (1664) which sets out rules for the application of reason in religious matters, stipulating the use of only those 'Philosophick theorems' which are 'solid and rational in themselves, nor really repugnant to the word of God'. Like Whichcote, Peter Sterry, John Smith and Nathaniel Culverwell are known only through posthumously published writings. The first published treatise by any of the Cambridge Platonists was Nathaniel Culverwell's *An Elegant and Learned Discourse of the Light of Nature* of 1652. Like the other Cambridge Platonists Culverwell emphasises the freedom of the will and proposes an innatist epistemology, according to which the mind is furnished with 'clear and indelible Principles' and reason an 'intellectual lamp' placed in the soul by God to enable it to understand God's will promulgated in the law of nature. These innate principles of the mind also moral principles. The soul is a divine spark, which derives knowledge by inward contemplation, not outward observation. Drawing on Suarez and Aquinas, as well as Platonist and contemporary philosophy, Culverwell was the only Cambridge Platonist to invoke natural law.

After studying at Emmanuel College, John Smith taught mathematics at Queen's College until his premature death in 1652. His *Select Discourses* (1659) discusses a number of metaphysical and epistemological issues relating to Chrstian belief - the existence of God, immortality of the soul and the rationality of religion. Smith outlines a hierarchy of four grades of cognitive ascent from sense combined with reason, through reason in conjunction with innate notions, and, thirdly, through disembodied, self-reflective reason; and finally divine love.

Peter Sterry too was educated at Emmanuel College, Cambridge. His *A Discourse of the Freedom of the Will* (1675) is the most visionary of all the writings of the Cambridge Platonists. But Sterry was more involved with events outside Cambridge than most of the others, on account of the fact that he was chaplain first to Lord Brooke and then to Oliver Cromwell. After the death of Cromwell he retired to a Christian community in East Sheen.. In his *Discourse* Sterry argues that freedom consists in acting in accordance with ones nature, appropriately to ones level of being, be it plant, animal or intellectual entity. Human liberty is grounded in the divine essence and entails liberty of the understanding and of the will.

Henry More

A life-long fellow of Christ's College, Cambridge, Henry More was the most prolific of the Cambridge Platonists. He was also the most directly engaged in contemporary philosophical debate: not only did he enter into correspondence with Descartes (between 1648 and 1649) but he also wrote against Hobbes, and was one of the earliest English critics of Spinoza (*Demonstrationem duarum* and *Epistola altera* both published in his *Opera omnia*, 1671). Although he eventually became a critic of Cartesianism he initially

advocated the teaching of Cartesianism in English Universities. More's published writings included, besides philosophy, poetry, theology and bible commentary. His main philosophical works are his *An Antidote Against Atheism* (1653), his *Of the Immortality of the Soul* (1659), *Enchiridion metaphysicum* (1671), and *Enchiridion ethicum* (1667). In these writings, More elaborated a philosophy of spirit which explained all the phenomena of mind and of the physical world as the activity of spiritual substance controlling inert matter. More conceived of both spirit and body as spatially extended, but defined spiritual substance as the obverse of material extension: where body is inert and solid, but divisible; spirit is active and penetrable, but indivisible. It was in his correspondence with Descartes that he first expounded his view that all substance, whether material or immaterial, is extended. As an example of non-material extension he proposed space, within which material extension is contained. He went on to argue space is infinite, anticipating that other native of Grantham, Isaac Newton. More also argued that God who is an infinite spirit is an extended being (*res extensa*). There are, therefore, conceptual parallels between the idea of God and the idea of space, a view which he elaborates in *Enchiridion metaphysicum*, where he argues that the properties of space are analogous to the attributes of God (infinity, immateriality, immobility etc.).

Within the category of spiritual substance More includes not just the souls of living creatures and God himself but the main intermediate causal agent of the cosmos, the Spirit of Nature (or 'Hylarchic Principle'). According to More the Spirit of Nature is the interface between the divine and the material. As a concept, it has affinities with Plato's *anima mundi* (world soul), and the Stoics' *pneuma*. The Spirit of Nature can also be understood as encapsulating 'certain general Modes and Lawes of Nature' (More, *A Collection*, Preface, p. xvi) since it is the Spirit of Nature that is responsible for uniting individual souls with bodies, and for ensuring the regular operation of non-animate nature. It is a 'Superintendant Cause' which combines efficient and telelological causality to ensure the smooth-running of the universe according to God's plan. More sought, by this hypothesis, to account for phenomena that apparently defy the laws of mechanical physics (for example the inter-vortical trajectory of comets, the sympathetic vibration of strings and tidal motion). More underpinned his soul-body dualism by his theory of 'vital congruity' which explains soul-body interaction as a sympathetic attraction between soul and body engineered by the operation the Spirit of Nature.

Like the other Cambridge Platonists, More was a religious apologist who used philosophy in defence of theism against the claims of rational atheists. The most important statement of More's theological position his *An Explanation of the Grand Mystery of Godliness* appeared in 1664, and propounds, in opposition to Calvinist pessimistic voluntarism, a moral, rational providentialism in which he vindicates the goodness and justice of God by invoking the Origenist doctrine of the pre-existence of the soul. The most consistent theme of his philosophical writings, are arguments for demonstrating the existence and providential nature of God. Indeed the foundation stone of More's apologetic enterprise is his philosophy of spirit, especially his arguments for the existence of incorporeal causal agents, that is, souls or spirits. Furthermore, More attempted to answer materialists like Thomas Hobbes whom he perceived as an atheist on account of his dismissal of the idea of incorporeal substance as non-sensical. More's strategy was to show that the same arguments that materialists use demonstrate the existence and properties of body, also support the obverse, the existence of incorporeal substances. In this way More sought to demonstrate that the *idea* of incorporeal substance, or spirit, was as intelligible as that of corporeal

substance, i.e. body. Like Plato (in *Laws* 10), More argues that the operations of the nature cannot be explained simply in terms of the chance collision of material particles. Rather we must posit some other source of activity, which More identifies as 'spirit'. It is a short step, he argues, from grasping the concept of spirit, to accepting the idea of an infinite spirit, namely God.

More underpins these *a priori* arguments for the existence of spirit, with a wide range of *a posteriori* arguments, taken from observed phenomena of nature to demonstrate the actions of spirit. Through this excursus into observational method he accumulated a wide variety of data ranging from experiments conducted by Robert Boyle and members of the Royal Society, to supernatural effects including cases of witchcraft and demons. He was censured by Boyle for misappropriating his experiments to endorse his hypothesis of the Spirit of Nature, and his apparent credulity, appears inconsistent with his otherwise rational philosophy, though it must be said that belief in witchcraft was not unusual in his time, and, secondly, was entirely consistent with the theory of spirit according to which to deny the existence of spirits good or evil, leads, logically to the denial of the existence of God. As he put it, alluding to James I's defence of episcopacy, 'That saying is no less true in Politicks " *No Bishop, no King*," than this in Metaphysicks, "*No Spirit, no God*" ' (More, 1662, *Antidote*, p. 142). His most well-known fellow-believer was Royal Society member, Joseph Glanvill (1636-1680), whose *Sadducismus triumphatus*, More edited.

More also published a short treatise on ethics entitled *Enchiridion Ethicum* (1667, translated as *An Account of Virtue*). Indebted to Descartes' theory of the passions this argues that knowledge of virtue is attainable by reason, and the pursuit of virtue entails the control of the passions by the soul. Motivation to good is supplied by rightly-directed emotion, while virtue is achieved by the exercise free will or *autoexousy* (More uses the same term as Cudworth), that is the 'Power to act or not act within ourselves'. Anticipating Shaftesbury's concept of moral sense More posits a special faculty of the soul combining reason and sensation which he calls the 'Boniform Faculty'.

More used a number of different genres for conveying his philosophical ideas. The most popular among these were his *Philosophical Poems* (1647) and his *Divine Dialogues* (1668). In *Conjectura cabbalistica* (1653), he presented core themes of his philosophy in the form of an exposition of occulted truths contained in the first book of Genesis. Subsequently he undertook a detailed study of the Jewish kabbalah which were published in Knorr von Rosenroth's *Kabbala denudata* (1679). These studies were based on the belief, then current, that kabbalistic writings contained, in symbolic form, original truths of philosophy, as well as of religion. Kabbalism therefore exemplified the compatibility of philosophy and faith. In addition to philosophy More published several studies of biblical prophecy (e.g. *Apocalypsis apocalypseos*, 1680, *Paralipomena prophetica*, 1685). In 1675, More prepared a Latin translation of his works, *Opera omnia* which ensured his philosophy reached a European audience as well as an English one.

Ralph Cudworth

Like his friend Henry More, Cudworth spent his entire career was as a teacher at the University of

Cambridge, where, in 1647, he was appointed Regius Professor of Hebrew and Master of Clare College. In 1654 he was elected Master of Christ's College, a post he held until his death. Cudworth published only one major work of philosophy in his lifetime, *The True Intellectual System of the Universe* (1678). Among the papers he left at his death, were the treatises published posthumously as *A Treatise Concerning Eternal and Immutable Morality* (1731) and his *A Treatise of Freewill* (1848). These papers also included two further manuscript treatises on the topic of 'Liberty and Necessity', which have never been published. Cudworth's humanistic erudition and baroque style have occluded the originality of his contribution to English philosophy and helped ensure to his undeserved neglect in the annals of English philosophy.

Cudworth's *True Intellectual System* propounds a an anti-determinist system of philosophy grounded in his conception of God as a fully perfect being, infinitely wise and good. The created world reflects the perfection, wisdom and goodness of its creator. It must, therefore be orderly, intelligible, and organised for the best. This anti-voluntarist understanding of God's attributes is also the foundation of epistemology and ethics, since God's wisdom and goodness are the guarantors of truth and of moral principles. By contrast, a philosophy founded on a voluntaristic conception of the deity would have no ground of certainty or of morality because it would depend on the arbitrary will of God who could, by arbitrary fiat, decree non-sense to be true and wrong to be right. It follows that misconceptions of God's attributes, which emphasise his power and will, result by definition in false philosopical systems with sceptical and atheistic implications.

Much of *The True Intellectual System* amounts to an extended *consensus gentium* argument for the existence of a supreme deity. By demonstrating from ancient sources that most ancient philosophers were theists, Cudworth sought to argue that theism is compatible with philosophy. Among the non-theists, Cudworth identifies four schools of atheistic philosophy, each of which is a type of materialism - Hylozoic atheism which attributes life to matter, Hylopathian atheism, which attributes all to matter, Cosmo-plastic atheism which makes the world-soul the highest numen. Each of these ancient brands of atheism has its latter-day manifestations in philosophers such as Hobbes (an example of a Hylopathian atheist) and Spinoza (a latter-day Hylozoist).

The true philosophy of Cudworth's intellectual system combines mechanistic atomism with Platonic metaphysics. Cudworth conceives this as having originated with Moses from whom it was transmitted via Pythagoras to Greek and other philosophers. Cudworth subscribed to the atomistic hypothesis that the created universe is constituted of particles of inert matter - indeed he regarded Cartesian mechanical philosophy as a recently revived variety of ancient atomism of Moses. But, for Cudworth, as for Plato, soul is ontologically prior to the physical world. Since motion, thought and action cannot be explained in terms of material particles, haphazardly jolted together, there must be some guiding originator, namely soul or spirit. Part of the appeal to Cudworth of the new 'mechanical' philosophy of Descartes was that, it was the dualism which underscored it. This Cudworth therefore interpreted Descartes' dualism with some latitude to explain all movement, life and action in terms of the activity of spirits operating on inert matter. Although he was critical of aspects of Cartesianism, he seized on its value for religious apologetics as instrumental in persuading materialists of the existence of spiritual agency.

In place of the mechanical explanation of the operations of nature, but proposed instead his hypothesis of 'the Plastick Life of Nature'. Similar in conception to More's Hylarchic Principle, Cudworth's Plastic Nature is a formative principle which acts as an intermediary between the divine and the natural world, maintaining the mundane operations of the physical universe. Plastic Nature, is the means whereby God imprints His presence on his creation and makes His wisdom and goodness manifest (and therefore intelligible) throughout created nature. In one respect, as Cudworth points out, Plastic Nature is a summation of all the laws of motion. At the same time he conceives it as having substance, as being some kind of spirit which carries out its functions unconsciously. Plastic Nature therefore has affinities with the Platonic *anima mundi*. The hypothesis of Plastic Nature also entails a teleological principle, which accounts for the design and purpose in the natural world. Plastic nature enables Cudworth to account for the providential ordering of the universe without falling into the trap of occasionalism. For by it he explains God's immanence in the world, without requiring immediate divine intervention in the minutiae of day-to-day operations in the natural world.

The Platonist principle that mind precedes the world lies at the foundation of Cudworth's epistemology which is discussed in *A Treatise of Eternal and Immutable Morality*. This is the most fully developed theory of knowledge by any of the Cambridge Platonists, and the most extensive treatment of innatism by any seventeenth-century philosopher. For Cudworth, as for Plato, ideas and moral principles 'are eternal and self-subsistent things'. The knowability of the world is explained in terms of the basic Platonic principles of archetype and ectype (form and copy). Cognition depends on the same principles, for just as the created world is a copy of the divine archetype, so also human minds contain the imprint of Divine wisdom and knowledge. The ideas in each individual mind are therefore the same in all minds. Since the human mind mirrors the mind of God, it is ready furnished with ideas and the ability to reason. Cognition therefore entails recollection and the ideas of things with which the mind thinks are therefore anticipations - Cudworth adopts the Stoic term *prolepsis* to denote them. But cognition is not a passive process. Rather it entails active participation of the mind. 'Knowledge', writes Cudworth, 'is not a passion from anything without the mind, but an active exertion of the inward strength, vigour, and power of the mind, displaying itself from within' (Cudworth, 1996, p. 74). Although innate knowledge is the only true knowledge, Cudworth's epistemology does not reject sense knowledge. On the contrary, sensory input is essential for knowledge of the body and the external world. And the external world is, intrinsically, intelligible, since it bears the imprint of its creator in the order and relationship of its component parts. However, raw sense data is not, by itself, knowledge. But it requires mental processing in order to become knowledge. As Cudworth puts it, we cannot understand the book of nature unless we know how to read.

Cudworth's theory of the mind as active is matched by an anti-determinist ethics of action, according to which the soul freely directs itself towards the good. In *A Treatise* Cudworth argues not only that ideas exist independently of human minds, but also the principles of morality are eternal and immutable. In a concerted attack on Hobbesian moral relativism, Cudworth, argues that the criteria of right and wrong, good and evil, justice and injustice are not a matter of convention, but are founded in the goodness and justice of God. Like Plato in the *Euthyphro*, Cudworth argues that it is not God's will that determines goodness, but that God wills things because they are good. The exercise of virtue is not, however, a passive process, but requires the free exercise of the individual will. Cudworth sets out his theory of free

will in three treatises on 'Liberty and Necessity', only one of which has been published, and that posthumously - *A Treatise of Freewill* (1848). According to Cudworth, the will is not a faculty of the soul, distinct from reason, but a power of the soul which combines the functions of both reason and will in order to direct the soul towards the good. Cudworth's use of the terms 'hegemonikon' (taken from Stoicism) and 'autexousion' (taken from Plotinus) underlines the fact that the exercise of will entails the power to act. It is internal direction, not external compulsion that induces us to act either morally or immorally. Without the freedom (and therefore power) to of act, there would be no moral responsibility. Moral conduct is active, not passive. Virtuous action is therefore a matter of active internal self-determination, rather than determination from without.

According to the moral psychology outlined in *A Treatise of Freewill*, the 'hegemonikon' has an integrative function within the soul, combining, on the one hand, the functions of will and reason, and, on the other the lower, animal, appetites of the soul with to the higher intellectual functions of the soul. In this way Cudworth bridges the divide between soul and body that characterises Cartesianism. Furthermore, Cudworth conceives of the *hegemonikon* not simply as the soul but the whole person, 'that which is properly we ourselves' (Cudworth, 1996, p. 178). Cudworth's concept of *hegemonikon* lays the basis for a concept of self identity founded in a subject that is at once thinking, autonomous and end-directed. Cudworth did not (as far as is known) develop a political philosophy. Nevertheless, the political implications of his ethical theory set him against Hobbes, but also, in many ways anticipate John Locke.

Legacy

After Hobbes and Locke, the Cambridge Platonists deserve to be considered an important third strand in English seventeenth-century philosophy. Their critique of Descartes, Hobbes, Spinoza has ensured that they are never ignored in philosophical history but they have yet to receive full recognition in their own right. Evidence from publication and citation suggests that their philosophical influence was more far-reaching than is normally recognised in modern histories of philosophy. The works of Cudworth and More were available to a European readership through Latin translation -More's *Opera omnia* appeared in 1675-9, and Cudworth's entire published works were translated into Latin by Johann Lorenz Mosheim and published in 1733. Among the immediate philosophical heirs of the Cambridge Platonists, mention should be made of Henry More's pupil, Anne Conway (1631-1679), one of the very few female philosophers of the period. Her *Principles of the Most Ancient and Modern Philosophy* (1692) entails a critique of More's dualistic philosophy of spirit, proposing instead a metaphysical monism that anticipates Leibniz. Another figure linked to More was John Norris (1657-1711) who was to become the leading English exponent of the philosophy of Malebranche. Whichcote's philosophical wisdom was admired by Anthony Ashley Cooper, third Earl of Shaftesbury who published his *Select Sermons* in 1698. Shaftesbury's tutor, John Locke was the intimate friend of Cudworth's philosophical daughter, Damaris Masham. The impact of Cudworth on Locke has yet to be fully investigated. Richard Price, and Thomas Reid were both indebted to Cudworth, whose theory of Plastic Nature was taken up in vitalist debates in the French enlightenment. Leibniz certainly read Cudworth and More. The intellectual legacy, the Cambridge Platonists extends not just to philosophical debate in seventeenth-century England but into European and Scottish Enlightenment thought and beyond.

Bibliography

Primary Sources

- Conway, Anne. *The Principles of the Most Ancient and Modern Philosophy*, London, 1692. Modern translation by T. Corse and A. Coudert. Cambridge: Cambridge University Press, 1996.
- Cragg, G. R., (ed.). *The Cambridge Platonists*. New York: Oxford University Press, 1968.
- Cudworth, Ralph *A Treatise Concerning Eternal and Immutable Morality*, London, 731, Modern edition ed. S. Hutton (Cambridge: Cambridge University Press, 1996). Patrides, C. A. (ed.). *The Cambridge Platonists*. (London: Arnold, 1969).
- Cudworth , Ralph. *The True Intellectual System of the Universe*. (London, 1678*)*. Facsimile reprint, Stuttgart-Bad Canstatt: Friedrich Frommann Verlag, 1964.
- Culverwell, Nathaniel *An Elegant and Learned Discourse of the Light of Nature*, London, 1652. Modern edition by R.A. Greene and H. McCallum. (Toronto, 1971).
- More, Henry. *A Collection of Several Philosophical Writings*, London, 1662.
- More, Henry. *Opera omnia*. 3 vols. London 1675-1679. Facsimile reprint, Hildesheim: Olms, 1966).
- Smith, John *Select Discourses*, ed. J. Worthington. London 1660. Facsimile reprint, New York and London: Garland, 1978.
- Sterry, Peter*, A Discourse of the Freedom of the Will*. London, 1675.
- Whichcote, Benjamin, *The Works of the Learned Benjamin Whichcote*, 4 vols. Aberdeen, 1751. Facsimile reprint New York, 1977.
- Whichcote, Benjamin, *Some Select Notions*. London, 1685.
- Whichcote, Benjamin, *Select Sermons*, with a Preface by Anthony Ashley Cooper, Third Earl of Shaftesbury. London, 1698.

Secondary Sources

- Darwall, S. *British Moralists and the Internal Ought*. Cambridge: Cambridge University Press. 1992.
- Gabbey, Alan. '*Philosophia cartesiana triumphata*: Henry More and Descartes, 1646-71'. In T.M. Lennon et al., *Problems in Cartesianism* , pp. 171-249. Kingston and Montreal: Queens McGill University Press, 1982.
- Hall, Rupert. *Henry More. Magic Religion and Experiment*. Oxford: Blackwell, 1990.
- Hutton, Sarah (ed.). *Henry More (1614-1687). Tercentenary Studies*. Dordrecht: Kluwer, 1990.
- Hutton, Sarah 'Lord Herbert and the Cambridge Platonists'. In S. Brown (ed.). *British Philosophy and the Age of Enlightenment*, *Routledge History of Philosophy*, vol. 5. London: Routledge, 1995.
- Hutton, Sarah 'The Cambridge Platonists'. In S. Nadler (ed.). *Blackwell Companion to Early Modern Philosophy*. Oxford: Blackwell, forthcoming.
- Koyré, Alexander. *From the Closed World to the Infinite Universe*. Baltimore, MA: Johns Hopkins University Press, 1957.

- Passmore, J.A. *Ralph Cudworth, an Interpretation*. Cambridge: Cambridge University Press, 1951.
- Rogers, G.A.J., J.-M. Vienne, Y.-C. Zarka (eds). *The Cambridge Platonists in Philosophical Context. Politics, Metaphysics and Religion.* Dordrecht: Kluwer Academic Publishers, 1997.
- Scott, Dominic *Recollection and Explanation. Plato's Theory of Learning and its Successors*. Cambridge University Press, 1990.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Conway, Lady Anne

Sarah Hutton
Middlesex University
S.Hutton@mdx.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 3, 2001
Content last modified: October 3, 2001

Eighteenth Century German Philosophy Prior to Kant

In Germany, the eighteenth century was the age of enlightenment, the age, that is, that called for the independence of reason. Although the ethos of this age found its clearest (and certainly its most famous) articulation towards the end of the century with Immanuel Kant and his critical philosophy, he was not the first to issue this call. Instead, that task fell to Christian Thomasius (Thomas) at the end of the seventeenth century. It was then taken up and further developed in a theological (pietist) direction by a number of minor figures, the Thomasians, and reissued in a rationalist direction in the early and middle part of the eighteenth century by Christian Wolff and his followers. The development of their position(s) as well as their philosophical (dis)agreements took place by and large at the University of Halle and against the context of pietism.

Christian Thomasius (1655-1724)

Although Thomasius is now largely forgotten, he was a pivotal figure in early eighteenth century enlightenment thought. In this context, however, he was also a somewhat ambiguous figure. On the one hand, he was clearly an innovator. Both a lawyer and a philosophy professor, he advocated the independent use of healthy reason, fought against prejudice, against belief in any of the then prevailing superstitions, against any form of (religious) persecution, against the witch-hunt and the use of torture, and in general, against any form of intolerance. He took issue with dependence on authority and the school philosophy's dependence on the syllogism. He lectured in German rather than the traditional Latin or the then fashionable French and was the first person, in Germany, to found a popular monthly journal, written by and large in German, devoted to book reviews, the *Monatsgespräche* (*Monthly Conversations*). On the other hand, especially when it came to matters of morality, he was more of a traditionalist. Here he retained distinctly non-Enlightenment ideas, particularly the belief in an evil will and the belief in the necessity of God's salvation.

Biography/Work

By way of background to both Thomasius's and Wolff's life it is important to note that the Germany of the eighteenth century was a country split into numerous states, each of which had its own government. There was no central government. Reeling from the effects of the Thirty Years War, many of its states did not have freedom of speech, freedom of religion, or, for that matter, a ‘national’ culture, though, given the variety of state governments, some states had relatively more freedom than others. It may seem surprising, therefore, that Christian Thomasius was able to issue the call to the enlightenment at the end of the seventeenth century, though not at all surprising that both Thomasius and Wolff were subject to arbitrary political power.

The son of jurist and philosopher Jakob Thomasius, Christian received his education at the University of Leipzig and his law degree at the University of Frankfurt an der Oder (in Eastern Germany) in 1679. He spent the early part of his career in his hometown Leipzig (in the state of Saxony), as a lawyer and (private) lecturer at the university there, but his controversial views and manner of expressing them, in particular, in the monthly journal *Monatsgespräche*, led to the prohibition to publish and hold lectures (private and academic) in Leipzig (and Saxony) in 1690. He was, however, welcome in the neighboring and comparatively more open-minded Halle (in the state of Brandenburg/Prussia), and was instrumental in founding the university there in 1694. He remained in Halle for the rest of his life, refusing an invitation to return to Leipzig in 1709.

Thomasius's body of work can be roughly divided into three parts. In his early Leipzig years, he was primarily interested in matters of law, particularly, following his father, in Pufendorf's natural law theory. This period ends around 1688 with the publication of *Institutiones jurisprudentiae divinae* (*Institutions of Divine Jurisprudence*) in which he sought to complete Pufendorf's project of divorcing natural law from theology. This year, as well, saw the publication of *Introductio ad philosophiam auliam* (*Introduction to Court Philosophy*), a text that is somewhat misnamed since it has less to do with proper conduct or even thought at court than with the proper use of reason, a topic that Thomasius would take up in greater detail in his 1691 *Introduction to the Doctrine of Reason*.

In general, 1687-8 seems to have marked an endpoint of sorts and the beginning of the second major stage, the more clearly philosophical one, of Thomasius's life. Even though he would remain in Leipzig for another two years, he had clearly broken with tradition by 1687, when he began lecturing and publishing in German. In 1688, he began the publication of the controversial *Monatsgespräche* (which appeared monthly until April 1690) and turned his attention to matters of theoretical and practical philosophy. Here two sets of books, the *Einleitung* and *Ausübung der Vernunftlehre* (*Introduction* and *Application of the Doctrine of Reason*), and the *Einleitung* and *Ausübung der Sittenlehre (Introduction* and *Application of Moral Theory*) that appeared in Halle between 1691 and 1696 mark the second part of his career. Written by and large in German with a minimum of technical terminology, these books were intended not for an audience of experts, but instead, as the subtitle to the *Introduction to the Doctrine of Reason* specifies, for a general audience of “all rational persons of whatever social standing and sex…”

During the late 1690s (and after a religious crisis that led to an at least temporary (re)affirmation of his pietist beliefs), Thomasius produced two works on metaphysics that endorsed a mystical variety of vitalism. In subsequent years, his interests shifted back to matters of law. This was the third part of his life and will not be further considered in this context.

Philosophy

Thomasius's philosophical stance was an empiricist one, not the rationalism that we find in much of the philosophical tradition and with Wolff. It is true that his belief in natural human reason and its capacity to find truth suggests a mild rationalism, but Thomasius abhorred innate ideas and maintained that all knowledge, all thought, begins with sense perception. This strong sensationism (which has similarities with Locke's position) was coupled, as has already been noted, with an enlightenment stance, in the sense that it was governed by the conviction that knowledge, truth and morality are the purview of everyone, not merely the elect few. The latter is particularly evident in the differentiation between *Gelehrtheit* and *Gelahrtheit* that he drew at the outset of the *Introduction to the Doctrine of Reason*. *Gelehrtheit* or academic learning is the domain of experts who are familiar with syllogistic logic, metaphysics, epistemology, and theology, but *Gelahrtheit* or practical learning is available to everyone with a healthy reason who pursues knowledge not for its own sake but for the use-value it has in daily life.

Thomasius's enlightenment convictions are similarly evident in his eclecticism. Though generally deemed a negative stance, this is not the case for Thomasius. He considers it positively as a corrective to any form of sectarian dogmatism. By his own account, he was influenced by several of his predecessors, notably, in Germany, Grotius and Pufendorf and, in England, Hobbes and Locke, and he appropriated those aspects of their theories that he found conducive to his overall aim: the spread of the Enlightenment ethos, understood here as the project of ensuring a healthy reason, one that can discover truth, that can lay open contradictions and fight prejudices.

Given Thomasius's basic presuppositions of where knowledge is likely to be found (in daily life rather than abstract speculation) and who is most likely to attain it (the person who has a healthy reason, not one

corrupted by prejudices), it is likely not surprising that his epistemology was not a theoretical one. His two books on theoretical philosophy, the *Introduction to the Doctrine of Reason* and the *Application of the Doctrine of Reason*, are books on truth. They are not, however, books on truth in the traditional sense. Thomasius did not develop a philosophical conception of truth or of the condition of its possibility. He seems to have simply adopted a correspondence theory of truth and to have taken the harmony of thought and thing as a given. Certainly, this harmony was not the problem for him that it was for 17th century thought and that it would be again in the later part of the 18th century (with Wolff, Knutsen and Kant). What mattered to Thomasius is the enlightenment optimism that truth is possible and, moreover, accessible to everyone. His *Introduction*, accordingly, was presented, as specified by the book's subtitle, as providing the means by which "all rational persons, of whatever social standing and sex, are shown in an understandable manner, and without the aid of syllogisms, how to differentiate between the true, the probable and the false, and to find new truths." It is a book of instruction in proper or correct use of healthy reason.

His *Application* continued this theme, though this time by providing people with the means of avoiding error. Avoiding error involves the eradication of prejudices, which are among the causes of the corruption of reason. That, in turn, is accomplished through what he identifies as dogmatic doubt, not the Cartesian doubt that deems everything false so as to find a first indubitable principle, a useless enterprise, according to Thomasius. Dogmatic doubt is the doubt about particular things, beliefs, and opinions, and this he found healthy and conducive to preventing error.

Thomasius's enlightenment convictions are not, however, as straightforward as might appear. He did believe in natural reason's capacity to overcome corruption, but even as he adhered to this view, he held that an evil will is at the root of this corruption, and that we require God's grace. This conflict is particularly evident in his moral philosophy. While his initial presentation was an optimistic affirmation of the viability of a moral position he identified as one involving "rational love" (*vernünftige Liebe*) as the only means to a "happy, courteous and cheerful life," as indicated by the subtitle of the *Introduction to Moral Theory*, by 1696, he had become disenchanted with this view. Human self-interest and an evil will stand in its way.

Thomasius's moral theory is a theory of the will. He held that in moral matters, the will dominates reason. Though human beings have free choice if not externally constrained, the will is not free. Rather, it is dominated by human affects; our passions, impulses, and desires. Like Hobbes, Thomasius believed that even though subject to such inner (psychological) constraints, the will still chooses (with the aid of reason); it consciously wills. And a conscious choice is precisely what is required for a (good) action to be considered moral, a good instinct or good inclinations may make us good, may even be desirable, but by itself this is not enough to make us moral. Morality requires a conscious act of will. The trouble with morality arises because the will is determined by evil desires, in particular, lust, ambition, and avarice. Although there are noble sentiments as well, which similarly influence the will, they are in conflict with the negative dispositions. The conflict can be brought to a positive conclusion only by appeal to divine grace (God's salvation).

This ambivalence and return to theology aside, Thomasius's moral position is an interesting one. The

theory of rational love is based on the fundamental equality of human beings as well as on their ability to think and choose independently (of authority). Ultimately, Thomasius's ethics is a social ethics. The theory is other-directed, and given the absence of laws and principles, constitutes a nice contrast to the formalist universalist ethics Kant would develop by the end of the next century. At the same time, the lack of any way of making this theory applicable in a context governed not by similar but instead by conflicting interests, makes something like a fomalist ethics an inevitability. By the end of the *Introduction to Moral Theory*, even Thomasius recognises that “rational love” will function only in relatively harmonious contexts, in others, particularly those characterised by unequal power positions, justice may well be required.

Christian Wolff (1679-1754)

It is without question that Christian Wolff was the most important German philosopher in the early and middle portion of the eighteenth century. Dissatisfied with the, to his mind, arbitrary eclecticism advocated by Thomasius and his ‘school,’ and equally dissatisfied with the scholastic school metaphysics which, he thought, lacked rigour, he produced a systematic philosophical system in reply. Indeed, his philosophy is reputed to constitute the most coherent systematic whole produced in the 18th century prior to Kant. As an enlightenment thinker, albeit a rationalist one, he was, like Thomasius, committed to public education. He saw philosophy, which he conceived as world-wisdom (*Welt-Weisheit*), as the means to public enlightenment and, in line with the mood of the time, the purview of everyone, not just of philosophers or experts. To make it accessible, he wrote in German, at least during his Halle years, and introduced much of the German philosophical terminology that is still used (the concepts of *Bewußtsein*, *Vorstellung*, and *Begriff* have their origin here).

Biography/Work

Wolff's birthplace was Breslau (then in Eastern Germany, now in Poland). At an early age, dissatisfied with orthodox theology, he turned to the study of mathematics as offering the best means to certainty. He studied theology and mathematics at the University of Jena and gained his Master's degree in Leizpig in 1702. Influenced by von Tschirnhaus and while working in Leipzig as a private lecturer, he wrote a dissertation seeking to apply the mathematical method to problems in practical philosophy. This drew Leibniz's attention to him. They began a correspondence that continued until Leibniz's death in 1716. Sponsored by Leibniz and von Tschirnhaus, he was appointed Professor of Mathematics and Natural Science at the University of Halle in 1707.

Since his position in Halle was predominantly as a teacher of mathematics -- a task at which he excelled -- Wolff did not begin lecturing and writing on philosophical matters until 1710. In mathematics, he produced textbooks, a four volume history of mathematics in 1710 and a mathematical dictionary in 1711. His philosophical lectures were in the first instance expositions of Leibniz's philosophy, which led his opponents (the representatives of the Thomasius ‘school’) to identify his philosophy as the Leibniz-Wolffian philosophy, a designation that remained despite his objection. He began an extensive series of publications, in German and often identified as his German works, on philosophical topics -- the “German

Logic” in 1713, the “German Metaphysics” in 1719, the “German Ethics” in 1720, the “German Politics” in 1721, the “German Cosmology” in 1723, and the “German Theology” in 1724, along with numerous short essays. Given his success as a teacher and fame as an author, he gained increasing prominence in Halle, much to the dismay of the “Thomasians” who had dominated philosophical instruction there. It did not contribute to a smooth collegial relationship that Wolff did not like Thomasius's eclectic philosophy and did not hide his dislike. Matters came to a head in 1721 when some political maneuvering on the part of his opponents (the “Thomasians” and the pietists) likely prompted by his own increasing fame and popularity with the students combined with his apparently difficult personality, brought him to the attention of the emperor in Berlin, Friedrich Wilhelm I, who expelled him from Prussia in 1723 on threat of hanging.

Having earlier been invited to the University of Marburg, he took up the position there. At the time Marburg was a more cosmopolitan place than Halle, and Wolff now had students from other countries. He saw himself as speaking to Europeans, not merely Germans, and began writing in Latin. In fact, he produced a second series of books, in Latin and identified as his Latin works, going over the same subject matter that he had treated in his German texts, albeit in more detail. Though even more scholastic than the earlier German texts, these books contributed to his fame in a much broader context (Europe rather than Germany). In 1733, Friedrich Wilhelm I invited him to return to Prussia, but Wolff declined this invitation. He became increasingly well established, so much so that a cabinet order of 1739 required candidates for the ministry to study Wolff's books, particularly his logic. In short order, Wolffians and Wolff societies could be found everywhere, even in Prussia. In 1740, Frederick the Great recalled him to Prussia, offering a permanent fellowship in the Berlin Academy. Rather than accept the invitation to Berlin, he returned to Halle to great acclaim and public approbation. To his disappointment, however, the mood in Halle had changed, the residing Wolffians, in particular Baumgarten, had begun to develop his thought, and his lectures were not successful. Complaining about the poor quality of the students, he gave up lecturing but continued writing.

Philosophy

Wolff was not an original philosopher, but a modernizer and systematiser. Rather than reject scholastic school philosophy outright, as Thomasius had done, he modernized and systematised it (and philosophy as a whole). Systematizing philosophy meant integrating different ideas from the philosophical tradition -- Descartes's concept of substance, for instance, and Leibniz's theory of pre-established harmony. But while eclectic in this sense, unlike Thomasius's thought, Wolff's was anything but arbitrary. Rather, he combined those ingredients into a comprehensive system on the model of mathematics. Mathematics was, for Wolff, a systematic science operating by definition and syllogistic proof, and this was the method he strove to make applicable to philosophy. In philosophy, the method was cashed out in a combined analytic/synthetic manner. Definitions were arrived at analytically -- the analysis was to be of empirical matters and was to convey simple ideas through a process of clarification, abstraction, and analysis. These were then combined into definitions. The definitions were to function as ingredients in the syllogisms that returned, synthetically, to the empirical starting point, though it is presumably now understood why things are as they are, an understanding that was, for Wolff, the goal of philosophy.

The first philosophy text Wolff produced was his “German Logic” (*Vernünfftige Gedancken von den Kräften des menschlichen Verstandes und ihrem richtigen Gebrauche in Erkenntnis der Wahrheit*). Logic is of central importance to Wolff because it sets out the rules for thought, which is understanding's ability to forge connections, according to Wolff a uniquely human ability. All human beings have natural understanding, but by itself this is not sufficient. Logic or the “art of demonstration” serves to refine this natural capacity and functions, as well, as the condition of science. From the point of view of the enlightenment, it is instructive here that Wolff insisted, as Thomasius had before him, that book and memory learning is not the same as knowledge. That requires the use of the powers of the understanding, and above all, much practice in the art of thought. While innate, the powers of understanding have to be honed through practice/experience.

Wolff's second philosophical treatise, the “German Metaphysics” (*Vernünfftige Gedancken von Gott, der Welt und der Seele des Menschen, auch allen Dingen überhaupt*) appeared some seven years after the “German Logic”. Just as that had vindicated the discipline of logic against the early enlightenment attacks, so this vindicated the discipline of metaphysics, understood as the “science of the possible as possible.” After a brief introductory chapter identifying his Cartesian stance by linking existence to consciousness (we must exist because we are conscious of ourselves), the “German Metaphysics” treats ontology, empirical psychology, cosmology, rational psychology and natural theology. In the second chapter Wolff sets out the two (Leibnizian) principles governing his philosophical thought: the principle of contradiction (“something cannot both be and not be at the same time” §10, 6) and the principle of sufficient reason (“everything that is must have a sufficient reason why it is” §30, 17). These are significant not only from the point of view of Wolff's thought but also in light of the role they would play in early Kant-criticism. In the chapter on rational psychology (chapter 5), he defends the Leibnizian conception of pre-established harmony (§765, 478-9) and, in the final chapter on natural theology, he tells his readers that the world mirrors God's perfection (§1045, 648). These are aspects of Wolff's position with which the (pietist) Thomasians would take issue.

After the “German Metaphysics” appeared, Wolff published about a book a year dealing with and integrating into his system other central philosophical matters: ethics in 1720, politics in 1721, physics in 1723, teleology in 1724 and biology in 1724.

The ethics (*Vernünfftige Gedancken von der Menschen Thun und Lassen, zu Berförderung ihrer Glückseligkeit*) is composed of four parts, a theoretical part that treats the foundation of practical philosophy and three practical parts that present a doctrine of duties that human beings have to themselves, to God, and to others. Not surprisingly, given that Wolff believes the world to mirror God's perfection, the issue in the ethics is perfection as well and not, as with Thomasius, happiness (that is left for his politics). Moral perfection is the guideline by which we ought to choose between two (or more) equally possible actions. That is to say, when making a free choice we ought to consider whether the action “promotes the perfection of our inner and outer state” (§2, 5) and that means considering whether the state of the soul and the body accords with the prior state or contradicts it. The outcome has greater perfection to the extent that it contributes to the continued “natural human state and its harmonious preservation over time”(§2, 5). The natural human state Wolff envisions is the state of the soul in its manifold efforts to find truth, and everything has to be done to maximize that state (see “German

Metaphysics” §152, 79). It so happens, that this is where happiness lies as well, and as Wolff indicates at the end of the ethics, it is incumbent upon human beings to ensure not only their own perfection/happiness, but to “contribute as much as possible to the happiness of others” (§767, 539).

In the “German Politics” (*Vernünfftige Gedancken von dem gesellschaftlichen Leben der Menschen und insonderheit dem gemeinen Wesen*), he proceeds to investigate the varieties of human societies and to specify how they ought to be set up so as to “promote the uninhibited progression to the common best” (§3, 3). A society must accord with the laws of nature, otherwise it cannot be considered a society, and accord with the laws of nature surely means perfection/happiness.

Wolff's Latin works appeared with equal regularity in the 1730, with the “Latin Logic” (*Philosophia rationalis sive logica*) beginning the series in 1728. In spite of their greater thoroughness, or perhaps because of it, they were not as widely read either in Germany or Europe as a whole. Whereas the German texts went through several reprints (14 for the Logic, 10 for the Metaphysics), the “Latin Logic” was only reprinted three times, and some of the other Latin texts perhaps twice. By the 1730s the German texts had established Wolff's influence and the Latin texts did little to change that. They will, accordingly, not be further considered in this context.

Context, Influences and Disciples

Both Thomasius and Wolff had followers. In Thomasius's case this was in spite of the fact that he claimed to have no desire to have disciples or found a school. But given the influence he had on his contemporaries, it would have been surprising if there had not been anyone who saw himself as following in his footsteps and, more importantly, taking up his cause (against Wolff). Among Thomasius's contemporaries to do so were Franz Budde (1667-1729), Joachim Lange (1670-1744), Andreas Rüdiger (1673-1731), and Adolf Friedrich Hoffmann (1707-1741). Christian August Crusius (?1715-1775) was a later follower (Hoffmann's student) who came on the philosophical scene when Wolffianism was already starting to decline.

Wolff's followers were perhaps more varied than those following and rejecting Thomasius. Indeed, by the 1740s, Wolffianism had become the leading German philosophy and had spread to the major German universities: to Halle (with Wolff, A. G. Baumgarten, though he moved on to Frankfurt an der Oder), G. F. Meier, and J. A. Eberhard, who is known particularly for the role he played in early Kant-criticism), to Marburg where virtually every academic was a Wolffian, to Giessen (with J. F. Müller, and Böhm), to Tübingen (with Georg Bernhard Bilfinger, Israel Gottlob Ganz and Gottfried Plouquet), to Leipzig (with Gottsched and Ludovici), to Jena (with Johann Peter Reusch and Joachim George Darjes), and to Königsberg (with Knutsen, Kant's teacher).

The reception of Wolff's philosophy is interesting not only in view of its widespread acceptance, but also in light of the disputes it generated and the further developments it gave rise to. Chief among these disputes were (1) the attack by the pietists (Budde and Lange) that led to Wolff's dismissal from Halle and (2) the attack by the Thomasians (Hoffmann and Crusius). Further developments of the Wolffian

philosophy can be found particularly in the domain of aesthetics with Gottsched and Baumgarten.

Pietism

The development of early and even mid-enlightenment thought in 18th century Germany proceeded hand in hand with the then relatively new (protestant) religious trend: pietism. Like the discontent that the representatives of the early enlightenment had with authority and, at least in the early years, intellectual life, those of pietism took issue with the (religious) orthodoxy and its intellectualism. Furthermore, rather than endorse obedience and conformity to the establishment, the pietist movement emphasized the subjective aspect of faith: a person's experience, feeling, and, above all, personal participation in religious matters and performances. The emphasis on the subjective and personal, and on people's actual participation in religion made pietism an ideal companion for early enlightenment thought. At issue here was not academic competence, not the cognitive aspect of religion, but its affective aspect with emphasis on devotion and practical service, just as what was at issue in the early enlightenment was not the intellectual aspect of reason, but its practical performance and service. And both movements were characterized by a commitment to egalitarianism.

This is not to say that the parallel development was always or even necessarily harmonious. In Halle, the chief representative of pietism was August Hermann Francke (1663-1722), who had been brought there by Thomasius. But Thomasius and Francke did not see eye to eye on all matters. While Thomasius had endorsed Francke's practical activism (he was the founder of the Halle orphanage), he broke with Francke by 1699, criticising his educational policies for producing “uneducated, melancholy, fantastic, obstinate, recalcitrant, and spiteful men” (cited in Beck, 253). It is not clear what, if any, philosophical reasons Thomasius had for turning against Francke. However, the details of their disagreement are not important for the development of 18th century philosophy.

the Thomasians

Budde and Lange, both ultimately Professors of Theology at Jena and Halle respectively, developed Thomasius's thought in a theological (pietist) direction. That is to say, both, and Lange more so than Budde, emphasized the need for revelation to a greater extent that Thomasius had done, as they also claimed that the source of evil was the will and that the root cause of evil was original sin. Though Thomasius could be found to hold these views as well, particularly during the completion of his *Application of Moral Theory*, their difference from Thomasius lies in the emphasis they placed on these views. Moreover, in his major work, *Medicina Mentis,* Lange devoted a great deal of attention to demonstrating how the ill or corrupt will and mind might be healed.

It is interesting that there were differences not just from the theological dimension of Thomasius's thought, but also in the philosophical respect. Here it is particularly noteworthy that in his philosophical texts Budde sought a systematic whole and was not content to adopt Thomasius's more eclectic style. Though he claimed to be an eclectic, he thought this did not entail that he could not provide a unified whole. That whole is provided by his three books in philosophy. The theoretical philosophy can be found

in the *Institutiones Philosophiae Eclecticae* that appeared in two parts in Halle in 1703. The two parts are the “Elementa philosophiae theoreticae,” roughly a metaphysics, and the “Elementa philosophiae instrumentalis,” Budde's logic. Along with the earlier work in practical philosophy, the *Elementae Philosophiae Practicae*, originally published in 1697 but significantly revised for the 1703 edition, these texts constitute Budde's philosophical system.

Rüdiger, Hoffmann, and Crusius were related through teacher-student relationships with Rüdiger teaching Hoffmann who then taught Crusius. See below (Disputes) for an account of the role they, and Crusius in particular, played in the disagreement of the Thomasians with Wolff.

the Wolffians

Wolff's enlightenment rationalism had made a decisive impact on the philosophical scene in Germany in the 1720s. Over the middle years of the 18thcentury, and against objections from both Thomasians and pietism, Wolffianism took hold, at least for a time. Certainly, the events that led to his expulsion from Halle contributed to his notoriety and fame. Philosophers took up his cause in virtually all the universities in Eastern and Northern Germany, occasionally developing his philosophical system in different directions. Among the chief representatives of Wolffianism were, in Leipzig, J. C. Gottsched (1700-1766), in Frankfurt am Main, Alexander Gottlieb Baumgarten (1714-62) and H. F. Meier, who were instrumental in founding Aesthetics, in Königsberg, Kant's teacher Martin Knutsen (1713-51). Wolffianism also played a central role in the early criticisms of Kant's critical philosophy. There it was represented primarily by Eberhard, Maaß, and Schwab and in the major review journal at the time, Nicolai's *Allgemeine deutsche Bibliothek*. (See Allison, *The Kant-Eberhard Controversy*). There can be no question that Wolffianism dominated German universities during the 1720s to the 1740s, though it disintegrated to some extent soon after his death in 1754 with, among other things, the development of aesthetics.

Disputes

Given the two radically different approaches to issuing the enlightenment call to the independent use of reason that we find in the 18th century, it is not surprising that the age should have its share of disputes. This was particularly evident in the disagreements surrounding Wolff and Wolffianism. It has already been indicated above that two attacks in particular stand out (1) the attack that led to his expulsion from Halle and (2) the attack that was mounted by Hoffmann and Crusius.

The first attack was mounted by Thomasius's pietist followers. They, in particular Budde and Lange, centered their opposition to Wolffianism around the fatalism and Spinozism they thought implied by his system. Also, Wolff's dependence on the mathematical method and the subordination of the divine will to necessity were questioned. Similarly, the Leibnizian doctrine of pre-established harmony of soul and body that the critics attributed to Wolff was thought to be at odds with the possibility of free will, and with that, responsibility for our actions. If the interaction of soul and body has to be pre-ordained or pre-established, then it is not possible that we could act otherwise than we have been determined to act, and

the notions of responsibility and sin are only a chimera. It must be added here that the critique of the belief in a free will is justified. Wolff struggled with this possibility as much as Leibniz had done, and it is simply not clear that given their adherence to pre-established harmony, either one managed to resolve the conundrum on the ontological level.

It is an interesting footnote to these disputes that while Thomasius himself did not participate in these disputes, Wolff did. In 1724 he published a detailed reply to Budde and Lange's critique of the German Metaphysics (*Kleine Kontroversschriften mit Joachim Lange und Johann Franz Budde*) and he also took their concerns up in the *Anmerkungen* to the German Metaphysics.

The second attack was mounted by those Thomasians who did not foreground their pietism, Andreas Rüdiger (1673-1731), Adolf Friedrich Hoffmann (1707-1741), and Christian August Crusius (?1715-1775). Rüdiger and Hoffmann attacked Wolff on a number of issues (the relation of mind and body, for instance, which they took to constitute a unity, not two distinct substances, the difference between the mathematical and the philosophical methods, and the question of the will's role in ethics). At least some of these issues were taken up by Crusius, who was the last Thomasian to take issue with Wolff.

Crusius did his philosophical work in Leipzig in the 1740s before taking on the professorship of Theology in 1751. Like earlier Thomasians, he questioned Wolff's subordination of the will to reason, arguing for the freedom of the former against what he takes to be Wolff's fatalism, though he was willing to concede that the will requires understanding. Perhaps the most significant aspect of Crusius' thought was his differentiation between the philosophical and mathematical methods. Whereas mathematics is said to be grounded on the principle of contradiction, this is not the case for philosophy. That has to do with real objects, with nature, its structures and forces, not with the abstract and artificial concepts of mathematics. At the same time, he also thought there was a role for mathematics in the analysis of nature -- it could provide concepts of bodies and their relations. Of course, as his contemporaries pointed out, one wonders why mathematics should have a role in the investigation of the “real” world if the nature of its abstraction made it distinct from the (concrete) real world. Crusius did not have an answer to these questions, but it was a topic that Kant would take on in his pre-critical project.

Beyond Wolff

With the decline of Wolffianism after 1754, German philosophy was for a time at loose ends. The decline was prompted not only by the criticisms offered by the Thomasians, but also by the internal developments of Wolffianism proposed by various of Wolff's disciples. Hermann Samuel Reimarus (1694-1768), for instance, developed a rationalist critique of revelation, arguing that for the rational person religion had to be based on reason (*Apologie oder Schützschrift für die vernünftigen Verehrer Gottes*, published posthumously). Others began to extend Wolffianism into areas Wolff himself had not considered, thereby contributing to significant development. Here the area of aesthetics is paramount. Moreover, by the middle of the 18th century, French and English (particularly) Scottish philosophy (Hume's *Inquiry*, Locke's *Essay* along with other texts by Hume, Reid, Hutcheson, Beattie, Condillac) began to be available in translation, thus offering a clear (empiricist) alternative to Wolffian rationalism. These texts were

immensely influential. Recall Kant's claim that his recollection of David Hume had awoken him from his dogmatic slumber (*Prolegomena*, *Ak*. IV, 260). The empiricists, for instance Feder and Garve also played a significant role in early Kant-criticism (see Sassen, *Kant's Early Critics*). By and large, the period of 18th century German philosophy after 1750 and before 1781 was a period not of any one dominant school, but of individuals loosely associated with a number of different trends. Among individuals who cannot be clearly assigned anywhere were Johann Heinrich Lambert, Moses Mendelssohn and Johann Nikolaus Tetens. Not either Thomasians or Wolffians, they must be considered in relation to Kant's pre-critical and critical philosophy, not from the point of view of German philosophy prior to Kant.

Aesthetics

Even though Wolff sought to integrate virtually all aspects of philosophy into his system, he had nothing to say about the philosophy of art. In some way this is not surprising if we consider that in the Germany of the early Enlightenment, the arts, particularly poetry and literature, were as good as nonexistent. The German literary greats had not even been born when Wolff wrote his German texts (Lessing was born in 1729, Herder in 1744, Goethe in 1749). First to develop a theory of the arts, particularly poetry, was Johann Christoph Gottsched (1700-1766), who published his ‘Critical Poetry’ (*Versuch einer critischen Dichtkunst vor die Deutschen*) in 1730. Treating poetry scientifically, he set out a set of rules that were to guide the composition. Given his conception of what a poem was (a moral fable) and given as well his idea of what was involved in its composition (a set of rules), there was little room here for beauty and even less for sentiment and inspiration. But sentiment and inspiration was precisely the direction in which poetry was going by the middle of the century. Gottsched's Wolffian philosophy of poetry, accordingly, was quickly supplanted by Johann Jakob Breitinger's *Critische Abhandlung* (1740), a book that emphasized the *a posteriori* experience of poetry, not its rule-bound composition.

A quasi-Wolffian synthesis of these two approaches was brought about by Alexander Gottlieb Baumgarten (1714-1762), probably Wolff's most famous disciple. Baumgarten produced two major texts, his metaphysics in 1739 (*Metaphysica*), which Kant would use as a textbook, and, in 1750 and 1758, his aesthetics *(Aesthetica*). That he designated the term ‘aesthetics’ to identify the philosophy of art, particularly poetry, would hardly be thought important had he done nothing else. However, in developing the ‘field’ of aesthetics, he also contributed significantly to the study of the senses. Whereas Wolff had conceived of the senses merely as providing the raw material for processing, a task performed by understanding and governed by the rules of logic, Baumgarten thought that the senses had their own rules and their own perfection, rules and perfection that differ from logical rules and the knowledge generated by logical processing. The rules of sensation are studied by the science of perception, which Baumgarten called aesthetics. But aesthetics is also the study of poetry. In *Reflections on Poetry* he set out the notion of aesthetic clarity. Although poetic representations are confused representations (§14, 42), they have sensuous clarity. Baumgarten identified this clarity as *extensive* clarity to differentiate it from the *intensive* clarity of logic (§17, 43). Appealing to the senses, poetic representation can be more illuminating than representations produced by logical processing and reasoning. Moreover, appealing to the affects, such representations are connected to pleasure. One cannot but see the influence that Baumgarten likely had on Kant's critical philosophy -- his vindication of the senses reappears in an inherently Kantian way both in the Transcendental Aesthetic of the *Critique of Pure Reason* and in the

Critique of Judgment.

Bibliography

Primary Texts

- Baumgarten, Alexander Gottlieb. *Meditationes philosophicae de nonnullis ad poema pertininentibus*, tr. Karl Aschenbrenner and William B. Holther, *Reflections on Poetry*. Berkeley: University of California Press, 1954.
- -----. *Metaphysica*. Halle, 1739.
- -----. *Aesthetica*. Halle, 1750, 1758.
- Breitinger, Johann Jakob. *Critische Abhandlung von der Natur, den Absichten und dem Gebrauche der Gleichnisse*, Zürich, 1740.
- Budde, Franz. *Institutiones Philosophiae Eclecticae*, 2 parts. Halle, 1703.
- -----. *Elementae Philosophiae Practicae*. Halle, 1697, 1703.
- Crusius, Christian August. *Die philosophischen Hauptwerke*, ed. Giorgio Tonelli. Hildesheim: Georg Olms, 1964ff.
- -----. *Anweisung vernünftig zu leben*. Leizig, 1744. *Werke* vol1.
- -----. *Entwurf der nothwendigen Vernunftwahrheiten*. Leipzig, 1745. *Werke* vol.2.
- -----. *Weg zur Gewißheit und Zuverlässigkeit*. Leipzig, 1747. *Werke* vol. 3
- -----. *Kleinere philosophische Schriften*, ed. Sonia Carboncini and Reinhold Finster. *Werke* vol 4.
- Gottsched, Johann Christoph. *Erste Gründe der gesamten Weltweisheit, darinn alle philosophische Wissenschaften in ihrer natürlichen Verknüpfung abgehandelt werden*. 2 volumes. Leipzig, 1733-34.
- -----. *Versuch einer critischen Dichtkunst vor die Deutschen*. Leipzig, 1730
- Kant, Immanuel. *Prolegomena zu einer jeden künftigen Metaphysik, die als Wissenschaft wird auftreten können*, ed.. Benno Erdmann. *Gesammelte Schriften* Berlin: de Gruyter and predecessors, 1900ff, vol IV
- Lange, Joachim. *Medicina Mentis*. London, 1715.
- -----. *Caussa Dei et religionis naturalis adversis atheismum*. Halle, 1727.
- Ludovici, Carl Günther. *Entwurf einer vollständigen Historie der Wolffschen* Philosophie. 3 volumes. Leipzig, 1737-38.
- Rüdiger, Andreas. *Physica divina, recta via, eademque inter supersttionem et atheismum media ad ultramque hominis felicitatem, naturalem, atque moralem ducens*. Frankfurt, 1716.
- -----. *De sensu vedri et falsi*. Leipzig, 1722.
- Reimarus, Hermann Samuel. *Apologie oder Schützschrift für die venünftigen Verehrer Gottes*. Published posthumously.
- Thomasius, Christian. *Ausgewählte Werke*, ed. W. Schneiders. Hildesheim: Georg Olms, 1993 ff.
- -----. *Institutiones jurisprudentiae divinae*. Leipzig, 1688.
- -----. *Instruductio ad philosophiam auliam*. Leipzig, 1688 (*Werke* vol.1)
- -----. *Einleitung zur Vernunftlehre*. Halle, 1691 (*Werke* vol.8)
- -----. *Ausübung der Vernunftlehre*. Halle, 1691 (*Werke* vol.9)

- -----. *Einleitung zur Sittenlehre*. Halle, 1692 (*Werke* vol. 10)
- -----. *Ausübung der Sittenlehre*. Halle, 1696 (*Werke* vol. 11)
- -----. *Kleine deutsche Schriften*. Halle, 1701 (*Werke* vol. 22)
- Wolff, Christian. *Gesammelte Werke*, eds. J.École, H. W. Arndt, Ch. A. Corr, J. E. Hoffmann, M. Thomann. Hildesheim: Georg Olms, 1965ff.
- -----. *Vernünfftige Gedancken von den Kräften des menschlichen Verstandes und ihrem richtigen Gebrauche in Erkenntnis der Wahrheit*, (*Deutsche Logik*), ed. Hans Werner Arndt. (*Werke* vol.1)
- -----. *Vernünfftige Gedancken von Gott, der Welt und der Seele des Menschen, auch allen Dingen überhaupt* (*Deutsche Metaphysik*), ed. Charles Corr. *(Werke* vol. 2)
- -----. *Der vernünfftigen Gedancken von Gott, der Welt und der Seele des Menschen, auch allen Dingen üuberhaupt, anderer Teil, bestehend in ausführlichen Anmerckungen* (*Anmerkungen zur Deutschen Metaphysik*), ed. Charles Corr (*Werke* vol. 3)
- -----. *Vernünfftige Gedancken von der Menschen Thun und Lassen, zu Berförderung ihrer Glückseligkeit* (*Deutsche Ethik*), ed. Hans Werner Arndt. (*Werke* vol. 4)
- -----. *Vernünfftige Gedancken von dem Gesellschaftlichen Leben der Menschen und insonderheit dem gemeinen Wesen* (*Deutsche Politik*), ed. Hans Werner Arndt. (*Werke* vol. 5)
- -----. *Kleine Kontroversschriften mit Joachim Lange und Johann Franz Budde*, ed. Jean École. (*Werke* vol. 17

Secondary Texts

- Allison, Henry. *The Kant-Eberhard Controversy*. Baltimore: The Johns Hopkins University Press, 1973.
- Barnard, F. M. “The ‘Practical Philosophy’ of Christian Thomasius,” *Journal of the History of Ideas* 31/2 (1971): 221-246.
- Beck, Lewis White. *Early German Philosophy: Kant and His Predecessors*. Thoemmes Press, 1996.
- Cassirer, Ernst. *The Philosophy of the Enlightenment*. Boston: Beacon Press, 1951.
- Hinrichs, Carl. *Preußentum und Pietismus: Der Pietismus in Brandenburg-Preußen als religiös-soziale Reformbewegung*. Göttingen: Vanderheck & Ruprecht, 1971.
- Hinske, Norbert ed. *Halle: Aufklärung und Pietismus*. Heidelberg: Verlag Lambert Schneiders, 1989
- Holzhey, Helmut. “Philosophie als Eklektik,” *Studia Leibnitiana*, 35/1 (1983):19-29.
- Hunter, Ian. *Rival Enlightenments: Civil and Metaphysical Philosophy in Early Modern Germany*. Cambridge: Cambridge University Press, 2001.
- Kuehn, Manfred. *Scottish Common Sense in Germany, 1768-1800*: A Contribution to the History of Critical Philosophy. .Kingston and Montreal: McGill-Queen's University Press, 1987.
- Raabe, Paul and Wilhelm Schmidt-Biggemann eds. *Aufklärung in Deutschland*. Bonn: Hohwacht Verlag, 1979.
- Sassen, Brigitte. *Kant's Early Critics: The Empiricist Critique of the Theoretical Philosophy*. New York: Cambridge University Press, 2000.
- Schneiders, Werner. *Christian Thomasius 1655-1728: Interpretation zu Werk und Wirkung*. Hamburg: Felix Meiner Verlag, 1989. (Contains bibliography)

- -----. *Christian Wolff 1679-1754: Interpretation zu seiner Philosophie und deren Wirkung*. Hambutg: Felix Meiner Verlag, 1983. (Contains bibliography).
- Schröder, Peter. “Thomas Hobbes, Christian Thomasius and the Seventeenth Century Debate on the Church and the State,” *History of European Ideas*, 223/2-4 (1998): 59-79.
- Schönfeld, Martin. *The Philosophy of the Young Kant: The Precritical Project*. Oxford: Oxford University Press, 2000.
- Vollhardt, Friedrich ed. *Christian Thomasius (1655-1728): Neue Forschungen im Kontext der Frühaufklärung*. Tübingen: Max Niemeyer Verlag, 1997. (Contains bibliography).
- Wilson, Holly. “Kant's Experiential Enlightenment and Court Philosophy in the 18th Century,” *History of Philosophy Quarterly*, 18/2 (April 2001): 179-205.
- Wundt, Max. *Die deutsche Schulphilosophie im Zeitalter der Aufklärung*. Hildesheim, 1964.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Hobbes, Thomas | Kant, Immanuel | legal philosophy | Leibniz, Gottfried Wilhelm | Locke, John

Brigitte Sassen
sassenb@mcmaster.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 10, 2002 Content last modified: March 10, 2002

Justice as a Virtue

When we speak of justice as a virtue, we are usually referring to a trait of individuals, even if we conceive the justice of individuals as having some (grounding) reference to social justice. But Rawls and others regard justice as "the first virtue of social institutions" (1971, p. 3), so "justice as a virtue" is actually ambiguous as between individual and social applications. This essay will reflect and explore that ambiguity, though the principal focus will understandably be on the justice of individuals.

However, even the idea of individual justice seems ambiguous in regard to scope. Plato in the Republic treats justice as an overarching virtue of individuals (and of societies), meaning that almost every issue he (or we) would regard as ethical comes in under the notion of justice (dikaosoune). But in modern usages justice covers only part of individual morality, and we don't readily think of someone as unjust if they lie or neglect their children--other epithets more readily spring to mind. What individual justice most naturally refers to are moral issues having to do with goods or property. It is, we say, unjust for someone to steal from people or not to give them what he owes them, and it is also unjust if someone called upon to distribute something good (or bad or both) among members of a group uses an arbitrary or unjustified basis for making the distribution (this last aspect of individual justice obviously has reference to social or at least group justice). Discussion of justice as an individual virtue standardly (at least) centers on questions, therefore, about property and other distributable goods.

1. History

Although the idea of social justice based in a social contract is mentioned in Plato's Republic and was

known even earlier, the Republic's conception of individual justice is distinctively virtue ethical. To be sure, Plato understands individual justice on analogy with justice "writ large" in the state, but he views the state, or republic, as a kind of organism or beehive, and the justice of individuals is not thought of as primarily involving conformity to just institutions and laws. Rather, the just individual is someone whose soul is guided by a vision of the Good, someone in whom reason governs passion and ambition through such a vision. When, but only when, this is the case, is the soul harmonious, strong, beautiful, and healthy, and individual justice precisely consists in such a state of the soul. Actions are then just if the sustain or are consonant with such harmony.

Such a conception of individual justice is virtue ethical because it ties justice (acting justly) to an internal state of the person rather than to (adherence to) social norms or to good consequences; but Plato's view is also quite radical because it at least initially leaves it an open question whether the just individual refrains from such socially proscribed actions as lying, killing, and stealing. Plato eventually seeks to show that someone with a healthy, harmonious soul wouldn't lie, kill, or steal, but most commentators consider his argument to that effect to be highly deficient.

Aristotle is generally regarded as a virtue ethicist par excellence, but his account of justice as a virtue is less purely virtue ethical than Plato's because it anchors individual justice in situational factors that are largely external to the just individual. Situations and communities are just, according to Aristotle, when individuals receive benefits according to their merits, or virtue: those most virtuous deserve more of whatever goods society is in a position to distribute (exemptions from various burdens or evils counting as goods). This desert-based conception of social justice then treats the virtue of individual justice as a matter of being disposed to properly respect and promote just social arrangements. An individual who seeks more than her fair share of various goods has the vice of greediness (pleonexia), and a just individual is one who has rational insight into her own deserts in various situations and who habitually (and without having to make heroic efforts to control contrary impulses) takes no more than what she deserves, no more than her fair share of good things. Since Aristotle treats all individual virtues as (learned dispositions) lying in a mean between extremes (courage, e. g., is between cowardice and foolhardiness), he also doesn't think it is virtuous to take less than one's fair share of things (though the issue is somewhat complicated for him).

However (and as William Frankena once noted), this account of justice seems circular or ungrounded, because if one's just deserts depend on how virtuous one is, the issue of what one's fair share is cannot be decided independently of whether one is being virtuous in actually taking some particular share and the latter issue, in turn, depends for Aristotle on one's being able to know independently what one's fair share is. We have reason to doubt, therefore, whether Aristotle has really given us a determinate conception of justice either as an individual or as a situational virtue.

Both Plato and Aristotle were rationalists as regards both human knowledge and moral reasons, and what they say about the virtue of justice clearly reflects the commitment to rationalism. Much subsequent thinking about justice (especially in the Middle Ages) was influenced by Plato and Aristotle and likewise emphasized the role of reason both in perceiving what is just and in allowing us to act justly rather than give in to contrary impulses or desires. But to the extent Christian writers allied themselves with Plato

and Aristotle, they were downplaying another central element in Christian thought and morality, the emphasis on agapic love. Such love seems to be a matter of motivationally active feeling rather than of being rational, and some writers on morality (eventually) allowed this side of Christianity to have a major influence on what they had to say about virtue.

In particular, the so-called moral sentimentalists Hutcheson and Hume (not to mention Shaftesbury and Adam Smith) treated morality as grounded in something other than reason, and the influence of Christian ideas and ideals of agapic love on Hutcheson (at least) is well documented. For Hutcheson, universal (i.e., impartial) benevolence is the highest and best of human motives, but we know this, not through reason, but through a moral sense (or sensitivity). Also, according to Hutcheson, the individual virtue of justice (ultimately) consists in being motivated by universal benevolence, and he explicitly denies that benevolence can ever conflict with true justice.

Hume saw (or believed he saw), however, that individual justice at least sometimes conflicts with what benevolence would motivate us to do. He is as much as sentimentalist as Hutcheson, believing that judgments about virtue and rightness depend on our capacity for sympathy rather than on some form of reason (or on a distinct moral sense) and holding that being virtuous depends on feelings and feelingful motives like benevolence and sympathy rather than on reason. But he thinks that the sentimentalist owes us an account of how a sense of justice that is sometimes opposed to benevolence and sympathy can nonetheless develop out of such motives. Motives like benevolence, curiosity, and prudence Hume calls natural in the twofold sense that they exist apart from social convention(s) and that they do not require explicitly ethical thinking (or conscience) or in order to issue in action. But the virtue of justice is not natural, but rather should be considered "artificial," according to Hume, because it depends for its existence on human conventions and artifices and because the primary motive to justice is a sense of justice (or of duty).

Now Hume thinks of the individual virtue of justice quite narrowly as comprising a certain kind of respect for (other people’s) property. The just person doesn’t steal from others and returns what he has borrowed (and Hume points out the similarity of this usage to the Aristotelian notion that justice consists in everyone getting his due, what he deserves). But there are other artificial virtues, according to Hume, among them, fidelity to promises (keeping one’s promises), law-abidingness, and (female) modesty/chastity. Hume thinks that it is difficult to account for any one, or all, of these artificial virtues in (his own) empiricist anti-rationalist terms, and there are at least two reasons for this.

One has to do with the inadequacy of natural motives like benevolence or prudence for grounding the requirements of justice. In primitive or simple societies, there may always be reasons of prudence to act justly with respect to the property of others: violations of justice are always likely to be detected by others and to lead to consequences one would prefer to avoid. In such circumstances honesty (a term Hume tends to use narrowly as synonymous with "justice") really is the best policy. Furthermore, within the narrow confines of a small group, personal affection and benevolent concern for those one knows and lives with may lead one to refrain from violations of their property. But in a larger (and more advanced) society, things will be different.

In large-scale modern societies (of the sort Hume lived in) we may not know our neighbors, much less all those people our actions might affect, and the people whose property justice calls on us to respect include a vast majority who don't know us either. Under conditions of such relative anonymity and complexity, the identity of a thief may be much more difficult to ascertain, and if one (knows one) can get away with stealing on some occasions, then prudence is presumably incapable of motivating a just refusal or unwillingness to steal. More importantly, perhaps, the conditions of a modern society leave us without strong ties of affection to many of the people we interact with or may affect by our actions, and Hume thinks that normal humanity or humane benevolence isn't a strong enough motive to get us to refrain from a theft that would greatly benefit ourselves or our families (those we do have strong affections toward). So if we refrain from such theft, we cannot explain or justify such refraining by reference to any actual natural motives.

But of course we do (many of us) justly refrain from taking or violating other people's property on occasions when natural motives as such may seem incapable of explaining why we do, and Hume believes (not uncommonsensically) that it is artificial motives or motivation that explain why we do so. Someone who can get away with stealing and who has stronger reasons/motives of affection to steal than not to steal may nonetheless refrain from stealing because she thinks it unjust or wrong to steal. A sense of duty or conscience is thus for Hume absolutely essential to understanding the virtue and obligation of justice/honesty.

Moreover, and this is perhaps the most important point for Hume, even if human beings were capable of the strong universal/impartial benevolence that Hutcheson regarded as the cornerstone of moral virtue, such benevolence would not in all instances suffice for us to fulfill our intuitive obligations of justice. Justice and moral obligation sometimes seem opposed to the dictates of (what would be motivated by) universal benevolence, and Hume cites one's obligation (of justice) to return what one owes to a "seditious bigot" as one glaring instance of this point. Concern for the good of society or of humanity would presumably dictate that one not return to the bigot money or property he would use to subvert and corrupt society or the state, but, according to Hume, we nonetheless think it obligatory to return what we owe, and what gets us to do so, therefore, is a(n artificial) sense of duty (or justice) rather than any (even hypothetically imaginable) natural motive.

It is possible to question this, and Bentham, for example, claimed that his disagreement with Hume's view of our actual obligations in such cases was what initially led him to utilitarianism. However, the idea that we should return the money or property despite what free-wheeling prudence or benevolence would lead us to do is intuitively forceful, and Hume shows himself aware of the potential clash here between benevolence and justice/morality in a way that Hutcheson was not. Moreover, I am dwelling at some length on Hume in part because the question of whether justice can be understood entirely in terms of natural motives is (as Hume rightly saw) absolutely crucial to understanding the nature of justice as an individual (or, for that matter, social) virtue.

Now if we agree that the right and just thing to do in cases like that of the seditious bigot is to return what one owes, then the importance of artificial motives looms large. But this leaves Hume with a

problem (or a set of problems). Hume is an empiricist and an anti-rationalist who emphasizes feeling/sentiment as the basis of morality. But if natural motives can't explain the virtue of justice, then the sentimentalist owes us an account of how a sense of duty, obligation or justice develops. (For the moment, I leave aside the other artificial virtues, which raises similar issues.) Rationalists tend to think of the sense of duty as a response of reason to certain moral facts or relations, so if Hume is to maintain his sentimentalism, he needs (among other things) to explain to us (in terms compatible with his empiricist premises) how a sense of duty develops out of (or can exist as a) feeling or feelingful motivation. And, as is well known, Hume finds this extremely difficult to do.

Hume seeks to explain moral judgment and a sense of duty or conscientiousness based in such judgment in terms of the same mechanisms of sympathy that operate within and through the natural virtues. However, we tend to be more sympathetic with those near and dear to us, and moral judgment seeks or presupposes some sort of impartiality regarding those affected by or engaging in actions (it is no more wrong for someone to kill a member of my family than for someone to kill someone I don't know). So Hume argues that we in various ways (try to) correct for personal (or temporal) bias when making moral judgments and take, in particular, the view of a sympathetic but impartial spectator in doing so. But this still can't explain why we should in all conscience and justice return what we owe to the seditious bigot--here the most extensive and impartial sympathy would seem to dictate acting for the greatest happiness rather than justly in Hume's terms. (The utilitarian typically regards acting for the greatest happiness as the essence of true justice, but the point remains that we intuitively regard justice as *conflicting* with the dictates of utilitarianism in the kind of case just mentioned.)

Hume goes through a number of (what seem to me and many others to be) contortions in an effort to explain the possibility of (anti-benevolent) conscientiousness and justice (or promise keeping, etc.). At times, he seems to think parents or educators can influence (or even psychologically force) children to disapprove of injustice (even when it promotes the general happiness). But this presupposes, I believe, that the parents and educators already themselves have acquired such a strict sense of conscientiousness, whereas it is not clear in the first place that or how Hume's theory allows this to happen. And it is also not clear how such disapproval, either in the children or in the parents/educators, can be inculcated via the mechanisms of sympathy (or immediate agreeability) that Hume says are required for moral approval and judgment.

Matters get even more problematic for Hume's theory of justice and the other artificial virtues because Hume makes it clear that he is (what we would call) a virtue ethicist. He says that the moral status of an action depends entirely on the goodness or badness of the motive that lies behind it, so that, e. g., it is only because certain helpful actions were intended to be helpful (were motivated by the natural virtue of benevolence) that we morally approve of them or judge them to be right and good. However, it is difficult to apply this virtue-ethical assumption to the artificial virtues, because the good motive operative in their instance is *the conscientious desire to do one's duty or what is right or obligatory*. According to Hume, if I return what I owe to the seditious bigot, my only just motive is the desire to do what is right and obligatory, but, in that case, the morally good motive that is supposed (according to Hume's virtue ethics) to *explain* the rightness or goodness of returning what I owe to the seditious bigot *already makes essential reference to the rightness or goodness or obligatoriness of doing so*. As Hume

himself tells us, this seems to be arguing (explaining) in a circle, and Hume makes the same point (perhaps even more forcefully) about fidelity to promises. Although some subsequent commentators think Hume has a way out of this circle, many have not thought this to be possible (I tend to agree with them), and if that is so, then Hume's attempt to justify or explain justice as an individual virtue via empiricist sentimentalist (associationist) mechanisms cannot succeed.

2. Rationalism and Justice

Such a conclusion has led many subsequent ethical thinkers to think that justice cannot be based in sentiment but requires a more intellectually constructive rational(ist) basis, and in recent times this view of the matter seems to have been held, most influentially, by John Rawls in *A Theory of Justice*. Rawls makes clear his belief in the inadequacy of benevolence or sympathetic human sentiment in formulating an adequate conception of social justice. (He says in particular that sentiment leaves unanswered or indeterminate various important issues of justice that a good theory of justice ought to be able to resolve).

Rawls's positive view of justice is concerned primarily with the justice of institutions or (what he calls) the "basic structure" of society: justice as an individual virtue is derivative from justice as a social virtue defined via certain principles of justice. The principles, famously, are derived from an "original position" in which (very roughly) rational contractors under a "veil of ignorance" decide how they wish to commit themselves to being governed in their actual lives. Rawls deliberately invokes Kantian rationalism (or anti-sentimentalism) in explaining the intellectual or theoretical motivation behind his construction, and the two principles of justice that he argues would be agreed upon under the contractual conditions he specifies represent a kind of egalitarian political liberalism. Roughly, those principles stress (equality of) basic liberties and opportunities for self-advancement over considerations of social welfare, and the distribution of goods in society (according to the the so-called difference principle) is then supposed to work to the advantage of all (especially the worst-off members of society). Rawls argues that a utilitarian principle of justice dictating simply the maximization of overall social well-being would not be accepted in his original position and is accordingly less plausible than the conception of justice embodied in his own two principles and the construction that leads to them. He also says that the idea of what people distributively deserve is derivative from social justice rather (as with Aristotle and much common-sense thinking) providing the basis for thinking about social justice.

However, it is not merely social justice that Rawls understands in (predominantly) rationalist fashion. When he explains how individuals (within a just society) develop a sense and/or the virtue of justice, he invokes the work of Piaget, who saw moral development as akin to the other sorts of development he so famously studied. Those other sorts were of course various forms of cognitive or intellectual development, and Piaget treats moral development, therefore, as principally involving increasing cognitive sophistication. More particularly, Piaget sees that sophistication as a matter of taking more and more general or universal views of moral issues, and the Kantian and rationalist idea that morality rests on and can be justified in terms of considerations of universality (if it is right for me, it is right for everyone similarly situated) or universalizability (could I will this to be a rule governing everyone's actions?) seems to underlie or to be presupposed in much that Piaget says about moral development.

Now Rawls lays more stress than Piaget does on the role our affective nature (sympathy and the desire for self-mastery) plays in the acquisition of moral virtue. But, like Piaget, he stresses the need for a *sufficiently general appreciation and rational understanding* of social relations as the grounding basis of sense of duty or of justice and he explicitly classifies his account of moral development as falling within the "rationalist tradition." Rawls also gives distinct arguments for believing that it is rational to *retain and act upon* a sense of justice.

According to Rawls, individual justice is theoretically derivative from social justice because the just individual is to be understood as someone with an effective or "regulative" desire to comply with the principles of justice. This makes the individual virtue of justice an artificial one in Hume's sense, but, in part because he doesn't assume virtue ethics, Rawls doesn't get caught up in the Humean circle we described above. Other questions about Rawls's rationalist account of the virtue of justice, however, can lead us back in the direction of a non-circular and non-Humean form of *sentimentalism* about that virtue.

3. Stages of Moral Development

Rawls is far from the only thinker to conceive of moral development in terms substantially derived from Jean Piaget's work, and at the time Rawls was writing *A Theory of Justice*, educational psychologist Lawrence Kohlberg was working out a Piaget-inspired conception of moral development that postulated six stages of normal human moral development. Kohlberg claimed that the highest stage of such development involves a concern for justice and human rights based on universal principles and he relegated sheer concern for relationships and for individual human well-being to lower stages of the process. Moreover, he saw the ordering of the different stages in Piagetian fashion as basically reflecting differences in rational understanding: those whose moral thinking involved the invoking of universal principles of justice and rights were said to show a more advanced cognitive development than those whose moral thought appeals primarily to the importance of relationships and of human well-being or suffering.

This treats utiltarianism as less cognitively advanced (more primitive) than rationalist views like Rawls's and Kant's, and utilitarians (like Hare) have naturally called into question the objectivity and intellectual fairness of Kohlberg's account. (In fact, Rawls also questions whether any purely psychological theory of moral development could ever undercut utilitarianism in the way Kohlberg sought to do.) More significantly, perhaps, the evidence for Kohlberg's stage sequence was drawn from studies of boys, and when one applies the sequence to the study of young girls, it turns out that girls on average end up at a less advanced stage of moral development than boys do.

What came next is well known. In her 1982 book *In a Different Voice: Psychological Theory and Women's Development,* Carol Gilligan responded to Kohlberg's views by questioning whether a theory of moral development based solely on a sample of males could reasonably be used to draw conclusions about the inferior moral development of women. Gilligan argued that her own studies of women's development indicated that the moral development of girls and women proceeds and ends in a different

fashion from that of boys and men, but that that proves nothing about inferiority or superiority: it is merely a fact of difference. In particular, Gilligan claimed that women tend to think morally in terms of connection to others (relationships) and in terms of caring about (responsibility for) those with whom they are connected; men, by contrast and in line with Kohlberg's studies, tend to think more in terms of general principles of justice and of individual rights against (or individual autonomy from) other people.

4. Caring and Justice

Subsequently, many have questioned the empirical validity or accuracy of the studies Gilligan relied upon, but others have pointed out that the idea of a "different voice" need not be tied to specific assumptions about differences between the sexes. The voice of justice and principle represents a different style of moral thinking (and of an overall moral life) from that of caring for and connection with others, and later writers (notably Nel Noddings, but also Gilligan herself in later work) have tried to elaborate what a morality (moral life) based in caring would be like and also to show that such a morality may be superior to that embodied in traditional thinking about justice and rights and universal(izable) moral principles.

The primary fulcrum for articulation of any ethic of caring seems to lie in an ideal that stresses connection over separateness. The Kantian emphasis on the autonomy of the moral person and the Rawlsian/contractarian assumption of separate individuals coming together to forge a social contract see us a basically separate from others, whereas an ideal of caring concern for others sees our (initial) actual historical and personal connections with others as the basis for a positive and caring response to such connection. (However, an ethic of caring doesn't favor social conservatism in the way much communitarian thought does: any social structure that shows insufficient concern for one group or another can arguably be criticized via the ideal of mutual caring.)

In addition, an ethic of justice and rights tells us to regulate our actions or lives in accordance with certain general moral principles (or explicitly moral insights), whereas the ethic of caring stresses the good of a concern for the welfare of others that is *unmediated by principles, rules, or judgments that tell us that we ought to be concerned about their welfare*. In an ethic of caring, therefore, caring is treated as a natural virtue in Hume's sense, but this further highlights the way in which such an ethic involves us in connection with, rather than separateness from, other people. If we are concerned about others on the basis of a conscientious desire to do our duty or adhere to certain moral principles, then our concern for them is mediated by moral thinking, and someone, therefore, who cares about the welfare of others without having to rely on or be guided by explicit moral principles (or thinking) is more connected with those others than someone who acts only on the basis of such mediating principles (or thought). So the ethic of caring stresses connection with others both in what it says about the normative basis of morality and in what it says about the ways in which moral goodness shows itself within a morally good life; and by the same token, traditional Kantian or contractarian views of rights and justice give a double importance to separateness or autonomy from others through the grounds they adduce for moral/political obligation and the stress they place on being guided by moral principles or judgments within the moral life.

As I indicated above, defenders of an ethic of caring have increasingly come to view caring as grounding (offering a normative basis for) morality as a whole. That means that ideas about justice and rights either have no validity or can actually be reinterpreted and given an arguably firmer justification in terms of (what we originally regarded as the opposed notion of) caring. But it is difficult to believe that morality can properly or plausibly be confined to intimate relations of caring. For better or worse, we have to learn to live together in larger social units, and we cannot be intimate or even acquainted with every human being whose actions and fate are morally significant for us. So an ethic of caring that seeks to account for individual and social morality generally needs to say something in its own voice about social and international justice and about how given individuals can realize the virtue of justice.

In answer to this more or less explicit challenge, some caring ethicists have highlighted potential analogies between the way a mother cares for her children and the kinds of care a government, state, or society can offer to its citizens or inhabitants (who presumably cannot provide everything they need and want on their own). Others have noted that the notion of caring doesn't have to be restricted to close personal relationships and that one can intuitively speak of caring, in humanitarian fashion, about people one doesn't know (except by description). This then allows there to be obligations of caring both toward near and dear and toward humanity more generally, though the issue of how to balance these concerns becomes very important at that point.

But all these ways of developing and extending an ethic of caring seem united in stressing (what Hume calls) natural motives over artificial ones. If someone who doesn't care about his family or about human beings in general, always fails to act helpfully toward others, he exhibits a lack of caring, and an ethic of caring regards acts which display such morally deficient motivation as morally criticizable and wrong. This is virtue ethical because, as with Hume, the criterion of the rightness of an action has to do with the inner state or motive that lies behind it. But by the same token individuals who demonstrate the virtue of caring act in ways that show how much they care or are concerned about others, in ways that demonstrate their emotional connectedness with others, and this means in particular that such people don't have to remind themselves of moral ideals and obligations in order to get themselves to help those they care about. They help because they care, not because conscience or some sort of (abstract) love of the Good tells them (how virtuous or dutiful it would be) to do so.

But is this sort of natural virtue really adequate to those moral/political concerns that transcend intimate personal relationships? Hume didn't think so, and it is certainly not obvious on the face of it how an ethic of caring could handle such issues. Still, Hume makes things difficult for natural virtue by conceiving individual and social justice in highly conservative terms. According to Hume we have a strict obligation of justice to allow people to keep what they have earned through their own diligence and ingenuity rather than (say) tax it away (or, presumably, force unionization on factory owners). And since (given facts about diminishing marginal utility that Hume himself assumes to be familar to us) progressive taxation and the closed union shop arguably are on the whole good for society and known to be so, the natural motive of benevolent concern for the welfare of one's society might actually lead in the direction of genuine social justice. But such justice would then have to be conceived in less austerely rigid and more humane terms than those assumed and relied on by Hume.

However, when it comes to the individual virtue of justice, Hume himself supplies some of the means toward a (virtue-ethical) account relying solely on natural virtue (and thus not subject to "Hume's circle"). He points out that it is easier to bear not having (or being given) something than to bear having something taken away from one (thus anticipating what later psychologists have said about "adaptation levels"). And this in and of itself gives someone who is benevolent or caring about the well-being of others some reason (or motive) not to steal or allow stealing from others. However, there is also the fact that stealing (as opposed to merely allowing a theft to occur) is a positive commission, and a natural virtues approach to individual justice (vis-a-vis property) would certainly be helped along, if the distinction, say, between commission and omission, or between doing and allowing, could somehow be captured in *non-artificial* or *natural* sentimentalist terms.

This seems a tall order, and in fact the suspicion or belief that only rationalism can account for such distinctions (as, e. g., in Kant's "formula of humanity") may represent one of the largest challenges a sentimentalist ethic of caring needs to face. (Hume doesn't focus on the commission/omission distinction, although, in any event, his artificial virtues approach to the remaining aspects of deontology may at this point seem a dead end.) My own belief is that sentimentalism is up to this challenge because of the way caring about and for others is grounded in naturally developing human empathy.

Normal human caring isn't impartial (in the manner of "universal benevolence"), because it is easier to empathize with those near and dear to us, i. e., those with whom we share thoughts, lives, roots, or familial (or ethnic or national) traditions. But recent psychological studies of empathy and its relation to altruism indicate that we also tend to empathize more with those whose problems are immediate for us. We respond more to a child drowning right before our eyes than to the plight of a child we don't see and whom we know (only by description) to be in danger of dying of starvation in some distant country; and, similarly, we respond more to the "clear and present" danger faced by miners we hear are trapped underground than to dangers we know will arise in some indefinite future.

But if such perceptual and temporal immediacy make such differences (respectively) to empathic concern for others, it is arguable that causal immediacy does as well. The harm I could cause is more immediate for me than harm that I might merely allow to occur. We naturally flinch from (inflicting) the former more than from (allowing) the latter, and I say "naturally" because, as with cases involving perceptual or temporal immediacy, this is not a matter of being guided by moral principles or strictures, but of responsiveness to non-moral situational differences. If we are in this way more sensitive and responsive to differences in the strength of our potential causal relation to some harm or evil, then a moral sentimentalism that restricts itself to natural virtues may possess the resources to distinguish between commission and omission, and it may be able to use that distinction (among other things) to explain why stealing, promise-breaking, and killing are worse than allowing others to do these things. This might well then allow such an approach to account for the virtue of individual justice more successfully than Hume's theory of artificial virtues can.

5. Conclusion

In any event, there are many different conceptions of the virtue of justice, and only some of them are distinctively virtue ethical. Many non-virtual ethical approaches put forward theories of virtue, and what distinguishes them from virtue ethics is that the given theory of virtue comes later in the order of explanation, rather than itself serving as the basis for understanding (all of) morality. Rawls's conception of justice as an individual virtue is a good example of a non-virtue-ethical account of a virtue, since, as we saw, it treats individual justice as a matter of accepting and complying with independently defended moral/political principles or rules. I have tried above to offer a representative sampling of the variety of possible views about justice as a virtue, but if we have learned anything from the history of philosophy, we ought to recognize that there are probably other ways of conceiving the virtue of justice we haven't yet thought of.

Bibliography

- Aristotle, *Nicomachean Ethics*.
- Frankena, William. *Ethics*, 2nd edit., Prentice-Hall, 1973.
- Gilligan, Carol. *In a Different Voice: Psychological Theory and Women's Development*, Harvard, 1982 and (with a new preface) 1993.
- Hoffman, Martin. *Empathy and Moral Development: Implications for Caring and Justice*, Cambridge, 2000.
- Hume, David. *A Treatise of Human Nature*.
- Hutcheson, Francis. *On the Original of Our Ideas of Beauty and Virtue*.
- Kohlberg, Lawrence. *Essays in Moral Development*, vol. 1, *The Philosophy of Moral Development*, Harper and Row, 1981, and vol. 2, *The Psychology of Moral Development*, Harper and Row, 1984.
- Noddings, Nel. *Caring: a Feminine Approach to Ethics and Moral Education*, Univ. of California, 1984.
- Plato, *Republic*.
- Rawls, John. *A Theory of Justice*, Harvard, 1971.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

justice: distributive | justice: intergenerational | justice: international

mslote@mail.as.miami.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 7, 2002
Content last modified: March 7, 2002

George Santayana

Philosopher, poet, literary and cultural critic, George Santayana is a principal figure in Classical American Philosophy. His naturalism and emphasis on creative imagination were harbingers of important intellectual turns on both sides of the Atlantic. He was a naturalist before naturalism grew popular; he appreciated multiple perfections before multiculturalism became an issue; he thought of philosophy as literature before it became a theme in American and European scholarly circles; and he managed to naturalize Platonism, update Aristotle, fight off idealisms, and provide a striking and sensitive account of the spiritual life without being a religious believer. His Hispanic heritage, shaded by his sense of being an outsider in America, captures many qualities of American life missed by insiders, and presents views equal to Tocqueville in quality and importance. Beyond philosophy, only Emerson may match his literary production. As a public figure, he appeared on the front cover of *Time* (3 February 1936), and his autobiography (*Persons and Places*, 1944) and only novel (*The Last Puritan*, 1936) were the best-selling books in the United States as Book-of-the-Month Club selections. The novel was nominated for a Pulitzer Prize, and Edmund Wilson ranked *Persons and Places* among the few first-rate autobiographies, comparing it favorably to Yeats's memoirs, *The Education of Henry Adams*, and Proust's *Remembrance of Things Past*. Remarkably, Santayana achieved this stature in American thought without being an American citizen. He proudly retained his Spanish citizenship throughout his life. Yet, as he readily admitted, it is as an American that his philosophical and literary corpuses are to be judged. Using contemporary classifications, Santayana is the first and foremost Hispanic-American philosopher

1. Biography

Santayana's heritage is rooted in the Spanish diplomatic society with its stress on high education and familiarity with the world community. He was born in Madrid, Spain, on 16 December 1863. His father, Agustín Santayana, was born in 1812. The father studied law and practiced for a short time before entering the colonial service for posting to the Philippines. While studying law, Agustín served an apprenticeship to a professional painter of the school of Goya and a number of his paintings remain in the private possession of the family. He translated four Senecan tragedies into Spanish, wrote an unpublished book about the island of Mindanao, had an extensive library, and made three trips around the world. In 1845, he became the governor of Batang, a small island in the Philippines. He took over the governorship from the recently deceased José Borrás y Bofarull, who was the father of Josefina Borrás, later to become Agustín's wife in 1861 and the mother of George Santayana. His mother, Josefina Sturgis (formerly Josefina Borrás y Carbonell), was born in Scotland and was the daughter of a Spanish diplomat. Previously she married George Sturgis (d. 1857), a Boston merchant, whose early death left her alone with children in Manila. There were five children from this first marriage, three of whom survived infancy. She promised her first husband to raise the children in Boston where she moved her family. During a holiday in Spain, Josefina met Agustín again, and they were married in 1861. He was fifty years of age and she was probably thirty-five. In 1863, Santayana was christened Jorge Agustín Nicolás Ruiz de Santayana y Borrás. His half sister, Susan, insisted that he be called "George," after her Boston father. Santayana, in turn, always referred to his sister in the Spanish, "Susana."

1863-1886. Santayana lived eight years in Spain, forty years in Boston, and forty years in Europe. In his autobiography, *Persons and* Places, Santayana divides his life into three phases. The background (1863-1886) encompasses his childhood in Spain through his undergraduate years at Harvard. The second period (1886-1912) is that of the Harvard graduate student and professor with a trans-Atlantic penchant for traveling to Europe. The third period (1912-1952) is the retired professor writing and traveling in Europe and eventually establishing Rome as his home.

The family moved from Madrid to Ávila where Santayana spent his boyhood. In 1869, Santayana's mother left Spain in order to raise the Sturgis children in Boston, keeping her pledge to her first husband. In 1872, his father realized the opportunities for his son were better in Boston, and he moved there with his son. Finding Boston inhospitable, Puritanical, and cold, the father returned alone to Ávila within a few months. The separation between father and mother was permanent. In 1888 Agustín wrote to Josefina: "When we were married I felt as if it were written that I should be reunited with you, yielding to the force of destiny. Strange marriage, this of ours! So you say, and so it is in fact. I love you very much, and you too have cared for me, yet we do not live together" (*Persons and Places*, 9).

Until his father's death (1893), Santayana regularly corresponded with his father and he visited him after Santayana's first year at Harvard College. In Boston, Santayana's family spoke only Spanish in their home. Santayana first attended Mrs. Welchman's Kindergarten to learn English from the younger children, then he was a student at the Boston Latin School, and he completed his B.A. and Ph.D. at Harvard College (1882-1889), including eighteen months of study in Germany on a Walker Fellowship. His undergraduate years at Harvard reveal an energetic student with an active social life. He was a member of eleven organizations including *The Lampoon* (largely as a cartoonist), the *Harvard Monthly* (a founding member), the Philosophical Club (President), and the Hasty Pudding.

Some scholars conclude that Santayana was an active homosexual based on allusions in Santayana's early poetry (McCormick, 49-52) and Santayana's association with known homosexual and bisexual friends. Santayana provides no clear indication of his sexual preferences, and he never married. Attraction to both women and men seems apparent in his undergraduate and graduate correspondence. The one documented comment about his homosexuality occurs when he was sixty-five. After a discussion of A. E. Housman's poetry and homosexuality, Santayana remarked, "I think I must have been that way in my Harvard days -- although I was unconscious of it at the time" (Cory, *Santayana: The Later Years*, 40). Because of Santayana's well-known frankness, many scholars consider Santayana a latent homosexual based on this evidence.

1886-1912. Santayana received his Ph.D. from Harvard in 1889 and became a faculty member at Harvard University (1889-1912) and eventually a central figure in the era now called Classical American Philosophy. He was a highly respected and popular teacher, and his students included poets (Conrad Aiken, T. S. Eliot, Robert Frost, Wallace Stevens), journalists and writers (Walter Lippmann, Max Eastman, Van Wyck Brooks), professors (Samuel Eliot Morison, Harry Austryn Wolfson), a Supreme Court Justice (Felix Frankfurter), many diplomats (including his friend Bronson Cutting), and a university president (James B. Conant). He retired from Harvard in 1912 at the age of forty-eight and lived the remainder of his life in England and Europe, never returning to the U.S. and rejecting academic posts offered at a number of universities, including Harvard, Columbia, and Cambridge.

Santayana cherished academic life for its freedom to pursue intellectual interests and curiosity, but he found that many aspects of being a professor infringed on that freedom. Faculty meetings and university committees seemed primarily partisan heat over false issues, so he rarely attended them. The general corporate and businesslike adaptation of universities was increasingly less conducive to intellectual development and growth. He expressed concern about the evolving Harvard goal of producing muscular intellectuals to lead America as statesmen in business and government. Were not delight and celebration also a central aspect of education? He wrote to a friend in 1892, expressing the hope that his academic life would be "resolutely unconventional" and noted that he could only be a professor *per accidens*, saying that "I would rather beg than be one essentially" (GS to H. W. Abbot, Stoughton Hall, Harvard, 15 February 1892. Columbia).

In 1893, Santayana experienced a *metanoia*, a change of heart. Gradually he altered his style of life from that of an active student turned professor to one focused on the imaginative celebration of life. In doing so, he began planning for his early retirement, finding university life increasingly less conducive to intellectual pursuits and delight in living. Three events preceded his *metanoia*: the unexpected death of a young student, witnessing his father's death, and the marriage of his sister Susana. Santayana's reflections on these events led to the ancient wisdom that acceptance of the tragic leads to a lyrical release. "Cultivate imagination, love it, give it endless forms, but do not let it deceive you. Enjoy the world, travel over it, and learn its ways, but do not let it hold you … . To possess things and persons in idea is the only pure good to be got out of them; to possess them physically or legally is a burden and a snare (*Persons and Places*, 427-28)."

Increasingly, naturalism and the lyrical cry of human imagination became the focal points of Santayana's life and thought. Pragmatism, as developed by Peirce and James, was an undercurrent in his naturalism, particularly as an approach to how we ascertain knowledge, but there are aspects of his naturalism more aligned with European and Greek thought that presage developments in the late twentieth century. His naturalism had its historical roots primarily in Aristotle and Spinoza and its contemporary background in James's pragmatism and Royce's idealism. His focus on and celebration of creative imagination in all human endeavors (particularly in art, philosophy, religion, literature, and science) is one of Santayana's major contributions to American thought. This focus, along with his Spanish heritage, Catholic upbringing, and European suspicion of American industry, set him apart in the Harvard Yard.

Santayana's strong interest in literature and aesthetics is evident throughout this early period, but by 1904, his attention turned almost fully to philosophical pursuits. During this period his publications include: *Lotze's System of Philosophy* (dissertation), *Sonnets and Other Verses* (1894), *The Sense of Beauty* (1896), *Lucifer: A Theological Tragedy* (1899), *Interpretations of Poetry and Religion* (1900), *A Hermit of Carmel, and Other Poems* (1901), *The Life of Reason* (five books, 1905-1906), *Three Philosophical Poets: Lucretius, Dante, and Goethe* (1910).

In May 1911, Santayana formally announced his long-planned retirement from Harvard. President Lowell asked him to reconsider. By now Santayana was a highly recognized philosopher, cultural critic, poet, and teacher, and his desire to be free from academic confinement was also well known. Lowell indicated he was open to any arrangement that provided Santayana the time he desired for writing and for travel in Europe. Initially Santayana agreed to alternate years in Europe and the U.S., but in 1912, his resolve to retire overtook his sense of obligation to Harvard. The year before his retirement, he had presented at least six lectures at a variety of universities including Berkeley, Wisconsin, Columbia, and Williams. His books were selling well and his publishers were asking for more. Two major universities were courting him. At forty-eight, he left Harvard to become a full-time writer and to escape the academic professionalism that nurtured a university overgrown with "thistles of trivial and narrow scholarship."

1912-1952. As Santayana sailed for Europe, his mother died, apparently of Alzheimer's disease. Always attentive to his family, Santayana visited her weekly, then daily, during his last years at Harvard. Knowing his mother's death was imminent, he arranged for Josephine, his half sister, to live in Spain with Susana, who previously had married a well-to-do Ávilan. An inheritance of $10,000 from his mother, coupled with his steady income from publications and his early planning, made retirement easier. He arranged for his half brother, Robert, to manage his finances with the agreement that upon Santayana's death, Robert or his heirs would receive the bulk of Santayana's estate. Hence, in January 1912, at age forty-eight, Santayana was free from the constraints of university regimen and expectations and, more importantly, free to write, to travel, and to choose his residence and country.

Santayana's book publications after leaving Harvard is remarkable: *Winds of Doctrine* (1913), *Egotism in German Philosophy* (1915), *Character and Opinion in the United States* (1920), *Soliloquies in England and Later Soliloquies* (1922), *Scepticism and Animal Faith* (1923), *Dialogues in Limbo* (1926),

Platonism and the Spiritual Life (1927), the four books of *The Realms of Being* (1927, 1930, 1938, 1940), *The Genteel Tradition at Bay* (1931), *Some Turns of Thought in Modern Philosophy* (1933), *The Last Puritan* (1935), *Persons and Places* (1944), *The Middle Span* (1945), *The Idea of Christ in the Gospels* (1946), *Dominations and Powers* (1951), and *My Host the World* (1953, posthumous).

Harvard attempted to bring him back to the United States, offering him several professorships beginning in 1917. As late as 1929, he was offered the Norton Chair in Poetry, one of Harvard's most respected chairs. In 1931, he received an invitation from Brown University, and Harvard later asked him to accept the William James Lecturer in Philosophy, a newly established honorary post. But Santayana never returned to Harvard or to America. Believing that the academic life was not a place for him to cultivate intellectual achievement or scholarly work, Santayana also refused academic appointments both at Oxford University and Cambridge University.

At first, Santayana planned to reside in Europe, and after numerous exploratory trips to several cities, he decided on Paris. However, while he was in England, World War I broke out and he was unable to return to the mainland. First, he lived in London and then primarily at Oxford and Cambridge. After the war, he was more of a traveling scholar, and his principal locales included Paris, Madrid, Ávila, the Riviera, Florence, and Rome. By the late 1920s, he settled principally in Rome, and during the summers, he often retreated to Cortina d'Ampezzo in northern Italy to write and to escape the heat. Because of his success as a writer, he assisted friends and scholars when they found themselves in need of financial support. For example, when Bertrand Russell was unable to find a teaching post in the U.S. or England because of his views regarding pacifism and marriage, Santayana displayed a characteristic generosity in his plan to make an anonymous gift to Bertrand Russell of the $25,000 royalty earnings from *The Last Puritan*, at the rate of $5,000 per year, in the letter to George Sturgis (15 July 1937). Despite the fact that he and Russell disagreed radically both politically and philosophically, his memory of their earlier friendship and his regard for Russell's genius moved him to compassion for Russell's financial plight.

The rise of Mussolini in the 1930s initially seemed positive to Santayana. He viewed the Italian civil society as chaotic and thought Mussolini might bring order where needed. But Santayana soon noted the rise of a tyrant. Trying to leave Italy by train for Switzerland, he was not permitted to cross the border because he did not have the proper papers. With most of his funds coming from the United States and England, his case was complicated by his Spanish citizenship and his age. He returned to Rome, and on 14 October 1941 he entered the Clinica della Piccola Compagna di Maria, a hospital-clinic run by a Catholic order of nuns, where he lived until his death eleven years later. This arrangement was not unusual. The hospital periodically received distinguished guests and cared for them in an assisted-living environment. Santayana died of cancer on 26 September 1952.

Santayana asked that he be buried in unconsecrated ground, affirming his naturalism to the end. However, the only such cemetery ground in Rome was reserved for criminals. The Spanish Consulate at Rome would not permit Santayana to be buried in such a place and provided the "Panteon de la Obra Pia espanola" in the Campo Verano cemetery as a suitable burial ground, turning it into a memorial for the lifelong Spanish citizen. At the graveside, Daniel Cory read lines from Santayana's "The Poet's Testament," a poem affirming his naturalistic outlook:

> I give back to the earth what the earth gave,
> All to the furrow, nothing to the grave.
> The candle's out, the spirit's vigil spent;
> Sight may not follow where the vision went.

In the United States, Wallace Stevens commemorated his teacher in "To an Old Philosopher in Rome."

> Total grandeur of a total edifice,
> Chosen by an inquisitor of structures
> For himself. He stops upon this threshold,
> As if the design of all his words takes form
> And frame from thinking and is realized.

2. Philosophy, Literature, and Culture

Throughout his life, Santayana's literary achievements are evident. As an eight-year-old Spaniard, he wrote *Un matrimonio* (A Married Couple), describing the trip of a newly married couple that meets the Queen of Spain. Later in Boston, he wrote a poetic parody of *The Aeneid*; "A Short History of the Class of '82"; and "Lines on Leaving the Bedford St. Schoolhouse." His first book, *Sonnets and Other Verses* (1894), is a book of poems, not philosophy. And, until the turn of the century, much of his intellectual life was directed to the writing of verse and drama. He was a principal figure in making modernism possible but was not a modernist in poetry or literature. His naturalism and emphasis on constructive imagination influenced both T. S. Eliot and Wallace Stevens. Eliot's notion of the "objective correlative" is drawn from Santayana, and Stevens follows Santayana in his refined naturalism by incorporating both Platonism and Christianity without any nostalgia for God or dogma.

Santayana was among the leaders in transforming the American literary canon, dislodging the dominant Longfellow, Lowell, Whittier, Holmes, Bryant canon. Santayana's essay "The Genteel Tradition in American Philosophy" (presented to the Philosophical Union of the University of California in 1911) greatly affected Van Wyck Brooks's *America's Coming-of-Age*, a book that set the tone for modernism. Brooks drew on Santayana's essay, adapting Santayana's idea of two Americas to fit his notion of an America split between highbrow and lowbrow culture.

By the turn of the century, Santayana's interests largely centered on his philosophical inquiries, and although he never abandoned writing poetry, he no longer considered it his central work. Even so, some of his most moving poetry came later and was inspired by the trench warfare and casualties of World War I: "A Premonition: Cambridge, October, 1913"; "The Undergraduate Killed in Battle: Oxford, 1915"; "Sonnet: Oxford, 1916"; and "The Darkest Hour: Oxford, 1917." Throughout his life, even near death, he recited and translated long fragments of Horace, Racine, Leopardi, and others.

The relationships between literature, art, religion and philosophy are prominent themes throughout Santayana's writings. *The Sense of Beauty* (1896) is a primary source for the study of aesthetics. Philip Blair Rice wrote in the foreword to the 1955 Modern Library edition: "To say that aesthetic theory in America reached maturity with *The Sense of Beauty* is in no way an overstatement. Only John Dewey's *Art as Experience* has competed with it in the esteem of philosophical students of aesthetics and has approached its suggestiveness for artists, critics and the public which takes a thoughtful interest in the arts." Santayana's groundbreaking approach to aesthetics is emphasized in Arthur Danto's "Introduction" to the 1988 critical edition. Danto writes that Santayana brings "beauty down to earth" by treating it as a subject for science and giving it a central role in human conduct, in contrast to the preceding intellectualist tradition of aesthetics. "The exaltation of emotion and the naturalization of beauty -- *especially* of beauty -- imply a revolutionary impulse for a book it takes a certain violent act of historical imagination to recover" (*Sense of Beauty*, xxviii). This naturalistic approach to aesthetics is expanded in his philosophical explication of art found in *The Life of Reason: Reason in Art* (1905).

In 1900, Santayana's *Interpretations of Poetry and Religion* develops his view that religion and poetry are expressive celebrations of life. Each in its own right is of great value, but if either is mistaken for science, the art of life is lost along with the beauty of poetry and religion. Science provides explanations of natural phenomena, but poetry and religion are festive celebrations of human life born of consciousness generated from the interaction of one's psyche (the natural structure and heritable traits of ones physical body) and the physical environment. As expressions of human values, poetry and religion are identical in origin. Understanding the naturalistic base for poetry and religion and valuing their expressive character enable one to appreciate them without being hoodwinked: "poetry loses its frivolity and ceases to demoralise, while religion surrenders its illusions and ceases to deceive" (172). Interestingly, his father expressed similar views in his letters to his son, providing the genesis of his son's reflections, and this conclusion is expressed as late as the 1946 publication of *The Idea of Christ in the Gospels* where Santayana presents the idea of Christ as poetic and imaginative, contrasted with attempts at historical, factual accounts of the Christ figure. The impact of Santayana's view was significant, and Henry James (after reading *Interpretations of Poetry and Religion*) wrote that he would "crawl across London" if need be to meet Santayana.

Three Philosophical Poets (1910) was the first volume of the Harvard Studies in Comparative Literature. Santayana employs a naturalistic account of poetry and philosophy, attempting to combine comparative structures with as few embedded parochial assumptions as possible while making explicit our material boundness to particular worlds and perspectives. His analyses of Lucretius, Dante, and Goethe are described by one biographer as "a classical work and one of the few written in America to be genuinely comparative in conception and execution, for its absence of national bias and its intellectual, linguistic, and aesthetic range" (McCormick, 193).

Initially, Santayana appears optimistic about the youthful America. In his Berkeley lecture, "The Genteel Tradition in American Philosophy," he declared "the American Will inhabits the sky-scraper; the American Intellect inhabits the colonial mansion." ("The Genteel Tradition in American Philosophy," Triton Edition, vol. VII. P. 129.) European transcendentalism and Calvinism are the American intellectual traditions, but they no longer suit the American drive for success in industry, business, and

football. Hence, the youthful willfulness of the country has outrun the old wits, but there remains a chance for wisdom and energy to be coupled in a future coherent and rich tradition, and he sees the beginnings of such a tradition in James's pragmatism.

Within a decade, he is less optimistic. *Character and Opinion in the United States* (1920) is his valediction to America. It includes frank, intellectual portraits of his Harvard colleagues and of American culture. From his residence in Cambridge, he praises the English emphasis on social cooperation and personal integrity and contrasts them with America where "You must wave, you must cheer, you must push with the irresistible crowd; otherwise you will feel like a traitor, a soulless outcast, a deserted ship high and dry on the shore … . This national faith and morality are vague in idea, but inexorable in spirit; they are the gospel of work and the belief in progress. By them, in a country where all men are free, every man finds that what most matters has been settled beforehand" (211).

Santayana's standing as a literary figure reached its zenith with the publication of *The Last Puritan* (1936). *The Last Puritan* is Santayana's only novel, and it was an international success. It was compared positively with Goethe's *Wilhelm Meister*, Pater's *Marius*, and Mann's *The Magic Mountain*. Its provenance lies in the 1890s when Santayana began a series of sketches on college life that, broadened through his experience and travel, resulted in *The Last Puritan*. Essentially, it is about the life and early death of an American youth, Oliver Alden, who is sadly restricted by his Puritanism. Santayana draws a sharp contrast with the European Mario, who delights in all matters without a narrow moralism. Mario is a carefree, naturally gifted and likeable young man who by American standards appears too focused on the peripheral aspects of life: travel, opera, love affairs, and architecture. And the American perspective is embodied in the tragic hero, Oliver Alden, who is the last puritan. He does what is right, based on his duties to his family, school, and friends. Life is a slow, powerful flow of tasks and responsibilities. He is intelligent and knows there is more than obligation, and he senses his guilt at not being able to achieve the natural abundant life, but knowing this only nourishes his Puritanism and causes him to feel guilty about being guilty. In a charming scene in the novel, Oliver introduces Mario to Professor Santayana at Harvard. Oliver is a dedicated student and football player, thoroughly a first rate American taking matters seriously and doing his best. After only a short visit with the Professor, Mario, it is decided by Santayana, does not need to take a course from the Professor. Mario already has the natural, instinctual approach of a cultivated person. Oliver, on the other hand, knows he must work to achieve his goal, which will be only a succession of goals, and ends tragically. Santayana's Hispanic and Catholic background play a central role in his critique of American life: too bound by past traditions and obligations that are not understood or rooted in one's own culture.

The fear that Santayana's autobiography would be lost or destroyed during World War II, led Scribner's, the publisher, to conspire with the U.S. Department of State, the Vatican, and the Spanish government to bring the manuscript of the first part (*Persons and Places*) out of Rome *sub rosa*, despite the Italian government's refusal to allow any mail to the U.S. The manuscript for the second part (*The Middle Span*, 1945) also was conveyed surreptitiously to New York. The third part (*My Host the World*, 1953) was published after Santayana's death. His autobiography provides the basis for understanding the development of his philosophy

3. Development of Santayana's Philosophy

In his autobiography, *Persons and Places*, Santayana describes the development of his thought as a movement from the idealisms of boyhood to the intellectual materialism of a traveling student, and finally to the complete, naturalistic outlook of the adult Santayana. He emphasizes the continuity of his life and beliefs, contrasting what may appear to be disparate views with the overall unity of his thought: "The more I change the more I am the same person" (*Persons and Places*, 159).

As a young man of the nineteenth century, he was influenced by the idealism of the age and of his age, but he claims to have always been a realist or naturalist at heart.

But those ideal universes in my head did not produce any firm convictions or actual duties. They had nothing to do with the wretched poverty-stricken real world in which I was condemned to live. That the real was rotten and only the imaginary at all interesting seemed to me axiomatic. That was too sweeping; yet allowing for the rash generalisations of youth, it is still what I think. My philosophy has never changed. (*Persons and Places*, 167)

Hence he notes, that in spite "of my religious and other day-dreams, I was at bottom a young realist; I knew I was dreaming, and so was awake. A sure proof of this was that I was never anxious about what those dreams would have involved if they had been true. I never had the least touch of superstition" (*Persons and Places*, 167). Santayana cites poems, "To the Moon" and "To the Host," written when he was fifteen or sixteen, as revealing this early realism, and he quotes from memory one stanza of "At the Church Door" where the realistic sentiment is the same (*Persons and Places*, 169).

By the time he was a traveling student seeing the world in Germany, England, and Spain his "intellectual materialism" was firmly established with little change in his religious affections.

From the boy dreaming awake in the church of the Immaculate Conception, to the travelling student seeing the world in Germany, England, and Spain there had been no great change in sentiment. I was still "at the church door". Yet in belief, in the clarification of my philosophy, I had taken an important step. I no longer wavered between alternative vies of the world, to be put on or taken off like alternative plays at the theatre. I now saw that there was only one possible play, the actual history of nature and of mankind, although there might well be ghosts among the characters and soliloquies among the speeches. Religions, all religions, and idealistic philosophies, all idealistic philosophies, were the soliloquies and the ghosts. The might be eloquent and profound. Like Hamlet's soliloquy they might be excellent reflective criticisms of the play as a whole. Nevertheless they were only parts of it, and their value as criticisms lay entirely in their fidelity to the facts, and to the sentiments which those facts aroused in the critic. (*Persons and Places*, 169)

The full statement and development of his materialism did not occur until later in his life. It was certainly in place by the time of *Scepticism and Animal Faith* (1923) but not fully so at the time of *The Life of*

Reason (1905). The influence of the Harvard philosophers, particularly James and Royce is evident in Santayana's thought, but he was hardly a mere follower and often advanced his philosophy more along European and Greek lines rather than the American tradition, which he thought was both too derivative and too tied to the advancement of business and capitalism.

The move from Harvard marked not only a geographical shift but a philosophical one as well. Henry Levinson in *Santayana, Pragmatism, and the Spiritual Life* provides a well-balanced account of this gradual but distinctive move from the Harvard philosophical mentality. Leaving Harvard also meant that Santayana abandoned the view of a philosopher as a public, philosophical statesman and of language as being representative. This philosophical turn placed makes him a forerunner of many issues in the next two centuries. Removing himself, literally and philosophically, from the American scene, Santayana increasingly came to believe that the "brimstone" sensibility of pragmatism was wrong-headed (*Character and Opinion in the United States*, 53). A major aspect of this sensibility was the view that philosophers must be engaged fundamentally in social and cultural policy formulation, and if they are not, they are not pulling their civic weight. In this fashion, Santayana believes the pragmatists came to belie "the genuinely expressive, poetic, meditative, and festive character of their vocation" (Levinson,165). A condition that James took seriously in his "On a Certain Blindness in Human Beings," suggesting that the world of practical responsibility fosters a blindness to multiplural ways of living that can only be escaped by catching sight of "the world of impersonal worths as such" -- "only your mystic, your dreamer, or your insolent loafer or tramp can afford so sympathetic an occupation" (James, *Talks to Teachers on Psychology and to Students on Some of Life's Ideals*, 141). Interestingly, America's imperialistic actions toward the Philippines during the Spanish-American War sparked James' remarks; this was a war that had a much deeper ancestral and historical aspect for Santayana and led to his poem, "Young Sammy's First Wild Oats." Whether connected or not, Santayana later came to identify himself as an intellectual vagabond or tramp, not isolated in the specific perspectives of an ideology, hosted by the world, and devoted to spiritual disciplines that "appear irresponsible to philosophers hoping to command representative or some otherwise privileged authority at the center of society" (Levinson, 167).

Building on his naturalism, institutional pragmatism, social realism, and poetic religion, Santayana on leaving Harvard moves even farther from the role of philosophical statesman by removing the representative authority of language from the quest for a comprehensive synthesis, by narrowing the line between literature and philosophy (as he had earlier done between religion and poetry), and by wrestling more with the influence of James than of Emerson. Santayana's stay in Oxford during the Great War led to his famous counter to Wilson's war to end all wars: "Only the dead have seen the end of war." (*Soliloquies in England and Later Soliloquies*, 102)

Santayana's message is clear: The epistemological project that Russell's *Problems* epitomizes is diseased. The renewed quest to establish unmediated Knowledge of Reality simply leads to "intellectual cramp" (*Soliloquies*, 216). Philosophy has itself become spiritually *disordered* by blinding its practitioners from their traditional and proper task, which is to celebrate the good life. If the spiritual disciplines of philosophy are to thrive, philosophers have to take off the bandages of epistemology and metaphysics altogether, accept the finite and fallible status of their knowledge claims, and get on with confessing their belief in the things that make life worth living. (Levinson, 204)

Leaving Harvard and America enabled Santayana to develop his naturalism.

4. Naturalism

Scepticism and Animal Faith (1923) introduces Santayana's mature naturalism. In summary, he maintains that knowledge and belief are not the result of reasoning. They are inescapable beliefs essential for action. Epistemological foundationalism is a futile approach to knowledge. A more promising approach is to discern the underlying belief structures assumed in animal action and imposed by natural circumstances. The foundations for this approach are rooted in Aristotle's concept of activity and the pragmatic approach to action and knowledge. Explanations of natural events are the proper purview of scientists, while explications of the meaning and value of action may be the proper sphere of historians and philosophers. Even so, both scientific explanation and philosophical explication are based in the natural world. Meaning and value are generated by the interaction of our physical makeup, which Santayana calls "psyche," and our material environment.

Santayana's critique of epistemological foundationalism is as unique as his heritage. With Spanish irony, he structures his argument after Descartes' *Meditations* but arrives at an anti-foundationalist conclusion. Drawing attention to what is given in an instant of awareness (the smallest conceivable moment of consciousness), he maintains that any knowledge or recognition found in such an instant would have to be characterized by a concept (or "essence" to use Santayana's term). Concepts cannot be limited to particular instances; rather the particular object is seen as an instance of the concept (essence). Thus, pursuing doubt to its ultimate end, one is confined by the "solipsism of the present moment." That is, in a single instant of awareness there can be no knowledge or belief, since both require concepts not bounded by a moment of awareness. Hence, the ultimate end of doubt, an instance of awareness, is empty. It is the vacant awareness of a given without a basis for belief, knowledge or action. Santayana concludes that if one attempts to find the bedrock of certainty, one may rest his claim only after he has, at least theoretically, recognized that knowledge is composed of instances of awareness that in themselves do not contain the prerequisites for knowledge, i.e., concepts, universals, or essences. That both skepticism and proofs against skepticism lead nowhere is precisely Santayana's point.

Philosophy must begin *in medias res* (in the middle of things), in action itself, where there is an instinctive and arational belief in the natural world: "animal faith." For Santayana, animal faith is the arational basis for any knowledge claims. It is the nether world of biological order operating through our physical, non-conscious being generating beliefs that are "radically incapable of proof" (*Scepticism*, 35).

It is a vital constitutional necessity, to belief in discourse, in experience … . All these objects may conceivably be illusory. Belief in them however, is not grounded on a prior probability, but all judgements of probability are grounded on them. They express a rational instinct or instinctive reason, the waxing faith of an animal living in a world which he can observe and sometimes remodel (*Scepticism*, 308-309).

He describes these prejudices as "animal" in an effort to emphasize our biological base and community. This emphasis is similar to Wittgenstein's reference to convictions that are beyond being justified or unjustified as "something animal" (*On Certainty* 359). Ours is a long-standing primitive credulity, and our most basic beliefs are those of an animal creed: "that there is a world, that there is a future, that things sought can be found, and things seen can be eaten" (*Scepticism* 180).

Santayana (like Hume, Wittgenstein, and Strawson) holds there are certain inevitable beliefs; they are inescapable given nature and our individual physical history. And like Wittgenstein, he maintains that these beliefs are various and variable. They are determined by the interplay between environment and psyche, i.e., between our natural conditions and the inherited, physical "organisation of the animal" (the psyche). That we now inescapably believe in external objects and the general reliability of inductive reasoning, for example, is a result of physical history and the natural conditions of our world and ourselves. Since these beliefs are relative to our physical histories, if our history and biological order had been different, our natural beliefs would also be different.

The environment determines the occasions on which intuitions arise, the psyche -- the inherited organisation of the animal -- determines their form, and ancient conditions of life on earth no doubt determined which psyches should arise and prosper; and probably many forms of intuition, unthinkable to man, express the facts and the rhythms of nature to other animal minds (*Scepticism*, 88).

By displacing privileged mentalistic accounts with his pragmatic naturalism, Santayana challenges then prevailing structures in both American and English philosophies.

Santayana explicates the primary distinguishable characteristics of our knowledge in his four-book *Realms of Being*. Believing that philosophical terminology should have historical roots, Santayana employed classical terminology for these characteristics: matter, essence, spirit, and truth. And although these terms are central to many philosophical traditions, he views his work as "a revision of the categories of common sense, faithful in spirit to orthodox human tradition, and endeavouring only to clarify those categories and disentangle the confusions that inevitably arise … ." (*Realms* 826)

Within Santayana's naturalism, the origins of all events in the world are arbitrary, temporal, and contingent. Matter (by whatever name it is called) is the principle of existence. It is "often untoward, and an occasion of imperfection or conflict in things." (*Realm of Matter*, v) Hence, a "sour moralist" may consider it evil, but, according to Santayana, if one takes a wider view "matter would seem a good … because it is the principle of existence: it is all things in their potentiality and therefore the condition of all their excellence or possible perfection." (*Realm of Matter*, v) Matter is the non-discursive, natural foundation for all that is. In itself, it is neither good nor evil but may be perceived as such when viewed from the vested interest of animal life. Latent animal interests convert matter's non-discernible, neutral face to a smile or frown. But "moral values cannot preside over nature." (*Realm of Matter*, 134) Principled values are the products of natural forces: "The germination, definition, and prevalence of any good must be grounded in nature herself, not in human eloquence." (*Realm of Matter*, 131) From the point of view of origins, therefore, the realm of matter is the matrix and the source of everything: it is

nature, the sphere of genesis, the universal mother.

“Essence” is Santayana’s term for concepts and meanings. He draws on Aristotle’s notion of essence but removes all capacities for producing effects. An essence is a universal, an object of thought, not a material force. However, consciousness of an essence is generated by the interaction of a psyche and the material environment. Hence, matter remains as the origin of existence and the arena of action, and the realm of essence encompasses all possible thought.

“Truth,” if some disinterested observer could ascertain it, would constitute all the essences that genuinely characterize the natural world and all activities within it. Since all living beings have natural interests and preferences, no such knowledge of truth can exist. All conscious beings must ascertain belief about truth based on the success of actions that sustain life and permit periods of delight and joy.

Santayana uses the term “spirit” to mean consciousness or awareness. As early as 19 April 1909, Santayana wrote to his sister that he was writing a brand new system of philosophy to be called “The realms of Being” -- not the mineral vegetable and animal, but something far more metaphysical, namely Essence, Matter, and Consciousness. It will not be a long book, but very technical.” When the book was published in the 1930s, he had added his notion of truth and substituted “spirit” for “consciousness.” From his perspective, the substitution did not alter the meaning of consciousness but rather captured an entire tradition of philosophical and religious inquiry as well as borrowed associated ideas from eastern religions. But to the consternation of traditional views, many found the identity of spirit with consciousness a troublesome idea. And so they should, for with this identity Santayana removes the spiritual from the field of agency as well as from an alternative lifestyle. Following the tracks of Aristotle, he makes the spiritual life one form of culminating experiences arising from fulfilling activity.

Awareness evolved through the natural development of the physical world, and he demurs to scientific accounts for explanations of that development. Almost poetically, he sees spirit as emerging in moments of harmony between the psyche and the environment. Such harmony is temporary, and the disorganized natural forces permit spirit to arise "only spasmodically, to suffer and to fail. For just as the birth of spirit is joyous, because some nascent harmony evokes it, so the rending or smothering of that harmony, if not sudden, imposes useless struggles and suffering (*Birth of Reason,*53)." Accepting the world’s insecure equilibrium enables one to celebrate the birth of spirit. Reasoning, particularly reasoning associated with action, is a signal of the nascent activities of the psyche working to harmonize its actions with the environment, and if successful, reason permits individual and social organization to prosper while spirit leads to the delight of imagination and artistry.

Some commentators characterize Santayana as an epiphenomenalist, and there are some commonalities, specifically the view that spirit is not efficacious. But there also are considerable differences. Santayana does not characterize his view as one-way interactionism, primarily because he does not think of spirit as an object to be acted upon. Spirit is rather a distinguishable aspect of thought, generated in activity, and may be viewed more as a relational property. Santayana sometimes speaks of spirit and essence as supervening on material events. But lacking the distinctions of contemporary philosophy, it is difficult to

characterize Santayana’s philosophy of mind accurately. His view of consciousness is more celebrational, as opposed to being a burden or eliciting action. Spirit is "precisely the voice of order in nature, the music, as full of light as of motion, of joy as of peace, that comes with an even partial and momentary perfection in some vital rhythm (*Birth of Reason*, 53).”

Santayana’s account of spirit and essence, may lead one to wonder how Santayana can be included as a pragmatist, and this classification is accurate only if one includes an extended notion of pragmatic naturalism. For Santayana, explanations of human life, including reason and spirit, lie within the sciences. The nature of truth simply is correspondence with what is, but since humans, nor any other conscious being, are able to see beyond the determinant limits of their nature and environment, pragmatism becomes the test of truth rather than correspondence. In short, the nature of truth is correspondence while the test of truth is pragmatic. If an explanation continues to bear fruit over the long run, then it is accepted as truth until it is replaced by a better explanation. In this, Santayana’s account of pragmatic truth is more closely aligned with Peirce’s conception than that of James or Dewey, including a tripartite account of knowledge consisting of the subject, symbol, and object. Pragmatism properly is focused on scientific inquiry and explanations, and it is severely limited, even useless, in spiritual and aesthetic matters. Pragmatism is rooted in animal life, the need to know the world in a way that fosters successful action. If all life were constituted only by successful or unsuccessful activities, one’s fated circumstances would govern. But consciousness makes liberation possible and brings delight and festivity in material circumstances.

Santayana’s anti-foundationalism, non-reductive materialism, and pragmatic naturalism coupled with his emphasis on the spiritual life and his view of philosophy as literature anticipated many developments in philosophy and literary criticism that occurred in the latter half of the twentieth century, and these served as a challenge to the more humanistic naturalisms of John Dewey and other American naturalists. These views also provide the foundation for his view of ethics, political philosophy, and the spiritual life.

5. Ethics, Politics, and the Spiritual Life

Santayana’s moral philosophy is based on his naturalism. Most commentators classify Santayana as an extreme moral relativist who maintains that all individual moral perspectives have equal standing and are based on the heritable traits and environmental circumstances of individuals. This naturalistic approach applies to all living organisms. Nature does not establish a moral hierarchy of goods between animal populations nor between individual animals. However, this same moral relativism is also the basis for Santayana’s claim that the good of individual animals is clear and is subject to naturalistic or biological investigation.

Two tenets of his ethics are (1) the forms of the good are diverse, and (2) the good of each animal is definite and final. The moral terrain of animals, viewed from a neutral perspective, places all animal interests and goods as equal. Each good stems from heritable physical traits and is shaped by adaptations to the environment. Concluding that the “forms of the good are divergent,” Santayana holds that the good for each animal may differ, depending on the nature of the psyche and the circumstances, and may be

different for an individual animal in different times and environments. There is no one good for all, or even for an individual.

Seen as a whole, animal goods are not logically or morally ordered, that are natural, morally neutral forces. But no living being can observe all interests with such neutrality. Situated in a particular place and time with heritable traits, all living beings have interests originating from their physiology and physical environment. For Santayana, one may reasonably note that a neutral observer could view all moral perspectives as equal, but such a view must be balanced by the understanding that no animal stands on neutral ground. There is a polarity between the ideal neutral, objective understanding of behavior on the one hand and the committed and vested interest of particular living beings on the other hand. One may recognize that every animal good has its own standing, and one may respect that ideal, but "the right of alien natures to pursue their proper aims can never abolish our right to pursue ours."(*Persons and Places, 179)*

Santayana's second moral insight is that for each animal the good is definite and final. There are specific goods for each animal depending on the specific heritable traits and interests of the psyche and on the specific circumstances of the environment. Self-knowledge, then, is the distinguishing moral mark. The extent to which one knows one's interests, their complexity and centrality, will determine whether one can achieve a good life, provided the environment is accommodating. Santayana's philosophy rests on his naturalism and on his humane and sympathetic appreciation for the excellence of each life. But from the perspective of autobiography, Santayana's clear notion of self-knowledge, in the sense of the Greeks, is his most distinguishing mark. For Santayana, "integrity or self-definition is and remains first and fundamental in morals … ."(*Persons and Places* 170)

Self-knowledge requires a critical appreciation of one's culture and physical inheritance, and the ability to shape one's life in streams of conflicting goods within oneself and within one's community. Although this position is common to many considerations of political philosophy, Santayana's approach to politics was much more conservative than that normally associated with the founders of American pragmatism, such as James and Dewey.

Santayana's political conservatism is founded on his naturalism and his emphasis on self-realization and spirituality. He is concerned that liberal democracy may not provide a consistent basis for individual freedom and spirituality. The twin fears of private anarchy and public uniformity are the grounds for his criticisms of democracy, and his account of social justice focuses on the individual rather than the society. Santayana's inattentiveness to social inequality is perhaps understandable in the context of his naturalism where the final cause is the "authority of things." His basic contention that individual suffering is the worse feature of human life, not social inequality, causes him to focus more on the natural dilemmas of the individual rather than on social action. Coupling this argument with the view that all institutions, including governments, are inextricably rooted in their culture and background perhaps makes it understandable that he would not readily see how particular views of social inequality could be transferred readily from one culture to another. In addition, Santayana's European and particularly Spanish background influenced his attitudes toward social action. His repeated "Latin" perspective caused him to look with considerable suspicion toward forcing Anglo-Saxon outlooks on other cultures.

Yet, in individual matters he was remarkably forthcoming as when he provided financial support to numerous friends, often of quite different philosophical, literary, and political persuasions than his own.

Within the natural order every living entity stands on the same natural ground bathing equally in the impartial light of nature. No one can claim a central place above others. But each entity also has an embodied set of values, and the art of life is to structure one's environment in such a fashion as to best realize those embodied values, i.e., to place in harmony the natural forces of one's life and one's environment.

American democracy has an exacting challenge. Lacking the time to live in the mind, Americans use quantity as a justification for lack of quality in their achievements. Quantity is potentially infinite and assures unrivaled busy-ness, but is it worth it? No, according to Santayana, if self-realization is the goal of individual life. Of course, circumstances make it difficult, perhaps impossible, for some individuals to order their lives reasonably and attain the practical wisdom to achieve individual happiness. America's economic success would appear to make this possible for many, but to succeed Americans must abandon servility to mechanism and economics. What is needed is a life made free by a recovery of the capacity to have a vision of the good life (*Persons and Places*, xxxiv). According to Santayana, the fanatic is a person who has lost sight of their goals and redoubled their efforts. To supplant this busy, blind, relentlessly quantitative existence, we must regain sight of our goals. Individual life should be structured in light of those goals.

Santayana's focus is on the individual, and the role of the state is to protect and to enable the individual to flourish. The goal is not something far off to be worked toward. It is not a task to be accomplished and then supplanted by another task, as it thought much of American enterprise does. Rather it is the celebration of life in its festivities. It is Aristotle's practical wisdom: structuring individual life as it is, living it joyfully, and assuring that one's commitments are conducive to the delights of the intellect and consistent with the demands of the time and tradition. It is the exercise of one's free choice, shaping one's life through material well-being, but doing so to appreciate the poetic, dramatic quality of our own existence. To rush through life and die without the joy of living, that is the tragedy of American life.

For some, though perhaps not for many, the spiritual life will be an organizing good. The cultural background for the spiritual life is the religious life, primarily as found within the Catholic Church and informed by the late nineteenth century and early twentieth century accounts of eastern religions. But Santayana is not interested in an historical or doctrinal explication of the elements of traditional religion, rather the philosophical task is to discern the elements giving rise to such traditional views, and, in his own case, to explicate the aspects of these origins without the dogmatism of traditional religious belief.

Introducing the concept of a spiritual life, led some to see an inherent conflict between Santayana's life of reason and the spiritual life. In a letter to Milton Karl Munitz (23 July 1939), Santayana explains the different perspectives of the life of reason and the spiritual life:

I admit gladly that religion (= the "Spiritual life") is a natural interest, to be collated within the life of

reason with every other interest; but it is an interest in the ultimate, an adjustment to life, death, science, and politics; and though cultivated specially by certain minds at certain hours, it has no moral or natural claim to predominance. The races and ages in which it is absent will inevitably regard it as unnecessary and obstructive, because they tend to arrange their moral economy without religion at all. Those to whom religion is absorbing (e.g. the Indians) will on the contrary think a moral economy inferior in which no place and no influence is given to the monition of ultimate facts. I think you would not find my two voices inharmonious (I agree that they are different in pitch) if you did not live in America in the XXth century when the "dominance of the foreground" is so pronounced. The dominance of the distance or background would impose a different synthesis.

If the spiritual life was considered a dominating or guiding influence in structuring one's life, the way Santayana views reason, then one would be forced to choose between the life of reason and the life of the spirit as a monk or a nun much choose between the life of the world and that of the religious order. But for Santayana, no such conflict exists; spirituality is not choosing a way of living over an extended period of time. Indeed, any effort to choose such a life would be short lived, since the spiritual life is a life of receptivity to all that comes in the moment while suspending animal interests. Suspending one's specific natural interests, such as eating or sleeping, for any extended period would be both detrimental and tragic.

Consciousness essentially is only awareness, an attention to what is given, rather than being an instrument in reshaping the world. Consciousness, emerging late in the evolutionary pathway, is a flowering of happy circumstances that celebrates what is given, and when truly recognized, does only that. It is joyful, delighting in what is presented, and not troubled by where it leads or what it means. The more dower, moralistic, and evangelical aspects of religion he saw as confused efforts to make religion a science, a social club, or a political movement. Spirit, or consciousness, is momentary, fleeting, and depends on the physical forces of our bodies and environment in order to exist. Shaping one's life to enhance these spiritual, fleeting moments, extending them as long as is practical, is one of the delights of living for some people, but it is certainly not a goal for all, nor should it be.

Bibliography

Primary Sources

The Works of George Santayana. Herman J. Saatkamp, Jr. (General Editor) and William G. Holzberger (Textual Editor). Cambridge, Mass., and London: The MIT Press: *Persons and Places: Fragments of Autobiography*, vol. 1 (1986); *The Sense of Beauty: Being the Outlines of Aesthetic Theory*, vol. 2 (1988); *Interpretations of Poetry and Religion*, vol. 3 (1989); *The Last Puritan: A Memoir in the Form of a Novel*, vol. 4 (1994); *The Letters of George Santayana*, vol. 5, *Book One: [1868]-1909* (2001),*Book Two: 1910-1920* (2002). [Translations of these works include *Interpretaciones de poesía y religión*, Carmen García Trevijano and Susana Nuccetelli, translators, with Introduction by Manuel Garrido, Madrid: Catedra, 1993; *Il senso della Bellezza*, Guiseppe Patella, translator, Palermo: Aesthetica

Edizioni, 1997; and *El Sentido de la belleza*, Carmen García Trevijano, translator, Madrid: Editorial Tecnos, 1999.]

'*Bibliographic Update*,' *Overheard in Seville: Bulletin of the Santayana Society*, edited by Angus Kerr-Lawson and Herman J. Saatkamp, Jr. Indianapolis: Indiana University Purdue University Indianapolis. [Published annually since 1983 containing updated bibliographical information regarding primary and secondary sources.]

- *Animal Faith and Spiritual Life: Previously Unpublished and Uncollected Writings by George Santayana With Critical Essays on His Thought*. Edited by John Lachs. New York: Appleton-Century-Crofts (1967).
- *Atoms of Thought: An Anthology of Thoughts From George Santayana*. Selected and edited, with and introduction, by Ira D. Cardiff. New York: Philosophical Library (1950).
- *Character and Opinion in the United States: With Reminiscences of William James and Josiah Royce and Academic Life in America*. New York: Charles Scribner's Sons (1920).
- *The Birth of Reason and Other Essays by George Santayana*. Edited by Daniel Cory and with an Introduction by Herman J. Saatkamp, Jr. New York: Columbia University Press (1995).
- *The Complete Poems of George Santayana: A Critical Edition*. Edited and with an introduction by William G. Holzberger. Lewisburg: Bucknell University Press; London: Associated University Press (1979).
- *Dialogues in Limbo*. London: Constable and Co. (1925); New York: Scribner's (1926).
- *Dialogues in Limbo, With Three New Dialogues*. New York: Scribner's (1948).
- *Dominations and Powers: Reflections on Liberty, Society, and Government*. New York: Scribner's; London: Constable (1951).
- *Egotism in German Philosophy*. New York: Scribner's (1915).
- *Essays in Literary Criticism of George Santayana*. Selected and edited, with an introduction, by Irving Singer. New York: Scribner's (1956).
- *The Genteel Tradition at Bay*. New York: Scribner's; London: "The Adelphi" (1931).
- The Genteel Tradition: Nine Essays by George Santayana. Edited by Douglas L. Wilson. Cambridge: Harvard University Press (1967).
- *George Santayana's America: Essays on Literature and Culture*. Collected and with an introduction by James Ballowe. Urbana: University of Illinois Press (1967).
- *A Hermit of Carmel, and Other Poems*. New York: Scribner's (1901).
- *The Idea of Christ in the Gospels; or, God in Man: A Critical Essay*. New York: Scribner's; Toronto: Saunders (1946).
- *The Idler and His Works, and Other Essays*. Edited and with a preface by Daniel Cory. New York: Braziller (1957).
- *Interpretations of Poetry and Religion*. New York: Scribner's; London: A. and C. Black (1900).
- *The Last Puritan: A Memoir in the Form of a Novel*. London: Constable (1935); New York: Scribner's (1936).
- *The Life of Reason: Or, The Phases of Human Progress*. New York: Scribner's; London: Constable. Introduction and Reason in Common Sense (1905), Reason in Society (1905), Reason in Religion (1905), Reason in Art (1905), and Reason in Science (1906).

- *Little Essays, Drawn From the Writings of George Santayana by Logan Pearsall Smith, With the Collaboration of the Author*. New York: Scribner's; London: Constable (1920).
- *The Letters of George Santayana*. Edited by Daniel Cory. New York: Scribner's; London: Constable (1955).
- *Lotze's System of Philosophy*. Edited, with an introduction and Lotze bibliography, by Paul Grimley Kuntz. Bloomington: Indiana University Press (1971).
- *Lucifer: A Theological Tragedy*. Chicago and New York: Herbert S. Stone (1899).
- *The Middle Span*. New York: Scribner's (1945); London: Constable (1947).
- *My Host the World*. New York: Scribner's; London: Cresset Press (1953).
- *Obiter Scripta: Lectures, Essays and Reviews*. Edited by Justus Buchler and Benjamin Schwartz. New York: Scribner's; London: Constable (1936).
- *Persons and Places: The Background of My Life*. New York: Scribner's; London: Constable (1944).
- *Philosophy of Santayana: Selections From the Works of George Santayana*. Edited, with an introductory essay, by Irwin Edman. New York: Scribner's (1936).
- *Physical Order and Moral Liberty: Previously Unpublished Essays of George Santayana*. Edited by John and Shirley Lachs. Nashville: Vanderbilt University Press (1969).
- *Platonism and the Spiritual Life*. New York: Scribner's; London: Constable (1927).
- *Poems: Selected by the Author and Revised*. New York: Scribner's; London: Constable (1923).
- *The Poet's Testament: Poems and Two Plays*. New York: Scribner's (1953).
- *Realms of Being*. New York: Scribner's; London: Constable. *The Realm of Essence: Book First* (1927), *The Realm of Matter: Book Second* (1930), *The Realm of Truth: Book Third* (1938, 1937), *The Realm of Spirit: Book Fourth* (1940).
- *Realms of Being*. One-volume edition, with a new introduction by the author. New York: Scribner's (1942).
- *Santayana on America: Essays, Notes, and Letters on American Life, Literature, and Philosophy*. Edited and with an introduction by Richard Colton Lyon (1968).
- *Scepticism and Animal Faith: Introduction to a System of Philosophy*. New York: Scribner's; London: Constable (1923).
- *Selected Critical Writings of George Santayana*. Edited by Norman Henfrey. 2 vols. Cambridge: Cambridge University Press (1968).
- *The Sense of Beauty: Being the Outlines of Aesthetic Theory*. New York: Scribner's; London: A. and C. Black (1896).
- *Soliloquies in England and Later Soliloquies*. New York: Scribner's; London: Constable (1922).
- *Some Turns of Thought in Modern Philosophy: Five Essays*. New York: Scribner's; Cambridge: Cambridge University Press (1933).
- *Sonnets and Other Verses*. Cambridge and Chicago: Stone and Kimball (1894).
- *Three Philosophical Poets: Lucretius, Dante, and Goethe*. Cambridge: Harvard University Press; London: Oxford University Press, (1910).
- *Winds of Doctrine: Studies in Contemporary Opinion*. New York: Scribner's; London: Dent (1913).

Secondary Sources

- Abellán, José Luis. *George Santayana, 1863-1956* [sic]. Madrid: Del Orto (1996).
- Alonso Gamo, José María. *Un español en el mundo: Santayana; poesía y poética*. Madrid: Ediciones Cultura Hispánica (1966).
- Arnett, Willard. *George Santayana*. New York: Washington Square Press (1968).
- Arnett, Willard. *Santayana and the Sense of Beauty*. Bloomington: Indiana University Press; London: M. Paterson (1955, 1984).
- Bosco, Nynfa. *Invito al pensiero di George Santayana*. Milano: Mursia (1987).
- Carter, David. *George Santayana*. New York and Philadelphia: Chelsea House Publishers (1992). A children’s book in the "Hispanics of Achievement" series.
- Cory, Daniel M. *Santayana: The Later Years, A Portrait with Letters*. New York: Braziller (1963).
- Dawidoff, Robert. *The Genteel Tradition and the Sacred Rage*. Chapel Hill: The University of North Carolina Press (1992).
- Duron, Jacques. *La Pensée de George Santayana*: Santayana en Amérique. Paris: Nizet (1949).
- Estébanez Estébanez, Cayetano. *La Obra Literaria de George Santayana*. Valladolid: Secretariado de Publicaciones e Intercambio Editorial, Universidad de Valladolid (2000).
- García Martín, Pedro. *El sustrato abulense de Jorge Santayana*. Avila: Institución “Gran Duque de Alba” de la Excma. Diputación Provincial de Avila (1989).
- Howgate, George W. *George Santayana*. Philadelphia: University of Pennsylvania Press; London: Oxford University Press (1938).
- Hughson, Lois. *Thresholds of Reality: George Santayana and Modernist Poetics*. Port Washington, New York: Kennikat Press (1977).
- Kirby-Smith, H. T. *A Philosophical Novelist: George Santayana and “The Last Puritan”*. Carbondale: Southern Illinois University Press (1997).
- Lachs, John. *George Santayana*. Boston: Twayne Publishers (1988). Twayne’s United States Authors Series.
- Lachs, John and Michael Hodges. *Thinking in the Ruins: Wittgenstein and Santayana on Contingency*. Nashville: Vanderbilt University Press (1999).
- Levinson, Henry Samuel. *Santayana, Pragmatism, and the Spiritual Life*. Chapel Hill: University of North Carolina Press (1992).
- Lind, Bruno. *Vagabond Scholar: A Venture Into the Privacy of George Santayana*. New York: Bridgehead (1962).
- McCormick, John. *George Santayana: A Biography*. New York: Alfred A. Knopf (1987).
- Patella, Giuseppe. *Bellezza, arte e vita. L’estetica mediterranea di George Santayana*. Milano: Mimesis (2000).
- Price, Kenneth M. and Robert C. Leitz. *Critical Essays on George Santayana*. Boston: G. K. Hall and Co. (1991).
- Schilpp, Paul Arthur. *The Philosophy of George Santayana*. Evanston, Illinois: Northwestern University Press (1940).
- Shook, John. *Pragmatism: An Annotated Bibliography 1898-1940*. Amsterdam and Atlanta,

Georgia: Rodopi (1998).

- Singer, Beth. *The Rational Society*. Cleveland: Press of Case Western Reserve University (1970).
- Singer, Irving. *George Santayana, Literary Philosopher*. New Haven: Yale University Press (2000).
- Singer, Irving. *Santayana's Aesthetics: A Critical Introduction*. Cambridge: Harvard University Press (1957).
- *The Southern Journal of Philosophy: Special Issue on Santayana* (summer 1972).
- Sprigge, Timothy. *Santayana: An Examination of his Philosophy*. London and Boston: Routledge & Kegan Paul (1974, 1995).
- Woodward, Anthony. *Living in the Eternal*. Nashville: Vanderbilt University Press (1988).

Other Internet Resources

- Web pages of the Santayana Edition (Indiana University/Purdue University)
- *Overheard in Seville: Bulletin of the Santayana Society*, edited by Angus Kerr-Lawson and Herman J. Saatkamp, Jr. Indianapolis: Indiana University Purdue University Indianapolis.

Related Entries

Dewey, John | epiphenomenalism | James, William | Peirce, Charles Sanders

Herman J. Saatkamp, Jr.
hsaatkam@iupui.edu

First published: February 11, 2002
Content last modified: February 11, 2002

Formal Learning Theory

Formal learning theory is the mathematical embodiment of a normative epistemology. It deals with the question of how an agent should use observations about her environment to arrive at correct and informative conclusions. Philosophers such as Putnam, Glymour and Kelly have developed learning theory as a normative framework for scientific reasoning and inductive inference.

Terminology. Cognitive science and related fields typically use the term "learning" for the process of gaining information through observation - hence the name "learning theory". To most cognitive scientists, the term "learning theory" suggests the empirical study of human and animal learning stemming from the behaviourist paradigm in psychology. The epithet "formal" distinguishes the subject of this entry from behaviourist learning theory. Because many developments in, and applications of, formal learning theory come from computer science, the term "computational learning theory" is also common. Philosophical terms for learning-theoretic epistemology include "logical reliability" (Kelly [1996], Glymour [1991]) and "means-ends epistemology" (Schulte [1999a]).

This entry focuses on the philosophical ideas and insights behind learning theory. It eschews theorems and definitions in favour of examples and informal arguments. Those interested in the mathematical substance of learning theory will find some references in the Bibliography, and a summary of the basic definitions in the Supplementary Document.

Philosophical characteristics. We can categorize normative epistemologies according to two criteria: (1) what are the objects of normative evaluation, and (2) what are the evaluation criteria to be employed? In learning theory, the basic object of normative evaluation is an inquirer's disposition to form beliefs given some evidence. The normative question that drives the theory is whether a given doxastic disposition serves the goals of inquiry or not. Most of learning theory examines which investigative strategies reliably lead to correct beliefs about the world.

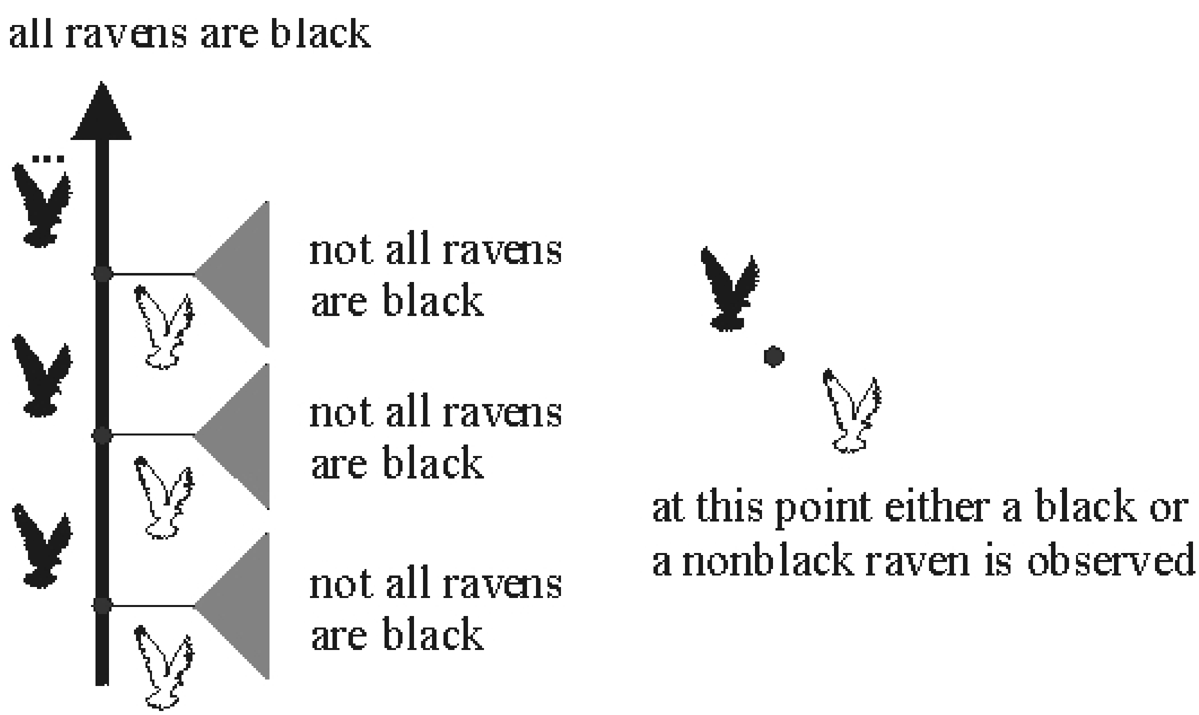
Overview. A number of examples will illustrate how means-ends analysis can lead us to endorse some ways of drawing inductive inferences and to reject others. Then I outline some of the general insights that lie behind such examples. Finally I relate means-ends epistemology to some other traditions in inductive epistemology.

1. Some Basic Examples

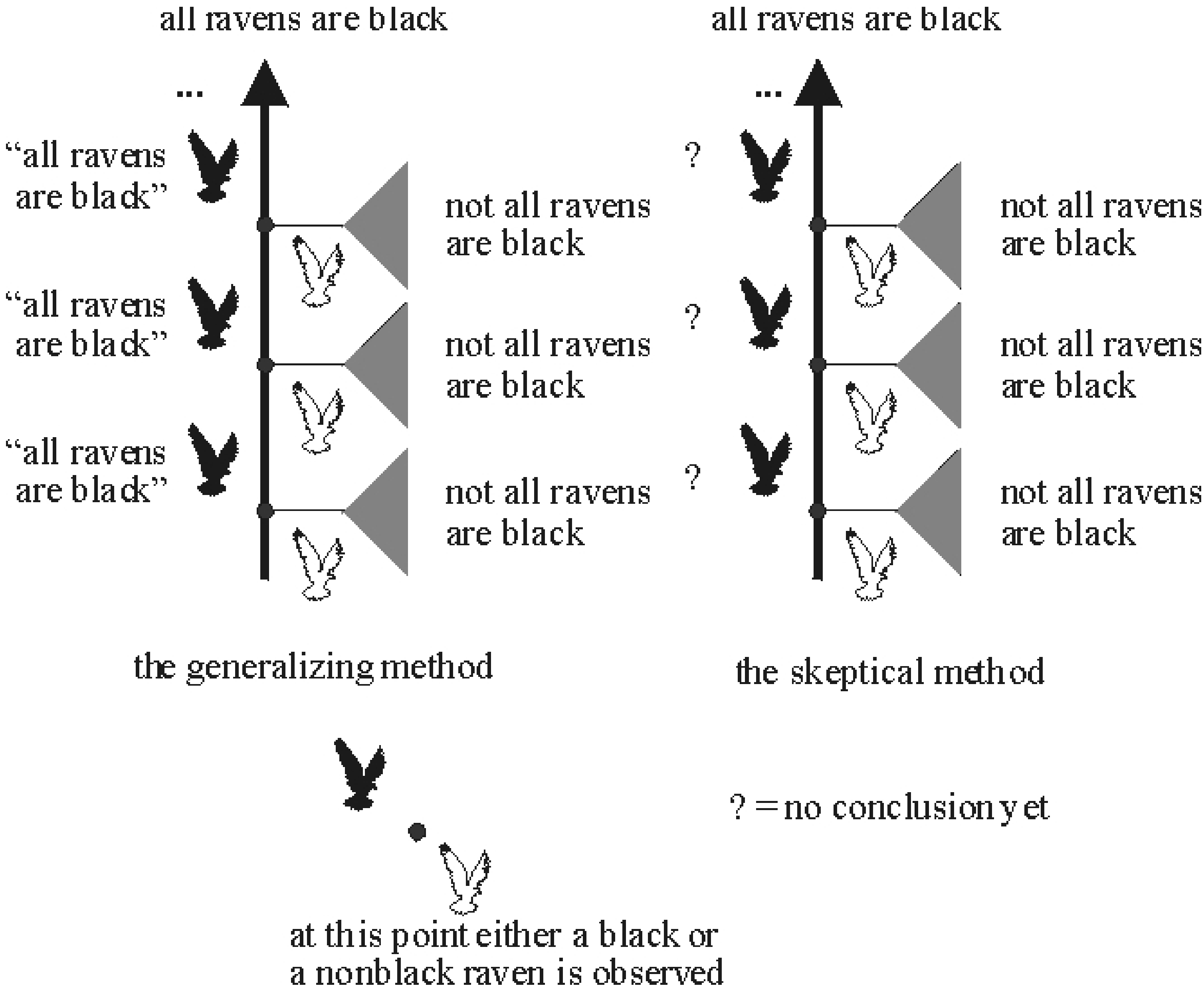
Learning-theoretic analysis assesses doxastic dispositions. Several terms for doxastic dispositions are in common use in philosophy; I will use "inductive strategy", "inference method" and most frequently "inductive method" to mean the same thing. The best way to understand how learning theory evaluates inductive methods is to work through some examples. The following presentation begins with some very simple inductive problems and moves on to more complicated -- and more realistic -- settings.

Universal Generalization

Let’s revisit the classic question of whether all ravens are black. Imagine an ornithologist who tackles this problem by examining one raven after another. There is exactly one observation sequence in which only black ravens are found; all others feature at least one nonblack raven. The figure below illustrates the possible observation sequences. Dots in the figure denote points at which an observation may be made. A black bird to the right of a dot indicates that at this stage, a black raven is observed; similarly for a white bird to the left of a dot. Given a complete sequence of observations, either all observed ravens are black or not; the figure labels complete observation sequences with the statement that is true of them. The gray fan indicates that after the observation of a white raven, the claim that not all ravens are black holds on all observation sequences resulting from further observations.



If the world is such that only black ravens are found, we would like the ornithologist to settle on this generalization. (It may be possible that some nonblack ravens remain forever hidden from sight, but even then the generalization "all ravens are black" at least gets the observations right.) If the world is such that eventually a nonblack raven is found, then we would like the ornithologist to arrive at the conclusion that not all ravens are black. This specifies a set of goals of inquiry. For any given inductive method that might represent the ornithologist’s disposition to adopt conjectures in the light of the evidence, we can ask whether that method measures up to these goals or not. There are infinitely many possible methods to consider; we’ll look at just two, a sceptical one and one that boldly generalizes. The bold method conjectures that all ravens are black after seeing that the first raven is black. It hangs on to this conjecture unless some nonblack raven appears. The sceptical method does not go beyond what is entailed by the evidence. So if a nonblack raven is found, the skeptical method concludes that not all ravens are black, but otherwise the method does not make a conjecture one way or another. The figure below illustrates both the generalizing and the sceptical method.



Do these methods attain the goals we set out? Consider the bold method. There are two possibilities: either all observed ravens are black, or some nonblack raven is found. In the first case, the method conjectures that all ravens are black and *never abandons this conjecture*. In the second case, the method concludes that not all ravens are black as soon as the first nonblack raven is found. Hence *no matter how* the evidence comes in, eventually the method gives the right answer as to whether all ravens are black and *sticks* with this answer. Learning theorists call such methods *reliable* because they settle on the right answer no matter what observations the world provides.

The skeptical method does not measure up so well. If a nonblack raven appears, then the method does arrive at the correct conclusion that not all ravens are black. But if all ravens are black, the skeptic never takes an "inductive leap" to adopt this generalization. So in that case, the skeptic fails to provide the right answer to the question of whether all ravens are black.

This illustrates how means-ends analysis can evaluate methods: the bold method meets the goal of reliably arriving at the right answer, whereas the skeptical method does not. Note the character of this argument against the skeptic: The problem, in this view, is not that the skeptic violates some canon of rationality, or fails to appreciate the "uniformity of nature". The learning-theoretic analysis concedes to the skeptic that no matter how many black ravens have been observed in the past, the next one could be white. The issue is that if all observed ravens are indeed black, then the skeptic *never answers* the question "are all ravens black?". Getting the right answer to that question requires generalizing from the evidence *even though* the

generalization could be wrong.

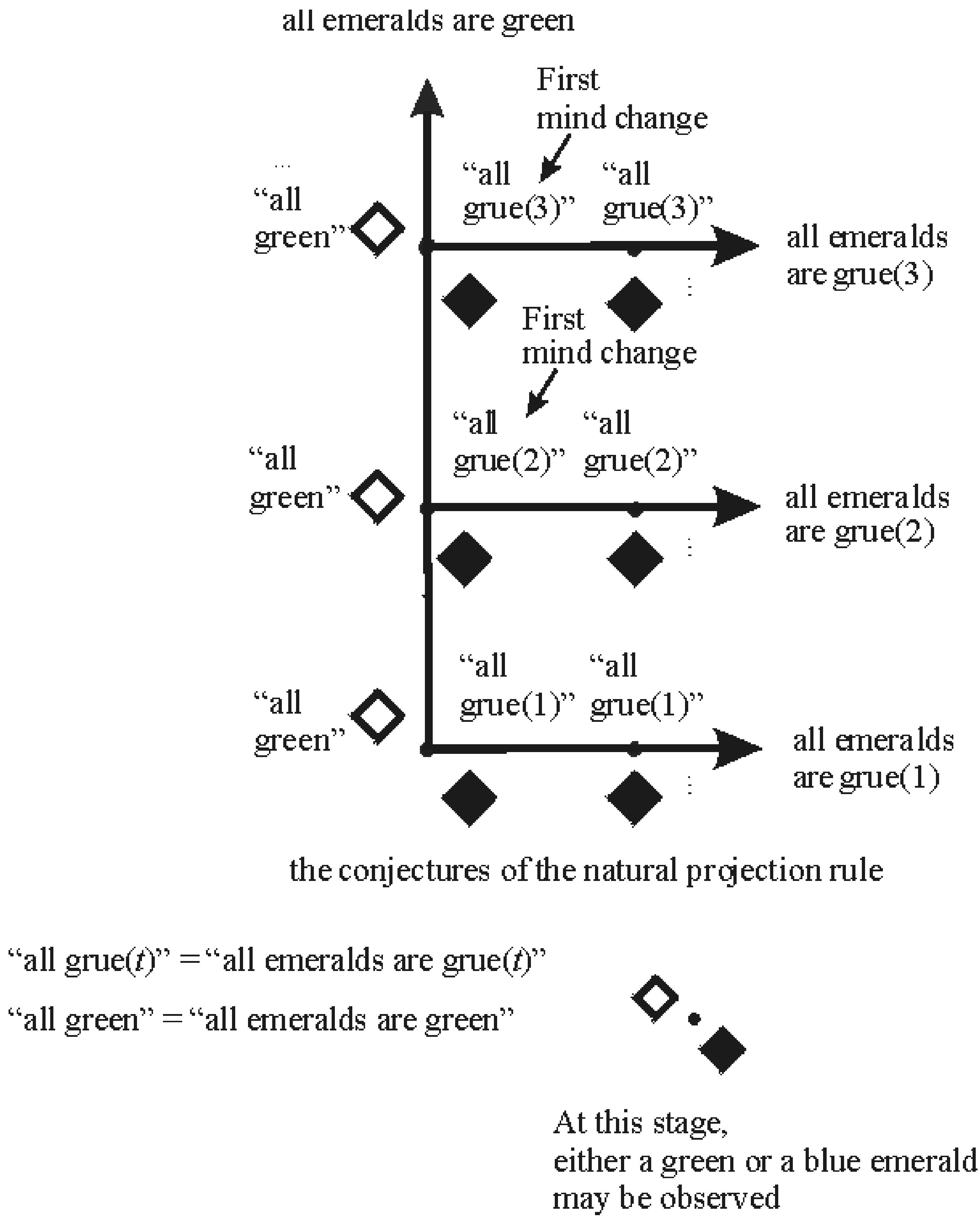
As for the bold method, it’s important to be clear on what it does and does not achieve. The method will eventually settle on the right answer -- but it (or we) may never *be certain* that it has done so. As William James put it, "no bell tolls" when science has found the right answer. We are certain that the method will eventually settle on the right answer; but we may never be certain that the current answer is the right one. This is a subtle point. The next example illustrates this point further.

A Riddle of Induction

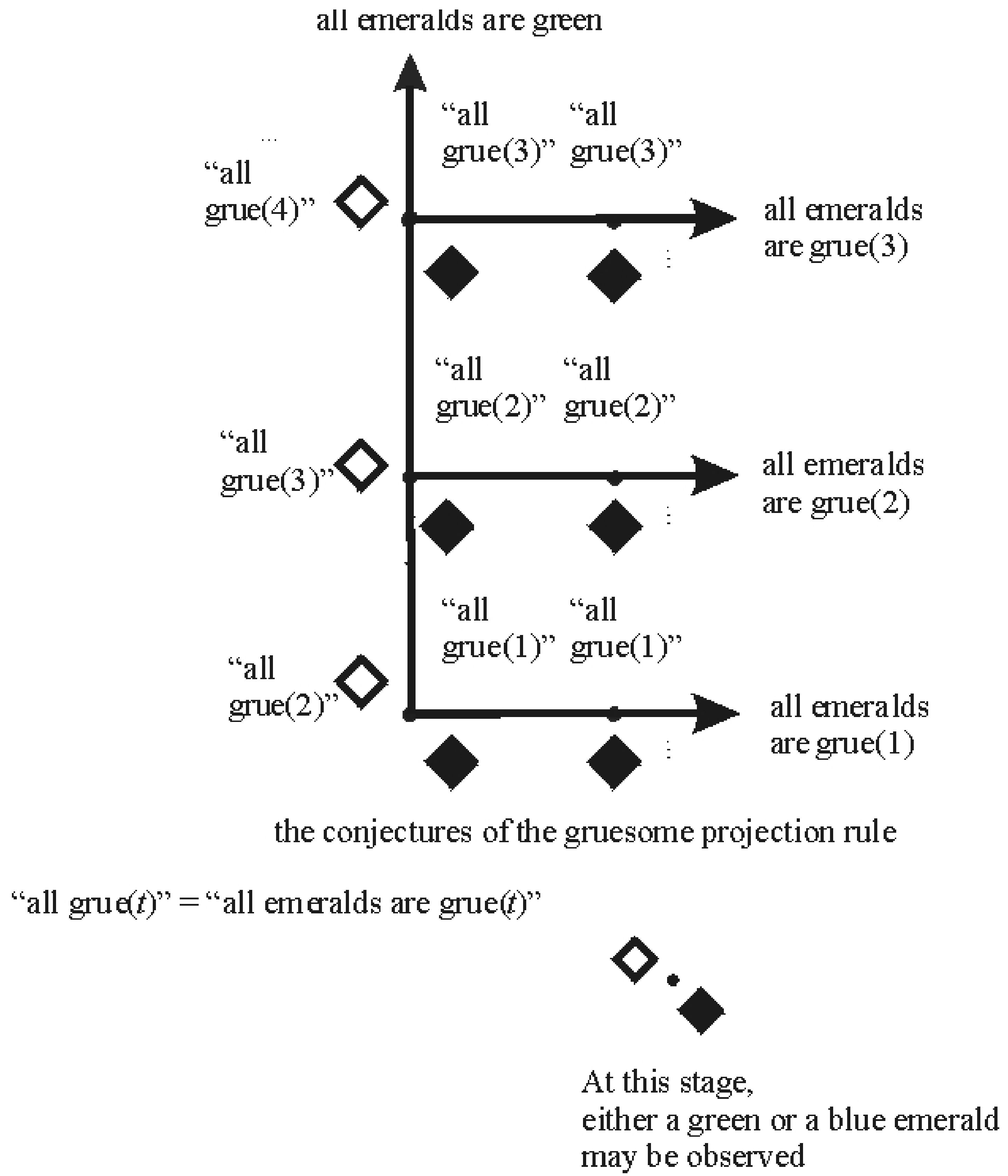
Nelson Goodman posed a famous puzzle about inductive inference known as the (New) Riddle of Induction ([Goodman 1983]). Our next example is inspired by his puzzle. Goodman considered generalizations about emeralds, involving the familiar colours of green and blue, as well as certain unusual ones:

> Suppose that all emeralds examined before a certain time *t* are green ... Our evidence statements assert that emerald *a* is green, that emerald *b* is green, and so on...
>
> Now let us introduce another predicate less familiar than "green’’. It is the predicate "grue’’ and it applies to all things examined before *t* just in case they are green but to other things just in case they are blue. Then at time *t* we have, for each evidence statement asserting that a given emerald is green, a parallel evidence statement asserting that emerald is grue.
>
> The question is whether we should conjecture that all emeralds are green rather than that all emeralds are grue when we obtain a sample of green emeralds examined before time *t*, and if so, why.

Clearly we have a family of grue predicates in this problem, corresponding to different "critical times" *t*; let’s write grue(*t*) to denote these. Following Goodman, let us refer to methods as projection rules in discussing this example. A projection rule succeeds in a world just in case it settles on a generalization that is correct in that world. Thus in a world in which all examined emeralds are found to be green, we want our projection rule to converge to the proposition that all emeralds are green. If all examined emeralds are grue(*t*), we want our projection rule to converge to the proposition that all emeralds are grue(*t*). Note that this stipulation treats green and grue predicates completely on a par, with no bias towards either. As before, let us consider two rules: the "natural" projection rule which conjectures that all emeralds are green as long as only green emeralds are found, and the "gruesome" rule which keeps projecting the next grue predicate consistent with the available evidence. Expressed in the green-blue vocabulary, the gruesome projection rule conjectures that after observing some number of *n* green emeralds, all future ones will be blue. The figure below illustrates the possible observation sequences and the natural projection rule in this model of the New Riddle of Induction.



The following figure shows the gruesome projection rule.



How do these rules measure up to the goal of arriving at a true generalization? Suppose for the sake of the example that the only serious possibilities under consideration are that either all emeralds are green or that all emeralds are grue(t) for some critical time t. Then the natural projection rule settles on the correct generalization no matter what the correct generalization is. For if all emeralds are green, the natural projection rule asserts this fact from the beginning. And suppose that all emeralds are grue(t) for some critical time t. Then at time t, a blue emerald will be observed. At this point the natural projection rule settles on the conjecture that all emeralds are grue(t), which must be correct given our assumption about the possible

observation sequences. Thus no matter what evidence is obtained in the course of inquiry -- consistent with our background assumptions -- the natural projection rule eventually settles on a correct generalization about the colour of emeralds.

The gruesome rule does not do as well. For if all emeralds are green, the rule will never conjecture this fact because it keeps projecting grue predicates. Hence there is a possible observation sequence -- namely those on which all emeralds are green -- on which the gruesome rule fails to converge to the right generalization. So means-ends analysis would recommend the natural projection rule over the gruesome rule. Some comments are in order.

1. As in the previous example, nothing in this argument hinges on arguments to the effect that certain possibilities are not to be taken seriously a priori. In particular, nothing in the argument says that generalizations with grue predicates are ill-formed, unlawlike, or in some other way a priori inferior to "all emeralds are green".

2. The analysis does not depend on the vocabulary in which the evidence and generalizations are framed. For ease of exposition, I have mostly used the green-blue reference frame. However, grue-bleen speakers would agree that the aim of reliably settling on a correct generalization requires the natural projection rule rather than the gruesome one, even if they would want to express the conjectures of the natural rule in their grue-bleen language rather than the blue-green language that we have used so far. (For more on the language-invariance of means-ends analysis see Section 4 (The Limits of Inquiry and the Complexity of Empirical Problems), as well as Schulte [1999a, 1999b]).

3. Though the analysis does not depend on language, it does depend on assumptions about what the possible observation sequences are. The example as described above seems to comprise the possibilities that correspond to the colour predicates Goodman himself discussed. But means-ends analysis applies just as much to other sets of possible predicates. Schulte [1999a, 1999b] and Chart [2000] discuss a number of other versions of the Riddle of Induction, in some of which means-ends analysis favours projecting that all emeralds are grue on a sample of all green emeralds.

4. Even with the assumptions granted so far, there are reliable projection rules that project that all emeralds are grue(t) on a sample of all green emeralds. For example, the projection rule "conjecture that all emeralds are grue(3) until 3 green emeralds are observed; then conjecture that all emeralds are green until a blue emerald is observed" is guaranteed to eventually settle on a correct generalization just like the natural projection rule. (It’s a worthwhile exercise to verify the reliability of this rule.) I will discuss criteria for further restricting the space of rules in Section 3 (The Long Run in The Short Run).

Generalizations with Exceptions

Let’s return to the world of ravens. This time the ornithological community is more guarded in its generalizations concerning the colour of ravens. Two competing hypotheses are under investigation: (1) That basically all ravens are black, but there may be a finite number of exceptions to that rule, and (2) that basically all ravens are white, but there may be a finite number of exceptions to that rule. Assuming that one or the other of these hypotheses is correct, is there an inductive method that reliably settles on the right one? What makes this problem more difficult than our first two is that each hypothesis under investigation is consistent with any finite amount of evidence. If 100 white ravens and 50 black ravens are found, either the 50 black ravens or the 100 white ravens may be the exception to the rule. In terminology made familiar by Karl Popper’s work, we may say that neither hypothesis is falsifiable. As a consequence, the inductive strategy from the previous two examples will not work here. This strategy was basically to adopt a "bold" universal generalization, such as "all ravens are black" or "all emeralds are green", and to hang on to this conjecture as long as it "passes muster". However, when rules with possible exceptions are under investigation, this strategy is unreliable. For example, suppose that an inquirer first adopts the hypothesis that "all but finitely many ravens are white". It may be the case that from then on, only black ravens are found. But each of these apparent counterinstances can be "explained away" as an exception. If the inquirer follows the principle of hanging on to her conjecture until the evidence is logically inconsistent with the conjecture, she will never abandon her false belief that all but finitely many ravens are white, much less arrive at the correct belief that all but finitely many ravens are black.

Reliable inquiry requires a more subtle investigative strategy. Here is one (of many). Begin inquiry with either competing hypothesis, say "all but finitely many ravens are black". Choose some cut-off ratio to represent a "clear majority"; for definiteness, let's say 70%. If the current conjecture is that all but finitely many ravens are black, change your mind to conjecture that all but finitely many ravens are white just in case over 70% of observed ravens are in fact white. Proceed likewise if the current conjecture is that all but finitely many ravens are white when over 70% of observed ravens are in fact black.

A bit of thought shows that this rule reliably identifies the correct hypothesis in the long run, no matter which of the two competing hypotheses is correct. For if all but finitely many ravens are black, eventually the nonblack exceptions to the rule will be exhausted, and an arbitrarily large majority of observed ravens will be black. Similarly if all but finitely many ravens are white.

The way in which this reliable method is sensitive to the frequency of occurrences of black resp. white ravens is reminiscent of statistical methods. This suggests considering statistical generalizations such as "the percentage of white ravens in the total population of ravens is 20%". Hans Reichenbach held that the central aim of inductive inference were generalizations of precisely that sort. More generally, we may consider hypotheses about proportions such as "the proportion of white ravens is between 10% and 20%". Kelly examines in detail various hypotheses of this kind and establishes when there are reliable methods for testing them. He also provides a learning-theoretic interpretation of classical statistical tests of a parameter for a probability distribution [Kelly 1996, Ch.3.4].

Generalizations with exceptions illustrate some subtle nuances in the relationship between Popperian falsificationism and the learning-theoretic idea of reliable convergence to the truth. In some settings of inquiry, notably those involving universal generalizations, a naively Popperian "conjectures-and-refutations" approach of hanging on to conjectures until the evidence falsifies them does yield a reliable inductive method. In other problems, like the current example, it does not. Generally speaking problems with unfalsifiable hypotheses require something other than the conjectures-and-refutations recipe for reliable methods (this assertion hinges on what exactly one means by "falsifiable hypothesis"; see Section 4 (The Limits of Inquiry and the Complexity of Empirical Problems) as well as [Schulte and Juhl 1996]). A Popperian might respond that such hypotheses are "unscientific" and hence it is no concern that the conjectures-and-refutations approach fails to reliably identify a correct hypothesis when unfalsifiable hypotheses are involved. But intuitively, a claim like "all but finitely many ravens are black" appears to be a respectable empirical hypothesis. More importantly than intuition, at least from the point of view of means-ends epistemology, it is possible to reliably assess the truth of such claims in the long run, even though all hypotheses under investigation are consistent with any finite amount of evidence. This constitutes a clear sense in which inquiry can test these hypotheses against empirical evidence in order to find the truth. If we allow that we would want inductive methodology to extend to problems like this one, the moral is that relying on falsifications is sometimes, but *not always*, the best way for inquiry to proceed.

2. Case Studies in Scientific Practice

This section provides further examples to illustrate learning-theoretic analysis. The examples in this section are more realistic and address methodological issues arising in scientific practice. The space constraints of the encyclopedia format allow only an outline of the full analysis; there are references to more detailed discussions below.

Conservation Laws in Particle Physics

One of the hallmarks of elementary particle physics is the discovery of new conservation laws that apply only in the subatomic realm [Ford 1963, Ne'eman and Kirsh 1983, Feynman 1965]. (Feynman groups one of them, the conservation of Baryon Number, with the other "great conservation laws" of energy, charge and momentum.) Simplifying somewhat, conservation principles serve to explain why certain processes involving elementary particles do not occur: the explanation is

that some conservation principle was violated (cf. Omnes [1971, Ch.2]). So a goal of particle inquiry is to find a set of conservation principles, such that for every process that is possible according to the (already known) laws of physics, there is some conservation principle that rules out that process. And if a process is in fact observed to occur, then it ought to satisfy all conservation laws that we have introduced.

This constitutes an inference problem to which we may apply means-ends analysis. An inference method produces a set of conservation principles in response to reports of observed processes. Means-ends analysis asks which methods are guaranteed to settle on conservation principles that account for all observations, that is, that rule out unobserved processes and allow observed processes. [Schulte 2000] describes an inductive method that accomplishes this goal. Interestingly, the conservation principles that this method would posit on the currently available evidence appear to be close to the ones that physicists have introduced.

It turns out that for some physical processes, the only way to get empirically adequate conservation principles is by positing that some hidden particles have gone undetected. It is remarkable that to find conservation principles that are consistent with what is observed, sometimes the only option is to form hypotheses about what is *un*observed. It is easy to miss this phenomenon without the scrutiny that means-ends analysis requires. Extending our problem so that inference methods not only posit conservation laws but also hidden particles makes inquiry more difficult. But if we grant the particle theorist the assumption that there are only finitely many types of hidden particles, then there is a still a method that is guaranteed to settle eventually on a theory that makes correct predictions about the observable phenomena, using a combination of conservation laws and hidden particles. A detailed discussion of the conservation law inference problem is in [Schulte 2000].

Models of Cognitive Architecture

Some philosophers of mind have argued that the mind is composed of fairly independent modules. Each module has its own "input" from other modules and sends "output" to other modules. For example, an "auditory analysis system" module might take as input a heard word and send a phonetic analysis to an “auditory input lexicon”. The idea of modular organization raises the empirical question of what mental modules there are and how they are linked to each other. A prominent tradition of research in cognitive neuroscience has attempted to develop a model of mental architecture along these lines by studying the responses of normal and abnormal subjects to various stimuli. The idea is to compare normal reactions with abnormal ones -- often caused by brain damage -- so as to draw inferences about which mental capacities depend on each other and how.

Glymour [1994] asked the reliabilist question whether there are inference methods that are guaranteed to eventually settle on a true theory of mental organization, given exhaustive evidence about normal and abnormal capacities and reactions. He argued that for some possible mental architectures, no amount of evidence of the stimulus-response kind can distinguish between them. Since the available evidence determines the conjectures of an inductive method, it follows that there is no guarantee that a method will settle on the true model of cognitive architecture.

In further discussion, Bub [1994] showed that if we grant certain restrictive assumptions about how mental modules are connected, then a complete set of behavioural observations would allow a neuropsychologist to ascertain the module structure of a (normal) mind. In fact, under Bub’s assumptions there is a reliable method for identifying the modular structure. Glymour has also explored to what extent richer kinds of evidence would resolve underdetermination of mental architecture by behavioural evidence. (One example of richer evidence are double disassocations. An example of a double dissocation would be a pair of patients, one who has a normal capacity for understanding spoken words, but fails to understand written ones, and another who understands written words but not spoken ones.)

Some more case studies of this sort may be found in [Kelly 1996, Ch. 7.7, Harell 2000]. The main interest of such studies lies of course in what they say about the particular domain under investigation. But they also illustrate some general features of learning theory:

1. *Generality*. The basic notions of the theory are very general. Essentially, the theory applies whenever one has a question that prompts inquiry, a number of candidate answers, and some evidence for deciding among the answers. Thus means-ends analysis can be applied in any discipline aimed at empirical knowledge, for example physics or psychology.

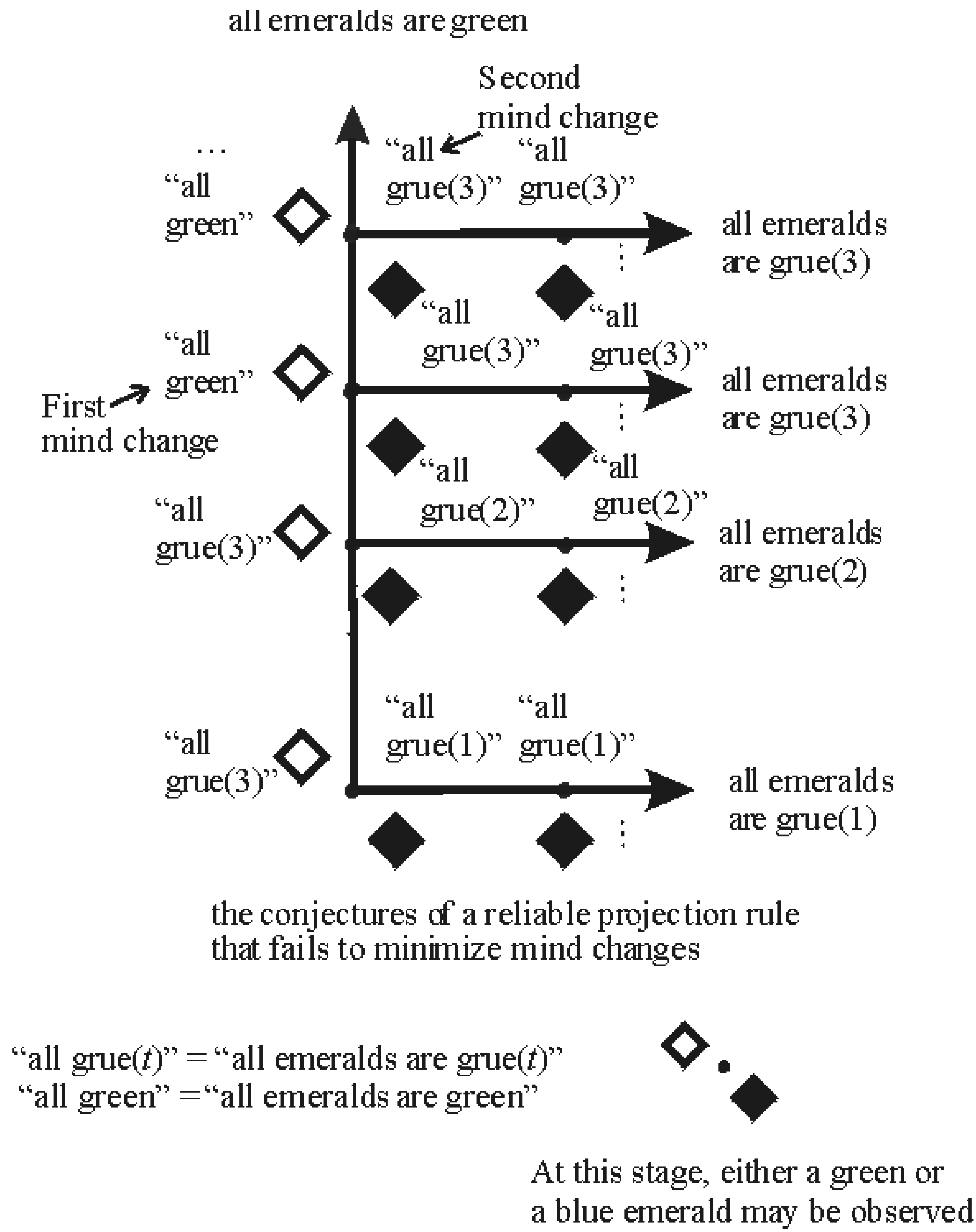
2. *Context Dependence*. Learning theory is pure normative a priori epistemology, in the sense that it deals with standards for assessing methods in possible settings of inquiry. But the approach does not aim for universal, context-free methodological maxims. The methodological recommendations depend on contingent factors, such as the operative methodological norms, the questions under investigation, the background assumptions that the agent brings to inquiry, the observational means at her disposal, her cognitive capacities, and her epistemic aims. As a consequence, to evaluate specific methods in a given domain, as in the case studies mentioned, one has to study the details of the case in question. The means-ends analysis often rewards this study by pointing out what the crucial methodological features of a given scientific enterprise are, and by explaining precisely why and how these features are connected to the success of the enterprise in attaining its epistemic aims.

3. *Trade-offs*. In the perspective of means-ends epistemology, inquiry involves an ongoing struggle with hard choices, rather than the execution of a universal "scientific method". The inquirer has to balance conflicting values, and may consider various strategies such as accepting difficulties in the short run hoping to resolve them in the long run. For example in the conservation law problem, there can be conflicts between theoretical parsimony, i.e., positing fewer conservation laws, and ontological parsimony, i.e., introducing fewer hidden particles. For another example, a particle theorist may accept positing undetected neutrinos in the hopes that they will eventually be observed as science progresses. An important learning-theoretic project is to examine when such tradeoffs arise and what the options for resolving them are. The next section deals with some aspects of this topic.

3. The Long Run in the Short Run

A longstanding criticism of convergence to the truth as an aim of inquiry is that, while fine in itself, this aim is consistent with any crazy behaviour in the short run [Salmon 1991]. For example, we saw in the New Riddle of Induction that a reliable projection rule can conjecture that the next emerald will be blue no matter how many green emeralds have been found -- as long as *eventually* the rule projects "all emeralds are green". One response is that if means-ends analysis takes into account other epistemic aims *in addition* to long-run convergence, then it *can* provide strong guidance for what to conjecture in the short run.

To illustrate this point, let us return to the Goodmanian Riddle of Induction. Since Plato, philosophers have considered the idea that *stable* true belief is better than unstable true belief, and epistemologists such as Sklar [1975] have advocated similar principles of "epistemic conservatism". Kuhn tells us that a major reason for conservatism in paradigm debates is the cost of changing scientific beliefs [Kuhn 1970]. In this spirit, learning theorists have examined methods that minimize the number of times that they change their theories before settling on their final conjecture. Such methods are said to minimize *mind changes*. The New Riddle of Induction turns out to be a nice illustration of this idea. Consider the natural projection rule (conjecture that all emeralds are green on a sample of green emeralds). If all emeralds are green, this rule never changes its conjecture. And if all emeralds are grue(*t*) for some critical time *t*, then the natural projection rule abandons its conjecture "all emeralds are green" at time *t* -- one mind change -- and thereafter correctly projects "all emeralds are grue(*t*)". Remarkably, rules that project grue rather than green do not do as well. For example, consider a rule that conjectures that all emeralds are grue(3) after observing one green emerald. If two more green emeralds are observed, the rule's conjecture is falsified and it must eventually change its mind, say to conjecture that all emeralds are green (suppose that green emeralds continue to be found). But then at that point, a blue emerald may appear, forcing a second mind change. This argument can be generalized to show that the aim of minimizing mind changes allows only the green predicate to be projected on a sample of all green emeralds [Schulte 1999a]. We saw in a figure above how the natural projection rule changes its mind at most once; the figure below illustrates in a typical case how an unnatural projection rule may have to change its mind twice or more.



The conservation law problem discussed in the previous section provides another illustration of how additional epistemic aims can lead to constraints in the short run. In this problem, reliability and minimizing mind changes require a particle theorist to adopt a conservation theory that rules out as many unobserved reactions as possible. In certain circumstances, the only way to attain this aim is to introduce hidden particles. Long-run reliability by itself may require that the theorist introduce hidden particles *eventually*; the additional goal of minimizing mind changes -- stable belief -- dictates exactly when this should occur [Schulte 2000].

Learning theorists have examined other epistemic aims, such as *fast convergence* to a right answer and *avoiding errors* before settling on the truth. Some of these desiderata come into conflict whereas others turn out to stand and fall together, often in surprising ways. For example, Kelly has shown that under very general circumstances, minimizing mind changes and minimizing errors vindicate the same inductive methods [Kelly 2001].

4. The Limits of Inquiry and the Complexity of Empirical Problems

After seeing a number of examples like the ones above, one begins to wonder what the pattern is. What is it about an empirical question that allows inquiry to reliably arrive at the correct answer? What general insights can we gain into how reliable methods go about testing hypotheses? Learning theorists answer these questions with *characterization theorems*. Characterization theorems are generally of the form “it is possible to attain this standard of empirical success in a given inductive problem if and only if the inductive problem meets the following conditions”. There are a number of standards of success in inquiry -- reliable convergence to the truth, fast reliable convergence, reliable convergence with few mind changes -- and hence correspondingly many characterization theorems.

A fundamental result describes the conditions under which a method can reliably find the correct hypothesis among a countably infinite or finite number H_1, H_2, …, H_n, …. of mutually exclusive hypotheses that jointly cover all possibilities consistent with the inquirer’s background assumptions. This is possible just in case each of the hypotheses is a countable disjunction of refutable empirical claims. By “refutable” I mean that if the claim is false, the evidence combined with the inquirer’s background assumptions will eventually conclusively falsify the hypothesis (see Schulte and Juhl [1996], Kelly [1996, Ch. 3.3]). For illustration, let’s return to the ornithological example with two alternative hypotheses: (1) all but finitely many swans are white, and (2) all but finitely many swans are black. As we saw, it is possible in the long run to reliably settle which of these two hypotheses is correct. Hence by the characterization theorem, each of the two hypotheses must be a disjunction of refutable empirical claims. To see that this indeed is so, observe that “all but finitely many swans are white” is logically equivalent to the disjunction

“at most 1 swan is black or at most 2 swans are black … or at most n swans are black … or…”,

and similarly for “all but finitely many swans are black”. Each of the claims in the disjunction is refutable, in the sense of being eventually falsified whenever it is false. For example, take the claim that “at most 3 swans are black”. If this is false, more than 3 black swans will be found, at which point the claim is conclusively falsified.

A few points will help explain the significance of characterization theorems like this one.

1. *Structure of Reliable Methods*. Characterization theorems tell us how the structure of reliable methods corresponds to the structure of the hypotheses under investigation. For example, the theorem mentioned establishes a connection between falsifiability and testability, but one that is more attenuated than the naïve Popperian envisions: it is not necessary that the hypotheses under test be directly falsifiable; rather, there must be ways of strengthening each hypothesis that yield a countable number of refutable “subhypotheses”. We can think of these refutable subhypotheses as different ways in which the main hypothesis may be true. (For example, one way in which "all but finitely many ravens are white" is true is if there are are at most 10 black ravens,; another if there are at most 100 black ravens, etc.)

2. *Import of Background Assumptions.* The characterization result draws a line between the solvable and unsolvable problems. Background knowledge reduces the inductive complexity of a problem; with enough background knowledge, the problem crosses the threshold between the unsolvable and the solvable. In many domains of empirical inquiry, the pivotal background assumptions are those that make reliable inquiry feasible. (Kuhn [1970] makes similar points). For example, in the particle dynamics problem, it is the assumption that some set of linear conservation laws is empirically adequate that

permits us to reliably find an empirically adequate theory of particle reactions. It seems that, conversely, in domains in which the available background assumptions do not reduce the complexity of the empirical problems enough, there is a sense that inquiry is not sufficiently constrained for steady progress. One example might be Chomsky's program of universal grammar whose inductive complexity hinges on how broad we assume the set of humanly learnable languages to be. It is an interesting question how much the complexity of an empirical investigation, as determined by the strength of the available background assumptions, connects with the practitioners' sense of progress and feasibility. In the domains considered so far, learning-theoretic complexity corresponds quite well to intuitive methodological difficulty, but more case studies are required. In any case, learning-theoretic characterization theorems help us to pinpoint the sources of inductive complexity, and thus the methodologically central assumptions and the weak spots in a given empirical enterprise.

3. *Universal Measure of Inductive Complexity*. There are characterization theorems for a number of standards of empirical success. The more demanding the standard, the more stringent the conditions that such standards require. It turns out that a number of natural standards of inductive success fall into a hierarchy of feasibility, in the sense that standards higher in the hierarchy are attainable if standards lower in the hierarchy are. For a trivial example, if it is possible to settle on a correct hypothesis with at most 5 mind changes, then a fortiori it is possible to succeed with 10 mind changes. For other cognitive aims the inclusion is more surprising; see Schulte [1999a]. We may think of the hierarchy of standards of empirical success as establishing a scale for inductive problems: The more difficult the problem, the less we can expect from inquiry. The hierarchy of empirical success allows us to weigh such diverse problems as Goodman's Riddle, particle dynamics and language learning on the same scale. For example, it should be clear that the characteristic condition for reliable inquiry -- - that each hypothesis be equivalent to a countable disjunction of refutable assertions -- applies in every domain of investigation. Thus means-ends analysis uncovers common structure among seemingly disparate problems.

4. *Language Invariance*. Learning-theoretic characterization theorems concern what Kelly calls the "temporal entanglement" of various observation sequences [Kelly 2000] (see also [Schulte and Juhl 1996]). Ultimately they rest on entailment relations between given evidence, background assumptions and empirical claims. Since logical entailment does not depend on the language we use to frame evidence and hypotheses, the inductive complexity of an empirical problem as determined by the characterization theorems is language-invariant. Indeed, it turns out that inductive complexity can be captured in terms of point-set topology and corresponds to a scale of topological complexity that has much importance in mathematics (the Borel and finite-difference hierarchies [Kelly 1996]).

5. Categorical vs. Hypothetical Imperatives

Kant distinguished between categorical imperatives that one ought to follow regardless of one's personal aim and circumstances, and hypothetical imperatives that direct us to employ our means towards our chosen end. One way to think of learning theory is as the study of hypothetical imperatives for empirical inquiry. Many epistemologists have proposed various categorical imperatives for inductive inquiry, for example in the form of an "inductive logic" or norms of "epistemic rationality". In principle, there are three possible relationships between hypothetical and categorical imperatives for empirical inquiry.

1. The categorical imperative will lead an inquirer to obtain his cognitive goals. In that case means-ends analysis *vindicates* the categorical imperative. For example, when faced with a simple universal generalization such as "all ravens are black", we saw above that following the Popperian recipe of adopting the falsifiable generalization and sticking to it until a counter example appears leads to a reliable method.

2. The categorical imperative may *prevent* an inquirer from achieving his aims. In that case the categorical imperative *restricts* the scope of inquiry. For example, in the case of the two alternative generalizations with exceptions, the principle of maintaining a universal generalization until it is falsified leads to an unreliable method (cf. [Kelly 1996, Ch. 9.4]).

3. Some methods meet *both* the categorical imperative and the goals of inquiry, and others don't. Then we may take the best of both worlds and choose those methods that attain the goals of inquiry and satisfy categorical imperatives. (See the further

discussion in this section.)

For a proposed norm of inquiry, we can apply means-ends analysis to ask whether the norm helps or hinders the aims of inquiry. This was the spirit of Putnam's critique of Carnap's confirmation functions [Putnam 1963]: the thrust of his essay was that Carnap's methods were not as reliable in detecting general patterns as other methods would be. More recently, learning theorists have investigated the power of Bayesian conditioning (see the entry on Bayesian epistemology). John Earman has conjectured that if there is any reliable method for a given problem, then there is a reliable method that proceeds by Bayesian updating [Earman 1992, Ch.9, Sec.6]. Cory Juhl [1997] provided a partial confirmation of Earman's conjecture: He proved that it holds when there are only two potential evidence items (e.g., "emerald is green" vs. "emerald is blue"). The general case is still open.

Epistemic conservatism is a methodological norm that has been prominent in philosophy at least since Quine's notion of "minimal mutilation" of our beliefs [1951]. One version of epistemic conservatism, as we saw above, holds that inquiry should seek stable belief. Another formulation, closer to Quine's, is the general precept that belief changes in light of new evidence should be minimal. Fairly recent work in philosophical logic has proposed a number of criteria for *minimal belief change* known as the AGM axioms [Gärdenfors 1988]. Learning theorists have shown that whenever there is a reliable method for investigating an empirical question, there is one that proceeds via minimal changes (as defined by the AGM postulates). The properties of reliable inquiry with minimal belief changes are investigated in [Kelly et al. 1995, Martin and Osherson 1998, Kelly 1999].

Much of computational learning theory focuses on inquirers with *bounded rationality*, that is, agents with cognitive limitations such as a finite memory or bounded computational capacities. Many categorical norms that do not interfere with empirical success for logically omniscient agents nonetheless limit the scope of cognitively bounded agents. For example, consider the norm of consistency: Believe that a hypothesis is false as soon as the evidence is logically inconsistent with it. The consistency principle is part of both Bayesian confirmation theory and AGM belief revision. Kelly and Schulte [1995] show that consistency prevents even agents with infinitely uncomputable cognitive powers from reliably assessing certain hypotheses. The moral is that if a theory is sufficiently complex, agents who are not logically omniscient may be unable to determine immediately whether a given piece of evidence is consistent with the theory, and need to collect more data to detect the inconsistency. But the consistency principle -- and a fortiori, Bayesian updating and AGM belief revision -- rule out this kind of scientific strategy.

More reflection on this and other philosophical issues in means-ends epistemology can be found in sources such as [Glymour 1991], [Kelly 1996, Chs. 2,3], [Glymour and Kelly 1992], [Kelly *et al*. 1997], [Schulte and Juhl 1996], [Glymour 1994], [Bub 1994]. Of particular interest in the philosophy of science may be learning-theoretic models that accommodate historicist and relativist conceptions of inquiry, chiefly by expanding the notion of an inductive method so that methods may actively select paradigms for inquiry; for more details on this topic, see [Kelly 2000, Kelly 1996, Ch.13]. Booklength introductions to the mathematics of learning theory are [Kelly 1996, Martin and Osherson 1998, Jain et al. 1999].

Supplementary Document: Basic Formal Definitions

Bibliography

- Bub, J. [1994]: 'Testing Models of Cognition Through the Analysis of Brain-Damaged Performance', *British Journal for the Philosophy of Science*, 45, pp.837-55.
- Chart, D. [2000]: 'Schulte and Goodman's Riddle', *British Journal for the Philosophy of Science*, 51, pp.837-55.
- Earman, J. [1992]: Bayes or Bust?. Cambridge, Mass.: MIT Press.
- Feynman, R. [1965; 19th ed. 1990]: *The Character of Physical Law*, Cambridge, Mass.: MIT Press.
- Ford, K. [1963]: *The World of Elementary Particles*, New York: Blaisdell Publishing.
- Gärdenfors, P. [1988]: *Knowledge In Flux: modeling the dynamics of epistemic states*, Cambridge, Mass.: MIT Press.

- Glymour, C. [1991]: 'The Hierarchies of Knowledge and the Mathematics of Discovery', *Minds and Machines* 1, pp. 75-95.
- ----- [1994]: 'On the Methods of Cognitive Neuropsychology', *British Journal for the Philosophy of Science* 45, pp. 815-35.
- Glymour, C. and Kelly, K. [1992]: 'Thoroughly Modern Meno', in: *Inference, Explanation and Other Frustrations*, ed. John Earman, University of California Press.
- Goodman, N. [1983]. *Fact, Fiction and Forecast.* Cambridge, MA: Harvard University Press.
- Harrell, M. [2000]: *Chaos and Reliable Knowledge*, Ph.D. Thesis, University of California at San Diego.
- Jain, S. et al [1999]: *Systems That Learn* 2nd ed. Cambridge, MA: MIT Press.
- James, W. [1982]: 'The Will To Believe', in *Pragmatism*, ed. H.S. Thayer. Indianapolis: Hackett.
- Juhl, C. [1997]: 'Objectively Reliable Subjective Probabilities', *Synthese* 109, pp. 293-309.
- Kelly, K. [1996]: *The Logic of Reliable Inquiry*, Oxford: Oxford University Press.
- ----- [1999]: ' Iterated Belief Revision, Reliability, and Inductive Amnesia', *Erkenntnis* 50: 11-58.
- ----- [2000]: 'The Logic of Success', *British Journal for the Philosophy of Science* 51:4, 639-660.
- ----- [2001]: 'A Close Shave with Realism: Ockham's Razor Derived From Efficient Convergence', unpublished manuscript.
- Kelly, K., and Schulte, O. [1995]: 'The Computable Testability of Theories Making Uncomputable Predictions', *Erkenntnis* 43, pp. 29-66.
- Kelly, K., Schulte, O. and Juhl, C. [1997]: 'Learning Theory and the Philosophy of Science', *Philosophy of Science* 64, 245-67.
- Kelly, K., Schulte, O. and Hendricks, V. [1995]: 'Reliable Belief Revision'**.** *Proceedings of the XII Joint International Congress for Logic, Methodology and the Philosophy of Science*.
- Kuhn, T. [1970]: *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Martin, E. and Osherson, D. [1998]: *Elements of Scientific Inquiry*. Cambridge, MA: MIT Press.
- Ne'eman, Y. and Kirsh, Y. [1983]: *The Particle Hunters*, Cambridge: Cambridge University Press.
- Omnes, R. [1971]: *Introduction to Particle Physics*, London, New York: Wiley Interscience.
- Putnam, H. [1963]: "Degree of Confirmation' and Inductive Logic', in *The Philosophy of Rudolf Carnap*, ed. P.a. Schilpp, La Salle, Ill: Open Court.
- Quine, W.: [1951]: 'Two Dogmas of Empiricism', *Philosophical Review* 60, 20-43.
- Salmon, W. [1991]: 'Hans Reichenbach's Vindication of Induction,' *Erkenntnis* 35:99-122.
- Schulte, O. [1999a]: 'Means-Ends Epistemology', *The British Journal for the Philosophy of Science*, 50, 1-31.
- ----- [1999b]: 'The Logic of Reliable and Efficient Inquiry', *Journal of Philosophical Logic* 28, 399-438.
- ----- [2000]: 'Inferring Conservation Principles in Particle Physics: A Case Study in the Problem of Induction'**,** *The British Journal for the Philosophy of Science*, 51: 771-806.
- Schulte, O., and Juhl, C. [1996]: 'Topology as Epistemology', *The Monist* 79, 1:141-147.
- Sklar, L. [1975]: 'Methodological Conservatism', *Philosophical Review* LXXXIV, pp. 374-400.

Other Internet Resources

- Learning Theory in Computer Science
- Inductive Logic Website on Formal Learning Theory and Belief Revision

Related Entries

confirmation | epistemology: Bayesian | induction: new problem of | induction: problem of | James, William | Peirce, Charles Sanders | Popper, Karl

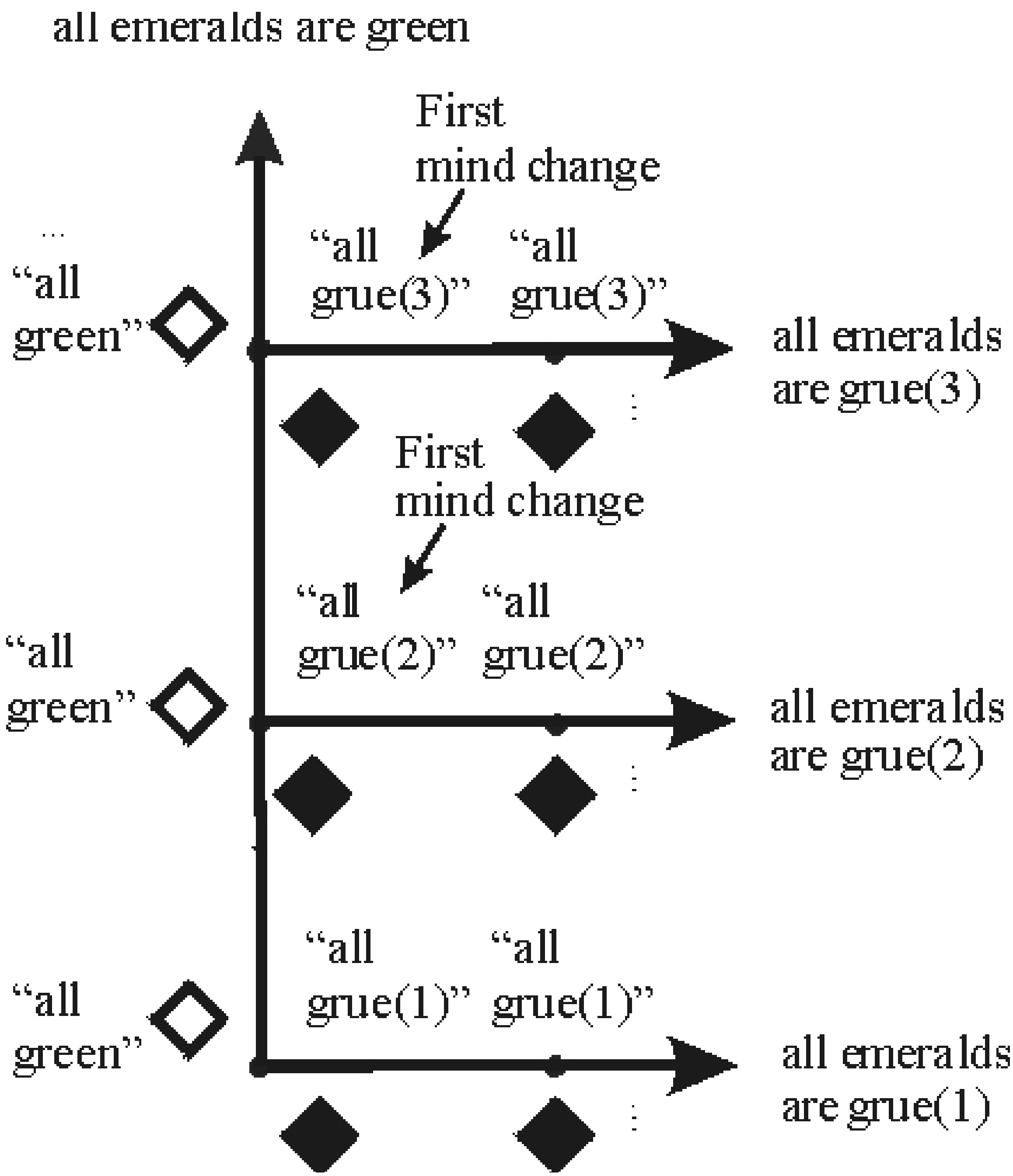
oschulte@sfu.ca

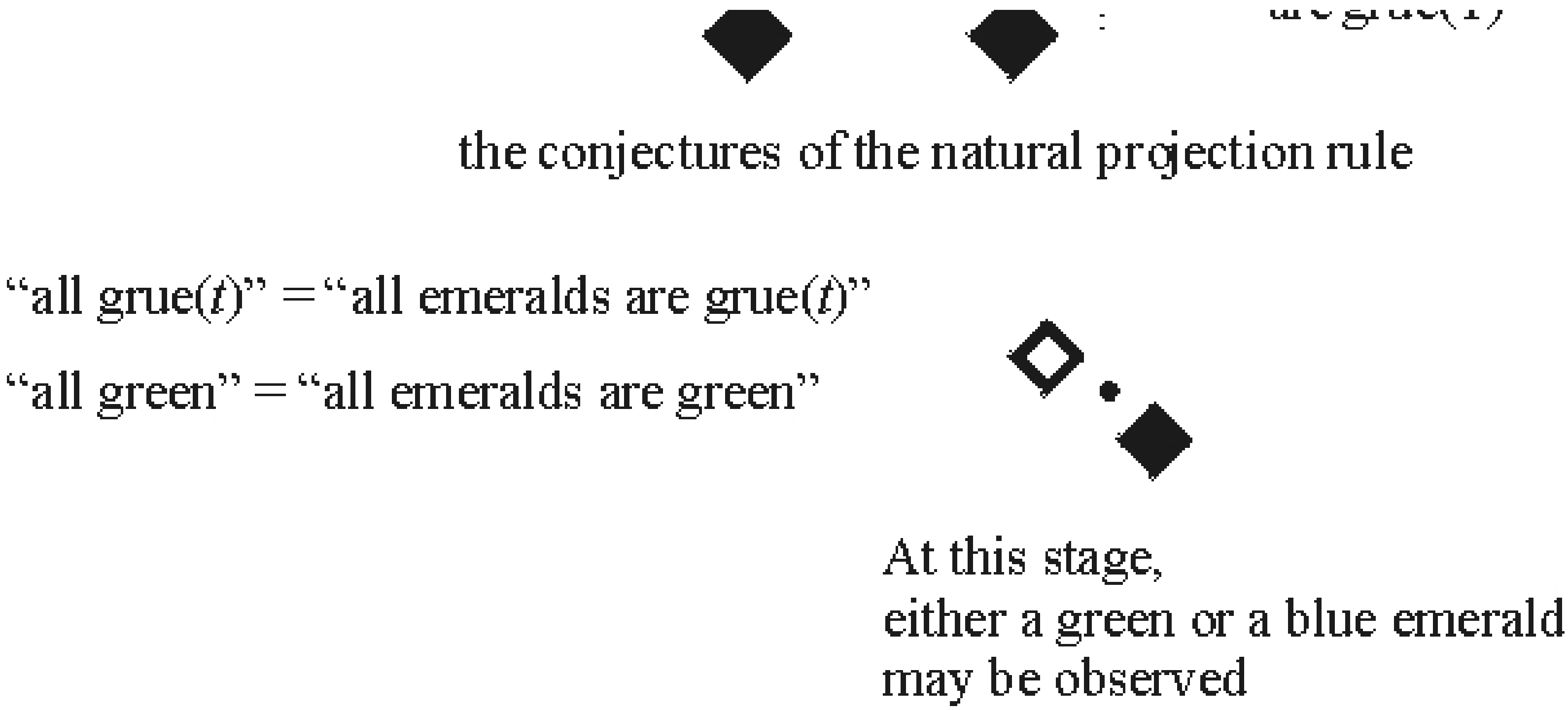
First published: February 2, 2002
Content last modified: February 2, 2002

Stanford Encyclopedia of Philosophy
Supplement to Formal Learning Theory

Basic Formal Definitions

The purpose of this supplement is a concise formal development of the basic notions of learning theory so as to make mathematical treatments of the subject more accessible to the reader. Also, the supplement develops some of the concepts discussed in the main entry in formal language. For the most part I follow Kelly's [1996] treatment which is at once more general and simpler than other approaches. The discussion below illustrates concepts by reference to the Goodmanian Riddle of Induction; the figure illustrating this inductive problem is reproduced here.





Evidence, Hypotheses and Discovery Problems

The basic building block of formal learning theory is the notion of an **evidence item**. For a general formulation, we may simply begin with a set *E* of evidence items. In general, nothing need be assumed about this set; in what follows, I will assume that *E* is at most countable, that is, that there are at most countably many evidence items. Some authors assume that evidence is formulated in first-order logic, typically as literals (e.g., [Earman 1992], [Martin and Osherson 1988]). In formal models of language learning, the evidence items are strings, representing grammatical strings from the language to be learned. In the example of the Riddle of Induction, the evidence items are G and B, respectively represented in the picture by a transparent and by a filled diamond, so *E* = {G,B}.

Given the basic set *E* of evidence items, we have the notion of a **finite evidence sequence**. A finite evidence sequence is a sequence (e_1, e_2, ..., e_n) of evidence items, that is, members of *E*. For example, the observation that the first three emeralds are green corresponds to the evidence sequence (G,G,G). A typical notation for a finite evidence sequence is *e*. If a finite evidence sequence *e* has *n* members, we say that the sequence is of length *n* and write *lh*(*e*) = *n*.

The next step is to consider an **infinite evidence sequence**. An infinite evidence sequence is a sequence (e_1, e_2, ..., e_n, ...) that continues indefinitely. For example, the infinite sequence (G,G,G,...,G,....) represents the circumstance in which all observed emeralds are green. A typical notation for an infinite evidence sequence is ε. Following Kelly [1996], the remainder of this supplement refers to an infinite evidence sequence as a **data stream**. Even though the notion of an infinite data sequence is mathematically straightforward, it takes some practice to get used to employing it. We often have occasion to refer to finite initial segments of a data stream, and introduce some special notation for this purpose: Let ε|*n* denote the first *n* evidence items in the data stream ε. For example if ε = (G,G,G,...,G,...) is the data stream featuring only green emeralds, then ε|3 = (G,G,G) is the finite evidence sequence corresponding to the observation

that the first three emeralds are green. We also write ε_n to denote the *n*-th evidence item observed in ε. For example, if ε = (G,G,G,...,G,...), then ε_2 = G.

An **empirical hypothesis** is a claim whose truth supervenes on a data stream. That is, a complete infinite sequence of observations settles whether or not an empirical hypothesis is true. For example, the hypothesis that "all observed emeralds are green" is true on the data stream featuring only green emeralds, and false on any data stream featuring a nongreen emerald. In general, we assume that a **correctness relation** *C* has been specified, where $C(\varepsilon,H)$ holds just in case hypothesis *H* is correct an data stream ε. What hypotheses are taken as correct on which data streams is a matter of the particular application. Given a correctness relation, we can define the empirical content of a hypothesis *H* as the set of data streams on which *H* is correct. Thus the empirical content of hypothesis *H* is given by $\{\varepsilon: C(\varepsilon, H)\}$. For formal purposes, it is often easiest to dispense with the correctness relation and simply to identify hypotheses with their empirical content. With that understanding, in what follows hypotheses will often be viewed as **sets of data streams**. For ease of exposition, I do not always distinguish between a hypothesis viewed as a set of data streams and an expression denoting that hypothesis, such as "all emeralds are green".

An inquirer typically does not begin inquiry as a tabula rasa, but has background assumptions about what the world is like. To the extent that such background assumptions help in inductive inquiry, they restrict the space of possible observations. For example in the discussion of the Riddle of Induction above, I assumed that that no data stream will be obtained that has green emeralds followed by blue emeralds followed by green emeralds. In the conservation principle problem discussed in the main entry, the operative background assumption is that the complete particle dynamics can be accounted for with conservation principles. As with hypotheses, we can represent the empirical content of given background assumptions by a set of data streams. Again it is simplest to identify **background knowledge** *K* with a set of data streams, namely the ones consistent with the background knowledge.

In a logical setting in which evidence statements are literals, learning theorists typically assume that a given data stream will feature all literals of the given first-order language (statements such as P(a) or ¬P(a)), and that the total set of evidence statements obtained during inquiry is consistent. With that background assumption, we may view the formula $\forall$x.P(x) as an empirical hypothesis that is correct on an infinite evidence sequence ε just in case no literal ¬P(a) appears on ε, that is for all *n* it is the case that $\varepsilon_n \neq \neg P(a)$. More generally, a data stream with a complete, consistent enumeration of literals determines the truth of every quantified statement in the given first-order language.

Inductive Methods and Inductive Success

An **inductive method** is a function that assigns hypotheses to finite evidence sequences. Following Kelly [1996], I use the symbol δ for an inductive method. Thus if *e* is a finite evidence sequence, then $\delta(e) = H$ expresses the fact that on finite evidence sequence *e*, the method δ outputs hypothesis *H*. It is also possible to have a method δ assign probabilities to hypotheses rather than choose a single conjecture, but I leave this complication aside here. Inductive methods are also called "learners" or "scientists"; no matter what the label is, the mathematical concept is the same. In the Goodmanian Riddle above, the natural projection

rule outputs the hypothesis "all emeralds are green" on any finite sequence of green emeralds. Thus if we denote the natural projection rule by δ, and the hypothesis that all emeralds are green by "all G", we have that δ(G) = "all G", δ(GG) = "all G", and so forth. Letting ε = (G,G,G,...,G,...) be the data stream with all green emeralds, we can write $\varepsilon|1$ = (G), $\varepsilon|2$ = (GG), etc., so we have that $\delta(\varepsilon|1)$ = "all G", $\delta(\varepsilon|2)$ = "all G", and more generally that $\delta(\varepsilon|n)$ = "all G" for all *n*.

An inductive method δ **converges to** a hypothesis *H* on a data stream ε **by time *n*** just in case for all later times $n' \geq n$, we have that $\delta(\varepsilon|n') = H$. This is a central definition for defining empirical success, as we will see shortly. To illustrate, the natural projection rule converges to "all G" by time 1 on the data stream ε = (G,G,G,...,G,...). It converges to "all emeralds are grue(3)" by time 3 on the data stream (G,G,B,B,B, ...). An inductive method δ **converges to** a hypothesis *H* on a data stream ε just in case there is a time *n* such that δ converges to *H* on ε by time *n*. Thus on the data stream (G,G,G...), the natural projection δ converges to "all G" whereas on the data stream (G,G,B,B,...) this rule converges to "all emeralds are grue(3)".

A **discovery problem** is a pair (**H**, *K*) where *K* is a set of data streams representing background knowledge and **H** is a mutually exclusive set of hypotheses that covers *K*. That is, for any two hypotheses *H*, *H'* in **H**, viewed as two sets of data streams, we have that $H \cap H' = \varnothing$. And for any data stream ε in K, there is a (unique) hypothesis *H* in **H** such that $\varepsilon \in H$. For example, in the Goodmanian Riddle of Induction, each alternative hypothesis is a singleton containing just one data stream, for example {(G,G,G,...)} for the empirical content of "all emeralds are green". The background knowledge *K* is just the union of the alternative hypotheses. In the problem involving the generalizations "all but finitely many ravens are white" and "all but finitely many ravens are black", the former hypothesis corresponds to the set of data streams featuring only finitely many black ravens, and the latter to the set of data streams featuring only finitely many white ravens. The background knowledge *K* corresponds to the set of data streams that eventually feature only white ravens or eventually feature only black ravens. Since each alternative hypothesis in a discovery problem (**H**, *K*) is mutually exclusive, for a given data stream ε in *K* there is exactly one hypothesis correct for that data stream; I write $H(\varepsilon)$ to denote that hypothesis.

In a discovery problem (**H**, *K*), an inductive method δ **succeeds** on a data stream ε in *K* iff δ converges to the hypothesis correct for ε; more formally, δ **succeeds** on a data stream ε in *K* iff δ converges to $H(\varepsilon)$ on ε. An inductive method δ **solves** the discovery problem (**H**, *K*) iff δ succeeds on all data streams in *K*. If δ solves a discovery problem (**H**, *K*), then we also say that δ is **reliable** for (**H**, *K*). If there is a reliable inductive method δ for a discovery problem (**H**, *K*), we say that the problem (**H**, *K*) is **solvable**. The main entry presented several solvable discovery problems. Characterization theorems like the one discussed there give conditions under which a discovery problem is solvable.

Efficient inductive inquiry is concerned with maximizing epistemic values other than convergence to the truth. Minimizing the number of mind changes is a topic in the main entry; what follows defines this measure of inductive performance as well as error and convergence time. Consider a discovery problem (**H**, *K*) and a data stream ε in *K*.

1. The **convergence time**, or **modulus**, of a method δ on ε is the least time *n* by which δ converges to a hypothesis *H* on ε. If δ is a reliable method for (**H**, *K*), then δ converges to a hypothesis on every data stream ε consistent with background knowledge *K* -- more specifically, δ converges to the correct hypothesis $H(\varepsilon)$ -- and the convergence time of δ is well-defined.
2. An inductive method δ **commits an error** at time *n* on ε iff $\delta(\varepsilon|n)$ is false, i.e., if $\delta(\varepsilon|n) \neq H(\varepsilon)$. As with convergence time, if δ is reliable, then it makes only finitely many errors on any data stream consistent with background knowledge. The number of errors commited by δ on a data stream ε is thus given by $|\{n: \delta(\varepsilon|n) \neq H(\varepsilon)\}|$.
3. To count mind changes (and errors) properly, it is useful to allow methods to produce an "uninformative conjecture" ?, which we may think of as a tautologous proposition. The point is that we don't want to count a change from "no opinion" to an informative hypothesis as a mind change. This device allows us to represent methods that "wait" until further evidence before taking an "inductive leap". Formally we say that an inductive method δ **changes its mind** at time *n*+1 on ε iff the method's previous conjecture at time *n* was informative and changes at time *n*+1. In symbols, δ **changes its mind** at time *n*+1 on ε iff: $\delta(\varepsilon|n) \neq ?$ and $\delta(\varepsilon|n) \neq \delta(\varepsilon|n+1)$. The number of mind changes made by δ on a data stream ε is thus given by $|\{n: \delta$ changes its mind on ε at time $n\}|$.

As we saw in the main entry, assessing methods by how well they do vis-a-vis these criteria of cognitive success leads to restrictions on inductive inferences in the short run, sometimes very strong restrictions. Learning-theoretic characterization theorems specify the structure of problems in which efficient inquiry is possible, and what kind of inferences lead to inductive success when it is attainable.

Oliver Schulte
oschulte@sfu.ca

Return to Formal Learning Theory

First published: February 2, 2002
Content last modified: February 2, 2002

Bayesian Epistemology

‘Bayesian epistemology’ became an epistemological movement in the 20th century, though its two main features can be traced back to the eponymous Reverend Thomas Bayes (c. 1701-61). Those two features are: (1) the introduction of a *formal apparatus* for inductive logic; (2) the introduction of a *pragmatic self-defeat test* (as illustrated by Dutch Book Arguments) for *epistemic* rationality as a way of extending the justification of the laws of deductive logic to include a justification for the laws of inductive logic. The formal apparatus itself has two main elements: the use of the laws of probability as coherence constraints on rational degrees of belief (or degrees of confidence) and the introduction of a rule of probabilistic inference, a rule or principle of *conditionalization*.

Bayesian epistemology did not emerge as a philosophical program until the first formal axiomatizations of probability theory in the first half of the 20th century. One important application of Bayesian epistemology has been to the analysis of scientific practice in *Bayesian Confirmation Theory*. In addition, a major branch of statistics, *Bayesian statistics*, is based on Bayesian principles. In psychology, an important branch of learning theory, *Bayesian learning theory*, is also based on Bayesian principles. Finally, the idea of analyzing rational degrees of belief in terms of rational betting behavior led to the 20th century development of a new kind of decision theory, *Bayesian decision theory*, which is now the dominant theoretical model for the both the descriptive and normative analysis of decisions. The combination of its precise formal apparatus and its novel pragmatic self-defeat test for justification makes Bayesian epistemology one of the most important developments in epistemology in the 20th century, and one of the most promising avenues for further progress in epistemology in the 21st century.

1. Deductive and Probabilistic Coherence and Deductive and Probabilistic Rules of Inference

There are two ways that the laws of deductive logic have been thought to provide rational constraints on belief: (1) Synchronically, the laws of deductive logic can be used to define the notion of deductive consistency and inconsistency. Deductive inconsistency so defined determines one kind of incoherence in belief, which I refer to as *deductive incoherence*. (2) Diachronically, the laws of deductive logic can constrain admissible changes in belief by providing the *deductive rules of inference*. For example, *modus ponens* is a deductive rule of inference that requires that one infer Q from premises P and $P \rightarrow Q$.

Bayesians propose additional standards of synchronic coherence -- standards of *probabilistic coherence* -- and additional rules of inference -- *probabilistic rules of inference* -- in both cases, to apply not to beliefs, but degrees of belief (degrees of confidence). For Bayesians, the most important standards of probabilistic coherence are the laws of probability. For more on the laws of probability, see the following supplementary article:

> Supplement on Probability Laws

For Bayesians, the most important probabilistic rule of inference is given by a p*rinciple of conditionalization*.

2. A Simple Principle of Conditionalization

If unconditional probabilities (e.g. $P(S)$) are taken as primitive, the conditional probability of S on T can be defined as follows:

> **Conditional Probability**:
> $P(S/T) = P(S\&T)/P(T)$.

By itself, the definition of conditional probability is of little epistemological significance. It acquires epistemological significance only in conjunction with a further epistemological assumption:

> **Simple Principle of Conditionalization**:
> If one begins with initial or *prior* probabilities P_i, and one acquires new evidence which can be represented as becoming certain of an evidentiary statement E (assumed to state the totality of one's new evidence and to have initial probability greater than zero), then rationality requires that one systematically transform one's initial probabilities to generate final or *posterior* probabilities P_f by conditionalizing on E -- that is: Where S is any statement, $P_f(S) = P_i(S/E)$.[1]

In epistemological terms, this Simple Principle of Conditionalization requires that the effects of evidence on rational degrees be analyzed in two stages: The first is non-inferential. It is the change in the probability of the evidence statement E from $P_i(E)$, assumed to be greater than zero and less than one, to $P_f(E) = 1$. The second is a probabilistic inference of conditionalizing on E from initial probabilities (e.g., $P_i(S)$) to final probabilities (e.g., $P_f(S) = P_i(S/E)$).

Problems with the Simple Principle (to be discussed below) have led many Bayesians to qualify the Simple Principle by limiting its scope. In addition, some Bayesians follow Jeffrey in generalizing the Simple Principle to apply to cases in which one's new evidence is less than certain (also discussed below). What unifies Bayesian epistemology is a conviction that conditionalizing (perhaps of a generalized sort) is rationally required in some important contexts -- that is, that some sort of conditionalization principle is an important principle governing rational changes in degrees of belief.

3. Dutch Book Arguments

Many arguments have been given for regarding the probability laws as coherence conditions on degrees of belief and for taking some principle of conditionalization to be a rule of probabilistic inference. The most distinctively Bayesian are those referred to as *Dutch Book Arguments*. Dutch Book Arguments represent the possibility of a new kind of justification for epistemological principles.

A Dutch Book Argument relies on some descriptive or normative assumptions to connect degrees of belief with willingness to wager -- for example, a person with degree of belief p in sentence S is assumed to be willing to pay up to and including $\$p$ for a unit wager on S (i.e., a wager that pays \$1 if S is true) and is willing to sell such a wager for any price equal to or greater than $\$p$ (one is assumed to be equally willing to buy or sell such a wager when the price is exactly $\$p$).[2] A *Dutch Book* is a combination of wagers which, on the basis of deductive logic alone, can be shown to entail a sure loss. A *synchronic Dutch Book* is a Dutch Book combination of wagers that one would accept all at the same time. A *diachronic Dutch Book* is a Dutch Book combination of wagers that one will be motivated to enter into at different times.

Ramsey and de Finetti first employed synchronic Dutch Book Arguments in support of the probability laws as standards of synchronic coherence for degrees of belief. The first diachronic Dutch Book Argument in support of a principle of conditionalization was reported by Teller, who credited David Lewis. The Lewis/Teller argument depends on a further descriptive or normative assumption about conditional probabilities due to de Finetti: An agent with conditional probability $P(S/T) = p$ is assumed to be willing to pay any price up to and including $\$p$ for a unit wager on S conditional on T. (A unit wager on S conditional on T is one that is called off, with the purchase price returned to the purchaser, if T is not true. If T is true, the wager is not called off and the wager pays \$1 if S is also true.) On this interpretation of conditional probabilities, Lewis, as reported by Teller, was able to show how to construct a diachronic Dutch Book against anyone who, on learning only that T, would predictably change his/her degree of belief in S to $P_f(S) > P_i(S/T)$; and how to construct a diachronic Dutch Book against anyone who, on

learning only that *T*, would predictably change his/her degree of belief in *S* to $P_f(S) < P_i(S/T)$. For illustrations of the strategy of the Ramsey/de Finetti and the Lewis/Teller arguments, see the following supplementary article:

> Supplement on Dutch Book Arguments

There has been much discussion of exactly what it is that Dutch Book Arguments are supposed to show. On the *literal-minded interpretation*, their significance is that they show that those whose degrees of belief violate the probability laws or those whose probabilistic inferences predictably violate a principle of conditionalization are liable to enter into wagers on which they are sure to lose. There is very little to be said for the literal-minded interpretation, because there is no basis for claiming that rationality requires that one be willing to wager in accordance with the behavioral assumptions described above. An agent could simply refuse to accept Dutch Book combinations of wagers.

A more plausible interpretation of Dutch Book Arguments is that they are to be understood hypothetically, as symptomatic of what has been termed *pragmatic self-defeat*. On this interpretation, Dutch Book Arguments are a kind of heuristic for determining when one's degrees of belief have the potential to be *pragmatically self-defeating*. The problem is not that one who violates the Bayesian constraints is likely to enter into a combination of wagers that constitute a Dutch Book, but that, on any reasonable way of translating one's degrees of belief into action, there is a potential for one's degrees of belief to motivate one to act in ways that make things worse than they might have been, when, as a matter of logic alone, it can be determined that alternative actions would have made things better (on one's own evaluations of better and worse).

Another way of understanding the problem of susceptibility to a Dutch Book is due to Ramsey: Someone who is susceptible to a Dutch Book evaluates identical bets differently based on how they are described. Putting it this way makes susceptibility to Dutch Books sound irrational. But this standard of rationality would make it irrational not to recognize all the logical consequences of what one believes. This is the *assumption of logical omniscience* (discussed below).

If successful, Dutch Book Arguments would reduce the justification of the principles of Bayesian epistemology to two elements: (1) an account of the appropriate relationship between degrees of belief and choice; and (2) the laws of deductive logic. Because it would seem that the truth about the appropriate relationship between the degrees of belief and choice is independent of epistemology, Dutch Book Arguments hold out the potential of justifying the principles of Bayesian epistemology in a way that requires no other epistemological resources than the laws of deductive logic. For this reason, it makes sense to think of Dutch Book Arguments as indirect, pragmatic arguments for according the principles of Bayesian epistemology much the same epistemological status as the laws of deductive logic. Dutch Book Arguments are a truly distinctive contribution made by Bayesians to the methodology of epistemology.

It should also be mentioned that some Bayesians have defended their principles more directly, with non-pragmatic arguments. In addition to reporting Lewis's Dutch Book Argument, Teller offers a non-

pragmatic defense of Conditionalization. There have been many proposed non-pragmatic defenses of the probability laws, the most compelling of which is due to Joyce. All such defenses, whether pragmatic or non-pragmatic, produce a puzzle for Bayesian epistemology: The principles of Bayesian epistemology are typically proposed as principles of *inductive* reasoning. But if the principles of Bayesian epistemology depend ultimately for their justification solely on the laws of deductive logic, what reason is there to think that they have any *inductive* content? That is to say, what reason is there to believe that they do anything more than extend the laws of deductive logic from beliefs to degrees of belief? It should be mentioned, however, that even if Bayesian epistemology only extended the laws of deductive logic to degrees of belief, that alone would represent an extremely important advance in epistemology.

4. Bayes' Theorem and Bayesian Confirmation Theory

This section reviews some of the most important results in the Bayesian analysis of scientific practice -- *Bayesian Confirmation Theory*. It is assumed that all statements to be evaluated have prior probability greater than zero and less than one.

Bayes' Theorem and a Corollary

Bayes' Theorem is a straightforward consequence of the probability axioms and the definition of conditional probability:

> **Bayes' Theorem**:
> $P(S/T) = P(T/S) \times P(S)/P(T)$ [where $P(T)$ is assumed to be greater than zero]

The epistemological significance of Bayes' Theorem is that it provides a straightforward corollary to the Simple Principle of Conditionalization. Where the final probability of a hypothesis *H* is generated by conditionalizing on evidence *E*, Bayes' Theorem provides a formula for the final probability of *H* in terms of the prior or initial *likelihood* of *H* on *E* ($P_i(E/H)$) and the prior or initial probabilities of *H* and *E*:

> **Corollary of the Simple Principle of Conditionalization**:
> $P_f(H) = P_i(H/E) = P_i(E/H) \times P_i(H)/P_i(E)$.

Due to the influence of Bayesianism, *likelihood* is now a technical term of art in confirmation theory. As used in this technical sense, likelihoods can be very useful. Often, when the conditional probability of *H* on *E* is in doubt, the likelihood of *H* on *E* can be computed from the theoretical assumptions of *H*.

Bayesian Confirmation Theory

A. Confirmation and disconfirmation. In Bayesian Confirmation Theory, it is said that evidence

confirms (or would confirm) hypothesis *H* (to at least some degree) just in case the prior probability of *H* conditional on *E* is greater than the prior unconditional probability of *H*: $P_i(H/E) > P_i(H)$. *E* disconfirms (or would disconfirm) *H* if the prior probability of *H* conditional on *E* is less than the prior unconditional probability of *H*.

B. Confirmation and disconfirmation by entailment. Whenever a hypothesis *H* logically entails evidence *E*, *E* confirms *H*. This follows from the fact that to determine the truth of *E* is to rule out a possibility assumed to have non-zero prior probability that is incompatible with *H* -- the possibility that ~*E*. A corollary is that, where *H* entails *E*, ~*E* would disconfirm *H*, by reducing its probability to zero. The most influential model of explanation in science is the hypothetico-deductive model (e.g., Hempel). Thus, one of the most important sources of support for Bayesian Confirmation Theory is that it can explain the role of hypothetico-deductive explanation in confirmation.

C. Confirmation of logical equivalents. If two hypotheses H1 and H2 are logically equivalent, then evidence *E* will confirm both equally. This follows from the fact that logically equivalent statements always are assigned the same probability.

D. The confirmatory effect of surprising or diverse evidence. From the corollary above, it follows that whether *E* confirms (or disconfirms) *H* depends on whether *E* is more probable (or less probable) conditional on *H* than it is unconditionally -- that is, on whether:

(b1) $P(E/H)/P(E) > 1$.

An intuitive way of understanding (b1) is to say that it states that *E* would be more expected (or less surprising) if it were known that *H* were true. So if *E* is surprising, but would not be surprising if we knew *H* were true, then *E* will significantly confirm *H*. Thus, Bayesians explain the tendency of surprising evidence to confirm hypotheses on which the evidence would be expected.

Similarly, because it is reasonable to think that evidence E_1 makes other evidence of the same kind much more probable, after E_1 has been determined to be true, other evidence of the same kind E_2 will generally not confirm hypothesis *H* as much as other diverse evidence E_3, even if *H* is equally likely on both E_2 and E_3. The explanation is that where E_1 makes E_2 much more probable than E_3 ($P_i(E_2/E_1) >> P_i(E_3/E_1)$), there is less potential for the discovery that E_2 is true to raise the probability of *H* than there is for the discovery that E_3 is true to do so.

E. Relative confirmation and likelihood ratios. Often it is important to be able to compare the effect of evidence *E* on two competing hypotheses, H_j and H_k, without having also to consider its effect on other hypotheses that may not be so easy to formulate or to compare with H_j and H_k. From the first corollary above, the ratio of the final probabilities of H_j and H_k would be given by:

Ratio Formula:

$$P_f(H_j)/P_f(H_k) = [P_i(E/H_j) \times P_i(H_j)]/[P_i(E/H_k) \times P_i(H_k)]$$

If the *odds of* H_j *relative to* H_k are defined as ratio of their probabilities, then from the Ratio Formula it follows that, in a case in which change in degrees of belief results from conditionalizing on E, the final odds ($P_f(H_j)/P_f(H_k)$) result from multiplying the initial odds ($P_i(H_j)/P_i(H_k)$) by the *likelihood ratio* ($P_i(E/H_j)/P_i(E/H_k)$). Thus, in pairwise comparisons of the odds of hypotheses, the likelihood ratio is the crucial determinant of the effect of the evidence on the odds.

F. The typical differential effect of positive evidence and negative evidence. Hempel first pointed out that we typically expect the hypothesis that all ravens are black to be confirmed to some degree by the observation of a black raven, but not by the observation of a non-black, non-raven. Let H be the hypothesis that all ravens are black. Let E_1 describe the observation of a non-black, non-raven. Let E_2 describe the observation of a black raven. Bayesian Confirmation Theory actually holds that both E_1 and E_2 may provide some confirmation for H. Recall that E_1 supports H just in case $P_i(E_1/H)/P_i(E_1) > 1$. It is plausible to think that this ratio is ever so slightly greater than one. On the other hand, E_2 would seem to provide much greater confirmation to H, because, in this example, it would be expected that $P_i(E_2/H)/P_i(E_2) >> P_i(E_1/H)/P_i(E_1)$.

These are only a sample of the results that have provided support for Bayesian Confirmation Theory as a theory of rational inference for science. For further examples, see Howson and Urbach. It should also be mentioned that an important branch of statistics, *Bayesian statistics* is based on the principles of Bayesian epistemology.

5. Potential Problems

This section reviews some of the most important potential problems for Bayesian Confirmation Theory and for Bayesian epistemology generally. No attempt is made to evaluate their seriousness here, though there is no generally agreed upon Bayesian solution to any of them.

5.1 Objections to the Probability Laws as Standards of Synchronic Coherence

A. The assumption of logical omniscience. The assumption that degrees of belief satisfy the probability laws implies omniscience about deductive logic, because the probability laws require that all deductive logical truths have probability one, all deductive inconsistencies have probability zero, and the probability of any conjunction of sentences be no greater than *any* of its deductive consequences. This seems to be an unrealistic standard for human beings. Hacking and Garber have made proposals to relax the assumption of logical omniscience. Because relaxing that assumption would block the derivation of almost all the important results in Bayesian epistemology, most Bayesians maintain the assumption of logical omniscience and treat it as an ideal to which human beings can only more or less approximate.

B. The problem of the priors. Are there constraints on prior probabilities other than the probability laws? Consider Goodman's "new riddle of induction": In the past all observed emeralds have been green. Do those observations provide any more support for the generalization that all emeralds are green than they do for the generalization that all emeralds are grue (green if observed before now; blue if observed later); or do they provide any more support for the prediction that the next emerald observed will be green than for the prediction that the next emerald observed will be grue (i.e., blue)? This question divides Bayesians into two categories:

> (a) *Objective Bayesians* (e.g., Rosenkrantz) hold that there are rational constraints on prior probabilities that require that observations support the green-generalization and the green-prediction much more strongly than the grue-generalization and the grue-prediction. Objective Bayesians are the intellectual heirs of the advocates of a Principle of Indifference for probability. Rosenkrantz builds his account on the maximum entropy rule proposed by E.T. Jaynes. The difficulties in formulating an acceptable Principle of Indifference have led most Bayesians to abandon Objective Bayesianism.
>
> (b) *Subjective Bayesians* (e.g., de Finetti) do not believe that rationality alone places enough constraints on one's prior probabilities to make them objective. For Subjective Bayesians, it is up to our own free choice or to evolution or to socialization or some other non-rational process to determine one's prior probabilities. Rationality only requires that the prior probabilities satisfy relatively modest synchronic coherence conditions.

Subjective Bayesians believe that their position is not objectionably subjective, because of results (e.g., Doob or Gaifman and Snir) proving that even subjects beginning with very different prior probabilities will tend to converge in their final probabilities, given a suitably long series of shared observations. These *convergence results* are not completely reassuring, however, because they only apply to agents who already have significant agreement in their priors and they do not assure convergence in any reasonable amount of time. Also, they typically only guarantee convergence on the probability of predictions, not on the probability of theoretical hypotheses. For example, Carnap favored prior probabilities that would never raise above zero the probability of a generalization over a potentially infinite number of instances (e.g., that all crows are black), no matter how many observations of positive instances (e.g., black crows) one might make without finding any negative instances (i.e., non-black crows). In addition, the convergence results depend on the assumption that the *only* changes in probabilities that occur are those that are the non-inferential results of observation on evidential statements and those that result from conditionalization on such evidential statements.

Because of the problem of the priors, it is an open question whether Bayesian Confirmation Theory has inductive content, or whether it merely translates the framework for rational belief provided by deductive logic into a corresponding framework for rational degrees of belief.

5.2 Objections to The Simple Principle of Conditionalization as a Rule

of Inference, Especially as an Explanation of Theory Confirmation in Science

A. The problem of uncertain evidence. The Simple Principle of Conditionalization requires that the acquisition of evidence be representable as changing one's degree of belief in a statement *E* to one -- that is, to certainty. But many philosophers would object to assigning probability of one to any contingent statement, even an evidential statement, because, for example, it is well-known that scientists sometimes give up previously accepted evidence. Jeffrey has proposed a generalization of the Principle of Conditionalization that yields that principle as a special case. Jeffrey's idea is that what is crucial about observation is not that it yields certainty, but that it generates a non-inferential change in the probability of an evidential statement *E* and its negation ~*E* (assumed to be the locus of all the non-inferential changes in probability) from initial probabilities between zero and one to $P_f(E)$ and $P_f(\sim E) = [1 - P_f(E)]$. Then on Jeffrey's account, after the observation, the rational degree of belief to place in an hypothesis *H* would be given by the following principle:

> **Principle of Jeffrey Conditionalization**:
> $P_f(H) = P_i(H/E) \times P_f(E) + P_i(H/\sim E) \times P_f(\sim E)$ [where *E* and *H* are both assumed to have prior probabilities between zero and one]

Counting in favor of Jeffrey's Principle is its theoretical elegance. Counting against it is the practical problem that it requires that one be able to completely specify the direct non-inferential effects of an observation, something it is doubtful that anyone has ever done. Skyrms has given it a Dutch Book defense.

B. The problem of old evidence. On a Bayesian account, the effect of evidence *E* in confirming (or disconfirming) a hypothesis is solely a function of the increase in probability that accrues to *E* when it is first determined to be true. This raises the following puzzle for Bayesian Confirmation Theory discussed extensively by Glymour: Suppose that *E* is an evidentiary statement that has been known for some time -- that is, that it is *old evidence*; and suppose that *H* is a scientific theory that has been under consideration for some time. One day it is discovered that *H* implies *E*. In scientific practice, the discovery that *H* implied *E* would typically be taken to provide some degree of confirmatory support for *H*. But Bayesian Confirmation Theory seems unable to explain how a previously known evidentiary statement *E* could provide any new support for H. For conditionalization to come into play, there must be a change in the probability of the evidence statement *E*. Where *E* is old evidence, there is no change in its probability. Some Bayesians who have tried to solve this problem (e.g., Garber) have typically tried to weaken the logical omniscience assumption to allow for the possibility of discovering logical relations (e.g., that *H* and suitable auxiliary assumptions imply *E*). As mentioned above, relaxing the logical omniscience assumption threatens to block the derivation of almost all of the important results in Bayesian epistemology, so there is no general agreement among Bayesians on how to solve this problem. Other Bayesians (e.g., Lange) employ the Bayesian formalism as a tool in the *rational reconstruction* of the evidentiary support for a scientific hypothesis, where it is irrelevant to the rational reconstruction whether the evidence was discovered before or after the theory was initially formulated.

C. The problem of rigid conditional probabilities. When one conditionalizes, one applies the initial conditional probabilities to determine final unconditional probabilities. Throughout, the conditional probabilities themselves do not change; they remain rigid. Examples of the Problem of Old Evidence are but one of a variety of cases in which it seems that it can be rational to change one's initial conditional probabilities. Thus, many Bayesians reject the Simple Principle of Conditionalization in favor of a qualified principle, limited to situations in which one does not change one's initial conditional probabilities. There is no generally accepted account of when it is rational to maintain rigid initial conditional probabilities and when it is not.

D. The problem of prediction vs. accommodation. Related to the problem of Old Evidence is the following potential problem: Consider two different scenarios. In the first, theory *H* was developed in part to *accommodate* (i.e., to imply) some previously known evidence E. In the second, theory *H* was developed at a time when *E* was not known. It was because *E* was derived as a *prediction* from *H* that a test was performed and *E* was found to be true. It seems that E's being true would provide a greater degree of confirmation for *H* if the truth of *E* had been *predicted* by *H* than if *H* had been developed to *accommodate* the truth of *E*. There is no general agreement among Bayesians about how to resolve this problem. Some (e.g., Horwich) argue that Bayesianism implies that there is no important difference between prediction and accommodation, and try to defend that implication. Others (e.g., Maher) argue that there is a way to understand Bayesianism so as to explain why there is an important difference between prediction and accommodation.

E. The problem of new theories. Suppose that there is one theory H_1 that is generally regarded as highly confirmed by the available evidence *E*. It is possible that simply the introduction of an alternative theory H_2 can lead to an erosion of H_1's support. It is plausible to think that Copernicus' introduction of the heliocentric hypothesis had this effect on the previously unchallenged Ptolemaic earth-centered astronomy. This sort of change cannot be explained by conditionalization. It is for this reason that many Bayesians prefer to focus on probability ratios of hypotheses (see the Ratio Formula above), rather than their absolute probability; but it is clear that the introduction of a new theory could also alter the probability ratio of two hypotheses -- for example, if it implied one of them as a special case.

6. Other Principles of Bayesian Epistemology

Other principles of Bayesian epistemology have been proposed, but none has garnered anywhere near a majority of support among Bayesians. The most important proposals are merely mentioned here. It is beyond the scope of this entry to discuss them in any detail.

A. Other principles of synchronic coherence. Are the probability laws the only standards of synchronic coherence for degrees of belief? Van Fraassen has proposed an additional principle (Reflection or Special Reflection), which he now regards as a special case of an even more general principle (General Reflection).[3]

B. Other probabilistic rules of inference. There seem to be at least two different concepts of probability: the probability that is involved in degrees of belief (epistemic or subjective probability) and the probability that is involved in random events, such as the tossing of a coin (chance). De Finetti thought this was a mistake and that there was only one kind of probability, subjective probability. For Bayesians who believe in both kinds of probability, an important question is: What is (or should be) the relation between them? The answer can be found in the various proposals for principles of direct inference in the literature. Typically, principles of direct inference are proposed as principles for inferring subjective or epistemic probabilities from beliefs about objective chance (e.g., Pollock). Lewis reverses the direction of inference, and proposes to infer beliefs about objective chance from subjective or epistemic probabilities, via his (Reformulated) Principal Principle.[4]

C. Principles of rational acceptance. What is the relation between beliefs and degrees of belief? Jeffrey proposes to give up the notion of belief (at least for empirical statements) and make do with only degrees of belief. Other authors (e.g., Levi, Maher, Kaplan) propose principles of rational acceptance as part of accounts of when it is rational to accept a statement as true, not merely to regard it as probable.

Bibliography

- Bayes, Thomas, "An Essay Towards Solving a Problem in the Doctrine of Chances", *Philosophical Transactions of the Royal Society of London* (1764) 53: 37-418, reprinted in E.S. Pearson and M.G. Kendall, eds., Studies in the History of Statistics and probability (London: Charles Griffin, 1970).
- Carnap, Rudolf, *Logical Foundations of Probability* (Chicago: University of Chicago Press; 1950).
- Carnap, Rudolf, *The Continuum of Inductive Methods* (Chicago: University of Chicago Press; 1952).
- de Finetti, Bruno, "La Prevision: ses lois logiques, se sources subjectives" (Annales de l'Institut Henri Poincare 7 (1937): 1-68. Translated into English and reprinted in Kyburg and Smokler, *Studies in Subjective Probability* (Huntington, NY: Krieger; 1980).
- Doob, J.L., "What is a Martingale?", *American Mathematical Monthly* 78 (1971): 451-462.
- Earman, John, *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory* (Cambridge, MA: MIT Press; 1992).
- Gaifman, H., and Snir, M., "Probabilities over Rich Languages", *Journal of Symbolic Logic* 47 (1982): 495-548.
- Garber, Daniel, "Old Evidence and Logical Omniscience in Bayesian Confirmation Theory", in J. Earman, ed., *Testing Scientific Theories, Midwest Studies in the Philosophy of Science*, Vol. X (Minneapolis: University of Minnesota Press; 1983): 99-131.
- Goodman, Nelson, *Fact, Fiction, and Forecast* (Cambridge: Harvard University Press; 1983).
- Glymour, Clark, *Theory and Evidence* (Princeton: Princeton University Press; 1980).
- Hacking, Ian, "Slightly More Realistic Personal Probability", *Philosophy of Science* 34 (1967): 311-325.
- Hempel, Carl G., *Aspects of Scientific Explanation* (New York: Free Press; 1965).
- Horwich, Paul, *Probability and Evidence* (Cambridge: Cambridge University Press; 1982).

- Howson, Colin, and Peter Urbach, *Scientific Reasoning: The Bayesian Approach, 2nd ed.* (Chicago: Open Court; 1993).
- Jaynes, E.T., "Prior Probabilities", *Institute of Electrical and Electronic Engineers Transactions on Systems Science and Cybernetics,* SSC-4 (1968): 227-241.
- Jeffrey, Richard, *The Logic of Decision, 2nd ed.* (Chicago: University of Chicago Press; 1983).
- Jeffrey, Richard, *Probability and the Art of Judgment* (Cambridge: Cambridge University Press; 1992).
- Joyce, James M., "A Nonpragmatic Vindication of Probabilism", *Philosophy of Science* 65 (1998): 575-603.
- Joyce, James M., *The Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press; 1999).
- Kaplan, Mark, *Decision Theory as Philosophy* (Cambridge: Cambridge University Press; 1996).
- Lange, Marc, "Calibration and the Epistemological Role of Bayesian Conditionalization", *Journal of Philosophy* 96 (1999): 294-324.
- Levi, Isaac, *The Enterprise of Knowledge* (Cambridge, Mass.: MIT Press; 1980)
- Levi, Isaac, *The Fixation Of Belief And Its Undoing* (Cambridge: Cambridge University Press; 1991).
- Lewis, David, "A Subjectivist's Guide to Objective Chance", in Richard C. Jeffrey, ed., *Studies in Inductive Logic and Probability, vol. 2* (Berkeley: University of California Press; 1980): 263-293.
- Maher, Patrick, "Prediction, Accommodation, and the Logic of Discovery", *PSA*, vol. 1 (1988): 273-285.
- Maher, Patrick, *Betting on Theories* (Cambridge: Cambridge University Press; 1993).
- Pollock, John L., *Nomic Probability and the Foundations of Induction* (Oxford: Oxford University Press; 1990).
- Popper, Karl, *The Logic of Scientific Discovery*, 3rd ed. (London: Hutchinson; 1968).
- Ramsey, Frank P., "Truth and Probability," in Richard B. Braithwaite (ed.), *Foundations of Mathematics and Other Logical Essay* (London: Routledge and Kegan Paul; Check on 1931 publication date), pp. 156-198.
- Réyni, A., "On a New Axiomatic Theory of Probability", *Acta Mathematica Academiae Scientiarium Hungaricae* 6 (1955): 285-385.
- Rosenkrantz, R.D., *Foundations and Applications of Inductive Probability* (Atascadero, CA: Ridgeview Publishing; 1981).
- Savage, Leonard, *The Foundations of Statistics, 2nd ed.* (New York: Dover; 1972).
- Skyrms, Brian, *Pragmatics and Empiricism* (New Haven: Yale University Press; 1984).
- Skyrms, Brian, *TheDynamics of Rational Deliberation* (Cambridge, Mass.: Harvard University Press; 1990).
- Teller, Paul, "Conditionalization, Observation, and Change of Preference", in W. Harper and C.A. Hooker, eds., *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (Dordrecht: D. Reidel; 1976).
- Van Fraassen, Bas C., "Calibration: A Frequency Justification for Personal Probability", in R.S. Cohen and L. Laudan, eds., *Physics, Philosophy, and Psychoanalysis: Essays in Honor of Adolf Grunbaum* (Dordrecht: Reidel; 1983).
- Van Fraassen, Bas C., "Belief and the Will", *Journal of Philosophy* 81 (1984): 235-256.

- Van Fraassen, Bas C., "Belief and the Problem of Ulysses and the Sirens", *Philosophical Studies* 77 (1995): 7-37.
- Zynda, Lyle, "Old Evidence and New Theories", *Philosophical Studies* 77 (1995): 67-95.

Other Internet Resources

[Please contact the author with suggestions]

Related Entries

Bayes' Theorem | logic: inductive | probability calculus: interpretations of

Acknowledgements

In the preparation of this article, I have benefited from comments from Marc Lange, Stephen Glaister, Laurence BonJour, and James Joyce.

William J. Talbott
wtalbott@u.washington.edu

First published: July 12, 2001
Content last modified: July 12, 2001

Stanford Encyclopedia of Philosophy
Supplement to Bayesian Epistemology

Probability Laws

There are many different versions of the probability laws. Probability can be defined over sentences or over sets; it can be defined as conditional or unconditional. This article assumes the following laws of unconditional probability defined over sentences:

> (A1) All probabilities are between zero and one -- that is, for any sentence S: $0 \leq P(S) \leq 1$.
>
> (A2) Logical truths have probability one -- that is, for any logical truth L: $P(L) = 1$.
>
> (A3) Where S and T, are mutually exclusive, the probability of S or T ($S \vee T$) is the sum of their individual probabilities -- that is: $P(S \vee T) = P(S) + P(T)$.

Using these laws, it is possible to derive as theorems many of the standards truths of probability -- for example, that the probability of a sentence and its negation sum to one -- in other words: $P(\sim S) = 1 - P(S)$.

(A3) is referred to as the *Principle of Finite Additivity*, because it involves only finite sums. Most mathematical treatments of probability require an extension of (A3) to cover countably infinite sums. The result is a *Principle of Countable Additivity*. The standard way of stating this axiom involves translating the axioms into set theory, where countably infinite unions (corresponding to infinite disjunctions) are defined. The relation to (A3) is clearer if one simply extends the ordinary notion of sentence to include infinitely long expressions formed in accordance with the formation rules of the language -- thus allowing for the possibility of infinite disjunctions:

> (A4) Where S_1, S_2, ... is a countably infinite sequence of mutually exclusive sentences: $P(S_1 \vee S_2 \vee \dots) = P(S_1) + P(S_2) + \dots$.

The discussion of the probability axioms in the text assumes only Finite Additivity, because it is only Axioms (A1)-(A3) that can be given a Dutch Book justification.

William J. Talbott
wtalbott@u.washington.edu

[Return to Bayesian Epistemology]

First published: July 12, 2001
Content last modified: July 12, 2001

Stanford Encyclopedia of Philosophy
Notes to Bayesian Epistemology

Notes

1. Some authors (e.g., de Finetti) take conditional probabilities rather than unconditional probabilities as primitive. Following ideas of Popper and Réyni, it is possible to take unconditional probabilities as primitive and define conditional probabilities for statements with prior probability of zero.

2. The assumptions are those necessary to assure that gambles are ranked by their expected monetary value. The idea of explaining rational behavior in terms of maximizing the expected value of gambles led to the 20th century development of Bayesian Decision Theory. See, for example, Ramsey, de Finetti, Savage, and Jeffrey.

3. Van Fraasen's Special Reflection Principle is: $P(A|p_t(A) = x) = x$ (where $t \geq 0$). His General Reflection Principle is that one's current opinion about an event E must lie in the range spanned by the possible opinions (based on one's present opinion) that one may come to have about E at later time t. Van Fraassen provides diachronic Dutch Book Arguments for both, though he downplays the significance of Dutch Book Arguments generally.

4. Lewis's (Reformulated) Principal Principle relates Chance for a world w at time t (P_{tw}) to Credence (C) as a function of the complete history of world w up to time t (H_{tw}) and the complete theory of chance for world w (T_w) as follows: $P_{tw}(A) = C(A/H_{tw}T_w)$.

William J. Talbott
wtalbott@u.washington.edu

First published: July 12, 2001
Content last modified: July 12, 2001

Stanford Encyclopedia of Philosophy
Supplement to Bayesian Epistemology

Dutch Book Arguments

The Ramsey/de Finetti argument can be illustrated by an example. Suppose that agent *A*'s degrees of belief in *S* and ~*S* (written *db*(*S*) and *db*(~*S*)) are each .51, and, thus that their sum 1.02 (greater than one). On the behavioral interpretation of degrees of belief introduced above, *A* would be willing to pay *db*(*S*) × $1 for a unit wager on *S* and *db*(~*S*) × $1 for a unit wager on ~*S*. If a bookie *B* sells both wagers to *A* for a total of $1.02, the combination would be a synchronic Dutch Book -- synchronic because the wagers could both be entered into at the same time, and a Dutch Book because *A* would have paid $1.02 on a combination of wagers guaranteed to pay exactly $1. Thus, *A* would have a guaranteed net loss of $.02

The Lewis/Teller argument can also be illustrated by an example. Suppose that agent *A*'s degrees of belief satisfy the synchronic probabilistic coherence conditions -- that is, the probability laws. Suppose also that *A* has the following initial probabilities:

$P_i(S) = 1/5$

$P_i(T) = 1/5$

$P_i(S\&T) = 1/10$

$P_i(S/T) = 1/2$

A is about to learn whether or not *T* is true (nothing more). If *A* learns that *T* is true, the Simple Principle of Conditionalization would require *A* to change her probability assignment to *S* ($P_f(S)$) to equal $P_i(S/T)$ = ½. Suppose *A* realizes that, if she learns that *T* is true, she will change her probability assignment to *S* to $P_f(S) = 6/10 > P_i(S/T)$ [a parallel argument applies to the case in which *A* knows in advance that were she to learn that *T*, $P_f(S)$ would be less than $P_i(S/T)$].

Initially, bookie *B* can make the following wagers with *A*:

(1) *B* sells *A* an unconditional wager that pays $.10 if *T* is true for $P_i(T)$ × $.10 = 1/5 × $.10 = $.02.

(2) *B* buys from *A* a unit wager on *S* conditional on *T* for $P_i(S/T) \times \$1 = ½ \times \$1 = \$.50$.

After it is determined whether or not *T* is true, there are two possibilities:

(a) *T* is not true.

In that case, *A* loses $.02 on the first wager and the second wager is called off, so no one wins or loses anything on the second wager. The result is a net loss of $.02 for *A*.

(b) *T* is true.

In that case, *B* makes an additional wager with *A*:

(3) *B* sells to *A* an unconditional unit wager on *S* for $P_f(S) \times \$1 = 6/10 \times \$1 = \$.60$.

Then there are two further sub-possibilities:

(b1) *S* is true. *A* gains $.08 on wager 1 (the $.10 pay-off, less the $.02 that *A* paid for the wager); *A* loses $.50 on wager 2 (*B* paid *A* $.50 for the wager, but *A* must pay $1 to *B*); *A* gains $.40 on wager 3 (*A* paid *B* $.60 for the wager, but *B* must pay *A* $1). The net result of all three wagers is a $.02 loss for *A*.

(b2) *S* is not true. Again *A* gains $.08 on wager 1 (the $.10 pay-off, less the $.02 that *A* paid for the wager); *A* gains $.50 on wager 2 (*B* paid *A* $.50 for the wager, and *A* does not pay *B* anything); *A* loses $.60 on wager 3 (*A* paid *B* $.60 for the wager, and *B* does not pay *A* anything). Again the net result of all three wagers is a $.02 loss for *A*.

Because (a), (b1), and (b2) exhaust all the logical possibilities, the example is one in which *A* is guaranteed to lose $.02, no matter what happens. Because wager 3 cannot be made at the same time as wagers 1 and 2, the combination of wagers 1-3 is a diachronic Dutch Book.

William J. Talbott
wtalbott@u.washington.edu

Return to Bayesian Epistemology

First published: July 12, 2001
Content last modified: July 12, 2001

Scottish Philosophy in the Nineteenth Century

Philosophical debate in 19th century Scotland was very vigorous, its agenda being set in large part by the impact of Kant and German Idealism on the philosophical tradition of the Scottish Enlightenment. The principal figures are Sir William Hamilton, James Frederick Ferrier and Alexander Bain, and later in the century, the so-called ‘Glasgow Idealists’ notably Edward Caird and Sir Henry Jones.

1. The Enlightenment Background

While Scottish philosophy of the 18th century is studied extensively, Scottish philosophy in the 19th century is neglected to the point of being virtually unknown. Francis Hutcheson, David Hume and Thomas Reid and are names familiar to almost all philosophers; Sir William Hamilton, James Frederick Ferrier and Alexander Bain to hardly any. Yet in their day, the names of these philosophers were not only prominent in Scotland, but widely known across Europe. To understand this decline in reputation, it is necessary to see 19th century Scottish philosophy against the background of the century that preceded it.

According to George Davie there is an

> opposition … between … two contrasting positions that in their tension provided Scottish philosophy with its central problem: the Berkeleian system, according to which, in the interests of reconciling progress with traditional standards, we are to set aside the instincts

> of the farmer in favour of the sophistication of the philosopher and to think with the learned while we talk with the vulgar; and the Hutchesonian system, according to which, with the same aim of reconciling material advance with the intellectual principle, we are to respect the instincts of the farmer as against the sophistication of the philosopher and initiate a sort of dialogue between the vulgar and the learned, instead of talking down to the farmer from the standpoint of the philosopher. (Davie 1994: 41-2)

Cast in these terms it is easy to place the two most famous philosophers of the Scottish Enlightenment on either side of the divide. On the side of the first is Hume, whose skeptical conclusions arise from the Berkleyan presupposition asserted in the very first sentence of his *Treatise of Human Nature*

> All the perceptions of the human mind resolve themselves into two distinct kinds, which I shall call IMPRESSIONS and IDEAS. The difference betwixt these consists in the degrees of force and liveliness with which they strike upon the mind. (Hume 1888: 1)

On the other side is Thomas Reid, for whom the errors of Hume result from the boldness of his starting point.

> It is genius, and not the want of it, that adulterates philosophy, and fills it with error and false theory. A creative imagination disdains the mean offices of digging for a foundation, of removing rubbish, and carrying materials: leaving these servile employments to the drudges in science, it plans a design, and raises a fabric. (Reid 1997: 15)

The problem as Reid saw it was that a highly theoretical philosophy was trying to run before it could walk, because in sharp contrast to subjects that are "really sciences" -- mechanics, astronomy and optics are the examples he gives --

> when we turn our attention inward and consider the phaenomena of human thoughts, opinions and perceptions, and endeavour to trace them to the general laws and first principles of our constitution, we are immediately involved in darkness and perplexity. And if common sense, or the principles of education, happen not to be stubborn, it is odds but we end in absolute skepticism. (Reid 1997: 16)

It is well known that, on Reid's analysis, Hume's skepticism derives in large part from his implicit subscription to the 'way of ideas', a conception of knowledge and experience that finds its origins in Descartes, Malebranche and Locke, and its most dramatic exposition in Berkeley who, though no skeptic, "proved by unanswerable arguments what no man in his senses could believe" (Reid 1997: 20). The antidote to such skepticism is common sense, but not of the robust sort displayed by Dr Johnson when he purported to refute Berkeley by kicking a stone. 'Common sense' can mean two things, in fact: widespread popular conviction on the one hand, or the basic principles at work in human reasoning and belief formation on the other. Widespread conviction can be false, of course, which is why the method of the School of Common Sense was thought suspect by many, described by Kant, for example, as a

stratagem by which 'the stalest windbag can confidently take up with the soundest thinker" (Kant 1951: 259). But in Reid at any rate, philosophical inquiry into the human mind is not a matter of making popular opinion the test of truth, but of initiating a 'dialogue between the vulgar and the learned,' (to repeat Davie's happy phrase) in which proper weight is attached to actual minds at work.

There is, then, this deep division within the philosophy of the Scottish Enlightenment, yet it occurs within a context of striking unanimity also. "Wise men now agree, or ought to agree in this, that there is but one way to the knowledge of nature's works; the way of observation and experiment" Reid writes (Reid 1997: 11), thereby endorsing the express intention of Hume to "introduce the experimental method of reasoning into moral subjects' (the subtitle of the *Treatise*). Both remarks reflect a commitment to the project of a 'science of mind', a project common to all the major Scottish philosophers of the period. Thus George Turnbull (Reid's teacher) writing in 1740 says "I was led long ago to apply myself to the study of the human mind in the same way as to that of the human body" (quoted in Davie 1994: 24)

In short, both division and unanimity are present within eighteenth century Scottish philosophy, unanimity with respect to aim – a science of mind – and division with respect to method – the 'principles of common sense' versus 'the way of ideas'. This is a tension, however, within only one part of 18th century Scottish philosophy, namely the philosophy of sensation and perception, and not perhaps the most influential part. The Scottish Enlightenment is in many ways more marked by the type of thinking about social and political topics that we find in Adam Smith and Adam Ferguson, as well as Hume, who in this respect take their cue from Hutcheson. The example *par excellence* is Smith's *Theory of Moral Sentiments*(1759; 6th edition 1790), where the Humean ambition of countering the "books of divinity and abstruse metaphysics" (*Treatise*) was further by a sympathetic attention to how human beings in society actually are, and what social forms and political arrangements will best work to their happiness and well being.

In the 19th century, this strand of Enlightenment thinking ceased to be an important part of the philosophical agenda. That agenda was dominated, rather, by the 'science of mind' more narrowly conceived, that is to say logic (i.e the philosophy of truth and reason) and the philosophy of perception. Consequently, from 1810 onwards, when Thomas Brown (1778-1820) took up the Chair of Moral Philosophy at Edinburgh, the story of Scottish philosophy is that of repeated attempts to resolve the tension that lay within that 'science'. It is also a story of remarkable continuity. Brown (1778-1820) was a student of Dugald Stewart (1753-1828), who in turn was a student and friend of Reid and himself held the Chair of Moral Philosophy at both Glasgow and Edinburgh. Stewart was enormously highly thought of in his own day, but in retrospect his contribution to the central debate in Scottish philosophy is limited. Brown died prematurely, but he was a prolific writer from an early age and left behind voluminous lectures. These lectures are critical of Reid (though on certain issues Brown way be said to side with Reid against Hume). On publication they were widely and rapturously received, but fell into almost total neglect by 1840. Perhaps their most enduring effect on the debate arose from the re-interpretation and defence of Reid that they induced on the part of the most prominent philosopher of the period – Sir William Hamilton.

2. Sir William Hamilton (1788-1856)

Sir William Hamilton was a graduate of the Universities of Glasgow and Oxford. At Glasgow he studied logic and moral philosophy under George Jardine and James Mylne both of whom are figures included in James McCosh's *The Scottish Philosophy*. In 1807 at Balliol College Oxford he held the Snell Exhibition, a scholarship that regularly allowed Scottish students of philosophy to spend time at England's oldest university, and he gained an extensive knowledge of Aristotelianism there. From 1811-21 he worked at the Scottish Bar (not altogether successfully) until being appointed Professor of Universal and Civil History at the University of Edinburgh, where he transferred to the Chair of Logic and Metaphysics in 1836, a post he held until his death in 1856.

At the height of his powers, Hamilton was regarded as a major European intellectual figure, and evidence of his stature lies in the fact that Hamilton was included in the series *Philosophical Classics*, edited by William Knight, Professor of Moral Philosophy at St Andrews, and thus ranked alongside Descartes, Berkeley, Locke, Kant and Hegel. Such an estimation must now strike us as bizarre, yet there is point in asking why his times regarded him in such a favorable light. The answer is that, thanks to two trips he made to Germany during his years as a lawyer, Hamilton acquired an extensive knowledge of German philosophy, little of which had been translated into English and which he could read in the original language. At the same time, he was not only thoroughly versed in the Scottish tradition of philosophy that he had acquired from Jardine and Mill, but an enthusiastic exponent of Reid, whose collected works he edited and annotated extensively. He was thus perfectly placed to broaden the horizons of Scottish philosophy, to push it beyond the narrower confines of Common Sense by bringing to wider attention the importance of Kant, and yet to do so as one profoundly sympathetic to the native tradition. It is precisely for these reasons, in fact, that he is praised by John Veitch in the *Philosophical Classics* volume devoted to his philosophy.

Hamilton's writings are extensive but arguably his views can be adequately ascertained from three long essays which appeared in the *Edinburgh Review* -- 'The Philosophy of the Unconditioned' (1829), 'The Philosophy of Perception' (1830) and 'Logic' (1833), subsequently republished in a collection of his writings. In the first of these Hamilton recounts the course philosophy had taken in France after "the philosophy of Descartes and Malebranche had sunk into oblivion"(Hamilton 1853: 2). At first there emerged a highly materialist version of Lockean empiricism "a doctrine so melancholy in its consequences, and founded on principles thus partial and exaggerated, [that it] could not be permanent" (*ibid*., 3). Rescue came from two sources. The first of these was the Scottish Philosophy of Common Sense which showed that there are mental phenomena that cannot be interpreted as any form of sensation and that " intelligence supposed principles, which, as *conditions* of its activity, cannot be the *results* of its operation" (*ibid*., 3). The other source of renewal was German philosophy after Kant, and in particular the Absolute Idealism that was "founded by Fichte, but evolved by Schelling" (*ibid*., 6). 'The Philosophy of the Unconditioned' is an examination of the most prominent French philosopher to make use of this second source – Victor Cousin – but this provides an occasion for Hamilton to formulate his own solution to the tension between the philosophy of common sense and the way of ideas.

The question at issue can be expressed in a number of different ways. Kant held that we can only have knowledge of phenomena, never of noumena or things in themselves. Clearly this version of phenomenalism, though in many ways the antithesis of empiricism, has elements in common with the 'way of ideas' to which Reid objected, which holds that the mind apprehends the world indirectly, through 'impressions'. The alternative position, referred to in the 19thcentury as 'presentationism' is often called 'direct realism' and holds, as Reid contends, that we directly apprehend the world of real things. Both positions have their difficulties. Those who followed Kant, notably Fichte and Schelling, sought to escape the 'scandal' of unknowable things-in-themselves, and those who followed Reid sought to overcome the contention implicit in his approach that our knowledge of the world is 'conditioned' by the principle of common sense. Hence the pursuit of a philosophy of the 'unconditioned'.

Hamilton's solution, ultimately, is to combine phenomenalism and presentationism. In 'The Philosophy of Perception' he engages in the debate by defending Reid against the criticism brought against him in Thomas Brown's postumously published *Lectures*, and in a very vigorous manner – "It is always unlucky to stumble on the threshold. The paragraph (Lect.xxvii) in which Dr Brown opens his attack on Reid contains more mistakes than sentences" (*ibid*., 69). Brown claimed that a close analysis of Reid's writings showed that his position on perception was not really that of direct realism but "hypothetical realism", that is the belief in an external world that cannot be known directly. It is this contention that Hamilton aims to refute, but it is arguable that he misinterprets Brown. Moreover in his notes to Reid's *Collected Works*, which were composed rather later, he appears to come round to something very like Brown's interpretation and to hold that Reid was not, strictly speaking, a direct realist after all.

If we construe Reid as holding that in the act of perception there are three elements – the physiological modification of the organ, a mental sensation and the perception of an object – then we can contrast this with Hamilton's position which holds that the mental sensation and the perception are simultaneous and in a sense two sides of the same coin. Reid holds, of course, that we do not reason from sensation to perception; the apprehending mind moves from one to the other by a natural, inbuilt instinct – one of the principles of common sense. Hamilton too holds that there is no reasoning process here, but he also thinks that the continuing division that Reid is employing between sensation and perception is incompatible with the idea of immediate perception or direct realism between. Hence his amendment, which so to speak ties the sensation and perception together. But how is this further contention to be sustained? Is it a conceptual truth of some kind, or an empirical observation about how the mind works? Hamilton's writings in general tend to assertion more than argument, and while he has a great deal to say on this point, it does seem that his 'solution' to the problem of perception is a an arbitrary stipulation designed to overcome it. At any rate, if we do press the question of its defence, we quickly encounter a new version of the old division, namely whether the perception is to be identified as a manifestation of self-evident principles of common sense, or as a psychological association of ideas. In this sense Hamilton's thesis is still set within the fundamental parameters of the Hume/Reid debate. It was the next major figure in 19th century Scottish philosophy – Hamilton's student and friend James Frederick Ferrier – who made the most strenuous effort to take a different tack.

3. James Frederick Ferrier (1808-1864)

It is a notable fact that the identification of 'Scottish philosophy' with 'Common Sense' is not one that the 18th century philosophers themselves made. Indeed, it was only in the 19th century that something called 'Scottish philosophy' came to self-consciousness, and only then that books with 'Scottish philosophy' in their titles began to appear. The most famous of these was James McCosh's encyclopaedic *The Scottish Philosophy* (1875), and perhaps the most insightful *Scottish Philosophy* (1885) by Andrew Seth Pringle Pattison But from the point of view of the century's principle philosophical debate, the most interesting is J F Ferrier's *Scottish Philosophy, the Old and the New*, (1854). This is because it was expressly written in defence of the contention that it is possible to engage in something called 'Scottish philosophy' while departing radically from the tenets of Reid, Stewart and so on. Ferrier writes with great force and feeling.

> It has been asserted, that my philosophy is of Germanic origin and complexion. A broader fabrication than that never dropped from human lips or dribbled from the point of a pen. My philosophy is Scottish to the very core; it is national in every fibre and articulation of its frame. It is a natural growth of old Scotland's soil and has drunk in no nourishment from any other land. Are we to judge the productions of Scotland by merely looking to what Scotland has hitherto produced? May a philosopher not be, heart and soul, a Scotsman – may he not be a Scotsman in all his intellectual movements, even though he should have the misfortune to differ in certain respects, from Dr Reid and Sir William Hamilton (Ferrier 1854: 12)

The explanation of the feeling with which Ferrier writes lies in the fact his little book is a response to the charge levelled against him in the contest for the Chair of Moral Philosophy at Edinburgh (then still in the gift of the Town Council), when he was accused by the Free Church party of departing from 'the Scottish philosophy' in favour of some sort of Hegelianism. This charge was almost certainly motivated by the ecclesiastical rivalries generated by the Disruption in the Church of Scotland that took place in 1843, but it is nonetheless true that Ferrier expressly denounces a certain conception of 'Common Sense' philosophy, and one which he identifies closely with Reid. Indeed he is not afraid to repeat his objections in his defence of himself.

> Suppose we are discussing the subject of salt, and that we say 'salt is white and gritty, it is in some degree moist, it is sometimes put into a salt cellar and placed on the dinner table …' … No man would be considered much of a chemist, who was merely acquainted with these and other such circumstances, concerning salt … . So, in philosophy, no man can be called a philosopher who merely knows and says, that he and other people exist, that there is an external world, that a man is the same to day as he was yesterday, and so forth. These are undoubtedly truths, but I maintain that they are not truths in philosophy, any more than those just mentioned are truths in chemistry. Our old Scottish school, however, is of a different way of thinking. It represents these and similar facts as the first truths of philosophy, and to these it has recourse in handling the deeper questions of metaphysics. I have no objections to this, for those who like it -- only my system deals with first truths of a very different order; and it denies that the first truths of the old Scottish school are truths

> in philosophy at all. This is one very fundamental point of difference between the old and the new Scottish system of metaphysics (*ibid*., 7)

It is important to note that Ferrier thinks this castigation of one version of ‘Common Sense’ philosophy is quite compatible with claiming the right to be the inheritor of, though not restricted by, the programme of Reid and Hamilton. And there are indeed several points of contact to be observed. The first is this. Ferrier shares with the school of Reid and Hamilton an almost unspoken assumption that the question of mind and world lies at the heart of philosophy. In this they all differ from the alternative conception of moral philosophy as social inquiry, which as we have already noted, is to be found in Ferguson, parts of Hume, and above all Adam Smith. Second, and more importantly perhaps, Ferrier’s own philosophical reflections continue to fit Davies description of Scottish philosophy as a ‘dialogue between the vulgar and the learned’.

Ferrier’s reputation rested upon an earlier series of essays on *The Philosophy of Consciousness* which appeared in Blackwood’s Magazine between 1838 and 1843. In these essays he took his stand on the contention that consciousness implies the impossibility of a naturalistic science of mind, and in a later essay robustly defends a version of Berkeleyan idealism. While Reid thought that Berkeley’s philosophical position was one that "no man in his senses could believe", somewhat surprisingly perhaps, Ferrier describes Berkeley as "the champion of common sense . . . who could have foiled the prince of skeptics at his own weapons" (Ferrier 1865: 301). "Among all philosophers ancient or modern, we are acquainted with none who presents fewer vulnerable points that Bishop Berkeley. His language it is true, has sometimes the appearance of paradox; but there is nothing paradoxical in his thoughts, and time has proved the adamantine solidity of his principles." (*ibid*., p.291) By Ferrier’s account Berkeley settles the issue of sensation and perception with which Hamilton struggled, by seeing that there is a false abstraction here.

> The external world *in itself*, and the external world in relation *to us*, was a philosophic distinction which he [Berkeley] refused to recognize. In his creed, the substantive and phenomenal were one. And though he has been accused of sacrificing the substance to the shadow, and though he still continues to be charged, by every philosophical writer, with reducing all things to ideas in the mind, he was guilty of no such absurdity … There does not appear to be much justice in the ordinary allegation, that Berkeley discredited the testimony of the senses, and denied the existence of the material universe. He merely denied the distinction between things and their appearances, and maintained that the thing *was* the appearance and the appearance *was* the thing. (*ibid*., 302-3)

On this interpretation Berkeley espouses a sort of idealism but

> genuine idealism, looking only to the fact, and instructed by the unadulterated dictates of common sense, denies … that we can separate in thought objects and perceptions *at all*; hence this system has nothing whatever to do either with the preservation or the destruction of the material universe; and hence, too, it is identical … with genuine

unperverted realism. (*ibid*., 309)

In this way Ferrier, despite his disagreements, actually concurs with Reid's strictures on the kind of philosophical theorizing that tries to deploy Newtonian methods in the way that Hume does. Indeed, Ferrier thinks that "the inert and lifeless character of modern philosophy is ultimately attributable to her having degenerated into a physical science" (*ibid*., 191), and he condemns the resulting "picture of man" as "a wretched association machine, through which ideas pass linked only by laws over which the machine has no control" (*ibid*., 196). His alternative to this externalist conception of 'the science of mind' is a return to the introspective examination of human consciousness. "Consciousness is philosophy nascent; philosophy is consciousness in full bloom and blow. The difference between them is only one of degree, and not one of kind; and thus all conscious men are to a certain extent philosophers, although they may not know it" (*ibid*., 197) In short, the proper engagement of philosophy is a matter of bringing consciousness to a better understanding of itself, which is at least one interpretation of the ambition of Reid's *Inquiry*.

Ferrier's philosophy, then, constitutes a further excursion in the common sense tradition, but one which sets itself at some considerable distance from Reid. For Reid, Berkeley is one of the chief architects of 'the way of ideas', and hence though not himself a skeptic, the purveyor of a philosophy that makes radical skepticism inevitable. In sharp contrast, for Ferrier, Berkeley's philosophy (with some additions of Ferrier's own) is the *answer*to skepticism. It hardly needs to be said that this was a highly controversial position. Moreover, it throws the whole subject of mind and consciousness back into the realms of metaphysical philosophy and hence seems to abandon the shared methodological assumption that, to quote Reid again, "there is but one way to the knowledge of nature's works; the way of observation and experiment" a supposition he wholeheartedly shared with Hume. This implication – that the methods of the sciences inapplicable to philosophy -- somewhat isolated Ferrier within Scottish philosophy. Though he was regarded with great acclaim in continental Europe, Scottish philosophers moved in different directions, some to an intensification of the experimental method, and some to Absolute Idealism. Of the first group, the most prominent and influential was Alexander Bain.

4. Alexander Bain (1818-1903)

Alexander Bain was Regius Professor of Logic at the University of Aberdeen from 1860 to 1880. A man of remarkable gifts, he was appointed to the Chair largely on the strength of distinguished philosophical work he had published while working as a journalist in London. *Dissertations on Leading Philosophical Question*s(1903), is a collection of his essays published in retirement, though almost all had originally appeared in the journal *Mind*, a journal he was instrumental in founding, In several of these essays, Bain takes Reid and Hamilton as his starting point and, broadly, follows the same methods. But he pushes them in a much more strongly empirical direction. The most interesting of his *Dissertations*, in this connection, is entitled 'Associationist Controversies' and at the heart of these controversies we can find a distinction between philosophy and psychology which both reveals the significant difference between Bain and Ferrier, and establishes the discipline of experimental psychology in its own right.

> We are, at the moment, in the midst of a conflict of views as to the priority of Metaphysics and Psychology. If indeed the two are closely identified as some suppose, there is no conflict; there is in fact, but one study. If, on the other hand, there are two subjects, each ought to be carried on apart for a certain length, before they can either confirm or weaken each other. I believe that in strictness, a disinterested Psychology should come first in order, and that, after going on a little way in amassing the facts, it should revise its fundamental assumptions … I do not see any mode of attaining a correct Metaphysics until Psychology has at least made some way upon a provisional Metaphysics (Bain 1903: 38)

Bain can be interpreted as a practitioner of the 'science of mind' no less than Reid or Hume. But whereas in Reid and Hume the distinction between philosophy and psychology as the modern world understands it, was unclear, it is one of Bain's chief claims to enduring significance that, as this quotation reveals, he brought the distinction between psychological and metaphysical questions to prominence, and in what we would call his research programme he gave priority to the former. The conclusion to be drawn is that Bain, like Ferrier, can be seen to stand in the tradition of Scottish philosophy in the sense that he adopted its methods. But in contrast to Ferrier, he did so in ways that further removed the question of sensation and perception from the realms of traditional metaphysics, and pressed the study of the mind in the direction of empirical psychology.

One notable feature of this development lies in the fact that Bain was one of the principal exponents and defenders of 'associationism', whose origins, arguably, are to be found most clearly in Hume's *Treatise*. Associationism is the application of empirical observation to the relation between ideas and experiences. What it seeks is observed regularities, in the hope of formulating psychological laws that will enable us to order the contents of mind. Two such principles -- Contiguity and Similarity -- were widely accepted, and identified by Bain as being employed by Reid and Hamilton. A third -- Contrast -- was more disputable, and in 'Associationist Controversies' Bain is principally concerned with the nature and identifiable independence of principles such as these.

However, for present purposes his arguments are interesting chiefly not so much for their elaboration of associationism, but for the light they throw on the development of Scottish philosophy in the nineteenth century. One point in particular seems to me illuminating. In the dispute between Reid and Hume with respect to the operations of the mind one of the fundamental points of difference is this. Reid is trying, in the main, to establish basic principles of the mind's operation which will vindicate its rationality, and hence avoid the depths of skepticism into which Hume's account forces it. Hume, on the contrary, declares that "reason is nothing but a wonderful and unintelligible instinct in our souls which carries us along a certain train of ideas … [and that this] habit is nothing but one of the principles of nature, and derives all its force from that origin" (Hume 1888 : 179), Reid's purpose is precisely to show that the basic operations of the mind are those of intelligibility. Now in terms of this difference, Bain is of Hume's persuasion. This is revealed not merely in his striking deployment of decidedly Humean terminology when, for instance, he contrasts the perception and the memory of a thing in terms of 'vividness' (Bain 1903: 42). It is even more evident when he asserts that "The flow of representations in dreaming and madness offers the best field of observation for the study of associations as such" (*ibid*., 45).

What this remark reveals is that Bain is interested first in establishing empirical laws with respect to the contents of the human mind. The reason that he thinks dreaming and madness are the best places to start is precisely because he sees that the pursuit of rational principles, that is to say, philosophically coherent principles, is likely to distort our observation by inclining us to see rational connections rather than empirical associations, or as he puts it "associations as such." In this respect he is employing Hume's rather than Reid's conception of human nature. Certainly he reserves judgement on the final outcome of these investigations with respect to philosophy, arguing only for the priority of psychology over metaphysics and not, as Hume may be said to do, for the elimination of the second by the first. But so far as the science of mind that had been such a marked feature of Scottish philosophy goes, Bain clearsightedly pursues its more empirical ambitions.

For Ferrier the empirical laws of association that Bain seeks are not 'truths in philosophy'. No one can be called a philosopher who merely knows and says, that in dreaming or madness this mental representation tends to be associated with that. The philosopher aspires, rather, to make sense of experience, and the whole point about the experience of the dreamer or the madman is that no sense is to be made of it. By contrast, the empirical psychologist, seriously committed to the experimental method, does not, in the end, render consciousness intelligible; he or she simply describes how the mind works.

With Ferrier and Bain, then, the tension within Scottish philosophy that Davie has identified is resolved in radically different ways, the first by a return to metaphysics, the second by an advance to psychology. Both can claim to be inheritors of the Scottish tradition, but both in their different ways may be said to have brought about its demise. With Bain, the nature of the demise is evident; the philosophy of mind is replaced by empirical psychology. With Ferrier, the nature of the demise is rather different. Faced with the prospect of returning to Berkeleyan metaphysics, several prominent Scottish philosophers preferred to look elsewhere, namely to Germany and Hegel. The result was that as the century ended a small group known as the Glasgow Idealists came to prominence.

5. The Scottish Idealists

In his illuminating study *Scottish Philosophy*, importantly subtitled *A comparison of the Scottish and German answers to Hume*, Andrew Seth Pringle Pattison remarks:

> The thread of national tradition, it is tolerably well known, has been but loosely held of late by many of our best Scottish students of philosophy. It will hardly be denied that the philosophical productions of the younger generation of our University men are more strongly impressed with a German than with a native stamp (Seth Pringle Pattison 1885:1-2)

Pringle Pattison does not say who it is he has in mind, but a knowledge of the period makes it relatively easy to guess. There is a danger, though, that the reference to Germany be somewhat misleading. An interest in, and a knowledge of, Kant can be found to go back to Hamilton, and far from being regarded

as a threat to the Scottish tradition was recognized (by Veitch, for instance) as an important part of its enrichment. The German philosophy referred to here, then, is that which emanated from Hegel.

The Secret of Hegel is the title of a very large book by James Hutchison Stirling, first published in 1864. Stirling is credited with bringing Hegel to the attention of British (and not just Scottish) philosophy for the first time, though a wit at the time remarked that if Stirling did know the secret of Hegel, he had kept it to himself! Though Stirling was, in modern terms, a layman (he held no university post) the book was well received, and it is a matter of some consequence that it contained significant criticism of Hamilton. In fact, Stirling subsequently published a short but highly critical volume entitled *Sir William Hamilton: being the philosophy of perception. An analysis* (1865). With these two books we can chart the diminishing interest in and influence of the Common Sense tradition within Scottish philosophy and the increasing influence of German Idealism and Hegel in particular.

Of the Scottish Idealists the most prominent and influential was unquestionably Edward Caird (1835-1908). A graduate of the University of Glasgow, after a period at Oxford he returned in 1866 to become Professor of Moral Philosophy at Glasgow, a post he held for almost 30 years, before returning to Oxford to become Master of Balliol. Caird was an admirer of Kant, who believed nevertheless that Kant had failed to capitalize fully on his own insights, and that the full import of his philosophy could be uncovered with the help of Hegel. The aim of philosophy, on this interpretation, was the ultimate reconciliation of seemingly incompatible elements in human experience – religion and science, freedom and causality, reason and desire, for instance. These things are made compatible by reconception, and what makes Caird’s view a version or application of Hegelianism is the idea that our knowledge of objects is perfected as we more adequately conceive of them as parts of a whole.

Caird has been generally regarded as a something of a fifth columnist as far as Scottish philosophy was concerned. George Davie describes him as "a very untypical Scotsman and one quite exceptionally apathetic to educational customs of the country" (Davie 1961: 86). However, this remark reflects the fact that over the course of the nineteenth century Scottish philosophers were concerned not only with philosophical debates of the kind reviewed here, but with the place of philosophy in the university curriculum. As a result, from Hamilton onwards, several of them wrote essays on educational reform and gave evidence to the many commissions of inquiry into the universities that were held. Caird’s indifference to the national tradition of philosophical education, if that is what it was, was simply the other side of his desire to bring Scottish philosophers into the wider context of contemporary European philosophy which was, of course, dominated by German Idealism. That he was neither alone nor unsuccessful in this ambition is evidenced by the fact that he inspired and recruited several other distinguished figures, most notably Sir Henry Jones, who also came to occupy the Chair of Moral Philosophy at Glasgow, and J H Muirhead, founder of the enduring Muirhead Library of Philosophy, a long series of major philosophical works published in London.

Nevertheless, it is true that between Caird and the Scottish (or Glasgow) Idealists on the one hand, and the tradition known as ‘The School of Common Sense’ there is such a marked difference, that the former can hardly be said to be a continuation of the latter, but rather to signal the end of a philosophical project

which had lasted the larger part of 200 years. The 19th century, in short, may be said to be the century in which a distinctively Scottish tradition in philosophy came to an end.

Bibliography

- Cross, R. C. (1971) 'Alexander Bain',*Aberdeen University Review* 44 pp 1-9
- Davie, George (1961) *The Democratic Intellect*, Edinburgh University Press, Edinburgh
- ----- (1994) *A Passion for Ideas: Essays on the Scottish Enlightenment Vol. II*, Polygon, Edinburgh
- ----- (2001) *The Scotch Metaphysics: A Century of Enlightenment in Scotland*, Routledge, London and New York
- Ferrier, J. F. (1875) *Philosophical Works of Ferrier*, Blackwood, Edinburgh
- Ferrier, J. F. (1856) *Scottish Philosophy, the Old and the New*, Edinburgh and London
- Hamilton, William (1853) *Discussions on Philosophy and Literature, Education and University Reform*, MacLachlan and Stewart, Edinburgh
- McCosh, James (1875) *The Scottish Philosophy, Biographical, Expository, Critical, from Hutcheson to Hamilton*, MacMillan and Co., London
- Haldane, Elizabeth S (1991) *James Frederick Ferrier*, Thommes Books Bristol. Reprint of 1899 edition with a new introduction by John Haldane
- Hume, David (1888) *A Treatise of Human Nature*, L. A. Selby-Bigge, Clarendon Press, Oxford
- Kant, Immanuel (1951) *Prolegomena to any future metaphysic* trans. L W Beck, New York
- Reid, Thomas (1997) *An Inquiry into the Human Mind on the Principles of Common Sense* (a critical edition, edited by Derek R. Brookes), Edinburgh University Press, Edinburgh
- Seth, Andrew (1899) *Scottish Philosophy: A Comparison of the Scottish and German Answers to Hume* (third edition), Blackwood, Edinburgh and London
- Stirling, James H. (1865) *Sir William Hamilton: The Philosophy of Perception (an analysis)*, Longmans, Green and Co., London
- Thomson, Arthur (1985) *Ferrier of St Andrews: an academic tragedy*, Scottish Academic Press
- Veitch, John (1882) *Hamilton*, Blackwood, Edinburgh and London.

Other Internet Resources

- Entry on William Hamilton, (Internet Encyclopedia of Philosophy, J. Fieser, ed., U. Tennessee)
- Entry on Frederick Ferrier, (Internet Encyclopedia of Philosophy, J. Fieser, ed., U. Tennessee)
- Entry on Edward Caird (Internet Encyclopedia of Philosophy, J. Fieser, ed., U. Tennessee)
- Chapter on Alexander Bain, in *Mind, Brain and Adaptation in the Nineteenth Century*, by Robert M. Young

Related Entries

Hume, David | Reid, Thomas

Gordon Graham
g.graham@abdn.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 29, 2002
Content last modified: January 29, 2002

Robert Boyle

Boyle was one of a leading intellectual figures of the seventeenth century. He was a dedicated experimenter, unwilling to construct abstract theories to which his results had to conform. "Our Boyle," Oldenburg wrote to Spinoza, "is one of those who are distrustful enough of their reasoning to wish that the phenomena should agree with it" (Hall & Hall 1965-1977, 2:38). Boyle, though a champion of the corpuscularian doctrine, preferred simply to report the results of his experiments, including negative results, and frequently lamented the fact that we lacked "histories" (collections of experimental results and accurate observations) in various fields of scientific endeavour. He performed so many experiments that he was able, at one "time to loose … at once near five centuries of Experiments of my own" (BP 9:28). Nor was this an isolated loss; nonetheless the number, variety and scope of his experiments were such that he carried on working and publishing with no particular difficulty. "His books," as Huygens remarked to Leibniz immediately after Boyle’s death, "are full of experiments" (Huygens 1888, 10:239). Moreover, experiments were *exactly* what he was interested in, he had a certain missionary zeal in spreading the corpuscularian gospel, but he was not himself interested in detailed system building, a fact that was commonly noted. Leibniz told Huygens that he was "astonished" that Boyle "who has so many fine experiments, [had] not come to some theory of chemistry after meditating so long on them. Yet in his books, and for all the consequences that he draws from his observations, he concludes only what we all know, that everything happens mechanically. He is perhaps too reserved. Excellent men should leave us even their conjectures; they are wrong if they wish to give us only those truths that are certain" (Leibniz to Huygens, Dec. 29, 1691, in Huygens 1888, 10:228). Boyle was ahead of his time. In the next century d’Alembert wrote "the taste for systems … is today almost entirely banished from works of merit … a writer among us who praised systems would have come too late" (d’Alembert 1751, 94).[1]

Boyle was a corpuscularian, a term he employed to paper over the differences between believers in a vacuum, and believers in a plenum, given that both of them agreed that the explanation of natural occurrences should be solely in terms of particles of matter, their motion and interaction. Boyle consistently refused to pronounce on the question of whether these *minima naturalia* should be considered *atoms*, in the strict sense of that term, or not.

Even a metaphysical non-corpuscularian such as Leibniz agreed with Boyle in practical terms. "However much I agree with the Scholastics in this general and, so to speak, metaphysical explanation of the principles of bodies," he wrote to Arnauld in July 1686, "I am as corpuscular as one can be in the explanation of particular phenomena, and it is saying nothing to allege that they have forms or qualities" (Gerhardt 1875, 2:58, trans. Mason 1967).

Boyle's scientific range was wide. Besides his well known work in mechanics, medicine, hydrodynamics and a wide variety of experiments with his vacuum pump, he was interested both theoretically and practically in alchemy (see Principe 1998), where his interest seems to have been fueled more by his constant desire to acquire knowledge of God and the world than by any desire for riches. He "cultivated Chymistry with a disinterested mind," seeking the improvement of his own knowledge, "*the* gratifying the Curious & the Industrious; and *the* Acquist of some useful helps to make good & uncommon Medicins."[2] As a corpuscularian he believed that transmutation was physically possible. As a person he believed that it actually occurred. He believed that in his own laboratory gold had been transmuted into a "baser metal" with a specific gravity about two thirds of that of gold, and since he was, he said, more interested in the luciferous aspect of discovery than the lucriferous, he found the process equally interesting in either direction.[3]

During the course of his life he sought constantly to improve the lot of humanity. He was interested in the improvement of agricultural methods, in the possibility of extracting fresh water from salt, in the improvement of medicines and medicinal practice, in the possibility of preserving food by vacuum packing, and in a number of other useful results, actual or potential, of experimental philosophy. He viewed his theological interests and his work in natural philosophy as forming a seamless whole and constantly used results from the one area to enlighten matters in the other.

Convinced that Christianity was the religion instituted by God, Boyle was concerned that the Bible should be widely promulgated and he devoted time and energy to having it translated into a variety of languages such as Irish, Turkish, and various native American languages. He viewed such conversion attempts as being on all fours with his attempts to find more efficacious medicines for:

> To convert Infidels to the Christian Religion is a work of great Charity and kindnes to men. I. In regard of the evills it frees them from, such as, (1) the gross errors and prejudices they had entertain'd before they were instructed in it. (2ly) The vices and polutions they securely liv'd in, before they receiv'd the Gospel; *some* of which were unworthy of men as such; *others* very prejudicial to humane society's; and *others* very mischievous to the vicious persons themselves; and *others* <againe> great hinderances to the discovery and reception of usefull and noble truths. (3ly) The unexpressible Infelicitys that attend the greatest part of such Infidels & wicked Persons, in the future state.
>
> II. The Christian Religion brings mankind diverse positive Benefits, such as are, more cleare and extensive knowledg of God, and divine things; the Remission of Sins; the Favour of God; severall graces and vertues suitable to mens respective needs and conditions; and above all, a happy Immortality in the Life to come. (Boyle Papers [BP] 5:73-4)

---

1. Life[4]

Robert Boyle was born in Lismore, in Ireland, on the January 25, 1627. He was his parents' fourteenth, penultimate, child, and the last to survive to adulthood. Boyle was the youngest son and, after his sister Margaret died when he was 10, the youngest child of the family. Boyle speaks fondly of his parents, but he could not have known them well. His mother died in childbirth a few weeks after Boyle's third birthday, and he last saw his father just before he and his brother Francis left for a continental tour when Boyle was twelve.

Like many children Boyle had his share of near escapes from serious injury as a child, but the time and Boyle being what they were he saw in each of them the hand of God. Michael Hunter has pointed out that "the spiritual autobiography, aimed at chronicling God's purpose for the individual in question by recounting providential escapes, spiritual trials and conversion experiences," was "a characteristic genre of autobiographical writing in seventeenth-century England" (Hunter 1994a, xx), and Nicolas Canny notes that "The invocation of providence as an explanation for accidental or chance happenings in this life was so commonplace among sincere Protestants in the early seventeenth century that it had come to be considered irreverent or profane not so to attribute them" (Canny 1982, 28). But, though common enough, there is no doubt that, in Boyle's case at least, the protestations were sincere. He continued to believe in this divine attention, though in a more intellectual realm, throughout his life. In 1663 he wrote,

> And though I dare not affirm … that God discloses to Men the Great Mystery of Chymistry by Good Angels, or by Nocturnal Visions … yet perswaded I am, that the favor of God does (much more than most Men are aware of) vouchsafe to promote some Mens Proficiency in the study of Nature (Boyle Works [BW], 3:276, Birch 1772, II:61).[5]

In Geneva on his continental tour Boyle underwent what he clearly felt to be a conversion from nominal or at least unthinking Christianity to committed Christianity. One summer night, he "was suddenly waked in a Fright with such loud Claps of Thunder … & every clap … both preceded & attended with Flashes of lightning so numerous … & so dazling, that [he] began to imagine … the Day of Judgment's being come." This led him to vow that "all further additions to his life shud be more Religiously & carefully employ'd." Realizing the inefficacy of a promise exacted under duress, Boyle repeated the performance under a serene and cloudless sky "so solemnly that from that Day he dated his Conversion; renewing now

he was past Danger, the vow he had made whilst he fancy'd himselfe to be in it: that tho his Feare was (& he blush't it was so) the … occasion of his Resolution of Amendment; yet at least he might not owe his more deliberate consecration … of himselfe to Piety, to any less noble Motive then that of it's owne Excellence" (BP 37:181r-v).

The promise seems never to have been broken, and indeed the later Boyle stressed the need to have an *examined* faith. He pointed out that "usually, such as are born in such a place, espouse the opinions true or false, that obtain there" (BW, 12:421, Birch 1772, VI:712), indeed, "the greatest number of those that pass for Christians, profess themselves such only because Christianity is the religion of their Parents, or their Country, or their Prince, or those that have been, or may be, their Benefactors; which is in effect to say, that they are Christians, but upon the same grounds that would have made them Mahometans, if they had been born and bred in Turky" (BP 7:233). Boyle felt that more was required of the thinking believer. Locke agreed: often a child's notion of God does more "resemble the Opinion, and Notion of the Teacher, than represent the True God" (Essay, 1.14.13).

Hard on the heels of Boyle's enlightenment, doubts about his faith began to trouble him,[6] and these "distracting Doubts of some of the Fundamentals of Christianity" continued: "never after did these fleeting Clouds, cease now & then to darken … the clearest serenity of his quiet: which made him often say that Injections of this Nature were such a Disease to his Faith as the Tooth-ach is to the Body; for tho it be not mortall, 'tis very troublesome" (BP 37:182r).

Leaving Switzerland, Boyle, along with Marcombes and his brother crossed the Alps and entered Italy in September 1641 where, in Florence, he spent the winter. "In *Italy* he read over the lives of the ancient philosophers with the utmost attention," presumably in Diogenes Laertius, and "[t]he sect, which then struck him most, was that of the Stoics; and he tried his proficiency in their philosophy, by enduring a long fit of the tooth-ach with great unconcernedness."[7] Still in Italy he had (in the winter of 41-42) what seems to have been one of the very few sexual encounters of his life. Writing about himself in the third person as Philaretus (sometimes P., or Filaretus) he says:

> Nor did he sometimes scruple, in his Governor's Company, to visit the famousest Bordellos; whither resorting out of bare Curiosity, he retain'd there an unblemish't Chastity, & still return'd thence as honest as he went thither. Professing that he never found any such sermons against them, as they were against themselves. The Impudent Nakednesse of vice, clothing it with a Difformity, Description cannot reach, & the worst of Epithetes cannot but flatter. But tho P. were noe Fewell for forbidden Flames, he prov'd the Object of unnaturall ones. For being at that Time not above 15, & the Cares of the World having not yet faded a Complexion naturally fresh enuf; as he was once unaccompany'd diverting himselfe abroad, he was somewhat rudely presst by the Preposterous Courtship of 2 of those Fryers, whose Lust makes no distinction of Sexes; but that which it's Preference of their owne creates; & not without Difficulty, & Danger, forc't a scape from these gown'd Sodomites. Whose Goatish Heates, serv'd not a little to arme Filaretus against such Peoples specious Hyprocrisy; & heightn'd & fortify'd in him an Aversenesse for Opinions, which now the Religieux discredit as well as the Religion (BP

> 37:184r-v).

Leaving Italy Marcombes and the two boys found on arrival in Marseilles that the monies the Great Earl had been in the habit of sending were no longer to arrive and that, indeed, the last quarter's payment had been held up by Cork's London agent. Moreover, there was a letter from the Earl, unaware of the mischance affecting the quarterly payment, telling them that, as a result of the rebellion in Ireland, no more money was to be forthcoming: in the "dangerous and poore estate whereunto by gods providence" he had been reduced, he had "with much difficulty gott together two hundred and fifty pounds by selling of plate," but to pay Marcombes' bills punctually as he had in the past "I am noe waies able." So he advised Marcombes to use the money to bring the two boys

> out of some meet port in France to land either at dublin, Corke, or Youghall, (for all other Cities and Sea townes are possessed by the enemy), or else my two sonnes [must] travaile into Holland, and putt themselves into entertaynement under the service and conduct of the Prince of Orange; for they must henceforward maintayne themselves by such entertaynements as they gett in the warres (Earl of Cork to Marcombes 9 March 1641/2, Maddison 1969, 47).

In the event Francis decided to return to Ireland, arriving in time to fight in the Battle of Liscarrol (September 3, 1642), at which another Boyle brother, Lewis, was killed. Meanwhile Robert decided that his health and lack of money ruled out a return to Ireland, and his age made soldiering in Holland an untempting and indeed implausible prospect.[8] He therefore decided to accept Marcombes' offer of hospitality in Geneva, and did not make his way to England until the summer of 1644.

Before leaving Geneva Boyle had a conversation with François Perreaud (1572-1657), who later wrote *Démonographie, ou traité des démons*, which Boyle then arranged to have translated into English by Peter du Moulin (the younger, 1601-1684). In a letter prefixed to the English edition Boyle recalled that "the conversation I had with that pious author during my stay at Geneva, and the present he was pleased to make me of this treatise before it was printed, in a place where I had opportunities to enquire both after the writer, and some passages of the book, did at length overcome in me (as to this narrative) all my settled indisposedness to believe strange things." Acceptance of at least the *possibility* of diabolic or angelic intervention was common among the intelligentsia in the second half of the seventeenth century. Cudworth pointed out one expedient reason for the belief:

> all these *Extraordinary Phænomena*, of Apparitions, Witchcraft, Possessions, Miracles, and *Prophecies*, do Evince that *Spirits, Angels or Demons*, though *Invisible* to us, are no *Phancies*, but *Real and Substantial* Inhabitants of the World; which favours not the *Atheistick Hypothesis;* but some of them, as the Higher kind of Miracles, and Predictions, do also immediatly enforce the acknowledgment of a *Deity*: a Being superiour to *Nature*, which therefore can check and controul it; and which comprehending the whole, foreknows the most *Remotely distant*, and *Contingent Events* (Cudworth 1678, 715).[9].

In a manuscript draft ("Loose papers whence some things are to be extracted for the Discourse of the causes of Atheism") Boyle considered three objections that might be made against such a belief: the implausibility of the standard means of bringing about such intervention; the unreliability of the witnesses; and the impossibility of incorporeal beings interacting with matter. He agrees that the first objection, "urg'd with great confidence, and not without much show of Reason" is a strong one, but suggests that "we men understand very little of the nature, customes, & government of the Intelligent creatures of the spirituall world: and particularly what concernes the Falne Angells or bad Daemons. And therefore they being themselves invisible to us, and capable of working in wayes that our sences cannot discerne; and being Agents of great craft & long experience; tis no wonder that many of their actions, thô never so pollytickly contrived & carried on, should seem irrationall to us: who know so little of their particular inclinations & designes, and the subtil & secret methods in which they carry them on."

The second objections he also accepts, though not wholly: "thô upon particular & cogent proofe I beleeve some of them to be true … yet I reject or distrust far the greatest part, as not being soe attested." The third he rejects as being simply inconsistent, for the human soul is accepted as incorporeal, and it works (though we know not how) on matter (BP 2:105).

When Boyle arrived back in England in mid-1644 at the age of 17 he was quickly reunited with his sister Katherine who seems immediately to have re-adopted the semi-maternal role she had no doubt often played after the death of their mother. She was concerned in a variety of other ways to look after his welfare, both spiritual and worldly. She was, for example, the immediate cause of his getting to know the members of the Hartlib circle.[10]

At this time Boyle settled in Stalbridge (on an estate left to him by his father) and occupied himself mainly in writing or planning works in ethics and theology. Much of his time during the early part of this Stalbridge period was spent in moral philosophy -- "My Ethics go very slowly on," he wrote to his sister Katherine on March 30, 1646 (Birch 1772, I:xxx) -- and there was at this stage no reason to think that he would become one of the great natural philosophers of his time. (See further, Hunter 1995b.) He already approximated to the "lay-bishop" that Aubrey was later to find him to be. He was, in fact, a serious, somewhat priggish young man, though he often gave signs of light-heartedness both as a boy and in later life. After his death Gilbert Burnet claimed that "As for Joy, he had indeed nothing of Frolick and Levity in him," a judgement accepted by Steven Shapin, but this fails to allow for the lighter moments that Boyle undoubtedly enjoyed.[11]

At Stalbridge, about 1649, Boyle began to be interested in experimenting, but was hindered by the fact that he could not obtain a furnace. Stalbridge was far enough away from tradespeople who could make such an item and the furnaces Boyle ordered tended to arrive "crumbled into as many pieces, as we into sects," leaving Boyle to attempt "such experiments, as the unfurnishedness of the place, and the present distractedness of my mind, will permit me" (Birch 1772, I:xxxxvi; xli).

Boyle was troubled throughout his life by the fragmentation of Christianity. Among "the giddy multitude … this multiplicity of religions will end in none at all," he wrote to John Mallet in 1652,[12] and at the

very end of his life he expressed in his Will the wish that the Boyle lecturers should, when "proveing the Christian Religion against notorious Infidels (*viz*), Theists, Pagans, Jews and Mahometans, not [descend] lower to any Controversies that are among Christians themselves."[13]

Eventually, however, a furnace did arrive, and Boyle found himself "so transported and bewitched [as to] fancy my laboratory a kind of Elysium … . I there forget my standish and my books, and almost all things" (Boyle to Katherine Ranelagh, Aug 31, 1649, Birch 1772, I:xlv).

Boyle was never a student at a university. Nor was he ever a fellow of an Oxford College, though that too has been claimed on his behalf (Dutton 1951, 20), but it was to Oxford that he removed after his time at Stalbridge, and it was there that his interest in natural philosophy first flowered. Before taking up residence in Oxford however he paid two lengthy visits to Ireland during the early 50s (for a year from June 1652, and then for eight months from Oct 1653), and it was from that "illiterate country" that he wrote to Clodius, probably toward the end of his second Irish visit, in the spring of 1654:

> For my part, that I may not live wholly useless, or altogether a stranger in the study of nature, since I want glasses and furnaces to make a chemical analysis of inanimate bodies, I am exercising myself in making anatomical dissections of living animals: wherein (being assisted by your father-in-law's friend Dr *Petty*, our general's physician) I have satisfied myself of the circulation of the blood, and the (freshly discovered and hardly discoverable) receptaculum chyli, made by the confluence of the venae lactae;[14] and have seen (especially in the dissections of fishes) more of the variety and contrivances of nature, and the majesty and wisdom of her author, than all the books I ever read in my life could give me convincing notions of (Birch 1772, VI:55).

It was also during this period, no doubt in large part due to Cromwell's extremely harsh treatment of the Irish, that Boyle's Irish properties were made secure and began returning rents to him, ultimately reaching almost £3000 p.a., Hooke told Aubrey. The fact that Boyle's friend Petty conducted the survey on which the disposal of the lands was based can hardly have been to Boyle's disadvantage.

On October 12, 1655, Katherine was in Oxford investigating the suitability of possible lodgings for Boyle. He was to lodge with the apothecary John Crosse, whom Birch felt worthy of mention because he had "a great acquaintance with Dr. *John Fell*,"[15] and the question was, which was the best room for his purposes, and how was it to be furnished?

> My Brother,
> It has pleased God to bring us safe to *Oxford*, and I am lodged at Mr. *Crosse*'s, with design to be able to give you from experience an account which is the warmest room; and indeed I am satisfied with neither of them as to that point, because the doors are placed so just by the chimnies, that if you have the benefit of the fire, you must venture having the inconvenience of the wind, which yet may be helped in either by a folding skreen; and then I think that which looks into the garden will be the more comfortable… (Birch 1772,

VI:523).

The house in question stood on the site where the Shelley Memorial now stands, and his two rooms there seem to have served Boyle admirably, though he later set up a retreat at Stanton St John's, where he could retire when the press of society grew too great in Oxford.

In Oxford Boyle's tremendous output of works in philosophy, theology, and experimental philosophy began. It was here that he published *New Experiments Physico-Mechanical, Touching the Spring of Air and its Effects, Certain Physiological Essays, The Sceptical Chymist, Some Considerations touching the Usefulness of Experimental Natural Philosophy*, and a number of others including *The Origine of Forms and Qualities.*

Boyle's years in London (from 1668 to his death) saw the continuation of his experimental work, along with a number of works on philosophy and theology, including *The Excellency of Theology, Compar'd with Natural Philosophy*, *Considerations About the Excellency and Grounds of the Mechanical Hypothesis*, the *Free Enquiry into the Vulgarly Receiv'd Notion of Nature*, the *Discourse of things above Reason*, *Disquisition about the Final Causes of Natural Things*, and *The Christian Virtuoso*.

In 1691 Katherine died, and Boyle, whose health throughout his life had been poor, died the following week.

2. Religious Views

The seventeenth century is notable not only for the number but also for the variety of arguments which were offered to prove God's existence. Although writers such as Pascal and Bayle felt such arguments to be both unnecessary and unavailable, demonstrations of God's existence were felt by many to be not only possible but desirable, since they were necessary in the fight against atheism.[16]

Descartes, as is well known, felt that God's existence could be, and epistemologically had to be, demonstrated, and offered a variety of proofs to provide such a demonstration. His version of the ontological argument, his proof from the supposed innate idea of God, his proof from the need for an eternal conscious being,[17] and his proof from the need for continuing creation, all found supporters in the later seventeenth century, though the first two were generally held to be unlikely to convince anyone.

Apart from a brief reference to it in the printed works Boyle does not mention the ontological argument (BW 9:413; Boyle 1772, 4:461-2). This distancing was not uncommon at the time. Ralph Cudworth, though clearly fascinated by the ontological argument, recognized that most would "Distrust, the *Firmness* and *Solidity* of such *thin* and *Subtle Cobwebs*," and offered an alternative argument in the hope that it would prove more "Convictive of the *Existence* of a God to the Generality" (Cudworth 1678, 725).

The language of the ontological argument was acceptable, even when the argument's validity was

rejected. Gassendi, for example, agreed that God is that than which nothing greater can be conceived, but denied the validity of the argument which offers this as its main premise (See further Osler 1994, ch. 2).

Nor was the argument from innate ideas more popular. After his dialogue character Cuphophron has espoused it, Henry More has his down to earth Hylobares burst out:

> Well, *Cuphophron*, you may hug your self in your high *Metaphysicall Acropolis* as much as you will, and deem those Arguments fetched from the frame of Nature mean and popular: but for my part, I look upon them as the most sound and solid Philosophicall Arguments that are for the proving the Existence of a God (More 1668, 53).[18]

Boyle, too, held that design arguments were both available and the most likely to persuade rational, open-minded hearers. Such arguments were intended to form a large part of a book on atheism, something he worked on throughout his adult life but never published, though parts of it were used in various other works of a theological nature. Boyle did, however, leave a plan of the intended work, and in the manuscript remains -- seven volumes of correspondence, forty-six volumes of miscellaneous papers and eighteen volumes of notebooks -- there are still a number of unpublished fragments and some longer selections which he intended for this work.

Boyle intended "the little Tract about Atheism" to have three sections:

> In the First of these, the Author represents some Reasons why it should not be thought strange if it be found somewhat difficult to demonstrate the Existence of a Deity.
>
> > 1 The First of these Reasons is, that by reason of the selfe existence and Primity of God, his Essence cannot be Causable.
> >
> > 2 The Vitious Affections & Habits and the depravd frame of mind to be met with in most Atheists do very much indispose them to be convinc'd by the proofes of a Deity that might other wise be sufficient.
> >
> > 3 Since God is a Being whose Nature is the most singular of all, there must necessarily belong to him divers things, not to be paralleld.
> >
> > 4 The Difficulty of such speculations as belong to the Contemplations of Gods Attributes keeps the generality of Atheists & Libertines from being qualifyed for such Enquiries.
>
> In the second section <the Author> haveing premisd, that the foregoing Reasons make it Equitable not to expect metaphysical or rigid Demonstrations of a Deity, but to be content with a moral one, if no better can be had, proceeds to the mediums whereof such a Demonstration can be made up. Such as are

The innate Idea of a Deity

The general Consent of mankind. (To one of which or both may be referrd the Epicurean Anticipation.)

The Reproaches or Boadings & Disquieting Terrors of a Guilty Conscience

The Fabrick & Conservation of the world, especially of Animals

The Nature & Propertys of the Soul of Man

The Lawes of its Union with the Body[19]

The Universal Providence that directs the Affairs of Mankind.

Supernatural Effects whether of good or Evil Spirits (as their Apparitions Action Oracles Predictions &c)

The Patefactions that God has made of himselfe by true miracles. (To which Prophecies are reduc'd.)

The Third section is spent in shewing some of the Reasons why the Arguments proposd in the Second are often unprevalent.

1 And among the Intellectual Impediments the First is, That Atheists often injuriously attribute to the notion of a Deity the fond Opinions or rash Assertions of unskillfull men.

2 Atheists on the other side do sometimes no less injuriously father their owne Errors & mistakes on the notion of a Deity.

3 They do not equitably consider the Nature of the Thing to be proved, & the necessity that thence arises, that the Theory of the Divine Attributes should be lyable to specious Objections.

4 They do not duely consider, that their owne Hypothesis is lyable to some of the same difficulties & Objections, and to others that they cannot solve.

5 The Objections are more popular & easy that are to be made against the notion of a Deity than the Answers to those Objections & the Arguments

> which prove that Notion.

Well, that was the plan, and Boyle certainly thought that the design arguments he intended for section two should convince the open-minded. Moreover, he thought, such arguments should particularly convince those who were knowledgeable about nature, who knew enough about the *details* of the world to be impressed by the intricacy of the presumed workmanship. "[T]here are," he wrote, "positive Reasons afforded by Philosophy to prove a Deity, namely … the Cartesian Idæa, the Originall of Motion, the use of Parts in Animalls, especially the Eye, the valves of the heart, the musculi perforantes & perforati,[20] & the temporary [parts][21] of a foetus <& the Mother>."

The argument from design, said Kant, "always deserves to be mentioned with respect. It is the oldest, the clearest, and the most accordant with the common reason of mankind" (Kant 1781, A623/B651). But, he pointed out, there was a problem with it, a problem which in fact had already been pointed not only by David Hume,[22] but by Boyle's younger contemporary, Charles Blount, who wrote,

> could we conclude any thing from *Miracles*, yet we could never thence conclude of the *Existence* of God. For since a Miracle is Work *limited*, and never implies any but a certain and limited Power: most certain and evident it is, that from such an Effect we cannot rightly infer the Existence of a Cause whose Power is *infinite*, but at most of a Cause whose Power is greater: I say, *at most*; because from many Causes concurring there may follow some Work, whose Force and Power is indeed less than the Power of all its Causes put together, but far greater than the Power of any one of them taken singularly (Blount 1683, 11).

To a large extent Boyle accepted these points. He notes explicitly that none of the proofs he was prepared to offer amounted to a *demonstration* of God's existence, and indeed he felt that a *demonstration* was not possible. A *demonstration* was typically held to proceed from necessarily true premises (often Aristotelian *principles* [23]) via a valid argument to a necessarily true conclusion, and part of what was at issue was whether we should be looking for a demonstration of God's existence, or something less which would nonetheless still be useful in the fight against atheism . "To haue the Science of a thing," said Pierre du Moulin the elder, "two certainties are required. The one is, that the thing be certaine of it selfe and vnchangeable. The second is, that the perswasion which wee haue of it be firme and cleare" (Du Moulin 1624, 162). Gassendi agreed, as did Arnauld and Nicole in the Port Royal Logic (Gassendi 1658, Canon XVI, 144; Arnauld 1662, part IV, ch 8, 323-4, pagination as in the 5th, 1683 edition).

The persuasive alternative to a demonstration was sometimes styled a *proof*, but often people spoke of *moral* demonstrations as opposed to strict, or mathematical, demonstrations. Boyle wrote:

> besides the Demonstrations wont to be treated of in vulgar Logick, there are among Philosophers three distinct, whether *kinds* or *degrees*, of Demonstration. For there is a *Metaphysical* Demonstration, as we may call that, where the Conclusion is manifestly built on those general Metaphysical Axioms, that can never be other than true; such as *Nihil*

> *potest simul esse & non esse* … *&c.* (Nothing can both be and not be at the same time) There are also *Physical* Demonstrations, where the Conclusion is evidently deduc'd from Physical Principles; such as … *Ex nihilo nihil fit* … *&c.* (From nothing, nothing comes) which are not so absolutely certain as the former, because, if there be a God, He may (at least for ought we know) be able to create & annihilate Substances …. And lastly, there are *Moral* Demonstrations … where the Conclusion is built, either upon some one *such* proof cogent in its kind, or some concurrence of Probabilities, that it cannot but be allowed, supposing the truth of the most receiv'd Rules of Prudence and Principles of Practical Philosophy.
>
> And this *third* kind of Probation, though it come behind the two others in certainty, yet it is the surest guide, which the *Actions of Men*, though not their *Contemplations*, have regularly allow'd them to follow (BW, 8:281; Boyle 1772, 4:182-3).

This moral certainty, Locke remarked, "is not only as great as our frame can attain to, but as our Condition needs" (Locke 1690, 4.11.8).

When we look in detail at Boyle's discussions of the various moral demonstrations outlined in his proposed second section it becomes clear that he fancied some considerably more than others. He often mentions the importance of conscience, but concentrates, as far as proofs go, on various design arguments, on arguments based on the incorporeality of the human soul, and on arguments involving miracles. Unlike the clerical authors of the time he pays little or no attention to the arguments from "the innate Idea of the Deity," from "the general Consent of Mankind," or from "the Universal Providence that directs the Affairs of Mankind."

Boyle did not expect his (or anyone's) proofs to convince most atheists:

> you need not thinke it strange, that I never pretended to convert resolved Atheists. For, besides the difficulty of treating clearly and cogently of such abstruse subjects as are many that relate to Atheism; the Will and Affections have so great an influence upon some mens Understandings, that 'tis almost as difficult to make them *beleive*, as to make them *Love*, against their Will. And it must be a very dazzleing Light, that makes an impression upon those that obstinately shut their Eyes against it. 'Tis not by Gods ordinary workes, but by his Extraordinary Power, that such men must be reclaimd to an acknowledgement of his <Existence>. For they that would find the Truth, especially in matters of Religion, must be diligent Inquirers after it, and *may be* strict Examiners of it, but *must not be* resolved Enemies to it. For to such, if to any, God is a Sun, that is not to be discover'd but by <his> owne Light (BP 2:64).

Boyle was aware that most believers held their belief on insufficient grounds (see BP 4:60), but felt himself fortunate in that sound philosophy showed that the religion to which he was born was the correct one. For Boyle, miracles (in particular the miracle of Pentecost) were a crucial factor in opting for

Christianity. The Christian miracles, he felt, clearly bore the stamp of God upon them. There were, he agreed, other miracles or apparent miracles, but the miracles which purported to establish Christianity were neither pretenses nor diabolical, and they *were* miraculous. Locke believed that "Mahomet having none to produce, pretends to no miracles for the vouching of his mission," but Boyle was aware of the argument that the *Koran* itself is miraculous (in view of the disparity between it and what might reasonably be expected of its author in the absence of divine inspiration). He felt, not that this argument was inappropriate, but that it failed the test empirically:

> [T]he *Saracens*, press'd with their Religions being destitute of attesting Miracles, … reply, That though there were no other Miracle to manifest the Excellency of their Religion …, yet the Alcoran it Self were sufficient, as being a Lasting Miracle that transcends all other Miracles. How Charming its Eloquence may be in its Original, I confesse my self too unskilfull in the Arabick Tongue, to be a competent Judge … but the Recent Translations I have seen of it in French, and … Latin, elaborated by great Scholars, and accurate Arabicians, by making it very Conformable to its Eastern Original, have not so rendred it, but that Persons that judge of Rhetorick by the Rules of it current in these Western Parts of the World, would instead of extolling it for the Superlative, not allow it the Positive Degree of Eloquence; [and] would think the Style as destitute of Graces, as the Theology of Truth … .[24]

Boyle does not deny that the style could have been miraculous, and indeed he runs a formally similar argument concerning the Apostles and the miracle of Pentecost: the Apostles' "Hearers … knew it was <not> naturally possible, that uninspir'd Persons, and especially illiterate Fishermen, should <grow> able, in a trice, to make <weighty> discourses to many differing Nations, in their respective Languages" (BP 7:99). Boyle accepts, and indeed uses, the *form* of the Koran argument: it is the premise he disputes.

Boyle is ambivalent about the function of miracles. Generally he regards them as being philosophically relevant after we have a proof of God's existence, something which natural theology will afford us. (Boyle fastens on two main types of design arguments: those involving the complexity of animate beings, particularly very small animate beings,[25] and those which highlight the need to explain the origin and continuing function of natural laws: God must not only sustain God's creatures, Boyle argues, he must also sustain the regularities which we recognize as lawlike.)

Having convinced ourselves of God's existence through the considerations which natural theology makes available to us, and realizing that God is likely to institute a religion to make his nature and requirements known to us, we look to miracles to see which instituted religion is the correct one. However, though in general Boyle argues that accepting something as a miracle *presupposes* God's existence, and so miracles are to be used to *institute* the correct religion rather than to ground its metaphysical basis, he does sometimes urge an argument from miracles which will yield not merely the correctness of Christianity, but the acceptability of religious belief as such. Briefly: we have good historical testimony for the occurrence of miracles, but miracles are possible only if God intervenes in nature (and thereby exists) (BP 5:106-7).

3. Boyle's World View

Boyle had a straightforward notion of creation. First of all God, at a particular, fairly recent, point in absolute time, made matter. Boyle was an admirer of "our *Irish* St. *Austin*"(Birch 1772, I:xxxiii), James Ussher, archbishop of Armagh, who famously propounded what seems to us, though not to his contemporaries, to be a very late date (4004 b.c.) for creation. Boyle saw "no just reason to embrace their opinion, that would so turn the two first chapters of *Genesis*, into an allegory, as to overthrow the literal and historical sense of them" and, noting the implausibility of the claims of "some extravagant ambitious People, such as those fabulous *Chaldeans*, whose fond account reach'd up to 40000 or 50000 years," held that "Theology teaches us, that the World is very far from being so old by 30 or 40 thousand years as they … have presum'd: and does, from the Scripture, give us such an account of the age of the World, that it has set us certain Limits, within which so long a Duration may be bounded, without mistaking in our Reckoning. Whereas Philosophy leaves us to the vastness of Indeterminate Duration, without any certain Limits at all" (BW, 8:21; Birch 1772, IV:11). Revelation gives us (1) truths of which we would otherwise be ignorant, such as "the order and time of the Creation of the World and of the first man and woman"; (2) details of truths which we can otherwise obtain "but very dimly, incogently, and defectively"

> such as … That the World had a beginning, that 'tis upheld and govern'd by Gods general concourse & providence; that God has a peculiar regard to mankind; and a propitious one to good men; that he foresees those future things, we call *contingent*: that mens souls shall not dye with their bodyes, and many other articles of the Philosophers, as well as the Christians Creed (BP 7:242).

Additionally there are (3) "*hints*" which lead us to truths we would otherwise miss, such as "that whatever men have generally believ'd, Vegetables had their Origine independent from the Sun, the earth having producd all kinds of plants a day before God made that Luminary" (BP 7:243). Moreover,

> it ought much to recommend many of the things that Revelation discovers to us, that they are congruous, and if I may so speak Symmetrical to what reason it self teaches us; and this Supernatural Light does not only confirm, but advance and compleat the truths discoverable by the light of nature. For God has so excellently orderd the discoverys he makes of Theological <Veritys> by meer reason, and by the holy Scriptures, that what Revelation superadds to Reason, does both very well agree with it, and supply what was wanting to it, that from them both might result as compleat a body of Theological Verities, as is either necessary or fit for us in our Mortal State (BP 7:245-6). (See further MacIntosh 1992.)

Unsurprisingly, then, Boyle makes a point of attempting to bring his creation story into line with a literal interpretation of the Genesis story. God, he believes, could have started things off earlier or later, but chose not to. Having created matter, he broke it up and started it moving. Sometimes Boyle says he broke it up *by* starting it moving.[26] *Then* he gave it laws, since the "casual justlings of atoms" would not,

Boyle thinks, have given rise to this world.[27] Hooke, explicating Genesis, argued for the same ordering. For Boyle and Hooke, that is, a world without laws is not only possible, our world was such a world for a time. In the previous century Konrad Daspodius held that comets "drift without laws," but by 1686 Leibniz was writing, "God does nothing out of order, and it is impossible even to feign events which are not regular."[28] Leibniz's point was that just as any 'random' sequence of points will determine a curve (actually an infinite set of curves, but Leibniz only needed the lesser claim), so any sequence of events will conform to a regular pattern. Generalizing, we might say that just as points underdetermine equations, so facts underdetermine theories.

It should be noted however that for Leibniz, as for a number of his contemporaries, not all such laws need be laws of *nature*.[29] Malebranche, for example, in his *Dialogues on Metaphysics*, offered no fewer than five distinct types of law: "general laws of the communication of motion, … laws of the union of soul and body, … laws of the union of the soul with God, with the intelligible substance of universal Reason, … general laws which give good and bad Angels power over bodies, … finally, the laws by which Jesus Christ received sovereign power in Heaven and on earth, over minds as well as bodies, not only to distribute temporal goods … but to diffuse (*répandre*) internal grace in our hearts" (Entretiens sur la Métaphysique, in Malebranche 1962, 12:319-320).[30]

Boyle's position is an intermediate one between the claim that some objects or events are lawless, and the claim that lawlessness is impossible. For Boyle, physical objects do exhibit nomological regularities, but this is a contingent fact about the world, or rather, for Boyle was cautious about generalizing, about the spatio-temporal portion of it we occupy. He agrees, however, that there are laws that are *not* laws of nature, with the laws of interconnection between body and soul providing, for him, an obvious example. This interconnection also provides a clear example of a state which God constantly preserves.

After having made matter, started it moving, and given it laws, God then formed the matter into particular structures and shapes, including certain "seeds." Then he added some "seminal" principles. Boyle does not clearly indicate whether or not these are a special subclass of natural laws affecting matter, or whether they are in some way what he sometimes calls "supra-mechanical," though he does point out that if there is a mechanism for animal inheritance, then it would seem to require a framing intelligent agent (BW, 12:445-6, Birch 1772, VI:728-9). His older contemporary Harvey, much admired by Boyle, was in no doubt about the matter, pouring scorn on those who talk

> As if (forsooth) Generation were nothing in the world but a meer separation, or Collection, or Order of things. I do not indeed deny that to the Production of one thing out of another, these forementioned things are requisite, but Generation herself is a thing quite distinct from them all … (Quoted Toulmin and Goodfield 1962, 146).

All this holds for the corporeal universe, as opposed to the three sorts of incorporeal creatures God created or, in our case, continues to create: angels, evil demons, and human souls: the good, the bad, and the imprisoned.[31] The angels were created "before the visible World … was half compleated," but God creates new human souls daily, and moreover works a "physical miracle" to attach them to their

respective bodies (BP 7:243, BP 2:62).[32] Sometimes Boyle felt that although humans are made in God's image they, like other created beings, are "at their best but umbratile, and Arbitrary Pictures of God their Creatour" (BP 4:4). Elsewhere, however, he offers a more traditional account of the soul as the image of God:

> The Christian virtuoso considers the rational soul, not barely as it guides the motions of that living engine, we call the body, but as it is a kind of imprisoned angel, that bears the image of God, and is capable of knowing, both ourselves and him; and by a consciousness of her being his production, is capable of acknowledgeing, loving, and obeying him, and referring to his glory all the excellencies she discovers, both in herself, and in the body she is united to; by which just reference, she is, by his goodness, in his divine Son, made capable of becoming incomparably more knowing, than here she is, and eternally happy with him (BW, 12:504, Birch 1772, VI:775).

Although humans are made in the image of God, they are considerably less clever than the angels,[33] and since it is quite possible that God's primary end in making the universe was to provide a universe for the angels, and not centrally for humans, it is thereby quite possible that the universe will be too complicated for us to understand:

> [I]f God be allowed to be, as indeed he is, the Author of the Universe, how will it appear that He, whose Knowledge infinitely transcends ours, and who may be suppos'd to operate according to the Dictates of his own immense Wisdom, should, in his Creating of things, have respect to the measure and ease of Humane Understandings; and not rather, if of any, of Angelical Intellects? So that whether it be to God or to Chance, that we ascribe the Production of things, that way may often be fittest or likelyest for Nature to work by, which is not easiest for us to understand (BW, 3:257, Birch 1772, II:46).

"[W]e presume too much of our own abilities," Boyle wrote, c. 1680, "if we imagine that the omniscient God can have no other Ends in the framing & managing of Things Corporeal, than such as we Men can discover" (Boyle MS 198, fol. 120). It follows at once that while simplicity may often be our best guide as to what working hypothesis to choose, we should not think it to be inevitably a reliable guide to truth.

Why did God create the universe in this piecemeal way? What is Boyle's rationale for thinking that God didn't just start off by creating matter in motion with the proper directed velocities and letting it give rise to the present world in its own good time? Or why not suppose that he created the *present* world as a going concern? Boyle doesn't tell us, but two points stand out.

First of all, it certainly fits the fact that Boyle has a very limited view of omnipotence.[34] Here he is, for example, bemused by the swiftness of God's creative ability:

> As great a Number & variety of parts as a living Humane Body consists of, 'tis highly probable that the Lump of Stupid[35] matter out of which they were fashion'd, was

> contriv'd into this admirable System; if not in a moment, yet in a very short time. For the sacred story relates, that man was not created till <about> the end of the six dayes work; and since in One day God created all the four footed Beasts, (wilde & tame,) and all the numerous Reptiles that creep upon the Earth after their kind; 'tis no way improbable that among so great a multitude of differing *species* of Animals, or Living Engines, that were made in one Part of the same sixth day, God should make a Humane Animal in an extreamly short time, not to say *in a trice* (BP 4:85). [Here and throughout in quotations, the emphasis is in the original.]

Doing things step by step, fairly quickly, and with moderate success, was quite enough to excite Boyle's admiration for the Almighty and, though he had certainly read Descartes's *Principles* (in which Descartes remarks explicitly that although God *could* have let things work their way from a very different initial state to the present world, *in fact* God started the world off *in medias res*), it is quite possible that the alternative did not strike him as likely: he was not, after all, a mathematician, or even a mathematical physicist.[36]

Given this, and the fact that motion was not natural to matter, but had to be added to it, it would seem plausible to Boyle that God created matter first, and then gave it a push, particularly since the push had to be precisely fine tuned in order to yield just the world we now have. (That matter is not naturally in motion forms the basis of one of Boyle's criticisms of Epicurus. Boyle takes Epicurus to hold that motion *is* an innate property of matter. How then, asks Boyle, are we to explain the fact that it is lost or changed as a result of collision between particles? (BP 2:5).) Additionally, Boyle notices that no system of laws can offer a complete explanation: we also need an account of the initial parameters.[37] But then, since they are logically distinct, why not have them chronologically distinct as well?

Secondly, there are, perhaps, historical reasons. For Boyle is conscious of himself as building on past views, and such views typically treated matter as giving rise to the present world, and, in the case of some past thinkers, at least, as having existed in a constant state for some time before the initiating changes that led to the present world occurred. The notion of a piecemeal creation, that is, fits Boyle's views of God's abilities, fits *Genesis*, and fits the views of previous thinkers. Probably we do not need to look farther for an explanation for his adopting such a view.

(b) *What kind of world is it that God created?*

God created a material world in time and space, but what kind of matter, what kind of space, what kind of time? As to the matter, Boyle agreed with contemporaries such as Huygens and Newton that "Matter [is] in its own Nature but one."[38] However Boyle, cautious as ever, explicitly allows God the possibility of creating matter which is *not* like ordinary matter, and instituting laws which are quite unlike the laws that obtain on earth. His views are worth quoting at length:

> [T]he World must every way have bounds, and consequently be finite; or it must not have bounds, and so be … *infinite*. And if the World be bounded, then those that believe a

> Deity,[39] to whose Nature it belongs to be of infinite Power, must not deny that God was, and still is,[40] able to make other Worlds than this of ours. …
>
> Now if we grant, with some modern Philosophers, that God has made other Worlds besides this of ours, it will be highly probable that he has there display'd His manifold Wisdom, in productions very differing from those wherein we here admire it. And even without supposing any more than one Universe, as all that portion of it that is visible to us, makes but a part of that vastly extended aggregate of bodies: So if we but suppose, that some of the Celestial Globes, whether visible to us, or plac'd beyond the reach of our sight, are peculiar Systemes, the consideration will not be very different. For since the fix'd stars are many of them incomparably more remote than the Planets, 'tis not absurd to suppose that as the Sun, who is the fix'd star nearest to us, has a whole Systeme of Planets that move about him, so some of the other fix'd Stars may be each of them the Centre, as it were, of another Systeme of Celestial Globes … . Now, in case there be other Mundane Systemes (if I may so speak) besides this visible one of ours, I think it may be probably suppos'd that God may have given peculiar and admirable instances of His inexhausted Wisedom in the Contrivance and Government of Systemes, that for aught we know may be fram'd and manag'd in a manner quite differing, from what is observ'd in that part of the Universe that is known to us. … [H]ere on Earth the Loadstone is a Mineral so differing in divers affections, not onely from all other Stones, but from all other bodies, that are not Magnetical, that this Heteroclite[41] Mineral scarce seems to be Originary of this World of ours, but to have come into it, by a remove from some other World or Systeme … .
>
> Now in these other Worlds, *besides* that we may suppose that the Original Fabrick … into which the Omniscient Architect at first contriv'd the parts of their matter, was very differing from the structure of our Systeme; *besides* this, I say, we may conceive that there may be a vast difference betwixt the subsequent *Phænomena*, and productions observable in one of *those* Systemes, from what regularly happens in *ours*, though we should suppose no more, than that two or three Laws of Local Motion may be differing in those unknown Worlds, from the Laws that obtain in ours (BW, 10:172-3, Birch 1772, V:138-139).

Boyle, that is, sees three distinct possibilities: the initial set up may differ, the matter involved may differ, and the laws in question may differ. Moreover the laws, as well as the matter, could have been formed differently by God, and could indeed vary from part to part of the current universe *within* the universe. Clearly the case of varying laws and the case of varying matter may run into each other, but Boyle treats them as distinct possibilities, and gives as an example the possibility of a combination of conservation and non-conservation possibilities: we can envisage bodies with the "power of exciting Motion in another Body, without the Movents loosing its own." Were this to be the case the resulting phenomena would be "strangely diversifyed." Moreover, God may have made a universe, or a part of this universe, which was such that "*some* parts of matter [would] be of themselves *quiescent* … and determin'd to continue at rest till some outward Agent force it into motion [while] *other* parts of the matter [may have] a Power … of restlessly moving themselves, without loosing that power by the motion they excite in quiescent bodies. … Nor is it so extravagant a thing, as at first it may seem, to entertain such suspicions as these. For in the

common Philosophy, besides that the Notion and Theory of Local Motion are but very imperfectly propos'd, there are Laws or Rules of it not *well*, not to *say at all*, establish'd."[42]

Boyle does not use the terminology of absolute space and time, but he remarks that God could have made the world earlier:

> Nor was it his Indigence, that forc't him to make the World, thereby to make new Acquisitions, but his Goodnesse, that prest him to manifest, and to impart his Glory; and the goods, which he so over-flowingly abounds with. Witness his Suspension of the World's Creation, which certainly had had an earlier Date, were the Deity capable of Want, and the Creatures of Supplying it (BW, 1:97, Birch 1772, I:270).

Boyle's general view about both space and time is that since they are

> Primary & Heteroclite … 'tis noe wonder that our Limited & Imperfect understandings should not be able to reach to a full & clear comprehension of them, but should be swallowd up with the <Scruples &> Difficulties that may be suggested by a <bold &> nice Inquiry into things, <to> which there seemes to belong a kind of Infinity (BP 2:53).

He also remarks about the world that

> *if* it be Finite [which Boyle allows as a possibility], then 'tis not in a place (such as the Schools define) after the manner of other Bodys, since there is no ambient Body whose inward surface determines it; and we may conceive it to move several wayes, as upwards or downwards, and yet not to change place, because (as was just now said,) it is in none, and all its Extream parts may keep the same situation in reference to one another (BP 1:64).[43]

Moreover, there is a

> rigid and Philosophical Notion … of rest, which for distinction sake may be called Absolute or Perfect Rest; which imports a continuance of a Body in the same place *precisely*, and includes an absolute Negation of all local Motion, though never so slow or imperceptible; … in this rigid sense of the word Rest, I durst not affirm, that there are any Bodies at Rest in the Universe (at least for any long time) but willingly [allow] it to be made a Problem, whether there be any or no; … perhaps I [incline] to the Negative part of the Question (BW, 6:194, Birch 1772, I:444).[44]

Again, Boyle writes: "Suppose a Ball were in motion, & all the world should be on a sudden annihilated about it; why may not the motion of that Ball be continu'd? there being nothing to stop it; & if it be continu'd, we have a motion where the mobile dos not quitt the neighborhood of som bodys, and approach nearer to others" (BP 1:3).

We have, then, Boyle's view that a body can continue "in the same place *precisely*," that the whole (finite) universe might move in space, and that God could have created the world earlier than he did. Such views do not at least amount to a *rejection* of absolute space and time. Boyle's contemporary, Leibniz, who did reject absolute space, explicitly drew the conclusion that a finite universe could *not* move as a whole in space, and could *not* have been created earlier in time.[45]

The universe God created, then, contains a number of finite *incorporeal* entities, for whom the writ of physics does not run, and a number of material entities, compounded as far as we are aware, of the same matter in every case, set in a space and time independent of them, and subject to a number of God given laws.

4. Laws of Nature

On a number of occasions Boyle assures us that "God [is] the Author of the universe & free Establisher of the Laws of motion, whose generall Concourse is necessary to the conservation & Efficacy of every particular Physicall Agent" (BP 2:132). The trouble is, he seems to have thought that this remark was fairly transparent, and does not trouble to explain it to us. Moreover, he tends to use much the same phraseology on each occasion he discusses the issue. It was a commonplace of the time that Boyle was no stylist -- it was obvious at times to Boyle himself -- but though even the obsequious Budgell remarked that Boyle was "too wordy and prolix," in this case at least he was not wordy enough (Budgell 1732, 124).

It is tempting to suppose that Boyle must have had some reasonably well thought out views on the question of *how* God sustains the world. He was after all one of the most impatient of thinkers when it came to fake or non-explanations, and he was in general very aware of the danger of letting verbal 'explanations' get in the way of real ones. He objected against the scholastics, for example that

> to explicate a *Phænomenon*, being to deduce it from something else in Nature more known to Us, than the thing to be explain'd by It, how can the imploying of Incomprehensible (or at least Uncomprehended) substantial Forms help Us to explain intelligibly This or That particular *Phænomenon*? For to say, that such an Effect proceeds not from this or that Quality of the Agent, but from its substantial Form, is to take an easie way to resolve all difficulties in general, without rightly resolving any one in particular; and would make a rare Philosophy, if it were not far more easie than satisfactory … (BW, 5:351-2, Birch 1772, III:46-7).[46]

On the other hand, in theology he was more likely than elsewhere to let things get by, since he was convinced in advance that his theological picture was the right one, and he was used to stifling doubts about theological claims. Writing about himself in the third person as a young man he speaks of the "fleeting Clouds" of doubts which never ceased "now & then to darken /obscure/ the clearest serenity of

his quiet: which made him often say that Injections of this Nature were such a Disease to his Faith as the Tooth-ach is to the Body; for tho it be not mortall, 'tis very troublesome" (BP 37:182r). (These doubts persisted: see Hunter 1990, 410.)

Moreover, he had a well worked out doctrine concerning the limitations of reason, and often points out that we should not expect fully to understand God's workings, for God is, after all, "<a Being> of a most Primary and most singular Nature" (BP 2:107). (For Boyle on the limits of reason see Wojcik 1997.) Furthermore, he was willing to admit the impossibility of our understanding -- at least given the present limitations on our intellects -- even quite ordinary and lawlike matters, e.g., the way in which the human soul and human body interact. How God could have created the world, and how it is that he can intervene in it, are matters as mysterious to us as how mind and body can interact, and that is a total mystery.

Sometimes, Boyle remarks, our ignorance of things has to do simply with our lack of knowledge of the inner or hidden workings of a thing. He offers his, and indeed the century's, standard example of clocks which may have various internal mechanisms to produce the same outer effects. Thus he remarks that "we know in general, that digestion is made by some *Menstruum* or subtile substance in the Stomach; thô we know not the particular nature of that substance, (as whether it be Acid, Urinous, &c.)". Sometimes, though, our ignorance is of a deeper, richer variety:

> *sometimes* … we are not able to conceive the *Modus* of a thing, soe much as *in general*, or, as to the *possibility* of it, (abstracting from the positive Proofs that such a thing is) as, when we cannot conceive, how the Rational Soul can stop or determine the motions of the <humane> Body. And in this latter case, our not knowing the *Modus* of a thing, is usually more than a bare Ignorance, and inclines us to frame Objections against the Truth or Existence of the thing: because oftentimes the Incomprehensiblenes of the *Modus*, is grounded upon some thing that we conceive to be in the case, repugnant to the Laws or Course of Nature, or to some Dictate of right Reason: as, in the instance newly mention'd, it seems repugnant to the nature of things, that an Immaterial substance, not being Impenetrable, can resist or reflect the motion of a Body &c (BP 1:129).

Boyle was impatient with the Cartesian suggestion that we might be able to alter the direction though not the quantity of motion, not for Leibniz's reason that the notion of quantity of motion required a confusion between momentum and kinetic energy,[47] but for the straightforward reason that interfering with the directed velocity required as mysterious an interaction as altering the 'quantity of motion' would. He was aware of the Cartesian claim "that the rational Soul doth [not] give any motion to the parts of the Body, but only *guide* or *regulate* that which she finds in them already" (BW, 9:379, Birch 1772, IV:416), but that, he felt, did not really solve the problem, for that interaction was as mysterious as an energy introducing one: "I do as little conceive how the motions of the *Conarion* can work upon an Immaterial soul, as how any other part of the Body can do it. Nor do I conceive how an Immaterial Soul can work upon the *Conarion* its self, more then it can upon any other part of the Body" (BP 1:128), and he notes that it will not "suffice to object, that the human will does, in these cases, not produce any new motion, but only determine[s] the motions of the spirits, and by their means of the locomotive organs. For to put a check, at pleasure, to the motion of a body, that does already actually move in one line, and determine its

motion to continue in another, that is perhaps differing from it, or even opposite to it; to do this, I say, without opposing to the moving body, some other body, which, by its resistance and situation may change its former course, is not a mechanical operation" (BW, 12:480, Birch 1772, VI:756). A change of direction, just as much as a change in the 'quantity of motion,' is in fact as mysterious and inexplicable, if done by incorporeal means, as the introduction of energy into the system would be.

5. Boyle's Law

Many of us learned at school that Boyle's Law holds for ideal gases and can be summarized as PV = k, where k is a constant, and P and V are pressure and volume respectively. This law does stem from Boyle's work, but it is not what Boyle took himself to have demonstrated.[48]

Boyle was arguing specifically against a Jesuit scientist, Franciscus Linus, who claimed, not that ordinary atmospheric air does not have any pressure (a spring), but that its pressure was not sufficiently powerful for it to do all the things it does in fact do. So Boyle decided on an experiment to show the way in which, as we would say, the pressure and the volume of the air vary, when the air is, in Boyle's words, either 'compressed or dilated.'

He and his assistant, at the time Robert Hooke, made a J shaped tube and began to make a few measurements, but "were hindered from prosecuting the trial at that time by the casual breaking of the tube."

Subsequently they made another, larger, better piece of apparatus, and taking particular care that the measurements should be accurate, tested the hypothesis "that supposes the pressures and expansions to be in reciprocal proportion." The results are set out, with misprints, in two tables, and Boyle's conclusion was that the experimental findings matched the predicted results very well in the case of compression, less well in the case of rarefaction. Boyle suggested that the divergence from the expected result in the case of rarefaction may have been due to "some little aerial bubbles in the quicksilver" ("so easy is it in such nice experiments to miss of exactness," he added).

Now, what did Boyle take himself to have shown? First, that there is, as a matter of experimental fact, a spring to the air: this is not in the sense in which Boyle understands the term, any longer an hypothesis: it is now obvious from the experimental results: what explains, or purports to explain this fact will be a theory or an hypothesis, but the result itself is in no sense an hypothesis. As Boyle said

> …to determine whether the motion of restitution in bodies proceed from this, that the parts of a body of a peculiar structure are put into motion by the bending of the spring, or from the endeavour of some subtle ambient body, whose passage may be opposed or obstructed, or else its pressure unequally resisted by reason of the new shape or magnitude, which the bending of a spring may give the pores of it seems to me a matter of more difficulty, than at first sight one would easily imagine it. Wherefore I shall decline meddling with a

> subject, which is much more hard to be explicated than necessary to be so by him, whose business it is not … to assign the adequate cause of the spring of the air, but only to manifest, that the air hath a spring, and to relate some of its effects (BW, 1:166, Birch 1772, I:12).

Secondly, Boyle takes himself to have shown that, for atmospheric air, within the limits of his experimental set-up, "the pressures and expansions [are] in reciprocal proportion," or, as we would say, pressure and volume vary inversely. He doesn’t take himself to have shown anything more than this. He does remark that further experiments may show that the relationship holds outside the boundary conditions imposed by the experimental set-up, but the experiments he has just made certainly don’t show that. What Boyle expressly said was,

> till further trial hath more clearly informed me, I shall not venture to determine, whether or no the intimated theory will hold universally and precisely, either in condensation of air, or rarefaction: all that I shall now urge being, that…the trial already made sufficiently proves the main thing, for which I here allege it; since by it, it is evident, that as common air, when reduced to half its wonted extent, obtained near about twice as forcible a spring as it had before, so this thus compress air being further thrust into half this narrow room, obtained thereby a spring about as strong again as that it last had, and consequently four times as strong as that of the common air (BW, 3:60, Birch 1772, I:159).

Thus Boyle’s Law, for Boyle, was not a universal generalization about ideal gases: it was a strictly limited claim about common or atmospheric air. Boyle did add that "there is no cause to doubt, that if we had been here furnished with a greater quantity of quicksilver and a very strong tube, we might, by a further compression of the included air, have made it counter balance the pressure of a far taller and heavier cylinder of mercury."

But he did not claim that the same ratio between pressure and volume would hold in such more extreme cases. Nor did he claim that there are no limits to the possible compression. It is worth stressing that Boyle had this limited view of his result, for Shapin and Schaffer 1985 suggest that

> The work Boyle undertook in reply to Linus was … done … with a specially constructed J-shaped tube in which pressures higher than atmospheric could be attained. Using this apparatus Boyle showed that if he compressed air twice as strongly as usual he could produce twice as strong a spring. He concluded that the process could go on indefinitely, so that there were no limits to the power of the air’s spring (Shapin and Schaffer 1985, 168-9).

But Boyle was quite happy not to draw such conclusions, simply because his experiments didn’t allow that kind of jump. There are other important ways in which he thought that generalizations about nature might fail of universality. He had a very healthy notion of the complexity of the world, and an acute sense of the difficulties to which even apparently simple experiments could give rise, and perhaps in consequence had more sympathy than stricter theologians for the problems which universe construction

might involve, even for the Almighty.

6. Perception and the Soul

Two distinct notions of the soul occupied centre stage in the seventeenth century. One, stemming from Plato and the Pythagoreans, with theological trimmings by Augustine, had been given immense prestige by Descartes' championing of it. This view was what Geach has called the "savage superstition ... that a man consists of two pieces, body and soul, which come apart at death." Geach adds, "the superstition is not mended but rather aggravated by conceptual confusion, if the soul-piece is supposed to be immaterial" (Geach 1969, 38).

The second main account, stemming from Aristotle, had been taken over and made Christian by St Thomas Aquinas.[49] In this account the soul was, though incorporeal, not simply a separate bit attached to the body, but was the form of the individual animal in question, whether human or not. Aquinas presented arguments to show that human souls were subsistent in view of various capacities they had, and proceeded from there to argue for the possibility of the continuing existence of human souls in the absence of the body. He was, however, clear that the human person (even when the person in question was Christ in human form) was not merely a soul with an attached body, but was the body informed by the soul: if your soul alone were to survive death you would not. Bodily resurrection is essential to the survival and immortality of humans.

Cartesians and Thomists alike believed on scriptural grounds that there were actual cases of separated souls, namely, the angels, fallen and unfallen, so the possibility that the human soul might itself be subsistent was simply the possibility that it might sufficiently resemble an already accepted ontological group: despite the problems that substance dualism raises, a number of which presented themselves clearly to Boyle, there was no *general* problem concerning incorporeal entities, and there were, Boyle felt, strong arguments for the incorporeality of the human soul.

For Boyle, as for other leading seventeenth century figures, perception was a matter of information entering the brain as a result of causal interaction between the perceiver and the perceived object. Arriving at the brain the information was processed by a subsystem or set of subsystems devoted to presenting it to the cognitive system, and to storing it thereafter. The initial processing was done by a system, the *common* sense, that combined the inputs from the various sense organs (left eye + right eye; eyes + ears; etc.) and it was then *imagined* -- that is, an image was formed in the brain though, as Kepler and Descartes noted explicitly, the image was not a literal, optical, image. That apart, seventeenth century thinkers accepted in general outline the position which had already been set out in the thirteenth century by Roger Bacon, who was in turn simply collating the views of earlier Islamic writers on the subject, though of course the details, particularly the details of the causal interaction between percipient and perceived, varied from writer to writer (See further Lindberg 1976, MacIntosh 1983, and Sutton 1998).

Imagination was a matter of material images being formed in the brain. But, it was held by Boyle and others, we have knowledge of things which are literally unimaginable -- that is, they cannot be accurately

represented by a corporeal image in the central nervous system. There were a variety of reasons for this belief. First it was held that there were things which were too large, and things which were too small to be imagined, that is, imaged. Hence some non-material faculty was needed to account for this ability. Additionally, there are things which are not image-able because they cannot be represented accurately by *any* physical system. Boyle's stock example in this area, though not his only example, is the incommensurability of the sides and diagonal of a square. Since $\sqrt{2}$ is irrational no discrete (corpuscular) system can accurately represent both. But we do have knowledge of squares. Therefore we must be employing a non-material system. Also, there were things such as Descartes' chiliagon which, while they could be represented physically, could not (as experience shows) be represented accurately by *our* physical imaging system.

Additionally (a familiar Aristotelian point), our ability to abstract -- to consider universals and not merely particular instances -- was held to provide further evidence for the incorporeality of the soul and hence for the possibility at least of human immortality.

Boyle also noted, as did his contemporary Henry More, the Cambridge Platonist, the occurrence of ecstasies. Boyle, like More, took the existence of ecstasies seriously and, accepting the literal meaning of the term, thought that such experiences showed the actuality of non-corporeal, out-of-body, experiences. Boyle indeed offers the case as a refutation of the Aristotelian view that images are required for human thinking. Locke was more cautious on the issue: "whether that, which we call *Extasy*, be not dreaming with the Eyes open, I leave to be examined."[50]

Thus, for Boyle, souls were almost certainly Cartesian souls, though as mentioned earlier he hesitates about whether or not the human soul may not be a substantial form (BW, 5:300, Birch 1772, III:12).

Given that souls are incorporeal adjuncts of the body, it follows that they are not *materially* destructible, and that the laws of interaction between soul and body are not laws of natural philosophy. Why grass *looks* green is a feature of the world which God decided upon and upholds. His reasons for this decision, says Boyle, were no doubt weighty, but they are, as to us, arbitrary (see, e.g., BP 2:62; 2:105; 9:40; 36.46v). Now, if our souls are non-material, that demolishes at least one philosophical barrier to a belief in an after-life such as is promised by Christianity.

Bibliography

Primary Sources

BP Boyle Papers: The Boyle Papers, Boyle Letters, and Boyle Notebooks in the Royal Society Library, London; now available on microfilm, Michael Hunter (ed.), as *Letters and Papers of Robert Boyle*, Bethsada MD: University Publications of America, 1990

BW Boyle Works, *The Works of Robert Boyle*, Hunter, M., and Davis, E. B. (eds.), 14 vols., London: Pickering and Chatto, 1999-2000

Secondary Sources

- Adam, C., and Tannery, P., eds., 1964-1976, *Oeuvres de Descartes*, 11 vols. Paris: J. Vrin, 1964-76
- Alexander, P., 1985, *Ideas, Qualities and Corpuscles: Locke and Boyle on the External World*, Cambridge: Cambridge University Press, 1985
- Anstey, P., 1999, "Boyle on occasionalism: an unexamined source," *Journal of the History of Ideas*, 60, 1999, 57-81.
- -----, 2000, *The Philosophy of Robert Boyle*, London: Routledge, 2000
- Aristotle, *The Complete Works of Aristotle*, Jonathan Barnes (ed.), Princeton: Princeton University Press, 1984
- Arnauld A., and Nicole, P., 1662, *Logic or the Art of Thinking*, J. V. Buroker (trans., ed.) Cambridge: Cambridge University Press, 1996
- Banester, J., 1578, *The Historie of Man, Sucked From the Sappe of the Most Approued Anathomistes*, London, 1578
- Ben-Chaim, M., 2000a, "Locke's ideology of 'common sense'," *Studies in History and Philosophy of Science*, 31, 2000, 473-501
- -----, 2000b, "The value of facts in Boyle's experimental philosophy," *History of Science*, 38, 2000, 57-77
- Birch, T., 1772, *The Works of the Honourable Robert Boyle*, Thomas Birch, (ed.), 6 vols. (London, 1772; reprinted Hildesheim: George Olms, 1966), a reprinting of the five volume 1744 edition.
- Blount, C., 1683, *Miracles, No Violations of the Laws of Nature*, London, 1683
- Brown, S., 1990, "Leibniz and the Fashion for Systems and Hypotheses," Gilmour, P. (ed.), *Philosophers of the Enlightenment*, Totowa, NJ: Barnes & Noble, 1990, 8-30
- Budgell, E., 1732, *Memoirs of the Life and Character of the Late Earl of Orrery And of the Family of the Boyles*, London, 1732
- Burton, J. D., 1994, "Crimson missionaries: the Robert Boyle legacy and Harvard College," *The New England Quarterly*, 67, 1994, 132-40
- Canny, N., 1982, *The upstart earl: A study of the social and mental world of Richard Boyle, first Earl of Cork, 1566-1643*, Cambridge: Cambridge University Press, 1982
- Cantor, Geoffrey, 1999, "Boyling over: a commentary on the preceding papers," *The British Journal for the History of Science*, 32, 1999, 315-24
- Chalmers, Alan, 1993, "The lack of excellency of Boyle's mechanical philosophy," *Studies in History and Philosophy of Science*, 24, 1993, 541-64
- Charnock, S., 1699, *The works of the late learned divine Stephen Charnock, B.D: Being Several Discourses upon the Existence and Attributes of God, His Discourse of Divine Providence, and a Supplement of Several Discourses on Various Divine Subjects*, 2 vols., London, 1699, the 3rd edition corrected

- Clarke, S., 1738, *The Works of Samuel Clarke, D.D.*, 4 vols., London, 1738, reprinted New York: Garland, 1978
- Clay, J., 1999,"Robert Boyle: a Jungian perspective," *The British Journal for the History of Science*, 32, 1999, 285-98, 315-24
- Collins, G. P., 2001, "Plus Ça Change: Has a Fundamental Constant Varied over the Aeons?" *Scientific American*, November 2001, 16-18
- Columbus, M. Realdus, 1559, *De Re Anatomica*, Venice, 1559
- Cook, M. G., 2001, "Divine Artifice and Natural Mechanism: Robert Boyle's Mechanical Philosophy of Nature," in Brooke, J. H., Osler, M. J., and van der Meer, J. M. (eds.), *Science in Theistic Contexts: Cognitive Dimensions*, *Osiris*, 16, 2001
- Cranston, M., 1957, *John Locke: A Biography*, London: Longmans, Green and Co, 1957
- Cudworth, Ralph, 1678, *The True Intellectual System of the Universe*, London, 1678, reprinted Stuttgart-Bad Cannstatt: Friedrich Frommann Verlag, 1964
- Cunningham, Andrew, 1997, *The Anatomical Renaissance: The Resurrection of the Anatomical Projects of the Ancients*, Aldershot: Scholar Press, 1997
- d'Alembert, J., 1751, "Preliminary Discourse to the *Encylopaedia* of Diderot," Schwab, R. (ed.), Indianapolis: Bobbs-Merrill, 1963
- Davis, E. B., 1994, "The anonymous works of Robert Boyle and the Reasons why a Protestant should not turn Papist 1687," *Journal of the History of Ideas*, 55, 1994, 611-29
- Debus, Allen G., 1965, *The English Paracelsians*, London, 1965
- Descartes, *Oeuvres de Descartes*, edited by C. Adam and P. Tannery, 11 vols. Paris: J. Vrin, 1964-76
- Dijksterhuis, E. J., 1961, *The Mechanization of the World Picture*, Oxford: Oxford University Press, 1961
- Dove, J., 1605: John Dove, *A Confutation of Atheisme*, London, 1605
- Dryden, J., 1958, *The Poems of John Dryden*, Kinsley, J., (ed.), 4 vols., Oxford: Clarendon Press, 1958
- Du Moulin, P., 1624, *The Elements of Logick*, De-Lawne, N., (trans.), London, 1624
- Dutton, R., 1951, *The Age of Wren*, London: Batsford, 1951
- Edwards, J., 1696, *A demonstration of the existence and providence of God, from the contemplation of the visible structure of the greater and the lesser world in two parts, the first strewing the excellent contrivance of the heavens, earth, sea, &c., the second the wonderful formation of the body of man*, London, 1696
- Edwards, Sandra, 1979, "Saint Thomas Aquinas on 'the same man'," *Southwest Journal of Philosophy*, 10, 1979, 89-97
- -----, 1985, "Aquinas on individuals and their essences," *Philosophical Topics*, 13, 1985, 155-163
- Figlio, Karl, 1999, "Psychoanalysis and the scientific mind: Robert Boyle," *The British Journal for the History of Science*, 32, 1999, 299-324
- Fleming, Stuart, 1987, "The search for nothing: Boyle's experience with a vacuum," *Archaeology*, 40, 1987, 72-3+.
- Foster, Michael, 1924, *Lectures on the History of Physiology During the Sixteenth, Seventeenth and Eighteenth Centuries*, Cambridge: Cambridge University Press, 1924
- Gassendi, Pierre, 1658: Pierre Gassendi, *Institutio Logica*, Jones, H. (trans.) Assen: van Gorcum,

1981

- Geach, P., 1969, *God and the Soul*, London: Routledge & Kegan Paul, 1969
- Gerhardt, C. I., (ed.), 1875, *Die Philosophische Schriften von G. W. Leibniz*, 7 vols. Berlin, 1875-90
- Glanville, J., 1689, *Sadducismus Triumphatus*, London, 1689 (3rd edition)
- Hall, A. R., and Hall, M. B., 1965, *The Correspondence of Henry Oldenburg*, Hall, A. R., and Hall, M. B., (eds., trans.), 11 vols., vols. 1-10 Madison: The University of Wisconsin Press, 1965-1973; vols. 10-11, London: Mansell, 1975, 1977
- Hooke, Robert, 1705, "Of Comets & Gravity," in Waller, R., (ed.), *Posthumous Works of Robert Hooke*, London, 1705; reprinted London: Frank Cass, 1971
- Hume, D., 1779, *Dialogues Concerning Natural Religion*, Kemp Smith, N. (ed.), Oxford: Clarendon Press, 1935
- Hunter, M., 1990, "Alchemy, magic and moralism in the thought of Robert Boyle," *British Journal for the History of Science*, 23, 1990, 387-410
- -----, 1993, "Casuistry in Action: Robert Boyle's Confessional Interviews with Gilbert Burnet and Edward Stillingfleet, 1691," *The Journal of Ecclesiastical History*, 44 1993, 80-98
- -----, 1994a, *Robert Boyle by Himself and His Friends, with a fragment of William Wotton's lost Life of Boyle*, Hunter, M. (intro., ed.), London: William Pickering, 1994
- -----, 1994b, *Robert Boyle Reconsidered*, Hunter, M. (ed.), Cambridge: Cambridge University Press, 1994
- -----, 1995a, *Science and the Shape of Orthodoxy: Intellectual Change in Late Seventeenth-Century Britain*, Woodbridge: The Boydell Press, 1995
- -----, 1995b, "How Boyle Became a Scientist," *History of Science*, 33, 1995, 59-103
- -----, 1998, *Archives of the Scientific Revolution*, Hunter, M., (ed.), Woodbridge: the Boydell Press, 1998
- -----, 1999a, "Introduction to a special issue assessing the appropriateness of practicing retrospective psychoanalysis on figures from earlier historical periods, using the 17th-century scientist Robert Boyle as a test case," *The British Journal for the History of Science*, 32, 1999, 257-60.
- -----, 1999b, "Robert Boyle 1627-91: a suitable case for treatment?" *The British Journal for the History of Science*, 32, 1999, 261-75, 315-24.
- -----, 2000, *Robert Boyle (1627-91): Scrupulosity and Science*, Woodbridge: The Boydell Press, 2000
- -----, 2001, "The discovery of second sight in late 17th-century Scotland," *History Today*, 51, 2001, 48-53
- -----, 2001, "The Work-Diaries of Robert Boyle: A Newly Discovered source and its Internet Publication," *Notes and Records of the Royal Society*, 55, 373-90
- Hunter, M., and Wootton, D., 1992, (eds.), *Atheism from the Reformation to the Enlightenment*, Oxford: Clarendon Press, 1992
- Huygens, Christiaan., 1888, *Oeuvres Complètes*, 30 vols., La Haye: Martinus Nijhoff, 1888-1950
- Ihde, A. J., 1964, "Alchemy in Reverse: Robert Boyle on the Degradation of Gold," *Chymia*, 9, 1964, 47-57.
- Johnston, Samuel, 1752, *Noetica*, in *Elementa Philosophica: Containing chiefly, Noetica, or*

Things relating to the Mind or Understanding: and Ethica, or Things Relating to the Moral Behaviour, Philadelphia, 1752; reprinted New York: Kraus Reprint Co., 1969

- Kahr, Brett, 1999, "Robert Boyle: a Freudian perspective on an eminent scientist," *The British Journal for the History of Science*, 32, 1999, 277-84, 315-24.
- Kant, I., 1781, *Critique of Pure Reason*, trans. N. Kemp Smith London: Macmillan, 1933, 1st ed. 1781, 2nd ed. 1787
- Kenny, A., 1993, *Aquinas on Mind*, London: Routledge, 1993
- Leibniz, G., 1685, *Discourse on Metaphysics*, Lucas, P. G., and Grint, L., (eds., trans.), Manchester: Manchester University Press, 1953 [first printed 1846, written 1685-6]
- Leibniz, G., 1686, *The Leibniz-Arnauld Correspondence*, Mason, H. T., (ed.), Manchester: Manchester University Press, 1967, correspondence spans 1686-7
- Leibniz, G., 1704, *New Essays on Human Understanding*, Remnant, P., and Bennett, J., (eds., trans.), Cambridge: Cambridge University Press, 1981, written c1704, published posthumously 1765
- Lennon, T. M., Nicholas, J. M., Davis, J. W., 1982, (eds.), *Problems of Cartesianism*, Kingston and Montreal: McGill-Queen's University Press, 1982
- Lindberg, D. C., *Theories of Vision from al-Kindi to Kepler*, Chicago: University of Chicago Press, 1976
- Locke, John, 1690, *An Essay Concerning Human Understanding*, Nidditch, P. H. (ed.), Oxford: Clarendon Press, 1975, 1st ed., London, 1690
- Locke, John, 1823, *The Works of John Locke*, 10 vols., London, 1823; reprinted Aalen: Scientia Verlag, 1963
- MacIntosh, J. J., 1976, "Primary and Secondary Qualities," *Studia Leibnitiana*, 1976, 8, 88-104
- -----, 1983, "Perception and Imagination in Descartes, Boyle and Hooke," *Canadian Journal of Philosophy*, 13, 1983, 327-352
- -----, 1991, "Robert Boyle on Epicurean Atomism and Atheism," in Osler 1991, 197-219
- -----, 1992, "Robert Boyle's Epistemology: The Interaction between Scientific and Religious Knowledge," *International Studies in the Philosophy of Science*, 6, 1992, 91-121
- -----, 1994, "Locke and Boyle on Miracles and God's Existence," in Hunter 1994b, 193-214
- -----, 1996, "Animals, Morality, and Robert Boyle," *Dialogue*, 35, 1996, 435-72
- -----, 1997, "The argument from the need for similar or 'higher; qualities: Cudworth, Locke, and Clarke on God's existence," *Enlightenment and Dissent*, 16, 1997, 29-59.
- -----, 2001, "Boyle, Bentley and Clarke on God, Necessity, Frigorifick Atoms and the Void," *International Studies in the Philosophy of Science*, 15, 2001, 33-47
- Maddison, R. E. W., 1969, *The Life of the Honourable Robert Boyle F. R. S.*, London, Taylor & Francis, 1969
- Malebranche, N., 1688, *Entretiens sur la Métaphysique*, in *Oeuvres de Malebranche*, Robinet, A., (ed.), 21 vols., Paris: J. Vrin, 1962-70
- Mallet, C. E., 1924, *A History of the University of Oxford*, 3 vols., Oxford: Clarendon Press, 1924
- McGuire, J. E., 1972, "Boyle's Conception of Nature," *Journal of the History of Ideas*, 33, 1972, 523-542
- Moore, Leslie, 1985, "'Instructive trees': Swift's Broom-stick, Boyle's Reflections, and satiric figuration," *Eighteenth Century Studies*, 19, 1985-6, 313-32.

- More, Henry, 1662, *An Antidote Against Atheism*, in *A Collection of Several Philosophical Writings*, 2 vols., 2nd ed., London 1662, reprinted New York: Garland, 1978
- More, Henry, 1668, *Divine Dialogues, Containing sundry disquisitions & Instructions Concerning the Attributes and Providence of God*, London, 1668.
- Newton, I., 1962, *Unpublished Scientific Papers*, Hall, A. R., and Hall, M. B., (eds.) Cambridge: Cambridge University Press, 1962
- Opie, I, and Opie, P., 1951, *The Oxford Dictionary of Nursery Rhymes* Oxford: Oxford University Press, 1951
- Osler, M. J., 1991 (ed.), *Atoms, Pneuma, and Tranquillity: Epicurean and Stoic Themes in European Thought*, New York: Cambridge University Press, 1991
- -----, 1992, "The intellectual sources of Robert Boyle's philosophy of nature: Gassendi's voluntarism and Boyle's physico-theological project," in Kroll, R, Ashcraft, R., and Zagorin, P., (eds.), *Philosophy, Science, and Religion in England 1640-1700*, Cambridge: Cambridge University Press, 1992, 178-98
- -----, 1994, *Divine Will and the Mechanical Philosophy*, New York: Cambridge University Press, 1994
- -----, 1996 "From Immanent Natures to Nature as Artifice: The Reinterpretation of Final Causes in Seventeenth-Century Natural Philosophy," *The Monist*, 79, 1996, 388-407v
- -----, 2001a, "Robert Boyle on Knowledge of Nature in the Afterlife," in Force, J. E., and Popkin, R. H., (eds.), *Millenarianism and Messianism in Early Modern European Culture: The Millenarian Turn*, Dordrecht: Kluwer, 2001, 43-54
- -----, 2001b, "Whose Ends? Teleology in Early Modern Natural Philosophy," in Brooke, John Hedley, Margaret J. Osler, and Jitse M. van der Meer, eds., *Science in Theistic Contexts: Cognitive Dimensions, Osiris*, 16, 2001
- -----, 1992, "The Scholar and the Craftsman Revisited: Robert Boyle as Aristocrat and Artisan," *Annals of Science*, 49, 1992, 255-276
- -----, 1993, "Biography, Culture, and Science: The Formative Years of Robert Boyle, *History of Science*, 31, 1993, 177-225
- Owens, Joseph, 1974, "Soul as Agent in Aquinas," *The New Scholasticism*, 48, 1974, 40-72
- Pagel, W., 1935, "Religious Motives in the Medical Biology of the XVIIth Century," *Bull. Inst. Hist. Med.*, 3, 1935
- Pelling, E., 1696, *A Discourse Concerning the Existence of God*, London, 1696
- Perreaud, F., 1653, *Démonographie, ou traité des démons*, Geneva, 1653
- Polanyi, M., 1958, *Personal Knowledge*, London: Routledge and Kegan Paul, 1958
- Principe, L. M., 1992, "Robert Boyle's Alchemical Secrecy: Codes, Ciphers and Concealments," *Ambix*, 39, 1992, 63-74
- -----, 1994a, "Boyle's alchemical pursuits," in Hunter 1994b, 91-105
- -----, 1994b, "Style and Thought of the Early Boyle: Discovery of the 1648 Manuscript of *Seraphic Love*," *Isis*, 85, 1994, 247-260
- -----, 1995, "Virtuous Romance and Romantic Virtuoso: The Shaping of Robert Boyle's Literary Style," *Journal of the History of Ideas*, 1995, 377-397
- -----, 1998, *The Aspiring Adept: Robert Boyle and his Alchemical Quest*, Princeton: Princeton University Press, 1998

- Roger, Jacques, 1982, "The Cartesian Model and Its Role in Eighteenth-Century 'Theory of the Earth'", in Lennon 1982, 95-112
- Ruby, J. E., 1986, "The Origins of Scientific "Law"," *Journal of the History of Ideas*, 47, 1986, 341-359
- Schubach, W., 1982, *The Paradox of Rembrandt's 'Anatomy of Dr. Tulp,'*, Medical History, Supplement No. 2; London: Wellcome Institute for the History of Medicine, 1982
- Shapin, S., and Schaffer, S., 1985, *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*, Princeton: Princeton University Press, 1985
- Shapin, S., 1993,"Personal development and intellectual biography: the case of Robert Boyle," *The British Journal for the History of Science*, 26, 1993, 335-346
- -----, 1994, *A Social History of Truth*, Chicago: University of Chicago Press, 1994
- Sleigh, R. C., 1990, *Leibniz and Arnauld*, New Haven: Yale University Press, 1990
- Sutton, J., 1998, *Philosophy and Memory Traces: Descartes to Connectionism*, Cambridge: Cambridge University Press, 1998
- Thomas, K., 1973, *Religion and the Decline of Magic*, Harmondsworth: Penguin, 1973
- Toulmin, S., and Goodfield, J., 1962, *The Architecture of Matter*, London: Hutchinson, 1962
- Wilson, C., 1995, *The Invisible World: Early Modern Philosophy and the Invention of the Microscope*, Princeton: Princeton University Press, 1995
- Wintroub, M, 1997, "The looking glass of facts: collecting, rhetoric and citing the self in the experimental natural philosophy of Robert Boyle," *History of Science*, 35, 1997, 189-217
- Wojcik, J., *Robert Boyle and the Limits of Reason*, New York: Cambridge University Press, 1997

Other Internet Resources

- The Robert Boyle Project (Michael Hunter, U. London/Birkbeck College)
  This is absolutely the first place to go. The page is not only valuable in itself, but is provided with a multitude of useful, relevant links, including:
  - Robert Boyle Work-diaries Project (Part of the Robert Boyle Project pages)
  - Bibliography of recent publications on Boyle (Part of *On the Boyle* newsletter)
- A facsimile reproduction of *The Sceptical Chymist* (U. Pennsylvania Library/Schoenberg Center for Electronic Text and Image)
- The text of Boyle's *Degradation of Gold*
- A brief, but accurate, account of Boyle (in French)
- Stanford Encyclopedia of Philosophy: Ralph Cudworth in The Cambridge Platonists

Related Entries

afterlife | Arnauld, Antoine | atheism and agnosticism | Bayle, Pierre | Descartes, René | Gassendi, Pierre | Hume, David | Kant, Immanuel | laws of nature | Leibniz, Gottfried Wilhelm | Malebranche, Nicolas | matter | mental imagery | miracles | omnipotence | ontological arguments | Pascal, Blaise | perception |

Plato | providence, divine | Spinoza, Baruch [Benedict] | theology: natural

J.J. MacIntosh
macintos@ucalgary.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 15, 2002
Content last modified: January 15, 2002

Stanford Encyclopedia of Philosophy
Notes to Robert Boyle

Notes

1. See further Brown 1990, who notes Hume's disparagement of the "passion for hypotheses and systems in natural philosophy" as well as d'Alembert's.

2. Boyle Letters 1:131r; reprinted in Birch 1772, I:cxxx. Birch notes, concerning Boyle's interest in alchemy, that he was instrumental in having "a statute made in the fifth year of King Henry IV against the multiplying of gold and silver" repealed. Every "article, or sentence, contained in the said act, and every word, matter, and thing, contained in the said branch or sentence, shall be repealed, annulled, revoked, and for ever made void; any thing in the said act to the contrary in any wise whatsoever notwithstanding." (Birch 1772, I:cxxxii).

3. See *An Historical Account of a Degradation of Gold, made by an Anti-Elixir: A strange Chemical Narrative*, BW 9:5-17; Birch 1772, IV:371-374. For discussion of this piece see Ihde 1964 and Principe 1998. The Hunter and Davis edition of Boyle's works is now the standard edition, deservedly, but Birch's 1772 edition is still widely available, so I am quoting from Hunter and Davis, but giving page references to both editions.

4. There is no definitive full-scale biography of Boyle. Among the most easily accessible sources are:

1. Hunter 1994a. This contains an excellent introduction by Hunter, as well as the best transcription of
2. Boyle's own third person autobiographical "An Account of Philaretus During His Minority" (BP 37:170r-184v, with some misordering), written between Boyle's coming of age in January, 1648, and July of the next year.
3. Birch's 1744 *Life* (Thomas Birch, *The Life of the Honourable Robert Boyle*, in Birch 1772, I: vi-ccxviii) which includes a shortened version of "An Account of Philaretus."
4. Maddison 1969, which includes an annotated transcription of "An Account of Philaretus …," 2-54.

Recent, important, accounts of shorter stretches of Boyle's life include Hunter 1993; Principe 1994 and 1995; and Shapin 1993.

5. Michael Hunter has pointed out the same suggestion coming from Meric Casaubon: ‘it is not improbable that divers secrets of [chemistry] came to the knowledg of man by the Revelation of Spirits’ (Hunter 1990, 398-9), and Debus notes Pagel generalizing the point: ‘by means of unprejudiced

experiment inspired by divine revelation, the adept may attain his end. Thus, knowledge is a divine favour, science and research divine service, the connecting link with divinity. Grace from above meets human aspiration for knowledge from below. Natural research is the search for God.' (Pagel 1935, 98f, quoted Debus 1965, 21).

6. Such doubts were not less common in the seventeenth century than in any other. See, e.g., Thomas 1973, 6.4, "Scepticism," 198-206.

7. Birch, *Life*, Birch 1772, I:xxvi; see BW 2:86, Birch 1772, I:355.

8. Boyle to his father, 25 May, 1642; Maddison 1969, 48-9.

9. Glanville, in *Sadducismus Triumphatus* (Glanville 1689), argued the same point at length. See further, Hunter 1979; Jacob 1974; and Prior 1932.

10. For further details, see Maddison 1969, 61.

11. Gilbert Burnet, *A Sermon at the Funeral of the Honourable Robert Boyle*, in Hunter 1994a, 55; Shapin 1994, 187.

12. Boyle to John Mallet, 2 Mar. 1641/2: B.M. Add. MS. 32093 (Mallet MSS.), fol. 293, quoted in Underdown 1971, 331.

13. Maddison 1969, 274. Hunter points out that the earliest draft of Boyle's Will mentions only "Atheists and Theists." (Hunter 1994a, lxxxiv, n 76. (Boyle's 'theists' are what we would now call Deists.)

14. The "freshly discovered ... receptaculum chyli (receptacle of the chyle, also known as the receptaculum commune and the receptaculum Pecqueti) made by the confluence of the venae lactae" is now more commonly known as the thoracic duct. The beginning of this duct is the cisterna chyli. The first publication of the discovery was made by Jean Pecquet (1624-74) in 1651 in his Experimenta nova anatomica, though he remarked that he made the discovery some years earlier. In the following year Johannes van Horn (1621-70) published the same discovery, having apparently made it independently (Foster 1924, 49). The venae lactae [the "milky" veins] are the vessels that carry the chyle (lymph given a milky look from the absorption of emulsified fats) from the small intestine to the thoracic duct. (Thanks to Andrew Cunningham for this and other information concerning the history of medicine.)

15. Birch 1772, I:liv. Fell was the Dean of Christ Church who was unorthodox enough to be "determined that even the young bloods of the House should work." (Mallet 1924, 2:427, quoted Cranston 1957, 70). Philosophers remember him as the unfortunate intermediary in 1684 when the King determined that Locke, who had behaved "undutifully to the government" should be deprived of his Studentship, but he lives in popular memory as a result of Tom Brown's impromptu rendering of Martial's "Non amo te,

Sabidi" (*Epigrammata* 1.32) as "I do not like thee Doctor Fell" (etc.). For details see Opie and Opie 1951, 169.

16. See further Berman 1988. The various clerical tracts and published sermons are often philosophically unsophisticated but are nonetheless interesting. See, among others, Charnock 1699, Dove 1605, Edwards 1696, and Pelling 1696.

17. This argument, which can be found in Zeno of Citium (*Nihil quod animi quodque rationis est expers, id generare ex se potest animantem conpotemque rationis*–"Nothing lacking consciousness and reason can produce out of itself beings with consciousness and reason," quoted Cicero, *De nat. deor*. ii 22.), is barely sketched by Descartes, but it was developed at length later in the century by Cudworth, Locke, Bentley and Clarke. See further MacIntosh 1997.

18. Cuphophron is "A zealous, but Airie-minded, *Platonist* and Cartesian, or Mechanist"; Philopolis is "The pious and loyall Politician"; Hylobares is "A young, witty, and well-moralized *Materialist*," who has earlier been converted by the design argument: "The weight of Reason and the vehemence of *Philotheus* [A zealous and sincere Lover of God and *Christ*, and of the whole Creation] his Zeal does for the present bear me down into this belief whether I will or no." (More 1668, 28)

19. For Boyle, as for most of his contemporaries, there are laws that are *not* laws of nature, with the laws of interconnection between body and soul providing, for him, an obvious example. This interconnection also provides a clear example of a state which God constantly preserves: "the very conditions of the *Union* of the Soul and Body; which being setled at first by God's *arbitrary institution*, and having nothing in all Nature parallel to them, the manner and Terms of that strange Union is a Riddle to Philosophers, but must needs be clearly known to *Him*, that alone did Institute it, and, (all the while it lasts) does preserve it." (BW, 10:188-9; Boyle 1772, 5:150. See also BW, 12:380, Boyle 1772, 6:681; BP 9:40; BP 36:46v).

20. These are the muscles (*m. flexor digitorum superficialis* and *m flexor digitorum profundus*) which, with their associated tendons move the fingers of the hand. The tendons of *m. flexor digitorum superficialis* are perforated and are penetrated by the lower tendons. Schupbach 1982 (60-64) provides evidence from a number of sixteenth and seventeenth century texts to show the way in which these tendons were considered "a thing notable and marueilous to behold [which] prudent nature [hath] wrought (Banester 1578, following Columbus 1559)." Paley's early nineteenth century *Natural Theology; or evidences of the existence and attributes of the Deity* shows an equal enthusiasm: "There is nothing, I believe, in a silk or cotton mill, in the belts, or straps, or ropes, by which motion is communicated from one part of the machine to another, that is more artificial, or more evidently so, than this *perforation* (quoted Schupbach 1982, 64)." Both Schupbach 1982, 15, and Cunningham 1997, 109, have an impressive 1685 drawing by G. de Lairesse showing the intersection of the tendons. (Thanks to Andrew Cunningham for drawing these works to my attention.)

21. Conjectured–the edge of the page is torn away. Cf. BP 4:67 where Boyle in a column headed

"Against Atheism" lists "The Eyes of Hawkes and Fishes and the Temporary parts belonging to the Foetus."

22. Now, Cleanthes, said Philo, with an air of alacrity and triumph, mark the consequences. *First*, By this method of reasoning, you renounce all claim to infinity in any of the attributes of the Deity. For as the cause ought only to be proportioned to the effect, and the effect, so far as it falls under our cognisance, is not infinite; what pretensions have we, upon your suppositions, to ascribe that attribute to the divine Being? ... Secondly, you have no reason ... for ascribing perfection to the Deity, even in his finite capacity, or for supposing him free from every error, mistake, or incoherence, in his undertakings (Hume, 1779, Part V).

23. Aristotelian principles were held to be necessarily true, though garnered from experience. After noting that "a deductive proposition ... will be demonstrative, if it is true and assumed on the basis of the first principles of its science (*Pr An*, 24a27-30)," Aristotle tells us that "it is the business of experience to give the principles which belong to each subject. I mean for example that astronomical experience supplies the principles of astronomical science ... . Similarly with any other art or science (*Pr An*, 46a17-21)." *How* these necessary truths are to be derived from experience is not completely clear in Aristotle.

24. John Locke, *A Discourse of Miracles*, (1706, written 1702), in Locke 1823, 9.258; BW 2:452-3; Boyle 1772, 2:298.

25. "God, in these little Creatures, oftentimes draws traces of Omniscience, too delicate to be liable to be ascrib'd to any other Cause. ...my wonder dwells not so much on Natures Clocks (is I may so speak) as on her Watches." (BW 3:223, Boyle 1772, 2:22, a slightly different version is at BP 8:139.) Berman points out (Berman 1988, 7, and 44 n 11) that the concentration on insects is not uncommon, but it is worth noting that many seventeenth century writers emphasized the meanness and contemptibility of insects whereas Boyle genuinely admires the workmanship involved. The virtuoso, he often emphasizes, is, by reason of his expertise, in a better position than the uninformed to see the strength of such design considerations.

26. See, e.g., BP 7:192. Boyle sometimes notices the logical possibility that God might have created matter "incoherent" (e.g., at BW 3:248, Birch 1772, II:38-9), but in general he adopts the position outlined.

27. Boyle makes this point in a number of places. See, e.g., BP, 1:16, 1:18, 2:14; BW 6:194, Birch 1772 I:445; BW 3:259, Birch 1772, II:48; BW 5:353-4, Birch 1772, III:48; BW 11:130, Birch 1772, V:428. The point was generally accepted. Dryden wrote, "No Atoms casually together hurl'd / Could e're produce so beautifull a world." (Dryden 1958, 1:13)

28. Konrad Daspodius, *Brevis doctrina de cometis, et cometarus effectibus* (Strasbourg, 1578), CIII(3), quoted in Ruby 1986, 356. Leibniz 1685, §6.

29. Earlier Descartes had remarked that the laws of nature simply were the laws of mechanics (*les regles des Mechaniques … sont les mesmes que celles de la nature, Discourse* V, AT 6:54). Leibniz thought it was a conceptual truth that the universe was lawlike–whatever happened was *the* law that governed things or, in particular, *that* thing. However, as noted earlier, he too agreed that, *in practice*, we should search for *intelligible*, that is, corpuscular, laws.

30. For further discussion see Sleigh 1990, 158 ff. In the seventeenth century these laws were generally felt to be, from a human point of view at least, the result of arbitrary decisions by God. The view was still common in the mid-eighteenth century when the American Samuel Johnson wrote: "We are, at present, *Spirits* or *Minds* connected with gross, *tangible Bodies*, in such a Manner, that as our Bodies, can perceive and act nothing but by our Minds, so, on the other Hand, our Minds perceive and act by Means of our bodily Organs. Such is the present Law of our Nature, which I conceive to be no other than a meer arbitrary Constitution or Establishment of Him that hath made us to be what we are" (Johnson 1752, §3, 3).

31. Human souls are "imprisoned" by contrast with "the angelical Community … of <Rational & Immortal beings> not clog'd with visible Bodys" (BP 1:66v). See further BW 8:33, Birch 1772, IV:19. In the apparently fragmentary "A Dialogue between the Soul and Body" Boyle's contemporary Andrew Marvell agreed: the soul is "hung up, as 'twere, in chains / Of nerves, and arteries, and veins," but Marvell nicely allows the body a similar plaint: "O who shall me deliver whole / From bonds of this tyrannic soul?"

32. In an earlier tradition time and the angels, along with the heavens and the earth, were co-created, so that there was never a time when the angels did not exist. See St Thomas Aquinas, *Summa Theologiae*, 1a 61.3 c.; 1a 66.4 c; *QD de Potentia Dei*, 3.18 ad 20.

33. This is stressed in *Of the High Veneration Man's Intellect owes to God, Peculiarly for His Wisdom and Power*, BW 10:176-7, Birch 1772, V:142.

34. On this issue see MacIntosh 1991, 1992, and 1994.

35. For Boyle, as for his contemporaries, 'stupid' in such context meant simply 'insensible,' or 'non-sensory.'

36. R. Descartes, *Principles of Philosophy*, 3.45, AT 7A:100. For a discussion of the way in which Descartes's counterfactual views came be taken chronologically see Roger 1982.

37. Boyle argues that this is generally true for scientific explanation, not just for explaining the beginning of the universe. Thus, he writes, "every distinct portion of Matter, … [is in] an Innumerable company of other Bodies … all … governed as well by … *The Vniversall fabrick of things*, as by *the Laws of Motion*

(BW 6:275, Birch 1772, III:298). The philosophical implications of this point are discussed by Michael Polanyi in Polanyi 1958, 328-331. Polanyi makes the Boylean point that "*the class of things defined by a common operational principle cannot be even approximately specified in terms of physics and chemistry*" (329, Polanyi's emphasis).

38. BW 5:305, Birch 1772, III:15. Cf. Newton: "The matter of all things is one and all the same, which is transmuted into countless forms by the operations of nature." (Newton 1962, 341) and Huygens: "[It is] accepted by almost all modern philosophers that it is only the figure and motion of the corpuscles of which all things are composed that produces all the admirable effects we see in nature" (Huygens to Paul Pellisson, August 15, 1679, Huygens 1888, 8:198).

39. 'believe a Deity' = 'believe in a Deity'. Boyle's usage was standard at the time.

40. Reading "was, and still is," for Birch's and the first edition's "is, and still was."

41. Something is *heteroclite* if it is unusual or in some way anomalous. Boyle also refers to God as a being that is "heteroclite."

42. BW 10:174, Birch 1772, V:140, reading "not well" for Birch's and the first edition's "well." The *possibility* that the laws of nature could vary over time receives serious scientific consideration today: see Collins 2001.

43. Already in 1277 Stephen Tempier, Bishop of Paris, had condemned the view that God could not move the Cosmos in a straight line (though doing so would not change its Aristotelian 'place'). The Aristotelian claim that such movement would leave a vacuum was not relevant, for God can bring about states of affairs that are "impossible according to nature." (See propositions 66 and 17 of the 219 condemned.)

44. See further, BW 9:409-10, Birch 1772, IV:459. "To be convinced that there is never a body without movement," Leibniz suggested, "one need only consult the distinguished Mr Boyle's book attacking absolute rest (Leibniz 1704, 53)," and recently Catherine Wilson has concurred: "For Boyle, intestine motion is occurring always, in solids as well as liquids (Wilson 1995, 52)," but Boyle himself was characteristically more cautious: "since I consider that we are not yet sure, but that though many of the parts of solid Bodies may not be *always* moveless, yet some others of them may *sometimes* for a while at least, be at perfect Rest; I shall conclude as I began, and without resolutely denying that there can be any such thing *in rerum naturà*, as absolute Rest, I shall content my self to say, That 'tis not either absurd to doubt whether there be or no; nor improbable to think that there is not, since we have not found it in those very Bodies, where with the greatest likelihood it might have been expected." (BW 6:210-11, Birch 1772, I:457.]

45. See, e.g., §§29 and 55 of Leibniz's fifth letter in the Clarke-Leibniz correspondence, Clark 1738,

4:639, 4:651. For suggestions that Boyle did reject absolute space and time see McGuire 1972, 532, and Alexander 1985, 75.

46. Substantial forms are useless for scientific explanations, and they have as well a number of internal difficulties (see, e.g., BW 5:454, Birch 1772, III:117). However, Boyle does not on this account rule out substantial forms altogether: "when ever I shall speake indefinitely of Substantiall forms, I would always be understood to except the Reasonable Soule, that is said to inform the humane Body; which Declaration I here desire may be taken notice of, once for all (BW 5:300, Birch 1772, III:12)."

47. Strictly, Leibniz distinguished between momentum, mv, and mv^2, that is, twice the kinetic energy. See *Discourse on Metaphysics*, §17.

48. In the mid-nineteenth century James Joule demonstrated that, given certain assumptions about the number, size, and random motion of molecules, if we identify the pressure P on the wall of any vessel containing a gas with the force per unit area exerted by the molecules in their collisions with the container, we can then prove

$$PV = \frac{Nm_0\bar{v}^2}{3}$$

where N is the number of molecules in the container, m_0 is the mass of each molecule, $\bar{v}^2$ is the mean-square speed of the molecules and V is the volume of the container. If we further assume that $\bar{v}^2$ remains constant when the temperature of the gas remains constant, then this formula expresses Boyle’s law, for all the quantities on the right-hand side of the equation are constant.

49. Four important discussions of the issues raised by the Thomistic/Aristotelian account of the soul are Edwards 1979 and 1985, Owens 1974, and Kenny 1993.

50. More 1662, 2:121; BW 12:390, Birch 1772, VI.689 (BP 1.125 has "dare not" for "cannot"); Locke, *Essay*, 2.19.1. That "the soul never thinks without an image," is *De Anima*, 431a, 15-17; St Thomas agreed: "In the present state of life in which the soul is united to a passible body, it is impossible for our intellect to understand anything actually, except by turning to phantasms (*Summa Theologiae* 1:84.7 c)."

J.J. MacIntosh
macintos@ucalgary.ca

First published: January 15, 2002

Content last modified: January 15, 2002

Divine Providence

Traditional theism holds that God is the creator of heaven and earth, and that all that occurs in the universe takes place under *Divine Providence*-that is, under God's sovereign guidance and control. According to believers, God governs creation as a loving father, working all things for good. Moreover, it is said, God is an absolutely perfect being. He is, first of all, omniscient or all-knowing: he knows of all truths that they are true, and of all falsehoods that they are false, whether they pertain to past, present or future. And God's knowledge does not change. Nothing is learned or forgotten with him; what he knows, he knows from eternity and infallibly. Second, God is omnipotent or all-powerful: anything that is logically possible, he can do. Finally, God is perfectly good: in all circumstances he acts for the best, intending the best possible outcome. Given these suppositions, our initial expectation would be that creation is ordained to perfect good: that as creator God pitches his efforts, which none can resist, toward accomplishing the greatest good imaginable, and hence that the world in which we find ourselves is, as Leibniz put it, the best of all possible worlds. But alas, the evidence is otherwise. The world may contain much good, but it is also a place of suffering, destruction, and death. Life is brief, and afflicted with sorrows of every kind-as often as not with no discernible purpose at all, much less a good one. And it ends for each of us in personal destruction-in death, which trumps all worldly hopes, and conceals impenetrably any experience that may lie beyond. Nor are these mere human hardships. Every living thing dies, all that is beautiful perishes, everything nature builds is destroyed. Indeed, if science is right not an atom, not a photon will escape the cauldron of the universe's final collapse. How can all of this be, if God's nature is as tradition postulates?

- Related Entries

---

1. Logical Consistency and Inductive Evidence

What is described above is the *problem of evil*. In its classical formulation, it is a problem of logical consistency. The opponent of theism alleges that a triad of properties traditionally held to belong to God's nature-omniscience, omnipotence and omnibenevolence-are not jointly consistent with the existence of evil in the world. An omniscient God, we must assume, would have knowledge of the evil in the world. An omnibenevolent God would desire to halt or prevent it, and an omnipotent God should be able to do so. Yet evil is rife. It must be, then, that God lacks at least one of the triad of attributes, and perhaps all of them. Perhaps as creator he is somewhat in the dark as to what evils may occur, and once they appear it is too late to forestall them. On the other hand, it may be that evil is endemic-built into the structure of any world, so that even God is powerless to prevent it. Or, maybe he just doesn't care, and has long since turned his attention to projects more interesting than nurturing our feeble destinies. Whatever the reason, the argument runs, he is not the God of Abraham, of Jesus, and of Mohammed. Their God simply would not permit the suffering and duress under which all creation labors. So while the presence of evil in the world does not serve to prove there is no God at all, it does show there is no God of the kind adumbrated in religious tradition.

The logical problem of evil may be countered with a logical rejoinder. The fact is that the three perfections described above are not by themselves sufficient to exclude the existence of evil in creation. To get that result we have to add a crucial premise to the argument put forth by opponents of theism: that there can be no justification for evil, no good reason why a God with the attributes in question would create a world that contained it. But why suppose this is so? Perhaps there is some good or goods that are possible only in a world that contains or at least permits evil, and without which creation would be vastly inferior to what it is. If that were true, then an all-good and all-loving God would not shrink from creating a world that contained the evil necessary for that good or goods to be achieved.[1] So there is no final inconsistency here, and the theist may wish to conclude on that basis that the problem of evil is resolved. But not yet. For the opponent may concede that the presence of evil in the world does not *entail* that there is no God of the kind religious tradition postulates. Still, he may hold, it gives us good inductive reason for thinking there is no such God.[2] The pervasiveness and profundity of the evil that occurs, the fact that it so often falls upon the innocent and helpless, and the simple fact that we can see no good coming from most of it are more than enough reason, according to this argument, for any rational person to reject the God of tradition. What good could possibly justify the Holocaust, or wholesale destruction of civilian populations in war? Or, lest the numbers submerge the agony, consider just a single case of innocent suffering, posed by William Rowe-a fawn burned horribly in a forest fire somewhere removed from any human awareness, doomed to days of lingering suffering before inevitable death.[3] We are unable to discern any good coming from this single instance of evil, and the same could doubtless be said for millions of others. What more reason could a rational person demand for rejecting the God of our fathers?

This so-called experiential argument from evil may be met with a response similar to the one directed against the logical argument, for the fact is that it too involves assumptions which, when brought to light, seem questionable. It assumes that for each instance of evil that occurs, we humans will be able to detect any good toward which it is directed, and that we will be able to tell whether the good is achieved, whether it was worth the evil sustained in reaching it, and whether it could better have been achieved without the attendant evil. Again, however, why assume any of this is so? An all-powerful, all-knowing and all-loving God could easily have aims exceeding any we have ever imagined. How they are achieved at all, much less the role sin and suffering may play in their achievement, could in principle escape us utterly.[4] If so, then we would be in no position to make the kinds of determinations about the role evil plays in the world, and how dispensable it may be, that the experiential argument presupposes. Still less should we expect to be able to make such determinations in every case, which is what the argument demands. As with the logical problem of evil, then, the theist may greet the experiential problem with a stand-pat position. Neither argument goes through unless we make assumptions we have no reason to make, and which when brought to light seem positively implausible.

Still, evil is troubling, and anyone troubled by it is likely to be left unsatisfied by the stand-pat response. The Western religious tradition is at home with the concept of mystery: it speaks often of aspects of God and his relationship to the world that outreach us, in that our intellects are not finally able to grasp them. But seldom if ever does the tradition treat mystery as totally impenetrable. Just the opposite: the whole point of the theological enterprise is to enable the believer to understand, however imperfectly, the nature of God and the plan of salvation. It is hard to see how this aim can be achieved if a phenomenon as central as evil must be held to escape all comprehension, nor is there any special reason to expect such a thing. Rather, it is in keeping with the hope of the believer that there should be available some glimpse of the good accomplished through the presence of sin and suffering in the world, even if in the actual struggle with them one's best ally is faith. If this is correct, then the theist should not limit his options to the negative. Beyond pointing out the shortcomings of the opponent's arguments, he can and should try to offer a positive *theodicy*-that is, an account of the role evil plays in creation, and a justification for its presence. Such an effort is likely, of course, to end up incomplete. In particular, the theist may be unable in many cases where evil occurs to point to a good to which it is indispensable. But he may be able to offer a general justification for the presence of evil, and to describe some good or goods which but for the occurrence of evil could not be achieved. The question is whether he can do so without compromising divine perfection.

2. The Free Will Defense

Not all theodicies are plausible, and some are so misguided as hardly to be worth pausing over. It will not do, for example, to claim that evil is nothing but a subjective illusion, that if only we could perceive things clearly we would see evil does not exist. As many have pointed out, even if this were true the illusion itself would still be an evil. Moreover, this stance is hardly in accord with the scriptural tradition, in which evil is cause for the utmost concern, both divine and human-even, in Christian theology, to the point that God's own Son must die to set things right. Nor will it do to say that the amount of evil in the world is negligible, that most people are nice most of the time, and that human sufferings are small and

inconsequential compared with the joys of life. This too makes light of the situation. Perhaps the amount of evil in the world can be overemphasized, but it is certainly sufficient to deserve our attention. Not every life contains more joy than sorrow, and evils such as genocide and full scale war are simply horrendous. Furthermore, evil would be a problem philosophically no matter how little of it there were, unless we could see justification for it.[5] A perfect God does not make mistakes, even little ones.

It may be possible, however, to minimize God's involvement in the evil of the universe. That is the aim of what is perhaps the most prominent strategy employed in recent theodicy, which is based the concept of free will, and its importance in the plan of creation. The free will defense begins by distinguishing two kinds of evil. *Moral evil* is evil that occurs through rational action-that is, through wrongful exercises of will on the part of rational beings. *Natural evil*, by contrast, is owing entirely to the operation of natural causes. To see how this distinction works, we need to realize that moral evil can itself be divided into several categories. First come exercises of will that are sinful in themselves, and these are of two kinds. They include wrongful acts of intention formation, as when one maliciously decides to kill another, and the volitional activity through which we execute wrong intentions-e.g., the effort of will aimed at carrying out the intention to murder. The moral wrong of these exercises of will is *intrinsic* to them. They are sinful in themselves, and would be so even if, through some fortuitous circumstance, the attempt to kill went awry, and the intended victim was not harmed at all. Suppose, however, that the action succeeds, as it does in most instances of wrongful willing. If so, further evil will occur-in the present case, the death of the victim. Now if the victim had died entirely as a result of natural causes, his death would have counted as a natural evil. In this case, however, it counts as a kind of moral evil, for while its occurrence requires the cooperation of natural causes, those causes are set in motion by the killer's volitional activity. Harm and suffering that are caused by wrongful willing count as *extrinsic* moral evil, in that they are caused by acts of will that are morally evil in themselves, or intrinsically.

The significance and pervasiveness of extrinsic moral evil is easy to underestimate, because a lot of the suffering and hardship that belongs in this category tends to masquerade as merely part of the human condition, and hence as natural evil. But it is not so. Many of the hardships that befall humankind-disease, ignorance, poverty and the like-owe their existence at least in part to wrongful willing. The poverty of some is owing to the greed of others; suffering and deprivation may occur because of institutionalized racial and ethnic hatred, or because leaders use their positions to advance their own power and prosperity at the expense of their citizenry, or simply because the cost of defense against foreign enemies brings economic hardship to a nation. In other cases the cause is sheer laziness, or the fact that time and talent that might have been devoted to good are instead consumed by selfish ends. Who can estimate how much of suffering and disease, of poverty and ignorance, or of the threat posed by natural disasters would by now have been conquered were not so much of our energy and resources diverted either to the pursuit of wrongful goals, or to guarding ourselves against those who do pursue them, and mending as well as we can the harm they cause? If human wills were not so often misdirected, human life would be transformed, and the struggles against those evils that seem to us no one's fault much further advanced. A great deal, then, of what we are likely to view as natural evil actually falls under the heading of extrinsic moral evil.

That all of sin and so much of suffering counts as moral evil is advantageous to free will theodicy, for

according to the free will defense moral evil is not to be blamed upon God. It is entirely our fault-that is, entirely the fault of rational beings who employ their wills to pursue evil. This is because we have *free will,* which is to be understood here in what is known as the *libertarian* sense. We exercise libertarian freedom in forming or executing an intention only if our deciding or willing is not the product of deterministic causation-that is, provided there is no set of conditions independent of our exercise of will which, together with scientific law, make it certain that we shall decide or will as we do. Independent conditions-our motives and beliefs, for example-may incline us toward one or another intention or action. But they cannot guarantee it, because what we decide and what we strive to achieve is finally up to us. Were it not so, we could not be held accountable for our actions. We would be no more responsible than someone who acted out of a psychological compulsion such as kleptomania, or who was a victim of addiction, hypnosis or the like.

Given the nature of libertarian freedom, then, our actions are up to us, in that they are not brought about by independent events. And because this is so, according to free will theodicy, moral evil is entirely our fault. God is not to be blamed for it, because it owes its existence to our wills, not to his.[6] God is, of course, responsible for having created a world that contains beings with free will. But proponents of the free will defense can point to two reasons that could justify God in populating the universe with such creatures. First, they are an enhancement to creation. Creatures with free will are sources of spontaneity in the world, able to choose for themselves the principles by which their conduct will be guided. As such, they display the kind of liberty we take God himself to have, and so are made in the image of their creator. Second, God endows us with this power because he desires creatures who will accept him freely, who will love and obey him not because they are programmed to do so, but as a matter of spontaneous choice. That we should come to love God in this way is far more satisfactory than that we should be driven to accept him. As in strictly human affairs, forced affection is a pale substitute for love voluntarily bestowed. But, the argument goes, God cannot endow us with free will without running the risk that some of us, at least, will turn against him, and use our freedom to seek evil ends. True, such ends are usually not achieved unless nature, which is God's creation, cooperates. The bullet could not find its mark or the poison be effective unless the relevant natural laws stayed in place. But freedom would be a sham if an evil will could never have its way, and the locus of sin lies not in the consequences of evil willing but in the willing itself. Moreover, even though the harm we cause through our actions requires God's cooperation, it too would not occur but for our choosing and willing as we do. The price of freedom, then, is moral evil. But moral evil is to be laid at our doorstep, not God's, for it is we who choose it. God merely permits our choices and makes them efficacious. In this he is justified, first because of the good of there being free creatures in the universe, and second because some, and perhaps many or even all, of these creatures will come to enjoy God's eternal friendship, by choosing freely to love and serve him.

The free will defense does not, in most formulations, offer a complete solution to the problem of evil. It deals only with moral evil, and although we have seen that this category covers more than might at first be supposed, it does not appear that all of the sorrows and failures of the world can be gathered under it. Even if the free will defense succeeds then, there will remain a residuum of natural evil to be addressed. It may, however, be questioned whether the defense does succeed. One criticism of it raises a question about the relation between creaturely freedom and God's activity as creator. J. L. Mackie has argued that if God is truly all-powerful, he ought to have been able to create creatures who possessed free will, but

who never did wrong.[7] If that were possible, then we would have had a universe free of moral evil, even though it contained creatures with free will. Perhaps the sinful populace that presently inhabits the world would have lost out on such a scenario: maybe God would have had to create an entirely different crowd. Still, moral evil would have been banished, and the condition of the world doubtless vastly improved. But could God have exerted such control over creation? Proponents of the free will defense have tended to think not. In order for God to provide creatures with meaningful freedom, they argue, God must relinquish control over how that freedom is exercised. Were it not so, libertarian freedom would be destroyed: our decisions and actions would not finally be up to us, but would instead be manipulated by God.[8] Even if they were exempt from natural causation, they would still be determined by God's will, and so would be unfree. Indeed, the argument runs, it would be logically impossible for God to create creatures possessed of libertarian freedom, and at the same time have the operations of their will fall under his creative *fiat*. Now it is not usually considered a failure of omnipotence for God to be unable to do what is logically impossible. We need not, therefore, relinquish the claim that God is all-powerful. Rather, the theist concludes, Mackie is simply mistaken in thinking such a God could create free creatures with a guarantee that they would never sin.

It should be pointed out that in giving this response, proponents of the free will defense are making an important assumption about the relationship between God's will as creator and ours as creatures-namely, that God's will operates in the same way natural causes do. That is, his *fiat* as creator counts as an independent condition or event, which causes the occurrence of what he wills in just the way natural causes produce their effects. So if, as creator, God wills that I decide to attend a concert this evening, then my decision to do so is causally determined, just as it would be had I been driven to it by an insatiable desire for Beethoven. Otherwise, we would not have a violation of the criterion for libertarian free will given earlier. Now we shall eventually see that this model of the relation between God's will and the world is at best unlikely, but let us suppose for now that the theist's answer to Mackie's complaint stands. Is the free will defense then successful? Again, it seems not, for the antitheist can still raise two complaints. First, he may argue, even if the free will defense does not violate God's omnipotence, it still violates his sovereignty. If God were fully sovereign over the universe his rule would be complete. All that occurs would be under his direct control, down to the smallest detail. According to the free will defense, however, this is not so. Rather, as we have just seen, proponents of this kind of theodicy insist that some of what goes on in the world is not under God's control, but under that of his creatures. To be sure, God need not have created free beings, and when they engage in sinful willing he can always thwart their ends by manipulating natural causes. But he cannot stop them from sinning (or, for that matter, from willing well), for both of these lie with the will itself. So in creating free creatures God relinquishes part of his sovereignty over the universe.[9]

Second, the antitheist may argue, the free will defense violates divine omniscience. For if I possess libertarian freedom, then whether I decide to go to the concert tonight is neither under God's direct control nor controlled by natural causes. And if that is the case then God has no way of knowing what I will decide. Like anyone, he can make a lucky guess: he may believe devoutly that I will decide to attend the concert, and that may turn out to be correct. But lucky guesses do not count as knowledge. Knowledge requires reasons, and God can have no satisfactory reason for his belief, if by "satisfactory" we understand what omniscience would seem to require-namely, a reason that guarantees correctness.

Rather, like any observer, God must wait to learn what my decision will be in order to be sure of it. But then throughout the time prior to my act, God is not omniscient. There is a truth about the future he does not know. Thus, the antitheist may conclude, the free will defense is in fact a failure. It exonerates God from direct responsibility for sin, but it does so only by surrendering part of God's sovereignty, and by making him fail, as creator, to be omniscient.

3. God's Knowledge of the Future

It is fair to say that philosophers responding to these difficulties have been more concerned to preserve God's omniscience than his sovereignty. Perhaps this is in part because philosophy is itself a matter of pursuing knowledge, so that philosophers are led to value omniscience more highly. Were we generals, say, or politicians, our priorities might be quite the opposite. In any case, most discussions of the seeming conflict between creaturely freedom and divine perfection have concentrated on the task of reconciling as far as possible the assertion that we have free will with the claim that God is all-knowing. One tactic for so doing is to hold that God cannot be faulted for not knowing in advance how we will exercise our freedom, since until we do there is simply nothing to know. According to views of this kind, not all propositions about the future have a truth value. Some do, of course: it is a necessary truth that 2 + 2 = 4, and this proposition has as much bearing on the future as it does on the present and past. Similarly, a proposition concerning the future may have a truth value when its truth is causally determined. Consider, for example, the proposition that the sun will rise tomorrow. Most likely, it is true. True or false, however, this proposition's truth value is fixed by causes already in place, causes that determine either that the sun will rise tomorrow or that it will not. But now consider the claim that I will decide an hour from now to attend a concert this evening. If I have free will, there are no conditions presently in place that determine whether I will so decide. This being the case, according to the present view, the proposition that I will decide in an hour to attend the concert is neither true nor false. It has no truth value at all, nor does any other proposition that describes a future free decision or action. But then, the argument runs, it is not a mark against his omniscience that in creating us, God does not know how we will exercise our freedom. It is logically impossible to know of a proposition that it is true or that it is false if it is neither. And inability to know what logically cannot be known does not harm God's omniscience, any more than inability to do what logically cannot be done harms his omnipotence.

If correct, this view would indeed reconcile divine omniscience and creaturely freedom, leaving only the problem of sovereignty to be addressed. But there are telling arguments against it. Propositions that venture to predict future free decisions and actions do appear to have truth values. One indication of this is that we believe and disbelieve such propositions, and what is it to believe a proposition but to believe it is true, or to disbelieve it but to believe it is false? Nor does it seem possible to worm our way out of this. Let *p* be the proposition that I will decide to attend a concert this evening. It might be protested that for someone to believe I will so decide is only to believe *p* will *become* true at the appointed time-i.e., at the time of my decision. But that does not solve the problem, for to believe this is simply to believe another future-tensed proposition-namely, "*p* will become true"-which is just as dependent on my decision for its truth as is *p*. Similarly, it will not do to claim that to believe *p* is not to believe it is true but only that it is likely or probable. For to hold these beliefs is just to hold, respectively, that is likely that *p* is true, or

probable that it is true. In short, there seems no avoiding the fact that to believe *p* is to be committed to its truth, pure and simple. Moreover, anyone thus committed would, if I later decide to attend the concert, be justified in saying they had been right about what I would decide, that their earlier belief had been correct. And again, what is it for a belief to be right or correct except for it to be true?

It seems unlikely, then, that one can save God's omniscience in creating the world by holding that propositions about future exercises of creaturely freedom have no truth value. But this is hardly the only option. Another alternative is to hold that God's position with respect to time is such that unlike us, he does not have to wait for the future to unfold in order to know its contents. Such was the view of Boethius, who held that God exists entirely outside of time, in a kind of eternal present to which all that occurs in time is equally accessible.[10] Thus, God is able in a single act of awareness to comprehend all of history, the past and future as well as the present, just as though they were now occurring. Many philosophers have followed Boethius in this, holding that God is in no way a temporal being, but is rather the creator of time, with complete and equal access to all of its contents.[11] And it may well appear that on such a view God's omniscience is restored, in that he has immediate cognitive access to everything that will ever occur. Moreover, there is no conflict with libertarian freedom, since the vantage point from which God knows our decisions and actions is completely external to time. This makes all talk about "when" God knows about our actions pointless. He simply knows them, in a unified, timeless and unchanging act of comprehension that comprises all that ever was or will be.

Some philosophers have criticized the idea that God is timeless.[12] But even if the Boethian position is correct on this score, the usefulness of this means of reconciling divine omniscience and human freedom is highly questionable. The difficulty is that in order for God to exercise full providence over the world, he needs to know *as creator* how the decisions and actions of creatures with libertarian freedom will go. It is hard to see how that is possible on the Boethian view, for even if God is outside of time, his activity as creator is still *ontologically* prior to the activities of free creatures on this account, whereas his knowledge of those activities is posterior to them. Thus, it seems impossible that God's creative will could be guided by his knowledge of our actions, even if, from his timeless perspective, such knowledge is finally available to him. If this is correct, then even the Boethian God runs an immense risk in creating the world. He can only hope that we will use our freedom justly and wisely, perhaps making some allowance for the possibility that we will not, but otherwise simply trusting in the outcome. So although God's omniscience may be restored by placing him outside of time, he is in no way empowered by it. It may be questioned, furthermore, whether this view of things really is consistent with the claim that God is timeless. The Boethian picture appears to call for a kind of transition, wherein God first creates free creatures in ignorance of what their actions will be and then learns about those actions by observation. But if that is so then there appears to be change in God, in which case he would have to be a temporal being after all. Now perhaps there is some way around this problem: maybe as creator God somehow operates in isolation from certain parts of his knowledge, while having access to all of it in his role as knower-and yet remains timeless in both capacities. Still, it is not satisfying that God should be limited in this way. His activity as creator ought to be completely unhampered. Moreover, the bifurcation in God's thinking called for by this adjustment only reinforces the difficulty of treating creaturely decisions and actions as falling under divine providence. As creator, on this account, God really does not know what kind of world he is creating: how evil it will be, whence the evil will arise, and how to anticipate it in

detail in the plan of creation. A better solution would be preferable, if one can be had.

4. Middle Knowledge

A possible solution to this problem was posed by the sixteenth-century Spanish Jesuit, Luis de Molina.[13] According to Molina, God is able to know *as creator* how any exercise of creaturely freedom will go. That is, God knows, for any creature he might create, how that creature will behave in whatever circumstances he might be placed. God is able to know this, moreover, even though the creatures in question will, if created, enjoy libertarian freedom. This kind of knowledge, which Molina called *middle knowledge*,[14] is comprised in what we may call *subjunctives of freedom*. Consider, for example, the situation in which I will find myself later today, when I deliberate about whether to attend the concert tonight. It is possible to formulate two subjunctive conditional propositions about that situation. The first states that if ever I were placed in the circumstances (call them C) that will then obtain, I would decide (freely) to attend the concert; the second states that in those circumstances, I would not so decide. Let these be symbolized as $C \Box\!\!\rightarrow p$ and $C \Box\!\!\rightarrow {\sim}p$, respectively. Both $C \Box\!\!\rightarrow p$ and $C \Box\!\!\rightarrow {\sim}p$ count as subjunctives of freedom, and since we are imagining that I will be placed in C later today, it is plausible to think that one or the other of them is true. Let us suppose it is $C \Box\!\!\rightarrow p$. According to defenders of middle knowledge, God knows prior to any creative act on his part that $C \Box\!\!\rightarrow p$ is true; hence he knows this prior to my existence, and prior to any act of mine. We need not take the "prior" here as temporal, if we hold that God exists outside of time. The idea, rather, is that the truth of $C \Box\!\!\rightarrow p$ is *logically prior* to any creative decision on God's part, and to any doing of mine, in that its truth is fixed independently of these matters. Further, God is able to *know* that $C \Box\!\!\rightarrow p$ is true independently of any decision of his, and without appealing to any actual decision on my part as evidence. Finally, God is armed with true subjunctives of freedom for every other set of circumstances in which I might ever have been placed, and the same for every other free individual he has the option of creating, whether he actually chooses to create the creature or not. In effect, then, middle knowledge gives God advance notice of every free decision or action that would ever occur, on the part of any creature he might create.[15]

Assuming it is a legitimate notion, middle knowledge does much to restore God's providence in creating free creatures. Once armed with information about how such a creature would decide and act in the various circumstances in which he might be placed, God has the option of not creating the creature, or of creating him in whatever circumstances are called for by the subjunctives of freedom God wishes to be reflected in the actual world. Now of course the circumstances in which one creature is placed may depend in part on how others choose to exercise their freedom. But the willings of those others can in turn be providentially arranged, since they too fall under middle knowledge. In principle, then, nothing need occur in the actual world that does not have God's prior recognition and consent, at least. There may, of course, be much that does not go as God would prefer. It is important to realize that middle knowledge does not restore complete sovereignty to God. If $C \Box\!\!\rightarrow p$ is true, then there is no way for God to create me in circumstances C and have me do anything but decide to go to the concert. The best he can do is alter my circumstances to fit some true subjunctive of freedom that has another outcome. And the same goes for the subjunctives of freedom that hold of all other creatures God might create. This means there is quite a range of worlds which, though logically possible, are not *feasible* for God, in that they are beyond

his reach as creator.[16] From God's point of view, free creatures will behave as they will behave, and that is that. Still, God can know in advance of creation what worlds are feasible, and can plan accordingly, which is a vast improvement over the Boethian view. The position as regards omniscience is also improved. There is still a kind of transition called for-this time commencing from a point at which God merely contemplates the possibilities of how things *might* go with creation, to a point at which, having decided what creatures and circumstances will populate the world, he knows how things *will* go. Again, however, it might be possible to work out a way in which the transition can be understood non-temporally. And in any case the Molinist position represents an advance over that of Boethius, in that now God's knowledge about how exercises of creaturely freedom will actually go can be derived from his decision about what creatures and circumstances he will create, rather than awaiting the actual decisions and actions of free beings. This diminishes God's passivity; it enables him to know *as creator* how the history of creation will unfold.

But *is* middle knowledge a legitimate notion? Many have thought not. Perhaps the most serious objection against it is that there does not appear to be any way God could come by such knowledge. Knowledge is not merely a matter of conceiving a proposition and correctly believing it to be true. It requires *justification*: one must have good reasons for believing. But what justification could God have for believing the propositions that are supposed to constitute middle knowledge? The truth of subjunctives of freedom cannot be discerned a priori, for they are contingent. It is not a necessary truth that if placed in circumstances *C*, I will decide to attend the concert tonight. Nor can we allow that God might learn the truth of $C \mathbin{\square\!\!\rightarrow} p$ from my actual behavior-that is, by observing that I actually do, in circumstances *C*, decide to attend the concert. For God could not make observations like this without also finding out what creative decisions he is actually going to make, which would destroy the whole purpose of middle knowledge. Instead of being guided in his creative choices by knowing what decisions creatures would make *if* they were created, God would be presented from the beginning with a *fait accompli*-with the reality that he was *going* to create certain creatures, and they were going to behave in certain ways. For God's options as creator to remain truly open, middle knowledge must have some other justification. Furthermore, it seems clear that observation of the actual behavior of creatures could not possibly inform God of the truth of those subjunctives of freedom that delineate the behavior of creatures he will *not* choose to create, for in their case there is no pertinent reality to consult. Yet Molinism wishes to allow for the possibility of such creatures. It is apparent, then, that neither conceptual resources nor resources founded in the concrete world will enable God to know in advance of his decisions as creator which counterfactuals of freedom are true. If there is a third resource, no one has said what it is. Thus, while God may firmly believe certain subjunctives of freedom, there appears to be no justification available to him that would allow such beliefs to constitute middle knowledge.[17]

Note that the above objection is not based on the claim that subjunctives of freedom lack truth values, or that their truth is not properly grounded.[18] That may indeed be a problem for some subjunctives of freedom, but it is not a problem for $C \mathbin{\square\!\!\rightarrow} p$. On the usual understanding, a subjunctive of freedom counts as true provided that, among worlds in which its antecedent is satisfied, there is at least one in which the consequent is satisfied as well, and which is more similar to our world than any in which the consequent is not satisfied. Now no world can be as similar to the actual world as that world is to itself, and we are assuming that *C* and *p* are true in the actual world. Accordingly, $C \mathbin{\square\!\!\rightarrow} p$ must be true as well. The only

way to avoid this outcome is to deny that propositions like *p*-that is, propositions which describe future free decisions in the actual world-have truth values, and we have already seen that this will not do. So $C \,\square\!\!\rightarrow p$ is perfectly well grounded. The problem is only that it is not grounded in the way it needs to be to serve as middle knowledge. It is not grounded *independently* of God's or my free decisions. In light of this, we can only conclude that the Molinist effort to reconcile creaturely freedom with God's omniscience and sovereignty as creator fails. We have yet to see how God can know as creator what decisions and actions his creatures will engage in, while at the same time upholding the idea that those decisions and actions are manifestations of libertarian freedom.

5. A Weaker Alternative

One may be tempted at this point simply to throw in the towel, to give up the endeavor to reconcile libertarian freedom with divine sovereignty and omniscience. If so, we may still insist on libertarian freedom for creatures. But if we are convinced this is incompatible with holding that God is omniscient, and that everything that takes place in the created world falls under his complete governance, then these claims will go by the board. On this type of view, God is a temporal being who, like us, must await the actions of free creatures in order to know with certainty what they will be. And much that occurs, most especially sinful decisions and willings, will not be of his choosing. Not that he is completely in the dark: God can still have probabilistic knowledge of how his creatures will act, and he can contrive to place them in circumstances designed to elicit if possible whatever behavior will achieve the most good. And of course he still has the power to motivate and punish, so creatures may be guided toward right paths. But on this scenario God's aims as creator can only be achieved-assuming they will be achieved at all-by taking risks. Inevitably, creaturely free will makes for a setting of uncertainty, and only within that setting can God attempt to bring creation to a happy outcome. Yet he proceeds, and his doing so is a measure of his love for us.[19]

Such a position may appeal to philosophers who find the God of perfect being theology too remote and mysterious to equate with the God of scripture. But this viewpoint faces serious problems. Some are relatively specific. For example, it is hard to see how, if even God does not know what they will be, the actions of free creatures could be the subject of prophecy. Yet they often are, in scripture.[20] Also, there will no doubt be many cases where multiple free actions impinge on some outcome God desires. When that is so, the probabilities of those actions need to be multiplied to determine God's assurance of the outcome, which as a result could be minuscule.[21] But the biggest difficulty is that this view places God's fate as creator almost completely in the hands of his creatures. No matter how concerned and loving he may be, or how powerfully he attempts to win us over, we are on this view out of God's control. There is always the chance, therefore, that his plans as creator will be utterly dashed, that his overtures to us will be rejected-even to the point, one supposes, of our all being lost-, that we will use our freedom and advancing knowledge to wreak ever greater horror, and that creation will turn out to be a disaster. Willingness to take chances may be laudable in some cases, but surely this level of risk is irresponsible. Moreover, it is completely out of keeping with both scripture and tradition, both of which portray God as above the fray of the world, unperturbed by its mishaps, and governing its course with complete power and assurance. On the present view, divine governance is a hit or miss affair, in which we can only wait

to see whether a somewhat poorly informed God will manage to bootstrap his way to his objectives. Surely, opponents argue, this gives away too much of the traditional notion of providence.

6. The Traditional Solution

If the views considered thus far all fail, there is no choice but to place the decisions and willings of rational creatures under God's creative authority. Only by so doing can we restore to him complete control over the course of events in the world, and only in this way can he know *as creator* what world he is creating, and so be omniscient. If all of our decisions and actions occur by God's creative decree, then all possible worlds are made feasible for him. He can create as he wishes, with full assurance as to the outcome. And he can know how things will go, in particular how we will decide and act, simply by knowing his own intentions as to what our decisions and actions will be. Clearly, there are respects in which this approach is to be preferred. From the perspective of piety, the versions of the free will defense we have seen so far are all troublesome: they seem to place our concern for ourselves above our regard for God, by maximizing our options at the expense of his. One can readily anticipate the response that if complete sovereignty for God and libertarian freedom for his creatures cannot both be had, then the devout (not to say Godfearing) philosopher would be well served to endorse the former, that anything less is not just out of keeping with the mainstream of theological tradition, but actually borders on blasphemy. Yet we have seen that free creatures are of greater value than the unfree, if only because their greater likeness to God makes them a desirable enhancement to creation. The question, then, is whether placing our decisions and actions under God's creative *fiat* leaves in place anything of creaturely freedom, or of the free will defense.

It may seem obvious that neither can survive: that once the operations of creaturely wills are subordinated to God's will, libertarian freedom disappears, and with it any hope of absolving God of moral evil. But at least where freedom is concerned, traditional theology asserts the opposite. Augustine, for example, held that God moves our wills, working in us both to will and to do of his good pleasure, as scripture says (Phil. 2:13). Yet he insists that this does not diminish our freedom, for if it did we would not be told in the same passage to work out our salvation in fear and trembling.[22] Similarly, Thomas Aquinas maintains that all of our doings, even those in which we sin, are on a par with the rest of creation in having God as their first cause. Only the defect of those actions which is their sinfulness derives from us. Sin, he says, is like limping, in which the defective motion arises from the crookedness of the limb, rather than the power of locomotion that impels it.[23] Like Augustine, moreover, Aquinas sees no conflict between God's activity as creator and ours as free creatures. On the contrary: he holds that God's activity as first cause is actually the cause of our freedom, since he moves us in accordance with our voluntary nature.[24] As to how this can be, it has to be said that Aquinas offers little by way of explanation. There is, however, an interesting suggestion in the *Summa Contra Gentiles*, where Thomas maintains that if the will were moved by an external principle as agent, the movement would be violent, and moves to the conclusion that God *alone* is able to move the creaturely will as agent without violence, since he alone is the cause and sustainer of its being, and thus is able to move it from within.[25]

This suggestion cannot be explored fully here,[26] but there are two things to be said in its favor. The first

concerns libertarian free agency itself, which is often portrayed as a power by which we cause, or confer existence upon, our own actions. Such a view would not accord with Aquinas's claim that only God can move the creaturely will in this way, and in fact it is hard to see how this portrayal of agency can be right. If I confer existence on my decision to attend the concert tonight, I must do so either through some act separate from the decision, or as an aspect of the decision itself. If it is through a separate act, the problem of freedom simply shifts its location. We have to be convinced that this act in turn receives its existence from me, and we appear headed for a vicious regress. But neither does it seem possible for me to confer existence on my act of deciding as an aspect of the act itself. For prior to the act's appearance there is nothing to do the conferring, and once it appears, the conferral is no longer needed. So whatever agency and voluntariness consist in, it does not appear to be an ability to confer existence on our own actions. That is important, because it means that unless we can find a way to ground libertarian freedom in the creative activity of God, the decisions and actions in which it is manifested are likely to have no accounting *whatever*-a situation few philosophers are likely to find satisfying, and which hardly puts me in control of my act of deciding.[27]

A second important consideration has to do with the relationship between God's creative will and the things he creates. We are prone to think of this as an event-causal relation, in which God issues a kind of command, and the command in turn produces the mandated effect. Applied to our example, this would mean that God creatively wills that I decide to attend the concert, and his willing then causes me to decide. And of course this sounds exactly like what Aquinas describes as the violent operation of an external principle. On this scenario my deciding is passive, and hence involuntary, because God's creative *fiat* is an event independent of my decision, which by acting upon me robs me of my autonomy. I am reduced to a puppet manipulated by God, rather than a free agent. Clearly, however, this scenario does not reflect the way Aquinas thinks creation works, and on that score there is reason to think Aquinas is right. For consider again the causal relation alleged to obtain between God's willing and mine. Whatever we take this to consist in,[28] it must exist contingently, for causal connections are not necessary beings. But then this supposed causal connection must also be created by God, and if that occurs through another process of command and causation, we would be facing another regress. The only way out is to hold that God *directly creates* the causal nexus-which is to say that in its creation, the nexus itself, not some command, is the *first* manifestation of God's creative activity. But if God's creative will can be directly efficacious in this task, then it can also be so in his creation of us and our actions. There is, then, no need for a nexus to explain the efficacy of God's creative will, nor is there any causal distance whatever between God and either us or our behavior. Rather, we and all that we do have our being *in* God, and the first manifestation of his creative activity regarding our decisions and actions is nothing short of the acts themselves.

If this is correct, then as Augustine and Aquinas both insist, God's creative activity does not violate libertarian freedom, for it does not count as an independent determining condition of creaturely decision and action. On the contrary: assuming God's own will is free, there is no event in heaven or earth that is independent of my deciding to attend the concert tonight, and which causes my decision. God's creative activity does not act upon me or render me passive in any way, for it consists solely in God's freely giving himself over to being the *ground of being* for me and all that I do. Accordingly, I can still display libertarian freedom. My decision is a spontaneous display of creaturely agency, free in the libertarian

sense because it does not occur through event causality, and because in it I am fully and intentionally committed both to deciding and to deciding exactly as I do. There are no further legitimate requirements for libertarian freedom. There is, of course, something that cannot happen on this view: it is not possible for God's activity as creator to be devoted to my deciding to attend the concert, and yet that I should forebear to decide at all, or decide to do something else. But that is not because if I were to try it, I would find myself in a losing battle with God's efficacious will. It is because there is no manifestation of that will regarding my decision short of the decision itself. The impossibility that God's will as creator and mine as creature should diverge suggests trouble because we can view what goes on either from God's perspective or from mine. That suggests two events, and a potential conflict between them. Properly interpreted, however, the traditional view appears to call for only one event, and as far as *it* is concerned, all the impossibility comes to is that I cannot at once both make a decision and not make it. To be incapable of the logically impossible is not a failure of libertarian freedom.

7. Sin

The traditional view is enigmatic, and a lot more would need to be said to make it convincing. But if something like the traditional view can be made to work, creaturely freedom is indeed reconcilable with divine sovereignty and omniscience. Our destinies are entirely subordinate to God's creative will; he exercises full control in all that we do, notwithstanding the fact that our deeds are fully voluntary, and we have every reason to expect that all that takes place in the world will reflect the providence of a perfectly loving father. As for omniscience, here too there is no difficulty. God knows about our decisions and actions simply by knowing his own intentions, for he wills that they occur. Nor is his will exercised from the fastidious distance preferred by Molinists, in which God creates us knowing what we will do, but has no hand in our actually doing it. Rather, God is as much the cause of our sinful actions as of our virtuous ones, or of any other event. Yet, Augustine and Aquinas would insist, he remains perfectly good and absolutely holy, a being deserving of our complete reverence and absolute devotion. How is such a thing possible?

Part of the answer lies in the fact that even if my act of deciding to go to the concert tonight has its existence grounded in God's creatively willing that I so decide, it is still *I* who act, still I who decide. God's willing that I decide as I do does not make my decision God's. Indeed, if it did, if my decision were predicated of God rather than me, his will would fail to achieve its object. But it is not possible for God's will to be frustrated, as long as what he wills is consistent. So regardless of what we may think of the traditional view's contention that divine sovereignty and libertarian freedom are fully compatible, that view does not take the operations of our will, or the actions founded upon them, away from us. They remain our own. Consequently, any sin they involve remains ours also. Thus, if I decide sinfully to go to the concert tonight-if, say, I am neglecting duties I know should take priority-the sin is mine, not God's. If he is to be faulted, it must be for some other reason. It should be noted, moreover, that God's position in this respect is not much different from what it is on the Molinist account. True, that view takes certain things out of God's hands. Whether I would decide to go to the concert in the circumstances in which I will find myself tonight is not, according to Molinism, up to God. But it is up to him whether I shall be created in those circumstances, and indeed whether I shall exist at all. On both views, God knowingly and

willingly creates a world in which rational creatures sin. The difference is that on the traditional view God does have complete control: he can create any possible world, and he is as much involved as creator in those acts in which we sin as he is in any others. And that means the standard free will defense, which works only by diminishing God's authority and circumscribing his providence, is not available.

But another may be. It must be remembered that even though the actions of free creatures do not escape providence, such creatures are still an enhancement to creation, in that their nature reflects more closely what we suppose to be God's own nature. As such, free creatures are more suited to the kind of fellowship with God that believers understand to be their ultimate destiny. It may be, however, that the achievement of that destiny inevitably involves sin. In part, this is because to have free will is to have a nature that is incomplete, in the sense that what we are never fully determines what we shall do. Rather, we have to complete our own nature, by establishing an identity for ourselves-that is, by adopting the patterns of behavior, and the long term objectives that define our lives. And it may be that in so doing we inevitably find ourselves in rebellion against God, simply because as free beings we take ourselves to be establishing our own destiny-and so make that, instead of obedience to God, our first priority. Second, it has to be remembered that true friendship is always voluntary. If God only *exacts* devotion from us, we are reduced to being his subjects. To be friends with him requires a meaningful and responsible decision on our part to accept the offer of friendship he presents to us. But a responsible choice in God's favor requires that we understand the alternative-which is to be at enmity with him. And there is good reason to think such an understanding requires that we sin. Guilt, remorse, a sense of defilement, and the hopeless desolation of being cut off from God cannot be understood in the abstract, because if they are only understood abstractly they are not *ours*. Only through experience can we understand what it means to be in rebellion against God, and we gain that experience by sinning. By turning away from God we realize what it means to be alone, and we learn that however successful they may be, our own projects cannot satisfy us. Only then are we in a position to choose responsibly to accept God's offer of fellowship. Finally, we must remember the simple fact that if we are to experience a transition that ends in our being united to God, that transition can only begin from a place where we are separated from him. It is plausible to think, however, that there is no morally neutral ground here: that unless we are within the circle of God's love, we must be outside it, and that once we choose voluntarily to stand alone, we are already in an attitude of hostility toward God. In short, if we are to come to God as voluntary agents, it may well be that we can only approach him from a position of sinfulness.

If this is correct, there is a great good that God, as a loving creator, is able blamelessly to will for us, but which in its exercise inevitably leads us into blameworthiness. That good is our autonomy-the thing that makes us most like God, and is the sole means by which we are able to reach friendship with him, but which can be responsibly exercised to enter that friendship only if first employed in a conceit of rebellion, wherein we learn our limitations, and come to appreciate the emptiness of a life based on subjective independence. Only thus are we able to reach a position of moral autonomy from which an authentic choice to enter into fellowship with God is possible. Thus, freedom is indeed crucial to moral evil: the implications of libertarian agency are such that its purpose in God's plan could not be achieved without the occurrence of sin. Unlike the standard free will defense, however, this approach does not endanger God's sovereignty or omniscience. As creator, he is fully involved in those acts in which we sin, for they can occur only through his will. But he incurs no blame for them, for they are our acts, not his, and

although they place us in rebellion against him, they do not put God in rebellion against himself. Indeed, no individual can be in rebellion against his own will. So if the essence of sin is rebellion against the will of God, then even though God is the first cause of those acts in which we sin, it is not possible that he himself sin in their occurrence. It is worth noting, too, that the present view makes it possible to explain what, on the standard free will defense, can only be a mystery- namely, that although all of us possess libertarian freedom, and so have the option of serving God, still *all* humans sin. The reason for this is not that God suffers a terrible run of bad luck in a grand lottery of his own institution. Rather, it is because only by passing through sin that the saved are able to achieve their destiny. By creating us in our sinfulness, God assures that each individual will develop an authentic moral identity, and, if the theistic tradition concerning divine justice is correct, prepares each for the eternal recompense appropriate to his character.

Two objections may be raised at this point. First, one may doubt that the kind of moral autonomy described above is truly authentic, given God's role in our actions. As long as God is in the controlling position this view assigns to him, it might be argued, the apparent moral autonomy of free creatures, and the destiny they achieve through it, is a sham-something visited upon them, rather than legitimately chosen. The second difficulty concerns the destiny itself. It is not a part of our religious tradition that all are saved. St. Paul, for example, seems clearly to have believed that some, the elect, are destined from the beginning for salvation, and others not (*Rom.* 9:10-24), and the same appears true of Jesus himself (*Matt.* 26:24, *Luke* 10:20). And it is part of standard theology that, after death, the saved are joined to God in the beatific vision, a state of eternal and indescribable joy. The lost fare far worse. They are condemned to the bitter and devastating frustration of permanent separation from their creator, and on many accounts to a lot of other miseries as well. Now on the present view, God is as much involved in the rebellion of the reprobate as in the conversion of the saved. And one may well wonder what could justify this. Why should a loving God create creatures destined for damnation?

To the first of these objections it may be responded that although all of our destinies are fully in the hands of God, and all of our actions under his authority as first cause, it does not follow that the choices on our part through which our destiny is achieved are anything but fully authentic, or that the moral identity in which we are created by God is imposed upon us. God's will does, of course, determine our decisions in the logical sense: if we know his will for a creature, we can always infer that creature's fate. But that is only a deductive relation among propositions; in itself, it does not necessarily imply any relation among real entities that curtails our freedom. And if, as was suggested above, the first expression of God's will regarding our choices is those very choices themselves, then there is no independent occurrence in the world or in God before which we are rendered passive. On the contrary: our acts of will are fully endowed by God with the characteristics of agency. Thus, our moral identity is in no way forced upon us. It is an identity fully of our own choosing, adopted by us in complete freedom and integrity, notwithstanding the fact that our doing so is entirely within the providence of God. Believers should not, then, fear that their destiny is a sham, or find fault with God for creating them what they are. Rather, the appropriate response would be to follow the scriptural injunction to work out one's salvation in fear and trembling, knowing that an all-powerful and provident God is also working through us to achieve his ends.

As for the second objection, whatever the sufferings of the lost may be, theologians have always agreed that the greatest evil they sustain is final and irremediable separation from God. Nothing could be worse than to be cut off from the love and friendship of a father whose power extends to every detail of the universe, and who invites us to a share in his very life. But if this is the greatest evil of damnation, then no one who ends that way is treated unfairly, for this separation is precisely what one chooses by insisting on a life of rebellion rather than seeking reconciliation with God. Indeed, having once created beings destined to be lost, it is hard to see how a loving God could do anything but honor their choice in the matter.[29] What is troubling, rather, is that he should create such beings at all, much less will their performance of the very actions through which they reject him. It may be argued, however, that even here God's love is at work. He cannot, of course, directly intend the rebellion of sinners, nor the destruction of the finally unrepentant. But the lost are full participants in securing their tragic destiny; and while a life ruined by final rebellion is morally indefensible, it is still morally meaningful. Through their actions, the lost carve out for themselves a character which, though not upright, represents a real option for a free creature. Thus, the argument runs, to the extent that moral autonomy is a good it can be willed for a creature by God even when it takes this form. Furthermore, it is claimed, it is a mistake to think that God is not lovingly involved in the lives of the reprobate, or that he would have been more loving had he not created them. On the contrary: what is meaningless is to suppose that God would have shown greater love toward the lost by omitting them from creation. What is not there cannot be loved. Equally, it is meaningless to think the lost would be better off had they not existed.[30] What does not exist is neither well nor poorly off, nor anywhere in between; and it is as good for the reprobate to have life, the opportunity for salvation, and an autonomous choice as to whether to accept it, as it is for the saved. What is not good for them is the use they make of the opportunity, in choosing to be without God. But that is fully their decision, and its consequences are fully earned.

8. Moral Evil and Defeasibility

If the sort of view outlined above is correct, there is ample reason for God to create a universe in which there is moral evil, for only through the presence of moral evil is it possible for creatures like us to develop a legitimate moral identity, and make an informed and responsible choice to accept or reject God's offer of friendship with us. It is important to see that on such a view, moral evil is not treated as a causal means to the good of our having friendship with God. If that were so, opponents could justly object that God could simply have created us in such friendship from the start, and the defense would fail. Rather, the good on which this kind of theodicy is based is a free and informed choice that can only be made from a position of sinfulness. Rather than constituting a causal *means* to our establishing our own stance toward God, sin is an indispensable *part* of the process-something without which a legitimate choice for or against God's friendship is not just causally but conceptually impossible. But that is not all. In the case of those who choose in God's favor, at least, moral evil is *defeated* in that it is bound up in a total state of affairs that counts as a far greater good, in which the evil is addressed and overcome. The very autonomy that the sinner once insisted upon is surrendered to God, thus becoming the foundation for a new understanding and a richer relationship, in which one is able to act as an informed and wholehearted participant in the divine enterprise of working good. Sin does not function as a causal means in this process, nor is it simply overbalanced by some other good with which it coexists. It

encountered and overcome, through the providential operation of God in creaturely freedom.

The idea that evil is defeasible was developed first by Roderick Chisholm.[31] It is an especially useful notion for theodicy, in that when evil is defeated, the usual objections to the presence of evil in the plan of providence are turned aside. If evil were only a means to good, and were simply outweighed by the goods to which it leads, antitheists could legitimately object that God could have created a better world simply by omitting the evil and creating those goods outright, or obtaining them through means that were not evil. By contrast, when evil is defeated it is caught up in a larger state of affairs that constitutes a far greater good, but not by containing components that might have occurred independent of the evil in question, and simply outweigh it. Rather, the evil is addressed within the larger state of affairs in such a way that it becomes integral to the good through which it is defeated. The defeat of evil is, moreover, an especially impressive sort of good. That Beethoven should have overcome the natural evil of his deafness to write the music he did, for example, strikes us as an amazing good-one far greater than what would have been accomplished had Beethoven written the same music (assuming that to have been possible) with good hearing. Similarly, that moral evil is overcome through the process by which sinners are brought into a right relationship with God may be considered a far greater good than would have been accomplished had he made us a community of spiritual lotus eaters, whose relationship to him, if any, was founded on no meaningful decision, but simply upon our never having had the experience that would make anything else possible. Indeed, given the moral vacuity of such an existence, it is not obvious that such creatures would even be fit for divine friendship, much less able to make the decision through which that relationship is brought to pass.

But of course the same does not occur in the case of the lost. Their rebellion is permanent, and is not overcome through any action of theirs. The antitheist may wish to argue that in this case it is evil that triumphs over good. The character of the lost is permanently corrupted: any virtue that was theirs is turned to wrongdoing, and any hope that they might even achieve peace with God, much less be of useful service to him, comes to nothing. Thus, it might be claimed, the very existence of the reprobate stands as a gratuitous and unanswered defilement of creation, in which evil is victorious. Can the theist point to anything that might reverse this verdict? One possibility is that God himself could take action specifically aimed at defeating all moral evil. This, of course is the defining theme of Christian soteriology. In that tradition, all sin is defeated through the paramount manifestation of God's love for the world, the redemptive suffering of Christ, which could not have occurred unless there were sin, and which makes possible God's offer of salvation to humankind. Not all sinners may accept the offer, but God is reconciled to all, in the sense that the substitutionary atonement of Christ covers all wrongfulness. For the Christian believer this is a *sine qua non*: nothing but the sacrifice of a being who participates fully in the divine nature is sufficient to satisfy fully the demand of divine justice, and unless that demand is satisfied friendship between God and humankind is impossible. Thus the redemptive suffering of Christ is for Christians the ultimate act of mercy and compassion on God's part toward all sinners, and the ultimate defeat of moral evil.

This kind of answer is, however, confined to a particular religious tradition. Is a more general solution possible? There are at least two lines of response the theist can take up at this point. First, he can argue that the moral evil wrought by unrepentant sinners is defeated through divine justice. Evildoers who

refuse reconciliation with God receive a recompense addressed precisely to their offense: a permanent state of separation from the God they reject. The destruction this entails befits their situation, but it is also a destruction the wrongdoer chooses, and his choice is honored by God in the outcome. Thus, religious apologists argue, the fate of the reprobate is a manifestation not only of God's love for them, but also of his justice. Without sin, it is held, much of God's goodness could be displayed in the world, but his justice could not. Not that God does not forgive the unrepentant; he does, but his forgiveness is rejected, and that is what seals the sinner's fate. So while the sinfulness of the unrepentant is not overcome through their accepting salvation, it is defeated through God's justice in honoring their rebellion. Secondly, there is the fact that the sinner's deeds are visited upon others in the world, who must cope not only with the suffering and hardship that results, but also with the very fact of sin: with knowing they have been denied the dignity appropriate to rational beings, and instead made the object of malice. Here too, the theist might argue, it is possible for sin to be defeated, by being pointed out and corrected, and above all by being forgiven. In admonishing and forgiving those who sin toward us, we ally ourselves with God in the struggle against moral evil, by refusing to lapse into vengefulness and self-pity, and instead focusing ourselves, and if possible the sinner as well, on the higher things of God. This also is a good to which sin is integral, rather than constituting a mere causal means, and which makes the world far better than it would be if sin never occurred.

9. Suffering

The position outlined above depends heavily on the idea that in the plan of providence moral evil is defeated, rather than simply being outweighed by some good to which it is a means. The evil of sinful willing is overcome when autonomy that is wrongfully willed is surrendered to God in repentance and conversion, or when the sinner is left to the just deserts of willing to be separated from God. It may fairly be argued, however, that the defeat here pertains primarily to intrinsic moral evil-that is, to sinful willing itself-rather than to the harm caused by it. Both conversion and reprobation may refute the sinful will, but they do not seem to alleviate or otherwise overcome the suffering caused by it. The idea of forgiveness may address the latter to some extent, but it is not entirely clear how. And even so, it may be claimed, much suffering is not, or at least not obviously, the result of wrongdoing. The pains and anxieties of daily living would doubtless be lessened if wrongful willing did not occur, but there would still be danger and disease, accidents, natural disasters, occasional deprivation, the sufferings of old age, and eventual death. Theodicy has to deal with these evils too, and as in the case of sin, it will not do to claim simply that they lead to some greater good that outweighs them. For as in the case of sin, the antitheist could then argue that God could as easily have created the world so that the resultant good would be achieved by means that did not involve suffering, or would simply appear without any means at all. How, then, might the theist respond here? Can the concept of defeasibility be developed so as to cover suffering as well as sin?

One indication that it can is the disdain we would have for a world in which there were no suffering or hardship to be faced, but instead only endless gratification. As usually formulated, the argument from evil is based on what appears to be a false presumption. It imagines that the ideal world for a loving and compassionate God to create must be what John Hick describes as a hedonistic paradise: a place devoted to human enjoyment, in which comfort and convenience are maximized, and pain and deprivation evil

have little or no place.[32] If not banished completely, they must be held to the minimum necessary to guarantee to God's creatures the most pleasant existence possible. Now obviously, that is not the sort of world we have. The amount of suffering is immense-far more, certainly, than it would be if God's aim were to maximize worldly joy. But the appropriate conclusion, the theist may argue, is not that the universe is not the creation of a provident God. For consider how we react to people whose lives have little to distinguish them except that they appear-perhaps deceptively-to be filled with enjoyment. There is a tendency, when we suffer one or another of life's ills, to envy such people: to wish our own existence could be as theirs seems to be, rather than the painful drudgery of the moment. But the truth is that we seldom admire those who appear to have a life of ease, nor are we likely to consider that kind of life very well spent. What we admire are lives of courage and sacrifice: persons who overcome hardship, deprivation, or weakness to achieve some notable success; who stand, perhaps not even successfully, against some great evil; or who relinquish their own happiness to alleviate the suffering of others. How would such lives be possible if natural evil did not exist?

Still less would we respect an entire world devoted to nothing but enjoyment. Imagine a society in which everyone has an electrode implanted in their brain, which, when a current is passed through it, causes intense euphoria, unmatched by any other pleasure. One simply needs to be attached to a power source, and the simple push of a button yields ecstasy. And that is all anyone cares about. Agriculture, commerce, government, and social institutions are organized toward but one goal: to maximize the time each person can spend plugged in, lost in self-stimulation. Individual lives are conducted with the same aim. Work is still necessary, but it is held to a minimum, and contact with fellow human beings has no purpose other than to keep things running smoothly, so that the pleasure of all can be maximized. Now if the antitheist ideal of creation were correct, this type of society ought to represent a high order of human existence-better by far than the world in which we presently find ourselves. In fact, however, it is beneath contempt, a level of existence so low as to be barely human. The enterprises we value most would shrivel to near nothingness in such a world: there would be no art or culture, no important public works, little technology and science-above all, no real human fellowship, no caring, no sacrifice.[33] Perversely enough, in fact, all we have to add to a world like this is war, and we get a situation not at all unlike Hobbes's state of nature, the very antithesis of anything we could value. Clearly, says the theist, we wish more for ourselves than this. A life without challenge is a life without interest.

If theism is correct on these matters, then a God interested in creating the best of worlds cannot have as his top priority the maximization of creaturely pleasure. Rather, a significant part of the enterprise of creation itself ought to be the confrontation and defeat evil-an accomplishment far greater than merely guaranteeing the unperturbed pleasure of all. And if human beings are created in God's image, and called to friendship with him, it is to be expected that they will have an important share in this enterprise. The central role in every human life of the struggle against evil bears this out. The battle is fought within each of us: the foremost challenge we face is that posed by our own sinfulness, which is overcome when we acknowledge that control of our destiny lies finally with God, and give up our false claim to ourselves. But for believers that is by no means the end of the matter. In the wake of repentance there should occur a gradual transformation of the individual, in which the damage wrought by sin is repaired, and the character traits appropriate for friendship with God are nourished. Remorse, anger and bitterness have to be replaced by gratitude, peace and hope; attitudes of failure must be supplanted by a sense of worth;

rationalization has to give way to self-understanding. Above all, the believer has to develop such virtues as humility, patience, courage, and concern for others-to give up selfishness in favor of charity. Hick calls this process "soul-making."[34] In it, the individual is transformed into a being suited for full friendship with God, because through the achievement of virtue he is made over into God's likeness. And much of the process takes place through our learning to deal with natural evil, with pain, sorrow and deprivation, in ourselves and others. By having to cope with our own suffering, we develop peace, humility, perseverance, and trust in God. We also learn sympathy for others who suffer, and by working to improve their lot we establish mercy and justice, in ourselves and in society. Indeed, much of human fellowship and solidarity is founded upon the support and comfort we lend to each other in times of need, and in the common enterprises by which we seek to secure ourselves and one another against the depredations of natural evil.

According to soul-making theodicy, then, by undergoing the soul-making process we develop the traits required for true friendship with God, in the only way that is suitable for humans. Some might object that God could have created virtuous people without their having to go through the troubles that afflict our present existence, that dispositions to good behavior need not be established through suffering.[35] Theists can respond, though, that this objection fails to grasp the nature of soul-making. Human virtues are not mere behavioral dispositions, of the kind found in the inanimate world. There are, no doubt, natural dispositions of patience, courage, kindness and the like-behavioral propensities we are born with, and that vary in strength from one individual to another. But these are not what we have in mind when we speak of moral virtue in the proper sense. True virtue has to be tested and refined. Someone with the virtue of patience must have tasted affliction and disappointment, and seen things through; the courageous person has to have endured danger and risk; the compassionate must have struggled with temptation, sorrow and hardship. The point of such experiences is not merely to strengthen our tendency to act rightly. Virtue is much more than an abiding behavioral propensity. It is a matter of practical wisdom. It requires that we know trial and suffering, and human weakness in the face of them, in the only way they truly can be known: through experience. Suffering, like sin, cannot be understood in the abstract. Only first hand awareness of the world's pain enables us to comprehend fully the options for good and ill in the situations we face, to judge correctly what action is called for, and to perform it with an attitude of humble submission to God, and loving concern for others. In short, true virtue requires knowledge of good and evil-not just as they are manifested in our own struggle with sin, but as they are played out in the travail of the whole world. As we gain this knowledge, we become more suited for God's friendship; indeed, the process of gaining it is the beginning of the friendship. For to address virtuously the hardships of life is not only to improve ourselves; it is to take up the role God has planned for us in the crucial creative enterprise of overcoming natural evil.

10. Suffering and Defeasibility

Soul-making, according to its defenders, is not possible except through the experience of suffering. Because this is so, and because a world in which humans are brought to spiritual maturity through this process is incomparably better than a hedonistic paradise, there is every reason to expect that a perfectly good and loving God would create a world in which there is suffering. Still, the opponent may object that

this answer is at best incomplete. For, he will argue, not all of the suffering of the world enters into soul-making. Consider again the case that was mentioned earlier, of the fawn caught in a forest fire. By and large, the sufferings of lower animals pass without even being remarked by rational beings, and seem to serve no purpose whatever. Even among humans, intense suffering is often followed simply by death, and contributes to no apparent moral development. Or, it may simply be that a person dies suddenly. What end of soul-making does that serve? And in any case, the complaint continues, surely the sheer *amount* of natural evil that exists in the world is incommensurate with the purposes described. In much of human pain and hardship we see little or nothing of the heroic, but only misery. Virtue is a fine thing, but could not God have contrived to purchase it less expensively?

It is difficult for the theist to give a fully satisfying answer at this point. Perhaps the best strategy is to invoke again the concept of defeasibility. For a theodicy that emphasizes that notion, God's full purpose in creating a world that includes pain and affliction includes more than the spiritual maturation of his creatures. Indeed, that effect is only concomitant to the objective that is first and foremost: that evil be defeated. We should expect, therefore, that evil will be found in the world to such a degree and in such variety as best to serve this overarching purpose. If anything less were so-if God, as creator, should shrink from some evil as too diabolical to be overcome, or too bitter to be endured-he would already have fled the field, and the battle would be lost before it had begun. But as to the actual degree and diversity of evil necessary for its defeat, that is very difficult for humans to judge. Our perspective on creation is vastly incomplete, and in any case we have no reason to expect to be privy to God's purposes regarding each detail of creation-especially, as we are about to see, when it comes to suffering. Still, some helpful points are available to the theist.

First, as regards the claim that there is too much evil, or that much of it serves no purpose, it is not obvious that each instance of suffering must be defeated separately from all others, or in the experience of the individual who endures it. The death of another can be as much a reminder to me of my finitude as my own impending death, sudden or lingering. The sufferings of others-even of lower animals-can be effective in arousing me to works of mercy and charity. Indeed, while particular acts of compassion must perforce be directed toward this or that specific instance of suffering, the attitude of compassion is not. Its object is *all* suffering, even suffering of which none but the sufferer will ever be specifically aware; and compassion can be awakened and reinforced by the very realization that much suffering goes unnoticed and unrequited. The same goes for other virtues. We may be prompted to greater efforts to secure justice in the world precisely because we know that much injustice is never redressed; we may seek to accomplish some difficult goal for which others suffered without success, and in whose memory we want to see it attained. One of the benefits of evil that seems unrequited, then, is that it elicits greater effort to deal with those ills we can address, and greater urgency to assist the forgotten. In this sense, at least, there is no such thing as suffering that serves no purpose.

A second important point concerns the *way* in which suffering is defeated in the process of soul-making. It is not simply that it is caught up in a larger process that is very good. That would be compatible with suffering merely being outweighed in the larger process, or with it serving only as a causal means to the achievement of virtue-neither of which is enough to secure its defeat. Rather, suffering is addressed in the process of soul-making and, as it were, refuted. To see how this occurs, we need to understand the nature

of suffering. It is rarely a matter of sheer physical pain, and even when it is, we are often quite prepared to bear it. A distance runner might endure a lot of pain, or a student considerable hardship, in pursuit of their goals. Yet both might maintain an optimistic spirit, and even deny they were really suffering. True suffering occurs when reason is left groping, and hope is called into question. We suffer when pain seems too great to bear and to serve no commensurate end: when the loss of a loved one leaves us dazed, empty and alone; when hateful assaults leave us feeling wounded and violated; when we are distressed by disease, or the decline of old age. Experiences like these tend to crowd all else out of our consciousness. They make us feel dismayed, vulnerable and incomplete; they make our projects appear trivial, our ambitions unreachable. Above all, the experience undermines our confidence in the essential goodness of the world, and our hope that all will be well. Indeed, what suffering does is to raise for each of us, in a way mere argument cannot, the problem of evil-and with it the inevitable temptation to sullenness, self-pity, and enmity toward whatever God there may be. But when we face the problem, and work through suffering, it is defeated. The defeat can manifest itself in many ways. Almost always, our capacity to endure future hardships, and to comfort others in theirs, is increased. We may come to see more accurately our place in the universe, to deepen our devotion to God, to resolve to assist in attacking the sources of suffering, in human wrongdoing as well as the vagaries of nature. In all of this, the theist may argue, natural evil is defeated: the very experience that threatens to overwhelm confidence in the good becomes the indispensable foundation for attitudes of character that serve to diminish its effects and aim at wiping it out. Above all, however, suffering is defeated in the simple fact that it is borne with good grace-that we refuse to submit to the temptation to descend into repining and bitterness, and resolve instead to continue in hope. Much of the heroism of those who suffer lies simply in this: that they strive to go on-something even a fawn can do.

The problem of suffering is not just an intellectual challenge, but a moral one as well. Evil calls for a certain reaction on our part, a reaction of striving, and it is defeated when that reaction occurs. It made be protested that this reaction is available only to believers, and hence that this approach to the problem of evil begs the question. That is not true. The theist is not required to produce a theodicy that will work even if there is no God, only one that will work if there is. And in any case, there seems to be no reason to think nonbelievers are precluded from participating in the defeat of evil. They can and do gain from suffering much as believers do-even to the point of reacting to it with hope in the final goodness of things, and with reverence for any source it may have, whatever its nature. It does have to be admitted, of course, that particular instances of pain and misery do not always meet with a positive response in the sufferer. There may not be sufficient time, and when there is the individual may lack the intelligence or understanding to react in ways we would recognize. Where we do understand the reaction, there may be a mix between good and bad manifestations of character. And of course there are cases where natural evil is borne with anything but good will-with furious anger, or morose despair. Faced with this reality, the theist can only point out what was mentioned above, that the evils which that befall some can and do work to encourage virtue in others, and this can occur by negative example also. If this seems insufficient, the theist can argue that in a world where each an every instance of evil was thoroughly and obviously defeated-so that a perfectly satisfying response to opponents of theism would always be available-an important evil would be missing. It is essential to the challenge of evil that it frequently appears gratuitous, that there seems to be too much of it, that as far as we can see, it often goes unaddressed. Anything less could not bring out the best in us. And then we would be far less suited to

God's friendship, and this world would be far less than is needed for it to be a creation worthy of God: the best of all possible worlds.

11.Conclusion

The hallmark of the traditional free will defense is its fastidiousness: it seeks to distance God from as much evil as possible, so that his goodness will not be tainted by it. The God of the Judeo-Christian tradition is not so fastidious. He is active in all our deeds, turning our hearts where he wills (*Prov*. 21:1), and working in us to will and to do as he pleases. Part of his purpose in this, the tradition holds, is that we be creatures with the moral authenticity that can only come with free will. The inevitable accompaniment, however, is that we sin. God does not will this for its own sake, but if God's providence is complete he does will for us the independence that amounts to our rebellion, for it is indispensable to his purpose. The question the tradition faces is whether God's providence can be complete here, whether he can have full sovereignty if we are truly free. The second focus of concern is the fact of suffering, which also falls under God's will. According to theodicies that emphasize soul-making and defeasibility, this is not because God is malevolent, but so that we can share with him the knowledge that evil is creation's enemy, and partake in the glory of its defeat. The scriptural God evinces no fear that he will be tainted by any of this, nor does he distance himself from evil in any way. On the contrary: even after Adam's sin God remains fully engaged with humankind, sparing no effort to secure our rescue, and treating our suffering with healing concern and compassion. In the Christian tradition, he is even willing to send his Son to bear our sorrows with us, and to be sacrificed so that we may again find acceptance with God in repentance. The fallenness of creation is not, then, an object of heavenly disdain, and for defenders of divine providence it is not cause for philosophical disappointment. Rather, they hold, the task of overcoming evil is central to the creative enterprise. We sin and suffer because God is out to defeat sin and suffering, and to see that all who are ordained to share in the victory do so. The theist is forced to admit, however, that we do not always understand in detail how this occurs. In that respect, at least, any theodicy has to be incomplete.

Bibliography

- Aquinas. *Basic Writings of St. Thomas Aquinas*, 2 vols., ed. A. C. Pegis. New York: Random House, 1944.
- Augustine. *On Grace and Free Will*, in *Basic Writings of Saint Augustine*, vol. I, ed. W. J. Oates. Grand Rapids, Michigan: Baker Book House, 1976.
- Adams, Robert M. "Middle Knowledge and the Problem of Evil," *American Philosophical Quarterly* 14 (1877), 109-14.
- Boethius, *The Theological Tractates and The Consolation of Philosophy*, ed. E. K. Rand. Cambridge, Massachusetts: Harvard University Press, 1973.
- Chisholm, Roderick M. "The Defeat of Good and Evil," *Proceedings of the American Philosophical Association* 42 (1968-69), pp. 21-38.
- Flint, Thomas P. *Divine Providence: The Molinist Account*. Ithaca, New York: Cornell University

Press, 1998.

- Hasker, William. *God, Time, and Knowledge*. Ithaca, New York: Cornell University Press, 1989.
- Helm, Paul. *Eternal God*. York: Oxford University Press, 1988.
- Hick, John. *Evil and the God of Love*. New York: Harper & Row, 1966.
- Hume, David. *Dialogues Concerning Natural Religion*, ed. M. Bell. New York: Penguin Books, 1990.
- -------. *A Treatise of Human Nature*, ed. L. A. Selby-Bigge. Oxford: Oxford University Press, 1888.
- Howard-Snyder, Daniel. "The Argument From Inscrutable Evil," in *The Evidential Argument From Evil*, ed. D. Howard-Snyder. Bloomington, Indiana: Indiana University Press, 1996, pp. 286-310.
- Kvanvig, Jonathan. *The Problem of Hell*. New York: Oxford University Press, 1993.
- -------, and McCann, Hugh J. "The Occasionalist Proselytizer: A Modified Catechism," in *Philosophical Perspectives* 5, ed. J. E. Tomberlin. Atascadero, California: Ridgeview Publishing, 1991, pp. 587-615.
- Leftow, Brian. *Time and Eternity*. Ithaca, New York: Cornell University Press, 1991.
- Mackie, J. L. "Evil and Omnipotence," *Mind* 64 (1955), 200-12.
- McCann, Hugh J. "Divine Sovereignty and the Freedom of the Will," *Faith and Philosophy* 12 (1995), 582-98.
- de Molina, Luis. *On Divine Foreknowledge: Part IV of the Concordia*, tr. Alfred J. Freddoso. New York: Cornell University Press, 1988.
- O'Connor, Timothy. "The Impossibility of Middle Knowledge," *Philosophical Studies* 66 (1992), 139-66.
- Plantinga, Alvin C. God, Freedom, and Evil. York: Harper & Row, 1974.
- Rowe, William L. "The Problem of Evil and Some Varieties of Atheism," *American Philosophical Quarterly* 16 (1979), 335-41.
- Wolterstorff, Nicholas. "God Everlasting," in *Contemporary Philosophy of Religion*, ed. S. M. Cahn and D. Shatz. New York: Oxford University Press, 1982, pp. 77-98.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

evil, problem of

Hugh McCann
h-mccann@tamu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 1, 2001
Content last modified: August 1, 2001

Stanford Encyclopedia of Philosophy
Notes to Divine Providence

Notes

1. Alvin C. Plantinga, *God, Freedom, and Evil* (New York: Harper & Row, 1974), p. 26.

2. William L. Rowe, "The Problem of Evil and Some Varieties of Atheism," *American Philosophical Quarterly* 16 (1979), 335-41.

3. *Ibid*. p. 337.

4. Daniel Howard-Snyder, "The Argument From Inscrutable Evil," in *The Evidential Argument From Evil*, ed. D. Howard-Snyder (Bloomington, Indiana: Indiana University Press, 1996, pp. 286-310.

5. David Hume, *Dialogues Concerning Natural Religion*, ed. M. Bell. (New York: Penguin Books, 1990, p. 111.

6. Plantinga, *God, Freedom, and Evil*, p. 30.

7. J. L. Mackie, "Evil and Omnipotence," *Mind* 64 (1955), 200-12, p. 209.

8. Plantinga, *God, Freedom, and Evil*, pp. 41-42; Thomas P. Flint, *Divine Providence: The Molinist Account* (Ithaca, New York: Cornell University Press, 1998), pp. 84-90.

9. Mackie, "Evil and Omnipotence," pp. 209-10; Flint, *Divine Providence*, pp. 84-85.

10. Boethius, *The Consolation of Philosophy, Bk. V, pr. 6.*

11. Contemporary defenders of the view that God is timelessly eternal include Paul Helm, *Eternal God* (New York: Oxford University Press, 1988); and Brian Leftow, *Time and Eternity* (Ithaca, New York: Cornell University Press, 1991).

12. See especially Nicholas Wolterstorff, "God Everlasting," in *Contemporary Philosophy of Religion*, ed. S. M. Cahn and D. Shatz (New York: Oxford University Press, 1982), pp. 77-98.

13. Luis de Molina, *On Divine Foreknowledge: Part IV of the Concordia*, tr. Alfred J. Freddoso (Ithaca,

New York: Cornell University Press, 1988). The Molinist view receives a thorough and careful defense from Thomas Flint in his *Divine Providence* (op. cit.), on which the present discussion is based.

14. Molina attributes three kinds of knowledge to God. *Natural* knowledge consists of logical and conceptual truths-e.g., that no bachelor is married-which are recognized by God as a matter of his essential nature. *Free* knowledge consists of contingent truths that are settled by God's will as creator-e.g., that there are tigers. Middle knowledge is so named because it falls between these two. The truths of which it is composed are, like those of free knowledge, contingent; but like natural knowledge they are held to be settled independently of God's will.

15. Flint, *Divine Providence*, pp. 37-41.

16. *Ibid*., p. 71.

17. That God could not know counterfactuals of freedom independently of his creative decisions is argued by Timothy O'Connor in "The Impossibility of Middle Knowledge," *Philosophical Studies* 66 (1992), 139-66.

18. For this objection see Robert M. Adams, "Middle Knowledge and the Problem of Evil," *American Philosophical Quarterly* 14 (1877), 109-14; and William Hasker, *God, Time, and Knowledge* (Ithaca, New York: Cornell University Press, 1989), pp. 29-52.

19. For a defense of this view of providence, see Clark Pinnock, Richard Rice, John Sanders. William Hasker, and David Bassinger, eds., *The Openness of God* (Downers Grove, Illinois: InterVarsity Press, 1994).

20. Flint, *Divine Providence*, pp. 100-02.

21. *Ibid*., p. 104.

22. On Grace and Free Will. Basic Writings of Saint Augustine, vol. I, ed. W. J. Oates (Grand Rapids, Michigan: Baker Book House, 1976), p. 750.

23. *Summa Theologica* I-II, Q. 79, Art. 2.

24. *Ibid*., I, Q. 83, Art. 1, *ad* 3.

25. Aquinas, *Summa Contra Gentiles*, Book III, Ch. 88.

26. For a defense see my "Divine Sovereignty and the Freedom of the Will," *Faith and Philosophy* 12

(1995), 582-98.

27. This is a form of the classic objection to libertarian freedom, articulated by David Hume in *A Treatise of Human Nature*, ed. L. A. Selby-Bigge (Oxford: Oxford University Press, 1888), Bk. II, Pt. III, Sec. II.

28. For a critical look at the idea of causal connection, see Jonathan Kvanvig's and my, "The Occasionalist Proselytizer: A Modified Catechism," in *Philosophical Perspectives* 5, ed. J. E. Tomberlin (Atascadero, California: Ridgeview Publishing, 1991), pp. 587-615.

29. The idea that in the first instance, damnation is not a matter of punishment but of God honoring the individual's right to self-determination is developed in Jonathan Kvanvig, *The Problem of Hell*(New York: Oxford University Press, 1993), Ch. 4.

30. It might be thought that Jesus' remark (*Matt*. 26:24) that it would have been better for his betrayer that he not have been born contradicts this. But that cannot be right, for if Judas had not existed, there would have been no "him" to refer to. It is far more plausible to take the remark as saying it would have been better for Judas had he died in his mother's womb.

31. Roderick M. Chisholm, "The Defeat of Good and Evil," *Proceedings of the American Philosophical Association* 42 (1968-69), pp. 21-38.

32. John Hick, *Evil and the God of Love* (New York: Harper & Row, 1966), pp. 292-93.

33. *Ibid*., pp. 359-61.

34. *Ibid*., pp. 289-97.

35. Cf. Mackie, "Evil and Omnipotence," pp. 205-06.

Hugh McCann
hjm7157@unix.tamu.edu

First published: August 1, 2001
Content last modified: August 1, 2001

Jonathan Edwards

Jonathan Edwards (1703-1758) is widely acknowledged to be America's most important and original philosophical theologian. His work as a whole is an expression of two themes -- the absolute sovereignty of God and the beauty of God's holiness. The first is articulated in Edwards' defense of theological determinism, in a doctrine of occasionalism, and in his insistence that physical objects are only collections of sensible "ideas" while finite minds are mere assemblages of "thoughts" or "perceptions." As the only real cause or substance underlying physical and mental phenomena, God is "being in general," the "sum of all being."

Edwards' second theme is articulated in accounts of God's end in creation, and of the nature of true virtue and true beauty. God creates in order to manifest a holiness which consists in a benevolence which alone is truly beautiful. Genuine human virtue is an imitation of divine benevolence and all finite beauty is an image of divine loveliness. True virtue is needed to discern this beauty, however, and to reason rightly about "divine things."

Edwards' projected *History of Redemption* would have drawn these themes together, for it is in his redemptive work in history that God's sovereignty, holiness, and beauty are most clearly exhibited.

1. Life

Edwards was born into a family of prominent Congregational ministers in East Windsor, Connecticut in 1703. In 1716 Edwards enrolled in Yale where he read Newton and Locke, and began "Notes on the Mind" and "Notes on Natural Science." Locke's influence on his epistemology, philosophy of language, and philosophical psychology was profound. Edwards' metaphysics, however, appears more strongly influenced by Malebranche and, to a lesser extent, the Cambridge Platonists, and bears little resemblance to Locke's. After briefly serving congregations in New York and Bolton, Connecticut, Edwards returned to Yale where he completed his Masters of Arts degree and became senior tutor in 1724. In 1725, the church in Northampton chose Edwards to succeed his grandfather, Solomon Stoddard -- the so-called "pope of the Connecticut valley." The most notable events of his tenure were the revivals of 1734 and 1740-41, the latter of which came to be known as the Great Awakening. Edwards' defense of the revivals and criticisms of its excesses culminated in his first major treatise, the *Religious Affections* (1746). Worsening relations with his congregation came to a head in a dispute over qualifications for church membership. Rejecting the less rigorous standards of his grandfather, Edwards insisted on a public profession of saving faith based on the candidate's religious experiences as a qualification not only for Holy Communion but also for church membership. He was dismissed in 1750 by a margin of one vote. After refusing invitations to pulpits in North America and Scotland, Edwards retreated to the Indian mission at Stockbridge where he had charge of two difficult congregations, supervised a boarding school for Indian boys, and completed his last major works -- *Freedom of the Will* (1754), *Original Sin* (1758), *End of Creation*, and *True Virtue* (both published posthumously in 1765). Edwards accepted an appointment as President of the College of New Jersey (now Princeton) in 1757. He died from complications arising from a smallpox inoculation on March 22, 1758, less than five weeks after his inauguration. Edwards' published works were primarily designed to defend the Puritan version of Calvinist orthodoxy and his influence on Congregational and Presbyterian theology was profound. His extensive notebooks reveal an interest in philosophical problems for their own sake, however, and his deployment of philosophical arguments in his private papers and published works are both sophisticated and frequently original.

2. Metaphysics

2.1 Theological Determinism

Edwards believed that indeterminism is incompatible with our dependence on God and hence with his sovereignty. If our responses to God's grace are contra-causally free, then our salvation depends partly on us and God's sovereignty isn't "absolute and universal." *Freedom of the Will* defends theological determinism. Edwards begins by attempting to show that libertarianism is incoherent. For example, he argues that by 'self-determination' the libertarian must mean either that one's actions including one's acts

of willing are preceded by an act of free will or that one's acts of will lack sufficient causes. The first leads to an infinite regress while the second implies that acts of will happen accidentally and hence can't make someone "better or worse, any more than a tree is better than other trees because it oftener happens to be lit upon by a swan or nightingale; or a rock more vicious than other rocks, because rattlesnakes have happened oftener to crawl over it." (*Freedom of the Will*, 1754; Edwards 1957-, vol. 1, 327) On the second alternative, acts of choosing (volitions) are neither chosen by us nor determined by reasons or our character or by other states of the soul. But if they are not, then they aren't truly ours and we cannot be held responsible for them. Edwards also argues that libertarianism is inconsistent with ordinary moral concepts. If, for example, the necessity of sinning wholly excuses, then a bias to sin should partially excuse. But it doesn't; a person who acts from settled habits of maliciousness is deemed "so much the more worthy to be detested and condemned." (*Freedom of the Will*, 1754; Edwards 1957-, vol. 1, 360) Since libertarianism implies that necessity excuses, it is inconsistent with the way we attribute blame.

In Edwards' opinion, libertarianism's specious aura of plausibility is grounded in a systematic confusion of "philosophical" and "vulgar" (ordinary) usage. For example, in ordinary usage something (e.g., remaining seated) is said to be "necessary to us...when we can't help it, let us do what we will." *(Freedom of the Will*, 1754; Edwards 1957-, vol. 1, 150) Causal necessity doesn't entail "vulgar necessity," however. Ingrained habits, deeply felt resentment, and the like may causally necessitate a malicious action. It doesn't follow that the agent wouldn't have refrained from acting maliciously if she had chosen not to act maliciously. Hence, the fact that she was causally unable to act other than she did does not imply that she was unable to do so in the "vulgar" or ordinary sense. Libertarians are therefore mistaken in thinking that because vulgar necessity excuses, so does causal necessity. Again, 'freedom' or 'liberty' in common speech refer only to "that power and opportunity for one to do and conduct as he will, or according to his choice," and contains no reference to the "cause or original" of the act of will. (*Freedom of* the *Will*, 1754; Edwards 1957-vol. 1, 163-64) Hence, that liberty in the ordinary sense is essential to moral agency does not entail that contra-causal freedom is. It is also important to note that action in the ordinary sense is "some motion or exertion of power, that is voluntary, or that is the *effect* of the will...[the term is] most commonly used to signify outward actions." *(Freedom of the Will*, 1754; Edwards 1957-, vol. 1, 346) Improperly extending the term 'action' to movements of the will has led some libertarians to conclude that since external behavior must be preceded by an act of will to be voluntary, and an appropriate object of moral appraisal, so too must acts of will. A consequence is the libertarian's misleading talk of the will's self-determination.

Edwards' principle reasons *for* theological determinism are God's sovereignty, the principle of sufficient reason (which requires that everything that begins to be have a complete cause), the nature of motivation, and God's foreknowledge. The latter two are discussed at length.

The argument from motivation depends upon Edwards' identification of willing or choosing with one's strongest inclination or preference. Since choosing just *is* a prevailing inclination, it is logically impossible to choose in the absence of a prevailing motive. If there is a prevailing motive, however, then the will is necessarily determined by it, for if the will were to choose contrary to a prevailing motive, the agent would have two opposed preponderant inclinations at the same time. All choices, therefore, are necessarily determined.

Edwards' most impressive arguments from divine foreknowledge are based on the impossibility of knowing future contingents and on the necessity of the past. One knows p only if one has evidence for it, and evidence "must be one of...two sorts, either *self-evidence* or *proof*." Propositions about future contingents can't be self-evident, however, because the states of affairs they represent are neither present to the mind nor necessary. But they can't be proved either, for if the state of affairs expressed by the proposition is genuinely contingent, "there is nothing now existent with which the future existence of the contingent event is [necessarily] connected." Future contingents are thus necessarily unknowable. (*Freedom of the Will*, 1754; Edwards 1957-, vol. 1, 259) Since God's knowledge of the future is comprehensive, it follows that no future event (and so no future human action) is genuinely contingent.

The conclusion also follows from the necessity of the past. Suppose I make a decision D at time t. Since God is omniscient, he has always believed that D occurs at t. Since he can't be mistaken, God's believing at some earlier time t-n that D occurs at t entails that D occurs at t. But God's forebelief is past in relation to t and is therefore "now necessary" in the sense that nothing done at t can alter it. What is entailed by a necessary fact is itself necessary, however. Therefore, D could not fail to occur at t. Nor can one evade this conclusion by appealing to God's timelessness as some do. For even if God's 'forebeliefs' are timeless and so don't precede the events they are about, divinely inspired prophecies are not. Yet divinely inspired prophecies, too, are necessarily connected with the human actions they foretell and they are clearly past (and hence necessary) in relation to them.

Necessity is consistent with moral responsibility, however. We are said to be responsible for our actions when we act as we choose and determinism does not deny that our actions often spring from our choices. Nor is necessity incompatible with praise and blame. Even though God and Christ necessarily act for the best, their actions are eminently praiseworthy.

It is worth noting that the aim of Edwards' philosophically sophisticated arguments is theological. He saw that "if modern divines...can maintain their peculiar notion of freedom, consisting in the *self-determining power* of the will, as necessary to moral agency..., then they have an impregnable castle, to which they may repair, and remain invincible, in all the controversies they have with the reformed divines concerning *original sin*, the *sovereignty* of grace, election..., and other principles of like kind." (*Freedom of the Will*, 1754; Edwards 1957-, vol. 3, 376) Edwards recognizes that "modern divines" pretend that doctrines like these undermine "the very foundation of all religion and morality. (*Freedom of the Will*, 1754; Edwards 1957, vol. 1, 422) *Freedom of the Will* concludes by arguing that, on the contrary, they do a much better job of supporting them.

2.2 Occasionalism, Idealism, Mental Phenomenalism, and Views on Identity

Edwards' occasionalism, idealism, and mental phenomenalism provide a philosophical interpretation of God's absolute sovereignty: God is the only real cause and the only true substance.

Edwards implicitly distinguishes between a real or true cause and a cause in the ordinary or "vulgar" sense. The latter is "that, after or upon the existence of which, or the existence of it after such a manner, the existence of another thing follows." ("The Mind," no. 26; Edwards 1957-, vol. 6, 350) Vulgar causes aren't real causes, however. In the first place, so-called second causes are spatially or temporally distinct from their effects, and "no [real] cause can produce effects in a time and place on which itself is not." (*Original Sin*, 1758; Edwards 1957-, vol. 3, 400) In the second, real causes necessitate their effects and second causes do not. "It don't at all necessarily follow," for example, "that because there was...color, or resistance,...or thought, or any other dependent thing at the last moment, that therefore there shall be the like at the next." *(Original Sin*, 1758; Edwards 1957-, vol. 3, 404). Finally, if second causes were real causes they would be sufficient to produce their effects. If they were sufficient, however, then God's activity would be redundant and it is not. Unlike second causes, God's causal activity meets all three conditions. Since God is not in time or space, there is no temporal or spatial separation between his activity and its effects. Since God is essentially omnipotent, his will is necessarily effective; it is logically impossible for him to will s and s not take place. The third condition is also met. Because God is omnipotent he doesn't need the cooperation of other causal powers to produce his effects. And because sovereignty belongs to him alone he doesn't share his causal power with others. God's decrees are thus fully sufficient for their effects. God alone, then, is the only real cause. Vulgar causes (e.g., heating water) are simply the occasions upon which God produces effects (e.g., the water's boiling) according to "methods and laws" which express his customary manner of acting.

In an early paper ("Of Atoms") Edwards pointed out that the concept of a material substance is the concept of something subsisting by itself, standing "underneath," and keeping "up solidity and all other [physical] properties." (Edwards 1957-, vol. 6, 215) He then argued that God alone meets these conditions, and concluded that if the concept of material substance refers to anything, it refers to God's causal activity.

Edwards also thought that "nothing has existence any where else...but either in created or uncreated consciousness." It follows that "the material universe exists only in the mind;" "the existence of all corporeal things is only ideas." ("Of Being," "The Mind," no. 51, and "Miscellanies," no. 179; Edwards 1957, vol. 6, 204, 368, and vol. 13, 327)

Edwards' arguments for idealism are similar to (but apparently uninfluenced by) Berkeley's. One of the best examples occurs in "The Mind," no. 27. Edwards first argues that the idea of a body can be resolved into ideas of color and resistance. Figure, for example, is the termination of color or resistance. Solidity *is* resistance, while motion is "the communication of this resistance from space to space." "Every knowing philosopher" agrees that colors exist only in minds. 'Resistance' refers either to instances in which one body resists another or to a power, namely, a body's disposition to resist other bodies. The first is a mode or property of ideas; it is ideas which are "resisted...move and stop, and rebound." For example, our observation of a billiard ball's ricocheting from the cushion can be resolved into impressions of a particular configuration of color and figure (the billiard ball) moving closer to another (the cushion), touching it, and then moving away from it. The power of resistance is no more than a divine "establishment," namely, "the constant law or method" of "the actual exertion of God's power" producing instances of resistance. So *instances* of resistance are qualities of ideas and the *power* of resistance is a

stable divine intention to act in certain ways. Resistance, therefore, exists only in relation to minds. Since the idea of a body can be reduced to ideas of color and resistance, and color and resistance have only mental existence, "the world is...an ideal one." (Edwards 1957-, vol. 6, 350-51)

Edwards' mental phenomenalism is a natural extension of his occasionalism and views on substance. If God is the only real cause of spatio-temporal phenomena, he is the only real cause of "thoughts" or "perceptions." If a substance is what "subsists by itself," "stands underneath," and "keeps up" a set of properties, then a mental substance can only be what subsists by itself, stands underneath, and keeps up mental properties. It follows that the concept of mental substance either denotes nothing or refers to God's causal activity. "What we call spirit," then, "is nothing but a composition and series of perceptions [mental events]...connected by...laws." ("Notes on Knowledge and Existence"; Edwards 1957-, vol. 6, 398)

Mental and physical substance are thus identical with God's causal production of the mental events constituting minds and the sensible ideas or "sensations" which constitute bodies "according to...methods and laws " which he has freely established. ("The Mind," no. 13; Edwards 1957-, vol. 6, 344) God is thus the only true substance as well as the only true cause.

God's sovereignty also extends to criteria of identity. "Species" (kinds or natures) are the ways we classify things. But our classifications depend on our needs and interests, and the character of the world we live in. Hence, in determining every feature of the spatio-temporal world, God has determined how things will be classified, that is, what counts as a "species" or kind. Since a thing's criteria of identity are determined by its nature or kind, God is their ultimate ground. In short, laws determine kinds and kinds determine criteria of identity. In determining laws God has therefore determined criteria of identity. (One implication is that God can so arrange things that Adam and his posterity count as one thing for purposes of punishment and reward.)

2.3 God as Being in General

God is "being in general." He "is the sum of all being and there is no being without His being. All things are in Him and He in all." ("Miscellanies," no. 880; Edwards 1955, 87) Edwards appears to have borrowed the phrase "being in general" from Malebranche. What does he mean by it?

He does not mean that God is the power of being or being as such as earlier commentators like Clyde Holbrook and Douglas Elwood have suggested. God is neither a power nor a universal but a concrete entity or substance -- a necessarily existing "intelligent willing agent such as our souls, only without our imperfections, and not some inconceivable, unintelligent, necessary agent." ("Miscellanies," no. 383; Edwards 1957-, vol. 13, 452)

True Virtue associates being with capacity or power, and asserts that "degree of existence" is a function of "greater capacity or power," of having "every faculty and every positive quality in an higher degree. An *archangel* must be supposed to have more existence, and to be every way further removed from

nonentity than a *worm* or a *flea*." (*True Virtue*, 1765; Edwards 1957-, vol. 8, 546) Miscellany 94 identifies perfect entity and perfect activity. "God is a pure act...because that which acts perfectly is all act, and nothing but act. There is an image of this in created beings that approach to perfect action." Thus, "the saints of heaven are all transfigured into love. dissolved into joy, become activity itself, changed into mere ecstasy." (Edwards 1957-, vol. 13, 260f.) "An Essay on the Trinity" argues that God’s essence is a love which subsists "in pure act and perfect energy," his holy will or activity. (Edwards 1971, 99-100, 110-11) "Of Being" and "The Mind," no. 45 identify being with consciousness. "Perceiving being only is properly being." (Edwards 1957-, vol. 6, 363) Although Edwards never systematically developed or integrated these scattered observations, their drift is toward the identification of being with mind in act, and of degree of being with degree of mind or consciousness and the comparative perfection of the activity in which it is engaged. God’s consciousness and power are unlimited, and his activity is perfect. His being is therefore unlimited.

Why, though, is God being in general? Because finite beings are absolutely and immediately dependent upon him for both their being and properties. Indeed, as the only true substance and only true cause, created beings are no more than God’s "shadows" or "images." (While "particular minds" deliberate and choose, and so possess a kind of agency, they lack real power and are thus no more than images of divine agency. Because they lack not only power but also consciousness and will, bodies are even further removed from real agency and hence are, as Edwards says, mere shadows of being.) As the only true substance and only true cause, God is the "head" of the system of beings, its "chief part," an absolute sovereign whose power and perfection are so great that "all other beings are as nothing to him, and all other excellency...as nothing and less than nothing,...in comparison of his." (*End of Creation*, 1765; Edwards 1957-, vol. 8, 451) "The whole system of created beings, in comparison of Him, is as the light dust of the balance." ("Miscellanies," no. 1208; Edwards 1955, 142) ‘Being in general,’ then, refers to the system of beings -- primarily to God but to "particular beings," too, in so far as they depend upon and more or less adequately reflect him.

The claim that God is the only real substance, the "proper entity" of things, has led to accusations of pantheism. Students of Edwards have responded by insisting on a distinction in Edwards between God and creatures. The distinction is real but insufficient to refute charges of pantheism. For, historically, pantheisms do not identify the divine with nature as such but, rather, with nature’s substance or essence or inner being or power. Natural phenomena aren’t identical with the divine. They are its modes or properties or parts. Edwards clearly believes that God is the world’s real substance. However, the sense of his assertion is very different from that of the pantheists. In claiming that God is the world’s substance Edwards means that God’s decrees are the only cause of an entity’s being and characteristics. He isn’t a pantheist because the relation between God and the world is construed as a relation between a creative volition and its immediate effects. Edwards’ model is not a whole and its parts, or a substance (a bearer of properties) and its properties, or an essence and its accidents, but agent causality.

2.4 God’s End in Creation

Edwards never doubted that God’s end is himself. Since true virtue consists in benevolence to being and

"complacence" or delight in moral excellence, and since God is the "chief part" of being and the fount of all excellence, a truly virtuous agent "must necessarily have a supreme love to God, both of benevolence and complacence." (*True Virtue,* 1765; Edwards 1957-, vol. 8, 551) It follows that *God's* rectitude and holiness "chiefly consists in a respect or regard to himself, infinitely above his regard to all other beings" and that, as a consequence, his works must be "so wrought as to show this supreme respect to himself." (*End of Creation*, 1765; Edwards 1957-, vol. 8, 422) God's ultimate aim in all his works must therefore be himself. Edwards concludes that he creates the world for his own glory. But Edwards also believed that because the essence of goodness is to communicate good for it own sake, "happiness is the end of the creation." ("Miscellanies," no. 3; Edwards 1957-, vol. 13, 199)

End of Creation reconciles these claims. God's glory is defined as "the emanation and true external expression of God's internal glory and fullness." It includes (1) "the exercise of God's perfections to produce a proper effect," (2) "the manifestation of his internal glory to created understandings," (3) "the communication of the infinite fullness of God to the creature," and (4) "the creature's high esteem of God, love to God, and complacence and joy in God; and the proper exercises and expressions of these." (*End of Creation*, 1765; Edwards 1957-, vol. 8, 527)

There is no ontological distinction between the first and third "parts" of God's glory since the principal effect of God's exercising his perfections is "his fullness communicated." Furthermore, the third part includes the second and fourth. For God's internal fullness or glory is the "fullness of his understanding consisting in his knowledge" of himself "and the fullness of his will consisting in his virtue and happiness." His "external glory...consists in the communication of these," i. e., in bringing it about that "particular minds" know and love God, and delight in him. The four "parts" are thus "one thing, in a variety of views and relations." (*End of Creation*, 1765; Edwards 1957-, vol. 8, 527)

In pursuing his own glory, God thus takes both himself *and* the creature's good as ultimate aims. Happiness consists in the knowledge and love of God, and joy in him. The creature's happiness is an ultimate end because it is *included* in God's ultimate end, namely, the communication of his internal glory "ad extra;" rather than being a means to God's glory, it is part of it.

An apparent consequence is that God *must* create a world to display his glory. *End of Creation* contends both that God's perfections include "a propensity of nature to diffuse of his own fullness" and that it isn't "possible for him to be hindered in the exercise of his goodness and his other perfections in their proper effect." (*End of Creation*, 1765; Edwards 1957-, vol. 8, 447) It follows that God must diffuse his own fullness, i. e., God must create. Edwards also appears committed to the claim that God necessarily creates *this* world (call it w*). God necessarily does what is "fittest and best." It is thus necessarily true that God creates the best possible world. Now God has created w*. Hence, w* is the best possible world. 'Being the best possible world' is an essential property of whatever world has it, however. It is therefore necessarily true that w* is the best possible world. It follows that it is necessarily true that God creates w*.

Whether Edwards was aware of these consequences is uncertain. The two most common objections to them, however, -- that they imply that there isn't any real contingency and that God isn't free -- would

not have troubled him. For Edwards thought that our world displays neither contra-causal freedom nor real indeterminacy. He also believed that moral agency and freedom are compatible with metaphysical necessity. God can only do what is "fittest and best." He is nevertheless free in the sense that he is aware of alternatives (the array of possible worlds), has the ability (i. e., the power and "skill") to actualize any of them, is neither forced, constrained nor influenced by any other being, and does precisely what he wishes. Edwards believes that this is the only kind of freedom that is either relevant to moral agency or worth having.

3. Value Theory

3.1 Ethics

True virtue aims at the good of being in general and therefore also prizes the disposition that promotes it. Truly virtuous people thus love two things -- being and benevolence. They not only value benevolence because it promotes the general good, however; they also "relish" or delight in it for its own sake. Hence, while virtue "most essentially consists in benevolence to being" (*True Virtue*, 1765; Edwards 1957-, vol. 8, 540), in a wider sense it includes not only benevolence but also "complacence" in benevolence's intrinsic excellence or beauty.

God, though, "is infinitely the greatest being," and "infinitely the most beautiful and excellent." True virtue thus principally consists "in a supreme love to God, both of benevolence and complacence." (*True Virtue*, 1765; Edwards 1957-, vol. 8, 550-51) It follows that "a determination of mind to union and benevolence to a *particular person* or *private system* [whether one's self, one's family, one's nation, or even humanity], which is but a small part of the universal system of being...is not of the nature of true virtue" unless it is dependent on or "subordinate to, benevolence to *Being in general*." (*True Virtue*, 1765; Edwards 1957-, vol. 8, 554)

One of the principal concerns of Shaftesbury, Hutcheson, et al., was to refute the contention that action is always motivated by self-love. Edwards' attitude toward these attempts is ambivalent. On the one hand, he denies that the *truly* benevolent are motivated by self-love. On the other, Edwards argues (against, e.g., Hutcheson) that most conscientious and other regarding behavior is, indeed, a form of self-love and that, in any case, acts motivated by rational self-love, conscience, or natural other regarding instincts such as parental affection or pity aren't genuinely virtuous.

Conscience, for instance, is the product of a power of placing ourselves in the situation of others (which is needed for any sort of mutual understanding), a sense of the natural fitness of certain responses (injury and punishment or disapproval, benefit and reward or approval), and self-love. Placing ourselves in the situation of those we have injured, we recognize that being treated in that way would not merely anger us but seem unfitting or undeserved, and that we are therefore inconsistent in approving of our treating others in ways we would not wish to be treated ourselves. The result is a sense of "inconsistence" or "self-opposition" between feelings of approval and disapproval toward the same action. This makes us

"uneasy" since "self-love implies an inclination to feel and act as one with ourselves." (*True Virtue*, 1765; Edwards 1957-, vol. 8, 589)

What, though, about instinctual other regarding impulses such as parental affection, "mutual affection between the sexes" (as distinct from simple sexual attraction), and pity? Edwards is inclined to think that all except pity are forms of self-love. The important point, however, is that even if they aren't, actions motivated by them aren't truly virtuous. To see why consider pity. If *truly* virtuous actions are motivated by benevolence to being in general, then actions motivated by other regarded impulses which are ultimately directed to "some particular persons or private system" aren't truly virtuous. (*True Virtue*, 1765; Edwards 1957-, vol. 8, 601) Now pity is directed to those in extreme distress whose suffering appears undeserved or excessive. Its object is therefore restricted to only part of being in general. Furthermore, since instinctual affections aren't "dependent" on "general benevolence," they are in potential conflict with it. Pity, for example, may motivate a judge to act unjustly.

We should not conclude that pity or other instinctual affections, or even rational self-love, are bad. Since they tend toward "the preservation of mankind and their comfortably subsisting in the world," things would be much worse without them. (*True Virtue*, 1765; Edwards 1957-, vol. 8, 600) Edwards point (like Kant's) is merely that their goodness isn't a truly moral goodness. The implication is nonetheless clear. Natural virtues are either tainted with self-love or fail to extend to being in general. They are therefore counterfeits or simulacra of true virtue. While they prompt us to promote the good of others, and to condemn vice, they fall infinitely "short of the extent of true virtuous benevolence, both in...nature and object." (*True Virtue*, 1765; Edwards 1957-, vol. 8, 609) Edwards concludes that true virtue is a supernatural gift.

3.2 Aesthetics

In Edwards' view, beauty or "excellency" "consists in the similarness of one being to another -- not merely equality and proportion, but any kind of similarness....This is an universal definition of excellency: The consent of being to being...". ("The Mind," no. 1; Edwards 1957-, vol. 6, 336) One who loves others, for instance, or actively desires their welfare, "agrees" with them or "consents" to them. Love's scope can be narrower or wider, however. Agreement or consent is "comprehensive" or "universal" only when directed towards being in general. Only true benevolence, therefore, is truly beautiful.

"Secondary" beauty is a mere "image" or "resemblance" of true beauty. It consists in "symmetry," "harmony," or "proportion," or "as Mr. Hutcheson" says, in "agreement of different things in form, manner, quantity, and visible end or design," i. e., in "regularity." The beauty of well-ordered societies, of "wisdom...consisting in the united tendency of thoughts, ideas, and particular volitions to one general purpose," of the natural fitness of actions and circumstances (having made a promise, for example, and keeping it), "of a building, of a flower, or of the rainbow" are examples. (*True Virtue*, 1765; Edwards 1957-, vol. 8, 561-62)

Since God's benevolence alone is perfect, he is the only thing that is (truly) beautiful without qualification. The fitness of God's dispensations, the harmony of his providential design, and so on, also exhibit the highest degree of secondary beauty. God is thus "infinitely the most beautiful and excellent," the measure of both primary and secondary beauty. Moreover, he is the "foundation and fountain of all beauty." "All the beauty to be found throughout the whole creation is...the reflection of the diffused beams of that being who hath an infinite fullness of brightness and glory." (*True Virtue*, 1765; Edwards 1957-, vol. 8, 550-51) And God's world is indeed *saturated* with beauty -- not only the "harmony of sounds, and the beauties of nature" (which bear the greatest resemblance to true or primary beauty, and to which Edwards was especially sensitive) but also (and primarily) the beauty of the Gospel, of God's providential work in history, and of the saints (the elect). The saints alone, however, can discern true beauty.

4. Epistemology

4.1 A Sense of the Heart

Because their hearts have been regenerated by the indwelling of the Holy Spirit, the saints love being in general. Their love is the basis of a new "spiritual sense" whose "immediate object" is "the beauty of holiness" -- a "new simple idea" that can't "be produced by exalting, varying or compounding" ideas "which they had before," and that truly "represents" divine reality. (*Religious Affections*, 1746 and *True Virtue*, 1765; Edwards 1957-, vol. 2, 205, 260, and vol. 8, 622)

Edwards sometimes identifies true beauty with the pleasure that holy things evoke in people with spiritual "frames" or "tempers" or with the tendency they have to evoke it. At other times he identifies it with the consent of being to being, i.e., with true benevolence or holiness. His view appears to be this. True beauty is identical with benevolence or agreement in somewhat the same way in which water is identical with H_2O or heat with molecular motion. But benevolence is also the objective basis of a dispositional property, namely, a tendency to produce a new simple idea in the savingly converted. This idea is a delight or pleasure in being's consent to being which somehow "represents" or is a "perception" of it. Edwards' account of true beauty thus resembles some accounts of color or extension. Spiritual delight is a simple idea or sensation like our ideas of color or extension. The dispositional property is a power objects have to produce these ideas in our understandings. Benevolence is the objective configuration underlying this power and corresponds to the microstructure of bodies that underlie their tendency to excite ideas of color or extension in minds like ours. Like simple ideas of redness, say, or extension, the new spiritual sensation "represents" or is a "perception" of its object. Just as 'red' or 'extension' can refer to the idea, the power, or the physical configuration that is the basis of the power, so "true beauty" can refer to the spiritual sensation, to the relevant dispositional property, or to benevolence.

Edwards calls the new mode of spiritual understanding a "sense" because the apprehension of spiritual beauty is (1) non-inferential and (2) involuntary, and Edwards, like Hutcheson, associates sensation with immediacy and passivity. (3) It also involves relish or delight, and Edwards followed Locke and

Hutcheson in thinking that, like a feeling of tactual pressure or an impression of redness, being pleased or pained is a kind of sensation or perception. Finally, (4) the new mode of understanding is the source of a new simple idea, and Edwards shared Locke's and Hutcheson's conviction that simple ideas come "from experience."

The saints alone are in an epistemic position to discern truths of religion that are dependent on the "excellency of divine things. For example, a conviction of Christ's sufficiency as a mediator depends on an apprehension of his beauty and excellency. Or, again, one must see the beauty of holiness to appreciate the "hatefulness of sin," and thus be convinced of the justice of divine punishment and our inability to make restitution. The new sense also helps us grasp the truth of the gospel scheme as a whole. A conviction of its truth is an immediate inference from a perception of the beauty or splendor of what it depicts, namely, "God and Jesus Christ...the work of redemption, and the ways and works of God." (*A Divine and Supernatural Light*, 1734; Edwards 1957-, vol. 17, 413)

Edwards' defense of the objectivity of the new spiritual sense has four steps. (1) Benevolence agrees with the nature of things. The world is an interconnected system of minds and ideas in which the only true substance and cause is an infinite and omnipotent love. Human benevolence is thus an appropriate or fitting response to reality. (2) Benevolence is pleased by benevolence; it relishes it, or delights in it, for its own sake. Since benevolence is an appropriate response to reality, so too is benevolence's delight in benevolence. (3) But a delight in benevolence just is a perception of its spiritual beauty. It follows that (4) the redeemed's spiritual perceptions are veridical -- "representations" of something "besides what [is] in [their] own minds." (*True Virtue*, 1765; Edwards 1957-, vol. 8, 622)

4.2 Sanctified Reason

Edwards thinks that reason can prove that God exists, establish many of his attributes, discern our obligations to him, and mount a probable case for the credibility of scripture. But he also believes that grace is needed both to help the natural principles "against those things that tend to stupefy [them] and to hinder [their] free exercise," and to sanctify "the reasoning faculty and" assist "it to see the clear evidence there is of the truth of religion in rational arguments." ("Miscellanies," nos. 626, 628; Edwards 1957-, vol. 18, 155, 156f.)

His view is briefly this. "Actual ideas" are ideas that are lively, clear, and distinct. Thought has a tendency to substitute "signs" (i. e., words or images) for actual ideas. While this tendency is useful and normally quite harmless, it impedes reasoning when "we are at a loss concerning a connection or consequence [between ideas], or have a new inference to draw, or would see the force of some new argument." ("Miscellanies," no. 782; Edwards 1957-, vol. 18, 457) Since accurate reasoning about a subject matter requires attending to actual ideas of it, one can't accurately reason about religion if one lack the relevant actual ideas. To have an actual idea of God, for example, one must have actual ideas of the ideas that compose it. But most of us don't. Those parts of the idea of God that everyone has (ideas of knowledge, power, and justice, for instance) either aren't attended to or, if they are, fail to elicit the appropriate affective reaction. In addition, we can't fully understand ideas of affections which we haven't

experienced and so can’t properly understand God’s benevolence if we aren’t benevolent ourselves. And without the simple idea of true beauty, one can understand neither God’s holiness nor the facts that depend on it.

True benevolence remedies these deficiencies. Because the desires of the truly benevolent are properly ordered, they attend to ideas of religion and are suitably affected by the ideas of God’s attributes and activities that everyone has. (They fear his wrath, for example, and are grateful for his benefits.) Furthermore, they understand God’s benevolence because their own benevolence mirrors it. Finally, the truly benevolent delight in the benevolence in which holiness consists, i. e., they "perceive" or "taste" or "relish" its beauty. Edwards’ claim, then, is that to reason accurately about God one must have an actual idea of him, and to have that one must be truly benevolent. Right reasoning about religious matters requires right affections.

Edwards is an evidentialist. Rational religious beliefs are either properly basic or rest on good evidence. A belief that the gospel scheme exhibits true beauty is an example of the former. But most religious beliefs depend on evidence. Sometimes this evidence includes the idea of true beauty. Even when it does not, however, the right affections are needed to appreciate its force. In either case, only those with properly disposed hearts can read the evidence correctly.

5. The History of Redemption

The trustees of the College of New Jersey invited Edwards to become its third president in 1753. In his reply Edwards gave a number of reasons why he hesitated to accept their offer. Among these was the fear that doing so would interfere with the completion of "a great work" which he had long contemplated "which I call a *History of the Work of Redemption*, a body of divinity in an entire new method, being thrown into the form of an history; considering the affair of Christian theology, as the whole of it, in each part, stands in reference to the great work of redemption by Jesus Christ...." (Edwards 1957-, vol. 16, 727f.) Although Edwards’ project was aborted by his untimely death, it would undoubtedly have been based on a sermon series delivered in 1739 which traces the work of redemption "from the fall of man to the end of the world." The proposed history would have been the culmination of the project begun in *True Virtue* and *End of Creation*. For creation and providence are subordinate to a redemption which is itself subordinate only to God’s glory. The history of redemption is "the *summum* and *ultimum* of all the divine operations and decrees," the manifestation of God’s internal glory in time. (Edwards 1957-, vol. 16, 728) Edwards’ *History* would also have provided a fitting climax to his intellectual career as a whole. For it is in his work of redemption that God’s sovereignty, holiness, and splendor are most fully displayed.

It is doubtful, however, that Edwards’ work would have anticipated modern historiography as some claim. For one thing, the sermon series is essentially a *doctrinal* work. (The section on Christ’s earthly ministry, for instance, is a discussion of the incarnation and atonement, not a life of Jesus.) For another, Edwards’ sources include not only biblical and "profane" histories but biblical prophecy as well. Finally, Edwards doesn’t restrict himself to natural causes in explaining events but also appeals to divine decrees and typology.

Whatever novelty the sermon series possesses is literary and theological. It partly consists in the rich skein of images Edwards uses to connect the events of redemption history. These include the model of a river and its tributaries, a tree and its branches, the construction of a building, the conduct of war, and "a wheel," or "a machine composed of wheels" with its reminiscences both of Ezekiel's vision of the divine throne chariot and of clockwork. ("Images of Divine Things," no. 89; Edwards 1957-, vol. 11, 86) It also consists in Edwards' extension of typology, the practice of interpreting things, persons, or events (the "type") as symbols or prefigurations of future realities (the "antitype"). Protestant divines had tended to restrict typology to figures, actions, and objects in the Old Testament which in their view shadowed forth Christ as their antitype. Edwards interprets the New Testament typologically as well, arguing that relevant passages prefigure events in the church's later history. Most radically, Edwards construes nature typologically. (Whether this constitutes a step towards Emerson and Thoreau, as some claim, is a moot point.) Finally, Edwards' emphasis on the objective side of God's act of redemption is comparatively rare in a Puritanism which tended to stress the redemption's application to individual souls. (The subjective side is extensively treated in a number of works of the 1730's and 1740's, however, the most important of which is *Religious Affections*.)

Bibliography

References

- Edwards, J., 1955, *The Philosophy of Jonathan Edwards from His Private Notebooks*,
- Harvey G. Townsend (ed.), Eugene, Oregon: University of Oregon Monographs. (A selection from Edwards' philosophical notebooks.)
- ------, 1971, *Treatise on Grace and other posthumously published writings*, Paul Helm (ed.), Cambridge: James Clarke & Co. (Contains papers not yet included in the collected works.)
- ------, 1829-30, *The Works of President Edwards*, 10 vols., Sereno E. Dwight (ed.), New York: G. & C. & H. Carvill. (A widely available edition of Edwards' work.)
- ------, 1968, *The Works of President Edwards*, Edward Williams and Edward Parsons (eds.), New York: B. Franklin. (A reprint of the 1817 ed. [8 vols.] and the two supplementary volumes published in 1847.)
- ------, 1957-, *The Works of Jonathan Edwards*, gen. eds. Perry Miller (vols. 1-2), John E. Smith (vols. 3-9), and Harry S. Stout (vol. 10-). New Haven: Yale University Press. (Will supersede earlier editions when completed. The extensive introductions are especially helpful.)

Further Reading

- Brown, Robert E., 1999, 'Edwards, Locke, and the Bible', *The Journal of Religion* 79: 361-84
- Cooey, Paula M., 1989, 'Eros and Intimacy in Edwards', *The Journal of Religion* 69: 484-501
- Daniel, Stephen H., 1994, *The Philosophy of Jonathan Edwards: A Study in Divine Semiotics*, Bloomington: Indiana University Press

- Delattre, Roland A., 1968, *Beauty and Sensibility in the Thought of Jonathan Edwards*, New Haven:Yale University Press
- Elwood, Douglas J., 1960, *The Philosophical Theology of Jonathan Edwards*, New York: Columbia University Press
- Fiering, Norman, 1981, *Jonathan Edwards' Moral Thought and its British Context*, Chapel Hill: University of North Carolina Press
- Helm, Paul, 1969, 'John Locke and Jonathan Edwards: A Reconsideration', *Journal of the History of Philosophy* 8: 51-61
- ------, 1979,'Jonathan Edwards and the Doctrine of Temporal Parts', *Archiv fur Geschichte der Philosophie* 61: 37-51
- Holbrook, Clyde A., 1973, *The Ethics of Jonathan Edwards: Morality and Aesthetics*., Ann Arbor: University of Michigan Press
- Jenson, Robert W., 1988, *America's Theologian: A Recommendation of Jonathan Edwards*, New York: Oxford University Press
- Lee, Sang Hyun, 1988, *The Philosophical Theology of Jonathan Edwards*, Princeton: Princeton University Press
- Lewis, Paul, 1994, '"The Springs of Motion": Jonathan Edwards on Emotions', *Journal of Religious Ethics* 22: 275-97
- McClymond, Michael, J., 1998, *Encounters with God: An Approach to the Theology of Jonathan Edwards*, Oxford: Oxford University Press
- McDermott, Gerald R., 2000, *Jonathan Edwards Confronts the Gods: Christian Theology, Enlightenment Religion, and Non-Christian Faiths*, Oxford: Oxford University Press
- Miller, Perry, 1949, *Jonathan Edwards*, New York: W. Sloane Associates
- Post, Stephen, 1986, 'Disinterested Benevolence: An American Debate Over the Nature of Christian Love', *Journal of Religious Ethics* 14: 356-68
- Schmidt, Lawrence K., 1988, 'Jonathan Edwards' Idealistic Argument from Resistance', *Southwest Philosophy Review* 4: 39-47
- Smith, John E., 1992, *Jonathan Edwards: Puritan, Preacher, Philosopher*, London: Geoffrey Chapman and Notre Dame, Ind.: University of Notre Dame Press
- Wainwright, William J., 1980, 'Jonathan Edwards and the Language of God', *The Journal of the American Academy of Religion* 48: 520-30
- ------, 1982, 'Jonathan Edwards, Atoms, and Immaterialism', *Idealistic Studies*: 79-89
- ------, 1988, 'Original Sin', in *Philosophy and the Christian Faith*, Thomas V. Morris (ed.), Notre Dame, Ind.: University of Notre Dame Press
- ------, 1995, *Reason and the Heart*, chapter 1, Ithaca, N. Y.: Cornell University Press
- ------, 1996, 'Jonathan Edwards, William Rowe, and the Necessity of Creation', in *Faith, Freedom, and Responsibility*, Jeff Jordan and Daniel Howard-Snyder (eds.), Lanham, Md.: Rowman and Littlefield
- ------, 2001, 'Jonathan Edwards and the Hiddenness of God', in *New Essays on the Divine Hiddenness*, Paul Moser and Daniel Howard-Snyder (eds.), Cambridge: Cambridge University Press
- ------, 2001, 'Theological Determinism and the Problem of Evil: Are Arminians Any Better Off?', *International Journal for Philosophy of Religion* 50: 81-96

- Winslow, Ola Elizabeth, 1940, *Jonathan Edwards 1703-1758: A Biography*, New York: Macmillan

Other Internet Resources

- Jonathan Edwards on the Web (Don Westblade, Department of Philosophy and Religion, Hillsdale College)
- The Works of Jonathan Edwards (Yale Divinity School)
- Jonathan Edwards Website (Mark Trigsted)
  This site has more complete bibliographical entries than other sites, and and its major virtue is that many writings by Edwards can be downloaded. The religious opinions expressed at the site, however, are not of a purely academic character.
- Stanford Encyclopedia of Philosophy: Francis Hutcheson in Scottish Philosophy in the Eighteenth Century by Alexander Broadie (University of Glasgow)

Related Entries

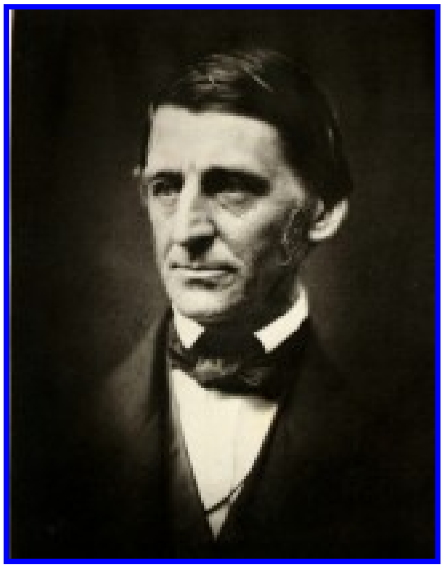
Berkeley, George | Cambridge Platonists | free will | Locke, John | Malebranche, Nicolas | occasionalism | omniscience | religious experience

William Wainwright
University of Wisconsin/Milwaukee
wjwain@uwm.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 15, 2002
Content last modified: January 15, 2002



(From Amos Bronson Alcott, *Ralph Waldo Emerson: An Estimate of His Character and Genius: In Prose and in Verse*, Boston: A. Williams and Co., 1882)

Ralph Waldo Emerson

An American essayist, poet, and popular philosopher, Ralph Waldo Emerson (1803-82) began his career as a Unitarian minister in Boston, but achieved worldwide fame as a lecturer and the author of such essays as "Self-Reliance," "History," "The Over-Soul," and "Fate." Drawing on English and German Romanticism, Neoplatonism, Kantianism, and Hinduism, Emerson developed a metaphysics of process, an epistemology of moods, and an "existentialist" ethics of self-improvement. He influenced generations of Americans, from his friend Henry David Thoreau to John Dewey, and in Europe, Friedrich Nietzsche, who takes up such Emersonian themes as power, fate, the uses of poetry and history, and the critique of Christianity.

1. Chronology of Emerson’s Life

1803, Born in Boston to William and Ruth Haskins Emerson.

1811, Father dies, probably of tuberculosis.

1812, Enters Boston Public Latin School

1817, Begins study at Harvard College: Greek, Latin, History, Rhetoric.

1820, Starts first journal, entitled "The Wide World."

1821, Graduates from Harvard and begins teaching at his brother William’s school for young ladies in Boston.

1825, Enters Harvard Divinity School.

1829, Marries Ellen Tucker and is ordained minister at Boston’s Second Church.

1831, Ellen Tucker Emerson dies, at age 19.

1832, Resigns position as minister and sails for Europe.

1833, Meets Wordsworth, Coleridge, J. S. Mill, and Thomas Carlyle. Returns to Boston in November, where he begins a career as a lecturer.

1834, Receives first half of a substantial inheritance from Ellen’s estate (second half comes in 1837).

1835, Marries Lidian Jackson.

1836, Publishes first book, *Nature*.

1838, Delivers the "Divinity School Address." Protests relocation of the Cherokees in letter to President Van Buren.

1841, *Essays* published (contains "Self-Reliance," "The Over-Soul," "Circles," "History").

1842, Son Waldo dies of scarlet fever at the age of 5.

1844, *Essays, Second Series* published (contains "The Poet," "Experience," "Nominalist and Realist").

1847-8, Lectures in England.

1850, Publishes *Representative Men* (essays on Plato, Swedenborg, Montaigne, Goethe, Napoleon).

1851-60, Speaks against Fugitive Slave Law and in support of anti-slavery candidates in Concord, Boston, New York, Philadelphia.

1856, Publishes *English Traits*.

1860, Publishes *The Conduct of Life* (contains "Culture" and "Fate").

1867, Lectures in nine western states.

1870, Publishes *Society and Solitude.* Presents sixteen lectures in Harvard’s Philosophy Department.

1872-3, After a period of failing health, travels to Europe, Egypt.

1875, Journal entries cease.

1882, Dies in Concord.

2. Major Themes in Emerson’s Philosophy

2.1 Education

In "The American Scholar," delivered as the Phi Beta Kappa Address in 1837, Emerson maintains that the scholar is educated by nature, books, and action. Nature is the first in time (since it is always there) and the first in importance of the three. Nature’s variety conceals underlying laws that are at the same time laws of the human mind: "the ancient precept, ‘Know thyself,’ and the modern precept, ‘Study nature,’ become at last one maxim" (87). Books, the second component of the scholar’s education, offer us the influence of the past. Yet much of what passes for education is mere idolization of books -- transferring the "sacredness which applies to the act of creation...to the record." The proper relation to books is not that of the "bookworm" or "bibliomaniac," but that of the "creative" reader who uses books as a stimulus to attain "his own sight of principles." Used well, books "inspire…the active soul" (88). Great books are mere records of such inspiration, and their value derives only, Emerson holds, from their role in inspiring or recording such states of the soul. The "end" Emerson finds in nature is not a vast collection of books, but, as he puts it in "The Poet," "the production of new individuals,… or the passage of the soul into higher forms" (CW3:14)

The third component of the scholar’s education is action. Without it, thought never "ripens into truth." Action is the process whereby what is not fully formed passes into expressive consciousness (91-2). Action is also the scholar’s "dictionary," the source for what she has to say. The true scholar speaks from experience, not in imitation of others; her words, as Emerson puts it, are "are loaded with life…." (Z: 92). The scholar’s education in original experience and self-expression is appropriate, according to Emerson, not only for a small class of people, but for everyone. Its goal is the creation of a democratic nation. Only when we learn to "walk on our own feet" and to "speak our own minds," he holds, will a nation "for the first time exist" (Z: 104-5).

Emerson returned to the topic of education late in his career in "Education," an address he gave in various versions at graduation exercises in the 1860’s. Self-reliance appears in the essay in his discussion of respect. The "secret of Education," he states, "lies in respecting the pupil." It is not for the teacher to choose what the pupil will know and do, but for the pupil to discover "his own secret." The teacher must therefore "wait and see the new product of Nature" (L: 143), guiding and disciplining when appropriate-not with the aim of encouraging repetition or imitation, but with that of finding the new power that is each child’s gift to the world. The aim of education is to "keep" the child’s "nature and arm it with knowledge in the very direction in which it points" (L: 144). This aim is sacrificed in mass education, Emerson warns. Instead of educating "masses," we must educate "reverently, one by one," with the attitude that "the whole world is needed for the tuition of each pupil" (L: 154).

2.2 Process

Emerson is in many ways a process philosopher, for whom the universe is fundamentally in flux and "permanence is but a word of degrees" (CW 2: 179). Even as he talks of "Being," Emerson represents it not as a stable "wall" but as a series of "interminable oceans" (CW3: 42). This metaphysical position has epistemological correlates: that there is no final explanation of any fact, and that each law will be incorporated in "some more general law presently to disclose itself" (CW2: 181). Process is the basis for the succession of moods Emerson describes in "Experience," (CW3: 30), and for the emphasis on the present throughout his philosophy.

Some of Emerson's most striking ideas about morality and truth follow from his process metaphysics: that no virtues are final or eternal, all being "initial," (CW2: 187); that truth is a matter of glimpses, not steady views. We have a choice, Emerson writes in "Intellect," "between truth and repose," but we cannot have both (CW2: 202). Fresh truth, like the thoughts of genius, comes always as a surprise, as what Emerson calls "the newness" (CW3: 40). He therefore looks for a "certain brief experience, which surprise[s] me in the highway or in the market, in some place, at some time..." (Z: 253). This is an experience that cannot be repeated by simply returning to a place or to an object such as a painting. A great disappointment of life, Emerson finds, is that one can only "see" certain pictures once, and that the stories and people who fill a day or an hour with pleasure and insight are not able to repeat the performance.

Emerson's basic view of religion also coheres with his emphasis on process, for he holds that one finds God only in the present: "God is, not was" (Z: 123). In contrast, what Emerson calls "historical Christianity" (114) proceeds "as if God were dead" (Z: 116). Even history, which seems obviously about the past, has its true use, Emerson holds, as the servant of the present: "The student is to read history actively and not passively; to esteem his own life the text, and books the commentary" (CW2:5).

2.3 Morality

Emerson's views about morality are intertwined with his metaphysics of process, and with his perfectionism, his idea that life has the goal of passing into "higher forms" (CW3:14). The goal remains, but the forms of human life, including the virtues, are all "initial" (CW2: 187). The word "initial" suggests the verb "initiate," and one interpretation of Emerson's claim that "all virtues are initial" is that virtues initiate historically developing forms of life, such as those of the Roman nobility or the Confucian *junxi*. Emerson does have a sense of morality as developing historically, but in the context in "Circles" where his statement appears he presses a more radical and skeptical position: that our virtues often must be abandoned rather than developed. "The terror of reform," he writes, "is the discovery that we must cast away our virtues, or what we have always esteemed such, into the same pit that has consumed our grosser vices" (CW2: 187). The qualifying phrase "or what we have always esteemed such" means that Emerson does not embrace an easy relativism, according to which what is taken to be a virtue at any time must actually be a virtue. Yet he does cast a pall of suspicion over all established modes of thinking and acting. The proper standpoint from which to survey the virtues is the ‘new moment‘ "the moment of truth

rather than repose" (CW2:202), in which what once seemed important may appear "trivial" or "vain." From this perspective (or more properly the developing set of such perspectives) the virtues do not disappear, but they may be fundamentally altered and rearranged.

Although Emerson is thus in no position to set forth a system of morality, he nevertheless delineates throughout his work a set of virtues and heroes, and a corresponding set of vices and villains. In "Circles" the vices are "forms of old age," and the hero the "receptive, aspiring" youth (CW2:189). In the "Divinity School Address," the villain is the "spectral" preacher whose sermons offer no hint that he has ever lived. "Self Reliance" condemns virtues that are really "penances" (CW2: 31), and the philanthropy of abolitionists who display an idealized "love" for those far away, but are full of hatred for those close by (CW2: 30).

Conformity is the chief Emersonian vice, the opposite or "aversion" of the virtue of "self-reliance." We conform when we pay unearned respect to clothing and other symbols of status, when we show "the foolish face of praise" or the "forced smile which we put on in company where we do not feel at ease in answer to conversation which does not interest us" (CW2: 32). Emerson criticizes our conformity even to our own past actions-when they no longer fit the needs or aspirations of the present. This is the context in which he states that "a foolish consistency is the hobgoblin of little minds, adored by little statesmen, philosophers and divines" (CW2: 33). There is wise and there is foolish consistency, and it is foolish to be consistent if that interferes with the "main enterprise of the world for splendor, for extent, … the upbuilding of a man" (99).

If Emerson criticizes much of human life, he nevertheless devotes most of his attention to the virtues. Chief among these is what he calls "self-reliance." The phrase connotes originality and spontaneity, and is memorably represented in the image of a group of nonchalant boys, "sure of a dinner…who would disdain as much as a lord to do or say aught to conciliate one...." The boys sit in judgment on the world and the people in it, offering a free, "irresponsible" condemnation of those they see as "silly" or "troublesome," and praise for those they find "interesting" or "eloquent." (CW2: 29). The figure of the boys illustrates Emerson’s characteristic combination of the romantic (in the glorification of children) and the classical (in the idea of a hierarchy in which the boys occupy the place of lords or nobles).

Speaking of "self-reliance," Emerson nevertheless warns-undermining his own previous statements-can be a "poor external way of speaking" (CW 2:40). For it can be taken to mean that there is a self already formed on which we may rely. The "self" on which we are to "rely" is, in contrast, the original self that we are in the process of creating. Such a self, to use a phrase from Nietzsche’s Ecce Homo, "becomes what it is."

For Emerson, the best human relationships require the confident and independent nature of the self-reliant. Emerson’s ideal society is a confrontation of powerful, independent "gods, talking from peak to peak all round Olympus." There will be a proper distance between these gods, who, Emerson advises, "should meet each morning, as from foreign countries, and spending the day together should depart, as into foreign countries" (CW 3:81). Even "lovers," he advises, "should guard their strangeness" (CW3:

82). Emerson portrays himself as preserving such distance in the cool confession with which he closes "Nominalist and Realist," the last of the *Essays, Second Series*:

> I talked yesterday with a pair of philosophers: I endeavored to show my good men that I liked everything by turns and nothing long.... Could they but once understand, that I loved to know that they existed, and heartily wished them Godspeed, yet, out of my poverty of life and thought, had no word or welcome for them when they came to see me, and could well consent to their living in Oregon, for any claim I felt on them, it would be a great satisfaction (CW 3:145).

The self-reliant person will "publish" her results, but she must first learn to detect that spark of originality or genius that is her particular gift to the world. It is not a gift that is available on demand, however, and a major task of life is to meld genius with its expression. "The man," Emerson states "is only half himself, the other half is his expression" (CW 3:4). There are young people of genius, Emerson laments in "Experience," who promise "a new world" but never deliver: they fail to find the focus for their genius "within the actual horizon of human life" (CW 3:31). Although Emerson emphasizes our independence and even distance from one another, then, the payoff for self-reliance is public and social. The scholar finds that the most private and secret of his thoughts turn out to be "the most acceptable, most public, and universally true" (Z: 97). And the great "representative men" Emerson identifies are marked by their influence on the world. Their names-Plato, Moses, Jesus, Luther, Copernicus, even Napoleon-are "ploughed into the history of the world" (Z: 112).

Although self-reliance is central, it is not the only Emersonian virtue. Emerson also praises a kind of trust, and the practice of a "wise skepticism." There are times, he holds, when we must let go and trust to the nature of the universe: "As the traveler who has lost his way, throws his reins on his horse's neck, and trusts to the instinct of the animal to find his road, so must we do with the divine animal who carries us through this world" (3:15). But the world of flux and conflicting evidence also requires a kind of epistemological and practical flexibility that Emerson calls "wise skepticism" (318). His representative skeptic of this sort is Michel de Montaigne, who as portrayed in *Representative Men* is no unbeliever, but a man with a strong sense of self, rooted in the earth and common life, whose quest is for knowledge. He wants "a near view of the best game and the chief players; what is best in the planet; art and nature, places and events; but mainly men" (CW4: 91). Yet he knows that life is perilous and uncertain, "a storm of many elements," the navigation through which requires a flexible ship, "fit to the form of man." (CW4: 91).

2.4 Christianity

The son of a Unitarian minister, Emerson attended Harvard Divinity School and was employed as a minister for almost three years. Yet he offers a deeply felt and deeply reaching critique of Christianity in the "Divinity School Address," flowing from a line of argument he establishes in "The American Scholar." If the one thing in the world of value is the active soul, then religious institutions, no less than educational institutions, must be judged by that standard. Emerson finds that contemporary Christianity

deadens rather than activates the spirit. It is an "Eastern monarchy of a Christianity" in which Jesus, originally the "friend of man," is made the enemy and oppressor of man. A Christianity true to the life and teachings of Jesus should inspire "the religious sentiment"-a joyous seeing that is more likely to be found in "the pastures," or "a boat in the pond" than in a church. Although Emerson thinks it is a calamity for a nation to lose the capacity to worship (Z: 122) he finds it strange that, given the "famine of our churches" (Z: 117) anyone should attend them. He therefore calls on the Divinity School graduates to breathe new life into the old forms of their religion, to be friends and exemplars to their parishioners, and to remember "that all men have sublime thoughts; that all men value the few real hours of life; they love to be heard; they love to be caught up into the vision of principles" (Z: 124).

2.5 Power

Power is a theme in Emerson’s early writing, but it becomes especially prominent in such middle- and late-career essays as "Experience," "Montaigne; or the Skeptic" "Napoleon," and "Power." Power is related to action in "The American Scholar," where Emerson holds that a "true scholar grudges every opportunity of action passed by, as a loss of power" (Z: 92). It is also a subject of "Self-Reliance," where Emerson writes of each person that "the power which resides in him is new in nature" (CW2:28). In "Experience" Emerson speaks of a life which "is not intellectual or critical, but sturdy" (CW3:294); and in "Power" he celebrates the "bruisers" (P: 372) of the world who express themselves rudely and get their way. The power in which Emerson is interested, however, is more artistic and intellectual than political or military. In a characteristic passage from "Power," he states:

> In history the great moment is, when the savage is just ceasing to be a savage, with all his hairy Pelasgic strength directed on his opening sense of beauty:-and you have Pericles and Phidias,-not yet passed over into the Corinthian civility. Everything good in nature and the world is in that moment of transition, when the swarthy juices still flow plentifully from nature, but their astringency or acridity is got out by ethics and humanity" (P: 375).

Power is all around us, but it cannot always be controlled. It is like "a bird which alights nowhere," hopping "perpetually from bough to bough" (CW3:34). Moreover, we often cannot tell at the time when we exercise our power that we are doing so: happily we sometimes find that much is accomplished in "times when we thought ourselves indolent" (CW3:28).

2.6 Unity and Moods

At some point in many of his essays and addresses, Emerson enunciates, or at least refers to, a great vision of unity. He speaks in "The American Scholar" of an "original unit" or "fountain of power" (Z: 84), of which each of us is a part. He writes in "The Divinity School Address" that each of us is "an inlet into the deeps of Reason." And in "Self-Reliance," the essay that more than any other celebrates individuality, he writes of "the resolution of all into the ever-blessed ONE" (CW 2:40). "The Oversoul" is Emerson’s most sustained discussion of "the ONE," but he does not, even there, shy away from the seeming conflict between the reality of process and the reality of an ultimate metaphysical unity. How

can the vision of succession and the vision of unity be reconciled?

Emerson never comes to a clear or final answer. One solution he both suggests and rejects is an unambiguous idealism, according to which a nontemporal "One" or "Oversoul" is the only reality, and all else is illusion. He suggests this, for example, in the many places where he speaks of waking up out of our dreams or nightmares. But he then portrays that to which we awake not simply as an unchanging "ONE," but as a process or succession: a "growth" or "movement of the soul" (CW2: 189); or a "new yet unapproachable America" (CW3: 259).

Emerson undercuts his visions of unity (as of everything else) through what Stanley Cavell calls his "epistemology of moods." According to this epistemology, most fully developed in "Experience" but present in all of Emerson's writing, we never apprehend anything "straight" or in-itself, but only under an aspect or mood. Emerson writes that life is "a train of moods like a string of beads," through which we see only what lies in each bead's focus (CW3: 30). The beads include our temperaments, our changing moods, and the "Lords of Life" which govern all human experience. The Lords include "Succession," "Surface, "Dream," "Reality," and "Surprise." Are the great visions of unity, then, simply aspects under which we view the world?

Emerson's most direct attempt to reconcile succession and unity, or the one and the many, occurs in the last essay in the *Essays, Second Series*, entitled "Nominalist and Realist." There he speaks of the universe as an "old Two-face…of which any proposition may be affirmed or denied" (CW3: 144). As in "Experience," Emerson leaves us with the whirling succession of moods. "I am always insincere," he skeptically concludes, "as always knowing there are other moods" (CW3: 145). But Emerson enacts as well as describes the succession of moods, and he ends "Nominalist and Realist" with the "feeling that all is yet unsaid," and with at least the idea of some universal truth (CW3: 363).

3. Some Questions about Emerson

3.1 Consistency

Emerson routinely invites charges of inconsistency. He says the world is fundamentally a process and fundamentally a unity; that it resists the imposition of our will and that it flows with the power of our imagination; that travel is good for us, since it adds to our experience, and that it does us no good, since we wake up in the new place only to find the same " sad self" we thought we had left behind (CW2: 46).

Emerson's "epistemology of moods" is an attempt to construct a framework for encompassing what might otherwise seem contradictory outlooks, viewpoints, or doctrines. Emerson really means to "accept," as he puts it, "the clangor and jangle of contrary tendencies" (CW3: 36). He means to be irresponsible to all that holds him back from his self-development. That is why, at the end of "Circles," he writes that he is "only an experimenter…with no Past at my back" (CW2: 188). In the world of flux that he depicts in that essay, there is nothing stable to be responsible to: "every moment is new; the past

is always swallowed and forgotten, the coming only is sacred" (CW2: 189).

Despite this claim, there is considerable consistency in Emerson's essays and among his ideas. To take just one example, the idea of the "active soul"-mentioned as the "one thing in the world, of value" in "The American Scholar-is a presupposition of Emerson's attack on "the famine of the churches" (for not feeding or activating the souls of those who attend them); it is an element in his understanding of a poem as "a thought so passionate and alive, that, like the spirit of a plant or an animal, it has an architecture of its own …." (CW3: 6); and, of course, it is at the center of Emerson's idea of self-reliance. There are in fact multiple paths of coherence through Emerson's philosophy, guided by ideas discussed previously: process, education, self-reliance, and the present.

3.2 Early and Late Emerson

It is hard for an attentive reader not to feel that there are important differences between early and late Emerson: for example, between the buoyant *Nature* (1836) and the weary ending of "Experience" (1844); between the expansive author of "Self-Reliance" (1841) and the burdened writer of "Fate" (1860). Emerson himself seems to advert to such differences when he writes in "Fate": "Once we thought positive power was all. Now we learn that negative power, or circumstance, is half" (Z: 369). Is "Fate" the record of a lesson Emerson had not absorbed in his early writing, concerning the multiple ways in which circumstances over which we have no control -- plagues, hurricanes, temperament, sexuality, old age-constrain self-reliance or self-development?

"Experience" is a key transitional essay. "Where do we find ourselves?" is the question with which it begins. The answer is not a happy one, for Emerson finds that we occupy a place of dislocation and obscurity, where "sleep lingers all our lifetime about our eyes, as night hovers all day in the boughs of the fir-tree" (CW3: 27). An event hovering over the essay, but not disclosed until its third paragraph, is the death of his five-year old son Waldo. Emerson finds in this episode and his reaction to it an example of an "unhandsome" general character of existence-it is forever slipping away from us, like his little boy.

"Experience" presents many moods. It has its moments of illumination, and its considered judgment that there is an "Ideal journeying always with us, the heaven without rent or seam" (CW3: 41). It offers wise counsel about "skating over the surfaces of life" and confining our existence to the "mid-world." But even its upbeat ending takes place in a setting of substantial "defeat." "Up again, old heart!" a somewhat battered voice states in the last sentence of the essay. Yet the essay ends with an assertion that in its great hope and underlying confidence chimes with some of the more expansive passages in Emerson's writing. The "true romance which the world exists to realize," he states, "will be the transformation of genius into practical power" (CW3: 49).

Despite important differences in tone and emphasis, Emerson's assessment of our condition remains much the same throughout his writing. There are no more dire indictments of ordinary human life than in the early work, "The American Scholar," where Emerson states that "Men in history, men in the world of today, are bugs, are spawn, and are called 'the mass' and 'the herd.' In a century, in a millennium, one or

two men; that is to say, one or two approximations to the right state of every man" (99). Conversely, there is no more idealistic statement in his early work than the statement in "Fate" that "[t]hought dissolves the material universe by carrying the mind up into a sphere where all is plastic" (377). All in all, the earlier work expresses a sunnier hope for human possibilities, the sense that Emerson and his contemporaries were poised for a great step forward and upward; and the later work, still hopeful and assured, operates under a weight or burden, a stronger sense of the dumb resistance of the world.

3.3 Sources and Influence

Emerson read widely, and gave credit in his essays to the scores of writers from whom he learned. He kept lists of literary, philosophical, and religious thinkers in his journals and worked at categorizing them.

Among the most important writers for the shape of Emerson’s philosophy are Plato and the Neoplatonist line extending through Plotinus, Proclus, Iamblichus, and the Cambridge Platonists. Equally important are writers in the Kantian and Romantic traditions (which Emerson probably learned most about from Coleridge’s *Biographia Literaria*). Emerson read avidly in Indian, especially Hindu, philosophy, and in Confucianism. There are also multiple empiricist, or experience-based influences, flowing from Berkeley, Wordsworth and other English Romantics, Newton’s physics, and the new sciences of geology and comparative anatomy. Other writers whom Emerson often mentions are Anaxagoras, St. Augustine, Francis Bacon, Jacob Behmen, Cicero, Goethe, Heraclitus, Lucretius, Mencius, Pythagoras, Schiller, Thoreau, August and Friedrich Schlegel, Shakespeare, Socrates, Madame de Staël and Emanuel Swedenborg.

Emerson’s works were well known throughout the United States and Europe in his day. Nietzsche read German translations of Emerson’s essays, copied passages from "History" and "Self-Reliance" in his journals, and wrote of the *Essays*: that he had never "felt so much at home in a book." Emerson’s ideas about "strong, overflowing" heroes, friendship as a battle, education, and relinquishing control in order to gain it, can be traced in Nietzsche’s writings. Other Emersonian ideas-about transition, the ideal in the commonplace, and the power of human will permeate the writings of such classical American pragmatists as William James and John Dewey; and his philosophy is a primary source for Stanley Cavell’s contemporary writing on "moral perfectionism."

Bibliography

Works by Emerson

CW *The Collected Works of Ralph Waldo Emerson*, ed. Robert Spiller et al, Cambridge, Mass: Harvard University Press, 1971-

Z *Selected Essays and Addresses*, ed. L. Ziff, New York: Penguin, 1982

- **P** *Ralph Waldo Emerson*, ed Richard Poirier (The Oxford Authors), Oxford and New York: Oxford University Press, 1990
- "Education," in *Lectures and Biographical Sketches*, in *The Complete Works of Ralph Waldo Emerson*, ed. Edward Waldo Emerson, Boston: Houghton Mifflin, 1883, pp. 125-59
- *The Complete Works of Ralph Waldo Emerson*, ed. Edward Waldo Emerson, Boston: Houghton Mifflin, 12 volumes, 1903-4
- *The Journals of Ralph Waldo Emerson*, ed. Edward Waldo Emerson and Waldo Emerson Forbes. 10 vols., Boston and New York: Houghton Mifflin, 1910-14
- *The Journals and Miscellaneous Notebooks of Ralph Waldo Emerson*, ed. William Gillman, et. al., Cambridge: Belknap Press, Harvard University Press
- *The Early Lectures of Ralph Waldo Emerson*, 3 vols, Stephen E. Whicher, Robert E. Spiller, and Wallace E. Williams, eds., Cambridge, Mass: Harvard University Press, 1961-72
- *The Letters of Ralph Waldo Emerson*, ed. Ralph L. Rusk and Eleanor M. Tilton . 10 vols. New York: Columbia University Press, 1964-95
- (with Thomas Carlyle), *The Correspondence of Emerson and Carlyle*, ed. Joseph Slater, New York: Columbia University Press, 1964

(See Chronology for original dates of publication.)

Selected Writings on Emerson

- Allen, Gay Wilson, 1981,*Waldo Emerson*, New York: Viking Press
- Bishop, Jonathan, 1964, *Emerson on the Soul*, Cambridge, MA: Harvard University Press
- Cameron, Sharon, 1986, "Representing Grief: Emerson's 'Experience'," *Representations* 15:15-41.
- Carpenter, Frederick Ives, 1930, *Emerson and Asia*, Cambridge, MA: Harvard University Press
- Cavell, Stanley, 1981, "Thinking of Emerson" and "An Emerson Mood," in *The Senses of Walden, An Expanded Edition,* San Francisco: North Point Press
- ------, 1988, *In Quest of the Ordinary: Lines of Skepticism and Romanticism*, Chicago: University of Chicago Press
- ------, 1989, "Finding as Founding: Taking Steps in Emerson's 'Experience'", in *This New Yet Unapproachable America*, Albuquerque, NM: Living Batch Press
- ------, 1990, "Introduction" and "Aversive Thinking," in *Conditions Handsome and Unhandsome: The Constitution of Emersonian Perfectionism*, Chicago: University of Chicago Press
- Conant, James, 1997, "Emerson as Educator," *ESQ: A Journal of the American Renaissance*, 43:181-206
- Ellison, Julie, 1984, *Emerson's Romantic Style*, Princeton: Princeton University Press
- Firkins, Oscar W., 1915, *Ralph Waldo Emerson*, Boston: Houghton Mifflin
- Goodman, Russell B., 1990a, *American Philosophy and the Romantic Tradition*, Cambridge: Cambridge University Press, Chapter 2
- ------, 1990b, "East-West Philosophy in Nineteenth Century America: Emerson and Hinduism," *Journal of the History of Ideas*, 625-45

- ------, 1997a, "Emerson's Mystical Empiricism," in *The Perennial Tradition of Neoplatonism*, ed. John J. Cleary, Leuven: Leuven University Press, 456-78.
- -----, 1997b, "Moral Perfectionism and Democracy in Emerson and Nietzsche," *ESQ: A Journal of the American Renaissance*, 43:159-80
- Holmes, Oliver Wendell, 1885, *Ralph Waldo Emerson*, Boston: Houghton Mifflin
- Matthiessen, F. O., 1941, *American Renaissance: Art and Expression in the Age of Emerson and Whitman*. New York: Oxford University Press
- Packer, B. L., 1982, *Emerson's Fall*, New York: Continuum
- -----, "The Transcendentalists," in *The Cambridge History of American Literature*, ed. Sacvan Bercovitch, Vol. 2, pp. 329-604.
- Poirier, Richard, 1987,*The Renewal of Literature: Emersonian Reflections*, New York: Random House
- ------, 1992, *Poetry and Pragmatism*, Cambridge, Mass.: Harvard University Press
- Porte, Joel, and Morris, Saundra (eds.), 1999, *The Cambridge Companion to Ralph Waldo Emerson*. Cambridge: Cambridge University Press
- Richardson, Robert D. Jr., 1995, *Emerson: The Mind on Fire*, Berkeley and Los Angeles: University of California Press
- Rusk, Ralph L., 1949, *The Life of Ralph Waldo Emerson*, New York: Scribners
- Whicher, Stephen, 1953, *Freedom and Fate: An Inner Life of Ralph Waldo Emerson*, Philadelphia: University of Pennsylvania Press

Other Internet Resources

- Guide to Resources on Emerson (maintained by Jone Johnson Lewis)

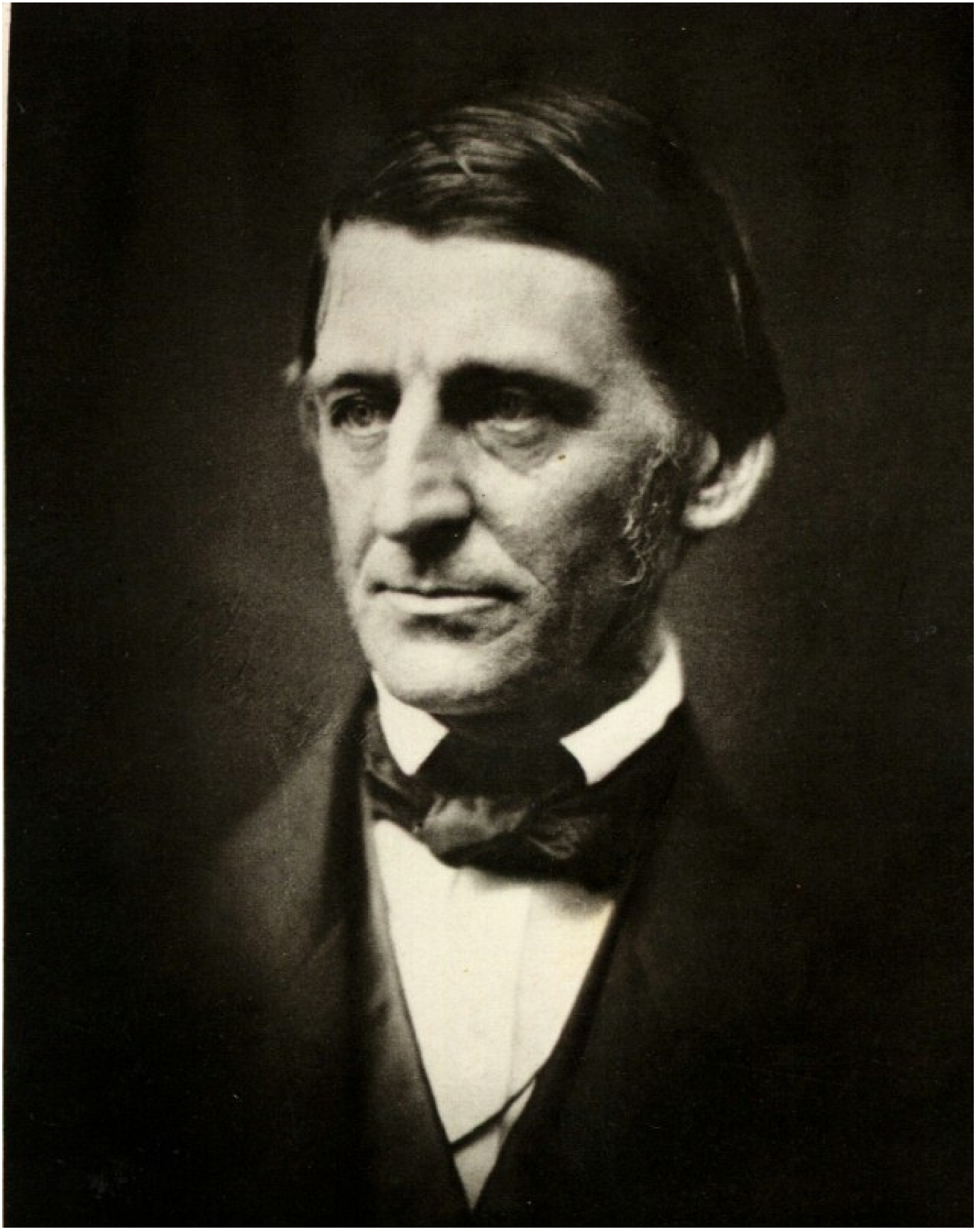
Related Entries

Anaxagoras | Bacon, Francis | Behmen, Jacob | Cambridge Platonists | Cicero | Coleridge, Samuel Taylor | Dewey, John | Goethe | Heraclitus | Iamblichus | James, William | Lucretius | Mencius | Nietzsche, Friedrich | Plotinus | Pythagoras | Schiller, Friedrich | Schlegel, Friedrich | Socrates | St. Augustine | Staël, Madame de | Swedenborg, Emanuel | Thoreau, Henry David | transcendentalism

Russell B. Goodman
U. New Mexico
rgoodman@unm.edu

Table of Contents

First published: January 3, 2002
Content last modified: January 3, 2002



The Epistemic Closure Principle

Most of us think we can always enlarge our knowledge base by accepting things that are entailed by (or logically implied by) things we know. The set of things we know is closed under entailment (or under deduction or logical implication), which means roughly that we know anything that follows from what we know. However, some theorists deny that knowledge is, in fact, closed under entailment, and the issue remains controversial. One might speak of two main camps: those who take closure as a firm datum -- as obvious enough to rule out any understanding of knowledge that undermines closure; and those who want to resolve the controversy by analyzing knowledge and working out the implications for closure. The matter is important, not just because of its bearing on the analysis of knowledge, but also because some theorists say that rejecting the closure principle is the key to defeating skepticism.

---

1. Is Knowledge Closed Under Entailment?

Precisely what is meant by the claim that knowledge is closed under entailment? One response is that the following straight principle of closure of knowledge under entailment is true:

> If person *S* knows *p*, and *p* entails *q*, then *S* knows *q*.

However, the straight principle is obviously false, since we can know one thing, *p*, but fail to see that *p* entails *q*, or for some other reason fail to believe *q*. Since knowledge entails belief (according to nearly all theorists), we fail to know *q*. A less obvious worry is that we might reason badly in coming to believe that *p* entails *q*. Perhaps we think that *p* entails *q* because we think everything entails everything, or because we have a warm tingly feeling between our toes.

Obviously, the straight principle needs qualifying, but this should not concern us so long as the qualifications are natural given the idea we are trying to capture, namely, that we can extend our

knowledge by recognizing, and accepting thereby, things that follow from our knowledge. The qualifications embedded in the following principle seem natural enough:

> If *S* knows *p*, and believes *q* by recognizing that *p* entails *q*, then *S* knows *q*.

If we continue in this way, qualifying the closure principle to handle counterexamples, can we finally devise a version that is true?

1.1 Tracking and Closure

Perhaps, but Fred Dretske, Robert Nozick and others have made a powerful case for rejecting any such principle, however qualified. They attack the closure principle on the basis of a distinctive analysis of knowledge -- what Nozick calls the tracking theory. To know *p* is to track the fact that *p*; simplifying greatly, the claim is that *S* tracks (and hence knows) *p* if and only if there is a reason *R* such that (a) *S*'s belief *p* is based on the fact that R holds, and (b) if *p* were false, *R* would not hold. Somewhat surprisingly, it turns out that we can track a fact, believe a second fact by seeing, correctly, that it follows from the first, yet fail to track the second fact. The main reason is that the following inference is invalid:

1. If *p* were false then *R* would not hold (i.e., we track *p*)

2. *p* entails *q*.

3. So if *q* were false then *R* would not hold (i.e., we track *q*).

Consider one of Dretske's own illustrations: suppose you are in an ordinary zoo standing in front of a cage marked 'zebra'. There is, in fact, a zebra there, in plain sight, and you believe *z*: the animal in the cage in front of you is a zebra. You believe *z* on the basis of *R*, your having zebra-like sensory impressions. You track the fact that *z*; if the animal were not a zebra, you would have different sensory impressions -- *R* would not hold. But now notice that *z* entails *not-m*: the animal in the cage is not a mule that is cleverly disguised to look like a zebra. Seeing this, you believe *not-m*. Your basis for believing *not-m* is your having zebra-like sensory impressions, together with the recognition that these support the belief *z* and that *not-m* follows from *z*. Nonetheless, you fail to track the fact that *not-m*, for if the animal were a cleverly disguised mule, you would still have zebra impressions, leading you to believe *z* and ultimately *not-m*.

As Dretske and Nozick point out, we will probably reject the closure principle if we equate knowledge with tracking. However, many theorists prefer to argue in the other direction: they have assumed, somewhat in the style of G. E. Moore (1959), that the closure principle, suitably qualified, is obviously true, so any argument to the contrary must be flawed. Hence it is the tracking account that must be rejected.

1.2 Indication, Safety, and Closure

Sosa (1999, 2001) and Luper (1984, 1987b) sustain the closure principle by replacing (b) of the tracking account with its contraposition -- namely, *R* would hold only if *p* were true. When this condition is met, *S*'s belief *p* has a property that Sosa calls *safety*; alternatively, one might say that *R indicates* that *p* is true. Suppose that *S* knows *p* if and only if *S* bases the belief *p* on a reason *R* that indicates that *p* is true. Then knowledge is closed, since *R* indicates that a proposition *p* is true only if *R* also indicates that *p*'s consequences are true. For the following inference (via strengthening the consequent) is valid:

1. *R* would hold only if *p* were true (i.e., *R* indicates *p*)

2. *p* entails *q*

3. So *R* would hold only if *q* were true (i.e., *R* indicates *q*).

1.3 Reliabilism and Closure

According to the reliabilist theory of knowledge, *S* knows *p* if and only if *S*'s true belief *p* is caused or sustained by a reliable belief-formation mechanism. Is the reliabilist committed to closure? The answer depends on precisely how the relevant notion of reliability is understood. One of the first reliabilist theories, offered by Alvin Goldman, is very similar to the tracking view, for Goldman argued that knowing *p* entails having the capacity to discriminate between the situation in which *p* is true, on the one hand, and alternative situations (in which *p* is false) that might arise given the circumstances at hand. If we understand reliability as tracking theorists do, we will reject closure. But there are other versions of reliabilism, which sustain closure. For example, the indicator account is a type of reliabilism. Also, we could say that a true belief *p* is reliably formed if and only if based on an event that *usually* would occur only if *p* (or a *p*-type belief) were true. Any event that, in this sense, reliably indicates that *p* is true will also reliably indicate that *p*'s consequences are true.

1.4 Internalism and Closure

Reliabilists adopt externalist accounts of knowledge: they say that factors outside of a subject's perspective can help determine whether she knows things. Other theorists embrace internalism: the view that only factors internal to a subject's perspective can affect the epistemic status of her beliefs. Will internalists accept the principle of closure? The answer is not obvious, since there is no agreed-upon internalist account of knowledge. However, almost all internalists will agree that knowledge entails justified true belief. (Some externalists will accept this too, usually after giving justification an externalist analysis, say in terms of reliable indication.) Assuming they are correct, then the issue of closure rests on whether or not justification is itself closed under entailment. Is it? The matter is not settled. Some argue that justification is not closed, using counterexamples like Dretske's own zebra case: because the zebra is in plain sight, you seem fully justified in believing *z*, but it is not so clear that you are justified in

believing *not-m*, even if you deduce this belief from *z*. Your zebra-like experiences constitute good evidence for *z*, but are they good evidence for *not-m*? The following principle seems questionable:

> If *E* is evidence for *p*, and *p* entails *q*, then *E* is evidence for *q*.

But even if we reject this principle (of the transmissibility of evidence or justification), it does not follow that justification is not closed under entailment, as Peter Klein pointed out. For closure of justification, all that is necessary is that when, given all of our relevant evidence *E*, we are justified in believing *p*, we also have sufficient justification for *p*'s consequences. Our justification for *p*'s consequences need not be *E*. Instead, it might be *p* itself, which is, after all, a justified belief. And since *p* entails its consequences, it is sufficient to justify them. Moreover, any good evidence we have against a consequence of *p* counts against *p* itself, preventing us from being justified in believing *p* in the first place, so if we are justified in believing *p*, considering all our evidence, pro and con, we will not have overwhelming evidence against propositions entailed by *p*. (A similar move could be defended against the tracking theorists when they deny the closure of knowledge: if we track *p*, and believe *q* by deducing it from *p*, then we track *q* if we take *p* as our basis for believing *q*.) Looked at in this way, the following principle of closure of justification seems correct:

> If, considering all of *S*'s evidence, *S* is justified in believing *p*, and *S* believes *q* by deducing it from *p*, then *S* is justified (on the basis of *p*, or alternatively, on the basis of *p* plus the deduction itself) in believing *q*.

However, to avoid paradox, it must be understood that our principle applies only to the implications of individual propositions, not to conjunctions of propositions. We are not always justified in believing the conjunction of claims that are individually justified. To see why, notice that if the chances of winning a lottery are sufficiently remote, then I am justified in believing that ticket 1 will lose. I am also justified in believing that ticket 2 will lose, and that 3 will lose, and so on. However, I am not justified in believing the conjunction of these propositions. If I were, I would justifiably believe that no ticket will win. Yet I know that some ticket will.

2. Is Denying Closure the Key to Resolving the Issue of Skepticism?

According to tracking theorists, we can account for the appeal of skepticism and explain where it goes wrong if we accept their view of knowledge and reject the principle of closure of knowledge. Rejecting closure is therefore the key to resolving skepticism. Given the importance of insight into the problem of skepticism, they would seem to have a good case for denying closure. Let us consider the story they present, and some worries about its acceptability.

2.1 Tracking and Skepticism

According to tracking theorists, skepticism is appealing because skeptics are partially right. They are correct when they say that we do not know that skeptical hypotheses fail to hold. For example, I do not know *not-biv*: I am not a brain in a vat on a planet far from earth being deceiving by alien scientists. For I do not track *not-biv*: if *biv* were true, I would still have the experiences that lead me to believe that *biv* is false. Skeptics are also correct when they point out that *not-biv* is entailed by all sorts of commonsense claims, such as *h*: I am in San Antonio. Having gotten this far, skeptics appeal to the principle of closure, and argue that since I would know *not-biv* if I knew *h*, then I must not know *h* after all. But this is precisely where skeptics go wrong: having accepted the tracking view -- as they do when they deny that we know skeptical hypotheses are false -- skeptics cannot appeal to the principle of closure, which is false on the tracking theory. We track (hence know) the truth of ordinary knowledge claims yet fail to track (or know) the truth of things that follow, such as that incompatible skeptical hypotheses are false.

One problem with this story is that it cannot come to terms with all types of skepticism. There are two main forms of skepticism (and various sub-categories): regress (or Pyrrhonian) skepticism, and indiscernability (or Cartesian) skepticism. At best, Dretske and Nozick have provided a way of dealing with indiscernability skepticism, not regress skepticism.

Another worry about the tracking theorist's account of indiscernability skepticism is that it forces us to give up the principle of closure. Given the intuitive appeal of this principle, some theorists have looked for alternative ways of explaining skepticism. Consider two possibilities, one offered by advocates of the indicator theory and one by contextualists.

2.2 Indication and Skepticism

Advocates of the indicator theory (Sosa 1999, Luper 1987b, 2001) accept the gist of the tracking theorist explanation of the appeal of skepticism but retain the principle of closure. The reason skepticism tempts us is that we tend to confuse the tracking account with the indicator account. After all, the tracking condition -- if *p* were false, *R* would not hold -- closely resembles the indication condition -- *R* would hold only if *p* were true. When we run the two together, we sometimes apply the tracking account and conclude that we do not know that skeptical scenarios do not hold. Then we shift back to the indicator account, and go along with skeptics when they appeal to the principle of entailment, which is sustained by the indicator account, and conclude that ordinary knowledge claims are false. But skeptics are wrong when they say we do not know that skeptical hypotheses are false. Roughly, we know skeptical possibilities do not hold since (given our circumstances) they are remote.

2.3 Contextualism and Skepticism

Contextualists, such as Stewart Cohen (1988, 1999), Keith DeRose (1995) and David Lewis (1996), offer yet another way of explaining skepticism without denying closure. According to contextualists, whether it is correct for a judge to attribute knowledge to someone depends on that judge's context, and the standards for knowledge differ from context to context. When the man on the street judges knowledge, the applicable standards are relatively modest. But an epistemologist takes all sorts of possibilities

seriously that are ignored by ordinary folk, and so must apply quite stringent standards in order to reach correct assessments. What passes for knowledge in ordinary contexts does not qualify for knowledge in contexts where heightened criteria apply. Skepticism is explained by the fact that the contextual variation of epistemic standards is easily overlooked. Skeptics note that in the epistemic context it is inappropriate to grant anyone knowledge. However, skeptics assume -- falsely -- that what goes in the epistemic context goes in all contexts. They assume that since those who take skepticism seriously must deny anyone knowledge, then everyone, regardless of context, should deny anyone knowledge. Yet people in ordinary contexts are perfectly correct in claiming that they know all sorts of things.

Furthermore, the closure principle is correct, contextualists say, so long as it is understood to operate within given contexts, not across contexts. That is, so long as we stay within a given context, we know the things we deduce from other things we know. But if I am in an ordinary context, knowing I am in San Antonio, I cannot come to know, via deduction, that I am not a brain in a vat on a distant planet, since the moment I take that skeptical possibility seriously, I transform my context into one in which heightened epistemic standards apply. When I take the vat possibility seriously, I must wield demanding standards that rule out my knowing I am not a brain in a vat. By the same token, these standards preclude my knowing I am in San Antonio. Thinking seriously about knowledge undermines our knowledge.

Bibliography

- Audi, R., 1995, "Deductive Closure, Defeasibility and Scepticism: A Reply to Feldman." *Philosophical Quarterly* 45: 494-499.
- Bogdan, R.J., 1985, "Cognition and Epistemic Closure," *American Philosophical Quarterly* 22: 55-63.
- Brueckner, A., 1985a, "Losing Track of the Sceptic," *Analysis* 45: 103-104.
- -----, 1985b, "Skepticism and Epistemic Closure," *Philosophical Topics* 13: 89-117.
- -----, 1985c, "Transmission for Knowledge Not Established," *Philosophical Quarterly* 35: 193-196.
- Cohen, S., 1987, "Knowledge, Context, and Social Standards," *Synthese* 73: 3-26.
- -----, 1988, "How to be a Fallibilist," *Philosophical Perspectives 2: Epistemology,* Atascadero, CA: Ridgeview, 91-123.
- -----, 1999, "Contextualism, Skepticism, and the Structure of Reasons," *Philosophical Perspectives 13: Epistemology*, Atascadero, CA: Ridgeview, 57-89.
- DeRose, K., 1995, "Solving the Skeptical Problem," *Philosophical Review* 104: 1-52.
- Dretske, F., 1970, "Epistemic Operators," *Journal of Philosophy* 67: 1007-1023.
- -----, 1971, "Conclusive Reasons," *Australasian Journal of Philosophy* 49: 1-22.
- -----, 1972, "Contrastive Statements," *Philosophical Review* 81: 411-430.
- Feldman, R., 1995, "In Defense of Closure," *Philosophical Quarterly* 45: 487-494.
- Goldman, A., 1976, "Discrimination and Perceptual Knowledge," *Journal of Philosophy* 73: 771-791.
- -----, 1979, "What is Justified Belief?," in *Justification and Knowledge,* G.S. Pappas (ed.), Dordrecht: D. Reidel.

- Hume, David, 1739-1740, *A Treatise of Human Nature*, L.A. Selby-Bigge (ed.), 2nd edition, Oxford: Clarendon Press, 1978.
- Klein, P., 1981. *Certainty: A Refutation of Skepticism*, Minneapolis, MN: University of Minnesota Press.
- -----, 1995, "Skepticism and Closure: Why the Evil Genius Argument Fails," *Philosophical Topics* 23: 213-236.
- Lewis, D., 1979, "Scorekeeping in a Language Game," *Journal of Philosophical Logic* 8: 339-359.
- -----, 1996, "Elusive Knowledge," *Australasian Journal of Philosophy* 74: 549-567.
- Luper(-Foy), S., 1984, "The Epistemic Predicament: Knowledge, Nozickian Tracking, and Skepticism," *Australasian Journal of Philosophy* 62: 26-50.
- -----, (ed.), 1987a, *The Possibility of Knowledge: Nozick and His Critics*, Totowa, NJ: Rowman and Littlefield.
- -----, 1987b, "The Causal Indicator Analysis of Knowledge," *Philosophy and Phenomenological Research* 47: 563-587.
- -----, 2001, "Indiscernability Skepticism," in *The Skeptics*, S. Luper (ed.), Hampshire: Ashgate Publishing, Limited, forthcoming.
- Moore, G. E., 1959, "Proof of an External World," and "Certainty," in *Philosophical Papers*, London: George Allen & Unwin, Ltd.
- Nozick, R., 1981, *Philosophical Explanations*, Cambridge: Cambridge University Press.
- Sextus Empiricus, 1933a, *Outlines of Pyrrhonism*, R.G. Bury (trans), London: W. Heinemann, Loeb Classical Library.
- Sosa, E., 1999, "How to Defeat Opposition to Moore," *Philosophical Perspectives* 13: 141-152.
- -----, 2001, "Neither Contextualism Nor Skepticism," in *The Skeptics*, S. Luper (ed.), Hampshire: Ashgate Publishing, Limited, forthcoming.
- Stine, G.C., 1971, "Dretske on Knowing the Logical Consequences," *Journal of Philosophy* 68: 296-299.
- -----, 1976, "Skepticism, Relevant Alternatives, and Deductive Closure," *Philosophical Studies* 29: 249-261.
- Vogel, J., 1990, "Are There Counterexamples to the Closure Principle? in *Doubting: Contemporary Perspectives on Skepticism*, M. Roth and G. Ross (eds.), Dordrecht: Kluwer Academic Publishers.

Related Entries

knowledge: analysis of | skepticism | skepticism: ancient

Steven Luper
sluper@trinity.edu

Table of Contents

First published: December 31, 2001
Content last modified: December 31, 2001

Antonio Rosmini

Antonio Rosmini (1797-1855), Italian priest, philosopher, theologian and patriot, and founder of a religious congregation, aimed principally in his philosophical work at re-addressing the balance between reason and religion which had largely been lost as a result of the Enlightenment. To this purpose, he absorbed the tradition of *philosophia perennis*, read extensively the works of post-Renaissance philosophers, and developed his own views on philosophical fundamentals and many of their applications. Best known in Italy, but a controversial figure there during his life and for a century or more after his death, his philosophical work, centred upon the notion of being and the dignity of the human person, can be summarised under the headings: aims and method, the objectivity of thought and the concept of certainty; the dignity of the human person; morality; human rights; the nature of human society; natural theology; and being. The following article will examine Rosmini's work under these titles, which are of perennial relevance and broad enough to embrace more particular themes, such as art, politics, education and marriage, which form a constant preoccupation of many of his lesser works, but can only be mentioned in passing here. His theological principles, other than those pertaining to natural theology, are considered only in so far as they throw light on the origin and development of his philosophical tenets.

1. Life and Works

Rosmini was born in 1797 at Rovereto, Italy, a staunchly Italian-speaking town, but at the time part of the Austro-Hungarian Empire and ruled from Vienna. The Rosmini family, citizens of Rovereto for several centuries, had become rich through the manufacture of silk, and enjoyed the way of life proper to the lesser aristocracy. Antonio's primary and secondary education was, however, catered for at the public school, and through his own intensive reading. Tertiary education was completed at the University of Padua. After ordination to the priesthood in 1821, Rosmini studied and wrote at Rovereto until 1826, and from 1826-28 at Milan. Despite his instinctive distaste for the excesses of the French Revolution, which inevitably played a large part in the cultural formation of persons growing to maturity in the first quarter of the 19th century, Rosmini was not afraid to take a stand against State interference in religious affairs. His *Panegyric for Pius VII* (1823) was considerably censored by the Austrian-Hungarian government and published only in 1831. By this time, Rosmini had founded his religious order (1828), and published at Rome his fundamental philosophical work, *A New Essay concerning the Origin of Ideas* (1830). As the title suggests, this was intended to supersede Locke's famous *Essay concerning Human Understanding*. From then on, a torrent of philosophical and theological works poured out until his death at Stresa. From 1836 until 1855, Rosmini was involved in constant controversy. The approbation of his religious order (1836-38), his work on conscience (1840), theological disputes (1843-48) and his participation in the political events of 1848, occasioned strong debate which led in 1849 to the inclusion of two of his works, *The Five Wounds of Holy Church* and the *Civil Constitution according to Social Justice*, in the *Index of Prohibited Books*. Strengthened by this, Rosmini's theological and political opponents pressed for an examination of all his works which led, however, to his exoneration (1854), a year before his death at Stresa in northern Italy (1855). Condemnation came posthumously (1888-9) when forty propositions, taken mostly from books published after his death, were included in the decree *Post Obitum* of Leo XIII. A remarkable swing in ecclesiastical opinion took place in 2001 when a Vatican *Note* stated: 'The motives for preoccupation and for doctrinal and prudential difficulties which determined the promulgation of the decree *Post Obitum* condemning the "Forty Propositions" drawn from the works of Antonio Rosmini can now be considered as surmounted.' (CDF, *Osservatore Romano*, 1 July 2001)

2. Aim and Method

Rosmini describes his aim and method at some length in his *On the Studies of the Author*. He sets out to combat error, to systematise the truth, to present a philosophy that can serve as a basis for the various branches of knowledge, and to offer philosophy as an aid to theology. To achieve this, he upholds freedom to philosophise, and sets out to reconcile, whenever possible, apparently contrasting views. His intention throughout is to present an image of knowledge as one, simple and indivisible.

Combating error

No one, he maintains, would err for the sake of erring. Philosophical tradition in particular provides an object lesson in the movement towards truth, and the elimination of error. Nevertheless, the path forward is not pursued without error because the movement towards higher levels of reflection takes place unevenly. Responses to questions at level A are no longer adequate in form to resolve questions at level

B, which inevitably take on new aspects through their application to new circumstances. The role of the philosopher is to distinguish the form of difficulties, which may vary from age to age or generation to generation, and to formulate questions in such a way that it is possible to see both their historical antecedents and the underlying principles to be employed in solving them. The process, however, will never be complete. The same principles will always cry out for application to new cases, and the same struggle to avoid error will ensue.

Systematising the truth

The negative task of combating error is not, however, sufficient. What is needed in addition, says Rosmini, is a 'system of truth', that is, a system which shows clearly how the passage is made from the most general, self-evident principles to more particular levels of knowledge. Knowledge-wise, we move instinctively from the more general to the particular. A mother does not begin by naming roses, carnation and other species for her children; she first indicates them all as 'flowers' before calling them by their particular names. So 'being', which is the most general of all notions, is the fundamental principle of human knowledge which draws together every aspect of being. When 'being' is seen as the supreme principle of unity on which all knowledge depends, truth -- 'being as known' -- has been systematised and is seen in all its beauty. Since, however, the full application of being is never seen once and for all, it is better to ensure adherence to principle than to grasp at unconnected facts which serve at best as a ragbag of erudition -- better to grasp the principle of the wheel, for example, than to know only a number of its applications.

Philosophy as a basis for the various branches of knowledge

Philosophy as 'the study of the final reasons' is thus central to Rosmini's understanding. For him, the Enlightenment, with its sensistic, subjectivist attitude and devotion to the act of reasoning, rather than to the light of reason, degenerates inevitably into a hotchpotch of negation and ignorance, leading to radical corruption in ethics, and every other branch of philosophical endeavour.

Philosophy as an aid to theology

The fragmentation of philosophy and its consequent separation from theology is, according to Rosmini, a necessary consequence of sensationalist thinking. There can be no place for revealed doctrine to be expounded as true science unless certain truths are already demonstrated, in the logical order at least, by philosophical reason. On the other hand, theology itself often cannot make progress unless it is prepared to turn to philosophy for assistance. The notions of body, of person and of many other matters essential to theology, cannot be adequately expressed in isolation from philosophical teaching. In its turn, according to Rosmini, divine revelation does not cancel, but completes and ennobles reason, especially by drawing its attention to problems such as the relationship between person and nature which would otherwise escape its attention.

Freedom to Philosophise

Error, the antithesis of knowledge, is the only intellectual impediment to free, philosophical thought. From this principle, Rosmini concludes that assent to erroneous prejudices, not assent to prejudice as such, is the principal obstacle to be overcome by philosophers. Their work consists in examining preconceptions and determining their truth in order to provide grounds for rational persuasion about what they know. To maintain, as many do, that the possession of some unproven truth is inimical to philosophical thought is tantamount to requiring nil knowledge in the prospective philosopher. Rather, a person who knows something, but has not yet come to grips with the reasons leading to it, is like a person who knows the answer to a problem, but still has to consider the reason for the answer. In this case, freedom is not constrained. The point at issue, therefore, in the case of religion, is not that Christians, or Buddhists, or Muslims, or any other religious persons, are necessarily hampered by their beliefs, but whether these beliefs are true, and to what extent they are true. It is not sufficient to state simply that only persons who are devoid of any belief are capable of philosophising freely. Rather

> total freedom is a necessary condition of the truth of faith. If faith were considered divine although in conflict with reason, it would impose an impossible obligation and totally inhibit our reasoning activity. We would be unable to give our assent to either reason or faith, and would thus remain deprived of truth (*IP* 39 [All numbers in references to Rosmini's works are to paragraphs, not to pages]).

It is not the case, Rosmini would affirm, that only non-believers have the capacity to enter the world of philosophical enquiry.

Reconciliation of Conclusions

According to Rosmini, eclecticism, especially that upheld by Victor Cousin, is not the way to promote reconciliation between conclusions. Philosophical systems are not brought together as a result of arbitrary choice between what they offer. Each system, if it is truly such, will have a principle from which deductions are made, and will be able to be reconciled, despite accidental differences, with every other system sharing the same principle. On the other hand, systems will not be reconcilable, despite their accidental agreement, if their basic principles differ. In the former case, agreement will be possible by working back to the principle, and setting out once more from there; in the second case, apparent agreement will only be skin-deep. Only shared principles allow for effective reconciliation between systems.

3. The Objectivity of Thought and the Concept of Certainty

Rosmini sets out to establish the nature of thought and certainty in his *A New Essay concerning the*

Origin of Ideas. Faced by the critical philosophy of Kant on the one hand, and by British empiricism on the other, Rosmini reaches back to the pre-Socratics, to Plato and Aristotle, to Augustine, Thomas Aquinas and Bonaventure, in an endeavour to establish the nature of thought and the basis of certainty in human existence. As his guiding principles he takes the following rules:

> In explaining facts connected with the human spirit, we must not make fewer assumptions than are required to explain them... [nor must we] make more assumptions than are needed to explain the facts (*NE*, vol. 1, 26-7).

With this as his methodological foundation, he places Locke, Condillac, Reid and Dugald Stewart among those whose explanation of the fact of thought is deficient; Plato, Aristotle, Leibniz and Kant are listed amongst those whose explanation is in many ways excessive. In other words, he distinguishes between sensationalists who, according to him, cannot explain the origin of ideas, and idealists who posit in their explanation more ideas or forms than are necessary. For him, however, human thought must depend upon the innate idea of being, without which nothing is intelligible. We cannot think of what is not. At the same time, nothing more is needed than the single idea of being, and its possible determinations, brought about through sensation, to explain all intellectual principles and the ramifications of thought.

Objectivity

After observing the fact of thought, Rosmini concludes that its absolute basis, without which nothing is thought and thought is nothing, must be the knowledge of being. Reflection can remove everything from thought, and still leave it embryonically sound provided the mind is granted the idea of being as its governing light. This idea, which possesses the divine attributes of universality, infinity, necessity and possibility, is not God himself but merely the possibility of things. Moreover, it acts in the mind, but without becoming a subjective part of the mind. As intellectual light, it illuminates, but from outside the mind. And it illuminates without revealing its source, as natural light can be seen without our looking at the sun. It is, as Dante would say, ‘the light connecting intellect and truth’ (*Purgatorio* 6: 45) and, as such, is the quasi-form of the intellect and the image of truth. Determinations of this idea, all of which possess in some way the characteristics of being, especially its objectivity, are provided through sensations experienced in the animal part of human existence and illuminated, being-wise, through the innate idea in which they are seen. Of themselves, sensations do not constitute knowledge but, when felt in the human subject, provide the matter of knowledge which determines the idea of being, that is, the form of knowledge. Knowledge consists first in the intuition of being, the universal, and then in a subsequent series of judgments, or direct perceptions, through which knowing subjects affirm the actual existence of what they have experienced sensationally. These direct perceptions cannot err, although reflection upon them, and a subsequent series of judgments, may be the source of error. In a word, Rosmini holds that basic knowledge, consisting of the idea of being and its immediate determinations, provides all that is needed for objective thought. Against idealists, he reduces the formal requirements of thought to the intuition of being; against sensationalists, he maintains the *per se* inadequacy of the senses to provide more than the matter of thought.

Certainty

Objectivity, therefore, is essentially a characteristic of what is known. Certainty, in Rosmini's view, is a characteristic of the person who knows, and can be defined as 'a firm and reasonable persuasion that conforms to the truth' (*NE*, vol. 3, 1044). In other words, we can be certain only of knowledge, not of error, and this because, according to Rosmini, to know and to know the truth is the same thing. The person who does not know the truth, does not know. There is no doubt, of course, that it is possible to be persuaded, and firmly persuaded, of error. But rational persuasion of error arrived at through one's own reasoning is not possible. In this case, either the premiss is wrong, or the argument is erroneous. On the other hand, the persuasion must be firm. Certainty is not achieved without energy directed into persuasion. Certainty requires that we know something to be true, to be what it is, that we are persuaded that it is what we know it to be, and that we have an adequate reason for our persuasion. And precisely because error attempts to alter the being of things, formal error will not be found rooted in the intellect nor in the senses nor in involuntary reflection. It begins with the will, the only human faculty capable of drawing the reason to invent what it does not see, or to deny what it sees. Under pressure from the will, reason will falsely affirm that being is not, or deny that being is.

4. The Dignity of the Human Person

It is already clear from what has been said that Rosmini, in his solution to the basic problem of knowledge, has offered a perspective which places human dignity on a transcendent level. Human beings are made such, he would maintain, by the intuition of being which accompanies them from the first moment of their existence. Through this intuition, they share in the finality of being itself and in some finite way participate in its infinite characteristics. Nevertheless, the subjective element proper to the human being neither can nor should be denied. Indeed, it has to be examined thoroughly if a rational account is to be rendered of the essential unitary make-up of human nature. This examination was carried out initially in Rosmini's *A New Essay concerning the Origin of Ideas* as part of his explanation of what he called 'impure' ideas, that is, ideas which, as perceptive (not simply intellective), require for their origin some sense experience. He pursued the investigation in his *Anthropology as an Aid to Moral Science* and in *Psychology* where in four volumes entitled *The Essence of the Human Soul*, *The Development of the Human Soul*, *Laws of Animality* and *Opinions about the Human Soul*, he observed and discussed at length the animal, as well as the intellectual side of human nature.

The first fact presented by observation on ourselves is the essential distinction in human beings between that which feels and that which is felt. According to Rosmini, these are quite different and unconfusible elements with opposite characteristics. That which feels is an immaterial principle (soul); that which is felt is the term (body) of this principle. Together feeler and felt constitute feeling, the underlying subject of reflection when attention is concentrated on 'myself' and those elements which constitute 'myself'. The chief action of body is to produce an undetermined, shapeless extension which enables that which feels to experience determined sensations of various kinds. This 'fundamental feeling', with its permanent perception of unlimited space, is that in which all other feelings are perceived, and runs parallel, as it were, with the idea of being in which all other ideas are intuited. Reaction to perceived

feelings constitutes vital instinct (relative to the fundamental feeling) and sensuous instinct (relative to the adventitious sensations). All feeling and instinct is *per se* unknown in Rosmini's view, and comes to be known through the formation in the knowing subject of ideas and perceptions which depend upon the illumination provided by the idea of being. Thus, for Rosmini, the human being is a knowing and feeling subject, having within itself a principle which, formed by the light of being, knows and reacts to what it knows through the faculties of intellect and will, and feels and reacts to what it feels through the faculties of sense and instinct. Within the human subject, the will - which reacts to what is known -- is the supreme active principle and as such constitutes 'person' in the individual. The dignity of the human being lies within the will as such, in the first place, and then within the will's choice to second whatever the intellect knows. This is also the foundation of genuine freedom within the human subject. Free to adhere to or reject what is known, human beings cannot be coerced by attempted external pressure or used as a means by others without prejudice to the inviolable truth in which they share innately through their participation in the light of being and which they attain adventitiously through the direct perception that unfolds determined truths to their intellectual gaze.

5. Morality

Rosmini's ethical philosophy springs directly from the analysis summarised in the previous section. For him, 'the human being is a knowing and feeling subject whose will, as supreme principle of activity, provides the basis of the incommunicable individuality that constitutes each real human nature as a person' (*Life*, p. 26). 'Person', as supreme principle, is also the subject of moral activity, and is to be distinguished from all those habits and acts within human beings which take place without the necessary intervention of the person or, at most, effect the person indirectly. For example, a good pianist is not necessarily a good person. Morality deals necessarily with what people do as persons, with what affects themselves as subjects who cannot step aside from the truth they know without violating their adherence to truth. Rosmini deals with ethics, the science of morality, in his *Principles of Ethics* and in *Comparative and Critical History of Systems dealing with the Principle of Morality*. The *History* first offers an overview of systems which throughout history deny morality, or make it impossible, or provide it with a subjective foundation. It then moves on to summarise the objective view expressed by philosophers and developed by Rosmini in *Principles of Ethics* .

Moral Obligation

This view, as stated by Rosmini, depends upon what he sees as a self-evident principle that cannot be denied by any sane person. We are bound, he says, to acknowledge (recognise) what we know for what we know it to be. Granted that knowledge, according to Rosmini, is co-terminus with truth, never with error, it is clear that Rosmini's affirmation depends upon the self-evident need to affirm what one knows. This, in turn, is an act by which the human subject is elevated, at a reflective level, to the level of the truth which, already known in the essential idea of being, is as sublime as being itself. On the other hand, to deny what one knows is equivalent to stepping into non-being, the antithesis of dignity. In more philosophical terms, Rosmini maintains that the principle he enunciates, in which every other moral dictate is implicitly included, may be characterised as follows: it expresses moral essence because it calls

for agreement between the intellective act of the will and the ideal and real entities of things known; it is simple and as such able to be participated by every moral activity; it is evident because it is nothing more than an expression of the principle of contradiction; it is universal because all moral effects, denoted through external actions, depend upon it; it is supreme because it offers no possibility of further investigation; it provides the foundation for the recognition of the human subject as an essentially moral being, that is, one who, through the innate presence of the idea of being and the principle of will, is furnished with the object of morality and essential adherence to it. Finally, Rosmini's distinction between the subject and object of morality opens a way, according to him, between the extremes of ethical theory. On the one hand, the limitation of the human subject provides for the possibility of moral error; on the other, the necessity and immutability of the object, the idea of being, furnish morality with its undeniable sense of obligation. It is a fatal but all too frequent mistake, Rosmini contends, to attribute characteristics proper to the subject (fallibility, error, and so on) to the object, and characteristics of the object (necessity, immutability and so on) to the subject.

Moral Good

A way is now present to describe adequately the nature of moral good as opposed to eudemonological good. That which is good is desirable, but what is desirable may either be desirable in itself, that is, as it stands in the order of being, or simply desirable in so far it is good, or imagined as good, for the individual. Moral good, says Rosmini, is found when the will adheres to what is good according to the order of being; eudemonological or utilitarian good is what is desired as good for the individual, without reference to what is good in itself. Human dignity is preserved only when, through an act of will, individuals adhering in practice to beings as they are in their order, implicitly adhere to the whole of being and to their presence in that order. Immorality, by which entities are appreciated or desired but not in their known order and thus not as they are, implies an essential rebellion against the order of being and thus against being itself. Self-imposed human indignity can go no lower than this.

Conscience

Having dealt with the nature of morality and moral obligation, Rosmini turns his attention to conscience. He himself admits that such a treatise is almost universally neglected by philosophers, and feels himself constrained to justify such a study, which he undertakes with an almost inevitable admixture of religious and theological elements. Nevertheless, it is not difficult to filter the philosophical principles in his work from characteristics proper to the faith he professes. This is especially apparent in his approach to the nature of conscience, an area in which his views first prompted the dissension's between him and some of his co-religionists that would go on until the end of his life. Rosmini defines conscience as a 'speculative judgement that a person makes about the morality of his practical judgement' (*C*19), that is, a judgement by which individuals come to know the moral value of their actions without necessarily acting upon it.

Several consequences can be drawn from this. First, morality in the individual is prior to conscience, and can be present without conscience; second, conscience can be mistaken (it is possible to form a incorrect

notion of the moral value of one's own action); third, it does not follow that conscience, once formed, will give rise to action in the person who forms it; fourth, conscience, if incorrect in its judgement, must as far as possible be reformed. Conscience, therefore, is only an adequate guide to morality when it provides accurate information of the moral state of a subject's past, present or future action. In the light of these affirmations, it is possible to see how Rosmini lays the groundwork for overcoming the dilemma posed by the question: must conscience always be followed? While it is certain that the dictates of conscience can never be morally disregarded, it is equally certain that a deliberately misleading act of conscience cannot morally be followed. Sometimes, therefore, it will be morally imperative to correct conscience, which is always possible through proper reflection on the moral value of the human act posited or about to be posited by the individual. Rosmini goes on to distinguish between problems about conscience and problems connected with the formation of conscience. In fact, conscience, according to him, is not present as long as judgement is suspended about the moral value of an individual's own action. Difficulties at this point are connected with the formation of the judgement, not with the judgement itself.

6. Human Rights

Rosmini's view of 'person', seen as an inviolable end which can never be reduced to the status of 'means', leads spontaneously to what today is seen as paramount in human existence, that is, the question of human rights. These rights are studied at length in *The Philosophy of Right*, a six-volume treatise in the only extant English translation. The general title of the work shows immediately that for Rosmini all rights are founded in a single element called 'right' from which all 'rights' emanate, some innate in human beings, others springing from the determined circumstances of individuals or societies. The treatise can be summed up under three heads: the essence of right, individual rights, social rights.

The Essence of Right

The basis of Rosmini's teaching on human rights is a consequence of his moral theory, of which he gives a careful synthesis in a preface to *The Philosophy of Right*. If each person is morally obliged to recognise in practice what is known for what it is known to be, every other human being will be recognised as essentially on a par with the knowing person, and will have to be acknowledged as such. But because each person is obliged to act in accordance with moral propriety, every person is obliged to respect this obligation in the other on pain of violating the moral law itself. Granted this principle as foundation, and 'right' as a relationship between one person and others by which a person has a claim to what is his own, Rosmini maintains that 'person' is subsistent right. In other words, all rights are founded on that to which persons have a claim in so far as they are acting morally or are at least not acting immorally. Such activity cannot be the object of attack on the part of others without violation of persons as ends. 'Right', as he says, 'is a moral governance or authority to act, or: right is a faculty to do what we please, protected by the moral law which obliges others to respect that faculty' (*ER*, vol. 1, 237).

The essence of right is, therefore, the activity of a person or persons relative to other persons. This

activity, however, can be exercised either by individuals or by persons acting as members of a society.

Rights of individuals

When a person's activity is actuated in a moral way, the object of that activity becomes the person's own, that is, becomes proper to the person in such a way that it cannot be violated without damage to the person in whose ownership it is. Practical experience of this is found in what Rosmini calls 'jural resentment', the injured feeling that occurs on the occasion of violation of some right and gives rise to an instinct for repossession or restitution. Such an experience is obviously not a fact related solely to matter; it is fundamentally a fact of the spirit where alone it can be felt. It is also an indication of the sphere of jural freedom within which a person is and must be left free. More importantly, 'person' does not possess right, but -- because formed by the light of being -- is right itself; does not possess freedom, but is freedom. The divisions of activity give rise to two major kinds of rights in people. If what is possessed is such from the very beginning of existence in a human being (life, for example), the individual possesses innate right(s) which may be called 'natural' (pertaining essentially to human nature) or 'rational' (the rights are what they are and cannot be otherwise). But rights may also be acquired by human beings during the course of life through adventitious activity. When rights have been established in this way, they too are inviolable although there are circumstances, such as lack of use, which dissolve the relationship of ownership and thus leave the field open to others who may wish to extend their activity to the matter in question. Leaving these circumstances aside, however, individual rights cannot be absorbed by others. The State, for example, cannot absorb the inalienable rights proper to persons, nor can it be considered as more than its individual members in such a way that persons can be sacrificed for the sake of society.

Social right

Nevertheless, societies exist within which rights arise from the bonds between intellectual beings. According to Rosmini, these societies fall under three headings: theocratic society, that is, society between God and his creatures; domestic society, which is divided into conjugal and parental society; and civil society, that is, the communion desired by several families who wish to entrust the preservation and the regulation of their rights to a single or collective mind called 'government'. Rosmini considers at some length the rights arising in these societies. In particular, he ponders the title of rights possessed by the Creator over human beings, the rights proper to husband and wife, and to parents. Of special interest is his description of the State as a society which, while it has the duty to influence for the common good only the modality and exercise of rights in its citizens, has no power to create or destroy human rights. In fact, the general purpose of the State is to arrange the exercise of individual rights in such a way that individuals are better able to enjoy the use of their innate and acquired rights. Thus, although in time of war the exercise of certain rights may be curtailed or even suspended, the rights remain invested in individuals to whose exercise they must be restored in normal circumstances. It is clear that Rosmini's view of civil society is completely anti-totalitarian. He does not, however, espouse the cause of modern democracy According to him, the principle of democracy is not 'one person, one vote', but would depend rather upon the contribution made by citizens to the well-being of the State. Difficult to arrange in

practice, the principle is nevertheless important, and can be considered the obverse of ‘No taxation without representation’.

7. The Nature and Purpose of Society

It is clear that Rosmini’s view of rights in human society depends to a great extent on his views about the nature of society as such. In fact, his earliest work as a philosopher (1818-1826) was almost totally taken up with a study of society, and was only abandoned when he saw that his ideas would lack a solid foundation unless the problem of knowledge had first been confronted. Eventually (1837), he published his ideas on society in *The Philosophy of Politics* in which he deals with the principle according to which societies stand or fall, and the end to which societies are directed. . However, despite the universality of the principles examined in this work, their application is restricted to civil society. According to Rosmini, the first rule and criterion for governing any society whatsoever is this: That which constitutes the existence or substance of a society is to be preserved and strengthened even at the cost of neglect to accidental refinement. This is also the first rule of politics. It follows that the greatest errors of government are those by which the government of a society, because of its excessive concern for the society’s accidental progress, loses sight of that which constitutes the substance of the society. The steps taken towards decline, that is, towards the substitution of essential matters by accidental, can be considered on four levels: the periods of founders of societies and basic legislation; of genuine development; of external splendour; of frivolity. Beginning from the first stage, when attention is inevitably fixed on the nature itself of a society, there is a gradual diminution of interest in underlying societal values until weakeness, manifested in attention to frivolities and inability to concentrate on weightier matters, undermines the society’s inner cohesion and its ability to withstand external inimical pressures. This would explain the profound truth lying behind Machiavelli’s observation: ‘If a sect or republic is to survive for any length of time, it must return frequently to its beginning’ (quoted in *PP*, vol. 1, 41)

For Rosmini, every society is simply the union of two or more people undertaken with the intention of obtaining a common advantage. ‘All the persons in this union [forming a society] together have the role of end, and the advantage expected from the association is applied equally to all (*PP*, vol. 2, 39). In other words, there must be in every society a moral element, an element of justice, which affects the behaviour of the members towards one another, even if collectively (as in Plato’s case of a band of robbers) they are unjust towards non-members. Hence the excellence of what Rosmini calls ‘the social bond’; where it is present and actuated, there is no injustice; injustice begins in its absence. In other words, the nature of society requires that those who form it, enjoy within it the personal dignity of end. There is, therefore, a moral element inherent in every society. Using some persons, some members, as means, even for the apparent good of the whole, is repugnant to the very nature of society. In a coherent universal society, such as that described by Cicero -- ‘This entire world is to be considered simply as a city common to both gods and human beings’ (quoted in *PP*, vol. 2, 49) -- there will inevitably be a tendency to maximum justice. This means, however, that no one in any society can make one person subservient to another. All that is pertinent only to nature can be used as means; all that which is proper to person, or to which persons have extended themselves so that some thing or things have become proper to them, must

be respected as end. Here again, though, a distinction must be made between that is proper to a member of a society as a person, and the modality of that which is proper. In certain cases, the government of a society may change the modality of what is proper if this is for the common good. A piece of land required for a road may, for instance, be substituted with an adequate sum of momey. But those things which have no modality, such as innocent life, cannot be violated under the pretext of commmon good. It is clear that these principles must direct the government of every society.

Rosmini also deals with the nature of the good which is the aim or end of society. For him, this good is human good which 'resides in virtue and the eudemenological appurtenances of virtue, and in general in every good in so far as it is connected with virtue.' He concludes, therefore, that every society, in so far as it is contrary to virtue, is illegitimate because its aim is contrary to the essence of society. At the same time, every law of society is invalid if, or in so far as, it prevents members from achieving virtue. 'Without virtue there is no human good, the end for which society is established' (*PP*, vol. 2, 189). But while anti-virtue is essentially detrimental to the good of any society, virtue of itself is not the only element forming the good in question. Contentment of spirit, that is, of the whole person, as distinct from passing pleasure confined to parts of human nature, is also included in every society's essential aim. Anything, therefore, opposed to contentment is inevitably detrimental to society, whatever favour it may have found in public opinion. Rosmini finds support for this affirmation in Hamilton, whom he quotes approvingly (*PP*, vol. 2, 195).

> When occasions present themselves in which the interests of the people are at variance with their inclinations, it is the duty of the persons whom they have appointed to be the guardians of those interests to withstand the temporary delusion in order to give them time and opportunity for more cool and sedate reflection.

Public opinion, it would seem, is not an infallible criterion of public good.

8. Natural theology

Rosmini's natural theology, in the conventional philosophical sense of natural theology, can be summarised under two headings: proofs for the existence of God, found scattered throughout his philosophical writings, and theodicy, developed at length in a work of that name. But it will also be considered, under the heading 'Being' according to the meaning given it by Rosmini.

The existence of God

In his natural theology, Rosmini offers considerations which are consistent with the basic principles of his philosophical teaching. Although he does not deny the validity of *a posteriori* proofs of the existence of God, he affirms that an *a priori* method is more satisfactory because it sets off from the idea of being, the foundation of his philosophy, and argues to the necessity of God's existence. First, however, he posits a fundamental barrier to the perception of God at a purely natural level of human nature. For Rosmini,

what is real can only be perceived through feeling, which indeed is part of human existence but only at a finite level where it cannot be the vehicle of the perception of God’s infinite reality. Ideal being, however, while expressing only the possibility, not the reality of things, is shown on analysis to be characterised by necessity, eternity and immateriality, factors which are intrinsic marks of possibility. The idea of being, therefore, with its divine characteristics serves as a bridge between God and human beings, enabling us to posit proofs of God’s existence without our knowing him through perception, or real contact. One example of the *a priori* proofs may be given here. If infinite, intelligible being is present to the human mind -- and Rosmini would maintain that it is -- an infinite mind capable of giving this idea to humans must exist. But such a mind cannot not be God. In this and all similar proofs, there is a common mode of procedure. The existence of God is necessary for the existence of intelligible being; but intelligible being certainly exists; the existence of God is therefore necessary. Having determined the existence of God, Rosmini then makes use of divinely revealed truth as a basis on which to offer some reflections about the nature of what he has proved to exist. These reflections, he would maintain, lie within the scope of philosophy because they depend methodologically solely on reason, and conclude with rational, not authoritative affirmations. Their basis in revelation, however, precludes any examination of them here.

Theodicy

Rosmini provides an exact description of his aim in his work on theodicy. ‘Theodicy (theou dike) means "justice of God". The intent of this work, therefore, is simply this: to justify God’s equity and goodness in the distribution of good and evil in the world.’ (*Theodicy* preface).

He also provides a clear indication of the method he will follow in the three books composing his *Theodicy*. The first prescribes the norms to be followed in judging about the disposition of divine providence if error is to be avoided; the second considers the laws of nature, the necessary limitations of what is created and the chain of causes operating in the universe; the third is devoted to the laws according to which God’s action takes place in the world he has created. These laws all spring from a single norm, the law of the least means, which Rosmini, following Aquinas, posits as follows: ‘The wise worker carries out his work in the briefest way possible’ (*Sapiens operator perficit opus suum breviori via qua potest*. *ST*, III, q. 4, art. 5, ad 3um). From this law, another ten are deduced whose titles provide a useful indication of their possible application. They are the laws of: excluded superfluity, the permission of evil, excluded equality, unity in the divine work, heroism, antagonism, rapidity in work, accumulation of good, and germ. These laws, according to Rosmini, are concerned with God’s providence relative to universal good. At the same time, but relative to particular good, the law followed is that of: supreme justice, equity, fittingness and conformity to God’s divine attributes. Divine governance consists in acting in such a way that the aspects of universal and particular good are harmonised.

9. Being

The last nine years of Rosmini’s life, with the exception of 1848-9 when he was actively engaged in

political affairs, were spent in great part working on his *Theosophy* (*Teosofia*). Despite its size (five volumes), this monumental work remained unfinished at Rosmini's death and was published posthumously. The title itself, despite its obvious etymological implications, is today unhelpful. It would seem to indicate a philosophy 'professing to achieve a knowledge of God by spiritual ecstasy, direct intuition or special individual relations' (*COD*). Rosmini, however, is concerned here only with reasoning about God, and takes the word 'theosophy' in its fundamental meaning of 'wisdom about God' in so far as God is the supreme Being and the apex of philosophical speculation. Such speculation must sooner or later deal with the problem of the One Being and many beings, with unity and plurality in all their manifestations. In many ways, the book is a direct challenge to Hegel and Schelling with whose philosophy Rosmini had become thoroughly familiar as a result of the great editions of their works published in the 1830s and early 1840s. Kant's teaching, already challenged in Rosmini's *A New Essay concerning the Origin of Ideas*, was now to be opposed in what Rosmini perceived as the delirium and exaggerations of Kant's Idealist successors.

> Whenever we try to reduce the whole human being to speculation, and substitute the part for the whole, we presume that all human good must lie solely in speculation. As result, we make every effort to turn what is real into an idea; we try to derive from the idea the matter which constitutes the sensible world, together with the Spirit and finally God himself… But although Schelling and Hegel claimed that they had reached such total science, they still needed to teach it publicly not only for the sake of attaining the practice of virtue (which would make it worthwhile), but even to draw a salary. This proves without doubt that their absolute idea did not contain everything. If the world were present in it, as they said it was, wheat, bread and wine would have been there also. My theosophy certainly cannot give the public such magnificent and wonderful promises, but it will explain how the speculative human mind is inclined to find everything in itself. In other words, it will demonstrate that there must be an object which contains effectively within itself the universality of things, and that this object is not the idea in our mind. Nevertheless, the idea which shines in the human mind draws its form as object from that object. Hence, because the idea also is *per se* object, we easily confuse it, in our speculation, with the complete, subsistent object. A strong desire then arises in us of attributing to the idea which we intuit the attributes of the subsistent object which we know must exist, although we do not intuit it. The tendency to unity, an essential element of every intellect, causes this error and forces us towards an abyss of unseen absurdities in the hope that these will satisfy our desperate purpose. We should acknowledge (and this theosophy will demonstrate) that if being itself has an objective existence, it is *per se* intelligible, and that if it contains everything (that which is not being is nothing), everything must also be contained in that which is intelligible. Theosophy will also clearly show that, although being must actually have this primal form, human nature cannot intuit *the intelligible which contains all*. Human nature arrives at this solely by reasoning, which can provide only a formal, negative concept of it. We cannot therefore have either the absolute knowledge which Schelling attributes to us through direct intuition, or the absolute idea which his disciple, George Hegel (who was opposed to all immediacy), promised us by mediate reasoning… (*TH*, vol. 1, Preface, 9-10)

But the main thrust of the work is positive, rather than critical.

> Theosophy will not be wasting time by demonstrating that behind this well-conceived and ingenious error [of Hegel] lies a great truth which those courageous speculative minds [Schelling and Hegel] tried in vain to grasp but could not. This truth is precisely the necessity I spoke of: there must be ‘something intelligible and eternal which contains everything’... Nevertheless, although absolute knowledge is proper to God but not to us, we do have an absolute knowledge relative to form, but not to matter (cf. *NE*, vol. 1, 325, 474-476). This kind of absoluteness of human knowledge caused errors in the German school, which I have already discussed. Theosophy must speak at length about absolute human knowledge, indeed it must use it and more importantly be it. Theosophy is simply the Theory of Ens (this definition is not to be despised, despite its being only two words). Because ens is first of all infinite and absolute, and only later enclosed and existing within limits as finite, no thought could attain it unless thought itself somehow became absolute. A thought informed by an object which is in some way absolute, is itself in some way made absolute. Plato therefore rightly called the treatise on what is greatest, the treatise about ens. There is nothing in the universe or in our mind antecedent to ens or being. When, in the order of things, we remove being, nothing remains except darkness in the order of cognition. For this reason the doctrine of ens, which I call ‘theosophy’, corresponds to the concept of philosophy in the ancients. According to them, philosophy differs from other sciences in that all other sciences suppose undemonstrated principles. Philosophy, however, which borrows nothing from anywhere, uses its own materials to construct itself. It starts from no gratuitous hypothesis or supposition -- on the contrary, it seeks and establishes what is undemonstrable, which gives it an unshakeable basis, and admits only what is necessary (*Ibid*., 10-12).

For Rosmini, the problem of being is finally considered under the three divisions of the science of theosophy: ontology, rational theology and cosmology, each of which must enter into the other if the science is to be complete. In fact, he maintains, it is impossible to speak of being in all its universal essence (ontology), without regard to the infinity and absoluteness of Being (rational theology), just as it is impossible to consider the world philosophically (cosmology) without taking its cause into consideration. Each of the three divisions of theosophy is as essential to the whole as, according to his example, the various vital organs are necessary to the existence of an animal. But the centre and substance of the whole treatise is teaching about God, without which there is no final explanation of the being or the world. Theosophy, therefore, is a single science which, through its division into three parts, is both one and three. Such an affirmation is paradigmatic of the thrust of Rosmini’s theosophy which eventually enables him to confront this rational science with the doctrine of the Trinity.

Ontology

Ontology, according to Rosmini’s description of it, considers being in all its universality, but only as the object of human thought which puts limited, intellectual beings in contact with the possibility of all that

is, not with actuality (it is impossible, for instance, to assuage hunger by thinking about a meal). Thought must, therefore, reach out to absolute reality not through direct perception of God, but by means of concepts. Ontology, dealing with these concepts analytically and synthetically is, as it were, an immense preface to rational theology. Ontology as the theory of being in all its possibility is the necessary propedeutic to theology as the theory of absolute being without which ontology itself is inevitably incomplete, and cannot progress beyond a treatise on categories and on dialectic.

Rational theology

If, as Rosmini suggests, ontology is the ‘theory of abstract being’, rational theology is the ‘theory of subsistent Being’. These two parts of theosophy cannot, however, can be distinguished only if the same universal concepts, of which both are formed, can in some way be differentiated. This is done by considering that ontology serves to find, review, and describe the nature and relationship of these concepts; theology, in Rosmini’s understanding, synthesises them to form a single concept of the infinite Being. It is true that such a concept remains abstract (the essence contained in the concept is not beheld by the knowing subject), but the ‘theologian’ passes from the concept to affirm subsistent Being, which is therefore no longer a mere concept, but something which must exist in itself. An effort is then made in rational theology to see how the ontological concepts have their truth and their foundation in first subsistent Being. While this takes place, the concepts themselves are identified and become one. Their separation, now considered anew, is seen to be relative not to first subsistent Being, but to the beholding, finite mind. The problem of the One and the many is rooted not in the nature of absolute, subsistent Being, but in being by participation.

Cosmology

Ontology, while a preface to rational theology, is also necessary for knowledge of the intimate nature of finite being and the world. Rosmini posits three reasons for this. First, although finite being is perceived, perception is limited to very few beings. If the essential conditions of this being are to be known, it is necessary to know the common conditions of all finite beings. This cannot be done without recourse to universal principles which enable us to deduce what is lacking to limited experience. Second, the same principles are needed to provide a notion of finite beings in themselves, divested of the sensory phenomena that accompany their perception. Third, subsistent Being, if known, would be intelligible *per se*. This is not the case with finite beings, which are known through their participation in intellectual light, and consequently remain unknown without some relationship to concepts.

Bibliography

References

References in this article to Rosmini’s works use the following abbreviations:

NE = *A New Essay concerning the Origin of Ideas*
IP = *Introduction to Philosophy*
PR = *The Philosophy of Right*
PP = *The Philosophy of Politics*
TH = *Theosophy*

Other abbreviations:

Life = *Antonio Rosmini: Introduction to his Life and Teaching*
CFD = *Congregation for the Doctrine of the Faith*
COD = *Concise Oxford Dictionary*

Critical Edition

- *Opere edite e inedite di Antonio Rosmini*. Rome-Stresa. 1966-. Città Nuova Editrice (80 projected volumes -- including 19 of correspondence -- of which approximately 40 have been published [2001])

Other Editions

- For a complete list of editions of all Rosmini's published works in Italian or in translations, cf. Bergamaschi, Cirillo.*Bibliografia degli scritti editi di Antonio Rosmini Serbati*, 1815-1998, 4 vols., Milan-Stresa, 1970-98.

Translations

- Cleary, D., Watson, T., and Murphy, R., (2001). *A New Essay concerning the Origin of Ideas*. 3 vols. Durham: Rosmini House, 2001
- Cleary, D., and Watson, T., (1988-). Durham: Rosmini House.
  - *Principles of Ethics*. 1988.
  - *Conscience*. 1989.
  - *Anthropology as an Aid to Moral Science*. 1991.
  - *The Philosophy of Politics*.

    vol. 1. *The Summary Cause for the Stability or Downfall of Human Societies*;

    vol. 2. *Society and Its Purpose*. 1994.
  - *The Philosophy of Right*.

    vol. 1. *The Essence of Right*. 1993.

vol. 2. *Rights of the Individual*. 1993.

vol. 3. *Universal Social Right*. 1995.

vol. 4. *Rights in God's Church*. 1995.

vol. 5. *Rights in the Family*. 1995.

vol. 6. *Rights in Civil Society*. 1996.

- *Psychology*. 1999.

vol. 1. *Essence of the Human Soul*.

vol. 2. *Development of the Human Soul*.

vol. 3. *Laws of Animality*.

vol. 4. *Opinions about the Human Soul*.

- *Theosophy* (5 vols. projected).

vol. 1. forthcoming

- Grey, M., 1887, *The ruling Principle of Method applied to Education*. Boston.
- Murphy, R., *Introduction to Philosophy*, forthcoming
- Signini, F., 1912, *Theodicy*. 3 vols. London

Secondary Works

- Bergamaschi, Cirillo.*Grande dizionario antologico del pensiero di Antonio Rosmini*. (4 vols of explanations of words and phrases from Rosmini's works in Rosmini's own words.) Rome. Città Nuova Editrice. 2001. CD version available.
- Bergamaschi, Cirillo. *Bibliografia Rosminiana*. 8 vols. Milan-Stresa, 1967-96 (covers Rosminian bibliography in all languages from 1814-1995).
- Cleary, Denis. *Antonio Rosmini: Introduction to his Life and Teaching*. Durham: Rosmini House, 1992 (2nd edition forthcoming).
- Davidson, Thomas. *The Philosophical System of Antonio Rosmini Serbati*, London, 1882.
- Pozzo, Riccardo. *The Philosophical Works of Antonio Rosmini in Translation* in *American Catholic Philosophical Quarterly*, LXXIII (1999), no. 4

Other Internet Resources

- Rosmini in English (Rosmini House, Durham, UK)

Related Entries

Aquinas, Saint Thomas | Augustine, Saint | Hegel, Georg Wilhelm Friedrich | Kant, Immanuel | Locke, John | Plato | Schelling, Friedrich Wilhelm Joseph von

Denis Cleary
deniscleary@rosmini-in-english.org

First published: December 28, 2001
Content last modified: December 28, 2001

Denis Anthony Cleary
Curriculum vitae.

Address:

Rosmini House
Woodbine Road
Durham, DH1 5DR, UK

Education

1951-54. Philosophy studies at the Pontifical Lateran University, Rome. Licentiate in Philosophy, 1954

1954-58. Theology studies at the Pontifical Lateran University, Rome. Language studies in Arabic and Hebrew. Licentiate in Theology, 1958.

1958-59. Philosophy studies at the Pontifical Gregorian University, Rome.

1960. Ph. D at the Pontifical Gregorian University, Rome (Thesis: Sensation in Bertrand Russell).

1959-1966. Adminstrator and lecturer in philosophy at the Rosminian house of studies, Wonersh, UK

1966-1973. Rector of the Rosminian house of studies, Wonersh, UK, and lecturer in philosophy at St. Johns Regional Seminary, Wonersh, UK

1973-1979. Rector of St. John at the Latin Gate, Rome, Italy, and of the Collegio Missionario Antonio Rosmini, Rome, Italy. Lecturer in philosoophy and theology at the Pontifical Beda College, Rome.

1979-1983. Lecturer in church history and Old Testament Scripture at Kibosho Regional Seminary, Moshi, Tanzania.

1983-1985. Resident at the Centro Internazionale di Studi Rosminiani, Stresa, Italy. Visitng lecturer in philosophy (*pour la recyclage benedictine)* at St. Anselms University, Rome, Italy.

1985- Director of Rosmini House, Durham, UK (Centre for translation into English of Rosminis works. For further information on this project, see www.rosmini-in-english.org and the feature review, The Philosophical Works of Antonio Rosmini in English Translation, Riccardo Pozzo, American Catholic Philosophical Quarterly, Vol. LXXIII, No. 4, pp. 609-637).

1993-98. Lecturer in philosophy at Ampleforth Abbey, UK.

1998. Speaking tour, with others, on aspects of Rosminis work, at American Universities (see An Introduction to the Thought of Antonio Rosmini, ed. Bernard A. Cook, Loyala University New Orleans 2000)

Publications

Books:

Antonio Rosmini: Introduction to his life and teaching. Durham 1993, pp 80.

Various Articles published in USA

Antonio Rosmini, "Ontologism", *New Catholic Encyclopedia*

Antonio Rosmini, Vincenzo Gioberti, in *Biographical Dictionary of Christian Theologians*, ed. Carey and Lienhard, Westport 2000

Rosmini on Natural Law and Right (*Vera Lex*, vol. XIII, numbers 1 & 2, 1993, pp. 6-12)

Translations of Rosmini's Work

Le Cinque piaghe della santa chiesa (The Five Wounds of the Church, Leominster 1987, pp. 256); Constitutiones Societatis a Caritate nuncuptae (Constitutions of the Society of Charity, Durham 1988, pp. 502);

Translations (with Terence P. Watson) of Rosmini's Work

Principi della scienza morale (Principles of Ethics, Durham 1989, pp. 111)

Nuovo Saggio sullorigine delle idee (vol 2, The Origin of Ideas, Durham 1987, pp. 335; vol. 3, Certainty, Durham 1993, pp. 346);

Tratttato della coscienza morale (Conscience, Durham 1989, pp. 438);

Filosofia della politica (vol. 1, The Summary Cause for the Stability or Downfall of Human Societies, Durham 1994, pp. 96; vol. 2, Society and its Purpose, Durham 1995, pp. 455)

Filosofia del diritto (vol. 1, The Essence of Right, Durham 1993, pp. 216; vol. 2, Rights of the Individual, Durham 1993, pp. 596; vol. 3, Universal Social Right, Durham 1995, pp. 144; vol. 4, Rights in Gods

Church, Durham 1995, pp. 176; vol. 5, Rights in the Family, Durham 1995, pp. 248; vol. 6, Rights in Civil Society, Durham 1996, pp. 487);

Psicologia (vol. 1, Essence of the Human soul, Durham 1999, pp. 392; vol. 2, Development of the Human soul, Durham 1999, pp. 560; vol. 3, Laws of Animality, Durham 1999, pp. 263; vol. 4, Opinions about the Human Soul, Durham 1999, pp. 160)

Work in progress on translations of Rosmini's Work

Il Linguaggio teologico, Introduzione alla filosofia

Affirmative Action

"Affirmative action" means positive steps taken to increase the representation of women and minorities in areas of employment, education, and business from which they have been historically excluded. When those steps involve *preferential* selection -- selection on the basis of race, gender, or ethnicity -- affirmative action generates intense controversy.

The development, defense, and contestation of preferential affirmative action has proceeded in two streams. One has been legal and administrative, as courts, legislatures, and executive departments of government have applied laws and rules requiring affirmative action. The other has been public debate, where the practice of preferential treatment has spawned a vast literature, pro and con. Often enough, the two streams have failed to make adequate contact, with the public quarrels not always very securely anchored in any existing legal basis or practice.

The ebb and flow of public controversy over affirmative action can be pictured as two spikes on a line, the first spike representing a period of passionate debate that began around 1972 and tapered off after 1980, and the second indicating a resurgence of debate in the 1990s. The first spike encompassed controversy about gender and racial preferences alike. This is because in the beginning, affirmative action was as much about the factory, firehouse, and corporate suite as about the university campus. The second spike represents a quarrel about race. This is because the only burning issue now is about preferential admissions in higher education.[1]

1. In the Beginning

In 1972, affirmative action became an inflammatory public issue. True enough, the Civil Rights Act of 1964 had made something called "affirmative action" a remedy federal courts could impose on violators of the Act. Likewise, since 1965, federal contractors had been subject to President Lyndon Johnson's Executive Order 11246, requiring them to take "affirmative action" to make sure they were not discriminating. But what did this 1965 mandate amount to? The Executive Order assigned to the Secretary of Labor the job of specifying rules of implementation. In the meantime, as the federal courts were enforcing the Civil Rights Act against discriminating companies, unions, and other institutions, the Department of Labor mounted an ad hoc attack on the construction industry, cajoling, threatening, negotiating, and generally strong-arming reluctant construction firms into a series of region-wide "plans," in which they committed themselves to numerical hiring goals. Through these contractor commitments, the Department could indirectly pressure recalcitrant labor unions, who supplied the employees at job sites.

While the occasional court case and government initiative made the news and stirred some controversy, affirmative action was pretty far down the list of public excitements until the autumn of 1972, when the Secretary of Labor's Revised Order No. 4, fully implementing the Executive Order, landed on campus by way of directives from the Department of Health, Education, and Welfare. Its predecessor, Order No. 4, first promulgated in 1970, cast a wide net over American institutions, both public and private. By extending to all contractors the basic apparatus of the construction industry "plans," the Order imposed a one-size-fits-all system of "underutilization analyses," "goals," and "timetables" on hospitals, banks, trucking companies, steel mills, printers, aircraft manufacturers -- indeed, on all the scores of thousands of institutions, large and small, that did business with the government -- including a special set of institutions with a particularly voluble and articulate constituency, namely, American universities.

At first, university administrators and faculty found the new rules murky but hardly a threat to the established order. The number of racial and ethnic minorities receiving PhDs each year was tiny. Any mandate to increase their representation on faculties would require more diligent searches by universities, to be sure, but searches nevertheless fated largely to mirror past results. The Revised Order, on the other hand, effected a change that punctured any campus complacency: it included women among the "protected classes" whose "underutilization" demanded the setting of "goals" and "timetables" for their "full utilization."[2] Unlike blacks and Hispanics, women were getting PhDs in substantial and growing numbers. If the affirmative action required of federal contractors was a recipe for "proportional representation," then Revised Order No. 4 was bound to leave a large footprint on campus. Some among the professoriate exploded in a fury of opposition to the new rules, while others responded with an equally vehement defense of them.[3]

As it happened, these events coincided with another development, namely the "public turn" in philosophy. For several decades Anglo-American philosophy had treated moral and political questions obliquely. On the prevailing view, philosophers were competent to do "conceptual analysis" -- to lay bare, for example,

the conceptual architecture of the idea of justice -- but they were not competent to suggest political principles, constitutional arrangements, or social policies that actually did justice. Philosophers might do "meta-ethics" but not "normative ethics." This viewed collapsed in the 1970s under the weight of two counter-blows. First, John Rawls published in 1971 *A Theory of Justice*,[4] an elaborate, elegant, and inspiring defense of a *normative* theory of justice. Second, in the same year *Philosophy & Public Affairs*, with Princeton University's impeccable pedigree, began life, a few months after Florida State's *Social Theory and Practice*. These journals, along with a re-tooled older periodical, *Ethics*, became self-conscious platforms for socially and politically engaged philosophical writing, born out of the feeling that in time of war (the Vietnam War) and social tumult (the Civil Rights Movement, women's liberation), philosophers ought to do, not simply talk about, ethics. In 1973, *Philosophy & Public Affairs* published Thomas Nagel's "Equal Treatment and Compensatory Justice"[5] and Judith Jarvis Thomson's "Preferential Hiring,"[6] and the philosophical literature on affirmative action burgeoned forth.[7]

In contention was the nature of those "goals" and "timetables" imposed on every contractor by Revised Order No. 4. Weren't the "goals" tantamount to "quotas," requiring institutions to use racial or gender preferences in their selection processes? Some answered "no."[8] Properly understood, affirmative action did not require (or even permit) the use of gender or racial preferences. Others said "yes."[9] Affirmative action, if it did not impose preferences outright, at least countenanced them. Among the yea-sayers, opinion divided between those who said preferences were morally indefensible and those who said they were not. Within this last set, different people put forward different justifications.

2. The Controversy Engaged

The essays by Thomson and Nagel both defended the use of preferences but on different grounds. Thomson defended job preferences for women and blacks as a form of compensation for their past exclusion from the academy and the workplace. Preferential policies, in her view, worked a kind of justice. Nagel, by contrast, thought that preferences might work a kind of social good, and without doing violence to justice. Institutions could for one or another good reason properly depart from standard meritocratic selection criteria because the whole system of tying economic reward to earned credentials was itself indefensible.

Justice and desert lay at the heart of subsequent arguments. Several writers took to task Thomson's argument that preferential hiring justifiably makes up for past wrongs. Preferential hiring seen as compensation looks perverse, they contended, since it benefits individuals (blacks and women possessing good educational credentials) least likely harmed by past wrongs while it burdens individuals (younger white male applicants) least likely to be responsible for past wrongs.[10] Instead of doing justice, preferential treatment violates rights (the right of an applicant "to equal consideration,"[11] the right of the maximally competent to a position,[12] the right of everyone to equal opportunity[13]) and confounds desert (by severing reward from a "person's character, talents, choices and abilities;"[14] by "subordinating merit, conduct, and character to race;"[15] by disconnecting outcomes from actual liability and damage[16]).

Defenders of preferences were no less quick to enlist justice and desert in their cause. Mary Anne Warren, for example, argued that in a context of entrenched gender discrimination, gender preferences might improve the "overall fairness" of job selections. Justice and individual desert need not be violated.

> If individual men's careers are temporarily set back because of . . . [job preferences given to women], the odds are good that these same men will have benefited in the past and/or will benefit in the future -- not necessarily in the job competition, but in some ways -- from sexist discrimination against women. Conversely, if individual women receive apparently unearned bonuses [through preferential selection], it is highly likely that these same women will have suffered in the past and/or will suffer in the future from . . . sexist attitudes.[17]

Likewise, James Rachels defended racial preferences as devices to neutralize unearned advantages by whites. Given the pervasiveness of racial discrimination, it is likely, he argued, that the superior credentials offered by white applicants do not reflect their greater effort, desert, or even ability. Rather, the credentials reflect their mere luck at being born white. "Some white . . . [applicants] have better qualifications . . . only because they have not had to contend with the obstacles faced by their black competitors."[18] Rachels was less confident than Warren that preferences worked uniformly accurate offsets. Reverse discrimination might do injustice to some whites; yet its absence would result in injustices to blacks who have been unfairly handicapped by their lesser advantages.

Rachels' diffidence was warranted in light of the counter-responses. If racial and gender preferences for jobs (or college admissions) were supposed to neutralize unfair competitive advantages, they needed to be calibrated to fit the variety of backgrounds aspirants brought to any competition for these goods. Simply giving blanket preferences to blacks or women seemed a much too ham-handed approach if the point was to micro-distribute opportunities fairly.[19]

3. Rights and Consistency

To many of its critics, reverse discrimination was simply incoherent. When "the employers and the schools *favor* women and blacks," objected Lisa Newton, they commit the same injustice perpetrated by Jim Crow discrimination. "Just as the previous discrimination did, this reverse discrimination violates the public equality which defines citizenship."[20]

William Bennett and Terry Eastland likewise saw racial preferences as in some sense illogical:

> To count by race, to use the means of numerical equality to achieve the end of moral equality, is counterproductive, for to count by race is to deny the end by virtue of the means. The means of race counting will not, cannot, issue in an end where race does not matter.[21]

When Eastland and Bennett alluded to those who favored using race to get to a point where race doesn't

count, they had in mind specifically the Supreme Court's Justice Blackmun who, in the famous 1978 *Bakke* case (discussed below), put his own views in just those simple terms. The legitimacy of racial preferences was to be measured by how fast using them moved us toward a society where race doesn't matter (a view developed in subtle detail by the philosopher Richard Wasserstrom).[22] While the critics of preferences feigned to find the very idea of using race to end racism illogical and incoherent, they also fell back on principle to block this instrumental defense should it actually prove both reasonable and plausible. "The moral issue comes in classic form," wrote Carl Cohen. "Terribly important objectives . . . appear to require impermissible means. Might we not wink at the Constitution this once" and allow preferences to do their good work?[23] Neither Cohen nor the other critics thought so. Principle must hold firm. "[I]n the distribution of benefits under the laws *all* racial classifications are invidious."[24]

But what, exactly, *is* the principle -- Constitutional or moral -- that bars the use of race as a means to "terribly important objectives"? Alan Goldman did more than anyone in the early debate to formulate and ground a relevant principle. Using a contractualist framework, he surmised that rational contractors would choose a rule of justice requiring positions to be awarded by competence. They would choose this rule because it instantiates a principle of equal opportunity which in turn instantiates a broad right to equal consideration of interests, this last principle springing from the basic condition of the contracting parties as rational, self-interested, and equally situated choosers. On its face, the rule of competence would seem to preclude filling positions by reference to factors like race and gender that are unrelated to competence. However, Goldman's "rule" blocked preferences only under certain empirical conditions. Goldman explained the derivation of the rule and its consequent limit this way:

> The rule for hiring the most competent was justified as part of a right to equal opportunity to succeed through socially productive effort, and on grounds of increased welfare for all members of society. Since it is justified in relation to a right to equal opportunity, and since the application of the rule may simply compound injustices when opportunities are unequal elsewhere in the system, the creation of more equal opportunities takes precedence when in conflict with the rule for awarding positions. Thus short-run violations of the rule are justified to create a more just distribution of benefits by applying the rule itself in future years.[25]

In other words, if "terribly important objectives" -- especially objectives having to do with equalizing opportunities in a system rife with inequality -- could in fact be furthered by measured and targeted reverse discrimination, justice wouldn't stand in the way. Thus, Goldman's rule did not have the adamantine character Cohen and other critics sought in a bar to preferences. Where can such an unyielding principle be found? I postpone further examination of this question until I discuss the *Bakke* case, below, whose split opinions constitute an extended debate on the meaning of Constitutional equality.

4. Real-World Affirmative Action -- The Workplace

The terms of the popular debate over racial and gender preferences mirrored the arguments philosophers and other academics were making to each other. Preference's defenders offered many reasons to justify

them, reasons having to do with compensatory or distributive justice, as well as reasons having to do with social utility (more blacks in the police department would enable it better to serve the community, more female professors in the classroom would inspire young women to greater achievements). Critics of preferences retorted by pointing to the law. And well they should, since the text of the Civil Rights Act of 1964 seemed a solid anchor even if general principle proved elusive. Title VI of the Act promised that "[n]o person . . . shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance."[26] Title VII prohibited all employment practices that discriminated on the basis of race, gender, religion, or national origin.[27] However, unlike Title VI, Title VII went on to spell out some exceptions. Under special circumstances, the Title permitted the use of gender, religion, and national origin as legitimate bases for employer selection. But it made no such exception for race. While being a woman or being a Roman Catholic could sometimes count as a legitimate occupational qualification, being black could not.

In face of the plain language of Titles VI and VII, how did preferential hiring and promotion ever arise in the first place? How could it be justified legally? The answer lay in the meaning of "discrimination." The Civil Rights Act did not define the term. The federal courts had to do that job themselves, and the cases before them drove the definition in a particular direction. Many factories and businesses prior to 1964, especially in the South, had in place facially discriminatory policies and rules. For example, a company's policy might have relegated blacks to the maintenance department and channeled whites into operations, sales, and management departments, where the pay and opportunities for advancement were far better. After passage of the Civil Rights Act, suppose the company willingly abandoned its facially segregative policy. Yet it could still carry forward the effects of its past segregation through other already-existing facially neutral rules. For example, a rule requiring workers to give up their seniority in one department if they transferred to another would have locked in place older black maintenance workers as effectively as the prior segregative rule that made them ineligible to transfer at all. Consequently, courts began striking down facially neutral rules that carried through the effects of an employer's past discrimination, regardless of the original intent or provenance of the rules. "Intent" was effectively decoupled from "discrimination." In 1971, the Supreme Court ratified this process, giving in the *Griggs* decision the following construction of Title VII:

> The objective of Congress in the enactment of Title VII . . . was to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees. Under the Act, practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they operate to "freeze" the status quo of prior discriminatory employment practices.
>
> ...
>
> What is required by Congress is the removal of artificial, arbitrary, and unnecessary barriers to employment when the barriers operate invidiously to exclude on the basis of racial or other impermissible classification.[28]

In a few short paragraphs the Court moved from proscribing practices that froze in place the effects of a firm's own past discrimination to proscribing practices that carried through the effects of past discrimination generally. The Court characterized statutory discrimination as *any exclusionary practice not necessary to an institution's activities*. Since many practices in most institutions were likely to be exclusionary, rejecting minorities and women in greater proportion than white men, all institutions needed to reassess the full range of their practices to look for, and correct, discriminatory effect. Against this backdrop, the generic idea of affirmative action took form:

> Each institution should effectively monitor its practices for exclusionary effect and revise those that cannot be defended as "necessary" to doing business. In order to make its monitoring and revising effective, an institution ought to predict, as best it can, how many minorities and women it would select over time, were it successfully nondiscriminating. These predictions constitute the institution's affirmative action "goals," and failure to meet the goals signals to the institution (and to the government) that it needs to revisit its efforts at eliminating exclusionary practices. There may still remain practices that ought to be modified or eliminated.[29]

The point of such affirmative action: induce change in institutions so that they could comply with the nondiscrimination mandate of the Civil Rights Act.

However, what if this self-monitoring and revising fell short? In early litigation under the Civil Rights Act, courts concluded that some institutions, because of their past exclusionary histories and present failure to find qualified women or minorities, needed stronger medicine. Courts ordered these institutions to adopt "quotas," to take in specific numbers of formerly excluded groups, on the assumption that once these new workers were securely lodged in place, the institutions would adapt to this new reality.[30]

Throughout the 1970s, courts and government enforcement agencies extended this idea across the board, requiring a wide range of firms and organizations -- from AT&T to the Alabama Highway Patrol -- temporarily to select by the numbers. In all these cases, the use of preferences was tied to a single purpose: to prevent ongoing and future discrimination. Courts carved out this justification for preferences not through caprice but through necessity. They found themselves confronted with a practical dilemma that Congress had never envisaged and thus never addressed when it wrote the Act. The dilemma was this: courts could impose racial preferences to change foot-dragging or inept defendants (thus transgressing the text of Title VII) or they could order less onerous steps they knew to be ineffective, thus letting discrimination continue (and violating their duty under Title VII). Reasonably enough, the federal courts resolved this dilemma by appeal to the broad purposes of the Civil Rights Act and justified racial preferences where needed to prevent ongoing and future discrimination.[31]

Thus, preferential affirmative action served the same rationale as the non-preferential sort. Its purpose was *not* to compensate for past wrongs, offset unfair advantage, appropriately reward the deserving and undeserving, or yield a variety of social goods; its purpose was to change institutions so they could

comply with the nondiscrimination mandate of the Civil Rights Act.

5. Real-World Affirmative Action -- The University

In the 1970s, while campuses were embroiled in debate about how to increase blacks and women on the faculty, universities were also putting into effect schemes to increase minority presence within the student body. Very selective universities, in particular, needed new initiatives because only a relative handful of black and Hispanic high school students possessed test scores and grades good enough to make them eligible for admission. These institutions faced a choice: retain their admissions criteria unchanged and accept the upshot -- hardly any blacks and Hispanics on campus -- or fiddle with their criteria to get a more substantial representation. Most elected the second path.

The Medical School of the University of California at Davis was typical. It reserved sixteen of the one hundred slots in its entering classes for minorities. In 1973 and again in 1974, Allan Bakke, a white applicant, was denied admission although his test scores and grades were better than most or all of those admitted through the special program. He sued. In 1977, his case, *Regents of the University of California v. Bakke*, reached the Supreme Court. The Court rendered its decision a year later.[32]

An attentive reader of Title VI of the Civil Rights Act might have thought this case was an easy call. So, too, thought four justices on the Supreme Court, who voted to order Bakke admitted to the Medical School. Led by Justice Stevens, they saw the racially segregated, two-track scheme at the Medical School, which was a recipient of federal funds, as a clear violation of the plain language of the Title.

Four other members of the Court, led by Justice Brennan, wanted very keenly to save the Medical School program. To find a more attractive terrain for doing battle, they made an end-run around Title VI, arguing that, whatever its language, it had no independent meaning itself. It meant in regard to race only what the Constitution meant.[33] Thus, instead of having to parse the stingy and unyielding language of Title VI ("no person shall be subjected to discrimination . . on the ground of race"), the Brennan group could turn their creative energies to interpreting the broad and vague language of the Fourteenth Amendment ("no person shall be denied the equal protection of the laws"), which provided much more wiggle-room for justifying racial preferences. The Brennan group persuaded one other member, Justice Powell, to join them in their view of Title VI. But Powell didn't agree with their view of the Constitution. He argued that the Medical School's policy was unconstitutional and voted that Bakke must be admitted. His vote, added to the four votes of the Stevens group, meant that Allan Bakke won his case and that Powell got to write the opinion of the Court. The Brennan strategy didn't reap the fruit it intended.

Against the leanings of the Brennan group, who would distinguish between "benign" and "malign" uses of race and deal leniently with the former, Powell insisted that the Fourteenth Amendment's promise of "equal protection of the law" must mean the same thing for all, black and white alike. To paraphrase Powell:

> The Constitution can tolerate no "two-class" theory of equal protection. There is no

> principled basis for deciding between classes that deserve special judicial attention and those that don't. To think otherwise would involve the Court in making all kinds of "political" decisions it is not competent to make. In expounding the Constitution, the Court's role is to discern "principles sufficiently absolute to give them roots throughout the community and continuity over significant periods of time, and to lift them above the pragmatic political judgments of a particular time and place . . . ."[34]

What, then, was the practical meaning of a "sufficiently absolute" rendering of the principle of equal protection? It was this: when the decisions of state agents "touch upon an individual's race or ethnic background, he is entitled to a judicial determination that the burden he is asked to bear on that basis is precisely tailored to serve a compelling governmental interest."[35]

Powell, with this standard in hand, then turned to look at the four reasons the Medical School offered for its special program: (i) to reduce "the historic deficit of traditionally disfavored minorities in medical schools and the medical profession;" (ii) to counter "the effects of societal discrimination;" (iii) to increase "the number of physicians who will practice in communities currently underserved;" and (iv) to obtain "the educational benefits that flow from an ethnically diverse student body."[36] Did any or all of them specify a goal precisely tailored to serve a compelling governmental interest?

As to the first reason, Powell dismissed it out of hand.

> If [the School's] purpose is to assure within its student body some specified percentage of a particular group merely because of its race or ethnic origin, such a preferential purpose must be rejected not as insubstantial but as facially invalid. Preferring members of any one group for no reason other than race or ethnic origin is discrimination for its own sake.

As to the second reason, Powell allowed it more force. A state has a legitimate interest in ameliorating the effects of past discrimination. Even so, contended Powell, the Court

> has never approved a classification that aids persons perceived as members of relatively victimized groups at the expense of other innocent individuals in the absence of judicial, legislative, or administrative findings of constitutional or statutory violations.[37]

And the Medical School

> does not purport to have made, and is in no position to make, such findings. Its broad mission is education, not the formulation of any legislative policy or the adjudication of particular claims of illegality. . . . [I]solated segments of our vast governmental structures are not competent to make those decisions, at least in the absence of legislative mandates and legislatively determined criteria.[38]

As to the third reason, Powell found it, too, insufficient. The Medical School provided no evidence that

the best way it could contribute increased medical services to underserved communities was by employing a racially preferential admissions scheme. Indeed, the Medical School provided no evidence that its scheme would result in any benefits at all to such communities.[39]

This left the fourth reason. Here Powell found merit. A university's interest in a diverse student body is legitimated by the First Amendment's implied protection of academic freedom. This Constitutional halo makes the interest "compelling." However, the Medical School's use of a racial and ethnic classification scheme was not "precisely tailored" to effect the School's interest in diversity, argued Powell.

> The diversity that furthers a compelling state interest encompasses a far broader array of qualifications and characteristics of which racial or ethnic origin is but a single though important element. [The Medical School's] special admissions program, focused solely on ethnic diversity would hinder rather than further attainment of genuine diversity.[40]

The diversity which provides an educational atmosphere "conducive to speculation, experiment and creation" feeds upon a nearly endless range of experiences, talents, and attributes that students might bring to campus. In reducing diversity to racial and ethnic quotas, the Medical School wholly misconceived this important educational interest.

In sum, although the last of the Medical School's four reasons encompassed a "compelling governmental interest," the School's special admissions program was not necessary to effect that interest. The special admissions program was unconstitutional. So concluded Justice Powell.

How, then, did the *Bakke* decision become the basis upon which universities across the land enacted -- or maintained -- racially preferential admissions policies?

If Powell had concluded with his assessment of the Medical School's four reasons, *Bakke* would have left university affirmative action in a precarious situation. However, when the California Supreme Court had earlier ruled on Bakke's lawsuit, it had ordered Bakke admitted and forbidden the Medical School to make any use of race or ethnicity in its admissions decisions. Powell thought this went too far. Given higher education's protected interest in "diversity," and given that a student's race or ethnicity might add to diversity just in the same way that her age, work and travel experiences, family background, special talents, fluency in several languages, athletic prowess, military service, and unusual accomplishments might add, Justice Powell vacated that portion of the California Supreme Court's order.

Then he added some dicta for guidance. If universities want to understand diversity and the role that race and ethnicity might play in achieving it, they should look to Harvard, proposed Powell, and he appended to his opinion a long statement of Harvard's diversity program. In such a program, Powell contended, racial or ethnic background might

> be deemed a "plus" in a particular applicant's file, yet it does not insulate the individual from comparison with all other candidates for the available seats. . . . This kind of program

> treats each applicant as an individual in the admissions process. The applicant who loses out on the last available seat to another candidate receiving a "plus" on the basis of ethnic background will not have been foreclosed from all consideration for that seat simply because he was not the right color or had the wrong surname. It would mean only that his combined qualifications . . . did not outweigh those of the other applicant. His qualifications would have been weighed fairly and competitively, and he would have had no basis to complain of unequal treatment under the Fourteenth Amendment.[41]

In these off-hand comments, universities saw a green light for pushing ahead aggressively with their affirmative action programs. Although Justice Powell's basic holding could not have been plainer (any system like the Medical School's that makes race a consistently decisive factor, or that assesses applications along two different tracks defined by race, or that uses numerical quotas fails Constitutional muster), by the mid-1980s universities across the land had in place systems of admissions and scholarship awards that exhibited some or all of these features. When the University of Maryland's Banneker scholarships -- awarded only to African American students -- were held in violation of the Constitution in 1974,[42] the house of cards forming university affirmative action began to fall. In 1996, the Court of Appeals for the Fifth Circuit struck down the University of Texas Law School's admissions program,[43] and in November of the same year the voters of California adopted Proposition 209, forbidding among other things all uses of race in the public university admissions system. In 1998, the voters of Washington enacted a similar measure. Also in 1998, the Court of Appeals for the First Circuit struck down a Boston plan assigning students to selective high schools by race.[44] In 2001, two more schools saw their admissions programs invalidated: the University of Georgia[45] and the University of Michigan Law School.[46] In many of the cases, universities were using schemes that contravened Justice Powell's own holding; they were giving more than a "plus" to race. However, the Fifth Circuit Court in *Hopwood* threw a cloud even over Justice Powell's small window for affirmative action, boldly asserting that the *Bakke* holding was now dead as law and that race could not be used at all in admissions.

6. Equality

Given Justice Powell's singular opinion, supported by no one else on the Court, and given the drift of Supreme Court decisions on racial preferences since 1978,[47] the *Hopwood* court was not outlandish, if a bit presumptuous, in declaring Powell's holding in *Bakke* dead. If the Supreme Court eventually confirms that Powell's holding is no longer the law, it is not likely either to exhume the arguments of Justice Brennan in *Bakke*. This would be a misfortune. They convey an interpretation of Constitutional equality that Justice Powell never fully engaged.

Brennan agreed with Powell that "equal protection" must mean the same thing -- that is, remain one rule -- whether applied to black or white. But the same rule applied to different circumstances need not yield the same results. Racial preferences created for different reasons and producing different outcomes need not all be judged in the same harsh, virtually fatal, manner.

Powell thought there was no principled way to distinguish "benign" from "malign" discrimination, but Brennan insisted there was. He argued that if the Court looked carefully at its past cases striking down Jim Crow laws, it would see the principle at work. What the Court found wrong in Jim Crow was that it served no purpose except to mark out and stigmatize one group of people as inferior. The "cardinal principle" operating in the Court's decisions condemned racial classifications "drawn on the presumption that one race is inferior to another" or that "put the weight of government behind racial hatred and separation."[48] Brennan agreed with Powell that no public racial classification motivated by racial animus, no classification whose purpose is to stigmatize people with the "badge of inferiority," could withstand judicial scrutiny. However, the Medical School's policy, even if ill-advised or mistaken, reflected a public purpose far different from that found in Jim Crow. The policy ought not be treated as though it were cut from the same cloth.

Brennan granted that if a state adopted a racial classification for the purpose of humiliating whites, or stigmatizing Allan Bakke as inferior and confining him to second-class citizenship, that classification would be as odious as Jim Crow. But the Medical School's policy had neither this purpose nor this effect. Allan Bakke may have been upset and resentful at losing out under the special plan, but he wasn't "in any sense stamped as an inferior by the Medical School's rejection of him." Nor did his loss constitute a "pervasive injury," in the sense that wherever he went he would be treated as a "second-class citizen" because of his color.[49]

In short, argued Brennan, the principle embedded in the Equal Protection Clause should be viewed as an *anti-caste principle*, a principle that uniformly and consistently rejects all public law whose purpose is to subject people to an inferior and degraded station in life, whether they are black or white. Of course, given the asymmetrical position of whites and blacks in our country, we are not likely to encounter laws that try to stigmatize whites as an inferior caste (much less succeed at it). But this merely shows that a principle applied to different circumstances produces different results. Given that the Medical School's program reflected an effort to undo the effects of a racial caste system long-enduring in America, it expressed a purpose of great social importance and should not be found Constitutionally infirm.[50]

Powell never successfully engaged this way of reading "Constitutional equality." His insistence on clear, plain, unitary, absolute principle does not cut against the Brennan view. The issue between them is not the consistency and stringency of the principle but its content. Does the Constitution say, "The state cannot deliberately burden someone by race unless it passes an almost-always fatal test," or does it say, "The state cannot deliberately burden someone by race if its purpose is to create or maintain caste"?[51]

If Powell did not answer Brennan, neither, in turn, did Brennan successfully address one of Powell's worries. Were the Brennan view to prevail, Powell feared, then people could be subjected without check to the half-baked plans of any public agency in the country determined to do its bit to "remedy" the effects of historical discrimination. People would have no protection against arbitrary and over-reaching "remedial" policies.

However, there was an opening offered in the *Bakke* opinion for protecting people against runaway

preferences without outlawing programs like the Medical School's, a middle way never seized by either side. Recall that Powell dismissed the second reason offered by the Medical School -- that the state has an interest in ameliorating the effects of past discrimination -- by dismissing the School itself as neither competent nor authorized to make findings of past harm and adopt remedies for them. In both respects he was right. But on his own terms, then, he was required to give a respectful hearing to a body that *was* competent and authorized to inquire and legislate. Thus, suppose the legislature of California, after due deliberation and inquiry, had decided that the state's public universities should use special admissions plans like the Medical School's to temper in small part the evils attendant on California's own past history of discrimination. The legislature's interest in such ameliorative goals was conceded by Powell to be weighty. Moreover, limited and modest race-conscious policies would work in harmony with the legislature's aim, not at cross-purposes with it, as the Medical School's quota scheme worked at cross-purposes with achieving "diversity" (understood in Powell's sense). Why shouldn't Powell accede to the Medical School's admissions program under this hypothetical? Indeed, had he offered dicta on this point rather than on diversity, he might have moved the affirmative action debate into a more fruitful arena of debate. Letting state legislatures, and only state legislatures, decide on the make-up of state affirmative action programs would have left the decision in the hands of those who are representative of the people (as university faculties are not), competent to take account of a full range of relevant political factors (as courts are not), and authorized to legislate for the whole public good (as neither courts nor subordinate public agencies are). Of course, as subsequent political events in California and Washington indicate, some states might have foresworn any kind of preferential affirmative action. But other state legislatures might have explicitly authorized the kind of affirmative action admissions schemes their own public universities have carried on in a veiled, if not *sub rosa*, manner for two decades.

If we turn away from Constitutional exegesis, are we likely to find in political theory itself any principle of equality implying that *every* use of racial preferences in *every* circumstance works an intolerable injustice? There is reason to think not. To see why, consider John Rawls' theory of justice-as-fairness. For our purposes, what is striking about the theory is the division of labor it involves. Its very broadest principles of liberty and equality are themselves unable to single out proper micro-allocations of social benefits and burdens. This is not a defect; this is their nature. What they *can* do is structure roles and institutions which then create the social and legal machinery for assigning rights and responsibilities. Rawls' principles require a constitution to secure equality of citizenship to each member of society, but leaves most other matters to legislative judgment. Thus, law that in form and fact makes some people "second-class citizens" would be unconstitutional, clearly, but this limitation doesn't block asking people to bear unequal burdens for the common good, not even unequal burdens premised on their race or ethnicity. Nor does Rawls' principle of fair equality of opportunity block such burdens, either, for, while ordinarily discouraging selection based on race or ethnicity, it can itself be limited in the name of achieving greater equality of opportunity (a point conceded by Goldman).

Will putting aside Rawls and looking farther afield likely yield an understanding of general equality adamantly inhospitable to every use of preferences? The prospects seem dim. As Georgia Warnke has recently argued, a general notion of equality can argue as much for affirmative action (and the social inclusion it effects) as against it (and the racial non-neutrality it involves).[52]

7. Diversity

In the second wave of public controversy, spanning the 1990s, the affirmative action debate has narrowed both in its terms and focus. The focus is now race and ethnicity, because the central quarrel is now about university admissions. The terms are concomitantly constrained, with defenders of affirmative action emphasizing the virtues of diversity[53] and opponents emphasizing the harms affirmative action imposes upon its very beneficiaries.[54]

This second wave of public debate has not produced a flurry of philosophical articles like the first one. Indeed, in two 1990s collections edited by philosophers, *Affirmative Action: Social Justice or Reverse Discrimination?* and *The Affirmative Action Debate*, most of the entries by philosophical writers represent work from the 1970s, with only a smattering from the 1980s and 1990s.[55] The debate of the 1990s has generated its share of print, but most of it has flowed from the pens of public intellectuals of one stripe or another or found home in the pages of law reviews. Nor has any of this work, by philosophers and nonphilosophers alike, departed from the templates established nearly thirty years ago.[56]

That "diversity" has played a prominent role in the revived debates about preferences is no surprise, given that "diversity" was the legal hook proffered to universities in 1978. How does diversity support university affirmative action? In a widely circulated report in 1996, Neil Rudenstine, president of Harvard University, justified Harvard’s commitment to diversity by invoking John Stuart Mill, who stressed the value of bringing "human beings in contact with persons dissimilar to themselves, and with modes of thought and action unlike those with which they are familiar." A diverse student body, argued Rudenstine, is as much an "educational resource" as a university’s faculty, library, and laboratories.[57]

This is the diversity spoken of by Justice Powell, a diversity of opinions, experiences, backgrounds, talents, aspirations, and perspectives represented on campus that fosters intense intellectual exchange, exploration, and growth among all students. Obviously, an individual’s ethnicity, race, or gender can bear on this sort of diversity just as her being a devout Christian, tuba player, fluent speaker of Farsi, reserve military officer, former Peace Corps volunteer, champion swimmer, and self-taught auto mechanic can bear on it. Thus, a college seeking to admit a diverse entering class would not want to be utterly blind to race, ethnicity, or gender.[58]

In recently defending itself in *Gratz v. Bollinger* against an attack on the way it admits students into its College of Literature, Science, and Arts, the University of Michigan relied heavily on the diversity argument, insisting that it had a compelling educational interest in achieving racial and ethnic diversity. It put into evidence findings by one of its psychology professors, Patricia Gurin, showing that

> students learn better in a diverse educational environment, and they are better prepared to become active participants in our pluralistic, democratic society once they leave such a setting. . . . [S]tudents who experienced the most racial and ethnic diversity in classroom settings and in informal interactions with peers showed the greatest engagement in active

> thinking processes, growth in intellectual engagement and motivation, and growth in intellectual and academic skills.[59]

These findings might seem dispositive. Racial and ethnic diversity furthers good education. But how much so? Is such diversity literally compelling or merely legally so.[60] In defending their affirmative action policies, universities must be careful not to commit themselves to claims that look dubious on second thought.

For example, is the University of Michigan contending that, without a steady 8 or 9 percent representation of African American undergraduates on campus, it wouldn't be able to offer an education adequately developing its students' intellectual skills and civic commitments? If the University makes this argument, does it then imply that students attending Morehouse College or Florida A&M University cannot receive excellent educations, intellectually and civically? After all, these campuses are not at all racially and ethnically diverse. Indeed, Morehouse is lacking in gender diversity as well, since it is a college for men.

Once we begin to attend to the extraordinary variety among institutions of higher education in America, we might conclude that no single pattern of diversity within a school is a *sine qua non* for students' intellectual growth and civic development. Having 8 or 9 percent African American undergraduates on campus may be an educational desideratum but not an imperative.

In fact, it is clear enough that, had Patricia Gurin's findings come out differently, the University of Michigan was not about to relax its target of 8 or 9 percent African American undergraduates. Although racial and ethnic diversity at Ann Arbor might enrich students' educational experiences, the main reason the University of Michigan strives for a reasonable representation of minorities on campus is because of the way it conceives of its mission: to prepare Michigan's future leaders.

The argument is straightforward:

- The leadership of the state ought roughly to represent the state's population.
- As the state's premier training ground for leadership, the University ought to graduate rising generations of future leaders that conform to this representational goal.
- To graduate such rising generations, it needs to admit racially and ethnically representative classes.

This is the "Michigan Mandate."[61] Racial and ethnic diversity aren't incidental contributors to a distinct academic mission; they are part of the mission of the University, just as educating young people from Michigan is part of the mission.

Likewise, the principal aim of the elite universities studied by William Bowen and Derek Bok in *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions* was not, through vigorous affirmative action, to enhance the liberal learning of their students (although they welcomed this gain for all students). Their main motive for assuring that the numbers of blacks on their campuses would be more than token derived from their self-conceptions as institutions training

individuals who would some day take up leadership roles in the professions, arts, science, education, politics, and government.[62] The nation, they believe, will be stronger and more just with a leadership reflecting a broader racial and ethnic profile than it does now (and than it did twenty-five years ago).

But at what cost? Stephan and Abigail Thernstrom have argued that the cost is high and falls on the very persons affirmative action is supposed to benefit. Under-prepared blacks are thrown into academic environments where they cannot compete.[63] In the Thernstroms' view, race-blind admissions policies would result in a desirable "cascading," with African American students ending up at colleges and universities where the academic credentials of entering students matches their own. In their own study, Bowen and Bok show that cascading isn't necessarily a valuable phenomenon. In fact, at the schools they studied, the better the institution a student entered, whatever his academic credentials, the more likely he was to graduate and go on to further education and earn a good income.[64]

Of course, the select schools Bowen and Bok studied may be quite unrepresentative of the full range of colleges and universities that resort to racial preferences, and the cost-benefit ratio that holds for these schools may not hold for the rest.[65] Nor does cascading in general look like a bad thing, if James Traub's portrait of the University of California system after Proposition 209 indicates a generalizable effect.[66]

Although the full facts about affirmative action in the university remain contentious and under debate, the data provided by Bowen and Bok settle at least one matter. Switching to class-based affirmative action would not be a proxy for race-based affirmative action.[67] At every income level, white students possess better grades and SAT scores than blacks at that same level. Giving preference by social class would result, thus, in disproportionately more whites than blacks entering selective universities.

8. Desert Confounded, Desert Misapplied

The affirmative action debate throws up many ironies but one in particular should be noted. From the time in 1973 when Judith Jarvis Thomson conjectured that it was "not entirely inappropriate" that white males bear the costs of the community's "making amends" to blacks and women through preferential affirmative action, the affirmative action debate has been distracted by intense quarrels over *who deserves what*. Do the beneficiaries of affirmative action deserve their benefits? Do the losers deserve their loss?

Christopher Edley, the White House assistant put in charge of President Clinton's review of affirmative action policy in 1994-95, speaks of how, during the long sessions he and his co-workers put in around the conference table, the discussion of affirmative action kept circling back to the "coal miner's son" question.

> Imagine a college admissions committee trying to decide between the white [son] of an Appalachian coal miner's family and the African American son of a successful Pittsburgh neurosurgeon. Why should the black applicant get preference over the white applicant?[68]

Why, indeed? This is a hard question if one defends affirmative action in terms of compensatory or distributive justice. If directly doing justice is what affirmative action is about, then its mechanisms must be adjusted as best they can to reward individual desert and true merit. The "coal miner's son" example is meant to throw desert in the defender's face: here is affirmative action at work thwarting desert, for surely the coal miner's son -- from the hard scrabble of Harlan County, say -- has lived with far less advantage than the neurosurgeon's son who, we may suppose, has reaped all the advantages of his father's (or mother's) standing. Why should the latter get a preference?

A defender might answer in the way that Charles Lawrence and Mari Matsuda do: "All the talk about class, the endless citings of the 'poor white male from Appalachia,' cannot avoid the reality of race and gender privilege."[69] White privilege means that racial preferences really do balance the scales. Male privilege means that gender preferences really do make selections fairer. There must be *no* concession: in every case the loser in affirmative action is *not* the more deserving.[70]

Even Justice Brennan tried his hand at this argument, writing in *Bakke*:

> If it was reasonable to conclude -- as we hold that it was -- that the failure of minorities to qualify for admission at Davis under regular procedures was due principally to the effects of past discrimination, then there is a reasonable likelihood that, but for pervasive racial discrimination, . . . [Bakke] would have failed to qualify for admission even in the absence of Davis' special admissions program.[71]

Counterfactually, Bakke was not denied anything to which he had moral claim in the first place.

Just as Mary Anne Warren and James Rachels in the 1970s thought that the losers under affirmative action were losing only illicit privileges, and the gainers merely gaining what should have been theirs to start with, so Michel Rosenfeld in the 1990s, in his extended "dialogic" defense of affirmative action, echoed the same thought:

> Although affirmative action treats innocent white males unequally, it need not deprive them of any genuine equal opportunity rights. *Provided an affirmative action plan is precisely tailored to redress the losses in prospects of success [by blacks and women] attributable to racism and sexism, it only deprives innocent white males of the corresponding undeserved increases in their prospects of success* . . . . [R]emedial affirmative action does not take away from innocent white males anything that they have rightfully earned or that they should be entitled to keep.[72]

But preferential programs that give blanket preferences by race or gender are hardly precisely tailored to match desert and reward since, as Lawrence and Matsuda acknowledge, the white male "privilege" is "statistical."[73] Yet it is individuals, not statistical averages, who gain or lose in the admissions committee decisions and employment office selections.

More pointedly, why the persistence of this obdurate strategy of defense when real-world affirmative action has had no truck with it? The affirmative action programs legitimated under the Civil Rights Act, in both their nonpreferential and preferential forms, had -- and have -- a specific aim: to change institutions so that they can meet the nondiscrimination mandate of the Act. Selection by race or gender was -- and is -- a means to such change. To the extent that such selection also compensated individuals for past wrongs or put people in places they really deserved*, these are incidental by-products of a process aimed at something else*.

Similarly, when the Medical School offered four reasons in defense of the special admissions program that left Bakke on the outside, none of these reasons said anything about matching admissions and desert. The criteria of the special admissions program -- race and ethnicity -- were instruments to further ends: integrating the classroom, the profession, and the delivery of medical services; and breaking the chain of self-reproducing societal discrimination. If the neurosurgeon's son *because of his race* can advance each of these goals and the coal miner's son can not, then the selection decision is easy: pick the *black* neurosurgeon's son (however advantaged he may have been) over the *white* coal miner's son (even were he the most deserving creature imaginable). The aims of real-world affirmative action make race and ethnicity (and sometimes gender) salient, not personal desert or merit.

Bibliography

- Beckwith, Francis J. and Jones, Todd E. (eds). *Affirmative Action: Social Justice or Reverse Discrimination?* Amherst, New York: Prometheus Books, 1997.
- Bolick, Clint. *The Affirmative Action Fraud: Can We Restore the American Civil Rights Vision?* Washington, D.C.: Cato Institute, 1996.
- Boxill, Bernard R. *Blacks and Social Justice*. Totowa, New Jersey: Rowman & Allanheld, 1984.
- Cahn, Steven M. (ed). *Affirmative Action and the University: A Philosophical Inquiry*. Philadelphia: Temple University Press, 1993.
- Cahn, Steven M. (ed.). *The Affirmative Action Debate*. New York: Routledge, 1995. [Contains many of the main articles from the 1970s.]
- Capaldi, Nicholas. *Out of Order: Affirmative Action and the Crisis of Doctrinaire Liberalism*. Buffalo, New York: Prometheus Books, 1985.
- Carter, Stephen L. *Reflections of an Affirmative Action Baby*. New York: Basic Books, 1991.
- Cohen, Carl. *Naked Racial Preference*. Lanham, Maryland: Madison Books, 1995.
- -----. "The Corruption That is Group Preference," *Academic Questions*, 11 (Summer 1998), 14-22.
- Cohen, Marshall et al. (eds). *Equality and Preferential Treatment. Princeton, New Jersey: Princeton University Press, 1977.* [Contains the early articles in *Philosophy & Public Affairs.*]
- Curry, George E. (ed). *The Affirmative Action Debate*. Reading, Massachusetts: Addison-Wesley Publishing Company, 1996.
- Dworkin, Ronald. *A Matter of Principle*. Cambridge, Massachusetts: Harvard University Press, 1985.
- Eastland, Terry. *Ending Affirmative Action: The Case for Colorblind Justice*. New York: Basic Books, 1996.

- Eastland, Terry and Bennett, William J. *Counting By Race: Equality from the Founding Fathers to* Bakke *and* Weber. New York: Basic Books, 1979.
- Edley, Christopher, Jr. *Not All Black and White: Affirmative Action and American Values*. New York: Hill and Wang, 1996.
- Edwards, John. *When Race Counts: The Morality of Racial Preference in Britain and America.* London: Routledge, 1995.
- Ezorsky, Gertrude. *Racism and Justice: The Case for Affirmative Action*. Ithaca, New York: Cornell University Press, 1991.
- Fullinwider, Robert K. *The Reverse Discrimination Controversy: A Moral and Legal Analysis*. Totowa, New Jersey: Rowman and Littlefield, 1980.
- Fullinwider, Robert K. and Mills, Claudia (eds). *The Moral Foundations of Civil Rights*. Totawa, New Jersey: Rowman & Littlefield, 1986.
- Glazer, Nathan. *Affirmative Discrimination: Ethnic Inequality and Public Policy*. New York: Basic Books, 1975.
- -----. "For Racial Dispensation in Admissions," *Academic Questions*, 11 (Summer 1998), 22-32.
- Goldman, Alan H. *Justice and Reverse Discrimination*. Princeton, New Jersey: Princeton University Press, 1979.
- Greenawalt, Kent. *Discrimination and Reverse Discrimination*. New York: Alfred A. Knopf, 1983.
- Gross, Barry R. *Discrimination in Reverse: Is Turnabout Fair Play?* New York: New York University Press, 1978.
- Gross, Barry R. (ed). *Reverse Discrimination*. Buffalo, New York: Prometheus Books, 1977.
- Gutmann, Amy and Thompson, Dennis. *Democracy and Disagreement*. Cambridge, Massachusetts: Harvard University Press, 1996.
- Kahlenberg, Richard D. *The Remedy: Class, Race, and Affirmative Action*. New York: Basic Books, 1996.
- Lawrence, Charles R. III and Matsuda, Mari J. *We Won't Go Back: Making the Case for Affirmative Action*. Boston: Hougton Mifflin Company, 1997.
- Mills, Nicolaus (ed). *Debating Affirmative Action: Race, Gender, Ethnicity, and the Politics of Inclusion*. New York: Dell Publishing, 1994.
- Mosley, Albert G. and Capaldi, Nicholas. *Affirmative Action: Social Justice or Unfair Preference?* Lanham, Maryland: Rowman and Littlefield, 1996.
- Nieli, Russell (ed). *Racial Preference and Racial Justice: The New Affirmative Action Controversy*. Washington, D.C.: Ethics and Public Policy Center, 1991.
- O'Neil, Robert M. *Discriminating Against Discrimination: Preferential Admissions and the* DeFunis *Case*. Bloomington, Indiana: Indiana University Press, 1975.
- Paul, Ellen Frankel et al. (eds). *Reassessing Civil Rights*. Oxford: Blackwell Publishers, 1991.
- Rosenfeld, Michel. *Affirmative Action and Justice: A Philosophical and Constitutional Inquiry*. New Haven, Connecticut: Yale University Press, 1991.
- Skrentny, John David. *The Ironies of Affirmative Action: Politics, Culture, and Justice in America*. Chicago: University of Chicago Press, 1996.
- Symposium: *Bakke* -- Civil Rights Perspectives. *Harvard Civil Rights-Civil Liberties Law Review*, 14 (Spring 1979).
- Symposium: Race-Based Remedies. *California Law Review*, 84 (July 1996).

- Symposium: The Meanings of Merit -- Affirmative Action and the California Civil Rights Initiative. *Hastings Constitutional Law Quarterly*, 23 (Summer 1996).
- Valls, Andrew. "The Libertarian Case for Affirmative Action," *Social Theory and Practice*, 25 (Summer 1999), 299-323.
- Young, Iris Marion. *Justice and the Politics of Difference*. Princeton, New Jersey: Princeton University Press, 1990.
- Waldron, Jeremy. "Humility and the Curse of Injustice," in Robert Post and Michael Rogin, eds., *Race and Representation: Affirmative Action* (New York: Zone Books, 1998), 385-389.

Other Internet Resources

- United States Commission on Civil Rights
- University of Michigan Law Suit Site
- American Association for Affirmative Action
- Center for Individual Rights
- Diversity Web
- Leadership Conference on Civil Rights
- Campaign for a Color-Blind America

Related Entries

equality | justice | rights

Robert K. Fullinwider
rf30@umail.umd.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 28, 2001
Content last modified: December 28, 2001

Stanford Encyclopedia of Philosophy
Notes to Affirmative Action

Notes

[1.] The legality of racial "set-asides" in public construction contracting remains an unresolved issue but the public in general is not aroused by the vicissitudes of the contracting process and the good or ill fortunes of construction firms.

[2.] Hugh Davis Graham, *The Civil Rights Era: Origins and Development of National Policy 1960-1972* (New York: Oxford University Press, 1990), p. 413.

[3.] So profound was the shock to the academy that Nicholas Capaldi, writing in 1985, remained under the impression that "[a]ffirmative action as a public policy was first applied on a massive and national scale to institutions of higher learning." Nicholas Capaldi, *Out of Order: Affirmative Action and the Crisis of Doctrinaire Liberalism* (Buffalo, New York: Prometheus Books, 1985), p. 1.

[4.] John Rawls, *A Theory of Justice* (Cambridge, Massachusetts: Harvard University Press, 1971).

[5.] Thomas Nagel, "Equal Treatment and Compensatory Discrimination," *Philosophy & Public Affairs*, 2 (Summer 1973), 348-363.

[6.] Judith Jarvis Thomson, "Preferential Hiring," *Philosophy & Public Affairs*, 2 (Summer 1973), 364-484.

[7.] Ironically enough, the first discussions of "inverse" discrimination began in one of the prime sites of analytical philosophy, *Analysis*. For the full record of exchanges, see James W. Nickel, "Discrimination and Morally Relevant Characteristics," *Analysis*, 32 (March 1972), 113-14; J. L. Cowan, "Inverse Discrimination," *Analysis*, 33 (October 1972), 10-12; Paul Taylor, "Reverse Discrimination and Compensatory Justice," *Analysis*, 33 (June 1973), 177-82; Roger Shiner, "Individuals, Groups and Inverse Discrimination," *Analysis*, 33 (June 1973), 185-87; Philip Silvestri, "The Justification of Inverse Discrimination," *Analysis*, 34 (October 1973), 31; William A. Nunn, "Reverse Discrimination," *Analysis*, 34 (April 1974), 151-54; James W. Nickel, "Should Reparations Be to Individuals or Groups?" *Analysis*, 34 (April 1974), 154-60; Alan Goldman, "Reparations to Individuals or Groups?" *Analysis*, 35 (April 1975), 168-70; Sara Ann Ketchum and Christine Pierce, "Implicit Racism," *Analysis*, 36 (January, 1976), 91-5; Paul Woodruff, "What's Wrong with Discrimination?" *Analysis*, 36 (March 1976), 158-60; Robert L. Simon, "Statistical Justification of Discrimination," *Analysis*, 38 (January 1978), 37-42.

[8.] See, for example, Gertrude Ezorsky, "Hiring Women Faculty," *Philosophy & Public Affairs*, 7 (Autumn 1977), 86.

[9.] Alan Goldman, "Affirmative Action," *Philosophy & Public Affairs*, 5 (Winter 1976), 182-83.

[10.] Robert Simon, "Preferential Hiring: A Reply to Judith Jarvis Thomson," *Philosophy & Public Affairs*, 3 (Spring 1974), 315-19; George Sher, Justifying Reverse Discrimination in Employment," *Philosophy & Public Affairs*, 4 (Winter 1975), 162; George Sher, "Reverse Discrimination, the Future, and the Past," *Ethics*, 90 (October 1979), 81-2; Alan Goldman, "Affirmative Action," *Philosophy & Public Affairs*, 5 (Winter 1976), 190-1. See also Robert K. Fullinwider, "Preferential Hiring and Compensation," *Social Theory and Practice* (Spring 1975), 307-320; Alan Goldman, *Justice and Reverse Discrimination* (Princeton, New Jersey: Princeton University Press, 1979), 65-102; Robert K. Fullinwider, *The Reverse Discrimination Controversy: A Moral and Legal Analysis* (Totowa, New Jersey: Rowman & Littlefield, 1980), pp. 30-44.

[11.] Thomson, "Preferential Hiring," p. 377; Simon, "Preferential Hiring: A Reply," 312.

[12.] Goldman, "Affirmative Action," p. 191; Goldman, *Justice and Reverse Discrimination*, pp. 24-8.

[13.] Barry R. Gross, "Is Turn About Fair Play?" in Barry R. Gross, ed., *Reverse Discrimination* (Buffalo, New York: Prometheus Books, 1977), p. 382; Barry R. Gross, *Discrimination in Reverse: Is Turnabout Fair Play* (New York: New York University Press, 1978), p. 97.

[14.] Robert L. Simon, "Individual Rights and ‘Benign’ Discrimination," *Ethics*, 90 (October 1979), 96.

[15.] Terry Eastland and William Bennett, *Counting By Race: Equality from the Founding Fathers to* Bakke *and* Weber (New York: Basic Books, 1979), 144.

[16.] Gross, *Discrimination in Reverse*, 125-42.

[17.] Mary Anne Warren, "Secondary Sexism and Quota Hiring," *Philosophy & Public Affairs*, 6 (Spring 1977), 256.

[18.] James Rachels, "What People Deserve," in John Arthur and William Shaw, eds., *Justice and Economic Distribution* (Englewood Cliffs, New Jersey, Prentice-Hall, 1978), 162.

[19.] See Sher, "Justifying Reverse Discrimination," 165ff.

[20.] Lisa Newton, "Reverse Discrimination as Unjustified," *Ethics*, 83 (July 1973), 310. Similar sentiments were expressed by Virginia Black:

> If it is irrational and unjust and cruel to fire someone because he is a black or she is a woman -- cases whose absurdity seems obvious -- then it is equally irrational and unjust and cruel to hire someone because be is a black or she is a woman. To appreciate the parallel, one has only to remember that to hire X *because* of color is, ipso facto, *not* to hire Y because of color. When inscribed in law, this is racism.

Virginia Black, "The Erosion of Legal Principles in the Creation of Legal Policies," *Ethics*, 84 (January 1974), 106

[21.] Eastland and Bennett, *Counting By Race*, 149. This idea that using racial preferences involved a kind of practical contradiction was given voice and support at the highest levels of government in the 1980s. William Bradford Reynolds, during his tenure as Assistant Attorney General for Civil Rights in the Reagan Administration, contended:

> [T]o those who argue that we must use race to get beyond racism . . . [h]istory teaches us all too well that such an approach does not work. It is wrong when the government bestows advantages on whites at the expense of innocent blacks; it assumes no greater claim of morality if the tables are turned … . Whatever group membership one inherits, it carries with it no entitlement to preferential treatment over those not similarly endowed with the same immutable characteristics. Any compromise of this principle is discrimination, plain and simple, and such behavior is no more tolerable when employed remedially, in the name of "affirmative action" or "racial balance," to bestow a gratuitous advantage on members of a particular group, than when it is divorced from such beneficence and for the most invidious of reasons works to one's disadvantage.

William Bradford Reynolds, "Individualism vs. Group Rights: The Legacy of *Brown*," *Yale Law Journal*, 93 (May 1984), 1004. While Reynolds found the proposition, "Use race to achieve a colorblind society," an assault on logic, he belonged to an administration whose defense policy -- like that of a long line of administrations -- was grounded on the proposition, "Prepare for war in order to have peace."

[22.] Richard Wasserstrom, "Racism, Sexism, and Preferential Treatment: An Approach to the Topics," *UCLA Law Review*, 24 (February 197), 581-622.

[23.] Carl Cohen, *Naked Racial Preference* (Lanham, Maryland: Madison Books, 1995), 20 (from "Race and the Constitution," first published in *The Nation*, 220 [February 8, 1975]).

[24.] Cohen, *Naked Racial Preference*, 52 (from "Who Are Equals?" first published in *National Forum: The Phi Kappa Phi Journal*, 58 (Winter 1978).

[25.] Goldman, *Justice and Reverse Discrimination*, 164-65.

[26.] Title 42 United States Code Sec. 2000d. Title IX of the Education Amendments of 1972 promised the same protection against gender discrimination. See 20 USC 1681.

[27.] "(a) … It shall be … unlawful for an employer -- (1) to . . . refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin; or (2) to limit, segregate, or classify his employees or applicants in any way which would tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin … ." Title 42 USC Sec 2000e-2

[28.] *Griggs v. Duke Power Company*, 401 U.S. 424 (1971), at 430, 431. At issue in the case was the use of an aptitude test and a high-school graduation requirement to screen job applicants. Duke Power Company did not succeed in showing that the results of the aptitude test or the possession of a high school diploma bore any demonstrable relation to performance at such of its jobs as janitor, maintenance worker, and the like.

[29.] For a more extended discussion of goals and quotas, see Fullinwider, *The Reverse Discrimination Controversy*, 162-177.

[30.] Consider, for example, this 1969 court decision involving a union with a record of excluding blacks and Mexican-Americans. the court imposed an injunction that

> prohibit[ed] discrimination in excluding persons from union membership or referring persons for work; prohibit[ed] use of member's endorsements, family relationship or elections as criteria for membership; … ordered the development of objective membership criteria and prohibited new members … until developed; and ordered continuation of chronological referrals for work, with alternating white and Negro referrals until objective membership criteria are developed.

On appeal, the Court of Appeals for the Fifth Circuit upheld the lower court. In its view,

> [t]he District Court did no more than prevent future discrimination when it prohibited a continuing exclusion of Negroes through the application of an apparently neutral membership provision … which served no significant trade-related purpose. [Further] the District Court did no more [in barring new membership until objective criteria were developed] than ensure that the injunction against further racial discrimination would be fairly administered. Absent objective criteria … covert subversion of the purpose of the injunction could occur. The same administrative reasons support alternating white and Negro referrals … .

Asbestos Workers v. Vogler, 407 F. 2d 1047 (1960), at 1051, 1055. Each part of the lower court's order, including the part that required racially balanced referrals, harkened back to a single ground: the part was necessary to prevent future discrimination.

[31.] For a list of Circuit Court decisions embracing this theory, see Robert K. Fullinwider, "Achieving Equal Opportunity," in Robert K. Fullinwider and Claudia Mills, eds., *The Moral Foundations of Civil Rights* (Totowa, New Jersey: Rowman & Littlefield, 1986), 106-08 and accompanying footnotes.

[32.] Regents of the University of California v. Bakke, 438, U.S. 265 (1978). In 1974, the Court accepted for decision a similar case involving a race-preference policy at the University of Washington Law School. However, perhaps experiencing second thoughts, the Court then dismissed the case as moot (the plaintiff, admitted to the Law School by a lower court, was about to graduate). See *De Funis v. Odegaard*, 416 U.S. 312 (1974).

[33.] Unlike Title VII, which was grounded in the Interstate Commerce Clause of the Constitution, Title VI was grounded in the federal government's spending powers. Brennan relied on Congressional debate to argue for the substantive identity of Title VI, on the one hand, and the Fourteenth Amendment to the Constitution, which forbids states to discriminate, and the Fifth Amendment, which incorporates the Fourteenth Amendment's strictures and applies them to the federal government, on the other hand. The debate can be summed up in this rhetorical question: "Why should the government through its spending be subsidizing acts of private discrimination that it would be forbidden by the Constitution to do itself?" See 438 U.S. 265, at 329-336.

[34.] 438 U. S. 265, at 295-300 (Powell quoting from Archibald Cox, *The Role of the Supreme Court in American Government* [New York: Oxford University Press, 1976], 114).

[35.] 438 U.S. 265, at 300.

[36.] 438 U.S. 265, at 307.

[37.] 438 U.S. 265, at 308.

[38.] 438 U.S. 265, at 310. Here Powell suggests that the threshold for justified preferences under the Constitution is exactly the same as the one for their use under Title VII: preferences may be used by an institution to prevent its own present and future discrimination. Justice Scalia takes this suggestion to stand for the Court's current settled doctrine. See *Richmond v. J. A. Croson Company*, 488 U.S. 469 (1989) (Scalia concurring).

[39.] 438 U.S. 265, at 311.

[40.] 438 U.S. 265, at 316.

[41.] 438 U. S. 265, at 318, 319.

[42.] *Podberesky v. Kirwan*, 38 F. 3d 147 (Fourth Circuit, 1994).

[43.] *Hopwood v. Texas*, 78 F 3d 932 (Fifth Circuit, 1996).

[44.] *Wessmann v. Gittens*, 106 F 3d 798 (First Circuit, 1998).

[45.] *Johnson v. Board of Regents*, 263 F 3d 1234 (Eleventh Circuit, 2001).

[46. *Grutter v. Bollinger*, 137 F. Supp. 2d 821 (2001).

[47.] See *Wygant v. Jackson*, 476 U.S. 267 (1986); *Richmond v. J. A. Croson Company*, 488 U.S. 469 (1989); *Adarand Constructors v. Pena*, 515 U.S. 200 (1995).

[48.] 438 U.S. 265, at 358.

[49.] 438 U.S. 265, at 376.

[50.] 438 U.S. 265, at 363.

[51.] A clear discussion of *Bakke* and equality's dictates can be found in "Bakke's Case: Are Quotas Really Unfair?" and "What Did *Bakke* Really Decide?" in Ronald Dworkin, *A Matter of Principle* (Cambridge, Massachusetts: Harvard University Press, 1985).

[52.] Georgia Warnke, "Affirmative Action, Neutrality, and Integration," *Journal of Social Philosophy*, 29 (1998), 87-103.

[53.] Chang-Lin Tien, "A Personal Perspective on Affirmative Action," in Robert Post and Michael Rogin, eds., *Race and Representation: Affirmative Action* (New York: Zone Books, 1998), 379-83.

[54.] Carl Cohen, "The Corruption That Is Group Preference," *Academic Questions*, 11 (Summer 1998), 14-21.

[55.] Francis J. Beckwith and Todd E. Jones, eds., *Affirmative Action: Social Justice or Reverse Discrimination?*" (Amherst, New York: Prometheus Books, 1997); Steven M. Cahn, ed., *The Affirmative Action Debate* (New York: Routledge, 1995). Of the twenty-two chapters in Cahn, only four postdate 1990; of the fifteen chapters in Beckwith and Jones, only one by a philosopher postdates 1990 (one other

reworks material written in the 1980s).

[56.] Some recent philosophical work on affirmative action: Albert G. Mosley and Nicholas Capaldi, *Affirmative Action: Social Justice or Unfair Preference?* (Lanham, Maryland: Rowman & Littlefield, 1996); Albert G. Mosley, "Policies of Straw or Policies of Inclusion? A Review of Pojman's 'Case Against Affirmative Action'," *International Journal of Applied Philosophy*, 12 (1998), 161-168; Louis Pojman, "Straw Man or Straw Theory? A Reply to Albert Mosley," *International Journal of Applied Philosophy*, 12 (1998), 169-80; Francis J. Beckwith, "The 'No One Deserves His or Her Talents' Argument for Affirmative Action," *Social Theory and Practice*, 25 (1999), 53-60; Sarah Stroud, "The Aim of Affirmative Action," *Social Theory and Practice*, 25 (1999), 385-408; Stephen Kershnar, "Strong Affirmative Action Programs and Disproportionate Burdens," *Journal of Value Inquiry*, 33 (1999), 201-209.

[57.] Neil Rudenstine, "Why a Diverse Student Body Is So Important," *Chronicle of Higher Education*, 42 (April 19, 1996), B1.

[58.] Elizabeth Anderson, in "The Democratic University: The Role of Justice in the Production of Knowledge," in Ellen Frankel Paul et al., eds. *The Just Society* (New York: Cambridge University Press, 1995), 186-219, makes a complementary argument. It parallels Rudenstine's, but emphasizes epistemic rather than educational considerations and applies to faculty and graduate students rather than undergraduates. "A knowledge claim gains objectivity and warrant," Anderson insists, "to the degree that it is the product of exposure to the fullest range of criticisms and perspectives … . Universities [should] recruit students and faculty to ensure broad representation of people from all walks of life, so that the products of inquiry are open to critical scrutiny and influence from the widest range of viewpoints, and so that the subjects and direction of inquiry are responsive to the widest range of interests" (203, 198). But Anderson goes further than simply commending broad representation. She maintains that

> [t]he internal knowledge-promoting aims of the university call for measures to promote equality of access by all groups in society to membership in its ranks. This is an argument for affirmative action in university admissions and faculty hiring that recognizes the positive contributions that members of oppressed groups can and do make to enhancing the objectivity of research. Equality of access through affirmative action policies is not, therefore, an external political goal that threatens to compromise the quality of research. It is a means to promote the objectivity of that research (198).

For a discussion of her argument, see Robert K. Fullinwider, "Diversity and Affirmative Action," *Report from the Institute for Philosophy and Public Policy*, 17 (Winter/ Spring 1997), 26-31.

[59.] *Gratz v. Bollinger*, 122 F. Supp. 2d 811 (2000), at 823.

[60.] Recall that Justice Powell put the university's interest in diversity under the umbrella of academic freedom which is protected by the First Amendment of the Constitution. It is this latter connection that

makes the university's interest legally "compelling." In this sense, the university's interest in admitting tuba players is as compelling as its interest in admitting racial minorities.

[61.] *Gratz v. Bollinger*, 135 F. Supp. 2d 790 (2001), at 796-797.

[62.] William G. Bowen and Derek Bok, *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions* (Princeton, New Jersey: Princeton University Press, 1998), 7.

[63.] Stephan Thernstrom and Abigail Thernstrom, *America in Black and White: One Nation, Indivisible* (New York" Simon and Schuster, 1997), 395-411.

[64.] Bowen and Bok, *The Shape of the River*, 63, 114, 144.

[65.] For critiques of *The Shape of the River*, see Terrance Sandalow, "Identity and Equality: Minority Preferences Reconsidered," *Michigan Law Review*, 97 (May 1999), 1874-1916, and Stephan Thernstrom and Abigail Thernstrom, "Reflections on The Shape of the River," *UCLA Law Review*, 46 (June 1999), 1583-1631.

[66.] James Traub, "The Class of Prop 209," *New York Times Sunday Magazine*, May 2, 1999, 44ff.

[67.] Bowen and Bok, *The Shape of the River*, 46-50.

[68.] Christopher Edley, Jr., *Not All Black and White: Affirmative Action and American Values* (New York: Hill and Wang, 1996), 132ff. Edley's example involved a coal miner's daughter. I've changed the daughter to a son for expository purposes.

[69.] Charles R. Lawrence III and Mari J. Matsuda, *We Won't Go Back: Making the Case for Affirmative Action* (Boston: Houghton Mifflin Company, 1997), 199-1.

[70.] Bernard Boxill has insisted that merely receiving benefits produced by injustice is enough to make one personally liable to compensate the victim of injustice (Bernard Boxill, "The Morality of Reparations," *Social Theory and Practice*, 2 [Spring 1972], 113-123). And who is the victim of injustice? "We *know* that all blacks, lower class, middle class, and upper class, have been wronged by racial injustice" (Bernard Boxill, *Blacks and Social Justice* [Totowa, New Jersey: Rowman and Allanheld, 1984], 164). In "The Morality of Preferential Hiring," *Philosophy & Public Affairs*, 7 (Spring 1978), 251, Boxill argues that in affirmative action the correlation between "preferences received" and "compensation deserved," though not perfect, is very high.

[71.] 438 U.S. 265, at 365-6.

[72.] Michel Rosenfeld, *Affirmative Action and Justice: A Philosophical Inquiry* (New Haven, Connecticut: Yale University Press, 1991), 307-8. Emphasis added.

[73.] Lawrence and Matsuda, *We Won't Go Back*, 252.

Robert K. Fullinwider
rf30@umail.umd.edu

First published: December 28, 2001
Content last modified: December 28, 2001

Exploitation

To exploit others is to take unfair advantage of them. Although 'exploitation' has figured prominently in Marxist theories, it is frequently invoked in ordinary moral and political discourse. This entry surveys various definitions that have appeared in the literature, attempts to identify the core elements of exploitation, and then considers its moral force. .

Introduction

Consider these examples of alleged exploitation.

> 1. The president of Stanford University claimed that big-time college athletics "reeks of exploitation," because the universities gain a great deal of revenue from the services of the athletes while the athletes (whose graduation rate is much lower than that of non-athletes) gain little from their college experience. (Kennedy 1990)
>
> 2. Advocates for people with AIDS protested the (then) annual price of AZT ($8,000), a drug that was thought to slow the progress of HIV. It was claimed that the Burroughs-Welcome Co., which produces AZT, was exploiting people who were already suffering.
>
> 3. A newspaper article described the proliferation of "posh strip clubs," at which topless young women dance for the customers. It was both claimed and denied that these clubs were loci of exploitation. The article noted that many believe that such bars "exploit women." At the same time, "many dancers say their work is no more exploitative than

> most other forms of employment," and the owner of one bar remarked that "If anyone is being exploited it is the men, the guys buying into the fantasy of she really likes me." (*New York Times*, April 15, 1992, C.12)
>
> 4. *USA Today* featured an article advocating the legalization of organ sales, whereby a person could be paid cash for a kidney. One reply maintained that such a policy would "open wide the door to exploitation." (September 14, 1991)
>
> 5. When Mary Beth Whitehead refused to surrender her daughter (who was known as "Baby M"), as specified in her $10,000 surrogacy contract with William Stern, the case touched off extended public, legal, and philosophical discussion of surrogate motherhood. One commentator observed that "one of the most serious charges against surrogate motherhood contracts is that they exploit women." (Field 1989, 25)

Although we frequently claim that some act, practice, or transaction is exploitative, the concept of exploitation is typically invoked without much analysis or argument, as if its meaning and moral force were self-evident. They are not. Even if some or even all of these sorts of claims are true, we still need to ask why are they true? And if they are true, what follows? More precisely, we can ask two questions: (1) what are the truth conditions of an exploitation claim? (2) what is the moral force of an exploitation claim? Let me explain.

For present purposes, an *exploitation claim* refers to statements that A’s interaction with B is (or is not) wrongfully exploitative or to statements that presuppose such a claim. To say that colleges exploit student athletes is to make an exploitation claim. Susan Okin makes an exploitation claim when she says that our family system constitutes "the pivot of a societal system of gender that renders women vulnerable to dependency, exploitation, and abuse," for we must know what exploitation involves to determine whether this claim is true. (Okin 1989, 135-36).

The first task of a theory of exploitation is to provide the truth conditions for an exploitation claim. At least one such condition is a moral criterion: a transaction is exploitative only if it is *unfair*. Interestingly, however, the (moral) "fact" of exploitation settles less than meets the eye. We must also consider the *moral force* of exploitation. In particular, we can ask whether the state should prohibit exploitative transactions or refuse to enforce exploitative agreements. The wrongness of exploitation does not dictate the way in which these moral questions should be answered.

This entry focuses on exploitative transactions or relations rather than "systemic" or macro level exploitation. It also has little to say about the Marxist view of exploitation. There are two major reasons for taking a different tack. First, the moral core of the Marxist view of exploitation is not unique to Marxism. When Marxism claims that the capitalist class exploits the proletariat, it employs the ordinary notion that one party exploits another when it gets unfair and undeserved benefits from its transactions or relationships. On that, the entry will have something to say. Second, what is unique to Marxism -- its approach to *measuring* exploitation through calculations of surplus value -- is very problematic.

This entry does the following. It first surveys the definitional landscape that has been marked out and highlights the conceptual quarrels which have arisen in the literature. It then sketches a very rough and preliminary account of the elements of exploitation. Finally, it makes some brief remarks about the moral force of exploitation.

The Definitional Landscape

At the most general level, A exploits B when A takes unfair advantage of B. (I shall always refer to the alleged exploiter as A and to the alleged exploitee as B). One problem with such a broad account, as Arneson notes, is that there will "be as many competing conceptions of exploitation as theories of what persons owe to each other by way of fair treatment." (Arneson 1992, 350). We can gain a somewhat sharper view of the issues that we must confront if we consider a sampling of the accounts that appear in the literature.

> 1. "[T]o exploit a person involves the *harmful, merely instrumental utilization* of him or his capacities, for one's own advantage or for the sake of one's own ends." (Buchanan 1985, 87).
>
> 2. "It is the fact that the [capitalist's] income is derived through *forced, unpaid, surplus* [wage] labor, *the product of which the workers do not control*, which makes [wage labor] exploitive." (Holmstrom 1997, 357).
>
> 3. "Exploitation necessarily involves benefits or gains of some kind to someone … Exploitation resembles a zero-sum game, viz. what the exploiter gains, the exploitee loses; or, minimally, for the exploiter to gain, the exploitee must lose." (Tormey 1974, 207-08)
>
> 4. "Exploitation [in exchange] demands. . .that there is no reasonably eligible alternative [for the exploitee] and that the consideration or advantage received is incommensurate with the price paid. One is not exploited if one is offered what one desperately needs at a fair and reasonable price." (Benn 1988, 138).
>
> 5. "Exploitation of persons consists in … wrongful behavior [that violates] the moral norm of *protecting the vulnerable*." (Goodin 1988a, 147).
>
> 6. "There are four conditions, all of which must be present if dependencies are to be exploitable. First, the relationship must be *asymmetrical* … Second, … the subordinate party must *need* the resource that the superordinate supplies … Third, … the subordinate party must depend upon some *particular* superordinate for the supply of needed resources … Fourth, the superordinate … enjoys discretionary control over the resources that the subordinate needs from him…" (Goodin 1988b, 37).

7. "Common to all exploitation of one person (B) by another (A) . . . is that A makes a profit or gain by turning some characteristic of B to his own advantage…exploitation … can occur in morally unsavory forms without harming the exploitee's interests and … despite the exploitee's fully voluntary consent to the exploitative behavior. ." (Feinberg 1988, 176-79).

8. "Persons are exploited if (1) others secure a benefit by (2) using them as a tool or resource so as (3) to cause them serious harm." (Munzer 1990, 171)

9. "A society is exploitative when its social structure is organized so that unpaid labor is systematically forced out of one class and put at the disposal of another … On the force-inclusive definition of exploitation, any exploitative society is a form of slavery." (Reiman 1987, 3-4).

10. "[A] group is exploited if it has some conditionally feasible alternative under which its members would be better off." (Roemer 1986, 136)

11. "[E]xploitation is seen as the failure to pay labour its marginal product…" (Brewer 1987, 86).

12. "An exploitative exchange is… an exchange in which the exploited party gets less than the exploiting party, who does better at the exploited party's expense… [T]he exchange must result from social relations of unequal power … exploitation can be entered into voluntarily; and can even, in some sense, be advantageous to the exploited party." (Levine 1988, 66-67).

13. "[Capitalist] social relations … are exploitative, not only in the specific sense of extracting surplus labour, but in the more general sense of using someone as a means, utilizing her to detriment as a way of promoting one's own good… " (Kymlicka 1989, 114).

14. "Workers are exploited if they work longer hours than the number of labor hours employed in the goods they consume." (Elster 1986, 121).

15. "[E]xploitation forms part of an exchange of goods and services when 1) the goods and services exchanged are quite obviously not of equivalent value, and 2) one party to the exchange uses a substantial degree of coercion." (Moore 1973, 53).

16. "[E]xploitation is a psychological, rather than a social or an economic, concept. For an offer to be exploitative, it must serve to create or to take advantage of some recognized psychological vulnerability which, in turn, disturbs the offeree's ability to reason effectively." (Hill 1994, 637).

All these accounts are compatible with the view that "A wrongfully exploits B when A takes unfair advantage of B." But there are some important differences among them. Some accounts (10, 14) are technical definitions of exploitation that are specific to a Marxist approach. Although none of the accounts denies that exploitation requires a gain to the exploiter, only some (3, 8) specifically mention that criterion. Some accounts invoke the Kantian notion that one wrongfully exploits when one treats another instrumentally or merely as a means (1, 8, 13). On some accounts, the exploited party must be harmed (1, 2, 3, 8, 9, 12), whereas other accounts allow that the exploited party may gain from the relationship (4, 7, 11, 12, 15). On some accounts, the exploited party must be coerced (2, 4, 6, 9, 15), whereas others require at least a defect in the quality of the consent (12, 16), and another maintains that exploitation can be fully voluntary (7).

We should not put rigid constraints on what counts as exploitation, at least at the outset. While some exploitative transactions are harmful to the exploitee, we often call exploitative cases in which the exploitee seems to gain from the transaction. Indeed, it is arguable that exploitation would be of much less theoretical interest on a "no harm, no exploitation" rule. It is trivially true that it is wrong for A to gain from an action that unjustifiably harms or coerces B. And even a libertarian will grant that some harmful exploitation may be legitimately prohibited by the state, if only because it is harmful (or rights violating) rather than because it is exploitative. By contrast, it is more difficult to explain when and why it might be wrong for A to gain from an action that benefits B and to which B voluntarily consents. And it is certainly more difficult to explain why society might be justified in prohibiting such transactions or refusing to enforce some such agreements.

For these reasons, it will be useful to make two sets of distinctions. First, we can distinguish between *harmful exploitation* and *mutually advantageous exploitation*. By mutually advantageous exploitation, we refer to those cases in which the exploit*ee* gains from the transaction as well as the exploiter. The advantageousness of the transaction is mutual, not the exploitation. To use somewhat different terminology, exploitation is mutually advantageous only when the transaction is Pareto Superior, that is, a transaction that leaves all parties better off. We can similarly distinguish between *nonconsensual exploitation*, where the exploited party does not give voluntary (or valid) consent, say because of coercion or fraud, and *consensual exploitation*, where it appears that the exploited party has given voluntary and appropriately informed consent to the transaction.

It might be argued that it begs the question to assume that exploitation *can* be mutually advantageous and consensual. The objection fails. If we were to assume, for the sake of argument, that the word exploitation is best limited to cases in which the exploitee is harmed, nothing would have changed. We would *still* have to ask whether there are important distinctions between those cases which are (*ex hypothesi*) wrongly referred to as mutually advantageous *exploitation* and those mutually advantageous transactions that are not described in that way. It would remain an open question as to whether some mutually advantageous arrangements are wrongful and why they are wrongful. If one wants to claim that a mutually advantageous and consensual transaction cannot be *unfair*, then that is not a dispute over language. That is a substantive claim, but there is no reason to think that position is correct.

The Elements of Exploitation

Let us start with the claim that A exploits B when A takes unfair advantage of B. Taking unfair advantage could be understood in two ways. First, it may refer to some dimension of the *outcome* of the exploitative act or transaction, that is, the transaction is substantively unfair. And this, it seems has two elements: (1) the benefit to A and (2) the effect on B. We may say that the benefit to A is unfair because it is wrong for A to benefit at all from his act (e.g. by harming B) or because A's benefit is excessive relative to the benefit to B. Second, to say that A takes unfair advantage of B may imply that there is some sort of defect in the *process* by which the unfair outcome has come about, for example, that A has coerced B or defrauded B or has manipulated B. In the final analysis we may find that these three elements are not all necessary to account for exploitation, but they provide us with a way to begin.

Benefit to A

A cannot take *unfair* advantage of B unless A gets some *advantage* from B. We can see the relevance of the "benefit to A" by contrasting exploitation with other forms of wrongdoing, such as discrimination, abuse, and oppression. Let us say that A discriminates against B when A wrongly deprives B of some opportunity or benefit because of some characteristic of B that is not relevant to A's action. There was a period in American history in which many women became public school teachers because they were denied the opportunity to enter other professions such as law and medicine. To the extent that society benefitted (in one way) from the pool of highly qualified public school teachers, the discrimination may have been exploitative, even if unintentionally so. But if A refuses to hire B solely because of B's race, then it would be odd to say that A exploits B, for A does not gain from the wrong to B.

Consider abuse. It has been alleged that medical students are frequently abused by verbal insults and denigration and that this abuse may leave long-lasting emotional scars. It is also sometimes claimed that medical interns are exploited, that they work long hours for low pay. The contrast is just right. There is no reason to think that anyone gains (in any normal sense) from abuse, but it is at least plausible to think that the hospitals or patients gain from the exploitation of interns.

Let us say that A oppresses B when A deprives B of freedoms or opportunities to which B is entitled. If A gains from the oppressive relationship, as when A enslaves B, then A may both oppress and exploit B. But if A does not gain from the oppression, the oppression is wrong but not exploitative. We might say that the unemployed are oppressed, but unless we could specify the ways in which some gain from their lack of employment, the unemployed are not exploited. Marxists would claim that capitalists pay exploitative wages to the employed precisely because there is a "reserve army" of the unemployed with whom the employed must compete. But that merely confirms that they are exploited because the oppression generates a gain to the capitalist class, and it is the employed who are exploited and not the unemployed that make such exploitation possible.

The Effect on B

As our definitional survey indicated, some commentators maintain that exploitation resembles a zero-sum game, that the exploiter gains what the exploitee loses. (Tormey, 207) Others maintain that exploitation is always harmful to the exploitee, even if the gains and losses do not cancel out. It is relatively uncontroversial that exploitation *can* be harmful to B, as in slavery. Other cases are more controversial. There are cases in which B is not directly affected by A's utilization of B, what Feinberg refers to as *harmless parasitism*, as when A follows B's taillights in a dense fog. A uses B to his own advantage, but does not render B worse off (assume that B is not bothered by A's headlights in B's mirror). (Feinberg, 14) In other cases of non-harmful exploitation, the transaction appears to benefit both A and B, as may be true of organ sales or commercial surrogacy.

Now in asking how A's action affects B's interests, we must be careful to adopt an *all things considered* point of view. There are, after all, negative *elements* in virtually all uncontroversially beneficial transactions. Paying money for a good that is clearly worth the price is still a negative element in the transaction. It would be better to get it for free. If A and B enter into a cooperative agreement where A gives B $100 for a book that is worth a lot to A (because it completes a collection) but is worth little to B, we do not say that B has been harmed by the transaction just because B has lost her book any more than we say that A has been harmed because the transaction required A to pay $100. Similarly, we do not say that a worker is harmed by employment merely because the worker prefers leisure to work. If the benefits to B from employment are greater than the costs to B, then employment is beneficial to B, all things considered. So in deciding whether a case of alleged exploitation should be classified as harmful exploitation or mutually advantageous exploitation, we must look at its net effect on B. If the benefits of a transaction exceed its costs, then it is not harmful even if it is exploitative, as might be true of organ sales and working as a stripper.

Joel Feinberg argues that if a transaction is mutually advantageous, then A does not gain at B's expense. (Feinberg 1988, 178). Not quite. There is an important sense in which any *marginal* gain to one party *within* a "zone of agreement" is always at the other party's expense. For while the parties may prefer any outcome within the zone of agreement to the non-agreement solution, they are not indifferent to the terms of the agreement. Mutually advantageous exploitation occurs when A and B gain relative to the non-cooperation baseline, but where the distribution of the benefits between A and B is unfair to B. Consider a garden variety case of alleged exploitation. An unexpected blizzard hits an area and people rush to the hardware store to buy a shovel. The hardware store owner sees the opportunity to make an abnormal profit and raises the price of a shovel from $15 to $30. If B agrees to pay $30 for the shovel, because the shovel is worth more than $30 to B under the circumstances, then the transaction is advantageous to both parties. If B is exploited, it is because B has paid too much. A similar structure applies to some of the other cases of alleged exploitation with which we began -- AZT for AIDS, surrogacy, organ sales. We need not deny that B benefits from these transactions, all things considered. Rather, A may exploit B if B pays too a high price for what she gains or does not receive enough for what she gives.

A mutually advantageous transaction is arguably (wrongly) exploitative only if the outcome is (in some way) unfair to B. This is not merely definitional. After all, it may be thought that a transaction is exploitative whenever takes advantage of B's vulnerabilities or desperate situation to strike a deal. That

is false. For if A makes a *reasonable* proposal that B has no alternative but to accept given B's desperate situation, A does not exploit B. If a doctor proposes to perform life-saving surgery for a reasonable fee, the patient is hardly exploited, even though the patient would not have agreed but for the fact that her life was in danger.

It might be said that "mutually advantageous" exploitation can and should be understood as a form of harmful exploitation. If we evaluate a transaction by reference to a *fairness baseline* as contrasted with a *no-transaction baseline*, then B is harmed when she pays $30 for a shovel by comparison with the fairness baseline (say, where B pays $15 for a shovel) even if B gains by comparison with the no-transaction baseline. Such relabeling would not change anything, for we would still have to distinguish between those cases in which B is harmed relative to both the fairness baseline and the no-transaction baseline and those cases where B is harmed only by reference to the fairness baseline but not by reference to the no-transaction baseline.

It may also be objected that the proposed distinction between harmful exploitation and mutually advantageous exploitation ignores a deeper -- Kantian -- way in which "mutually advantageous" exploitation is harmful to B, namely that A treats B merely as a means to be utilized to his own advantage rather than as an end in herself -- if so treating a person is to harm her. Allen Buchanan argues that exploitation occurs "whenever persons are harmfully utilized as mere instruments for private gain," and adds that this could apply to business transactions between two affluent bankers --"Each harmfully utilizes the other as a mere means to his own advantage." (Buchanan, 1984, 44.)

It is not clear what to make of this view. First, on one plausible reading of the Kantian maxim, one treats another as a mere means only when one treats "him in a way to which he could not possibly consent," as in cases of coercion and fraud, where A seeks to undermine B's capacity as an autonomous decision-maker. (Korsgaard 1993, 40). There is no reason to think that each banker could not possibly consent to be so treated by the other banker. Second, to say that A exploits B when A "harmfully utilizes" B as a "mere means" is equivocal as to whether "harmfully" is a reinforcing or modifying adverb. On one view, "harmfully" is merely reinforcing because the utilization of B merely as a means *constitutes* an independent harm to B. On another view, "harmfully" is a modifying adverb, because we can contrast the cases in which A harmfully utilizes B as a mere means with cases in which A non-harmfully utilizes B as a mere means. If we accept the first interpretation, we would still want to distinguish between those cases in which B is harmed apart from being treated merely as a means from those in which B is not harmed apart from the harm that derives from being treated merely as a means. On the second view, the bankers may utilize each other as means, but absent an independent form of harm, there is no reason to think that they are harmed by their utilization as a means itself. So that Kantian view does nothing to deny the distinction between harmful exploitation and mutually advantageous exploitation.

And so it seems better simply to grant that some allegedly exploitative transactions are mutually advantageous and go on to ask what makes a mutually advantageous transaction unfair. This is not easy because there is no non-problematic account of fair transactions. (Wertheimer 1996). Here are several possibilities.

We might say that a transaction is unfair when the goods exchanged are "incommensurable," as might be thought of the exchange of an organ for money. There are two problems here. First, it is not clear whether goods are ultimately incommensurable (Chang 1997). Second, if goods are incommensurable, it is not why an exchange of those goods is unfair.

Assuming that we can compare the gains of the parties, it is frequently suggested that a transaction is exploitative when A gains much more than B. But if we measure the parties' gains in terms of marginal utility from the no transaction baseline, the exploitee often gains more than the exploiter. If a doctor overcharges for life-saving surgery, exploiting the patient's situation, the doctor gains less than the patient. If a store owner charges $30 for a shovel, the buyer may well get more utility from the shovel than the seller gets for the money. Indeed, the exploiter's power over the exploitee stems precisely from the fact that he does not stand to gain too much. He can easily walk away from the transaction, whereas the exploitee cannot.

This suggests that we cannot evaluate the fairness of a transaction solely by comparing the gains of the parties. Rather, we must measure the fairness of their gains against a normative baseline as to how much the parties ought to gain, and that baseline is not easy to specify. A promising but not unproblematic candidate is to measure the parties' gains against what they would have gained in a "hypothetical competitive market," where there was relatively complete information. On this view, there is no independent standard of a "just price" for goods such as a shovel or a kidney, nor need we accept whatever the actual market yields, given the market's sundry imperfections. Rather, we evaluate the parties' gains by what they would have received under relatively perfect market conditions, just as we may try to determine the "fair market value" or a home by what the home would sell for under relatively perfect market conditions in that locale.

It might be thought that exploitation (at least when it is morally objectionable) is cofined to cases in which the exploitee is less well-off than the exploiter. Although most cases of exploitation will probably fit this pattern, exploitation is not confined to such cases. We might think, for example, that a store owner who charges an exorbitant price for the snow shovel is exploiting the customer, even if the customer is much wealthier than the store owner. On the present view, exploitation is transaction specific.

Although I am not sure that any available account of fair transactions provides the solution we need, some transactions are intuitively quite unfair even though they are advantageous to both parties. So let us assume that, in principle, some account of unfair transactions can be given. The question now arises as to whether an unfair transaction is always exploitative or whether A exploits B only if there is some defect in the process that culminates in B's decision.

Process

As we have seen, it seems plausible to argue that A does not exploit B simply because there is unfairness in the distribution of rewards. If B voluntarily agrees to what might otherwise be a maldistribution of advantages, as when B voluntarily decides to make a gift of goods or labor to A, then it seems wrong to

say that A has exploited B. It would, for example, be odd (although perhaps not impossible) to claim that a hospital exploits its volunteer workers just because the workers are volunteers rather than paid employees. So it seems that a relationship or transaction is exploitative only if there is some defect in the process by which it came about.

Interestingly, both Marxists and libertarians accept the view that voluntary transactions cannot be exploitative. Marxists tend to adopt a "force inclusive definition" of exploitation. Marxists do not say that capitalists exploit their workers *in spite* of the fact that the workers voluntarily agree to their employment status. They argue that workers are exploited only because they do not voluntarily agree to their employment status. Marxists concede that the proletariat is not enslaved, because they are not tied to any particular employer, but they transfer their labor to the capitalist under the "dull compulsion of economic relations." (Elster 1983, 277-78). Libertarians can be understood as accepting this "force inclusive" definition of exploitation, but come to the opposite conclusion. They maintain that since market transactions are not coerced, the workers are therefore not exploited. We do not need to accept these alternatives. Leaving aside just how to distinguish between nonconsensual and consensual exploitation, A can arguably exploit B even if B is not coerced (or defrauded), even if there is nothing untoward about B’s decision within her objective situation.

Let us press this issue a bit further. There are some instances of alleged exploitation in which the issue of consent does not seem to arise. There are cases in which the exploitee may be entirely passive. A may sell photographs of B without B’s knowledge, or rob a purse from a sleeping B or follow B’s taillights in a dense fog. In these cases, B’s will is not involved. Call this nonvolitional exploitation. If nonvolitional exploitation operates *without* the engagement of B’s will, then nonconsensual exploitation operates *against* B’s will, as when A coerces B or deceives B. The question now arises as to when there is such a procedural defect.

In general, A coerces B to do X only if A proposes (threatens) to make B worse off with reference to some baseline condition if B chooses not do X, although specifying the appropriate baseline against which to measure the proposal can be a complicated matter. (Wertheimer, 1987) If A gets B to pay A $100 per week by threatening to bomb B’s store if he does not pay up, then A coerces B into paying $100 a week. By contrast, if A gets B to pay A $100 per week by proposing to clean B’s store each night, then A has made a non-coercive (or inducive) offer to B. A does not propose to worsen B’s situation if B rejects A’s proposal. On this view, A does not coerce B in the cases involving organ sales or commercial surrogacy, because A does not propose to worsen B’s situation if B rejects A’s proposal.

Fraud also undermines the validity of B’s consent. Suppose that A offers to sell B a car for $10,000. A tells B that the car has been driven only 50,000 miles, but has set back the odometer from 90,000 to 50,000. B has not given valid consent, because valid consent must be informed (or not misinformed) as well as uncoerced.

By contrast with cases of coercion and fraud, there are at least some cases of alleged exploitation in which B’s consent is not defective in either of these ways. In many cases of alleged exploitation, A gets

B to agree to a mutually advantageous transaction to which B would not have agreed under better or perhaps more just background conditions, where A has played no direct causal role in creating those circumstances, where A has no special obligation to repair those conditions, and where B is fully informed as to the consequences of various choices. Although B might prefer to have a different range of options available to him, she can make a perfectly rational choice among the various options. Such conditions may (or may not) obtain in cases such as commercial surrogacy, organ sales, and, say, our snow shovel case.

It might be objected that perfectly rational and (otherwise) uncoerced choices are not appropriately consensual if made under conditions of desperation or from an inequality of bargaining power, or under unjust background conditions. But even if we refer to such transactions as nonconsensual, we would still have to contrast the cases that are nonconsensual because of coercion or fraud and those that are allegedly nonconsensual in other ways. And we will still have to ask what the moral force of such exploitation amounts to: Should we prohibit A from making such proposals? Should we refuse to enforce agreements made under such conditions? And that brings us to the moral force of exploitation.

The Moral Force of Exploitation

I have suggested that exploitation provides a moral description of a transaction, but that its moral force is less clear. The moral force of harmful and nonconsensual exploitation is relatively unproblematic. Whatever the added moral importance of the gain to A from the harm to B, it is certainly at least *prima facie* wrong for A to harm B and it seems that the state is at least *prima facie* justified in prohibiting or refusing to enforce such transactions.

Mutually advantageous transactions present a more difficult set of problems. Even if a transaction between A and B is unfair, it might be thought that there can be nothing *seriously* wrong about an agreement from which both parties benefit, particularly if A has no obligation to enter into any transaction with B. Moreover, even if there is something seriously wrong about such transactions, it might be argued that it could not be a wrong that would justify state intervention. Recall the snow shovel case. Even if A acts wrongly or fails to act virtuously, it is arguable that A does not harm anyone or violate anyone's rights, and only harm or rights violations justify state intervention. If the state cannot force A to sell the shovel to B, it might be thought completely *irrational* for the state to prohibit A and B from entering into a consensual and mutually advantageous transaction.

Perhaps this view is correct. Bracketing arguments based on externalities, it seems perfectly plausible to maintain that the state is justified in interfering with transactions only if one party is violating the other's rights. That said, those who invoke the concept of exploitation frequently maintain that such exploitation provides a reason for state intervention. For example, when it is claimed that commercial surrogacy exploits the birth mothers, the critics typically argue that surrogacy contracts should be unenforceable or entirely prohibited. Similar things are said about the sale of bodily organs. Those who make such arguments do frequently claim that the transactions are nonconsensual or harmful, but they seem prepared to make such arguments even if the transactions are consensual and mutually advantageous.

On what grounds might we justify interfering with consensual and mutually advantageous exploitative transactions? It might be thought that we could interfere on paternalistic grounds. A paternalistic argument could not justify interfering with exploitative transactions if the exploitative transaction is advantageous to the exploitee and if interference is not likely to result in a transaction that is more beneficial to B. For paternalism justifies interfering for someone's good, and this interference would not be to the target's benefit. But there might be situations in which B knows enough to agree only to those exploitative transactions that are beneficial (as compared with no transaction), but does not know that less exploitative transactions are available. And so there may be a "soft paternalist" justification for interference with some mutually advantageous exploitative transactions.

We might also justify interfering with exploitative transactions on strategic grounds. Suppose that A enjoys a monopoly position, say, as a potential rescuer of B. If we prohibit A from charging an exorbitant price for his services, then A might offer his services for a reasonable price. This argument would not justify interfering in a highly competitive market, for, under such conditions, A would not and could not offer his services for a better price. But there may be numerous situations in which such strategic arguments can work. (Wertheimer 1996).

Can we justify interfering with mutually advantageous and consensual transactions on perfectionist or moralistic grounds? That is more difficult. Joel Feinberg has maintained that because mutually advantageous exploitation is not harmful, such exploitation would constitute a "free-floating evil," a wrong that is bad for no one. "In these cases there is no wrongful loss for the exploitee, who can himself have no grievance." (Feinberg 1988, 176). There are two questions here: is mutually advantageous exploitation a free-floating evil? and free-floating or not, can we justify interfering with immoral transactions on the ground that they are immoral?

Is mutually advantageous exploitation a free-floating evil? I think not. Suppose that B and C both need blood transfusions, and that the only available blood is compatible with B's blood type, but not with C's blood type. There are only two possible worlds: (1) No Transfusion, where neither B nor C gets a transfusion; (2) Transfusion, where B gets the transfusion and C does not. In this case, to say that it would be wrong -- in any way -- to give the transfusion to B would seem to involve a "free-floating evil." Giving the transfusion to B is good for B and bad for no one, including C, for there is no feasible alternative world in which C could have gained.

But the world of mutually advantageous exploitation is not like this. Recall the "snow shovel" example. Here there are, let us say, three feasible alternative worlds: (1) Transaction 1, where A sells B a shovel for an exorbitant $30; (2) Transaction 2, where A sells B a shovel for the normal price of $15; (3) no transaction. Is Transaction 1 better for A and worse for no one? Yes and No. Yes, when compared to the No Transaction baseline. No, when compared with Transaction 2. By comparison with Transaction 2, the "wrong" in Transaction 1 is not free floating. B is not harmed in Transaction 1, but B's interests are clearly negatively affected by A's choice to engage in Transaction 1 as contrasted with Transaction 2. To say that the wrong involved in mutually advantageous exploitation is not free floating does not establish its moral force. There might be good reason for the state to stay its hand. But even if there are good

reasons not to interfere with most cases of mutually advantageous exploitation, it does not follow that exploitation is morally trivial. The disposition not to take unfair advantage of one's fellows may be among the more important moral virtues and a necessary condition of civilized life, even if there are also good reasons for not penalizing the failure to display that virtue.

Even if exploitation is seriously wrong, it may not be the worst form of injustice or inequality. Suppose that social justice requires a relatively egalitarian distribution of resources. If an inequality between A and B is exploitative only if there is some causal relationship between A's and B's positions, then many injustices will have nothing to do with exploitation. Although it may be unjust that A has much more than B, A's having more may have nothing to do with B's having little. Consider the exploitation of labor. Some people are more productive than others, albeit often because of morally irrelevant factors, such as social background or native talents.

If we exploit people when we fail to reward them in proportion to their productive contribution, then the low contributors may not be exploited. (Nagel 1991, 99-100) Indeed, if the high contributors are taxed to provide for the needs of the low contributors, they might maintain that it is they who are being exploited. Consider the unemployed. If the unemployed are not exploited because they "do not produce any surplus value for the capitalist to appropriate," we may well conclude that being excluded from the labor system is much worse than being exploited in the system. (Kymlicka, 176). Still, even if exploitative inequality is not always morally worse than non-exploitative inequality, it is an interesting and important question as to whether and in what ways the inequalities and suffering that arise from exploitation have a special call on our moral attention.

Bibliography

- Arneson, R., 1992, "Exploitation," in Lawrence C. Becker (ed.) *Encyclopedia of Ethics*, New York: Garland, pp. 350-52.
- Benn, S., 1988, *A Theory of Freedom*, Cambridge: Cambridge University Press
- Brewer, J., 1987, "Exploitation in the New Marxism of Collective Action," *Sociological Review*, 35, pp. 84-96.
- Buchanan, A., 1985, *Ethics, Efficiency, and the Market*, Totowa: N.J.: Rowman and Allanheld
- ------, 1984, *Marx and Justice*, Totowa: N.J.: Rowman and Allanheld
- Chang, R. (ed.), 1997, *Incommensurability, Incomparability, and Practical Reason*, Cambridge: Harvard University Press
- Elster, J., 1983, "Exploitation, Freedom, and Justice," in J. Roland Pennock and John Chapman (eds.) *Nomos XXVI: Marxism*, New York: New York University Press, pp. 227-52.
- ------ (ed.), 1986, *Karl Marx: A Reader*, Cambridge: Cambridge University Press
- Feinberg, J., 1988, *Harmless Wrongdoing*, Oxford: Oxford University Press
- Field, M., 1989, *Surrogate Motherhood*, Cambridge: Harvard University Press
- Goodin, R., 1988a, *Reasons for Welfare*, Princeton: Princeton University Press
- ------, 1988b, "Reasons for Welfare: Economic, Sociological, and Political -- But Ultimately Moral," in J. Donald Moon (ed.) *Responsibility, Rights, and Welfare*, Boulder, Co.: Westview

Press, pp. 19-54.

- Hill, J. L., 1994, "Exploitation," *Cornell Law Review*, 79, pp. 631-99.
- Holsmstrom, N., 1997, "Exploitation," *Canadian Journal of Philosophy*, 7, pp. 353-69.
- Kennedy, D., 1990, "So What if College Players Turn Pro Early?" *New York Times*, January 19, 1990, Section B, p. 7
- Korsgaard, C., 1993, "The Reasons We Can Share" *Social Philosophy and Policy*, 10, pp. 24-51.
- Kymlicka, W., 1989, *Liberalism, Community and Culture*, Oxford: Clarendon Press
- Levine, A., 1988, *Arguing for Socialism*, London: Verso
- Moore, B., 1973, *Reflections on the Causes of Human Misery*, Boston: Beacon Press
- Munzer, S., 1990, *A Theory of Property*, Cambridge: Cambridge University Press
- Nagel, T., 1991, *Equality and Partiality*, New York: Oxford University Press
- Nozick, R., 1974, *Anarchy, State, and Utopia*, New York: Basic Books
- Okin, S., 1989, *Justice, Gender and the Family*, New York: Basic Books
- Rawls, J., 1971, *A Theory of Justice*, Cambridge: Harvard University Press
- Reiman, J., 1987, "Exploitation, Force, and the Moral Assessment of Capitalism: Thoughts on Roemer and Cohen," *Philosophy and Public Affairs*, 16, pp. 3-41.
- Roemer, J., 1985, "Should Marxists Be Interested in Exploitation?" *Philosophy and Public Affairs*, 14, pp. 30-65.
- ------, 1986, "An Historical Materialist Alternative to Welfarism," in J. Elster and A. Hylland (eds.), *Foundations of Social Choice Theory*, Cambridge: Cambridge University Press
- Temkin, L., 1993, *Equality*, New York: Oxford University Press
- Tormey, J. F., 1974, "Exploitation, Oppression and Self-Sacrifice," *Philosophical Forum*, 5, pp. 206-21.
- Wertheimer, A., 1987, *Coercion*, Princeton: Princeton University Press
- ------, 1996, *Exploitation*, Princeton: Princeton University Press
- Wood, A., 1995, "Exploitation," *Social Philosophy and Policy*, 12, pp. 136-58.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

equality | justice: distributive | Marxism

Alan Wertheimer
alan.wertheimer@uvm.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 20, 2001
Content last modified: December 20, 2001

Non-Monotonic Logic

The term "non-monotonic logic" covers a family of formal frameworks devised to capture and represent *defeasible inference*, i.e., that kind of inference of everyday life in which reasoners draw conclusions tentatively, reserving the right to retract them in the light of further information. Such inferences are called "non-monotonic" because the set of conclusions warranted on the basis of a given knowledge base does not increase (in fact, it can shrink) with the size of the knowledge base itself. This is in contrast to classical (first-order) logic, whose inferences, being deductively valid, can never be "undone" by new information.

Abstract Consequence relations

Classical first-order logic (henceforth: FOL) is monotonic: if a sentence φ can be inferred in FOL from a set Γ of premises, then it can also be inferred from any set Δ of premises containing Γ as a subset. In other words, FOL provides a relation $\models$ of *consequence* between sets of premises and single sentences with the property that if $\Gamma \models \varphi$ and $\Gamma \subseteq \Delta$, then $\Delta \models \varphi$. This follows immediately from the nature of the relation $\models$, for $\Gamma \models \varphi$ holds precisely when φ is true on every interpretation on which all sentences in Γ are true (see the entry on classical logic for details on the relation $\models$).

From an abstract point of view, we can consider the formal properties a consequence relation. Let $\mid\!\sim$ be any relation between sets of premises and single sentences. It is natural to consider the following properties, all of which are satisfied by the consequence relation $\models$ of FOL:

- **Supraclassicality**: if $\Gamma \models \varphi$ then $\Gamma \mid\!\sim \varphi$.

- **Reflexivity**: if $\varphi \in \Gamma$ then $\Gamma \mathrel{|\!\!\sim} \varphi$;
- **Cut**: If $\Gamma \mathrel{|\!\!\sim} \varphi$ and $\Gamma, \varphi \mathrel{|\!\!\sim} \psi$ then $\Gamma \mathrel{|\!\!\sim} \psi$;
- **Monotony**: If $\Gamma \mathrel{|\!\!\sim} \varphi$ and $\Gamma \subseteq \Delta$ then $\Delta \mathrel{|\!\!\sim} \varphi$.

Supraclassicality just requires that $\mathrel{|\!\!\sim}$ be an extension of $\models$, i.e., that if φ follows from Γ in FOL, then it must also follow according to $\mathrel{|\!\!\sim}$. (The relation $\models$ is trivially supraclassical).

The most straightforward of the remaining conditions is reflexivity: we certainly would like all sentences in Γ to be inferable from Γ. This translates to the requirement that if φ belongs to the set Γ, then φ is a consequence of Γ. Reflexivity is a rather minimal requirement on a relation of logical consequence: It is hard to imagine in what sense a relation that fails to satisfy reflexivity, can still be considered considered a *consequence* relation.

Cut, a form of transitivity, is another crucial feature of consequence relations. Cut is a conservativity principle: if φ is a consequence of Γ, then ψ is a consequence of Γ together with φ only if it is already a consequence of Γ alone. In other words, by adjoining to Γ something which is already a consequence of Γ does not lead to any *increase* in inferential power.

Cut is best regarded as a condition on the "length" of a proof (where "length" is to be understood in some intuitive sense related to the complexity of the argument supporting a given conclusion). When viewed in these terms, Cut amounts to the requirement that the length of the proof does not affect the degree to which the assumptions support the conclusion. Where φ is already a consequence of Γ, if ψ can be inferred from Γ together with φ, then ψ can also be obtained via a longer "proof" that proceeds indirectly by first inferring φ.

In many forms of probabilistic reasoning the degree to which the premises support the conclusion is inversely correlated to the length of the proof. For this reason, many forms of probabilistic reasoning fail to satisfy Cut. To see this, we adapt a well-known counter-example: let *Ax* abbreviate "*x* is a Pennsylvania Dutch", *Bx* abbreviate "*x* is a native speaker of German", and *Cx* abbreviate "*x* was born in Germany". Further, let Γ comprise the statements "Most *A*s are *B*s," "Most *B*s are *C*s," and *Ax*. Then Γ supports *Bx*, and Γ together with *Bx* supports *Cx*, but Γ by itself does not support *Cx*. (Here statements of the form "Most *A*s are *B*s" are interpreted probabilistically, as saying that the conditional probability of *B* given *A* is, say, greater that 50%.) To the extent that Cut is a necessary feature of a well-behaved consequence relation, examples of inductive reasoning such as the one just given cast some doubt on the possibility of coming up with a well-behaved relation of probabilistic consequence.

For our purposes, Monotony is the central property. Monotony states that if φ is a consequence of Γ then it is also a consequence of any set containing Γ (as a subset). The import of monotony is that one cannot pre-empt conclusions by adding new premises. However, there are many inferences typical of everyday (as opposed to mathematical or formal) reasoning, that do not satisfy monotony. These are cases in which we reach our conclusions *defeasibly* (i.e., tentatively), reserving the right to retract them in the light of

further information. Perhaps the clearest examples are derived from legal reasoning, in which defeasible assumptions abound. In the judicial system, the principle of *presumption of innocence* leads us to infer (defeasibly) from the fact that *x* is to stand trial, the conclusion that *x* is innocent; but clearly the conclusion can be retracted in the light of further information.

Other examples are driven by typicality considerations. If being a B is a typical trait of A's, then from the fact that *x* is an A we infer the conclusion that *x* is a B. But the conclusion is defeasible, in that B-ness is not a necessary trait of A's, but only (perhaps) of 90% of them. For example, mammals, by and large, don't fly, so that lack of flight can be considered a typical trait of mammals. Thus, when supplied with information that *x* is a mammal, we naturally infer that *x* does not fly. But this conclusion is defeasible, and can be undermined by the acquisition of new information, for example by the information that *x* is a bat. This inferential process can be further iterated. We can learn, for instance, that *x* is a baby bat, that does not know how to fly yet.

Defeasible reasoning, not unlike FOL, can follow complex patterns. However, such patterns are beyond reach for FOL, which is, by its very nature, monotonic. The challenge then is to provide for defeasible reasoning what FOL provides for formal or mathematical reasoning, i.e., an account that is both materially adequate and formally precise.

The conclusion of the preceding discussion is that monotony has to be abandoned, if we want to give a formal account of these patterns of defeasible reasoning. Then the question naturally arises of what formal properties of the consequence relation are to replace monotony. Two such properties have been considered in the literature, for an arbitrary consequence relation $\mid\sim$:

- **Cautious Monotony**: If $\Gamma \mid\sim \varphi$ and $\Gamma \mid\sim \psi$, then $\Gamma, \varphi \mid\sim \psi$.
- **Rational Monotony**: If it's not the case that $\Gamma \mid\sim \neg \varphi$, and moreover $\Gamma \mid\sim \psi$, then $\Gamma, \varphi \mid\sim \psi$.

Both Cautious Monotony and the stronger principle of Rational Monotony are special cases of Monotony, and are therefore not in the foreground as long as we restrict ourselves to the classical consequence relation $\models$ of FOL.

Although superficially similar, these principles are in reality quite different. Cautious Monotony is the converse of Cut: it states that adding a consequence φ back into the premise-set Γ does not lead to any *decrease* in inferential power. Cautious Monotony tells us that inference is a cumulative enterprise: we can keep drawing consequences that can in turn be used as additional premises, without affecting the set of conclusion. Together with Cut, Cautious Monotony says that if φ is a consequence of Γ then for any proposition ψ, ψ is a consequence of Γ if and only if it is a consequence of Γ together with φ. It has been often pointed out, most notably by Dov Gabbay, that Reflexivity, Cut and Cautious Monotony are critical properties for any well-behaved non-monotonic consequence relation.

The status of Rational Monotony is much more problematic. As we observed, Rational Monotony can be regarded as a strengthening of Cautious Monotony, and like the latter it is a special case of Monotony.

However, there are reason to think that Rational Monotony might not be a correct feature of a non-monotonic consequence relation. A counter-example due to Robert Stalnaker (see Stalnaker (1994)) involves three composers: Verdi, Bizet, and Satie. Suppose that we initially accept (correctly but defeasibly) that Verdi is Italian, while Bizet and Satie are French. We assume that these defeasible conclusions are built into whatever inferential mechanism implements the non-monotonic relation $\mid\!\sim$.

Suppose now that we are told by a reliable source of information that that Verdi and Bizet are compatriots. This lead us no longer to endorse the propositions that Verdi is Italian (because he could be French), and that Bizet is French (because he could be Italian); but we would still draw the defeasible consequence that Satie is French, since nothing that we have learned conflicts with it. By letting I(v), F(b), and F(s) represent our initial beliefs about the composers' nationalities, and C(v,b) represent the proposition that Verdi and Bizet are compatriots, the situation could be represented as follows:

$$C(v,b) \mid\!\sim F(s),$$

Now consider the proposition C(v,s) that Verdi and Satie are compatriots. Before learning that C(v,b) we would be inclined to reject the proposition C(v,s) because we endorse and I(v) and F(s), but after learning that Verdi and Bizet are compatriots, we can no longer endorse I(v), and therefore no longer reject C(v,s). Then the following *fails*:

$$C(v,b) \mid\!\sim \neg C(v,s).$$

However, if we added C(v,s) to our stock of beliefs, we would lose the inference to F(s): in the context of C(v,b), the proposition C(v,s) is equivalent to the statement that all three composers have the same nationality. This leads us to suspend our assent to the proposition F(s). In other words, and contrary to Rational Monotony, the following also *fails*:

$$C(v,b), C(v,s) \mid\!\sim F(s).$$

The previous discussion gives a rather clear picture of the desirable features of a non-monotonic consequence relation. Such a relation should satisfy Supraclassicality, Reflexivity, Cut, and Cautious Monotony.

Skeptical or credulous?

A separate issue from the formal properties of a non-monotonic consequence relation, although one that is strictly intertwined with it, is the issue of how *conflicts* between potential defeasible conclusions are to be handled.

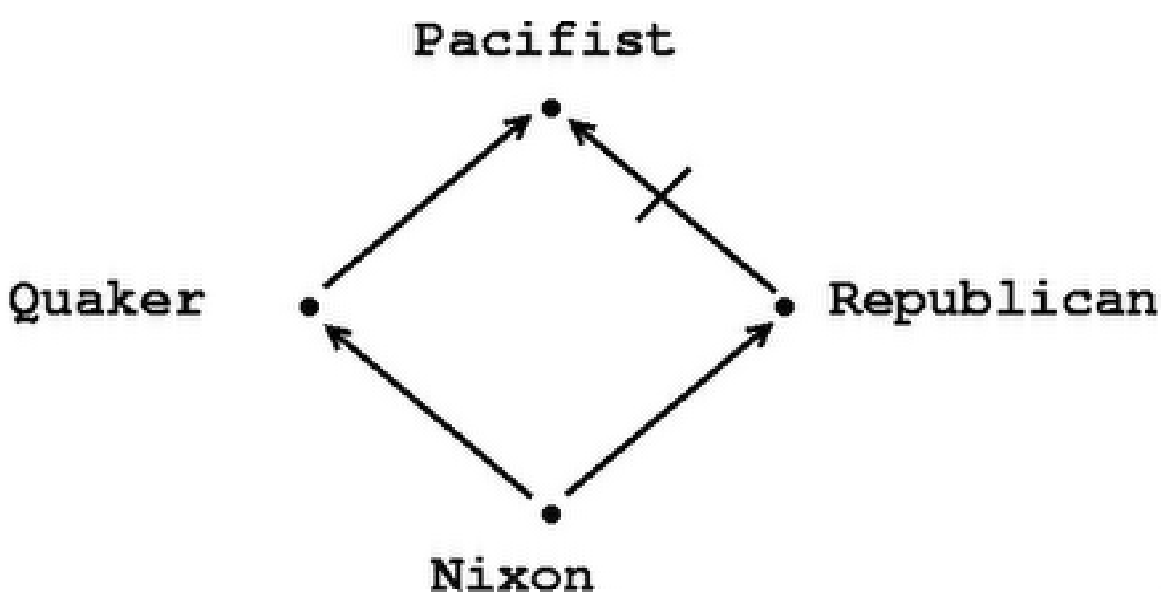
There are two different kinds of conflicts that can arise within a given non-monotonic framework: (*i*) conflicts between defeasible conclusions and "hard facts;" and (*ii*) conflicts between one potential

defeasible conclusion and another (many formalisms, for instance, provide some form of defeasible inference rules, and such rules might have conflicting conclusions). When a conflict (of either kind) arises, steps have to be taken to preserve or restore consistency.

All non-monotonic logics handle conflicts of the first kind in the same way: indeed, it is the very essence of defeasible reasoning that conclusions can be retracted when new facts are learned. But conflicts of the second kind can be handled in two different ways: one can draw inferences either in a "cautious" or "bold" fashion (also known as "skeptical" or, respectively, "credulous"). These two options correspond to significantly different ways to construe a given body of defeasible knowledge, and yield different results as to what defeasible conclusions are warranted on the basis of such a knowledge base.

The difference between these basic attitudes comes to this. In the presence of potentially conflicting defeasible inferences (and in the absence of further considerations such as specificity -- see below), the credulous reasoner always commits to as many defeasible conclusions as possible, subject to a consistency requirement, whereas the skeptical reasoner withholds assent from potentially conflicted defeasible conclusions.

A famous example from the literature, the so-called "Nixon diamond," will help make the distinction clear. Suppose our knowledge base contains (defeasible) information to the effect that a given individual, Nixon, is both a Quaker and a Republican. Quakers, by and large, are pacifists, whereas Republicans, by and large are not. The question is what defeasible conclusions are warranted on the basis of this body of knowledge, and in particular whether we should infer that Nixon is a pacifist or that he is not pacifist. The following figure provides a schematic representation of this state of affairs in the form a (defeasible) network:



The credulous reasoner has no reason to prefer either conclusion ("Nixon is a pacifist;" "Nixon is not a pacifist") to the other one, but will definitely commit to one or the other. The skeptical reasoner recognizes that this is a conflict not between hard facts and defeasible inferences, but between two different defeasible inferences. Since the two possible inferences in some sense "cancel out," the skeptical

reasoner will refrain from drawing either one.

Whereas many of the early formulations of defeasible reasoning have been credulous, skepticism has gradually emerged as a viable alternative, which can, at times, be better behaved. Arguments have been given in favor of both skeptical and credulous inference. Some have argued that credulity seems better to capture a certain class of intuitions, while others have objected that although a certain degree of "jumping to conclusion" is by definition built into any non-monotonic formalism, such jumping to conclusions needs to be regimented, and that skepticism provides precisely the required regimentation. (A further issue in the skeptical/credulous debate is the question of whether so-called "floating conclusions" should be allowed; see Horty (forthcoming) for a review of the literature and a substantial argument that they should not.)

Non-monotonic formalisms

One of the most significant developments both in logic and artificial intelligence, is the emergence of a number of non-monotonic formalisms, devised expressly for the purpose of capturing defeasible reasoning in a mathematically precise manner. The fact that patterns of defeasible reasoning have been accounted for in such a rigorous fashion has wide-ranging consequences for our conceptual understanding of argumentation and inference.

Pioneering work in the field of non-monotonic logics began with the realization that ordinary first-order logic is inadequate for the representation of defeasible reasoning accompanied by the effort to reproduce the success of FOL in the representation of mathematical, or formal, reasoning. Among the pioneers of the field in the late 1970's were (among others) J. McCarthy, D. McDermott & J. Doyle, and R. Reiter (see Ginsberg (1987) for a collection of early papers in the field and Gabbay et al (1994) for a more recent collection of excellent survey papers). In 1980, the *Artificial Intelligence Journal* published an issue (vol. 13, 1980) dedicated to these new formalisms, an event that has come to be regarded as the "coming of age" of non-monotonic logic.

If one of the goals of non-monotonic logic is to provide a materially adequate account of defeasible reasoning, it is important to rely on a rich supply of examples to guide and hone intuitions. Database theory was one of the earliest sources of such examples, especially as regards the *closed world assumption*. Suppose a travel agent has access to a flight database, and needs to answer a client's query about the best way to get from Oshkosh to Minsk. The agents queries the database and, not surprisingly, responds that there are no direct flights. How does the travel agent know?

It is quite clear that, in a strong sense of "know", the travel agent does not *know* that there are no such flights. What is at work here is a tacit assumption that the database is *complete*, and that since the database does not list any direct flights between the two cities, then there are none. A useful way to look at this process is as a kind of *minimization*, i.e., an attempt to minimize the extension of a given predicate ("flight-between," in this case). Moreover, on pain of inconsistencies, such a minimization needs to take place not with respect to what the database explicitly contains but with respect to what it implies.

The idea of minimization is at the basis of one of the earliest non-monotonic formalisms, McCarthy's *circumscription*. Circumscription makes explicit the intuition that, all other things being equal, extensions of predicates should be *minimal*. Again, consider principles such as "all normal birds fly". Implicit in this principle is the idea that the extension of the abnormality predicate should be minimal, and that specimens should not be considered to be abnormal unless there is positive information to that effect. McCarthy's idea was to represent this formally, using second-order logic (SOL). In SOL, in contrast to FOL, one is allowed to explicitly quantify over predicates, forming sentences such as $\exists P \forall x Px$ ("there is a universal predicate") or $\forall P(Pa \leftrightarrow Pb)$ ("*a* and *b* are indiscernible").

In circumscription, given predicates *P* and *Q*, we abbreviate $\forall x(Px \rightarrow Qx)$ as $P \leq Q$; similarly, we abbreviate $P \leq Q$ & $\neg Q \leq P$ as $P<Q$. If $A(P)$ is a formula containing occurrences of a predicate *P*, then the circumscription of *P* in *A* is the second-order sentence $A^*(P)$:

$$A(P) \mathbin{\&} \neg \exists Q\, [A(Q) \mathbin{\&} Q < P\,]$$

$A^*(P)$ says that *P* satisfies *A*, and that no smaller predicate does. Let *Px* be the predicate "*x* is abnormal," and let $A(P)$ be the sentence "All normal birds fly." Then the sentence "Tweety is a bird," together with $A^*(P)$ implies "Tweety flies," for the circumscription axiom forces the extension of *P* to be empty, so that "Tweety is normal" is automatically true.

In terms of consequence relations, circumscription allows us to define, for each predicate *P*, a non-monotonic relation $A(P) \mid\!\sim \varphi$ that holds precisely when $A^*(P) \models \varphi$. (This basic form of circumscription has been generalized, for in practice, one needs to minimize the extension of a predicate, while allowing the extension of certain other predicates to vary.) From the point of view of applications, however, circumscription has a major computational shortcoming, which is due to the nature of second-order language in which circumscription is formulated. The problem is that SOL, contrary to FOL, lacks a complete inference procedure: the price one pays for the greater expressive power of second-order logic is that there are no complete axiomatizations, as we have for FOL. It follows that there is no way to list, in an effective manner, all SOL validities, and hence to determine whether $A(P) \mid\!\sim \varphi$.

Another non-monotonic formalism inspired by the intuition of minimization of abnormalities is *non-monotonic inheritance*. Whenever we have a taxonomically organized body of knowledge, we presuppose that subclasses inherit properties from their superclasses: dogs have lungs because they are mammals, and mammals have lungs. However, there can be exceptions, which can interact in complex ways. To use an example already introduced, mammals, by and large, don't fly; since bats are mammals, in the absence of any information to the contrary, we are justified in inferring that bats do not fly. But then we learn that bats are exceptional mammals, in that they do fly: the conclusion that they don't fly is retracted, and the conclusion that they fly is drawn instead. Things can be more complicated still, for in turn, as we have seen, baby bats are exceptional bats, in that they do not fly (does that make them unexceptional mammals?). Here we have potentially conflicting inferences. When we infer that Stellaluna, being a baby bat, does not fly, we are resolving all these potential conflicts based on a *specificity* principle: more

specific information overrides more generic information.

Non-monotonic inheritance networks were developed for the purpose of capturing taxonomic examples such as the above. Such networks are collections of nodes and directed ("IS-A") links representing taxonomic information. When exceptions are allowed, the network is interpreted *defeasibly*. The following figure gives a network representing this state of affair:

stellaluna diagram

In such a network, links of the form $A \rightarrow B$, represent the fact that, typically and for the most part, A's are B's, and that therefore information about A's is more specific than information about B's. More specific information overrides more generic information. Research on non-monotonic inheritance focuses on the different ways in which one can make this idea precise.

The main issue in defeasible inheritance is to characterize the set of assertions that are supported by a given network. It is of course not enough to devise a representational formalism, one also needs to specify how the formalism is to be interpreted. Such a characterization is accomplished through the notion of *extension* of a given network. There are two competing characterizations of extension for this kind of networks, one that follows the credulous strategy and one that follows the skeptical one. Both proceed by first defining the *degree* of path through the network as the length of the longest sequence of links connecting its endpoints; extensions are then constructed by considering paths in ascending order of their degrees. We are not going to review the details, since many of the same issues arise in connection with default logic (which is treated to greater length below), but Horty (1994) provides an extensive survey. It is worth mentioning that since the notion of degree makes sense only in the case of acyclic networks, special issues arise when networks contain cycles (see Antonelli (1997) for a treatment of inheritance on cyclic networks).

Although the language of non-monotonic networks is expressively limited by design (in that only links of the form "IS-A" can be represented in a natural fashion), such networks provide an extremely useful setting in which to test and hone one's intuitions and methods for handling defeasible information. Such intuitions and methods are then applied to more expressive formalisms. Among the latter is Reiter's *Default Logic*, perhaps the most flexible among non-monotonic frameworks.

In Default Logic, the main representational tool is that of a *default rule*, or simply a *default*. A default is a *defeasible inference rule* of the form

$$\frac{\gamma : \theta}{\tau}$$

(where γ, θ, τ are sentences in a given language, respectively called the pre-requisite, the justification

and the conclusion of the default). The interpretation of the default is that if γ is known, and there is no evidence that θ might be false, then τ can be inferred.

As is clear, application of the rule requires that a consistency condition be satisfied. What makes meeting the condition complicated is the fact that rules can interact in complex ways. In particular, it is possible that application of some rule might cause the consistency condition to fail for some, not necessarily distinct, rule. For instance, if θ is $\neg\tau$ then application of the rule is in a sense self-defeating, in that application of the rule itself causes the consistency condition to fail.

Reiter's default logic uses the notion of an *extension* to make precise the idea that the consistency condition has to be met both before and after the rule is applied. Given a set Γ of defaults, an extension for Γ represents a set of inferences that can be *reasonably* and *consistently* drawn using defaults from Γ. Such inferences are those that are warranted on the basis of a maximal set of defaults whose consistency condition is met both before and after their being triggered.

More in particular (and in typical circular fashion), an extension for Γ is a maximal subset Δ of Γ the conclusions of whose defaults both imply all the pre-requisites of defaults in Γ and are consistent with all the justifications of defaults in Γ. This definition can be made precise as follows. By a *default theory* we mean a pair (W,Δ), where Δ is a (finite) set of defaults, and W is a set of sentences (a world description). The idea is that W represents the strict or background information, whereas Δ specifies the defeasible information. Given a pair (T_1,T_2) of sets of sentences, a default such as the above is *triggered* by (T_1,T_2) if and only if $T_1 \models \gamma$ and it's not the case that $T_2 \models \neg\theta$ (i.e., θ is consistent with T_2). (Notice how this definition is built "on top" of $\models$: we could, conceivably, employ a different relation here.)

Finally we say that a set of sentences E is an *extension* for a default theory (W,Δ) if and only if

$$E = E_0 \cup E_1 \cup \ldots \cup E_n \cup \ldots ,$$

where: $E_0 = W$, and

$$E_{n+1} = E_n \cup \{ \tau : (\gamma : \theta) / \tau \in \Delta \text{ is triggered by } (E_n, E) \}.$$

It is important to notice the occurrence of the limit E in the definition of E_{n+1}): the condition above is not a garden-variety recursive defintion, but a truly circular characterization of extensions.

This circularity can be made explicit by giving an equivalent definition of extension as solutions of fixpoint equations. Given a default theory, let **S** be an operator defined on set of sentences such that for any set S of sentences, $\mathbf{S}(S)$ is the smallest set containing W, deductively closed (i.e., such that if $\mathbf{S}(S) \models \varphi$ then $\varphi \in \mathbf{S}(S)$), and such that if a default with consequent τ is triggered by (S,S) then $\tau \in \mathbf{S}(S)$. Then one can show that E is an extension for (W,Δ) if and only if E is a fixed point of **S**, i.e., if $\mathbf{S}(E) = E$.

For any given default theory, extensions need not exist, and even when they exist, they need not be unique. Let us consider a couple of examples. Our first example is a default theory that has no extension: let *W* contain the sentence γ let Δ comprise the single default

$$\frac{\gamma : \theta}{\neg\theta}$$

If *E* were an extension, then the default above would have to be either triggered or not triggered by it, and either case is impossible. If the default were triggered, then the consistency condition would fail, and if if were not triggered then the consistency condition would hold, and hence the default would have to be triggered by maximality of extensions.

Let us now consider an example of a default theory with multiple extensions. As before, let *W* contain the sentence γ, and suppose Δ comprises the following two defaults

$$\frac{\gamma : \theta}{\neg\tau}$$

and

$$\frac{\gamma : \tau}{\neg\theta}$$

This theory has exactly two extensions, one in which the first default is triggered and one in which the second one is. It is easy to see that at least a default has to be triggered in any extension, and that both defaults cannot be triggered by the same extension.

These example are enough to bring out a number of features. First, it should be noted that neither one of the two characterizations of default logic given above gives us a way to "construct" extension by means of anything resembling an iterative process. Essentially, one has to "guess" a set of sentences *E*, and then verify that it satisfies the definition of an extension.

Further, the fact that default theories can have zero, one, or more extensions raises the issues of what inferences one is warranted in drawing from a given default theory. The problem can be presented as follows: given a default theory (W,Δ), what sentences can be regarded as *defeasible consequences* of the theory? At first sight, there are several options available.

One option is to take the union of the extensions of the theory, and consider a sentence φ a consequence

of a given default theory if and only if φ belongs to *any* extension of the theory. But this option is immediately ruled out, in that it leads to endorsing contradictory conclusion, as in the second example above. It is widely believed that any viable notion of defeasible consequence for default logic must have the property that the set

$$\{\varphi : (W,\Delta) \mid\!\sim \varphi \}$$

must be consistent whenever W is. Once this option is ruled out, only two alternatives are left:

The first alternative, known as the "credulous" or "bold" strategy, is to pick an extension E for the theory, and say that φ is a defeasible consequence if and only if $\varphi \in E$. The second alternative, known as the "skeptical" or "cautious" strategy, is to endorse a conclusion φ if and only if φ is contained in *every* extension of the theory.

Both the credulous and the skeptical strategy have problems. The problem with the credulous strategy is that the choice of E is arbitrary: with the notion of extension introduced by Reiter, extensions are *orthogonal*: of any two distinct extensions, neither one contains the other. Hence, there seems to be no principled way to pick an extension over any other one. This has led a number of researcher to endorse the skeptical strategy as a viable approach to the problem of defeasible consequence. But as showed by Makinson, skeptical consequence, as based on Reiter's notion of extension, fails to be cautiously monotonic. To see this, consider the default theory $(W\Delta)$, where W is empty, and Δ comprises the following two defaults:

$$\frac{:\theta}{\theta}$$

and

$$\frac{\theta \vee \gamma : \neg\theta}{\neg\theta}$$

This theory has only one extension, coinciding with the deductive closure of $\{\theta\}$. Hence, if we define defeasible consequence by putting $(W,\Delta) \mid\!\sim \varphi$ if and only if φ belongs to every extension of (W,Δ), we have $(W,\Delta) \mid\!\sim \theta$, as well as $(W,\Delta) \mid\!\sim \theta \vee \gamma$ (by the deductive closure of extensions).

Now consider the theory with Δ as before, but with W containing the sentence $\theta \vee \gamma$. This theory has two extensions: one the same as before, but also another one coinciding with the deductive closure of $\{\neg\theta\}$, and hence not containing θ. It follows that the intersection of the extensions no longer contains θ, so that $(\{\theta \vee \gamma\}\ \Delta) \mid\!\sim \theta$ now *fails*, against cautious monotony. (Notice that the same example establishes a

counter-example for Cut for the credulous strategy, when we pick the extension of $(\{\theta \vee \gamma\}, \Delta)$ that contains $\neg\theta$.)

It is clear that the issue of how to define a non-monotonic consequence relation for default logic is intertwined with the way that *conflicts* are handled. The problem of course is that in this case neither the skeptical nor the credulous strategy yield an adequate relation of defeasible consequence. In Antonelli (1999) a notion of *general extension* for default logic is introduced, showing that this notion yields a well-behaved relation of defeasible consequence that satisfies all four requirements of Supraclassicality, Reflexivity, Cut, and Cautious Monotony.

A different set of issues arises in connection with the behavior of default logic from the point of view of computation. For a given semi-decidable set Γ of sentences, the set of all φ that are a consequence of Γ in FOL is itself semi-decidable (see "Computability", this Encyclopedia). In the case of default logic, to formulate the corresponding problem one extends (in the obvious way) the notion of (semi-)decidability to sets of defaults. The problem, then, is to decide, given a default theory (W, Δ) and a sentence φ whether $(W, \Delta) \mid\!\sim \varphi$, where $\mid\!\sim$ is defined, say, skeptically. Such a problem is not even semi-decidable, the essential reason being that in general, in order to determine whether a default is triggered by a pair of sets of sentences, one has to perform a consistency check, and such checks are not computable.

But the consistency checks are not the only source of complexity in default logic. For instance, we could restrict our language to conjunctions of atomic sentences and their negations (making consistency checks feasible). Even so, the problem of determining whether a given default theory has an extension would still be highly intractable (NP-complete, to be precise, as shown by Kautz & Selman (1991)}), seemingly because the problem requires checking all possible sequences of firings of defaults

Default logic is intimately connected with certain *modal* approaches to non-monotonic reasoning, which belong to the family of *autoepistemic logics*. Modal logics in general have proved to be one of the most flexible tools for modelling all sorts of dynamic processes and their complex interactions (see the entry "modal logic", this Encyclopedia). Beside the applications in knowledge representation, which we are going to treat below, there are modal frameworks, known as *dynamic logics*, that play a crucial role, for instance, in the modelling of serial or parallel computation. The basic idea of modal logic is that the language is interpreted with respect to a given set of *states*, and that sentences are evaluated relative to one of these states. What these states are taken to represent depends on the particular application under consideration (they could be epistemic states, or states in the evolution of a dynamical system, etc.), but the important thing is that there are *transitions* (of one or more different kinds) between states.

In the case of one transition that is both *transitive* (i.e., such that if $a \rightarrow b$ and $b \rightarrow c$ then $a \rightarrow c$) and *euclidean* (if $a \rightarrow b$ and $a \rightarrow c$ then $b \rightarrow c$), the resulting modal system is referred to as K45. Associated with each kind of state transition there is a corresponding modality in the language, usually represented as a box $\Box$. A sentence of the form $\Box A$ is true at a state s if and only if A is true at every state s' reachable from s by the kind of transition associated with $\Box$ (see Chellas (1980) for a comprehensive introduction to modal logic).

In autoepistemic logic, the states involved are epistemic states of the agent (or agents). The intuition underlying autoepistemic logic is that we can sometimes draw inferences concerning the state of the world using information concerning our own knowledge or ignorance. For instance, I can conclude that I do not have a sister given that if I did I would probably know about it, and nothing to that effect is present in my "knowledge base". But such a conclusion is defeasible, since there is always the possibility of learning new facts. In order to make these intuitions precise, consider a modal language in which the necessity operator $\Box$ is interpreted as "it is known that". As in default logic or defeasible inheritance, the central notion in autoepistemic logic is that of an *extension* of a theory *S*, i.e., a consistent and self-supporting sets of beliefs that can reasonably be entertained on the basis of *S*. Given a set *S* of sentences, let S_0 be the subset of *S* composed of those sentences containing no occurrences of $\Box$; further, let the *positive introspective closure* PIC(S_0) of S_0 be the set

$$\{\Box\varphi : \varphi \in S_0\},$$

and the *negative introspective closure* NIC(S_0) of S_0 the set

$$\{\neg\Box\varphi : \varphi \notin S_0\}.$$

The set PIC(S_0) is called the introspective closure because it explicitly contains positive information about the agent’s epistemic status: PIC(S_0) expresses what is known (similarly, NIC(S_0) contains negative information about the agent’s epistemic status, stating explicitly what is not known).

With these notions in place, we define an extension for *S* to be a set *T* of sentences such that:

$$T = \{ \varphi : \varphi \text{ follows (in K45) from } S \cup \text{PIC}(T_0) \cup \text{NIC}(T_0) \}$$

Autoepistemic logic provides a rich language, with interesting mathematical properties and connections to other non-monotonic formalisms. It is faithfully intertranslatable with Reiter’s version of default logic, and provides a defeasible framework with well-understood modal properties.

Conclusion

There are three major issues connected with the development of logical frameworks that can adequately represent defeasible reasoning: (*i*) material adequacy; (*ii*) formal properties; and (*iii*) complexity. Material adequacy concerns the question of how broad a range of examples is captured by the framework, and the extent to which the framework can do justice to our intutions on the subject (at least the most entrenched ones). The question of formal properties has to do with the degree to which the framework allows for a relation of logical consequence that satisfies the above mentioned conditions of Supraclassicality, Reflexivity, Cut, and Cautious Monotony. The third set of issues has to do with computational

complexity of the most basic questions concerning the framework.

There is a potential tension between (*i*) and (ii): the desire to capture a broad range of intuitions can lead to *ad hoc* solutions that can sometimes undermine the desirable formal properties of the framework. In general, the development of non-monotonic logics and related formalisms has been driven, since its inception, by consideration (*i*) and has relied on a rich and well-chosen array of examples. Of course, there is some question as to whether any single framework can aspire to be universal in this respect.

More recently, researchers have started paying attention to consideration (*ii*), looking at the extent to which non-monotonic logics have generated well-behaved relations of logical consequence. As Makinson (1994) points out, practitioners of the field have encountered mixed success. In particular, one abstract property, Cautious Monotony, appears at the same time to be crucial and elusive for many of the frameworks to be found in the literature. This is a fact that is perhaps to be traced back, at least in part, to the above-mentioned tension between the requirement of material adequacy and the need to generate a well-behaved consequence relation. The complexity issue appears to be the most difficult among the ones that have been singled out. Non-monotonic logics appear to be stubbornly intractable with respect to the corresponding problem for classical logic. This is clear in the case of default logic, given the ubiquitous consistency checks. But beside consistency checks, there are other, often overlooked, sources of complexity that are purely combinatorial. Other forms of nonmonotonic reasoning, beside default logic, are far from immune from these combinatorial roots of intractability. Although some important work has been done trying to make various non-monotonic formalism more tractable, this is perhaps the problem on which progress has been slowest in coming.

Bibliography

- Antonelli, G.A., 1997, "Defeasible inheritance over cyclic networks", *Artificial Intelligence*, vol. 92 (1), pp. 1-23.
- -----, 1999, "A directly cautious theory of defeasible consequence for default logic via the notion of general extension", *Artificial Intelligence*, vol. 109 (1-2), pp. 71-109.
- Chellas, B., 1980, *Modal Logic: an introduction*, Cambridge: Cambridge University Press, 1980.
- Gabbay, D., Hogger, C., and Robinson, J., (eds.), 1994, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3, Oxford and New York: Oxford University Press
- Ginsberg, M., (ed.), 1987, *Readings in Nonmonotonic Reasoning*, , Los Altos, CA: Morgan Kauffman
- Horty, J.F., 1994, "Some direct theories of nonmonotonic inheritance", in Gabbay *et al*. (1994), pp. 111-187.
- -----, forthcoming, "Skepticism and floating conclusions", *Artificial Intelligence Journal*.
- Kautz, H., and Selman, B., 1991, "Hard problems for simple default logic", *Artificial Intelligence Journal*, vol. 49, pp. 243-279
- Makinson, D., 1994, "General patterns in nonmonotonic reasoning", in Gabbay *et al*. (1994), pp. 35-110.
- Stalnaker, R., 1994, "Nonmonotonic consequence relations", *Fundamenta Informaticae*, vol. 21,

pp. 7-21.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

artifical intelligence: logic and | artificial intelligence | computability theory | logic: classical | logic: modal | model theory: first-order

G. Aldo Antonelli
aldo@uci.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 10, 2001
Content last modified: December 10, 2001

First-order Model Theory

First-order model theory, also known as classical model theory, is a branch of mathematics that deals with the relationships between descriptions in first-order languages and the structures that satisfy these descriptions. From one point of view, this is a vibrant area of mathematical research that brings logical methods (in particular the theory of definition) to bear on deep problems of classical mathematics. From another point of view, first-order model theory is the paradigm for the rest of model theory; it is the area in which many of the broader ideas of model theory were first worked out.

1. First-order languages and structures

Mathematical model theory carries a heavy load of notation, and HTML is not the best container for it. In what follows, syntactic objects (languages, theories, sentences) are generally written in roman or greek letters (for example L, T, φ), and set-theoretic objects such as structures and their elements are written in italic (*A*, *a*). Two exceptions are that variables are italic (*x*, *y*) and that sequences of elements are written with lower case roman letters (a, b).

We recall and refine some definitions from the entries on classical logic and model theory. A *signature* is a set of individual constants, predicate symbols and function symbols; each of the predicate symbols and function symbols has an *arity* (for example it is binary if its arity is 2). Each signature K gives rise to a first-order language, by building up formulas from the symbols in the signature together with logical symbols (including =) and punctuation.

If K is a signature, then a *structure of signature* K, say *A*, consists of the following items:

1. A set called the *domain* of *A* and written dom(*A*); it is usually assumed to be nonempty;
2. for each individual constant *c* in K, an element c^A of dom(*A*);
3. for each predicate symbol *P* of arity *n*, an *n*-ary relation P^A on dom(*A*);
4. for each function symbol *F* of arity *n*, an *n*-ary function F^A from dom(*A*) to dom(*A*).

The *elements* of *A* are the elements of dom(*A*). Likewise the *cardinality* or *power* of *A* is the cardinality of its domain. Since we can recover the signature K from the first-order language L that it generates, we can and will refer to structures of signature K as L-*structures*. We think of *c* as a name for the element c^A in the structure *A*, and likewise with the other symbols.

For example the field of real numbers forms a structure ***R*** whose elements are the real numbers, with signature consisting of the individual constant 0 to name the number zero, a 1-ary function symbol - for minus, and two 2-ary function symbols + and . for plus and times. At first sight we can't add a symbol to express 1/*x*, since all the named functions have to be defined on the whole domain of the structure, and there is no such real number as 1/0. But on second thoughts this is not a serious problem; any competent mathematician puts the condition '*x* is not zero' before dividing by *x*, and so it never matters what the value of 1/0 is, and we can harmlessly take it to be 42. But most model theorists are uncomfortable with any kind of division by zero, so they stick with plus, times and minus.

If L is the first-order language of signature K, then Tarski's model-theoretic truth definition tells us when a sentence of L is true in *A*, and when an assignment of elements of *A* to variables satisfies a formula of L in *A*. Instead of talking of assignments satisfying a formula, model theorists often speak of the set of *n*-tuples of elements of *A* that is *defined* by a formula $\varphi(v_1,...,v_n)$; the connection is that an *n*-tuple $(a_1,...,a_n)$ is in the defined set if and only if the assignment taking each v_i to a_i satisfies the formula.

If φ is a sentence, we write

$$A \models \varphi$$

to mean that φ is true in *A*. If $\varphi(v_1,...,v_n)$ is a formula with free variables as shown, we write

$$A \models \varphi[\mathrm{a}]$$

to mean that the *n*-tuple a is in the set defined by φ. (The entry on classical logic uses the notation '$A,s \models \varphi$', where *s* is any assignment to all the variables of L that assigns to each variable v_i free in φ the *i*-th element in the *n*-tuple a.)

Two L-structures that are models of exactly the same sentences of L are said to be *elementarily equivalent*. Elementary equivalence is an equivalence relation on the class of all L-structures. The set of all the sentences of L that are true in the L-structure *A* is called the *complete theory* of *A*, in symbols Th(*A*). A theory that is Th(*A*) for some structure *A* is said to be *complete*. (By the completeness theorem for first-order logic, for which see the entry on classical logic, a theory is complete if and only if it is maximal syntactically consistent.) The two structures *A* and *B* are elementarily equivalent if and only if Th(*A*) = Th(*B*).

To continue the example of the field ***R*** of real numbers: It is often not at all obvious whether two given structures are or are not elementarily equivalent. One of the greatest achievements of the pre-history of model theory was Tarski's description in 1930 of Th(***R***) (which he published in full only after the war; see his book in the Bibliography below). This description implied among other things that the structures elementarily equivalent to ***R*** are exactly the real-closed fields, a class of fields which was already known to the algebraists in its own right.

When mathematicians introduce a class of structures, they like to define what they count as the basic maps between these structures. The basic maps between structures of the same signature K are called *homomorphisms*, defined as follows. A *homomorphism from structure A to structure B* is a function *f* from dom(*A*) to dom(*B*) with the property that for every atomic formula $\varphi(v_1,...,v_n)$ and any *n*-tuple a = $(a_1,...,a_n)$ of elements of *A*,

$$A \models \varphi[\mathrm{a}] \Rightarrow B \models \varphi[\mathrm{b}]$$

where b is $(f(a_1),...,f(a_n))$. If we have '⇔' in place of ' ⇒ ' in the quoted condition, we say that *f* is an *embedding* of *A* into *B*. Since the language includes =, an embedding of *A* into *B* is always one-to-one, though it need not be onto the domain of *B*. If it is onto, then the inverse map from dom(*B*) to dom(*A*) is also a homomorphism, and both the embedding and its inverse are said to be *isomorphisms*. We say that two structures are *isomorphic* if there is an isomorphism from one to the other. Isomorphism is an equivalence relation on the class of all structures of a fixed signature K. If two structures are isomorphic then they share all model-theoretic properties; in particular they are elementarily equivalent.

If *A* and *B* are structures of signature K with dom(*A*) a subset of dom(*B*), and the interpretations in *A* of the symbols in K are just the restrictions of their interpretations in *B*, then we say that *A* is a *substructure* of *B* and conversely *B* is an *extension* of *A*. If moreover *B* has some elements that are not in *A*, we say that *A* is a *proper substructure* of *B* and *B* is an *proper extension* of *A*. If *B* is a structure and *X* is a nonempty subset of dom(*B*), then there is a unique smallest substructure of *B* whose domain contains all of *X*. It is known as the *substructure of B* generated by X, and we find it by first adding to *X* all the elements c^B where *c* are individual constants, and then closing off under the functions F^B where *F* are function symbols.

For example the substructure of the field ***R*** generated by the number 1 consists of 1, 0 (since it is named by the constant 0), 1+1, 1+1+1 etc., -1, -2 etc., in other words the ring of integers. (There is no need to close off under multiplication too, since the set of integers is already closed under multiplication.) If we had included a symbol for $1/x$ too, the substructure generated by 1 would have been the field of rational numbers. So the notion of substructure is sensitive to the choice of signature.

2. Elementary maps

Let L be a first-order language and let *A* and *B* be L-structures. Suppose *e* is a function which takes some elements of *A* to elements of *B*. We say that *e* is an *elementary map* if whenever a sequence of elements $a_1, ..., a_n$ in the domain of *e* satisfy a formula $\varphi(x_1,...,x_n)$ of L in *A*, their images under *e* satisfy the same formula in *B*; in symbols

$$A \models \varphi(a_1,...,a_n) \Rightarrow B \models \varphi(e(a_1),...,e(a_n)).$$

We say that *e* is an *elementary embedding* of *A* into *B* if *e* is an elementary map and its domain is the whole domain of *A*. As the name implies, elementary embeddings are always embeddings.

If there is an elementary embedding from *A* to *B* then *A* and *B* are elementarily equivalent. On the other hand an embedding between elementarily equivalent structures, or even between isomorphic structures, need not be elementary. (For example, writing ***Z*** for the abelian group of the integers with signature consisting of 0 and +, the embedding from ***Z*** to ***Z*** that takes each integer *n* to 2*n* is an embedding, and of course ***Z*** is isomorphic to itself; but this embedding is not elementary, since 1 satisfies the formula $\neg\exists y(y + y = v_1)$, but 2 doesn't.)

We say that *A* is an *elementary substructure* of *B*, and *B* is an *elementary extension* of *A*, if *A* is a substructure of *B* and the inclusion map is an elementary embedding. It's immediate from the definitions that an elementary extension of an elementary extension of *A* is again an elementary extension of *A*.

Elementary embeddings are natural maps to consider within first-order model theory. Around 1950 Abraham Robinson was impressed that maps between algebraic structures in general seem hardly ever to be elementary, whereas some important maps (such as embeddings between two algebraically closed fields, or between two real-closed fields) turn out to be elementary. He was also surprised to find that this fact about algebraically closed fields is another way of stating a celebrated theorem called the Hilbert Nullstellensatz. These observations of Robinson have had a huge effect on the development of model theory. In Robinson's terminology, a first-order theory is *model-complete* if every embedding between models of the theory is elementary. This notion has found many uses, and it often appears in applications of model theory in algebra.

Elementary embeddings have a number of properties that make them useful. We have space for four.

> **The downward Loewenheim-Skolem theorem**:
> Suppose L is a first-order language which has κ formulas, *A* is an L-structure and λ is a cardinal which is at least κ but less than the cardinality of *A*. Suppose also that *X* is a set of at most λ elements of *A*. Then *A* has an elementary substructure which has cardinality exactly λ and contains all the elements in *X*.

There is a proof of this in the entry on classical logic, using Skolem hulls. Note that λ must be infinite since every first-order language has infinitely many formulas.

> **The elementary chain theorem**:
> Suppose that L is a first-order language and A_0, A_1, ... is a sequence (of any length) of L-structures such that any structure in the sequence is an elementary substructure of all the later structures in the sequence. Then there is a unique smallest L-structure *B* which contains all the structures in the sequence as substructures; this structure *B* is an elementary extension of all the structures in the sequence.

> **The elementary amalgamation theorem**:
> Suppose L is a first-order language, *A* is an L-structure and *B*, *C* are two elementary extensions of *A*. Then there are an elementary extension *D* of *B* and an elementary embedding *e* of *C* into *D* such that (i) for each element *a* of *A*, *e(a)* = *a*, and (ii) if *c* is an element of *C* but not of *A*, then *e(c)* is not in *B*.

The latter is a consequence of the compactness theorem in the next section.

> **The upward Loewenheim-Skolem theorem**:
> Suppose L is a first-order language which has κ formulas, *A* is an L-structure whose cardinality is an infinite cardinal μ, and λ is a cardinal which is at least as great as both κ and μ. Then *A* has an elementary extension whose cardinality is λ.

There is a proof of this in the entry on classical logic. The name of the theorem is a little unfortunate, since the theorem was first proved by Tarski, and Skolem didn't even believe it (because he didn't believe in uncountable cardinals).

There is another proof using the elementary amalgamation theorem and the elementary chain theorem. The compactness theorem and the diagram lemma (see below) allow us to prove that *A* has a proper elementary extension A'. Now use A' and again A' for the structures *B* and *C* in the elementary amalgamation theorem. Then *D* as in the theorem is an elementary extension of *A*, and by (ii) in the theorem, it must contain elements that are not in *A*, so that it is a proper elementary extension. Repeat to get a proper elementary extension of *D*, and so on until you have an infinite elementary chain. Use the elementary chain theorem to find an elementary extension of *A* that sits on top of this chain. Keep repeating these moves until you have an elementary extension of *A* that has cardinality at least λ. Then if

necessary use the downward Loewenheim-Skolem theorem to pull the cardinality down to exactly λ. This kind of argument is very common in first-order model theory.

3. Five grand theorems

The five theorems reported in this section are in some sense the pillars of classical model theory. All of them are theorems about first-order model theory. A great deal of the work done in the third quarter of the twentieth century was devoted to working out the consequences of these theorems within first-order model theory, and the extent to which similar theorems hold for languages that are not first-order.

3.1 The compactness theorem

> If T is a first-order theory, and every finite subset of T has a model, then T has a model.

There is a proof of this theorem in the entry on classical logic. The theorem has several useful paraphrases. For example it is equivalent to the following statement:

> Suppose T is a first-order theory and φ is a first-order sentence. If T $\models \varphi$ then there is a finite subset U of T such that U $\models \varphi$.

(See the entry Model Theory for the notion $\models$ of model-theoretic consequence. To derive the second statement from the first, note that 'T $\models \varphi$' is true if and only if there is no model of the theory T $\cup$ {$\neg \varphi$}.)

Anatolii Mal'tsev first gave the compactness theorem in 1938 (for first-order logic of any signature), and used it in 1940/1 to prove several theorems about groups; this seems to have been the first application of model theory within classical mathematics. Leon Henkin rediscovered the theorem a few years later and gave some further applications. The theorem fails badly for nearly all infinitary languages.

3.2 The diagram lemma

If *A* is an L-structure, then we form the *diagram* of *A* as follows. First add to L a supply of new individual constants to serve as names for all the elements of *A*. (This illustrates how in first-order model theory we easily find ourselves using uncountable signatures. The 'symbols' in these signatures are abstract set-theoretic objects, not marks on a page.) Then using L and these new constants, the *diagram* of *A* is the set of all the atomic sentences and negations of atomic sentences that are true in *A*.

> If B' is a model of the diagram of *A*, and *B* is B' with the new constants removed from the signature, then there is an embedding of *A* into *B*.

Namely, if an element of *A* is named by a new constant *c*, then map that element to the element of B' named *c*. A variant of this lemma is used in the proof of the elementary amalgamation theorem.

3.3 The Lyndon interpolation theorem

This theorem may have the longest pedigree of any theorem of model theory, since it generalises the Laws of Distribution for syllogisms, which go back at least to the early Renaissance. The theorem is easiest to state if we assume that our first-order languages have symbols $\wedge$, $\forall$ and $\neg$, but not $\rightarrow$ or $\leftrightarrow$. Then an occurrence of a predicate symbol R in a formula φ is said to be *positive* (resp. *negative*) if it lies within the scope of an even (resp. odd) number of occurrences of $\neg$.

> Suppose L and M are first-order languages, L $\cup$ M is the smallest first-order language containing both L and M, and L $\cap$ M is the language consisting of all the formulas which are in both L and M. Suppose T is a theory in L, U is a theory in M, and no (L $\cup$ M)-structure is both a model of T and a model of U. Then there is a sentence φ of L $\cap$ M which is true in all models of T and false in all models of U. (This sentence φ is called the *interpolant*.) Moreover every predicate symbol with a positive occurrence in φ has a positive occurrence in some sentence of T and a negative occurrence in some sentence of U, and conversely every predicate symbol with a negative occurrence in φ has a negative occurrence in some sentence of T and a positive occurrence in some sentence of U.

There are several proofs of this theorem, and not all of them are model-theoretic. Without the last sentence, the theorem is known as Craig's interpolation theorem, since William Craig proved this version a few years before Roger Lyndon found the full statement in 1959. As Craig noted at the time, his interpolation theorem gives a neat proof of Evert Beth's definability theorem, which runs as follows.

Suppose that L is a first-order language and M is the first-order language got by adding to L a new predicate symbol R. Suppose also that T is a theory in M. We say that T *implicitly defines* R if it is false that there are two M-structures which are models of T, have the same elements and interpret all the symbols of L in the same way but interpret the symbol R differently. We say that T *defines* R *explicitly* if there is a formula $\varphi(x_1,...,x_n)$ of L such that in every model of T, the formulas φ and $R(x_1,...,x_n)$ are satisfied by exactly the same *n*-tuples $(a_1,...,a_n)$ of elements. It is easy to see that if T defines R explicitly then it defines R implicitly. (This fact is known as *Padoa's method*; Padoa used the failure of implicit definability as a way of proving the failure of explicit definability.) Beth's theorem is the converse:

> Suppose that L, M, R and T are as above. If T defines R implicitly then T defines R explicitly.

3.4 The omitting types theorem

> Suppose L is a first-order language which has countably many formulas. Suppose T is a complete theory in L, and Φ is a set of formulas of L which all have the free variable *x*. Finally suppose that every model of T contains an element which satisfies all the formulas in Φ. Then there is a formula $\psi(x)$ of L such that in every model of T there is an element satisfying ψ, and every element that satisfies ψ in any model of T also satisfies all the formulas in Φ. (In other words, there is a finite reason why the 'type' Φ can't be 'omitted' in any model of T.)

This theorem, which goes back to the mid 1950s, very definitely depends on the language being first-order and countable. It has several useful generalisations, for example *model-theoretic forcing*, which is an analogue of the forcing construction in set theory. In fact the games used for model-theoretic forcing (see the entry on logic and games) can be adapted to prove the omitting types theorem too. There are similar but more complicated theorems for uncountable first-order languages; some of these can be paraphrased as omitting types theorems for infinitary languages.

3.5 The initial model theorem

A quantifier-free formula is said to be a *Horn formula* (after Alfred Horn) if it has one of the three forms

- ψ,
- $\varphi_1 \wedge \ldots \wedge \varphi_n \rightarrow \psi$,
- $\neg(\varphi_1 \wedge \ldots \wedge \varphi_n)$,

where the formulas $\varphi_1, \ldots, \varphi_n, \psi$ are all atomic. A *universal Horn sentence* (also known to the computer scientists as a *Horn clause*) is a sentence that consists of universal quantifiers followed by a quantifier-free Horn formula; it is said to be *strict* if no negation sign occurs in it (i.e. if it doesn't come from a quantifier-free Horn formula of the third kind).

> Let T be a theory consisting of strict universal Horn sentences. Then T has a model *A* with the property that for every model *B* of T there is a unique homomorphism from *A* to *B*. (Such a model *A* is called an *initial model* of T. It is unique up to isomorphism.)

This theorem is a generalisation, due to Mal'tsev, of a group-theoretic construction called *construction by generators and relations*. It is the main idea behind *algebraic specification*, which is one approach to the specification of systems in computer science. The required behaviour of the system is written down as a set of strict universal Horn sentences, and then the initial model of these sentences is an abstract version of the required system.

4. Three useful constructions

A construction is a procedure for building a structure. We have already seen several constructions in the theorems above: for example the omitting types construction and the initial model construction. Here are three more.

4.1 Products and reduced products

If *A* and *B* are L-structures, we form their *product* $C = A \times B$ as follows. The elements of *C* are the ordered pairs (*a*,*b*) where *a* is an element of *A* and *b* is an element of *B*. The predicate symbols are interpreted 'pointwise', i.e. so that for example

> *(a,b)* is in P^C if and only if *a* is in P^A and *b* is in P^B.

The structures *A* and *B* are called the *factors* of $A \times B$. In the same way we can form products of any number of structures. If all the factors of a product are the same structure *A*, the product is called a *power* of *A*. A theorem called the *Feferman-Vaught theorem* tells us how to work out the complete theory of the product from the complete theories of its factors.

This construction has some variants. We can define an equivalence relation on the domain of a product *C*, and then take a structure *D* whose elements are the equivalence classes; the predicate symbols are interpreted so as to make the natural map from dom(*C*) to dom(*D*) a homomorphism. In this case the structure *D* is called a *reduced product* of the factors of *C*. It is a *reduced power* of *A* if all the factors are equal to *A*; in this case the *diagonal map* from *A* to *D* is the one got by taking each element *a* to the equivalence class of the element *(a,a,...)*.

Suppose we use a set *I* to index the factors in a product *C*. An *ultrafilter* over *I* is a set *U* of subsets of *I* with the properties

- if sets *X* and *Y* are in *U* then so is their intersection $X \cap Y$;
- if *X* is in *U* and $X \subseteq Y \subseteq I$ then *Y* is in *U*;
- for each subset *X* of *I*, exactly one of *X* and its complement $I \setminus X$ is in *U*.

If we have an ultrafilter *U* over *I*, then we can construct a reduced product from *C* by making two elements of *C* equivalent if and only if the set of indices at which they are equal is a set in the ultrafilter *U*. This is indeed an equivalence relation on the domain of *C*, and the resulting reduced product is called an *ultraproduct* of the factors of *C*. If *C* is a power of *A* then this ultraproduct is called an *ultrapower* of *A*, and it is sometimes written *U*-prod *A*. A theorem called *Los's theorem* describes what sentences are true in an ultraproduct. Its most useful consequence is the following:

> If *U* is an ultrafilter then the diagonal map from *A* to *U*-prod *A* is an elementary

embedding.

If the ultrafilter *U* is *nonprincipal*, i.e. contains no finite sets, then the diagonal map is not onto the domain of *U*-prod *A*, and in fact *U*-prod *A* is generally much larger than *A*. So we have a way of constructing large elementary extensions. The axiom of choice guarantees that every infinite set has many nonprincipal ultrafilters over it. Ultrapowers are an essential tool for handling large cardinals in set theory (see the entry on set theory).

A remarkable (but in practice not terribly useful) theorem of Saharon Shelah tells us that a pair of structures *A* and *B* are elementarily equivalent if and only if they have ultrapowers that are isomorphic to each other.

4.2 Saturation

Suppose *A* is an L-structure, *X* is a set of elements of *A*, *B* is an elementary extension of *A* and *b*, *c* are two elements of *B*. Then *b* and *c* are said to have *the same type over X* if for every formula $\varphi(v_1,\ldots,v_{n+1})$ of L and every *n*-tuple d of elements of *X*,

$$B \models \varphi[b,\mathrm{d}] \Leftrightarrow B \models \varphi[c,\mathrm{d}].$$

We say that *A* is *saturated* if whenever *X* is a set of elements of *A*, of cardinality less than that of *A*, and *B* is any elementary extension of *A*, we always have that every element of *B* has the same type over *X* as some element already in *A*.

This rather heavy definition gives little clue how useful saturated structures are. If every structure had a saturated elementary extension, many of the results of model theory would be much easier to prove. Unfortunately the existence of saturated elementary extensions depends on features of the surrounding universe of sets. There are technical ways around this obstacle, for example using weakenings of the notion of saturation. We have two main ways of constructing saturated elementary extensions. One is by ultrapowers, using cleverly constructed ultrafilters. The other is by taking unions of elementary chains, generalising the proof we gave for the upward Loewenheim-Skolem theorem.

The existence of partially saturated elementary extensions of the field ***R*** of real numbers is the main technical fact behind Abraham Robinson's *nonstandard analysis*. See Section 4 of the entry on model theory for more information on this. Though model theory provided the first steps in nonstandard analysis, this branch of analysis rapidly became a subject in its own right, and its links with first-order model theory today are rather slim.

4.3 Ehrenfeucht-Mostowski models

Let A be an L-structure, X a set of elements of A and < a linear ordering of X (not necessarily definable by a first-order formula). We say that $(X,<)$ is an *indiscernible sequence* in A if for every natural number n, and all elements $a_1,...,a_n,b_1,...,b_n$ of A such that $a_1 < ... < a_n$ and $b_1 < ... < b_n$, the map taking each a_i to the corresponding b_i is an elementary map. If T is a theory with infinite models, then T has models that are the Skolem hulls (see the entry on classical logic) of indiscernible sequences. These models are known as *Ehrenfeucht-Mostowski models*, after the two Polish model theorists who first carried out this construction in the mid 1950s. These models tend to be the opposite of saturated; we can arrange that very few types over sets of elements are represented among their elements. Some important distinctions between different models of set theory can be expressed in terms of the indiscernible sequences within these models; see the entry on set theory.

5. Three successful programmes

Every healthy branch of mathematics needs a set of problems that form a serious challenge for its researchers. We close with a brief introduction to some of the research programmes that drove first-order model theory forwards in the second half of the twentieth century. There are other current programmes besides these; see for example the handbook edited by Yuri Ershov in the bibliography, which is about model theory when the structures are built recursively.

5.1. Categoricity and classification

In 1904 Oswald Veblen stole a term from Kant and described a theory as *categorical* if it has just one model up to isomorphism (i.e. it has a model and all its models are isomorphic to each other). The depressing news is that there are no categorical first-order theories with infinite models; we can see this at once from the upward Loewenheim-Skolem theorem. In fact if T is a first-order theory with infinite models, then the strongest kind of categoricity we can hope for in T is that for certain infinite cardinals κ, T has exactly one model of cardinality κ, up to isomorphism. This property of T is called κ-*categoricity*.

Now there is a heuristic principle that many people have used, though it seems to have no simple formulation. I suggest ‘Few is beautiful’. The principle says that if a first-order theory T constrains its models (of a particular cardinality) to be all similar to each other, this can only be because the models of T have few irregularities and asymmetries. So there should be a good structural description of these models. One should expect that they are ‘good structures’ from the point of view of classical mathematics. As a first step, one easily sees from the upward and downward Loewenheim-Skolem theorems that if T is κ-categorical for some κ at least as large as the number of formulas in the language of T, then T must be a complete theory. From now on, T is a complete theory with infinite models; and for simplicity we shall assume that the language of T is countable.

In 1954 Jerzy Los announced that he could only find three kinds of example of theories T that are κ-categorical. Namely:

- T is *totally categorical* if it is κ-categorical for every infinite cardinal κ. A typical example is the complete theory of an infinite-dimensional vector space over a finite field.
- T is *uncountably categorical* (but not totally categorical) if it is κ-categorical precisely when κ is uncountable. Essentially the only example that Los could find was the complete theory of an algebraically closed field; this is uncountably categorical by a well-known theorem of Steinitz.
- T is *countably categorical* (but not uncountably categorical) if it is κ-categorical precisely when κ is countable. A typical example is the complete theory of a dense linear ordering with no first or last element; this is countably categorical by a well-known theorem of Cantor.

Los asked whether there are any other possibilities besides these three. (Of course most complete theories are not κ-categorical for any κ.)

This question of Los was a tremendous stimulus to research, and it led to a classic paper of Michael Morley in 1965 which showed that Los's three possibilities are in fact the only ones. One central idea of Morley's analysis was that models of an uncountably categorical theory have the smallest possible number of types of element; this led directly to the branch of model theory called *stability theory*, which studies theories that have a limited number of element types. These theories have the remarkable property that every infinite indiscernible sequence in any of their models is indiscernible under any linear ordering whatever; so these sequences are a kind of generalisation of bases of vector spaces. Another idea implicit in Morley's work, but much clarified by later work of John Baldwin and Alistair Lachlan, was that in any model of an uncountably categorical theory there is a central core (called a *strongly minimal set*) which carries a dependence relation obeying similar laws to linear dependence in vector spaces. In terms of this dependence relation one can define a dimension for the model, and what remains of the model outside the core is so closely tied to the core that the dimension determines the model up to isomorphism.

Saharon Shelah developed Morley's ideas with tremendous resourcefulness and energy. His main aim was to stretch the 'Few is beautiful' idea by showing that there are clear dividing lines between kinds of theory T. On one side of a dividing line are theories with some good structural property that forces the number of nonisomorphic models of a given cardinality to be small. On the other side, every theory has (for example) two models of the same cardinality that are not isomorphic but are extremely hard to tell apart. Shelah coined the name *classification theory* for this research, though the name never caught on. The text of Lascar listed below is an elegant introduction to this whole programme, from Los to Shelah. Meanwhile Shelah himself has extended it far beyond first-order logic. Even in the first-order case, Shelah had to invent new set-theoretic techniques (such as proper forcing) to carry out his constructions.

5.2. Geometric model theory

Geometric model theory grew out of Michael Morley's 1965 paper, but in a different direction from Shelah's work (though today it makes regular use of technical tools developed by Shelah in his classification programme). Morley had shown that models of an uncountably categorical theory have structural properties that are interesting in their own right, regardless of the complete theory of the

structure; so it became the custom to talk of *uncountably categorical structures*, meaning models of uncountably categorical theories. (And likewise *totally categorical structures*.) Independently Boris Zil'ber in Siberia and Greg Cherlin in the United States noticed that any infinite group that is definable in an uncountably categorical structure must have many features in common with the algebraic groups studied by algebraic geometers. Zil'ber in particular showed that many methods from algebraic geometry generalise to the model-theoretic case. His secret weapon was Bezout's Theorem from geometry, which he used to guide him to solutions of very difficult model theoretic problems; for example his theorem that no totally categorical theory can be axiomatised by a finite number of axioms. (It was secret in the sense that it guided his intuition but never appeared explicitly in his results.) Zil'ber also noticed an important difference between the first and second of Los's examples above. Namely, in a vector space the subspaces (i.e. the subsets closed under linear dependence) form a modular lattice; but the algebraically closed subsets of an algebraically closed field form a lattice that is not modular.

Partly because of the difficulty of communications between Siberia and the West, these results of Zil'ber took some time to digest, and in part they had to be rediscovered in the West. But when the message did finally get through, the result was a new branch of model theory which has come to be known as *geometric model theory*. The programme is broadly to classify structures according to (a) what groups or fields are interpretable in them (in the sense sketched in the entry on model theory) and (b) whether or not the structures have 'modular geometries'; and then to use this classification to solve problems in model theory and geometry. Since the mid 1980s the leader of this research has been Ehud Hrushovski. In the early 1990s, using joint work with Zil'ber, Hrushovski gave a model-theoretic proof (the first complete proof to be found) of the geometric Mordell-Lang conjecture in all characteristics; this was a conjecture in classical diophantine geometry. The book edited by Bouscaren in the references below is devoted to Hrushovski's proof and the necessary background in model theory. Both (a) and (b) are fundamental to Hrushovski's argument.

5.3. O-minimality

Of the three programmes described here, this is the oldest, since it grew out of Tarski's description of the complete theory of the field of real numbers (which he proved by the method of quantifier elimination; see section 2.2 of the entry on Tarski's truth definitions). In the course of giving this description, Tarski had shown that every first-order formula $\varphi(x)$ in the relevant language, possibly with parameters, is satisfied by exactly the same assignments as some boolean combination of formulas of the form $x < s$ or $t < x$ where s, t are constant terms naming parameters. Another way of saying this is that

> Every set of elements definable by a first-order formula is a finite union of open intervals with named endpoints, together with some finite set of elements.

A linearly ordered structure with this property is said to be *o-minimal*. (The idea of the name is that o-minimality is an analogue of Morley's 'strong minimality', in a form that makes sense for structures that carry a linear ordering, whence 'o-' for ordering.)

In 1982 Lou van den Dries showed that the fact that the field of real numbers is o-minimal gives a large amount of useful information about the definable sets of higher dimension, such as the family of definable subsets of the real plane. Soon after this, Julia Knight, Anand Pillay and Charles Steinhorn noticed that if a structure *A* is o-minimal, then so is any structure elementarily equivalent to *A*, and that Van den Dries' analysis of higher-dimensional definable set applies to all these structures. These results led to much activity on the frontier between model theory and function theory. Several old problems from model theory and function theory were solved. Alex Wilkie showed that the field of real numbers with a symbol for exponentiation is o-minimal and has a model-complete complete theory, and thereby gave a positive answer to Tarski's old problem of whether this structure allows a quantifier elimination (though his method was very far from the syntactic analysis that Tarski had in mind). We now know a wide range of ways of adding interesting features to the field of real numbers in such a way that the resulting structure is still o-minimal (and hence in some sense mathematically tractable). Van den Dries has urged that o-minimal structures provide a good setting for developing the 'tame topology' programme of Alexander Grothendieck.

Bibliography

- Beth, E. 1953, "On Padoa's method in the theory of definition", *Nederl. Akad. Wetensch. Proc. Ser. A* **56**, 330-339.
- Bouscaren, E. ed. 1998, *Model Theory and Algebraic Geometry: An introduction to E. Hrushovski's proof of the geometric Mordell-Lang conjecture*, Lecture Notes in Mathematics **1696**, Berlin : Springer-Verlag.
- Buechler. S. 1996, *Essential Stability Theory*, Berlin : Springer-Verlag,.
- Chang, C. and Keisler, J. 1990, *Model Theory*, Amsterdam : North-Holland.
- Dries, L. van den 1998, *Tame Topology and O-minimal Structures*, Cambridge : Cambridge University Press.
- Ehrig, H. and Mahr, B. 1985, *Fundamentals of Algebraic Specification I: Equations and Initial Semantics*, Berlin : Springer-Verlag.
- Ershov, Y. ed. 1998, *Handbook of Recursive Mathematics I, Recursive Model Theory*, New York : Elsevier.
- Hart, B., Lachlan, A. and Valeriote, M. 1996, *Algebraic Model Theory*, Dordrecht : Kluwer.
- Haskell, D., Pillay, A. and Steinhorn, C. 2000, *Model Theory, Algebra, and Geometry*, Cambridge : Cambridge University Press.
- Hodges, W. 1993, *Model Theory*, Cambridge : Cambridge University Press.
- Hodges, W. 1998, "The laws of distribution for syllogisms", *Notre Dame Journal of Formal Logic* **39**, 221-230.
- Lascar, D. 1986, *Stability in Model Theory*, Harlow : Longman.
- Morley, M. 1965, "Categoricity in power", *Transactions of the American Mathematical Society* **114**, 514-538.
- Pillay, A. 1996, *Geometric Stability Theory*, Oxford : Oxford University Press.
- Poizat, B. 2000, *A Course in Model Theory*, New York : Springer.

- Shelah, S. 1990, *Classification Theory*, Amsterdam : North-Holland.
- Tarski, A. 1951, *A Decision Method for Elementary Algebra and Geometry*, Berkeley, University of California Press.
- Vaught, R. 1974, "Model theory before 1945", in *Proceedings of the Tarski Symposium*, ed. L. Henkin et al., Providence RI : American Mathematical Society, 153-172.

Other Internet Resources

[Please contact that author with suggestions.]

Related Entries

logic: and games | logic: classical | logic: infinitary | model theory | set theory | Tarski, Alfred: truth definitions

Wilfrid Hodges
School of Mathematical Sciences,
Queen Mary, University of London
w.hodges@qmw.ac.uk

First published: November 9, 2001
Content last modified: November 9, 2001

Model Theory

Model theory began with the study of formal languages and their interpretations, and of the kinds of classification that a particular formal language can make. Mainstream model theory is now a sophisticated branch of mathematics (see the entry on first-order model theory). But in a broader sense, model theory is the study of the interpretation of any language, formal or natural, by means of set-theoretic structures, with Alfred Tarski's truth definition as a paradigm. In this broader sense, model theory meets philosophy at several points, for example in the theory of logical consequence and in the semantics of natural languages.

1. Basic notions of model theory

Sometimes we write or speak a sentence S that expresses nothing either true or false, because some crucial information is missing about what the words mean. If we go on to add this information, so that S comes to express a true or false statement, we are said to *interpret* S, and the added information is called an *interpretation* of S. If the interpretation *I* happens to make S state something true, we say that *I* is a *model* of S, or that *I satisfies* S, in symbols '$I \models$ S'. Another way of saying that *I* is a model of S is to say that S is *true in I*, and so we have the notion of *model-theoretic truth*, which is truth in a particular interpretation. But one should remember that the statement 'S is true in *I*' is just a paraphrase of 'S, when interpreted as in *I*, is true'; so model-theoretic truth is parasitic on plain ordinary truth, and we can always paraphrase it away.

For example I might say

He is killing all of them,

and offer the interpretation that 'he' is Alfonso Arblaster of 35 The Crescent, Beetleford, and that 'them' are the pigeons in his loft. This interpretation explains (a) what objects some expressions refer to, and (b) what classes some quantifiers range over. (In this example there is one quantifier: 'all of them'). Interpretations that consist of items (a) and (b) appear very often in model theory, and they are known as *structures*. Particular kinds of model theory use particular kinds of structure; for example mathematical model theory tends to use so-called *first-order structures*, model theory of modal logics uses *Kripke structures*, and so on.

The structure *I* in the previous paragraph involves one fixed object and one fixed class. Since we described the structure today, the class is the class of pigeons in Alfonso's loft today, not those that will come tomorrow to replace them. If Alfonso Arblaster kills all the pigeons in his loft today, then *I* satisfies the quoted sentence today but won't satisfy it tomorrow, because Alfonso can't kill the same pigeons twice over. Depending on what you want to use model theory for, you may be happy to evaluate sentences today (the default time), or you may want to record how they are satisfied at one time and not at another. In the latter case you can relativise the notion of model and write '$I \models_t S$' to mean that *I* is a model of S at time *t*. The same applies to places, or to anything else that might be picked up by other implicit indexical features in the sentence. For example if you believe in possible worlds, you can index $\models$ by the possible world where the sentence is to be evaluated. Apart from using set theory, model theory is completely agnostic about what kinds of thing exist.

Note that the objects and classes in a structure carry labels that steer them to the right expressions in the sentence. These labels are an essential part of the structure.

If the same class is used to interpret all quantifiers, the class is called the *domain* or *universe* of the structure. But sometimes there are quantifiers ranging over different classes. For example if I say

One of those thingummy diseases is killing all the birds.

you will look for an interpretation that assigns a class of diseases to 'those thingummy diseases' and a class of birds to 'the birds'. Interpretations that give two or more classes for different quantifiers to range over are said to be *many-sorted*, and the classes are sometimes called the *sorts*.

The ideas above can still be useful if we start with a sentence S that does say something either true or false without needing further interpretation. (Model theorists say that such a sentence is *fully interpreted*.) For example we can consider *misinterpretations I* of a fully interpreted sentence S. A misinterpretation of S that makes it true is known as a *nonstandard* or *unintended* model of S. The branch of mathematics called nonstandard analysis is based on nonstandard models of mathematical statements about the real or complex number systems; see Section 4 below.

One also talks of *model-theoretic semantics* of natural languages, which is a way of *describing* the meanings of natural language sentences, not a way of *giving* them meanings. The connection between this semantics and model theory is a little indirect. It lies in Tarski's truth definition of 1933. See the entry on Tarski's truth definitions for more details.

2. Model-theoretic definition

A sentence S divides all its possible interpretations into two classes, those that are models of it and those that are not. In this way it defines a class, namely the class of all its models, written Mod(S). To take a legal example, the sentence

> The first person has transferred the property to the second person, who thereby holds the property for the benefit of the third person.

defines a class of structures which take the form of labelled 4-tuples, as for example (writing the label on the left):

- the first person = Alfonso Arblaster;
- the property = the derelict land behind Alfonso's house;
- the second person = John Doe;
- the third person = Richard Roe.

This is a typical model-theoretic definition, defining a class of structures (in this case, the class known to the lawyers as *trusts*).

We can extend the idea of model-theoretic definition from a single sentence S to a set T of sentences; Mod(T) is the class of all interpretations that are simultaneously models of all the sentences in T. When a set T of sentences is used to define a class in this way, mathematicians say that T is a *theory* or a *set of axioms*, and that T *axiomatises* the class Mod(T).

Take for example the following set of first-order sentences:

$$\forall x \forall y \forall z\ (x + (y + z) = (x + y) + z).$$
$$\forall x\ (x + 0 = x).$$
$$\forall x\ (x + (-x) = 0).$$
$$\forall x \forall y\ (x + y = y + x).$$

Here the labels are the addition symbol '+', the minus symbol '-' and the constant symbol '0'. An interpretation also needs to specify a domain for the quantifiers. With one proviso, the models of this set of sentences are precisely the structures that mathematicians know as *abelian groups*. The proviso is that in an abelian group *A*, the domain should contain the interpretation of the symbol 0, and it should be

closed under the interpretations of the symbols + and -. In mathematical model theory one builds this condition (or the corresponding one for other function and constant symbols) into the definition of a structure.

Each mathematical structure is tied to a particular first-order language. A structure contains interpretations of certain predicate, function and constant symbols; each predicate or function symbol has a fixed arity. The collection K of these symbols is called the *signature* of the structure. Symbols in the signature are often called *nonlogical constants*, and an older name for them is *primitives*. The first-order language of signature K is the first-order language built up using the symbols in K, together with the equality sign =, to build up its atomic formulas. (See the entry on classical logic.) If K is a signature, S is a sentence of the language of signature K and *A* is a structure whose signature is K, then because the symbols match up, we know that *A* makes S either true or false. So one defines the class of abelian groups to be the class of all those structures of signature +, -, 0 which are models of the sentences above. Apart from the fact that it uses a formal first-order language, this is exactly the algebraists' usual definition of the class of abelian groups; model theory formalises a kind of definition that is extremely common in mathematics.

Now the defining axioms for abelian groups have three kinds of symbol (apart from punctuation). First there is the logical symbol = with a fixed meaning. Second there are the nonlogical constants, which get their interpretation by being applied to a particular structure; one should group the quantifier symbols with them, because the structure also determines the domain over which the quantifiers range. And third there are the variables *x*, *y* etc. This three-level pattern of symbols allows us to define classes in a second way. Instead of looking for the interpretations of the nonlogical constants that will make a sentence true, we *fix* the interpretations of the nonlogical constants by choosing a particular structure *A*, and we look for assignments of elements of *A* to variables which will make a given formula true in *A*.

For example let ***Z*** be the additive group of integers. Its elements are the integers (positive, negative and 0), and the symbols +, -, 0 have their usual meanings. Consider the formula

$$v_1 + v_1 = v_2.$$

If we assign the number -3 to v_1 and the number -6 to v_2, the formula works out as true in ***Z***. We express this by saying that the pair (-3,-6) *satisfies* this formula *in **Z***. Likewise (15,30) and (0,0) satisfy it, but (2,-4) and (3,3) don't. Thus the formula *defines* a binary relation on the integers, namely the set of pairs of integers that satisfy it. A relation defined in this way in a structure *A* is called a *first-order definable relation in A*. A useful generalisation is to allow the defining formula to use added names for some specific elements of *A*; these elements are called *parameters* and the relation is then *definable with parameters*.

This second type of definition, defining relations inside a structure rather than classes of structure, also formalises a common mathematical practice. But this time the practice belongs to geometry rather than to algebra. You may recognise the relation in the field of real numbers defined by the formula

$$v_1^2 + v_2^2 = 1.$$

It's the circle of radius 1 around the origin in the real plane. Algebraic geometry is full of definitions of this kind.

During the 1940s it occurred to several people (chiefly Anatolii Mal'tsev in Russia, Alfred Tarski in the USA and Abraham Robinson in Britain) that the metatheorems of classical logic could be used to prove mathematical theorems about classes defined in the two ways we have just described. In 1950 both Robinson and Tarski were invited to address the International Congress of Mathematicians at Cambridge Mass. on this new discipline (which as yet had no name - Tarski proposed the name 'model theory' in 1954). The conclusion of Robinson's address to that Congress is worth quoting:

> [The] concrete examples produced in the present paper will have shown that contemporary symbolic logic can produce useful tools - though by no means omnipotent ones - for the development of actual mathematics, more particularly for the development of algebra and, it would appear, of algebraic geometry. This is the realisation of an ambition which was expressed by Leibnitz in a letter to Huyghens as long ago as 1679.

In fact Mal'tsev had already made quite deep applications of model theory in group theory several years earlier, but under the political conditions of the time his work in Russia was not yet known in the West. By the end of the twentieth century, Robinson's hopes had been amply fulfilled; see the entry on first-order model theory.

There are at least two other kinds of definition in model theory besides these two above. The third is known as *interpretation* (a special case of the interpretations that we began with). Here we start with a structure A, and we build another structure B whose signature need not be related to that of A, by defining the domain X of B and all the labelled relations and functions of B to be the relations definable in A by certain formulas with parameters. A further refinement is to find a definable equivalence relation on X and take the domain of B to be not X itself but the set of equivalence classes of this relation. The structure B built in this way is said to be *interpreted in* the structure A.

A simple example, again from standard mathematics, is the interpretation of the group $\boldsymbol{Z}$ of integers in the structure $\boldsymbol{N}$ consisting of the natural numbers 0, 1, 2 etc. with labels for 0, 1 and +. To construct the domain of $\boldsymbol{Z}$ we first take the set X of all ordered pairs of natural numbers (clearly a definable relation in $\boldsymbol{N}$), and on this set X we define the equivalence relation $\sim$ by

$$(a,b) \sim (c,d) \text{ if and only if } a + d = b + c$$

(again definable). The domain of $\boldsymbol{Z}$ consists of the equivalence classes of this relation. We define addition on $\boldsymbol{Z}$ by

$(a,b) + (c,d) = (e,f)$ if and only if $a + c + f = b + d + e$.

The equivalence class of (a,b) becomes the integer $a - b$.

When a structure *B* is interpreted in a structure *A*, every first-order statement about *B* can be translated back into a first-order statement about *A*, and in this way we can read off the complete theory of *B* from that of *A*. In fact if we carry out this construction not just for a single structure *A* but for a family of models of a theory T, always using the same defining formulas, then the resulting structures will all be models of a theory T′ that can be read off from T and the defining formulas. This gives a precise sense to the statement that the theory T′ is *interpreted in* the theory T. Philosophers of science have sometimes experimented with this notion of interpretation as a way of making precise what it means for one theory to be reducible to another. But realistic examples of reductions between scientific theories seem generally to be much subtler than this simple-minded model-theoretic idea will allow. See the entry on intertheory relations in physics.

The fourth kind of definability is a pair of notions, implicit definability and explicit definability of a particular relation in a theory. See section 3.3 of the entry on first-order model theory.

Unfortunately there used to be a very confused theory about model-theoretic axioms, that also went under the name of implicit definition. By the end of the nineteenth century, mathematical geometry had generally ceased to be a study of space, and it had become the study of classes of structures which satisfy certain 'geometric' axioms. Geometric terms like 'point', 'line' and 'between' survived, but only as the primitive symbols in axioms; they no longer had any meaning associated with them. So the old question, whether Euclid's parallel postulate (as a statement about space) was deducible from Euclid's other assumptions about space, was no longer interesting to geometers. Instead, geometers showed that if one wrote down an up-to-date version of Euclid's other assumptions, in the form of a theory T, then it was possible to find models of T which fail to satisfy the parallel postulate. (See the entry on geometry in the 19th century for the contributions of Lobachevski and Klein to this achievement.) In 1899 David Hilbert published a book in which he constructed such models, using exactly the method of interpretation that we have just described.

Problems arose because of the way that Hilbert and others described what they were doing. The history is complicated, but roughly the following happened. Around the middle of the nineteenth century people noticed, for example, that in an abelian group the minus function is definable in terms of 0 and + (namely: $-a$ is the element b such that $a + b = 0$). Since this description of minus is in fact one of the axioms defining abelian groups, we can say (using a term taken from J. D. Gergonne, who should not be held responsible for the later use made of it) that the axioms for abelian groups *implicitly define* minus. In the jargon of the time, one said not that the axioms define the function minus, but that they define the *concept* minus. Now suppose we switch around and try to define plus in terms of minus and 0. This way round it can't be done, since one can have two abelian groups with the same 0 and minus but different plus functions. Rather than say this, the nineteenth century mathematicians concluded that the axioms only partially define plus in terms of minus and 0. Having swallowed that much, they went on to say that

the axioms together form an implicit definition of the concepts plus, minus and 0 together, and that this implicit definition is only partial but it says about these concepts precisely as much as we need to know.

One wonders how it could happen that for fifty years nobody challenged this nonsense. In fact some people did challenge it, notably the geometer Moritz Pasch who in section 12 of his *Vorlesungen über Neuere Geometrie* (1882) insisted that geometric axioms tell us nothing whatever about the meanings of ‘point’, ‘line’ etc. Instead, he said, the axioms give us *relations* between the concepts. If one thinks of a structure as a kind of ordered *n*-tuple of sets etc., then a class Mod(T) becomes an *n*-ary relation, and Pasch’s account agrees with ours. But he was unable to spell out the details, and there is some evidence that his contemporaries (and some more recent commentators) thought he was saying that the axioms may not determine the meanings of ‘point’ and ‘line’, but they do determine those of relational terms such as ‘between’ and ‘incident with’! Frege’s demolition of the implicit definition doctrine was masterly, but it came too late to save Hilbert from saying, at the beginning of his *Grundlagen der Geometrie*, that his axioms give ‘the exact and mathematically adequate description’ of the relations ‘lie’, ‘between’ and ‘congruent’. Fortunately Hilbert’s mathematics speaks for itself, and one can simply bypass these philosophical faux pas. The model-theoretic account that we now take as a correct description of this line of work seems to have surfaced first in the group around Giuseppe Peano in the 1890s, and it reached the English-speaking world through Bertrand Russell’s *Principles of Mathematics* in 1903.

3. Model-theoretic consequence

Suppose L is a language of signature K, T is a set of sentences of L and φ is a sentence of L. Then the relation

$$\mathrm{Mod}(T) \subseteq \mathrm{Mod}(\varphi)$$

expresses that every structure of signature K which is a model of T is also a model of φ. This is known as the *model-theoretic consequence relation*, and it is written for short as

$$T \models \varphi$$

The double use of $\models$ is a misfortune. But in the particular case where L is first-order, the completeness theorem (see the entry on classical logic) tells us that ‘$T \models \varphi$’ holds if and only if there is a proof of φ from T, a relation commonly written

$$T \vdash \varphi$$

Since $\models$ and $\vdash$ express exactly the same relation in this case, model theorists often avoid the double use of $\models$ by using $\vdash$ for model-theoretic consequence. But since what follows is not confined to first-order languages, safety suggests we stick with $\models$ here.

Before the middle of the nineteenth century, textbooks of logic commonly taught the student how to check the validity of an argument (say in English) by showing that it has one of a number of standard forms, or by paraphrasing it into such a form. The standard forms were syntactic and/or semantic forms of argument in English. The process was hazardous: semantic forms are almost by definition not visible on the surface, and there is no purely syntactic form that guarantees validity of an argument. For this reason most of the old textbooks had a long section on ‘fallacies’ - ways in which an invalid argument may seem to be valid.

In 1847 George Boole changed this arrangement. For example, to validate the argument

> All monarchs are human beings. No human beings are infallible. Therefore no infallible beings are monarchs.

Boole would interpret the symbols P, Q, R as names of classes:

> P is the class of all monarchs.
> Q is the class of all human beings.
> R is the class of all infallible beings.

Then he would point out that the original argument paraphrases into a set-theoretic consequence:

$$(P \subseteq Q), (Q \cap R = 0) \models (R \cap P = 0)$$

(This example is from Stanley Jevons, 1869. Boole’s own account is idiosyncratic, but I believe Jevons’ example represents Boole’s intentions accurately.) Today we would write $\forall x\,(Px \rightarrow Qx)$ rather than $P \subseteq Q$, but this is essentially the standard definition of $P \subseteq Q$, so the difference between us and Boole is slight.

Insofar as they follow Boole, modern textbooks of logic establish that English arguments are valid by reducing them to model-theoretic consequences. Since the class of model-theoretic consequences, at least in first-order logic, has none of the vaguenesses of the old argument forms, textbooks of logic in this style have long since ceased to have a chapter on fallacies.

But there is one warning that survives from the old textbooks: If you formalise your argument in a way that is *not* a model-theoretic consequence, it doesn’t mean the argument is *not valid*. It may only mean that you failed to analyse the concepts in the argument deeply enough before you formalised. The old textbooks used to discuss this in a ragbag section called ‘topics’ (i.e. hints for finding arguments that you might have missed). Here is an example from Peter of Spain’s 13th century *Summulae Logicales*:

> ’There is a father. Therefore there is a child.’ ... Where does the validity of this argument

> come from? From the relation. The maxim is: When one of a correlated pair is posited, then so is the other.

Hilbert and Ackermann, possibly the textbook that did most to establish the modern style, discuss in their section III.3 a very similar example: ‘If there is a son, then there is a father’. They point out that any attempt to justify this by using the symbolism

$$\exists x \, S(x) \rightarrow \exists x \, F(x)$$

is doomed to failure. ‘A proof of this statement is possible only if we analyze conceptually the meanings of the two predicates which occur’, as they go on to illustrate. And of course the analysis finds precisely the relation that Peter of Spain referred to.

On the other hand if your English argument translates into an invalid model-theoretic consequence, a counterexample to the consequence may well give clues about how you can describe a situation that would make the premises of your argument true and the conclusion false. But this is not guaranteed.

One can raise a number of questions about whether the modern textbook procedure does really capture a sensible notion of logical consequence. For example in Boole’s case the set-theoretic consequences that he relies on are all easily provable by formal proofs in first-order logic, not even using any set-theoretic axioms; and by the completeness theorem (see the entry on classical logic) the same is true for first-order logic. But for some other logics it is certainly not true. For instance the model-theoretic consequence relation for some logics of time presupposes some facts about the physical structure of time. Also, as Boole himself pointed out, his translation from an English argument to its set-theoretic form requires us to believe that for every property used in the argument, there is a corresponding class of all the things that have the property. This comes dangerously close to Frege’s inconsistent comprehension axiom!

In 1936 Alfred Tarski proposed a definition of logical consequence for arguments in a fully interpreted formal language. His proposal was that an argument is valid if and only if: under any allowed reinterpretation of its nonlogical symbols, if the premises are true then so is the conclusion. Tarski assumed that the class of allowed reinterpretations could be read off from the semantics of the language, as set out in his truth definition. He left it undetermined what symbols count as nonlogical; in fact he hoped that this freedom would allow one to define different kinds of necessity, perhaps separating ‘logical’ from ‘analytic’. One thing that makes Tarski’s proposal difficult to evaluate is that he completely ignores the question we discussed above, of analysing the concepts to reach all the logical connections between them. The only plausible explanation I can see for this lies in his parenthetical remark about

> the necessity of eliminating any defined signs which may possibly occur in the sentences concerned, i.e. of replacing them by primitive signs.

This suggests to me that he wants his primitive signs to be *by stipulation* unanalysable. But then by

stipulation it will be purely accidental if his notion of logical consequence captures everything one would normally count as a logical consequence.

Historians note a resemblance between Tarski's proposal and one in section 147 of Bernard Bolzano's *Wissenschaftslehre* of 1837. Like Tarski, Bolzano defines the validity of a proposition in terms of the truth of a family of related propositions. Unlike Tarski, Bolzano makes his proposal for propositions in the vernacular, not for sentences of a formal language with a precisely defined semantics.

On all of this section, see also the entry on logical consequence.

4. Expressive strength

A sentence S defines its class Mod(S) of models. Given two languages L and L′, we can compare them by asking whether every class Mod(S), with S a sentence of L, is also a class of the form Mod(S′) where S′ is a sentence of L′. If the answer is Yes, we say that L is *reducible to* L′, or that L′ is *at least as expressive as* L.

For example if L is a first-order language with identity, whose signature consists of 1-ary predicate symbols, and L′ is the language whose sentences consist of the four syllogistic forms (All *A* are *B*, Some *A* are *B*, No *A* are *B*, Some *A* are not *B*) using the same predicate symbols, then L′ is reducible to L, because the syllogistic forms are expressible in first-order logic. (There are some quarrels about which is the right way to express them; see the entry on the traditional square of opposition.) But the first-order language L is certainly not reducible to the language L′ of syllogisms, since in L we can write down a sentence saying that exactly three elements satisfy P(*x*), and there is no way of saying this using just the syllogistic forms. Or moving the other way, if we form a third language L″ by adding to L the quantifier Q*x* with the meaning 'There are uncountably many elements such that ...', then trivially L is reducible to L″, but the downward Loewenheim-Skolem theorem shows at once that L″ is not reducible to *L*.

These notions are useful for analysing the strength of database query languages. We can think of the possible states of a database as structures, and a simple Yes/No query becomes a sentence that elicits the answer Yes if the database is a model of it and No otherwise. If one database query language is not reducible to another, then the second can express some query that can't be expressed in the first.

So we need techniques for comparing the expressive strengths of languages. One of the most powerful techniques available consists of the back-and-forth games of Ehrenfeucht and Fraïssé between the two players Spoiler and Duplicator; see the entry on logic and games for details. Imagine for example that we play the usual first-order back-and-forth game *G* between two structures *A* and *B*. The theory of these games establishes that if some first-order sentence φ is true in exactly one of *A* and *B*, then there is a number *n*, calculable from φ, with the property that Spoiler has a strategy for *G* that will guarantee that he wins in at most *n* steps. So conversely, to show that first-order logic can't distinguish between *A* and *B*, it suffices to show that for any finite *n*, Duplicator has a strategy that will guarantee she doesn't lose *G*

in the first n steps. If we succeed in showing this, it follows that any language which does distinguish between A and B is not reducible to the first-order language of the structures A and B.

These back-and-forth games are immensely flexible. For a start, they make just as much sense on finite structures as they do on infinite; many other techniques of classical model theory assume that the structures are infinite. They can also be adapted smoothly to many non-first-order languages.

In 1969 Per Lindström used back-and-forth games to give some abstract characterisations of first-order logic in terms of its expressive power. One of his theorems says that if L is a language with a signature K, L is closed under all the first-order syntactic operations, and L obeys the downward Loewenheim-Skolem theorem for single sentences, and the compactness theorem, then L is reducible to the first-order language of signature K. These theorems are very attractive; see Chapter XII of Ebbinghaus, Flum and Thomas for a good account. But they have never quite lived up to their promise. It has been hard to find any similar characterisations of other logics. Even for first-order logic it is a little hard to see exactly what the characterisations tell us. But very roughly speaking, they tell us that first-order logic is the unique logic with two properties: (1) we can use it to express arbitrarily complicated things about finite patterns, and (2) it is hopeless for discriminating between one infinite cardinal and another.

These two properties (1) and (2) are just the properties of first-order logic that allowed Abraham Robinson to build his *nonstandard analysis*. The background is that Leibniz, when he invented differential and integral calculus, used infinitesimals, i.e. numbers that are greater than 0 and smaller than all of 1/2, 1/3, 1/4 etc. Unfortunately there are no such real numbers. During the nineteenth century all definitions and proofs in the Leibniz style were rewritten to talk of limits instead of infinitesimals. Now let ***R*** be the structure consisting of the field of real numbers together with any structural features we care to give names to: certainly plus and times, maybe the ordering, the set of integers, the functions sin and log, etc. Let L be the first-order language whose signature is that of ***R***. Because of the expressive strength of L, we can write down any number of theorems of calculus as sentences of L. Because of the expressive weakness of L, there is no way that we can express in L that ***R*** has no infinitesimals. In fact Robinson used the compactness theorem to build a structure $\boldsymbol{R}'$ that is a model of exactly the same sentences of L as ***R***, but which has infinitesimals. As Robinson showed, we can copy Leibniz's arguments using the infinitesimals in $\boldsymbol{R}'$, and so prove that various theorems of calculus are true in $\boldsymbol{R}'$. But these theorems are expressible in L, so they must also be true in ***R***.

Since arguments using infinitesimals are usually easier to visualise than arguments using limits, nonstandard analysis is a helpful tool for mathematical analysts. Jacques Fleuriot in his recent PhD thesis automated the proof theory of nonstandard analysis and used it to mechanise some of the proofs in Newton's *Principia*.

5. Models and modelling

To *model* a phenomenon is to construct a formal theory that describes and explains it. In a closely related sense, you *model* a system or structure that you plan to build, by writing a description of it. These are

very different senses of ‘model’ from that in model theory: the ‘model’ of the phenomenon or the system is not a structure but a theory, often in a formal language. The *Universal Modeling Language*, UML for short, is a formal language designed for just this purpose. It’s reported that the Australian Navy once hired a model theorist for a job ‘modelling hydrodynamic phenomena’. (Please don’t enlighten them!)

In cognitive science the difference between models and modelling has become blurred. A central question of cognitive science is how we represent facts or possibilities in our minds. If one formalises these mental representations, they become something like ‘models of phenomena’. But it is a serious hypothesis that in fact our mental representations have a good deal in common with simple set-theoretic structures, so that they are ‘models’ in the model-theoretic sense too. In 1983 two influential works of cognitive science were published, both under the title *Mental Models*. The first, edited by Dedre Gentner and Albert Stevens, was about people’s ‘conceptualizations’ of the elementary facts of physics; it belongs squarely in the world of ‘modelling of phenomena’. The second, by Philip Johnson-Laird, is largely about reasoning, and makes considerable use of ‘model-theoretic semantics’ in our sense. Researchers in the Johnson-Laird tradition tend to refer to their approach as ‘model theory’, and to see it as allied in some sense to what we have called model theory. (The book by Alan Garnham in the references is a recent work in this line.)

Pictures and diagrams seem at first to hover in the middle ground between theories and models. In practice model theorists often draw themselves pictures of structures, and use the pictures to think about the structures. On the other hand pictures don’t generally carry the labelling that is an essential feature of model-theoretic structures. There is a fast growing body of work on reasoning with diagrams, and the overwhelming tendency of this work is to see pictures and diagrams as a form of language rather than as a form of structure. For example Eric Hammer and Norman Danner (in the book edited by Allwein and Barwise, see the Bibliography) describe a ‘model theory of Venn diagrams’; the Venn diagrams themselves are the syntax, and the model theory is a set-theoretical explanation of their meaning.

The model theorist Yuri Gurevich introduced *abstract state machines* (ASMs) as a way of using model-theoretic ideas for specification in computer science. According to the Abstract State Machine website (see Other Internet Resources below),

> One uses a specification methodology to describe a system by means of a particular syntax and associated semantics. If the semantics of the specification methodology is unclear, descriptions using the methodology may be no clearer than the original systems being described. ASMs use classical mathematical structures to describe states of a computation; structures are well-understood, precise models.

The book of Staerk et al. in the Bibliography is an example of ASMs at work.

Today you can make your name and fortune by finding a good representation system. There is no reason to expect that every such system will fit neatly into the syntax/semantics framework of model theory, but it will be surprising if model-theoretic ideas don’t continue to make a major contribution in this area.

Bibliography

Introductory texts

- Doets, K. 1996, *Basic Model Theory*, Stanford : CSLI Publications.
- Hodges, W. 1997, *A Shorter Model Theory*, Cambridge : Cambridge University Press.
- Manzano, M. 1999, *Model Theory*, Oxford : Oxford University Press.
- Rothmaler, P. 2000, *Introduction to Model Theory,* Amsterdam : Gordon and Breach.

Model-theoretic definition

- Gergonne, J. 1818, "Essai sur la théorie de la définition", *Annales de Mathématiques Pures et Appliquées*, **9**, 1-35.
- Hilbert, D. 1899, *Grundlagen der Geometrie*, Leipzig: Teubner.
- Lascar, D. 1998, "Perspective historique sur les rapports entre la théorie des modèles et l'algèbre', *Revue d'histoire des mathématiques* **4**, 237-260.
- Pasch, M. 1882, *Vorlesungen über Neuere Geometrie*, Springer-Verlag, Berlin.
- Robinson, A. 1952, "On the application of symbolic logic to algebra", *Proceedings of the International Congress of Mathematicians, Cambridge, Mass. 1950, Vol. 1,* Providence RI : American Mathematical Society, 686-694.
- Suppes, P. 1957, Chapter 8 "Theory of definition" in *Introduction to Logic*, Princeton NJ : Van Nostrand.
- Tarski, A. 1954, "Contributions to the theory of models, I", *Indagationes Mathematicae* **16**, 572-581.

Model-theoretic consequence

- Blanchette, P. 1996, "Frege and Hilbert on consistency", *The Journal of Philosophy* **93**, 317-336.
- Boole, G. 1847, *The Mathematical Analysis of Logic*, Cambridge : Macmillan, Barclay and Macmillan.
- Etchemendy, J. 1990, *The Concept of Logical Consequence*, Cambridge Mass. : Harvard University Press.
- Frege, G. 1971, *On the Foundations of Geometry, and Formal Theories of Arithmetic,* trans. Kluge, E., New Haven Conn. : Yale University Press.
- Gómez-Torrente, M. 1996, "Tarski on logical consequence", *Notre Dame Journal of Formal Logic* **37**, 125-151.
- Hilbert, D. and Ackermann, W. 1950, *Principles of Mathematical Logic* (translated from 1938 German edition), New York : Chelsea Publishing Company.
- Kreisel, G. 1969, "Informal rigour and completeness proofs", in Hintikka, J. ed., *The Philosophy of Mathematics*, London : Oxford University Press, 78-94.

- Tarski, A. 1983, "On the concept of logical consequence", translated in Tarski, A., *Logic, Semantics, Metamathematics*, ed. Corcoran, J., Indianapolis, Indiana : Hackett, 409-420.

Expressive strength

- Ebbinghaus, H.-D., and Flum, J. 1999, *Finite Model Theory*, Berlin : Springer-Verlag.
- Ebbinghaus, H.-D., Flum, J. and Thomas, W. 1984, *Mathematical Logic*, New York : Springer-Verlag.
- Fleuriot, J. 2001, *A Combination of Geometry Theorem Proving and Nonstandard Analysis, with Application to Newton's Principia*, New York : Springer-Verlag.
- Immerman, N. 1999, *Descriptive Complexity*, New York : Springer-Verlag.
- Loeb, P. and Wolff, M. eds. 2000, *Nonstandard Analysis for the Working Mathematician*, Dordrecht : Kluwer.
- Robinson, A. 1967, "The metaphysics of the calculus", in *Problems in the Philosophy of Mathematics*, ed. Lakatos I., Amsterdam : North-Holland, 28-40.

Models and modelling

- Allwein., G. and Barwise, J. eds. 1996, *Logical Reasoning with Diagrams*, New York, Oxford University Press.
- Fowler, M. 2000, *UML Distilled*, Boston : Addison-Wesley.
- Garnham, A. 2001, *Mental Models and the Interpretation of Anaphora*, Philadelphia PA : Taylor and Francis.
- Gentner, D. and Stevens, A. eds. 1983, *Mental Models*, Hillsdale NJ : Lawrence Erlbaum.
- Johnson-Laird, P. 1983, *Mental Models: Towards a cognitive science of language, inference, and consciousness*, Cambridge : Cambridge University Press.
- Staerk, R., Schmid, J. and Boerger, E. 2001, *Java and the Java Virtual Machine - Definition, Verification, Validation*, New York : Springer-Verlag.

Other Internet Resources

- Abstract State Machine website, by Jim Huggins (Computer Science Program, Science and Mathematics Department, Kettering University)

Related Entries

diagrams | geometry: in the 19th century | logic: classical | logical consequence | physics: intertheory relations in | square of opposition | Tarski, Alfred: truth definitions

Wilfrid Hodges
School of Mathematical Sciences,
Queen Mary, University of London
w.hodges@qmw.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 9, 2001
Content last modified: November 9, 2001

Tarski's Truth Definitions

In 1933 the Polish logician Alfred Tarski published a paper in which he discussed the criteria that a definition of 'true sentence' should meet, and gave examples of several such definitions for particular formal languages. In 1956 he and his colleague Robert Vaught published a revision of one of the 1933 truth definitions, to serve as a truth definition for model-theoretic languages. In this entry, we simply review the definitions and make no attempt to explore the implications of Tarski's work for semantics (natural language or programming languages) or for the philosophical study of truth.

1. The 1933 programme and the semantic conception

In the 1920s Alfred Tarski embarked on a project to give rigorous definitions for notions useful in scientific methodology. In 1933 he published (in Polish) his analysis of the notion of a true sentence. This long paper undertook two tasks: first to say what should count as a satisfactory definition of 'true sentence' for a given formal language, and second to show that there do exist satisfactory definitions of 'true sentence' for a range of formal languages. We begin with the first task; Section 2 will consider the second.

We say that a language is *fully interpreted* if all its sentences have meanings that make them either true or false. All the languages that Tarski considered in the 1933 paper were fully interpreted, with one exception described in Section 2.2 below. This was the main difference between the 1933 definition and the later model-theoretic definition of 1956, which we shall examine in Section 3.

Tarski described several conditions that a satisfactory definition of truth should meet.

1.1 Object language and metalanguage

If the language under discussion (the *object language*) is L, then the definition should be given in another language known as the *metalanguage*, call it M. The metalanguage should contain a copy of the object language (so that anything one can say in L can be said in M too), and M should also be able to talk about the sentences of L and their syntax. Finally Tarski allowed M to contain notions from set theory, and a 1-ary predicate symbol *True* with the intended reading 'is a true sentence of L'. The main purpose of the metalanguage was to formalise what was being said about the object language, and so Tarski also required that the metalanguage should carry with it a set of axioms expressing everything that one needs to assume for purposes of defining and justifying the truth definition. The truth definition itself was to be a definition of *True* in terms of the other expressions of the metalanguage. So the definition was to be in terms of syntax, set theory and the notions expressible in L, but not semantic notions like 'denote' or 'mean' (unless the object language happened to contain these notions).

Tarski assumed, in the manner of his time, that the object language L and the metalanguage M would be languages of some kind of higher order logic. Today it is more usual to take some kind of informal set theory as one's metalanguage; this would affect some details of Tarski's paper but not its main thrust. Also today it is usual to define syntax in set-theoretic terms, so that for example a string of letters becomes a sequence. In fact one must use a set-theoretic syntax if one wants to work with an object language that has uncountably many symbols, as model theorists have done freely for over half a century now.

1.2 Formal correctness

The definition of *True* should be 'formally correct'. This means that it should be a sentence of the form

For all x, *True*(x) if and only if $\varphi(x)$,

where *True* never occurs in φ; or failing this, that the definition should be provably equivalent to a sentence of this form. The equivalence must be provable using axioms of the metalanguage that don't contain *True*. Definitions of the kind displayed above are usually called *explicit*, though Tarski in 1933 called them *normal*.

1.3 Material adequacy

The definition should be 'materially adequate'. This means that the objects satisfying φ should be exactly the objects that we would intuitively count as being true sentences of L, and that this fact should be provable from the axioms of the metalanguage. At first sight this is a paradoxical requirement: if we can prove what Tarski asks for, just from the axioms of the metalanguage, then we must already have a materially adequate formalisation of 'true sentence of L' within the metalanguage, suggesting an infinite

regress. In fact Tarski escapes the paradox by using (in general) infinitely many sentences of M to express truth, namely all the sentences of the form

$\varphi(s)$ if and only if ψ

whenever *s* is the name of a sentence S of L and ψ is the copy of S in the metalanguage. So the technical problem is to find a single formula φ that allows us to deduce all these sentences from the axioms of M; this formula φ will serve to give the explicit definition of *True*.

Tarski's own name for this criterion of material adequacy was *Convention T*. More generally his name for his approach to defining truth, using this criterion, was *the semantic conception of truth*.

As Tarski himself emphasised, Convention T rapidly leads to the liar paradox if the language L has enough resources to talk about its own semantics. (See the entry on the revision theory of truth.) Tarski's own conclusion was that a truth definition for a language L has to be given in a metalanguage which is essentially stronger than L.

There is a consequence for the foundations of mathematics. First-order Zermelo-Fraenkel set theory is widely regarded as the standard of mathematical correctness, in the sense that a proof is correct if and only if it can be formalised as a formal proof in set theory. We would like to be able to give a truth definition for set theory; but by Tarski's result this truth definition can't be given in set theory itself. The usual solution is to give the truth definition informally in English. But there are a number of ways of giving limited formal truth definitions for set theory. For example Azriel Levy showed that for every natural number *n* there is a Σ_n formula that is satisfied by all and only the set-theoretic names of true Σ_n sentences of set theory. The definition of Σ_n is too technical to give here, but three points are worth making. First, every sentence of set theory is provably equivalent to a Σ_n sentence for any large enough *n*. Second, the class of Σ_n formulas is closed under adding existential quantifiers at the beginning, but not under adding universal quantifiers. Third, the class is not closed under negation; this is how Levy escapes Tarski's paradox. (See the entry on set theory.) Essentially the same devices allow Jaakko Hintikka to give an internal truth definition for his Independence-Friendly Logic; this logic shares the second and third properties of Levy's classes of formulas.

2. Some kinds of truth definition on the 1933 pattern

In his 1933 paper Tarski went on to show that many fully interpreted formal languages do have a truth definition that satisfies his conditions. He gave four examples in that paper. One was a trivial definition for a finite language; it simply listed the finitely many true sentences. One was a definition by quantifier elimination; see Section 2.2 below. The remaining two, for different classes of language, were examples

of what people today think of as the standard Tarski truth definition; they are forerunners of the 1956 model-theoretic definition.

2.1 The standard truth definitions

The two standard truth definitions are at first glance not definitions of truth at all, but definitions of a more complicated relation involving assignments *a* of objects to variables:

> *a* satisfies the formula F.

In fact satisfaction reduces to truth in this sense: *a* satisfies F if and only if taking each free variable in F as a name of the object assigned to it by *a* makes F into a true sentence. So it follows that our intuitions about when a sentence is true can guide our intuitions about when an assignment satisfies a formula. But none of this can enter into the formal definition of truth, because ‘taking a variable as a name of an object’ is a semantic notion, and Tarski’s truth definition has to be built only on notions from syntax and set theory (together with those in the object language); recall Section 1.1. In fact Tarski’s reduction goes in the other direction: if F has no free variables, then to say that F is true is to say that every assignment satisfies it.

The reason why Tarski defines satisfaction directly, and then deduces a definition of truth, is that satisfaction obeys *recursive conditions* in the following sense: if F is a compound formula, then to know which assignments satisfy F, it’s enough to know which assignments satisfy the immediate constituents of F. Here are two typical examples:

- The assignment *a* satisfies the formula ‘F *and* G’ if and only if *a* satisfies F and *a* satisfies G.
- The assignment *a* satisfies the formula ‘*For all x*, G’ if and only if for every individual *i*, if *b* is the assignment that assigns *i* to the variable *x* and is otherwise exactly like *a*, then *b* satisfies G.

We have to use a different approach for atomic formulas. But for these, at least assuming for simplicity that L has no function symbols, we can use the metalanguage copies #(R) of the predicate symbols R of the object language. Thus:

- The assignment *a* satisfies the formula R(*x*,*y*) if and only if #(R)(*a*(*x*),*a*(*y*)).

(Warning: the expression # is in the metametalanguage, not in the metalanguage M. We may or may not be able to find a formula of M that expresses # for predicate symbols; it depends on exactly what the language L is.)

One sometimes says that Tarski’s definition of satisfaction is *compositional*, meaning that the class of assignments which satisfy a compound formula F is determined solely by (1) the syntactic rule used to construct F from its immediate constituents and (2) the classes of assignments that satisfy these immediate constituents. (This is sometimes phrased loosely as: satisfaction is defined recursively. But

this formulation misses the central point, that (1) and (2) don't contain any syntactic information about the immediate constituents.) Compositionality explains why Tarski switched from truth to satisfaction. You can't define whether '*For all x*, G' is true in terms of whether G is true, because in general G has a free variable *x* and so it isn't either true or false.

The name 'compositionality' is from a paper of Katz and Fodor in 1963 on natural language semantics. In talking about compositionality, we have moved to thinking of Tarski's definition as a semantics, i.e. a way of assigning 'meanings' to formulas. (Here we take the meaning of a sentence to be its truth value.) Compositionality means essentially that the meanings assigned to formulas give *at least* enough information to determine the truth values of sentences containing them. One can ask conversely whether Tarski's semantics provides *only as much information as we need* about each formula, in order to reach the truth values of sentences. If the answer is yes, we say that the semantics is *fully abstract* (for truth). One can show fairly easily, for any of the standard languages of logic, that Tarski's definition of satisfaction is in fact fully abstract.

As it stands, Tarski's definition of satisfaction is not an explicit definition, because satisfaction for one formula is defined in terms of satisfaction for other formulas. So to show that it is formally correct, we need a way of converting it to an explicit definition. One way to do this is as follows, using either higher order logic or set theory. Suppose we write S for a binary relation between assignments and formulas. We say that S is a *satisfaction relation* if for every formula G, S meets the conditions put for satisfaction of G by Tarski's definition. For example, if G is 'G_1 *and* G_2', S should satisfy the following condition for every assignment *a*:

> $S(a,G)$ if and only if $S(a,G_1)$ and $S(a,G_2)$.

We can define 'satisfaction relation' formally, using the recursive clauses and the conditions for atomic formulas in Tarski's recursive definition. Now we prove, by induction on the complexity of formulas, that there is exactly one satisfaction relation S. (There are some technical subtleties, but it can be done.) Finally we define

> *a* satisfies F if and only if: there is a satisfaction relation S such that $S(a,F)$.

It is then a technical exercise to show that this definition of satisfaction is materially adequate. Actually one must first write out the counterpart of Convention T for satisfaction of formulas, but I leave this to the reader.

2.2 The truth definition by quantifier elimination

The remaining truth definition in Tarski's 1933 paper -- the third as they appear in the paper - is really a bundle of related truth definitions, all for the same object language L but in different interpretations. The quantifiers of L are assumed to range over a particular class, call it *A*; in fact they are second order

quantifiers, so that really they range over the collection of subclasses of *A*. The class *A* is not named explicitly in the object language, and thus one can give separate truth definitions for different values of *A*, as Tarski proceeds to do. So for this section of the paper, Tarski allows one and the same sentence to be given different interpretations; this is the exception to the general claim that his object language sentences are fully interpreted. But Tarski stays on the straight and narrow: he talks about 'truth' only in the special case where *A* is the class of all individuals. For other values of *A*, he speaks not of 'truth' but of 'correctness in the domain *A*'.

These truth or correctness definitions don't fall out of a definition of satisfaction. In fact they go by a much less direct route, which Tarski describes as a 'purely accidental' possibility that relies on the 'specific peculiarities' of the particular object language. It may be helpful to give a few more of the technical details than Tarski does, in a more familiar notation than Tarski's, in order to show what is involved. Tarski refers his readers to a paper of Thoralf Skolem in 1919 for the technicalities.

One can think of the language L as the first-order language with predicate symbols ⊆ and =. The language is interpreted as talking about the subclasses of the class *A*. In this language we can define:

- '*x* is the empty set' (viz. *x* ⊆ every class).
- '*x* is an atom' (viz. *x* is not empty, but every subclass of *x* not equal to *x* is empty).
- '*x* has exactly *k* members' (where *k* is a finite number; viz. there are exactly *k* distinct atoms ⊆ *x*).
- 'There are exactly *k* elements in *A*' (viz. there is a class with exactly *k* members, but there is no class with exactly *k*+1 members).

Now we aim to prove:

> *Lemma*. Every formula F of L is equivalent to (i.e. is satisfied by exactly the same assignments as) some boolean combination of sentences of the form 'There are exactly *k* elements in *A*' and formulas of the form 'There are exactly *k* elements that are in v_1, not in v_2, not in v_3 and in v_4' (or any other combination of this type, using only variables free in F).

The proof is by induction on the complexity of formulas. For atomic formulas it is easy. For boolean combinations of formulas it is easy, since a boolean combination of boolean combinations is again a boolean combination. For formulas beginning with ∀, we take the negation. This leaves just one case that involves any work, namely the case of a formula beginning with an existential quantifier. By induction hypothesis we can replace the part after the quantifier by a boolean combination of formulas of the kinds stated. So a typical case might be:

> ∃*z* (there are exactly two elements that are in *z* and *x* and not in *y*).

This holds if and only if there are at least two elements that are in *x* and not in *y*. We can write this in turn

as: The number of elements in *x* and not in *y* is not 0 and is not 1; which is a boolean combination of allowed formulas. The general proof is very similar but more complicated.

When the lemma has been proved, we look at what it says about a sentence. Since the sentence has no free variables, the lemma tells us that it is equivalent to a boolean combination of statements saying that *A* has a given finite number of elements. So if we know how many elements *A* has, we can immediately calculate whether the sentence is 'correct in the domain *A*'.

One more step and we are home. As we prove the lemma, we should gather up any facts that can be stated in L, are true in every domain, and are needed for proving the lemma. For example we shall almost certainly need the sentence saying that $\subseteq$ is transitive. Write T for the set of all these sentences. (In Tarski's presentation T vanishes, since he is using higher order logic and the required statements about classes become theorems of logic.) Thus we reach, for example:

> *Theorem*. If the domain *A* is infinite, then a sentence S of the language L is correct in *A* if and only if S is deducible from T and the sentences saying that the number of elements of *A* is not any finite number.

The class of *all* individuals is infinite (Tarski asserts), so the theorem applies when *A* is this class. And in this case Tarski has no inhibitions about saying not just 'correct in *A*' but 'true'; so we have our truth definition.

The method we have described revolves almost entirely around removing existential quantifiers from the beginnings of formulas; so it is known as *the method of quantifier elimination*. It is not as far as you might think from the two standard definitions. In all cases Tarski assigns to each formula, by induction on the complexity of formulas, a description of the class of assignments that satisfy the formula. In the two previous truth definitions this class is described directly; in the quantifier elimination case it is described in terms of a boolean combination of formulas of a simple kind.

At around the same time as he was writing the 1933 paper, Tarski gave a truth definition by quantifier elimination for the first-order language of the field of real numbers. He published it separately, and at first only as an interesting way of characterising the relations definable by formulas. Later he gave a fuller account, emphasising that his method provided not just a truth definition but an algorithm for determining which sentences about the real numbers are true and which are false.

3. The 1956 definition and its offspring

In 1933 Tarski assumed that the formal languages that he was dealing with had two kinds of symbol (apart from punctuation), namely constants and variables. The constants included logical constants, but also any other terms of fixed meaning. The variables had no independent meaning and were simply part of the apparatus of quantification.

Model theory by contrast works with three levels of symbol. There are the logical constants (=, ¬, ∧ for example), the variables (as before), and between these a middle group of symbols which have no fixed meaning but get a meaning through being applied to a particular structure. The symbols of this middle group include the nonlogical constants of the language, such as relation symbols, function symbols and constant individual symbols. They also include the quantifier symbols ∀ and ∃, since we need to refer to the structure to see what set they range over. This type of three-level language corresponds to mathematical usage; for example we write the addition operation of an abelian group as +, and this symbol stands for different functions in different groups.

So one has to work a little to apply the 1933 definition to model-theoretic languages. There are basically two approaches: (1) Take one structure *A* at a time, and regard the nonlogical constants as constants, interpreted in *A*. (2) Regard the nonlogical constants as variables, and use the 1933 definition to describe when a sentence is satisfied by an assignment of the ingredients of a structure *A* to these variables. There are problems with both these approaches, as Tarski himself describes in several places. The chief problem with (1) is that in model theory we very frequently want to use the same language in connection with two or more different structures - for example when we are defining elementary embeddings between structures (see the entry on first-order model theory). The problem with (2) is more abstract: it is disruptive and bad practice to talk of formulas with free variables being 'true'. (We saw in Section 2.2 how Tarski avoided talking about truth in connection with sentences that have varying interpretations.) What Tarski did in practice, from the appearance of his textbook in 1936 to the late 1940s, was to use a version of (2) and simply avoid talking about model-theoretic sentences being true in structures; instead he gave an indirect definition of what it is for a structure to be a 'model of' a sentence, and apologised that strictly this was an abuse of language. (Chapter VI of Tarski 1994 still contains relics of this old approach.)

By the late 1940s it had become clear that a direct model-theoretic truth definition was needed. Tarski and colleagues experimented with several ways of casting it. The version we use today is based on that published by Tarski and Robert Vaught in 1956. See the entry on classical logic for an exposition.

The right way to think of the model-theoretic definition is that we have sentences whose truth value varies according to the situation where they are used. So the nonlogical constants are not variables; they are definite descriptions whose reference depends on the context. Likewise the quantifiers have this indexical feature, that the domain over which they range depends on the context of use. In this spirit one can add other kinds of indexing. For example a Kripke structure is an indexed family of structures, with a relation on the index set; these structures and their close relatives are fundamental for the semantics of modal, temporal and intuitionist logic.

Already in the 1950s model theorists were interested in formal languages that include kinds of expression different from anything in Tarski's 1933 paper. Extending the truth definition to infinitary logics was no problem at all. Nor was there any serious problem about most of the generalised quantifiers proposed at the time. For example there is a quantifier Q*xy* with the intended meaning:

> $QxyF(x,y)$ if and only if there is an infinite set X of elements such that for all a and b in X, $F(a,b)$.

This definition itself shows at once how the required clause in the truth definition should go.

In 1961 Leon Henkin pointed out two sorts of model-theoretic language that didn't immediately have a truth definition of Tarski's kind. The first had infinite strings of quantifiers:

$$\forall v_1\ \exists v_2\ \forall v_3\ \exists v_4 \ldots R(v_1,v_2,v_3, v_4,\ldots).$$

The second had quantifiers that are not linearly ordered. For ease of writing I use Hintikka's later notation for these:

$$\forall v_1\ \exists v_2\ \forall v_3\ (\exists v_4/\forall v_1)\ R(v_1,v_2,v_3, v_4).$$

Here the slash after $\exists v_4$ means that this quantifier is outside the scope of the earlier quantifier $\forall v_1$ (and also outside that of the earlier existential quantifier).

Henkin pointed out that in both cases one could give a natural semantics in terms of Skolem functions. For example the second sentence can be paraphrased as

$$\exists f \exists g\ \forall v_1\ \forall v_3\ R(v_1,f(v_1),v_3,g(v_3)),$$

which has a straightforward Tarski truth condition in second order logic. Hintikka then observed that one can read the Skolem functions as winning strategies in a game, as in (the entry on) logic and games. In this way one can build up a compositional semantics, by assigning to each formula a game. A sentence is true if and only if the player Myself (in Hintikka's nomenclature) has a winning strategy for the game assigned to the sentence. This game semantics agrees with Tarski's on conventional first-order sentences. But it is far from fully abstract; probably one should think of it as an operational semantics, describing how a sentence is verified rather than whether it is true.

The problem of giving a Tarski-style semantics for Henkin's two languages turned out to be different in the two cases. With the first, the problem is that the syntax of the language is not well-founded: there is an infinite descending sequence of subformulas as one strips off the quantifiers one by one. Hence there is no hope of giving a definition of satisfaction by recursion on the complexity of formulas. The remedy is to note that the *explicit* form of Tarski's truth definition in Section 2.1 above didn't require a recursive definition; it needed only that the conditions on the satisfaction relation S pin it down uniquely. For Henkin's first style of language this is still true, though the reason is no longer the well-foundedness of the syntax.

For Henkin's second style of language, at least in Hintikka's notation, the syntax is well-founded, but the displacement of the quantifier scopes means that the usual quantifier clauses in the definition of satisfaction no longer work. To get a compositional and fully abstract semantics, one has to ask not what assignments of variables satisfy a formula, but what *sets* of assignments satisfy the formula 'uniformly', where 'uniformly' means 'independent of assignments to certain variables, as shown by the slashes on quantifiers inside the formula'. Henkin's second example is of more than theoretical interest, because clashes between the semantic and the syntactic scope of quantifiers occur very often in natural languages.

Bibliography

- Henkin, L. 1961, "Some remarks on infinitely long formulas", in *Infinitistic methods: Proceedings of the symposium on foundations of mathematics*, Oxford: Pergamon Press,167-183.
- Hintikka, J. 1996, *The Principles of Mathematics Revisited*, Cambridge : Cambridge University Press.
- Hodges, W. 1997, "Compositional semantics for a language of imperfect information", *Logic Journal of the Interest Group in Propositional and Predicate Logic* **5**, 539-563.
- Katz, J. and Fodor, J. 1963, "The structure of a semantic theory", *Language* **39**, 170-210.
- Levy, A. 1965, *A hierarchy of formulas in set theory*, Memoirs of American Mathematical Society **57**.
- Skolem, T. 1919, "Untersuchungen über die Axiome des Klassenkalküls und über Produktations- und Summationsprobleme, welche gewisse Klassen von Aussagen betreffen", *Videnskapsselskapets Skrifter, I. Matem.-naturv. klasse* **3**.
- Tarski, A. 1933, "The concept of truth in the languages of the deductive sciences" (Polish), *Prace Towarzystwa Naukowego Warszawskiego, Wydzial III Nauk Matematyczno-Fizycznych* **34**, Warsaw; reprinted in Zygmunt 1995, pages 13-172; expanded English translation in Tarski 1983, pages 152-278.
- Tarski, A. 1944, "The semantic conception of truth", *Philosophy and Phenomenlological Research* **4**, 13-47.
- Tarski, A. 1983, *Logic, Semantics, Metamathematics, papers from 1923 to 1938*, ed. John Corcoran, Indianapolis : Hackett Publishing Company.
- Tarski, A. 1994, *Introduction to Logic and to the Methodology of the Deductive Sciences* (new edition of book originally published in Polish in 1936), New York, Oxford University Press.
- Tarski, A. and Vaught, R. 1956, "Arithmetical extensions of relational systems", *Compositio Mathematica* **13**, 81-102.
- Zygmunt, J. ed. 1995, *Alfred Tarski, Pisma Logiczno-Filozoficzne, 1 Prawda*, Warsaw : Wydawnictwo Naukowe PWN.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

logic: and games | logic: infinitary | logic: intuitionistic | logic: temporal | model theory | model theory: first-order | truth: deflationary theory of | truth: revision theory of

Wilfrid Hodges
School of Mathematical Sciences,
Queen Mary, University of London
w.hodges@qmul.ac.uk

First published: November 9, 2001
Content last modified: November 9, 2001

Logic and games

Games between two players, of the kind where one player wins and one loses, became a familiar tool in many branches of logic during the second half of the twentieth century. Important examples are semantic games used to define truth, back-and-forth games used to compare structures, and dialogue games to express (and perhaps explain) formal proofs.

1. Games in the History of Logic

The links between logic and games go back a long way. If one thinks of a debate as a kind of game, then Aristotle already made the connection; his writings about syllogism are closely intertwined with his study of the aims and rules of debating. Aristotle's viewpoint survived into the common medieval name for logic: *dialectics*. Ramist logic, which drove out the old School logic in the mid sixteenth century, aimed to provide a set of practical tools for the orator and the debater. That view of logic had faded away by the eighteenth century. But in the mid twentieth century Charles Hamblin revived the link between dialogue and the rules of sound reasoning, soon after Paul Lorenzen had connected dialogue to constructive foundations of logic.

A different strand, not quite so old, is the use of games for teaching logic. This is probably the right way to think of the medieval game of ‘obligationes’, where a debater tries to drive his opponent into an unnecessary contradiction. We have at least two textbooks of logic from the early sixteenth century that

present it as a game for an individual student, and Lewis Carroll's *The Game of Logic* (1887) is a more recent example in the same genre.

Mathematical game theory was founded in the early twentieth century. Although no mathematical links with logic emerged until the 1950s, it is striking how many of the early pioneers of game theory are also known for their contributions to logic: John Kemeny, J. C. C. McKinsey, John von Neumann, Willard Quine, Julia Robinson, Ernst Zermelo and others. In 1953 David Gale and Frank Stewart made fruitful connections between set theory and games. Shortly afterwards Leon Henkin suggested a way of using games to give semantics for infinitary languages.

The first half of the twentieth century was an era of increasing rigour and professionalism in logic, and to most logicians of that period the use of games in logic would probably have seemed frivolous. The intuitionist L. E. J. Brouwer expressed this attitude when he accused his opponents of causing mathematics ‘to degenerate into a game’ (as David Hilbert quoted him in 1927). Wittgenstein's language games provoked little response from the logicians. But in the second half of the century the centre of gravity of logical research moved from foundations to techniques, and from about 1960 games were used more and more often in logical papers.

2. Logical Games

From the point of view of game theory, the games that appear in logic are not at all typical. They normally involve just two players, they often have infinite length, the only outcomes are winning and losing, and no probabilities are attached to actions or outcomes. The barest essentials of a logical game are as follows.

There are two players. In general we can call them $\forall$ and $\exists$. The pronunciations ‘Abelard’ and ‘Eloise’ go back to the mid 1980s and usefully fix the players as male and female (though feminist logicians have asked about the propriety of this type-casting). Other names are in common use for the players in particular types of logical game.

The players play by choosing elements of a set Ω, called the *domain* of the game. As they choose, they build up a sequence

$$a_0, a_1, a_2, \ldots$$

of elements of Ω. Infinite sequences of elements of Ω are called *plays*. Finite sequences of elements of Ω are called *positions*; they record where a play might have got to by a certain time. A function τ (the *turn function* or *player function*) takes each position **a** to either $\exists$ or $\forall$; if τ(**a**) = $\exists$, this means that when the game has reached **a**, player $\exists$ makes the next choice (and likewise with $\forall$). The game rules define two sets $W\forall$ and $W\exists$ consisting of positions and plays, with the following properties: if a position **a** is in $W\forall$ then so is any play or longer position that starts with **a** (and likewise with $W\exists$); and no play is in both W

∀ and W∃. We say that player ∀ *wins* a play **b**, and that **b** is a *win* for ∀, if **b** is in W∀; if some position **a** that is an initial segment of **b** is in W∀, then we say that player ∀ *wins already at* **a**. (And likewise with ∃ and W∃.) So to summarise, a logical game is a 4-tuple (Ω, τ, W∀, W∃) with the properties just described.

We say that a logical game is *total* if every play is in either W∀ or W∃, so that there are no draws. Unless one makes an explicit exception, logical games are always assumed to be total. (Don't confuse being total with the much stronger property of being determined -- see below.)

It is only for mathematical convenience that the definition above expects the game to continue to infinity even when a player has won at some finite position; there is no interest in anything that happens after a player has won. Many logical games have the property that in every play, one of the players has already won at some finite position; games of this sort are said to be *well-founded*. An even stronger condition is that there is some finite number *n* such that in every play, one of the players has already won by the *n*-th position; in this case we say that the game has *finite length*.

A *strategy* for a player is a set of rules that describe exactly how that player should choose, depending on how the other player has chosen at earlier moves. Mathematically, a strategy for ∀ consists of a function which takes each position **a** with τ(**a**) = ∀ to an element *b* of Ω; we think of it as an instruction to ∀ to choose *b* when the game has reached the position **a**. (Likewise with a strategy for ∃.) A strategy for a player is said to be *winning* if that player wins every play in which he or she uses the strategy, regardless of what the other player does. At most one of the players has a winning strategy (since otherwise the players could play their winning strategies against each other, and both would win, contradicting that W∀ and W∃ have no plays in common). Occasionally one meets situations in which it seems that two players have winning strategies (for example in the forcing games below), but closer inspection shows that the two players are in fact playing different games.

A game is said to be *determined* if one or other of the players has a winning strategy. There are many examples of games that are not determined, as Gale and Stewart showed in 1953 using the axiom of choice. This discovery led to important applications of the notion of determinacy in the foundations of set theory. Gale and Stewart also proved an important theorem that bears their name: Every well-founded game is determined. It follows that every game of finite length is determined -- a fact already known to Zermelo in 1913. (A more precise statement of the Gale-Stewart theorem is this. A game G is said to be *closed* if ∃ wins every play of G in which she hasn't lost at any finite position. The theorem states that every closed game is determined.)

In most applications of logical games, the central notion is that of a winning strategy for the player ∃. Often these strategies (or their existence) turn out to be equivalent to something of logical importance that could have been defined without using games -- for example a proof. But games are felt to give a better definition because they quite literally supply some motivation: ∃ is trying to win. This raises a question that is not of much interest mathematically, but it should concern philosophers who use logical games. If we want ∃'s motivation in a game G to have any explanatory value, then we need to understand

what is achieved if ∃ does win. In particular we should be able to tell a realistic story of a situation in which some agent called ∃ is trying to do something intelligible, and doing it is the same thing as winning in the game. As Richard Dawkins said, raising the corresponding question for the evolutionary games of Maynard Smith,

> The whole purpose of our search ... is to discover a suitable actor to play the leading role in our metaphors of purpose. We ... want to say, ‘It is for the good of ... ‘. Our quest in this chapter is for the right way to complete that sentence. (*The Extended Phenotype*, Oxford University Press, Oxford 1982, page 91.)

For future reference let us call this the *Dawkins question*. In many kinds of logical game it turns out to be distinctly harder to answer than the pioneers of these games realised.

Just as in classical game theory, the definition of logical games above serves as a clothes horse that we can hang other concepts onto. For example it is common to have some laws that describe what elements of Ω are available for a player to choose at a particular move. Strictly this refinement is unnecessary, because the winning strategies are not affected if we decree instead that a player who breaks the law loses immediately; but for many games this way of viewing them seems unnatural. (For example in Jaakko Hintikka's semantic games, some steps expect a player to choose an element of the structure, whereas other steps require a player to choose a formula. We may as well make it a law that the players must choose elements of these kinds.) A subtler extension is to restrict the information available to the players, so that the games are of imperfect information. Hintikka has explored this possibility in his Game-Theoretic Semantics.

3. Semantic Games

In the early 1930s Alfred Tarski proposed a definition of truth. His definition consisted of a necessary and sufficient condition for a sentence in the language of a typical formal theory to be true; his necessary and sufficient condition used only notions from syntax and set theory, together with the primitive notions of the formal theory in question. In fact Tarski defined the more general relation ‘formula $\varphi(x_1,...,x_n)$ is true of the elements $a_1,..., a_n$’; truth of a sentence is the special case where $n = 0$. For example the question whether

> ‘For all *x* there is *y* such that R(*x*,*y*)’ is true

reduces to the question whether the following holds:

> For every object *a* the sentence ‘There is *y* such that R(*a*,*y*)’ is true.

This in turn reduces to:

For every object *a* there is an object *b* such that the sentence ‘R(*a*,*b*)’ is true.

In this example, that's as far as Tarski's truth definition will take us.

In the late 1950s Leon Henkin noticed that we can intuitively understand some sentences which can't be handled by Tarski's definition. Take for example the infinitely long sentence

For all x_0 there is y_0 such that for all x_1 there is y_1 such that ... $R(x_0,y_0,x_1, y_1,...)$.

Tarski's approach fails because the string of quantifiers at the beginning is infinite, and we would never reach an end of stripping them off. Instead, Henkin suggested, we should consider the game where a person ∀ chooses an object a_0 for x_0, then a second peson ∃ chooses an object b_0 for y_0, then ∀ chooses a_1 for x_1, ∃ chooses b_1 for y_1 and so on. A play of this game is a win for ∃ if and only if the infinite atomic sentence

$R(a_0,b_0,a_1, b_1,...)$

is true. The original sentence is true if and only if player ∃ has a winning strategy for this game. Strictly Henkin used the game only as a metaphor, and the truth condition that he proposed was that the skolemised version of the sentence is true, i.e. that there are functions f_0, f_1, ... such that for every choice of a_0, a_1, a_2 etc. we have

$R(a_0,f_0 (a_0),a_1,f_1 (a_0,a_1),a_2, f_2(a_0,a_1,a_2), ...)$.

But this condition translates immediately into the language of games; the Skolem functions f_0 etc. are a winning strategy for ∃, telling her how to choose in the light of earlier choices by ∀. (It came to light sometime later that C. S. Peirce had already suggested explaining the difference between ‘every’ and ‘some’ in terms of who chooses the object; for example in his second Cambridge Conference lecture of 1898.)

Soon after Henkin's work, Jaakko Hintikka added that the same idea applies with conjunctions and disjunctions. We can regard a conjunction ‘$\varphi \wedge \psi$’ as a universally quantified statement expressing ‘Every one of the sentences φ, ψ holds’, so it should be for the player ∀ to choose one of the sentences. As Hintikka put it, to play the game $G(\varphi \wedge \psi)$, ∀ chooses whether the game should proceed as $G(\varphi)$ or as $G(\psi)$. Likewise disjunctions become existentially quantified statements about sets of sentences, and they mark moves where the player ∃ chooses how the game should proceed. To bring quantifiers into the same style, he proposed that the game $G(\forall x\, \varphi(x))$ proceeds thus: player ∀ chooses an object and provides a name *a* for it, and the game proceeds as $G(\varphi(a))$. (And likewise with existential quantifiers, except that ∃ chooses.) Hintikka also made an ingenious suggestion for introducing negation. Each game G has a *dual game* which is the same as G except that the players ∀ and ∃ are transposed in both the rules

for playing and the rules for winning. The game G(¬φ) is the dual of G(φ).

One can prove that for any first-order sentence φ, interpreted in a fixed structure *A*, player ∃ has a winning strategy for Hintikka's game G(φ) if and only if φ is true in *A* in the sense of Tarski. Two features of this proof are interesting. First, if φ is any first-order sentence then the game G(φ) has finite length, and so the Gale-Stewart theorem tells us that it is determined. We infer that ∃ has a winning strategy in exactly one of G(φ) and its dual; so she has a winning strategy in G(¬φ) if and only if she doesn't have one in G(φ). This takes care of negation. And second, if ∃ has a winning strategy for each game G(φ(*a*)), then after choosing one such strategy f_a for each *a*, she can string them together into a single winning strategy for G(∀*x* φ(*x*)) (namely, ‘Wait and see what element *a* ∀ chooses, then play f_a’). This takes care of the clause for universal quantifiers; but the argument uses the axiom of choice, and in fact it is not hard to see that the equivalence of Hintikka's and Tarski's definitions of truth is equivalent to the axiom of choice (given the other axioms of Zermelo-Fraenkel set theory).

It's puzzling that we have here two theories of when a sentence is true, and the theories are not equivalent if the axiom of choice fails. In fact the reason is not very deep. The axiom of choice is needed not because the Hintikka definition uses games, but because it assumes that strategies are deterministic, i.e. that they are single-valued functions giving the user no choice of options. A more natural way of translating the Tarski definition into game terms is to use nondeterministic strategies. (However, Hintikka insists that the correct explication of ‘true’ is the one using deterministic strategies, and so it is Tarski's definition that works only accidentally.)

Computer implementations of these games of Hintikka proved to be a very effective way of teaching the meanings of first-order sentences. One such package was designed by Jon Barwise and John Etchemendy at Stanford and marketed as ‘Tarski's World’. Independently another team at the University of Omsk constructed a Russian version for use at schools for gifted children.

In the published version of his John Locke lectures at Oxford, Hintikka in 1973 raised the Dawkins question (see above) for these games. His answer was that one should look to Wittgenstein's language games, and the language games for understanding quantifiers are those which revolve around seeking and finding. In the corresponding logical games one should think of ∃ as Myself and ∀ as a hostile Nature who can never be relied on to present the object I want; so to be sure of finding it, I need a winning strategy. This story was never very convincing; the motivation of Nature is irrelevant, and nothing in the logical game corresponds to seeking. In retrospect it is a little disappointing that nobody took the trouble to look for a better story. It may be more helpful to think of a winning strategy for ∃ in G(φ) as a kind of proof (in a suitable infinitary system) that φ is true.

Later Jaakko Hintikka extended the ideas of this section in two directions, namely to natural language semantics and to games of imperfect information. The name *Game-Theoretic Semantics*, GTS for short, has come to be used to cover both of these extensions.

The games described in this section adapt almost trivially to many-sorted logic: for example the

quantifier $\forall x$, where x is a variable of sort σ, is an instruction for player $\forall$ to choose an element of sort σ. This immediately gives us the corresponding games for second-order logic, if we think of the elements of a structure as one sort, the sets of elements as a second sort, the binary relations as a third and so on. It follows that we have, quite routinely, game rules for most generalised quantifiers too; we can find them by first translating the generalised quantifiers into second-order logic.

4. Modal Semantics

Structures of the following kind give rise to interesting games. The structure *A* consists of a set *S* of elements (which we shall call *states*, adding that they are often called *worlds*), a binary relation *R* on *S* (we shall read *R* as *arrow*), and a family P_1,..., P_n of subsets of *S*. The two players $\forall$ and $\exists$ play a game G on *A*, starting at a state *s* which is given them, by reading a suitable logical formula φ as a set of instructions for playing and for winning.

Thus if φ is P_i, then player $\exists$ wins at once if *s* is in P_i, and otherwise player $\forall$ wins at once. The formulas $\psi \wedge \theta$, $\psi \vee \theta$ and $\neg \psi$ behave as in Hintikka's games above; for example $\psi \wedge \theta$ instructs player $\forall$ to choose whether the game shall continue as for ψ or for θ. If the formula φ is $\Box \psi$, then player $\forall$ chooses an arrow from *s* to a state *t* (i.e. a state *t* such that the pair (*s*, *t*) is in the relation *R*), and the game then proceeds from the state *t* according to the instructions ψ. The rule for $\Diamond \psi$ is the same except that player $\exists$ makes the choice. Finally we say that the formula φ is *true at s in A* if player $\exists$ has a winning strategy for this game based on φ and starting at *s*.

These games stand to modal logic in very much the same way as Hintikka's games stand to first-order logic. In particular they are one way of giving a semantics for modal logic, and they agree with the usual Kripke-type semantics. Of course there are many types and generalisations of modal logic (including closely related logics such as temporal, epistemic and dynamic logics), and so the corresponding games come in many different forms. One example of interest is the computer-theoretic logic of Martin Hennessy and Robin Milner, used for describing the behaviour of systems; here the arrows come in more than one colour, and moving along an arrow of a particular colour represents performing a particular ‘action’ to change the state. In the related ’logic of games’, proposed by Rohit Parikh, games that move us between the states are the subject matter rather than a way of giving a truth definition.

5. Back-and-Forth Games

In 1930 Alfred Tarski formulated the notion of two structures *A* and *B* being *elementarily equivalent*, i.e. that exactly the same first-order sentences are true in *A* as are true in *B*. At a conference in Princeton in 1946 he described this notion and expressed the hope that it would be possible to develop a theory of it that would be ‘as deep as the notions of isomorphism, etc. now in use’ (Hourya Sinaceur ed., "Address at the Princeton University Bicentennial Conference on Problems of Mathematics (December 17-19, 1946), by Alfred Tarski", *Bulletin of Symbolic Logic* **6** (2000): 1-44).

One natural part of such a theory would be a purely structural necessary and sufficient condition for two structures to be elementarily equivalent. Roland Fraïssé, a French-Algerian, was the first to find a usable necessary and sufficient condition. It was rediscovered a few years later by the Kazak logician A. D. Taimanov, and it was reformulated in terms of games by the Polish logician Andrzej Ehrenfeucht. The games are now known as *Ehrenfeucht-Fraïssé* games, or sometimes as *back-and-forth* games. They have turned out to be one of the most versatile ideas in twentieth-century logic. They adapt fruitfully to a wide range of logics and structures.

In a back-and-forth game there are two structures *A* and *B*, and two players who are commonly called Spoiler and Duplicator. (The name is due to Joel Spencer in the early 1990s. More recently Neil Immerman suggested Samson and Delilah, using the same initials; this places Spoiler as the male player ∀ and Duplicator as the female ∃.) Each step in the game consists of a move of Spoiler, followed by a move of Duplicator. Spoiler chooses an element of one of the two structures, and Duplicator must then choose an element of the other structure. So after *n* steps, two sequences have been chosen, one from *A* and one from *B*:

$$(a_0,\ldots,a_{n-1};b_0,\ldots,b_{n-1}).$$

This position is a win for Spoiler if and only if some atomic formula (of one of the forms ‘$R(v_0,\ldots,v_{k-1})$’ or ‘$F(v_0,\ldots,v_{k-1}) = v_k$’ or ‘$v_0=v_1$’, or one of these with different variables) is satisfied by $(a_0,\ldots,a_{n-1})$ in *A* but not by $(b_0,\ldots,b_{n-1})$ in *B*, or vice versa. The condition for Duplicator to win is different in different forms of the game. In the simplest form, EF(*A*,*B*), a play is a win for Duplicator if and only if no initial part of it is a win for Spoiler (i.e. she wins if she hasn't lost by any finite stage). For each natural number *m* there is a game $EF_m(A,B)$; in this game Duplicator wins after *m* steps provided she has not yet lost. All these games are determined, by the Gale-Stewart Theorem. The two structures *A* and *B* are said to be *back-and-forth equivalent* if Duplicator has a winning strategy for EF(*A*,*B*), and *m-equivalent* if she has a winning strategy for $EF_m(A,B)$.

One can prove that if *A* and *B* are *m*-equivalent for every natural number *m*, then they are elementarily equivalent. On the other hand a winning strategy for Spoiler in $EF_m(A,B)$ can be converted into a first-order sentence that is true in exactly one of *A* and *B*, and in which the nesting of quantifier scopes has at most *m* levels. So we have our necessary and sufficient condition for elementary equivalence, and a bit more besides.

If *A* and *B* are back-and-forth equivalent, then certainly they are elementarily equivalent; but in fact back-and-forth equivalence turns out to be the same as elementary equivalence in an infinitary logic which is much more expressive than first-order logic. There are many adjustments of the game that give other kinds of equivalence. For example Barwise, Immerman and Bruno Poizat independently described a game in which the two players have exactly *p* numbered pebbles each; each player has to label his or her choices with a pebble, and the two choices in the same step must be labelled with pebbles carrying the

same number. As the game proceeds, the players will run out of pebbles and so they will have to re-use pebbles that were already used. The condition for Spoiler to win at a position (and all subsequent positions) is the same as before, except that only the elements carrying labels at that position count. The existence of a winning strategy for Duplicator in this game means that the two structures agree for sentences which use at most *p* variables (allowing these variables to occur any number of times).

The theory behind back-and-forth games uses very few assumptions about the logic in question. As a result, these games are one of the few model-theoretic techniques that apply as well to finite structures as they do to infinite ones, and this makes them one of the cornerstone of theoretical computer science. One can use them to measure the expressive strength of formal languages, for example database query languages. A typical result might say, for example, that a certain language can't distinguish between ‘even’ and ‘odd’; we would prove this by finding, for each level *n* of complexity of formulas of the language, a pair of finite structures for which Duplicator has a winning strategy in the back-and-forth game of level *n*, but one of the structures has an even number of elements and the other has an odd number.

There is also a kind of back-and-forth game that corresponds to our modal semantics above in the same way as Ehrenfeucht-Fraïssé games correspond to Hintikka's game semantics for first-order logic. The players start with a state *s* in the structure *A* and a state *t* in the structure *B*. Spoiler and Duplicator move alternately, as before. Each time he moves, Spoiler chooses whether to move in *A* or in *B*, and then Duplicator must move in the other structure. A move is always made by going forwards along an arrow from the current state. If between them the two players have just moved to a state *s*' in *A* and a state *t*' in *B*, and some predicate P_i holds at just one of *s*' and *t*', then Duplicator loses at once. Also she loses if there are no available arrows for her to move along; but if Spoiler finds there are no available arrows for him to move along in either structure, then Duplicator wins. If the two players play this game with given starting states *s* in *A* and *t* in *B*, and both structures have just finitely many states, then one can show that Duplicator has a winning strategy if and only if the same modal sentences are true at *s* in *A* as are true at *t* in *B*.

There are many generalisations of this result, some of them involving the following notion. Let *Z* be a binary relation which relates states of *A* to states of *B*. Then we call *Z* a *bisimulation* between *A* and *B* if Duplicator can use *Z* as a nondeterministic winning strategy in the back-and-forth game between *A* and *B* where the first pair of moves of the two players is to choose their starting states. In computer science the notion of a bisimulation is crucial for the understanding of *A* and *B* as systems; it expresses that the two systems interact with their environment in the same way as each other, step for step. But a little before the computer scientists introduced the notion, essentially the same concept appeared in Johan van Benthem's PhD thesis on the semantics of modal logic (1976).

6. Dialogue Games and Proof-theoretic Semantics

Imagine ∃ taking an oral examination in proof theory. The examiner gives her a sentence and invites her to start proving it. If the sentence has the form

$$\varphi \vee \psi$$

then she is entitled to choose one of the sentences and say ‘OK, I'll prove this one’. (In fact if the examiner is an intuitionist, he may insist that she choose one of the sentences to prove.) On the other hand if the sentence is

$$\varphi \wedge \psi$$

then the examiner, being an examiner, might well choose one of the conjuncts himself and invite her to prove that one. If she knows how to prove the conjunction then she certainly knows how to prove the conjunct.

The case of $\varphi \rightarrow \psi$ is a little subtler. She will probably want to start by assuming φ in order to deduce ψ; but there is some risk of confusion because the sentences that she has written down so far are all of them things to be proved, and φ is not a thing to be proved. The examiner can help her by saying ‘I'll assume φ, and let's see if you can get to φ from there’. At this point there is a chance that she sees a way of getting to φ by deducing a contradiction from ψ; so she may turn the tables on the examiner and invite him to show that his assumption is consistent, with a view to proving that it isn't. The symmetry is not perfect: he was asking her to show that a sentence is true everywhere, while she is inviting him to show that a sentence is true somewhere. Nevertheless we can see a sort of duality.

Ideas of this kind lie behind the dialectical games of Paul Lorenzen. He showed that with a certain amount of pushing and shoving, one can write rules for the game which have the property that ∃ has a winning strategy if and only if the sentence that she is presented with at the beginning is a theorem of intuitionistic logic. In a gesture towards medieval debates, he called ∃ the Proponent and the other player the Opponent. (In the medieval version they are usually Respondens and Opponens.) Almost as in the medieval obligationes, the Opponent wins by driving the Proponent to a point where the only moves available to her are blatant self-contradictions.

Lorenzen claimed that the rules of his games could be justified on a pre-logical basis, and so they formed a foundation for logic. Unfortunately any ‘justification’ involves a convincing answer to the Dawkins question, and this Lorenzen never provided. For example he spoke of moves as ‘attacks’, even when (like the examiner's choice at $\varphi \wedge \psi$ above) they look more like help than hostility. To repair Lorenzen's omission, one certainly needs to distinguish between different stances that a person might take in an argument: stating, assuming, conceding, querying, attacking, committing oneself. Whether it is really possible to define all these notions in a pre-logical way is a moot point. But perhaps this is unimportant. A more positive view is that this kind of refinement serves to link Lorenzen's dialogues to informal logic, and especially to the research that aims to systematise the possible structures of sound informal argument.

In any case, Lorenzen's games stand as an important paradigm of what recent proof theorists have called *semantics of proofs*. A semantics of proofs gives a ‘meaning’ not just to the notion of being provable, but to each separate step in a proof. It answers the question ‘What do we achieve by making this particular move in the proof?’ During the 1990s a number of workers at the logical end of computer science looked for games that would stand to linear logic and some other proof systems in the same way as Lorenzen's games stood to intuitionist logic. Andreas Blass, and then later Samson Abramsky and colleagues, gave games that corresponded to parts of linear logic, but at the time of writing we don't yet have a perfect correspondence between game and logic. This example is particularly interesting because the answer to the Dawkins question should give an intuitive interpretation of the laws of linear logic, a thing that this logic has badly needed. The games of Abramsky et al. tell a story about two interacting systems. But while he began with games in which the players politely take turns, Abramsky's more recent work allows the players to act ‘in a distributed, asynchronous fashion’, taking notice of each other only when they choose to. These games are no longer in the normal format of logical games, and their real-life interpretation raises a host of new questions.

7. Other Model-theoretic Games

The logical games in this section are mathematicians' tools, but they have some conceptually interesting features.

Forcing games

Forcing games are also known to descriptive set theorists as *Banach-Mazur games*; see the references by Kechris or Oxtoby below for more details of the mathematical background. Model theorists use them as a way of building infinite structures with controlled properties. To sketch the idea, imagine that a countably infinite team of builders are building a house *A*. Each builder has his or her own task to carry out: for example to install a bath or to wallpaper the entrance hall. Each builder has infinitely many chances to enter the site and add some finite amount of material to the house; these slots for the builders are interleaved so that the whole process takes place in a sequence of steps counted by the natural numbers.

To show that the house can be built to order, we need to show that each builder separately can carry out his or her appointed task, regardless of what the other builders do. So we imagine each builder as player ∃ in a game where all the other players are lumped together as ∀, and we aim to prove that ∃ has a winning strategy for this game. When we have proved this for each builder separately, we can imagine them going to work, each with their own winning strategy. They all win their respective games and the result is one beautiful house.

More technically, the elements of the structure *A* are fixed in advance, say as a_0, a_1, a_2 etc., but the properties of these elements have to be settled by the play. Each player moves by throwing in a set of atomic or negated atomic statements about the elements, subject only to the condition that the set

consisting of all the statements thrown in so far must be consistent with a fixed set of axioms written down before the game. (So throwing in a negated atomic sentence $\neg\varphi$ has the effect of preventing any player from adding φ at a later stage.) At the end of the joint play, the set of atomic sentences thrown in has a canonical model, and this is the structure *A*; there are ways of ensuring that it is a model of the fixed set of axioms. A possible property P of *A* is said to be *enforceable* if a builder who is given the task of making P true of *A* has a winning strategy. A central point (due essentially to Ehrenfeucht) is that the conjunction of a countably infinite set of enforceable properties is again enforceable.

The name ‘forcing’ comes from an application of related ideas by Paul Cohen to construct models of set theory in the early 1960s. Abraham Robinson adapted it to make a general method for building countable structures, and Martin Ziegler introduced the game setting. More recently Robin Hirsch and Ian Hodkinson have used related games to settle some old questions about relation algebras.

Forcing games are a healthy example to bear in mind when thinking about the Dawkins question. They remind us that in logical games it need not be helpful to think of the players as opposing each other.

Cut-and-choose games

In the traditional cut-and-choose game you take a piece of cake and cut it into two smaller pieces; then I choose one of the pieces and eat it, leaving the other one for you. This procedure is supposed to put pressure on you to cut the cake fairly. Mathematicians, not quite understanding the purpose of the exercise, insist on iterating it. Thus I make you cut the piece I chose into two, then I choose one of those two; then you cut this piece again, and so on indefinitely. Some even more unworldly mathematicians make you cut the cake into countably many pieces instead of two.

These games are important in the theory of definitions. Suppose we have a collection *A* of objects and a family *S* of properties; each property cuts *A* into the set of those objects that have the property and the set of those that don't. Let ∃ cut, starting with the whole set *A* and using a property in *S* as a knife; let ∀ choose one of the pieces (which are subsets of *A*) and give it back to ∃ to cut again, once more using a property in *S*; and so on. Let ∃ lose as soon as ∀ chooses an empty piece. We say that (*A*,*S*) has *rank* at most *m* if ∀ has a strategy which ensures that ∃ will lose before her *m*-th move. The rank of (*A*,*S*) gives valuable information about the family of subsets of *A* definable by properties in *S*.

Variations of this game, allowing a piece to be cut into infinitely many smaller pieces, are fundamental in the branch of model theory called *stability theory*. Broadly speaking, a theory is ‘good’ in the sense of stability theory if, whenever we take a model *A* of the theory and *S* the set of first-order formulas in one free variable with parameters from *A*, the structure (*A*,*S*) has ‘small’ rank. A different variation is to require that at each step, ∃ divides into two each of the pieces that have survived from earlier steps, and again she loses as soon as one of the cut fragments is empty. (In this version ∀ is redundant.) With this variation, the rank of (*A*,S) is called its *Vapnik-Chervonenkis dimension*; this notion is used in computational learning theory.

Games on the tree of two successor functions

Imagine a tree that has been built up in levels. At the bottom level there is a single root node, but a left branch and a right branch come up from it. At the next level up there are two nodes, one on each branch, and from each of these nodes a left branch and a right branch grow up. So on the next level up there are four nodes, and again the tree branches into left and right at each of these nodes. Continued to infinity, this tree is called the *tree of two successor functions* (namely left successor and right successor). Taking the nodes as elements and introducing two function symbols for left and right successor, we have a structure. A powerful theorem of Michael Rabin states that there is an algorithm which will tell us, for every monadic second-order sentence φ in the language appropriate for this structure, whether or not φ is true in the structure. ('Monadic second-order' means that the logic is like first-order, except that we can also quantify over sets of elements -- but not over binary relations on elements, for example.)

Rabin's theorem has any number of useful consequences. For example Dov Gabbay used it to prove the decidability of some modal logics. But Rabin's proof, using automata, was notoriously difficult to follow. Yuri Gurevich and Leo Harrington, and independently Andrei Muchnik, found much simpler proofs in which the automaton is a player in a game. This result is one of several that connect games with automata.

Bibliography

Logical Games in General

- David Gale and F. M. Stewart, "Infinite games with perfect information", in *Contributions to the Theory of Games II*, ed. H. W. Kuhn and A. W. Tucker, Annals of Mathematics Studies 28, Princeton University Press, Princeton NJ (1953): 245-266.
- Alexander S. Kechris, *Classical Descriptive Set Theory*, Springer, New York 1995.
- *Infinitistic Methods* (editors unnamed), Pergamon Press, Oxford 1961. (*Seminal papers of Henkin and Lorenzen appear in this book.*)

Semantic Games

- Jon Barwise and John Etchemendy, *Tarski's World*, software available from CSLI or Cambridge University Press.
- Leon Henkin, "Some remarks on infinitely long formulas", in *Infinitistic Methods* (above, 1961): 167-183.
- Jaakko Hintikka, *Logic, Language-Games and Information: Kantian Themes in the Philosophy of Logic*, Clarendon Press, Oxford 1973.
- Jaakko Hintikka and Gabriel Sandu, "Game-theoretical semantics", in *Handbook of Logic and Language*, ed. Johan van Benthem and Alice ter Meulen, Elsevier, Amsterdam (1997): 361-410.
- Wilfrid Hodges, "Elementary Predicate Logic 25: Skolem Functions", in *Handbook of*

Philosophical Logic I, Elements of Classical Logic, ed. Dov Gabbay and Franz Guenthner, Reidel, Dordrecht (1983): 92-97 (*Proof of equivalence of game and Tarski semantics*).
- Charles Sanders Peirce, *Reasoning and the Logic of Things: The Cambridge Conferences Lectures of 1898*, ed. Kenneth Laine Ketner, Harvard University Press, Cambridge Mass. 1992.

Modal Semantics

- Martin Hennessy and Robin Milner, "Algebraic laws for indeterminism and concurrency", *Journal of the ACM* **32** (1985): 137-162.
- Rohit Parikh, "The logic of games and its applications", in "Topics in the Theory of Computation", ed. Marek Karpinski and Jan van Leeuwen, *Annals of Discrete Mathematics* **24** (1985): 111-140.

Back-and-Forth

- Johan van Benthem, "Correspondence Theory", in *Handbook of Philosophical Logic II, Extensions of Classical Logic*, ed. Dov Gabbay and Franz Guenthner, Reidel, Dordrecht (1984): 167-247.
- Patrick Blackburn, Maarten de Rijke and Yde Venema, *Modal Logic*, Cambridge University Press, Cambridge (2001).
- Kees Doets, *Basic Model Theory*, CSLI Publications and FoLLI, Stanford 1996.
- Heinz-Dieter Ebbinghaus and Jörg Flum, *Finite Model Theory*, Springer, New York, Second edition 1999.
- Andrzej Ehrenfeucht, "An application of games to the completeness problem for formalized theories", *Fundamenta Mathematicae* **49** (1961): 129-141.
- Martin Otto, *Bounded Variable Logics and Counting -- A Study in Finite Models*, Lecture Notes in Logic **9**, Springer-Verlag 1997.

Dialogue and Communication Games

- Samson Abramsky and Radha Jagadeesan, "Games and full completeness for multiplicative linear logic", *Journal of Symbolic Logic* **59** (1994): 543-574.
- Samson Abramsky and Paul-André Melliès, "Concurrent games and full completeness’’, in *Proceedings of the Fourteenth International Symposium on Logic in Computer Science*, Computer Science Press of the IEEE (1999): 431-442.
- Andreas Blass, "A game semantics for linear logic", *Annals of Pure and Applied Logic* **56** (1992): 183-220.
- Walter Felscher, "Dialogues as a foundation for intuitionistic logic", in *Handbook of Philosophical Logic III, Alternatives to Classical Logic*, ed. Dov Gabbay and Franz Guenthner, Reidel, Dordrecht (1986): 341-372.
- Charles Hamblin, *Fallacies*, Methuen, London (1970).

- Paul Lorenzen, "Ein dialogisches Konstruktivitätskriterium", in *Infinitistic Methods* (above, 1961): 193-200.
- Douglas N. Walton and Erik C. W. Krabbe, *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*, State University of New York Press, Albany 1995.

Other Model-Theoretic Games

- Martin Anthony and Norman Biggs, *Computational Learning Theory*, Cambridge University Press, Cambridge (1992) (*for Vapnik-Chervonenkis dimension*).
- Yuri Gurevich and Leo Harrington, "Trees, automata, and games", in *Proceedings of the ACM Symposium on the Theory of Computing*, ed. H. R. Lewis, ACM, San Francisco (1984): 171-182.
- Robin Hirsch and Ian Hodkinson, *Relation Algebras by Games* (to appear).
- Wilfrid Hodges, *Building Models by Games*, Cambridge University Press, Cambridge (1985).
- Wilfrid Hodges, *Model Theory*, Cambridge University Press, Cambridge (1993).
- J. C. Oxtoby, *Measure and Category*, Springer-Verlag, New York (1971).
- Martin Ziegler, "Algebraisch abgeschlossene Gruppen", in *Word Problems II, The Oxford Book*, ed. S. I. Adian et al., North-Holland, Amsterdam (1980): 449-576.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

game theory | generalized quantifiers | logic: classical | logic: infinitary | logic: informal | logic: intuitionistic | logic: modal | model theory | proof theory | set theory

Wilfrid Hodges
w.hodges@qmw.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 27, 2001
Content last modified: July 27, 2001

Replication

The basic idea of replication has a long history in biology, beginning with the traditional distinction between homocatalysis and heterocatalysis and then later between transcription and translation. The basic distinction that these pairs of terms were designed to indicate is between like producing like (homocatalysis and transcription) and like producing unlike (heterocatalysis and translation). The paradigm example of this distinction is the contrast between genes and organisms. Genes perform two functions. They make other genes and, by means of the developmental process, they help produce organisms and their phenotypic traits (for a recent discussion of these two processes at the molecular level, see Rebek 1994 and Cook 1999).

As pervasive as this terminology has been, it engendered very little controversy until Richard Dawkins introduced the distinction between replicators and vehicles in his *The Selfish Gene* (1976). For his purposes Dawkins found the contrast between genes and organisms too restrictive. Everyone agrees that genes are replicators, but genes may not be the only replicators. Perhaps more inclusive entities than single genes might also function as replicators. At the very least, this possibility should not be defined out of existence. Hence, Dawkins adopted “replicator” as a more inclusive term than “gene.” He also introduced the term “vehicle” for those entities produced by replicators that help these replicators increase in numbers by interacting effectively with their environments. This distinction can be expressed in terms of either entities or processes. According to Dawkins replicators function in replication, while vehicles function in environmental interaction.

A long-standing dispute in evolutionary biology concerns the levels at which selection can occur (Brandon 1996, Keller 1999). As it turns out, there is no one process termed “selection.” Instead there are two -- “replication” and “environmental interaction”. Some authors see this dispute as concerning the levels at which replication can take place. Dawkins argues that replication in *biological* evolution occurs exclusively at the level of the genetic material. Hence, he is a gene replicationist. Other authors take the levels of selection dispute to concern environmental interaction and insist that environmental interaction can take place at a variety of levels from single genes, cells and organisms to colonies, demes and possibly entire species. Organisms certainly interact with their environments in ways that bias the transmission of their genes, but then so do entities that are both less inclusive than entire organisms (e.g., sperm cells) and more inclusive (e.g., beehives).

The term “replication” refers first and foremost to like producing like. The nature of this process of replication and the objections to it are the main topics of this entry. Genes are self-replicating molecules. Some critics interpret this claim to be asserting that a strand of DNA placed on a glass slide might all on

its own start replicating. Of course, no one has ever held such a nonsense view. Genes replicate themselves but only with the aid of highly complicated molecular machinery. Too often, however, the importance of this machinery goes unnoticed. To be sure, when we make copies on a Xerox machine, we are interested in the texts, figures or just scrawls that appear on these sheets of paper. We are not interested in how the Xerox machine works, even if it does *all* the work.

Another issue arises when replication is analyzed in terms of information. Numerous mechanisms exist for passing on information from one replicator to another. Sometimes the material in which the information is encoded gets passed on; sometimes not. For example, in the commonest sort of transcription in biology, replication is accompanied by the transmission of a substantial amount of physical material. However, in other sorts of information transfer, little or no matter is passed on. In this case, replication via DNA is aberrant, not typical. Other critics object to any use of terms associated with human languages to characterize biological replication. Talk about the genetic code is at best heuristic, at worst totally inappropriate. However, the major problem with treating replication in biological contexts is that we do not have an analysis of "information" and related terms up to the job.

As mentioned earlier, genes have two functions. One is replication. The other is the production of organisms and their phenotypic traits. The connections between genes and organisms (development) are notoriously controversial. The vast majority of literature on the topic of replication, environmental interaction and selection concerns the complexities involved in describing development. Does it make sense to refer to the "gene for eye color" or the "gene for homosexuality"? However, the topic of this contribution is replication, not development. In replication, like produces like, genes producing genes and, more generally, replicators producing replicators. Making sense of this topic is intimidating enough without introducing the relation between replication and environmental interaction. Unfortunately, replication cannot be explicated adequately without at least touching on development and its constituent difficulties. For example, if replication were simply the passing on of structure largely intact without any subsequent translation, we would not be tempted to employ terms such as "information" in connection with it.

---

1. The Conceptual Grip of Genes and Organisms

Those of us who are attempting to provide a more general analysis of “selection” try to free ourselves from the conventional way of viewing this process. For example, according to the Central Dogma of molecular biology, information can be transferred from nucleic acid to nucleic acid or from nucleic acid to proteins. It cannot be transferred from protein to protein or from proteins to nucleic acid. This asymmetry results from nucleic acids exhibiting what might be termed “potential information” while proteins do not. Given a particular stretch of DNA, it can give rise to a variety of proteins, depending on various environment factors. In a particular interaction, a particular protein will be produced. This protein will be only one of the numerous proteins possible given that stretch of DNA. In the translation of potential information into an actual protein, lots of information is lost. Hence, it is no longer available to be read back into the DNA that produced it.

But what if information flow occurs without the participation of genes? For example, from season to season a butterfly feeds on and lays her eggs on the same sort of nasty tasting host plant. In the process she transfers the nasty taste of the host plant to herself and to her progeny. Then her progeny does the same. This looks very much like information being transferred but not via the genes. Granted, genes play several important roles in this process but not in the transference of information. As another example, what if proteins served as templates for other proteins? We would have copying but nothing that might be termed a “genetic code” (Godfrey Smith 2000). The point of the preceding examples is to indicate how closely we are tied to traditional ways of viewing replication and translation. For example, just about everyone assumes that development is the only relation between replication and environmental interaction. Development is one possible relation between replicators and environmental interaction but not the only one. Freeing ourselves from the perspective of genes, organisms and development is easier said than done. It is too easy to slide back into “old think.”

In this article fundamentally different perspectives are introduced and compared - so fundamental that they deserve to be termed “metaphysical.” Although the weight of evidence can be brought to bear on these alternative formulations, it is rarely decisive. Rather the deciding factor is how well these different perspectives organize this data. What impact on our understanding of biological phenomena does switching to the Gene’s Eye view produce? Does extending the phenotype hurt or help? How about extending the notion of replication? Do these more general concepts facilitate the formation of a general analysis of replication that applies at the very least to gene-based biological evolution but also to selection in other contexts? Any answer to these questions entails a strange mix of science and philosophy.

I begin with Dawkins’ analysis because historically his writings gave rise to the present-day controversy over replication and replicators. The first issue concerns disagreements over the proper definition of “replicator.” Dawkins argues that genes and only genes function as replicators in biology. He also introduces vehicles as the entities that aid genes in their quest to become more prevalent, but gradually he reduces the role of vehicles until it is all but nonexistent. I then to proceed to evaluate the various criticisms of genes as replicators and Dawkins’ extended phenotype. Can various elements of the immune

system function as replicators? How about conceptual change? Can it be profitably viewed as a selection process? The idea of *copying* is essential to this discussion. I then turn to Developmental Systems Theory and the attempt to extend the replicator. Genes are replicators, but possibly they are not the only replicators. Finally, I discuss all too briefly the role of *information* in replication.

2. Dawkins on Replicators and Vehicles

For Dawkins selection results from an interplay between replicators and vehicles. Replicators produce copies of themselves, but they also produce entities that interact with their environments in such a way that replication is differential. Some replicators are passed on more profusely than others. For Dawkins, genes are the sole replicators in biological evolution. They are important in a second sense as well. As fascinating as all the complex adaptations that have arisen through selection may be, the results of this process matter in selection only if they are reflected in the content of their respective replicators. Dawkins is frequently (and inaccurately) castigated for being a genetic determinist, as if he thought that genes are sufficient to produce phenotypic traits, but he is aware of the crucial role that an organism's environment, including other organisms, plays in selection. A gene all by itself never did anything.

In Dawkins' early writings, replicators and vehicles played different but complementary and equally important roles in selection, but as Dawkins honed his view of the evolutionary process, vehicles became less and less fundamental. Initially, Dawkins was content to dethrone the organism from its pride of place in biology. It is an important focus of environmental interaction, but other entities, both below and above the organismic level, can also function as vehicles. In later writings Dawkins goes even further and argues that phenotypic traits are what really matter in selection and that they can be treated independently of their being organized into vehicles. More than that, features such as spider webs should be viewed as part of a spider's phenotype. Hence, Dawkins chose as the title of his second book *The Extended Phenotype* (1982a).

From the start Dawkins acknowledged that entities at a variety of levels can function as vehicles. The issue for Dawkins is replication, and according to Dawkins, genes and only genes can function as replicators in biological evolution. How big these genetic units are depends on such things as the prevalence of sex, the frequency of crossover and the intensity of selection. "If there were sex but no crossing-over, each chromosome would be a replicator, and we should speak of adaptations as being for the good of the chromosome. If there is no sex we can, by the same token, treat the entire genome of an asexual organism as a replicator. But the organism itself is not a replicator" (Dawkins 1982a:95).

Dawkins presents two reasons for organisms not being able to function as replicators. The first is the one he uses to delineate evolutionary genes. As in the case of long stretches of DNA, organisms are too easily and frequently broken up to be considered units of replication. A second reason why whole organisms cannot function as replicators is that they cannot pass on changes in their structure. Phenotypic change may result in phenotypic change (as in bad tasting butterflies), but these changes do not find their way to the genetic material. Like it or not, Lamarckian inheritance does not occur. The amount of DNA that counts as a replicator certainly varies, but according to Dawkins, nothing more inclusive than the genetic

material functions as replicators in biological evolution.

Dawkins never lost his fascination with vehicular adaptations, a fascination that his critics denigrate as Panglossian adaptationism. He fills his books with adaptationist scenarios, some more firmly supported by data than others, but from the perspective of the structure of evolutionary theory, replicators are much more important than vehicles. For example, Dawkins argues at some length that adaptations are always for the good of replicators, not vehicles. Vehicles *exhibit* these adaptations, but ultimately all adaptations must be *explicable* in terms of changes in gene frequencies. Thus, it comes as no surprise when Dawkins (1994:617) proclaims that he "coined the term 'vehicle' not to praise it but to bury it." As prevalent as organisms might be, as determinate as the causal roles that they play in selection are, reference to them can and must be omitted from any perspicuous characterization of selection in the evolutionary process. Dawkins is far from a genetic *determinist*, but he is certainly a genetic *reductionist*. Whether reductionism itself is good or bad is a moot question (Van Regenmortel and Hull 2002).

3. Replicators and Genes

Although Dawkins finds it desirable to replace genes with replicators in his general characterization of the evolutionary process, he says very little about this more general notion in his *The Selfish Gene*. Instead he discusses the special case of genes as replicators. The primacy of the genetic perspective in the characterization of replicators is one of the chief weaknesses not only in Dawkins' discussions but also in the work of his successors -- including this one. We claim to be providing a general notion of replication, adequate to handle all different sorts of replication, when too often we are simply reading peculiarities of genetic replication into our general analysis of replication.

According to Dawkins, replicators have three basic properties - longevity, fecundity and copying-fidelity. Longevity means longevity in the form of copies. No gene as a physical body lasts all that long. In mitosis, a gene loses half of its substance at each replication. What endures is not the entity itself but the information incorporated in its structure. It is this information that is copied with such high fidelity. Mutations do occur but at very low frequencies. Even so, in some organisms mutation rates must be too high because mechanisms have evolved that search out and repair such errors.

The main source of variation in genes, however, is not mutation but crossover. Pairs of homologous chromosomes line up next to each other at meiosis, crossover and recombine. For stretches of DNA in which different alleles exist, the result can be a change in information. Quite obviously, the shorter the stretch of DNA involved, the less likely that crossover will occur and the message changed. Dawkins appeals to such dismantling of entities to argue against organisms functioning as replicators. In sexual organisms, organisms themselves are torn apart and reassembled each generation. If long stretches of DNA lack the necessary identity by descent to function as replicators, then sexual organisms certainly lack it. However, some other explanation has to be provided for asexual organisms because they pass on their overall structure largely unchanged from generation to generation.

Dawkins bypasses classical Mendelian genes and even molecular genes to adopt G. C. Williams'

conception of evolutionary genes as the fundamental units of natural selection. Dawkins (1976:30) defines evolutionary genes as "any portion of chromosomal material which potentially lasts for enough generations to serve as a unit of natural selection." The limits of these genes need not be absolutely sharp. Nor must all genes be of the same length. The greater the selection pressure, the smaller the gene. At bottom selection takes place between alternative alleles residing at the same locus. "As far as a gene is concerned, its alleles are its deadly rivals, but other genes are just a part of its environment, comparable to temperature, food, and predators, or companions" (Dawkins 1976:40). Alleles cannot cooperate with each other; only compete. That is where "selfish" in the "selfish gene" comes in.

According to Dawkins (1976:95), the selfish gene is not just one physical bit of DNA. It is "all replicas of a particular bit of DNA, distributed throughout the world." Hence, genes do not form classes of spatiotemporally unrelated individuals but trees. They must be replicas. But being a replica is not enough. The linear repetition of the "same gene" in the form of several hundred copies is quite common. These replicas, however, do not reside at the same locus. As identical in structure as these genes may be, they do not compete with each other in the way that alleles at the same locus can. In the simplest and most basic sense, alleles compete with alternative alleles at the same locus. Any other sorts of competition and cooperation are merely extrapolations from this fundamental sense of allelic competition. Even though genes may cooperate with each other in very complicated ways in embryological development, in replication they can be treated as "separate and distinct." In development the effects of genes blend. In replication replicators do not blend.

Dawkins introduced the general notions of replicator and vehicle so that selection need not be limited exclusively to gene-based biological evolution. However, as the preceding discussion indicates, his later revisions to his general theoretical outlook were influenced strongly by the traditional perspective of genes and organisms. Genes contain the information necessary to produce organisms and their adaptations. Genes "ride around" in and "guide" organisms. As Dawkins describes them, vehicles are relatively discrete entities that "house" replicators and which can be regarded as machines programmed to preserve and propagate the replicators that ride inside them. Although these terms may be appropriate for the relations between genes and organisms, they interfere with a more general analysis of replication and selection. What really matters in selection is that entities at various levels of organization interact with their environments in such a way that the relevant replicators increase in relative frequency. The actual causal chain that connects replicators and vehicles need not be limited to development.

For example, Dawkins argues at some length that genes and only genes can function as replicators in biological evolution. He goes on to add that "all adaptations are for the preservation of DNA; DNA itself just is" (Dawkins 1982b:45). But DNA itself exhibits adaptations. Anyone who has spent much time examining the molecular structure of DNA soon realizes that it is adapted to replicate. In addition, the proliferation of junk DNA and meiotic drive are two examples in which the only phenotypes that matter are phenotypic characteristics of genes (Brandon 1996:133; Sterelny, Smith and Dickison 1996:388). Dawkins' characterizations of replicators, vehicles and the relations between the two are too closely tied to genes, organisms and development. DNA can certainly replicate itself, but it can also function as a "vehicle" even though it cannot code for, ride around in or direct itself. In sum, a more general characterization of selection is needed, a characterization that does not assume that the only causal

connection between replicators and vehicles is development.

4. Memes and the Immune System

Initially Dawkins" discusses only one example of replication that goes beyond gene-based biological evolution - memetic change. Later he adds the immune system. Does replication play the same role in other sorts of selection that it plays in gene-based biological evolution? The reaction of the immune system to antigens is closest to the standard biological example of replication and selection. Genes are the primary replicators, but these genes are somatic, not germ-line. Certainly immune systems arose through the same process as other functional systems. The basic structure of the human immune system is "built into our genes," and the only genes that count in evolution are the one that we received from our ancestors and can subsequently pass on to our progeny. In Dawkins' terminology, the genes that code for our immune system are active, germ-line replicators.

However, the genes that function in the reaction of the immune system to antigens have two peculiarities. First, they incorporate mechanisms designed to produce very high frequencies of mutation, and second, none of the genes involved in the functioning of the immune system are germ-line. The genes that give rise to B-cells for instance are designed to mutate extensively until one of these cells identifies an invader as foreign. It then proliferates extensively as it attacks the invader. As an organism matures, it accumulates more and more of the B-cells that have been successful in its past. More than this, as the process of proliferation continues, the strength of the affinities to binding sites increases. Initially the primary antibodies almost always exhibit a weak affinity for their targets, but as the reaction to the antigen continues, these affinities become stronger.

Within the confines of a single organism, the reaction of the immune system to antigens has all the characteristics of selection processes, but when the organism dies all of these adaptations are lost. In some species, females pass on not only the genes for the basic structure of their immune systems but also some of the machinery that past invasions of antigens have produced in her. However, these cells are rapidly removed from the offspring as it develops its own immune system. The reaction of the immune system to antigens departs from gene-based selection in biological evolution in two ways. First, from the organismal perspective, the genes that function in protecting an organism from invaders are not germ-line. They are somatic. Second, the relevant mutation rates are much, much higher in the immune system than in ordinary gene-based selection. Instead of mechanisms existing to discover mutations and repair them, mechanisms exist that encourage mutations - massively so. If the functioning of the immune system is to count as selection, then some changes must be made. One possibility is to clarify what counts as a "germ-line" replicator and a rejection of the notion that extremely low mutation rates are inherent in all selection processes.

The modifications needed to include both gene-based selection in ordinary biological evolution and the reaction to the immune system are relatively "minor." Such is not the case with Dawkins' second example - memetic evolution. Several apparent differences exist between it and the previous examples of replication and selection. Numerous critics have developed a cottage industry engaged in listing all the

differences that exist between biological evolution, on the one hand, and social-cultural-conceptual evolution on the other (henceforth, SCC). For example, Dawkins (1976) admits that he is on shaky ground when it comes to the high copying fidelity required of replicators. Memes seem to get changed much more frequently than genes. However, if differences in copying fidelity are enough to preclude memetic evolution from being a genuine case of selection, then the immune system must also be rejected. Mutation rates are much higher in the production of B-calls than in any area of SCC no matter how speed is calculated.

In general, the standards used to evaluate memetic evolution are much higher than those used to evaluate any other sort of selection. Time and again an overly idealized view of Mendelian genetics is contrasted to a much more realistic view of SCC change. So the story goes, one problem with memes is that they do not have discrete boundaries, do not all come in the same size, and in their functioning are strongly influenced by their environments. Genes, so the critics claim, have sharp boundaries, are all of the same size, and are immune to environmental influences. If memetic evolution is to be evaluated fairly, the same level of criticism must be applied to all putative examples of selection from gene-based selection in biological evolution and the reaction of the immune system to antigens to the development of the central nervous system and social learning (Hull, Langman and Glenn 2001).

Dawkins (1976) also places considerable emphasis on human brains as the “vehicles” for memetic evolution. He defines “meme” as an entity capable of being transmitted from one brain to another. Computers are also plausible vehicles for memes. Dawkins- discussion of memes is, once again, marred by the pervasiveness of the gene-organism perspective. For example, he defines “replicator” in terms of transmission of information - memes leaping from brain to brain or from brains to computers and back again. But memes do not leap from brain to brain or from computer to computer. Their content is transmitted in a variety of ways, including books, audiotapes, conversations and the like. As much as the physical basis changes, the message remains sufficiently unchanged. All instances of this message are equally memes, not just the ones residing in human brains and computers.

All the objections to the gene-meme analogy to one side, Dawkins (1976:211) finds the chief difference between genetic and memetic change is that biological evolution is at bottom a war between alleles residing at the same locus. “Memes seem to have nothing equivalent to chromosomes and nothing equivalent to alleles.” First, the usual depiction of alleles residing at the same locus on homologous chromosomes so central to Mendelian genetics is an over simplification, but more importantly, for at least half of life on earth, replication and selection took place in the absence of chromosomes, meiosis and the like. If gene-based biological evolution took place for so long in the absence of the Mendelian apparatus and still does so in many extant organisms, then just possibly we should not demand that memetic evolution proceed by this very special and possibly aberrant sort of inheritance. The cost of meiosis remains a serious problem in ordinary biological evolution. Demanding that SCC evolution incorporate this same highly problematic element in its own makeup seems strange in the extreme. If we are to develop a general analysis of selection, then we must distinguish between essential and contingent features of this process.

Numerous evolutionary biologists question how fundamental to selection the perspective of alleles at a

locus actually is. Everyone certainly says that evolution is changes in gene frequencies. However, many do not go on to add that evolution is *nothing but* changes in gene frequencies. When one looks at the work of evolutionary biologists, one discovers that it involves much more than changes in gene frequencies. Selection in meme-based evolution must be fleshed out. It remains to be seen how different well-worked-out versions of SCC theory will differ from more familiar forms of selection. And where they differ, it does not follow automatically that the meme-based theory must be modified. One might well change the traditional gene selectionist view of biological evolution. Selection in gene-based theories of evolution was worked out first, but historical precedence does not entail conceptual priority.

5. The Extended Replicator

The discussion thus far has involved criticisms of Dawkins' notion of genes and replication. Dawkins' critics think that genes play too important of a role in his notion of replication. Replication, so they argue, can occur at other levels of organization as well. Just as Dawkins extended the notion of the phenotype, these authors propose to make the notion of replicators more general as well - to extend the replicator so to speak. Dawkins introduced his *Extended Phenotype* conception for two reasons. First, he wanted to extend the notion of a "phenotypic trait." The sort of nest that a bird builds or the song that it sings can count as phenotypic traits just as much as the shape of its bill. Second, Dawkins wanted to break the hold that organisms have over how we conceptualize the living world. Traits do tend to come bundled into reasonably discrete entities, but for making inferences about the evolutionary process, traits can be treated as separate and distinct. However, as he proceeded, Dawkins eventually decided that extending the phenotype was not enough. He had to bury it. All sorts of fascinating mechanisms to one side, what really matters in selection takes place at the level of replicators.

Although such critics of Dawkins as Sterelny *et al*. (1996) decline to join with Dawkins in his demotion of environmental interaction, they do agree with him that replicators are special. They play a special role in development. However, they do not limit replicators to genes even in biological evolution. Sterelny *et al*. (1996) propose to extend the replicator to include nonstandard entities. For example, the sort of burrow that a particular organism digs is influenced by its genetic makeup, but if these burrows are used over and over again, characteristics of these burrows can themselves be viewed as replicators. The effects of these burrows get passed on from generation to generation but not via the genes.

As more and more nongenetic replicators are acknowledged, Dawkins' Gene's Eye view begins to gradate into the conception of the Extended Replicator. These two views are so general that any case that can be described in one can be redescribed in the other. Differences lie in ease of description. According to Sterelny et al., the burrows that some organisms dig can function as replicators - extended replicators - while Dawkins portrays them as instance of extended phenotypes. The contrast is between selfish burrows and selfish genes for burrowing.

Starting with Oyama's *The Ontogeny of Information* (1985), a view of biological evolution has arisen that emphasizes development (see also Griffiths and Gray 1994, Oyama 2000). In the 19th century development was an extremely active research program. The next great discoveries in biology were going

to be in the area of development. Such was not to be. First, evolutionary biology and then genetics took over biology, and they did so while avoiding development. Everyone knew that development was central to both evolution and reproduction, but no one could see how to integrate the masses of developmental data available into the emerging synthesis of evolutionary biology and genetics. As a result, development was left out of the New Synthesis. Considering how central development actually is in biology, the advances made while ignoring it are staggering.

Even so, developmental biology continued on its course until at long last we seem to understand development well enough to begin integrating it into the rest of biology. On the most conservative view, current versions of evolutionary theory can remain largely unchanged as development is grafted onto them. On a second view, both perspectives are likely to require some modification to bring off this integration. Our understanding of development may have to be modified, but so too for evolutionary theory. Finally, at the other extreme, development will all but replace evolutionary theory. In their more exuberant moments, advocates of Developmental Systems Theory (hereafter DST) seem to be claiming just that. Just as some molecular biologists think that molecular biology is rapidly replacing all the rest of biology, advocates of DST argue that developmental theories will simply replace current versions of evolutionary theory.

On the DST view, genes have no privileged role in repeated cycles of development. In fact, no element of the developmental matrix plays any privileged causal role - not genes, not organisms, not the environment, not anything. Everything counts equally as a resource, albeit in particular situations certain resources will play more important roles than other resources. In rejecting any privileged role for genes, advocates of DST are especially skeptical of one particular role supposedly played by genes - the transmission of information. According to some, information is central to developmentalism, but genes are not the only mechanisms for information transfer. According to others, information plays no role in the emerging developmentalist perspective. The developmental system as a whole is the unit of selection - the replicator.

In the continuing debate over DST versus traditional theories of evolutionary biology, Sterelny *et al*. (1996) hold a fairly conservative position. They agree with the developmentalists that genes play no privileged role in the development of phenotypes from genotypes. Genes play a role in this process, simply not a *privileged* role. Genes can serve as a causal bridge from phenotype to phenotype, but other entities can do so as well. Genes are not the only replicators in biological evolution. The repeated cycles in inheritance include many different sorts of constancies and repetitions - genes, cellular machinery, phenotypic traits including behaviors, and social structures. Information remains central to selection processes, but genes are not the only carriers of such information. Genes predict phenotypic characters only in the same sense that environmental factors predict them.

Both Dawkins (1976) and Hull (1980) distinguish between replicators and vehicles (or interactors). To the questions surrounding these two sorts of entities, Lloyd (1992) adds two more - who “owns” adaptations and who “benefits” from selection? In the previous literature, authors seem to assume a greater concordance in the answers to these four questions than can be justified. Some biologists argue that those entities that exhibit a particular adaptation are the beneficiaries of the selection process that

resulted in these adaptations. Dawkins disagrees. Such adaptations function only as intermediaries in a much more fundamental process - the spreading of copies of particular alleles. These surviving alleles are the long-term beneficiaries of selection processes. Sterelny *et al*. argue that replicators and only replicators are the beneficiaries of adaptations, keeping in mind that they count entities more inclusive than genes as replicators.

With respect to the version of replication formulated by Sterelny *et al*. (1996), both copying and biofunction are crucial. Copying is quite obviously a causal phenomenon, but not any old causal connection will do. Similarity of copies is necessary but not sufficient. These copies must be copies of copies. One copy must produce another copy and that copy produce still another copy and so on. For one entity to be a copy of another, it must be the output of a process whose biofunction is to conserve function. The function of copying is to produce from one token another token which is similar in the relevant respects. Genes fit this definition, but so do lots of examples of nongenetic transmission; e.g., habitat stability resulting from nest site imprinting, the song that a bird learns, various micro-organism symbionts, not to mention SCC transmission.

6. Information

The language of “codes” and “information” flows easily enough with respect to replication. Transcription, translation, punctuation, redundancy, synonymy, editing, proofreading, errors, repairing of errors, messages, copies, and information all sound natural enough. This ease of expression supports the contention that selection plays the same role in SCC evolution as it does in gene-based biological evolution. Some authors argue that both processes count as languages in the same sense of “language.” All the problems listed for treating gene-based biological evolution as a language are simply replicated in ordinary languages. If the genetic code is not a code, then neither is the Morse code or English for that matter. Differences of degree exist but not differences in kind. For example, mutation rates are very low in gene-based biological evolution, moderate in SCC evolution and massive with respect to the immune system. Concurrent variation is extensive in the immune system, greatly reduced in both gene-based biological evolution and SCC evolution, and reduced even further in operant learning.

The literature dealing with information is both extensive and factious. Several different formal analyses of information can be found and very little agreement about which analysis is best for which subjects. On one point these scholars tend to agree - cybernetic information and communication-theoretic information will not do for replication in biological contexts. The best bet is semantic information [Sterelny (2000), Godfrey-Smith (2000) and Sarkar (2000)]. The trouble is that no widely accepted version of semantic information exists. Winnie (2000) distinguishes between Classical and Algorithmic Information Theory and opts for a revised version of the Algorithmic Theory. But once again, the problem is that no such formal analysis currently exists. In the face of all this disagreement and unfinished business, biologists such as Maynard Smith (2000) maintain either that informal analyses of “information” are good enough or that some future formal version of information theory will justify the sorts of inferences that they make. For now, the most likely conclusion is that no version of information theory as currently formulated can handle “information” as it functions in biology (see Griffiths 2001 for further discussion).

7. Conclusion

In general the authors who have concerned themselves with the role of replication in selection strive to liberate our thinking from the hold that genes and organisms have over us. To the extent that replication can be separated off from development, it plays a central role in selection processes of all sorts, including traditional gene-based biological evolution, the reaction of the immune system to antigens and possibly even SCC evolution. In reaction to all sorts of nonstandard examples of replication, Dawkins is driven to extend the phenotype. In reaction to DST some critics of Dawkins extend the replicator as well. Replicators include genes, memes and even entities that are commonly thought of as parts of an organism's phenotype or environment. Some authors argue that replication in this extended sense is necessary for selection; others that it is not. Selection can occur in the absence of replication. Still others maintain that all of these biological phenomena are best explained without any reference to such traditional notions as replication, environmental interaction, selection, evolution, etc., in terms of DST.

Bibliography

- Brandon, R. R. 1996, *Concepts and Methods in Evolutionary Biology*, Cambridge: Cambridge University Press.
- Cook, P. R. 1999, The Organization of Replication and Transcription, *Science* 284: 1790-1797.
- Dawkins, R. 1976, *The Selfish Gene*, Oxford: Oxford University Press.
- Dawkins, R. 1982a. *The Extended Phenotype*, Oxford: Oxford University Press.
- Dawkins, R. 1982b. Replicators and Vehicles, In *Current Problems in Sociobiology* (ed. King's College Sociobiology Group). Cambridge: Cambridge University Press, pp. 45-64.
- Dawkins, R. 1994, Burying the Vehicle, *Behavioral and Brain Sciences* 17:616-617.
- Godfrey Smith, P. 2000, Information, Arbitrariness, and Information: Comments on Maynard Smith, *Philosophy of Science* 67: 202-207.
- Griffiths, P. 2001, Genetic Information: A Metaphor in Search of a Theory, *Philosophy of Science* 68:394-412.
- Griffiths, P., and R. Gray, 1994, Developmental Systems and Evolutionary Explanation, *Journal of Philosophy* 91:277-304.
- Hull, D. L., 1980, Individuality and Selection, *Annual Review of Ecology and Systematics* 11:311-332.
- Hull, D. L., R. E. Langman and S. S. Glenn, 2001, A General Account of Selection: Biology, Immunology, and Behavior, *Behavioral and Brain Sciences* 24:511-573.
- Keller. L. (ed.). 1999, *Levels of Selection*, Princeton: Princeton University Press.
- Lloyd, E. A. 1992. Unit of Selection, *Keywords in Evolutionary Biology* (ed. by E. Fox Keller and E. A. Lloyd), Cambridge, Mass.: Harvard University Press, pp. 334-340.
- Maynard Smith, J. 2000, The Concept of Information in Biology, *Philosophy of Science* 67: 177-194.
- Oyama, S., 1985, *The Ontogeny of Information: Developmental Systems and Evolution*, Cambridge: Cambridge University Press.

- Oyama, S., 2000, *The Ontogeny of Information: Developmental Systems and Evolution*, Second Edition, revised and expanded with an Introduction by Richard C. Lewontin, Durham, North Carolina: Duke University Press.
- Rebek, Jr., J. 1994, Synthetic Self-Replicating Molecules, *Scientific American* 271: 48-55.
- Sarkar, S. 2000, Information in Genetics and Developmental Biology: Comments on Maynard Smith, *Philosophy of Science* 67: 208-213.
- Sterelny, K. 2000, The “Genetic Program”: A commentary on Maynard Smith on Information in Biology, *Philosophy of Science* 67: 195-201.
- Sterelny, K., K. C. Smith and M. Dickison, 1996, The Extended Replicator, *Biology & Philosophy* 11:377-403.
- Van Regenmortel, M., and D. L. Hull, 2002, *Promises and Limits of Reductionism in the Biomedical Sciences*, Chichester: John Wiley and Sons.
- Winnie, J. A., 2000, Information and Structure in Molecular Biology: Comments on Maynard Smith, *Philosophy of Science* 67:517-526.

Other Internet Resources

- The Artificial Self-Replication page, by Dr. Moshe Sipper, Logic Systems Lab, Swiss Federal Institue of Technology)
  [The concept of replication has led a largely independent life in the formal sciences as a consequence of John von Neumann's seminal work on self-replicating automata. This page offers a good introduction to the literature.]

Related Entries

developmental biology | evolution: cultural | genetics | genetics: gene | heritability | information: biological | life | molecular biology | natural selection

David Hull
Northwestern University
d-hull@northwestern.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 4, 2001

Content last modified: December 4, 2001

Mental Illness

Psychiatry involves theories of the mind, theories of the causes of mental disorders, classification schemes for those disorders, research about the disorders, proven treatments and research into new treatments, and a number of professions whose job it is to work with or on behalf of people with mental disorders. The philosophical study of psychiatry discusses conceptual, ethical, metaphysical, social, and epistemological issues that arise in all these aspects of psychiatry Central to this study is the nature of mental illness.

The connection between philosophical issues in the study and treatment of mental illness and these other areas of philosophy is in many cases obvious. For example, it takes little thought to see how the question of when and how people with mental disorders are responsible for their actions is connected with the insanity defense in law, and the more general debate over the justification of punishment. Similarly, it is clear how studying the historical growth of the idea of madness and changes in the way societies treat those they classify as mad helps us assess claims that psychiatry today is a form of social control, and further, whether social control is a legitimate function for psychiatry.

The philosophical investigation of the nature of mental illness is therefore relevant to many other areas of philosophy. While there is no sharp divide between the philosophical discussion of the nature of mental illness and the wider philosophical discussion of psychiatry, we can focus on three major issues that have preoccupied the philosophical literature.

1. Does Mental Illness Exist

The English-speaking world has not always used medical language to describe the behavior we now label

as symptomatic of mental illness or mental disorder. Descriptions were sometimes framed in quite different terms, such as possession. What we now call mental illness was not always treated as a medical problem. Non-English-speaking nations in the West have had changes in their linguistic usage and their treatment of the mentally ill roughly parallel to Anglophone countries. Anthropological work in non-Western cultures suggests that there are many cases of behavior that psychiatry would classify as symptomatic of mental disorder, which are not seen within their own cultures as signs of mental illness. Indeed, other cultures may not even have a concept of mental illness that corresponds even approximately to the Western concept.

The mainstream view in the West is that the changes in our description and treatment of mental illness are a result of our increasing knowledge and greater conceptual sophistication. On this view, we have conquered our former ignorance and now know that mental illness exists, even though there is a great deal of further research to be done on the causes and treatment of mental illness. However, there are some thinkers who have challenged this mainstream view. Some have argued for a relativist view, that the reality of mental illness is not an absolute transcultural fact. The relativist view would have to be that statements about the existence of mental illness can be true in some cultures and false in other cultures.

A more extreme view is that there is no such thing as mental illness in any culture, and that there could not be, because the very notion of mental illness is based on a fundamental mistake or set of mistakes. This sort of view is most closely associated with the psychiatrist Thomas Szasz. It is this view that we will focus on here.

There are many arguments in the voluminous work of Szasz, and it is not easy to always keep them separated. One critic of Szasz separates out at least six main arguments against the existence of mental illness within his work. (Reznek, 1991, Chapter 5).

Sometimes he has compared psychiatry to alchemy or astrology (1974, pp. 1-2), and says they are all pseudo-sciences. On this criticism, it seems that the reason that mental illness does not exist is the same sort of reason that phlogiston or astral influences do not exist: it is an empirical mistake caused by flawed methodology. The continued belief in mental illness by psychiatrists is the result of dogmatism and a pseudoscientific approach using *ad hoc* defenses of their main claims. He also accuses psychiatrists of secrecy and obfuscation.

However, it seems that his most fundamental criticism is not of the scientific methodology of psychiatry, but of its concepts. His claim is that the concept of mental illness is based on a confusion.

> [The belief in mental illness] rests on a serious, albeit simple, error: it rests on mistaking or confusing what is real with what is imitation; literal meaning with metaphorical meaning; medicine with morals. (Ibid, p. x.)

Ssasz says that there cannot be mental illness, literally speaking, because it is no more than a metaphor. He argues that by definition, "disease means bodily disease." (Ibid, p. 74), and further, the mind is not

literally part of the body.

Szasz's critique of the foundations of psychiatry has attracted a great deal of attention from supporters and detractors. It has generated debate over the following issues:

1. Is it true that disease, by definition, must refer to bodily disease?
2. Is it true that the mind is not literally part of the body? Couldn't the mind be identified with the brain or the neural system?
3. Is it true that that medicine is, intrinsically, not about the moral evaluation of behavior, even if it might be used instrumentally as part of a moral evaluation?
4. Is it true that psychiatry is founded on pseudoscience?

Szasz's position has not gained any widespread credence. Given the progress of neuroscience and our increasing ability to affect emotions, thought, and behavior through medication, psychiatry has if anything gained in scientific credibility in the thirty years since Szasz first proposed his critique.

Few of Szasz's supporters have been willing to take as extreme position as him. Critics of psychiatry have been more focused on particular purported mental illnesses, such as alcoholism, psychopathy, multiple personality disorder, rather than the whole category of mental illness. There is still vigorous debate over the reality of such mental disorders. This debate has turned more on empirical issues than the more philosophical claims of Szasz.

2. Is There an Objective Way to Classify Mental Illnesses?

While Szasz's approach is of more historical importance than of current relevance to philosophy of psychiatry, there is vigorous ongoing debate concerning the way that mental illnesses should be classified. There are two aspects to this: what conditions get classified as mental illneses rather than normal conditions, and, among those conditions we agree are mental illnesses, how they are grouped together into different kinds. Controversial diagnostic categories have included homosexuality, psychopathy and personality disorders, attention deficit hyperactivity disorder, and dysthymia. The unitary nature of schizophrenia has come under special scrutiny, as has the unitary nature of autism. This debate spans both empirical and philosophical issues, and it is the former aspect, and the distinction between normality and psychopathology, that has gained the most philosophical scrutiny. The primary questions of concern are:

1. Will it be possible in the future to classify mental illnesses according to their causes, as we do in much of the rest of medicine?
2. Given that we currently classify most mental illnesses according to their symptoms rather than their causes, is there any reason to think that our current diagnostic categories (e.g., schizophrenia, depresssion, manic depression, anxiety disorders) correspond with natural kinds?

3. Is it possible for our current classification scheme in psychiatry to be in any important sense "atheoretical" and independent of any particular theories of the etiology of mental disorders?
4. Is it possible for any classification scheme of mental illnesses to be purely scientific, and is it possible for a classification scheme to be independent of values -- or to ask the reverse, do our classification schemes in psychiatry always rest on some non-scientific conception of what should count as a normal life?

This last question can be extended to all illnesses, and not just psychiatric classification. It is in psychiatry, though, that there is most reason to worry that values enter into the classification scheme, and that there is concern that the profession might be medicalizing what should be seen as normal conditions.

Before we discuss the main approaches here, we should note a couple of points. First, the concepts of disease, illness, abnormality, malady, disorder and malfunction are closely related, but they are not the same. Much careful work has been done trying to find if one of these is more basic than any of the others, or if some of these concepts can be completely analyzed in terms of the others. For our purposes here, we shall gloss over the differences between these concepts. For the most part, we will simply refer to the concept of illness.

Second, even if we could find an uncontroversial general criterion of illness, we would still need to do some work to find a criterion of mental illness. This might seem a simple task, since mental illnesses would simply be those that are illnesses of the mind. But often neurological disorders such as Alzheimer's are not classified with mental illnesses, and there is even resistance from some to classifying them in wider categories such as mental disorder. That seems to be due to the belief that neurological disorders are physical disorders resulting from identifiable damage to the neural system. As we become increasingly able to identify the brain dysfunctions associated with mental disorders, it may well be that the distinction between neurological disorders and mental illnesses starts to fade, as might the professional distinction between neurologists and psychiatrists. It may turn out that a defense of psychiatry does not need to find a clear conceptual distinction between the mental and the physical. Certainly, so far the main debate about finding a criterion of illness has not paid much attention to the problem of finding a criterion of "mental".

A main approach to psychiatric classification is the "medical model". This holds that psychiatric classification is capable of being both scientific and objective. The best-known defender of such an approach is Christopher Boorse, in a series of influential papers. At the other end of the spectrum are theories that psychiatric classification depends solely on the whim or values of those doing the classification, that there is nothing objective about it at all, and that there are no facts about what is normal. These subjective theories are generally proposed in a spirit of criticizing or undermining psychiatry, and are often very sympathetic to the Szaszian view that there is really no such thing as mental illness, and so there could not be a legitimate objective classification of different kinds of mental illness. Often the suggestion that goes with these views is that classification schemes are created to suit the needs of those in power. This view has not often been argued for explicitly, but is at least implicit in the work of Szasz, and it may be implicitly in the work of sociological theorists Peter Sedwick and Thomas Scheff. (See Reznek, 1991, Chapters 6 and 7). Some have suggested that this view underlies the

historical analysis of Michel Foucault. As for its plausibility, the view that the classification is totally subjective or arbitrary stands or falls with antirealism about mental illness, and it has not received much support in the last twenty years.

A middle range of views, sometimes called "mixed" (e.g., Wakefield 1992) hold that diagnostic categories do match real mental illnesses and that there are facts about the world that determine what should be labeled as a mental disorder, but that at the same time, there is an irreducible element of value or normativity in deciding psychiatric categories.

Most debate on classification has been between different versions of mixed models, and with some debate between mixed models and of the medical model. The medical model says that whether a particular condition is normal or abnormal is simply for science to say. It does not pretend that science has provided all the answers and it may concede that there is a great deal more research to be done. It might even concede that we will never be able to collect enough evidence to discover whether a particular condition is abnormal or not: it would conclude in such a case that we cannot know the truth.

It would be highly implausible for a defender of the medical model to insist that values never in fact enter into the psychiatric taxonomy -- a brief study of the history of various categories show that empirical research and neutral scientific facts are certainly not the only things that have been taken into consideration in classification schemes. The medical model claims (a) that it is possible to have a value-neutral classification scheme and (b) it is best to use a value-neutral classification scheme.

In justifying part (b) of their claim, some defenders of the medical model might claim we can discover a *conceptual truth* of the form:

> a disease/illness/malady/disorder/malfunction is a condition that ...

where the ellipsis is filled by some clause such as "reduces the lifespan of the organism," "reduces the productivity of the organism," or "reduces the ability of the genes of the organism to reproduce themselves." However, such an approach is highly problematic because it is very difficult to establish non-trivial conceptual truths about controversial concepts, and it seems clear that our actual usage for the last few centuries of words like disease, illness, or malady do no correspond well with such purported definitions. They are either too broad, too narrow, or both.

An alternative approach to defending (b) is to argue that medicine, and psychiatry especially, should be value-neutral and so its classification scheme should be value-neutral. Note, of course, that there are obvious ways in which we want medicine to not be neutral: for example, it should not be neutral about saving lives or improving health. So we need to be careful about the way in which we want medicine to be neutral. The most appealing interpretation of the idea that medicine, and psychiatry, should be value-neutral is that it should rise above political fighting, and that it should not be influenced by prejudice. However, it is an open question whether being value-neutral and being unbiased are the same. When it comes to deciding what conditions should count as illness, one could be unbiased but nevertheless adopt

a view laden with normative assumptions, or so at least many would argue.

The best way to explain the idea that taking a value-laden stance is not necessarily to be biased is to examine the criticisms of claim (a) above, from theorists who argue that psychiatry and the rest of medicine are inevitably normative. Most such theorists do not infer from this that medicine is always biased; instead, their view is that the nature of psychiatric classification requires that some assumptions are made about what counts as health and what counts as illness, and that these assumptions are not purely scientific. They generally go on to suggest that since medicine and psychiatry in particular have to make such assumptions, it should be as open and honest about it as possible, so that debates about certain categories of psychopathology are not based on a misunderstanding of the kind of enterprise involved. Often they suggest that in a democracy, there should be public debate about what values should be at the heart of medicine and psychiatry.

Those who argue that classification is necessarily value-laden rarely rest their argument on the claim that all science is value-laden, or even more controversially, that all science is subjective. For the sake of argument, it is possible for all sides of the debate to concede that we can know facts about the causes and consequences of the conditions we label as illnesses, and that these facts are entirely value-neutral. (There are of course some who would dispute the possibility of there being, or our knowing, any value-neutral facts, but this is an extreme view, and it does not single out medical classification as an interesting and unusual case of value-ladenness. So we will set it aside.)

We now can ask why those who think that psychiatric classification must be value-laden think so, and how those who think it can be value-neutral propose to find such a classification.

If a theory can, by itself, provide us with a way of demarcating human health from pathology, then the theory must, on its own, have some account of what healthy function is, and what should count as a malfunction of a human being. Those who believe in value-neutral classification generally argue that "health" can be defined scientifically, and thus without value-laden assumptions. Those who disagree think that the criteria used to define "health" are always value-laden, even if they are also based in scientific understanding.

Thus Boorse, who argues for the value-neutral view of classification, suggests that evolutionary theory can tell us what conditions are healthy. In one paper, he gives the following definition of health:

> An organism is *healthy* at any moment in proportion as it is not diseased; and a *disease* is a type of internal state of the organism which:
>
> - interferes with the performance of some natural function -- i.e., some species-typical contribution to survival and reproduction -- characteristic of the organism's age; and
> - is not simply in the nature of the species, i.e. is either atypical of the species or, if

typical, mainly due to environmental causes. (Boorse, 1976, page 62.)

This purported definition has received a great deal of critical discussion. Without setting out all the details of that discussion, we can at least explain the opposing view. There are several possible points to make; here are three:

(C1) In much of medicine, and especially psychiatry, we do not know with any certainty what is evolutionarily natural, because our scientific studies are still in their early stages, and it can be very difficult to find data that will settle scientific controversies. The most promising scientific theory is evolutionary psychology, but this is extremely programmatic. For many conditions, such as homosexual behavior or mild depression, it is not clear whether these conditions help or hinder the continuance of the species (or the continuance of whatever set of genes the theory says is fundamental). Therefore the idea of settling the debates of what should count as illnesses with science is at best a proposal for a distant future time. It is likely that many of the scientific questions will never receive satisfactory answers, in which case we will never be able to use science completely to determine our answers.

(C2) More fundamentally, even were we to have a complete theory of evolutionary psychology, it would still be controversial whether to use such a theory in determining whether particular conditions are normal or abnormal. That is to say, many dispute whether medicine should base its view of naturalness on conditions that help the promotion of the species. For instance, many would claim that medicine is far more individualistic these days, and that we are more concerned with what hinders a particular individual, whether or not it helps the rest of the species, or would have helped the species in times when we were developing evolutionarily. There are two points here: first, that evolutionary psychology does not provide the only possible model of health and illness, and second, that given a choice between two or more models, we have to employ values in making that choice. Thus, even if we adopt evolutionary psychology as the theory to determine what is health and what is illness, this choice is not one based on purely objective scientific criteria, but rather is normatively loaded. For example, Boorse’s evolutionary approach uses illness to mean an internal condition that interferes with a natural function that promotes the continuation of the species, but in deciding between this and an alternative approach, we have to ask whether this model of illness is the best one for us. No value-neutral scientific experiment or theory can answer that question, and we have to bring values into our considerations.

(C3) What’s more, the answers that evolutionary psychology seems to suggest on controversial cases often don’t match with our contemporary practice, and there is serious doubt that a model of health provided by evolutionary psychology is really the one that we should adopt.

Mixed models

Even if the medical model of illness is wrong, it may only require a small modification in order to become acceptable. This is what has been argued by Jerome Wakefield in a number of influential publications. Wakefield attempts to keep the concept of a natural function, and the concept of dysfunction, central in our understanding of mental disorder. Further, he proposes that evolutionary

psychology can explicate which conditions are functional and which are dysfunctional. He suggests that dysfunction is a purely factual scientific concept. So some conditions, even though they may be judged negatively, will not count as disorders because they are not dysfunctions, and are not caused by dysfunctions. For example, some have claimed that children who masturbate have "childhood masturbation disorder." Wakefield says that there is no such disorder, and whatever one's values, such behavior could never count as a disorder because it is not unnatural according to the scientific theory of evolutionary psychology.

However, he concedes that some values must enter in to our decision of what conditions are mental disorders: he argues that some dysfunctions are not judged negatively, and so do not need to be classified as disorders. For example, even if evolutionary theory could show that homosexuality was unnatural, we might not classify homosexuality as a disorder because we might decide that it is not harmful. Our society may have changed so much since the times when our natures were formed that even if a person lacks certain abilities, for example, to be a hunter, and was evolutionarily speaking unnatural, we could agree that the ability to be a hunter is no longer necessary in our society, and so lacking hunter abilities does not mean one has a disorder. Further, some deficits may make us less than perfect, but still we would not judge that we are so lacking as to have a disorder.

This leads Wakefield to the following analysis of a disorder:

> A condition is a disorder if and only if (a) the condition causes some harm or deprivation of benefit to the person as judged by the standards of the person's culture (the value criterion), and (b) the condition results from the inability of some internal mechanism to perform its natural function, wherein a natural function is an effect that is part of the evolutionary explanation of the existence and structure of the mechanism (the explanatory criterion). (Wakefield, 1992, in Edwards, 1997, pp. 87-8)

This analysis is still subject to criticisms C1 and C2 that we set out above. It is less clear that C3 applies to it, since Wakefield's allowing considerations of value to enter in helps the model to better match our intuitions and existing practice.

A different mixed model comes from Culver and Gert (1982, p. 81). This too has been influential.

> A person has a malady if and only if he has a condition, other than his rational beliefs and desires, such that he is suffering, or at increased risk of suffering, an evil (death, pain, disability, loss of freedom or opportunity, or loss of pleasure) in the absence of a distinct sustaining cause.

This model includes a role for objective fact both in the chance of death, pain, etc., and also in determining whether a condition that is caused by a distinct sustaining cause. For example, it would seem that being homosexual can often cause a person to be unhappy, but in at least many and probably most instances, the reason for this is societal prejudice against homosexuals, which would count as a distinct

sustaining cause. So this model would not necessarily entail that homosexuality is a malady, even if the condition makes people unhappy. On the other hand, the model makes it inevitable that values also enter into the determination of what counts as a malady, most obviously in the decision of which beliefs and desires are rational, and which are irrational.

Furthermore, while in a great many cases, it may simply be an empirical issue whether a person is suffering a disability, loss of opportunity or loss of pleasure, there are at least some cases where values can enter into determining what counts as disability, etc. This is clearest in cases where pragmatic considerations (for example, whether a managed care company should provide funding for treatment) require that a definite decision be made about whether a condition is a malady or not, but where it is debatable whether the condition really is a disability or a significant loss of opportunity. Recent debates over the use of medication for erectile dysfunction have vividly illustrated this for physical disorders; if a man has a condition that means that he is only able to have sex twice a week without taking medication, does that mean he has a disability? It is hard to see how one could provide an answer to this question without making assumptions about what is normal. Also, in recent years some advocates for the deaf have argued that deafness is not a disability, but is rather simply a difference from people who can hear; this too suggests that values can enter into our understanding of what counts as disability. In the realm of mental health, it can be argued that whether a person has lost freedom, for example in the controversial category of substance dependence, different observers may agree on the empirical facts concerning a person who takes drugs, but may still disagree whether she has a disability, and even whether she has diminished freedom; the difference can depend on what normative assumptions are made. For a different example, we can consider a case where a person who in her teen years experienced productive and pleasurable hypomanic episodes, but who by her twenties no longer has such episodes, has undergone a change that should count as a disability or a loss of opportunity. The approach of Culver and Gert does not rely on evolutionary psychology, and so avoids the problems set out in (C1) and (C2) above. If the preceding argument is correct, however, and the criteria for what count as an evil can depend on normative assumptions, then Culver and Gert's general definition of malady suffers from a problem of circularity.

Multiple Personality

There have been many other proposals for how to provide criteria of mental illness. We do not have the space to discuss them all in this section. Furthermore, there have been many contested mental disorders, such as schizophrenia, autism, addiction, homosexuality, attention deficit hyperactivity disorder, oppositional defiant disorder, antisocial personality disorder, and dysthymia. We do not have space to discuss them all, and indeed, the attention these conditions have received has rarely been from philosophers but from other psychiatric critics with different backgrounds.

We will discuss one condition that has gained philosophical scrutiny, that of multiple personality. Philosophers have been particularly interested in this phenomenon because it raises important issues of the understanding of the unity of consciousness and it may be an important test case for various theories of personal identity, maybe providing a counterexample to any theory that implies that there can no more

than one person "in" one body. But in order to discuss these aspects of the phenomenon, philosophers have had to first address what the phenomenon really is. In particular, various skeptics have argued that there is no such thing as multiple personality, or that it is in some way artificial or inauthentic.

In order to introduce the interesting philosophical work on the reality of multiple personality, it is necessary to give a very brief sketch of what it is taken to be.

In multiple personality, it seems that there are at least two distinct personalities within one person. These personalities, or "alters," apparently have profoundly different voices, speech patterns, self-descriptions, memories, character traits, beliefs, desires, and levels of education. Different alters within one body can describe themselves as being of different ages, genders, ethnicities, skin color, height, weight, and eye color. Different alters within one body can fail to be aware of each other, but there can be interaction between them. Sometimes awareness is one-directional: A is aware of the thoughts and actions of B without B being aware of A. Sometimes one alter can directly interfere with the thoughts or actions of another alter. Or so they claim. The number of people diagnosed with multiple personality has varied greatly over time and place: it was first described in the nineteenth century, especially in France and the USA; and after it seemed to become almost non-existent in the first half of the twentieth century, it grew again in the second half, especially in the USA.

There has been a great deal of empirical and methodological debate about the causes and the treatment of multiple personality. It seems that it is linked with childhood abuse, and it may be that splitting into two, or dissociating, is a way of coping with a traumatic experience while it is happening or after it is over. It is also linked with another controversial phenomenon, hypnotism. People who are highly hypnotizable seem to be especially prone to becoming multiple personalities. Sometimes hypnotism has been used by therapists as a way to discover "hidden" personalities.

Skeptics have raised different sorts of objections to multiple personality. Their objections are of several different kinds; here are three.

> (S1) Some have claimed that multiple personality does not exist at all, and it is a hoax by patients and therapists seeking attention or money, or hoping to use it as an excuse for criminal behavior.
>
> (S2) Some have claimed that multiple personality is not really a separate phenomenon, but rather is a unusual form of other mental disorders, such as manic depression, schizophrenia, or borderline personality disorder. This raises a taxonomic issue of when a condition should be classified as an atypical form of a known mental disorder rather than as an instance of a separate, independent mental disorder.
>
> (S3) Some have claimed that multiple personality, while a separate disorder, is caused not by traumatic childhood experiences but rather by overenthusiastic and irresponsible therapists. Vulnerable patients have been encouraged to believe that they have been abused

> as children. The treatment they have then received from therapists, especially hypnotism, has not discovered, but rather created separate personalities.

All of this so far is mostly an empirical matter, depending on a careful scrutiny of the existing data, careful thought about our present taxonomic categories, and probably pointing to further possible research. By far the most sophisticated philosophical work on the reality of multiple personality has been by Ian Hacking, in a series of papers and books since the mid 1980s. Hacking combines careful historical research, an understanding of statistical methods and scientific research, and a grasp of philosophical debates about realism, truth, and nominalism. He addresses the approach of the social sciences, and in particular the claim that psychiatric phenomena are socially constructed. This seems to amount to the claim that while in some sense there might always have been multiple personalities, it is a socially constructed phenomenon in the sense that we have started to bring certain concepts and patterns of ideas to bear on the phenomenon, and that a number of political and social factors are what has led us to frame our thought in this way. This view denies that a label such as multiple personality picks out a natural kind of people.

Hacking does not seem sympathetic to skeptics of the variety (S1) and (S2). In some of his work he seems quite sympathetic to (S3) and also to some of the insights of social constructionism. But he goes beyond most simple forms of social constructionism, and introduces the idea that the people classified by social categories will themselves be affected by the classification. So the issue is more than simply a matter of discussing what concepts we use in framing psychopathology.

> People of these kinds can become aware that they are classified as such. They can make tacit or even explicit choices, adapt or adopt ways of living so as to fit or get away from the classification applied to them. These very choices, adaptations, or adoptions have consequences for the group, for the kind of people that is invoked. The result may be particularly strong interactions. What was known about people of a kind may become false because people of that kind have changed in virtue of what they believe about themselves. I have called this phenomenon *the looping effects of human kinds*. (Hacking, 1999, p. 34).

Hacking has suggested that once we understand these interactions between our categories and the people categorized, we should stop wanting a simple yes or no answer to the question "is multiple personality real?" He argues that there has been a great deal of confusion in debate between the sides often labeled as constructionists and realists, not just about multiple personality, but a whole range of phenomena and categories, including subatomic particles, childhood, emotions, and women refugees. He argues that a central assumption for anyone who argues that X is socially constructed is

> [0] In the present state of affairs, X is taken for granted; X appears to be inevitable.

Those who argue for some forms of social construction of X argue that it is not in fact inevitable, and could be different. Some are content to be purely descriptive about this, while others, taking a stronger position against X, argue that we should construct our categories differently and do away with X, or at

least view the category of X with some suspicion and recognize its contingency.

Even though Hacking finds the language of "social construction" mostly unhelpful, he views dissociation and multiple personality with some suspicion. He calls it an example of an *interactive kind*, created by looping effects, and he explicitly hopes that the category dies away (Hacking 1998, p. 100). He further argues that it is problematic to use cases of multiple personality and dissociation to draw conclusions about the fundamental nature of the mind or personal identity. "Multiple personality teaches nothing about 'the self' except that it is an idea that can be exploited for many ends." (Ibid, p. 96).

Hacking has provided us with the most detailed and careful philosophical approach to addressing issues in classification of mental disorders. His work is bound to be influential.

3. When are People with Mental Illnesses Responsible for Symptomatic Behavior?

There are philosophical positions that say that people are never responsible for their behavior, because their behavior is always determined. There are also some philosophers, such as the early Sartre, who have made statements that seem to imply that people are always responsible for their behavior, or at least all their intentional actions. (See the entry on freewill.) These are, however, extreme positions, and a far more widespread view is that of common sense: people are normally responsible for their behavior, but there are sometimes circumstances that provide them with an excuse. There is still considerable debate as to what makes a good excuse. In past decades some sociological analyses have led to suggestions that some people turn to crime, for instance, because of poverty or discrimination, and that this provides some excuse for the criminal behavior. In more recent decades, such views have received less support. However, there is still considerable support for the idea that people with mental illnesses are not fully responsible for those actions symptomatic of their illness. To what extent people are responsible, and which mental illnesses provide excuses, are issues still very much up for debate.

Three mental illnesses that have received attention from philosophers and psychiatric theorists on the issue of responsibility are schizophrenia, psychopathy, and alcoholism. There are of course many other mental illnesses where the issue of responsibility arises: obvious examples are depression, obsessive-compulsive disorder, manic episodes, paraphilias, and borderline personality disorder. Despite the fact that the various theories of the etiology and nature of these disorders are very suggestive of ways to understand the responsibility of those with the disorders for their symptomatic behavior, these and other mental disorders have received surprisingly little discussion from philosophers vis-à-vis responsibility for action. So we will focus on the disorders that have been discussed.

Schizophrenia

One of the most typical symptoms of schizophrenia is the suffering of delusions. When people with

schizophrenia are suffering extreme and pervasive delusions, they do not understand what they are doing. It is no simple matter to define a delusion, and it is highly problematic to simply equate it with a false belief, but it is safe to say that paradigm cases of delusion imply a significant lack of, or distortion in, understanding of one's situation. In paranoid schizophrenia, for example, patients tend to interpret what other people say with what might be called a hermeneutics of fear and suspicion, and in extreme cases, will have elaborate and fixed fantastic theories about ways in which others are aiming to harm them.

It should be emphasized that schizophrenia causes far more than cognitive distortions. It is a disorder that causes great emotional problems as well, and these often contribute to the bizarre behavior of schizophrenics. Nevertheless, it is the distortions in belief and reasoning that provide the clearest excuse and make it plausible that often schizophrenics are not responsible for their behavior.

It is this sort of case that is central to the insanity defense in the law, and which has received considerable discussion by philosophers and psychiatrists interested in the justification of punishment. There are many different kinds of cases where mentally ill people seem to have some grasp of what they are doing, and that what they are doing is wrong, and it is very difficult to draw clear lines between somewhat similar cases.

Psychopathy

The category of psychopathy is one of the more controversial within psychiatry. The closest that the diagnostic manual DSM-IV-TR comes to this diagnosis is antisocial personality disorder, and the whole category of personality disorder has come under critical scrutiny. Antisocial personality disorder, and the corresponding diagnoses for youth, (behavioral disorders and oppositional defiant disorder), have been especially questioned because they include as symptoms destructive and often criminal behavior. There is a great deal of suspicion of any attempt to excuse the symptomatic behavior of psychopaths.

Some of the debate hangs on the correct explanation of the behavior of psychopaths. Psychopaths are often intelligent and calculating, yet they are also impulsive and pay as little regard for their own long-term interests as they do for that of other people. They can be very emotional, yet they also seem to lack some emotional capacities. In particular, it is still an open question to what extent they comprehend the wrongness of their actions, and can be said to have a conscience. If their moral understanding is extremely limited -- for example an ability to list the kinds of actions that would be classed as morally wrong, but no ability to empathize with those who suffer -- then there is still philosophical work to be done in deciding what this implies for moral responsibility, punishment or treatment.

Another characterization of psychopaths is that they are simply people with deeply flawed characters and no use for morality. This characterization is probably closer to media portrayals of psychopaths than clinical reality, but it still raises philosophical issues. In particular, we can ask, if a person has a bad character, and lacks any interest in or feeling for the welfare of others, then he may not be able to behave well. How can we blame someone for doing what is in his nature? This is an issue for moral theory generally, and arises especially for virtue theory. It is of particular practical consequence when it comes

to judging psychopaths, if this account of their behavior matches any real psychopaths.

To go further into the philosophical discussion of psychopathy would require setting out empirical discussion and ethical theory in far more detail, so we will leave this discussion of psychopathy.

Alcoholism

There has been a great deal of discussion of whether alcoholism should count as a disease, by physicians, philosophers, legal theorists and policy makers. Many factors enter into the discussion, but one way of understanding the literature is that it centers on the issue of whether alcoholics are responsible for their continued drinking. It is plausible that alcoholism is a disease if and only if alcoholics are in some significant way (in need of clarification) not responsible for their drinking. What is less clear is which is logically prior, the disease status of alcoholism or the responsibility of alcoholics for their drinking.

Often one finds claims in popular discourse to the effect that alcoholics are not responsible for their drinking because the drinking is a symptom of a disease, or because it is the disease that causes them to drink. If this is the case, there must be independent evidence that alcoholism is a disease: various sorts of evidence have been suggested, including the withdrawal symptoms that alcoholics experience when they abstain from drinking, physical changes that occur in the brain as a result of excessive long-term drinking, and epidemiological studies that show that there is a genetic component to alcoholism.

It takes little thought to see, however, that these sorts of evidence can't by themselves prove that alcoholism is a disease. How one proves that a condition is a disease depends partly on what criteria of disease we can agree upon, but even without giving a definition of disease, one can see that the claim that the empirical evidence entails that alcoholism is a disease is highly contestable. The existence of withdrawal symptoms does show that it is difficult to stop drinking, but there is a great logical distance between having a habit that is hard to give up and having a disease. The fact that brain abnormalities occur in excessive drinkers is suggestive of a physical disorder, but abnormalities in themselves do not constitute diseases or disorders. The fact that heavy drinking causes changes in people's brains is not in itself surprising. Further evidence about what the effect of the brain abnormality has on the person would be needed, and its correlation with heavy drinking is not enough. Finally, the fact that a habit such as heavy drinking has a genetic component again does not prove that it is a disease. Laziness and cowardice could also turn out to have genetic components, but that would not make them diseases.

The problem for the disease status of alcoholism is that a habitual drinker can be described with such strongly evaluative terms as weak, self-deceiving, selfish, self-destructive, shortsighted, uncaring about other people, and even pathetic. Some would claim that such psychological characteristics provide the best explanation of an alcoholic's problem drinking, and if this is right, then the alcoholism-as-disease explanation is at best secondary, and at worst, utterly wrong-headed. While it is hard to find a description of self-destructive heavy drinking that makes it simply a matter of personal decision, an expression of one's values, or a rational choice, it does seems that problem drinking can often be a self-perpetuating way of life. It is difficult or impossible to locate a specific single cause of the drinking, and it also seems

that the drinker has a role in perpetuating her problem. It is not simply something that happens to her.

Nevertheless, the testimonials and behavior of alcoholics also provide grounds for thinking that they have extreme difficulty in giving up drink, and often no simple exertion of willpower or resolution to give up will solve the problem. Often heavy drinkers try to stop or cut back but fail to do so, even when they know full well that terrible consequences will result from their continuing to drink, and when drinking does not provide pleasure or lasting benefit.

This sort of argument suggests that it is in fact the issue of personal responsibility that is logically prior, and the question of whether alcoholism is a disease depends on it. Thus, the argument would go, alcoholism is a disease because alcoholics cannot control their drinking.

Now, it is very plausible that self-control is a matter of degree; people have more or less control over their behavior. Some people are better at resisting temptation than others. It is rare for a person to have complete control over her actions, and it is almost unimaginable for a person to have no control whatsoever over her actions. It would seem to follow that there must be some degree to which alcoholism is a disease, or to put it another way, the disease status of alcoholism is not all-or-nothing.

However, at least for the purposes of public policy in present circumstances, there is no room for a concept of a condition that is a part or semi-disease. It tends to be that policy must come down on one side or the other on the question whether alcoholism is a disease, but that given the nature of alcoholism, the issue is not simply settled by the scientific and psychological facts. It follows that other considerations can come into play, such as considerations about the social and economic effects of labeling alcoholism a disease. These considerations may tip the scales one way or the other. This may provide a justification of current practice where alcoholism counts as a disease for some purposes but not others. For example, under US law, alcoholism is not a disability covered by the Americans with Disabilities Act, and so is not a condition employers must make allowances for. But treatment of alcoholism by federal health care organizations (such as the Veterans Administration) is mandated by law.

Philosophers have started to discuss the irrationality of alcoholics, how to explain their symptomatic behavior, and to what extent they are responsible for their behavior. Notable examples are Gary Watson (1999), Alfred Mele (1996), and Jon Elster (1999a). There is an astonishing amount of empirical research and psychological models aiming to explain alcoholism, and philosophers will probably find, as they have found with much work on emotion and social psychology, that the literature contains questionable assumptions about fundamental psychological concepts. Central to the philosophical discussion is the examination of the possibility of irresistible desires and the way that cravings can reduce an addict's self-control. Indeed, addiction can provide an important test case for any theory of the nature of action, since it is a prime example of irrationality, and it is important that theories about the nature of practical reasoning be able to give an adequate account of the nature of irrationality.

There is some overlap between this topic and that of the responsibility of weak willed agents, although so

far there has been little systematic discussion by philosophers of the relation between addiction and weak willed action.

Bibliography

Works Cited

- American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision,* DSM-IV-TR. (Washington, DC: American Psychiatric Association, 2000).
- Boorse, Christopher. "What a Theory of Mental Health Should Be." *Journal of the Theory of Social Behavior*, 6 (1976): 61-84.
- Culver, Charles M. and Gert, Bernard. *Philosophy in Medicine*. (New York: Oxford University Press, 1982).
- Edwards, Rem B. (editor). *Ethics of Psychiatry: Insanity, Rational Autonomy, and Mental Health Care.* (Amherst, NY: Prometheus Books, 1997).
- Elster, Jon. *Strong Feelings: Emotion, Addiction, and Human Behavior*. (Cambridge, MA: MIT Press, 1999a).
- Elster, Jon (editor). *Addiction: Entries and Exits*. (New York: Russell Sage, 1999b)
- Hacking, Ian. *Mad Travelers: Reflections on the Reality of Transient Mental Illnesses*. (Charlottesville: University Press of Virginia, 1998).
- Hacking, Ian. The Social Construction of What? (Cambridge, MA: Harvard University Press, 1999).
- Mele, Alfred R. "Addiction and Self-Control." *Behavior and Philosophy* 24 (1996) 99-117.
- Perring, Christian. "The Neuron Doctrine in Psychiatry," Peer Commentary on "A Neuron Doctrine in the Philosophy of Neuroscience" by Ian Gold and Daniel Stoljar, *Behavioral and Brain Sciences* 1999, 22
- Reznek, Lawrie. *The Philosophical Defense of Psychiatry*. (New York: Routledge, 1991).
- Szasz, Thomas. *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct* (New York: Harper and Row, 1974.)
- Wakefield, Jerome C. "The Concept of Mental Disorder: On the Boundary Between Biological Facts and Social Values." *American Psychologist* 47, no. 3. (1992): 373-88.
- Watson, Gary. "Disordered Appetites: Addiction, Compulsion, and Dependence." in Elster (1999b).

Further Reading

- Clark, Lee Anna (editor). "Special Section: The Concept of Disorder: Evolutionary Analysis and Critique"; *Journal of Abnormal Psychology*, August 1999, vol. 108, no. 3, pp. 371-472.
- Elliott, Carl. *The Rules of Insanity: Moral Responsibility and the Mentally Ill Offender*. (Albany: SUNY Press, 1996).

- Fulford, K. W. M. *Moral Theory and Medical Practice*. (Cambridge, UK: Cambridge University Press, 1989).
- Fulford, K. W. M. "Value, Action, Mental Illness, and the Law." In S. Shute, J. Gardner, and J. Horder, eds., *Action and Value in Criminal Law*. (Oxford: Clarendon Press, 1993), pp. 279-310.
- Graham, George, and G. Lynn Stephens (editors). *Philosophical Psychopathology*. (Cambridge, MA: MIT Press, 1994).
- Humber, J. M. and R. F. Almeder (editors). *What is Disease?* (Totowa, NJ: Humana Press, 1997).
- Lennox, James. G. "Health as an objective value." *The Journal of Medicine and Philosophy*, 1995, 20, 499-511.
- Reznek, Lawrie. *The Nature of Disease*. (London: Routledge & Kegan Paul, 1987).
- Sadler, John Z. and K. W. M. Fulford, (editors). "Special Issue: Aristotle, Function, and Mental Disorder." *Philosophy, Psychiatry, & Psychology*. 2000, vol. 7, no. 1.
- Sadler, John. Z., Osborne P. Wiggins, and Michael. A. Schwartz, (editors). *Philosophical Perspectives on Psychiatric Classification*. (Baltimore: Johns Hopkins University Press, 1994).
- Wilkinson, Stephen. "Is ‘Normal Grief’ a Mental Disorder?", *Philosophical Quarterly* 50, no. 200 (2000): 289-304.
- Zachar, Peter. *Psychological Concepts and Biological Psychiatry: A Philosophical Analysis.* (Amsterdam: John Benjamins, 2000).

Other Internet Resources

- Philosophy of Psychiatry Links (maintained by the author)

Related Entries

free will | neuroscience, philosophy of

Christian Perring
Dowling College
cperring@yahoo.com

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 30, 2001
Content last modified: November 30, 2001

Economic Analysis of the Law

Economic analysis of law applies the tools of microeconomic theory to the analysis of legal rules and institutions. Ronald Coase [1961] and Guido Calabresi [1961] are generally identified as the seminal articles but Commons [1924] and Hale [1952] among others had brought economic thinking to the study of law in the 1910s and 1920s. Moreover, as I will elaborate below, economic analysis of law derives from several different intellectual traditions in economics.

Richard Posner [1973] brought economic analysis of law to the attention of the general legal academy; by the late 1970s, his work had provoked a vigorous controversy within the legal academy. That controversy has usually defined the debate around the philosophical foundations of economic analysis of law. Posner made two claims: (I) Common law legal rules are, in fact, efficient; and (II) Legal rules ought to be efficient. In both claims, "efficient" means maximization of the social willingness-to-pay. In the course of the controversy, two other claims were articulated in Kornhauser [1984, 1985]: (III) Legal processes select for efficient rules; and (IV) individuals respond to legal rules economically. (In this third claim, "efficient" means "Pareto efficient.") Kornhauser identified this last, *behavioral* claim as central to the enterprise. A fifth claim is also implicit in the literature: (V) on the best interpretation of law, common law doctrines promote efficiency.[1] Notice that (V) differs from (I) in important respects. According to (V), an economic interpretation fits a doctrine not because, as asserted in (I), the legal rules in fact induce efficient behavior but because the rule would induce efficient behavior within the view of the world that seems to underlie the judicial decisions. (I) is an empirical claim that requires the analyst to determine whether the actual behavior induced by legal rules is efficient; it requires knowledge of how individuals do, in fact, behave and of which behavior in the real world would, in fact, be efficient (V) requires only knowledge of the content of judicial opinions; the analyst interprets these opinions to extract an economic model that underlies the decision. (V) might be true even though legal rules induced inefficient behavior in the real world because the announced legal rule might be efficient within the implicit model used by judges.

These five claims do not correspond directly to traditional questions in the philosophy of law. The *evaluative claim* (II) that legal rules ought to be efficient would, if directed to judges, qualify as a theory of adjudication, one of the central concerns of anglo-american philosophy of law. Central philosophic questions concerning the concept of law, of its normativity, and the obligation to obey the law, however, are not directly addressed. The behavioral claim as well as the evolutionary claim (III) and the positive claim (II), by contrast, concern empirical issues that philosophers of law generally neglect. Nevertheless, the controversy within the legal academy has generally regarded economic analysis of law as providing a comprehensive theory of law that challenges traditional approaches to law. Indeed, an explanation of the

vehemence of the controversy should identify differences in fundamental views concerning law.

---

1. Two Strands of Thought within Economic Analysis of Law

The vast literature of economic analysis of law is not easily characterized. For purposes of this essay, I identify two distinct strands of thought within economic analysis of law. I shall call one strand *policy analysis* and the second strand *political economy*.[2] These two strands may be differentiated along a number of dimensions.

First, policy analysis generally focuses on analysis of the effects of legal rules and institutions on *outcomes*. An outcome usually consists of the "objective" effects of the rule or institution on the behavior of "private" individuals. By contrast, political economy generally investigates the operation of political institutions such as courts, legislatures, the executive and administrative agencies; it usually focuses on the behavior of the public officials within those institutions. Ideally, one would trace the effects of different institutional rules and structures through the behavior of public officials to the effects on the behavior of private individuals. In practice, however, tracing effects of changes in institutional rules to final outcomes is too difficult and too uncertain. A change in the structure of legislative institutions, for example, would likely affect the content of the legislative programs enacted in the jurisdiction. To trace effects to final outcomes in terms of the behavior of private individuals would thus require the analyst to predict the set of statutes that would be enacted within various legislative structures.

Second, and related to the first, policy analysis generally assumes that public officials in general and judges in particular, are conscientious. Judges thus enforce the legal rules as they are announced, regardless of the judge’s own view of the desirability of the legal rule or its impact on her personally. Political economy, by contrast, assumes that public officials have the same motivation as private individuals; they are self-interested. In the context of adjudication, as will be elaborated below, the political economist interprets self-interested judicial behavior as decisions that promote the policy

preferences of the judge.

Third, policy analysis generally adopts a welfarist stance towards evaluation of legal rules while political economy has evolved from a more contractarian tradition. Policy analysts, when evaluating legal rules ask whether that legal rule induces behavior that satisfies some welfarist criterion, usually either Pareto efficiency or (constrained) social welfare maximization.[3] Political economy, however, has to a large extent emerged from an economic tradition, exemplified by James Buchanan,[4] that rejects the maximization of social welfare as a criterion and seeks to evaluate political institutions on grounds of consent or, more generally, within the contractarian tradition.

Fourth, we might understand the distinction between policy analysis and political economy as a difference in the view of the instrumentalism of law. Policy analysis tends to proceed legal rule by legal rule. It asks, for example, how does a change in the standard of care affect the behavior of tortfeasors and tort victims? Or how does contracting behavior differ if the measure of damages shifts from expectation damages to reliance damages? The analyst thus imputes a purpose (usually, but not necessarily, of maximization of social welfare) to the promulgator of the legal rule. The analyst then assumes that the policymaker has chosen the legal rule that best promotes her (imputed) objective. Legal rules are then instrumental to the achievement of the posited goal; call this approach *rule instrumentalism*.

The political economist, by contrast, generally denies that any purpose can be attributed to the promulgator of a legal rule largely because legal rules are not promulgated by a single individual with power to control unilaterally the content of the rule. Certainly, from the perspective of political economy, legislators have no common purpose and one should not assume or expect that any statute maximizes social welfare. Legislation results from the interplay of interest groups that do not reflect all interests within society. Even if the legislature did reflect all interests within society, each interest does not have an equal (or proportionate) say in the formulation of the statute.[5] Finally, even if each interest did play a "proportionate" role in the formulation of the statute, Arrow's General Possibility Theorem teaches that the aggregation of interests might still not yield a coherent purpose. Political economy thus rejects rule instrumentalism.

One might attribute the rejection of rule instrumentalism within political economy to a commitment to an explanatory rather than a normative project. At the level of constitutional political economy, however, the research program usually adopts the perspective of a constitutional designer and this designer arguably has a view of law that includes *institutional instrumentalism*: i.e., legal institutions, rather than specific legal rules, promote the specific goals of the constitutional designer.[6] The constitutional designer seeks a political structure that promotes her goals. The project of constitutional political economy is thus normative in nature. Indeed the normative nature of the project dominates any explanatory aim. Many within the project - see Brennan and Buchanan [1981, 1985] - argue that one ought to adopt an economic theory of behavior of public officials and private individuals even if that theory is not the best explanatory theory.

These last two differences suggest that the two strands of economic analysis of law may endorse radically

different positive and normative theories of adjudication. These theories are sketched and discussed in section 5 below.

More significantly, for purposes of this entry, however, is the basic similarity between these two strands of economic analysis of law. Both of these strands adopt the standard assumption of neo-classical economics that each individual seeks to maximize her preferences. Moreover, they generally assume that each individual acts in her own self-interest, narrowly defined. This approach presents the single, greatest obstacle to the articulation of a general theory of law that confronts economic analysis: it has no room for the *normative* aspect of law. It is this denial of the normativity of law that accounts for the vehement resistance that economic analysis provoked within the legal academy.

2. The Concept of Law

Though commentators often characterize economic analysis of law as providing a comprehensive theory of law, its narrower ambitions become apparent when one realizes that economic analysis of law has not explicitly addressed the question "What is law?". Indeed, economic analyses of law generally *presuppose* a concept of law in that the law is uncontroversially known to all actors.

One might extract two quite different accounts of the concept of law from the practice of economic analysts. On one interpretation, economic analysis of law relies on a straightforward theory of legal positivism. On the second interpretation, at which the end of the last section hinted, economic analysis of law assimilates the analysis of law to the analysis of social phenomena generally; it postulates, then, that there is no analytically useful concept of law.

2.1 Legal Positivism and Policy Analysis in Economic Analysis of Law

The policy analysis strand of economic analysis of law often implicitly adopts some variant of legal positivism as its understanding of the concept of law. Recall that the policy analysis treats the behavior of judges in particular (and sometimes public officials generally) differently from the behavior of those subject to the legal rules.

An economic analysis of the behavioral effects of a legal rule generally begins with the assumption that the legal rule is clearly known not only to judges and other public officials but also to those subject to the legal rule. This knowledge of private citizens might amount simply to the knowledge of what consequences follow from each possible action the agent might take. Actions that provoke a response from public officials generally, or judges in particular, have no special character to them; the citizen in her deliberations treats the consequences of rule-following or rule-breaking as she treats any other price. On Hart's account of legal positivism, however, a private citizen may adopt this detached attitude towards legal rules. The concept of law inherent in policy analysis is thus consistent with positivism.

The typical model, however, assumes that public officials conscientiously apply the legal rule under study. The public official does not identify the rule that would best promote her own preferences and then apply (or not apply) that rule; rather she "conscientiously" applies the rule that "ought" to govern the event. Conscientious application here simply implies that the official may uncontroversially apply an identified legal rule to the events in question.[7] This assumption might reflect a "partial equilibrium" approach to the analysis of the problem at hand. If the effects of the legal rule are the central focus of inquiry, the incentives and behavior of public officials who enforce that rule may be of less interest. The analysis of the institutional structures and processes that insure the "conscientious" application of law by public officials are left for later analysis.

Other aspects of the economic analysis of law are consistent with this positivist approach to the law. Economic analysis of social norms, for example, often provides a characterization of social norms that largely coincides with Hart's own scheme for distinguishing social rules that are legal rules from social rules that are not. Specifically, economic analysts of law point to the decentralized character of the promulgation and enforcement of legal rules as the properties that distinguish social norms from legal rules.

2.2 The Elimination of Distinctively Legal Phenomena

As noted earlier, economic analysis of law has no room in its theory of behavior for normative concepts, including ones as basic to law as duty and rule.[8] This aversion to normative concepts takes a moderate form within the policy strand; private individuals act only prudentially but public officials act conscientiously to meet their legal obligations. The political economy project within economic analysis of law adopts a more extreme ambition; both private citizens and public officials have only prudential motivations. Political economy aspires to eliminate normative motivations; or at least to reduce normative motivations to prudential ones. This ambition has dramatic consequences for one's understanding of law.

2.21 A Legal System as a fusion of a legal order and a legal regime

Legal phenomena present philosophical problems in significant part because they have both normative and institutional aspects. I shall call the normative aspect of a legal system the *legal order* and the institutional aspect the *legal regime*. Together, the legal order and the legal regime constitute the *legal system*. The legal order consists of the legal norms of the system; these norms are expressed in (but not co-extensive with) the Constitution (if any), statutes, administrative regulations, and judicial decisions of the jurisdiction. The legal regime consists of the legal institutions extant in the jurisdiction. These clearly include the legislatures, executives, administrative agencies, and courts but also may include such institutions as the bar and legal education.

One might create a taxonomy of legal theories in terms of the relation between the legal order and legal regime presumed by that theory. It is useful to characterize two extreme positions within this taxonomy. At one extreme, the legal system is understood completely in terms of the legal order. At the other extreme, the legal system is understood completely in terms of the legal regime.

What does it mean to understand the legal system solely in terms of the legal order? More specifically, how does one understand the legal regime solely in terms of the legal order? Hart's legal positivism provides a useful starting point. As Hart notes, the legal order consists of both primary and secondary rules. Secondary rules are constitutive of the legal regime; they define the powers and obligations of public officials. For secondary rules to explain fully the legal regime, they must create a comprehensive set of obligations; that is, the obligations of a public official must dictate her action in every instance. Hart, of course, denied that secondary rules could do this so his positivism does not lie at this extreme end of the taxonomy.

A comprehensive specification of the obligations of public officials is not sufficient to explain the legal regime solely in terms of the legal order. In addition, one must assume that public officials are *conscientious*.[9] A conscientious public official always *meets* her obligations. Moreover, she always meets them for an appropriate reason. Moral reasons are appropriate reasons for meeting an obligation. Prudential reasons are inappropriate in the sense that the secondary rules would then not serve as a motivation for the agent.

The parallel question posed by the other extreme is more easily understood. To understand the legal order solely in terms of the legal regime requires only that one defines norms of the legal order in terms of the behaviors of the public officials within the legal regime. The legal order plays no role in the motivation of public officials. More strongly, one might argue that the legal order does not determine which individuals are public officials. Rather, the behavior of a certain group of individuals determines what obligations other individuals have. Here "obligations" are defined solely in terms of the consequences that follow from non-compliance. That is, individuals have purely prudential obligations.

Notice that Hart's legal positivism might be understood as falling close to either extreme category of this taxonomy. I elaborated the explanation of the legal regime in terms of the legal order by reference to Hart's legal positivism. Consideration of the grounds of acceptance of the rule of recognition permits one to see how Hart's positivism might be understood to explain the legal order solely in terms of the legal regime. The rule of recognition is defined behaviorally in terms of the rule that law-applying public officials accept. The reason for acceptance is controversial. Hart [1961] seems to place no restriction on the motivation a public official might have. Others [e.g Holton, 1998], however, have argued that public officials must have a moral reason, not a prudential one.

2.22 Political economy and the concept of law

Political economy aims to explain all legal phenomena in terms of the self-interest of agents; it renounces the concept of normatively guided behavior. For the political economist, then, the legal regime explains completely the legal order because it interprets legal obligation solely in terms of the incentives that "legal rules" create for individuals.

On this account, of course, legal rules do not play any role in the explanation of behavior of either private individuals or public officials. An individual faced with a choice considers the costs and benefits that each

option presents to her. These costs and benefits will include "legal costs and benefits" but these costs and benefits are not determined by rules; they are the result of the incentives that private and public officials face. Rules are only rules of thumb that express the response of average individuals under normal circumstances to particular events. Which rules of thumb are used, of course, may greatly affect the social equilibrium achieved in a particular jurisdiction.

3. The Obligation to Obey the Law

In the previous section, I discussed the role that legal rules played in the practical deliberations of private agents. On that account, the sanctions imposed for non-compliance with the rule provided agents with prudential reasons to conform the legal rule. Independent of those sanctions, however, the agent had no reason to obey the law. I argued further that, within the political economy strand of economic analysis of law, public officials had only prudential reasons for conformity to their public obligations. From the perspective of the last section, then, no one has a general obligation to obey the law.

In this section, I suggest that economic analysis nonetheless offers a structure within which one may articulate prudential accounts for a general obligation to obey the law.

3.1 Equilibrium Selection

Legal positivism grounds law in social practice. Its difficulty in explaining the normativity of law emerges directly from this attention to the social nature of law; social facts themselves, it would seem, cannot give rise to any obligation. Various authors (e.g., Postema [1982]) have tried to resolve this conundrum through an analysis of convention as a coordination game.

In a coordination game, the interests of all players are coincident; each player ranks the possible outcomes of the game identically. The difficulty for the players arises because they do not know which of the multiple equilibria of the game to play. Consider, for example, an island in a pristine legal state-there are no legal rules. The individuals must decide on which side of the road to drive. Formulated as a game, each individual has two strategies: drive on the right (*R*) and drive on the left (*L*). Each has an identical evaluation of the outcomes of the strategy choices of everyone. Each ranks the outcomes in which everyone chooses the same side of the road -- all *R* or all *L* -- highest and each ranks the outcome in which half choose *R* and half choose *L* worst.[10] This game has two equilibria: all choose R and all choose L. Unfortunately, knowledge that two equilibria exist does not help agents determine which strategy to adopt, or, alternatively which equilibrium to play.

Announcement of a legal rule in this context can coordinate the players' actions. It gives each a reason to choose as the rule dictates if it affects the individual's beliefs concerning which strategy the other agents will adopt. On this account, the social fact that individuals accept the law provides each individual with a reason to act. This reason is independent of any sanction that the law might impose for non-compliance. Moreover, this reason is prudential in that it best promotes the agent's own welfare and moral, in the

sense that it best promotes the well-being of all. This coincidence results from the coincidence of interests of all agents.

Law understood as a coordination device at best provides a partial account of the grounds of normativity. Obviously, many if not most laws concern conduct in which the interests of agents do not coincide. Coordination cannot provide agents with a reason to act in this case. Moreover, it is not clear that the acceptance of a rule of recognition by public officials constitutes a coordination game. Consider a specific judge *J*. That judge may think rule R is the best rule of recognition. Consistent with that belief she considers a world in which all judges accept *R* as best. Nothing guarantees that a second judge J′ consider *R* best. He may think R′ better. J′ will also believe that a world in which all judges accept R' is best. But it is not obvious that *J* must consider a world in which all accept R′ as preferable to one in which a majority accept *R*.

Finally, note that the argument is incomplete. It requires that individuals have sufficiently common knowledge of the law and that others know the law for it to have any plausibility. Even under these circumstances, however, the argument still requires that each infer from this common knowledge that each individual will comply.

3.2 Bounded Rationality and the Conservation of Decision Resources.

In the simplest model in which such an account exists, agents face a cost to deliberation. The more complex the deliberative calculation, the more costs the agent incurs. When the marginal cost of deliberation is sufficiently high, the agent might do better to follow a rule of thumb that quickly, and cheaply, identifies a good but not optimal, action.[11] If the expected benefit from choosing the optimal action (relative to the good action) is less than the cost, it is prudent for the agent to adopt the rule of thumb. More sophisticated accounts of an economic rationale for rule-following rely on more complex models of bounded rationality.

To complete an economic account of the authority of law requires that one explain why the agent should consider legal rules as the relevant rules to which she should defer. One might argue that those who promulgate legal rules have special expertise that makes it likely that they will enact rules that are better than the rules that the agent herself would formulate. For some legal rules-technical rules concerning health and safety promulgated by administrative agencies-this argument may have merit. After all the decision at issue depends on a mass of technical data that is not easily assimilable or manipulable. For many other legal rules promulgated by legislatures and courts, however, this argument may not apply.

Several other features of this argument merit attention. First, it parallels the argument for authority offered by Joseph Raz.[12] Moreover, as in Raz's argument, authority is specific to legal rules rather than to law in general. An agent might believe the law more expert than she with respect to some decisions than to others. In fact agents with different expertise themselves would find different legal rules authoritative.

Second, on this account of authority, the legal rule affects the agent's deliberation not because of the sanction for non-compliance-as in the view of legal rules as incentives--but because compliance with the legal rule even in the absence of a sanction is in the agent's interest. This feature of the account of authority conforms to notions, developed further below, of the way in which rules enter the deliberative process. But this feature also limits the applicability of the account to those legal rules that bear on the agent's immediate interest. Many legal rules direct the agent to adopt actions that raise her own costs; in the absence of a sanction for non-compliance her own interest would dictate non-compliance. So, for example, a rule requiring that an agent adopt due care in certain activities may raise the agent's costs.[13]

The prudential account of authority outlined above primarily addresses private individuals. One might ask the parallel question concerning the obligation to obey the law of public officials. In some respects, this question has greater significance than the question concerning private individuals because many acknowledge that the motivation of private individuals to obey the law is usually prudential, the desire to avoid sanction. Moreover, on some jurisprudential accounts, most notably H.L.A. Hart's version of legal positivism, the attitudes and behavior of public officials determine the existence and nature of law.

The economic account of authority, however, does not provide a compelling explanation of official behavior. Consider how the economic account applies to public officials. The relevant obligations here are the official obligations of the individual: the judicial obligation to decide cases according to the dictates of *stare decisis* and other obligatory practices; the executive official's obligation to apply the law. Two difficulties arise immediately. How is compliance with these official obligations in the individual's interest? Why must the agent follow a rule rather than optimize in each instance? This second difficulty is less troublesome than the first; Ronald Heiner [1986], for example, has offered a prudential account, grounded in bounded rationality, of the judicial obligation to adhere to *stare decisis*.[14]

One might attempt to resolve the first difficulty concerning the agent's interest by arguing that compliance with official obligation is in the individual's interest because she desires to maintain her employment. But this explanation rests on an incentive argument. The sanction of dismissal induces compliance rather than a normative motivation to comply with one's obligation; it is another prudential account. The prudential account of authority thus fails to overcome this first difficulty. It is not clear then that the prudential account of authority can ground a positivist conception of law.

4. Theories of Adjudication

4.1 Adjudication in Political Economy

As noted earlier the political economy strand of economic analysis of law itself contains two strands that are in tension with each other. On the one hand, the political economy strand seeks only to explain legal phenomena rather than to prescribe either the structure of legal institutions nor the content of particular legal rules. One might find within this strand of political economy a positive theory of adjudication but not a normative theory. Indeed, the positive theory advanced argues that judges seek to promote their

interests. Usually, these interests are defined as policy interests, that is, an interest to promote particular policies.

The second strand of political economy, *constitutional political economy*, does have normative aims. It assumes that political actors will act in a self-interested fashion *within* existing political institutions but that agents will act more impartially in the *design* of the political institutions within which they will work.[15] A normative theory of adjudication does emerge from this strand of political economy but it differs significantly from the normative theory endorsed by the policy analysis strand of economic analysis of law. For constitutional political economy, a normative theory of adjudication must be a structural one; it should describe the structure of adjudication. The theory thus cannot dictate directly judicial motivation because, according to political economy, judges will always act self-interestedly. Adjudicative institutions, however, can be designed to align better the interests of judges with the interests of the designer of the constitution.

In 1975, Landes and Posner offered a justification for the independence of the judiciary that is often understood as a normative theory of adjudication within the tradition of constitutional political economy. On the account of Landes and Posner, an independent judiciary serves the interest of legislators who seek to impose their policies on the jurisdiction for periods that exceed the length of their majority in the legislature. As a consequence, they find it in their interest to have the judiciary enforce the original bargain struck in all legislation.

This argument contains a normative theory of statutory interpretation. Judges ought to enforce the bargains reached by the legislature that enacted the statute. On this account, a judge ignores the views of the current legislative majority. She also eschews interpretation of the statute in terms of her own policy preferences.

One should note that, from the perspective of constitutional political economy, the argument of Landes and Posner is incomplete. They ground their theory of judicial independence in the interests of legislators. The interests of legislators within extant legislative institutions may not coincide with the interests of the constitutional designer.

4.2 Adjudication in Policy Analysis

A normative theory of adjudication was among the earliest claims advanced in the economic analysis of law. Posner [1973, 1979, 1980, 1985, 1990, 1995] asserted claim II in the introduction: the common law *ought* to be efficient. He interpreted efficiency as "wealth maximization" but then interpreted wealth maximization as "willingness to pay." This interpretive stance yielded an argument that judges in (common law) cases ought to choose the legal rule that maximized the ratio of benefits to costs as measured by the sum of individual willing nesses to pay.

Posner’s claim evoked great controversy in the late 1970s and early 1980s. (See, e.g., *Symposium* [1980]). Twenty years later, Kaplow and Shavell [2001] revived and revised Posner’s claim. The revision had two

components. First, and most important, they chose *welfarism* generally rather than cost-benefit analysis in particular as the normative basis for adjudication. Welfarism requires that evaluation depend solely on the well-being of individuals. Cost-benefit analysis is thus a form of welfarist evaluation; but Kaplow and Shavell's argument allows them to avoid various criticisms of cost-benefit analysis. Second, Kaplow and Shavell do not argue primarily for a normative theory of adjudication. Rather they contend that evaluation of legal rules and institutions by scholars ought to be welfarist. They suggest however that judges by and large have the same evaluative obligation as the third party analyst.

4.21 A brief critique of cost-benefit analysis as a theory of adjudication

Cost-benefit analysis attempts to implement a Kaldor-Hicks evaluative criterion. According to the Kaldor-Hicks criterion, a distribution of goods (broadly understood) *X* is superior to a distribution of goods *Y* if and only if there exists a third distribution of goods *Z* such that (a) *Z* is a redistribution of the distribution *X*; and (b) *Z* is Pareto preferred to *Y*.[16]

Cost-benefit analysis proceeds in two steps. First, for each individual, it identifies a particular representation of the individual's ordinal ranking of the options open to the policy maker. Second, it aggregates these representations of each individual's preferences into a social ranking.

The first step is unproblematic. Consider agent *K*. *K* has preferences over states of the world. A representation of these preferences assigns a number to each state of the world such that *K* prefers state *X* to state *Y* if and only if the number assigned to state *X* is higher than the number assigned to state *Y*. Cost-benefit analysis assigns as numbers the agent's willingness to pay. This procedure thus links the range of numbers that the agent may assign to the agent's wealth as willingness to pay is defined in part in terms of the agent's ability to pay.[17] The procedure for assigning numbers on the basis of an individual's willingness to pay in fact yields a representation of that agent's preferences.

The second step of cost-benefit analysis is more problematic. To aggregates the individual willingnesses to pay, cost-benefit analysis simply sums the individual willingnesses to pay. One can see immediately several difficulties with this procedure. First, each ranking is ordinal; the numbers have no significance beyond the order. If *K* assigns a number 2 to state *X*, a number 4 to state *Y* and a number 16 to state *Z*, we cannot conclude anything about *K*'s intensity of preference; she does not prefer *Z* to *Y* six times as much as she prefers *Y* to *X*. It therefore seems odd that one can add agent *K*'s willingness to pay to agent J's willingness to pay.

Second, cost-benefit analysis adopts a method of interpersonal comparisons of well-being that is particularly unconvincing. Interpersonal comparison of well-being requires that one identify the appropriate representation of each individual's preference ordering and compare those representations. Cost-benefit analysis however does not identify representations on moral or political grounds; rather it chooses the representations that contingently arise from the actual distribution of wealth and income in the society.[18] If Tom is poor while Bill is wealthy, it is unclear why the representations of the well-being of each that derives from willingness to pay provide interpersonally comparable measures. Equally, if

Tom and Bill are equally wealthy but Tom is disabled and Bill is not, the willingness to pay of each may still not be interpersonally comparable.

4.22 A Structural Critique of Welfarist theories of adjudication

One might construct a normative theory of adjudication at either of two levels. First, one might take as given the general structure of adjudication within a particular judicial system and ask what obligations the judges within that system ought to have. Second, one might more fundamentally design the judicial system from scratch. On this second account, the institutional environment in which judges act as well as the obligations of judges within that institutional environment would be subject to evaluation.

Most normative theories of adjudication are of the first type. They take the institutional structure in which adjudication occurs largely as given and then identify the obligations of judges within that system. Phrased differently, normative theories of adjudication are interpretive of an ongoing practice rather than efforts to design a practice from scratch. Welfarist theories of adjudication face several difficulties when understood as interpretive theories of existing (common law) practice.[19]

First, the structure of adjudication does not generally provide adequate or appropriate information for the selection of rules that maximize social welfare. Adjudication in a common law system usually focuses on a past transaction between particular parties. This transaction may not be typical of transactions that were litigated; it certainly will not be typical of the entire population of transactions that a rule would govern. Under a given rule, for instance, the set of transactions that do not lead to litigation are likely to differ systematically from the set of transactions that do give rise to litigation. Equally important, different legal rules are apt to generate different sets of transactions. The current structure of adjudication does not provide any information that would help a decision maker assess these differences across potential legal rules.

Second, the selection procedure for judges does not identify individuals with the appropriate training and background to make accurate calculations of social welfare. Judges in common law countries have generally not been trained systematically in economics and statistics, two disciplines necessary (but not sufficient) for the determination of social welfare under alternative legal rules.

Third, and related, judges usually face severe constraints in the set of legal rules they may consider in any adjudication. When confronted by a tort case, for example, the court usually considers a limited number of legal regimes; perhaps it will reformulate the standard of care or shift from a regime of negligence to one of strict liability. A court, however, is unlikely to adopt a complex scheme of no-fault insurance or to impose a different insurance scheme even though these more radical transformations of social institutions would provide higher overall welfare.

5. Evaluation of Legal Rules and Institutions

The evaluative tradition in economics is resolutely welfarist. That tradition extends to the policy analysis

branch of economic analysis of law. In the prior section, I considered the manifestation of this tradition in the advocacy of cost-benefit analysis as a normative theory of adjudication. In this section, I consider arguments for welfarism as an evaluative standard against which to appraise legal rules and institutions. Welfarism here is not advocated as a theory of adjudication. Consequently, the structural critique of section 5.22 does not apply. The argument here thus lies almost wholly on philosophic territory.

The argument for welfarism as the evaluative criterion for legal rules and institutions has two key elements.[20] The first identifies an individual's well-being with her preferences. Thus, an individual K has greater well-being in state *X* than in state *Y*, if and only if she prefers state *X* to state *Y*. The second, key element of the argument welfarism is a strategy of incorporation. The strategy of incorporation includes within the agent's preference ordering anything that the agent considers relevant to her decisions. The individual's preferences thus correspond to her all-things-considered judgments; that is, *K* prefers state *X* to state *Y* if and only if *K* believes, all things considered, that state *X* is better than (or ought to be promoted rather than) state *B*. K's all-things-considered judgments, of course, will included many considerations that, in ordinary language, are not usually considered as in *K*'s self-interest or even as contributing to her well-being. So, for example, her all-things-considered judgments will incorporate concerns for the well-being of others as well as considerations of justice and deontological constraints on action.

The success of this argument for welfarism depends on the success of this strategy of incorporation of every reason for action into an agent's preference ordering. I shall raise here two objections to this strategy of incorporation. First, the resulting extended preference ordering does not correspond to a concept of well-being that is morally compelling because not all concerns incorporated into the individual's preferences play the appropriate role in her deliberations. Second, aggregation of these extended preference orderings does not treat the moral reasons incorporated in the extended preference ordering appropriately.

A complete argument that the extended preference ordering does not correspond to a morally compelling conception of well-being would require articulation of the concept of well-being, a task well beyond the scope of this entry. For purposes of this critique, it may suffice to note the varying levels of choice at which various concerns apply. The strategy of incorporation requires that one incorporate a moral concern into an individual's extended preference ordering when that moral concern motivates an individual's choices. While it may be the case that the individual in fact prefers a world in which the given moral concern motivates to one in which it does not, the preference for the moral concern arises in the choice of institutional or social arrangements not in the choice within a given set of institutions or social arrangements. The choice within the given institutions may reduce the agent's well-being rather than promote it.

Turn now to the second objection. Even if extended preference orderings do not correspond to well-being, a defender of welfarism might argue that it is appropriate to aggregate these orderings rather than ones that capture the conception of well-being. This argument, however, misunderstands the distinction between judgment and preference.[21] Many of the concerns that enter into an agent's all-things-considered decisions are judgments rather than preferences. Appropriate techniques of aggregation of

judgment differ from those of aggregation of preferences.

Crudely, preference differs from judgment in two respect. First, an expression of preference is personal. The individual expresses a preference that is valid for her; she makes no claim about the validity of this preference for others. Liza's statement that she prefers to be a jazz musician to a lawyer makes no claims concerning Henry's preferred profession. Judgments generally have a greater scope. If Liza claims that jazz musicians contribute more to social welfare than lawyers, she is not expressing a preference that this be so; she asserts that it is true for everyone (at least within current social arrangements).

Second, an individual is sovereign over her own preferences but not over her judgments. An individual has final say over her preferences but not over her judgments. Henry's assertion that he prefers to be a lawyer rather than a jazz musician provides no reason for Liza to reconsider her statement that she prefers to be a jazz musician. If, however, Henry asserts that lawyers contribute more to social welfare than jazz musicians, Liza does have reason to reconsider her contrary judgment. Only one of them can be correct.[22]

Many of the concerns that, under the strategy of incorporation, are included in an individual's extended preference ordering sound in judgment rather than preference. Concerns for distributive justice, for example, reflect moral judgments and not expressions of preference.[23] Similarly, respect for deontological constraints on action sound in judgment not preference. Moreover, these moral judgments depend on proof and argument.

The distinction between preference and judgment points to several difficulties with this argument for welfarism . First, to the extent that individual well-being is understood subjectively, it is likely to be a matter of preference not judgment. Yet the extended preference orderings that constitute the domain of the social welfare function include judgments; this conflation of judgment and preferences provides an argument in addition to the one above that these orderings are not equivalent to well-being.

Second, the methods of aggregation of judgment are likely to differ from the methods of aggregation of preference. Rational aggregation of judgments is likely to parallel rational aggregation of belief. When individuals have different beliefs, one does not generally resolve the conflict through some simple process of aggregation.[24] Rather, the individuals pool information; then each individual updates her beliefs in light of the new information. Ideally, this process leads to convergence of belief. Similarly, the process of moral argument ideally leads to the revision of individual moral judgments.

6. Concluding Remarks

Though the controversy over economic analysis of law has waned, its project continues to disquiet many scholars who study legal phenomena. The prior discussion identifies two distinct sources for that disquiet.

Many legal scholars object to the normative theory of adjudication advanced by policy analysts. These

scholars generally reject the welfarism to which policy analysis is committed. The prior discussion suggests, however, that a rejection of welfarism as a moral theory is neither necessary nor sufficient for the rejection of the normative theory of adjudication advanced by policy analysts.

The methodology of economic analysis of law poses a more significant challenge to traditional accounts of law. Economic analysis of law provokes disquiet because the model of self-interested maximization of preferences does not admit a concept of normativity but explaining the normativity of law is a central pre-occupation of philosophy of law. The logic of this commitment to self-interested maximization of preferences would appear to lead to a denial of the need for a distinct concept of law in the explanation and evaluation of social institutions.

Bibliography

- Blackburn, Simon [1998] *Ruling Passions: A Theory of Practical Reason* Oxford: Clarendon Press.
- Brennan, Geoffrey and and Alan Hamlin [2000] *Democratic Devices and Desires* Cambridge: Cambridge University Press.
- Brennan, Geoffrey and James Buchanan [1981] The Normative Purpose of Economic ’Science’" Rediscovery of an Eighteenth Century Method," 1 *International Review of Law and Economics* 155, 158
- Brennan, Geoffrey and James Buchanan [1985] *The Reason of Rules* New York: Cambridge University Press
- Calabresi, Guido [1961] Some Thoughts on Risk Distribution and the Law of Torts, 70 Yale L.J. 499
- Coase, Ronald [1961] "The Problem of Social Cost," 3 *Journal of Law and Economics* 1.
- Commons, John R. [1924] *Legal Foundations of Capitalism* New York: MacMillan
- Gibbard, Allan [1990] *Wise Choices, Apt Feelings: A Theory of Normative Judgment* Cambridge: Harvard University Press
- Green, Leslie [1985] "Authority and Convention," 35 *Philosophical Quarterly* 329-46.
- Hale, Robert [1952] *Freedom Through Law: Public Control of Private Governing Power* New York: Columbia University Press
- Hart, H.L.A. [1961] *The Concept of Law* Oxford: Oxford University Press
- Heiner, Ronald [1986] "Imperfect Decisions and the Law: On the evolution of Legal Precedent and Rules", 15 *Journal of Legal Studies* 227.
- Holton, Richard [1998] ’Positivism and the Internal Point of View,’ 17 *Law and Philosophy* 597-625.
- Kaplow, Louis and Steven Shavell [2001] "Fairness versus Welfare," 114 *Harvard Law Review* 961.
- Kornhauser, Lewis A. [1984] "The Great Image of Authority," 36 *Stanford Law Review* 349
- Kornhauser, Lewis A. [1985] "L’Analyse Economique du Droit," numeros 118-9 *La Revue de Synthese* 313.
- Kornhauser, Lewis A. [forthcoming 2000] "Three Roles for a Theory of behavior in a theory of

law," *Rechtstheorie*

- Kornhauser, Lewis A. [2001] "Virtue and Self-Interest in the Design of Constitutional Institutions,"
- Kornhauser, Lewis A. [1999] "The Normativity of Law," 1 *American Law and Economics Review* 3
- Kornhauser, Lewis A.[2001] "A Weaved-up Folly? Preference, Well-being and Morality in Social Decisions" mimeo
- Kornhauser, Lewis A. and Lawrence G. Sager [1986] "Unpacking the Court," 96 *Yale Law Journal* 82
- Landes, William and Richard A. Posner [1975] "The Independent Judiciary in an Interest Group Perspective" 4 *Journal of Legal Studies*
- Posner, Richard A. [1973] *Economic Analysis of Law* Boston: Little Brown (1st edition)
- Posner, Richard A. [1979] "Utilitarianism, Economics and Legal Theory," 8 *Journal of Legal Studies* 103-140.
- Posner, Richard A. [1980] "The Ethical and Political Basis of the Efficiency Norm in Common Law Adjudication," 8 *Hofstra Law Review* 487-598.
- Posner, Richard A. [1985] "Wealth Maximization Revisited," 2 *Notre Dame Journal of Law, Ethics, and Public Policy* 85-106.
- Posner, Richard A. [1990] *The Problems of Jurisprudence* (Cambridge MA: Harvard University Press.
- Posner, Richard A. [1995] "Wealth Maximization and Tort Law: A Philosophical Inquiry," in David G. Owen (ed.), *Philosophical Foundations of Tort Law* (Oxford: Clarendon Press)
- Postema, Gerald, "Coordination and Convention at the Foundations of Law," 11 *Journal of Legal Studies* 165 -202
- Rasmusen, Eric [1994] "Judicial legitimacy as a repeated game," 10 *Journal of Law, Economics, & Organization* 63-83
- Raz, Joseph [1994] "Authority, Law, and Morality," in Joseph Raz, *Ethics in the Public Domain* Oxford: Oxford University Press
- Symposium [1980], Symposium on Efficiency in the Law, 8 Hofstra Law Review

Other Internet Resources

- American Law and Economics Association
- Encyclopedia of Law and Economics, edited by Boudewijn Bouckaert (University of Ghent) and Gerrit De Geest (University of Ghent and Utrecht University)
- ‘Philosophy of Law’, by Kenneth Einar Himma (U. Washington), in the *Internet Encyclopedia of Philosophy*
- Law and Economic Resources (maintained by FindLaw.com, in Mountain View, CA)

Related Entries

consequentialism | game theory | game theory: and ethics | legal obligation and authority | legal philosophy | nature of law | nature of law: interpretivist theories | well being

Acknowledgements

Research for this entry was supported by the Filomen d'Agostino and Max E. Greenberg Research Fund of New York University School of Law.

Lewis Kornhauser
lewis.kornhauser@nyu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 26, 2001
Content last modified: November 26, 2001

Stanford Encyclopedia of Philosophy
Notes to Economic Analysis of the Law

Notes

1. I have phrased (*V*) in Dworkinian terms. Many economic analyses of specific doctrines argue that the economic interpretation fits the doctrinal development; many also contend that efficiency is valuable. I should note, however, that "fit" here has an odd sense. The interpretive claim does not assert that the legal rules in fact are efficient. Rather, it asserts that the rules are efficient within the context of some model that one might impute to judges. The imputation, however, does not rest on an interpretation of judicial opinions.

2. Economic analyses of law include both theoretical and empirical studies. One might thus differentiate strands in terms of the nature of the theory used - game theory, simple price theory, or behavioral economics (which abandons one or more of the rationality postulates of neoclassical economic theory). Alternatively, one might differentiate strands in terms of the aim of enterprise: policy analysis vs. explanation vs prediction.

3. In many articles that explicitly address policy issues, the analyst employs cost benefit analysis. Cost-benefit analysis is often regarded as an implementation of some welfarist social objective function. For criticism of this assumption, see Kornhauser [1998, 2000].

4. The intellectual geography of political economy is actually more complex. Buchanan is a founder and exemplar of the "Virginia" school of the theory of political institutions, sometimes called "public choice theory." Its evaluative commitments are indeed contractarian. The Rochester school of the theory of political institutions, sometimes called "positive political theory," however, has not in general rejected welfarism. Both schools apply economic techniques to the analysis of political institutions and both schools have influenced the economic analysis. (One might differentiate them crudely in the following way. Public choice theory fixes political institutions and asks how the distribution of preferences within the polity affects outcomes. Positive political theory, by contrast, fixes the distribution of preferences within the polity and asks how institutional design affects outcomes.

5. Political economic accounts of adjudication will also reject a purposive account of adjudication. Even if a judge pursues a specific purpose, no judge has complete control over the development of the law. Consequently, the legal rules that evolve are unlikely to be efficient or to further the aims of a specific individual. Nor can a particular aim be attributed to the courts as a whole.

6. Institutional instrumentalism is somewhat at odds with a contractarian evaluative position. Note that one might identify a third type of legal instrumentalism, *systemic instrumentalism*, in which the legal

system as a whole promotes some goal. One might understand Durkheim's theory of law in this manner. On legal instrumentalism generally see Kornhauser [2000].

7. Some models do study the consequences of uncertainty over the legal rule and some of these models interpret this uncertainty as judicial uncertainty. The formalism, however, admits equally the interpretation that the private agent is uncertain how the court will respond to her action.

8. Obligation might enter into an economic account of behavior as a concern that the agent integrates into her preference ordering. Breach of an obligation would impose a "psychic cost" that she weighs with other costs and benefits in her decision. Though under this approach obligations motivate, it is clearly ad hoc and unsatisfactory. For further comments see Kornhauser [2000].

9. Public officials must also be infallible in the sense that they always correctly understand and implement the obligations defined by the legal order.

10. In the standard exposition of this game, there are only two players so that identification of the highest and lowest ranked alternatives gives a complete ranking. The text suggests a game with more than two players. To fill out the ranking define $n(L)$ as the number of people who drive on the left and $n(R)$ as the number of people who drive on the right. Let $m = \min\{n(L), n(R)\}$ and $M = \max\{n(L), n(R)\}$. Let $r = m/M$. The ratio r is defined for each possible choice of strategy vector. Consider two strategy vectors s and s′ with r and r′ the associated ratios. Then each agent prefers s to s′ if and only if $r < r'$.

11. That is, the rule specifies an action that is not necessarily a best response, under perfect information and costless deliberation, to the specific situation. The rule might be an optimal response to the class of situations, given costly deliberation and imperfect information.

12. See Raz [1994]. Raz, at times, seems to ground legal authority in convention; such a grounding would provide a third, economic account of authority as the analysis of convention rests on self-interest. This argument, too, fails to provide an account of authority that treats legal rules as exclusionary reasons. For a clear exposition of this argument see Green [1985].

13. In accident situations where the agents are symmetrically placed, each taking due care may be in the interests of both but, in the non-cooperative solution to this strategic situation, each would, in the absence of a legal rule, adopt a suboptimal level of care. In the prudential account, by contrast, it is in each agent's interest to follow the rule because the costs of identifying the better action exceed the gains from such efforts.

14. Heiner [1986]. This account assumes that each judge has an interest in "correctly" deciding cases. It thus assumes away the first difficulty. Rasmusen [1994] provides an equilibrium account (see below) of *stare decisis*, on the other hand, that relies only on the interest of each individual judge. Rasmusen's theory of *stare decisis*, though it grounds the practice in judicial "self-interest" (understood as the desire

to influence policy), does not yield a rule that functions as an exclusionary reason to follow the law.

15. Constitutional political economy generally argues that, even at the design phase, agents act self-interestedly. Impartiality arises not from a shift in motivation but from the difference in the environment so that self-interest coincides with a more impartial perspective. See generally Brennan and Buchanan [1981, 1985] for this argument. Brennan has subsequently rejected his earlier views in Brennan and Hamlin [2000]. For a critique, see Kornhauser [2001].

16. *Z* is pareto preferred to *Y* if and only no one in the population prefers *Y* to *Z* and at least one person prefers *Z* to *Y*. Technically, the Kaldor-Hicks criterion is not a purely welfarist critierion as it requires reference to the underlying distribution of goods and not simply the underlying distribution of well-being.

17. So if agent *K* has wealth w_K, cost benefit analysis assigns a number [-, w_K] where negative numbers indicate the amount one must pay *K* in order for her to accept that state of affairs.

18. My critique of cost-benefit analysis here is thus related to but different from the standard critique that cost-benefit analyses depend on the underlying distribution of income. That critique contends that an unjust distribution would undermine the moral legitimacy of the cost benefit analysis. The argument in the text suggests that, even if the underlying distribution of income and wealth were morally acceptable, we would require additional argument to establish that the representation of preference derived by willingness to pay would yield interpersonally comparable measures of well-being.

19. They fare no better as interpretive theories of adjudication within civil law countries. Economic analysis of law, however, has concentrated predominantly on the study of common law systems of adjudication.

20. Kaplow and Shavell [2001] offers a defense of welfarism consistent with the characterization in the text. I provide a more comprehensive discussion and critique of Kaplow and Shavell's argument in Kornhauser [2001].

21. The distinction between judgment and preference is elaborated in Kornhauser and Sager [1986] and Kornhauser [2001].

22. Henry contradicts Liza's assertion that she prefers to be a jazz musician, Liza may have reason to reconsider her expression of preference if Henry is a close acquaintance of hers. If Henry is a stranger, however, Liza may regard his assertion as an unwelcome intrusion in her affairs.

23. This claim does not rely on a cognitive view of morality. Even expressivists such as Gibbard [1990] and Blackburn [1998] acknowledge that moral assertions are not personal and that the individual has no sovereignty over them. For a fuller discussion, see Kornhauser [2001].

24. In some contexts of course majority rule is used to resolve conflict in belief but we defend this aggregation method in a different way. Compare the defense of majority rule in jury decisions to its defense in interest group theories of politics.

Lewis Kornhauser
lewis.kornhauser@nyu.edu

First published: November 26, 2001
Content last modified: November 26, 2001

On The Nature of Law

Lawyers are typically interested in the question: What is *the law* on a particular issue? This is always a local question and answers to it are bound to differ according to the specific jurisdiction in which they are asked. In contrast, the philosophy of law is interested in the general question: What is Law? This general question about the nature of law presupposes that law is a unique social-political phenomenon, with more or less universal characteristics that can be discerned through philosophical analysis. General jurisprudence, as this philosophical inquiry about the nature of law is called, is meant to be universal. It assumes that law possesses certain features, and it possesses them by its very nature, or essence, as law, whenever and wherever it happens to exist. However, even if there are such universal characteristics of law, the reasons for a philosophical interest in elucidating them remain to be explained. First, there is the sheer intellectual interest in understanding such a complex social phenomenon which is, after all, one of the most intricate aspects of human culture. Law, however, is also a normative social practice: it purports to guide human behavior, giving rise to reasons for action. An attempt to explain this normative, reason-giving aspect of law is one of the main challenges of general jurisprudence. These two sources of interest in the nature of law are closely linked. Law is not the only normative domain in our culture; morality, religion, social conventions, etiquette, and so on, also guide human conduct in many ways which are similar to law. Therefore, part of what is involved in the understanding of the nature of law consists in an explanation of how law differs from these similar normative domains, how it interacts with them, and whether its intelligibility depends on such other normative orders, like morality or social conventions.

Contemporary legal theories define these two main interests in the nature of law in the following terms. First, we need to understand the general conditions which would render any putative norm legally valid. Is it, for example, just a matter of the source of the norm, such as its enactment by a particular political institution, or is it also a matter of the norm's content? This is the general question about the conditions of legal validity. Second, there is the interest in the normative aspect of law. This philosophical interest is twofold: A complete philosophical account of the normativity of law comprises both an explanatory and a normative-justificatory task. The explanatory task consists of an attempt to explain how legal norms can give rise to reasons for action, and what kinds of reasons are involved. The task of justification concerns the elucidation of the reasons people *ought* to have for acknowledging law's normative aspect. In other words, it is the attempt to explain the moral legitimacy of law. A theory about the nature of law, as opposed to critical theories of law, concentrates on the first of these two questions. It purports to explain what the normativity of law actually consists in.

Thus, elucidating the conditions of legal validity and explaining the normativity of law form the two main subjects of any general theory about the nature of law. In the course of the last few centuries, two

main rival philosophical traditions have emerged, providing different answers to these questions. The older one, dating back to late mediaeval Christian scholarship, is called the natural law tradition. Since the early 19th century, Natural Law theories have been fiercely challenged by the legal positivism tradition promulgated by such scholars as Jeremy Bentham and John Austin. The philosophical origins of Legal Positivism are much earlier, though, probably in the political philosophy of Thomas Hobbes. The main controversy between these two traditions concerns the conditions of legal validity. Basically, Legal Positivism asserts, and Natural Law denies, that the conditions of legal validity are purely a matter of social facts. In contrast to Positivism, Natural Law claims that the conditions of legal validity are not exhausted by social facts; the moral content of the putative norms also bears on their legal validity. As the famous dictum of Saint Augustine has it: '*lex iniusta non est lex*' (unjust law is not law).

The Conditions of Legal Validity

The main insight of Legal Positivism, that the conditions of legal validity are determined by social facts, involves two separate claims which have been labeled The Social Thesis and The Separation Thesis. The Social Thesis asserts that law is, profoundly, a social phenomenon, and that the conditions of legal validity consist of social facts. Early Legal Positivists followed Hobbes' insight that the law is, essentially, an instrument of political sovereignty, and they maintained that the basic source of legal validity resides in the facts constituting political sovereignty. Law, they thought, is basically the command of the sovereign. Later legal Positivists have modified this view, maintaining that social conventions, and not the facts about sovereignty, constitute the grounds of law. Most contemporary legal Positivists share the view that there are conventional rules of recognition, namely, social conventions which determine certain facts or events that provide the ways for the creation, modification, and annulment of legal standards. These facts, such as an act of legislation or a judicial decision, are the *sources of law* conventionally identified as such in each and every modern legal system.

Natural lawyers deny this insight, insisting that a putative norm cannot become legally valid unless it passes a certain threshold of morality. Positive law must confirm in its content to some basic precepts of Natural Law, that is, universal morality, in order to become law in the first place. In other words, Natural Lawyers maintain that the moral content of norms, and not just their social origins, also form part of the conditions of legal validity.

The Separation Thesis is an important negative implication of this Social Thesis, maintaining that there is

a conceptual separation between law and morality, that is, between what the law is, and what the law ought to be. The Separation Thesis, however, has often been overstated. It is sometimes thought that Natural Law asserts, and Legal Positivism denies, that the law is, by necessity, morally good or that the law must have some minimal moral content. The Social Thesis certainly does not entail the falsehood of the assumption that there is something necessarily good in the law. Legal Positivism can accept the claim that law is, by its very nature or its essential functions in society, something good that deserves our moral appreciation. Nor is Legal Positivism forced to deny the plausible claim that wherever law exists, it would have to have a great many prescriptions which coincide with morality. There is probably a considerable overlap, and perhaps necessarily so, between the actual content of law and morality. Once again, the Separation Thesis, properly understood, pertains only to the conditions of legal validity. It asserts that the conditions of legal validity do not depend on the moral content of the norms in question. What the law is cannot depend on what it ought to be in the relevant circumstances.

Many contemporary legal Positivists would not subscribe to this formulation of the Separation Thesis. A contemporary school of thought, called Inclusive Legal Positivism, endorses the Social Thesis, namely, that the basic conditions of legal validity derive from social facts, such as social rules or conventions which happen to prevail in a given community. But, Inclusive Legal Positivists maintain, legal validity is sometimes a matter of the moral content of the norms, depending on the particular conventions that happen to prevail in any given community. Those social conventions on the basis of which we identify the law may, but need not, contain reference to moral content as a condition of legality.

The NaturaLaw tradition has undergone a considerable refinement in the 20th century, mainly because its classical, popular version faced an obvious objection about its core insight: Basically, it is just difficult to maintain that morally bad law is not law. The idea that law must pass, as it were, a kind of moral filter in order to count as law strikes most jurists as incompatible with the legal world as we know it. Therefore, contemporary Natural Lawyers have suggested different and more subtle interpretations of the main tenets of Natural Law. For example, John Finnis views Natural Law (in its Thomist version) not as a constraint on the legal validity of positive laws, but mainly as an elucidation of an ideal of law in its fullest, or highest sense, concentrating on the ways in which law necessarily promotes the common good. As we have noted earlier, however, it is not clear that such a view about the necessary moral content of law is at odds with the main tenets of Legal Positivism.

The idea that the conditions of legal validity are at least partly a matter of the moral content of the norms is articulated in a sophisticated manner by Ronald Dworkin's legal theory. Dworkin is not a classical Natural Lawyer, however, and he does not maintain that morally acceptable content is a precondition of a norm's legality. His core idea is that the very distinction between facts and values in the legal domain, between what the law is and what it ought to be, is much more blurred than Legal Positivism would have it: Determining what the law is in particular cases inevitably depends on moral-political considerations about what it ought to be. Evaluative judgments partly determine what the law is.

Dworkin's legal theory is not based on a general repudiation of the classical fact-value distinction, as much as it is based on a certain conception of legal reasoning. This conception went through two main stages. In the 1970's Dworkin argued that the falsehood of Legal Positivism resides in the fact that it is

incapable of accounting for the important role that legal principles play in determining what the law is. Legal positivism envisaged, Dworkin claimed, that the law consists of rules only. However, this is a serious mistake, since in addition to rules, law is partly determined by legal principles. The distinction between rules and principles is basically a logical one. Rules, Dworkin maintained, apply in an 'all or nothing fashion'. If the rule applies to the circumstances, it determines a particular legal outcome. If it does not apply, it is simply irrelevant to the outcome. On the other hand, principles do not determine an outcome even if they clearly apply to the pertinent circumstances. Principles basically provide the judges with a reason to decide the case one way or the other, and hence they only have a dimension of weight. That is, the reasons provided by the principle may be relatively strong, or weak, but they are never 'absolute'. Such reasons, by themselves, cannot determine an outcome, as rules do.

The most interesting, and from a Positivist perspective, most problematic, aspect of legal principles, however, consists in their moral dimension. According to Dworkin's theory, unlike legal rules, which may or may not have something to do with morality, principles are essentially moral in their content. It is, in fact, partly a moral consideration which determines whether a legal principle exists or not. Why is that? Because a legal principle exists, according to Dworkin, if the principle follows from the best moral and political interpretation of past judicial and legislative decisions in the relevant domain. In other words, legal principles occupy an intermediary space between legal rules and moral principles. Legal rules are posited by recognized institutions and their validity derives from their enacted source. Moral principles are what they are due to their content, and their validity is purely content dependent. Legal principles, on the other hand, gain their validity from *a combination* of source-based and content-based considerations. As Dworkin put it in the most general terms: 'According to law as integrity, propositions of law are true if they figure in or follow from the principles of justice, fairness, and procedural due process that provide the best constructive interpretation of the community's legal practice.' (*Law's Empire*, at p. 225) The validity of a legal principle then, derives, from a combination of facts and moral considerations. The facts concern the past legal decisions which have taken place in the relevant domain, and the considerations of morals and politics concern the ways in which those past decisions can best be accounted for by the correct moral principles.

Needless to say, if such an account of legal principles is correct, the separation thesis can no longer be maintained. But many legal philosophers doubt that there are legal principles of the kind Dworkin envisaged. There is an alternative, more natural way to account for the distinction between rules and principles in the law: the relevant difference concerns the level of generality, or vagueness, of the norm-act prescribed by the pertinent legal norm. Legal norms can be more or less general, or vague, in their definition of the norm-act prescribed by the rule, and the more general or vague they are, the more they tend to have those quasi-logical features Dworkin attributes to principles.

In the 1980's Dworkin radicalized his views about these issues, striving to ground his anti-positivist legal theory on a general theory of interpretation, and emphasizing law's profound interpretative nature. Despite the fact that Dworkin's interpretative theory of law is extremely sophisticated and complex, the essence of his argument from interpretation can be summarized in a rather simple way. The main argument consists of two main premises. The first thesis maintains that determining what the law requires in each and every particular case necessarily involves an interpretative reasoning. Any statement

of the form "According to the law in S, *x* has a right/duty etc., to *y*" is a *conclusion* of some interpretation or other. Now, according to the second premise, interpretation always involves evaluative considerations. More precisely, perhaps, interpretation is neither purely a matter of determining facts, nor is it a matter of evaluative judgment *per se*, but an inseparable mixture of both. Clearly enough, one who accepts both these theses must conclude that the separation thesis is fundamentally flawed. If Dworkin is correct about both theses, it surely follows that determining what the law requires *always* involves evaluative considerations.

Both of Dworkin's two theses are highly contestable. Some legal philosophers have denied the first premise, insisting that legal reasoning is not as thoroughly interpretative as Dworkin assumes. Interpretation, according to this view, is an exception to the standard understanding of language and communication, rendered necessary only when the law is, for some reason, unclear. However, in most standard instances, the law can simply be understood, and applied, without the mediation of interpretation. Other legal philosophers denied the second premise, challenging Dworkin's thesis that interpretation is necessarily evaluative.

Dworkin's legal theory shares certain insights with the Inclusive version of Legal Positivism. Note, however, that although both Dworkin and Inclusive Legal Positivists share the view that morality and legal validity are closely related, they differ on the grounds of this relationship. Dworkin maintains that the dependence of legal validity on moral considerations is an *essential* feature of law which basically derives from law's profoundly interpretative nature. Inclusive Positivism, on the other hand, maintains that such a dependence of legal validity on moral considerations is a contingent matter; it does not derive from the nature of law or of legal reasoning as such. Inclusive Positivists claim that moral considerations affect legal validity only in those cases which follow from the social conventions which happen prevail in a given legal system. In other words, the relevance of morality is determined in any given legal system by the contingent content of that society's conventions. As opposed to both these views, traditional, or as it is now called, Exclusive Legal Positivism maintains that a norm is never rendered legally valid in virtue of its moral content. Legal validity, according to this view, is entirely dependent on the conventionally recognized factual sources of law.

It may be worth noting that those legal theories maintaining that legal validity partly depends on moral considerations must also share a certain view about the nature of morality. Namely, they must hold an objective stance with respect to the nature of moral values. Otherwise, if moral values are not objective and legality depends on morality, legality would also be rendered subjective, posing serious problems for the question of how to identify what the law is. Some legal theories, however, do insist on the subjectivity of moral judgements, thus embracing the skeptical conclusions which follow about the nature of law. According to these skeptical theories, law is, indeed, profoundly dependent on morality, but, as these theorists assume that morality is entirely subjective, it only demonstrates how the law is also profoundly subjective, always up for grabs, so to speak. This skeptical approach, fashionable in so called post-modernist literature, crucially depends on a subjectivist theory of values, which is rarely articulated in this literature in any sophisticated way.

The Normativity of Law

Throughout human history the law has been known as a coercive institution, enforcing its practical demands on its subjects by means of threats and violence. This conspicuous feature of law made it very tempting for some philosophers to assume that the normativity of law resides in its coercive aspect. Even within the legal positivist tradition, however, the coercive aspect of the law has given rise to fierce controversies. Early legal Positivists, such as Bentham and Austin, maintained that coercion is an *essential* feature of law, distinguishing it from other normative domains. Legal Positivists in the 20th century have tended to deny this, claiming that coercion is neither essential to law, nor, actually, pivotal to the fulfillment of its functions in society.

There are several issues entangled here, and we should carefully separate them. John Austin famously maintained that each and every legal norm, as such, must comprise a threat backed by sanction. This involves at least two separate claims: In one sense, it can be understood as a thesis about the concept of law, maintaining that what we call ‘law’ can only be those norms which are backed by sanctions of the political sovereign. In a second, though not less problematic sense, the intimate connection between the law and the threat of sanctions is a thesis about the normativity of law. Basically, it is a reductionist thesis, maintaining that the normativity of law consists in the subjects' ability to predict the chances of incurring punishment or evil.

In addition to this particular controversy, there is the further question, concerning the relative importance of sanctions for the ability of law to fulfill its social functions. *Hans Kelsen*, for instance, maintained that the monopolization of violence in society, and the law's ability to impose its demands by violent means, is the most important of law's functions in society. Twentieth century legal Positivists, like H.L.A. Hart and Joseph Raz, deny this, maintaining that the coercive aspect of law is much more marginal than their predecessors assumed. Once again, the controversy here is actually twofold: is coercion *essential* to what the law does? And even if it is not deemed essential, how important it is, compared with the other functions law fulfils in our lives?

Austin's reductionist account of the normativity of law, maintaining that the normative aspect of law simply consists in the subjects' ability to predict sanctions, was discussed extensively, and fiercely criticized, by H.L.A. Hart. Hart's fundamental objection to Austin's reductionist account of law's normativity is, on his own account, ‘that the predictive interpretation obscures the fact that, where rules exist, deviations from them are not merely grounds for prediction that hostile reactions will follow.... but are also a reason or justification for such reaction and for applying the sanctions.’ (*The Concept of Law*, at p. 82) This emphasis on the reason-giving function of rules is surely correct, but perhaps not enough. Supporters of the predictive account could claim that it only begs the further question of *why* people should regard the rules of law as reasons or justifications for actions. If it is, for example, *only* because the law happens to be an efficient sanction-provider, then the predictive model of the normativity of law may turn out to be correct after all. In other words, Hart's fundamental objection to the predictive model is actually a result of his vision about the main functions of law in society, holding, contra Austin and Kelsen, that those functions are not exclusively related to the ability of the law to impose sanctions.

It is arguable, however, that law's functions in our culture are more closely related to its coercive aspect than Hart seems to have assumed. Contemporary use of ‘game theory’ in the law tends to show that the rationale of a great variety of legal arrangements can be best explained by the function of law in solving problems of opportunism, like the so called Prisoner's Dilemma situations. In these cases, the law's main role is, indeed, one of providing coercive measures. Be this as it may, we should probably refrain from endorsing Austin's or Kelsen's position that providing sanctions is law's only function in society. Solving recurrent and multiple coordination problems, setting standards for desirable behavior, proclaiming symbolic expressions of communal values, resolving disputes about facts, and such, are important functions which the law serves in our society, and those have very little to do with law's coercive aspect and its sanction-providing functions.

The extent to which law can actually guide behavior by providing its subjects with reasons for action has been questioned by a very influential group of legal scholars in the first half of the 20th century, called the Legal Realism school. American Legal Realists claimed that our ability to predict the outcomes of legal cases on the basis of the rules of law is rather limited. In the more difficult cases which tend to be adjudicated in the appellant courts, legal rules, by themselves, are radically indeterminate as to the outcome of the cases. The Legal Realists thought that lawyers who are interested in the predictive question of what the courts will actually decide in difficult cases need to engage in sociological and psychological research, striving to develop theoretical tools which would enable us to predict legal outcomes. Thus Legal Realism was mainly an attempt to introduce the social sciences into the domain of jurisprudence for predictive purposes. To what extent this scientific project succeeded is a matter of controversy. Be this as it may, Legal Realism paid very little attention to the question of the normativity of law, that is, to the question of how the law does guide behavior in those cases in which it seems to be determinate enough.

A much more promising approach to the normativity of law is found in Joseph Raz's theory of authority, which also shows how such a theory about the normativity of law entails important conclusions with respect to the conditions of legal validity. The basic insight of Raz's argument is that the law is an authoritative social institution. The law, Raz claims, is a *de facto* authority. However, it is also essential to law that it must be held to claim legitimate authority. Any particular legal system may fail, of course, in its fulfillment of this claim. But the law is the kind of institution which necessarily claims to be a legitimate authority.

According to Raz, the essential role of authorities in our practical reasoning is to mediate between the putative subjects of the authority and *the right reasons* which apply to them in the relevant circumstances. An authority is legitimate only if its putative subjects are likely to comply better with the relevant reasons which apply to them by following the authoritative resolution than by trying to figure out or act on those reasons by themselves.

Now, it follows that for something to be able to claim legitimate authority, it must be of *the kind of thing* capable of claiming it, namely, capable of fulfilling such a mediating role. What kinds of things can claim legitimate authority? There are at least two such features necessary for authority-capacity: First, for

something to be able to claim legitimate authority, it must be the case that its directives are *identifiable* as authoritative directives, without the necessity of relying on those same reasons which the authoritative directive replaces. If this condition is not met, namely, if it is impossible to identify the authoritative directive as such without relying on those same reasons the authority was meant to rely on then the authority could not fulfill its essential, mediating role. In short, it could not make the practical difference it is there to make. Note that this argument does not concern the efficacy of authorities. The point is not that unless authoritative directives can be recognized as such, authorities could not function effectively. The argument is based on the rationale of authorities within our practical reasoning. Authorities are there to make a practical difference, and they could not make such a difference unless the authority's directive can be recognized as such without recourse to the reasons it is there to decide upon. In other words, it is nonsensical to have authorities if, to discover what is an authority and what is not, you have to engage in the same reasoning process that reliance on the authority is supposed to replace. Secondly, for something to be able to claim legitimate authority, it must be capable of forming an opinion on how its subjects ought to behave, distinct from the subjects' own reasoning about their reasons for action. In other words, a practical authority, like law, must be basically personal authority, in the sense that there cannot be an authority without an author.

Raz's conception of legal authority provides very strong support for Exclusive Legal Positivism since it requires that the law, *qua* an authoritative resolution, be identifiable on its own terms, that is, without having to rely on those same considerations which the law is there to settle. Therefore a norm is legally valid (i.e. authoritative) only if its validity does not derive from moral or other evaluative considerations about which it is there to settle. Notably, Raz's theory challenges both Dworkin's anti-positivist legal theory, and the Inclusive version of Legal Positivism. This challenge, and the controversies it gave rise to, form one of the main topics discussed in contemporary general jurisprudence.

Explaining the rationale of legal authority, however, is not the only component of a theory about the normativity of law. If we hold the Legal Positivist thesis that law is essentially founded on social conventions, another important question arises here: how can a conventional practice give rise to reasons for action and, in particular, to obligations? Some legal philosophers claimed that conventional rules cannot, by themselves, give rise to obligations. As Leslie Green observed, Hart's ‘view that the fundamental rules [of recognition] are ‘mere conventions’ continues to sit uneasily with any notion of obligation’, and this Green finds troubling, because the rules of recognition point to the ‘sources that judges are *legally* bound to apply.’ (‘The Concept of Law Revisited’, at p. 1697) The debate here is partly about the conventional nature of the rules of recognition, and partly about the ways in which conventions can figure in our reasons for action. According to one influential theory, inspired by David Lewis, conventional rules emerge as solutions to multiply and recurrent coordination problems. If the rules of recognition are, indeed, of such a coordination kind, it is relatively easy to explain how they may give rise to obligations. Coordination conventions would be obligatory if the norm subjects have an obligation to solve the coordination problem which initially gave rise to the emergence of the relevant convention. It is doubtful, however, that coordination conventions are at the foundations of law. In certain respects the law may be more like a structured game, or an artistic genre, which are actually constituted by social conventions. Such constitutive conventions are not explicable as solutions to some pre-existing recurrent coordination problem. The conventional rules constituting the game of chess, for

example, are not there to solve a coordination problem between potential players. Antecedent to the game of chess, there was no particular coordination problem to solve. The conventional rules of chess constitute the game itself as a kind of social activity people would find worthwhile engaging in. The constitutive conventions partly constitute the values inherent in the emergent social practice. Such values, however, are only there for those who care to see them. Constitutive conventions, by themselves, cannot ground an obligation to engage in the practice they constitute.

From a moral point of view, the rules of recognition, by themselves, cannot be regarded as sources of obligation to follow the law. Whether judges, or anybody else, should or should not respect the rules of recognition of a legal system, is basically a moral issue, that can only be resolved by moral arguments (concerning the age old issue of political obligation). And this is more generally so: the existence of a social practice, in itself, does not provide anyone with an obligation to engage in the practice. The rules of recognition only define what the practice is, and they can say nothing on the question of whether one should or should not engage in it. But of course, once one does engage in the practice, playing the judge, as it were, there are *legal* obligations defined by the rules of the game. In other words, there is nothing special in the idea of a *legal* obligation to follow the rules of recognition. The judge in a soccer game is equally obliged to follow the rules of his game, and the fact that the game is conventional poses no difficulty from this, let us say, ‘internal-player's’ perspective. But again, the constitutive rules of soccer cannot settle for anyone the question of whether they should play soccer or not. Similarly, the rules of recognition cannot settle for the judge, or anyone else for that matter, whether they should play by the rules of law, or not. They only tell the judges what the law *is*. Unlike chess or soccer, however, the law may well be a kind of game that people have an obligation to play, as it were. But if there is such an obligation, it must emerge from external, moral, considerations, that is, from a general moral obligation to obey the law. The complex question of whether there is such a general obligation to obey the law, and whether it depends on certain features of the relevant legal system, is extensively discussed in the literature on political obligation. A complete theory about the normativity of law must encompass these moral issues as well.

Bibliography

- Austin, John, *The Province of Jurisprudence Determined*, (first published 1832) (Weidenfeld & Nicolson, London, 1954.)
- Coleman, Jules, ‘Incorporationism, Conventionality, and The Practical Difference Thesis’, *Vol. 4 Legal Theory*, (1998), 381.
- ------, *The Practice of Principle* (Oxford, 2001).
- Dworkin, Ronald, *Taking Rights Seriously* (Duckworth, London, 1977).
- ------, *Law's Empire* (Fontana, 1986).
- Finnis, John, *Natural Law and Natural Rights*, (Oxford, 1980).
- Green, Leslie, ‘The Concept of Law Revisited’, *94 Michigan Law. Rev.*, (1996), 1687.
- Hart, H.L.A., *The Concept of Law* (Oxford, 1961). Second Edition with Postscript , Raz & Bulloch eds., (Oxford, 1994).
- Kelsen, Hans, *General Theory of Law and State* (Wedberg trans.), (New York, Russell & Russell,

1945, 1961).

- Marmor, Andrei, *Interpretation and Legal Theory*, (Oxford, 1992).
- ------, *Positive Law & Objective Values* (Oxford, 2001)
- Raz, Joseph, 'Legal Principles and the Limits of Law' in Cohen (ed.), *Ronald Dworkin and Contemporary Jurisprudence*, (Duckworth 1984), 73.
- ------, *The Authority of Law*, (Oxford, 1979).
- ------, 'Law, Authority and Morality', in his *Ethics In The Public Domain*, (Oxford, 1994), chapter 9.
- Waluchow, Wil, *Inclusive Legal Positivism*, (Oxford, 1994)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Austin, John | Bentham, Jeremy | nature of law: legal positivism | nature of law: natural law theories

Andrei Marmor
andrei_marmor@law.uchicago.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 27, 2001
Content last modified: May 27, 2001

John Austin

John Austin is considered by many to be the creator of the school of analytical jurisprudence, as well as, more specifically, the approach to law known as "legal positivism." Austin's particular command theory of law has been subject to pervasive criticism, but its simplicity gives it an evocative power that cannot be ignored.

1. Life

John Austin's life (1790-1859) was filled with disappointment and unfulfilled expectations. His influential friends (who included Jeremy Bentham, James Mill, John Stuart Mill and Thomas Carlyle) were impressed by his intellect and his conversation, and predicted he would go far. However, in public dealings, Austin's nervous disposition, shaky health, tendency towards melancholy, and perfectionism combined to end quickly careers at the Bar, in academia, and in government service. (Hamburger 1985, 1992)

Austin was born to a Suffolk merchant family, and served briefly in the military before beginning his legal training. He was called to the Bar in 1818, but he took on few cases, and quit the practice of law in 1825. Austin shortly thereafter obtained an appointment to the first Chair of Jurisprudence at the recently established University College London. He prepared for his lectures by study in Bonn, and evidence of the influence of continental legal and political ideas can be found scattered throughout Austin's writings.

Lectures from the course he gave were eventually published in 1832 as "Province of Jurisprudence

Determined." (Austin 1995) However, attendance at his courses was small and getting smaller, and he gave his last lecture in 1833. A short-lived effort to give a similar course of lectures at the Inner Temple met the same result. Austin resigned his University College London Chair in 1835. He later briefly served on the Criminal Law Commission, and as a Royal Commissioner to Malta, but he never found either success or contentment. He did some occasional writing on political themes, but his plans for longer works never came to anything during his lifetime, due apparently to some combination of perfectionism, melancholy, and writer's block. His changing views on moral, political, and legal matters also apparently hindered both the publication of a revised edition of "Province of Jurisprudence Determined," and the completion of a longer project started when his views had been different.

Much of whatever success Austin found during his life, and after, must be attributed to his wife Sarah, for her tireless support, both moral and economic (during the later years of their marriage, they lived primarily off her efforts as a translator and reviewer), and her work to publicize his writings after his death (including the publication of a more complete set of his Lectures on Jurisprudence) (Austin 1873).

While Austin's work was influential in the decades after his death, its impact seemed to subside substantially by the beginning of the twentieth century. A significant portion of Austin's current reputation derives from H.L.A. Hart's use (1958, 1994) of Austin's theory as a foil for the explanation of Hart's own, more nuanced approach to legal theory. In recent decades some theorists have revisited Austin's work, offering new characterizations and defenses of his ideas (e.g., Morison 1982, Rumble 1985).

2. Analytical Jurisprudence and Legal Positivism

Early in his career, Austin came under the influence of Jeremy Bentham, and Bentham's utilitarianism is evident (though with some differences) in the work for which Austin is best known today. On Austin's reading of utilitarianism, Divine will is equated with Utilitarian principles: "utility is the index to the law of God ... . To make a promise which general utility condemns, is an offense against the law of God" (Austin 1873: Lecture VI, p. 307; see also Austin 1995: Lecture II, p. 41). This particular reading of utilitarianism, however, has had little long-term influence, though it seems to have been the part of his work that received the most attention in his own day (Rumble 1995: p. xx). Austin early on shared many of the ideas of the Benthamite philosophical radicals; he was "a strong proponent of modern political economy, a believer in Hartleian metaphysics, and a most enthusiastic Malthusian." (Rumble 1985: pp. 16-17)

Austin's importance to legal theory lies elsewhere -- his theorizing about law was novel at three different levels of generality. First, he was arguably the first writer to approach the theory of law analytically (as contrasted with approaches to law more grounded in history or sociology, or arguments about law which were secondary to more general moral and political theories). Analytical jurisprudence emphasizes the analysis of key concepts, including "law," "(legal) right," "(legal) duty," and "legal validity." Though analytical jurisprudence has been challenged by some in recent years (e.g., Leiter 1998), it remains the dominant approach to discussing the nature of law. Analytical jurisprudence, an approach to theorizing about law, has sometimes been confused with what the American legal realists (an influential group of

theorists prominent in the early decades of the 20th century) called "legal formalism" -- a narrow approach to how judges should decide cases. The American legal realists saw Austin in particular, and analytical jurisprudence in general, as their opponents in their critical and reform-minded efforts. In this, the realists were simply mistaken; unfortunately, it is a mistake that can still be found in some contemporary legal commentators.

(There is some evidence that Austin's views later in his life may have moved away from analytical jurisprudence towards something more approximating the historical jurisprudence school. (Hamburger 1985: pp. 178-91))

Second, within analytical jurisprudence, Austin was the first systematic exponent of a view of law known as "legal positivism." Most of the important theoretical work on law prior to Austin had treated jurisprudence as though it were merely a branch of moral theory or political theory: asking how should the state govern? (and when were governments legitimate?), and under what circumstances did citizens have an obligation to obey the law? Austin specifically, and legal positivism generally, offered a quite different approach to law: as an object of "scientific" study, dominated neither by prescription nor by moral evaluation. Subtle jurisprudential questions aside, Austin's efforts to treat law systematically gained popularity in the late 19th century among English lawyers who wanted to approach their profession, and their professional training, in a more serious and rigorous manner (Cotterrell 1989: pp. 79-81).

Legal positivism asserts (or assumes) that it is both possible and valuable to have a morally neutral descriptive (or "conceptual" -- though this is not a term Austin used) theory of law. (The main competitor to legal positivism, in Austin's day as in our own, has been natural law theory.) Legal positivism does not deny that moral and political criticism of legal systems are important, but insists that a descriptive or conceptual approach to law is valuable, both on its own terms and as a necessary prelude to criticism.

There were theorists prior to Austin who arguably offered views similar to legal positivism or who at least foreshadowed legal positivism in some way. Among these would be Thomas Hobbes, with his amoral view of laws as the product of Leviathan (Hobbes 1996); David Hume, with his argument for separating "is" and "ought" (which worked as a sharp criticism for some forms of natural law theory, which purported to derive moral truths from statements about human nature) (Hume 2000); and Jeremy Bentham, with his attacks on judicial lawmaking and on those, like Sir William Blackstone, who justified such lawmaking with natural-law-like justifications (Bentham 1970, 1996).

Austin's famous formulation of what could be called the "dogma" of legal positivism is as follows:

> The existence of law is one thing; its merit or demerit is another. Whether it be or be not is one enquiry; whether it be or be not conformable to an assumed standard, is a different enquiry. A law, which actually exists, is a law, though we happen to dislike it, or though it vary from the text, by which we regulate our approbation and disapprobation. (Austin 1995: Lecture V, p. 157)

Third, Austin's version of legal positivism, a "command theory of law" (which will be detailed in the next

section) has also been influential. Austin's theory had similarities with the views developed by Jeremy Bentham, whose theory could also be characterized as a "command theory." However, Austen's work was more influential in this area, because Bentham's jurisprudential writings did not appear in an even-roughly systematic form until well after Austin's work had already been published. (Bentham 1970, 1996; Cotterrell 1989: pp. 52-53)

3. Austin's Views

Austin's basic approach was to ascertain what can be said generally, but still with interest, about all laws. Austin's analysis can be seen as either a paradigm of, or a caricature of, analytical philosophy, in that his discussions are dryly full of distinctions, but are thin in argument. The modern reader is forced to fill in much of the meta-theoretical, justificatory work, as it cannot be found in the text. Where Austin does articulate his methodology and objective, it is a fairly traditional one: he "endeavored to resolve a *law* (taken with the largest signification which can be given to that term *properly*) into the necessary and essential elements of which it is composed." (Austin 1995: Lecture V, p. 117)

As to what is the core nature of law, Austin's answer is that laws ("properly so called") are commands of a sovereign. He clarifies the concept of positive law (that is, man-made law) by analyzing the constituent concepts of his definition, and by distinguishing law from other concepts that are similar:

- "Commands" involve an expressed wish that something be done, and "an evil" to be imposed if that wish is not complied with.
- Rules are general commands (applying generally to a class), as contrasted with specific or individual commands ("drink wine today" or "John Major must drink wine").
- Positive law consisted of those commands laid down by a sovereign (or its agents), to be contrasted to other law- givers, like God's general commands, and the general commands of an employer.
- The "sovereign" was defined as a person (or collection of persons) who receives habitual obedience from the bulk of the population, but who does not habitually obey any other (earthly) person or institution. Austin thought that all independent political societies, by their nature, have a sovereign.
- Positive law should also be contrasted with "laws by a close analogy" (which includes positive morality, laws of honor, international law, customary law, and constitutional law) and "laws by remote analogy" (e.g., the laws of physics).
  (Austin 1995: Lecture I).

Austin also wanted to include within "the province of jurisprudence" certain "exceptions," items which did not fit his criteria but should nonetheless be studied with other "laws properly so called": repealing laws, declarative laws, and "imperfect laws" - laws prescribing action but without sanctions (a concept Austin ascribes to "Roman [law] jurists"). (Austin 1995: Lecture I, p. 36)

In the criteria set out above, Austin succeeded in delimiting law and legal rules from religion, morality,

convention, and custom. However, also excluded from "the province of jurisprudence" were customary law (except to the extent that the sovereign had, directly or indirectly, adopted such customs as law), public international law, and parts of constitutional law. (These exclusions alone would make Austin's theory problematic for most modern readers.)

Within Austin's approach, whether something is or is not "law" depends on which people have done what: the question turns on an empirical investigation, and it is a matter mostly of power, not of morality. Of course, Austin is not arguing that law should not be moral, nor is he implying that it rarely is. Austin is not playing the nihilist or the skeptic. He is merely pointing out that there is much that is law that is not moral, and what makes something law does nothing to guarantee its moral value. "The most pernicious laws, and therefore those which are most opposed to the will of God, have been and are continually enforced as laws by judicial tribunals." (Austin 1995: Lecture V, p. 158).

In contrast to his mentor Bentham, Austin had no objection to judicial lawmaking, which Austin called "highly beneficial and even absolutely necessary." (Austin, 1995: Lecture V, p. 163) Nor did Austin find any difficulty incorporating judicial lawmaking into his command theory: he characterized that form of lawmaking, along with the occasional legal/judicial recognition of customs by judges, as the "tacit commands" of the sovereign, the sovereign's affirming the "orders" by its acquiescence. (Austin 1995: Lecture 1, pp. 35-36).

4. Criticisms

As many readers come to Austin's theory mostly through its criticism by other writers (prominently, that of H.L.A. Hart), the weaknesses of the theory are almost better known than the theory itself:

- In many societies, it is hard to identify a "sovereign" in Austin's sense of the word (a difficulty Austin himself experienced, when he was forced to describe the British "sovereign" awkwardly as the combination of the King, the House of Lords, and all the electors of the House of Commons). Additionally, a focus on a "sovereign" makes it difficult to explain the continuity of legal systems: a new ruler will not come in with the kind of "habit of obedience" that Austen sets as a criterion for a system's rule-maker. However, one could argue (see Harris 1977) that the sovereign is best understood as a constructive metaphor: that law should be viewed as if it reflected the view of a single will (a similar view, that law should be interpreted as if it derived from a single will, can be found in Ronald Dworkin's work (1986)).
- A "command" model seems to fit some aspects of law poorly (e.g., rules which grant powers to officials and to private citizens - of the latter, the rules for making wills, trusts, and contracts are examples), while excluding other matters (e.g., international law) which we are not inclined to exclude in the category "law."
- More generally, it seems more distorting than enlightening to reduce all law to one type. For example, rules that empower people to make wills and contracts perhaps can be re-characterized as part of a long chain of reasoning for eventually imposing a sanction (Austin spoke in this context of the sanction of "nullity") on those who fail to comply with the relevant provisions. However,

such a re-characterization this misses the basic purpose of those sorts of laws - they are arguably about granting power and autonomy, not punishing wrongdoing.
- A theory which portrays law solely in terms of power fails to distinguish rules of terror from forms of governance sufficiently just that they are accepted as legitimate by their own citizens.

(Austin was aware of some of these lines of attack, and had responses ready; it is another matter whether his responses were adequate.) It should also be noted that Austin's work shows a silence on questions of methodology, though this may be forgivable, given the early stage of jurisprudence. As discussed in an earlier section, in many ways, Austin was blazing a new path.

When H.L.A. Hart revived legal positivism in the middle of the 20th century (Hart 1958, 1994), he did it by criticizing and building on Austin's theory: for example, Hart's theory did not try to reduce all laws to one kind of rule, but emphasized the varying types and functions of legal rules; and Hart's theory, grounded partly on the distinction between "obligation" and "being obliged," was built around the fact that some participants within legal systems "accepted" the legal rules as reasons for action, above and beyond the fear of sanctions.

5. A Revisionist View?

Some modern commentators appreciate in Austin elements that were probably not foremost in his mind (or that of his contemporary readers). For example, one occasionally sees Austin portrayed as the first "realist": in contrast both to the theorists that came before Austin and to some modern writers on law, Austin is seen as having a keener sense of the connection of law and power, and the importance of keeping that connection at the forefront of analysis. (cf. Cotterrell 1989: pp. 57-79) When circumstances seem to warrant a more critical, skeptical or cynical approach to law and government, Austin's equation of law and force will be attractive - however distant such a reading may be from Austin's own liberal-utilitarian views at the time of his writing, and his even more conservative political views later in his life. (Hamburger, 1985)

Bibliography

Primary Sources

- Austin, John, *The Province of Jurisprudence Determined*, W. Rumble (ed.), Cambridge: Cambridge University Press, 1995) (first published, 1832)
- -----, *Lectures on Jurisprudence*, or *The Philosophy of Positive Law*, two vols., R. Campbell (ed.), 4th edition, London: John Murray, 1873

Secondary Sources

- Bentham, Jeremy, *An Introduction to the Principles of Morals and Legislation* (J. H. Burns & H.L.A. Hart, eds., Oxford: Oxford University Press, 1996)
- -----, *Of Laws in General* (H.L.A. Hart, ed., London: Athlone Press, 1970)
- Cosgrove, Richard A., *Scholars of the Law: English Jurisprudence from Blackstone to Hart*, ch. 4 (New York: New York University Press, 1996)
- Cotterrell, Roger, *The Politics of Jurisprudence: A Critical Introduction to Legal Philosophy*, ch. 3 (London: Butterworths, 1989)
- Dworkin, Ronald, *Law's Empire* (Cambridge, Mass.: Harvard University Press, 1986)
- Hamburger, Lotte & Joseph, *Troubled Lives: John and Sarah Austin* (Toronto: University of Toronto Press, 1985)
- -----, *Contemplating Adultery: The Secret Life of a Victorian Woman* (London: Macmillan, 1992)
- Harris, J.W., "The Concept of Sovereign Will," *Acta Juridica*, 1977, pp. 1 ff
- Hart, H.L.A., *The Concept of Law*, 2nd edition (Oxford: Clarendon Press, 1994)
- -----, "Positivism and the Separation of Law and Morals," *Harvard Law Review*, 1958, vol. 71, pp. 593 ff
- Hobbes, Thomas, *Leviathan* (Richard Tuck, ed., Cambridge: Cambridge University Press, 1996) (first published, 1651)
- Hume, David, *A Treatise of Human Nature* (David Fate Norton & Mary J. Norton, eds., Oxford: Oxford University Press, 2000) (first published, 1739)
- Leiter, Brian, "Realism, Hard Positivism, and Conceptual Analysis," *Legal Theory*, 1998, vol. 4, pp. 533-47
- Mill, John Stuart, "Austin on Jurisprudence," *Edinburgh Review*, vol. 118 (Oct. 1863), pp. 438-82
- Moles, Robert N., *Definition and Rule in Legal Theory: A Reassessment of H.L.A. Hart and the Positivist Tradition* (Oxford: Basil Blackwell, 1987)
- Morison, W. L., *John Austin* (Stanford: Stanford University Press, 1982)
- Rumble, W. E., "Introduction," in Austin (1995), pp. vii-xxiv
- -----, *The Thought of John Austin: Jurisprudence, Colonial Reform, and the British Constitution* (London: Athlone Press, 1985)
- Tapper, Colin, "Austin on Sanctions," *Cambridge Law Journal*, 1965, pp. 271-87

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

legal philosophy | nature of law | nature of law: legal positivism

Brian Bix
bixxx002@umn.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: February 23, 2001
Content last modified: February 23, 2001

Well Being

Well-being is most commonly used in philosophy to describe what is non-instrumentally or ultimately good *for* a person. The question of what well-being consists in is of independent interest, but it is of great importance in moral philosophy, especially in the case of utilitarianism, according to which well-being is to be maximized. Signficant challenges to the very notion have been made, in particular by G.E. Moore and T.M. Scanlon. It has become standard to distinguish theories of well-being as either hedonist theories, desire theories, or objective list theories. According to the view known as welfarism, well-being is the only value. Also important in ethics is the question of how a person's moral character relates to their well-being.

1. The Concept

Popular use of the term 'well-being' usually relates to health. A doctor's surgery may run a 'Women's Well-being Clinic', for example. Philosophical use is broader, but related, and amounts to the notion of how well a person's life is going for that person. A person's well-being is what is 'good for' them. Health, then, might be said to be a constituent of my well-being, but it is not plausibly taken to be all that matters for my well-being. One correlate term worth noting here is 'self-interest': my self-interest is what

is in the interest of myself, and not others.

The philosophical use of the term also tends to encompass the ‘negative’ aspects of how a person’s life goes for them. So we may speak of the well-being of someone who is, and will remain in, the most terrible agony: their well-being is negative, and such that their life is worse for them than no life at all. The same is true of closely allied terms, such as ‘welfare’, which covers how a person is faring as a whole, whether well or badly, or ‘happiness’, which can be understood -- as it was by the classical utilitarians from Jeremy Bentham onwards, for example -- to be the balance between good and bad things in a person’s life. But note that philosophers also use such terms in the more standard ‘positive’ way, speaking of ‘ill-being’, ‘ill-faring’, or, of course, ‘unhappiness’ to capture the negative aspects of individuals’ lives.

‘Happiness’ is often used, in ordinary life, to refer to a short-lived state of a person, frequently a feeling of contentment: ‘You look happy today’; ‘I’m very happy for you’. Philosophically, its scope is more often wider, encompassing a whole life. And in philosophy it is possible to speak of the happiness of a person’s life, or of their happy life, even if that person was in fact usually pretty miserable. The point is that some good things in their life made it a happy one, even though they lacked contentment.

When discussing the notion of what makes life good for the individual living that life, it is preferable to use the term ‘well-being’ instead of ‘happiness’. For we want at least to allow conceptual space for the possibility that, for example, the life of a plant may be ‘good for’ that plant. And speaking of the happiness of a plant would be stretching language too far. (An alternative here might be ‘flourishing’, though this might be taken to bias the analysis of human well-being in the direction of some kind of natural teleology.) In that respect, the Greek word commonly translated ‘happiness’ (*eudaimonia*) might be thought to be superior. But, in fact, *eudaimonia* seems not only to have been restricted to conscious beings, but to human beings: animals cannot be *eudaimon*. This is because *eudaimonia* suggests that the gods, or fortune, have favoured one, and the idea that the gods could care about non-humans would not have occurred to most Greeks.

It is occasionally claimed that certain ancient ethical theories, such as Aristotle’s, result in the collapse of the very notion of well-being. On Aristotle’s view, if you are my friend, then my well-being is closely bound up with yours. It might be tempting, then, to say that ‘your’ well-being is ‘part’ of mine, in which case the distinction between what is good for me and what is good for others has broken down. But this temptation should be resisted. Your well-being concerns how well your life goes for you, and we can allow that my well-being depends on yours without introducing the confusing notion that my well-being is constituted by yours. There are signs in Aristotelian thought of an expansion of the subject or owner of well-being. A friend is ‘another self’, so that what benefits my friend benefits me. But this should be taken either as a metaphorical expression of the dependence claim, or as an identity claim which does not threaten the notion of well-being: if you really are the same person as I am, then of course what is good for you will be what is good for me, since there is no longer any metaphysically significant distinction between you and me.

Well-being is a kind of value, sometimes called ‘prudential value’, to be distinguished from, for example,

aesthetic value or moral value. What marks it out is the notion of 'good for'. The serenity of a Vermeer painting, for example, is a kind of goodness, but it is not 'good for' the painting. It may be good for us to contemplate such serenity, but contemplating serenity is not the same as the serenity itself. Likewise, my giving money to a development charity may have moral value, that is, be morally good. And the effects of my donation may be good for others. But it remains an open question whether my being morally good is good for me; and, if it is, its being good for me is still conceptually distinct from its being morally good.

2. Moore's Challenge

There is something mysterious about the notion of 'good for'. Consider a possible world that contains only a single item: a stunning Vermeer painting. Leave aside any doubts you might have about whether paintings can be good in a world without viewers, and accept for the sake of argument that this painting has aesthetic value in that world. Now it seems intuitively plausible to claim that the value of this world is constituted solely by the aesthetic value of the painting. But now consider a world which contains one invididual living a life that is good for them. How are to describe the relationship between the value of this world, and the value of the life lived in it for the individual? Are we to say that the world has a value at all? How can it, if the only value it contains is 'good for' as opposed to just 'good'? And yet we surely do want to say that this world is better ('more good') than some other empty world. Well, should we say that the world is good, and is so because of the good it contains 'for' the individual? This fails to capture the idea that there is in fact nothing of value in this world than what is good for the individual.

Thoughts such as these led G.E. Moore to object to the very idea of 'good for' (Moore 1903, pp. 98-9). Moore argued that the idea of 'my own good', which he saw as equivalent to what is 'good for me', makes no sense. When I speak of, say, pleasure as what is good for me, he claimed, I can mean only either that the pleasure I get is good, or that my getting it is good. Nothing is added by saying that the pleasure constitutes my good, or is good for me.

But the distinctions I drew between different categories of value above show that Moore's analysis of my claim that my own good consists in pleasure, is too narrow. Indeed Moore's argument rests on the very assumption that it seeks to prove: that only the notion of 'good' is necessary to make all the evaluative judgements we might wish to make. The claim that it is good that I get pleasure is, logically speaking, equivalent to the claim that the world containing the single Vermeer is good. It is, so to speak, 'impersonal', and leaves out of account the special feature of the value of well-being: that it is good for individuals.

Indeed, one way to respond both to Moore's challenge, and to the puzzles above, is to try, when appropriate, to do without the notion of 'good' and make do with 'good for', alongside the separate and non-evaluative notion of reasons for action. Thus, the world containing the single individual with a life worth living, might be said to contain nothing good *per se*, but a life that is good for that individual. And this fact may give us a reason to bring about such a world, given the opportunity.

3. Scanlon's Challenge

Moore's book was published in Cambridge, England, at the beginning of the twentieth century. At the end of the same century, a book was published in Cambridge, Mass., which also posed some serious challenges to the notion of well-being: *What Do We Owe to Each Other?*, by T.M. Scanlon.

Moore's ultimate aim in criticizing the idea of 'goodness for' was to attack egoism. Likewise, Scanlon has an ulterior motive in objecting to the notion of well-being -- to attack so-called 'telelogical' or end-based theories of ethics, in particular, utilitarianism, which in its standard form requires us to maximize well-being. But in both cases the critiques stand independently.

One immediately odd aspect of Scanlon's position that 'well-being' is an otiose notion in ethics is that he himself seems to have a view on what well-being is. It involves, he believes, among other things, success in one's rational aims, and personal relations. But Scanlon claims that his view is not a 'theory of well-being', since a theory must explain what unifies these different elements, and how they are to be compared. And, he adds, no such theory is ever likely to be available, since such matters depend so much on context.

Scanlon does, however, implicitly make a claim about what unites these values: they are all constituents of well-being, as opposed to other kinds of value, such as aesthetic or moral. Nor is it clear why Scanlon's view of well-being could not be developed so as to assist in making real-life choices between different values in one's own life.

Scanlon suggests that we often make claims about what is good in our lives without referring to the notion of well-being, and indeed that it would often be odd to do so. For example, I might say, 'I listen to Alison Krauss's music because I enjoy it', and that will be sufficient. I do not need to go on to say, 'And enjoyment adds to my well-being'.

But this latter claim sounds peculiar only because we already *know* that enjoyment makes a person's life better for them. And in some circumstances such a claim would anyway not be odd: consider an argument with someone who claims that aesthetic experience is worthless, or with an ascetic. Further, people do use the notion of well-being in practical thinking. For example, if I am given the opportunity to achieve something significant, which will involve considerable discomfort over several years, I may consider whether, from the point of view of my own well-being, the project is worth pursuing.

Scanlon argues also that the notion of well-being, if it is to be philosophically acceptable, ought to provide a 'sphere of compensation' -- a context in which it makes sense to say, for example, that I am losing one good in my life for the sake of gain over my life as a whole. And, he claims, there is no such sphere. For Scanlon, giving up present comfort for the sake of future health 'feels like a sacrifice'.

But this does not chime with my own experience. When I donate blood, this feels to me like a sacrifice. But when I visit the dentist, it feels to me just as if I am weighing up present pains against potential future

pains. And we can weigh up different components of well-being against one another. Consider a case in which you are offered a highly paid job, but many miles away from your friends and family.

Scanlon denies that we need an account of well-being to understand benevolence, since we do not have a general duty of benevolence, but merely duties to benefit others in specific ways, such as to relieve their pain. But, from the philosophical perspective, it may be quite useful to use the heading of ‘benevolence’ in order to group such duties. And, again, comparisons may be important: if I have several *pro tanto* duties of benevolence, not all of which can be fulfilled, I shall have to weigh up the various benefits I can provide against one another. And here the notion of well-being will again come into play.

Further, if morality includes so-called ’imperfect’ duties to benefit others, that is, duties that allow the agent some discretion as to when and how to assist, the lack of any overarching conception of well-being is likely to make the fulfillment of such duties problematic.

4. Theories of Well-being

4.1 Hedonism

On one view, human beings always act in pursuit of what they think will give them the greatest balance of pleasure over pain. This is ‘psychological hedonism’, and will not be my concern here. Rather, I intend to discuss ‘evaluative hedonism’ or ‘prudential hedonism’, according to which well-being consists in the greatest balance of pleasure over pain.

This view was first, and perhaps most famously, expressed by Socrates and Protagoras in the Platonic dialogue, *Protagoras* (Plato 1976 [C4 BCE]: 351b-c). Jeremy Bentham, perhaps the most well-known of the more recent hedonists, begins his *Introduction to the Principles of Morals and Legislation* thus: ‘Nature has placed mankind under the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do’.

In answer to the question, ‘What does well-being consist in?’, then, the hedonist will answer, ‘The greatest balance of pleasure over pain’. We might call this *substantive hedonism*. A complete hedonist position will involve also *formal hedonism*, which consists in an answer to the following question: ‘What *makes* pleasure good, and pain bad?’, the answer being, ‘The pleasantness of pleasure, and the painfulness of pain’. Consider a substantive hedonist who believed that what makes pleasure good for us is that it fulfills our nature. This theorist is not a formal hedonist.

Hedonism -- as is demonstrated by its ancient roots -- has long seemed an obviously plausible view. Well-being, what is good *for* me, might be thought to be naturally linked to what seems good *to* me, and pleasure does, to most people, seem good. And how could anything else benefit me if I did not enjoy it?

The simplest form of hedonism is Bentham’s, according to which the more pleasantness one can pack

into one's life, the better it will be, and the more painfulness one encounters, the worse it will be. How do we measure the value of the two experiences? The two central aspects of the respective experiences, according to Bentham, are their duration, and their intensity.

Bentham tended to think of pleasure and pain as a kind of sensation, as the notion of intensity might suggest. One problem with this kind of hedonism is that there does not appear to be a single common strand of pleasantness running through all the different experiences people enjoy, such as eating hamburgers, reading Shakespeare, or playing waterpolo. Rather, it seems, there are certain experiences we want to continue, and we might be prepared to call these -- for philosophical purposes -- pleasures (even though some of them, such as diving in a very deep and narrow cave, for example, would not normally be described as pleasurable).

But simple hedonism could survive this objection merely by incorporating whatever view of pleasure was thought to be plausible. A more serious objection is to the evaluative stance of hedonism itself. Thomas Carlyle, for example, described the hedonistic component of utilitarianism as the 'philosophy of swine', the point being that simple hedonism places all pleasures on a par, whether they be the lowest animal pleasures of sex or the highest of aesthetic appreciation. One might make this point with a thought experiment. Imagine that you are given the choice of living a very fulfilling human life, or that of a barely sentient oyster, which experiences some very low-level pleasure. Imagine also that the life of the oyster can be as long as you like, whereas the human life will be of eighty years only. If Bentham were right, there would have to be a length of oyster life such that you would choose it in preference to the human. And yet many say that they would choose the human life in preference to an oyster life of any length.

Now this is not a knockdown argument against simple hedonism. Indeed some people are ready to accept that at some length or other the oyster life becomes preferable. But there is an alternative to simple hedonism, outlined famously by J.S. Mill, using his distinction between 'higher' and 'lower' pleasures (1998 [1863], ch. 2). Mill added a third property to the two determinants of value identified by Bentham, duration and intensity. To distinguish it from these two 'quantitative' properties, Mill called his third property 'quality'. The claim is that some pleasures, by their very nature, are more valuable than others. For example, the pleasure of reading Shakespeare, by its very nature, is more valuable than any amount of basic animal pleasure. And we can see this, Mill suggests, if we note that those who have experienced both types, and are 'competent judges', will make their choices on this basis.

A long-standing objection to Mill's move here has been to claim that his position can no longer be described as formally hedonist. If higher pleasures are higher because of their nature, that aspect of their nature cannot be pleasantness, since that could be determined by duration and intensity alone. And Mill anyway speaks of properties such as 'nobility' as adding to the value of a pleasure. Now it has to be admitted that Mill is sailing close to the wind here. But there is logical space for a hedonist position which allows properties such as nobility to determine pleasantness, and insists that only pleasantness determines value. But one might well wonder how nobility could affect pleasantness, and why Mill did not just come out with the idea that nobility is itself a good-making property.

But there is a yet more weighty objection to hedonism of any kind: the so-called 'experience machine'.

Imagine that I have a machine that I could plug you into for the rest of your life. This machine would give you experiences of whatever kind you thought most valuable or enjoyable -- writing a great novel, bringing about world peace, attending an early Rolling Stones' gig. You would not know you were on the machine, and there is no worry about its breaking down or whatever. Would you plug in? Would it be wise, from the point of your own well-being, to do so? Robert Nozick thinks it would be a big mistake to plug in: 'We want to do certain things… we want to be a certain way… plugging into an experience machine limits us to a man-made reality' (Nozick 1974, p. 43).

One can make the machine sound more palatable, by allowing that genuine choices can be made on it, that those plugged in have access to a common 'virtual world' shared by other machine-users, a world in which 'ordinary' communication is possible, and so on. But this will not be enough for many anti-hedonists. A further line of response begins from so-called 'externalism' in the philosophy of mind, according to which the content of mental states is determined by facts external to the experiencer of those states. Thus, the experience of *really* writing a great novel is quite different from that of *apparently* writing a great novel, even though 'from the inside' they may be indistinguishable. But this is once again sailing close to the wind. If the world can affect the very content of my experience without my being in a position to be aware of it, why should it not affect the value of my experience?

The strongest tack for hedonists to take is to accept the apparent force of the experience machine objection, but to insist that it rests on 'common sense' intuitions, the place in our lives of which may itself be justified by hedonism. This is to adopt a strategy similar to that developed by 'two-level utilitarians' in response to alleged counter-examples based on common-sense morality. The hedonist will point out the so-called 'paradox of hedonism', that pleasure is most effectively pursued indirectly. If I consciously try to maximize my own pleasure, I will be unable to immerse myself in those activities, such as reading or playing games, which do give pleasure. And if we believe that those activities are valuable independently of the pleasure we gain from engaging in them, then we shall probably gain more pleasure overall.

These kinds of stand-off in moral philosophy are unfortunate, but should not be brushed aside. They raise questions concerning the epistemology of ethics, and the source and epistemic status of our deepest ethical beliefs, which we are further from answering than many would like to think. Certainly the current trend of quickly dismissing hedonism on the basis of a quick run-through of the experience machine objection is not methodologically sound.

4.2 Desire Theories

The experience machine is one motivation for the adoption of a desire theory. When you are on the machine, many of your central desires are likely to remain unfilled. Take your desire to write a great novel. You may believe that this is what you are doing, but in fact it is just a hallucination. And what you want, the argument goes, is to write a great novel, not the experience of writing a great novel.

Historically, however, the reason for the current dominance of desire theories lies in the emergence of welfare economics. Pleasure and pain are inside people's heads, and also hard to measure -- especially

when we have to start weighing different people's experiences against one another. So economists began to see people's well-being as consisting in the satisfaction of preferences or desires, the content of which could be revealed by their possessors. This made possible the ranking of preferences, the development of 'utility functions' for individuals, and methods for assessing the value of preference-satisfaction (using, for example, money as a standard).

The simplest version of a desire theory one might call the *present desire* theory, according to which someone is made better off to the extent that their current desires are fulfilled. This theory does succeed in avoiding the experience machine objection. But it has serious problems of its own. Consider the case of the *angry adolescent*. This boy's mother tells him he cannot attend a certain nightclub, so the boy holds a gun to his own head, wanting to pull the trigger and retaliate against his mother. Recall that the scope of theories of well-being should be the whole of a life. It is implausible that the boy will make his life go as well as possible by pulling the trigger. We might perhaps interpret the simple desire theory as a theory of well-being-at-at-a-particular-time. But even then it seems unsatisfactory. From whatever perspective, the boy would be better off if he put the gun down.

We should move, then, to a *comprehensive desire* theory, according to which what matters to a person's well-being is the overall level of desire-satisfaction in their life as a whole. A *summative* version of this theory suggests, straightforwardly enough, that the more desire-fulfilment in a life the better. But it runs into Derek Parfit's case of *addiction* (1984, p. 497). Imagine that you can start taking a highly addictive drug, which will cause a very strong desire in you for the drug every morning. Taking the drug will give you no pleasure; but not taking it will cause you quite severe suffering. There will be no problem with the availability of the drug, and it will cost you nothing. But what reason do you have to take it?

A *global* version of the comprehensive theory ranks desires, so that desires about the shape and content of one's life as a whole are given some priority. So, if I prefer not to become a drug addict, that will explain why it is better for me not to take Parfit's drug. But now consider the case of the *orphan monk*. This young man began training to be a monk at the earliest age, and has lived a very sheltered life. He is now offered three choices: he can remain as a monk, or become either a cook or a gardener outside the monastery, at a grange. He has no conception of the latter alternatives, so chooses to remain a monk. But surely it might be possible that he would have a better life were he to live outside?

So we now have to moved to an *informed desire* version of the comprehensive theory. According to the informed desire account, the best life is the one I would desire if I were fully informed about all the (non-evaluative) facts. But now consider a case suggested by John Rawls: the *grass-counter*. Imagine a brilliant Harvard mathematician, fully informed about the options available to her, who develops an overriding desire to count the blades of grass on the lawns of Harvard. Like the experience machine case, this case is another example of philosophical 'bedrock'. Some will believe that, if she really is informed, and not suffering from some neurosis, then the life of grass-counting will be the best for her.

But it does seem that all these problem cases for desire theories are symptoms of a more general difficulty. Recall again the distinction between substantive and formal theories of well-being. The former state the constituents of well-being (such as pleasure), while the latter state what makes these things good

for people (pleasantness, for example). Substantively, a desire theorist and a hedonist may agree on what makes life good for people: pleasurable experiences. But formally they will differ: the hedonist will refer to pleasantness as the good-maker, while the desire theorist must refer to desire-satisfaction. (It is worth pointing out here that if one characterizes pleasure as an experience the subject wants to continue, the distinction between hedonism and desire theories becomes quite hard to pin down.)

The idea that desire-satisfaction is a 'good-making property' is somewhat odd. As Aristotle says (1984 [C4 BCE], *Metaphysics* 1072a, tr. Ross): 'desire is consequent on opinion rather than opinion on desire'. In other words, we desire things, such as writing a great novel, because we think those things are independently good; we do not think they are good because they will satisfy our desire for them.

4.3 Objective List Theories

The threefold distinction I am using between different theories of well-being has become standard in contemporary ethics. There are problems with it, however, as with many classifications, since it can blind one to other ways of characterizing views. Objective list theories are usually understood as theories which list items constituting well-being that consist neither merely in pleasurable experience nor in desire-satisfaction. Such items might include, for example, knowledge or friendship. But it is worth remembering, for example, that hedonism might be seen as one kind of 'list' theory, and all list theories might then be opposed to desire theories as a whole.

What should go on the list? It is important that every good should be included. As Aristotle put it: 'We take what is self-sufficient to be that which on its own makes life worthy of choice and lacking in nothing. We think happiness to be such, and indeed the thing most of all worth choosing, not counted as just one thing among others' (2000 [C4 BCE], *Nicomachean Ethics* 1197b, tr. Crisp). In other words, if you claim that well-being consists only in friendship and pleasure, I can show your list to be unsatisfactory if I can demonstrate that enjoyment or pleasure is also something that makes people better off.

What is the 'good-maker', according to objective list theorists? This depends on the theory. One, influenced by Aristotle and recently developed by Thomas Hurka (1993), is *perfectionism*, according to which what makes things consituents of well-being is their perfecting human nature. If it is part of human nature to acquire knowledge, for example, then a perfectionist should claim that knowledge is a constituent of well-being. But there is nothing to prevent an objective list theorist's claiming that all that the items on her list have in common is that each, in its own way, advances well-being.

How do we decide what goes on the list? All we can work on is the deliverance of reflective judgement -- intuition, if you like. But one should not conclude from this that objective list theorists are, because they are intuitionist, less satisfactory than the other two theories. For those theories too can be based only on reflective judgement. Nor should one think that intuitionism rules out argument. Argument is one way to bring people to see the truth. Further, we should remember that intuitions can be mistaken. Indeed, as suggested above, this is the strongest line of defence available to hedonists: to attempt to undermine the

evidential weight of many of our natural beliefs about what is good for people.

One common objection to objective list theories is that they are élitist, since they appear to be claiming that certain things are good for people, even if those people will not enjoy them, and do not even want them. One strategy here might be to adopt a 'hybrid' account, according to which certain goods do benefit people independently of pleasure and desire-satisfaction, but only when they do in fact bring pleasure and/or satisfy desires. Another would be to bite the bullet, and point out that a theory could be both élitist and true.

It is also worth pointing out that objective list theories need not involve any kind of objectionable authoritarianism or perfectionism. First, one might wish to include autonomy on one's list, claiming that the informed and reflective living of one's own life for oneself itself constitutes a good. Secondly, and perhaps more significantly, one might note that any theory of well-being in itself has no direct moral implications. There is nothing logically to prevent one's holding a highly élitist conception of well-being alongside a strict liberal view that forbade paternalistic interference of any kind with a person's own life (indeed, on some interpretations, J.S. Mill's position is close to this).

One not implausible view, if desire theories are indeed mistaken in their reversal of the relation between desire and what is good, is that the debate is really between hedonism and objective list theories. And, as suggested above, what is most at stake here is the issue of the epistemic adequacy of our beliefs about well-being. The best way to resolve this matter would consist, in large part at least, in returning once again to the experience machine objection, and seeking to discover whether that objection really stands.

5. Well-being and Morality

5.1 Welfarism

Well-being obviously plays a central role in any moral theory. A theory which said that it just does not matter would be given no credence at all. Indeed, it is very tempting to think that well-being, in some ultimate sense, is all that can matter morally. Consider, for example, Joseph Raz's 'humanistic principle': 'the explanation and justification of the goodness or badness of anything derives ultimately from its contribution, actual or possible, to human life and its quality' (Raz 1986, p. 194). If we expand this principle to cover non-human well-being, it might be read as claiming that, ultimately speaking, the justificatory force of any moral reason rests on well-being. This view is *welfarism*.

Act-utilitarians, who believe that the right action is that which maximizes well-being overall, may attempt to use the intuitive plausibility of welfarism to support their position, arguing that any deviation from the maximization of well-being must be grounded on something distinct from well-being, such as equality or rights. But those defending equality may argue that egalitarians are concerned to give priority to those who are worse off, and that we do see here a link with concern for well-being. Likewise, those concerned with rights may note that rights are to certain goods, such as freedom, or the absence of 'bads', such as

suffering (in the case of the right not to be tortured, for example). In other words, the interpretation of welfarism is itself a matter of dispute. But, however it is understood, it does seem that welfarism poses a problem for those who believe that morality can require actions which benefit no one, and harm some, such as, for example, punishments intended to give individuals what they deserve.

5.2 Well-being and Virtue

Ancient ethics was, in a sense, more concerned with well-being than a good deal of modern ethics, the central question for many ancient moral philosophers being, ‘Which life is best for one?’. The rationality of egoism -- the view that my strongest reason is always to advance my own well-being -- was largely assumed. This posed a problem. Morality is naturally thought to concern the interests of others. So if egoism is correct, what reason do I have to be moral?

One obvious strategy to adopt in defence of morality is to claim that a person’s well-being is in some sense constituted by their virtue, or the exercise of virtue, and this strategy was adopted in subtly different ways by the three greatest ancient philosophers, Socrates, Plato, and Aristotle. At one point in his writings, Plato appears to allow for the rationality of moral self-sacrifice: the philosophers in his famous ‘cave’ analogy in the *Republic* (519-20) are required by morality to desist from contemplation of the sun outside the cave, and to descend once again into the cave to govern their fellow citizens. In the voluminous works of Aristotle, however, there is no recommendation of sacrifice. Aristotle believed that he could defend the virtuous choice as always being in the interest of the individual.

His primary argument is his notorious ‘function argument’, according to which the good for some being is to be identified through attention to its ‘function’ or characteristic activity. The characteristic activity of human beings is to exercise reason, and the good will lie in exercising reason well -- that is, in accordance with the virtues. This argument, which is stated by Aristotle very briefly and relies on assumptions from elsewhere in his philosophy and indeed that of Plato, appears to conflate the two ideas of what is good for a person, and what is morally good. I may agree that a ‘good’ example of humanity will be virtuous, but deny that this person is doing what is best for them. Rather, I may insist, reason requires one to advance one’s own good, and this good consists in, for example, pleasure, power, or honour. But much of Aristotle’s *Nicomachean Ethics* is taken up with portraits of the life of the virtuous and the vicious, which supply independent support for the claim that well-being is constituted by virtue. In particular, it is worth noting the emphasis placed by Aristotle on the value to a person of ‘nobility’ (*to kalon*), a quasi-aesthetic value which those sensitive to such qualities might not implausibly see as a constituent of well-being of more worth than any other. In this respect, the good of virtue is, in the Kantian sense, ‘unconditional’. Yet, for Aristotle, virtue or the ‘good will’ is not only morally good, but good for the individual.

Bibliography

Two significant recent works are Griffin (1986), which presents an objective list theory, and Sumner (1996), which rejects many current options and advocates a theory of well-being based on the idea of ‘life-satisfaction’. A collection of useful essays is Nussbaum and Sen (1993).

- Aristotle (1984 [C4 BCE]) *The Complete Works of Aristotle*, ed. J. Barnes (Princeton: Princeton University Press).
- Aristotle (2000 [C4 BCE]) *Nicomachean Ethics*, ed. and introd. R. Crisp (Cambridge: Cambridge University Press
- Bentham, J. (1996 [1789]) *An Introduction to the Principles of Morals and Legislation*, ed. J. Burns and H.L. A. Hart, introd. F. Rosen (Oxford: Clarendon Press).
- Griffin, J. (1986) *Well-being* (Oxford: Clarendon Press).
- Hurka, T. (1993) *Perfectionism* (Oxford: Clarendon Press).
- Mill, J.S. (1998 [1863]), *Utilitarianism*, ed. R. Crisp (Oxford: Oxford University Press).
- Moore, G.E. (1903), *Principia Ethica* (Cambridge: Cambridge University Press).
- Nozick, R. (1974) *Anarchy, State, and Utopia* (Oxford: Basil Blackwell).
- Nussbaum, M and A. Sen (ed.) (1993) *The Quality of Life*, Oxford: Clarendon Press.
- Parfit, D. (1984) *Reasons and Persons* (Oxford: Clarendon Press).
- Plato (1976 [C4 BCE]) *Protagoras*, ed. and trans. C.C.W. Taylor, Oxford: Clarendon Press.
- Raz, J. (1986), *The Morality of Freedom* (Oxford: Clarendon Press).
- Scanlon, T. (1998) *What Do We Owe to Each Other?*, Harvard: Belknap Press.
- Sumner, W. (1996) *Welfare, Happiness, and Ethics* (Oxford: Clarendon Press).

Other Internet Resources

- Aristotle and Virtue Theory (in *Ethics Updates*, L. Hinman, U. San Diego)
- Utilitarianism (in *Ethics Updates*, L. Hinman, U. San Diego)

Related Entries

Aristotle: ethics | autonomy: in moral and political philosophy | autonomy: personal | consequentialism | egoism | ethics: ancient | hedonism | Mill, John Stuart | Plato: ethics and politics in *The Republic*

Roger Crisp
St. Annes College/Oxford University
roger.crisp@st-annes.ox.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 6, 2001
Content last modified: November 6, 2001

Stanford Encyclopedia of Philosophy

Causation in the law

The basic questions dealt with in this entry are: (i) whether and to what extent causation in legal contexts differs from causation outside the law, for example in science or everyday life, and (ii) what are the appropriate criteria in law for deciding whether one action or event has caused another, (generally harmful) event. The importance of these questions is that responsibility in law very often depends on showing that a specific action or event or state of affairs has caused specific harm or loss to another. Are the criteria adopted in deciding these causal issues both objective and properly attuned to the function of fixing responsibility?

The entry covers the nature and functions of causation, the relation between causation and legal responsibility, and the criteria for the existence of causal connection in law. The last topic is treated in two parts: what are causally relevant conditions (‘causes-in-fact‘) and what are the grounds for limiting responsibility (the ‘proximate cause’ requirement).

1. Nature and Functions of Causation

Law is concerned with the application of causal ideas, embodied in the language of statutes and decisions, to particular situations. This involves, first, a conception of what a cause is outside the law. To this a variety of answers empirical (Hume) and metaphysical (Kant) have been given and each has its contemporary supporters.

Secondly, a theory is required of how causal notions should function in different contexts. In the context of application the notion of cause is a multi-purpose tool. One function, perhaps fundamental, is forward-looking: that of specifying what will happen and by what stages if certain conditions are present together. This use of cause serves to provide recipes and make predictions. It also yields the idea of a causal process. Another function is backward-looking and explanatory: that of showing which earlier conditions

best account for some later event or state of affairs. A third function is attributive: that of fixing the extent of responsibility of agents for the outcomes that follow on their agency or intervention in the world.

For the first of these purposes the emphasis falls on a cause as consisting of the whole complex of conditions required if a certain outcome is to follow (J.S. Mill). Even when applied to a specific situation this involves considering what generally happens when certain conditions are present. In the second, explanatory, context the focus is on selecting from the whole complex the particular condition or conditions that best explain a given outcome. The aim can be either to explain a class of events or a particular event. In the third, attributive, context the aim is again selective, but from a different point of view. It is to attribute responsibility to an agent for those outcomes that his, her or its agency serves to explain and that can therefore plausibly be treated as part of the agency's impact on the world. Here the purpose is to settle the extent of responsibility that attaches to a particular human action or other event or state of affairs. This responsibility is then attributed to an agent or, metaphorically, to the other event or state of affairs in question (e.g. outbreak of war, high unemployment).

In law the second and third of these functions of the notion of cause are prominent, often in combination. Many legal inquiries are concerned to explain how some event or state of affairs came about, especially an untoward event such as death or a state of affairs such as insolvency. But in law the third function is particularly salient and controversial. Whether someone is liable to punishment or to pay compensation or is entitled to claim compensation often depends on showing whether the person potentially liable or entitled has caused harm of a sort that the law seeks to avoid. For example, all systems of law hold that a person can be guilty of homicide only if he or she has caused another's death. All systems treat it as a more serious offence to cause death than to attempt to do so. It is a civil wrong to cause injury to another by negligence in driving a vehicle, but the claim is barred or reduced if the negligent conduct of the person injured is also a cause of the injury. An insurer is required to pay for losses caused by an event of the type defined in the insurance policy, such as fire or flooding, but not if the cause of the loss is something else.

The attribution of responsibility on causal grounds is not confined to law. Historians and moralists, for example, assess the responsibility of agents for the outcomes, political, social, economic or military of what they did or failed to do. Unlike lawyers, they are concerned with responsibility for good as well as bad outcomes. But whereas historians may aim to assess the outcome of an agent's conduct over a period or even a lifetime, lawyers focus on the harmful outcomes of particular actions. These uses of causation by historians, moralists and lawyers raise the question, adumbrated by Collingwood, of whether the attribution of responsibility requires a different conception of cause from that employed for prediction or explanation. In the legal theory of causation this problem is of central importance.

2. Causation and Legal Responsibility

When rules of law attributing responsibility for harm caused are formulated in statutes, regulations and judicial decisions, the word 'cause' is often used. The notion that causal connection between agency and

harm must be established is however often implied even when the word is not used. This is true, for example, of the use of verbs such as ‘damage’, which imply a causal relation between an agency and the harm done. In legal contexts the possible range of agency is not confined to human conduct, but may extend to damage done by the agency of juristic persons, animals, inanimate objects such as motor vehicles and inanimate forces such as fire. In all these instances the use of the notion of cause is central to the legal inquiry, since to establish responsibility it must be shown that the harm was done or brought about by the agency that the law treats as a potential basis for the existence or extent of liability.

The relationship between causing harm and legal responsibility is however complex. The complexities concern the *incidence* of responsibility, the *grounds* of responsibility, the *items* between which causal connection must be demonstrated, and the variety of *relationships* that can in some sense be regarded as causal. So far as the *incidence* of responsibility is concerned, while in law the relevant causes may be human or animal behaviour or natural events or processes, legal responsibility attaches in modern law only to natural persons (human beings) and juristic persons such as states, corporations and other institutions to which personality is ascribed in law.

As regards the *grounds* of responsibility it is important to grasp that for a person to cause harm or loss to another (the term ‘harm’ will be used for short) is in law neither a necessary nor a sufficient condition of being legally responsible for the harm. It is not a necessary condition for two reasons. First, in legal contexts people are often made responsible for harm caused by other persons (e.g. the vicarious liability of employers for employees), animals (e.g. the bite of a dangerous dog), inanimate objects (e.g. the collapse of buildings, the impact of vehicles) or processes (e.g. fire, subsidence). In these instances the ground of responsibility is, from the point of view of the person held responsible, not that he, she or it has caused harm but that they bear the risk that some other person, animal, thing or process may cause harm. The risk may be voluntarily assumed, as in insurance contracts, or may be imposed by law, as in the case of employers’ liability for wrongs committed by employees in the course of their employment. Much law is indeed concerned with the distribution of social risks. The responsibility of the person who bears the risk may be additional or alternative to the responsibility of the person (if any) who wrongfully caused the harm in question. Thus, if an employer is responsible for harm caused by his or her employee to another person the employee may or may not also be legally responsible for that harm. In law the main grounds of responsibility for harm are therefore (i) an agent’s personal responsibility for causing harm and (ii) a person’s responsibility arising from the fact that he, she or it bears the risk of having to answer in legal proceedings for the harm in question.

A second reason why causing harm is not a necessary condition of legal responsibility is that there are many contexts in which a person is civilly or criminally responsible irrespective of whether any harm has been caused by their conduct or that of an agency for which they are responsible. Thus, in Anglo-American law those who trespass on another’s land or who break a contract may be civilly liable and those who unlawfully possess firearms criminally liable though no tangible harm is thereby caused to anyone. Both inside and outside the law many actions are regarded as wrongful whether or not they cause tangible harm. Moreover the imposition of penalties in civil law and of punishments in criminal law need not bear any relation to the harm (if any) caused by the conduct for which the penalty or punishment is imposed.

To cause harm to another is also not a sufficient condition of legal responsibility, even in the eyes of those, such as the early Epstein, who would in general favour making agents strictly liable for the harm they cause. For a person to be legally responsible for causing harm to another requires, apart from a number of conditions relating to jurisdiction, procedure and proof, that the conduct should be of the sort that the law designates as unlawful (e.g. negligent driving) or as a potential source of liability (e.g. keeping a dangerous animal). It also requires that the purpose of the law should encompass harm of the sort for which a remedy is sought. Thus, in some contexts only physical, not economic or psychological harm grounds a legal remedy. Moreover considerations of morality must not rule out liability, as they well might if, for example, a burglar were to claim compensation for an injury suffered while breaking a window in order to enter the victim's house.

There is also a complication concerning the *items* between which causal connection must in law be shown to exist. The inquiries with which law is concerned relate to particular events. Did one action, event, process or state of affairs (event for short) cause another? The link that must be established in legal proceedings between events is of a special type. A person's conduct or a natural event or process can always be described in a number of different ways, but only certain descriptions of an alleged cause are crucial in legal proceedings. For example, if a claim for damages is brought against a motorist for causing injury to the claimant by driving negligently, only that description of his or her manner of driving that amounts to negligence is capable of constituting a relevant cause. Hence 'On 5 March at 5 p.m. Smith drove at sixty miles an hour in a built-up area' may be relevant while ';Smith drove a Mercedes' may not be, though both correctly describe Smith's act of driving a car on the occasion in question. In a legal context, therefore, the link to be established must be framed in terms of a link between particular aspects of events. The claimant in a civil action will typically argue, for example, that the fact that Smith drove at sixty miles an hour in a built-up area on such-and-such an occasion caused the collision that in turn caused the victim to suffer a broken leg. Though it is controversial whether causal connection is to be conceived as a relation between events or facts (Davidson), in law both are relevant. The events in issue must be identified from the point of view of the time, place and persons involved, but the aspect of the events between which a causal link must be shown has to be specified in such a way as to show that it falls within the relevant legal categories, such as (in the example given above) negligence and physical injury.

The relationship between causing harm and legal responsibility is also complex because of the great *variety of relationships* between agency and harm that can be regarded as in some sense causal, or analogous to a causal relationship. An omission to prevent harm when the person concerned has a legal duty to prevent it can ground legal responsibility but would ordinarily be described as 'not preventing' rather than causing the harm. Again, legal responsibility is often imposed, in the context of interpersonal relationships, on those who influence others by advising, encouraging, helping, permitting, coercing, deceiving, misinforming or providing opportunities to others that motivate or enable them to act in a way that is harmful to themselves or to others. In some cases (coercion, deceit) the persons held responsible would naturally be said to have caused the persons influenced to act as they did, while in others they would not, though the weaker interpersonal relationship is in some respects analogous to more plainly causal relationships. Failing to help or provide opportunities to others by advising, warning, informing or

rescuing them or supplying them with agreed goods and services are other grounds of responsibility for negative agency that, again, are at least analogous to causal relationships. The existence of this wide spectrum of causal or near-causal grounds of responsibility recognised in law and morality raises the question whether any uniform theory of causation is capable of accounting for all of them.

3. Criteria for the Existence of Causal Connection in Law

The theories concerning the criteria for the existence of causal connection in law fall into two classes. Some focus on the type of condition that the alleged cause must constitute in relation to the alleged consequence. Others are concerned with a specific feature that the cause must possess in relation to the consequence in order that causal connection may be made out. The first class of theory concerns the identification of the causally relevant conditions of an outcome, or, in the language of causal minimalists, ‘cause-in-fact’. Must the cause be a necessary condition, a sufficient condition or a necessary member of a set of conditions that are together sufficient for the outcome? In law these terms, much discussed in the philosophical literature, are interpreted as meaning ‘necessary or sufficient in the particular circumstances in issue’. The inquiry will be, for example, into what was a necessary or sufficient to cause a particular persons’ death, not what are in general the necessary or sufficient conditions of death.

The second type of theory concerns the criteria for determining the limits of legal responsibility for causing harm. Even supposing that the alleged cause constitutes the right sort of condition of the outcome (e.g. a necessary condition), responsibility cannot extend indefinitely. The failure of a doctor to prescribe an effective contraceptive cannot be held to be responsible for the death of the victim of a murder committed by the child conceived as a result of the doctor’s negligence. Some consequences are ‘too remote’. But what are the appropriate criteria of limitation?

In many legal contexts and in the view of many theorists a single criterion is called for. It should be remembered, however, that the search for a single criterion may be no more than a response to legal doctrine. This sometimes requires all the limiting factors to be brought under a single umbrella, such as ‘proximate cause’ or ‘adequate cause’ even though, underlying these phrases, there are a number of distinct reasons for imposing limits on the extent of responsibility. A number of expressions are used to describe the allegedly single limiting factor, in particular ‘proximate (adequate, direct, effective, operative, legal, responsible)’ cause in contrast with ‘remote, indirect or legally inoperative’ causes.

Some theorists (for example Leon Green and others since the 1920’s up to Wright and Stapleton today) hold that only the issue of causally relevant condition or cause-in-fact is genuinely causal. It alone raises questions to which an objective, scientifically valid, answer can be given (Becht and Miller). Even this has been questioned by Malone, who has pointed to the incorporation of normative considerations in the rules for proving cause-in-fact in civil law. The second type of theory concerns questions of responsibility that would in the view of these causal minimalists be better addressed directly rather than by asking whether on the facts a causal relation existed between agency and harm. One way of doing this

is to ask what would be the fairest way of distributing the relevant social risks. Another (Posner) would be to place responsibility, especially in civil law, on the person best placed to avoid the loss most cheaply. In practice legislators and judges have seldom abandoned the traditional terminology in discussing the second issue, but the proposal to do so has been repeatedly revived.

3.1 Causally relevant conditions: ‘Cause-in-fact’

What sort of condition must be attributed to an agency for its action or intervention (action for short) to count as causal? Opinion is divided between those to whom the action must in the circumstances be necessary to the outcome (a but-for condition), those to whom it must in the circumstances form a necessary part of a complex of conditions sufficient for the outcome (a NESS condition), and those who would describe the required connection in a more quantitative or scalar mode by requiring that the action be a ‘substantial factor in’ or ‘contribute to’ the outcome.

The but-for theory, endorsed by many legal and philosophical theorists including Mackie, has the heuristic advantage that a simple and often reliable way of ruling out the existence of causal connection between agency and harm is to ask whether the harm would in the circumstances have occurred in the absence of the agency. If the harm would have occurred in any event the agency is probably not its cause or one of its causes. If it would not have occurred in the absence of the agency the agency will be a causally relevant condition or, if one endorses causal minimalism, a cause-in-fact of the harm.

There are however cases in which the but-for test is difficult to reconcile with our intuitive judgements of responsibility. These concern two types of case in particular, those of over-determination and of joint determination. If two huntsmen independently but simultaneously shoot and kill a third person, or two contractors independently fail to deliver essential building supplies on time, it is intuitively clear that each should be held responsible for the death or building delay. Yet the but-for test seems to yield the conclusion that neither has caused the harm. Again, in interpersonal relationships it is often the case that advice etc. can be regarded as contributing to a person’s decision without its being shown that the person would not have acted as they did apart from the advice. Many reasons bear on the decisions we make. Sometimes it is not possible to be sure that in the absence of one of them the decision would have been different. We know only that to the person reaching the decision the reasons taken into account were jointly sufficient to induce him, her or it to decide as he or she did.

In reply it is argued (Mackie) that in these cases all the agencies that are singly or jointly sufficient for the outcome together constitute its cause. But in law this does not solve the problem because, unless the agents are acting in concert, the responsibility of each agency has to be independently established. This can be done either by an appeal to intuitive notions of responsibility or by recourse to an alternative ground of responsibility based on risk. On the alternative view an agency that provides an independently or jointly sufficient condition of harm bears the risk that that harm will eventuate even if it would in the circumstances have come about in any event.

Some of those who reject this approach (e.g. Hart and Honoré, Wright) have recourse to a theory based

on J.S.Mill's notion of a jointly sufficient set of conditions. The theory also draws on Mackie's idea, in the context of causal generalisations, of an INUS condition (insufficient but non-redundant part of an unnecessary but sufficient condition). They advocate the view that in a specific situation a causally relevant condition is a necessary element of a set of conditions jointly sufficient for the harmful outcome. For this Wright's term NESS condition (necessary element of a sufficient set) is currently used, a NESS condition being a specific instance of an INUS condition. NESS supporters therefore appeal to the idea that particular causal links are instances of generalisations about the way in which events are connected. They argue that in order to test whether an outcome would have occurred in the absence of the agency in question it is necessary to make a counterfactual calculation, which can only be done on the basis of such generalisations.

Those who reject the NESS theory either assert that singular causal judgments do not depend on generalisations or point to the fact that reliable generalisations of the sort presupposed by it are in practice virtually confined to inorganic physical processes. Organic processes, such as those involved in the development of disease, and, still more, in decision-making by human beings, do not conform to settled patterns. The NESS theory therefore has at most a narrow range of application.

Some of those who are impressed by what they see as the deficiencies of both the but-for and NESS theories prefer a more quantitative or scalar approach, according to which an agency can cause an outcome to a greater or less extent (Moore). They argue that an agency must be a 'substantial factor in' or 'contribute to' the harmful outcome in order to be legally a cause of it. This approach has a particular attraction when a number of processes (e.g. several fires or pollutants) merge to bring about harm. It enables distinctions to be made according to the extent of contribution of a particular process to the outcome. It also fits the rule that in most legal contexts an agency, in order to be responsible for the whole of the harm that ensues, need only be shown to be one of the causes of harm, not the sole cause. The criticism that can be made of this approach is that it presupposes an independent understanding of causes as necessary and/or sufficient conditions in relation to their consequences.

Difficult legal problems arise in certain cases of overdetermination, often termed those of 'overtaking causes' or 'causal preemption'. Suppose that a lethal dose of poison is given but the victim is fatally wounded before the poison takes effect. The pre-empting, not the pre-empted condition is taken to be the cause of the death. Which condition is taken to preempt the other is sometimes controversial but it is clear than in reaching a decision attention must be paid to the stages and processes by which the alleged causes lead to the harmful outcome.

The idea that responsibility should depend on the agent's having changed the course of events points in the direction of the but-for theory. The function of cause in relation to recipes and prediction points towards the NESS theory. The phenomenon of multiple causes, which have often to be weighed against one another, points to a quantitative theory. But whichever is favoured has to be applied in the light the law's commitment to vindicating rights and securing a fair distribution of risks.

3.2 'Proximate cause'

The theories about the specific qualities that an agency must possess in relation to the outcome in order to be its cause in law are in Anglo-American law often grouped under this rubric, though many other terms (e.g. adequate, direct, efficient, operative, legal, responsible) are also found in the literature. These limiting theories are invoked because if every causally relevant condition (cause-in-fact) is treated as grounding responsibility for the outcomes to which it is causally relevant the extent of legal responsibility will extend almost indefinitely. (This alarming scenario would however be subject to independent legal requirements as regards proof, type of damage and lapse of claims through the passage of time). The theories in question therefore embody reasons for limiting the extent of legal responsibility. The reasons adduced for limiting responsibility are however differently viewed by different theorists. Causal minimalists treat all these theories as non-causal, in the sense that they embody grounds of legal policy other than the policy of holding the agent responsible for the harm caused by their action or intervention. Others treat some of the suggested limiting factors as causal and others as non-causal. It is indeed not open to dispute that at least two non-causal factors limit the extent of legal responsibility. One is the scope and purpose of the rule of law in question. No rule is intended to give a remedy for every conceivable type of harm or loss. Another concerns the aspiration of the law to achieve results that are morally unobjectionable. This rules out certain claims that would be inequitable on the part of the claimant or unfair towards the agent. It needs to be stressed that the grounds for limiting responsibility will not necessarily be the same in every branch of the law. In particular, the greater the weight attached to considerations of risk distribution the more likely it is that different limits will be appropriate in, for example, criminal, civil and public law.

3.3 Allegedly causal grounds of limitation

Certain theorists reject causal minimalism, which involves a restricted notion of cause that is current in no extra-legal context. They propose grounds of limitation that reflect the causal judgements that would be made outside the law. They claim that these grounds have a basis in ordinary usage (Hart & Honoré) or in the metaphysics of causation (Moore). The chief grounds proposed are that responsibility is limited (i) when a later intervention of a certain type is a condition of the harmful outcome (ii) when the agency has not substantially increased the probability of the harmful outcome that in fact supervenes and (iii) when the causal link involves a series of steps and ultimately peters out, so that the outcome is too remotely connected with the alleged cause. They argue that in these cases the agency, though a causally relevant condition, did not cause the outcome.

The idea that responsibility is excluded when the harm in question was conditioned by a *later intervention* is conventionally expressed by saying that an intervening or superseding cause broke the causal link between agency and outcome. These ‘breaks’ are not conceived as physical discontinuities in the course of events. The metaphor derives rather from the fact that in an explanatory context a cause may be regarded as an intervention in the normal course of events. The most persuasive explanations of an outcome are those that point to a condition that is abnormal or unexpected in the context or to a deliberate action designed to bring the outcome about. If these criteria are then applied in attributive contexts, an agency will not be regarded as the cause of an outcome when that outcome is explained by a later abnormal action or conjunction of events or a deliberate intervention designed to bring it about. A

later event of this sort is contrasted with a state of affairs (e.g .victim’s thin skull) existing at the time of the alleged cause. The latter, however extraordinary, does not preclude the attribution of the outcome to which it contributes to the alleged cause. In practice this notion is widely applied in both civil and, as Kadish has shown, criminal law. The use of these criteria of intervention in legal systems is said to be derived from common sense and to be consistent with treating causal issues in law as questions of fact. It is also supported (Honoré) on the ground that to attribute only a limited range of outcomes, whether achievements or failures, to human agents fosters a sense of personal identity that would be lost if the attribution to agents was not limited in this way. If there were not such a limiting factor we should have to share our successes and failures with many other people of whom it could be said that but for their actions what we think of as ‘our’ distinctive successes and failures would not have occurred. For example the success of a student in an examination would be *equally* the achievement of all those (parent, teacher, doctor, grant-giver, girl/boy friend) who made it possible for the student to succeed. It would not be specially the student's.

The criticism of this notion of later intervention takes two forms. First, the criteria set out are too vague to govern decision in controversial cases. Suppose that a motorist negligently injures a pedestrian, who is then taken to hospital and wrongly treated for the injury. Instead of asking whether the mistaken treatment was so abnormal as not to be accounted a consequence of the motorist’s negligent driving it would, in the critics’ eyes, be better to ask whether the risk of medical mistreatment should be borne exclusively by the hospital authorities. Secondly, even if the criteria suggested for selecting certain conditions as causes are in place in explanatory inquiries they are not necessarily so in attributing responsibility. There is no good reason to transfer them from an explanatory to an attributive context. To do so in civil law may result in saddling a person guilty of momentary carelessness with massive losses (Waldron).

Another limiting notion that has some claim to be regarded as causal is that of *probability*. According to the adequate cause theory, put forward by the physiologist Von Kries in 1886, developed systematically by Träger and advocated in a contemporary form by Calabresi, an agency is a cause only if it significantly increases the objective probability of the outcome that in fact ensues. Objective probability is here contrasted with subjective foreseeability, but this probability must be relative to an assumed epistemic base. It is inevitably a matter of policy which base to choose, and whether to include information not known or not available to the agent when he or she or it acted. Responsibility is excluded in relation to an outcome the probability of which was not substantially increased by the agency in question. This theory, long orthodox in German civil law, but increasingly supplemented by policy-oriented criteria, is intuitively attractive when the agent wrongfully exposes someone to a risk of harm to which they would not otherwise be exposed. For example, the agent wrongfully obstructs a pathway so that the claimant is forced to take a more dangerous route along a canal, and falls into the canal, sustaining injury. The obstructer is then the adequate cause of the injury. But one who wrongfully delays a passenger who is as a result obliged to board a later airplane, which crashes, is not the adequate cause of the passenger’s death in the crash. At least on the basis of information available at the time, the probability of being killed in an air crash was not substantially increased by the delay.

There are however instances in which an agency substantially increases the probability of harm but the

harm that occurs would intuitively be attributed to a later intervention. Suppose, for example, that in the example given a passer-by deliberately threw the claimant into the canal. It would be natural to attribute any injury suffered by the claimant not to the obstruction of the pathway but to the act of the third person. This objection can be met by having recourse to the risk theory, a version of the probability theory with strong support in Anglo-American writing in both criminal and civil law (Keeton, Seavey, Glanville Williams). According to this theory responsibility for harmful outcomes is restricted to the type of harm the risk of which was increased by the agency's intervention. The harm must be 'within the risk'. But much then turns on how the agent's conduct and the risk are defined. Is the risk of falling into the canal different from the risk of being pushed into it?

As stated earlier, in law responsibility for harm can rest on risk allocation as well as on causation. The risk theory has merits that are independent of its claim to explain what it is for an agency to cause harm. It can be treated as illustrating a wider principle that responsibility for harm is confined to the type of harm envisaged by the purpose of the rule of law violated (Normzweck), a theory espoused in Germany (Von Cämmerer, J.G.Wolf). For example, if a rule requiring machinery to be fenced is designed to prevent harmful contact between the machinery and the bodies of workmen, a workman who suffers psychological harm from the noise made by the unfenced machine cannot ground a claim for compensation on the failure to fence. The fencing requirement was not designed to reduce noise, even though a proper barrier would have reduced the noise to such an extent as to avoid the psychological trauma.

The limitations set by the purposes of legal rules cannot be regarded as causal. They vary from one branch of the law and one legal system to another. It is true that sometimes the purpose of legal prohibition may be the simple one of imposing responsibility for the harm caused by a breach of that prohibition. In that case the limits set by causal and purposive criteria coincide. But even in such a case it is a matter of legal policy which types of harm are to be compensated or to lead to criminal liability. The purposive limits on responsibility have therefore either to be regarded as additional to those (later intervention, heightened probability) proposed by those who reject causal minimalism, or as replacing them. The latter view is consistent with causal minimalism.

Other proposed criteria of limitation are based on moral considerations. Theorists who regard fault as an essential condition of criminal or civil responsibility often argue that a person should not be liable for unintended and unforeseeable harm. There are problems about settling whether only the type of harm or the specific harm must be unforeseeable, and the moment at which foreseeability is to be judged. But foreseeability, though it bears some relation to probability, is clearly a non-causal criterion, and one that can apply only to human conduct, not to other alleged causes. Moreover some supporters of the risk theory argue that different criteria should govern the existence and extent of legal liability. Even if the foreseeability of harm is a condition of liability, sound principles of risk allocation place on the agent who is at fault in failing to foresee and take precautions against harm the risk that an unforeseeable extent of harm will result from his or her fault, provided that this is of the type that the rule of law in question seeks to prevent.

There is no reason to suppose that the law, when it engages in explanatory inquiries, adopts different

criteria of causation from those employed outside the law in the physical and social sciences and in everyday life. However, even here, requirements of proof may lead to a divergence, for example, between what would medically be treated as the cause of a disease and what counts in law as its cause. As regards attributive uses of cause, the fact that the law has to attend simultaneously both to the meaning of terms importing causal criteria and to the purposes of legal rules and their moral status makes the theory of causation a terrain of debate that is unlikely to yield solutions commanding general agreement (e.g. Stapleton, Wright 2001; Moore, forthcoming).

Bibliography

- Becht, A.C., and F.W. Millar, *The Test of factual Causation in Negligence and Strict Liability* (St. Louis, Miss., 1961).
- Calabresi, G., ‘Concerning Cause and the Law of Torts’ (1975) 43 *U.Chicago LR* 69-108.
- Cämmerer, E. von, (1956, Freiburg-im-Breisgau), *Das Problem des Kausalzusammenhange im Rechte, besonders im Strafrechte*.
- Coase, R.H., ‘The Problem of Social Cost’, (1960) 3 *J. Law and Economics* 1-44.
- Collingwood, R.G., *An Essay on Metaphysics* (1940)
- Donald Davidson, ‘Causal Relations’ in *Essays on Actions and Events* (Oxford 1980)
- Epstein, R.A., ‘A Theory of Strict Liability’, (1973) 2 *J.Legal Studies* 151-204.
- -----, ‘Causation and Corrective Justice. A reply to two Critics’, *J. Legal Studies* 8 (1979) 477-504.
- Green, L., *Rationale of Proximate Cause* (Kansas City, 1927).
- -----, ‘The Causal Relation Issue in Negligence Law’ (1962) 60 *Michigan LR* 543-576.
- Hart, H.L.A., and Tony Honoré, *Causation in the Law* (2nd ed. Oxford, Clarendon 1985).
- Honoré, A.M. (Tony), ‘Causation and Remoteness of Damage’, *International Encyclopedia of Comparative Law* (1971) XI ch.7.
- -----, ‘Necessary and Sufficient Conditions in Tort Law’, (ed. D.G.Owen) *Philosophical Foundations of Tort Law* (Clarendon, Oxford 1995) 363-385.
- Kadish, S., ‘Causation and Complicity: A Study in the Interpretation of Doctrine’, *California LR* 73 (1985) 323-410.
- -----, ‘A Theory of Complicity’, in *Issues in Contemporary Legal Philosophy*: *The Influence of H.L.A.Hart* (ed. Ruth Gavison 1987) 287-303.
- Keeton, R.E., *Legal Cause in the Law of Torts* (1963).
- Kries, J. von, *Über den Begriff der objektiven Möglichkeit und einiger Anwendungen desselben* (1888).
- Landes, W., and R. Posner, ‘Causation in Tort Law; An Economic Approach’, *J. Legal Studies* 12 (1983) 109-134.
- Mackie, J.L., *The Cement of the Universe. A study of Causation* (Oxford, Clarendon 1974, 1980).
- Malone, W.S., ‘Ruminations on Cause-in-fact’ (1956-7) 9 *Stanford LR* 60-99.
- Moore, M., ‘The Metaphysics of Causal Intervention’, California LR (2000) 827-878.
- -----, *Causation and responsibility* (forthcoming)
- Peczenik, A., *Causes and Damages* (Lund, 1979)

- Posner, R.A., A Theory of Negligence , J. Legal Studies 1 (1972) 29-96.
- Seavey, W., 'Mr Justice Cardozo and the Law of Torts', Harvard LLR 52 (1939) 371-407;
- Stapleton, J., 'Law, Causation and Common Sense', (1988) 8 *Oxford J. Legal Studies* 111-131
- -----, 'Legal cause: Cause-in-Fact and the Scope of Liability for Consequences', (2001) 54 *Vanderbilt LR* 941-000.
- Träger, *Der Kausalbegriff im Straf- und Zivilrecht* (1904)
- Waldron, J., 'Moments of Carelessness and Massive Loss', in *Philosophical Foundations of Tort Law* (ed. D.G.Owen, Clarendon, Oxford 1995) 387-408
- Williams, G., 'The Risk Principle', *Law Quarterly Review*, 77 (1961), 179-212
- Wolf, J.G., *Der Normzweck im Deliktsrecht. Eine Diskussionsbeitrag*. (Göttingen, 1962).
- Wright, R.W., 'Causation in Tort law', *California LR* 73 (1985) 1737-1828.
- -----, 'Actual Causation vs. Probabilistic Linkage; The Bane of Economic Analysis', *J. Legal Studies* 14 (1985) 435-548.
- -----, 'Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts', *Iowa LR*73 (1988) 1001-1077.
- -----, 'Once More into the Bramble Bush: Duty, Causal Contribution and the extent of Legal Responsibility', 54 *Vanderbilt LR* (2001) 1071-1132.

Other Internet Resources

- Fumerton, R., and Kress, K., 'Causation and the Law: Preemption, Lawful Sufficiency, and Causal Sufficiency, *Law and Contemporary Problems*, special issue on Law and Causation in Science, John M. Conley (ed.), Volume 64/Number 4 (August 2001)
- Spellman, B., and Kincannon, A., 'The Relation Between Counterfactual ("But For") and Causal Reasoning: Experimental Findings and Implications for Juror's Decisions, *Law and Contemporary Problems*, special issue on Law and Causation in Science, John M. Conley (ed.), Volume 64/Number 4 (August 2001)

Related Entries

causation: in science | causation: the metaphysics of | moral responsibility

A. M. (Tony) Honore
tony.honore@law.ox.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 8, 2001
Content last modified: November 8, 2001

Spinoza's Psychology

In Part III of his *Ethics*, "On the Origin and Nature of the Affects," Spinoza addresses two of the most serious challenges facing his thoroughgoing naturalism. First, he attempts to show that human beings follow the order of nature. Human beings, on Spinoza's view, have causal natures similar in kind to other ordinary objects, other "finite modes" in the technical language of the *Ethics*, so they ought to be analyzed and understood in the same way as the rest of nature. Second, Spinoza attempts to show that moral concepts, such as the concepts of good and evil, virtue, and perfection, have a basis in human psychology. Just as human beings are no different from the rest of nature, so moral concepts are no different from other concepts. They must be wholly explicable under the same laws which explain the rest of nature. Spinoza's detailed account of the human affects--the actions and passions of the human mind--is crucial to both tasks. For his argument to succeed, the theory of the affects must be both a plausible account of human psychology and a plausible basis for ethics.

---

1. The Human Being as Part of Nature

In the Preface to Part III, Spinoza states his view that all things alike must be understood to follow from the laws of nature:

> The laws and rules of nature, according to which all things happen, and change from one form to another, are always and everywhere the same. So the way of understanding the nature of anything, of whatever kind, must also be the same, viz. through the universal laws of nature.

Many philosophers have treated the human mind as an exception to otherwise universal natural laws, as a thing which is capable of good and evil or as an uncaused cause of action for example. Spinoza though insists that human beings are not "outside nature." Any features or deeds of human beings that seem exceptional, then, must have for Spinoza some explanation in terms of natural laws. That is, if there is any sense at all in saying that something which happens to a particular human being is good or evil, that must be a sense in which any other finite mode might also be said to have benefitted or have been harmed; and, if there is any sense at all in calling a human being free, that must be a sense in which other things of different kinds might also be said to be free.

Spinoza's thesis (IIIp7) that the essence of any finite mode, including any human mind (IIIp9), is a striving (*conatus*) to persevere in being is an attempt to give an account of nature under which human beings with their apparent peculiarities are natural. Spinoza argues that all finite modes strive to persevere in being (IIIp6), and he uses an analysis of human striving to explain human good and evil and human freedom in terms that might apply to any finite modes. An action of a human mind (IIp48) or of anything other than God (Ip17c2), cannot be free, for Spinoza, in the sense of being the effect of a free cause. There is human freedom for Spinoza, however, in the sense of freedom from external interference: I am free in producing some effect (for example, in doing something) if that effect follows from my essence alone, or, in other words, if it is the effect of my striving alone. (For discussions of action and human freedom, see IIId2 and V Preface.) 'Good' and 'evil' are labels that describe natural properties in the sense that they describe changes that might occur in any particular things at all (although we reserve the labels for these changes when they occur in human beings). An increase in the power with which a mind strives is good, for Spinoza, and a decrease evil (see IIIp11s, IIIp39s and IV Preface). Because the striving thesis thus explains both human freedom and good and evil, it is of central importance to Spinoza's psychology and ethics. One might raise questions about the validity of Spinoza's argument to the doctrine, about its plausibility as an account of the nature of particular objects, or about its plausibility as an account of human nature. The subsections which follow address these issues in turn.

1.1 The Argument to the Striving Doctrine

Spinoza's argument to IIIp6 is uncharacteristically insulated from the rest of the *Ethics*. As Spinoza presents the argument at IIIp6d, it depends principally upon IIIp4, a proposition which Spinoza takes to be self-evident, and IIIp5 which derives from IIIp4 alone. The argument also involves, less directly, IP25C and its gloss at IP34.

> I Proposition 25 Corollary: Particular things are nothing but affections of God's attributes, or modes by which God's attributes are expressed in a certain and determinate way.

> I Proposition 34: God’s power is his essence itself.
>
> III Proposition 4: No thing can be destroyed except through an external cause.
>
> III Proposition 5: Things are of a contrary nature, i.e., cannot be in the same subject, insofar as one can destroy the other.
>
> Therefore, I Proposition 6: Each thing, as far as it can by its own power, strives to persevere in its being.

A thing’s essence may not be absolutely equivalent to its nature for Spinoza, since a thing such as a square circle has a nature but cannot exist (Ip11) and one might interpret Spinoza as holding that anything which has an essence might exist (IId2). Still, the two terms might be taken interchangeably here because Spinoza is only describing existents. If this assumption is correct, then perhaps Spinoza’s reasoning runs like this:

> Particular things, subjects, are expressions of power, since they are modes of God’s attributes (Ip25) and God’s attributes constitute God’s essence (ID4) and God’s essence is his power (Ip34). It is self-evident that nothing can be destroyed except through an external cause (IIIp4), so an apparent particular thing which is self-destructive is in fact at least two (IIIp5). The power a genuine particular thing expresses, then, must therefore be directed toward its own perseverance in being (IIIp6).

Pace Spinoza, the claim at IIIp4 that no thing can be destroyed except through an external cause is not clearly self-evident. Even assuming IIIp4 to be true, however, one might raise questions about Spinoza’s argument. Why is it that, just because a thing does not strive to destroy itself, that thing must therefore strive to persevere in being? A thing might it seems not strive for anything, or perhaps it might strive to do something which is neither perseverance nor self-destruction. Spinoza’s use of IP34 and Ip25c seems intended to rule out the first of these possibilities. Although Spinoza’s term ‘express’ (*exprimere*) is notoriously unclear, it may mean something like "is a particular form of." In that case, because particular things are expressions of God’s essence, his power, they must be particular forms of power. So there cannot be a thing which does not strive at all or, in other words, there cannot be a thing which is not any expression of power at all.

The second version of the objection, the version which notes the possibility that a particular thing as described in IIIp4 might strive for something other than either self-destruction or perseverance, remains a challenge to sympathetic readers of the *Ethics* however. The negation of IIIp6, that some thing does not strive to persevere in its being, where perseverance in being is understood as a particular end among many possible options, is perfectly consistent with the truth of IIIp4. After all, not striving to persevere in being, which IIIp6 rules out, is not the same thing as striving not to persevere in being, which IIIp4 rules out. A sympathetic reader of Spinoza might try to resolve the difficulty through an understanding of what

it means to strive to persevere in being, under which striving to persevere in being comes to mean just the same thing as striving to do something other than destroy oneself, in particular, striving to maintain a present state. (See Curley 1, 109, for an interpretation similar to this one.) This reading comes closer to making the argument from IIIp4 to IIIp6 seem valid, but it raises a new problem: that of reconciling this interpretation of striving with Spinoza's accounts of human motivation which follow from IIIp6. For Spinoza consistently regards sane human beings as finite modes who, beyond merely not trying to kill themselves, actively try to preserve themselves. People do not merely resist changes to whatever state they are in; they strive to change their states in order to know more and in order to live with a greater force. One of the main problems Spinoza faces, then, is reconciling the most plausible version of IIIp6 as an account of the natures of ordinary objects (under which IIIp6 represents a version of the principle of inertia) with the most plausible version of IIIp6 as an account of human nature (under which IIIp6 represents a version of psychological egoism).

1.2 The Striving Doctrine as an Account of the Natures of Particular Objects

Despite worries that one might have about the validity of Spinoza's argument, the doctrine has at least some claim to plausibility as an account of the nature of particular things. The *Ethics* stands badly in need of some account of what finite modes are, after all, and IIIp4 provides at least one interesting way of distinguishing genuine objects from mere constructs: if the thing in question destroys itself it is not a genuine object. Thus, by IIIp4, a thing which destroys itself--one might think a lit candle or a time bomb such a thing--is not a genuine object but a thing which does not destroy itself is. To the extent that IIIp4 makes most of the things we intuitively consider particular objects particular objects, it captures ordinary views. To the extent that it rules out some clear class of things (lit candles and time bombs), it represents a provocative philosophical thesis. The plausibility of the doctrine depends on whether we find that there really is reason to find basic metaphysical differences in kind between "things" which tend to destroy themselves and things which do not.

IIIp6 introduces, perhaps, a slightly different thesis about what it means to be a particular thing: a particular thing is one which strives to persevere in being. IIIp6's dependence upon IIIp4 suggests that this thesis means that any object will remain in the same state unless external causes affect it. Such a thesis appears to be a very general form of the principle of inertia, and, indeed, Spinoza seems to invoke the principle of inertia in the terms he uses at IIIp6. '*Conatus*' is a technical term of Cartesian physics, referring to an object's motion. Spinoza himself uses the term in this way in his exposition of Descartes's *Principles of Philosophy*. (Compare, for example, Descartes's *Principles* II, art. 37, III, art. 56 and III, art. 58 to Spinoza's exposition IIp14c, IIID3 and IIp17, respectively). Moreover, at IIIp6, in addition to using the term '*conatus*' again, Spinoza also uses the same phrase that he uses in framing the principle of inertia at IIp14 of his exposition of Descartes: "as far as it can by its own power" *quantum in se est*. (Note however that there is some controversy over how this phrase is to be understood: see Curley's footnote to IIIp6 in his translation and Garrett, 1999, note 2.) So there is a good textual basis for the conclusion that IIIp6 indeed has this meaning.

Apart from the question of how the principle of inertia can give us an understanding of human nature, this interpretation of IIIp6 raises two difficult questions. First, one might object that Spinoza puts too much philosophical weight on a single fact about the nature of things. After all, at IIIp7, Spinoza argues that a thing's striving to persevere in being is its essence. One should conclude from an understanding of IIIp6 as a new law of inertia, then, that IIIp7 simply means that an object's current state, in particular its current state of motion, is its essence. But why should a thing's tendency, in the absence of interference, to maintain its current state represent its essence? To hold such a view would be, seemingly arbitrarily, to elevate one natural property above others. On the face of it, there is no reason to suppose that a thing's tendency to continue to hold the properties it now holds is any more characteristic of or essential to it than any of its various tendencies to interact with other kinds of objects to produce particular kinds of effects.

Second, one might object that, IIIp6, understood as a restatement of the principle of inertia, extends a physical principle to mind without sufficient clarity. In stating the principle of inertia, both Descartes and Spinoza are careful to limit the claim to a claim about bodies. Spinoza, for example, in his definition of *conatus ad motum*, IIId3 of his exposition of Descartes, writes:

> By striving for motion we do not understand any thought, but only that a part of matter is so placed and stirred to motion, that it really would go somewhere if it were not prevented by any cause.

In addition to characterizing matter, however, IIIp6 is a foundational claim about the nature of mind and, in particular, about human psychology. There is a basis, in Spinoza's dual aspect theory and parallelism, for thinking that whatever is true about bodies is true about minds also (and see IIIp10 and IIIp11 for Spinoza's account of striving and the mind/body relation). Striving in physics, however, is understood as a tendency to a certain kind of motion, and motion seems, if anything does, to belong to bodies alone. So Spinoza needs to supply an account of the mental correlate to the physical "striving for motion." But IIIp6 leaves open the question of what it means for mind to strive. On this objection, the striving doctrine uses a kind of metaphorical language, the term 'striving', where a precise and literal account of what it is that is characteristic of mind is required.

1.3 The Striving Doctrine as an Account of Human Nature

Spinoza's naturalism benefits rhetorically from his use of the term '*conatus*' to describe the essences of human beings and other finite modes alike. For the term is not only a technical term of Cartesian physics. Cicero uses the term in *De Natura Deorum* (and other Roman and Greek Stoics use close cognates) in a psychological sense, referring to human desire, and Hobbes in his physiology uses the term to refer to the physical causes of human desire (*Leviathan* VI). So '*conatus*' has both broad, physical and specifically human, psychological connotations which help to make the gap between nature and the human mind appear narrow.

Whether Spinoza successfully capitalizes on his rhetorical skill, however, and draws a plausible account of the nature of the human mind out of his general account of the essences of finite modes depends upon IIIp9:

> III Proposition 9: Both insofar as the mind has clear and distinct ideas, and insofar as it has confused ideas, it strives, for an indefinite duration, to persevere in its being and it is conscious of this striving it has.

IIIp9 suggests that Spinoza is a psychological egoist of some sort. That is, it suggests that he believes that what human beings desire to do is to secure their own interests (construed here as perseverance in being). Indeed Spinoza goes on to define desire at IIIp9 as human striving (or appetite) together with the consciousness of striving. So clearly human desire for Spinoza is part of the striving for perseverance in being and thus shares its character.

There is some question, however, about what variety of psychological egoism Spinoza holds. Desire might be part of a striving for perseverance, after all, without all desires being desires for perseverance. One might have a strong instinctual desire for things which are instrumental to perseverance in being without desiring perseverance itself, for example. Or one might desire perseverance in being but also desire other kinds of things.

IIIp9 might be supposed to support a very strong version of psychological egoism, orthodox egoism (perhaps Delahunty, 221, holds this view). Orthodox egoism, is the view that human beings are always consciously selfish. Under this view, A consciously desires only those objects which benefit A, B desires only those objects which benefit B, and so on for all human beings. At IIIp9, Spinoza writes that the human mind seeks to persevere in being both insofar as it has clear and distinct ideas and insofar as it has confused ideas. It is natural to understand this claim to mean something like the following:

> Sometimes people do things which conduce to their perseverance and other times people do things which fail to so conduce. In both types of case, though, people desire to persevere. When I do something that fails to help me to persevere, it's because the ideas on which I based my action were confused; that is, I thought I knew what would help me to persevere, but I was wrong. When I do something that does help me to persevere, though (unless I have simply been lucky in acting from an inadequate idea), it is because I acted on clear and distinct ideas or, in other words, genuine knowledge about what would help me to persevere.

The categorical language Spinoza uses in the Appendix to Part I provides explicit support for this interpretation of IIIp9: "men act always on account of an end, viz. on account of their advantage, which they want." Moreover there are other important passages in Spinoza's works which are strongly compatible with the interpretation of Spinoza as an orthodox egoist. These include *Ethics* IVp8d, and his political writings, especially *Ethics* IVp36s2, and his *Political Treatise*, chapter 2)

Other evidence suggests that Spinoza is not an orthodox egoist, however. In particular, there is reason to question whether the argument of the *Ethics* commits Spinoza to the account of actions following from confused ideas that the interpretation of IIIp9 above attributes to him. Part of IIIp39s concerns those agents who are the most confused. That passage is useful because it describes explicitly the conscious thought-processes that precede action:

> Though men are liable to a great many affects, so that one rarely finds them to be always agitated by one and the same affect, still there are those in whom one affect is stubbornly fixed. For we sometimes see that men are so affected by one object that, although it is not present, they still believe they have it with them. When this happens to a man who is not asleep, we say that he is mad or insane. Nor are they thought to be less mad who burn with Love,and dream, both night and day, only of a lover or a courtesan. For they usually provoke laughter. But when a greedy man thinks of nothing else but profit, or money, and an ambitious man of esteem, they are not thought to be mad, because they are usually troublesome and are considered worthy of Hate. But Greed, Ambition, and Lust really are species of madness, even though they are not numbered among the diseases.

In this scholium (and in other discussions of monomania such as III Definition of the Affects, XLVIII and IVp20s) Spinoza describes a variety of possible ends of human action, none of which is perseverance in being. Moreover, lest one think that the greedy man seeks profit because he mistakenly believes that it leads to perseverance, Spinoza emphasizes the point that it is always one and the same object that obsesses these men.

IIIp39s suggests that Spinoza holds a different kind of view, predominant egoism, the view that most people, most of the time consciously desire perseverance in their own being. The particular type of predominant egoism that IIIp39s suggests introduces important aspects of Spinoza's ethical theory: if the most confused people, people addled by greed or lust or ambition are those who always seek something other than perseverance in being, then perhaps Spinoza's view is that, for any of us who act on some similar sentiments occasionally, we do so just to the extent that we also have confused ideas. Thus human beings are predominantly egoistic because, by and large, we act on clear and distinct ideas: it is rational to seek to persevere in being (see also what "reason demands" at IVP18S). But we are not orthodox egoists, on this interpretation of Spinoza as a predominant egoist, because we are not fully rational. To the extent that we have confused ideas, we may indeed consciously pursue ends other than perseverance in being. On this interpretation of Spinoza, there is a right (or at least a rational) end to pursue--perseverance in being--and other ends are wrong (or at least irrational).

IVp20 provides support for this interpretation of Spinoza's predominant egoism:

> IV Proposition 20: The more each one strives, and is able, to seek his own advantage, i.e., to preserve his being, the more he is endowed with virtue; conversely, insofar as each one neglects his own advantage, i.e., neglects to preserve his own being, he lacks power.

Here Spinoza explicitly admits that a person may "neglect his own advantage." So IVp20 apparently contradicts the orthodox egoism of I Appendix. Moreover, IVp20 states that, to the extent that a person does seek perseverance in being, that person is virtuous. Virtue has a metaphysical connotation in the *Ethics*. A thing's virtue is just the same as its power (IVd8). But the term undeniably has moral connotations as well. So IVp20 suggests, as IIIp39s does, that consciously trying to preserve oneself is right and neglecting to preserve oneself is wrong.

IIIp9 admits of various interpretations. However, the weight of the textual evidence supports the view that he is a predominant, not an orthodox, egoist. Any particular human desire, then, even a desire that is not a desire for perseverance in being or its means, must on Spinoza's view be related to perseverance in being in some way (by IIIp6 and IIIp9s). Spinoza's association of desires for things other than perseverance in being in passages such as IIIp39s and IVp20 suggests moreover that such desires are part or product of confusion. So passionate desires, for Spinoza, are often desires for things other than perseverance in being, although they may be confused desires for perseverance as well (see IVp63s2 and other discussions of fear).

Further reading: For discussion of IIIp4, see Matson. For interpretations of Spinoza's argument from IIIp4-IIIp6, see Bennett (1); Curley (1); and Della Rocca. For discussions of the ancient roots of the striving doctine, see James and Wolfson. Lachterman covers Spinoza physics and his use of Descartes. Most book-length interpretations of the *Ethics* include detailed accounts of Spinoza's view of human nature. One recent account of interest is Yovel's. The best recent discussions of psychological egoism come in the context of the interpretation of Hobbes, to whom Spinoza is sometimes compared. See Kavka and Hampton.

2. The Affects

Spinoza's account of the affects (*affectus*) of the human mind is a response to one of the central problems for his naturalism. It is an attempt to show how the wide range of desires and emotions of the human mind can be produced by something which follows the order of nature. At the start of Part III (see also Chapter 2 of his *Political Treatise*), Spinoza notes that traditional accounts of the passions, with the exception of Descartes's, have rested on the assumption--one wholly baseless in Spinoza's view--that human beings are a separate "dominion" within the dominion of nature, with different kinds of constituents and governed by different sorts of laws. Spinoza's project continues what he finds to be Descartes's important innovation: seeking "to explain human affects through their first causes." So his account of the affects may be most profitably compared to Descartes's in his *Passions of the Soul*. It may also be usefully compared to accounts in the writings of Hobbes (especially *Leviathan* VI), a contemporary who shared many of Spinoza's philosophical commitments, or to some of the "traditional accounts" which Spinoza faults, such as Aquinas's *Summa Theologiae*. (Aquinas's treatments of the passions appear mainly between Ia75 and 2a2ae189.)

Spinoza, though, because he denies freedom of the will, is more thorough than Descartes in his commitment to naturalism. This commitment makes the task Spinoza undertakes in the *Ethics* an even

more dramatic revision of traditional understandings of the passions than that which Descartes produced. So Spinoza, even more than Descartes, is open to the sort of objection which traditional authors, those to whom it seems beyond question that human beings are outside nature, might raise: how can the full range of human psychological phenomena be produced by natural causes? For the argument of the *Ethics* to succeed, Spinoza must produce, first, an account of how human desires and emotions might be a part of nature as he has presented it in the *Ethics* and, second, a description of those human desires and emotions which is plausibly complex, that is, plausibly consistent with our experience of ourselves. The subsections which follow address these issues in turn.

2.1 The Affects and Striving

The human affects, for Spinoza, are a part of nature insofar as each can be redescribed in terms of striving, a property which all particular things in nature share. Desire and its varieties are striving itself, under a certain description. Human passion, in whatever form it takes, is for Spinoza a decrease in the power with which the we strive. Active affects are each increases in the power with which we strive.

Spinoza introduces the first of his primary affects, desire, at IIIp9s, directly after introducing the doctrine of human striving, which, in its most general form, he calls appetite.

> III Proposition 9, Scholium: ...Between appetite and desire there is no difference, expect that desire is generally related to men insofar as they are conscious of the appetite. So desire can be defined as appetite together with consciousness of the appetite.

Thus Spinoza identifies human desire with human essence and especially with consciousness of one’s essence, the striving for perseverance in being. Spinoza’s theory of consciousness is notoriously incomplete, and it is not clear whether, on his view, things other than human beings have consciousness. For human beings, at least, however, what seems to us to cause us to act, our desire, does, on Spinoza’s view, do just that. If I am asked for the proximate cause of my action in picking up my coffee cup, for example, I will respond that it was my desire for the coffee. In identifying the cause of human action, striving, with conscious desire, then, IIIp9s vindicates common sense to a degree. Had Spinoza identified desire with something other than striving, then he would commit himself to the view that my desire does not in fact cause me to pick up the cup. (Desire for Spinoza, in its narrow definition at IIIp9s, is psycho-physical and, in its broader definition at III, Definitions of the Affects, I, may be either. So, this example, perhaps despite appearances, does not run afoul of Spinoza’s denial of mind-body interaction.)

IIIp9s, then, goes a long way toward showing how the universal striving doctrine can be the basis for an account of human desire. A serious problem remains, however. Although we tend to see desire as the proximate cause of action, we tend also to conceive of desire as involving teleology or final causes. If desire causes me to pick up the cup, what causes my desire? The common-sense answer is teleological: I have, as an end, coffee, and I am, in a sense, drawn toward it. Spinoza, though well-aware of the fact that we commonly suppose that there are teleological causes of our actions, denies their reality:

> What is called a final cause is nothing but a human appetite insofar as it is considered as a principle, or primary cause, of some thing. For example, when we say that habitation was the final cause of this or that house, surely we understand nothing but that a man, because he imagined the conveniences of domestic life, had an appetite to build a house. So habitation, insofar as it is considered as a final cause, is nothing more than this singular appetite. It is really an efficient cause, which is considered as a first cause, because men are commonly ignorant of the causes of their appetites.

While he can accommodate, in a general way, the common-sense view that desire causes action, then, Spinoza must deny the other common-sense view that desire operates teleologically, and must explain what the ends of human action are, why they seem to human beings to explain actions, and what their real relation is to the processes of efficient causation involved in human behavior. (For an argument against this view, the view that Spinoza denies all teleology, see Garrett 2.)

For these tasks, Spinoza introduces the other primary affects and a number of psychological laws associated with them. He introduces the primary passions at IIIp11s.

> III Proposition 11, Scholium: We see, then, that the mind can undergo great changes, and pass now to a greater, now to a lesser perfection. These passions, indeed, explain to us the affects of Joy [*laetitia*] and Sadness [*tristitia*]. By Joy, therefore, I shall understand in what follows that passion by which the mind passes to a greater perfection. And by Sadness, that passion by which it passes to a lesser perfection.

The perfectionist language Spinoza uses is important for an understanding of the basis for ethics that he finds in psychology. Here, however, it may be understood in terms of striving. An increased power to persevere in being is for Spinoza a transition to greater perfection and a decreased power is a transition to lesser perfection (see IIIp11, the end of IV Preface, and especially III, Definitions of the Affects, III, Exp.). So joy is the passion one experiences in the transition to an increased power to strive, and sadness is the passion one experiences in the opposite transition. Spinoza thus provides, in his account of the affects, the basis for an explanation of how it is that introspection into our conscious experience of desire might fail to bring us accurate knowledge of our own psychological processes. Our conscious experience, both in reacting to ourselves and external objects and also in forming our desires, has an emotional component: we experience joy and sadness and varieties of these. But we may be unaware of why we feel joy or sadness or why, really, we desire what we desire. So Spinoza writes repeatedly, in the context of his criticisms of teleological reasoning and the introspective experiences of free will or mind/body causation (e.g., at IIIp2s): "men are conscious of their actions and ignorant of the causes by which they are determined."

Spinoza offers an explanation of the apparent teleology in desire, in particular, at IIIp28:

> We strive to promote the occurrence of whatever we imagine will lead to joy, and to avert

> or destroy what we imagine is contrary to it, or will lead to sadness.

Spinoza reserves the term 'imagine' [*imaginor*] for the description of conscious states, so IIIp28 describes, at least in part, the objects of desire. If I imagine that coffee will lead to joy, then I will desire that joy and so that coffee. IIIp28, strictly speaking, is not an exhaustive characterization of objects of desire. It implies only that we desire anything which we imagine will lead to joy and are averse to whatever we imagine will lead to sadness and not that we might not have other kinds of desires also, desires unrelated to either joy or sadness. A review of the particular forms of desire Spinoza catalogues in Part III suggests, however, that the view is still stronger than the limited claim of IIIp28: it seems that Spinoza does hold that anything I desire will be a thing which leads to joy or sadness.

What seems on introspection, then, to be a teleological cause of action, the end represented by an object of desire, is for Spinoza a peculiar manifestation in consciousness of striving, which in turn is the genuine (i.e., efficient) cause of action. I reach for the cup of coffee, I may think, because the joy that I anticipate in the coffee "pulls" me to it; in fact, however, I reach for the coffee because my characteristic striving (perhaps as a partial cause in combination with other partial causes such as the memory of past cups--IIIp36) has that effect. It "pushes" me toward the cup.

2.2 The Variety of Affects

Perhaps the psychological view that Spinoza introduces at IIIp28 is susceptible to the sort of objection which one might raise against psychological hedonism, the view that human beings only desire pleasure, the avoidance of pain, and what is instrumental to these things. It may seem to some people that IIIp28 is not consistent with their own experience of their motives in acting. So, someone with a strong sense of justice might say:

> It's not that I like Jones or would get any joy from having him walk. I think the guy's a jerk, and I hate to think of him out on the street. But I want him to be released from prison. He simply did not do what he's been convicted of, so he should be set free.

On the basis of introspective observations like this one, one might complain that, even if Spinoza's account of the affects can be shown to be consistent with the general theory of striving as it is presented at IIIp6, still the theory of affects is not itself a realistically complex account of human desire, since it cannot account for desires like this one which do not have emotional components. The plausibility of Spinoza's view depends upon the extent to which it can reasonably redescribe this desire, and other similarly troubling desires,in ways which are consistent with IIIp28.

Spinoza attempts to show that there many varieties of joy, sadness, and desire. Thus he might attempt to address the complaint by showing that its author offers a slightly inaccurate description of the situation:

> The author denies liking Jones. Let us suppose even that he hates Jones. Even so, that does

> not mean that the author anticipates no joy at all in Jones's release. Knowing that his society is just in at least this one case may reassure the author of the complaint to some degree that he might be fairly treated himself. So it might be a kind of hope (IIIp18s2) which motivates the desire. Or perhaps there are people the author likes, his fellow citizens generally perhaps, to whom he wishes a similar peace of mind. The author may want the release in order to find a kind of joy, whether it be out if his ambition to please them or simply out of his human kindness (IIIp29s) or nobility (IIIp59s), in the well-being of these other people.

Far from insisting that there is one particular kind of emotion that moves people, Spinoza writes that there is an innumerable variety of affects:

> III Proposition 56: There are as many species of Joy, Sadness and Desire, and consequently of each affect composed of these (like vacillation of mind) or derived from them (like love, hate, hope, fear, etc.), as there are objects by which we are affected.

IIIp51 assures us, moreover, that the same object might affect different people, or even the same person at different times, in different ways. So Spinoza protects himself from the charge that IIIp28 is obviously false (albeit at the risk of forwarding an unfalsifiable psychological claim) by arguing that, despite the seeming simplicity of that proposition, it cannot be falsified by the great variety of conscious human motives.

Although Spinoza repeatedly insists that the variety of affects is innumerable, he nevertheless does characterize, in his own terms, many of the traditional passions, each of which is either a kind of joy, sadness or desire or a hybrid of two different affects. A few Spinoza's particular accounts are notable.

Pity (*commiseratio*) is for Spinoza a species of sadness, sadness that arises from injury to another (IIIp22s), and so to feel pity, on Spinoza's view is to experience a decrease in one's own power to persevere in being. If continued perseverance in being is what virtuous agents seek, then, Spinoza will be commmited to the view that pity is not a virtue. Indeed, Spinoza writes at IVp50c, "A man who lives according to the dictates of reason, strives, as far as he can, not to be touched by pity." So Spinoza stands apart from traditional Christian views on this subject (and also on the subjects of humility and repentence), and with Hobbes who conceives of pity in *Leviathan* VI as a kind of grief and so a kind of displeasure. This revisionary tendency in his thought is tempered, however, by IIIp54, where he presents pity, and also the other traditional Christian virtues of humility and repentence, as, if not genuine virtues themselves, at least means to virtue, by which people are made more able to come to learn to follow the dictates of reason.

Self-esteem (*acquiescentia in se ipso*) which Spinoza introduces at IIIp30 as Joy accompanied by the idea of oneself as an internal cause becomes an important part of Spinoza's ethical theory, a species of which is even blessedness (*beatitudo*, see IV, App. 4), the highest form of human happiness. Human beings, as finite modes cannot, on Spinoza's view, avoid affecting and being affected by external objects.

Nevertheless, Spinoza's emphasis on self-esteem, and, in his ethical theory, on self-knowledge, suggests that to the extent we are able to bring about effects, including our own emotions, as whole or adequate causes of those effects, we are better off. His remarks concerning the impossibility of controlling the passions and the desireability of controlling them nevertheless to the extent that we can (V Preface) similarly emphasize the ethical importance of self-knowledge and freedom from external influences.

Finally, **active joy** and **active desire** which Spinoza introduces at IIIp58 represent a separate class of affects notable both for their novelty against the background of traditional accounts of the passions and also for their importance to Spinoza's ethical arguments of Parts IV and V. On traditional accounts of the passions, even Descartes's (*The Passions of the Soul*, I.1), actions and passions are the same thing, regarded from different perspectives: when A does X to B, X is an action for A but a passion for B. For Spinoza, however, anything which follows in a person where that person is an "inadequate" or partial cause of the thing, is a passion, and anything that follows where a person is an "adequate" or whole cause of the thing is an action. Thus Spinoza's class of active affects places a strong emphasis on people's roles as whole causes of what they do; because it becomes for Spinoza ethically important that a person be active rather than passive, that emphasis raises a host of questions about the extent to which a person, a particular thing interacting constantly with other things and indeed requiring some of them for sustenance, can come to resist passion and guide himself by means of joy and the active desires.

Because joy and sadness as introduced at IIIp11s are passions, all of the desires arising from them or species of them are passive as well, that is, they are not desires which arise from a person's striving alone but only as a partial cause in combination with other, ultimately external causes. Active joy, which must include at least some types of warranted self-esteem, and active desires, among which Spinoza lists at IIIp59s tenacity (*animositas*) and nobility (*generositas*) are wholly active however; that is, they are emotions and desires that people have only insofar as they are adequate causes, or genuine actors. (Notice that sadness cannot ever be an active emotion. People cannot, insofar as they are active bring it about that their power of acting is decreased, so passive sadness, unlike passive joy and desire, has no active counterpart.)

Of these active affects, the most important for an interpretation of Spinoza's ethics and political philosophy is likely nobility. Spinoza's predominant egoism, together with some of his still stronger statements of psychological egoism such as that at I Appendix, suggest that individuals are not, or are not often, altruistic. Moreover, his ethics, with its emphasis on self-esteem and self-knowledge appears in ways to be an individualistic one: the good, when I attain it, is a perfection of myself, not of society or the world. However, nobility, as Spinoza defines it, is a wholly active desire to join others in friendship and to aid them. Spinoza needs, in his ethics, to explain how aiding others is virtuous, and, in his political theory, to explain why a person, even a rational one, would want to come to the aid of others.

Further reading: For discussions of Spinoza's theory of affects in a comparative framework, see Hoffman and James (2). For discussions of the relationship between striving and the affects, see these works and also Davidson,and Schrijvers. Bennett (2), Curley (2), Della Rocca and Garrett (2) discuss Spinoza's views on teleology. Lloyd offers accounts of various particular affects. For a detailed discussion of self-esteem, see Rutherford.

3. The Psychological Basis for a Theory of Value

Spinoza’s insistence that human beings not be treated as a dominion within a dominion includes a commitment to ethical naturalism also. Just as he insists that the human mind must be explicable in terms of the laws which govern nature, so he insists that ethical properties, which he sometimes characterizes as human "modes of thinking," be explicable in terms of natural ones. The theory of the affects serves Spinoza’s ethical naturalism by introducing explanations of ethical concepts, most importantly the concepts of good, evil, and perfection, in psychological terms. In his ethics, Spinoza in some way "retains these words," although he may be understood to do so under some formal refinement or revision of them (See IV Preface). So his discussions of good and evil and of human perfection in Part III provide the basis for the formal ethical argument which follows in Parts IV and V.

3.1 Good and Evil as Modes of Thinking

Before defining ‘good’ and ‘evil’ formally, Spinoza at IV Preface regards good and evil as labels, "modes of thinking," that human beings apply to things but which really reveal little about the things to which they are applied:

> As far as good and evil are concerned, they also indicate nothing positive in things, considered in themselves, nor are they anything other than modes of thinking, or notions we form because we compare things to one another. For one and the same thing can be good, and [evil], and also indifferent. For example, Music is good for one who is melancholy, [evil to] one who is mourning, and nether good nor [evil] to one who is deaf.

The phrase "nothing positive in things" means perhaps that an observer of people would find that ‘good’ and ‘evil’, as people use them are two place predicates rather than one place predicates. If Martha calls music evil, then, what that indicates to one who knows about the human use of these terms is that the music is evil to Martha. Moreover, since the same music can be good or evil for different people, or for people in different states, the two place predication reveals more about Martha than about the music. It must be some fact about the person, rather than some fact about the thing called good or evil, that is of central importance to the understanding of the label.

IIIp9s suggests that the fact about the person which the label reveals is her conative state:

> It is clear that we neither strive for, nor will, neither want, nor desire anything because we judge it to be good; on the contrary, we judge something to be good because we strive for it, will it, want it, and desire it.

Spinoza finds that the designation of a thing as good follows from a person’s conative state: Martha is

averse to music and therefore she calls it evil. Should the music become good to another person, or perhaps Martha herself in different circumstances, it would not be because the music has changed, but because the person's conative state is different: she desires the music. (This analysis of the good is remarkable similar to Hobbes's at *Leviathan VI*. Maimonides, another of Spinoza's influences, also has a similar analysis: *Guide of the Perplexed*, III, 13.)

IIIp39s offers a similar analysis. There Spinoza writes, in the same vein as IIIp9s, that, "each one, from his own affect, judges, or evaluates, what is good and what is [evil]...So the greedy man judges an abundance of money best, and poverty worst. The ambitious man desires nothing so much as esteem and dreads nothing so much as shame." However, IIIp39s also uses Spinoza's theory of the affects to introduce new definitions of good and evil:

> By good here I understand every kind of joy, and whatever leads to it, and especially whatever satisfies any kind of longing, whatever that may be. And by evil, every kind of sadness and especially what frustrates longing.

IIIp28, the proposition establishing Spinoza's doctrine that human beings desire whatever will bring joy and are averse to whatever will lead to sadness, allows Spinoza to connect the objects of any human desire with joy or the avoidance of sadness. So, if it is true that we call a thing good only if we desire it, then it will also be true that anything we call good will be joy or what leads to it. Understood in this way, IIIp39s simply restates the doctrine of IIIp9s in light of IIIp28.

However, Spinoza might be extending rather than merely restating his position at IIIp39s. Every kind of joy we experience is not presumably a result of conscious desire, and Spinoza allows at IIIP39s that these instances of joy (i.e., those which do not satisfy any kind of longing) are also good. So not only is whatever Martha desires good for her, but, in addition, anything which she does not desire but which nonetheless might bring her joy will also be good. IIIp39s therefore identifies the good and evil for a person with broader classes of things, and makes possible an analysis of good and evil in terms of something other than an individual person's current desires. Because, in giving an account of the right way of living in Parts IV and V, Spinoza presumably urges people to desire and do things in a way different from what they desire and do already, this broadening of the application of the terms 'good' and 'evil' (to apply to things other than what people presently desire or are averse to) contributes to the plausibility of his ethical naturalism.

3.2 The Psychological Basis for Perfectionism

The meaning of 'perfection' (*perfectio*) in various passages of the *Ethics* is obscured by the fact that Spinoza uses the term both in what he takes to be its common meaning and also in a narrow, formal sense. In some passages, such as IV Preface, Spinoza treats perfection, like good and evil, to be a label that people apply to things, to be explained in terms of those people's use of the label: people call a thing perfect which conforms to the model of the thing that they create for themselves. In other passages, however, Spinoza treats perfection as whatever is "positive" or "real" in a thing, a genuine propety, and

this concept forms part of his formal apparatus (IId6).

Both senses of the term occur in Spinoza's analysis of human perfection. At IV Preface Spinoza invokes a concept of perfection based upon a model of human nature that we set before ourselves. And Spinoza describes perfection of the human mind in terms of its power of thinking, as we have already seen, at IIIp11 and its scholium: the mind's power of thinking is its perfection; joy is an increase in that power or a passage to a greater perfection; and sadness is a decrease in that power or a passage to a lesser perfection.

Thus Spinoza has two different accounts of human perfection which might contribute to the perfectionist language of the ethical argument that he develops in various ways in Parts IV and V. Although the first account contributes to both Spinoza's formal definitions of good and evil and also, probably, to the "free man" propositions of Part IV, it is perhaps the second account which is of greater importance. It, after all, resonates with Spinoza's account of the good at IIIp39s, and so comes to form part of a consistent moral view: if the good is any form of joy (IIIp39s) and joy is the passage to a greater perfection (IIIp11s), then the good is whatever makes us more perfect. On the other hand, the first account might imply a different set of norms from IIIp39s: it seems that we may have different models of human nature that we set before ourselves, and that they may or may not include the various things that give us joy. Perhaps the formal account of perfection at IId6 and later at IIIp11 give Spinoza a means of reformulating the idea of perfection as a model of human nature in a way which reconciles the two senses of the term: the ideal we set before ourselves will be a person who possesses the greatest possible power of action. This would be, in effect, to correlate our systematically distorted ways of perceiving ourselves--as free agents pursuing as an end a model of human nature--with the causes that really determine our actions.

Further Reading: Curley (1), Delahunty, Garrett (1) and Yovel offer accounts of Spinoza's ethical language as it relates to his psychology. For a discussion of Spinoza's views on perfection, see Garrett (1), Allison and Wolfson.

Bibliography

- Allison, H. (1987). *Benedict de Spinoza: An Introduction*. New Haven: Yale University Press.
- Bennett, J. (1984). *A Study of Spinoza's Ethics*. Indianapolis: Hackett.
- ------ (1990). "Spinoza and Teleology: A Reply to Curley," in *Spinoza: Issues and Directions*, ed. Edwin Curley and Pierre-Francois Moreau. Leiden: Brill.
- Cicero. (1933). *De Natura Deorum*. Cambridge: Loeb Classical Library.
- Curley, E. (1988). *Behind the Geometrical Method*. Princeton: Princeton University Press.
- ------ (1990). "On Bennett's Spinoza: The Issue of Teleology," in *Spinoza: Issues and Directions*, ed. Edwin Curley and Pierre-Francois Moreau. Leiden: Brill.
- Delahunty, R. (1985). *Spinoza*. London: Routledge and Kegan Paul.
- Davidson (1999). "Spinoza's Causal Theory of the Affects," in *Desire and Affect: Spinoza as Psychologist*, ed. Yirmiyaho Yovel. New York: Little Room Press.
- Della Rocca, M. (1996). "Spinoza's Metaphysical Psychology," in *The Cambridge Companion to*

Spinoza, ed. Don Garrett. Cambridge: Cambridge University Press.

- Descartes, R. (1985). *The Philosophical Writings of Descartes*, vol 1, trans. by John Cottingham, Robert Stoothoff and Dugald Murdoch. Cambridge: Cambridge University Press.
    - [This book includes both the *Principles of Philosophy* and *The Passions of the Soul*.]
- Garrett, D. (1991). "Spinoza's Ethical Theory," in *The Cambridge Companion to Spinoza*, ed. Don Garrett. Cambridge: Cambridge University Press.
- Garrett, D. (1999). "Teleology in Spinoza and Early Modern Rationalism," in *New Essays on the Rationalists*, ed. Rocco Gennaro and Charles Huenemann. Oxford: Oxford University Press.
- Hampton, J. (1986). *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press.
- Hobbes, T. (1839-1845). Thomae Hobbes Malmesburiensis -- Opera Philosophica quae Latine scripsit Omnia, ed. Sir William Molesworth. 5 Volumes. London: John Bohn.
- ------. (1839-185). The English Works of Thomas Hobbes, ed. Sir William Molesworth. 11 Volumes. London: John Bohn.
- Hoffman, P. (1991). "Three Dualist Theories of the Passions," Philosophical Topics, vol. 19, No. 1.
- James, S. (1993). "Spinoza the Stoic," in *The Rise of Modern Philosophy*, ed. Tom Sorell. Oxford: Clarendon Press.
- ------.(1997). *Passion and Action: The Emotions in Seventeenth-Century Philosophy*. New York: Oxford University Press.
- Jarrett, C. (1991). "Spinoza's Denial of Mind-Body Interaction and the Explanation of Human Action," The Southern Journal of Philosophy, 29(4).
- Kavka, G. (1986). *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.
- Lachterman, D. (1978). "The Physics of Spinoza's Ethics," in *Spinoza: New Perspectives*, ed. Robert Shahan and J.I. Biro. Norman: University of Oklahoma Press.
- Lloyd, G. (1994). *Self-Knowledge in Spinoza's 'Ethics'*. Ithaca: Cornell University Press.
- Maimonides, M. (1963). *The Guide of the Perplexed*, trans. Shlomo Pines. Chicago: University of Chicago Press.
- Matson, W. (1977). "Death and Destruction in Spinoza's Ethics," Inquiry, vol. 20.
- Rutherford, D. (1999). "Salvation as a State of Mind: The Place of Acquiescientia in Spinoza's Ethics," British Journal for the History of Philosophy, 7(3).
- Schrijvers, M. (1999). "The Conatus and the Mutual Relationship Between Active and Passive Affects in Spinoza," in *Desire and Affect: Spinoza as Psychologist*, ed. Yirmiyaho Yovel. New York: Little Room Press.
- Spinoza, B. (1925). *Spinoza Opera*, ed. Carl Gebhart. 4 volumes. Heidelberg: Carl Winter.
- ------. (1985). *The Collected Works of Spinoza*. Vol. I, ed. and trans. Edwin Curley. Princeton: Princeton University Press.

    [I use Curley's translation for the passages quoted here. Students of Spinoza in English should be especially attentive to the fact that different translators give different translations for Spinoza's particular affects.]
- Thomas Aquinas, Saint (1964). *Summa Theologiae*. New York: McGraw-Hill.
- Wolfson, H.A. (1934). *The Philosophy of Spinoza*. Cambridge: Harvard Univeristy Press.
- Yovel, Y. (1999). "Transcending Mere Survival," in *Desire and Affect: Spinoza as Psychologist*,

ed. Yirmiyaho Yovel. New York: Little Room Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Descartes, René | Hobbes, Thomas | Spinoza, Baruch [Benedict]

Michael LeBuffe
lebuffe@philosophy.tamu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 22, 2001
Content last modified: October 22, 2001

Law and Ideology

If law is a system of enforceable rules governing social relations and legislated by a political system, it might seem obvious that law is connected to ideology. Ideology refers, in a general sense, to a system of political ideas, and law and politics seem inextricably intertwined. Just as ideologies are dotted across the political spectrum, so too are legal systems. Thus we speak of both legal systems and ideologies as liberal, fascist, communist, and so on, and most people probably assume that a law is the legal expression of a political ideology. One would expect the practice and activity of law to be shaped by people's political beliefs, so law might seem to emanate from ideology in a straightforward and uncontroversial way.

However, the connection between law and ideology is both complex and contentious. This is because of the diversity of definitions of ideology, and the various ways in which ideology might be related to law. Moreover, whilst the observation about law's link with ideology might seem a sociological commonplace, the link between law and ideology is more often made in a critical spirit, in order to impugn law.

At issue is an understanding of ideology as a source of manipulation. Law as ideology directs its subjects in ways that are not transparent to the subjects themselves; law, on this view, cloaks power. The ideal of law, in contrast, involves a set of institutions that regulate or restrain power with reference to norms of justice. Thus the presence of the ideological in law must, in some sense, compromise law's integrity. Not only is the view of law as ideology at odds with a lot of mainstream thinking about law, it seems difficult to reconcile with the central philosophical positions on the nature of law, e.g. a positivist conception of law as a set of formal rules, or a natural law conception where law is identified with moral principles.

1. Liberal Concepts of Ideology

What is ideology? The term was likely coined by the French thinker Claude Destutt de Tracy at the turn of the nineteenth century, in his study of the Enlightenment. For De Tracy, ideology was the science of ideas and their origins. Ideology understands ideas to issue, not haphazardly from mind or consciousness, but as the result of forces in the material environment that shape what people think. De Tracy believed his view of ideology could be put to progressive political purposes, since understanding the source of ideas might enable efforts on behalf of human progress.

Ideology today is generally taken to mean not a science of ideas, but the ideas themselves, and moreover ideas of a particular kind. Ideologies are ideas whose purpose is not epistemic, but political. Thus an ideology exists to confirm a certain political viewpoint, serve the interests of certain people, or to perform a functional role in relation to social, economic, political and legal institutions. Daniel Bell dubbed ideology 'an action-oriented system of beliefs,' and the fact that ideology is action-oriented indicates its role is not to render reality transparent, but to motivate people to do or not do certain things. Such a role may involve a process of justification that requires the obfuscation of reality. Nonetheless, Bell and other liberal sociologists do not assume any particular relation between ideology and the status quo; some ideologies serve the status quo, others call for its reform or overthrow.

On this view, ideology can shape law, but a variety of ideologies might be vying for legal mastery; there is no necessary connection between law and a particular ideology. Law need not be understood as compromised, since law being ideological might just refer to the institutions of popular sovereignty, where public policy reflects citizens' principles and beliefs; ideology would in that case just be a shorthand way of referring to the views of citizens that are legitimately instantiated in the laws of the land. Nonetheless, Bell argued that a postwar consensus on capitalism and liberal democracy might spell the 'end of ideology.'

2. Radical Concepts of Ideology

A more critical understanding of law's relation to ideology, and the role and purposes that ideology serves, is found in the writings of Karl Marx and Friedrich Engels. Like De Tracy, Marx and Engels contend that ideas are shaped by the material world, but as historical materialists they understand the material to consist of relations of production that undergo change and development. Moreover, for Marx and Engels, it is the exploitative and alienating features of capitalist economic relations that prompt ideas they dub 'ideology.' Ideology only arises where there are social conditions such as those produced by private property that are vulnerable to criticism and protest; ideology exists to inure these social conditions from attack by those who are disadvantaged by them. Capitalist ideologies give an inverted explanation for market relations, for example, so that human beings perceive their actions as the consequence of economic factors, rather than the other way around, and moreover, thereby understand the market to be natural and inevitable. Members of the Frankfurt School such as Jurgen Habermas drew

on the Marxist idea of ideology as a distortion of reality to point to its role in communication, wherein interlocutors find that power relations prevent the open, uncoerced articulation of beliefs and values.

Thus ideology, far from being a science, as De Tracy contends, or any set of action-oriented beliefs as Bell puts it, is rather inherently conservative, quietist, and epistemically unreliable. Ideology conserves by camouflaging flawed social conditions, giving an illusory account of their rationale or function, in order to justify and win acceptance of them. Indeed, on this view of the ideological role of law, in a just society there would be no need for a mystifying account of reality, and thus no need for law. The concept of law as ideology is thus central to the Marxist view that law will wither away with the full flowering of communism.

The negative view of ideology taken by Marxists might suggest a crude conception where legal ideology is a tool cynically wielded by the powerful to ensure submission by the powerless. However, it offends the 'conception of right,' Marx argues, if 'a code of law is the blunt, unmitigated, unadulterated expression of the domination of a class.' And because ideology such as law takes a formal and normative form, the powerful are in its grips too, persuaded by an account of the inevitable and just order from which they profit. Moreover, ideology is no mere fiction; it is produced by real social conditions and reflects them. Ideology thus must succeed in constituting a consensus about capitalism, and it must do so by giving expression to capitalism's recognizable features. Equality before the law, for example, is both elicited by, and reflects, the reality of capitalist economic relations, even if it is an equality that is formal and incomplete. Consent will not be forthcoming if legal ideology bears no relation whatsoever to the social conditions it seeks to justify. The idea that ideology inverts reality is important here. In his camera obscura metaphor in *The German Ideology*, Marx contends that reality appears upside down in ideology, much like the photographic process provides an inverted image. The inverted image is telling; it is a recognisable depiction of reality, even if it is at the same time a distorted one. Karl Mannheim elaborated further on the idea of the complex relation between reality and ideology by pointing to the human need for ideology. Ideologies are neither true nor false but are a set of socially conditioned ideas that provide a truth that people, both the advantaged and the disadvantaged, want to hear.

In the 1920s, American jurisprudence came under the influence of another version of the critical view of ideology and law. The school of legal realism abandoned Marx's specifically historical materialist explanation, but took up the idea that social forces outside the law are central in determining what the law is. Realists opposed traditional 'formalist' accounts of adjudication, where judges are understood to rely on uniquely and distinctively legal materials in rendering their judgments. Instead, the realists contended that law is inherently indeterminate, and thus judicial decisions must be explained by factors outside the law. Ideology emerges as one kind of realist explanation, where judicial decisions are the effect of political ideas, be they of the judge, the legal profession more generally, societal elites, or majority public opinion. The realists aligned their critique of law with a progressive politics. The inevitable influence of factors external to the law meant that social and political changes augured by the emerging welfare state were no threat to the purity of law. Indeed, the expanding regulative power of the administrative state would make it more likely that the influences on the law were now those of popular sovereignty and social justice, rather than the more nefarious influences of the past.

The view that law is a reflection of ideology was taken up again in the 1970s and 80s, with the emergence of the Critical Legal Studies movement. Critical Legal Studies was a radical movement of lawyers shaped by a number of influences: the Marxist and realist traditions; the philosophical perspective of 'deconstruction;' and the politics of issues such as feminism, environmentalism and anti-racism. The movement takes up the realist idea that law is fundamentally indeterminate, and echoes Marxist views about the ways in which the interests of the powerful shape law. Exponents offer some astute observations about the way law is taught and practiced to give the misleading impression of law's certainty and legitimacy. Particular legal doctrines are targeted for papering over the inconsistent and arbitrary features of legal decision-making; the rule of law, for example, is criticized for a naïve view of the form of law as unaffected by law's content and the social context in which law operates. The indeterminacy of law can produce a variety of results; Duncan Kennedy, for example, points out the surprising ways in which the ideology of formal legal reasoning can remedy injustice, even if ideology often disables such remedies as well.

3. Ideology and the Sources of Law

The well-known debate about the sources of law appears to be radically undercut by a view of law as ideology. The sources debate has usually been posed in terms of the extent to which morality is intrinsic to the definition of law. Natural lawyers argue that what is law must partly depend on moral criteria. Following Thomas Aquinas, the traditional criteria have not strayed far from the teachings of the Roman Catholic Church, but more recent natural law arguments, such as those of Lon Fuller and Ronald Dworkin, have proffered secular standards emanating from the procedural ideals of the rule of law or the constitutionalism of American liberalism. All natural lawyers, however, are agreed that what the law is must be determined, in some sense, by what the law ought to be.

Positivists, in contrast, have argued that what is law is determined only by the institutional facts internal to a legal system, facts that may or may not meet moral standards. Early positivists, such as Thomas Hobbes and John Austin, argued that even the legitimacy of law did not depend on moral criteria; law must be obeyed, however much it falls short of moral ideals. More recent exponents, such as H.L.A. Hart and Joseph Raz, have offered as a justification for the positivist position the idea that because what is law is a factual question, law's legitimacy can be determined by moral criteria outside the law that might recommend disobedience. All positivists, however, are agreed that what the law is and what it ought to be should be kept distinct.

The natural law and legal positivist positions are united, however, in the aim to provide a concept of the essence of law. This endeavour supplies them with a common enemy in the view of law as ideology, which finds trying to determine the essence of law as fundamentally misconceived. After all, if law is inevitably shaped by ideas emanating from power relations outside of the law, then law has no essence, be it moral or institutional. If law is reduced to ideology, or seen as its mere effect, then legality looks contingent and unprincipled, having no necessary content or definition, no intrinsic character. If law both mirrors and distorts the realities of power, it is power, not principles of legality, which tell us what law is. Thus for most mainstream legal theorists, the ideological is no necessary feature of the law, and law

should certainly not be defined according to the radical conception where intrinsic to law is a mystification of reality, or an obfuscation of social relations in order to exact compliance.

The picture is more complicated, however. The ideology position does, after all, have some affinities with rival views on the sources of law. The ideology view concedes to the positivist, for example, that law emerges from the practices of society, though the practices are extra-legal -- political, economic and social -- rather than the practices of institutional facts internal to a legal system. Social forces are ultimately determining of the content and form of a legal system. Indeed, the Marxist Louis Althusser's idea of ideological state apparatuses has a definite positivist flavour in its insistence that political reality can be exhaustively described by reference to structures rather than norm-bearing agents. We might expect that the radical exponent of ideology would resist the combination of a positivist-ideology view. The radical would find in the positivist emphasis on institutions a too uncritical attitude to the ideological structures that shape those institutions. But it seems possible that the positivist position could be revised to remove any ascribing of legitimacy to the institutions that define law in order to accommodate the critique of the radical ideology position.

As for the natural law position, the ideology view concedes to the natural lawyer that law is normative. What is ideology, after all, but a set of values and ideals? However, on the ideology view, the norms are defined in terms of the interests they serve, rather than the justice they embody. Law is normative, but it is certainly not moral, the ideology view insists against the natural law view. The critical aspect of the radical ideology view suggests an impasse between the natural lawyer and the ideology position that is more difficult to overcome than in the positivist case.

Of course, natural lawyers and positivists could quite easily find room for the liberal view of ideology as an action-oriented system of beliefs as a supplement to their views about the sources of law, in the sense that ideology is part of the sociological landscape to which their concepts of law apply. Natural law can find popular expression in a society's ideology, and positivist legal institutions might reflect ideological beliefs.

4. Ideology and the Rule of Law

All this points to another and related tension. This is the tension between the ideology view and the concept of the rule of law, the centrepiece of a liberal legal order. At their most basic, the terms the rule of law, due process, procedural justice, legal formality, procedural rationality, justice as regularity, all refer to the idea that law should meet certain procedural requirements so that the individual is enabled to obey it. These requirements center on the principle that the law be general, that it take the form of *rules*. Law by definition should be directed to more than a particular situation or individual; the rule of law also requires that law be relatively certain, clearly expressed, open, prospective and adequately publicised.

The view of law as ideology, even in its radical variants, would not deny the presence of the rule of law in the liberal legal order; indeed, the rule of law is often invoked as a paradigmatic example of legal ideology. This is because, however, the rule of law is interpreted as a device that serves the interests of

the powerful; moreover, it is a device that dissembles itself. The rule of law, in its restraint on the exercise of governmental and judicial power, facilitates the aims of those with power of other kinds, particularly economic power. This is not a surprising argument, if one considers how right-wing thinkers like Frederick Hayek have lauded the rule of law for its essential role in buttressing the free market. Left wing and right wing thinkers are agreed, then, on the capitalist function of the rule of law.

For the left wing theorist of ideology, however, the rule of law also has ideological aspects that mean it serves capitalist purposes in more sinister ways. For in its restraint on political and legal power, the rule of law implies that these public forms of power are the only forms of power that exist, or at least the only ones that matter. Moreover, in assuring the subjects of the law that that law is applied with generality and certainty, the rule of law also implies that formal justice is the only relevant kind of justice; that equality before the law is identical to equality per se.

These claims about the rule of law and ideology are complex and need careful scrutiny. Does the rule of law necessarily involve manipulation on behalf of the capitalist order? Given its formal virtues, and its agnosticism on the content of law, the rule of law seems innocent of charges of a capitalist bias, or a bias of any kind. As Raz puts it, the rule of law's virtue is like the virtue of a sharp knife; it enables the law to fulfill its function, whatever the function might be. Moreover, it is hard to see how the rule of law itself is engaged in any project of deception. Generality in the law, for example, does not necessarily entail any particular commitments on how the economy or society should be organized; nor does it propagate falsity or error. Nonetheless, it is true that the proceduralism of the rule of law can be put to ideological purposes, to deflect social criticism and prevent radical change. And if enthusiasts of the rule of law place enough emphasis on procedural justice, this can reduce the likelihood that more substantive conceptions of justice will have success. Historically, societies governed by the rule of law have tended to be structured by capitalist markets, suggesting an affinity between the two sets of institutions. The rule of law can have an ideological effect even if it is not ideological in its essence.

5. Conclusion: Ideology and Justice

The idea that law is ideological is an important contribution to legal scholarship. First, it enables a more critical view of the law and its role, demystifying a set of vital social institutions. Second, it points to the importance of sociological and political factors in our understanding of the law. Legality is shaped and influenced by non-legal aspects of society, and law, in turn, has an impact on society and social change, not just in the obvious effects of particular judgments, but in the political culture that a legal system helps produce.

The ideology view risks, however, an unhelpful reductionism. Conceiving of law as ideological above all else can promote a crude and erroneous understanding of the relation between power and legality, where law serves only the interests of the powerful and where legal guarantees are mere shams. Moreover, this can license a cynicism about the law that is paradoxically contrary to the emancipatory aims of the radical politics that was the impetus for the critique of law as ideology in the first place. That is, radical critics risk dismissing altogether the possibility of legal recourse for remedying injustice.

Furthermore, the cynicism of the ideology view is in fact the fruits of a kind of utopianism about law, for it counters the bleak portrait of legal ideology manipulated on behalf of the powerful with an ideal society without ideology or law, where human beings' relations to each other and to reality are transparent and conflict-free. The 'end of ideology' thesis, advanced by Bell in a triumphalist spirit on behalf of liberal capitalism, but interestingly even more salient in Marxist ideals of communism, might be wrong in its assumption that human beings can transcend ideology. Indeed, the radical concept of ideology casts doubt on the likelihood that individuals' beliefs can provide an objective account of reality, untainted by distorted and self-justifying processes of inquiry.

How then, can the concept of ideology be deployed in legal scholarship? In fact, the more subtle critiques of ideology grasp the extent to which both liberation and manipulation can be embodied in the law. Recall the nuanced conception of Marx and Engels, where ideology gives an inverted image of reality, but a recognizable image nonetheless. This suggests that the ideals of legality are not a mere charade but are instantiated in the law, if only in a partial and incomplete form. The Marxist historian E.P. Thompson made this point in his argument for the universal value of the rule of law. Thompson contended that in order for law to function as ideology, it must proffer some genuine moral value.

To illustrate, consider how someone's cruelty might be masked by polite manners; this does not demonstrate that good manners have no worth. Legal ideology, too, might paper over injustice in ways that serve justice nonetheless. A functional argument about ideology, then, must concede the value of the phenomenon that serves ideological aims. Ideology cannot be devoid of emancipatory aspects altogether; if law trumpets justice, equality and freedom, then it must succeed in realizing these ideals, however imperfectly, in order for law to function as ideology. We can thus appreciate legal guarantees of a procedural kind for the genuine protection they offer the subjects of the law, whilst at the same time conceding the quietist politics that proceduralism might engender.

Moreover, understanding the ideological role of law need not undercut other conceptions of how law is to be defined or understood. This is particularly so if we recognize the unlikelihood of eliminating altogether ideological modes of understanding. A conception of law as having a moral source, or a source in a system's institutions, can be independent of a realistic appraisal of law's ideological function, or the ideological process in which laws are made. Both positivists and natural lawyers, so long as they do not insist that their conceptions of law are exhaustive of law's reality, can permit the influence of ideology, even in its more radical interpretations. Law can be ideology as well as other moral or institutional phenomena at the same time; indeed, law will probably not succeed as ideology unless it is multi-dimensional in just this way.

Bibliography

- Fisher, W.W. *et al.*, 1933, *American Legal Realism*, New York: Oxford University Press
- Kennedy, D., 1976, 'Form and Substance in Private Law Adjudication,' *Harvard Law Review*, Vol. 89

- Mannheim, K., 1936, *Ideology and Utopia*, New York: Harcourt, Brace and World
- Marx, K. and Engels, F., 1976, *The German Ideology*, Collected Works, Vol. 6, London: Lawrence and Wishart
- Sypnowich, C., 1990, *The Concept of Socialist Law*, Oxford: Clarendon

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Engels, Friedrich | Frankfurt School | liberalism | Marx, Karl | nature of law: legal realisms | nature of law: natural law theories | positivism | rule of law and procedural fairness | socialism

Christine Sypnowich
cs4@post.queensu.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: October 22, 2001
Content last modified: October 22, 2001

Antoine Le Grand

Antoine Le Grand (1629-1699) was a philosopher and catholic theologian who played an important role in propagating the Cartesian philosophy in England during the latter half of the seventeenth century. He was born in Douai, (at the time under rule by the Spanish Hapsburgs), and early in life was associated with an English community of Franciscans who had a college there. Le Grand became a Franciscan Recollect friar prior to leaving for England as a missionary in 1656. In England, he taught philosophy and theology, advocating Catholicism and eventually Cartesianism, the latter being as unpopular as the former was perilous. It is not clear how Le Grand came to Cartesianism, but the first evidence of his adoption of the new philosophy was in his *Institutio Philosophiae*, published in London in 1672. His early works show affinities to the philosophies of Seneca and Epicurus. He is noted for his polemical exchanges with Samuel Parker and John Sergeant, and for having given Descartes's work a Scholastic form so that it would be accepted in the schools.

1. Life and Writings

Le Grand lived in London for many years before retiring to Oxfordshire towards the close of his life in 1695. He was generally well-received at the University of Cambridge, perhaps owing to the influence of some leading Neo-Platonists at Cambridge such as John Smith, Henry More, and Ralph Cudworth, who were at least initially sympathetic to Descartes's ideas. John Smith, author of *Select Discourses* (1660), and the earliest recorded partisan of the Cartesian philosophy in England, was the first to introduce the study of Descartes to Cambridge. Henry More corresponded with Descartes, and was sympathetic until about 1665, when he launched the most vigorous attack on Cartesianism in the age, in his *Enchiridion Metaphysicum* (1671). Cudworth, like More although with less venom, objected to Descartes's

mechanistic account of the material world in his *The True Intellectual System of the Universe* (1678). It was Le Grand who debated and defended Descartes's philosophy to these English critics.

It is likely that Le Grand's edition of *Jacobi Rohaulti tractatus physicus* (1682) was the first of his Cartesian works accepted at Cambridge. This was a new edition, which included Le Grand's commentary, of Bonnet's 1672 Latin translation of Rohault's immensely popular tract of the Cartesian physics, *Traité de physique* (1671). In 1692, Samuel Clarke published his own Latin version of the text, which incorporates his and Le Grand's commentary in the form of footnotes. Clarke, though an adherent of Newton's physics, thought he could best propagate the new doctrine by publishing Rohault's text with suggestive notes directed at the necessity of modifying Cartesian theory. According to the biographical preface to Clarke's *Works*, at the time of his entrance to Cambridge in 1691, Rohault's *Traité* was the standard modern scientific text, and Newton's *Principia* (1687), had not yet been accepted: "The philosophy of Des Cartes was then the Established philosophy of that university, and the system of nature hardly allowed to be explained any otherwise than his principles … . The Great Sir Isaac Newton had indeed then published his Principia. But this book was but for the few." Eventually Clarke's translation of the *Traité*, which underwent four editions, became the new preferred Cambridge textbook, as the Cartesian physics gave way to that of Newton early in the eighteenth century.

At Oxford, Le Grand received a hostile reception. Samuel Parker aligned Hobbes's mechanism with that of Descartes, charging both with atheism. Parker's condemnations led to the banning of Descartes's philosophy at Oxford, quashing its public entrance at the University. Le Grand responded to Parker's charges of atheism in his *Apologia de Descartes* (1679), challenging Parker's criticisms with various proofs of God's existence. Another long-time, Oxford critic of Le Grand was the English secular priest and Aristotelian, John Sergeant. Sergeant, best known for his criticisms of Locke's philosophy, was also highly critical of the Cartesian philosophy. Le Grand responded to Sergeant's criticisms of the Cartesian criterion of truth in his *Dissertatio de ratione cognoscendi* … in 1679. A second major controversy occurred between the two authors late in Le Grand's life, this time over the nature of ideas. This dispute led Le Grand to write a series of short pieces, published later as *Several Smaller Pieces Against M. J. Sergeant* (1698). In response, Sergeant attacked the Cartesian idea of extension, to which the aging Le Grand never publicly responded. Le Grand died at the home of a wealthy farmer in Oxfordshire, where he had served as a tutor until his death in 1699.

Antoine Le Grand's most substantial work, *An Entire Body of Philosophy According to the Principles Of the Famous Renate des Cartes* (1694), is a Cartesian tract from beginning to end. It is based on a Latin text that underwent four editions before being translated into English by Richard Blome, and it includes corrections, alterations and additions by Le Grand himself. This work, as the title reveals, consists of three books. The first book, *The Institution*, is intended as a treatment of the general nature of things according to Descartes's principles; book two, *The History of Nature*, illustrates, by means of a great variety of reported experiments and examples, the operation of these first principles in nature. In this book, Le Grand applied the general Cartesian principles to his study of particular bodies and their qualities, showing how such principles can explain all natural phenomena. His extensive discussion includes bodies as various as the loadstone, plants, and insects. And finally, in his third book, *A Dissertation of the Want of Sense and Knowledge in Brute Animals*, he argued against the supposed link

of life and sense from Plato onwards, and after offering a brief survey of various hypotheses on the nature of soul by Aristotle, Gassendi, Fabri, and Descartes, he adopted Descartes's view. In the Preface, Le Grand wrote that, " … this whole work contains nothing else, but his [Descartes's] opinions, or what may clearly and distinctly be deduced from them."

2. Metaphysics

Le Grand defended the first principles of Descartes's philosophy with great fidelity. He held to Descartes's views that the essence of matter is extension; that the essence of mind is thought; that material substance and mental substance are essentially, and really distinct; that mind and body interact; that while humans have souls, brutes and other living things are mere machines; and that material things operate by means of moving parts according to the laws of motion. Le Grand did not make any substantial revisions to Cartesian metaphysics. However, he did make two important contributions: first, he attempted to clarify Descartes's account of motion which had direct consequences for the Cartesian account of matter, causation, and mind-body interaction; secondly, he extended the scope of Cartesian physics, treating such subjects as metals, plants, insects, animals, and the human body in detail.

Le Grand's contribution to the Cartesian account of motion may either be seen as an extension or a revision of Descartes's, sometimes ambiguous, treatment. Le Grand took seriously the claims that God is the total and efficient cause of motion in the universe, and that matter is entirely passive, and hence bodies are incapable of self-movement, or of moving other bodies. In his *Entire Body of Philosophy*, he argued that since a body may be in motion or at rest, motion must be a mode non-essential to matter. Moreover, given that matter itself is inert, it cannot be the source of the order and direction of motion. To provide order and direction, God laid down the laws of motion. Thus, motion itself as well as the orderly movement of bodies derives from God who acts as the effective principle. While the specific position, constitution, and configuration of the parts of a particular body determine how certain local motions are transferred, the source and ultimate direction of the motion itself is God. What this means for body-body interaction is that bodies function as secondary causes, directing local motions in virtue of the specific configurations of their parts. Bodies do not possess any causal power to produce or cease movement.

Le Grand's account of body-body interaction clears the way to explain mind-body interaction. In the same way as finite bodies, God functions as the effective principle of finite minds, providing the ultimate source of change: " … there is nothing, besides motion, which can strike the organs of the senses, or affect the mind itself." (1694, p. 284) Although mind and matter are substances sharing no common properties, it is in virtue of God acting as the effective principle that mind and body interact. This kind of interaction is no more or no less problematic than the interaction of two physical bodies. There was no real problem of interaction for Le Grand, since he believed that it was not the substances per se that acted on one another but in all cases God alone provided the motive force in the universe. Although things by their nature respond to this motive force in an orderly way, i.e., according to the laws of nature, the fact that they exist as they do and that they interact is a fact completely dependent on God's will. God's power is expressed as local motions in bodies and as passions/thoughts in minds. In short, it is in virtue of God, their effective principle, that a mind and body, or a body and body, or even a mind and mind, are said to

interact.

Le Grand described the mind-body union in terms directly borrowed from Descartes. However, Le Grand attempted to explicate further than did Descartes the nature of the mind-body union. According to Le Grand, there are three kinds of union, each possessing its own principle which effects that union: the first is that of two minds whose principle of union is love; the second is that of two physical bodies whose principle of union is local presence; and the third is that of the mind and body whose principle of union is actual dependence. Just as two physical bodies are joined by physical contact, and as two minds are joined by love, the mind and body are joined by a mutually dependent activity. So long as the body actually receives its specific motions dependently on the soul, and the soul actually receives its local motions (passions) dependently on the body, the spirit and the body are joined. Although there can be no mode common to mind and matter, there is this mutual action. While there is no mode shared by two different substances, there is a similitude and relation that exists between mind and body: "This similitude and relation we have formerly affirmed to consist in action and passion." (1694, p. 325) In other words, just as the body is capable of receiving and transmitting local motions since motion is a mode of matter, the mind is capable of varying passions since passions are a mode of the mind. It is by the mutual commerce of such motions and passions that the mind and body are said to be united. The mutual activity said to occur between the mind and the body is a property which follows only from the union of mind and body and cannot proceed from either alone, "And the truth is, since neither body can think, nor mind be capable of dimension, there can be no mode common to mind and body, except a mutual acting of each upon each, from which alone the properties of both can follow." (1694, p. 325)

Le Grand's extensions to Descartes's physics included phenomena now classified as metallurgy, entomology, botany, biology, physiology, medicine, and psychiatry. Part II of the *Entire Body of Philosophy*, entitled *The History of Nature*, catalogues and critically discusses the latest experiments of his time, as well as the theories of the ancients and moderns. Prominent in his discussions is the importance of secondary causes in nature (both exemplary and secondary efficient causes), and the need for experiments, not just as tools of confirmation, but also as a means of discovering the true nature of things. This was due to his application of mechanism to explain not only the behavior of material bodies but also the entire institution of nature. Le Grand believed that God laid down the laws of nature and the principles of being by acting as the primary efficient cause, and that the operation of these laws and principles manifested themselves in nature in the form of secondary causes and effects. Although the laws and their specific mechanisms of operation are not visible, secondary causes and their effects are. These causes and effects then are known by experience and are the starting point of all science, which is characterized by reasoning from effects (observed in nature) to causes (first principles discerned by reason).

3. Epistemology

Le Grand's account of sensations and ideas is orthodox Cartesianism. Sensory impressions are what mediate the external object and our mind's idea of it, and they consist in nothing more than the immediate motions of the sensory organs in the body. Such motions are produced by a natural necessity and they

share no similarity or affinity with the particular objects that cause them. Like Descartes, Le Grand used the example of the sword wounding the body to illustrate the non-resemblance or dissimilitude of the relations between external objects and sensations, and sensations and ideas. (1694, p. 327) The sword that produces pain in us is nothing like our sensation or idea of pain, nor is our idea of pain anything like our sensation of pain. Yet, we maintain that there is a causal and representational relation between the sword and the idea it produces. In addition, Legrand like Descartes made a clear distinction between sensory impressions, which are particular, quantifiable motions, and ideas, which are representational or propositional in character. Given that sensations are non-resembling and non-representational (they are mere patterns of local motions) it follows that ideas, which are essentially representational, could not be derived from them.

From the lack of any form of similitude or affinity between object/sensation, and sensation/idea, it follows that there is no such relation which holds between an idea and an external (material) object. From this lack of similitude, Le Grand concluded that adventitious ideas (coming from material objects outside us) must be innate or inbred in the mind. For, if the external object is not like the idea we form of it, then the only explanation remaining is that the mind is responsible for it. Likewise, fictitious ideas, like sirens and chimeras, have no exemplar outside the mind and so must be formed according to forms natural to the human mind. And finally, common notions such as substance, truth, goodness, equity, and God, as well as axioms such as the same thing cannot be and not be, are innate, that is, they proceed from the mind alone, since all corporeal motions are particular but these notions are universal. The sense in which they are innate differs from adventitious and fictitious ideas; innate ideas do not proceed from the senses or the imagination but "are congenite and inbred with the said mind, from their original." (1694, p. 328) By this Le Grand meant that the mind or thought itself, not any of its faculties such as sense or intellection, is the principle or original of such ideas. These ideas are formed in the mind by the mind and from the mind.

Thus as Descartes held, there are three kinds of ideas (adventitious, fictitious, and innate), which are distinguished by their differing sources as well as the way in which they are inbred in the mind. Adventitious ideas proceed from the senses, fictitious ideas proceed from the imagination and the intellect, and innate ideas proceed from thought itself, which acts as their ground or original. Nonetheless, all ideas, regardless of their source or origin, depend on the mind in some essential way for their form. But this gives rise to the problem of explaining how ideas can be said to represent, if they in no way resemble their objects. This problem is especially acute for Cartesians who held that there is a modal difference between what is found at the level of sensory perception and intellection, such that impressions cannot contain any of the properties found at the level of ideas. What arrives at our faculty of thinking from the senses is not ideas such as we form them in our thought, as the scholastic empiricists held, but rather only various particular motions emitted by external objects. (1694, p. 328)

Le Grand's solution to the problem of how ideas represent their objects employed the notion of substitution or ‘supplying a stead’ -- wherein the cause (the object) contains all the properties found in the effect (the idea) not actually but in virtue of its ability to supply the substitute properties or proxies. And a relation, according to Le Grand, "…is nothing else but a mode of our understanding, comparing one thing with others, because of some properties or acts that are found in them." (1694, p. 17) Descartes

himself never cashed out the notion of representation in terms of substitution, although he came close to suggesting it in the French version of the Third Meditation, in which he claimed that such things as extension, shape, position and movement may be contained in him eminently, "…and as it were the garments under which corporeal substance appears to us." (1985-91b, fn.1, p. 31) One could conceivably interpret this to mean that the garments of corporeal substance, namely extension, shape, position and movement are the garments supplied by the mind as the forms or the conceptions under which the mind grasps material things. Although the mind itself is not extended, shaped, locally positioned or moved, it dresses material substance in these properties in order to perceive particular material things. But there is no suggestion in Descartes regarding how the dressing is related to the material object grasped. Le Grand's notion of substitution was intended as an explication of this relation, and is his contribution to the Cartesian dialectic on ideas.

Le Grand was one of the few Cartesians to defend Descartes's doctrine of the creation of essences and eternal truths. The thesis is that God is the efficient cause of all things, both actual and possible, including all the truths we call eternal, "In like manner as a king is the maker of all laws in his kingdom. For all these truths are inborn in us from him; as a king also would have them so in his subjects, if he had power enough to write his laws in their hearts." (1694, p. 63) The main worry that critics, like Malebranche, had of this doctrine was that it would remove any necessary foundation for the propositions of science and theology, making them contingent and uncertain. To answer this worry, Le Grand added that there is one important difference between kings and God in the way they set down their laws, "A king can change his laws because his will is changeable, but God's will is unchangeable, for it is His perfection to be invariable in manner." (1694, p. 63) In this way, Le Grand attempted to account for the dependence of all things on God's will, while at the same time accounting for the immutable foundation of the truths of natural philosophy.

Whereas the creation of true and immutable natures was the work of God's freewill (not dictated according to his Wisdom, as Malebranche and other critics held), once created, they were necessary. In order to tie this necessity to the immutability of God's will without limiting God in any way, Le Grand drew on a Scholastic distinction between antecedent and consequent necessity. He argued that true and immutable natures, such as mathematical truths, only possessed a consequent necessity. God did not will that 6+4=10 because he saw it could not be otherwise, but in virtue of his free will, 6+4 [necessarily] = 10; therefore it could not be otherwise. As Descartes put this same point, "And even if God has willed that some truths should be necessary, this does not mean that he willed them necessarily; for it is one thing to will that they be necessary, and quite another to will this necessarily, or to be necessitated to will it." [1985-91c, p. 235]. For nothing outside of God, not even the eternal truths or immutable essences, necessitate that God act in one way or another, but rather, they are themselves eternal and immutable in virtue of the fact that God, whose existence is necessary and immutable, willed them in their essence and existence. Eternal truths and immutable essences are necessary only in that they presuppose and are consequent to the act of God who caused them.

Echoing Descartes in part VI of the *Discourse on Method*, Le Grand held that God implanted certain simple, true, and immutable ideas in our minds so that we could have a science of nature; yet, He also created a nature whose power is so ample and vast that only observation can close the gap between the

two. Thus, while Le Grand remains essentially a rationalist in his claim that knowledge of immutable essences, laws and truths remains the autonomous domain of reason, which is independent of the appearances of the senses, it is a rationalism tempered by his view that truths and laws are dependent on the will of God, and hence, are in some sense contingent. This dependence means that truth must be sought after in the effects of nature, and not in something independent of those effects, and that secondary causes, although dependent both on God's will and on the primary truths of nature, have a genuine role in causal explanation. In other words, our search for truth is in the specific operations of things, and even though our understanding of these truths is importantly independent of these specific operations, our discovery of them is not.

4. Ethics

Le Grand's early ethical and political writings are not Cartesian. In *Le Sage des Stoïques, ou l'Homme sans Passions, Selon les Sentiments de Sénèque* (1662), later translated and published as *Man Without Passions* (1675), he expounds the Stoical doctrines of Seneca, for which the goal of the moral person is to expunge the passions. He later rejected this view of the passions and argued the Cartesian view that the passions ought to be trained (not expunged) in the moral life. Le Grand also wrote a curious political treatise, *Scydromedia* (1669), which is a semi-fictional, utopian work describing his vision of the ideal state.

There is nothing innovative in Le Grand's moral theory, yet his discussions are rich with references to ancient and contemporary theories. He borrows from the ancient Atomists, the Stoics, the Scholastics, and the "Moralists" of his time, and frames it, where possible, in Cartesian terms. Le Grand acknowledges that Descartes himself wrote little on ethics, but he argues that Descartes's treatment of the soul and the passions provides a solid foundation for the treatment of moral matters. According to Le Grand, the object of ethics is right reason, its end is the perfection of man, and it is an active not a speculative science. An example of his reconciling project can be seen in Book I, Part X of his *Entire Body of Philosophy* (1694), where Le Grand attempted to reconcile the doctrines of Seneca and Epicurus on the role of pleasure in the virtuous life, by drawing on Descartes's theory of the passions. (1694, p. 347) He argues that pleasure has a role to play in the moral life, since virtue depends on freewill (as the Stoics held) and pleasure derives from the satisfaction of the mind in possessing the good (as Epicurus held). What Descartes's theory provided was an explanation of how pleasure (a passion) could aid the will in choosing the right course of action, while maintaining the voluntary nature of the will and virtue.

Le Grand, although not an innovator, is worthy of study for his contribution to the development of Cartesianism in the latter half of the seventeenth century. No less important is the fact that he spent most of his life in England, where his contact with members of the Royal Society and the universities of Cambridge and Oxford had a lasting impact on the reception of Descartes's ideas in England, Germany, and France.

Bibliography

Primary Texts

- Cudworth, Ralph (1678). *The True Intellectual System of the Universe,* London
- Descartes, Rene (1985-91a). *The Philosophical Writings of Descartes*, volume 1, Cambridge
- --------- (1985-91b). *The Philosophical Writings of Descartes, Meditations on First Philosophy with Objections and Replies*, volume II, Cambridge
- --------- (1985-91c). *The Philosophical Writings of Descartes, Descartes's Correspondence*, volume III, Cambridge
- Le Grand, Antoine. (1662) *Le sage des Stoïques ou l'homme sans passions , selon les sentiments de Sénèque*, The Hague; reprinted anonymously as: *Les caractères de l'homme sans passions, selon les sentiments de Sénèque*, Paris (1663, 1682); Lyons (1665); translated into English G. Richard, *Man without Passion: Or, the wise Stoick, according to the Sentiments of Seneca*, London (1675)
- ------. (1669) *L'Epicure spirituel, ou l'empire de la volupté sur les vertus*, Douai; Paris; translated into English by E. Cooke, *The Divine Epicurus, or, the Empire of Pleasure over the Virtues*, London (1676)
- ------. (1669) *Scydromedia seu sermo quem Alphonsus de la Vida habuit corram comite de Falmouth de monarchia liber primus*, London; Nuremberg, (1680); translated into German under same title U. Greiff, Bern: Lang (1991)
- ------. (1671) *Philosophia veterum, e mente Renati Descartes more scholastico breviter digesta*, London
- ------. (1672) *Institutio philosophiae secundum Principia D. Renati Descartes: Novo methodo adornata & explicata, cumque indice locupletissimo actua*, London (1675, 1678, 1680, 1683); Nuremberg (1679, 1683, 1695, 1711); Geneva, (1694)
- ------. (1673) *Historia naturae variis experimentis & ratiociniis elucidata*, London (1680); Nuremberg, (1678, 1680, 1702)
- ------. (1675) *Dissertatio de carentia sensus et cognitionis in brutis*, London; Lyons (1675); Nuremberg (1679)
- ------. (1679) *Apologia pro Renato Des-Cartes contra Samuelem Parkerum, S.T.P. archidiaconum cantuariensem, instituta & adornata*, London (1682); Nuremberg (1681)
- ------. (1682) *Jacobi Rohaulti tractatus physicus gallice emissus et recens latinitate donatus per Th. Bonetum D.M. Cum animadversionibus Antonii Le Grand*, London; Amsterdam (1691)
- ------. (1685) *Historia sacra a mundi exordio ad Constatini Magni imperium deducta*, London; Herborn (1686)
- ------. (1694) *An Entire Body of Philosophy, According to the Principles of the Famous Renate des Cartes*, in Three Books, I The Institution; II The History of Nature; III Dissertation on Brutes, trans. from the Latin into English R. Blome, London: Roycroft; reprinted with introduction R.A. Watson, New York: Johnson Reprint Corp. (1972)
- ------. (1698) *Censura Justissima Responsi, ut habetur, terribilis; cui titulus est idea cartesiana ad lydium veritatis lapidem*, London

- ------. (1698) *Dissertatio de ratione cognoscendi et appendix de mutatione formali, contra J.S. [John Sergeant] methodum sciendi*, London
- More, Henry (1671). *Enchiridion Metaphysicum*, London
- Newton, Isaac (1687). *Principia*, London
- Rohault, Jacques (1671). *Traité de physique,* Paris
- Sergeant, John. (1698) *Non ultra: or a letter to a learned Cartesian; settling the rule of truth, and first principles, upon their deepest grounds*, London
- Smith, John (1660). *Select Discourses*, London

Selected Studies and Critical Discussions

- Bouillier, Francisque (1854). *Histoire de la philosophie cartésienne*, 2 vols., Paris
- Rosenfield, Leonora Cohen (1968). *From beast-machine to man-machine; animal soul in French letters from Descartes to La Mettrie*, new and enlarged edition, New York: Oxford University Press
- Ryan, John K. (1935). "Anthony Legrand, 1629-99: Franciscan and Cartesian," *The New Scholasticism*, 9:226-250
- ----- (1936). "Scydromedia: Anthony Legrand's Ideal Commonwealth," *The New Scholasticism*, 10:39-55
- Watson, Richard A. (1966). *The Downfall of Cartesianism 1673-1712*, The Hague: Martinus Nijhoff

Other Internet Resources

- English Books On-Line.
  This is a subscription service; many university libraries have site licenses which allow patrons to view or download early modern texts (original editions) for free, in pdf format. The collection includes works from the first book printed in English by William Caxton, through the age of Spenser and Shakespeare and the tumult of the English Civil War. Early English Books Online (EEBO) contains over 125,000 titles listed in Pollard & Redgrave's *Short-Title Catalogue* (1475-1640), Wing's *Short-Title Catalogue* (1641-1700), and the *Thomason Tracts* (1640-1661). It includes a number of Le Grand's works, as well as those of Henry More, Ralph Cudworth, Samuel Parker, and John Sergeant.

Related Entries

Descartes, René | Malebranche, Nicolas

Patricia Easton

patricia.easton@cgu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: September 13, 2001
Content last modified: September 13, 2001

Godfrey of Fontaines

Godfrey of Fontaines was one of the major philosopher-theologians to serve as Regent Master at the University of Paris in the final quarter of the thirteenth century, along with Henry of Ghent and Giles of Rome. After completing philosophical studies in the Faculty of Arts and beginning to study theology probably by August, 1274, he became Regent Master in Theology in 1285 and retained this position until 1303/1304. His greatest contributions to philosophy were in the area of metaphysics. He was positively influenced by Thomas Aquinas, but differed with him on various major points, and also seems to have been influenced by the Arts Master, Siger of Brabant. His might well be described as a metaphysics of act and potency, since he often turns to this in seeking to resolve metaphysical problems such as, for instance, the relationship between essence and existence and between possible and actual being, the distinction between substance and accidents and between the soul and its powers, the causes of intellection and volition, or the nature of prime matter and its relationship to substantial form. Overall his philosophical thought and especially his metaphysics is somewhat more Aristotelian and less influenced by Neoplatonism than that of many of his contemporaries, including Aquinas.

1. Life and Writings

Godfrey of Fontaines was born in present day Belgium in the principality of Liège, very likely at the chateau of the noble family of which he was a member, at Fontaines-les Hozémont, probably shortly before 1250. This approximate date can be inferred from the fact that he conducted his first quodlibetal dispute as a Master of Theology at Paris in 1285, and that one could not become a Master in this faculty before reaching the age of 35. While nothing is known with certainty about his life prior to his arrival at Paris, he must have pursued philosophical studies there in the Faculty of Arts in the early 1270s. University statutes required one to spend at least eight years in studying theology before becoming a Master, and it is also known that he was inscribed at the Sorbonne before August 15, 1274. Hence his theological studies should have begun by that date (De Wulf, 1904, 3-16; Wippel, 1981, xv-xviii).

Evidently a great lover of books, among the surviving 37 manuscripts Godfrey left to the Sorbonne is a valuable "Student Notebook" (Paris: Bibl. Nat. lat. 16.297) which he himself compiled during his student days at Paris, and into which he entered many writings in his own hand. While not completed until ca. 1280 rather than by 1274 or earlier as previously proposed, it reflects his interests in the 1270s and includes works by Thomas Aquinas, Siger of Brabant, Boethius of Dacia, other anonymous Questions on various works of Aristotle which are clearly by some Master(s) of Arts of the time, still others by Giles of Rome, some extracts from Albert the Great and from Henry of Ghent, as well as some other purely theological writings (Wippel, 2001, 360-67; Duin, 1959). His interest in the writings of Radical Aristotelian Arts Masters at Paris in the 1260s and 1270s is also indicated by other works of Arts Masters included in additional manuscripts in his library (Bibl. Nat. lat. 15.819, 16.096). Also noteworthy is the inclusion in his Student Notebook of one of the oldest and most reliable copies of Thomas Aquinas's controversial *De aeternitate mundi*, which itself dates from 1270.

Godfrey continued to teach as a Master of Theology at the University until ca. 1303/1304 when he conducted his fifteenth and last Quodlibetal Disputation. He may have been away from Paris for some more extended periods after completing Quodlibet XIV ca. 1298-1299. It is known that he maintained close connections with Liège during his career at Paris and, indeed, that he served as Canon of Liège. He also served as Canon of Tournai and as Provost of San Severin at Cologne, and was elected Bishop of Tournai in 1300 but renounced his rights to the See when the election was contested. The year of his death is probably 1306 or 1309, but the day is known, i.e., October 29.

Godfrey selected the quodlibetal disputation as his major vehicle for publication. These solemn disputations were conducted twice during the academic year, i.e., before Christmas and before Easter. They were open to all members of the learned public and questions could be put to the presiding Master by anyone in attendance. During the first day's disputation, preliminary answers would be given to these many and disparate questions. On the following day or at least on some following day, the Master, having in the meantime imposed some logical organizing plan upon the various questions, would return for another session where he would present his definitive response or "determination" for each of them.

Subsequently he would prepare this version for publication and, when it was completed, would submit it to the University Bookseller. Masters were not required to conduct these open disputations and since they were regarded as onerous, not all did. Godfrey's 15 Quodlibets have all been edited, although only *reportationes* of the first four remain, i.e., copies taken down by an auditor. Brief versions (*abbreviationes)* of Quodlibets III and IV have also been edited. As a Master of theology Godfrey also conducted ordinary Disputed Questions, and some of these have been preserved.

2. Subject of Metaphysics

Godfrey was surely familiar with the controversy concerning whether one should with Avicenna stress the nonparticular and therefore the universal character of being as being and hence make this the subject of metaphysics or rather with Averroes emphasize it as the science which has the highest kind of being, the divine, as its subject. On this Godfrey clearly agrees with Avicenna, even though he devotes little explicit attention to the controversy itself. Thus in Quodlibet X, q. 11 he holds that being as being is the object (or subject) of metaphysics (PB 4.349), and in Quodlibet VI, q. 6 that the concept of being is first and simplest because it enters into every other concept (PB 3.137). Therefore it is also the most general concept or, one may say, transcendental. At times Godfrey also describes the object of the intellect as being as being (Quodlibet II, q. 8). He denies that God is the subject of metaphysics, even though God is the first and primary being and must be studied within metaphysics (Quodlibet I, q. 5). He distinguishes a metaphysical study of God, which he says may be described as a kind of theology, from the theology that is based on Sacred Scripture. This does not have being as being as its subject, but God himself (Quodlibet IX, q. 20; Wippel, 1981, 3-15).

3. Division of Being

In opposition to Henry of Ghent's division of real being into essential being (*esse essentiae*) and existential being (*esse existentiae*), in Quodlibet VIII, q. 3 Godfrey proposes his own division. Being (*esse*) may be divided into being in the mind ("cognitive" being) which is a lesser or diminished being, and real being, i.e., being outside the mind or knower. Real being is subdivided into real being in potency and real being in act. A thing possesses real being in potency insofar as it has being by reason of its cause or causes. It has real being in act insofar as it is realized in its own nature in completed or perfected form. A thing may have real being in potency either by reason of its intrinsic cause (e.g., if matter preexists which may enter into the thing's constitution), or by reason of an extrinsic cause (e.g., if an agent preexists which can produce it). Godfrey illustrates this with the example of a rose that does not yet actually exist. Prior to the creation of the world and hence of matter it possessed real being in potency only by reason of God, its extrinsic cause. After creation it possessed real being in potency by reason of preexisting matter as well, an intrinsic cause. And now it may also possess real potential being by reason of some created extrinsic cause (or causes). Both before and after the world's creation it enjoys cognitive being insofar as it was and is known by the divine mind (PB 4.38-40; Wippel, 1981, 15-17).

4. Analogy of Being

In describing being as being as the object of the intellect in Quodlibet II, q. 8, Godfrey also states that being is taken analogically and not univocally. It is affirmed first and foremost of substance, especially of first substance, and of everything else as related to substance. Hence both substance and accident are included under this analogous notion of being and this single object of the intellect (PB 2.135-36; Wippel, 1981, 19-24). In Quodlibet III, q. 1 he attempts to establish the analogical character of being by arguing at some length that it cannot be either univocal or purely equivocal. Central to his rejection of univocity of being is his denial that it is a supreme genus. It must apply not only to the generic and specific aspects in which different things share, but also to the differences, including the individual differences between them (PB 2.162-63). In Quodlibet XV, q. 3, while responding to an objection perhaps taken from Meister Eckhart's Parisian Disputed Question 1, he maintains that, if being is applied to accidents insofar as they are related to substance in some way, this does not mean that being is not intrinsically present in accidents. So, too, when it is applied to creatures viewed as effects and to God their cause, it applies intrinsically to God. There is an analogy and proportion in reality between these different instantiations of being and corresponding to this, there is analogy in meaning as well. It is this that is grasped by the analogous concept of being (PB 14.18-20; Wippel, 2001, 381-82).

5. Transcendentals

In Quodlibet III, q. 1 Godfrey identifies the one, the true, and the good as properties of being that are really identical and convertible with it, and hence as transcendental characteristics of being. These properties are not really distinct from being itself (PB 2.163-64). In Quodlibet VI, q. 16 he distinguishes between the one or the kind of unity that serves as a principle for number and which is based on discrete quantity (numerical unity in the strict sense), and the kind that is convertible with being. While the former is restricted to corporeal being, the latter applies to every subsisting substance and to every accident that exists in such a substance. Hence it alone is transcendental (PB 3.256-58). Regarding the true, in Quodlibet VI, q. 6 Godfrey indicates that truth adds nothing real to being but only a (conceptual) relationship to mind or intellect. To assign this kind of truth to a thing is simply to acknowledge that it can be grasped by intellect or that it is intelligible. Consequently, Godfrey holds that truth is present in being virtually insofar as it has the capacity (*virtus*) to produce truth in the intellect. But he favors the view that when taken formally truth resides in the intellect (PB 3.137-41; Wippel, 1981, 25-34). Hence here he is recognizing the distinction later referred to as that between ontological truth and logical truth.

6. Essence and Existence

Already during Godfrey's student days in Arts and in Theology at Paris there was considerable discussion concerning the exact relationship between essence and existence in finite or created beings. Closely associated with the metaphysical thought of Aquinas was the view that in all finite beings there is a real, i.e., not merely a conceptual or mind-dependent distinction and composition of an essence principle and

an act of existing (*esse*) or existence principle. Already in the 1270s Giles of Rome was developing his own theory of real distinction between essence and existence, and was soon engaged in ongoing controversy with Henry of Ghent. At times Giles referred to essence and existence as distinct "things" (*res*) and, while he denied that either could exist in separation from the other or that existence is an essence, his terminology left its mark on the theory and open to such misinterpretations; for it invited critics to view existence or the act of being as an entity rather than as a principle of an existing entity, as Aquinas had envisioned it. While Henry of Ghent rejected any real distinction between essence and existence, he defended something more than a merely conceptual distinction between them, namely a new and third type that would fall between the real distinction and the merely conceptual, an "intentional" distinction (Wippel, 1981, 40-45).

Godfrey briefly refers to this issue in Quodlibet II, q. 2, while attempting to determine whether the essence of a creature is indifferent to existence and nonexistence. He comments that either essence is really identical with existence and differs from it (1) only conceptually or (2) intentionally, or else (3) existence is a distinct thing, i.e., the act of the essence and really distinct from it (PB 2.60). In Quodlibet IV, q. 2, he was asked to determine whether to hold that predicamental things are eternal by reason of their quiddity is also to hold that the world is eternal. In preparing his response he presents in greater detail these three different views on the relationship between essence and existence (*esse*). According to some they are really distinct from one another and enter into real composition with one another. But one is not separable from the other so as to be able to exist apart from it. Consequently, if a thing lacks or loses its existential being, it also lacks or loses its essential being. According to a second view they are really identical, but differ intentionally. Hence when a thing loses its existential being, its essence cannot be said to exist; but it does retain its true predicamental or essential being (*esse essentiae*). Finally, a third position, which Godfrey himself embraces, maintains that they are really identical and differ only conceptually. They do not enter into composition with one another. Therefore, to the extent that something enjoys essential being, to that same degree it enjoys actual existence. And, adds Godfrey, whatever is understood of one is also understood of the other (PB 2.235).

In Quodlibet III, q. 1, dating from 1286, Godfrey considers the argumentation offered for each of these positions. First he presents the theory that distinguishes really between essence and existence in language that reflects the terminology of Giles of Rome. Thus he refers to existence both as "something" (*aliquid*) and as a "thing" (*res*) that is added to essence. He then presents a number of arguments in support of this view which seem to be taken from Giles, especially from his *Quaestiones disputatae de esse et essentia*, q. 11. One of these is reminiscent of the first part of Aquinas's much discussed reasoning in *De ente et essentia*, c. 4, which begins from the fact that one can understand what something is without knowing whether it actually exists. However, Giles's presentation (and Godfrey's repetition) of the argument makes a stronger claim. One can understand what something is and also know that it does not exist. But because nothing can be understood with the opposite of itself, essence and existence must be really distinct. After presenting a series of arguments against this theory and refuting the arguments he had initially offered in support of it, Godfrey resolutely rejects any real distinction between essence and existence. For him they are identical, and differ only in the way they signify, just as the concrete noun "a being" (*ens*), the abstract noun "essence" (*essentia*), and the verb "to be" or "to exist" (*esse*) differ in their mode of signifying, but designate one and the same reality (PB 3.304 [brief version]; Wippel, 1981, 45-

66).

In presenting Henry's theory of intentional distinction Godfrey traces this back in large measure to what he regards as an incorrect interpretation of Avicenna's notion of nature or essence when it is considered simply in itself or "absolutely" rather than as existing in the mind or in an individual entity. As Godfrey explains in his later Quodlibet VIII, q. 3, according to Henry's position real being is divided into essential being (*esse essentiae*) and existential being (*esse existentiae*). A thing possesses essential being from eternity insofar as it corresponds to its appropriate exemplar idea within the divine intellect. Because of this it is a true or real quiddity or essence from eternity and falls into its appropriate predicament even though it is not an actual existent. Existing entities receive actual existence only in the course of time when the divine will intervenes to cause this. Within an actually existing being, therefore, its essence and existence are not really distinct. But Henry denies that they are identical. They are "intentionally" distinct (Wippel, 1981, 67-79; Marrone, 2001, 39-52; Porro, 1996, 211-53). Godfrey rejects Henry's new and intermediary kind of distinction out of hand. A distinction must either be real or purely conceptual. Accordingly, in Quodlibet III, q. 1 he argues at length against Henry's application of the intentional distinction to essence and existence Wippel, 1981, 85-88).

As already noted, for Godfrey essence and existence are really identical and only differ conceptually. Whatever is true of essence is true of existence, and vice versa. It is not necessary to posit two really distinct or even two intentionally distinct principles to account for the fact that one may be aware of something as a possible existent when it does not actually exist. It is enough to distinguish between potential being and actual being. If something is in potency in terms of its essence, it is in potency in terms of its existence. And if it is actual in terms of its essence, it is actual in terms of its existence.

Godfrey proposes a different kind of act-potency "composition" in order to meet one argument in support of real distinction between essence and existence. If, as Aquinas, Giles, and Godfrey all hold, the angels of Christianity are purely spiritual and not composed of (spiritual) matter and form, then it seems that they must be composed of essence and existence. Otherwise they would be perfectly simple and equal to God. Godfrey responds that one and the same being, even if purely spiritual, may be regarded as actual insofar as it exists, but as potential insofar as it falls short of the actuality enjoyed by a higher being and, above all, by the First Being, God. He quotes Proposition 2 from Proclus's *Elementatio theologica*: "What participates in the One is both One and not-One." As Godfrey reads this, anything that is different from the One can fall short of it only by approaching (*accessus*) the not-one. Hence it is simply by reason of the fact that such a being recedes from the One that it is not the One itself. In this way more perfect beings such as angels are distinct from the One, or God, without being composed either of matter and form or of essence and existence. Nonetheless, actuality and potentiality are present in them because they possess a certain intermediary nature and hence are likened or "assimilated" to something higher and more actual, and to something lower and more potential. Therefore they are "composed" of potency and act, not really, but conceptually. This composition is not merely imaginary. It applies to such entities by their being related to something higher and to something lower (Quodlibet III, qq. 1, 3; Quodlibet VII, q. 7; Wippel, 1981, 90-97).

Godfrey's library contains two likely sources for this theory, first a rudimentary version found in the

abbreviated text of Siger of Brabant's *Quaestiones in Metaphysicam* included in Godfrey's Student Notebook, and second, an anonymous set of questions on the *Posterior Analytics* contained in the manuscript Bibl. Nat. lat. 16.096. While Godfrey literally borrows certain parts of his theory from the second text, he does not follow this anonymous Radical Aristotelian Arts Master when he goes on to argue that separate entities do not depend upon God as their efficient cause, but only as their final cause (Wippel, 1984), 231-44.

7. Philosophical Knowledge of God's Existence

According to Godfrey, insofar as a natural knowledge of God is accessible to human reason, it belongs more properly to metaphysics than does knowledge of any other being. And he clearly holds that God's existence can be established by philosophical reasoning. But Avicenna and Averroes had differed concerning whether it belongs to natural philosophy (physics) or to metaphysics to demonstrate this conclusion. Avicenna had maintained that this task pertains to metaphysics and only to metaphysics, whereas Averroes had defended the opposite position that only physics can demonstrate the existence of God, the First Mover (Wippel, 1981, 102-3). As will be seen more fully in the following section, Godfrey defends something of a compromise position. In Quodlibet XI, q. 1 he mentions that the metaphysician's consideration of God in himself is more perfect than that of the natural philosopher, who simply views him as the First Mover of the first mobile being, i.e., the outermost heavenly sphere. But, Godfrey adds, by reason of all that God is in himself, he is also the First Mover (PB 5.3). In Quodlibet V, q. 10 he writes that one can know by reasoning from natural things that God is the first being which depends on nothing whatsoever and upon which everything else depends and, therefore, that he is the causal and productive principle of all other things (PB 3.41; Wippel, 1981, 105). And in Quodlibet IX, q. 20, he refers to different things that natural reason can know in metaphysics with certainty about God—that because he is the first being he is simple; that he is being in actuality; that he is an intellectual being, etc. (PB 4.288).

8. Quidditative Knowledge of God and the Divine Attributes

In Quodlibet VII, q. 11 Godfrey considers the view of some, presumably Thomas Aquinas, who say that in this life we can know that God is, but not what he is. Godfrey makes a clear but implicit reference to ST I, q. 3, a. 4 where Thomas writes that even when we recognize that God is, the "is" which we understand is not the act of being whereby God subsists in himself, but only that indicating that the proposition "that he is" is true (PB 3.383).

Godfrey finds this too restrictive. And so after offering a detailed explanation of the different ways in which one can know of something "what it is" and "that it is," he writes that just as in knowing material things we move from more confused to less confused knowledge, so it is in the case of our natural knowledge of God. Just as we find that some things are the principal causes of others, and some are

governed by others, so we impose the name "God" to signify something in the universe which is the one first cause of everything else and than which nothing greater can be thought. But such nominal knowledge is not enough to prove that what we express by the name "God" exists in reality or "that he is." Next we may follow Aristotle's procedure in *Physics* VII where by eliminating recourse to an infinite regress of moved movers he concludes that one First Mover or God exists. And from the continuous motion of the first mobile sphere Aristotle shows in *Physics* VIII that God is perpetual and pure act. This tells us that God is in reality, but not what he is in any real sense. In the *Metaphysics* (Bk XII), continues Godfrey, Aristotle accepts the knowledge "that God is" as proved in the *Physics* and now proceeds to show that certain perfections are present in God to the preeminent degree. According to Godfrey, Aristotle uses these perfections as quasi-differences and thereby progresses from knowledge "that God is" to knowledge "what he is" by passing from a confused and quasi-generic knowledge to a more determined and quasi-specific knowledge. And so, too, proposes Godfrey, we may reason, for instance, first by knowing him as a substance, then as an incorporeal substance, then as a living and intelligent incorporeal substance. He acknowledges that God does not really fall into any genus or species. Nonetheless, he maintains against Aquinas that we can know "what God is" in some real sense, even though he recognizes that in this life such knowledge will always be imperfect (PB 3.384-86; Wippel, 1981, 108-15).

Consistent with the above, Godfrey defends the presence of a plurality of attributes in God even though, because of the divine simplicity, these are only conceptually distinct from the divine essence and from one another. In Quodlibet VII, q. 1 he distinguishes two ways in which the term "attribute" may be taken. It may be used to signify a divine perfection in the sense that some perfection in a creature which does not imply any imperfection in and of itself, in other words a pure perfection, is assigned to God to an eminent degree. Or it may be taken as signifying a pure perfection which is realized in God as a "quasi-quality" which perfects the divine substance in a "quasi-accidental" fashion. Godfrey comments that it is in this second sense that attributes are usually applied to God, even though this is not intended to imply any real distinction or composition of substance and attribute in him. When taken in this second sense, Godfrey holds that there are many such quasi-qualities which perfect God in this quasi-accidental way, and therefore, many divine attributes. They are assigned to God in preeminent fashion because of his infinite perfection (PB 3.265).

In this same question Godfrey was asked to resolve a still more fundamental issue. Because of God's absolute simplicity, divine attributes signify perfections that are really identical with the divine essence and with one another. The conceptual distinction between them can only arise from an intellect's consideration. But what is the ultimate foundation for this conceptual distinction? Does it arise from a consideration of God simply as he is in himself, or does it only result from some reference to really distinct realizations of these perfections in creatures? Godfrey replies that if one takes an attribute in the first sense as implying that every pure perfection present in creatures is to be assigned to God to an infinite degree, the answer is clear. The intellect, especially a created intellect, can arrive at such a conceptual distinction of divine attributes only by reasoning from the really distinct instantiations of such perfections in creatures (PB 3.267-70; Wippel, 1981, 116-18). But what about the divine intellect? Godfrey then refers to and rejects the view of Henry of Ghent, according to whom attributes, when taken in the second sense, can be recognized as distinct and as multiple by God himself insofar as he views

himself directly and without any reference to creatures. Against Henry, Godfrey maintains that not even God himself can be aware of this conceptual distinction between the divine attributes without referring to other beings in which these perfections are present in really distinct fashion. To hold otherwise would be to introduce too much distinction and diversity into the divine essence itself and would thereby compromise the divine simplicity (PB 3.267-73; Wippel, 1981, 118-23; Maurer, 1999, 192-200).

9. Eternity of the World

Together with his Christian contemporaries, Godfrey believed that the world began to be. But much debated was the question whether natural reason can prove this, or whether it can be held solely on the grounds of religious faith. Probably best known for holding that it cannot be demonstrated that the world began to be was Thomas Aquinas who in his *De aeternitate mundi* went a step beyond his earlier writings and maintained not only this but also that an eternally created world is possible (Wippel, 1984, 203-14). Bonaventure had presented a series of arguments to prove that the world began to be, and many others strongly defended this position, including Henry of Ghent (Dales, 1990; Wippel, 1981, 153-58). Thus it is not surprising that in his Quodlibet II, q. 3 of Lent, 1286, Godfrey was asked to determine whether the world or any creature could be or exist from eternity.

Godfrey develops his position in conscious opposition to Henry of Ghent, and with a considerable but unacknowledged dependency on Aquinas's *De aeternitate mundi* which, as noted above, was contained in his Student Notebook. But after he has, with Aquinas, shown that there is no contradiction in holding that something can be created and still have begun to be, Godfrey considers certain objections against this position. One of these objections clearly gives him pause. If the world had been created from eternity, on every given day extending backward into a beginningless past God could have created some material object such as a stone. But if that had happened, an actual infinity of stones would now exist and God could unite all of them into one infinite body. But an infinite body is an impossibility, and so too is an infinity of simultaneously existing finite bodies and, therefore, so is an eternally created world (PB 2.68-69, 76; Wippel, 1981, 160-63).

Godfrey comments that this objection can also be formulated more forcefully in terms of human souls, i.e., the actual infinity of human souls that would have resulted from an eternal world eternally populated by human beings with immortal souls. Aquinas had considered this form of the objection in his *De aeternitate mundi*, and had noted that it has not yet been demonstrated that God could not produce an actual infinity of spiritual beings. Godfrey does not adopt this solution, presumably because he is convinced that an actual infinity of entities whether spiritual or material is impossible. Instead he proposes as a possible alternative an eternally populated world involving the transmigration of a finite number of souls to an infinity of bodies, and ordered only to their natural perfection. But because this world seems to be intended primarily for human beings destined to enjoy eternal happiness in soul and body, Godfrey grants that it may be argued with probability that this world could not have been created from eternity under the present dispensation by God's ordained power. But, he also remarks, this does not prove that no creature or no world could have been eternally created. He concludes that it cannot be demonstrated either that an eternal world is not possible, or that it is possible. Either side may be

regarded as probable, and neither is to be rejected as theologically erroneous (PB 2.79-80; Wippel, 1981, 167-68).

10. Substance, Accidents, Human Action

In Quodlibet XIV, q. 5 Godfrey contrasts mental being with extramental being and then divides the latter into being *per se* and being *per accidens*. Being *per se* is divided into substance and the nine genera of accidents. The latter may be regarded as real modes of substance or of being in the unqualified sense (*ens simpliciter*). As Godfrey indicates elsewhere, a substance enjoys a separate being or exists in itself, while it is of the nature of an accident to be ordered to and to exist in something else. An accident is not so much a being as "of a being" (PB 5.427) For practical purposes Godfrey accepts the number of predicaments as ten, but indicates that determination of their precise number is a matter of probability rather than certainty (Wippel, 1981, 174-75). For Godfrey substance and accident are related as potency and act since substance serves as a subject for its accidents. Because of this he denies that any substance can be the efficient cause of the accidents that inhere in it, for it would then be in act with respect to them (as their efficient cause) and in potency (as receiving them) at one and the same time. He would always insist that nothing can be in act and in potency at one and the same time with respect to the same thing for this would be to assign being and nonbeing to it simultaneously. He also holds against many of his contemporaries that the powers of the soul are really distinct from the essence of the soul and from one another (Wippel, 1981, 176-84, 202-7).

If the soul's immanent operations such as thinking or willing inhere in their respective powers, the intellect and the will, those powers themselves cannot be the efficient causes of such acts. Against Henry of Ghent and later in Quodlibet XV against the Franciscan Gonsalvus of Spain, Godfrey denies that any exception can be made to the act-potency axiom. The will cannot reduce itself from potency to act or immediately efficiently cause its acts of volition. The efficient cause of volition can only be the object presented to the will by the intellect. Against the charge of intellectual determinism sometimes attributed to this position by its critics, Godfrey grounds human freedom in the radical indeterminacy, he even speaks of the freedom, of the intellect itself. And in his final Quodlibet XV, q. 4, he argues that if, *per impossibile*, the will could move itself directly without being moved by its object, freedom would be less well preserved than if one maintains with him that the will is moved by its object and then moves the apprehensive powers to their respective acts and, by means of such motions, indirectly moves itself with respect to secondary objects of volition (PB 14.20-23; Putallaz, 1995, 184-87, 198-208, 233-47).

11. Theory of Abstraction

According to Godfrey the agent intellect and the possible intellects are distinct powers of the individual human soul. His theory of intellectual knowledge is based on the agent intellect's ability to abstract potentially intelligible content from phantasms (images) produced by the imagination, an internal sense. The imagination depends upon the external senses for the data preserved in the phantasms. In Quodlibet V, q. 10 he makes a studied effort to explain in what the process of abstraction from phantasms consists

as he responds to the question whether the agent intellect produces any positive disposition in the phantasm. One of the functions of the agent intellect is to illuminate phantasms so that they can move the possible intellect to understand. Because the possible intellect is at times only in potency with respect to an intelligible object, and because nothing can reduce itself from potency to act, it must be reduced to the act of understanding by something else. Hence the agent intellect must in some way enable phantasms to move or to actualize the possible intellect.

But, argues Godfrey, because phantasms exist in the imagination and are organic and individuated and therefore incapable of moving a purely spiritual power which knows in universal fashion, the agent intellect does not introduce any positive disposition into the phantasms. Such a disposition would itself be individuated and organic and incapable of moving the possible intellect. He concludes, therefore, that the agent intellect operates on the phantasm simply by removing or separating or isolating one factor present therein--the quiddity of the thing, from another--its individuating characteristics. What has been so removed or separated or abstracted is thereby rendered universal and capable of moving the possible intellect. In what would become a frequently cited illustration, he draws an analogy with milk which possesses both color (white) and taste (sweet). Without the influence of light, however, it could not manifest itself or be perceived under the species of color (as white) without also manifesting itself under the species of taste (as sweet). Because of the influence of light one can speak of a kind of "abstraction" of the white from the sweet, although not in the sense that one would then exist apart from the other.

So, too, in the order of consideration although not in the order of reality, the agent intellect separates or frees the quiddity presented in a phantasm from its individuating conditions and thereby reduces it from being potentially intelligible to being actually intelligible and capable of moving the possible intellect to understand. This freeing or abstracting process takes place because of a kind of spiritual contact with the light of the agent intellect (Wippel, 1986).

12. Prime Matter

Godfrey defends the matter-form composition of corporeal beings but rejects any kind of spiritual matter and therefore any matter-form composition of spiritual entities. He also opposes the strong tendency within the Franciscan tradition, also promoted by Henry of Ghent, that would assign some minimum degree of actuality to prime matter. According to Godfrey prime matter is pure potentiality and can never be kept in existence without some substantial form, not even by God. Prime matter and substantial form are directly related to one another as potency principle and act principle. Neither is a being in its own right, but both are principles of one and the same composite entity (Quodlibet XIV, q. 5, PB 5.404-05).

13. Unicity vs. Plurality of Substantial Form

Heatedly debated in the 1270s and 1280s both at Paris and Oxford was the question concerning whether there is one or more than one substantial form in one substance, and especially in a human being. Godfrey devotes considerable attention to this, especially in Quodlibet II of Easter, 1286 and then again

in Quodlibet III of Christmas, 1286. These dates are significant because alleged theological difficulties with the theory of unicity of substantial form as advanced by Aquinas and others had led to prohibitions and condemnations of this by the Archbishop of Canterbury, Robert Kilwardby, O.P on March 18, 1277, a reissuing of this by his successor in that see, John Pecham, O.F.M. in 1284, and finally a new and resounding condemnation by Pecham on April 30, 1286.

In Quodlibet II, q. 7 Godfrey considers three theories that defend plurality of forms in all material substances, or at least of duality of substantial forms in human beings (Henry of Ghent). Godfrey severely criticizes all three theories. Most fundamental among his many arguments against them is his conviction that a substantial form confers substantial being upon a composite substance. Therefore, the presence of more than one substantial form in such a being would undermine its substantial unity. Already in this discussion Godfrey addresses alleged theological issues raised against unicity of substantial form, especially one concerning the continuing numerical identity of Christ's body during the period between his death on Good Friday and his resurrection on Easter Sunday. At this point Godfrey's philosophical preference is for unicity of form in all material entities, including humans. He regards theories that defend plurality of forms in all material substances as more improbable, and Henry's theory of duality of forms in human beings as less improbable. He maintains that the theological question remains open to a defense either of unicity or of plurality of substantial forms.

In Quodlibet III, q. 5, after Pecham's condemnation of April, 1286, Godfrey examines the theological aspects of this issue in detail. Again he insists on one's freedom, theologically speaking, to defend either unicity or plurality of forms in human beings, especially at Paris, but does not commit himself definitively to either side (PB 2.197-211). In sum, when he is primarily concerned with the philosophical issues involved, he favors unicity of substantial form in all material substances. When discussing the theological issues, he is more hesitant and, without embracing either side in any definitive way, continues to argue for one's freedom to defend either side pending some future decision by the Church (Wippel, 1981, pp, 321-47).

14. The Principle of Individuation

Together with many of his contemporaries, Godfrey attempts to identify the principle within material entities that accounts for their being multiplied numerically within species. In discussing this he returns to the distinction he had drawn in Quodlibet VI, q. 16 between the kind of unity which serves as a principle of number and which is based on discrete quantity, and the kind that is convertible with being, transcendental unity. Because the substantial form is the determining principle within a corporeal entity and makes its essence capable of being defined, Godfrey holds that it is by reason of its form that such a substance enjoys transcendental unity.

Nonetheless, in Quodlibet VII, q. 5 he comments that if different individuals within the same species share in the same specific nature, that nature cannot itself serve as the principle that renders them numerically distinct from one another. Something else seems to be needed, perhaps quantity. But to make quantity the principle of individuation will not resolve the problem, for this would be to reduce the

principle that distinguishes one substance from others within the same species to the level of an accident. Godfrey notes that in the case of created immaterial beings (angels), their substantial form is also that by which they are individuals. He proposes, therefore, that in corporeal substances it is also the substantial form that serves as their principle of individuation. Nonetheless, quantity has its role to play as well since it is required in order to divide matter into distinct parts and thereby enable it to receive and individuate substantial forms of the same kind. In other words, quantity disposes matter so as to enable it to function as the material principle of individuation. The formal cause or principle of a thing's individuation is its substantial form. Quantity is not the formal principle of individuation, but by disposing matter so that it can receive different substantial forms, it may be described as a quasi-material disposing cause of individuation (Wippel, 1981, 349-64).

Bibliography

- Dales, R. (1990). *Medieval Discussions of the Eternity of the World*. Leiden: E.J. Brill.
- De Wulf, M. (1904). *Un théologien-philosophe du XIIIe siècle. Étude sur la vie, les oeuvres et l'influence de Godefroid de Fontaines*. Brussels: M. Hayez.
- Duin, J.J. (1959). "La bibliothèque philosophique de Godefroid de Fontaines," *Estudios Lulianos* 3, pp. 21-36, 136-60.
- Godfrey of Fontaines. *Les Philosophes Belges* (Louvain: Institut Supérieur de Philosophie de l'Université). Vol. 2 (=PB 2): *Les quatres premiers Quodlibets de Godefroid de Fontaines* (1904), M. De Wulf, A. Pelzer, eds.; Vol. 3: *Les Quodlibets cinq, six et sept* (1914), M. De Wulf, J. Hoffmans, eds.; Vol. 4: *Le huitième Quodlibet, Le Neuvième Quodlibet, Le dixième Quodlibet* (1924, 1928, 1931), J. Hoffmans, ed.; Vol 5: *Les Quodlibets onze et douze, Les Quodlibets treize et quatorze* (1932, 1935), J. Hoffmans, ed.; Vol. 14: *Le Quodlibet XV et trois Questions ordinaires de Godefroid de Fontaines* (1937), O. Lottin, ed.
- Godfrey of Fontaines. Disputed Questions. Some have been edited in scattered publications. For a list see Wippel, 1981, pp. xxxi-xxxiii.
- Marrone, S. (2001). *The Light of Thy Countenance. Science and Knowledge of God in the Thirteenth Century. Vol. 2: God at the Core of Cognition*. Leiden: Brill.
- Maurer, A. (1999). *The Philosophy of William of Ockham in Light of its Principles*. Toronto: Pontifical Institute of Mediaeval Studies.
- Paulus, J. (1938). *Henri de Gand. Essai sur les tendances de sa métaphysique*. Paris: J. Vrin.
- Porro, P. (1996). "Possibilità ed *Esse essentiae* in Enrico di Gand," in W. Vanhamel, ed., *Henry of Ghent. Proceedings of the International Colloquium on the Occasion of the 700th Anniversary of his Death (1293)*. Leuven: Leuven University Press, pp. 211-53.
- Putallaz, F.X. (1995). *Insolente liberté. Controverses et condemnations au XIIIe siècle*. Fribourg: Éditions Universitaires/Paris: Éditions du Cerf.
- Wippel, J.F. (1981). *The Metaphysical Thought of Godfrey of Fontaines. A Study in Late Thirteenth-Century Philosophy*. Washington, D.C.: The Catholic University of America Press.
- Wippel, J.F. (1984). "Possible Sources for Godfrey of Fontaines' Views on the Act-Potency Composition of Simple Creatures," *Mediaeval Studies* 44 (1984), pp. 222-44.
- Wippel, J.F. (1986). "The Role of the Phantasm in Godfrey of Fontaines' Theory of Intellection,"

in C. Wenin, ed., *L'homme et son univers au moyen âge (*Actes du septième congrès internationale de philosophie médiévale [30 Août-4 Septembre 1982]), Vol. 2, pp. 573-82.

- Wippel, J.F. (2001). "Godfrey of Fontaines at the University of Paris in the Last Quarter of the Thirteenth Century," in J.A. Aertsen, K. Emery, Andreas Speer, eds., *Nach der Verurteilung von 1277. Philosophie und Theologie an der Universität von Paris im letzten Viertel des 13. Jahrhunderts. Studien und Texte* (Miscellanea Mediaevalia, 28) Berlin-New York: Walter de Gruyter, pp. 359-89.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aquinas, Saint Thomas | Giles of Rome | Henry of Ghent

John Wippel
wippel@cua.edu

First published: August 17, 2001
Content last modified: August 17, 2001

Causation and Manipulability

Manipulablity theories of causation, according to which causes are to be regarded as handles or devices for manipulating effects, have considerable intuitive appeal and are popular among social scientists and statisticians. This article surveys several prominent versions of such theories advocated by philosophers, and the many difficulties they face. Philosophical statements of the manipulationist approach are generally reductionist in aspiration and assign a central role to human action. These contrast with recent discussions employing a broadly manipulationist framework for understanding causation, such as those due to the computer scientist Judea Pearl and others, which are non-reductionist and rely instead on the notion of an intervention. This is simply an appropriately exogenous causal process; it has no essential connection with human action. This interventionist framework manages to avoid at least some of these difficulties faced by traditional philosophical versions of the manipulability theory and helps to clarify the content of causal claims.

1. Introduction

A commonsensical idea about causation is that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control: very roughly, if *C* is genuinely a cause of *E*, then if I can manipulate *C* in the right way, this should be a way of manipulating or changing *E*. This idea is the cornerstone of manipulability theories of causation developed by philosophers such as Gasking (1955), Collingwood (1940), von Wright (1971), and Menzies and Price (1993). It is also an idea that is advocated by many non-philosophers. For example, in their extremely influential text on experimental design (1979) Cook and Campbell write:

> *The paradigmatic assertion in causal relationships is that manipulation of a cause will result in the manipulation of an effect.* … Causation implies that by varying one factor I can make another vary. [Cook & Campbell, 1979, p. 36, emphasis in original.]

Similar ideas are commonplace in econometrics and in the so-called structural equations or causal modeling literature, and very recently have been forcefully reiterated by the computer scientist Judea Pearl in an impressive book length treatment of causality (Pearl, 2000).

To a large extent, however, recent philosophical discussion has been unsympathetic to manipulability theories: it is claimed both that they are unilluminatingly circular and that they lead to a conception of causation that is unacceptably anthropocentric or at least insufficiently general in the sense that it is linked much too closely to the practical possibility of human manipulation. (See, e.g., Hausman, 1986, 1998). Both objections seem prima-facie plausible. Suppose that *X* is a variable that takes one of two different values, 0 and 1, depending on whether some event of interest occurs. Then for an event or process *M* to qualify as a manipulation of *X*, it would appear that there must be a causal connection between *M* and *X*: to manipulate *X*, one must *cause* it to change in value. How then can we use the notion of manipulation to provide an account of causation? Moreover, it is uncontroversial that causal relationships can obtain in circumstances in which manipulation of the cause by human beings is not practically possible -- think of the causal relationship between the gravitational attraction of the moon and the motion of the tides or causal relationships in the very early universe. How can a manipulability theory avoid generating a notion of causation that is so closely tied to what humans can do that it is inapplicable to such cases?

As remarked above, the generally negative assessment of manipulability theories among philosophers contrasts sharply with the widespread view among statisticians, theorists of experimental design, and many social and natural scientists that an appreciation of the connection between causation and manipulation can play an important role in clarifying the meaning of causal claims and understanding their distinctive features. This in turn generates a puzzle. Are non-philosophers simply mistaken in thinking that focusing on the connection between causation and manipulation can tell us something valuable about causation? Does the widespread invocation of something like a manipulability conception among practicing scientist show that the usual philosophical criticisms of manipulability theories of causation are misguided?

The ensuing discussion is organized as follows. §§2 and 3 describe two of the best known philosophical

formulations of the manipulability theory -- those due to von Wright (1971) and Menzies and Price (1993) -- and explore certain difficulties with them. §4 argues that the notion of a free action cannot play the central role it is assigned in traditional versions of manipulability theories. §5 introduces the notion of an intervention which allows for a more adequate statement of the manipulability approach to causation and which has figured prominently in recent discussion. §6 considers Pearl's "interventionist" formulation of a manipulability theory and an alternative to it, due to Woodward. §§7 and 8 take up the charge that manipulability theories are circular. §9 returns to the relationship between interventions and human actions, while §10 compares manipulability accounts with David Lewis' closely related counterfactual theory of causation. Finally, §11 considers whether there are meaningful causal claims that cannot be captured by a manipulability theory.

As we shall see, the different assessments of manipulability accounts of causation within and outside of philosophy derive from the different goals or aspirations that underlie the versions of the theory developed by these two groups. Philosophical defenders of the manipulability conception have typically attempted to turn the connection between causation and manipulability into a reductive analysis: their strategy has been to take as primitive the notion of manipulation (or some related notion like agency or bringing about an outcome as a result of a free action), to argue that this notion is not itself causal (or at least does not presuppose all of the features of causality the investigator is trying to analyze), and to then attempt to use this notion to construct a non-circular reductive definition of what it is for a relationship to be causal. Philosophical critics have (quite reasonably) assessed such approaches in terms of this aspiration (i.e., they have tended to think that manipulability accounts are of interest only insofar as they lead to a non-circular analysis of causal claims) and have found the claim of a successful reduction unconvincing. By contrast, statisticians and other non-philosophers who have explored the link between causation and manipulation generally have not had reductionist aspirations-instead their interest has been in unpacking what causal claims mean and in showing how they figure in inference by tracing their interconnections with other related concepts (such as manipulation) but without suggesting that the notion of manipulation is itself a causally innocent notion.

It is the impulse toward reduction that generates the other feature that critics have found objectionable in standard formulations of the manipulability theory. To carry through the reduction, one needs to show that the notion of agency is independent of or prior to the notion of causality and this in turn requires that human actions or manipulations be given a special status-they can't be ordinary causal transactions, but must instead be an independent fundamental feature of the world in their own right. This both seems problematic on its own terms (it is prima-facie inconsistent with various naturalizing programs) and leads directly to the problem of anthropocentricity: if the only way in which we understand causation is by means of our prior grasp of an independent notion of agency, then it is hard to see what could justify us in extending the notion of causation to circumstances in which manipulation by human beings is not possible and the relevant experience of agency unavailable. As we shall see, both von Wright and Menzies and Price struggle, not entirely successfully, with this difficulty.

The way out of these problems is to follow writers like Pearl in reformulating the manipulability approach in terms of the notion of an intervention, where this is characterized in purely causal terms that make no essential reference to human action. Some human actions will qualify as interventions but they

will do so in virtue of their causal characteristics, not because they are free or carried out by humans. This "interventionist" reformulation allows the manipulability theory to avoid a number of counterexamples to more traditional versions of the theory. Moreover, when so reformulated, the theory may be extended readily to capture causal claims in contexts in which human manipulation is impossible. However, the price of such a reformulation is that we lose the possibility of a reduction of causal claims to claims that are non-causal. Fortunately (or so §§7 and 8 argue) an interventionist formulation of a manipulability theory may be non-trivial and illuminating even if it fails to be reductive.

2. von Wright

In his (1971) von Wright describes the basic idea of the manipulability approach as follows:

> … to think of a relation between events as causal is to think of it under the aspect of (possible) action. It is therefore true, but at the same time a little misleading to say that if *p* is a (sufficient) cause of *q*, then if I could produce *p* I could bring about *q*. For *that p* is the cause of *q*, I have endeavored to say here, *means* that I could bring about *q*, if I could do (so that) *p*. (p. 74)

To the objection that "doing" or "producing" is already a causal notion and hence not something to which we can legitimately appeal to elucidate the notion of causation, von Wright responds as follows:

> The connection between an action and its result is intrinsic, logical and not causal (extrinsic). If the result does not materialize, the action simply has not been performed. The result is an essential "part" of the action. It is a *bad* mistake to think of the act(ion) itself as a cause of its result. (pp. 67-8)

Here we see a very explicit attempt to rebut the charge that an account of causation based on agency is circular by contending that the relation between an action (or a human manipulation) and its result is not an ordinary causal relation. Moreover, von Wright readily embraces the further conclusion that seems to follow from this: human action must be a concept which, in our understanding of the world, is just as "basic" as the notion of causality (p. 74).

Given the logical structure of von Wright's views, it is also not surprising, to find him struggling to make sense of the idea that there can be causal relations involving events that human beings cannot in fact manipulate. He writes:

> The eruption of Vesuvius was the cause of the destruction of Pompeii. Man can through his action destroy cities, but he cannot, we think, make volcanoes erupt. Does this not prove that the cause-factor is not distinguished from the effect-factor by being in a certain sense capable of manipulation? The answer is negative. The eruption of a volcano and the destruction of a city are two very complex events. Within each of them a number of events

> or phases and causal connections between them may be distinguished. For example, that when a stone from high above hits a man on his head, it kills him. Or that the roof of a house will collapse under a given load. Or that a man cannot stand heat above a certain temperature. All these are causal connections with which we are familiar from experience and which are such that the cause-factor typically satisfies the requirement of manipulability. (p. 70)

von Wright's view is that to understand a causal claim involving a cause that human beings cannot in fact manipulate (e.g., the eruption of a volcano) we must interpret it in terms of claims about causes that human beings *can* manipulate (impacts of falling stones on human heads and so on). We will return to this general idea below in connection with Price and Menzies but it is worth noting that it faces an obvious problem. If we try to explain what it means to say that different galaxies attract one another gravitationally by contending that such interactions are in some relevant respects similar to gravitational interactions with which we are familiar or have experience (people and projectiles falling to earth), we need to explain what "similar" means and it is very hard to see how to do this within the framework of an agency theory. The relevant notion of similarity does not seem to be a notion that can be spelled out in terms of similarities in people's experiences of agency. Either we explain the relevant notion of similarity in straightforwardly causal terms that seem to have nothing to do with agency (e.g., we say that the similarity consists in the fact that the same gravitational force law is operative in both cases), in which case we have effectively abandoned the agency theory, or else we are led to the conclusion that causal claims involving unmanipulable causes like galaxies involve a conception of causality which is fundamentally different from the conception that is applicable to manipulable causes.

3. Menzies and Price

A very similar dialectic is at work in an extremely interesting recent paper by Peter Menzies and Huw Price (1993) (and in a series of papers written by Price alone, 1991, 1992) which represents the most detailed and sustained attempt in the recent philosophical literature to develop an "agency" theory of causation. Price and Menzies basic thesis is that:

> … an event *A* is a cause of a distinct event *B* just in case bringing about the occurrence of *A* would be an effective means by which a free agent could bring about the occurrence of *B*. (1993, p. 187)

They take this connection between free agency and causation to support a probabilistic analysis of causation (according to which "*A* causes *B*" can be plausibly identified with "*A* raises the probability of *B*") provided that the probabilities appealed to are what they call "agent probabilities," where

> [a]gent probabilities are to be thought of as conditional probabilities, assessed from the agent's perspective under the supposition that antecedent condition is realized ab initio , as a free act of the agent concerned. Thus the agent probability that one should ascribe to *B* conditional on *A* is the probability that *B* would hold were one to choose to realize *A*.

> (1993, p. 190)

The idea is thus that the agent probability of *B* conditional on *A* is the probability that *B* would have conditional on the assumption that *A* has a special sort of status or history -- in particular, on the assumption that A is realized by a free act. *A* will be a cause of *B* just in case the probability of *B* conditional on the assumption that *A* is realized by a free act is greater than the unconditional probability of *B*; *A* will be a spurious cause of *B* just in case these two probabilities are equal. As an illustration, consider a stock example of philosophers -- a structure in which atmospheric pressure, represented by a variable *Z*, is a common cause of the reading *X* of a barometer and the occurrence of a storm *Y*, with no causal relationship between *X* and *Y*. *X* and *Y* will be correlated, but Price's and Menzies' intuitive idea is that conditional on the realization of *X* by a free act, this correlation will disappear, indicating that the correlation between *X* and *Y* is spurious and does not reflect a causal connection from *X* to *Y*. If, by contrast, this correlation were to persist, this would be an indication that *X* was after all a cause of *Y*. (What "free act" might mean in this context will be explored below, but I take it that what is *intended* -- as opposed to what Price and Menzies actually say -- is that the manipulation of *X* should satisfy the conditions we would associate with an ideal experiment designed to determine whether *X* causes *Y* -- thus, for example, the experimenter should manipulate the position of the barometer dial in a way that is independent of the atmospheric pressure *Z*, perhaps by setting its value after consulting the output of some randomizing device.)

Like von Wright, Price and Menzies attempt to appeal to this notion of agency to provide a non-circular, reductive analysis of causation. They claim that circularity is avoided because we have a grasp of the *experience* of agency that is independent of our grasp of the general notion of causation.

> The basic premise is that from an early age, we all have direct experience of acting as agents. That is, we have direct experience not merely of the Humean succession of events in the external world, but of a very special class of such successions: those in which the earlier event is an action of our own, performed in circumstances in which we both desire the later event, and believe that it is more probable given the act in question than it would be otherwise. To put it more simply, we all have direct personal experience of doing one thing and thence achieving another. … It is this common and commonplace experience that licenses what amounts to an ostensive definition of the notion of 'bringing about'. In other words, these cases provide direct non-linguistic acquaintance with the concept of bringing about an event; acquaintance which does not depend on prior acquisition of any causal notion. An agency theory thus escapes the threat of circularity. (1993, p. 194-5)

Again like von Wright, Menzies and Price recognize that, once the notion of causation has been tied in this way to our "personal experience of doing one thing and hence achieving another" (1993, p. 194), a problem arises concerning unmanipulable causes. To use their own example, what can it mean to say that "the 1989 San Francisco earthquake was caused by friction between continental plates" (p. 195) if no one has (or given the present state of human capabilities could have) the direct personal experience of bringing about an earthquake by manipulating these plates? Their response to this difficulty is complex, but the central idea is captured in the following passages

> … we would argue that when an agent can bring about one event as a means to bringing about another, this is true in virtue of certain basic intrinsic features of the situation involved, these features being essentially non-causal though not necessarily physical in character. Accordingly, when we are presented with another situation involving a pair of events which resembles the given situation with respect to its intrinsic features, we infer that the pair of events are causally related even though they may not be manipulable. (1993, p. 197)

> Clearly, the agency account, so weakened, allows us to make causal claims about unmanipulable events such as the claim that the 1989 San Francisco earthquake was caused by friction between continental plates. We can make such causal claims because we believe that there is another situation that models the circumstances surrounding the earthquake in the essential respects and does support a means-end relation between an appropriate pair of events. The paradigm example of such a situation would be that created by seismologists in their artificial simulations of the movement of continental plates. (1993, p. 197)

The problem with this strategy parallels the difficulty with von Wright's broadly similar suggestion. What is the nature of the "intrinsic" but (allegedly) "non-causal" features in virtue of which the movements of the continental plates "resemble" the artificial models which the seismologists are able to manipulate? It is well-known that small scale models and simulations of naturally occurring phenomena that superficially resemble or mimic those phenomena may nonetheless fail to capture their causally relevant features because, for example, the models fail to "scale up" -- because causal processes that are not represented in the model become quite important at the length scales that characterize the naturally occurring phenomena. Thus, when we ask what it is for a model or simulation which contains manipulable causes to "resemble" phenomena involving unmanipulable causes, the relevant notion of resemblance seems to require that the same *causal* processes are operative in both. Price and Menzies provide no reason to believe that this notion of resemblance can be characterized in non-causal terms. But if the extension of their account to unmanipulable causes requires a notion of resemblance that is already causal in character and which, ex hypothesi cannot be explained in terms of our experience of agency, then their reduction fails.

It might be thought the difficulty under discussion can be avoided by the simple expedient of adhering to a counterfactual formulation of the manipulability theory. Indeed, it is clear that *some* counterfactual formulation is required if the theory is to be even remotely plausible: after all, no one supposes that *A* can only be a cause of *B* if *A* is in fact manipulated. Instead, the intuitive core of the manipulability theory should be formulated as the claim (**CF**):

> (**CF**) *A* causes *B* if and only if *B* would change if an appropriate manipulation on *A* *were* to be carried out.

The suggestion under consideration attempts to avoid the difficulties posed by causes that are not manipulable by human beings by contending that for (**CF**) to be true, it is not required that the manipulation in question be practically possible for human beings to carry out or even that human beings exist. Instead all that is required is that *if* human beings were to exist and to carry out the requisite manipulation of *A* (e.g. the continental plates), *B* (whether or not an earthquake occurs) would change. (The possibility of adopting such a counterfactual formulation is sympathetically explored, but not fully endorsed by Ernest Sosa and Michael Tooley in the introduction to their (1993))

One fundamental problem with this suggestion is that, independently of whether a counterfactual formulation is adopted, the notion of a free action or human manipulation cannot by itself, for reasons to be described in Section 4, do the work (that of distinguishing between genuine and spurious causal relationships) that Menzies and Price wish it to do. But in addition to this, a counterfactual formulation along the lines of (**CF**) seems completely unilluminating unless accompanied by some sort of account of how we are to understand and assess such counterfactuals and, more specifically, what sort of situation or possibility we are supposed to envision when we imagine that the antecedent of (**CF**) is true. Consider, for example, a causal claim about the very early universe during which temperatures are so high that atoms and molecules and presumably anything we can recognize as an agent cannot exist. What counterfactual scenario or possible world are we supposed to envision when we ask, along the lines of (**CF**), what would happen if human beings were to exist and were able to carry out certain manipulations in this situation? A satisfying version of an agency theory should give us an account of how our experience of agency in ordinary contexts gives us a purchase on how to understand and evaluate such counterfactuals. To their credit, von Wright and Price and Menzies attempt to do this, but in my view they are unsuccessful.

4. Causation and Free Action

As we have seen, Menzies and Price assign a central role to "free action" in the elucidation of causation. They do not further explain what they mean by this phrase preferring instead, as the passage quoted above indicates, to point to a characteristic experience we have as agents. It seems clear, however, that whether (as soft determinists would have it) a free action is understood as an action that is uncoerced or unconstrained or due to voluntary choices of the agent, or whether, as libertarians would have it, a free action is an action that is uncaused or not deterministically caused, the persistence of a correlation between *A* and *B* when *A* is realized as a "free act" is *not* sufficient for *A* to cause *B*. Suppose that, in the example described above, the position of the barometer dial *X* is set by a free act (in either of the above senses) of the experimenter but that that this free act (and hence *X*) is correlated with *Z*, the variable measuring atmospheric pressure, perhaps because the experimenter observes the atmospheric pressure and freely chooses to set *X* in a way that is correlated with *Z*. (This possibility is compatible with the experimenter's act of setting *X* being free in either of the above two senses.) In this case, *X* will remain correlated with *Y* when produced by a free act, even though *X* does not cause *Y*. Suppose, then, that we respond to this difficulty by adding to our characterization of *A's* being realized by a free act the idea that this act must not itself be correlated with any other cause of *A*. (Passages in Price, 1991 suggest such an additional proviso, although the condition in question seems to have nothing to do with the usual

understanding of free action.) Even with this proviso, it need not be the case that *A* causes *B* if *A* remains correlated with *B* when *A* is produced by an act that is free in this sense, since it still remains possible that the free act that produces *A* also causes *B* via a route that does not go through *A*. As an illustration, consider a case in which an experimenter's administration of a drug to a treatment group (by inducing patients to ingest it) has a placebo effect that enhances recovery, even though the drug itself has no effect on recovery. There is a correlation between ingestion of the drug and recovery that persists under the experimenter's free act of administering the drug even though ingestion of the drug does not cause recovery.

5. Interventions

Examples like those just described show that if we wish to follow Menzies and Price in defending the claim that if an association between *A* and *B* persists when *A* is given the right sort of "independent causal history" or is "manipulated" in the right way, then *A* causes *B*, we need to be much more precise by what we mean by the quoted phases. There have been a number of attempts to do this in the recent literature on causation. The basic idea that all of these discussions attempt to capture is that of a "surgical" change in *A* which is of such a character that if any change occurs in *B*, it occurs only as a result of its causal connection, if any, to *A* and not in any other way. In other words, the change in *B*, if any, that is produced by the manipulation of *A* should be produced only via a causal route that goes through *A*. Manipulations or changes in the value of a variable that have the right sort of surgical features have come to be called *interventions* in the recent literature (e.g. Spirtes, Glymour and Scheines, 1993, Meek and Glymour, 1994, Hausman, 1998, Pearl, 2000, Woodward, 1997, 2000, Woodward and Hitchcock, forthcoming, Cartwright, forthcoming) and I will follow this practice. The characterization of the notion of an intervention is rightly seen by many writers as central to the development of a plausible version of a manipulability theory . One of the most detailed attempts to think systematically about interventions and their significance for understanding causation is due to Pearl, 2000 and I turn now to a discussion of his views.

6. Pearl

Pearl uses systems of equations and directed graphs to represent causal relationships and his work provides a striking illustration of the heuristic usefulness of a manipulationist framework in specifying what it is to give such systems a causal interpretation.[1] Pearl characterizes the notion of an intervention by reference to a primitive notion of a causal mechanism. A functional causal model is a system of equations $X_i = F(Pa_i, U_i)$ where Pa_i represents the parents or direct causes of X_i that are explicitly included in the model and U_i represents an error variable that summarizes the impact of all excluded variables. Each equation represents a distinct causal mechanism which is understood to be "autonomous" in the sense in which that notion is used in econometrics; this means roughly that it is possible to interfere with or disrupt each mechanism (and the corresponding equation) without disrupting any of the others. The simplest sort of intervention in which some variable X_i is set to some particular value x_i

amounts, in Pearl's words, to "lifting X_i from the influence of the old functional mechanism $X_i = F_i$ (Pa_i, U_i) and placing it under the influence of a new mechanism that sets the value x_i while keeping all other mechanisms undisturbed." (Pearl, 2000, p. 70; I have altered the notation slightly). In other words, the intervention disrupts completely the relationship between X_i and its parents so that the value of X_i is determined entirely by the intervention. Furthermore, the intervention is surgical in the sense that no other causal relationships in the system are changed. Formally, this amounts to replacing the equation governing X_i with a new equation $X_i = x_i$, substituting for this new value of X_i in all the equations in which X_i occurs but leaving the other equations themselves unaltered. Pearl's assumption is that the other variables that change in value under this intervention will do so only if they are effects of X_i.

Following Pearl, let us represent the proposition that the value of *X* has been set by an intervention to some particular value, x_0, by means of a "*do*" operator ($do(X=x_0)$, or more simply, $do\ x_0$). It is important to understand that conditioning on the information that the value of *X* has been set to x_0 will in general be quite different from conditioning on the information that the value of *X* has been observed to be x_0. (See Meek and Glymour, 1994; Pearl, 2000.) For example, in the case in which *X* and *Y* are joint effects of the common cause *Z*, $P(Y/X=x_0) \neq P(Y)$; that is, *Y* and *X* are not independent. However, $P(Y/do(X=x_0) = P(Y)$; that is, *Y* will be independent of *X*, if the value of *X* is set by an intervention. This is because the intervention on *X* will break the causal connection from *Z* to *X*, so that the probabilistic dependence between *Y* and *X* that is produced by *Z* in the undisturbed system will no longer hold once the intervention occurs. In this way, we may capture Menzies' and Price's idea that *X* causes *Y* if and only if the correlation between *X* and *Y* would persist under the right sort of manipulation of *X*.

This framework allows for a simple definitions of various causal notions. For example, Pearl defines the "causal effect" of *X* on *Y* associated with the "realization" of a particular value *x* of *X* as:

(**C**) $P(y/do\ x)$,

that is, as the distribution that *Y* would assume under an intervention that sets the value of *X* to the value *x*. It is obvious that this is a version of a counterfactual account of causation.

One of the many attractions of this approach is that it yields a very natural account of what it is to give a causal interpretation to a system of equations of the sort employed in the so-called causal modeling literature. For example, if a linear regression equation $Y = aX + U$ makes a causal claim, it is to be understood as claiming that if an intervention were to occur that sets the value of $X=x_0$ in circumstances $U=u_0$, the value of *Y* would be $y = ax_0 + u_0$, or alternatively that an intervention that changes *X* by amount *dx* will change *Y* by amount *a dx*. As another illustration consider the system of equations

(1) $Y = aX + U$
(2) $Z = bX + cY + V$

We may rewrite these as follows

(1) $Y = aX + U$
(3) $Z = dX + W$

where $d = b + ac$ and $W = cU + V$. Since (3) has been obtained by substituting (1) into (2), the system (1)-(2) has exactly the same solutions in *X* , *Y* , and *Z* as the system (1)-(3). Since *X*, *Y* and *Z* are the only measured variables, (1)-(2) and (1)-(3) are "observationally equivalent" in the sense that they imply or represent exactly the same facts about the patterns of correlations that obtain among the measured variables. Nonetheless two systems correspond to different causal structures. (1)-(2) says that *X* is a direct cause of *Y* and that *X* and *Y* are direct causes of *Z*. By contrast, (1)-(3) says that *X* is a direct cause of *Y* and that *X* is a direct cause of *Z* but says nothing about a causal relation between *Y* and *Z*. We can cash this difference out within the interventionist/manipulationist framework described above -- (2) claims that an intervention on *Y* will change *Z* while (3) denies this. (Recall that an intervention on *Y* with respect to *Z* must not be correlated with any other cause of *Z* such as *X*, and will break any causal connection between *X* and *Y*.) Thus while the two systems of equations agree about the correlations so far observed, they disagree about what would happen under an intervention on *Y*. According to an interventionist/manipulationist account of causation, it is the system that gets such counterfactuals right that correctly represents the causal facts.

One possible limitation of Pearl's characterization of an intervention concerns the scope of the requirement that an intervention on X_i leave intact *all* other mechanisms besides the mechanism that previously determined the value of X_i. If, as Pearl apparently intends, we understand this to include the requirement that an intervention on X_i must leave intact the causal mechanism if any, that connects X_i to its possible effects *Y*, then an obvious worry about circularity arises, at least if we want to use the notion of an intervention to characterize what it is for X_i to cause *Y*. A closely related problem is that given the way Pearl characterizes the notion of an intervention, his definition (**C**) of the causal effect of *X* on *Y*, seems to give us not the causal contribution made by $X = x$ alone to *Y* but rather the combined impact on *Y* of this contribution and whatever contribution is made to the value of *Y* by other causes of *Y* besides *X*. For example, in the case of the the regression equation $Y = aX+U$, the causal effect in Pearl's sense of *X=x on Y* is apparently $P(Y) = ax + U$, rather than, as one might expect, just *ax*. In part for these reasons,Woodward (1997, 2000) and Woodward and Hitchcock (forthcoming)explore a different way of characterizing the notion of an intervention which does not make reference to the relationship between the variable intervened on and its effects. For Woodward and Hitchcock, in contrast to Pearl, an intervention *I* on a variable *X* is always defined with respect to a second variable *Y* (the intent being to use the notion of an intervention on *X* with respect to *Y* to characterize what it is for *X* to cause *Y*). Such an intervention *I* must meet the following requirements (**M1**) - (**M4**):

(**M1**) *I* must be the only cause of *X*; i.e., as with Pearl, the intervention must completely disrupt the causal relationship between *X* and its previous causes so that the value of *X* is set entirely by *I*,

(**M2**) *I* must not directly cause *Y* via a route that does not go through *X* as in the placebo example,

(**M3**) *I* should not itself be caused by any cause that affects *Y* via a route that does not go through *X*, and

(**M4**) *I* leaves the values taken by any causes of *Y* except those that are on the directed path from *I* to *X* to *Y* (should this exist) unchanged.

Within this framework, the most natural way of defining the notion of causal effect is in terms of the *difference* made to the value of *Y* by a change or difference in the value of *X*. Focusing on differences in this way allows us to isolate the contribution made to *Y* by *X* alone from the contribution made to *Y* by its other causes. Moreover, since in the non-linear case, the change in the value of *Y* caused by a given change in the value of *X* will depend on the values of the other causes of *Y*, it seems to follow that the notion of causal effect must be relativized to a background context B_i which incorporates information about these other values. In deterministic contexts, we might thus define the causal effect on *Y* of a change in the value of *X* from $X=x$ to $X=x'$ in circumstances B_i as:

(**CD**) $Y_{do\ x,\ B_i} - Y_{do\ x',\ B_i}$,

that is, as the difference between the value that *Y* would take under an intervention that sets $X=x$ in circumstances B_i and the value that *Y* would take under an intervention that sets $X=x'$ in B_i, where the notion of an intervention is now understood in terms of (**M1**) - (**M4**) rather than in the way recommended by Pearl. In non-deterministic contexts, the definition of causal effect is, analogously, $P_{do\ x,B_i}(Y) - P_{do\ x',B_i}(Y)$. In the deterministic case, *X* will then be a cause of *Y* in B_i if and only if the causal effect of *X* on *Y* in B_i is non-zero for some pair of values of *X* - that is, if and only if there are distinct values of *X*, *x* and x' such that the value of *Y* under an intervention that sets $X=x$ in B_i is different from the value of *Y* under an intervention that sets $X=x'$. In probablistic contexts, *X* will be a cause of *Y* if the probability distribution of *Y* is different for two different values of *X*, when these are set by interventions.

I will not attempt to adjudicate here among these and various other proposals concerning the best way to characterize the notions of intervention and causal effect. Instead, I want to comment on the general strategy they embody and to compare it with the approach to causation associated with theorists like Menzies and Price. Note first that the notion of an intervention, when understood along either of the lines described above, is an unambiguously causal notion in the sense that causal notions are required for its characterization -- thus the proposals variously speak of an intervention on *X* as breaking the causal connection between *X* and its causes while leaving other causal mechanisms intact or as not affecting *Y* via a causal route that does not go through *X*. This has the immediate consequence that one cannot use the notion of an intervention to provide a reduction of causal claims to non-causal claims. Moreover, to the extent that reliance on some notion like that of an intervention is unavoidable in any satisfactory version of a manipulability theory (as I believe that it is), any such theory must be non-reductionist. Indeed, we can now see that critics who have charged manipulability theories with circularity have in one

important sense understated their case: manipulability theories turn out to be "circular" not just in the obvious sense that for an action or event *I* to constitute a an intervention on a variable *X*, there must be a causal relationship between *I* and *X*, but in the sense that *I* must meet a number of other causal conditions as well.

7. Is Circularity a Problem?

Suppose that we agree that any plausible version of a manipulability theory must make use of the notion of an intervention and that this must be characterized in causal terms. Does this sort of "circularity" make any such theory trivial and unilluminating? It seems to me that it does not, for at least two reasons. First, it may be, as writers like Woodward contend, that in characterizing what it is for a process *I* to qualify as an intervention on *X* for the purposes of characterizing what it is for *X* to cause *Y*, we need not make use of information about the causal relationship, if any, between *X* and *Y*. Instead, it may be that we need only to make use of *other* sorts of causal information, e.g., about the causal relationship between *I* and *Y* or about whether *I* is caused by causes that cause *Y* without causing *X*, as in (**M1**) - (**M4**) above. To the extent that this is so, we may use one set of claims about causal relationships (e.g., that *X* has been changed by a process that meets the conditions for an intervention) together with correlational information (that *X* and *Y* remain correlated under this change) to characterize what it is for a different relationship (the relationship between *X* and *Y*) to be causal. This does not yield a reduction of causal talk to non-causal talk, but it is also not viciously circular in the sense that it presupposes that we already have causal information about the very relationship that we are trying to characterize. One reason for thinking that there must be *some* way of characterizing the notion of an intervention along the lines just described is that we do sometimes learn about causal relationships by performing experiments -- and it is not easy to see how this is possible if to characterize the notion of an intervention on *X* we had to make reference to the causal relationship between *X* and its effects.

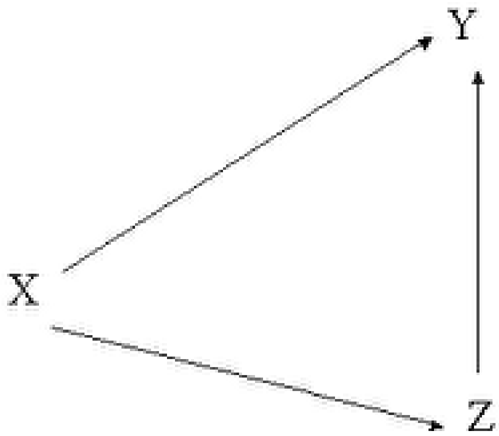
8. The Plurality of Causal Concepts

A second respect in which reliance on the notion of an intervention need not be thought of as introducing a vicious circularity is this: So far, I have been following von Wright and Menzies and Price in assuming that there is just one causal notion or locution (*A* causes *B*, where *A* and *B* are types of events) that we are trying to analyze. But in fact there are many such notions. For example, among causal notions belonging to the family of so-called type causal notions (i.e., causal claims that relate types of events or variables) there is a distinction to be drawn between what we might call claims about total or net causes and claims about direct causes. Even if the notion of an intervention presupposes some causal notion such as some notion of type causation, it may be that we can use it to characterize other causal notions.

As an illustration consider the causal structure represented by the following equations and associated directed graph

$$Y = aX + cZ$$

$Z = bX$



In this structure, there are two different causal routes from *X* to *Y* -- a direct causal relationship and an indirect relationship with *Z* as an intermediate variable. If $a = -bc$, there is cancellation along these two routes. This means that no intervention on *X* will change the value of *Y*. In one natural sense, this seems to mean that *X* does not cause *Y*, just as (**C**) (§6) suggests. In another natural sense, however, *X* does seem to be a cause -- indeed a direct cause- of *Y*. We can resolve this apparent inconsistency by distinguishing between two kinds of causal claims[2] -- the claim *X* is a total or net cause of *Y*, where this is captured by (**C**)or (**CD**), and the claim that *X* is a direct cause of *Y*, where this is understood along the following lines: *X* is a direct cause of *Y* if and only if under some intervention that changes the value of *X*, the value of *Y* changes when all other variables in the system of interest besides *X* and *Y* including those that are on some causal route from *X* to *Y*, are held fixed at some value, also by interventions. (For related, but different, characterizations of direct causation along these lines, see Pearl, 2000 and Woodward, forthcoming) Fixing the other values of other variables means that each of these values are determined by separate processes, each meeting the conditions for an intervention, that are appropriately independent of each other and of the intervention that changes the value of *X*. The effect of intervening to fix the values of these variables is thus that each variable intervened on is disconnected from its causes, including *X*. In the example under discussion, *X* qualifies as a direct cause of *Y* because if we were to fix the value of *Z* in a way that disconnects it from the value of *X*, and then intervene to change the value of *X*, the value of *Y* would change. This idea can then be generalized to provide a characterization of "contributing" causation or causation along a causal route, i.e., to capture the sense in which *X* is an indirect cause of *Y* along the route that goes through *Z*. (Woodward, forthcoming).

One can also use the notion of an intervention to characterize what it is for the fact or event of *X*'s taking on some particular value to be an actual or token cause of *Y*'s taking on a particular value (as opposed to *X*'s being a type cause of *Y*) and in this way help to clarify the relationship between type and token causation. (See Pearl 2000, Hitchcock, 2001, Woodward, forthcoming). Thus even if a "manipulationist" or "interventionist" framework does not yield a reduction of causal talk to non-causal talk, it provides a natural way of marking the distinctions among a number of different causal notions and exhibiting their interrelations. More generally, even if a manipulationist account of causation does not yield a reduction but instead simply connects "causation" (or better, various more specific causal concepts) with other concepts within the same circle, we still face many non-trivial choices about how the concepts on this

circle are to be elucidated and connected up with one another. For example, it is far from obvious how to characterize the notion of an intervention so as to avoid the various counterexamples to standard statements of the manipulability theory such as the theory of Menzies and Price. It is in part because the notion of manipulation/intervention has an interesting and complex fine structure-a structure that is left largely unexplored in traditional manipulability theories-that working out the connection between causation and manipulation turns out to be interesting and non-trivial rather than banal and obvious.

9. Interventions That Do Not Involve Human Action

We noted above that a free action need not meet the conditions for an intervention, on any of the conceptions of intervention described in §6. It is also true that a process or event can qualify as an intervention even if it does not involve human action or intention at any point. This should be apparent from the way in the notion of an intervention has been characterized, for this is entirely in terms of causal and correlational concepts and makes no reference to human beings or their activities. In other words, a purely "natural" process involving no animate beings at all can qualify as an intervention as long as it has the right sort of causal history -- indeed, this sort of possibility is often described by scientists as a natural experiment. Moreover, even when manipulations are carried out by human beings, it is the causal features of those manipulations and not the fact that they are carried out by human beings or are free or are attended by a special experience of agency that matters for recognizing and characterizing causal relationships. Thus, by giving up any attempt at reduction and characterizing the notion of an intervention in causal terms, an "interventionist approach of the sort described under §§5 and 6 avoids the second classical problem besetting manipulability theories -- that of anthropocentrism and commitment to a privileged status for human action. For example, under this approach *X* will qualify as a (total) cause of *Y* as long as it is true that for some value of *X* that if *X* were to be changed to that value by a process having the right sort of causal characteristics, the value of *Y* would change. Obviously, this claim can be true even if human beings lack the power to manipulate *X* or even in a world in which human beings do not or could not exist. There is nothing in the interventionist version of a manipulability theory that commits us to the view that all causal claims are in some way dependent for their truth on the existence of human beings or involve a "projection" on to the world of our experience of agency.

10. Interventions and Counterfactuals

We noted above that interventionist versions of manipulability theories are counterfactual theories. What is the relationship between such theories and more familiar versions of counterfactual theories such as the theory of David Lewis? Lewis' theory is an account of what it is for one individual token event to cause another while (**C**) is formulated in terms of variables or types of events , but abstracting away from this and certain other differences, there are a number of striking similarities between the two approaches. As readers of Lewis will be aware, any counterfactual theory must explain what we should envision as changed and what should be held fixed when we evaluate a counterfactual the antecedent of which is not true of the actual world -- within Lewis' framework, this is the issue of which worlds in which the antecedent of the counterfactual holds are "closest" or "most similar" to the actual world. Lewis' answer

to this question invokes a "similarity" ordering that ranks the importance of various respects of resemblance between worlds in assessing overall similarity. (Lewis, 1979). For example, avoiding diverse, widespread violations of law is said to be the most important consideration, preserving perfect match of particular fact over the largest possible spatio-temporal region is next in importance and more important than avoiding small localized violations of law, and so on. As is well-known the effect of this similarity ordering is, at least in most situations, to rule out so-called "back-tracking" counterfactuals (e.g., the sort of counterfactual that is involved in reasoning that if the effect of some cause had not occurred, then the cause would not have occurred). When the antecedent of a counterfactual is not true of the actual world, Lewis' similarity metric leads us (at least in deterministic contexts) to think of that antecedent as made true by a "small" miracle.

The notion of an intervention plays a very similar role within manipulability theories of causation to Lewis' similarity ordering. Like Lewis' ordering, the characterization of an intervention tells us what should be envisioned as changed and what should be held fixed when we evaluate a counterfactual like "If *X* were to be changed by an intervention to such and such a value, the value of *Y* would change". (For example, on Pearl's understanding of an intervention, in evaluating this counterfactual, we are to consider a situation in which the previously existing causal relationship between *X* and its causes is disrupted, but all other causal relationships in the system of interest are left unchanged.) A moment's thought will also show that, as in Lewis' account, both Pearl's and Woodward's characterizations of interventions rule out backtracking counterfactuals -- for example, in evaluating a counterfactual of the form "if an intervention were to occur that changes *E*, (where *E* is an effect of *C*), then *C* would change", Pearl holds that we should consider a situation in which the relationship between *E* and its causes (in this case, *C*) is disrupted, but all other causal relationships are left unchanged, so that *C* still occurs, and the above counterfactual is false, as it should be. Moreover, there is a clear similarity between Lewis' idea that the appropriate counterfactuals for analyzing causation are counterfactuals the antecedents of which are made true by miracles, and the idea of an intervention as an exogenous change that disrupts the mechanism that was previously responsible for the cause event *C*. Indeed, one might think of an interventionist treatment of causation as explaining why Lewis' account with its somewhat counterintuitive similarity ordering works as well as it does -- Lewis' account works because his similarity ordering picks out roughly those relationships that are stable under interventions and hence exploitable for purposes of manipulation and control and, as a manipulability theory claims, it is just these relationships that are causal. This is not to say, however, that the two approaches always yield identical assessments of particular causal and counterfactual claims -- Hitchcock and Woodward, forthcoming and Woodward, forthcoming describe cases in which the two approaches diverge and in which the interventionist approach seems more satisfactory.[3].

11. The Scope of a Manipulability Theory

In the version of a manipulability theory considered under §6 above, causal claims are elucidated in terms of counterfactuals about what would happen under interventions. As we have seen, the notion of an intervention should be understood without reference to human action, and this permits formulation of a manipulability theory that applies to causal claims in situations in which manipulation by human beings

is not a practical possibility. Moreover, the counterfactual formulation allows us to make sense of causal claims in contexts in which interventions do not in fact occur and arguably even in cases in which they are causally impossible, as long as we have some principled basis for answers to questions about what *would* happen to the value of some variable *if* an intervention were to occur on another variable. Consider, for example, the (presumably true) causal claim (**G**):

> (**G**) The gravitational attraction of the moon causes the motion of the tidest.

Human beings cannot at present alter the attractive force exerted by the moon on the tides (e.g., by altering its orbit). More interestingly, it may well be that there is no physically possible process that will meet the conditions for an intervention on the moon's position with respect to the tides -- all possible processes that would alter the gravitational force exerted by the moon may be insufficiently "surgical". For example, it may very well be that any possible process that alters the position of the moon by altering the position of some other massive object will have an independent impact on the tides in violation of condition (**M2**) for an intervention. It is nonetheless arguable we have a principled basis in Newtonian mechanics and gravitational theory themselves for answering questions about what would happen if such a surgical intervention were to occur and that this is enough to vindicate the causal claim (**G**).

Although a properly formulated version of a manipulability theory will thus allow us to talk about causal relationships in some contexts in which interventions are not physically possible, there are plausible arguments that such theories do place some restrictions on which relationships qualify as causal. I conclude by considering several cases of this sort, which will also serve to bring out an additional distinctive feature of manipulability theories.

12. (Alleged) Causes That Are Unmanipulable for Logical, Conceptual, or Metaphysical Reasons

Several statisticians (e.g., Holland, 1986, Rubin, 1986) who advocate manipulationist or counterfactual accounts of causation have held that causal claims involving causes that are unmanipulable in principle are defective or lack a clear meaning -- they think of this conclusion as following directly from a manipulationist approach to causation. What is meant by an unmanipulable cause is not made very clear, but the examples discussed typically involve alleged causes (e.g., race, or membership in a particular species, or perhaps gender) for which we lack any clear conception of what would be involved in manipulating them or any basis for assessing what would happen under such a manipulation. Such cases contrast with the case involving (**G**) above, where the notion of manipulating the moon's orbit seems perfectly clear and well-defined, and the problem is simply that the world happens to be arranged in such a way that an intervention that produces such a change is not physically possible.

A sympathetic reconstruction of the position under discussion might go as follows. On a manipulationist account of causation, causes (whether we think of them as events, types of events, properties, facts, or what have you) must be representable by means of *variables* -- where this means, at a minimum, that it

must be possible for the cause to change or to assume different values. This is required if we are to have a well-defined notion of manipulating a cause and well-defined answers to counterfactual queries about what would happen if the cause were to be manipulated in some way -- matters which are central to what causal claims mean on any version of a manipulability theory worthy of the name. Philosophers tend to think of causes as properties or events but in many cases, it is straightforward to move back and forth between such talk and a representation in terms of variables, as we have been doing throughout this entry. For example, rather than saying that the impact of the baseball caused the window to shatter or that impacts of baseballs cause window shatterings, we may introduce two indicator variables -- *I* which takes the values 0 and 1 for {no impact, impact} and *S* which takes the values 0 and 1 for {no shattering, shattering} and use these variables to express the idea that whether or not the window shatters is counterfactually dependent on (interventions that determine) whether the impact occurs. Both *I* and *S* describe causes that are straightforwardly manipulable. However, for some putative causes, there may be no well-defined notion of change or variation in value and if so, a manipulability theory will not count these as genuine causes. For example, if it is metaphysically necessary that everything that exists is a physical object or if we lack any coherent conception of what it is for something to exist but to be non-physical, then there will be no well-defined notion of intervening to change whether something is a physical object. While there are true (and even lawful) generalizations about all physical objects, on a manipulability theory these will not describe causal relationships. Thus, although to the best of our knowledge, it is a law of nature that (**L**) no physical object can be accelerated from a velocity less than that of light to a velocity greater than light, (**L**) is not, according to a manipuability theory, a *causal* generalization.

Moreover, even with respect to variables that can take more than one value, the notion of an intervention or manipulation will not be well-defined if there is no well-defined notion of *changing* the values of that variable. Suppose that we introduce a variable "animal" which takes the values {lizard, kitten, raven}. By construction, this variable has more than one value, but if, as seems plausible, we have no coherent idea of what it is to change a raven into lizard or kitten, there will be no well-defined notion of an intervention for this variable and being an animal (or being a raven) will not be the sort of thing that can count as a bona-fide cause on a manipulability theory. The notion of changing the value of a variable seems to involve the idea of an alteration from one value of the variable to another in circumstances in which the very same system or entity can possess both values and this notion seems inapplicable to the case under discussion.

Some readers will take it to be intuitively obvious that being a raven can be a cause, e.g., of some particular organism's being black. Many standard theories of causation also endorse this conclusion, for example, if we are willing to assume it is a law that all ravens are black, then nomological theories of causation will support the claim (**R**):

> (**R**) Ravenness causes blackness.

Similarly, ravenness raises the probability of blackness and hence (**R**) qualifies as causal on probabilistic theories of causation, and depending on how the relevant similarity ordering is understood, (**R**) may also qualify as causal on a Lewis style counterfactual theory. If causal claims like (**R**) are true, it is an

important inadequacy in manipulability theories that they seem unable to capture such claims. By contrast, others will think that claims like (**R**) are, if not false, at least unclear and unperspicuous, and that it is a point in favor of manipulability theories that they explain why this is the case. Those who take this second view will think that claims like (**R**) should be replaced by claims that involve causes that are straightforwardly manipulable. For example, (**R**) might be replaced by a claim that identified the genetic factors and biochemical pathways that are responsible for raven pigmentation -- factors and pathways for which there is a well-defined notion of manipulation and which are such that if they were appropriately manipulated, this would lead to changes in pigmentation. Manipulability theorists like Rubin and Holland will think that such a replacement would be clearer and more perspicuous than the original claim (**R**). In any case, claims involving causes that are unmanipulable in the sense that we seem to lack any clear conception of what would be involved in manipulating them are one important sort of case in which a manipulability approach will diverge from many other standard theories of causation.

Consider an additional illustration of this general theme. Holland (1986) appeals to a manipulability theory of causation to argue that the following claim is fundamentally unclear.

> (**F**) Being female causes one to be discriminated against in hiring and/or salary

In contrast to the previous cases, the problem here is not so much that under all interpretations of the putative cause ("being female") we lack any clear idea of what it would be like to manipulate it, but rather that there are several rather different things that might be meant by manipulation of "being female" (which from the perspective of a manipulability theory is to say that there several quite different variables we might have in mind when we talk about being female as a cause) and the consequences for discrimination of manipulating each of these may be quite different. For example, (**F**) might be interpreted as claiming that a literal manipulation of gender, as in a sex change operation, that leaves an applicant's qualifications otherwise unchanged, will change expected salary or probability of hiring. Alternatively, and more plausibly, (**F**) might be interpreted as claiming that manipulation of a potential employer's *beliefs* about applicant's gender will change salary and hiring probability, in which case (**F**) would be more perspicuously expressed as the claim that employer beliefs about gender cause discrimination. Still another possible interpretation -- in fact what Holland claims one *ought* to mean by (**F**) -- is that differentials in salary and hiring between men and women would disappear (or at least be reduced substantially) under a regime in which various sorts of biased practices were effectively eliminated, presumably as the result of changes in law and custom. While I see no reason to follow Holland in thinking that this is the only legitimate interpretation of (**F**), it is plainly a legitimate interpretation. Moreover, Holland is also correct to think that this last hypothetical experiment which involves manipulating the legal and cultural framework in which discrimination takes place is a quite different experiment from an experiment involving manipulating gender itself or employee beliefs about gender and that each of these experiments is likely to lead to different outcomes. From the perspective of a manipulability theory, these different experiments thus correspond to different causal claims. As this example illustrates, part of the heuristic usefulness of a manipulability theory is that encourages us to clarify or disambiguate causal claims by explicitly distinguishing among different possible claims about the outcomes of hypothetical experiments that might be associated with them. That we can clarify the meaning of a causal claim in this way is just what we would expect if a manipulability account of

causation is correct.

Bibliography

- Cartwright, N. (forthcoming): "Two Notions of Intervention, Two Notions of Invariance and Two Notions of What it is For a Causal Claim To Be Correct"
- Collingwood, R.(1940): *An Essay on Metaphysics*. Oxford: Clarendon Press.
- Cook, T. and Campbell, D. (1979): *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Miflin Company.
- Gasking, D. (1955): "Causation and Recipes", *Mind*, **64**, pp. 479-487.
- Haavelmo, T. (1944): "The Probability Approach in Econometrics", *Econometrica*, **12** (Supplement).
- Hauseman, D. (1986): "Causation and Experimentation" *American Philosophical Quarterly* **23**, pp. 143-54
- Hausman, D. (1998): *Causal Asymmetries*. Cambridge: Cambridge University Press
- Hitchcock, C. (2001): " The Intransitivity of Causation Revealed in Equations and Graphs", *The Journal of Philosophy*, pp. **98**, 273- 299.
- Hitchcock, Forthcoming, "A Tale of Two Effects".
- Holland, P. (1986): "Statistics and Causal Inference", *Journal of the American Statistical Association*, **81**,pp. 945-960.
- Lewis, D. (1973): "Causation", *Journal of Philosophy*, **70**, pp. 556-567.
- Lewis, D. (1979): "Counterfactuals Dependence and Time's Arrow", *Nous*, **13**, pp. 455-76.
- Meek, C. and Glymour, C. (1994): "Conditioning and Intervening", *British Journal for the Philosophy of Science*, **45**, pp. 1001-1021.
- Menzies, P. and Price, H. (1993): "Causation as a Secondary Quality", *British Journal for the Philosophy of Science*, **44**, pp. 187-203.
- Pearl, J. (2000): *Causality*. New York: Cambridge University Press, New York.
- Price, H. (1991): "Agency and Probabalistic Causality", *British Journal for the Philosophy of Science*, **42**, pp. 157 -76.
- Rubin, D. (1986): "Comment: Which Ifs Have Causal Answers?", *Journal of the American Statistical Association*, **81**, pp. 961-962.
- Sosa, E. and Tooley, M. (eds.)(1993): *Causation*. Oxford: Oxford University Press.
- Spirtes, P., Glymour, C. and Scheines, R.(1993): *Causation, Prediction and Search*. New York: Springer-Verlag.,
- von Wright, G.(1971): *Explanation and Understanding*. Ithica, New York: Cornell University Press.
- Woodward, J. (1997): "Explanation, Invariance, and Intervention" *PSA* 1996, volume 2, pp.S26-41.
- Woodward, J. (2000): "Explanation and Invariance in the Special Sciences’, *British Journal for the Philosophy of Science*, **51**, pp. 197-254.
- Woodward, J. and Hitchcock, C. (Forthcoming): "Explanatory Generalizations: A Counterfactual Account", *Nous*.

- Woodward, J. (Forthcoming): *A Theory of Explanation*. New York: Oxford University Press.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

causation: causal processes | causation: counterfactual theories of | causation: probabilistic

James Woodward
jfw@hss.caltech.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 17, 2001
Content last modified: August 17, 2001

Stanford Encyclopedia of Philosophy
Notes to Manipulability and Causation

Notes

1. As Pearl readily acknowledges, his work draws on a long tradition in econometrics of interpreting equations that express causal claims as claims about the outcomes of hypothetical experiments - see, e.g., Haavelmo, 1944.

2. For a related distinction, see Hitchcock, forthcoming.

3. More accurately, the interventionist account of type causation diverges from what seems to be the natural way of extending Lewis' theory to such causes. Consider a simple example discussed in Woodward, forthcoming. *C* is a deterministic direct (type) cause of *E* but also deterministically causes *E* indirectly by means of *n* causal routes that go through $C_1,\ldots,C_n$. Consider the counterfactual (1) "If $C_1,\ldots,C_n$ had not occurred, *E* would not have occurred". As explained above, any counterfactual theory will need to employ such counterfactuals to capture the notion of direct cause or causation along a route. On the interventionist account of the relationship between causal claims and counterfactuals, (1) is false, since under the assumption of the antecedent of (1), *C* will still occur and will cause *E*. Intuitively, this the correct assessment of (1). Under Lewis' theory, we have a choice between two different possible worlds that realize the antecedent of (1). In the first *C* occurs and each of the *n* links between *C* and $C_1,\ldots,C_n$ are broken. This requires *n* distinct miracles. In the second world, *C* fails to occur and hence $C_1,\ldots,C_n$ also fail to occur. This second world has less perfect match with the actual world than the first world, but involves only one miracle. At least for large *n*, Lewis' similarity ordering tells us that this second world is closer to the actual world. Thus (1) comes out true.

James Woodward
jfw@hss.caltech.edu

First published: August 17, 2001
Content last modified: August 17, 2001

Gersonides

Perhaps no other medieval Jewish philosopher has been so maligned over the centuries as Gersonides (Levi ben Gerson, acronym Ralbag). Indeed, his major philosophical work, *Sefer Milhamot Ha-Shem* (*The War of the Lord*, 1329), was called "*Wars against the Lord*" by one of his opponents. Despite the vilification of his position, Gersonides emerges as one of the most significant and comprehensive thinkers in the medieval Jewish tradition. He has been constantly quoted (even if only to be criticized), and, through the works of Hasdai Crecas and others, Gersonides' ideas have influenced such thinkers as Gottfried Wilhelm Leibiniz and Benedict de Spinoza. This article will survey his major contributions to medieval philosophy.

1. Introduction

In the introduction to his recently completed translation of *Wars of the Lord*, Feldman suggests that the significance of Gersonides lies in his emphasis upon "religious rationalism in Judaism." According to Feldman, we see a man who "has taken seriously the fact that he has reason, who believes that this faculty is God-given, and who attempts to understand God with this instrument" (*Wars*, p. 52). Gersonides is the philosopher who attempted to show that philosophy and Torah, that reason and revelation are co-extensive; he is a philosophical optimist who believes that reason was fully competent to attain all the important and essential truths. Thus, according to Feldman, Gersonides is "a most vigorous and consistent defender of human reason in religion"(*Wars*, p. 53).

This trust is reflected in Gersonides' introductory remarks to *Wars*. There, Gersonides upholds the primacy of reason, attributing to Maimonides the position that "we must believe what reason has determined to be true. If the literal sense of the Torah differs from reason, it is necessary to interpret those passages in accordance with the demands of reason" (*Wars*, p. 98). Gersonides believes that reason and Torah cannot be in opposition: "if reason causes to affirm doctrines that are incompatible with the literal sense of Scripture, we are not prohibited by the Torah to pronounce the truth on these matters, for reason is not incompatible with the true understanding of the Torah" (*ibid.*). Thus reason is upheld as a criterion for achieving truth.

2. Biography

Gersonides left few letters and does not talk about himself in his writings; nor is his life discussed at great length by his contemporaries. Hence, what is known of his biography is sketchy at best. Levi ben Gerson was born in 1288 in Provence and may have lived for a time in Bagnol sur-Ceze. It is probable that his father was Gershom ben Salomon de Beziers, a notable mentioned in medieval histories. With the decline of Spanish Judaism in the thirteenth century, Provence quickly became the cultural center for Jewish intellectual activity. The popes in Avignon had a lenient policy toward the Jews, whose creative life flourished, particularly in philosophy and theology. Jewish philosophers did not have direct access to the works of Aristotle, but Provençal Jews learned of Aristotle through the commentaries of Averroes, the twelfth-century Spanish Muslim philosopher. By the end of the thirteenth century these commentaries had been translated from Arabic into Hebrew, and Averroes' thought, as well as that of Aristotle, was being integrated into the mainstream of Jewish philosophy.

Gersonides may have married a distant cousin; it is not known whether he had any offspring. Although Gersonides spoke Provençal, his works are all written in Hebrew, and all of his quotations from Averroes, Aristotle, and Moses Maimonides are in Hebrew as well. He may have had a reading knowledge of Latin; he appears to manifest an awareness of contemporary Scholastic discussions. He might, however, have learned of such discussions in oral conversations with his Christian contemporaries. Apart from several trips to Avignon, Gersonides most likely resided his entire life in Orange. There is some evidence that he may have followed the traditional occupation of his family, moneylending. He died on 20 April 1344.

In addition to Averroes and Aristotle, Gersonides was influenced by Moses Maimonides, his greatest Jewish philosophical predecessor. Gersonides' works can be seen as an attempt to integrate the teachings of Aristotle, as mediated through Averroes and Maimonides, with those of Judaism. In *Milhamot Ha-Shem* he laid down the general rule that "the Law cannot prevent us from considering to be true that which our reason urges us to believe." His adherence to this principle is reflected throughout his work.

3. Major Works

What distinguished Gersonides from his predecessors was his reliance upon and consummate knowledge

of mathematics, coupled with his belief in the accuracy of observations achieved by the use of good instruments. Because of this rootedness in empirical observation bolstered by mathematics, Gersonides believed that he had the tools to succeed where others had failed. Only when he has resolved the problems in astronomy does Gersonides apply their findings to theological cosmology. As we shall see, Gersonides' theology and astronomy are deeply involved with each other.

This realist stance is stated in the context of examining al-Bitruji's astronomical proposals. Gersonides' contention is that "no argument can nullify the reality that is perceived by the senses, for true opinion must follow reality but reality need not conform to opinion" (Goldstein, 1974, p. 24). That Gersonides clearly considered his own observations to be the ultimate test of his system is explicit from his attitude towards Ptolemy. The importance of empirical observation cannot be underestimated, he claims, and he values his own observations over those of others. "We did not find among our predecessors from Ptolemy to the present day observations that are helpful for this investigation except our own" (*Wars*, V.1.3, p. 27), he says in describing his method of collecting astronomical data. Often his observations do not agree with those of Ptolemy, and in those cases he tells us explicitly that he prefers his own. Gersonides lists the many inaccuracies he has found trying to follow Ptolemy's calculations. Having investigated the positions of the planets, for example, Gersonides encountered "confusion and disorder" which led him to deny several of Ptolemy's planetary principles (Goldstein, 1988, p. 386). He does warn his colleagues, however, to dissent from Ptolemy only after great diligence and scrutiny.

Gersonides' scientific works comprise mathematics and astronomy. His *Sefer Ma'aseh Hoshev* (*The Work of a Counter*, 1321) is concerned with arithmetical operations and uses of a symbolic notation for numerical variables. Gersonides' major scientific contributions were in astronomy; his works were known by his contemporaries and influenced later astronomers. His astronomical writings are contained primarily in book 5, part 1 of *Milhamot Ha-Shem*. In 136 chapters Gersonides reviews and criticizes astronomical theories of the day, compiles astronomical tables, and describes one of his astronomical inventions. This instrument, which he called *Megalle 'amuqqot* (*Revealer of Profundities*) and which was called *Bacullus Jacobi* (*Jacob's staff*) by his Christian contemporaries, was used to measure the heights of stars above the horizon. The astronomical parts of *Milhamot Ha-Shem* were translated into Latin during Gersonides' lifetime. One of the craters of the moon, *Rabbi Levi*, is named after him.

Gersonides was well known as a Halakhist, one who deals with the intricacies of Jewish law. From this respect, his greatest contribution to Judaica was in the area of biblical commentary. His commentary on the Book of Job, completed in 1325, proved to be one of his most popular works and was one of the earliest Hebrew books to be published (in Ferrara, 1477). The commentary, which complements book 4 of *Milhamot Ha-Shem*, is concerned with the problem of divine providence. Each of the characters in the Book of Job represents a different theory of divine providence; Gersonides' own position is a restatement of Elihu's theory that providence is not directed to particulars but rather to groups of individuals, or universals.

Gersonides also wrote a logical treatise, *Sefer Ha-heqesh Ha-yashar* (*On Valid Syllogisms*, 1319), in which he examines problems associated with Aristotle's modal logic as developed in the Prior Analytics. This treatise was translated into Latin at an early date, although Gersonides' name was not attached to it.

Gersonides' major philosophical work, *Milhamot Ha-Shem*, was completed in 1329; it had been twelve years in the making. In 1317 Gersonides began an essay on the problem of creation. This problem, which has vexed Jewish philosophers since Philo Judaeus, had recently received elaborate treatment by Maimonides. Gersonides was dissatisfied with Maimonides' discussion and proposed to reopen the issue. This project was soon laid aside, however, for he felt that it could not be adequately discussed without proper grounding in the issues of time, motion, and the infinite. By 1325 his manuscript had developed to include discussion not only of creation but also of immortality, divination, and prophecy. By 1328 it included a chapter on providence as well. Books 5 and 6 were completed, by Gersonides' own dating, by 1329.

As Isaac Husik has pointed out, Gersonides "has no use for rhetorical flourishes and figures of speech … the effect upon the reader is monotonous and wearisome." His style has been compared to that of Thomas Aquinas and even of Aristotle in its use of a precise, technical vocabulary which eschews examples. In contradistinction to Maimonides, who introduced allegory, metaphor, and imprecise language into his work to convey the ambiguity of the subject matter, Gersonides saw it as his function to elucidate the issues as clearly as possible. Gersonides is the first Jewish philosopher to use this analytic, scholastic method.

4. Major Themes in *Milhamot Ha-shem*

In the introduction to *Milhamot*, Gersonides specifies six questions which he hopes to examine: Is the rational soul immortal? What is the nature of prophecy? Does God know particulars? Does divine providence extend to individuals? What is the nature of astronomical bodies? Is the universe eternal or created? Each question occupies a separate book. Gersonides attempts to reconcile traditional Jewish beliefs with what he feels are the strongest points in Aristotle's philosophy. Although a synthesis of these systems is his ultimate goal, philosophy often wins out at the expense of theology.

Gersonides' attitude toward previous astronomers, coupled with his faith in human reason, are reflected in his discussion of creation. Maimonides went to great lengths to maintain that the topic of creation is beyond rational demonstration. Gersonides, on the other hand, devotes many chapters in *Milhamot* VI to proving that the Platonic theory of creation out of an eternal formless matter is rationally demonstrable. The question of whether the universe was created or had existed from eternity had been treated by Maimonides in an ambiguous manner; scholars still disagree over whether Maimonides ultimately upheld an Aristotelian, Platonic, or scriptural doctrine of creation. Gersonides' position is unambiguously Platonic. Gersonides argues that the world was created outside of time by a freely willing agent. He must then decide whether the world was engendered *ex nihilo* or out of a preexistent matter. Arguing that *ex nihilo* creation is incompatible with physical reality, he adopts a model drawn from Plato's *Timaeus*. Gersonides interprets the opening of Genesis to refer to two types of matter. *Geshem* is the primordial matter out of which the universe was created; not capable of motion or rest, it was characterized by negation and was inert and chaotic. This matter is identified with the primeval waters described in Genesis. *Homer* is prime matter, in the Aristotelian sense of a substratum always aligned with form. It

contains within itself the potentiality to receive forms but is not an ontologically independent entity. Gersonides compares this matter to darkness: just as darkness is the absence of light, this matter represents the absence of form or shape. On this basis Gersonides argues that the world was created out of an eternally preexistent matter.

Gersonides' cosmology forms the backdrop of the other books of *Milhamot*. His predecessor Maimonides had claimed that no valid inference can be drawn from the nature of the sublunar sphere to that of the superlunar sphere. Gersonides, however, rejects the metaphysical bite to the distinction, and argues that inasmuch as both spheres contain material elements, what we know about creation is based on astronomy, and astronomy is fundamentally no different a human science than physics. Astronomy can only be pursued as a science by "one who is both a mathematician and a natural philosopher, for he can be aided by both of these sciences and take from them whatever is needed to perfect his work" (*Wars*, V.1.1, p. 23). Gersonides sees the ultimate function of astronomy to understand God. Astronomy, he tells us, is instructive not only by virtue of its exalted subject matter, but also because of its utility in the other sciences. By studying the orbs and stars, we are led ineluctably to a fuller knowledge and appreciation of God. Astronomy thus functions as the underpinning of the rest of the work.

Gersonides' discussion of immortality of the soul in book 1 must be understood against the backdrop of a notoriously difficult passage in Aristotle's *On the Soul*, book 3, chapter 5 (430a22-25). In this passage Aristotle seems to postulate the existence of an active intellect which is separable from the passive intellect and which is primarily responsible for the intellectual activities of the human soul. But what is the relation between the active and passive intellects, and which, if either, is immortal? Gersonides states and rejects three positions that elucidate a version of the unity of intellect. The import of Gersonides' critique of his predecessors can be reduced to three main issues. From a theological perspective, it is clear that the doctrine of unity of intellect threatens the notion of personal immortality. For if all humans share the same intellect, then upon physical death, all that remains of the person is an unindividualized intellect. Epistemologically, the doctrine is unable to account for how it is that two (or more) knowers can entertain contrary items of knowledge; or, more stringently, how one person can be mistaken about something another person knows. And from a metaphysical perspective, the main problem is how to individuate this separate intellect when it is manifested in many individuals: for if it is individuated materially on the basis that individual bodies differ, then the substance is no longer incorporeal or separate. As Feldman has pointed out, on this theory an incorporeal substance is either a unique member of a species or is not a member of a species at all (*Wars*, I.4, p. 79).

Gersonides avoids these untoward consequences by adopting the position of Alexander of Aphrodisias. Gersonides agrees with Alexander of Aphrodisias that immortality consists in the intellectual perfection of the material intellect. He disagrees with Alexander, however, over the precise nature of this intellectual attainment. For Alexander (according to Gersonides) had claimed that immortality is achieved when the intellect acquires knowledge of the Agent Intellect (hence the term "acquired intellect" is introduced). Immortality is thus understood by Alexander to be a form of conjunction between the Agent and acquired intellects. "They [the followers of Alexander] maintain that the material intellect is capable of immortality and subsistence when it reaches that level of perfection where the objects of knowledge that it apprehends are themselves intellects, in particular the Agent

Intellect…[material intellect] is immortal when it is united with the Agent Intellect" (*Wars*, I.8, p. 170).

Gersonides rejects this notion of conjunction, however, and replaces it with a model of immortality according to which it is the content of knowledge of the acquired intellect that matters. When the content of the acquired intellect mirrors the rational ordering of the Agent Intellect, immortality is achieved. What is the content of this knowledge? The Agent Intellect must possess complete knowledge of the sublunary world; that is, it "contains a conception of the rational order obtaining in all individuals" (*Wars*, I.4, p. 136). The anti-Platonic tenor of this position is emphasized when Gersonides describes in more detail what it is that the Agent Intellect knows. For according to Gersonides, the knowledge of the Agent Intellect must be grounded in the domain of particulars. Thus Gersonides' position avoids the epistemological difficulties apparent in a realist ontology. Inasmuch as the material intellect reflects the knowledge inherent in the Agent Intellect, and inasmuch as this knowledge is grounded in particulars, it follows that humans can have knowledge of particulars; in this acquisition of knowledge lies immortality.

Books 2 to 4 focus on the relation between God and the world. The general problem is whether God's knowledge is limited to necessary states of affairs or extends to the domain of contingency as well. If the former, then God could not be said to have knowledge of humans, and so divine providence would not be efficacious. But if God does know contingents -- in particular, future contingent events -- then it would appear that human freedom is curtailed by God's prior knowledge of human actions. The problem of the apparent conflict between divine omniscience and human freedom was discussed by many medieval philosophers. Gersonides does not follow the majority opinion on this issue: rather than claim that God does know particulars and that this knowledge somehow does not affect human freedom, Gersonides argues that God knows particulars only in a certain sense. In an apparent attempt to mediate between the view of Aristotle, who said that God does not know particulars, and that of Maimonides, who said that he does, Gersonides holds that God knows particulars only insofar as they are ordered. That is, God knows that certain states of affairs are particular, but he does not know in what their particularity consists. God knows individual persons, for example, only through knowing the species humanity.

Whereas Maimonides claimed that God's knowledge does not render the objects of his knowledge necessary, Gersonides maintains that divine knowledge precludes contingency. To retain the domain of contingency, he adopts the one option open to him: namely, that God does not have prior knowledge of future contingents. According to Gersonides, God knows that certain states of affairs may or may not be actualized. But insofar as they are contingent states, he does not know which of the alternatives will be the case. For if God did know future contingents prior to their actualization, there could be no contingency in the world.

In book 2, in an attempt to explain how prophecies are possible in a system which denies the possibility of knowledge of future contingents, Gersonides claims that the prophet does not receive knowledge of particular future events; rather his knowledge is of a general form, and he must instantiate this knowledge with particular facts. What distinguishes prophets from ordinary persons is that the former are more attuned to receive these universal messages and are in a position to apply them to particular circumstances.

A further dilemma surrounds the doctrine of divine providence. If God does not have knowledge of future contingents, how can he be said to bestow providence on his creatures? This problem is discussed by Gersonides both in his commentary on Job and in book 4 of *Milhamot*. In both texts he argues that providence is general in nature; it primarily appertains to species and only incidentally to particulars of the species. God, for example, does not know the particular individual Levi ben Gerson and does not bestow particular providence on him. Rather, inasmuch as Levi ben Gerson is a member of the species humanity and the species philosopher, he is in a position to receive the providential care accorded to those groups.

For Gersonides, the issues of prophecy, omniscience and providence are developed against the backdrop of astrological determinism. Like many thinkers of the late Middle Ages, Gersonides had to confront two opposing sets of traditions: on the one hand, attacks by religious authorities (e.g. Augustine's attack in *City of God*; Maimonides' letters) on the grounds that astrology compromised human free will; on the other hand, the wide scale acceptance of astrology from the 12th century on. In the 12th and 13th centuries, most Jewish and Christian philosophers supported natural astrology, the view that the celestial bodies affect sublunar life and existence to some extent at least. That the sun and moon both affect natural cycles and events on earth is unequivocal and represents a classic paradigm of natural astrology. The calculations of natural astrology overlapped those of astronomy, and could be utilized for practical purposes such as fixing the calendar. According to astrologers, each planet and sign of the zodiac has its own character, power and attributes. Inasmuch as the characters of the planets and the signs of the zodiac are opposed to each other, they are engaged in a perpetual power struggle. Thus the position of the planets and their interrelation with the signs of the zodiac, regulate the fate of both individuals and nations. Astrological predictions could apply, then to both individuals as well as to the history of Israel and its place in universal history.

An attack upon astrology as a whole belonged to a much larger conflict, that between the roles of reason and faith. Thus, astrology should not be situated within the context of magic or the occult, but rather should be construed as a robust contender to science. Based on a precise scientific astronomy, astrology was a science accepted from the second to the seventeenth centuries. On the scientific level it prevailed almost uncontested until and including Newton.

5. Gersonides' Astrological Determinism

Even a summary reading of Gersonides' major philosophical work evinces an explicit "belief in" astrology. Gersonides develops his astral determinism in two contexts: in book II of *Wars* he interweaves astrological motifs into his discussion of divine providence and prophecy, while in Book V astrology occupies a central role in the context of his cosmological speculations. His major concern is the extent to which the stars and planets exerted an influence over human events in general, or more particularly, over those actions that entail human choice. Judicial astrology was based on the assumption that the entire world of nature was governed and directed by the movement of the heavens and the celestial bodies, and that man, as an animal naturally generated and living in the world of nature, was also naturally under

their rule.

Langermann emphasizes the teleological nature of astrology for Gersonides, its chief merit being its ability to provide "teleological explanations for the wide variety of stellar motions that are observed to take place" (*Wars*, Vol III, p. 506). This teleology is reflected in V.2 ch 7-9 where, after listing 27 problems raised by the heavenly bodies, Gersonides suggests that only astrological considerations can furnish satisfactory replies; it is astrology alone that can explain the connection between the two realms. It is worth noting that on this point, Gersonides disagrees with Maimonides over the ultimate purpose of the celestial bodies. For Maimonides it is not possible that a greater entity, the heavens, would exist for the sake of the sublunar universe. Gersonides disagrees, maintaining that it is not inappropriate that the more noble exist for the less noble. The stars, he argues, exist for the sake of things in the sublunar world (*Wars*, V.2.3, p. 194). More explicitly, the heavenly bodies are designed for the benefit of sublunar existence, and they guarantee the perpetuation of life on earth.

This teleology is spelled out in *Milhamot* II, in which Gersonides is concerned to explain how divine knowledge operates, and to what extent divine foreknowledge of future contingents affects human choice. His major thesis is that divine knowledge is predicated to a great extent upon knowledge of the heavenly bodies, which bodies are in turn "systematically directed toward his [man's] preservation and guidance so that all his activities and thoughts are ordered by them" (*Wars*, II.2, p. 33). In support of this teleological cosmology, Gersonides presents an extensive argument to the effect that the celestial bodies have a purpose. On the basis of this argument Gersonides concludes that from the perspective of the teleological structure of the universe, we can understand why the heavenly bodies behave the way they do. This teleology is reflected by a "law, order and rightness" in the universe, implying the existence of an intellect that orders the nature of things: "you see that the domain of the spheres provides, in the best way possible, for the sub-lunar world" (*Wars*, V.2.5, p. 137).

As we have seen, the existence of a connection between celestial and terrestrial events was admitted by most everybody, but not everybody agreed on the nature of this connection. Gersonides as well must account for the type of relation obtaining between celestial and terrestrial events. Having articulated the ordering power of the astral bodies, Gersonides describes in *Milhamot* V.3 the separate intellects and the spheres that they move. The main characteristic of the astral bodies is their luminosity (*nitzutz*). This luminosity affects their actions and effects (*Wars*, V.2.3, p. 137). Gersonides is very much aware of the problem of accounting for how the astral bodies can affect actions at a distance. The sun, for example, functions as a paradigm for action at a distance. Once we understand, Gersonides claims, how the activity of heating reaches earth from the sun, we can understand how the particular activities of the other stars reach the sublunar realm as well. By explaining the efficient cause as the light or radiation of the stars, Gersonides can account for weak or strong effects. As Langerman has pointed out, Gersonides' account furnishes the basis for the introduction of astrological causation into natural philosophy.

In *Milhamot* V.2.8 Gersonides lays out six astrological principles that affect his general cosmological scheme. These can be summarized as follows. First, each astral body exercises a different influence specific to it. Second, astral influence depends upon its position in the zodiac (*galgal hamazalot*). Third, the longer a star stays in one place in the zodiac, the greater its effect because of the strength of its

luminosity. Fourth, astral influence is dependent upon its inclination to the north or to the south; its effect will be strongest when it is in the middle, as evidenced by the sun, whose heat is strongest when it is at the Tropic of Cancer as opposed to being at the Tropic of Capricorn. Fifth, the greater the radiation or luminosity of a star, the stronger its influence. And finally, the closer to earth a star is, the stronger will be its influence (*Wars*, V.2.8, p. 207-8). These principles function as the underpinnings of his general astronomy as well.

In light of the original problems posed by astrology above, let me propose that the most important piece of Levi's astrology is what Langermann calls the variety of the heavens (*ribbui hayahasim*). Gersonides must be able to account for individual variety in the sublunar realm. Inasmuch as stellar radiation is the means by which stellar influences are conveyed, the wide variety of mixtures of stellar radiation guarantees a sufficient variety of "influences" on terrestrial processes. The movers emanate from God who is construed as the "First Separate Intellect" (*Wars*, V.3.8, p. 272). They are ordered in a rational system that governs the sublunar domain. If there were no one first intellect, Gersonides argues, the rational order we see in the heavens would be the result of chance, which is unacceptable. The agent intellect thus functions as the link between these celestial bodies and human affairs. The kinds of information it transmits are of an astronomical type, as evidenced in the following example: "it [the agent intellect] knows how many revolutions of the sun, or of the diurnal sphere, or of any other sphere [have transpired] from the time at which someone, who falls under a particular pattern, had a particular level of good or ill fortune…" (*Wars*, II.6, p. 64). The agent intellect serves as the repository of information communicated by the heavenly bodies. The patterns revealed in this communication between agent intellect and diviner (astrologer, prophet) are from the heavenly bodies which themselves are endowed with intellects and so "apprehend the pattern that derives from them." Each mover apprehends the order deriving from the heavenly body it moves, and not patterns that emanate from other heavenly bodies. As a result, the imaginative faculty receives the "pattern inherent in the intellects of the heavenly bodies from the influence deriving from them. This influence derives from the position of the heavenly bodies "by the ascendent degree or the dominant planet [in a particular zodiacal position]" (*Wars*, II.6, p. 64). However, inasmuch as the heavenly bodies do not jointly cooperate with one another (*lo yishtatfu*) in this process, it is possible for the communication to be misconstrued.

Of course, as we all know, astrologers often err in their predictions. Astrological errors can be due to several factors. In general, Gersonides claims, we know very little of the order of the heavenly bodies. "In general, it is impossible for man to know the [complete] truth of the order of the sublunar world. This is nicely illustrated in astrology, where frequently false predictions are made. All the more so is it impossible for man to know the general order of the sublunar world by means of its causes so that his knowledge would be perfect" (*Wars*, I.12, p. 219). In some cases, the information is not transmitted clearly. Why is it that certain communications are received more clearly than others? A constitutionally perfect imaginative faculty receives information from both dominant and weak heavenly bodies. By 'weak', Gersonides means that certain celestial bodies are too weak both to bring about events on earth as well as to transmit information about these events. Hence he concludes that information about the future emanates "from the dominant body in the particular proper face (*panim*) in which it has dominance but not from any of the attending planets (*ha-meshartim*)" (*Wars*, II.7, pp. 69-70). But to constitutionally imperfect imaginative faculties, the information received is only from the dominant heavenly bodies.

Hence the overall quality of the information received will differ in the two cases. More specifically, because of the difficulty of obtaining the necessary positions of these bodies by observation, astrologers are often unable to verify their data. Furthermore, since the zodiacal position of a heavenly body at any given time is only repeated once in many thousands of years, astrologers have no access to the repeatability of those events that would be required to verify their knowledge. Furthermore, humans simply do not have sufficient knowledge about the heavenly bodies.

The final cause for error has to do with human free will: as we have seen above, our intellect and choice "have the power to move us contrary to that which is determined by the heavenly bodies" (*Wars*, II.2, p. 34). Although he admits that on occasion human choice is able to contravene the celestial bodies, nevertheless this intervention is rare, and true contingency is a rare state of affairs indeed in Gersonides' ontology. Gersonides presents an argument to show that human choice guided by reason can subvert the celestial bodies despite their general ordering of our lives. The heavenly bodies can order human affairs either by virtue of their difference of position in the heavens, or from the difference of the bodies among themselves. Astral bodies, however, will affect different individuals in different ways; they can also affect an individual differently at different times; and finally, two or more bodies can affect a single individual, resulting in multiple influences that can have contrary effects, echoing the scholastic phrase, "*sapiens dominabitur astris* [the wise man will be ruled by the stars]". Gersonides notes that humans can contravene these effects: God has provided humans with "the intellectual capacity (*sekhel ba'al takhlit*) that enables us both to act contrary to what has been ordered by the heavenly bodies and to correct, as far as possible, the [astrally ordained] misfortunes that befall us" (*Wars*, II.2, p. 35). Nevertheless, he assures us that whatever happens by chance is "determined and ordered according to this type of determinateness and order" (*Wars*, II.2, p. 34). Outdoing even Plato's hierarchical structuring in *Republic* IV, Gersonides argues that the ultimate perfection and ordering of society is due to astrological influence.

The commensurability of the motion of heavenly bodies raises an additional concern, having to do with the uniqueness of individual beings and the doctrine of eternal return. Gersonides' immediate 13th century predecessors Shem-Tov ibn Falaquera and Judah ben Solomon ha-Cohen discussed this issue against the backdrop of Aristotle's *Gen. Animalia*. In *Gen. Animalia,* Aristotle had established a connection between the life spans and gestations periods of animals and the revolutions of the sun and moon (*Gen. Anim.,* IV.10, 777b17-778a10). Thus the revolutions of the sun measure not only time but also produce the alternating periods of growth and decay. Eschatological predictions are thus tied to the cyclicity of the heavenly bodies. Gersonides did not [to our knowledge] indulge in eschatological and millennial predictions. In fact, Gersonides wrote only one astrological text that has survived, a prognostication based on the conjunction of Saturn and Jupiter to take place in March 1345. Gersonides himself died in 1344, a year before the event in question. As Goldstein has demonstrated, this conjunction was predicted already by Ibn Ezra, and repeated by Abraham Bar Hiyya in his *Megillat ha-Megalleh* where the conjunction was associated with a date of messianic significance that would supposedly take effect in 1358 (Goldstein, 1990, p. 3). The conjunction was codified by Levi ben Abraham ben Hayyim in his encyclopedia *Livyat Hen*, indicating an awareness in the Jewish community of the messianic significance of this conjunction. According to North, Ibn Ezra was the earliest scholar to record one of the seven methods for the setting up of the astrological houses; Gersonides then computed the astrological houses for the prognostication of 1345 according to Ibn Ezra's method. (See North, 1986,

p. 25.) Goldstein suggests that as the date 1345 approached, the Papal court might have become interested as well in the conjunction. We do know that Gersonides' text was translated into Latin with the aid of his brother shortly after Gersonides' death in 1344.

In his prognostication, Gersonides predicts that there will be "extraordinary evil with many wars, visions and miraculous signs;" "Diseases and death, and the evil will endure for a long time;" "the absence of good, pleasure and happiness for most of the inhabited world;" "the spilling of much blood and increasing enmity, jealousy, strife, famine, various diseases, drought and dearth" (Goldstein, 1990). The Black Death, which arrived in Europe in 1347, was thus provided with numerous astrological credentials. The official statement of the medical faculty of the University of Paris, presented to the king in 1348, reported on the conjunction of Saturn and Jupiter in the house of Aquarius on 20 March 1345, which was seen to spread "death and disaster". It is not hard to see how the conjunction of 1345 came to be associated with the Black Death.

6. Conclusion

Gersonides' philosophical ideas went against the grain of traditional Jewish thought. Gersonides reflects the following characteristics: first, his writings demonstrate a fundamental interplay and harmony between astrological and theological beliefs. It is clear that the appeal of astrology lay in the fact that it offered useful information, while it looked and operated like a science. Even the critics of astrology had to agree that the heavens exerted a real influence upon terrestrial events. The complexity of the rules of astrology and internal disagreement among its followers served to increase the respect accorded to the science. Failures did not cause the astrologer to lose faith, just as failures among modern physicists do not lead to loss of faith in science. Gersonides believed that life on earth had a meaning, and that terrestrial events had an order. Astrology was a means of ascertaining that meaning. Gersonides' views on prophecy, providence, free-will and evil reflected ingredients of this philosophical determinism. Whereas his commentaries occupied a central place in Jewish theology, his philosophical work was rejected. Jewish philosophers such as Hasdai Crescas and Isaac Abrabanel felt obliged to subject his works to lengthy criticism. Only in recent years has Gersonides received his rightful place in the history of philosophy. As scholars have rediscovered his thought and have made his corpus available to a modern audience, Gersonides is once again appreciated as an insightful, ruthlessly consistent philosopher.

Bibliography

Principal Works

- *Sefer Ha-heqesh Ha-yashar* (On Valid Syllogisms, written 1319); translated into Latin as *Liber Syllogismi Recti*.
- *Sefer Ma'aseh Hoshev* (*The Work of a Counter*, written 1321; edited and translated into German by Gerson Lange (Frankfurt am Main: Golde, 1909).

- *Perush 'al Sefer lyob* (*Commentary on Job*, written 1325; Ferrara, 1477).
- *Sefer Milhamot Ha-Shem* (*The Wars of the Lord*, written 1329; Riva di Trento, 1560; Leipzig, 1866; Berlin, 1923).
- *Perush 'al Sefer Ha-Torah* (*Commentary on the Pentateuch*, written 1329-1338; Venice, 1547; Jerusalem, 1967).

Editions in English

- *The Commentary of Levi ben Gerson on the Book of Job*, translated by Abraham L. Lassen. New York: Bloch, 1946.
- *Providence and the Philosophy of Gersonides*, translated by David Bleich. New York: Yeshiva University Press, 1973.
- *Gersonides' The Wars of the Lord. Treatise Three: On God's Knowledge*, translated by Norbert M. Samuelson. Toronto: Pontifical Institute of Mediaeval Studies, 1977.
- *Creation of the World According to Gersonides*, translated by Jacob Staub. Chico, Cal.: Scholars Press, 1982.
- *The Wars of the Lord*. Translated by Seymour Feldman. 3 vols. Philadelphia: Jewish Publication Society, 1984-1999.
- *Commentary on Song of Songs*. Trans and ed. by Menachem Kellner, New Haven: Yale University Press, 1998.

Selected Secondary Literature

- Carlebach, Salomon. *Levi Ben Gerson Als Mathematiker*. Berlin, 1910.
- Dahan, Gilbert, ed. *Gersonide En Son Temps*. Louvain-Paris: E. Peeters, 1991.
- Eisen, Robert. *Gersonides on Providence, Covenant, and the Jewish People*. Albany: State University of New York Press, 1995.
- Feldman, Seymour. "Gersonides' Proofs for the Creation of the Universe." *Proceedings of the American Academy for Jewish Research* (1967): 113-37.
- -----. "A Debate Concerning Determinism in Late Medieval Jewish Philosophy." *Proceedings of the American Academy for Jewish Research* 51 (1984): 15-54.
- -----. "Platonic Themes in Gersonides' Doctrine of the Active Intellect." In *Neoplatonism and Jewish Thought*, edited by Lenn Goodman, 255-77. Albany: State University of New York Press, 1992.
- Freudenthal, Gad. "Épistemologie, astronomie et astrologie chez Gersonide." *Revue des études juives* 146, no. 3-4 (1987): 357-65.
- -----. "Les Sciences dans les communautés juives médiévales de Provence: Leur Appropriation, Leur Rôle." *Revue des études juives* CLLII (1993): 29-136.
- Freudenthal, Gad. Ed. *Studies on Gersonides: A Fourteenth-Century Jewish Philosopher-Scientist*. Leiden: Brill, 1992.
- Goldstein, Bernard R. "The Astronomical Tables of Rabbi Levi Ben Gerson." In *Transactions of the Connecticut Academy of Arts and Sciences*. Hamden, CT: Shoestring Press, 1974.

- -----. "The Status of Models in Ancient and Medieval Astronomy." *Centaurus* 24 (1980): 132-47.
- -----. *The Astronomy of Levi Ben Gerson 1288-1344: A Critical Edition of Chapters 1-20*. New York: Springer Verlag, 1985.
- -----. "Levi Ben Gerson's Astrology in Historical Perspective." In *Gersonide En Son Temps*, edited by Gilbert Dahan, 287-300. Louvain-Paris: E. Peeters, 1991.
- -----. Levi Ben Gerson's Contributions to Astronomy in *Studies on Gersonides*. Edited by Gad Freudenthal. Leiden: E.J. Brill, 1992.
- -----. "Astronomy and Astrology in the Works of Abraham Ibn Ezra." *Arabic Sciences and Philosophy* 6 (1996): 9-21.
- Goldstein, Bernard R., and David Pingree. "Levi Ben Gerson's Prognostication for the Conjunction of 1345." *Transactions of the American Philosophical Society* 80 (1990): 1-60.
- Husik, Isaac, "Studies in Gersonides," *Jewish Quarterly Review*, new series 7 (1916-1917): 553-594; 8 (1917-1918): 113-156, 231-268.
- Kellner, Menachem. "R. Levi Ben Gerson: A Bibliographical Essay." *Studies in Bibliography and Booklore* 12 (1979): 13-23.
- Klein-Braslavy, Sara. "Determinism, Possibility, Choice and Foreknowledge in Ralbag." *Da' at* 22 (1989): 4-53.
- -----. "Gersonides on the Magnet and the Heat of the Sun." In *Studies on Gersonides* Ed. Gad Freudenthal., 267-84. Leiden: E.J. Brill, 1992.
- -----. "Gersonides and Astrology." In *Levi Ben Gershom: The Wars of the Lord*, edited by Seymour Feldman, 506-19. New York: JPS, 1999.
- Manekin, Charles H. *The Logic of Gersonides: An Analysis of Selected Doctrines*. Dordrecht: Kluwer Academic, 1992.
- -----. "On the Limited-Omniscience Interpretation of Gersonides' Theory of Divine Knowledge." In *Perspectives on Jewish Thought and Mysticism*, edited by Elliot R. Wolfson, Alfred L. Ivry, Allan Arkush, 135-70. Amsterdam: Harwood Academic, 1998.
- North, J.D. *Stars, Minds and Fate: Essays in Ancient and Medieval Cosmology*. London: The Hambledon Press, 1989.
- Rudavsky, T.M. ""Divine Omniscience and Future Contingents in Gersonides"." *Journal of the History of Philosophy* 21 (1983): 513-36.
- -----. "Divine Omniscience, Contingency and Prophecy in Gersonides." In *Divine Omniscience and Omnipotence in Medieval Philosophy*, edited by T. M. Rudavsky, 161-81. Dordecht: D. Reidel, 1984.
- -----. "Creation, Time and Infinity in Gersonides." *Journal of the History of Philosophy* 26, no. 1 (1988): 25-44.
- -----. "The Jewish Tradition: Maimonides, Gersonides and Bedersi." In *Individuation in Late Scholasticism and the Counter-Reformation*, edited by J. J. Gracia, 69-96. Albany: State University of New York Press, 1993.
- -----. *Time Matters: Time, Creation and Cosmology in Medieval Jewish Philosophy*. Albany: State University of New York Press, 2000.
- Samuelson, Norbert. "Gersonides' Account of God's Knowledge of Particulars." *Journal of the History of Philosophy* 10 (1972): 399-416.
- Touati, Charles. *La Pensée Philosophique et Théologique de Gersonide*. Paris, 1973.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Judaic Philosophy | Maimonides [Moses ben Maimon]

Tamar Rudavsky
rudavsky.1@osu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 17, 2001
Content last modified: August 17, 2001

Computer Ethics: Basic Concepts and Historical Overview

Computer ethics is a new branch of ethics that is growing and changing rapidly as computer technology also grows and develops. The term "computer ethics" is open to interpretations both broad and narrow. On the one hand, for example, computer ethics might be understood very narrowly as the efforts of professional philosophers to apply traditional ethical theories like utilitarianism, Kantianism, or virtue ethics to issues regarding the use of computer technology. On the other hand, it is possible to construe computer ethics in a very broad way to include, as well, standards of professional practice, codes of conduct, aspects of computer law, public policy, corporate ethics--even certain topics in the sociology and psychology of computing.

In the industrialized nations of the world, the "information revolution" already has significantly altered many aspects of life -- in banking and commerce, work and employment, medical care, national defense, transportation and entertainment. Consequently, information technology has begun to affect (in both good and bad ways) community life, family life, human relationships, education, freedom, democracy, and so on (to name a few examples). Computer ethics in the broadest sense can be understood as that branch of applied ethics which studies and analyzes such social and ethical impacts of information technology.

In recent years, this robust new field has led to new university courses, conferences, workshops, professional organizations, curriculum materials, books, articles, journals, and research centers. And in the age of the world-wide-web, computer ethics is quickly being transformed into "global information ethics".

---

1. Some Historical Milestones

1940s and 1950s

Computer ethics as a field of study has its roots in the work of MIT professor Norbert Wiener during World War II (early 1940s), in which he helped to develop an antiaircraft cannon capable of shooting down fast warplanes. The engineering challenge of this project caused Wiener and some colleagues to create a new field of research that Wiener called "cybernetics" -- the science of information feedback systems. The concepts of cybernetics, when combined with digital computers under development at that time, led Wiener to draw some remarkably insightful ethical conclusions about the technology that we now call ICT (information and communication technology). He perceptively foresaw revolutionary social and ethical consequences. In 1948, for example, in his book *Cybernetics: or control and communication in the animal and the machine*, he said the following:

> It has long been clear to me that the modern ultra-rapid computing machine was in principle an ideal central nervous system to an apparatus for automatic control; and that its input and output need not be in the form of numbers or diagrams. It might very well be, respectively, the readings of artificial sense organs, such as photoelectric cells or thermometers, and the performance of motors or solenoids ... . we are already in a position to construct artificial machines of almost any degree of elaborateness of performance. Long before Nagasaki and the public awareness of the atomic bomb, it had occurred to me that we were here in the presence of another social potentiality of unheard-of importance for good and for evil. (pp. 27-28)

In 1950 Wiener published his monumental book, *The Human Use of Human Beings*. Although Wiener did not use the term "computer ethics" (which came into common use more than two decades later), he laid down a comprehensive foundation which remains today a powerful basis for computer ethics research and analysis.

Wiener's book included (1) an account of the purpose of a human life, (2) four principles of justice, (3) a powerful method for doing applied ethics, (4) discussions of the fundamental questions of computer ethics, and (5) examples of key computer ethics topics. [Wiener 1950/1954, see also Bynum 1999]

Wiener's foundation of computer ethics was far ahead of its time, and it was virtually ignored for

decades. On his view, the integration of computer technology into society will eventually constitute the remaking of society -- the "second industrial revolution". It will require a multi-faceted process taking decades of effort, and it will radically change everything. A project so vast will necessarily include a wide diversity of tasks and challenges. Workers must adjust to radical changes in the work place; governments must establish new laws and regulations; industry and businesses must create new policies and practices; professional organizations must develop new codes of conduct for their members; sociologists and psychologists must study and understand new social and psychological phenomena; and philosophers must rethink and redefine old social and ethical concepts.

1960s

In the mid 1960s, Donn Parker of SRI International in Menlo Park, California began to examine unethical and illegal uses of computers by computer professionals. "It seemed," Parker said, "that when people entered the computer center they left their ethics at the door." [See Fodor and Bynum, 1992] He collected examples of computer crime and other unethical computerized activities. He published "Rules of Ethics in Information Processing" in *Communications of the ACM* in 1968, and headed the development of the first Code of Professional Conduct for the Association for Computing Machinery (eventually adopted by the ACM in 1973). Over the next two decades, Parker went on to produce books, articles, speeches and workshops that re-launched the field of computer ethics, giving it momentum and importance that continue to grow today. Although Parker's work was not informed by a general theoretical framework, it is the next important milestone in the history of computer ethics after Wiener. [See Parker, 1968; Parker, 1979; and Parker et al., 1990.]

1970s

During the late 1960s, Joseph Weizenbaum, a computer scientist at MIT in Boston, created a computer program that he called ELIZA. In his first experiment with ELIZA, he scripted it to provide a crude imitation of "a Rogerian psychotherapist engaged in an initial interview with a patient". Weizenbaum was shocked at the reactions people had to his simple computer program: some practicing psychiatrists saw it as evidence that computers would soon be performing automated psychotherapy. Even computer scholars at MIT became emotionally involved with the computer, sharing their intimate thoughts with it. Weizenbaum was extremely concerned that an "information processing model" of human beings was reinforcing an already growing tendency among scientists, and even the general public, to see humans as mere machines. Weizenbaum's book, *Computer Power and Human Reason* [Weizenbaum, 1976], forcefully expresses many of these ideas. Weizenbaum's book, plus the courses he offered at MIT and the many speeches he gave around the country in the 1970s, inspired many thinkers and projects in computer ethics.

In the mid 1970s, Walter Maner (then of Old Dominion University in Virginia; now at Bowling Green State University in Ohio) began to use the term "computer ethics" to refer to that field of inquiry dealing with ethical problems aggravated, transformed or created by computer technology. Maner offered an experimental course on the subject at Old Dominion University. During the late 1970s (and indeed into

the mid 1980s), Maner generated much interest in university-level computer ethics courses. He offered a variety of workshops and lectures at computer science conferences and philosophy conferences across America. In 1978 he also self-published and disseminated his *Starter Kit in Computer Ethics*, which contained curriculum materials and pedagogical advice for university teachers to develop computer ethics courses. The *Starter Kit* included suggested course descriptions for university catalogs, a rationale for offering such a course in the university curriculum, a list of course objectives, some teaching tips and discussions of topics like privacy and confidentiality, computer crime, computer decisions, technological dependence and professional codes of ethics. Maner's trailblazing course, plus his *Starter Kit* and the many conference workshops he conducted, had a significant impact upon the teaching of computer ethics across America. Many university courses were put in place because of him, and several important scholars were attracted into the field.

1980s

By the 1980s, a number of social and ethical consequences of information technology were becoming public issues in America and Europe: issues like computer-enabled crime, disasters caused by computer failures, invasions of privacy via computer databases, and major law suits regarding software ownership. Because of the work of Parker, Weizenbaum, Maner and others, the foundation had been laid for computer ethics as an academic discipline. (Unhappily, Wiener's ground-breaking achievements were essentially ignored.) The time was right, therefore, for an explosion of activities in computer ethics.

In the mid-80s, James Moor of Dartmouth College published his influential article "What Is Computer Ethics?" (see discussion below) in *Computers and Ethics*, a special issue of the journal *Metaphilosophy* [Moor, 1985]. In addition, Deborah Johnson of Rensselaer Polytechnic Institute published *Computer Ethics* [Johnson, 1985], the first textbook -- and for more than a decade, the defining textbook -- in the field. There were also relevant books published in psychology and sociology: for example, Sherry Turkle of MIT wrote *The Second Self* [Turkle, 1984], a book on the impact of computing on the human psyche; and Judith Perrolle produced *Computers and Social Change: Information, Property and Power* [Perrolle, 1987], a sociological approach to computing and human values.

In the early 80s, the present author (Terrell Ward Bynum) assisted Maner in publishing his *Starter Kit in Computer Ethics* [Maner, 1980] at a time when most philosophers and computer scientists considered the field to be unimportant [See Maner, 1996]. Bynum furthered Maner's mission of developing courses and organizing workshops, and in 1985, edited a special issue of *Metaphilosophy* devoted to computer ethics [Bynum, 1985]. In 1991 Bynum and Maner convened the first international multidisciplinary conference on computer ethics, which was seen by many as a major milestone of the field. It brought together, for the first time, philosophers, computer professionals, sociologists, psychologists, lawyers, business leaders, news reporters and government officials. It generated a set of monographs, video programs and curriculum materials [see van Speybroeck, July 1994].

1990s

During the 1990s, new university courses, research centers, conferences, journals, articles and textbooks appeared, and a wide diversity of additional scholars and topics became involved. For example, thinkers like Donald Gotterbarn, Keith Miller, Simon Rogerson, and Dianne Martin -- as well as organizations like Computer Professionals for Social Responsibility, the Electronic Frontier Foundation, ACM-SIGCAS -- spearheaded projects relevant to computing and professional responsibility. Developments in Europe and Australia were especially noteworthy, including new research centers in England, Poland, Holland, and Italy; the ETHICOMP series of conferences led by Simon Rogerson and the present author; the CEPE conferences founded by Jeroen van den Hoven; and the Australian Institute of Computer Ethics headed by Chris Simpson and John Weckert.

These important developments were significantly aided by the pioneering work of Simon Rogerson of De Montfort University (UK), who established the Centre for Computing and Social Responsibility there. In Rogerson's view, there was need in the mid-1990s for a "second generation" of computer ethics developments:

> The mid-1990s has heralded the beginning of a second generation of Computer Ethics. The time has come to build upon and elaborate the conceptual foundation whilst, in parallel, developing the frameworks within which practical action can occur, thus reducing the probability of unforeseen effects of information technology application [Rogerson, Spring 1996, 2; Rogerson and Bynum, 1997].

2. Defining the Field of Computer Ethics

From the 1940s through the 1960s, therefore, there was no discipline known as "computer ethics" (notwithstanding the work of Wiener and Parker). However, beginning with Walter Maner in the 1970s, active thinkers in computer ethics began trying to delineate and define computer ethics as a field of study. Let us briefly consider five such attempts:

When he decided to use the term "computer ethics" in the mid-70s, Walter Maner defined the field as one which examines "ethical problems aggravated, transformed or created by computer technology". Some old ethical problems, he said, are made worse by computers, while others are wholly new because of information technology. By analogy with the more developed field of medical ethics, Maner focused attention upon applications of traditional ethical theories used by philosophers doing "applied ethics" -- especially analyses using the utilitarian ethics of the English philosophers Jeremy Bentham and John Stuart Mill, or the rationalist ethics of the German philosopher Immanual Kant.

In her book, *Computer Ethics*, Deborah Johnson [1985] defined the field as one which studies the way in which computers "pose new versions of standard moral problems and moral dilemmas, exacerbating the old problems, and forcing us to apply ordinary moral norms in uncharted realms," [Johnson, page 1]. Like Maner before her, Johnson recommended the "applied ethics" approach of using procedures and concepts from utilitarianism and Kantianism. But, unlike Maner, she did not believe that computers create wholly new moral problems. Rather, she thought that computers gave a "new twist" to old ethical

issues which were already well known.

James Moor's definition of computer ethics in his article "What Is Computer Ethics?" [Moor, 1985] was much broader and more wide-ranging than that of Maner or Johnson. It is independent of any specific philosopher's theory; and it is compatible with a wide variety of methodological approaches to ethical problem-solving. Over the past decade, Moor's definition has been the most influential one. He defined computer ethics as a field concerned with "policy vacuums" and "conceptual muddles" regarding the social and ethical use of information technology:

> A typical problem in computer ethics arises because there is a policy vacuum about how computer technology should be used. Computers provide us with new capabilities and these in turn give us new choices for action. Often, either no policies for conduct in these situations exist or existing policies seem inadequate. A central task of computer ethics is to determine what we should do in such cases, that is, formulate policies to guide our actions.... One difficulty is that along with a policy vacuum there is often a conceptual vacuum. Although a problem in computer ethics may seem clear initially, a little reflection reveals a conceptual muddle. What is needed in such cases is an analysis that provides a coherent conceptual framework within which to formulate a policy for action [Moor, 1985, 266].

Moor said that computer technology is genuinely revolutionary because it is "logically malleable":

> Computers are logically malleable in that they can be shaped and molded to do any activity that can be characterized in terms of inputs, outputs and connecting logical operations....Because logic applies everywhere, the potential applications of computer technology appear limitless. The computer is the nearest thing we have to a universal tool. Indeed, the limits of computers are largely the limits of our own creativity [Moor, 1985, 269]

According to Moor, the computer revolution is occurring in two stages. The first stage was that of "technological introduction" in which computer technology was developed and refined. This already occurred in America during the first forty years after the Second World War. The second stage -- one that the industrialized world has only recently entered -- is that of "technological permeation" in which technology gets integrated into everyday human activities and into social institutions, changing the very meaning of fundamental concepts, such as "money", "education", "work", and "fair elections".

Moor's way of defining the field of computer ethics is very powerful and suggestive. It is broad enough to be compatible with a wide range of philosophical theories and methodologies, and it is rooted in a perceptive understanding of how technological revolutions proceed. Currently it is the best available definition of the field.

Nevertheless, there is yet another way of understanding computer ethics that is also very helpful--and

compatible with a wide variety of theories and approaches. This "other way" was the approach taken by Wiener in 1950 in his book *The Human Use of Human Beings*, and Moor also discussed it briefly in "What Is Computer Ethics?" [1985]. According to this alternative account, computer ethics identifies and analyzes the impacts of information technology upon human values like health, wealth, opportunity, freedom, democracy, knowledge, privacy, security, self-fulfillment, and so on. This very broad view of computer ethics embraces applied ethics, sociology of computing, technology assessment, computer law, and related fields; and it employs concepts, theories and methodologies from these and other relevant disciplines [Bynum, 1993]. The fruitfulness of this way of understanding computer ethics is reflected in the fact that it has served as the organizing theme of major conferences like the National Conference on Computing and Values (1991), and it is the basis of recent developments such as Brey's "disclosive computer ethics" methodology [Brey 2000] and the emerging research field of "value-sensitive computer design". (See, for example, [Friedman, 1997], [Friedman and Nissenbaum, 1996], [Introna and Nissenbaum, 2000].)

In the 1990s, Donald Gotterbarn became a strong advocate for a different approach to defining the field of computer ethics. In Gotterbarn's view, computer ethics should be viewed as a branch of professional ethics, which is concerned primarily with standards of practice and codes of conduct of computing professionals:

> There is little attention paid to the domain of professional ethics -- the values that guide the day-to-day activities of computing professionals in their role as professionals. By computing professional I mean anyone involved in the design and development of computer artifacts... The ethical decisions made during the development of these artifacts have a direct relationship to many of the issues discussed under the broader concept of computer ethics [Gotterbarn, 1991].

With this professional-ethics definition of computer ethics in mind, Gotterbarn has been involved in a number of related activities, such as co-authoring the third version of the ACM Code of Ethics and Professional Conduct and working to establish licensing standards for software engineers [Gotterbarn, 1992; Anderson, et al., 1993; Gotterbarn, et al., 1997].

3. Example Topics in Computer Ethics

No matter which re-definition of computer ethics one chooses, the best way to understand the nature of the field is through some representative examples of the issues and problems that have attracted research and scholarship. Consider, for example, the following topics:

(See also the wide range of topics included in the recent anthology [Spinello and Tavani, 2001].)

3.1 Computers in the Workplace

As a "universal tool" that can, in principle, perform almost any task, computers obviously pose a threat to jobs. Although they occasionally need repair, computers don't require sleep, they don't get tired, they don't go home ill or take time off for rest and relaxation. At the same time, computers are often far more efficient than humans in performing many tasks. Therefore, economic incentives to replace humans with computerized devices are very high. Indeed, in the industrialized world many workers already have been replaced by computerized devices -- bank tellers, auto workers, telephone operators, typists, graphic artists, security guards, assembly-line workers, and on and on. In addition, even professionals like medical doctors, lawyers, teachers, accountants and psychologists are finding that computers can perform many of their traditional professional duties quite effectively.

The employment outlook, however, is not all bad. Consider, for example, the fact that the computer industry already has generated a wide variety of new jobs: hardware engineers, software engineers, systems analysts, webmasters, information technology teachers, computer sales clerks, and so on. Thus it appears that, in the short run, computer-generated unemployment will be an important social problem; but in the long run, information technology will create many more jobs than it eliminates.

Even when a job is not eliminated by computers, it can be radically altered. For example, airline pilots still sit at the controls of commercial airplanes; but during much of a flight the pilot simply watches as a computer flies the plane. Similarly, those who prepare food in restaurants or make products in factories may still have jobs; but often they simply push buttons and watch as computerized devices actually perform the needed tasks. In this way, it is possible for computers to cause "de-skilling" of workers, turning them into passive observers and button pushers. Again, however, the picture is not all bad because computers also have generated new jobs which require new sophisticated skills to perform -- for example, "computer assisted drafting" and "keyhole" surgery.

Another workplace issue concerns health and safety. As Forester and Morrison point out [Forester and Morrison, 140-72, Chapter 8], when information technology is introduced into a workplace, it is important to consider likely impacts upon health and job satisfaction of workers who will use it. It is possible, for example, that such workers will feel stressed trying to keep up with high-speed computerized devices -- or they may be injured by repeating the same physical movement over and over -- or their health may be threatened by radiation emanating from computer monitors. These are just a few of the social and ethical issues that arise when information technology is introduced into the workplace.

3.2 Computer Crime

In this era of computer "viruses" and international spying by "hackers" who are thousands of miles away, it is clear that computer security is a topic of concern in the field of Computer Ethics. The problem is not so much the physical security of the hardware (protecting it from theft, fire, flood, etc.), but rather "logical security", which Spafford, Heaphy and Ferbrache [Spafford, et al, 1989] divide into five aspects:

1. Privacy and confidentiality
2. Integrity -- assuring that data and programs are not modified without proper authority
3. Unimpaired service
4. Consistency -- ensuring that the data and behavior we see today will be the same tomorrow
5. Controlling access to resources

Malicious kinds of software, or "programmed threats", provide a significant challenge to computer security. These include "viruses", which cannot run on their own, but rather are inserted into other computer programs; "worms" which can move from machine to machine across networks, and may have parts of themselves running on different machines; "Trojan horses" which appear to be one sort of program, but actually are doing damage behind the scenes; "logic bombs" which check for particular conditions and then execute when those conditions arise; and "bacteria" or "rabbits" which multiply rapidly and fill up the computer's memory.

Computer crimes, such as embezzlement or planting of logic bombs, are normally committed by trusted personnel who have permission to use the computer system. Computer security, therefore, must also be concerned with the actions of trusted computer users.

Another major risk to computer security is the so-called "hacker" who breaks into someone's computer system without permission. Some hackers intentionally steal data or commit vandalism, while others merely "explore" the system to see how it works and what files it contains. These "explorers" often claim to be benevolent defenders of freedom and fighters against rip-offs by major corporations or spying by government agents. These self-appointed vigilantes of cyberspace say they do no harm, and claim to be helpful to society by exposing security risks. However every act of hacking is harmful, because any known successful penetration of a computer system requires the owner to thoroughly check for damaged or lost data and programs. Even if the hacker did indeed make no changes, the computer's owner must run through a costly and time-consuming investigation of the compromised system [Spafford, 1992].

3.3 Privacy and Anonymity

One of the earliest computer ethics topics to arouse public interest was privacy. For example, in the mid-1960s the American government already had created large databases of information about private citizens (census data, tax records, military service records, welfare records, and so on). In the US Congress, bills were introduced to assign a personal identification number to every citizen and then gather all the government’s data about each citizen under the corresponding ID number. A public outcry about "big-

brother government" caused Congress to scrap this plan and led the US President to appoint committees to recommend privacy legislation. In the early 1970s, major computer privacy laws were passed in the USA. Ever since then, computer-threatened privacy has remained as a topic of public concern. The ease and efficiency with which computers and computer networks can be used to gather, store, search, compare, retrieve and share personal information make computer technology especially threatening to anyone who wishes to keep various kinds of "sensitive" information (e.g., medical records) out of the public domain or out of the hands of those who are perceived as potential threats. During the past decade, commercialization and rapid growth of the internet; the rise of the world-wide-web; increasing "user-friendliness" and processing power of computers; and decreasing costs of computer technology have led to new privacy issues, such as data-mining, data matching, recording of "click trails" on the web, and so on [see Tavani, 1999].

The variety of privacy-related issues generated by computer technology has led philosophers and other thinkers to re-examine the concept of privacy itself. Since the mid-1960s, for example, a number of scholars have elaborated a theory of privacy defined as "control over personal information" (see, for example, [Westin, 1967], [Miller, 1971], [Fried, 1984] and [Elgesem, 1996]). On the other hand, philosophers Moor and Tavani have argued that control of personal information is insufficient to establish or protect privacy, and "the concept of privacy itself is best defined in terms of restricted access, not control" [Tavani and Moor, 2001] (see also [Moor, 1997]). In addition, Nissenbaum has argued that there is even a sense of privacy in public spaces, or circumstances "other than the intimate." An adequate definition of privacy, therefore, must take account of "privacy in public" [Nissenbaum, 1998]. As computer technology rapidly advances -- creating ever new possibilities for compiling, storing, accessing and analyzing information -- philosophical debates about the meaning of ‘privacy’ will likely continue (see also [Introna, 1997]).

Questions of anonymity on the internet are sometimes discussed in the same context with questions of privacy and the internet, because anonymity can provide many of the same benefits as privacy. For example, if someone is using the internet to obtain medical or psychological counseling, or to discuss sensitive topics (for example, AIDS, abortion, gay rights, venereal disease, political dissent), anonymity can afford protection similar to that of privacy. Similarly, both anonymity and privacy on the internet can be helpful in preserving human values such as security, mental health, self-fulfillment and peace of mind. Unfortunately, privacy and anonymity also can be exploited to facilitate unwanted and undesirable computer-aided activities in cyberspace, such as money laundering, drug trading, terrorism, or preying upon the vulnerable (see [Marx, 2001] and [Nissenbaum, 1999]).

3.4 Intellectual Property

One of the more controversial areas of computer ethics concerns the intellectual property rights connected with software ownership. Some people, like Richard Stallman who started the Free Software Foundation, believe that software ownership should not be allowed at all. He claims that all information should be free, and all programs should be available for copying, studying and modifying by anyone who wishes to do so [Stallman, 1993]. Others argue that software companies or programmers would not

invest weeks and months of work and significant funds in the development of software if they could not get the investment back in the form of license fees or sales [Johnson, 1992]. Today's software industry is a multibillion dollar part of the economy; and software companies claim to lose billions of dollars per year through illegal copying ("software piracy"). Many people think that software should be ownable, but "casual copying" of personally owned programs for one's friends should also be permitted (see [Nissenbaum, 1995]). The software industry claims that millions of dollars in sales are lost because of such copying. Ownership is a complex matter, since there are several different aspects of software that can be owned and three different types of ownership: copyrights, trade secrets, and patents. One can own the following aspects of a program:

1. The "source code" which is written by the programmer(s) in a high-level computer language like Java or C++.
2. The "object code", which is a machine-language translation of the source code.
3. The "algorithm", which is the sequence of machine commands that the source code and object code represent.
4. The "look and feel" of a program, which is the way the program appears on the screen and interfaces with users.

A very controversial issue today is owning a patent on a computer algorithm. A patent provides an exclusive monopoly on the use of the patented item, so the owner of an algorithm can deny others use of the mathematical formulas that are part of the algorithm. Mathematicians and scientists are outraged, claiming that algorithm patents effectively remove parts of mathematics from the public domain, and thereby threaten to cripple science. In addition, running a preliminary "patent search" to make sure that your "new" program does not violate anyone's software patent is a costly and time-consuming process. As a result, only very large companies with big budgets can afford to run such a search. This effectively eliminates many small software companies, stifling competition and decreasing the variety of programs available to the society [The League for Programming Freedom, 1992].

3.5 Professional Responsibility

Computer professionals have specialized knowledge and often have positions with authority and respect in the community. For this reason, they are able to have a significant impact upon the world, including many of the things that people value. Along with such power to change the world comes the duty to exercise that power responsibly [Gotterbarn, 2001]. Computer professionals find themselves in a variety of professional relationships with other people [Johnson, 1994], including:

employer	-- employee
client	-- professional
professional	-- professional
society	-- professional

These relationships involve a diversity of interests, and sometimes these interests can come into conflict with each other. Responsible computer professionals, therefore, will be aware of possible conflicts of interest and try to avoid them.

Professional organizations in the USA, like the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronic Engineers (IEEE), have established codes of ethics, curriculum guidelines and accreditation requirements to help computer professionals understand and manage ethical responsibilities. For example, in 1991 a Joint Curriculum Task Force of the ACM and IEEE adopted a set of guidelines ("Curriculum 1991") for college programs in computer science. The guidelines say that a significant component of computer ethics (in the broad sense) should be included in undergraduate education in computer science [Turner, 1991].

In addition, both the ACM and IEEE have adopted Codes of Ethics for their members. The most recent ACM Code (1992), for example, includes "general moral imperatives", such as "avoid harm to others" and "be honest and trustworthy". And also included are "more specific professional responsibilities" like "acquire and maintain professional competence" and "know and respect existing laws pertaining to professional work." The IEEE Code of Ethics (1990) includes such principles as "avoid real or perceived conflicts of interest whenever possible" and "be honest and realistic in stating claims or estimates based on available data."

The Accreditation Board for Engineering Technologies (ABET) has long required an ethics component in the computer engineering curriculum. And in 1991, the Computer Sciences Accreditation Commission/Computer Sciences Accreditation Board (CSAC/CSAB) also adopted the requirement that a significant component of computer ethics be included in any computer sciences degree granting program that is nationally accredited [Conry, 1992].

It is clear that professional organizations in computer science recognize and insist upon standards of professional responsibility for their members.

3.6 Globalization

Computer ethics today is rapidly evolving into a broader and even more important field, which might reasonably be called "global information ethics". Global networks like the Internet and especially the world-wide-web are connecting people all over the earth. As Krystyna Gorniak-Kocikowska perceptively notes in her paper, "The Computer Revolution and the Problem of Global Ethics" [Gorniak-Kocikowska, 1996], for the first time in history, efforts to develop mutually agreed standards of conduct, and efforts to advance and defend human values, are being made in a truly global context. So, for the first time in the history of the earth, ethics and values will be debated and transformed in a context that is not limited to a particular geographic region, or constrained by a specific religion or culture. This may very well be one of the most important social developments in history. Consider just a few of the global issues:

Global Laws

If computer users in the United States, for example, wish to protect their freedom of speech on the internet, whose laws apply? Nearly two hundred countries are already interconnected by the internet, so the United States Constitution (with its First Amendment protection for freedom of speech) is just a "local law" on the internet -- it does not apply to the rest of the world. How can issues like freedom of speech, control of "pornography", protection of intellectual property, invasions of privacy, and many others to be governed by law when so many countries are involved? If a citizen in a European country, for example, has internet dealings with someone in a far-away land, and the government of that land considers those dealings to be illegal, can the European be tried by the courts in the far-away country?

Global Cyberbusiness

The world is very close to having technology that can provide electronic privacy and security on the internet sufficient to safely conduct international business transactions. Once this technology is in place, there will be a rapid expansion of global "cyberbusiness". Nations with a technological infrastructure already in place will enjoy rapid economic growth, while the rest of the world lags behind. What will be the political and economic fallout from rapid growth of global cyberbusiness? Will accepted business practices in one part of the world be perceived as "cheating" or "fraud" in other parts of the world? Will a few wealthy nations widen the already big gap between rich and poor? Will political and even military confrontations emerge?

Global Education

If inexpensive access to the global information net is provided to rich and poor alike -- to poverty-stricken people in ghettos, to poor nations in the "third world", etc.-- for the first time in history, nearly everyone on earth will have access to daily news from a free press; to texts, documents and art works from great libraries and museums of the world; to political, religious and social practices of peoples everywhere. What will be the impact of this sudden and profound "global education" upon political dictatorships, isolated communities, coherent cultures, religious practices, etc.? As great universities of the world begin to offer degrees and knowledge modules via the internet, will "lesser" universities be damaged or even forced out of business?

Information Rich and Information Poor

The gap between rich and poor nations, and even between rich and poor citizens in industrialized countries, is already disturbingly wide. As educational opportunities, business and employment opportunities, medical services and many other necessities of life move more and more into cyberspace, will gaps between the rich and the poor become even worse?

3.7 The Metaethics of Computer Ethics

Given the explosive growth of Computer ethics during the past two decades, the field appears to have a

very robust and significant future. Two important thinkers, however, Krystyna Gorniak-Kocikowska and Deborah Johnson, have recently argued that computer ethics will disappear as a separate branch of ethics. In 1996 Gorniak-Kocikowska predicted that computer ethics, which is currently considered a branch of applied ethics, will eventually evolve into something much more.[1] According to her hypothesis, "local" ethical theories like Europe's Benthamite and Kantian systems and the ethical systems of other cultures in Asia, Africa, the Pacific Islands, etc., will eventually be superceded by a global ethics evolving from today's computer ethics. "Computer" ethics, then, will become the "ordinary" ethics of the information age.

In her 1999 ETHICOMP paper [Johnson, 1999], Johnson expressed a view which, upon first sight, may seem to be the same as Gorniak's.[2] A closer look at the Johnson hypothesis reveals that it is a different kind of claim than Gorniak's, though not inconsistent with it. Johnson's hypothesis addresses the question of whether or not the name "computer ethics" (or perhaps "information ethics") will continue to be used by ethicists and others to refer to ethical questions and problems generated by information technology. On Johnson's view, as information technology becomes very commonplace -- as it gets integrated and absorbed into our everyday surroundings and is perceived simply as an aspect of ordinary life -- we may no longer notice its presence. At that point, we would no longer need a term like "computer ethics" to single out a subset of ethical issues arising from the use of information technology. Computer technology would be absorbed into the fabric of life, and computer ethics would thus be effectively absorbed into ordinary ethics.

Taken together, the Gorniak and Johnson hypotheses look to a future in which what we call "computer ethics" today is globally important and a vital aspect of everyday life, but the name "computer ethics" may no longer be used.

Bibliography

- Anderson, Ronald, Deborah Johnson, Donald Gotterbarn and Judith Perrolle (February 1993) "Using the New ACM Code of Ethics in Decision Making," Communications of the ACM, Vol. 36, 98-107.
- Brey, Philip (2001) "Disclosive Computer Ethics." In R. A. Spinello and H. T. Tavani, eds., Readings in CyberEthics, Jones and Bartlett.
- Bynum, Terrell Ward (1993) "Computer Ethics in the Computer Science Curriculum." In Bynum, Terrell Ward, Walter Maner and John L. Fodor, eds. (1993) Teaching Computer Ethics, Research Center on Computing & Society.
- Bynum, Terrell Ward (1999), "The Foundation of Computer Ethics", a keynote address at the AICEC99 Conference, Melbourne, Australia, July 1999. Published in the June 2000 issue of Computers and Society.
- Conry, Susan (1992) "Interview on Computer Science Accreditation." In Bynum, Terrell Ward and John L. Fodor, creators, Computer Ethics in the Computer Science Curriculum (a video program), Educational Media Resources.
- Elgesem, Dag (1996) "Privacy, Respect for Persons, and Risk." In Ess, Charles, ed., Philosophical

Perspectives on Computer-Mediated Communication, State University of New York Press.

- Fodor, John L. and Terrell Ward Bynum, creators. (1992) What Is Computer Ethics? [a video program], Educational Media Resources.
- Forester, Tom and Perry Morrison (1990) Computer Ethics: Cautionary Tales and Ethical Dilemmas in Computing, MIT Press.
- Fried, Charles (1984) "Privacy." In Schoeman, F. D., ed., Philosophical Dimensions of Privacy, Cambridge University Press.
- Friedman, Batya, ed. (1997) Human Values and the Design of Computer Technology, Cambridge University Press.
- Friedman, Batya and Helen Nissenbaum (1996) "Bias in Computer Systems", ACM Transactions on Information Systems, Vol. 14, No. 3, 330-347.
- Gorniak-Kocikowska, Krystyna (1996) "The Computer Revolution and the Problem of Global Ethics." In Bynum and Rogerson (1996) Global Information Ethics, Opragen Publications, 177-90.
- Gotterbarn, Donald (1991) "Computer Ethics: Responsibility Regained," National Forum: The Phi Beta Kappa Journal, Vol. 71, 26-31.
- Gotterbarn, Donald (2001) "Informatics and Professional Responsibility", Science and Engineering Ethics, Vol. 7, No. 2.
- Gotterbarn, Donald, Keith Miller, and Simon Rogerson (1997) "Software Engineering Code of Ethics," Information Society, Vol. 40, No. 11, 110-118.
- Introna, Lucas D. (1997) "Privacy and the Computer: Why We Need Privacy in the Information Society," Metaphilosophy, Vol. 28, No. 3, 259-275.
- Introna, Lucas D. and Helen Nissenbaum (2000) "Shaping the Web: Why the Politics of Search Engines Matters", The Information Society, Vol. 16, No.3, 1-17.
- Johnson, Deborah G. (1985) Computer Ethics, Prentice-Hall, 2nd Edition, 1994.
- Johnson, Deborah G. (1992) "Proprietary Rights in Computer Software: Individual and Policy Issues." In Bynum, Terrell Ward, Walter Maner and John L. Fodor, eds. (1992) Software Ownership and Intellectual Property Rights, Research Center on Computing & Society.
- Johnson, Deborah G. (1999) "Computer Ethics in the 21st Century", a keynote address at the ETHICOMP99 Conference, Rome, Italy, October 1999. Published in Spinello, Richard A. and Herman T. Tavani, eds. (2001) Readings in CyberEthics, Jones and Bartlett.
- Kocikowski, Andrzej (1996) "Geography and Computer Ethics: An Eastern European Perspective." In Bynum, Terrell Ward and Simon Rogerson, eds. (1996) Global Information Ethics, Opragen Publications, 201-10. (The April 1996 issue of Science and Engineering Ethics)
- The League for Programming Freedom (1992) "Against Software Patents." In Bynum, Terrell Ward, Walter Maner and John L. Fodor, eds. (1992) Software Ownership and Intellectual Property Rights, Research Center on Computing & Society.
- Maner, Walter (1980) Starter Kit in Computer Ethics, Helvetia Press (published in cooperation with the National Information and Resource Center for Teaching Philosophy). [Originally self-published by Maner in 1978.]
- Maner, Walter (1996) "Unique Ethical Problems in Information Technology," In Bynum and Rogerson. (1996) 137-52.
- Marx, Gary T. (2001) "Identity and Anonymity: Some Conceptual Distinctions and Issues for

Research". In J. Caplan and J. Torpey, Documenting Individual Identity. Princeton University Press.

- Miller, A. R. (1971) The Assault on Privacy: Computers, Data Banks, and Dossiers, University of Michigan Press.
- Moor, James H. (1985) "What Is Computer Ethics?" In Bynum, Terrell Ward, ed. (1985) Computers and Ethics, Blackwell, 266-75. [Published as the October 1985 issue of Metaphilosophy.]
- Moor, James H. (1997) "Towards a Theory of Privacy in the Information Age," Computers and Society, Vol. 27, No. 3, 27-32.
- Nissenbaum, Helen (1995) "Should I Copy My Neighbor's Software?" In D. Johnson and H. Nissenbaum, eds., Computers, Ethics, and Social Responsibility, Prentice Hall.
- Nissenbaum, Helen (1998) "Protecting Privacy in an Information Age: The Problem of Privacy in Public," Law and Philosophy, Vol. 17, 559-596.
- Nissenbaum, Helen (1999) "The Meaning of Anonymity in an Information Age," The Information Society, Vol. 15, 141-144.
- Parker, Donn (1968) "Rules of Ethics in Information Processing," Communications of the ACM, Vol. 11., 198-201.
- Parker, Donn (1979) Ethical Conflicts in Computer Science and Technology. AFIPS Press.
- Parker, Donn, S. Swope and B.N. Baker (1990) Ethical Conflicts in Information & Computer Science, Technology & Business, QED Information Sciences.
- Perrolle, Judith A. (1987) Computers and Social Change: Information, Property, and Power. Wadsworth.
- Rogerson, Simon (Spring 1996) "The Ethics of Computing: The First and Second Generations," The UK Business Ethics Network News.
- Rogerson, Simon and Terrell Ward Bynum (June 9, 1995) "Cyberspace: The Ethical Frontier," Times Higher Education Supplement, The London Times.
- Sojka, Jacek (1996) "Business Ethics and Computer Ethics: The View from Poland." In Bynum and Rogerson. (1996) Global Information Ethics, Opragen Publications, 191-200.
- Spafford, Eugene, et al. (1989) Computer Viruses: Dealing with Electronic Vandalism and Programmed Threats, ADAPSO.
- Spafford, Eugene (1992) "Are Computer Hacker Break-Ins Ethical?" Journal of Systems and Software, January 1992, Vol. 17, 41-47.
- Spinello, Richard A. and Herman T. Tavani, eds. (2001) Readings in CyberEthics, Jones and Bartlett.
- Stallman, Richard (1992) "Why Software Should Be Free." In Bynum, Terrell Ward, Walter Maner and John L. Fodor, eds. (1992) Software Ownership and Intellectual Property Rights, Research Center on Computing & Society, 35-52.
- Tavani, Herman T. (1999) "Privacy On-Line," Computers and Society, Vol. 29, No. 4, 11-19.
- Tavani, Herman T. and James H. Moor (2001) "Privacy Protection, Control of Information, and Privacy-Enhancing Technologies", Computers and Society, Vol. 31, No. 1, 6-11.
- Turkle, Sherry (1984) The Second Self: Computers and the Human Spirit, Simon & Schuster.
- Turner, A. Joseph (1991) "Summary of the ACM/IEEE-CS Joint Curriculum Task force Report: Computing Curricula, 1991," Communications of the ACM, Vol. 34, No. 6., 69-84.

- van Speybroeck, James (July 1994) "Review of Starter Kit on Teaching Computer Ethics" (Terrell Ward Bynum, Walter Maner and John L. Fodor, eds.) Computing Reviews, 357-8.
- Weizenbaum, Joseph (1976) Computer Power and Human Reason: From Judgment to Calculation, Freeman.
- Westin, Alan R.(1967) Privacy and Freedom, Atheneum.
- Wiener, Norbert (1948) Cybernetics: or Control and Communication in the Animal and the Machine, Technology Press.
- Wiener, Norbert (1950/1954) The Human Use of Human Beings: Cybernetics and Society, Houghton Mifflin, 1950. (Second Edition Revised, Doubleday Anchor, 1954.)

Other Internet Resources

- ACM SIGCAS, Special Interest Group on Computers and Society, Association for Computing Machinery
- Australian Institute of Computer Ethics
- Centre for Computing and Social Responsibility, (De Montfort University, UK)
- Centre for Philosophy of Information and Communication Technology, (Erasmus University, the Netherlands)
- Computer Ethics Bibliography, by Herman Tavani (Rivier College)
- Electronic Frontier Foundation
- Electronic Privacy Information Center
- Research Center on Computing & Society, (Southern Connecticut State University)
- Software Engineering Ethics Research Institute, (East Tennessee State University)

Related Entries

privacy | property

Terrell Ward Bynum
bynum@southernct.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

A to Z Table of Contents

First published: August 13, 2001

Content last modified: August 13, 2001

Stanford Encyclopedia of Philosophy
Notes to Computer Ethics

Notes

1. It will evolve, she said, into a system of global ethics applicable in every culture on earth. In Gorniak-Kocikowska [1996], we find:

> Just as the major ethical theories of Bentham and Kant were developed in response to the printing press revolution, so a new ethical theory is likely to emerge from computer ethics in response to the computer revolution. The newly emerging field of information ethics, therefore, is much more important than even its founders and advocates believe. (p. 177)
>
> ... The very nature of the Computer Revolution indicates that the ethic of the future will have a global character. It will be global in a spatial sense, since it will encompass the entire Globe. It will also be global in the sense that it will address the totality of human actions and relations. (p. 179)
>
> ... the rules of computer ethics, no matter how well thought through, will be ineffective unless respected by the vast majority of or maybe even all computer users. This means that in the future, the rules of computer ethics should be respected by the majority (or all) of the human inhabitants of the Earth ... . In other words, computer ethics will become universal, it will be a global ethic. (p. 187)

2. In Johnson [1999], we find:

> I offer you a picture of computer ethics in which computer ethics as such disappears. ... We will be able to say both that computer ethics has become ordinary ethics and that ordinary ethics has become computer ethics. (pp. 17-18)

Terrell Ward Bynum
bynum@southernct.edu

First published: August 13, 2001
Content last modified: August 13, 2001

Medieval Theories of Causality

Causality plays an important role in medieval philosophical writing: the dominant genre of medieval academic writing was the commentary on an authoritative work, very often a work of Aristotle. Of the works of Aristotle thus commented on, the *Physics* plays a central role. Other of Aristotle's scientific works -- *On the Heavens and the Earth*, *On Generation and Corruption* - are also significant: so there is a rather daunting body of work to survey.

One might, though, be tempted to argue that this concentration on causality is simply an effect of reading Aristotle, but this would be too hasty. Medieval thinkers were attracted to the problem of causality long before most of Aristotle's texts became available in the thirteenth century: already in the twelfth century the created universe was seen as a rational manifestation of God (Wetherbee 1988, p. 25), and, consequently, the rational investigation of the universe was seen as a way of approaching God: "In the creation of things", says William of Conches, "divine power, wisdom and goodness are beheld" (William of Conches, *Glosa super Platonem*, p. 60).

Even apart from direct literary influence, the nature of the philosophical and theological themes which were popular in the Middle Ages also led to an emphasis on causality. Writers studied the interrelationship of divine grace and natural processes, the role of the will in ethics, free will and determinism: all of these problems have an important causal component. These questions were often handled by methods which might seem to us to be extraordinarily naturalistic - naturalistic, of course, in the sense of the modes of natural investigation which were current at the time. It comes as no surprise to know that many medieval thinkers discussed the question of whether divine grace can increase: what *is* surprising is that many of the discussions use the technical tools of Aristotle's physical and biological works, tools which were originally developed to discuss problems of continuity and change in the natural world. What is even more surprising is the technical proficiency of many of these discussions: fourteenth-century work on this topic gave rise to very acute analyses of the variation of continuous quantities (see Murdoch 1975).

What should become evident during this survey is the extremely tight and complex interconnection between medieval causal theories and medieval ontology. After Aristotle's texts had been assimilated, almost all medieval academic theories had an ontology which was basically hylomorphic: substances were composites of matter and form, and change was described as the loss of one form and the acquisition of another. Form was not merely shape, but an active principle: the form of a thing was responsible for its causal role (White 1984). Furthermore, in any causal interaction, the allocation of active and passive roles to the individuals involved tended to be thought of as unproblematic. Although many aspects of

Aristotle's causal theories were extensively and critically debated, this basic hylomorphism persisted throughout; and it is this, rather than anything more arcane, which often poses the greatest problems in assimilating, or evaluating, medieval thought on these topics.

Causality and Motion

The term ‘motion’, in Aristotelean philosophy, can stand for a wide range of changes of state, and not simply changes of place (the latter is usually known as *local* motion). Aristotle's *Physics* is basically an exhaustive study of motion in this very wide sense. However, local motion is an interesting topic, and we shall start with it.

Motions are, in Aristotle's physics, classified into *natural* and *violent*. A paradigmatic example of natural (local) motion is the motion of a freely falling body, whereas an example of violent (local) motion would be the motion of a thrown body. If we throw a body, then it is relatively unproblematic to account for the motion when it is in contact with our hand: what is difficult is to account for its continued motion thereafter. Aristotle's theory accounts for it by saying that, when it is moving, a temporary vacuum is caused behind it, and, in order to fill in this vacuum, air rushes around from the front, thus leaving a void in front of the projectile which is filled by the continued motion of the projectile. This explanation was vulnerable to a large number of objections -- for example, it is clearly easier to throw a moderately heavy object, such as a stone, than a light object, such as a bean, whereas light objects ought to be more susceptible than others to motions of the air. And Aristotle's theory, when confronted with the example of two stones thrown in opposite directions so as to pass near to one another, cannot consistently say how the air is supposed to move in the neighbourhood of their close encounter. These objections were made by numerous medieval authors, most significantly by John Buridan (*De Caelo et Mundo* III, qu. 22, pp. 227ff.) and Nicole Oresme (*Du ciel et du monde* II, ch. 25ff., pp. 525ff.).

This critique of Aristotle's theory of projectile motion did not come out of nowhere. Aristotle relied on a concept of natural motion, and *that*, in turn, relied on a concept of natural place: natural motion was motion towards the natural place of a body (i.e. motion downwards in the case of earth, and motion upwards in the case of fire). (Aristotle, *Physics* IV.5, 212b30-213a5) Ockham is already quite equivocal

about the concept of natural place: and this is for several reasons.

1. One is that -- as we shall see later -- he is generally quite suspicious about teleology, and the concept of natural place is basically a teleological one. Correspondingly, Ockham attempts -- not very successfully -- to explain the kinematics associated with natural place in terms of *efficient* causality. (Ockham, *Expositio Physicorum* IV, c6: *Opera Philosophica* V, p. 78; Goddu 1984, pp. 122ff.).
2. Another reason is because of several examples which tend to undermine the difference between rest and motion. Ockham, and many other medievals, have reductionist accounts of place in terms of contact between bodies; the place of a body is just the surfaces of the bodies surrounding it (Ockham, *Expositio Physicorum* IV, c6: *Opera Philosophica* V, pp. 55ff.). So, if we have a ship in a river, which is flowing, is the place of the ship the surface of the surrounding water? Is this, then, a moving place? And what, then, would be the relation between that moving place and fixed places? Ockham eventually decides that there are only fixed places, but his arguments are not very strong, and one is left with the impression that the very ideas of rest and motion have become somewhat problematic. (Ockham, *Expositio Physicorum* IV, c7: *Opera Philosophica* V, pp. 79ff.)
3. The final reason is motivated by a theological example: we can suppose that God could create another world than this, but, in that case, what would the earth of that world do? Would it move towards the centre of *this* world (which seems to us to be the natural place of earth)? Or towards the centre of the other world? (Ockham, I *Sent.*, d. 44: *Opera Theologica* IV, pp. 655-56; Goddu 1984, p. 124. See also Marsilius of Inghen, *Si essent plures mundi*.)

Correspondingly, both Buridan and Oresme are sceptical, not only about Aristotle's theory of projectile motion, but also about the related notions of natural place, motion, and rest. They both state -- Oresme far more emphatically -- that it would be consistent with all we observe if the earth were to rotate while the heavens remained at rest; Oresme and Buridan have, on these grounds, been described as "precursors of Galileo".

However, what is more interesting for us are the alternative causal accounts that Buridan and Oresme adopted: they both said that projectiles move violently because of a form inherent in them, which led them to move in a non-natural direction, and which naturally decayed. This form was known as ‘impetus’, and was a common theme in the philosophy of the thirteenth and fourteenth centuries; some version of an impetus theory goes back to the early thirteenth century (Wood 1992). There was, especially in the fourteenth century, a considerable amount of quantitative work on impetus which attempted to establish such things as the law according to which impetus decayed (Weisheipl 1982, pp. 535ff.).

What is significant here is that -- despite the radical changes in cosmology -- this is is still an extremely *medieval* theory: causation is due to forms inhering in substances, and there is a division of the substances involved into agents and patients. Instead of there being a single form involved in projectile motion -- the form of heaviness, responsible for natural motion downwards -- there are two, weight and impetus, and the two conflict. The basic ontology is still the same, and the division into agents and patients, though its details may have changed, still persists. Furthermore, despite persistent doubts, there is still something of a distinction between motion and rest, and motion can only be the result of agency. Contrast this with

Galileo's or -- still more -- Newton's account: here uniform motion and rest are treated on an equal footing, and, consequently, there can be no unequivocal distinction between motion and rest. So, although Buridan and Galileo are -- in some sense -- precursors of Galileo, their causal ontology is still, in important respects, thoroughly medieval (Maier 1964)

Causality, Self-Motion, and the Will

An example of motion in the wider sense is an act of the will: it is a change of state of some entity (namely the mind or soul), but would not have been thought of as *local* motion by most medieval thinkers -- thought and will were generally regarded as immaterial processes (see Cross 1999, p. 75).

Aristotle has a picture of willed action in which actions are caused by combinations of beliefs and desires: these belief-desire states are not, of course, themselves actions (Normore 1998). This picture of the will fits with one of Aristotle's major causal doctrines: that nothing causes a change in itself.

However, Aristotle's picture of the will was not undisputed in the Middle Ages: as early as the twelfth century, Anselm had outlined a theory in which the will was a self-mover, and in which moral conflict was explained by the presence of two wills in the same person (Normore 1998, p. 28). This position was later taken up, in conscious opposition to Aristotle, by thinkers in the Franciscan school -- Peter Olivi, and then Scotus and Ockham.

Scotus follows a modified Anselmian line, speaking of a single will, with two inclinations: one towards self-fulfillment, the other towards justice). It is the presence of these two inclinations which distinguishes willed causes from natural causes: natural causes are determined to perform their acts (unless impeded), whereas the will is not thus determined (Scotus, *Metaphysics* IV, 9: in Scotus, *On the Will and Morality*, pp. 136ff.; Lee 1998; Cross 1999, pp. 84ff.). The will is thus self-determining, rather than determined by its end, and so Scotus affirms self-motion in psychology. In fact, he goes further, and admits self-motion in physical cases as well: for example, a falling object is actively moving towards its goal, and its motion is caused by itself (because it is heavy); so this, too, is an instance of self-motion (Effler 1962).

Ockham expands on Scotus's theory of the will to deny that actions are properly explained by their ends: we are *influenced* by ends, but our actions are not *necessitated* by them and thus not caused by them (Ockham, *Quodlibet* I, qu. 16: *Opera Theologica* IX, pp. 87ff.). A free agent is one which, under exactly the same circumstances, could have chosen otherwise; and so a free agent can reject the Beatific Vision (and, in fact, actively turn to any other object whatever). (Ockham, *Quodlibet* IV, q. 1: *Opera Theologica* IX, pp. 292ff.).

Causal Accounts of Perception

Perception was, throughout the Middle Ages, a contentious topic, and it was also a topic in which the answers to strictly *causal* questions could influence philosophical positions in other areas (for example,

on whether certain knowledge of external entities was attainable). The "traditional" view, dating back to Roger Bacon in the mid-thirteenth century, was that physical objects were known because they caused a succession of likenesses, or *species*, first in the medium between the object and the perceiver, then in the senses, and finally in the intellect, of the perceiver (Tachau 1988, pp. 3ff.). This position was attacked by thinkers such as Henry of Ghent, Peter Olivi, and Duns Scotus. Interestingly, many of these criticisms tend towards a relational account of perception, in which -- although *species* still play a role -- the role that they play is to be a means by which we know things, and in which the *species* themselves are not known directly but only by reflection. (Tachau 1988, p. 66)

Ockham then radicalised these critiques by denying that there are any such *species* at all: perception and other phenomena which were usually explained by *species* -- the sun heating or illuminating physical objects, for example -- were now explained by action at a distance (Tachau 1988, pp. 130ff.). There was a similar debate about the causal mechanisms behind memory, where, again, Ockham denied a *species*-based account; however, in the case of memory, he replaced species not with action at a distance but with habits (Wolter and Adams 1993).

Ockham denies species not on the basis of empirical evidence, or on the basis of epistemological arguments, but purely and simply on the basis of his razor: if we deny *species*, then we can give an account of the phenomena which uses fewer entities, because *species* are entities. Although this position of Ockham's did not have much influence on his contemporaries or followers -- it is, after all, extremely implausible -- it is a good example of how causal reasoning is affected by tacit ontological assumptions: the fact that *species* were seen as entities and the fact that Ockham had a programme of reducing the number of entities, led to an account of perception which tried to do away with *species*. On the other hand, action at a distance was, despite its implausibility, entirely unaffected by Ockham's critique. And, similarly, Ockham's account was not noticeably *simpler* than the accounts it criticised, which shows how far Ockham's own razor was from the principles of simplicity and the like, which are usually considered to be its modern equivalents.

Causality, Knowledge and Necessity

There is a persistent supposition -- see, for example, (Gilson 1937) - that Ockham, and many of his fourteenth-century followers, had a basically Humean position on causality; this supposition has deep historical roots (Nadler 1996), but is inaccurate (Adams 1987, pp. 741ff.).

The supposedly Humean position has three basic assertions: that there is nothing more to causality than the regular sequence of phenomena, that such a regular sequence cannot give a necessary connection, and that, consequently, we can have no certain knowledge of causal relations.

One item in this chain of argument has some textual support in Ockham: he did not believe that the relation of efficient causality was a *thing* distinct from its *relata* (Ockham, *Quodlibet* VI, qu. 12: *Opera Theologica* IX, pp. 629ff.) However, one can still believe this and hold that causality is a real relation, and Ockham *did* so believe (Adams 1987, p. 744; White 1990b). So this link in the chain is not found in

Ockham.

The "Humean" argument, in addition, makes a detour through psychology: as Adams analyses it, it relies on a premise like "There can be nothing more in concepts than there actually is in intuitions" (Adams 1987, p. 744). But such a detour through psychology, though widely practiced in the eighteenth century, was somewhat foreign to medieval thought (White 1990a).

Even though pseudo-Humean arguments of this sort cannot reasonably be ascribed to Ockham or to most other medieval thinkers -- with the possible exception of Nicholas of Autrecourt -- there still remains the question of what their views on these questions actually were. Since the medievals generally did not conflate ontological and epistemological issues, there are two questions: first about the necessity of causality, and second about whether we can know causal propositions with certainty.

Causality and Necessity

Medieval thinkers believed that the world was created by God, and so a question like "Is proposition P contingent?" were seen as equivalent to the question "Could God have created a world in which P does not hold?". So our question can be reduced to one about divine power.

A very common theme in medieval thought is the distinction between God's absolute and ordered, or ordained, power (*potentia absoluta* and *potentia ordinata*). This distinction goes back to early medieval thought (Moonan 1994), and was extensively used in later medieval philosophy (Courtenay 1971; Adams 1987, pp. 1186ff.).

God's absolute power is unrestricted power. According to this power, God can create a huge variety of possible worlds. One frequently used principle is this: given two distinct entities, God can create a world in which one of them, but not the other, exists, or, in *this* world, God can destroy one of them, leaving the other intact. We should note that this is not exactly innocuous; ontologically, it amounts to some sort of logical atomism. See (White 1990b).

But God will, in practice, not exercise absolute power: as Aquinas puts it, "what is attributed to the divine power insofar as the command of a just will executes it, God is said to be able to do with respect to His ordered power". (Aquinas, *Summa theologiae* I, qu. 25, a. 5, ad 1) So there are limits to God's ordained power (which come from the concept of a just agent): inside the space of worlds which God could create by absolute power, there is a space of worlds which could be created by ordered power. It is this smaller space of worlds which is relevant for our question of the necessity of causal connections. And, with respect to God's ordered power, there was a wide range of causal assertions which were regarded as necessary by medieval thinkers.

Knowing Causal Propositions: Demonstration

As far as our knowledge of causal propositions is concerned, we can again draw a distinction. One

question is this: do medieval thinkers, in practice, establish causal propositions on the basis of argument? And the other is this: what sort of metatheory of causal argument do the medievals have?

The answer to the first question is quite straightforward. Ockham, like other fourteenth-century theologians, frequently gives instances where we can make reliable causal inferences and come to know causal propositions on the basis of experience (Ockham, *Ordinatio* Prologue, qu. 2: *Opera Theologica* I, p. 87) These arguments frequently rely on a theory of natural kinds: for example, Ockham writes

> Because someone sees that, after eating such an herb, health follows for someone with a fever, and because he can eliminate all other causes of health for that person, he knows evidently that that herb was the cause of health; and thus he has knowledge (*experimentum*) in the singular case. It is, however, obvious to him that all individuals of the same kind have an effect of the same kind in a patient of the same kind; and thus he assents evidently, as to a principle, that every herb of such a kind cures fever. (Ockham, *Ordinatio* prologue, qu. 2: *Opera Theologica* I, p. 87)

The second question is that of a metatheory. Here the story becomes somewhat more complicated. There was a generally accepted metatheory, namely that of Aristotle's *Posterior Analytics*, according to which scientific demonstrations were syllogistic proofs, based on necessary and self-evident premises. There were two sorts of these: proofs of the simple fact (*demonstrationes quia*) and proofs of the reasoned fact (*demonstrationes propter quid*). In the latter, the syllogisms involved must have middle terms that are *causes* of the state of affairs which is to be demonstrated. This gives a theory of scientific reasoning in which the structure of the arguments is intimately tied up with the structure of the causal chains that they demonstrate.

There is, indeed, an extensive literature of medieval commentaries on the *Posterior Analytics*, and much of this literature is very important; we find in it a great deal of material on the authors' attitudes to necessity, the structure of science, the relation between various sciences, the autonomy of philosophy *vis-à-vis* theology, and the like. However, it cannot be taken to be automatically relevant to the *practice* of reasoning in the Middle Ages: the logical metatheory (that of the syllogism) is far too restrictive, and the conditions placed on scientific demonstrations are far too stringent, for it to be a plausible description of very many actual processes of reasoning, in the Middle Ages or at any other time.

Final Causes

We often find in Aristotle and in the literature influenced by him an enumeration of four types of cause: formal, material, efficient and final. The first two are uses of ‘cause’ in a somewhat wider sense than is current nowadays: the term here simply means ‘explanation in general’ (Ockham, *Expositio Physicorum* II, c11: *Opera Philosophica* IV, p. 348), and explanations by means of matter and form were common both in Aristotle and in the literature. Efficient causes are what we would now simply call ‘causes’. Final causes, however, are problematic: a final cause is an end or a purpose, and, whereas it is clear that *rational* agents act for the sake of ends, it is not clear that much else does. Furthermore, it also seems

clear to us that the causality of a rationally pursued goal can be reduced to efficient causality.

Aristotle, however, has a much stronger position on final causality: he believes that there are processes in nature (the growing of a tree, for example) which are completed and regulated by a final state, or end, towards which they tend. As Adams puts it,

> According to Aristotelian metaphysics, natures are complexes of powers. When appropriately coordinated, the collective exercise of such powers converges on an end. In the sublunary world, elemental powers are simple and deterministic. Even where more complex living things are concerned, the "coordination" of their powers is "built-in" in such a fashion that -- given relevant circumstances -- they function to achieve their end. (Adams 1996, p. 499)

Aristotle's natural science tends to be governed by the biological paradigm, and it is clear that, for him, final causes in this strong sense are extremely pervasive. He also argues in the *Physics* that natural processes cannot all be explained by final causality alone, which implies that final causality cannot, in general, be reduced to efficient causality.

The medieval literature is far from unanimous on these questions. William of Ockham, for example, who wrote several commentaries on Aristotle's *Physics*, and who discusses these questions at numerous places in his commentary on Aristotle's *Physics*, hardly has a uniform position. He is quite happy with explanations of natural phenomena by means of efficient causes in general, but he will also often speak of final causes: what is unclear is whether the final causes he speaks of (with varying degrees of strength in different works) have any explanatory role to play that cannot be reduced to efficient causality (Adams 1998).

Bibliography

Texts

- John Buridan (1942), *Quaestiones super Libros Quattuor de Caelo et Mundo*, ed. E. A. Moody, Cambridge, MA: Medieval Academy of America.
- John Duns Scotus (1997), *On the Will and Morality*, selected and translated by Allan B. Wolter, Washington, DC: Catholic University of America Press.
- Marsilius of Inghen (1992), "*Si essent plures mundi*, (*Quaestiones libri de caelo et mundo* I, qu. xiv)," in Braakhuis and Hoenen (1992), 108-116.
- Nicole Oresme (1968), *Le Livre du ciel et du monde*, tr. A. J. Menut, Madison: University of Wisconsin Press.
- Thomas Aquinas (1952-6), *Summa Theologiae*, Turin: Marietti.
- William of Conches (1965), *Glosa super Platonem*, ed. F. Jeauneau, Paris: Vrin.
- William of Ockham (1985), *Expositio in Libros Physicorum Aristotelis*, in *Opera Philosophica* IV-V, St. Bonaventure, NY: St. Bonaventure University Press.

- William of Ockham (1984), *de Fine* ("Utrum ex hoc quod aliquid moveat ut finis sequatur ipsum habere aliquod esse reale extra animam"), in *Opera Theologica* VIII, St. Bonaventure, NY: St. Bonaventure University Press, pp. 98-154.
- William of Ockham (1967-77), *Scriptum in Librum Primum Sententiarum: Ordinatio*, in *Opera Theologica* I-IV, St. Bonaventure, NY: St. Bonaventure University Press.
- William of Ockham (1980), *Quodlibeta*, in *Opera Theologica* IX, St. Bonaventure: St. Bonaventure University Press.

Secondary Literature

- Adams, Marilyn McCord (1979), "Was Ockham a Humean about Efficient Causality?," *Franciscan Studies* 39, 5-48.
- Adams, Marilyn McCord (1987), *William Ockham*, Notre Dame, IN: University of Notre Dame Press.
- Adams, Marilyn McCord (1996), "Scotus and Ockham on the Connection of the Virtues," in Honnefelder *et al.* 1996, 499-522
- Adams, Marilyn McCord (1998), "Ockham on Final Causality: Muddying the Waters" *Franciscan Studies* 56, 1-46 Mediaeval Academy of America.
- Braakhuis, H. A. G. and M. J. F. M. Hoenen (1992), *Marsilius of Inghen: Acts of the International Marsilius of Inghen Symposium* Nijmegen: Ingenium.
- Courtenay, William J. (1971), "Covenant and Causality in Pierre d'Ailly," *Speculum* 46, 94-119. Reprinted in William J. Courtenay, *Covenant and Causality in Medieval Thought*, London: Variorum Reprints.
- Craig, William Lane (1980), *The Cosmological Argument from Plato to Leibniz*, London: Macmillan.
- Cross, Richard (1999), *Duns Scotus*, Oxford: Oxford University Press.
- Effler, Roy R. (1962), *John Duns Scotus and the Principle "Omne quod movetur ab alio movetur,"* St. Bonaventure, NY: Franciscan Institute.
- Gilson, E. (1937), *The Unity of Philosophical Experience* New York: Scribners.
- Honnefelder, L., R. Wood, and M. Dreyer (1996), *John Duns Scotus: Metaphysics and Ethics*, Leiden: Brill.
- Goddu, A. (1984), *The Physics of William of Ockham*, Leiden: Brill.
- Kretzmann, N., A. Kenny, and J. Pinborg (1982), *The Cambridge History of Later Medieval Philosophy*, Cambridge: Cambridge University Press.
- Lee, Sukjae (1998), "Scotus on the Will: The Rational Power and the Dual Affections," *Vivarium* 36, 40-54.
- Maier, Anneliese (1964), "‘Ergebnisse’ der Spätscholastischen Naturphilosophie," in *Ausgehendes Mittelalter: Gesammelte Aufsätze zur Geistesgeschichte des 14. Jahrhunderts*, Roma: Edizioni di Storia e Letteratura, 425-57.
- Moonan, Lawrence (1994), *Divine Power: The Medieval Power Distinction up to its Adoption by Albert, Bonaventure, and Aquinas* Oxford: Clarendon Press.
- Murdoch, John E. (1975), "From Social into Intellectual Factors: An Aspect of the Unitary Character of Medieval Learning," in *The Cultural Context of Medieval Learning*, ed. John E.

Murdoch and Edith Sylla, Dordrecht: Reidel, 271-348.

- Nadler, Steven (1996), "'No Necessary Connection': The Medieval Roots of the Occasionalist Roots of Hume," *The Monist* 79, 448-466.
- Normore, Calvin (1998), "Picking and Choosing: Anselm and Ockham on Choice," *Vivarium* 36, 23-39.
- Tachau, Katherine H. (1988), *Vision and Certitude in the Age of Ockham*, Leiden: Brill.
- Weisheipl, J. (1982), "The Interpretation of Aristotle's *Physics* and the Science of Motion," in Kretzmann *et al.* 1982, 521-536.
- Wetherbee, Winthrop (1988), "Philosophy, Cosmology and the Renaissance," in *A History of Twelfth-Century Philosophy*, ed. Peter Dronke, Cambridge: Cambridge University Press, 21-53.
- Wolter, A. B., and M. M. Adams (1993), "Memory and Intuition: A Focal Debate in Fourteenth Century Cognitive Psychology," *Franciscan Studies* 53, 175-230.
- White, Graham (1990a), "Ockham and Hume's Question," in *Knowledge and the Sciences in Medieval Philosophy. Proceedings of the Eighth International Congress of the SIEPM*, ed. Simo Knuuttila, Reijo Työrinoja, and Sten Ebbesen, Helsinki: Yliopistopaino.
- White, Graham (1990b), "Ockham and Wittgenstein", in *Die Gegenwart Ockhams*, ed. W. Vossenkuhl & R. Schönberger, Weinheim, 165-188.
- White, Graham (1984), "Ockham's Real Distinction Between Form and Matter," *Franciscan Studies* 44, 211-25.
- Wood, Rega (1992), "Richard Rufus of Cornwall and Aristotle's *Physics*", *Franciscan Studies* 52, 247-281.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Buridan, John [Jean] | Duns Scotus, John | Marsilius of Inghen | Ockham [Occam], William | Olivi, Peter John | Oresme, Nicole

Graham White
graham@dcs.qmul.ac.uk

First published: August 10, 2001
Content last modified: August 10, 2001

Conditionals

Take a sentence in the indicative mood, suitable for making a statement: "We'll be home by ten", "Tom cooked the dinner". Attach a conditional clause to it, and you have a sentence which makes a conditional statement: "We'll be home by ten if the train is on time", "If Mary didn't cook the dinner, Tom cooked it". A conditional sentence "If *A*, *C*" or "*C* if *A*" thus has two contained sentences or sentence-like clauses. *A* is called the antecedent, *C* the consequent. If you understand *A* and *C*, and you have mastered the conditional construction (as we all do at an early age), you understand "If *A*, *C*". What does "if" mean? Consulting the dictionary yields "on condition that; provided that; supposing that". These are adequate synonyms. But we want more than synonyms. A theory of conditionals aims to give an account of the conditional construction which explains when conditional judgements are acceptable, which inferences involving conditionals are good inferences, and why this linguistic construction is so important. Despite intensive work of great ingenuity, this remains a highly controversial subject.

1. Introduction

First let us delimit our field. The examples with which we began are traditionally called "indicative conditionals". There are also "subjunctive" or "counterfactual" conditionals like "Tom would have cooked the dinner if Mary had not done so", "We would have been home by ten if the train had been on time". Counterfactuals are the subject of a separate entry, and theories addressing them will not be discussed here. That there is some difference between indicatives and counterfactuals is shown by pairs of examples like "If Oswald didn't kill Kennedy, someone else did" and "If Oswald hadn't killed Kennedy, someone else would have": you can accept the first yet reject the second (Adams (1970)). That

there is not a huge difference between them is shown by examples like the following: "Don't go in there", I say, "If you go in you will get hurt". You look sceptical but stay outside, when there is large crash as the roof collapses. "You see", I say, "if you had gone in you would have got hurt. I told you so."

It is controversial how best to classify conditionals. According to some theorists, the forward-looking "indicatives" (those with a "will" in the main clause) belong with the "subjunctives" (those with a "would" in the main clause), and not with the other "indicatives". (See Gibbard (1981, pp. 222-6), Dudman (1984, 1988), Bennett (1988). Bennett (1995) changed his mind. Jackson (1990) defends the traditional view.) The easy transition from typical "wills" to "woulds" is indeed a datum to be explained. Still, straightforward statements about the past, present or future, to which a conditional clause is attached -- the traditional class of indicative conditionals -- do (in my view) constitute a single semantic kind. The theories to be discussed do not fare better or worse when restricted to a particular subspecies.

As well as conditional statements, there are conditional commands, promises, offers, questions, etc.. As well as conditional beliefs, there are conditional desires, hopes, fears, etc.. Our focus will be on conditional statements and what they express -- conditional beliefs; but we will consider which of the theories we have examined extends most naturally to these other kinds of conditional.

Three kinds of theory will be discussed. In §2 we compare truth-functional and non-truth-functional accounts of the truth conditions of conditionals. In §3 we examine what I call the suppositional theory: that conditional judgements essentially involve suppositions. On development, it turns out to be incompatible with construing conditionals as statements with truth conditions. §4 looks at some responses from advocates of truth conditions. In §5 we consider a wider variety of conditional speech acts and propositional attitudes.

Where I need to distinguish between different interpretations, I write "$A \supset B$" for the truth-functional conditional, "$A \rightarrow B$" for a non-truth-functional conditional and "$A \Rightarrow B$" for the conditional as interpreted by the suppositional theory; and for brevity I call protagonists of the three theories Hook, Arrow and Supp, respectively. I use "~" for negation.

2. Truth Conditions for Indicative Conditionals

2.1 Two Kinds of Truth Condition

The generally most fruitful, and time-honoured, approach to specifying the meaning of a complex sentence in terms of the meanings of its parts, is to specify the truth conditions of the complex sentence, in terms of the truth conditions of its parts. A semantics of this kind yields an account of the validity of arguments involving the complex sentence, given the conception of validity as necessary preservation of truth. Throughout this section we assume that this approach to conditionals is correct. Let *A* and *B* be two sentences such as "Ann is in Paris" and "Bob is in Paris". Our question will be: are the truth conditions of "If *A*, *B*" of the simple, extensional, truth-functional kind, like those of "*A* and *B*", "*A* or *B*" and "It is not

the case that *A*"? That is, do the truth values of *A* and of *B* determine the truth value of "If *A*, *B*"? Or are they non-truth-functional, like those of "*A* because *B*", "*A* before *B*", "It is possible that *A*"? That is, are they such that the truth values of *A* and *B* may, in some cases, leave open the truth value of "If *A*, *B*"?

The truth-functional theory of the conditional was integral to Frege's new logic (1879). It was taken up enthusiastically by Russell (who called it "material implication"), Wittgenstein in the *Tractatus*, and the logical positivists, and it is now found in every logic text. It is the first theory of conditionals which students encounter. Typically, it does not strike students as *obviously* correct. It is logic's first surprise. Yet, as the textbooks testify, it does a creditable job in many circumstances. And it has many defenders. It is a strikingly simple theory: "If *A*, *B*" is false when *A* is true and *B* is false. In all other cases, "If *A*, *B*" is true. It is thus equivalent to "~(*A*&~*B*)" and to "~*A* or *B*". "$A \supset B$" has, by stipulation, these truth conditions.

If "if" is truth-functional, this is the right truth function to assign to it: of the sixteen possible truth-functions of *A* and *B*, it is the only serious candidate. First, it is uncontroversial that when *A* is true and *B* is false, "If *A*, *B*" is false. A basic rule of inference is modus ponens: from "If *A*, *B*" and *A*, we can infer *B*. If it were possible to have *A* true, *B* false and "If *A*, *B*" true, this inference would be invalid. Second, it is uncontroversial that "If *A*, *B*" is *sometimes* true when *A* and *B* are respectively (true, true), or (false, true), or (false, false). "If it's a square, it has four sides", said of an unseen geometric figure, is true, whether the figure is a square, a rectangle or a triangle. Assuming truth-functionality -- that the truth value of the conditional is *determined* by the truth values of its parts -- it follows that a conditional is *always* true when its components have these combinations of truth values.

Non-truth-functional accounts agree that "If *A*, *B*" is false when *A* is true and *B* is false; and they agree that the conditional is sometimes true for the other three combinations of truth-values for the components; but they deny that the conditional is always true in each of these three cases. Some agree with the truth-functionalist that when *A* and *B* are both true, "If *A*, *B*" must be true. Some do not, demanding a further relation between the facts that *A* and that *B* (see Read (1995)). This dispute need not concern us, as the arguments which follow depend only on the feature on which non-truth-functionalists agree: that when *A* is false, "If *A*, *B*" may be either true or false. For instance, I say (*) "If you touch that wire, you will get an electric shock". You don't touch it. Was my remark true or false? According to the non-truth-functionalist, it depends on whether the wire is live or dead, on whether you are insulated, and so forth. Robert Stalnaker's (1968) account is of this type: consider a possible situation in which you touch the wire, and which otherwise differs minimally from the actual situation. (*) is true (false) according to whether or not you get a shock in that possible situation.

Let *A* and *B* be two logically independent propositions. The four lines below represent the four incompatible logical possibilities for the truth values of *A* and *B*. "If *A*, *B*", "If ~*A*, *B*" and "If *A*, ~*B*" are interpreted truth-functionally in columns (i)-(iii), and non-truth-functionally (when their antecedents are false) in columns (iv)-(vi). The non-truth-functional interpretation we write "$A \rightarrow B$". "T/F" means both truth values are possible for the corresponding assignment of truth values to *A* and *B*. For instance, line 4, column (iv), represents two possibilities for *A*, *B*, If *A*, *B*, (F, F, T) and (F, F, F).

Truth-Functional Interpretation

			(i)	(ii)	(iii)
	A	*B*	$A \supset B$	$\sim A \supset B$	$A \supset \sim B$
1.	T	T	T	T	F
2.	T	F	F	T	T
3.	F	T	T	T	T
4.	F	F	T	F	T

Non-Truth-Functional Interpretation

			(iv)	(v)	(vi)
	A	*B*	$A \rightarrow B$	$\sim A \rightarrow B$	$A \rightarrow \sim B$
1.	T	T	T	T/F	F
2.	T	F	F	T/F	T
3.	F	T	T/F	T	T/F
4.	F	F	T/F	F	T/F

2.2 Arguments for Truth-Functionality

The main argument points to the fact that minimal knowledge that the truth-functional truth condition is satisfied is enough for knowledge that if *A*, *B*. Suppose there are two balls in a bag, labelled *x* and *y*. All you know about their colour is that at least one of them is red. That's enough to know that if *x* isn't red, *y* is red. Or: all you know is that they are not both red. That's enough to know that if *x* is red, *y* is not red.

Suppose you start off with no information about which of the four possible combinations of truth values for *A* and *B* obtains. You then acquire compelling reason to think that either *A* or *B* is true. You don't have any stronger belief about the matter. In particular, you have no firm belief as to whether *A* is true or not. You have ruled out line 4. The other possibilities remain open. Then, intuitively, you are justified in inferring that if ~*A*, *B*. Look at the possibilities for *A* and *B* on the left. You have eliminated the possibility that both *A* and *B* are false. So if *A* is false, only one possibility remains: *B* is true.

The truth-functionalist (call him Hook) gets this right. Look at column (ii). Eliminate line 4 and line 4 only, and you have eliminated the only possibility in which "$\sim A \supset B$" is false. You know enough to

conclude that "$\sim A \supset B$" is true.

The non-truth-functionalist (call her Arrow) gets this wrong. Look at column (v). Eliminate line 4 and line 4 only, and some possibility of falsity remains in other cases which have not been ruled out. By eliminating just line 4, you do not thereby eliminate these further possibilities, incompatible with line 4, in which "$\sim A \rightarrow B$" is false.

The same point can be made with negated conjunctions. You discover for sure that ~(*A*&*B*), but nothing stronger than that. In particular, you don't know whether *A*. You rule out line 1, nothing more. You may justifiably infer that if *A*, ~*B*. Hook gets this right. In column (iii), if we eliminate line 1, we are left only with cases in which "$A \supset \sim B$" is true. Arrow gets this wrong. In column (vi), eliminating line 1 leaves open the possibility that "$A \rightarrow \sim B$" is false.

The same argument renders compelling the thought that if we eliminate *just A*&~*B*, nothing stronger, i.e., we don't eliminate *A*, then we have sufficient reason to conclude that if *A*, *B*.

Here is a second argument in favour of Hook, in the style of Natural Deduction. The rule of Conditional Proof (CP) says that if *Z* follows from premises *X* and *Y*, then "If *Y*, *Z*" follows from premise *X*. Now the three premises ~(*A*&*B*), *A* and *B* entail a contradiction. So, by Reductio Ad Absurdum, from ~(*A*&*B*) and *A*, we can conclude ~*B*. So by CP, ~(*A*&*B*) entails "If *A*, ~*B*". Substitute "~*C*" for *B*, and we have a proof of "If *A*, then ~~*C*" from "~(*A*&~*C*)". And provided we also accept Double Negation Elimination, we can derive "If *A*, then *C*" from "~(*A*&~*C*)".

Conditional Proof seems sound: "From *X* and *Y*, it follows that *Z*. So from *X* it follows that if *Y*, *Z*". Yet *for no reading of "if" which is stronger than the truth-functional reading is CP valid* -- at least this is so if we treat "&" and "~" in the classical way and accept the validity of the inference: (I) ~(*A*&~*B*); *A*; therefore *B*. Suppose CP is valid for some interpretation of "If *A*, *B*". Apply CP to (I), and we get ~(*A*&~*B*); therefore if *A*, *B*, i.e., $A \supset B$ entails if *A*, *B*.

2.3 Arguments Against Truth-Functionality

The best-known objection to the truth-functional account, one of the "paradoxes of material implication", is that according to Hook, the falsity of *A* is sufficient for the truth of "If *A*, *B*". Look at the last two lines of column (i). In every possible situation in which *A* is false, "$A \supset B$" is true. Can it be right that the falsity of "She touched the wire" entails the truth of "If she touched the wire she got a shock"?

Hook might respond as follows. How do we test our intuitions about the validity of an inference? The direct way is to imagine that we know for sure that the premise is true, and to consider what we would then think about the conclusion. Now when we know for sure that ~*A*, we have no use for thoughts beginning "If *A*, ...". When you know for sure that Harry didn't do it, you don't go in for "If Harry did it ..." thoughts or remarks. In this circumstance conditionals have no role to play, and we have no practice in assessing them. The direct intuitive test is, therefore, silent on whether "If *A*, *B*" follows from ~*A*. If

our smoothest, simplest, generally satisfactory theory has the consequence that it does follow, perhaps we should learn to live with that consequence.

There may, of course, be further consequences of this feature of Hook's theory which jar with intuition. That needs investigating. But, Hook may add, even if we come to the conclusion that "$\supset$" does not match perfectly our natural-language "if", it comes close, and it has the virtues of simplicity and clarity. We have seen that rival theories also have counterintuitive consequences. Natural language is a fluid affair, and we cannot expect our theories to achieve better than approximate fit. Perhaps, in the interests of precision and clarity, in serious reasoning we should replace the elusive "if" with its neat, close relative, $\supset$.

This was no doubt Frege's attitude. Frege's primary concern was to construct a system of logic, formulated in an idealized language, which was adequate for mathematical reasoning. If "$A \supset B$" doesn't translate perfectly our natural-language "If *A*, *B*", but plays its intended role, so much the worse for natural language.

For the purpose of doing mathematics, Frege's judgement was probably correct. The main defects of $\supset$ don't show up in mathematics. There are some peculiarities, but as long as we are aware of them, they can be lived with. And arguably, the gain in simplicity and clarity more than offsets the oddities.

The oddities are harder to tolerate when we consider conditional judgements about empirical matters. The difference is this: in thinking about the empirical world, we often accept and reject propositions with degrees of confidence less than certainty. "I think, but am not sure, that *A*" plays no central role in mathematical thinking. We can, perhaps, ignore as unimportant the use of indicative conditionals in circumstances in which we are *certain* that the antecedent is false. But we cannot ignore our use of conditionals whose antecedent we think is likely to be false. We use them often, accepting some, rejecting others. "I think I won't need to get in touch, but if I do, I shall need a phone number", you say as your partner is about to go away; not "If I do I'll manage by telepathy". "I think John spoke to Mary; if he didn't he wrote to her"; not "If he didn't he shot her". Hook's theory has the unhappy consequence that *all* conditionals with unlikely antecedents are likely to be true. To think it likely that $\sim A$ is to think it likely that a sufficient condition for the truth of "$A \supset B$" obtains. Take someone who thinks that the Republicans won't win the election ($\sim R$), and who rejects the thought that if they do win, they will double income tax (*D*). According to Hook, this person has grossly inconsistent opinions. For if she thinks it's likely that $\sim R$, she must think it likely that at least one of the propositions, $\{\sim R, D\}$ is true. But that is just to think it likely that $R \supset D$. (Put the other way round, to reject $R \supset D$ is to accept $R\&\sim D$; for this is the only case in which $R \supset D$ is false. How can someone accept $R\&\sim D$ yet reject *R*?) Not only does Hook's theory fit badly the patterns of thought of competent, intelligent people. It cannot be claimed that we would be better off with $\supset$. On the contrary, we would be intellectually disabled: we would not have the power to discriminate between believable and unbelievable conditionals whose antecedent we think is likely to be false.

Arrow does not have this problem. Her theory is designed to avoid it, by allowing that "$A \rightarrow B$" may be

false when *A* is false.

The other paradox of material implication is that according to Hook all conditionals with true consequents are true: from *B* it follows that $A \supset B$. This is perhaps less obviously unacceptable: if I'm sure that *B*, and treat *A* as an epistemic possibility, I must be sure that if *A*, *B*. Again the problem becomes vivid when we consider the case when I'm only nearly sure, but not quite sure, that *B*. I think *B* *may* be false, and will be false if certain, in my view unlikely, circumstances obtain. For example, I think Sue is giving a lecture right now. I don't think that if she was seriously injured on her way to work, she is giving a lecture right now. I reject that conditional. But on Hook's account, the conditional is false only if the consequent is false. I think the consequent is true: I think a sufficient condition for the truth of the conditional obtains.

2.4 Grice's Pragmatic Defence of Truth-Functionality

H. P. Grice famously defended the truth-functional account, in his William James lectures, "Logic and Conversation", delivered in 1967 (see Grice (1989); see also Thomson (1990)). There are many ways of speaking the truth yet misleading your audience, given the standard to which you are expected to conform in conversational exchange. One way is to say something weaker than some other relevant thing you are in a position to say. Consider disjunctions. I am asked where John is. I am sure that he is in the pub, and know that he never goes near libraries. Inclined to be unhelpful but not wishing to lie, I say "He is either in the pub or in the library". My hearer naturally assumes that this is the most precise information I am in a position to give, and also concludes from the truth (let us assume) that I told him "If he's not in the pub he's in the library". The conditional, like the disjunction, according to Grice, is true if he's in the pub, but misleadingly asserted on that ground.

Another example, from David Lewis (1976, p. 143): "You won't eat those and live", I say of some wholesome and delicious mushrooms -- knowing that you will now leave them alone, deferring to my expertise. I told no lie -- for indeed you don't eat them -- but of course I misled you.

Grice drew attention, then, to situations in which a person is *justified in believing* a proposition, which would nevertheless be an unreasonable thing for the person to *say*, in normal circumstances. His lesson was salutary and important. He is right, I think, about disjunctions and negated conjunctions. Believing that John is in the pub, I can't consistently *disbelieve* "He's either in the pub or the library"; if I have any epistemic attitude to this proposition, it should be one of belief, however inappropriate for me to assert it. Similarly for "You won't eat those and live" when I know you won't eat them. But it is implausible that the difficulties with the truth-functional conditional can be explained away in terms of what is an inappropriate conversational remark. They arise at the level of belief. Thinking that John is in the pub, I may without irrationality disbelieve "If he's not in the pub he's in the library". Thinking you won't eat the mushrooms, I may without irrationality reject "If you eat them you will die". As facts about the norms to which people defer, these claims can be tested. A good enough test is to take a co-operative person, who understands that you are merely interested in her opinions about the propositions you put to her, as opposed to what would be a reasonable remark to make, and note which conditionals she assents

to. Are we really to brand as illogical someone who dissents from both "The Republicans will win" and "If the Republicans win, income tax will double"?

The Gricean phenomenon is a real one. On anyone's account of conditionals, there will be circumstances when a conditional is justifiably believed, but is liable to mislead if stated. For instance, I believe that the match will be cancelled, because all the players have 'flu. I believe that whether or not it rains, the match will be cancelled: if it rains, the match will be cancelled, and if it doesn't rain, the match will be cancelled. Someone asks me whether the match will go ahead. I say, "If it rains, the match will be cancelled". I say something I believe, but I mislead my audience -- why should I say that, when I think it will be cancelled whether or not it rains? This does not demonstrate that Hook is correct. Although I believe that the match will be cancelled, I don't believe that if all the players make a very speedy recovery, the match will be cancelled.

2.5 Compounds of Conditionals: Problems for Hook and Arrow

$\sim(A \supset B)$ is equivalent to $A\&\sim B$. Intuitively, you may safely say, of an unseen geometric figure, "It's not the case that if it's a pentagon, it has six sides". But by Hook's lights, you may well be wrong; for it may not be a pentagon, and in that case it is true that if it's a pentagon, it has six sides.

Another example, due to Gibbard (1981, pp. 235-6): of a glass that had been held a foot above the floor, you say (having left the scene) "If it broke if it was dropped, it was fragile". Intuitively this seems reasonable. But by Hook's lights, if the glass was not dropped, and was not fragile, the conditional has a true (conditional) antecedent and false consequent, and is hence false.

Grice's strategy was to explain why we don't assert certain conditionals which (by Hook's lights) we have reason to believe true. In the above two cases, the problem is reversed: there are compounds of conditionals which we confidently assert and accept which, by Hook's lights, we do not have reason to believe true.

The above examples are not a problem for Arrow. But other cases of embedded conditionals count in the opposite direction. Here are two sentence forms which are, intuitively, equivalent:

> (i) If ($A\&B$), C.
> (ii) If A, then if B, C.

(Following Vann McGee (1985) I'll call the principle that (i) and (ii) are equivalent the Import-Export Principle, or "Import-Export" for short.) Try any example: "If Mary comes then if John doesn't have to leave early we will play Bridge"; "If Mary comes and John doesn't have to leave early we will play Bridge". "If they were outside and it rained, they got wet"; "If they were outside, then if it rained, they got wet". For Hook, Import-Export holds. (Exercise: do a truth table, or construct a proof.) Gibbard (1981, pp. 234-5) has proved that for no conditional with truth conditions stronger than $\supset$ does Import-Export hold. Assume Import-Export holds for some reading of "if". The key to the proof is to consider

the formula

> (1) If ($A \supset B$) then (if A, B).

By Import-Export, (1) is equivalent to

> (2) If (($A \supset B$) & A)) then B.

The antecedent of (2) entails its consequent. So (2) is a logical truth. So by Import-Export, (1) is a logical truth. On any reading of "if", "if A, B" entails ($A \supset B$). So (1) entails

> (3) ($A \supset B$) $\supset$ (if A, B).

So (3) is a logical truth. That is, there is no possible situation in which its antecedent ($A \supset B$) is true and its consequent (if A, B) is false. That is, ($A \supset B$) entails "If A, B".

Neither kind of truth condition has proved entirely satisfactory. We still have to consider Jackson's defence of Hook, and Stalnaker's response to the problem about non-truth-functional truth conditions raised in §2.2. These are deferred to §4, because they depend on the considerations developed in §3.

3. The Suppositional Theory

3.1 Conditional Belief and Conditional Probability

Let us put truth conditions aside for a while, and ask what it is to believe, or to be more or less certain, that B if A -- that John cooked the dinner if Mary didn't, that you will recover if you have the operation, and so forth. How do you make such a judgement? You suppose (assume, hypothesise) that A, and make a hypothetical judgement about B, under the supposition that A, in the light of your other beliefs. Frank Ramsey put it like this:

> If two people are arguing "If *p*, will *q*?" and are both in doubt as to *p*, they are adding *p* hypothetically to their stock of knowledge, and arguing on that basis about *q*; ... they are fixing their degrees of belief in *q* given *p* (1929, p. 247).

A suppositional theory was advanced by J. L. Mackie (1973, chapter 4). Peter Gärdenfors's work (1986, 1988) could also come under this heading. But the most fruitful development of the idea (in my view) takes seriously the last part of the above quote from Ramsey, and emphasises the fact that conditionals can be accepted with different degrees of closeness to certainty. Ernest Adams (1965, 1966, 1975) has developed such a theory.

When we are neither certain that *B* nor certain that ~*B*, there remains a range of epistemic attitudes we may have to *B*: we may be nearly certain that *B*, think *B* more likely than not, etc.. Similarly, we may be certain, nearly certain, etc. that *B* given the supposition that *A*. Make the idealizing assumption that degrees of closeness to certainty can be quantified: 100% certain, 90% certain, etc.; and we can turn to probability theory for what Ramsey called the "logic of partial belief". There we find a well-established, indispensable concept, "the conditional probability of *B* given *A*". It is to this notion that Ramsey refers by the phrase "degrees of belief in *q* given *p*".

It is, at first sight, rather curious that the best-developed and most illuminating suppositional theory should place emphasis on uncertain conditional judgements. If we knew the truth conditions of conditionals, we would handle uncertainty about conditionals in terms of a general theory of what it is to be uncertain of the truth of a proposition. But there is no consensus about the truth conditions of conditionals. It happens that when we turn to the theory of uncertain judgements, we find a concept of conditionality in use. It is worth seeing what we can learn from it.

The notion of conditional probability entered probability theory at an early stage because it was needed to compute the probability of a conjunction. Thomas Bayes (1763) wrote:

> The probability that two ... events will both happen is ... the probability of the first [multiplied by] the probability of the second *on the supposition that* the first happens [my emphasis].

A simple example: a ball is picked at random. 70% of the balls are red (so the probability that a red ball is picked is 70%). 60% of the red balls have a black spot (so the probability that a ball with a black spot is picked, on the supposition that a red ball is picked, is 60%). The probability that a red ball with a black spot is picked is 60% of 70%, i.e. 42%.

Ramsey, arguing that "degrees of belief" should conform to probability theory, stated the same "fundamental law of partial belief":

> Degree of belief in (*p* and *q*) = degree of belief in *p* × degree of belief in *q* given *p*. (1926, p. 77)

For example, you are about 50% certain that the test will be on conditionals, and about 80% certain that you will pass, on the supposition that it is on conditionals. So you are about 40% certain that the test will be on conditionals and you will pass.

Accepting Ramsey's suggestion that "if", "given that", "on the supposition that" come to the same thing, writing "$\mathbf{p}(B)$" for "degree of belief in *B*", and "$\mathbf{p}_A(B)$" for "degree of belief in *B* given *A*", and rearranging the basic law, we have:

> $\mathbf{p}(B \text{ if } A) = \mathbf{p}_A(B) = \mathbf{p}(A\&B)/\mathbf{p}(A)$, provided $\mathbf{p}(A)$ is not 0.

Call a set of mutually exclusive and jointly exhaustive propositions a partition. The lines of a truth table constitute a partition. One’s degrees of belief in the members of a partition, idealized as precise, should sum to 100%. That is all there is to the claim that degrees of belief should have the structure of probabilities. Consider a partition of the form {*A*&*B*, *A*&~*B*, ~*A*}. Suppose someone X thinks it 50% likely that ~*A* (hence 50% likely that *A*), 40% likely that *A*&*B*, and 10% likely that *A*&~*B*. Think of this distribution as displayed geometrically, as follows. Draw a long narrow horizontal rectangle. Divide it in half by a vertical line. Write "~*A*" in the right-hand half. Divide the left-hand half with another vertical line, in the ratio 4:1, with the larger part on the left. Write "*A*&*B*" and "*A*&~*B*" in the larger and smaller cells respectively.

A&*B*	*A*&~*B*	~*A*

(Note that as {*A*&*B*, *A*&~*B*, ~*A*} and {*A*, ~*A*} are both partitions, it follows that **p**(*A*) = **p**(*A*&*B*) + **p**(*A*&~*B*).)

How does X evaluate "If *A*, *B*"? She assumes that *A*, that is, hypothetically eliminates ~*A*. In the part of the partition that remains, in which *A* is true, *B* is four times as likely as ~*B*; that is, on the assumption that *A*, it is four to one that *B*: **p**(*B* if *A*) is 80%, **p**(~*B* if *A*) is 20%. Equivalently, as *A*&*B* is four times as likely as *A*&~*B*, **p**(*B* if *A*) is 4/5, or 80%. Equivalently, **p**(*A*&*B*) is 4/5 of **p**(*A*). In non-numerical terms: you believe that if *A*, *B* to the extent that you think that *A*&*B* is nearly as likely as *A*; or, to the extent that you think *A*&*B* is much more likely than *A*&~*B*. If you think *A*&*B* is as likely as *A*, you are certain that if *A*, *B*. In this case, your **p**(*A*&~*B*) = 0.

Go back to the truth table. You are wondering whether if *A*, *B*. Assume *A*. That is, ignore lines 3 and 4 in which *A* is false. Ask yourself about the relative probabilities of lines 1 and 2. Suppose you think line 1 is about 100 times more likely than line 2. Then you think it is about 100 to 1 that *B* if *A*.

Note: these thought-experiments can only be performed when **p**(*A*) is not 0. On this approach, indicative conditionals only have a role when the thinker takes *A* to be an epistemic possibility. If you take yourself to know for sure that Ann is in Paris, you don’t go in for "If Ann is not in Paris ..." thoughts (though of course you can think "If Ann had not been in Paris ..."). In conversation, you can pretend to take something as an epistemic possibility, temporarily, to comply with the epistemic state of the hearer. When playing the sceptic, there are not many limits on what you *can*, at a pinch, take as an epistemic possibility -- as not already ruled out. But there are some limits, as Descartes found. Is there a conditional thought that begins "If I don’t exist now ..."?

On Hook’s account, to be close to certain that if *A*, *B* is to give a high value to **p**(*A* ⊃ *B*). How does **p**(*A* ⊃ *B*) compare with $\mathbf{p}_A(B)$? In two special cases, they are equal: first, if **p**(*A*&~*B*) = 0 (and **p**(*A*) is not 0), **p**(*A* ⊃ *B*) = $\mathbf{p}_A(B)$ = 1 (i.e. 100%). Second, if **p**(*A*) = 100%, **p**(*A* ⊃ *B*) = $\mathbf{p}_A(B)$ = **p**(*B*). In all other cases, **p**(*A* ⊃ *B*) is greater than $\mathbf{p}_A(B)$. To see this we need to compare **p**(*A*&~*B*) and **p**(*A*&~*B*)/**p**(*A*). Consider

again the partition {$A\&B$, $A\&{\sim}B$, ${\sim}A$}. $\mathbf{p}(A\&{\sim}B)$ is a smaller proportion of the whole space than it is of the A-part -- the part of the space in which A is true -- except in the special cases in which $\mathbf{p}(A\&{\sim}B) = 0$, or $\mathbf{p}({\sim}A) = 0$. So, except in these special cases, $\mathbf{p}_A({\sim}B)$ is greater than $\mathbf{p}(A\&{\sim}B)$. Now $\mathbf{p}(A \supset B) = \mathbf{p}({\sim}(A\&{\sim}B))$; and $\mathbf{p}(A\&{\sim}B) + \mathbf{p}({\sim}(A\&{\sim}B)) = 1$. Also $\mathbf{p}_A(B) + \mathbf{p}_A({\sim}B) = 1$. So from $\mathbf{p}_A({\sim}B) > \mathbf{p}(A\&{\sim}B)$ it follows that $\mathbf{p}(A \supset B) > \mathbf{p}_A(B)$.

Hook and the suppositional theorist (call her Supp) come spectacularly apart when $\mathbf{p}({\sim}A)$ is high and $\mathbf{p}(A\&B)$ is much smaller than $\mathbf{p}(A\&{\sim}B)$. Let $\mathbf{p}({\sim}A) = 90\%$, $\mathbf{p}(A\&B) = 1\%$, $\mathbf{p}(A\&{\sim}B) = 9\%$. $\mathbf{p}_A(B) = 10\%$. $\mathbf{p}(A \supset B) = 91\%$. For instance, I am 90% certain that Sue won't be offered the job (${\sim}O$), and think it only 10% likely that she will decline the offer (D) if it is made, that is $\mathbf{p}_O(D) = 10\%$. $\mathbf{p}(O \supset D) = \mathbf{p}({\sim}O$ or $(O\&D)) = 91\%$.

Now let us compare Hook, Arrow, and Supp with respect to two questions raised in §2.

- Question 1. You are certain that ${\sim}(A\&{\sim}B)$, but not certain that ${\sim}A$. Should you be certain that if A, B?

  Hook: yes. Because "$A \supset B$" is true whenever $A\&{\sim}B$ is false.

  Supp: yes. Because $A\&B$ is as likely as A. $\mathbf{p}_A(B) = 1$.

  Arrow: no, not necessarily. For "$A \rightarrow B$" may be false when $A\&{\sim}B$ is false. With just the information that $A\&{\sim}B$ is false, I should not be certain that if A, B.
- Question 2. If you think it likely that ${\sim}A$, might you still think it unlikely that if A, B?

  Hook: no. "$A \supset B$" is true in all the possible situations in which ${\sim}A$ is true. If I think it likely that ${\sim}A$, I think it likely that a sufficient condition for the truth of "$A \supset B$" obtains. I must, therefore, think it likely that if A, B.

  Supp: yes. We had an example above. That most of my probability goes to ${\sim}A$ leaves open the question whether or not $A\&B$ is more probable than $A\&{\sim}B$. If $\mathbf{p}(A\&{\sim}B)$ is greater than $\mathbf{p}(A\&B)$, I think it's unlikely that if A, B. That's compatible with thinking it likely that ${\sim}A$.

  Arrow: yes. "If A, B" may be false when A is false. And I might well think it likely that that possibility obtains, i.e. unlikely that "If A, B" is true.

Supp has squared the circle: she gets the intuitively right answer to both questions. In this she differs from both Hook and Arrow. Supp's way of assessing conditionals is incompatible with the truth-functional way (they answer Question 2 differently); and incompatible with stronger-than-truth-functional truth conditions (they answer Question 1 differently). It follows that Supp's way of assessing conditionals is incompatible with the claim that conditionals have truth conditions of any kind. $\mathbf{p}_A(B)$

does not measure the probability of the truth of any proposition. Suppose it did measure the probability of the truth of some proposition *A***B*. Either *A***B* is entailed by "*A* ⊃ *B*", or it is not. If it is, it is true whenever ~*A* is true, and hence cannot be improbable when ~*A* is probable. That is, it cannot agree with Supp in its answer to Question 2. If *A***B* is not entailed by "If *A*, *B*", it may be false when ~(*A*&~*B*) is true, and hence certainty that ~(*A*&~*B*) (in the absence of certainty that ~*A*) is insufficient for certainty that *A***B*; it cannot agree with Supp in its answer to Question 1.

To make the point in a slightly different way, let me adopt the following as an expository, heuristic device, a harmless fiction. Imagine a partition as carved into a large finite number of equally-probable chunks, such that the propositions with which we are concerned are true in an exact number of them. The probability of any proposition is the proportion of chunks in which it is true. The probability of *B* on the supposition that *A* is the proportion *of the A-chunks* (the chunks in which *A* is true) which are *B*-chunks. With some misgivings, I succumb to the temptation to call these chunks "worlds": they are equally probable, mutually incompatible and jointly exhaustive epistemic possibilities, enough of them for the propositions with which we are concerned to be true, or false, at each world. The heuristic value is that judgements of probability and conditional probability then translate into statements about proportions.

Although Supp and Hook give the same answer to Question 1, their reasons are different. Supp answers "yes" *not* because a proposition, *A***B*, is true whenever *A*&~*B* is false; but because *B* is true in all the "worlds" which matter for the assessment of "If *A*, *B*": the *A*-worlds. Although Supp and Arrow give the same answer to Question 2, their reasons are different. Supp answers "yes", not because a proposition *A***B* may be false when *A* is false; but because the fact that most worlds are ~*A*-worlds is irrelevant to whether most *of the A-worlds* are *B*-worlds. To judge that *B* is true *on the supposition that A* is true, it turns out, is not to judge that something-or-other, *A***B*, is true.

By a different argument, David Lewis (1976) was the first to prove this remarkable result: there is no proposition *A***B* such that, in all probability distributions, $\mathbf{p}(A{*}B) = \mathbf{p}_A(B)$. A conditional probability does not measure the probability of the truth of any proposition. If a conditional has truth conditions, one should believe it to the extent that one thinks it is probably true. If Supp is correct, that one believes "If *A*, *B*" to the extent that one thinks it probable that *B* on the supposition that *A*, then this is not equivalent to believing some proposition to be probably true. Hence, if Supp is right, conditionals shouldn't be construed as having truth conditions at all. A conditional judgement involves two propositions, which play different roles. One is the content of a supposition. The other is the content of a judgement made under that supposition. They do not combine to yield a single proposition which is judged to be likely to be true just when the second is judged likely to be true on the supposition of the first.

Note: ways of restoring truth conditions, compatible with Supp's thesis, are considered in §4.

3.2 Validity

Ernest Adams, in two articles (1965, 1966) and a subsequent book (1975), gave a theory of the validity of arguments involving conditionals as construed by Supp. He taught us something important about

classically valid arguments as well: that they are, in a special sense to be made precise, probability-preserving. This property can be generalized to apply to arguments with conditionals. The valid ones are those which, in the special sense, preserve probability or conditional probability.

First consider classically valid (that is, necessarily truth-preserving) arguments which don't involve conditionals. We use them in arguing from contingent premises about which we are often less than completely certain. The question arises: how certain can we be of the conclusion of the argument, given that we think, but are not sure, that the premises are true? Call the improbability of a statement one minus its probability. Adams showed this: if (and only if) an argument is valid, then in no probability distribution does the improbability of its conclusion exceed the sum of the improbabilities of its premises. Call this the Probability Preservation Principle (PPP).

The proof of PPP rests on the Partition Principle -- that the probabilities of the members of a partition sum to 100% -- nothing else, beyond the fact that if *A* entails *B*, $\mathbf{p}(A\&{\sim}B) = 0$. Here are three consequences:

1. if *A* entails *B*, $\mathbf{p}(A) \leq \mathbf{p}(B)$
2. $\mathbf{p}(A \text{ or } B) = \mathbf{p}(A) + \mathbf{p}(B) - \mathbf{p}(A\&B) \leq \mathbf{p}(A) + \mathbf{p}(B)$
3. For all *n*, $\mathbf{p}(A_1 \text{ or ... or } A_n) \leq \mathbf{p}(A_1) + ... + \mathbf{p}(A_n)$

Suppose A_1, ... A_n entail *B*. Then ~*B* entails ${\sim}A_1$ or ... or ${\sim}A_n$. Therefore $\mathbf{p}({\sim}B) \leq \mathbf{p}({\sim}A_1) + ... + \mathbf{p}({\sim}A_n)$: the improbability of the conclusion of a valid argument cannot exceed the sum of the improbabilities of the premises.

The result is useful to know: if you have two premises of which you are at least 99% certain, they entitle you to be at least 98% certain of a conclusion validly drawn from them. Of course, if you have 100 premises each at least 99% certain, your conclusion may have zero probability. That is the lesson of the "Lottery Paradox". Still, Adams's result vindicates deductive reasoning from uncertain premises, provided that they are not too uncertain, and there are not too many of them.

So far, we have a very useful consequence of the classical notion of validity. Now Adams extends this consequence to arguments involving conditionals. Take a language with "and", "or", "not" and "if" -- but with "if" occurring only as the main connective in a sentence. (We put aside compounds of conditionals.) Take any argument formulated in this language. Consider any probability function over the sentences of this argument which assigns non-zero probability to the antecedents of all conditionals -- that is, any assignment of numbers to the non-conditional sentences which conforms to the Partition Principle, and to the conditional sentences which conforms to Supp's thesis: $\mathbf{p}(B \text{ if } A) = \mathbf{p}_A(B) = \mathbf{p}(A\&B)/\mathbf{p}(A)$. Let the improbability of the conditional "If *A*, *B*" be $1 - \mathbf{p}_A(B)$. *Define* a valid argument as one such that there is no probability function in which the improbability of the conclusion exceeds the sum of the improbabilities of the premises. And a nice logic emerges, which is now well known. It is the same as Stalnaker's logic over this domain (see §4.1). There are rules of proof, a decision procedure, consistency and completeness can be proved. See Adams (1998 and 1975).

I shall write the conditional which satisfies Adams's criterion of validity "$A \Rightarrow B$". We have already seen that in all distributions, $\mathbf{p}_A(B) \leq \mathbf{p}(A \supset B)$. Therefore, $A \Rightarrow B$ entails $A \supset B$: the former cannot be probable and the latter improbable. Call a non-conditional sentence a factual sentence. If an argument has a factual conclusion, and is classically valid with the conditional interpreted as $\supset$, it is valid with the conditional interpreted as the stronger $\Rightarrow$. The following patterns of inference are therefore valid:

> *A*; $A \Rightarrow B$; so *B* (modus ponens)
> $A \Rightarrow B$; ~*B*; so ~*A* (modus tollens)
> *A* or *B*; $A \Rightarrow C$; $B \Rightarrow C$; so *C*.

We cannot consistently have their premises highly probable and their conclusion highly improbable.

Arguments with conditional conclusions, however, may be valid when the conditional is interpreted as the weaker $A \supset B$, but invalid when it is interpreted as the stronger $A \Rightarrow B$. Here are some examples.

> *B*; so $A \Rightarrow B$.

I can consistently be close to certain that Sue is lecturing right now, while thinking it highly unlikely that if she had a heart attack on her way to work, she is lecturing just now.

> ~*A*; so $A \Rightarrow B$.

You can consistently be close to certain that the Republicans won't win, while thinking it highly unlikely that if they win they will double income tax.

> ~(*A*&*B*); so $A \Rightarrow$ ~*B*

I can consistently be close to certain that it's not the case that I will be hit by a bomb and injured today, while thinking it highly unlikely that if I am hit by a bomb, I won't be injured.

> *A* or *B*; so ~$A \Rightarrow B$.

As I think it is very likely to rain tomorrow, I think it's very likely to be true that it will rain or snow tomorrow. But I think it's very unlikely that if it doesn't rain, it will snow.

> $A \Rightarrow B$; so $(C\&A) \Rightarrow B$ (strengthening of the antecedent).

I can think it's highly likely that if you strike the match, it will light; but highly unlikely that if you dip it in water and strike it, it will light.

Strengthening is a special case of transitivity, in which the missing premise is a tautology: if *C*&*A* then *A*; if *A*, *B*; so if *C*&*A*, *B*. So transitivity also fails:

$A \Rightarrow B$; $B \Rightarrow C$; so $A \Rightarrow C$.

Adams gave this example (1966): I can think it highly likely that if Jones is elected, Brown will resign immediately afterwards; I can also think it highly likely that if Brown dies before the election, Jones will be elected; but I do not think it at all likely that if Brown dies before the election, Brown will resign immediately after the election!

We saw in §2.2 that Conditional Proof (CP) is invalid for any conditional stronger than $\supset$. It is invalid in Adams's logic. For instance, "~(*A*&*B*); *A*; so ~*B*" is valid. It contains no conditionals. Any necessarily truth-preserving argument satisfies PPP. If I'm close to certain that I won't be hit by a bomb and injured, *and close to certain that I will be hit by a bomb*, then I must be close to certain that I won't be injured. But, as we saw, "~(*A*&*B*); so $A \Rightarrow$ ~*B*" is invalid. Yet we can get the latter from the former by CP.

Why does CP fail on this conception of conditionals? After all, Supp's idea is to treat the antecedent of a conditional as an *assumption*. What is the difference between the roles of a premise, and of the antecedent of a conditional in the conclusion?

The antecedent of the conditional is indeed treated as an assumption. On this conception of validity, the premises are not treated as assumptions. Indeed, it is not immediately clear what it would be to treat a conditional, construed according to Supp, as an assumption: to assume something, as ordinarily understood, is to assume that it is true; and conditionals are not being construed as ordinary statement of fact. But we could approximate the idea of taking the premises as assumptions: so doing is, in most contexts, tantamount to treating them, hypothetically, as certainties. So treating the premises would be to require of a valid argument that it preserve certainty: that there must be no probability distributions in which all the premises (conditional or otherwise) are assigned 1 and the conclusion is assigned less than 1. Call this the certainty-preservation principle (CPP).

The conception of validity we have been using (PPP) takes as central the fact that premises are accepted with degrees of confidence less than certainty. Now, anything which satisfies PPP satisfies CPP. And for argument involving only factual propositions, the converse is also true: the same class of arguments necessarily preserves truth, necessarily preserves certainty and necessarily preserves probability in the sense of PPP. But arguments involving conditionals can satisfy CPP without satisfying PPP. The invalid argument forms above do preserve certainty: if you assign probability 1 to the premises, then you are constrained to assign probability 1 to the conclusion (in all probability distributions in which the antecedent of any conditional gets non-zero probability). But they do not preserve high probability. They do not satisfy PPP. If at least one premise falls short of certainty by however small an amount, the conclusion can plummet to zero.

The logico-mathematical fact behind this is the difference in logical powers between "All" and "Almost

all". If all *A*-worlds are *B*-worlds (and there are some *C*&*A*-worlds) then all *C*&*A*-worlds are *B*-worlds. But we can have: almost all *A*-worlds are *B*-worlds but no *C*&*A*-world is a *B*-world. If all *A*-worlds are *B*-worlds and all *B*-worlds are *C*-worlds, then all *A*-worlds are *C*-worlds. But we can have: all *A*-worlds are *B*-worlds, almost all *B*-worlds are *C*-worlds, yet no *A*-world is a *C*-world; just as we can have, all kiwis are birds, almost all birds fly, but no kiwi flies.

Someone might react as follows: "All I want of a valid argument is that it preserve certainty. I'm not bothered if an argument can have premises close to certain and a conclusion far from certain, as long as the conclusion is certain when the premises are certain".

We *could* use the word "valid" in such a way that an argument is valid provided it preserves certainty. If our interest in logic is confined to its application to mathematics or other a priori matters, that is fine. Further, when our arguments do not contain conditionals, if we have certainty-preservation, probability-preservation comes free. But if we use conditionals when arguing about contingent matters, then great caution will be required. Unless we are 100% certain of the premises, the arguments above which are invalid on Adams's criterion guarantee nothing about what you are entitled to think about the conclusion. The line between 100% certainty and something very close is hard to make out: it's not clear how you tell which side of it you are on. The epistemically cautious might admit that they are never, or only very rarely, 100% certain of contingent conditionals. So it would be useful to have another category of argument, the "super-valid", which preserves high probability as well as certainty. Adams has shown us which arguments (on Supp's reading of "if") are super-valid.

4. Truth Conditions Revisited: Stalnaker and Jackson

4.1 Stalnaker

Adams's theory of validity emerged in the mid-1960s. "Nearest possible worlds" theories were not yet in evidence. Nor was Lewis's result that conditional probabilities are not probabilities of the truth of a proposition. (Adams expressed scepticism about truth conditions for conditionals, but the question was still open.) Stalnaker's (1968) semantics for conditionals was an attempt to provide truth conditions which were compatible with Ramsey's and Adams's thesis about conditional belief. (See also Stalnaker (1970)). That is, he sought truth conditions for a proposition *A*>*B* (his notation) such that $\mathbf{p}(A>B)$ must equal $\mathbf{p}_A(B)$:

> Now that we have found an answer to the question, "How do we decide whether or not we believe a conditional statement?" [Ramsey's and Adams's answer] the problem is to make the transition from belief conditions to truth conditions; ... . The concept of a *possible world* is just what we need to make the transition, since a possible world is the ontological analogue of a stock of hypothetical beliefs. The following ... is a first approximation to the

> account I shall propose: Consider a possible world in which *A* is true and otherwise differs minimally from the actual world. *"If A, then B" is true (false) just in case B is true (false) in that possible world*. (1968, pp. 33-4)

If an argument is necessarily truth-preserving, the improbability of its conclusion cannot exceed the sum of the improbabilities of the premises. The latter was the criterion Adams used in constructing his logic. So Stalnaker’s logic for conditionals must agree with Adams’s over their common domain. And it does. The argument forms we showed to be invalid in Adams’s logic (§3.2) are invalid on Stalnaker’s semantics. For instance, the following is possible: in the nearest possible world in which you strike the match, it lights; in the nearest world in which you dip the match in water and strike it, it doesn’t light. So Strengthening fails. (By "nearest world in which ..." I mean the possible world which is minimally different from the actual world in which ... .)

Conditional Proof fails for Stalnaker’s semantics. "*A* or *B*; ~*A*; so *B*" is of course valid. But (*) "*A* or *B*, therefore ~*A*>*B*" is not: it can be true that Ann or Mary cooked the dinner (for Ann cooked it); yet false that in the nearest world to the actual world in which Ann did not cook it, Mary cooked it.

Stalnaker (1975) tried to show that although the above argument form (*) is invalid, it is nevertheless a "reasonable inference" when "*A* or *B*" is assertable, that is, when the speaker has ruled out ~*A*&~*B*, but ~*A*&*B* and *A*&~*B* remain open possibilities. Indicative conditionals, he claims, are used only when their antecedents are epistemically possible for the speaker (here he agrees with Supp). Then comes the crucial claim: *worlds which are epistemically possible for the speaker count as closer to the actual world than those which are not*. All ~*A*&~*B*-worlds have been eliminated. Not all ~*A*&*B*-worlds have been eliminated. All the speaker’s epistemically possible ~*A*-worlds are *B*-worlds. So the nearest ~*A*-world is a *B*-world. "*A*>*B*" is true.

This makes the truth conditions of a conditional, e.g. "If Ann didn’t cook the dinner, Bob cooked it" *dependent on what the speaker believes*. All that is common to different utterances of "*A*>*B*" is that they say that a certain *A*-world is a *B*-world. That is not news: provided that *A* and *B* are compatible, some *A*-world is a *B*-world. Which world is being said to be a *B*-world depends on the speaker’s beliefs. With fixed meanings for *A* and *B*, there is no single proposition *A*>*B*, but a different one for each belief state: we might write $A>_{p}B$, where "p" is a probability function indexed to a person and a time.

This enables Stalnaker to avoid the argument against non-truth-functional truth conditions given in §2.2. The argument was as follows. There are six incompatible logically possible combinations of truth values for *A*, *B* and ~*A* → *B*. You start off with no firm beliefs about which obtains. Now you eliminate ~*A*&~*B*, i.e. establish *A* or *B*. That leaves five remaining possibilities, including two in which "~*A* → *B*" is false. So you can’t be certain that ~*A* → *B*. Stalnaker replies: you can’t, indeed, be certain that the proposition you were wondering about earlier is true. But in your new epistemic state, you express a new proposition by "~*A* → *B*", with different truth conditions, governed by a new nearness relation, and you know that that new proposition is true.

Disagreement and change of mind give way to equivocation. Suppose you and I start off knowing *A* or *B* or *C*. You then eliminate *C*. You accept "If ~*A*, *B*" and reject "If ~*A*, *C*". I eliminate *B*. I accept "If ~*A*, *C*", and reject "If ~*A*, *B*". I assent to a sentence from which you dissent, and vice versa. We do not disagree. We express different propositions, with different truth conditions, governed by our different epistemic states. Worlds which are near for me are far for you.

Are belief-relative truth conditions better than no truth conditions? They account for the validity of arguments; but Adams's logic has its own rationale, without them. They account for sentences with conditional constituents. But we saw (§2.4) that they sometimes give counterintuitive results. Are we able to escape Lewis's result that a conditional probability is not the probability of the truth of a proposition, by making the proposition dependent on the believer's epistemic state? Lewis showed that there is no proposition $A*B$ such that in every belief state $\mathbf{p}(A*B) = \mathbf{p}_A(B)$. He did not rule out that in every belief state there is some proposition or other, $A*B$, such that $\mathbf{p}(A*B) = \mathbf{p}_A(B)$. However, in the wake of Lewis, Stalnaker himself proved this stronger result, for his conditional connective: the equation $\mathbf{p}(A>B) = \mathbf{p}_A(B)$ cannot hold for all propositions in a single belief state. If it holds for *A* and *B*, we can find two other propositions, *C* and *D* (truth-functional compounds of *A*, *B* and $A>B$) for which, demonstrably, it does not hold. (See Stalnaker's letter to van Fraassen published in van Fraassen (1976, pp. 303-4), Gibbard (1981, pp. 219-20), and Edgington (1995, pp. 276-8).)

It was Gibbard (1981, pp. 231-4) who showed just how belief-sensitive Stalnaker's truth conditions would be. Later (1984), Stalnaker abandoned the claim that conditionals express belief-relative propositions, writing "It follows that the conditional ... expresses one proposition when it is asserted, and a different one when it is denied" (p. 110).

4.2 Jackson

Frank Jackson holds that "If *A*, *B*" has the truth conditions of "$A \supset B$", i.e. "~*A* or *B*"; but it is part of its meaning that it is governed by a special rule of assertability. "If" is assimilated to words like "but", "nevertheless" and "even". "*A* but *B*" has the same truth conditions as "*A* and *B*", yet they differ in meaning: "but" is used to signal a contrast between *A* and *B*. When *A* and *B* are true and the contrast is lacking, "*A* but *B*" is true but inappropriate. Likewise, "Even John can understand this proof" is true when John can understand this proof, but inappropriate when John is a world-class logician.

According to Jackson, in asserting "If *A*, *B*" the speaker expresses his belief that $A \supset B$, and also indicates that this belief is "robust" with respect to the antecedent *A*. In Jackson's early work (1979, 1980) "robustness" was explained thus: the speaker would not abandon his belief that $A \supset B$ if he were to learn that *A*. This, it was claimed, amounted to the speaker's having a high probability for $A \supset B$ given *A*, i.e. for (~*A* or *B*) given *A*, which is just to have a high probability for *B* given *A*. Thus, assertability goes by conditional probability. Robustness was meant to ensure that an assertable conditional is fit for modus ponens. Robustness is not satisfied if you believe $A \supset B$ solely on the grounds that ~*A*. Then, if you discover that *A*, you will abandon your belief in $A \supset B$ rather than conclude that *B*.

Jackson came to realise, however, that there are assertable conditionals which one would not continue to believe if one learned the antecedent. I say "If Reagan worked for the KGB, I'll never find out" (Lewis's example (1986, p. 155)). My conditional probability for consequent given antecedent is high. But if I were to discover that the antecedent is true, I would abandon the conditional belief, rather than conclude that I will never find out that the antecedent is true. So, in Jackson's later work (1987), robustness with respect to *A* is simply defined as $\mathbf{p}_A(A \supset B)$ being high, which is trivially equivalent to $\mathbf{p}_A(B)$ being high. In most cases, though, the earlier explanation will hold good.

What do we need the truth-functional truth conditions for? Do they explain the meaning of compounds of conditionals? According to Jackson, they do not (1987, p. 129). We know what "$A \supset B$" means, as a constituent in complex sentences. But "$A \supset B$" does not mean the same as "If *A*, *B*". The latter has a special assertability condition. And his theory has no implications about what, if anything, "if *A*, *B*" means when it occurs, unasserted, as a constituent in a longer sentence.

(Here his analogy with "but" etc. fails. "But" can occur in unasserted clauses: "Either he arrived on time but didn't wait for us, or he never arrived at all" (see Woods (1997, p. 61)). It also occurs in questions and commands: "Shut the door but leave the window open". "Does anyone want eggs but no ham?". "But" means "and in contrast". Its meaning is not given by an "assertability condition".)

Do the truth-functional truth conditions explain the validity of arguments involving conditionals? Not in a way that accords well with intuition, we have seen. Jackson claims that our intuitions are at fault here: we confuse preservation of truth and preservation of assertability (1987, pp. 50-1).

Nor is there any direct evidence for Jackson's theory. Nobody who thinks the Republicans won't win treats "If the Republicans win, they will double income tax" as inappropriate but probably true, in the same category as "Even Gödel understood truth-functional logic". Jackson is aware of this. He seems to advocate an error theory of conditionals: ordinary linguistic behaviour fits the false theory that there is a proposition $A*B$ such that $\mathbf{p}(A*B) = \mathbf{p}_A(B)$ (1987, pp. 39-40). If this is his view, he cannot hold that his own theory is a psychologically accurate account of what people do when they use conditionals. Perhaps it is an account of how we *should* use conditionals, and would if we were free from error: we *should* accept that "If the Republicans win they will double income tax" is probably true when it is probable that the Republicans won't win. Would we gain anything from following this prescription? It is hard to see that we would: we would deprive ourselves of the ability to discriminate between believable and unbelievable conditionals whose antecedents we think false.

4.3 Compounds

A common complaint against Supp's theory is that if conditionals do not express propositions with truth conditions, we have no account of the behaviour of compound sentences with conditionals as parts (see e.g. Lewis (1976, p. 142)). However, no theory has an intuitively adequate account of compounds of conditionals: we saw in §2.4 that there are compounds which Hook gets wrong; and compounds which Arrow gets wrong. Grice's and Jackson's defences of Hook focus on what more is needed to justify the

assertion of a conditional, beyond belief that it is true. This is no help when it occurs, unasserted, as a constituent of a longer sentence, as Jackson accepts. And with negations of conditionals and conditionals in antecedents, we saw, the problem is reversed: we assert conditionals which we would not believe if we construed them truth-functionally.

There have been ambitious attempts to construct a general theory of compounds of conditionals, compatible with Supp's thesis. They are based on a partial restoration of truth values, which has some merit. Note that the difficulties for Hook and Arrow in §§2 and 3 were focused on the last two lines of the truth table -- the cases in which the antecedent is false. No problems arose in virtue of the cases in which the antecedent is true. Perhaps we can say that "If *A*, *B*" is true when *A* and *B* are both true, is false when *A* is true and *B* is false, and has no truth value when *A* is false. We must immediately add that to believe (or assert) that if *A*, *B*, is not to believe (assert) that it is true; for it is true only if *A*&*B*; and one might believe that if *A*, *B*, and properly assert it, without believing that *A*&*B* -- indeed, while thinking that it is very likely not true. If I say "If you press that button, there will be an explosion", I hope and expect that you will not press it, and hence that my remark is not true.

Instead, one must say that to believe "If *A*, *B*" is to believe that it is true *rather than false*; it is to believe that *A*&*B* is much more likely than *A*&~*B*; i.e., to believe that it is true given that it is true or false. This is just to say that one's confidence in a conditional is measured by $\mathbf{p}_A(B)$. Note that for a bivalent proposition, belief that it is true coincides with belief that it is true rather than false. But the latter, not the former, generalizes to conditionals.

This has some minor advantages. It allows one to be right by luck, and wrong by bad luck: however strong my grounds for thinking that *B* if *A*, if it turns out that *A*&~*B*, I was wrong. However poor my grounds, if it turns out that *A*&*B*, I was vindicated.

Now in principle one could handle negations, conjunctions and disjunctions of conditionals by three-valued truth tables; and continue to say that a complex statement is believable to the extent that it is judged probably true given that it is true or false. For a conjunction, $((A \Rightarrow B)\&(C \Rightarrow D))$, the most natural truth table would seem to be: the conjunction is true iff both conjuncts are true; false iff at least one conjunct is false; otherwise it lacks a truth value. This has unappetizing consequences. Consider a conjunction of two conditionals whose antecedents are *A* and ~*A* respectively, such that the first conditional is 100% certain and the second 99% certain, for instance, $((A \Rightarrow A)\&(\sim A \Rightarrow B))$ where $\mathbf{p}_{\sim A}(B)$ = 0.99. This looks like something about which you should be close to certain. But it cannot be true (for one of the antecedents is false), and it may be false, in the unlucky event that it turns out that ~*A*&~*B*. So the probability of its truth, given that it has a truth value, is 0. One can try other truth tables: make the conjunction true provided that it has at least one true conjunct and no false conjunct, false if it has at least one false conjunct, lacking truth value otherwise. And one can come up with equally unappetizing consequences. For work in this tradition and valuable surveys of related work, see McDermott (1996) and Milne (1997).

A variant of this approach gives "semantic values" to conditionals as follows: 1 (= true) if *A*&*B*; 0 (=

false) if A&~B; $\mathbf{p}_A(B)$ if ~A. Thus we have a belief-relative three-valued entity. Its probability is its "expected value". For instance, I'm to pick a ball from a bag. 50% of the balls are red. 80% of the red balls have black spots. Consider "If I pick a red ball (R) it will have a black spot (B)". $\mathbf{p}_R(B)$ = 80%. If R&B, the conditional gets semantic value 1, if R&~B, it gets semantic value 0. What does it get if ~R? One way of motivating this approach is to treat it as a refinement of Stalnaker's truth conditions. Is the nearest R-world a B-world or not? Well, if I actually don't pick a red ball, there isn't any difference, in nearness to the actual world, between the worlds in which I do; but 80% of them are B-worlds. Select an R-world at random; then it's 80% likely that it is a B-world. So "If R, B" gets 80% if ~R. You don't divide the ~R-worlds into those in which "If R, B" is true and those in which it is false. Instead you make the conditional "80%-true" in all of them. The expected value of "If R, B" is $(\mathbf{p}(R\&B) \times 1) + (\mathbf{p}(R\&{\sim}B) \times 0) + (\mathbf{p}({\sim}R) \times 0.8)) = (0.4 \times 1) + (0.1 \times 0) + (0.5 \times 0.8) = 0.8 = \mathbf{p}_R(B)$. Ways of handling compounds of conditionals have been proposed on the basis of these semantic values. But again, they sometimes give implausible results. For developments of this approach, see van Fraassen (1976), McGee (1989), Jeffrey (1991), Stalnaker and Jeffrey (1994). For some counterintuitive consequences, see Edgington (1991, pp. 200-2), Lance (1991), McDermott (1996, pp. 25-28).

Thus, no general algorithmic approach to complex statements with conditional components has yet met with success. Many followers of Adams take (by default) a more relaxed approach to the problem. They try to show that when a sentence with a conditional subsentence is intelligible, it can be paraphrased, at least in context, by a sentence without a conditional subsentence. As conditionals are not ordinary propositions, in that they essentially involve suppositions, this (it is claimed) is good enough. They also point out that some constructions are rarer, and harder to understand, and more peculiar, than would be expected if conditionals had truth conditions and embedded in a standard way. See Appiah (1985, pp. 205-10), Gibbard (1981, pp. 234-8), Edgington (1995, pp. 280-4), Woods (1997, pp. 58-68 and 120-4); see also Jackson (1987, pp. 127-37).

For some constructions the paraphrase can be done in a general, uniform way. For example, "If A, then if B, C" can be paraphrased "If A&B, C". "It's not the case that if A, B" is probably best paraphrased as "If A, it's not the case that B". The alternative would be something like "If A, it might well be the case that ~B", expressing the judgement that the probability of B given A is not particularly high. But with a categorical statement, like "It will rain today", it is when one disagrees strongly that one says "No it won't" or "It's not the case that it will rain today". When one disagrees weakly, one says something like "It might well not" or "I wouldn't be so sure". By analogy, then, it seems that it is strong disagreement with "If A, B" that deserves the negation operator. If someone asserts two or more conditionals joined by "and", each conditional can be assessed as a separate assertion.

Disjunctions of conditionals are peculiar. Of course, it is a sufficient condition for accepting such a disjunction that one accepts one disjunct. But this is rather uninteresting. "Or" is a very useful word when it connects things we are uncertain about, for often we can be confident that A or B, while not knowing which. We are often uncertain about conditionals. Yet "Either (if A, B) or (if C, D) -- but I don't know which" is a form of thought that is rarely if ever instantiated in real life. If conditionals are ordinary statements of fact, this is odd. The problem is not merely one of syntactic complexity: "Either (A&B) or

(*C*&*D*) -- I don’t know which" is just as syntactically complex, but is quite commonplace. Several agile minds have risen to the challenge of providing me with examples of the kind I claim are virtually non-existent: "Either, if I go out I’ll get wet, or, if I turn the television on I’ll see tennis -- I don’t know which": for, either it’s raining or it isn’t. If it’s raining and I go out, I’ll get wet. If it isn’t raining and I turn the television on I’ll see tennis. "Either, if you open box *A*, you’ll get ten dollars, or, if you open box *B*, you’ll get a button -- I don’t know which"; for, if Fred is in a good mood he has put ten dollars in box *A* and twenty dollars in box *B*; if Fred is not in a good mood he has put a paper clip in box *A* and a button in box *B*. All right. But the disjunction of conditionals is an exceedingly bad way of conveying the information you have, and once the necessary background is filled in, we see that the disjunction belongs elsewhere. So we have little use for them. On the other hand, our genuine need for disjunctions shows up naturally inside a conditional "If *A*, either *B* or *C* (I don’t know which)". Some apparent disjunctions of conditionals are really no such thing: "Either we’ll have fish, if John arrives, or we’ll have leftovers, if he doesn’t". Note that both disjuncts are asserted. Note that it doesn’t seem to matter whether one uses "or" or "and" in "If it’s fine, we’ll have a picnic, or [and] if it isn’t, we’ll go to the movies". I conclude that disjunctions proper of conditionals are of little use -- the best we can do, with my early examples, is to discern some disjunction of categorical propositions each disjunct of which supports one or other conditional.

Conditionals in antecedents are also problematic. Gibbard suggests (1981, pp. 234-8) that we have no general way of decoding them, and some cannot be deciphered, for example, said of a recent conference, "If Kripke was there if Strawson was there, then Anscombe was there". "Do you know what you have been told?", he asks (p. 235). When we do understand utterances of this form, he suggests, it is because we can identify some obvious basis, *D*, for an assertion of "If *A*, *B*" and interpret "If (*B* if *A*), *C*" as "If *D*, *C*". For instance, "If the light will go on if you press the switch, the electrician has come": if the power is on, the electrician has come.

As said above, "If *A*, then if *B*, *C*" is to be paraphrased as "If *A*&*B*, then *C*". For to suppose that *A*, then to suppose that *B* and make a judgement about *C* under those suppositions, is the same as to make a judgement about *C* under the supposition that *A*&*B*. Let’s consider this as applied to a problem raised by McGee (1985) with the following example. Before Reagan’s first election, Reagan was hot favourite, a second Republican, Anderson, was a complete outsider, and Carter was lagging well behind Reagan. Consider first

> (1) If a Republican wins and Reagan does not win, then Anderson will win.

As these are the only two Republicans in the race, (1) is unassailable. Now consider

> (2) If a Republican wins, then if Reagan does not win, Anderson will win.

We read (2) as equivalent to (1), hence also unassailable.

Suppose I’m close to certain (say, 90% certain) that Reagan will win, and hence close to certain that

(3) A Republican will win.

But I don’t believe

(4) If Reagan does not win, Anderson will win.

I’m less than 1% certain that (4). On the contrary, I believe that if Reagan doesn’t win, Carter will win. As these opinions seem sensible, we have a prima facie counterexample to modus ponens: I accept (2) and (3), but reject (4). Truth conditions or not, valid arguments obey the probability-preservation principle. I’m 100% certain that (2), 90% certain that (3), but less than 1% certain that (4).

Hook saves modus ponens by claiming that I must accept (4). For Hook, (4) is equivalent to "Either Reagan will win or Anderson will win". As I’m 90% certain that Reagan will win, I must accept this disjunction, and hence accept (4). Hook’s reading of (4) is, of course, implausible.

Arrow saves modus ponens by claiming that, although (1) is certain, (2) is not equivalent to (1), and (2) is almost certainly false. For Stalnaker,

(5) If a Republican wins, then if Reagan doesn’t win, Carter will win

is true. To assess (5), we need to consider the nearest world in which a Republican wins (call it *w*), and ask whether the conditional consequent is true at *w*. At *w*, almost certainly, it is Reagan who wins. We need now to consider the nearest world to *w* in which Reagan does not win. Call it w'. In w', almost certainly, Carter wins.

Stalnaker’s reading of (2) is implausible; intuitively, we accept (2) as equivalent to (1), and do not accept (5).

Supp saves modus ponens by denying that the argument is really of that form. "$A \Rightarrow B$; *A*; so *B*" is demonstrably valid when *A* and *B* are propositions. For instance, if $\mathbf{p}(A)$ = 90% and $\mathbf{p}_A(B)$ = 90% the lowest possible value for $\mathbf{p}(B)$ is 81%. The "consequent" of (2), "If Reagan doesn’t win, Anderson will win", is not a proposition. The argument is really of the form "If *A*&*B*, then *C*; *A*; so if *B* then *C*". This argument form is invalid (Supp and Stalnaker agree). Take the case where *C* = *A*, and we have "If *A*&*B* then *A*; *A*; so if *B* then *A*". The first premise is a tautology and falls out as redundant; and we are left with "*A*; so if *B* then *A*". We have already seen that this is invalid: I can think it very likely that Sue is lecturing right now, without thinking that if she was injured on her way to work, she is lecturing right now.

Compounds of conditionals are a hard problem for everyone. It is difficult to see why it should be so hard if conditionals have truth conditions. Supp is not at a unilateral disadvantage.

5. Other Conditional Speech Acts and Propositional Attitudes

As well as conditional beliefs, there are conditional desires, hopes, fears, etc.. As well as conditional statements, there are conditional commands, questions, offers, promises, bets, etc.. "If he calls" plays the same role in "If he calls, what shall I say?", "If he calls, tell him I'm out" and "If he calls, Mary will be pleased". Which of our theories extends to these other kinds of conditional?

One believes that *B* to the extent that one thinks *B* more likely than not *B*; according to Supp, one believes that *B* if *A* to the extent that one believes that *B* under the supposition that *A*, i.e. to the extent that one thinks *A*&*B* more likely than *A*&~*B*; and there is no proposition *X* such that one must believe *X* more likely than ~*X*, just to the extent that one believes *A*&*B* more likely than *A*&~*B*. Conditional desires appear to be like conditional beliefs: to desire that *B* is to prefer *B* to ~*B*; to desire that *B* if *A* is to prefer *A*&*B* to *A*&~*B*; there is no proposition *X* such that one prefers *X* to ~*X* just to the extent that one prefers *A*&*B* to *A*&~*B*. I have entered a competition and have a very small chance of winning. I express the desire that if I win the prize (*W*), you tell Fred straight away (*T*). I prefer *W*&*T* to *W*&~*T*. I do not necessarily prefer ($W \supset T$) to ~($W \supset T$), i.e. (~*W* or *W*&*T*) to *W*&~*T*. For I also want to win the prize, and much the most likely way for (~*W* or *W*&*T*) to be true is that I don't win the prize. Nor is my conditional desire satisfied if I don't win but in the nearest possible world in which I win, you tell Fred straight away.

If I believe that *B* if *A*, i.e. (according to Supp) think *A*&*B* much more likely than *A*&~*B*, this puts me in a position to make a conditional commitment to *B*: to assert that *B*, conditionally upon *A*. If *A* is found to be true, my conditional assertion has the force of an assertion of *B*. If *A* is false, there is no proposition that I asserted. I did, however, express my conditional belief -- it is not as though I said nothing. Suppose I say "If you press that switch, there will be an explosion", and my hearer takes me to have made a conditional assertion of the consequent, one which will have the force of an assertion of the consequent if she presses the button. Provided she takes me to be trustworthy and reliable, she thinks that if she presses the switch, the consequent is likely to be true. That is, she acquires a reason to think that if she presses it, there will be an explosion; and hence a reason not to press it.

Conditional commands can, likewise, be construed as having the force of a command of the consequent, conditional upon the antecedent's being true. The doctor says to the nurse in the emergency ward, "If the patient is still alive in the morning, change the dressing". Considered as a command to make Hook's conditional true, this is equivalent to "Make it the case that either the patient is not alive in the morning, or you change the dressing". The nurse puts a pillow over the patient's face and kills her. On the truth-functional interpretation, the nurse can claim that he was carrying out the doctor's order. Extending Jackson's account to conditional commands, the doctor said "Make it the case that either the patient is not alive in the morning, or you change the dressing", and indicated that she would still command this if she knew that the patient would be alive. This doesn't help. The nurse who kills the patient still carried out an order. Why should the nurse be concerned with what the doctor would command in a counterfactual situation?

Hook will reply to the above argument about conditional commands that we need to appeal to pragmatics. Typically, for any command, conditional or not, there are tacitly understood reasonable and unreasonable ways of obeying it; and killing the patient is to be tacitly understood as a totally unreasonable way of making the truth-functional conditional true -- as, indeed, would be changing the dressing in such an incompetent way that you almost strangle the patient in the process. The latter clearly is obeying the command, but not in the intended manner. But it is stretching pragmatics rather far to say the same of the former. To take a less dramatic example, at Fred's request, the Head of Department agrees to bring it about that he gives the Kant lectures if his appointment is extended. She then puts every effort into making sure that his appointment is not extended. Is it plausible to say that this is doing what she was asked to do, albeit not in the intended way?

Extending Stalnaker's account to conditional commands, "If it rains, take your umbrella" becomes "In the nearest possible world in which it rains, take your umbrella". Suppose I have forgotten your command or alternatively am inclined to disregard it. However, it doesn't rain. In the nearest world in which it rains, I don't take my umbrella. On Stalnaker's account, I disobeyed you. Similarly for conditional promises: on this analysis I could break my promise to go to the doctor if the pain gets worse, even if the pain gets better. This is wrong: conditional commands and promises are not requirements on my behaviour in other possible worlds.

Among conditional questions we can distinguish those in which the addressee is presumed to know whether the antecedent is true, and those in which he is not. In the latter case, the addressee is being asked to suppose that the antecedent is true, and give his opinion about the consequent: "If it rains, will the match be cancelled?". In the former case -- "If you have been to London, did you like it?" -- he is expected to answer the consequent-question if the antecedent is true. If the antecedent is false, the question lapses: there is no conditional belief for him to express. "Not applicable" as the childless might write on a form which asks "If you have children, how many children do you have?". You are not being asked how many children you have in the nearest possible world in which you have children. Nor is it permissible to answer "17" on the grounds that "I have children ⊃ I have 17 children" is true. Nor are you being asked what you would believe about the consequent if you came to believe that you did have children.

Widening our perspective to include these other conditionals tends to confirm Supp's view. Any propositional attitude can be held categorically, or under a supposition. Any speech act can be performed unconditionally, or conditionally upon something else. Our uses of "if", on the whole, seem to be better and more uniformly explained without invoking conditional propositions.

Bibliography

6.1 General Overviews

- Edgington, Dorothy 1995: "On Conditionals". *Mind* 104, pp. 235-329.

- Harper, W. L., Stalnaker, R., and Pearce, C. T. (eds.) 1981: *Ifs*. Dordrecht: Reidel.
- Jackson, Frank ed. 1991: *Conditionals*. Oxford: Clarendon Press.
- Sanford, David H. 1989: *If P, then Q: Conditionals and the Foundations of Reasoning*. London: Routledge.
- Woods, Michael 1997: *Conditionals*. Oxford: Clarendon Press.

6.2 Other Works Referred to in the Text

- Adams, E. W. 1965: "A Logic of Conditionals". *Inquiry*, 8, pp. 166-97.
- Adams, E. W. 1966: "Probability and the Logic of Conditionals", in Hintikka, J. and Suppes, P. eds., *Aspects of Inductive Logic*. Amsterdam: North Holland, pp. 256-316.
- Adams, E. W. 1970: "Subjunctive and Indicative Conditionals". *Foundations of Language*, 6, pp. 89-94.
- Adams, E. W. 1975: *The Logic of Conditionals*. Dordrecht: Reidel.
- Adams, E. W. 1998: *A Primer of Probability Logic*. Stanford: CLSI Publications.
- Appiah, A. 1985: *Assertion and Conditionals*. Cambridge: Cambridge University Press.
- Bayes, Thomas 1763: "An Essay Towards Solving a Problem in the Doctrine of Chances". *Transactions of the Royal Society of London*, 53, pp. 370-418.
- Bennett, Jonathan 1988: "Farewell to the Phlogiston Theory of Conditionals". *Mind*, 97, pp. 509-27.
- Bennett, Jonathan 1995: "Classifying Conditionals: the Traditional Way is Right". *Mind*, 104, pp. 331-44.
- Dudman, V. H. 1984: "Parsing "If"-sentences". *Analysis*, 44, pp. 145-53.
- Dudman, V. H. 1988: "Indicative and Subjunctive". *Analysis*, 48, pp. 113-22.
- Edgington, Dorothy 1991: "The Mystery of the Missing Matter of Fact". *Proceedings of the Aristotelian Society Supplementary Volume* 65, pp. 185-209.
- Frege, G. 1879: *Begriffsschrift* in Geach, Peter and Black, Max 1960: *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Basil Blackwell.
- Gärdenfors, Peter 1986: "Belief Revisions and the Ramsey Test for Conditionals". *Philosophical Review* 95, pp. 81-93.
- Gärdenfors, Peter 1988. *Knowledge in Flux*. Cambridge MA: MIT Press.
- Gibbard, A. 1981: "Two Recent Theories of Conditionals" in Harper, Stalnaker and Pearce (eds.) 1981.
- Grice, H. P. 1989: *Studies in the Way of Words*. Cambridge MA: Harvard University Press.
- Jackson, Frank 1979: "On Assertion and Indicative Conditionals". *Philosophical Review*, 88, pp. 565-589.
- Jackson, Frank 1981: "Conditionals and Possibilia". *Proceedings of the Aristotelian Society* 81, pp. 125-137.
- Jackson, Frank 1987: *Conditionals*. Oxford: Basil Blackwell.
- Jackson, Frank 1990: "Classifying Conditionals I", *Analysis*, 50, pp. 134-47, reprinted in Jackson 1998.
- Jackson, Frank 1998: *Mind, Method and Conditionals*. London: Routledge.
- Jeffrey, Richard 1991: "Matter of Fact Conditionals". *Proceedings of the Aristotelian Society*

Supplementary Volume 65, pp. 161-183.
- Lance, Mark 1991: "Probabilistic Dependence among Conditionals". *Philosophical Review*, 100, pp. 269-76.
- Lewis, David 1973: *Counterfactuals*. Oxford: Basil Blackwell.
- Lewis, David 1976: "Probabilities of Conditionals and Conditional Probabilities". *Philosophical Review*, 85, pp. 297-315. Page references to Lewis 1986.
- Lewis, David 1986: *Philosophical Papers* Volume 2. Oxford: Oxford University Press.
- Mackie, J. 1973: *Truth, Probability and Paradox*. Oxford: Clarendon Press.
- McDermott, Michael 1996: "On the Truth Conditions of Certain 'If'-Sentences". *Philosophical Review* 105, pp. 1-37.
- McGee, Vann 1985: "A Counterexample to Modus Ponens". *Journal of Philosophy*, 82, pp. 462-71.
- McGee, Vann 1989: "Conditional Probabilities and Compounds of Conditionals". *Philosophical Review* 98, pp. 485-542.
- Milne, Peter 1997: "Bruno de Finetti and the Logic of Conditional Events". *British Journal for the Philosophy of Science*, 48, pp. 195-232.
- Ramsey, F. P. 1926: "Truth and Probability" in Ramsey 1990 pp. 52-94.
- Ramsey, F. P. 1929: "General Propositions and Causality" in Ramsey 1990 pp. 145-63.
- Ramsey, F. P. 1990: *Philosophical Papers* ed. by D. H. Mellor. Cambridge University Press.
- Read, Stephen 1995: "Conditionals and the Ramsey Test". *Proceedings of the Aristotelian Society Supplementary Volume*, 69, pp. 47-65.
- Stalnaker, R. 1968: "A Theory of Conditionals" in *Studies in Logical Theory, American Philosophical Quarterly* Monograph Series, 2. Oxford: Blackwell, pp. 98-112. Reprinted in Jackson, Frank ed. 1991. Page references to 1991.
- Stalnaker, R. 1970: "Probability and Conditionals". *Philosophy of Science*, 37, pp. 64-80. Reprinted in Harper, W. L., Stalnaker, R. and Pearce, G. eds. 1981.
- Stalnaker, R. 1975: "Indicative Conditionals", *Philosophia*, 5, pp. 269-86, reprinted in Jackson, F. ed. 1991.
- Stalnaker, R. 1984: *Inquiry*. Cambridge MA: MIT Press.
- Stalnaker, R. and Jeffrey, R. 1994: "Conditionals as Random Variables", in Eells, E. and Skyrms, B. eds., *Probability and Conditionals*. Cambridge: Cambridge University Press.
- Thomson, James 1990: "In Defense of ⊃". *Journal of Philosophy*, 87, pp. 56-70.
- van Fraassen, Bas 1976: "Probabilities of Conditionals", in Harper, W. and Hooker, C. eds., *Foundations of Probability theory, Statistical Inference, and Statistical Theories of Science*, Volume I. Dordrecht: Reidel, pp. 261-308.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

conditionals: counterfactual | logic: classical | logic: conditional | probability calculus: interpretations of | Ramsey, Frank

Dorothy Edgington
University College, Oxford University
dorothy.edgington@university-college.oxford.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: August 8, 2001
Content last modified: August 8, 2001

Comparative Philosophy, Chinese and Western

Comparative philosophy brings together philosophical traditions that have developed in relative isolation from one another and that are defined quite broadly along cultural and regional lines -- Chinese versus Western, for example. Several main issues about the commensurability of philosophical traditions make up the subject matter of comparative philosophy. One issue is methodological commensurability -- whether and how comparisons between different philosophical traditions, in this case the Chinese and Western, are to be conducted. Views run the gamut from those holding that meaningful comparisons cannot be conducted at all to those holding that the content of traditions must largely be the same. Other issues concerning commensurability concern specific subject matters of traditions. The issue of metaphysical and epistemological commensurability involves the comparison of traditions on their conceptions of the real and their modes of inquiry and justification. Ethical commensurability involves the comparison of these traditions on the matters of how people ought to live their lives, whether both traditions have moralities and if so how similar and dissimilar they are. The separation between these main issues is somewhat artificial, given that a discussion of methodological commensurability will inevitably involve the comparison of traditions on metaphysical, epistemological, and ethical matters. There is some heuristic value, however, in beginning with a general discussion of views on methodological commensurability with a brief illustration of how these views might be applied to some Chinese/Western comparisons. Subsequent sections will address Chinese-Western comparisons in metaphysics, epistemology, and ethics that have assumed special prominence in the literature.

Doing comparative philosophy well can be very difficult because of the vast range of texts and their intellectual and historical contexts it requires its practioners to cover. Oversimplifications, excessively stark contrasts, and illicit assimilations count as the most frequent sins. One benefit of comparative philosophy lies in the way that it forces reflection on the most deeply entrenched and otherwise unquestioned agendas and assumptions of one's own tradition. Another benefit at which its practioners often aim is that the traditions actually interact and enrich one another. Demands for rigor and depth of scholarship obviously rank as some of the most important standards applying to philosophy inquiry. The task of meeting these standards becomes more manageable as the field of inquiry narrows. Such a result can be legitimate but sometimes myopic and impoverishing.

---

1. Methodological Commensurability

Those arguing for radical incommensurability -- the view that the questions and answers in one tradition cannot sustain meaningful statement in the other tradition -- rely on the recognition of radical difference in basic concepts and modes of inquiry. Given such radical differences, they argue, there can be no cross-traditional reference to a common subject matter and to a truth about that subject matter that is independent of the basic conceptual vocabulary and theories and justificatory practices of a particular tradition (see Rorty, 1989, and Shweder, 1989). Looking for a possible Chinese-Western instance of radical incommensurability, one might go to Daoist texts such as the *Daodejing* and the *Zhuangzi*. When it is said in chapter one (as traditionally arranged and on one translation) of the *Daodejing* that "The Way that can be spoken of is not the constant Way," or in chapter two that the "sages abide in nonaction and practice the teaching that is without words," one is finding something different from the usual in Western texts infused with ideals of discursive rationality and argumentation. Or consider the *Lunyu* or *Analects* 1:2, where the importance of rightly ordered family relations is emphasized for right order in the state ("He who has grown to be a filial son and respectful younger brother will be unlikely to defy his superiors and there has never been the case of someone inclined to defy his superiors and stir up a rebellion"). The prominent and enduring place of this theme of state as family writ large sets that tradition apart from Western contractual traditions that have come to emphasize right order in the state as that which can be ratified by an uncoerced agreement among equals concerned to protect their private interests. Opponents of radical incommensurability will level the charge that it presupposes a hyperdramatic contrast between traditions. For example, the Western tradition has not lacked for skeptics on the power of discusive rationality, and some of these skeptics have nevertheless believed in a mode of veridical access to something of supreme significance -- to a powerful experience within themselves or to something much larger outside themselves. At the same time, it must be noted that the positive theme is more recessive in the Western tradition and appears mainly in theistic versions, as it does in Plotinus (*Enneads*), Meister Eckhart (*Von unsagbaren Dingen*) and Hildegaard of Bingen (*Scivias*). Similarly, the Western tradition has certainly housed strains of thought that do view the state as more of a natural outgrowth of small human groups such as family and community. Even a major modern figure such as David Hume (*Treatise of Human Nature*, 3.2.2) explicitly rejects the idea of contract as central to understanding the origin or justification of social and political bonds. There might still be a difference between Chinese and Western traditions with respect to which strains of thought become dominant or at least prevalent, but that difference does not appear to come under the heading of radical incommensurability..

Samuel Fleischacker (1992) proposes a more moderate version of incommensurability -- sometimes we can understand others just well enough to know that we don't understand them. His argument has roots in

Wittgenstein's view that knowledge depends on a background of shared assumptions and standards of evidence. "World pictures" are embedded within cultures. Our world picture involves not only a distinctive set of beliefs about the world but also an ordering of interests that determines how we go about trying to have reliable beliefs. This ordering differs from those dominant in other cultures. We in the West have given precedence to our interests in "egalitarian knowledge" (wanting and believing that people have roughly equal access to the truth) and in prediction and control of this-worldly objects. The world pictures of other cultures embody other interests, and we may not be able to prove that they are wrong, or indeed, we may be unable to fully understand why it is that they value the interests they value so highly. Nevertheless, we understand that they do value these interests highly and think they are wrong to do so. We make such judgments despite our merely partial understanding because we tend to see a certain set of interests as the proper guide for a minimally decent or sensible human life.

The idea that the distinctive character of a "world picture" lies partly in the interests they embody seems plausible. Daoism and Confucianism, at least after a certain stage in the development of these schools, exemplify the way that a set of interests intertwine with beliefs about the world. Both schools exemplify in different ways a conception of understanding the world that is inseparable from the interest in coming into "attunement" with it, to use a felicitous word from Charles Taylor (1982). To become attuned to the world is to see its goodness and to know one's place in the order of the world. To say that Daoists exemplify this theme about seeing the goodness of the world is in a way misleading, since there is in the *Daodejing* and *Zhuangzi* a profound mistrust of our conceptual separations between opposites such as good and bad. However, as has been noted many times, the notion of "nonaction" or *wuwei* does not denote literal inaction but presupposes something like the possibility of an unforced acting *with* the grain of things, and that presupposes that it is possible to become attuned to that grain while in a state of awareness that is not cluttered by distorting conceptual oppositions. Both Daoist texts straightforwardly recommend *wuwei* and in this sense presuppose the goodness of the world and the way its "grain" goes. Confucian texts uphold the ideal of a different kind of attunement, under which the world and its order can be called good without the ambiguity that Daoist skepticism with conceptual opposition creates. For example, the *Mencius* (the name is actually a latinization given by Christian priests for '*Mengzi*') presents a theory that human nature contains the germs or sprouts of goodness, tendencies to certain feelings and judgments such as manifested by the feeling of compassion for a child about to fall into a well. These sprouts are in human nature because they were sent by *tian* (literally meaning 'sky' but most often translated as 'Heaven' and perhaps best conceived as an impersonal ordering force of the universe). Taylor thinks that modern science has severed the connection between understanding and attunement. In Fleischacker's terms, modern science is predicated on different interests, prediction and control foremost among these interests. Taylor believes that the severance of understanding and attunement resulted in superior understanding at least of physical nature. But as he is careful to point out, no single argument can prove global superiority. If we can take attunement as an ideal, we have failed miserably, even as our technological control of nature has increased immeasurably. Perhaps, then, the contrast between a Chinese world picture in which attunement figures prominently and a modern scientific view predicated on the interests of prediction and control serves as an example of the sort of moderate incommensurability Fleischacker has in mind. One question to be raised about this kind of incommensurability, however, is whether it truly involves lack of understanding between traditions. Taylor himself draws examples of the theme of attunement from Plato and the European cultural tradition. Do we really fail to understand the

appeal behind world pictures of attunement? Taylor's complex assessment suggests no inability on our part to understand the force behind both kinds of world pictures. Even to take the stance that attunement pictures are comforting illusions (an assessment less complex than the one Taylor adopts) is to suppose that one does understand the appeal behind them.

A different kind of incommensurability that may arise is not incomprehensibility between traditions but lack of common standards sufficient for settling significant conflicts between them. Alasdair MacIntyre (1988, 1989) has illustrated the difference between being able to understand a tradition and being able to translate all its claims into the language of another tradition by pointing to the possibility of "bilinguals" -- people who, for example, might have been raised within one community and its tradition, and then through migration or conquest, become a member of another community and its different tradition. Such bilinguals might very well understand each tradition, and such understanding might include knowledge of those parts of each that cannot be translated into the language of the other. Such bilinguals would not encounter the sort of radical incommensurability constituted by incomprehensibility, but they may be *unable to resolve conflicts of belief* between the traditions, instead having to relativize the claims of each in some such form as "seems true to this particular community" or "seems justified to this particular community." This kind of "evaluational" incommensurability, rather than meaning incommensurability, might fit better the case of world pictures based on attunement versus world pictures that sever the connection between understanding and attunement. MacIntyre himself presents a possibility for resolving such conflicts in case one tradition continually fails in its attempts to address certain key problems or issues. If another tradition has the conceptual resources to explain why it is that the first tradition continues to fail, then advocates of that first tradition may have to acknowledge its limitations and even transfer their allegiances. Whether that is the case for conflicts over attunement is not obvious (See Wong, 1989).

Moving to the end opposite from various forms of incommensurability, we find views holding that there must be substantial agreeement between traditions. The argument stems from a conception of the way interpretation works. We proceed on the assumption that the others we are interpreting live in the same world as we do.We subvert this assumption, however, if we attribute to them beliefs that substantially differ from our own. As Donald Davidson (1980) has argued, a belief is identified by its location in a pattern of beliefs, and it is this pattern that determines the subject matter of the belief, what the belief is about. If we attribute to others a pattern of beliefs that are different from our own, i.e., false beliefs, this tends to undermine the identification of the subject matter; to undermine, therefore, the validity of the belief as being about that subject. David Cooper (1978) applies a Davidsonian principle of charity to the question of whether different cultures have more or less the same morality. We can only identify others' beliefs as moral beliefs about a given subject matter if there is a massive degree of agreement between their and our beliefs about that subject matter. For Cooper this implies that the moral beliefs we attribute to others must be about something connected with welfare, happiness, suffering, security, and the good life. Michele Moody-Adams (1997) gives a more recent version of the same argument, starting with the premise that understanding others requires that there be quite substantial agreement about many of the basic concepts that are relevant to moral reflection. She concludes that “ultimate” or fundamental” moral disagreement is not possible.

The way that an earlier example of putative radical difference can be questioned serves as some confirmation of this argument from charity for strong agreement between traditions. Confucian conceptions of the state as family writ large are not especially puzzling even when we do not subscribe to them. We find similar themes that have arisen within the Western tradition, and again, it is possible to conceive what the appeal would be. However, these points also raise certain doubts about the use of charity to argue for strong agreement. . The fact that we can point to similar beliefs within our own tradition and that we can imagine what the appeal would lie behind such beliefs does not mean that we share those beliefs with others even as we attribute to them. It may be replied that the principle of charity does not require complete agreement, but only "substantial" agreement or agreement on "ultimate" or "fundamental" disagreement. A further question that will pressed against this position is how much agreement is sufficient for identification of others' beliefs as being about the same subject matter as ours. We do attribute to others error and simple difference (without judging that someone is in error) of belief about the same subject matter. We can attribute error to others if we believe them to be in circumstances that encourage error, and we identify such types of circumstance from past experience of discovering ourselves to be in error. Furthermore, we attribute simple difference of belief when we recognize that there is a range of reasonable interpretations or weightings to be given to evidence. Henry Richardson (1997) has argued that the principle of charity itself needs interpretation and hence cannot be the ultimate standard for interpretation. Interpreting a a philosophical text, requires taking account of the cognitive aims the authors had in writing what they did. Is it more charitable, Richardson asks, for a translator of Machiavelli's *The Prince* to resolve ambiguities and seek to maximize agreement between Machiavelli and ourselves? Or is it more charitable to set him out as intentionally provocative and deliberately cryptic?

An alternative to both the radical incommensurability and strong agreement positions is that there is no general answer on how much agreement or disagreement is to be found between complex and hetereogeneous traditions such as the Chinese and Western. As emerged above, one must distinguish between meaning and evaluational incommensurabilities. Furthermore, how much disagreement we will find not only depends on the particular subject matter but also on the answers we find most plausible within a larger context of inquiry: the attempt to explain others, their actions, practices, institutitions, and history. In such a larger context, attribution to others of beliefs different from ours may be more or less reasonable depending on how it fits into a reasonable explanation of them. And what counts as a reasonable explanation in particular cases will be set within the context of one's larger theories about persons and societies, among other things. One's explanation of people can reasonably attribute error or simple difference of belief on an indefinite number of important matters that might even deserve to be called ’fundamental’ as long as it seems plausible to attribute error of simple difference of belief to them in their epistemic situation as we construe it (see also Grandy, 1973). Or one's explanation can reasonably imply convergence of similarity of belief. An intermediate possibility is that the difference or similarity becomes more prominent as one defines the subject matter in a ‘thicker’ or ‘thinner’ way. Does the Chinese tradition recognize individual rights? The answer, as shall emerge below, may depend on how specifically one defines the notion of a right. If one defines a right rather thinly as what one has whenever one has justifiable claims on others to assure one's possession of things or one's exercise of certain capacities, then one can plausibly argue that there is a common notion of rights between the Chinese and Western traditions. If, on the other hand, one includes within one's definition of rights the idea that they

justifiable independently of what is a good and worthwhile life for human beings, then one plausibly argue against a common notion of rights.

2. Metaphysical and Epistemology Commensurability

One common portrait of the difference between the Chinese and Western traditions posits a radical incommensurability on the very nature of philosophical inquiry. Chinese philosophy is "wisdom" literature, composed primarily of stories and sayings designed to move the audience to adopt a way of life or to confirm its adoption of that way of life. Western philosophy is systematic argumentation and theory. Is there such a difference? One reason to think so is the fairly widespread wariness in Chinese philosophy of a discursive rationality that operates by deduction of conclusions about the particular from high-level generalizations. The seventeenth chapter of the *Zhuangzi* notes that the sage-king Yao looked for a suitable successor, found the perfect candidate in Shun, and then abdicated so that Shun could take over the throne. The result was glorious. However, when Kuai imitated Yao the result was disastrous. Tang and Wu were kings who fought and conquered. But Duke Bo also acted on that rule, fought, and lost. That is why it is impossible to establish “any constant rule.” Inspired by the achievement of insight or wisdom in some particular cases, we create general rules that we believe will work for many other cases in the future. The unfortunate result is that our original insights and wisdom are magnified beyond the scope of their applicability. Confucians are more willing to articulate their teachings in the form of principles, but such principles seem to function as designators of values or general considerations that ought to be given weight in judgments about what to do. Never lost is recognition of the necessity for the exercise of discretion in judgment according to the particular circumstances at hand. The best rules lose applicability in unusual circumstances. Rules and values conflict in many circumstances, and there are no "super-principles" to supply ready answers. The appropriate resolution to each conflict depends very much on the situation. In *Mencius* 7A35, Mencius is asked what the legendary sage-king Shun would have done if his father had killed a man.Mencius replies that the only thing to do would be to apprehend him. Shun could not interfere with the judge, who was acting on the law. However, Mencius continues, Shun would then have abdicated and fled with his father to the seacoast. As Mencius portrays it, then, Shun's actions strike a balance between the different values in tension with one another. The refusal to interfere with the judge is a way of acknowledging the necessity of impartially administering a social order. At the same time, fleeing with one's father is honoring the value of greater loyalty to family. Shun manages to honor both values at different moments in his dealing with the situation. Deduction from a principle could not yield such a balance. We are expected, however, to learn from stories such as Shun's, precisely because they function as concrete paradigms for judgment-making in the future. When we encounter situations that pose similar-looking conflicts between impartial concern and familial loyalties, we have Shun's judgment as a resource and a model. That model is not the same as a general principle that would deductively yield a judgment about what to do in the present situation. We must exercise judgment in determining whether new situations are similar enough to the case of Shun, and we must exercise judgment as to what actions would be parallel to Shun's actions.

Naes and Hanay (1972) have characterized Chinese philosophy as "invitational" in its method of

persuasion, meaning that it portrays a way of life in a vivid fashion so as to invite the audience to consider its adoption. The *Analects*, for example, portrays the ideal of the *junzi* (often translated as "gentleman" but perhaps more accurately glossed as the noble person) as realized by persons of genuine substance who are undisturbed by the failure of others to recognize their merits (1.1: "To be undisturbed by others yet not complain, is this not the mark of the *junzi*?"). In the *Mencius* (2A2), such a person possesses a kind of equanimity or heart that is unperturbed by the prospects of fame and success. This unperturbed heart corresponds to the cultivation of one's *qi* (vital energies) by uprightness. One might be able to see such passages as appealing to experiences the audience might have in its encounters with persons who do seem to possess special strength, substance, and tranquillity through identification with and commitment to a cause they perceive to be far greater than themselves. One need not interpret such sayings as attempting to persuade by the pure emotive effect of certain words, as in propaganda. Rather, they may correspond to a way of doing philosophy that attempts to say something about values in life that can be supported by experience, even if not all testimony will agree (Kupperman, 1999). The Daoists recommend a way of life that they explicitly characterize as one that cannot be argued for, but their recommendation receives some support through commonly shared experience. Consider again the notion of *wuwei* and its illustrations in the *Zhuangzi* through stories of exemplary craft. Most famously, Zhungzi's Cook Ding cuts up an oxen so smoothly and effortlessly that his knife never dulls, and it is if he is doing a dance with his knife as it zips through the spaces between the joints. He does this not through "perception and understanding" but through the *qi*, the vital energies of the body. His marvelous skill is knowledge of how to adjust his own movements to the spaces within oxen that he and the oxen form seamless wholes. Similarly, Woodcarver Qing has learned to prepare for carving his marvelous bellstands in such a way that he clears his mind of all distraction and sees the stand within the timber he has selected. Suggested here is a portrait of acting in the world that consists of complete and full attention to present circumstances so that the agent can act with the grain of things (the Cook Ding passage refers to *tianli* or heavenly patterns). Such a portrait does resonate with the actual experience of craftspeople, artists, athletes, musicians and dancers who have advanced beyond self-conscious technique and rule-following, who become fully absorbed in the experience of working with the material, the instruments or in the movement of their bodies, and who experience their actions as an effortless flow and in fact perform at very high levels. In such ways, Chinese thinkers draw a picture of the world that must in the end by evaluated by explanatory power in some very broad sense. We must ask whether the picture helps make sense of our experience of the world (again in a broad sense of ‘experience’ not limited to quantifiable observations in replicable experiments) and whether it preserves features of that experience that think are *prima facie* genuine.

So then, is it right to say that Chinese philosophy is invitational while Western philosophy is argumentative? One answer is that there is a difference but that it is more a matter of degree than an absolute contrast. It was Aristotle, after all, who said that discussions about the good in human life cannot be properly assimilated by the young because they do not have enough experience of life (*Nichomachean Ethics* I.3). And Plato despite his insistence on the centrality of argumentation to philosophy, dispatches the short analytical arguments presented in Book I of the *Republic* in favor of lengthy expository portraits of the ideal city-state and the harmonious soul for the rest of that work. Those portraits sometimes present only the thinnest of arguments for crucial premises, and at other times no argument at all. Some of his claims, about the divisive effects of family loyalties and the ill-effects of democracy, obviously appeal to experience, even if not all testimony will agree. In fact, it is hard work to find an acknowledged great in

the Western tradition to whom such characterizations do *not* apply, at least to some degree. Sometimes, as in Spinoza (*The Ethics*), the contrast is glaring between the aspiration to prove points by way of deductive argument from self-evident axioms and the obvious source of those points from experience of life. It is true that much Western philosophy, especially of the late modern variety, and most especially emanating from the United Kingdom and North America, attempts to establish its claims through argumentation that is more rigorous than appeals to experience and explanatory power in the broad sense. But it must also be noted that there is argument in Chinese philosophy. Chad Hansen (1992) has pointed out the pivotal role of the philosopher Mozi, who criticized the Confucians for an uncritical acceptance of tradition and who explicitly introduced standards for the evaluations of belief. This introduction of argumentation required response in kind. Mencius gives a Confucian response to to the Mohists and argues on behalf of his theory of human nature as containing the germs or sprouts of the ethical virtues, in the form of natural dispositions to have certain kinds of feeling and judging reactions to situations, such as compassion for a child about to fall into a well (2A6, and see Shun, 1997 for an extensive analysis of argumentation in the *Mencius* text). He defends himself against the arguments of rival theorists who hold that human nature has no innate ethical predispositions but is neutral (6A). Xunzi, a later Confucian thinker, attempts to give a refutation of Mencius's theory in favor of his own theory that human nature has dispositions that get us into trouble and that ethical norms are an invention designed to avoid that trouble (*Xunzi*, chapter 23). Methods of argumentation reach their most sophisticated state of development in Xunzi (See Cua, 1985).

Differences in the way philosophy is conceived may reflect differences in the interests philosophy is meant to satisfy. Chad Hansen (1992) points to another possible difference in interests -- this time in interests that language is meant to satisfy, arguing that the classical Chinese thinkers did not conceive of the primary function of language to be descriptive and as attempting to match propositions with states of affairs, but rather as a pragmatic instrument for guiding behavior. In fact, Hansen sees the Chinese tradition as centrally concerned with the conflict of *daos*, which he defines as sets of behavior-guiding practices, including discourses. Western interpreters have been unable to see this, argues Hansen, because they have imposed their own concerns with correspondence truth and metaphysics on the Chinese tradition. They have as a result imposed upon Daoism an irrational mysticism focused on a metaphysically absolute *Dao*. Michael LaFargue (1992) also argues that the *Daodejing* is not to be interpreted as as concerning some metaphysical entity called the *Dao*, but is rather concerned with self cultivation that allows one to have a transforming experience of deep and peaceful stillness within one's personal center. *Wuwei* is the style of action that is rooted in such an experience. David Hall and Roger Ames (1987) give a related interpretation of Confucius, in part reacting against Herbert Fingarette's (1972) influential interpretation of Confucius' *Dao* as an ideal normative order transcending the contingencies of time, place, history, and culture. Hall and Ames argue Confucius's *Dao* was not conceived as a tradition and language-independent reality against which linguistically formulated beliefs were to be measured as reliable or unreliable, but in fact a cumulative creation of individuals working from within a context provided by a society's tradition, consisting of customs, conventions, conceptions of proper behavior and good manners, conceptions of right conduct and of what is of ultimate value and of what lives are worth living.

These interpretations perform valuable functions in questioning what is sometimes an unreflective imposition of Western philosophical agendas on Chinese thinkers. The debate will go on, however.

Concerning Confucius, it is true that the *Analects* often displays an attitude of tolerance and flexibility in judging where the *Dao* lies. On the other side, it can be pointed out that in sayings such as 1.2 (filial piety and brotherly love are at the root of the virtue of *ren* or humanity), there is no indication that the claim is limited to Chinese culture but extended to human beings generally. One way to understand it is to take it as saying that human beings have to learn to respond to a kind of authority that is not based on force and coercion, but respect and care. Or consider the consistent Confucian theme that rulers cannot hold power simply on the basis of law and punishment. There is no sign such judgments are meant to be limited in scope to one's own time and place. Concerning the *Daodejing*, it is clear that there are very strong practical concerns underlying the text. A way of life is being recommended (as in Hansen), and perhaps that way of life is rooted in a certain kind of transforming experience (as in LaFargue). On the other side, it could be argued that such practical and experiential concerns do not exclude metaphysical concerns. Consider chapter four of that text where *Dao* is described as being empty, as seeming something like the ancestor of the myriad of things, as appearing to precede the Lord (*di)*. For something that at least *looks* metaphysical in the *Mencius,* consider aforementioned 2A2, concerning the unperturbed heart that can be achieved by cultivating one's floodlike *qi.* Such *qi* is vast and unyielding, and if cultivated with uprightness will fill up the space between *tian* (Heaven) and earth. Perhaps the lesson to draw is not that Chinese thinkers lacked metaphysical concerns but that they did not separate practical from metaphysical concerns in the way that contemporary Western thinkers might.

However this issue is resolved with respect to the classical Chinese thinkers, few have disputed that classical concepts such as that of nonbeing *eventually* acquired frankly metaphysical meanings in the Chinese tradition, where it refers at the least to an indeterminate ground in which the determinate “ten thousand things” are incipient (Neville, 1989). This embrace of an indeterminate ground of the determinate may reflect the decision to give the phenomenon of change a fundamental place in ontology, rather than an absolutely stable being as in Parmenidean ontology and as later reflected in Aristotelian and Cartesian notions of substance (Cheng, 1989, 1991). The revival of interest in Chinese metaphysics has partly been fueled by the perception that twentieth century physics has in fact undermined the strategy of giving determinate being ontological primacy (Zukov, 1979). The Neo-Confucian Chu Hsi (*Zhuzi yulei)* reinterpreted ethical themes inherited from the classical thinkers and grounded them in a cosmology and metaphysics. On his conception of *ren* as an all inclusive virtue, it constitutes the *Dao* and consists of the fact that the mind of Heaven and Earth to produce things is present in everything, including the mind of human beings. Another great Neo-Confucian, Wang Yang-Ming (*Quan xilu)* does seem more pragmatic than metaphysical, He taught of the sage who formed one body with Heaven and Earth and the myriad things, but he showed little of Chu's interest in the *li* or principle of existent things. The investigation of things prescribed in the *Great Learning* (*Da Xue*) was not the empirical inquiry Chu envisioned but a rectification of the mind with evil thoughts. Perhaps Chu and Wang represent development of tendencies that were present from the beginning, and between which there was never conceived to be a mutually exclusive choice.

When we get to Chinese Buddhism, there is more evidence for metaphysical concerns that at the same time are urgently practical. It is difficult to view as anything but a metaphysical doctrine the Buddha's view of the self as a floating collection of various psychophysical reactions and responses with no fixed center or unchanging ego entity. He did not deny that we think of the self as a fixed and unchanging

center, but considered such a self a delusion. Our bodily attributes, various feelings, perceptions, ideas, wishes, dreams, and in general a consciousness of the world display a constant interplay and interconnection that leads us to the belief that there is some definite ‘I’ that underlies and is independent of the ever-shifting series. But there is only the interacting and interconnected series. This metaphysical concern, of course, had deep practical implications for the Buddha. It points toward the answer to human suffering, which ultimately stems from a concern for the existence and pleasures and pains of the kind of self that never existed in the first place. The recognition that none of the "things" of ordinary life are fixed and separate entities, anymore than the self is, leads to a recognition of all of life as an interdependent whole and to the practical attitude of compassion for all of life. In a comparative perspective, one cannot help but be struck by the similarity between the Buddhist view of the self and David Hume's doubts about the existence of a unitary and stable self (*Treatise of Human Nature*, 1.4.6). Consider also Derek Parfit's (1984) point that acceptance of a Humean or a Buddhist view of the self can lead to sense that one is less separate from other selves and to a wider concern when "my" projects seem not so absolutely different from "your" projects. However one might regard the argument for impersonal concern from such a view of the self, the view itself may seem to have a claim to a our renewed attention. It certainly fits better with a naturalized conception of human beings as part of this world and not as Cartesian thinking substances that somehow operate apart from the rest of nature.

3. Ethical Commensurability

Confucianism is a perfectionist virtue ethic if such an ethic is distinguished by its central focus on three subjects: character traits identified as the virtues; the good and worthwhile life; and particularist modes of ethical reasoning. These three subjects are interrelated. The virtues are traits of character necessary for living a good life. The virtues typically involve acting on particularist modes of ethical reasoning that do not rely on deducing specific action-guiding conclusions about how to act from general principles but rather on judging in the context at hand what needs to be done. Consider some of the virtues that belong to the *junzi* (the noble person): *ren* (humanity, benevolence), *xiao* (filial piety), *yi* (righteousness), and *li* (acting according to ceremonial ritual or more generally propriety). The very concept of y*i* connotes the ability to identify and perform the action that is appropriate to the particular context (*Analects* 4:10 says that the *junzi* is not predisposed to be for or against anything, but rather goes with what is *yi*). While traditional rules of ritual provide one with a sense of what is courteous and respectful action given standard contexts, the virtue of *yi* allows one to identify when those rules need to be set aside in exigent circumstances (see Cua, 1997). The previously discussed example in the *Mencius* of Shun and his father shows how a ruler's more general concern for his subjects and his filial duties to his fathers must be balanced in ways that cannot be given by principle but only by reflection on what the particular circumstances suggest and allow. Finally, consider that another example from the Mencius about the time when Shun wanted to marry. He knew that his parents, not the wisest nor the best of parents, would not permit him to marry if told of Shun's intention. Shun went ahead and got married without telling his parents, an act that normally would be a grave offense against filial piety. Two reasons are given for the justifiability of this act. One is that the worst way of being a bad son is to provide no heir (4A26); the other is that letting his parents thwart his desire to realize the greatest of human relationships which in turn would cause bitterness toward his parents (5A2). Hence an act that normally would be a grave

offense against filial piety constitutes filial piety in the particular circumstances. Particularist modes of reasoning are needed, then, to judge when the usual rules apply, to balance conflicting values, and to specify the concrete meaning of single values in application to context.

The parallels to ancient Greek virtue ethics, medieval virtue ethics, and also to contemporary virtue ethics in the West are striking, and help to account for the renewal of Western interest in Confucianism. Eastern and Western virtue ethics converge in focusing on certain virtues as crucial for ethical development of the person. There are particularly interesting discussions of courage and the possible role of fear in Mencius and Aristotle (see Van Norden, 1997). Particularist modes of reasoning in Confucianism parallel the Aristotelian notion of a *phronesis* or practical wisdom that depends significantly on knowledge of particulars acquired through experience. A good example is his doctrine of the mean, which holds that virtuous action and feeling consists of avoiding the extremes of deficit and excess. The doctrine does not imply that we ought always to act moderately and with moderate feeling. Aristotle says that the mean is “relative to us,” giving the illustrative analogy that too little food for Milo is too much for the beginner in athletic exercises (*Nichomachean Ethics*, 2.6). Depending on the situation, the appropriate action and feeling may be extreme on a common sense understanding but appropriate given the agent and the circumstances. Part of the contemporary revival of virtue ethics is premised partly on a reaction against the ambition of modern ethical theory to guide primarily through general principles of action rather than through the specification of ideal character traits. Virtue ethics also tend to embody the theme that the ethical life of right (and in the case of Chinese and contemporary Western virtue ethics) caring relationship to others is necessary for human flourishing. In the *Mencius* this theme emerges in identification of the distinctively human potentials with the incipient tendencies to develop the moral virtues (*Mencius* 2A6, 6A1, 6A3, 6A7). Aristotle held that reason makes us distinctively human and that our reason and social nature compel recognition of the desirability of the ethical life for human beings (see Nivison, 1996 for comparisons of Aristotle and Mencius; and Yearley, 1990 for comparisons of Acquinas and Mencius). Xunzi is equally emphatic about the necessity of right and caring relationship to others for human flourishing, even though he denies (at least when he is criticizing Mencius) that human nature contains tendencies to engage in such relationships (see Ivanhoe, 1991, on the way ethical norms help human beings to flourish; and Nivison, 1996a, 1996b, Van Norden, 1992, Wong, 1996b, and Kline, 2000, on the difference between Mencius and Xunzi's theories of human nature; see Goldin, 1999, for a book-length treatment of Xunzi's philosophy).

The similarities coexist with significant differences, however. There is no parallel in Greek virtue ethics for the centrality of family life in the Confucian conception of the good life. Part of the reason for this lies in the Confucian appreciation for the family as the first arena in which care, respect, and deference to legitimate authority is learned (*Analects* 1.2). The way in which particularist reasoning is illustrated in historical stories such as those about Shun is also a distinctive feature of Confucian ethics. These stories present paradigms of good judgment and of good individuals, from which persons engaged in ethical cultivation of themselves should learn. Another distinctive feature of Confucian ethics is the emphasis it gives to an aesthetic dimension of the good life. To act according to ritual propriety is not simply to conform to notions of appropriate behavior in this or that context. It is to act with the right attitude, reverence, say, in the case of serving one's parents in an appropriately respectful manner, and tt is to express such an attitude so gracefully and without internal conflict that doing so has become second

nature (see Kupperman, 1999 and Cua, 1997). The importance attached to *li*, to ritual propriety itself, indicates the Confucian appreciation for the role of culture and convention in enabling human beings to express ethical attitudes toward each another such as care and respect. It is not as if bowing naturally *means* deferential respect; it must be agreed through convention that it does mean something like this (see Fingarette, 1972; see Shun, 1993, for a discussion of the relationship of *li* to the important virtue of *ren*). Mencius and Xunzi engaged in a vigorous, provocative debate over human nature and whether there are natural tendencies that form the basis for development of a good person. They debate in a highly sophisticated manner issues as to whether ethical norms and values are discovered or invented are debated, and their arguments are based partly on what would make for a plausible explanation of how human beings develop into goodness and of how they become bad. Taking all these distinctive features together, it is fair to say that Confucianism offers an especially rich moral psychology (see Nivison, 1996, for several influential essays on moral psychology in comparative perspective, including "Motivation and Moral Action in Mencius," "Philosophical Voluntarism in Fourth-Century China," "Two Roots or One?" and "Xunzi on 'Human Nature'").

One debate that arises within the comparative perspective, however, is whether the Confucians had anything that fits the Western notion of morality. The words 'ethics' and 'ethical' have been used to in this piece to remain neutral on this matter. Some contemporary thinkers (most prominently, Williams, 1985) have tended to confine the 'moral' to a relatively narrow set of characteristics associated with Kant's moral philosophy -- a belief in universal laws validated by pure reason, a belief that responsibility for one's actions requires a freedom from determination from external causes -- and on this view of the moral, it could be argued that there is no equivalent in Chinese philosophy (see Rosemont, 1988). If such a contrast is made, then Confucians are acknowledged to have an ethics as opposed to a morality, where an ethics does embrace questions about value, how one ought to live one's life, and what the good life consists in. One question to be debated here is whether Western notions of the moral are so uniform and narrow as to conform to one philosopher's (even a great one) specific conception of the moral. On the criteria given by Williams, Hume wouldn't have a morality, even though he used the term. It is true that a dominant strain of modern Western morality makes use of a crucial distinction between a morally significant sphere of life that has to do primarily with one's relationships with others and a "private" sphere of life in which one has moral "time-off" that is no one's business but one's own. If it is essential to one's conception of the moral that it be set off from such a private sphere in which moral judgments supposedly have no application, then Confucians would have no morality. Here again, however, it has to be asked whether this is too narrowing. It would eliminate certain utilitarians, for example, who insist that there is no purely private sphere because any type of action or omission could have substantial impact on others given the right circumstances.

Another potential contrast arises from the focus in modern Western moralities on individual rights to liberty and to other goods, where the basis for attributing such rights to persons lies in a moral worth attributed to each individual independently of what conduces to individual's responsibilities to self and others. Confucianism lacks a comparable concept, given its assumption that the ethical life of responsibility to others and individual flourishing are inextricably intertwined. A frequently-made criticism from the Western side is that Confucianism fails to provide adequate protection to those legitimate interests an individual has that may conflict with community interests. On the other side, some

advocates of Confucian ethics criticize rights-focused moralities for ignoring the social nature of human beings and of portraying human life in an excessively "atomistic" or "individualist" conception of persons (e.g., Rosemont, 1986). Against those who argue that Confucianism does not protect the individual enough, it could be replied that the Confucian framework of responsibilities to others can afford significant protections to the individual and arguably addresses the human need for community and belonging better than rights frameworks (Rosemont, 1991). Moreover, it is possible that rights in some sense can play a role in the Confucian tradition, even if such rights are not grounded in the idea of the independent moral worth of the autonomous individual. Within that tradition, rights may be seen as necessary for protecting individuals' interests when the right relationships of care irretrievably break down (Chan, 1999). Rights in the sense of justified claims to be protected in one's speech even when protest and dissenting against authority can be justified as conducive to the health of the community. Mencius recognized a right to revolution against tyrannical kings (1B8); he furthermore advised kings to attach more weight to the opinions of his people than to those of his ministers and officers in making certain crucial decisions (1B7). Xunzi recognized the need for subordinates to speak their views freely to their superiors (*Xunzi*, *Zigong*, *Way of the Son*). If we carry the reasoning in Mencius and Xunzi one step further, we see the need to protect a space in which they may speak freely without fear of suppression, and hence a derive a right in the "thin" sense of what one has whenever one has justifiable claims on others to assure one's possession of things or one's exercise of certain capacities. On the other side, one must be wary of oversimplifications of Western rights-oriented ethical codes. The social nature of persons is not denied by all such codes (of the major theories only Hobbes seems to take an unambiguously “atomistic” view of human beings, and Rousseau and Locke seem to require no such view).

The fact that there are developments of each tradition that bring each closer to the other may suggest that each could learn from the other. Even if not all rights-oriented codes are "atomistic," an increasing worry about Western culture as it has actually developed in practice is the prevalence of an individualism that Tocqueville defined as a "calm and considered feeling which disposes each citizen to isolate himself from the mass of his fellows and withdraw into the circle of family and friends," such that "with this little society formed to his taste he gladly leaves the greater society to look after itself." Such people, Tocqueville observed, form "the habit of thinking of themselves in isolation and imagine that their whole destiny is in their hands." They come to "forget their ancestors" and also their descendants, as well as isolating themselves from their contemporaries. "Each man is forever thrown back on himself alone, and there is danger that he may be shut up in the solitude of his own heart" (1969, pp. 506, 508). Those impressed with this worry and connect it with gross inequality in the most affluent nation in the world would do well to look to a tradition that appreciates the way we thrive or falter within specific communities that nurture or shut us out. On the other side, a tradition that has tended to value the idea of social harmony at the cost of sufficiently protecting dissenters who desire to point out abuses of power or just plain bad thinking by authorities would do well to look at another tradition that does not value social harmony as highly but has endured and is vigorous.

These arguments for greater compatibility between Chinese and Western ethics do not eliminate all significant differences between them on the subject of rights. It is possible to argue that even if responsibility-frameworks are developed and institutionalized to provide genuine protection to dissenting individuals, they cannot provide as much protection as rights frameworks when individual interests

seriously threaten communal or social interests. And if Western ethics sometimes provides more protection to the individual against communal or social interests, this could be seen as unacceptable from a Confucian standpoint. One possible stance on these kinds of differences is evaluational incommensurability. The stance is that each tradition is not wrong to emphasize different values, that no judgment of superiority can be made here. The argument for this may start with the claim that each sort of ethic focuses on a good that may reasonably occupy the center of an ethical ideal for human life. On the one hand, there is the good of belonging to and contributing to a community; on the other, there is the good of respect for the individual apart from any potential contribution to community. It would be surprising, the argument goes, if there were just one justifiable way of setting a priority with respect to the two goods, even when we take into account the justifiable ways in which the two kinds of ethics could be brought closer together. On this view, comparative ethics teaches us about the diversity and richness of what human beings may reasonably prize, and about the impossibility of reconciling all they prize in just a single ethical ideal (e.g., Wong, 1984, 1996).

Daoist ethics are often cited as exemplars of radical difference with the Western tradition. A lot of the case for radical difference rests on the strong skepticism in these ethics about the benefits of conceptualized distinctions between good and bad, right and wrong. Yet these ethics make recommendations that add up to putting forward a way of life. In the *Zhuangzi*, that way of life involves not taking taking oneself and one's ideas so seriously. Chapter two contains a story of monkeys who were furious with their keeper when he announced the policy of three nuts every morning and four in the evening. The keeper then announced a change: "four nuts in the morning and three in the evening," this time to the great delight of the monkeys. Kupperman (1999) suggests that we are invited to view our own urgent concerns in the humorous light of the monkeys' concern about the difference between the two policies. At the same time we are invited to question our own judgment of the monkeys. Why think the monkeys are silly if our urgent concerns might look to some other kind of creature the way the monkeys' concerns look to us? The way of life recommended in the *Zhuangzi*, then, includes openness to what might escape our current conceptualizations and preconceptions. We are invited to see that our conceptualizations of the world are inevitably incomplete and distorting. We attempt to order the world by sorting its features under pairs of opposites, but opposites in the real world never match up neatly with our conceptual opposites. Real "opposites" escape our attempts to cleanly separate them. Despite our best efforts, they switch places in our conceptual maps, blur, and merge into one another. That is why chapter two of the *Zhuangzi* says that the sage recognizes a "this," but a "this" which is also "that," a "that" which is also "this."The appropriate response to the inadequacies of our conceptual structures is to remain open to what those structures distort or hide. In chapter five, men who have had their feet amputated as criminal punishment are scorned by society, but not by their Daoist masters, who see what is of *worth* in them. In chapter one, Zhuangzi chastises his friend Huizi for failing to see beyond the ordinary, humdrum uses of some large gourds. Huizi tried using one of the gourds for a water container, but it was so heavy he couldn't lift it. He then tried to make dippers from them, but they were too large and unwieldy. He deemed the gourds of no use and smashed them to pieces. Zhuangzi asks why he didn't think of making the gourd into a great tub so he could go floating around the rivers and lakes, instead of worrying because it was too big and unwieldy to dip into things! "Obviously you still have a lot of underbrush in your head!" concluded Zhuangzi. He does not deny that the more ordinary uses are genuine uses for the gourds, and indeed, they are. Rather, Zhuangzi's point is to clear the underbrush from our heads and to get

an *enlarged* view of what is of value.

Much of the value of Daoist ethics lies in its warnings against the constricting effects of conventional ethical codes, the blinkering of vision that comes with what we might otherwise regard as admirable integrity and dedication. The other side of this warning is an intriguing positive: that there is a way we can pay attention to the world that allows us to "see around" our blinkering conceptualizations, to see what is of value that we have missed, to move with the grain of things. It is not easy to find an analogue to someone such as Zhuangzi in the Western tradition. There are some important parallels in the Hellenistic Stoics and Epicureans to certain themes in Zhuangzi. They, like him, emphasized the need to accept the inevitable in human life, the need to dampen one's desires to achieve tranquillity in the face of the inevitable, and to identify with the world that makes acceptance and dampening of desires possible (Nussbaum, 1994). On the other hand, there is contrast in the Stoic belief in *logos* as the basis of order in the universe. There are parallels to some Zhuangist themes in Nietzsche: the skepticism about the adequacy of our conceptualizations of the world and of value; the warnings against conventional moralities as constricting, and the awareness of how the application of ethical judgment to others can be a means of asserting power over them (*The Genealogy of Morals*). However, Zhuangzi's vision possesses a generosity and inclusiveness, an embrace of the lowly and cast-off, that seems anathema to Nietzsche's celebration of the übermensch and his project of fashioning himself like a work of art. Nietzsche proposes strong, vital desire at the heart of this aesthetic project, rather than the dampening of desire. He represents a kind of radical individualism that did not find a congenial home in the Chinese tradition (Solomon, 1995).

In Buddhism, especially Chan Buddhism, there also is rejection of conventional ethical values as blinkering and distorting (see Hui Neng, *Liuxu tanjing*, and The *Recorded Conversations of I-Hsüan*), and also a sense that one can become attuned to the world so as to move with its grain. This is not surprising since Buddhism was profoundly influenced by Daoism upon its importation into China. However, Buddhism may have especially challenging implications for Western ethics in its special emphasis on the elimination of suffering and on the way it explains suffering by referring to the human attachment to self as fixed ego entity. As noted above in discussing Parfit's connection to the Buddhist view of the self, realization that the self is not a bounded and discrete entity may encourage a much more impersonal view of oneself and one's projects and desires. One's concern widens to all of life, and one dampens one's desires so as to lessen attachment to the self's cares and concerns. Some may find this an unacceptably demanding ethic. It may seem to drain all passion from life, and it requires that we dampen the attachment we have not only to our selves but also to special others. This negative reaction may fit with Kupperman's (1999) characterization of much Western ethics as upholding only a limited altruism that allows one a private sphere of life free from moral demands and in which one gives much more weight to the cares and concerns of the self and those close to the self. However, as the above comparison with Parfit suggests, there are themes in Western philosophy that parallel the kind of impersonal altruism urged upon us by Buddhism. Some utilitarians have strongly held to the theme that each counts for one in calculating what produces the greatest good, and they have derived challenging consequences from that theme for the question of what one should be prepared to give to alleviate the suffering of strangers (see Singer, 1972, and Unger, 1996), arguing that the way many in affluent nations indulge themselves and their own is simply insupportable in a world of widespread and severe suffering. Some have seen the sort of

impersonal concern that utilitarianism may demand as an indication that it unsuitable for human beings, who are so strongly partial to themselves and their own (see Williams, 1981 and Wolf, 1982). Buddhism presses for the possibility that impersonal concern is humanly possible, and the fact that it is a vibrant and long-lived tradition with many committed practitioners provides some support for the viability of impersonal concern as a ideal that is capable of claiming allegiance and influencing how people try to live their lives (see Flanagan, for a reference to Buddhism in support of the viability of such an ideal .

4. Why Do Comparative Philosophy If It's So Hard?

The most obvious sin of doing comparative philosophy is assimilating another tradition to one's own by unreflectively importing assumptions, frameworks, and agendas into one's reading of that other tradition. Sometimes too much charity (as a principle of interpretation) is insufficiently respectful of the distinctness of the other tradition. There are more subtle dangers. If one is a dissident from main trends in one's home tradition, one is tempted to find another tradition that "got it right." This is fine as long as the desire to find another such tradition does not lead to distortion or oversimplification of that other tradition. These are dangers one can recognize but not always avoid successfully, because success may require knowing a lot about the other tradition when it is hard enough to master a single tradition to the point where one can avoid saying silly things about it. Are the risks worth it?

One reason for thinking so is that comparative philosophy is an instance of a sound and sensible strategy for doing philosophy. When facing hard problems it is simply a good strategy to consider a wide range of enduring, respected ideas bearing on those problems. We of course must be wary of the possibility that the other tradition is not really addressing the same problem we are, or that it is addressing only part of the problem we are addressing. But when there is common address of a problem, it is not always the case that one tradition must be adjudicated as entirely right and the other as entirely wrong. There is a good possibility that each tradition has something insightful to say about some aspect of the problem and that each tradition could incorporate something of what the other tradition has to say (see Yu and Bunnin, 2001). When one crosses traditions in enacting this strategy, there is the opportunity for fruitful interaction and mutual influence. Already discussed above is the opportunity for fruitful interaction between rights-oriented traditions and traditions such as Confucianism focused on the value of relationship and community. Consider another area of potential interaction: when a tradition as long-lived and as sophisticated as the Confucian tradition brims full with writing on the nature of agency, human motivation, and the problem of developing ethical excellence and commitment, one cannot afford to ignore it (see Ivanhoe, 2000, for portraits of the tradition of Confucian and neo-Confucian thinkers on these subjects). Confucianism and Daoism emphasize the need to to pay attention to the concrete details of the situation at hand and displays healthy skepticism about the power of general principle to reveal what sort of action is suitable to the situation. Confucianism displays an appreciation for the power of custom, convention and ceremony for helping to make vivid and concrete and meaningful such ethical abstractions as love, respect, and care. From these themes, there is much to be learned, especially if one's home tradition has had a strong focus for a good portion of the modern period on the top-down deduction of specific ethical judgments from extremely abstract principles. At the same time, an appreciation for the concrete and for the culturally specific may obscure the possibility of general and transcultural principles

that help to evaluate the concrete and culturally specific, and the Chinese tradition could benefit from interaction with the Western (Cua, 1985). Finally, recall the case of the Buddhist ideal of impersonal concern being brought to bear on a debate in Western philosophy over the viability of ideals of impersonal concern. This illustrates the point that comparative philosophy can stretch one's sense of possibility, in this case, of human possibility. That is a benefit in itself.

Bibliography

- Chan, Joseph (1999), "A Confucian Perspective on Human Rights for Contemporary China," in *The East Asian Challenge for Human Rights*, ed. Joanne R. Bauer and Daniel A. Bell, Cambridge: Cambridge University Press.
- Cooper, David E. (1978), "Moral Relativism," in *Midwest Studies in Philosophy* 3.
- Cheng, Chung-ying (1989) “Chinese Metaphysics as Non-metaphysics: Confucian and Daoist Insights into the Nature of Reality,” in *Understanding the Chinese Mind: The Philosophical Roots*, ed. Robert E. Allinson, Hong Kong: Oxford University Press.
- Cheng, Chung-ying (1991), *New Dimensions of Confucian and Neo-Confucian Philosophy*, Albany: State University of New York Press.
- Cua, Antonio (1985) *Ethical Argumentation: A Study In Hsün Tzu's Moral Epistemology,* Honolulu: University of Hawaii Press.
- Cua, Antonio (1997) *Moral Vision and Tradition: Essays in Chinese Ethics*, Washington, D.C.: Catholic University of America Press.
- Davidson, Donald (1969), "Truth and Meaning," in *Philosophical Logic*, ed. J.W. Davis, D.J. Hockney, and W.K. Wilson, Dordrecht: D. Reidel.
- Davidson, Donald (1980), "Thought and Talk," in *Mind and Language*, ed. Samuel Guttenplan, London: Routledge & Kegan Paul.
- Fingarette, Herbert (1972), *Confucius: The Secular as Sacred*, New York: Harper.
- Flanagan, Owen (1991), *Varieties of Moral Personality*, Cambridge, Mass.: Harvard University Press.
- Fleischacker, Samuel (1992) *Integrity and Moral Relativism*, Leiden: E.J. Brill.
- Goldin, Paul Rakita (1999), *Rituals of the Way: The Philosophy of Xunzi*, La Salle, Ill.: Open Court.
- Grandy, Richard (1973), "Reference Meaning and Belief," *Journal of Philosophy* 70: 439-452.
- Hall, David L., and Ames, Roger T. (1987), *Thinking Through Confucius*, Albany: State University of New York Press.
- Hansen, Chad (1992), *A Daoist Theory of Chinese Thought*: *A Philosophical Interpretation*, New York: Oxford University Press.
- Ivanhoe, Philip J. (1991) "A Happy Symmetry: Xunzi's Ethical Thought," *Journal of the American Academy of Religion* 59: 309-322.
- Ivanhoe, Philip J. (2000) *Confucian Moral Self Cultivation*, second edition, Indianapolis: Hackett Publishing.
- Kline III, T.C. (2000), "Moral Agency and Motivation in the *Xunzi*," in *Virtue, Nature, and Moral Agency in the Xunzi*, ed. T.C. Kline III and Philip J. Ivanhoe, Indianapolis: Hackett Publishing Co.

- Kupperman, Joel J. (1999), *Learning from Asian Philosophy*, New York: Oxford University Press.
- Lafargue, Michael (1992), *The Tao of the Tao Te Ching*, Albany: State University of New York Press.
- MacIntyre, Alasdair (1988), *Whose Justice? Which Rationality?* Notre Dame: University of Notre Dame Press.
- MacIntyre, Alasdair (1989), "Relativism, Power,and Philosophy," in *Relativism: Interpretation and Confrontation*, ed. Michael Krausz, Notre Dame: University of Notre Dame Press.
- Moody-Adams, Michele M.(1997), *Fieldwork in Familiar Places*: *Morality, Culture, & Philosophy*, Cambridge, Mass.: Harvard University Press.
- Neville, Robert (1989) “The Chinese Case in a Philosophy of World Religions” in *Understanding the Chinese Mind: The Philosophical Roots,* ed. Robert E. Allinson, Hong Kong: Oxford University Press.
- Nivison, David S. (1996a), *The Ways of Confucianism,* ed. Bryan Van Nordan, La Salle, Ill.: Open Court.
- Nivison, David S. (1996b), "Response to Wong," in *Chinese Language, Thought and Culture: Nivison and His Critics,* ed. P.J. Ivanhoe, La Salle, Ill.: Open Court.
- Nussbaum, Martha C. (1994), *The Therapy of Desire*: *Theory and Practice in Hellenistic Ethics*, Princeton, N.J.: Princeton University Press.
- Parfit, Derek (1984), *Reasons and Persons*, Oxford: Clarendon Press.
- Richardson, Henry, *Practical Reasoning about Final Ends*, Cambridge, Cambridge University Press: 1997.
- Rorty, Richard (1989), *Contingency, irony, and solidarity*, Cambridge: Cambridge University Press.
- Rosemont, Henry (1988), "Against Relativism," in *Interpreting Across Boundaries*, Princeton, N.J.: Princeton University Press.
- Rosemont, Henry (1991), *A Chinese Mirror: Moral Reflections on Political Economy and Society*, La Salle, Ill.: Open Court.
- Shun, Kwong-loi (1993), "*Jen* and *Li* in the *Analects*," *Philosophy East & West* 43: 457-479.
- Shun, Kwong-loi (1997), *Mencius and Early Chinese Thought*, Stanford: Stanford University Press.
- Shweder, Richard (1989), "Post-Nietzschean Anthropology: The Idea of Multiple Objective Worlds," in *Relativism: Interpretation and Confrontation*, ed. Michael Krausz, Notre Dame: University of Notre Dame Press.
- Singer, Peter (1972), "Famine, Affluence, and Morality," *Philosophy and Public Affairs* 1: 229-243.
- Solomon, Robert C. (1995), "The Cross-Cultural Comparison of Emotion," in *Emotions in Asian Thought*, eds. Joel Marks and Roger T. Ames, Albany: State University of New York Press.
- Unger, Peter (1996), *Living High and Letting Die: Our Illusion of Innocence*, New York: Oxford University Press.
- Van Norden, Bryan (1992), "Mengzi and Xunzi: Two Views of Human Agency," *International Philosophical Quarterly* 32:161-184.
- Van Norden, Bryan (1997), "Mencius on Courage," *Midwest Studies in Philosophy: v. 21, The Philosophy of Religion*, Notre Dame: University of Notre Dame Press.

- Van Norden, Bryan (2001), "Mencius and Augustine on Evil: A Test Case for Comparative Philosophy," in *Two Roads to Wisdom? Chinese and Analytic Philosophical Traditions*, ed. Bo Mou, Chicago: Open Court.
- Taylor, Charles (1982), "Rationality," in *Rationality and Relativism*, ed. Martin Hollis and Steven Lukes, Cambridge, Mass.: MIT Press.
- Tocqueville, Alexis de (1969), *Democracy in America*, trans. George Lawrence, ed. J. P. Mayer, New York: Doubleday.
- Williams, Bernard (1981), "Persons, Character, and Morality," in *Moral Luck*, Cambridge: Cambridge University Press.
- Wolf, Susan (1982), "Moral Saints," *Journal of Philosophy* 79: 419-439.
- Wong, David (1989), "Three Kinds of Incommensurability," in *Relativism: Interpretation and Confrontation*, ed. Michael Krausz, Notre Dame: University of Notre Dame Press.
- Wong, David (1996a) "Xunzi on Moral Motivation," in *Chinese Language*, *Thought and Culture: Nivison and His Critics*, ed. P.J. Ivanhoe, La Salle, Ill.: Open Court.
- Wong, David (1996b), "Pluralistic Relativism," *Midwest Studies in Philosophy*, *v. 20*, *Moral Concepts*, Notre Dame: University of Notre Dame Press.
- Yearley, Lee H. (1990), *Mencius and Aquinas: Theories of Virtue and Conceptions of Courage*, Albany: State University of New York Press.
- Yu, Ji-yuan, and Bunnin, Nicholas (2001), "Saving the Phenomena: An Aristotelian Method in Comparative Philosophy," in *Two Roads to Wisdom? Chinese and Analytic Philosophical Traditions*, ed. Bo Mou, Chicago: Open Court.
- Zukav, Gary (1979), *The Dancing Wu Li Masters*, New York: Morrow.

Other Internet Resources

- Wesleyan Chinese Philosophical Etext Archive, A site containing electronic versions of the Chinese texts
- Essential Readings on Chinese Philosophy, An informative bibliography by Bryan Van Norden, including recommendations on translations, articles and books on Chinese Philosophy
- Chad Hansen's Chinese Philosophy Pages, Information and links related to Hansen's "Daoist-oriented" interpretation of the Chinese tradition

Related Entries

Confucius | Mencius | Mohism | Taoism | Zhuangzi

David B. Wong
Duke University
dbwong@duke.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 31, 2001
Content last modified: July 31, 2001

Stanford Encyclopedia of Philosophy

Robert Holkot [Holcot]

Robert Holkot, OP (d. 1349) belonged to the first generation of scholars to absorb and develop the views of William Ockham. He is particularly known for his "covenantal theology" and his views on human freedom within the framework of a divine command ethics. He developed an original theology grounded in Ockham's logic and metaphysics, and his works were influential into the sixteenth century.

1. Life and Work

Robert Holkot was from the village of Holcot (or "cot in the rock" as he glossed it) near Northampton, and apparently a commoner: he spoke of how the most capable men seemed to come from humbler backgrounds. He joined the Dominican order, and if he received the usual training, obtained his initial education in arts, logic, Aristotelian philosophy, and theology within the Dominican schools. He studied at Oxford, commenting on Peter Lombard's *Sentences* in the years 1331-1333. Once he obtained his doctorate in theology, he served as Dominican regent master there. Subsequently, Richard of Bury, the Bishop of Durham, chose Holkot as one of his clerks to work with him in London. Tradition also places Holkot at Cambridge, where he may have served as a Dominican lecturer or regent master in theology prior to 1343, when he is known to have returned to the Dominican priory of Northampton. He remained at Northampton, teaching and writing, until his death of the plague in 1349, acquired, as the story has it, while ministering to the sick.

Holkot produced a number of works over his lifetime. While he was at Oxford, he lectured on Peter

Lombard's *Sentences*, on Matthew and the Book of the Twelve Prophets, and engaged in ordinary and quodlibetal debates. He also engaged in a dispute with his fellow students about epistemology, published as the *Sex articuli*, and probably wrote another work, *De imputabilitate peccati* or *On the imputability of sin*. A text, *De stellis, On the stars*, a rough commentary on Aristotle's *De caelo* was probably originally intended as part of his commentary on the *Sentences*, but circulated as a separate tract. His *Sermo finalis*, the final sermon given at the time of passing on the lectureship on the *Sentences* to the next Dominican, also survives. While in London, Holkot helped Richard of Bury with the book, the *Philobiblon*. Two works for preachers, the *Moralitates* and the *Convertimini*, date from his later years. His most famous Biblical lectures, on the book of Wisdom, are associated with Cambridge, and survive as the *Postilla super librum Sapientiae*. Portions of lectures on Ecclesiastes also survive, most likely from his time in Northampton, and he was known to be giving lectures on Ecclesiasticus when he died. A sermon collection, spanning his career, has also been preserved. Most of these texts exist (if they have come down to us) only in manuscript or early sixteenth century editions. Modern editions are available, however, of selected portions, sermons and questions, and of the *Sex articuli*.

2. Relation to Ockham

2.1 Ockham's influence

Although Holkot was a Dominican, well versed in the texts of Aquinas, his philosophy and theology owe much more to the scholastics of the fourteenth century than to the thirteenth. William Ockham exercised the most important influence. The hallmarks of Ockham's philosophy are: his reduction of Aristotle's ten categories of being to substance and quality; his analysis of the other eight categories and many other terms of philosophical art as connotative terms, best understood as exponible into more fundamental absolute terms denoting substances and qualities; his rejection of Aristotle's final, formal and material causes as properly causal, keeping only efficient causality; his conception of mental language as a logical thought structure existing independently of spoken language; his reformulation of the prevailing views about reference (supposition theory) to accommodate his spare metaphysics; his rejection of species as necessary for knowledge in favor of intuitive cognition or the direct intellectual cognition of objects; and his view that the ethical precepts of the Ten Commandments are not absolute but subject to divine will, such that God could, without contradiction, have created a system in which moral good involves obeying the opposite of each of the traditional commands. Holkot assumed most of Ockham's philosophical positions as foundational, taking them for granted in the development of his theology.

Holkot was not much concerned with defending or exploring his Ockhamist philosophical presumptions. They appear as premises, scattered throughout his texts, rather than subjects of extended analysis.

2.2 Differences with Ockham on epistemology

Holkot did differ with Ockham in the details of his epistemology. Holkot, like Ockham, adopted the terms "intuitive" and "abstractive" cognition to designate the basic forms of human understanding. But

Holkot's treatment of intuitive cognition differed from Ockham's on the question of the possibility of intuitive cognition of non-existents. For Ockham, intuitive cognition was the direct intellectual cognition of the presence and existence of an object. Holkot used Ockham's own style of analysis to develop his critique. He noted that "intuitive cognition" was a connotative term, connoting both a kind of quality, which is cognition, and the cognized object as it exists and is present in itself. The term stands for the co-presence of cognition with its object. This led Holkot to argue against Ockham's contention that God's omnipotent power to cause directly whatever is ordinarily caused through secondary causes would enable God to conserve the intuitive cognition of an object even after the object has been destroyed. Holkot objected that given the meaning of the term "intuitive cognition," if God were to conserve cognition of an object after destroying it, that cognition by definition could no longer be an intuitive cognition. It would be an abstractive cognition, the kind of cognition present in the absence of an object.

Holkot also differed with Ockham about the nature of abstractive cognition. He argued in favor of retaining species as part of natural and cognitive processes. In *De stellis* he refers to the sun propagating the natural species of light through the medium of the air. He did not consider the species operative in cognition to be such natural species, however. The term ‘species’ when used to refer to the whiteness in one external object and to the whiteness in another could be called univocal, having the same meaning in each case, but the term ‘species’ used to refer to the whiteness in an object as a quality and to the whiteness representing the object in the intellect was equivocal. The intellectual species is only a likeness of the thing in the sense of representing it (like a statue of Hercules in relation to Hercules), and we experience it in ourselves as it enables us to think about an external object in the absence of that object. Holkot was not much concerned whether such "spiritual qualities" were called "species," "idols," "images," or "exemplars" as long as they were understood to serve as representatives of things or even "knowledge habits" and not as the natural qualities that exist in extramental reality. Holkot's opponent was not Ockham, here, however, but his Dominican contemporary William Crathorn, who had argued for the view that natural and cognitive species were the same in kind. Holkot ridiculed Crathorn's position at length in the *Sex articuli*, on the grounds that if Crathorn were right, our minds would become white or black, hot or cold, depending on what we were thinking about. Crathorn was arguing in line with a long tradition stretching back to Roger Bacon. Holkot's sharp disjunction between natural and spiritual "likenesses," natural and spiritual qualities, went beyond the traditional distinction between sensible and intelligible species and seems to show the effects of the Ockhamist critique, even while he retained remnants of the Aristotelian vocabulary.

3. Natural Theology

3.1 What reason cannot do

Ockham had argued for stringent limits on the ability of reason to establish the existence of God. While an argument for the existence of God as "first conserver" of things could be made, Ockham had argued against the ability of natural reason to prove there was only one divine being. Holkot developed such strictures, arguing that unaided human reason could not prove through a strict demonstration that any incorporeal being like an angel or God existed. The consequence for Holkot was that any reference to

such incorporeal beings found in the texts of ancient philosophers must have come down to them from their predecessors passing on a vestige of knowledge about God acquired ultimately from Adam and Eve. Holkot also contended that some pagans, who lacked the law of Moses, still received faith and grace from God outside the Mosaic Law because they did their best to live according to the principles of natural law. Holkot's sanguine view of pagan philosophers like Hermes Trismegistus and Aristotle rested not on their ability to use natural reason to discern theological truths, but on his confidence that God had accorded a measure of revelation to more than those who had the texts of scripture.

3.2 What reason can do

If basic theological premises require revelation for human beings to know them, then the arena of human reason in theology is restricted to reasoning about what is revealed. Some of the tenets of Christian doctrine, like the doctrines of the Trinity, Incarnation and Eucharist, offer particular challenges to logic. There was a general belief among medieval scholastics that Aristotelian logic exemplified natural reason at its best and was universally applicable to all domains because its rules held through formal relation to the principle of non-contradiction. If key Christian doctrines were not amenable to Aristotelian logical principles, however, it would seem to imply that God is not subject to the principle of non-contradiction and that Aristotelian logic is not universal. Holkot took up these issues in his discussion of the doctrine of the Trinity.

Difficulty arises in the doctrine of the Trinity over doctrinally true premises that seem to give rise to doctrinally false conclusions:

> The divine Essence is the Father,
>
> The divine Essence is the Son,
>
> Therefore, the Father is the Son.

Prior to Holkot, a variety of distinctions had been proposed to modify the identity relation of the copula in such premises and to block the conclusion. But Holkot objected that the divine Essence was in no way "really," "modally," "formally," "rationally," "convertibly," nor in any other way distinguished from the divine Persons or the divine relations of paternity, filiation and spiration. This put him back face to face with the dilemma.

Holkot responded in a passage for which he is perhaps best known, that there must be two systems of logic, a logic appropriate to the natural order, exemplified in Aristotle's works, and a logic appropriate to the supernatural order, a logic of faith, whose rules are supplementary to those of Aristotle. He concluded that Aristotelian logic did not hold universally, but only for the natural order unless additions were made to take into account theological cases. This did not mean that he abandoned the principle of non-contradiction in matters of faith. Rather the nature of the divine being meant that syllogisms involving Trinitarian terms functioned like expository syllogisms about particulars when unquantified universal

terms are substituted for particular ones:

> Human being is running,
>
> Human being is bald.
>
> Therefore, bald human being is running.

The conclusion is invalid because the subject term in each premise might stand for different people, like Plato and Socrates.

Holkot argued that since Aristotle could not have known about God as three Persons and one divine Essence, he could not have foreseen the need to adjust his logic for such cases, but with some supplementary rules taken from religious authority, like: "every absolute is predicated in the singular and not in the plural about the three persons," and "unity holds its consequent where the opposition of relation does not stand in the way" (*Sent.* I, q. 5, f. f2ra), Holkot believed the Trinitarian cases could be covered. The logic of faith does not have a large number of additional principles, and it, like Aristotelian logic, is rational because it is subject to the principle of non-contradiction.

Holcot's view of the relation between faith and reason was very much in the tradition of Anselm, of faith seeking understanding. His adherence to the principle of non-contradiction was uncompromising: "no intellect can assent to the opposite of the first principle or believe that contradictories are true at the same time" (*Quod.* I, q. 2, in *Exploring*, 38, ll. 165-166). Faith required that reason believe that all of the truths of the faith are compatible, even when at times they could not be demonstrated or shown to be so.

4. Necessity and Contingency

4.1 Historical context

The Condemnations of 1277 and John Duns Scotus impelled the view that the world could be other than it is. The idea that God's omnipotent power provides him with an infinity of choices out of which he chooses to create only one set of possibilities became a governing idea among subsequent English schoolmen. Scotus also argued forcefully for the idea that each moment was open to contingent possibility, such that for any time t, the events at t were possible not to be the events at t. Contingency, traditionally assigned to the future, in Scotus' view superceded or governed even the hypothetical necessity of the present. Ockham retreated from Scotus' view, reassigning contingency to future events and reasserting the full force of hypothetical necessity for events in the present. However, working out the implications for philosophy and theology of a contingent world order was the central intellectual challenge for Holkot's generation.

4.2 God's absolute and ordained power

Divine omnipotence involves the absolute power to enact anything that does not involve a contradiction. But among the multitude of possibilities open to divine enactment, God chooses or ordains a subset of compatible possibilities that constitute the world and its history as we know it. The relationship between God's absolute power and the ordained system in place at any given time provided a fault line for exploring questions of necessity and contingency. Thirteenth century theologians formulated the relationship as one between what God has done and what he could have done otherwise, safely relegating contingency to a now foreclosed past. Canon lawyers, however, appropriated the distinction to describe the powers of the pope to set aside "ordained" or enacted Church law through the "plenitude" or "absolute" power of his office. Because papal power transcended enacted law, and popes (and monarchs) who enacted laws were in some sense not subject to those laws, they could provide for exceptions or change enacted laws without contradiction. Such application of the distinction between absolute and ordained power raised the possibility that God might intervene in the ordained system through his absolute power. Beginning with Scotus, the formulation of the canonists began to enter into discussions of God's exercise of absolute power. The appropriation of the legal tradition did not lead to the conclusion (at least for Scotus, Ockham and Holkot) that God uses his absolute power to act inordinately in the ordained system, but rather it enables God (as it did in the change from the Old to the New Law) to set aside one ordained system and replace it with another. Several different and incompatible systems of divine legislation have operated at different times during human history. God's absolute capacity to transcend any given ordained system and replace it with another has made such a switch possible without involving God in a contradiction of his nature. Holkot also invoked this dialectical relationship between God's absolute and ordained power to explain how God provides for dispensations from his laws in particular cases. God never acts inordinately, but the system of divine ordinations is complex and involves multiple incompatible subsets capable of being in place at any given time.

Holkot analyzed God's power in terms of sets of compatible propositions.

> If all the propositions that can exist, were to exist, God cannot do what would entail contradictory propositions being true at the same time, and He can do all those things that, having posed them perfectly instantiated in being, entail no contradictory propositions being true at the same time. (*Sent*. II, q. 2, art. 6, f. i4va)

He then argued that talk of God's absolute and ordained power was not about a two-fold power, but two ways of modifying the proposition: "God can produce A." The proposition "God can produce A from his ordained power," means that it is possible for God to produce A, and A will be compatible with his existing statutes. The proposition "God can produce A from his absolute power," means that it is possible for God to produce A (because A in itself entails no contradictory propositions being simultaneously true), and A is not compatible with his existing statutes. God has only one power, which is God himself, and which human beings can understand in two different ways: ordinately and absolutely. (The stricture on propositional existence results from his view, which he shared with Ockham and a number of his contemporaries, that only propositional tokens counted as real propositions capable of producing a logical contradiction.)

The principle of non-contradiction served as the ultimate safeguard of rationality and certainty in Holkot's system. The role of the principle was particularly important because Holkot, more than perhaps any other late medieval theologian, underscored God's freedom to set aside ordained laws without incurring any fault or obstacle.

God can be obliged to no law but that without its observance he can be morally good, because otherwise the divine goodness would depend on creatures, and God would be less good than he is if he were to destroy every creature; and similarly God would begin to be better than he was before the observance of the law. Whence, just as a prince who is above the law can perform some act without sin or evil, which those existing under the law in no way can do without sin, so God in not fulfilling what he promised acts without the evil of falsity or perjury, which someone existing under the law could in no way do. (*Quodl*. III, q. 8, in *Seeing the Future*, 103, ll. 537-546)

Divine promises, revelations, and enactments were all in Holkot's view ungrounded in divine goodness in the sense that God was not under obligation because of his goodness to fulfill them or keep them in being. The contingency of the ordained system was a fact of the human condition. So what reassurance could human beings have that keeping faith with God's precepts would result in their salvation? What would happen if God were to set aside the current law and enact some incompatible alternative, as it would clearly seem to be in God's power to do? If God did not inform people about such a change, then invincible ignorance would protect them from being held accountable for not following the new laws. Holkot did not believe that God could ask people to obey laws of which they were ignorant because that would require them to do what is impossible and contradictory. And if God did inform people of the new laws, then these laws would supercede the incompatible old set, and the faithful could obey God without being held to contradictory commands.

5. Divine Command Ethics

5.1 Covenantal Theology

In a system of divine command ethics, human beings are obliged to do what God asks them to do because God commands it, not because there is some underlying system of absolute goodness that ethical precepts should ideally mirror. William Ockham had subscribed to such a view. He had argued that no contradiction would arise if God were to command that the Ten Commandments, the precepts fundamental to both the Old and New Laws, were no longer in effect and that from then on people would be obliged to obey their opposites. Most of Ockham's sympathizers backed away from the idea that God could command people to hate him, on the grounds that that command, at least, would be contradictory. But Holkot followed Ockham in subscribing to the ultimate contingency of the decalogue.

With no act having intrinsic worth, the meritoriousness of human behavior was grounded in a covenant between God and the human faithful. Within the terms of the New Law, God would not deny salvation to all those who did their best to obey his commands and adhere to the Articles of the Faith. The causal

effect of meritorious acts in effecting salvation was a secondary form of causality, functioning like money, as an agreed upon medium of exchange in the economy of salvation. Because God's goodness was not a guarantee of covenant, however, Holkot stressed that human adherence to the terms of the covenant constituted an act of faith that God would indeed uphold his promises, even knowing that nothing compels God to do so.

5.2 The significance of intention

Where the fact of command matters more than the substance of what is commanded, human intention to obey has greater significance than the substantive enactments of obedience. Holkot perceived the connection between divine command and the human intention to obey as at the heart of the relationship between human beings and God. For instance, Holkot posed the case of a simple old woman, who in good faith comes to church to hear a new church doctrine from her bishop. If the bishop gets the doctrine backwards, explaining to his congregation that they should believe just the opposite of what the new article of belief contains, is the elderly laywoman required to accept the words of her bishop as true? One of Holkot's fellow scholars had argued that she would only be in this position if she were being punished for sin, but Holkot responded that it was not the substance of her belief that mattered, but her intention to do what was right and to obey God. Her intention to do her best to conform her will to God would be sufficient under the covenant to ensure her salvation if she were to persevere in that intention. God would not deny her salvation because those on whom she necessarily depended for knowledge about God's will were misinformed or confused.

Debates over the place of deception in terms of both the absolute and ordained systems of possibility involved Holkot and a number of his immediate contemporaries. If the world is a contingent place that can be other than it is, then do God's revelations constrain the scope of his possible future actions? If not, can what God says be deceptive or false? Discussion took up scriptural instances in which God seemed to deceive. Holkot, against a number of his contemporaries, argued that God can deceive human beings even in the ordained system, has deceived them as shown in scripture, and has deceived them for no redeeming good apparent to human beings. If God's words to human beings might be deceptive, such that it is not just the bishop who may impart false information, but even God, then the human intention to believe God's words as true and to obey them takes on even more importance. Holkot did not believe that God was playing the role of Descartes' deceiving demon, but Holkot also did not know how to rule out the possibility that he might be deceived about any given thing he believed. The important thing was that even if he were deceived, God had promised that his intention to believe what was revealed and to do what he understood God wanted him to do provided security under the covenant. Faith in the covenant was the source of certainty, not rational demonstration.

The place of intentionality in Holkot's theology and his generous view of divine graciousness provide the context for his use of a version of what has come to be called "Pascal's wager." Holkot passed on a story about a learned heretic who was converted to a belief in immortality by a challenge from a Dominican lay brother: if you believe in immortality and it is true, you will have gained a great deal, and if you believe in immortality and it is not true, you will lose nothing. Forming an intention to believe could

constitute doing his best on the part of the heretic, and God would reward such an intention with the grace necessary for conversion to the belief.

6. Divine Foreknowledge

Discussions about the contingency of the created order and the various ways necessity might impinge on that contingency tended to focus on the challenge God's foreknowledge of future events posed to the contingency of events.

In his commentary on Peter Lombard's *Sentences*, Holkot put forward an elaborate argument:

> If *a* is a sin that Socrates will freely commit tomorrow.
>
> Then it is argued: God knows that *a* will be, therefore from eternity He knew that *a* will be or he began to know that *a* will be.
>
> It cannot be said that He began to know that *a* will be, because then He could know or foreknow something anew and as a result of time. . . .
>
> If He knew *a* from eternity, I pose that "*a* will be" was written on a wall yesterday. Therefore, the proposition "that written on the wall was true" is true, and . . . consequently necessary because it is a true proposition about the past. Therefore, it is necessary that it be the case as the proposition denotes, i. e., it is necessary that Socrates sin. (*Sent*. II, q. 2, in *Seeing the Future*, 126, ll. 307-317.)

Holkot contended that the common response of his era to such an argument, was to pose the possibility of a counterfactual past: to say that the proposition "*a* will be" is true, yet contingently true, and therefore, although it is true, it can never have been true. Holkot argued, the possibility of a counterfactual past differentiated propositions about the future on contingent matters and their equivalents--whether set in the past or present--from propositions about the past and present that are not about such contingent matters. The propositions "*a* was known by God," and "*a* is known by God," although set in the past and present, are true and yet can never have been true, just like other propositions about the future, because they are about *a*, and *a*, as a future contingent, may still not happen. Holkot's response is recognizable as a version of what in modern discussions is called the Ockhamist solution (although the argument traces back at least to Bonaventure).

6.1 The "obligational" model

What Holkot added to the discussion was an elaborate analysis of such puzzles using the rules and structure of obligational debates to explore counterfactual possibility. Debates *de obligatione* were a commonplace in the medieval university curriculum and involved one person, the "opponent" posing a

proposition to another, the "respondent," that, if accepted, would form the basis for a continuing exchange. The posed proposition was usually a counterfactual or a proposition whose truth status was uncertain. The opponent then proposed further propositions to the respondent, each of which might follow from, contradict or be irrelevant to the first. The respondent was to take the first proposition as true for the time of the debate (understood in Holkot's version to take place in a single hypothetical instant of time), and to respond with agreement or rejection depending on whether the succeeding propositions followed from what had been agreed on before or contradicted the preceding concessions. If a proposed proposition was unconnected with any of the preceding propositions, the respondent would respond with agreement, rejection or doubt depending on what he understood the actual state of affairs in the world to be. The forms and rules of obligational debate suggested a rigorous format for exploring the contingent possibilities, which Holkot adopted.

A simple form of the puzzle might proceed as follows:

> Opponent: Let it be the case that God knows *a* will be, where *a* is a future contingent.
>
> Respondent: I accept.
>
> Opponent: Everything that is possible to be is also possible not to be (by the definition of contingency).
>
> Respondent: I accept.
>
> Opponent: As a future contingent, *a* is possible to be and possible not to be.
>
> Respondent: I accept.
>
> Opponent (from the Aristotelian rule that the impossible does not follow from the possible): Let it be the case that *a* will not be.
>
> Respondent: I accept.
>
> Opponent: Then God is deceived.

In resolving such puzzles, Holkot invoked a series of rules, one of which has significance for his moral philosophy, as well. Holkot argued that when the opponent proposed the initial proposition, he was also implicitly posing the rejection of its contradictory. The Aristotelian rule that the impossible does not follow from the possible, seems to allow the contradictory of the initial proposition to enter the debate. But Holkot argued that such a move in effect amounted to starting the debate all over again with a new starting point, a proposition contradictory to the first. The respondent would now be obliged, if he continued with the debate, to answer in accordance with the new contradictory proposition, and would refuse to concede that "God is deceived."

Holkot viewed the human relationship to divine revelation as equivalent to engaging in an obligational debate. The faithful obligated themselves to accept divine revelations as true for the time of this life (even though as contingents it was possible that they might not be true), and if God commanded them to act in a way contrary to previous commands, the new commandment would supercede the old, just if a new obligational debate had begun. Those who accepted the obligation to obey, would also be obligated to live in a way that was consistent with the obligations incurred, even if God did not reveal the details. Human reason was required to discern how to act in uncertain cases.

6.2 The modal ground of experience

In dealing with the problem of God's knowledge of future contingents, Ockham had proposed thinking about time as the modal feature of language. Propositions in the past tense are necessary *per accidens*: they refer to events that could have been otherwise before they happened, but that now could not have been other than they were because of the necessity of the past. Propositions in the present tense are hypothetically necessary: they refer to events that could be otherwise, but given that they are what they are, if they are, cannot not be what they are. Propositions in the future tense are contingent: they refer to events that are possible to be and also possible not to be. Ockham argued that God's knowledge of events tracked this modal arrow just as human knowledge of events does, reintroducing an arrow of "time" for God as well as human beings.

In the ensuing years, Ockham's modal view of "time" was joined to a way of speaking about truth traceable to Richard of Campsall, an Oxford master of arts and theology teaching in the years just before Ockham. Holkot exemplifies this way of thinking.

In *De Interpretatione*, chapter 9, Aristotle had bequeathed a difficult problem to the medieval debate about divine foreknowledge. His contention that in order to avoid attaching necessity to all events, propositions about future events were not yet true or false seemed to deny to God the possibility of knowing the future or to rule out the contingency of events. Boethius had provided a response that held until the fourteenth century, but after Scotus subjected his response to a severe critique, new discussion of the Aristotelian three-valued logic appeared. Campsall distinguished between propositions about the past and present that were "determinately true or false" and propositions about the future that were "indeterminately true or false." Holkot adopted this way of modally dividing up determinations of truth and falsity:

> . . . future contingents are said to be propositions about the future of which there is no determinate truth or falsity, because although they are true or false, yet those which are true can never have been true and those which are false can never have been false. (*Quod.* III, q. 1, in *Seeing the Future*, 63, ll. 93-96.)

By the time of Holcot, the analysis of future contingency in terms of a possible counterfactual past, and the identification of such an analysis with a multi-valued logic, had attained the status of an identifiable

tradition. Ockham had not adopted the terminology of "indeterminately true or false" and "contingently true or false" to speak of the truth status of future contingent propositions. He had insisted on a two-valued system in which all propositions are determinately true or false. But Holkot departed from him in this. Holkot's position reflects a view of modality as primary. Necessity and contingency are fundamental, and judgments of truth mean something different in each modal context, rather than truth being primary, and necessity and contingency providing a different valence to otherwise true propositions. The efforts to grapple with the implications of contingency had come a long way.

Bibliography

Texts

- *In quatuor libros sententiarum quaestiones*. Lyons, 1518. Reprinted by Minerva GMBH Frankfurt, 1967. Also contains *De imputabilitate peccati*, and *Determinationes* I-XII, the first of which is not by Holkot.
- *Quaestiones quodlibetales*. In *Exploring the Boundaries of Reason: Three Questions on the Nature of God by Robert Holcot, OP*, ed. Hester Goodenough Gelber (Toronto, 1983).
- *Quaestiones quodlibetales*. In *Roberto Holcot O.P.: Dottrina della grazia e della giustificazione con due questioni quodlibetali inedite,* edited by Paulo Molteni. Pinerolo, 1967.
- *Quaestiones quodlibetales*. In *Seeing the Future Clearly: Questions on Future Contingents*, ed. Paul Streveler and Katherine Tachau with William J. Courtenay and Hester Goodenough Gelber. Toronto, 1995.
- *Quaestiones quodlibetales*. In J. T. Muckle, "Utrum Theologia sit scientia. A Quodlibetal Question of Robert Holcot, O.P." *Mediaeval Studies* 20 (1958): 127-153.
- *Quaestiones quodlibetales*. In Ernest A. Moody "A Quodlibetal Question of Robert Holcot, O. P. on the Problem of the Objects of Knowledge and of Belief," *Speculum* 39 (1964), 53-74. Reprinted in idem, *Studies in Medieval Philosophy, Science and Logic: Collected Papers, 1933-1969* (Berkeley and Los Angeles, 1975), 321-352.
- *Quaestiones quodlibetales*. In William J. Courtenay, "A Revised Text of Robert Holcot's Quodlibetal Dispute on Whether God is Able to Know More Than He Knows," *Archiv für Geschichte der Philosophie* 53 (1971): 1-21. [A revision of Moody's edition.]
- *Quaestiones quodlibetales*. In Kurt Villads Jensen, "Robert Holkot's Questio on Killing Infidels: A Reevaluation and an Edition," *Archivum Fratrum Praedicatorum* 63 (1993): 207-228.
- *Sermo finalis*. In J. C. Wey, "The *Sermo finalis* of Robert Holcot," *Mediaeval Studies* 11 (1949): 219-224.
- *Sex articuli*. In *Die "Conferentiae" des Robert Holcot O.P. und die akademischen Auseinandersetzungen an der Universität Oxford 1330-1332*, ed. Fritz Hoffmann, 65-127. Beiträge zur Geschichte der Philosophie und Theologie Mittelalters, n. s. no. 36. Münster, 1993.
- *Tractatus de stellis*. In Lynn Thorndike, "A New Work by Robert Holcot (Corpus Christi College, Oxford, MS 138)," *Archives internationales d'histoire des sciences* 10 (1957): 227-235.
- *Super libros Sapientiae*. Hagenau, 1494. Reprinted by Minerva G.M.B.H. Frankfurt, 1974.

Translations

- *Super libros Sapientiae*, chap. 3, lects. 35 and 52; chap. 12, lect. 145. In *Forerunners of the Reformation: The Shape of Late Medieval Thought Illustrated by Key Documents*, ed. and trans, Heiko Oberman (Philadelphia, 1981) 142-150.

Secondary Literature

- Allen, Judson B. (1969). "The Library of a Classiciser: The Sources of Robert Holkot's Mythographic Learning," in *Arts libéraux et philosophie au moyen âge*, Actes du IVe congrès internationale de philosophie médiévale (Paris and Montreal) 721-729.
- -----, (1971). *The Friar as Critic: Literary Attitudes in the Later Middle Ages* (Nashville).
- Courtenay, William J. (1972). "The King and the Leaden Coin: The Economic Background of 'Sine qua non' Causality." *Traditio* 28:185-209. Reprinted in William J. Courtenay, *Covenant and Causality in Medieval Thought: Studies in Philosophy, Theology and Economic Practice* (London, 1984).
- -----, (1980). "The Lost Matthew Commentary of Robert Holcot, O.P." *Archivum Fratrum Praedicatorum* 50:103-112.
- -----, (1985). "The Dialectic of Omnipotence in the High and Late Middle Ages," in *Divine Omniscience and Omnipotence in Medieval Philosophy*, ed. Tamar Rudavsky, 243-269. Synthese Historical Library, no. 25 (Dordrecht).
- -----, (1990). *Capacity and Volition: A History of the Distinction of Absolute and Ordained Power* (Bergamo).
- Del Pra, Mario (1956). "Linguaggio e conoscenza assertiva nel pensiero di Roberto Holcot," *Rivista critica di storia della filosofia* 11 (1956): 15-40.
- -----, (1974) "La proposizione come oggetto della conoscenza scientifica nel pensiero di Roberto Holcot," in *Logica e realtà: momenti del pensiero medievale* (Bari) 83-119.
- Gillespie, Richard E. (1971). "Robert Holcot's Quodlibeta," *Traditio* 27: 480-490.
- Grassi, Onorato (1979). "Le tesi di Robert Holcot sul valore non scientifico della conoscenza teologica," *Rivista di filosofia neo-scolastica* 71:49-79.
- Hoffmann, Fritz (1963). "Robert Holcot--Die Logik in der Theologie," in *Die Metaphysik im Mittelalter*, Miscellanea mediaevalia 2 (Berlin) 624-639.
- -----, (1971)"Der Satz als Zeichen der theologischen Aussage bei Holcot, Crathorn und Gregor von Rimini," in *Der Begriff der Repraesentatio im Mittelalter: Stellvertretung, Symbol, Zeichen, Bild*, ed. Albert Zimmermann, 296-313. Miscellanea mediaevalia, no. 8. (Berlin).
- -----, (1972). *Die theologische Methode des Oxforder Dominikanerlehrers Robert Holcot*. Beiträge zur Geschichte der Philosophie und Theologie des Mittelalters, n. s. no. 5. (Münster).
- -----, (1974). "Thomas -Rezeption bei Robert Holcot?" *Theologie und Philosophie* 49:236-251.
- -----, (1995). "Der Wandel in der scholastischen Argumentation vom 13. zum 14. Jahrhundert, aufgeseigt an zwei Beispielen: Robert Holcot und William (Johannes?) Crathorn (1330-1332 in Oxford)," in *Die Bibliotheca Amploniana: Ihre Bedeutung im Spannungsfeld von Aristotelismus, Nominalismus und Humanismus,* ed. Andreas Speer, 301-322. Miscellanea Mediaevalia, no. 23,

(Berlin).

- Incandela, Joseph M. (1994). "Robert Holcot, O.P., on Prophecy, the Contingency of Revelation, and the Freedom of God," *Medieval Philosophy and Theology* 4:165-188.
- Kennedy, Leonard A. (1993). *The Philosophy of Robert Holcot, Fourteenth-Century Skeptic*. Studies in the History of Philosophy, no. 27 (Lewiston, N. Y.).
- Kirjavainen, Heikki (1990). "Existential Presuppositions in Semantics According to Ockham and Holcot," in *Knowledge and the Sciences in Medieval Philosophy*, vol. 2, eds. Simo Knuuttila, Reijo Työrinoija, and Sten Ebbesen, 196-209. Eighth International Congresss of Medieval Philosophy (Helsinki).
- -----, (1993). "Trinitarian Sophisms in Robert Holcot's Theology," in *Sophisms in Medieval Logic and Grammar*, ed. Stephen Read, Acts of the Ninth European Symposium for Medieval Logic and Semantics (Dordrecht) 348-356.
- Meissner, Alois (1953*). Gotteserkenntnis und Gotteslehre: Nach dem Englischen Dominikanertheologen Robert Holcot* (Limburt/Lahn).
- Molteni, Paulo (1967). *Roberto Holcot O.P.: Dottrina della grazia e della giustificazione con due questioni quodlibetali inedite* (Pinerolo).
- Normore, Calvin (1982). "Future Contingents," in *The Cambridge History of Later Medieval Philosophy: From the Rediscovery of Aristotle to the Disintegration of Scholasticism 1100-1600*, eds. Norman Kretzmann, Anthony Kenny and Jan Pinborg, 358-381 (Cambridge).
- -----, (1985). "Divine Omniscience, Omnipotence and Future Contingents: An Overview," in *Divine Omniscience and Omnipotence in Medieval Philosophy*, ed. Tamar Rudavsky, 3-22. Synthese Historical Library, no. 25 (Dordrecht).
- Nuchelmans, Gabriel (1973). *Theories of the Proposition: Ancient and Medieval Conceptions of the Bearers of truth and Falsity*. North-Holland Linguistic Series 8 (Amsterdam) 195-208.
- Oberman, Heiko A. (1962). "Facientibus quod in se est Deus non denegat gratiam: Robert Holcot, OP and the Beginnings of Luther's Theology," *Harvard Theological Review* 55:317-342.
- -----, (1963). *The Harvest of Medieval Theology: Gabriel Biel and Late Medieval Nominalism* (Cambridge, Mass.) 235-248.
- Schepers, Heinrich (1970). "Holkot contra dicta Crathorn: I. Quellenkritik und biographische Auswertung der Bakkalareatsschriften zweier Oxforder Dominikaner des XIV. Jahrhunderts." *Philosophisches Jahrbuch* 77: 320-354.
- -----, (1972). "Holkot contra dicta Crathorn: II. Das ‘significatum per propositionem’. Aufbau und Kritik einer nominalistischen Theorie über den Gegenstand des Wissens." *Philosophisches Jahrbuch* 79:106-136.
- Smalley, Beryl (1950-51). "Some Latin Commentaries on the Sapiential Books in the Late Thirteenth and Early Fourteenth Centuries," *Archives d'histoire doctrinale et littéraire du moyen âge* 25-26:103-128.
- -----, (1956)"Robert Holcot, OP," *Archivum Fratrum Praedicatorum* 26:5-97.
- -----, (1960) *English Friars and Antiquity in the Early Fourteenth Century* (Oxford) 133-202.
- Spade, Paul Vincent (1992). "If *Obligationes* Were Counterfactuals," *Philosophical Topics* 20:171-188.
- -----, (1993). "Opposing and Responding: A New Look at *Positio*," *Medioevo* 19:232-257.
- Streveler, Paul (1976). "Robert Holkot on Future Contingencies: A Preliminary Account," in

Studies in Medieval Culture, no. 8-9, eds. John R. Sommerfeldt and E. Rozanne Elder, 163-171 (Kalamazoo).

- Tachau, Katherine (1982). "The Problem of the *Species in medio* at Oxford in the Generation after Ockham," *Mediaeval Studies* 44: 394-443.
- -----, (1988). *Vision and Certitude in the Age of Ockham: Optics, Epistemology and the Foundations of Semantics 1250-1345*. Studien und Texte zur geistesgeschichte des Mittelalters, no. 22 (Leiden).
- -----, (1987). "The Influence of Richard Campsall on 14th-Century Oxford Thought," in *From Ockham to Wyclif*, eds. Anne Hudson and Michael Wilks, 109-123. Studies in Church History 5. (Oxford).
- -----, (1987). "Wodeham, Crathorn, and Holcot: The Development of the *complexe significabile*," in *Logos and Pragma: Essays on the Philosophy of Language in Honor of Professor Gabriel Nuchelmans*, eds. L. M. de Riijk and H. A. G. Braakhuis, 161-187 (Nijmegen).
- -----, (1991). "Looking Gravely at Dominican Puns: The 'Sermons' of Robert Holcot and Ralph Friseby," *Traditio* 46:337-345.
- -----, (1991). "Richard Campsall as a Theologian: New Evidence," in *Historia Philosophiae Medii Aevi: Studien zur Geschichte der Philosophie des Mittelalters*, eds. Burkhard Mojsisch and Olaf Pluta, 2:979-1002 (Amsterdam/Philadelphia).
- -----, (1994). "Robert Holcot on Contingency and Divine Deception," in *Filosofia e teologia nel trecento: Studi in ricordo di Eugenio Randi*, eds. Luca Bianchi, 157-196 (Louvain-la-neuve).
- -----, (1995). "Introduction," in *Seeing the Future Clearly: Questions on Future Contingents*. Eds. Paul Streveler and Katherine Tachau with William J. Courtenay and Hester Goodenough Gelber (Toronto).
- -----, (1996). "Logic's God and the Natural Order in Late Medieval Oxford: the Teaching of Robert Holcot," *Annals of Science* 53: 235-267.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Aristotle: on non-contradiction | categories: medieval theories of | causation: medieval theories of | condemnation of 1277 | conscience: medieval theories of | Descartes, René | Duns Scotus, John | future contingents: medieval theories of | medieval philosophy | modality: medieval theories of | Ockham [Occam], William | Pascal's wager | practical reason: medieval theories of

Acknowledgements

All translations are by the author.

Hester Gelber
hgelber@stanford.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 23, 2001
Content last modified: July 23, 2001

Epistemological Problems of Perception

The historically most central epistemological issue concerning perception, to which this article will be almost entirely devoted, is whether and how beliefs about physical objects and about the physical world generally can be justified or warranted on the basis of sensory or perceptual experience -- where it is *internalist* justification, roughly having a *reason* to think that the belief in question is true, that is mainly in question (see the entry justification, epistemic: internalist vs. externalist conceptions of). This issue, commonly referred to as "the problem of the external world," divides into two closely related sub-issues, which correspond to the first two main sections below. The first of these issues has to do with the nature of sensory experience and its relation to the physical world; it is typically (though as we shall see not altogether perspicuously) formulated as the question of what are the *immediate* objects of awareness in sensory experience or, in a variant but essentially equivalent terminology, of what is *given* in such experience. Perhaps the most historically standard, though not currently the most popular answer to this question has been that it is *sense-data* (private, non-physical entities actually having the experienced sensory qualities) that are the immediate objects of awareness or that are given. The second issue has to do with the way in which beliefs about the physical world are justified on the basis of such sensory experience. If it is concluded that physical objects are not themselves given, the two main answers to this question are *representationalism* (the view that the immediate objects of experience represent or depict physical objects in a way that allows one to infer justifiably from such experience to the existence of the corresponding "external" objects) and *phenomenalism* (the view that physical objects are reducible to or definable in terms of the occurrence and obtainability of such experience). A third alternative view that has received attention in recent discussion is *direct realism*: the view that physical objects are after all directly or immediately perceived in a way that allegedly avoids the need for any sort of justificatory inference from sensory experience to physical reality. In addition to these views concerning the internalist justification of beliefs about physical objects, there are also externalist accounts of how such beliefs are justified; these will be briefly considered at the end of the article.

---

1. The Nature of Sensory Experience

What is it that we are *immediately* or *directly* aware of in sensory or perceptual experience? Is it public physical objects, private sensory entities of some sort, or perhaps some still further sort of entity (or state)?

1.1 The idea of immediacy or givenness

Before considering answers to this question, it is important to become clearer about the meaning of the question itself. What is it for something to be an object of *immediate* (or *direct*) awareness or to be given? (For brevity I will mostly employ the latter term.)[1] Historically, most of those (beginning with Descartes and Locke) who have attempted to answer this question have concluded that it is something other than a physical object that is given, but the reason for this conclusion cannot be understood without becoming clearer about the idea of immediacy or givenness itself.

There are two more or less standard criteria of givenness that have at various times been offered, either jointly or separately, but these unfortunately do not seem to work very well. The first of these appeals to the idea of *inference*: something is immediately experienced or is given if the cognitive consciousness of it is not arrived at via any sort of inferential process. The obvious problem with this is that it makes it hard to make sense of the historically common view that physical objects are not given, since the sensory awareness of physical objects in the most normal sorts of cases does not *seem* on the surface to be a product of inference. Certainly the person who has such an awareness is not normally conscious of having made an inference; and to insist, as is sometimes done, that there must have been an "unconscious inference" of some sort could only be justifiable if some other criterion of givenness is being at least tacitly invoked. It is somewhat more plausible to hold that beliefs about physical objects, even if not arrived at via inference, must still be inferentially *justified*, but neither the rationale for such a claim nor its relation to the idea of givenness or immediacy is clear at this point.

The second of the standard criteria appeals to the idea of *certainty*: something is immediately experienced or given if the awareness of it is certain, incapable of being mistaken. But while it is plausible enough that all perceptual awareness of physical objects is at least in principle subject to error,

it is less clear that there is *anything* generally present in sensory or perceptual experience about which error is impossible; beliefs about any aspect of experience, involving as they do the need for conceptual classification, are always capable in principle of being mistaken. Nor, for that matter, is it clear why if some sort of item did have this status, this would show that the awareness of it is more fundamental in the way that the idea of immediacy or givenness seems to suggest, and in particular why beliefs about physical objects must be somehow based on the awareness of this other sort of item.

On the basis of difficulties like these, it has sometimes been concluded that the idea of immediacy or givenness is hopelessly obscure, with the implication being that there is no reason why the perceptual awareness of physical objects should not be itself regarded as epistemically fundamental, not needing to be justified by appeal to a more basic awareness of sense-data or of anything else. This is at least part of the motivation for direct realist views (see below).

But such a conclusion (as is suggested in part by the difficulties that arise in making clear sense of direct realism) seems too hasty. My suggestion instead would be that the underlying conception of immediacy or givenness, of which the foregoing criteria are best regarded as merely symptoms, is that for something to be given is simply for it to be an aspect or feature of the content of conscious experience itself (as distinct from the intentional objects that such conscious content may in part be about). On this conception, much more is given than sensory content narrowly construed: conscious feelings, the conscious aspects of emotions, and most importantly the conscious contents of thoughts or beliefs. But it also seems clear that there is a narrowly sensory component, consisting of or involving colors, shapes, sounds, tactile qualities of various sorts, etc. And although more detailed arguments will be considered shortly, it also seems more or less undeniable that physical objects, at least as commonsensically construed, are not themselves literally part of that conscious content, even though they are depicted or represented by various aspects of it.

1.2 The Sense-Datum Theory

As already remarked, the most widely held view historically concerning the strictly sensory aspect of immediate or given experience is that what is given in such experience is, not public physical objects, but rather *sense-data* (or sometimes *sensa*): private, non-physical entities that actually possess the various sensory qualities that a person experiences. (For a variant usage of this term, see Moore [1953], who there uses the term "sense-datum" to stand for *whatever* it is that is immediately experienced or given, possibly even a public physical object, and then argues somewhat tentatively that the entities that actually have this status are sense-data in the more usual sense, rather than physical objects.)

Two main arguments have been offered for this view.

The Argument from Illusion

(Or, perhaps better, the argument from illusion/perceptual relativity/hallucination.) This very widely advocated argument (first offered explicitly in Berkeley [1713]) appeals to the immense variety of cases

in which: (i) what is immediately perceived or given has different qualities from different perspectives or under different perceptual conditions, even though the relevant physical object does not change (perceptual relativity); or (ii) in which qualities are immediately experienced that the relevant object clearly does not possess (illusion); or (iii) in which qualities are experienced in a situation in which there is no physical object of the relevant sort present at all (hallucination). Some fairly standard examples: viewing a circular coin from different angles, resulting (allegedly) in the experience of a variety of elliptical shapes; viewing a white or colored object under different kinds of lighting; feeling the temperature of a luke-warm bucket of water with a hand that has previously become inured to water of substantially higher or lower temperatures; viewing a straight stick that is immersed in water and so looks bent; hallucinating a non-existent object such as pink rat or a dagger.

The basic claim is that in cases of illusion or hallucination, the object that is immediately experienced or given has qualities that no public physical object in that situation has and so must be distinct from any such object. And in cases of perceptual relativity, since objects with different qualities are experienced from each of the different perspectives or under each of the relevant conditions, at most one of these various immediately experienced or given objects could be the physical object itself; it is then further argued that since there is no apparent experiential basis for regarding one out of any such set of related perceptual experiences as the one in which the relevant physical object is immediately experienced, the most reasonable conclusion is that the immediately experienced or given object is always distinct from the physical object (or, significantly more weakly, that there is no way to identify which, if any, of the immediately experienced objects is the physical object itself, so that the evidential force of the experience is in this respect the same in all cases, and it is epistemologically as though physical objects were never given, whether or not that is in fact the case).

The cogency of this argument has been challenged in a number of different ways, of which the most important are the following. First, it has been questioned whether there is any reason to suppose that in cases of these kinds there must be some *object* present that actually has the experienced qualities, which would then seemingly have to be something like a sense-datum. Why couldn't it be the case that the perceiver is simply in a state of seeming to experience such an object without any object actually being present? (See the discussion below of the adverbial theory.) Second, it has been argued that in cases of illusion and perceptual relativity at least, there is an object present, namely the relevant physical object, which is simply misperceived, for the most part in readily explainable ways. Why, it is asked, is there any need to suppose that an additional object is also involved? Third, the last part of the argument has been challenged, both (i) by questioning whether it is really true that there is no experiential difference between veridical and non-veridical perception; and (ii) by arguing that even if sense-data are experienced in non-veridical cases and if the difference between veridical and non-veridical cases is, after all, experientially indiscernible, there is still no reason to think that sense-data are the immediate objects of experience in veridical cases. Fourth, various puzzling questions have been raised about the nature of sense-data: Do they exist through time or are they momentary? Can they exist when not being perceived? Are they public or private? Can they be themselves misperceived? Do they exist in minds or are they extra-mental even if not physical? On the basis of the intractability of these questions, it has often been argued that the conclusion of the argument from illusion is clearly unacceptable or even ultimately unintelligible, even in the absence of a clear diagnosis of exactly where and how it goes

wrong.[2]

The Argument from the Scientific Account of Perception

A second argument for the conclusion that sense-data, rather than physical objects, are the direct or immediate objects of even veridical perceptual experience appeals to the causal account of the perceptual process offered by natural science. The main aspects of that account that are cited in this connection are: (i) the fact that the character of the resulting experience and of the physical object that it seems to present can be altered in major ways by changes in the conditions of perception or the condition of the relevant sense-organs and the resulting neurophysiological processes, with no change in the external physical object (if any) that initiates this process and that may seem to be depicted by the experience that results; (ii) the related fact that any process that terminates with the same sensory and neural results will yield the same perceptual experience, no matter what the physical object (if any) that initiated the process may have been like; and (iii) the fact that the causal process that intervenes between the external object and the perceptual experience takes at least a small amount of time, so that insofar as the character of the experience reflects the features of that object, they reflect an earlier stage of that object rather than the one actually existing at that moment. (In extreme cases, as in astronomical observation, the external object may long ago have ceased to exist by the time it is perceived.) These facts are claimed to point inexorably to the conclusion that the direct or immediate object of such an experience, the object that is *given*, is an entity produced at the end of this causal process and is thus distinct from the physical object, if any, that initiates the process.

It is difficult to resist the conclusion that there is a fundamental distinction between the external object, if any, that initiates the perceptual process and the perceptual experience that eventually results, thus amounting to a fundamental *dualism* between perceptual experience and the external object and raising the issue of how the latter can be known on the basis of the former. What can and has been resisted, by the adverbial theory in particular, is the idea that this dualism is a dualism of *objects*, with perceptual experience being a more direct experience of objects of a diffeent sort, sense-data. (Though we will eventually see that the result of balking at this point may have less epistemological significance than it has usually be accorded.)

1.3 The Adverbial Theory

The sense-datum theory is often characterized as an *act-object* theory of the nature of immediate experience: it accounts for such experience by postulating both an *act* of awareness or apprehension and an *object* (the sense-datum) which that act apprehends or is aware of. The fundamental idea of the adverbial theory, in contrast, is that there is no need for such objects and the problems (such as whether they are physical or mental or somehow neither) that they bring with them. Instead, it is suggested, merely a mental act or mental state with its own intrinsic character is enough to account for the character of immediate experience.

According to the adverbial theory, what happens when, for example, I immediately experience a silver

elliptical shape (as when viewing a coin from an angle) is that I am in a certain specific state of sensing or sensory awareness or of *being appeared to*: I sense in a certain manner or am appeared to in a certain way, and it is that specific manner of sensing or way of being appeared to that accounts for the specific content of my immediate experience. This content can be verbally indicated by attaching an *adverbial* modifier to the verb that expresses the act of sensing (which is where the label for the view comes from). Thus in the example just mentioned, it might be said that I sense or am appeared to *silver-elliptical-ly* -- where this rather artificial term is supposed to express the idea that the qualitative content that is treated by the sense-datum theory as involving features or properties of an object should instead be thought of as somehow nothing more than the specific manner in which I sense or the specific way in which I am appeared to. Similarly, when I hallucinate a pink rat, I sense or am appeared to *a-pink-rat-ly* -- or, perhaps better, *a-pink-ratshape-ly*. And analogously for other examples of immediate experience.

The essential point here is that when I sense or am appeared to silver-elliptical-ly, there need be nothing more going on than that I am in a certain distinctive sort of experiential state. In particular, there need be no object or entity of any sort that is literally silver and elliptical -- not in the material world, not in my mind, and not even in the realm (if there is such a realm) of things that are neither physical nor mental.

1.4 Conclusion Concerning the Sense-Datum and Adverbial Theories

Each of these two views has fairly obvious virtues and equally obvious drawbacks. The sense-datum theory accounts more straightforwardly for the character of immediate experience. I experience a silver and elliptical shape because an object or entity that literally has that color and shape is directly before my mind. But both the nature of these entities and (as we will see further below) the way in which they are related to the mind are difficult to understand.

The adverbial theory, on the other hand, has the advantage of being metaphysically simpler and of avoiding difficult issues about the nature of sense-data.[3] The problem with it is that we seem to have no real understanding of the nature of the states in question or of how exactly they account for the character of immediate experience. It is easy, with a little practice, to construct the adverbial modifiers. But it is doubtful that anyone has a very clear idea of the meaning of such an adverb, of what exactly it says about the character of the state -- beyond saying merely and unhelpfully that it is such as to *somehow* account for the specific character of the experience.

Here I will limit myself to a brief consideration of one further, less obvious argument against the sense-datum theory and in favor of the adverbial theory, and to pointing out why the issue between these two views, though of great metaphysical significance, may not matter very much if at all for epistemological purposes. As we have so far characterized it, the sense-datum theory is incomplete in one fundamental way. In addition to arguing that sense-data exist, a sense-datum theorist needs some account of the relation between a person and a sense-datum when the former immediately experiences the latter. It does not seem acceptable to say that simply that the sense-datum is itself in the mind, a mental entity, since it has features which neither an immaterial mind nor the physical brain seem capable of possessing. The natural thing to say is that the sense-datum somehow influences the internal state of the person (that is, of

his or her mind) in a way that reflects the sense-datum's specific character. But the resulting state of mind would then be just the sort of state that the adverbial theory describes, one which is such that a person who is in it will thereby experience the properties in question. And there would then be no apparent reason why such a state could not be produced directly by whatever process is supposed to produce the sense-datum, with the latter thus becoming an unnecessary intermediary. Thus the sense-datum theorist must apparently say that the immediate experience of the sense-datum does not involve any distinct internal state of the person that reflects its character, but is instead an essentially and irreducibly *relational* state of affairs. The person simply experiences the sense-datum, but without there being any corresponding change in his or her internal states that would adequately reflect the character of the supposed sense-datum and so make its existence unnecessary in the way suggested. But does this really make good metaphysical sense, and, more importantly, would it allow the person to grasp or apprehend the nature of the sense-datum in a way that could be the basis for further justification and knowledge?

Both views thus have serious problems, though, in light of the last argument, I would assess the problems of the sense-datum theory as the more serious. Fortunately, however, as already suggested, it does not seem necessary for strictly *epistemological* purposes to decide between these two views. The reason is that while they give very different accounts of what is ultimately going on in a situation of immediate experience, they make no difference with respect to the experienced content of that experience. And it is on that experienced content, not on the further metaphysical explanation of it, that the justificatory power, if any, of such an experience depends.

2. The Justification of Beliefs about the Physical World

In considering this second issue, it will be useful to begin by assuming, subject to later reconsideration, that *either* the sense-datum theory or the adverbial theory is correct: that what we are *immediately* aware of in sensory experience is never an external material object, but is either a sense-datum or else the content of a state of sensing or being appeared to. It will be useful to have a brief label for this disjunctive position, and I will refer to it here as *perceptual subjectivism*.[4] In fact, I will usually, for the sake of simplicity, formulate the issues to be considered in terms of sense-data, leaving the adverbial variant to be supplied by the reader.[5]

Assuming perceptual subjectivism, there are two main non-skeptical alternatives regarding the justification of beliefs concerning material objects on the basis of immediate experience. The more obvious and historically prior view, at least approximated by Descartes and Locke, is *representationalism* or *representative realism*: the view that our subjective sensory experience (and the beliefs that we adopt on the basis of it) constitute a *representation* of the external material world, one that is caused by that world and that we are justified, on the basis of something like a causal or explanatory inference, in thinking to be at least approximately accurate. The second main view is that (i) we can have no knowledge (or perhaps even no intelligible conception) of a realm of external causes of our experience, but also (ii) that our beliefs about the material world can still be in general justified and true

because their content pertains only to the features and order of our subjective experience. This is the view that has come to be known as *phenomenalism*.[6] It will be convenient to begin with the latter of these two views, which was widely held for at least a good part of the 20th century.

2.1 Phenomenalism

As just briefly formulated, the phenomenalist view is that the content of propositions about material objects and the material world is *entirely* concerned with features and relations of the immediate objects of our perceptual experience, that is, the features and relations of our sense-data.[7] According to the phenomenalist, to believe that a physical or material object of a certain sort exists is just to believe that sense-data of various sorts have been experienced, are being experienced, will be experienced, and/or would be experienced under certain specifiable conditions. Thus, for example, to believe that there is a large brown table in a certain room is to believe, roughly, (i) that the sorts of sense-data that seem from a common-sense standpoint to reflect the presence of such a table either have been, are presently, or will in the future be experienced in the context of other sense-data, themselves experienced concurrently or immediately before or after, that reflect the location as the room in question; and in addition -- or instead, if the table has never in fact been perceived and never in fact will be perceived -- (ii) that such sense-data *would* be experienced *if* other sense-data that reflect the perceiver's going to that room were experienced.

In a fairly standard formula, to believe that such a material object exists is, according to the phenomenalist, to believe nothing more than that sense-data of the appropriate sort are actual (in the past, present, or future) and/or possible -- where to say that certain sense-data are *possible* is to say, not just that it is *logically* possible for them to be experienced (which would apparently always be so as long as the description of them is not contradictory), but that they would in fact be experienced under certain specified circumstances (themselves specified in sense-datum terms); thus it would be clearer to speak of actual and *obtainable* sense-data. John Stuart Mill put this point by saying that material objects are "permanent possibilities of sensation,"[8] that is, of sense-data -- where, of course, the possibilities in question are only *relatively* permanent, since objects can change or be destroyed.

The main argument for this commonsensically implausible view derives from Hume [1739-40]. One premise is the Humean idea that causal relations can be known only by experience, so that there is no way in which a causal relation between the immediate content of experience and something outside that immediate content could be known (and hence also no way to justifiably invoke such external causes as explanations of that experience). The other main premise is simply the common-sense conviction that skepticism is false, that we do *obviously* have justified beliefs and knowledge concerning ordinary objects like trees and rocks and buildings and about the material world in which they exist. And the argument is then just that the only way that such justified beliefs and knowledge are possible, given that no causal or explanatory inference from immediate experience to material objects that are genuinely external to that experience could ever be justified, is if the content of our beliefs about the material world does not really have to do with objects existing outside our immediate experience, but instead pertains just to that experience and the order that it manifests. Most phenomenalists will admit that this seems initially implausible, but will try to argue that this apparent implausibility is in some way an illusion, one

that can be explained away once the phenomenalist view and the considerations in favor of it have been fully understood.[9]

Objections to Phenomenalism

Many, many objections and problems that have been advanced in relation to phenomenalism. Here we will be content with a few of the most interesting ones:

Consider, first, what is perhaps the most obvious question about the phenomenalist view: *Why*, according to the phenomenalist, are the orderly sense-data in question obtainable or "permanently possible?" What is the *explanation* for the pattern of actual and obtainable sense experiences that constitutes the existence of a material object or of the material world as a whole, if this is not to be explained by appeal to genuinely external objects? The only possible phenomenalist response to this question is to say that the fact that sensory experience reflects this sort of order is simply the most fundamental fact about reality, not further explainable in terms of anything else. For *any* attempted further explanation, since it would obviously have to appeal to something outside of that experience, would be (for the reasons already discussed) unjustified and unknowable. (The phenomenalist will add that it is obvious anyway that not everything can be explained, since each explanation just introduces some further fact for which an explanation might be demanded.)

But it seems both quite implausible to suppose that something as large and complicated as the total order of our immediate experience has no explanation at all -- and also very obvious that common sense (at least if it accepted perceptual subjectivism) would regard claims about material objects as providing such an explanation, rather than as just a redescription of the experiential order itself (as the phenomenalist claims). Perhaps, for all we have seen so far, the phenomenalist is right that we cannot ever know that any such explanation is correct, but this, if so, is an argument for *skepticism* about the material world, not a justification for perversely reinterpreting the meaning or content of claims about material objects. (Here it is important to be clear that phenomenalism is *not* supposed to be a skeptical view, but rather an account of how beliefs about material objects are indeed justified and do constitute knowledge -- given the phenomenalist account of the content of such beliefs.)

A second problem (or rather a set of related problems) has to do with the specification of the conditions under which the various sense-data that (according to phenomenalism) are what a material-object proposition is about either are or would be experienced. It is clear that such conditions must be specified to have even a hope of capturing the content of at least most such propositions in sense-datum terms. To recur to our earlier example, to say merely that the sense-data that are characteristic of a brown table are actual or obtainable in some circumstances or other may perhaps capture the content of the claim that the world contains at least one brown table (though even that is very doubtful), but surely not of any more specific claim, such as the one about such a table being in a particular room. For that, conditions must be specified that say, as it were, that it is in relation to that particular room that the sense-data are or would be experienced. (But for the phenomenalist, the room does not of course exist as a mind-external place; talk of a room or of any physical location is to be understood merely as a way of indicating one aspect of

the order of immediate experience.)

What makes this problem extremely difficult is that for phenomenalism to be a viable position, the conditions under which sense-data are experienced or obtainable must themselves be specifiable in terms of *other sense-data*, not in terms of material objects and structures such as the library or room in question. For the essential claim of phenomenalism is that the content of propositions about material objects can be *entirely* specified in terms of sense-data. If in specifying the conditions under which the actual and obtainable sense-data relevant to one material-object proposition would occur, it were necessary to make essential reference to other material objects, then the account of the content of the first proposition would not yet be completely in sense-datum terms. And if in specifying the conditions relevant to claims about those other material objects, still other material objects would have to be mentioned, and so on, then the phenomenalist account would never be complete. If the content of propositions about material objects cannot be given *entirely* in terms of sense-data, if that content involves essential and ineliminable reference to further such objects, then phenomenalism fails.

How then can the idea that sense-data are or would be observed in a certain location be adequately captured in purely sense-datum terms? The natural response, which was in effect invoked when the example was originally discussed, is to appeal to the idea of a *sensory route*: a series of juxtaposed and often overlapping sense-data that would be experienced in what we think of intuitively as moving to or approaching the location in question. But there are at least two serious problems about this answer, however. One is that there are normally *many* different sensory routes to a given location, depending on where one starts; and if the starting location is itself determined by a previous sensory route, then a regress threatens, in which the sensory conditions must go further and further back in time without ever reaching a place from which they can unproblematically begin. A second problem is that it seems clear that we can often understand the claim that a certain material object or set of objects exists at a certain physical location without having any clear idea of the relevant sensory route: for example, I understand the claim that there are penguins at the South Pole, but have no idea of the sensory route that I would have to follow to guarantee that I have reached the South Pole. (Note that it is a *guarantee* that is required, for otherwise the content of the claim in question as not been fully captured.)[10]

A related, but still much more difficult problem of what the phenomenalist can say about the content of propositions about material objects and events in the past, perhaps the very distant past. Under what sensory conditions would sense-data of a tree have to have been obtainable to make it true that there was a pine tree in the place now occupied by my house in 1000 B.C.? It is thus very doubtful that the sort of specification of conditions that the phenomenalist needs is possible in general.

A generalization of this objection is offered by Roderick Chisholm.[11] Chisholm argues that there is in fact *no* conditional proposition in sense-datum terms, however long and complicated the set of conditions in the "if" part, that is *ever* even part of the content of a material-object proposition. This is shown, he claims, by the fact that for any such sense-datum proposition, it is *always* possible to describe conditions of observation (including conditions having to do with the state of the observer) under which the sense-datum proposition would be false, but the material-object proposition might still be true. The idea here is

to describe various sorts of abnormalities pertaining to the conditions or the observer: for example, having followed the sensory route to the room in the library, I am suddenly struck blind or knocked unconscious or injected with a mind-altering drug at just the instant before I would experience the distinctive table sense-data, which thus are not experienced (or the lighting is so altered as to make it impossible to see the table or to make it look very different in color; or the table is dropped through a trap door in the floor, to be restored only after I leave; etc.). Chisholm's suggestion is that the only way to guarantee that the sense-data that are experienced reflect the object that is actually there is to specify the conditions in *material* terms. But in that case, for the reason already discussed, the phenomenalist project cannot succeed.

A third, somewhat related, but deeper problem[12] arises by reflecting that it is apparently a condition for the success of phenomenalism that the realm of sense-data have an intrinsic order of its own, one that can be recognized and described solely in terms of the sense-data themselves. For how could we (without invoking independent material objects) have any justification for thinking that further sense-data will, under various conditions, occur, except by finding regularities in those we actually experience and reasoning inductively? But does such an intrinsic order of sense-data really exist? It is obvious that our sense-data are not merely chaotic, but far less obvious that they have an order that can be captured without making reference to material objects. And this is not something that the phenomenalist can just assume, for it is utterly essential to his whole position.

A fourth and final objection to phenomenalism, one that is much simpler and more straightforward, concerns what the phenomenalist must apparently say about the knowledge of the mental states of people other than myself (or other than whoever is thinking about the issue). The whole thrust of the phenomenalist position, as we have seen, is that *any* inference beyond immediate experience is impossible, that claims that might seem to be about things outside of experience must, if they are to be justified and knowable, be understood as pertaining only to features and orderly patterns of that experience. But the mental states of *other* people, their experiences and feelings and conscious thoughts, are surely outside of *my* immediate experience. Indeed, to reach justified conclusions about what people distinct from me are genuinely thinking and experiencing apparently requires *two* inferences: first, an inference from my immediate experience of sense-data pertaining to their physical bodies to conclusions about those bodies; and then, second, an inference from the facts about those bodies thus arrived at to further conclusions about the minds and mental states of the people in question. *Both* of these inferences depend on causal relations that are, according to the phenomenalist, unknowable, because we cannot experience both sides, or in the second case even one side, of the relation; and thus neither inference, construed in that way, is justified according to the basic phenomenalist outlook.

What phenomenalism must apparently say here, in order to be consistent, is: (i) that the content of propositions about the conditions and behavior of other people's bodies, like that of all other material object propositions, pertains only to facts about *my* immediate experience; and (ii) that the content of further claims about the mental states associated with those bodies is only a further, more complicated and less direct description of, once again, *my* experience. Though the phenomenalist would perhaps resist putting it this way, the upshot is that *my* mind and mental states, including my immediate experience, is the only mind and the only collection of mental states that genuinely exist, with claims that are

apparently about other minds amounting only to further descriptions of this one mind and its experiences. This is the view known as *solipsism*. It seems clearly to be an absurd consequence, thus yielding a really decisive objection, if one were still needed, to phenomenalism.

2.2 Representationalism

If phenomenalism is indeed untenable, and assuming that we continue to accept perceptual subjectivism, then the only non-skeptical alternative apparently left is *representationalism*: the view, restating it a bit, that our immediately experienced sense-data, together with the further beliefs that we arrive at on the basis of them, constitute a *representation* or depiction of an independent realm of material objects -- one that we are in general, according to the representationalist, justified in believing to be true.

Defenses of representationalism have taken a variety of forms, but I will assume here that the best general sort of defense for such a view is one along the lines suggested, albeit not very explicitly, in Locke (and indeed also, though even less explicitly, in Descartes). The central idea is, first, that (contrary to the claim of the phenomenalist) some *explanation* is needed for the complicated and intricate order that we find in our (involuntarily experienced) sense-data (or adverbial contents); and, second, that the *best* explanation, that is, the one most likely to be correct, is that those experiences are caused by and, with certain qualifications, systematically reflect the character of a world of genuinely independent material objects, which we accordingly have good reasons for believing to exist.

It is this appeal to what has come to be referred to as "inference to the best explanation" that allegedly provides an answer to the Humean argument, by allowing the supposed causal and explanatory relation between material objects and sensory experience to be known or justifiably believed in despite the fact that it cannot itself be immediately experienced. A consideration of the merits of this general idea of explanatory or abductive inference is beyond the scope of the present article [see abduction]. Here we will be concerned solely with whether and to what extent representationalism can be defended, given the fairly widely accepted assumption that such reasoning is in general cogent.

The Representationalist Explanation of Experience

Before considering whether the representationalist explanation of sensory experience is really the *best* one, we need to consider in more detail what the rationale for that explanation might be.

The place to start is to ask what it is about the character of our immediate sensory experience that points to or perhaps even seems to demand such an explanation. In perhaps the earliest very explicit discussion of this issue, Locke[13] points to two features of our experience in this connection: (i) its involuntary character, i.e., the fact that it simply occurs without any choice or control on the part of the person having the experience; (ii) the systematic order or coherence of that experience. But while these features may indeed demand *some* sort of explanation, they do not, at least when described at that level of abstraction, point at all clearly at the specific one that the representationalist favors. Why shouldn't these features of experience be explained in some quite different way, such as the one proposed by Berkeley (and

considered by Descartes): by appeal to a deity or similar being who causes the experience in us?[14] If anything about experience does points to Locke's explanation rather than Berkeley's, it will have to be something more specific than the features mentioned by Locke.

My tentative suggestion is that there are at least two aspects of the order of experience that suggest and might indeed seem to demand the representationalist's explanation in terms of external material objects. The first is the presence in immediate experience of repeatable sequences of experienced qualities that overlap and often shade gradually into one another. Here I have in mind something like the "sensory routes" that were, as discussed earlier, invoked by the phenomenalist. While these "sensory routes" cannot ultimately do the job that the phenomenalist needs them to do, for the reasons given there, they are nonetheless very real and pervasive. Think of the ways in which such "sensory routes" can be experienced in opposite orders (imagine here what common sense would regard as walking from one place to another and then returning to the first place by the same route -- perhaps even walking backwards, so as to make the two sequences as similar as possible). Think of the ways in which such "sensory routes" intersect with each other, thus, for example, allowing one to get from one end to the other without going through the "route" itself, thereby delineating a sensory loop. Think of the resulting structure of a whole set of overlapping and intersecting "sensory routes."[15]

The idea is then that at least the most obvious and natural explanation of these features of our experience is that we are located in a 3-dimensional spatial realm of objects through which we move and of which we can perceive at any given moment only the limited portion that is close enough to be accessible to our various senses (what this requires differs from sense to sense). Our experience reflects both the qualities of these objects and the different perspectives from which they are perceived as we gradually approach them from different directions, at different speeds, under different conditions of perception, etc. It is thus the relatively permanent structure of this spatial array of objects that is reflected in the much more temporary and variable, but broadly repeatable features of our immediate experience.

The second, and even more important aspect of immediate experience that points to the representationalist explanation is the fact, already noticed in our discussion of phenomenalism, that the order just indicated, though undeniably impressive, is in fact *incomplete* or *fragmentary* in a number of related ways. The easiest way to indicate these is by reference to the sorts of situations that, from a common-sense standpoint, produce and explain them (though the representationalist cannot, of course, assume at this stage, without begging the question, that these situations are what is actually occurring). Imagine then traversing a "sensory route" of the sort just indicated, but doing so: (i) with one's eyes closed (or one's ears plugged, etc.) during some of the time required, or perhaps while asleep during part of the time (traveling in a car or train); or (ii) while the conditions of perception, including those pertaining to the functioning of your sense-organs and to your mental "processing", are changing or being varied (involving such things as changing lighting, including complete darkness; jaundice and similar diseases that affect perception; objects and conditions that temporarily block or interfere with perception; even something as simple as turning one's head in a different direction, blinking, or wiping one's eyes). Reflection on cases like these show that the sensory sequences that define the various "sensory routes" are in fact substantially less regular and dependable than they might at first seem.

Thus the representationalist can in effect argue that the realm of immediate sensory experience, of sense-data (or adverbial contents), is both *too* orderly not to demand an explanation and *not orderly enough* for that explanation to be that the sense-data have an intrinsic order of their own. What this strongly suggests, the representationalist will argue, is an independent realm of objects outside our experience, having its own patterns of (mainly spatial) order, with the partial and fragmentary order of our experience resulting from our partial and intermittent perceptual contact with that larger and more stable realm.

The discussion so far provides only an initial and highly schematic picture of the representationalist's proposed explanation. It would have to be filled out in a number of ways in order to be even approximately complete. Here I will be content with three further points. First, the main focus of the discussion so far has been on *spatial* properties of material objects and the features of immediate experience that seem to suggest them. Thus the result so far is at best only a kind of skeletal picture of the material world, one that would have to be "fleshed out" in various ways in order to even approximate the common-sense picture of the world. In fact, it is useful to think of the representationalist explanation as starting with spatial properties as a first and most fundamental stage and then adding further refinements to that starting point.

Second, the most important addition to this initial spatial picture of the world would be various sorts of *causal* relations among material objects and between such objects and perceivers, together with the causal and dispositional properties of objects (flammability, solubility, malleability, brittleness, toxicity, etc., etc.) that underlie such relations. These are warranted by the need to explain apparent changes in material objects that are reflected in relatively permanent changes in the otherwise stable "sensory routes". Here it is important to note that like the stable spatial order, the causal regularities that pertain to material objects are only intermittently and fragmentarily reflected in immediate experience, partially for the reasons already considered, but also because any given perceiver may simply not be in the right position to observe the beginning or end or some intermediate part of a given causal sequence, even though other parts are experienced. Simple examples would include throwing a rock into the air without seeing or hearing it land, pulling on a string without observing the movement of an object at the other end (or seeing the object move but without observing the movement of part of the intervening string), or planting a seed and returning later to find a well-developed plant.[16]

Third, there is the issue of primary and secondary qualities. Locke's view was that material objects have *primary* qualities like size, shape, and motion through space, but not *secondary* qualities like color, smell, taste, and felt temperature; and most other representationalists have tended to agree with him on this. Here it will suffice to focus on color, surely the most apparently pervasive and interesting of the secondary qualities.[17] Clearly the denial that material objects are genuinely colored seriously complicates the representationalist's proposed explanation by making the relation between material objects and our immediate experiences much less straightforward than it would otherwise be: according to such a view, while our immediate experiences of spatial properties are caused more or less directly by closely related spatial properties of objects (allowing, importantly, for perspective), our immediate experiences of color properties are caused by utterly different properties of material objects, primarily by

how their surfaces differentially reflect wavelengths of light.

Locke offers little real argument for this view, but the argument he seems to have in mind is that as the causal account of the material world develops, it turns out that ascribing a property like color (construed as the "sensuous" property that is present in immediate visual experience) to material objects is in fact quite useless for explaining our experiences of colors. What colors we experience depends on the properties of the light that strikes our eyes and this in turn, in the most standard cases, depends on how material objects reflect and absorb light, which yet in turn depends on the structure of their surfaces as constituted by primary and causal properties. I think that this is correct as a matter of science, but the important point for the moment is that *if* it is correct, then the denial that material objects are really colored simply follows from the basic logic of the representationalist position: according to representationalism, the only justification for ascribing *any* property to the material world is that it is required to explain some aspect of our immediate experience, so that the ascription of properties that cannot figure in such explanations is automatically unjustified.

Alternatives to the Representationalist Explanation

The discussion so far has perhaps made a reasonable case, though of course nothing like a conclusive one, that the representationalist's proposed explanation of the order of our immediate experience has a good deal of plausibility in relation to that experience. But this is still not enough to show that it is the *best* explanation and hence the one, assuming the general acceptability of theoretical reasoning, whose acceptance is thereby justified. Why, if at all, should the explanation of our experience that invokes external, mind-independent material objects be preferred to other possible explanations such as Berkeley's view that our sensory experience is caused by God, who shapes and maintains the order that it reflects (or the very similar if not identical one that appeals to Descartes's evil genius)?

It should be obvious that Berkeley's explanatory hypothesis is capable of explaining the very same features of immediate experience that the representationalist appeals to. All that is needed is for God to have an ideally complete conception or picture of the representationalist's material world and then to systematically cause experiences in perceivers that reflect their apparent location in and movement through such a world. (This assumes that God can recognize intentions to "move" in various directions and adjust the person's perceptions accordingly; of course, no genuine movement really takes place, nor does the perceiver really have a physical location.) A different, but essentially parallel explanatory hypothesis, is provided by a science fiction scenario: the perceiver is a disembodied brain floating in a vat of brain nutrients and receiving electrical impulses from a computer that again contains an ideally complete model or representation of a material world and generates the impulses accordingly, taking account of motor impulses received from the brain that reflect the person's intended movements. And further explanatory hypotheses can be generated according to the same basic formula: there must be some sort of a representation or model of a material world together with some sort of mechanism (which need not be mechanical in the ordinary sense) that systematically produces experience in perceivers, allowing for their subjectively intended movements. *Any* pattern of immediate experience that can be explained by the representationalist's explanatory hypothesis can thus automatically be also explained by explanatory hypotheses of this latter sort, probably indefinitely many of them, with no possible

experiential basis for deciding between them or between any one of them and the representationalist hypothesis.

If there is to be a reason for favoring the representationalist hypothesis, it will therefore have to be *a priori* in character, and it is more than a little difficult to see what it might be. Here I will limit myself to one fairly tentative suggestion.

One striking contrast between the representationalist's explanatory hypothesis and the others we have looked at is that under the representationalist view there is a clear intuitive sense in which the qualities of the objects that explain our immediate experience are reflected in the character of that experience itself, so that the latter can be said to be, allowing for perspective and perhaps other sorts of distortion, experiences of the former, albeit indirect ones. Once again this applies most straightforwardly to spatial properties: thus, for example, the rectangular or trapezoidal shape that is immediately experienced can be said to be an indirect perception of a rectangular face of the material object that causes that experience. In contrast, the features of the elements in the other explanatory hypotheses that are responsible for the various features of our experience are not directly reflected in that experience. For example, what is responsible in these other hypotheses for the rectangular or trapezoidal shape in my immediate experience is one aspect of God's total picture or conception of a material world, or perhaps one aspect of a representation of such a world stored in a computer. This aspect has in itself no shape of any sort (or at least, in the case of the computer, none that is at all relevant to the shape that I experience); it is merely a *representation* of a related shape, according to some system of representation or coding. Thus its relation to the character of the experience that it is supposed to explain is inherently less direct, more convoluted than is the case for the representationalist's explanation.

My tentative suggestion is that the inherently less direct character of the way that these competing explanatory hypotheses account for the features of our immediate experience may yield a reason for preferring the more direct and thus in a sense simpler representationalist explanatory hypothesis, for regarding it as more likely to be true. But how, exactly?[18] The idea is that an explanatory hypothesis like Berkeley's, at least as we have construed it, depends for its explanatory success on the truth of two equally essential claims: first, the claim that a material world of the sort postulated by the representationalist *could* account for the features of our experience, for it is precisely by emulating or mimicking the action of such a world that God (or the computer) decides just what experiences to produce in us; and, second, that God (or the computer) can indeed successfully produce the required emulation. But the representationalist view requires only the truth of the first of these two claims. It is thus, I suggest, inherently less vulnerable to problems and challenges and so more likely to be true. And this is an apparent reason for regarding the representationalist's explanatory hypothesis as providing the best of these competing explanations.

Is this a successful argument for representationalism? There are at least two questions about it that need to be considered. First, the argument assumes that the competitors to representationalism are all parasitic upon the representationalist explanatory hypothesis in the way indicated, and it is worth asking whether this is really so. Is there an explanation of our immediate experience that does not in this way rely on an

emulation of the way in which a material world would produce that experience? Second, even if the argument succeeds to a degree, *how* probable or likely does it make the material world hypothesis in comparison to these others? Is the resulting degree of probability or likelihood high enough to agree approximately with our common-sense convictions in this regard?

2.3 Direct Realism

The upshot of our discussion so far is that phenomenalism appears entirely untenable, and that at least a better defense than many have supposed possible can be offered for representationalism. Many recent philosophers, however, have thought that there is a *third* alternative that is superior to either of these: one usually referred to as *direct realism*. The central idea of direct realism is that the view we have called perceptual subjectivism is false, that is, that instead of immediately experiencing either sense-data or adverbial contents, we instead *directly* experience external material objects, without the mediation of these other sorts of entities. And the suggestion often seems to be, though this is usually not explained very fully, that such a view can simply bypass the representationalist's problem of justifying the inference from immediate experience to the material world and do so without having to advocate anything as outlandish as phenomenalism.[19]

For anyone who has struggled with the idea of sense-data (or the adverbial alternative) and with the difficulties and complexities of representationalism and phenomenalism, the apparent simplicity of direct realism, the way in which it seems to make extremely difficult or even intractable problems simply vanish, may be difficult to resist. We must be cautious, however. What does such a view amount to, and can it really deliver the results that it promises?

We may begin with a point that is often advanced in arguments for direct realism, one that turns out however to be of much less help than has sometimes been thought in either defending or even explaining the view. Think about an ordinary example of perceptual experience: standing in my back yard, I watch my dogs chasing each other in a large circle around some bushes, weaving in and out of the sunshine and shadows, as a car drives by on the street. The direct realist's claim is that in such a case (assuming that I am in a normal, non-philosophical frame of mind), the picture that it is easy to find in or read into some representationalists, according to which I *first* have thoughts or occurrent beliefs about the character of my experience (whether understood in sense-datum or in adverbial content terms) and *then* infer explicitly from these to thoughts or beliefs about material objects is simply and flatly wrong as a description of my actual conscious state. In fact, the only things that I think about at all *directly* and explicitly in such a case are things like dogs and bushes and cars and sunlight, not anything as subtle and abstruse as sense-data or adverbial contents. The direct realist need not deny (though some have seemed to) that my sensory experience somehow involves the various qualities, such as complicated patterns of shape and color, that these other views have spoken of, or even that I am in *some* way aware of conscious of these. His point is that whatever may be said about these other matters, from an intuitive standpoint it is material objects and nothing else that are "directly before my mind" -- and that any view that denies this obvious truth is simply mistaken about the facts.

Almost everyone will agree that the direct realist is right about this. What happens most centrally in perceptual experience is that we have explicit thoughts or "perceptual judgments" about what we are perceiving; and in normal cases (apart from very special artistic or perhaps philosophical contexts), these perceptual judgments are directly and entirely about things (and processes and qualities) in the external material world. Philosophers speak of that which a propositional state of mind is directly about as its *intentional object*, and we can accordingly say that the intentional objects of our basic perceptual judgments are normally material objects. In this way, the relation of such judgments to material objects is, it might be said, *intentionally* direct.

But what bearing, if any, does this intentional directness have on the central epistemological question of what reason or justification we have for thinking that such perceptual judgments about the material world are *true*? *Perhaps* the sort of direct presence to the mind that is involved in the idea of immediate experience considered earlier yields the result that one's beliefs or awarenesses concerning the objects of such experience are automatically justified, simply because there is no room for error to creep in. But is there any way in which it follows from the mere fact that perceptual judgments about material objects are *intentionally* direct that they are also *justified*? It still seems obvious that both a perceptual judgment and the total state of mind of which it is a part are quite distinct from the material object, if any, that is its intentionally direct object. This is shown by the fact that in cases like hallucination, the object in question need not exist at all, but it would be clear enough even without such cases -- phenomenalist views having been rejected, the material object does not somehow literally enter the mind. Thus even though perceptual judgments are directly about such objects in the intentional sense, the question of whether they represent them correctly still arises in exactly the same way that it does for the representationalist. And this question must apparently still be answered, if at all, by appeal to the immediately experienced features of that state of mind, with the specific character of the sensory experience being the only obvious thing to appeal to.

Thus while the idea of intentional directness contributes to a view that presents a somewhat more accurate picture of the perceptual state of mind, the view that results seems to still be fundamentally a version of representationalism in that it faces the same essential problem of justifying the transition (whether it is an explicit inference or not) from the character of the person's experience to beliefs or judgments about the material world. If this is all that direct realism amounts to, then it is not a genuinely distinct third alternative.

Is there any further way to make sense of the "directness" to which direct realist appeals, one that might yield more interesting epistemological results? It is far from obvious what it would be. Some proponents of this supposed view have tried to deny that we have any awareness of the character of our immediate experience that is both distinct from our judgmental awareness of material objects and of the sort that could provide the basis for the justification of material object claims. Such a challenge raises subtle and difficult issues about different kinds of awareness, but it is hard to see how it could really be correct. Moreover, the correctness of this challenge, while it would surely constitute a serious or perhaps even conclusive objection to representationalism, would not yield in any obvious way a positive direct realist account of how beliefs about the material world are justified, if not in the representationalist way.

My tentative conclusion is that the idea that direct realism represents a further alternative on the present issue is a chimera. Thus, once phenomenalism is rejected as hopeless, the only alternatives with regard to knowledge of the external world appear to be skepticism and some version of representationalism, perhaps one that recognizes and incorporates the view that perceptual judgments about the material world are intentionally direct.

2.4 Conclusion

The following two conclusions seem to me to be fairly strongly supported by the foregoing discussion. First, some version of perceptual subjectivism is probably correct, with the adverbial theory being the more promising of the two main alternatives. Second, assuming that perceptual subjectivism is indeed correct, representationalism is by far the most defensible non-skeptical account of how beliefs concerning material objects are justified. But there are many philosophers who regard direct realism as a more promising alternative than has been suggested here. And there are perhaps many more (though not the present author) who believe that some epistemological view at a different level, such a a coherentist or contextualist theory of justification, can circumvent these problems.

3. Externalist Theories of Perception

As noted at the beginning, the foregoing discussion is entirely concerned with the views and issues that arise under an *internalist* account of epistemic justification. The adoption of an *externalist* account of justification alters and greatly simplifies the account of the justification of perceptual beliefs. Here I wll focus on the most standard version of exernalism, namely *reliabilism*, according to which a belief is justfied if it results from a reliable causal process.[20]

On such a reliabilist view, the justification of a perceptual belief depends only on the reliability of the perceptual process that produces it, that is, on the fact (*assuming that it is a fact*) that this process leads to a suitably high proportion of true beliefs. (Note that it is *not* required that the believer or anyone else know that the process is reliable or have any sort of cognitive access to its reliability -- all that is required it is that is in fact reliable.) The justification of such a belief thus requires no appeal to sensory experience at all, thus effectively short-circuiting the issue that divides representationalism and phenomenalism. Such reliabilist views might in a way be viewed as versions of direct realism, but it is less misleading to simply regard them as rejecting the issue which all three of the more traditional theories attempt to respond to: the issue of how sensory experience provides a reason for thinking that perceptual beliefs are true.

Reliabilism thus offers a very sraightforward and seemingly unproblematic account of how perceptual beliefs about physical objects and the physical world are, in a sense, justified -- again, on the assumption that our perceptual processes are in fact reliable in the way that we take them to be. But it is important to realize that even if this last assumption is correct, the reliabilist account is quite compatible with our having no reason at all for thinking that such processes are indeed reliable and so no reason at all for

thinking that our perceptual beliefs are either justified or likely to be true. It is for this reason that the ease with which the standard issues are avoided on a reliabilist view may seem to constitute a defect rather than a virtue of the position.

Bibliography

- Alston, William P., *The Reliability of Sense Perception* (Ithaca, NY: Cornell University Press, 1993).
- Armstrong, D. M., *Perception and the Physical World* (London: Routledge, 1961).
- Ayer, A. J., 1946-7. "Phenomenalism," *Proceedings of the Aristotelian Society*, vol. 47 (1946-7), pp. 163-96.
- Austin, J. L., 1962. *Sense and Sensibilia* (London: Oxford University Press).
- Berkeley, George, 1710. *Principles of Human Knowledge*, many editions.
- Berkeley, George, 1713. *Three Dialogues concerning Hylas and Philonous*, many editions.
- BonJour, Laurence, 2000. "Foundationalism and the External World," in James Tomberlin (ed.), *Philosophical Perspectives*, vol. 13 (London: Blackwell).
- Broad, C. D., 1923. *Scientific Thought* (London: Routledge & Kegan Paul), Part II.
- Chisholm, Roderick, 1957. *Perceiving* (Ithaca, N.Y.: Cornell University Press).
- Descartes, Rene, 1641. *Meditations on First Philosophy*, many editions.
- Dretske, Fred, *Seeing and Knowing* (London: Routledge, 1969).
- Dretske, Fred, *Knowledge and the Flow of Information* (Oxford:Blackwell, 1981).
- Goldman, Alvin, *Epistemology and Cognition* (Cambridge, Mass.: Harvard Univerity Press, 1986).
- Hume, David, 1739-40. *A Treatise of Human Nature*, many editions.
- Jackson, Frank, 1977. *Perception: A Representative Theory* (Cambridge: Cambridge University Press).
- Locke, John, 1689. *An Essay concerning Human Understanding*, many editions.
- Mackie, J. L., 1976. *Problems from Locke* (Oxford: Oxford University Press), ch. 2.
- Mill, J. S., 1865. *An Examination of Sir William Hamilton's Philosophy* (London: Longmans, Green), ch. 11 and appendix to ch. 12.
- Moore, G. E., 1953. *Some Main Problems of Philosophy* (London: Allen & Unwin), chapter 2.
- Pitcher, George, *A Theory of Perception* (Princeton: Princeton University Press, 1971).
- Price, H. H., 1950. *Perception*, 2nd ed. (London, Methuen).
- Robinson, Howard, *Perception* (London: Routledge, 1994).
- Russell, Bertrand, 1912, *The Problems of Philosophy* (London: Oxford University Press, 1912), chapters 1-5.
- Sellars, Wilfrid, 1963. "Phenomenalism," in Sellars, *Science, Perception and Reality* (London: Routledge & Kegan Paul), chapter 3.
- Whiteley, C. H., 1959. "Physical Objects," *Philosophy*, vol. 34 (1959).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Berkeley, George | Hume, David | justification, epistemic: coherentist theories of | justification, epistemic: contextualist theories of | justification, epistemic: foundationalist theories of | justification, epistemic: internalist vs. externalist conceptions of | Locke, John | perception

Laurence BonJour
bonjour@u.washington.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: July 12, 2001
Content Last modified:

Stanford Encyclopedia of Philosophy
Notes to Epistemological Problems of Perception

Notes

1. Yet another way to refer to this same basic idea is to speak of objects with which we are acquainted (or, sometimes, directly acquainted). See Russell [1912], chapter 5.

2. For a fairly comprehensive but rather one-sided discussion of responses to the argument from illusion, see J. L. Austin [1962].

3. A further advantage often claimed for the adverbial theory is that it is compatible with materialist views of the mind: while it is clear that the brain does not contain entities having the properties ascribed to sense-data and at best obscure how it could stand in a relation of apprehension to such entities, there is no clear reason why a state of being appeared to silver-elliptical-ly could not just be a brain state. (To which it might be responded: (i) that we have no real understanding of how it could be a brain state either, of what features of a brain state would make it such a state of being appeared to; and (ii) that the absence of any clear difficulty here is simply a reflection of the obscurity of the nature of the supposed adverbial state.)

4. This term is my own coinage and not one that is generally used.

5. One major proponent of the sense-datum theory has advanced the argument that the adverbial theory cannot adequately capture all of the cases that can be described in terms of sense-data, in particular that it cannot adequately describe cases in which we experience a number of different apparent objects having a variety of different properties in a way that keeps straight which object has which property. Thus compare a case in which I am experiencing a red circle and a green square with one in which I am experiencing a green circle and a red square. In both cases, I might be said to be sensing or to be appeared to red-and-green-and-round-and-square-ly, thus apparently failing to capture the distinction between the two cases. And the suggestion is that only the sense-datum theory can successfully distinguish what is going on in such cases, by making explicit reference to each of the apparent objects. See Jackson [1977], pp. 64-68. But this objection seriously underestimates the resources available to the adverbial theory. In the cases in question, the adverbialist can say that I sense red-circle-and-green-square-ly in the first case and green-circle-and-red-square-ly in the second case, thus capturing the difference. More generally, if it is possible to capture the content of a particular immediate experience adequately in sense-datum terms, as the proponent of sense-data must agree that it is, then the adverbialist can construct a description that is equally adequate insofar as the present issue is concerned by simply making the entire sense-datum description the basis for his adverbial modifier, that is, by saying that the person is sensing or being appeared to [such and such sense-data]-ly, with the appropriate

sense-datum description going into the brackets.

6. It is harder to identify the historical sources of phenomenalism. Contrary to what is often suggested, Berkeley's "idealist" view (in Berkeley [1710] and [1713]) is not in fact in any clear way an anticipation of phenomenalism, but rather in effect a curious version of representationalism, in which our perceptual ideas constitute partial representations of the much more complete picture of the material world constituted by God's much more complete ideas. The phenomenalist view is at least suggested, but never quite arrived at in Hume [1739-40], Part IV, Sections 2-4. Perhaps the earliest reasonably clear statement is Mill [1865].

7. Or the features reflected in immediately experienced adverbial contents. But, as noted above, I will mostly leave this alternative possibility to be supplied by the reader.

8. Mill [1865].

9. There is also, as briefly mentioned earlier, a second fairly widely advocated argument for phenomenalism, one that starts from the premise that all intelligible ideas or concepts are derived by "abstraction" from immediate experience, so that we arguably could not even understand the idea of objects existing outside of that experience. If this were so, and if (as again seems obvious) we do understand the idea or concept of a physical or material object, then it would follow that this idea or concept is not about trans-experiential objects, but can only be about some feature or aspect of experience itself. The problem with this argument is that the initial premise about the derivation of concepts is far less obviously correct than is the claim that we do in fact obviously have ideas or concepts, indeed lots of them, that are about things outside immediate experience, making it far more reasonable to reject the conclusion than to accept the premise.

10. See Ayer [1946-7] for a discussion of this example.

11. Chisholm [1957], Appendix.

12. Advanced in Sellars [1963].

13. Locke [1689], Book IV, chapter 11.

14. Berkeley [1710] and [1713].

15. For a much, much more extensive discussion of this general sort of point, see Price [1950]. A very condensed summary of Price's account is offered in BonJour [2000].

16. For a good discussion of the general representationalist argument, especially with reference to the

causal regularities in the material world, see Whiteley [1959]. Whiteley, however, eventually arrives at a more extreme view according to which material objects explain our experience but cannot be known to have *any* of the properties actually manifested in that experience.

17. Locke adds solidity to the list of primary qualities. What he has in mind by this is not entirely clear, but solidity seems to be either the feeling of resistance that a rigid object produces when touched (in which case, it seems to belong with the secondary qualities) or else the causal capacity of preventing other objects from occupying the same space (in which case it is a causal property, what Locke calls a "power," and again not a primary quality on a par with the others, all of which are directly reflected in experience).

18. Sometimes philosophers appeal in such discussions to a general standard of simplicity, according to which it is just a fundamental principle that the simpler explanation is more likely to be true. The problem with this is twofold: first, the justification or rationale for the principle in question is anything but clear; and, second, the way in which it would apply to the case with which we are concerned is at least debatable, since Berkeley's explanation, for example, might be claimed to be simpler on the grounds that it invokes only one entity, albeit an extremely complicated one, rather than the many objects that make up a material world.

19. In earlier discussions of these issues, views in the general direction of direct realism were often referred to as "naïve realism" and ascribed to unsophisticated common sense.

20. The classic presentation of externalist reliabilism is Goldman [1986]. See also Dretske [1981] and, for an earlier version of the same general approach with more explicit consideration of perception, Dretske [1969].

Laurence BonJour
bonjour@u.washington.edu

First published: July 12, 2001
Content last modified: July 12, 2001

Interpretation and Coherence in Legal Reasoning

The subject of legal reasoning appears to occupy the more practical end of the spectrum of jurisprudential theorising. Surely if anything matters in our attempts to understand law, it matters how judges do and/or should decide cases, and that we have an account which adequately explains and can perhaps be used to guide their activities. The recent history of legal philosophy abounds with many and various attempts to address these issues and others which have been viewed as falling within the ambit of legal reasoning. Is legal reasoning an activity which is exclusive to the adjudicative institutions of legal systems or is any reasoning about the law to be regarded as legal reasoning, no matter where or by whom it is undertaken? Does legal reasoning take on a special character when it is undertaken in courts and by judges? Are there special methods or modes of reasoning which are unique to or at least distinctive of the law, or is legal reasoning just like reasoning in any other sphere of human activity, distinctive only in the subject matter to which it is applied? This last question is particularly relevant to present concerns, as it is one task of this entry to discuss various views concerning whether and to what extent interpretation and coherence have a special role to play in legal reasoning, because of the nature of law itself.

After a brief clarificatory consideration of the ambit of the term, ‘legal reasoning’, the entry deals first with interpretation and then with coherence, and discusses various views concerning these concepts and their relevance for law. Throughout, the discussion focuses upon the role which interpretation and coherence play within legal reasoning, and the reasons why these concepts are regarded by some as being distinctive of reasoning about the law.

What Do Legal Theorists Mean By 'Legal Reasoning'?

This may seem like an easy question, for surely legal reasoning is simply reasoning about the law, or about how judges should decide cases. On closer inspection, however, our ease may evaporate, for both of these formulations are ambiguous, at least according to some ways of thinking about the law. Some legal theorists regard the questions, 'what is the law?', and 'how should judges decide cases?' as distinct questions with distinct answers (see e.g. Hart 1994; Kelsen 1967; Raz 1979 & 1994). That is to say, their accounts of law and their accounts of adjudication are not one and the same, and they contend that in settling disputes which come before them, the remit of judges is wider than merely trying to establish what the law is as regards the issues in the case at hand. In adjudication, such theorists claim, extra-legal considerations can come into play, and judges may have discretion to modify existing law or to fill in gaps where existing law is indeterminate. This being so, for some legal theorists, the first formulation above, that legal reasoning is reasoning about the law, is ambiguous between: (a) reasoning to establish the content of the law as it presently exists, and (b) reasoning from that content to the decision which a court should reach in a case which comes before it.

Moreover, the second formulation of the ambit of legal reasoning given above, i.e. that legal reasoning is about how judges should decide cases, is also ambiguous on some approaches to legal theory. This is because the answer to the question, "how should a court decide a case, reasoning from the existing law applicable to it?" (i.e. legal reasoning in the sense given in (b) above) and the answer to the question, "how should a court decide a case, all things considered?", may sometimes come apart. A particular instance might be the kind of situation which could arise for a judge in a 'wicked' legal system where the law on some issue is so morally odious that, all things considered, the judge should not decide the case according to the law at all, but rather should refuse to apply the law (see Hart 1958; Hart 1994, chapter 9, section 3; Raz 1994, essay 14. This possibility is also noted by Dworkin 1986, chapter 3, 101-108, in discussing whether the Nazis had law).

There are thus three things (at least, there may be others) which legal theorists could mean by legal reasoning: (a) reasoning to establish the existing content of the law on a given issue, (b) reasoning from the existing content of the law to the decision which a court should reach in a case involving that issue which comes before it, and (c) reasoning about the decision which a court should reach in a case, all things considered.

It should be noted that some legal theorists, most notably for present purposes Ronald Dworkin, do not carve up the questions and issues on this topic in the way outlined above. For Dworkin, when judges decide a case according to law, they do no more than ascertain the content of the law and apply it to the facts of the case. In other words, judges never resort to extra-legal considerations in deciding cases according to law: all the considerations which they are entitled to take into account are part of the law. This means that, according to Dworkin, when judges reason about the law in sense (b), what they are doing amounts to no more nor less than reasoning about the law in sense (a), i.e. reasoning to establish the content of the law (see Dworkin 1977 & 1986).

This entry is concerned with legal reasoning in senses (a) and (b), and with sense (b) in particular. It should be noted that the discussion does not directly address the different accounts of the nature and limits of law which are revealed by those varying views mentioned above regarding what it is that judges do when they reason about the law in sense (b). Where such differences have a bearing upon issues pertaining to the role of interpretation and coherence in legal reasoning, they will be mentioned in the text. For further discussion of the nature and limits of law, see various entries under *nature of law* in this volume.

2. The Role of Interpretation in Legal Reasoning:

2.1 Some Intellectual Roots Considered

Recent interest in the role of interpretation in legal reasoning springs from several sources. For some, interpretation is where we should look in order to find the solution, or at least the only possible response, to the problem of linguistic indeterminacy in law which they perceive (in turn, renewed interest in the problems of linguistic indeterminacy in law seems to have stemmed at least in part from the resurgence in the last twenty years in scholarship addressing Wittgenstein's remarks on rule-following in the Philosophical Investigations (see e.g. Holtzman & Leich (eds.) 1981; Kripke 1982; Baker & Hacker, 1984 & 1985), and the migration of these concerns from philosophy of language into philosophy of law (see e.g. Marmor 1992; Stone 1995; Smith 1990). Certain aspects of this trend, most notably that of academic lawyers employing ill-digested Wittgenstein to dubious ends, are criticised by Bix 1993). According to this line of thinking, if words, and legal rules composed of words, have no intrinsic meaning and hence cannot, in and of themselves, constrain legal reasoning, then it must be we -- readers or interpreters -- who supply such meaning via the process of interpretation (see Fish 1989; Cornell 1992. Stone 1995 criticises this understanding of the role of interpretation in legal reasoning, but notes its adoption by various legal theorists). In the case of other theorists, interest in this topic stems from a wish to investigate the parallels and divergences between interpretation in law and interpretation in literature (Levinson 1982; Dworkin 1985; Fish 1989). The arrival on the jurisprudential scene in the mid-1980s of Ronald Dworkin's powerful new account of law as an interpretive concept, with concomitant implications for the activities of both judges and legal theorists (see Dworkin 1986) also did much to contribute to interest in the role of interpretation in legal reasoning. Moreover, this account seemed to rouse Dworkin's legal positivist adversaries into elaborating more fully upon something which he has

always claimed has been seriously underdeveloped in their work (see e.g. Dworkin 1977), namely an account of the nature of adjudication, and of the role of interpretation within it (Marmor 1992; Raz 1995; Raz 1996a & 1996b).

2.2 An Initial View of the Nature of Interpretation: Conservation and Creativity

As might be expected, as a result of these different (although often intertwining) intellectual backgrounds and sources of interest in interpretation, legal theorists approach this subject with very different questions and concerns to which they give concomitantly different answers. For all this, however, a surprising number of legal theorists agree -- at least at an abstract level -- about one central characteristic of interpretation, namely that interpretation is a Janus-faced concept, encompassing both a backward-looking conserving component, and a forward-looking creative one. In other words, an interpretation of something is an interpretation of *something* -- it presupposes that there is a something, or an original, there to be interpreted, and to which any valid interpretation must be faithful to some extent, thus differentiating interpretation from pure invention -- but it is also an *interpretation* of something, i.e. an attempt not merely to reproduce but to make something of or bring something out of an original. (See e.g. Fiss 1982; Dworkin 1986; Marmor 1992; Endicott 1994; Raz 1996b & 1996c.)

Much jurisprudential writing on interpretation in legal reasoning is concerned with how to strike the right balance between the conserving and creative elements in interpretation, and with the constraints which are and/or should be operative upon judges as they undertake this balancing act. Some theorists claim that such concerns about how one ought to interpret the law indicate that it is part of the way that we think about this practice that we regard rival interpretations as subject to objective evaluation as good or bad, better or worse, correct or incorrect (Dworkin 1986; Raz 1996b and 1996c). On this view, characterisations of interpretation which attempt to impugn the objectivity of such evaluations (e.g. Levinson 1982) are to be understood as revisionist accounts which attempt to persuade us that all is not as it appears to be with our practice of judging interpretations to be good or bad, better or worse, correct or incorrect as we currently understand it (see e.g. Raz 1996b).

2.3 Locating Interpretation in Legal Reasoning

It is important to consider how interpretation, as characterised in subsection 2.2 above, fits into the discussion of the ambit of the term legal reasoning in the opening section of this entry. The key to this issue lies in interpretation’s dualistic nature, i.e. that it has both a backward-looking conserving aspect and a forward-looking creative one. This dualism would seem to indicate that in interpreting the law, judges both seek to capture and be faithful to the content of the law as it currently exists, and to supplement, modify, or bring out something new in the law, in the course of reasoning from the content of the law to a decision in a particular case. In turn, this would seem to indicate that interpretation, because of its dualistic nature, has a role to play in both legal reasoning in sense (a), i.e. reasoning to establish the existing content of the law on a given issue, and legal reasoning in sense (b), namely reasoning from the existing content of the law to the decision which a court should reach in a case

involving that issue which comes before it.

One legal theorist who adopts exactly this approach, and so views interpretation in legal reasoning as ‘straddling the divide’ between identifying existing law, and developing and modifying the law, is Joseph Raz (see Raz 1996a and 1996b). According to Raz, the fact that interpretation has a role to play in both of these activities assists in explaining why we do not find a two-stage or clearly bifurcated approach to legal reasoning in judicial decisions. Judges do not first of all engage in legal reasoning in sense (a), having recourse only to legal materials, and then, having established what the existing law is and determined how far it can take them in resolving the instant case, then move on to a separate stage of legal reasoning in sense (b) which requires them to look to extra-legal materials in order to complete the job, because much of their reasoning is interpretive and interpretation straddles the divide between legal reasoning in senses (a) and (b). This point may assist Raz in defusing some of the criticisms which have been levelled at the legal positivist approach to legal reasoning such as that positivism’s account is phenomenologically inaccurate because when we examine cases, we do not find two distinct stages to judicial reasoning, one to establish whether any legal rules bear upon the problem at hand, and one wherein judges effectively legislate to fill in the gaps when the legal rules ‘run out’ (see e.g. Dworkin 1977 & 1986). As Raz himself notes, however (especially in Raz 1996b), this ‘straddling the divide’ approach may in fact seem to undermine the very ideas that there is a tenable distinction between legal reasoning in senses (a) and (b), and that there are gaps in the law. Interpretation appears to blur or even erase the line between the separate law-finding and law-creating roles which many legal positivists ascribe to judges, and the fact that courts always seem to be able to decide cases by interpreting the law may also seem to cast doubt on the idea that the law is incomplete, and hence that judges sometimes have to reach outside of the law in the adjudication process. Interest in the pervasiveness of interpretation in legal reasoning, and in the Janus-faced nature of interpretation may thus form part of the background which has led legal theorists like Dworkin to deny that the distinction between identifying existing law, and developing and changing the law, as understood by certain legal positivists, is a tenable or coherent one. That interpretation appears to operate at every stage in the legal reasoning process may also have influenced Dworkin’s denial that there are gaps in the law, and his counter-claim, contra Hart and Raz, that everything which a judge is entitled to rely on in deciding a case is already part of the law (see Dworkin 1986).

2.4 Some Points of Disagreement

Legal theorists disagree about the proper characterisation of many aspects of the schematic account of the interpretive process given in subsections 2.2 and 2.3 above. A few of these disagreements will be surveyed here in order to give a fuller picture of some of the issues and views which are relevant to this topic. Accounts of the role of interpretation in legal reasoning, then, differ regarding the following matters (n.b. several of the following points overlap to some extent):

(1) What exactly is the original or object which is being interpreted in the case of law: the law as a whole?; certain aspects of legal practice?; statutes?; judicial decisions?; authoritative legal decisions?; legal texts? Raz 1996b claims that the primary objects of interpretation are the decisions of legal authorities. He reaches this conclusion as a result of his view that law is an institutionalised normative

system wherein the institutions concerned operate by issuing purportedly authoritative directives concerning what ought to be done. According to Raz, then, the central role which authority plays in law means that when we come to interpret the law, what we are primarily seeking to do is to establish the existence and meaning of any purportedly authoritative directives of legal institutions, and it is, therefore, the decisions of those institutions which constitute the originals to be interpreted in the case of law. (See further the discussion of Raz's views in subsection 2.5 below). This stance can be contrasted with that adopted by Ronald Dworkin. The process of 'constructive interpretation' (Dworkin 1986, and see also the entry interpretivist theories of law) which plays such a central role in Dworkin's jurisprudential thought involves interpreters, 'imposing purpose on an object or practice in order to make of it the best possible example of the form or genre to which it is taken to belong.' (Dworkin 1986, p52). Moreover, in the case of legal interpretation, Dworkin appears to settle for the argumentative social practice of law as the original to be interpreted (Dworkin 1986, p63). This being so, for Dworkin, the object of legal interpretation appears to be broader than that adopted by Raz. For Raz, we interpret in order to establish whether any authoritative legal directives are currently in force and bear upon the legal issue at hand, and so we should look to the decisions establishing those directives in getting the interpretive process off the ground. Dworkinian interpretation, however, seems to have a more abstract and global feel to it, in the sense that it is the social practice of law as a whole (Dworkin 1986, 87-88), including the entire legal history of a given jurisdiction, as well as any data speaking to the point or purpose of legal practice in general, which constitutes the original to be interpreted.

(2) How much emphasis is to be placed on the conserving vs. the creative elements in interpretation? At the conserving extreme, we find accounts such as 'originalism' in US constitutional interpretation which claims that in interpreting a particular provision of the Constitution, judges should seek to establish the way in which the provision was originally understood by those who ratified it (see e.g. Bork 1990). As close as possible conformity with those original intentions thus furnishes us with the standard of correctness in constitutional interpretation on this approach. According to Bork 1990, adherence to originalism in US constitutional interpretation is necessary to ensure that the judiciary confines itself to its proper sphere of authority, thus preserving the separation of powers and structure of government in the form in which the founders of the US intended. Accounts such as that offered by Levinson (1982) reject originalism and give far more weight to the role of innovation in legal interpretation. Indeed, Levinson contends that all US constitutional interpretation is necessarily creative, due to radical and pervasive linguistic indeterminacy in the law. If such approaches wish to claim that there are standards by which we can judge interpretations to be better or worse, correct or incorrect, then they must find such standards other than in the conserving aspect of interpretation, as the requirement that one be faithful to the meaning of an original seems to be obliterated on such views. As was noted at the end of subsection 2.2 above, some such views may wish to claim that our practice of accounting interpretations as good or bad, correct or incorrect is incoherent. One other possibility, however, is that the notion of 'correctness' can be salvaged by being pegged to the communal reactions of the relevant interpretive community; our interpretations are the right ones when they accord sufficiently with those of our similarly situated fellow interpreters (see e.g. Kripke 1982 and those works discussed in point **(3)** below). Some commentators have poured scorn on the idea that this sort of approach could yield standards of correctness worthy of the name. See for example Baker & Hacker 1984, who contend that Kripke's position amounts to the view that, 'an unjustified stab in the dark is unobjectionable as long as it is made in good company.' (Baker

and Hacker 1982, 81-82).

(3) Following on from point **(2)** above: how big a role does the requirement that judges must be faithful to an original play in constraining legal interpretation, and are there any additional constraints which supplement the constraints generated by the need to be faithful to an original which guide judges as they interpret the law? For example, for Owen Fiss (1982), ‘disciplining rules’ in the form of those standards which are constitutive of the profession of judging supply constraints upon judicial interpretation which supplement the rules of language which already constrain all language users in their attempts to understand texts. According to Fiss, then, judges are constrained both by the need to be faithful to the original legal text which they are interpreting, and by supplementary norms of interpretation which are constitutive of the judicial role (Fiss lists the requirement that judges must always consult history when interpreting the law as an example of a ‘disciplining rule’). Fiss’ view is criticised by Stanley Fish (1989), who contends that Fiss’ ‘disciplining rules’ would themselves require interpretation in order for judges to know what they mean and require of them, and hence cannot supply constraints upon judicial interpretation. Fish’s contention that all potential candidates which might constrain interpretation are themselves susceptible to being interpreted in a variety of ways results in his claiming that texts or originals cannot constrain judges at all in the way in which is commonly supposed, as texts do not have meanings in advance of particular interpretations of them. This seemingly radical indeterminacy is deceptive, however, for although Fish removes the constraints on interpretation provided by legal texts or supplementary norms of the judicial profession, he replaces them with the conditioning and training processes of ‘interpretive communities’, which ensure that, ‘...readers are already and always thinking within the norms, standards, criteria of evidence, purposes and goals of a shared enterprise’, such that, ‘the meanings available to them have been preselected by their professional training.’ (Fish 1989, 133).

(4) Whether or not it is possible to have a general *theory* of interpretation. Raz 1996a rejects the possibility of two types of general theories of interpretation: ‘operational’ or recipe-like theories which are designed to guide judges to the right decision in a case which comes before them, and theories which, although they may not aim to guide judges to the correct decision, nonetheless claim to provide us with criteria via which to distinguish good interpretations from bad, and so enable us to check the correctness of decisions which have been made. According to Raz, theories of the former kind are impossible because morality (to which recourse must be had as regards the innovative aspect of legal interpretation) is not susceptible to explanation via ‘operational’ theories, i.e., ‘theories which would enable a person whose moral understanding and judgement are suspect to come to the right moral conclusions regarding situations he may face by consulting the theory.’ (Raz 1996a, p21. In this he follows similar claims made by several contemporary moral philosophers, e.g. Williams 1985; Dancy 1993). Moreover, claims Raz, theories which purport to tell us how to differentiate good interpretations from bad are also impossible because, by its very nature, innovation defies generalisation, such that it is futile to attempt to construct a general theory which differentiates good interpretations from bad as regards the forward-looking aspect of interpretation. It should be noted that (as is apparent from the discussions in this entry as a whole), Raz does believe that it is possible to have an account which explains certain aspects of the nature and role of interpretation in legal reasoning; what he doubts is that it is possible to have a certain kind of account of interpretation, namely an account in the form of a theory which purports either to operate as a recipe for concocting good interpretations, or to provide us with a general account of how to evaluate interpretations

as good or bad, right or wrong.

Ronald Dworkin, by contrast, does purport to offer judges a general theory of legal interpretation which they can use to guide their interpretive activities, and which, if followed correctly, will lead them to the 'one right answer' in the case before them (on Dworkin's 'one right answer' thesis, see further point **(7)** below). For Dworkin, it is the aim of all legal interpretation to 'constructively interpret' the social practice of law, by imposing purpose upon it such as, 'to make of it the best possible example of the form or genre to which it is taken to belong.' (Dworkin 1986, p52). The more specific theory which he believes that judges should follow in fulfilling this task -- 'law as integrity' -- 'instructs judges to identify legal rights and duties, so far as possible, on the assumption that they were created by a single author -- the community personified' (see subsections 2.5 and 3.4, and the entry [interpretivist theories of law]). It should be noted, however, that while Dworkin's general theory of interpretation is designed to assist in guiding judges to the one right answer in a case which comes before them, he claims that it is not recipe-like in the sense of providing judges with a detailed step by step programme for correct judicial decision-making: 'I have not devised an algorithm for the courtroom. No electronic magician could design from my arguments a computer program that would supply a verdict everyone would accept once the facts of the case and the text of all past statutes and judicial decisions were put at the computer's disposal' (Dworkin 1986, p. 412).

Dworkin's pro-theory stance has attracted criticism from a variety of quarters. Fish (1989, essays 4, 5 and 16) claims that 'law as integrity' is not a theory which judges can use to guide their interpretive activities because it is a strategy which they cannot help but put into practice, and to which they are always and already committed simply in virtue of their membership in the judicial interpretive community. Sunstein 1996 also warns against the kind of 'high-level theories' which Dworkin instructs judges to construct and follow in deciding cases. Sunstein is suspicious of the value of such theorising on the grounds that, 'it takes too much time and may be unnecessary; because it may go wrong insofar as it operates without close reference to actual cases; because it often prevents people from getting along at all; and because general theorizing can seem or be disrespectful insofar as it forces people to contend, unnecessarily, over their deepest and most defining moral commitments.' (Sunstein 1996, p50). Instead, Sunstein advocates a special role for 'incompletely theorized agreements' in judicial decision-making. Such agreements can occur where judges agree on the outcomes of individual cases even though they disagree on which general theory best accounts for those outcomes, or agree on a general principle, but not on what that principle requires in particular cases, or agree on a 'mid-level' principle (see Sunstein 1996, p36) but disagree about both the general theory underlying it and particular cases falling under it. Sunstein regards incompletely theorised agreements as vital to legal reasoning, because they allow the diverse individuals who constitute the judiciary to agree on outcomes against the background of certain institutional constraints such as that they have, 'a weak democratic pedigree and limited fact-finding capacity' (Sunstein 1996, p6), must make many decisions, make them fairly quickly, show adequate respect, to litigants, and to each other, and avoid error insofar as is possible. Sunstein presents a strong case for the role of incompletely theorised agreements in law in general, pointing out that the institution of such agreements is one of the most important social functions of legal rules, as rules are capable of allowing agreement in the face of disagreement, in the sense that sometimes judges can agree on outcomes to cases governed by a rule whilst disagreeing about the rule's justification (on this important function of legal

rules see also Raz 2001).

(5) Whether or not interpretation is always of something which to some extent already has meaning, or whether interpretation is the *fundamental* determinant of the meaning of linguistic expressions in, for example, legal texts. Marmor 1992 & Stone 1995 deny that interpretation is the fundamental determinant of the meaning of linguistic expressions and contend that, following a certain reading of Wittgenstein's remarks on rule-following (namely the kind of reading offered by McDowell 1984 and Baker and Hacker 1985), it must be possible for us to grasp the meaning of, for example, a legal rule, in a way which does not require recourse to interpretation. Cornell 1992 & Fish 1989 deny this, and contend that interpretation is all-pervasive, and is the fundamental and inescapable determinant of meaning, 'all the way down', in all cases. In the minds of some legal theorists at least, this point has strong links with with the issue of linguistic indeterminacy in law mentioned in subsection 2.1 above (see e.g. Fish 1989; Cornell 1992). Those theorists who contend that interpretation is the fundamental determinant of the meaning of linguistic expressions often claim that such interpretation is necessary because legal rules expressed in language do not have determinate meanings and hence cannot determine their own correct application; in John McDowell's terminology, those rules, by themselves, cannot have 'normative reach.' (McDowell 1984 & 1992). This being the case, so this line of thinking goes, interpretation is required in order to bridge the gap between the inert legal rule and the situations to which it applies. For these theorists, the pertinent issue then becomes: how do we know that one interpretation rather than another is in accordance with the rule, if the rule itself cannot determine its own correct application? As Wittgenstein notes in his remarks on rule-following, it seems that: ' 'Whatever I do is, on some interpretation, in accord with the rule'', such that we can, '...give one interpretation after another; as if each one contented us at least for a moment, until we thought of yet another standing behind it.' (Wittgenstein 1967, §198 and §201 respectively). Some theorists, e.g. Fish 1989, seem to embrace this potential infinite regress of alternative interpretations of rules (see also point **(3)** above) and attempt to avoid the radical linguistic indeterminacy which it seems to entail by replacing the standards of correctness demarcated by rules with the conditioning and training processes of interpretive communities. For theorists like Marmor (1992) and Stone (1995) who deny that interpretation is the fundamental determinant of the meaning of linguistic expressions, the Wittgensteinian challenge is seen as a kind of *reductio ad absurdum* which indicates that we have gone astray in our understanding of how rules operate. Such theorists seek to avoid linguistic indeterminacy, and reject interpretation as the fundamental determinant of meaning by denying that there is a gap which needs to be bridged between grasping a rule and understanding those actions which it requires. As was noted above, this denial usually proceeds via a non-sceptical reading of Wittgenstein's remarks on rule-following, along the lines of that offered by McDowell 1984 and Baker and Hacker 1985.

(6) Which values judges should attempt to realise in legal interpretation, and how those values are to be balanced against one another. One debate on this issue is that between Dworkin (1986), who champions the role of the value of integrity in interpreting the law, and those who, like Raz (1994a), and Réaume (1989), doubt whether Dworkinian integrity *is* a value which should be pursued in legal interpretation. The value of coherence in legal reasoning is addressed further in Section 3 of this entry.

(7) Whether interpretation in legal reasoning can lead judges to the 'one right answer' as regards the legal issue at hand. For example, Finnis 1987 denies that it is possible for interpretation in legal reasoning to

lead judges to one right answer in the sense claimed in Dworkin 1986, because of pervasive incommensurabilities in the criteria by reference to which we are supposed to adjudge one interpretation to be better than another. Finnis argues, *contra* Dworkin, that while we should seek good answers and avoid bad ones, we should not delude ourselves into dreaming of uniquely correct answers to issues of legal interpretation, for to do so commits us to, ‘utilitarianism’s deepest and most flawed assumption: the assumption of the commensurability of basic goods and thus of the states of affairs which instantiate them.’ (Finnis 1987, p. 375). Dworkin (1986 & 1991) remains firmly committed to the one right answer thesis, although it should be noted that in Dworkin 1986 chapter 11 he makes the point that in a sense there can be different ‘right answers’ for different interpreters: ‘For every route that Hercules took from that general conception to a particular verdict, another lawyer or judge who began in the same conception would find a different route and end in a different place, as several of the judges in our sample cases did. He would end differently because he would take leave of Hercules, following his own lights, at some branching point sooner or later in the argument.’ (Dworkin, 1986, p. 412).

2.5 Interpretation: Desirable or Necessary? or Why Is Legal Reasoning Interpretive?

The points of disagreement surveyed above speak to differing views regarding how judges should go about interpreting the law, and how we should understand their activities. Such concerns, however, do not directly address the important question of whether there is something about the nature of law which makes it either desirable or necessary that interpretation should play a role in legal reasoning in the first place. In other words: why is legal reasoning interpretive at all?

Raz 1996c contends that while some conventions of legal interpretation vary according to time and place, there are other features which legal interpretation necessarily exhibits, owing to the nature of law itself. While we can debate about the desirability of conventions of interpretation falling into the former category (e.g. we can consider the value of allowing the work of legal academics, or records of Parliamentary debates to serve as aids to interpretation in a particular jurisdiction), the latter category of features leave us no room for manoeuvre: courts cannot help but have recourse to them in interpreting the law. According to Raz, assigning a limited role to the intentions of legislators in the interpretation of legislation is one such necessary feature of legal interpretation. It is, he claims, simply part of our way of thinking about legislative institutions that their procedures and modes of operation are designed so as to allow legislators to make the law which they intend to make. To assume otherwise, Raz contends, is to render unintelligible any possible justification for entrusting law-making powers to those institutions. This being so, when judges come to interpret the decisions of legislative institutions, they must do so such that the law thus interpreted reflects the intentions of those who made it.

These considerations may also seem to speak mainly to the issue of how we are to go about interpreting aspects of the law. However, the reasons why it is important to pay attention to the intentions of law-making institutions when we interpret the law also furnish us with Raz’s answer to the question of why legal reasoning is interpretive at all. We pay attention to the intentions of law-making institutions because it is important to establish which legal rules those institutions have laid down, and what they mean. In

turn, it is important to establish the existence and meaning of legal rules laid down by law-making institutions because of law's purportedly authoritative nature. For Raz, legal institutions claim to express binding and authoritative judgements regarding what ought to be done which are designed to allow people to better conform to reason if they follow the decisions of the authority than if they try to follow those other reasons which apply to them directly (see Raz 1994, ch.10). In deciding cases according to law, then, we have a responsibility to try to establish the existence and meaning of any purportedly authoritatively binding legal rules which have a bearing on the situation under consideration, and we do so by interpreting the decisions of law-making institutions in a way which accords with the intentions of those institutions in making the decisions in question. For Raz, then, it is the authoritative nature of law which explains why legal reasoning is interpretive, whereas, for example, moral reasoning is not. Law, unlike morality, stems from social sources (on the role of social sources in understanding law, see Raz 1979 and the entry on legal positivism), from institutions issuing purportedly authoritative directives which claim to express a binding judgement about what ought to be done. Part of our task in reasoning about the law is thus to establish the existence and meaning of those directives, and, in order to do so, we must interpret the decisions of law-making institutions in accordance with the intentions of the law-makers in order to try to establish the content and meaning of the law which they intended to make (see also Raz 1996a and 1996b).

It is interesting to compare Raz's stance on the reasons why legal reasoning is necessarily interpretive with Ronald Dworkin's views on this topic. Rather than being based on the view that in ascertaining the content and meaning of the law, we should look to authoritative social sources, Dworkin's contention that legal reasoning is necessarily interpretive rests on an account of law which expressly repudiates the Razian understanding of law as source-based. According to Dworkin, the view that law is to be identified by reference to authoritative social sources yields a grossly inadequate account of the argumentative nature of legal practice, and of the nature and depth of disagreement within it (see Dworkin 1986 ch.1). He contends that an adequate account of these features of legal practice can only be gained when we understand that law is an interpretive concept, i.e. that it is a social practice wherein a certain interpretive attitude has taken hold. The attitude in question comprises two components: the assumption that the practice does not merely exist, but has a purpose or point, and the further assumption that the rules of the practice are not necessarily what they have always been taken to be, but rather are sensitive to, and can be revised in light of, its point (Dworkin 1986 ch.2). For Dworkin, then, it is these features of the social practice of law: that members of that practice dispute and disagree about what the best interpretation of the rules of the practice are, in light of its point, which dictate that legal reasoning is necessarily interpretive. Once the interpretive attitude has taken hold amongst the participants in a social practice, the only way to understand it adequately is to do as the participants in that practice do: i.e. join the practice and make the same kind of interpretive claims concerning the point of the practice, and what the rules of it are in light of that point, as they do. For Dworkin, this point holds good for the activities of judges and legal theorists alike: anyone reasoning about the law is required to treat it as an interpretive social practice and offer interpretations of what it requires in light of the purpose or point which they assign to it.

3. The Role of Coherence in Legal Reasoning:

As several commentators have noted (see Kress 1984; Marmor 1992; Raz 1994a), coherence theories, long influential in other areas of philosophy (see, for example, the entries on the coherence theory of truth and coherentist theories of epistemic justification) have more recently found their way into the philosophy of law. While this migration may be attributed in part to the frequent influence of the general philosophical climate upon the intellectual weather systems of jurisprudential theorising, it also makes sense to ask whether there is something about the nature of law which makes it particularly ripe for explanation via coherence accounts. For example, those commentators who view Ronald Dworkin's theory of law as integrity as a coherence account appear to answer this question in the affirmative (see e.g. Kress 1984; Hurley 1989): coherence, in the sense of interpreting the law as speaking with one voice as integrity requires, is a value which is supposed to have special relevance in the legal realm, in terms of the role which it should play in guiding judges seeking to interpret the law correctly. It has also been noted that features of the law such as the doctrine of precedent, arguments from analogy, and the requirement that like cases be treated alike seem particularly apt to be illuminated via some kind of coherence explanation. (See Kress 1984. Raz 1994a notes the temptation here, but contends that there is nothing inherent in arguments from analogy or in the requirement that like cases be treated alike which demands that they be understood in terms of a coherence account of adjudication. See also the entry precedent and analogy in legal reasoning.) Moreover, the idea of coherence as a special virtue of interpretation in legal reasoning plays an important role in the work of several major continental legal philosophers (see e.g. Peczenik 1989; Alexy 1989; Aarnio 1987; Alexy & Peczenik 1990).

The following discussion attempts to explore some of these issues concerning whether and why considerations of coherence have an important role to play in understanding law. As this entry seeks to illuminate the role of coherence in legal reasoning, the emphasis here is on coherence accounts of adjudication, and on examining the role which coherence plays in courts' reasoning about how to decide cases according to law. This being so, this part of the entry discusses legal reasoning in the sense outlined in formulation (b) in Section 1 ("What Do Legal Theorists Mean By 'Legal Reasoning'?"), i.e. reasoning from the content of the existing law on a given issue to the decision which a court should reach in a case involving that issue which comes before it.

3.1 What Constitutes Coherence?

Two central questions must be addressed in considering the role of coherence in legal reasoning: what is the nature of the coherence relation which features in coherence accounts of adjudication, and what role does coherence play in explaining or justifying judicial decisions in such accounts?

Amongst those legal theorists taking an interest in the role of coherence in legal reasoning, there is general agreement both that the coherence in question must amount to more than logical consistency amongst propositions (see Kress 1984; MacCormick 1984; Marmor 1992; Alexy & Peczenik 1990) and that it is not clear from many coherence accounts exactly what this something more amounts to (see Kress 1984; Peczenik 1989; Marmor 1992). MacCormick 1984 views coherence in terms of unity of principle in a legal system, contending that the coherence of a set of legal norms consists in their being related either in virtue of being the realisation of some common value or values, or in virtue of fulfilling some

common principle or principles. Raz 1994a also characterises coherence in law in terms of unity of principle. On his view, the more unified the set of principles underlying those court decisions and legislative acts which make up the law, the more coherent law is.

Other writers have attempted to supply a more formal definition of, for example, a minimally coherent legal system (see Levenbook 1984), or otherwise to flesh out in a more detailed manner the criteria of coherence. Alexy and Peczenik 1990 define coherence in terms of the degree of approximation to a perfect supportive structure exhibited by a set of propositions, and list ten criteria by reference to which coherence thus defined can be evaluated (the criteria are: (1) the number of supportive relations, (2) the length of the supportive chains, (3) the strength of the support, (4) the connections between supportive chains, (5) priority orders between reasons, (6) reciprocal justification, (7) generality, (8) conceptual cross-connections, (9) number of cases a theory covers, and (10) diversity of fields of life to which the theory is applicable). Such an approach raises many questions, such as how these various criteria of coherence are to be weighed and balanced against each other, and whether it is always the case that the weighing operation will result in a complete ranking of given sets of propositions as either more or less coherent than each other, so that when faced with competing such sets, it is always possible to find the most coherent set of propositions according to the ten criteria. Alexy and Peczenik recognise that weighing and balancing the criteria of coherence will be a complex matter, but appear to assume that it will always be possible to establish which is the most coherent of rival sets of propositions.

A further characterisation of the kind of coherence which is to be sought in legal reasoning may be found in Ronald Dworkin's work. Many writers regard Dworkin's account of integrity in adjudication as an example of a coherence account. (See Hurley 1989 & 1990; Marmor 1992. Kress 1984, although writing before Dworkin had fully developed his account of law as integrity, also views Dworkin as offering a coherence account of adjudication. Raz 1994a disputes the idea that Dworkin's account of law should be understood as a coherence account.) On this view, judges should try to realise the value of coherence in judicial decisions by interpreting the law as 'speaking with one voice', i.e. they should identify legal rights and duties on the basis that they were all created by a single author, the community personified.

3.2 Coherence of What?

The next issues to consider are (1) what is to be made coherent in coherence accounts of legal reasoning, and (2) what role coherence plays in explaining or justifying judicial decisions on such accounts. Regarding the question of what is to be made coherent in coherence accounts of legal reasoning, Raz 1994a contends that coherence accounts, when applied to law, require a 'base' or something which is to be made coherent, which differs in character in some crucial respects from the sort of base which features in coherence accounts in other areas of philosophy. Raz points out that while coherence accounts of justified belief take each person's belief set as their 'base' or as that which is to be made coherent, coherence accounts of law cannot be person-relative in this way, on pain of failing to offer an account which is in touch with the concrete reality of law in the jurisdiction under consideration. Raz's contention is that the law of a given jurisdiction does not vary with the beliefs of those subject to it, and in his view, that law is objective in this way means that there must be a common base to which coherence accounts in law are addressed. His suggestion in this regard is that coherence accounts in law take court decisions and

legislative and regulatory acts as their base, and hold law to be the set of principles that makes the most coherent sense of that base. Raz further distinguishes between coherence accounts of law and coherence accounts of adjudication. The essential difference between them is the stage at which considerations of coherence come into play. In the case of a coherence account of law, the whole of what the law *is* is determined by applying a coherence test to those court decisions and legislative and regulatory acts of a given jurisdiction. A coherence account of adjudication, however, accepts that the vagaries of politics and the influence of political considerations on legislative and judicial decisions make it unlikely that the settled law of a jurisdiction will exhibit coherence to any great extent. This being so, if we are to apply a coherence account in order to determine how judges ought to decide cases according to law (legal reasoning in sense (b)), then we should assume a coherence-independent test to identify the settled law of a jurisdiction first, and then bring in considerations of coherence at a later stage, and hold that courts ought to adopt that outcome to a case which is favoured by the most coherent set of propositions which, were the settled rules of the system justified, would justify them.

MacCormick 1984 espouses a similar view of the role which coherence can play in adjudication and gives an indication of how we might think of the links between interpretation and coherence in legal reasoning. According to MacCormick, in deciding a case according to law, courts should first of all interpret the existing law in order to establish a coherent view of some branch of the law, and they should do this by showing how that branch of law is justified according to some coherent set of principles or values which underlie it. The court should then use this view of the law in order to justify its decision in a new case which comes before it. On such an approach, then, once courts establish what the settled law is, they should then interpret law in applying it to a new case such that their decision is in accord with the most coherent account which justifies that settled law.

3.3 Coherence in Legal Reasoning: Necessary, Sufficient or Desirable?

Once a stance has been taken on the nature of the coherence relation in the case of law, many further questions concerning the role which considerations of coherence are to play in legal reasoning come to the fore. One important issue is that of how much emphasis is to be placed on coherence in justifying a judicial decision. Is exhibiting some degree of coherence with the existing law a necessary requirement of any justified judicial decision? Is it both a necessary and sufficient requirement, such that an account of the role of coherence in adjudication supplies us with a complete explanation of how judges should decide cases according to law? Or is coherence rather to be regarded as a desirable feature of judicial decision making, but one which can be overridden by other considerations in certain circumstances?

At this point in the discussion, it is possible to draw out some further possible links between the two concepts with which this entry is concerned, namely interpretation and coherence. If we hold that legal reasoning in sense (b), namely reasoning from the content of the law to the outcome which judges should adopt in a case before them, is wholly (Dworkin 1986) or mainly (Raz 1996a) interpretive, then we can reformulate some of the questions raised above concerning how much emphasis is to be placed upon coherence in adjudication in terms of the extent to which, in interpreting the law, we should interpret it in

such a way as to realise the value of coherence in judicial decisions. So, for example, is coherence the sole desideratum which should guide judges in interpreting the law, or is it merely one feature of a successful such interpretation, and, moreover, is it a necessary feature, or one which, although desirable, may be overridden by competing values which judges should also try to realise in interpreting the law?

Levenbook 1984 contends that it is a necessary condition for a judicial decision to be legally justified that it coheres with some part of the established law. She contrasts her understanding of this requirement with that adopted by MacCormick 1978. According to Levenbook, while MacCormick also holds that minimal coherence with some part of the established law is a necessary condition of a judicial decision being justified, he nevertheless contends that, so long as this minimal standard is met, further considerations of coherence which are also relevant to the decision can be defeated on consequentialist grounds. Levenbook's view is that this approach gives coherence too modest a role to play in legal reasoning; once a very minimal requirement of coherence is met, this value is too easily defeasible by other considerations. Levenbook finds this account of adjudication troubling because, in her view, it fails to do justice to judges' responsibility to be faithful to pre-existing law, a responsibility which places the judiciary in a quite different situation from the legislature when it comes to the question of how law ought to be developed. She contends that a judge who, within the limits allowed to him by the law, adopts a decision which is better on moral grounds over one which displays greater congruence with the trend or spirit of the existing law has made a mistake, and has adopted a legally unjustified decision. Levenbook very succinctly focuses the dilemma which judges must confront in deciding how much weight is to be placed on considerations of coherence in judicial decisions: should judges always adopt the outcome to a case which best coheres with the pre-existing law, or can they ever be justified in adopting an outcome which is less coherent but morally preferable? This way of focusing the dilemma which judges may face brings out another important aspect of coherence in legal reasoning, namely that granting a strong role to considerations of coherence is to place considerable emphasis on the backward-looking aspect of adjudication, as such an approach may require judges to place greater value on adhering to what has gone before, rather than on doing what would otherwise, on moral grounds, be the right thing.

Raz (1994a), who contends that coherence in legal reasoning is sometimes desirable, but certainly defeasible, poses essentially the same dilemma, but seems to place the burden of proof on those grasping the other horn of it. That is to say, Raz asks not how could it ever be justified for judges to deviate from the trend of the existing law in order to adopt a less coherent but otherwise morally preferable decision, but rather why should judges ever deviate from what is otherwise the morally best solution to a case before them on grounds of coherence? The burden of proof point is important because Raz's view seems to be that arguments in favour of a strong role for coherence in legal reasoning go through too easily, or are too readily adopted as a default position, perhaps because of the fact that reasoning by analogy is a common feature of many legal systems and seems to lend itself to being characterised in coherence terms, such that the facts speak for themselves in favour of the conclusion that considerations of coherence have a special role to play in legal reasoning. Pointing out that reasoning by analogy is not a necessary fact of life in all legal systems, and that, even where it does feature, it is still necessary to provide an explanation of the rationale of arguments from analogy, and the links between such arguments and coherence accounts of adjudication, Raz seeks to shift the burden of proof onto those who champion coherence. If judges are sometimes to deviate from what would otherwise be, according to law, the morally best

outcome to a case before them on grounds of coherence, then, Raz contends, we are in need of a convincing positive argument why this should be so.

3.4 Why Should Coherence Play a Role in Legal Reasoning?

One such argument may well be found in a point already mentioned in passing above, i.e. that judges are in a different position from legislators when it comes to deciding how the law ought to be developed. Raz (1979 and 1994a) argues that when faced with the choice of adopting the (according to law) morally best outcome to a case over an outcome which coheres better with the settled law, courts have to bear in mind that if they choose the former route, then some problematic consequences may ensue, such as that they may introduce conflicting rules reflecting conflicting social and economic purposes into the law, and hence create a considerable degree of dissonance with regard to existing legal doctrines in a given area of law. Such consequences need not dog legislative attempts to develop the law in a way which deviates from the doctrinal past, because legislative institutions have the power to sweep away the past in introducing new legislation, and to do so in such a way as to radically reform a whole area of law at one stroke. Courts, by contrast, can only make decisions concerning the issues arising in the case before them, and have considerably less opportunity to engage in radical reform of the law. These factors mean that judicial reform of the law will always be partial in nature, and, as was noted above, such partial reform brings with it the possibility of introducing dissonance and conflict into the law in the meantime. This may provide one reason why sometimes courts should give greater weight to considerations of coherence with pre-existing law in deciding cases which come before them, rather than striking out in a (albeit otherwise morally preferable) direction which coheres less well with settled law.

On Dworkin’s account of adjudication -- at least when that account is understood as a coherence account (see subsection 3.1 above) -- we find a different kind of answer to the question of why coherence has a special role to play in legal reasoning. As was noted in subsection 2.4, for Dworkin, in adjudicating cases, judges should seek to constructively interpret the law, i.e. to impose purpose on it in order to make of it the best possible example of the form or genre to which it is taken to belong (see Dworkin 1986, chs.2 & 3, and the entry on interpretivist theories of law). For Dworkin, the form or genre of law is to provide a convincing justification for the exercise of state coercion (see Dworkin 1986, *passim*; Dickson 2001, chs. 5 & 6), and, as he regards both judges and legal theorists as engaging in constructive interpretation (see Dworkin 1986, p90), Dworkin further contends that any adequate jurisprudential account of law must explain how what it takes to be law provides a general justification for the exercise of the coercive power of the state. In *Law’s Empire*, Dworkin argues that such a justification can best be provided when the law is viewed as the organised and coherent voice of what he refers to as a ‘community of principle’ i.e. a community whose members accept that their fates are linked by virtue of the fact that their rights and responsibilities are governed by common principles. So for Dworkin, we must interpret law as coherent, in the sense of speaking with one voice, because by so doing, we understand law as the voice of a community of principle, and so as capable of providing a general justification for the exercise of state coercion (see Dworkin 1986).

Some theorists are not so concerned to provide a ‘law-specific’ explanation of why coherence has an

important role to play in legal reasoning. For example, although acknowledging that there may be much more to be said about what is distinctive about legal reasoning, Hurley 1990 is largely content to explore the consequences for legal reasoning which ensue from the coherentist account of general practical reasoning which she espouses.

3.5 Coherence in Legal Reasoning: Global or Local?

In considering the role of coherence in legal reasoning, a final point to mention is that of how much of the law is to be made coherent according to various jurisprudential accounts granting a role to considerations of coherence. Are we talking of global coherence, such that judges should strive to reach judicial decisions which cohere to some extent with the settled law of an entire legal system, or should the coherence we seek be more local in nature, e.g. coherence with particular branches or areas of law?

Levenbook 1984 is a supporter of local coherence, and criticises those who, like Sartorius (1968 and 1971) and Dworkin (1977, and, although not yet written at the time of Levenbook's article, Dworkin 1986), hold that justified judicial decisions are those which best cohere with the law as a whole. According to Levenbook, champions of global coherence ignore the fact that sometimes a legally justified decision is supported by, in the sense of cohering with, principles which are distinctive of one area or branch of the law, but the principles concerned differ substantially from, and hence do not cohere well with principles from other branches of law. On this line of thinking, the judicial decision which coheres best with the principles underlying some specific field of law may not result in increased coherence of the entire system of law. That global coherence theorists may well be led to reject such a decision, and to hold that an alternative decision which coheres well with the overall system of law, but which increases incoherence locally is the more strongly justified, is, for Levenbook, a good reason to reject their theories: she contends that any plausible account of adjudication must make room for the kind of 'area-specific coherence' which she believes is necessary in the case of law. Raz 1994a also champions local over global coherence in adjudication, and his argument mentioned in the previous section concerning the limitations on the reforming role of courts, and the way in which this sometimes militates in favour of coherence playing a role in judicial decision making, is intended to support local coherence only. Peczenik 1994 claims that while the goal of the kind of doctrinal interpretation undertaken by legal scholars (which he refers to as 'legal dogmatics') is to establish the unity of an entire legal system, the judicial interpretation undertaken by judges is of a far more local variety, as it is concerned merely with the norms applicable to the case in question, and because a coherent interpretation of those norms may decrease their coherence with other legal norms.

While, as Levenbook 1984 notes, Dworkin's account of integrity in adjudication requires judges to attempt to view the legal system as a whole as exhibiting coherence and speaking with one voice in interpreting the law, Dworkin does also recognise that compartmentalisation into different branches or areas of law is an indisputable feature of legal practice, and he accordingly attempts to integrate it within his vision of adjudicative integrity. He does so via his doctrine of local priority in interpretation, i.e. that if a given principle justifying a judicial decision does not fit at all well with the area of law which the case is classified as falling under, then this counts dramatically against deciding the case in accordance with that principle, no matter how well such an interpretation coheres with other areas of the law (see Dworkin

1986, ch. 7). However, because of the strong pull toward global coherence in law as integrity -- expressed in Dworkin's claim that it is necessary to strive to view the legal system as a whole as speaking with one voice, the voice of an authentic political community, in order that law can be seen as justifying state coercion -- the current compartmentalisation of the law is not an unrevisable given for judges deciding cases, rather, it, too, is something which is subject to the Dworkinian process of constructive interpretation. This being so, he claims, when the compartmentalisation of the law does not track widely held principles of those subject to the law accounting current classificatory boundaries as important, then the doctrine of local priority is to be given much less force. Indeed, where there is a serious mismatch between the current compartmentalisation and actual views about relevant similarities and differences between areas of law held by those subject to it, it might be possible for judges to discard the doctrine of local priority altogether, and undertake radical reform of some departments of law in order to make them cohere better with others, or even to erase entire the alleged boundaries between certain branches of law in the course of interpreting the law and applying it to new cases.

Bibliography

- Aarnio, A., 1987, *The Rational as Reasonable*, Reidel, Dordrecht, Boston & Lancaster.
- Alexy, R., 1989, *A Theory of Legal Argumentation*, Clarendon Press, Oxford.
- Alexy, R. & Peczenik, A., 1990, 'The Concept of Coherence and Its Significance for Discursive Rationality', 3 *Ratio Juris*, 130-47.
- Baker, G.P., & Hacker, P.M.S., 1984, *Scepticism, Rules and Language*, Blackwell, Oxford.
- Baker, G.P. & Hacker, P.M.S., 1985, *Wittgenstein: Rules, Grammar and Necessity, Vol. 2 of an Analytical Commentary on the Philosophical Investigations*, Blackwell, Oxford.
- Bix, B., 1993, *Law, Language, and Legal Determinacy*, Clarendon Press, Oxford.
- Bork, R., 1990, *The Tempting of America*, New York Free Press, New York.
- Brest, P., 1980, 'The Misconceived Quest for the Original Understanding', 60 *Boston University Law Review*, 204-38.
- Coleman, J.C., & Leiter, B., 1995, 'Determinacy, Objectivity and Authority' in Marmor, A., *Law and Interpretation*, Clarendon Press, Oxford.
- Cornell, D., 1992, *The Philosophy of the Limit*, Routledge & Kegan Paul, New York, NY.
- Dancy, J., 1993, *Moral Reasons*, Blackwell, Oxford.
- Dickson, J., 2001, *Evaluation and Legal Theory*, Hart Publishing, Oxford.
- Dworkin, R., 1977, *Taking Rights Seriously*, Harvard University Press, Cambridge, Mass.
- Dworkin, R., 1985, 'How Law is like Literature' in Dworkin. R., *A Matter of Principle*, Harvard University Press, Cambridge, Mass.
- Dworkin, R., 1986, *Law's Empire*, Fontana Press, London.
- Dworkin, R., 1991, 'On Gaps in the Law' in Amselek and MacCormick, eds. *Controversies About Law's Ontology*, Edinburgh University Press, Edinburgh.
- Endicott, T.A.O., 1994, 'Putting Interpretation In Its Place', 13 *Law and Philosophy*, 451-79.
- Finnis, J., 1987, 'On Reason and Authority in Law's Empire' 6 *Law and Philosophy*, 357-380.
- Fish, S., 1989, *Doing What Comes Naturally: Change, Rhetoric and the Practice of Theory in Literary and Legal Studies*, Duke University Press, Durham, N.C. & Clarendon Press, Oxford.

- Fiss, O., 1982, ‘Objectivity and Interpretation’, 34 *Stanford Law Review*, 739-763.
- Hart, H.L.A., 1958, ‘Positivism and the Separation of Law and Morals’, 71 *Harvard Law Review*, 593-629.
- Hart, H.L.A., 1994, *The Concept of Law*, 2nd edn., with a postscript edited by P.A. Bulloch & J. Raz, Clarendon Press, Oxford.
- Holtzman, S. & Leich, C. (eds.), 1981, *Wittgenstein: To Follow A Rule*, Routledge and Kegan Paul, London.
- Hurley, S., 1989, *Natural Reasons*, Oxford University Press, Oxford.
- Hurley, S., 1990, ‘Coherence, Hypothetical Cases, and Precedent’, 10 *Oxford Journal of Legal Studies*, 221-51.
- Kelsen, H., 1967, *The Pure Theory of Law*, 2nd edn. trans. M. Knight, University of California Press, Berkeley, Ca.
- Kress, K., 1984, ‘Legal Reasoning and Coherence Theories: Dworkin’s Rights Thesis, Retroactivity, and the Linear Order of Decisions’, 72 *California Law Review*, 369-402.
- Kripke, S., 1982, *Wittgenstein on Rules and Private Language: an Elementary Exposition*, Blackwell, Oxford.
- Levenbook, B.B., 1984, ‘The Role of Coherence in Legal Reasoning’, 3 *Law and Philosophy*, 355-74.
- Levinson, L., 1982, ‘Law as Literature’, 60 *Texas Law Review*, 392-402.
- Marmor, A., 1992, *Interpretation and Legal Theory*, Clarendon Press, Oxford.
- Marmor, A. (ed.), 1995, *Law and Interpretation*, Clarendon Press, Oxford.
- MacCallum, G.C., 1968, ‘Legislative Intent’ in R.S. Summers (ed.), *Essays in Legal Philosophy*, Blackwell, Oxford.
- MacCormick, N., 1978, *Legal Reasoning and Legal Theory*, Clarendon Press, Oxford.
- MacCormick, N., 1984, ‘Coherence in Legal Justification’, in A. Peczenik et al. (eds.), *Theory of Legal Science*, D. Reidel Publishing, Dordrecht.
- McDowell, J., 1984, ‘Wittgenstein on Following a Rule’, 58 *Synthèse*, 325-363.
- McDowell, J., 1992, ‘Meaning and Intentionality in Wittgenstein’s Later Philosophy’, XVII *Midwest Studies in Philosophy*, 40-52.
- Moore, M., 1985, ‘A Natural Law Theory of Interpretation’ 58 *Southern California Law Review*, 277-398.
- Peczenik, A., 1989, *On Law and Reason*, Kluwer, Dordrecht.
- Peczenik, A., 1994, ‘Law, Morality, Coherence and Truth’, 7 *Ratio Juris*, 146-76.
- Raz, J., 1979, *The Authority of Law*, Clarendon Press, Oxford.
- Raz, J., 1994, *Ethics in the Public Domain*, Clarendon Press, Oxford.
- Raz, J., 1994a, ‘The Relevance of Coherence’ in Raz, J., *Ethics in the Public Domain*, Clarendon Press, Oxford.
- Raz, J., 1995, ‘Interpretation Without Retrieval’, in Marmor, A. (ed.), *Law and Interpretation*, Clarendon Press, Oxford.
- Raz, J., 1996a, ‘On The Nature of Law’, 82 *Archive fur Rechts und Sozialphilosophie*, 1-25.
- Raz, J., 1996b, ‘Why Interpret?’, 9 *Ratio Juris*, 349-63.
- Raz, J., 1996c, ‘Intention in Interpretation’, in George, R.P., (ed.), *The Autonomy of Law*, Clarendon Press, Oxford.

- Raz, J., 1998, ‘Postema on Law’s Autonomy and Public Practical Reasons: A Critical Comment’, 4 *Legal Theory*, 1-20.
- Raz, J., 2001, ‘Reasoning With Rules’, 54 *Current Legal Problems*, Oxford University Press, Oxford, 1 (forthcoming).
- Réaume, D., 1989, ‘Is integrity a virtue? Dworkin’s theory of legal obligation’, 39 *University of Toronto Law Journal*, 380-409.
- Sartorius, R., 1968, ‘The Justification of the Judicial Decision’, 78 *Ethics*, 171-87.
- Sartorius, R., 1971, ‘Social Policy and Judicial Legislation’, 8 *American Philosophical Quarterly*, 151-60.
- Smith, G.A., 1990, ‘Wittgenstein and the Sceptical Fallacy’, 3 *Canadian Journal of Law and Jurisprudence*, 155-186.
- Stone, M., 1995, ‘Focusing the Law: What Legal Interpretation is Not’ in Marmor, A. (ed.), *Law and Interpretation*, Clarendon Press, Oxford.
- Sunstein, C., 1996, *Legal Reasoning and Political Conflict*, Oxford University Press, New York & Oxford.
- Williams, B., 1985, *Ethics and the Limits of Philosophy*, Fontana, London.
- Wittgenstein, L., 1967, *Philosophical Investigations*, 3rd edn. trans. G.E.M. Anscombe, Blackwell, Oxford.

Other Internet Resources

[Please contact the author with suggestions]

Related Entries

justification, epistemic: coherentist theories of | legal reasoning: precedent and analogy | nature of law | nature of law: interpretivist theories | nature of law: legal positivism | truth: coherence theory of

Julie Dickson
j.dickson@ucl.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: May 29, 2001
Content last modified: May 29, 2001

Integrity

Integrity is one of the most important and oft-cited of virtue terms. It is also perhaps the most puzzling. For example, while it is sometimes used virtually synonymously with 'moral,' we also at times distinguish acting morally from acting with integrity. The person of integrity may in fact act immorally -- though they would usually not know they are acting immorally. Thus one may acknowledge a person to have integrity even though that person may hold importantly mistaken moral views.

When used as a virtue term, 'integrity' refers to a quality of a person's character, however, there are other uses of the term. One may speak of the integrity of a wilderness region or an ecosystem, a computerized database, a defense system, a work of art, and so on. When it is applied to objects, integrity refers to the wholeness, intactness or purity of a thing -- meanings that are sometimes carried over when it is applied to people. A wilderness region has integrity when it has not been corrupted by development or by the side-effects of development, when it remains intact as wilderness. A database maintains its integrity as long as it remains uncorrupted by error; a defense system as long as it is not breached. A musical work might be said to have integrity when its musical structure has a certain completeness that is not intruded upon by uncoordinated, unrelated musical ideas, that is, when it possesses a kind of musical wholeness, intactness and purity.

Integrity is also attributed to various parts or aspects of a person's life. We speak of attributes such as professional, intellectual and artistic integrity. However, the most philosophically important sense of the term 'integrity' relates to general character. Philosophers have been particularly concerned to understand what it is for a person to exhibit integrity throughout their life. Acting with integrity on some particularly important occasion will, philosophically speaking, always be explained in terms of broader features of a person's character and life. What is it to be a person *of* integrity? Ordinary discourse about integrity involves two fundamental intuitions: first, that integrity is primarily a formal relation one has to oneself, or between parts or aspects of one's self; and second, that integrity is connected in an important way to acting morally, in other words, there are some substantive or normative constraints on what it is to act with integrity. How these two intuitions can be incorporated into a consistent theory of integrity is not obvious, and most accounts of integrity tend to focus on one of these intuitions to the detriment of the other. A number of accounts have been advanced, the most important of them being: (i) integrity as the integration of self; (ii) integrity as maintenance of identity; (iii) integrity as standing for something; and (iv) integrity as moral purpose. These accounts are reviewed below. We then examine an issue that has been of central concern to philosophers exploring the concept of integrity: the relation between integrity and moral theory. Do moral theories such as utilitarianism allow room for people to live with integrity?

Integrity as Self-Integration

On the self-integration view of integrity, integrity is a matter of persons integrating various parts of their personality into a harmonious, intact whole. Understood in this way, the integrity of persons is analogous to the integrity of things: integrity is primarily a matter of keeping the self intact and uncorrupted. There are a variety of ways of developing this picture, depending on the concept of self and of integration that are employed.

One instructive attempt to describe the fully integrated self is Harry Frankfurt's. (Frankfurt 1987, pp. 33-34) Frankfurt does not explicitly address himself to the problem of defining integrity, nonetheless he does describe an important and influential account of self-integration. According to Frankfurt, desires and volitions (acts of will) are arranged in a hierarchy. First-order desires are desires for various goods; second-order desires are desires that one desire certain goods, or that one act on one first-order desire rather than another. Similarly, one may will a particular action (first-order volition) or one may will that one's first order volitions are of a particular sort (second-order volition). Second-order desires and volitions pave the way for third-order desires and volitions, and so on. According to Frankfurt, wholly integrated persons bring these various levels of volition and desire into harmony and fully identify with them at the highest level. There are various ideas as to what it means to fully identify with higher-level desires and volitions. However, such identification appears to involve knowing them; not deceiving oneself about them; and acting on them (usually).

A person is subject to many conflicting desires. If one simply acted at each moment out of the strongest current desire, with no deliberation or discrimination between more or less worthwhile desires, then one clearly acts without integrity. Frankfurt calls such a person a ‘wanton’ (Frankfurt 1971). Integrity thus requires that one discriminate between first-order desires. One may do this by endorsing certain first-order desires and ‘outlawing’ others. For instance, one may endorse a desire to study and ‘outlaw’ a desire to party, and do so by reference to a higher order desire ranking success over fun. Second-order desires may conflict. One may value success over fun, but also both fear that a ruthless pursuit of success will make one boring and value being fun over being boring. Fully integrated persons will not fall victim to such conflict; they will either avoid it altogether (if they can) or resolve the conflict in some way.

Resolution of self-conflict may be achieved by appeal to yet higher level desires or volitions, or by deciding to endorse one set of desires and outlawing others. At some point the full integration of one's self will require that one decide upon a certain structure of higher level desires and order one's lower level desires and volitions in light of it. As Frankfurt puts it, when a person unreservedly decides to endorse a particular desire:

> the person no longer holds himself at all apart from the desire to which he has committed himself. It is no longer unsettled or uncertain whether the object of that desire -- that is, what he wants -- is what he really wants: The decision determines what the person really wants by making the desires upon which he decides fully his own. To this extent the person, in making a decision by which he identifies with a desire, *constitutes himself*. (Frankfurt 1987, p. 38)

When agents thus constitute themselves without ambivalence (that is, unresolved desire for a thing and against it) or inconsistency (that is, unresolved desire for incompatible things), then the agent has what Frankfurt calls wholeheartedness. On one way of developing the integrated-self view of integrity, wholeheartedness is equated with integrity. It should be noted that self-conflict is not limited to desire. Conflict also ranges over commitments, principles, values, and wishes. Furthermore, all of these things -- desires, commitments, values, and so on -- are in flux. They change over time so that achieving the kind of ‘wholeheartedness’ that Frankfurt describes is a never-ending process and task. Self-knowledge is crucial to this process in so far as one must know what one's values, for example, are if one is to order them.

Frankfurt's account illustrates one way of describing the fully-integrated self. (See Taylor 1981 for a different approach.) The key question, however, is whether the idea of a fully-integrated self adequately captures the quality we ascribe when we say of someone that they are a person of integrity. There have been a number of criticisms of the integrated-self view of integrity. First, it places only formal limits on the kind of person who may be said to have integrity. People of integrity, however, are plausibly thought to be generally honest and genuine in their dealings with others. (See Halfon 1989, pp. 7-8.) Imagine a person who sells used-cars for a living and is wholeheartedly dedicated to selling cars for as much money as possible. Such a person will be prepared to blatantly lie in order to set up a deal. The person may well be perfectly integrated in Frankfurt's sense, but we should feel no temptation at all to describe them as having exemplary integrity.

Second, a person of integrity is plausibly said to make reasonable judgements about the relative importance of various desires and commitments. Yet, again, the self-integration view places only formal limits on the kind of desires that constitute a self. (See McFall 1987 pp. 9-11, Calhoun 1995 pp. 237-38). As McFall notes, one cannot say with a straight face something like: ‘Harold demonstrates great integrity in his single-minded pursuit of approval.’ (McFall 1987 p.9.) If integrity is nothing more than the perfect integration of self, however, it is hard to see how one can automatically deny Harold's integrity.

Third, on some accounts, the fully and perfectly integrated person is not able to experience genuine temptation. Temptation requires that the full force of an ‘outlaw’ desire be experienced, but successful

integration of the self may mean that such desires are fully subordinated to wholeheartedly endorsed desires and this may preclude an agent fully experiencing them. (See Taylor 1981 p.151 for an example of a view like this.) That a person experiences, and overcomes, temptation would count against their integrity on such a view. One might think, however, that a capacity to overcome temptation and display strength of character is in fact a sign of a person's integrity, not its lack. (Halfon 1989 pp. 44-7 urges this criticism.)

Fourth, Cheshire Calhoun argues that agents may find themselves in situations in which wholeheartedness tends to undermine their integrity rather than constitute it. (Calhoun 1995 pp.238-41. Analogously, Victoria Davion 1991 pp.180-192 argues that a person may change radically and yet maintain integrity.) In the midst of a complex and multifaceted life one may have compelling reasons to avoid neatly resolving incompatible desires. The cost of the resolution of all self-conflict may be a withdrawal from aspects of life that make genuine claims upon us. Resolving self-conflict at the expense of fully engaging with different parts of one's life does not seem to contribute to one's integrity. It seems rather like the sort of cop-out that undermines integrity. (One should not confuse integrity with *neatness*.)

The Identity View of Integrity

A related approach to integrity is to think of it primarily in terms of a person's holding steadfastly true to their commitments, rather than ordering and endorsing desires. ‘Commitment’ is used as a broad umbrella term covering many different kinds of intentions, promises, convictions and relationships of trust and expectation. One may be, and usually is, committed in many different ways to many different kinds of thing: people, institutions, traditions, causes, ideals, principles, projects, and so on. Commitments can be explicitly, self-consciously, publicly entered into or implicit, unself-conscious, private. Some are relatively superficial and unimportant, like casual support of a sporting team; others are very deep, like the commitment implicit in genuine love or friendship.

Because we find ourselves with so many commitments, of so many different kinds, and because commitments inevitably clash and change over time, it will not do to define integrity merely in terms of remaining steadfastly true to one's commitments. It matters which commitments we expect a person of integrity to remain true to. Philosophers have developed different accounts of integrity in response to this need to specify the kind of commitments that are centrally important to a person's integrity.

One option here is to define integrity in terms of the commitments that a person identifies with most deeply, as constituting what they consider their life is fundamentally about. Commitments of this kind are called ‘identity-conferring commitments’ or sometimes ‘ground projects’. This view of integrity, the identity view, is associated most closely with Bernard Williams. It is implicit in his discussion of integrity and utilitarianism (Williams 1973; we examine this discussion below) and also features in his criticism of Kantian moral theory (1981b). The idea is that for a person to abandon an identity-conferring commitment is for them to lose grip on what gives their life its identity, or individual character. An identity-conferring commitment, according to Williams, is ‘the condition of my existence, in the sense that unless I am propelled forward by the conatus of desire, project and interest, it is unclear why I should

go on at all.’ (Williams 1981b p.12).

One apparent consequence of defining integrity as maintenance of identity-conferring commitments is that integrity cannot really be a virtue. This is Williams's view. He argues that although it is an admirable quality, integrity is not related to motivation as virtues are. A virtue either motivates a person to act in desirable ways (as benevolence moves a person to act for another's good), or it enables a person to act in desirable ways (as courage enables a person to act well). If integrity is no more than maintenance of identity, however, it can play neither of these roles. On the identity view of integrity, to act with integrity is just to act in a way that accurately reflects your sense of who you are; to act from motives, interests and commitments that are most deeply your own. (Williams 1981a p.49) A further consequence of this view of integrity as maintenance of identity-conferring commitments is that there appears to be no normative constraints either on what such commitments may be, or on what the person of integrity can do in the pursuit of those commitments. The person of integrity can do horrific things and maintain their integrity so long as they are acting accordance with their core commitments.

A number of criticisms of the identity view of integrity have been made. First, integrity is usually regarded as something worth striving for and the identity account of integrity fails to make sense of this. (See Cox, La Caze, Levine 1999.) It disconnects integrity from the prevalent view that it is a virtue of some kind and generally praiseworthy. Second, the identity theory of integrity ties integrity to commitments with which an agent identifies, but acts of identification can be ill-informed, superficial and foolish. A person may, through ignorance or self-deception, fail to understand or properly acknowledge the source of their deepest commitments and convictions and we are unlikely to attribute integrity to a person who held true to a false and unrealistic picture of themselves. (On the other hand, this view of integrity as maintenance of identify-conferring commitments, recognizes the relevance of self-knowledge to *acting* with integrity. If a person fails to act on their core commitments, through self-deception, weakness of will, cowardice, or even ignorance, then they lack integrity.)

Third, on the identity view of integrity, a person's integrity is only at issue when their deepest, most characteristic, or core convictions and aspirations are brought into play. However, we expect persons of integrity to behave with integrity in many different contexts, not only those of central importance to them. (See Calhoun 1995, p.245.)

Fourth, as noted above, the identity view of integrity places only formal conditions upon the kind of person that might be said to possess integrity. The identity view of integrity shares this feature with the self-integration view of integrity and similar criticism can be made of it on this ground. It seems plausible to observe certain substantive limits on the kinds of commitments had by a person of integrity.

Integrity as Standing for Something

The self-integration and identity views of integrity see it as primarily a personal virtue: a quality defined by a person's care of the self. Cheshire Calhoun argues that integrity is primarily a social virtue, one that is defined by a person's relations to others (Calhoun 1995). The social character of integrity is, Calhoun

claims, a matter of a person's proper regard for their own best judgement. A person of integrity does not just act consistently with their endorsements, they stand for something: they stand up for their best judgement within a community of people trying to discover what in life is worth doing. As she puts it:

> Persons of integrity treat their own endorsements as ones that matter, or ought to matter, to fellow deliberators. Absent a special sort of story, lying about one's views, concealing them, recanting them under pressure, selling them out for rewards or to avoid penalties, and pandering to what one regards as the bad views of others, all indicate a failure to regard one's own judgment as one that should matter to others. (Calhoun 1995 p. 258)

On Calhoun's view, integrity is a matter of having proper regard for one's role in a community process of deliberation over what is valuable and what is worth doing. This, she claims, entails not only that one stand up, unhypocritically, for one's best judgement, but also that one have proper respect for the judgement of others.

Calhoun's account of integrity promises to explain why it is that the fanatic lacks integrity. It seems intuitively very plausible to distinguish between fanatical zeal and integrity, but the self-integration and identity views of integrity threaten to make the fanatic a paradigm case of a person of integrity. Fanatics integrate desires and volitions of various orders in an intimidatingly coherent package; they remain steadfastly true to their deepest commitments like no others. On Calhoun's view of integrity, however, we can locate a distinction between integrity and fanaticism. Fanatics lack one very important quality that, on Calhoun's view, is centrally important to integrity: they lack proper respect for the deliberations of others. What is not clear in Calhoun's account, and is in fact very hard to get clear on in any case, is what the *proper* respect for other's views in the end amounts to. Exemplary figures of integrity often stand by their judgement in the face of enormous pressure to recant. How, then, is one to understand the difference between standing up for one's views under great pressure and fanatically standing by them? Calhoun's claim that the fanatic lacks integrity because they fail to properly respect the social character of judgement and deliberation sounds right, but most of the work is done by the idea of ‘proper respect’ -- and it is not clear what this in the end comes to. Her view appears to allow for the possibility that integrity can accommodate the very kind of fanaticism (for example, the virtuous Nazi) that she wishes her account of integrity to eschew.

Calhoun's account of integrity places no conceptual constraints on the kinds of commitments that a person of integrity may endorse. It does not seem necessary on her view that a person of integrity have a special concern with acting morally. Although they have a special concern to understand what in life is worth doing, the person of integrity is not constrained to give moral, other-regarding answers to this question. The following account of integrity is explicitly concerned with attitudes towards morality.

Integrity as Moral Purpose

Another way of thinking about integrity places moral constraints upon the kinds of commitment to which a person of integrity must remain true. There are several ways of doing this. Elizabeth Ashford argues for

a virtue she calls ‘objective integrity’. Objective integrity requires that agents have a sure grasp of their real moral obligations. (Ashford 2000 p. 246.) A person of integrity cannot, therefore, be morally mistaken. Understood in this way, one only ascribes integrity to a person with whom one finds oneself completely in moral agreement. This concept of integrity does not, however, closely match ordinary use of the term. The point of attributing integrity to another is not to signal unambiguous moral agreement. It is often to ameliorate criticism of another's moral judgement. For example, we may disagree strongly with the Pope's views of the role of women in the Church, take this to be a significant moral criticism of him, and yet admit that he is a man of integrity. In such a case it is largely the point of attributing integrity to open a space for substantial moral disagreement without launching a wholesale attack upon another's moral character.

Mark Halfon offers a different way of defining integrity in terms of moral purpose. Halfon describes integrity in terms of a person's dedication to the pursuit of a moral life and their intellectual responsibility in seeking to understand the demands of such a life. He writes that persons of integrity:

> embrace a moral point of view that urges them to be conceptually clear, logically consistent, apprised of relevant empirical evidence, and careful about acknowledging as well as weighing relevant moral considerations. Persons of integrity impose these restrictions on themselves since they are concerned, not simply with taking any moral position, but with pursuing a commitment to do what is best. (Halfon 1989, p. 37.)

Halfon's view allows that integrity is not necessarily ‘objective’, as Ashford claims, and is similar in a number of respects to Calhoun's. Both see integrity as centrally concerned with deliberation about how to live. However, Halfon conceives this task in more narrowly moral terms and ties integrity to personal intellectual virtues exercised in pursuit of a morally good life. Halfon speaks of a person confronting ‘all relevant moral considerations’, but this turns out to be quite a formal constraint. What counts as a relevant moral consideration, on Halfon's view, depends upon the moral point of view of the agent. Persons of integrity may thus be responsible for acts others would regard as grossly immoral. What is important is that they act with moral purpose and display intellectual integrity in moral deliberation. This leads Halfon to admit that, on his conception of integrity, it is possible for a Nazi bent on genocide of the entire Jewish people to be a person of moral integrity. Halfon thinks it possible, but not at all likely. (Halfon 1989 pp. 134-36)

Other philosophers object to this consequence. If the genocidal Nazi is a possible object of ascriptions of moral integrity, then we can properly ascribe integrity to people whose moral viewpoint is bizarrely remote from any we find intelligible or defensible. (See McFall 1987. Putnam 1996 draws on the work of Carol Gilligan 1982 to suggest a different way of overcoming the problem of the Nazi of integrity.) Moral constraints upon attributions of integrity need not take the form of Ashford's ‘moralized’ view or Halfon's more limited formal view. One might say instead that attributions of integrity involve the judgement that an agent acts from a moral point of view those attributing integrity find intelligible and defensible (though not necessarily right) -- and that this formal constraint does have substantive implications. It prohibits attributing integrity to, for example, those who advocate genocide, or deny the moral standing of people on, for example, sex-based or racial grounds. There are things which a person of

integrity cannot do. The Nazis and other perpetrators of great evil were either committed to what they were doing, in which case they were immoral (or not moral agents at all) and lacked integrity; or else they lacked integrity because they were self-deceived or dissembling and never had the Nazi commitments they alleged to have, and acted upon, at all. Judgements of integrity would thus involve judgement about the reasonableness of other's moral points of view, rather than the absolute correctness of their view (Ashford) or the intellectual responsibility with which they generally approach the task of thinking about moral questions (Halfon).

Defining integrity in terms of moral purpose has the advantage of capturing intuitions of the moral seriousness of questions of integrity. However, the approach appears too narrow. Halfon's identification of integrity and moral integrity appears to leave out important personal aspects of integrity, aspects better captured by the other views of integrity we have examined. Integrity does not seem to be exclusively a matter of how people approach plainly moral concerns. Other matters like love, friendship and personal projects seem to be centrally important. Imagine a person who sets great store in writing a novel, but who postpones the writing of it for years on one excuse or another and then abandons the idea of novel-writing after one difficult experience with a first chapter. We would think this person's integrity diminished by their failure to make a serious attempt to see the project through, yet the writing of a novel need not be a moral project. (See McFall 1987 on the difficulty of bringing together personal and moral aspects of integrity.)

All of the accounts of integrity we have examined have a certain intuitive appeal and capture some important feature of our concept of integrity. There is, however, no philosophical consensus on the best account. It may be that the concept of integrity does not lend itself to a single coherent description. Integrity may be a cluster concept, tying together different, overlapping qualities of character under the one term. On the other hand, it may be that a fully adequate account of integrity is simply yet to emerge.

Integrity and Moral Theory

Despite the fact that it is somewhat troublesome, the concept of integrity has played an important role in contemporary discussion of moral theory. An important and influential line of argument, first developed by Bernard Williams, seeks to show that certain moral theories do not sufficiently respect the integrity of moral agents. (See Williams 1973 & 1981.) This has become an important avenue of critique of modern moral theory. (See, for example, Scheffler 1993 and Lomasky 1987.)

Modern moral theories, the most representative of which are utilitarianism and Kantian moral theory, do not concern themselves directly with virtue and character. Instead, they are primarily concerned to describe morally correct action. Theories of morally correct action generally aspire to develop criteria by which to categorize actions as morally obligatory, morally permissible, or morally impermissible. Some theories of morally correct action also introduce the category of the supererogatory: an action is supererogatory if and only if it is morally praiseworthy, but not obligatory. The two theories of primary concern to Williams are utilitarianism and Kantian moral theory, and both of these are usually interpreted as eschewing the category of the supererogatory. (See Baron 1995 for an argument that Kantian moral

theory has no need for the category of the supererogatory.) Williams maintains that both utilitarianism and Kantian moral theory are deeply implausible because of their integrity undermining effects. His argument against utilitarianism makes the more transparent appeal to the concept of integrity and it is this argument that we examine here. (But see Herman 1983, Rogerson 1983, Jensen 1989, and Baron 1995, chapter four, for critical discussion of the Williams's argument against Kantian moral theory.)

Williams's argument against utilitarianism is directed against a particular version of utilitarianism -- act-utilitarianism. This is, very roughly, the view that an agent is to regard as morally obligatory all and only actions that maximize general well-being. The act-utilitarian theory that Williams criticizes has an important feature: it aspires to describe the correct form of moral *deliberation*. It does more than specify what it is for an action to be morally correct, it specifies how an agent should think about moral decisions. Agents should think about which of the actions available to them will maximize general well-being and decide to act accordingly. Notice that this theory is completely impartial and that it makes no room for an agent to give special weight to personal commitments, causes, projects, and the like. Act-utilitarianism recognizes no personal sphere of activity in which moral reflection operates merely as a side-constraint.

According to Williams, an agent who adopted this version of utilitarianism would find themselves unable to live with integrity. As he puts it, to become genuinely committed to act-utilitarianism is for a person to become alienated:

> in a real sense from his actions and the source of his actions in his own convictions. It is to make him into a channel between the input of everyone's projects, including his own, and an output of optimific decision; but this is to neglect the extent to which *his* actions and *his* decisions have to be seen as the actions and decisions which flow from the projects and attitudes with which he is most closely identified. It is thus, in the most literal sense, an attack on his integrity. [Williams (1973, p.117)]

Williams's argument is based on the identity theory of integrity, discussed above. Integrity, on this view, requires that persons act out of their own convictions, that is, out of commitments with which they deeply identify. Act-utilitarianism seeks to replace personal motivations of this kind with impartial utilitarian reasoning. Williams's argument appears to make acting with integrity incompatible with acting in accordance with act-utilitarianism.

Williams develops the point with two famous and much discussed examples. (1972, pp.97-99). The example which best illustrates his argument involves the figure of George, a recent doctoral graduate in chemistry who is having difficulty finding work. George has young children. He also has poor health, limiting his job opportunities. George's (unnamed) wife must work to support the family and on Williams's story this causes a great deal of strain on the family. George has a strong commitment to pacifism, a conviction amounting to an identity-conferring commitment. A dilemma arises for George when more senior colleague tells him about a decently paid job in a laboratory doing work on biological and chemical warfare. If George does not take up the job, it will almost certainly go to another chemist, one without George's pacifist commitment, who will pursue the development of biological and chemical

weapons more vigorously than George. Should George take the job or not?

The most likely act-utilitarian conclusion here is that George should accept the job. This would contribute greatly to the well-being of his family as well as probably contributing to general welfare by forestalling some relatively zealous development of weapons of mass destruction. Weighed in the balance are George's feelings in the matter. The utilitarian calculation, if it really does come out this way, is demanding a sacrifice of George: that he put aside his opposition to, and distaste for, biological and chemical weapons development and deal with the anguish and alienation that may result from working in the laboratory.

According to Williams, however, act-utilitarianism in fact demands a different kind of sacrifice from George. It demands that he act without integrity, abandoning or ignoring a longstanding, identity-conferring commitment to pacifism simply because maximum general well-being is to be found elsewhere. This is just one, particularly acute, example of the tendency of impartial utilitarian deliberation to run roughshod over identity-conferring commitments, treating them as no more than one source of utility among others. In general, Williams concludes, identity-conferring commitments cannot play the kind of role in act-utilitarian moral deliberation that is required for an agent to act with integrity, that is, for an agent to act with genuine conviction in matters of grave, identity-determining importance to them.

Williams's critique of utilitarianism has spawned a large and important literature in which the argument has been interpreted and reinterpreted, redrafted, and much criticized. There are, nonetheless, three main lines of response to the Williams's critique of utilitarianism. We consider them in turn. The first reply essentially concedes the point and offers in response a development of utilitarian moral theory, one aimed at avoiding the flaws that Williams sought to demonstrate. One way to do this is by watering down the impartiality of utilitarian theory, explicitly factoring in the permissibility of giving extra weight to one's own personal projects, commitments, and so on. (See Scheffler 1993 for a development of this view, and Harris 1989a and 1989b for criticism of the adequacy of this response.)

Another way to try and improve utilitarianism in response to Williams's argument is to advance a less ambitious form of utilitarian moral theory. Recall that Williams criticizes a version of act-utilitarian *moral deliberation*, so one may respond to it by describing a version of act-utilitarianism that does not dictate the form of moral deliberation. A moral theory, on this view, primarily describes morally correct action and does not automatically entail a theory of correct moral deliberation. Thus one might subscribe to an act-utilitarian account of morally correct action whilst not demanding that someone like George approach life by deliberating in strictly utilitarian ways. There are however, a number of difficulties with separating out theories of morally correct action and correct moral deliberation in this way. For one thing, it appears to deprive a theory of morally correct action of much point. What is the point, one might ask, of subscribing to a moral theory if it offers no clear practical guidance on how one should act? (See Williams 1981a for a discussion of this point.) Nonetheless, there have been attempts to develop and to motivate versions of utilitarianism not prescribing methods of moral deliberation. (See Railton 1986 for development of such a view and Harcourt 1998 for criticism of it.)

A second possible line of response to the argument is to deny the presupposition of Williams's argument that it is absurd for a moral theory to undermine integrity. It may just be that moral demands upon us really are very stringent, and identity-conferring commitments must sometimes (perhaps often) be sacrificed in the interests of, say, our acting to ameliorate preventable suffering. One might even consider it a virtue of utilitarianism that it demonstrates how genuinely difficult it is to preserve one's integrity when confronting a world of massive and easily preventable suffering. (See Ashford 2000 for an argument along these lines.)

The third, and most influential, line of response argues directly against the idea that utilitarianism demands that agents act against their convictions. Utilitarianism demands that agents adopt utilitarian ideals; that agents give utilitarian ideals the kind of priority that would have them function as the central identity-conferring commitments of their life. Thus utilitarianism does not demand that one live without identity-conferring commitments at all, but that one live with *utilitarian* identity-conferring commitments. Were George a utilitarian, he would not have been acting against his convictions by taking a job in the chemical weapons factory. He does not lose his integrity simply in virtue of his commitment to utilitarianism. Williams appears to confuse the case in which a utilitarian George acts against his personal interests (in which case his integrity would be preserved) with the case in which a non-utilitarian George is somehow persuaded to act as a utilitarian (in which case his integrity would not be preserved). Acting as a utilitarian when one has no sympathy with utilitarianism may well diminish one's integrity, but such a loss of integrity is not attributable to utilitarianism and has no bearing on utilitarianism's plausibility as a moral theory. (See Carr 1976, Trianosky 1986 and Blustein 1991 for versions of this criticism.)

The matter is not finally settled, however, for notice that Williams's critique is premised on a version of the identity theory of integrity. As we have seen, there are other plausible candidates for an account of integrity and the critique of utilitarianism may well succeed better in their terms. The key issues are whether utilitarian commitment is compatible with a fully satisfactory account of integrity, and if so, whether integrity is of such value and importance that the clash between integrity and utilitarian commitment undermines the plausibility of utilitarian moral theory. An adequate account of integrity needs to deal with these issues and to capture basic intuitions about the nature of integrity: that persons of integrity may differ about what is right but a moral monster cannot have integrity.

Bibliography

- Ashford, Elizabeth (2000). ‘Utilitarianism, Integrity and Partiality. ‘ *Journal of Philosophy* 97, pp. 421-439.
- Baron, Marcia (1995). *Kantian Ethics Almost without Apology*. Ithaca: Cornell University Press.
- Blustein, Jeffrey (1991). *Care and Commitment: Taking the Personal Point of View* New York: Oxford University Press.
- Calhoun, Chesire (1995). ‘Standing for Something.’ *Journal of Philosophy* XCII, pp. 235-260
- Carr, Spencer (1976). ‘The Integrity of a Utilitarian.’ *Ethics* 86, pp. 241-46.
- Cox, Damian; La Caze, Marguerite; Levine, Michael P. (1999). ‘Should We Strive for Integrity?,’

Journal of Value Inquiry Vol. 33, No. 4.

- Davion, Victoria (1991) ‘Integrity and Radical Change’, *Feminist Ethics*, Ed. Claudia Card, Lawrence, Kansas: University of Kansas Press, pp.180-192.
- Frankfurt, Harry (1971). ‘Freedom of the Will and the Concept of a Person.’ *Journal of Philosophy* LXVIII, pp. 5-20.
- ------ (1987). ‘Identification and Wholeheartedness.’ Ferdinand Schoeman, ed. *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. New York: Cambridge University Press.
- Gilligan, Carol. (1982) *In a Different Voice*, Cambridge, Mass: Harvard University Press
- Halfon, Mark (1989). *Integrity: A Philosophical Inquiry*. Philadelphia: Temple University Press.
- Harcourt, Edward (1998). ‘Integrity, Practical Deliberation and Utilitarianism.’ *Philosophical Quarterly* 48, pp. 189-198.
- Harris, George W. (1989a). ‘Integrity and Agent Centered Restrictions.’ *Nous*, 23, pp. 437-456.
- ------ (1989b). ‘A Paradoxical Departure from Consequentialism’ *Journal of Philosophy* 86, pp. 90-102.
- Herman, Barbara (1983). ‘Integrity and Impartiality.’ *Monist* 66, pp. 233-250.
- Jensen, Henning (1989). ‘Kant and Moral Integrity.’ *Philosophical Studies* 57, pp. 193-205.
- Lomasky, Loren (1987). *Persons, Rights, and the Moral Community*. Oxford: Oxford University Press.
- McFall, Lynne (1987). ‘Integrity.’ *Ethics* 98, pp. 5-20. Reprinted in John Deigh (ed.), *Ethics and Personality*, Chicago: University of Chicago Press, 1992, pp. 79-94.
- Putman, Daniel (1996). ‘Integrity and Moral Development.’ *The Journal of Value Inquiry* 30, pp. 237-246.
- Railton, Peter (1984) ‘Alienation, Consequentialism and the Demands of Morality’ *Philosophy and Public Affairs* 13, pp. 134-72.
- Rogerson, Kenneth (1983). ‘Williams and Kant on Integrity.’ *Dialogue* 22, pp. 461-478.
- Scheffler, Samuel (1993) *The Rejection of Consequentialism*, Revised Edition, Oxford: Oxford University Press.
- Taylor, Gabriele (1981). ‘Integrity.’ *Proceedings of the Aristotelian Society*, Supplementary Volume 55, pp. 143-159.
- ------ (1985). ‘Integrity.’ *Pride, Shame and Guilt: Emotions of Self-Assessment*. Oxford: Oxford University Press, pp.108-141.
- Trianosky, Gregory W. (1986). ‘Moral Integrity and Moral Psychology: A Refutation of Two Accounts of the Conflict Between Utilitarianism and Integrity’ *Journal of Value Inquiry* 20, pp. 279-288.
- Williams, Bernard (1973). ‘Integrity.’ J.J.C. Smart and Bernard Williams, *Utilitarianism: For and Against* New York: Cambridge, 108-117.
- ------ (1981a). ‘Utilitarianism and Moral Self-Indulgence.’ *Moral Luck: Philosophical Papers 1973-1980*. Cambridge: Cambridge University Press, 40-53.
- ------ (1981b). ‘Persons, Character and Morality.’ *Moral Luck*, 1-19.

Other Internet Resources

[Please contact the authors with suggestions.]

Related Entries

consequentialism | ethics: virtue | impartiality | moral psychology

Damian Cox | **Marguerite La Caze** | **Michael P. Levine**
D.Cox@cowan.edu.au | *m.lacaze@mailbox.uq.edu.au* | *mlevine@arts.uwa.edu.au*

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 9, 2001
Content last modified: April 9, 2001

Francis of Marchia

Francis of Marchia was perhaps the most exciting theologian active at the University of Paris in the quarter century between the Franciscan Peter Auriol (fl. 1318) and the Augustinian Hermit Gregory of Rimini (fl. 1343). Although he had innovative and even influential ideas in philosophical theology, natural philosophy, and political philosophy, until recently he has been little studied.

Life and Work

Francis of Marchia (a.k.a. de Appignano, de Pignano, de Esculo, de Ascoli, Franciscus Rubeus, and the *Doctor Succinctus*) was born ca. 1290 in the village of Appignano del Tronto in the province of Ascoli Piceno, east of Rome and near the Adriatic. He became a Franciscan and rose in the order's educational hierarchy until he lectured on the *Sentences* at the Franciscans' Paris *studium* in 1319–20. From his Paris years (1319–23 or 24) we have a popular commentary on the four books of the *Sentences*, extant in at least ten manuscripts for each book, and in at least two distinct redactions for the first two books. From the same period stems Marchia's short commentary on the first two books of the *Metaphysics*. By 1324, when he was lector at the Franciscans' *studium* in Avignon, Marchia had become Master of Theology. Perhaps his *Quodlibet* derives from theological debates held in Avignon. Marchia's literal commentary on the *Physics* and his more independent long commentary or questions on the first seven books of the *Metaphysics* probably belong to his Paris or Avignon period. In 1328, Marchia fell out with Pope John XXII for supporting the Franciscan Minister General Michael of Cesena on the issue of apostolic poverty, left Avignon, and eventually took refuge with the Holy Roman Emperor Louis of Bavaria along with William of Ockham, Marsilius of Padua, and others. During this period Marchia wrote his

Improbatio against John XXII. Marchia was eventually captured, and he confessed and retracted his errors before the Inquisition in 1341, or perhaps 1336 (a possibility suggested in Wittneben-Lambertini 1999). He was reconciled with the Church, and died some time after 1344.

Until 1991 almost none of Marchia's writings were published and only Anneliese Maier devoted serious and sustained attention to his thought. Since then, however, Nazareno Mariani has published three volumes of Marchia's works (Marchia 1993, 1998, and 1997b): the *Improbatio*, the *Physics* commentary, and the *Quodlibet*, the last of which includes roughly a quarter of the main version — the *Scriptum* — of book I of the *Sentences* commentary, and two questions from book II, in appendices. In a 1991 monograph on Marchia's cosmological thought (Schneider 1991, 29), Notker Schneider announced the edition of the *Metaphysics* commentary, building on the work of Albert Zimmermann (Marchia 1965, Zimmermann 1966). Schneider, Russell Friedman, and Christopher Schabel have added to Mariani's efforts in editing sections of the *Sentences* commentary (Marchia 1991b, 1997a, 1999a, 1999b, 2000, 2001). Currently, Mariani is editing the *Reportatio* for book I, while Friedman and Schabel are occupied with the *Scriptum* on the same book. The Centro Studi Francesco d'Appignano was established in Marchia's hometown in May of 2001 during the First International Conference on Francis of Marchia. The second such conference is planned for 2003. Finally, the journal *Picenum Seraphicum*, edited by Roberto Lambertini, promises to be a forum for Marchia studies.

Philosophical Theology

Marchia's main and most popular writing, larger than all his other works combined, is the *Sentences* commentary. Although this genre of scholastic writing contains much material that we would call pure science, especially in Marchia's case (see Natural Philosophy below), it is primarily a vehicle for philosophical theology. Here it must be stressed at the outset that Francis of Marchia was **not** a faithful Scotist, contrary to a common opinion based on misconceptions from early in the twentieth century. Most recent research has proven that Marchia generally rejected or severely modified John Duns Scotus's doctrines, rather than followed them, even in the specific contexts like Trinitarian theology where he was claimed to have been a loyal Scotist (Marchia 1999a). For example, Marchia was uncomfortable with Scotus's stress on and use of a strong distinction between the divine intellect and will, and this led Marchia to oppose Scotus on issues such as the procession of the Holy Spirit and the mechanism of divine foreknowledge. Indeed, it seems that Marchia was perhaps less of a Scotist than any of the other continental Franciscans active between Peter Auriol and the Black Death. Nevertheless, Scotus forms much of the backdrop for Marchia's theology. The other thinker against whose theories Marchia often developed his own doctrine was Auriol, whose *Sentences* commentary Marchia appears to have known in a *Reportatio* version.

A critical edition of one redaction of Marchia's *Sentences* commentary would probably take up about five large volumes. It is roughly equal in size to that of his contemporary and confrère Landulph Caracciolo. Whereas Caracciolo's commentary is frequently a disjointed, point-by-point refutation of Peter Auriol and defense of Scotus on many issues, however, Marchia was more selective in choosing his topics, more independent in giving his determination, and generally more eloquent. For example, Marchia asks only

one brief question on the Immaculate Conception, a favorite Franciscan topic (book III, q. 8, and because of a missing quire this question is not present in the most studied manuscript, Vat. Chigi. lat. B VII 113), showing himself to support the immaculatist position, but he asks a great number of questions in books I, II, and IV on the relationship between the will and intellect.

Thus, although Auriol's controversial opinions on epistemology and divine foreknowledge drew fire from Caracciolo and most other Paris theologians up to Gregory of Rimini's day, even from Oxford Franciscans, Marchia devoted much energy to foreknowledge but almost ignored the great debate over intuitive and abstractive cognition. Aside from an isolated mention of "intuitive cognition" in *Scriptum* I, d. 3, q. 4, and of "intuitive vision" in the very last question in book IV, Marchia merely discusses the problem in passing on two brief occasions: first, in book II, q. 25 (Marchia 1997b, p. 322), while treating angelic knowledge, and second, in book III, q. 13, in the context of the beatific vision of the Word (not yet edited). In the latter case, the more substantial passage, he gives a somewhat Scotistic definition: "Intuitive and abstractive cognition are not distinguished according to having a species or not, but only according to the disposition of the object, because if the object is present, the species represents it intuitively; if absent, it represents it abstractively." Therefore, "the same species that is intuitive in the presence of the object is abstractive in the absence of the object." Marchia adds in agreement with Auriol that God "can cause the act of seeing without the object," and that a species of a created object "indifferently represents" a present or absent object. Although this might provoke some epistemological questions, Marchia turns to the vision of the Essence and leaves human cognition *in via* aside.

In contrast, Marchia singled out divine foreknowledge for special treatment, devoting three entire distinctions (*Scriptum* I, dd. 35, 36, and 38) to the issue and concentrating most of his discussion on the reconcilation of foreknowledge with human free will (Marchia 1999b, 2000; Schabel 2000). Here Marchia was opposed in some way to just about everyone else who had written on the issue, but mainly Auriol, who had claimed that any determination prior to the coming about of a contingent event destroyed contingency, including the truth or falsity of future-contingent propositions. Marchia's defense of the application of the Principle of Bivalence to propositions about the contingent future was the model for Gregory of Rimini. But having shown that such propositions are either determinately true or false, Marchia went on to articulate a type of prior determination that saved foreknowledge while preserving contingency. In fact, according to Marchia, there *must* be some determination in the causes of future contingent events prior to their actual occurrence, otherwise nothing would occur:

> And I ask about that determination in the cause, was it in the cause before the placing of the effect [into reality] or not? If it was, then I have my point. If not, I ask, how is the effect determined in its cause before it is put into being, necessarily or contingently? If necessarily, then it comes about necessarily, according to this opinion. If contingently and a contingent is not determined to one side in its cause, then that determination is not determined except through some prior contingent determination. And I would ask of this just as before, will it go on infinitely, or is it necessary to stop at some contingent determination in the cause before the effect? (d. 35: Marchia 1999, p. 75)

Marchia, however, was aware that Auriol had claimed that any such prior determination was fatal for

contingency, so Marchia draws a distinction between different indeterminations and determinations, perhaps expanding on isolated remarks made by Scotus. There is (1) an indetermination ‘about the possible’ (*de possibili*), with respect to being able to act and being able not to act. With this indetermination, we are *not* determined *de possibili* before an event, so we are free and act contingently and not necessarily. There is also (2) a posterior indetermination ‘about inhering’ (*de inesse*), with respect to what will be the case in reality. This indetermination toward what will inhere in reality, however, would be an *obstacle* to foreknowledge and, for us, to acting. Thus it must be replaced by (3) a determination in the contingent cause toward acting, both for the future to be known and for us to act. The (4) determination *de possibili*, toward being able to act or being able not to act, is absent from free causes until the event occurs, at which time our freedom and power with respect to that event are removed.

An obvious objection is that, for Marchia, the effect is determinate in the cause before the action of the cause, and thus that determination is ‘presupposed’ in the subsequent action of the cause. Since it is ‘presupposed’, that determination is not in the cause's power, and thus is not contingent. Marchia replies with another distinction:

> ‘Action’ can be taken in three ways: either it can be taken actually, namely when an agent is actually acting; or it can be taken virtually, when an agent can act although he is not acting; or it can be taken in a middle way, not purely actually nor purely virtually, but in a middle way as ‘dispositionally’ or ‘aptitudinally’, namely when an agent is not acting but is determined toward acting, although in actuality he is not acting — and he not only can act, but is determined to be acting later. Similarly there is a threefold ‘determination’ of the agent: one actual, by which an agent actually determinately puts one part of a contradiction into effect; a second is a potential determination by which an agent posits or can determine any part of a contradiction dividedly; the other is, as it were, a ‘dispositional’ or ‘aptitudinal’ determination, by which an agent is determined with respect to the future to putting one part of a contradiction [into effect]. Each determination presupposes the action corresponding to it, because an actual determination follows the action in actuality; the dispositional determination follows the action dispositionally, although it precedes the actual action; the potential determination follows the potential action, although it precedes that actual and dispositional action. (d. 35: Marchia 1999, pp. 89–90)

Thus when an agent is determined *de inesse* to doing something in the future, that determination is like a disposition, and neither actual, because the event has not yet occurred, nor potential, because the possibility to do otherwise is not removed. Such a determination is not ‘actually’ in the agent's power, Marchia grants, but it is in his power ‘dispositionally’, for although the agent cannot act before he acts, he can be disposed to act so that he will in fact act.

The foregoing example is representative of Marchia's thought in many ways. He frequently draws clever and original distinctions, and the *de possibili/ de inesse* division is employed in other contexts such as predestination (book I, d. 40: Marchia 2001). He makes similar innovative distinctions when discussing the different types of human and divine willing (*Reportatio* I, dd. 45-48: see attached edition). Although

one could argue about the cogency of Marchia's arguments, in the case of the *de possibili/ de inesse* distinction Marchia found a favorable response among the following, who adopted the device: his Franciscan successors at Paris in the next decade Aufredo Gonteri Brito, William Rubio, and William of Brienne; the Augustinian Hermits Michael of Massa and Gregory of Rimini; and, in the later fifteenth century, Fernando of Cordoba and Francesco della Rovere, who was to become Pope Sixtus IV. Thus in philosophical theology one could and often did look to Marchia for an alternative to Scotus and an innovative response to Auriol.

Natural Philosophy

Since the time of Pierre Duhem in the early twentieth century, Francis of Marchia has been known as a scientist, and looking through the titles of his questions, one finds an abundance of scientific topics related to such things as the infinite and the psychology of willing (Friedman-Schabel, 2001). In general, Marchia displays a great interest in the causal process. One thing that helps explain his popularity among historians of medieval science, and perhaps his own interest in scientific matters, is his clear and sharp distinction between natural causation that works necessarily and the contingent causation of human, angelic, and divine free will. In an influential passage containing echoes of Siger of Brabant, a passage to which Anneliese Maier first drew attention (*Scriptum* I, d. 36: Maier 1949; Marchia 2000), Marchia explains that there are two types of contingency in the world: first, there is contingency *per se, simpliciter*, *positiva*, and *intrinseca*. This is the contingency by which something is still able to occur or not occur even when all the required accidental, natural causes have been posited. There is only one source of such contingency: free will. Second, there is contingency *per accidens, secundum quid*, *privativa*, and *extrinseca*. This is the contingency of natural causation. A natural effect takes place as the result of many accidental causes. Some of these causes may be impeded by other natural, accidental causes, and so with respect to a small, limited number of causes a natural effect may be considered ‘contingent’. This does not mean that the natural effect is really and truly contingent without qualification in the first way, however, because if we take *all* of the natural effect's causes into account, the effect will *necessarily* follow, or not follow, as the case may be. That is, assuming God's contingent creation in the first place, and His ‘general influence’ that keeps the chain of causation in existence, natural causation works necessarily, and so with all of an effect's causes taken together, what happens in nature is necessary. Of crucial important for science is Marchia's further assertion that these ‘contingent’ effects can even be known by a *created* intellect. This is because the number of natural causes is not infinite. Thus a finite, created intellect can know the natural future with certainty. The only problem, says Marchia, is that we humans have a short life and an intellect that is bound with the body.

One of the most important innovations of the mature Galileo was the assertion that the celestial and terrestial realms are made of the same fundamental matter and therefore follow the same basic natural laws. Francis of Marchia put forth a similar hypothesis in his commentary on book II, qq. 29-32 (Marchia 1991b). Contrary to contemporary Aristotelian theory, Marchia argues that the heavens are not made up of a fifth, incorruptible, nobler element, which radically differentiates the supralunar realm from the sublunar one. On the contrary, the basic matter is the same everywhere, and just as Marchia considers the natural world to follow predictable patterns, he also thinks that those patterns are universally applicable.

These two tenets have important implications for the practice of natural philosophy (Schneider 1991).

With this attitude it is no wonder Marchia's physical theories drew the attention of medieval and modern scholars alike. The first important study on Marchia's thought was Anneliese Maier's partial edition and analysis of book IV, q. 1, of Marchia's *Sentences* commentary (Marchia 1940). There Marchia puts forth his famous forerunner to John Buridan's impetus theory of projectile motion. Aristotle had not provided a satisfactory explanation for why, when we throw a ball, for example, the ball keeps going even after we have released it. Marchia's explanation, which had its own partial predecessors, was that we leave behind a force in the ball — a *virtus derelicta* — that keeps the ball in motion. The fact that this force was temporary rather than permanent meant that it was not akin to the inertia theory of classical mechanics, but, as Marchia stated explicitly, it was a simple theory and did explain the phenomena in temporary projectile motion. Marchia also applied the theory to celestial motion, but he does not appear to have reconciled the semi-permanent nature of the motion of the heavens with the more ephemeral *virtus derelicta*. Maier saw reactions to Marchia's treatment in the works of Francis of Meyronnes, his follower Himbert of Garda, Nicholas Bonet, John Canonicus, William of Ockham, and Buridan, although it is hard to say whether he had much positive influence in this context.

Perhaps medieval scholars did not arrive at inertia because of their reluctance to consider motion in a vacuum. In observable projectile motion, at least in the fourteenth century, the projectile always ended up back on the ground. One of Aristotle's arguments for the impossibility of a vacuum was the lack of resistance of the medium. If velocity was a function of the proportion of force to the resistance of the medium, then with no resistance there would be motion in an instant, something usually considered an impossibility. In book II, q. 16, a. 5 of his *Sentences* commentary, however, Marchia has to wonder why an angel, which is not a "*corpus quantum*," or bodily mass, cannot move instantaneously, i.e. from one place to another without any temporal duration. Part of the answer is simply that it is a contradiction to be two places at once, but Marchia adds that there must be some sort of internal resistance in angels that makes instantaneous motion impossible. The notion of internal resistance unrelated to natural place (even though no *corpus quantum* is involved) and the concept of impetus (Buridan's making the *virtus derelicta* permanent) appear to be primitive versions of the ingredients of a theory of inertial mass. Since Marchia's writings predate those of his more famous successors, the Oxford Calculators and Buridan and Nicole Oresme at Paris, and some of his ideas at least resemble Galileo's in some way, Marchia's possible impact on later scientists is a good topic for future research.

Political and Social Thought

We are only just beginning to appreciate Marchia's thought in philosophical theology and natural philosophy. A fuller understanding requires, first of all, the critical edition of the *Sentences* commentary. We are in a better position to investigate where Marchia stands in political theory because of Mariani's publication of the *Improbatio* (Marchia 1993), probably written in the beginning of 1330. As in the case of Ockham, circumstances drove Marchia into opposition to Pope John XXII, and as a result Marchia's later writings focus on more worldly affairs. The group of scholars supporting the ex-minister general of the Franciscans Michael of Cesena gathered at Louis of Bavaria's court in Munich and collaborated on

their anti-John XXII tracts. It has been shown that Marchia's *Improbatio* influenced Ockham's *Opus nonaginta dierum* (Miethke 1969; Lambertini 2000) and most probably the Cesena group's *Appellatio magna monacensis* (Lambertini, forthcoming). Thus, as in philosophical theology and natural philosophy, Marchia had an historical impact.

The Franciscans maintained that they lived the most perfect life that was humanly possible, following the model of Christ and the apostles who, they claimed, possessed nothing either as individuals or in common. John XXII not only denied that Christ and the apostles had no possessions, but he also declared the Franciscan position to be heretical. The resulting quarrel came to touch on such issues as usury, ownership of property, disposal of property, natural and divine law and rights, papal infallibility, and ultimately the basis of sovereignty.

On the question on dominion, or possession, of property, Marchia accepted the pope's assertion that, even in the state of nature before the fall, Adam had dominion over the things he used, but Marchia denied that this type of dominion had much at all in common with post-lapsarian dominion: they are *alterius generis*, differing like violent and natural, like corruptible and incorruptible (Lambertini 2000, VII). Before the fall, Adam had "dominion of natural liberty and perfection," according to the "primaeval natural law"; afterwards, although Marchia admitted that there was a "remnant of natural law," basically it had to be replaced by positive law and the dominion of "servile necessity" and the "power of compulsion." This is because, if left to his own devices, post-lapsarian man would grab all he could get.

Marchia describes the situation in the state of nature thus:

> In the state of nature all things had been common to all people, not only with respect to the dominion of things but also with respect to use, whether *de iure* or *de facto* — with respect to use de iure because the right of using whatever was suitable to them had been common to all and proper to none. (Marchia 1993, pp. 154-5)

Moreover, even before the creation of Eve, Adam had no dominion over things that was "proper" to him. Therefore Pope John erred in claiming that the eventual "division of things" into private possessions stemmed from divine law, since it came from human or positive law, necessitated by human iniquity. The first division, by human law, occurred already with Cain and Abel before the Flood. Of course God restored common dominion to Noah and his sons with the Deluge, but humans and human law soon reinstated the division of goods. This was not by divine will, for "when they began to build the tower [of Babel] it was by human counsel, not by divine — indeed God was not pleased." By extension, the laws of emperors and princes derive from human law, not divine, with the exception of Hebrew law in the time before Christ, a special status lost at the crucifixion.

The upshot is that private property is not divinely instituted, for Marchia: it is a human introduction. In the state of nature, by divine law, humans had a common "dominion" over goods, but this dominion is not at all the dominion of human law, of private property. Moreover, it is this common dominion that Christ and the apostles followed. It is therefore humanly possible to do so, indeed the best possible way

to live, and the Franciscans approached it more closely than anyone else.

On the question of ecclesiastical power, of course, Marchia was also in disagreement with Pope John XXII (Lambertini 2000, IX). In order to defend Christ's poverty and that of the Franciscans by extension, Marchia interpreted Christ's remark to Pilate that "my kingdom is not of this world" as meaning not only that the origins of his kingdom and the source of his power are not of this world, but also that his kingdom and his power are not with respect to the things of this world. Indeed, for Marchia, Caesar had legitimate sovereignty over this world. Marchia thus denied that Christ, as a man, possessed any temporal power, and hence he rejected John XXII's claim that Christ was *rex* and *dominus* in a temporal sense. The implication is that the pope could not inherit temporal power over the entire earth, although Marchia admitted that the temporal sword could pertain to the pope indirectly and mediately.

Given that the few studies that have been done demonstrate that Marchia had an impact on his successors in many areas of philosophical thought, and given that it is only in the past decade that any substantial part of his works has been available in print, we can in the near future expect a flood of investigations of the ideas of this interesting and important scholastic thinker.

Critical Editions of Texts

The following texts are critical editions, using all known manuscript witnesses, which will be published in the future with apparatus criticus and fontium. Distinctions 35-48 of book I of the *Sentences* concern the unified themes of divine knowledge, power, and will. Marchia's *Scriptum* for book I ends prematurely at distinction 40, and distinction 39 on divine ideas is much trunctated. Since distinctions 35-40 of the Scriptum have been or are being published (Marchia 1999b, 2000, 2001), the following texts from the *Reportatio* for book I, distinctions 39, 42-44, and 45-48, will complement the *Scriptum* material on these subjects.

- Francis of Marchia, *Reportatio in primum librum Sententiarum*, Distinctio 39, 42-44, 45-48
- Francis of Marchia, *In secundum librum Sententiarum*, Quaestio 16, Articulus 5

Bibliography

Texts

- Marchia 1940: IV *Sent.*, q. 1: ed. A. Maier, “Franciscus de Marchia,” A. Maier, *Die Impetustheorie* (Vienna-Leipzig 1940), 45–77; reprinted A. Maier, *Zwei Grundprobleme der scholastischen Naturphilosophie* (= Storia e Letteratura, 37) (Rome 1968), 161–200; English trans. of text in M. Clagett, *The Science of Mechanics in the Middle Ages* (Madison 1959), 526–30.

- Marchia 1965: *Metaphysics* I, q. 1, and VI, q. 16: ed. A. Zimmermann, *Ontologie oder Metaphysik? Die Diskussion über den Gegenstand der Metaphysik im 13. und 14. Jahrhundert. Texte und Untersuchungen* (= Studien und Texte zur Geistesgeschichte des Mittelalters, 8) (Leiden 1965).
- Marchia 1986: *Metaphysics* II, q. 5: ed. N. Schneider, "Eine ungedruckte Quaestio zur Erkennbarkeit des Unendlichen in einem Metaphysik-Kommentar des 14. Jahrhunderts," in A. Zimmermann, ed., *Miscellanea Mediaevalia* 18 (Berlin 1986), 96–118.
- Marchia 1991a: *Metaphysics* III, q. 9, in Schneider 1991.
- Marchia 1991b: II *Sent.*, qq. 29–32, in Schneider 1991.
- Marchia 1993: *Francisci de Esculo, OFM, Improbatio contra libellum Domini Johannis qui incipit Quia vir reprobus*, ed. N. Mariani (= Spicilegium Bonaventurianum, 28) (Grottaferrata 1993).
- Marchia 1997a: *Scriptum* I, d. 27: ed. R.L. Friedman "*In principio erat Verbum*: The Incorporation of Philosophical Psychology into Trinitarian Theology, 1250–1325" (PhD Dissertation, University of Iowa 1997), 555–72.
- Marchia 1997b: *Francisci de Marchia sive de Esculo, OFM, Quodlibet cum quaestionibus selectis ex commentario in librum Sententiarum*, ed. N. Mariani (= Spicilegium Bonaventurianum, 29) (Grottaferrata 1997).
- Marchia 1998: *Francisci de Marchia sive de Esculo, OFM, Sententia et compilatio super libros Physicorum Aristotelis*, ed. N. Mariani (= Spicilegium Bonaventurianum, 30) (Grottaferrata 1998).
- Marchia 1999a: *Scriptum* I, d. 11: ed. R.L. Friedman, "Francis of Marchia and John Duns Scotus on the Psychological Model of the Trinity," *Picenum Seraphicum* 18 n.s. (1999), 11–56.
- Marchia 1999b: *Scriptum* I, d. 35: ed. C. Schabel, "Il determinismo di Francesco di Marchia (Parte I)," *Picenum Seraphicum* 18 n.s. (1999), 57–95.
- Marchia 2000: *Scriptum* I, dd. 36, and 38: ed. C. Schabel, "Il determinismo di Francesco di Marchia (Parte II)," *Picenum Seraphicum* 19 n.s. (2000), 3–55.
- Marchia 2001: *Scriptum* I, dd. 39–40: ed. C. Schabel, "La dottrina sulla predestinazione di Francesco di Marchia," *Picenum Seraphicum* 20 n.s. (2001), forthcoming.

Secondary Literature

- Bakker, P.J.J.M. 1999: *La raison et miracle. Les doctrines Eucharistiques (c. 1250 – c. 1400)*, 2 vols. (Nijmegen 1999).
- Duhem, P. 1985: *Medieval Cosmology. Theories of Infinity, Place, Time, Void, and the Plurality of Worlds*, trans. R. Ariew (Chicago 1985).
- Friedman, R.L., and C. Schabel 2001: "Francis of Marchia's Commentaries on the Sentences: Question List and State of Research," *Mediaeval Studies* 63 (2001), forthcoming.
- Kürzinger, J. 1930: *Alfonsus Vargas Toletanus und seine theologische Einleitungslehre. Ein Beitrag zur Geschichte der Scholastik im 14. Jahrhundert* (= BGPTM 22.5-6) (Münster i. W. 1930).
- Lambertini, R. 2000: *La povertà pensata. Evoluzione storica della definizione dell'identità minoritica da Bonaventura ad Ockham* (Modena 2000), papers VII, VIII, and IX.

- Lambertini, R., forthcoming: “Francesco d'Ascoli e la polemica francescana contro Giovanni XXII: A proposito dei rapporti tra l'*Improbatio e l'Appellatio magna monacensis*.”
- Lang, A. 1930: *Die Wege der Glaubensbedrüngen bei den Scholastikern des 14. Jahrhunderts* (= BGPTM 30.1-2) (Münster i. W. 1930).
- Maier, A. 1949: *Die Vorläufer Galileis im 14. Jahrhundert* (= Storia e Letteratura, 22) (Rome 1949).
- Maier, A. 1950: *An der Grenze von Scholastik und Naturwissenschaft* (= Storia e Letteratura, 41) (Rome 1952).
- Maier, A. 1955: *Metaphysische Hintergründe der spätscholastischen Naturphilosophie* (= Storia e Letteratura, 52) (Rome 1955).
- Maier, A. 1958: *Zwischen Philosophie und Mechanik* (= Storia e Letteratura, 69) (Rome 1958).
- Miethke, J. 1969: *Ockhams Weg zur Sozialphilosophie* (Berlin 1969).
- Schabel, C. 2000: *Theology at Paris 1316–1345: Peter Auriol and the Problem of Divine Foreknoweldge and Future Contingents* (= Ashgate Studies in Medieval Philosophy, 1) (Aldershot 2000).
- Schneider, N. 1991: *Die Kosmologie des Franciscus de Marchia: Texte, Quellen, und Untersuchungen zur Naturphilosophie des 14. Jahrhunderts* (= Studien und Texte zur Geistesgeschichte des Mittelalters, 28) (Leiden 1991).
- Teetaert, A. 1933: “Pignano (François de),” *Dictionnaire de Théologie Catholique* XII (Paris 1933), cols. 2104–2109.
- Wittneben, E.L., and R. Lambertini 1999: “Un teologo francescano alle strette. Osservazioni sul testimone manoscritto del processo a Francesco d'Ascoli,” *Picenum Seraphicum* 18 n.s. (1999), 97–122.
- Zimmermann, A. 1966: “Allgemeine Metaphysik und Teilmetaphysik nach einen anonymen Kommentar zur aristotelischen Ersten Philosophie aus dem 14. Jahrhundert,” *Archiv für Geschichte der Philosophie* 48 (1966), 190–206.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Auriol [Aureol, Aureoli], Peter | Buridan, John [Jean] | Duns Scotus, John | Gregory of Rimini

Christopher Schabel
University of Cyprus
schabel@ucy.ac.cy

First published: March 23, 2001
Content last modified: March 23, 2001

In secundum librum Sententiarum by Francis of Marchia

Quaestio 16, Articulus 5

Quantum ad quintum, utrum videlicet angelus possit moveri in instanti, videtur quod sic, immo quod non possit successive moveri, et hoc per rationes quas facit Philosophus, quarto Physicorum, ad probandum quod non possit aliquid moveri in vacuo. Arguit enim ibi, capitulo de vacuo, primo sic: qualis est proportio medii ad medium in densitate et raritate, talis est proportio motus ad motum in velocitate et tarditate; sed vacui ad plenum nulla est proportio in raritate et densitate; ergo nec motus facti in pleno ad motum factum in vacuo erit aliqua proportio in velocitate et tarditate. Constat autem quod ista ratio Philosophi tenet ex hoc quod medium magis vel minus resistit motori seu etiam mobili secundum quod est magis vel minus densum. Quanto enim est in medio minor resistentia, tanto motus est velocior, et econverso tanto motus tardior quanto resistentia maior.

<1> Tunc potest argui ad propositum primo sic: ubi nulla est proportio medii ad medium quantum ad resistentiam, nec est aliqua proportio motus ad motum quantum ad velocitatem et tarditatem; sed medii per quod angelus movetur ad medium per quod corpus movetur nulla est proportio quantum ad resistentiam, cum medium per quod angelus movetur nullo modo resistat angelo; ergo nec motus angeli ad motum corporis erit aliqua proportio in velocitate; ergo est in instanti.

Confirmatur, quia secundum Commentatorem ibidem, divisio sive successio non causatur in motu nisi ex vel resistentia medii ad mobile vel ex resistentia mobilis ad motorem; sed in motu angeli non est aliqua resistentia medii ad mobile nec mobilis ad motorem; ergo etc.

<2> Preterea secundo, Philosophus arguit ibi deducendo ad inconveniens, quod si fieret aliquis motus in vacuo, sequeretur quod in eodem et in equali tempore posset esse motus in vacuo et in pleno.

Consimiliter, ego ostendo quod si angelus movetur in tempore, sequitur quod aliquod corpus sensibile possit moveri in eodem et in equali tempore sicut angelus, quod tamen est impossibile. Probatio assumpti: quanto aliquod corpus est minus, tanto potest moveri in minori tempore ab eadem virtute; sed corpus quantum quodcumque est divisibile in infinitum; ergo si angelus movetur in tempore, corpus etiam. Accipio aliquod corpus, puta B. In tantum B poterit dividi quod eque velociter movebitur et in equali tempore ab eodem motore sicut ipse angelus, immo etiam in minori. Hoc autem est impossibile; ergo etc.

Dico tamen quantum ad hoc quod angelus virtute sua non potest naturaliter moveri localiter in instanti. Quod probo sic: illud quod in eodem instanti movetur de loco ad locum per medium in eodem instanti est in termino a quo et in termino ad quem; sed angelus non potest simul esse in pluribus locis sibi equalibus; ergo non potest de loco sibi equali et proportionato moveri in instanti ad alium distantem ab illo loco priori.

<-1p> Ad rationem primam in oppositum, forte posset dici uno modo quod duplex est successio, una essentialis, et ista non est ex aliqua resistentia medii ad mobile nec mobilis ad motorem, sed est ex divisione medii per quod fit motus. Alia est successio accidentalis motus penes quam attenditur velocitas et tarditas in motu. Sic enim se videntur habere velocitas et tarditas in motu sicut se habent raritas et densitas in quantitate permanente. Et ideo, sicut quantitas permanens videtur habere certum terminum quantitatis quantum ad maius et quantum ad minus — aliter non diceret Philosophus quod est dare minimam carnem et minimum os — et quod omnium natura constantium est certus terminus et magnitudo, et ita potest aliquod quantum permanens in tantum rarefieri quod non potest plus vel etiam condensari, ita etiam videtur quod licet motus possit intendi et remitti quantum ad velocitatem et tarditatem, est tamen dare terminum utrobique. Et ita diceretur quod motus angeli habet successionem absque aliqua resistentia.

<-1r> Sed tamen dico aliter sequendo Commentatorem, et concedo cum ipso quod omnis motus successivus est successivus propter aliquam resistentiam. Unde dico quod motus angeli est successivus propter resistentiam mobilis ad motorem. Ubi tamen est advertendum quod mobile resistere motori potest esse duplici de causa: uno modo aliquod mobile resistit motori ex hoc quod habet inclinationem naturalem ad aliquod ubi oppositum illi ubi ad quod movetur, sicut grave resistit moventi ipsum sursum quia habet inclinationem ad oppositum, puta ad ubi deorsum. Isto modo celum non resistit angelo moventi ipsum, quia tunc motus celi esset violentus. Alio autem modo aliquod mobile potest resistere suo motori non quia habeat inclinationem ad oppositum, sed solum quia non habet perfectam obedientiam ad ipsum. Quia enim istud mobile, quodcumque sit, non potest simul esse naturaliter in pluribus locis, ideo quando est in uno loco non potest esse in alio. Nec est in perfecta obedientia respectu alicuius agentis finiti quod possit moveri ab isto loco et poni in alio in quacumque mensura. Et sic, quando movetur ab uno loco ad alium, resistit motori, cum non sit in perfecta obedientia eius ut moveatur sive transferatur de loco ad locum in quacumque mensura.

Dico ergo quod non solum repugnantia contraria mobilis ad motorem, qualis est in motu violento propter inclinationem mobilis ad ubi oppositum, est causa successionis motus, sed etiam resistentia privativa, qualis est in quocumque motu locali cuiuscumque rei finite, sive corporalis sive spiritualis, facto a virtute

finita. Ex quo concedo quod angelus potest movere et se ipsum et alia successive et non in instanti propter rationem iam dictam, quia licet in eius motu quo movet se localiter non sit resistentia mobilis ad motorem positiva contraria, est tamen ibi, ut dictum est, resistentia privativa.

<-2p> Ad secundum, dico quod impossibile est corpus aliquod posse moveri in tam brevi tempore sicut angelus movetur. Ad probationem, quando arguitur, ‘quanto corpus est minus, tanto potest in minori tempore moveri, quia tanto minus resistit motori’, hic posset dici quod non est verum. Licet enim requiratur debita proportio inter movens et mobile, non tamen semper minus mobile potest ab eodem motore moveri citius, quia non semper minus est sibi magis proportionatum. Citius enim et facilius movet idem movens aliquem maiorem lapidem quam minimum.

<-2r> Hic tamen dico aliter, et magis ad propositum, quia quod maius mobile eadem virtus velocius et citius moveat, hoc forte provenit aliunde, videlicet ex maiori et minori resistentia medii ad mobile, vel ex aliquo alio. Et ideo dico sic: quandocumque aliqua sunt alterius rationis, quantumcumque unum augeatur in infinitum, numquam tamen propter hoc potest pertingere ad perfectionem alterius perfectioris illo secundum suam rationem specificam et formalem. Exemplum: angulus rectus et acutus sunt alterius rationis, et ideo si acutus crescat in infinitum, numquam propter hoc perveniet ad equalitatem anguli recti, nec erit sibi equalis. In proposito etiam, subtilitas corporis et subtilitas angeli sunt alterius rationis, et ideo, licet corpus quantum quanto magis rarefit, tanto fit subtilius, tamen esto quod in infinitum rarefieret, numquam adhuc ad subtilitatem angeli pervenire posset.

Tunc per hoc dico ad propositum quod corpus mobile resistit virtuti motive angeli, angelus etiam resistit sibi ipsi. Sed ista resistentia qua angelus ut mobile resistit sue virtuti motive alterius rationis est ab illa resistentia qua corpus resistit sibi ut moventi vel cuicumque alteri, et hec est minor illa. Et ideo, quantumcumque corpus diminueretur sive divideretur in infinitum, et sic eius resistentia qua resistit motori per consequens diminueretur, numquam propter hoc ista resistentia posset adequari illi. Nec corpus posset esse in illa perfecta obedientia ad motum localem respectu angeli, nec etiam respectu alicuius alterius, sicut est ipse angelus. Sic ad questionem.

Return to section "Natural Philosophy" in the entry Francis of Marchia
Return to section "Critical Edition of Texts" in the entry Francis of Marchia

Stanford Encyclopedia of Philosophy
Supplement to Francis of Marchia

Reportatio in primum librum Sententiarum by Francis of Marchia

Disinctio 39

Circa distinctionem 39 quero utrum in Deo sint plures vel infinite idee.

Et videtur quod non, quoniam illud quod est imperfectionis non est ponendum in Deo in quo nulla imperfectio potest esse; sed multitudo infinitarum idearum dicit imperfectionem, dicit enim confusionem sine ordine; ergo non sunt in Deo infinite idee. Sed si non sunt infinite, non oportet in eo ponere finitas nec aliquam per consequens, cum ipse intelligat infinita et ipse idee ponantur propter obiecta cognita; ergo etc.

Contra: Augustinus, in libro *83 questionum*, questione 46, "tanta," inquit, "ibi vis est in ideis ut nemo sine eis sapiens esse possit."

Item, 12 *De civitate Dei*, capitulo 18, dicit quod quamvis infinitorum non sit numerus, non est tamen apud illum incomprehensibilis infinitas numerorum cuius scientie non est numerus.

Circa istam questionem sic procedam: primo enim excludam aliquos modos dicendi; secundo dicam aliter.

[Articulus primus]

Quantum ad primum, dicitur quod in Deo sunt omnium intelligibilium distincte et proprie idee. Cuius ratio est quoniam omnis intellectus qui intelligit aliquid indiget ratione intelligendi ad intelligendum illud, per quam quidem intelligit illud; sed intellectus divinus intelligit essentiam suam et etiam omnia alia a se;

ergo oportet quod habeat rationem intelligendi se et etiam alia a se. Sed idem sub eadem ratione non potest esse ratio intelligendi immediate plura ut plura et distincta; essentia autem est formaliter unica, alia autem ab essentia, que quidem intelligit, sunt plura, immo etiam infinita; ergo preter essentiam oportet ponere in Deo rationes intelligendi proprias singulorum que intelligit.

Ad hoc etiam videtur esse intentio Augustini, *83 questionibus* ubi supra, ubi loquens de ideis sic dicit: "Idee sunt forme principales rerum stabiles atque incomprehensibiles, ac per hoc nec formate nec create sunt, sed tamen per eas formari dicitur quicquid formatur et oritur."

[Primus modus dicendi]

Et in ista conclusione concordant multi, licet diversimode. Unde iuxta hoc est hic triplex modus dicendi. Primus modus est quod idee sunt rationes intelligendi, et ideo precedunt ipsum actum intelligendi divinum in ipsa essentia formaliter existentes.

Modus autem ponendi istorum talis est: Dicunt enim quod in essentia divina correspondet propria perfectio cuiuslibet nature specifice distincta a perfectione correspondente alteri specierum, sicut aliqua perfectio propria est in divina essentia formaliter que correspondet perfectioni lapidis et alia distincta ab ista que correspondet perfectioni hominis, et sic de omnibus aliis speciebus. Tunc dicunt quod iste perfectiones in divina essentia formaliter representant perfectiones specierum quibus correspondent, et iste sunt idee ipsarum, sicut perfectio in divina essentia correspondens perfectioni hominis et ipsam representans distincte est eius idea, et sic de aliis. Ita quod secundum hoc idee non sunt nisi quedam divine perfectiones speciales in essentia divina distinctis perfectionibus creaturarum quarum sunt idee correspondentes et ipsas representantes, que quidem sunt rationes intelligendi illa obiecta que representant distincte incommutabiles et eterne.

Pro ista opinione est ratio et auctoritas Augustini. Ratio est ista: ratio intelligendi videtur esse prior ipso actu intelligendi, cum sit medium eius; sed idee sunt rationes intelligendi et sunt distincte; ergo sunt in divina essentia ante actum.

Preterea, omne quod cognoscitur cognoscitur per similitudinem aliquam; sed idem inquantum idem non potest esse similitudo plurium ut plura sunt; ergo cum divinus intellectus cognoscat plura perfecte et distincte, oportet quod habeat in essentia sua aliqua plura distincta in quibus fundetur similitudo ad diversa obiecta cognita. Ista autem non sunt aliud quam huiusmodi predicte plures perfectiones correspondentes distinctis creaturarum perfectionibus et ipsas representantes, ergo etc.

Ad hoc etiam secundo est auctoritas Augustini, *83 questionibus*, questione 48, ubi sic dicit: "Omnia sunt condita propriis rationibus. Nec enim," inquit, "eadem ratione conditus est homo qua equus." Unde concludit singula, ergo propriis sunt creata rationibus.

Preterea, 6 *De trinitate*, capitulo ultimo, loquens de verbo divino, dicit quod est ars "Dei Patris plena omnium rationum viventium et incommutabilium."

Sed iste modus dicendi non videtur esse verus. Primo, quia videtur ponere aliqua limitata in Deo formaliter. Secundo, quia videtur ponere pluritatem sine necessitate.

Primum ostendo sic: omnis perfectio representans determinate unam perfectionem tantum et non aliam est limitata in representando, et per consequens in essendo, quoniam limitatio in representando non venit nisi ex limitatione sui in essendo; sed secundum istam opinionem perfectio cuiuslibet rei, puta perfectio correspondens homini, in divina essentia representat tantum hominem et non alia, et sic de aliis perfectionibus in divina essentia correspondentibus aliis; ergo quelibet perfectio correspondens cuilibet perfectioni creaturarum in divina essentia est limitata in representando, et per consequens in essendo.

Secundum, videlicet quod ponat pluritatem sine necessitate, probatur: quoniam constat quod divinus intellectus intelligit huiusmodi perfectiones proprias quas vocat ideas, quero per quid? Aut enim intelligit eas per essentiam suam aut per seipsas. Si per essentiam suam, ergo idem sub eadem ratione potest esse ratio intelligendi plura distincta, et ita per consequens non oportet ponere ad intelligendum plura aliqua plura distincta que sint rationes intelligendi ipsa. Si autem dicas quod non per essentiam, sed per seipsas, contra, quia intellectus intelligens plura eque primo per rationes proprias intelligit ea distinctis actibus; sed intellectus divinus eodem actu numero quo intelligit se intelligit quecumque alia a se; ergo et eadem ratione intelligendi intelligit omnia. Non enim actus intelligendi potest esse magis illimitatus quam medium sive ratio intelligendi; ergo necessario unico actu semper correspondet unica ratio intelligendi, licet non econverso, quia ratio intelligendi potest esse illimitatior ipso actu, cum possit esse eadem ratio multorum actuum; ergo etc.

Confirmatur, quia actus intelligendi est similitudo obiecti expressior specie intelligibili quacumque; sed per unicam speciem non possunt representari plura sub propriis rationibus eorum; ergo nec per unicum actum possunt plura distincte intelligi per rationes proprias singulorum; ergo etc.

Preterea, non est illimitatior actus intelligendi in genere actus secundi quam essentia in genere actus primi; ergo sicut divina intellectio est unica respectu omnium intelligibilium, ita et ratio intelligendi omnia erit unica, puta essentia ipsa; ergo etc.

[Secundus modus dicendi]

Alius modus dicendi est qui ponit huiusmodi rationes intelligendi creaturas, que sunt idee, non esse in divina essentia ante actum intelligendi formaliter, ut ponit prima opinio, sed tantum ponit eas esse in ea obiective. Ante enim actum intelligendi quodlibet relucet in divina essentia propria perfectio eius, que quidem ut in divina essentia relucentes ante intellectionem cuiuscumque sunt rationes intelligendi ipsas res ut extra existentes. Et ita isti duo modi concordant in hoc, quia uterque ponit ideas esse rationes intelligendi obiecta secundaria precedentes in divina essentia intellectionem cuiuslibet. Discordant autem in hoc, quia primus modus dicendi ponit huiusmodi rationes esse in essentia formaliter, secundus autem non, sed tantum obiective.

Contra istum modum arguo sicut contra primum, quia cum intellectus divinus intelligat huiusmodi

rationes quas ponis in essentia obiective, aut intelligit eas per essentiam divinam aut per seipsas. Non per se, quia non sunt idem cum essentia, non potest autem intelligere aliquid per se quod non est idem cum essentia, quia tunc acciperet notitiam ab alio a se, et ita perficeretur ab ipso aliquo modo, quod est impossibile. Ergo per essentiam. Ergo, cum illa sint plura et ab essentia distincta, idem potest esse ratio intelligendi plura distincte. Ergo etc.

Preterea, intellectus intelligens unico actu omne intelligibile unica ratione intelligit omnia; sed intellectus divinus est huiusmodi; ergo etc. Ergo huiusmodi rationes plures ideales non sunt rationes intelligendi, cum ratio intelligendi istius intellectus sit unica numero sicut et actus eius.

[Tertius modus dicendi]

Tertius modus dicendi est quod idee sunt rationes non precedentes actum intelligendi, sed magis ipsum sequentes. Sunt enim, ut dicunt, quedam relationes rationis facte per intellectum ad ipsa obiecta secundaria, ita quod diverse idee et proprie rerum non sunt nisi diverse relationes rationis formate per intellectum ad diversa obiecta secundaria. Que quidem non sunt rationes intelligendi ipsa obiecta secundaria, immo essentia est ratio intelligendi ea, sed sunt rationes sub quibus essentia est ratio propria intelligendi quodlibet. Nec enim determinata per unam relationem rationis terminatam ad aliquid est propria ratio intelligendi illud ut per aliam aliud et sic de aliis. Dicunt enim quod essentia divina potest considerari dupliciter, quia vel in se absolute — et sic non est ratio intelligendi nisi se ipsam tantum; vel ut est sub distinctis relationibus rationis ad distincta obiecta intelligibilia secundaria — et ut sic essentia est ratio intelligendi illa, et ideo ipsa essentia est ratio intelligendi. Idee autem sunt rationes sub quibus ipsa essentia est ratio intelligendi.

Et iste modus dicendi est tripertitus. Quidem enim ponunt huiusmodi relationes rationis ad obiecta secundaria, que quidem sunt idee, esse in essentia divina per modum obiecti. Dicunt enim quod essentia divina est aliquo modo imitabilis a qualibet creatura. Et ita dicunt quod essentia ut imitabilis habet rationem idee, ut idea formaliter non sit nec dicat aliquid aliud ab essentia divina quam respectum imitabilitatis, secundum quorum diversitatem et multiplicitatem multiplicantur et plurificantur ipse idee.

Alii vero ponunt huiusmodi respectus esse in essentia divina ut ipsa habet rationem intelligendi sive ut se tenet ex parte rationis intelligendi.

Alii dicunt quod sunt sive fundantur in ipso actu intellectionis. Ut enim actus intelligendi divinus habet ad istud obiectum secundarium proprium respectum habet rationem idee respectu illius et intelligit illud per ipsum actum ut sub illo respectu.

Sed contra omnes istos tres modos arguo. Primo sic: Omnis respectus rationis presupponit terminum proprium seu obiectum; sed termini seu obiecta istorum respectuum sunt ipsa intelligibilia extra; ergo prius ordine nature divinus intellectus intelligit distincte ipsa obiecta secundaria quam ponat huiusmodi respectus rationis ad ipsa; ergo essentia non est ratio intelligendi ea sub istis respectibus, sed per se.

Confirmatur, quia intellectus non ponit huiusmodi respectus rationis nisi circa obiectum preconceptum; ergo prius natura oportet divinum intellectum obiecta secundaria preconcipere quam huiusmodi respectus rationis ad ipsa in essentia ponere. Quero tunc per quid ea intelligit? Constat quod non per essentiam sub istis respectibus, quia nondum sunt. Ergo per essentiam tantum; quare etc.

Preterea, constat quod divinus intellectus intelligit huiusmodi respectus rationis primo et distincte. Quero per quid? Aut enim intelligit eos per essentiam immediate aut per eam ut sub aliis respectibus rationis. Si primum, ergo idem potest esse ratio intelligendi plura distincte. Frustra ergo ponuntur huiusmodi plures rationes intelligendi ad hoc quod divinus intellectus distincte intelligere possit plura. Si autem per alios respectus, queram de illis, per quid eos intelligit? Et sic in infinitum.

[Articulus secundus: Opinio propria]

Et ideo dico aliter quod Deus quecumque intelligit, sive se sive alia a se, omnia intelligit immediate per essentiam suam, non per aliquid quomodocumque aliud a se. Et ideo non pono ideas ut rationes intelligendi divino intellectui alia, quia nec sic invenio quod eas posuerit Augustinus. Essentia enim absolutissime accepta sine determinatione quacumque est ratio intelligendi quecumque intelligibilia. Et ideo idee non ponuntur ut rationes intelligendi nec etiam producendi, sed ponuntur tantum propter ordinem actuum, ut inferius ostendetur.

Unde circa hoc sunt tria videnda: primo, quomodo idem sub eadem ratione potest esse ratio intelligendi plura distincte; secundo, ad quid ponuntur idee; tertio, quale esse habent idee.

[Prima quaestio secundi articuli]

Quantum ad primum ostendo hoc primo ex parte obiecti, ubi advertendum quod triplex est genus obiectorum: aliquid enim est obiectum finitum actu et potentia, ut quodlibet individuum, ut Sortes vel Plato, et tale obiectum est ratio intelligendi tantum finita actu et potentia. Aliud est infinitum in potentia, sed finitum in actu, ut quodlibet universale, saltem species que continet sub se infinita individua in potentia, et tale est ratio intelligendi infinita in potentia, modo videlicet quo ipsum est infinitum, non autem infinita in actu distincte. Tertio est aliud quod est in actu simpliciter infinitum, ut divina essentia, que est infinita in actu infinite et extra genus.

Tunc potest ex istis argui ad propositum sic: sicut se habet infinitum in potentia quantum ad hoc quod est esse rationem intelligendi infinita in potentia, ita infinitum in actu quantum ad hoc quod est esse rationem intelligendi infinita in actu; sed infinitum in potentia, puta homo vel quecumque alia species, est ratio intelligendi infinita in potentia, licet confuse et indistincte; ergo infinitum in actu simpliciter, cuius est divina essentia, potest esse ratio intelligendi distincte et perfectissime infinita in actu.

Preterea, obiectum habens unam rationem equivalentem in perfectione infinitis rationibus distinctis potest esse sufficientissime ratio intelligendi infinita intelligibilia que per illas rationes infinitas, quibus ista ratio unica equivalet, possunt intelligi — ista patet, continens enim aliqua eminenter potest esse ratio

intelligendi illa acsi ea formaliter contineret; sed divina essentia unica equivalet perfectionibus creaturarum, immo per unicam rationem supereminenter continet omnes; ergo ipsa per unicam rationem potest esse ratio intelligendi omnia.

Secundo declaro hoc idem ex parte potentie. Ubi advertendum quod nos per eandem numero potentiam sine distinctione aliqua ex parte potentie possumus successive intelligere infinita cum sola distinctione et successione actuum. Licet enim eadem potentia omnia intelligamus, tamen non eodem actu, nec eadem ratione intelligendi, sed alio et alio, et hoc est ex limitatione nostri actus intelligendi, sicut ex illimitatione potentie, quod eadem indistincta possit in infinita obiecta. Unde si nos haberemus unum actum intelligendi eque illimitatum in genere actus sicut est potentia in genere potentie, et semper stantem ut ipsa sicut erit in patria — quia ibi, secundum Augustinum, non erunt volubiles cogitationes, sed omnia unico intuitu sive actu videbimus — tunc quicquid intelligimus, intelligeremus per illum actum potentie adequatum, sicut per eandem potentiam. Nunc autem non, quia potentia est illimitatior ipso actu.

Consimiliter dico de ratione intelligendi sicut de actu, quoniam si ipsa esset eque illimitata sicut potentia, intelligeremus per unicam rationem intelligendi; sed divinus intellectus, quia illimitatior nostro, cum sit infinitus, habet in se unicum actum semper stantem sibi adequatum et eque illimitatum sicut ipse habet unicam rationem intelligendi adequatam et illimitatam, videlicet essentiam suam, que quidem est eque illimitata in ordine rationis intelligendi sicut est potentia intellectiva in ratione principii intellectivi. Ex quo patet quod sicut non oportet in Deo ponere plures actus intelligendi ad hoc ut plura distincte intelligat, ita eadem ratione nec plures rationes intelligendi, cum possit per unicam rationem sicut et per unicum actum intelligere quecumque intelligibilia distinctissime et perfecte.

[Secunda quaestio secundi articuli]

Quantum ad secundum, ubi est videndum ad quid ponuntur idee, ex quo non ponuntur sicut rationes intelligendi, ut dictum est, dico quod ponuntur propter ordinem actuum.

Ad cuius evidentiam sciendum est quod agens per intellectum habens duas vel plures operationes subordinatas, vel unam continentem duas sic subordinatas sicut habens operationem immanentem et transeuntem, vel unam utramque continentem, propter ordinem istarum, sicut operatio immanens est prior, ita prius agens habet terminum operationis immanentis quam terminum operationis transeuntis. Vel si sit idem terminus utriusque, primo habet ipsum ut est terminus operationis immanentis quam ut est terminus transeuntis. Tunc ad propositum, Deus habet circa creaturam duplicem operationem, videlicet immanentem, ut velle et intelligere, et transeuntem, sicut actum creandi, qui est quasi transiens secundum rationem, non tamen realiter magis quam operatio immanens, quia nec iste due operationes sunt, ut puto, distincte in Deo. Deus ergo propter ordinem istorum actuum predictorum prius habet terminum intellectionis quam terminum creationis. Idee autem non sunt nisi ipsa obiecta secundaria actus intelligendi ut actum ipsum terminantia, ita quod lapis ut cognitus est idea lapidis ut lapis est terminus actionis sive operationis transeuntis, puta creationis. Et ita de aliis secundariis obiectis omnibus divini intellectus, ita quod sicut operatio immanens est quasi via ad operationem transeuntem, ita terminus operationis immanentis ad terminum operationis transeuntis. Et ita idee non sunt rationes proxime intelligendi, nec rationes etiam producendi ex parte producentis, ut imaginatur quidam. Deus enim eadem

volitione qua ab eterno voluit res contingenter, illa semper stante eadem ponit res in esse pro tempore pro quo ponit.

Hec videtur esse intentio Augustini, qui dicit quod idee sunt forme rerum principales atque incommutabiles que divina intelligentia continentur, et hoc per modum obiectorum, ut dictum est, non formaliter.

Tunc secundum hoc dico quod in Deo sunt infinite idee. Quot enim sunt obiecta cognita tot sunt idee, quodlibet enim habet propriam ideam; obiecta autem cognita sunt in Deo infinita, cum cognoscat infinita; igitur etc.

[Tertia quaestio secundi articuli]

Quantum ad tertium, quale videlicet esse habeant huiusmodi idee, utrum videlicet quidditates rerum cognite ab eterno ab intellectu divino, que quidem, ut dictum est, sunt idee ipse, habeant aliquod esse distinctum ab actu intelligendi, vel sint penitus idem cum actu.

Aliqui dicunt quod non. Dicunt enim quod lapis intellectus vel volitus ut sic non est aliud ab actu divine intellectionis vel volitionis. Et hec etiam videtur esse de intentione Augustini 5 *Super Genesim*, capitulo XV, et 4 *De trinitate*. 5 enim *Super Genesim*, super illo verbo: "Quod factum est in ipso vita erat," probat Augustinus dupliciter quod ista littera non debet sic punctuari: ‘Quod factum est in ipso, vita erat’, sed sic: ‘Quod factum est, in ipso vita erat’. Cuius ratio sua prima ibi est quod primo modo punctundo videtur innui quod illud quod factum est non sit factum per ipsum, sed tantum in ipso, cum tamen omnia sint facta per ipsum et sint in ipso.

Secunda ratio sua ibi est quia tunc sequeretur quod terra vel quodcumque aliud inanimatum esset vita, cum quodlibet sit factum in ipso. Et concludit postea sic: "Ergo quod factum est, in ipso vita erat antequam fieret. Et non qualitercumque vita, quia non qualis est vita pecudum, sed que erat lux homini," hec est intellectualis; sed si creature cognite haberent esse distinctum ab ipso, constat quod quantum ad illud esse non essent vita intellectualis; ergo etc.

Preterea, secundum eundem, idee sunt forme principales, non facte seu formate, <sed> incommutabiles et eterne; constat autem quod solus Deus est talis; ergo sequitur quod quidditates rerum cognite non habent aliquod esse distinctum ab actu divine cognitionis.

Confirmatur, quia si haberent aliquod aliud esse, illud esset formatum seu constitutum per actum intelligendi seu volendi; sed secundum Augustinum, idee non sunt formate; ergo etc.

Sed contra: quia idem secundum idem non potest esse necessarium et contingens; sed actus intelligendi divinus est necessarius, cum sit idem quod Deus, esse autem intellectum vel volitum est contingens — contingenter enim Deus vult alia a se, et ita potest ea non velle nec intelligere, et ita tam esse intellectum quam volitum est contingens; ergo etc.

Preterea, que non sunt eadem inter se, nec alicui tertio; sed esse volitum et non-volitum et intellectum et non-intellectum non sunt idem inter se; ergo nec cum actu; sed quicquid extra se est volitum et intellectum potest esse non-intellectum et non-volitum; ergo etc.

Ad hoc etiam videtur esse intentio Augustini in libro *83 questionum*, questione 46, ubi dicit quod "alia ratione conditus est homo, alia equus" etc., ubi supra. Ex quo potest argui, quoniam illud quod multiplicatur non potest esse idem cum eo quod est immultiplicabile; sed actus divinus intelligendi vel volendi est unus vel unicus ipsis intellectis et volitis multiplicatis — secundum enim Augustinum ubi supra, singula singulis et suis propriis rationibus communicantur; ergo etc.

Posset dici quod habent aliquod esse distinctum ab actu. Ad cuius evidentiam posset dici quod sicut est duplex actio, videlicet transiens et immanens, ita utrisque correspondet aliquis terminus, cum nulla actio sit sine aliquo termino sibi correspondente — omnis enim actio est nata habere aliquem terminum; nunc autem intellectio et volitio in Deo sunt actiones sive operationes immanentes; ergo habent terminum aliquem accipientem esse per ipsas. Terminus autem istorum actuum non est res producta secundum esse quod habet in re extra, quoniam ut sic, secundum istud esse, res sunt terminus actionis transeuntis. Nec sunt ab eterno, sed ex tempore, cum tamen ab eterno fuerint intellecte et volite. Ergo sicut per actum creationis transeuntem obiectum capit esse simpliciter, ita etiam per actum immanentem, puta per intellectionem et volitionem, obiectum capit aliquod esse diminutum et secundum quid.

Confirmatur, quia divina intellectio maioris perfectionis est quam sit nostra; sed per intellectionem nostram, obiectum intellectum capit aliquod esse, puta esse intellectum, distinctum ab ipso actu intelligendi — quod patet, quia huiusmodi esse est obiective in intellectu, non autem isto modo est in ipso intellectio, sed tantum formaliter, nisi in intellectione reflexa; ergo etc.

Secundum hoc ergo, tenendo hoc posset dici quod esse cognitum vel volitum dicit duo. Dicit enim concretum denominativum intellectionis, et quantum ad hoc esse cognitum non differt ab actu cognitionis nisi secundum rationem, sicut nec album ab albedine vel quodcumque aliud concretum a suo abstracto. Secundo dicit substratum, quod quidem denominatur ab ipso cognitionis actu, quod quidem est esse diminutum factum vel formatum per ipsum actum cognitionis, qui fuit ab eterno, sicut et huiusmodi esse diminutum.

Ad argumentum in contrarium, ad illud Augustini, 5 *Super Genesim*, concedo quod obiectum quodcumque, antequam fieret, erat in intellectu divino vita intellectualis, et hoc participative, non formaliter et intrinsece. Obiectum enim cognitum ut obiective in actu participat aliquo modo vitam sui actus non formaliter, sed tantum obiective.

Vel potest dici aliter quod obiectum ab eterno habuit duplex esse in Deo, sive dupliciter fuit in eo. Fuit enim in ipso virtualiter, et quantum ad istud esse, videlicet virtuale, obiectum intellectum et actus sunt omnino idem, quia huiusmodi esse virtuale non est nisi ipse Deus a quo non distinguitur ipse actus. Nec quantum ad istud esse res sive rerum quidditates factibiles sunt in potestate Dei, sicut nec ipse Deus. Licet enim Deus posset ponere et non ponere divisim res in esse, et per consequens sit in eius potestate

producere vel non producere circa eas, non tamen posse producere. Licet enim contingenter producat, necessario tamen potest eas producere. Et isto modo, puta quantum ad huiusmodi esse virtuale, videtur loqui Augustinus, 6 *De trinitate*, ubi dicit quod filius est ars "Dei patris plena omnium rationum vivientium." Et subdit ibi quod omnes huiusmodi rationes sunt "in ea unum <sicut> unum de uno."

Alio autem modo obiectum habet esse in Deo, videlicet intentionaliter seu obiective. Et isto modo habet esse aliquod distinctum ab actu. Et hoc modo quantum ad istud esse loquitur Augustinus in libro *83 questionum*, questione 46, ubi supra, quando dicit quod "non eadem ratione conditus est homo qua equus," etc. Et ita Augustinus videtur innuere quod est duplex genus intentionum sive rationum. Sunt enim quedam virtuales, alie obiective, sive quod quidditates rerum habent duplex esse, videlicet virtuale et obiectivum. Secundum esse virtuale non sunt formate per actum aliquem, sed sunt idem quod actus, et ita ut sic dicit quod sunt incommutabiles et eterne et unum. Secundum autem esse obiectivum sunt formate per actum, et quantum ad tale esse dicit quod "non eadem ratione conditus est homo qua equus."

Tunc ergo secundum hoc potest dici quod quelibet creatura, antequam fieret, in illo vita erat formaliter, et hoc quantum ad esse virtuale. Erat etiam vita non formaliter, sed participative obiective, sive quantum ad esse obiectivum.

Ad illud quod dicitur, quod sunt incommutabiles etc., dico quod verum est quantum ad esse virtuale, non autem quantum ad esse obiectivum, puta quantum ad esse intellectum vel volitum. Immo quantum ad istud esse sunt contingentes, et in tali esse contingenter posite seu formate. Sic ad questionem.

[Ad argumentum principale]

Ad rationem in principio, quando dicitur quod non est ponendum in Deo aliquid quod sit imperfectionis, concedo. Sed tunc nego minorem. Infinitas enim idearum eo modo quo dictum est superius ipsam esse in Deo non dicit imperfectionem.

Ad probationem, dico quod ista multitudo non dicit confusionem sine ordine, immo est ibi ordo perfectionis, cum omnes ad rationem essentie reducantur.

Distinctiones 42-44

Circa distinctiones 42, 43, et 44 quero primo utrum ratio possibilitatis et ratio impossibilitatis creature sumatur ex parte Dei vel ex parte creature.

Videtur primo quod ex parte creature, quoniam a puro actu non potest accipi ratio potentie, nec etiam ratio opposita rationi potentie, cuius est ratio impossibilitatis; sed Deus est actus purus; ergo, etc.

Contra, quia ratio habentis causam magis debet sumi ex prima causa quam a secunda; omnis possibilitas et impossibilitas rei habent causam aliquam possibilitatis et impossibilitatis; ergo ratio cuiuslibet debet accipi ex parte Dei, qui est omnium prima causa.

Circa istam questionem primo videndum est quomodo Deus sit omnipotens; secundo, unde sumatur possibilitas et impossibilitas quarumdam rerum secundum quam dicimus quod Deus non potest peccare vel mentiri et huiusmodi.

[Articulus primus]

Quantum ad primum dico quod omnipotens est qui potest in omne possibile. Ad cuius evidentiam sciendum est quod possibile sumitur dupliciter. Uno enim modo accipitur ut distinguitur contra impossibile, et sic isto modo omne necessarium est possibile, quia non est impossibile -- immo necessarium est maxime possibile isto modo, quia ita est possibile quod nullo modo potest non esse. Unde isto modo possibile non opponitur necessario, sed opponitur contradictorie impossibili. Alio modo accipitur possibile ut distinguitur contra necessarium, secundum quod possibile est cui non repugnat esse et non esse, quia potest esse et non esse. Necessarium autem est cui repugnat non esse.

Tunc dicunt aliqui quod possibile quod est obiectum omnipotentie est possibile sumptum secundo modo, non autem primo modo, videlicet non ut dividitur contra impossibile, sed tantum ut dividitur contra necessarium, ut omnipotens dicatur qui potest in omne possible -- sit possibile quod potest esse et non esse. Et hoc est commune tribus personis, cum quelibet possit in quodcumque possibile isto modo.

Sed tamen dico aliter quod obiectum omnipotentie est quodcumque aliud ab omnipotente, sive sit possibile sive necessarium. Unde dico quod omnipotens est qui potest in omne aliud a se. Nihil enim potest in se ipsum. Et in hoc conveniunt omnes tres persone divine. Quelibet enim potest in quodcumque aliud a se, licet non in quodcumque alium a se masculine. Filius enim non est aliud a Patre, secundum Augustinum, sed tantum alius. Idem de Spiritu Sancto. Et ideo, quia in quodcumque aliud potest una personarum, potest alia, licet non in quemcumque alium. Idcirco omnes tres persone sunt omnipotentes et unum omnipotens.

Hoc de primo.

[Articulus secundus]

Quantum ad secundum, sunt tres modi dicendi. Primus modus dicendi est quod aliter est loquendum de causa possibilitatis et aliter de causa sive ratione impossibilitatis, quoniam possibilitas dicit aliquo modo perfectionem. Omne enim possibile ut sit, dicitur aliquo modo ens, et ita dicit perfectionem. Impossibilitas autem dicit imperfectionem, quoniam impossibile est non ens. Secus autem est dicendum de hiis que includunt perfectionem et secus de eo quod includit imperfectionem, quoniam omne includens vel dicens perfectionem aliquam habet primo ortum a Deo quam a creatura. Omnis enim perfectio et omne bonum primo est a Deo quam ab aliquo alio. Illud autem quod dicit imperfectionem et defectum non habet primo ortum a Deo, sed a creatura. Quia ergo possibilitas dicit bonum et perfectionem, ideo dicunt quod ratio possibilitatis est primo ex parte Dei, non autem ex parte creature. Sed quia per oppositum impossibilitas dicit imperfectionem, ideo etiam per contrarium ratio impossibilitatis sumitur primo ex parte creature, non ex parte Dei. Unde dicunt quod ideo hoc est possibile fieri quia Deus est

potens illud facere; non autem e converso Deus potest hoc facere quia ipsum est possibile fieri. Ita quod potentia vel potestas Dei activa est ratio possibilitatis creature passive. De impossibilitate autem est contrarium, quoniam ideo Deus non potest hoc facere quia hoc est impossibile fieri; non autem e converso quia hoc Deus non potest facere ideo hoc est impossibile fieri. Ita quod ratio impossibilitatis sumitur primo ex parte creature et ratio possibilitatis ex parte Dei.

Pro ista opinione arguitur, quia quicquid perfectionis est in creatura venit primo ex parte Dei, sic et quicquid est imperfectionis venit ex parte creature; sed esse possibile est perfectionis; esse autem impossibile est imperfectionis; ergo etc.

Contra, quia in causis precisis, si affirmatio est causa affirmationis, et negatio est causa negationis, secundum Philosophum in primo *Posteriorum*; sed secundum istam opinionem, quia Deus potest aliquid facere, ideo illud est possibile; ergo negatio illius affirmationis, que est quod Deus non potest hoc facere, erit causa quare illud sit impossibile fieri, que est consimiliter negatio illius affirmationis cuius illa affirmatio erat causa; ergo si prima ratio possibilitatis est ex parte Dei, erit etiam et prima ratio impossibilitatis.

Preterea, illud quod est posterius numquam potest esse causa prioris; sed si possibilitas ex parte Dei est causa possibilitatis ex parte creature, negatio possibilitatis ex parte Dei erit prior negatione possibilitatis ex parte creature. Probatio: Quia in negationibus disparatis, cuius sunt iste, negatio prioris affirmationis est prior negatione posterioris affirmationis; sed possibilitas ex parte Dei est prior possibilitate ex parte creature, cum sit per te causa eius; ergo et negatio eius erit prior negatione illius; ergo negatio possibilitatis sive impossibilitas creature non est causa impossibilitatis Dei, cum sit posterior ea, ut probatum est.

Alius modus dicendi est quod ratio utriusque, videlicet tam possibilitatis quam impossibilitatis, sumitur ex parte Dei. Ideo enim aliquid est possibile fieri quia Deus potest illud facere, et ideo impossibile quia non potest Deus illud facere; non autem e converso.

Et si queratur quare Deus hoc non potest facere, hic est status. Non enim dicendum est quod ideo non potest quia est impossibile, quia tunc sic dicendo esset | circulus in causis, cum hoc ponatur impossibile quia Ipse non potest illud facere, sed non potest quia hoc non sibi convenit, <quia> hoc posse non est posse.

Sed contra hoc arguo, quia nihil est impossibile nisi quod includit contradictionem, quia omne non includens contradictionem est possibile. Secundum quod aliqua includant contradictionem vel non includant, hoc non est ex parte Dei, sed ex parte terminorum. Deo enim per impossibile non existente, homo et non homo includerent contradictionem et repugnantiam, et homo et animal non includerent. Ergo ratio possibilitatis vel impossibilitatis venit non ex parte Dei, sed ex ratione intrinsica terminorum.

Tertius modus dicendi est quod quecumque possibilia vel impossibilia sunt possibilia vel impossibilia non ex parte Dei nec ex aliquo intrinsico, sed ex natura terminorum. Quod enim homo et album sint possibilia,

hoc est ex natura ipsorum, puta quia homo est homo et album est album. Quod etiam homo et non homo sint impossibilia, hoc est ex natura terminorum, quia videlicet homo est homo, etc.

Contra hoc arguo, quia si termini sunt causa possibilitatis et impossibilitatis opposite, quero in quo genere cause universalis istius impossibilitatis qua est impossibile quod homo sit lapis? Quero in quo genere cause isti termini ‘homo’ et ‘lapis’ sunt causa eius? Aut enim sunt causa eius in genere cause materialis aut formalis aut efficientis aut finalis. Non formalis, quia lapis et homo non contradicunt formaliter, sed sunt quedam disparata. Nec materialis, quia omnis causa materialis reducitur ad causam effectivam. Materia enim per se non habet effectum sine efficiente. Ergo si termini ponantur causa huius impossibilitatis in genere cause materialis, vel quicquid aliud ponatur causa isto modo, oportet quod reducatur ad aliquid quod sit causa in genere cause efficientis. Talis autem est Deus. Cuiuscumque enim effectus Deus est causa efficiens principalis.

Si autem dicas quod in genere cause efficientis, idem sequitur, puta quod Deus sit principalior et nobiliori modo causa huius impossibilitatis seu repugnantie. Quodcumque enim aliud a Deo ponatur causa effectiva alicuius, semper Deus est illius causa principalior in genere sive ordine cause efficientis.

Dico tamen quod quelibet predictarum opinionum est vera si bene intelligatur. Ad cuius evidentiam sciendum est quod potentia est triplex, sive tripliciter accipitur. Uno enim modo potentia accipitur pro eo quod est principium entis, et isto modo accipitur a Philosopho, secundo *De anima* et 8 *Metaphysice*. In secundo enim *De anima* dicit Philosophus quod ideo ex anima et corpore fit per se unum quia unum est in potentia et aliud in actu. Idem sententialiter dicit in 8 *Metaphysice*. Hoc etiam modo loquitur de potentia Commentator in *De substantia orbis*, ubi dicit quod materia substantiatur per posse. Et ad istum modum potentie reducitur modus potentie active et passive, de quo Philosophus loquitur 8 et 9 *Metaphysice*, dicens diffiniendo potentiam activam quod potentia activa est principium transumendi aliud in quantum aliud et passiva principium transumendi ab alio. Differt tamen ista potentia a prima, quoniam prima potentia est principium entis per modum partis. Totum enim est principium componentis*. Ista autem secunda, videlicet activa et passiva, est principium entis non per modum partis sed per modum agentis et pacientis. Activa enim est principium per modum agentis et passiva per modum pacientis.

Alia est potentia que nec est principium per modum partis nec per modum agentis vel pacientis, sed est differentia entis accidentalis vel essentialis, et hec est potentia que est actu dividens ens, secundum quod dicitur quod omne quod est vel est ens in actu vel in potentia. Et ista potentia non est eadem cum prima, quoniam prima stat simul cum suo actu contra quem dividitur. Nec sibi repugnant, cum constituat unum per se cum eo in quo, ut in quodam effectu communi utriusque est simul utuumque. Illa autem potentia que est differentia entis non potest esse simul cum actu condiviso sibi. Et de ista potentia loquitur Philosophus, 8, 9, et 12 *Metaphysice*, dicens quod illud idem quod est primo in potentia postea est in actu. Commentator etiam, super 12* *Metaphysice*, de ista potentia loquens, dicit quod transumptio non largitur mere multitudinem sed perfectionem. Ista etiam potentia differt in hoc a prima quoniam prima potentia non est semper in eodem genere cum actu suo, sicut subiectum quod est in potentia passiva respectu accidentis est alterius generis ab ipso. Ista autem secunda potentia que est differentia entis est semper eiusdem generis et etiam eiusdem speciei cum actu suo. Eadem enim albedo que prius fuit in potentia postea est in actu.

Tertia potentia est que nec est principium nec differentia entis, sed est tantum modus compositionis, ut illa potentia de qua loquitur Philosophus in primo *Periarmenias*, quando dicit quod possibile est hominem currere et possibile est hominem non currere. Huiusmodi enim potentia non est nisi modus quidam compositionis predicamenti cum subiecto. Et ista potentia est potentia logica.

Nunc autem inter huiusmodi dictas tres potentias est ordo, quoniam potentia que est differentia entis supponit potentiam que est principium entis, non e converso. Potentia enim que est principium entis non presupponit potentiam que est differentia entis. Sed potentia tertia, que videlicet est tantum modus compositionis, neutram potentiam presupponit, videlicet nec illam que est principium nec etiam secundam que est differentia entis, sed est potentia que sequitur quoscumque terminos ex natura terminorum. Et ideo generatio naturalis et creatio distinguntur per hoc quoniam generatio presupponit duplicem potentiam, quia tam potentiam que est principium entis quam etiam illam que est differentia. Presupponit enim generale esse in potentia, et ita potentiam que est differentia entis et etiam potentiam subiectivam termini generationis. Et ista est principium entis. Creatio autem neutram presupponit. Nec enim presupponit illam que est principium entis nec illam que est differentia entis, sed tantum presupponit potentiam tertiam, videlicet illam que est modus compositionis, sicut creatio celi presupponit ipsum celum esse creabile. Celum autem esse creabile non est ipsum celum habere potentiam que est principium seu differentia entis, sed tantum celum esse creabile est ipsum creari fore possibile, que quidem possibilitas non est aliud quam modus sive modificatio propositionis.

Ex istis patet falsitas opinionis eorum qui ponunt quod si materia poneretur sine forma, quod tunc non haberet aliquem actum, quoniam, licet non haberet aliquem actum oppositum potentie que est principium entis, haberet tamen actum oppositum potentie que opponitur impossibili sive impossibilitati. Impossibile enim privat omnem actum, quo tamen non esset privativa universaliter materia posita sine omni forma. Solum enim esset privativa actu opposito potentie que est principium entis et etiam potentie que est differentia. Non autem actu opposito impossibilitati.

Nunc ad propositum dico quod si loquamur de illa potentia que est principium entis vel etiam de illa que est differentia entis, dico quod utraque istarum est simplex et incomplexa, et utraque venit ex potentia Dei activa -- non autem ex parte creature, quia utraque istarum sequitur creationem. Loquendo autem de potentia tertia, que requiritur et presupponitur in creatione, qua quidem creabile non fuit impossibile sed possibile creari, dico quod ista venit tam ex parte terminorum quam etiam aliquo modo ex causa extrinseca, videlicet ex parte Dei. Ubi advertendum quod dupliciter potest intelligi aliquid esse causam complexionis terminorum. Aliquid enim est causa complexionis aliquorum terminorum per modum complexi, sicut complexio principiorum sive premissarum est causa sive ratio complexionis conclusionum. Aliquid etiam secundo est causa complexionis non isto modo, sed simpliciter et modo incomplexo, sicut intellectus dicitur esse causa veritatis et complexionis conclusionis quam infert effective ex premissis et ipsarum etiam premissarum.

Tunc ex hoc ad propositum dico quod istius possibilitatis, qua quidem dicitur celum possibile esse, est causa, licet diversimode, tam Deus quam etiam termini huius complexionis, puta celum ibi esse. Termini enim sunt causa per modum complexionis istius possibilitatis sive non repugnantie. Deus autem, qui est

ens incomplexum, est causa tam non repugnantie sive possibilitatis quam etiam ipsius complexionis per modum incomplexi. Sunt tamen alique complexiones in Deo que sunt cause aliarum per modum complexi, ut quia Deus agit ideo passum patitur, sicut quia non agit ideo non patitur paciens, ita quod complexio affirmativa est causa veritatis complexionis affirmative et negativa negative. Sic ergo dico quod Deus est causa per modum incomplexi terminorum et complexionis eorum et etiam cuiuscumque possibilitatis et impossibilitatis ipsorum.

Et ita dico quod possibilitas et impossibilitas accipiuntur tam ex parte terminorum quam etiam ex parte Dei diversimode, ut dictum est.

Ex quo patet quod vera est illa opinio que dicebat quod possibilitas et impossibitas rei sumuntur ex rationibus terminorum.

Vera etiam est illa alia que dicebat quod ex parte Dei.

Per hoc etiam patet ad omnes rationes factas in oppositum.

Distinctiones 45-48

[Quaestio prima]

Circa distinctiones 45, 46, 47, et 48, quero primo utrum voluntas Dei semper impleatur.

Videtur quod sic, quia "Omnia quecumque voluit, Dominus fecit," etc.

Contrarium habetur ubi dicitur, "Quotiens volui congregare filios tuos, quemadmodum gallina congregat pullos suos et noluisti."

Ad istam question respondet Augustinus, et est in *Enchiridion*, capitulo 61, et Anselmus in libro *Cur Deus homo*, dicentes ambo quod semper voluntas Dei impletur. Licet enim aliquando videatur impediri uno modo, impletur tamen alio modo.

Ad cuius evidentiam distinguit Anselmus triplicem actum divine voluntatis, quod est intelligendum non ex parte actus, cum sit unicus in se, sed tantum ex parte obiecti. Sicut enim in nobis actus voluntatis est triplex -- est enim quidam quo voluntas aliquid imperat vel consulit, et iste est actus imperandi aliquid affirmative vel negative vel annuendi; alius est actus renumerandi sive reddendi bonum pro obediente et consilio assentiente, vel malum pro non obediente nec consilio assentiente; tertius actus est actus permittendi fieri bonum vel malum cum posset ne fieret prohiberi -- ita consimiliter ex parte divine voluntatis correspondet predicto triplici actui: Unus actus qui quidem, licet in se sit indistinctus et unicus, distinguitur tamen ex parte obiecti isto triplici modo dicto, quoniam aliquid Deus precipit affirmative ut fiat, sicut illud: "Honora patrem tuum," etc., vel negative ut non fiat, sicut illud: "Non furaberis," etc. Omnia enim precepta sunt affirmativa vel negativa; aliquid etiam consulit, ut illud: "Si vis perfectus esse,

vade et vende," etc. Secundo habet alium actum non per modum precepti nec per modum consilii, sed per modum reddendi, sive actum qui est reddere bonum bonis et malum malis. Et iste actus est actus iustitie.

Tertius eius actus est actus permittendi, quo quidem permittit malum fieri, quod tamen posset si vellet ne fieret impedire. Unde secundum Augustinum hec est ratio quia plures sunt mali quam boni, quia quanto plures sunt mali tanto magis Dei misericordia relucet in bonis qui preservantur a malis. Plusquam enim primus homo, qui pro se et aliis hominibus iustitiam acceperat, peccavit, omnes sumus facti una massa peccati. Propter quod, ut ait Augustinus, non nullus liberaretur nec a malo preservaretur nec posset conqueri. Quod autem nunc aliqui et pauci liberentur, hoc est ex maxima et summa Dei misericordia.

Tunc ad propositum, ad videndum quomodo voluntas Dei semper impleratur, Anselmus ponit quoddam exemplum de terra. Dicit enim quod consimiliter se habet creatura rationalis respectu Dei sive voluntatis eius sicut se habet terra respectu celi, quoniam sicut terra, que est in medio celi, continetur undique a qualibet parte celi, ita quod quanto magis recedit ab una parte celi tanto magis appropinquat ad aliam, nec potest ab una recedere quin ad aliam appropinquet, ita consimiliter creatura rationalis concluditur undique voluntate et omnipotentia Dei. Et ideo si vis fugere voluntatem Dei iubentem, statim incideris in voluntatem eius punientem, transeundo per voluntatem eius permittentem, et ita incideris in peius.

Ex quo concludit Anselmus, sicut et potest concludi, quod voluntas Dei semper impleretur, puta vel voluntas iubens vel puniens vel permittens. Sicut enim in contradictione semper altera pars est vera, non utraque, ita de divina voluntate semper vel impleretur ista vel illa voluntas, videlicet iubens vel puniens, licet non omnes simul. Unde ista sunt opera eius magna et exquisita in omnes voluntates eius ut nullus possit transire nec subterfugere eius voluntatem. Et ideo dicebat Psalmus, "Si ascendeo in celum tu illic es; si descendeo in infernum ades. Si sumpsero pinnas diluculo," etc.

Tunc ergo sumendo dicta sanctorum ad istam questionem et reducendo ea ad formam, quando queritur utrum voluntas Dei semper impleratur, dico quod voluntarium est triplex, ut in precedentibus dictum fuit: Est enim quoddam voluntarium permixtum cum involuntario, aliud permixtum cum non voluntario, tertium simpliciter impermixtum.

Secundum hoc dico quod, licet actus divine voluntatis sit secundum se indistinctus et unicus, tamen ex parte obiecti distinguitur, quoniam ad aliquid se habet per modum voluntarii permixti cum involuntario, ad aliquid per modum voluntarii permixti cum non voluntario, et tertio ad aliquid per modum voluntarii simpliciter et totaliter impermixti.

Ex hoc respondeo ad questionem. Et dico quod voluntas sive velle Dei permixtum cum involuntario non semper impletur, nec etiam velle permixtum cum non voluntario, cuiusmodi est illud de quo dicitur, "Quotiens volui congregare filios tuos et noluisti." Voluntas autem Dei sive velle simpliciter et totaliter impermixtum semper impletur, et de ista loquitur Psalmus, "Omnia quecumque voluit, Dominus fecit," et Apostolus dicens, "Voluntati eius quis resistet?"

Ad argumenta patet.

[Quaestio secunda]

Secundo quero iuxta hoc utrum teneamur conformare voluntatem nostram divine voluntati in quocumque volito vel nolito ab ipsa.

Hic ad istam questionem respondetur communiter et bene, si bene intelligatur, distinguendo de voluntate Dei. Dicitur enim quod duplex est voluntas Dei, intelligendo semper ex parte obiecti: Quedam enim est voluntas signi, alia vero est voluntas beneplaciti. Voluntas quidem signi est illa que explicatur per signum, puta per consilium aliquod vel preceptum, sive expresse et explicite sive implicite. Voluntas autem beneplaciti est voluntas secreta, qua quidem Ipse vult aliquid, quod tamen nobis non innotescit exterius per aliquod signum, videlicet consilium vel preceptum.

Tunc dicitur quod nos tenemur conformare nos sive voluntatem nostram divine voluntati, loquendo de voluntate signi. Debemus enim facere et tenemur ad illud quod precipit, esto etiam quod Ipse illud non velit voluntate beneplaciti quod bene interdum contingit, videlicet quod Deus aliquid precepit quod tamen fieri noluit; ut patet de Abraam cui precepit ut immolaret filium suum, quod tamen non volebat voluntate beneplaciti, sicut patuit postea per effectum. Et tamen hoc non obstante, Abraam tenebatur obedire ex quo sibi precipiebatur.

Sed contra istam distinctionem de voluntate signi et beneplaciti arguo. Videtur enim quod contradictionem includat nisi sane intelligatur. Et arguo sic: Omne signum non concordans signato est falsum; sed si est aliquid volitum voluntate beneplaciti quod non sit volitum voluntate signi, illud signum non concordat signato, immo discordat ab eo; ergo huiusmodi voluntas signi est falsa.

Et ideo dico aliter. Ad cuius evidentiam sciendum est quod in Deo, ut sepe dictum est, est triplex voluntarium: Quoddam simpliciter et totaliter impermixtum, ut illud quo Deus vult salvare bonos sive electos. Aliud est permixtum cum non-voluntario, hoc est, cum negatione alicuius gradus voluntarii. Et isto voluntario sic permixto Deus vult generaliter et indifferenter omnes salvare. Omnibus enim vult beatitudinem, non tamen perfecte seu intense, cum non removeat omne impedimentum ad hoc in omnibus, quod tamen posset facere si vellet. Et ideo ratione impedimenti quod non removet, cum tamen removere posset, velle divinum quo eis vult beatitudinem non est efficax seu intensum, immo remissum, et ideo cum non-voluntario mixtum.

Tertium est voluntarium permixtum cum involuntario, cuiusmodi est illud quo Deus vult dampnare malos seu reprobos. Et hoc involuntarium superat voluntarium propter illorum malitiam, sicut cum aliquis propter periculum tempestatis proicit merces in mari. Talis quidem vult proicere et non vult. Unde habet simul actum volendi et nolendi respectu diversorum. Vult enim proicere propter periculum iminens tempestatis; non vult autem propter preciositatem mercium. Et tamen hoc involuntarium cui voluntarium est admixtum superat voluntarium.

Consimiliter dico de Deo quantum ad velle dampnare malos. Deus enim ab eterno previdit eos sicut et alios ad sui similitudinem et imaginem esse futuros. Et ideo ordinavit eos ad beatitudinem ut ad proprium

eorum finem, et voluit eis ipsam. Tamen quia previdit alios perseverare debere in malitia, ideo ratione huius, eius velle fuit permixtum cum involuntario, quia noluit eos salvare, sed dampnare, quod quidem <est> involuntarium, superavit voluntarium ratione malitie finalis ipsorum.

Tunc ad propositum, respondeo. Ad questionem dico quod nos tenemur conformare voluntatem nostram voluntati divine quantum ad illud quod nobis est notum Deum velle, non autem quantum ad aliud. Et hinc est quod, cum velle sive voluntarium totaliter impermixtum, quo Deus vult electos salvare, sit nobis ignotum -- ignotum enim est nobis qui sunt illi quos tali actu impermixto vult ad beatitudinem pervenire -- cum etiam non solum istud voluntarium totaliter impermixtum, sed etiam illud quod est permixtum cum involuntario, quo quidem reprobos vult dampnare, sit consimiliter nobis ignotum, per consequens dico quod non tenemur conformare sive conformiter velle voluntati divine nec quantum ad velle sive voluntarium simpliciter impermixtum nec etiam quantum ad permixtum cum involuntario. Nullus enim tenetur ad impossibile; velle autem conformiter aliqui quantum ad illud quod mihi non est notum ipsum velle est impossibile. Nunc autem in vita ista non est nobis certum quid est a Deo volitum isto duplici voluntario.

Exemplum: Deus voluit filium suum mori, licet non per Iudeos. Esto etiam quod Deus voluisset Iudeos Christum occidere, ipsi tamen non tenebantur eum occidere, immo peccassent occidendo eum cum eos hoc Deum velle lateret.

Et ideo per oppositum, quia velle sive voluntarium quo Deus vult medio modo illud, videlicet quod nec est totaliter impermixtum nec etiam permixtum cum involuntario, sed tantum cum non-voluntario, quo quidem vult omnibus beatitudinem, est nobis ex scriptura certum, ideo quantum ad hoc tenemur conformare voluntatem nostram voluntati divine et conformiter velle ei nulli malum volendo, sed cuilibet eternam beatitudinem ut communem finem omnium appetendo, ad quem Ipse nos producat, qui est benedictus in secula secolorum. Amen. Amen. Amen.

Explicit lectura fratris Francisci de Marchia super primum, secundum reportationem factam sub eo tempore quo legit Sententias Parisius, anno Domini MCCCXX.

Return to Francis of Marchia

Dante Alighieri

Dante's engagement with philosophy cannot be studied apart from his vocation as a writer, in which he sought to raise the level of public discourse by educating his countrymen and inspiring them to pursue happiness in the contemplative life. He was one of the most learned Italian laymen of his day, intimately familiar with Aristotelian logic and natural philosophy, theology (he had a special affinity for the thought of Albert the Great and Thomas Aquinas), and classical literature. His writings reflect this in its mingling of philosophical and theological language, invoking Aristotle and the neo-Platonists side by side with the poet of the psalms. Like Aquinas, Dante wished to summon his audience to the practice of philosophical wisdom, though by means of truths embedded in his own poetry, rather than mysteriously embodied in scripture.

1. Life

Dante was born in 1265 in Florence. At the age of 9 he met for the first time the eight-year-old Beatrice Portinari, who became in effect his Muse, and remained, after her death in 1290, the central inspiration for his major poems. Between 1285, when he married and began a family, and 1302, when he was exiled from Florence, he was active in the cultural and civic life of Florence, served as a soldier and held several political offices.

Since the early thirteenth century two great factions, the Guelfs and the Ghibellines, had competed for control of Florence. The Guelfs, with whom Dante was allied, were identified with Florentine political

autonomy, and with the interests of the Papacy in its long struggle against the centralizing ambitions of the Hohenstaufen emperors, who were supported by the Ghibellines. After Charles of Anjou, with the blessing of the Papacy and strong Guelf support, defeated Hohenstaufen armies at Benevento (1265/6) and Tagliacozzo (1268), the Guelfs became the dominant force in Florence. By the end of the century, the Guelfs were themselves riven by faction, grounded largely in family and economic interests, but determined also by differing degrees of loyalty to the papacy and to Guelf allegiances.

In 1301, when conflict arose between the "Blacks," the faction most strongly committed to Guelf and papal interests, and the more moderate Whites, Pope Boniface VIII instigated a partisan settlement which allowed the Blacks to exile the White leadership, of whom Dante was one. He never returned to Florence, and played no further role in public life, though he remained passionately interested in Italian politics, and became virtually the prophet of world empire in the years leading up to the coronation of Henry VII of Luxemburg as head of the Holy Roman Empire (1312). The development of Dante's almost messianic sense of the imperial role is hard to trace, but it was doubtless affected by his bitterness over what he saw as the autocratic and treacherous conduct of Pope Boniface, and a growing conviction that only a strong central authority could bring order to Italy.

During the next twenty years Dante lived in several Italian cities, spending at least two long periods at the court of Can Grande della Scala, lord of Verona. In 1319 he moved from Verona to Ravenna, where he completed the *Paradiso*, and where he died in 1321.

Dante's engagement with philosophy cannot be studied apart from his vocation as a writer -- as a poet whose theme, from first to last is the significance of his love for Beatrice, but also as an intellectual strongly committed to raising the level of public discourse. After his banishment he addressed himself to Italians generally, and devoted much of his long exile to transmitting the riches of ancient thought and learning, as these informed contemporary scholastic culture, to an increasingly sophisticated lay readership in their own vernacular.

This project was Dante's contribution to a long-standing Italian cultural tradition. His reading in philosophy began, he tells us, with Cicero and Boethius, whose writings are in large part the record of their dedication to the task of establishing a Latinate intellectual culture in Italy. The *Convivio* and the *De vulgari eloquentia* preserve also the somewhat idealized memory of the Neapolitan court of Frederick II of Sicily (1195-1250) and his son Manfred (1232-66), intellectuals in their own right as well as patrons of poets and philosophers, whom Dante viewed as having revived the ancient tradition of the statesman-philosopher [Van Cleve, 299-332; Morpurgo]. Dante himself probably studied under Brunetto Latini (1220-94), whose encyclopedic *Livres dou Tresor* (1262-66), written while Brunetto was a political exile in France, provided vernacular readers with a compendium of the Liberal Arts and a digest of Aristotelian ethical and political thought [Meier; Imbach, 37-47; Davis (1984), 166-97].

But the fullest medieval embodiment of Dante's ideal is his own writings. In them we see for the first time a powerful thinker, solidly grounded in Aristotle, patristic theology, and thirteenth-century scholastic debate, bringing these resources directly to bear on educating his countrymen and inspiring them to pursue the happiness that rewards the philosopher.

2. Early Poetry

Though he evidently did not begin serious study of philosophy until his mid-twenties, Dante had already been intellectually challenged by the work of a remarkable group of poets, practitioners of what he would later recall as the *dolce stil novo*, in whose hands a lyric poetry modelled on the *canso* of the Provençal troubadours became a vehicle for serious enquiry into the nature of love and human psychology. A generation earlier Guido Guinizzelli (1230-1276) had puzzled contemporaries with poems treating love in terms of the technicalities of medicine and the cosmology of the schools, while celebrating in quasi-mystical language his lady's power to elevate the spirit of her poet-lover:

> *Splende in la intelligenzïa del cielo*
> *Deo crïator, più che 'n nostri occhi 'l sole;*
> *ella intende suo fattor oltra 'l cielo,*
> *e 'l ciel volgiando, a lui obedir tole;*
> . . .
> *così dar dovria, al vero,*
> *la bella donna, poi che 'n gli occhi splende*
> *del suo gentil, talento,*
> *chi mai da lei obedir non si disprende.*
>
> [*Al cor gentil rempaira sempre amore*, 41-44, 47-50]
>
> Translation:
> God the creator shines in the intelligence of heaven more than the sun in our eyes, and this [intelligence] understands her maker beyond the universe. Making the heavens turn, she submits to obey Him . . . So truly should the beautiful lady, when she shines on the eyes of her gentle [lover], impart the desire that his obedience to her never fail.

The Lady, exerting on her lover a power derived from the participation of her understanding in the divine, plays the role of the celestial *intelligenze*, who transmit the influence of the First Mover to the universe at large. The poet is thus caught up in a circular process through which his understanding, like theirs, is drawn toward the divine as manifested in the lady's divinely inspired radiance. For Guinizelli this exploitation of the idea of celestial hierarchy is perhaps only a daring poetic conceit. For Dante it will become a means to the articulation of his deepest intuitions.

Guido Cavalcanti, Dante's older contemporary and the single strongest influence on his early poetry, was renowned not only as a poet, but for his knowledge of natural philosophy. His great *canzone*, "*Donna mi prega*," which became the subject of learned Latin commentaries, deals with ideas commonly associated with the "radical Aristotelianism" or "Averroism" of his day. The purpose of this astonishing poem is to describe in precise philosophical terms ("*naturale dimostramento*") the experience of love.

For Guido there is an absolute cleavage between the sensory and intellectual aspects of the response to a loved object. Once the phantasma of the object becomes an abstracted form in the possible intellect, it is wholly insulated from the *diletto* of the *anima sensitiva* (21-28). This has seemed to modern commentators to imply an Averroist view of the intellect as a separate, universal entity [Corti (1983), 3-37], and the lines which follow (30-56), where the *vertú* of the sensitive soul displaces reason and "assumes its function," presenting to the will an object whose desirability threatens a fatal disorientation, sustain this impression. Love is still the aristocratic vocation of the troubadours, and Guido acknowledges that noble spirits are aroused by it to prove their merit. But they work in darkness, for the force that moves them obscures the light of intellectual contemplation (57-68). The *canzone* is so exclusively an exercise in "natural philosophy," so centered on biological necessity, that consciousness itself is wholly excluded from consideration. The ethical dimension of love consists in the challenge its blind urgency presents to reason. "Nobility" is a matter of self-control, and the precarious happiness that such love affords has no ideal dimension.

Guido's influence on Dante was profound. But the *Vita nuova*, an anthology of Dante's early poetry interspersed with a narrative combining commentary on his poetic development with the history of his devotion to Beatrice during her earthly life, reveals a growing realization that his own conception of poetry and love differ fundamentally from Guido's. Like Guido Dante accepted love as being, for better or worse, fundamental to the noble life, and his early lyrics express a sense like Guido's of the internally divisive power of desire. But as the *Vita nuova* unfolds there is a gradual shift of focus: having failed to win his lady's favor by dramatizing his own sufferings, Dante resolves to devote his poetry henceforth wholly to praise of her [VN, c. 18.4-6]. The result of this new resolve is a *canzone*, "*Donne ch'avete intelletto d'amore*" ("Ladies who have intelligence of love"), which returns to the source of his inspiration and Guido's in the poetry of Guinizelli, and makes a wholly new departure. For Guido, the "heavenly" allure of the lady is a deception perpetrated by the senses, all the more dangerous as the lover's *gentilezza* responds more fully to the attraction of her beauty and subjects itself to the "fierce accident" of passion. Dante, too, sees that the experience his early, tormented lyrics depict is "an accident occurring in a substance" [VN 25.1-2], but the "fiery spirits of love" which strike the eyes of those on whom his lady bestows her greeting are not just goads to desire:

> *E quando trova alcun che degno sia*
> *di veder lei. quei prova sua vertute,*
> *ché li avvien, ciò che li dona, in salute,*
> *e sì l'umilia ch'ogni offesa oblia.*
> *Ancor l'ha Dio per maggior grazia dato*
> *che non pò mal finir chi l'ha parlato.*
>
> [*Donne ch'avete intelletto d'amore*, 37-42, VN 19.10]
>
> Translation:
> And when she finds one who is worthy to behold her, he feels her power, for what she bestows on him is restorative, and humbles him, so that he forgets any injury. Moreover God has made the power of her grace even greater, for no one who has spoken with her can

come to a bad end.

Pursuit of the lady's favor has become a test, not just of nobility, but of virtue. Her beauty is perfect, the fullest possible exampling of nature's power to reveal God's creative love. The climax of the *Vita nuova* occurs when Dante encounters Guido's lady, Giovanna, followed by his own Beatrice, "one marvel," as he says, "following the other" [VN 24.8]. At once he realizes that Giovanna's beauty, like the prophecy of the biblical Giovanni, is a precursor, heralding the "true light" of Beatrice, just as Guido's poetry of earthly love is finally a foil to his own celebration of the transcendent love revealed to him in Beatrice.

3. Philosophical Training

The philosophical content of the *Vita nuova* is minimal, a skeletal version of contemporary faculty psychology and a few brief references to metaphysics. But while finding his orientation as a poet Dante was also engaged in the study of philosophy, and spent "some thirty months" frequenting "the schools of the religious orders and the disputations of the philosophers" [*Conv*. 2.12.7]. This period must have included study in the Dominican school at Santa Maria Novella, where Dante could have learned logic and natural philosophy, and heard Fra Remigio de' Girolami (d. 1319) expound a theology based on Thomas and Aristotle [Panella; Davis (1984), 198-223]. Remigio, like Dante, read widely in classical literature of all sorts, and he was fond of drawing lessons in political and ethical conduct from his reading. For both Remigio and Dante, moreover, Thomas was primarily the author of the *Summa contra Gentiles* and the commentary on the *Ethics*, concerned, like Aristotle himself, to demonstrate the capacities of human reason as a means to truth.

Dante cites a dozen works of Aristotle, apparently at first hand, and shows a particularly intimate knowledge of the *Ethics*, largely derived, no doubt, from Thomas [Minio-Paluello]. But his Aristotelianism was nourished by other sources as well. Bruno Nardi has argued persuasively that his attitude toward the study of philosophy also owes a great deal to the more eclectic Albert the Great [Nardi (1967), 63-72; (1992), 28-29]. In Albert he encountered a wide-ranging encyclopedism which included original work, experimental and theoretical, in natural science, and treated Aristotelian natural philosophy and psychology in the light of a neo-Platonism derived from Arabic philosophers and such Greco-Arab sources as the *Liber de Causis*, as well as the Christian neo-Platonist tradition of the Pseudo-Dionysius. Albert aimed to discover Aristotle's own meaning, with the help of Greek and Arab commentators who led him into disagreement with other *Latini*, including at certain points his pupil Thomas, and he asserts more than once that philosophy and theology are separate spheres of knowledge. It was doubtless this willingness to pursue philosophy on its own terms that appealled to Dante, who also sought to distinguish philosophical and religious knowledge without simply subordinating the former to the latter.

Albert's view of the procession of the universe from the "substantial light" of the divine intellect through the operation of a hierarchy of lesser intelligences is clearly perceptible in Dante's treatment of the cosmic *intelligenze* or *sostanze separate* in the *Convivio* [*Conv*. 2.4-5; Nardi (1992), 47-62]. It shows up again in his treatment of the growth of the human embryo, which seems to imply, not a sequence of animations by nutritive, sensitive and intellective powers, as for Thomas, but the continuous operation of a single *virtus*

formativa, whose operation Albert compares to that of the *prima intelligentia* in the soul [*De intellectu & intelligibili* 2.2], and which is responsible not only for the development of the human creature but for effecting its union with an essentially external *anima intellectiva* [Boyde 270-79; Nardi (1960), 9-68; (1967), 67-70].

Albert is thus a likely conduit for seemingly Averroist elements in Dante's thought. He regards intellectual activity as the operation of the *intellectus agens*, through which the human soul is illumined by the divine Intelligence. Each soul possesses its own intellect, but this intellect is a "reflection" (*resultatio*) of the light of the primal mind, which thus, in effect, becomes itself the true agent intellect. Albert explicitly rejects the Averroist view of the active intellect as itself a celestial intelligence, a single, separate substance which actualizes in the passive intellect phantasms supplied by individual human minds. But he argues that only an intellect universal in nature can produce an understanding of universal forms. The intellect and the soul of which it is a function thus partake of the character of the separate intelligences. Soul is not the actualizing essence of the human creature, as in Thomas, but is related to body through the mediation of its organic faculties. In itself, through its agent intellect, the soul is drawn to contemplate the intelligences which order the universe at large, is informed by them with the transcendent knowledge they manifest, and finally "stands" in the divine intellect. In this way certain men are enabled to fulfil the innate human desire for understanding and attain a natural beatitude, "substantiated and formed in the divine being" [Albert, *De intellectu & intelligibili* 2.2-12; Nardi (1960), 145-50].

That this fulfillment is attained through natural understanding, with no recourse to the theology of grace and revelation, marks a crucial difference between Albert and Thomas, who devotes several chapters of the *Summa contra gentiles* to a forceful refutation of the notion that final happiness as defined by Aristotle is possible in this life [SCG 3.37-48]. For Thomas the desire to know is one and the same at all levels, and philosophy, seeking the causes of things, is ultimately "ordered entirely to the knowing of God" [SCG 3.25.9] Dante's own position on this question is difficult to define precisely. The poet of the *Paradiso* is at one with Thomas on the value of philosophy as consisting finally in its power to prepare the mind for faith [*Par*. 4.118-32; 29.13-45], but he shares Albert's fascination with natural understanding, and in earlier writings his willingness to grant philosophy a "beatitude" of its own hints at a latent dualism in his thought [Foster (1965), 51-71; (1977), 193-208].

Dante was surely aware also of a "radical" Aristotelianism centered in Bologna, where masters influenced by Siger of Brabant and Boethius of Dacia were affirming the autonomy of human reason and its capacity to attain happiness though its own powers [Corti (1981), 9-31; Vanni Rovighi]. But these thinkers, too, were following paths first taken by Albert, and his influence, together with that of Thomas, is sufficient to account for the distinctive features of Dante's use of philosophy. Whatever the precise channels, Dante was unquestionably one of the most learned Italian laymen of his day, aware of the issues contested in the schools, and at home with the modes of discourse in which they were discussed.

But there is also an old-fashioned strain in Dante's thinking, an idealistic, Platonizing view of the mental universe which recalls not just the neo-Platonized Aristotle of the *Liber de causis*, but the more primitive encyclopedism of twelfth- century thinkers like Bernardus Silvestris and Alan of Lille, poet-philosophers whose world view, inherited from late-antique neo-Platonism, was defined by the Liberal Arts and the

cosmology of Plato's *Timaeus* [Vasoli, 83-102; Garin, 64-70]. In Bernardus' *Cosmographia* and Alan's *Anticlaudianus*, the unfolding of the secrets of nature by the enquiring mind generates an allegory of intellectual pilgrimage toward truth. Dante's experience of philosophy, though defined in more dynamic and sophisticated terms, is a version of the same journey. The experience of love becomes a means to self-realization, and an awareness of the hierarchy of forces operative in the universe at large, which makes possible an *ascensus mentis ad sapientiam*, to that "amoroso uso della sapienza" which enables the human mind to participate in the divine.

4. The *Convivio*

The record of Dante's thirty months of study, and the fullest expression of his philosophical thought, is the *Convivio*, in which commentary on a series of his own *canzoni* is the occasion for the expression of a range of ideas on ethics, politics, and metaphysics, as well as for extended discussion of philosophy itself. Dante describes the genesis of his love of philosophy, and reflects on the ability of philosophical understanding to mediate religious truth, tracing the desire for knowledge from its origin as an inherent trait of human nature to the point at which the love of wisdom expresses itself directly as love of God.

Philosophy itself is the "love of Wisdom," and Dante's central metaphor for representing it is the poetic celebration of a noble lady, a *donna gentile*, an act which, like Guinizelli, he sees as involving the influence of cosmic powers. His poetry, "materiated" out of love and virtue [*Conv*. 1.1.14] comes into being because his nature is responsive to the influence of the "movers" of the universe, the intelligences, whose loving understanding determines "the most noble form of heaven" as they in turn respond to "the love of the Holy Spirit" [2.5.13, 18]. Their cosmic activity is a continual translation of understanding into love and natural process, and it is this which causes Dante to sing [2, Canzone, 1-9]:

> *Voi che 'ntendendo il terzo ciel movete,*
> *udite il ragionar ch'è nel mio core,*
> *ch'io nol so dire altrui, sì mi par novo.*
> *El ciel che segue lo vostro valore,*
> *gentili creature che voi sete,*
> *mi tragge ne lo stato ov'io mi trovo.*
> *Onde 'l parlar de la vita ch'io provo,*
> *par che si drizzi degnamente a vui:*
> *però vi priego che lo mi 'ntendiate.*
>
> Translation:
> You who by understanding move the third heaven, hear the discourse which is in my heart, and which seems so strange to me that I know not how to say it to others. The heaven which responds to your power, noble creatures that you are, draws me into the state in which I find myself, and so it seems that speech about the life I am experiencing is most appropriately addresses to you. Therefore I pray that you will understand me.

The intellective power or *intendimento* of the intelligences moves Dante to an utterance which only these same powers can fully understand. Thus there is a continuum, a process of *circulazione* which begins in the mind of God and descends through the work of the *intelligenze* to draw Dante's nature into that praise of the *donna gentile* which constitutes the fulfillment of his own nature, the highest expression of which his desire and intellect are capable [2.5.15, 18; 2.6.5].

Of the four books or *trattati* of the *Convivio* the first is largely a defense of Dante's decision to write his prose commentaries, as well as the poems they expound, in the Tuscan vernacular rather than in Latin. The second book provides a delineation of the Ptolemaic universe which the *intelligenze* govern, capped by a description of the Empyrean Heaven [2.3.8-11]:

> . . . outside all of these [spheres, heavens] the Catholics place the Empyrean heaven, which is to say, "the heaven of flame," or "luminous heaven"; and they hold it to be motionless because it has in itself, with respect to each of its parts, that which its matter desires. This is why the Primum Mobile has the swiftest movement; for because of the most fervent desire that each part of the ninth heaven has to be conjoined with every part of that divinest, tranquil heaven, to which it is contiguous, it revolves beneath it with such desire that its velocity is almost incomprehensible. Stillness and peace are the qualities of the place of that Supreme Deity which alone completely beholds itself. This is the place of the blessed spirits, according to the will of the Holy Church, which cannot lie. Aristotle, to anyone who rightly understands him, seems to hold the same opinion in the first book of *Heaven and the World* [i.e. *De caelo*]. This is the supreme edifice of the universe in which all the world is enclosed and beyond which there is nothing; it is not itself in space but was formed solely in the Primal Mind, which the Greeks call Protonoe. This is that magnificence of which the Psalmist spoke when he says to God: "Your magnificence is exaltled above the heavens."

The role of the Empyrean in thirteenth-century thought is equivocal. Some thinkers attempt to explain it scientifically, as a comprehensive cosmic principle, while for Thomas and Albert any such realm must be spiritual in nature, and can bear no natural relation to the astronomical universe, though both at times seem to grant it a certain influence on the natural order [Nardi (1967), 196-214; Vasoli, 94-102]. Dante's account reflects these uncertainties. He begins by citing "the Catholics," or orthodox belief, as authority for his account of this "abode of the supreme deity," but then goes on to treat the Empyrean as a created thing, "formed in the Primal Mind," and as the motionless cause of motion in the physical universe. If God dwells in this place, the Empyrean resides equally in Him, and the universe at large is encompassed, causally and locally, by the Empyrean. Dante deploys the Aristotelian physics of desire to explain the relationship of the Empyrean to the lesser heavens, yet it is at the same time beyond space, a wholly spiritual realm where blessed spirits participate in the divine mind. Dante seems to emphasize this double status by mingling theological and philosophical language, and invoking Aristotle and the neo-Platonists side by side with the poet of the Psalms. In the *Paradiso* the problems raised here will be implicitly resolved by a brilliant recourse to the "metaphysics of light"; when Dante and Beatrice, emerging from the "greatest body," the crystalline sphere or Primum Mobile, pass on "*al ciel ch'è pura luce, / Luce intellettual piena d'amore*" [*Par*. 30.39-40], we know that we are at the precise point at which the *bonum diffusivum sui* that is God's love transforms itself to cosmic energy, "the love that moves the sun and the

other stars." But poetry is perhaps the only means of defining this threshold [Bonaventure, *Sent*. 2. d. 2, a. 2, q. 1, c. 4; Thomas, *Quodl*. 6, q. 11, a. unicus 19].

Similar ambiguities appear in Dante's discussion of the *intelligenze* themselves. Since in governing the several heavens the intelligences engage in a kind of civil life, they must enjoy an active as well as a contemplative existence. But the latter is of a higher order than the former, and no single intelligence can partake of both. Influenced perhaps by Thomas's commentary, Dante imputes to Aristotle in the *Ethics* the view that such divine beings must know only a contemplative life [2.4.13; cp. Aristotle, NE 10.8, 1178b; Thomas, *Exp. Eth.* 10, lect. 12, 2125]. Dante's attempt to resolve the issue is oddly unpersuasive. He argues that the circular motion of the heavens, by which the world is governed, is really a function of the contemplative activity of the intelligences [2.4.13]. Here, as in the case of the Empyrean which they inhabit, we can see Aristotle's celestial movers undergoing a neo-Platonizing transformation, but Dante ends this stage of his discussion by noting that the truth concerning the Intelligences can not be fully grasped by our earthly understanding [2.4.16-17].

The second book concludes with an extended allegory in which the concentric "heavens" or planetary spheres are identified with the seven Liberal Arts, the "starry sphere" with physics and metaphysics, the *Primum mobile* with moral philosophy, and the Empyrean beyond with theology. This synthesis of the natural and the intellectual universe expresses an ideal of education which harks back to the late-antique sources of twelfth-century Platonism, but which Dante has imbued with new life. His emphasis on the ordering function of moral wisdom, and on the happiness attainable through intellectual contemplation, reflects an engagement with the philosophical tradition, and a commitment to philosophy as such, which belong to the later thirteenth century. The final chapter of Book Two affirms the beauty that consists in seeing the causes of those "wonders" which, as the opening of the *Metaphysics* declares, draw us to philosophy.

The third book is perhaps the most important for the student of Dante's knowledge and use of philosophy. Its central theme is praise of philosophy's power, as "l'amoroso uso della sapienza," "the loving use of wisdom," to impart the highest happiness to those who love her, perfecting their natures and drawing them close to God, of whose majesty and wisdom her beauty is the expression. It is largely a meditation on love, understood as Dante's response, intellectual, poetic and psychological, to his enlightenment at the hands of the beautiful lady whom he celebrates as Philosophy.

Early in the third book Dante cites the *Liber de Causis*: Every "substantial form" proceeds from the first cause, God, and participates in His divine nature according to its nobility [3.2.4-7; LC 1.1]. The human soul, noblest of all created forms, loves all things to the degree that they manifest the divine goodness, but desires above all to be united with God. Philosophy is the expression of this desire: Its "form" is "an almost divine love of knowledge" [3.11.13] which leads to "the spiritual uniting of the soul with what it loves" [3.2.3]. It is through philosophy that humanity perfects its "truly human or, better, angelic nature, that is to say the rational [nature]" [3.3.11], discovering in itself "that distinguished and most precious part which is deity" and "participating in the divine nature as an everlasting intelligence" [3.2.14, 19]. As such it mirrors the nobility, wisdom and love of the divine essence and its "loving use of wisdom" becomes by participation "marriage" with God [3.12.11-14].

All of this may appear sheer fantasy, but we should remember that the aim of philosophy as the *Convivio* pursues it is to attain, through natural reason, the greatest happiness of which we are capable in our earthly state. Such felicity is of course circumscribed by our mortality, and the Dante who can celebrate philosophical understanding as a quasi-mystical union with God knows at the same time that true union is granted only through grace, to a soul made receptive by the infusion of virtues which wholly transcend the workings of rational, natural virtue. For as Thomas says, the rational virtues "are dispositions by which man is fittingly disposed with reference to the nature by which he is a man. But the infused virtues dispose man in a higher way, and in view of a higher end; and also, it follows, with reference to some higher nature" [ST 1.2.110.3r]. This "higher nature" is of course the divine nature "through participation in which we are reborn in grace."

Dante acknowledges Thomas's distinction when he speaks of the soul after death as "more than human" [2.8.6], and asserts that to perceive God is not possible for our nature [3.15.10]. For both Dante and Thomas humanness is defined by the conjoining of soul and body, and human knowledge depends on the evidence of the senses [Foster (1965), 69-71; Thomas, ST 1.89a1]. Aristotle had similarly argued that a life of pure contemplation is beyond our strictly human capacity; we can live in this way only to the extent that we have in us "something divine" [NE 10.7, 1177b]. Thomas argues more subtly that the *modus essendi* of the soul joined to the body differs from that of the soul in separation; though they are the same in nature, the separated soul understands, not by means of sensory images, but "through species which it participates in by virtue of the divine light" [ST 1.89.1r]. In the meantime, as Dante acknowledges, there are truths which we can apprehend only as if in a dream, "*come sognando*," [*Conv*. 3.15.6; Nardi (1944), 81-90], and our desire for perfect understanding is necessarily limited, "proportionate to the wisdom which can be acquired here"; for to desire what is beyond the capacity of our intellectual nature would be ethically and rationally incoherent, a desire for imperfection rather than perfection of understanding [3.15.8-10].

But the *Convivio* continually strains against these limits. For Dante, first and foremost a poet of love, the experience of acquiring philosophical understanding has an important psychological component. By enabling us to analyze the processes of perception, philosophy brings us into contact with the true nature of things, and for Dante, as Kenelm Foster observes, the slightest such contact could have a metaphysical value [Foster (1965), 59-60]: "It did not in one sense matter to Dante what the particular object of his knowing might be, since the joy of knowing it was already a foretaste of all conceivable knowledge and all joy; and this precisely because, in knowing, the mind seized truth. . . . once intelligence, the truth-faculty, had tasted truth as such, that is, its own correspondence with reality, it could not help desiring truth whole and entire, that is, its correspondence with all reality." At this point knowledge and the joy of possessing it combine to prepare the ground for faith. By explaining phenomena which without her guidance would merely astonish us, philosophy inspires us to believe "that every miracle can be perceived by a superior intellect to have a reasonable cause" [3.14.14]:

> Our good faith has its origin in this, from which comes the hope that longs for things foreseen; and from this springs the activity of charity. By these three virtues we ascend to philosophize in that celestial Athens where Stoics and Peripatetics and Epicureans, by the

light of eternal truth, join ranks in a single harmonious will.

Philosophy thus conceived can still be regarded as the handmaid of theology, but as Dante develops his philosophical ideal metaphorically in terms of the beauty of the Donna Gentile, it assumes a religious value of its own. Since the wisdom she embodies is the consummation of human self-realization, the Donna Gentile resides in the divine mind as "the intentional exemplar of the human essence" [3.6.6]. In desiring her we desire our own perfection, for she is "as supremely perfect as the human essence can be." When at this point Dante adds a reminder that nothing in our human experience can fully satisfy this desire, he seems to be acknowledging that what Thomas' *Ethics* commentary calls "the ultimate end of desire's natural inclination" is unattainable in this life, since it would require an understanding more complete than any human being can possess [Thomas, *Exp. Eth*. 1, lect. 9, 107; SCG 3.48.2].

But having provided this caution, Dante seems to ignore it, as if unable to resist the conviction that philosophy satisfies our desire in a manner proper to itself. Everything naturally desires its own perfection, and for human beings this is "the perfection of reason" [3.15.3-4; cp. Thomas, *Exp. Eth*. 9, lect. 9, 1872]. But philosophy, as embodied in the Donna Gentile, is not just the consummation of natural understanding. For Dante, as for Aristotle, the human intellect as such is somehow more than human, and he is at times similarly unclear on the question of whether human beings can attain happiness through the exercise of virtue, and to what extent it is a gift of the gods [Foster (1977), 198-201]. Repeatedly he draws a distinction between merely human happiness and that attainable through grace, only to seemingly disregard it in subsequent discussion. Thus in the final chapter of the third treatise he acknowledges the "strong misgivings" that one might have about the happiness attainable through philosophy. Since certain things -- God, eternity, and primal matter are named -- exceed the capacity of our intellect, our natural desire to know must remain unfulfilled in this life [3.15.7]. Dante answers this by affirming, as noted above, that the natural desire for perfection is always proportionate to our capacity to attain it; for to desire the unattainable would be to desire our imperfection [3.15.8-11]. Human happiness, then, consists in the attainment of Aristotle's "human good," through the exercise of the virtues. This is what Dante calls "*l'umana operazione*," and its end is the highest that human beings can attain through their own powers.

Yet philosophy offers the promise of more. The same chapter is climaxed by the vision of Wisdom as "the mother of all things," the origin of all motion and order in the created universe, guiding the quest of human wisdom by the light of the divine intellect. When the human mind is fully informed by philosophy, it would appear, it becomes virtually one of the *intelligenze*, who know both what is above them and what is below, God as cause and the created universe as effect [3.6.4-6]. Thus Dante can speak of our rational nature as our "truly human, or, to speak more exactly, our angelic nature" [3.3.11], as if it enjoyed a more or less mystical existence of a higher order as well as that of the "merely" human nature that pursues the active life of virtue.

The *Liber de causis* says that each cause infuses into its effect the goodness it receives from its own cause, or, in the case of the soul, from God [*Conv*. 3.6.11; LC 4.48]. When in gazing on the body of the Donna Gentile we behold *maravigliose cose*, we are perceiving the effect of a cause which is ultimately God, and thus, Dante asserts [3.6.12-13]:

> it is evident that her form (that is, her soul), which directs the body as its proper cause, miraculously receives the goodness of God's grace. Thus outward appearance provides proof that this lady has been endowed and ennobled by God beyond what is due to our nature . . .

Thus in effect the Donna Gentile is the perfection we desire. Through her we experience the divine goodness, by an outflowing, a *discorrimento* which Dante glosses with a further reference to the *Liber de Causis* [3.7.2; LC 20.157], in terms of the hierarchical emanation of the divine goodness. In the quasi-continuous series of gradations that descends from angel to brute animal, there is no intervening grade between man and angel, so that some human beings are so noble as to be nothing less than angels [Aristotle, NE 7.1, 1145a]. Such is the Donna Gentile; she receives divine virtue just as the angels do [3.7.7]. She is a thing *visibilmente miraculosa*, ordained from eternity by God in *testimonio de la fede* for us [3.7.16-17; Foster (1965), 56]. Philosophy has "wisdom for her subject matter and love for her form" [3.14.1], and God, by instilling his radiance in her, "reduces" that love as nearly as possible to his own similitude [3.14.3; cp. Thomas, SCG 1.91].

Philosophy has clearly become far more than the means whereby human nature achieves self-realization, though this ideal continues to provide a framework for Dante's praise of her. She has assumed the status of Wisdom, *sapientia*, the divine mind as expressed in the order and harmony of creation. Her beauty can only be described in terms of its effects, like the separate substances and God Himself. The true philosopher "loves every part of wisdom, and wisdom every part of the philosopher, since she draws him to herself in full measure" [3.11.12]. Here we may recall Dante's account of how the swift motion of the Primum Mobile expresses its desire for total participation in the divinity of the Empyrean [2.3.8]. And it is in such terms that Dante ends his account of philosophy-as-wisdom. In the final chapter of the third treatise she is explicitly identified with the all-creating Wisdom of God [3.15.15], and Dante concludes in prophetic exhortation [3.15.17]:

> O worse than dead are you who flee her friendship! Open your eyes, and gaze forth! For she loved you before you existed, preparing and ordering your coming; and after you were made, she came to you in your own likeness in order to place you on the straight way.

The fourth treatise of the *Convivio* seems to have been written later than the first three, and it is markedly different in orientation. The principal theme of its *canzone* is the true nature of nobility. Introducing his prose discussion, Dante gives a curious account of how an interruption in his philosophical studies, caused by what the *canzone* calls "disdainful and harsh" behavior on the part of the Donna Gentile, provided an occasion for taking up this topic [4.1.8]:

> Since this lady of mine had somewhat altered the tenderness of her looks at me, especially in those features at which I would gaze when seeking to learn *whether the primal matter of the elements was intended by God* -- and for this reason I refrained for a short period of time from coming into the presence of her countenance -- while living, as it were, in her absence, I set about contemplating the shortcoming within man regarding the above-mentioned error [i.e. a false perception of the bases of human nobility].

That God is the creator of prime matter was an article of faith, and Thomas had dealt decisively with the role of divine will and intellect in the creative act [SCG 2.20.7, 21-24]. That Dante should admit to having entertained doubts about such a question is perhaps a way of indicating his awareness of a danger inherent in his philosophical studies. Deeply concerned to affirm the dignity of reason and the truth embodied in material creation, he may have sensed himself idolizing the secondary powers in whose hierarchical *circulazione* he felt himself, as poet, to be in a special sense participant, and allowing these preoccupations to cloud his awareness of God's omnipotence. The anger of the Donna Gentile would then express his sense of a corresponding loss of focus, a failure to affirm her unique and transcendent role in the expression of the divine will.

Whatever the precise nature of the dilemma to which Dante alludes, the fourth treatise is marked by a noticeable shift away from metaphysics in the direction of ethics and rhetoric. Philosophical knowledge is redirected to the purposes of social and political life, and the treatise, while punctuated like the others by numerous digressions, pursues a single sustained argument. Dante begins by explaining that social order as a condition of human happiness, and that it requires a single governor whose authority embraces that of all particular governors and directs their several efforts to a single end [4.4]. After a long digression on the role of Rome in the providential design of human history, he turns from political to philosophical authority, citing Aristotle as in effect the governor of the mind, "master and leader of human reason insofar as it is directed to man's highest work" [4.6.8]. He then proceeds to qualify both political and philosophical authority, justifying himself at length as he does so. Imputing to Aristotle the statement that "whatever appears true to the majority cannot be entirely false" [*Topica* 1.1, 100b? NE 1.9. 1098b?], he explains that this must be understood to apply, not to sense perception, but only to acts of the mind [4.8.6]. An emperor's authority, too, must be circumscribed; the art of ruling and the laws it creates cannot overrule rational judgment based on the laws of nature [4.9].

On this basis Dante proceeds to refute the view that nobility consists in wealth and ancestry, a view which he here attributes to Frederick II, "the last emperor of the Romans," and for which he will elsewhere cite Aristotle's *Politics* [*Mon*. 2.3.4; Pol. 4.8, 1294a]. Perhaps as significant as the arguments he musters to show the treacherous nature of riches and the uncertain course of nobility from one generation to another is the assertion of Dante's own authority, as philosopher and citizen, that is implied by his elaborate apology for speaking as he does [Ascoli, 35-41]. The gesture nicely epitomizes the project of the *Convivio*, a vernacular discourse which defines for its lay audience the limits of political and scholastic authority, and affirms the autonomy and potential dignity of individual human reason.

The later portions of the fourth treatise are grounded in another Aristotelian definition of nobility, as the perfection of a thing according to its nature [*Conv*. 4.16.7; *Physics* 7.3.246a]. The human expression of this perfection is virtue, moral and intellectual. Electing to address the moral virtues, as more accessible to a lay understanding, Dante begins by describing how nobility is implanted in the nascent soul as the seed of virtue, from which spring the two branches of the active and the contemplative life. The final chapters of the *Convivio* show how the virtues that stem from nobility can direct "the natural appetite of the mind," enabling it to evolve through love of them to the happiness which is the end of virtue [*Conv*. 4.17.8-9; NE 1.13, 1102a].

In the final stanza of the canzone analyzed in the fourth treatise, Dante addresses the poem itself as "Contra-li-erranti mia," "my song against-the-erring ones," and the final chapter of the commentary explains this as an allusion to the *Summa contra gentiles* of Thomas, written "to confound all those who stray from our faith" [*Conv*. 4.30.3]. By thus declaring himself the follower of so fine a craftsman, Dante suggests, he hopes to "ennoble" his own undertaking.

The *Contra gentiles* may seem an odd choice of model. Bruno Nardi considers that Dante had at most a superficial knowledge of this work at the time when he wrote the *Convivio*, and it is certainly the case that he is fundamentally at odds with Thomas over such specific matters as the origin of the soul, the role of the celestial intelligences in creation, and, more important, in claiming for philosophy the power to fulfil the human desire for knowledge in this life [Nardi (1992), 28-29]. On all of these matters Dante is closer to the position of Albert.

On the broader question of the nature of the human desire for knowledge, and the extent to which this desire can be fulfilled by the rational intellect, Dante remains, throughout the *Convivio*, sharply at odds with Thomas. The fourth treatise offers what we may take as his final word, as philosopher, on this question. Having dwelt at length on the insatiability of the base desire for riches, Dante addresses the question of whether our desire for knowledge, too, since it continues to grow as knowledge is acquired, is not similarly base. Dante begins his answer by asserting that "the supreme desire of each thing, and the one that is first given to it by nature, is to return to its first cause," and illustrates this proposition by the images of a traveller on an unfamiliar road, who imagines each house he encounters to be the inn he seeks, and the desires of youth, which focus first on an apple or a pet bird, then evolve to encompass love and prosperity [*Conv*. 4.12.15-16]. But while this may seem to evoke Thomas's view of a single desire which seeks to grow continuously toward union with God, Dante's point is that the path to fulfillment involves multiple desires and the attainment of multiple perfections [*Conv*. 34.13.2]:

> For if I desire to know the principles of natural things, as soon as I know them this desire is fulfilled and brought to an end. If I then desire to know what each of these principles is and how each exists, this is a new and separate desire. Nor by the appearance of this desire am I dispossessed of the perfection to which I was brought by the other, and this growth is not the cause of imperfection but of greater perfection.

Thomas can speak of the natural desire to know as a force like gravity, whose attraction intensifies as it approaches its object [SCG 3.25.13]. In contrast Dante's insistence on types and stages of knowing may seem almost perverse, a matter of emphasizing the stages of the mind's ascent rather than the desire that leads it forward from stage to stage. But what is at stake for Dante is the need to acknowledge human ends as having a definite value of their own, and this need will play an equally important role in Dante's other major philosophical work, the *Monarchia*.

Before leaving the *Convivio*, however, I would like to suggest a way in which Dante's citation of the *Summa contra gentiles* is, after all, an appropriate way of labelling his own undertaking. The *Contra gentiles* is unique among medieval *summae* in aiming to demonstrate, not just the compatibility of

Aristotelian physics and metaphysics with revealed truth, but the extent to which the *invisibilia Dei* can be understood without recourse to that truth. In Norman Kretzmann's phrase it is "a risky *tour de force*" that actively engages unbelievers in metaphysical argument, and spends more time undoing mistakes than affirming Christian doctrine. Revealed truth provides a means of determining the topics to be discussed, and the harmony of natural demonstration with revelation is repeatedly noted, but the basis for demonstration is provided by Aristotle, and what the first three of Thomas's four books present is a case, not for Christianity, but for theism [Kretzmann, pp. 43-53].

Dante seems to acknowledge the pioneering aspect of Thomas's undertaking. Like Thomas, he is testing philosophy, privileging Aristotle as a unique resource capable of helping him discover truth by natural means. Gauthier sees Thomas "nel mezzo del cammin" as he composes the *Contra gentiles*, adopting the position of the Aristotelian *sapiens* to reflect on his own ongoing work and justify it to contemporaries [Gauthier, 179-81]. Dante, too, is deeply concerned to define and justify his own position as a voice of wisdom for his contemporaries. The truths he affirms are encoded in his own poetry, rather than mysteriously embodied in Scripture, and he addresses a cultured but non-Latinate audience unschooled in philosophy. But in substituting the Donna Gentile of philosophical wisdom for Beatrice beata, the "authentic," salvific Beatrice who will reemerge as the voice of truth in the *Commedia*, Dante is establishing a relationship between secular knowledge and the truth that Beatrice embodies analogous to the relation Thomas establishes between philosophy and theology proper.

5. The *Monarchia*

The *Monarchia* is in its own way as idiosyncratic as the *Convivio*. Its purpose, foreshadowed in the discussion of empire in *Convivio* IV, is to demonstrate the necessity of a single ruling power, reverent toward but independent of the Church, capable of ordering the will of collective humanity in peace and concord. Under such a power the potential intellect of humanity can be fully actuated -- the intellect, that is, of collective humanity, existent throughout the world, acting as one. For just as a multitude of species must continually be generated to actualize the full potentiality of prime matter, so the full intellectual capacity of humanity cannot be realized at one time nor in a single individual [*Mon*. 1.3.3-8]. Here Dante adds his own further particularization of this Aristotelian doctrine [*De Anima* 3.5, 430a10-15], asserting that no single household, community, or city can bring it to realization. The ordering of the collective human will to the goal of realizing its intellectual potential requires universal peace [1.4], and this in turn requires a single ordering power through whose authority humanity may achieve unity and so realize the intention and likeness of God [1.8].

The basis of this argument for empire is evidently the first sentence of the Prologue to Thomas' literal commentary on the *Metaphysics*, where he declares that when several things are ordered to a single end, one of them must govern, "as the Philosopher teaches in his *Politics*" [Thomas, *Exp. Metaph.*, Proemium; Aristotle, *Politics* 1.5, 1254a-55a]. For Thomas this is only an analogy, a way of introducing the theme of order as it applies to the soul and its pursuit of happiness. The passage he cites from the *Politics* is concerned only with the rudiments of hierarchy; the idea of "ordering of things to one end" is present only by implication, and Aristotle makes no attempt to develop its metaphysical implications. Dante, however,

seems clearly to associate with Aristotle, or with Thomas' reference to Aristotle, the idea of "a political organization which leads in its way to 'beatitudo' for the whole human race" [Minio-Paluello, 74-77]. One may wonder if Dante's erroneous impression of the Aristotelian passage, which he cites directly with no reference to Thomas in both the *Convivio* and the *Monarchia* [*Conv*. 4.4.5; *Mon*. 1.5.3], is not a symptom of his intense need to draw the Philosopher into support of his view of world empire.

The second of the *Monarchia*'s three books deals with the great example of Rome, describing the city's providential role in world history, largely by way of citations from Roman literature aimed at demonstrating the consistent dedication of Roman power to the public good, and the conformity of Roman *imperium* with the order of nature and the will of God. The third book deals with the crucial issue of the relationship between political and ecclesiastical authority. Dante argues on various grounds that power in the temporal realm is neither derived from nor dependent on spiritual authority, though it benefits from the power of the Papacy to bless its activity. These arguments consist largely in refutations of traditional claims for the temporal authority of the Papacy, but the final chapter makes the argument on positive grounds. Since man consists of soul and body, his nature partakes of both the corruptible and the incorruptible. Uniting two natures, his existence must necessarily be ordered to the goals of both these natures [*Mon*. 3.16.7-9]:

> Ineffable providence has thus set before us two goals to aim at: i.e. happiness in this life, which consists in the exercise of our own powers and is figured in the earthly paradise; and happiness in the eternal life, which consists in the enjoyment of the vision of God (to which our own powers cannot raise us except with the help of God's light) and which is signified by the heavenly paradise. Now these two kinds of happiness must be reached by different means, as representing different ends. For we attain the first through the teachings of philosophy, provided that we follow them putting into practice the moral and intellectual virtues; whereas we attain the second through spiritual teachings which transcend human reason, provided that we follow them putting into practice the theological virtues, i.e. faith, hope, and charity. These ends and the means to attain them have been shown to us on the one hand by human reason, which has been entirely revealed to us by the philosophers, and on the other by the Holy Spirit . . .

This is Dante's most explicit, uncompromising claim for the autonomy of reason, reinforced by the entire world-historical argument of the *Monarchia* and constituting its final justification for world empire. Dante here goes well beyond Augustine's sense of the stabilizing function of empire, and eliminates any hint of the anti-Roman emphasis in Augustine's separation of the earthly and heavenly cities. In the final sentences of the *Monarchia* the temporal monarch becomes, like the aspiring intellect of the *Convivio*, the uniquely privileged beneficiary of a divine bounty which, "without any intermediary, descends into him from the Fountainhead of universal authority" [*Mon*. 3.16.15]. Like the Averroistic reasoning of his earlier claim that only under a world empire can humanity realize its intellectual destiny, this crowning claim shows Dante appropriating Aristotle to the service of a unique and almost desperate vision of empire as a redemptive force. But whether we consider the world view of the *Monarchia* an aberration [D'Entreves, 51] or take it as Dante's straightforward exposition of his views on the relations of secular and religious authority, its categorical definition of the twofold purpose of human life is impossible to

explain away. In the *Paradiso* [8.115-17] as in the *Monarchia*, to be a "citizen" is essential to human happiness, and the idea of an imperial authority independent of papal control remained fundamental to his political thought.

6. The *Commedia* (*The Divine Comedy*)

The *Monarchia*'s crowning vision is not Dante's last word on the subject of human happiness, nor on the possibility of achieving happiness by natural means. The "earthly paradise" which we attain for ourselves through philosophy is certainly not the paradise Dante the pilgrim will discover at the summit of Purgatory. To the philosopher the *Commedia* promises only the cold light and enamelled greenery of Limbo, the somber Elysium where Dante encounters Aristotle and the "philosophic family" who look to him as their master, living out an eternity, not of happiness, but of desire without hope [*Inf*. 4.111-20, 130-44].

The contrast expresses the difference in orientation between the *Commedia* on the one hand and the *Convivio* and *Monarchia* on the other. The *Commedia* is concerned always with the ultimate, eternal destiny of human life, with the transcendence, rather than the fulfillment of human understanding. When Beatrice at the summit of Purgatory utters prophetic words which "soar" far beyond Dante's power to envision her meaning, she explains that his limitations are those of "that school which you have followed," whose teachings are as far from the divine way as the earth from the *Primum Mobile* [*Purg*. 33.82-90]. The "school" in question is the study of philosophy as Dante had pursued and celebrated it in earlier writings. It is his training in this school that makes possible the luminous precision of the great doctrinal passages in the *Purgatorio* and *Paradiso* [*Purg*. 17.90-139; 25.37-87; *Par*. 2.112-48; 7.64-77; 13.52-78; 29.13-45; 30; 97-108], but it is a training that harbors the danger of rationalism and intellectual pride. In the *Convivio* God is the highest good, but remains the distant, unchanging focus of the aspiring mind. In the *Commedia* God assumes an active role as the dispenser of that grace without which the intellectual quest is futile:

> *Io veggio ben che già mai non si sazia*
> *nostro intelletto, se 'l ver non lo illustra*
> *di fuor dal qual nessun vero si spazia.*
>
> [*Par*. 4.124-26]
>
> Translation:
> I see well that never is our intellect satisfied, unless that truth illumines it beyond which no truth may soar.

Bibliography

Editions of Italian Texts

- Contini, Gianfranco, ed., *Poeti del duecento* (2 vols., Milan and Naples, 1960). Vol. II contains the poetry of Guido Guinizzelli and Guido Cavalcanti
- Dante, *Opere minori* (2 vols, Milan and Naples). Includes copiously annotated editions of the *Vita nuova* (Domenico de Robertis); *Convivio* (Cesare Vasoli); *Monarchia* (Bruno Nardi).
- Singleton, Charles, ed., *The Divine Comedy*, 6 volumes (Princeton, 1970-75). Clear prose facing-page translation, along with commentary volumes that quote generously from Thomas Aquinas and others.

Secondary Sources

- Ascoli, Albert, "The Vowels of Authority (Dante's *Convivio* IV.vi.3-4)," in *Discourses of Authority in Medieval and Renaissance Literature*, eds. Kevin Brownlee, Walter Stephens (Hanover, NH and London, 1989), pp. 23-46.
- Boyde, Patrick, *Dante Philomythes and Philosopher. Man in the Cosmos* (Cambridge, 1981)
- Corti, Maria, *Dante a un nuovo crocevia* (Florence, 1981)
- ----, *La felicità mentale. Nuove prospettive per Cavalcanti e Dante* (Turin, 1983)
- d'Entrèves, A. Passerin, *Dante as a Political Thinker* (Oxford, 1952)
- Davis, Charles Till, *Dante and the Idea of Rome* (Oxford, 1957)
- ----, *Dante's Italy and Other Essays* (Philadelphia, 1984)
- Foster, Kenelm, *The Two Dantes and Other Studies* (Berkeley, CA and Los Angeles, 1977)
- ----, "Religion and Philosophy in Dante," in *The Mind of Dante*, ed. Uberto Limentani (Oxford, 1965), pp. 47-78
- Garin, Eugenio, *Rinascite e rivoluzioni. Movimenti culturali dal XIV al XVII secolo* (ed. 2, Rome, 1976)
- Gauthier, René-Antoine, *Saint Thomas d'Aquin, Somme contre les Gentils* (Paris, 1993)
- Gilson, Etienne, *Dante and Philosophy* (tr. David Moore, London, 1949)
- Imbach, Ruedi, *Dante, la philosophie et les laïcs* (Fribourg and Paris, 1996)
- Kretzmann, Norman, *The Metaphysics of Theism: Aquinas's Natural Theology in ’Summa Contra Gentiles" I* (Oxford, 1997)
- Meier, Christel, "*Cosmos Politicus*: Der Funktionswandel der Enzyklopädie bei Brunetto Latini," *Fruhmittelalterliche Studien* 22 (1988): 315-56.
- Minio-Paluello, Lorenzo, "Dante's Reading of Aristotle," in *The World of Dante*, ed. Cecil Grayson (Oxford, 1980), pp. 61-80
- Nardi, Bruno, *Nel mondo di Dante* (Rome, 1944)
- ----, *Studi di filosofia medievale* (Rome, 1960)
- ----, *Saggi di filosofia dantesca* (ed. 2, 1967)
- ----, *Dante e la cultura medievale* (ed. 2, 1983)
- ----, *Dal "Convivio" alla "Commedia"* (ed. 2, Rome, 1992)
- Panella, Emilio, *Per lo studio di fra Remigio dei Girolami, Memorie domenicane*, n.s. 10 (1979)
- Van Cleve, Thomas C., *The Emperor Frederick II of Hohenstaufen. Immutator Mundi.* (Oxford, 1972)
- Vanni Rovighi, Sofia, "Le ‘disputazioni de li filosofanti’," in *Dante e Bologna nei tempi di Dante*

(Bologna, 1967), pp. 179-92
- Vasoli, Cesare, *Otto saggi per Dante* (Firenze, 1995)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Albert the Great [= Albertus magnus] | Book of Causes [= *Liber de causis*]

Winthrop Wetherbee
Cornell University
ww22@cornell.edu

First published: January 29, 2001
Content last modified: March 1, 2001

Constitutionalism

Constitutionalism is the idea, often associated with the political theories of John Locke and the "founders" of the American republic, that government can and should be legally limited in its powers, and that its authority depends on its observing these limitations. This idea brings with it a host of vexing questions of interest not only to legal scholars, but to anyone keen to explore the legal and philosophical foundations of the state. How can a government be legally limited if law is the creation of government? Does this mean that a government can be "self-limiting," or is there some way of avoiding this implication? If meaningful limitation is to be possible, must constitutional constraints be somehow "entrenched"? Must they be enshrined in written rules? If so, how are they to be interpreted? In terms of literal meaning or the intentions of their authors, or in terms of the, possibly ever-changing, values they express? How one answers these questions depends crucially on how one conceives the nature, identity and authority of constitutions. Does a constitution establish a stable framework for the exercise of public power which is in some way fixed by factors like the original meaning or intentions? Or is it a "living tree" which grows and develops in tandem with changing political values and principles? These and other such questions are explored below.

1. Constitutionalism: a Minimal and a Rich Sense

In some minimal sense of the term, a "constitution" consists of a set of rules or norms creating, structuring and defining the limits of, government power or authority. Understood in this way, all states have constitutions and all states are constitutional states. Anything recognisable as a state must have some acknowledged means of constituting and specifying the limits (or lack thereof) placed upon the three basic forms of government power: legislative power (making new laws), executive power (implementing laws) and judicial power (adjudicating disputes under laws). Take the extreme case of an absolute monarch, Rex, who combines unlimited power in all three domains. If it is widely acknowledged that Rex has these powers, as well as the authority to exercise them at his pleasure, then the constitution of this state could be said to contain only one rule, which grants unlimited power to Rex. He is not *legally* answerable for the wisdom or morality of his decrees, nor is he bound by procedures, or any other kinds of limitations or requirements, in exercising his powers. Whatever he decrees is constitutionally valid.

When scholars talk of constitutionalism, however, they normally mean something that rules out Rex's case. They mean not only that there are rules creating legislative, executive and judicial powers, but that these rules impose limits on those powers.[1] Often these limitations are in the form of individual or group rights against government, rights to things like free expression, association, equality and due process of law. But constitutional limits come in a variety of forms. They can concern such things as the *scope* of authority (e.g. in a federal system, provincial or state governments may have authority over health care and education while the federal government's jurisdiction extends to national defence and transportation); the *mechanisms* used in exercising the relevant power (e.g. procedural requirements governing the form and manner of legislation); and of course *civil rights* (e.g. in a Charter or Bill of Rights). Constitutionalism in this richer sense of the term is the idea that government can/should be limited in its powers and that its authority depends on its observing these limitations. In this richer sense of the term, there is no "constitution" in Rex's society because the rules defining his authority impose no such limits. Compare a second state in which Regina has all the powers possessed by Rex except that she lacks authority to legislate on matters concerning religion. Suppose further that Regina also lacks authority to implement, or to adjudicate on the basis of, any law which exceeds the scope of her legislative competence. We have here the seeds of constitutionalism as that notion has come to be understood in Western legal thought.

In discussing the history and nature of constitutionalism, a comparison is often drawn between Thomas Hobbes and John Locke who are thought to have defended, respectively, the notion of a constitutionally unlimited sovereign (e.g. Rex) versus that of a sovereign limited by the terms of a social contract containing substantive limitations on her authority (e.g. Regina).[2] But an equally good focal point is the English legal theorist John Austin who, like Hobbes, thought that the very notion of limited sovereignty is incoherent. For Austin, all law is the command of a sovereign, and so the notion that the sovereign could be limited by law requires a sovereign who is self-binding, who commands him/her/itself. But no

one can "command" himself, except in some figurative sense, so the notion of limited sovereignty is, for Austin (and Hobbes), as incoherent as the idea of a square circle.[3] Though this feature of Austin's theory has some plausibility when applied to the British Parliamentary system, where Parliament is often said to be "supreme" and constitutionally unlimited,[4] it faces serious difficulty when applied to most other constitutional democracies such as one finds in the United States and Germany, where it is clear that the powers of government are legally limited by a constitution. Austin's answer was to say that sovereignty may lie with the people, or some other person or body whose authority is unlimited. *Government bodies* -- e.g. Parliament or the judiciary -- can be limited by constitutional law, but the sovereign -- i.e. "the people" -- remains unlimited. Whether this provides Austin with an adequate means of dealing with constitutional democracies is highly questionable. For Austin's sovereign is a determinate individual or group of individuals whose commands *to others* constitute law. But if we identify the commanders with "the people", then we have the paradoxical result identified by H.L.A. Hart -- the commanders are commanding the commanders. In short, we lapse into incoherence.[5]

2. Sovereign versus Government

Though there are serious difficulties inherent in Austin's attempt to make sense of "the people's sovereignty," his account does bring out the need to distinguish between two different concepts: sovereignty and government. Roughly speaking, we might define "sovereignty" as the possession of supreme (and possibly unlimited) power and authority over some domain, and "government" as those persons or bodies through whom sovereignty is exercised. Once some such distinction is drawn, we see immediately that sovereignty might lie somewhere other than with the government. And once this implication is accepted, we can coherently go on to speak of *limited* government coupled with *unlimited* sovereignty. Arguably this is what one should say about constitutional democracies where the people's sovereignty is thought to be unlimited but the government's power is constitutionally limited. As Locke held, unlimited sovereignty remains with the people who have the normative power to void the authority of their government (or some part thereof) if it exceeds its constitutional limitations.

Though sovereignty and government are different notions, it does seem possible for them to apply to the same individual or body. It is arguable that Hobbes insisted on the identification of sovereign and government insofar as he seemed to require a (virtually) complete transfer of all rights and powers from sovereign individuals to a political sovereign whose authority was to be absolute, thus rendering it possible to emerge from the wretched state of nature in which life is "solitary, poor, nasty, brutish and short."[6] In Hobbes' theory, supreme sovereignty must reside in the supreme governmental person or body who enjoys unlimited power and authority to rule the commonwealth. Anything less than unlimited government would, given human nature and the world we inhabit, destroy the very possibility of stable government. So even if "sovereignty" and "government" are different notions, this neither means nor implies that the two could not apply to one and the same individual(s).

3. Entrenchment

According to most theorists, a further important feature of constitutionalism is that the rules imposing limits upon government power must be in some way be *entrenched*, either by law or by way of "constitutional convention."[7] In other words, those whose powers are constitutionally limited -- i.e. the organs of government -- must not be legally entitled to change or expunge those limits at their pleasure. Most written constitutions contain amending formulae which can be triggered by, and require the participation of, the government bodies whose powers they limit. But these formulae invariably require something more than a simple decision on the part of the present government to invoke a change. Sometimes constitutional assemblies are required, or super-majority votes, referendums, or the agreement of not only the central government in a federal system but also some number or percentage of the governments or regional units within the federal system. Entrenchment not only facilitates a degree of stability over time (a characteristic aspiration of constitutional regimes), it is arguably a requirement of *the very possibility* of constitutionally limited government. Were a government entitled, at its pleasure, to change the very terms of its constitutional limitations, there would, in reality, be no such limitations.

Consider Regina once again. Were she entitled, at her discretion, to remove (and perhaps later reinstate) the constitutional restriction preventing her from legislating on religious matters, then it is questionable whether she could sensibly be said to be "bound" by this requirement. On the other hand, were there a constitutional rule or convention specifying that Regina is entitled to remove this restriction only if she succeeds in convincing two thirds of her subjects to vote for the change, then we might meaningfully speak of constitutional limitation. Of course this constitutional meta-rule or convention is itself subject to change or elimination -- a fact which raises a host of further puzzles. For example, does such an act require application of the very rule in question -- i.e. two third's majority vote -- or are "the people," as sovereign, at liberty to change or expunge it at *their* pleasure? If we accept the distinction between government and sovereignty urged above, as well as the proposition that sovereignty cannot be self-limiting, (X cannot limit X) then we seem led to the conclusion that the constitutional meta-rule -- and hence the constitutional regime of which it is an integral part -- both exist at the pleasure of the people. Entrenchment may be an essential element of constitutional regimes, but constitutions cannot be entrenched against the actions of "the sovereign people" at whose pleasure they exist.

4. "Writtenness"

Some scholars believe that constitutional rules do not exist unless they are in some way enshrined in a written document.[8] Others argue that constitutions can be unwritten, and cite, as an obvious example of this possibility, the constitution of the United Kingdom. One must be careful here, however. Though the UK has nothing resembling the American Constitution and its Bill of Rights, it nevertheless contains a number of written instruments which arguably form a central element of its constitution. Magna Carta (1215 A.D.) is perhaps the earliest document of the British constitution, while others include The Petition of Right (1628) and the Bill of Rights (1689). Furthermore, constitutional limits are also said to be found in certain principles of the common law, explicitly cited in landmark cases concerning the limits of government power. The fact remains, however, that Britain seems largely to have an unwritten constitution, suggesting strongly that writtenness is not a defining feature of constitutionalism.

Why would one think that constitutional norms must be written rules, as opposed to more informal conventions or social rules? One possible reason is that unwritten rules are sometimes less precise and therefore more open to "interpretation," gradual change, and ultimately avoidance, than written ones. If this were true, then an unwritten rule could not, as a practical matter, serve adequately to limit government power. But this need not be the case. Long standing social rules and conventions are often clear and precise, as well as more rigid and entrenched than written ones, if only because their elimination, alteration or re-interpretation typically requires widespread changes in traditional attitudes, beliefs and behaviour.

5. Montesquieu and the Separation of Powers

Does the idea of constitutionalism require, as a matter of conceptual or practical necessity, the division of powers urged by Montesquieu? In Regina's case, there is no such separation. But how, it might be asked, can she be the one (qua judge) who determines whether her legislation satisfies the prescribed constitutional limitation? Even if, *in theory*, Regina's constitution prohibits her from removing her constitutional restriction at will (because she must observe the 2/3rds meta-rule)she can always choose to ignore her restrictions, or to "interpret" them so as to escape their binding force. Perhaps Bishop Hoadly was right when he said (1717) in a sermon before the English King: "Whoever hath an ultimate authority to interpret any written or spoken laws, it is he who is truly the Law-giver to all intents and purposes, and not the person who first wrote or spoke them."[9] Although some constitutional limits, e.g. one which restricts the Mexican President to a single term of office, seldom raise questions of interpretation, many others are ripe for such questions. Regina might argue that a decree requiring all shops to close on Sundays (the common Sabbath) does not concern a religious matter because its aim is a common day of rest, not religious observance. Others might argue, with seemingly equal plausibility, that it does concern a religious matter and therefore lies outside Regina's legislative competence. That constitutions often raise such interpretive questions gives rise to an important question: Does the possibility of constitutional limitation on supreme legislative (and executive) power require, as a matter of practical politics, that judicial power reside in some individual or group of individuals distinct from that in which legislative and executive powers are vested? In modern terms, must constitutional limits on a legislative body like Parliament, the Duma or Congress, or an executive body like the President or her Cabinet, be subject to interpretation and enforcement by an independent judiciary?

Marbury v Madison settled this question in the affirmative as a matter of American law, and most nations follow *Marbury* (and Montesquieu) in accepting the practical necessity of some such arrangement. But it is not clear that the arrangement truly is practically necessary, let alone conceptually so. Bishop Hoadly notwithstanding, there is nothing nonsensical in the suggestion that X might be bound by an entrenched rule, *R*, whose interpretation and implementation is left to *X*. What *R* actually requires is not necessarily identical with what *X* thinks or says that it requires, any more than what the American Constitution requires is necessarily identical with what the American Supreme Court says that it requires. This is so even when there is no superior institution to correct *X*'s judgment, or that of the American Supreme Court, when they go wrong. That constitutional limits can sometimes be interpreted so as to avoid their effect, and no recourse be available to correct mistaken interpretations and abuses of power, does not,

then, imply the absence of constitutional limitation. But does it imply the absence of *effective* limitation? Perhaps so, but even here there is reason to be cautious in drawing general conclusions. There is a long-standing tradition within the British Parliamentary system according to which Parliament is alone in being able not only to create, but also to interpret and implement its own constitutional limits. And whatever its faults, there is little doubt that Parliament typically acts responsibly in observing its own constitutional limits.

6. Constitutional Law versus Constitutional Convention

The idea of constitutionalism is usually thought to require *legal* limitation on government power and authority. But according to most constitutional scholars, there is more to a constitution than constitutional law. Many people will find this suggestion puzzling, believing their constitution to be nothing more (and nothing less) than a formal document, possibly adopted at a special constitutional assembly, which contains the nation's supreme law. But there is a long-standing tradition of conceiving of constitutions as containing much more than constitutional law. Dicey is famous for proposing that, in addition to constitutional law, the British constitutional system contains a number of "constitutional conventions" which effectively limit government in the absence of legal limitation. These are, in effect, social rules arising within the practices of the *political* community and which impose important, but *non-legal*, limits on government powers. An example of a British constitutional convention is the rule that the Queen may not refuse Royal Assent to any bill passed by both Houses of the UK Parliament. Perhaps another example lies in a convention that individuals chosen to represent the State of Florida in the American Electoral College (the body which actually chooses the American President by majority vote) must vote for the Presidential candidate for whom a plurality of Floridians voted on election night. Owing to the fact that they are political conventions, unenforceable in courts of law, constitutional conventions are said to be distinguishable from constitutional laws, which can indeed be legally enforced. If we accept Dicey's distinction, we must not identify the constitution with constitutional law. It includes constitutional conventions as well. We must further recognize the possibility that a government, though *legally* within its power to embark upon a particular course of action, might nevertheless be *constitutionally* prohibited from doing so. It is possible that, as a matter of law, Regina might enjoy unlimited legislative, executive and judicial powers which are nonetheless limited by constitutional conventions specifying how those powers are to be exercised. Should she violate one of these conventions, she would be acting legally, but unconstitutionally, and her subjects might well feel warranted in removing her from office -- a puzzling result only if one thinks that all there is to a constitution is constitutional law.

7. Constitutional Interpretation and Constitutional Theories

As we have just seen, there is (often) more to a constitution than constitutional law. As we have also

seen, constitutional norms need not always be written rules. Despite these important observations, two facts must be acknowledged: (1) the vast majority of constitutional cases hinge on questions of constitutional law; and (2) modern constitutions are predominantly written documents.[11] Consequently, constitutional cases often raise theoretical issues concerning the proper approach to the interpretation of written instruments -- coloured, of course, by the special role of constitutions in defining and limiting the authority and powers of government.

8. The Fixed View and the Living Tree

Although theories of constitutional interpretation are many and varied, they all seem, in one way or another, to ascribe importance to a select number of key factors: textual meaning, political and legal history, intention, and moral/political theory. The roles played by these factors in a theory depend crucially on how the theorist conceives of a constitution and its role in limiting government power. For example, if a theorist views a constitution as foundational law whose existence, meaning and authority derive from the determinate, historical acts of its authors and/or those they represent(ed), and whose principal point is to *fix* a framework within which government power is to be exercised, she may be inclined towards an interpretative theory which accords pride of place to factors like authors' intentions, and literal or plain meaning insofar as the latter is considered the best guide to the former. On what we will call the "*fixed view*" of a constitution, it is natural to think that such factors should govern whenever these are clear and consistent. On the fixed view, a constitution sets a framework for law and politics which is fixed by the historical acts of its authors; it is therefore wholly inappropriate for a judge to ignore such factors when she interprets its provisions, even when doing so would allow her to avoid results which appear unacceptable, perhaps even unjust.

If, on the other hand, one views a constitution as a "*living tree*", which by its nature grows and adapts to contemporary circumstances and beliefs, and whose authority resides in its justice, or in the consent, commitment or sovereignty of "the people *now*," not "the people *then*," then one will be far less likely to find such appeals persuasive, let alone conclusive. One inclined towards the living tree conception will tend to spurn appeals to textual meaning and authors' intentions as attempts to impose the dead hand of the (possibly distant) past upon contemporary society and practice. Government must be limited in power, but the terms of these limitations should be allowed to evolve and adapt in light of changing circumstances and political beliefs. Despite its undoubted appeal to some, the living tree conception faces tough questions: is viewing a constitution as a "living tree", malleable in the hands of contemporary interpreters -- particularly judges -- consistent with its status as foundational law, and with the entrenchment, stability and protection from unwarranted state power which seem to be crucial, if not essential, aspects of the very idea of constitutionally limited government? Different theories of constitutional interpretation split on how they answer this important question.

9. Textualism: The Meaning of a Constitution's Text

No one denies that the literal meanings of the actual words chosen in drafting a constitution play a key

role in determining its impact upon decisions, just as they do in the interpretation of statutes, wills, consent forms, and any other written (and sometimes unwritten) legal instruments. Despite factors such as vagueness, open texture, indeterminacy and the like, the semantic content of a constitutional provision, as a rule or norm intended to convey meaning through the use of words, sets limits to its proper interpretation. As Alice said, words can't just mean whatever one wants them to mean.

Textualism appeals to many, but especially those who accept the fixed view of the constitution, coupled with a belief that a constitution is, principally, one important device through which citizens are protected from unwarranted state power, including unwarranted *judicial* power. Requiring that judges interpret constitutional provisions in light of the meaning of the constitution's text (particularly the "original meaning" it bore at the time of his adoption) respects the role of the founders in fixing, on behalf of the community, the basic framework of government and the limits within which state power is to be exercised. Political decisions about that proper framework and its constituent limits have, on this theory, already been made in a proper forum by those in whose hands such decisions were rightly placed. Their decisions have been communicated and should not, lest stability and legitimacy be threatened, be subject to continuous revisiting and review, particularly by (typically unelected) judges who lack the authority enjoyed by the constitution's authors. The discovery of textual meaning is (it is thought) a largely factual matter, requiring none of the moral and political reasoning appropriately undertaken by the founders. If constitutional change is required, the constitution itself sets procedures through which such changes can be affected. Should these prove ineffective, and yet change still be warranted, then the people, as the sovereign power underlying constitutional democracies, have the authority to abandon the constitution, through revolution, peaceful or not, and to substitute something else. But so long as the constitution remains in force, the semantic content of its rules must be taken as governing all matters of constitutional law.

Despite its obvious appeal, Textualism -- or as it is sometimes called, "strict constuctionism" -- faces a number of difficulties. First, semantic content is not always fully determinate or stable from one generation to the next. This is especially true of words and phrases like "equality," "due process of law," "fundamental justice," "free and democratic society," "freedom of religion" and so on. These seem to lack the determinate and relatively stable semantic content of phrases like "five year term" or "two-thirds majority." The evaluative concepts expressed by the former are highly contestable politically, perhaps even "essentially contestable," and cannot therefore serve the role suggested by the fixed view.

Textualism faces a further difficulty. Even when the meaning of a word or phrase used in a constitution is plain for all to see, it is not always the case that it is considered dispositive. For example, taken in terms of both its original and (perhaps different) contemporary meaning, the First Amendment of the American Constitution is clearly violated by a whole host of American laws, e.g., those proscribing incitement, perjury and libel. Taken literally the First Amendment renders unconstitutional *any law* which in *any way* restricts freedom of speech. If so, then it is unconstitutional in the United States to punish untruthful witnesses, prevent primary school teachers from uttering vicious racial slurs against their minority students, or convict those who incite crowds to violence. But such actions have never been understood to violate the First amendment, leading to the inevitable conclusion that more than semantic meaning governs its interpretation and application. And this is generally, if not universally, true of

modern states and their constitutions. But if more than meaning governs, what else counts? The most obvious choice, especially for those attracted to the fixed view, are the "intentions" of the framers. In response to the suggestion that the American First Amendment prohibits laws against perjury, a defender of the fixed view is likely to reply: "But that can't possibly be what the framers had in mind -- what they intended -- in choosing the words they did." This leads us to a second type of interpretive theory, Originalism, which focusses, not on word meaning, but on the *intentions* of those by whose actions the constitution's various provisions came into existence.

10. Originalism

An Originalist might claim that Textualism is partially correct but doesn't go far enough. The original intentions of a constitution's authors are what really count; and the reason that textual meaning is so important is that it's often the most reliable guide to those intentions. The drafters of a constitution may be presumed to have known and had in mind the standard applications of the words they used, and to have intended the results suggested by those applications, together with the goals and values those applications were best suited to achieve. But when textual meaning fails, direct appeal to the relevant intentions is necessary. In both kinds of cases, however, the ultimate aim is to respect original intentions.

Whatever its precise contours, an Originalist theory is, like Textualism, likely to rest on the fixed view of a constitution. To be sure, the constitution's rules are fixed by the authors' intentions in deciding as they did, and not by the semantic content of the words chosen to communicate those intentions. But they are fixed nonetheless, and must, as a result, not be revisited and revised lest the authority and stability of the constitution be threatened. The intentions of those by whose authority a constitution is made must always govern its interpretation, not the new value judgments and decisions of contemporary judges (or any other interpreters) asking the very same questions the founders intentions were supposed to have settled.

Originalism faces a number of difficulties, some shared with Textualism. For example, original intentions are often unclear, if not completely indeterminate, leaving the interpreter with the need to appeal to other factors. Original intentions can vary from one author to the next, and can range from the very general to the highly specific. At one end of the spectrum are the various, and sometimes conflicting *goals and values* the authors of a provision intended their creation to achieve. At the other end are the very specific *applications* the authors might have had in mind when they chose the provision they did. Did the intended applications of an equality provision encompass *equal access to the legal system* by all groups within society? Or only something more specific like equal access to *fairness at trial*? Did they perhaps include *equal economic and social opportunities* for all groups within society? Different authors might have "intended" all, none, or some of these applications when they chose the equality provision. And as with the general goals and values underlying a provision, there is room for inconsistency and conflict. Constitutional authors, no less than legislators, union activists, or the members of a church synod, can have different goals and applications in mind *and yet settle on the same set of words*. In light of this fact, it is often unhelpful to rely on original intentions when interpreting a constitution.

11. Hypothetical Intent Theory

One of the most serious difficulties faced by Originalism is that contemporary life is often very different from the life contemplated by the authors of a constitution. As a result, many intended applications may now seem absurd or highly undesirable in light of new scientific and social developments and improved moral understanding. Modern life includes countless situations which the authors of a constitution could not possibly have contemplated, let alone intended to be dealt with in any particular way. The right to free speech which found its way into many constitutions in the early modern period, could not possibly have been intended by its defenders to encompass, e.g., pornography on the internet. In response to such difficulties, an Originalist might appeal to what we can call "hypothetical intent." The basic idea is that we should always consider, in such instances, the hypothetical question of what the original authors *would have intended* to be done in the case at hand had they known what we now know to be true. We are, on this view, to put ourselves imaginatively in the authors' shoes, and determine, in light of their intended goals and values, and possibly by way of analogy with their intended applications, what they would have wanted to be done in the new circumstances.

The Hypothetical Intent Theory faces difficulties too. First, the theory presupposes that we can single out one, consistent set of values, goals and applications attributable to the authors, in terms of which we are to ask the question: What would they have wanted to have done given these (intended) values, goals and applications? But as we have already seen, the authors of a constitution invariably have different things in mind when they agree on a constitutional text. Second, even if we could single out, at some appropriate level of generality, a set of goals, values and applications from which our hypothetical inquiry is to proceed, it is unlikely that there will always be a uniquely correct answer to the question of what the authors would have intended in these cases which they did not anticipate and could not possibly have imagined. What would an 18th century founder, firmly in favour of freedom of speech, have thought about child pornography on the internet? Thirdly, and perhaps most importantly, we are left with the question of why it much matters what a long dead group of individuals might have wanted done were they apprised of what we now know. The main appeal of the original intent theory is that it appears to tie constitutional interpretation to *historical decisions actually made* by individuals with authority to decide questions concerning the proper limits of government power. If we are now to consider, not what they *did* decide, but what they *might have decided* had they known what we now know, then the question naturally arises: Why not just forget this theoretically suspect, hypothetical exercise and make the decisions ourselves? There is some plausibility in the claim that the decision should be made in light of the very general goals and values probably intended by the authors -- if, that is, one could discover what these were and if they could all be rendered consistent. But why should we wish to perpetuate their possibly misguided views about the appropriate ways in which to secure these goals and values? Unless we reject completely the idea that there might be moral progress, or the idea that any such progress must always be dismissed for the sake of a fixedness allegedly guaranteed by adherence to authors' intent, there seems little reason to believe that we should be so tied. To think otherwise might well be to allow the dead hand of the past to govern the affairs of today.

True enough, it might be replied. But the alternative is one which undermines the very point of

constitutions. If we view a constitution as a living tree whose limitations are constantly open to revisiting and revision in light of changing times and improved moral/political understanding, then it can no longer function as a stable instrument whose very point and purpose is to limit the power of government -- particularly, though not exclusively, arbitrary judicial power. Arguments of political morality may be necessary to frame a constitution, but if judges and other contemporary interpreters are allowed to construe it in light of how they choose to understand those limits, then the possibility of limitation vanishes. But does it? One theorist who thinks not is Ronald Dworkin, whose theory of constitutional interpretation attempts to do justice to both these points of view.

12. Dworkin: Moral Theory

For Dworkin, historical factors like semantic meaning and intention, though always important, are in no way dispositive. They in no way *fix* the limits of government until such time as an amendment passes or a revolution occurs. On the contrary, constitutions frame the terms of an ongoing political debate about the moral principles of justice, fairness and due process underlying a nation's constitutional limits on government power. And as the political community's understanding of these principles develops and improves, the very content of the constitution develops and, it is hoped, improves along with it.

A crucial element in Dworkin's constitutional theory is his general claim that the law of a community includes more than any explicit rules and decisions authoritatively adopted in accordance with accepted procedures. It does, of course, include many such rules and decisions and these can be found, paradigmatically, in statute books, judicial decisions and, of course, written constitutions. These are often termed "positive law." But the positive law in no way exhausts the law according to Dworkin. Most importantly, for our purposes, it in no way exhausts that part of law we call "the constitution." In Dworkin's view, a constitution includes the principles of political morality which provide the best explanation and moral justification -- i.e. the best interpretation -- of whatever limits have been expressed in positive law. Hence, constitutional interpretation must always invoke a theory of political morality. One concerned to interpret the limits upon government power and authority imposed by a constitution must look to an interpretive theory which provides the positive constitutional law with its morally best explanation and justification.

The development of an interpretive theory of the constitution is, Dworkin acknowledges, an extremely difficult task, and people of good will and integrity will reasonably disagree about which theory is best. There is no mechanical, morally neutral test to apply, only the competing interpretations of those whose task it is to interpret. This does not mean, however, that attempting to evaluate theories is foolish, or that there really is no such thing as a best theory since there is no mechanical way of discovering it. The presence of disagreement, controversy, and uncertainty in constitutional cases, does not entail that there are no right answers to the questions posed, and no uniquely correct theory which determines what those answers are and hence what the constitution actually requires. The presence of such factors entails only that interpreters must, as they must do in all interpretive enterprises, including the arts, the sciences, and the law, exercise judgment in fashioning their interpretive theories. Dworkin goes so far as to argue that in a mature legal system there almost always is a best constitutional theory, and judges (and legislators)

are duty-bound to try their best to discern and implement its requirements in making their decisions.

There are, for our purposes, three important implications of Dworkin's theory of constitutional interpretation. First, original intentions and semantic meaning at best set the stage for the debates in political morality which constitutional cases both require and licence. They seldom, if ever, settle matters. Second, constitutional cases require the kind of decision-making which is, on the Originalist and Textualist theories, properly undertaken only by those who have already fixed the standards contained within the constitution -- i.e. its authors or framers. The kind of morally neutral decision-making, under standards set by other responsible agents, to which the Originalist and Textualist theories aspire, is simply impossible on Dworkin's theory. Dworkin's theory requires wholesale rejection of the fixed view.The constitution is not a finished product handed down in a form fixed till such time as its amending formula is invoked successfully or a revolution occurs. Rather it is a work in progress requiring continual revisiting and reworking as our theories concerning its limits are refined and improved. It is, in short, a living tree.

A third, related implication of Dworkin's theory is that judges in constitutional cases are not merely agents of the authors in carrying out their explicit decisions. On the contrary, they are partners with the authors in an ongoing project, one which requires participants, both then and now, to engage in the kind of moral decision-making which, on the fixed view, settled matters when the constitution was adopted (and/or amended). The limits to government power are, on Dworkin's theory, essentially contestable, ad infinitum. If there is a correct theory of a constitution, it requires, for its development and elaboration, an interpreter of super-human powers of moral, political and legal reasoning. In short, it requires Dworkin's judge Hercules.[12] But Hercules is a product of Dworkin's imagination, and so the project of interpreting the contestable terms of constitutions is an ongoing one, requiring each and every interpreter to provide her own best, and undoubtedly imperfect, interpretation of the limits placed upon government by her constitution. The latter is never fixed.

13. Critical Theory

That it requires the skill, acumen and insight of a Hercules is seen by many theorists as a serious drawback of Dworkin's approach to constitutional interpretation. If ordinary judges, with their limited skill, integrity and objectivity are at liberty to decide in terms of their own, highly contestable moral theories of the constitution, then the inevitable result is a kind of unbridled judicial activism which threatens the stability and legitimacy of the constitution and the limits on government power which it is supposed to represent. Instead of limitations properly fixed and settled by apolitical factors like historical intentions and meanings, we would have "limitations" continually in flux and subject to different interpretations by different judges with their own theories of political morality. Those of an originalist or texualist bent will see in such consequences sufficient reason to reject Dworkin's theory in favour of their alternative. But for many constitutional scholars, originalism and textualism are as problematic as Dworkin's interpretive theory. For these "critical theorists", semantic meaning, historical intentions, and herculean interpretive theory, all fail, in one way or the other, to fix meaningful limits upon government power.[13] As a result, reliance on such factors in constitutional adjudication only serves to rationalize the

purely political decisions by judges pursuing, consciously or not, their own political ideologies. A further consequence is suppression of those -- women, minority racial groups, the poor, and so on -- whose interests are not supported by these ideologies. Instead of the curbing of arbitrary government power for which the idea of constitutionalism is supposed to stand, we have political suppression disguised in a cloak of false constitutional legitimacy.

So critical theorists are highly skeptical of constitutional practice and theories which applaud constitutionalism as a bulwark against oppression.[14] As we saw at the outset, a key element in the idea of constitutionalism is that government can/should be limited in its powers and that its authority depends on observance of those limits. We further noted that the authority of constitutions in liberal democracies is generally thought to lie in "the people." One important implication of critical theory is that the concept of " the people" is as much a fabrication as is Dworkin's Hercules. Instead of "we the people", western societies are comprised of various groups competing either for domination (e.g., white males and the wealthy) or for recognition and the elimination of oppression (e.g. the poor, women, and racial minorities). The law, including constitutional law, is a powerful tool which has, historically, been utilized by dominant groups to secure and maintain their superior status. As such, a constitution is anything but the protection from unwarranted power that its champions have heralded over the centuries. What is taken to be the plain meaning of the word "equality" is what the dominant group understands it to be. What is taken to be the obvious historical intentions of the framers is whatever intentions fit the ideologies of the dominant groups. What is taken to be the best moral theory underlying the constitution is nothing more than a rationalization of current social structures, all of which systematically oppress the interests of women, minorities and the poor.

Critical theories represent a serious challenge not only to conventional theories and established practices of constitutional interpretation, but to the very idea of constitutionalism itself -- the idea that government can and should be limited in ways which serve to protect us from unwarranted state power. According to originalists and textualists, the constitution protects us from judges and other officials by restricting them to morally neutral decisions about historical intentions and semantic meanings. According to Dworkin, it is Hercules' best moral theory of the constitution which serves as the bulwark against oppression. One crucial feature of Hercules' theory is that it is often at odds with received opinion, in particular with the self-serving convictions and prejudices of the various dominant groups within society. Following Hercules' moral theory of the constitution will lead a judge to protect the rights of oppressed groups from the power of dominant groups, especially when that power has the sanction of legislation. But the ordinary judge is not, critical theorists will insist, identical with Hercules. On the contrary, he is more often than not himself a member of the dominant group (e.g. wealthy, white males), and shares the social background, education, perspective, and values of that group. As a result, his conceptions of the relevant contested concepts (e.g., equality or freedom of expression) will be their conceptions -- i.e. conceptions which serve the interests of the dominant groups against whom the constitution is meant to serve as protection. But if semantic meaning, intentions and Hercules' best theory are all at the mercy of dominant ideologies, then the kind of protections heralded by the idea of constitutionalism may be a myth, and a harmful one at that. So what is the solution according to critical theorists? The proffered solutions vary considerably from one critical theorist to the next, depending on how radical or skeptical the theorist tends to be. A revolutionary communist might advocate the complete overthrow of constitutional,

democratic government, while many liberal feminists are content to work within existing constitutional systems to eradicate the vestiges of patriarchy which have survived recent feminist movements.[15] But all seem to agree that progress can be made only if the myths surrounding constitutional protection -- the constraining force of meaning, intention, and objective moral theory -- are all exposed, and that the true political forces at work in constitutional practice are acknowledged and dealt with openly. Whether the idea of constitutionalism can survive the lessons of critical theory is a very good question.

Bibliography

- Ackerman, Bruce, *We The People: Foundations* (Cambridge: Harvard University Press, 1991)
- Alexander, Larry, ed., *Constitutionalism* (Cambridge: Cambridge University Press, 1998)
- Altman, Andrew, *Critical Legal Studies: A Liberal Critique* (Princeton: Princeton University Press, 1990)
- Bentham, Jeremy, "Constitutional Code", in Vol. 9, *Works of Jeremy Bentham*, pp. 119-124, Bowring ed. (1838-1843)
- Bickel, Alexander, *The Least Dangerous Branch: The Supreme Court at the Bar of Politics* (New Haven: Yale University Press, 1962)
- Blackstone, Sir William, *Commentaries on the Laws of England*, Vol 1-4 (1765-69)
- Blaustein and Gisbert, ed., *Constitutions of the Countries of the World* (Dobbs Ferry NY: Oceana Publications, 1994)
- Bobbit, Philip, *Constitutional Fate: Theory of the Constitution* (New York: Oxford University Press, 1982)
- _____, *Constitutional Interpretation* (Oxford: Blackwell, 1991)
- Bork, Robert, "Neutral Principles and Some First Amendment Problems," I*ndiana Law Journal*, Vol. 47, no. 1 (1971), 17
- _____, *The Tempting of America: The Political Seduction of the Law* (New York: Macmillan, 1990).
- Brison, S & Sinnott-Armstrong, W, *Contemporary Perspectives on Constitutional Interpretation (*Boulder Co.: Westview Press, 1993)
- *Buffalo Law Review*, Vol. 36 (1987) "Constitutional Law from a Critical Legal Perspective: A Symposium"
- Dicey, A.V., *Introduction to the Study of the Law of the Constitution*, 9th ed. (London: Macmillan, 1948)
- Dworkin, Ronald, *Taking Rights* Seriously (London: Duckworth, 1978)
- _____, *A Matter of Principle* (Cambridge: Harvard University Press, 1985)
- _____, *Law's Empire* (Cambridge: Harvard University Press, 1986)
- _____, *A Bill of Rights for Britain* (London: Chatto & Windus, 1990)
- _____, *Freedom's Law: The Moral Reading of the American Constitution*(Cambridge: Harvard University Press, 1996)
- Ely, John, *Democracy and Distrust* (Cambridge: Harvard University Press, 1980)
- Gray, John Chipman, "A Realist Conception of Law", *The Philosophy of Law* (eds. Feinberg & Gross), 3rd ed.

- Grey, Thomas, "Do We Have An Unwritten Constitution?" 27 *Stanford Law Review* (1975)
- _____, "Constitutionalism: An Analytic Framework", in *Constitutionalism: (Nomos XX)*, ed. R. Pennock & J. Chapman (New York: New York University Press, 1979), 189
- Hart, H.L.A., *The Concept of Law*, 2nd ed. (Oxford: Oxford University Press: 1994)
- Hobbes, Thomas, *De Cive* (The Philosophical Rudiments Concerning Government and Society)(1642) (various editions)
- _____, *Leviathan* (1651) (various editions)
- Hogg, Peter, *Constitutional Law of Canada* (Toronto: Carswell, 1999)
- Locke, John, *Two Treatises of Government* (1690) (various editions) especially Book II, Chapters XI-XIV
- Lyons, David, "Constitutional Interpretation and Original Meaning", in *Moral Aspects of Legal Theory: Essays on Law, Justice and Political Responsibility* (Cambridge: Cambridge University Press, 1993)
- _____, "A Preface to Constitutional Theory" in *Moral Aspects of Legal Theory: Essays on Law, Justice and Political Responsibility* (Cambridge: Cambridge University Press, 1993)
- MacKinnon, Catherine, *Feminism Unmodified: Discourses on Life and Law* (Cambridge: Harvard University Press, 1987)
- _____, *Toward a Feminist Theory of the State*, (Cambridge: Harvard University Press, 1989)
- _____, *Only Words* (Cambridge: Harvard University Press, 1993)
- Marshall, Geoffrey, *Constitutional Theory* (Oxford: Oxford University Press, 1971)
- _____, *Constitutional Conventions: The Rules and Forms of Political Accountability* (Oxford: Oxford University Press, 1984)
- McIlwain, Charles, *Constitutionalism: Ancient and Modern* (Ithica, NY: Cornell University Press,1947)
- Michelman, Frank, "Constitutional Authorship," in ed. L. Alexander, *Constitutionalism* (Cambridge University Press, 1998)
- Montesquieu, Baron de, *The Spirit of the Laws*, tr. Thomas Nugent, ed., F. Neumann (New York, 1949)
- Rehnquist, William, "The Notion of a Living Constitution," *Texas Law Journal* 64 (1976), 695
- Richards, David, *Toleration and the Constitution* (New York: Oxford University Press, 1986)
- Paine, Tom, "The Rights of Man," in *The Essential Tom Paine* (1969)
- Pennock, R & J. Chapman, ed., *Constitutionalism: (Nomos XX)* (New York: New York University Press, 1979), 189
- Rakove, Jack, ed., *Interpreting the Constitution: The Debate Over Original Intent* (Boston: Northeastern University Press, 1990)
- Raz, Joseph, "On the Authority and Interpretation of Constitutions: Some Preliminaries" in ed. L. Alexander, *Constitutionalism*(Cambridge: Cambridge University Press, 1998)
- Rubenfeld, Jed, "Legitimacy and Interpretation," in ed. L. Alexander,*Constitutionalism*, (Cambridge: Cambridge University Press, 1998)
- Shklar, Judith, *Montesquieu* (Oxford: Oxford University Press, 1987)
- Smith, Patricia, ed., *Feminist Jurisprudence* (New York: Oxford University Press, 1992)
- Strossen, Nadine, *Defending Pornography* (New York: Scribner, 1995)
- Tribe, Laurence, *American Constitutional Law* (1978)

- _____, "Taking Text and Structure Seriously: Reflections on Free-Form Method in Constitutional Interpretation," *Harvard Law Review* 108 (1995)
- Tushnet, Mark, *Red, White and Blue: A Critical Analysis of Constitutional Law* (Cambridge: Harvard University Press, 1988)
- _____, "Constitutional Interpretation, Character and Experience", *Boston University Law Review*, Vol. 72, no. 4
- Unger, Roberto, *The Critical Legal Studies Movement* (Cambridge: Harvard University Press, 1986)
- Jeremy Waldron, "A Rights-Based Critique of Constitutional Rights,"*Oxford Journal of Legal Studies*, Vol. 13 (1993)
- _____, "Precommitment and Disagreement," in ed. Alexander, *Constitutionalism*(Cambridge: Cambridge University Press, 1998)
- _____, *Law and Disagreement* (Oxford: Oxford University Press, 1999)

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

legal obligation and authority | legal reasoning: interpretation and coherence | legal reasoning: precedent and analogy

Wil Waluchow
McMaster University
walucho@mcmaster.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 10, 2001
Content last modified: January 10, 2001

Stanford Encyclopedia of Philosophy
Notes to Constitutionalism

Notes

1. Unless otherwise indicated, the term "constitutional" (and its cognate terms "constitutionalism", "constitution", and so on) should henceforth be understood to carry this richer meaning.

2. Whether Locke and Hobbes are properly invoked in this way is perhaps open to question. There is reason to believe that Locke's argument defends political, as opposed to strictly legal, limitations upon the sovereign. But as we shall see, constitutionalism seems to require legal limitation. It might be argued that *effective* political limitation requires legal limitation as well, but this does not seem strictly necessary. More on this later.

3. "For to be subject to laws is to be subject to the commonwealth - that is to the sovereign - that is, to himself, which is not subjection but freedom from the laws." (Leviathan, Ch. 29, 255)

4. What Parliament does "no authority upon Earth can undo." (Sir William Blackstone)

5. See Hart, *The Concept of Law*, pp 73-78. For Austin, see *The Province of Jurisprudence Determined*, Lecture VI.

6. *Leviathan*, Part 1, Ch. 13. Although Hobbes's sovereign is constitutionally unlimited, Hobbes insisted that individuals retained the right to self-preservation. It would be incoherent, Hobbes thought, for individuals to give up that right the protection of which is the very reason people have for creating a sovereign power. Although individuals retain the right to self-preservation, it is also true that Hobbes' unlimited sovereign has the right to take anyone's life if, in the sovereign's judgment, this is necessary to preserve the well being of the commonwealth.

7. Constitutional conventions are explored in Sec. 6 below. Although entrenchment is an almost universal characteristic of modern constitutions, and although one could plausibly argue that it is practically desirable, it may not be absolutely necessary. Some constitutional norms are ordinary statutes amenable to introduction and change by ordinary legislative procedures. Indeed, some constitutions are almost wholly statutory, e.g. the 1848 Italian Constitution and the constitution of New Zealand.

8. See, e.g., J. Rubenfeld, "Legitimacy and Interpretation."

9. As quoted in John Chipman Gray, "A Realist Conception of Law," p. 12.

10. It is arguable that the people of the United Kingdom, in virtue of their membership in the European Community and the fact that British Courts now enforce, as binding, Community law, have in fact relinquished their unlimited sovereignty. If the law of member states (e.g. France, Denmark, and the UK) must now be consistent with Community law, and the latter is immune from legislative change or repeal through legislative acts on the part of member governments, then it can be argued that the sovereignty of the member states within the European Community has been replaced by the sovereignty of "the people of Europe."

11. Henceforth, and unless otherwise indicated, all uses of the word "constitution" (and cognate terms) should be understood as referring to constitutional *law*.

12. Hercules is first introduced by Dworkin in Ch. 4 of *Taking Rights Seriously* and reappears in subsequent writings, most notable, *Law's Empire*.

13. Critical theories come in a variety of forms, the most influential ones being the Critical Legal Studies movement and feminist jurisprudence.

14. "I no longer believe that constitutional theory constrains, or is supposed to constrain judges. Rather...it serves primarily to provide a set of rhetorical devices that judges can deploy as they believe effective." (Mark Tushnet, "Constitutional Interpretation, Character and Experience, p. 759.)

15. See, e.g. Nadine Strossen, *Defending Pornography*.

Wil Waluchow
McMaster University
walucho@mcmaster.ca

First published: January 10, 2001
Content last modified: January 10, 2001

Curry's Paradox

Curry's paradox, so named for its discoverer, namely Haskell B. Curry, is a paradox within the family of so-called paradoxes of self-reference (or paradoxes of circularity). Like the liar paradox (e.g., 'this sentence is false') and Russell's paradox, Curry's paradox challenges familiar naive theories, including naive truth theory (unrestricted T-schema) and naive set theory (unrestricted axiom of abstraction), respectively. If one accepts naive truth theory (or naive set theory), then Curry's paradox becomes a direct challenge to one's theory of logical implication or entailment. Unlike the liar and Russell paradoxes Curry's paradox is negation-free; it may be generated irrespective of one's theory of negation. An intuitive version of the paradox runs as follows.

Consider the following list of sentences, named 'The List':

1. Tasmanian devils have strong jaws.
2. The second sentence on The List is circular.
3. If the third sentence on The List is true, then *every sentence* is true.
4. The List comprises exactly four sentences.

Although The List itself is not paradoxical, the third sentence (a conditional) is. Is it true? Well, suppose, for conditional proof, that its antecedent is true. Then

> the third sentence of The List is true

is true. By substitution, it follows that

> If the third sentence of The List is true, then every sentence is true

is true. But, then, Modus Ponens on the above two sentences yields that

> every sentence is true

is true. So, by conditional proof, we conclude that

> If the third sentence of The List is true, then every sentence is true

is true. By substitution, it follows that

> the third sentence of The List is true

is true. But, now, by Modus Ponens on the above two sentences we get that

> every sentence is true

is true. By naive truth theory we disquote (or, in this case, dis-display, as it were) to conclude: Every sentence is true! So goes (one version of) Curry's paradox.

---

1. Brief History and Some Caveats

In 1942 Haskell B. Curry presented what is now called *Curry's paradox*. Perhaps the most intuitive version of the paradox is due to Arthur N. Prior (1955), who recast Curry's paradox as a "proof" of God's existence. (Let C = ‘If C is true then God exists’.) The version presented above is (in effect) Prior's version.

There are basically two different versions of Curry's paradox, a truth-theoretic (or proof-theoretic) and a set-theoretic version; these versions will be presented below. For now, however, there are a few caveats that need to be issued.

Caveat 1. *Loeb's Paradox*. Prior's version is (in effect) rehearsed by Boolos and Jeffrey (1989), where neither Prior nor Curry is given credit; rather, Boolos and Jeffrey point out the similarity of the paradox to reasoning used within the proof of Loeb's Theorem; and subsequent authors, notably Barwise & Etchemendy (1984), have called the paradox *Loeb's paradox*. While there is no doubt strong justification for the alternative name (given the similarity of Curry's paradox to the reasoning involved in proving Loeb's Theorem) the paradox does appear to have been first discovered by Curry.

Caveat 2. *Geometrical Curry Paradox (Jigsaw Paradox)*. This is *not* the same Curry paradox under discussion; it is a well-known paradox, due to *Paul* Curry, having to do with so-called geometrical dissection. (The so-called Banach-Tarski geometrical paradox is related to *Paul* Curry's geometrical

paradox.) See Gardner 1956 and Fredrickson 1997 for full discussion of *this* (geometrical) Curry paradox.

2. Curry's Paradox: Truth- and Set-Theoretic Versions

Truth-Theoretic Version

Assume that our truth predicate satisfies the following T-schema:

> *T-Schema*: $\mathrm{T}[A] \leftrightarrow A$,

where '[ ]' is a name-forming device. Assume, too, that we have the principle called *Assertion* (also known as *pseudo modus ponens*):

> *Assertion*: $(A \,\&\, (A \rightarrow B)) \rightarrow B$

(NB: We could also use the principle called *Contraction*: $((A \rightarrow (A \rightarrow B)) \rightarrow (A \rightarrow B)$.) Curry's paradox quickly generates *triviality*, the case in which everything is true.

By diagonalization, self-reference or the like we can get an arbitrary sentence, C, such that:

> $C \;=\; \mathrm{T}[C] \rightarrow \boldsymbol{F}$,

where $\boldsymbol{F}$ is anything you like. (For effect, though, make $\boldsymbol{F}$ something obviously false.) By an instance of the T-schema ('$\mathrm{T}[C] \leftrightarrow C$') we immediately get:

> $\mathrm{T}[C] \leftrightarrow (\mathrm{T}[C] \rightarrow \boldsymbol{F})$,

Again, using the same instance of the T-Schema, we can substitute C for T[C in the above to get (1):

1. $C \leftrightarrow (C \rightarrow \boldsymbol{F})$	[by T-schema and Substitution]
2. $(C \,\&\, (C \rightarrow \boldsymbol{F})) \rightarrow \boldsymbol{F}$	[by Assertion]
3. $(C \,\&\, C) \rightarrow \boldsymbol{F}$	[by Substitution, from 2]
4. $C \rightarrow \boldsymbol{F}$	[by Equivalence of C and $C\&C$, from 3]
5. C	[by Modus Ponens, from 1 and 4]
6. $\boldsymbol{F}$	[by Modus Ponens, from 4 and 5]

Letting $\boldsymbol{F}$ be anything entailing triviality Curry's paradox quickly "shows" that the world is trivial!

Set-Theoretic Version

The same result ensues within naive set theory. Assume, in particular, the (unrestricted) axiom of abstraction (or comprehension):

> *Unrestricted Abstraction*: $x \in \{y \mid A(y)\} \leftrightarrow A(x)$.

Moreover, assume that our conditional, $\rightarrow$, satisfies Contraction (as above), which permits the deduction of

> $(s \in s \rightarrow A)$

from

> $s \in s \rightarrow (s \in s \rightarrow A)$.

In the set-theoretic case, let $C =_{\mathrm{df}} \{x \mid x \in x \rightarrow \boldsymbol{F}\}$, where $\boldsymbol{F}$ remains as you please (but something obviously false, for effect). From here we reason thus:

1. $x \in C \leftrightarrow (x \in x \rightarrow \boldsymbol{F})$	[by Naive Abstraction]
2. $C \in C \leftrightarrow (C \in C \rightarrow \boldsymbol{F})$	[by Universal Specification, from 1]
3. $C \in C \rightarrow (C \in C \rightarrow \boldsymbol{F})$	[by Simplification, from 2]
4. $C \in C \rightarrow \boldsymbol{F}$	[by Contraction, from 3]
5. $C \in C$	[by Modus Ponens, from 2 and 4]
6. $\boldsymbol{F}$	[by Modus Ponens, from 4 and 5]

So, coupling Contraction with the naive abstraction schema yields, via Curry's paradox, triviality.

Significance, Solutions, and Open Problems

Significance

What is the significance of Curry's paradox? The answer depends on one's approach to paradox in general. Any comprehensive theory of language has to give some sort account of the paradoxes (e.g., the liar, or Russell's, or etc.). Classical approaches tend to fiddle with the T-schema (or naive abstraction) or reject the existence of certain (paradoxical) sentences. Such classical approaches tend to respond to Curry's paradox in the same fashion — by rejecting the existence of Curry sentences or fiddling with the unrestricted T-schema (or naive abstraction). Some popular variations of these two options include Gupta-

Belnap revision theory (1993), Tarski's familiar hierarchical theory (or Russellian type theory), Simmons's singularity theory (1993), Burge's indexical theory (1979), Kripke's fixed point semantics (1975), Gaifman's pointer semantics (1988), Barwise-Etchemendy situation-cum-Aczel-set-theory (1984), and others. (NB: These theories are quite different from each other; however, each of them fits under one of the two so-called classical options mentioned above; they either modify the naive T-schema or reject the existence of so-called strengthened liar sentences.) Where Curry's paradox becomes especially significant is not with classical approaches but rather with certain non-classical approaches; specifically, Curry's paradox is a direct challenge to any non-classical approach that attempts to preserve naive truth (or set) theory in full. Such approaches attempt to preserve naive truth (or set) theory, preserve the apparent existence of Curry sentences, and avoid the apparent non-triviality of the world. Satisfying these desiderata requires a paraconsistent logic, one that affords inconsistent but non-trivial theories. What Curry's paradox shows is that not just any old paraconsistent logic will do; in particular, on pain of triviality, no connective in the language can satisfy contraction or absorption and support the T-scheme or Naive comprehension scheme. Among other things, this constraint rules out quite a few popular candidates for implicative conditionals -- including, for example, various popular relevant conditionals, including those of ***E*** and ***R***.

A Solution

There is great interest in resolving the paradoxes in the sort of non-classical fashion suggested above. Such interest, coupled with Curry's paradox, has fostered ongoing interest in non-classical (paraconsistent) semantics for entailment. One area in which such research is growing is substructural logic. While there is no generally accepted (non-classical) solution to Curry's paradox one approach is particularly promising, an approach due to Graham Priest (1992) and based upon Kripke's invocation of non-normal worlds. (Kripke invoked such worlds for purposes of modeling Lewis systems weaker than *S4*, not for purposes of solving Curry's paradox.) The idea may be seen easily through its semantics, as follows.

Setting negation aside (for purposes of Curry), we assume a propositional language with the following connectives: conjunction (&), disjunction ($\vee$), and entailment ($\rightarrow$). (For purposes of resolving Curry's paradox, negation may be set aside; however, the current semantics allow for a variety of approaches to negation, as well as quantifiers.) An interpretation is a 4-tuple, (W,N,[ ], f), where W is a non-empty set of worlds (index points), N is a non-empty subset of W, [ ] is a function from propositional parameters to the powerset of W; we may, for convenience, see the range of [ ] as comprising propositions (sets of worlds at which various sentences are true), and so call the values of [ ] *propositions*. We let NN be the set of so-called non-normal worlds, namely $NN = W - N$. In turn, f is a function from (ordered) pairs of propositions to NN. Now, [ ] is extended to all sentences (A, B, ...) via the following clauses:

$[A\&B] = [A]\cap[B]$

$[A\vee B] = [A]\cup[B]$

The value of an entailment is the union of two sets: *N*, the class of normal worlds where the entailment is true, and *NN*, the of non-normal worlds where the entailment is true. Assuming the usual *S5* truth conditions, *N* and *NN* are specified thus:

$N = W$, if $[A] \subseteq [B]$; otherwise, $N = \varnothing$.

$NN = f([A],[B])$.

With all this in hand, validity is defined in the usual way: namely, as truth-preservation at all *normal* worlds of all interpretations.

Why restrict the definition merely to normal worlds? The explanation goes hand-in-hand with the informal interpretation of non-normal worlds; according to Priest's suggestion, non-normal worlds should be understood to be worlds where the laws of logic are different — different from the actual laws, where such laws are expressed by (true) entailment claims. Accordingly, since our definition of validity is an attempt to capture our (actual) logical laws, we need not, and should not, worry about worlds where the logical laws are different, at least not in our definition of validity. Such worlds, however, are otherwise very important; as one can easily verify, such worlds afford the usual logical laws (within the positive fragment at issue) but do not sanction the unwanted "laws" — e.g., Assertion and the like. In this way, one can enjoy naive truth theory (or naive set theory) without tripping into triviality as a result of Curry sentences.

Priest (1992) gives a sound and complete proof theory for the given semantics, but this is left for the reader to consult.

Open Issues and Problems

With the foregoing semantics one sneed not reject the existence of Curry sentences (which are difficult to reject when one's language is a natural language) or naive truth theory; however, there are various philosophical issues that need to be addressed, a few of which are canvassed below.

One philosophical issue confronting the given semantics is the very nature of such non-normal worlds. What are they? As intimated, Priest's suggestion is that they are simply (impossible) worlds where the laws of logic are different. But is there any reason, independent of Curry's paradox, to admit such worlds? Fortunately, the answer seems to be ‘yes’. One reason has to do with the common (natural language) reasoning involving counter-logicals, including, for example, sentences such as ‘If intuitionistic logic is correct, then double negation elimination is invalid’. Invoking non-normal worlds provides a simple way of modelling such sentences and the reasoning involving them.

Another objection also arises. Notice that, on the foregoing semantics, there are (non-normal) worlds where the law of simplification, i.e., $A \& B \rightarrow B$, is false; however, there is no world (normal or otherwise) at which we have a false *B* but true *A*&*B*. Likewise for all other worlds where the logical laws differ; the

worlds themselves, as it were, do not break the laws, even though the laws are false at such worlds. What explains this "lack of supervenience" at non-normal worlds? Priest himself offers no explanation, and the problem remains an open one. None the less, here is a suggestion (which has yet to be explored in print): What would it take for logical laws to fail? Most philosophers will agree that it is hard to imagine worlds in which there are events that contravene logical laws. My suggestion is that the only way for logical laws to fail is via arbitrary "fiat", as it were. No world (possible or otherwise) comprises events that refute, contravene, or otherwise show the actual logical laws to be false; what is required to falsify logical laws is mere arbitrariness; and such arbitrariness is precisely what one gets from the function, *f*. The suggestion, then, is simply this: For logical laws to fail at any world (and, hence, at non-normal worlds) one requires arbitrariness and thereby a lack of the supervenience at issue. Whether this suggestion solves the (philosophical) problem at hand is an (other) open problem.

There are other philosophical (and logical) problems that remain open. One of the most important recent papers discussing such problems is Restall's "Costing Non-Classical Solutions to Paradoxes of Self-Reference" (see Other Internet Resources). Restall shows that the sorts of non-classical approach discussed above must give up either transitivity of entailment, infinitary disjunction or distributive lattice logic (i.e., an infinitary disjunction operator distributing over finite conjunction); otherwise, as Restall shows, Curry's paradox arises immediately and triviality ensues. The importance of Restall's point lies not only in the formal constraints imposed on suitable non-classical approaches to Curry; its importance lies especially in the philosophical awkwardness imposed by such constraints. For example, one (formal) upshot of Restall's point is that, on a natural way of modelling propositions (e.g., in familiar world-semantics), some classes of propositions will not have disjunctions on the (given sort of) non-classical approach; the philosophical upshot (and important open problem) is that there is no known explanation for why such classes lack such a disjunction. (Needless to say, it is not a sufficient explanation to note that the presence of such a disjunction would otherwise generate triviality via Curry's paradox.)

The foregoing issues and open problems confront various non-classical approaches to paradox, problems that arise particularly sharply in the face of Curry's paradox. It should be understood, however, that such problems may remain pressing even for those who are firmly committed to classical approaches to paradox; for one might be interested not so much in *accepting* or *believing* such non-classical proposals but, rather, merely in using such proposals to model various naive but non-trivial theories — naive truth theory, naive set theory, naive denotation theory, etc.. One need not believe or accept such theories to have an interest in modeling them accurately. If one has such an interest, then the foregoing problems arising from Curry's paradox must be addressed. (See Slaney 1989, and the classic Meyer, Dunn, and Routley 1979, and also Restall 2000 for further discussion.)

Bibliography

Works Cited or Further Reading

- Barwise, J., and Etchemendy, J., 1984. *The Liar*, New York: Oxford University Press.
- Boolos, G., and Jeffrey, R., 1989. *Computability and Logic*, 3rd edition, New York: Cambridge

University Press.

- Burge, T., 1979. "Semantical Paradox", *Journal of Philosophy* 76:169-198.
- Curry, H., 1942, "The inconsistency of certain formal logics", *Journal of Symbolic Logic* 7, pp. 115-117.
- Frederickson, G., 1997, *Dissections: Plane and Fancy*, Cambridge: CUP.
- Gardner, M., 1956. *Mathematics, Magic and Mystery*, New York, Dover Publ.
- Gaifman, H., 1988. "Operational pointer semantics: Solution to self-referential puzzles I", in Vardi, M., ed, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, Morgan Kaufmann.
- Goldstein, L. 1986. "Epimenides and Curry". *Analysis* 463:117-121.
- Gupta and Belnap, 1993. *The Revision Theory of Truth*, Cambridge, MA: MIT Press.
- Kripke, S., 1975. "Outline of a theory of truth", *Journal of Philosophy* 72:690-716.
- Meyer, R. K., Routley, R. and Dunn, J.M., 1979, "Curry's paradox", *Analysis* 39, pp. 124- 128.
- Myhill, J., 1975. "Levels of Implication". In A. R. Anderson, R. C. Barcan-Marcus, annd R. M. Martin, editors, *The Logical Enterprise*, pp. 179-185. Yale Univ. Press.
- Myhill, J., 1984. "Paradoxes". *Synthese*, 60:129-143.
- Priest, G., 1987. *In Contradiction*, Martinus Nijhoff.
- Priest, G., 1992. "What is a non-normal world?", *Logique & Analyse* 139-40:291-302.
- Prior, A. N., 1955. "Curry's Paradox and 3-Valued Logic", *Australasian Journal of Philosophy* 33:177-82.
- Read, S., 2001, "Self-Reference and Validity Revisited", in *Medieval Formal Logic*, M. Yrjonsuuri (ed.), Kluwer 2001, pp. 183-96. (Preprint (in PDF) available online)
- Restall, G., 2000, *An Introduction to Substructural Logics*, Routledge. (online précis)
- Simmons, K., 1993. *Universaliity and the Liar*, New York: Cambridge University Press.
- Moh Skaw-Kwei. "Logical Paradoxes for Many-Valued Systems". *Journal of Symbolic Logic*, 19 (1954), pp. 37-39.
- Slaney, John. 1989. "RWX is not Curry Paraconsistent", in G. Priest, R. Routley, and J. Norman (eds.), *Paraconsistent Logic: Essays on the Inconsistent*, Philosophia Verlag, 472--480.

Other Internet Resources

- Restall, G., "Costing Non-Classical Solutions to Paradoxes of Self-Reference".
- Restall, G., 1994, *On Logics Without Contraction*, (Ph.D. dissertation, University of Queensland)

Related Entries

logic: paraconsistent | logic: relevance | logic: substructural | Russell's paradox

JC Beall
beall@uconn.edu

First published: January 10, 2001
Content last modified: January 10, 2001

Disjunction

Disjunction is a binary truth-function, the output of which is a sentence true if at least one of the input sentences (disjuncts) is true, and false otherwise. Disjunction, together with negation, provide sufficient means to define all other truth-functions. Its supposed connection with the *or* words of natural language has intrigued and mystified philosophers for many centuries, and the subject has inspired much creative myth-making, particularly since the advent of truth-tables early in the twentieth century. In this article some of those myths are set out and dispelled.

Introduction

A disjunction is a kind of compound sentence historically associated by English-speaking logicians and their students with indicative sentences compounded with *either* ... *or*, such as

> Either I am very rich or someone is playing a cruel joke.

But nowadays the term *disjunction* is more often used in reference to sentences (or well-formed formulae) of associated form occurring in formal languages. Logicians distinguish between

> (a) the abstracted *form* of such sentences and the roles that sentences of that form play in

arguments and proofs,

(b) the *meanings* that must be assigned to such sentences to account for those roles.

The former represents their *syntactic* and *proof-theoretic* interests, the latter their *semantic* or *truth-theoretic* interest in disjunction. Introductory logic texts are sometimes a little unclear as to which should provide the defining characteristics of disjunction. Nor are they clear as to whether disjunctions are primarily features of natural or of formal languages. Here we consider formal languages first.

Syntax

The definition of a formal system, either axiomatic or natural deductive, requires the definition of a language, and here the formal vocabulary of disjunction makes its first appearance. If the disjunctive constant $\vee$ (historically suggestive of Latin *vel* (*or*)) is a primitive constant of the language, there will be a clause, here labeled [$\vee$] in the inductive definition of the set of well-formed formulae (wffs). Using α and β as metalogical variables, ranging over wffs, such a clause would read:

[$\vee$] If α is a wff and β is a wff, then $\alpha \vee \beta$ is a wff

perhaps accompanied by an instruction that $\alpha \vee \beta$ is to be referred to as the *disjunction* of the wffs α and β, and read as "[name of first wff] vel (or 'vee', or 'or') [name of second wff]". Thus, on this instruction, the wff $p \vee q$ is the *disjunction* of p and q, and is pronounced as 'pea vel queue' or 'pea vee queue' or 'pea or queue'. In this case, p and q are the disjuncts of the disjunction.

If $\vee$ is a non-primitive constant of the language, then typically it will be introduced by an abbreviative definition. In presentations of classical systems in which the conditional constant $\rightarrow$ or $\supset$ and the negational constant $\neg$ are taken as primitive, the disjunctive constant $\vee$ might be introduced in the abbreviation of a wff $\neg\alpha \rightarrow \beta$ (or $\neg\alpha \supset \beta$) as $\alpha \vee \beta$. Alternatively, if the conjunctive & has already been introduced either as a primitive or as a defined constant, $\vee$ might be introduced in the abbreviation of a wff $\neg(\neg\alpha \,\&\, \neg\beta)$ as $\alpha \vee \beta$.

Proof Theory

Much as we would understand the conversational significance of vocabulary more generally if we had a complete set of instructions for initiating its use in a conversation, and for suitable responses to its introduction by an interlocutor, we give the proof-theoretic significance of a connective by providing rules for its introduction into a proof and for its elimination. In the case of $\vee$, these might be the following:

[$\vee$-introduction] For any wffs α and β, a proof having a subproof of α from an ensemble

Σ of wffs, can be extended to a proof of $\alpha \vee \beta$ from Σ.

[$\vee$-elimination] For any wffs α, β, γ, a proof that includes

- a subproof of $\alpha \vee \beta$ from an ensemble of wffs Σ,
- a subproof of γ from an ensemble $\Delta \cup \{\alpha\}$, and
- a subproof of γ from an ensemble $\Theta \cup \{\beta\}$,

can be extended to a proof of γ from $\Sigma \cup \Delta \cup \Theta$.

Intuitively, the former would correspond to a rule of conversation that permitted us to assert *A* or *B* (for any *B*) given the assertion that *A*. Thus if we are told that Nicholas is in Paris, we can infer that Nicholas is either in Paris or in Toulouse.

Intuitively, the latter rule would correspond to a rule that, given the assertion that *A* or *B*, would permit the assertion of anything that is permitted both by the assertion of *A* and by the assertion of *B*. For example, given the assertion on certain grounds that Nicholas is in Paris or Toulouse, we are warranted in asserting on the same grounds plus some geographical information, that Nicholas is in France, since that assertion is warranted (a) by the assertion that Nicholas is in Paris together with some of the geographical information and (b) by the assertion that Nicholas is in Toulouse together with the rest of the geographical information. More generally we may sum the matter up by saying that the rule corresponds to the conversational rule that lets us extract information from an *or*-sentence without the information of either of its clauses. In the example, we are given the information that Nicholas is in Paris or Toulouse, but we are given neither the information that Nicholas is in Paris nor the information that he is in Toulouse.

Semantics

In its simplest, classical, semantic analysis, a disjunction is understood by reference to the conditions under which it is true, and under which it is false. Central to the definition is a *valuation*, a function that assigns to every atomic, or unanalysable sentence of the language a value in the set {1, 0}. In general the inductive truth-definition for a language corresponds, clause by clause to the definition of its well-formed formulae. Thus for a propositional language it will take as its basis, a clause according to which an atom is true or false accordingly as the valuation maps it to 1 or to 0. In systems in which $\vee$ is a primitive constant, the clause corresponding to disjunction takes $\alpha \vee \beta$ to be true if at least one of α, β is true, and takes it to be false otherwise. Where $\vee$ is introduced by either of the definitions earlier mentioned, that truth-condition can be computed for $\alpha \vee \beta$ from those of the conditional ($\rightarrow$ or $\supset$) or conjunction (&) and negation ($\neg$).

Now the truth-definition can be regarded as an extension of the valuation from the atoms of the language to the entire set of wffs with 1 understood as the truth-value, true, and 0 understood as the truth-value,

false. Thus, classically, disjunction is semantically interpreted as a binary truth-function from the set of pairs of truth-values to the set {0, 1}. The tabulated graph of this function, as dictated by the truth-definition, is called the truth-table for disjunction. That table is the following:

α β	$\alpha \vee \beta$
1 1 1 0 0 1 0 0	1 1 1 0

Inclusive and Exclusive disjunctions

Authors of introductory logic texts generally take this opportunity to distinguish the disjunction we have been discussing from another binary truth-function $\veebar$ whose graph is tabulated by the table:

α β	$\alpha \veebar \beta$
1 1 1 0 0 1 0 0	0 1 1 0

where $\alpha \veebar \beta$ is read α *xor* β. This truth-function is referred to variously as exclusive disjunction, as 0110 disjunction (after the succession of values in its main column), and as logical difference. The wff $\alpha \veebar \beta$ is true when exactly one of α, β is true; false otherwise. To make matters explicit, the earlier discussed truth-function $\vee$ is called inclusive, or non-exclusive or 1110 disjunction.

Natural Language

It is an assumption, at any rate a claim, of many textbook authors that there are both uses of *or* in English that correspond to 1110 disjunction and uses that correspond to 0110 disjunction, and this supposition generally motivates the introduction and discussion of the xor connective. Since we are following the usual order of textbook exposition, this is perhaps the moment to make a few observations on this score. The first are purely syntactic. The *or* of English that such authors cite is a coordinator (or coordinating conjunction). It can coordinate syntactic elements of virtually any grammatical type, not merely whole sentences. Moreover, if we consider only its uses joining whole sentences, we must notice that it can join sentences of virtually any mood: interrogative sentences and imperatives as well as indicative sentences can be joined by *or* in English. And again, if we restrict our attention to its uses joining indicative sentences, we must note that *or* is by no means restricted to the binary cases in this role. Indeed, there is

no theoretical finite limit to the number of clauses that it can join. This is perhaps the most fundamental relevant syntactic difference between *or* on the one hand and $\vee$ and $\veebar$ on the other. The sentence

> Nathalie has been and gone or Nathalie will arrive today or Nathalie will not arrive at all

is a perfectly correct sentence and not ambiguous as between

> (Nathalie has been and gone or Nathalie will arrive today) or Nathalie will not arrive at all

and

> Nathalie has been and gone or (Nathalie will arrive today or Nathalie will not arrive at all).

By contrast, the wff $p \vee q \vee r$, far from being ambiguous as between $(p \vee q) \vee r$ and $p \vee (q \vee r)$, is, on the inductive definition of well-formedness, not a wff. If the parenthesis-free notation is tolerated in general logical exposition, this is because $\vee$ is *associative*, that is, the wffs $(p \vee q) \vee r$ and $p \vee (q \vee r)$ are syntactically interderivable, and semantically have identical truth-conditions. The formal account of disjunction could readily be liberalized to accommodate that fact, and even conveniently in languages in which $\vee$ was primitive. In that case our inductive definition of the language could permit any such string as $\vee(\alpha_1, \ldots, \alpha_i, \ldots, \alpha_n)$ to be well-formed if $\alpha_1, \ldots, \alpha_i, \ldots$ and α_n are. The relevant clause of the truth-definition would accordingly be modified in such a way as to give $\vee(\alpha_1, \ldots, \alpha_i, \ldots, \alpha_n)$ the maximum of the truth-values of $\alpha_1, \ldots, \alpha_i, \ldots$ and α_n. Moreover, this accords well with such cases as the one cited in which *or* joins more than two simple clauses: such a sentence is true if at least one of its clauses is true; false otherwise.

The fact that English *or* is not binary does not accord so well with the claim made by many textbook authors that there are uses of *or* that require representation by 0110 disjunction. To be sure, $\veebar$ is associative, so that a notational liberalization would be possible, parallel to the one described for $\vee$. But, as Hans Reichenbach seems first to have pointed out (in Reichenbach [1947]), the truth-definition for $\veebar(\alpha_1, \ldots, \alpha_i, \ldots, \alpha_n)$ would have to be such as to give it the value 1 if any odd number of $\alpha_1, \ldots, \alpha_i, \ldots, \alpha_n$ have the value 1; the value 0 otherwise. The result is evident from the truth-table where $n > 2$. For $n = 3$, suppose that $\alpha \veebar \beta \veebar \gamma$ has the value 1. The truth-definition as given by the table requires that exactly one of $\alpha \veebar \beta, \gamma$ has the value 1. Let γ have the value 1; then $\alpha \veebar \beta$ has the value 0. Then α and β have the same value. That is, either both α and β have the value 0, or both α and β have the value 1. In the former case exactly one of α, β, γ has the value 1; in the latter, all three have the value 1. That is, the disjunction will take the value 1 if and only if an odd number of disjuncts have the value 1. A simple induction will prove that this result holds for an exclusive disjunction of any finite length. It is sufficient for present purposes to note that, in the case where $n = 3$, $\veebar(\alpha_1, \alpha_2, \alpha_3)$ will be true if all of its disjuncts are true. Now there is no naturally occurring coordinator in any natural language matching the truth-conditional profile of such a connective. There is certainly no use of *or* in English in accordance with which five sentences *A*, *B*, *C*, *D*, and *E* can be joined to form a sentence *A* or *B* or *C* or *D* or *E*, which is

true if and only if either exactly one of the component sentences is true, or exactly three of them are true or exactly five of them are true.

Most of the texts make no claims about exclusive disjunctive uses of either English or Latin *or*-words beyond the two-disjunct case. But it is a fair presumption that the belief in exclusive disjunctive uses of *or* in English includes just such three-disjunct uses of *or*. Such a use of *or*, would be one in accordance with which three sentences *A*, *B*, and *C* can be joined to form a sentence *A* or *B* or *C*, which is true if and only if exactly one of the component sentences is true. Though not a 0110-disjunctive use of *or*, this would be a general use representable as 0110 disjunction in the two-disjunct case.

The question as to whether there is such a use of *or* in English, or any other natural language goes to the very heart of the conception of truth conditional semantics. For it seems certain that there are conversational uses of *or* that invite the inference of exclusivity, but which do not seem to require exclusivity for their truth. Thus, for example, if one says (as in Tarski [1941], 21) 'We are going on a hike or we are going to a theater', even with charged emphasis upon the *or*, one will have spoken falsely if in the event we do both, unless, as in Tarski's example, one has also denied the conjunction.

Some authors have sought examples of 0110 disjunction in *or*-sentences whose clauses are mutually exclusive. For example, Kegley and Kegley discuss the case (Kegley and Kegley [1978], 232):

> John is at the play, or he is studying in the library

of which the authors remark, "There is no mistaking the sense of *or* here: John cannot be in both places at once". If their example were an example of exclusive disjunction, we could safely infer from it that the play is not being performed in the library, that the theatre is not in the library, that John is not swotting in the stalls between acts while his companion fights her way to the bar to fetch the drinks. In fact, even, perhaps particularly, when the disjuncts are genuinely mutually exclusive, there are no grounds for the supposition that the *or* represents 0110 disjunction. Were there such grounds the $\vee$ of formal logic would require distinct semantic accounts for the wffs $p \vee q$ and $p \vee \neg p$. As Barrett and Stenner point out (Barrett and Stenner [1971]), the case requires quite the reverse. Since the truth-tables of $\vee$ and $\veebar$ differ exactly in the output value of the first row, what alone would clinch the case for the existence of an exclusive *or* would be a sentence in which both disjuncts were true, and the disjunction therefore false. No author has yet produced such an example.

The Myth of *Vel* and *Aut*

If the logic texts dictate the structure and content of our discussion, it is perhaps as well to dispel another current myth -- namely that the notational choice of $\vee$, (read as *vel*) as the connective of inclusive disjunction, and the claim that the English *or* has 0110-disjunctive uses are supported by the facts of the Latin language. I. Copi is as explicit as any (Copi [1971], 241):

> The Latin word "vel" expresses weak or inclusive disjunction, and the Latin word "aut" corresponds to the word "or" in its strong or exclusive sense.

The idea is, first, that whereas English has only one *or*-word, Latin has two: *vel* and *aut*, and secondly, that the uses of *vel* in Latin would be representable as 1110 disjunction and the uses of *aut* as 0110 disjunction. As to the first, the very shape of the claim is likely to mislead. The case is not that Latin had two words for *or*, but rather that Latin had more than one word that gets translated into English as *or*. In fact, Latin had *many* words that are translated into English as *or*, including, besides the two listed, at least *seu*, *sive* and the enclitic *ve*. So does English have many words that can be translated into English as *or*, including *unless*, *if ... not*, *but* (It does not rain but it pours) and so on. All vocabulary has a history, and languages accumulate vocabulary that becomes adapted to nuanced uses.

Now the supposition that Latin had a 0110 coordinator must suffer from the same implausibilities as the corresponding supposition about English. What of the two-disjunct case? If any general tendency can be detected in actual Latin usage, say in the classical period, that would distinguish the uses of *vel* from those of *aut*, it is that *aut* tended to be brought into use in the formation of lists of disjoint or contrasted or opposed items, categories or classes or states, as for example

> *Omne enuntiatum aut verum aut falsum est* [Every statement is either true or false] (Cicero, *De Fato*, 222).

The difficulty with these examples is that the exclusiveness of the states independendently of the choice of connective must mask any disjointness that the connective could itself impose. That it does not impose *any* disjointness itself is best seen in its list-forming uses. Consider the list (Cicero, *De Officiis*):

> *tribunos aut plebes* [the magistrates or the mob, (*accusative plural*)]

to be sure the categories are disjoint, and this fact might be supposed to contribute to the selection of *aut*. But the mutual exclusion in such cases need not survive the addition of a verb.

> *Timebat tribunos aut plebes* [one feared the magistrates or the mob]

does not exclude the case in which one feared both. However, what clinches the refutation of this mythical supposition is that if that whole clause is brought within the scope of a negator, the resulting sentence will expect a reading along the lines of 1110 disjunction.

> *Nemo timebat tribunos aut plebes* [No one feared the magistrates or the mob]

just means no one feared either. It does not mean everyone either feared neither or feared both. Since the negation of a 0110 disjunction is a 1110 disjunction (either both disjuncts are true or both disjuncts are false), this use of *aut* cannot be a 0110 disjunctive use.

In fact, in classical Latin, *aut* was favoured over *vel* in constructions involving negations, and in that use, *aut* behaves analogously to ⊻. But pretty well anywhere an *aut* could be used, a *vel* could be substituted, and vice versa. The resulting sentence would have a different flavour, and in some instances would be mildly eccentric, but would not have a different truth condition. The uses of *vel* reflected its origins as an imperative form of *volo*. The flavour of

> *Nemo timebat vel tribunos vel plebes*

would be closer to that of

> Name which group (of the two) you will: no one feared them.

Aut was adversative: no one feared either social extremity. (For more examples and a more detailed discussion, see Jennings [1994], 239-251.)

The *Or* of Natural Language

There are undoubtedly disjunctive uses of *or* in English, and of corresponding vocabulary in other natural languages. But the uses of *or* after the pattern of the logic texts:

> Either Argentina will boycott the conference or the value of lead will diminish

and so on constitute only a very small proportion, certainly fewer than 5% of the occurrences of *or* in English, and, it can be supposed, of corresponding words in all other natural languages as well. It is therefore not surprising that it should be some of these non-disjunctive uses that have been misidentified as instances of exclusive disjunction. The example cited (in Richards [1978], 84) is a good representative example of one such common misidentification:

> So how can we find a clear-cut case of the exclusive ‘or’? Imagine a boy who asks for ice cream *and* strawberries for tea. He is told as a sort of refusal:
>
> > ’You can have ice cream *or* strawberries for tea’.
>
> Here there is no doubt: not both may be had.

Once again there is a difficulty in trying to account for the exclusivity by reference to truth-conditions, though, if we are permitted to consult the intentions of the speaker (as Richards himself does) we may be in no doubt as to the prohibition of strawberries and icecream, however curious such a prohibition might seem. But this example, in company with the many others like it (which this author has sometimes referred to collectively as *the argument from confection*) suffers from the even more serious flaw that it

is not a disjunction at all. The problem is not that the *or* does not join whole clauses. Even if we expand the example to

> ’You can have ice cream for tea *or* you can have strawberries for tea’,

the sentence cannot be construed as a disjunction. The reason is that the child would be correct in inferring that he can have ice cream for tea, and would be correct in inferring that he can have strawberries for tea. Such sentences are elliptical for conjunctions, not for disjunctions, even on a truth-conditional construal. It just happens that for such conjunctions, questions of exclusivity, or rather non-combinativity also arise.

Not every *or* of English (nor every counterpart of *or* in other languages) is disjunctive, even among those that join pairs of indicative sentences.

Bibliography

- Barrett, Robert B., and Stenner, Alfred J., "The Myth of the Exclusive ‘Or’", *Mind*, **80 (317)**, 1971, 116-121.
- Cicero, Marcus Tullius, *De Fato*, Translated by H. Rackham. Cambridge, Mass 1942.
- Cicero, Marcus Tullius, *De Officiis*, Translated by Walter Miller. Cambridge, Mass 1975.
- Copi, I.M., *Introduction to Logic*, New York, 1971.
- Jennings, R.E., *The Genealogy of Disjunction*, New York: Oxford University Press, 1994.
- Kegley, Charles W., and Kegley, Jacquelyn Ann, *Introduction to Logic*, Lanham, Md., 1978.
- Reichenbach, Hans, *Elements of Symbolic Logic*, New York: MacMillan, 1947.
- Richards, T.A., *The Language of Reason*, Rushcutters Bay, NSW, 1978.
- Alfred Tarski, *Introduction to Logic and to the Methodology of the Deductive Sciences*, New York 1941, (Revised 1946 edition).

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

connectives

R.E. Jennings
jennings@sfu.ca

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 5, 2001
Content last modified: January 5, 2001

Legal Punishment

The question of whether, and how, legal punishment can be justified is central to both legal and political philosophy: what could justify a state in using the apparatus of the law to inflict burdensome sanctions on its citizens? Radically different answers to this question are offered by consequentialist and by retributivist theorists -- and by those who seek to combine consequentialist with retributivist considerations in ‘mixed’ theories of punishment; an important strand in recent theorising has been the idea of punishment as a communicative enterprise. Meanwhile, abolitionist theorists argue that we should aim to replace legal punishment rather than to justify it.

1. Legal Punishment and its Justification

The central question asked by philosophers of punishment is: What can justify punishment? More precisely, since they do not usually talk much about punishment in such contexts as the family or the workplace, their question is: What can justify formal, legal punishment imposed by the state on criminals? We will also focus on legal punishment here: not because the other species of punishment do not raise important normative questions (they do), nor because such questions can be answered in terms of an initial justification of legal punishment as being the paradigm case (since it's not clear that they can be), but because legal punishment, apart from being more dramatically coercive and burdensome than other species of punishment usually are, raises distinctive issues about the role of the state and its

relationship to its citizens, and about the role of the criminal law. Future references to ‘punishment’ should therefore be read, unless otherwise specified, as references to legal or criminal punishment.

What then are we to justify in justifying punishment? The search for a precise definition of punishment that exercised some philosophers (for discussion and references see Scheid 1980) is likely to prove futile: but we can say that legal punishment involves the imposition of something that is intended to be burdensome or painful, on a supposed offender for a supposed crime, by a person or body who claims the authority to do so. How can such a practice, which infringes the freedom of those subjected to it, which not only causes but aims to cause them suffering, be justified?

We should not assume, however, that there is only one question of justification, which can receive just one answer. As Hart famously pointed out (Hart 1968: 1-27), we must distinguish at least three justificatory issues. First, what is the ‘general justifying aim’ of a system of punishment: what justifies the creation and maintenance of such a system -- what good does it achieve, what duty does it fulfil? Second, who may properly be punished: what principles or aims should determine the allocations of punishments to individuals? Third, how should the appropriate amount of punishment be determined: how should sentencers go about deciding how severe a sentence they should impose? We can add a fourth issue, which is insufficiently discussed by philosophers: what concrete modes of punishment are appropriate, in general or for particular crimes? It might of course turn out that answers to all these questions will flow from a single theoretical foundation -- for instance from a unitary consequentialist principle specifying the good that punishment should achieve, or from some version of the retributivist principle that the sole proper aim of punishment is to impose on the guilty the suffering they deserve. But, in this as in other matters of normative political theory, matters might not be as easy and simple as that: we might find that quite different and conflicting values are relevant to different issues about punishment; and that any complete normative account of punishment will have to find a place for these values -- and to help us find some no doubt uncomfortable compromises between them when they conflict.

Even this way of putting the matter oversimplifies it, by implying that we can hope to find a ‘complete normative account of punishment’: an account, that is, of how punishment can be justified. That is certainly an implicit assumption of much philosophical and legal discussion: punishment can -- of course -- be justified, and the theorists' task is to establish and explicate that justification. But it is an illegitimate assumption: normative theorists must be open to the possibility, startling and disturbing as it might be, that this pervasive human practice cannot be justified. Nor is this merely the kind of fantastical scepticism that moral philosophers are sometimes prone to imagine (‘suppose someone denied that killing for pleasure was wrong’): there is a significant strand of ‘abolitionist’ penal theorising (to which insufficient attention is paid in the philosophical literature) which argues precisely that legal punishment cannot be justified and should be abolished. The abolitionist claim is not merely that our existing penal practices are not justified: viewed in the light of many normative penal theories (one might almost say, of any plausible normative penal theory) our existing penal practices, especially those involving imprisonment, are not merely imperfect, but so radically inconsistent with the values that should inform a practice of punishment that they cannot claim to be justified. For those who think that punishment can in principle be justified, this means simply (and hardly surprisingly) that our penal practices need radical reform if they are to become justified: but the abolitionist critique goes much deeper than that, to argue that legal

punishment cannot be justified even in principle (see e.g. Christie 1977, 1981; Hulsman 1986, 1991; Bianchi & van Swaaningen 1986; de Haan 1990; Bianchi 1994. For a critical survey see Duff 1996: 67-87).

We will attend to some abolitionist arguments in what follows. Even if those arguments can be met, even if legal punishment can be justified, at least in principle, the abolitionist challenge is one that must be met, rather than ignored; and it will help to remind us of the ways in which any practice of legal punishment is bound to be morally problematic.

2. Punishment, Crime and the State

Legal punishment presupposes crime as that for which punishment is imposed, and a criminal law as that which defines crimes as crimes; and a system of criminal law presupposes a state, which has the political authority to make and enforce the law and to impose punishments. A normative account of legal punishment and its justification must thus at least presuppose, and should perhaps make explicit, a normative account of the criminal law (why should we have a criminal law at all?) and of the proper powers and functions of the state (by what authority or right does the state make and declare law, and impose punishments on those who break it?).

How far it matters, in this context, to make explicit a political theory of the state depends on how far different plausible political theories will generate very different accounts of how punishment can be justified and should be used. We cannot pursue this question here (for two sharply contrasting views on it, see Philips 1986, Davis 1989), but should look briefly at the concept of crime, since that is one focus of the abolitionist critique of punishment.

On a simple positivist view of law, crimes are kinds of conduct that are prohibited, on pain of threatened sanctions, by the law; and for positivists like Bentham, who combine positivism with a normative consequentialism, the questions of whether we should maintain a criminal law at all, and of what kinds of conduct should be criminalised, are to be answered by trying to determine whether and when this method of controlling human conduct is likely to produce a net increase in good. Such a perspective seems inadequate, however: inadequate both to the claims of the criminal law, which presents its demands as something other or more than those of a gunman writ large -- as something other or more than ‘Behave thus, or else!’; and to the normative issues at stake when we ask what kinds of conduct should be criminalised. For the criminal law portrays crime not merely as conduct which has been prohibited, but as a species of wrongdoing: whether our inquiry is analytical (into the concept of crime) or normative (as to what kinds of conduct, if any, should be criminal), we must therefore focus on that notion of wrongdoing.

Crimes are, at least, socially proscribed wrongs -- kinds of conduct which are condemned as wrong by some purportedly authoritative social norm. That is to say that they are wrongs which are not merely ‘private’ affairs, which properly concern only those directly involved in them: the community as a whole -- in this case the political community speaking through the law -- claims the right to declare them to be wrongs. But crimes are ‘public’ wrongs in a sense that goes beyond this. Civil law deals in part with

wrongs which are non-private in that they are legally and socially declared as wrongs -- with the wrong constituted by libel, for instance: but they are still treated as ‘private’ wrongs in the sense that it is up to the person who was wronged to seek legal redress. She must decide to bring, or not to bring, a civil case against the person who wronged her; and although she can appeal to the law to protect her rights, the case is still between her and the defendant. By contrast, a criminal case is between the whole political community -- the state or the people -- and the defendant (or, in politically backward polities such as Britain, between the monarch and the defendant): the wrong is ‘public’ in the sense that it is one for which the wrongdoer must answer not just to the individual victim, but to the whole polity through its criminal courts.

It is notoriously difficult to give a clear and plausible account of the distinction between civil and criminal law, between ‘private’ and ‘public’ legal wrongs, whether our interest is in the analytical question of what the distinction amounts to, or in the normative question of which kinds of wrong should fall into which category (see Murphy & Coleman 1984, ch. 3; a symposium in Boston University Law Review vol. 76 (1996): 1-373). It might be tempting to say that crimes are ‘public’ wrongs in the sense that they injure the whole community: they threaten social order, for instance, or cause ‘social volatility’ (Becker 1974); or they involve taking unfair advantage over those who obey the law (Murphy 1973); or they undermine the trust on which social life depends (Dimock 1997). But such accounts distract our attention from the wrongs done to the individual victims that most crimes have, when it is those wrongs that should be our central concern: we should condemn the rapist or murderer, we should see the wrong he has done as our concern, because of what he has done to his victim. Another suggestion is that ‘public’ wrongs are those which flout the community's essential or most basic values, in which all members of the community should see themselves as sharing: the wrong is done to ‘us’, not merely to its individual victim, in the sense that we identify ourselves with the victim as a fellow citizen (see Marshall & Duff 1998).

Abolitionists, however, sometimes argue that we should seek to eliminate the concept of crime from our social vocabulary: we should talk and think not of ‘crimes’, but of ‘conflicts’ or ‘troubles’ (Christie 1977; Hulsman 1986). One motivation for this might be the thought that ‘crime’ entails punishment as the appropriate response: but that is not so, since we could imagine a system of criminal law without punishment. To define something as a ‘crime’ does indeed imply that some kind of public response is appropriate, since it is to define it as a kind of wrong that properly concerns the whole community: but that public response could consist in nothing more than, for instance, some version of a criminal trial which calls the alleged wrongdoer to answer for her alleged wrongdoing, and condemns her for it, through a criminal conviction, if she is proved guilty. One can of course count a criminal conviction as a kind of punishment: but it does not entail the kind of punishment, imposed after conviction, with which penal theorists are primarily concerned.

Another possible motivation for the abolitionist objection to the concept of crime is a kind of moral relativism which objects to the ‘imposition’ of values on those who might not share them (Bianchi 1994: 71-97): but since abolitionists are very ready to tell us, insistently, how we ought to respond to conflicts or troubles, and how a state ought or ought not to treat its citizens, such an appeal to relativism reflects serious confusion (see Williams 1976: 34-39).

Another abolitionist concern is that by defining and treating conduct as 'criminal', the law 'steals' the conflicts which crime involves from those to whom they properly belong (Christie 1977): instead of allowing, and helping, those who find themselves in conflict to resolve their trouble, the law takes the matter over and translates it into the professionalised context of the criminal justice system, in which neither 'victim' nor 'offender' is allowed any appropriate or productive role. We will look later at the suggestion, which arises from this kind of view, that our response to crimes should consist not in punishment, but in a process of mediation or 'restoration' between victim and offender: all we need note here is that at least sometimes it is important to insist that the situation involves not just people in 'conflict', but a victim who has been wronged and an offender who has done the wrong. Faced, for instance, by feuding neighbours who persistently accuse each other of more or less trivial wrongs, it might indeed be appropriate to suggest that they should forget about condemning each other and look for a way of resolving their conflict. But faced by a rapist and the person he raped, or by a violent husband and the wife he has been beating up, it would a betrayal both of the victim and of the values to which we are supposedly committed to portray the situation merely as a 'conflict' which the parties should seek to resolve: whatever else or more we can do, we must recognise and declare that here is a victim who has been seriously wronged; and we must be ready to censure the offender's action as a wrong.

However, to argue that we should retain the concept of crime, that we should maintain a criminal law which defines and condemns a category of 'public' wrongs, is not yet to say that we should maintain a penal system which punishes those who commit such wrongs; as I have noted, while a system of criminal law might require something like a system of criminal trials which will authoritatively identify and condemn criminal wrongdoers, it does not of its nature require the imposition of further sanctions on such wrongdoers. So we must turn at last to the question of what could justify such a system of punishment.

3. Pure Consequentialism and Punishment

Many people, including those who do not take a consequentialist view of other matters, think that any adequate justification of punishment must be basically consequentialist. For we have here a practice which inflicts, indeed seeks to inflict, significant hardship or pain: how else could we hope to justify it than by showing that it brings consequential benefits sufficiently large to outweigh, and thus to justify, that hardship and pain? However, when we try to flesh out that simple consequentialist thought into something closer to a full normative account of punishment, problems begin to appear.

A consequentialist must justify punishment (if she is to justify it at all) as a cost-effective means to certain independently identifiable goods. Whatever account she gives of the final good or goods at which all action ultimately aims, the most plausible immediate good that a system of punishment can bring is the prevention of crime: a rational consequentialist system of law will define as criminal only conduct that is in some way harmful; in preventing crime we will thus be preventing the harms that crime causes; and punishment can prevent crime by incapacitating, or deterring, or reforming potential offenders.

It is a contingent question whether punishment can be an efficient method of preventing crime in any of these ways, and some objections to punishment rest on the empirical claim that it cannot be -- that there

are other and more efficient methods of crime-prevention, for instance those involving a therapeutic approach to offenders (see Wootton 1963; Menninger 1968). Our focus here, however, must be on the moral objections to consequentialist accounts of punishment -- on objections to the effect that crime-preventive efficiency does not suffice to justify a system of punishment.

The most familiar objections to consequentialist penal theories are objections to purely consequentialist theories which hold that only the consequences are relevant to question of justification (for two simple examples of such theories, see Wilson 1983; Walker 1991). For, critics argue, it could turn out that manifestly unjust punishments (the punishment of those known to be innocent, for instance, or excessively harsh punishment of the guilty) would efficiently serve the aim of crime prevention, and consequentialists must then regard such punishments as in principle justified: but they would be wrong, just because they would be unjust (see e.g. Hart 1968, chs. 1-2; Ten 1987; Primoratz 1999, chs. 2-3).

There are some equally familiar consequentialist responses to this familiar objection. One is to argue that such ‘unjust’ punishments would be justified if they would really produce the best consequences (see e.g. Smart 1973: 69-72; Bagaric & Amarasekara 2000) -- to which the critic will reply that we cannot thus put aside the moral significance of injustice. Another is to argue that in the real world it is extremely unlikely that such punishments would ever be for the best, and even less likely that the agents involved could be trusted reliably to pick out those rare cases in which they would be: thus we, and especially our penal officials, will do best if we think and act as if such punishments are intrinsically wrong and unjustifiable (see e.g. Rawls 1955; Hare 1981, chs. 3, 9.7) -- to which the critic will respond that this still makes the wrongness of punishing a known innocent contingent on its effects, and fails to recognise the intrinsic wrong that such punishment does (see e.g. Duff 1986: 151-64; Primoratz 1999, chs. 3.3, 6.5). Another response is to argue that a richer or subtler account of the ends that the criminal law should serve will generate suitable protection against unjust punishments (see Braithwaite & Pettit 1990, especially 71-6, on ‘dominion’ as the end of criminal law): but the objection remains that any purely consequentialist account will make the protection of the innocent against injustice contingent on its instrumental contribution to the system's aims (on Braithwaite & Pettit, see von Hirsch & Ashworth 1992; Duff 1996: 20-25; Pettit 1997).

4. Side-Constrained Consequentialism and Punishment

The most familiar response to such objections is to abandon pure consequentialism, in favour of a side-constrained consequentialism: to insist that the positive justification of any system of punishment -- its ‘general justifying aim’ (Hart 1968: 8-11) -- lies in its beneficial effects, but to argue that our pursuit of that aim is subject to non-consequentialist constraints which forbid, for instance, the deliberate punishment of the innocent, or the excessively harsh punishment of the guilty. (See most famously Hart 1968, and Scheid 1997 for a sophisticated Hartian theory; on Hart see Lacey 1988: 46-56; Morison 1988; Primoratz 1999, ch. 6.6.)

One question about such accounts concerns the grounding of these side-constraints. If they are derived from a ‘negative’ retributivism which insists that punishment is justified only if it is deserved (see Dolinko 1991: 539-43), they face the problem of explaining this retributivist notion of desert (see s. 4 below): but it is not clear whether they can be justified without such an appeal to retributivist desert (see Hart 1968: 44-48; Walker 1991, ch. 11; Feinberg 1988: 144-55). Even if such side-constraints can be securely grounded, however, consequentialist theories of punishment face further objections, focused on the moral character of punishment within those constraints. For the side-constrained consequentialist, so long as punishment is deserved it may and should be used to serve consequentialist ends -- most obviously the end of crime prevention: but, the critic now objects, to use punishment thus is to use those who are punished ‘merely as means’ to those further ends, which is to deny them the respect, the moral standing, that is their due as responsible agents (see Murphy 1973: 218). Objections to purely consequentialist theories often focused on the rights of the innocent: the objections to side-constrained consequentialism, which clearly protects the rights of innocent, focus on the rights or moral standing of the guilty.

The Kantian prohibition on treating each other ‘merely as means’ is admittedly unclear in its implications; and it can be argued that if punishment is reserved for those who voluntarily break the law, it does not treat them merely as means (see Walker 1980: 80-85). But a version of it does seem to have force against purely reformative punishments that aim simply so to modify offenders' dispositions that they will in future willingly obey the law; against purely incapacitative punishments that aim simply to prevent offenders from committing further crimes; and against purely deterrent punishments that aim simply to give potential offenders prudential reason to obey the law. For, the Kantian can argue, if we are to treat another ‘as an end’, with the respect due to her as a rational and responsible agent, we must seek to modify her conduct only by offering her good and relevant reasons to modify it for herself. These modes of purely consequentialist punishment, however, do not satisfy that demand. A purely reformative system treats those subjected to it not as rational, self-determining agents, but as objects to be re-formed by whatever efficient (and humane) techniques we can find. A purely incapacitative system does not leave those subjected to it free, as responsible agents should be left free, to determine their own future conduct, but seeks to pre-empt their future choices by incapacitating them. And although a purely deterrent system does, unlike the others, offer potential offenders reason to obey the law, it offers them the wrong kind of reason: instead of addressing them as responsible moral agents, in terms of the moral reasons which justify the law's demands on them, it addresses them as merely self-interested beings, in the coercive language of threat; deterrence treats ‘a man like a dog instead of with the freedom and respect due to him as a man’ (Hegel 1821: 246. For these objections see Lewis 1953; H Morris 1968; von Hirsch 1993: 9-14; von Hirsch & Ashworth 1998, chs. 1, 3).

Such objections leave many people unpersuaded -- in particular the Hegelian objection to a side-constrained system of deterrent punishments that punishes only the guilty. One response is to argue that those who voluntarily break the law thereby forfeit at least some of the rights that citizens can normally claim: their wrongdoing legitimises kinds of treatment (reformative or incapacitative treatment, for instance) that would normally be wrong as right-violating (see Goldman 1982; C Morris 1991). Another response is to argue that a side-constrained system of deterrent punishments can be shown to be consistent with a proper respect for those who are punished, or threatened with punishment, by portraying it as a

species of societal (self-) defence (for versions of this kind of argument see Alexander 1980; Quinn 1985; Farrell 1985, 1995; Montague 1995); or by pointing out that it offers self-interested agents who are deaf to the law's moral appeal prudential reasons which they can grasp and see as relevant (Baker 1992). (On these arguments see Duff 2000, chs. 1.3, 3.1-3.) Even those who find the Hegelian objection to deterrent punishment persuasive can recognise such an account of punishment as a system of side-constrained deterrence offering those who are unmoved by the moral reasons for refraining from crime prudential reason to do so as the most plausible form that a consequentialist (but not purely consequentialist) theory of legal punishment can take.

However, we have already noted that a side-constrained consequentialism might need to appeal to a negative retributivist notion of desert to ground its side-constraints; we must now examine some of the ways in which that notion of penal desert has been used to ground, not merely negative side-constraints on the pursuit of some consequentialist end, but a positive justification of punishment.

5. ‘Positive’ Retributivism and the Meaning of Desert

‘Positive’ retributivism holds not merely that we must not punish the innocent (or punish the guilty more than they deserve), but that we should punish the guilty (to the extent that they deserve): penal desert constitutes not just a necessary, but a sufficient reason for punishment, or at least a strong positive reason for it. A striking feature of penal theorising during the last three decades of the twentieth century was a revival of positive retributivism -- of the idea that the positive justification of punishment is to be found in its intrinsic character as a deserved response to crime (see H. Morris 1968; N. Morris 1974; Murphy 1973; von Hirsch 1976).

Retributivism comes in very different forms (Cottingham 1979). All can be understood, however, as attempting to answer the two central questions faced by any retributivist theory of punishment. First, what is the justificatory relationship between crime and punishment that the idea of desert is supposed to capture: why do the guilty ‘deserve to suffer’ (see L. Davis 1972) -- and what do they deserve to suffer (see Ardal 1984; Honderich 1984, ch. 2)? Second, even if they deserve to suffer, why should it be for the state to inflict that suffering on them through a system of criminal punishment (Murphy 1985; Husak 1992; Shafer-Landau 1996)?

One retributivist answer to these questions which was popular for a time was that crime involves taking an unfair advantage over the law-abiding, and that punishment removes that unfair advantage. The criminal law benefits all citizens by protecting them from certain kinds of harm: but this benefit depends upon citizens accepting the burden of self-restraint involved in obeying the law. The criminal takes the benefit of the self-restraint of others, but refuses to accept that burden herself: she has gained an unfair advantage, which punishment removes by imposing some additional burden on her. (See H. Morris 1968; Murphy 1973; Sadurski 1985; Sher 1987, ch. 5; Adler 1992, chs. 5-8; Dagger 1993: for criticism, see Burgh 1982; Falls 1987; Dolinko 1991; Anderson 1997).

This kind of account does indeed answer the two questions noted above. What the criminal deserves to suffer is the loss of her unfair advantage, and she deserves that because it is unfair that she should get away with taking the benefits of the law without accepting the burdens on which those benefits depend; it is the state's job to inflict this suffering on her, because it is the author or guarantor of the criminal law. However, such accounts have internal difficulties: for instance, how are we to determine how great was the unfair advantage gained by a crime; how far are such measurements of unfair advantage likely to correlate with our judgements of the seriousness of crimes? (For a detailed defence of the 'unfair advantage' theory as a theory of sentencing, see M. Davis 1992, 1996; for criticism see Scheid 1990, 1995; von Hirsch 1990.) Furthermore, they seem to misrepresent what it is about crime that makes it deserving of punishment: what makes murder, or rape, or theft, or assault a criminal wrong, deserving of punishment, is surely the wrongful harm that it does to the individual victim -- not (as on this kind of account) the supposed unfair advantage that the criminal takes over all those who obey the law.

A different retributivist account appeals not to the abstract notion of unfair advantage, but to our (normal, appropriate) emotional responses to crime: to, for instance, the resentment of 'retributive hatred', involving a desire to make the wrongdoer suffer, that crime may arouse (see Murphy & Hampton 1988, chs. 1, 3); or to the guilt, involving the judgement that I ought to be punished, that my own wrongdoing would arouse in me (see Moore 1997, ch. 4). Such accounts try to answer the first of the two questions noted above: crime deserves punishment in the sense that it makes appropriate certain emotions (resentment, guilt) which are satisfied by or expressed in punishment. However, they do not yet show why it should be the state's task to satisfy or provide formal expression for such emotions (but see Stephen 1873: 152); and their answers to the first question are also problematic. Criminal wrongdoing should, we can agree, provoke certain kinds of emotion, such as self-directed guilt and other-directed indignation; and such emotions might typically involve a desire to make those at whom they are directed suffer. But just as we can agree that anger is an appropriate response to wrongs done to me, whilst also arguing that we should resist the desire to hit back which anger often, even typically, involves (see Horder 1992:194-7): so we could argue that we should resist the desire for suffering that guilt and resentment typically involve. At the least we need to know more than we are told by these accounts about just what wrongdoers deserve to suffer, and why the infliction of suffering should be an appropriate way to express such proper emotions. (For critical discussions of Murphy, see Murphy & Hampton 1988, ch. 2; Duff 1996: 29-31; Murphy 1999. On Moore, see Dolinko 1991: 555-9; Knowles 1993; Murphy 1999.)

A third kind of account seeks the meaning and justification of punishment as a deserved response to crime in its expressive or communicative character. (On the expressive dimension of punishment, see generally Feinberg 1970, Primoratz 1989: for critical discussion see Hart 1963: 60-69; Skillen 1980; M. Davis 1996: 169-81.) Consequentialists can of course portray punishment as useful partly in virtue of its expressive character (see Lacey 1988; Braithwaite & Pettit 1990): but a portrayal of punishment as a mode of moral communication has been central to some recent versions of retributivism.

6. Punishment as Communication

The central meaning and purpose of punishment, on such accounts, is to communicate to offenders the

censure or condemnation that they deserve for their crimes. Once we recognise, as we should, that punishment can serve this communicative purpose, we can see how such accounts begin to answer the two questions that retributivists face. First, there is an obviously intelligible justificatory relationship between wrongdoing and censure -- as a response which is intended to bring pain (the pain of condemnation by one's fellows) to an offender for his offence: whatever puzzles there might be about other attempts to explain the idea of penal desert, the idea that wrongdoers deserve to suffer censure is surely unpuzzling. Second, it is appropriate for the state to ensure that such censure is formally administered through the criminal justice system: for crimes are public wrongs, breaches of the political community's authoritative code; as such, they merit public censure by the community. Furthermore, whilst internal to censure is the intention, or hope, that the person censured will accept the censure as justified and will thus be motivated to avoid crime in future, this kind of account can avoid the charge (as brought against consequentialist theories) that it seeks to coerce or manipulate offenders into obeying the law. For censure addresses, and respects, the person censured as a rational and responsible agent: it constitutes an appropriate, deserved response to the wrong that she did, and seeks to bring her to modify her future conduct only by reminding her of the good moral reasons that she has for refraining from crime (see von Hirsch 1993, ch.2; Duff 2000, chs. 1.4.4, 3.2).

However, an obvious and crucial question faces any such justification of punishment as a communicative enterprise. Censure can be communicated through a formal conviction in a criminal court; or it could be communicated by some further formal denunciation issued by a judge or some other representative of the legal community, or by a system of purely symbolic punishments which were painful only in virtue of their censorial meaning. It can, of course, also be communicated by ‘hard treatment’ punishments of the kinds imposed by our courts -- by imprisonment, by compulsory community service, by fines and the like, which are painful or burdensome independently of their censorial meaning (on ‘hard treatment’, see Feinberg 1970): but why should we choose such methods of communication, rather than methods that do not involve hard treatment (see Christie 1981: 98-105)? Is it because they will make the communication more effective (see Falls 1987; Primoratz 1989; Kleinig 1991)? But why is it so important to make the communication effective -- and is there not a serious danger that the hard treatment will conceal, rather than highlight, the moral censure it should communicate (see Mathiesen 1990: 58-73)?

One kind of answer to this question brings consequentialism and deterrence back into the picture: we should communicate censure through penal hard treatment because this will give those who are insufficiently impressed by the moral appeal of censure prudential reason to refrain from crime; because, that is, the prospect of such punishment might deter those who are not susceptible to moral persuasion. (See Lipkin 1988, Baker 1992; and for a sophisticated revision of this idea, which makes deterrence firmly secondary to censure, see von Hirsch 1993, ch. 2; Narayan 1993: for critical discussion see Bottoms 1998; Duff 2000, ch. 3.3.) This kind of account differs from the side-constrained consequentialist accounts discussed earlier, since the (retributivist) imposition of deserved censure is now part of the positive justifying aim of punishment; and it can claim, in response to the Hegelian objection to deterrence, that it does not address potential offenders merely ‘like dogs’, since the laws initial appeal to the citizen is in the appropriate moral terms: the prudential, coercive reasons constituted by penal hard treatment as deterrence are relevant only to those who are deaf, or at least insufficiently attentive, to the law's moral appeal. It is still true, however, that on this account the law, in speaking to those who are not

persuaded by its moral appeal, is to abandon the attempt at moral communication in favour of the brute language of threats; and, for those who take seriously the Kantian demand that the state and its law should address its citizens as responsible moral agents, this slide into deterrence is still morally problematic.

A different answer to the ‘Why hard treatment?’ question explains penal hard treatment as an essential aspect of the enterprise of moral communication itself. Punishment, on this view, should aim not merely to communicate censure to the offender, but to persuade the offender to recognise and repent the wrong he has done, and so to recognise the need to reform himself and his future conduct, and to make apologetic reparation to those whom he wronged. His punishment then constitutes a kind of secular penance that he is required to undergo for his crime: its hard treatment aspects, the burden it imposes on him, should serve both to assist the process of repentance and reform, by focusing his attention on his crime and its implications, and as a way of making the apologetic reparation that he owes. (See Duff 2000. This kind of account has some relation to accounts that portray punishment as a kind of moral education: see H. Morris 1981; Hampton 1984; for criticism see Deigh 1984; Shafer-Landau 1991). This account faces serious objections (see Bickenbach 1988; Ten 1990; von Hirsch 1999; Bagaric & Amarasekara 2000): in particular that it cannot show penal hard treatment to be a necessary aspect of a communicative enterprise which is still to respect offenders as responsible and rational agents who must be left free to remain unpersuaded; that apologetic reparation must be voluntary if it is to be of any real value; and that a liberal state should not take this kind of intrusive interest in its citizens' moral characters. We cannot discuss these objections here, but should turn to a currently prominent strand in abolitionist thought, with which this communicative account of punishment can be usefully compared.

7. ‘Restorative Justice’

The ‘restorative justice’ movement has been growing in strength: although there are different and conflicting conceptions of what ‘restorative justice’ means or involves, the central theme is that what crime makes necessary is a process of reparation or restoration between offender, victim and other interested parties; and that this is achieved not through a criminal process of trial and punishment, but through mediation or reconciliation programmes that bring together the victim, offender and other interested parties to discuss what was done and how to deal with it (see generally Matthews 1988; Daly & Immarigeon 1998; von Hirsch & Ashworth 1998, ch. 7; Braithwaite 1999).

Now advocates of restorative justice often contrast it with ‘retributive’ justice, and argue that we should look for restoration rather than retribution or punishment. But it could be argued that this is a mistake. For when we ask what it is that requires ‘restoration’ or repair, the answer must refer not only to whatever material harm was caused by the crime, but to the wrong that was done: that was what fractured the relationship between offender and victim (and the broader community), and that is what must be recognised and ‘repaired’ or made up for if a genuine reconciliation is to be achieved. A restorative process that is to be appropriate to crime must therefore be one that seeks an adequate recognition, by the offender and by others, of the wrong done -- a recognition that must for the offender, if genuine, be repentant; and that seeks an appropriate apologetic reparation for that wrong from the offender. But those are also the aims of punishment as a species of secular penance, as sketched above.

This argument does not, of course, support that account of punishment against its critics. What it might suggest, however, is that whilst we can learn much from the restorative justice movement, especially about the role that processes of mediation and reparation can play in our responses to crime, its aim should not be the abolition or replacement of punishment: ‘restoration’ is better understood, in this context, as the proper aim of punishment, not as an alternative to it (see further Duff 2000, ch. 3.4-6, but also Zedner 1994).

8. Further Issues

The previous sections sketched the central contemporary accounts of whether and how legal punishment can be justified -- and some of the objections and difficulties that they face. A number of further important questions face any theory of punishment, which can only be noted here.

First, there are questions about sentencing. Who should decide what kinds and what levels of sentence should be attached to different offences or kinds of offence: what should be the respective roles of legislatures, of sentencing councils or commissions, of appellate courts, of trial judges, of juries? By what criteria should such decisions be made: how far should they be guided by a retributivist principle of proportionality, requiring punishments to be ‘proportionate’ in their severity to the seriousness of the crime; how far by consequentialist considerations of efficient crime-prevention? What kinds of punishment should be available to sentencers, and how should they decide which mode of punishment is appropriate for the particular offence (considerations of the meaning of different modes of punishment should be central to this question)? (On sentencing see generally Robinson 1987; Morris & Tonry 1990; von Hirsch 1993; Tonry 1996; von Hirsch & Ashworth 1998)

Second, there are questions about the relation between theory and practice -- between the ideal, as portrayed by a normative theory of punishment, and the actualities of existing penal practice. Suppose we have come to believe, as a matter of normative theory, that a system of legal punishment could in principle be justified -- that the abolitionist challenge can be met. It is, to put it mildly, unlikely that our normative theory of justified punishment will justify our existing penal institutions and practices: it is far more likely that such a theory will show our existing practices to be radically imperfect -- that legal punishment as it is now imposed is far from meaning or achieving what it should mean or achieve if it is to be adequately justified. If our normative theorising is to be anything more than an empty intellectual exercise, if it is to engage with actual practice, we then face the question of what we can or should do about our current practices. The obvious answer is that we should strive so to reform them that they can be in practice justified, and that answer is certainly available to consequentialists, on the plausible assumption that maintaining our present practices, whilst also seeking their reform, is likely to do more good or less harm than abandoning them. But for retributivists who insist that punishment is justified only if it is just, and for communicative theorists who insist that punishment is just and justified only if it communicates an appropriate censure to those who deserve it, the matter is harder: for to maintain our present practices, even while seeking their radical reform, will be to maintain practices which perpetrate serious injustice (see Murphy 1973; Duff 2000, ch. 5).

Third, the relation between the ideal and the actual is especially problematic in the context of punishment partly because it involves the preconditions of just punishment. That is to say, what makes an actual system of punishment unjust(ified) might be not its own operations as such (what punishment is or achieves within that system), but the absence of certain political, legal and moral conditions on which the whole system depends for its legitimacy (see Duff 2000, ch. 5.2). For instance, a just system of criminal law must convict and punish only those who are responsible, in the sense of being answerable for their crimes: only those who have the capacities necessary to answer for their actions, who are bound by this criminal law, and who are answerable to the political community whose law it is and whose courts call them to answer. There is much work to be done in spelling out such preconditions of just punishment: but it is at least arguable that they are far from satisfied for many of those who are convicted and punished by our own systems of criminal justice. To the extent that they are not satisfied, however, those systems lack legitimacy, and the punishments they inflict are unjustified. This conclusion should not surprise us: but it challenges any comfortable assumption that we can support and rely on our existing systems of criminal justice with a clear conscience.

Bibliography

Honderich 1984, Ten 1987, and Primoratz 1999 are useful introductory books. Duff & Garland 1994 and von Hirsch & Ashworth 1998 are useful collections of readings.

- Adler, J. (1992), The Urgings of Conscience. Philadelphia: Temple University Press.
- Alexander, L. (1980), ‘The Doomsday Machine: Proportionality, Punishment and Prevention’. The Monist 63: 199-227.
- Anderson, J. L. (1997), ‘Reciprocity as a Justification for Retributivism’. Criminal Justice Ethics 16: 13-25.
- Ardal, P. (1984), ‘Does Anyone ever Deserve to Suffer?’ Queen's Quarterly 91-2: 241-57.
- Bagaric, M., & Amarasekara, K. (2000), ‘The Errors of Retributivism’. Melbourne University Law Review 24: 1-66.
- Baker, B. M. (1992), ‘Consequentialism, Punishment and Autonomy’. In Retributivism and Its Critics, ed. W. Cragg. Stuttgart: Franz Steiner, 149-61.
- Becker, L. (1974), ‘Criminal Attempts and the Theory of the Law of Crimes’. Philosophy and Public Affairs 3: 262-94.
- Bianchi, H. (1994), Justice as Sanctuary: Toward a New System of Crime Control. Bloomington: Indiana University Press.
- Bianchi, H , & van Swaaningen, R. (eds.), (1986), Abolitionism: Towards a Non-Repressive Approach to Crime. Amsterdam: Free University Press.
- Bickenbach, J. E. (1988), Critical Notice of R. A. Duff, Trials and Punishments. Canadian Journal of Philosophy 18: 765-86.
- Bottoms, A. (1998), ‘Five Puzzles in von Hirsch's Theory of Punishment’. In Fundamentals of Sentencing Theory, ed. A. J. Ashworth & M. Wasik. Oxford: Oxford University Press, 53-100.
- Braithwaite, J. (1999), ‘Restorative Justice: Assessing Optimistic and Pessimistic Accounts’. In

Crime and Justice: A Review of Research, vol. 23, ed. M. Tonry. Chicago: University of Chicago Press, 241-367.

- Braithwaite, J., & Pettit, P. (1990), Not Just Deserts. Oxford: Oxford University Press.
- Burgh, R.W. (1982), 'Do the Guilty Deserve Punishment?' Journal of Philosophy 79: 193-210.
- Christie, N. (1977), 'Conflicts as Property'. British Journal of Criminology 17: 1-15.
- Christie, N. (1981), Limits to Pain. London: Martin Robertson.
- Cottingham, J. (1979), 'Varieties of Retribution'. Philosophical Quarterly 29: 238-46.
- Dagger, R. (1993), 'Playing Fair with Punishment'. Ethics 103: 473-88.
- Daly, K. & Immarigeon, R. (1998), 'The Past, Present, and Future of Restorative Justice'. Contemporary Justice Review 1: 21-45.
- Davis, L. H. (1972), 'They Deserve to Suffer'. Analysis 32: 136-40.
- Davis, M. (1989), 'The Relative Independence of Punishment Theory'. Law and Philosophy 7: 321-50.
- Davis, M. (1992), To Make the Punishment Fit the Crime. Boulder, Colorado: Westview Press.
- Davis, M. (1996), Justice in the Shadow of Death: Rethinking Capital and Lesser Punishments. Lanham: Rowman & Littlefield
- de Haan, W. (1990), The Politics of Redress: Crime, Punishment and Penal Abolition. London: Unwin Hyman.
- Deigh, J. (1984), 'On the Right to be Punished: Some Doubts'. Ethics 94: 191-211.
- Dimock, S. (1997), 'Retributivism and Trust'. Law and Philosophy 16: 37-62.
- Dolinko, D. (1991), 'Some Thoughts about Retributivism'. Ethics 101: 537-59.
- Duff, R. A. (1986), Trials and Punishments. Cambridge: Cambridge University Press.
- Duff, R. A. (1996), 'Penal Communications: Recent Work in the Philosophy of Punishment'. Crime and Justice: A Review of Research 20: 1-97.
- Duff, R.A. (2000), Punishment, Communication, and Community. New York: Oxford University Press.
- Duff, R.A., & Garland, D. (eds.) (1994), A Reader on Punishment. Oxford: Oxford University Press.
- Falls, M. M. (1987), 'Retribution, Reciprocity, and Respect for Persons'. Law and Philosophy 6: 25-51.
- Farrell, D. M. (1985), 'The Justification of General Deterrence'. Philosophical Review 94: 367-94.
- Farrell, D. M. (1995), 'Deterrence and the Just Distribution of Harm'. Social Philosophy and Policy 12: 220-240.
- Feinberg, J. (1970), 'The Expressive Function of Punishment'. In his Doing and Deserving, Princeton, N. J.: Princeton University Press, 95-118
- Feinberg, J. (1988), Harmless Wrongdoing (vol. IV of The Moral Limits of the Criminal Law). New York: Oxford University Press.
- Goldman, A. H. (1982), 'Toward a New Theory of Punishment'. Law and Philosophy 1: 57-76.
- Hampton, J. (1984), 'The Moral Education Theory of Punishment'. Philosophy and Public Affairs 13: 208-38.
- Hare, R.M. (1981), Moral Thinking: Its Levels, Methods and Point. Oxford: Oxford University Press.
- Hart, H. L. A. (1963), Law, Liberty and Morality. New York: Random House.

- Hart, H. L. A. (1968), Punishment and Responsibility. Oxford: Oxford University Press.
- Hegel, G.W.F. (1821), The Philosophy of Right, trans. T Knox. Oxford: Oxford University Press (1942).
- Honderich, T. (1984), Punishment: The Supposed Justifications (rev. ed.). Harmondsworth, Middlesex: Penguin Books.
- Horder, J. (1992), Provocation and Responsibility. Oxford: Oxford University Press.
- Hulsman, L. (1986), 'Critical Criminology and the Concept of Crime'. Contemporary Crises 10: 63-80.
- Hulsman, L. (1991), 'The Abolitionist Case: Alternative Crime Policies'. Israel Law Review 25: 681-709.
- Husak, D. (1992), 'Why Punish the Deserving?' Nous 26: 447-64.
- Kleinig, J. (1991), 'Punishment and Moral Seriousness'. Israel Law Review 25: 401-21.
- Knowles, D. (1993), 'Unjustified Retribution'. Israel Law Review 27: 50-58.
- Lacey, N. (1988), State Punishment: Political Principles and Community Values. London: Routledge.
- Lewis, C. S. (1953), 'The Humanitarian Theory of Punishment'. Res Judicatae 6; reprinted in Readings in Ethical Theory (2nd ed.), ed. W Sellars & J Hospers. New York: Appleton-Century-Crofts (1970), 646-50.
- Lipkin, R. J. (1988), 'Punishment, Penance and Respect for Autonomy'. Social Theory and Practice 14: 87-104.
- Marshall, S. E., & Duff, R. A. (1998), 'Criminalization and Sharing Wrongs'. Canadian Journal of Law & Jurisprudence 11: 7-22.
- Mathiesen, T. (1990), Prison on Trial. London: Sage.
- Matthews, R., ed. (1988), Informal Justice. London: Sage.
- Menninger, K. (1968), The Crime of Punishment. New York: Viking Press.
- Montague, P. (1995), Punishment as Societal Defense. Lanham: Rowman & Littlefield.
- Moore, M. S. (1997. Placing Blame: A Theory of Criminal Law. Oxford: Oxford University Press.
- Morison, J. (1988), 'Hart's Excuses: Problems with a Compromise Theory of Punishment'. In The Jurisprudence of Orthodoxy, ed. P. Leith & P. Ingram. London: Routledge, 117-46.
- Morris, C. W. (1991), 'Punishment and Loss of Moral Standing'. Canadian Journal of Philosophy 21: 53-79.
- Morris, H. (1968), 'Persons and Punishment'. The Monist 52: 475-501.
- Morris, H. (1981), 'A Paternalistic Theory of Punishment'. American Philosophical Quarterly 18: 263-71.
- Morris, N. (1974), The Future of Imprisonment. Chicago: University of Chicago Press.
- Morris, N., & Tonry, M. (1990), Between Prison and Probation: Intermediate Punishments in a Rational Sentencing System. New York: Oxford University Press.
- Murphy, J. G. (1973), 'Marxism and Retribution'. Philosophy and Public Affairs 2: 217-43.
- Murphy, J. G. (1985), 'Retributivism, Moral Education and the Liberal State'. Criminal Justice Ethics 4: 3-11.
- Murphy, J. G. (1999), 'Moral Epistemology, the Retributive Emotions, and the "Clumsy Moral Philosophy" of Jesus Christ'. In The Passions of Law, ed. S. Bandes. New York: NYU Press, 149-67.

- Murphy, J. G., & Coleman, J. (1984), The Philosophy of Law. Totowa, N. J.: Rowman & Littlefield.
- Murphy, J. G., & Hampton, J. (1988), Forgiveness and Mercy. Cambridge: Cambridge University Press
- Narayan, U. (1993), ‘Appropriate Responses and Preventive Benefits: Justifying Censure and Hard Treatment in Legal Punishment’. Oxford Journal of Legal Studies 13: 166-82.
- Pettit, P. (1997), ‘Republican Theory and Criminal Punishment’. Utilitas 9: 59-79.
- Philips, M. (1986), ‘The Justification of Punishment and the Justification of Political Authority’. Law and Philosophy 5: 393-416.
- Primoratz, I. (1989), ‘Punishment as Language’. Philosophy 64: 187-205.
- Primoratz, I. (1999), Justifying Legal Punishment (2nd ed.). New Jersey: Humanities Press.
- Quinn, W. (1985), ‘The Right to Threaten and the Right to Punish’. Philosophy and Public Affairs 14: 327-73.
- Rawls, J. (1955), ‘Two Concepts of Rules’. The Philosophical Review 64: 3-32.
- Robinson, P. (1987), ‘A Sentencing System for the 21st Century?’ Texas Law Review 66: 1-61.
- Sadurski, W. (1985), ‘Distributive Justice and the Theory of Punishment’. Oxford Journal of Legal Studies 5: 47-59.
- Scheid, D. E. (1980), ‘Note on Defining "Punishment"’. Canadian Journal of Philosophy 10: 453-462.
- Scheid, D. (1990), ‘Davis and the Unfair-Advantage Theory of Punishment: A Critique’. Philosophical Topics 18: 143-70.
- Scheid, D. (1995), ‘Davis, Unfair Advantage Theory and Criminal Desert’. Law and Philosophy 14: 375-409.
- Scheid, D. (1997), ‘Constructing a Theory of Punishment, Desert, and the Distribution of Punishments’. Canadian Journal of Law and Jurisprudence 10: 441-506.
- Shafer-Landau, R. (1991), ‘Can Punishment Morally Educate?’ Law and Philosophy 10: 189-219.
- Shafer-Landau, R. (1996), ‘The Failure of Retributivism’. Philosophical Studies 82: 289-316.
- Sher, G. (1987), Desert. Princeton: Princeton University Press.
- Skillen, A. J. (1980), ‘How to Say Things with Walls’. Philosophy 55: 509-23.
- Smart, J. J. C. (1973), ‘An Outline of a System of Utilitarian Ethics’. In Smart & B Williams, Utilitarianism: For and Against, by. Cambridge: Cambridge University Press, 1-74.
- Stephen, J.F. (1873), Liberty, Equality, Fraternity, edited by J White. Cambridge: Cambridge University Press (1967).
- Ten, C. L. (1987), Crime, Guilt and Punishment. Oxford: Oxford University Press.
- Ten, C. L. (1990), ‘Positive Retributivism’. Social Philosophy and Policy 7: 194-208.
- Tonry, M. (1996), Sentencing Matters. New York: Oxford University Press.
- von Hirsch, A. (1976), Doing Justice: The Choice of Punishments. New York: Hill & Wang.
- von Hirsch, A. (1990), ‘Proportionality in the Philosophy of Punishment: From "Why Punish?" to "How Much?"’ Criminal Law Forum 1: 259-90.
- von Hirsch, A. (1993), Censure and Sanctions. Oxford: Oxford University Press.
- von Hirsch, A. (1999), ‘Punishment, Penance and the State’. In Punishment and Political Theory, ed. M. Matravers. Oxford: Hart Publishing, 69-82.
- von Hirsch, A., & Ashworth, A. J. (1992), ‘Not Not Just Deserts: A Response to Braithwaite and

Pettit'. Oxford Journal of Legal Studies 12: 83-98.

- von Hirsch, A., & Ashworth, A. J., eds. (1998), Principled Sentencing, 2nd ed. Oxford: Hart Publishing
- Walker, N. (1980), Punishment, Danger and Stigma. Oxford: Blackwell.
- Walker, N. (1991), Why Punish? Oxford: Oxford University Press.
- Williams, B. (1976), Morality. Cambridge: Cambridge University Press.
- Wilson, J. Q. (1983), Thinking about Crime (rev. ed.). New York: Basic Books.
- Wootton, B. (1963), Crime and the Criminal Law. London: Stevens.
- Zedner, L. (1994), 'Reparation and Retribution: Are They Reconcilable?' Modern Law Review 57: 228-50.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

legal obligation and authority | legal reasoning: interpretation and coherence | legal reasoning: precedent and analogy

Antony Duff
r.a.duff@stir.ac.uk

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 2, 2001
Content last modified: January 2, 2001

Brentano's Theory of Judgement

One of Brentano's foremost aims in philosophy was to provide a new foundation for epistemology and logic as two closely related disciplines. He tried to achieve this by a systematic analysis of the mental phenomena involved in attaining knowledge and in drawing inferences. For Brentano knowledge is reached by judgements that are directly or indirectly evident, and logical inferences can contribute to our knowledge because they can make a judgement indirectly evident for us. Hence both epistemology and logic rely on a conception of judgements, how they differ from other mental phenomena, and how they are related to each other.

Brentano's view of the nature of judgement differs significantly from other views that can be found in Aristotle, Kant, or Frege. In contrast to Aristotle, Brentano emphasizes the importance of existential judgements with only one term, and claims that predicative judgements are a special case of existential ones. In contrast to Kant, he emphasizes the difference between presentations and judgements, rejecting their unification in the single category ‘thinking’. In contrast to Frege, he holds that judgements do not require the existence of complete thoughts or propositions which have to be grasped before a judgement can be made. It is the mental act of judging, not its object or content, which is the bearer of truth-values. In view of these differences Brentano's theory of judgement has been called *existential* (non-predicative), *idiogenetic* (non-reductionist), and *reistic* (non-propositional).

Today Brentano's theory does not have many adherents. The now dominant view is that propositions or sentences are the objects of belief, and that judgements occur when beliefs are acquired, manifested, or changed. Logical inferences are then defined as relations between propositions or sentences, abstracting from the mental attitudes that go along with them. Although this anti-psychological approach is widely accepted today, there is still an open question concerning the order of explanation here: Are beliefs and judgements true because they are directed at true propositions, or should we say that propositions (and sentences) are true because they express true beliefs and judgements? Once this question is raised, Brentano's theory of judgement remains an interesting alternative to the current mainstream in logic and epistemology.

The Nature of Judgement

The main elements of Brentano's theory of judgement can be found in chapter 7 and appendix IX of his *Psychology from an Empirical Standpoint* (1874; the appendix was added in the second edition 1911). A more elaborate exposition of his theory is contained in the logic lectures which Brentano held at the University of Wuerzburg (1869-71) and at the University of Vienna (1875-1889). Unfortunately Brentano never realized his plan -- announced in the second edition of the *Psychology* (Vol. II, p.77n/p.230n) -- to publish his elementary logic as a whole. Up to now only a small selection of his lecture notes, mixed with excerpts from other writings by Brentano and his pupil Franz Hillebrand, has been published in *Die Lehre vom richtigen Urteil* (1956).

Brentano's leading question was a psychological one: What happens in our minds when we make a judgement? Introspectively it is an act quite similar to making a decision, although its behavioral effects are different. Suppose you are uncertain what to think about the existence of extraterrestrial life. Some data suggest that life exists only on earth, others suggest that there may be intelligent beings somewhere else in the universe. Eventually you may become convinced one way or the other, and you either accept or reject the existence of extraterrestrial life. That is when you *judge*.

This example illustrates three crucial claims that Brentano makes:

> (1) Judgements require that something (some object) is given in presentation, but not that something is predicated of it.
>
> (2) Judgements are either positive or negative, depending on whether the presented object is accepted as existing, or rejected as fictitious or non-existing.
>
> (3) Judgements are best expressed in sentences of the form ‘A exists’ or ‘A does not exist’, where the term ‘A’ denotes the presented object which is also the object of the judgement, and the rest of the sentence indicates its quality.

These three claims form the core of Brentano's theory of judgement: The *foundational thesis* (1) concerns the relation between judgement and predication, the *polarity thesis* (2) determines the place of negation in judgements, and the *existential thesis* (3) determines a canonical form in which all judgements can be expressed. Of course, these claims must be seen in the context of Brentano's overall theory of mental phenomena, in particular in the context of his account of intentionality. This background cannot be

discussed here, but it is worth mentioning that the term ‘object of judgement’, as it is used here, always refers to an entity which is distinct from the judgement itself and not contained in it. It is also assumed here that judgements have a content or subject matter, which is not separable from the act itself, and which Brentano originally called the ‘immanent objectivity of a mental phenomenon’. The content of a judgement must not be conceived as a propositional entity, however, since Brentano explicitly denied that judgements have such entities as their contents. (Complex entities which are not propositional and which are just as ephemeral as the content of a judgement can already be found in Aristotle; see G. B. Matthews, 1982.)

All three of Brentano's claims above were already highly controversial among his immediate pupils. We find for instance in Husserl's fifth Logical Investigation an account of judgements which deviates from Brentano in all three respects. According to Husserl judgements are intentional acts with a propositional content directed at proposition-like entities which he calls *Sachverhalte*. Why Husserl deviated from his teacher in such a radical way, and whether he did so for good reasons, are questions still in discussion today. (See for instance Mulligan 1988).

The Foundational Thesis and the Judgement/Predication Distinction

The claim that judgements are based on presentations is a commonplace in philosophy, but it is a matter of controversy how this relationship should exactly be spelled out. Traditional logic suggests that two presentations must be involved in every judgement, since a judgement is made when something is attributed or denied of something else. Therefore the sentences that are traditionally used for expressing judgements have the subject-predicate form ‘S is P’ and ‘S is not P’.

Brentano rejects this traditional view by pointing out that judgements may arise also from a single presentation. When someone judges that extraterrestrials exist, he does not connect the notion of extraterrestrial life with the notion of existence. He merely thinks of such beings and accepts their existence, i.e. he has a presentation of such beings and accepts it as a presentation of something existing. Existential judgements are therefore not to be expressed in the subject-predicate form ‘S is P’, but in the simple form ‘A exists’, when ‘A’ is a singular term, and ‘A's exist’ or ‘Some A exist’, when ‘A’ is a general term. (Brentano mentions in a footnote that Aristotle himself may have acknowledged simple judgements of this form. *Psycholog*y Vol. II, p.54n/p.211n).

Existential judgements show that predication is not *necessary* for forming a judgement, but neither is it sufficient according to Brentano. Many philosophers have assumed that a predicative judgement is nothing more than ‘the putting together of two ideas’ -- in the case of ‘S is P’ -- or ‘the separating of two ideas’ -- in the case of ‘S is not P’. This view is sometimes called the ‘combinatorial theory of judgement’, and Brentano was not the first to point out the deficiences of this view. He refers to John Stuart Mill who already denied that judgements arise from a habit of associating or dissociating ideas. What Brentano adds to Mill's criticism is a precise diagnosis of the mistake: the combinatorial theory

tries to locate the characteristic feature of a judgement in its *content* instead of locating it in its *quality*. When we combine a subject- and a predicate-term we just form a more complex idea which is again the content of a presentation. What is still missing is the qualitative moment of acceptance or rejection (see *Psychology*, Vol II, p.63/p.221).

Thus Brentano's theory draws a sharp line between judgement and predication in recognizing judgements with a non-predicational content and in taking subjectless sentences at face value. Sentences like ‘It is raining’ or ‘There is no water on the moon’ need not be paraphrased into subject-predicate form along the lines of ‘The weather is rainy’ or ‘The moon is lacking water’. They directly express a judgement by specifying an object which is given in presentation (rain, water on the moon) and by indicating whether this object is accepted or rejected. (This advantage of Brentano's theory was especially exploited by Marty 1884-1895).

Things get more complicated, however, when Brentano later (in appendix IX of the second edition of the *Psychology*) introduces so-called ‘double judgements’. In making a double judgement one first accepts the existence of something, and then adds to this first judgement a second one to the effect that the object, whose existence one already has accepted, either has or lacks some property. According to this refined view, a predication is made not by combining two ideas or presentations, but by combining two judgements.

The introduction of double judgements leaves the analysis of existential judgements intact, since in judging that S exists we do not first accept S as existing and then attribute existence or non-existence to it. However, one can now predicate P of S in two different ways: either by first forming the complex presentation of an object S which is P and then accepting this object, or by first accepting the existence of S and then attributing P to it, thus making the double judgement that S is P. In this latter case too, the attribution of P involves two steps: first the predicate P is connected merely in presentation with object S whose existence has been accepted, and then the object S is accepted once more, but this time together with P as one of its properties. That predication and judgement remain distinct acts also in the case of double judgements can be seen from the following fact: When we imagine a person (perhaps oneself) who is double-judging that S is P, we can disagree only with the second part of her judgement, and still form the complex presentation of an S which is P. And conversely, we can form the complex presentation of an S which is P and yet agree with the double judgement that S is not P. (See *Psychologie*, Vol. II, p.164/p.295. This point is further elaborated in Terrell 1976).

In its final form Brentano's account of the relation between judgement and predication turns out to be less straightforward than the standard Fregean account with its simple distinction between ‘grasping a proposition’ and ‘judging it to be true’. At no point did Brentano, however, lose sight of the claim that predication is not essentially connected with judging.

The Polarity Thesis and the Judgement/Negation Distinction

According to Brentano's second thesis, judgements are always positive or negative. In this respect they are like preferences and emotional attitudes which are for or against something. (Brentano holds the controversial view that feelings and acts of will belong to the same category and that they all involve such a polarity.) Presentations, on the other hand, are neither positive nor negative. They simply present an object to the mind without taking a stance towards it. This happens when we simply see or hear something, or when we imagine something in our phantasy. As long as no judgement is made (and no emotional evaluation and no preference is involved), there is nothing positive or negative about an act of presentation.

This essential difference tends to be overlooked when one uses the single category of ‘thinking’ for both judgements and presentations, as does the Kantian tradition. According to Brentano presentations and judgements are as different from each other as they are different from feelings and acts of will. Their difference is not just external -- having to do with the way in which they influence our actions -- it is an internal difference lying in the distinctive quality of judgements. Therefore, if one acknowledges that feelings or acts of will form a separate category besides the category of ‘thinking’, one should accept for the same reason that judgements and presentations form distinct categories as well.

With his polarity thesis Brentano not only dismisses the Kantian tradition, he also rejects a view that Frege made popular, namely that there are no negative judgements. When we deny the existence of something, e.g. the existence of extraterrestrial life, we still accept something as true, Frege would say, namely the negative thought that there are no extraterrestrials. Negation enters the formation of *thoughts*, it does not divide *judgements* into positive and negative.

Frege's elimination of negative judgements rests on the assumption that thoughts (or judgement-contents) can be true or false independently of being accepted or rejected, and therefore can also be negated. Brentano does not explicitly discuss this view, but his objection to it seems clear: The polarity between truth and falsity must be grounded in our ability to form opposite judgements. We first have to realize that from two opposing judgements with respect to the same subject matter, one will be true and the other one false. Only then can we understand what it means for a sentence, (a judgement content, a proposition, a thought, or whatever), to be true or false. (These issues are further discussed in Reinach 1911).

Brentano's treatment of negation has important further consequences. First, if the contrast between truth and falsity is explained along these lines, then the contrast between positive and negative concepts must also be explained at the level of judgements, not at the level of presentations. In his later writings Brentano took up this challenge when he tried to show that only positively conceived ‘things’ are properly regarded as objects of presentations. This became his ontological doctrine called *reism*. (On this issue see Körner 1978).

Secondly, if negation is completely eliminated from the level of presentaions, the analysis of categorial judgements has to be revised accordingly. Initially, Brentano paraphrased these judgements in existential form as follows:

I: Some S are P	There is an S which is P
E: No S is P	There is no S which is P
O: Some S are not P	There is an S which is a non-P
A: All S are P	There is no S which is a non-P

The negation in E-judgements poses no problem: it properly indicates that a negative judgement is made. The negative concept ‘non-P’ used in the paraphrases of O- and A-judgements is more problematic, however. Here a negation enters at the level of presentations, not at the level of judgement as the polarity thesis requires.

A more complicated analysis is required to get around this difficulty. In the case of O-judgements the introduction of double judgements will help. It then turns out that an O-judgement does not consist in predicating non-P of S, but in first accepting S and then making a negative judgement to the effect that S is not P, i.e. a judgement that denies the application of P to S. This still leaves the A-judgements as a problem case. At this point Brentano again invokes a higher-level presentation, namely the presentation of someone whose judgements are evaluated as right or wrong. With these additional tools at hand, Brentano arrives at the following analysis of the four categorical judgements (see *Psychology*, Vol. II, 164-169/pp.295-298):

I: Some S are P	There is an S and that S is P
E: No S is P	There is no one who correctly judges ‘Some S is P’
O: Some S are not P	There is an S and that S is not P
A: All S are P	There is no one who correctly judges ‘Some S are not P’

All negations here indicate that a negative *judgement* is made. This vindicates the claim that the polarity between positive and negative judgements is *basic* and provides the distinguishing mark that separates judgements from presentations. Brentano admits, however, that for practical reasons it may be convenient to use negative concepts, e.g. for simplifying inferences. When one does so, one should keep in mind however that these concepts do not properly pick out objects of presentation. Along these lines one could also justify the use of propositional clauses and thereby avoid all the complications of the existential analysis; but Brentano does not seem to have considered this more radical simplification (see *Psychology*, Vol. II, p.169/p.299).

The Existential Thesis

Brentano's third thesis says that all simple judgements (that involve only a simple act of judging) can be expressed in sentences of the form ‘A exists’ or ‘A does not exist’ (or ‘A's exist’ and ‘A's do not exist’ respectively). This thesis marks the contrast to all propositional theories of judgement. Propositional

theories assume that a complete sentence (or a that-clause) is needed for expressing the content of a judgement. That a proposition (or sentence) is actually accepted, i.e. that a judgement is made, must therefore be indicated by an additional sign -- like Frege's judgement-stroke -- or it remains implicit in the assertive use of a declarative sentence.

On Brentano's theory, by contrast, only a simple or complex *term* is needed to express the content of a judgement, and hence a complete sentence can express both the *content* and the *quality* of a judgement. In making this claim, Brentano relies on the distinction between categorematic and syncategorematic expressions, i.e. between terms that purport to denote entities, and expressions like ‘is’, ‘and’, ‘or’, etc. that do not. The former specify the content of a judgement, whereas the latter are used for specifying its quality. This distinction also applies to sentences of the form ‘A exists’. Here the ‘exists’ does not purport to denote anything -- the property of existence -- rather it indicates which judgement is made: A positive judgement in present tense in the case of ‘A exists (now)’, a negative judgement in the present tense in the case of ‘A does not exist now’, a positive judgement in the past tense in the case of ‘A existed’, a negative apodictic judgement in the case of ‘A does necessarily not exist’, etc. (I consider here throughout only the most basic distinction between positive and negative cases.)

Brentano also introduces two special signs to separate those sentence parts that specify the content of a judgement from those that specify its quality. He uses the sign ‘+A’ to express the positive judgement that A exists, and the sign ‘−A’ to express the negative judgement that A does not exist. These signs remind one of Frege's judgement stroke, but the theory behind them is quite different. Two important differences should be noted here:

Firstly, ‘+A’ is not be read as ‘it is accepted that A exists’. This would suggest that the sign ‘+’ functions as the operator ‘it is accepted that’, and that the term ‘exists’ expresses part of the content of the judgement. But the whole point of Brentano's theory is that the term ‘exists’ is syncategorematic and merely expresses the *quality* of the judgement. ‘A’ alone must therefore express the whole (non-propositional) content. This also tells against a suggestion made by Arthur Prior, namely to read ‘A exists’ as ‘Something is A’. It is not enough to treat ‘existence’ as a second-level predicate to avoid the misinterpretation that it contributes to the content of the judgement (see Prior 1976, p.115).

Secondly, ‘−A’ should not to be read as ‘the existence of A is rejected’. This would suggest that there is a difference between ‘the existence of A is rejected’ and ‘the non-existence of A is accepted’, and equally between ‘A is rejected as existing’ and ‘A is accepted as non-existing’. Brentano's theory leaves no room for such distinctions. Otherwise it would reduce to the (non-controversial) claim that all categorial judgements are expressible in the form of existential propositions. Brentano's much stronger claim is however that no propositions at all are accepted in such judgements, not even existential ones.

What, then, is the best way to read the formulas ‘+A’ and ‘−A’? There is no better way than reading them as ‘A exists/does not exist’ or as ‘A is accepted/rejected’. Whatever term we use for the symbols ‘+’ and ‘−’, they will have no specific meaning beyond their function of indicating the quality of the judgement expressed.

Having noted these differences between Brentano's and Frege's symbolism, one may wonder whether Brentano really has a consistent theory here.

One problematic fact is that it is unclear how to interpret the formulas ‘+A’ and ‘−A’ when they are not used, but merely mentioned. When such a formula is quoted, the expression ‘A’ is still meaningful and expresses the content of a judgement, but the signs ‘+’ and ‘−’ become completely idle. This, of course, is also true of Frege's judgement stroke, which loses its function when it is not used to make an assertion.

However, there seems to be further difficulty that is peculiar only to Brentano's symbols. Whereas Frege's judgement-stroke is added to complete sentences, Brentano's symbols are *parts* of complete sentences. But every complete sentence can be used without expressing a judgement, for instance as the antecedent or consequent of a conditional. There is no obstacle in forming the complex judgement ‘If A exists, then B does not exist’, and yet we cannot symbolize it as ‘If +A, then −B’. Apparently, then, the term ‘exist’ is not (or not merely) an indicator of the judgement-quality, as Brentano would have it. (This objection was raised in Geach 1965.)

In dealing with this objection one might appeal to Brentano's own treatment of conditional (or hypothetical) judgements. He reduces them to *single* existential judgements with a complex object. Thus, a judgement of the form ‘If A exists, then B does not exist’ gets analysed as ‘An A together with a B does not exist’, where ‘A together with B’ denotes the complex object which is rejected (see *Psychology* Vol. II, p.170/p.299; see also *Lehre* p.123).

But there is more to Geach's objection. It shows that on Brentano's theory the term ‘exists’, like the copula ‘is’, can be used in two different ways. It can either be used to *express* a judgement or to *talk about* a judgement made by someone (possibly by oneself). We have already seen how Brentano uses this distinction for separating judgement and presentation, and for analysing A-judgements without invoking negative concepts. He also needs to make use of this distinction when it comes to conditional judgements. The judgement ‘If A exists, then B does not exist’ might then be analysed as ‘It is impossible correctly both to accept A and to reject B’, which can be expressed in existential form as ‘Someone who can correctly accept A and reject B does not exist’. (This analysis is suggested in Chisholm 1982, p.36).

In this way Brentano's theory of judgement may be applicable to a wider range of complex judgements (see Pasquarella 1987). Even if these extensions are rejected as unnecessarily complicated however, Brentano's existential analysis offers a viable alternative to the propositional theory at least for some basic kinds of judgements, like the ones used in syllogistic. This may not be very significant from the point of view of modern logic, which does not distinguish between basic and non-basic judgements in this way, but it may have a considerable *ontological* significance. Brentano's theory shows how a commitment to propositional entities can be avoided at least within certain limits. Entities like ‘propositions’, ‘states of affairs’, ‘facts’, ‘Meinongian objectives’, etc. might therefore be introduced only for convenience, but they need not be taken ontologically seriously. Any stronger commitment to such entities remains therefore dubious, and it is for this reason that Brentano came to reject the

correspondence theory of truth. Judgements are true, according to his existential thesis, because certain entities *exist* (or do not *exist*), not because certain entities ‘correspond’ to our judgements. (Advocates of a correspondence theory have criticized Brentano precisely for this reason. See Schlick 1925, pp.60ff and 176ff).

The Reform of Logic

In the second half of the 19th century logic freed itself from the constraints of the Aristotelian tradition. This move is often linked with the demise of ‘psychologism’, the view that logic needs to be based on psychology. Mathematical logicians like Bolzano and Frege established modern logic as a strictly non-psychological, ‘objective’ discipline. From this point of view Brentano appears as one of the last advocates of the ‘old logic’, and his theory as a final attempt at providing a psychological foundation for logic.

It is true that Brentano rejected the idea of a ‘mathematical logic’ as he found it in the writings of George Boole (see *Psychology*, Appendix X). Nevertheless, as we have noticed, there are important points of convergence between Brentano's and Frege's views (of which neither of them seems to have been aware): (1) judgement is distinct from predication, (2) existence is not a first-level predicate, (3) logical analysis must penetrate the linguistic expressions which often disguise the form of our judgements. But this is not all. There is even more agreement between Brentano and modern logic, however, when one compares them with the old syllogistic logic.

This further convergence becomes visible when one considers Brentano's criticism of the traditional *square of opposition*. This square is made up of the four categorial judgements A (‘All S are P’), E (‘No S are P’), I (‘Some S are P’), and O (‘Some S are not P’), among which the following relations have been claimed to hold:

(i) A contradicts O, and vice versa.

(ii) E contradicts I, and vice versa.

(iii) A and E can be false but not true together (= law of contrariety)

(iv) I and O can be true but not false together (= law of subcontrariety)

(v) A implies I (= subalternation)

(vi) E implies O (= subalternation)

(vii) I converts into ‘Some P are S’ (simple conversion)

(viii) E converts into ‘No P is S’ (simple conversion)

(ix) A converts into ‘All non-P are non-S’ (conversion by contraposition)

(x) O converts into ‘Some non-P are not non-S’ (conversion by contraposition)

Brentano rejects almost all of these claims. After translating the categorical judgements into existential form (leaving aside double-judgements for the moment), he reaches the following conclusions:

(i) and (ii) are the only logical relationships correctly identified by traditional logic.

(iii) to (vi) are mistaken: If S is an empty term, both A and E are true, and both I and O are false.

(vii) and (viii) are correct, but not because of a conversion of one judgement into another one, but only because *one* judgement is expressed in two ways.

(ix) and (x) are correct, but no contraposition is needed; only a simple conversion is used.

All these results emerge from one major shift in the underlying theory of judgements: Traditional logic takes A and I to be positive judgements, and E and O to be negative ones. According to Brentano all universal judgements (both A and E) are negative and therefore lack any existential import, whereas all particular judgements (both I and O) are positive and have such import. Once this ‘mistake’ is corrected, most of the traditional disputes about their logical relationships become obsolete. This is why Brentano said that his theory "leads to nothing less than a complete overthrow, and at the same time, a reconstruction of elementary logic. Everything then becomes simpler, clearer, and more exact" (*Psychology* Vol.II 77/230). (For a critical survey of Brentano's logic reform see Prior 1962, pp.166ff. and Simons 1987).

When we compare Brentano's results with the doctrines of modern logic, we see that they are in complete agreement concerning (i) - (vi). With respect to (vii) and (x) there is at least no major disagreement. It is still acceptable to say that the simple conversion of terms is only a change in the linguistic expression of a judgement, not in the judgement itself, and the same can be said about the conversion of an A-judgement. Here, too, no contraposition is needed, since in predicate logic an A-judgement can be either expressed as an implication or a negated conjunction.

In conclusion one may say that Brentano's psychological approach to logic did not prevent him from arriving at results very close to what modern logic teaches us. Perhaps, then, the major difference between Brentano and modern logic should not be seen in his psychologism, but rather in his focus on general terms. Frege changed this focus with his function/argument analysis of sentences, thereby replacing general terms with unsaturated expressions. This step is missing in Brentano's theory, and that is what sets it apart from the mainstream in contemporary logic and epistemology.

Bibliography

- Brentano, F., 1874 and 1911, *Psychologie vom empirischen Standpunkt*. Duncker & Humblot, Leipzig. (2nd ed. of engl. trans., 1995, *Psychology from an Empirical Standpoint*, Routledge, London.)
- Brentano, F., 1956, *Die Lehre vom richtigen Urteil*. Francke Verlag, Bern.
- Chisholm, R., 1982, ‘Brentano's Theory of Judgement’ in R. Chisholm, *Brentano and Meinong Studies*, Rodopi, Amsterdam, pp.17-36.
- Dölling, E., 1993, ‘Brentanos und Freges Urteilslehre -- Ein Vergleich’, in: W. Steltzner (ed.), *Philosophie und Logik*. Walter de Gruyter, Berlin, pp.24-32.
- Geach, P., 1965, ‘Assertion’, reprinted in P. Geach, 1972, *Logic Matters*, Basil Blackwell, Oxford, pp.254-269.
- Hillebrand, F., 1891, *Die neuen Theorien der kategorischen Schlüsse*. Wien, Hölder.
- Husserl, E., 1900-1901, *Logische Untersuchungen*. Niemeyer: Halle a. d. Saale. (Eng. trans. of 2nd. ed., *Logical Investigations*, Routledge, London 1970.
- Körner, St., 1978, ‘Über Brentanos Reismus und die extensionale Logik’ *Grazer Philosophische Studien* 5, pp.29-43.
- Marty, A., 1884-1895 ‘Über subjectlose Sätze und das Verhältnis der Grammatik zu Logik und Psychologie’, *Vierteljahresschrift für Philosophie* 8, pp. 56-94, 161-192, 292-340; 18, pp.320-356, 421-471; 19, pp.19-87, 263-334.
- Matthews, G.B., 1982, ‘Accidental Unities’, in M. Shofield and M.Nussbaum: *Language and Logos*, Cambridge University Press, Cambridge, 223-240.
- Mulligan, K., 1988, ‘Judgings: Their Parts and Counterparts’ *Topoi Supplementa 2*, pp.117-148
- Pasquarella, L., 1987, ‘Intensional Logic and Brentano's Non-propositional Theory of Judgement’, *Grazer Philosophische Studien*, 29, pp.59-62.
- Prior, A., 1962, *Formal Logic*, The Clarendon Press, Oxford.
- Prior, A., 1976, *The Doctrine of Propositions and Terms*. Duckworth, London.
- Reinach, A., 1911, ‘Zur Theorie des negativen Urteils’, Eng. trans. in: B. Smith (ed.), 1982, *Parts and Moments. Studies in Logic and Formal Ontology*. Philosophia Verlag, München, pp.315-377.
- Rojszczak, A., and Smith, B., ‘Theories of Judgement’, in T. Baldwin. (ed.), *The Cambridge History of Nineteenth-Century Philosophy*. Cambridge University Press, Cambridge (in press).
- Rothenberg, B., 1962, *Studien zur Logik Franz Brentanos*. Frankfurt (Dissertation).
- Schlick, M., 1925, *Allgemeine Erkenntnislehre*. Frankfurt. (Eng. trans. *General Theory of Knowledge*, New York 1974).
- Schmit, R., 1985, ‘Allgemeinheit und Existenz. Zur Analyse des kategorischen Urteils bei Herbart, Sigwart, Brentano und Frege’, *Grazer Philosophische Studien* 23, pp.58-78.
- Simons, P., 1987, ‘Brentano's Reform of Logic’, *Topoi* 6, pp.25-38.
- Terrell, B., 1978, ‘Quantification and Brentano's Logic’, *Grazer Philosophische Studien* 5, pp. 45-65.

Other Internet Resources

- Franz Brentano and his main pupils (Scott Moore, Baylor University)

Related Entries

assertion | negation | predication and instantiation | propositions | psychologism | reism | square of opposition | truth: correspondence theory of

Johannes Brandl
Johannes.Brandl@sbg.ac.at

First published: November 22, 2000
Content last modified: November 22, 2000

Feminist History of Philosophy

The past two decades have seen an explosion of feminist writing on the philosophical canon, a development that has clear parallels in other disciplines like literature and art history. Since most of the writing is, in one way or another, critical of the tradition, a natural question to ask is: Why does the history of philosophy have importance for feminist philosophers? This question assumes that the history of philosophy is of importance for feminists, an assumption that is warranted by the sheer volume of recent feminist writing on the canon. This entry explores the different ways that feminist philosophers are interacting with the Western philosophical tradition.

Feminist philosophers engaged in a project of re-reading and re-forming the philosophical canon have noticed two significant areas of concern. The first is the problem of historical exclusion. Feminist philosophers are faced with a tradition that believes that there are no women philosophers and, if there are any, they are unimportant. Of course, women are not entirely absent from the history of philosophy, and that brings us to the second challenge we face. Canonical philosophers have had plenty to say about women and what we are like. In general terms, we often find that philosophical norms like reason and objectivity are defined in contrast to matter, the irrational or whatever a given philosopher associates with women and the feminine. Our tradition tells us, either implicitly through images and metaphors, or explicitly in so many words, that philosophy itself, and its norms of reason and objectivity, exclude everything that is feminine or associated with women.

In response, feminist philosophers have criticized both the historical exclusion of women from the philosophical tradition, and the negative characterization of women or the feminine in it. Feminist historians of philosophy have argued that the historical record is incomplete because it omits women philosophers, and it is biased because it devalues any women philosophers it forgot to omit. In addition, feminist philosophers have argued that the philosophical tradition is conceptually flawed because of the way that its fundamental norms like reason and objectivity are gendered male.[1] By means of these criticisms, feminist philosophers are enlarging the philosophical canon and re-evaluating its norms, in order to include women in the philosophical "us".

The following entry contains 3 major sections. Section 1 ("Feminist Criticisms of the Canon as Misogynist") describes feminist readings of the philosophical canon that challenge its derogatory characterizations of women. These are of three kinds: (a) readings that record the explicit misogyny of great philosophers (like Aristotle's description of a female as a deformed male); (b) readings that argue for gendered interpretations of theoretical concepts (like matter and form in Aristotle); (c) synoptic interpretations of the canon (like the view that, historically, reason and objectivity are gendered male).

The third category of feminist criticisms of the canon diagnoses where philosophy as a whole went most deeply wrong, and, in doing so, it constructs a negative canon of philosophy. The negative canon exposes the ways in which the views of canonical philosophers throughout the history of philosophy are explicitly or implicitly misogynist or sexist. Section 2 ("Feminist Revisions of the History of Philosophy") discusses the response of feminist philosophy to the myths that there are no women philosophers and, in any case, no important ones. One response has been the retrieval of women philosophers for the historical record. A related development is the elevation to the canon of women philosophers like Mary Wollstonecraft, Hannah Arendt and Simone de Beauvoir. Section 3 ("Feminist Appropriation of Canonical Philosophers") examines the way that feminist philosophers have been engaged in rereading the canon looking for antecedents to feminist philosophy in the work of those philosophers (e.g. Hume) and those theories (e.g. Arisotle's virtue ethics) that are most congenial to current trends in feminism or which provide most fuel for feminist thought. This is to use the canon as other movements have done--as a resource, and as confirmation that a feminist perspective or problem is securely rooted in our philosophical culture.

---

1. Feminist Criticisms of the Canon as Misogynist

> Women are capable of education, but they are not made for activities which demand a universal faculty such as the more advanced sciences, philosophy and certain forms of artistic production. ... Women regulate their actions not by the demands of universality, but by arbitrary inclinations and opinions.
>
> (Hegel 1973: 263)

The idea that the gender of philosophers is important or even relevant to their work is a thought that runs counter to the self-image of philosophy. So, it is interesting to explore how and why feminist philosophers came to the realization that gender is a useful analytic category to apply to the history of philosophy. We can distinguish two aspects to this process although, in many cases, the two aspects

merge into a single project. The first stage of realizing the importance of gender consisted of the cataloguing of the explicit misogyny of most of the canon. The second stage consisted of probing the theories of canonical philosophers in order to uncover the gender bias lurking in their supposedly universal theories. The second stage, the discovery that a philosopher's supposedly universal and objective theories were gender specific, raised the further question of whether or not the theoretical gender bias was intrinsic to the theory or extrinsic to it. Let me illustrate these points with Aristotle.

1.1 Explicit Statements of Misogyny in Philosophical Texts

There is no doubt that Aristotle's texts are misogynist; he thought that women were inferior to men and he said so explicitly. For example, to cite Cynthia Freeland's catalogue: "Aristotle says that the courage of a man lies in commanding, a woman's lies in obeying; that "matter yearns for form, as the female for the male and the ugly for the beautiful;" that women have fewer teeth than men; that a female is an incomplete male or "as it were, a deformity": which contributes only matter and not form to the generation of offspring; that in general "a woman is perhaps an inferior being"; that female characters in a tragedy will be inappropriate if they are too brave or too clever"(Freeland 1994: 145-46). However dispiriting or annoying this litany is, and whatever problems it presents to a woman studying or teaching Aristotle, it can be argued that Aristotle simply held a mistaken view about women and their capacities (as did most Athenians of his time). But, if this is so, then Aristotle's theories, or most of them, are not tarnished by his statements about women, and we can ignore them, since they are false.

I have chosen Aristotle as my example, but similar feminist critiques are available chronicling the explicit misogyny of other canonical figures like Plato and Kant. Feminist criticisms of Plato's theories stress dialogues (like the *Timaeus* and *Laws*) that characterize women as inferior to men rather than the egalitarian *Republic*. Kant's writings, like Aristotle's, provide the ideal target for feminist criticism because they contain both overt statements of sexism and racism, and a theoretical framework that can be interpreted along gender lines.[2]

1.2 Gendered Interpretations of Philosophical Concepts

If we consider Aristotle's theory of hylomorphism we find a connection between form and being male, and matter and being female. That is, we find that matter and form are gendered notions in Aristotle (Witt 1998). By a gendered notion I mean a notion that is connected either overtly or covertly, either explicitly or metaphorically with gender or sexual difference.[3] Furthermore, matter and form are not equal partners in Aristotle's metaphysics; form is better than matter. And since hylomorphism is the conceptual framework that underlies most of Aristotelian theory from metaphysics and philosophy of mind to biology and literary theory, it looks as if his supposedly universal and objective theories are gendered, and it looks as if his negative charactization of women tarnishes his philosophical theories.

Are Aristotle's theories intrinsically gendered and sexist, so that gender cannot be removed without altering the theories themselves? Several feminist philosophers have developed this thesis. For example, in "Woman Is Not a Rational Animal", Lynda Lange argues that Aristotle's theory of sex difference is

implicated in every piece of Aristotle's metaphysical jargon, and she concludes that "it is not at all clear that it [Aristotle's theory of sex difference] can simply be cut away without any reflection on the status of the rest of the philosophy"(Harding and Hintikka 1983: 2). Elizabeth Spelman has argued that Aristotle's politicized metaphysics is reflected in his theory of soul, which, in turn, is used to justify the subordination of women in the Politics (Spelman 1983). And, finally, Susan Okin has argued that Aristotle's functionalist theory of form was devised by Aristotle in order to legitimate the political status quo in Athens, including slavery and the inequality of women (Okin 1979: ch. 4).

If these scholars are right, then Aristotle's theories are intrinsically biased against women, and it is unlikely that they can have any value for feminists beyond the project of learning about the ways in which the philosophical tradition has devalued women. Alternatively, I have argued that the suspect gender associations with Aristotelian matter and form are extrinsic tothese concepts, and therefore removable from Aristotle's theories without substantially altering them (Witt 1998). The argument that Aristotle's gender associations are not intrinsic to his concepts of matter (female) and form (male) turns on the incompatibility of the position that matter is intrinsically female, and form intrinsically male with the position that every composite substance is a unity of matter and form. If every composite substance is a complex of matter and form, then each would be a hermaphrodite, rather than a male or a female as is the case with animals. Moreoever,whatever plausibility gender associations with matter and form might have with regard to animals, is lost entirely when we consider artifacts, like shoes and beds. If intrinsic gender associations with matter and form are incompatible with Aristotle's theory of substance and extrinsic gender associations are compatible with that theory, then the principle of charity dictates that we opt for the consistent interpretation.

Sometimes, as in the case of Descartes, the feminist argument in favor of a gendered theory is subtle since, unlike Aristotle, he expresses both a personal and a theoretical committment to equality. Further, his theories are not stated using gendered notions. Yet, some feminists have argued that his theory of mind-body dualism, and his abstract characterization of reason resonate with gender implications-- on the assumption that women are emotional and bodily creatures (e.g. Scheman 1993; Bordo 1987; Lloyd 1993b, ch. 3).

1.3 Synoptic Interpretations of the Philosophical Canon

The philosophical canon can allow the luster some of its members to be tarnished by feminist criticism, just as it has weathered criticisms from analytic or continental perspectives. The most radical feminist critics, however, have urged that the canon's central philosophical norms and values, like reason and objectivity, are gendered notions. The synoptic approch considers the Western philosophical tradition as a whole, and argues that its core concepts are gendered male. But, if this is so, then the Western philosophical tradition as a whole, and the central concepts that we have inherited from it, requires critical scrutiny by feminists. Moreover, philosophy's self-image as universal and objective, rather than particular and biased, is mistaken.

Feminist synoptic interpretations of the canon take several forms. The first, exemplified by Genevieve

Lloyd's *Man of Reason*, argues that reason and objectivity in the history of philosophy are gendered male.[4] The way that reason and objectivity are gendered male varies as philosophical theory and historical period varies, but the fact that they are gendered is a constant. From Aristotle to Hume, from Plato to Sartre, reason is associated with maleness. Therefore, the notion of reason that we have inherited, whether we are empiricists or existentialists, requires critical scrutiny.

The second form of synoptic interpretation, exemplified by Susan Bordo's *The Flight to Objectivity*, argues that the modern period in philosophy, and, in particular, the philosophy of Descartes, is the source of our ideals of reason and objectivity that are gendered male. In other words, this story chronicles a turn in philosophy coincident with the rise of modern science, which generated ideals of reason and objectivity that are deeply antagonistic to women and feminism.[5] Cartesian rationalism and the norms of modern science mark a decisive break with a philosophical and cultural tradition that was more accommodating of female characteristics and powers.

It is important to note that Lloyd and Bordo differ not only with regard to the historical story they tell concerning the maleness of reason, but also with regard to the way they understand that maleness. For Lloyd, the maleness of reason is symbolic and metaphorical rather than cultural or psychological. Lloyd does not intend the maleness of reason to refer to either a socially constituted gender category or a psychological orientation shared by males. "This book is not a direct study of gender identity. It seeks rather to contribute to the understanding of how the male-female distinction operates as a symbol in traditional philosophical texts, and of its interactions with explicit philosophical views of reason".[6] In understanding the maleness of reason as symbolic rather than as psychological or social, Lloyd avoids making a theoretical committment to any particular psychology of sex differences or any particular account of the social formation of gender identity. What she gains in flexibility, however, she loses in content, since it is difficult to specify exactly what metaphorical maleness is, and how it is related to psychological or social maleness. Other feminists have attempted to develop an account of how male metaphors and symbols undermine philosophical arguments (Rooney 1991).

For Bordo, however, the maleness of Cartesian reason is given both a social meaning and a psychological content. First, the social meaning of maleness: "In the seventeenth century it [the feminine orientation toward the world] was decisively purged from the dominant intellectual culture, through the Cartesian ‘rebirthing’ and restructuring of knowledge and the world as masculine"(Bordo 1987: 100). This social meaning is paired with a psychological consequence: "The ‘great Cartesian anxiety,’ although manifestly expresssed in epistemological terms, discloses itself as anxiety over separation from the organic female universe"(Bordo 1987: 5). Cartesian ‘anxiety’ is separation anxiety from mother nature; the rational norms of clarity and distinctness are read as symptoms of this anxiety.[7] Bordo's social-psychological notion of maleness while rich and explicit, provides a large target for critics because it is based on a controversial historical thesis (that the 17th century showed a marked increase in gynophobia) and a disputed psychological theory of the family (Object Relations Theory).

Luce Irigaray takes a radical stance towards the history of philosophy by trying to indicate what is suppressed and hidden in the tradition rather than cataloguing its evident "maleness". Her work, like

Bordo's, makes use of psychoanalytic theory in interpreting texts and, like Lloyd's, it explores the symbolic associations of philosophical images and concepts. However, unlike Bordo and Lloyd, Irigaray uses highly unconventional methods of interpreting canonical philosophical texts in order to uncover the ways in which the feminine or sexual difference is repressed in them. For example, Irigaray uses humor and parody rather than straightforward exegesis, and she points to instabilities (contraditions) in philosophical texts as symptoms of patriarchal thinking. According to Irigaray, patriarchal thinking attempts to achieve universality by repressing sexual difference. But, the presence of contradictions or instabilities in a philosophical text is symptomatic of the failure of patriarchal thinking to contain sexual difference. For example, Irigaray might look at the argument I described above for considering gender associations with form and matter in Aristotle to be extrinsic, rather than intrinsic, to those concepts, and argue that the fact that Aristotle's hylomorphism as a universal theory is incompatible with gender associations is a symptom of patriarchal thinking rather than evidence that the proposed interpretation is mistaken.[8]

Despite their different historical stories, and the different ways that they understand the maleness of reason, each of these panoramic visions of the history of philosophy deliver the same moral, which is that the central norms that inform our philosophical culture today are gendered male.[9] Hence, these synoptic narratives of the philosophical tradition provide historical justifications for feminist philosophers who are critical of our central philosophical norms of reason and objectivity. Does the feminist synoptic critical reading of the history of philosophy justify either the conclusion that traditional conceptions of reason ought to be flat-out rejected by feminists or the conclusion that traditional conceptions of reason ought to be subjected to critical scrutiny?

Even if feminist historical arguments are successful in showing that philosophical norms like reason and objectivity are gendered male, this conclusion does not justify a flat-out rejection of either traditional philosophy or its norms of reason and objectivity (Witt 1993). Recall the distinction introduced above between intrinsically and extrinsically gendered notions. An intrinsically gendered notion is one that necessarily carries implications regarding gender, i.e., if one were to cancel all implications concerning gender, one would be left with a different notion than the original. In contrast, an extrinsically gendered notion typically does carry implications concerning gender, but not necessarily so. If the maleness of reason is extrinsic to the traditional concept of reason, then the historical fact that it was a gendered notion does not justify or require its rejection by feminists. If on the other hand, it can be shown that the maleness of reason is intrinsic to it, it still does not follow that reason ought to be rejected by feminists. For, the idea the reason is intrinsically male-biased would justify a rejection of it only if it ought to be other than it is. So, what needs to be argued is that reason and objectivity would be different, and better, if they were not gendered male, but were gender-neutral, gender-inclusive or female. But, if feminist philosophers develop this argument, which they need to buttress the historical argument, then they are reconceptualizing traditional notions of reason and objectivity rather than rejecting them.

Even though the work that feminist philosophers have done to show the ways in which traditional conceptions of reason and objectivity are associated with maleness falls short of justifying their rejection, their work has been valuable in two respects. First, it has established that gender is associated with the central norms of philosophy, a conclusion that warrants attention from anyone attempting to understand

our philosophical tradition. Second, the historical studies raise questions about reason and objectivity that are valuable areas of inquiry for contemporary philosophers.

2. Feminist Revisions of the History of Philosophy

> These women are not women on the fringes of philosophy, but philosophers on the fringes of history.
>
> ---Mary Ellen Waithe

Feminist canon revision is most distinctive, and most radical, in its retrieval of women philosophers for the historical record, and in its placement of women in the canon of great philosophers. It is a distinctive project because there is no comparable activity undertaken by other contemporary philosophical movements, for whom canon creation has been largely a process of selection from an already established list of male philosophers. It is a radical project because by uncovering a history of women philosophers, it has destroyed the alienating myth that philosophy was, and by implication is or ought to be, a male preserve.

In *A History of Women Philosophers* Mary Ellen Waithe has documented at least 16 women philosophers in the classical world, 17 women philosophers from 500-1600, and over 30 from 1600-1900.

And, in the recent feminist series *Re-reading the Canon* three of the fourteen canonical philosophers are women: Mary Wollstonecraft, Hannah Arendt and Simone de Beauvoir. What is crucial to understand is that none of the three is canonical--if by that you mean included in the history of philosophy as it is told in philosophy department curricula, in histories of philosophy, and in scholarly writing.

Indeed, *The Encyclopedia of Philosophy*, published in 1967, which contains articles on over 900 philosophers, does not include an entry for any of them. Moreover, if the index is to be believed, de Beauvoir and Wollstonecraft are not mentioned at all in any article, and Hannah Arendt merits a single mention in an article on "Authority". Far from being canonical, these women philosophers are scarcely even marginal, warranting perhaps a passing reference in a survey of existentialism or political philosophy, but little more.[10] Hence, the feminist series *Re-reading the Canon* is not only engaged in a critical re-reading of canonical figures like Plato and Hegel, but is also, by fiat, changing the contours of the canon.

The project of retrieving women philosophers has a paradoxical relationship with contemporary feminist theory, however. On the one hand, it is clearly a feminist project; its originators were interested in establishing that women have been philosophers throughout the history of the discipline despite their routine omission from standard histories and encyclopedias of philosophy. However, the newly - recovered women philosophers suggest that there is little overlap among three groups: women philosophers, feminine philosophers and feminist philosophers. For most of the newly discovered women philosophers were not feminist thinkers nor did they write philosophy in a feminine voice, different from

their male counterparts. Indeed, their breadth of philosophical interests is comparable to that of male philosophers although their domain of application sometimes differs. In her introduction to *A History of Women Philosophers* Mary Ellen Waithe comments "If we except the Pythagorean women, we find little differences in the ways men and women did philosophy. Both have been concerned with ethics, metaphysics, cosmology, epistemology and other areas of philosophic inquiry"(Waithe 1987-1991 Vol. 1: xxi). And another editor, Mary Warnock, comments "In the end, I have not found any clear "voice" shared by women philosophers (Warnock 1996: xlvii). The women philosophers restored to the tradition by feminist hands are not all proto-feminists nor do they speak in a uniform, and different, voice from their male peers.

Similarly, women philosophers who are candidates for initiation into the philosophical canon--like Mary Wollstonecraft, Hannah Arendt and Simone de Beauvoir--are a diverse crew. According to Elizabeth Young-Bruehl "That Hannah Arendt should have become a provocative subject for feminists is startling" presumably because of Arendt's explicit criticism of feminism. And while Wollstonecraft and de Beauvoir were both feminists, they share neither a common philosophical voice nor common philosophical principles. In *The Vindication of the Right's of Women* Wollstonecraft argued for the education of women using Enlightenment principles, while Beauvoir's *The Second Sex* reflects her marxist and existentialist roots.

The diversity of women philosophers raises the question why their recovery or re-valuation is an important project for contemporary feminist theory. What the retrieval of women philosophers, and their inclusion in the philosophical canon has done is to challenge the myth that there are no women in the history of philosophy and the fallback position that if there are any women philosophers, they are unimportant. Lovers of wisdom that we all are, we all benefit from the correction of these mistaken beliefs. Moreover, as feminists, we are interested in correcting the effects of discrimination against women philosophers, who were written out of history, unfairly, because of their gender not their philosophical ideas.

However, what is really at issue is not philosophy's past, but its present; its self-image as male. That self-image is created and maintained in part by a tacit historical justification. It is a damaging self-image for women philosophers today, and for women who aspire to be philosophers. The real significance of uncovering the presence of women in our history, and in placing women in our canon is the effect that has on the way we think about the "us" of philosophy.

3. Feminist Appropriation of Canonical Philosophers

Feminist philosophers have also changed the history of philosophy by appropriating its ideas for feminist purposes. From the perspective of negative canon formation, the history of philosophy is a resource only in so far as it describes the theories and thinkers that were most deeply mistaken about women. Other feminist historians of philosophy have found important resources for feminism in canonical philosophers.

Indeed, they have found valuable concepts even in the worst offenders of the negative canon, like Aristotle and Descartes.

For example, in *The Fragility of Goodness* Martha Nussbaum has described the virtues of an Aristotelian ethics with its emphasis on the importance of concrete context, emotion and care for others in an ethical life (Nussbaum 1986). And Marcia Homiak has argued that Aristotle's rational ideal, far from being antithetical to feminists, actually captures some of feminism's deepest ethical insights (Homiak 1993). With regard to Descartes, Margaret Atherton has argued that his concept of reason was interpreted in egalitarian rather than masculinist terms by several women philosophers of the 18th century, and was used in their arguments for equal education for women.[11]

Other feminists have urged the reconsideration of the views of canonical figures, like Hume and Dewey, who have played only a minor role in the negative feminist canon. For example, Annette Baier has argued at length for the value of a Humean perspective in both epistemology and in ethics for feminist theory (Baier 1987; Baier 1993). And, in *Pragmatism and Feminism* Charlene Seigfried argues for the value of pragmatism for feminism; a position also taken by Richard Rorty (Seigfried 1996; Rorty 1991).

It is interesting to note that some of the very same philosophers who were cast as the villains of the negative canon are also mined by feminist theorists for useful ideas. Indeed, it is likely that every philosopher, from Plato to Nietzsche, who has been condemned to the negative canon also appears in some feminist's positive canon. This is perplexing. After all, if feminists evaluate canonical texts so differently, it raises questions about the coherence of feminist interpretations of texts. Is Aristotle a feminist hero or villain? Are Descartes' ideas dangerous for feminists or useful to them? If feminists have argued both positions, we begin to suspect that there is no such thing as a feminist interpretation of a philosopher. And this might lead us to wonder about the coherence and unity of the project of feminist canon revision.

Why is it that feminist philosophers have reached different, and even sometimes incompatible interpretations of the history of philosophy? In my view, the multiple and contrary readings of the philosophical canon by feminists reflects the contested nature of the "us" of contemporary feminism. The fact that feminist interpretations of canonical figures is diverse reflects, and is a part of, on-going debates within feminism over its identity and self-image. Disagreements among feminist historians of philosophy over the value of canonical philosophers, and the appropriate categories to use to interpret them, are, in the final analysis, the result of debate within feminist philosophy over what feminism is, and what is theoretical committments should be, and what its core values are.

Bibliography

Comprehensive Bibliography

Supplementary Document:

Bibliography of Feminist Philosophers Writing about the History of Philosophy [by Abigail Gosselin]

References

- Amorós, Celia. 1994. "Cartesianism and Feminism. What Reason Has Forgotten; Reasons for Forgetting" in Hypatia vol. 9, no. 1 (Winter 1994).
- Antony, Louise, Witt, Charlotte ed. 1993. *A Mind of One's Own: Feminist Essays on Reason and Objectivity* (Westview Press).
- Atherton, Margaret. 1993. "Cartesian Reason and Gendered Reason" in Antony and Witt 1993.
- _______. 1994. *Women Philosophers of the Early Modern Period* (Hackett Publishing Co.).
- Baier, Annette. 1987. "Hume, the Women's Moral Theorist" *Women and Moral Theory* ed. Eva Feder Kittay and Diana T. Meyers (Roman and Littlefield)
- _______. 1993. "Hume, the Reflective Women's Epistemologist?" in A Mind of One's Own.
- Bar On, Bat-Ami. 1994. *Engendering Origins: Critical Feminist Readings in Plato and Aristotle* (Albany SUNY Press).
- _______. 1994. *Modern Engendering: Critical Feminist Readings in Modern Western Philosophy* (Albany SUNY Press).
- Bordo, Susan R. 1987. *The Flight to Objectivity: Essays on Cartesianism and Culture* (State University of New York Press).
- Butler, Judith. 1990. *Gender Trouble* (Routledge 1990).
- Cornell, Drucilla, 1993. *Transformations* (New York 1993).
- Deutscher, Penelope, 1997. *Yielding Gender: Feminism, Deconstruction and the History of Philosophy* (London and New York: Routledge).
- Falco, Maria J., ed. 1996. *Feminist Interpretations of Mary Wollstoncraft* (The Pennsylvania State University Press).
- Freeland, Cynthia, 1994. "Nourishing Speculation: A Feminist Reading of Aristotelian Science" in *Engendering Origins: Critical Feminist Readings in Plato and Aristotle* ed. By Bat-Ami Bar On (State University of New York Press, Albany).
- _______, ed. 1998. *Feminist Interpretations of Aristotle* (The Pennsylvania State University Press)
- Frye, Marilyn, 1992. *The Possibility of Feminist Theory* (California, The Crossing Press).
- Fuss, Diana. 1989. *Essentially Speaking* (Routledge).
- Harding, Sandra and Hintikka, Merrill eds. 1983. *Discovering Reality* (Dordrecht, Holland D. Reidel).
- _______. 1986. *The Science Question in Feminism* (Cornell University Press).
- Hegel, G.W. F.1973. *The Philosophy of Right*, T.M. Knox (trans) (New York, Oxford University Press).
- Hekman, Susan, ed. 1996. *Feminist Interpretatons of Michel Foucault* (The Pennsylvania State University Press)
- Holland, Nancy J., ed. 1997 *Feminist Interpretations of Jacques Derrida* (The Pennsylvania University State Press).

- Homiak, Marcia. 1993. "Feminism and Aristotle's Rational Ideal" in Antony and Witt 1993.
- Honig, Bonnie, ed. 1995. *Feminist Interpretations of Hannah Arendt* (The Pennsylvania State University Press)
- Keller, Evelyn Fox, 1985. *Reflections on Gender and Science* (Yale University Press)
- Leon, Celine and Walsh Sylvia, ed., 1997. *Feminist Interpretations of Soren Kierkegaard* (The Pennsylvania State University Press)
- Lloyd, Genevieve. 1993a. "Maleness, Metaphor, and the 'Crisis' of Reason" in Antony and Witt 1993.
- ________.1993b. *The Man of Reason: "Male" and "Female" in Western Philosophy* (The University of Minnesota Press, Minneapolis).
- Mills, Patricia Jagentowicz, ed. 1996 *Feminist Interpretations of G.W.F.Hegel* (The Pennsylvania State University Press).
- Nussbaum, Martha. 1986. *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy* (Cambridge UP)
- Okin, Susan Moller, 1979. *Women in Western Political Thought* (Princeton)
- O'Neill, Eileen, 1978. "Disappearing Ink: Early Modern Women Philosophers and Their Fate in History" in Kourany, Janet, ed. *Philosophy in a Feminist Voice: Critiques and Reconstructions* (Princeton University Press)
- Rooney, Phyllis. 1991. "Gendered Reason: Sex, Metaphor and Conceptions of Reason" by Phyllis Rooney in *Hypatia* 6:2 (Summer 1991): 77-103.
- _______.1994. "Recent Work in Feminist Discussions of Reason." *American Philosophical Quarterly* Vol. 31, Number 1, (January 1994).
- Rorty, Richard. 1991. "Feminism and Pragmatism" *Michigan Quarterly Review* 30/2 (Spring 1991) 231-58
- Rorty, Richard, Schneewind, J.B., Skinner, Quentin eds. 1984. *Philosophy in History* (Cambridge, Cambridge University Press).
- Scheman, Naomi. 1993. "Though This Be Method, Yet There Is Madness in It: Paranoia and Liberal Epistemology" in Antony and Witt 1993.
- Schott, Robin. 1988. *Eros and Cognition* (Boston).
- Seigfried, Charlene Haddock. 1996. *Pragmatism and Feminism* (Chicago & London, University of Chicago Press).
- Simons, Margaret A., ed. 1995. *Feminist Interpretations of Simone de Beauvoir* (The Pennsylvania State University Press).
- Soper, Kate, 1995. *What is Nature?: Culture, politics and the non-human* (Oxford, Blackwell).
- Spelman, Elizabeth, 1983. "Aristotle and the Politicization of the Soul" in Harding and Hintikka 1983.
- _______. 1988. *Inessential Woman* (Boston).
- Tuana, Nancy, 1992. *Woman and the History of Philosophy* (New York, Paragon Press).
- _______. ed. 1994. *Feminist Interpretations of Plato* (The Pennsylvania State University Press).
- Waithe, Mary Ellen ed. 1987-1991. *A History of Women Philosophers* Vol. 1-3 (Kluwer Academic Publishing).
- Warnock, Mary ed. 1996. *Women Philosophers* (London J.M. Dent).
- Witt, Charlotte. 1993. "Feminist Metaphysics" in Antony and Witt 1993.

- ________.1998. "Form, Normativity and Gender in Aristotle: A Feminist Perspective" in Freeland 1998.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

feminism, approaches to | feminism, history of

Charlotte Witt
University of New Hampshire
cewitt@cisunix.unh.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: November 3, 2000
Content last modified: November 3, 2000

Stanford Encyclopedia of Philosophy
Notes to Feminist History of Philosophy

Notes

1. Phyllis Rooney summarizes feminist criticisms of the "maleness" of reason in (Rooney 1994).

2. Kant's derogatory remarks about women are in his pre-critical work *Observations on the Beautiful and Sublime* p. 111 (Goldthwaite ed.). Robin Schott presents a critical, feminist reading of Kant's conceptual framework in (Schott 1988).

3. Feminists use the idea of a gendered notion to mean different things. In this article I draw a distinction between holding that a notion is intrinsically gendered, and holding that it is extrinsically gendered. Other feminists, however, use the idea of a gendered notion in ways that do not map easily onto my distinction. For example, as I discuss below, Genevieve Lloyd argues for the symbolic gendering of philosophical notions, and it may be that her interpretation does not map onto the intrinsic/extrinsic distinction. And Sally Haslanger argues that objectivity is a gendered notion in the philosophy of Catharine MacKinnon in a way that is neither extrinsic nor intrinsic.

4. (Tuana 1992) also provides a feminist reading of the history of philosophy.

5. (Keller 1985) is a classic source for feminist criticism of the rise of modern science.

6. (Lloyd 1993b, ix). In another essay Lloyd explains that in her view there is an important connection between gender metaphors in philosophical texts and real world gender divisions and the way that gender identity is formed in a culture. See (Lloyd 1993a).

7. In (Scheman 1993), Naomi Scheman develops parallels between the Cartesian subject--disembodied, rational, and unitary--and the repression and projection characteristic of paranoia. Her analysis, like Bordo's, makes use of psychoanalytic categories but it does not make a historical, cultural claim as Bordo does.

8. For a discussion of Irigaray's contribution to feminist scholarship on the history of philosophy see (Deutscher 1997).

9. Bordo and Lloyd differ in other important respects as well. Bordo is interested in providing a social and psychological explanation for the masculinization of philosophy by Descartes. Why did Descartes conceive of reason and objectivity in a masculine guise? The social answer is that during his life

European culture was undergoing a gynophobic spasm. The psychological answer depends upon object relations theory, and the development of that theory along gender lines by Chodorow and others. Lloyd is not interested primarily in the causal question addressed by Bordo. Moreover, she thinks that the maleness of reason in the philosophical tradition is primarily symbolic or metaphorical rather than social or psychological. Ultimately, then Bordo and Lloyd differ as to what is meant by the maleness of reason.

10. For a discussion of the omission of Simone de Beauvoir from the philosophical canon see the introduction to (Simons 1995).

11. 11. (Atherton 1993). Celia Amorós has retrieved the arguments in support of the equality of the sexes made by the 17th century Cartesian philosopher Francois Poullain de la Barre (1647-1723). See (Amorós 1994).

Charlotte Witt
University of New Hampshire
cewitt@cisunix.unh.edu

First published: November 3, 2000
Content last modified: November 3, 2000

Stanford Encyclopedia of Philosophy
Supplement to Feminist History of Philosophy

Bibliography of Feminist Philosophers

General

Books

Antony, Louise and Charlotte Witt, eds. 1993. *A Mind of One's Own: Feminist Essays on Reason and Objectivity* (Boulder: Westview Press)

Bar On, Bat-Ami, ed. 1994. *Modern Engendering: Critical Feminist Readings in Modern Western Philosophy* (Albany: SUNY Press)

Coole, Diana H., 1988. *Women in Political Theory: From Ancient Misogyny to Contemporary Feminism* (Sussex: Wheatsheaf Books)

Deutscher, Penelope, 1997. *Yielding Gender: Feminism, Deconstruction, and the History of Philosophy* (London and New York: Routledge)

Elshtain, Jean Bethke, ed. 1982. *The Family in Political Thought* (Amherst: University of Massachusetts Press)

Gatens, Moira, 1991. *Feminism and Philosophy: Perspectives on Difference and Equality* (Bloomington: Indiana University Press)

Gould, Carol C. and Marx W. Wartofsky, eds. 1976. *Women and Philosophy: Toward a Theory of Liberation* (New York: G. P. Putnam's Sons)

Grimshaw, Jean, 1986. *Feminist Philosophers: Women's Perspectives on Philosophical Traditions* (Brighton: Wheatsheaf Books)

Harding, Sandra and Merrill B. Hintikka, eds. 1983. *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. (Dordrecht: Reidel)

Jones, Gregory L. and Stephen E. Fowl, eds., 1995. *Rethinking Metaphysics* (Cambridge: Blackwell)

Keller, Evelyn Fox, 1985. *Reflections on Gender and Science* (New Haven: Yale University Press)

Lloyd, Genevieve, 1993. *The Man of Reason: "Male" and "Female" in Western Philosophy* (Minneapolis: University of Minnesota Press)

Mahowald, Mary, 1983. *The Philosophy of Woman* (Indianapolis: Hackett)

McAlister, Linda Lopez, ed. 1996. *Hypatia's Daughters: Fifteen Hundred Years of Women Philosophers* (Bloomington: Indiana University Press)

Menage, Gilles, 1984. *The History of Women Philosophers* Trans. Beatrice H. Zedler (Lanham: University Press of America)

Moscovici, Claudia, 1996. *From Sex Objects to Sexual Subjects* (New York: Routledge)

Nye, Andrea, 1988. *Feminist Theory and the Philosophies of Man* (London: Croom Helm)

Okin, Susan Moller, 1979. *Women in Western Political Thought* (Princeton: University of Princeton Press)

Rorty, Richard, J. B. Schneewind, and Quentin Skinner, eds. 1984. *Philosophy in History* (Cambridge: Cambridge University Press)

Soper, Kate, 1995. *What is Nature? Culture, Politics, and the Non Human* (Oxford: Blackwell)

Spelman, Elizabeth, 1988. *Inessential Woman: Problems of Exclusion in Feminist Thought* (Boston: Beacon Press))

Tuana, Nancy, 1993. *The Less Noble Sex: Scientific, Religious, and Philosophical Conceptions of Woman's Nature* (Bloomington: Indiana University Press)

Tuana, Nancy, 1992. *Woman and the History of Philosophy* (New York: Paragon House)

Waithe, Mary Ellen, ed. 1987-1991. *A History of Women Philosophers, Vol. 1-3* (Kluwer Academic Publishing)

Warnock, Mary, ed. 1996. *Women Philosophers* (London: J. M. Dent)

Warren, Mary Anne, 1980. *The Nature of Woman: An Encyclopaedia and Guide to the Literature* (Reyes CA: Edgepress)

Articles

Atherton, Margaret, "Doing the History of Philosophy as a Feminist," *American Philosophical Association Newsletteron Feminism and Philosophy* Ed. L. Antony and D. Meyers.

Bell, Linda A., "Gallantry: What it is and Why it Should Not Survive," *Southwestern Journal of Philosophy* 22 (1984), 165-174.

Code, Lorraine, "Simple Equality is Not Enough," *Australasian Journal of Philosophy* Supp. 64 (1986)

Lloyd, Genevieve, "The Man of Reason," *Metaphilosophy* 10, 1 (January 1979), 18-37.

Olkowski, Dorothea, "Materiality and Language: Butler's Interrogation of the History of Philosophy," *Philosophy and Social Criticism* 23, 3 (1997), 37-53.

Tuana, Nancy, "The Weaker Seed: The Sexist Bias of Reproductive Theory," *Hypatia* 3 (1988), 35-59.

Wolff, Robert Paul, "There's Nobody Here But Us Persons," in *Women and Philosophy: Toward a Theory of Liberation* Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Ancient

Books

Archer, L. S. Fischler and M. Wyke, eds. 1994. *Women in Ancient Socieities* (London: Routledge)

Bar-On, Bat Ami, ed. 1994. *Engendering Origins: Critical Feminist Readings in Plato and Aristotle* (Albany: SUNY Press)

DuBois, Page, 1998. *Sowing the Body: Psychoanalysis and Ancient Representations of Women* (Chicago: University of Chicago Press)

Mahowald, Mary Briody, ed. 1983. *Philosophy of Woman: An Anthology of Clasic and Current Concepts, Second Edition* (Indianapolis: Hackett)

Peradotto, J. and J. P. Sullivan, eds. 1984. *Women in the Ancient World: The Arethusa Papers* (Albany: SUNY Press)

Pomeroy, Sarah, 1975. *Goddesses, Whores, Wives, and Slaves: Women in Classical Antiquity* (New York: Schocken Books)

Pomeroy, Sarah, 1984. *Women in Hellenistic Egypt: From Alexander to Cleopatra* (New York: Schocken Books)

Rabinowitz, Nancy, 1993. *Anxiety Veiled: Euripides and the Traffic in Women* (Ithaca: Cornell University Press)

Rabinowitz, Nancy, 1993. *Feminist Theory and the Classics* (New York and London: Routledge)

Snyder, Jane, 1988. *The Women and the Lyre: Women Writers in Classical Greece and Rome* (Carbondale: South Illinois University Press)

Ward, Julie K., 1996. *Feminism and Ancient Philosophy* (New York and London: Routledge)

Wright, F. A., 1969. *Feminism in Greek Literature: From Homer to Aristotle*. (Port Washington: Kennikat Press)

Articles

Arthur, Marilyn, "Early Greece: The Origin of the Western Attitude Toward Women" in *Women and the Ancient World: The Arethusa Paper*, J. Perradotto and J. P. Sullivan, eds. (Albany: SUNY Press, 1984)

Asmis, Elizabeth, "The Stoics on Women," in *Feminism and Ancient Philosophy*, Julie K. Ward, ed. (New York and London: Routledge, 1996)

Brumbaugh, Robert and John Burnham, "Coins and Classical Philohophy," *Teaching Philosophy* 12 (1989), 243-255.

Gottner Abendroth, Heide, 1991. *The Dancing Goddess: Principles of a Matriarchal Aesthetic*, Maureen T. Krause, trans. (Boston: Beacon Press)

Hawkesworth, Mary E., "Re/Vision: Feminist Theory Confronts the Polis," *Social Theory and Practice* 13 (1987), 155-186.

Katz, Marilyn, "Ideology and ‘The Status of Women’ in Ancient Greece," *History and Theory* 31, 4 (1992), 70-97.

Kotzin, Rhoda Hadassah, "Ancient Greek Philosophy" in *A Companion to Feminist Philosophy,* Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Molinaro, Ursule, "A Christian Martyr in Reverse Hypatia: 370-415 A.D.," *Hypatia* 4 (1989), 6-8.

Nussbaum, Martha, "Therapeutic Arguments and Structures of Desire," in *Feminism and Ancient Philosophy*, Julie K. Ward (New York and London: Routledge, 1996)

Perez-Estevez, Antonio, "Feminidad Y Racionalidad En El Pensamiento Griego," *Rev. Filosof (Venezuela)* 9 (1986), 167-199. (Spanish)

Skinner, Marilyn, ed. "Rescuing Creusa: New Methodological Approaches to Women in Antiquity," *Special Issue of Helios* 13, 2 (1987)

Smith, Nicholas, "Plato and Aristotle on the Nature of Women," *Journal of the History of Philosophy* 21 (1983), 467-478.

Spelman, Elizabeth, "Anger and Insubordination," in *Beyond Domination: New Perspectives on Women and Philosophy*, Carol Gould, ed. (Totowa NJ: Rowman & Allanheld, 1984)

Spelman, Elizabeth V., "Woman as Body: Ancient and Contemporary View," *Feminist Studies* 8 (1982), 109-131.

Thompson, Patricia J., "Re-Claiming Hestia: Goddess of Everyday Life," *Philosophy in the Contemporary World* 3, 4 (1996), 20-28.

Wartenberg, Thomas E., "Teaching Women Philosophy," *Teaching Philosophy* 11 (1988), 15-24.

Whitbeck, Caroline, "Theories of Sex Difference," in *Women and Philosophy: Toward a Theory of Liberation*, Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Wider, Kathleen, "Women Philosophers in the Ancient World: Donning the Mantle," *Hypatia* 1 (1986), 21-62.

Wiseman, Mary Bittner, "Beautiful Exiles in Aesthetics," in *Aesthetics in Feminist Perspective*, Hilde Hein, ed. (Bloomington: Indiana University Press, 1993)

Plato

Books

Bluestone, Natalie Harris, 1987. *Women and the Ideal Society: Plato's* Republic *and Modern Myths of Gender* (Amherst: University of Massachusetts Press)

Tuana, Nancy, ed. 1994. *Feminist Interpretations of Plato* (University Park: Pennsylvania University Press)

Articles

Allen, Christine Garside, "Plato on Women," *Feminist Studies*, 2, 2-3 (1975), 132

Annas, Julia, "Plato's *Republic* and Feminism," *Philosophy*, 51 (1976), 309. Reprinted in *Feminism and Ancient Philosophy*, Julie K. Ward (New York and London: Routledge, 1996)

Bowery, Anne-Marie, "Diotima Tells a Story: A Narrative Analysis of Plato's 'Symposium,'" in *Feminism and Ancient Philosophy*, Julie K. Ward (New York and London: Routledge, 1996)

Bowery, Anne-Marie, "Plato Visits Postmodernity," *Southwest Philosophy Review* 11 (1995), 135-142.

Brown, Wendy, "'Supposing Truth Were a Woman': Plato's Subversion of Masculine Discourse," *Political Theory* 16 (1988), 594-616.

Calvert, Brian, "Plato and the Equality of Women," *Phoenix*, 29, 3 (1975)

Cappelletti, Angel J., "Sobre El Feminismo De Platon," *Revista de Filosofio (Venezuela)* 12 (1980), 87-96. (Spanish)

Darling, John, "Are Women Good Enough: Plato's Feminism Re-Examined," *Journal of Philosophy in Education* 20 (1986), 123-128.

De Pater, W. and W. Van Langendonck, "Natuurlijkheid Van De Taal En Iconiciteit: Plato En Hedendaagse Taaltheorieen," *Tijdschr Filosof* 51 (1989), 256-297. (Dutch/Flemish)

Dickason, Anne, "Anatomy and Destiny: The Role of Biology in Plato's Views of Women," *The Philosophical Forum*, V (Fall-Winter 1973-1974). Reprinted in *Women and Philosophy: Toward a Theory of Liberation*, Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Fortenbaugh, W. W., "On Plato's Feminism in 'Republic V,'" *Apeiron*, IX, 2 (1975)

Freeman, Barbara, "(Re)writing Patriarchal Texts: The Symposium," in *Postmodernism and Continental Philosophy*, Hugh J. Silverman and Donn Welton, eds. (Albany: SUNY Press, 1988)

Genova, Judith, "Feminist Dialectics: Plato and Dualism," in *Engendering Origins*, Bat-Ami Bar On, ed. (Albany: SUNY Press, 1994)

Gould, Timothy, "Intensity and its Audiences: Notes Towards a Feminist Perspective," *Canadian Journal of Philosophy* 12 (1982), 287-302.

Hampton, Cynthia, "Overcoming Dualism: The Importance of the Intermediate in Plato's *Philebus*," in *Engendering Origins*, Bat-Ami Bar On, ed. (Albany: SUNY Press, 1994)

Hawthorne, Susan, "Diotima Speaks Through the Body," in *Engendering Origins* Bat-Ami Bar On, ed. (Albany: SUNY Press, 1994)

Irigaray, Luce, "Sorcerer's Love: A Reading of Plato's ‘Symposium,’" Trans. Eleanor H. Kuykendall, in *Feminism and Philosophy: Essential Readings in Theory, Reinterpretation, and Application* (Boulder: Westview Press, 1995)

Jacobs, William, "Plato on Female Emancipation and the Traditional Family," *Apeiron* 12 (1978), 24-31.

Joo, Maria, "The Platonic ‘Eros’ and Its Feminist Interpretations," *Magyar Filozofiai Szemle* 1-2-3 (1996), 1-30. (Hungarian)

Lange, Lynda, "The Function of Equal Education in Plato's ‘Republic’ and ‘Laws,’" in *The Sexism of Social and Political Theory*. L. Clark and L. Lange, eds. (Toronto: University of Toronto Press, 1979)

Lesser, Harry, "Plato's Feminism," *Philosophy* 54 (1979), 113-117.

Levin, Susan B., "Women's Nature and Role in the Ideal *Polis*: ‘Republic V’ Revisited," in *Feminism and Ancient Philosophy* Julie K. Ward (New York and London: Routledge, 1996)

Lovibond, Sabina, "An Ancient Theory of Gender: Plato and the Pythagorean Table," in *Women in Ancient Societies*, Archer, Fischler, and Wyke, eds. (London: Routledge), 88-101.

Mansfeld, Jaap, "Plato Over De Vrouw," *Alg. Ned. Tijdschr Wijs* 79 (1987), 199-120. (Dutch/Flemish)

Marquez, Alvaro, "El Tema De Lo Femenino En Platon," *Revista de Filosofio (Venezuela)* 9 (1986), 33-41. (Spanish)

Martin, Jane R, "Equality and Education in Plato," in *Feminism and Philosophy*, M. Vetterling-Braggin, F. A. Elliston, J. English, eds. (Totowa NJ: Littlefield, 1977)

Nye, Andrea, "The Hidden Host: Irigaray and Diotima at Plato's Symposium," *Hypatia* 3 (1989), 45-61.

Okin, Susan Moller, "Philosopher Queens and Private Wives: Plato on Women and the Family," *Philosophy and Public Affairs*, 6 (summer 1977)

Osborne, Martha Lee, "Plato's Unchanging View of Woman: A Denial That Anatomy Spells Destiny," *The Philosophical Forum* (summer 1975)

Pierce, Christine, "Equality: ‘Republic V,’" *The Monist*, 57 (January 1973)

Pierce, Christine, "Eros and Epistemology," in *Engendering Origins*, Bat-Ami Bar On, ed. (Albany: SUNY Press, 1994)

Pomeroy, Sarah, "Feminism in Book V of Plato's ‘Republic,’" *Apeiron*, VIII, 1 (1974)

Saxenhouse, Arlene W., "Eros and the Female in Greek Political Thought: An Interpretation of Plato's ‘Symposium,’" *Political Theory* 12 (1984), 5-27.

Saxenhouse, Arlene W., "The Philosopher and the Female in the Political Thought of Plato," *Political Theory*, 4 (May 1976), 195-212.

Senter, Nell W., "Plato on Women," *Southwest Philosophical Studies* 2 (1977), 4-13.

Smith, Janet Farrell, "Plato, Irony and Equality," *Hypatia* WSIF 1 (1983), 597-607.

Smith, Nicholas, "The Logic of Plato's Feminism," *Journal of Social Philosophy* 11 (1980), 5-11.

Tress, Daryl McGowan, "Relations in Plato's ‘Timaeus,’" *Journal of Neoplatonic Studies* 3 (1994), 93-139.

Vlastos, Gregory, "Was Plato a Feminist?" *Times Literary Supplement* 276 (March 17, 1989), 276, 288-289. Cited from *Studies in Greek Philosophy* ed. Daniel W. Graham, 1995. Volume 2: *Socrates, Plato, and Their Tradition* (Princeton: Princeton University Press), 133-143.

Wender, Dorothea, "Plato: Misogynist, Paedophile and Feminist," *Arethusa*, VI (spring 1973)

Aristotle

Books

Bickford, Susan, 1996. *The Dissonance of Democracy: Listening, Conflict, and Citizenship* (Ithaca: Cornell University Press)

Fortenbaugh, W. W., 1975. *Aristotle on Emotion: A Contribution to Philosophical Psychology, Rhetoric, Poetics, Politics, and Ethics* (New York: Harper & Row)

Freeland, Cynthia, 1998. F*eminist Interpretations of Aristotle* (University Park: Pennsylvania University Press)

Holland, Nancy, 1998. *The Madwoman's Dream: The Concept of the Appropriate in Ethical Thought* (University Park: Pennsylvania State University Press)

Articles

Achtenberg, Deborah, "Aristotelian Resources for Feminist Thinking," in *Feminism and Ancient Philosophy,* Julie K. Ward (New York and London: Routledge, 1996)

Achtenberg, Deborah, "The Role of the Ergon Argument in Aristotle's Nichomachean Ethics," *Ancient Philosophy* 9, 1 (1989),

Allen, Christine Garside, "Can a Woman be Good in the Same Way as a Man?" *Dialogue* 10 (1971), 534-544.

Berman, Ruth, "From Aristotle's Dualism to Materialist Dialectics: Feminist Transformation of Science and Society," in *Gender/Body/Knowledge: Feminist Reconstructions of Being and Knowing,* Alison M. Jaggar and Susan R. Bordo, eds. (New Brunswick: Rutgers University Press, 1989)

Cavarero, Adriana, "Equality and Sexual Difference," in *Beyond Equality and Difference*, Gisela Bock, ed. (New York: Routledge, 1992)

Cole, Eve Browning, "'Women, Slaves, and Love of Toil' in Aristotle's Moral Philosophy," in *Engendering Origins,* Bat-Ami Bar On, ed. (Albany: SUNY Press, 1994)

Cook, Kathleen C., "Sexual Inequality in Aristotle's Theories of Reproduction and Inheritance," in *Feminism and Ancient Philosophy,* Julie K. Ward (New York and London: Routledge, 1996)

Curd, Patricia, "Aristotelian Visions of Moral Character in Virginia Woolf's *Mrs. Dalloway*," in *Feminism and Ancient Philosophy,* Julie K. Ward (New York and London: Routledge, 1996)

Curran, Angela, "Feminism and the Narrative Structures of the 'Poetics,'" in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Deslauriers, Marguerite, "Sex and Essence in Aristotle's 'Metaphysics' and 'Biology,'" in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Fememias, Maria Luisa, "Women and Natural Hierarchy in Aristotle," *Hypatia* 9, 1 (1994), 164-172.

Fortenbaugh, W. W., "Aristotle on Slaves and Women," in *Articles on Aristotle: 2, Ethics and Politics,* J. Barnes, J. Schofield, and R. Sorabji, eds. (London: Duckworth, 1977)

Freeland, Cynthia A., "Nourishing Speculation: A Feminist Reading of Aristotelian Science," in *Engendering Origins*, Bat-Ami Bar On, ed. (Albany: SUNY Press, 1994)

Freeland, Cynthia A., "On Irigaray on Aristotle," in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Green, Judith, "Aristotle on Necessary Verticality, Body Heat, and Gendered Proper Places in the Polis: A Feminist Critique," *Hypatia* 7, 1 (1992), 70-96.

Groenhout, Ruth, "The Virtue of Care: Aristotelian Ethics and Contemporary Ethics of Care," in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Hass, Marjorie, "Feminist Readings of Aristotelian Logic," in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (university Park: Pennsylvania University Press, 1998)

Hein, Hilde, "Liberating Philosophy: An End to the Dichotomy of Spirit and Matter," in *Women, Knowledge, and Reality: Explorations in Feminist Philosophy* Ann Garry and Marilyn Pearsall, eds. (Boston: Unwin Hyman, 1989)

Hirschman, Linda Redlick, "The Book of ‘A,’" in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Homiak, Marcia, "Feminism and Aristotle's Rational Ideal," in *Feminism and Ancient Philosophy*, Julie K. Ward (New York and London: Routledge, 1996)

Horowitz, Maryanne Cline, "Aristotle and Women," *Journal of the History of Biology* 9 (1976), 183-213.

Irigaray, Luce, "Place, Interval: A Reading of Aristotle's ‘Physics IV,’" in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Koziak, Barbara, "Tragedy, Citizens, and Strangers: The Configuration of Aristotelian Political Emotion," in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Lange, Lynda, "Woman is Not a Rational Animal: On Aristotle's Biology of Reproduction," in *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. Sandra Harding and Merrill B. Hintikka, eds. (Dordrecht: Reidel, 1983)

Matthews, Gareth B., "Gender and Essence in Aristotle," *Australasian Journal of Philosophy* Supp. 64 (1986), 16-25.

Modrak, Deborah K. W., "Aristotle's Theory of Knowledge and Feminist Epistemology," in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Modrak, Deborah, "Aristotle: Women, Deliberation, and Nature," in *Engendering Origins*, Bat-Ami Bar On, ed. (Albany: SUNY Press, 1994)

Morsink, Johannes, "Was Aristotle's Biology Sexist?" *Journal of the History of Biology* 12,1 (1979), 83-112.

Mulgan, Richard, "Aristotle and the Political Role of Women," *History of Political Thought* 15, 2 (1994), 179-202.

Nussbaum, Martha, "Aristotle, Feminism, and Needs for Functioning," in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Poster, Carol, "(Re) Positioning Pedagogy: A Feminist Historiography of Aristotle's ‘Rhetorica,’" in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

Rosenberg, Rosalind, "In Search of Woman's Nature, 1850-1920," *Feminist Studies*, 3 (1975), 141-154.

Sakezles, Priscilla K., "Feminism and Aristotle," *Apeiron* 32, 1 (1999), 67-74.

Senack, Christine M., "Aristotle on the Woman's Soul,’ in *Engendering Origins*, Bat-Ami Bar On, ed. (Albany: SUNY Press, 1994)

Spelman, Elizabeth V., "Aristotle and the Politicization of the Soul, " in *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. Sandra Harding and Merrill B. Hintikka, eds. (Dordrecht: Reidel, 1983)

Stiehm, Judith Hicks, "The Unit of Political Analysis: Our Aristotelian Hangover," in *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. Sandra Harding and Merrill B. Hintikka, eds. (Dordrecht: Reidel, 1983)

Thom, P., "Stiff Cheese For Women," *Philosophical Forum* 8, 1 (1976), 94-107.

Tress, Daryl McGowan, "The Metaphysical Science of Aristotle's ‘Generation of Animals,’ and It Feminist Critics," *Review of Metaphysics* 46, 2 (1992), 307-341. Reprinted in *Feminism and Ancient Philosophy* Julie K. Ward (New York and London: Routledge, 1996)

Tuana, Nancy, "Aristotle and the Politics of Reproduction," in *Engendering Origins*, Bat-Ami Bar On, ed.

(Albany: SUNY Press, 1994)

Tumulty, Peter, "Aristotle, Feminism, and Natural Law Theory," *New Scholars* 55 (1981), 450-464.

Ward, Julia K., "Aristotle on *Philia*: The Beginning of a Feminist Ideal of Friendship," in *Feminism and Ancient Philosophy* Julie K. Ward (New York and London: Routledge, 1996)

Whitbeck, Caroline, "Theories of Sex Difference," in *Women and Philosophy: Toward a Theory of Liberation* Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Witt, Charlotte, "Form, Normativity, and Gender in Aristotle: A Feminist Perspective," in *Feminist Interpretations of Aristotle*, Cynthia A. Freeland, ed. (University Park: Pennsylvania University Press, 1998)

St. Paul

Dubarle, A. M., "Paul et L'Antifeminisme," *Rev. Sci. Phi. Theol.* 60 (1976), 261-280. (French)

Medieval Philosophy

Books

Borresen, Kari Elisabeth, ed. 1991. *Images of God and Gender Models: in Judaeo-Christian Tradition* (Atlantic Highlands: Humanities Press)

Brabant, Margaret, ed. 1992. *Politics, Gender, and Genre: The Political Thought of Christine de Pizan* (Boulder: Westview Press)

Articles

Green, Karen, "Christine de Pisan and Thomas Hobbes," *Philosophical Quarterly* 44, 177 (1994), 456-475.

Hollywood, Amy M., "Beauvoir, Irigaray, and the Mystical," *Hypatia* 9, 4 (1994), 158-185.

John, Helen J., "Hildegard of Bingen: A New Medieval Philosopher?" *Hypatia* 7, 1 (1992), 115-123.

McLaughlin, Eleanor, "Equality of Souls, Inequality of Sexes: Women in Medieval Theology," in *Religion and Sexism*, Rosemary Ruether, ed. (New York: Simon & Schuster, 1974)

Ruether, Rosemary, "Misogynism and Virginal Feminism in the Fathers of the Church," in *Religion and*

Sexism, Rosemary Ruether, ed. (New York: Simon & Schuster, 1974)

St. Augustine

Articles

Borresen, Kari Elisabeth, "Patristic ‘Feminism’: The Case of Augustine,’ *Augustinian Studies* 25 (1994), 139-152.

Duval, Shannon, "Augustine's Radiant Confessional--Theatre of Prophecy,’ *Contemporary Philosophy* 15, 2 (1993), 1-4.

St. Thomas Aquinas

Books

Traina, Cristina L. H., 1999. *Feminist Ethics and Natural Law: The End of the Anathemas* (Washington DC: Georgetown University Press)

Articles

Hartel, Joseph, "The Integral Feminism of St. Thomas Aquinas," *Gregorianum* 77, 3 (1996), 527-547.

Hein, Hilde, "Liberating Philosophy: An End to the Dichotomy of Spirit and Matter," in *Women, Knowledge, and Reality: Explorations in Feminist Philosophy* Ann Garry and Marilyn Pearsall, eds. (Boston: Unwin Hyman, 1989)

Lavaud, B., "Toward a Theology of Woman," *Thomist* 2 (1940), 459-518.

Renaissance Philosophy

Books

Maclean, Ian, 1980. *The Renaissance Notion of Woman: A Study in the Fortunes of Scholasticism and Medical Science in European Intellectual Life* (Cambridge: Cambridge University Press)

Articles

Gibson, Joan, "Educating for Silence: Renaissance Women and the Language Arts,’ *Hypatia* 4 (1989), 9-27.

Zedler, Beatrice H., "Marie le Jars de Gournay," in *A History of Women Philosophers, Volume II: Medieval, Renaissance and Enlightenment, A. D. 500-1600* (Norwell: Kluwer, 1989)

Francois Poullain de la Barre, 1647-1723

Amoros, Celia, "Cartesianism and Feminism: What Reason has Forgotten; Reasons for Forgetting," *Hypatia* 9, 1 (1994), 147-163.

Seidel, Michael A., "Poullain de la Barre's ‘The Woman as Good as the Man,’ *Journal of the History of Ideas* 35 (1974), 499-508.

Francis Bacon and the Scientific Revolution

Books

Merchant, Carolyn, 1980. *The Death of Nature: Women, Ecology, and the Scientific Revolution* (San Francisco: Harper & Row)

Articles

Keller, Evelyn Fox, "Baconian Science: A Hermaphroditic Birth," *Philosophical Forum* 11, 3 (1980), 299-308.

Potter, Elizabeth, "Modeling the Gender Politics in Science," *Hypatia* 3 (1988), 19-33.

Rene Descartes and Cartesianism

Books

Bordo, Susan, ed. 1999. *Feminist Interpretations of Rene Descartes* (University Park: Pennsylvania University Press)

Bordo, Susan, 1987. *The Flight to Objectivity: Essays on Cartesianism and Culture* (Albany: SUNY Press)

Scheman, Naomi, 1993. *Engenderings: Constructions of Knowledge, Authority, and Privilege* (New York: Routledge)

Articles

Amoros, Celia, "Cartesianism and Feminism: What Reason has Forgotten; Reasons for Forgetting,"

Hypatia 9, 1 (1994), 147-163

Atherton, Margaret, "Cartesian Reason and Gendered Reason," in *A Mind of One's Own: Feminist Essays on Reason and Objectivity* Louise Antony and Charlotte Witt, eds. (Boulder: Westview Press) 19-34.

Berman, Ruth, "From Aristotle's Dualism to Materialist Dialectics: Feminist Transformation of Science and Society," in *Gender/Body/Knowledge: Feminist Reconstructions ofBeing and Knowing*, Alison M. Jaggar and Susan R. Bordo, eds. (New Brunswick: Rutgers University Press, 1989)

Bordo, Susan, "The Cartesian Masculinization of Thought," *Signs* 11 (1986), 439-456.

Bordo, Susan and Mario Moussa, "Rehabilitating the ‘I,’" in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Cantrell, Carol H., "Analogy as Destiny: Cartesian Man and the Woman Reader," *Hypatia* 5, 2 (1990), 7-19.

Clarke, Stanley, "Descartes' ‘Gender,’" in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

David, Anthony, "Le Doeuff and Irigaray on Descartes," *Philosophy Today* 41, 3-4 (1997), 367-382.

Gatens, Moira, "Modern Rationalism," in *A Companion to Feminist Philosophy*, Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Harth, Erica, "Cartesian Women," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Heywood, Leslie, "When Descartes Met the Fitness Babe: Academic Cartesianism and the Late Twentieth-Century Cult of the Body," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Hodge, Joanna, "Subject, Body, and the Exclusion of Women from Philosophy," in *Feminist Perspectives in Philosophy*, Morwenna Griffiths, ed. (Bloomington: Indiana University Press, 1988) 152-168.

Irigaray, Luce, "Wonder: A Reading of Descartes, ‘The Passions of the Soul,’" Carolyn Burke and Gillian C. Gill, trans., in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Lloyd, Genevieve, "Reason as Attainment," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

O'Neill, Eileen, "Women Cartesians, ‘Feminine Philosophy,’ and Historical Exclusion," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Paliyenko, Adrianna, "Postmodern Turns Against the Cartesian Subject: Descrtaes' ‘I’, Lacan's Other," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Perry, Ruth, "Radical Doubt and the Liberation of Women," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Saenz, Mario, "Cartesian Autobiography/Post-Cartesian Testimonials," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Stern, Karl, "Descartes," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Thompson, J., "Women and the High Priests of Reason," *Radical Philosophy* 34 (summer 1983), 10-14.

Wartenberg, Thomas E., "Descartes's Mood: The Question of Feminism in the Correspondence with Elisabeth," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Winders, James A., "Writing Like a Man (?): Descartes, Science, and Madness," in *Feminist Interpretations of Rene Descartes*, Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Wiseman, Mary Bittner, "Beautiful Exiles in Aesthetics," in *Aesthetics in Feminist Perspective,* Hilde Hein, ed. (Bloomington: Indiana University Press, 1993)

Spinoza

Gatens, Moira, "Modern Rationalism," in *A Companion to Feminist Philosophy*, Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Leibniz

Gatens, Moira, "Modern Rationalism," in *A Companion to Feminist Philosophy*, Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Hobbes

Dalitz, Renee J., "The Subjection of Women in the Contractual Society," in *Empirical Logic and Public Debate*, Erik C. W. Krabbe, ed. (Amsterdam: Rodopi, 1993)

Green, Karen, "Christine de Pisan and Thomas Hobbes," *Philosophical Quarterly* 44, 177 (1994), 456-475.

Stefano, Christine Di, "Masculinity as Ideology in Political Theory: Hobbesian Man Considered," *Women's Studies International Forum* 6, 6 (1983)

Seventeenth Century Woman Philosophers

Books

Atherton, Margaret, ed. 1994. *Women Philosophers of the Early Modern Period* (Indianapolis:Hackett)

Smith, Hilda L., 1982. *Reason's Disciples: Seventeenth-Century English Feminists* (Urbana: University of Illinois Press)

Frankel, Lois, "Damaris Dudworth Masham: A Seventeenth Century Feminist Philosopher," *Hypatia* 4 (1989), 80-90.

O'Neill, Eileen, "Disappearing Ink: Early Modern Philosophers and Their Fate in History," in *Philosophy in a Feminist Voice*, Janet Kourany, ed. (Princeton: Princeton University Press, 1998)

O'Neill, Eileen, "Women Cartesians, ‘Feminine Philosophy,’ and Historical Exclusion’ in *Feminist Interpretations of Rene Descartes* Susan Bordo, ed. (University Park: Pennsylvania University Press, 1999)

Nye, Andrea, "Polity and Prudence:The Ethics of Elisabeth, Princess Palatine," in *Hypatia's Daughters: Fifteen Hundred Years of Women Philosophers* McAlister, Linda Lopez, ed. (Bloomington: Indiana University Press, 1996)

Pateman, Carole, "Patriarchal Confusions," *International Journal of Moral and Social Studies* 3 (1988), 127-143.

Shanley, Mary Lyndon, "Marriage Contract and Social Contract in Seventeenth Century English Political Thought," in *The Family in Political Thought*, Jean Bethke Elshtain, ed. (Brighton: Harvester Press, 1982)

Weinberg, Sue, "Damaris Cudworth Masham: A Learned Lady of the Seventeenth Century," in *Norms and Values: Essays on the Work of Virginia Held* Joram Graf Haber, ed. (Lanham: Rowman & Littlefield)

David Hume

Books

Baier, Annette, 1994. *Moral Prejudices* (Cambridge MA: Harvard University Press)

Richards, Janet Radcliffe, 1980. *The Sceptical Feminist: A Philosophical Inquiry* (Boston: Routledge and K. Paul)

Articles

Battersby, C., "An Enquiry Concerning the Humean Woman," *Philosophy* 56 (1981), 303-312.

Burns, S., "The Humean Female," *Dialogue* 15, 3 (1976), 414-424.

Burns, S. and L. Marcil-Lacoste, "Hume on Women," in *The Sexism of Social and Political Theory* L. Clark and L. Lange, eds. (Toronto: University of Toronto Press, 1979)

Gowans, Christopher W., "After Kant: Ventures in Morality Without Respect for Persons," *Social Theory and Practice* 22, 1 (1996), 105-129.

Marcil-Lacoste, L., "The Consistency of Hume's Position Concerning Women," *Dialogue* 15, 3 (1976), 425-440.

Immanuel Kant

Books

Baier, Annette, 1994. *Moral Prejudices* (Cambridge MA: Harvard University Press)

Holland, Nancy, 1998. *The Madwoman's Dream: The Concept of the Appropriate in Ethical Thought* (University Park: Pennsylvania State University Press)

Hutchings, Kimberly, 1996. *Kant, Critique, and Politics* (New York: Routledge)

Jauch, Ursula Pia, 1989. *Immanuel Kant zur Geschlechterdifferenz: Aufklarerische Vorurteilskritik und burgerliche Geschlechtsvormundschaft* (Vienna: Passagen) (German)

Moscovici, Claudia, 1996. *From Sex Objects to Sexual Subjects* (New York: Routledge)

Schott, Robin May, 1993. *A Feminist Critique of the Kantian Paradigm* (University Park: Pennsylvania

State University Press)

Schott, Robin May, ed. 1997. *Feminist Interpretations of Immanuel Kant* (University Park: Pennsylvania State University Press)

Articles

Baron, Marcia, "Kantian Ethics and Claims of Detachment," in *Feminist Interpretations of Immanuel Kant* Robin May Schott, ed. (University Park: Pennsylvania State University Press, 1997)

Blum, Lawrence, "Kant's and Hegel's Moral Rationalism: A Feminist Perspective," C*anadian Journal of Philosophy*, 12, 2 (1982), 287-302.

David-Menard, Monique, "Kant, the Law, and Desire," Leslie Lykes de Halbert, trans., in *Feminist Interpretations of Immanuel Kant* Robin May Schott, ed. (University Park: Pennsylvania State University Press, 1997)

Gould, Timothy, "Intensity and its Audiences: Notes Towards a Feminist Perspective," *Canadian Journal of Philosophy* 12 (1982), 287-302.

Gowans, Christopher W., "After Kant: Ventures in Morality Without Respect for Persons," *Social Theory and Practice* 22, 1 (1996), 105-129.

Hall, Kin, "Sensus Communis and Violence: A Feminist Reading of Kant's ‘Critique of Judgment,’" in *Feminist Interpretations of Immanuel Kant* Robin May Schott, ed. (University Park: Pennsylvania State University Press, 1997)

Hermann, Barbara, "Ob es sich lohnen konnte, uber Kants Auffassungen von Sexualitat und Ehe nachzudenken?" *Deutsche Zeitschrift fur Philosophie* 43, 6 (1995), 967-988. (German)

Heinrichs, Thomas, "Die Ehe als Ort gleichberechtigter Lust," *Kant Studien* 86, 1 (1995)), 41-53. (German)

Kneller, Jane, "The Aesthetic Dimension of Kantian Autonomy," in *Feminist Interpretations of Immanuel Kant* Robin May Schott, ed. (University Park: Pennsylvania State University Press, 1997)

Kneller, Jane, "Discipline and Silence: Women and Imagination in Kant's Theory of Taste," in *Aesthetics in Feminist Perspective* Hilde Hein, ed. (Bloomington: Indiana University Press, 1993)

Lango, John W., "Does Kant's Ethics Ignore Relations Between Persons?" in *Norms and Values: Essays on the Work of Virginia Held* Joram Graf Haber, ed. (Lanham: Rowman & Littlefield, 1998)

Mendus, Susan, "Kant: An Honest but Narrow-Minded Bourgeois?" in *Women in Western Political Philosophy*, Ellen Kennedy and Susan Mendus, eds. (Brighton: Wheatsheaf Books, 1987)

Moen, Marcia, "Feminist Themes in Unlikely Places: Re-Reading Kant's 'Critique of Judgment,'" in *Feminist Interpretations of Immanuel Kant* Robin May Schott, ed. (University Park: Pennsylvania State University Press, 1997)

Nagl-Docekal, Herta, "Feminist Ethics: How it Could Benefit from Kant's Moral Philosophy," Stephanie Morgenstern, trans., in *Feminist Interpretations of Immanuel Kant* Robin May Schott, ed. (University Park: Pennsylvania State University Press, 1997)

Okin, Susan Moller, "Reason and Feeling in Thinking about Justice," *Ethics* 99 (1989), 229-249.

Rumsey, Jean, "Re-Visions of Agency in Kant's Moral Theory," in *Feminist Interpretations of Immanuel Kant* Robin May Schott, ed. (University Park: Pennsylvania State University Press, 1997)

Schott, Robin May, "Feminism and Kant: Antipathy or Sympathy?" in *Autonomy and Community* Jane E. Kneller, ed. (Albany: SUNY Press, 1998)

Schott, Robin May, "Kant," in *A Companion to Feminist Philosophy,* Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Schroder, Hannelore, "Kant's Patriarchal Order," Rita Gircour, trans., in *Feminist Interpretations of Immanuel Kant* Robin May Schott, ed. (University Park: Pennsylvania State University Press, 1997)

Sedgwick, Sally, "Can Kant's Ethics Survive the Feminist Critique?" *Pacific Philosophical Quarterly* 71, 1 (1990), 60-79.

Wilson, Holly L., "Rethinking Kant from the Perspective of Ecofeminism," in *Feminist Interpretations of Immanuel Kant* Robin May Schott, ed. (University Park: Pennsylvania State University Press, 1997)

Wilson, Holyn, "Kant and Ecofeminism," in *Ecofeminism: Women, Culture, Nature* Karen J. Warren, ed. (Bloomington: Indiana University Press)

Wiseman, Mary Bittner, "Beautiful Exiles in Aesthetics," in *Aesthetics in Feminist Perspective* Hilde Hein, ed. (Bloomington: Indiana University Press, 1993)

Zweig, Arnulf, "Kant and the Family," in *Kindred Matters* Diana Tietjens Meyers, ed. (Ithaca: Cornell University Press, 1993)

Sapp, Vicki G., "The Philosopher's Seduction: Hume and the Fair Sex," *Philosophy and Literature* 19, 1 (1995), 1-15.

Liberalism

Brennan, Theresa and Carole Pateman, "'Mere Auxiliaries to the Commonwealth': Women and the Origins of Liberalism," *Political Studies* XXVII 2, (1968)

Tapper, Marion, "Can a Feminist Be a Liberal?" *Australasian Journal of Philosophy* Supp. 64 (1986)

John Locke

Butler, Melissa, "Early Political Roots of Feminism: John Locke and the Attack on Patriarchy," *American Political Science Review* 72 (1978) Reprinted in *Feminism and Philosophy: Essential Readings in Theory, Reinterpretation, and Application* (Boulder: Westview Press, 1995)

Clark, Lorenne, "Women and Locke: Who Owns the Apples in the Garden of Eden?" in *The Sexism of Social and Political Theory*. L. Clark and L. Lange, eds. (Toronto: University of Toronto Press, 1979)

Simons, Martin, "Why Can't a Man Be More Like a Woman? (A Note on John Locke's Educational Thought)," *Educational Theory* 40, 1 (1990), 135-145.

Jean -Jacques Rousseau

Articles

Bloch, M. and J. H. Bloch, "Women and the Dialectics of Nature in Eighteenth Century French Thought," in *Nature, Culture, and Gender*, C. MacCormack and M. Strathern, eds. (Cambridge: Cambridge University Press, 1980)

Canovan, Margaret, "Rousseau's Two Concepts of Citizenship," in *Women in Western Political Philosophy*, Ellen Kennedy and Susan Mendus, eds. (Brighton: Wheatsheaf Books, 1987)

Gatens, Moira, "Rousseau and Wollstonecraft: Nature vs. Reason," *Australasian Journal of Philosophy* Supp. 64 (1986), 1-15.

Green, Karen, "Rousseau's Women," *International Journal of Philosophical Studies* 4, 1 (1996), 87-109.

Holland, Nancy J., "Introduction to Kofman's 'Rousseau's Phallocratic Ends,'" *Hypatia* 3 (1989). 119-122.

Lange, Lynda, "Rousseau and Modern Feminism," *Soc. Theor. Pract.* 7 (1981), 245-277.

Lange, Lynda, "Rousseau: Women and the General Will," in *The Sexism of Social and Political Theory*. L. Clark and L. Lange, eds. (Toronto: University of Toronto Press, 1979)

Lloyd, Genevieve, "Rousseau on Reason, Nature, and Women," *Metaphilosophy* 14, 3 & 4 (July & October 1983)

Martin, J., "Sophie and Emile: A Case Study of Sex Bias in the History of Educational Thought," *Harvard Educational Review* 51, 3 (1981), 357-371.

Pateman, Carole, "'The Disorder of Women': Women, Love, and the Sense of Justice," *Ethics* 91 (1980), 20-31.

Rapaport, Elizabeth, "On the Future of Love: Rousseau and the Radical Feminists," *The Philosophical Forum* 5, 1-2 (1973-1974), 185-205. Reprinted in *Women and Philosophy: Toward a Theory of Liberation* Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Thomas, Paul, "Jean-Jacques Rousseau, Sexist?" *Feminist Studies* (1991), 195-218.

Weiss, Penny, "Rousseau's Political Defense of the Sex-Roled Family," *Hypatia* (1990), 90-109.

Wexler, Victor, "'Made for Man's Delight': Rousseau as Anti-Feminist," *American Historical Review* 81, 2 (April 1976)

Mary Wollstonecraft

Books

Falco, Maria J., ed., 1996. *Feminist Interpretations of Mary Wollstonecraft* (University Park: Pennsylvania University Press)

Sabrosky, Judith A., 1979. *From Rationality to Liberation* (Westport: Greenwood Press)

Articles

Barker-Benfield, G. J., "Mary Wollstonecraft: Eighteenth-Century Commonwealthwoman," *Journal of the History of Ideas* 50 (1989), 95-115.

Brody, Miriam, "Mary Wollstonecraft: Sexuality and Women's Rights," in *Feminist Theories,* Dale Spender, ed. (London: The Women's Press, 1983)

Coddetta, Carolina, "The Problem of Power in the Feminist Theory," *Fronesis* 2, 2 (1995), 59-95.

(Spanish)

Disch, Lisa, "Claire Loves Julie: Reading the Story of Women's Friendship in ‘La Nouvelle Heloise,’" *Hypatia* 9, 3 (1994), 19-45.

Duhan, Laura, "Feminism and Peace Theory: Women as Nurturers versus Women as Public Citizens," in *In the Interest of Peace: A Spectrum of Philosophical Views* (Wolfeboro: Longwood, 1990)

Gatens, Moira, "Rousseau and Wollstonecraft: Nature vs. Reason," *Australasian Journal of Philosophy* Supp. 64 (1986), 1-15.

Grimshaw, Jean, "Mary Wollstonecraft and the Tensions in Feminist Philosophy," *Radical Philosophy* 52 (1989), 11-17.

Gubar, Susan, "Feminist Misogyny: Mary Wollstonecraft and the Paradox of ‘It Takes One to Know One,’" *Feminist Studies* 20, 3 (1994), 453-473.

Janes, R. M., "On the Reception of Mary Wollstonecraft's ‘A Vindication of the Rights of Women,’" *Journal of the History of Ideas* 39 (1978), 293-302.

Korsmeyer, Carolyn, "Reason and Morals in the Early Feminist Movement: Mary Wollstonecraft," *Philosophical Forum (Boston)* 5, 1-2 (1973-1974). Reprinted in *Women and Philosophy: Toward a Theory of Liberation* Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Larson, Elizabeth, "Mary Wollstonecraft and Women's Rights," *Free Inquiry* 12, 2 (1992). 45-48.

Mackenzie, Catriona, "Reason and Sensibility: The Ideal of Women's Self-Governance in the Writings of M. Wollstonecraft," in *Hypatia's Daughters: Fifteen Hundred Years of Women Philosophers* McAlister, Linda Lopez, ed. (Bloomington: Indiana University Press, 1996)

McCrystal, John, "Revolting Women: The Use of Revolutionary Discourse in Mary Astell and Mary Wollstonecraft Compared," *History of Political Thought* 14, 2 (1993), 189-203.

Wexler, Alice, "Mary Wollstonecraft, Her Tragic Life and Her Passionate Struggle for Freedom," *Feminist Studies* 7 (1981), 114-133.

Catharine Macaulay

Gardner, Catherine, "Catharine Macaulay's ‘Letters on Education’: Odd but Equal," *Hypatia* 13, 1 (1998), 118-137.

Mary Astell

Bryson, Cynthia B., "Mary Astell: Defender of the ‘Disembodied Mind,’" *Hypatia* 13, 4 (1998) 40-62.

McCrystal, John, "Revolting Women: The Use of Revolutionary Discourse in Mary Astell and Mary Wollstonecraft Compared," *History of Political Thought* 14, 2 (1993), 189-203.

Utilitarianism

Boralevi, Lea Campos, "Utilitarianism and Feminism," in *Women in Western Political Philosophy*. Ellen Kennedy and Susan Mendus, eds. (Brighton: Wheatsheaf Books, 1987)

James Mill

Ball, Terence, "Utilitarianism, Feminism and the Franchise: James Mill and His Critics," *History of Political Thought* 1 (1980)

Jeremy Bentham

Books

Campos-Boralevi, Lea, 1984. *Bentham and the Oppressed* (West: Berlin-De-Gruyter)

Articles

Williford, Miriram, "Bentham and the Rights of Women," *Journal of the History of Ideas* 36 (January-March 1975)

John Stuart Mill

Books

Himmelfarb, Gertrude, 1974. *On Liberty and Liberalism: The case of John Stuart Mill* (New York: Knopf)

Morales, Maria H., 1996. *Perfect Equality: John Stuart Mill on Well-Constituted Communities* (Lanham: Rowman & Littlefield)

Pappe, H. O., 1962. *John Stuart Mill and the Harriet Taylor Myth* (New York: Cambridge University Press)

Pyle, Andrew, ed. 1995. *'The Subjection of Women': Contemporary Responses to John Stuart Mill* (Bristol: Thoemmes)

Articles

Annas, Julia, "Mill and the Subjection of Women," *Philosophy* 52 (1977)

Brecher, Bob, "Why Patronize Feminists? A reply to Stove on Mill," *Philosophy* 68, 265 (1993), 397-400.

Burgess Jackson, Keith, "John Stuart Mill, Radical Feminist," *Social Theory and Practice* 21, 3 (1995), 389-396.

Donner, Wendy, "John Stuart Mill's Liberal Feminism," *Philosophical Studies* 69, 2-3 (1993), 155-166.

Dyzanhaus, David, "John Staurt Mill and the Harm of Pornography," *Ethics* 102, 3 (1992), 534-551.

Goldstein, Leslie, "Mill, Marx, and Women's Liberation," *Journal of the History of Philosophy* XVIII, 3 (July 1980)

Hornsby, Jennifer and Rae Langton, "Free Speech and Illocution," *Legal Theory* 4, 1 (1998), 21-37.

Howes, John, "Mill on Women and Human Development," *Australasian Journal of Philosophy* Supp. 64 (1986), 66-74.

Knight, Jamie K., "With Liberty and Justice for Some," *International Journal of Applied Philosophy* 2 (1984), 85-90.

Mahowald, Mary B., "Against Paternalism: A Developmental View," *Philosophy Research Archives* 6, 1386 (1980)

Mahowald, Mary, "Freedom versus Happiness, and 'Women's Lib,'" *Journal of Social Philosophy* 6 (1975), 10-13.

Mendus, Susan, "John Stuart Mill and Harriet Taylor on Women and Marriage," *Utilitas* 6, 2 (1994), 287-299.

Nubiola, Jaime, "Emancipacion, magnanimidad y mujeres," *Anuario Filosofico* 27, 2 (1994), 641-654 (Spanish)

Robson, John M., "'Feminine' and 'Masculine': Mill vs. Grote," *Mill Newsletter* 12 (1977), 18-22.

Shanley, Mary Lyndon, "Marital Slavery and Friendship: John Stuart Mills' 'The Subjection of Women," *Political Theory* 9 (1981), 229-247.

Skipper, Robert, "Mill and Pornography," *Ethics* 103, 4 (1993), 726-730.

Tulloch, Gail, "Mill's Epistemology in Practice in His Liberal Feminism," *Educational Philosophy and Theory* 21, 2 91989), 32-39.

G. W. F. Hegel

Books

Gauthier, Jeffrey A., 1997. *Hegel and Feminist Social Criticism: Justice, Recognition, and the Feminine* (Albany: SUNY Press)

Mills, Patricia Jagentowicz, 1996. *Feminist Interpretations of G. W. F. Hegel* (University Park: Pennsylvania State University Press)

Mills, Patricia Jagentowicz, 1987. *Woman, Nature, and Psyche* (New Haven: Yale University Press)

Articles

Armstrong, Susan, "A Feminist Reading of Hegel and Kierkegaard," in *Hegel, History, and Interpretation* Shaun Gallagher, ed. (Albany: SUNY Press, 1997)

Arthur, Chris, "Hegel as Lord and Master," *Radical Philosophy* 50 (1988), 19-25.

Assiter, Alison, "Autonomy and Pornography," in *Feminist Perspectives in Philosophy* Morwenna Griffiths, ed. (Bloomington: Indiana University Press, 1988)

Blum, Lawrence, "Kant's and Hegel's Moral Rationalism: A Feminist Perspective," C*anadian Journal of Philosophy*, 12, 2 (1982), 287-302.

Brod, Harry, "Pornography and the Alienation of Male Sexuality," *Social Theory and Practice* 14 (1988), 265-284.

Cutrofello, Andrew, "Hegel's Confessions; or, Why We Need a Sequel to the 'Phenomenology of Spirit,'" *Owl of Minerva* 26, 1 (1994), 21-28.

Easton, Susan, "Hegel and Feminism," *Radical Philosophy* 38 (1984)

Easton, Susan, "Slavery and Freedom: A Feminist Reading of Hegel," *Politics* 5, 2 (October 1985)

Fuchs, Jo Ann Pilardi, "On the War Path and Beyond: Hegel, Freud, and Feminist Theory," *Hypatia* WSIF 1 (1986), 565-572.

Gould, Timothy, "Intensity and its Audiences: Notes Towards a Feminist Perspective," *Canadian Journal of Philosophy* 12 (1982), 287-302.

Harrington, Thea, "The Speaking Abject in Kristeva's 'Powers of Horror,'" *Hypatia* 13, 1 (1998), 138-157.

Hayim, Gila J., "Hegel's Critical Theory and Feminist Concerns," *Philosophy and Social Criticism* (1990), 1-21.

Hodge, Joanna, "Women and the Hegelian State," in *Women in Western Political Philosophy*, Ellen Kennedy and Susan Mendus, eds. (Brighton: Wheatsheaf Books, 1987)

Holland, Nancy J., "Convergence on Whose Truth?: Feminist Philosophy and the 'Masculine Intellect' of Pragmatism," *Journal of Social Philosophy* 26, 2 (1995), 170-183.

James, Christine, "Hegel, Harding, and Objectivity," *Southwest Philosophy Review* 14, 1 (1997), 111-122.

Mills, Patricia Jagentowicz, "'Feminist' Sympathy and Other Serious Crimes: A Reply to Swindle," *Owl of Minerva* 24, 1 (1992), 55-62.

Mills, Patricia Jagentowicz, "Hegel and 'the Woman Question': Recognition and Intersubjectivity," in *The Sexism of Social and Political Theory*. L. Clark and L. Lange, eds. (Toronto: University of Toronto Press, 1979)

Mills, Patricia Jagentowicz, "Hegel's 'Antigone,'" *Owl of Minerva* 17 (1986), 131-152.

Mills, Patricia Jagentowicz, "Woman's Experience: Re the Dialectic of Desire and Recognition," in *Writing the Politics of Difference* (Albany: SUNY Press, 1991)

Oliver, Kelly, "Antigone's Ghost: Undoing Hegel's 'Phenomenology of Spirit,'" *Hypatia* 11, 1 (1996), 67-90.

Parente, Alfredo, "Una Feminista che Esortava a Sputare su Hegel," *Riv. Stud. Croce* 19 (1982), 204-205. (Italian)

Perez Estevez, Antonio, "Lo Femenino en la Filosofia del Derecho de Hegel," *Revista de Filosofio (Venezuela)* 15 (1991), 11-20. (Spanish)

Ravven, Heidi M., "Has Hegel Anything to Say to Feminists?" *Owl of Minerva* 19 (1988), 149-168.

Ravven, Heidi M., "A Response to 'Why Feminists Should Take the 'Phenomenology of Spirit' Seriously," *Owl of Minerva* 24, 1 (1992), 63-68.

Rosenthal, Abigail, "Feminism Without Contradictions," *Monist* 57 (1973), 28-42.

Swindle, Stuart, "Why Feminists Should Take the 'Phenomenology of Spirit' Seriously," *Owl of Minerva* 24, 1 (1992), 41-54.

Marx and Marxism

Books

Aronson, Ronald, 1994. *After Marxism* (New York: Guilford)

Barrett, Michele. 1980. *Women's Oppression Today* (London: Redwood Burn LTD)

Carver, Terrell and Paul Thomas, eds. 1995. *Rational Choice Marxism* (University Park: Pennsylvania State University Press)

Cooke, Brett, George E. Slusser, and Jaume Marti-Olivella, eds. 1998. *The Fantastic Other: An Interface of Perspectives* (Amsterdam: Rodopi)

Ferguson, Kathy E., 1980. *Self, Society, and Womankind: The Dialectic of Liberation* (Wesport: Greenwood Press)

Gottlieb, Roger D., ed. 1989. *An Anthology of Western Marxism* (New York: Oxford University Press)

Hartsock, Nancy C. M., 1983. *Money, Sex, and Power: Toward a Feminist Historical Materialism* (New York: Longman)

Kain, Philip J., 1993. *Marx and Modern Political Theory* (Lanham: Rowman and Littlefield) Chapter 7

Messerschmidt, James W., 1986. *Capitalism, Patriarchy, and Crime: Toward a Socialist Feminist Criminology* (Totowa: Rowman & Littlefield)

Sayers, Janet, 1982. *Biological Politics* (London: Tavistock)

Stevernagel, Gertrude A., 1979. *Political Philosophy as Therapy: Marcuse Reconsidered* (Westport:

Greenwood)

Vogel, Lise, 1983. *Marxism and the Oppression of Women: Toward a Unitary Theory* (London: Pluto)

Articles

Aveling, Eleanor Marx and Edward Aveling, "The Woman Question: From a Socialist Point of View," *Westminster Review* 125 (1886)

Bologh, Roslyn Wallach, "Marx, Weber, and Masculine Theorizing: A Feminist Analysis," in *The Marx-Weber Debate* Norbert Wiley, ed. (Newbury Park: Sage, 1987)

Brod, Harry, "Pornography and the Alienation of Male Sexuality," *Social Theory and Practice* 14 (1988), 265-284.

Cocks, Joan, "Cultural Theory Looks at Identity and Contradiction," *Quest* (1990), 38-60.

Dunayevskaya, Raya, "Marx's ‘New Humanism’ and the Dialectics of Women's Liberation in Primitive and Modern Societies," *Praxis International* 3-4 (1984)

Glass, Marvin and Ernie Thompson, "Reproduction for Money: Marxist Feminism and Surrogate Motherhood," *Nature, Society, and Thought* 7, 3 (1994), 281-297.

Goldstein, Leslie, "Mill, Marx, and Women's Liberation," *Journal of the History of Philosophy* XVIII, 3 (July 1980)

Hartsock, Nancy C. M., "The Feminist Standpoint: Developing the Ground for a Specifically Feminist Historical Materialism," in *Feminism and Philosophy: Essential Readings in Theory, Reinterpretation, and Application* (Boulder: Westview Press, 1995)

Held, Virginia, "Marx, Sex, and the Transformation of Society," *The Philosophical Forum* 5, 1-2 (1973-1974), 168-184. Reprinted in *Women and Philosophy: Toward a Theory of Liberation* Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Henderson, Janet, "An Eco-Feminist Critique of Marx," *Dialogue (PST)* 31 (1989), 58-64.

Jaggar, Alison, "Human Biology in Feminist Theory: Sexual Equality Reconsidered," in *Beyond Domination* Carol C. Gould., ed. (Totowa: Rowman & Allanheld, 1984)

Jones, Kathleen B., "Socialist-Feminist Theories of the Family," *Praxis International* 8 (1988), 284-300.

Kain, Philip J., "Marx, Housework, and Alienation," *Hypatia* 8, 1 (1993), 121-144.

Kain, Philip J., "Modern Feminism and Marx," *Studies in Soviet Thought* 44, 3 (1992), 159-192. (German)

Marcuse, Herbert, "Marxism and Feminism," *Women's Studies* 2, 3 (1974), 279-288.

Nicholson, Linda, "Feminism and Marx: Integrating Kinship with the Economic," *Praxis International* 5 (1986), 367-380.

O'Brien, Mary, "Reproducing Marxist Man," in *Feminism and Philosophy: Essential Readings in Theory, Reinterpretation, and Application* (Boulder: Westview Press, 1995)

Pasquinelli, Carla, "Beyond the Longest Revolution: The Impact of the Italian Women's Movement on Cultural and Social Change," *Praxis International* 4 (1984), 131-136.

Schmitt, Richard, "Alienation and Autonomy," *Praxis International* 8 (1988), 222-236.

Schmitt, Richard, "A New Hypothesis about the Relations of Class, Race, and Gender: Capitalism as a Dependent System," *Social Theory and Practice* 14 (1988), 345-365.

Friedrich Engels

Books

Sayers, Janet, Mary Evans, and Nenneke Redclift, eds. 1987. *Engels Revisited: New Feminist Essays* (London: Travistock)

Vogel, Lise, 1983. *Marxism and the Oppression of Women: Toward a Unitary Theory* (London: Pluto)

Articles

Carling, Alan, "Rational Choice Marxism and Postmodern Feminism: Towards a More Meaningful Incomprehension," in *Rational Choice Marxism* Terrell Carver, ed. (University Park: Pennsylvania State University Press, 1995)

Carvel, Terrel, "Engels' Feminism," *History of Political Thought* VI, 3 (winter 1985)

Dunayevskaya, Raya, "Marx's ‘New Humanism’ and the Dialectics of Women's Liberation in Primitive and Modern Societies," *Praxis International* 3 (1984), 369-381.

Soren Kierkegaard

Books

Leon, Celine and Sylvia Walsh, eds. 1997. *Feminist Interpretations of Soren Kierkegaard* (University Park: Pennsylvania State University Press)

Articles

Armstrong, Susan, "A Feminist Reading of Hegel and Kierkegaard," in *Hegel, History, and Interpretation* Shaun Gallagher, ed. (Albany: SUNY Press, 1997)

Berry, Wanda Warren, "Kierkegaard and Feminism: Apologetic, Repetition, and Dialogue in Kierkegaard," in *Post/Modernity* Martin J. Matustik, ed. (Bloomington: Indiana University Press, 1995)

Cahoy, William J., "One Species or Two: Kierkegaard's Anthropology and the Feminist Critique of the Concept of Sin," *Modern Theology* 11, 4 (1995), 429-454.

Howe, Leslie A., "Kierkegaard and the Feminine Self," *Hypatia* 9, 4 (1994), 131-157.

Kruks, Sonia, "Existentialism and Phenomenology," in *A Companion to Feminist Philosophy,* Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Makarushka, Irena, "Reflections on the ‘Other’ in Dineson, Kierkegaard, and Nietzsche," in *Kierkegaard on Art and Communication* George Pattison, ed. (New York: St. Martin's Press, 1992)

McBride, William L., "Sartre's Debts to Kierkegaard: A Partial Reckoning," in *Post/Modernity* Martin J. Matustik, ed. (Bloomington: Indiana University Press, 1995)

Walsh, Sylvia I., "On ‘Feminine’ and ‘Masculine’ Forms of Despair," in *The Sickness Unto Death* Robert L. Perkins, ed. (Macon: Mercer University Press, 1987)

Walsh, Sylvia I., "Subjectivity versus Objectivity: Kierkegaard's ‘Postscript’ and Feminist Epistemology," in *International Kierkegaard Commentary* Robert L. Perkins, ed. (Macon: Mercer University Press, 1997)

Friedrich Nietzsche

Books

Graybeal, Jean, 1990. *Language and ‘The Feminine’ in Nietzsche and Heidegger* (Bloomington: Indiana University Press)

Perez Estevez, Antonio, 1989. *El individuo y la feminidad* (Zulia: University of Zulia) (Portugese)

Schutte, Ofelia, 1984. *Beyond Nihilism: Nietzsche Without Masks* (Chicago: University of Chicago Press)

Articles

Ainley, Alison, "Ideal Selfishness: Nietzsche's Metaphor of Maternity," in *Exceedingly Nietzsche* David Farrell Krell, ed. (London: Routledge and K. Paul, 1988) 116-130.

Addelson, Kathryn Pyne, "Awakening," *Hypatia* WSIF 1 (1983), 583-595.

Armour, Ellen T., "Questions of Proximity: ‘Woman's Place’ in Derrida and Irigaray," *Hypatia* 12,1 (1997), 63-78.

Behler, Diana, "Nietzsche and Postfeminism," *Nietzsche Studien* 22 (1993), 355-370.

Bergoffen, Debra B., "Nietzsche Was No Feminist..." *International Studies in Philosophy* 26, 3 91994), 23-31.

Bergoffen, Debra B., "On the Advantage and Disadvantage of Nietzsche for Women," in *The Question of the Other* Arleen B. Dallery, ed. (Albany: SUNY Press, 1989) 77-88.

Bertram, Maryanne J., "’God's "Second" Blunder’--Serpent Woman and the ‘Gestalt’ in Nietzsche's Thought," *S. J. Phil.* 19 (1981), 259-278.

Booth, David, "Nietzsche's ‘Woman’ Rhetoric," *History of Philosophy Quarterly* 8, 3 (1991), 311-325.

Burney Davis, Terri, "The Vita Femina and Truth," *History of European Ideas* (1989), 841-847.

Card, Claudia, "Genealogies and Perspectives: Feminist and Lesbian Reflections," *International Studies in Philosophy* 28, 3 (1996), 99-111.

Clark, Maudemarie, "Nietzsche's Misogyny," *International Studies in Philosophy* 26, 3 (1994), 3-12.

Diethe, Carol, "Nietzsche and the Woman Question," *History of European Ideas* (1989), 865-875.

Diprose, Rosalyn, "Nietzsche, Ethics, and Sexual Difference," *Radical Philosophy* 52 (1989), 27-33.

Freyelberg, Bernard D., "Nietzsche in Derrida's Spurs: Deconstruction as Deracination," *History of European Ideas* (1989), 685-692.

Higgins, Kathleen Marie, "Gender in 'The Gay Science,'" *Philosophy and Literature* 19, 2 (1995), 227-247.

Holm, Elly and Paul Cilliers, "Beyond the Politics of Positionality: Deconstruction and Feminism," *South African Journal of Philosophy* 17, 4 (1998), 377-394.

Irigaray, Luce, 1991. *Marine Lover of Friedrich Nietzsche* Gillian C. Gill, trans. (New York: Columbia University Press)

Johnson, Pauline, "Nietzsche Reception Today," *Radical Philosophy* 80 (1996), 24-33.

Joos, Ernest, "Nietzsche et les Femmes," *Laval Theol. Phil.* 41 (1985), 305-315. (French)

Kaufman, Cynthia, "Knowledge as Masculine Heroism or Embodied Perception: Knowledge, Will, and Desire in Nietzsche," *Hypatia* 13, 4 (1998), 63-87.

Lorraine, Tamsin, "Nietzsche and Feminism: Transvaluing Women in 'Thus Spoke Zarathustra," *International Studies in Philosophy* 26, 3 (1994), 13-21.

Makarushka, Irena, "Reflections on the 'Other' in Dineson, Kierkegaard, and Nietzsche," in *Kierkegaard on Art and Communication* George Pattison, ed. (New York: St. Martin's Press, 1992)

Malet, N., "L'Homme et la Femme dans la Philosophie de Nietzsche," *Rev. Metaph. Morale* 82 (1977), 38-63.

Marion, Jean Luc, "The Exactitude of the 'Ego,'" *American Catholic Philosophical Quarterly* 67, 4 (1993), 561-568.

Menck, Christoph, "Schwerpunkt: Nietzsche und die Praktische Philosophie," *Deutsche Zeitschrift fur Philosophie* 41, 5 (1993), 828-830. (German)

Mortensen, Ellen, "Irigaray and Nietzsche: Echo and Narcissus Revisited?" in *The Fate of the New Nietzsche* Keith Ansell-Pearson, ed. (Brookfield: Avebury, 1993)

Munnich, Susana, "En torno a la frase de Nietzsche 'Le verdad es mujar,'" *Convivium* 9 (1996), 77-91. (Spanish)

Oliver, Kelly, "The Plaint of Ariadne: Luce Irigaray's 'Amante Marine de Friedrich Nietzsche," in *The Fate of the New Nietzsche* Keith Ansell-Pearson, ed. (Brookfield: Avebury, 1993)

Orniston, Gayle, "Traces of Derrida: Nietzsche's Image of Woman," *Philosophy Today* 28 (1984), 178-

188.

Owen, David, "Nietzsche's Squandered Seductions: Feminism, the Body, and the Politics of Genealogy," in *The Fate of the New Nietzsche* Keith Ansell-Pearson, ed. (Brookfield: Avebury, 1993)

Parens, Erik, "Derrida, 'Woman,' and Politics: A Reading of 'Spurs,'" *Philosophy Today* 33, 4 (1989), 291-301.

Pasons, Katherine Pyne, "Nietzsche and Moral Change," *Feminist Studies* 2 (1974), 57-76.

Reguera, Isidoro, "El Nietzsche practico de Nolte: Nietzsche, profeta tragico de la guerra y organizador politico de la aniquilacion," *Revista de Filosofia (Spain)* 8, 14 91996), 127-157. (Spanish)

Schrift, Alan D., "On the Gift-Giving Virtue: Nietzsche's Unacknowledged Feminine Economy," *International Studies in Philosophy* 26, 3 (1994), 33-44.

Thompson, J. L., "Nietzsche on Woman," *International Journal of Moral and Social Studies* 1990, 207-220.

Wischke, Mirko, "The Conflict of Morality with Basic Life Instincts," *Filozofska Istrazivanja* 15, 4 (1995), 673-681. (Serbo-Croatian)

Wischke, Mirko, "The Conflict of Morality with Basic Life Instincts," *Synthesis Philosophica* 11, 1 (1996), 39-48.

Charles Darwin

Books

Sayers, Janet, 1982. *Biological Politics* (London: Tavistock)

Articles

Gates, Barbara T., "Revisioning Darwin, with Sympathy," *History of European Ideas* 19, 4-6 (1994), 761-768.

Richards, Evelleen, "Darwin and the Descent of Women," in *The Wider Domain of Evolutionary Thought* David Oldroyd, ed. (Dordrecht: Reidel, 1983), 57-112.

Other Nineteenth Century Philosophy

Books

Benstock, Shari, ed. 1987. *Feminist Issues in Literary Scholarship* (Bloomington: Indiana University Press)

Boller, Paul Jr., 1974. *American Transcendentalism, 1830-1860: An Intellectual Inquiry* (New York: Putnam)

Eisenstein, Zillah R. 1981. *The Radical Future of Liberal Feminism* (New York: Longman)

McElroy, Wendy, ed. 1982. *Freedom, Feminism, and the State* (Washington DC: Cato Institute)

Articles

Altman, Elizabeth C., "The Philosophical Bases of Feminism: The Feminist Doctrines of the Saint-Simonians and Charles Fourier," *Philosophical Forum (Boston)* 7 (1976), 277-293.

Bar-On, Bat Ami, "The Feminist ‘Sexuality Debates’ and the Transformation of the Political," *Hypatia* 7, 4 (1992), 45-58.

Berry, Edmund G., "Margaret Fuller Ossoli, 1810-1850," *Dalhousie Review* 30 (1950), 369-376.

Brouwer, Christien, "Nature in Terms of Femininity: The Case of Nineteenth Century Plant Geography," *Commun. Cog*. 21 (1988), 129-132.

Bruland, Esther Byle, "Evangelical and Feminist Ethics: Complex Solidarities," *Journal of Religious Ethics* 17, 2 (1989), 139-160.

DuBois, Ellen, "The Radicalism of the Woman Suffrage Movement: Notes Toward the Reconstruction of Nineteenth-Century Feminism," *Feminist Studies* 3 (1975), 63-71.

Goldstein, Leslie Friedman, "Early European Feminism and American Women," in *Women's Rights and the Rights of Man*, A. J. Arnaud, ed. (Oxford: Aberdeen, 1990)

Goldstein, Leslie Friedman, "Early Feminist Theories in French Utopian Socialism: The St. Simonians and Fourier," *Journal of the History of Ideas* 43 (1982), 91-108.

Gordon, Linda and Ellen DuBois, "Seeking Ecstasy on the Battlefield: Danger and Pleasures in Nineteenth Century Feminist Sexual Thought," *Feminist Studies* 9 (1983), 7-26.

McLaren, Angus, "Sex and Socialism: The Opposition of the French Lefts to Birth Control in the

Nineteenth Century," *Journal of the History of Ideas* 37 (1976), 475-492.

Palmer, L. M., "The End of the End of Ideology," *History of European Ideas* 49, 4-6 (1994), 709-713.

Pedersen, Joyce Senders, "Education, gender, and Social Change in Victorian Liberal Feminist Theory," *History of European Ideas* 8 (1987), 503-519.

Polyakov, L. V., "Women's Emancipation and the Theology of Sex in Nineteenth Century Russia," *Philosophy East and West* 42, 2 (1992), 297-308.

Riot-Sarcey, Michele and Eleni Varikas, "Feminist Consciousness in the Nineteenth Century: The Consciousness of a Pariah?" *Praxis International* 5 (1986), 443-465.

Schafer, Sylvia, "When the Child is the Father of the Man: Work, Sexual Difference, and the Guardian-State in Third Republic France," *History and Theory* 31, 4 (1992), 98-115.

Sears, W. P., "The Educational Theories of Louisa May Alcott," *Dalhousie Review* 27 (1947), 327-334.

Wosk, Julie, "The ‘Electric Eve’: Galvanizing Women in Nineteenth and Twentieth Century Art and Technology," *Research in Philosophy and Technology* 13 (1993), 43-56.

Freud

Books

Cooke, Brett, George E. Slusser, and Jaume Marti-Olivella, eds. 1998. *The Fantastic Other: An Interface of Perspectives* (Amsterdam: Rodopi)

Flax, Jane, 1989. *Thinking Fragments: Psychoanalysis, Feminism, and Postmodernism in the Contemporary West* (Berkeley: University of California Press)

Mills, Patricia Jagentowicz, 1987. *Woman, Nature, and Psyche* (New Haven: Yale University Press)

O'Neill, John, ed. 1996. *Freud and the Passions* (University Park: Pennsylvania University Press)

Sayers, Janet, 1982. *Biological Politics* (London: Tavistock)

Articles

Benjamin, Jessica, "The Shadow of the Other (Subject): Intersubjectivity and Feminist Theory," *Constellations* 1, 2 (1994), 231-254.

Brennan, Teresa, "Essence Against Identity," *Metaphilosophy* 27, 1-2 (1996), 92-103.

Davis, Karen Elizabeth, "I Love Myself When I am Laughing: A New Paradigm for Sex," *Journal of Social Philosophy* (1990), 5-24.

Ferguson, Ann, "Motherhood and Sexuality: Some Feminist Questions," *Hypatia* 1 (1986), 3-22.

Fuchs, Jo Ann Pilardi, "On the War Path and Beyond: Hegel, Freud, and Feminist Theory," *Hypatia* WSIF 1 (1986), 565-572.

Harrington, Thea, "The Speaking Abject in Kristeva's ‘Powers of Horror,’" *Hypatia* 13, 1 (1998), 138-157.

Hengehold, Laura, "Rape and Communicative Agency: Reflections in the Lake at L-" *Hypatia* 8, 4 (1993), 56-71.

Keiser, R. Melvin, "Postcritical Religion and the Latent Freud," *Zygon* (1990), 433-447.

Kittay, Eva Feder, "Rereading Freud on ‘Femininity,’ or Why Not ‘Womb’ Envy," *Hypatia* WSIF 2 (1984), 385-391.

Mothersill, Mary, "Notes on Feminism," *Monist* 57 (1973), 105-114.

Nissim Sabat, Marilyn, "Freud, feminism, and faith," *Listening* 20 (1985), 208-220.

Oliver, Kelly, "Fleshy Memory," *Radical Philosophy* 65 (1993), 30-32.

Pawlowski, Pawel Maciej, "On Some Philsophical Problems of Sigmund Freud's Psychoanalytic Theory," *Kwartalnik Filozoficzny* 26, 2 (1998), 101-114. (Polish)

Whitbeck, Caroline, "Theories of Sex Difference," in *Women and Philosophy: Toward a Theory of Liberation* Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Jung

Books

Cooke, Brett, George E. Slusser, and Jaume Marti-Olivella, eds. 1998. *The Fantastic Other: An Interface of Perspectives* (Amsterdam: Rodopi)

Stevernagel, Gertrude A., 1979. *Political Philosophy as Therapy: Marcuse Reconsidered* (Westport: Greenwood)

Articles

Keller, Catherine, "Reconnecting: A Reply to Robert Moore," in *Archetypal Process: Self and Divine in Whitehead, Jung, and Hillman* (Evanston: Northwestern University Press)

Valle, Valerie A. and Elizabeth L. Kruger, "The Nature and Expression of Feminine Consciousness," in *The Metaphors of Consciousness* Rolf von Eckartsberg, ed. (New York: Plenum Press, 1981)

Whitbeck, Caroline, "Theories of Sex Difference," in *Women and Philosophy: Toward a Theory of Liberation* Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

American Philosophy

Pragmatism

Books

Seigfried, Charlene Haddock, 1996. *Pragmatism and Feminism: Reweaving the Social Fabric* (Chicago: University of Chicago Press)

Articles

Mahowald, Mary Briody, "What Classical American Philosophers Missed: Jane Addams, Critical Pragmatism, and Cultural Feminism," *Journal of Value Inquiry* 31, 1 (1997), 39-54.

Miller, Marjorie C., "Feminism and Pragmatism," *Monist* 75, 4 (1992), 445-457

Miller, Marjorie C., "Response to Eugenie Gatens-Robinson, Marcia K. Moen, and Felicia E. Kruse," *Transactions of the Charles S. Peirce Society* (1991), 465-474.

Rooney, Phyllis, "Feminist-Pragmatist Revisionings of Reason, Knowledge, and Philosophy," *Hypatia* 8, 2 91993), 15-37.

Seigfried, Charlene Haddock, "The Missing Perspective: Feminist Pragmatism," *Transactions of the Charles S. Peirce Society* (1991), 329-337.

Seigfried, Charlene Haddock, "Pragmatism," in *A Companion to Feminist Philosophy*, Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

William James

Radin, Margaret Jane, "The Pragmatist and the Feminist," in *Pragmatism in Law and Society* Michael Brint, ed. (Boulder: Westview Press, 1991)

Santayana

Miller, Marjorie C., "Essence and Identity: Santayana and the Category 'Women,'" *Transactions of the Charles S. Peirce Society* 30, 1 (1994), 33-50.

Charles S. Peirce

Ayim, Maryann, "The Implications of Sexually Stereotypic Language as Seen Through Peirce's Theory of Signs," *Transactions of the Charles S. Peirce Society* 19 (1983), 183-198.

Chopp, Rebecca S., "Feminist Queries and Metaphysical Musings," *Modern Theology* 11, 1 (1995), 47-63.

Moen, Marcia K., "Peirce's Pragmatism as a Resource for Feminism," *Transactions of the Charles S. Peirce Society* (1991), 435-450.

Sharp, Ann Margaret, "Peirce, Feminism, and Philosophy for Children," in *Children: Thinking and Philosophy* Daniela G. Camhy, ed. (Sankt Augustin: Academia, 1994)

John Dewey

Boisvert, Raymond D., "Heteronomous Freedom," in *Philosophy and the Reconstruction of Culture* John Stuhr, ed. (Albany: SUNY Press, 1994)

Capps, John, "Pragmatism, Feminism, and the Sameness-Difference Debate," *Transactions of the Charles S. Peirce Society* 32, 1 (1997), 39-54.

Clark, Ann, "The Quest for Certainty in Feminist Thought," *Hypatia* 8, 3 (1993), 84-93.

Duhan, Laura, "Feminism and Peace Theory: Women as Nurturers versus Women as Public Citizens," in *In the Interest of Peace: A Spectrum of Philosophical Views* (Wolfeboro: Longwood, 1990)

Gatens-Robinson, Eugenie, "Dewey and the Feminist Successor Science Project," *Transactions of the Charles S. Peirce Society* (1991), 417-433.

Giarelli, James M., "Dewey and the Feminist Successor Pragmatism Project," *Free Inquiry* 13, 1 (1993),

30-31.

Hart, Carroll Guen, "’Power in the Service of Love’: John Dewey's ‘Logic’ and the Dream of a Common Language," *Hypatia* 8, 2 (1993), 1990-214.

Heldke, Lisa and Stephen Kellers, "Objectivity as Responsibility," *Metaphilosophy* 26, 4 (1995), 360-378.

Leach, Mary, "(Re)searching Dewey for Feminist Imaginaries: Linguistic Continuity, Discourse, and Gossip," *Studies in Philosophy and Education* 13, 3-4 (1995), 291-306.

Pappas, Gregory Fernando, "Dewey and Feminism: The Affective and Relationships in Dewey's Ethics," *Hypatia* 8, 2 (1993), 78-95.

Rethorst, John C., "Myth and Morality," *Journal of Moral Education* (1991), 329-337.

Seigfried, Charlene Haddock, "John Dewey's Pragmatist Feminism, " in *Reading Dewey* Larry A. Hickman, ed. (Bloomington: Indiana University Press, 1998)

Seigfried, Charlene Haddock, "Validating Women's Experience Pragmatically," in *Philosophy and the Reconstruction of Culture* John Stuhr, ed. (Albany: SUNY Press, 1994)

Sorrell, Kory Spencer, "Feminist Ethics and Dewey's Moral Theory," *Transactions of the Charles S. Peirce Society* 35, 1 (1999), 89-114.

Sullivan, Shannon, "Teaching as a Pragmatist: Relating Non-Foundational Theory and Classroom Practice," *Teaching Philosophy* 20, 4 (1997), 401-419.

Upin, Jane, "Charlotte Perkins Gilman: Instrumentalism Beyond Dewey," *Hypatia* 8, 2 (1993), 15-37.

G. E. Moore

Martin, Bill, "’To the Lighthouse’ and the Feminist Path to Posmodernity," *Philosophy and Literature* 13 (1989), 307-315.

Bertrand Russell

Harrison, Brian, "Bertrand Russell: The False Consciousness of a Feminist," *Russell* 4 (1984), 157-206.

Martin, Bill, "’To the Lighthouse’ and the Feminist Path to Posmodernity," *Philosophy and Literature* 13 (1989), 307-315.

Tait, Katharine, "Russell and Feminism," *Russell* 29 (1978), 5-16.

Alfred North Whitehead

Books

Davaney, Sheila Greeve, ed. 1981. *Feminism and Process Thought* (New York: Mellen Press)

Articles

Cobb, John Jr., "Whiteheadian Thought," *Dialogue and Humanism* 1,2 (1991), 79-91.

Havell, Nancy R., "The Promise of a Process Feminist Theory of Relations," *Process Studies* 17 (1988), 78-87.

Thie, Marilyn, "Feminist Concerns and Whitehead's Theory of Perception," *Process Studies* 8 (1978), 186-191.

Ayn Rand

Books

Goldstein, Mimi Reisel and Chris Matthews Sciabarra, eds. 1991. *Feminist Interpretations of Ayn Rand* (University Park: Pennsylvania University Press)

Merrill, Ronald E., 1991. *The Ideas of Ayn Rand* (Peru: Open Court)

Rawls

Books

Richards, Janet Radcliffe, 1980. *The Sceptical Feminist: A Philosophical Inquiry* (Boston: Routledge and K. Paul)

Articles

Agra, Maria Xose, "Justicia y Genero: Algunas cuestiones relevantes en torno a la teoria de la justicia de J. Rawls," *Anales de la Catedra Francisco Suarez* 31 (1994), 123-145. (Spanish)

Anderson, David, "False Stability in Rawlsian Liberalism," *Contemporary Philosophy* 14, 5 (1992), 11-

16.

Baehr, Amy R., "Toward a New Feminist Liberalism: Okin, Rawls, and Habermas," *Hypatia* 11,1 (1996), 49-66.

Cornell, Drucilla, "Response to Thomas McCarthy: The Political Alliance Between Ethical Feminism and Rawls's Kantian Constructivism," *Constellations* 2,2 (1995), 189-206.

Gatens, Moira, "Between the Sexes: Care or Justice?" in *Introducing Applied Ethics,* Brenda Almond, ed. (Cambridge: Blackwell, 1995)

Green, Karen, "Rawls, Women, and the Priority of Liberty," *Australasian Journal of Philosophy* Supp. 64 (1986), 26-36.

Hedman, Carl, "Ethics and Group Conflict: Between Marxism and Liberalism," *Radical Philosophy* 46 (1987), 8-15.

Meyers, Diana Tietjens, "Moral Reflections: Beyond Impartial Reason," *Hypatia* 8, 3 (1993), 21-47.

Okin, Susan Moller, "Justice and Gender," *Philosophy and Public Affairs* 16 (1987), 42-72.

Okin, Susan Moller, "'Political Liberalism,' Justice, and Gender," *Ethics* 105, 1 (1994), 23-43.

Okin, Susan Moller, "Reason and Feeling in Thinking about Justice," *Ethics* 99 (1989), 229-249.

Pateman, Carole, "'The Disorder of Women': Women, Love, and the Sense of Justice," *Ethics* 91 (1980), 20-31.

Russell, J. S., "Okin's Rawlsian Feminism? Justice in the Family and Another Liberalism," *Social Theory and Practice* 21, 3 (1995), 397-426.

Sehon, Scott, "Okin on Feminism and Rawls," *Philosophical Forum* 27, 4 (1996), 321-332.

Shaw, Beverly, "Sexual Justice and the Sceptical Feminist," *Journal of Applied Philosophy* 1 (1984), 115-122.

Thompson, Janna, "What Do Women Want? Rewriting the Social Contract," *International Journal of Moral and Social Studies* 8, 3 (1993), 257-272.

Richard Rorty

Books

Fraser, Nancy, 1989. *Unruly Practices: Power, Discourse, and Gender in Contemporary Social Theory* (Minneapolis: University of Minnesota Press)

Olthius, James H., ed. 1997. *Knowing Other-Wise: Philosophy at the Threshold of Spirituality* (New York: Fordham University Press)

Rothleder, Dianne, 1999. *The Work of Friendship: Rorty, His Critics, and the Project of Solidarity* (Albany: SUNY Press)

Articles

Amoros, Celia, "Richard Rorty and the 'Tricoteuses,'" *Constellations* 3, 3 (1997), 364-376.

Fraser, Nancy, "Solidarity or Singularity? Richard Rorty Bteween Romanticism and Technocracy," *Praxis International* 8 (1988), 257-272. Reprinted in *Reading Rorty* (Cambridge: Blackwell, 1991)

Fritzman, J. M., "Thinking With Fraser About Rorty, Feminism, and Pragmatism," *Praxis International* 13, 2 (1993), 113-125.

Kaufman-Osborn, Timothy W., "Teasing Feminist Sense From Experience," *Hypatia* 8, 2 (1993), 124-144.

Leland, Dorothy, "Rorty on the Moral Concern Philosophy: A Critique Froma Feminist Point of View," *Praxis International* 8 (1988), 273-283.

Schultz, Bart, "Comment: The Private and Its Problems--Pragmatism, Pragmatic Feminism, and Homophobia," *Philosophy of the Social Sciences* 29, 2 (1999), 281-305

Skillen, Tony, "Reply to Richard Rorty's 'Feminism and Pragmatism': Richard Rorty--Knight Errant," *Radical Philosophy* 62 (1992), 24-26.

Wilson, Catherine, "Reply to Richard Rorty's 'Feminism and Pragmatism': How Did the Dinosaurs Die Out? How Did the Poets Survive?" *Radical Philosophy* 62 (1992), 20-23.

Ludwig Wittgenstein

Lampshire, Wendy Lee, "Decisions of Identity: Feminist Subjects and Grammars of Sexuality," *Hypatia* 10, 4 (1995), 32-45.

Lampshire, Wendy Lee, "History as Genealogy: Wittgenstein and the Feminist Deconstruction of Objectivity," *Philosophy and Theology* (1991), 313-331.

Lampshire, Wendy Lee, "Women--Animals--Machines: A Grammar for a Wittgensteinian Ecofeminism," *Journal of Value Inquiry* 29, 1 (1995), 89-101.

Martin, Bill, "'To the Lighthouse' and the Feminist Path to Postmodernity," *Philosophy and Literature* 13 (1989), 307-315.

Orr, Deborah, "On Logic and Moral Voice," *Informal Logic* 17, 3 (1995), 347-363.

German Philosophy

Edmund Husserl

Books

Casebien, Allan, 1992. *Film and Phenomenology: Toward a Realist Theory of Cinematic Representation* (New York: Cambridge University Press)

Articles

Bergoffen, Debra B., "From Husserl to de Beauvoir: Gendering the Perceived Subject," *Metaphilosophy* 27, 1-2 91996), 53-62.

Nissim Sabat, Marilyn, "The Crisis in Psychoanalysis: Resolution Through Husserlian Phenomenology and Feminism," *Human Studies* (1991), 33-66.

Willis, Clyde E., "The Phenomenology of Pornography: A Comment on Catharine MacKinnon's 'Only Words,'" *Law and Philosophy* 16, 2 (1997), 177-199.

Heidegger

Books

Graybeal, Jean, 1990. *Language and "The Feminine" in Nietzsche and Heidegger* (Bloomington: Indiana University Press)

Holland, Nancy, 1998. *The Madwoman's Dream: The Concept of the Appropriate in Ethical Thought* (University Park: Pennsylvania State University Press)

Huntington, Patricia J., 1998. *Ecstatic Subjects, Utopia, and Recognition: Kristeva, Heidegger, Irigaray* (Albany: SUNY Press)

Articles

Holland, Nancy J., "Heidegger and Derrida Redux: A Close Reading," in *Hermeneutics and Deconstruction,* Hugh J. Silverman, ed. (Albany: SUNY Press, 1985)

Klawiter, Maren, "Using Arendt and Heidegger to Consider Feminist Thinking on Women and Reproductive/Infertility Technologies," *Hypatia* (1990), 65-89.

Kruks, Sonia, "Existentialism and Phenomenology," in *A Companion to Feminist Philosophy,* Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Vasey, Craig R., "Faceless Women and Serious Others," in *Ethics and Danger,* Arleen B. Dallery, ed. (Albany: SUNY Press, 1992)

Hannah Arendt

Books

Bickford, Susan, 1996. *The Dissonance of Democracy: Listening, Conflict, and Citizenship* (Ithaca: Cornell University Press)

Disch, Lisa J., 1994. *Hannah Arendt and the Limits of Philosophy: With a New Preface* (Ithaca: Cornell University Press)

Honig, Bonnie, ed. 1995. *Feminist Interpretations of Hannah Arendt* (University Park: Pennsylvania State University Press)

Hutchings, Kimberly, 1996. *Kant, Critique, and Politics* (New York: Routledge)

Articles

Allen, Amy, "Solidarity After Identity Politics: Hannah Arendt and the Power of Feminist Theory," *Philosophy and Social Criticism* 25, 1 (1999), 97-118.

Benhabib, Seyla, "Feminist Theory and Hannah Arendt's Concept of Public Space," *History of the Human Sciences* 6, no. 2: 97-114.

Cutting Gray, Joanne, "Hannah Arendt, Feminism, and the Politics of Alterity: ‘What Will We Lose if

We Win,'" in *Hypatia's Daughters: Fifteen Hundred Years of Women Philosophers* McAlister, Linda Lopez, ed. (Bloomington: Indiana University Press, 1996)

Duhan, Laura, "Feminism and Peace Theory: Women as Nurturers versus Women as Public Citizens," in *In the Interest of Peace: A Spectrum of Philosophical Views* (Wolfeboro: Longwood, 1990)

Franco, Vittoria, "Agnes Heller, una vita per l'autonomia e la liberta," *Iride* 8, 16 (1995), 544-602. (Italian)

Klawiter, Maren, "Using Arendt and Heidegger to Consider Feminist Thinking on Women and Reproductive/Infertility Technologies," *Hypatia* (1990), 65-89.

Long, Christopher Philip, "A Fissure in the Distinction: Hannah Arendt, the Family, and the Public/Private Dichotomy," *Philosophy and Social Criticism* 24, 5 (1998), 85-104.

MacCannell, Juliet Flower, "Facing Fascism: A Feminine Politics of Jouissance," *Topoi* 12, 2 (1993), 137-151.

Mann, Patricia S., "Toward a Postpatriarchal Society," in *Norms and Values: Essays on the Work of Virginia Held,* Joram Graf Haber, ed. (Lanham: Rowman and Littlefield, 1998)

Moynagh, Patricia, "A Politics of Enlarged Mentality: Hannah Arendt, Citizenship Responsibility, and Feminism," *Hypatia* 12, 4 (1997), 27-53.

Nye, Andrea, "Friendship Across Generations," *Hypatia* 11, 3 (1996), 154-160.

Winant, Terry, "The Feminist Standpoint: A Matter of Language," *Hypatia* 2 (1987), 123-148.

Young-Bruehl, Elisabeth, "Hannah Arendt Among Feminists," in *Hannah Arendt: Twenty Years Later,* Larry May, ed. (Cambridge: MIT Press, 1996)

Frankfurt School (Horkheimer, Adorno, Marcuse)

Books

Alway, Joan, 1995. *Critical Theory and Political Possibilities: Conceptions of Emancipatory Politics in the Works of Horkheimer, Adorno, Marcuse, and Habermas* (Westport: Greenwood Press)

Holub, Renate, 1992. *Antonio Gramsci: Beyond Marxism and Postmodernism* (New York: Routledge)

Ingram, David, 1990. *Critical Theory and Philosophy* (New York: Paragon House)

Mills, Patricia Jagentoowicz, 1987. *Woman, Nature, and Psyche* (New Haven: Yale University Press)

Articles

Ryle, Martin, "Histories of Cultural Populism," *Radical Philosophy* 78 (1996), 27-33.

Adorno

Donovan, Josephine, "Everyday Use and Moments of Being: Toward a Nondominative Aesthetic," in *Aesthetics in Feminist Perspective*, Hilde Hein, ed. (Bloomington: Indiana University Press, 1993)

Phelan, Shane, "The Jargon of Authenticity: Adorno and Feminist Essentialism," *Philosophy and Social Criticism* (1990), 39-54.

Wilke, Sabine and Heidi Schlipphacke, "Construction of Gendered Subject: A Feminist Reading of Adorno's ‘Aesthetic Reading,’" in *The Semblance of Subjectivity*, Tom Huhn, ed. (Cambridge: MIT Press, 1997)

Max Horkheimer

Rumpf, Mechthild, "’Mystical Aura’: Imagination and the Reality of Maternal in Horkheimer's Writings," in *On Max Horkheimer* Seyla Benhabib, ed. (Cambridge: MIT Press, 1993)

Marcuse

Books

Perez Estevez, Antonio, 1989. *El individuo y la feminidad* (Zulia: University of Zulia) (Portugese)

Stevernagel, Gertrude A., 1979. *Political Philosophy as Therapy: Marcuse Reconsidered* (Westport: Greenwood)

Articles

Farganis, Sondra, "Liberty: Two Perspectives on the Women's Movement," *Ethics* 88 (1977), 62-73.

Jurgen Habermas

Books

Fraser, Nancy, 1989. *Unruly Practices: Power, Discourse, and Gender in Contemporary Social Theory* (Minneapolis: University of Minnesota Press)

Hutchings, Kimberly, 1996. *Kant, Critique, and Politics* (New York: Routledge)

Kelly, Michael, ed. 1994. *Critique and Power* (Cambridge: MIT Press)

Articles

Baehr, Amy R., "Toward a New Feminist Liberalism: Okin, Rawls, and Habermas," *Hypatia* 11,1 91996), 49-66.

Couture, Tony, "Feminist Criticisms of Habermas' Ethics and Politics," *Dialogue* 34, 2 (1995), 259-279.

Crocker, Nancy, "The Problem of Community," *Southwest Philosophical Studies* 19 (1992), 50-62.

Fleming, Marie, "Women and the ‘Public Use of Reason,’" *Social Theory and Practice* 19, 1 (1993), 27-50.

Fraser, Nancy, "Michel Foucault: A ‘Young Conservative’?" *Ethics* 96 (1985), 165-184.

Herrera, Maria, "Equal Respect Among Unequal Partners: Gender Difference and the Constitution of Moral Subjects," *Philosophy East and West* 42, 2 (1992), 263-275.

Landes, Joan B., "Jurgen Habermas's ‘The Structural Transformation of the Public Sphere’: A Feminist Inquiry," *Praxis International* 12, 1 (1992), 106-127.

Pamerleau, William C., "Can Habermas' Discourse Ethics Accommodate the Feminist Perspective?" in *Rending and Renewing the Social Order,* Yeager Hudson, ed. (Lewiston: Mellen Press, 1996)

Still, Judith, "What Foucault Fails to Acknowledge...: Feminism and ‘The History of Sexuality,’" *History of Human Sciences* 7, 2 (1994), 150-157.

Young, Iris Marion, "Impartiality and the Civic Public: Some Implications of Feminist Critiques of Moral and Political Theory," *Praxis International* 5 (1986), 381-401.

French 20th Century Philosophy

Books

Allen, Jeffner and Iris Marion Young, eds. 1989. *The Thinking Muse: Feminism and Modern French*

Philosophy (Bloomington: Indiana University Press)

French, Patrick and Roland Francois Lack, eds. 1998. *The Tel Quel Reader* (New York: Routledge)

Le Doeuff, Michele, 1990. *The Philosophical Imaginary*, Colin Gordon, trans. (Stanford: Stanford University Press)

Mathy, Jean Philippe, 1993. *Extreme--Occident: French Intellectuals and America* (Chicago: University of Chicago Press)

Matthews, Eric, 1996. *Twentieth-Century French Philosophy* (New York: Oxford University Press)

Mortley, Raoul, 1991. *French Philosophies in Conversation: Levinas, Schneider, Serres, Irigaray, Le Doeuff, Derrida* (New York: Routledge)

Merleau-Ponty

Books

Allen, Jeffner and Iris Marion Young, eds. 1989. *The Thinking Muse: Feminism and Modern French Philosophy* (Bloomington: Indiana University Press)

Mazis, Glen, 1993. *Emotion and Embodiment: Fragile Ontology* (New York: Lang)

Perez Estevez, Antonio, 1989. *El individuo y la feminidad* (Zulia: University of Zulia) (Portugese)

Articles

Bigwood, Carol, "Renaturalizing the Body (With a Little Help from Merleau-Ponty)," *Hypatia* 1991, 54-73.

Fielding, Helen, "Grounding Agency in Depth: The Implications of Merleau-Ponty's Thought for the Politics of Feminism," *Human Studies* 19, 2 (1996), 175-184.

Kozel, Susan, "The Diabolical Strategy of Mimesis: Luce Irigaray's Reading of Maurice Merleau-Ponty," *Hypatia* 11, 3 (1996), 114-129.

O'Loughlin, Marjorie, "Intelligent Bodies and Ecological Subjectivists: Merleau-Ponty's Corrective to Postmodernism's ‘Subjects’ of Education," in *Philosophy of Education,* Alven Neiman, ed. (Urbana: Philosophy of Education and Society, 1995)

Popen, Shari, "Merleau-Ponty Confronts Postmodernism: A Reply to O'Loughlin," in *Philosophy of Education,* Alven Neiman, ed. (Urbana: Philosophy of Education and Society, 1995)

Preston, Beth, "Merleau-Ponty and the Feminine Embodied Eixstence," *Man and World* 29, 2 (1996), 167-186.

Reineke, Martha J., "Lacan, Merleau-Ponty, and Irigaray: Reflections on a Specular Drama," *Auslegung* 14 (1987), 67-85.

Sullivan, Shannon, "Domination and Dialogue in Merleau-Ponty's ‘Phenomenology of Perception,’" *Hypatia* 12, 1 (1997), 1-19.

Young, Iris Marion, "Throwing Like a Girl: A Phenomenology of Feminine Body Comportments, Motility, and Spatiality," *Human Studies* 3 (1980), 137-156.

Albert Camus

Books

Allen, Jeffner and Iris Marion Young, eds. 1989. *The Thinking Muse: Feminism and Modern French Philosophy* (Bloomington: Indiana University Press)

Articles

Bartlett, Elizabeth Ann, "Beyond Either/Or: Justice and Care in the Ethics of Albert Camus," in *Explorations in Feminist Ethics* Eve Browning, ed. (Bloomington: Indiana University Press, 1992)

Kruks, Sonia, "Existentialism and Phenomenology," in *A Companion to Feminist Philosophy,* Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Jean-Paul Sartre

Books

Allen, Jeffner and Iris Marion Young, eds. 1989. *The Thinking Muse: Feminism and Modern French Philosophy* (Bloomington: Indiana University Press)

Barrett, W., 1962. *Irrational Man* (New York: Doubleday Anchor) Part III, Chapter 10.

Articles

Barnes, Hazel E., "Sartre and Sexism," *Philosophy and Literature* (1990), 340-347.

Bergoffen, Debra B., "The Look as Bad Faith," *Philosophy Today* 36, 3 (1992), 221-227.

Collins, Margery and Christine Pierce, "Holes and Slime: Sexism in Sartre's Psychoanalysis," in *Women and Philosophy: Toward a Theory of Liberation*, Carol Gould and Marx X. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Comesana, Gloria M., "Analisis de las figuras femeninas en el teatro de Sartre," *Revista de Filosofia (Venezuela)* 9 (1986), 103-133. (Spanish)

Comesana Santalices, Gloria, "'Behind Closed Doors': Analysis of the Feminine Figures in Sartrean Theater," *Revista de Filosofia (Venezuela)* 24, 2 (1996), 53-79. (Spanish)

Fullbrook, Kate and Edward Fullbrook, "Sartre's Secret Key," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Keat, R., "Masculinity in Philosophy," *Radical Philosophy* 34 (summer 1983), 15-20.

Kruks, Sonia, "Existentialism and Phenomenology," in *A Companion to Feminist Philosophy,* Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Kruks, Sonia, "Simone de Beauvoir: Teaching Sartre About Freedom," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Mui, Constance, "Sartre's Sexism Reconsidered," *Auslegung* 16, 1 (1990), 31-41.

Murphy, Julien S., "The Look in Sartre and Rich," *Hypatia* 2 (1987), 113-124.

Simone de Beauvoir

Ascher, Carole, 1981. *Simone de Beauvoir: A Life of Freedom*. (Brighton: Harvester Press)

Bergoffen, Debra B., 1997. *The Philosophy of Simone de Beauvoir: Gendered Phenomenologies, Erotic Generosities* (Albany: SUNY Press)

Bieber, Konrad, 1979. *Simone de Beauvoir* (Boston: Hall)

Easlea, B., 1981. *Science and Sexual Oppression* (London: Weidenfeld & Nicholson), Chapter 2

Keefe, Terry, 1983. *Simone de Beauvoir: A Study of Her Writings* (Totowa: Barnes and Noble)

Moi, Toril, 1994. *Simone de Beauvoir: The Making of an Intellectual Woman* (Cambridge: Blackwell)

Nordquist, Joan, 1991. *Social Theory: A Bibliographic Series, No. 23--Simone de Beauvoir: A Bibliography* (Santa Cruz: Reference and Research)

Okely, Judith, 1986. *Simone de Beauvoir* (New York: Pantheon)

Richards, Janet Radcliffe, 1980. *The Sceptical Feminist: A Philosophical Inquiry* (Boston: Routledge and K. Paul)

Sabrosky, Judith A., 1979. *From Rationality to Liberation* (Westport: Greenwood Press)

Savage Brosman, Catharine, 1991. *Simone de Beauvoir Revisited* (Boston: Twayne)

Simons, Margaret A., ed. 1995. *Feminist Interpretations of Simone de Beauvoir* (University Park: Pennsylvania State University Press)

Wenzel, Helene Vivienne, ed. 1986. *Simone de Beauvoir: Witness to a Century* (New Haven: Yale University Press)

Articles

Alexander, Anna, "The Eclipse of Gender: Simone de Beauvoir and the ‘Differance’ of Translation," *Philosophy Today* 41, 1-4 (1997), 112-120.

Allen, Jeffner, "A Response to a Letter from Peg Simons, December 1993," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Arp, Kristana, "Beauvoir's Concept of Bodily Alienation," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Barber, Michael D., "Autobiography: Precarious Totality," in *Alfred Schutz's ‘Sociological Aspect of Literature* Lester Embree, ed. (Dordrecht: Kluwer)

Bergoffen, Debra B., "From Husserl to de Beauvoir: Gendering the Perceived Subject," *Metaphilosophy* 27, 1-2 91996), 53-62.

Bergoffen, Debra B., "The Look as Bad Faith," *Philosophy Today* 36, 3 (1992), 221-227.

Bergoffen, Debra B., "Out From Under: Beauvoir's Philosophy of the Erotic," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Bordo, Susan, "The Feminist as Other," *Metaphilosophy* 27, 1-2 (1996), 10-27.

Butler, Judith, "Gendering the Body: Beauvoir's Philosophical Contribution," in *Beyond Domination: New Perspectives on Women and Philosophy,* Carol Gould, ed. (Totowa NJ: Rowman & Allanheld, 1984)

Coddetta, Carolina, "The Problem of Power in the Feminist Theory," *Fronesis* 2, 2 (1995), 59-95. (Spanish)

Comesana Santalices, Gloria M., "'El segundo sexo,' vigentia y proyeccion," *Revista de Filosofia (Venezuela)* (1989), 45-72. (Spanish)

Dallery, Arleen B., "Sexual Embodiment: Beauvoir and French Feminist ('Ecriture Feminine')," *Hypatia* WSIF 3 (1985), 197-202.

Dijkstra, Sandra, "Simone de Beauvoir and Betty Friedan," *Feminist Studies* 6 (1980), 290-303.

Farrell Smith, Janet, "Possessive Power," *Hypatia* 1 (1986), 103-120.

Felstiner, Mary Lowenthal, "Seeing 'The Second Sex' Through the Second Wave," *feminist Studies* 6 (1980), 247-276.

Ferguson, Ann, "Lesbian Identity: Beauvoir and History," *Hypatia* WSIF 3 (1985), 203-208.

Fraser, Miriam, "Feminism, Foucault, and Deleuze," *Theory, Culture, and Society* 14, 2 (1997) 23-37.

Fullbrook, Kate and Edward Fullbrook, "Sartre's Secret Key," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Godard, Linda, "Pour une nouvelle lecture de la question de la femme: essai a partir de la pensee de Jacques Derrida," *Philosophiques* 12 (1985), 147-165. (French)

Hatcher, Donald L., "Existential Ethics and Why It's Immoral to be a Housewife," *Journal of Values Inquiry* 23 (1989), 59-68.

Hollywood, Amy M., "Beauvoir, Irigaray, and the Mystical," *Hypatia* 9, 4 (1994), 158-185.

Holveck, Eleanore, "Can a Woman Be a Philosopher? Reflections of a Beauvoirian Housemaid," in

Feminist Interpretations of Simone de Beauvoir Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Idt, Genevieve, "Simone de Beauvoir's Adieux: A Funeral Rite and a Literary Challenge," in *Sartre Alive* Ronald Aronson, ed. (Detroit: Wayne State University Press, 1991)

Klaw, Barbara, "Sexuality in Beauvoir's ‘Les Mandarins,’" in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Kruks, Sonia, "Existentialism and Phenomenology," in *A Companion to Feminist Philosophy,* Alison M. Jaggar, ed. (Cambridge: Blackwell, 1998)

Kruks, Sonia, "Simone de Beauvoir: Teaching Sartre About Freedom," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Langer, Monika, "A Philosophical retrieval of Simone de Beauvoir's ‘Pour Une Morale de l'amiguite,’" *Philosophy Today* 38, 2 (1994), 181-190.

Lazaro, Reyes, "Feminism and Motherhood: O'Brien vs. Beauvoir," *Hypatia* 1 (1986), 87-102.

Le Doeuff, Michele, "Operative Philosophy: Simone de Beauvoir and Existentialism," *Ideology and Consciousness* 6 (autumn 1979), 47-58.

Le Doeuff, Michele, "Simone de Beauvoir: Falling into (Ambiguous) Line," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Levaux, Michele, "Simone de Beauvoir, une feministe exceptionnelle," *Etudes* 360 (1984), 493-498. (French)

Lundgren Gothlin, Eva, "Ethics, Feminism, and Postmodernism: Seyla Benhabib and Simone de Beauvoir," in *The Postmodern Critique of the Project of the Enlightenment* Sven Eric Liedman, ed. (Amsterdam: Rodopi, 1997)

Lundgren Gothlin, Eva, "Simone de Beauvoir and Ethics," *History of European Ideas* 19, 4-6 (1994), 899-903.

Mahowald, Mary Briody, "To be or Not to Be a Woman: Anorexia Nervosa, Normative Gender Roles, and Feminism," *Journal of Medicine and Philosophy* 17, 2 (1992), 233-251.

Malion, Joseph, "Existentialism, Feminism, and Simone de Beauvoir," *History of European Ideas* 17, 5

(1993), 651-658.

McCall, D. Kaufman, "Simone de Beauvoir, *The Second Sex* and Jean-Paul Sartre," *Signs* 5, 2 (winter 1979), 209-223.

Morgan, Kathryn Pauly, "Romantic Love, Altrusim, and Self-Repsect," *Hypatia* 1 (1986), 117-148.

Murphy, Julien, "Beauvoir and the Algerian War: Toward a Postcolonial Ethics," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Nye, Andrea, "Preparing the Way for a Feminist Praxis," *Hypatia* 1 (1986), 101-116.

Pilardi, Jo Ann, "The Changing Critical Fortunes of ‘The Second Sex,’" *History and Theory* 32 (1993), 51-73.

Schutte, Ofelia, "A Critique of Normative Heterosexuality: Identity, Embodiment, and Sexual Difference in Beauvoir and Irigaray," *Hypatia* 12, 1 (1997), 40-62.

Seigfried, Charlene Haddock, "’Second Sex’: Second Thoughts," *Hypatia* WSIF 3 (1985), 219-229.

Simons, Margaret A., "Beauvoir and the Roots of Radical feminism," in *Reinterpreting the Politics: Continental Philosophy and Political Theory* Lenore Langsdorf, ed. (Albany: SUNY Press, 1998)

Simons, Margaret A. and Jessica Benjamin, "Simone de Beauvoir: An Interview," *feminist Studies* 5 (1979), 330-345.

Simons, Margaret A., "The Second Sex: From Marxism to Radical Feminism," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Simons, Margaret A., "Sexism and the Philosophical Canon: On Reading Beauvoir's ‘The Second Sex,’" *Journal of the History of Ideas* 51, 3 (1990), 487-504.

Simons, Margaret A., "Two Interviews With Simone de Beauvoir," *Hypatia* 3 (1989), 11-27.

Singer, Linda, "Interpretation and Retrieval: Rereading Beauvoir," *Hypatia* WSIF 3 (1985), 231-238.

Slattery, Patrick and Marla Morris, "Simone de Beauvoir's Ethics and Postmodern Ambiguity: The Assertion of Freedom in the Face of the Absurd," *Educational Theory* 49, 1 (1999), 21-36.

Spelman, Elizabeth, "Women as Body: Ancient and Contemporary Views," *Feminist Studies,* 8, 1 (1982), 109-131.

Tirrell, Lynne, "Definition and Power: Toward Authority Without Privilege," *Hypatia* 8, 4 (1993), 1-34.

Vintges, Karen," The Second Sex and Philosophy," in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Ward, Julie K., "Beauvoir's Two Sense of Body in 'The Second Sex,'" in *Feminist Interpretations of Simone de Beauvoir* Margaret A. Simons, ed. (University Park: Pennsylvania State University Press, 1995)

Young, Iris Marion, "Throwing Like a Girl: A Phenomenology of Feminine Body Comportments, Motility, and Spatiality," *Human Studies* 3 (1980), 137-156.

Zephyr, Jacques J., "Simone de Beauvoir et la femme," *Rev. Univ. Ottowa* 54 (1984), 37-53. (French)

Zerilli, Linda M. G., "Doing Without Knowing: Feminism's Politics of the Ordinary," *Political Theory* 26, 4 (1998), 435-458.

Other Twentieth Century Philosophy

Books

Benstock, Shari, ed. 1987. *Feminist Issues in Literary Scholarship* (Bloomington: Indiana University Press)

Articles

Cocks, Joan, "Cultural Theory Looks at Identity and Contradiction," *Quest* (1990), 38-60.

Hall, Diana Long, "Biology, Sex Hormones, and Sexism in the 1920s." in *Women and Philosophy: Toward a Theory of Liberation,* Carol C. Gould and Marx W. Wartofsky, eds. (New York: G. P. Putnam's Sons, 1976)

Melander, Ellinor, "Toward the Sexual and Economic Emancipation of Women: The Philosophy of Grete Meisel-Hess," *History of European Ideas* 14, 5 (1992), 695-713.

Pakszys, Elzbieta, "Women, Women's Issues, and Feminism in Polish Philosophy," in *Philosophy in Post-Communist Europe*, Dane R. Gordon, ed. (Amsterdam: Rodopi, 1998) (Lvar-Warsaw School--Analytic Philosophy, of the First Half of the 20th Century)

Sarvasy, Wendy, "Social Citizenship From a Feminist Perspective," *Hypatia* 12, 4 (1997), 54-73 (Early 20th Century Feminism)

Wosk, Julie, "The 'Electric Eve': Galvanizing Women in Nineteenth and Twentieth Century Art and Technology," *Research in Philosophy and Technology* 13 (1993), 43-56.

Abigail Gosselin
University of New Hampshire

Return to Feminist History of Philosophy

First published: November 3, 2000
Content last modified: November 3, 2000

Many-Valued Logic

Many-valued logics are non-classical logics. They are similar to classical logic because they accept the principle of truth-functionality, namely, that the truth of a compound sentence is determined by the truth values of its component sentences (and so remains unaffected when one of its component sentences is replaced by another sentence with the same truth value). But they differ from classical logic by the fundamental fact that they do not restrict the number of truth values to only two: they allow for a larger set *W* of truth degrees.

Just as the notion of ‘possible worlds’ in the semantics of modal logic can be reinterpreted (e.g., as ‘moments of time’ in the semantics of tense logic or as ‘states’ in the semantics of dynamic logic), there does not exist a standard interpretation of the truth degrees. How they are to be understood depends on the actual field of application. It is general usage, however, to assume that there are two particular truth degrees, usually denoted by "0" and "1", respectively, which act like the traditional truth values "falsum" and "verum".

The formalized languages for systems of *many-valued logic* (MVL) follow the two standard patterns for propositional and predicate logic, respectively:

- there are propositional variables together with connectives and (possibly also) truth degree constants in the case of propositional languages,
- there are object variables together with predicate symbols, possibly also object constants and function symbols, as well as quantifiers, connectives, and (possibly also) truth degree constants in the case of first-order languages.

As usual in logic, these languages are the basis for semantically as well as syntactically founded systems of logic.

Semantics

There are two kinds of semantics for systems of many-valued logic.

We discuss these in turn.

Standard Logical Matrices

The most suitable way of defining a system **S** of many-valued logic is to fix the characteristic logical matrix for its language, i.e. to fix:

* the set of truth degrees,
* the truth degree functions which interpret the propositional connectives,
* the meaning of the truth degree constants,
* the semantical interpretation of the quantifiers,

and additionally,

* the *designated truth degrees*, which form a subset of the set of truth degrees and act as substitutes for the traditional truth value "verum".

A well-formed formula *A* of a propositional language counts as *valid* under some valuation α (which maps the set of propositional variables into the set of truth degrees) iff it has a designated truth degree under α. And *A* is *logically valid* or a *tautology* iff it is valid under all valuations.

In the case of a first-order language, such a well-formed formula *A* counts as *valid* under an interpretation α of the language iff it has a designated truth degree under this interpretation and all assignments of objects from the universe of discourse of this interpretation to the object variables. *A* counts as *logically valid* iff it is valid under all interpretations.

Like in classical logic, such an interpretation has to provide

* a (non empty) universe of discourse,

- the meaning of the object constants of the language,
- the meaning of the predicate letters and the function symbols of the language.

A *model* of some set Σ of well-formed formulas is a valuation α or an interpretation α such that all $A \in \Sigma$ are valid under α . That Σ *entails* *A* means that each model of Σ is also a model of *A*.

Algebraic Semantics

There is a second type of semantics for systems **S** of many-valued logic which is based on a whole characteristic class **K** of (similar) algebraic structures. Each such algebraic structure has to provide all the data which have to be provided by a characteristic logical matrix for the language of **S**.

The notion of validity of a formula *A* with respect to an algebraic structure from **K** is defined as if this structure would form a logical matrix. And *logical validity* here means validity for all structures from the class **K**.

The type of algebraic structures which may form such a characteristic class **K** for some system **S** of MVL is usually determined by the (syntactical or semantical) Lindenbaum algebra of **S**, and often plays also a crucial role within an algebraic completeness proof. The algebraic structures in **K** have a similar role for **S** as the Boolean algebras do for classical logic.

For particular systems of MVL one has e.g. the following characteristic classes of algebraic structures:

- for infinite valued *Lukasiewicz logic* the class of MV-algebras,
- for infinite valued *Gödel logic* the class of all Heyting algebras which satisfy prelinearity $(x \rightarrow y) \cup (y \rightarrow x) = 1$,
- for Hajek's *basic t-norm logic* the class of all divisible residuated lattices which satisfy prelinearity.

From a philosophical point of view, it would be preferable to have a semantic foundation for a system of MVL which uses a characteristic logical matrix. However, from a formal point of view, both approaches are equally important, and the algebraic semantics turns out to be the more general approach.

[Return to Table of Contents]

Proof Theory

The main types of logical calculi are all available for systems of MVL:

- Hilbert type calculi

- Gentzen type sequent calculi
- Tableau calculi

However, some of the above are available only for finitely valued systems.

Hilbert type calculi. These calculi are formed in the same way as the corresponding calculi for classical logic: some set of *axioms* is used together with a set of *inference rules*. The notion of derivation is the usual one.

Gentzen type sequent calculi. In addition to the usual types of sequent calculi, researchers have also recently started to discuss ‘hypersequent’ calculi for systems of MVL. Hypersequents are finite sequences of ordinary sequents.

For finitely valued systems, particularly *m*-valued ones, there are also sequent calculi which work with *generalized sequents*. In the *m*-valued case, these are sequences of length *m*.

Tableau calculi. The tree structure of the tableaux remains the same in these calculi as in the tableau calculi for classical logic. The labels of the nodes become more general objects, namely, *signed formulas*. A signed formula is a pair, consisting of a *sign* and a well-formed formula. A sign is either a truth degree, or a set of truth degrees.

Tableau calculi with signed formulas are usually restricted to finite-valued systems of MVL, so that they can be dealt with in an effective way.

[Return to Table of Contents]

Systems of Many-Valued Logic

The main systems of MVL often come as families which comprise uniformly defined finite-valued as well as infinite-valued systems. Here is a list:

- Lukasiewicz logics
- Gödel logics
- t-Norm related systems
- 3-valued systems
- Dunn/Belnap's 4-valued system
- Product systems

Lukasiewicz logics

The systems L_m and L_∞ are defined by the logical matrix which has either some finite set

$$W_m = \{k/m-1 \mid 0 \leq k \leq m-1\}$$

of rationals within the real unit interval, or the whole unit interval

$$W_\infty = [0,1] = \{x \in R \mid 0 \leq x \leq 1\}$$

as the truth degree set. The degree 1 is the only designated truth degree.

The main connectives of these systems are a strong and a weak conjunction, & and $\wedge$, respectively, given by the truth degree functions

$$u \mathbin{\&} v = \max\{0, u + v - 1\},$$

$$u \wedge v = \min\{u, v\},$$

a negation connective $\neg$ determined by

$$\neg u = 1 - u,$$

and an implication connective $\rightarrow$ with truth degree function

$$u \rightarrow v = \min\{1, 1 - u + v\}.$$

Often, two disjunction connectives are also used. These are defined in terms of & and $\wedge$, respectively, via the usual de Morgan laws using $\neg$. For the first-order Lukasiewicz systems one adds two quantifiers $\forall$, $\exists$ in such a way that the truth degree of $\forall x H(x)$ is the *infimum* of all the relevant truth degrees of $H(x)$, and that the truth degree of $\exists x H(x)$ is the *supremum* of all the relevant truth degrees of $H(x)$.

Gödel logics

The systems G_m and G_∞ are defined by the logical matrix which has either some finite set

$$W_m = \{k/m-1 \mid 0 \leq k \leq m-1\}$$

of rationals within the real unit interval, or the whole unit interval

$$W_\infty = [0,1] = \{x \in R \mid 0 \leq x \leq 1\}$$

as the truth degree set. The degree 1 is the only designated truth degree.

The main connectives of these systems are a conjunction $\wedge$ and a disjunction $\vee$ determined by the truth degree functions

$$u \wedge v = \min\{u, v\},$$

$$u \vee v = \max\{u, v\},$$

an implication connective $\rightarrow$ with truth degree function

$$u \rightarrow v = \begin{cases} 1, & \text{if } u \leq v \\ v, & \text{if } u > v \end{cases}$$

and a negation connective $\sim$ with truth degree function

$$\sim u = \begin{cases} 1, & \text{if } u = 0 \\ 0, & \text{if } u \neq 0 \end{cases}$$

For the first-order Gödel systems one adds two quantifiers $\forall$, $\exists$ in such a way that the truth degree of $\forall x H(x)$ is the *infimum* of all the relevant truth degrees of $H(x)$, and that the truth degree of $\exists x H(x)$ is the *supremum* of all the relevant truth degrees of $H(x)$.

t-Norm related systems

For infinite valued systems with truth degree set

$$W_\infty = [0,1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$$

the influence of fuzzy set theory quite recently initiated the study of a whole class of such systems of MVL.

These systems are basically determined by a (possibly non-idempotent) strong conjunction connective $\&_T$ which has as corresponding truth degree function a *t-norm* T, i.e. a binary operation T in the unit interval which is associative, commutative, non-decreasing, and has the degree 1 as a neutral element:

- $T(u,T(v,w)) = T(T(u,v),w)$,
- $T(u,v) = T(v,u)$,

- $u \leq v \Rightarrow \mathrm{T}(u,w) \leq \mathrm{T}(v,w)$,
- $\mathrm{T}(u,1) = u$.

For all those t-norms which have the *sup-preservation property*

$$\mathrm{T}(u, \sup_i v_i) = \sup_i \mathrm{T}(u,v_i),$$

there is a standard way to introduce a related implication connective $\rightarrow_{\mathrm{T}}$ with the truth degree function

$$u \rightarrow_{\mathrm{T}} v = \sup \{z \mid \mathrm{T}(u,z) \leq v\}.$$

This is connected with the t-norm T by the crucial *adjointness condition*

$$\mathrm{T}(u,v) \leq w \Leftrightarrow u \leq (v \rightarrow_{\mathrm{T}} w),$$

which determines $\rightarrow_{\mathrm{T}}$ uniquely for each T with sup-preservation property.

The language is further enriched with a negation connective, $-_{\mathrm{T}}$, determined by the truth degree function

$$-_{\mathrm{T}} u = u \rightarrow_{\mathrm{T}} 0.$$

This forces the language to have also a truth degree constant $\underline{0}$ to denote the truth degree 0 because then $-_{\mathrm{T}}$ becomes a definable connective.

Usually one adds as two further connectives a (weak) conjunction $\wedge$ and a disjunction $\vee$ with truth degree functions.

$$u \wedge v = \min \{u, v\},$$

$$u \vee v = \max \{u, v\}.$$

Particular cases of such t-norm related systems are the infinite valued Lukasiewicz and Gödel systems L_∞, G_∞, and also the *product logic* which has the usual arithmetic product as its basic t-norm.

The class of all t-norms, even of those which have the sup-preservation property, is very large. Actually one is able to axiomatize t-norm based systems for some particular classes of t-norms. And Hajek (1998) has given an axiomatization of the logic which has as its algebraic semantics the class of all t-norm based structures whose t-norm is a continuous function.

The axiomatization of further t-norm based systems, as well as the question for t-norm based quantifiers,

are recent research problems.

3-valued systems

3-valued systems seem to be particularly simple cases which offer intuitive interpretations of the truth degrees; these systems include only one additional degree besides the classical truth values.

The mathematician and logician Kleene used a third truth degree for "undefined" in the context of partial recursive functions. His connectives were the negation, the weak conjunction, and the weak disjunction of the 3-valued Lukasiewicz system together with a conjunction $\wedge_+$ and an implication $\rightarrow_+$ determined by truth degree functions with the following function tables:

$\wedge_+$	0	½	1
0	0	½	0
½	½	½	½
1	0	½	1

$\rightarrow_+$	0	½	1
0	1	1	1
½	½	½	½
1	0	½	1

Here ½ is the third truth degree "undefined". In this Kleene system, the degree 1 is the only designated truth degree.

Blau (1978) used a different system as an inherent logic of natural language. In Blau's system, both degrees 1 and ½ are designated. Other interpretations of the third truth degree ½, for example as "senseless", "undetermined", or "paradoxical", motivated the study of other 3-valued systems.

Dunn/Belnap's 4-valued system

This particularly interesting system of MVL was the result of research on relevance logic, but it also has significance for computer science applications. Its truth degree set may be taken as

$$W^* = \{\emptyset, \{\bot\}, \{\top\}, \{\bot, \top\}\},$$

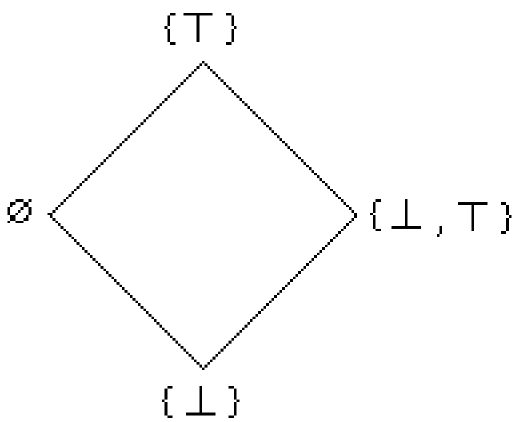
and the truth degrees interpreted as indicating (e.g. with respect to a database query for some particular

state of affairs) that there is

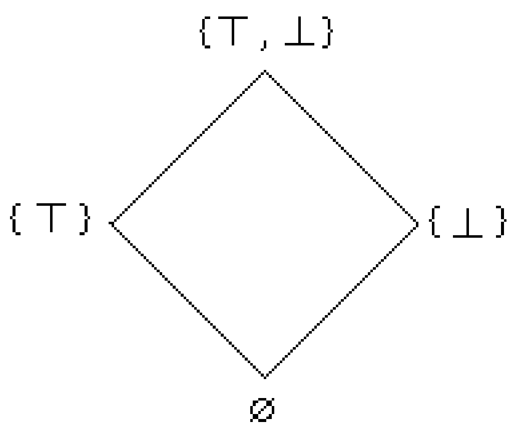
- no information concerning this state of affairs,
- information saying that the state of affairs fails,
- information saying that the state of affairs obtains,
- conflicting information saying that the state of affairs obtains as well as fails.

This set of truth degrees has two natural (lattice) orderings:

- a *truth ordering* which has $\{\top\}$ on top of the incomparable degrees $\emptyset$, $\{\bot , \top\}$, and has $\{\bot\}$ at the bottom; i.e.,



- an *information* (or: *knowledge*) *ordering* which has $\{\bot , \top\}$ on top of the incomparable degrees $\{\bot\}$, $\{\top\}$, and has $\emptyset$ at the bottom; i.e.,



Given the inf and the sup under the truth ordering, there are truth degree functions for a conjunction and a disjunction connective. A negation is, in a natural way, determined by a truth degree function which exchanges the degrees $\{\bot\}$ and $\{\top\}$, and which leaves the degrees $\{\bot , \top\}$ and $\emptyset$ fixed.

Actually, there is no standard candidate for a implication connective, and the choice of the designated truth degrees depends on the intended applications:

- for computer science applications it is natural to have $\{\top\}$ as the only designated degree,

- for applications to relevance logic the choice of {⊤}, {⊥ , ⊤} as designated degrees proved to be adequate.

The choice of suitable entailment relations is still an open research topic.

Product systems

The general problem of finding an intuitive understanding of the truth degrees occasionally has a nice solution: one can consider them as comprising different aspects of the evaluation of sentences. In such a case of, say, *k* different aspects the truth degrees may be chosen as *k*-tuples of values which evaluate the single aspects. (And these, e.g., may be standard truth values.)

The truth degree functions over such *k*-tuples additionally can be defined "componentwise" from truth degree (or: truth value) functions for the values of the single components. In this manner, *k* logical systems may be combined into one many-valued *product system*.

In this way, the truth degrees of Dunn/Belnap's 4-valued system can be considered as evaluating two aspects of a state of affairs (SOA) related to a database:

1. whether there is positive information about the truth of this SOA or not, and
2. whether there is positive information about the falsity of this SOA or not.

Both aspects can use standard truth values for this evaluation.

In this case, the conjunction, disjunction, and negation of Dunn/Belnap's 4-valued system are componentwise definable by conjunction, disjunction, or negation, respectively, of classical logic, i.e. this 4-valued system is a product of two copies of classical two-valued logic.

[Return to Table of Contents]

Applications of Many-Valued Logic

Many-valued logic was motivated in part by philosophical goals which were never achieved, and in part by formal considerations concerning functional completeness. In the earlier years of development, this caused some doubts about the usefulness of MVL. In the meantime, however, interesting applications were found in diverse fields. Some of these shall now be mentioned.

Applications to Linguistics. A challenging problem is the treatment of presuppositions in linguistics, i.e. of assumptions that are only implicit in a given sentence. So, for example, the sentence "The present king of Canada was born in Vienna" has the *existential presupposition* that there is a present king of Canada.

It is not a simple task to understand the propositional treatment of such sentences, e.g. to give criteria for forming their negation, or understanding the truth conditions of implications.

One type of solution for these problems refers to the use of many truth degrees, e.g. to *product systems* with ordered pairs as truth degrees: meaning that their components evaluate in parallel whether the presupposition is met, and whether the sentence is true or false. But 3-valued approaches have also been discussed.

Applications to Logic. A first type of application of systems of MVL to logic itself is to use them to gain a better understanding of other systems of logic. In this way the Gödel systems arose out of an approach to test whether intuitionistic logic may be understood as a finitely valued logic. The introduction of systems of MVL by Lukasiewicz (1920) was initially guided by the (finally unsuccessful) idea of understanding the notion of possibility, i.e. modal logic, in a 3-valued way.

A second type of application to logic is the merging of different types of logical systems, e.g. the formulation of systems with graded modalities. Melvin Fitting (1991/92) considers systems that define such modalities by merging modal and many-valued logic, with intended applications to problems of Artificial Intelligence.

A third type of application to logic is the modeling of partial predicates and truth value gaps. However, this is possible only in so far as these truth value gaps behave "truth functionally", i.e. in so far as the behavior of the truth value gaps in compound sentences can be described by suitable truth functions. (This is not always the case, e.g. it is not the case in formulations which use *supervaluations*.)

Applications to Philosophical Problems . How to understand the meaning of "truth" is an old philosophical problem. A logical approach toward this problem consists in enriching a formalized language L with a truth predicate T, to be applied to sentences of L -- or, even better, to be applied to sentences of the extension L_T of L with the predicate T.

Based upon this idea, a reasonable theory of such languages which contain truth predicates was developed in the mid-1930s by A. Tarski. One of the results was that such a language L_T, which contains its own truth predicate T and has a certain richness in expressive power, is necessarily inconsistent.

Another approach toward such languages L_T which contain their own truth predicate T was offered by S. Kripke (1975) and is essentially based upon the idea of considering T as a partial predicate, i.e. as a predicate which has "truth value gaps". In a case Kripke (1975) considers, these truth value gaps behave "truth functionally" and so can be treated like a third truth degree. Their propagation in compound sentences then becomes describable by suitable truth degree functions of three-valued systems. In Kripke's (1975) approach this reference was to three-valued systems which S. C. Kleene (1938) had considered in the (mathematical) context of partial functions and predicates in recursion theory.

A second application of MVL inside philosophy is to the old paradoxes like the *Sorites* (heap) or the *falakros* (bald man). (See the entry Sorites paradox.) In the case of the Sorites, the paradox is as follows:

> (i) One grain of sand is not a heap of sand. And (ii) adding one grain of sand to something which is not a heap does not turn it into a heap. Hence (iii) a single grain of sand can never turn into a heap of sand, no matter how many grains of sand are added to it.

Thus the true premise (i) gives a false conclusion (iii) via a sequence of inferences using (ii). A rather natural solution inside an extension of MVL with a graded notion of inference, often called *fuzzy logic*, is to take the notion of heap as a *vague* one, i.e. as a notion which may hold true of given objects only to some (truth) degree. Additionally it is suitable to consider premise (ii) as only partially true, however to a degree which is quite near to the maximal degree 1. Then each single inference step is of the form:

- (a): *k grains of sand do not make a heap.*
- (ii): *Adding one grain of sand to k grains does not make (k+1) grains into a heap.*
- Hence (b): *(k+1) grains of sand do not make a heap.*

However, this inference has to involve truth degrees for the premises (a) and (ii), and has to provide a truth degree for the conclusion (b). The crucial idea for the modeling of this type of reasoning inside MVL is to make sure that the truth degree for (b) is smaller than the truth degree for (a) in case the truth degree for (ii) is smaller than the maximal one. In effect, then, the sentence *n grains of sand do not make a heap* tends toward being false for an increasing number *n* of grains.

Applications to Hardware Design. Classical propositional logic is used as a technical tool for the analysis and synthesis of some types of electrical circuits built up from "switches" with two stable states, i.e. voltage levels. A rather straightforward generalization allows the use of an *m*-valued logic to discuss circuits built from similar "switches" with *m* stable states. This whole field of application of many-valued logic is called many-valued (or even: fuzzy) switching. A good introduction is Epstein (1993).

Applications to Artificial Intelligence. AI is actually the most promising field of applications, which offers a series of different areas in which systems of MVL have been used.

A first area of application concerns vague notions and commonsense reasoning, e.g. in expert systems. Both topics are modeled via fuzzy sets and fuzzy logic, and these refer to suitable systems of MVL. Also,

in databases and in knowledge-based systems one likes to store vague information.

A second area of application is strongly tied with this first one: the automatization of data and knowledge mining. Here clustering methods come into consideration; these refer via unsharp clusters to fuzzy sets and MVL. In this context one is also interested in automated theorem proving techniques for systems of MVL, as well as in methods of logic programming for systems of MVL.

Applications to Mathematics. There are three main topics inside mathematics which are related to many-valued logic. The first one is the mathematical theory of fuzzy sets, and the mathematical analysis of "fuzzy", or approximate reasoning. In both cases one refers to systems of MVL. The second topic has been approaches toward consistency proofs for set theory using a suitable system of MVL. And there is an -- often only implicit -- reference to the basic ideas of MVL in independence proofs (e.g. for systems of axioms) which often refer to logical matrices with more than two truth degrees. However, here MVL is more a purely technical tool because in these independence proofs one is not interested in an intuitive understanding of the truth degrees at all.

[Return to Table of Contents]

History of Many-Valued Logic

Many-valued logic as a separate subject was created by the Polish logician and philosopher Lukasiewicz (1920), and developed first in Poland. His first intention was to use a third, additional truth value for "possible", and to model in this way the modalities "it is necessary that" and "it is possible that". This intended application to modal logic did not materialize. The outcome of these investigations are, however, the Lukasiewicz systems, and a series of theoretical results concerning these systems.

Essentially parallel to the Lukasiewicz approach, the American mathematician Post (1921) introduced the basic idea of additional truth degrees, and applied it to problems of the representability of functions.

Later on, Gödel (1932) tried to understand intuitionistic logic in terms of many truth degrees. The outcome was the family of Gödel systems, and a result, namely, that intuitionistic logic does not have a characteristic logical matrix with only finitely many truth degrees. A few years later, Jaskowski (1936) constructed an infinite valued characteristic matrix for intuitionistic logic. It seems, however, that the truth degrees of this matrix do not have a nice and simple intuitive interpretation.

A philosophical application of 3-valued logic to the discussion of paradoxes was proposed by the Russian logician Bochvar (1938), and a mathematical one to partial function and relations by the American logician Kleene (1938). Much later Kleene's connectives also became philosophically interesting as a technical tool to determine fixed points in the revision theory of truth initiated by Kripke (1975).

The 1950s saw (i) an analytical characterization of the class of truth degree functions definable in the infinite valued propositional Lukasiewicz system by McNaughton (1951), (ii) a completeness proof for the same system by Chang (1958, 1959) introducing the notion of MV-algebra and a more traditional one by Rose/Rosser (1958), as well as (iii) a completeness proof for the infinite valued propositional Gödel system by Dummett (1959). The 1950s also saw an approach of Skolem (1957) toward proving the consistency of set theory in the realm of infinite valued logic.

In the 1960s, Scarpellini (1962) made clear that the first-order infinite valued Lukasiewicz system is not (recursively) axiomatizable. Hay (1963) as well as Belluce/Chang (1963) proved that the addition of one infinitary inference rule leads to an axiomatization of L_∞. And Horn (1969) presented a completeness proof for first-order infinite valued Gödel logic. Besides these developments inside pure many-valued logic, Zadeh (1965) started an (application oriented) approach toward the formalization of vague notions by generalized set theoretic means, which soon was related by Goguen (1968/69) to philosophical applications, and which later on inspired also a lot of theoretical considerations inside MVL.

The 1970s mark a period of restricted activity in pure many-valued logic. There was, however, a lot of work in the closely related area of (computer science) applications of vague notions formalized as fuzzy sets, initiated e.g. by Zadeh (1975, 1979). And there was an important extension of MVL by a graded notion of inference and entailment in Pavelka (1979).

In the 1980s, fuzzy sets and their applications remained a hot topic that called for theoretical foundations by methods of many-valued logic. In addition, there were the first complexity results e.g. concerning the set of logically valid formulas in first-order infinite valued Lukasiewicz logic, by Ragaz (1983). Mundici (1986) started a deeper study of MV-algebras.

These trends have continued since the 1980s. Research has included applications of MVL to fuzzy set theory and their applications, detailed investigations of algebraic structures related to systems of MVL, the study of graded notions of entailment, and investigations into complexity issues for different problems in systems of MVL. This research was complemented by interesting work on proof theory, on automated theorem proving, by different applications in artificial intelligence matters, and by a detailed study of infinite valued systems based on t-norms.

[Return to Table of Contents]

Bibliography

Monographs and Survey Papers

- Ackermann, R. (1967): *An Introduction to Many-Valued Logics.* Routledge and Kegan Paul, London.
- Bolc, L. and Borowik, P. (1992): *Many-Valued Logics,* 1. Theoretical Foundations. Springer,

Berlin.
- Cignoli, R., d'Ottaviano, I. and Mundici, D. (2000): *Algebraic Foundations of Many-Valued Reasoning.* Kluwer Acad. Publ., Dordrecht.
- Epstein G. (1993): *Multiple-Valued Logic Design.* Institute of Physics Publishing, Bristol.
- Gottwald, S. (1999): Many-valued logic and fuzzy set theory, in: U. Höhle, S.E. Rodabaugh (eds.) *Mathematics of Fuzzy Sets.* Logic, Topology, and Measure Theory. The Handbooks of Fuzzy Sets Series, Kluwer Acad. Publ., Boston 1999, 5-89.
- Gottwald, S. (2001): *A Treatise on Many-Valued Logics.* Studies in Logic and Computation, vol. 9, Research Studies Press Ltd., Baldock.
- Hähnle, R. (1993): *Automated Deduction in Multiple-Valued Logics.* Clarendon Press, Oxford.
- Hähnle, R. (1999): Tableaux for many-valued logics, in: M. d'Agostino et al. (eds.) *Handbook of Tableau Methods.* Kluwer Acad. Publ., Dordrecht, 529-580.
- Hähnle, R. (2001): Advanced many-valued logics, in: D. Gabbay, F. Guenthner (eds.), *Handbook of Philosophical Logic.* 2nd ed., vol. 2, Kluwer Acad. Publ., 297-395.
- Hajek, P. (1998): *Metamathematics of Fuzzy Logic*. Kluwer Acad. Publ., Dordrecht.
- Karpenko, A.S. (1997): *Mnogoznacnye Logiki.* Logika i Kompjuter, vol. 4, Nauka, Moscow.
- Malinowski, G. (1993): *Many-Valued Logics.* Clarendon Press, Oxford.
- Novak, V., Perfilieva, I. and Močkoř, J. (1999): *Mathematical Principles of Fuzzy Logic.* Kluwer Acad. Publ., Boston.
- Panti, G. (1998): Multi-valued logics, in: D. Gabbay, P. Smets (eds.) *Handbook of Defeasible Reasoning and Uncertainty Management Systems.* vol. 1: P. Smets (ed.) *Quantified Representation of Uncertainty and Imprecision.* Kluwer Acad. Publ., Dordrecht, 25-74.
- Rescher, N (1969): *Many-Valued Logic.* McGraw Hill, New York.
- Rine, D.C. (ed.) (1977): *Computer Science and Multiple Valued Logic.* North-Holland Publ. Comp., Amsterdam [2nd rev. ed. 1984].
- Rosser, J.B. and Turquette, A.R. (1952): *Many-Valued Logics.* North-Holland Publ. Comp., Amsterdam.
- Urquhart, A. (2001): Basic many-valued logic, in: D. Gabbay, F. Guenthner (eds.), *Handbook of Philosophical Logic*, 2nd ed., vol. 2, Kluwer Acad. Publ., Dordrecht, 249-295.
- Wojcicki, R. and Malinowski, G. (eds.) (1977): *Selected Papers on Lukasiewicz Sentential Calculi.* Ossolineum, Wroclaw.
- Wolf, R.G. (1977): A survey of many-valued logic (1966-1974), in: J.M. Dunn, G. Epstein (eds.), *Modern Uses of Multiple-Valued Logic.* Reidel, Dordrecht, 167-323.
- Zinovev, A.A. (1963): *Philosophical Problems of Many-Valued Logic.* Reidel, Dordrecht.

Other Works Cited

- Belluce, L.P. and Chang, C.C. (1963): A weak completeness theorem for infinite valued first-order logic, *Journal Symbolic Logic* 28, 43-50.
- Belnap, N.D. (1977): How a computer should think, in: G. Ryle (ed.) *Contemporary Aspects of Philosophy*. Oriel Press, Stockfield, 30-56.
- Belnap, N.D. (1977): A useful four-valued logic, in: J.M. Dunn, G. Epstein (eds.), *Modern Uses of Multiple-Valued Logic*. Reidel, Dordrecht, 8-37.

- Blau, U. (1978): *Die dreiwertige Logik der Sprache: ihre Syntax, Semantik und Anwendung in der Sprachanalyse*. de Gruyter, Berlin.
- Bochvar, D.A. (1938): Ob odnom trechznacnom iscislenii i ego primenenii k analizu paradoksov klassiceskogo rassirennogo funkcional'nogo iscislenija, *Matematiceskij Sbornik* 4 (46), 287-308. [English translation: Bochvar, D.A., On a three-valued logical calculus and its application to the analysis of the paradoxes of the classical extended functional calculus, *History and Philosophy of Logic* 2, 87-112.]
- Chang, C.C. (1958): Algebraic analysis of many valued logics, *Transactions American Mathematical Society* 88, 476-490.
- Chang, C.C. (1959): A new proof of the completeness of the Lukasiewicz axioms, *Transactions American Mathematical Society* 93, 74-80.
- Dummett, M. (1959): A propositional calculus with denumerable matrix, *Journal Symbolic Logic* 24, 97-106.
- Dunn, J.M. (1976): Intuitive semantics for first-degree entailments and ‘coupled trees’, *Philosophical Studies* 29, 149-168.
- Fitting, M.C. (1991/92): Many-valued modal logics. I-II, *Fundamenta Informaticae* 15, 235-254; 17, 55-73.
- Gödel, K. (1932): Zum intuitionistischen Aussagenkalkül, *Anzeiger Akademie der Wissenschaften Wien,* Math.-naturwiss. Klasse 69, 65-66;
- ---- (1933), *Ergebnisse eines mathematischen Kolloquiums* 4 (1933), 40.
- Goguen, J.A. (1968-69): The logic of inexact concepts, *Synthese* 19, 325-373.
- Hay, L.S. (1963): Axiomatization of the infinite-valued predicate calculus, *Journal Symbolic Logic* 28, 77-86.
- Jaskowski, S. (1936): Recherches sur le système de la logique intuitioniste, in: *Actes du Congrès Internationale de Philosophie Scientifique 1936*. vol. 6, Paris, 58-61. [English translation: *Studia Logica* 34 (1975), 117-120.]
- Kleene, S.C. (1938): On notation for ordinal numbers, *Journal Symbolic Logic* 3, 150-155.
- Kripke, S.A. (1975): Outline of a theory of truth, *J. of Philosophy* 72, 690-716.
- Lukasiewicz, J. (1920): O logice trojwartosciowej, *Ruch Filozoficny* 5, 170-171. [English translation in: Lukasiewicz (1970).]
- Lukasiewicz, J. (1970): *Selected Works*. (ed.: L. Borkowski), North-Holland Publ. Comp., Amsterdam and PWN, Warsaw.
- McNaughton, R. (1951): A theorem about infinite-valued sentential logic, *Journal Symbolic Logic* 16, 1-13.
- Mundici, D. (1986): Interpretation of AF C*-algebras in Lukasiewicz sentential calculus. *J. Functional Analysis* 65, 15-63.
- Post, E. L. (1920): Determination of all closed systems of truth tables, *Bulletin American Mathematical Society* 26, 437.
- Post, E. L. (1921): Introduction to a general theory of elementary propositions, *American Journal Mathematics* 43, 163-185.
- Ragaz, M. (1983): Die Unentscheidbarkeit der einstelligen unendlichwertigen Prädikatenlogik, *Archiv mathematische Logik Grundlagenforschung* 23, 129-139.
- Rose, A. and Rosser, J.B. (1958): Fragments of many-valued statement calculi, *Transactions*

American Mathematical Society 87, 1-53.
- Scarpellini, B. (1962): Die Nichtaxiomatisierbarkeit des unendlichwertigen Prädikatenkalküls von Lukasiewicz, *Journal Symbolic Logic* 27, 159-170.
- Skolem, Th. (1957): Bemerkungen zum Komprehensionsaxiom, *Zeitschrift mathematische Logik Grundlagen Mathematik* 3, 1-17.
- White, R.B. (1979): The consistency of the axiom of comprehension in the infinite-valued predicate logic of Lukasiewicz, *J. Philosophical Logic* 8, 509-534.
- Zadeh, L.A. (1965): Fuzzy sets, *Information and Control* 8, 338-353.
- Zadeh, L.A. (1975): Fuzzy logic and approximate reasoning, *Synthese* 30, 407-428.
- Zadeh, L.A. (1978): Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets Systems* 1, 3-28.
- Zadeh, L.A. (1979): A theory of approximate reasoning, in: J.E. Hayes, D. Michie, L.I. Mikulich (eds.), *Machine Intelligence 9*. Halstead Press, New York, 149-194.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

logic: classical | logic: fuzzy | logic: paraconsistent | logic: relevance | Prior, Arthur | Sorites paradox

Siegfried Gottwald
gottwald@rz.uni-leipzig.de

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: April 25, 2000
Content last modified: April 25, 2000

Philip the Chancellor

Philip the Chancellor was an influential figure in a number of different circles in the first half of the thirteenth century. He enjoyed a long though rather turbulent ecclesiastical career and was famous for his sermons and his lyric poetry, the latter of which has received attention by a number of musicologists in recent years. In the areas of philosophy and theology, his major work, *Summa de bono*, which was composed sometime in the 1220s-1230s, was a ground-breaking achievement in many ways. Philip was one of the first to organize a *Summa* around a central foundational principle, the notion of the good. *Summa de bono* also contains most likely the earliest treatment of a topic that rose to prominence in the later medieval period, the doctrine of the transcendentals. Elements of Philip's theory of action drew comments from such later notables as Albert the Great. *Summa de bono* was a well-respected and influential work in the thirteenth century.

Philip's Life

Philip was born in Paris probably in the 1160s, the exact date unknown. He was a member of a prominent Parisian family; many of his relatives held important positions, in service to the French kings or to the church. Several were bishops of prominent sees, in particular, his uncle, Peter of Nemours, who was bishop of Paris from 1208 to 1218 and who helped to foster Philip's career. Most likely, Philip received his education at the nascent University of Paris, where he also taught. Sometime in the early 1200s, he was appointed the archdeacon of Noyon, a position he held even after becoming chancellor of the Cathedral of Notre Dame in Paris in 1217. Philip died in 1236, most likely on December 26.

Philip's Career as Chancellor

The position of chancellor at the cathedral of Notre Dame in Paris was one of some importance, although it was the lowest ranking office in the administrative hierarchy of the chapter. Minor duties included keeping the official seal of the chapter, executing the official decrees of the administrative body, maintaining the non-musical books of the library, and serving in the liturgy of the daily office. His most significant responsibility was in the area of education. The chancellor was originally the head of the cathedral school. As masters began to flock to Paris in the late twelfth and early thirteenth centuries, especially to the Ile-de-la-Cite, where Notre Dame is located, the chancellor's supervision was extended to include these masters as well. His power lay in his authority to grant the teaching license, which was required in order to hold classes in one of the schools springing up on the Ile-de-la-Cite. It was the chancellor's duty to evaluate candidates applying for this license. Moreover, the chancellor had the right to revoke the license should a master prove unworthy or incompetent, as well as the responsibility for maintaining order and discipline among the scholars within his jurisdiction.

By the time Philip became chancellor in 1217, the masters of the various schools in Paris had begun to seek autonomy from the cathedral chapter and had won a number of important concessions by papal decree. In 1215, the papal legate, Robert of Courcon, drew up a number of formal statutes, codifying the *de facto* practices regarding matters such as the examinations for teaching licenses, accepted dress and behavior, curriculum, and discipline of students. As a result, by the time that Philip became chancellor, at least on paper, it appeared as if the power of his office had been greatly reduced, even with respect to granting the teaching license. For although the chancellor retained the power to grant these licenses, the statutes dictated that he could not turn down anyone the masters deemed fit to teach. But in reality, a long struggle ensued between the masters and the chancellor, who sought to retain his power, beginning with Philip's predecessor (Stephen of Reims) and continuing into much of Philip's own tenure. Finally, in the late 1220s and early 1230s, Philip made his peace with the masters, who had gone on strike and left Paris along with many of their students in response to a conflict with the secular authorities. No doubt recognizing that their departure imperiled the continuing prestige of Paris as a center of education, as well as his own position, Philip worked hard to convince the scholars to return to Paris and reconvene their classes. His efforts were successful, and the masters returned in 1231.

Philip's Philosophical Significance

In his major philosophical work, *Summa de bono*, Philip uses the notion of the good as an organizing principle for his study. He divides the text roughly into four sections. Philip discusses first the nature of good in general terms, and then the highest good and its relationship to created goods. After this rather short introduction to the notion of the good, Philip goes on to examine in some depth the various sorts of created goods. He divides his discussion of created goods into three parts. First, he looks at the good retained by creatures by virtue of their natures (*bonum naturae*). In turn, these goods fall into two categories: those goods that cannot be diminished by evil and those goods that can be lost through evil. In his discussion of the former, Philip focuses first on the angels and their properties, and then on human

beings and their properties. Those goods that can be affected by evil are discussed in conjunction with Adam's fall from grace in the Garden and its consequences. Following the discussion of *bonum naturae*, Philip considers what he calls *bonum in genere*. Although this sort of good has a rather peculiar title, the *bonum in genere* represents goods that come about as a result of an agent's actions. These goods have this title because what determines whether a given act is good depends not only on the sort of act it is (its "form" so to speak) but also what the act has to do with (its "matter" so to speak), thus suggesting that these sorts of goods can be classified along the lines of genera and species. Moreover, these sorts of generic goods contrast with the meritorious goods brought about as the result of God's grace. After discussing the *bonum in genere*, Philip goes on to look at the good that is associated with grace. Here, he divides his treatment into the graces that pertain to angels (as well as their ministries) and the graces that pertain to human beings. Philip includes the virtues in his discussion of human graces. Although he denies that the virtues are a type of grace, he includes them in this section because virtues come about as a result of grace working within human beings. This idea is most naturally associated with the theological virtues of faith, hope, and charity, and indeed, Philip discusses these virtues here at some length. But he also includes in this section a lengthy discussion of the cardinal virtues: prudence, fortitude, temperance (and the associated virtues of modesty, sobriety, continence, and virginity), and justice, virtues which one might argue have no direct connection with grace. Philip admits that strictly speaking cardinal virtues are not divine virtues since they have to do with what is for the sake of the end and not directly with the end itself (the end of course being God). But he argues that justice has to do both with God and with human governance; perhaps because of this connection, Philip felt justified in including them in a broader discussion of grace. Philip ends his work with another source of grace, the seven gifts of the Holy Spirit. Philip's discussion of the good of grace in human beings occupies over half of *Summa de bono* with his unit on the virtues accounting for a large portion of this section.

Philip's *Summa de bono* represents a significant innovation in medieval philosophical work. His use of the good as an organizing principle is a departure from the explicitly theological structure of other well-known texts of the time including Peter Lombard's *Sentences* and William of Auxerre's *Summa aurea*. Philip is certainly concerned with theological issues, but unlike the *Sentences* or the *Summa aurea*, Philip devotes no significant sections of his work to topics such as God's nature, the Incarnation, or the sacraments. Furthermore, whereas often William's or Lombard's examination of a philosophical issue arises in the context of a larger theological issue, this is not always the case for Philip. For example, in William's *Summa aurea*, the issues of human action and its freedom arise within a discussion of Adam's fall in the Garden. William raises these issues here because he recognizes that holding Adam responsible for his sin requires that Adam acted freely. Philip places his treatment of free action within an examination of human psychology. It is not until after Philip discusses the character and the powers of the human soul that he gets around to examining theological issues such as the origin and the immortality of the soul. Philip is one of the first major thinkers in the Latin West whose work reflects the influence of the recent influx of newly translated texts from Aristotle and his Arabic commentators, particularly in the area of metaphysics. This may help to explain the distinctive character of his work. Philip's use of Aristotelian metaphysics is especially interesting given that many of Aristotle's metaphysical treatises and work in natural philosophy were officially banned at Paris during the time that Philip was working on *Summa de bono*.

Philip's Innovative Theory of the Transcendentals

Philip's influence in the thirteenth century was especially felt in the area of the transcendentals and in action theory. In this section, I will consider his theory of the transcendentals. I will examine his theory of action in the following section. To some extent, the doctrine of the transcendentals has its roots in the Christian-Platonic discussions of the relationship between created and divine being and goodness, in particular, in William of Auxerre's treatment of goodness prefatory to his examination of the virtues in *Summa aurea*. However, its more immediate ancestor is Aristotle's doctrine of the categories. Certain properties fall into none of Aristotle's categories; rather they are properties of all of the things to which the categories are applicable. For this reason, these properties are said to "transcend" the categories. Although there is some variation in what is counted as a transcendental, the list generally included being, unity, truth, and goodness. Thus, everything that falls into any of Aristotle's categories is a being, has a certain sort of unity, and is true and good to a certain extent.

Not only do these properties transcend the categories and as a result, apply to everything classified by the categories, but they are held to be convertible with each other as well. This could mean one of two things. The transcendentals could be coextensional, so that whatever has being also has unity, truth, and goodness. This leaves open the possibility that the transcendentals are separate and distinct from one another. The second option of the convertibility thesis involves a stronger claim, namely, the idea that the transcendentals differ from one another only in concept, not in reality. Unity, truth, and goodness add nothing to a particular being over and above what is already there; everything that is a being is also one, true, and good in virtue of the very same characteristics. But to describe something as a being and to describe it as, say, good is to express two different things about it since the concept of a being and the concept of a good are two very different concepts. Thus, while being and goodness are extensionally equivalent, they are intensionally distinct.

Philip adopts the second notion of convertibility. The various transcendentals do not differ in reality, only in concept. The concept of being is fundamental in that the concepts of the other transcendentals presuppose it. However, the concepts of all of the other transcendentals add a certain basic notion to the notion of being in order to differentiate them from being. This basic notion is the notion of being that is undivided. Because this is a purely negative notion, it picks out no additional property in reality. The addition of indivision alone yields the concept of unity. To derive the concepts of the true and the good, one adds further the notion of the appropriate cause. The concept of truth involves the idea of the formal cause, that is, the cause in virtue of which matter is enformed, and a thing becomes what it is. Things are true, that is, genuine instances of the kind of thing they are to the extent that they instantiate the form of things of that kind. Thus, the concept of truth is the concept of being that is undivided from a formal cause. Goodness, on the other hand, has to do with being that is undivided from a final cause, that is, a cause that has to do with goals or ends, especially those goals that have been brought to fulfillment. Everything has a particular nature, that is, properties that make that thing a thing of that type. But things can exemplify those properties to a greater or lesser extent. Philip claims that everything has as its goal its own perfection, which means that things move toward exemplifying their specifying characteristics to the greatest extent possible. To the extent that a thing does so, that thing will be good. But that thing will

also have being to the same extent. Thus, goodness and being in a given thing coincide in reality, and a thing's goodness adds nothing over and above the thing's being. But of course, goodness and being involve two different concepts. Thus, being and goodness have the same extension while differing intensionally.

It has been argued (by Pouillion and others) that Philip's discussion of the transcendentals in *Summa de bono* represents the earliest formal treatise on the transcendentals in the history of Western philosophy. About ten years earlier, William of Auxerre also discussed the relationship between being and goodness; he raises the issue of whether being and being good are the same. But in resolving this issue, William never considers the central issue of the transcendentals, the idea that they are extensionally equivalent while intensionally different. It appears that Philip is the first to do so. MacDonald argues that Philip's greater familiarity with the metaphysical works of Aristotle and the Arabic commentators accounts for the startling innovations of his work on the transcendentals. It is likely that Philip encountered the fundamental notion of extensional equivalence and intensional difference in the work of the Arabic commentators of Aristotle; both Avicenna and Averroes argue that unity and being have the same extension while differing conceptually. Philip extends this idea to include not only being and unity, but also the true and the good. Philip's work in turn sets the stage for the development of this topic throughout the thirteenth century, influencing the work of such notable thinkers as Alexander of Hales, Albert the Great, and Thomas Aquinas.

Philip's Psychology

Although not as well known, Philip's psychology also exhibits some innovative features and influenced later thinkers, in particular, Albert the Great. In the early thirteenth century, theorists accounted for human abilities by arguing that in order to do what they do, agents must possess certain capacities or powers. Thus for every ability, there must be a corresponding and separate power. Human beings have the ability to think; therefore, they must have a cognitive power, often called intellect or reason. Medieval philosophers also adopted Aristotle's distinction between practical intellect, which discerns what to do, and speculative intellect, which discerns the truth about the way things are. Human beings have the capacity for desire; therefore, they must have appetitive powers. Medieval philosophers distinguish between two kinds of appetite; a rational appetite, called the will, that is responsive to the dictates of the intellect, and an appetite that is responsive to sensory apprehension, called the sensory appetite. Since activities such as thinking and desiring are different sorts of activities, most theorists thought of their corresponding powers, the (practical) intellect and the will, as separate. However, on Philip's account, the practical intellect and the will are not separate powers. He argues that with respect to the capacity for performing actions, there is only one power with two separate acts. Philip gives a number of arguments for this position. One of them is particularly interesting because Philip uses the doctrine of the transcendentals to establish his conclusion that the practical intellect and will are one and the same power. Philip first notes that the intellect and the will have different ends. The intellect, even as practical intellect, has truth as its goal or end since its job, so to speak, is to ascertain the way things are, to make judgments about the state of reality, including what alternatives for action are available for the agent. The will, on the other hand, is an appetite for the good; it inclines the agent toward what she

judges to be good. Thus, its end is the good. But according to the doctrine of the transcendentals, the true and the good differ only intensionally, not extensionally. Philip thinks that if the ends of the powers do not differ extensionally, then the powers themselves do not differ extensionally as well. Thus, with respect to action, there is only one power with two different acts, acts of conceiving and judging on the one hand, and acts of desiring (the good) and willing on the other. Talking about the will is merely shorthand for referring to acts of willing or desiring. Talking about the intellect is merely shorthand for referring to certain cognitive judgments about what to do or how to act.

Although Philip denies that the practical intellect and the will are separate powers, he argues for a distinction among the various apprehensive powers. Thus, he sees a genuine distinction between the speculative intellect, the practical intellect, imagination, and the sensory apprehensive powers. This is because what is apprehended by each of these powers is different in nature. Philip denies that any such distinction is present among motive powers. Insofar as the appetite is moved by a sensory apprehension, we call it the sensory appetite. Insofar as the appetite is moved by the judgment of intellect, we call it the will. However, in reality, there is only one motive power to account for both of these sorts of desires, according to Philip. Philip's position on the inseparability of practical intellect and will does not appear to have convinced his thirteenth-century contemporaries or near-contemporaries. Albert the Great in *Summa de homine* (*circa* 1245) addresses the arguments given by Philip for his position although as was the custom of the time with respect to one's contemporaries or near-contemporaries, he does not refer to Philip by name. Albert rejects Philip's position, but the fact that Albert examines the issue at all indicates something about the esteem given to Philip's work.

An important part of thirteenth-century psychology was the development of a theory of free action. This was especially important both for theology and for ethics. Medieval thinkers, beginning with Augustine, recognized that moral responsibility requires freedom and so the possibility of that freedom needed to be explained. Along those same lines, they argued that unless human beings are able to act freely, God is not justified in punishing sins. Furthermore, human freedom plays a major role in theodicy; for example, Augustine argues that God is not responsible for the evil found in the world because that evil is perpetuated by the free choices of human beings. Thus, given this background and these commitments, it was common for medieval philosophers to examine the topic of freedom somewhere in their writings. In the first half of the thirteenth century, it was customary to examine these issues in the context of a treatise on what became known as *liberum arbitrium*. In medieval theories of action in the early part of the thirteenth century, "*liberum arbitrium*" is a technical term. It is a placeholder for whatever it is that enables human beings to act freely. The term originates in the work of Augustine who wrote a treatise entitled *De libero arbitrio*. The starting point for thirteenth-century treatises on *liberum arbitrium* was a definition taken from Peter Lombard's *Sentences*: "*liberum arbitrium* is a faculty of reason and will, by which good is chosen with the assistance of grace, or evil, when grace is not there to assist. And it is called '*liberum*' with respect to the will, which can be turned toward either [good or bad], while [it is called] '*arbitrium*' with respect to reason, as it has to do with that power or faculty to which the discerning between good and evil belongs." Although Lombard's main discussion of *liberum arbitrium* is found in book two, distinction twenty-five of his *Sententiae in IV libris distinctae*, this definition is found in the twenty-fourth distinction of book two, chapter three. This definition was commonly but mistakenly attributed to Augustine by commentators in the thirteenth century. Lombard himself does not reveal his

source. References to Augustine dominate his discussion of *liberum arbitrium* which might account for the association of the definition with Augustine.

Philosophers in the early thirteenth century faced the task of how to understand this definition. Although it is obvious from Lombard's formulation that both intellect and will have something to do with *liberum arbitrium*, their exact relationship is unclear. The phrase "*liberum arbitrium*" itself contributes to the uncertainty. The first part of the phrase, "*liberum*," is uncontroversial; it simply means "free." Difficulties arise with respect to the notion of "*arbitrium*." This notion has both cognitive and appetitive connotations, for it can have meanings as diverse as "judgment," "decision," "wish," or "inclination." It can also refer to a power or ability to make judgments or decisions or to the very agent who makes these judgments or decisions. Thus, the term covers a lot of territory, territory that has to do with both cognitive and appetitive capabilities. Accordingly, it is natural to connect *liberum arbitrium* with both intellect and will. In writing treatises on *liberum arbitrium*, thirteenth-century philosophers sought to sort out the connections between intellect and will on the one hand and the production and freedom of human action on the other. Some of these philosophers argued that free action results from the interaction of intellect and will, while others argued that although the intellect is an important precondition for an action's being free, the will is the true instrument that brings about a free action. Still others argued that *liberum arbitrium* is a separate faculty altogether although it is closely linked with and interacts with the intellect and will in the production of a free action. The practice of writing treatises on *liberum arbitrium* began to die off toward the later decades of the thirteenth century when philosophers started examining the topic of *voluntas libera* (free will) instead.

In his treatise on *liberum arbitrium* in *Summa de bono*, Philip adopts John Damascene's basic description of action. According to Damascene, an eighth-century patristic, a number of different stages come together in the production of an action. These stages include desiring, considering the various courses of action which will satisfy one's desires, deliberating over those courses of action, judging which one is to be performed, willing and choosing a particular alternative, and initiating the action. Since each of these activities are activities of the will and of the intellect, it follows that actions result from the activities of the will and the intellect. Damascene also claims that each of these stages are performed freely. Because each stage is performed freely, the resulting action is also free.

Philip modifies this position. He thinks that only the final activity of the intellect is performed freely, that is, the final judgment about what course of action to take. This is because Philip thinks that with respect to the previous activities of the intellect, activities such as identifying possible courses of action and deliberating over them, the intellect suffers from certain constraints. These constraints have to do with the structure of the world around us which in turn structures our beliefs. Beliefs play a role in our deliberations over what to do; thus the constraints placed upon beliefs in turn constrain the intellect in its activities. But according to Philip, the intellect retains some freedom for the final judgment about what to do is made freely. Thus, the intellect need not judge that a particular course of action be performed; it could have judged differently. In Philip's view, the will does not suffer from such constraints. The will is an appetite for the good so that whatever it wills, it wills a good. But a thing's being a good is merely a necessary condition, not a sufficient one. A thing's being good does not compel the will's act. Moreover, no judgment of the intellect constrains the will's choice; the will is free either to will the alternative put

forth by the intellect or to reject it and will something else. It is because of this capacity of the will that Philip sees the will as the primary source of freedom in a human being. For in the final analysis, it is the will rather than the judgment of the intellect that determines the action the agent performs. Because the will wills freely, the agent performs the action freely. Thus, while both intellect and will have important roles to play in the production of a free action, freedom is primarily a function of the will. Philip's theory of action helps to set the stage for the prominent voluntarist movement that rose later in the thirteenth century with thinkers such as Peter John Olivi and John Duns Scotus.

Bibliography

Primary Sources

- Philip the Chancellor. *Philippi Cancellarii Summa de bono*. N. Wicki, ed. Berne: Francke, 1985.

References and Further Reading

- Callus, D. A., O.P. "Philip the Chancellor and the *De anima* ascribed to Robert Grosseteste." *Medieval and Renaissance Studies* 1: 105-27.
- Gracia, Jorge J. E. "The Transcendentals in the Middle Ages: An Introduction." *Topoi* 11: 113-20.
- Kent, Bonnie. *Virtues of the Will: The Transformation of Ethics in the Late Thirteenth Century*. Washington, D.C.: Catholic University of America Press, 1995.
- Korolec, J.B. "Free will and free choice." *The Cambridge History of Later Medieval Philosophy*. Kretzmann *et al*, eds. Cambridge University Press: 1982, pp. 629-641.
- Lottin, Odon. *Psychologie et morale aux XIIe et XIIIe Siecles*. Gembloux, Belgium: J. Duculot, S.A. Editeur, 1957.
- MacDonald, Scott. *Being and Goodness: The Concept of the Good in Metaphysics and Philosophical Theology*. Ithaca: Cornell University Press, 1991.
- MacDonald, Scott. "Goodness as Transcendental: The Early Thirteenth-Century Recovery of an Aristotelian Idea." *Topoi* 11: 173-86.
- Payne, Thomas Blackburn II. *Poetry, Politics, and Polyphony: Philip the Chancellor's Contribution to the Music of the Notre Dame School*, vol.1, ch.1, "The Life of Philip the Chancellor." Ph.D. diss. University of Chicago, 1991.
- Potts, Timothy. *Conscience in Medieval Philosophy*. Cambridge: Cambridge University Press, 1980.
- Wicki, N. *Philippi Cancellarii Summa de bono*, vol.1, ch.1 "Vie de Philippe le Chancelier." Berne: Francke, 1985.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

free choice: medieval theories of[= *liberum arbitrium*] | free will | transcendentals

Colleen McCluskey
St. Louis University
mcclusc@slu.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: March 20, 1999
Content last modified: March 20, 1999

Artifact

An artifact may be defined as an object that has been intentionally made or produced for a certain purpose. Often the word ‘artifact’ is used in a more restricted sense to refer to simple, hand-made objects (for example, tools) which represent a particular culture. (This might be termed the ‘archaeological sense’ of the word.) In experimental science, the expression ‘artifact’ is sometimes used to refer to experimental results which are not manifestations of the natural phenomena under investigation, but are due to the particular experimental arrangement.

Artifacts

Artifacts are contrasted to natural objects; they are products of human actions. Consequently an artifact has necessarily a maker or an author. Using the word ‘author’ in a somewhat generalized sense, we may thus adopt the principle:

> (A1) If an object is an artifact, it has an author.

Can (A1) be strengthened to an equivalence? Experimental artifacts are unintended products of the experimenter's plans and actions, but otherwise the word is usually applied only to intended products: not all products of an agent's actions are artifacts. If we restrict the application of the expressions ‘author’ and ‘authorship’ in a similar way, we might strengthen (A1) to

> (A2) An object is an artifact if and only if it has an author.

According to (A2), *artifact* and *author* are correlative concepts (Hilpinen 1993). It should be observed

that (A2) allows the possibility that an artifact has more than one author: such objects may be termed ‘collective artifacts’. (A2) makes the concept of artifact equivalent to that of *work* (as product as opposed to activity); for example, according to (A2), all works of art, including musical and literary works, should be called ‘artifacts’ insofar as they have authors. In aesthetics, the expression ‘artifact’ has been used in this wide sense when it has been argued, as many philosophers have done, that works of art are necessarily artifacts. In a more restricted sense which is closer to the "archaeological" meaning mentioned above, artifacts are physical objects or in any case inhabitants of the physical, spatio-temporal world. If we say that an author can create a work only by making some artifact (e.g., to write a novel, one has to produce a manuscript), the expression ‘artifact’ is used in this narrow (or primary) sense.

An object which is an artifact in this primary sense is usually made from some pre-existing object or objects by successive intentional modifications; this activity is called *work*. This feature of artifacts is reflected in the definition of an artifact as an object "showing human workmanship or modification".

When a person intends to make an object, his productive intention has as its content some description of the intended object; the agent intends to make an object of a certain kind. An author's intention "ties" to an artifact a number of predicates which determine the *intended character* of the object. The existence and some of the properties of the artifact are dependent on its intended character. This is expressed by the following *Dependence Condition*:

> (DEP) The existence and some of the properties of an artifact depend on an agent's (or author's) intention to make an object of certain kind.

The causal tie between an artifact and its intended character -- or, strictly speaking, between an artifact and the author's productive intention -- is mediated by the author's actions, that is, by his work on the object. The actual properties of an artifact constitute its *actual character*. The success of the author's productive activity depends on the degree of fit or agreement between the intended and the actual character of the object. The actual character of an artifact is of course always much richer than the intended character; the artifact fits the author's intentions if and only if the former includes the latter. At least one of the descriptions included in the intended character must be a sortal predicate which determines the identity of the object and the criteria by which it can be distinguished from other objects. For example, ‘painting’ and ‘chair’ are sortal descriptions, but ‘red thing’ is not: it is possible to give a definite answer to the question of how many chairs there are in a given room, but not to the question of how many red things there are in the room.

The Evaluation of Artifacts

Often an artifact is identified by a sortal description which refers to its intended function (e.g., ‘hammer’). But this need not always be the case: for example, ‘painting’ is an artifact sortal which is not derived from the purpose or function of the object, but from the way in which it has been produced. An object that has been made for a purpose *F* may be termed ‘an *F*-object’. The properties of an *F*-object can be divided into two classes: (i) those relevant to the functioning of the object as an *F*-object, and (ii) the

properties irrelevant to the purpose *F*. The former properties may be termed the *significant* properties of the object (or its *F-significant* properties); they may also be called the "good-making properties" of the object. For example, the weight of a hammer is one of its significant features, but its color is not. In addition to an identifying (sortal) description *F*, the content of an author's productive intention includes the properties that he regards as significant for the purpose *F*. The latter properties depend on this purpose (or purposes); thus the intended character of an artifact is not simply a collection of predicates, but has a hierarchical structure. In many cases an object is expected to serve many different purposes; thus the description *F* may be quite complex.

An author's productive activity may be *evaluated* on the basis of the relationships among the intended character of an artifact, its actual character, and a purpose *F*:

> (E1) The degree of fit or agreement between the intended character and the actual character of an object,

> (E2) The degree of fit between the intended character of an object and the purpose *F*, in other words, the suitability of an object of the intended kind for the purpose *F*,

and

> (E3) The degree of fit between the actual character of an object and the purpose *F*, that is, the suitability of an artifact for *F*.

(E1) determines whether an artifact is a successful embodiment of the author's intentions, (E2) determines whether the character that the author intends to give to an artifact is suitable for the purpose *F*, and (E3) tells whether the author has succeeded in making an object that is in fact suitable for the purpose *F*. The study of artifacts (*as* artifacts) is intrinsically evaluative, since viewing an object as an artifact means viewing it in the light of intentions and purposes.

The purpose *F* on which the evaluation of an artifact and its design is based need not be the purpose that the author had in mind; it can be any purpose for which the artifact might be used. The direction of evaluation may be reversed so that the maker or owner of an artifact tries to find new uses for it. In addition, we should distinguish the actual character of an artifact from the author's conception of it. If the author's conception of an object agrees with its intended character, the artifact is subjectively satisfactory for the author, but it may fail to fit the author's productive intentions if he has a mistaken conception of it.

If the author's productive activity is successful, the character of a completed artifact both depends on and agrees with his productive intentions so that it can be regarded as an embodiment of these intentions.If the actual character of an object does not agree with its intended character, it is unsatisfactory from the author's point of view, and if the author's conception of an object does not agree with its intended character, the artifact is subjectively unsatisfactory from the author's point of view. In the latter case the

author has a reason to try to improve the object until it satisfies his productive intentions. A change in an object which improves the fit between its actual and intended character is, from the standpoint of the author's intentions, a *progressive* change.

It seems plausible to regard an object as a proper artifact only if its maker's productive activity has some degree of success, for example, satisfies some sortal predicate included in his productive intention. This condition may be termed the *Success Condition*:

> (SUC) An object is an artifact made by an author only if it satisfies some sortal description included in the author's productive intention.

If an agent's activity fails in every respect, the agent does not accomplish anything, but produces only "scrap". But even if the object does not fit the author's productive intention, but he *accepts* it as a satisfactory realization of his intention, it may be regarded as a proper artifact; this is expressed by the following *Acceptance Condition*:

> (ACC) An object is an artifact made by an author only if the author accepts it as satisfying some sortal description included in his productive intention.

If an artifact has several authors, the Acceptance Condition should hold for at least one of them. According to the Acceptance Condition, an object is an artifact only if its maker regards it as such, that is, accepts it as a product of his intentional activity. The Success Condition concerns the fit between the actual and the intended character of an object; the Acceptance Condition the fit between the author's conception of an object and its intended character. In this context it should be observed that the author's intention may change during his productive activity. In the above conditions, ‘productive intention’ should be regarded as referring to the content of the author's "final" intentions concerning the artifact.

The conditions listed above provide a partial characterization of the concept of artifact. We might say that different intentionally modified objects exhibit different degrees of artifactuality, depending on how well they satisfy these conditions.

Randall Dipert's (1993) theory of artifacts includes the condition that an artifact (in the strict sense) should be intended by its author to be recognized as having been intentionally modified for a certain purpose. This is a plausible condition, since an *F*-object can presumably be a good *F*-object only if its potential users recognize it as such. However, this recognizability should not be taken to mean general recognizability: a mechanical shark used in making an adventure film is an artifact, but its authors do not wish the audience to recognize it as such, on the contrary; the condition of recognizability concerns only the persons who are using it in the making of the film.

Works of Art

As was mentioned above, artifactuality is often regarded as a defining characteristic of works of art (Stephen Davies 1991); for example, this is an essential condition in George Dickie's (1984) analysis, according to which a work of art is an "artifact of a kind created to be presented to an Artworld public". The condition of artifactuality is plausible only if the concept of artifact is understood in a wide sense in which intentionally created events and processes (e.g., performances) and works which have instances (for example, musical and literary works) can be regarded as artifacts. According to condition (A2), the condition of artifactuality in this sense is equivalent to the requirement that a work of art should have an author. Some philosophers of art have rejected the condition of artifactuality, using instances of "driftwood art" and analogous examples as counterexamples. According to condition (A2), this view has the seemingly paradoxical consequence that a work of art need not be a product of anyone's work and need not have an author. Other philosophers have responded to such examples by extending the concept of artifactuality in such a way that the presentation of a natural object as an object of aesthetic appreciation counts as an "intentional modification" required for artifactuality. If the expression ‘artifact’ is used in a sufficiently wide sense, the condition of artifactuality clearly holds for artworks, but it is equally obvious that not all works of art (or works in general) are artifacts in the narrow sense of the word. In aesthetic evaluation and criticism, however, they are treated as if they were artifacts.

Bibliography

- Stephen Davies, 1991. *Definitions of Art*, Ch. 5. Ithaca and London: Cornell University Press.
- George Dickie, 1984. *The Art Circle: A Theory of Art*. New York: Haven.
- Randall R. Dipert, 1993. *Artifacts, Art Works, and Agency*. Philadelphia: Temple University Press.
- Risto Hilpinen, 1992. ‘Artifacts and Works of Art’. *Theoria* 58, 58-82.
- Risto Hilpinen, 1993. ‘Authors and Artifacts’. *Proceedings of the Aristotelian Society* 93, 155-178.

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

aesthetics

Risto Hilpinen
hilpinen@miami.edu

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: January 5, 1999
Content last modified: January 5, 1999

Indispensability Arguments in the Philosophy of Mathematics

One of the most intriguing features of mathematics is its applicability to empirical science. Every branch of science draws upon large and often diverse portions of mathematics, from the use of Hilbert spaces in quantum mechanics to the use of differential geometry in general relativity. It's not just the physical sciences that avail themselves of the services of mathematics either. Biology, for instance, makes extensive use of difference equations and statistics. The roles mathematics plays in these theories is also varied. Not only does mathematics help with empirical predictions, it allows elegant and economical statement of many theories. Indeed, so important is the language of mathematics to science, that it is hard to imagine how theories such as quantum mechanics and general relativity could even be stated without employing a substantial amount of mathematics.

From the rather remarkable but seemingly uncontroversial fact that mathematics is indispensable to science, some philosophers have drawn serious metaphysical conclusions. In particular, Quine (1976; 1980a; 1980b; 1981a; 1981c) and Putnam (1979a; 1979b) have argued that the indispensability of mathematics to empirical science gives us good reason to believe in the existence of mathematical entities. According to this line of argument, reference to (or quantification over) mathematical entities such as sets, numbers, functions and such is indispensable to our best scientific theories, and so we ought to be committed to the existence of these mathematical entities. To do otherwise is to be guilty of what Putnam has called "intellectual dishonesty" (Putnam 1979b, p. 347). Moreover, mathematical entities are seen to be on an epistemic par with the other theoretical entities of science, since belief in the existence of the former is justified by the same evidence that confirms the theory as a whole (and hence belief in the latter). This argument is known as the Quine-Putnam indispensability argument for mathematical realism. There are other indispensability arguments,[1] but this one is by far the most influential, and so in what follows I'll concentrate on it.

Spelling Out the Quine-Putnam Indispensability Argument

The Quine-Putnam indispensability argument has attracted a great deal of attention, in part because many see it as the best argument for mathematical realism (or platonism). Thus anti-realists about mathematical entities (or nominalists) need to identify where the Quine-Putnam argument goes wrong. Many platonists, on the other hand, rely very heavily on this argument to justify their belief in mathematical entities. The argument places nominalists who wish to be realist about other theoretical entities of science (quarks, electrons, black holes and such) in a particularly difficult position. For typically they accept something quite like the Quine-Putnam argument[2]) as justification for realism about quarks and black holes. (This is what Quine (1980b, p. 45) calls holding a "double standard" with regard to ontology.)

For future reference I'll state the Quine-Putnam indispensability argument in the following explicit form:

> (P1) We ought to have ontological commitment to all and only the entities that are indispensable to our best scientific theories.
>
> (P2) Mathematical entities are indispensable to our best scientific theories.
>
> (C) We ought to have ontological commitment to mathematical entities.

Thus formulated, the argument is valid. This forces the focus onto the two premises. In particular, a couple of important questions naturally arise. The first concerns how we are to understand the claim that mathematics is indispensable. I address this in the next section. The second question concerns the first premise. It is nowhere near as self-evident as the second and it clearly needs some defense. I'll discuss its defense in the following section. I'll then present some of the more important objections to the argument, before considering the Quine-Putnam argument's role in the larger scheme of things - where it stands in relation to other influential arguments for and against mathematical realism.

What is it to be Indispensable?

The question of how we should understand 'indispensability' in the present context is crucial to the Quine-Putnam argument, and yet it has received surprisingly little attention. Quine actually speaks in terms of the entities quantified over in the canonical form of our best scientific theories rather than indispensability. Still, the debate continues in terms of indispensability, so we would be well served to clarify this term.

The first thing to note is that 'dispensability' is not the same as 'eliminability'. If this were not so, *every* entity would be dispensable (due to a theorem of Craig).[3] What we require for an entity to be 'dispensable' is for it to be eliminable *and* that the theory resulting from the entity's elimination be an attractive theory. (Perhaps, even stronger, we require that the resulting theory be *more* attractive than the original.) We will need to spell out what counts as an attractive theory but for this we can appeal to the standard desiderata for good scientific theories: empirical success; unificatory power; simplicity; explanatory power; fertility and so on. Of course there will be debate over what desiderata are appropriate and over their relative weightings, but such issues need to be addressed and resolved independently of issues of indispensability. (See Burgess (1983) and Colyvan (1999b) for more on these issues.)

These issues naturally prompt the question of *how much* mathematics is indispensable (and hence how much mathematics carries ontological commitment). It seems that the indispensability argument only justifies belief in enough mathematics to serve the needs of science. Thus we find Putnam speaking of "the set theoretic 'needs' of physics" (Putnam 1979b, p. 346) and Quine claiming that the higher reaches of set theory are "mathematical recreation ... without ontological rights" (Quine 1986, p. 400) since they do not find physical applications. One could take a less restrictive line and claim that the higher reaches of set theory, although without physical applications, do carry ontological commitment by virtue of the fact that they have applications *in other parts of mathematics*. So long as the chain of applications eventually "bottoms out" in physical science, we could rightfully claim that the whole chain carries ontological commitment. Quine himself justifies some transfinite set theory along these lines (Quine 1984, p. 788), but he sees no reason to go beyond the constructible sets (Quine 1986, p. 400). His reasons for this restriction, however, have little to do with the indispensability argument and so supporters of this argument need not side with Quine on this issue.

Naturalism and Holism

Although both premises of the Quine-Putnam indispensability argument have been questioned, it's the first premise that is most obviously in need of support. This support comes from the doctrines of naturalism and holism.

Following Quine, naturalism is usually taken to be the philosophical doctrine that there is no first philosophy and that the philosophical enterprise is continuous with the scientific enterprise (Quine 1981b). By this Quine means that philosophy is neither prior to nor privileged over science. What is more, science, thus construed (i.e. with philosophy as a continuous part) is taken to be the complete story of the world. This doctrine arises out of a deep respect for scientific methodology and an acknowledgment of the undeniable success of this methodology as a way of answering fundamental questions about all nature of things. As Quine suggests, its source lies in "unregenerate realism, the robust state of mind of the natural scientist who has never felt any qualms beyond the negotiable uncertainties internal to science" (Quine 1981b, p.72). For the metaphysician this means looking to our best scientific theories to determine what exists, or, perhaps more accurately, what we ought to believe to

exist. In short, naturalism rules out unscientific ways of determining what exists. For example, naturalism rules out believing in the transmigration of souls for mystical reasons. Naturalism would not, however, rule out the transmigration of souls if our best scientific theories were to require the truth of this doctrine.[4]

Naturalism, then, gives us a reason for believing in the entities in our best scientific theories and no other entities. Depending on exactly how you conceive of naturalism, it may or may not tell you whether to believe in *all* the entities of your best scientific theories. I take it that naturalism does give us *some* reason to believe in all such entities, but that this is defeasible. This is where holism comes to the fore: in particular, confirmational holism.

Confirmational holism is the view that theories are confirmed or disconfirmed as wholes (Quine 1980b, p. 41). So, if a theory is *confirmed* by empirical findings, the *whole* theory is confirmed. In particular, whatever mathematics is made use of in the theory is also confirmed (Quine 1976, pp. 120-122). Furthermore, as Putnam (1979a) has stressed, it is the same evidence that is appealed to in justifying belief in the mathematical components of the theory that is appealed to in justifying the empirical portion of the theory (if indeed the empirical can be separated from the mathematical at all). Naturalism and holism taken together then justify P1. Roughly, naturalism gives us the "only" and holism gives us the "all" in P1.

It is worth noting that in Quine's writings there are at least two holist themes. The first is the confirmational holism discussed above (often called the Quine-Duhem thesis). The other is semantic holism which is the view that the unit of meaning is not the single sentence, but systems of sentences (and in some extreme cases the whole of language). This latter holism is closely related to Quine's well-known denial of the analytic-synthetic distinction (Quine 1980b) and his equally famous indeterminacy of translation thesis (Quine 1960). Although for Quine, semantic holism and confirmational holism are closely related, there is good reason to distinguish them, since the former is generally thought to be highly controversial while the latter is considered relatively uncontroversial.

Why this is important to the present debate is that Quine explicitly invokes the controversial semantic holism in support of the indispensability argument (Quine 1980b, pp. 45-46). Most commentators, however, are of the view that only confirmational holism is required to make the indispensability argument fly (see, for example, Colyvan (1998); Field (1989, pp. 14-20); Hellman (199?); Resnik (1995a; 1997); Maddy (1992)) and my presentation here follows that accepted wisdom. It should be kept in mind, however, that while the argument, thus construed, is Quinean in flavor it is not, strictly speaking, Quine's argument.

Objections

There have been many objections to the indispensability argument, including Charles Parsons' (1980) concern that the obviousness of basic mathematical statements is left unaccounted for by the Quinean

picture and Philip Kitcher's (1984, pp. 104-105) worry that the indispensability argument doesn't explain *why* mathematics is indispensable to science. The objections that have received the most attention, however, are those due to Hartry Field, Penelope Maddy and Elliott Sober. In particular, Field's nominalisation program has dominated recent discussions of the ontology of mathematics.

Field (1980) presents a case for denying the second premise of the Quine-Putnam argument. That is, he suggests that despite appearances mathematics is not indispensable to science. There are two parts to Field's project. The first is to argue that mathematical theories don't have to be true to be useful in applications, they need merely to be *conservative*. (This is, roughly, that if a mathematical theory is added to a nominalist scientific theory, no nominalist consequences follow that wouldn't follow from the nominalist scientific theory alone.) This explains why mathematics *can* be used in science but it does not explain why it *is* used. The latter is due to the fact that mathematics makes calculation and statement of various theories much simpler. Thus, for Field, the utility of mathematics is merely pragmatic - mathematics is not indispensable after all.

The second part of Field's program is to demonstrate that our best scientific theories can be suitably nominalised. That is, he attempts to show that we could do without quantification over mathematical entities and that what we would be left with would be reasonably attractive theories. To this end he is content to nominalise a large fragment of Newtonian gravitational theory. Although this is a far cry from showing that *all* our current best scientific theories can be nominalised, it is certainly not trivial. The hope is that once one sees how the elimination of reference to mathematical entities can be achieved for a typical physical theory, it will seem plausible that the project could be completed for the rest of science.[5]

There has been a great deal of debate over the likelihood of the success of Field's program but few have doubted its significance. Recently, however, Penelope Maddy, has pointed out that if P1 is false, Field's project may turn out to be irrelevant to the realism/anti-realism debate in mathematics.

Maddy presents some serious objections to the first premise of the indispensability argument (Maddy 1992; 1995; 1997). In particular, she suggests that we ought not have ontological commitment to *all* the entities indispensable to our best scientific theories. Her objections draw attention to problems of reconciling naturalism with confirmational holism. In particular, she points out how a holistic view of scientific theories has problems explaining the legitimacy of certain aspects of scientific and mathematical practices. Practices which, presumably, ought to be legitimate given the high regard for scientific practice that naturalism recommends. It is important to appreciate that her objections, for the most part, are concerned with methodological consequences of accepting the Quinean doctrines of naturalism and holism - the doctrines used to support the first premise. The first premise is thus called into question by undermining its support.

Maddy's first objection to the indispensability argument is that the actual attitudes of working scientists towards the components of well-confirmed theories vary from belief, through tolerance, to outright rejection (Maddy 1992, p. 280). The point is that naturalism counsels us to respect the methods of

working scientists, and yet holism is apparently telling us that working scientists ought not have such differential support to the entities in their theories. Maddy suggests that we should side with naturalism and not holism here. Thus we should endorse the attitudes of working scientists who apparently do not believe in *all* the entities posited by our best theories. We should thus reject P1.

The next problem follows from the first. Once one rejects the picture of scientific theories as homogeneous units, the question arises whether the mathematical portions of theories fall within the true elements of the confirmed theories or within the idealized elements. Maddy suggests the latter. Her reason for this is that scientists themselves do not seem to take the indispensable application of a mathematical theory to be an indication of the truth of the mathematics in question. For example, the false assumption that water is infinitely deep is often invoked in the analysis of water waves, or the assumption that matter is continuous is commonly made in fluid dynamics (Maddy 1992, pp. 281-282). Such cases indicate that scientists will invoke whatever mathematics is required to get the job done, without regard to the truth of the mathematical theory in question (Maddy 1995, p. 255). Again it seems that confirmational holism is in conflict with actual scientific practice, and hence with naturalism. And again Maddy sides with naturalism. (See also Parsons (1983) for some related worries about Quinean holism.) The point here is that if naturalism counsels us to side with the attitudes of working scientists on such matters, then it seems that we ought not take the indispensability of some mathematical theory in a physical application as an indication of the truth of the mathematical theory. Furthermore, since we have no reason to believe that the mathematical theory in question is true, we have no reason to believe that the entities posited by the (mathematical) theory are real. So once again we ought to reject P1.

Maddy's third objection is that it is hard to make sense of what working mathematicians are doing when they try to settle independent questions. These are questions, that are independent of the standard axioms of set theory - the ZFC axioms.[6] In order to settle some of these questions, new axiom candidates have been proposed to supplement ZFC, and arguments have been advanced in support of these candidates. The problem is that the arguments advanced seem to have nothing to do with applications in physical science: they are typically intra-mathematical arguments. According to indispensability theory, however, the new axioms should be assessed on how well they cohere with our current best scientific theories. That is, set theorists should be assessing the new axiom candidates with one eye on the latest developments in physics. Given that set theorists do not do this, confirmational holism again seems to be advocating a revision of standard mathematical practice, and this too, claims Maddy, is at odds with naturalism (Maddy 1992, pp. 286-289).

Although Maddy does not formulate this objection in a way that directly conflicts with P1 it certainly illustrates a tension between naturalism and confirmational holism.[7] And since both these are required to support P1, the objection indirectly casts doubt on P1. Maddy, however, endorses naturalism and so takes the objection to demonstrate that confirmational holism is false. I'll leave the discussion of the impact the rejection of confirmational holism would have on the indispensability argument until after I outline Sober's objection, because Sober arrives at much the same conclusion.

Elliott Sober's objection is closely related to Maddy's second and third objections. Sober (1993) takes

issue with the claim that mathematical theories share the empirical support accrued by our best scientific theories. In essence, he argues that mathematical theories are not being tested in the same way as the clearly empirical theories of science. He points out that hypotheses are confirmed relative to competing hypotheses. Thus if mathematics is confirmed along with our best empirical hypotheses (as indispensability theory claims), there must be mathematics-free competitors. But Sober points out that *all* scientific theories employ a common mathematical core. Thus, since there are no competing hypotheses, it is a mistake to think that mathematics receives confirmational support from empirical evidence in the way other scientific hypotheses do.

This in itself does not constitute an objection to P1 of the indispensability argument, as Sober is quick to point out (Sober 1993, p. 53), although it does constitute an objection to Quine's overall view that mathematics is part of empirical science. As with Maddy's third objection, it gives us some cause to reject confirmational holism. The impact of these objections on P1 depends on how crucial you think confirmational holism is to that premise. Certainly much of the intuitive appeal of P1 is eroded if confirmational holism is rejected. In any case, to subscribe to the conclusion of the indispensability argument in the face of Sober's or Maddy's objections is to hold the position that it's permissible at least to have ontological commitment to entities that receive no empirical support. This, if not outright untenable, is certainly not in the spirit of the original Quine-Putnam argument.

Conclusion

It is not clear how damaging the above criticisms are to the indispensability argument. Indeed, the debate is very much alive, with many recent articles devoted to the topic. (See bibliography notes below.) Closely related to this debate is the question of whether there are any other decent arguments for platonism. If, as some believe, the indispensability argument is the *only* argument for platonism worthy of consideration, then if it fails, platonism in the philosophy of mathematics seems bankrupt. Of relevance then is the status of other arguments for and against mathematical realism. In any case, it is worth noting that the indispensability argument is one of a small number of arguments that have dominated discussions of the ontology of mathematics. It is therefore important that this argument not be viewed in isolation.

The two most important arguments *against* mathematical realism are the epistemological problem for platonism - how do we come by knowledge of causally inert mathematical entities? (Benacerraf 1983b) - and the indeterminacy problem for the reduction of numbers to sets - if numbers are sets, which sets are they (Benacerraf 1983a)? Apart from the indispensability argument, the other major argument *for* mathematical realism is that it is desirable to provide a uniform semantics for *all* discourse: mathematical and non-mathematical alike (Benacerraf 1983b). Mathematical realism, of course, meets this challenge easily, since it explains the truth of mathematical statements in exactly the same way as in other domains.[8] It is not so clear, however, how nominalism can provide a uniform semantics.

Finally, it is worth stressing that even if the indispensability argument *is* the only good argument for platonism, the failure of this argument does not necessarily authorize nominalism, for the latter too may

be without support. It does seem fair to say, however, that if the objections to the indispensability argument are sustained then one of the most important arguments for platonism is undermined. This would leave platonism on rather shaky ground.[9]

Bibliography

Although the indispensability argument is to be found in many places in Quine's writings (including 1976; 1980a; 1980b; 1981a; 1981c), the *locus classicus* is Putnam's short monograph *Philosophy of Logic* (included as a chapter of the second edition of the third volume of his collected papers (Putnam, 1979b)). See also Putnam (1979a) and the introduction of Field (1989) which has an excellent outline of the argument. Colyvan (2001) is a sustained defence of the argument.

See Chihara (1973), and Field (1980; 1989) for attacks on the second premise and Colyvan (1999b; 2001), Maddy (1990), Malament (1982), Resnik (1985), Shapiro (1983) and Urquhart (1990) for criticisms of Field's program. For a fairly comprehensive look at nominalist strategies in the philosophy of mathematics (including a good discussion of Field's program), see Burgess and Rosen (1997), while Feferman (1993) questions the amount of mathematics required for empirical science. See Azzouni (1997), Balaguer (1996b; 1998), Maddy (1992; 1995; 1997), Melia (2000), Peressini (1997), Sober (1993) and Vineberg (1996) for attacks on the first premise. Colyvan (1998; 1999a; 2001; 2002), Hellman (1999) and Resnik (1995a; 1997) reply to some of these objections.

For variants of the Quinean indispensability argument see Maddy (1992) and Resnik (1995a).

- Azzouni, J., 1997, "Applied Mathematics, Existential Commitment and the Quine-Putnam Indispensability Thesis", *Philosophia Mathematica (3)* **5**/3 (October): 193-209
- Balaguer, M., 1996a, "Towards a Nominalization of Quantum Mechanics", *Mind* **105**/418 (April): 209-226
- Balaguer, M., 1996b, "A Fictionalist Account of the Indispensable Applications of Mathematics", *Philosophical Studies* **83**/3 (September): 291-314
- Balaguer, M., 1998, *Platonism and Anti-Platonism in Mathematics,* New York: Oxford University Press
- Benacerraf, P., 1983a, "What Numbers Could Not Be", reprinted in Benacerraf and Putnam (1983), pp. 272-294
- Benacerraf, P., 1983b, "Mathematical Truth", reprinted in Benacerraf and Putnam (1983), pp. 403-420 and in Hart (1996), pp. 14-30
- Benacerraf, P. and Putnam, H. (eds.), 1983, *Philosophy of Mathematics: Selected Readings,* 2nd edition, Cambridge: Cambridge University Press
- Burgess, J., 1983, "Why I Am Not a Nominalist", *Notre Dame Journal of Formal Logic* **24**/1 (January): 93-105
- Burgess, J. and Rosen, G., 1997, *A Subject with No Object: Strategies for Nominalistic Interpretation of Mathematics,* Oxford: Clarendon
- Chihara, C., 1973, *Ontology and the Vicious Circle Principle,* Ithaca, NY: Cornell University

Press
- Colyvan, M., 1998, "In Defence of Indispensability", *Philosophia Mathematica (3)* **6**/1 (February): 39-62
- Colyvan, M., 1999a, "Contrastive Empiricism and Indispensability", *Erkenntnis* **51**/2-3 (September): 323-332
- Colyvan, M., 1999b, "Confirmation Theory and Indispensability", *Philosophical Studies* **96**/1 (October): 1-19
- Colyvan, M., 2001, *The Indispensability of Mathematics*, New York: Oxford University Press
- Colyvan, M., 2002, "Mathematics and Aesthetic Considerations in Science", *Mind* forthcoming
- Feferman, S., 1993, "Why a Little Bit Goes a Long Way: Logical Foundations of Scientifically Applicable Mathematics", *Proceedings of the Philosophy of Science Association* **2**: 442-455
- Field, H.H., 1980, *Science Without Numbers: A Defence of Nominalism*, Oxford: Blackwell
- Field, H.H., 1989, *Realism, Mathematics and Modality*, Oxford: Blackwell
- Hart, W.D. (ed.), 1996, *The Philosophy of Mathematics*, Oxford: Oxford University Press
- Hellman, G., 1999, "Some Ins and Outs of Indispensability: A Modal-Structural Perspective", in A. Cantini, E. Casari and P. Minari (eds.), *Logic and Foundations of Mathematics*, Dordrecht: Kluwer, pp. 25-39
- Irvine, A.D. (ed.), 1990, *Physicalism in Mathematics*, Dordrecht: Kluwer
- Kitcher, P., 1984, *The Nature of Mathematical Knowledge*, New York: Oxford University Press
- Maddy, P., 1990, "Physicalistic Platonism", in A.D. Irvine (ed.), *Physicalism in Mathematics*, Dordrecht: Kluwer, pp. 259-289
- Maddy, P., 1992, "Indispensability and Practice", *Journal of Philosophy* **89**/6 (June): 275-289
- Maddy, P., 1995, "Naturalism and Ontology", *Philosophia Mathematica (3)* **3**/3 (September): 248-270
- Maddy, P., 1997, *Naturalism in Mathematics*, Oxford: Clarendon Press
- Maddy, P., 1998, "How to be a Naturalist about Mathematics", in H.G. Dales and G. Oliveri (eds.), *Truth in Mathematics*, Oxford: Clarendon, pp. 161-180
- Malament, D., 1982, "Review of Field's *Science Without Numbers*", *Journal of Philosophy* **79**/9 (September): 523-534 and reprinted in Resnik (1995b), pp. 75-86
- Melia, J., 2000, "Weaseling Away the Indispensability Argument", *Mind* **109**/435 (July): 455-479
- Parsons, C., 1980, "Mathematical Intuition", *Proceedings of the Aristotelian Society* **80** (1979-1980): 145-168 and reprinted in Resnik (1995b), pp. 589-612 and in Hart (1996), pp. 95-113
- Parsons, C., 1983, "Quine on the Philosophy of Mathematics", in *Mathematics in Philosophy: Selected Essays*, Ithaca, NY: Cornell University Press, pp. 176-205
- Peressini, A., 1997, "Troubles with Indispensability: Applying Pure Mathematics in Physical Theory", *Philosophia Mathematica (3)* **5**/3 (October): 210-227
- Putnam, H., 1979a, "What is Mathematical Truth", in *Mathematics Matter and Method: Philosophical Papers Vol. 1*, 2nd edition, Cambridge: Cambridge University Press, pp. 60-78
- Putnam, H., 1979b, "Philosophy of Logic", reprinted in *Mathematics Matter and Method: Philosophical Papers Vol. 1*, 2nd edition, Cambridge: Cambridge University Press, pp. 323-357
- Quine, W.V., 1960, *Word and Object*, Cambridge, MA: Massachusetts Institute of Technology Press
- Quine, W.V., 1976, "Carnap and Logical Truth" reprinted in *The Ways of Paradox and Other*

Essays, revised edition, Cambridge, MA: Harvard University Press, pp. 107-132 and in Benacerraf and Putnam (1983), pp. 355-376

- Quine, W.V., 1980a, "On What There Is", reprinted in *From a Logical Point of View,* 2nd edition, Cambridge, MA: Harvard University Press, pp. 1-19
- Quine, W.V., 1980b, "Two Dogmas of Empiricism", reprinted in *From a Logical Point of View,* 2nd edition, Cambridge, MA: Harvard University Press, pp. 20-46 and in Hart (1996), pp. 31-51 (Page references are to the first reprinting)
- Quine, W.V., 1981a, "Things and Their Place in Theories", in *Theories and Things,* Cambridge, MA: Harvard University Press, pp. 1-23
- Quine, W.V., 1981b, "Five Milestones of Empiricism", in *Theories and Things*, Cambridge, MA: Harvard University Press, pp. 67-72
- Quine, W.V., 1981c, "Success and Limits of Mathematization", in *Theories and Things,* Cambridge, MA: Harvard University Press, pp. 148-155
- Quine, W.V., 1984, "Review of Parsons', *Mathematics in Philosophy", Journal of Philosophy* **81**/12 (December): 783-794
- Quine, W.V., 1986, "Reply to Charles Parsons", in L. Hahn and P. Schilpp (eds.), *The Philosophy of W.V. Quine,* La Salle, ILL: Open Court, pp. 396-403
- Resnik, M.D., 1985, "How Nominalist is Hartry Field's Nominalism", *Philosophical Studies* **47** (March): 163-181
- Resnik, M.D., 1995a, "Scientific Vs Mathematical Realism: The Indispensability Argument", *Philosophia Mathematica (3)* **3**/2 (May): 166-174
- Resnik, M.D. (ed.), 1995b, *Mathematical Objects and Mathematical Knowledge,* Aldershot (UK): Dartmouth
- Resnik, M.D., 1997, *Mathematics as a Science of Patterns,* Oxford: Clarendon Press
- Shapiro, S., 1983, "Conservativeness and Incompleteness", *Journal of Philosophy* **80**/9 (September): 521-531 and reprinted in Resnik (1995b), pp. 87-97 and in Hart (1996), pp. 225-234
- Sober, E., 1993, "Mathematics and Indispensability", *Philosophical Review* **102**/1 (January): 35-57
- Urquhart, A., 1990, "The Logic of Physical Theory", in A.D. Irvine (ed.), *Physicalism in Mathematics,* Dordrecht: Kluwer, pp. 145-154
- Vineberg, S., 1996, "Confirmation and the Indispensability of Mathematics to Science" *PSA 1996* (Philosophy of Science, supplement to vol. 63), pp. 256-263

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

abduction | meaning holism | naturalism | nominalism: in metaphysics | Platonism: in metaphysics | Quine, Willard van Orman | realism

Mark Colyvan
mark.colyvan@utas.edu.au

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Table of Contents

First published: December 21, 1998
Content last modified: July 5, 2001

Stanford Encyclopedia of Philosophy
Notes to Indispensability Arguments in the Philosophy of Mathematics

---

Notes

1. In general, an indispensability argument is an argument that purports to establish the truth of some claim based on the indispensability of the claim in question for certain purposes (to be specified by the particular argument). For example, if *explanation* is specified as the purpose, then we have an explanatory indispensability argument. Thus we see that inference to the best explanation is a special case of an indispensability argument. See the introduction of Field (1989, pp. 14-20) for a nice discussion of indispensability arguments and inference to the best explanation. See also Maddy (1992) and Resnik 1995a) for variations on the Quine-Putnam version of the argument.

2. Most scientific realists accept inference to the best explanation. Indeed, it might be said that inference to the best explanation is the cornerstone of scientific realism. But, as we saw in note 1, inference to the best explanation may be seen as a kind of indispensability argument, so any realist who accepts the former while rejecting the latter is in a somewhat delicate position.

3. This theorem states that relative to a partition of the vocabulary of an axiomatizable theory *T* into two classes, *t* and *o* (theoretical and observational, say) there exists an axiomatizable theory *T** in the language whose only non-logical vocabulary is *o*, of all and only the consequences of *T* that are expressible in *o* alone. If the vocabulary of the theory can be partitioned in the way that Craig's theorem requires, then the theory can be reaxiomatized so that apparent reference to any given theoretical entity is eliminated. See Field (1980, p. 8) for further details.

4. It turns out that the details of the formulation of naturalism are crucial to the argument. See Maddy (1998) for a slightly different formulation that doesn't support the conclusion of the Quine-Putnam argument.

5. The issue of how likely modern theories such as general relativity and quantum mechanics are to yield to nominalisation is a very interesting and controversial matter. David Malament (1982) argues that quantum mechanics, for one, is likely to resist nominalisation because of the central role infinite-dimensional Hilbert spaces play in the theory. Mark Balaguer (1996a; 1998), however, suggests a way of nominalising the Hilbert spaces in question.

6. For example, the continuum hypothesis is the assertion that the cardinality of the real numbers is the first non-denumerable cardinal. It turns out that neither this hypothesis nor its negation is provable from the ZFC axioms; the question of the size of the continuum is *independent* of ZFC.

7. It is tempting to formulate Maddy's third objection as follows:

> Naturalism endorses actual mathematical practice. This practice typically does not rely on extra-mathematical canons of justification. So it seems that naturalism endorses belief in some mathematical theories with or without finding physical applications for them. Presumably this belief carries with it certain ontological commitments. So it seems that we ought to have ontological commitment to at least some entities that are not indispensable to our best (empirical) theories. Thus it might be argued that a thoroughgoing naturalist is not committed to *only* the entities of our best scientific theories and thus is not committed to P1.

This, however, I think is to misrepresent Maddy's argument. For starters, the indispensability argument can be made to work with P1 reading: 'We ought to have ontological commitment to all the entities that are indispensable to our best scientific theories'. That is, the 'only' part argued against in the above formulation of Maddy's objection is unnecessary. More importantly, I think that Maddy's objection is subtler than this formulation. The objection is directed at the motivation for P1 - not at P1 itself. Her aim is to show that confirmational holism ought to be rejected. This would rob P1 of much of its plausibility, as we shall see in the discussion following that dealing with Sober's objection.

8. For example, 'There is a city larger than Hobart' and 'There is a natural number larger than 100' are true just in case there exist objects in the respective domains of quantification with the properties of being larger than Hobart and being larger than the natural number 100 respectively. Such a strategy is not open to anyone who holds that (i) both these sentences are true, (ii) numbers do not exist and (iii) cities do exist. Most nominalists thus find themselves unable to employ such a semantics. Field is a notable exception here (Field, 1980; 1989); he denies that sentences such as 'There is a natural number larger than 100' are true. He claims that they are *true-in-the-story-of-mathematics* but that this is not true *simpliciter*. He thus avoids the difficult problem of providing a nominalisticlly acceptable account of the truth of such sentences. (He does, however, have a residual worry. According to Field, sentences such as 'In Peano number theory there is a number larger than 100' *are* true *simpliciter*, but it's not clear that he has a fully-developed semantics for these.)

9. I'd like to thank Helen Regan, Angela Rosier and Ed Zalta for comments on earlier versions of this entry.

Mark Colyvan
mark.colyvan@utas.edu.au

First published: December 21, 1998
Content last modified: February 5, 1999

Thought Experiments

Thought experiments are devices of the imagination used to investigate nature. We need only list a few of the well-known thought experiments to be reminded of their enormous influence and importance in the sciences: Newton's bucket, Maxwell's demon, Einstein's elevator, Heisenberg's gamma-ray microscope, Schrödinger's cat. The 17th century saw some of its most brilliant practitioners in Galileo, Descartes, Newton, and Leibniz. And in our own time, the creation of quantum mechanics and relativity are almost unthinkable without the crucial role played by thought experiments. Galileo and Einstein were, arguably, the most impressive thought experimenters, but they were by no means the first. Thought experiments existed throughout the middle ages, and can be found in antiquity, too.

Examples of Thought Experiments

One of the most beautiful early examples (Lucretius, *De Rerum Natura*) attempts to show that space is infinite: If there is a boundary to the universe, we can toss a spear at it. If the spear flies through, it isn't a boundary after all; if the spear bounces back, then there must be something beyond the supposed edge of space, a cosmic wall which is itself in space that stopped the spear. Either way, there is no edge of the universe; space is infinite. This example nicely illustrates many of the common features of thought experiments: We visualize some situation; we carry out an operation; we see what happens. It also illustrates their fallibility. (In this case we've learned how to conceptualize space so that it is both finite and unbounded.)

Often a real experiment that is the analogue of a thought experiment is impossible for physical, technological, or just plain practical reasons; but this needn't be a defining condition of thought experiments. The main point is that we seem able to get a grip on nature just by thinking, and therein lies the great interest for philosophy. How is it possible to learn (apparently) new things about nature without new empirical data?

Ernst Mach (who seems to have coined the expression *Gedankenexperiment*) developed an interesting empiricist view in his classic, *The Science of Mechanics*. We possess, he says, a great store of "instinctive knowledge" picked up from experience. This needn't be articulated at all, but comes to the fore when we consider certain situations. One of his favourite examples is due to Simon Stevin. When a chain is draped over a double frictionless plane, as in Fig. 1a, how will it move? Add some links as in Fig. 1b.

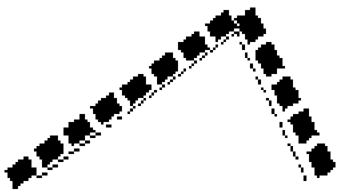


Figure 1a

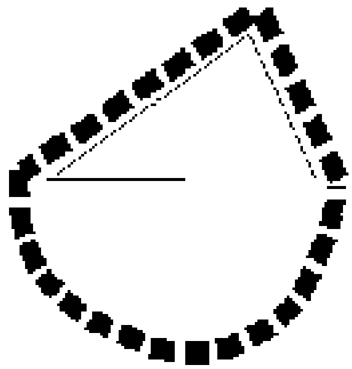


Figure 1b

Now it is obvious. The initial setup must have been in static equilibrium. Otherwise, we would have a perpetual motion machine; and according to our experience-based "instinctive knowledge", says Mach, this is impossible.

Recent Work on Thought Experiments

Thomas Kuhn's "A Function for Thought Experiments" employs many of the concepts (but not the terminology) of his well-known Structure of Scientific Revolutions. On his view a well-conceived thought experiment can bring on a crisis or at least create an anomaly in the reigning theory and so contribute to paradigm change. So thought experiments can teach us something new about the world, even though we have no new data, by helping us to reconceptualize the world in a better way.

Recent years have seen a sudden growth of interest in thought experiments. The views of Brown (1991) and Norton (1991, 1996) represent the extremes of platonic rationalism and classic empiricism, respectively. Norton claims that any thought experiment is really a (possibly disguised) argument; it starts with premisses grounded in experience and follows deductive or inductive rules of inference in arriving at its conclusion. The picturesque features of any thought experiment which give it an experimental flavour might be psychologically helpful, but are strictly redundant. Thus, says Norton, we never go beyond the empirical premisses in a way to which any empiricist would object. (For criticisms see Brown (1991, 1993) and for a defense see Norton (1996).)

By contrast, Brown holds that in a few special cases we do go well beyond the old data to acquire *a*

priori knowledge of nature. Galileo showed that all bodies fall at the same speed with a brilliant thought experiment that started by destroying the then reigning Aristotelian account. The latter holds that heavy bodies fall faster than light ones (H > L). But consider (Fig. 2), in which a heavy canon ball (H) and light musket ball (L) are attached together to form a compound object (H+L); the latter must fall faster than the cannon ball alone. Yet the compound object must also fall slower, since the light part will act as a drag on the heavy part. Now we have a contradiction. (H+L > H and H > H+L) That's the end of Aristotle's theory; but there is a bonus, since the right account is now obvious: they all fall at the same speed (H = L = H+L).

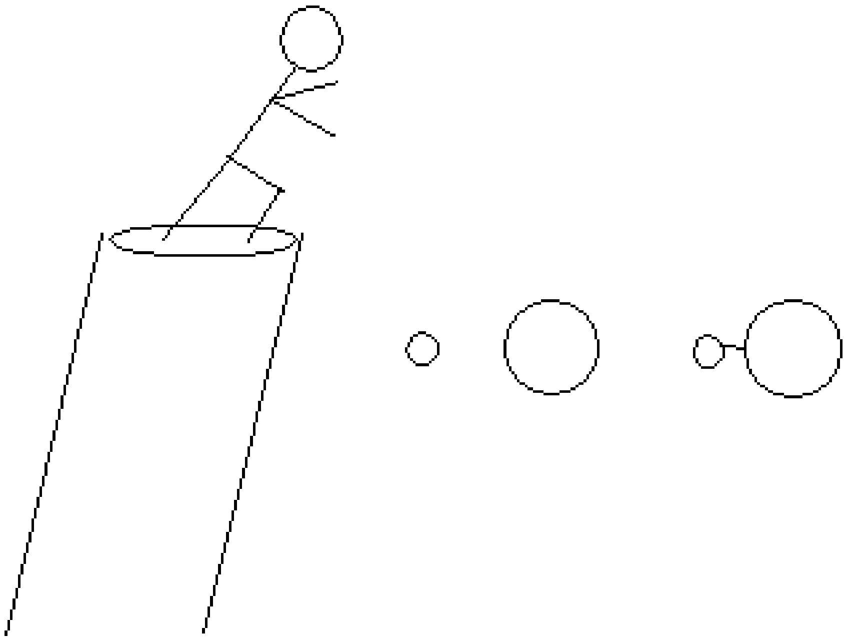


Figure 2

This is said to be *a priori* (though still fallible) knowledge of nature since there are no new data involved, nor is the conclusion derived from old data, nor is it some sort of logical truth. This account of thought experiments is further developed by linking the a priori epistemology to a recent account of laws of nature which holds that laws are relations between objectively existing abstract entities. It is thus a rather platonistic view, not unlike platonistic accounts of mathematics such as that urged by Gödel. (For details see Brown 1991.)

The two views just sketched might occupy the opposite ends of a spectrum of positions on thought experiments. Some of the promising new alternative views include those of Sorensen (somewhat in the spirit of Mach) who holds that thought experiments are a "limiting case" of ordinary experiments; they can achieve their aim, he says, without being executed. (Sorensen's book is also valuable for its extensive discussion of thought experiments in philosophy of mind, ethics, and other areas of philosophy, as well as the sciences.) Other promising views include those of Gooding (who stresses the similar procedural nature of thought experiments and real experiments), Miscevic and Nersessian (each of whom tie thought experiments to "mental models"), and several of the accounts in Horowitz and Massey (1991). More recent excellent discussions include: Arthur (1999), Gendler (1998), Haggqvist (1996), Humphreys

(1994), Genz (1999), McAllister (1996). The literature on thought experiments continues to grow rapidly.

Bibliography

- Arthur, R. (1999) "On Thought Experiments as A Priori Science," *International Studies in the Philosophy of Science*, 13, 3, 215-229
- Brown, J.R. (1991) *Laboratory of the Mind: Thought Experiments in the Natural Sciences*, London: Routledge
- Brown, J.R. (1993) "Why Empiricism Won't Work", in D. Hull, M. Forbes, and K. Okruhlik (eds.) *PSA 1992*, vol. 2, East Lansing, MI: Philosophy of Science Association
- Gendler, T.S.(1998) "Galileo and the Indispensibility of Scientific Thought Experiment", *The British Journal for the Philosophy of Science*, Vol.49, No.3, (September), 397-424
- Genz, H. (1999) *Gedanken-experimente*, Wiley-VCH, Weinheim, (in German)
- Gooding, D. (1993) "What is Experimental About Thought Experiments?" in D. Hull, M. Forbes, and K. Okruhlik (eds.) *PSA 1992*, vol. 2, East Lansing, MI: Philosophy of Science Association
- Hacking, I. (1993) "Do Thought Experiments have a Life of Their Own?" in D. Hull, M. Forbes, and K. Okruhlik (eds.) *PSA 1992*, vol. 2, East Lansing, MI: Philosophy of Science Association
- Haggqvist, S. (1996) *Thought Experiments in Philosophy*, (Stockholm: Almqvist & Wiksell International)
- Horowitz, T. and G. Massey (eds.) (1991) *Thought Experiments in Science and Philosophy*, Savage MD: Rowman and Littlefield
- Humphries, P. (1994) "Seven Theses on Thought Experiments", in Earman *et al.*, (eds) *Philosophical Problems of the Internal and External World*, Pittsburgh: University of Pittsburgh Press
- Kuhn, T. (1964) "A Function for Thought Experiments", reprinted in Kuhn, *The Essential Tension*, Chicago: University of Chicago Press, 1977
- Mach, E. (1960) *The Science of Mechanics*, (Trans by J. McCormack), sixth edition, LaSalle Illinois: Open Court
- Mach, E. (1976) "On Thought Experiments", in *Knowledge and Error*, Dordrecht: Reidel
- McAllister, J. (1996) "The Evidential Significance of Thought Experiments in Science", *Studies in History and Philosophy of Science*, vol 27, no. 2, 233-250
- Miscevic , N. (1992) "Mental Models and Thought Experiments", *International Studies in the Philosophy of Science*, Vol. 6, No. 3, pp. 215-226
- Nersessian, N. (1993) "In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling" in D. Hull, M. Forbes, and K. Okruhlik (eds.) *PSA 1992*, vol. 2, East Lansing, MI: Philosophy of Science Association
- Norton, J. (1991) "Thought Experiments in Einstein's Work", in Horowitz and Massey (1991)
- Norton, J. (1996) "Are Thought Experiments Just What You Always Thought?" *Canadian Journal of Philosophy*
- Sorensen, R. (1992) *Thought Experiments*, Oxford: Oxford University Press

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

Descartes, René | Leibniz, Gottfried Wilhelm | Mach, Ernst | science, philosophy of

James Robert Brown
jrbrown@chass.utoronto.ca

First published: December 28, 1996
Content last modified: May 1, 2002

Category Theory

Category theory is a general mathematical theory of structures and sytems of structures. It allows us to see, among other things, how structures of different kinds are related to one another as well as the universal components of a family of structures of a given kind. The theory is philosophically relevant in more than one way. For one thing, it is considered by many as being an alternative to set theory as a foundation for mathematics. Furthermore, it can be thought of as constituting a theory of concepts. Finally, it sheds a new light on many traditional philosophical questions, for instance on the nature of reference and truth.

General Definitions

Category theory is a generalized mathematical theory of structures. One of its goals is to reveal the universal properties of structures of a given kind via their relationships with one another. Formally, a category **C** can be described as a collection **Ob**, the *objects* of **C**, which satisfy the following conditions:

> For every pair *a*, *b* of objects, there is a collection **Mor**(*a*, *b*), namely, the morphisms from *a* to *b* in **C** (when **f** is a morphism from *a* to *b*, we write **f**: $a \rightarrow b$);
>
> For every triple *a*, *b* and *c* of objects, there is a partial operation from pairs of morphisms in **Mor**(*a*, *b*) X **Mor**(*b*, *c*) to morphisms in **Mor**(*a*, *c*), called the composition of morphisms in **C**
> (when **f**: $a \rightarrow b$ and **g**: $b \rightarrow c$, (**g** o **f**): $a \rightarrow c$ is their composition);
>
> For every object *a*, there is a morphism **id***a* in Mor(*a*, *a*), called the identity on *a*.

Furthermore, morphisms have to satisfy two axioms:

> Associativity: if **f**: $a \rightarrow b$, **g**: $b \rightarrow c$ and **h**: $c \rightarrow d$, then **h** o (**g** o **f**) = (**h** o **g**) o **f**
>
> Identity: if **f**: $a \rightarrow b$, then (**id***b* o **f**) = **f** and (**f** o **id***a*) = **f**.

One of the interesting features of category theory is that it provides a uniform treatment of the notion of structure. This can be seen, first, by considering the variety of examples of categories. Almost every known example of a mathematical structure with the appropriate structure preserving map yields a category. Thus, sets with functions between them constitute a category. Groups with group homomorphisms constitute a category. Topological spaces with continuous maps constitute a category. Vector spaces and linear maps constitute a category. Differential manifolds with smooth maps constitute a category. And so on. It is important to note that what characterizes a category is its morphisms and not its objects. Thus, the category of topological spaces with *open* maps is a different category than the category of topological spaces with continuous maps. In particular, the latter will have different properties as a category than the former. The previous examples have something in common: the objects are all structured sets with structure preserving maps. However, any entity satisfying the conditions given in the definition is a category. Thus, any preordered set is a category. For, given two elements *p*, *q* of the preordered set, there is a morphism **f**: $p \rightarrow q$ if and only if *p* is smaller or equal to *q*. Hence, a preordered set is a category in which there is at most one morphism between any two objects. A deductive system such that the entailment relation is reflexive and transitive is a category. Any monoid (and thus any group) can be seen as a category: in this case, the category has only one object and the morphisms of the category are given by the elements of the monoid. Composition of morphisms corresponds to multiplication of elements of the monoid. It is easily checked that the axioms of a monoid corresponds to the axioms of a category in this particular case. It can therefore be said that the notion of a category is a generalization of the concept of preorder *and at the same time* a generalization of the notion of monoid.

Category theory unifies mathematical structures in a second, perhaps even more important, manner. Once a type of structure has been defined, it quickly becomes imperative to determine how new structures can be constructed out of the given one and how given structures can be decomposed into more elementary substructures. For instance, given two sets A and B, set theory allows us to construct their cartesian product A X B. For an example of the second sort, given a finite abelian group, it can be decomposed into a product of some of its subgroups. In both cases, it is necessary to know how structures of a certain kind combine. The nature of these combinations might appear to be considerably different when looked at from too close. Category theory reveals that many of these constructions are in fact special cases of objects in a category with what is called a "universal property". Indeed, from a categorical point of view, a set-theoretical cartesian product, a direct product of groups, a direct product of abelian groups, a product of topological spaces and a conjunction of propositions in a deductive system are all instances of a categorical concept: the categorical product. What characterizes the latter is a universal property. Formally, a product for two objects *a* and *b* in a category **C** is an object *c* of **C** *together with* two morphisms, called the projections, **p**: $c \rightarrow a$ and **q**: $c \rightarrow b$ such that, and this is the universal property, for all object *d* with morphisms **f**: $d \rightarrow a$ and **g**: $d \rightarrow b$, there is a unique morphism **h**: $d \rightarrow c$ such that **p** o **h**

= **f** and **q** o **h** = **g**. Notice that we have defined a product for *a* and *b* and not the product for *a* and *b*. Indeed, products and, in fact, every object with a universal property, are defined up to (a unique) isomorphism. Thus, in category theory, the nature of the elements constituting a certain construction is irrelevant. What matters is the way an object is related to the other objects of the category, that is, the morphisms going in and the morphisms going out, or, put differently, how certain structures can be mapped into it and how it can map its structure into other structures of the same kind.

Another crucial aspect of category theory is that it allows to see how different kind of structures are related to one another. For instance, in algebraic topology, topological spaces are related to groups by various means (homology, cohomology, homotopy, K-theory). It was precisely in order to clarify how these connections are made and to compare them with one another that Eilenberg and Mac Lane invented category theory. Indeed, topological spaces with continuous maps constitute a category and similarly groups with group homomorphisms. In the very spirit of category theory, what should matter here are the morphisms between categories. These are given by functors and are informally structure preserving maps between categories. This simply means that, given two categories **C** and **D**, a functor F from **C** to **D**, should send objects of **C** to objects of **D** and morphisms of **C** to morphisms of **D** in such a way that composition of morphisms in **C** is preserved, i.e. F(**g** o **f**) = F(**g**) o F(**f**), and identity morphisms are preserved, i.e. F(**id***a*) = **id**F*a*. It follows immediately that a functor preserves commutativity of diagrams between categories. Homology, cohomology, homotopy, K-theory are all example of functors. A more direct example is provided by the power set operation which yields two functors on the category of sets, depending on how one defines its action on functions. Thus, given a set X, P(X) is the usual set of subsets of X and given a function f: X → Y, P(f): P(X) → P(Y) takes a subset A of X and maps it to B = f(A), the image of f restricted to A in Y. It is easily verified that it is a functor. There are in general many functors between two given categories and it becomes natural to ask how they are connected. For instance, given a category **C**, there is always the identity functor from **C** to **C** which sends every object of **C** to itself and every morphism of **C** to itself. In particular, there is the identity functor over the category of sets. Now, the identity functor is related to the power set functor described above in a natural manner. Indeed, given a set X and its power set P(X), there is a function hX which takes an element x of X and send it to the singleton set {x}, a subset of X, i.e. an element of P(X). This function in fact belongs to a family of functions indexed by the objects of the category of sets {hY: Y → P(Y)| Y in **Ob(Set)**}. Moreover, it satisfies the following commutativity condition. Given any function f: X → Y, the identity functor yields the same function Id(f): Id(X) → Id(Y). The commutativity condition thus becomes: hY o Id(f) = P(f) o hX. Thus the family of functions h(-) relates the two functors in a natural manner. Such families of morphisms are called *natural transformations* between functors.

The above notions constitute the elementary concepts of category theory. However it should be noted that they are not fundamental notions of category theory. These are arguably the notions of limits/colimits which are, in turn, special cases of what is certainly the cornerstone of the theory, the concept of adjoint functors. We will not present the definition here. Suffice it to say that adjoint functors pervade mathematics and this pervasiveness is certainly one of the most mysterious fact that category theory reveals about mathematics and probably thinking in general.

Brief Historical Sketch

It is difficult to do justice to the short but intricate history of the field, in particular it is not possible to mention all those who have contributed to its rapid development. This warning being said, here are some of the main threads that have to be mentioned.

Categories, functors, natural transformations, limits and colimits appeared almost out of nowhere in 1945 in Eilenberg & Mac Lane's paper entitled "General Theory of Natural Equivalences". We said "almost", because when one looks at their 1942 paper "Group Extensions and Homology", one discovers specific functors and natural transformations at work. In fact, it was basically the need to clarify and abstract from their 1942 results that Eilenberg & Mac Lane came up with the notions of category theory. The central notion for them, as the title indicates, was the notion of natural transformation. In order to give a general definition of the latter, they defined the notion of functor, borrowing the terminology from Carnap, and in order to give a general definition of functor, they defined the notion of category, borrowing this time from Kant and Aristotle. After their 1945 paper, it was not clear that the concepts of category theory would be more than a convenient language. Then in 1950, Mac Lane used the language of categories to give an arrow theoretic definition of the notion of product in general. However, it was only in 1957 and in 1958 that the situation radically changed. Indeed, in 1957 Grothendieck published his landmark "Sur quelques points d'algebre homologique" in which categories are used intrinsically to define and construct more general theories which are then applied to specific fields, in particular, in the following years, algebraic geometry, and in 1958 Kan published "Adjoint functors" and showed that the latter concept subsumes the important concepts of limits and colimits and also captures crucial conceptual situations. From then on, category theory became more than a convenient language.

Indeed, in the sixties, category theory developed rapidly, mainly in the context of algebraic geometry, algebraic topology and universal algebra. One of the main features which appears on the scene then and which is the direct development of Grothendieck's work is the characterization of abstract categories in which various branches of mathematics can be developed. For instance, what is called an abelian category is given by categorical conditions, usually the existence of certain adjoint functors, which garantee the development, in this case, of a large portion of homological algebra. In the same period, and this time mostly under the influence of William Lawvere, the idea that category theory could be used as a foundation of mathematics came to be considered and developed seriously. This approach culminated in the development of an other idea due to Grothendieck and his school: the notion of a topos.

Even though the concept of a topos was presented in the 1960s, it was certainly Lawvere & Tierney's work on the elementary axiomatization of the concept, published in the early 1970s, which gave to the notion its foundational status and impetus. Very roughly, a topos is a category which also possess a rich logical structure, rich enough to develop most of "ordinary mathematics", that is, most of what is taught in an undergraduate degree in mathematics. But it is also a generalized topological space and thus provides a direct connection between logic and geometry. The 1970s saw the development and application of the concept. (For more on the history of topos theory, see McLarty 1992.)

Finally, from the 1980s to this day, category theory found new applications. On the one hand, it now has many applications to theoretical computer science where it has firm roots and contributes, among other things, to the development of the semantics of programming and the development of new logical systems. On the other hand, its applications to mathematics are becoming more diversified and it even touches upon theoretical physics where higher-dimensional category theory, which is to category theory what higher-dimensional geometry is to plane geometry, is used in the study of knots.

Philosophical Significance

Category theory challenges philosophers in two non-exclusive ways. On the one hand, it is certainly the task of philosophy to clarify the general epistemological status of category theory and, in particular, its foundational status. On the other hand, category theory can be used by philosophers in their exploration of philosophical and logical problems. These two aspects can be illustrated briefly in turn.

Category theory is now a common tool in the toolbox of mathematicians. That much is clear. It is also clear that category theory unifies and provides a fruitful organization of mathematics. Given these simple facts, it remains to be seen whether category theory should be "on the same plane", so to speak, with set theory, whether it should be considered seriously as providing a foundational alternative to set theory or whether it is foundational in a different sense altogether. (The same question applies, in fact and even with more force, to topos theory, but we will unfortunately ignore this area here.) Arguments in favor of category theory and arguments against category theory as a foundational framework have been advanced. (See Marquis 1995 for a quick overview and a proposal.) This is in itself a complicated issue which is rendered even more difficult by the fact that the foundations of category theory itself still have to be clarified. Given that most of philosophy of mathematics of the last 50 years or so has been done under the assumption that mathematics is more or less set theory in disguise, the retreat of set theory in favor of category theory would necessarily have an important impact on philosophical thinking.

The use of category theory for logical and philosophical studies is already well underway. Indeed, categorical logic, the study of logic with the help of categorical means, has been around for about 30 years now and is still vigorous. On that front, many important results have been obtained but are still largely ignored by philosophers. Suffice it to mention the generalization of Kripke-Beth semantics for intuitionistic logic to sheaf semantics by Joyal, the discovery of the so-called geometric or coherent logic, whose practical and conceptual significance still has to be exposed, the notion and theorems of strong conceptual completeness, the geometric proofs of the independence of the continuum hypothesis and other strong axioms of set theory, the development of synthetic differential geometry which provides an alternative to standard and non-standard analysis, the construction of the so-called effective topos in which every funtion on the natural numbers is recursive, categorical models of linear logic, modal logic and higher-order type theories in general and the development of a graphical syntax, called sketches. Category theory also provides relevant information to more general philosophical questions. For instance, Ellerman 1987 has tried to show that category theory constitutes a theory of universals which has properties radically different from set theory considered as a theory of universals. If we move from universals to concepts in general, we can see how category theory could be useful even in cognitive

science. Indeed, Macnamara and Reyes have already tried to use categorical logic to provide a different logic of reference. (See, for instance, Macnamara & Reyes 1994.) McLarty, Marquis and Awodey have tried to show how it sheds an interesting light on structuralists approach to mathematical knowledge.

Thus, category theory is philosophically relevant in many ways and which will undoubtedly have to be taken into account in the years to come.

Bibliography

- Awodey, S., 1996, "Structure in Mathematics and Logic: A Categorical Perspective", *Philosophia Mathematica*, **(3)**, 4, 209-237.
- Bell, J., 1981, "Category Theory and the Foundations of Mathematics", *British Journal for the Philosophy of Science*, **32**, 349-358.
- Bell, J., 1986, "From Absolute to Local Mathematics", *Synthese*, **69**, 409-426.
- Bell, J., 1988, *Toposes and Local Set Theories: An Introduction*, Oxford: Oxford University Press.
- Eilenberg, S. & Mac Lane, S., 1942, "Group Extensions and Homology", *Annals of Mathematics*, **43**, 757-831.
- Eilenberg, S. & Mac Lane, S., 1945, "General Theory of Natural Equivalences", *Transactions of the American Mathematical Society*, **58**, 231-294.
- Ellerman, D., 1987, "Category Theory and Concrete Universals", *Synthese*, **28**, 409-429.
- Feferman, S., 1977, "Categorical Foundations and Foundations of Category Theory", *Logic, Foundations of Mathematics and Computability*, R. Butts (ed.), Reidel, 149-169.
- Grothendieck, A., 1957, "Sur Quelques Points d'Algebre Homologique", *Tçhoku Mathematics Journal*, **9**, 119-221.
- Kan, D. M., 1958, "Adjoint Functors", *Transactions of the American Mathematical Society*, **87**, 294-329.
- Lambek, J. & Scott, P.J., 1986, *Introduction to Higher Order Categorical Logic*, Cambridge: Cambridge University Press.
- Lawvere, F. W., 1966, "The Category of Categories as a Foundation for Mathematics", *Proceedings of the Conference on Categorical Algebra*, La Jolla, New York: Springer Verlag, 1-21.
- Lawvere, F. W., 1969, "Adjointness in Foundations", *Dialectica*, **23**, 281-295.
- Lawvere, F. W., 1972, "Introduction", *Toposes, Algebraic Geometry and Logic*, Lecture Notes in Mathematics, 274, Springer-Verlag, 1-12.
- Lawvere, F. W., 1975, "Continuously Variable Sets: Algebraic Geometry = Geometric Logic", *Proceedings of the Logic Colloquium Bristol 1973*, Amsterdam: North Holland, 135- 153.
- Mac Lane, S., 1950, "Dualities for Groups", *Bulletin of the American Mathematical Society*, **56**, 485-516.
- Mac Lane, S., 1971, *Categories for the Working Mathematician*, New York: Springer Verlag.
- Mac Lane, S. & Moerdijk, I., 1992, *Sheaves in Geometry and Logic*, New York: Springer-Verlag.
- Macnamara, J. & Reyes, G., (eds.), 1994, *The Logical Foundation of Cognition*, Oxford: Oxford University Press.

- Makkai, M, & Reyes, G., 1977, *First-Order Categorical Logic*, Springer Lecture Notes in Mathematics 611, New York: Springer.
- Marquis, J.-P., 1993, "Russell's Logicism and Categorical Logicisms", *Russell and Analytic Philosophy*, A. D. Irvine & G. A. Wedekind, (eds.), Toronto, University of Toronto Press, 293-324.
- Marquis, J.-P., 1995, "Category Theory and the Foundations of Mathematics: Philosophical Excavations", *Synthese*, **103**, 421-447.
- McLarty, C., 1990, "Uses and Abuses of the History of Topos Theory", *British Journal for the Philosophy of Science*, **41**, 351-375.
- McLarty, C., 1992, *Elementary Categories, Elementary Toposes*, Oxford: Oxford University Press.
- McLarty, C., 1993, "Numbers Can be Just What They Have to", *No–s*, **27**, 487-498.
- Moerdijk, I. & Reyes, G., 1991, *Models for Smooth Infinitesimal Analysis*, New York: Springer Verlag.
- Tierney, M., 1972, "Sheaf Theory and the Continuum Hypothesis", *Toposes, Algebraic Geometry and Logic*, F.W. Lawvere (ed.), Springer Lecture Notes in Mathematics 274, 13-42.

Other Internet Resources

- Web page of McGill's "Centre de recherches en theorie des categories"
- Luca Mauri's page on Category Theory
- The category theory mailing list with many links and useful information

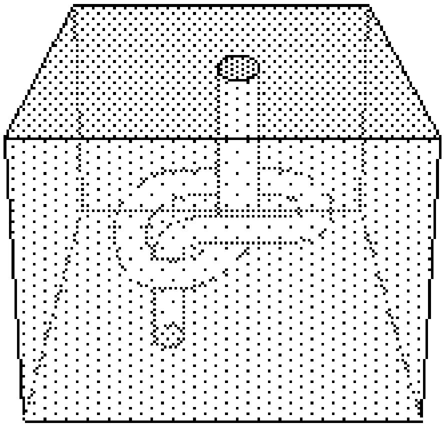
Related Entries

mathematics, philosophy of | set theory

Jean-Pierre Marquis
Jean-Pierre.Marquis@umontreal.ca

First published: December 6, 1996
Content last modified: July 15, 1997

Stanford Encyclopedia of Philosophy



Holes

Holes are an interesting case-study for ontologists and epistemologists. Naive, untutored descriptions of the world and explanations of facts in the world often make essential reference to holes. A hole explains why water flowed out of the reservoir. A colander wouldn't be what it is, without all those holes in it. It is because there is a hole that somebody has the impression of seeing a hole. Yet it might be argued that commitment to these entities is only illusory:

Problems with Holes

1. For any explanation of a physical interaction that might be offered in terms of holes, there must be some accompanying explanations that invoke material objects; but then, it seems that these explanations alone could be enough. That water flowed out of the reservoir is explained by a number of facts about water flow and the way it can be confined, and at no point in these explanations need the concept of a hole appear; instead, one might happen to talk of the shape of the reservoir.

2. Locke implied that a causal theory of perception is incompatible with the perception of holes; since

holes are not material they cannot be the source of any causal flow. (This might be considered an instance of the argument ad 1.) That we do have the impression of perceiving holes should then be considered a sort of systematic illusion. (Unless one rejects causal accounts of perception.)

Theories of Holes

If, on account of such concerns, holes are not taken at face value, a number of options are available.

> (a) Holes do not exist. This requires at least a systematic way of paraphrasing every hole-committing sentence by means of a sentence that does not refer to or quantify over holes. (The donut is holed, but there is no hole in it). Provided the language contains all the necessary shape-predicates, this might well be a favourable strategy: after all, holes are a paradigm example of nothings.
>
> (b) Holes exist, but they are something else. For instance, they are (parts of) material objects, say, hole-linings or hole-surrounds. This calls for an account of the altered meaning of certain predicates or prepositions. (What would ‘inside’ and ‘outside’ mean? What would it mean to ‘enlarge’ a hole?)

Or holes are negative, missing parts. On this account, a donut would be a sort of mereological sum of a pie and a mysterious missing bit in the middle. Or again, holes are not categorically homogeneous with their hosts. They are not particulars, but relations between a material object and a volume of space. (But now how can we account for the shape and size of holes? Relations do not have shapes and sizes.)

On the other hand, the possibility remains that holes be taken for what they are. They are full-fledged countable entities, like stones and chunks of cheese. But unlike stones and chunks of cheese, holes are ontologically parasitic: they are always in or through something else, and cannot be detached from their hosts. Holes are immaterial; localized at --but not identical with-- regions of space; fillable; and somehow causally liable. They are subject to part/whole structures. Yet holes are always in one piece--there is no such thing as half a hole.

Holes are topologically assorted: superficial hollows are distinguished from internal cavities; straight perforations are distinguished from knotted tunnels. But the hole realist will not fail to notice the unity of this assortment. These are all species of the same genus. Thus the underlying topology must depart from the basic account of handlebodies and calls for a direct analysis of their topological complements. Look at the donut, but keep an eye on the hole--or on what could fill it.

Bibliography

- Casati, Roberto, and Varzi, Achille. C.: Holes and Other Superficialities, Cambridge, MA, and London: MIT Press (Bradford Books), 1994.

- Lewis, David K., and Lewis, Stephanie R.: "Holes", Australasian Journal of Philosophy, 48 (1970), 206-212; reprinted in David K. Lewis, Philosophical Papers. Volume 1, New York and Oxford: Oxford University Press, 1983, pp. 3-9.
- Tucholsky, Kurt: "Zur soziologischen Psychologie der Löcher" (signed Kaspar Hauser), Die Weltbühne, March 17, 1931, p. 389; now in Gesammelte Werke, ed. by Mary Gerold-Tucholsky and Fritz J. Raddatz, Reinbek bei Hamburg: Rowohlt Verlag, 1960, Vol. 9, pp. 152-153 (English translation by Harry Zohn: "The Social Psychology of Holes", in Germany? Germany! The Kurt Tucholsky Reader; Manchester: Carcanet Press, 1990, pp. 100-101).

Other Internet Resources

[Please contact the authors with suggestions.]

Related Entries

ontology

Roberto Casati
casati@ehess.fr
and
Achille C. Varzi
achille.varzi@columbia.edu

First published: December 5, 1996
Content last modified: December 5, 1996

Private Language

The idea of a private language was made famous in philosophy by Ludwig Wittgenstein, who in section 243 of his book *Philosophical Investigations* explains it thus: ‘The words of this language are to refer to what can be known only to the speaker; to his immediate, private, sensations. So another cannot understand the language.’ [My translation.] This is not intended to cover (easily imaginable) cases of recording one's experiences in a personal code, for such a code, however obscure in fact, could in principle be deciphered. What Wittgenstein had in mind is a language conceived as *necessarily* comprehensible only to its single originator because the things which define its vocabulary are necessarily inaccessible to others.

Immediately after introducing the idea, Wittgenstein goes on to argue that there cannot be such a language. The importance of drawing philosophers' attention to a largely unheard-of notion and then arguing that it is unrealizable lies in the fact that an unformulated reliance on the possibility of a private language is arguably essential to mainstream epistemology, philosophy of mind and metaphysics from Descartes to versions of the representational theory of mind which have been prominent in late twentieth century cognitive science.

Section Links:

Overview: Wittgenstein's Argument and its Interpretations

Wittgenstein's main attack on the idea of a private language is contained in sections 244-271 of

Philosophical Investigations. These passages, especially those from section 256 onwards, are now commonly known as ‘the private language argument’, despite the fact that he brings further considerations to bear on the topic in other places in his writings.

The argument is quickly summarized. The conclusion is that a language in principle unintelligible to anyone but its originating user is impossible. The reason for this is that such a so-called language would, necessarily, be unintelligible to its supposed originator too, for he would be unable to establish meanings for its putative signs.

Nevertheless, there has been fundamental and widespread disagreement over the details, the significance and even the intended conclusion of the argument, let alone over its soundness. Some of this disagreement has arisen because of the notorious difficulty and occasional elusiveness of Wittgenstein's own text (sometimes augmented by problems of translation). For example, some philosophers have questioned the very existence in the relevant passages of a unified structure properly identifiable as a sustained argument. But much derives from the tendency of philosophers to read into the text their own preconceptions without making them explicit and asking themselves whether its author shared them. Some commentators, for instance, supposing it obvious that sensations are private, have interpreted the argument as intended to show they cannot be talked about; some, supposing the argument to be an obvious but unsustainable attempt to wrest special advantage from scepticism about memory, have maintained it to be unsound because it self-defeatingly implies the impossibility of public discourse as well as private; some have assumed it to be a direct attack on the problem of other minds; some have claimed it to commit Wittgenstein to behaviourism; some have thought it to imply that language is, of necessity, not merely potentially but actually social.

The early history of the secondary literature is largely one of disputation over these matters. Yet what these earlier commentators have in common is significant enough to outweigh their differences and make it possible to speak of them as largely sharing an Orthodox understanding of the argument. Since the publication in 1982 of Saul Kripke's definitely unOrthodox book, however, in which he suggested that the argument poses a sceptical problem about the whole notion of meaning, public or private, disputation conducted by Orthodox rules of engagement has been largely displaced by a debate on the issues arising from Kripke's interpretation. Both debates, though, show a tendency to proceed with only the most cursory attention to the original argument which started them off.

This rush to judgment about what is at stake, compounded by a widespread willingness to discuss commentators' more accessible accounts of the text rather than confront its difficulties directly, has made it difficult to recover the original from the accretion of more or less tendentious interpretation which has grown up around it. Such a recovery is one of the tasks attempted in this article.

[Return to Section Links]

The Significance of the Issue

The issue's significance can be seen by considering how the argument is embedded in the structure of *Philosophical Investigations*. Immediately prior to the introduction of the argument (sec. 241 f.), Wittgenstein suggests that the existence of the rules governing the use of language and making communication possible depends on agreement in human behaviour -- such as the uniformity in normal human reaction which makes it possible to train most children to look at something by pointing at it. (Unlike cats, which react in a seemingly random variety of ways to pointing.) One function of the private language argument is to show that not only actual languages but the very possibility of language and concept formation depends on the possibility of such agreement.

Another, related, function is to oppose the idea that metaphysical absolutes are within our reach, that we can find at least part of the world as it really is in the sense that any other way of conceiving that part must be wrong (cf. *Philosophical Investigations* p. 230). Philosophers are especially tempted to suppose that numbers and sensations are examples of such absolutes, self-identifying objects which themselves force upon us the rules for the use of their names. Wittgenstein discusses numbers in earlier sections on rules (185-242). Some of his points have analogues in his discussion of sensations, for there is a common underlying confusion about how the act of meaning determines the future application of a formula or name. In the case of numbers, one temptation is to confuse the mathematical sense of ‘determine’ in which, say, the formula $y = 2x$ determines the numerical value of y for a given value of x (in contrast with $y > 2x$, which does not) with a causal sense in which a certain training in mathematics determines that normal people will always write the same value for y given both the first formula and a value for x -- in contrast with creatures for which such training might produce a variety of outcomes (cf. sec. 189). This confusion produces the illusion that the result of an actual properly conducted calculation is the inevitable outcome of the mathematical determining, as though the formula's meaning itself were shaping the course of events.

In the case of sensations, the parallel temptation is to suppose that they are self-intimating. Itching, for example, seems like this: one just feels what it is directly; if one then gives the sensation a name, the rules for that name's subsequent use are already determined by the sensation itself. Wittgenstein tries to show that this impression is illusory, that even itching derives its identity only from a sharable practice of expression, reaction and use of language. If itching were a metaphysical absolute, forcing its identity upon me in the way described, then the possibility of such a shared practice would be irrelevant to the concept of itching: the nature of itching would be revealed to me in a single mental act of naming it (the kind of mental act which Russell called ‘acquaintance’); all subsequent facts concerning the use of the name would be irrelevant to how that name was meant; and the name could be private. The private language argument is intended to show that such subsequent facts could not be irrelevant, that no names could be private, and that the notion of having the true identity of a sensation revealed in a single act of acquaintance is a confusion.

The suggestion that a language could be private in the way described appears most openly in the second of Bertrand Russell's published lectures ‘The Philosophy of Logical Atomism’, where Russell says:

> In a logically perfect language, there will be one word and no more for every simple

> object, and everything that is not simple will be expressed by a combination of words, by a combination derived, of course, from the words for the simple things that enter in, one word for each simple component. A language of that sort will be completely analytic, and will show at a glance the logical structure of the facts asserted or denied. ... A logically perfect language, if it could be constructed, would not only be intolerably prolix, but, as regards its vocabulary, would be very largely private to one speaker. That is to say, all the names that it would use would be private to that speaker and could not enter into the language of another speaker.
>
> ... A name, in the narrow logical sense of a word whose meaning is a particular, can only be applied to a particular with which the speaker is acquainted, because you cannot name anything you are not acquainted with.
>
> ... One can use ‘this’ as a name to stand for a particular with which one is acquainted at the moment. We say ‘This is white’. ... But if you try to apprehend the proposition that I am expressing when I say ‘This is white’, you cannot do it. If you mean this piece of chalk as a physical object, then you are not using a proper name. It is only when you use ‘this’ quite strictly, to stand for an actual object of sense [i.e. a sense-datum], that it is really a proper name. And in that it has a very odd property for a proper name, namely that it seldom means the same thing two moments running and does not mean the same thing to the speaker and to the hearer.
>
> ... [I]n order to understand a name for a particular, the only thing necessary is to be acquainted with that particular. When you are acquainted with that particular, you have a full, adequate and complete understanding of the name, and no further information is required.

Although Wittgenstein does not explicitly say so, it is likely that this is the inspiration of his argument: his writing is marked in many places by criticism of Russell.

But the idea of a private language is more usually hidden: the confusions supposed to belong to it allegedly underlie a range of articulated philosophical notions and theories, without themselves being so articulated. The argument is thus perhaps most profitably read, not as refuting any particular theory, but as removing the motivation for considering a range of apparently independent or even competing theories along with their associated tasks, problems and solutions.

For example, a still very common idea, often attributed to John Locke and openly embraced by Jerry Fodor in the nineteen seventies, is that interpersonal spoken communication works by speakers' translation of their internal mental vocabularies into sounds followed by hearers' re-translation into their own internal vocabularies. Again, Descartes considered himself able to talk to himself about his experiences while claiming to be justified in saying that he does not know (or not until he has produced a reassuring philosophical argument) anything at all about an external world conceived as something

independent of them. And he and others have thought: while I may make mistakes about the external world, I can infallibly avoid error if I confine my judgments to my immediate sensations. (Compare *The Principles of Philosophy*, I, 9.) Again, many philosophers, including John Stuart Mill, have supposed there to be a problem of other minds, according to which I may reasonably doubt the legitimacy of applying, say, sensation-words to beings other than myself.

In each of these examples, the implication is that the internal vehicle of my musings could in principle be private: for these problems and theories even to make sense, sharability must be irrelevant to meaning and it must be at least conceivable that my knowledge, even my understanding, is necessarily confined to my own case. This is especially clear with Descartes: for his sceptical question to be raised without being immediately self-defeating, he must hold it possible to identify his experiences inwardly -- where 'inwardly' means without relying on resources supplied by his essential embodiment in a world whose existence is independent of his own mind and accessible to others (e.g. such resources as the concepts acquired in a normal upbringing). The question which accordingly looms large in the private language argument is: How is this identification of one's experiences to be achieved?

[Return to Section Links]

The Private Language Argument Expounded

Preliminaries

Having introduced the idea of a private language in the way already quoted, Wittgenstein goes on to argue in a preliminary discussion (sections 244-255) that there are two senses of 'private' which a philosopher might have in mind in suggesting that sensations are private, and that sensations as they are talked about in natural languages (such as English and German) are in fact private in neither of them. He then turns, at section 256, to the question whether there could be a private language at all. He continues to talk of sensations, and of pain as an example, but one should remember that these are not *our* sensations, the everyday facts of human existence, but the sensations of something like a Cartesian soul (perhaps one associated with a physical body, as indicated in sections 257 and 283), something which has no publicly available life and whose "experiences" are accordingly private -- that is, they are supposed exemplars of *philosophical accounts* of the everyday facts of human existence, not those facts themselves. So in section 256 Wittgenstein suggests that one cannot arrive at the idea of a private language by considering a natural language: natural languages are not private, for our sensations are expressed. But neither can we arrive at the idea by starting with a natural language and just subtracting from it all expression of sensations (temporary paralysis is clearly not in question), as he considers next, for as he says in section 257, even if there could be language in such a situation as this where teaching is impossible, the earlier argument of *Philosophical Investigations* (sections 33-35), concerning ostensive definition, has shown that mere "mental association" of one thing with another is not alone enough to make the one into a name of the other. Naming one's sensation requires a place for the new word: that is, a notion of sensation. The attempt to name a sensation in a conceptual vacuum merely raises the

questions of what this business is supposed to consist in, and what is its *point*. But, for the sake of getting to the heart of the matter, Wittgenstein puts the first of these questions on one side and pretends that it is sufficient for the second to imagine himself in the position of establishing a private language for the purpose of keeping a diary of his sensations.

However, to investigate the possibility of the imagined diary case by exploring it from the inside (the only way, he thinks, really to expose the confusions involved) requires him to use certain words when it is just the right to use these words which is in question. Thus he is forced to mention in section 258 actions like ostensive definition, concentrating the attention, speaking, writing, remembering, believing and so on, in the very process of suggesting that none of these can really be done in the situation under consideration (section 261).

This difficulty has often gone unnoticed by commentators on the argument, with particularly unhappy results for the understanding of the discussion of the diary example. Fogelin, for instance, a paradigm representative of Orthodoxy, treats this as a case where *he himself*, a living embodied human being, keeps a diary and records the occurrences of a sensation which he finds it impossible to describe to anyone else. But we are not to assume that the description of the keeping of the diary is a description of a possible or even ultimately intelligible case. In particular, we are not to think of such a human being's keeping a real diary, but of something like the Cartesian internal equivalent. It is thus vital to the argument that the diary case is presented in the first person, without our pressing the question, ‘Who is speaking?’ At this stage we are simply not to worry about whether the diary story ultimately makes sense or not. But the fact that it may not make sense must be remembered in reading what follows, which in strictness should constantly be disfigured with scare quotes. (I shall, as I have already, occasionally supply them as a reminder, reserving double quotes for this purpose.)

To summarize the argument's preliminary stage: In section 256 Wittgenstein asked of the "private language", ‘*How* do I use words to stand for my sensations?’, and reminded us in section 257 that we cannot answer ‘As we ordinarily do’. So this question, which is the same question as ‘How do I obtain meaning for the expressions in a "private language"?’ is still open; and the answer must be independent of our *actual* connections between words and sensations. In the attempt to arrive at an answer, and explore the question in its full depth, he temporarily allows the use of the notions of sensation and diary-keeping (despite the objections of section 257), and imagines himself in the position of a private linguist recording his sensations in a diary. The aim is to show that even if this concession is made, meaning for a sensation-word still cannot be secured and maintained by such a linguist. The crucial central part of the argument begins here, at section 258.

The Central Argument

Wittgenstein points out of the diary case ‘first of all that a definition of the sign cannot be formulated’. (The translation here obscures the reason why. Wittgenstein's word is ‘aussprechen’, better translated as ‘expressed’ than ‘formulated’: the point follows by definition from the fact that the case is one where the definition is private.) So if meaning is to be obtained for the "sign", this must be achieved through a

private exercise of ostensive definition, where I concentrate on the sensation and produce the sign at the same time. But if this exercise is to be genuine and successful ostensive definition, it must *establish* the connection between sign and sensation, and this connection must *persist*. As Wittgenstein says, ‘"I impress [the connection] on myself" can only mean: this process brings it about that I remember the connection *right* in the future’. For I do not define anything, even to myself let alone anyone else, by merely attending to something and making a mark, unless this episode has the appropriate consequences.

Interlude: the Rejection of Orthodoxy

At this point we should suspend our exposition of the argument, in order to examine closely the remark ‘this process brings it about that I remember the connection *right* in the future’.

This remark has usually been interpreted as a demand that, for the sign ‘S’ to have been given a meaning, it must always figure thereafter (if used affirmatively and sincerely) as a true statement: that is, I must use the sign ‘S’ affirmatively only when I really do have the sensation S. And it has usually been thought that the subsequent argument concerns the adequacy of memory to ensure that I do not later misidentify my sensations and call a different kind of sensation ‘S’ in the future. This account of the argument and its history is summed up by Anthony Kenny as follows:

> Many philosophers have taken ‘I remember the connection right’ to mean ‘I use "S" when and only when I really have S’. They then take Wittgenstein's argument to be based on scepticism about memory: how can you be sure that you have remembered aright when next you call a sensation ‘S’? ...
>
> Critics of Wittgenstein have found the argument, so interpreted, quite unconvincing. Surely, they say, the untrustworthiness of memory presents no more and no less a problem for the user of a private language than for the user of a public one. No, Wittgenstein's defenders have said, for memory-mistakes about public objects may be corrected, memory-mistakes about private sensations cannot; and where correction is impossible, talk of correctness is out of place. At this point critics of Wittgenstein have either denied that truth demands corrigibility, or have sought to show that checking is possible in the private case too. (Kenny, pp. 191-192)

This interplay of criticism and defence characterizes the Orthodox interpretation of the argument. (See Fogelin, pp. 162-4, for a good example.) There seem to be at least two reasons why this interpretation should have become established. First, philosophers committed to the idea of a private language are generally looking for an arrangement in which mistakes of fact are impossible; that is, they are trying to overcome scepticism by finding absolute certainty. (Descartes is the example usually cited.) And this would make sceptical arguments appear to be natural weapons to use in reply to them. (See, e.g., Fogelin p. 153.) Secondly, it is plausible -- which is not the same as correct -- to suppose that one cannot be mistaken concerning the natures of one's present sensations, and a supposed proof that the idea of a private language entails that one is just as fallible on this subject as on any other could thus seem

crippling to that idea.

But, as Kenny first showed, the question of factual infallibility in future uses of the sign ‘S’ is not the issue. If we look closely at section 258, we see that ‘I remember the connection *right*’ refers to remembering a *meaning*, namely, the meaning of the sign ‘S’, not to making sure that I infallibly apply ‘S’ only to S's in the future. (Nor does the private language argument depend on taking the latter to be an effect of the former.)

The Central Argument Continued

Now that we are clearer about what the connection is which has to be remembered right, we can return to the exposition of the argument. I am to imagine that I am a private linguist. I have a sensation, and make the mark ‘S’ at the same time, as I might in an ordinary case introduce a sign by ostensive definition. Afterwards, I "believe" myself to have established a meaning for this sign ‘S’, and I now use it to judge that I am again experiencing the same sensation. What do I mean by ‘S’ on this second occasion? Wittgenstein considers two possible answers.

The First Answer

One of the answers is that what I mean by ‘S’ is just the sort of sensation I am now having. Of this Wittgenstein says merely:

> ... whatever is going to seem right to me is right. And that only means that here we can't talk about ‘right’.

The point is highly condensed. Here is a more explicit version. For there to be factual assertion, there must be the distinction between truth and falsehood, between saying what is the case and saying what is not. For there to be the distinction between truth and falsehood, there must be a further distinction between the source of the meaning, and the source of the truth, of what is said. Suppose that I confront some object and say of it ‘This is S’. If I must also appeal to this very object to explain the meaning of the sign ‘S’, I deprive my initial utterance of any claim to the status of factual assertion -- it becomes, at best, ostensive definition. (The ‘at best’ is important here, for the same reason that the diary example is not to be assumed genuinely possible.)

The Second Answer

The second answer Wittgenstein considers to the question of what I mean by ‘S’ is this: I mean by ‘S’, not this current sensation, but the sensation I named ‘S’ in the past. We have already seen, in Kenny's rejection of the Orthodox reading of the argument, that scepticism about memory has no place in the discussion of "private language"; the text simply does not support it. But at this point we must break with Kenny too. For according to his account the crucial claim becomes: ‘If it is possible for me to misremember my previous ostensive definition of "S", then I do not really know what "S" means.’ (See,

e.g., Kenny p. 194.) This is just conventional scepticism about memory extended to include meanings as well as judgments. And it is an elementary point of epistemology that knowing something does not obviously entail just as a result of the definition of knowledge that it is *impossible* for one to be wrong about that thing, only that one is not in fact wrong.

What has gone wrong? The answer is that Kenny's and the Orthodox accounts share an unnoticed assumption: that even in the circumstances of the "private language" there is actually an application of a sign to a private sensation by a private linguist. The problem as Kenny then conceives it is one of later remembering this earlier application in order that ‘S’ should have retained its meaning. The question then seems to be whether one's admittedly fallible memory is adequate for the maintenance of meaning. But why should this assumption be allowed? What entitles us to assume that a private linguist could even ostensively define his sign to himself in the first place? As we have seen, this is one of the matters in question; and sections 260 and 261 show that Wittgenstein was not prepared to let an argument in favour of private language proceed from this assumption. In these two sections Wittgenstein reminds us that his arguments in the earlier sections (e.g. 33-35) of *Philosophical Investigations* showed that ostensive definition was not achieved by any performance unless certain circumstantial conditions are fulfilled; and nothing about the diary case as so far described shows them to be fulfilled. It is only later (sections 270-271) that Wittgenstein imagines a partial fulfilment of them, and the result there is to render the language public.

One cause of the muddle is Wittgenstein's insistence that there must be a distinction between obeying a rule and merely thinking that one has. This does not result, as the Orthodox have supposed, in a demand for, and eventual rejection of, ‘memory-infallibility in a private language’: demand and rejection being based respectively on the grounds that without infallibility one could always be going wrong and would never know if one were, and with infallibility one would collapse the distinction between obeying a rule and merely thinking one was obeying it. Rather, the argument is this. The private linguist cannot legislate a meaning for a sign by "private ostensive definition" *merely* -- for this has to establish a technique of using the sign (section 260). The technique cannot function by means of repeated "ostensive definitions", as we saw in examining the first answer, since this collapses the distinction between meaning and truth and thus destroys the possibility of making factual judgments. So the so-called "definition" has on some other basis to establish a constancy in use of the sign.

But this is just what is in question. What would be *constancy* here? What would *be* using the sign in the same way as before? How *was* the sign used in the first place? As there cannot be assumed to be a way of using the sign which the linguist succeeds even in determining, let alone establishing, and which is the correct way, independent of the linguist's later impression of the correct way, then a defender of "private language" would have to show that there was. It might now seem as if one could show this by appealing to the private linguist's memory. He simply remembers how he used the sign before. And this looks straightforward enough, because one thinks: he certainly did something before, for he remembers it. And we do not require his memory to be infallible. But the memory does at least have to be a *memory*: that is, accurate or not, it has to be of something determinate which existed independently of the memory of it; and the "memory" alone cannot bring such a thing into existence.

This is the argument of section 265, which has often been mistakenly given an epistemological interpretation. Again we cannot assume that there has been an actual table (even a mental one) of meanings in the case of the private linguist, a table which is now recalled and about which the linguist must rely on recall since the original has gone. Rather, as sections 260-264 show, there may be nothing determinate other than this "remembering of the table". So when we think that a private linguist could remember the meaning of ‘S’ by remembering a past correlation of the sign ‘S’ with a sensation, we are supposing what needs to be itself established -- that there was indeed some independent correlation to be remembered. Fallibility of memory, even of memory of meaning, is neither here nor there: the point is not that there is doubt *now* about the trustworthiness of memory, but that there was doubt *then* about the status of what occurred. And this original, *non*-epistemological, doubt cannot later be removed by "recollections" of a status inherently dubious in the first place. That is, if there was no genuine original correlation in the first place, a "memory" will not create one. But if, alternatively, we do not suppose that there was something independent of the memory to be remembered, again ‘what seems right is right’; the "memory" of the "correlation" is being employed to confirm itself, for there is no independent access to the "remembered correlation". (Not even the independent access that we have as posers of the example, since the question is, can we pose such an example? The typical mistake commentators make here is to disguise the problem by thinking of S in terms of some concept, such as *pain*, which they bring to the example themselves.) This is why Wittgenstein says (section 265), ‘As if someone were to buy several copies of the morning paper to assure himself that what it said was true’.

The Closing Stages

So far the argument has been conducted in terms of souls unrelated to bodies or related only to inert bodies. At section 269, however, it moves to examples where there is bodily behaviour but despite this there is still the temptation to think of private meanings for words independent of their public use. This suggests a further chance for a defender of the idea of a private language: that a private linguist might secure a meaning for his sign ‘S’ by correlating its private use with some public phenomenon. This would apparently serve to provide a function for the noting of ‘S’ in the diary (section 260) and thus give a place for ostensive definition, and would give as well a guarantee that there is some constancy in the linguist's use of the term ‘S’ independent of his impression of such constancy. Wittgenstein uses the example of the manometer in sections 270-271 to consider this idea, and his criticism of it is in effect that this method of securing meaning works, but that the secured meaning is public: the so-called "private object", even if there were such a thing, is revealed to be irrelevant to meaning. Presumably a defender of "private language" would hope that the example would work like this: if I keep saying, on the basis of my sensation, that my blood pressure is rising, and the manometer shows that I am right, then this success in judging my own blood pressure shows that I had in fact established a private meaning for the sign ‘S’ and was using the sign in the same way each time to judge that my sensation was the same each time. However, all the example really shows is that just thinking that I have the same sensation now as I had when my blood pressure rose formerly, can be a good guide to the rising of my blood pressure. Whether in some "private sense" the sensation was "actually the same" or not becomes completely irrelevant to the question of constancy in the use of ‘S’ -- that is, there is no gap between the actual nature of the sensation and my impression of it, and ‘S’ in this case could mean no more than ‘sensation of the rising of the blood pressure’; indeed, for all we are told of the sign's role, it could even mean

merely ‘blood pressure rising’.

Are the Orthodox Objections Met?

Does the ruling out of memory-scepticism as irrelevant to the private language argument mean that two associated Orthodox objections to it are likewise irrelevant? The first of these is that the argument, self-defeatingly, rules out a public language as well. The second is that the argument, equally self-defeatingly, rules out as impossible something perfectly conceivable: namely, the case of a so-called ‘Robinson Crusoe’, a human being who, unlike Defoe's original Crusoe, is isolated from birth but devises a language for his own purposes without his having first been taught another language by someone else. Wittgenstein's Orthodox defenders, faced with this second objection, looked to be on shaky ground, often being forced into the position of conceding that the argument did indeed exclude the case, but claiming (not very plausibly) that such a Crusoe is after all impossible so that the concession was not damaging.

The question as it concerns the first objection has already been answered. The supposed threat to public language arose entirely from the the claim that memory-scepticism could not be confined to the private case. But since scepticism concerning memory is no part of the argument, there is no reason to suppose that any question of such confinement arises, and thus there is no question of the argument's being self-defeating by excluding the possibility of something we know to be actual, i.e. the language we already have. Now showing the absence of any appeal to memory-scepticism involved transferring the burden of the argument from the question of whether or not an ostensive definition could be remembered or not to the question of whether there could be an ostensive definition in the first place.

This enables us to answer the question as it concerns the second objection. It is clear that an argument which has as its focus the question of ostensive definition is not committed to ruling out in advance all hypothetical cases of ‘Robinson Crusoes’. For there is no a priori barrier to imagining a form of life complex enough for us to be assured that a determinate ostensive definition had been accomplished by such a being. Such a Crusoe, unlike a private linguist, lives in a world independent of his impressions of it, and thus there could be definite occurrences in it which he could remember or forget; and some of those occurrences could be correlations of signs with objects. (There are, however, further complications here. See Canfield 1996.)

[Return to Section Links]

Kripke's Sceptical Wittgenstein

The Orthodox domination of the secondary literature on private language was largely ended by Saul Kripke's account of Wittgenstein's treatment of rules and private language, in which Wittgenstein appears as a sceptic concerning meaning. Kripke (p. 5) denies commitment to the identity of this sceptical figure with its historical source, and, appropriately, his account has spawned a literature of its own in which discussion often proceeds largely independently of the original private language argument: Kripke's

Wittgenstein, real or fictional, has become a philosopher in his own right, and for many people, it is not an issue whether the historical Wittgenstein's original ideas about private language are faithfully captured in this version. The complexities of the subsequent discussion of the philosophical -- as opposed to interpretative -- questions raised by Kripke's Wittgenstein need a separate article to themselves. (For a survey, see Boghossian.) All that will be settled here is the interpretative question.

Kripke's account resembles that given here in its rejection of Orthodoxy and in its emphasis on the logical priority of the discussion of rule-following to that of private language. It differs in the prominence it gives to the opening sentence of *Philosophical Investigations* section 201: ‘This was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule.’ Kripke says of this (p. 68), ‘The impossibility of private language emerges as a corollary of [Wittgenstein's] sceptical solution of his own paradox’. Wittgenstein himself immediately brushed this "paradox" aside in his very next paragraph: ‘It can be seen that there is a misunderstanding here ...’; but Kripke takes the paradox to pose a genuine and profound sceptical problem about meaning.

The example Kripke chooses to illustrate the problem is that of addition. What is it to grasp the rule of addition? The application of the rule is potentially infinite, and bizarre interpretations of the rule, as well as the standard one, are compatible with any finite set of applications of the usual sort such as 7 + 14 = 21. So what is it which makes it true that when I say ‘plus’ I mean the usual addition function and not some other? Kripke understands this question as containing a Humean problem to which, he claims, Wittgenstein gives a Humean, ‘sceptical’ solution. Kripke formulates the problem in two different ways.

The first way is this: ‘there is no fact about me that distinguishes my meaning a definite function by "plus" ... and my meaning nothing at all’ (p. 21). The absence of this fact, in Kripke's view, leads Wittgenstein to abandon the explanation of the meanings of statements like ‘By "plus", I meant addition’ in terms of truth-conditions, and to replace it with explanation in terms of assertibility-conditions, which involve actual (not merely potential) community agreement. (Hence the claim that this is a ‘sceptical solution’: Wittgenstein is supposed to concede to the sceptic the absence of truth-conditions for such statements.) This agreement, on Kripke's account, legitimizes the assertion that I meant addition by ‘plus’ despite there having been no fact of the matter.

This requirement of community agreement for meaning obviously rules out the possibility of private language immediately, thereby making the argument of *Philosophical Investigations* sections 256-271 superfluous. This superfluity makes for an odd reading of the text; and the oddness is highlighted by the observation that this first formulation of the sceptical problem relies on Kripke's assumption that we have some idea of what a fact is, independent of a statement's being true. For one of the themes of *Philosophical Investigations* is that there is no such idea, that the only route to the identification of facts is through the uses of the expressions in which those facts are stated, uses which give us the truth-conditions. These uses are often very different from what we would expect -- hence the impression that truth-conditions are lacking -- and it is a matter of some philosophical difficulty to see them clearly.

The other formulation of the problem is this (Kripke p. 62): ‘Wittgenstein questions the nexus between

past "intention" or "meanings" and present practice: for example, between my past "intentions" with regard to "plus" and my present computation ... .' The idea is that my grasp of the rule governing the use of 'plus' does not determine that I shall produce a unique answer for each of indefinitely many new additions in the future. The impression that something is missing here, though, is a result of just that kind of confusion about determination identified in the section headed 'The Significance of the Issue' above.

Kripke's account of the private language argument is thus vitiated by his unargued reliance on ideas which Wittgenstein argued against. This of course does not show that he has not hit upon a new and more interesting notion of private language than that expounded here.

[Return to Section Links]

Bibliography

- Boghossian, P.A. (1989) 'The rule-following considerations', *Mind* 98 (392): 507-49.
- Candlish, S. (1997) 'Wittgensteins Privatsprachenargumentation', in E. von Savigny (ed) *Wittgensteins Philosophische Untersuchungen*, Berlin: Akademie Verlag.
- Canfield, J.V., ed., (1986) *The Philosophy of Wittgenstein, Volume 9: The Private Language Argument*, New York: Garland.
- Canfield, J., ed., (1986) *The Philosophy of Wittgenstein, Volume 10: Logical Necessity and Rules*, New York: Garland.
- Canfield, J. (1991) 'Private language: *Philosophical Investigations* section 258 and environs', in R. Arrington and H.-J. Glock (eds) *Wittgenstein's Philosophical Investigations: Text and Context*, London and New York: Routledge, pp. 120-37.
- Canfield, J. (1996) 'The community view', *The Philosophical Review*, 105, pp. 469-488.
- Fogelin, R.J. (1976) *Wittgenstein*, London: Routledge.
- Hacker, P.M.S. (1990) *Wittgenstein: Meaning and Mind, Volume 3 of an Analytical Commentary on the Philosophical Investigations*, Oxford: Blackwell, pp. 1-286.
- Jones, O.R., ed., (1971) *The Private Language Argument*, London: Macmillan.
- Kenny, A. (1973) *Wittgenstein*, London: Allen Lane, Ch. 10.
- Kripke, S. (1982) *Wittgenstein on Rules and Private Language*, Oxford: Blackwell.
- Russell, B. (1918) 'The philosophy of Logical Atomism' in *The Collected Papers of Bertrand Russell, Volume 8: The Philosophy of Logical Atomism and Other Essays 1914-19*, London: George Allen and Unwin, 1986.
- Winch, P. (1983) 'Facts and superfacts', *The Philosophical Quarterly* 33 (133): 398-404, reprinted 1987 (slightly altered) in his *Trying to Make Sense*, Oxford: Blackwell, pp. 54-63.
- Wittgenstein, L. (1953) *Philosophical Investigations*, translated by G. E. M. Anscombe, 3rd edition, 1967, Oxford: Blackwell.
- Wittgenstein, L. (1934-6) 'Notes for lectures on "private experience" and "sense-data"', *The Philosophical Review* LXXVII (1968): 275-320. (Also included in Jones 1971.)

[Return to Section Links]

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

consciousness | Russell, Bertrand | Wittgenstein, Ludwig

Stewart Candlish
candlish@arts.uwa.edu.au

First published: July 26, 1996
Content last modified: September 28, 1998

Stanford Encyclopedia of Philosophy

How to Cite This Encyclopedia

To cite the *Stanford Encyclopedia of Philosophy*, we recommend the following bibliographic format, which you may need to adapt to the meet the style requirements of the publication for which you are writing.

Typically, users read the current version of each online entry. This is the version you reach directly from our Table of Contents. However, because Encyclopedia entries are subject to periodic revision, it is more appropriate to cite the most recent archived version. When you quote from an entry, it is particularly important to cite an archived version. Be sure to verify that the passage you are quoting matches what is in the archived version.

If the material you wish to cite has not been archived (because the entry is new or has been recently modified), you should, if possible, wait for the next archived version of the Encyclopedia. Fixed editions of the Encyclopedia are created and archived every three months, on the 21st of September, December, March, and June.

As an example, you would cite Andrew Irvine's entry on Bertrand Russell by finding the most recent archived edition containing this entry. Your citation might look like this:

> Irvine, A., "Bertrand Russell", *The Stanford Encyclopedia of Philosophy* (Fall 1999 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall1999/entries/russell/

Note that the above URL ends in the name "russell" followed by a slash "/". "russell" is the name of the directory which contains the entry. The filename of the entry itself is "index.html" (indeed, all the entries in the Encyclopedia are named "index.html"). However, you need not include the entry filename at the end of the URL. By default, index.html is displayed when a web browser requests a URL such as the one displayed above.

If you require other kinds of bibliographic information, you may find the following facts helpful.

- Title: *The Stanford Encyclopedia of Philosophy*
- Principal Editor: Edward N. Zalta
- World Wide Web URL: http://plato.stanford.edu/
- Publisher:

  The Metaphysics Research Lab
  Center for the Study of Language and Information

Stanford University
Stanford, CA 94305-4115

- International Standard Serial Number: ISSN 1095-5054

If you have any further questions, you may write to:

Principal Editor
Stanford Encyclopedia of Philosophy
CSLI/Ventura Hall
Stanford University
Stanford, CA 94305-4115

or send email to

editors@plato.stanford.edu